

TANGIBLE VISUAL ANALYTICS: THE INTEGRATION OF TANGIBLE INTERACTIONS
AND COMPUTATIONAL TECHNIQUES FOR BIOLOGICAL DATA VISUALIZATION AND
MODELLING WITH EXPERTS IN THE LOOP

by

Roozbeh Manshaei

M.Sc. K.N. Toosi University of Technology, Tehran, Iran 2009

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2017

© Roozbeh Manshaei 2017

Author's Declaration

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Abstract

Tangible Visual Analytics: the integration of tangible interactions and computational techniques for biological data visualization and modelling with experts in the loop

Roozbeh Manshaei

Doctor of Philosophy, Electrical and Computer Engineering

Ryerson University, 2017

Understanding and interpreting the inherently uncertain nature of complex biological systems, as well as the time to an event in these systems, are notable challenges in the field of bioinformatics. Overcoming these challenges could potentially lead to scientific discoveries, for example paving the path for the design of new drugs to target specific diseases such as cancer, or helping to apply more effective treatment for these diseases. In general, reverse engineering of these types of biological systems using online datasets is difficult. In particular, finding a unique solution to these systems is hard due to their complexity and the small sample size of datasets. This remains an unsolved problem due to such uncertainty, and the often intractable solution space of these systems.

The term "uncertainty" describes the application-based margin of significance, validity, and efficiency of inferred or predictive models in their ability to extract characteristic properties and features describing the observed state of a given biological system. In this work, uncertainties within two specific bioinformatics domains are considered, namely "gene regulatory network reconstruction" (in which gene interactions/relationships within a biological entity are inferred from gene expression data); and "cancer survivorship prediction" (in which patient survival rates are predicted based on clinical factors and treatment outcomes). One

approach to reduce uncertainty is to apply different constraints that have particular relevance to each application domain. In gene network reconstruction for instance, the consideration of constraints such as sparsity, stability and modularity, can inform and reduce uncertainty in the inferred reconstructions. While in cancer survival prediction, there is uncertainty in determining which clinical features (or feature aggregates) can improve associated prediction models. The inherent lack of understanding of how, why and when such constraints should be applied, however, prompts the need for a radically new approach.

In this dissertation, a new approach is thus considered to aid human expert users in understanding and exploring inherent uncertainties associated with these two bioinformatics domains. Specifically, a novel set of tools is introduced and developed to assist in evidence gathering, constraint definition, and refinement of models toward the discovery of better solutions. This dissertation employs computational approaches, including convex optimization and feature selection/aggregation, in order to increase the chances of finding a unique solution. These approaches are realized through three novel interactive tools that employ tangible interaction in combination with graphical visualization to enable experts to query and manipulate the data. Tangible interaction provides physical embodiments of data and computational functions in support of learning and collaboration. Using these approaches, the dissertation demonstrates: (1) a modified stability constraint for reconstructing gene regulatory network that shows improvement in accuracy of predicted networks, (2) a novel modularity constraint (neighbor norm) for extracting available structures in the data which is validated with Laplacian eigenvalue spectrum, and (3) a hybrid method for estimating overall survival and inferring effective prognosis factors for patients with advanced prostate cancer that improves the accuracy of survival analysis.

"If you can't explain it simply,
you don't understand it well enough."

– Albert Einstein

To my parents Akbar and Shahpar
and my lovely wife for their endless love and support ...

Acknowledgements

Every thesis is a journey, and I would like to thank all the people that accompanied and supported me during mine. I would first like to express my sincere gratitude to my advisors, Dr. Matthew Kyan and Dr. Ali Mazalek, who have been incredible mentors guiding me in my research. I could not have imagined having this thesis completed without their immense knowledge and expert suggestions.

During my thesis, I had the chance of working together with many exceptional people. Most notably, I would like to thank my colleagues at the Synaesthetic Media lab (Synlab) for their constructive criticism and helpful suggestions regarding this work as well as for their support. I thank my friends and co-authors Sean Delong, Shahin Khayyer, Uzair Mayat, Justin Digregorio, Masoud Hashemi for sharing all the laughs, struggles, good times, and midnight coding hours. Their support and the successful cooperation formed the very basis of this thesis.

Last but not least, my deepest gratitude goes to my parents and my wife who have provided me with their love and affection and have believed in my abilities. They been a pillar of strength behind me through the years allowing me to focus and achieve my goals.

Toronto, Ontario, Canada, 2017

R. M.

Contents

Abstract	iii
Acknowledgements	vi
List of tables	xi
List of figures	xiii
1 Introduction	1
1.1 Research Overview	1
1.2 Research Motivation	2
1.3 Research Challenges	3
1.3.1 Challenge 1 - Gene regulatory networks reconstruction	3
1.3.2 Challenge 2 - Prognosis Models Reconstruction	11
1.3.3 Challenge 3 - Grasping Trends in Biological Systems Modeling	12
1.4 Approaches to Face the Challenges	13
1.4.1 Approaches to Encountering Challenge 1	13
1.4.2 Approaches to Encountering Challenge 2	15
1.4.3 Approaches to Encountering Challenge 3	17
1.5 Research Contributions	19
1.6 Outline of the Dissertation	24
2 Background and Foundations	26
2.1 Visual Analytics	26

Contents

2.1.1	Visual Analytics vs. Information/Data Visualization	28
2.1.2	Visual Analytics related fields	29
2.1.3	Visual Analytics Data Structures	35
2.2	Situation Awareness, sense-making and mental models	37
2.2.1	Situation Awareness Definition	37
2.2.2	Sense-making Definition	38
2.2.3	Mental Models Definition	41
2.3	Tangible User Interfaces (TUIs)	43
2.3.1	Graphical User Interfaces vs. Tangible User Interfaces	43
2.3.2	Interactions	44
2.3.3	TUIs Survey - Strengths and Limitations	46
3	Model Refinement and Inference, and Interactive Biological Networks Visualization	49
3.1	Stability Constraint Modification	50
3.1.1	Reconstruction Algorithm	52
3.1.2	Results and Discussion	60
3.2	Structural Norm Minimization based on Neighbourhoods	66
3.2.1	Problem Formulation	67
3.2.2	Result	68
3.3	Tangible Biological Networks	68
3.3.1	Exploration of Gene Interaction	72
3.3.2	Related Works	73
3.3.3	Tangible Biological Networks	75
3.3.4	Basic Interactions and Visualization	76
3.3.5	Query Construction	80
3.3.6	Technical Implementation	90
4	Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization	94

4.1	A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC	95
4.1.1	Materials and Method	96
4.1.2	Results	107
4.1.3	Discussion	122
4.2	Graspable Prognosis Factors Visualization for Interpreting Cancer Data	126
4.3	Materials and Design	128
4.3.1	Data Description	134
4.3.2	Analysis, Modeling and Visualization	135
4.4	Discussion	159
5	Tangible Tensors	162
5.1	Tangible Tensors	163
5.1.1	Related Works	164
5.1.2	Research Rationale	167
5.1.3	Toolkit Design	169
5.1.4	Active Pathways Case Studys	179
5.1.5	Discussion	180
6	Conclusion and Future Directions	184
6.1	Summary of Contributions and Lessons Learned	184
6.2	Future Work	188
	Appendices	191
A	Multi-touch Surfaces	192
A.1	Purchasing a Multi-Touch Screen	192
A.1.1	Infrastructure Requirements	193
A.1.2	Wall Design	193
A.1.3	3-tiles Tables	194
A.2	Windows Computer for interfacing the MultiTaction Cells	194

Contents

A.3 Cell Setup	195
B Tangible Devices	199
B.1 Active Tangibles	199
B.1.1 Connecting Smartwatches to the Network	200
C Communication	203
C.1 Required Software	203
C.1.1 Installing XAMPP Control Panel	204
C.1.2 Installing Apache through XAMPP	205
C.1.3 Installing Node.js	206
C.1.4 Running the Node.js TCB Server	206
C.1.5 Changing Mobile Menu socket.io Addresses	207
Bibliography	209

List of Tables

3.1	GRN Reconstruction Algorithm.	60
3.2	Comparison of our algorithm performance by other methods in literature in detecting interactions among experimentally known gene interactions.	64
3.3	Comparison of the proposed algorithm with other methods using statistical criteria.	66
4.1	Clinical and pathological parameters of the study cohort and comparative analysis results.	112
4.2	Univariate and multivariate our hybrid method results for evaluating predictors associated with overall survival	114
4.3	Univariate and multivariate results based on LDH predictor imputation on overall survival	119
4.4	The C-Index, AUC at 12, 18, 24 months, and iAUC for Halabi et al.[Halabi <i>et al</i> , 2014] model and also our methods including and excluding LDH imputation.	121
4.5	The prognostic factors with p_Value < 0.05 for CRPC patients reported in previous studies.	123
4.6	Feature dataset that includes the clinical and pathological features of all patients. Each row is assigned to a single patient and each column to a single feature. . .	135
4.7	Censoring dataset that includes the censoring information of the events and also time to events. Rows show the patients ID and columns represents censoring information and time to events.	136

List of Tables

4.8	Normal range information for clinical features. This dataset includes two rows and the columns show all features. If there is no defined normal range or it is not available for some features, user should consider zero values for those specific features.	136
4.9	Features pre-clustering information that classifies the features in different categories (e.g. lab value, lesion measure, etc.). This dataset consists of two rows. The first row lists all the features present in the feature dataset, and the second row contains the pre-clustering label for each feature. If the user does not define any pre-clustering labels, those features will be considered individually as single un-clustered features.	137
4.10	Systematic breakdown of the user experience in Phase 1 of the TMVV application.	139

List of Figures

1.1	a) The schematic of chromosome and genes location (left), b) Propagation of cellular information [GeneticsRef., 2016].	4
1.2	Transcription (left) and translation (right) sub-processes [GeneticsRef., 2016]. .	4
1.3	Machinery of mRNA production and utilization and the used simple model in GRNs.	5
1.4	GRN representation that is compressed in time.	6
1.5	DNA microarray process.	7
1.6	Time series microarray data [Parmar, 2013].	8
1.7	Flow of interactive modeling and visualization design related to Tangible Biological Networks (TBNs), Tangible MultiVariate Visualization (TMVV), Tangible Tensors (TTs) platforms in dissertation	21
2.1	Visual analytics cycle	29
2.2	GUI and TUI interaction models - a) MVC model; b) MCRpd model [Ullmer, 2000]	44
3.1	A sample problem segmentation.	54
3.2	Weighted l_1 – <i>norm</i> relaxation interpretation and its solution: To illustrate the property of weighted l_1 – <i>norm</i> relaxation, let us consider the problem of finding the sparsest interaction network that provides a given level of performance $\epsilon^{(p)} \geq 0$. The solution of the relaxed formulation is the intersection of the constraints set $\Omega = \{W^{(p \rightarrow p+1)} \Lambda(W^{(p \rightarrow p+1)}) \leq \epsilon^{(p)}\}$ and the smallest sub-level set of $\xi = \sum_{i,j=1}^n b_{ij} w_{ij} $ that touches Ω	56

List of Figures

3.3	Regions for D_j , $j = 1, \dots, 4$ for the eigenvalues of an assumed stable matrix $W \in R^{4 \times 4}$	59
3.4	An example of inferred sub-networks in $m - 1$ ranges of time. *Symbols " \rightarrow " and " $-o$ " illustrate activator and repressor interactions, respectively.	61
3.5	As a sample, in this figure, inferred genetic interaction network from sub-network between time point 28 and 29 for twelve yeast cell cycle regulatory genes is shown. Each node represents a gene and the presence of an edge between the two nodes represents the existence of interaction between the two genes. Symbols " \rightarrow " and " $-o$ ", shown by blue and red edges, illustrate activator and repressor interactions, respectively. Dashed edges represent interactions that have been verified. In contrast, dotted edges are incorrect extracted interactions.	63
3.6	(A) Blocks of the matrix, (B) The clusters shown on a graph, (C) Noisy clusters, (D) Noisy adjacency matrix, (E) Recovered by method in [Richard <i>et al</i> , 2012], and (F) Recovered by our proposed method.	69
3.7	Laplacian spectrum obtained (A) by method in [Richard <i>et al</i> , 2012], (B) by our proposed method.	70
3.8	Active tangible interactions: a) users start to select the interested organism, swipe to select the type of exploration, gene-based, linked based and network based, b) users select one of them by tapping, which displays a list of developed and related filters.	77
3.9	When an active tangible is placed on the table after selecting the organism or gene parameter, the system displays all available networks involving that organism or gene, respectively, in a doughnut chart. The outer slice is a magnifier. When users point the tangible towards a zone like a dial, the system magnifies the respective slices for better visualization.	78
3.10	A screenshot of the visualization for gene networks within four organisms in force-directed graphs.	79
3.11	The eigenvalue space related to HIV organism.	82
3.12	The multiplicity of laplacian eigenvalues in HIV network.	83

3.13 The comparison of eigenvalues space and their multiplicities.	84
3.14 The removal effect of genes with various degree on network energy in E.coli organism.	86
3.15 The eigenvector centrality spectrum for HIV network.	87
3.16 Eigenvector centrality representation by changing nodes' size on Bos Taurus network.	87
3.17 Assortativity plot for undirected, in-in, in-out, out-in, and out-out interactions in Gallus Gallus organism.	89
3.18 Clustering Coefficient related to each gene is shown by changing the size of each node on the tabletop/wall. Also, the network clustering coefficient that is the average of this metric for all genes is illustrated on mobile menu on active tangible.	90
4.1 Comparison of missing factors within datasets D1, D2, D3 and D4. The datasets under comparison are indicated in the legend, with the colour assignment indicating the features missing within compared datasets, as indicated in the vertical axis as f_i where i is the index of the clinical predictor. In addition, the number of missing features is indicated in the horizontal axis below each bar. For each dataset combination, dark blue indicates common features between the datasets.	99
4.2 a) Logarithmic distribution of continuous clinical features within the possible clinical low, medium and high range for databases Celgene (D2), Sanofi (D3), and AstraZeneca (D4). b) Deviation of the clinical values from the normal range, where values within a normal range are assigned a value of 0, values above the normal range are assigned a positive value, and values below the normal range are assigned a negative value. The features under review are as follows: CA, AST, ALP, ALB, ALT, CREAT, HB, WBC, TPRO, NEU, PHOS, PLT, and TBILI.	103
4.3 Rules in phase one of factors selection for dimensionality reduction based on restricting model's $p_values < 0.05$	105
4.4 Graphical representation of the feature selection process detailed in phase 1. Red boxes represent the extracted feature in each layer.	106

List of Figures

- 4.5 Rules in phase two of factors selection for dimensionality reduction based on maximizing model's accuracy 108
- 4.6 Graphical representation of the feature selection process detailed in phase 2. Purple boxes represent the extracted feature in each layer. 109
- 4.7 a) Results of phase one of the method, clinical feature combination vs log of the p_value. Feature combinations extracted from phase one were inputted into phase two of the method shown in (b). b) Results of phase two of the method, ALP is selected in first layer as most effective factor for obtaining accuracy. Additional prognosis factors were appended onto ALP until an increase or plateau in accuracy. c) Graphical representation clinical and pathological factors space depicts the consideration of p_value and accuracy for highlighting the selection of various clinico-pathological factors combinations. Tracked black line shows our method's path in determining the final selected combination. 111
- 4.8 Kaplan–Meier curves for overall survival (OS) according to ALP, DIFF_ALP, AST and Liver Metastases. a) The blue and red lines indicate survival for patients with $0 < \text{ALP} < 83$ and $83 < \text{ALP} < 130.5$, respectively, the orange and gray lines indicate patients' OS with $130.5 < \text{ALP} < 266$, $266 < \text{ALP} < 398.3$, respectively; b) we stratified the patients into three cohorts with different ranges of ALP named DIFF_ALP; OS according to low, normal, and high are indicated by blue, red, and orange lines; c) the patients were divided into four subgroups according to the AST; the blue, red, and orange lines indicate overall survival of three cohorts with $0 < \text{AST} < 32$, and the purple line indicates patients with $32 < \text{AST} < 328$; d) liver metastases (blue line) is associated with shorter OS time in comparison with no liver leisure (red line). 115
- 4.9 Kaplan-Meier curve for OS according to DIFF_ALB and DIFF_HB. a) We stratified the patients into three cohorts with low, normal, and high ALB. The blue red, and orange lines indicate survival for patients with low, normal, and high, respectively; b) the patients with normal range of HB (shown in red) has a shorter OS than those with lower (shown in blue). 116

- 4.10 Kaplan-Meier survival curves are shown for the patients according to prior use of ACE inhibitors, Adrenal metastases, presence of lymph node lesion, and ECOG. a) OS according to the prior use and no prior use of ACE inhibitors which are indicated with blue and red respectively; b) the patients with adrenal metastasis (shown in blue) has a shorter OS than those without (shown in red); c) the patients based on the lymph node metastases size are categorized in three cohorts (no lesion, <52mm, >52mm) which their OS are indicated by blue, red, and orange lines respectively; d) The patients are stratified into three ECOG classes 0, 1, and 2 (blue, red and orange curves) which represent low, intermediate and high risk, respectively. 117
- 4.11 Kaplan-Meier survival curves according to LDH factor for all the training and testing samples. The patients are divided into four quartile groups according to the LDH. The OS for patients with a LDH>278 (gray curve) were significantly shorter than those in other quartile groups (blue, red and orange curves). . . . 120
- 4.12 Flowchart illustrating the rationale architecture for TMVV research and design. 129
- 4.13 Phase 1 flowchart. Directed arrows show the interaction path of the user through Phase 1 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters a new phase of the TMVV platform. 131

List of Figures

4.14 Phase 2 flowchart. Directed arrows show the interaction path of the user through Phase 2 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters or returns to a new phase of the TMVV platform.	132
4.15 Phase 3 flowchart. Directed arrows show the interaction path of the user through Phase 3 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters or returns to a new phase of the TMVV platform.	133
4.16 Visual depiction of the user interfaces (menus) that the user will be able to interact with within Phase 1 of the TMVV application.	138
4.17 Menu d of Figure 4.12. The user is being asked to select the underlying database (prostate or breast cancer). The local host is representative of the interactive tabletop which is blank at this stage because no analysis is yet to be performed.	140
4.18 Pre-clustered zoomable feature selection platform.	141
4.19 The various analytical tools are shown on the application interface. The implemented Kaplan-Meier plot and Box-Plot are shown. Note the easily identifiable outliers on the Box-Plot.	145

4.20 Implemented Kaplan-Meier plot and Box-Plot following categorization by normal range. Alternative categorizations are shown in menu I of Figure 4.16. . . .	146
4.21 An example for multivariate correlation scatterplot [Zerbini, Alexandre N., et al. "Baleen whale abundance and distribution in relation to environmental variables and prey density in the Eastern Bering Sea." Deep Sea Research Part II: Topical Studies in Oceanography (2015)].	147
4.22 Right- Implemented cluster visualization of related features; Left- Multivariate correlation scatterplot.	149
4.23 The multivariate correlation scatterplot can be selected through the interactive tabletop to reveal the underlying correlation parameter of each scatterplot. . .	150
4.24 Univariate p-value and hazard ratio for a single feature.	151
4.25 Visual depiction of the various Phase 2 user interfaces. All the menus available for user interaction are shown.	153
4.26 Hovering over the feature interaction network reveals the selection rationale for each feature node.	154
4.27 Right - Feature iteration summary; Left - Feature solution space p-value matrix. Notice the selected features as small dots on the solution space.	156
4.28 Right - Feature iteration summary; Left - Feature solution space accuracy matrix. Notice the selected features as small dots on the solution space.	157
4.29 Left - Feature solution space p-value matrix; Right - Feature iteration summary. Notice the variation of significant features over multiple iterations.	158
4.30 Visual depiction of the various phase 3 user interfaces. All the menus available for user interaction are shown.	159
4.31 Right - Global hazard risk displayed on smartphone; Left - Hazard risk over time displayed on tabletop platform.	160
5.1 General flowchart showcasing the Tangible Tensors toolkit.	168

List of Figures

5.2	(a) General view of the project screen in the tangible tensors platform. Overview of the multi-section arrangement; (b) Functions available to users in main menu of toolkit as displayed on the active tangible.	171
5.3	(a) Overview of the tensor menu screen. All the user tensors are displayed; (b) Visualization of the classification function. Tensors belonging to groups of users such as Biomedical Engineers (BE), computer engineers (CE) or Biologists (B) are all clustered together; (c) Visualization of the sort tensor function; (d) Functions available to users in the tensor menu of toolkit as displayed on the active tangible.	173
5.4	(a) Visualization of the select frame function; (b) Visualization of the delete frame function; (c) Visualization of the select molecule function; (d) Visualization of the rotate tensor function; (e) Visualization of the sort tensor frames function; (f) Functions available to users in the tensor screen of toolkit as displayed on the active tangible.	175
5.5	(a) Main menu of the pathways project when first opened; (b) Following the manipulation of parameters, the plots will show the experimental results in real time and generate an iteration of the user's tensor in the tensor bar.	177
5.6	(a) Visualization of the user classification function; (b) Visualization of the delete tensor function; (c) Visualization of the sort tensor function.	181
5.7	(a) Visualization of the select frame function; (b) Visualization of the delete frame function; (c) Visualization of the select molecule function; (d) Visualization of the sort frames function; (e) Visualization of the rotate tensor function.	182
A.1	Front and top views of 3-cell wall section angles	194
A.2	MultiTaction 3-tiles table and 12-tiles wall designs	195
A.3	General network configuration of server and cell communication	196
A.4	Enabling 4x4 markers on each cell	197
A.5	Configuring the cell's network address	197
A.6	Configuring the server's network address	198
B.1	PCB design of the active tangible	201

B.2 Case design and final 3D print of the active tangible	201
C.1 General flow of application communication	204
C.2 Running XAMPP control panel as administrator	205
C.3 Installing the Apache service	206
C.4 Node.js server running	207

1 Introduction

1.1 Research Overview

This dissertation focuses on reverse engineering of biological and clinical prognosis systems. The results from this research could be applied to better understanding these systems, to develop new reconstruction tools, to improve data collection and training factors of reconstruction, and to make better decision upon a sense-making process.

In addition, this dissertation develops three complement paradigms of the conceptual and procedural aspects involved with:

1. How interactive visualization makes sense of big biological network data and provides a platform to understand and compare the structural features.
2. How tangible modelling and visualization introduces a survival analysis framework and deployable system to predict hazard risk of cancer patients and help clinicians to decide about the proper treatment.
3. How graspable modelling and visualization provides an effective tool to gain insight into large system architecture and component interconnections while also helping to verify and evaluate a system's behaviour visually and allow one to highlight the most effective molecules, or features of a system and sort through them based on different validation

terms for system interpretation.

1.2 Research Motivation

Biological and clinical prognosis systems are extremely complex and their functionality depends on their components (such as genes, proteins or clinicopathological features) and also the interactions among them. Reverse engineering of biological and clinical prognosis systems is a backward reasoning process which, when explored through observation of their behavior, allows one to analyze and assess the components of these systems, understand their structure and how they work [Cho, 2016, Zhang, 2015], modify hypotheses, and evaluate. Gene regulatory network inference and effective clinical factors discovery are two such challenges, prevalent in modern bioinformatics.

To date, research in these areas has favored a purely statistical approach, the results of which are often highly contingent on the availability of adequate ground truth data. Accordingly, methods are predominantly evaluated based on synthetic data alone. The size, complexity and fragmented nature of existing real-world bioinformatic datasets, along with the lack of interactive tools to assist in the visualization and exploration of such data, and limited availability of ground truth, hinders progress in these areas. In addition, the structural nature of such problems permits immediate opportunities for embedding and leveraging visualization into the analysis pipeline.

As such, there is a need to develop integrated reconstruction, visualization and statistical analysis approaches by using the existing datasets to assist biologists and physicians for having a better understanding from these biological systems. Furthermore, the lack of interactive visualization tools which utilize existing big datasets as the starting point for the extraction of evidence and prior information can improve the reconstruction and statistical analyses considered. Indeed, an important prerequisite in modeling, inference and prediction, is to examine the structural properties of these systems before attempting to reconstruct them and understand their function.

In recent years, interactive visualization techniques have been implemented to explore differ-

ent representations of these systems [Dang, 2015, Blevins, 2016], but the cognitive aspects of dynamic visualization and reconstruction have received little consideration.

In this thesis, we address such limitations, by pursuing an integrated approach toward modeling, inference and prediction tasks deployed across several bioinformatic contexts. In particular, we investigate and propose more unified frameworks and toolsets that support different users (scientists, biologists and physicians) in the task of understanding and analyzing biological data and systems, by leveraging interactive visualization, stimulating cognition and supporting collaborative analyses.

1.3 Research Challenges

1.3.1 Challenge 1 - Gene regulatory networks reconstruction

DNA molecules within chromosomes in the nucleus of a cell represent the nucleic genetic material and genes as basic inheritance units are encoded in discrete segments of DNA molecules (Figure 1.1a). The genetic information for the development and functioning of all living organs are located on the mentioned segments. All cells of our body (such as heart, liver, brain, etc.) comprise the same genetic information, but only some specific genes would be active in each of these different organs.

The gene activation means its information is propagated or so called “expressed”. Expressed genes provide a vision about molecular functions in the cell. Indeed, a gene is expressed when it makes a new product (protein). Genetic information expression is a multi-step process that proceeds from DeoxyriboNucleic Acid (DNA) to messenger RiboNucleic Acid (mRNA) to protein (Figure 1.1b). Put simply, two main sub-processes “Transcription and Translation” play a principle role in genetic information expression. Indeed, transcription and translation rates control the flow of information from DNA to mRNA (mRNA production), and from mRNA to protein (mRNA utilization) (Figure 1.2).

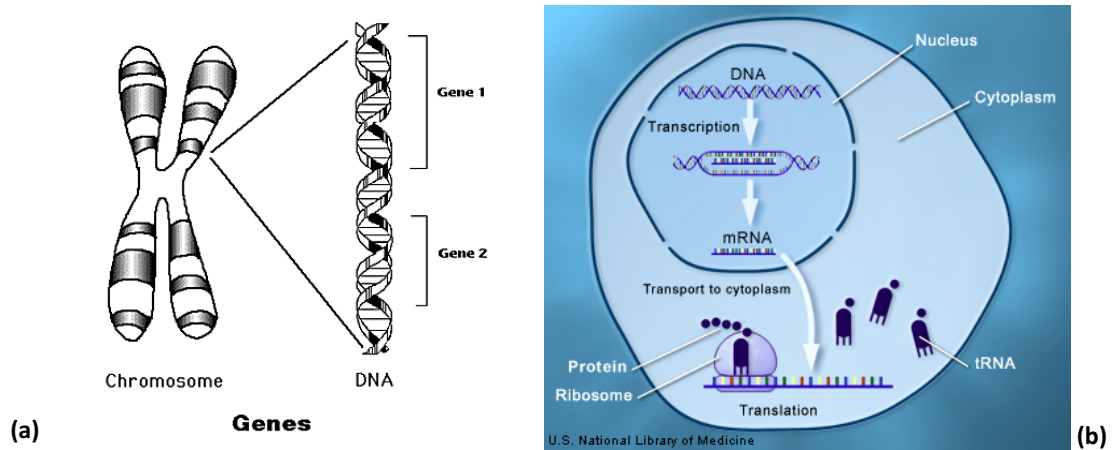


Figure 1.1: a) The schematic of chromosome and genes location (left), b) Propagation of cellular information [GeneticsRef., 2016].

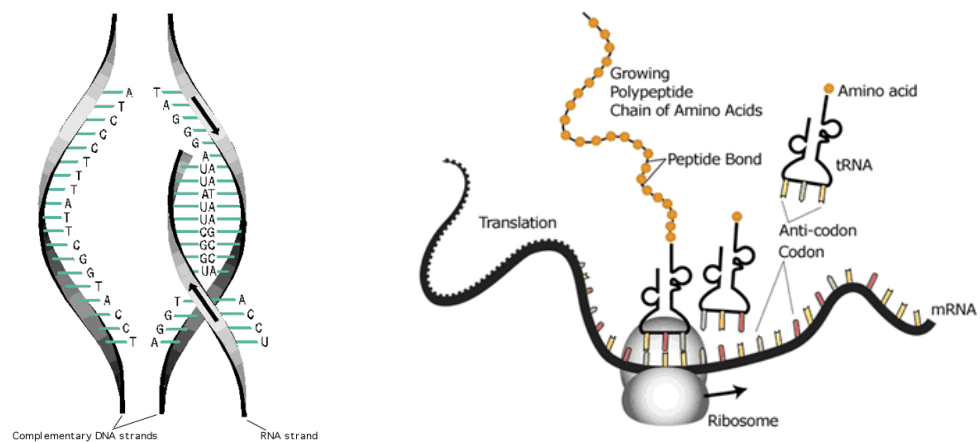


Figure 1.2: Transcription (left) and translation (right) sub-processes [GeneticsRef., 2016].

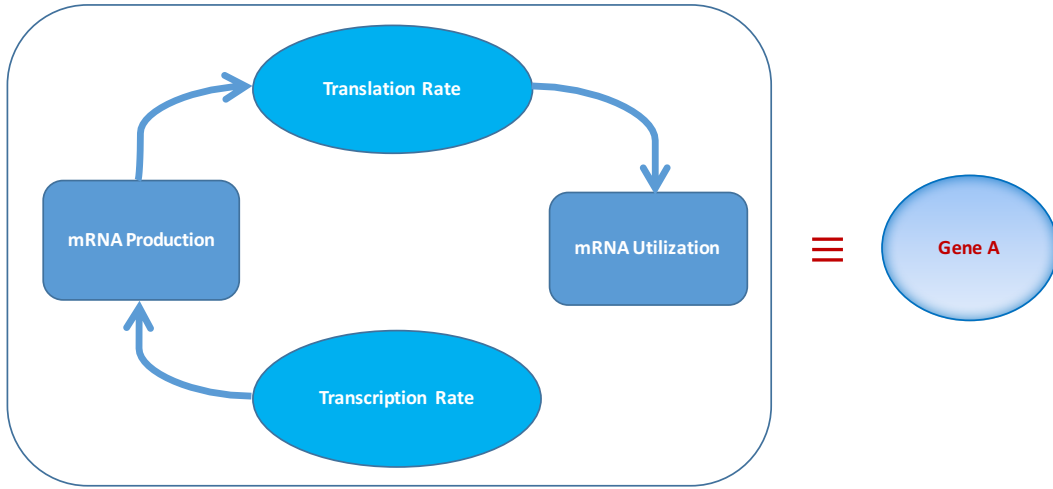


Figure 1.3: Machinery of mRNA production and utilization and the used simple model in GRNs.

Gene regulatory networks (GRNs) are represented as a graph where each node depicts the machinery of mRNA production and utilization (gene expression) (Figure 1.3). As shown in the above sub-processes, mRNA plays a key role as a mediator of gene expression; its role is carrying certain part(s) of information from the genetic library (encoded in the DNA) to the machinery that creates proteins. Regulation of mRNA production and utilization by controlling the transcription and translation rates is the main occurrence in regulating gene expression. The dynamic nature of these events has a key effect on the genetic interactions in GRNs (as shown in Figure 1.4).

At this point, there is a question: “Is it possible to measure this dynamic in gene expression?”. In the past, biologists have been able to measure gene expression levels only for a few genes at one certain time, but with DNA microarray development and similar technologies, recently they can measure the activity (expression) of thousands of genes at any given time, or profile expression changes for thousands of genes over some desired duration.

DNA microarrays are created by robotic machines. These machines arrange thousands of synthetic gene sequences on a microscopic slide. A microarray, like a slide, typically includes thousands of micro-spots consisting of probe sequences (each unique to a particular target

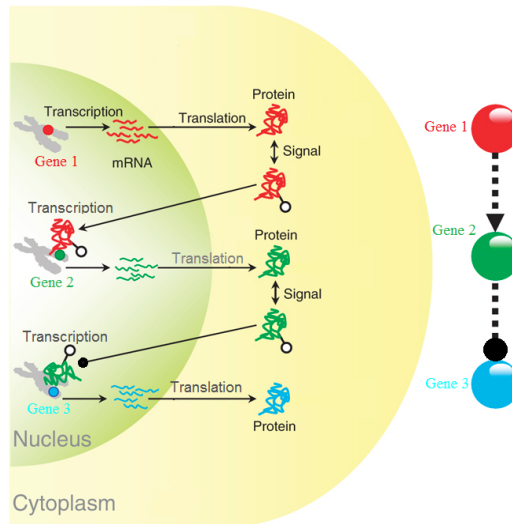


Figure 1.4: GRN representation that is compressed in time.

gene). As an example, it is possible to place a large number of probes (sometimes more than 10,000) on a small surface about one square centimeter. In this technology, mRNA from the particular samples of interest (e.g. normal cells vs. cancer cells) would be isolated, and by using a reverse transcriptase (RT) enzyme, converted to a complementary DNA (cDNA) labeled with different fluorescent dyes (different dyes for each sample of interest). Then, biologists place the labeled cDNAs on each slide, where the cDNAs hybridize (bind) to their synthetic complementary DNAs attached on the spots of microarray. The fluorescent intensity of a spot is directly proportional to the amount of mRNA expressed in the cellular sample relating to a particular gene. If a particular gene is very active (e.g. produces many labeled cDNA), a very bright fluorescent area is generated. On the other hand, a gene which is less active results in dimmer fluorescent spots. A light scanner can detect the different fluorescent dyes by scanning the surface of the array for hybridized material. A summary of microarray technology for a two-channel experiment is shown in Figure 1.5. As a common experiment, two different samples (normal and cancer) are labeled with different dyes (green and red), and then bound to the same microarray. Both labeled samples are mixed and hybridized on the slide. Green spots appear when only cDNA from the normal cell is bound, while red spots show only cDNA from the cancer sample is bound, whereas yellow spots demonstrate that cDNA from both are

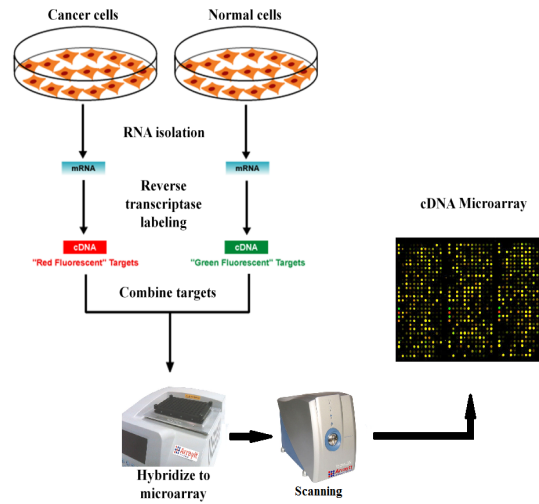


Figure 1.5: DNA microarray process.

bound in equal amounts, and grey spots indicate no hybridization (the latter two indicative of similar genetic response within both types of cell).

In this dissertation, we are generally interested in time series microarray data in which the arrays are collected over a time course (Figure 1.6). Usually, microarray experiments are noisy and some sources of error (such as systematic biases in scanner settings, laser saturation effects, and sample plate origin) have an influence on the final image, hence, biologists replicate the experiment at least three times to ensure the stability of the obtained results. Microarray data measurements are then reproduced to form repeatable results. In final step, some preprocessing methods are utilized to adjust microarray intensities related to slides extracted through a course so that comparisons can be made within and between slides in the experiment over time. In addition, adjustments are necessary to remove alterations which are created technically and do not demonstrate true genetic changes. Also, some complex algorithms are developed to remove the bias artifacts. After these preprocessing steps, microarray analysis methods are applied to extract meaningful gene expression levels of all genes over time from this microarray data.

Although hybridization based microarrays are still the main mechanism to extract large gene

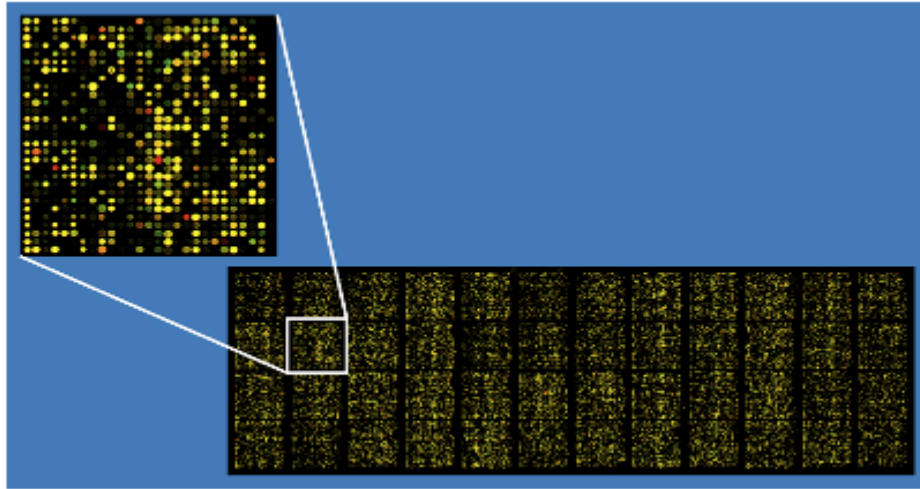


Figure 1.6: Time series microarray data [Parmar, 2013].

expression levels with advantages in repeatability of experiments and sensitivity, new technologies that are capable of extracting time-series gene expression data continue to emerge: such as probe-based methods (e.g. Nano-string) [Geiss *et al*, 2008] and RNA sequencing (RNA-Seq) [Wang 09]. Nano-string technology [Geiss *et al*, 2008] is a conversion of the DNA microarray using molecular barcodes and microscopic imaging to detect and count up to several hundred mRNA in one hybridization reaction. The used protocol in this method eliminates any enzymatic reactions that might create bias in the output results. RNA-Seq [Wang *et al*, 2009] is a technology that uses next-generation sequencing to extract the quantity of mRNA at a given time. It has the potential to replace microarrays in transcriptome analysis due to its advantages in capturing more changes from gene expression levels, delivering data in familiar array format, less bias in experiments, and at lower cost per sample. This being said, the analysis that will be conducted in the proposed work will focus on hybridization based microarray databases as they are more readily available. It is important to note though, that such analysis can extend to the data that becomes available from these more recent technologies.

It is vital to characterize changes in gene expression levels over time since the gene expression regulation is a dynamic process. In the proposed research, we are interested in time series gene expression data. The key challenge for time course data is that the number of time points

is small and the number of genes is very large. Another significant challenge is the sheer complexity and lack of formal structure that pose gene regulatory networks. In mechanics, there is adequate data and the system is really well known, but we encounter big data from cellular processes and the data is gathered before we even have any hypothesis. The data is ready to be queried with all mentioned challenges. The main tasks of interest in microarray data analysis are: gene regulatory network reconstruction, and visualization of genetic interconnection networks.

The aim of gene regulatory network reconstruction is to detect the most likely interactions by identifying sets of relevant model parameters that are required to obtain an appropriate correspondence between measured data and model output. Today's gene expression data is noisy, unstructured, and dynamic rather than static, whilst it may also be corrupted or incomplete. Therefore, biologists need new computational tools in order to extract useful information from the datasets. Then, it is necessary to translate this big data into vectors or matrices in a more sophisticated manner and develop more generalized methods for analyzing them. The number of computational methods that are being developed to reconstruct gene regulatory networks (GRNs) from genome-wide expression data is rapidly increasing, often referred to as expression-centered methods. Methods that infer GRNs go one step beyond and infer causality relationships in the network by also identifying the programmes of the genes or modules (gene clusters), to describe how mRNAs cause the observed changes in expression of their cognate target genes. In general, there are two representations of networks in computational biology. The first one is so called a co-expression network in which the nodes and undirected edges represent the genes and the degree of similarity in the expression profiles of the genes, respectively. Either a correlation or dependency relationship between the nodes is displayed in these networks, while the cause of the relationship is not represented. The second is termed a gene regulatory network represented as a directed graphical network in which the nodes represent genes (or modules) and edges are directed. A causal relationship is shown by this kind of network [De Smet & Marchal, 2010], where directed edges can represent activation (the promotion of expression in the associated gene), or inhibition (the suppression of expression in the associated gene). Inference of causality is important in identifying poten-

tial signaling pathways to be analyzed or exploited by biologists in future investigations.

The objective for achieving systematic, comprehensive and accurate reverse engineering of biological systems makes bioinformatics much more demanding for biologists. The reconstruction approaches consider the biological system as a black box and then tries to identify the details of the components of the systems (such as genes and proteins), and infer their interactions. They use related datasets as the starting point and statistical data mining algorithms for a comprehensive understanding and interpretation of the biological system. These systems can be represented as networks in which nodes represent the genes/proteins and edges show the interconnection among them. This format of representation allows to capture their structural features of these systems.

Although the focus of most recent studies is on the reconstruction of these systems to infer their functionality, it is a prerequisite to go through the structural features of these networks before attempting to recover them and interpret their functioning. For example, (1) *Sparsity*: it has shown that the distribution of interactions in gene regulatory and protein–protein networks is very sparse. Thus, the sparsity property is considered as a feature of genetic networks. This feature is foundation for many structural features. (2) *Dynamics*: the genes of a regulatory network exhibit a dynamic behavior over time. These components of the system respond to changes in their environment and also cell state, and express their responses over time that can be measured via high throughput gene expression technologies. (3) *Robustness*: it is a common feature of biological systems which maintains their functionality caused by perturbations inside and outside of the cell. Although this property is directly connected to the stability and the energy of the system in signal processing, it is poorly understood as to whether it depends on the changes in energy states or the network's stability (as theoretically proposed in network science), and there remains no clear interpretation of network's stability and energy. (4) *Modularity*: a module in a network is composed of a subset of nodes that are more connected among themselves than with other network nodes. Such structural features make this field of study really interesting and also challenging.

There is a need to have tools to allow us to utilize existing databases to explore more accurate structural features of biological systems and have a better understanding of the laws which

may be operating and evolving in these systems. In addition, discovering assessment terms to validate the performance of models according to the inferred structural features needs to be more adequately defined.

1.3.2 Challenge 2 - Prognosis Models Reconstruction

After lung cancer and colorectal cancer, prostate cancer is the third leading cause of cancer mortality in developed countries and is the most frequently diagnosed cancer for men [Siegel *et al*, 2016]. Almost 15% of the 2 million men diagnosed in the US with prostate cancer over the last 10 years had metastatic disease at the diagnosis time [Dream Challenge 9.5].

It has been demonstrated from 70 years ago that deprivation of androgen hormone leads to regression of prostate cancer and alleviation of pain in patients. Androgen deprivation therapy (ADT) has become the main treatment for patients with prostate cancer which results in a high rate of response [Gupta *et al*, 2014]. Despite such blockages, prostate cancer is known to progress to a state called metastatic, castration-resistant prostate cancer (mCRPC) account for one third of all patients.

Although significant improvements in outcome for men with mCRPC are now possible due to recent drug approvals including various therapies - how best to deploy them has not been found to produce any major improvement in mCRPC patient overall survivability. The survey of clinicopathological features related to mCRPC patients independent of treatment will facilitate the design of future trials, especially through homogenizing risk, and will enable smaller trials for the assessment of treatment effects. There is therefore, a need to develop new models for predicting survival in mCRPC patients treated with docetaxel. Over the last decade, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) Consortium (www.dreamchallenge.org) has hosted over 40 challenges focused on difficult and impactful questions related to systems biology and biomedical science. [Dream Challenge 9.5] has actively pursued the above goal for prostate cancer patients.

Survival analysis is a branch of statistics used to analyze datasets where the output is the time to an event of interest [Shafipour & Abdolvahhab, 2016]. In the clinical setting, survival

analysis is often used to study the time to death of patients in studies. The difficulty of these studies is due to dealing with right censored datasets where patients may drop out before the conclusion of the test or patients do not experience the event before the end of the test, thereby resulting in an unobservable survival time for the patient.

The tools to tackle these difficulties entail the use of algorithms from survival analysis domain. In addition, different feature selection methods used in survival analysis are purely statistical approaches and ultimate users often lose intuition about why a feature is important. Furthermore, there is no way to construct and explore the hybrid features automatically. Therefore, there is a need to develop tools which allow experts in the loop to iteratively define and evaluate features to be integrated into the survivability prediction process.

1.3.3 Challenge 3 - Grasping Trends in Biological Systems Modeling

Understanding and interpreting the inherently uncertain nature of complex biological systems is a notable challenge. The data sets are large and incomplete, and the problems are interdisciplinary in that they require knowledge from multiple disciplines in order to gain understanding and draw new insights. For example, expertise from biological disciplines is needed in order to understand the biological functioning and implications of a given system, while knowledge from mathematics and engineering disciplines is needed in order to be able to analyze the data and construct models that can accurately represent the given system.

Currently, large datasets and modeling problems are typically tackled through algorithmic approaches. These can sometimes fail, since even if the algorithms find a solution, it is difficult for researchers to know how and why such a solution was obtained, and whether it is even a good solution. Also, these approaches do not easily support shared understanding between researchers from different backgrounds. To address these limitations, there is a need for platforms that can enable researchers to directly and iteratively build and manipulate models, informing their understanding of complex systems and enabling them to make new hypotheses and develop efficient constraints for further exploration of a given system. Such a

platform should allow researchers with differing disciplinary expertise to consecutively adjust the structure and variables of the system to gain a better understanding of the cause and effect relationships among the parameters and to also be surprised by the predicted effect. The aim is to find solution(s) that can address the problem context appropriately. Factors include whether the selected model, structure and parameters are an appropriate match for the given system, which requires confirmation from specialists. It may be that a solution appears good, but that the structure or parameters of the model do not accurately represent the specialist's understanding of the real system. Indeed, finding the key factors and also the representative pathway to frame the problem provide a flow to reach the suitable solution(s).

1.4 Approaches to Face the Challenges

1.4.1 Approaches to Encountering Challenge 1

The objectives of this part of dissertation are to have a better understanding of various cellular processes, using gene expression data to reverse engineer phenotypes; also to develop an interactive visualization tool for exploring structural features in gene regulatory networks and extracting the evidence that might be used to design new constraints or validation terms in future reconstruction methodologies. Toward this end, the following two key problems will be addressed:

1. How do the genes whose gene expressions change with time, interact?
2. How can structural features in genetic networks be visualized and interpreted interactively?

Answers to these questions are investigated in three parts: In the first part, I have focused on reconstruction of gene regulatory networks of the cell cycle as one example of a complex and dynamic process. One possibility to conquer the aforesaid challenges is to impose more biological constraints (such as sparsity, stability, etc.). Such constraints and prior knowledge play an essential role in discovering the dynamics of genetic regulation. Moreover, only a few

numbers of algorithms have focused on time-series gene expression data to track behavior changes through time, and others have concentrated on static gene expressions.

This work proposes a modification on GRN reconstruction algorithm using a stable network topology via sparsity-seeking convex optimization to capture behavioral patterns in such dynamic datasets. This method is directed towards obtaining greater insight into the cellular dynamics over time. The novelty of the proposed method is its consideration of a complete collection of constraints including sparsity, stability and a priori knowledge in network topology optimization. Firstly, we define a precise representation of the system by ordinary differential equation (ODE) models, used as functions to encode dynamical structure information. Secondly, we use discrete formalisms, which provide a coarse, but realistic representation of the system dynamics, whilst still highlighting fundamental features of the network structure. We then explore the properties of such a representation for the purpose of solving the reconstruction problem and finally analyze how the proposed methodology should be modified to incorporate such constraints.

The proposed method is applied to yeast *Saccharomyces Cerevisiae* as a biological dataset. The extracted knowledge of possible gene interconnections is shown to be more consistent than those from the literature in revealing known and experimentally validated gene interactions from real expression data. The sensitivity, precision, and F-score criteria are calculated to evaluate the inferred networks through this method. Resulting networks are compared with five previous studies in the literature on the same dataset.

In the second part, a novel methodology is developed which integrate two universal properties of the networks, sparsity, low-rank properties in optimization process with a novel modularity property. The modularity property implies that if a sample point or node is placed in a module in the network, its neighbouring sample points or nodes would also be located in the same module with high probability. A challenge will be raised when we want to validate the number of obtained modules. We don't have or better to say we don't know about the labels of samples or nodes in most existing databases (especially biological datasets). This challenge has been addressed by an interesting property in spectral graph theory field. The multiplicity of eigenvalue 0 in Laplacian spectrum shows us the number of modules.

Although the focus of above parts is on the reconstruction of these systems and their inferred functionality, there is a need to go through the structural features of these networks before attempting to recover them and interpret their functioning. It is due to the fact that databases that collate information on partial ground-truths can only give us a sense of “consensus” amongst researchers. These databases have many problems, of which incomplete data is a major one. Some other of these problems are structural. Structural problems often provide something more than just statistical analysis to do comparison. This requires an interactive visualization tool to allow human-in-the-loop processing to help in assessing the visual and structural aspects of data.

A tangible biological networks (TBNs) tool is thus developed to make sense of the biological networks and their structure, and to aide in making decisions about what should be investigated in the future. We can measure structural features, but the difficulty is in how to build it into the model. This tool takes this type of research out of black box in an attempt to elucidate some kind of intuition. Indeed, the interactive visualization is what allows us to look at these networks and obtain intuitions.

By using this tool, we can figure it out if there is an evidence in real networks to consider structural constraints in the modeling. This tool uses the advantages of tangible interactions in terms of collaboration and learning and attempts to create a common language among people in different fields of study, allowing experts the chance to clarify their own prior knowledge with their colleagues. This visual, interactive and iterative approach frames the reverse engineering problems, for a more effective environment in which such signal processing can be deployed, assessed, and in some way, validated or experienced by knowledge experts to enhance insight and sense making.

1.4.2 Approaches to Encountering Challenge 2

The first objective of this part of dissertation is to have a better understanding of effective prognosis factors related to mCRPC patients. [Halabi *et al*, 2014] have showed that by using the recent mCRPC patient datasets, better prognostic models can be developed. They extended

their study to investigate the site of metastatic disease as being informative prognostic factors. Their finding has demonstrated the importance of prognostic research to include current clinical trials and patient health status in determining the best treatment choices. The results of the Halabi study motivated the organizers at Sage Bionetworks to create the DREAM 9.5 [Dream Challenge 9.5]. Organizers hoped a crowd-sourced competition could lead to new models for predicting survival in mCRPC patients treated with docetaxel. With better models, the goal is to allow clinical researchers to decide if docetaxel is a viable first treatment option or not.

Participants are asked to build a prognostic model to predict a global risk prediction and optionally 3 separate optimized risk predictions at 12, 18, and 24 months for patients in the test data set. For the global risk prediction, a concordance index (c-index) and an integrated time-dependent AUC (iAUC), from 6 to 30 months, using the estimator of cumulative AUC [Hung and Chiang, 2010], are calculated for each model. For the time specific predictions (12, 18, 24 months), an AUC score using Hung and Chiang's estimators are calculated for these times. Further details of these performance measures are discussed in the chapter 4.

The scope of this part of dissertation encompasses using tools from survival analysis to develop a prognostic model for predicting risk in mCRPC patients treated with docetaxel. Since the historical median survival time for mCRPC patient is less than 2 years, an accurate prognostic model indicating whether docetaxel provides any tangible improvement would be of potential benefit in the decision making of treatment options. In the other word, this research investigates the influence of both clinical and pathologic features on the survival time of prostate cancer patients.

It is widely accepted that many factors play an influential role in determining the survival time of prostate cancer patients. Many analyses of cancer registry survival data use the Cox proportional hazards model [Cox D.R., 1972], which has had a profound influence on the development within the field of survival analysis. However, it relies on the assumption that the ratio between hazards is constant over time. Because of the long-term follow-up required for prostate cancer patients, the proportional hazards assumption is often violated, leading to poor model fit.

One of the possible ways to extend the Cox model is to include hybrid combinations of clinicopathological features, either as a linear function or as another polynomial function. The presence of experts-in-the-loop is very useful to factorize out a clinicopathological feature which does not satisfy proportionality and also use his prior knowledge to add informative hybrid combinations of clinicopathological features which could not be found in original dataset.

Currently, the research in this area signalizes the clinicopathological features that can be used to better understand patient prognosis, but there is no visual predictive model with good prediction power that can be presented to assist healthcare workers. Therefore, this thesis proposes an interactive modeling and visualization tool called Tangible MultiVariate Visualization (TMVV) to address these challenges.

Interactive visualization by assistive manipulative models allows to clinicians to move on from the traditional scoring protocols (such as Apgar ...). In addition, innovative modeling and visualization tools for data analysis, monitoring, and interpretation are needed for better understanding from different cancers and also for standard and robust designing of data collection systems. The objectives of the proposed system are to improve data collection, visualization of prediction results, and further enhance diagnosis and prognosis models.

1.4.3 Approaches to Encountering Challenge 3

The aims presented in Challenge 3 lay at the intersection of computer science and behavioral sciences (Human Computer Interaction (HCI)), computational biology, cognitive science, and interactive visualization. Indeed, we propose an exploration platform that builds an interactive 4-way relationship among the users, biological data, contexts, and decision-making. This multidisciplinary platform is a tool named Tangible Tensors (TTs) to manipulate biological data, capture information, and effectively use organizational knowledge.

Fortunately, recent progress in technology has provided the conditions to build active tangibles and also multi-touch tabletops and walls to interact with the data and system properly. Through the use of such tools, one can better explain the causality of a system, figure out

how to reach a successful outcome, avoid repeating previous non-successful experience, and in the model design process, define what kind of constraints and simplifications should be considered.

This tool creates a schema that facilitates insight development through the manipulation of this representation and proportional interaction based on obtained insights. In other words, it accentuates the role of knowledge-based structures obtained from expertise and experience by allowing the user to make sense of a problem and formulate a strategy to harvest the solutions. For introducing this tool, we need to determine the boundary between the problem/project and solution spaces and also the overlap between these two spaces. As a first step, we need to know the definition of these spaces:

1. problem/project-space is a context representation of a particular project requirements and its underlying components and predesignated behavior.
2. Solution-space is a conceptual representation of possible solutions and its relationships to the underlying components of problem/project space.

As we can see from the definitions, there are many contexts in which these spaces may overlap. We cannot define a real distinction between these two spaces. While we are looking for any graspable causality relationships in problem space, the solution space is providing us a domain to resolve this problem.

We should manage the solution pathways properly because exploring the appropriate solution for a shared problem can bring up conflict when we have different users. Also, we should consider each solution can be proper in some conditions and inappropriate in others. The user should have this chance to explore under what conditions the solutions of users with different expertise is acceptable. We aim to develop a tool to help people capture more knowledge from the shortcomings of their solution-selection narrative patterns and also the privilege of all the alternatives obtained from other users. In other words, there is a general rule in this tool that the user should indicate the faults of his/her solutions and also consider the alternative solutions of other users.

According to the above discussion, we make motivation for an individual to use the tool (after only brief instruction), comes from the instant informative feedback provided for the user to make a sense of his/her tactic for model adjustment, with inferred results shedding light to guide the user for applying better tactic. This feedback is measurable. Therefore, do not impose specific restrictions on the number of parameters or dimensions of the system.

The unique design thinking of each user gets the informative feedback to understand the desirability and also add the feasibility to this aspect. Indeed, the focus of design thinking is on problem-space exploration and diverse perspectives instead of addressing the solution. The tool brings different experience levels related to various expertise in one place and facilitate to bring many ideas before extracting meaningful insights and deciding about an ultimate solution.

Generally, in this dissertation, tangible visualization and interactive modeling embedded within the sense-making process address three main interactive visualization challenges:

1. The developed system and technology provide an interactive and graspable environment of the present components in a certain biological or clinical prognosis system.
2. Various users could interact and think about the system they choose to analyze (independently, yet within a shared visual space).
3. They can make sense of effective biological or clinicopathological factors in system performance and also use information from various visualization tools.

A brief description of the structure of the thesis is provided along with specific contributions are given in the following section.

1.5 Research Contributions

There are seven research contributions in this dissertation:

1. Stable gene regulatory network topology via sparsity-seeking convex optimization,

2. Structural Norm Minimization based on Neighbourhoods,
3. The description of the interactive visualization gap for biological networks, conceptualization of structural features and scalability as a "sense-making" task, and the development of Tangible Biological Networks (TBNs),
4. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC,
5. Graspable prognosis factors visualization for interpreting cancer data named Tangible MultiVariate Visualization (TMVV),
6. Tangible Tensors (TTs): An interactive toolkit for grasping trends in biological systems modeling,

Contributions 3, 5, and 6 provide a novel cornerstone which start to bridge a gap between low-level detailed representations of biological and clinico-pathological systems and high-level abstract construals that expert people (like bioinformaticians, biologists, physicians, clinicians, and pathologists) use in practice (Figure 1.7). Abstract reasoning and interactive visualization in systems reconstruction and interpretation are fundamental to connect high- and low level representations in order to achieve the system biology goal of “augmented biological networks and survival analysis” in the next decades.

One of interesting topics in computational biology is reconstruction of gene regulatory networks. These networks represent the regulatory and physical interactions present among genes of an organism. In this application, we are encountered with gene expression patterns over time, from which an interconnected topology describing the regulatory interactions between genes must be inferred. To arrive at this target, the first contribution proposes a framework for studying gene interactions inference using a combination of sparsity, prior knowledge and stability techniques to explain time series patterns obtained from high-throughput microarray technology (Chapter 3). The stability criteria is modified based on the obtained results from the gene networks exploration. The reconstruction methodology is developed by minimizing

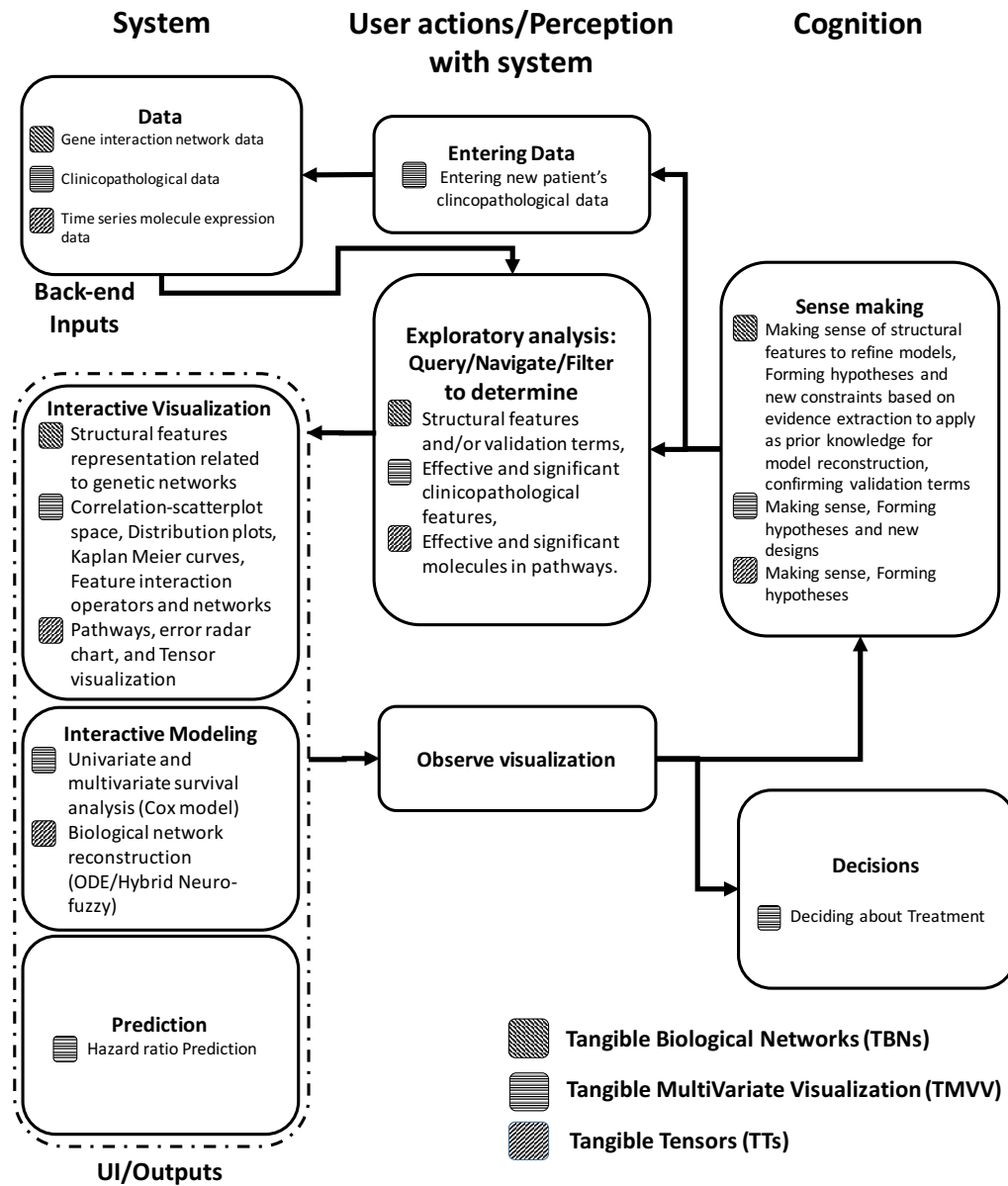


Figure 1.7: Flow of interactive modeling and visualization design related to Tangible Biological Networks (TBNs), Tangible MultiVariate Visualization (TMVV), Tangible Tensors (TTs) platforms in dissertation

the trade-off between of the sparsity of gene interactions in the network and the best model accuracy, where stability and prior knowledge are considered as constraints for the genetic network structure. Our algorithm is applied to cell-cycle gene expression data in *Saccharomyces cerevisiae*, and the results show the reconstruction performance. The convex nature of the model proposed, makes it suitable for application to large-scale networks.

The second contribution of this dissertation is an inference from the exploration of various networks. A novel method is developed which not only consider two universal properties of the networks such as sparsity, low-rank property in optimization process but also consider clustering property. The clustering property implies that if a sample point or node is placed in a cluster in the network, its neighbouring sample points or nodes would also be placed in that cluster with high probability. The obtained results are confirmed by the analysis of laplacian spectrum of the reconstructed network (Chapter 3).

