Eurographics Workshop on Visual Computing for Biology and Medicine (2015)
K. Bühler, L. Linsen, and N. W. John (Editors)

Discovering Medical Knowledge Using Visual Analytics
– a survey on methods for systems biology and ⋆omics data –

W. Sturm[1,2], T. Schreck[1], A. Holzinger[3,4] and T. Ullrich[1,2]

[1] Institut für ComputerGraphik & WissensVisualisierung (CGV), TU Graz, Austria
[2]Fraunhofer Austria Research GmbH, Visual Computing, Graz, Austria
[3]Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Austria
[4]CBmed, Center for Biomarker Research in Medicine, Graz, Austria

Abstract

Due to advanced technologies, the amount of biomedical data has been increasing drastically. Such large data sets might be obtained from hospitals, medical practices or laboratories and can be used to discover unknown knowledge and to find and reflect hypotheses. Based on this fact, knowledge discovery systems can support experts to make further decisions, explore the data or to predict future events. To analyze and communicate such a vast amount of information to the user, advanced techniques such as knowledge discovery and information visualization are necessary. Visual analytics combines these fields and supports users to integrate domain knowledge into the knowledge discovery process.

This article gives a state-of-the-art overview on visual analytics reseach with a focus on the biomedical domain, systems biology and ⋆omics data.

Categories and Subject Descriptors (according to ACM CCS): H.1.2 [Information Systems]: User/Machine Systems—Human information processing J.3 [Computer Applications]: Life and Medical Sciences—Biology and genetics J.3 [Computer Applications]: Life and Medical Sciences—Medical information systems

1. Introduction

Due to the emerging trend towards personalized medicine (P4: Personalized, Predictive, Preventive, Participatory), European health systems are challenged by increasingly big and complex sets of heterogeneous, high-dimensional data and increasing amounts of unstructured information. Thus, cognitive complexity and high-level visualizations challenge the appropriate understanding of information in the clinical context. User-centered design and the tailoring of information representations to the specificity of human information processing is crucial. This is still more important facing the increasing diversity of end users in the increasingly complex biomedical domain, which have to understand and handle complex information in the medical field for the purpose of decision making. This challenge is addressed by biomedical visual analytics [HJ14].

This article reviews and categorizes state-of-the-art approaches of knowledge discovery and visual analytics for the biomedical domain. It also reviews the novel biomedical approach of systems biology which makes use of so-called "⋆omics" data (genomics, proteomics, metabolomics, transcriptomics, etc.) to analyze biological properties of genomes, proteins and metabolites and to understand biological and pathological processes.

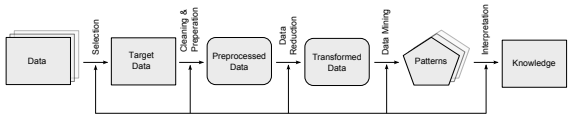

Figure 1: *The simplified iterative KDD process depicts how new knowledge can be extracted from multiple data sources* [FPSS96b].

The knowledge discovery process – also knwon as knowledge discovery in databases (KDD) – is outlined in Figure 1. It consists of several important steps:

DOI: 10.2312/vcbm.20151210

Domain Knowledge This step includes understanding of the domain by gathering necessary state-of-the-art information and defining a final goal of the process.

Target Data set The creation of a data set by acquainting data from several sources is vital in order to unify values. Moreover, the data and variables, which should be used in the further process, should be selected.

Data Cleaning and Preparation In general, large data sets are noisy, inconsistent and might come from heterogeneous sources, so that cleansing of the data is essential. The quality of a performed knowledge discovery is directly dependent on the quality of the underlying data set [HK06]. Cleaning includes handling missing values, removing outliers, smoothing noise and resolving inconsistency. Data cleaning is an essential element of data mining but experts have to be aware that each manipulation of the data set might lead to a different result and interpretation of the data. Therefore, the final finding might deviate even more from the real model.

Data Reduction The data can be reduced by dimensionality reduction such as principle component analysis [WEG87], multi-dimensional scaling [CC00] and independent component analysis [HKO04]. Furthermore, additional approaches to reduce the number of variables are specific transformation methods and the assortment of features that represent the data set best.

According to FAYYAD ET AL., data mining tasks can be classified into six different types [FPSS96a], namely *clustering*, *classification*, *association rule mining*, *regression* and *summarization*. Mostly, these techniques are derived or reused from various research fields (e.g., machine learning, statistics and pattern recognition).

Clustering Clustering algorithms assign every data item to one class of a predefined set of classes to describe the data. In other words, such algorithms determine a set of categories or clusters to distinguish and to heap together data points. Depending on the algorithm, clusters can be mutually exhaustive, hierarchical or overlapping [FPSS96a]. *k-means*, *hierarchical clustering* or *clique* are just a few examples of clustering algorithms. Basically, clustering algorithms need a similarity and dissimilarity function, also known as distance function, to distinguish data points. Examples of distance functions are *Euclidean distance* or *Minkowski distance* [XW*05].

Classification Classification is about learning a function (classifier) which assigns new data items into one of the predefined classes. The decision is based on the learned knowledge from a labeled past data set. Thus, classification algorithms are trained by supervised learning techniques. There exist many applications of classification in various domains. Basically, algorithms are subdivided into binary classifications (positive and negative outcome) and multi class classifications [Alp04]. Some examples of commonly accepted techniques are *Neural Networks* [Gro88], *Naive Bayes Classifier* [Ris01], *Decision Trees* [SL91], *K-nearest Neighbor* [CH67] and *Support Vector Machines* [HDO*98].

Association Rule Mining Association rule mining (also known as Dependency modeling) intends to find a model which represents major dependencies between variables in large databases. Two levels of dependency models can be distinguished: the *structural* model shows local dependencies of variables while *quantitative* models describe the strength of dependency as a numerical value [FPSS96a, LHM98].

Regression Regression involves the search of a linear and higher dimensional function, which approximates the given data with a minimal distance error (e.g., mean square error). A so-called regression function models the relation between one or several predictor variables (multiple regression) and a single dependent response variable. Regressions are usually used for prediction tasks. However, a low-dimensional regression function can also represent the dependency in a human-understandable way (e.g, plot) [FPSS96a, Alp04].

