

Report from Dagstuhl Seminar 18161

Visualization of Biological Data – Crossroads

Edited by

Jan Aerts¹, Nils Gehlenborg², Georgeta Elisabeta Marai³, and
Kay Katja Nieselt⁴

1 KU Leuven, BE, jan.aerts@kuleuven.be

2 Harvard University, US, nils@hms.harvard.edu

3 University of Illinois – Chicago, US, gmarai@uic.edu

4 Universität Tübingen, DE, kay.nieselt@uni-tuebingen.de

Abstract

Our ability to generate and collect biological data has accelerated significantly in the past two decades. In response, many novel computational and statistical analysis techniques have been developed to process and integrate biological data sets. However, in addition to computational and statistical approaches, visualization techniques are needed to enable the interpretation of data as well as the communication of results. The design and implementation of such techniques lies at the intersection of the biology, bioinformatics, and data visualization fields. The purpose of Dagstuhl Seminar 18161 “Visualization of Biological Data – Crossroads” was to bring together researchers from all three fields, to identify opportunities and challenges, and to develop a path forward for biological data visualization research.

Seminar April 15 – 20, 2018 – <http://www.dagstuhl.de/18161>

2012 ACM Subject Classification Applied computing → Bioinformatics, Applied computing → Life and medical sciences, Applied computing → Life and medical sciences, Applied computing → Bioinformatics

Keywords and phrases imaging, omics, sequence analysis, visual analytics, visualisation

Digital Object Identifier 10.4230/DagRep.8.4.32

1 Executive Summary

Jan Aerts (KU Leuven, BE, jan.aerts@kuleuven.be)

Nils Gehlenborg (Harvard University, US, nils@hms.harvard.edu)

Georgeta Elisabeta Marai (University of Illinois – Chicago, US, gmarai@uic.edu)

Kay Katja Nieselt (Universität Tübingen, DE, kay.nieselt@uni-tuebingen.de)

License  Creative Commons BY 3.0 Unported license

© Jan Aerts, Nils Gehlenborg, Georgeta Elisabeta Marai, and Kay Katja Nieselt

The rapidly expanding application of experimental high-throughput and high-resolution methods in biology is creating enormous challenges for the visualization of biological data. To meet these challenges, a large variety of expertise from the visualization, bioinformatics and biology domains is required. These encompass visualization and design knowledge, algorithm design, strong implementation skills for analyzing and visualizing big data, statistical knowledge, and specific domain knowledge for different application problems. In particular, it is of increasing importance to develop powerful and integrative visualization methods combined with computational analytical methods. Furthermore, because of the growing relevance of visualization for bioinformatics, teaching visualization should also become part of the bioinformatics curriculum.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Visualization of Biological Data – Crossroads, *Dagstuhl Reports*, Vol. 8, Issue 04, pp. 32–71

Editors: Jan Aerts, Nils Gehlenborg, Georgeta Elisabeta Marai, and Kay Katja Nieselt



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

With this Dagstuhl Seminar we wanted to continue the process of community building across the disciplines of biology, bioinformatics, and visualization. We aim to bring together researchers from the different domains to discuss how to continue the BioVis interdisciplinary dialogue, to foster the development of an international community, to discuss the state-of-the-art and identify areas of research that might benefit from joint efforts of all groups involved.

Based on the topics identified in the seminar proposal, as well as the interest and expertise of the confirmed participants, the following four topics were chosen as focus areas for the seminar, in addition to the overarching topic of collaboration between the data visualization, bioinformatics, and biology communities:

Visualization challenges related to high-dimensional medical data. Patient data is increasingly available in many forms including genomic, transcriptomic, epigenetic, proteomic, histologic, radiologic, and clinical, resulting in large (100s of TBs, 1000s of patients), heterogeneous (dozens of data types per patient) data repositories. Repositories such as The Cancer Genome Atlas (TCGA) contain a multitude of patient records which can be used for patient stratification, for high-risk group and response to treatment discoveries, or for disease subtype/biomarker discoveries. Still, patient records from the clinic are used singularly to diagnose patients in the clinic without including likely insights from other sources. Similarly, molecular expression signatures from the omic sources barely impinge on the clinical observations. There is an urgent need to bridge the divide the precision medicine gap between the laboratory and the clinic, as well as a need to bridge the quantitative sciences with biology. Additionally, many precision medicine studies plan to include sensor data (e.g. physical activity, sleep, and other patient-worn sensors) that will add another dimension of complexity that analysis and visualization tools need to take into account.

This highly relevant topic focused on visual analytic tools and collaborations that will promote and leverage notions of patient similarity across the phenotypical scales. Scalable and robust machine learning methods will need to work synergistically to integrate evidence of similarity while meaningful visual encodings should simultaneously summarize and illuminate patient similitude at the individual and group level. This topic is closely related to some of the topics below.

Visualization of biological networks. Modeling the stochasticity of genetic circuits is an important field of research in systems biology, and can help elucidate the mechanisms of cell behavior, which in turn can be the basis of diseases. These models can further enable predictions of important phenotypic cellular states. However, the analysis of stochastic probability distributions is difficult due to their spatiotemporal and multidimensional nature, and due to the typically large number of simulations run under varying settings. Moreover, stochastic network researchers often emphasize that what is of biological significance is often not of statistical significance – numerical analyses often miss small or rare events of particular biological relevance. A visual approach can help, in contrast, in mining the network dynamics through the landscape defined by these probability distributions.

Another major challenge relates to finding “stable behavior” of networks, including those recruited in signal transduction. Multistability and bistability have been often studied in metabolic chemically reactive networks. Necessary conditions have been formulated to imply the emergence of stable phenotypes. However, these methods have been deployed on small networks. Recently many groups have recognized that scalable methods can be explored using steady state or quasi steady state models that are derived from stoichiometry and rate-action kinetics. These unfortunately suffer from the lack of methods that will examine

the large parametric space. Consider this: N interacting molecules imply N^2 interactions and in turn the same order of the governing “parameters” (activation rates and abundances). For even mid-size portions of salient pathways (EGFR, B-cell Receptor activation, etc.) finding stable states is challenging. It is certainly the case that a complete graph is never realized and sparsity and network mining can be used to glean the necessary structure. Design of experiments followed by visualization of parametric spaces will be required to search for these stable points. Furthermore, the huge size of this space needs possibly new scalable approaches for the visualization.

Visualization for pan-genomics. With the advent of next-generation sequencing we can observe the increase of genome data both in the field of metagenomics (simultaneous assessment of many species) as well as within the field of pan-genomics. In metagenomics, the aim is to understand the composition and operation of complex microbial consortia in environmental samples. On the other hand in pangenomics genomes within a species are studied. While originally a pan-genome has been referred to as the full complement of genes in a clade (mainly a species in bacteria or archaea), this has recently been generalized to considering a pan-genome as any collection of genomic sequences to be analyzed jointly or to be used as a reference rather than a single genome.

In bioinformatics, both topics impose a number of computational challenges. For example, a recent review paper by Marschall et al. on “Computational Pan-Genomics: Status, Promises and Challenges” (DOI: 10.1093/bib/bbw089) addresses current efforts in this sub-area of bioinformatics. This area needs novel, qualitatively different computational methods and paradigms. While the development of new promising computational methods and new data structures both in metagenomics and pangenomics can be observed, a number of open challenges exist. One of them in the area of pangenomics is for example the transition from the representation of reference genomes as strings to representations as graphs. However, the important topic of pan-genome visualization has not been addressed in the aforementioned review. Interestingly this has been taken up in a break-out session in a recent Dagstuhl seminar on “Next Generation Sequencing - Algorithms, and Software For Biomedical Applications”, and identified as a topic of urgent interest and demand. One observation for example is that in pan-genomes there are segments of conserved regions interspersed by highly variable regions. Open question here is how to visualize the highly variable regions, or how to interpret its content in the context of its neighborhood. Other open visualization topics involve the visual representation of the graph structure underlying pangenomes.

In the field of metagenomics some common visualization approaches, such as heatmaps or scatter plots in combination with principal component analyses, are used, however, many open challenges exist. In particular those visualization tools that are developed for genomics studies fall short in representing large-scale, high dimensional metagenomics studies. Especially the magnitude of the data presents a challenge to meaningfully represent biologically valuable information from complex analysis results. Thus also in this topic the question of large-scale and heterogeneous data visualization is of central importance.

Curriculum development of biological data visualization. Parallel to the recognized need to teach bioinformatics students about big data in biology, there is a growing need to familiarise students with modern visual analytics methodologies applied to biological data, and to provide hands-on training. While several community members are teaching summer camps, tutorials, and workshops on biological data visualization, many of these educational sessions take the form of an introduction to specific tools. We find ourselves handling similar questions: what is exploratory data visualization, what is visual analytics, which frameworks

to think about visualization exist, how can we explore design space, and how can we visualise biological data to gain insight into them, so that hypotheses can be generated or explored and further targeted analyses can be defined?

Despite the increasing importance of visualization for bioinformatics, there is currently a general lack of integration into the bioinformatics education, and a useful and appropriate curriculum has not yet been developed. In this topic the following questions will be addressed: What should a modern and seminal curriculum for visualization in bioinformatics look like? How far along the introductory visualization courses should this curriculum go, while allowing biological data topics as well? What are the essential topics, and how can comprehensive training be achieved?

The schedule for the seminar was developed by the organizers based on previous successful Dagstuhl seminars. Emphasis was given to a balance between prepared talks and panels and break groups for less structured discussions focused on a selection of highly relevant topics. Three types of plenary presentations were available to participants who had indicated interest in presenting during the seminar: overview talks (20 minutes plus 10 minutes for questions), regular talks (10 minutes plus 5 minutes for questions), and panel presentations (5 minutes per speaker followed by a 20 – 25 minute discussion). The break out groups met multiple times for several hours during the week and reported back to the overall group on several occasions. This format successfully brought bioinformatics and visualization researchers onto the same platform, and enabled researchers to reach a common, deep understanding through their questions and answers. It also stimulated very long, intense, and fruitful discussions that were deeply appreciated by all participants.

This report describes in detail the outcomes of this meeting. Our outcomes include a set of white papers summarizing the breakout sessions, overviews of the talks, and a detailed curriculum for biological data visualization courses.