The third contribution is the fulfillment of the representational gap in network analysis and conceptualization of structural constraints related to biological network reconstruction as a “sense-making” process (Chapter 3). This contribution relates the fields of sense-making and mental models with Human Computer Interaction (HCI) and network science from where the concept comes from or what it really means to argue why prior knowledge or visual experience of the viewer is necessary for defining structural constraints in biological network reconstruction. This contribution provides a tangible and collaborative platform in which the users can apply well-designed statistical methods associated with interactive visual representations to achieve qualitative and quantitative comparative analyses and also knowledge extraction. The obtained results from this toolbox could improve the visual and conceptual understanding from the structural features of biological networks and prevent fault or incomplete constraints to reconstruct gene regulatory networks.

The forth contribution is a hybrid survival analysis method to estimate hazard risks of metastatic castration-resistant prostate cancer (mCRPC) patients (Chapter 4). The two phase method was then employed in order to determine the most effective clinical features for hazard risk prediction. Effective prognosis factors identification and knowledge about hazard risk could have significant impact on clinicians and patients. Predictive algorithms, may provide clini-

cians with the tools necessary to make more accurate decisions about treatment strategies, most importantly patients may be more aware their prospective future.

A fifth contribution of this dissertation is named Tangible MultiVariate Visualization (TMVV). Indeed, we present a tabletop and smartphone based framework to design a cross-platform and collaborative system to discover effective factors in predicting hazard risk of patients in a given disease and classifying the patients using a Web interface (Chapter 4). It uses smartphones as active tangibles to allow covariates selection, survey the selected covariates by statistical tools (box plot, Kaplan-Meier plot and re-categorization), and decide about the effectiveness of each prognosis factor based on the extracted results from the univariate and multivariate survival analysis. In addition, the user could see multivariate correlation scatter-plot chart of the selected covariates on tabletop and could track the significance and accuracy of bivariate modeling on this chart simultaneously. We also introduce a new presentation for visualizing the effective interactions among clinico-pathological covariates of prostate cancer patients using covariate-link networks and the users could track the reasons of factors selection by clicking on the link between factors. In addition, the solution spaces based on p_value and accuracy of the hazard ratio prediction is developed in which the users could follow their performance and possible thought rotations.

Finally, our framework allows clinicians to import the covariates of a new patient and use the effective factors obtained during manipulating of the system to predict the hazard risk for a new patient. The ultimate user of this system could be clinicians, physicians, and health workers. This framework involves extensive interaction between users's actions and mental reasoning process that occurs in the human mind. This contribution characterizes the different abstractions that are used in survival analysis to represent and analyze the prognosis factors related to the given disease and also interpret the effect of factors combination on a specific event like death. This contribution is a necessary platform in order to survey the procedures and interactions which are inspired to design novel and general methodologies for analyzing survival time in various databases.

The sixth contribution of the dissertation introduces the concept of bridging the gap between research related to data analysis and research involving around the visualization of data called

Tangible Tensors (TTs). Specifically, concepts related to tensors were used as the backbone for developing a unique visually based tool for manipulating and spotting trends in a set of data. First, we introduce the concept of tensors and their relevance to visualizing data. Next, we present the system design for a toolkit that utilizes the tensors concept. Finally, we demonstrate the importance of this technology for providing progressive innovation in the field of biological pathways analysis. This framework provides a systematic deduction of the pathways structure and the procedures involved in modelling (Chapter 5). This tool would establish the knowledge tracking which performs effective reasoning and problem-solving process in modelling domain. The theory of how people make sense of a system and what is their strategy to manipulate modelling's parameter and achieve a good performance is presented. This theory includes sensing data and model, goal representation, manipulating data and model, sensing information, updating the mental model, comparing the results with other users' results, interpreting the obtained information, and generating new hypotheses. Consequently, the cycle that people with different expertise uses to interact with a problem/system and deduce the sense-making loop for intelligence analysis is considered in this contribution.

1.6 Outline of the Dissertation

Chapter 2 provides a literature review which details the concepts of visual analytics, situation awareness, sense-making, and mental models, and tangible user interfaces (TUIs).

Chapter 3 presents a modification on stability property according to the exploration of the structural features related to biological networks in convex optimization by using of time-series gene expression datasets. In addition, a new inspired property from TBNs (called neighbour norm) is developed as a modularity constraint included in an optimization methodology. In addition, an interactive and tangible visualization is developed to explore biological networks and the characteristics of their structure, and also to improve cognitive performance of interdisciplinary researchers in defining valuable constraints at biological networks reconstruction. We have named it Tangible Biological Networks (TBNs). The user could explore how a struc-

tural feature changes in different biological networks and what kind of relationships there are among various features. The system provides a platform to survey how conceptual knowledge could help to better understanding from solution space in reconstruction methods.

In chapter 4, a general hybrid method as an achievement of using the visualization platform is designed to find significant prognosis factors to predict survival time and hazard risk of more than 2000 prostate cancer patients. In addition, a graspable prognosis factors visualization to survey, analyze, and interpret cancer datasets is developed and described in this chapter to identify how these factors relate to each other.

In chapter 5, we present an interactive toolkit for grasping the trends in biological systems modelling called Tangible Tensors.

Finally, chapter 6 summarizes the overall findings of the dissertation into biological and clinical systems analyses. It also discusses the implications of the developed methodologies and systems and indicates areas for future research.

2 Background and Foundations

In this chapter we present the concepts related to visual analytics, situation awareness, sense-making, mental models and also tangible user interfaces.

2.1 Visual Analytics

Our world currently bombards us with enormous amount of data that is constantly generated by industry and business branches, government and science labs [Hashem *et al*, 2015]. Here we are facing two challenges, first of which is data storage and second one is the extraction of useful information from the raw data. Although we have had progress during past decade to overcome the data collection and storage, the rate of data generation is still way faster than the ability to analyze them for advanced purposes such as decision making [Ernst and Young Global Limited, 2014]. Not being able to use the data in a proper way results in the waste of money, time and resources. This problem can affect people in both their professional and private lives. For example, data analysts and decision makers can be lost and come up with conflicting result due to massive data that comes from different sources [Elliott *et al*, 2015]. Currently we lack models, methods and technologies that turn raw data to reliable and solid knowledge at the right time [Berman *et al*, 2015].

Knowing how important is to analyze and understand data leads to establishment of different

powerful and fully automated tools. However, finding the path that connects data to decisions or new hypotheses is way far more complex. Even having these automated tools has not been able to have a huge impact on our understanding [Vanky *et al*, 2016]. Indeed, they have added a complexity to our analyses, in the other words, we need to to rethink our processes of analysis. Besides, fully-automated data processors lack the ability to communicate their knowledge which cause a new problem. This issue is pivotal since these methods reflect their creator's knowledge and if decisions are made using such methods turn out to be wrong, we need to test the procedures. This is where visual analytics comes to play important role. Instead of dealing with the final result, it provides communication language. Decision makers and data analysts need a mechanism to effectively communicate their thought processes toward decisions that are data driven in order to reach the ultimate goal, which is improvement of our knowledge and decisions.

Visual analytics is the science of analytical reasoning that takes advantage of capabilities of both human and data processing to achieve the most accurate and reliable results [Wong *et al*, 2012]. Visual analytics is interdisciplinary in that combines various related research areas such as visualization, machine learning, statistical analysis, cognition science, and human computer interaction (HCI). When human decision making is integrated with modelling analyses, there emerges opportunities to augment and increase the human's natural processing ability in terms of fostering (re)-evaluation of designs, implementation, and knowledge [Sun *et al*, 2013]. The main challenge of visual analytics approach is that visualization techniques are not flexible and interactive enough to effectively interface with complex data [Wong *et al*, 2012]. Novel techniques are a necessity for computational processing, visualizing different aspects of data, interacting with the information, and re-evaluating the knowledge. From this point of view, there is a need to develop new technologies in HCI side to address this challenge. That is a motivation for the research in this dissertation. In the following sections, a brief survey on visual analytics is presented.

2.1.1 Visual Analytics vs. Information/Data Visualization

The boundary between data visualization and visual analytics is not clear for most people. Data visualization is a field of study that uses graphical displays to present measured or/and simulated data and does not use statistical data analysis or machine learning algorithms [Weissgerber *et al*, 2016]. The objective of this field of science is to facilitate sense-making to understand and discover the hidden stories in the data and communicate/present such stories to particular audiences. In other words, data visualization enables us to understand data significance by placing it in a visual format instead of a verbal one and also share and exchange our insights with other people. The main challenge is that potential users could not express their needs directly to solve a defined problem.

Data visualization, which makes an effort to produce different views and create valuable communication among people, is a component of visual analytics. Indeed, visual analytics is a multi-disciplinary field of study that integrates machine learning algorithms with data visualization techniques interactively by having the users involved directly in the processing loop to understand various aspects of complex datasets and make new ideas and hypotheses [Keim *et al*, 2008]. This field of study opens the door to focus on the parts of problems that cannot be analyzed automatically because of massive, fuzzy, and conflicting nature of datasets. Visual analytics thus enables a user to combine the desirable information, indicate the expected outcomes, and discover unanticipated results.

While data visualization and machine learning are commonly used in various disciplines during recent years, advanced knowledge discovery and hypothesis making algorithms have been rarely utilized. The objective of visual analytics is to assistively explore hidden parts of each problem that cannot be elucidated or solved automatically. This field of research attempts to open a new window to solve the problems which researchers were not be able to find a feasible solution for them.

Visual Analytics Cycle

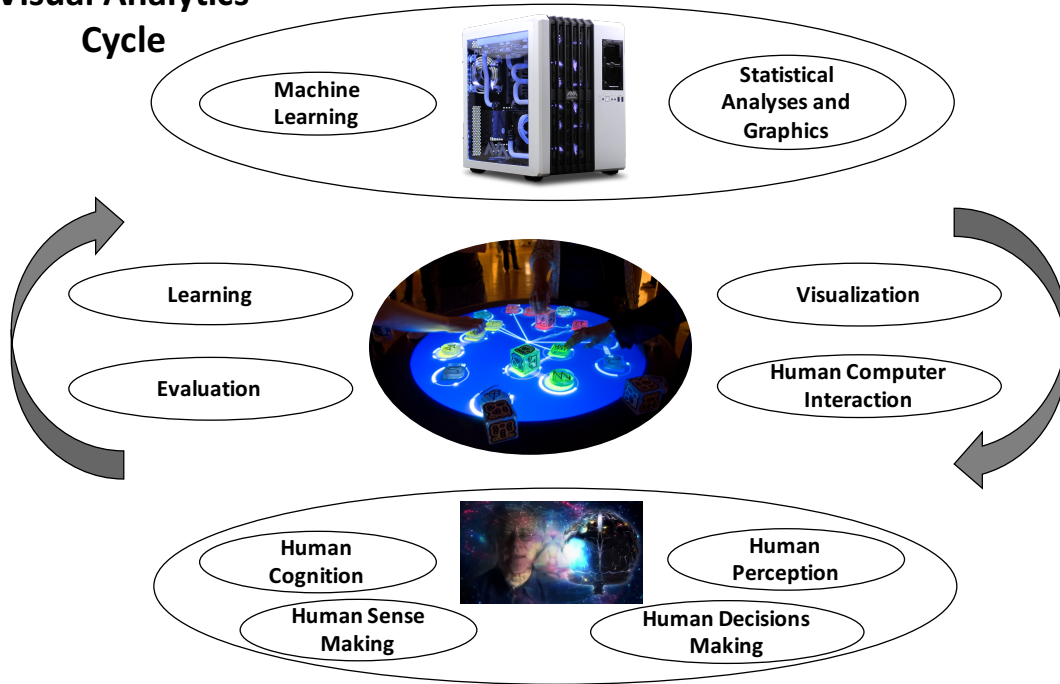


Figure 2.1: Visual analytics cycle

2.1.2 Visual Analytics related fields

This field of research draws inspiration from four sets of sciences [Brehmer *et al*, 2016] including visualization and data mining (machine driven), sense-making/cognitive science (human driven), and finally, communication/interaction (human/machine driven). Also, learning and assessment are built up in this intermediate area by utilizing the evaluation methodologies (Fig 2.1). In the following subsections, we briefly review four important disciplines.

Visualization

A general definition of visualization is any methodology for presenting an image, a diagram or an animation to communicate different ideas (such as gene sequence visualization [Waese *et al*, 2016], gene expression profile visualization [Ståhl *et al*, 2016], etc.). The reason why we use visualization is that it can address demands in three levels that these levels can be

categorized to "consume and produce", "search", and "query" [Kohlhammer *et al*, 2011].

In first level, data visualization could be used in order to not only consume the existing information but also produce new information. The information consumption could categorize in three aspects of presentation, discovery, and enjoyment. Presentation may occur in training, designing, decision making, and prediction processes [Marchionini *et al*, 2006], [Roth, 2013]. Discovery is focused on scientific inquiry [Bybee *et al*, 2005] for hypothesis generation and deciphering unexpected information [De Robertis *et al*, 2014]. Finally, in enjoyment, the target is to be faced with unexpected experience [Sprague *et al*, 2012] instead of making a decision or verifying a hypothesis from visualized data [Dörk *et al*, 2011, Sprague *et al*, 2012]. Visualization can be used to produce information through data generation, transformation, derivation, and interpretation [Willett *et al*, 2012]. For instance, in our tangible multivariate visualization project, a clinician who is familiar with clinicopathological features could produce new interactions by applying different operators on features; these interactions could be used in survival analysis.

The second level, without paying attention to first level, focuses on search. Searching has various categories including "looking up", "locating", and "exploring". Searching for a known reference in search environment is named looking up or locating [Andrienko, 2006]. As an example for searching the known cases, there are 200 clinical features which are categorized in 10 clusters; a clinician is looking for a specific feature and he has knowledge about these 10 clusters; therefore, searching in the feature space is named "looking up". However, a non-clinical person wants to do the same task without any prior knowledge in terms of the clusters; this category of searching is called "locating". On the other hand, searching for features instead of references is named "exploration". Features could be specific value, minimum and maximum of a function, patterns, paths, and the ranges [Andrienko, 2006]. An example for these cases, a biologist is searching in a gene regulatory network to find the shortest path between genes toward Identifying genetic determinants of longevity [Managbanag *et al*, 2008].

The third level is "querying" which occurs when one obtains an overview of data, and also identifies, compares, and interprets the discovered features. Querying increases the search number of considered target/s which can be referred to discover different aspects of a certain target, to

compare multiple targets, and interpret the relationships among targets [Tweedie, 1997].

Two important questions need to be addressed about the visualization. The first one is “what can be defined as inputs and outputs for visualization”. This question usually refers to the existing datasets, including the networks composed of nodes and edges [Shneiderman, 1996] (such as genes and their interactions in a gene regulatory network), the tables composed of samples and their characteristics [Shneiderman, 1996] (such as patients and their clinical features and time to events), and the temporal datasets composed from samples and their performance over time [Lopes and Bontempi, 2015] (such as genes, and their time series gene expression patterns). However, each individual according to his unique design thinking can suggest our own input and output. In other words, this question brings up unique mentality to define new problems according to the existing datasets or to suggest new data collection for surveying new input and outputs.

The second question is how the visualization technique supports a defined task. For answering to this question, “encoding”, “manipulation”, and “Introducing” actions are required [Shaw *et al*, 1998]. In encoding action, visualization techniques initially convert the data as a visual representation (e.g. scatterplot) [Munzner, 2014]. Manipulation action applies on visual representation obtained from encoding. It includes six subactions, “selection”, “navigation”, “change”, “filtering”, and “aggregation”. Selection action could cover from directly clicking on samples to highlight significant elements in representation [Weaver, 2007]. Navigation action could assist to alter the views. Arranging action refers to re-ordering the axes (such as arranging the samples based on a certain feature again). Change action varies the size and transparency of nodes or links in a network or the colour of samples in a scatterplot [Heer and Robertson, 2007]. Filtering actions pertain to temporary and permanent adjustments which allow one to add or remove some elements in visual representation (such as a set of nodes or edges carrying a certain characteristic). Aggregating action considers the level and scale of detail present in a dataset which allow one form a class or cluster. Aggregating action could be considered as a solution for dealing with big data. Introducing action allows one to add new elements to the representation and includes “annotating”, “importing”, and “recording” subactions. Annotating pertains to a note of explanation or comment for assigning

to one or several elements as a characteristic [Nazemi, 2016]. Importing refers to add new samples into the representation which could be obtained from the previous knowledge of an individual. Recording action points to saving and capturing actions (such as a snapshot from the achievements if they are desirable for further investigations) [Shrinivasan *et al*, 2008].

Data Mining

Data mining is an intersection of Artificial Intelligence (AI), Machine learning (ML), and Statistical Analysis (SA). AI and ML are subfields of computer science which refer to algorithms that can extract and learn the structures from the data with and without relying on rule-based programming respectively. Also, SA is a subfile of mathematics which focuses on data collection and organization and also information analysis and interpretation based on finding the relationships between variables and desired outcomes. Data mining is categorized into predictive and descriptive models [Bertolucci, 2013]. The predictive modelling is a supervised learning technique utilized to predict future behaviours of interested variables [Walsh *et al*, 2012] such as classification algorithms [Aggarwal, 2014], regression methods [Harrell, 2015], and etc. The descriptive model is a unsupervised learning process to discover the relationships of patterns in existing data [Patel and Patel, 2016] such as clustering methods [NafeesAhmed and Razak, 2014], correlation analyses ([Rodríguez-Fernández *et al*, 2016]), and etc. The predictive models are further used in healthcare research. Generally, extracting the structures is defined as an expectation from data mining. These structures may be real or/and not be what it purports to be. We should note that there is no distinction between these two types of obtained structures. Therefore, the challenge is how to reduce the number of spurious structure extracted [Ratner, 2011]. In addition, a user cannot adjust the parameters analytically in data mining. Many groups in different discipline use data mining techniques like neural networks [Hoyt *et al*, 2016], fuzzy logic [Singh and Wayal, 2012], support vector machines (SVMs) [Hamel, 2011], genetic algorithm [Verma and Vineeta 2012], genetic programming [Gandomi, 2016] and etc. One of these groups are the researchers in health care who use data mining increasingly [Chen *et al*, 2010, Liu *et al*, 2014, Chen, 2014]. The clinical

datasets are used to do survival analysis [Duan *et al*, 2011], improve the quality of treatment [Sun *et al*, 2013], etc. Generally, data mining has played a significant role in healthcare field especially in medical diagnosis. With a rapid growth of genetic and clinical data, an important challenge is to deal with this amount of data, explore useful information, extract knowledge, and make hypothesis for future actions [Hu *et al*, 2012]. Indeed, we need to integrate different datasets with various formats (such as continuous, binary, and string) to explore the relationships among the variables. It is not easy to extract the hidden patterns in big data because the quantity of data is massive but the quality is usually low [Chen *et al*, 2015]. Therefore, we need to develop data mining algorithms which not only analyze big data properties, but also could extract the relationships between these properties. Finding the highest accuracy by using these methodologies is a open challenge in medical diagnosis [Peker, 2016] and survival analysis [Montes-Torres *et al*, 2016] to assist the clinicians for applying the appropriate treatment. This challenge brings up due to massive, uncertain and incomplete nature of the big genetic and clinical datasets [Chen *et al*, 2015]. There is a need to develop hybrid methods to enhance the predictions by using data cleaning, filtering, and reduction and also finding incorrect samples and structures in data. In addition, parallel processing is necessity to deal with the massive size of data.

Cognition Science

Cognition is composed from three modules “learning”, “memory”, and “experience” to connect perception and action in a meaningful path. This categorization is obtained from Mike Denham’s definition for cognition [Mike Denham, 2016]. Indeed, perception is the gate of cognition. Concepts manipulation in the mind provides a better understanding in order to move forward to a target [Christian Bauckhage, 2016]. In addition, Patrick Courtney [euCognition, 2016] states that cognition ability could be improved over time using various senses (such as vision, touch, hearing, and etc) and communication to maintain motivations and achieve the targets. This improvement is affected by planning, inductive and/or deductive reasoning, and modifying capabilities to make a connection between variables. The commu-

Chapter 2. Background and Foundations

nication aspect of cognition is highlighted by [Katerina Pastra, 2016] who declares cognition is a ability to represent intentionality when engaged in communication and also understand purposive activities of others which named “Theory of Mind (ToM)” [Mahy *et al*, 2014]. In other words, ToM pertains to the cognitive ability to infer and understand other’s mental thoughts ([Martin *et al*, 2016]).

[Cecilio Angulo, 2016] thinks cognition ability derives from the nature of self-awareness processing of the human mind. He thinks consciousness is a sense of self-awareness and the communication aspect of cognition is caused by consciousness. Intention and decision have direct connections with consciousness [Dehaene, 2014]. Indeed, they are direct strikes of consciousness that what you intend to create and how you could make decision about how good and bad it is. Maria Petrou [Petrou *et al*, 2010] thinks the important thing in terms of cognition is that it could be understandable by human mind but could not be defined as a term or formula. If it was possible to define the cognition as a formula, it could be possible to implement it as a method in data mining. She states cognition is just a state of mind and could not consider it as a stable factor. In other world, human behaviour is not understandable and predictable at each moment.

Memory as one of main models of cognition is categorized in two different types "declarative" and "procedural". Declarative refers to facts and events memories which could be consciously declared and the kind of information handled is verbalized easily [Riedel and Blokland, 2016]. However, procedural is a part of long-term memory that knows how to perform certain procedures. Property of procedural memory is different and it is difficult to convert it into the words [Brill-Schuetz, 2014].

Human Computer Interaction

Human computer Interaction (HCI) is a multidisciplinary field of research which studies the interactions between human actors and actions and computer [Gaines and Monk 2015]. Indeed, HCI is an interactive system including three main components “Human”, “Computer”, and “Interaction”. Human could point a single user or a group of users in general or with a

specific expertise. Computer includes any computational devices which could embed human knowledge into existing tools. Interaction is every human activity done consciously and intentionally [Bakker *et al*, 2016]. HCI field consists of various activities such as the development of new algorithms for analysis, new design for applications, novel frameworks for new technologies design. In addition, the development of platforms that support the interaction between human and various systems is a branch of HCI. Moreover, HCI attempt to evaluate user experiments and new methodologies by considering various inputs and outputs sensors and devices. Therefore, the HCI function is related to input and output equipments (such as keyboard, mouse, touch screen and pattern recognition devices) and related software to control the system and understand commands and requests.

2.1.3 Visual Analytics Data Structures

Data structure is a specific way of data organization to effectively provide different amount of data in the databases. There are many types of data structure in general but three common ones are considered in this dissertation to be applied on three visual analytics platform.

Network Data

We are living inside of a connected world and all pieces of information around us could be considered as components of a network. Network data is a database that uses graph structures (such as nodes and edges and their possible properties) to explore information and knowledge [Pratt *et al*, 2015]. A highlighted aspect of this type of database is that the communication among the nodes in the graph is stored as edges. The nodes could be a representation for people, companies, and countries in social networks and genes, proteins, and enzymes in the biological networks. In addition, the edges are the samples for representing the relationship between the nodes and could allow to explore meaningful patterns from the network. Moreover, the properties could be related to the nodes and/or also edges. For example, if a given gene was one of the nodes in a biological network, biologists might highlight it as a member of a gene family that can be used for gene modularity visualization. Another

example is when there are information in terms of genetic and physical interactions among genes in the database and it can be used as edges properties on network representation to specify the obtained experimental results with and without genetic mutations. Also, this type of data might include information in terms of the direction and weight of flow along edges.

Categorical Data

Categorical data consists of a list of numeric values that can be divided into groups. These values might be observed directly from the events or might be measured within given intervals. For example, the clinical features related to the patients with a specific disease could be observed from a specific stage or could be measured in monthly intervals over running a treatment. There are two types of categorical data, dichotomous and polytomous [Millsap, 2012]. While dichotomous values are binary, polytomous variables have more than two possible variables.

While numerical format of data by using exact values are valuable, it could be more informative if we categorize these variables into small number of groups. The measured data might not be numeric (for example the name of applied treatment for patients) but it is possible to assign a quantity to these values. This leads to better analyze and manipulate data.

Time Series Data

Our world is multidimensional and time is one of these dimensions. Time series data refers to any data related to time. This type of data varies over time and represents the changes of an variable or a state over a period of time. Time values represent when an event occurs over time and time series data is a collection of the snapshots from an event over time [Fu, 2011]. Time series are used in many fields of knowledge scubas finance, economics, biology, and etc. There are three important questions about this type of data. First, is there any increasing or decreasing trend observed over time for a variable? Second, is there any unexpected changes occur for each state over time? If so, is there any relationships between these trends and unexpected changes for a variable with other variables in the database?

2.2 Situation Awareness, sense-making and mental models

2.2.1 Situation Awareness Definition

Situation Awareness (SA) is the ability of information extraction from the unstructured data, elements identification, mental model creation, and finally hypothesis making about future experiments [Green, 2008]. This concept is directly connected to understand and grasp something related to a situation which is mentally interpreted [Kokar and Endsley, 2012]. SA is a key element that is the result of human's attention and working memory [Gutzwiller and Clegg, 2013]. Indeed, in all human's activities, this concept is involved in interdependent decision making [Gonzalez and Wimisberg, 2007].

One of important literatures attempting to formalize the SA concept is presented by [Endsley, 1995] which has integrated SA as a three-level process between a certain situation and decision/action. In the first level, a decision maker should perceive the perceptual elements in the situation which are distinguishable without providing mental processing for the person. For example for a survival analysis case, a physician should identify the clinical and pathological features, the events such as death, and the applied treatments without any thinking. We should note that the elements could be replaced according to different people with various goals and knowledge [Macrae and Bodenhausen, 2001].

In second level, the decision maker should create a comprehension of the situation. In other words, the user integrates the elements with the intended goals. This level provides interpretation ability for the user to answer an important question about why these elements are important for achieving the given goals. For instance, in survival analysis case, the physician should find the relationships of clinicopathological features with each other, and assess the significance of these features, and distinguish the feature patterns. In this level, mental processing of the user plays an effective role to make the elements meaningful [Gheisari and Irizarry, 2011]. Mental processing help to recognize something related to the situation which might be stored in working or long-term memory of the user as a valuable knowledge and can be transferred to a new concept [Kihlstrom, 2011]. For example, the experience of physician which could be

an evidence for the significance of a clinical feature according to the defined goal. This level is important because a person without prior knowledge about the problem domain could not be able to make a mental concept and interpret.

In final level, the person performing the task should project the future status of a given situation. If one individual element could predict the other elements action or future status, the user might be able to predict what happens next and make hypotheses. As an example, the physician could predict hazard risk of patients with significant features and determine whether a treatment will be successful for a given disease or not. As it is obvious, having prior background and knowledge from the status is a need for performing mental activities such as making sense of causal structure, prediction and interpretation in this level [Johnson and Keil, 2014].

We should note that quality of data, urgency of the situation, and the expertise of user are really effective in performing SA. Consider we are presenting a situation for a user as a HCI task; we present the situation by a simple example for the user to perceive the necessity of the situation and also allow the user to manipulate some parts of situation according to his skills and experiences. Two activities are happening during this process. First one is related to everything which the user is doing by interacting with the situation and second one is about everything which is happening by system independently or in response to the user's action [Zhang and Norman, 1994]. The information is extracted from the interaction between user and situation [Groome, 2016]. This information is assigned to each user and could be including several levels of meaning which might attract different users's attention [Wickens *et al*, 2015]. According to the SA process, second and third levels are important in terms of decision making. Also, sense-making could be considered as a series of strategies to explore effectively the connections between the elements and defined events. The next section discusses background literature on sense-making process.

2.2.2 Sense-making Definition

Everyday we face with the sense-making process to make our environment meaningful. When a person faces with a new situation which has been unknown or uncertain for him, sense-

making process is getting activate to understand the situation, make clear unknown events, fill the existing mental gap, and enable to make intuitive decisions [Dervin, 1992]. The activity of seeking and finding gaps involves learning. The sense-making process targets to select and combine the elements in the situation somehow make the user's mental figure meaningful. Sense-making is a multi-level procedure, including cognitive process, metacognitive process, and collaborative process. The cognitive process is a cognitive activity performance that selects and combines the elements of a situation to affect the user's mental contents for improving the thinking process, remembering something, and learning [Sorden, 2012]. The metacognitive process is a thinking about thinking procedure in which occurs real learning [Magno, 2010]. Metacognitive strategies manage the selections, combinations and interpretations occurred in cognitive process. The collaborative process uses the HCI technologies to create a platform for interdisciplinary people to share their perspectives in order to manage cognitive activities and the proper decisions usually occurs here [Azuma *et al*, 2006].

Sense-making is an internal mental process which is not possible to measure it directly. Sense-making could be categorized in two classes [Fatemieh *et al*, 2010]. First class is individual make-sensing. While the individual user manipulates the elements of a situation, it might not only be identified by the system to record the sense as an output of sense-making and apply it on future status, but data collection is also biased by the individual's experience and experience [Wahl *et al*, 2016]. Therefore, an important question comes through here how sense could be measured and assessed. We should pay attention that the structure for storing any sense or decision does not record every experience in detail but summarize it into a framework to express the user's perspective. Sense could be a reason for why an element selection is done in a given situation or how a combination is selected as a determinative factor on prediction of an event or an verification on what an element or a relationship might be meaningful and significant. These explored information could be merged into an existing framework or leads to generate a new framework. The feedback from the framework could help the user to consider or dismiss the obtained information. Non-fit information in the framework could be considered as an evidence to reform or change the criteria. The second class is collaborative sense-making which is explored when two or more users together inter-

act with the system [Baber *et al*, 2016]. We should note that sense might not be made upon a shared understanding that can be considered as an achievement in terms of surveying the quality of data, reviewing situation definition or assessing the expertise of users about the situation. collaboration provides a platform for sharing the hypotheses and comparing the frames used to make the decision and hypothesis. In this dissertation, we have presented three interactive platforms to collaboratively explore and analyze the biological and clinical systems by considering multi-level sense-making procedure in design. In chapter 4, we are visualizing the sense-making network as a set of individual sense-making processes in various iterations. In interdisciplinary research, sense making is a way to obtain a shared understanding from the situation that support the plausibility idea. The obtained sense should be plausible enough for all members in the team to move toward a decision or hypothesis [Mikesa and Kaplan, 2014]. The sense-makers should explore with considering a question if the actions would be considered plausible by their other colleagues. As it is obvious, the plausibility criterion should be accepted by people with various expertise, backgrounds and perspectives. In other words, this plausibility is provided by an evaluation frame and consensus will obtain based on the agreement of enough team members. The process of achieving to a good consensus is challenging and important. The main challenge is that the frames should be acceptable for all sense-makers in the team or at least they should have access to their suitable evaluation terms. Three different inferences occur when a person attempts to solve a problem. One of them does not create any new mental perception. In other words, you, as a sense-maker, interact with the elements in a given situation and the obtained inference is completely match with your expectations and previous knowledge [Maitlis and Sonenshein, 1988]. According to your perspective, everything is clear and there is no reasoning and sense-making required for situation interpretation based on the elements. It could be useful for evaluation system assessment. However, consider another inference when the decision-maker faces with unexpected results [Karni and Vierø, 2014]. This inference includes two states; i) it not only confirms the prior knowledge but also shows hidden aspects not noted or discovered before. ii) it might be partially or completely unmatched with the expectations which forces the sense-maker to expend mental effort on interpreting the obtained results, keeping or ignoring the new inferences,

retaining the inconsistent thoughts [Tourish and Robson, 2006].

A Data/Frame Theory of sense making is presented by [Klein *et al*, 2006] that has emphasized several point of views. They present sense-making as a process of fitting the information into a frame and also fitting frame around the obtained information. The information is inferred using the frame instead of recording a mental perception. The frame is extracted from some data elements. The sense-making relies on insight and creative problem solving and also logical inference. Sense-making is stopped when the data and frame confirm each others. The users with more experience might argue almost similar with novices but the obtained set of frames is richer in comparison. The achievement of sense making would be what the user should do in a situation and also an abstract mental understanding. The users prefer to make mental models according to their expertise instead of entering to the other domains. Sense-making takes different forms and activities such as reframing, questioning the frame, elaborating the frame, preserving the frame, seeking a frame and comparing the frame [Sieck *et al*, 2007]. The only weakness about this theory is how the collaboration sense-making process could apply to this definition. In this dissertation, we present Tangible Tensors as a novel collaborative and interactive visualization platform for reconstructing biological pathways where tangible user interface (TUI) technology has assisted us to not only apply all steps of this theory but also attempt to fill its collaboration weakness.

According to the above discussion, we can conclude that a key factor in sense-making process is a mental model to add or subtract the obtained information from manipulating elements in a situation. The next section surveys this subject.

2.2.3 Mental Models Definition

A mental model is an explanation of people's thoughts in reality relate to elements, situation, and inferences [Khella, 2002]. During situation exploration, a sense-maker who manipulates the system to predict the future of a situation, remodel his mental figure simultaneously. The user's understanding of a situation is effective on the quality of his mental model [Nakarada-Kordic *et al*, 2016]. For example, a biologist could perform on surveying biologi-

cal pathways better than non-biologist people because he has a structured memory and an organized knowledge related to molecules expression, their interactions, and their effects on various events. Therefore, they could monitor the process better and perform faster, and interpret the results more correctly rather than non-expert people [Bruer, 1993]. In general, they could make a better mental figure. We should note that if the design structure is far from the format in expert mind, they could not interact with the system and would not be able to apply memorized patterns that they have had experienced with [Sargent, 2013].

As we pointed before, the mental models are not directly explainable and consequently not easily storable. Designing a general formularization to capture and store the mental model is a open challenge [Birkhofer, 2011]. Indeed, each formularization has pros and cons which can make it more or less appropriate to capture the inference knowledge from the user's memory [Whitenton, 2013]. In general there are two types of knowledge to capture. One of them is procedural knowledge which is extracted during performing some tasks in situation [Rittle-Johnson *et al*, 2014]. Another one is declarative knowledge which refers to the factual information that a user knows based on his experience [McNeil, 2015]. We should pay attention that mental models are updated when user interacts with the system according to the user performance and inference. In mental models updating process, existing knowledge could be modified or changed and new aspects and relationships might be added. Following the trend of the mental changes could leads to improve the learning and inference abilities.

In complex problems (such as network science), the target might not be defined or be really fuzzy in the beginning. This happens due to people do not know what they are looking for. Therefore, an interactive representation could allow users to explore the different aspects of data and make a mental model from problem space is essential. The reason is that a user needs to learn and increase his information about the problem space and then he can define a specific goal according to the obtained mental model.

2.3 Tangible User Interfaces (TUIs)

In this section, a brief background about Tangible User Interfaces (TUIs) is presented. The theory behind of TUIs is published in many literatures. These literatures review interaction techniques [Hinckley, 2014, Van Beurden, 2013, Furtado De Mendonca Monco, 2015, Gulliksson, 2012, Wiethoff, 2012], introduce the combination of activity theory and interaction design [Daniels *et al*, 2013, Kaptelinin and Nardi, 2012], movement-based interaction design [Loke and Robertson, 2013, Cruz Mendoza, 2015] and etc. We cannot present a detailed study of TUIs in this short section but we introduce the important aspects.

2.3.1 Graphical User Interfaces vs. Tangible User Interfaces

Interactions with digital information are largely limited to Graphical User Interfaces (GUIs) devices such as personal computers and laptops and etc. GUI has existed from the 70's and was appeared for the first time in Xerox commercially [Smith *et al*, 1987]. Nowadays, GUI has become as a symbol for human computer interaction and it has regularly used on Microsoft and Macintosh systems. Generally, GUIs visualize a system on a display and this graphical representation could be manipulated with mouse and keyboard to explore design alternatives as improvable models. However, mice and keyboards create indirect interaction instead of physical interaction with the system and GUI is separated from the way interaction is accomplished in physical world by human. In addition, although GUI could give this chance to one person to use a device for doing several tasks, its collaboration ability is effected [Magerkurth and Tandler, 2002].

On the other hand, the exponential growth of digital and physical technologies leads to develop the interfaces that allow us to interact with computers via tangibles in the same way that we interact with our environment. These technologies have presented Tangible User Interfaces (TUIs). TUI is built according to human skills in sensing and manipulating of physical world. For the first time, TUI is presented by [Ullmer and Ishii, 1997, Ullmer, 2000] as a interaction model which is named Model-Control-Representation physical and digital (MCRpd). This

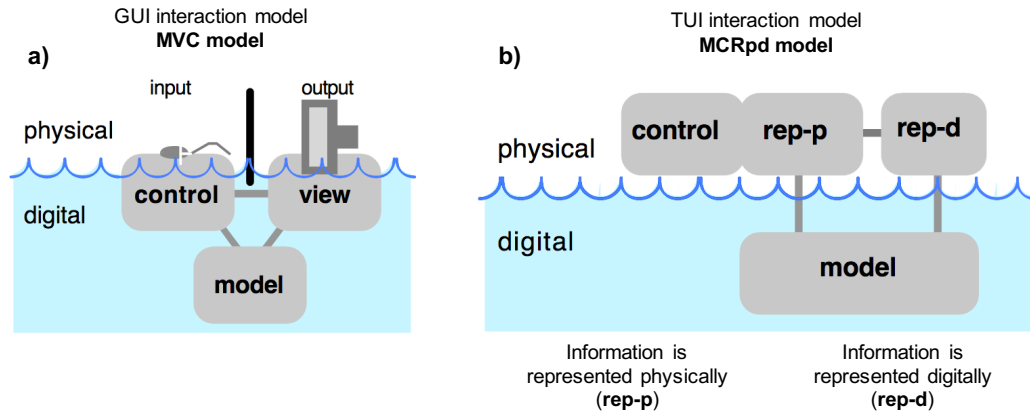


Figure 2.2: GUI and TUI interaction models - a) MVC model; b) MCRpd model [Ullmer, 2000]

model is an extension for Model-View-Controller (MVC) model used for GUIs [Burbeck, 1992]. While MVC model separates digital representation from control by GUI's mice and keyboards (figure 2.2a), MCRpd integrates representation and control in both conceptual and physical levels [Ullmer, 2000] (figure 2.2b). Representing physically (rep-p) and digitally (rep-d) are shown as tangible objects and video projection respectively. According to TUI model in [Ullmer, 2000], visual feedback is a part of proposed model but several TUI systems have been developed which use haptic feedback instead of visual feedback [Velloso *et al*, 2015]. However, our focus on this dissertation is on TUI interaction model with visual feedback. TUI provides an environment to manipulate the information with our hands and it could be an alternative to graphical interfaces. The quality of a TUI system is determined by the strength of coupling between the TUI and the task it is designed to support, rather than the coupling between physical and digital representations [Sharlin *et al*, 2004].

2.3.2 Interactions

Tangible Interaction

Tangible interaction pursues a terminology that focuses on user experience and interaction with a system [Baskinger and Gross, 2010, E. Van Den Hoven *et al*, 2007]. Tangible inter-

actions are social and collaborative and these characteristics are important in designing TUIs [Hornecker and Buur, 2006, Shaer and Hornecker, 2010]. Social interaction, collaborative learning, and sharing knowledge & skills as a nature of TUIs provide a different perspective of design in comparison with GUIs [Fernaes *et al*, 2008]. Indeed, the framework suggested by [Hornecker and Buur, 2006] emphasizes on the characteristics of TUIs, helps to understand and compare them between different systems. A interaction designer could develop tangible interactions which can be related to physical objects or a gesture. Tangible interactions include basic & complex techniques and gestures. Basic techniques are basic interaction styles with the tangibles, including dragging, rotating, and etc. Complex techniques are complex interaction styles with the tangibles, including dragging & strobing, stacking, and etc. Gestures are physical events linked to hands and body actions [Xambó 2015].

Situated Interaction

An system-oriented definition of context is presented in [Abowd *et al*, 1999, Dey, 2001, Baldauf, 2007]. The system is able to detect multiple objects or people in the situation and track their poses (for example object's orientation toward another object and a person). This information related to objects, users, system, and their interactions is named context. The system can understand situation (surroundings people and social settings) just like a person would. Situation determines how a user acts with physical world, what are the commonalities, and how contexts help the user to select an action as the best suitable [McCullough, 2001, Jumisko-Pyykkö and Vainio, 2012]. Situated interactions are finding the ways to capture the contexts that affect the users' ability to attain their planned activities. In other word, they are efficient interactions between several users and situation-aware system.

Hands-on Interaction

One of main factors to develop intelligent behaviour is physical actions in the world [Piaget, 1952]. Children could understand and solve a problem by acting physically and without expressing the concepts verbally [Goldin-Meadow and Wagner, 2014]. This fact used in traditional edu-

cation systems is called “hands-on” which emphasizes on the cognition based on physical engagement [Dewey, 1938, Bull, 2005, Kolb, 2014]. TUI is a “hands-on” interface in which the tangibles and multitouch surfaces play the role of physical objects [Shaer and Hornecker, 2010, North *et al*, 2009]. Theoretical foundation of hands-on interaction design is presented in field of HCI to investigate the cooperation of the reactions of hands [Fitzmaurice *et al*, 1995, Hauptmann, 1989, Kurtenbach, 1997]. Hands-on gestures allow a user to do two tasks in parallel [Buxton and Myers, 1986]. In addition, they could improve the cognitive performance of a user in terms of thinking about a task [Hinckley *et al*, 1997, Hinckley *et al*, 1994]. The effect of hands-on interactions needs to be explored more to consider in designing novel interfaces that can find the solution in various domains.

2.3.3 TUIs Survey - Strengths and Limitations

TUIs have beneficial aspects in comparison with traditional user interfaces. Shaer and Hornecker have highlighted these aspects in their survey [Shaer and Hornecker, 2010]. We have pointed out most important strengths such as collaboration, situatedness, tangible thinking, and Space-Multiplexing in the following:

- **Support for collaboration:** Mouse, keyboard, and display as GUI interfaces could be utilized only by an individual user. In contrast, it has shown in different applications [Maquil *et al*, 2012, Schneider and Blikstein, 2014, Flyckt, 2013] that TUI-based systems are a suitable choice for planning the collaboration context. Indeed, TUIs make the intended data and model accessible to multiple users to manipulate them at the same time. The observation of the performed physical interactions applied by different users assists to improve individual and group awareness and coordinate their group’s interaction [Klemmer *et al*, 2006]. The users’ familiarity with the physical interactions and also the affordances of physical tangibles for effective manipulation are two important factors that should consider in collaboration tasks [Yiannoudes, 2016].
- **Situatedness:** Human could use the passive and active tangibles situated in the physical

context to retain the superiority of physical world [Van Campenhout *et al*, 2013]. Indeed, situatedness is the ability of human to interact with physical world. “Physicality plays an important role in interpersonal communication” [Brave *et al*, 1998] and is a aspect of TUIs. According to this context, situatedness provides fundamental information in terms of multiple users interaction and assists to design interfaces which capture physicality [Dourish, 2004, ?, Fernaeus *et al*, 2008].

- **Support for Tangible Thinking:** Amy Leidtke defines tangible thinking as an exhibit that inquires how a person utilizes various methods (such as observation, perception, experimentation, and visualization) to explore, understand, interpret and communicate concepts and systems in different domains. TUIs attempt to support thinking based on this definition [Klemmer *et al*, 2006].
- **Space-Multiplexing:** We are living in a world defined based on space and time The analytic philosophy was inspired from space and time philosophy and was mainly focused on whether these two are independently existed from each other in the mind or not [Beaney, 2013]. The analytic tools and technical devices (such as TUIs) we use to interact and navigate the information depend on space or time. Therefore, input devices are classified to time-multiplexing and space-multiplexing [Fitzmaurice and Buxton, 1997]. While time-multiplexing provides one controller for all functions over time, space-multiplexing offers one device for each function which allows to apply different functions simultaneously [Schaper, 2013]. For example, a mouse performs as a time-multiplexing input device and controls various actions in GUI at different time points. In contrast, users could manipulate different parameters in different positions of multitouch surface simultaneously and independently by using a TUI-based system [Fitzmaurice, 1996].

Share and Honecker have surveyed the limitations of TUIs along to their benefits. TUIs limitations include scalability, versatility, malleability, and user fatigue [Shaer and Hornecker, 2010] which are pointed in the following briefly.

- **Scalability:** One of limitations of TUI is in terms of scalability [Shaer and Hornecker, 2010].

During recent two decades, research on TUIs has grown up but unfortunately the developed TUI-based systems are usually used for small studies and are rarely scaled up to real and large problems. The reason is that the large number of manipulations does not allow to handle the complex commands easily [Bellotti *et al*, 2002, Hornecker, 2012]. For example, when we are going to compare different solutions of a problem on the screen, it might be impossible because of physical space limitation based on the size of multitouch screen [Edge and Blackwell, 2006, Gelineck *et al*, 2013]. It could be considered as a mutability limitation of TUI-based systems on supporting the tangibles usage for applying adequate manipulations in comparison with GUI systems which their screens could be scrolled by a mouse [Bergner *et al*, 2011]. In this dissertation, we tackle with this limitation by creating three web applications in chapter 3, 4, and 5 based on HTML5 standards because these standards are more compatible with other browsers, Javascript is recently used more due to access to open source 2D and 3D libraries, it assists to increase the scalability of TUIs.

- **Malleability and Versatility:** Digital objects in comparison with physical objects are easy to build, modify and distribute. In other words, GUIs are more malleable and versatile in comparison with TUIs [Riedenklau, 2014]. While GUIs allow users to undo an action automatically, show a actions history, and replay a scenario, TUIs are unable to survey applied scenarios and invert them [Carandang and Campbell, 2013]. We attempt to address this limitation by proposing Tangible Tensors platform in Chapter 5. In addition, although GUIs are general-purpose and can be utilized for a broad range of tasks due to their flexibility and adaptability, TUIs are typically special-purpose.

3 Model Refinement and Inference, and Interactive Biological Networks Visualization

In this chapter we represent three contributions. In the first contribution, the stability criteria is modified in reconstruction and a framework is developed for studying gene interactions inference using a combination of sparsity, prior knowledge and stability constraints. The second contribution of this dissertation is an inference from the exploration of various networks. A novel method is developed which not only consider two universal properties of the networks such as sparsity, low-rank property in optimization process but also consider modularity property. The modularity property implies that if a sample point or node is placed in a cluster in the network, its neighbouring sample points or nodes would also be placed in that cluster with high probability. The obtained results are confirmed by the analysis of laplacian spectrum of the reconstructed model. The third contribution relates the fields of sense-making and mental models with Human Computer Interaction (HCI) and network science from where the concept comes from or what it really means to argue why prior knowledge or visual experience of the viewer is necessary for defining structural constraints in biological network reconstruction. This contribution provides a tangible and collaborative platform in which the users can apply well-designed statistical methods associated with interactive visual representations to achieve qualitative and quantitative comparative analyses and also knowledge extraction. The obtained results from this toolbox could improve the visual and conceptual understanding from the structural features of biological networks and prevent fault or incomplete constraints

to reconstruct gene regulatory networks. This contribution opens the door to design and modify the constraints inspired from the reality in model reconstruction.

3.1 Stability Constraint Modification

Research in bioinformatics and computational biology is faced with a vast volume of uncertainty based on the stochastic and fuzzy nature of biological systems. For instance, growth and development as well as environmental stresses can all contribute to change in gene expression levels. In addition, under such conditions, some genes influence the expression of other genes and their functionalities.

Nowadays, the ability to measure the expression of genes on a genome-wide scale has grabbed biologists' attention for deciphering the dynamic of gene interaction networks based on time-series gene expressions. Thus, optimized methods are needed to uncover the role of many genes with uncertain functions from discrete datasets of continuous biological processes.

The goal of optimization is estimating and designing a system in the most effective way possible, and as such, forms the basis of much research addressing the reconstruction of gene regulatory networks. In this application, finding the best compromise among several conflicting demands subject to predetermined constraints is considered. Indeed, optimization not only allows biologists to realize current adaptations of biological systems, but also helps them to predict new designs that may have yet to be encountered [Alexander, 1996, Sutherland, 2005]. Since a significant role in the progress of systems biology is played by engineering methods [Doyle, 2006, Kremling, 2007, Wolkenhauer, 2005, Sontag, 2005], it is expected that model-based optimization as a key methodology that will contribute to this evolution.

A remarkable question pertains to which criteria (i.e. objective functions) are being optimized in genetic networks [Schuetz, 2007, Nielsen, 2007]. In [Alon, 2007], the question of how constrained evolutionary optimization has been used to understand optimal biological system design is studied. Moreover, complexity and robustness in biochemical networks have been better understood and explained by optimization principles [Stelling, 2004, Carlson and Doyle, 2002, Tanaka, 2005].

Optimization has been used for inferring important bio-molecular networks, such as transcriptional regulatory networks [Wang, 2007] and gene regulatory networks (GRNs) [Wang, 2006, Thomas, 2007, Cho, 2007].

Due to the time consuming nature in which samples are produced and the high costs associated with biological experiments, especially in clinical studies, it would be perfect if one could minimize cost while maximizing the amount of information to extract from biological experiments. Indeed, we need efficient multi-criteria algorithms for optimization under uncertainty and complexity to better scale up from the level of the genome.

In the rest of this part, we have developed an algorithm for reconstructing the sparsest gene interconnection network using time-series genetic experimental data. To avoid the concave (nonlinear) nature of the problem, we employ a weighted convex relaxation, which leads to a conversion of our concave problem into one that is convex. On one hand, our reconstruction algorithm is able to use known a-priori knowledge about the gene interactions network. Using this prior knowledge, we can realize whether one gene affects another gene or not, or whether this effect is positive (activation) or negative (inhibition). In addition, stability and sparsity of the genetic networks and the best of our knowledge about these networks are considered in the modeling formulation. Indeed, the convex optimization formulation in our approach is preserved by the stability condition, which we capture by linear constraints that arise from Gershgorin's theorem. Further linear constraints are also considered by achieving both a best fit on the genetic data while also satisfying a priori knowledge on the gene interconnected network.

The rest of this section is organized as follows. We describe the genetic network reconstruction problem, develop the proposed convex relaxation and discuss the stability issue. Finally, we illustrate efficiency of our approach on experimental data for the yeast cell-cycle.

3.1.1 Reconstruction Algorithm

Ordinary Differential Equations Model

The transcription and translation sub-processes in a regulatory network consisting of n genes can be represented as the following system:

$$\begin{cases} \dot{g} = Dg + T\mathbb{R} \\ \mathbb{R} = f(g) \end{cases} \quad (3.1)$$

where: $g = [g_1, g_2, \dots, g_n]^T \in R^n$ represents the concentration of mRNAs of n genes; $D = \text{diag}[-d_1, -d_2, \dots, -d_n] \in R^{n \times n}$ with d_i as the degradation rate of gene i ; $\mathbb{R} = [r_1, r_2, \dots, r_n]^T \in R^n$ represents the reaction rates; and $T \in R^{n \times n}$ is the reaction topology of the regulatory network. Assuming the reaction rate \mathbb{R} is a linear combination of g and $E \in R^{n \times n}$ is the coefficient matrix. Then,

$$\mathbb{R} = Eg \quad (3.2)$$

Thus, we have,

$$\dot{g} = Dg + TEg \quad (3.3)$$

A standard discretization method like zero-order hold (ZOH) is used in this case on observation points for a sampling time Δt , thus:

$$g[k+1] = (e^{D\Delta t} + (e^{D\Delta t} - 1)D^{-1}TE)g[k] = Wg[k] \quad (3.4)$$

where TE describes the structure of gene regulatory network in following: $te_{ij} > 0$, $te_{ij} = 0$, and $te_{ij} < 0$ if gene j activates gene i , does not regulate gene i , and represses gene i respectively. To reconstruct the gene regulatory network, we need to determine the sign of elements in SE . Both matrices $e^{D\Delta t}$ and $(e^{D\Delta t} - 1)D^{-1}$ are diagonal and all diagonal elements of these two matrices are positive numbers. Consequently, the sign of all elements of matrix

W are the same as the corresponding elements in SE . Therefore, reconstruction of gene regulatory networks from gene expression patterns is akin to estimating the elements of matrix W .

Problem Formulation

The method that we propose in this part is based on convex optimization which minimizes a convex cost function over feasible solutions. Reconstruction problem formulation as convex optimization is suitable for researchers because there are useful techniques to solve problems in convex form efficiently [Boyd, 2004]. Key observations in modeling a gene regulatory network that we consider in our formulation are as follows:

1) Error Constraint Let $G \in R^{n \times m}$ be gene expression levels of genes at time points. Our problem is divided into $m - 1$ sub-problems (as shown Figure 3.1) solved individually. Let $G^{(p)}$ and $G^{(p+1)}$ ($p = 1, \dots, m - 1$) be the p th and $p + 1$ th columns of G , respectively. If we assume that the measurements are noise free, we can consider:

$$G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)} = 0 \quad (3.5)$$

However, a realistic assumption is in the presence of noise. Consequently, we do not assume that 3.5 is zero. Instead, we define:

$$G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)} = \epsilon^{(p)} \quad \epsilon = [\epsilon^{(1)}, \epsilon^{(2)}, \dots, \epsilon^{(m-1)}] \in R^{n \times (m-1)} \quad (3.6)$$

and try to minimize $\epsilon^{(p)}$ as a function of $W^{(p \rightarrow p+1)}$, while obtaining a minimal model for $W^{(p \rightarrow p+1)}$ and satisfying any a priori constraints that might be imposed on $W^{(p \rightarrow p+1)}$. In this step, we establish the error level that a model can attain. For instance, we employ the total

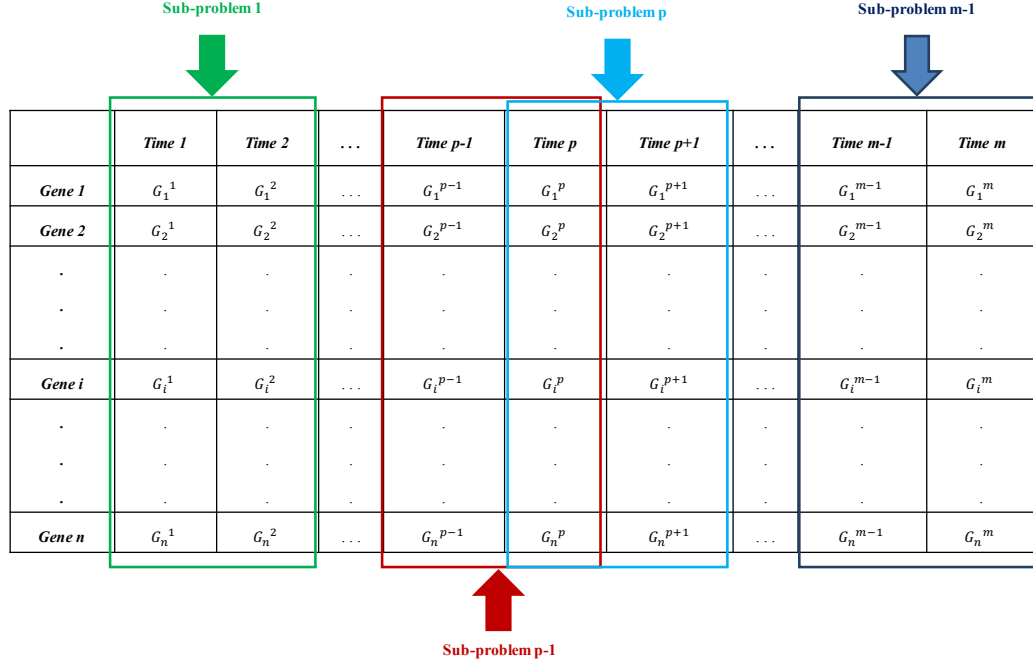


Figure 3.1: A sample problem segmentation.

squared error as the error criterion:

$$\|G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)}\|_2 < \epsilon^{(p)} \quad (3.7)$$

2) Sparsity Constraint The requirement that $W^{(p \rightarrow p+1)}$ be sparse is related to biological networks being sparse in nature. Thus, we consider sparsity constraints for our formulation. To reach this target, we use a cardinality function defined as the number of non-zero components of a matrix, i.e.,

$$Card(X) = \sum_{i,j=1}^n I(x_{ij}) \quad (3.8)$$

where I is an indicator function defined as:

$$I(z) = \begin{cases} 1, & z \neq 0 \\ 0, & \text{Otherwise} \end{cases} \quad (3.9)$$

In this step, the problem of reconstructing the sparsest W that satisfies the error constraints can be formulated as the following constrained optimization problem:

$$\begin{cases} \text{minimize} & \mu \text{Card}(W^{(p \rightarrow p+1)}) + (1 - \mu)\epsilon^{(p)} \\ \text{subject to} & \|G^{(p+1)} - W^{(p \rightarrow p+1)}G^{(p)}\|_2 < \epsilon^{(p)} \\ & \epsilon^{(p)} > 0 \end{cases} \quad (3.10)$$

The problem data are the matrix G and the parameter $0 < \mu < 1$ while the matrix $W^{(p \rightarrow p+1)}$ and fitting error $\epsilon^{(p)}$ are variables. μ is used to control the trade-off between sparsity ($\text{Card}(W^{(p \rightarrow p+1)})$), and best fitting ($\epsilon^{(p)}$).