Summarization Summarization aims to find a short description of the data which is commonly used for interactive exploratory data analysis and report generations [FPSS96a]. CHANDOLA ET AL. describe summarization as follows:

> "Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation." [CK07]

For summarization, various values can be representative while preserving the most information. For example the centroid of a cluster of documents is a good representative of all items within the cluster. Another summarization approach uses aggregation functions (calculation of maximum, average, etc.) [AK06].

Sequential Patterns The search for sequential patterns aims to find trends or to analyze the process generating patterns in time-dependent data sets [FPSS96b].

2. Visual Analytics

A novel approach combines and emphases the research fields human computer interaction (HCI) and Knowledge discovery in databases. The ultimate goal of this approach is to enhance human intelligence by computational power and intelligence [Hol13] – the visual analytics process.

The visual analytics process implies the selection of automated data mining algorithms combined with an appropriate visual presentation [KAF*08, KKEM10]. Therefore, it is a combination of traditional data mining and information visualization (see Figure 2).

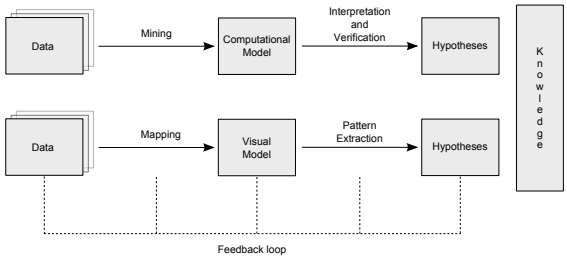

Figure 2: *A comparison of analytic prosesses between conventional data mining (top) and information visualization (bottom)* [KKEM10].

To emphasize the process, KEIM extended SCHNEIDERMAN's mantra as follows:

> "Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand." [KMS*08]

Moreover, an essential part of the overall visual analytics process is the sense-making loop [KAF*08]: the visualization process is iterative, where the user interface acts as link between data and user.

Visual analytics techniques can be categorized in several ways. The categorization used by BERTINI ET AL. [BL10] emphasizes whether the visualization or the analytical part plays the major role. For that, they used three categories, namely: *computationally enhanced visualization*, *visually enhanced mining* and *integrated visualization and mining*. TURKAY ET AL. [TJHH14] presented a 2-dimensional classification scheme. The first categorization distinguishes the type of analytical task which is classified in *summarizing information*, *finding groups & classification* and *investigating relations & prediction*. The second one categorizes the applied visualization technique according to its integration level of analytical and computational tools: *visualization as a presentation medium*, *semi-interactive use of computational methods* and *tight integration of interactive visual and computational tools*.

3. Systems Biology and ⋆omics Data

Concerning visual analytics techniques the bio-medical domain is faced with various challenges.

The combination of multiple data sets is often necessary and the data formats tend to be as diverse as its sources. Therefore, data pre-processing is needed to obtain a uniformly structured data set for performing further analysis. Each data source is likely to contain different records or some sources might be incomplete. Values may be continuous or discrete, stored in varied dimensions or even be acquainted under different measurement standards and conditions. Such conditions imply technical and environmental aspects (e.g., used equipment, ambient temperature, etc.) and require particular data transformations [Kob14, HK06]. If these influences are not considered carefully, the combined data set might lead to harmful divergences of values and furthermore to distorted results of the performed analysis.

In fact, the integration and linking of medical data from different temporal and observation scales is a huge challenge. For example, in "Image Analysis in Epidemiological Applications" [TGR*15] the challenges of visual feature extraction and comparison from a given scale (e.g., a given patient organ) in long-term studies are laid out. Similarity, linking data from different observation scales like the molecular scale, protein scale, and metabolism scale potentially needed for a given patient, remains complex (cf. Figure 6 and below).

Biomedical data sets usually contain personal information which has to be protected by applying to ethical policies. Third parties must not be able to identify patients in a single data set or even by linking multiple accessible data sets combined with potential background knowledge (linkage attack). To emphasize sensitivity, linkage-relevant attributes are divided into identifiers and so-called quasi identifiers (QI) [KHS*14]. While pure identifiers uniquely identify a person, a combination of QIs is needed for a confident identification. There exist multiple approaches to achieve anonymity like anonymization and pseudonymization.

Anonymization describes, besides the removal of personal information, the fragmentation of attributes and addition of ambiguity to protect privacy while retaining the data's quality for performing knowledge discovery.

Pseudonymization replaces all identifiers with non-related pseudonyms or hashes. Another approach is the generalization of values (e.g., usage of the birth year instead of the exact date) which weakens identifiers efficiently but might influence the data quality for further research as well.

Data cleansing includes removing noise, handling and mapping missing values within the data set to achieve better quality in knowledge discovery. Therefore, data cleansing is an essential step and it might take up to 80% of the time of the overall process [DND*02, MM10]. Besides the general data cleansing tasks of the KDD process, missing data fields can be filled by performing further additional information acquisitions. As data cleansing modifies the original data set, experts need to be aware of the fact, that any modification leads to a deviated interpretation of the data set.

Knowledge discovery implies the selection and application of data mining and machine learning algorithms to search for new patterns. Such patterns support experts to discover new knowledge and unknown relations within the data set. The result of the applied algorithm has to be visualized in a comprehensible way to allow experts to investigate the discovered knowledge. The visualization system should offer sophisticated interaction methods to explore the data set and adjust granularity. The biomedical domain chal-

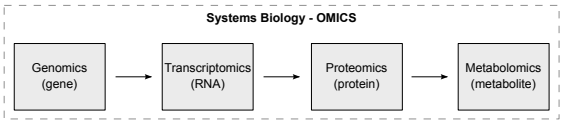

Figure 3: *This figure illustrates relations between different types of ⋆omics-data. Gene data (genomics) is transcribed to transcriptomics (RNA). RNA can be broken down to all proteins it consists of (proteomics) and each protein can be described by motabolites and its corresponding chemical process (metabolomics).*

lenges visualizations in multiple ways. First, because of the trend to data-centric medicine, systems have to cope with huge, complex and multidimensional volumes, which are likely to include unstructured and noisy data. Furthermore, precision medicine aims to integrate multiple data sources (e.g, ⋆omics-data, etc.) [TJHH14]. This fact dramatically increases complexity of the data set and adds an additional challenge for data analysts and appropriate visualizations.

