

DeCART Summer School 2018

Data Visualization

Day 2 - Morning

17 July 2018

Nils Gehlenborg, PhD - Harvard Medical School

Syllabus

- **Day 1**

- Morning: Introduction to Data Visualization
- Afternoon: Design Process, Evaluation, and Interaction

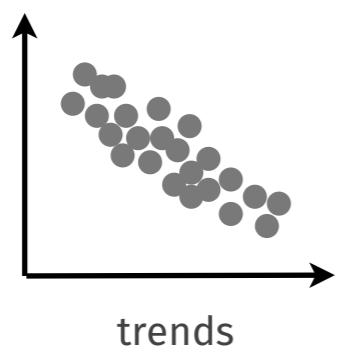
- **Day 2**

- **Morning: Visualization of high-dimensional, network, and temporal data**
- Afternoon: Genomic data visualization techniques
Introduction to Altair and advanced Altair

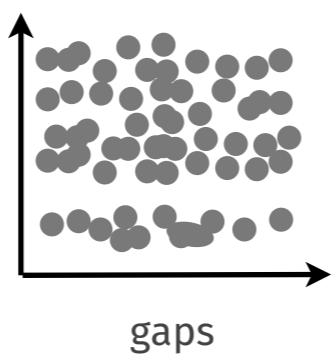
Syllabus - Day 2 (Morning)

- 9:00 - 9:15 | Review
- 9:15 - 10:00 | High-Dimensional Data
- 10:00 - 10:15 | Break
- 10:15 - 10:45 | Network Data
- 10:45 - 11:30 | Temporal Data