Weighted l_1 -norm relaxation: The sparsity function ($\text{Card}(W^{(p \rightarrow p+1)})$) in the objective function is concave. It is possible to solve this problem by some methods [Boyd, 2013] but these methods are very slow, and cannot apply to medium and large networks. Thus, a convex relaxation of the cardinality cost function is used to relax this concave function to a convex one thereby allowing for the application of this method to large networks. We replace the sparsity part of the objective function with the weighted l_1 -norm as follows (Figure 3.2):

$$\text{Card}(W^{(p \rightarrow p+1)}) = \sum_{i,j=1}^n b_{ij} |w_{ij}| \quad (3.11)$$

where b_{ij} s are chosen by:

$$b_{ij} = \left(\frac{0.1}{0.1 + |w_{ij}|} \right) \quad (3.12)$$

The main idea is to replace $\text{Card}(W^{(p \rightarrow p+1)})$ with 3.11, initialize the b_{ij} values, and then solve

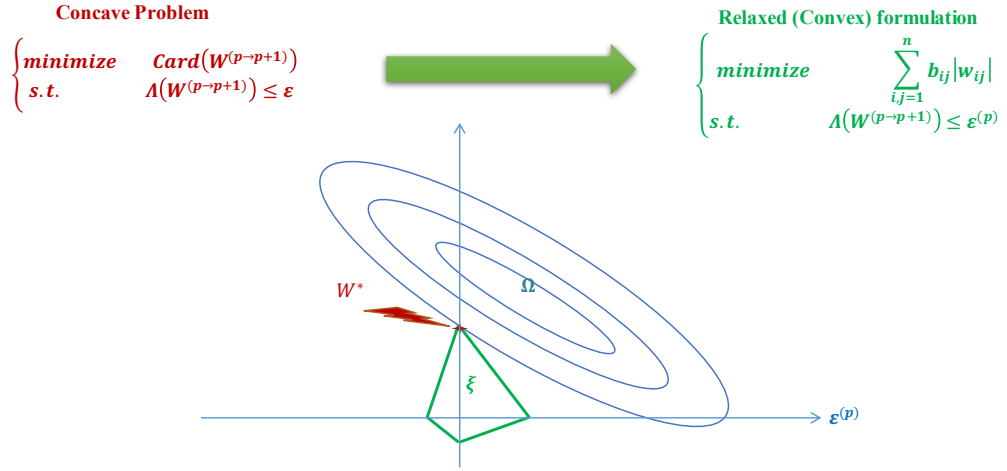


Figure 3.2: Weighted l_1 – norm relaxation interpretation and its solution: To illustrate the property of weighted l_1 – norm relaxation, let us consider the problem of finding the sparsest interaction network that provides a given level of performance $\epsilon^{(p)} \geq 0$. The solution of the relaxed formulation is the intersection of the constraints set $\Omega = \{W^{(p \rightarrow p+1)} | \Lambda(W^{(p \rightarrow p+1)}) \leq \epsilon^{(p)}\}$ and the smallest sub-level set of $\xi = \sum_{i,j=1}^n b_{ij} |w_{ij}|$ that touches Ω .

problem 3.10 and update the b_{ij} values frequently using 3.12. Consequently, we can clean the final inferred $W^{(p \rightarrow p+1)}$ from genetic interactions with weak connectivity. Therefore, our formulation is changed as follows:

$$\begin{cases} \text{minimize} & \mu \sum_{i,j=1}^n b_{ij} |w_{ij}| + (1 - \mu) \epsilon^{(p)} \\ \text{subject to} & \|G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)}\|_2 < \epsilon^{(p)} \\ & \epsilon^{(p)} > 0 \end{cases} \quad (3.13)$$

3) Prior Knowledge Constraint If a priori knowledge about the network is available, we can consider it as a sign matrix $H = (h_{ij}) \in \{0, +, -, ?\}^{n \times n}$, for which positive, negative and no interactions between any two genes are signed as (+), (-) and (0) respectively. In addition, if we do not have any prior knowledge about interactions it will be shown by (?) in the network. In this step, we consider prior knowledge in equation 3.13 with the use of some linear constraints

as follows:

$$W^{(p \rightarrow p+1)} \in H \Leftrightarrow \begin{cases} w_{ij} > 0 & \text{if } h_{ij} = + \\ w_{ij} < 0 & \text{if } h_{ij} = - \\ w_{ij} = 0 & \text{if } h_{ij} = 0 \\ w_{ij} \in 0 & \text{if } h_{ij} = \{+ \text{ or } - \text{ or } 0\} = ? \end{cases} \quad (3.14)$$

This results in the problem:

$$\begin{cases} \text{minimize} & \mu \sum_{i,j=1}^n b_{ij} |w_{ij}| + (1 - \mu) \epsilon^{(p)} \\ \text{subject to} & \|G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)}\|_2 < \epsilon^{(p)} \\ & \epsilon^{(p)} > 0 \\ & W^{(p \rightarrow p+1)} \in H \end{cases} \quad (3.15)$$

4) Stability Constraint From both theoretical and experimental viewpoints, stability behavior of GRNs has an important biological implication and a potential engineering application. Stability of linear dynamical systems can be determined from its eigenvalues. To show stability, it suffices to show that the eigenvalues of $W^{(p \rightarrow p+1)}$ are within the unit circle. To verify this for $W^{(p \rightarrow p+1)}$, we apply the Gershgorin theorem [Golub, 2013]. Gershgorin theorem provides a region for the eigenvalues of any matrix. The version of this which will be sufficient for our purpose is the following:

Theorem: if $W^{(p \rightarrow p+1)}$ is an $n \times n$ square matrix and $\lambda_i (i = 1, \dots, n)$ are the eigenvalues of $W^{(p \rightarrow p+1)}$, then every eigenvalue of $W^{(p \rightarrow p+1)}$ lies inside or at the boundary of at least one of the circle in the complex-plane defined as:

$$|z - w_{ii}| \leq C_i \quad (3.16)$$

$$C_i = \sum_{j=1, j \neq i}^n |w_{ij}|, \quad i = 1, \dots, n \quad (3.17)$$

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

Around every element w_{ii} on the diagonal of the matrix, a circle with radius equal to the sum of the norms of the other elements in the same row (C_i) is known as a Gershgorin closed disk (D_j). Every eigenvalue of a square matrix W lies in one of its Gershgorin disks. Then all eigenvalues lie within $\bigcup_j D_j$. The theorem provides the region of the eigenvalues of a matrix, and it can be used as a stability or asymptotic stability condition. State space models for discrete-time systems are stable if all eigenvalues of the state matrix lie inside the unit circle. Hence, if the all Gershgorin circles of the state matrix are inside the unit circle, then all eigenvalues are inside unit circle, and the system is stable (in this case, the eigenvalues are on the real line) (Figure 3.3). Indeed, a state matrix is stable if the following conditions hold for all $i = 1, \dots, n$:

$$\begin{cases} -\sum_{j=1, j \neq i}^n |w_{ij}| + w_{ii} \geq -1 & \text{if } w_{ii} < 0 \\ \sum_{j=1, j \neq i}^n |w_{ij}| + w_{ii} \leq 1 & \text{if } w_{ii} \geq 0 \end{cases} \quad (3.18)$$

According to theorem Gershgorin for discrete models, we can include stability in equation 3.15. Indeed, all Gershgorin discs are inside the unit circle and, based on equation 3.18, $W^{(p \rightarrow p+1)}$ is stable if:

$$\sum_{j=1, j \neq i}^n |w_{ij}| \leq 1, \quad i = 1, \dots, n \quad (3.19)$$

Therefore, we can improve our formulation as:

$$\begin{cases} \text{minimize} & \mu \sum_{i,j=1}^n b_{ij} |w_{ij}| + (1 - \mu) \epsilon^{(p)} \\ \text{subject to} & \|G^{(p+1)} - W^{(p \rightarrow p+1)} G^{(p)}\|_2 < \epsilon^{(p)} \\ & \epsilon^{(p)} > 0 \\ & W^{(p \rightarrow p+1)} \in H \\ & \sum_{j=1}^n |w_{ij}| \leq 1, \quad i = 1, \dots, n \end{cases} \quad (3.20)$$

Proposed Algorithm Finally, the sparsest stable $W^{(p \rightarrow p+1)}$ will be determined for each sub-network that results in sufficiently small $\epsilon^{(p)}$, while incorporating prior knowledge (Algorithm

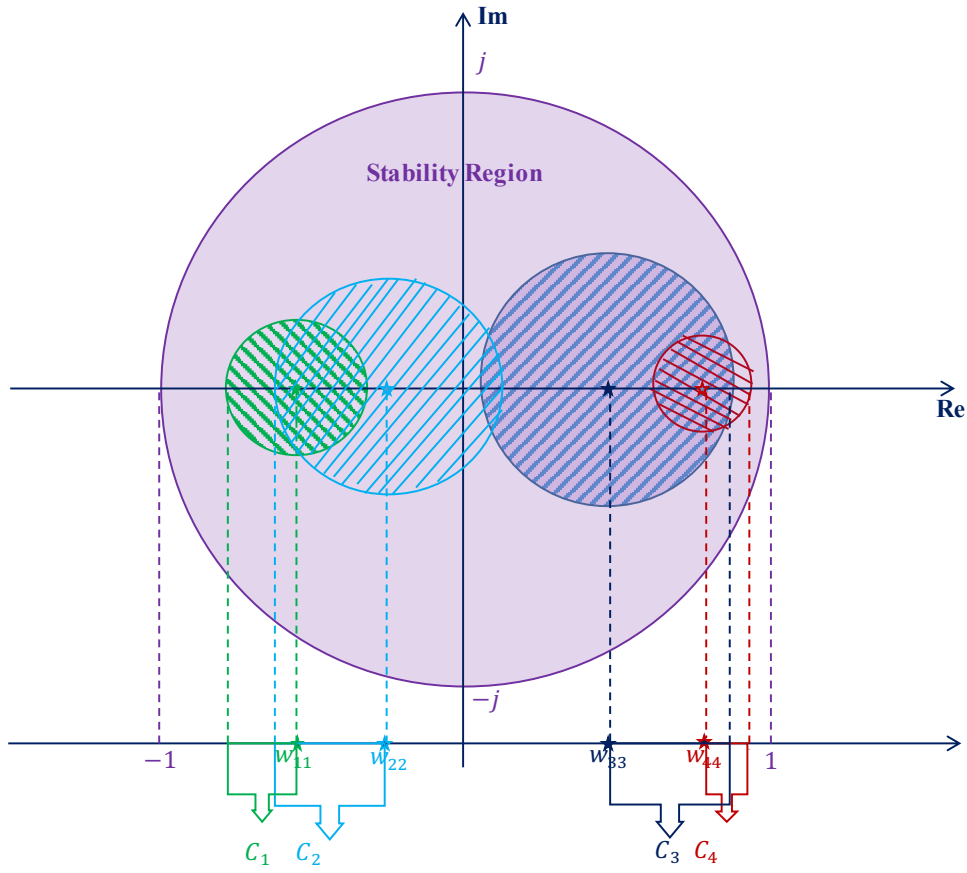


Figure 3.3: Regions for D_j , $j = 1, \dots, 4$ for the eigenvalues of an assumed stable matrix $W \in R^{4 \times 4}$.

Table 3.1: GRN Reconstruction Algorithm.

Algorithm 1	GRN Reconstruction Algorithm
	Control parameter $0 < \mu < 1$
	Initialize $b'_{ij}, s = 1, \gamma = 10^{-3}$
	$w'_{ij} s = 0, \quad i, j = 1, 2, \dots, n$
	While $\ W_{new}^{(p \rightarrow p+1)} - W_{old}^{(p \rightarrow p+1)}\ _2 \geq \gamma$
	<i>Solve the convex program 3.20</i>
	<i>Update $b'_{ij} s$ by 3.12</i>
	End

3.1).

In Algorithm 3.1: μ allows us to control the trade-off between model accuracy and model sparsity and it is found out that $\mu = 0.13$ provides the best result (most similarity between the extracted interactions and real ones) among $\mu = [0, 1]$. In the first stage, we initialize b_{ij} 's = 1 and w_{ij} 's = 0, $i, j = 1, 2, \dots, n$; and also consider $\gamma = 10^{-3}$. In second stage, find $W_{new}^{(p \rightarrow p+1)}$ by solving the convex optimization problem 3.20. In third stage, using obtained $W_{new}^{(p \rightarrow p+1)}$ from the second stage, we update b_{ij} by 3.12. In the final stage, if $\|W_{new}^{(p \rightarrow p+1)} - W_{old}^{(p \rightarrow p+1)}\|_2 \geq \gamma$, then we update $W_{old}^{(p \rightarrow p+1)} = W_{new}^{(p \rightarrow p+1)}$ and go to second stage, otherwise stop the iteration and return the current solution as the optimal value $W_{optimal}^{(p \rightarrow p+1)} = W_{new}^{(p \rightarrow p+1)}$. As mentioned, our problem is divided into $m-1$ sub-problems and $(m-1)$ sub-networks are identified using $(m-1)W_{optimal}^{(p \rightarrow p+1)}$ values inferred from Algorithm 3.1 as shown in Figure 3.4.

3.1.2 Results and Discussion

In this section, we present the results of the proposed algorithm for experimental data. Indeed, we will introduce yeast (*Saccharomyces cerevisiae*) cell cycle microarray time series data sets presented in [Gadkar, 2005, Spellman, 1998]. This data has been extensively exploited for both practical and academic applications.

Researchers frequently use these data sets to demonstrate and validate statistical and clus-

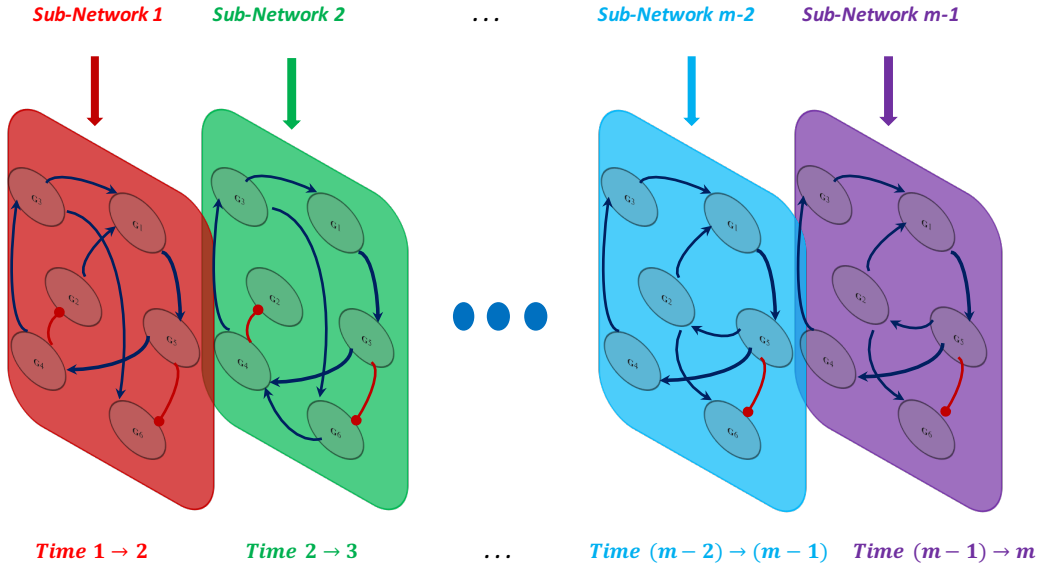


Figure 3.4: An example of inferred sub-networks in $m - 1$ ranges of time. Symbols " \rightarrow " and " \circ " illustrate activator and repressor interactions, respectively.

tering analysis (e.g., [Zeng, 2008, Shotwell and Slate, 2011, Grün and Slate, 2012, Tian, 2011]), and reverse engineering methods [Böck, 2012].

In the following sections, we present a sample of the identified interaction network, revealing the interactions in yeast cell cycle. In addition, the comparison of the results with experimentally evaluated network and previous works are presented.

Data: yeast cell cycle dataset

In order to evaluate our algorithm, we obtained optimal s for all sub-problems based on yeast cell cycle microarray time series data sets. We have focused on twelve yeast genes playing key roles in the control of cell cycle from the Yeast Proteome Database [Costanzo, 2000]. Our algorithm is used as an identification method to find all possible genetic interaction networks that fit the data for the set of twelve genes.

As described by Cho et al [Cho, 1998], gene expression profiles for the yeast cell cycle have

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

been studied through four microarray time series data sets: Alpha, Cdc15, Cdc28 and Elu with 18, 24, 17, and 14 time points respectively. We have used Alpha, Cdc15 data sets, in which still some samples were missed. The missing values are computed using an estimation method based on the K-Nearest Neighborhood (KNN) algorithm [Troyanskaya, 2001].

S.cerevisiae cell cycle regulatory protein-DNA interactions were also the subject of a recent extensive experimental study [Lee, 2002] for which a great deal of information has been compiled in KEGG pathway database [Kanehisa, 2010, Kanehisa, 2012].

Genetic interaction network reconstruction for yeast cell cycle data

To test the capability of our proposed algorithm, we used the algorithm 3.1 for 42 times (Alpha & Cdc15 datasets) to extract gene regulatory network (GRN) structures from inferred 's. By a part of yeast cell cycle regulatory network extracted from KEGG database [Kanehisa, 2010, Kanehisa, 2012], the results of GRN reconstruction were evaluated. Our algorithm was implemented in MATLAB using the CVX toolbox for convex optimization problems [Boyd, 2013] and run on an Intel Core i7, 3.40 GHz processor with 8 GB RAM. For problem of size $n = 12$ and 15% knowledge sign, algorithm 3.1 took approximately 25 seconds for each sub-problem.

To study the network's dynamics, we reconstructed the 41 networks using our proposed method. Figure 3.5 illustrates the extracted genetic network of activator and repressor interactions between time point 28 and 29 as a sample of 41 identified sub-networks. Our final network is extracted based on 50% of confirmed interactions present in the inferred networks over time in different phases (e.g. S phase); in other word, one interaction is considered in final network when 50% of reconstructed networks into the specific phase confirm its presence.

Discussion

We compared the reconstructed genetic interaction networks based on six algorithms from the same dataset. The results show that the identified networks from our proposed algorithm have the best performance among other five methods.

Table 3.2 show that the final extracted network from our algorithm demonstrates a suitable

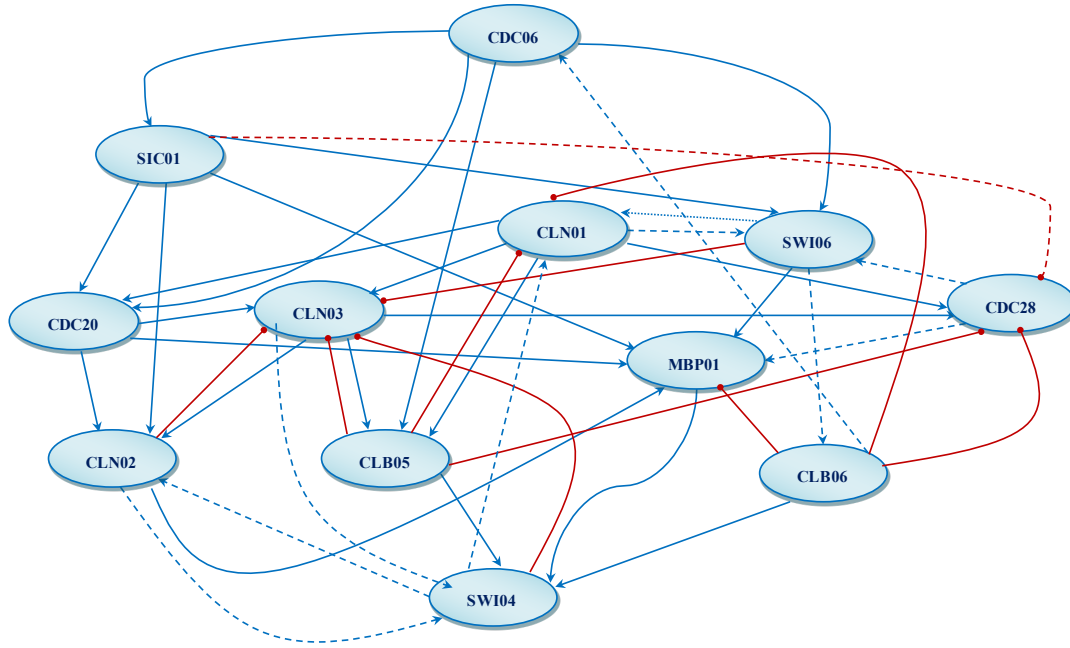


Figure 3.5: As a sample, in this figure, inferred genetic interaction network from sub-network between time point 28 and 29 for twelve yeast cell cycle regulatory genes is shown. Each node represents a gene and the presence of an edge between the two nodes represents the existence of interaction between the two genes. Symbols "→" and "—o", shown by blue and red edges, illustrate activator and repressor interactions, respectively. Dashed edges represent interactions that have been verified. In contrast, dotted edges are incorrect extracted interactions.

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

Table 3.2: Comparison of our algorithm performance by other methods in literature in detecting interactions among experimentally known gene interactions.

No	Interactions	DBN [Kim, 2004]	VBEM [Tienda-Luna, 2009]	Time delay-ARACNE [Zoppoli, 2010]	PF subjected to LASSO [Noor, 2012]	Propose Algorithm
1	<i>CLN03</i> → <i>SWI04</i>	Consistent	Consistent	Consistent	Not Found	Consistent
2	<i>CLN03</i> → <i>SWI06</i>	Not Found	Not Found	Consistent	Inconsistent	Consistent
3	<i>CLN03</i> → <i>MBP01</i>	Not Found	Not Found	Consistent	Not Found	Not Found
4	<i>CDC28</i> → <i>MBP01</i>	Not Found	Consistent	Consistent	Not Found	Consistent
5	<i>CDC28</i> → <i>SWI04</i>	Not Found	Not Found	Not Found	Not Found	Consistent
6	<i>CDC28</i> → <i>SWI06</i>	Not Found	Not Found	Consistent	Inconsistent	Consistent
7	<i>CDC28</i> → <i>CDC06</i>	Not Found	Consistent	Not Found	Not Found	Not Found
8	<i>CDC28</i> → <i>SIC01</i>	Not Found	Not Found	Not Found	Not Found	Not Found
9	<i>SWI04</i> → <i>CLN01</i>	Consistent	Consistent	Consistent	Not Found	Consistent
10	<i>SWI04</i> → <i>CLN02</i>	Consistent	Not Found	Not Found	Not Found	Consistent
11	<i>SWI04</i> → <i>CDC28</i>	Not Found	Not Found	Not Found	Not Found	Not Found
12	<i>SWI06</i> → <i>CLB05</i>	Not Found	Not Found	Not Found	Consistent	Not Found
13	<i>SWI06</i> → <i>CLB06</i>	Not Found	Consistent	Consistent	Not Found	Consistent
14	<i>SWI06</i> → <i>CLN1</i>	Inconsistent	Inconsistent	Consistent	Consistent	Inconsistent
15	<i>SWI06</i> → <i>CLN2</i>	Not Found	Not Found	Not Found	Consistent	Consistent
16	<i>SWI06</i> → <i>CDC28</i>	Not Found	Not Found	Not Found	Not Found	Not Found
17	<i>MBP01</i> → <i>CLB05</i>	Not Found	Not Found	Not Found	Not Found	Consistent
18	<i>MBP01</i> → <i>CLB06</i>	Not Found	Not Found	Not Found	Not Found	Not Found
19	<i>MBP01</i> → <i>CDC28</i>	Not Found	Not Found	Not Found	Not Found	Not Found
20	<i>CLN01</i> → <i>SWI06</i>	Not Found	Not Found	Not Found	Consistent	Consistent
21	<i>CLN01</i> → <i>SWI04</i>	Not Found	Not Found	Not Found	Not Found	Not Found
22	<i>CLN01</i> → <i>SIC01</i>	Consistent	Inconsistent	Inconsistent	Consistent	Inconsistent
23	<i>CLN02</i> → <i>SIC01</i>	Not Found	Not Found	Inconsistent	Not Found	Not Found
24	<i>CLN02</i> → <i>SWI04</i>	Not Found	Not Found	Consistent	Not Found	Consistent
25	<i>CLN02</i> → <i>SWI06</i>	Not Found	Not Found	Not Found	Consistent	Not Found
26	<i>SIC01</i> → <i>CDC28</i>	Not Found	Not Found	Not Found	Inconsistent	Consistent
27	<i>SIC01</i> → <i>CLB05</i>	Not Found	Not Found	Not Found	Not Found	Not Found
28	<i>SIC01</i> → <i>CLB06</i>	Not Found	Not Found	Not Found	Not Found	Not Found
29	<i>CDC20</i> → <i>CLB05</i>	Not Found	Not Found	Not Found	Not Found	Consistent
30	<i>CDC20</i> → <i>CLB06</i>	Not Found	Not Found	Not Found	Not Found	Not Found
31	<i>CDC20</i> → <i>CDC28</i>	Not Found	Not Found	Not Found	Consistent	Not Found
32	<i>CLB05</i> → <i>CDC06</i>	Not Found	Not Found	Not Found	Not Found	Not Found
33	<i>CLB06</i> → <i>CDC06</i>	Not Found	Inconsistent	Consistent	Not Found	Consistent

*Symbols "→" and "→o" illustrate activator and repressor interactions, respectively.

match with the KEGG pathway. Indeed, we observe that our algorithm is capable to extract 15 true connections out of 33 available experimentally illustrated connections, while 4, 5, 7, 10, and 13 true connections are captured by DBN [Kim, 2004, Dondelinger, 2012], VBEM [Tienda-Luna, 2009], PF subjected to LASSO [Noor, 2012], Time delay-ARACNE [Zoppoli, 2010] and RBNFN [Manshaei, 2012] respectively.

Some criteria are useful to evaluate the goodness of fit of the inferred network [Emmert-Streib, 2012]: the proportion of recovered true edges in the target network is called Sensitivity, while Precision corresponds to the expected success rate in the experimental validation of the predicted interactions.

We can compute sensitivity and precision from following equations:

$$Sensitivity = \frac{TP}{TP + FN} \quad (3.21)$$

$$Precision = \frac{TP}{TP + FP} \quad (3.22)$$

where true positive (TP) is the inferred number of edges identified correctly, false negative (FN) is the number of edges that were not identified, and false positives (FP) is the number of edges identified incorrectly.

Also, we calculated the *F*-score by using equation 3.23 to further quantify the performance of the algorithms [Emmert-Streib, 2012].

$$F - Score = \frac{1}{\alpha \left(\frac{1}{Precision} \right) + (1 - \alpha) \left(\frac{1}{Sensitivity} \right)} \quad (3.23)$$

where α is a weighting factor and here we consider $\alpha = 0.5$, otherwise known as the harmonic mean of precision and sensitivity since importance of precision and sensitivity is even. Consequently, the goodness of fit of the results based on our proposed algorithm and the other structure learning approaches in predicting the connectivity network of the KEGG pathway were compared using estimates of above criteria.

The results, presented in Table 3.3, show that evaluation criteria, precision, sensitivity and *F*-score are distinctively higher for our approach compared to previous methods indicating the efficiency of our approach.

Considering the results obtained from proposed method, it can be concluded that there is considerable agreement between the findings of our algorithm and experimental results reported in the literature. If gene expression data is not analyzed properly, it can be difficult to

Table 3.3: Comparison of the proposed algorithm with other methods using statistical criteria.

	TP	FP	FN	Sensitivity	Precision	F-Score
DBN [Kim, 2004]	4	1	29	12.1%	80%	21%
VBEM [Tienda-Luna, 2009]	5	3	28	15.2%	62.5%	23.5%
Time delay-ARACNE [Zoppoli, 2010]	10	2	23	30.3%	83.3%	44.4%
PF subjected to LASSO [Noor, 2012]	7	3	26	21.2%	70%	32.5%
Proposed Algorithm	15	2	18	45.4%	88.2%	59.9%

interpret and can easily be misconstrued. In comparison to other methods, our computational algorithm analyses the data in a manner that is efficient, unbiased, and fast.

3.2 Structural Norm Minimization based on Neighbourhoods

Extracting existent structures in complex data has attracted significant attention in network science, signal processing, machine learning and system identification. Various norm minimization models have been used to extract available structures in the data in recent years: l_0 and l_1 for sparse signal recovery, nuclear norm l_* for low rank matrix recovery, and mixed norms such as $l_{1,inf}$ and $l_{1,2}$ to recover group sparse structures. A combination of these norms could be used to promote more structure. For instance l_1 and l_* are used in robust PCA to recover a matrix $M = L + S$ as a summation of a low rank L matrix and a sparse S matrix [Candes *et al*, 2011]. Also, in a more recent work a simultaneous minimization of l_1 and l_* was used to recover the matrices that are sparse and low-rank at the same time, $\argmin_S \|M - S\|_F + \gamma_1 \|S\|_* + \gamma_2 \|S\|_1$ with $\|\cdot\|_F$ denoting the Frobenius norm [Richard *et al*, 2012]. Such structures arise in the context of block diagonal covariance matrices, unsupervised learning methods, and social or gene-gene interaction networks where underlying graphs have block-diagonal adjacency matrices. A lower bound is provided for simultaneous norm minimization in [Oymak *et al*, 2014], which implies that the minimization of these norms does not improve the recovery compared to minimizing each norm separately. However, based on the results reported by [Richard *et al*, 2012] and our investigations, we believe that the provided lower bound is not very tight, since the

improvements have been achieved by combining l_1 and l_* norms. In this abstract, we focus on minimizing a combination of a new neighbourhood norm l_N and a weighted nuclear norm to recover block-diagonal matrices from noisy measurements.

3.2.1 Problem Formulation

In our model, we define $l_N(S)$ as $\|S\|_N = \|S\|_1 - \sum_{i=1}^N (\alpha_i |s_i| + \sum_{j|(i,j) \in E} \beta_{i,j} |s_j|)$ where E is a set of neighbouring indexes, and α_i and $\beta_{i,j}$ are positive scalars. Minimizing the l_N recovers sparse matrices with non-zero elements being clustered. The clustering implies that if a point is placed in a cluster, its neighbouring points would also be placed in that cluster with high probability. Unlike the methods based on $l_{1,2}$ minimization there is no need to know the clusters/groups indexes and the number of clusters. Using this definition, we solve the following model to recover simultaneous low rank and block-diagonal matrices from noisy measurements:

$$\hat{S} = \underset{S}{\operatorname{argmin}} \left(\frac{1}{2} \|M - S\|_F^2 + \lambda (\gamma \|S\|_N + (1 - \gamma) \|S\|_{w,*}) \right) \quad (3.24)$$

where $\|S\|_{w,*} = \sum_i w_i |\sigma_i(s)|$ is the weighted nuclear norm, $w = [w_1, \dots, w_n]$, $w_i \geq 0$, $\sigma_i(s)$ is the i^{th} singular value of matrix S , and λ and γ are positive scalars.

Solving The Proposed Model

An alternating direction based method is used to solve 3.24, by splitting it into two smaller problems: a weighted nuclear norm minimization solved by an iterative singular value thresholding [Gu *et al*, 2014], and a sparsity inducing problem $\underset{S}{\operatorname{argmin}} (\frac{1}{2} \|M - S\|_F^2 + \lambda (\gamma \|S\|_N))$ solved by an orthogonal matching pursuit based algorithm [Huang *et al*, 2009].

3.2.2 Result

The proposed method is used to recover a block diagonal covariance matrix. Four clusters are generated by multivariate Gaussian distributions with means $\mu_1 = (-3, -3)$, $\mu_2 = (-3, 3)$, $\mu_3 = (3, -3)$, $\mu_4 = (3, 3)$ and covariance matrices of $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = [1, 0; 0, 1]$. An iid zero mean white Gaussian noise with standard deviation of 0.5 is added to the data. In this simulation, we consider $\lambda = 0.1$, $\gamma = 0.04$, $\gamma_1 = \lambda(1 - \gamma)$, $\gamma_2 = \lambda\gamma$, $\alpha = 1$, and $\beta = 0.7$. Also, E is a set of 4 neighbouring indexes.

The proposed method is applied on the adjacency matrix defined as $M(i, j) = \exp(-|s_i - s_j|/2)$, and is compared with the method proposed in [Richard *et al*, 2012]. As can be seen in Fig.3.6(F), compared to [Richard *et al*, 2012] (Fig.3.6(E)) our proposed method has recovered the clusters correctly, although the clusters in the noisy adjacency shown in Fig.3.6(C) are not easily distinguishable.

Furthermore, there is a relation between the number of connected components (islands) and Laplacian spectrum in spectral graph theory [Mohar, 1997]. Indeed, the multiplicity of the eigenvalue 0 in the Laplacian spectrum equals the number of connected components in the graph. According to this fact, let S be an adjacency matrix with distance between each pair of samples. Then the multiplicity of the eigenvalue 0 of its laplacian spectrum equals the number of clusters. It is shown in Fig. 3.7 that the obtained result from our proposed method would be able to confirm 4 clusters whereas the Laplacian spectrum of [Richard *et al*, 2012] is able to find 1 cluster.

3.3 Tangible Biological Networks

In the first two sections of this chapter, we introduced the idea of how evidences influence the way we might apply modifying model. We need an intuitive way to gather and visualize evidence from partial information. This section is sort of talking about sets of tools to deal

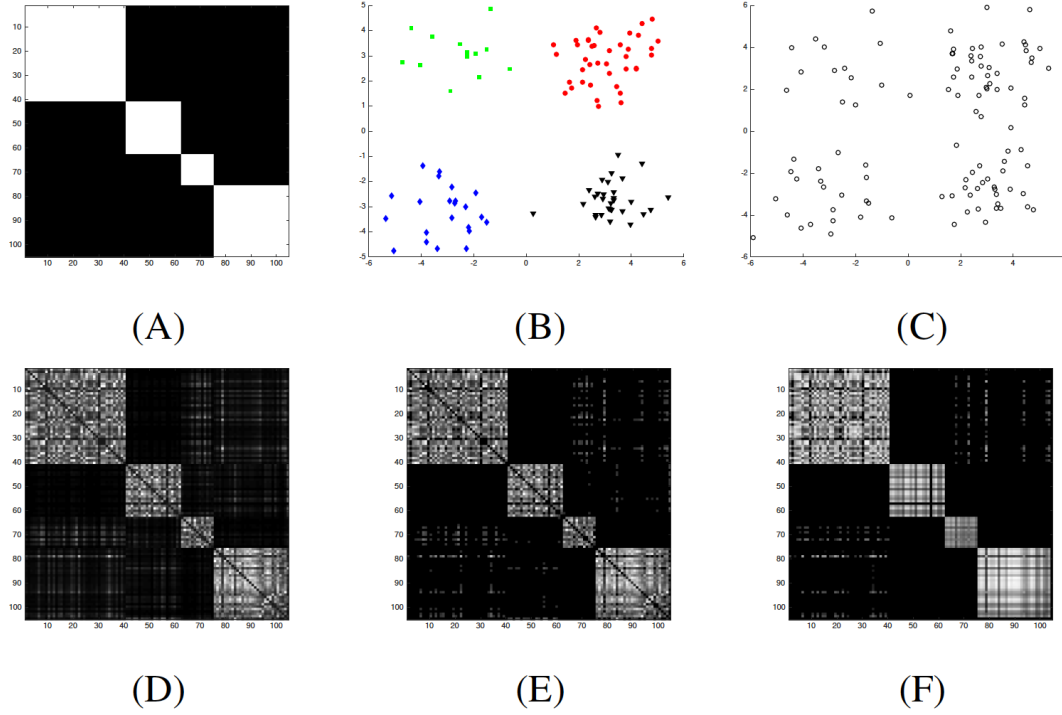


Figure 3.6: (A) Blocks of the matrix, (B) The clusters shown on a graph, (C) Noisy clusters, (D) Noisy adjacency matrix, (E) Recovered by method in [Richard *et al*, 2012], and (F) Recovered by our proposed method.

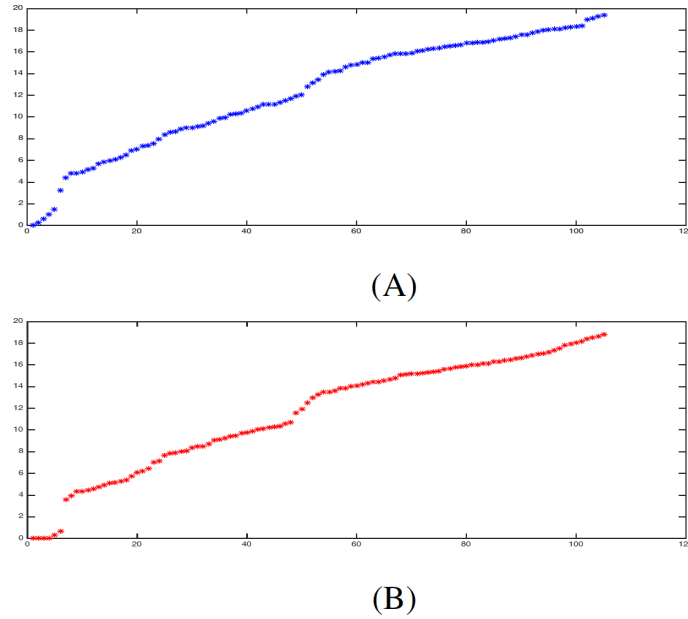


Figure 3.7: Laplacian spectrum obtained (A) by method in [Richard *et al*, 2012], (B) by our proposed method.

with evidence gathering, visualizing of the current state of knowledge plus the results of model that we might be plotting and then taking a step further to have tools that support input for refining model.

In other words, we want to build tools for evidence gathering, visualization of something that inherently special and providing kind of feedback visually on the performance of these networks versus other consensus. We want to provide a package and allow a person to navigate genes and interactions in these organisms, select them, aggregate them, apply several filters on the same network, also let them to tag genes and interactions with tags, and these tags appear as colored labels on nodes and edges of the network or nodes and edges with different sizes.

This particular visualization uses the network illustration to emphasize how strongly confirmation of the gene or protein connection is. Indeed, there are different ways that we can represent and they provide different levels of evidence or somehow shape the prior information and they are really important that how we should configure the model that we are attempting to apply, but also we could use this kind of visualization to illustrate how well the inference is

confirmed by the consensus.

We started to build a tool more on evidence gathering side and we did modelling in offline sense but we realized their importance because we got better results. After reviewing the evidence and realizing that the state of art models have problem with some of their assumptions, what we wanted to do was to build tools that we could gather evidence, feed that back into the design models and evaluate.

Gene network exploration is an integral part of uncovering the secrets and structure of genetic pathways. Biologists explore gene networks to find the genetic differences and structural evolution among divergent organisms, and to identify the structural limitations on the possible paths of evolution. Because gene interaction structure is sparse, the structural features of a gene interaction network are often generalizable across different organisms. As a result, gene exploration often points towards "probable" interactions, which results in new discoveries.

Currently there is no efficient gene exploration interface available. Most present systems either fail to work with big data sets or fail to provide the support for custom query construction. Most of these solutions are also system-dependent (that is, do not work with all operating systems), require the installation of proprietary software, or demand high processing power. Furthermore, almost all present solutions are desktop-based, which hinders collaborative exploration and limits the users to smaller screens.

Here, we present the final design of a novel tabletop/wall-active tangible-based framework that supports cross-platform and collaborative gene exploration using Web interfaces. The intention was not only to design and develop an easy to use gene exploration interface but also to provide the biology community with a flexible framework that can be used to add new visualizations and features. The intention was also to encourage collaborative exploration and support non-experts who are coming into an expert domain.

The contribution of our work is threefold. First, we discuss the design and development of a new approach for visualizing and manipulating gene networks that can aid in conceptual understanding in both learning and discovery contexts. Second, we demonstrate how smart-watches and tablets can be used as active tangibles on tabletops or touch wall. We allow users to construct custom queries using active tangibles by expanding their interaction vocabulary.

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

Finally, we introduce new exploration methods for surveying and comparing structural features of various biological networks. The strategy here is to lead with these features and why they are important in analysis of biological networks in general. We have developed 30 structural features based on nodes, edges, and networks as interactive filters (such as i) eigen space and laplacian eigen space; ii) multiplicity for each eigenvalue and laplacian eigenvalue; iii) network energy and the effect of gene and interaction removals on energy of gene regulatory networks; iv) eigen centrality; v) disassortativity; vi) clustering coefficient; vii) measuring the confidence level of protein and genetic interactions, based on the total number of articles confirming the interactions; etc).

3.3.1 Exploration of Gene Interaction

Gene interactions in different organisms provide an insight into functional and structural connections among genes and their produced proteins [Lehner, 2011], [Van Steen, 2011]. The expression of a gene can be either self-regulating or influenced by a collaboration between itself and other expressing genes.

Recently, uncovering the networks of gene interactions (i.e. regulatory relationships between genes) has become a systematic and large-scale phenomenon [Laufer, 2014]. Yet, the task of analyzing gene interactions is challenging due to the large number of genes, which increases the search space for possible interactions. In addition, the complexity and scale of such networks, either known or inferred, renders the task of surveying functional attributes and understanding biological processes within model organisms a challenge in and of itself.

The area of network science has addressed this issue by defining gene networks as a set of nodes (e.g. genes or their produced proteins) and edges forming linkages between them (e.g. interactions). This simplification makes it easier for practitioners to extract local and global structural features and provides a better understanding of the processes driving the growth and response of living organisms [Barabási, 2015].

An interaction could be either physical or genetic. A physical interaction refers to the experience where a direct physical association between two gene products (proteins) has been

identified, while a genetic interaction is identified when the effect of mutating one gene is reflected by the perturbation in another gene [Beyer, 2007]. Interactions can also be classified as inhibitory and excitatory. In an excitatory interaction, the expression of one gene increases/decreases the expression of another gene, while in inhibitory interaction, an increase in the expression of a gene decreases the expression of another gene [Patel *et al*, 2014]. Every gene in a regulatory network has one or more activator or inhibitor genes. Thus, to understand the structure of a gene interaction network, practitioners usually explore the network to discover the activators and inhibitors of the genes involved.

The knowledge of gene interactions also plays an important role in uncovering the structure of genetic pathways. It assists researchers and practitioners to better understand the path of evolution, including the genetic differences and structural evolution among divergent organisms [Phillips, 2008]. It also assists them to identify the structural limitations on the possible paths of evolution. Besides, as gene interaction structure is sparse, the number of actual interactions is far less than the maximum number of possible interactions [Leclerc, 2008]. Therefore, the structural features of a gene interaction network are often generalizable across different organisms, which also highlights the importance of exploring gene networks.

3.3.2 Related Works

In this section, we provide a brief survey of the existing work in the area.

Biological Data Visualization

With the increasing importance of and access to large data sets by diverse prospective users, several multi-touch and tangible interfaces have been created to support learning and understanding in data intense areas in informal and formal settings. For example, the G-nome Surfer is a tabletop multi-touch application for collaborative exploration of genomic databases [Shaer *et al*, 2012], while SynFlo exploits tangible interaction in combination with a tabletop to introduce synthetic biology concepts to non-experts [Xu *et al*, 2013]. However, their expressivity is limited, as users cannot define and set query operators directly.

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

Some projects have aimed at the visualization of biological pathways using traditional GUIs. VisANT [Hu, 2005], a Web-based system, provides a framework for visualizing and analyzing different types of networks of biological interactions and associations, while Cytoscape supports the visualization and integration of molecular interaction networks and biological pathways [Shannon, 2003].

NAViGaTOR [Brown, 2009] is a software package for visualizing and analyzing protein-protein interaction networks. It can query in two interaction databases and display networks in both 2D and 3D. Likewise, EINVis supports the visualization and exploration the interaction networks [Wu, 2013], but can only work with smaller networks. GeneMANIA, in contrast, can find genes that are related to a set of input genes using a very large set of functional association data [Warde-Farley, 2010]. None of these packages, however, support collaborative exploration. Schkolne et al. developed an immersive tangible interface for designing DNA components [Schkolne, 2004]. Their system, however, does not support search or comparison. Facet-Streams [Jetter, 2011] allows users to construct expressive queries using passive tokens on a tabletop display surface. While this system is able to simplify the construction of complex queries, it has some limitations. These include physical clutter, separation of query formulation from the browsing of results, and a lack of persistency for query parameters.

Tangible Queries and Active Tokens

Several tangible interface systems have explored the use of active tokens for constructing queries. An early example, Navigational Blocks [Brown, 2009], is a tangible interface that employs electronically augmented blocks to query a database. Each block represents one query parameter and its six faces represent possible values. In this case, the number of blocks and their fixed values limit the search space to a predefined number of queries. Another system, DataTiles [Rekimoto, 2001], uses transparent RFID-tagged tiles on a horizontal display surface to manipulate digital information. Each tile represents specific information or control structures, and the placement of tiles on the surface can trigger computational functions, such as submitting a query or launching an application. While the system provides an expressive

physical language, the interaction is constrained to a horizontal surface on which the tiles must be placed within predefined grid cells. Tangible Query Interfaces (TQI) [Ullmer, 2003] introduced two kinds of active token interfaces for query formulation: parameter wheels for fixed query parameters, and parameter bars that can be dynamically assigned to various parameters. In both cases, the tokens can represent either discrete or continuous values, and are manipulated and interpreted within physical constraints, such as racks. Although, this configuration affords expressive query construction, it limits the portability of the tokens, as well as the possibilities for collaboration.

Unlike most query interfaces, Stackables [Klum, 2012] explores the use of active tangibles that enable vertical stacking. Each token represents a query parameter and multiple tokens can be stacked to express a query consisting of logical AND or NOT. Results are visualized on an adjacent display screen. In this system, the interaction with the tangibles is limited to using its sliders and button. Sifteo Cubes, on the other hand, can detect shaking, flipping, tilting, neighboring, and tapping, but do not provide the support for multi-touch. A more recent project used smartphones as active tangibles on a tabletop display, allowing users to query for the common files on smartphones when they are in close proximity [Mazalek, 2014]. However, similar to Facet-Streams [Jetter, 2011] and the TQI [Ullmer, 2003], the tangibles must be on the table to perform the queries. By contrast, the active tangible utilized in this dissertation can support all the actions stated above, as well as multi-touch, in both on- and off-the-table scenarios. Recent work by [Valdes, 2014] investigated the use of gestural interaction with active tangibles for manipulating large data sets. They studied user expectations of a hybrid tangible and gestural language engaging this space. Based on the results, they provided a vocabulary of user-defined gestures for interaction with active tokens. We used this as a guideline to design the interactions for our system.

3.3.3 Tangible Biological Networks

Tangible Biological Networks allows users to explore gene and protein networks on an interactive tabletop using active tangibles. It supports both on- and off-the-table interactions.

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

On-the-table interactions utilize multi-touch and the position of the active tangibles on the tabletop. Off-the-table interactions utilize multi-touch on the active tangibles.

3.3.4 Basic Interactions and Visualization

Tangible Biological Networks allows practitioners to explore all verified physical and genetic interactions based on an organism or a gene, using a linear navigation approach. Thus, to explore a organism, first, users have to select the organism from a list, and then pick the one they are interested in (see Figure 3.8a). Likewise, to explore a network involving a particular gene, they first have to select the gene from a list, which will display all networks involving that gene (across all organisms), and then select the network they want to explore. Here, we discuss these basic interactions and visualizations.

Step 1: Navigation Initiation

Once users decide whether they want to explore gene and protein interactions based on an organism or a gene, they have to initiate the navigation for that particular parameter using the active tangibles. The start screen of the tangibles requests the users to Swipe to continue and allows them to select from three available parameters: gene, interaction, and network, by performing vertical swipe gestures. A down gesture moves the parameters forward and an up gesture moves them backward. The gene, interaction, and network parameters allow users to initiate the navigation based on a gene, or interaction or organism (see Figure 3.8(b)).

They can then navigate to a particular organism by performing vertical swipes. The list displays all organisms by their scientific names. We considered using symbols instead of names, but decided against it to avoid confusion, as there is no commonly accepted symbols available. Once an organism or gene is selected, users can either continue navigating further on the tangible screen or can place it on the table to visualize the current data. If an organism is selected, and the tangible is placed on the table, the system displays all available organisms on the table in a doughnut chart surrounding the active tangible. See Figure 3.9. The arc length of a slice and its central angle and area are proportional to the density of the networks, which we determine based on the total number of genes and proteins in a network. The system

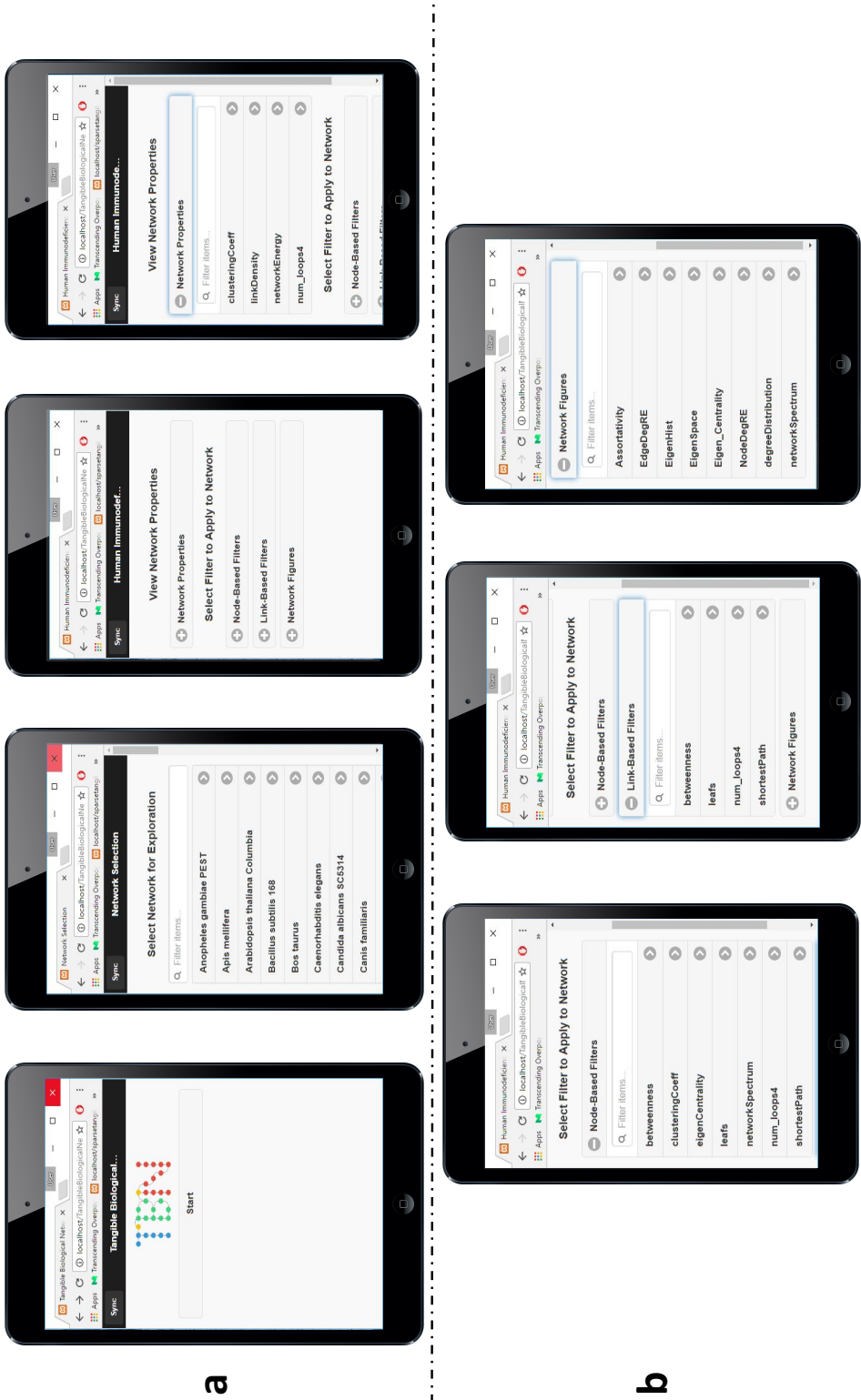


Figure 3.8: Active tangible interactions: a) users start to select the interested organism, swipe to select the type of exploration, gene-based, linked based and network based, b) users select one of them by tapping, which displays a list of developed and related filters.

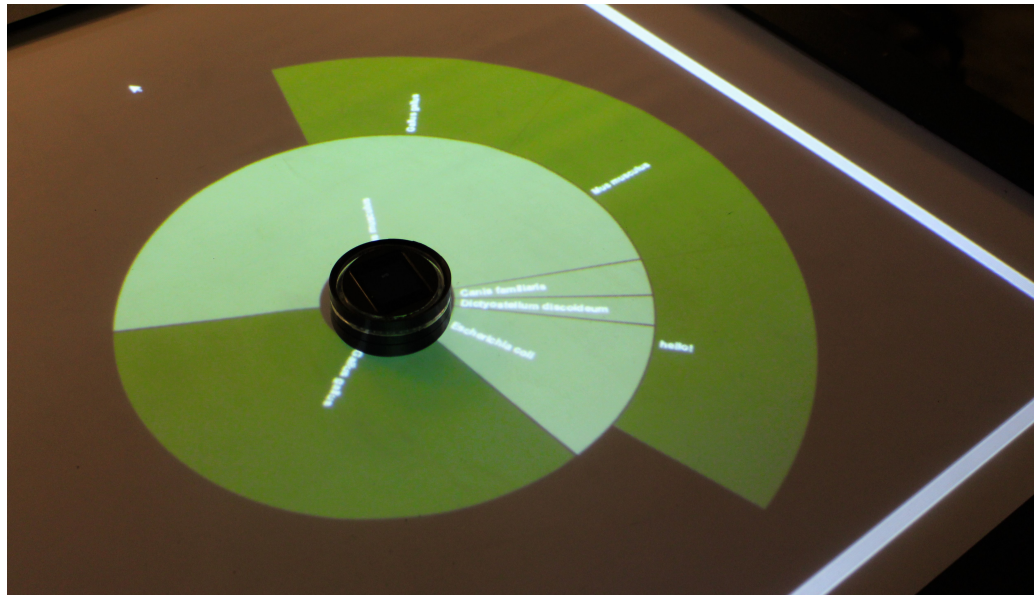


Figure 3.9: When an active tangible is placed on the table after selecting the organism or gene parameter, the system displays all available networks involving that organism or gene, respectively, in a doughnut chart. The outer slice is a magnifier. When users point the tangible towards a zone like a dial, the system magnifies the respective slices for better visualization.

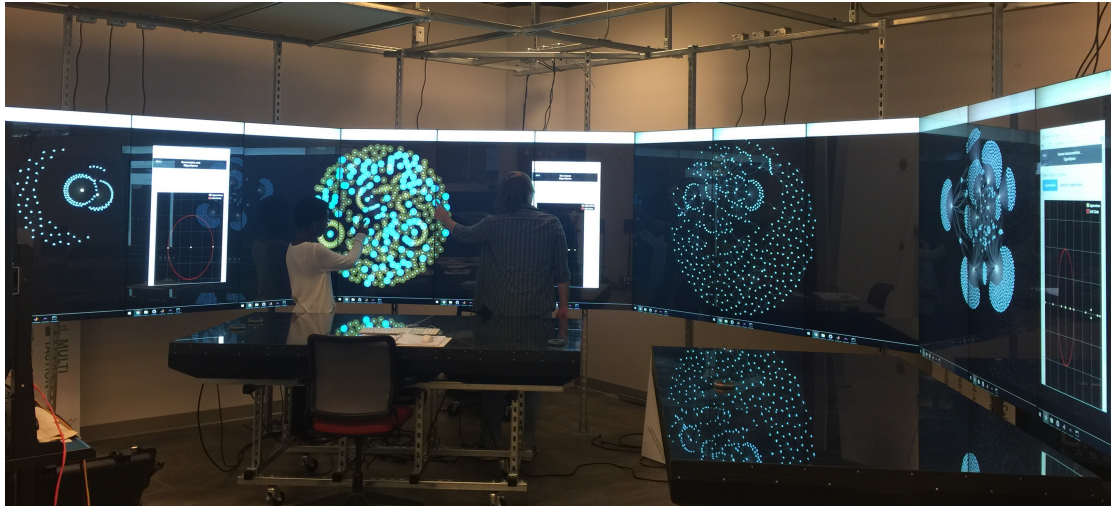


Figure 3.10: A screenshot of the visualization for gene networks within four organisms in force-directed graphs.

assigns each slice with a random shade of green and displays the labels inside the doughnut in white. We picked this style after testing several alternatives, where this yielded the best usability results. The system also displays all networks in force-directed graphs around the chart. These graphs use a physical simulation of charged particles and springs, placing the most common interactions (based on the literature) in closer proximity. See Figure 3.10. They repel each other to avoid overlaps and orient themselves to the outside of clusters. Similar to the chart, the system assigns each network with a random color, but displays the labels in black to distinct the genes from the networks.

Likewise, when an active tangible is placed on the tabletop/wall after selecting a gene, the system displays all available organisms containing that particular gene on the tabletop in a translucent doughnut chart surrounding the tangible. The system also displays all networks in force-directed graphs.

We assign each active tangible with a unique ID, thus the users can select multiple networks using multiple tangibles and compare them on the tabletop/wall (see Figure 3.10). Both the chart and the graphs follow their associated active tangibles as they are moved around the tabletop. They both disappear when the tangibles are lifted from the table and reappear when placed again on the table. To free an active tangible from its assigned organism or gene, users

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

have to shake the device once, which will bring back the initial state to allow the selection of a new parameter.

Step 2: Network Selection

Users can select a network either on- or off-the-table. To select a network autonomously, first, they have to navigate to the intended network on the active tangible by swiping and then have to tap on the screen. If a mistake was made, users can either go one step backward by performing a left swipe or reinitiate navigation by shaking the active tangible. Making selections off-the-table allows users to quickly bind several active tangibles to different parameters. The tangibles can be placed on the tabletop/wall at any time to visualize the associated data.

On-the-table, users can use the active tangible as a dial to focus on a particular zone of the chart, representing a group of networks or organisms. When the active tangible points towards a zone, the system magnifies the respective slices for better visualization (Figure 3.9). This is because, complex organisms often contain hundreds of subnetworks; also, very common genes are often prevalent in tens of organisms, which makes it difficult to read the labels, as they become too small. When magnified, the users can select the network of interest either by tapping on the corresponding slice or on the active tangible. Once selected, the system removes all networks and displays, leaving only the one picked by the users or the one present in the selected organism. The system also updates the doughnut chart accordingly. Similar to the previous step, users can correct their mistakes either by swiping on the tangible or by restarting the session by shaking it.

3.3.5 Query Construction

The Tangible Biological Networks is intended to allow expressive query construction using active tangibles. To construct a query, users have to select different parameters on different active tangibles, where each tangible represents a query parameter. Most techniques also require the tangibles to be on the tabletop to construct queries. As a result, all of them are either inflexible in terms of usability, or fail to support autonomously query construction. Tangible Biological Networks address these issues by detecting multi-touch gestures.

EigenLand Exploration

The survey of the relationships between eigenvalues and the structure of networks is main objective of spectral graph theory. By exploring these relationships, it could be possible to extract the information about the networks. In spectral graph theory, the focus of all studies are on few lowest and largest eigenvalues for networks with certain features and the remaining range of eigenvalues is almost ignored. This shows that how little we know about the spectrum of networks. Our interactive visualization could play an essential role in exploration of hidden relationships.

Eigenspace and laplacian eigenspace

Let O be $n * n$ gene or protein adjacency network of an organism. An eigenvector of this network is a vector V which $OV = \lambda V$ for real or complex number λ . λ is defined as the eigenvalue of O belonging to eigenvector V . λ is an eigenvalue if and only if the matrix $O - \lambda I$ is singular ($\det(O - \lambda I) = 0$). There are n roots with considering multiplicity for this equation. In other words, this genetic network has n eigenvalues.

If the network O is undirected graph with node set $V(O) = 1, \dots, n$, all eigenvalues ($\lambda_n \leq \dots \leq \lambda_1$) are real.

$$O_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have interaction} \\ 0 & \text{O.W.} \end{cases} \quad (3.25)$$

where $\lambda_1 + \lambda_2 + \dots + \lambda_n = \text{trace}(O) = \sum_i o_{i,i}$. We can see the summation of eigenvalues is zero when we don't have any self-regulation in the network (Figure 3.11). In addition, $\lambda_1^2 + \lambda_2^2 + \dots + \lambda_n^2 = 2e(O)$ where $e(O)$ is the number of network's edges. The summation of the squares of the eigenvalues is the same as the trace of O^2 which its diagonal entries count the number of closed walks of length 2 (which starts and ends at the same gene). Furthermore, the summation of the cubes of eigenvalues ($\lambda_1^3 + \lambda_2^3 + \dots + \lambda_n^3$) is the same as the trace of O^3 which is six times of the number of triangles (the number of closed walks of length 3) of the network.

A unsigned network has a non-negative real eigenvalues which has maximum absolute value among all eigenvalues. The product of all eigenvalues by considering their multiplicities is

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization



Figure 3.11: The eigenvalue space related to HIV organism.

$\prod_{i=1}^n \lambda_i = \det(O)$ and the number of non-zero eigenvalues (including multiplicities) is the rank of O .

$L(O) = D(O) - O$, where $D(O)$ is the diagonal matrix of the degrees of O . The laplacian of the network is defined as the $n * n$ matrix $L(O) = L_{i,j}$ in which:

$$L_{i,j} = \begin{cases} d_i & \text{if } i = j \\ -o_{ij} & \text{if } i \neq j \end{cases} \quad (3.26)$$

d_i denotes the degree of node i . (In the case of weighted graphs, $d_i = \sum_j o_{i,j}$). Let Laplacian eigenvalues be $\mu_n \geq \dots \geq \mu_1$.

We have seen when the matrices O and $L(O)$ are symmetric (undirected network), their eigenvalues are real.

In signal processing, the analogy could be to represent a 'signal' with a truncated set of spectral components (like Fourier and wavelet transforms [Yuan *et al*, 2014]). In the other words, shape implies structure. In this dissertation, we are interested in structural information about the graphs. The network eigenvalues and laplacian eigenvalues can be used to give information

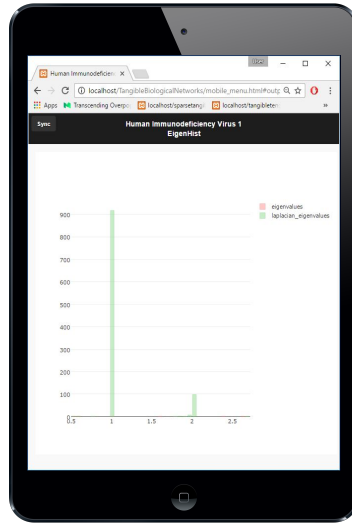


Figure 3.12: The multiplicity of laplacian eigenvalues in HIV network.

about the structure of a graph. We would then expect that if two networks have many eigenvalues in common then they probably share some structure.

Multiplicity for each eigenvalue and laplacian eigenvalue

When O is connected, the largest eigenvalue of O has multiplicity of 1 (see Figure 3.12). It is a kind of average degree. For the laplacian $L(O)$, this corresponds to the smallest eigenvalue. Indeed, if O is connected, then μ_1 as an smallest eigenvalue of $L(O)$ like the largest eigenvalue of network has multiplicity 1.

For a general network, the multiplicity of the laplacian eigenvalue 0 is equal to the number of connected components. Similar statement is not true for the adjacency matrix. This illustrates the phenomenon that the laplacian is often better behaved algebraically then the adjacency matrix. Tangible Biological Networks allow users to compare different filters by visualizing different graphs (see Figure 3.13.).

The normalized laplacian has eigenvalues $(\zeta_n \geq \dots \geq \zeta_1)$ always laying in the range between 0 and 2 [Chung, 1996]. One advantage of this is that it makes it easier to compare the distribution

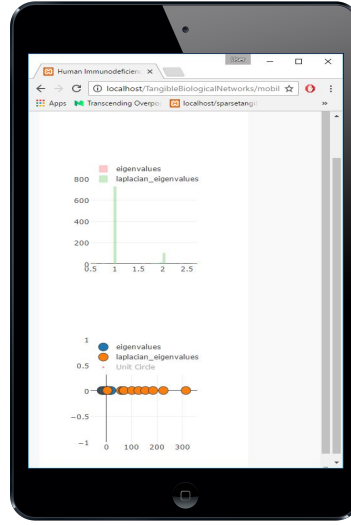


Figure 3.13: The comparison of eigenvalues space and their multiplicities.

of the eigenvalues for two different networks especially if we are dealing with big networks.

$$\max_{i \neq 0} |1 - \zeta_i| \leq 1 \quad (3.27)$$

For a network without isolated nodes, the multiplicity of 0 as an eigenvalue of adjacency matrix is the same as the multiplicity of 1 as an eigenvalue of the normalized laplacian [Butler, 2008]. In addition, the number of negative eigenvalues for the adjacency matrix is the same as the number of eigenvalues of the normalized laplacian greater than 1 and the number of positive eigenvalues for O is the same as the number of eigenvalues of the normalized laplacian less than 1.

Network energy

The network energy is equal to the sum of the absolute values of its eigenvalues. This concept has proposed in the paper [Gutman, 2001]. It has recently become a popular topic of research. Details of the theory of network energy can be found in literatures [Cvetkovic and Gutman, 2009, Cvetkovic and Gutman, 2011, Gutman, 2009, Gutman, 2006, Li 2012]. Since the network energy has showed to be a mathematically interesting concept, and since tens of significant results on it could be obtained [Li 2012], our objective is to use eigenvalues of various networks,

and consider track network energies.

The energy of network O is computed as:

$$Energy(O) = \sum_{i=1}^n |\lambda_i| \quad (3.28)$$

Laplacian version of network energy is defined as:

$$Energy(L(O)) = \sum_{i=1}^n |\mu_i - 2e(O)/n| \quad (3.29)$$

In our context, we have seen $Energy(L(O)) \geq Energy(O)$ for almost all organisms. Du, Li and Li has proved this claim for almost all networks in their paper [Du 2010].

Within our context, the gene regulatory network is a graph where genes equate to nodes and interactions equate to edges. The user could look at the effect on the network energy or laplacian energy when one or several nodes or edges are removed from the network (see Figure 3.14). Indeed, a user could measure the vulnerability of a network to different types of damages (for instance, node or edge deletion would cause a considerable damage to the network). In the other words, the user explores whether the efficiency or robustness of the network is increased by measuring the energy of a network before and after a change (such as deletion) or not; and also an expert person could analyze whether the change/s may become permanent or will be reverted.

Eigenvector centrality

In our context, eigenvector centrality shows influence measurement of a gene/protein in a genetic network. The eigenvector centrality value for each gene/protein is extracted individually. It shows how connected a gene/protein is, how influential a gene/protein is in terms of its neighbourhood, how easily a gene/protein can propagate an information, and how intermediary is a gene/protein as a linkage between genes/proteins in the genetic network [Alvarez-Socorro 2015].

In general, we have seen a gene/protein is important if it is connected by other important



Figure 3.15: The eigenvector centrality spectrum for HIV network.