Users and experts may use the discovered knowledge to make decisions for further actions or document the result. Generally, decision support systems represent extracted knowledge from the analyzed data, so it does not offer a complete solution for a given problem. The main expertise for making further decisions and solving problems is still the experts experience and knowledge [HJ14, SGG*01].

Within this article, we will focus on the visualization of ⋆omics-data. The term "⋆omics" describes the combination of several research fields which are called *genomics*, *transcriptomics*, *proteomics* and *metabolomics* [HK11]. Lately, these research fields have advanced significantly due to high-throughput technologies such as *microarray technology* [Hel02], *Next-Generation Sequencing* (NGS) [Mar08] and *mass spectrometry* [AM03]. Due to these techniques, a vast amount of data has been generated and enables experts to perform detailed research. As depicted in Figure 3, all mentioned types of ⋆omics-data depend on each other in a sequential manner. The most important ⋆omics-data types (in terms of data volume) are *genomics*, *proteomics*, and *metabolomics*.

Genomics In general terms, genomics is the research field of genes and gene expressions (DNA). Microarray techniques are one of the key technologies which significantly advanced genomics. Microarray data sets usually are of high dimensionality, so that dimensionality reduction may be applied to simplify the data set before using it for further analysis [WvdL11]. The most common visualization techniques are scatter plots, parallel coordinates plots [Ins85] and heat maps [GOB*10].

Parallel coordinate plots are a flexible way to analyze multivariate gene data. It supports users to find correlations between samples and expression levels. Conditions (brushes) are used to highlight a specific subset of the data. A disadvantage of the parallel coordinate plots is that the order of the axes influences the graphical representation significantly. To avoid too many intersections, a limited amount of samples may be used. Moreover, quality metrics can support the system to find a more preferred order.

Figure 4 shows various examples of using heat maps to analyze microarray gene expression data. A clustering of rows and columns leads to an ordered matrix, which simplifies the investigation of relations and values. In addition to that, threshold values can be used to hide uninteresting values and highlight a specific range of values [KPH*12].

Proteomics An understanding of relations between proteins is essential in systems biology as biological processes of a cell are controlled by protein interactions. Data sets containing information about protein interactions are usually large and complex because a single protein can interact with up to several dozens proteins [RP12, SMM*14]. BU ET AL. state:

> "It is believed that all biological processes are essentially and accurately carried out through protein–protein interactions." [BZC*03]

As protein–protein interactions are usually visualized by graphs, a complete representation of all interactions is overwhelming for users. Therefore, tools try to visualize specific proteins or important subsets at a time (see Figure 5). Due to its high complexity, common tools use very different methods to visually represent such graphs (no standard method has been recognized yet) [BZC*03, SMM*14].

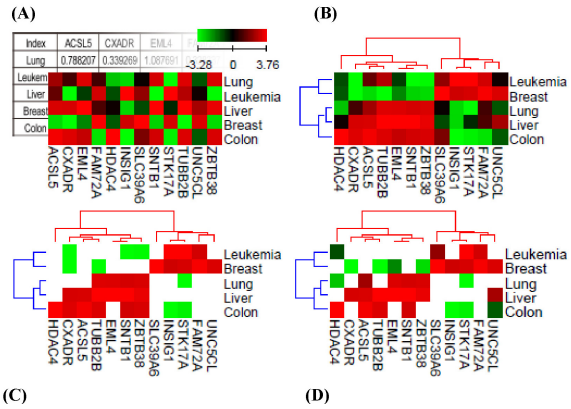

Figure 4: *Illustration of heat maps depicting microarray data for 12 genes and 5 cancer samples. Up-regulated gene expressions are shown in red and down-regulated ones in green. (a) The input data is shown as a standard heat map. (b) Cancer samples (rows) and genes (columns) have been reordered by clustering. Adjacent dendrograms represent the cluster result. (c) Selective depiction of high and low expressions. (d) Selected depiction of genes controlled by a threshold value. (Image source:* KIM ET AL. [KPH*12]*).*

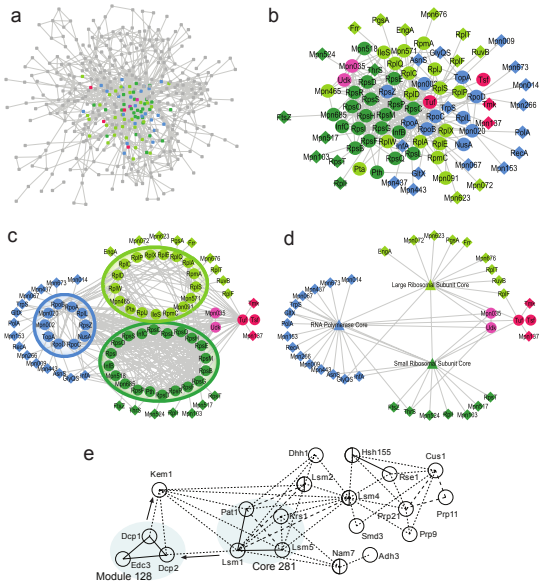

Figure 5: *Examples of visualized protein interaction networks. (a) A protein interaction network with more than 400 proteins placed by using a force-directed algorithm. (b) Simplified graph by removing unimportant nodes. (c) Manual replacement of nodes of the network to emphasize structure and interactions. (d) All core nodes of one type have been collapsed to a single meta node to simplify the network. (e) A representation of stages in deadenylation-dependent mRNA degradation. (Image source:* GEHLENBORG ET AL. [GOB*10]*).*

A drawback of visualized protein interactions is the fact, that only already-known interactions can be visualized. If the underlying protein complex purification techniques (e.g., *mass spectrometry* [AM03], *correlated messenger RNA expression profiles* [HMJ*00]) does not detect any interaction, it will not be visualized afterwards. However, protein networks can still be used to understand and to find biological functions by graph mining. For example, finding quasi-cliques or quasi-bipartites might reveal unknown knowledge [BZC*03].

Metabolomics Metabolomics is about analyzing metabolites and their associated chemical reactions within a cell. To represent such chemical chain reactions, metabolic pathways are used. Such pathways are usually represented as acyclic graphs.