2 Table of Contents

Executive Summary

<i>Jan Aerts, Nils Gehlenborg, Georgeta Elisabeta Marai, and Kay Katja Nieselt</i>	32
--	----

Program and Participants	38
---------------------------------	----

Discussions and Outcomes	39
---------------------------------	----

Conclusion and Next Steps	56
----------------------------------	----

Overview of Talks	56
--------------------------	----

Spatial Networks in Neuroscience	
----------------------------------	--

<i>Katja Bühlert</i>	56
----------------------	----

Visualizing Public Health Data	
--------------------------------	--

<i>Anamaria Crisan</i>	57
------------------------	----

Big Mechanism Visualization	
-----------------------------	--

<i>Angus Forbes</i>	57
---------------------	----

Network Visualization Challenges	
----------------------------------	--

<i>Karsten Klein</i>	58
----------------------	----

Biological Networks	
---------------------	--

<i>Alexander Lex</i>	59
----------------------	----

Telling Stories With High-D Data in the Clinic	
--	--

<i>Raghuram Machiraju</i>	59
---------------------------	----

Curriculum Panel: Teaching Visualization	
--	--

<i>Lennart Martens</i>	59
------------------------	----

Visualizing and Interpreting Metabolite-Gene Relationships with RaMP	
--	--

<i>Ewy Mathé</i>	60
------------------	----

Visualization of Single Cell Cancer Genomes	
---	--

<i>Cydney Nielsen</i>	60
-----------------------	----

Strategic Graph Rewriting, Network Analysis, and Visual Analytics: challenges and thoughts	
---	--

<i>Bruno Pinaud</i>	61
---------------------	----

The Bio/Life-Sciences need better visualization of statistical network structures	
---	--

<i>William Ray</i>	61
--------------------	----

Making Sense of Large Scale Image Data	
--	--

<i>Jens Rittscher</i>	62
-----------------------	----

High-Dimensional Medical Data Panel: Exploration and Communication in Bio-medical Visualization	
---	--

<i>Timo Ropinski</i>	62
----------------------	----

Metronome – Connecting genotypes and phenotypes	
---	--

<i>Christian Stolte</i>	63
-------------------------	----

Collaborations between VIS / Bioinformatics / Bio Communities	
---	--

<i>Marc Streit</i>	63
--------------------	----

Visualization for Pan-genomes and Meta-genomes <i>Granger Sutton</i>	64
Democratization of Data Science <i>Blaz Zupan</i>	64
Panel discussions	64
Collaboration Panel: From Genomics/Bioinformatics to Visualization – in 5 minutes <i>Jan Aerts</i>	64
Collaboration Panel: Mutual Respect <i>Sheelagh Carpendale</i>	65
Biological Networks Panel: Matching the User’s Mental Map <i>Carsten Görg</i>	65
Biological Networks Panel: Visualising Biological Networks: comparison of trees to graphs <i>Jessie Kennedy</i>	65
Collaboration Panel: Visualization and Visual Analysis of Biomolecular Structures <i>Barbora Kozlíková</i>	66
Curriculum Panel: Designing a curriculum for teaching visualization in bioinformatics <i>Michael Krone</i>	67
Curriculum Panel: Vis is a large number of small problems <i>Martin Krzywinski</i>	67
Biological Networks Panel: Scaffolding Bionetwork Visualization with models and theories <i>Georgeta Elisabeta Marai</i>	67
Pan-Genomics Panel: Some questions and challenges about comparative genomics and pan-genomics <i>Kay Katja Nieselt</i>	68
High-Dimensional Medical Data Panel: High-Dimensional Medical Data <i>Jos B.T.M. Roerdink</i>	68
Biological Networks Panel: Visualisation of Biological Networks: Past, Present, and Future <i>Falk Schreiber</i>	69
Pan-Genomics Panel: Scaling Sequence Comparison for Pan & Metagenomics <i>Danielle Szafir</i>	69
High-Dimensional Medical Data Panel: Living with Algorithms <i>Cagatay Turkyay</i>	70
Acknowledgements	70
Participants	71

Table 1 Schedule of Dagstuhl Seminar 18161 from April 15th through April 20th, 2018.

Monday	Tuesday	Wednesday	Thursday	Friday
Introductions	High-D Medical Data Talks	Curriculum Panel	Curriculum Discussion	Breakouts
Collaboration Talks & Panel	High-D Medical Data Panel	Breakout Reports	Breakout Reports	Breakout Reports
Biological Networks Talks	Pan-Genomics Talks & Panel	Trip to Villa Borg	Breakouts	
Biological Networks Panel	Breakouts	Cloef Hike	Breakout Reports	

3 Program and Participants

An overview of the schedule for the seminar is provided in Table 1.

During the five days of the seminar, a total of 30 prepared presentations were given across five focus areas:

- **Collaboration**
 - *Overview Talk* – Marc Streit
 - *Panel* – Sheelagh Carpendale, Jan Aerts, Barbora Kozlikova
 - **Biological Networks**
 - *Overview Talk* – Falk Schreiber
 - *Regular Talks* – Bruno Pinaud, Katja Bühler, Alexander Lex, Angus Forbes, Karsten Klein
 - *Panel* – Will Ray, Jessie Kennedy, Carsten Görg, Liz Marai
 - **High-Dimensional Medical Data**
 - *Overview Talk* – Raghu Machiraju
 - *Regular Talks* – Cydney Nielsen, Jens Rittscher, Ewy Mathe, Ana Crisan, Christian Stolte, Blaž Zupan
 - *Panel* – Jos Roerdink, Timo Ropinski, Cagatay Turkay, Raghu Machiraju
 - **Pan-Genomics**
 - *Overview Talk* – Granger Sutton
 - *Panel* – Danielle Szafir, Kay Nieselt, Granger Sutton
 - **Curriculum**
 - *Panel* – Lennart Martens, Martin Krzywinski, Michael Krone
- On the second day, participants joined one or in rare cases two breakout groups that focused on problems in these areas. The break out groups met multiple times for several hours during the week and reported back to the overall group on three occasions.
- The breakout groups received detailed instructions to guide their discussions towards tangible outcomes. Specifically, the breakout groups were given the following tasks in addition to the discussion of their focus topic:
- **Day 2**
 - Identify driving questions for a publication
 - Decide what type of publication and venue would be appropriate
 - Create a timeline for the remainder of the week
 - Identify a speaker for the breakout group
 - **Day 3**
 - Create a rough outline of the manuscript and finalize paper type
 - Review closest related work

- **Day 4**
 - Finalize outline
 - Assign manuscript sections to breakout group participants
 - Formulate one paragraph outlining the contributions of the manuscript
- **Day 5**
 - Agree on timeline for deliverables post-seminar

Based on feedback provided at the end of the seminar, this structured approach was well received by the participants and helped them to focus their discussions.

4 Discussions and Outcomes

High-Dimensional Medical Data

The topic on High-Dimensional Medical Data was split into three subtopics: patient similarity, trust, and awareness.

A. Patient similarity

The patient similarity workgroup included the following members: Jan Aerts, James Chen, Arlene Chung, Mirjam Figaschewski, David Gotz, Raghu Machiraju, Jens Rittscher.

Comparing individuals is a common aspect in different levels of working with patient-related data. First, all-versus-all comparisons are relevant in patient stratification (e.g. to select a patient subgroup which is relevant for inclusion in a clinical trial, or to identify how patient populations behave in a public health context) as well as disease stratification. Second, a single patient can also be compared to larger groups, for example to identify the cases that a new patient resembles so that adequate treatment can be selected. In this workgroup, we discussed the context of calculating these similarities, their different types and constituent parts, and developed some recommendations including how visual analytics can fit into the process.

Patient information is collected in a long list of features (pathology, genotype, EHR-based features, lifestyle, treatment response, prognosis, etc), and different approaches were discussed for combining this information. In traditional stratification methods, such as risk scores, focus often lies on ranking patients in a linear order. However, because there is a mismatch between the linear ordering and the high-dimensional nature of patient data, patients with similar rank may be very different. For example, patients with the same risk for hospital readmission may be at risk for very different reasons (see Figure 1).

Early integration of features, on the one hand, allows for generating a more holistic overview of a patient population which allows the identification of e.g. subgroups and the stratification thereof. A prime example of this is the use of topological data analysis, which aims at discerning the underlying “shape” of a complex dataset (see Figure 2).

Other use cases – such as for point-of-care decisions – require a more hierarchical approach where features are considered one or a few at a time, as in a decision tree. Nevertheless, also for the point-of-care use case the placement of a patient in their broader context can be very beneficial. For example, capturing similarities and building reference libraries can allow for a more systematic approach to clinical grading.

Calculating these similarities however does have its challenges, especially when combining several across different categories and scales. We identified a number of issues with the

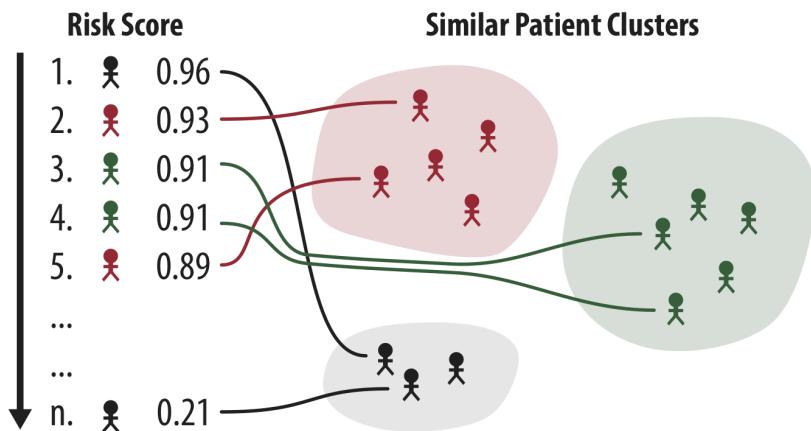


Figure 1 In traditional stratification methods, a mismatch often exists between the linearity of e.g. risk scores, and the high-dimensionality and richness of the underlying patient data.

calculation of similarities. First, we may have clearly-defined similarity in specific cases but a family of similarities can give inconsistent partial ordering. Second, what if the corpus of data is dynamic? Third, data may be sparse and there may be individuals that do not have similarities to any other. Finally, there is a significant problem with missing data, as different patients will have data available for different features.

We believe that a visual analytics (VA) approach – i.e. combining interactive data visualization with automated analysis methods (both statistics and machine learning) – can alleviate some of the issues present in this field, and will allow for opportunities for more informed decisions, more trust in these decisions, and greater objectivity. The VA approach is particularly useful for so-called wicked problems such as this. Wicked problems are described as suffering from finitude in resources and/or knowledge, having very complex interactions between components, and partially depending on values and norms of the people involved. In this case, visual analytics can help in making the patient-comparison process more transparent, interpretable and contextualized for users who leverage those insights into their normal workflows. In particular, VA aims to help generate insights and uncover biases and issues with unknown assumptions as they can make these explicit.

B. Revealing biases in the biomedical research process through visualization

Group consisting of Ryo Sakai, Anamaria Crisan, Ewy Mathé, Torsten Möller, Christian Stolte and Jos Roerdink.

Modern biomedical research has become a complex process involving a growing arsenal of technical devices to generate data. It requires collaboration between disciplines to design experiments, manage and process the collected data, and interpret and analyse the results.

Trust is an essential ingredient for successful collaborative projects, and needs to exist on many levels, in each phase of a project: trust in protocols and measurement accuracy for data collection; in algorithms, models, and processes for data processing; in decision-making and result-finding; and, finally, trust in people and their willingness and ability to collaborate to disseminate the results. We believe that visualization could be used to build trust in the research process and confidence in its outcomes by addressing and revealing biases that can enter the process at each stage.

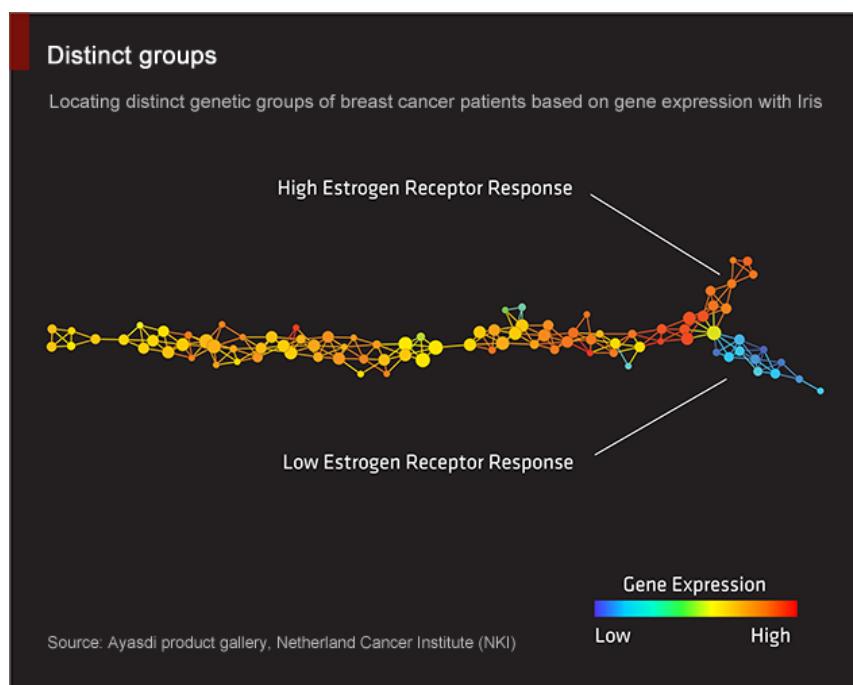


Figure 2 Shape of NKI cancer patient population, coloured by ESR1 expression, and indicating a subpopulation of patients who survived with low ESR1 levels. (taken from Lum et al, 2013; for full details see reference)

Biases can be categorized by their source, along a gradient from machine-centered to human-centered bias (see Figure 3). Over time, the extent and source of a particular bias may change; overlaps can indicate a multitude of factors that need to be taken into account.