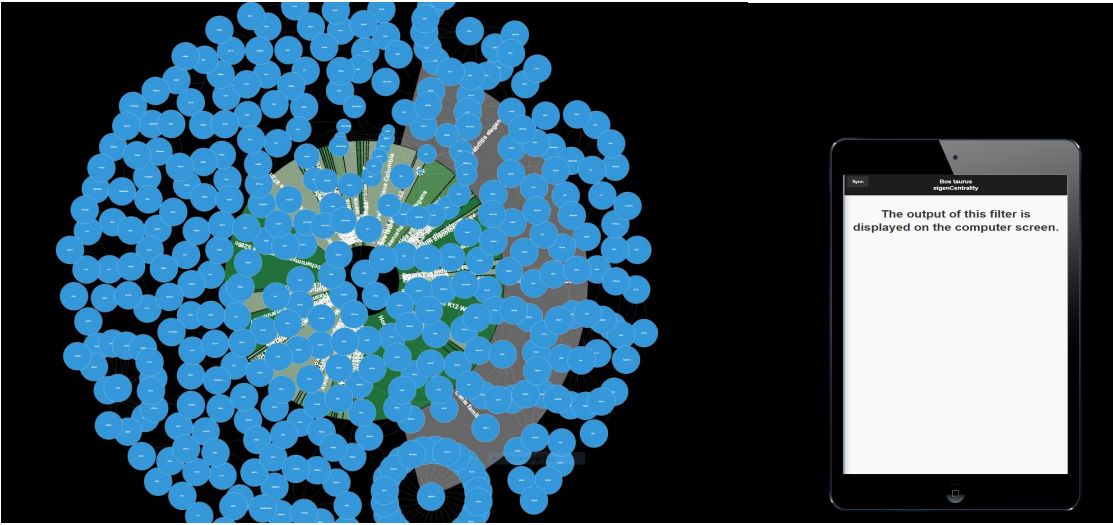


Figure 3.16: Eigenvector centrality representation by changing nodes' size on Bos Taurus network.

Assortativity

Assortativity is a network metric defined by Newman [Newman 2002]. In our context, it represents how genes/proteins are associated with other genes/proteins in a network related to an organism. The connected genes/proteins have the similar or opposite degree. Generally, assortativity ρ is defined as a range from -1 to 1. When high (low) degree genes/proteins are on average connected to low (high) degree ones, the network is named disassortative (*rho* has negative values). In addition, when the genes/proteins with the similar degrees tend to connect together in average, the network is named assortative (*rho* has positive values).

Biologists believe this metric not only provides useful information in terms of network structure, but also allow them to explore the dynamic behaviour and robustness of a network. Assortativity for undirected and weighted networks specially needs to survey further and our interactive system is a suitable platform for exploration and comparison. For example, when we encounter with a directed network, we will have four combinations for calculating assortativity. Indeed, we can consider incoming and outgoing interactions and consequently we can calculate assortativity based on incoming-incoming interactions (in-in), incoming-outgoing interactions (in-out), outgoing-incoming interactions (out-in) and also, outgoing-outgoing interactions (see figure 3.17).

Clustering Coefficient

In network science, clustering coefficient is a degree measurement which genes/proteins in a network tend to cluster together. In undirected networks, this metric is defined for gene/protein i (Figure 3.18) as:

$$C_i = 2\mathbb{N}_i / (\kappa_i(\kappa_i - 1)) \quad (3.31)$$

where \mathbb{N}_i is the number of interactions between all neighbors of gene/protein i and κ_i is the number of neighbors of gene/protein n [Barabasi, 2004]. In contrast, $C_n(directed) = C_n(Undirected)/2$. By taking average from the clustering coefficients of all genes/proteins in

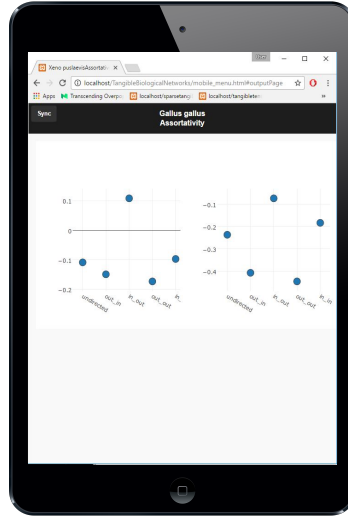


Figure 3.17: Assortativity plot for undirected, in-in, in-out, out-in, and out-out interactions in Gallus Gallus organism.

the network, one can obtain network clustering coefficient (see the mobile menu in Figure 3.18). We have seen that the genes/proteins with less than two neighbors do not have clustering coefficient; indeed the value is 0.

Agreement Score Calculation

Tangible Biological Networks is intended to allow practitioners to filter the data based on agreement score, which is a normalized score of the number of times an interaction was confirmed in the literature and the impact factor of the venues where they were published. This is a way to quantify "consensus". We calculate this for all interactions by multiplying each publication with their respective impact score, adding the results of all multiplications, and then normalizing the scores by dividing them with the highest impact score in the network. Thus, the agreement score for each interaction ranges from 0 to 1. We use the following equation to calculate the actual agreement score for an interaction i .

$$AgreementScore_i = \sum_{p=1}^n (p \times impact_p) \quad (3.32)$$

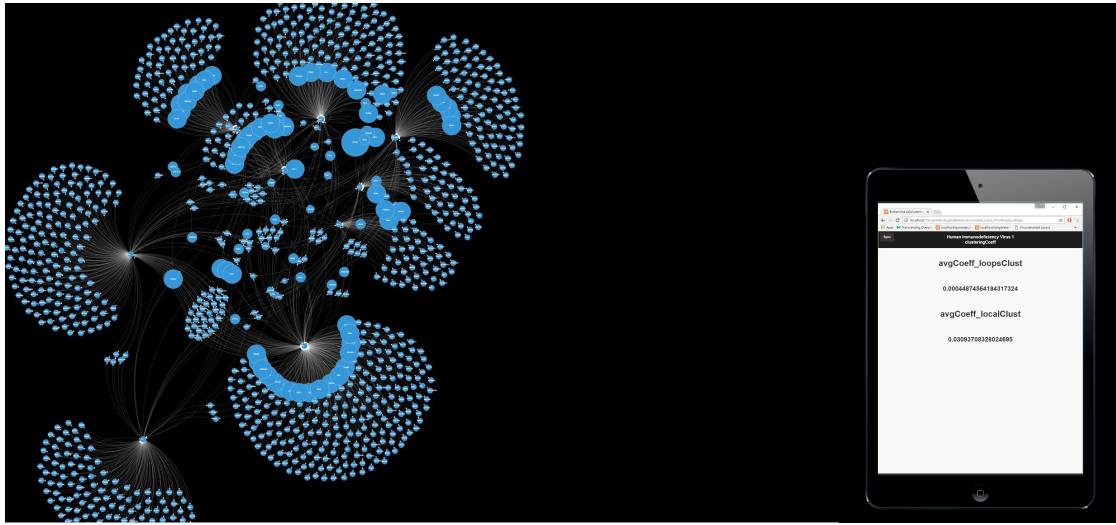


Figure 3.18: Clustering Coefficient related to each gene is shown by changing the size of each node on the tabletop/wall. Also, the network clustering coefficient that is the average of this metric for all genes is illustrated on mobile menu on active tangible.

We then normalize the score by dividing the result with the highest impact score in the network. Here, n signifies the total number of publications reporting an interaction i , p signifies a publication, $impact_p$ signifies the impact factor of the venue where p was published.

We hope that this metric will provide the researchers with a "confidence level" for each interaction. Currently, the only way to acquire a sense of this is by going through different genetic databases that report different interactions. Typically, due to the absence of a metric, validation of an interaction often depends on whether or not the investigator finds enough evidence of that interaction in the literature and subjectively deems the results reported in the publications reliable. The agreement score offers a less tedious, and more objective alternative.

3.3.6 Technical Implementation

Tangible Biological Networks is a mainly web-based platform made using front-end web development tools such as HTML, CSS, and JavaScript. There are back-end tools involved as well, such as Node.js and Python scripts, used to handle, process, and transfer project-related data. There are several reasons for implementing the system using web-based development

tools: (i) Most current generation devices support HTML/JavaScript and are web-page friendly; (ii) JavaScript has a large collection of libraries that allow for easy and productive development; and (iii) It requires the least amount of client-side tools, and can run solely on a web browser. We describe the implementation of the different system components, including the active tangibles, the interactive tabletop, and the communication processes (more details is provided in appendices chapter).

Active Tangibles

Not only we utilize tablets (such as iPad 2) as a active tangibles in our system, but we also use a platform called “Actibles” developed in our lab as the active tangibles in this system (appendix B). Actibles are hand-held devices that are made for the purpose of user-interaction with tabletop-based projects. They employ LG G smartwatches as their core. These come with an IPS LCD Display, 9 Axis sensors, 512 MB RAM, and 4GB internal storage. Each smartwatch is contained within a 3D printed custom case that includes additional hardware and allows for query-based interactions such as stacking, rotating, shaking, and bumping.

The cases are built using a plastic polymer and have a rounded cuboid shape. This shape promotes a greater diversity of interactions compared to a cylindrical shape, as it has clear edges for neighboring, while still maintaining the rounded feel of a dial. The cases also accommodate for an on-board printed circuit board (PCB), which allows for communication between built-in sensors, the smartwatch and the interactive table. The Actibles employ magnetic sensors to detect stacking, neighboring and bumping interactions, as well as an array of RGB LEDs for visual feedback via patterns and animations of colored light.

Aside from the sensors and the LEDs, the Actibles also include a Wi-Fi module, which can send and receive data packets through the UDP protocol. This allows for a more direct and versatile form of communication between the tangibles and the Node.js server. More specifically, the Wi-Fi connection is used to control the LEDs and send any interaction information, such as stacking, to the server. The Node.js server acts as a communication link between the active tangibles and the interactive tabletop display.

Chapter 3. Model Refinement and Inference, and Interactive Biological Networks Visualization

The use of smartwatches within the Actibles has a number of advantages: (i) They enable web-page viewing using an Android Wear compatible web-browser, which allows for cross-device communication; (ii) They come with built-in inertial sensors such as an accelerometer, a compass, and a gyroscope, which is especially useful for the shaking gesture; and (iii) They run on Android Wear OS which is a stable and globally supported mobile operating system. Mobile-friendly JavaScript/HTML5 web-pages were implemented on the smartwatches using jQuery mobile to display the mobile menus and read any events such as swipe, touch, and shake. In order to detect the Actibles on the interactive tabletop, unique fiducial markers were attached to the bottom of each one. These markers are read by the IR sensors within the tabletop display and transmitted to the server using the TUIO protocol.

Interactive Tabletop

The interactive tabletop used within Tangible Biological Networks project is the MultiTaction Cell Ultra-Thin Bezel (MT553UTB) display (appendix A). The MT553UTB is a 55-Inch-High Definition LCD display that features several built-in infrared cameras for touch and marker detection. The display has its own internal computer running the Linux OS. It includes a Hybrid Tracking Engine that allows for several interaction methods such as multi-finger touch, hold, drag, as well as fiducial marker tracking using the MultiTaction Codice markers. The MultiTaction display is connected to a computer running the Windows 10 OS. A web browser is used to open the HTML5/JavaScript web-pages of the Tangible Biological Networks system, which display the tabletop visualization interface. The system predominantly uses the Data Driven Documents (D3.js) library, which stores data using Document Object Model (DOM) elements and JSON files, and visualizes it using Scalable Vector Graphics (SVG).

Communications

To communicate between the devices, a Node.js server is used in conjunction with Socket.IO. Socket.IO is a JavaScript library which allows for real-time bidirectional communication between devices via web-browsers. When a device-based event such as tap or swipe occurs on

a device, Socket.IO broadcasts a message which is received by the central Node.js server. The server handles the message appropriately by communicating with the tabletop. Instead of a TUIO server, a client-based Node.js server is used as it allows us to broadcast custom events. Although the server is located on the same computer that drives the tabletop display, it runs separately from the tabletop interface. More details are given in appendix C.

Database Communication

Our system collects the protein/gene interaction data from the Biological General Repository for Interaction Datasets (BioGRID). This public database archives gene/protein interaction data from model organisms and humans in the CSV (Comma Separated Values) format. Our system pulls all data from the database and converts them into optimized JSON files. The system also calculates and records the total number of networks in organisms and the agreement scores for each interaction. These operations are performed on the server using a .NET application. We decided to use this database instead of alternatives such as IRefWeb and GEMMA, because BioGRID is arguably the most complete database. It holds over 770,000 interactions selected from both high-throughput datasets and individual focused studies, derived from over 54,000 publications in the primary literature [Chatr-Aryamontri, 2015].

The first platform in chapter 3 is a tool particularly around evidence gathering and still on the way and we are working on to close the loop fully by including the modelling in online process, but the other application that we considered in chapter 4, produce much more closed loop system which was tackled problem of cancer survivorship prediction.

4 Effective Prognosis Factors Inference in Prostate Cancer and Tangible Multi- Variate Visualization

In this chapter, a hybrid survival analysis method to estimate hazard risks of metastatic castration-resistant prostate cancer (mCRPC) patients is developed. In order to extract the most significant clinical features from the dataset, data was cleaned via removal of incomplete data, and correction of units across databases. Imputation was used to estimate missing data values and new features were extracted based on clinical relevance. The two phase method was then employed in order to determine the most effective clinical features for hazard risk prediction.

In addition, we present a tabletop and smartphone based framework to design a cross-platform and collaborative system to discover effective covariates in predicting hazard risk of patients in a given disease and classifying the patients using a Web interface. It uses smartphones as active tangibles to allow covariates selection, survey the selected covariates by statistical tools (box plot, Kaplan-Meier plot and re-categorization), and decide about the effectiveness of each covariate based on the extracted results from the univariate and multivariate survival analysis. In addition, the user could see multivariate correlation scatterplot chart of the selected covariates on tabletop and could track the significance and accuracy of bivariate modeling on this chart simultaneously. We also introduce a new presentation for visualizing the effective interactions among clinicopathological covariates of prostate cancer patients using covariate-link networks. Finally, our framework allows clinicians to import the covariates

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

of a new patient and use the effective covariates obtained during manipulating of the system to predict the hazard risk for a new patient.

4.1 A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Prostate cancer is the most frequently diagnosed cancer amongst men in North America, subsequently accounting for the third most common cause of cancer related deaths [IARC 2016]. 15% of men diagnosed with prostate cancer are further categorized as having metastatic castrate resistant prostate cancer (mCRPC) [Wu *et al*, 2014]. mCRPC is defined as prostate cancer progression despite androgen deprivation therapy [IARC 2016]. Extensive research, pertaining to the most effective treatment method for mCRPC patients are currently ongoing [IARC 2016]. Treatment options include but are not limited to combinations of chemotherapy, radiotherapy and supportive care. Unfortunately, the most effective treatment regimens are still unknown [Ghavamian].

In order to better understand treatment outcomes, survival analysis results are often used. Depending on the number of clinical predictors used survival analysis studies may be categorized into univariate, and multivariate studies. Majority of current studies fall into the univariate category, paying particular attention to specific treatment outcomes, without survival predictions. For example, the study by [Shao *et al*, 2014] is a univariate survival analysis study, which uses the SEER medicare database to determine for prostate cancer specific survival (PCSS) after metastasis [Shao *et al*, 2014]. The study aimed to compare the PCSS of patients after Radical Prostatectomy (RP) or Radiation Therapy (RT), using survival times as an indicator of success [Shao *et al*, 2014]. In [Joniau *et al*, 2015] study, COX multivariate and univariate regression models were used in order to determine the prostate cancer specific survival (PCSS)[Joniau *et al*, 2015]. The study [Joniau *et al*, 2015] was conducted using 1632 patients classified as high risk, whom had undergone radical prostatectomy followed over a period of 12 years. Another study is a multivariate one that used the Statistical report Epidemiology and End Results (SEER) web based tool to provide a nomograph of the patients

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

risks [Feuer *et al*, 2014]. The SEER software also provides the capability of providing crude risk, which includes survival in the case of competing risk factors. [Feuer *et al*, 2014] uses the factors of age, race, tumour characteristics and comorbidity profiles (disease status, pathological state, gleason score, medical history). The study population was provided by the united states Medicare data profiles [Feuer *et al*, 2014].

A number of recent studies consider the role of specific clinical predictors in estimating the overall survival of prostate cancer patients [Nuhn *et al*, 2014, Chen, 2014, Koo *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Zhao *et al*, 2015, Koo *et al*, 2015, Gravis *et al*, 2015, Chi *et al*, 2015, Uemura *et al*, 2016, Yamashita *et al*, 2016]. A study in this particular field was conducted by [Halabi *et al*, 2014] using eight clinical predictors to determine the overall survival of patients. Survival prediction algorithms, directly affect treatment decisions and are therefore of great importance in the clinical setting. The direct clinical applicability of this nature of research, has prompted organizations such as [Dream Challenge 9.5] to design research competitions that aim at predicting overall survival of patients, using a wide variety of clinical predictors. In congruence with current clinical needs, our research aimed at building an effective and robust model that could be used for predicting patient survival. The problem was divided into data pre-processing, feature selection and modelling. In order to assess patient survival with greatest efficacy, the most important clinical and pathological features were then extracted from the given data.

4.1.1 Materials and Method

Study Design and Population

To identify factors associated with mCRPC, a cohort study was designed by the Prostate Cancer Dream Challenge 9.5 (PCC) [Dream Challenge 9.5], using phase 3 prostate cancer clinical trial data provided by the Project Data Sphere platform [Project Data Sphere, 2016] to predict overall patient survival. The main outcome measure was the time to event (days till death). The database consisted of around 150 baseline clinical covariates of over 2000 mCRPC patients treated with first-line docetaxel.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

The data represents 4 cancer trials of first line metastatic Hormone Refractory Prostate Cancer (HRPC) patients, where all patients received docetaxel treatment in the comparator arm [Project Data Sphere, 2016]. A total of 2070 patient information was provided from the following biotechnology and cancer institutes: AstraZeneca [Astrazeneca, 2016] (470 patients), Celgene [Celgene Canada 2016] (526 patients), Memorial Sloan Kettering Cancer center [MSKCC, 2016] (476 patients), and Sanofi [Sanofi, 2016] (598 patients). Where AstraZeneca is used for validation dataset, and the last three datasets are used for training and testing datasets.

Possible predictive factors

The recorded features included basic patient information such as age and race, along with patient medical history, previous medication use, vital sign information, and various lesion measurements.

Preprocessing

Handling Inconsistent data

The data provided had many data recording inconsistencies, such as uneven predictor categorization and measurement units. In the first step of data preprocessing these inconsistencies were eliminated. In order to handle inconsistencies in the first step new categories were extracted to achieve consistency between different databases. For example, different clinical predictors such as patient age and racial groups were recategorized in the process. In the second step, measurement units were unified across the databases to allow for proper comparison. In the final step, predictors with recording errors (such as a PSA of 0 which is physiologically impossible) were considered as missing values.

Handling Incomplete data

The next step in data pre-processing involved manual inspection of the data for missing information. Missing information was categorized in two distinct cases delineated below.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

1. Dataset removal: Of the 121 predictors provided, not all predictors were recorded in the datasets (datasets are Memorial Sloan Kettering Cancer center (D1), Celgene (D2), Sanofi (D3), and AstraZeneca (D4)). In order to have a complete data set, we either had to completely remove these features, or impute (extrapolate) them. Imputation refers to the process of extrapolating missing data using the patterns of existing data [Gelman and Hill, 2006]. Because of the enormity of missing information, data imputation would not provide accurate results [Gelman and Hill, 2006].

Figure 4.1 depicts the comparison between the number of missing features in the datasets of the database. In each bar graph, the common features between two and three of the datasets are depicted in dark blue, and the missing features are written on the horizontal axis below the bar, with the datasets under comparison indicated in the legend. For example in the third bar graph, the combination of features in D2 and D3 are under comparison, the number of missing features between these datasets is 14, as indicated in the horizontal axis, subsequently D2 and D3 have the least number of missing features amongst all other combinations of datasets. Subsequently datasets D2 and D3 have been chosen for the training and testing datasets. Comparison of D2 and D3 with the validation dataset D4 confirms this decision as the number of missing features within these three datasets is only 18, providing the greatest amount of information for training, testing and validation purposes.

In order to incorporate as much of the original data as possible, without utilizing imputation, it was necessary to eliminate the D1 dataset all together. As indicated in Figure 4.1, comparison of D1 with any other datasets, results in a high number of missing features. In addition to eliminating a dataset, particular attention was provided with respect to missing data in the training datasets (D2 and D3) and validation dataset D4. Clinical features with a large proportion of missing data or greatly lacking in a particular dataset were removed. Information that is missing from all three of the datasets, D2, D3, and D4 do not provide any significant information, and were eliminated.

2. Data Imputation: The second case of missing information consisted of sparse missing recorded data amongst a number of patients. Imputation was used to estimate missing values. Imputation was conducted using the 'MICE' package incorporated into the R programming

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

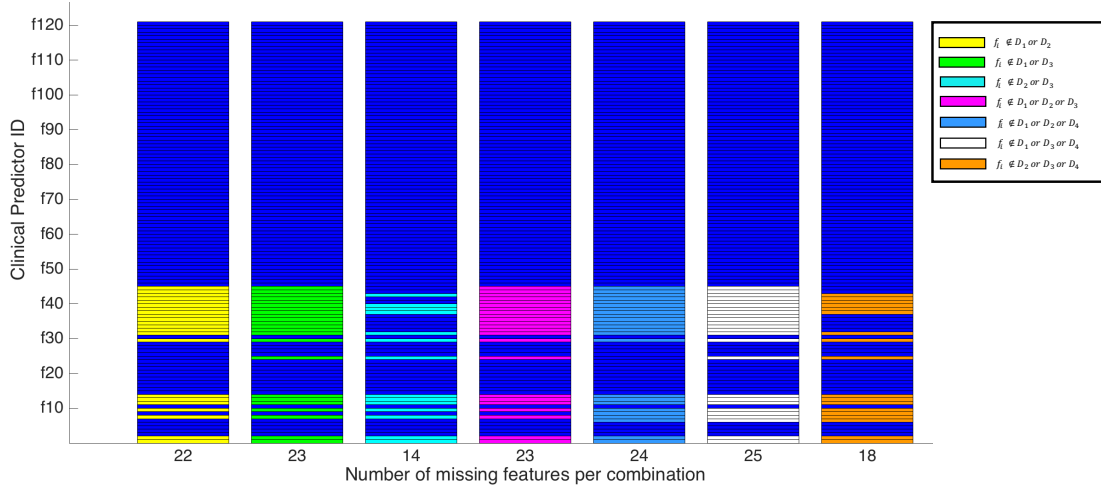


Figure 4.1: Comparison of missing factors within datasets D1, D2, D3 and D4. The datasets under comparison are indicated in the legend, with the colour assignment indicating the features missing within compared datasets, as indicated in the vertical axis as f_i where i is the index of the clinical predictor. In addition, the number of missing features is indicated in the horizontal axis below each bar. For each dataset combination, dark blue indicates common features between the datasets.

language [Buuren and Groothuis-Oudshoorn, 2011].

Additional feature extraction

The data provided was pre-categorized into primary and supplementary data. Based on clinical relevance and applicability, data from the supplementary data set was added to the primary data and used for training and validation purposes. Supplementary data in Project Data Sphere platform [Project Data Sphere, 2016] include additional event related information regarding patient lab tests, and lesion measurements. Lesion measurements include the target tumour sizes (tumours that have been clinically measured), along with non target tumours (non clinically measured tumours) and other qualitative information such as tumour progression, regression or stability. Lab test information in the primary data set includes all patient screening results previous to treatment, such as complete blood cell count test, prostate specific antigen etc. Further information regarding these primary data values was extracted from the supplementary lab test data, such as the normal range of clinical predictors

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

mentioned in the primary data set. It is important to note that all event related information included baseline (information acquired previous to study start date) and non baseline measurements. For model training and validation purposes only the baseline values were used. Using the information provided, additional clinical predictors were extracted and created. Deviation of the clinical features from the normal range, summation of target lesion sizes categorized according to site, and manipulation of specific features were used as new clinical predictors and appended to the primary data set.

1. Deviation from normal range: The normal range of features in each dataset is dependant on the dictionaries provided in the database. In order to extract this set of features, the following rules were utilized:

- If the lab value is within the normal range, the feature was assigned a value "0".
- If the value was above the normal range, the deviation from the upper limit was assigned a positive sign.
- If the value was below the normal range, the deviation from the lower limit was assigned a negative sign.

A diagrammatic representation of the assigned rules is shown in Figure 4.2, where column a in Figure 4.2 shows the distribution of clinical features that lie within low, medium and high range for each database D2, D3 and D4 with a logarithmic range. Column b in Figure 4.2 shows the deviation of these values from the normal range, in accordance with the rules above. For example, ALP (Alkaline phosphatase) has a distribution of ALP values in all three ranges (high, medium and low) as shown in column a of Figure 4.2. According to rule 1 whichever value that is within the normal range will be assigned a value of 0, shown in column b of Figure 4.2, this results in the normal data points to fall on the vertical axis. In the case of ALP there are many values in the high range, in accordance with rule 2, this corresponds to a high ratio of positive values shown in column b of Figure 4.2. Conversely, there are very few

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

data points in the low value range, corresponding to no negative data points in column b of Figure 4.2 resulting in a right handed distribution. On the other hand, Hemoglobin values are either low or normal, and do not exceed the normal range, causing a left handed distribution with all data points falling below the median shown in column b of Figure 4.2. By graphically representing the clinical features with a high distribution range, in column b of Figure 4.2, may be used to immediately distinguish clinical features with high prognostic value.

2. Lesion summation: The second set of extracted features consists of baseline target lesion measurements. For each patient the following rules were applied:

- If there was no lesion information available, the feature was assigned a value of 0.
- If there was only a single lesion measurement available, the measurement value was assigned to the feature.
- If there were multiple lesion measurements available, the sum of all the lesion measurements was assigned to the feature.

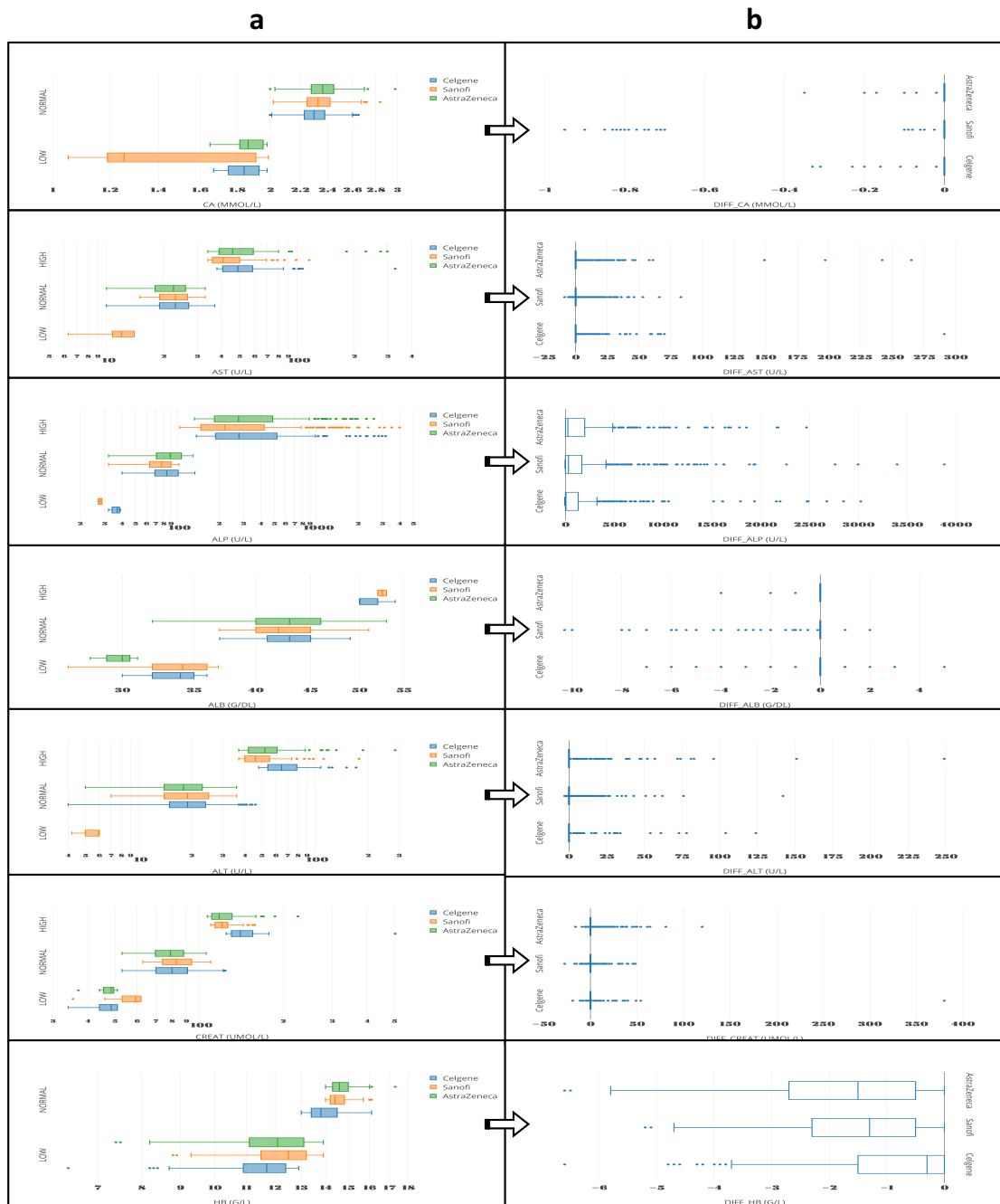
Lesions in different sites were categorized and summed accordingly (bone, lymph node, etc).

3. Feature manipulation: In addition to extracting novel features, manipulation was completed on certain features. For example, the log of the Prostate Specific Antigen PSA was obtained as a feature. The log of the PSA ensures that the values obtained do not contain any outliers, and are more indicative of clinical results[Nuhn *et al*, 2014, Chen, 2014, Rathkopf *et al*, 2014].

Statistical analysis

For statistical analysis, the Cox proportional hazards model was used. In survival analysis studies, the cox model examines the relationship between covariates and time-to-event data. For each covariate, this relationship is expressed as the model coefficient or parameter, which

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization



4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

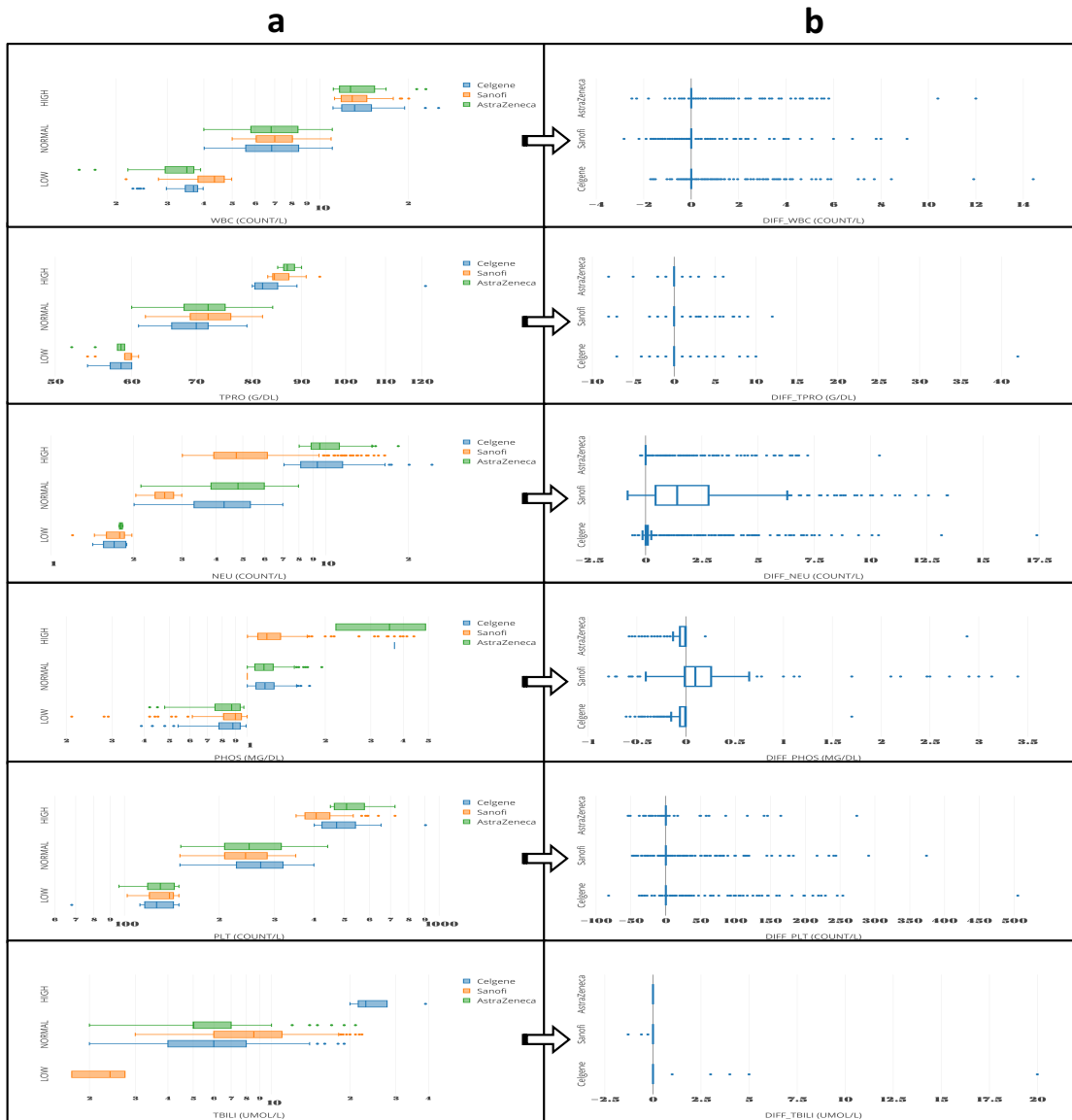


Figure 4.2: a) Logarithmic distribution of continuous clinical features within the possible clinical low, medium and high range for databases Celgene (D2), Sanofi (D3), and AstraZeneca (D4). b) Deviation of the clinical values from the normal range, where values within a normal range are assigned a value of 0, values above the normal range are assigned a positive value, and values below the normal range are assigned a negative value. The features under review are as follows: CA, AST, ALP, ALB, ALT, CREAT, HB, WBC, TPRO, NEU, PHOS, PLT, and TBILI.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

is directly proportional to the effect of the covariate on overall patient survival[George, 2014]. The cox model served as the basis for our two phase feature selection model designed to identify covariates which were both statistically significant (represented by covariate p_value) in predicting mCRPC patients' time-to-events and also produced the most accurate models. The first phase of the feature selection process was designed to extract statistically significant covariates, to reduce the dimensionality of the feature space and improve our understanding of the relationship between clinical features and patient survival. In phase two, the dimensionality of the feature space obtained from phase one was further reduced to obtain the most accurate survival prediction model.

Phase One

Individual layers make up the basic feature selection structure in phase one. In the first layer, each covariate is separately analyzed based on their p_value, obtained using the Cox model. This value represents the level of association between the covariate and length of survival. A smaller value indicates a higher level of association. In this layer, the covariate with the smallest p_value is extracted from the overall feature space and appended onto the empty feature set Alpha. This newly created set serves as input to the following layer. In layer two, remaining covariates are individually appended to the set Alpha from layer one. In this layer, the Alpha set with the lowest p_value less than 0.05 is selected as the layer output. This process is repeated in subsequent layers and continues to grow the set Alpha until the stipulated criteria threshold is no longer met. The final resulting set Alpha at the end of phase one represents the minimum number of features with the highest level of association to the length of survival. This process is formulated in Table 4.3 and graphically represented in Figure 4.4.

Phase Two

While all the features extracted in the first phase are strongly associated with patient survival times, they are all not necessarily effective in constructing the most accurate prediction model. Therefore, the purpose of phase two is to select the most effective covariates from

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Figure 4.3: Rules in phase one of factors selection for dimensionality reduction based on restricting model's $p_values < 0.05$

Phase 1	P_value Statement	If-Then Condition
Layer 1	$\min_{i_1 \in \{1, \dots, n\}} (pvalue(f_{i_1})) = pvalue(f_{i_1'})$	if $pvalue(f_{i_1'}) < 0.05$, then go to layer 2.
Layer 2	$\min_{i_2 \in \{1, \dots, n\} - \{i_1\}} (pvalue(f_{i_1'} \& f_{i_2})) = pvalue(f_{i_1'} \& f_{i_2'})$	if $pvalue(f_{i_1'} \& f_{i_2'}) < 0.05$, then go to layer 3.
Layer 3	$\min_{i_3 \in \{1, \dots, n\} - \{i_1, i_2\}} (pvalue(f_{i_1'} \& f_{i_2'} \& f_{i_3})) = pvalue(f_{i_1'} \& f_{i_2'} \& f_{i_3'})$	if $pvalue(f_{i_1'} \& f_{i_2'} \& f_{i_3'}) < 0.05$, then go to layer 4.
.	.	.
Layer $\gamma - 1$	$\min_{i_{\gamma-1} \in \{1, \dots, n\} - \{i_1, i_2, \dots, i_{\gamma-2}\}} (pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-2}'} \& f_{i_{\gamma-1}})) = pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-2}'} \& f_{i_{\gamma-1}'})$	if $pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-2}'} \& f_{i_{\gamma-1}'}) < 0.05$, then go to layer γ .
Layer γ	$\min_{i_\gamma \in \{1, \dots, n\} - \{i_1, i_2, \dots, i_{\gamma-1}\}} (pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-1}'} \& f_{i_\gamma})) = pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-1}'} \& f_{i_\gamma'})$	if $pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_{\gamma-1}'} \& f_{i_\gamma'}) < 0.05$, then go to layer $\gamma + 1$.
Layer $\gamma + 1$	$\min_{i_{\gamma+1} \in \{1, \dots, n\} - \{i_1, i_2, \dots, i_\gamma\}} (pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_\gamma'} \& f_{i_{\gamma+1}})) = pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_\gamma'} \& f_{i_{\gamma+1}'})$	if $pvalue(f_{i_1'} \& f_{i_2'} \& \dots \& f_{i_\gamma'} \& f_{i_{\gamma+1}'}) > 0.05$, then stop phase 1 and ignore the selected predictor in the current layer.

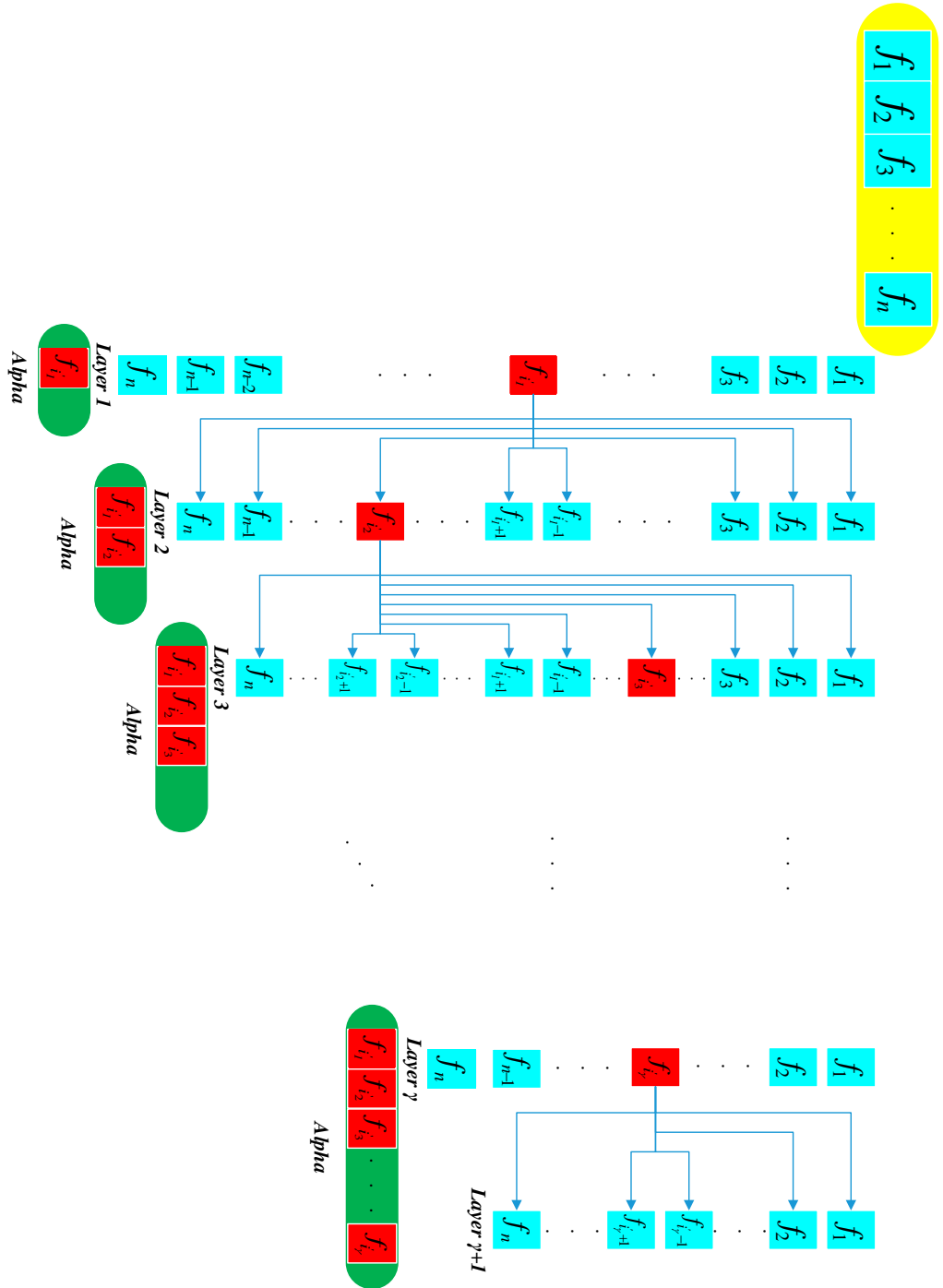


Figure 4.4: Graphical representation of the feature selection process detailed in phase 1. Red boxes represent the extracted feature in each layer.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

the first phase's output Alpha set. This phase is similar to phase one in terms of structure, with a few key differences. In both phases, different layers are used to extract covariates from the available feature space. However, in phase two, the output set Beta is evaluated not in terms of p_value , but in terms of model prediction accuracy, measured using the integrated time-dependent area under the receiving operating characteristic curve (iAUC) [Saha-Chaudhuri and Heagerty, 2013]. In addition, while in phase one the algorithm was designed to minimize the p_value of Alpha set, the set Beta in phase two is expanded by maximizing iAUC. This process is repeated until the set accuracy is no longer increased in value or decreases. Where no change in accuracy indicates redundancy of the iterated feature; and a decrease in accuracy is indicative of an informative conflict between extracted features. The algorithm for phase two is mathematically represented in Table 4.5 and visualized in Figure 4.6.

4.1.2 Results

Prognostic model results

Figure 4.7(a) on the top left is obtained from the results of phase one of the method, it illustrates the clinical feature combinations vs the log of the p_value . As evident in Figure 4.7a the combination of values with the lowest log p_value is AST, the second significant combination is AST and DIFF_HB which is the deviation from the normal range of hemoglobin. Predictor combination extraction continued until p_values lower than 0.05, resulting in eleven clinical feature combinations. The eleven feature combinations extracted from phase one were then used as an input set for phase two of the method which focuses on the accuracy of survival analysis shown in Figure 4.7b. The results of Figure 4.7(b) show that ALP alone has the accuracy close to 66. Evident in Figure 4.7(b) additional features were appended to ALP, until an increase or plateau in accuracy. This resulted in ten clinical feature combinations. Figure 4.7(c) is a graphical representation of the feature space, simultaneously showing the consideration of p_value and accuracy for determination of the clinical feature combination. The tracked black line represents the method's path in determining the final selected clinical feature

Figure 4.5: Rules in phase two of factors selection for dimensionality reduction based on maximizing model's accuracy

Phase 2	Accuracy Statement	If-Then Condition
Layer 1	$\max_{j_1 \in \{1, \dots, p\}} (accuracy(f_{i_{j_1}})) = accuracy(f_{i_{j_1}})$	Go to layer 2.
Layer 2	$\max_{j_2 \in \{1, \dots, p\} - j_1} (accuracy(f_{i_{j_1}} \& f_{i_{j_2}})) = accuracy(f_{i_{j_1}} \& f_{i_{j_2}})$	if $accuracy(f_{i_{j_1}}) < accuracy(f_{i_{j_1}} \& f_{i_{j_2}})$, then go to layer 3.
Layer 3	$\max_{j_3 \in \{1, \dots, p\} - \{j_1, j_2\}} (accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& f_{i_{j_3}})) = accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& f_{i_{j_3}})$	if $accuracy(f_{i_{j_1}} \& f_{i_{j_2}}) < accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& f_{i_{j_3}})$, then go to layer 4.
.	.	.
Layer $\lambda - 1$	$\max_{j_{\lambda-1} \in \{1, \dots, p\} - \{j_1, j_2, \dots, j_{\lambda-2}\}} (accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-2}}} \& f_{i_{j_{\lambda-1}}})) = accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-1}}})$	if $accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-2}}}) < accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-1}}})$, then go to layer λ .
Layer λ	$\max_{j_{\lambda} \in \{1, \dots, p\} - \{j_1, j_2, \dots, j_{\lambda-1}\}} (accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-1}}} \& f_{i_{j_{\lambda}}})) = accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda}}})$	if $accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda-1}}}) < accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda}}})$, then go to layer $\lambda + 1$.
Layer $\lambda + 1$	$\max_{j_{\lambda+1} \in \{1, \dots, p\} - \{j_1, j_2, \dots, j_{\lambda}\}} (accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda}}} \& f_{i_{j_{\lambda+1}}})) = accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda+1}}})$	if $accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda}}}) > accuracy(f_{i_{j_1}} \& f_{i_{j_2}} \& \dots \& f_{i_{j_{\lambda+1}}})$, then stop phase 2 and ignore the selected predictor in the current layer.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

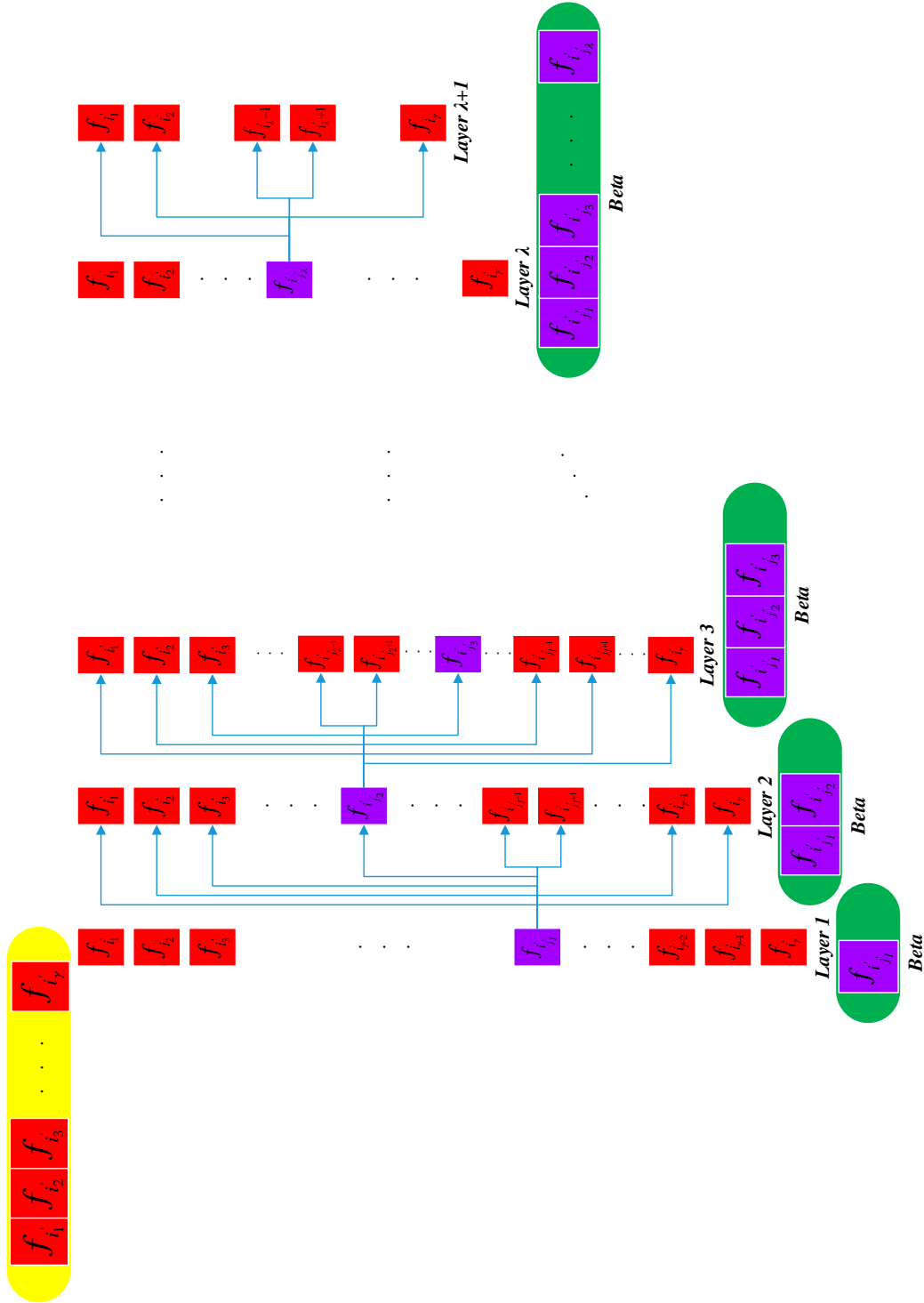


Figure 4.6: Graphical representation of the feature selection process detailed in phase 2. Purple boxes represent the extracted feature in each layer.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

combination.

Best Performing Prognostic Factors

In total 1592 mCRPC patients were treated with docetaxel-containing chemotherapy. The baseline characteristics of clinical feature of these patients in the training, testing and validation sets which were extracted from the presented algorithm are displayed in Table 4.1.

The mean age at diagnosis was 68. As evident in the table lymph node metastasis size is 44.5% of patients in the training set for greater than and less than 52 mm lymph node size. Liver metastasis is present in 10.9% of patients in the training set, and adrenal metastasis is present in 3% of patients in the training set. Despite the training set value especially for the adrenal metastasis having a low percentage of patients in the training set, the values in the testing and validation are also significantly low, therefore allowing for accurate comparison. This trend is also observed in subsequent clinical features. The sparse clinical features as presented in Table 4.1, are DIFF_ALP in the lower than normal range, presence of adrenal metastasis and DIFF_ALB in the higher than normal range. It is notable to mention that the DIFF_HB does not have values that deviate higher than the normal range.

Univariate and Multivariate Analysis

The focus of our study is on multivariate analysis, for comparative reasons, an univariate analysis was conducted using the proposed method. As presented in Table 4.2, multivariate analysis was performed on the clinico-pathological variables. The extracted factors of the our multivariate model that showed a significant risk of death are contained in the following: ALP (alkaline phosphatase), HB_DIFF (deviation from normal range of hemoglobin), Lymph_size (the cumulative lymph node size), AST (aspartate aminotransferase), ECOG_C (baseline patient performance status), DIFF_ALP (deviation from normal range of alkaline phosphatase), DIFF_ALB (deviation from normal range of albumin), Liver (baseline liver lesions), Adrenal (baseline adrenal lesions) and ACE_inhibitors (prior ACE inhibitors). Hazard ratio values below 1 in table 4.2 are indicative of an inverse relationship with risk of death which include high

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

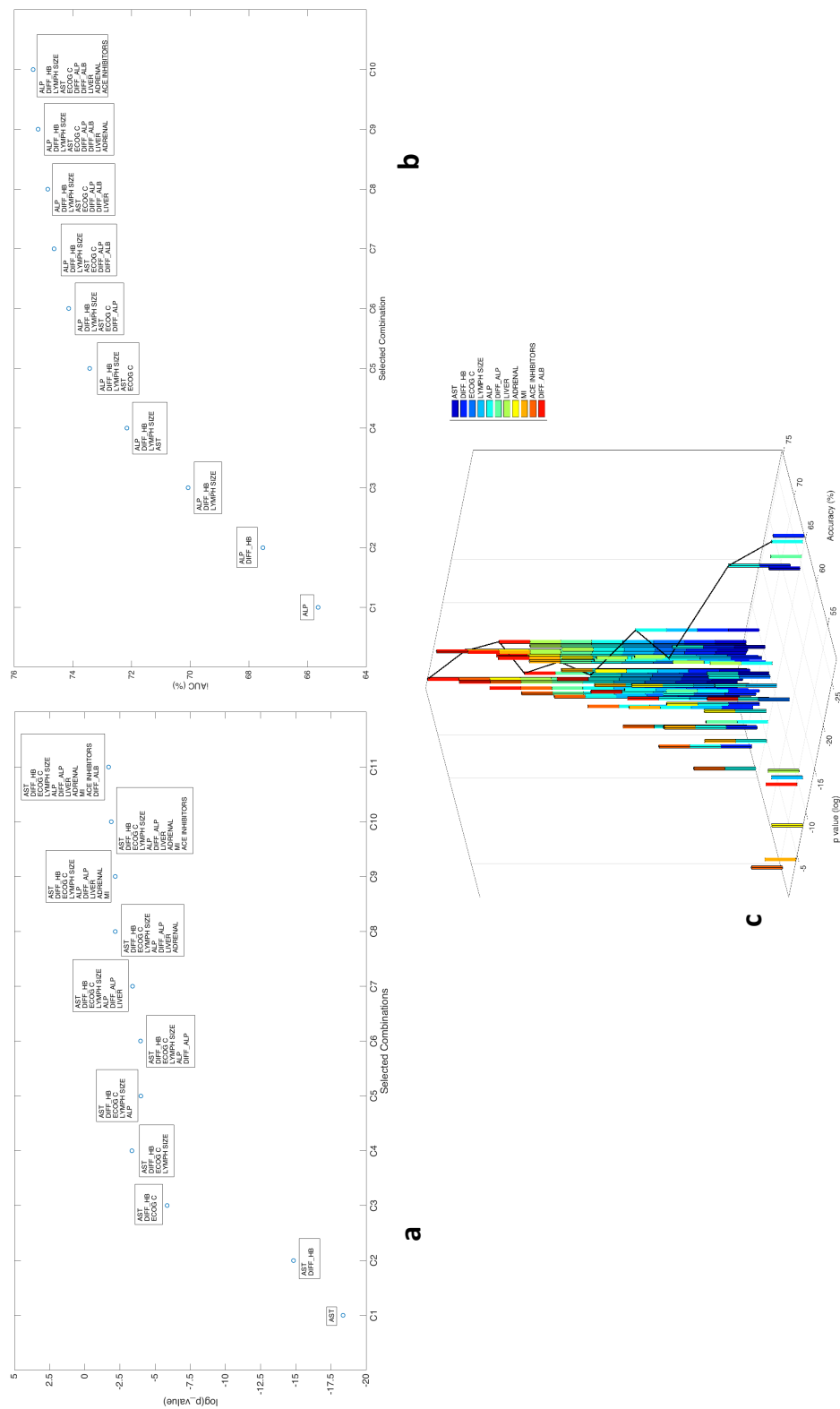


Figure 4.7: a) Results of phase one of the method, clinical feature combination vs log of the p_value. Feature combinations extracted from phase one were inputted into phase two of the method shown in (b). b) Results of phase two of the method, ALP is selected in first layer as most effective factor for obtaining accuracy. Additional prognosis factors were appended onto ALP until an increase or plateau in accuracy. c) Graphical representation clinical and pathological factors space depicts the consideration of p_value and accuracy for highlighting the selection of various clinico-pathological factors combinations. Tracked black line shows our method's path in determining the final selected combination.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

Table 4.1: Clinical and pathological parameters of the study cohort and comparative analysis results.

		Training (n=841)		Testing (n=281)		Validation (n=470)	
ALP		n	%	n	%	n	%
0<ALP<83		208	24.7	82	29.2	84	17.8
83<ALP<130.5		212	25.2	59	20.1	113	24
130.5<ALP<266		210	24.9	72	25.6	130	27.6
266<ALP<3983		211	25.1	68	24.2	143	30.4
DIFF_ALP							
Normal		386	45.9	113	40.2	198	42.1
Low		4	0.5	2	0.7	0	0
High		451	53.6	166	59.1	272	57.9
AST							
0<AST<20		210	24.9	60	21.3	110	23.4
20<AST<25		174	20.1	60	21.3	93	19.8
25<AST<32		247	29.4	80	28.5	141	30
32<AST<328		201	23.9	81	28.8	126	26.8
ACE_Inhibitors							
No prior use		654	77.7	225	80.1	413	87.9
Prior use		187	22.3	56	19.9	57	12.1
Adrenal Metastasis							
Absent		816	97	274	97.5	446	94.9
Present		25	3	7	2.5	24	5.1
Liver Metastasis							
Absent		749	89.1	255	90.7	406	86.4
Present		92	10.9	26	9.3	64	13.6
Lymph Node Metastasis Size							
Absent		466	55.4	143	50.9	307	65.3
<52mm		187	22.2	68	24.2	81	17.2
>52mm		188	22.3	70	24.9	82	17.4
DIFF_ALB							
Normal		755	89.7	238	84.7	466	99.1
Low		72	8.6	42	14.9	4	0.8
High		14	1.6	1	0.3	0	0
DIFF_HB							
Normal		271	32.2	40	14.2	72	15.3
Low		570	67.7	241	85.8	398	84.7
ECOG							
0		407	48.4	129	45.9	247	52.6
1		398	47.3	142	50.5	223	47.4
2		36	4.3	10	3.6	0	0
LDH							
LDH<185		210	25	39	13.9	132	28
185<LDH<218		206	24.5	78	27.7	111	39.5
218<LDH<278		215	25.5	82	29.2	97	20.6
LDH>278		210	25	82	29.2	130	27.6

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

range of DIFF_ALP (HR:0.4339, 95% CI: 0.4314 to 0.4363) and prior use of ACE_Inhibitors (HR:0.8627, 95% CI: 0.6572 to 1.1325). Conversely hazard ratios deviating from 1 proportionally indicate a greater risk of death. The largest hazard ratio value obtained was indicated for ALP>269 (HR: 22.66, 95% CI: 22.5391 to 22.7943) shows a significant risk of death, in addition ALP between 131 and 269 has a high HR of 2.26 (95% CI: 2.2466 to 2.2720) in comparison with ALP values between 0 and 85. This is indicative of the overall importance of ALP in survival analysis for patients with mCRPC. The presence of adrenal metastasis and liver metastasis have HR values of 2.2570 (95% CI: 1.3040 to 3.9065) and 1.7219 (95% CI: 1.2541-2.3643) respectively, in comparison with non lesion patients in these categories. DIFF_HB hazard ratio that are lower than the normal range is 1.3738 (95% CI: 1.2479 to 1.5124). The HRs of patients with ECOG performance status of 1 or 2 are 1.19 and 1.43 respectively compared with patients with a performance status of 0.

In addition, the obtained results from univariate analyses are presented in Table 4.2. The following factors were associated with greater risk of death: ALP > 269 (Hazard Ratio (HR): 1.3520, 95% Confidence Interval (CI): 1.3518 to 1.3523), AST > 31.25 (HR: 1.3336, 95% CI: 1.3287 to 1.3385), presence of adrenal metastasis (HR: 2.7402, 95% CI: 1.5975 to 4.7001), presence of liver metastasis (HR: 1.9425, 95% CI: 1.4349 to 2.6296), lymph node metastasis size>52mm (HR: 1.3393, 95% CI: 1.3378 to 1.3408), low range of DIFF_ALB (HR: 1.4278, 95% CI: 1.3466 to 1.5138), low range of DIFF_HB (HR: 1.5553, 95% CI: 1.4259 to 1.6965), and ECOG performance status 1 and 2 (HR: 1.5556, 95% CI: 1.2972 to 1.8656 & HR: 2.42, 95% CI: 2.0179 to 2.9023, respectively). In addition, all clinical and pathological factors in Table 4.2 except ACE_Inhibitors (p_value=0.618) were also significant predictors of overall survival. These results disclosed that the presence of adrenal and liver metastases had higher risk for mortality.