There exist many stand-alone tools to explore a specific type of data but it does not support the user to link the gained knowledge to other data sets [Lin11]. Therefore, the ultimate goal of systems biology is to support biologists to gain insight into whole organisms by linking all abstraction levels to a single system (e.g., from organs to molecules). This can only be achieved by an integrative framework which combines several visualizations of interlinked heterogeneous data sets (see Figure 6). Currently, this goal remains a considerable way off. The first steps have been done and already show the high potential for visual analytics applications [BSM*15], but in order to reach the ultimate goal several political and social hurdles have to be surmounted: questions of standardization, data access, data security and privacy have to be answered.

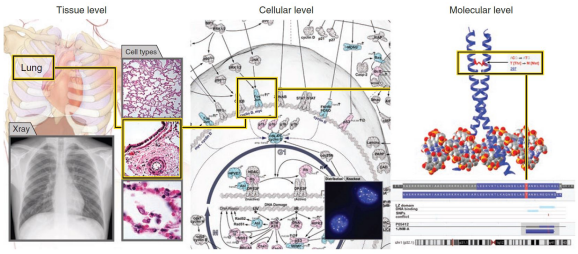

Figure 6: *The ultimate goal of systems biology is to link heterogeneous data sets to support biologists and bio-medical experts to gain insight into the whole biological system. Such visualizations might depict X-ray scans, tissues, cellular and molecular data, genomes and metabolic pathways. (Image source:* O'DONOGHUE ET AL. [OGG*10]*).*

4. Visual Analytics in Biomedical Domain

We performed an analysis of 73 recent visual analytics papers. Our review is based on the state-of-the-art report of TURKAY ET AL. [TJHH14] and it extends the given analysis by classifying all scientific papers into the categorizations *data type* and *visualization techniques*. Moreover, several additional visual analytics papers are included.

All papers are categorized into four dimensions, where the first two are inherited from the analysis of TURKAY ET AL. [TJHH14]:

- type of analysis
- level of integration
- visualization technique
- data type

Each dimension is divided into the following subcategories:

Type of analysis: Summarizing information, groups & classification, dependence & prediction.

As discussed in Section 2, the type of analysis categorizes papers according to analytical task which the presented approach is supposed to carry out.

Level of integration: Visualization as presentation, semi-interactive methods, tight integration.

The level of integration describes how tightly computational tools and algorithms are integrated into the visual analytics system to enable the user to steer the automated analytical process (see Section 2).

Visualization technique: Geometric, table-based, icon/glyph-based, pixel-based, graph.

Visualization techniques are categorized according to KEIM ET AL. [KK96, Kei01] and in addition to that, the category *table-based* has been added to emphasize common table-based visualizations, such as table lens and heat maps.

Data type: Genomics, proteomics, metabolomics, text, graph, image, multivariate data.

Besides common data types in the bio-medical domain (text, image), the category *data type* contains all main ⋆omics-data types (genomics, proteomics, metabolomics). For general and novel visual analytic approaches, which do not target the bio-medical domain in particular, the general categories *multivariate data* and *graph analysis* were used.

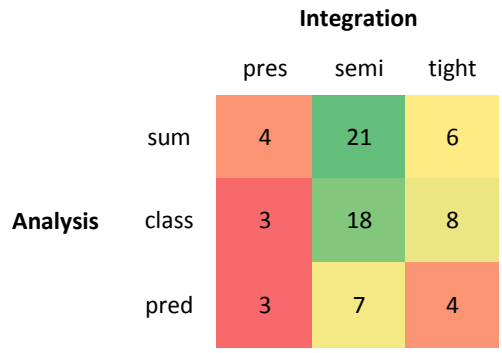

Figure 7: *Integration level vs. type of analysis: Most visual analytics systems are of the integration level* semi-interactive methods *for both analysis task (*summarizing information *and* groups & classification*). There is still a lack of prediction systems that tightly integrate the user.*

Table 1 summarizes the surveyed works across the level of integration and type of analysis dimensions. It appears that a majority of techniques integrates analysis and visualization to some degree, with a good amount of works even with higher levels of integration.

If we look at the level of integration by visualization type according to Table 2, we find that a majority of methods are in the class of geometric transform-based and table-based techniques, and for these works, also semi- or tight integration levels are observed.

This indicates to us a trend towards higher levels of integration of visualization, interaction and data analysis, a trend which appears natural in face of growing data volumes. We also observe that there are rather few works in icon-based techniques and with tight integration. Generally, icon- and pixel-oriented techniques realize high-dense information displays, eventually utilizing every pixel to represent a data record or dimension. One explanation for the lower level of integration could be, that pixel and some icon displays are hard to interact with directly, as precise selection may be more difficult than with other, less dense visual representations.

We point out that while we have done this selection and categorization of works to the best of our knowledge, there are of course many cases where one could argue for one category instead of the other. As this is a difficult task, and as demonstration videos are not available for all of the works, it remains challenging to assess e.g., the level of integration. Also, while we aimed for a representative literature selection in the field, we may well have missed relevant works of researchers. Therefore, the given categorization represents our understanding, but may be subject to further refinement, reorganization, and extension by dimensions and approaches in future work.

5. Open Problems

There is still a huge demand for specialized and highly integrative visual analytics approaches in the biomedical domain. Many highly integrative approaches are general approaches, but it can also be applied on particular sub-fields of bio-medicine. Therefore, there is a need of further research on specialized applications that integrate the users' knowledge to the analytical process.

As many approaches support a single data type, there is an even larger lack of solutions, which integrate multiple data sets to analyze them in parallel. Based on this analysis, an even broader and more detailed investigation of current research would reveal, how many systems already support multiple data sets.

As therapy outcomes as natural text and a lot of medical knowledge is located in books, the automated analysis of text is still a hot topic and needs further research. In addition to that, new approaches for graph analysis and graph mining are needed to analyze complex graphs (hairballs) in a comprehensible way.

However, systems biology aims to combine multiple data sets to analyze multiple layers of a biological system at once. The ultimate goal of such biomedical systems is to understand biological or pathological processes as a whole. Such a system would interlink all related data sets (e.g., images, text, measured values, scans) and offer visual analytics to support experts to explore the data while integrating personal domain knowledge. Such sophisticated visual analytics systems will boost evidence-based medicine to a new level.