At different stages of a research project, the same bias may require different visualizations to reveal its effect in each particular context.

Careful selection of visualizations to highlight each potential bias will help make the analysis transparent, establish a solid basis for quality control and validation, and may also be useful for explaining methods in a publication.

Figure 4 is an example of a structure that can be populated with specific visualization types to address each kind of bias, at each project phase.

By identifying and quantifying potential biases visually, we believe we can help researchers become vigilant to flaws and pitfalls to mitigate risks in the biomedical research process.

Additional Sources:

- <https://eagereyes.org/basics/encoding-vs-decoding>
- http://decisive-workshop.dbvis.de/?page_id=555#101
- <https://www.computer.org/csdl/trans/tg/2006/04/v0421.html>
- https://en.wikipedia.org/wiki/List_of_cognitive_biases
- <https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18>

C. VISard: a card game

Group consisting of Martin Krzywinski, Timo Ropinski, Marc Streit, Cagatay Turkay, Michel A. Westenberg, Blaz Zupan.

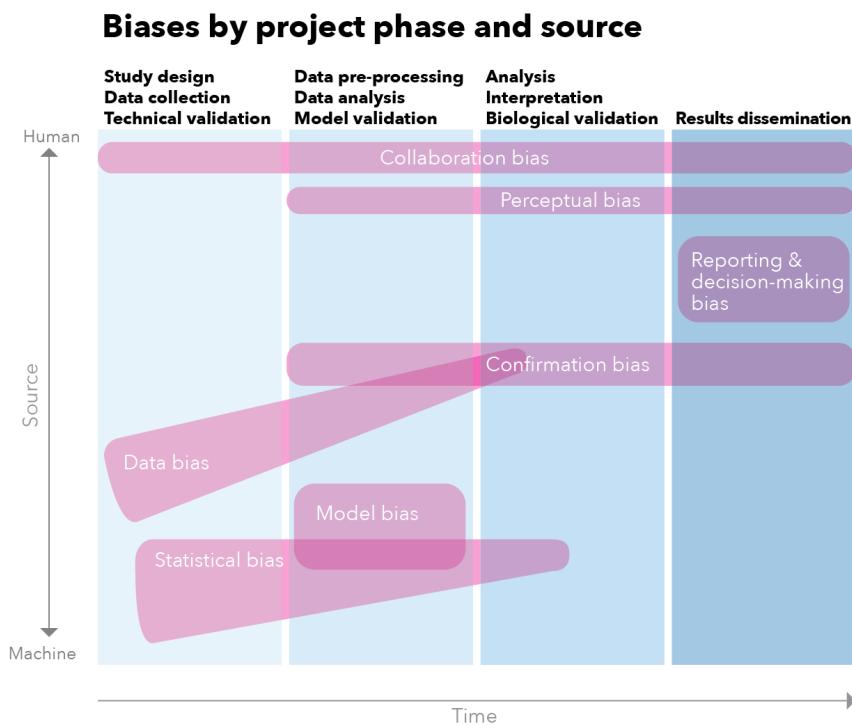


Figure 3 Categorization of bias types.

VISard is a card game and playful take on data visualization education and engagement. It teaches players about the vicissitudes of creating visualizations, dealing with data, users and tasks and the fortunes (good and bad) of practical aspects of computing, design and publishing.

Game goal. The goal of this multiplayer game is to be the first to create a visualization that satisfies requirements, while being subject to constraints, benefitting from lucky breaks and suffering setbacks due to unfortunate events. The game may be played cooperatively or competitively – you can hinder other players to avoid being scooped as you race to publish your visualization.

Game process. Each player creates a visualization by playing various cards. The visualization must meet an acceptable level of accuracy, intuitiveness and engagement.

These acceptable levels are defined by a combination of data set, user and task. These levels are the same for each player and generated at the beginning by randomly drawing requirement cards.

Visualization requirements. The requirements for a successful visualization are determined by three cards drawn at random from the requirements pile at the start of the game.

Data requirement cards (Figure 5) describe a dataset or analysis context such as protein interactions or bacterial phylogeny. Each of these data sets is associated with a specific type, such as a network or tree. These types influence the behaviour of other cards.

Each data set contributes uniquely to the accuracy, intuition and engagement requirement. For example, the bacterial phylogeny card (Figure 6) adds +4 to accuracy, +2 to intuition and +1 to engagement.

	Data collection	Algorithm/ Process/ Models	Decision making/ Results	Collaboration
Statistical bias				
Cognitive bias				
Model/ selection bias/ method				
Perceptual bias				
Data bias (sample bias)				
Reporting bias				
Etc. ...				

Figure 4 Categorization of bias types.

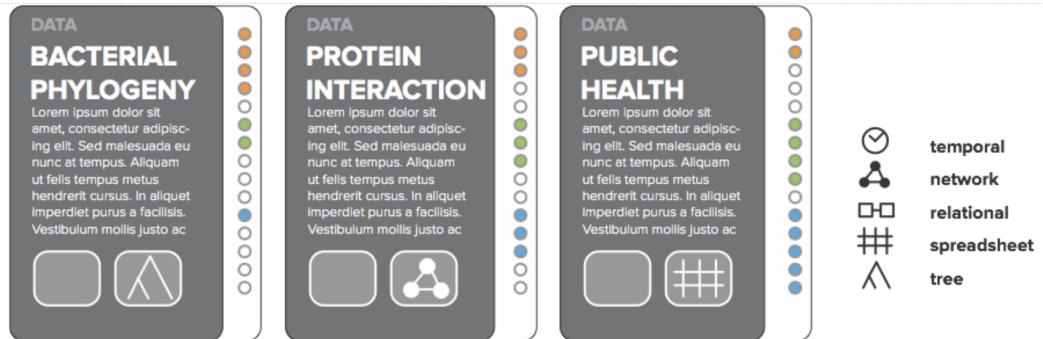


Figure 5 A sample of data requirement cards.

The second requirement is the user. Each of these cards has its own requirements (Figure 7). For example, designing a visualization for a scientist calls for high accuracy (+5) but low engagement (+1). On the other hand, kids require high accuracy (+5) but low engagement (+1).

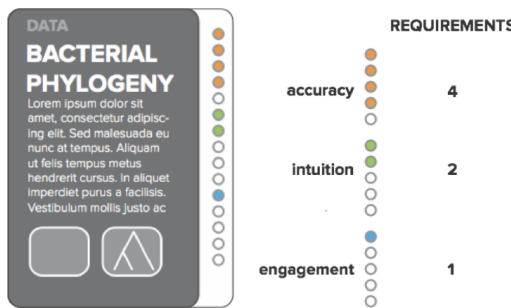
The final requirement card represents the task (Figure 8). Just as for data and user, the task cards contribute to the overall requirement.

The requirement cards are drawn at random – unusual combinations of data, user and tasks are possible! For example, consider the following requirement set: exploring bacterial phylogeny with kids (Figure 9). The total requirements for a visualization for this set of cards is 6 accuracy, 8 intuition and 8 engagement.

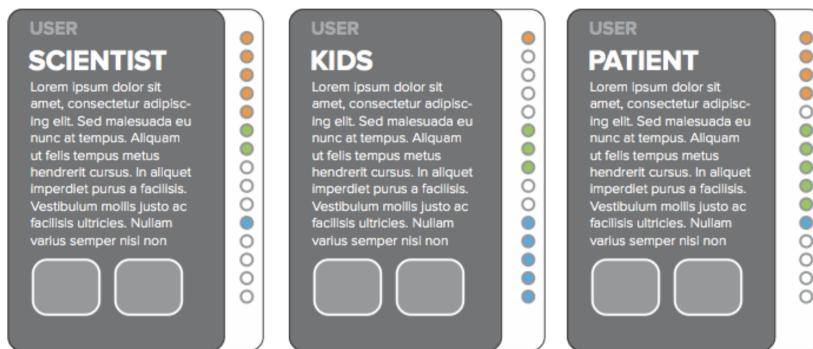
The cards are designed so that they can be stacked to show only the requirement tab to assist in counting the requirements (Figure 10).

Examples of requirement cards with scores. Scores are (accuracy,intuition,engagement,class) where class is an (optional) data type that influences the behaviour of other player cards (e.g. plot type).

- User: scientist (1,5,2), kids (5,1,4), patient (2,3,5), politician/decision maker (3,2,4), student of engineering (4,4,2)



■ **Figure 6** Each requirement card contributes towards a requirement in accuracy, intuition and engagement. A user's visualization must meet or exceed each of these requirements for a successful visualization.



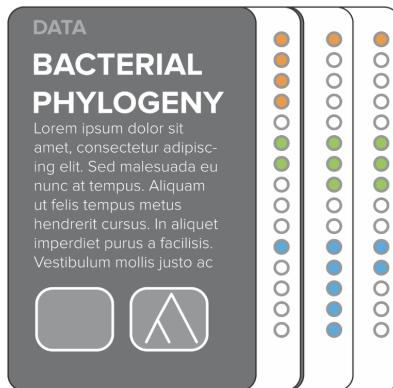
■ **Figure 7** A sample of user requirement cards.



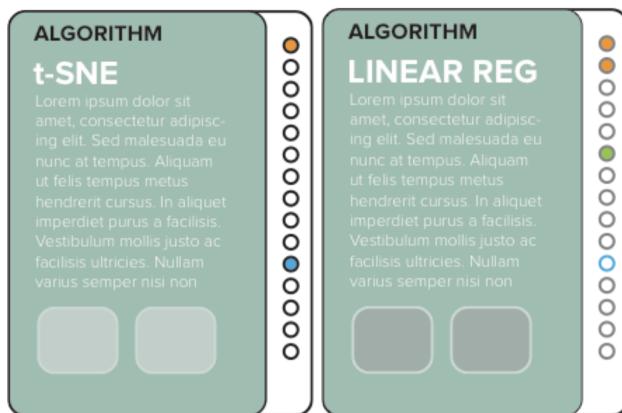
■ **Figure 8** A sample of task requirement cards.



■ **Figure 9** A sample set of data, user and task requirement cards that define the requirements for a successful visualization.



■ **Figure 10** Requirements are shown as circles on the card's tab, allowing cards of a similar type to be stacked.



■ **Figure 11** Example algorithm cards.

- Data: ECG (5,1,4, time series), Country happiness index (5,2,3,geo), Protein interactions (3,2,1, network), Health demographics (5,2,4,table+geo), Bacterial phylogeny (3,4,1,tree), Gene expression (1,5,4,table), Patient collection (1,5,2,text)
- Task: Outlier detection (4,2,2), Trends (2,1,2), Correlation (5,2,5), Cluster analysis (1,2,4), Pedigree analysis (2,3,4)

Building the visualization. Once the requirements have been determined, each player builds a visualization using a combination of analysis, plot, encoding and design cards. Each of these contributes uniquely towards the requirements (Figure 11).

For example, the t-SNE algorithm card contributes +1 to accuracy and +1 to engagement. However, it does not contribute to intuition. On the other hand, linear regression card contributes +2 to accuracy, +1 to intuition but imposes a penalty of -1 to engagement.

The requirement values selected for each card are a combination of our perception of the method, how it might be perceived by users and, importantly, of aspects of the algorithm that we wish to emphasize to the player. For example, the t-SNE card would briefly describe the algorithm and indicate that distance between projected points is not interpretable.

Similarly the plot (visualization) and encoding cards contribute to a user's visualization (Figure 12). Some cards may be incompatible with others – for example, a scatter plot cannot be used on time-course data.