Kaplan-Meier Curves

In order to adjust for potential confounding factors in survival prediction, multivariate analysis was performed in our hybrid model. Variables considered as potential predictors for

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

Table 4.2: Univariate and multivariate our hybrid method results for evaluating predictors associated with overall survival

		Univariate Analysis			Multivariate Analysis		
		HR	95% CI	P_Value	HR	95% CI	P_Value
ALP				2.2689e-12			0.0153
	0 < ALP < 85	ref	-		ref	-	
	85 < ALP < 131	1.0225	1.0223-1.0227		1.2585	1.2514-1.2656	
	131 < ALP < 269	1.0820	1.0818-1.0822		2.2593	2.2466-2.2720	
	ALP > 269	1.3520	1.3518-1.3523		22.6663	22.5391-22.7943	
DIFF_ALP				7.9469e-12			0.0268
	Normal	ref	-		ref	-	
	Low	0.9980	0.9978-0.9982		1.0195	1.0137-1.0253	
	High	1.0910	1.0907-1.0912		0.4339	0.4314-0.4363	
AST				2.5372e-11			1.6810e-04
	0 < AST < 19.075	ref	-		ref	-	
	19.075 < AST < 24	1.0513	1.0475-1.0552		1.0358	1.0310-1.0405	
	24 < AST < 31.25	1.1192	1.1151-1.1234		1.0823	1.0774-1.0873	
	AST > 31.25	1.3336	1.3287-1.3385		1.2240	1.2184-1.2296	
ACE_Inhibitors				0.6180			0.2875
	No prior use	ref	-		ref	-	
	Prior use	0.9346	0.7163-1.2193		0.8627	0.6572-1.1325	
Adrenal Metastasis				2.5053e-04			0.0036
	Absent	ref	-		ref	-	
	Present	2.7402	1.5975-4.7001		2.2570	1.3040-3.9065	
Liver Metastasis				1.7337e-05			7.8031e-04
	Absent	ref	-		ref	-	
	Present	1.9425	1.4349-2.6296		1.7219	1.2541-2.3643	
Lymph Node Metastasis Size				1.1194e-05			1.7588e-04
	Absent	ref	-		ref	-	
	<52mm	1.0607	1.0595-1.0619		1.0526	1.0514-1.0539	
	>52mm	1.3393	1.3378-1.3408		1.2895	1.2880-1.2910	
DIFF_ALB				7.0008e-05			0.0672
	Normal	ref	-		ref	-	
	Low	1.4278	1.3466-1.5138		1.2111	1.1311-1.2968	
	High	0.8369	0.7893-0.8873		0.9087	0.8486-0.9730	
DIFF_HB				1.0953e-12			3.7647e-06
	Normal	ref	-		ref	-	
	Low	1.5553	1.4259-1.6965		1.3738	1.2479-1.5124	
ECOG				1.8749e-06			0.0709
	0	ref	-		ref	-	
	1	1.5556	1.2972-1.8656		1.1972	0.9848-1.4555	
	2	2.4200	2.0179-2.9023		1.4333	1.1790-1.7425	

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

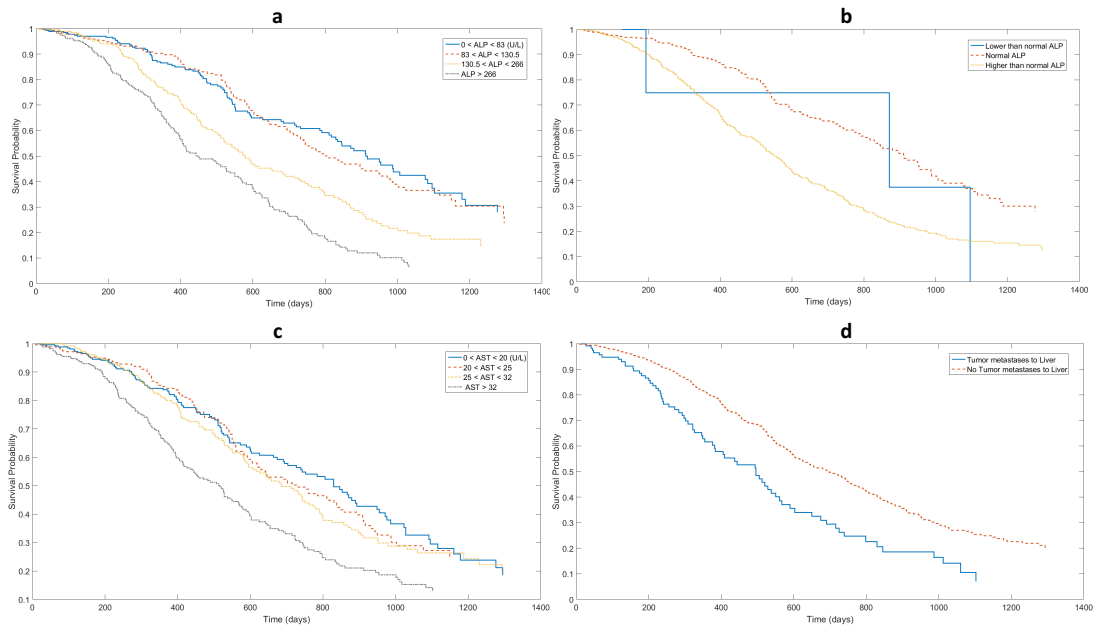


Figure 4.8: Kaplan–Meier curves for overall survival (OS) according to ALP, DIFF_ALP, AST and Liver Metastases. a) The blue and red lines indicate survival for patients with $0 < \text{ALP} < 83$ and $83 < \text{ALP} < 130.5$, respectively, the orange and gray lines indicate patients' OS with $130.5 < \text{ALP} < 266$, $266 < \text{ALP} < 398.3$, respectively; b) we stratified the patients into three cohorts with different ranges of ALP named DIFF_ALP; OS according to low, normal, and high are indicated by blue, red, and orange lines; c) the patients were divided into four subgroups according to the AST; the blue, red, and orange lines indicate overall survival of three cohorts with $0 < \text{AST} < 32$, and the purple line indicates patients with $32 < \text{AST} < 328$; d) liver metastases (blue line) is associated with shorter OS time in comparison with no liver metastases (red line).

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

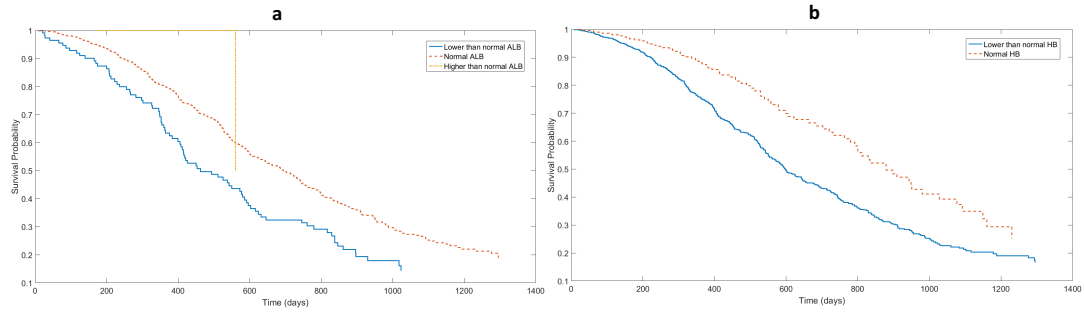


Figure 4.9: Kaplan-Meier curve for OS according to DIFF_ALB and DIFF_HB. a) We stratified the patients into three cohorts with low, normal, and high ALB. The blue red, and orange lines indicate survival for patients with low, normal, and high, respectively; b) the patients with normal range of HB (shown in red) has a shorter OS than those with lower (shown in blue).

multivariate modeling were selected by two phase analyses and subsequently tested in a stepwise manner outlined in statistical analysis section. Entry and retention of variables in the model were set at a statistically significant p -value of 0.05 and a maximum level of accuracy.

Kaplan-Meier curves were used to estimate the overall survival (OS) according to the extracted clinico-pathologic factors for a hypothetical cohort, to indicate statistical significance. All factors were stratified using the following cut-off values: existence of metastasis, quartile division of population, and deviation of values from the normal range.

Indeed, the clinico-pathological factors with Hazard Ratio (HR) on the time to death in the univariate analyses were used to generate Kaplan-Meier curves (continuous variables were categorized in K-M curves, similar to data presented in Table 4.2). Kaplan-Meier curves may be used for survival estimation, despite patient discontinuation or varying lengths of studied times. The trend of the K-M curve indicates that death occurs more frequently at the beginning of time, this is attributable to the high survival rate and large population of patients at the beginning. Consequently greater precision of estimates is achieved at the beginning of time when there are a greater number of patients, rather than at the end of the study, when the patient population has diminished, due to patient deaths and discontinuations.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

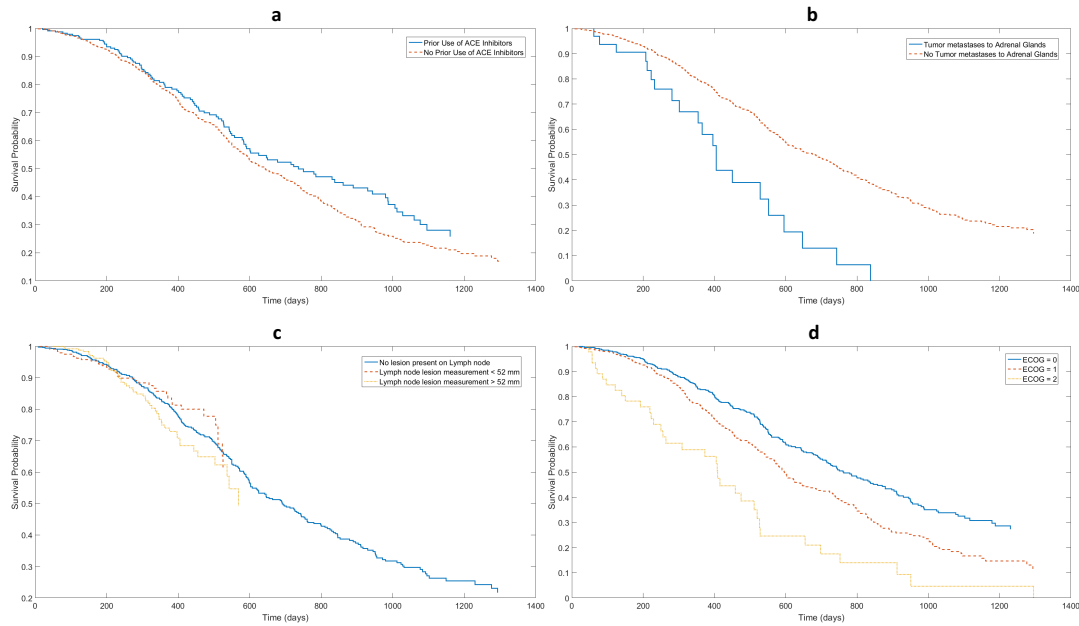


Figure 4.10: Kaplan-Meier survival curves are shown for the patients according to prior use of ACE inhibitors, Adrenal metastases, presence of lymph node lesion, and ECOG. a) OS according to the prior use and no prior use of ACE inhibitors which are indicated with blue and red respectively; b) the patients with adrenal metastasis (shown in blue) has a shorter OS than those without (shown in red); c) the patients based on the lymph node metastases size are categorized in three cohorts (no lesion, <52mm, >52mm) which their OS are indicated by blue, red, and orange lines respectively; d) The patients are stratified into three ECOG classes 0, 1, and 2 (blue, red and orange curves) which represent low, intermediate and high risk, respectively.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

As shown in Figure 4.8(a) the patients with $130.5 < \text{ALP} < 266$ and $266 < \text{ALP} < 398.3$ (shown in orange and gray respectively) had a greater probability of mortality with respect to patients having $0 < \text{ALP} < 83$, and $83 < \text{ALP} < 130.5$ (shown in blue and red respectively). Moreover, patients who achieved a higher DIFF_ALP range (shown in orange), revealed worse OS with respect to the normal range shown in red ($p_value = 7.9469e-12$). As shown in figure 4.8(b) patients with a lower than normal DIFF_ALP value (shown in blue) do not demonstrate a predictable trend, this is attributable to the low number of patients in this cohort (refer to Table 4.1). In addition, compared to cases with first three quartiles (shown in blue, red, orange), cases with $32 < \text{AST} < 328$ (shown in gray) was definitely associated poorer OS time (Figure 4.8(c)). Finally, figure 4.8(d) shows that patients with liver metastasis (shown in red) has a shorter overall survival (OS) than those without (shown in blue).

Patients with low DIFF_ALB levels (shown in blue) were found to have significantly worse OS ($P_value = 7.9469e-12$) compared to patients with normal DIFF_ALB levels indicated with red (Figure 4.9(a)). As shown in figure 4.9(a) patients with a higher than normal DIFF_ALB value indicated with orange, do not present an interpretable trend, which is once again attributed to 14 patients in this cohort (refer to Table 4.1). Furthermore, Figure 4.9(b) indicates the DIFF_HB values, both the lower than normal and normal values follow a predictable trend and are indicated with blue and red respectively. Low DIFF_HB values are associated with an increased risk of mortality ($p_value = 1.0953e-12$).

Although ACE Inhibitors was not a significant clinical feature ($p_value = 0.6180$), patients with prior use of ACE inhibitors (shown in blue) had a higher probability of OS as presented in Figure 4.10(a). Univariate analysis demonstrated that, compared with cases without Adrenal metastases, Adrenal metastases was still significantly associated with shorter OS time (shown in blue) (Figure 4.10(b)). Patients with lymph node lesion measurement $> 52\text{mm}$ (shown in orange) showed worse OS ($p_value = 1.1194e-05$) compared to those with $< 52\text{mm}$ or no lesion presence (shown in red and blue respectively) (Figure 4.10(c)). We compared the survival probability by ECOG categories. They were stratified into three cohorts 0, 1, and 2. Patients with ECOG = 2 had a significantly shorter OS than patients with ECOG performance status 0 and 1 (Figure 4.10(d)). The hazard ratios in each group were shown in Table 4.2.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Table 4.3: Univariate and multivariate results based on LDH predictor imputation on overall survival

	Univariate Analysis			Multivariate Analysis		
	HR	95% CI	P_Value	HR	95% CI	P_Value
LDH			1.0182e-18			0.0047
LDH<185	ref	-		ref	-	
185<LDH<218	1.0721	1.0716 - 1.0726		1.0320	1.034 - 1.0327	
218<LDH<278	1.1673	1.1667 - 1.1678		1.0726	1.0719 - 1.0732	
LDH>278	1.4946	1.4940 - 1.4953		1.1996	1.1989 - 1.2004	

LDH Imputation performance

Since the datasets used in our model were collected individually, there is non-random missing clinico-pathological factors due to not having them recorded. For example, information regarding LDH is available in Memorial Sloan Kettering Cancer center (D1) and Celgene (D2) datasets (1002 patients), however there is no information regarding LDH in Sanofi (D3) dataset having a total of 598 patients.. Since LDH covariate is one of the prognosis factors which has a confirmed role in mortality probability of mCRPC patients [Nuhn *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Gravis *et al*, 2015, Chi *et al*, 2015, Uemura *et al*, 2016], we applied Multivariate Imputation by Chained Equations (MICE)[Buuren and Groothuis-Oudshoorn, 2011] to impute the missing values related to LDH in third dataset according to other datasets, where this missing factor is iteratively imputed. However, the imputation trend needs to be explained and clarified percisely: the complete records not included in the validation set are used to impute LDH factor of the training and testing sets, therefore information distribution of validation records is not used during imputation of training and testing sets. Consequently, the obtained results of MICE provides a completed version of the training and testing datasets with LDH factor and we are able to evaluate the impact of LDH imputation on the predictors' performance.

As shown in Table 4.3, the p_value related to univariate analysis is obtained 1.0182e-18 and also the multivariate analysis indicates a p_value of 0.0047, meaning that both analyses confirm

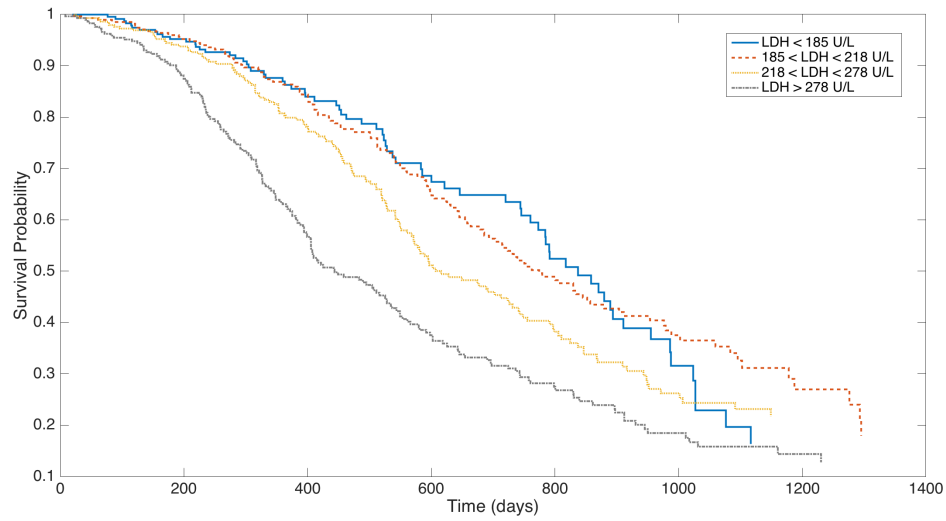


Figure 4.11: Kaplan-Meier survival curves according to LDH factor for all the training and testing samples. The patients are divided into four quartile groups according to the LDH. The OS for patients with a LDH>278 (gray curve) were significantly shorter than those in other quartile groups (blue, red and orange curves).

statistically significance of LDH marker for estimating survival time. Univariate analysis indicates that patients with LDH>278 (HR: 1.4946, 95% CI: 1.4940 to 1.4953) are significantly associated with mortality. In addition, multivariate analysis shows the hazard ratio of patients with LDH level more than 278 is 1.1996 that is another confirmation on overall importance of LDH in survival analysis for patients with mCRPC.

The patients were stratified into four subgroups according to LDH level: 1) LDH<185 (210 patients); 2) 185<LDH<218 (206 patients); 3) 218<LDH<278 (215 patients); 4) LDH>278 (210 patients). The OS for the patients with LDH>278 U/L (shown in gray) were significantly shorter than those cohorts with a lower LDH (depicted in blue, red, orange curves) (Figure 4.11). Accuracy of our methodology is affected by imputing LDH factor. Indeed, adding imputed LDH factor to incomplete datasets has an enhancing effect on AUC (shown in Table 4.4).

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Table 4.4: The C-Index, AUC at 12, 18, 24 months, and iAUC for Halabi et al.[Halabi *et al*, 2014] model and also our methods including and excluding LDH imputation.

	AUC 12	AUC 18	AUC 24	C-Index	iAUC
Halabi et al. Method	77.41%	73.7%	76.13%	70.92%	75.37%
Our model without LDH Imputation	76.83%	74.36%	78%	71.09%	76.6%
Our model with LDH Imputation	78.63%	75.12%	77.95%	72.1%	77.21%

Model Performance

A hybrid model by searching the feature space based on p_value and accuracy, for finding prognosis factors, is designed in this study. Not only p_value as a statistically significant term, but also accuracy as a criterion for tracking conflict among prognostic factors are considered in this method to search the factors combination space efficiently. The performance of this method is assessed for its discriminative ability by computing AUC, C_Index and iAUC. Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) have been recently extended to assess prognosis models. The AUC score ranges from 1 to 0.5 with 1 being perfect prediction and 0.5 being no better than luck. Concordance Index (C_Index) is an extension of AUC but it is independent of threshold and suitable for survival analysis[Steck *et al*, 2008]. C_Index is a measure to assess the separation of two survival distributions[Koziol and Zhenyu, 2009]. Also, the integrated time-dependent area under the curve (iAUC) is a weighted average of the AUC across a follow-up period which provides a measure for model accuracy[Saha-Chaudhuri and Heagerty, 2013]. The iAUC was computed for the validation dataset based on extracted coefficients from the training dataset to assess the prognostic utility and measure the model predictive power.

Our model utility is tabulated by comparing this model to recent prognostic model [Halabi *et al*, 2014] (Table 4.4). To evaluate the prognosis value of the LDH feature to the hazard score, we calculated iAUC for two models including and excluding LDH factor. The iAUC value of the model

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

including LDH factor is generally larger than that of the model excluding LDH. As a result, the mortality probability at 12, 18, and 24 months versus observed probability is obtained. The AUC values at all these months for model with LDH imputation are higher than the AUC values obtained from Halabi method [Halabi *et al*, 2014] (shown in Table 4.4). The iAUC and C-Index for risk score obtained from our method when we impute LDH factor for all patients are 77.21% and 72.1% on the validation set, respectively. This indicates that the use of LDH factor can improve predictability of the hazard score even when we are imputing LDH values for 598 of patients in training dataset.

20-fold cross-validation is applied to validate which the model is not overfitted. The robustness of the method on the validation set is assessed by calculating the average accuracy to predict hazard risk which is 74.25% with the variance of 2.77%. It can guarantee that overfitting is not happening and should not be considered as a major concern.

4.1.3 Discussion

Recently, many studies have attempted to extract the clinico-pathologic factors for better overall survival (OS) prediction of mCRPC patients. Unfortunately, there is great variance in reporting of effective clinico-pathological features for predicting OS of mCRPC patients in different studies. Discrepancy in data procurement may be due to different standards in recording the related information of target patients or disease progression level related to target patients. In order to assess, statistical significant clinical and pathological factors have been collected from various univariate and multivariate studies in the past three years, this information is depicted in Table 4.5. The blue and green colours are indicative of significant factors obtained from univariate and multivariate analyses, respectively. Furthermore, the yellow and orange colours are presenting non-significant univariate and multivariate analyses, respectively. The normalized and logarithmic clinical factors are shaded with left-handed and right-handed diagonals respectively. White colour indicated that either no analysis has been done in the given study or that specific feature were not available in the given study database.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Table 4.5: The prognostic factors with p_Value < 0.05 for CRPC patients reported in previous studies.

	Patient Characteristics		Prognostic factors																	
	Number of patients	Age median	ALP	AST	ACEI	AM	LNM	LM	BM	ALB	HB	ECOG	LDH	PSA	WBC	GS	PS	BMI	Age	
Nuhn et al., 2014	238	68.3	<div></div>	<div></div>			<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	
Chen et al., 2014	107	73.75						<div></div>	<div></div>		<div></div>			<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Koo et al., 2014	440	71.75	<div></div>						<div></div>		<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Rathkopf et al., 2014	1088	70.5	<div></div>						<div></div>		<div></div>	<div></div>	<div></div>	<div></div>	<div></div>				<div></div>	<div></div>
Halabi et al., 2014	1050	69	<div></div>	<div></div>			<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Zhao et al., 2015	278	72.1	<div></div>								<div></div>	<div></div>		<div></div>	<div></div>		<div></div>		<div></div>	<div></div>
Koo et al., 2015	248	69.3	<div></div>						<div></div>			<div></div>		<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>
Gravis et al., 2015	385	63	<div></div>						<div></div>		<div></div>		<div></div>	<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>
Chi at al., 2015	762	69	<div></div>					<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>		<div></div>		<div></div>	<div></div>
Uemura et al., 2016	41	73						<div></div>		<div></div>	<div></div>		<div></div>	<div></div>	<div></div>				<div></div>	<div></div>
Yamashita et al., 2016	79	72	<div></div>				<div></div>		<div></div>		<div></div>	<div></div>	<div></div>	<div></div>	<div></div>		<div></div>		<div></div>	<div></div>
Our study	1122	68	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>
Our study with LDH Imputation	1122	68	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>		<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>	<div></div>

Significant Univariable

Significant Multivariable

Non Significant Univariable

Non Significant Multivariable

No analysis or non available

Logarithm of clinical factor

Normalized clinical factor

Significant Univariable
 Significant Multivariable
 Non Significant Univariable
 Non Significant Multivariable
 No analysis or non available
 Logarithm of clinical factor
 Normalized clinical factor

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

The goal of our study was to increase the efficacy of hazard ratio estimation for mCRPC patients to provide an assured system for determining the best treatment methods. The inferred informative predictors from our study are consistent with the findings from recent studies [Nuhn *et al*, 2014, Chen, 2014, Koo *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Zhao *et al*, 2015, Koo *et al*, 2015, Gravis *et al*, 2015, Chi *et al*, 2015, Yamashita *et al*, 2016, Uemura *et al*, 2016]. As indicated in Table 4.5, the statistical effectiveness of ALP, ALB, HB, ECOG, and LDH on OS prediction are confirmed by the majority of the previously published methods [Nuhn *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Chi *et al*, 2015] further validating our inferred results. Specifically, [Rathkopf *et al*, 2014] have shown that log(ALP), HB, and log(LDH) are independent clinical features on multivariate analysis. In addition, [Halabi *et al*, 2014] and [Chi *et al*, 2015] reported that ALP, ALB, HB, ECOG, and LDH are associated with better OS prediction in univariate and multivariate analysis of mCRPC patients. Our method suggests that AST is a useful clinical feature in prognosis which is also confirmed in the [Nuhn *et al*, 2014] study .

In addition to significant clinical features, effectiveness of cancer metastasis is also useful in prediction of survival for mCRPC patients. As is the case with clinical features, selection of significant markers are varied amongst previous literature. Our study confirms the effect of Adrenal Metastasis (AM), Liver Metastasis (LM) and also lymph node (LN) tumor size on increasing hazard risk in both univariate and multivariate survival analyses. The lymph node metastasis has been reported by [Halabi *et al*, 2014] and bone metastasis is suggested by [Koo *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Gravis *et al*, 2015] as significant cancer progression features in increasing hazard risk in mCRPC patients. Also, liver metastasis is confirmed as a valuable factor in hazard prognosis by recent studies [Chi *et al*, 2015, Uemura *et al*, 2016].

The differences in findings are most likely related to the used protocol and missing data in various databases. For instance in our study, LDH factor which is missing for dataset D3 was a challenge that we have encountered and used the MICE imputation to predict them from the patterns of datasets D1, D2. Our study and other findings [Nuhn *et al*, 2014, Rathkopf *et al*, 2014, Halabi *et al*, 2014, Gravis *et al*, 2015, Chi *et al*, 2015, Yamashita *et al*, 2016] confirm the effectiveness of this marker on hazard risk analysis but it will not be really accurate by imputation.

4.1. A Hybrid Method for Estimating Overall Survival and Inferring Effective Prognosis Factors for Patients with mCRPC

Incomplete clinico-pathological datasets will continue to be a main challenge until we have access to richer and more standard datasets by golden collection systems.

The inferred results from our method is highly reliable, due to the large analysis sample of 1122 mCRPC patients treated with docetaxel therapy, this analysis has been internally validated using 470 mCRPC patients.

Our next interactive and visualizable system will demonstrate how systematic information interpretation could lead to actionable knowledge, which in turn would lead to better data collection, visualization tool and further enhancing the diagnosis and prognosis models.

The model that we developed was performed well as it has shown in the above sections in comparison with a recent publication in this domain and also the extracted effective factors from our method are compared with the other publications. There were many groups in DREAM challenge competition that they were looking at hybrid features. There is no truly way necessarily to construct those features that a lot of them depend on prior knowledge (some intuition about the nature of data that have been collected).

Determining of these features that lead to prediction improvement and it is also about sense making in the process, and supporting the understanding. The human role in that is a key. Therefore, that let us to build the system around how could support feature selection and more importantly feature construction and interactively feed these into the decision making process and test model. Again evidence from the community actually lead to a different approach. That is again brings up the idea of an interactive system to do this, and could evaluate how user's modification improves the results or share them with others.

The other step is if you could not create hybrid features then you've got an explosion of kind of sense of features that could be considered. This is really intractable with the original features set. Technically, we need the combination of human intelligent with machine intelligent. Hybrid features set might work well if some intuition is behind of selected features. Therefore, if you look at the original set of features alone you may not be able to see the impact of that feature by using the hybrid sense. There is a kind of selection phase for the way of getting prior

information that would be fit into this. We need a system in order to propose new features to lie down through this type of validations and that is where we need to combine human intelligence with machine intelligence. Consequently, it leads to our proposed system Tangible MultiVariate Visualization (TMVV).

4.2 Graspable Prognosis Factors Visualization for Interpreting Cancer Data

Various data sources have their unique structure that can be extremely informative when displayed by visualization techniques. The main challenge is finding representations that will effectively combine data to discover important features in modeling and prediction. Multivariate visualizations provide a multi-feature combinations environment to explore relationships between various features and the output of interest.

The goal of blending multiple features into a single visualization is to discover potential relationships that would be challenging to deduce from the constituent features. The design of a composed visualization depends on what the goal of survey is about. Take, for instance, a matrix containing 3 cancer trials of first line metastatic Hormone Refractory Prostate Cancer (HRPC) patients, where all patients received docetaxel treatment. The data include clinical covariates such as patient demographics, lesion measure, medical history, prior surgery and radiation, prior medicine, vital sign, etc. Although our interactive visualization is useful for following changes between two-covariate combinations, one of the challenges being addressed is how pair-wise comparison can be comfortably achieved for the user. To address this challenge and understand how changes in covariate combinations might be effective to classify an output, considering a network and replacing the nodes in the network with the covariates could be useful. This covariate-link network not only allows one to study the behavior of individual covariates in the contents of a network but also minimizes the user's need to remember effective combinations. In this way, one can repeatedly switch between the combinations to understand the differences in output classification. With all of this in mind, the user is able to

4.2. Graspable Prognosis Factors Visualization for Interpreting Cancer Data

analyze the statistical terms of each feature relevant to their research. The research rationale for this part of dissertation was built around an original database pertaining to prostate cancer. While this served as both the original database and template for visualization modeling, it is important to note that the theoretical concepts and implemented technologies can be translated across a multitude of research oriented fields. The proposed technology allows a user to analyze the statistical terms of each feature connected to an underlying database. After a feature is deemed significant by the user, the significance and hazard ratios of each feature (univariate analysis) and their combinations (multivariate analysis) can be presented. Univariate analysis can be used as a factor to determine if a feature is significant or not after a user has expressed interest in it. If the application is medical treatment, the user has this chance to use this manipulated model to determine hazard risk and decide about the proper treatment for a new patient.

Interactive technology that is deployable on various mediums provides an effective and accessible alternative for professionals seeking visually based research tools. There has been a minimal amount of studies dedicated to visualizing and interacting with big data on an accessible framework. The development of technologies dedicated to this field is important for the following reasons; data analysis should be an accessible and convenient process and researchers should have experience with various analytical tools so that they can confidently conclude the most appropriate strategy for addressing a given scenario. This concept also touches upon the fact that different sets of data have different forms which may be easily identified through visualization in comparison to traditional text or purely numerical displays. The technology developed here is accessible on a smart phone; arguably the most accessible and convenient medium available. Additionally, the technology has been implemented with a wide variety of visualizations that are intuitive and customizable. Both of the concerns leading to the rationale behind this research have been addressed in the implementation of the technology.

Assuming that the user is convinced of the advantage in visualizing their data, the research presented here seeks to grant the user with a greater degree of control and customization. With the use of TMVV and based on the output that the user is interested in, they can choose rele-

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

vant features, how those features are visualized and how they can be modeled in conjunction with one another to reveal an underlying theme. The advantage of this, and hence the rationale behind developing this technology, is that by incorporating multiple features into a single visualization, users are able to make important observations that would be more arduous to conclude via traditional methodologies. This part of dissertation will propose a design for a three phase system/toolkit that will approach data analysis as previously described. A graphical representation of the logic and architecture behind each phase in the toolkit can be seen in Figure 4.12.

The system architecture shown in Figure 4.12 is deployed within the typical framework for a web-based application. Users have access to the system with a web browser. While the technical parameters and system functionality will be explained in greater detail in later sections of this dissertation, Figure 4.12 gives a high level overview of the system using a unique UML model. The TMVV toolkit/application was designed with three phases that each serve a unique purpose and have a unique set of functions for the user. The purpose of each phase is iterated in each segment of Figure 4.12, while the user operations specific to a given phase are shown under the Operation Toolbox headings. Phase 4.12 is centered around modelling and visualizing data from an analytical point of view. If the user wishes to customize their analytical process to be more suitable to their particular data set, they would move to Phase 2 where the main purpose is refinement of methodologies. Finally, when users are satisfied with the setup of their software tools, they can move to Phase 3 where the main purpose is decision making based on the findings of the analysis. It is also worth mentioning that users focused on univariate analysis will be primarily using Phase 1, and in the event that multivariate analysis is desired Phase 2 will be used.

4.3 Materials and Design

The proposed technology can be used in the same manner as a smart phone application. After being installed on a device, the user can easily open the application and intuitively navigate through the series of menus and functions. The analytical functions are accessible when

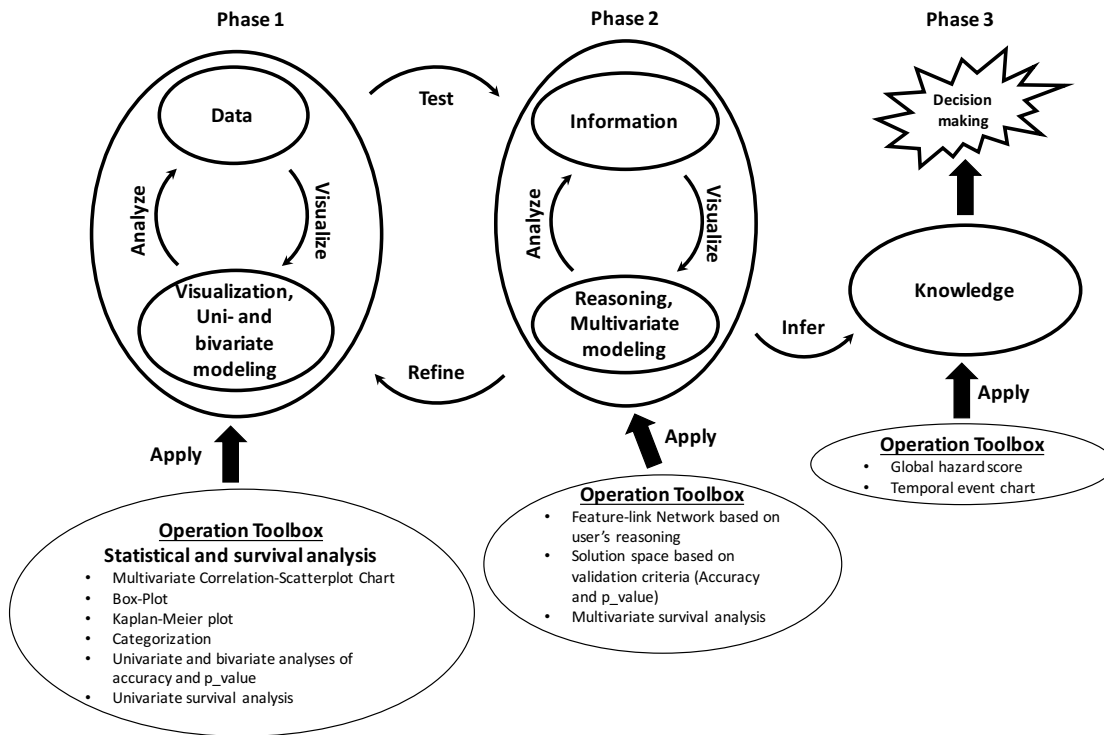


Figure 4.12: Flowchart illustrating the rationale architecture for TMVV research and design.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

the smartphone application interfaces with an interactive tabletop. At the moment there are two versions of the proposed system/application. There is a tangible version in which fiducials have been attached to smart phones that interact with MultiTaction screens in order to manipulate data and features. However, most users do not have access to the fiducial or multi-touch screen technologies. As a result, a desktop version has also been created. In the desktop version the smart phone is still used, however, there is no physical interaction between the smartphone and MultiTaction screens. This means that rather than using touch, the user uses a mouse. In the figures showcasing the implementation, the desktop version will be used. The implementation shown in the figures is the desktop version where the described menus and functions are accessed via HTML files. One of the focuses of this dissertation will be a breakdown of the design that has been implemented thus far and the usefulness of the various visual and analytical tools. By showcasing how the application looks and feels readers should gain a more concrete sense of the TMVV application's innovative approach and functionality. This is particularly important for the professionals who are looking for alternative forms of data analysis. Developers are not always aware of the important relationships that exist between various entities in a data set. As such, one of the overarching design goals was to hand over a plethora of tools over to the user who may find use for them. The TMVV application can be divided into three phases, each with their own set of navigations. The TMVV application can be thought of as a series of paths. Depending on the option that is selected in each menu the application will take you to a different part of the current phase or a new phase entirely. Flowcharts depicting the flow of data, Figure 4.13, 4.14, and 4.15, will serve to showcase the user process.

The analysis and numerical values behind the visualizations throughout the application are based on underlying datasets that the user will have the option to incorporate into the application as needed. The figures depicted throughout the description of the TMVV application are intended to showcase the analytical tools that have been developed and the resulting visualizations are the result of sample database. For the purpose of visualizing data and showcasing the

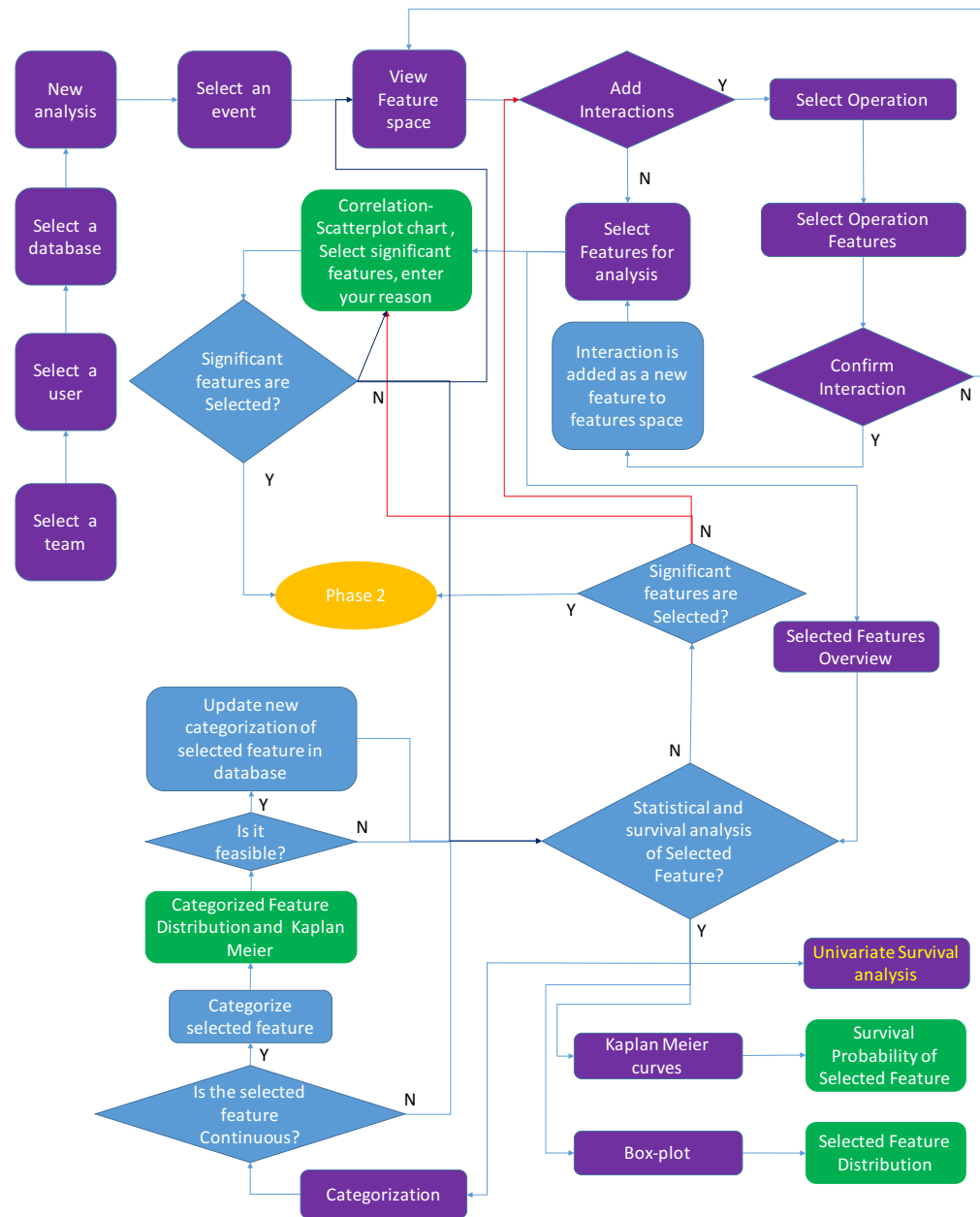


Figure 4.13: Phase 1 flowchart. Directed arrows show the interaction path of the user through Phase 1 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters a new phase of the TMVV platform.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

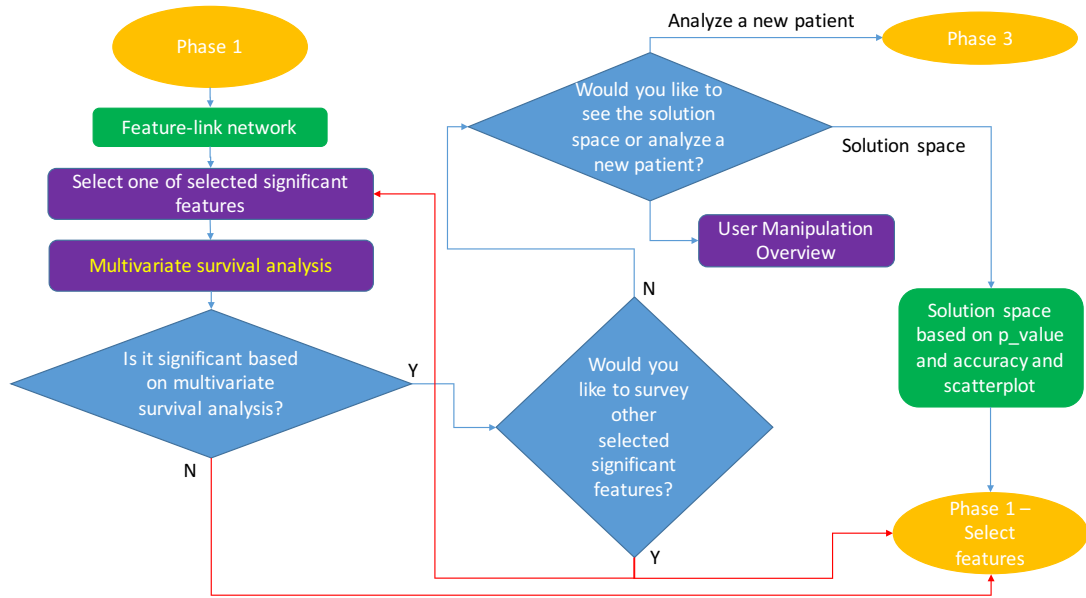


Figure 4.14: Phase 2 flowchart. Directed arrows show the interaction path of the user through Phase 2 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters or returns to a new phase of the TMVV platform.

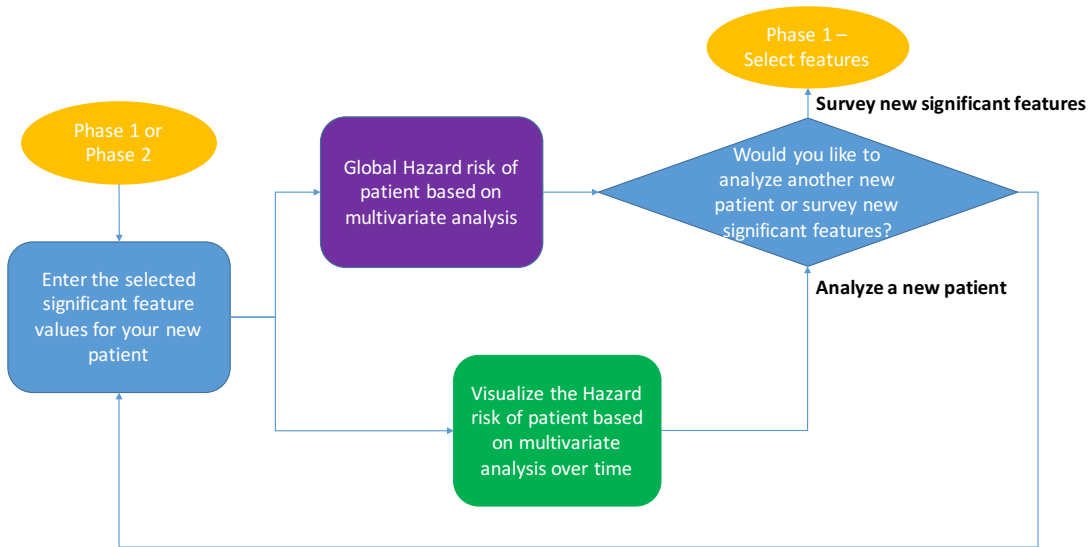


Figure 4.15: Phase 3 flowchart. Directed arrows show the interaction path of the user through Phase 3 of the TMVV platform. The fill color signifies the surface of interaction and/or visualization, where purple represents interactions and/or visualization which occur on the smartphone, green represents interactions and/or visualization on the tabletop, and blue represents one of two things: actions which neither occur on the smartphone and tabletop, or the actions which occur internally in the TMVV system. The shape signifies the type of user interaction: rectangles represent a process, action, or visualization, diamonds represent a decision the user must make, and an oval represents when a user enters or returns to a new phase of the TMVV platform.

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

capabilities of the TMVV application, a sample prostate cancer database was used. Readers should keep in mind that the same tools and principles can be applied to databases pertaining to their own research. While beyond the scope of this article, a tutorial explaining how to structure and then infuse a personal database with the TMVV application will be available. An explanation of the system architecture begins below with a breakdown of the data that interfaces with the TMVV application.

4.3.1 Data Description

Input Data Format

The input database format follows a right censored format. With this in mind, when a user wishes to perform analysis with unique data, they should keep this format in mind.

Input Data Structure and Personal Data Usage

The following points iterate the structure of the input data and how a user may go about generating input files using their own data (possibly patient related).

- Feature dataset that includes the clinical and pathological features of all patients. Rows are assigned to patients with different ID and columns show the features (Table 4.6).
- Censoring dataset that includes the censoring information of the events and also time to events. Rows show the patients ID and columns represents censoring information and time to events (Table 4.7).
- Normal range information for continuous clinical features: it includes two rows (the first one is assigned to lower bound and second one is related to upper bound of each feature) and the columns show all features (Table 4.8). If there is no defined normal range or it is not available for some features, user should consider zero values for those specific features.

Table 4.6: Feature dataset that includes the clinical and pathological features of all patients. Each row is assigned to a single patient and each column to a single feature.

Patients	Feature 1	Feature 2	Feature 3	...	Feature k
ID_1	F_1^{ID1}	F_2^{ID1}	F_3^{ID1}	...	F_k^{ID1}
ID_2	F_1^{ID2}	F_2^{ID2}	F_3^{ID2}	...	F_k^{ID2}
ID_3	F_1^{ID3}	F_2^{ID3}	F_3^{ID3}	...	F_k^{ID3}
.
.
.
ID_n	F_1^{IDn}	F_2^{IDn}	F_3^{IDn}	...	F_k^{IDn}

- Also, there is a file with the name of features pre-clustering that classifies the features in different categories (e.g. lab value, lesion measure, etc.) (Table 4.9).

Prostate Cancer Database Explanation

To better understand the concepts and visualizations presented throughout the dissertation for TMVV functionality, an explanation of the sample prostate cancer database is in order. The prostate cancer database contains 3 trials of first line metastatic Hormone Refractory Prostate Cancer (HRPC) patients who all received docetaxel treatment. The data also includes clinical covariates such as patient demographics, lesion measurements, medical history, prior surgery and radiation and vital signs. More details can be found in the information from the website <<https://www.synapse.org/#!/Synapse:syn2813558/wiki/209583>>.

4.3.2 Analysis, Modeling and Visualization

The toolkit based on the proposed design has been implemented with server and client side JavaScript. The source can be found via GitHub or on the Synlab website for insight into

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

Table 4.7: Censoring dataset that includes the censoring information of the events and also time to events. Rows show the patients ID and columns represents censoring information and time to events.

Patients	Censoring information related to event 1	Time to event 1	Censoring information related to event 2	Time to event 2	...	Censoring information related to event m	Time to event m
ID_1	C_1^{ID1}	T_1^{ID1}	C_2^{ID1}	T_2^{ID1}	...	C_m^{ID1}	T_m^{ID1}
ID_2	C_1^{ID2}	T_1^{ID2}	C_2^{ID2}	T_2^{ID2}	...	C_m^{ID2}	T_m^{ID2}
ID_3	C_1^{ID3}	T_1^{ID3}	C_2^{ID3}	T_2^{ID3}	...	C_m^{ID3}	T_m^{ID3}
.
.
.
ID_n	C_1^{IDn}	T_1^{IDn}	C_2^{IDn}	T_2^{IDn}	...	C_m^{IDn}	T_m^{IDn}

Table 4.8: Normal range information for clinical features. This dataset includes two rows and the columns show all features. If there is no defined normal range or it is not available for some features, user should consider zero values for those specific features.

Feature 1	Feature 2	Feature 3	...	Feature k
L_1	L_2	L_3	...	L_k
U_1	U_2	U_3	...	U_k

$$L_i = \begin{cases} b & \text{if we have a binary feature} \\ c & \text{if we have a categorical feature} \\ \text{Lower bound of normal range} & \text{if we have a continuous feature} \end{cases}$$

$$U_i = \begin{cases} b & \text{if we have a binary feature} \\ c & \text{if we have a categorical feature} \\ \text{Upper bound of normal range} & \text{if we have a continuous feature} \end{cases}$$

Table 4.9: Features pre-clustering information that classifies the features in different categories (e.g. lab value, lesion measure, etc.). This dataset consists of two rows. The first row lists all the features present in the feature dataset, and the second row contains the pre-clustering label for each feature. If the user does not define any pre-clustering labels, those features will be considered individually as single un-clustered features.

	Feature 1	Feature 2	Feature 3	...	Feature k
Pre-Cluster labels	PC_1	PC_2	PC_3	...	PC_k

personal implementation. As previously described, the toolkit is divided into a three phase process; each of which will be described next in greater detail. Phase 1 primarily deals with the statistical analysis and visualization of data based on the chosen underlying database. Phases 2 and 3 are more abstract in the sense that they deal with the customization of analytical tools. These are particularly important for making the method of analysis suitable to the type of data being used or the nature of the subject matter. Flowcharts depicting the user experience for Phases 1, 2 and 3 are shown in the Figures 4.13, 4.14, and 4.15 and are good tools for guiding one's train of thought in understanding the TMVV functionality.

Phase 1 – Data Statistical Analysis and Visualization

The various menus and options available in Phase 1 of the TMVV application can be seen in Figure 4.16. Phase 1 is focused on data analysis and visualization through the use of scatter-plots, box plots, Kaplan-Meier's plots and univariate survival analysis.

A systematic breakdown of the various user options and functions available in Phase 1 can be seen in Table 4.10. The various visualizations and analytical tools accessible through Phase 1 will also be described in detail. However, to see how the application progresses depending on the chosen option, readers should refer to Figure 4.16. A more detailed discussion of each menu in Phase 1 functionality can be read following Table 4.10.

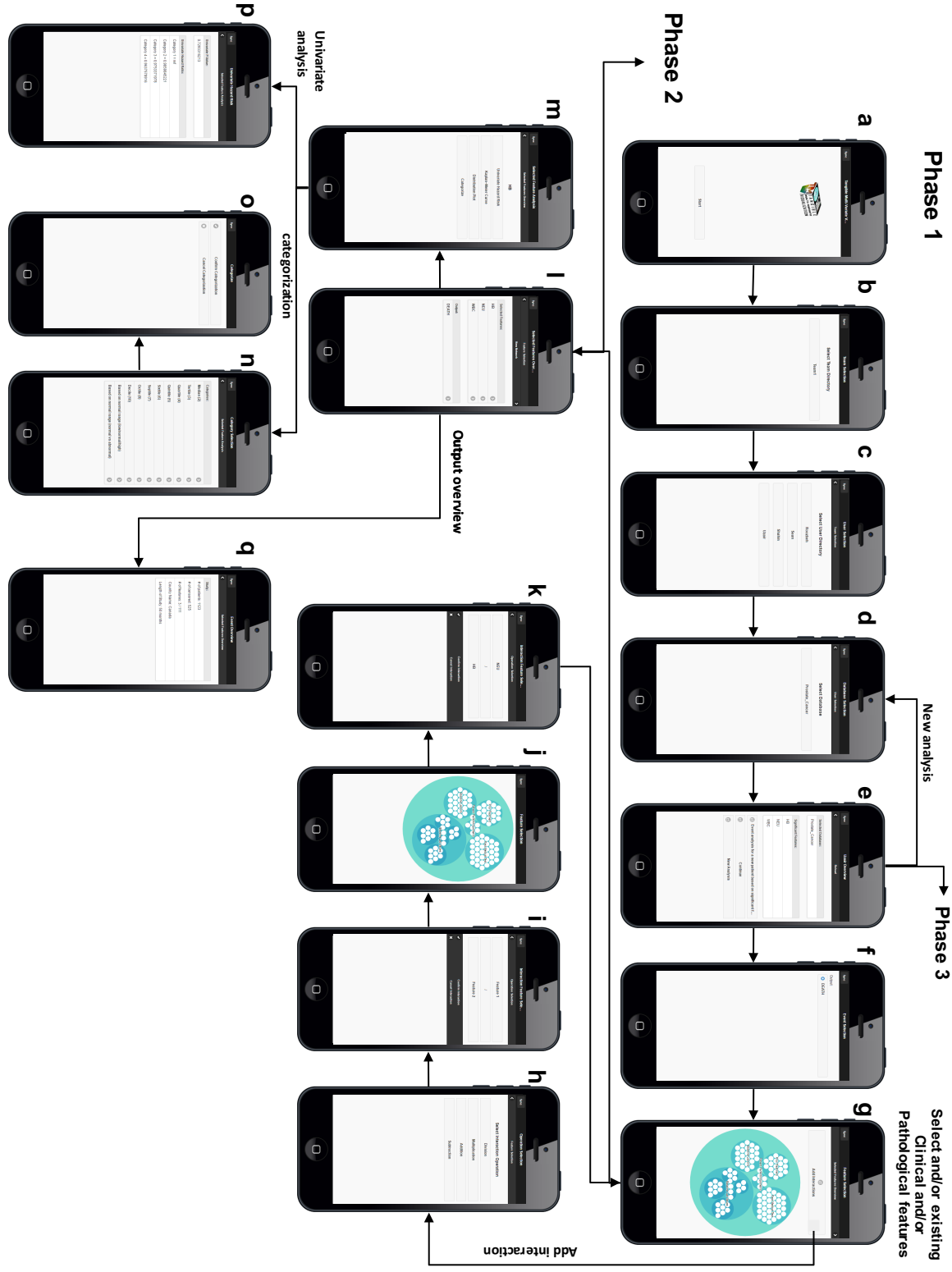


Figure 4.16: Visual depiction of the user interfaces (menus) that the user will be able to interact with within Phase 1 of the TMV application.

Table 4.10: Systematic breakdown of the user experience in Phase 1 of the TMVV application.

Menu(s)	Interaction	Description
a, b, c, d	User Configuration	Select team folder, user folder, and Setup and configure the underlying database for analysis and visualization.
e	Proceed to Analysis or Configure New Analysis	If the user has used the TMVV application before, are they continuing with their previous analysis or configuring a new analysis?
f	Event Selection	Select events corresponding to the underlying chosen database.
g	Clinicopathological Factor Selection	The first visualization schema unique to the TMVV application. Zoomable platform presenting users with factor selection.
h, i, j, k	Create New Feature Interaction	Series of menus dealing with the configuration of interactions between features specific to the underlying dataset.
l, m, n, o, p, q	Analytics	Series of menus dealing with the analysis of data based on the underlying dataset, selected events and configured interactions between features.

User Configuration Menu

When the TMVV application is initiated the first three menu asks the user to select their team and user folders. The system determines if the user is a new user or a returning user by examining user specific folder.

As previously stated additional databases can be integrated into the TMVV application. The database that is chosen will provide the underlying numerical data that will produce modeling and visualizations during other portions of the application. The users will be able to perform univariate and multivariate analysis to evaluate the significance of their selected features. This operation as seen in the desktop implementation is shown in Figure 4.17.

Proceed to Analysis or Configure New Analysis Menu

In menu e, the user will be presented with the options to perform New Patient Analysis, Continue Analysis and New Analysis. The descriptions of these options are seen here.

- **New Patient Analysis** – This option will lead to Phase 3, the focus of which is on decision



Figure 4.17: Menu d of Figure 4.12. The user is being asked to select the underlying database (prostate or breast cancer). The local host is representative of the interactive tabletop which is blank at this stage because no analysis is yet to be performed.

making and will be discussed later.

- **Continue Analysis** - In the case where the user has previously used the application they will have the option to continue progressing through the application with the options that have been previously configured. Users selecting this option will be able to see a list on the screen of their smartphone showing the underlying dataset as well as the features that were selected as significant features at previous system's manipulation.
- **New Analysis** - In the case where the user is new to the system or a previous user desires to conduct a new analysis, they will have the option to perform a new analysis altogether.

Event Selection Menu

The user will be able to view a list of events that correspond with the underlying dataset. For example, in relation to the database for prostate cancer the events include data on death rates, disease reoccurrence and discharges from hospitals. Event selection perhaps gives the best understanding of the advantages of the TMVV application. Each data set will grant the user

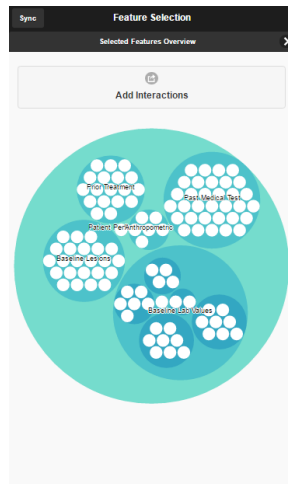


Figure 4.18: Pre-clustered zoomable feature selection platform.

access to a variety of related events that serve as a building block for custom visualization and analytical tools.

Clinicopathological Factor Selection Menu

This menu presents users with the first visualization schema unique to the TMVV application. During consultation with working physicians it was suggested that related features be pre-clustered on a zoomable platform. This visualization can be seen in Figure 4.18. Each individual unit corresponds to a feature that the user has the option of performing visual analysis with. However, only features that have been configured by the user through the application will respond to interaction. At this stage, the user will generally be presented with two options; the first is to add an interaction between two of the selected features, the second is to select one of the individual unit's active within the visualization and initiate the analytical portion of the TMVV application.

Create New Feature Interaction Menus

In the event that the user selects "Add Interaction" in menu g of Figure 4.16 they will have

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

initiated the portion of the TMVV application dedicated to aggregating the analysis of custom features. The features available for selection in menu g of Figure 4.16 are based on an event related to a selected underlying database. Sometimes it is useful for the user to be able to create a new feature that is built around the relationship between two features. The addition of interactions to the TMVV application allows users to survey the effect of new interactions obtained from applying the different operations on the original features from a given event. This addresses the concern that developers are not always aware of the importance of certain interactions between features. In this case, the consumer are professionals involved in research dealing with large sets of data. While the connections between different features may not be obvious to developers, it is important to make tools that address these relationships available to the users who may find importance in them. Opting to add a feature interaction will prompt the user through menus h to k as shown in Figure 4.16.

Menu h in Figure 4.16 lists a variety of operations that the user is able to perform between two features. The operations that have been included in this implementation of the TMVV application include multiplication, division, addition and subtraction. To reiterate, while the need to perform these mathematical operations between features may not be clear to developers, researchers may find usefulness in finding summations, ratios and so forth.

Following the selection of an operation the user is directed to menu i of Figure 4.16. In this menu, the user selects the two features that the operation will be performed with. In the example shown in Figure 4.16, the user is dividing feature 1 by feature 4.16.

Finally, as seen in menu k of Figure 4.16, the user will have the option to add this custom feature/interaction to their personal database. In other words, the interaction between two features will now exist as its own unique feature which is the result of whichever operation was performed. When the interaction is confirmed it will be added to the cluster mechanized visualization of existing clinical and pathological features as shown in menu g of Figure 4.16. Whether the user opt to add this feature to their analytical pool, or if the user is no longer interested in the relationship that was configured between two features they will return back to menu g of Figure 4.16.

Analytics Menus

In the event that the user selects one of the configured features in menu g of Figure 4.16 the TMVV application will be routed to menu l. A notable feature of menu l is that it shows the event related to the feature that was selected, in addition to other features associated with that event. Features that are independent are visualized as "Features" while features that are the result of interactions are visualized as "Interactions". For example, in menu l of Figure 4.16, the selected features are entitled Feature 1, Feature 2, and Interaction 1, to account for the independent entities as well as the interaction that was configured in menu i of Figure 4.16. One can also see that the event that all the selected features are related to is entitled "Death". At this point, the user has the option to go to the feature network (Phase 2), perform analytics (menus l through p), or view the parameters of the event. Phase 2 will be described later and explains the feature selection process as only database and event selection has been covered thus far. By selecting the event the user is brought to menu q of Figure 4.16. This menu is static and simply displays the parameters of the event associated with the selected features. In the example shown in Figure 4.16, the event is related to the prostate cancer database and thus deals with patient oriented information. Some of the event study parameters include the number of patients, features and county names while the length of the study is indicated to be fifty months. The defining design element of the TMVV application are the visual analytical tools which are accessed and customized in menus l through p. The visualizations are only viewable when the smartphone is interfacing with an interactive tabletop platform.

The user will select the features they wish to display in menu l of Figure 4.16 and then be directed to menus m through p of Figure 4.16 where they will have the ability to select their visualization preference and categorization.

The proposed TMVV application/toolkit currently visualizes patient data and designed features using a variety of plots that were briefly described in the "Analytics" section of Table 4.10.

Kaplan-Meier Plots and Box Plots

The intent behind the visualizations is to be able to quickly and intuitively identify outliers or

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

other relevant parameters in a data set. Kaplan-Meier curves and Box-Plots were identified as suitable tools for meeting this requirement and can be seen in Figures 4.19 and 4.20. These figures show how the Kaplan-Meier curves and box plots are visualized when the TMVV application interfaces with an interactive tabletop. Figure 4.19 shows the default quartile form for plotting box-plots and Kaplan-Meier graphs. Menu n of Figure 4.16 also shows the option to perform different forms of analysis (i.e. – median, quartile. . . etc.) which is represented by Figure 4.20.

Multivariate Correlation-Scatterplot Charts

Imagine that you are interested in studying clinical and/or pathological patterns in patients with advanced prostate cancer. A big database from 1122 patients is collected which include 110 clinicopathological features. This data is entered in a file with .csv format which can be overwhelming to extract information and track patterns from. Therefore, we create a multivariate correlation-scatterplot chart to see how the features relate to each other. Half of this chart is plotted by scatterplot blocks and half of it represents the correlation between various features. Scatterplots are ideal candidates for this type of data because quantitative features can be plotted on both the vertical and horizontal axis. Each patient appears as a point on the graph while the scatterplot gives a visual representation of the feature's relationship. Another expression for showing this relationship is the correlation that is graphically plotted as a correlation matrix. The scatterplots and correlation matrix are symmetric in the data they represent and therefore can be visualized on a single platform where the related pairs are symmetrical to one another across a diagonal. This concept is visually realized in Figure 4.21.