	Visualization as Presentation	Semi-interactive Methods	Tight Integration
Summarizing Information	[DCP*10], [MTW*08], [NCD*10], [SMM*14]	[BSK*15], [BTK11], [BZC*03], [CHB*12], [CK07], [FJA*11], [FSF*13], [FWG09], [HMJ*00], [JBS08], [JJ09], [KFH10], [KKM13], [KHK12], [MMDP10], [ODH*07], [PS09], [TRM12], [TGR*15], [Wea04], [YHW*07], [YWRH03]	[EBN13], [EHM*11], [IMI*10], [NM13], [TFH11], [WM04]
Groups & Classification	[DLZ07], [KBH06], [TA08]	[AEEK99], [DGN06], [GRVE07], [GWR09], [Kan12], [KPH*12], [KKM13], [LSP*10], [LSS*12], [MBD*11], [MK08], [PLS*12], [RK04], [RPN*08], [SBVLK09], [SS02], [WFH*01], [YNM*13]	[AW12], [CLKP10], [DWHM14], [PTRV13], [RWH*10], [TPRH11a], [TPRH11b], [vdEvW11]
Dependence & Prediction	[KSM*12], [KSB*09], [KHK12]	[BMPM12], [EDF08], [MMP09], [MWS*10], [MP13], [PBK10], [YWRH03]	[BPFG11], [DWHM14], [MME*12], [TLLH12]

Table 1: *Level of integration vs. type of analysis*

	Visualization as Presentation	Semi-interactive Methods	Tight Integration
Geometric	[DLZ07], [KSM*12], [KSB*09], [KHK12], [MTW*08], [NCD*10]	[BSK*15], [BTK11], [BMPM12], [BZC*03], [CHB*12], [CK07], [DGN06], [EDF08], [FJA*11], [FWG09], [GRVE07], [GWR09], [HMJ*00], [JBS08], [JJ09], [Kan12], [KFH10], [KKM13], [KHK12], [LSS*12], [MBD*11], [MK08], [MMP09], [MMDP10], [MWS*10], [MP13], [ODH*07], [PLS*12], [PS09], [PBK10], [RK04], [RPN*08], [SBVLK09], [SS02], [TRM12], [WFH*01], [Wea04], [YHW*07], [YWRH03], [YNM*13]	[AW12], [BPFG11], [CLKP10], [DWHM14], [EBN13], [EHM*11], [IMI*10], [MME*12], [NM13], [PTRV13], [RWH*10], [TFH11], [TLLH12], [TPRH11a], [TPRH11b], [vdEvW11], [WM04]
Table-based	[KSM*12], [KSB*09], [KHK12], [NCD*10]	[CHB*12], [DGN06], [HMJ*00], [KPH*12], [KKM13], [KHK12], [LSP*10], [LSS*12], [MBD*11], [MMP09], [RK04], [SS02], [TRM12], [Wea04], [YNM*13]	[CLKP10], [DWHM14], [EBN13], [MME*12], [RWH*10], [TPRH11a]
Icon- & Pixel-based	[KHK12], [KSB*09]	[AEEK99], [CHB*12], [GRVE07], [KHK12], [MMDP10], [RPN*08], [SBVLK09], [TRM12], [YHW*07]	[EBN13], [NM13]
Graph	[DCP*10], [DLZ07], [KSM*12], [KBH06], [KSB*09], [SMM*14], [TA08]	[BZC*03], [DGN06], [FSF*13], [HMJ*00], [KKM13], [LSP*10], [LSS*12], [MMP09], [MWS*10], [PLS*12], [PS09], [RK04], [SS02], [WFH*01], [Wea04], [YWRH03]	[AW12], [DWHM14], [GKN*15], [PTRV13], [TPRH11a], [vdEvW11]

Table 2: *Level of integration vs. visualization technique*

Acknowledgements

Parts of this work have been carried out with the K1 COMET Competence Center CBmed, funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG.

References

[AEEK99] ANKERST M., ELSEN C., ESTER M., KRIEGEL H.-P.: Visual classification: an interactive approach to decision tree construction. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), ACM, pp. 392–396. 7

[AK06] ALFRED R., KAZAKOV D.: Data summarization approach to relational domain learning based on frequent pattern to support the development of decision making. In *Advanced Data Mining and Applications*. Springer, 2006, pp. 889–898. 2

[Alp04] ALPAYDIN E.: *Introduction to machine learning*. MIT press, 2004. 2

[AM03] AEBERSOLD R., MANN M.: Mass spectrometry-based proteomics. *Nature 422*, 6928 (2003), 198–207. 4, 5

[AW12] AHMED Z., WEAVER C.: An adaptive parameter space-filling algorithm for highly interactive cluster exploration. In *Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 13–22. 7

[BL10] BERTINI E., LALANNE D.: Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 9–18. 3

[BMPM12] BOOSHEHRIAN M., MÖLLER T., PETERMAN R. M., MUNZNER T.: Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1235–1244. 7

[BPFG11] BERGER W., PIRINGER H., FILZMOSER P., GRÖLLER E.: Uncertainty-aware exploration of continu-

ous parameter spaces using multivariate prediction. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 911–920. 7

[BSK*15] BEHRISCH M., SHAO L., KWON B. C., SCHRECK T., SIPIRAN I., KEIM D.: Quality Metrics Driven Approach to Visualize Multidimensional Data in Scatterplot Matrix. *Proceedings of the Eurographics Conference on Visualization (Poster paper) 17* (2015), 6. 7

[BSM*15] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A Visual-Interactive System for Prostate Cancer Cohort Analysis. *IEEE Computer Graphics and Applications 35*, 3 (2015), 44–55. 5

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: an overview and systematization. *Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2203–2212. 7

[BZC*03] BU D., ZHAO Y., CAI L., XUE H., ZHU X., LU H., ZHANG J., SUN S., LING L., ZHANG N., ET AL.: Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic acids research 31*, 9 (2003), 2443–2450. 4, 5, 7

[CC00] COX T. F., COX M. A.: *Multidimensional scaling*. CRC Press, 2000. 2

[CH67] COVER T., HART P.: Nearest neighbor pattern classification. *Transactions on Information Theory 13*, 1 (1967), 21–27. 2