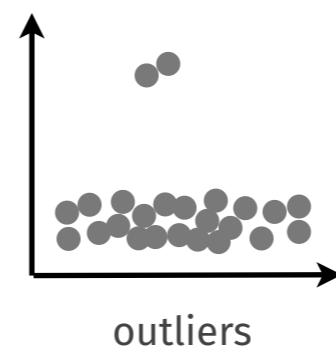
Review of Day 1



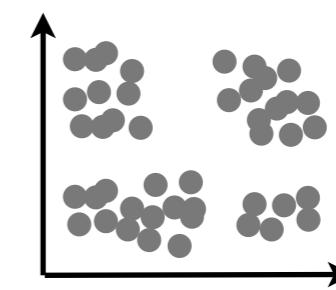
trends



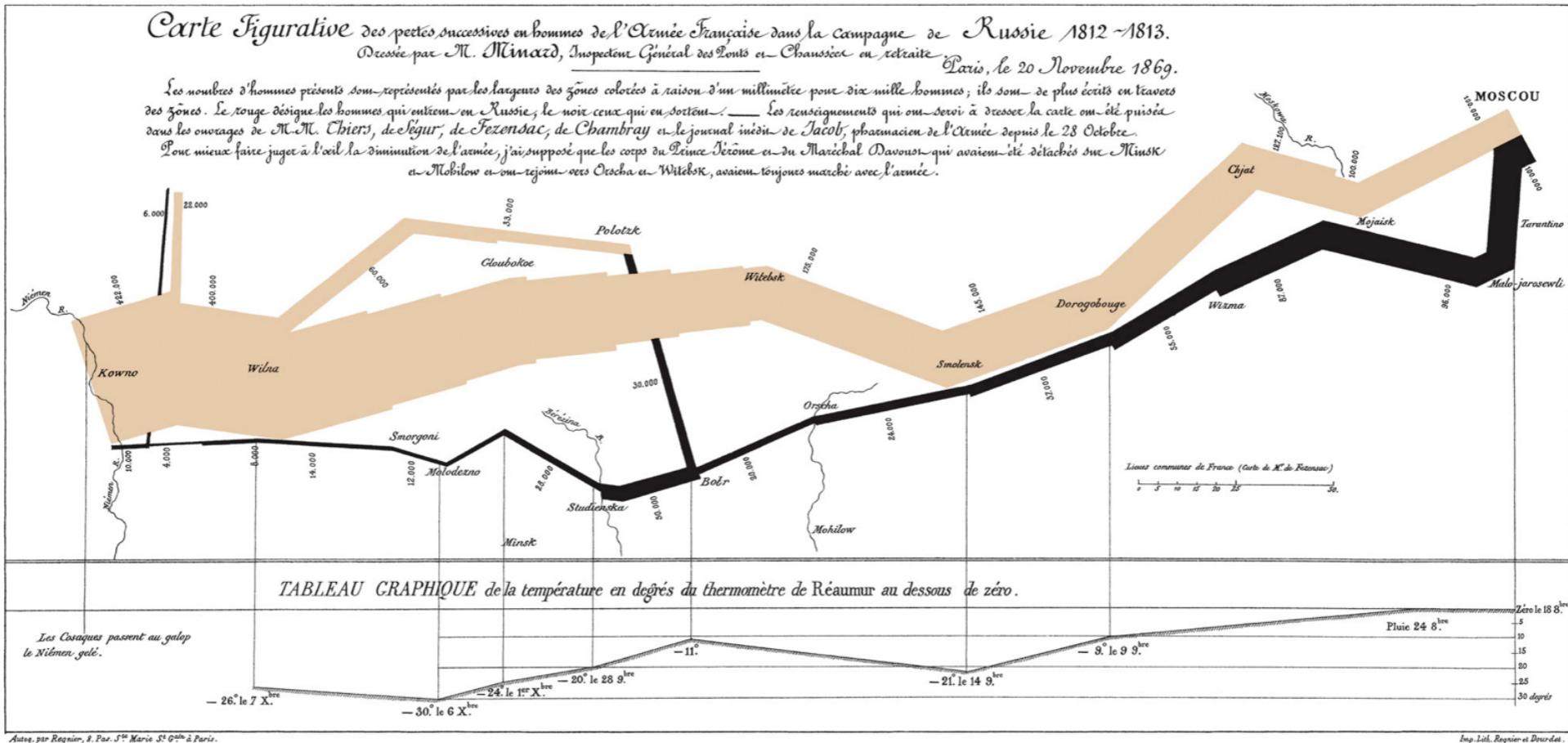
gaps



outliers



clusters



Minard 1869

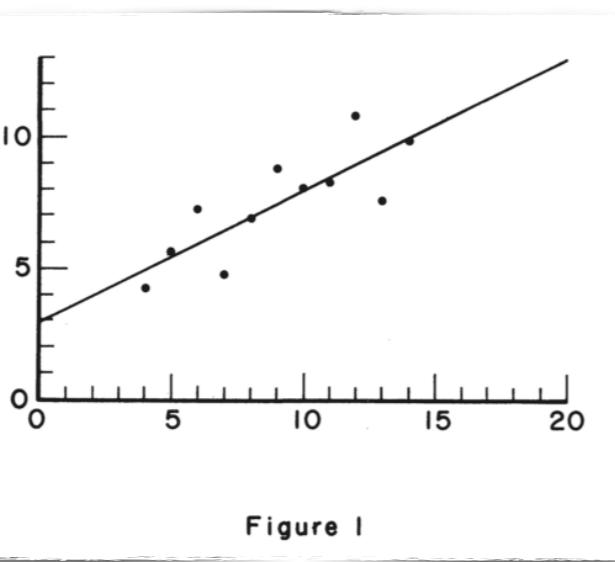


Figure 1

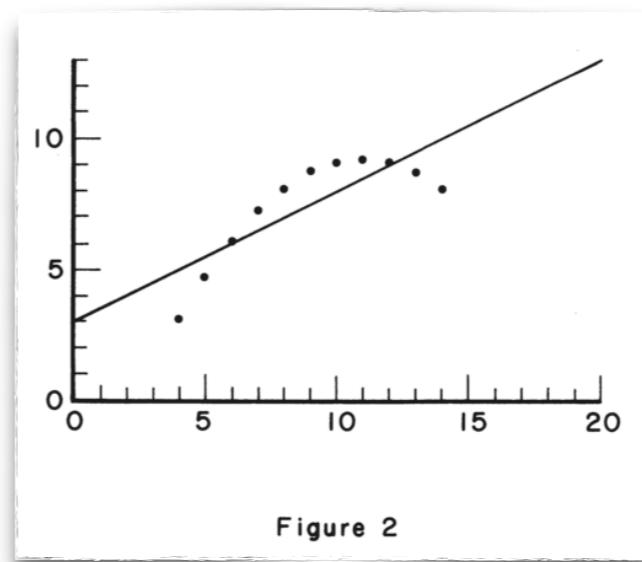


Figure 2

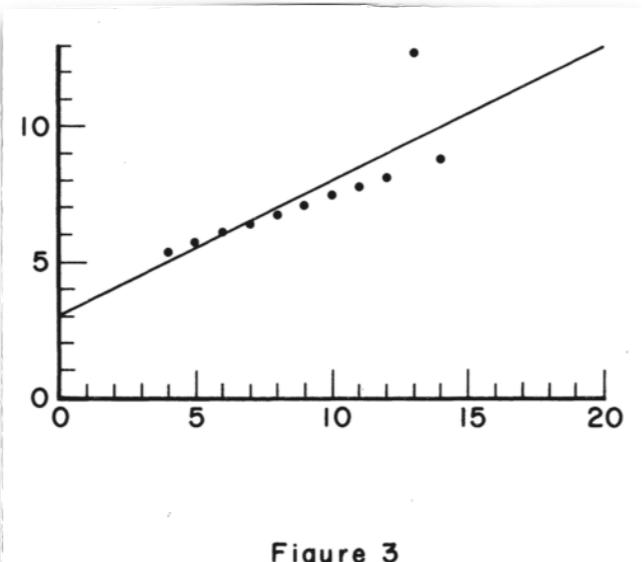


Figure 3

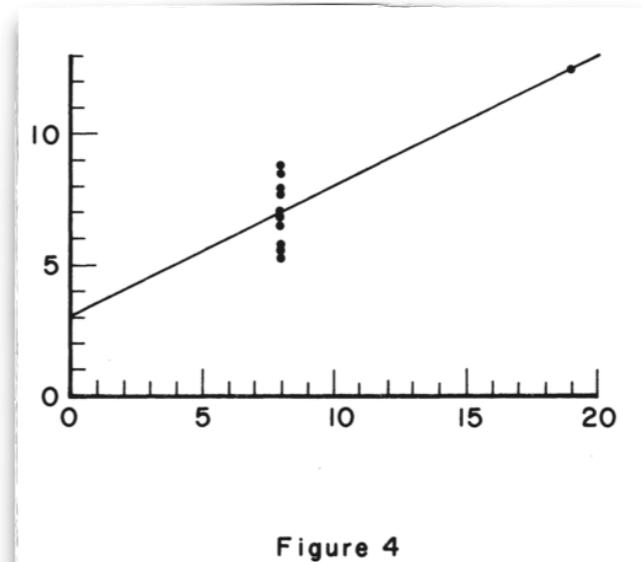


Figure 4