All correlational studies have three properties. One of them is direction, another one is strength, and the last one is form, all of which are helpful to investigate the relationship between features. In the representation shown in Figure 4.22, we are covering the direction by demonstration of scatterplots in the lower triangular segment and showing the strength by correlation matrix in the upper triangular segment. The direction of the correlation is determined by positive

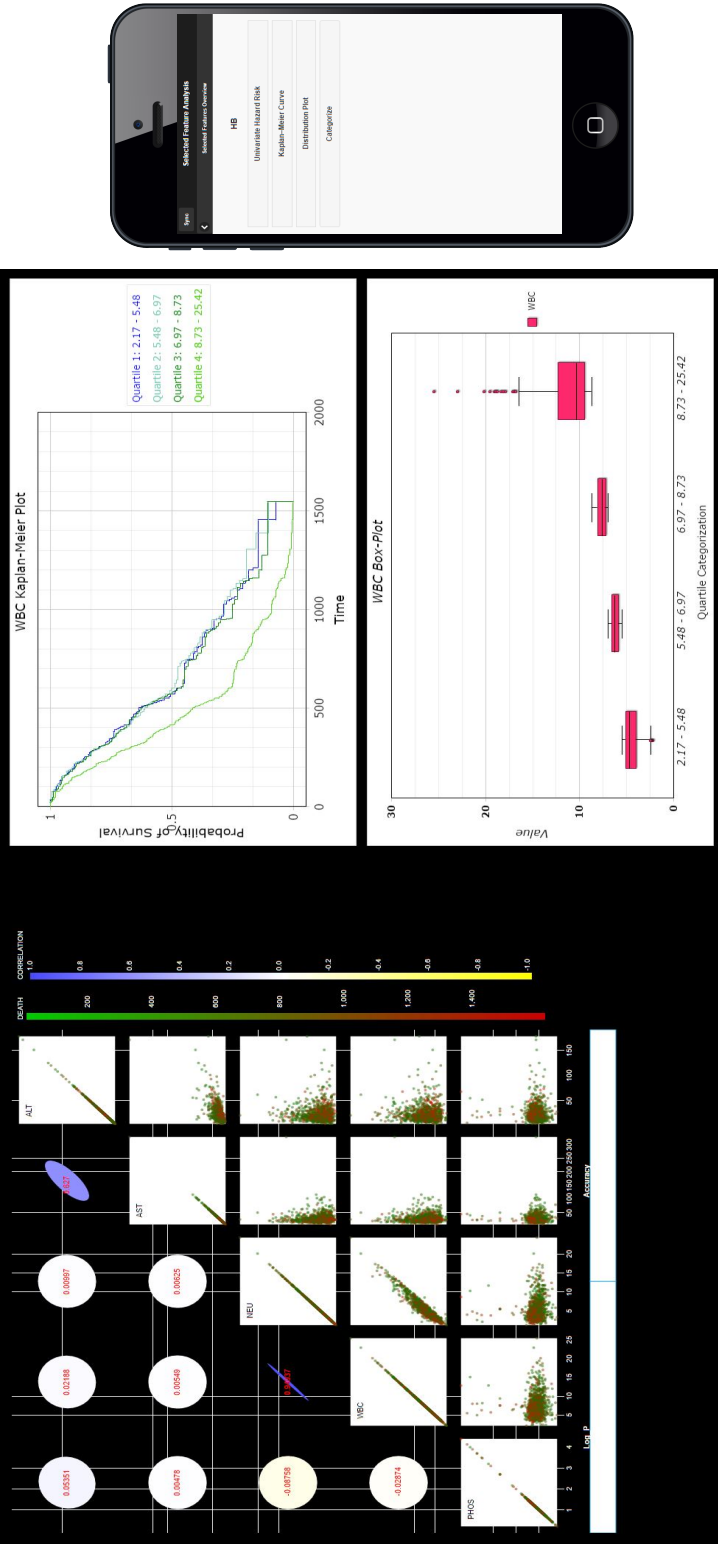


Figure 4.19: The various analytical tools are shown on the application interface. The implemented Kaplan-Meier plot and Box-Plot are shown. Note the easily identifiable outliers on the Box-Plot.

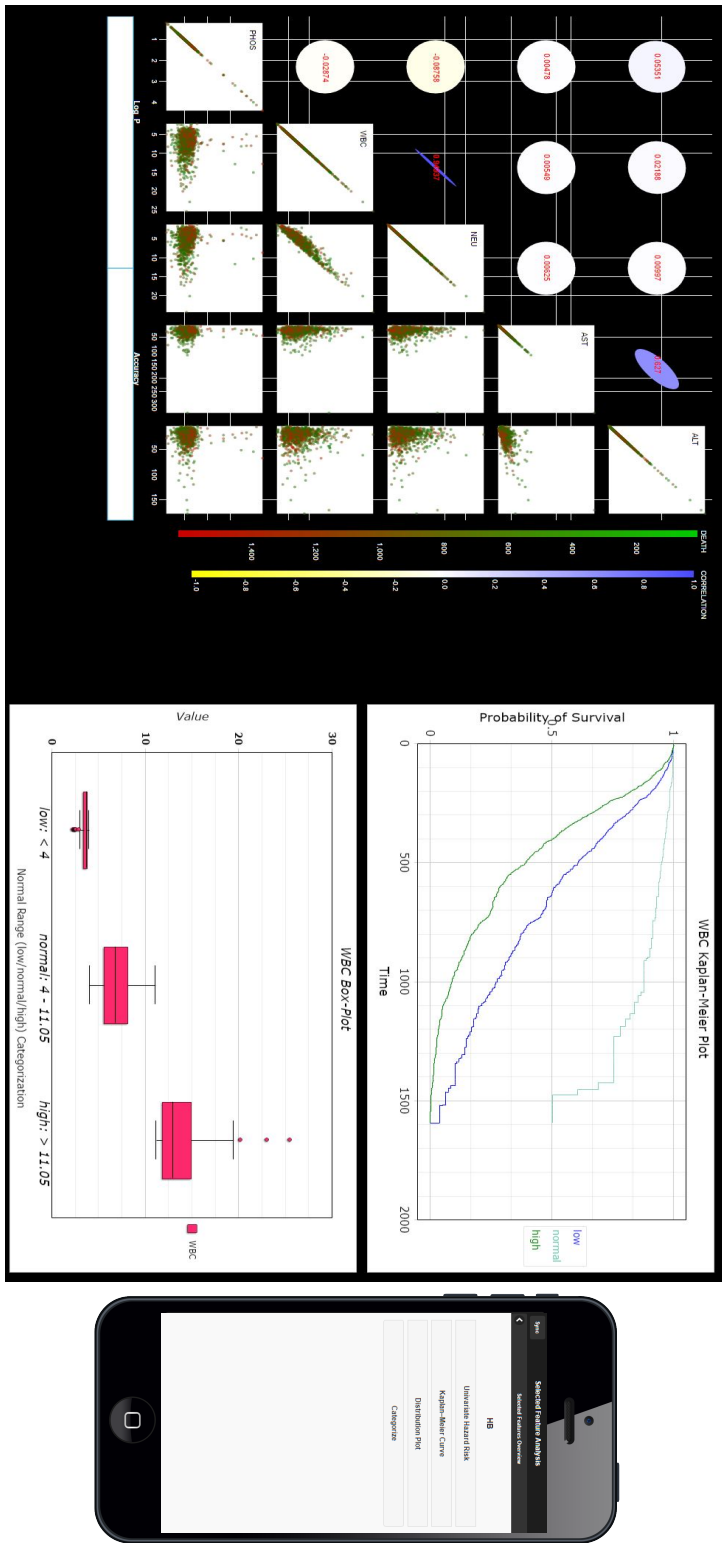


Figure 4.20: Implemented Kaplan-Meier plot and Box-Plot following categorization by normal range. Alternative categorizations are shown in menu I of Figure 4.16.

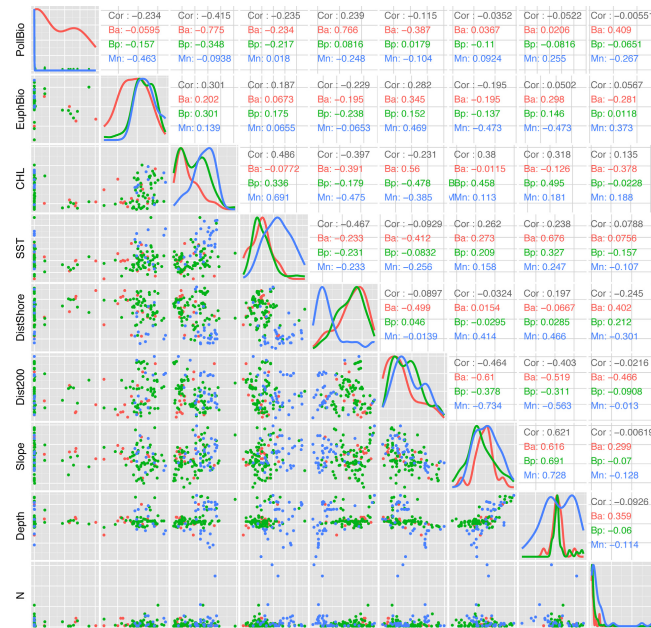


Figure 4.21: An example for multivariate correlation scatterplot [Zerbini, Alexandre N., et al. "Baleen whale abundance and distribution in relation to environmental variables and prey density in the Eastern Bering Sea." Deep Sea Research Part II: Topical Studies in Oceanography (2015)].

Chapter 4. Effective Prognosis Factors Inference in Prostate Cancer and Tangible MultiVariate Visualization

and negative numbers. In positive correlation, when a feature is increasing, another feature is increasing too. For negative correlation, when a feature is increasing, another feature is decreasing. We are visualizing the positive correlation with blue circles and negative ones with yellow circles. The second aforementioned property is strength, which is a crucial piece of information. The correlation measures the strength of linear relationship and is always between -1 and 1. When we want to show the strength between two features, there are two general approaches that can be taken. The first is to use different circle sizes with larger circles pertaining to a higher degree of strength. The second is to use color intensity with stronger shades of yellow or blue pertaining to a higher degree of strength. As an example, the correlation of 1 is a perfect correlation and in reverse the correlation of 0 or close to 0 is showing no relationship or very weak linear relationship between two features. In other words, the correlation with the highest numerical value is the strongest. We should note that no causation evidence is provided by correlations. Rather, it is just indicating evidence of association.

It should be considered that the correlation value is meaningful if the relationship between two features is linear. Therefore, the form property becomes important and is provided for the user through the scatterplots. For instance, one might come across a curved form rather than a linear line. This can occur when a feature does not increase at a constant rate or fluctuates between periods of increase and decrease. The implantation of the multivariate scatterplot in the TMVV application can be seen in Figures 4.22 and 4.23. Figure 4.23 also demonstrates the application's ability to obtain p-value and accuracy from the modeling process. Underneath the scatterplot the user has the option to view p-value or accuracy for any selected combination.

Univariate Feature Analysis

In addition to statistical visualizations, the user also has the opportunity from menu m of Figure 4.16 to analyze the univariate survival analysis. The univariate analysis provides the

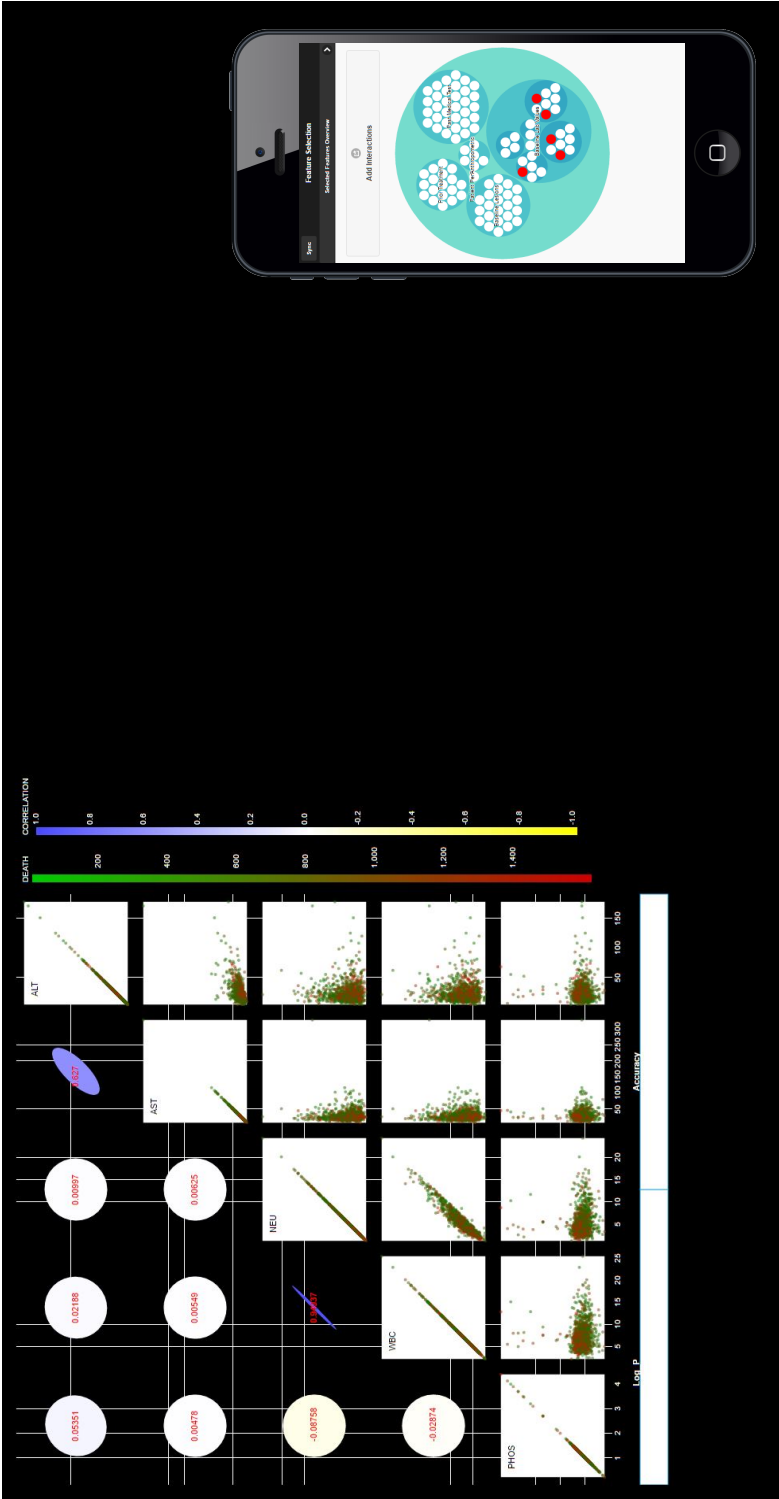


Figure 4.22: Right- Implemented cluster visualization of related features; Left- Multivariate correlation scatterplot.

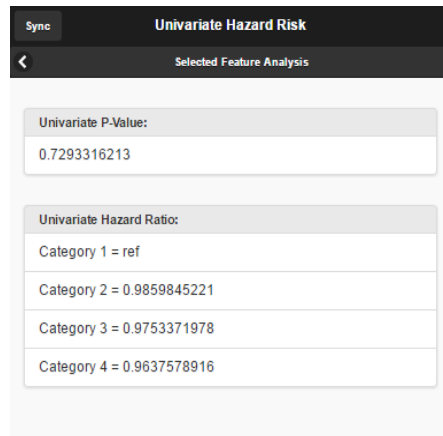


Figure 4.24: Univariate p-value and hazard ratio for a single feature.

p-values of the univariate model for each individual feature. The results related to hazard ratios are obtained according to the selected category by the user. For example, when we have four groups (quartile categorization), one group is considered as a reference in the modeling and the other groups' hazard ratios are obtained based on the reference. It is worth noting that categorization between binary features holds no meaning. As a result, for a binary feature, one group which has the value of 0 is considered as a reference and the hazard ratio of the other group with value 1 is determined based on the first group. Similar is the case for categorical features. Figure 4.24 demonstrates the outcome of the univariate survival analysis for a single feature.

The user has the opportunity to use all of the statistical analysis presented in the models to make a decision regarding the significance of features or the bivariate combinations of selected features. This is done by clicking on a block of the correlation scatterplot. When the user selects a block, a menu appears to record why the user made the selection in the first place. By doing this, the application keeps a digital mental map of the users thought and analytical process. This is useful for recalling why decisions were made at an earlier point in the modeling process.

Phase 2 – Information Analysis and Visualization

When the user selects "Feature Network" in menu i of Figure 4.16 they are sent to Phase 2 of the TMVV application. Phase 2 can be seen in Figure 4.25. Phase 2 is centered on the statistical parameters of the selected features. Phase 2 also includes multivariate survival results based on each of the features in menu a of Figure 4.25 when we are operating within the contexts of an underlying prostate cancer database. These results include the p-values of the multivariate model for each individual feature in addition to the selected set of features. In the example shown in Figure 4.25, these p-values are primarily associated with a predication for survival time for the patients. The results related to hazard ratios are obtained according to the selected categories by the user in Phase 1. One group is used as a reference and the hazard ratios for the other groups are obtained through this reference (the user has the ability to change the categorization for each continuous feature in phase 1 and the methodology with provide a comparison of hazard ratios from the previous iteration in menu b of Phase 2). It can be mentioned that a categorization between binary features is meaningless. For example, when we use lymph node metastasis as a pathological feature, patients with no metastasis are used as the reference to find the hazard ratio patients with this condition. The user is then presented with a final opportunity to check the significance and effectiveness of each feature individually and also as a member of a selected set before saving the result of the manipulation. This can be seen in menus a and b of Phase 2 in Figure 4.25. Computations are carried out through MATLAB and R programming.

When the smartphone is interfacing with an interactive tabletop platform, the feature interaction network is also visualized on the tabletop platform. The feature interaction network can be seen in Figure 4.26. In the context of the prostate cancer database, each node in the interaction network is representing a feature. The selection of a combination of features as a significant factor by the user manifests itself as an edge in the interaction network. In this manner, the user is able to visually gauge each significant factor or combinations of significant factors. If the user hovers over a node of the feature network, the rationale or reasons for



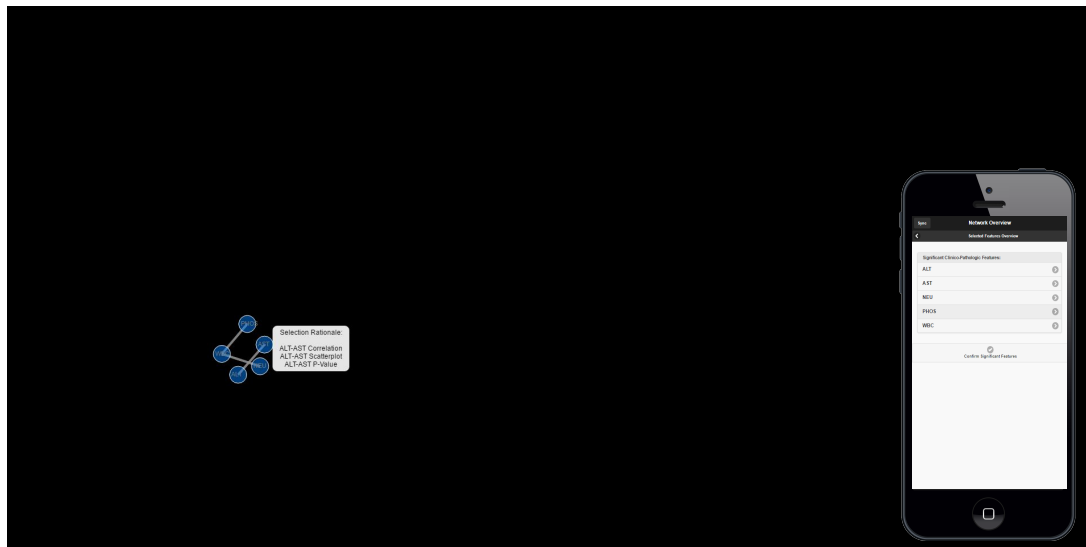


Figure 4.26: Hovering over the feature interaction network reveals the selection rationale for each feature node.

selecting this feature or combination of features is shown. This serves as a very powerful tool for tracking the thinking path of the user over time.

The user has the opportunity to use this additional statistical analysis presented in phase 2 to solidify his decision regarding the significance of features or the bivariate combinations of selected features. When the user is confident and confirms the significant features in menu a of Figure 4.25, the user is presented with three options shown in menu c of Figure 4.25 Iteration Summary, New Patient Event Analysis, and Run Another Iteration. The descriptions of these options are seen here.

- **Iteration Summary** – This option will return the user to menu e of Figure 4.16 from where the user can proceed to Phase 1, Phase 3, or choose to end the current session by closing the browser window.
- **New Patient Event Analysis** - This option will lead to Phase 3, the focus of which is on

decision making.

- **Run Another Iteration** - In the case where the user wishes to continue and perform another iteration of feature analysis. For example, in the first iteration the user chooses 4 features and at the end of Phase 2 they decide to analyze additional features. Users selecting this menu will be able to visualize the feature solution space. The data processing in this option is best explained through the use of an example. Imagine that there are 110 features which would result in a $110 * 110$ solution space matrix. Axes show each feature and each value in this matrix shows the p-value or accuracy for all bivariate feature combinations. The solution space matrix can be visualized based on accuracy values or p-values by simply selecting the solution space. This methodology shows the user the best combinations between selected features and allows them instantly spot correlations. This visualization also gives the user a tactic regarding which features to remove and how to select additional features for effective results if desired. Also, the user is given the chance to see the scatter plot of a selected feature in the right hand box. Figure 4.27 and 4.28 show the feature solution space in the context of the underlying prostate cancer database.

Alongside the solution space matrix, users are also presented with a visual summary of all previous iterations. Figure 4.29 shows the feature iteration summary, which highlights the features deemed significant by the user in each iteration. In this manner the user can recall his previous iterations, and strengthen their tactic regarding which features to select or remove to achieve effective results.

Phase 3 - Knowledge Analysis and Decision Making

When the user selects "Event Analysis for a New Patient Based on Selected Factors" in menu c of Figure 4.16 they are sent to Phase 3 of the TMVV application. Phase 3 can be seen in Figure 4.30. Phase 3 is centered on event analysis by providing patient and feature menus

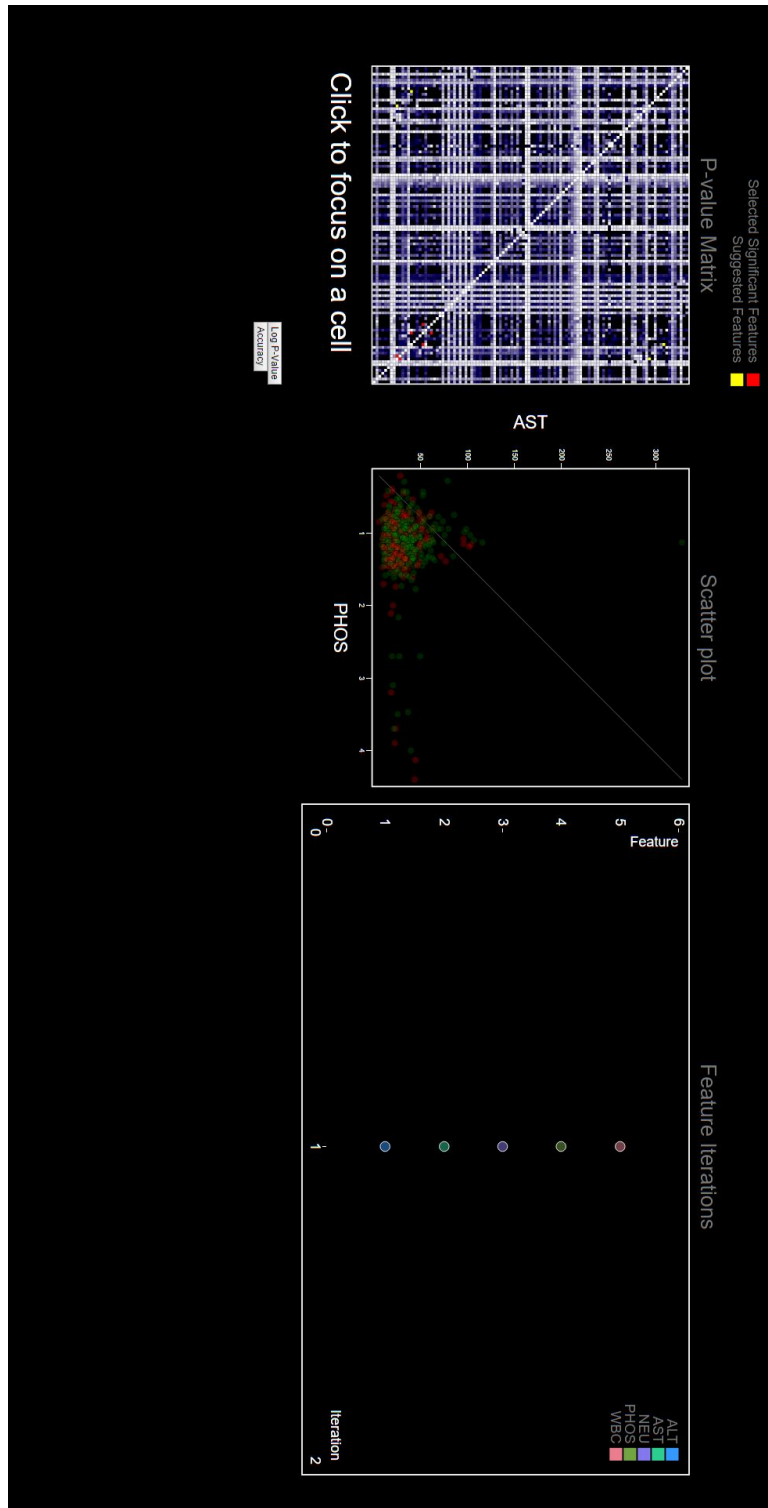


Figure 4.27: Right - Feature iteration summary; Left - Feature solution space p-value matrix. Notice the selected features as small dots on the solution space.

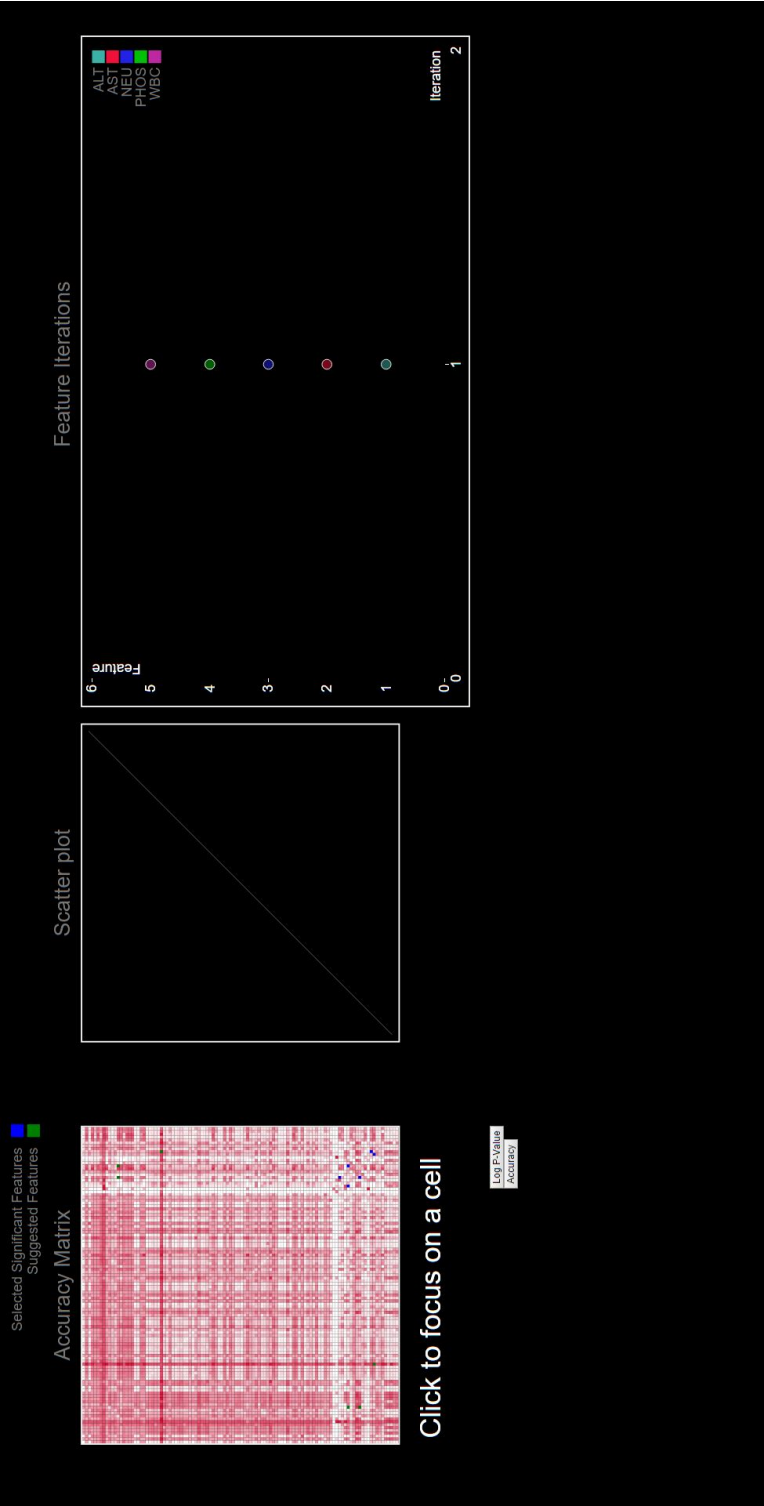


Figure 4.28: Right - Feature iteration summary; Left - Feature solution space accuracy matrix. Notice the selected features as small dots on the solution space.

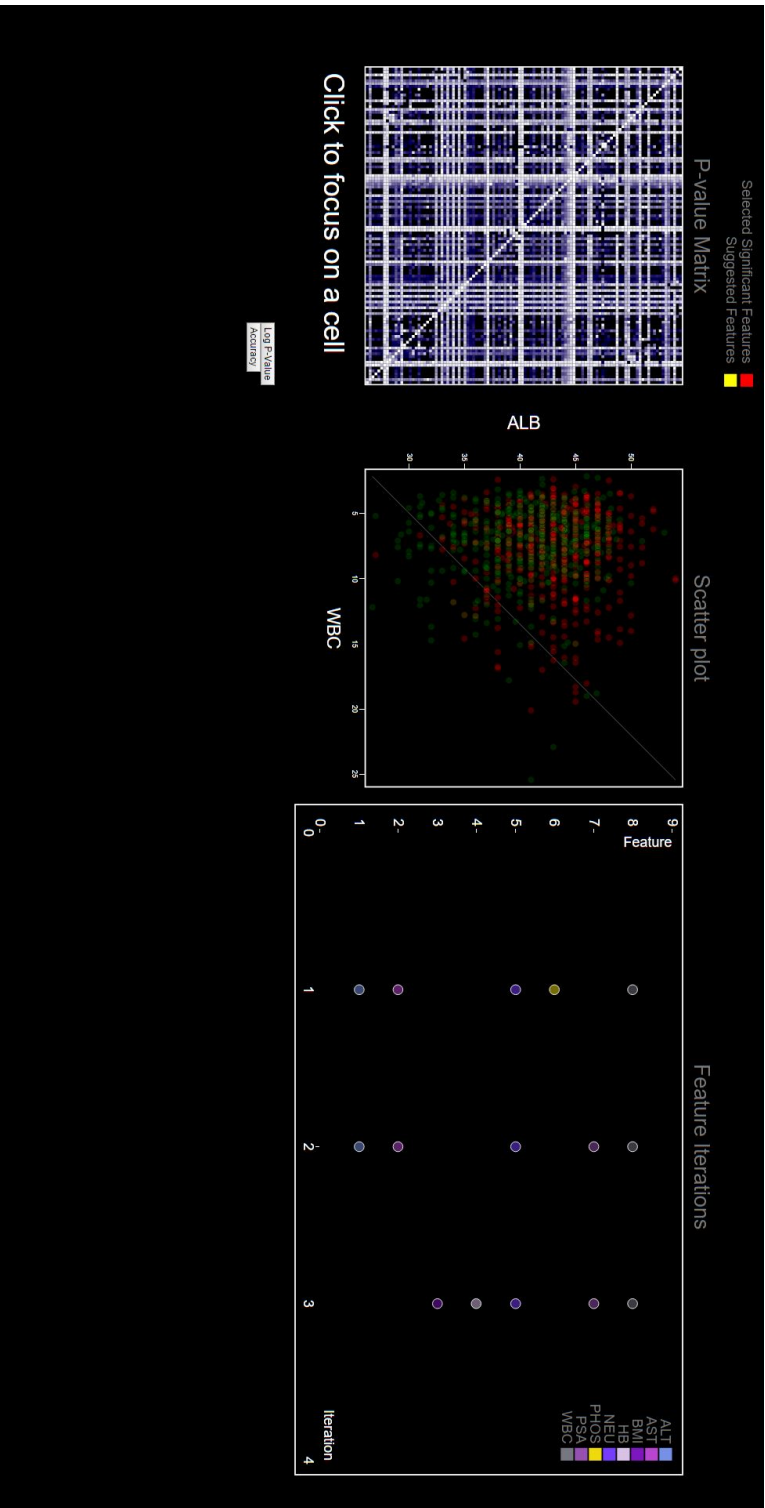


Figure 4.29: Left - Feature solution space p-value matrix; Right - Feature iteration summary. Notice the variation of significant features over multiple iterations.

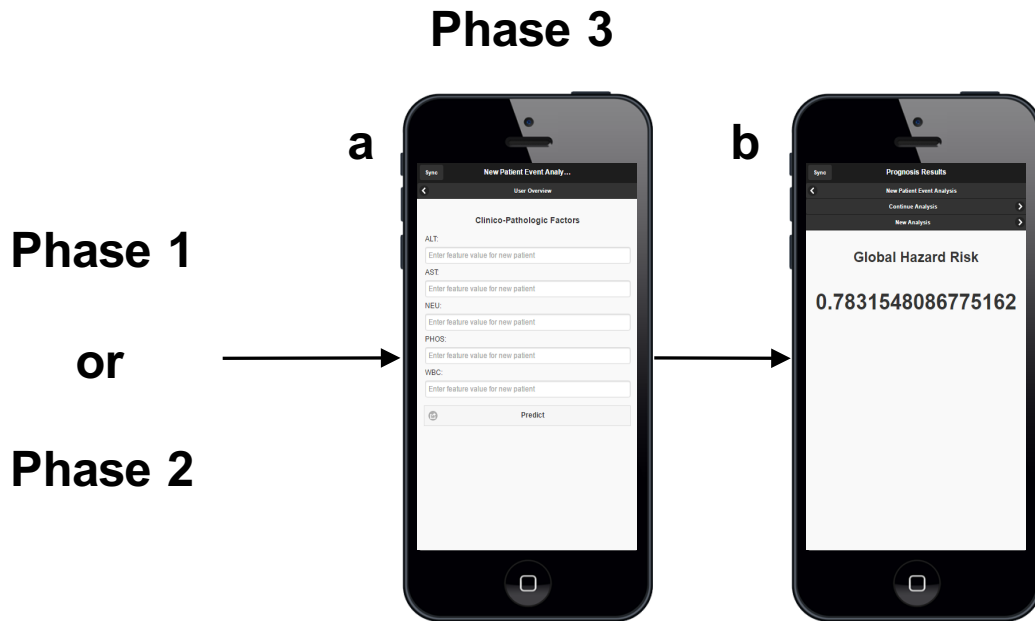


Figure 4.30: Visual depiction of the various phase 3 user interfaces. All the menus available for user interaction are shown.

with additional functionalities. Users are given the opportunity to manipulate the features from Phase 1 and select significant and effective features for predicting the time to event related conditions related to other parameters within the database. In the context of a prostate cancer database, users will have a fixed model to predict hazard risk for each new patient. The coefficient and baseline hazard risks of the model are saved based on the selected significant features and the user is able to find the global hazard risk and also hazard risk over time of new patients by testing or importing their own external features (Figure 4.31).

4.4 Discussion

Understanding and interpreting the uncertain nature of complex systems is a notable challenge that is a central focus of the TMVV application. To address this challenge, we have aimed to provide a platform that guides users towards applicable solutions and interpretations for a given problem. Inadvertently, this serves to also give users a better understanding regarding

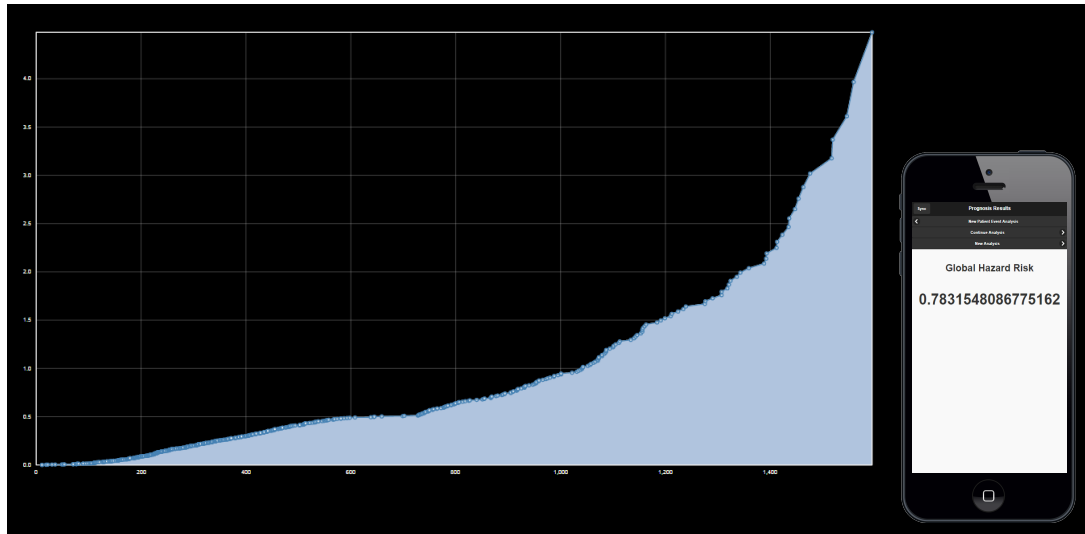


Figure 4.31: Right - Global hazard risk displayed on smartphone; Left - Hazard risk over time displayed on tabletop platform.

uncertainty and how to develop efficient constraints for the further surveillance of a system. These aims propel the presented research in a direction that promotes the integration of various technical fields that would better represent individuals from a multitude of backgrounds. There is not a protocol for the analysis of complicated systems. Rather, individuals have unique designs and mentalities that incorporate their own methods, perspectives and understanding of the systems data into their analysis. This distinct thinking makes it possible for people to implement their own ideas and successfully discover further aspects of the system being analyzed. With the technology that has been proposed we would like to incorporate the aforementioned thought into the user experience. This has been executed by creating an environment where users can consecutively adjust the variables of a system to better understand the cause and effect relationships amongst the parameters and reveal the unpredicted correlations. A point of debate is why there should be a change in the field of data analytics. More specifically, why argue that the TMVV application should serve as an alternative to existing systems and methods that analyze systems and obtain the best accuracy? The TMVV application is an interactive system that finds and then operates within a solution space. This serves as an

important model because one is often not searching for the best predictive accuracy but rather a set of solutions that approach the system from different perspectives. Fortunately, recent progressions in technology have provided tools and mechanisms for bringing such a system into fruition. The functionality of the TMVV application is dependent upon the use of active tangibles and multi-touch surfaces. By deploying a framework that utilizes these elements the user can better explain the causality of a system, determine how to reach a successful outcome, avoid the repetition of unsuccessful experiences, and reveal what kinds of constraints and simplifications should be considered during the model design process.

The proposed framework creates a schema that facilitates insightful development through the manipulation of parameters based on conditional knowledge gained by the user through experience. In other words, it accentuates the role of knowledge based structures obtained through expertise and experience by allowing the user to make sense of a problem and formulate a strategy to harvest the solutions.

We would not have known to even think about trying hybrid features if we were not from the DREAM challenge community. The feedback came from the DREAM challenge participants who got better performance by considering the features that they are not in the original set of data. There is a question here: How and Why we would have even thought to construct features? This is really interesting kind of direction. That is a collaborative process that we were being aware of that. Therefore, what if we take that approach and build the system that can support this learning and collaboration procedure. We are going to focus on this design in chapter 5.

5 Tangible Tensors

In this chapter, we introduce the concept of bridging the gap between research related to data analysis and research revolving around the visualization of data. Specifically, concepts related to tensors were used as the backbone for developing a unique visually based tool for manipulating and spotting trends in a set of data. First, we introduce the concept of tensors and their relevance to visualizing data. Next, we present the system design for a toolkit that utilizes the tensors concept. Finally, we demonstrate the importance of this technology for providing progressive innovation in the field of biological pathways analysis.

There are two things that we need to address leading to this section. First, one probably is wondering why tangible? Why not visual analytics system that is not tangible? The answer is that the whole idea behind tangible or one of the major ideas is that we can pair off and have people interact with the system independent from another. Yes there are some space where we can support collaboration and we can see what other people are doing but the interactions are isolated. How do we actually build a system to support that? That is what led to Tangible Tensors. Another question is, why tensor? It is due to we could represent the trend of users' thoughts in 3D dimension and we could easily interact with.

5.1 Tangible Tensors

Big data visualization has been implemented for a variety of real-world applications such as demographics and economics. Presenting data in a way that is both intuitive and interactive may pose an alternative to extensive statistical analysis. By giving professionals the tools to organize and manipulate large data sets, they can bypass some aspects of tedious interpretation and instantly view trends, develop theories and extrapolate meaning from their subject matter. The field of computational biology is currently lacking effective tools to visually explore and manipulate data. This may be, in part, due to the lack of an effective model or toolkit that has the potential to radically change how we view the importance of data visualization. As developers, not all the important connections between factors in nature and healthcare are clear to us. However, to a biologist, clinician, or other healthcare professional who understands the importance between certain genes or molecules, being able to interact with these connections in whatever manner suits their needs is powerful and progressive. Tangible Tensors poses as a suitable platform and starting point for some of the tools that could be of use to the aforementioned professionals.

Tensor representation is particularly amenable to visualization, and offers unique opportunities and flexibility in this regard. When implemented in a digital environment, they use color, scale and shape to pose as an alternative to traditional numerical data. Given their 2D or 3D structure, the means by which tensors can be visualized and manipulated is only limited by the design requirements of the developer. Tensors are suitable candidates for data visualization because of their high-dimensional structure and their ability to be transformed and reconstituted relatively easily.

The main advantage of developing a generic software toolkit is that there is no real limitation to the types of data sets that can be explored. To reiterate, the goal is to provide people who know their field with the ability to manipulate and visualize data in the manner that best suits them. The more tools and interactions that the software allows for, the more successful it is in assisting professionals and removing some of the need for tedious analysis. Tensors, as a visualization tool, are intuitive to interact with and we can more readily 'view' information

embedded within a tensor in a variety of ways.

As it currently stands, there is limited work and research dedicated to the development of a platform that hands versatility over to the user. Current research is primarily geared to the development of platforms that deal with specific and singular sets of data [Isenberg *et al*, 2008], or rather, with the design of tensors themselves. These works offer insight into useful tools and techniques for tensor design that can aid the design process for a Tangible Tensors toolkit. The toolkit being explored in this dissertation takes a more multifaceted approach. Where current research seeks to visualize data in the form of a tensor [Zhukov and Barr, 2003], we aim to allow the user to decide on the application and pick from an array of tensor based tools that will assist in their analysis. Another differing factor is that the proposed Tangible Tensors toolkit has been designed with the intent that it will be used in a touchscreen-based interactive environment. Existing research [Deiros *et al*, 2016, Olshannikova *et al*, 2015] has been dedicated to effective visualization, which heightens user experience and provides an alternative for professionals working with large sets of data. Adding an interactive element aims to push this boundary even further, hence the emphasis on the implementation of a touchscreen environment for the toolkit throughout this dissertation.

5.1.1 Related Works

The Tangible Tensors toolkit aims to extract existing effective strategies for tensor technology and fill in the gaps left by existing research. Currently, existing research centers on tensor visualization techniques, human-computer interaction platforms, and reviewing existing tensor technologies and processes.

Among reviewed works there seems to be a theme of increasing sophistication in visualization techniques. Advanced tensor visualization techniques have been used to study, for example, 3D diffusion tensor MRI data of the heart [Zhukov and Barr, 2003]. This would serve the purpose of recovering and visualizing the helical structure and orientation of the heart muscle and its fibers. Visualization techniques have also been applied to neuroscience and tensor realization of brain tissue [Hadjwiger *et al*, 2012]. Visualization techniques pertaining to biological

applications were of particular interest due to the fact that healthcare experts are among the target demographic for the proposed toolkit. Biological data is often given in a discrete form, which results in the need for interpolation. Tensor visualization compensates for this need by incorporating interpolation errors into the data processing pipeline.

Another recurring theme among existing research is that of human computer interaction. Specifically, platforms have been developed for intuitive interaction on touch screens dealing with 2D and 3D animations. One paper in particular explored several techniques to interactively represent 2D vector fields [Isenberg *et al*, 2008]. The proposed techniques worked in conjunction with touch screen displays that allow users to custom design glyphs in the form of arrows or lines, that best reveal patterns of an underlying dataset. This concept has been translated into a field of research dedicated to developing visualizations and interactive tangible objects that can be easily manipulated by a user. In particular, emerging evidence from the area of embodied cognition [Chandrasekharan, 2009] supports the idea that action (through the motor system) and perception (through effective visualization) can be coupled to support and augment conceptual understanding in data-driven domains [Shaer *et al*, 2013]. Recent research in tangible and gestural interactions provides a means to support users in forming insights from large or complex datasets. Examples of this include the G-nome surfer which employed multi-touch tabletop interactions to navigate genomic data [Shaer *et al*, 2012], and SynFlo which utilized active tokens, programmable physical objects with integrated display, sensing, or actuation technologies, to simulate biological experiments [Xu *et al*, 2013]. Systems such as these build on and support a scientist's use of external artifacts to address complex problems. These artifacts come in the form of models, diagrams and instruments and are often pivotal in the support of their experimental reasoning [Nersessian, 2010].

Understanding the usefulness of tensors in real-life can come from a high level overview of how they are designed to represent data. Tensor visualizations are primarily implemented with the use of algorithms that have the ability to show local or global information in the context of a personal set of a data. Tensors have been used, for instance, to guide the rendering of volumetric datasets, by guiding transfer functions for scalability, color assignment and opacity [Vilanova *et al*, 2006]. Complimentary research explores visualization techniques

for 3D tensors [Jeong *et al.*, 2013]. Properly implemented 3D tensor visualization results in models that are truly depictive of that the data that they represent. The respective papers in this area of research emphasize that 3D visualization is a challenging task traditionally carried out by algorithms. However, no single algorithm for 3D tensors exists that is sufficient for visualizing all of the vital aspects of a data set on a digital platform. To summarize, while many papers explore visualization techniques, actual applications for these techniques have not been thoroughly explored.

Authors who have explored tensor fields through a critical lens offer insight into the areas in need of innovation. Tensors are used extensively in numerous engineering disciplines for modeling, simulation, and analysis [Hlawitschka *et al.*, 2014]. Engineering tensor data tends to be complex and large in volume [Carpendale, 2008], hence challenges are faced in creating effective visualization tools. This fact has fostered a recent interest in the visualization community to advance visualization and processing techniques for engineering tensor fields [Carpendale, 2008]. Designing effective visualizations for engineering tensor fields is a multifaceted problem, which includes factors like visual intuitiveness [Franklin, 2015], scalability [Yalcin *et al.*, 2016], interactivity [Boy *et al.*, 2016], merging detail and context [Tax *et al.*, 2015], integrating modeling and simulation [Nowke *et al.*, 2015], and overcoming terminology barriers [Thielea *et al.*, 2015], among many others. All this complexity usually results in trade-offs among different visualization strategies. Several authors have stressed the importance of validating research in information visualization due to its increasing importance and relevance to engineering and healthcare applications. This body of work argues that not enough research has been completed to justify how best to use information visualization tools such as tensors for real life applications. Particularly, research makes a case for the scale on which information visualization is tested on, arguing that development tests deal with data sets that are much smaller and more ideal than the data collected from real life applications or demographics. The authors contributing to this body of work, which summarizes the challenges in creating effective visualizations [Hlawitschka *et al.*, 2014], aim at initiating a discussion among the developer community about new ways of creating solutions to the problems in existing tensor technologies.

5.1.2 Research Rationale

Tensor visualization is a central topic among the scientific community because it can carry data and information in higher dimensions such as 3D and above, as dynamic tensors offer the analysis of information from a multitude of perspectives. The main challenge in dynamic tensor visualization is balancing the volume of information with visualizations that are optimal and easy for the user to understand. A major goal in tensor research is the development of a visualization that can indicate the most significant and effective features of data and also help query the changes made to the data set into a solution space over time.

A main focus of recent research in biomedical engineering is modeling, optimization, and interpretation of molecules, features or devices in systems. Tangible tensor visualization can play a significant role in such processes by representing the relationships among the various components of a complex system (e.g. biological circuit or signaling pathway). Varying behaviors of system components changes the results of modeling and optimization, which leads to various interpretations that can be tangibly visualized. Thus, tangible tensors are an effective tool to gain insight into system architecture and component interconnections while also helping to verify and evaluate a system's behavior visually. The proposed toolkit will allow one to highlight the most effective molecules, features, or devices of a system and sort through them based on different validation terms for system interpretation.

The review process strengthened the argument that there is a need for tools such as Tangible Tensors to be applied to biology. As it stands, no works deal with the organization and manipulation of biological data to various degrees of detail and scale. Tangible Tensors poses a means to bridge the gap between data visualization and biology; however, it is not only limited to biological applications. The research process and study design of Tangible Tensors will address the following questions in response to the lack of exploration in the field:

- How can data be visualized and manipulated to increase meaning and better support the formation of insights?
- How can interaction in digital space be integrated with the needs of professionals from

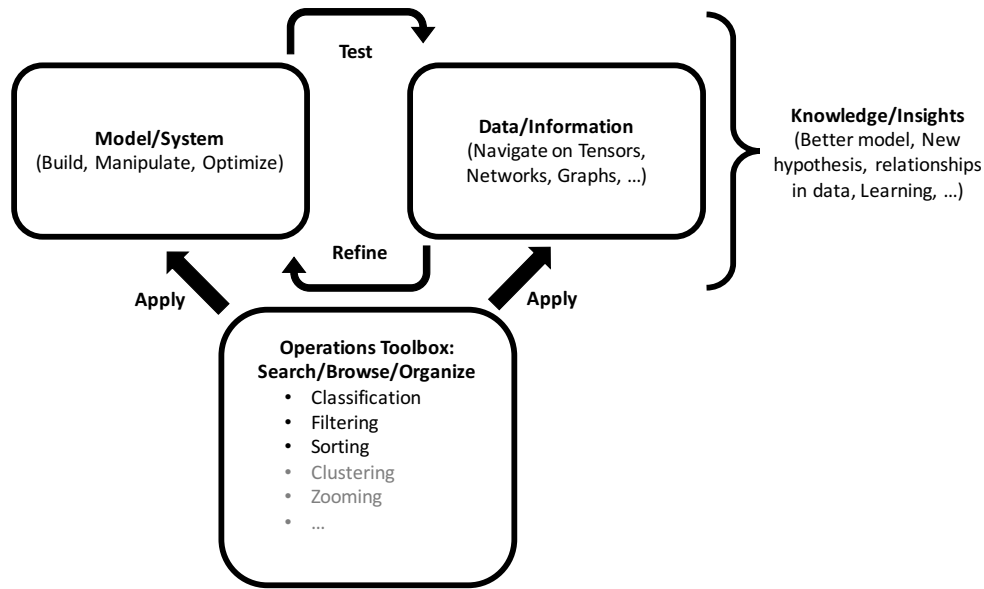


Figure 5.1: General flowchart showcasing the Tangible Tensors toolkit.

a multitude of industries?

In response to the discrepancies in the field of tensor visualization, we have developed a toolkit that provides users with a practical means of analyzing data (specially biological data). The platform was designed for use with multiple forms of data, meaning that it would be convenient and useful to professionals from a variety of fields. Further, the interactive element introduces touch screens and a variety of tensor menus. The menus allow users to view data from multiple perspectives while maintaining an organized workflow. This is especially important when a user is looking to manipulate data and compare it to a previously manipulated set. The toolkit will not only provide a tensor visualization of the existing data set, but a basis for comparison between the original and manipulated sets derived from modeling. With that said, the gaps in current research are being addressed with less emphasis on advanced visualization techniques and more emphasis on multiple perspective visualization [Li, 2004], organization and interaction [Ragan *et al*, 2016].

Fig.5.1 provides a general sense of how the toolkit addresses the gaps left by current research

on tensor visualization. The user navigates through a variety of menus with the use of an active tangible, an interactive physical object that includes display, sensing and/or actuation capabilities. Each menu provides the user with functions that allow them to explore their data and visualize it from a variety of different perspectives. Fig. 5.1 shows the general operation of the toolkit, explained in further detail in the next section. The model/system component represents the project screen or main menu that the user sees upon opening the toolkit. Data analysis works in conjunction with the data/information screen, which provides a more detailed medium for interaction and analysis. At any point in the interaction, the user has access to a variety of functions that will carry out many of the traditional tasks associated with data and statistical analysis. Some of these functions include granting the user the ability to classify, sort and cluster data. The design section will discuss how the graphical interfaces of the proposed design are integrated with the functions to visualize tangible tensors and interact with them.

5.1.3 Toolkit Design

The Tangible Tensors toolkit provides a series of menus that showcase data visualization in varying degrees of detail. The user inputs data corresponding to the subject matter of their choice along with their expected results. They are then able to modify parameters of this data and observe the outcome. The cause and effect relationship for each individual test is displayed on an error radars chart. Multiple tests, resulting in multiple radar charts, form the 3D tensor. Each tensor corresponds to a different user. Users make use of the functions present in the Tangible Tensors toolkit with the use of personal active tangibles [Arif *et al*, 2016]. The proposed active tangibles are uniquely shaped with touch screens and sensitivity/responsiveness to a variety of strategically selected motions. An interactive touch display is being used along with the active tangibles. Active tangible interactions can occur on a tabletop surface or a multi-touch wall display. For users without access to touch screen platforms the toolkit can be used with an ordinary desktop/laptop screen in which case equivalent interactions can be achieved through traditional input devices like a mouse.

For development purposes, the toolkit was built around the relationships between parameters. It is important to keep in mind that the parameters are merely representative of quantities or variables that can be determined by the user depending on their field of study or research interests. The data initially entered into the system includes parameters of interest and their expected values. The main menu of the Tangible Tensors platform provides a multi-section arrangement of screens. This interface provides the user with a complete overview of the project.

Project View (Phase 1)

The main menu of the toolkit is divided into multiple views, each corresponding to a visualization and series of operations that can be carried out by the user. Fig. 5.2(a) shows the general layout of the project main menu and will be used as a reference for the description of each segment's functions. All parameters and experimental data of the system of focus, in the form of comma-separated values (CSV) file, are loaded into Tangible Tensors through the initial data-selection menu. Once the parameters are loaded and the toolkit is started, the parameters become available for model-building and simulation in the project view. Parameters can be added to the project and edited via context or active tangible menus. The parameters visualization view gives a visualization of these parameters, all of which are subject to change by the user. The user has the option to change all, some or none of these parameters and run a test based on these changes. The parameter concentration-time graphs depict the most recent iteration of a test run by a user. It is important to keep in mind that the toolkit is being presented to show the result of changing parameter concentrations; however, the same principles can be applied to any form of data. The Root Mean Square (RMS) error radar chart segment displays a radar chart generated in real-time, which quantifies the degree of error in parameter concentration for the most recent iteration. Finally, the tensor bar displays each user's personal tensor, which is essentially a collection of parameter modifications modeled as a 3D visualization. The tensor can be thought of as a collection of 'slices' that are unique radar charts representing solutions/degrees of error based on the manipulation of parameters. With

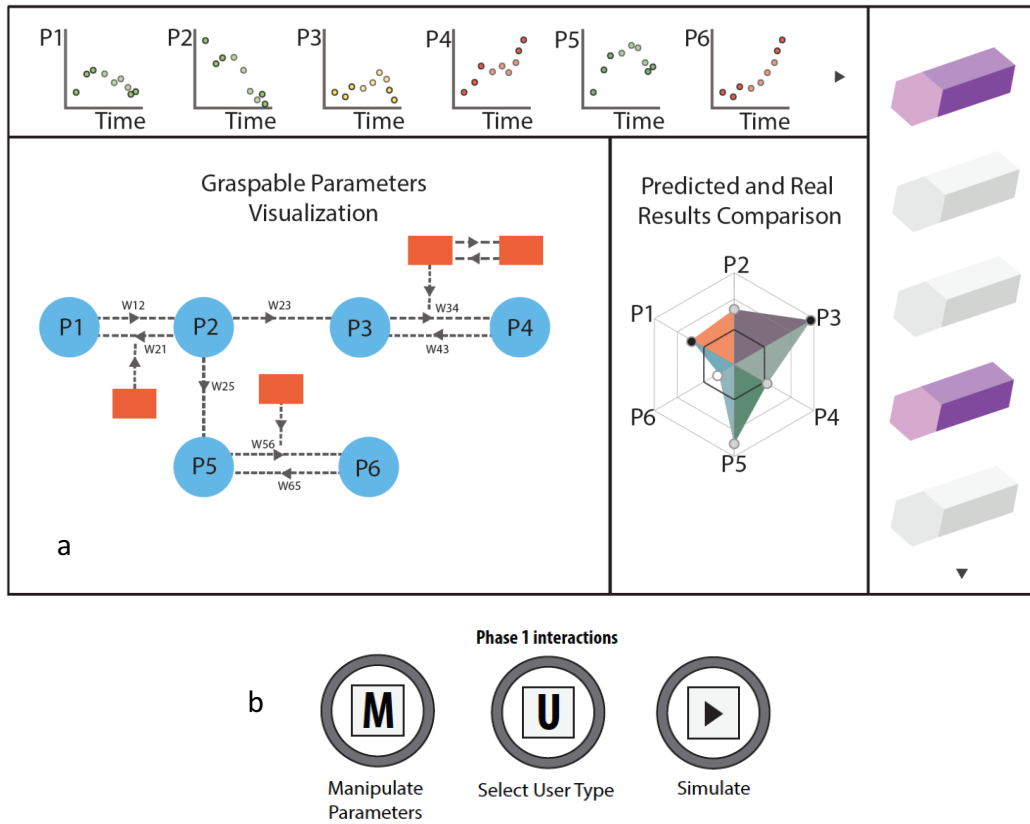


Figure 5.2: (a) General view of the project screen in the tangible tensors platform. Overview of the multi-section arrangement; (b) Functions available to users in main menu of toolkit as displayed on the active tangible.

this in mind, the user is building their own tensor with the number of 'slices' being equivalent to the number of tests or manipulations that were run.

Parameter Visualization Interactions In the parameter visualization screen, the user is granted access to specific functions that work in conjunction with their own active tangible and the touchscreen on which the toolkit is being run. An overview of the available functions can be seen in Fig. 5.2(b). The defined functions in the parameter visualization screen are as follows:

1. **Select User Type** - the user has the option to select their profession and apply it to their active tangible. By doing so, all the manipulations made by that user will be organized under the title that they specified. This effectively functions to label the process of manipulation followed by different groups of users. It can be noted that the application of this function is not limited to classifying different users according to their title or profession. Rather, if desired, it can be used as an organizational tool for a single user analyzing multiple sets of data.
2. **Manipulate Parameters** - in the parameter visualization segment the active tangible gives the user the ability to change the parameters corresponding to their data set. The user would use the touchscreen of the active tangible to select the manipulation function and then place it over the parameter of interest. Once placed, the active tangible's display will allow the user to change this parameter (this will usually consist of an increase or decrease in value that was predetermined by the user).
3. **Simulation** - after certain parameters have been manipulated the user will have the option to select a simulation function on the active tangible and run the simulation corresponding to the changes that were made. The simulation will be shown in real time with the "concentration over time" graphs, in the form of a series of plots, corresponding to each parameter. Following the completion of the simulation, the iteration will be displayed as a radar chart that is part of the user's personal tensor in the "tensor" menu bar. This can be seen in Fig. 5.2 as the "play" icon on the active tangible in phase 1.
4. **Selection** - the selection function makes use of the touchscreen on which the toolkit is being run. Without the need for their active tangible, the user is able to touch the "tensor" menu bar, which opens a new menu. This menu will be explained in further detail throughout the upcoming sections of this dissertation; however, essentially allows the user to examine, in further detail, their tensor and the results of their manipulations.

The purpose of the simulation process can seem somewhat abstract when explained out of context. Each simulation generates a radar chart that visualizes the error or results of changing

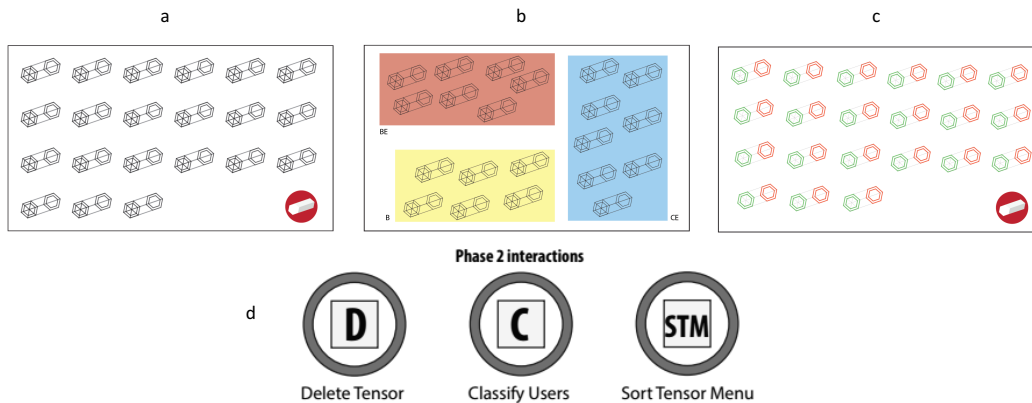


Figure 5.3: (a) Overview of the tensor menu screen. All the user tensors are displayed; (b) Visualization of the classification function. Tensors belonging to groups of users such as Biomedical Engineers (BE), computer engineers (CE) or Biologists (B) are all clustered together; (c) Visualization of the sort tensor function; (d) Functions available to users in the tensor menu of toolkit as displayed on the active tangible.

certain parameters in a system. With this in mind, depending on the application, the user is able to interact in a cause-and-effect environment. They are able to change parameters, and visually see the results of those changes. This goes back to the original motivation for the toolkit; a demand for making the interpretation of data intuitive and transitioning away from solely numerical analysis. The tensor is representative of a solution space, with each 'slice' or radar chart being representative of a particular solution.