[CHB*12] CARVER T., HARRIS S. R., BERRIMAN M., PARKHILL J., MCQUILLAN J. A.: Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics 28*, 4 (2012), 464–469. 7

[CK07] CHANDOLA V., KUMAR V.: Summarization–compressing data into an informative representation. *Knowledge and Information Systems 12*, 3 (2007), 355–378. 2, 7

[CLKP10] CHOO J., LEE H., KIHM J., PARK H.: ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Symposium on Visual Analytics Science and Technology (VAST)* (2010), IEEE, pp. 27–34. 7

[DCP*10] DEMIR E., CARY M. P., PALEY S., FUKUDA K., LEMER C., VASTRIK I., WU G., D'EUSTACHIO P., SCHAEFER C., LUCIANO J., ET AL.: The BioPAX community standard for pathway data sharing. *Nature biotechnology 28*, 9 (2010), 935–942. 7

[DGN06] DIETZSCH J., GEHLENBORG N., NIESELT K.: Mayday-a microarray data analysis workbench. *Bioinformatics 22*, 8 (2006), 1010–1012. 7

[DLZ07] DEMŠAR J., LEBAN G., ZUPAN B.: FreevizâĂŤan intelligent multivariate visualization approach to explorative analysis of biomedical data. *Journal of biomedical informatics 40*, 6 (2007), 661–671. 7

[DND*02] DUHAMEL A., NUTTENS M., DEVOS P., PICAVET M., BEUSCART R.: A preprocessing method for improving data mining techniques. Application to a large medical diabetes database. *Studies in health technology and informatics 95* (2002), 269–274. 3

[DWHM14] DING H., WANG C., HUANG K., MACHIRAJU R.: iGPSe: A visual analytic system for integrative genomic based cancer patient stratification. *BMC bioinformatics 15*, 1 (2014), 203. 7

[EBN13] ENDERT A., BRADEL L., NORTH C.: Beyond control panels: Direct manipulation for visual analytics. *Computer Graphics and Applications 33*, 4 (2013), 6–13. 7

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (2008), 1539–1148. 7

[EHM*11] ENDERT A., HAN C., MAITI D., HOUSE L., LEMAN S., NORTH C.: Observation-level interaction with statistical models for visual analytics. In *Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 121–130. 7

[FJA*11] FERNSTAD S. J., JOHANSSON J., ADAMS S., SHAW J., TAYLOR D.: Visual exploration of microbial populations. In *Symposium on Biological Data Visualization (BioVis)* (2011), IEEE, pp. 127–134. 7

[FPSS96a] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI magazine 17*, 3 (1996), 37. 2

[FPSS96b] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM 39*, 11 (Nov. 1996), 27–34. 1, 2

[FSF*13] FRANCESCHINI A., SZKLARCZYK D., FRANKILD S., KUHN M., SIMONOVIC M., ROTH A., LIN J., MINGUEZ P., BORK P., VON MERING C., ET AL.: String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research 41*, D1 (2013), D808–D815. 7

[FWG09] FUCHS R., WASER J., GROLLER M. E.: Visual human+ machine learning. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 1327–1334. 7

[GKN*15] GERASCH A., KÜNTZER J., NIERMANN P., STÖCKEL D., KAUFMANN M., KOHLBACHER O., LENHOF H.-P.: Network-based interactive navigation and analysis of large biological datasets. *it-Information Technology 57*, 1 (2015), 37–48. 7

[GOB*10] GEHLENBORG N., O'DONOGHUE S. I., BALIGA N. S., GOESMANN A., HIBBS M. A., KITANO H., KOHLBACHER O., NEUWEGER H., SCHNEIDER R., TENENBAUM D., ET AL.: Visualization of omics data for systems biology. *Nature methods 7* (2010), S56–S68. 4, 5

[Gro88] GROSSBERG S.: Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks 1*, 1 (1988), 17–61. 2

[GRVE07] GROTTEL S., REINA G., VRABEC J., ERTL T.: Visual verification and analysis of cluster detection for molecular dynamics. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1624–1631. 7

[GWR09] GUO Z., WARD M. O., RUNDENSTEINER E. A.: Model space visualization for multivariate linear trend discovery. In *Symposium on Visual Analytics Science and Technology* (2009), IEEE, pp. 75–82. 7

[HDO*98] HEARST M. A., DUMAIS S. T., OSMAN E., PLATT J., SCHOLKOPF B.: Support vector machines. *Intelligent Systems and their Applications, IEEE 13*, 4 (1998), 18–28. 2

[Hel02] HELLER M. J.: DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering 4*, 1 (2002), 129–153. 4

[HJ14] HOLZINGER A., JURISICA I.: *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges*, vol. 8401. Springer Berlin Heidelberg, 2014. 1, 4

[HK06] HAN J., KAMBER M.: *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006. 2, 3

[HK11] HORGAN R. P., KENNY L. C.: 'Omic' technologies:

genomics, transcriptomics, proteomics and metabolomicssbml):. *The Obstetrician & Gynaecologist 13*, 3 (2011), 189–195. 4

[HKO04] HYVÄRINEN A., KARHUNEN J., OJA E.: *Independent component analysis*, vol. 46. John Wiley & Sons, 2004. 2

[HMJ*00] HUGHES T. R., MARTON M. J., JONES A. R., ROBERTS C. J., STOUGHTON R., ARMOUR C. D., BENNETT H. A., COFFEY E., DAI H., HE Y. D., ET AL.: Functional discovery via a compendium of expression profiles. *Cell 102*, 1 (2000), 109–126. 5, 7

[Hol13] HOLZINGER A.: Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What Is the Benefit of Bringing Those Two Fields to Work Together? In *Availability, Reliability, and Security in Information Systems and HCI*, Cuzzocrea A., Kittl C., Simos D., Weippl E., Xu L., (Eds.), vol. 8127 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 319–328. 2

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MOLLER T.: Dimstiller: Workflows for dimensional analysis and reduction. In *Symposium on Visual Analytics Science and Technology (VAST)* (2010), IEEE, pp. 3–10. 7

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91. 4

[JBS08] JANICKE H., BOTTINGER M., SCHEUERMANN G.: Brushing of attribute clouds for the visualization of multivariate data. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (2008), 1459–1466. 7