Figure 12 Examples of visualization, encoding and design cards.

Examples of visualization-building cards.

- Design: hot metal colormap (4,3,1), rainbow colormap (5,1,4), log scale (1,3,1), area encoding (4,2,5), length encoding (5,5,5), shape encoding (5,4,4)
- Transformation: k-means, regression analysis, t-SNE, MDS, missing value imputation, PCA
- Visualization type: scatterplot, force-directed layout, treemap, heatmap, matrix layout, mosaic diagram, circos, pie chart, bar chart, parallel coordinates, silhouette plot

Game mechanics – Single player.

1. Draw a data, user and task card randomly. These are your requirements.
2. From a deck of all other cards, player draws 6 cards to make a hand.
3. A round begins by playing a card towards the requirements. Every played card adds or subtracts from the running total of each of the 3 requirement categories. Cards are organized based on type and you can only have one card of each type in play at any given time.
4. Some cards have additional requirements that must be fulfilled. This may prevent playing certain cards or cards of a certain type.
5. Anytime a card played successfully, the user may discard up to 2 cards and draw to complete the hand.

Various end game scenarios are possible. 1. endless play (game ends whenever the player chooses). 2. when a fixed number of cards have been played (e.g. 5, 7, whatever).

Game mechanics – Multiplayer.

1. Draw a data, user and task card for the group and place them in the center. These are the requirements which every player attempts to meet.
2. Each player draws 5 cards from the deck to make a hand.
3. The group chooses who goes first and order proceeds clockwise.
4. A round begins by a player performing one of these tasks (user cannot pass): play a card to build up their solution or perform a task made possible by an action or event card. Cards are organized based on type and you can only have one card of each type in play at any given time. This can be facilitated by stacking the cards of a given type, with the active card placed on top of the stack.

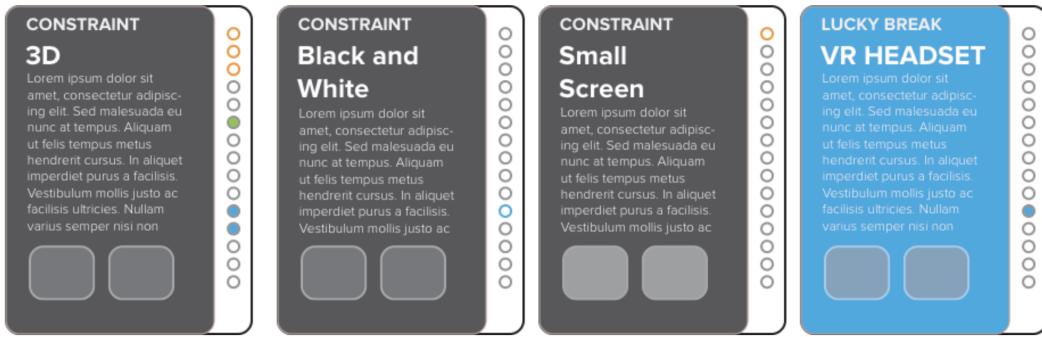
The game ends only when a player chooses to play a publish action card. In order to complete a visualization a player must have at least one card of each type.

Player-specific constraints. To teach users about the challenges of constraints and benefits of new technologies or approaches, mixed with the visualization-building cards are constraint cards (Figure 13).

For example, if a user draws a 3D constraint card then they are penalized for accuracy but obtain a bonus for intuition and engagement – the constraint card contributes -3 to accuracy, +1 to intuition and +2 to engagement. Constraint card examples: Excel, PowerPoint, Tableau, Cytoscape, 3D, black and white, small screen, VR headset.

Events. Some cards in the deck act as events (Figure 14). When they're drawn, the player may be forced (or choose) to perform an action.

The event cards represent fortunate or unfortunate events that may occur during the process of building a visualization. Some cards penalize the player (e.g. data loss, which requires that a player discard one of the active cards in their hand) while others benefit the player (e.g., scoop, which allows the player to steal any card from another player).



■ **Figure 13** Examples of constraint cards which impose a bonus and/or penalty to the user. These are specific to a user and modify their requirements for a successful visualization.



■ **Figure 14** Examples of event cards.

The publish card is a special event. It is required for a user to be able to trigger the end of the game, assuming that they have met the requirements of the visualization.

Event cards.

- Publish – ends game at user’s discretion
- Scoop – take a cards of a certain type from any other player’s visualization. Steal card may allow for more than one type. The stolen card must be immediately played or discarded.
- Swap – exchange a card of a certain type from any other player. Swapped card must be played immediately. There are different kinds of swaps: single card from top of type pile, entire type pile, or entire visualization.
- Data loss – discard the number of cards on the data loss card from the top of type stack
- Reflect disaster – cancel action of another player on your hand or a drawn action card
- Change requirement card – replace one of the requirement cards by a new randomly drawn card
- Modify requirement card – alters the requirements for a given player, place this card face up near your visualization

Game modes. The game may be played with the requirements visible to all players (Figure 15).

The requirement cards are double-sided, with one side printed without the explicit requirements. In this mode, the players must anticipate the requirements of each card. The player who triggers the end of the game takes the chance of meeting the requirements. If



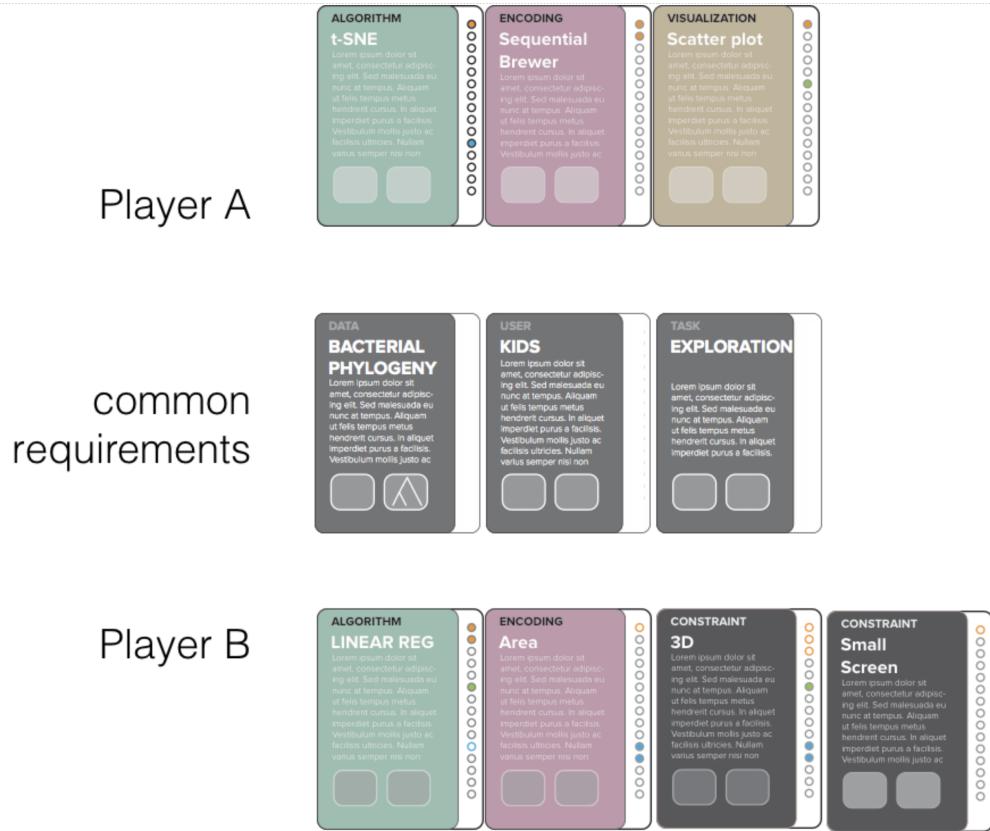
■ **Figure 15** An example game state in which player A and B both see the visualization requirements.

they do not, then the amount that they fall short off is added to the visualizations of other players when tallying the score (Figure 16).

Expansions. The game is scalable through expansion packs. Themes in current news, for example, can be made available as additional data or task cards (Figure 17). Similarly, newly published algorithms or visualizations can be accommodated.

Biological Networks

The biological networks group included the following members: Alexander Lex, Scooter Morris, Jessie Kennedy, Carsten Gorg, Bruno Pinaud, Anne Knudsen, Katja Buhler, Angus Forbes, William Ray, and Liz Marai. In a common meeting, the group brainstormed for open research topics. As a group, we then assessed both individual member interest and expertise in the resulting list of questions and culled the list. After several passes, we converged towards six main topics and the following subgroups; each subgroup produced next a list of keywords to better crystallize the topic, and a list of potential publication venues, along with the publication type (survey versus position versus guidelines etc.):



■ **Figure 16** An example game state in which player A and B do not know the precise requirements. Each player must guess the requirements for each card.



■ **Figure 17** Examples from an ethics expansion pack.

- Topic: Query-networks (details-first, expand on demand) for thousands of nodes
 - Members: A. Lex (Lead), S. Morris, C. Goerg, J. Kennedy, A. Knudsen
 - Keywords: zoom-into-detail vs reorient-for-detail, semantic vs geometric zoom, “too many” nodes vs “too many” edges, “meaning” of thresholds (weight vs relative weight)
 - Venue: Perspective PLoS CompBio
- Topic: Spatiality in neural networks
 - Members: L. Marai (Lead), K. Buhler, A. Forbes
 - Keywords: biological networks, spatial data, 3D coordinates, neurons, atlases, data integration, spatial nonspatial integration, survey, review, design, guidelines, neuroscience, connectome visualization, Hypergraphs, Multilayer networks, constrained layout, multidimensional projections (parallel coords, etc)
 - Venue: Review followed by Taxonomy/Position: TVCG, Nature Methods, Neuroinformatics
- Topic: Visualization for the Rule-Based Modeling of Biological Systems
 - Members: A. Forbes (Lead), B. Pinaud, L. Marai
 - Keywords: rule based model, rule inference, graph rewriting
 - Venue: Review, Bioinformatics/BMC Bioinformatics
- Topic: 10 simple rules to create biological network figures for communication
 - Members: L.Marai (Lead), S. Morris, A. Lex, J. Kennedy, C. Goerg, K. Buhler, B. Pinaud
 - Keywords: Visualization design, Ideas that need keywords/better searches: (data-centric vs user-centric vs task-centric, what to sacrifice for simplicity/informing the user); Is the visualization of both nodes and edges always necessary?
 - Venue: Guideline; PLoS CompBio
- Topic: What Cytoscape needs to do to get vis researchers to work in its web-browser
 - Members: S. Morris (Lead) and everybody else in the group
 - Keywords: Cytoscape, network visualization
 - Venue: (non publishable)
- Topic: Spatial Networks in Bioinformatics
 - Members: W. Ray (Lead), A. Forbes, B. Pinaud, L. Marai
 - Keywords: bioinformatics network visualization
 - Venue: Position/Survey/Guidelines

Additional topics that the group considered were: Comparison, Dynamic Nets, Spatiality in protein networks, Aggregation, Provenance of nets, Multi-attribute nets, and Hypernetwork graphs. These topics of interest could not be tackled due to time constraints, and were slated for discussion at future meetings.

Over the following energetic breakout sessions, the smaller groups converged towards an outline and an abstract for each publication, as well as which group member was in charge of which section. Group leads then contacted editors at potential publication venues. Each group agreed on a timeline for finalizing their target publication, on the platform they were going to use for drafts, and on their preferred means of communication.

Pangenomics

The working group for this topic consisted of the following four members (in alphabetical order):

Kay Nieselt, Jim Procter, Granger Sutton, Danielle Szafir.