Tensor Menu View (Phase 2)

As previously explained, the tensors bar in the main menu displays all the currently active tensors. Each tensor within the tensors bar corresponds to a single user, and interactions within it are defined by the active tangible assigned to that user, as described above. This process allows the Tangible Tensors platform to run multiple projects simultaneously. When this menu is selected through the touchscreen the tensor menu screen, it will appear as shown in Fig. 5.3. The tensor menu screen displays each user's tensor and grants access to another

set of unique functions through the active tangibles.

Tensor Menu Interactions The tensor menu screen gives a visual list of all the user tensors that have been active at some point in time regardless of their current state. Fig. 5.3(a) shows how the collection of tensors will be displayed. The functions as they will be displayed on the active tangibles can be seen in Fig. 5.3(d). With the use of the active tangible the user has access to the following functions in this menu:

1. **Classify Users** - the user classification function is used to organize the tensor menu into clusters. One scenario could have this screen sort the tensors based on user type (label). For example, a cluster of tensors that belong to Geneticists (G) alongside another cluster that represents all of the tensors belonging to Computational Biologists (CB) and so on. This classification into the labeled groups allows for an easy overview of all the tensors in relation to various users. Outside the purpose of user classification, this also presents an opportunity for the user to investigate the performance of different groups as they participate in the modeling of the same set of parameters. This can be seen in Fig 5.3(b).
2. **Sort Tensor Menu** - as mentioned earlier each tensor is a compilation of radar charts or 'measures of error' based on how the user manipulated parameters in the project screen. This sorting function will sort all of the tensors based on a specific parameter. This function treats all tensors equally regardless of which class they belong to. The user is able to sort all of the tensors according to the different validation terms such that tensors are organized from the best predicted outcome to the least desired outcome. Color coding the best outcomes and least desired outcomes using a color gradient from green to red respectively provides additional visual aid to decipher the inter-sorted tensors. This can be seen in Fig. 5.3(c). It is fair to presume that an effective data analysis toolkit would be centered on finding the most optimal solution or outcome. However, complex problems may have several solutions with acceptable data fits for a set of predefined criteria. Hence, it is beneficial to provide the user with all of these solutions so that their specific parameters can be compared.

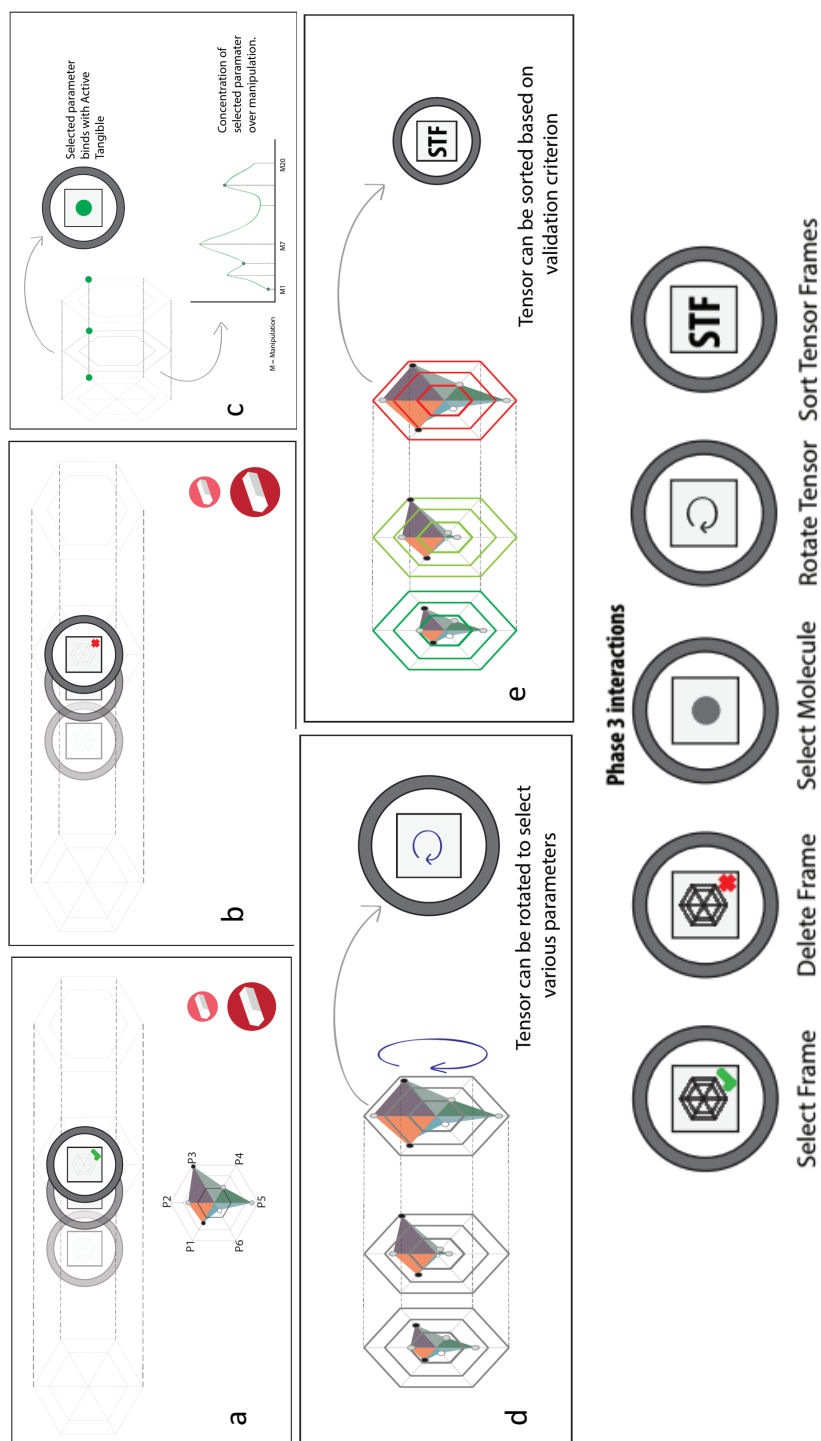


Figure 5.4: (a) Visualization of the select frame function; (b) Visualization of the rotate frame function; (c) Visualization of the delete frame function; (d) Visualization of the sort tensor frames function; (e) Visualization of the sort tensor frames function; (f) Functions available to users in the tensor screen of toolkit as displayed on the active tangible.

3. **Delete Tensor** - the user has the ability to delete a whole set of tensors that they feel are not useful to their work. This function is the final function that the user is able to select with their active tangible in the tensor menu screen.

Similar to the project screen, the tensor menu screen allows for a degree of functionality without the use of the active tangible. The touchscreen can be used to select a specific tensor in the menu and initiate phase 3 of the tangible tensors toolkit.

Tensor View (Phase 3)

The tensor screen is the most detailed level of visualization in the Tangible Tensors platform. Selecting a specific tensor from the tensor menu screen brings up the tensor screen. The tensor screen provides a 3D visualization of the tensor, in addition to detailed information regarding the selected tensor. This information is presented from multiple perspectives and accessed through a variety of functions on the active tangibles. The functions available via the active tangibles are shown in Fig. 5.4.

Tensor Screen Interactions Interaction with the tensor screen is achieved, once again, through the active tangible. These vary from simple interactions to more complex information-rich interactions, as shown in Fig. 5.4(f). These interactions can be summarized as follows:

1. **Select Frame** - individual radar charts can be selected and bound using the active tangible for further manipulation. Selecting a specific frame or "slice" displays the selected radar chart in a menu beside the respective tensor. This provides an enlarged visualization of the radar chart and makes the degree of error and parameters much more visible to the user. This can be seen in Fig. 5.4(a).
2. **Delete Frame** - frames that yielded undesired results have the option to be deleted. This can be seen in Fig. 5.4(b).
3. **Select Molecule** - translating the active tangible across the tensor brings up the cor-

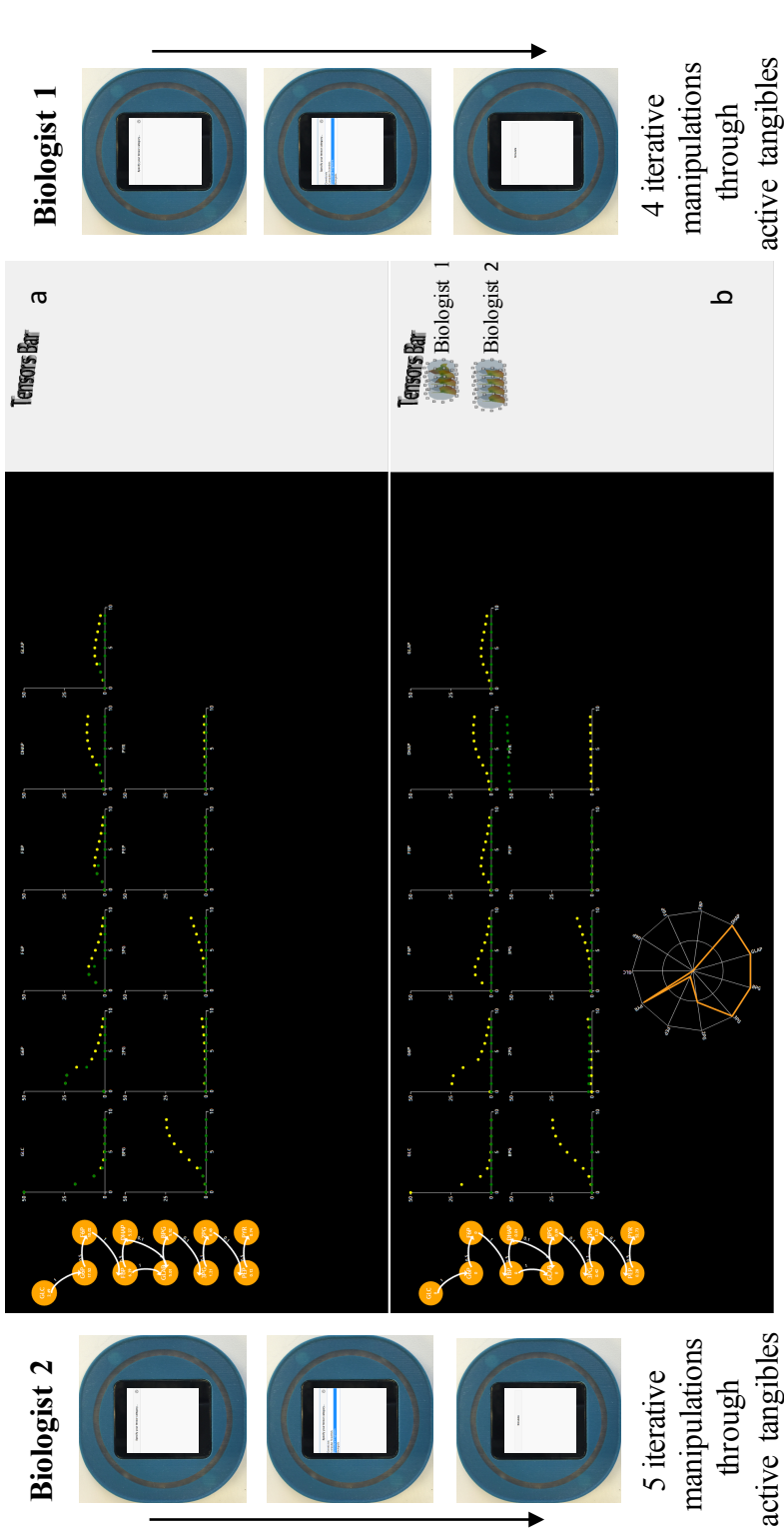


Figure 5.5: (a) Main menu of the pathways project when first opened; (b) Following the manipulation of parameters, the plots will show the experimental results in real time and generate an iteration of the user's tensor in the tensor bar.

responding data point, which can be selected for further manipulation. Selection of individual data points or parameters reveals the variation of this parameter across the number of manipulations/iterations that composed the user's tensor. This can be seen in Fig. 5.4(c) as a plot.

4. **Rotate Tensor** - the tensor can be rotated or moved around in a 3D plane for easier viewing by rotating or moving the active tangible. This can be seen in Fig. 5.4(d). Through rotation the user is able to analyze parameters that are in their field of view by default in addition to parameters that are initially more difficult to see. Large data benefits from multi-perspective analysis that is not possible with static 3D modeling where components of the data are not in the forefront of the visualization.
5. **Sort Tensor Frames** - one of the primary objectives of the toolkit is to find the best configuration of parameters (the simulation that yields the least amount of error in the solution space). While general color coding makes it easy to distinguish good solutions from bad solutions, sorting will allow the user to gauge strong solutions among other positive results in the solution space. This function will arrange the simulations from best to worst. This can be seen in Fig. 5.4(e). Due to the fact that the toolkit has the ability to present the user with several solutions and then sort them from best to worst it is fair to wonder what kind of scenario would lead someone to select a worse solution over a seemingly better one. By providing multiple solutions, we are adhering to original ideology that as developers we are not aware of what may be relevant to a user from a specific profession. Further, creativity and understanding in the context of system modeling is more dynamically achieved when surveillance is done outside of what is typically considered the best solution space. Analyzing solutions that will worsen the result challenges the ideas of the user to ponder the strengths and weaknesses of their system as a whole. Additionally, the best solution may not be plausible due to resource constraints or even feasibility. Providing users with alternatives creates an adaptive environment where the user can find creative ways to maximize their resources.

5.1.4 Active Pathways Case Studys

The details discussed in the design section were brought to fruition through the Active Pathways project; a platform designed to mimic the functionality of the proposed Tangible Tensors toolkit with a specific scenario in mind. The Active Pathways project is an active tangible and interactive tabletop system that supports modeling of biochemical reaction networks. Specifically, it grants the user the ability to manipulate the concentrations of interconnected molecules and visually observe the outcome in real-time in relation to one or more experimental datasets ([Shaer *et al*, 2013],[Wu *et al*, 2011]).

Active Pathways: System Overview

The interaction in Active Pathways is defined around two major tasks involved in modeling biochemical pathways. First, building the model, and second, fitting the model to experimental data by adjusting parameters. Multiple active tangibles are used for both model construction and fitting tasks. The active tangibles provide the user with a hierarchy of menu options that are directly related to the functions outlined in the design proposal. The menu is initially categorized into a "model" option and "data" option. Users switch between these options by tilting their active tangible and a selection is made by tapping the active tangible's screen. When the desired function is selected, the user is able to place their active tangible on a touch screen displaying the pathways menu and change parameters as desired. Additionally, the active tangibles provide "play" and "stop" options, which allow users to start and stop the simulation when ready.

Tangible Tensors applied to Active Pathways

Many of the described visualizations have been implemented for each of the menus. The project also has a high degree of functionality. As it stands, many of the required functions for each menu are accessed through direct interaction with the touch screen display. The end goal is to transition the project to a system that has functionality through the user's active tangible.

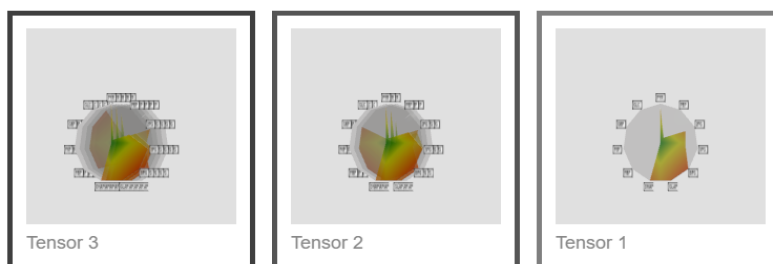
The upcoming series of figures will show the implementation of the pathways project and describe its connection to the design for the Tangible Tensors toolkit.

Fig. 5.5 shows the pathways menu as displayed on a touch screen platform. The menu was designed to resemble Fig. 5.2(a) visually and in terms of all the functionality that was mentioned in the toolkit design. The active tangibles are functional in this phase of the pathways project. The user is able to use their active tangible to manipulate molecular concentrations. Following a manipulation, a simulation can be initiated, in which case the results will be shown in real time via the plots in the top portion of the pathways menu. Following a simulation, the user will see the first iteration of their tensor appear to the right of the pathways menu in the tensor bar. Once a user selects the tensor bar it takes them to the tensor menu. Fig. 5.6 depicts the tensor menu as displayed on a touch screen platform. While all the desired functions are active, they are not accessible through the active tangible. All the active user tensors are displayed, and through the touch screen capabilities the user has the ability to cluster, sort and select tensors. Once a specific tensor is selected from the tensor menu, the user is taken to the tensor screen. Fig. 5.7 showcases how the tensor screen has been visualized and implemented. The functions described in the design proposal are all active, however, at this stage, depend on the capabilities of the touch screen display and are not accessible through the active tangible. In the tensor screen the user is able to select frames, delete frames, select molecules, rotate the tensor and sort the frames of the selected tensor based on the strength of the solution.

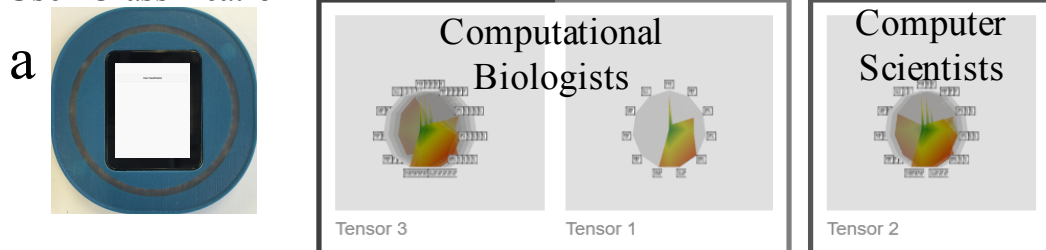
5.1.5 Discussion

The Tangible Tensors toolkit is a new tensor based visualization and manipulation tool that serves to improve functionality over previous data analytics approaches. To address this goal, we aimed to provide a platform that guides the user towards solutions for a problem/system and how to better interpret and develop efficient constraints for the system at hand. These aims propel research in a direction that encourages the integration of fields including human

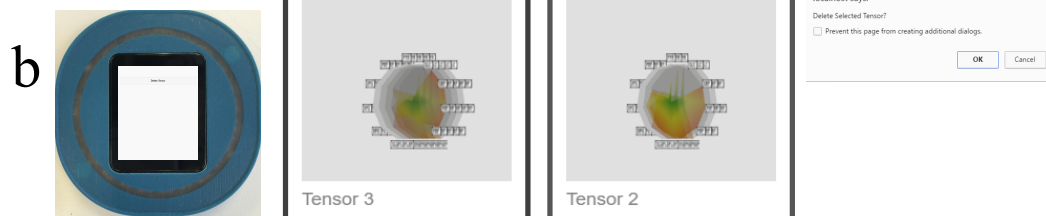
After pathways
manipulations by
three users



User Classification



Delete Tensor



Sort Tensors

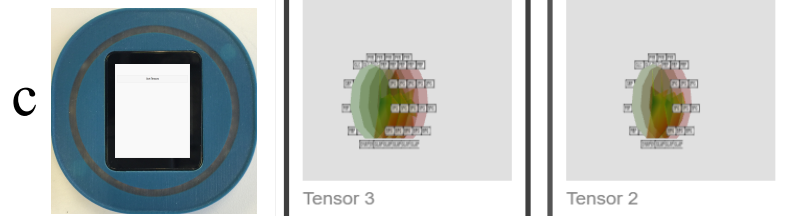


Figure 5.6: (a) Visualization of the user classification function; (b) Visualization of the delete tensor function; (c) Visualization of the sort tensor function.

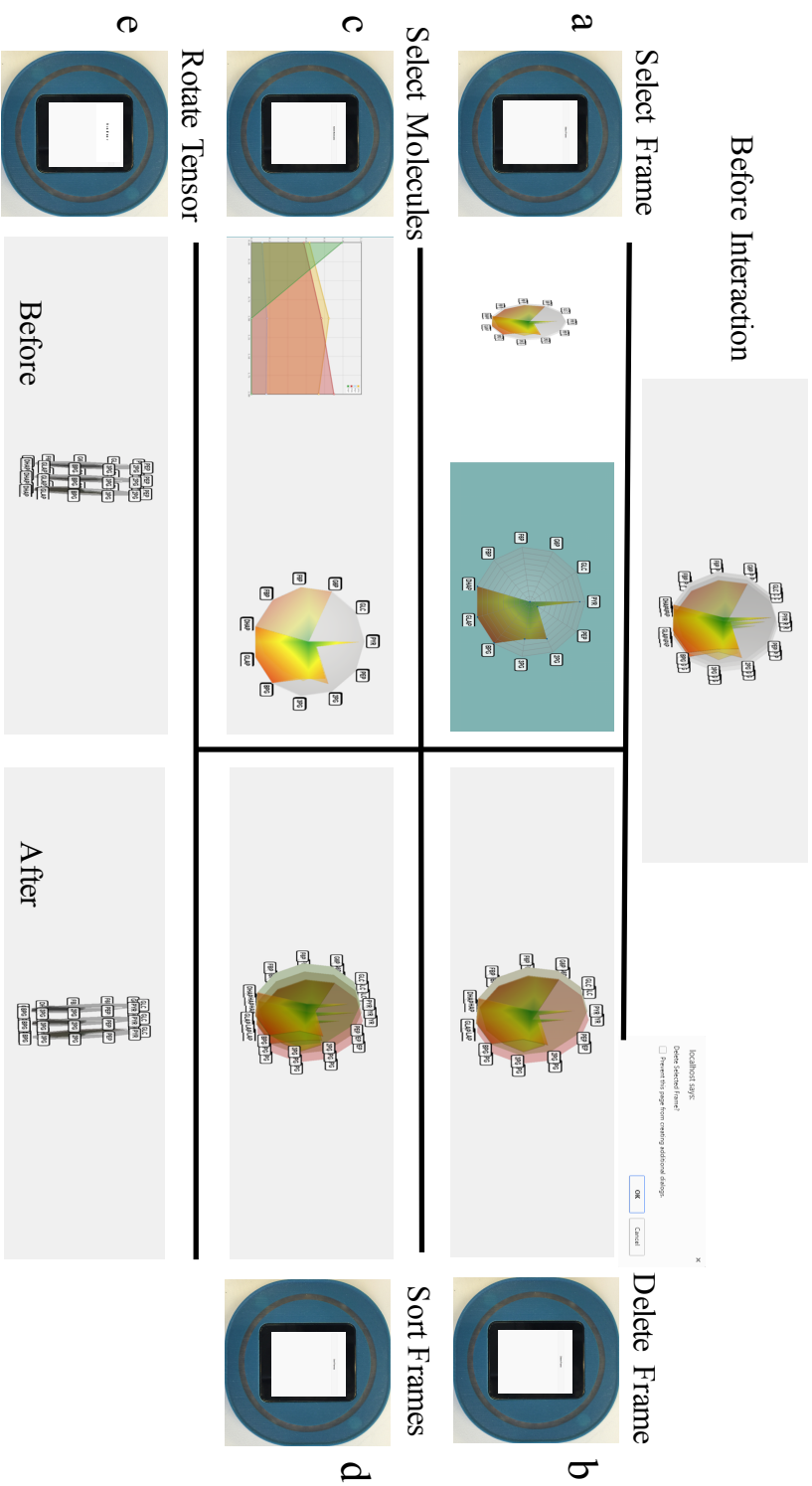


Figure 5.7: (a) Visualization of the select frame function; (b) Visualization of the delete frame function; (c) Visualization of the select molecule function; (d) Visualization of the sort frames function; (e) Visualization of the rotate tensor function.

computer interaction, computer science and other engineering and science disciplines depending on the needs and backgrounds of the development team.

One usually does not have direct experience in the analysis of a complex system. However, individuals all have a unique idea regarding the design of a system that integrates their own perspectives and experiences with the specific parameters of the system. The tangible tensors framework fully embraces this approach and allows people to manipulate and design their analysis in such a way that they are able to discover more aspects of the system. By using this framework, one can better explain the causality of the system, figure out how to reach a successful outcome, avoid repeating previous non-successful experiences, and in the model design process, gain insight into what kinds of constraints and simplifications should be considered.

6 Conclusion and Future Directions

TBNs, TMVV, and TTs are prototypical implementations of the interactive gene regulatory networks exploration, interactive survival analysis, and interactive biochemical networks modeling presented in this dissertation respectively. The reconstruction approaches themselves can be applied as patterns for facing the various problems in achieving situation awareness from biological and clinical prognosis systems. It has shown how visual analytics and tangible user interface fields can be used to solve big data challenges. This chapter summarizes the contributions of this dissertation, presents lessons learned, and recommends the future directions.

6.1 Summary of Contributions and Lessons Learned

In chapter 1 of this dissertation, three main research challenges were indicated based on requirements of domain experts. They comprise the questions of how the genes whose gene expressions change with time, interact; how structural features in genetic networks can be visualized and interpreted interactively; how we can extract the effective set of prognosis factors to predict hazard ratio of advanced prostate cancer patients more accurately; how processing and visualizing of prognosis factors can provide better understanding of their effectiveness on time to event (death) prediction; how one can better explore the causality of a system,

figure out how to reach a successful outcome, and avoid repeating previous non-successful experience. Following these requirements, the dissertation has presented four corresponding methods, three tangible visual analytics platforms.

Chapter 3 has highlighted a modification on gene regulatory network reconstruction algorithm using a stable network topology via sparsity-seeking convex optimization. Recently, many recovering methods are focused on discovering the dynamics of genetic regulation. Each of these methods apply one or several structural properties (e.g. prior knowledge such as sparsity) as constraint(s) to their optimization method. Although some of these constraints are defined according to real features of biological networks, a few of them are inspired directly from the signal processing concepts and no validity evidence is obtained from gene regulatory networks to confirm them. Stability criterion is one of these constraints which is defined based on its definition for continuous domain in control systems field (e.g. if the real parts of eigenvalues are in the left-hand plane, the system is stable). Based on the obtained results from eigenvalue spectrum of 30 gene regulatory networks, we revealed evidence that the eigenvalue distribution of these networks is following a circle shape around origin axis in complex plane. To this end, I decided to utilize the discrete definition of stability criterion (e.g. a system is asymptotically stable if all its eigenvalues are located in the unit circle of the complex plane). The inferred results show the effect of new constraint on the performance of reconstruction method.

At the same time, a comprehensive survey on existing convex constraints prompted the design a new constraint to be applied to optimization methods. This new constraint is inspired from this simple concept in mathematics that if a sample point is placed in a cluster, its neighbouring sample points would also be placed in that cluster with high probability. This modularity criterion is named neighbor norm. I have tested the performance of this constraint along with sparsity and low rank properties inside of an optimization method by using a labeled synthetic dataset. By sorting the samples based on the predefined labels, the reconstructed result shows the performance improvement in comparison with previous methods (which utilized sparsity and low rank properties) qualitatively. I needed to compare the obtained results quantitatively. Ultimately, I could find a theorem in spectral graph theory which creates

a relation between the number of connected components and Laplacian eigenvalue spectrum. It shows the multiplicity of the eigenvalue 0 in the Laplacian spectrum equals the number of connected components in the graph. By performing this criterion, we could show obviously that our method infers more accurate result.

The obtained experiences in designing the mentioned reconstruction methods encouraged me to focus on how we can provide a platform to visualize and interpret the structural features of biological networks interactively; how a user can enter an analysis loop that continuously advances their query dictionary and his own understanding of the networks with different sizes; how users could collaborate and share their own prior knowledge about these networks. To address this questions, a novel platform called Tangible Biological Networks (TBNs) has been developed by using of multi-touch screens, actibles (customized active tangibles in our lab [www.synlab.ca]) and tablets (e.g. iPad2) presented in chapter 3. This platform allows us to explore structural features in biological networks by a wide range of filters and compare them to reveal hidden and unexpected concepts from the reality of data. An informal user study showed the usefulness and applicability of this platform for users with different expertise. Although some of the filters were not easy to apply for the users specially biologists and they might need some training, the users from different disciplines were convinced that TBNs would significantly enhance their understanding and interpretation abilities compared with traditional platforms.

The reconstruction problem for extracting key clinical and pathological features to reach better time to event (death in this dissertation) prediction. In addition, a visual analytics tool to provide a platform for exploring the reasons behind of effective features selection and creating novel feature interactions based on users' prior knowledge interactively were addressed in chapter 4. In the part of reconstruction, a general method was presented that finds effective prognosis factors by utilizing a novel two-phase feature selection and Cox model. This method automatically infers the effective ones to predict hazard ratio for advanced prostate cancer patients by considering possible inconsistency between clinicopathological factors. The strength of this approach is its generality.

The obtained experience in designing this method showed me that understanding of statistical

6.1. Summary of Contributions and Lessons Learned

terms' validity and also the factor combinations and recategorization as new features that are not included in database originally could improve the accuracy performance of prediction. Making these factor combinations needs prior knowledge regarding the given disease (e.g. prostate cancer in this dissertation). We got consultations from two physicians to combine and recategorize the original features. This fact encouraged me to focus on how we can design a platform to visualize and interpret prognosis factors of a given disease; how a user could enter an analysis loop and not only get access to different formats of visualization and representation but also have access to the results of univariate and multivariate survival analyses to better understand the role of clinical factors individually or multi-selected; how we could record the prior knowledge of different users during their manipulation process and gain from their reasons for categorization and combinations of clinical factors. To address these questions, a novel platform called Tangible MultiVariate Visualization (TMVV) has been developed by using multi-touch screens and smartphones presented in chapter 4. This platform allows us to investigate the role of prognosis factors in hazard ratio prediction of a given cancer dataset by using a proper range of visualization tools and record the clinicopathological factors selected by users and their selection reasons. These selected factors are represented as interactive features network (nodes present prognosis factors and edge show bivariate combinations of selected factors). Users have this chance to detect existing faults and incomplete parts of database by different visualization tools. An informal user study allowed us to figure it out what kind of visualizations and modeling results a clinician would like to use when interacting with this system. The suggested feedbacks allowed us to improve the usefulness and applicability of this platform for users with different expertise.

As mentioned in the dissertation, modeling problems are typically tackled through algorithmic approaches. These can sometimes fail, since even if the algorithms find a solution, it is difficult for researchers to know how and why such a solution was obtained, and whether it is even a good solution. Also, these approaches do not easily support shared understanding between researchers from different backgrounds. To address these challenges and questions, we have developed a novel platform that can enable researchers to directly and iteratively build and manipulate models, informing their understanding of complex systems and enabling them

to make new hypotheses and develop efficient constraints for further exploration of a given system. We named this platform, “Tangible Tensors (TTs)” This platform allows researchers with differing disciplinary expertise to consecutively adjust the structure and variables of the system to gain a better understanding of the cause and effect relationships among the parameters and to also be surprised by the predicted effect. The aim is to find solution(s) that can address the problem context appropriately. Factors include whether the selected model, structure and parameters are an appropriate match for the given system, which requires confirmation from specialists. It may be that a solution appears good, but that the structure or parameters of the model do not accurately represent the specialist’s understanding of the real system. Indeed, finding the key factors and also the representative pathway to frame the problem provide a flow to reach the suitable solution(s). This platform also provides an interactive environment for the users to explore the solutions space related to other users who used the system before and learn and make a mental model before manipulating the system.

6.2 Future Work

Ongoing Tasks on Stability Constraints: For applying first reconstruction method on big networks, I am going to apply a relaxation on the utilized stability property because I could explore that although the eigenvalues distribution in complex plane related to big networks follows the circular shape, all eigenvalues are not inside of unit circle. Indeed, they are in a circle with a bigger radius which depends on that certain organism. I would like to apply the stability relaxation in reconstruction method and have an adaptive radius as an undirected prior knowledge from each biological network. Also, we should consider that there are few big positive and real eigenvalues in large networks which should consider these somehow within the stability constraint or as a complement constraint.

Ongoing Tasks on Modularity Constraints: I am going to expand the modularity property on network science and consider if a gene is placed in a module in the network, the other genes with high similar expression patterns would be also placed in that module with high

probability. I would like to apply this as a new structural feature on reconstruction of gene regulatory networks and test the results based on Laplacian spectrum validation term.

Ongoing Tasks for Tangible Biological Networks: The current version of Tangible Biological Networks platform is a prototypical implementation including more than 30 mathematical filters which are applicable and comparable interactively. Based on a vast study on existing features in mathematics, I have collected around 100 structural features applicable on the networks. I am going to add all of them in a framework and make a commercial tangible and web-based toolbox. Also, more comprehensive user studies will be conducted on this platform to survey the validity of the users' design decisions.

Ongoing Tasks for Hybrid Method for Estimating Overall Survival: We have applied this method on two different databases (e.g. breast cancer and Alzheimer's) that their results are not reported in this dissertation and could achieve really high accuracy in terms of time to event prediction. These results could be another proof beside of using 20-fold cross-validation and independent validation dataset in chapter 4 to show the method is not overfitted. Because of these achievements, I claimed that the designed method is a general method and independent of databases. I am going to extract the details of survival analysis related to these two new databases (like prostate cancer in this dissertation).

Ongoing Tasks for Tangible MultiVariate Modeling and Visualization: The current version of Tangible MultiVariate Visualization platform is a complete framework which is applicable interactively and user could choose his own dataset. This feature of the system allows users to utilize their own private data. This addresses one of the existing challenge in Bioinformatics which scientists prefer to do not share their private datasets on other servers. They will be able to install the app and does not Internet connectivity to run the framework. We are going to package visual analytics framework into an iOS mobile application which will be integrated with an application on a multi-touch surface or desktop PC to be available in clinical care settings. Also, I am going to run a comprehensive user study on this platform to survey the thoughts of different clinicians around effective factors on different cancer and then integrate the obtained results to share with the beneficiary communities. Indeed, we will be able to provide a systematic data collection could lead to a better knowledge collection tool.

Ongoing Tasks for Tangible Tensors: The current version of this system has been developed specifically to deal with biological and biochemical systems, the current methodology is Ordinary Differential Equations (ODE), and validation terms are defined based on the obtained RMS errors from the simulation. I am going to add a complete library of different methodologies and validation terms to this system to make it as a general toolkit to be applicable for integrating with various problems. I am going to utilize my novel hybrid neuro fuzzy reconstruction modelling [Manshaei, 2012] as an alternative method in Tangible Tensors (TTs) platform. The reason for this decision is that the logic related to this type of methodologies which use the fuzzy logic, starts with human language rules that could help users to understand and follow the scenario of modeling in TTs platform.

Appendices

System Overview

The entire system is largely web-based, being comprised mostly of HTML, JavaScript, and Cascading Style Sheets for visualizations and interactions. The system uses multi-touch surfaces, alone or in combination, to enable touch and object tracking across the applications. Active tangibles, along with other mobile devices, are used to create additional levels of interactions.

In summary, there are various interactions with many devices, but the system can be broken down into main categories:

- Multi-touch Surfaces
- Tangible Devices
- Communication

The following sections describe each of these categories in detail, along with step-by-step instructions on setting up each component.

A Multi-touch Surfaces

Multi-touch surfaces are touchable surfaces that can recognize multiple points of contact simultaneously. Typical capacitive-based multi-touch surfaces can detect multiple points of finger contact on the surface, but a vision-based multi-touch surface is required to detect the presence of more complex markers. Our applications require the ability to detect both visual markers and finger touches, in any quantity.

These multi-touch surfaces capable of detecting markers and fingers can be custom-built or purchased pre-built, with the latter option being more finely tuned and expensive. If designing a customized surface, the best option is to encapsulate everything within a table, with the multi-touch surface on top. In the following sections, we discuss how to set up purchased pre-built multi-touch surfaces to make them ready for use.

A.1 Purchasing a Multi-Touch Screen

Designing a multi-touch table, although less expensive and more customizable, takes much more planning and setup time and usually requires a dedicated team. The alternative to building your own table is to purchase a pre-built multi-touch screen to serve as the primary tracking system.

For our applications, we have designed infrastructure around the use of [MultiTaction, 2016] cells in a variety of different arrangements. MultiTaction cells have a relatively short setup time compared to a custom designed multi-touch table, and their modularity allows multiple cells to be arranged in customizable configurations. In the following sections, we describe the general process for setting up a MultiTaction platform usable for all multi-touch purposes.

A.1.1 Infrastructure Requirements

Preplanning is needed before purchasing MultiTaction screens, to ensure both structural and electrical requirements are met. The amount of planning needed is based on the complexity of the desired setup. A MultiTaction wall, for example, would require planning for additional electrical breakers and metal framing to physically support the weight of all the screens.

Although there are many possibilities for the framing to support the cell structures, we generally used [Unistrut, 2016] in all designs. This is primarily because of its versatility and structural strength. The specifications of the MultiTaction cells can be found in the MultiTaction Cell 55" manual [MultiTaction, 2016], but the specifications of interest for infrastructure design of each cell are:

- Its weight of 40 kg (90 lbs)
- Its power supply of 100-240 VAC 50/60 Hz and consumption of 450W

A.1.2 Wall Design

The wall we designed uses 12 MultiTaction tiles in portrait orientation, with a 45° bend after every 3 tiles (meaning there are 4 sections of flat 3x1 tiles) (Fig. A.1). To support these, we used 8 vertical beams of Unistrut running from the ground to the ceiling (2 per 3-cell section), which were attached to the screens using 2 horizontal beams running across each flat 3-tiles section.

Each of the 12 MultiTaction Cells draw a maximum of $\frac{450W}{100V} = 4.5A$ from its power supply.

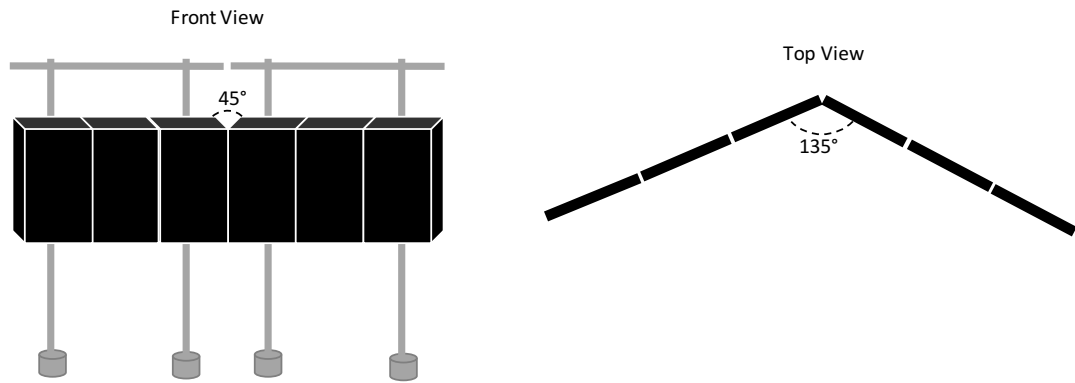


Figure A.1: Front and top views of 3-cell wall section angles

That means on a typical 15 A breaker, a maximum of 3 tiles can be connected (and on a 20 A breaker, 4 can be connected) safely. Because of this, additional electrical wiring will probably be needed for designing such a large-scale platform.

A.1.3 3-tiles Tables

Alongside the MultiTaction wall, we also designed two 3-tiles MultiTaction tables. Similar to the wall, the tables are constructed out of Unistrut beams and fittings, along with wheels to allow easy transportation. Since the breakers supplying the tables allow up to 20 A, the table (and the computer it connects to) can be run entirely on a single breaker. Both the table and wall are shown in Fig. A.2.

The quantity and length of every part we used in our design can be found in the Supplementary Materials 2 spreadsheet.

A.2 Windows Computer for interfacing the MultiTaction Cells

To enable functionality across cells, they must be connected to a single computer. The computer must have at least as many graphical inputs as there are cells, and must have enough RAM



Figure A.2: MultiTaction 3-tiles table and 12-tiles wall designs

and CPU power to run the applications. The MultiTouch Computer Guide [TactionGuide, 2016] has a set of recommendations for the specifications of the computer.

For our 12-cell wall, 3 FirePro W9100 graphics cards were used, along with 64gb of RAM and i7 Intel processor. The computer is running Windows 10 with the MultiTouch Cornerstone SDK. Installing the SDK requires you to login to MultiTouch's download page [TactionGuide, 2016] and run the installer.

A.3 Cell Setup

MultiTouch's Taction guide [TactionGuide, 2016] is the best place to start for understanding how the cells integrate with each other and the computer. Essentially, every cell must be connected to the central computer by both a DVI (video) interface and an Ethernet (data) interface. In our case, the primary computer (with address 10.0.1.100) acted as the server, which accesses each of the screen's addresses through a network switch. The screen's IPs

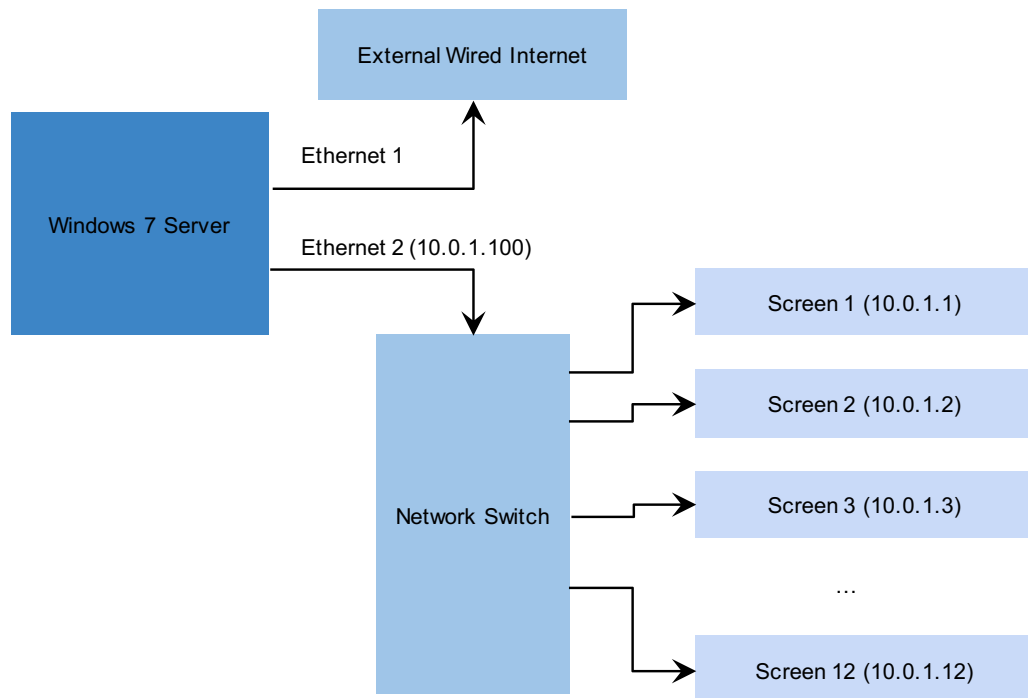


Figure A.3: General network configuration of server and cell communication

are incremental, starting at 10.0.1.1 and go through to however many screens there are. Fig. A.3 shows the general setup of how our devices are named, which later dictates how they communicate.

Some important settings need to be altered in the cell's firmware to enable this communication. Firstly, 4x4 markers need to be enabled for each of the cells. This can be done by navigating to Calibration > Markers and clicking 4x4 (in the top left). 'Save' must also be clicked after every change made (Fig. A.4).

Secondly, the cell must be assigned its own network configuration. This is done by navigating to Setup > Network and assigning the cell an Address, Netmask, Gateway, and DNS. In our setup, the central computer had an IP of 10.0.1.100, so we use that as both the DNS and



Figure A.4: Enabling 4x4 markers on each cell

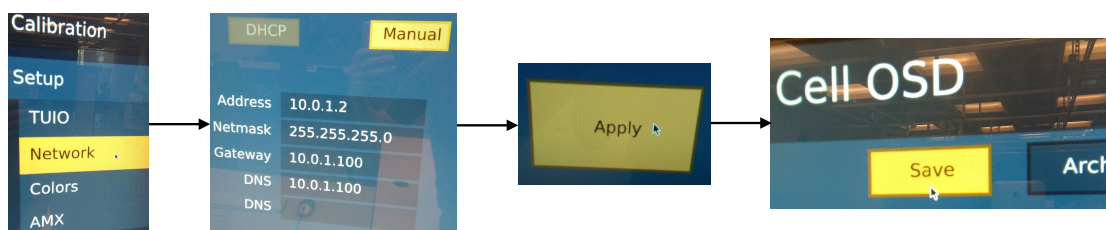


Figure A.5: Configuring the cell's network address

Gateway. By default, the Netmask is 255.255.255.0. The address changes with each cell, but could be incremental for each adjacent cell (i.e., cells from left to right could have addresses from 10.0.1.1 to 10.0.1.3) (Fig. A.5).

Finally, since all cells now point to 10.0.1.100 as their gateway, the computer needs to be assigned as 10.0.1.100 over the Ethernet port plugged into the cell network. This can be done by navigating to the Network and Sharing Center, clicking on Change adapter settings (in the left panel), right-clicking the Ethernet connection and selecting Properties, double clicking on Internet Protocol Version 4 (TCP/IPv4), modify the configuration like Fig. A.6.

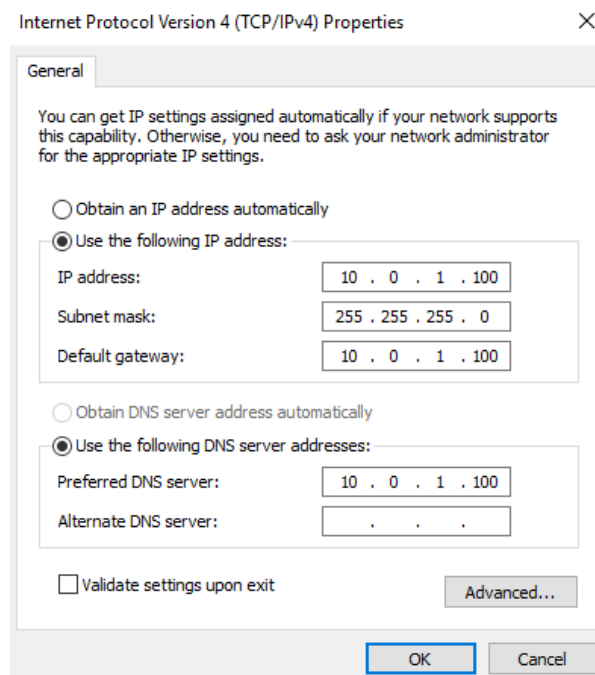


Figure A.6: Configuring the server's network address

B Tangible Devices

Tangible devices, in this context, refer to both our custom-designed active tangibles as well as any mobile device interfacing with the system. Both types of devices are used as a basis for an additional level of interaction with the platform. Both also rely on the use of a web browser for mobile applications, and thus they must be connected to the same network as the server.

B.1 Active Tangibles

The 'active tangible' is defined as the wireless interactive input device used to manipulate the GUI in our projects.

The active tangible consists of three major components; the case, smartwatch and active tangible printed circuit board (PCB) (Figs. B.1, B.2). The smart watch is embedded in the case and serves as the active tangible's graphical interface, which is a browser menu system. The smartwatch and the active tangibles' custom electronics are communicatively independent, as the smartwatch only relies on power from the PCB. The custom electronics PCB on the active tangible consists of six major components:

- LED Light Ring

Appendix B. Tangible Devices

- Wifi Communication
- Microcontroller
- Magnetic (Hall Effect) Sensors
- Power supervisory, regulation and charging
- Accelerometer

The LED light ring serves to provide feedback to the user while manipulating data. The LED feedback instructions are provided wirelessly by the server via the Wi-Fi module. Pre-loaded LED light ring commands are programmed into the microcontroller to allow for simple instructions from the server. A 2300 mAh galaxy S3 battery is used to provide up to 3 hours of continuous use, and can be charged using a USB cable. The smartwatch and the custom electronics are turned off and on using a single tactile button on the side of the active tangible. The eight Hall Effect sensors are staggered along the PCB in such a way, that the active tangible can detect other active tangibles in close proximity. Magnets are installed in the case of the active tangibles so interactions from the side, as well as stacking interactions are possible with multiple active tangibles. Each active tangible will send to the server which sensor is tripped, and the server determines which two tangibles are connected. An accelerometer is implemented to allow for unique movement events, as well as the capability to turn off or on an active tangible by a shaking movement.

The active tangibles can be re-programmed to customize their use via USB and an Arduino IDE.

B.1.1 Connecting Smartwatches to the Network

The LG Smartwatches are Android based, and require an Android-OS smartphone to tether to via Bluetooth, so that the smart phone's Wi-Fi can be streamed to the smartwatch. This can be an involved process, but the general steps for doing so are outlined below:

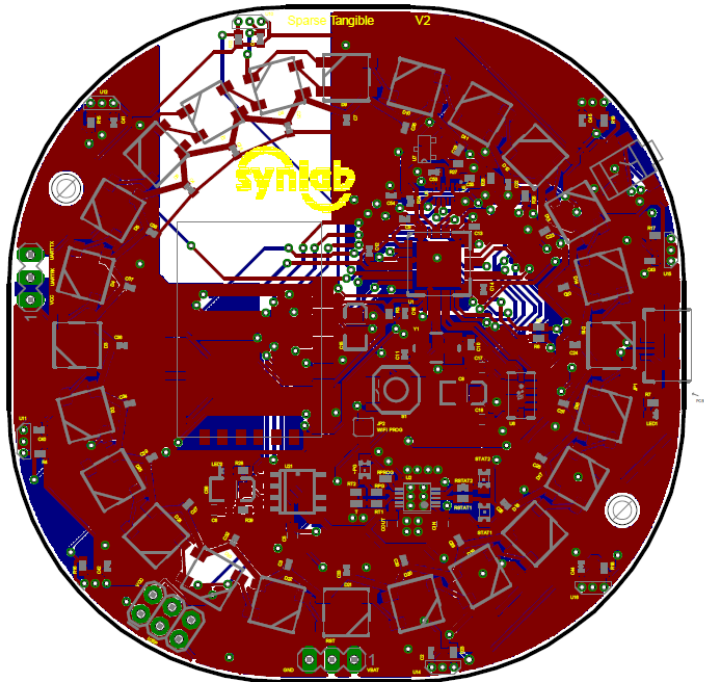


Figure B.1: PCB design of the active tangible

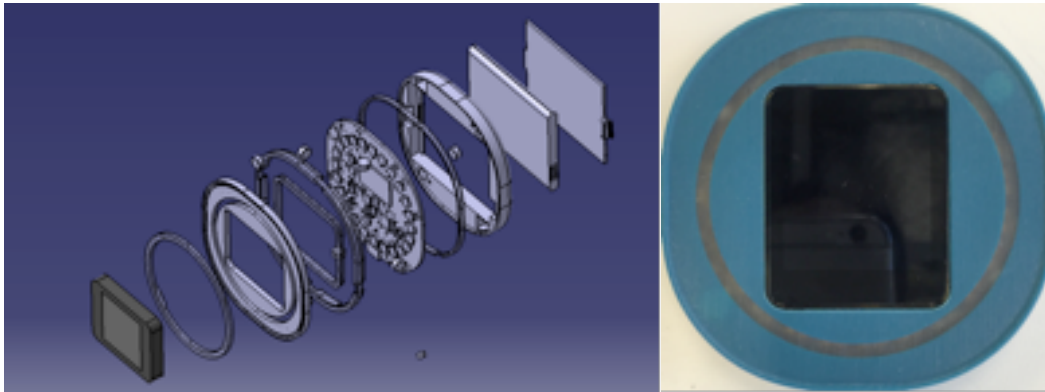


Figure B.2: Case design and final 3D print of the active tangible

Appendix B. Tangible Devices

- The Android Wear app must be installed through the Google Play store on the smartphone
- Sync the smartphone with the smartwatch through Android wear
- Update the smartwatch firmware through its system settings.
This may have to be done multiple times, as the smartwatch goes through different stages of updates
- The Wearable Internet Browser (WIB) app needs to be purchased from the Google Play store
- Install WIB on the smartphone
- While still synced, open up the WIB app on the smartphone.
The smartphone should now try and sync the WIB app to the smartwatch
If it doesn't, uninstall the WIB app and repeat from step 5
- On the smartphone WIB app, bookmark the mobile menu application. It might have a URL like: <serverIP/projectname/mobile_menu.html>
- The bookmark should sync across to the smartwatch, and can be opened through WIB

If at any point along the process the smartwatch needs to be resynced to any phone, the smartwatch needs to be factory reset. The updates will persist after the reset, but all apps will be wiped.

C Communication

The communication between the server and external devices is crucial for the functionality of the system. All communication is done through Node.js and TCP/UDP communication, with the general flow of information outlined in Fig. C.1.

The central Windows computer has two servers running simultaneously: one to capture TUIO multitouch events from the surface (this server could be Community Core Vision or MTServer.exe), and one to handle every other event and send information to connected clients (the central Node.js server). The Node.js server then communicates to each client over socket.io, causing the application across all devices respond to user input.

C.1 Required Software

All of our projects are run as a web-based system, being comprised of multiple HTML pages, Cascading Style Sheets (CSS) and JavaScript scripts. Thus, they need an HTTP server to operate properly- we use XAMPP to set up a basic local Apache HTTP server for running the system. Furthermore, the projects rely on client-server interactions (such as menu navigation in a separate webpage) to support a wider variety of interactions. Node.js, along with node modules such as socket.io, are used to create the server-client communication through the use of a node server.

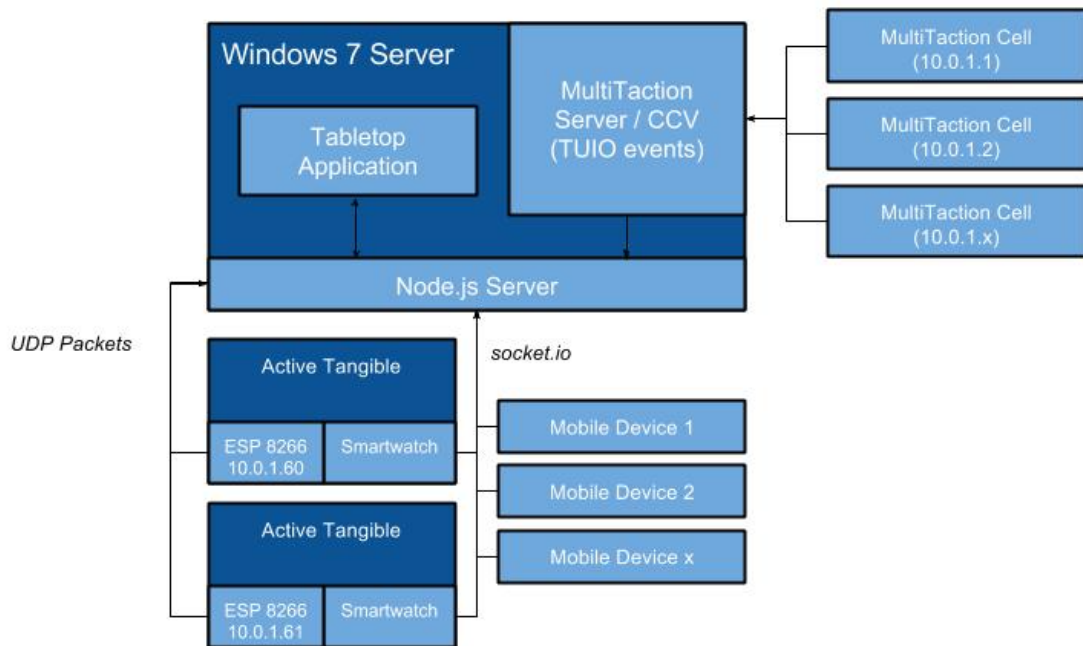


Figure C.1: General flow of application communication

C.1.1 Installing XAMPP Control Panel

XAMPP is an Apache distribution with a simple GUI for quickly creating a local webserver. The following outlines the steps to install XAMPP and get a webserver running:

- Download the XAMPP installer from their Homepage [XAMPP, 2016]:
- Run the installer executable, and click next until it asks for a folder install directory. The default directory used is <C:/xampp> for Windows, but if you want to change the installation to a folder elsewhere, make note of where you install it.
- Click next until it begins the setup, and wait for it to finish the installation. Click finish when the installer is done the setup.
- Close the installer without running the XAMPP control panel

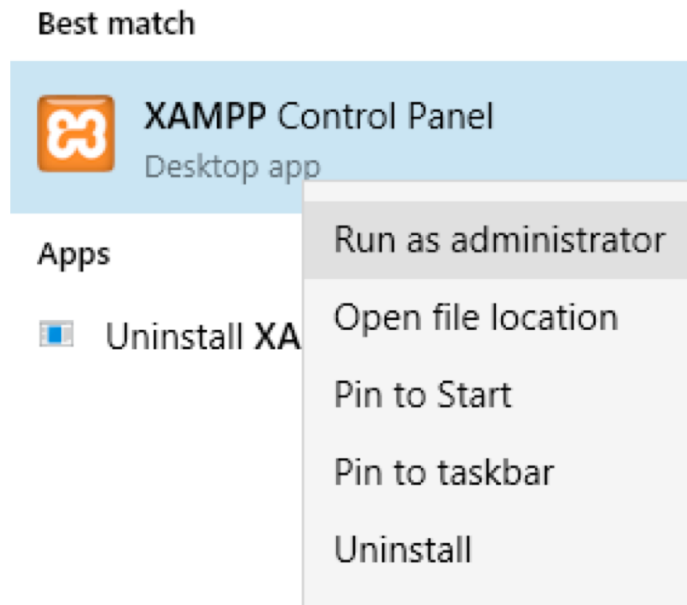


Figure C.2: Running XAMPP control panel as administrator

C.1.2 Installing Apache through XAMPP

To install and run the Apache HTTP hosting service,

- Search for 'XAMPP' in the start menu
- Right click 'XAMPP Control Panel' and Run it as administrator (Fig. C.2)
- Allow XAMPP through the network (if a Windows Firewall notification pops up)
- Click the 'x' next to the Apache module to install it (Fig. C.3)

Now, when 'localhost' is typed into any browser, a default XAMPP index page should be loaded. This means that the Apache server is running, and now any files (such as .html or .php) in <C:/xampp/htdocs> can be accessed through the HTTP protocol. At this point, any project source placed in the <C:/xampp/htdocs> (Windows) or </Applications/XAMPP/xamppfiles/htdocs> (OS X) directory can be run. However, the projects still require a Node.js server to be fully functional.

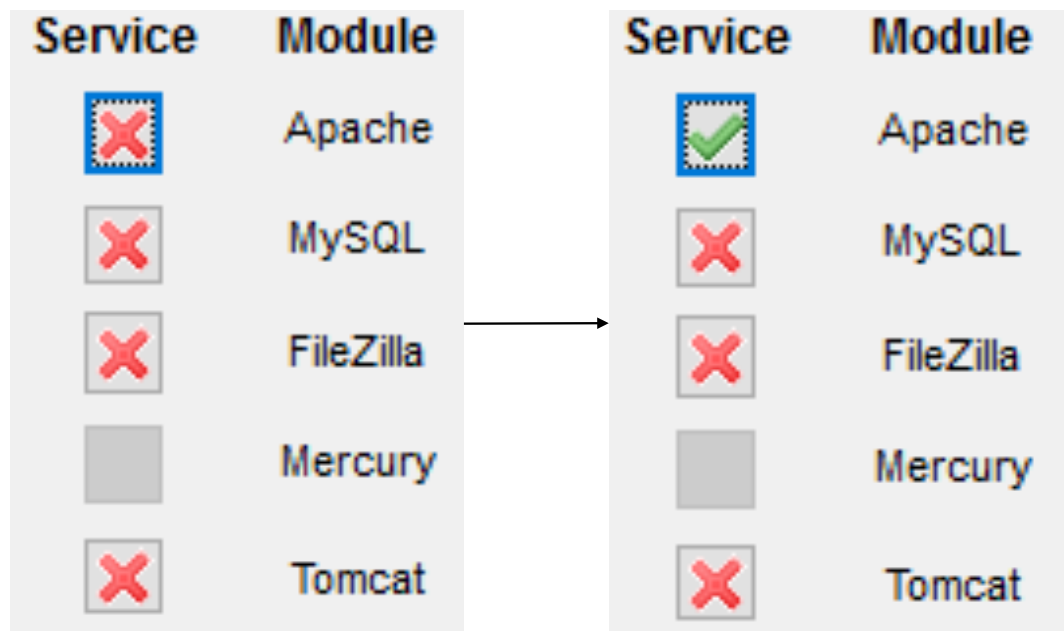


Figure C.3: Installing the Apache service

C.1.3 Installing Node.js

Node.js first needs to be downloaded from Node.js's website [Node.js, 2016] (both versions will work, but v4.4.5 LTS is better for long-term compatibility). Execute the .msi installer file, and proceed through the installation (choosing a desired directory, typically <C:/Program> Files (x86)/nodejs, for Node.js). After Node.js is installed, the server can now be run. All required modules should be included in the compressed TCB server file.