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 993–1000. 7

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: *Visual analytics: Definition, process, and challenges*. Springer, 2008. 2, 3

[Kan12] KANDOGAN E.: Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 73–82. 7

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics, IEEE Transactions on 12*, 4 (2006), 558–568. 7

[Kei01] KEIM D. A.: Visual exploration of large data sets. *Communications of the ACM 44*, 8 (2001), 38–44. 6

[KFH10] KEHRER J., FILZMOSER P., HAUSER H.: Brushing moments in interactive visual analysis. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 813–822. 7

[KHK12] KUHN R. M., HAUSSLER D., KENT W. J.: The ucsc genome browser and associated tools. *Briefings in bioinformatics* (2012), bbs038. 7

[KHS*14] KIESEBERG P., HOBEL H., SCHRITTWIESER S., WEIPPL E., HOLZINGER A.: Protecting anonymity in data-driven biomedical science. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer Berlin Heidelberg, 2014, pp. 301–316. 3

[KK96] KEIM D. A., KRIEGEL H.-P.: Visualization techniques for mining large databases: A comparison. *Knowledge and Data Engineering, IEEE Transactions on 8*, 6 (1996), 923–938. 6

[KKEM10] KEIM D. A., KOHLHAMMER J., ELLIS G., MANSMANN F.: *Mastering The Information Age – Solving Problems with Visual Analytics*. Eurographics Association, 2010. 2, 3

[KKM13] KLEIN K., KRIEGE N., MUTZEL P.: Scaffold hunter: facilitating drug discovery by visual analysis of chemical space. In *Computer Vision, Imaging and Computer Graphics. Theory and Application*. Springer, 2013, pp. 176–192. 7

[KMS*08] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: *Visual analytics: Scope and challenges*. Springer, 2008. 3

[Kob14] KOBAYASHI M.: Resources for Studying Statistical Analysis of Biomedical Data and R. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, Holzinger A., Jurisica I., (Eds.), vol. 8401 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2014, pp. 183–195. 3

[KPH*12] KIM N., PARK H., HE N., LEE H. Y., YOON S.: QCanvas: an advanced tool for data clustering and visualization of genomics data. *Genomics & informatics 10*, 4 (2012), 263–265. 4, 7

[KSB*09] KRZYWINSKI M., SCHEIN J., BIROL I., CONNORS J., GASCOYNE R., HORSMAN D., JONES S. J., MARRA M. A.: Circos: an information aesthetic for comparative genomics. *Genome research 19*, 9 (2009), 1639–1645. 7

[KSM*12] KARR J. R., SANGHVI J. C., MACKLIN D. N., GUTSCHOW M. V., JACOBS J. M., BOLIVAL B., ASSAD-GARCIA N., GLASS J. I., COVERT M. W.: A whole-cell computational model predicts phenotype from genotype. *Cell 150*, 2 (2012), 389–401. 7

[LHM98] LIU B., HSU W., MA Y.: Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (1998), pp. 337–341. 2

[Lin11] LIND A.: Interactive and Exploratory Visual Analysis In Biology. *Seminar in Visualization 358* (2011), 111–126. 5

[LSP*10] LEX A., STREIT M., PARTL C., KASHOFER K., SCHMALSTIEG D.: Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1027–1035. 7

[LSS*12] LEX A., STREIT M., SCHULZ H.-J., PARTL C., SCHMALSTIEG D., PARK P. J., GEHLENBORG N.: Stratomex: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1175–1184. 7

[Mar08] MARDIS E. R.: The impact of next-generation sequencing technology on genetics. *Trends in genetics 24*, 3 (2008), 133–141. 4

[MBD*11] MAY T., BANNACH A., DAVEY J., RUPPERT T., KOHLHAMMER J.: Guiding feature subset selection with an interactive visualization. In *Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 111–120. 7

[MK08] MAY T., KOHLHAMMER J.: Towards closing the analysis gap: Visual generation of decision supporting schemes from raw data. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 911–918. 7

[MM10] MALETIC J. I., MARCUS A.: Data cleansing: A prelude to knowledge discovery. In *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 19–32. 3

[MMDP10] MEYER M., MUNZNER T., DEPACE A., PFISTER H.: Multeesum: A tool for comparative spatial and temporal gene expression data. *IEEE Transactions on visualization and computer graphics 16*, 6 (2010), 908. 7

[MME*12] MALIK A., MACIEJEWSKI R., ELMQVIST N., JANG Y., EBERT D. S., HUANG W.: A correlative analysis process in a

visual analytics environment. In *Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 33–42. 7

[MMP09] MEYER M., MUNZNER T., PFISTER H.: Mizbee: a multiscale synteny browser. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 897–904. 7

[MP13] MUHLBACHER T., PIRINGER H.: A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 1962–1971. 7

[MTW*08] MARTONE M. E., TRAN J., WONG W. W., SARGIS J., FONG L., LARSON S., LAMONT S. P., GUPTA A., ELLISMAN M. H.: The cell centered database project: an update on building community resources for managing and sharing 3d imaging data. *Journal of structural biology 161*, 3 (2008), 220–231. 7

[MWS*10] MEYER M., WONG B., STYCZYNSKI M., MUNZNER T., PFISTER H.: Pathline: A tool for comparative functional genomics. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 1043–1052. 7

[NCD*10] NIELSEN C. B., CANTOR M., DUBCHAK I., GORDON D., WANG T.: Visualizing genomes: techniques and challenges. *Nature methods 7* (2010), S5–S15. 7

[NM13] NAM J. E., MUELLER K.: Tripadvisor^{ND}: A tourism-inspired high-dimensional space exploration framework with overview and detail. *Visualization and Computer Graphics, IEEE Transactions on 19*, 2 (2013), 291–305. 7

[ODH*07] OELTZE S., DOLEISCH H., HAUSER H., MUIGG P., PREIM B.: Interactive visual analysis of perfusion data. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1392–1399. 7

[OGG*10] O'DONOGHUE S. I., GAVIN A.-C., GEHLENBORG N., GOODSELL D. S., HÉRICHÉ J.-K., NIELSEN C. B., NORTH C., OLSON A. J., PROCTER J. B., SHATTUCK D. W., ET AL.: Visualizing biological data – now and in the future. *Nature methods 7* (2010), S2–S4. 5