After a first discussion round it was agreed to work on the topic ‘Open Visualisation challenges for Pangenomics’. In the first discussion round also the following tasks, questions and open challenges were identified:

- Which are the standard and most commonly used visualisations for pangenomes?
- Which analytics should be connected with the visualisations, in term of visual analytics (VA) software?
- Which questions do researchers in pangenomics ask?
- What type of data feeds into a Vis or VA tool?
- What commonly used visual encodings exist?
- In terms of scalability (growing pan-genomes), what type of aggregation methods are in place or missing?
- For a given pan-genome what is the best computational as well visual approach for an update of a given pan-genome?

In a separate document, many more details for each of these questions and open challenges have been collected.

Topic: ‘The 10 most important visualisations of pangenomes’

for the series ‘Ten simple rules’ in Plos Comp Biol.

The ten rules / most important visualisations are:

1. Central definition of the pan-genome: Gene content is the core of pan-genome. Thus the central visualisation is a matrix view of the gene content, possibly in conjunction with a synteny viewer (see below no. 4). A dichotomy of approaches exists:
 - a. Start at the whole genome alignment and layer features onto it
 - b. Start at the feature level and zoom into the nucleotide level
2. Overview and details on demand to represent gene organisation (an example tool is PanACEA (T. Clarke et al., BMC Bioinformatics 2018)).
3. Visualisation of clustering results for the visual identification of core genome as well as unique genes for individual members. Analytical approaches are for example bi-clustering. Visualisation should allow reordering of rows and columns of matrix.
4. Clustering of species as a dendrogram combined with the gene content matrix, for example as a heatmap. An example tool is ROARY (Page et al., 2015) for a gene-content heatmap or Sequence Surveyor (Albers, Dewey, & Gleicher, 2011) for a heatmap encoding synteny.
5. Allow possibility of tree comparison, for example to highlight leaves that have been rearranged between two trees (example tool is TreeJuxtaposer by Munzner et al., ACM Transactions on Graphics (TOG) 2003)
6. Visually analyse horizontal gene transfer and gene loss (an example tool is panX by Ding, Baumdicker & Neher, NAR 2017)
7. Visually compare intersection and uniqueness of genes between two (sub)sets of a pan-genome (given for example by a dendrogram, an example tool is Hierarchical Sets by Pedersen, 2017). If more than two sets (of species’ gene content) are compared, then UpSet (Lex et al., 2014) is a recommended tool.
8. Represent genomic architectures to visually show rearrangements in genomes (example tool is GenomeRing (Herbig et al., Bioinformatics 2012))
9. Curation of genomic annotation: a visualisation of a pan-genome together with provided gene annotation in each strain can help to identify poorly and even erroneously annotated features. An example tool is Pan-Tetris (Hennig et al., BMC Bioinformatics 2016).
10. Aggregation: a challenge in future is the issue of scalability. A visual tool for pangenomes should offer the possibility to aggregate species and pangenomes.

The group agreed to write such an article and it agreed on a timeline for finalizing the publication. The structure of the article is as follows: we start with (our) central definition of the pangenome with respect to its visual representation and mining possibilities. We will reflect the prokaryotic perspective more than the eukaryotic one and focus on a subset of tasks from prokaryotes. Then the ten rules will follow with example applications.

Collaboration

The following seminar participants joined the collaboration working group: Lennart Martens, Cydney Nielsen, Sheelagh Carpendale, Nils Gehlenborg, Michael Krone, Barbora Kozlikova, Helena Jambor, Falk Schreiber, Karsten Klein.

The collaboration breakout group focused on the question of how one can turn the visualization research conducted in a collaboration between visualization, bioinformatics, and biology researchers, into a stable tool that can serve the bioinformatics and biology community beyond the duration of the collaboration.

The group agreed that collaborations between data visualization, bioinformatics, and biology researchers can be productive and result in progress in all three fields. However, one of the major challenges encountered is that collaborations often end too early, when a prototype visualization has been created and potentially published, but not turned into a usable and maintainable tool for the bioinformatics and biology communities. The group identified the vastly different timescales of conducting and publishing research in these three communities as a major driver behind such undesirable outcomes. In essence, there is a period in which the visualization collaborator has already gained close to maximum cumulative value from the collaboration, when the bioinformatician and the biologist have not yet gained much value from the collaboration (see Figure 18). The group termed this period the “valley of death”, because it is the phase during which many collaborations fall apart.

Next, the group identified the reasons for why the valley of death exists, considering the perspectives of visualization, bioinformatics, and biology researchers. For visualization researchers, the main concerns are lack of incentives to move beyond a prototype and create a usable tool that would not result in a visualization venue paper, visualization researchers might lack the software engineering skills to build a production quality tool and the inability to attract and retain professional software developers, as well as a the lack of appreciation for a stable tool that could serve as a basis for future research. From the perspective of bioinformatics researchers, the biggest concerns are the need to publish a usable tool rather than a prototype, the time required to develop a usable tool, as well as the need to move from a feature-laden prototype to a streamlined tool that is focused on core functionality but stable. Biologists are generally less concerned about the valley of death, as the insight needed for their research question might be provided by the prototype – with the data loaded by the visualization researcher – or a a tool built based on the prototype.

Based on these observations, the group discussed what would need to happen in visualization, bioinformatics, and biology research for more collaborations to successfully cross the valley of death. A visualization researcher would be incentivized to continue a collaboration if there was recognition that a finished tool can offer value for future work, the consideration of publishable contributions to the tool (e.g. an evaluation or systems paper), if there was agreement that trans-community research is valuable, if there was more recognition for visualization work published outside the visualization community, and finally that usable tools will likely be helpful in establishing future collaborations. Bioinformatics and biology

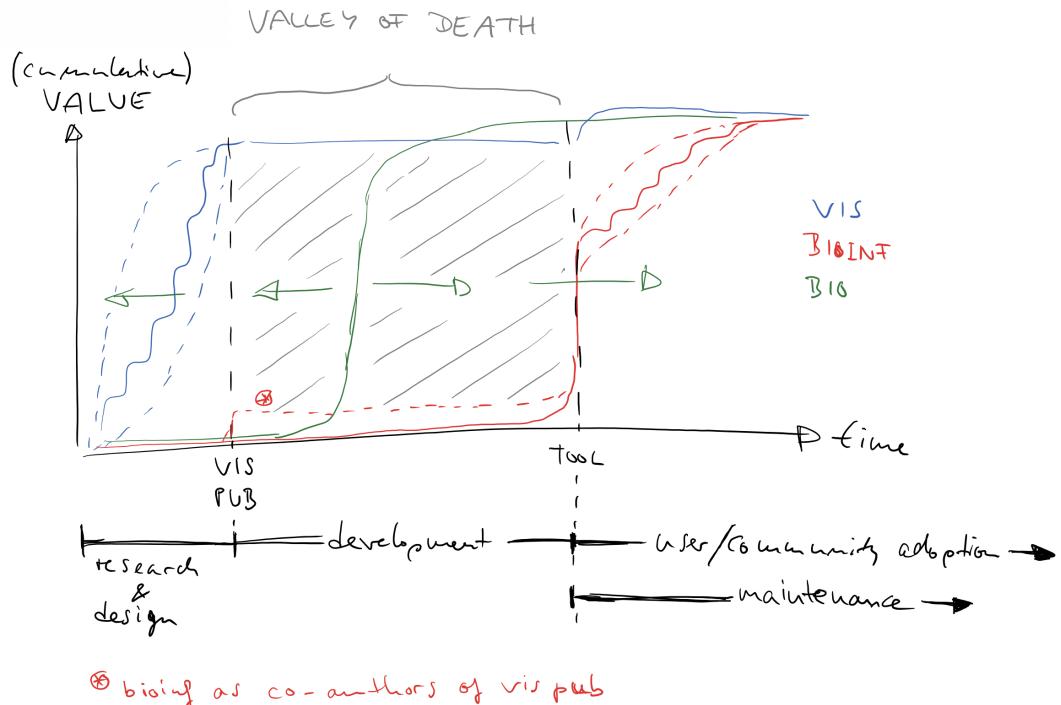


Figure 18 A draft diagram produced during the seminar that illustrates the concept of the valley of death.

researchers could support the process by allocating resources for the transition from prototype to tool, which is explicitly supported by some funding agencies. Additionally, biologists and bioinformatics should use the prototype to generate interest in the method in their communities, further incentivizing the development of a tool.

The key take away identified by the group is that everyone in the collaboration needs to have awareness about the valley of death and there needs to be agreement by among all collaborators about how the valley of death will be crossed in this particular context.

Finally, the group identified several representative examples from different biological and data visualization domains, that illustrate characteristics of both successful and unsuccessful collaborations. Based on these examples, the group formulated a structure for the ideal collaboration, which is broken down into three stages. The first stage would result in a prototype and publication in the visualization literature, the second stage would result in joint publications in biology and bioinformatics venues describing a tool based on the prototype and applications of that tool, and the third stage would be adoption of the tool by the community and transition from active development to long-term maintenance.

Curriculum

The seminar discussion of a curriculum took place in three phases. In the first phase, the seminar participants dedicated one breakout session and a plenary session to discussing the contents of a curriculum in biology visualization. For the first breakout session, the seminar participants partitioned spontaneously in five groups. We took advantage of the splendid weather and grouped around the tables in the yard outside the dining room. Each group

discussed the core competencies required in a biology visualization course, and reported back to the plenary.

In the second phase, a small subset of participants, representing each group, dedicated one additional breakout session to summarizing the output of the seminar's work in this direction. Representatives Torsten Möller, Liz Marai, Danielle Albers-Szafir, William Ray, and Bruno Pinaud collected the notes from all working groups and compiled a taxonomy for the contents of such a biology visualization course. One result that emerged from this discussion was that different audiences have different content needs: for example, someone teaching molecular biology visualization will need to cover the rendering pipeline, while someone teaching genomic visualization will not. The result of this intensive work and discussion was a matrix table of contents, with sections mandatory for all such courses, and with optional topics depending on the data type (see table below).

I. Cross-Cutting Processes.

1. Why Visualization?
2. Tasks+Data+Workflows
3. Design Principles + Typography (both process, e.g., prototyping, and visualization design)
4. Evaluation
5. Provenance (optional)
6. Ethics (optional)
7. Rendering Pipeline (optional)

II. Applications. Choose topic(s) of interest, e.g. a subset of Geospatial Data Images, Networks, Populations & Sets, Sequences & Genomes, Tables, Text, and Three-Dimensional Structures. Cover the following topics for each application:

1. Color
2. Perception
3. Visual Encodings
4. Facets
5. Interaction
6. Summarization

III. Additional Characteristics of Data.

1. Temporality
2. Scale / Multi-scale
3. Uncertainty

Each cell in the resulting matrix will be crowd-sourced and moderated by a designated contributor. Each such cell will contain metadata and links to example teaching materials, organized along the following categories:

- Moderator
- Table of Contents
- Instructional Videos (Brief, 5-15 minutes)
- Reading Materials
- Examples (good and bad)
- Exercises & Summative Evaluations
- Tools
- Tutorials
- Example Courses (including durations & schedules)

- Learning Outcomes (Bloom's Taxonomy)
- Lecture Materials

In the third phase, the results from this consolidating work were reported and discussed in another plenary session. The seminar participants were extremely pleased with the outcome. A lively discussion of the logistics for completing the cells of the curriculum table followed.

5 Conclusion and Next Steps

In the final plenary session all participants of the seminar discussed the possibility of a follow-up meeting or even the possibility to have a regular Dagstuhl seminar about the topic of large-scale biological data visualization. An overwhelming majority of the group voted for a follow-up or even regular meeting. This was also confirmed in the Dagstuhl survey. Finally, the result of the survey showed that the scientific quality of the seminar was rated as ‘outstanding’ (as a median). Thus, the organizers of this seminar would like to discuss possibilities for repeating this seminar with the Dagstuhl directors and staff.