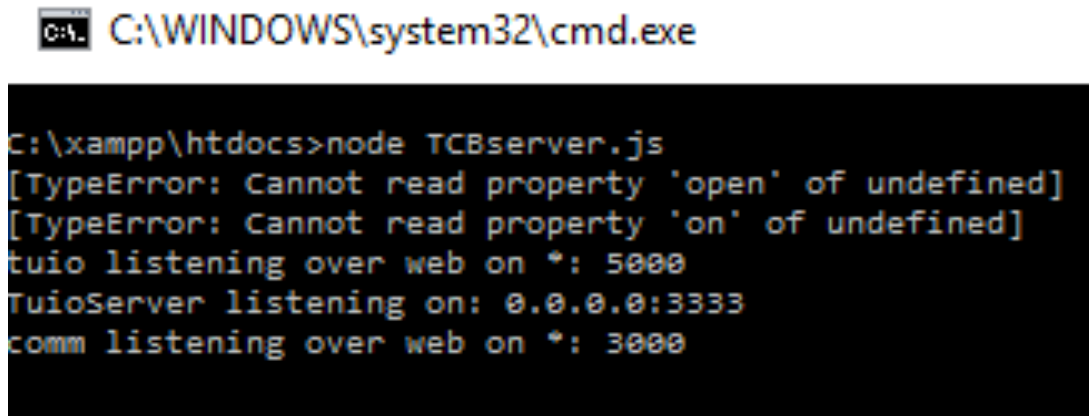
C.1.4 Running the Node.js TCB Server

To run the server, extract the contents of the .zip folder to any directory, and double click the runTCB.cmd script. Alternatively, you can open up the command prompt, and run the command:

```
node TCBserver.jsx
```

The command window might look something like Fig. C.4.

Once the server is running, all projects should be fully functional, with the exception of some



```
C:\WINDOWS\system32\cmd.exe

C:\xampp\htdocs>node TCBserver.js
[TypeError: Cannot read property 'open' of undefined]
[TypeError: Cannot read property 'on' of undefined]
tuio listening over web on *: 5000
TuioServer listening on: 0.0.0.0:3333
comm listening over web on *: 3000
```

Figure C.4: Node.js server running

mobile device communication (discussed next).

C.1.5 Changing Mobile Menu socket.io Addresses

In order for the project mobile menu applications to connect to the Node.js server previously setup, the one line of the project mobile menu scripts needs to be modified, to point to the IP address on which the server is running. Modifying the mobile IP's can be done by:

- Locating the `mobile_menu.js` file, which can be found in:
<projectname/scripts/mobile_menu.js>
- Finding your server's local IP address. This can be done in Windows by opening up a cmd terminal (Shift+RightClick in any window, and Open command window here) and typing in 'ipconfig'.
Your IP for every connected network will be displayed here, but you need to find the IPv4 address for the network to which both your server and mobile device will be connected. A connectify hotspot might be useful for setting this network.
- Open up the `mobile_menu.js` file, and replace the IP in the following line:
`var socket = io.connect('192.168.1.1:3000');`

Appendix C. Communication

To the address you found in step 2.

- Save the `mobile_menu.js` file with the changed IP

Now, whenever the mobile menu is loaded on any device, it should connect to the Node.js server through `socket.io`.

Bibliography

- [Abowd *et al*, 1999] Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999, September). Towards a better understanding of context and context-awareness. In International Symposium on Handheld and Ubiquitous Computing (pp. 304-307). Springer Berlin Heidelberg.
- [Aggarwal, 2014] Aggarwal, C. C. (Ed.). (2014). Data classification: algorithms and applications. CRC Press.
- [Alexander, 1996] Alexander, R. M. (1996). Optima for animals. Princeton University Press.
- [Alon, 2007] Alon, U. (2006). An introduction to systems biology: design principles of biological circuits. CRC press.
- [Alvarez-Socorro 2015] Alvarez-Socorro, A. J., Herrera-Almarza, G. C., & González-Díaz, L. A. (2015). Eigencentality based on dissimilarity measures reveals central nodes in complex networks. Scientific reports, 5.
- [Andrienko, 2006] Andrienko, N., & Andrienko, G. (2006). Exploratory analysis of spatial and temporal data: a systematic approach. Springer Science & Business Media.
- [Arif *et al*, 2016] Arif, A. S., Manshaei, R., Delong, S., East, B., Kyan, M., & Mazalek, A. (2016, February). Sparse Tangibles: Collaborative Exploration of Gene Networks using Active Tangibles and Interactive Tabletops. In Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction (pp. 287-295). ACM.

Bibliography

- [Astrazeneca, 2016] <http://www.astrazeneca.ca/en/Home/>
- [Azuma *et al*, 2006] Azuma, R., Daily, M., & Furmanski, C. (2006, July). A review of time critical decision making models and human cognitive processes. In 2006 IEEE aerospace conference (pp. 9-pp). IEEE.
- [Baber *et al*, 2016] Baber, C., Attfield, S., Conway, G., Rooney, C., & Kodagoda, N. (2016). Collaborative sense-making during simulated Intelligence Analysis Exercises. *International Journal of Human-Computer Studies*, 86, 94-108.
- [Bakker *et al*, 2016] Bakker, S., Hausen, D., & Selker, T. (Eds.). (2016). *Peripheral Interaction: Challenges and Opportunities for HCI in the Periphery of Attention*. Springer.
- [Baldauf, 2007] Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2(4), 263-277.
- [Barabasi, 2004] Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2), 101-113.
- [Barabási, 2015] Barabási, A. (2015). *Network Science*. Cambridge University. Retrieved July 21, 2015 from <http://www.barabasi.com/networksciencebook>
- [Baskinger and Gross, 2010] Baskinger, M., & Gross, M. D. (2010). Cover Story-Tangible interaction= form+ computing. *interactions*, 17(1), 6-11.
- [Beaney, 2013] Beaney, M. (Ed.). (2013). *The Oxford handbook of the history of analytic philosophy*. OUP Oxford.
- [Bellotti *et al*, 2002] Bellotti, V., Back, M., Edwards, W. K., Grinter, R. E., Henderson, A., & Lopes, C. (2002, April). Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 415-422). ACM.
- [Bergner *et al*, 2011] Bergner, S., Crider, M., Kirkpatrick, A. E., & Möller, T. (2011). Mixing board versus mouse interaction in value adjustment tasks. *arXiv preprint arXiv:1110.2520*.

- [Berman *et al*, 2015] Berman, H.M., Gabanyi, M.J., Groom, C.R., Johnson, J.E., Murshudov, G.N., Nicholls, R.A., Reddy, V., Schwede, T., Zimmerman, M.D., Westbrook, J. and Minor, W., (2015). Data to knowledge: how to get meaning from your result. *IUCrJ*, 2(1), 45-58.
- [Bertolucci, 2013] Bertolucci, J. (2013). Big data analytics: Descriptive vs. predictive vs. prescriptive. *Information Week*.
- [Beyer, 2007] Beyer, A., Bandyopadhyay, S., & Ideker, T. (2007). Integrating physical and genetic maps: from genomes to interaction networks. *Nature Reviews Genetics*, 8(9), 699-710.
- [Birkhofer, 2011] Birkhofer, H. (2011). *The future of design methodology*. London: Springer.
- [Blevins, 2016] Blevins, M., Wehbe, F. H., Rebeiro, P. F., Caro-Vega, Y., McGowan, C. C., & Shepherd, B. E. (2016). Interactive Data Visualization for HIV Cohorts: Leveraging Data Exchange Standards to Share and Reuse Research Tools. *PloS one*, 11(3), e0151201.
- [Böck, 2012] Böck, M., Ogishima, S., Tanaka, H., Kramer, S., & Kaderali, L. (2012). Hub-centered gene network reconstruction using automatic relevance determination. *PLoS One*, 7(5), e35077.
- [Boy *et al*, 2016] Boy, J., Eveillard, L., Detienne, F., & Fekete, J. D. (2016). Suggested Interactivity: Seeking Perceived Affordances for Information Visualization. *IEEE transactions on visualization and computer graphics*, 22(1), 639-648.
- [Boyd, 2004] Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.
- [Boyd, 2013] Boyd S. P. (2013). *l1-norm methods for convex cardinality problems*. Lecture Notes for EE364b Stanford University.
- [Brave *et al*, 1998] Brave, S., Ishii, H., & Dahley, A. (1998, November). Tangible interfaces for remote collaboration and communication. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (pp. 169-178). ACM.

Bibliography

- [Brehmer *et al*, 2016] Brehmer, M. M. (2016). Why visualization?: task abstraction for analysis and design (Doctoral dissertation, University of British Columbia).
- [Brill-Schuetz, 2014] Brill-Schuetz, K. A., & Morgan-Short K. (2014) The Role of Procedural Memory in Adult Second Language Acquisition. Annual Meeting of the Cognitive Science Society (COGSCI).
- [Brown, 2009] Brown, K.R., Otasek, D., Ali, M., McGuffin, M.J., Xie, W., Devani, B., van Toch, I.L. & Jurisica, I., (2009). NAViGaTOR: network analysis, visualization and graphing Toronto. *Bioinformatics*, 25(24), 3327-3329.
- [Brier, 1993] Brier, J. T. (1993). The mind's journey from novice to expert. *American Educator*, 17(2), 6-15.
- [Bull, 2005] Bull, G. (2005). Children, computers, and powerful ideas. *Contemporary Issues in Technology and Teacher Education*, 5(3/4), 349-352.
- [Burbeck, 1992] Burbeck, S. (1992). Applications programming in smalltalk-80 (tm): How to use model-view-controller (mvc). *Smalltalk-80 v2*, 5.
- [Butler, 2008] Butler, S. K. (2008). Eigenvalues and structures of graphs. ProQuest.
- [Buuren and Groothuis-Oudshoorn, 2011] Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3).
- [Buxton and Myers, 1986] Buxton, W., & Myers, B. (1986, April). A study in two-handed input. In *ACM SIGCHI Bulletin* (Vol. 17, No. 4, pp. 321-326). ACM.
- [Bybee *et al*, 2005] Bybee, R., Bloom, M., Phillips, J., & Knapp, N. (2005). *Doing Science: The process of Scientific Inquiry*. Center for Curriculum Development. Colorado Springs, CO.
- [Camarata, 2002] Camarata, K., Do, E. Y. L., Johnson, B. R., & Gross, M. D. (2002, January). Navigational blocks: navigating information space with tangible media. In *Proceedings of the 7th international conference on Intelligent user interfaces* (pp. 31-38). ACM.

- [Candes *et al*, 2011] Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM (JACM)*, 58(3), 11.
- [Carandang and Campbell, 2013] Carandang, X., & Campbell, J. (2013). The Design of a Tangible User Interface for a Real-Time Strategy Game.
- [Carlson and Doyle, 2002] Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences*, 99(suppl 1), 2538-2545.
- [Carpendale, 2008] Carpendale, S. (2008). Evaluating information visualizations. In *Information Visualization* (pp. 19-45). Springer Berlin Heidelberg.
- [Cecilio Angulo, 2016] <http://people-esaii.upc.edu/people/cangulo/>
- [Celgene Canada 2016] www.celgenecanada.net/en/index.aspx
- [Chandrasekharan, 2009] Chandrasekharan, S. (2009). Building to discover: a common coding model. *Cognitive Science*, 33(6), 1059-1086.
- [Chatr-Aryamontri, 2015] Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'donnell, L. & Reguly, T., (2015). The BioGRID interaction database: 2015 update. *Nucleic acids research*, 43(D1), D470-D478.
- [Chen *et al*, 2014] Chen, C. H., Hsieh, J. T., Huang, K. H., Pu, Y. S., & Chang, H. C. (2014). Predictive clinical indicators of biochemical progression in advanced prostate cancer patients receiving Leuplin depot as androgen deprivation therapy. *PloS one*, 9(8), e105091.
- [Chen *et al*, 2015] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, 2015, 12.
- [Chen *et al*, 2010] Chen, M., Gonzalez, S., Leung, V., Zhang, Q., & Li, M. (2010). A 2G-RFID-based e-healthcare system. *IEEE Wireless Communications*, 17(1), 37-43.

Bibliography

- [Chen, 2014] Chen, M. (2014). NDNC-BAN: supporting rich media healthcare services via named data networking in cloud-assisted wireless body area networks. *Information Sciences*, 284, 142-156.
- [Chi *et al*, 2015] Chi, K.N., Kheoh, T., Ryan, C.J., Molina, A., Bellmunt, J., Vogelzang, N.J., Rathkopf, D.E., Fizazi, K., Kantoff, P.W., Li, J. & Azad, A.A., (2015). A prognostic index model for predicting overall survival in patients with metastatic castration-resistant prostate cancer treated with abiraterone acetate after docetaxel. *Annals of Oncology*, mdv594.
- [Cho, 1998] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. & Davis, R.W., (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 2(1), 65-73.
- [Cho, 2007] Cho, K. H., Choo, S. M., Jung, S. H., Kim, J. R., Choi, H. S., & Kim, J. (2007). Reverse engineering of gene regulatory networks. *IET Systems Biology*, 1(3), 149-163.
- [Cho, 2016] Cho, H., Berger, B., & Peng, J. (2016). Reconstructing Causal Biological Networks through Active Learning. *PloS one*, 11(3), e0150611.
- [Christian Bauckhage, 2016] <https://www.iais.fraunhofer.de/de/institut/mitarbeiterprofile/christian-bauckhage.html>
- [Chung, 1996] Chung, F. (1996). Discrete Isoperimetric Inequalities. In *DMTCS* (pp. 24-24).
- [Costanzo, 2000] Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C. & Roberg-Perez, K.J., (2000). The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic acids research*, 28(1), 73-76.
- [Cox D.R., 1972] David, C. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34, 187-220.

- [Cruz Mendoza, 2015] Cruz Mendoza, R., Bianchi-Berthouze, N., Romero, P., & Lavín, G. C. (2015). A classification of user experience frameworks for movement-based interaction design. *The Design Journal*, 18(3), 393-420.
- [Cvetkovic and Gutman, 2009] Cvetkovic, D. M., & Gutman, I. (Eds.). (2009). *Applications of graph spectra*. Matematički institut SANU.
- [Cvetkovic and Gutman, 2011] Cvetković, D. M., & Gutman, I. (2011). *Selected topics on applications of graph spectra*. Beograd: Matematički institut SANU.
- [Daniels *et al*, 2013] Daniels, H., Edwards, A., Engeström, Y., Gallagher, T., & Ludvigsen, S. R. (Eds.). (2013). *Activity theory in practice: Promoting learning across boundaries and agencies*. Routledge.
- [Dang, 2015] Dang, T. N., Murray, P., Aurisano, J., & Forbes, A. G. (2015, August). Reaction-Flow: an interactive visualization tool for causality analysis in biological pathways. In *BMC proceedings* (Vol. 9, No. 6, p. 1). BioMed Central.
- [De Robertis *et al*, 2014] DE ROBERTIS, E. M. Deciphering Complexity in Biology: Induction of Embryonic Cell Differentiation by Morphogen Gradients. *Complexity and Analogy in Science: Theoretical, Methodological and Epistemological Aspects*, 161.
- [De Smet & Marchal, 2010] De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10), 717-729.
- [Dehaene, 2014] Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.
- [Deiros *et al*, 2016] Deiros, D. R., Gibbs, R. A., & Rogers, J. (2016). DNAism: exploring genomic datasets on the web with Horizon Charts. *BMC bioinformatics*, 17(1), 1.
- [Dervin, 1992] Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Qualitative research in information management*, 9, 61-84.

Bibliography

- [Dewey, 1938] Dewey, J. (1938). The theory of inquiry. New York: Holt, Rinehart & Wiston.
- [Dey, 2001] Dey, A. K. (2001). Understanding and using context. Personal and ubiquitous computing, 5(1), 4-7.
- [Dondelinger, 2012] Dondelinger, F., Husmeier, D., & Lèbre, S. (2012). Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. Euphytica, 183(3), 361-377.
- [Dörk *et al*, 2011] Dörk, M., Carpendale, S., & Williamson, C. (2011, May). The information flaneur: A fresh look at information seeking. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1215-1224). ACM.
- [Dourish, 2004] Dourish, P. (2004). Where the action is: the foundations of embodied interaction. MIT press.
- [Doyle, 2006] Doyle, F. J., & Stelling, J. (2006). Systems interface biology. Journal of the Royal Society Interface, 3(10), 603-616.
- [Dream Challenge 9.5] <https://www.synapse.org/#!Synapse:syn2813558/wiki/70844>
- [Du 2010] Du, W., Li, X., & Li, Y. (2010). The Laplacian energy of random graphs. Journal of Mathematical Analysis and Applications, 368(1), 311-319.
- [Duan *et al*, 2011] Duan, L., Street, W. N., & Xu, E. (2011). Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterprise Information Systems, 5(2), 169-181.
- [E. Van Den Hoven *et al*, 2007] Van Den Hoven, E., Frens, J., Aliakseyeu, D., Martens, J. B., Overbeeke, K., & Peters, P. (2007, February). Design research & tangible interaction. In Proceedings of the 1st international conference on Tangible and embedded interaction (pp. 109-115). ACM.
- [Edge and Blackwell, 2006] Edge, D., & Blackwell, A. (2006). Correlates of the cognitive dimensions for tangible user interface. Journal of Visual Languages & Computing, 17(4), 366-394.

- [Elliott *et al*, 2015] Elliott, J.H., Grimshaw, J., Altman, R., Bero, L., Goodman, S.N., Henry, D., Macleod, M., Tovey, D., Tugwell, P., White, H. & Sim, I., (2015). Informatics: Make sense of health data. *Nature*, 527(7576), 31.
- [Emmert-Streib, 2012] Emmert-Streib, F., Glazko, G., & De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in genetics*, 3, 8.
- [Endsley, 1995] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64.
- [Ernst and Young Global Limited, 2014] Ernst & Young Global Limited (2014). Big Data: Changing the way businesses compete and operate.
- [euCognition, 2016] <http://www.vernon.eu/euCognition/definitions.htm>
- [Fatemieh *et al*, 2010] Fatemieh, O., Chandra, R., & Gunter, C. A. (2010, April). Secure collaborative sensing for crowd sourcing spectrum data in white space networks. In *New Frontiers in Dynamic Spectrum*, 2010 IEEE Symposium on (pp. 1-12). IEEE.
- [Fernaes *et al*, 2008] Fernaeus, Y., Tholander, J., & Jonsson, M. (2008, February). Towards a new set of ideals: consequences of the practice turn in tangible interaction. In *Proceedings of the 2nd international conference on Tangible and embedded interaction* (pp. 223-230). ACM.
- [Fernaes *et al*, 2008] Fernaeus, Y., Tholander, J., & Jonsson, M. (2008). Beyond representations: Towards an action-centric perspective on tangible interaction. *International Journal of Arts and Technology*, 1(3-4), 249-267.
- [Feuer *et al*, 2014] Feuer, E.J., Rabin, B.A., Zou, Z., Wang, Z., Xiong, X., Ellis, J.L., Steiner, J.F., Cynkin, L., Nekhlyudov, L., Bayliss, E. & Hankey, B.F., (2014). The Surveillance, Epidemiology, and End Results Cancer Survival Calculator SEER* CSC: Validation in a Managed Care Setting. *Journal of the National Cancer Institute. Monographs*, 2014(49), 265.

Bibliography

- [Fitzmaurice *et al*, 1995] Fitzmaurice, G. W., Ishii, H., & Buxton, W. A. (1995, May). Bricks: laying the foundations for graspable user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 442-449). ACM Press/Addison-Wesley Publishing Co. .
- [Fitzmaurice, 1996] Fitzmaurice, G. W. (1996). Graspable user interfaces (Doctoral dissertation, University of Toronto).
- [Fitzmaurice and Buxton, 1997] Fitzmaurice, G. W., & Buxton, W. (1997, March). An empirical evaluation of graspable user interfaces: towards specialized, space-multiplexed input. In Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (pp. 43-50). ACM.
- [Flyckt, 2013] Flyckt, M. (2013). Cubieo: Ambiguity in Tangible Collaborative User Interfaces.
- [Franklin, 2015] Franklin, C. Space-Time Diffusion Visualization using Bayesian Inference, 2015 Pre-Conference Workshop: 36th International Conference on Information Systems.
- [Fu, 2011] Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164-181.
- [Furtado De Mendonca Monco, 2015] Furtado De Mendonca Monco, E. (2015). From head to toe: body movement for human-computer interaction (Doctoral dissertation).
- [Gadkar, 2005] Gadkar, K. G., Gunawan, R., & Doyle, F. J. (2005). Iterative approach to model identification of biological networks. *BMC bioinformatics*, 6(1), 1.
- [Gaines and Monk 2015] Gaines, B. R., & Monk, A. F. (2015). *Cognitive Ergonomics: Understanding, Learning, and Designing Human-Computer Interaction*. P. Falzon (Ed.). Academic Press.
- [Gandomi, 2016] Gandomi, A. H., Sajedi, S., Kiani, B., & Huang, Q. (2016). Genetic programming for experimental big data mining: A case study on concrete creep formulation. *Automation in Construction*, 70, 89-97.

- [Geiss *et al*, 2008] Geiss, G.K., Bumgarner, R.E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D.L., Fell, H.P., Ferree, S., George, R.D., Grogan, T. & James, J.J., (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3), 317-325.
- [Gelineck *et al*, 2013] Gelineck, S., Overholt, D., Büchert, M., & Andersen, J. (2013). Towards an interface for music mixing based on smart tangibles and multitouch. *NIME 2013*.
- [Gelman and Hill, 2006] Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- [GeneticsRef., 2016] "Genetics Home Reference, Your Guide To Understanding Genetic Conditions". Genetics Home Reference. N.p., 2016. Web. 26 Oct. 2016.
- [George, 2014] George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology*, 21(4), 686-694.
- [Ghavamian] Ghavamian, R. *Prostate Cancer Treatment Protocols*.
- [Gheisari and Irizarry, 2011] Gheisari, M., & Irizarry, J. (2011). Investigating facility managers' decision making process through a situation awareness approach. *International Journal of Facility Management*, 2(1).
- [Goldin-Meadow and Wagner, 2014] Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in cognitive sciences*, 9(5), 234-241.
- [Golub, 2013] Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (Vol. 3). JHU Press.
- [Gonzalez and Wismisberg, 2007] Gonzalez, C., & Wismisberg, J. (2007). Situation awareness in dynamic decision making: Effects of practice and working memory. *Journal of Cognitive Engineering and Decision Making*, 1(1), 56-74.
- [Gravis *et al*, 2015] Gravis, G., Boher, J.M., Fizazi, K., Joly, F., Priou, F., Marino, P., Latorzeff, I., Delva, R., Krakowski, I., Laguerre, B. & Walz, J., (2015). Prognostic factors for survival in

Bibliography

- noncastrate metastatic prostate cancer: validation of the glass model and development of a novel simplified prognostic model. *European urology*, 68(2), 196-204.
- [Green, 2008] Green, R. S. (2008). Cognitive task analyses for life science automation training program design. ProQuest.
- [Groome, 2016] Groome, D., & Eysenck, M. (2016). An introduction to applied cognitive psychology. Psychology Press.
- [Grün and Slate, 2012] Grün, B., Scharl, T., & Leisch, F. (2012). Modelling time course gene expression data with finite mixtures of linear additive models. *Bioinformatics*, 28(2), 222-228.
- [Gu *et al*, 2014] Gu, S., Zhang, L., Zuo, W., & Feng, X. (2014). Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2862-2869).
- [Gulliksson, 2012] Gulliksson, H. (2012). Human-information-thing interaction: technology and design.
- [Gupta *et al*, 2014] Gupta, E., Guthrie, T., & Tan, W. (2014). Changing paradigms in management of metastatic Castration Resistant Prostate Cancer (mCRPC). *BMC urology*, 14(1), p.55.
- [Gutman, 2001] Gutman, I. (2001). The energy of a graph: old and new results. In *Algebraic combinatorics and applications* (pp. 196-211). Springer Berlin Heidelberg.
- [Gutman, 2006] Gutman, I. (2006). Chemical graph theory - The mathematical connection. *Advances in Quantum Chemistry*, 51, 125-138.
- [Gutman, 2009] Gutman, I., Li, X., & Zhang, J. (2009). Graph energy. *Analysis of Complex Networks: From Biology to Linguistics*, 145-174.
- [Gutzwiller and Clegg, 2013] Gutzwiller, R. S., & Clegg, B. A. (2013). The role of working memory in levels of situation awareness. *Journal of Cognitive Engineering and Decision Making*, 7(2), 141-154.

- [Hadwiger *et al*, 2012] Hadwiger, M., Beyer, J., Jeong, W. K., & Pfister, H. (2012). Interactive volume exploration of petascale microscopy data streams using a visualization-driven virtual memory approach. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2285-2294.
- [Halabi *et al*, 2014] Halabi, S., Lin, C.Y., Kelly, W.K., Fizazi, K.S., Moul, J.W., Kaplan, E.B., Morris, M.J. & Small, E.J., (2014). Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology*, 32(7), 671-677.
- [Hamel, 2011] Hamel, L. H. (2011). *Knowledge discovery with support vector machines* (Vol. 3). John Wiley & Sons.
- [Harrell, 2015] Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.
- [Hashem *et al*, 2015] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98-115.
- [Hauptmann, 1989] Hauptmann, A. G. (1989, March). Speech and gestures for graphic image manipulation. In *ACM SIGCHI Bulletin* (Vol. 20, No. SI, pp. 241-245). ACM.
- [Heer and Robertson, 2007] Heer, J., & Robertson, G. (2007). Animated transitions in statistical data graphics. *IEEE transactions on visualization and computer graphics*, 13(6), 1240-1247.
- [Hinckley *et al*, 1994] Hinckley, K., Pausch, R., Goble, J. C., & Kassell, N. F. (1994, April). Passive real-world interface props for neurosurgical visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 452-458). ACM.
- [Hinckley *et al*, 1997] Hinckley, K., Pausch, R., Proffitt, D., Patten, J., & Kassell, N. (1997, March). Cooperative bimanual action. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 27-34). ACM.

Bibliography

- [Hinckley, 2014] Hinckley, K., Jacob, R. J., Ware, C., Wobbrock, J. O., & Wigdor, D. (2014). Input/Output Devices and Interaction Techniques.
- [Hlawitschka *et al*, 2014] Hlawitschka, M., Hotz, I., Kratz, A., Marai, G.E., Moreno, R., Scheuermann, G., Stommel, M., Wiebel, A. & Zhang, E., (2014). Top Challenges in the Visualization of Engineering Tensor Fields. In Visualization and Processing of Tensors and Higher Order Descriptors for Multi-Valued Data (pp. 3-15). Springer Berlin Heidelberg.
- [Hornecker and Buur, 2006] Hornecker, E., & Buur, J. (2006, April). Getting a grip on tangible interaction: a framework on physical space and social interaction. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 437-446). ACM.
- [Hornecker, 2012] Hornecker, E. (2012, February). Beyond affordance: tangibles' hybrid nature. In Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (pp. 175-182). ACM.
- [Hoyt *et al*, 2016] Hoyt, R., Linnville, S., Thaler, S., & Moore, J. (2016). Digital Family History Data Mining with Neural Networks: A Pilot Study. Perspectives in Health Information Management, 13(Winter).
- [Hu, 2005] Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., & DeLisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules. Nucleic acids research, 33(suppl 2), W352-W357.
- [Hu *et al*, 2012] Hu, T., Chen, H., Huang, L., & Zhu, X. (2012, August). A survey of mass data mining based on cloud-computing. In Anti-counterfeiting, Security, and Identification (pp. 1-4). IEEE.
- [Huang *et al*, 2009] Huang, J., Huang, X., & Metaxas, D. (2009, September). Learning with dynamic group sparsity. In 2009 IEEE 12th International Conference on Computer Vision (pp. 64-71). IEEE.
- [Hung and Chiang, 2010] Hung, H., & Chiang, C. T. (2010). Estimation methods for time-dependent AUC models with survival data. Canadian Journal of Statistics, 38(1), 8-26.

- [IARC 2016] IARC. International Agency for Research on Cancer (IARC). <http://www-dep.iarc.fr/>.
- [Isenberg *et al*, 2008] Isenberg, T., Everts, M. H., Grubert, J., & Carpendale, S. (2008, May). Interactive exploratory visualization of 2D vector fields. In *Computer Graphics Forum* (Vol. 27, No. 3, pp. 983-990). Blackwell Publishing Ltd.
- [Jeong *et al.*, 2013] Jeong, W., Schneider, J., Hansen, A., Lee, M., Turney, S.G., Faulkner-Jones, B.E., Hecht, J.L., Najarian, R., Yee, E., Lichtman, J.W. & Pfister, H., 2013, September. A Collaborative Digital Pathology System for Multi-Touch Mobile and Desktop Computing Platforms. In *Computer Graphics Forum* (Vol. 32, No. 6, pp. 227-242).
- [Jetter, 2011] Jetter, H. C., Gerken, J., Zöllner, M., Reiterer, H., & Milic-Frayling, N. (2011, May). Materializing the query with facet-streams: a hybrid surface for collaborative search on tabletops. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3013-3022). ACM.
- [Johnson and Keil, 2014] Johnson, S. G., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, 143(6), 2223.
- [Joniau *et al*, 2015] Joniau, S., Briganti, A., Gontero, P., Gandaglia, G., Tosco, L., Fieuws, S., Tombal, B., Marchioro, G., Walz, J., Kneitz, B. & Bader, P., (2015). Stratification of high-risk prostate cancer into prognostic categories: a European multi-institutional study. *European urology*, 67(1), 157-164.
- [Jumisko-Pyykkö and Vainio, 2012] Jumisko-Pyykkö, S., & Vainio, T. (2012). Framing the context of use for mobile HCI. *Social and Organizational Impacts of Emerging Mobile Devices: Evaluating Use: Evaluating Use*, 217.
- [Kaltenbrunner, 2007] Kaltenbrunner, M., & Bencina, R. (2007, February). *reactIVision*: a computer-vision framework for table-based tangible interaction. In *Proceedings of the 1st international conference on Tangible and embedded interaction* (pp. 69-74). ACM.

Bibliography

- [Kanehisa, 2010] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl 1), D355-D360.
- [Kanehisa, 2012] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, gkr988.
- [Kaptelinin and Nardi, 2012] Kaptelinin, V., & Nardi, B. (2012). Activity theory in HCI: Fundamentals and Reflections. *Synthesis Lectures Human-Centered Informatics*, 5(1), 1-105.
- [Karni and Vierø, 2014] Karni, E., & Vierø, M. L. (2014). Awareness of unawareness: a theory of decision making in the face of ignorance. *Queen's Economics Department Working Paper*, (1322).
- [Katerina Pastra, 2016] <http://csri.gr/researchers/katerina-pastra>
- [Keim *et al*, 2008] Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Information visualization* (pp. 154-175). Springer Berlin Heidelberg.
- [Khella, 2002] Khella, A. (2002). Knowledge and mental models in HCI. Retrieved November, 14, 2008.
- [Kihlstrom, 2011] Kihlstrom, J. (2011). How Students Learn and How We Can Help Them. *How Students Learn*.
- [Kim, 2004] Kim, S., Imoto, S., & Miyano, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75(1), 57-65.
- [Klein *et al*, 2006] Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), 88-92.

- [Klemmer *et al*, 2006] Klemmer, S. R., Hartmann, B., & Takayama, L. (2006, June). How bodies matter: five themes for interaction design. In Proceedings of the 6th conference on Designing Interactive systems (pp. 140-149). ACM.
- [Klum, 2012] Klum, S., Isenberg, P., Langner, R., Fekete, J. D., & Dachsel, R. (2012, May). Stackables: combining tangibles for faceted browsing. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 241-248). ACM.
- [Kohlhammer *et al*, 2011] Kohlhammer, J., Keim, D., Pohl, M., Santucci, G., & Andrienko, G. (2011). Solving problems with visual analytics. *Procedia Computer Science*, 7, 117-120.
- [Kokar and Endsley, 2012] Kokar, M. M., & Endsley, M. R. (2012). Situation awareness and cognitive modeling. *IEEE Intelligent Systems*, 27(3), 91-96.
- [Kolb, 2014] Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- [Koo *et al*, 2014] Koo, K. C., Park, S. U., Kim, K. H., Rha, K. H., Hong, S. J., Yang, S. C., & Chung, B. H. (2015). Prognostic impacts of metastatic site and pain on progression to castrate resistance and mortality in patients with metastatic prostate cancer. *Yonsei medical journal*, 56(5), 1206-1212.
- [Koo *et al*, 2015] Koo, K. C., Park, S. U., Kim, K. H., Rha, K. H., Hong, S. J., Yang, S. C., & Chung, B. H. (2015). Predictors of survival in prostate cancer patients with bone metastasis and extremely high prostate-specific antigen levels. *Prostate international*, 3(1), 10-15.
- [Koziol and Zhenyu, 2009] Koziol, J. A., & Jia, Z. (2009). The concordance index C and the Mann-Whitney parameter $\Pr(X > Y)$ with randomly censored data. *Biometrical Journal*, 51(3), 467-474.
- [Kratz and Jones, 2014] Kratz, A., Auer, C., & Hotz, I. (2014). Tensor Invariants and Glyph Design. In *Visualization and Processing of Tensors and Higher Order Descriptors for Multi-Valued Data* (pp. 17-34). Springer Berlin Heidelberg.

Bibliography

- [Kremling, 2007] Kremling, A., & Saez-Rodriguez, J. (2007). Systems biology - an engineering perspective. *Journal of biotechnology*, 129(2), 329-351.
- [Kurtenbach, 1997] Kurtenbach, G., Fitzmaurice, G., Baudel, T., & Buxton, B. (1997, March). The design of a GUI paradigm based on tablets, two-hands, and transparency. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 35-42). ACM.
- [Laufer, 2014] Laufer, C., Fischer, B., Huber, W., & Boutros, M. (2014). Measuring genetic interactions in human cells by RNAi and imaging. *Nature protocols*, 9(10), 2341-2353.
- [Leclerc, 2008] Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology*, 4(1), 213.
- [Lee, 2002] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. & Zeitlinger, J., (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *science*, 298(5594), 799-804.
- [Lehner, 2011] Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8), 323-331.
- [Li, 2004] Li, Y. (2004). Discovering structure of data to create multiple perspective visualization (Doctoral dissertation, Massachusetts Institute of Technology).
- [Li 2012] Li, X., Shi, Y., & Gutman, I. (2012). *Graph energy*. Springer Science & Business Media.
- [Liu *et al*, 2014] Liu, J., Wan, J., He, S., & Zhang, Y. (2014). E-healthcare supported by big data. *ZTE Communications*, 12(3), 46-52.
- [Loke and Robertson, 2013] Loke, L., & Robertson, T. (2013). Moving and making strange: An embodied approach to movement-based interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(1), 7.
- [Lopes and Bontempi, 2015] Lopes, M., & Bontempi, G. (2015). Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression

- data. Quantitative Assessment and Validation of Network Inference Methods in Bioinformatics, 99.
- [Macrae and Bodenhausen, 2001] Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British journal of psychology*, 92(1), 239-255.
- [Magerkurth and Tandler, 2002] Magerkurth, C., & Tandler, P. (2002). Augmenting tabletop design for computer-supported cooperative work. In *Workshop on Co-located Tabletop Collaboration: Technologies and Directions at CSCW* (Vol. 2, p. 2002).
- [Magno, 2010] Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning*, 5(2), 137-156.
- [Mahy *et al*, 2014] Mahy, C. E., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental cognitive neuroscience*, 9, 68-81.
- [Managbanag *et al*, 2008] Managbanag, J. R., Witten, T. M., Bonchev, D., Fox, L. A., Tsuchiya, M., Kennedy, B. K., & Kaeberlein, M. (2008). Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. *PloS one*, 3(11), e3802.
- [Manshaei, 2012] Manshaei, R., Sobhe Bidari, P., Aliyari Shoorehdeli, M., Feizi, A., Lohrasebi, T., Malboobi, M.A., Kyan, M. & Alirezaie, J., (2012). Hybrid-controlled neurofuzzy networks analysis resulting in genetic regulatory networks reconstruction. *ISRN bioinformatics*, 2012.
- [Maquil *et al*, 2012] Maquil, V., Zephir, O., & Ras, E. (2012). Creating metaphors for tangible user interfaces in collaborative urban planning: Questions for designers and developers. In *From Research to Practice in the Design of Cooperative Systems: Results and Open Challenges* (pp. 137-151). Springer London.
- [Marchionini *et al*, 2006] Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4), 41-46.
- [Martin *et al*, 2016] Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind?. *Trends in cognitive sciences*, 20(5), 375-382.

Bibliography

- [Mazalek, 2014] Mazalek, A., & Arif, A. S. (2014, September). Mobile-based tangible interaction techniques for shared displays. In Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (pp. 561-562). ACM.
- [McCullough, 2001] McCullough, M. (2001). On typologies of situated interaction. *Human-Computer Interaction*, 16(2), 337-349.
- [McNeil, 2015] McNeil, S. (2015). Visualizing mental models: understanding cognitive change to support teaching and learning of multimedia design and development. *Educational Technology Research and Development*, 63(1), 73-96.
- [Mike Denham, 2016] <https://www.plymouth.ac.uk/staff/mike-denham>
- [Mikesa and Kaplan, 2014] Mikes, A., & Kaplan, R. S. (2014, October). Towards a contingency theory of enterprise risk management. AAA.
- [Millsap, 2012] Millsap, R. E. (2012). Statistical approaches to measurement invariance. Routledge.
- [Maitlis and Sonenshein, 1988] Maitlis, S., & Sonenshein, S. (2010). Sensemaking in crisis and change: Inspiration and insights from Weick (1988). *Journal of management studies*, 47(3), 551-580.
- [Mohar, 1997] Mohar, B. (1997). Some applications of Laplace eigenvalues of graphs. In *Graph symmetry* (pp. 225-275). Springer Netherlands.
- [Montes-Torres *et al*, 2016] Montes-Torres, J., Subirats, J. L., Ribelles, N., Urda, D., Franco, L., Alba, E., & Jerez, J. M. (2016). Advanced Online Survival Analysis Tool for Predictive Modelling in Clinical Data Science. *PloS one*, 11(8), e0161135.
- [MSKCC, 2016] <https://www.mskcc.org/>
- [MultiTaction, 2016] "MultiTaction | Advanced Interactive Displays", Multitaction.com, 2016. [Online]. Available: <<https://www.multitaction.com/>>. [Accessed: 13- Jun- 2016].
- [Munzner, 2014] Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

- [NafeesAhmed and Razak, 2014] NafeesAhmed, K., & Abdul Razak, T. (2014). A Comparative Study of Different Density based Spatial Clustering Algorithms. *International Journal of Computer Applications*, 99(8), 18-25.
- [Nakarada-Kordic *et al*, 2016] Nakarada-Kordic, I., Weller, J. M., Webster, C. S., Cumin, D., Frampton, C., Boyd, M., & Merry, A. F. (2016). Assessing the similarity of mental models of operating room team members and implications for patient safety: a prospective, replicated study. *BMC Medical Education*, 16(1), 229.
- [Nazemi, 2016] Nazemi, K. (2016). *Adaptive semantics visualization* (Vol. 646). Springer.
- [Nersessian, 2010] Nersessian, N. J. (2010). *Creating scientific concepts*. MIT press.
- [Newman 2002] Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20), 208701.
- [Nielsen, 2007] Nielsen, J. (2007). Principles of optimal metabolic network operation. *Molecular Systems Biology*, 3(1), 126.
- [Node.js, 2016] "Node.js", Nodejs.org, 2016. [Online]. Available: <<https://nodejs.org/en/>>. [Accessed: 13- Jun- 2016].
- [Noor, 2012] Noor, A., Serpedin, E., Nounou, M., & Nounou, H. (2012). Inferring gene regulatory networks via nonlinear state-space models and exploiting sparsity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4), 1203-1211.
- [North *et al*, 2009] North, C., Dwyer, T., Lee, B., Fisher, D., Isenberg, P., Robertson, G., & Inkpen, K. (2009, August). Understanding multi-touch manipulation for surface computing. In *IFIP Conference on Human-Computer Interaction* (pp. 236-249). Springer Berlin Heidelberg.
- [Nowke *et al*, 2015] Nowke, C., Zielasko, D., Weyers, B., Peyser, A., Hentschel, B., & Kuhlen, T. W. (2015). Integrating Visualizations into Modeling NEST Simulations. *Frontiers in neuroinformatics*, 9.

Bibliography

- [Nuhn *et al*, 2014] Nuhn, P., Vaghasia, A. M., Goyal, J., Zhou, X. C., Carducci, M. A., Eisenberger, M. A., & Antonarakis, E. S. (2014). Association of pretreatment neutrophil-to-lymphocyte ratio (NLR) and overall survival (OS) in patients with metastatic castration-resistant prostate cancer (mCRPC) treated with first-line docetaxel. *BJU international*, 114(6b), E11-E17.
- [Olshannikova *et al*, 2015] Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data*, 2(1), 1.
- [Oymak *et al*, 2014] Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., & Hassibi, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5), 2886-2908.
- [Parmar, 2013] Parmar P.S., (2013) "Microarray Technologies in the Diagnosis and Treatment of Head and Neck Cancer."
- [Patel *et al*, 2014] Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P., & Baldi, P. (2014). How pervasive are circadian oscillations?. *Trends in cell biology*, 24(6), 329-331.
- [Patel and Patel, 2016] Patel, S., & Patel, H. (2016). Survey of Data Mining Techniques used in Healthcare Domain. *International Journal of Information*, 6(1/2).
- [Peker, 2016] Peker, M. (2016). A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. *Journal of medical systems*, 40(5), 1-16.
- [Petrou *et al*, 2010] Petrou, M., Tabacchi, M. E., & Piroddi, R. (2010). Networks of concepts and ideas. *The Computer Journal*, bxp113.
- [Phillips, 2008] Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11), 855-867.
- [Piaget, 1952] Piaget, J. (1952). *The origins of intelligence in children* (Vol. 8, No. 5, pp. 18-1952). New York: International Universities Press.

- [Pratt *et al*, 2015] Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S. & Stojmirovic, A., (2015). NDEx, the Network Data Exchange. *Cell systems*, 1(4), 302-305.
- [Project Data Sphere, 2016] www.projectdatasphere.org
- [Ragan *et al*, 2016] Ragan, E. D., Endert, A., Sanyal, J., & Chen, J. (2016). Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics*, 22(1), 31-40.
- [Rathkopf *et al*, 2014] Rathkopf, D.E., Smith, M.R., De Bono, J.S., Logothetis, C.J., Shore, N.D., De Souza, P., Fizazi, K., Mulders, P.F., Mainwaring, P., Hainsworth, J.D. & Beer, T.M., (2014). Updated interim efficacy analysis and long-term safety of abiraterone acetate in metastatic castration-resistant prostate cancer patients without prior chemotherapy (COU-AA-302). *European urology*, 66(5), 815-825.
- [Ratner, 2011] Ratner, B. (2011). *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data*. CRC Press.
- [Rekimoto, 2001] Rekimoto, J., Ullmer, B., & Oba, H. (2001, March). DataTiles: a modular platform for mixed physical and graphical interactions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 269-276). ACM.
- [Richard *et al*, 2012] Richard, E., Savalle, P. A., & Vayatis, N. (2012). Estimation of simultaneously sparse and low rank matrices. *arXiv preprint arXiv:1206.6474*.
- [Riedel and Blokland, 2016] Riedel, W. J., & Blokland, A. (2015). Declarative memory. In *Cognitive Enhancement* (pp. 215-236). Springer International Publishing.
- [Riedenklau, 2014] Riedenklau, E. (2016). *Development of actuated Tangible User Interfaces: new interaction concepts and evaluation methods* (Doctoral dissertation, Bielefeld University).

Bibliography

- [Rittle-Johnson *et al*, 2014] Rittle-Johnson, B., & Schneider, M. (2014). Developing conceptual and procedural knowledge of mathematics. Oxford handbook of numerical cognition. Oxford, UK: Oxford University Press. doi, 10.
- [Rodríguez-Fernández *et al*, 2016] Rodríguez-Fernández, A., Ramos-Díaz, E., Fernández-Zabala, A., Goni, E., Esnaola, I., & Goni, A. (2016). Contextual and psychological variables in a descriptive model of subjective well-being and school engagement. *International Journal of Clinical and Health Psychology*, 16(2), 166-174.
- [Roth, 2013] Roth, R. E. (2013). An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE transactions on visualization and computer graphics*, 19(12), 2356-2365.
- [Saha-Chaudhuri and Heagerty, 2013] Saha-Chaudhuri, P., & Heagerty, P. J. (2013). Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, 14(1), 42-59.
- [Sanofi, 2016] <http://www.sanofi.ca/l/ca/index.jsp>
- [Sargent, 2013] Sargent, C. S. (2013). Find it, fix it, and thrive: The impact of insisting on proficiency in prerequisite knowledge in intermediate accounting. *Issues in Accounting Education*, 28(3), 581-597.
- [Schaper, 2013] Schaper, H. (2013). Physical Widgets on Capacitive Touch Displays (Doctoral dissertation, RWTH Aachen University).
- [Schkolne, 2004] Schkolne, S., Ishii, H., & Schroder, P. (2004, October). Immersive design of DNA molecules with a tangible interface. In *Proceedings of the conference on Visualization'04* (pp. 227-234). IEEE Computer Society.
- [Schneider and Blikstein, 2014] Schneider, B., & Blikstein, P. (2015). Using exploratory tangible user interfaces for supporting collaborative learning of probability. *IEEE TLT*.

- [Schuetz, 2007] Schuetz, R., Kuepfer, L., & Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular systems biology*, 3(1), 119.
- [Shaer and Hornecker, 2010] Shaer, O., & Hornecker, E. (2010). Tangible user interfaces: past, present, and future directions. *Foundations and Trends in Human-Computer Interaction*, 3(12), 1-137.
- [Shaer *et al*, 2012] Shaer, O., Strait, M., Valdes, C., Wang, H., Feng, T., Lintz, M., Ferreira, M., Grote, C., Tempel, K. & Liu, S., (2012). The design, development, and deployment of a tabletop interface for collaborative exploration of genomic data. *International Journal of Human-Computer Studies*, 70(10), 746-764.
- [Shaer *et al*, 2013] Shaer, O., Mazalek, A., Ullmer, B., & Konkel, M. (2013, February). From big data to insights: opportunities and challenges for TEI in genomics. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction* (pp. 109-116). ACM.
- [Shafipour & Abdolvahhab, 2016] Shafipour, G., & Fetanat, A. (2016). Survival analysis in supply chains using statistical flowgraph models: predicting time to supply chain disruption. *Communications in Statistics-Theory and Methods*, (just-accepted), 00-00.
- [Shannon, 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T., (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), 2498-2504.
- [Shao *et al*, 2014] Shao, Y.H.J., Kim, S., Moore, D.F., Shih, W., Lin, Y., Stein, M., Kim, I.Y. & Lu-Yao, G.L., (2014). Cancer-specific survival after metastasis following primary radical prostatectomy compared with radiation therapy in prostate cancer patients: results of a population-based, propensity score-matched analysis. *European urology*, 65(4), 693-700.
- [Sharlin *et al*, 2004] Sharlin, E., Watson, B., Kitamura, Y., Kishino, F., & Itoh, Y. (2004). On tangible user interfaces, humans and spatiality. *Personal and Ubiquitous Computing*, 8(5), 338-346.

Bibliography

- [Shaw *et al*, 1998] Shaw, C. D., Ebert, D. S., Kukla, J. M., Zwa, A., Soboroff, I., & Roberts, D. A. (1998, May). Data visualization using automatic perceptually motivated shapes. In Photonics West'98 Electronic Imaging (pp. 208-213). International Society for Optics and Photonics.
- [Shneiderman, 1996] Shneiderman, B. (1996, September). The eyes have it: A task by data type taxonomy for information visualizations. In Visual Languages, 1996. Proceedings., IEEE Symposium on (pp. 336-343). IEEE.
- [Shotwell and Slate, 2011] Shotwell, M. S., & Slate, E. H. (2011). Bayesian outlier detection with dirichlet process mixtures. *Bayesian Analysis*, 6(4), 665-690.
- [Shrinivasan *et al*, 2008] Shrinivasan, Y. B., & van Wijk, J. J. (2008, April). Supporting the analytical reasoning process in information visualization. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1237-1246). ACM.
- [Sieck *et al*, 2007] Sieck, W. R., Klein, G., Peluso, D. A., Smith, J. L., Harris-Thompson, D., & Gade, P. A. (2007). FOCUS: A model of sensemaking. KLEIN ASSOCIATES INC FAIRBORN OH.
- [Siegel *et al*, 2016] Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1), 7-30.
- [Singh and Wayal, 2012] Singh, S. K., & Wayal, M. G. (2012, February). A review: data mining with fuzzy association rule mining. In *International Journal of Engineering Research and Technology* (Vol. 1, No. 5 (July-2012)). ESRSA Publications.
- [Smith *et al*, 1987] Smith, D. C., Irby, C., Kimball, R., Verplank, W. L., & Harslem, E. (1987, December). Designing the Star user interface. In *Human-computer interaction* (pp. 653-661). Morgan Kaufmann Publishers Inc.
- [Sontag, 2005] Sontag, E. D. (2005). Molecular systems biology and control. *European journal of control*, 11(4), 396-435.
- [Sorden, 2012] Sorden, S. D. (2012). The cognitive theory of multimedia learning. *Handbook of educational theories*. Charlotte, NC: Information Age Publishing.

- [Spellman, 1998] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. & Futcher, B., (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12), 3273-3297.
- [Spizzirri 2011] Spizzirri, L. (2011). Justification and application of eigenvector centrality. *Algebra in Geography: Eigenvectors of Network*.
- [Sprague *et al*, 2012] Sprague, D., & Tory, M. (2012). Exploring how and why people use visualizations in casual contexts: Modeling user goals and regulated motivations. *Information Visualization*, 11(2), 106-123.
- [Ståhl *et al*, 2016] Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F, Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M. & Mollbrink, A., (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294), 78-82.
- [Steck *et al*, 2008] Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., & Raykar, V. C. (2008). On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems* (pp. 1209-1216).
- [Stelling, 2004] Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & Doyle, J. (2004). Robustness of cellular functions. *Cell*, 118(6), 675-685.
- [Sun and Chandan, 2013] Sun, J., & Reddy, C. K. (2013, August). Big data analytics for healthcare. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1525-1525). ACM.
- [Sun *et al*, 2013] Sun, G. D., Wu, Y. C., Liang, R. H., & Liu, S. X. (2013). A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5), 852-867.
- [Sutherland, 2005] Sutherland, W. J. (2005). The best solution. *Nature*, 435(7042), 569-569.

Bibliography

- [TactionGuide, 2016] "TactionGuide | Multitouch Cornerstone", MultiTouch Support Center | Multitouch Cornerstone, 2016. [Online]. Available: <<https://cornerstone.multitouch.fi/tactionguide>>. [Accessed: 13- Jun- 2016].
- [Tanaka, 2005] Tanaka, R., Csete, M., & Doyle, J. (2005). Highly optimised global organisation of metabolic networks. *Systems biology*, 152(4), 179.
- [Tax *et al*, 2015] Tax, C.M., Chamberland, M., van Stralen, M., Viergever, M.A., Whittingstall, K., Fortin, D., Descoteaux, M. & Leemans, A., (2015). Seeing More by Showing Less: Orientation-Dependent Transparency Rendering for Fiber Tractography Visualization. *PloS one*, 10(10), e0139434.
- [Thielea *et al*, 2015] Thielea, T., Jooßa, C., Richerta, A., & Jeschkea, S. (2015, August). Terminology Based Visualization of Interfaces in Interdisciplinary Research Networks. In *Proceedings 19th Triennial Congress of the IEA* (Vol. 9, p. 14).
- [Thomas, 2007] Thomas, R., Paredes, C. J., Mehrotra, S., Hatzimanikatis, V., & Papoutsakis, E. T. (2007). A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data. *BMC bioinformatics*, 8(1), 1.
- [Tian, 2011] Tian, L. P., Liu, L. Z., Zhang, Q. W., & Wu, F. X. (2011). Nonlinear model-based method for clustering periodically expressed genes. *The Scientific World Journal*, 11, 2051-2061.
- [Tienda-Luna, 2009] Tienda-Luna, I. M., Perez, M. C. C., Padillo, D. P. R., Yin, Y., & Huang, Y. (2009). Sensitivity and specificity of inferring genetic regulatory interactions with the VBEM algorithm. *IADIS International Journal on Computer Science and Information Systems*, 4(1), 54-63.
- [Tourish and Robson, 2006] Tourish, D., & Robson, P. (2006). Sensemaking and the distortion of critical upward communication in organizations. *Journal of Management Studies*, 43(4), 711-730.

- [Troyanskaya, 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R.B., (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.
- [Tweedie, 1997] Tweedie, L. (1997, March). Characterizing interactive externalizations. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 375-382). ACM.
- [Uemura *et al*, 2016] Uemura, K., Miyoshi, Y., Kawahara, T., Yoneyama, S., Hattori, Y., Teranishi, J.I., Kondo, K., Moriyama, M., Takebayashi, S., Yokomizo, Y. & Yao, M., (2016). Prognostic value of a computer-aided diagnosis system involving bone scans among men treated with docetaxel for metastatic castration-resistant prostate cancer. *BMC cancer*, 16(1), 1.
- [Ullmer and Ishii, 1997] Ullmer, B., & Ishii, H. (1997, October). The metaDESK: models and prototypes for tangible user interfaces. In *Proceedings of the 10th annual ACM symposium on User interface software and technology* (pp. 223-232). ACM.
- [Ullmer, 2000] Ullmer, B., & Ishii, H. (2000). Emerging frameworks for tangible user interfaces. *IBM systems journal*, 39(3.4), 915-931.
- [Ullmer, 2003] Ullmer, B., Ishii, H., & Jacob, R. J. (2003). Tangible query interfaces: Physically constrained tokens for manipulating database queries. In *Proc. of INTERACT* (Vol. 3, pp. 279-286).
- [Unistrut, 2016] "Unistrut U.S.", Unistrut.us, 2016. [Online]. Available: <<http://www.unistrut.us/>>. [Accessed: 13- Jun- 2016].
- [Valdes, 2014] Valdes, C., Eastman, D., Grote, C., Thatte, S., Shaer, O., Mazalek, A., Ullmer, B. & Konkel, M.K., (2014, April). Exploring the design space of gestural interaction with active tokens through user-defined gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 4107-4116). ACM.
- [Van Beurden, 2013] Van Beurden, M. H. P. H. (2013). Interaction in depth (Doctoral dissertation, Thesis, Interaction group, Eindhoven University of Technology, the Netherlands).

Bibliography

- [Van Campenhout *et al*, 2013] Van Campenhout, L., Frens, J., Overbeeke, K., Standaert, A., & Peremans, H. (2013). Physical interaction in a dematerialized world. *International Journal of Design*, 7(1).
- [Van Steen, 2011] Van Steen, K. (2011). Travelling the world of gene-gene interactions. *Briefings in bioinformatics*, bbr012.
- [Vanky *et al*, 2016] Vanky, A. (2016) Make Data Make Sense: The Importance of Visualization in Data Analytics. *IQT QUARTERLY*, Vol. 7 No. 4.
- [Velloso *et al*, 2015] Velloso, E., Schmidt, D., Alexander, J., Gellersen, H., & Bulling, A. (2015). The feet in human-computer interaction: a survey of foot-based interaction. *ACM Computing Surveys (CSUR)*, 48(2), 21.
- [Verma and Vineeta 2012] Verma, G., & Verma, V. (2012). Role and applications of genetic algorithm in data mining. *International journal of computer applications*, 48(17), 5-8.
- [Vilanova *et al*, 2006] Vilanova, A., Zhang, S., Kindlmann, G., & Laidlaw, D. (2006). An introduction to visualization of diffusion tensor imaging and its applications. In *Visualization and Processing of Tensor Fields* (pp. 121-153). Springer Berlin Heidelberg.
- [Waese *et al*, 2016] Waese, J., Pasha, A., Wang, T. T., van Weringh, A., Guttman, D. S., & Provart, N. J. (2016). Gene Slider: sequence logo interactive data-visualization for education and research. *Bioinformatics*, btw525.
- [Wahl *et al*, 2016] Wahl, M., Krüger, J., & Frommer, J. (2016, July). Users' Sense-Making of an Affective Intervention in Human-Computer Interaction. In *International Conference on Human-Computer Interaction* (pp. 71-79). Springer International Publishing.
- [Walsh *et al*, 2012] Walsh, A. M., Hyde, M. K., Hamilton, K., & White, K. M. (2012). Predictive modelling: parents' decision making to use online child health information to increase their understanding and/or diagnose or treat their child's health. *BMC medical informatics and decision making*, 12(1), 1.

- [Wang, 2006] Wang, Y., Joshi, T., Zhang, X. S., Xu, D., & Chen, L. (2006). Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, 22(19), 2413-2420.
- [Wang, 2007] Wang, R. S., Wang, Y., Zhang, X. S., & Chen, L. (2007). Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics*, 23(22), 3056-3064.
- [Wang *et al*, 2009] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57-63.
- [Wang and Ma 2014] Wang, Y., & Ma, J. (2014). *Mobile Social Networking and Computing: A Multidisciplinary Integrated Perspective*. CRC Press.
- [Warde-Farley, 2010] Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, E., Lopes, C.T. & Maitland, A., (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl 2), W214-W220.
- [Weaver, 2007] Weaver, C. (2007, January). Patterns of coordination in Improvise visualizations. In *Proceedings of the IS&T/SPIE conference on visualization and data analysis*, San Jose.
- [Weissgerber *et al*, 2016] Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., & Milic, N. M. (2016). From Static to Interactive: Transforming Data Visualization to Improve Transparency. *PLoS Biol*, 14(6), e1002484.
- [Whitenton, 2013] Whitenton, K. (2013). Minimize cognitive load to maximize usability. *Pozyskano*, 4, 2014.
- [Wickens *et al*, 2015] Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology & human performance*. Psychology Press.
- [Wiethoff, 2012] Wiethoff, A. (2012). *Prototyping tools for hybrid interactions* (Doctoral dissertation, lmi).

Bibliography

- [Willett *et al*, 2012] Willett, W., Heer, J., & Agrawala, M. (2012, May). Strategies for crowdsourcing social data analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 227-236). ACM.
- [Wolkenhauer, 2005] Wolkenhauer, O., Ullah, M., Wellstead, P., & Cho, K. H. (2005). The dynamic systems approach to control and regulation of intracellular networks. *FEBS letters*, 579(8), 1846-1853.
- [Wong *et al*, 2012] Wong, P. C., Shen, H. W., Johnson, C. R., Chen, C., & Ross, R. B. (2012). The top 10 challenges in extreme-scale visual analytics. *IEEE computer graphics and applications*, 32(4), 63.
- [Wu *et al*, 2011] Wu, A., Yim, J. B., Caspary, E., Mazalek, A., Chandrasekharan, S., & Nersessian, N. J. (2011, November). Kinesthetic pathways: a tabletop visualization to support discovery in systems biology. In Proceedings of the 8th ACM conference on Creativity and cognition (pp. 21-30). ACM.
- [Wu, 2013] Wu, Y., Zhu, X., Chen, J., & Zhang, X. (2013). EINVis: a visualization tool for analyzing and exploring genetic interactions in large-scale association studies. *Genetic epidemiology*, 37(7), 675-685.
- [Wu *et al*, 2014] Wu, J. N., Fish, K. M., Evans, C. P., deVere White, R. W., & Dall'Era, M. A. (2014). No improvement noted in overall or cause-specific survival for men presenting with metastatic prostate cancer over a 20-year period. *Cancer*, 120(6), 818-823.
- [Xambó 2015] Xambó, A. (2015). Tabletop Tangible Interfaces for Music Performance: Design and Evaluation (Doctoral dissertation, The Open University).
- [XAMPP, 2016] "XAMPP Installers and Downloads for Apache Friends", [Apachefriends.org](https://www.apachefriends.org/), 2016. [Online]. Available: <<https://www.apachefriends.org/index.html>>. [Accessed: 13- Jun-2016].
- [Xu *et al*, 2013] Xu, W., Chang, K., Francisco, N., Valdes, C., Kincaid, R., & Shaer, O. (2013, February). From wet lab bench to tangible virtual experiment: SynFlo. In Proceedings of

- the 7th International Conference on Tangible, Embedded and Embodied Interaction (pp. 399-400). ACM.
- [Yalcin *et al*, 2016] Yalcin, M. A., Elmqvist, N., & Bederson, B. B. (2016). AggreSet: Rich and Scalable Set Exploration using Visualizations of Element Aggregations. *IEEE transactions on visualization and computer graphics*, 22(1), 688-697.
- [Yamashita *et al*, 2016] Yamashita, S., Kohjimoto, Y., Iguchi, T., Koike, H., Kusumoto, H., Iba, A., Kikkawa, K., Kodama, Y., Matsumura, N. & Hara, I., (2016). Prognostic factors and risk stratification in patients with castration-resistant prostate cancer receiving docetaxel-based chemotherapy. *BMC urology*, 16(1), 1.
- [Yiannoudes, 2016] Yiannoudes, S. (2016). *Architecture and Adaptation: From Cybernetics to Tangible Computing*. Routledge.
- [Yuan *et al*, 2014] Yuan, Z., Li, F., Zhang, P., & Chen, B. (2014). Description of shape characteristics through Fourier and wavelet analysis. *Chinese Journal of Aeronautics*, 27(1), 160-168.
- [Zeng, 2008] Zeng, L., Wu, J., & Xie, J. (2008). Statistical methods in integrative analysis for gene regulatory modules. *Statistical applications in genetics and molecular biology*, 7(1).
- [Zhang and Norman, 1994] Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive science*, 18(1), 87-122.
- [Zhang, 2015] Zhang, S., Jing, Y., Zhang, M., Zhang, Z., Ma, P., Peng, H., Shi, K., Gao, W.Q. & Zhuang, G., (2015). Stroma-associated master regulators of molecular subtypes predict patient prognosis in ovarian cancer. *Scientific reports*, 5.
- [Zhao *et al*, 2015] Zhao, T., Liao, B., Yao, J., Liu, J., Huang, R., Shen, P., Peng, Z., Gui, H., Chen, X., Zhang, P. & Zhu, Y., (2015). Is there any prognostic impact of intraductal carcinoma of prostate in initial diagnosed aggressively metastatic prostate cancer?. *The Prostate*, 75(3), 225-232.

Bibliography

- [Zhukov and Barr, 2003] Zhukov, L., & Barr, A. H. (2003, October). Heart-muscle fiber reconstruction from diffusion tensor MRI. In Proceedings of the 14th IEEE Visualization 2003 (VIS'03) (p. 79). IEEE Computer Society.
- [Zoppoli, 2010] Zoppoli, P., Morganella, S., & Ceccarelli, M. (2010). TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics, 11(1), 1.