[PBK10] PIRINGER H., BERGER W., KRASSER J.: Hypermoval: Interactive visual validation of regression models for real-time simulation. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 983–992. 7

[PLS*12] PARTL C., LEX A., STREIT M., KALKOFEN D., KASHOFER K., SCHMALSTIEG D.: enroute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis. In *Symposium on Biological Data Visualization (BioVis)* (2012), IEEE, pp. 107–114. 7

[PS09] PERER A., SHNEIDERMAN B.: Integrating statistics and visualization for exploratory power. 7

[PTRV13] PARULEK J., TURKAY C., REUTER N., VIOLA I.: Visual cavity analysis in molecular simulations. *BMC Bioinformatics 14*, Suppl 19 (2013), S4. 7

[Ris01] RISH I.: An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (2001), vol. 3, IBM New York, pp. 41–46. 2

[RK04] RASMUSSEN M., KARYPIS G.: gcluto: An interactive clustering, visualization, and analysis system. *UMN-CS TR-04 21* (2004). 7

[RP12] RAMESHWARI R., PRASAD T.: Systematic and Integrative Analysis of Proteomic Data using Bioinformatics Tools. *arXiv preprint arXiv:1211.2743* (2012). 4

[RPN*08] RINZIVILLO S., PEDRESCHI D., NANNI M., GIANNOTTI F., ANDRIENKO N., ANDRIENKO G.: Visually driven analysis of movement data by progressive clustering. *Information Visualization 7*, 3-4 (2008), 225–239. 7

[RWH*10] RUBEL O., WEBER G. H., HUANG M.-Y., BETHEL E. W., BIGGIN M. D., FOWLKES C. C., LUENGO HENDRIKS C. L., KERANEN S. V., EISEN M. B., KNOWLES D. W., ET AL.: Integrating data clustering and visualization for the analysis of 3d gene expression data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on 7*, 1 (2010), 64–79. 7

[SBVLK09] SCHRECK T., BERNARD J., VON LANDESBERGER T., KOHLHAMMER J.: Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization 8*, 1 (2009), 14–29. 7

[SGG*01] SIM I., GORMAN P., GREENES R. A., HAYNES R. B., KAPLAN B., LEHMANN H., TANG P. C.: Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association 8*, 6 (2001), 527–534. 4

[SL91] SAFAVIAN S. R., LANDGREBE D.: A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics 21*, 3 (1991), 660–674. 2

[SMM*14] SALAZAR G. A., MEINTJES A., MAZANDU G., RAPANOËL H. A., AKINOLA R. O., MULDER N. J.: A web-based protein interaction network visualizer. *BMC Bioinformatics 15*, 1 (2014), 129. 4, 7

[SS02] SEO J., SHNEIDERMAN B.: Interactively exploring hierarchical clustering results [gene identification]. *Computer 35*, 7 (2002), 80–86. 7

[TA08] TELEA A., AUBER D.: Code flows: Visualizing structural evolution of source code. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 831–838. 7

[TFH11] TURKAY C., FILZMOSER P., HAUSER H.: Brushing dimensions-a dual visual analysis model for high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2591–2599. 7

[TGR*15] TOENNIES K. D., GLOGER O., RAK M., WINKLER C., KLEMM P., PREIM B., VÖLZKE H.: Image analysis in epidemiological applications. *it-Information Technology 57*, 1 (2015), 22–29. 3, 7

[TJHH14] TURKAY C., JEANQUARTIER F., HOLZINGER A., HAUSER H.: On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, 2014, pp. 117–140. 3, 4, 5

[TLLH12] TURKAY C., LUNDERVOLD A., LUNDERVOLD A. J., HAUSER H.: Representative factor generation for the interactive visual analysis of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2621–2630. 7

[TPRH11a] TURKAY C., PARULEK J., REUTER N., HAUSER H.: Integrating cluster formation and cluster evaluation in interactive visual analysis. In *Proceedings of the 27th Spring Conference on Computer Graphics* (2011), ACM, pp. 77–86. 7

[TPRH11b] TURKAY C., PARULEK J., REUTER N., HAUSER H.: Interactive visual analysis of temporal cluster structures. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 711–720. 7

[TRM12] THORVALDSDÓTTIR H., ROBINSON J. T., MESIROV J. P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* (2012), bbs017. 7

[vdEvW11] VAN DEN ELZEN S., VAN WIJK J. J.: Baobabview: Interactive construction and analysis of decision trees. In *Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 151–160. 7

[Wea04] WEAVER C.: Building highly-coordinated visualizations in improvise. In *Symposium on Information Visualization (INFOVIS)* (2004), IEEE, pp. 159–166. 7

[WEG87] WOLD S., ESBENSEN K., GELADI P.: Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1 (1987), 37–52. 2

[WFH*01] WARE M., FRANK E., HOLMES G., HALL M., WITTEN I. H.: Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies 55*, 3 (2001), 281–292. 7

[WM04] WILLIAMS M., MUNZNER T.: Steerable, progressive multidimensional scaling. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (2004), IEEE, pp. 57–64. 7

[WvdL11] WANG H., VAN DER LAAN M. J.: Dimension reduction with gene expression data using targeted variable importance measurement. *BMC Bioinformatics 12*, 1 (2011), 312. 4

[XW*05] XU R., WUNSCH D., ET AL.: Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*, 3 (2005), 645–678. 2

[YHW*07] YANG J., HUBBALL D., WARD M. O., RUNDENSTEINER E. A., RIBARSKY W.: Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *Visualization and Computer Graphics, IEEE Transactions on 13*, 3 (2007), 494–507. 7

[YNM*13] YOUNESY H., NIELSEN C. B., MÖLLER T., ALDER O., CULLUM R., LORINCZ M. C., KARIMI M. M., JONES S. J.: An interactive analysis and exploration tool for epigenomic data. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 91–100. 7

[YWRH03] YANG J., WARD M. O., RUNDENSTEINER E. A., HUANG S.: Visual hierarchical dimension reduction for exploration of high dimensional datasets. *Proceedings of the Symposium on Data Visualisation 16* (2003), 19–28. 7