6 Overview of Talks

6.1 Spatial Networks in Neuroscience

Katja Bühler (VRVis – Wien, AT)

License  Creative Commons BY 3.0 Unported license
© Katja Bühler

In my talk, I addressed one of the questions given as a point for discussion in context of the Networks Panel. “What type of spatial data (3D coordinates), if any, show up in biological networks?” Understanding how the brain works is currently subject of large scale brain initiatives worldwide generating huge amounts of data at all scales. The brain is a spatial structure and consequently also the data underlying neurocircuit research is inherently spatial. I started with an overview on different kinds of spatial data and spatial networks being central to neurocircuit research and how standard brains and hierarchical brain parcellations are used to establish a spatial and semantic reference system. These reference systems allow us to integrate data across scales and types, to perform data aggregation to reduce complexity and to provide important anatomical and functional context for neuroscientists. I presented several examples illustrating how these spatial data characteristics can be exploited to create comprehensive network visualizations, to design data structures ensuring interactivity on large scale datasets and to fuse heterogeneous network and non-network data across scales. A discussion of open challenges and requirements on visualizations and interactive visual analytics systems for supporting neuroscientists in their daily research tasks concluded my talk.

6.2 Visualizing Public Health Data

Anamaria Crisan (*University of British Columbia – Vancouver, CA*)

License  Creative Commons BY 3.0 Unported license
© Anamaria Crisan

Background. Genomic epidemiology integrates next-generation sequencing data from surveillance programs and outbreak investigations with administrative datasets, providing a rich pool of data to inform public health decision-making. Bioinformatics pipelines culminating in data visualizations are often used to explore, interrogate, and communicate these complex integrated datasets, but while the bioinformatics tools underlying these platforms are rigorously tested and evaluated, the resulting data visualizations are often created on an ad hoc basis.

Methods. We have conducted a systematic review of the microbial genomic epidemiology literature from the past ten years to survey existing visualizations and to systematically characterize those visualizations by creating a why-what-how annotation code set that describes why the visualization was created (e.g. to show disease transmission in a hospital), what data were used (e.g. genomic data, event data, outcome data), and how the data was visualized (e.g. phylogenetic tree, timeline). To populate the why-what-how code set in a reproducible, transparent, and timely manner we have also created a pipeline that uses text mining and topic modelling to understand why a visualization was created followed by annotation using online open source software borrowed from image analysis research to derive the what and how components of the code set. Together the components of the why-what-how code set form the basis of a typology.

Results. We have developed GEViT (Genomic Epidemiology Visualization Typology), which allows researchers to systematically characterize and analyze visualizations developed specifically for microbial genomic epidemiology applications. Our initial findings show that different visualizations for a common objective (why) incorporated different data types (what) and used a variety of approaches to visualize these data (how), from colour to shapes to textual annotations. The preliminary data visualization corpus and associated code set have been compiled into a searchable gallery with suggestions of best practices that researchers and public health officials can use to guide data visualization efforts to communicate findings, or inform the design of data visualization components within analytic tools.

Conclusions. Through the development of GEViT, we demonstrate how it is possible to reason systematically about data visualization design and analysis. We anticipate that the GEViT resources will provide a comprehensive framework that allows researchers and healthcare stakeholders to design and analyze visualizations that facilitate the exploration and interpretation of complex healthcare datasets.

6.3 Big Mechanism Visualization

Angus Forbes (*University of California, Santa Cruz, US*)

License  Creative Commons BY 3.0 Unported license
© Angus Forbes

My talk presents a series of visualization projects related to the “Big Mechanism” program, supporting a range of tasks broadly relating to the assembly and execution of biological networks extracted from biomedical texts. Understanding the complex processes of life

requires multiple points of view, enabling a wide range of analysis tasks. This understanding, especially when drawn from contemporary heterogeneous big datasets, is often only a working model which is likely to undergo revision, and some of the visualization tools I present explicitly indicate where our knowledge comes from, i.e., which databases, which cell lines, which articles, which sentences. Designing accurate representations of biological data is in itself an interesting research topic, but it is also important to create representations that support useful ways of analyzing this data, and another series of tools I present utilize novel encodings to facilitate reasoning about the dynamics of and causal relationships within complex biological systems.

Overview of BioVis projects:

1. P Murray, F McGee, and AG Forbes. A taxonomy of visualization tasks for the analysis of biological pathway data. *BMC Bioinformatics* 18(2), 2017. https://creativecommons.soe.ucsc.edu/pdfs/Murray_BioPathTaxonomy_BMCBioinformatics2017.pdf
2. P Boutillier, M Maasha, X Li, HF Medina-Abarca, J Krivine, J Feret, I Cristescu, AG Forbes, and W Fontana. The Kappa platform for rule-based modeling. *Bioinformatics* 34(15), 2018.
https://creativecommons.soe.ucsc.edu/pdfs/Boutillier_KappaPlatform_BioVis2018.pdf
3. AG Forbes, A Burks, K Lee, X Li, P Boutillier, J Krivine, and W Fontana. Dynamic influence networks for rule-based models. *IEEE Transactions on Visualization and Computer Graphics* 24(1), 2018.
https://creativecommons.soe.ucsc.edu/pdfs/Forbes_DIN-Viz_VAST2017.pdf
4. AG Forbes, K Lee, G Hahn-Powell, MA Valenzuela-Escárcega, and M Surdeanu. Text annotation graphs: Annotating complex natural language phenomena. LREC 2018. https://creativecommons.soe.ucsc.edu/pdfs/Forbes_TAG_AnnotatingComplexNLPPhenomena_LREC_2018.pdf
5. TN Dang, P Murray, J Aurisano, and AG Forbes. ReactionFlow: An interactive visualization tool for causality analysis in biological pathways. *BMC Proceedings* 9(6), 2015.
https://creativecommons.soe.ucsc.edu/pdfs/Dang_ReactionFlow_BioVis2015.pdf

6.4 Network Visualization Challenges

Karsten Klein (Universität Konstanz, DE)

License  Creative Commons BY 3.0 Unported license
© Karsten Klein

Network visualization has come a long way and there are many solutions for problems that were posed 10-20 years ago. However, some problems are not really solved, and new issues arise as more and more data is available, integrated, and also tasks become more complex. Notable examples are visual comparison and visualization of dynamic networks. In addition, new technology is available the affordances and requirements of which are often ignored in the visualisation concept design. I give an overview on network visualisation from my point of view, list the most pressing challenges, and give a few examples from my current research.

6.5 Biological Networks

Alexander Lex (*University of Utah – Salt Lake City, US*)

License  Creative Commons BY 3.0 Unported license
© Alexander Lex

There is a variety of biological data types that can be modeled as networks. Most of these networks are more valuable if they are considered in the context of node and edge attributes. In this talk I present some layout adaption/linearization strategies to visualize such multivariate networks.

I also introduce a distinction between overview and local tasks, those that are concerned with specific nodes, and argue that local network analysis tasks are more common. I present some techniques that are optimized to ensure readability for local network analysis tasks.

6.6 Telling Stories With High-D Data in the Clinic

Raghu Machiraju (*The Ohio State University – Columbus, US*)

License  Creative Commons BY 3.0 Unported license
© Raghu Machiraju

It is typical for multiple data to be used in the clinic for diagnosis. However, diagnosis in the clinic is stymied by genetic heterogeneity which often results in different outcomes for the same treatment. Patient stratification and biomarker discovery is needed while using multiple data. In this talk, we discuss how “visual stories” can help with gleaning disease etiology and lead to better patient stratifications. Many of these visual stories use interpretable representations. A case is also made for data-driven and often uninterpretable representations especially obtained from deep learning methods.

6.7 Curriculum Panel: Teaching Visualization

Lennart Martens (*Ghent University, BE*)

License  Creative Commons BY 3.0 Unported license
© Lennart Martens

Teaching visualization can follow an overall scheme that works for most courses (or curricula). This schema focuses (in that order) on the following answers to the questions that a student would like answered. (i) What is the topic of the course (centred on definitions and/or description)? (ii) Why is this topic important? (iii) What is the objective of this course/curriculum (what are the learning outcomes)? (iv) What can be done to reach this course/curriculum objective (transferring the relevant knowledge teaching the actual skills)? (v) Learn how to apply these skills in practice. With regards to the content of such a course, some elements that come to mind are listed next. Start from poor examples of visualizations, and critique these. Teach the basics of human perception. List and describe visual elements, and what each is good for (possibly extending this with what each of these is not good for, similar to the format for software design patterns). List and describe graphical representations and their uses in the same overall way. List and describe existing frameworks for visualization, and teach elementary considerations related to how to make libraries or

plugins in such tools. The practical sessions can be based on improved visualizations for poor examples that the students have found.

6.8 Visualizing and Interpreting Metabolite-Gene Relationships with RaMP

Ewy Mathé (Ohio State University – Columbus, US)

License  Creative Commons BY 3.0 Unported license
© Ewy Mathé

Joint work of Ewy Mathé, Elizabeth Baskin, Senyang Hu, Andrew Patt, Jalal K. Siddiqui, Bofei Zhang
Main reference Bofei Zhang, Senyang Hu, Elizabeth Baskin, Andrew Patt, Jalal K. Siddiqui, Ewy Mathé: “RaMP: A Comprehensive Relational Database of Metabolomics Pathways for Pathway Enrichment Analysis of Genes and Metabolites”, *Metabolites*, 8(1). pii: E16, 2018.
URL <https://doi.org/10.3390/metabo8010016>

The value of metabolomics in translational research is undeniable and metabolomics data is increasingly being generated in large cohorts, alongside other omics data such as gene expression. Analysis of these integrated datasets and functional interpretation of disease-associated metabolites is difficult, and is often hampered by the lack of user-friendly computational tools. With this in mind, we developed RaMP (Relational database of Metabolomics Pathways), which integrates biological pathways from KEGG, Reactome, WikiPathways, and HMDB. The database is accessible directly (mysql dump) or through an R package that is publicly available via GitHub (<https://github.com/Mathelab/RaMP-DB>) and includes detailed documentation on installation and usage. The next steps are to visualize the contents of the database to evaluate metabolite annotations between different databases and to create visual approaches to enable toggling between different types of information (e.g. biological pathways, chemical information, etc.). During this process of developing tools, it is important to balance out generalized/simple tools, that are easier to implement and arguably more user-friendly, and domain-specific/tailored tools, that are powerful and flexible but require in-depth understanding. In all cases though, robustness and reproducibility should be integrated.

6.9 Visualization of Single Cell Cancer Genomes

Cydney Nielsen (BC Cancer Agency – Vancouver, CA)

License  Creative Commons BY 3.0 Unported license
© Cydney Nielsen

Joint work of Cydney Nielsen, Maia Smith, Samantha Leung, Viktoria Bojilova, Oleg Golovko, Daniel Machev, Sohrab Shah

Cancer development is an evolutionary process driven by mutation. Single cell genomics is changing our ability to quantify tumour heterogeneity and observe the dynamics of genetically distinct cells over time and anatomical space. This rich research domain offers many visualization challenges and I will highlight four pressing issues that potentially generalize to other areas of biomedical research: (1) designing new visual representations; (2) creating interfaces that serve a broad spectrum of users; (3) achieving responsive interactivity with large and varied data; and (4) integrating with the analytical process. In conclusion, I would encourage us as a community to further integrate our visualizations into the relevant analysis

workflows, such that interactive visualization is increasingly embraced by the bioinformatics and biology communities as a central analysis methodology, rather than niche.

6.10 Strategic Graph Rewriting, Network Analysis, and Visual Analytics: challenges and thoughts

Bruno Pinaud (University of Bordeaux, FR)

License  Creative Commons BY 3.0 Unported license
© Bruno Pinaud

Joint work of Maribel Fernández, Hélène Kirchner, Bruno Pinaud, Jason Vallet
Main reference Maribel Fernández, Hélène Kirchner, Bruno Pinaud, Jason Vallet: “Labelled graph strategic rewriting for social networks”, J. Log. Algebr. Meth. Program., Vol. 96, pp. 12–40, 2018.
URL <http://dx.doi.org/10.1016/j.jlamp.2017.12.005>

In my 10-minute talk I have presented some challenges and thoughts about rule-based modelling and the usage of visualisation to steer every step of the workflow: model design, simulation, then analysis. This talk is based on my work on Porgy (<http://porgy.labri.fr>) and the collaboration with Maribel Fernandez (King’s College London), Hélène Kirchner (Inria, France) and Guy Melançon (U. Bordeaux, France).

I started with a quick reminder about Graph Rewriting which is all about designing executable specifications of complex systems and in the end trying to understand how the behaviour of the system at a global scale emerges from rules specifying how local modifications operate. To create such a software, one big challenge is if there is a data-structure universal enough to handle efficiently all operations of the system and moreover, a data-structure powerful enough to support different type of networks (e.g., bio, social net, capital markets, relational database design).

To give my answer to this question, I have presented in a few slides our visual graph programming environment called Porgy and our data-structure called “labelled port graph”. However, this graph model has to be used along with a graph hierarchy mechanism to avoid duplicating nodes/edges/attributes like the one implemented in Tulip (Porgy is built upon Tulip). To conclude, I left some open questions about the usage of higher order rules (i.e., a node of the rule replaces a sub-graph) and from a more technical point of view the usage of graph database to improve the rule matching phase.

6.11 The Bio/Life-Sciences need better visualization of statistical network structures

William Ray (Ohio State University – Columbus, US)

License  Creative Commons BY 3.0 Unported license
© William Ray

Many biological systems possess properties such that there are natural elements that are thought of as nodes, with weighted connections between them that can be thought of as edges, but that do not fit into traditional graph-theoretic frameworks, and that therefore are difficult to represent using traditional graph-layout tools.

For example, if one looks at a protein family – a collection of proteins from different organisms that all perform the same function – one can learn quite a lot about why proteins that perform that function work, or don’t work.

However, that inspection requires looking not-only at the choices that evolution has made at each position in the protein, but also how these choices are interrelated.

Unfortunately the intuitive graph-theoretic representation for a protein treats each sequential residue as a node in a graph, and treats dependencies, distances, or other biophysical relationships between residues that can be determined for the family, as weighted edges between nodes. This representation can show, for example, the mutual information between different positions in the family alignment, but can't show amino-acid-specific relationships between position.

As a result, this position-centric view disguises many important dependencies, such as when the large majority of choices are independent, but some small subset have a strong dependence requirement.

Conditional Random Fields provide an interesting formalism for approaching this data. Structurally CRFs (and similar probabilistic networks) describe node-link networks where each node contains a set of categorical sub-nodes, and each edge is composed of conditionally-weighted sub-edges between the sub-nodes. This formalism maps conveniently to these biological networks, and it appears to be a natural mapping to many biological phenomenon with conditionally-interrelated features. As a result, visualizing and interacting with the structure of Conditional Random Fields can provide important insight into fundamental biology.

6.12 Making Sense of Large Scale Image Data

Jens Rittscher (University of Oxford, GB)

License  Creative Commons BY 3.0 Unported license
© Jens Rittscher

Three concrete examples for generating high-dimensional data from image data sets are being presented. The first illustrates visualisation tasks in high-throughput screening. The main challenge here to discover new cellular phenotypes. A human organotypic cell culture system that models epithelial interactions *in vitro* serves as a second example. Finally, I will use digital pathology to demonstrate how various different technology can be harnessed to visualize genomic and molecular information in the tissue architecture context. In summary, the talk will motivate three questions and challenges: (1) How can we integrate dynamic information over time, (2) How can we analyse and visualise expression of molecular markers in the tissue architecture context and (3) The need for developing metrics that capture patient similarity.

6.13 High-Dimensional Medical Data Panel: Exploration and Communication in Biomedical Visualization

Timo Ropinski (Universität Ulm, DE)

License  Creative Commons BY 3.0 Unported license
© Timo Ropinski

When designing visualizations, typically user, data and task need to be considered in order to obtain an effective visualization. Whereas in the area of biomedical visualization at least three different types of users need to be taken into account: medical doctors, medical researchers, and patients. Furthermore, with respect to data, often the high dimensionality in this context is challenging. Unfortunately, for many scenarios in this field, state-of-the-art high-dimensional visualization techniques are not appropriate. When for instance a medical

doctor analyses blood work, often the main task is to compare the data at hand with given reference values. Thus, no embedding of several data sets is required, but rather a comparative visualization of relatively few ones. Similar requirements must be met when a patient reviews his/her tracked health data. On the other hand, when a medical doctor communicates made findings, storytelling techniques seem to be the relevant technique of choice. Accordingly, biomedical visualization researchers need to look into these different requirements, when developing or selecting appropriate visualizations.

6.14 Metronome – Connecting genotypes and phenotypes

Christian Stolte (New York Genome Center, US)

License  Creative Commons BY 3.0 Unported license
© Christian Stolte

Joint work of Christian Stolte, Kevin Shi, Nathaniel Novod, Nina Lapchyk, Fred Criscuolo, Toby Bloom
URL <https://metronome.nygenome.org>

MetroNome is a web-based genotype/phenotype exploration platform with a data visualization interface. It is focused on enabling data sharing, data integration, data exploration, and identification of cohorts via complex combinations of genotypic and phenotypic traits, across diseases.

Metronome is intended to allow researchers to:

- explore data with minimal effort to generate and test hypotheses
- identify cohorts of interest by filtering across multiple types of data, including genotype and clinical data
- use data visualization to find relationships among genomic variants and subject or sample attributes
- share data among groups of collaborators, privately and securely
- combine private data with large public cohorts while retaining full control over that data

MetroNome is intended to hold all types of genomic variants, gene expression data, as well as de-identified subject data from medical records.

6.15 Collaborations between VIS / Bioinformatics / Bio Communities

Marc Streit (Johannes Kepler Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Marc Streit

In the first part of my talk, I summarize what advice researchers and practitioners can get from a theory of visualization. We – as a community – currently provide advice by publishing models and theories, by collecting techniques and methods, and by describing best practices. While this is very useful, it is often not actionable. A less explored possibility is to provide cheat sheets in the form of decision trees that can help practitioners to create effective visualizations. These decision trees could be created as a community effort, underpinned with our models, and carefully annotated. In the second part, I talk about why generalizing design studies is hard, why data and task abstraction is key to create impact in visualization through collaboration with domain experts, and what lessons I've learned from previous collaborations.

6.16 Visualization for Pan-genomes and Meta-genomes

Granger Sutton (J. Craig Venter Institute – Rockville, US)

License  Creative Commons BY 3.0 Unported license
© Granger Sutton

Pan-genomes share many characteristics with meta-genomes and can use the same visualization approaches in part but also have distinctive needs. At the highest level a pan-genome is a universe of genes distributed across a set of genomes where each genome contains a set of genes which is a subset of the universe. This is also true for meta-genomes but with genomes being replaced by samples or environments. The top level data representation is then a two dimensional matrix of genes across genomes or samples. A typical visualization is a heat map which has been bi-clustered to provide cladograms in both dimensions. Both can also be represented by metabolic networks of what functional capabilities are contained. In many cases a meta-genomic sample will in fact contain one or more pan-genomes. Meta-genomes tend to have much less complete or fractured genome representations than pan-genomes. Deconvoluting assembled contigs into species specific bins is unique to meta-genomes and often read based approaches rather than assembled contigs are used for meta-genomes.

6.17 Democratization of Data Science

Blaz Zupan (University of Ljubljana, SI)

License  Creative Commons BY 3.0 Unported license
© Blaz Zupan

I will talk about how visual programming, interactive visualization, and explorative data analysis can help us in making visual analytics and machine learning accessible to everyone. I will demo these concepts in the case of single cell data analytics in Orange (visit <http://orange.biolab.si> or <http://youtube.com/orangedatamining> for short videos).

7 Panel discussions

7.1 Collaboration Panel: From Genomics/Bioinformatics to Visualization – in 5 minutes

Jan Aerts (KU Leuven, BE)

License  Creative Commons BY 3.0 Unported license
© Jan Aerts

In this short talk in preparation on a panel discussion regarding collaboration between visualization experts and biology/bioinformatics researchers, I start with a quick overview of my own journey from genomics to visualization research. In addition, and more importantly, I indicate some challenges that we encounter in bridging the gap between these two domains. These include reusability of solutions, composability of bioinformatics tools versus often monolithic approach for visualization tools, and the (incorrectly) perceived nice-to-have view on visualization in omics projects.

7.2 Collaboration Panel: Mutual Respect

Sheelagh Carpendale (University of Calgary, CA)

License  Creative Commons BY 3.0 Unported license
© Sheelagh Carpendale

This talk is about mutual respect – just one of the many important factors in a good collaboration. One aspect of mutual respect is developing an understanding of how the different research communities think about the way they would like to make contributions to their discipline. While biologists' goals usually center around developing a better understanding of their data, ideally leading to new biological insights, visualization researchers' goals usually center around contributing to advancing visualization through creating new visual representations, new layout approaches and/or new exploration techniques. There can be a tendency to favour the more easily understand the global importance of new biological insights. However, it is important in a collaboration that one disciplines goals not over shadow the other. We should remember that there are visual representations that have fundamentally empowered society, for example, the alphabet is a 'visualization' of spoken language. It may be difficult to learn but is a very powerful visual representation.

7.3 Biological Networks Panel: Matching the User's Mental Map

Carsten Görg (University of Colorado – Aurora, US)

License  Creative Commons BY 3.0 Unported license
© Carsten Görg

Biologists tend to think about relationships between biological entities they study in a specific way. Often, they have a detailed mental map or even use an actual sketch or drawing that represents relationships between biological entities or biological processes. Computationally generated representations of networks typically don't match the biologists' mental or drawn representation. We propose a network layout approach that arranges the nodes not only based on the network topology but takes the underlying biological semantics into account to create a high-level blueprint of the network. Biologists can interactively rearrange the elements in the blueprint so that the representation matches their mental model. The detailed layout is then generated based on the constraints and structure defined in the blueprint.

7.4 Biological Networks Panel: Visualising Biological Networks: comparison of trees to graphs

Jessie Kennedy (Edinburgh Napier University, GB)

License  Creative Commons BY 3.0 Unported license
© Jessie Kennedy

Comparison in and between biological networks is a common problem in biological visualisation. In biological taxonomy visualisation the underlying data structure is a graph consisting of many overlapping trees, where one of the user tasks is to compare their taxonomy with pre-existing taxonomies. In 2000, we developed two visualisations to support comparison of multiple taxonomies, one was a force directed graph layout, where the user could add

as many of the taxonomies as desired. Taxonomies in the graph are identified by having different coloured edges and nodes contain coloured marks to show which taxonomies they belong to. The graph layout suffered from hairball issues therefor we introduced search and filter mechanisms to assist in understanding for the user. However, the users still found the graph layout difficult to comprehend due to the inability to easily identify and separate individual taxonomies and the difficulty in seeing the top down layout of the taxonomies. We also developed a small multiples visualisation with icicle plots representing each taxonomy with linking and focus & context techniques for exploring and comparing the taxonomies. The tool supported removal of ranks in the taxonomies to ease comparison of trees by forcing similar tree structures across the taxonomies. This approach was much preferred by the taxonomists. We then developed a combined approach based on a directed acyclic graph, which maintained the top down layout of the taxonomies, where the user could highlight one taxonomy in the context of the other taxonomies to easily show the differences.

More recently we have been addressing the problem of comparison of multiple networks faced by computational biologists trying to determine best network models for a range of biological networks such as gene interaction networks to ecological networks. In determining the biological network the computational biologists make use of Bayesian network inference algorithms to generate 100s-100s of candidates networks which are given a score based on their fit to the underlying model. The biologists need to examine and compare these hundreds of networks to understand the scores assigned to the networks, e.g. to determine if networks with similar scores are similar or different. If similar then it is likely that the highest scoring network will be the best, however if different then the user might want to generate a consensus network from a range of networks selected by the user. This problem concerns comparison of many small to medium networks rather than one or a few large networks. We have created Bayespiles which is adapted from small multipiles, a matrix based visualisation of many networks which piles and summarises networks. Bayespiles enables the exploration, organisation and comparison of hundreds of scored directed networks from multiple heuristic search runs. It features two matrix-based representation modes for directed networks (top-down and diamond), a normalised histogram that shows the distribution of scores in the solution space, flexible network ordering based on run ID, iteration or score, node reordering, interactive comparison of networks across groups, support for the manual construction of a consensus network, interactive graph filtering mechanisms and a summary view of all outgoing edges for selected nodes.

There remain many challenges in comparing networks including scalability to many large networks and comparing multilayer and multidimensional networks.

7.5 Collaboration Panel: Visualization and Visual Analysis of Biomolecular Structures

Barbora Kozlíková (Masaryk University – Brno, CZ)

License  Creative Commons BY 3.0 Unported license
© Barbora Kozlíková

In my five-minute talk I am introducing my experience in collaboration with protein engineering research group. With its members, we are focusing on analysis and visualization of protein structures, namely searching for tunnels connecting the protein outer environment with its active site. Such tunnels can be subsequently used for transportation of ligands to

the active site. In my talk I am stressing the importance of finding the common language with the domain experts and developing mutual trust. I also discuss different publication possibilities and options, as well as issues related to transferring the research results into practice, which requires more engineering and management skills than the research ones.

7.6 Curriculum Panel: Designing a curriculum for teaching visualization in bioinformatics

Michael Krone (Universität Stuttgart, DE)

License  Creative Commons BY 3.0 Unported license
© Michael Krone

My talk focusses on the requirements and contents of a curriculum (or course) for teaching visualization in bioinformatics. I will present my personal recommendation for core visualization principles that should be taught in the context of biological visualization. Key aspects that have to be considered are the background of the intended audience (computer science, bioinformatics, biology) and the level of their studies (bachelor, master, PhD). Students also need to learn certain technical skills in order to be able to create their own visualizations (e.g., programming). Libraries like D3 for non-spatial data or Three.js for spatial data can be powerful tools that lessen the programming burden. This leads to the question of how to teach students to use these tools efficiently in a reasonably short time. In summary, I think that teaching basic visualization concepts is more important than teaching using tools.

7.7 Curriculum Panel: Vis is a large number of small problems

Martin Krzywinski (BC Cancer Research Centre – Vancouver, CA)

License  Creative Commons BY 3.0 Unported license
© Martin Krzywinski

An effective way to teach visualization is to break a visualization task or challenge into a large number of small problems. Many of these small problems recur and for each there is a relatively small number of ways in which well-meaning users get it wrong. This talk demonstrates redesign examples of typical visualizations from biology and demonstrates the similarity across users' missteps in the context of this kind of divide-and-conquer strategy.

7.8 Biological Networks Panel: Scaffolding Bionetwork Visualization with models and theories

Georgeta Elisabeta Marai (University of Illinois – Chicago, US)

License  Creative Commons BY 3.0 Unported license
© Georgeta Elisabeta Marai

Four years ago, I joined a fearless research group at the Electronic Visualization Laboratory, who have the wisdom to question all existing paradigms. In that spirit, I question three definitions and paradigms for biological networks. First, a biological network is any network that applies to biological systems – for example, a functional network in the mouse audio

cortex is still a biological network. Second, some biological networks have spatial components – even when those components are not anchored in the physical (e.g., image) space, they bear meaning to the biologist. Third, in terms of principles that should guide the selection of a visualization technique for biological networks, “overview-first” is not the only possible design approach. There is also “Search-first” (van Ham and Perer 2009), “Details-first”, and so on.

7.9 Pan-Genomics Panel: Some questions and challenges about comparative genomics and pan-genomics

Kay Katja Nieselt (Universität Tübingen, DE)

License  Creative Commons BY 3.0 Unported license
© Kay Katja Nieselt

In my short panel talk I briefly outline some of the questions and challenges in pan-genomics. I start first with a generalized definition of a pangenome. Based on that I point out that pangenomics has and will influence a number of both traditional viewpoints in biology in the future, such as the definition of a species. One main point is the data structure that represents a pangenome, which depends on its definition as well as context that it is studied in. Depending on the data structure, different visualisations are needed that biologists would want to see when studying pangomes. There are many algorithmic as well as visualisation challenges in this field, such as scalability and update, which hopefully will be addressed during this seminar.

7.10 High-Dimensional Medical Data Panel: High-Dimensional Medical Data

Jos B.T.M. Roerdink (University of Groningen, NL)

License  Creative Commons BY 3.0 Unported license
© Jos B.T.M. Roerdink

In my presentation I will focus on a number of aspects when dealing with High-Dimensional Medical Data, such as:

How to collaborate? Important issues are spending time with collaborators, using simple visualizations with explanations, avoiding sophistication and information overload, and the need to build trust.

How to integrate into existing, complex data ecosystems? This is generally not possible in medicine, because of certification issues. But integrating tools in systems of medical researchers is. It is important to find a liaison person.

Who are the key influencers in this field? This concerns first of all people with a genuine interest who want to invest time and energy, funding agencies, societal drives.

Cross-cutting interests shared by other communities. High on my list are comparison of visualizations, workflows, and provenance.

7.11 Biological Networks Panel: Visualisation of Biological Networks: Past, Present, and Future

Falk Schreiber (*Universität Konstanz, DE*)

License  Creative Commons BY 3.0 Unported license
© Falk Schreiber

Networks play an important role in the life sciences. Networks can represent data and processes from chemistry (e.g. chemical structure graphs) to molecular biology (e.g. metabolic networks, co-expression networks) to ecology (e.g. food webs, animal behaviour networks) to medicine (e.g. infection networks) to other related areas. In the first part, this talk will present the history and state-of-the-art of network visualisation (layout) with a focus on metabolic networks. Here we will discuss benefits and disadvantages of common layout algorithms often used to visualise biological networks and look at some specific layout methods. The second part of the talk will investigate visualisation-related topics such as graphical standards for biological networks (e.g. SBGN) and visual analytics for biological network exploration and investigation. This presentation will finish with an outlook to future developments such as immersive analytics for biological networks.

References

- 1 O. Kohlbacher, F. Schreiber and M. O. Ward: ‘Multivariate networks in the life sciences’ in A. Kerren, H. C. Purchase and M. O. Ward (eds.), *Multivariate network visualization*, Springer LNCS 8380, 61–73, 2014.
- 2 C. Bachmaier, U. Brandes and F. Schreiber: ‘Biological networks’ in R. Tamassia (ed.), *Handbook of graph drawing and visualization*, Chapman & Hall/CRC Press, 621–652, 2013.
- 3 N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, K. Hideya, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn and H. Kitano: ‘The Systems Biology Graphical Notation’ *Nature Biotechnology* 27: 735–741, 2009.

7.12 Pan-Genomics Panel: Scaling Sequence Comparison for Pan & Metagenomics

Danielle Szafir (*University of Colorado – Boulder, US*)

License  Creative Commons BY 3.0 Unported license
© Danielle Szafir

Pangenome and metagenome comparisons require biologists to identify and interpret meaningful similarities and differences between organisms. This comparison problem requires analysis tools to support comparisons as the number and complexity of sequences increase and also introduce new questions unsupported by different tools. Existing sequence comparison tools enable scalability along at most two of these different dimensions. By understanding the needs and computational and visual challenges associated with comparison tasks in pan- and metagenomes, we can begin to create visualizations that support the needs of these analyses along all three dimensions.

7.13 High-Dimensional Medical Data Panel: Living with Algorithms

Cagatay Turkay (City – University of London, GB)

License  Creative Commons BY 3.0 Unported license
© Cagatay Turkay

The analysis and modelling of high-dimensional medical data is relevant for a wide spectrum of users (researchers/clinicians/patients) in different capacities and complexities. Wherever a user stands on this spectrum, due to the complexities that high dimensional data introduces (heterogeneity / sparsity / uncertainty), interacting with algorithms is unavoidable, be it in terms of getting a recommendation or in terms of building explanatory models. The pursuit for interpretable, comprehensible and explainable algorithms is getting interest in several domains currently including machine learning, knowledge discovery and data visualisation focusing on several different application domains. Visualisation has already shown great potential as an expressive and insightful medium with integrated and linked representations of several components of algorithms and engaging different users in various capacities. This talk investigates the different users in the context of high-dimensional medical data, touches upon a number of visual analytics techniques, and discusses a number of challenges that cuts across these different use cases at the intersection of algorithms and users.

References

- 1 Endert, A., W. Ribarsky, C. Turkay, B. L. Wong, Ian Nabney, I. Díaz Blanco, and F. Rossi. “The state of the art in integrating machine learning into visual analytics.” In Computer Graphics Forum, vol. 36, no. 8, pp. 458 – 486. 2017.
- 2 Hohman, F. M., Kahng, M., Pienta, R., and Chau, D. H. (2018). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. IEEE Transactions on Visualization and Computer Graphics.
- 3 Krause, J., Perer, A., and Ng, K. (2016, May). Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5686 – 5697). ACM.

8 Acknowledgements

We would like to thank all participants of the seminar for their contributions and lively discussions; we also would like to thank the scientific directorate of Dagstuhl for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the untiring help of the (scientific) staff of Dagstuhl, including Ms. Annette Beyer, Ms. Heike Clemens and Mr. Michael Gerke.

Participants

- Jan Aerts
KU Leuven, BE
- Katja Bühler
VRVis – Wien, AT
- Sheelagh Carpendale
University of Calgary, CA
- James L. Chen
Ohio State University – Columbus, US
- Arlene Chung
University of North Carolina – Chapel Hill, US
- Anamaria Crisan
University of British Columbia – Vancouver, CA
- Mirjam Figaschewski
Universität Tübingen, DE
- Angus Forbes
University of California at Santa Cruz, US
- Nils Gehlenborg
Harvard University, US
- Carsten Görg
University of Colorado – Aurora, US
- David H. Gotz
University of North Carolina – Chapel Hill, US
- Helena Jambor
TU Dresden, DE
- Jessie Kennedy
Edinburgh Napier University, GB
- Karsten Klein
Universität Konstanz, DE
- Anne Knudsen
University of Calgary, CA
- Barbora Kozlíková
Masaryk University – Brno, CZ
- Michael Krone
Universität Stuttgart, DE
- Martin Krzywinski
BC Cancer Research Centre – Vancouver, CA
- Alexander Lex
University of Utah – Salt Lake City, US
- Raghu Machiraju
The Ohio State University – Columbus, US
- Georgeta Elisabeta Marai
University of Illinois – Chicago, US
- Lennart Martens
Ghent University, BE
- Ewy Mathé
Ohio State University – Columbus, US
- Torsten Möller
Universität Wien, AT
- Scooter Morris
University of California – San Francisco, US
- Cydney Nielsen
BC Cancer Agency – Vancouver, CA
- Kay Katja Nieselt
Universität Tübingen, DE
- Bruno Pinaud
University of Bordeaux, FR
- James Procter
University of Dundee, GB
- William Ray
Ohio State University – Columbus, US
- Jens Rittscher
University of Oxford, GB
- Jos B.T.M. Roerdink
University of Groningen, NL
- Timo Ropinski
Universität Ulm, DE
- Ryo Sakai
PharmiWeb Solutions – Bracknell, GB
- Falk Schreiber
Universität Konstanz, DE
- Christian Stolte
New York Genome Center, US
- Marc Streit
Johannes Kepler Universität Linz, AT
- Granger Sutton
J. Craig Venter Institute – Rockville, US
- Danielle Szafrir
University of Colorado – Boulder, US
- Cagatay Turkay
City – University of London, GB
- Michel A. Westenberg
TU Eindhoven, NL
- Blaz Zupan
University of Ljubljana, SI