Anscombe 1973, The American Statistician

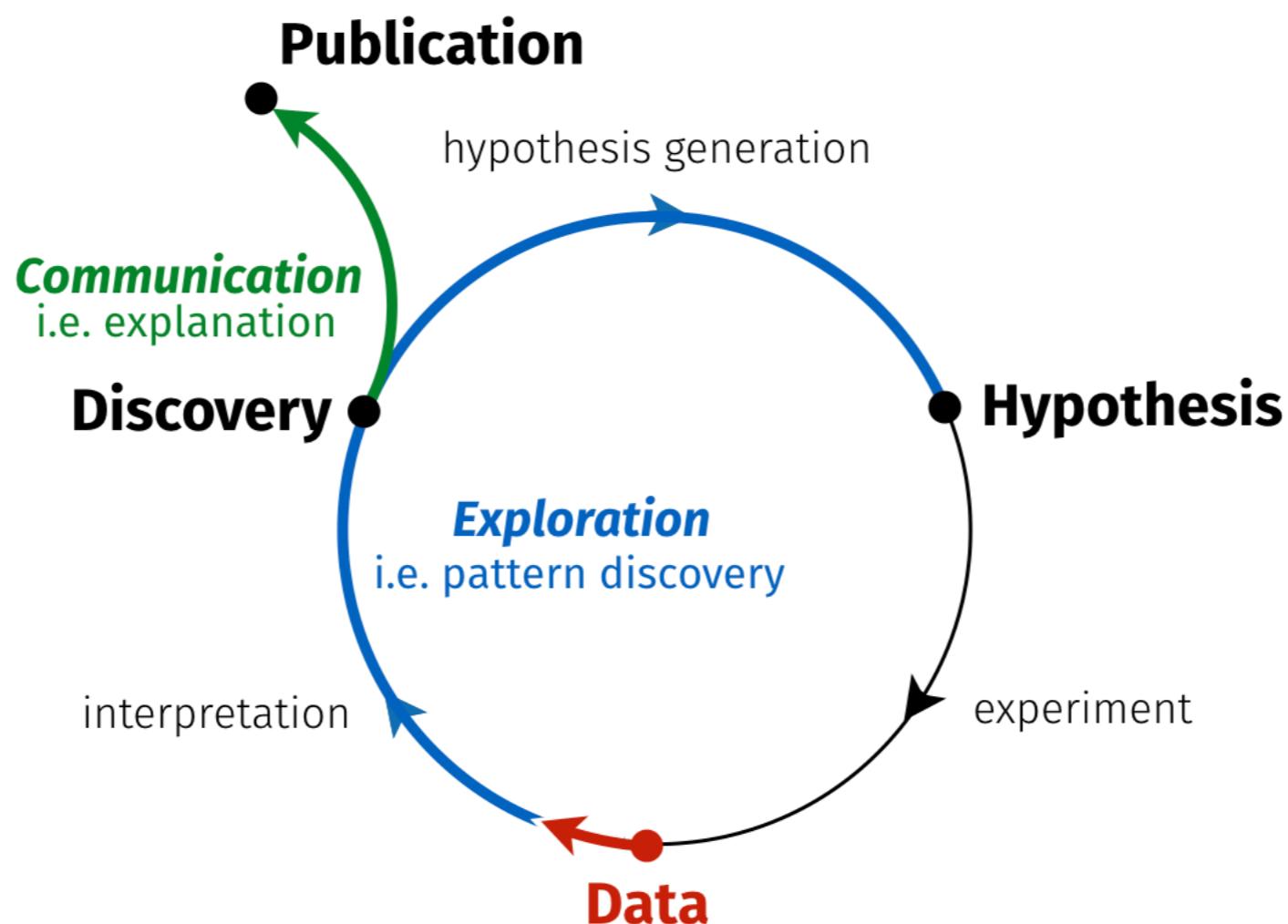
Visualization Use Cases

Exploration

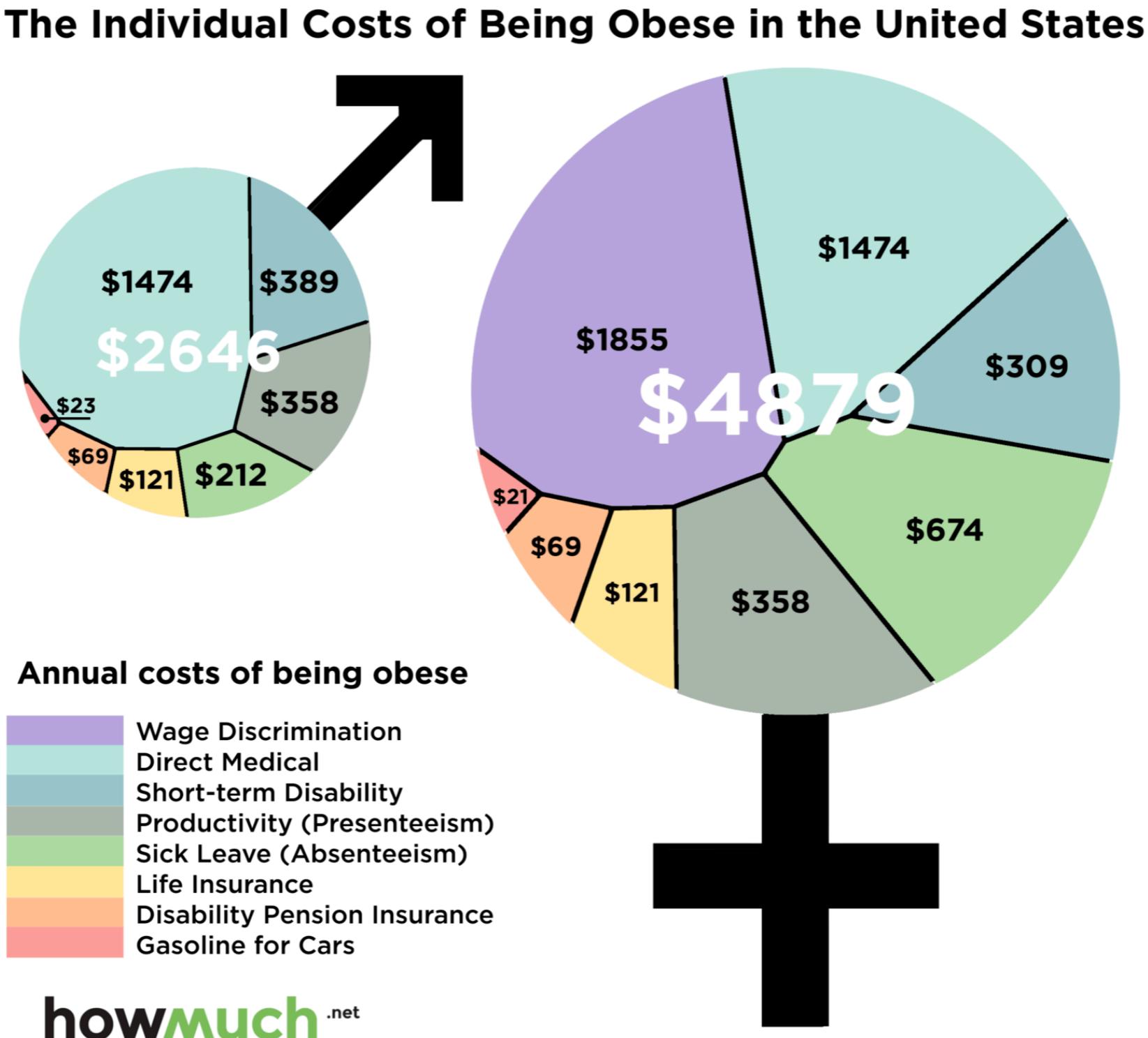
Confirmation

Communication

Visualization Use Cases



Redesign Exercise



Redesign Exercise

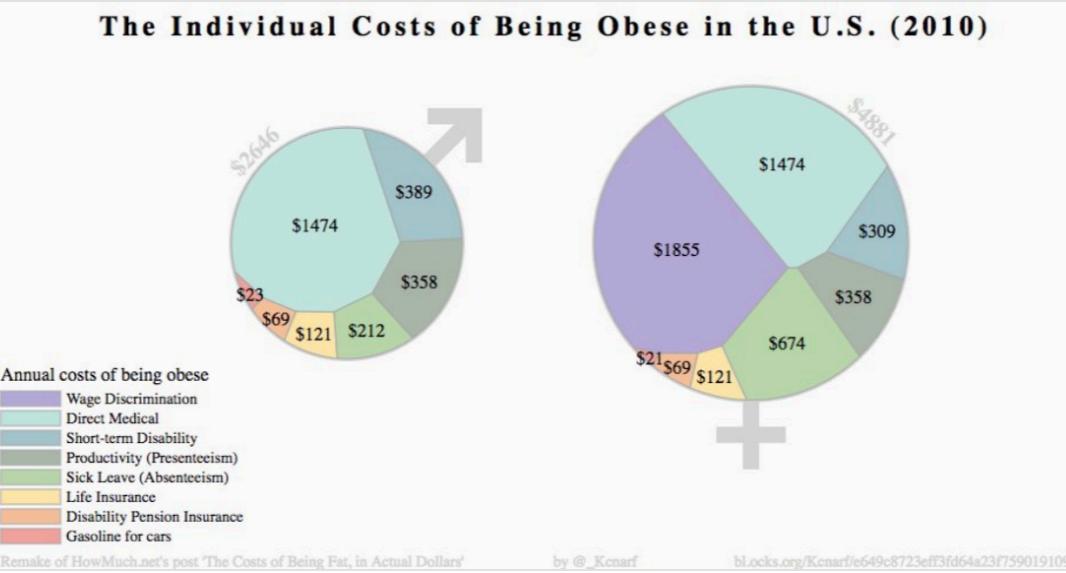
 **Franck Lebeau**
 @_Kcnarf

[Follow](#) ▾

🎉 Updated API for d3-voronoi-map (github.com/Kcnarf/d3-voro...), which now allows to control the final tessellation.

Here (bl.ocks.org/Kcnarf/e649c87..., #D3js) remake of a [@howmuch_net](#)'s viz (howmuch.net/articles/obesi...), placing same cell types at the same positions eases comparison.

The Individual Costs of Being Obese in the U.S. (2010)



Category	Cost (\$)
Wage Discrimination	\$1855
Direct Medical	\$1474
Short-term Disability	\$389
Productivity (Presenteeism)	\$358
Sick Leave (Absenteeism)	\$212
Life Insurance	\$23
Disability Pension Insurance	\$674
Gasoline for cars	\$309
Total	\$4881

Annual costs of being obese

- Wage Discrimination
- Direct Medical
- Short-term Disability
- Productivity (Presenteeism)
- Sick Leave (Absenteeism)
- Life Insurance
- Disability Pension Insurance
- Gasoline for cars

Remake of HowMuch.net's post 'The Costs of Being Fat, in Actual Dollars'.

by @_Kcnarf bl.ocks.org/Kcnarf/e649c8723eff3fd64a23f75901910930

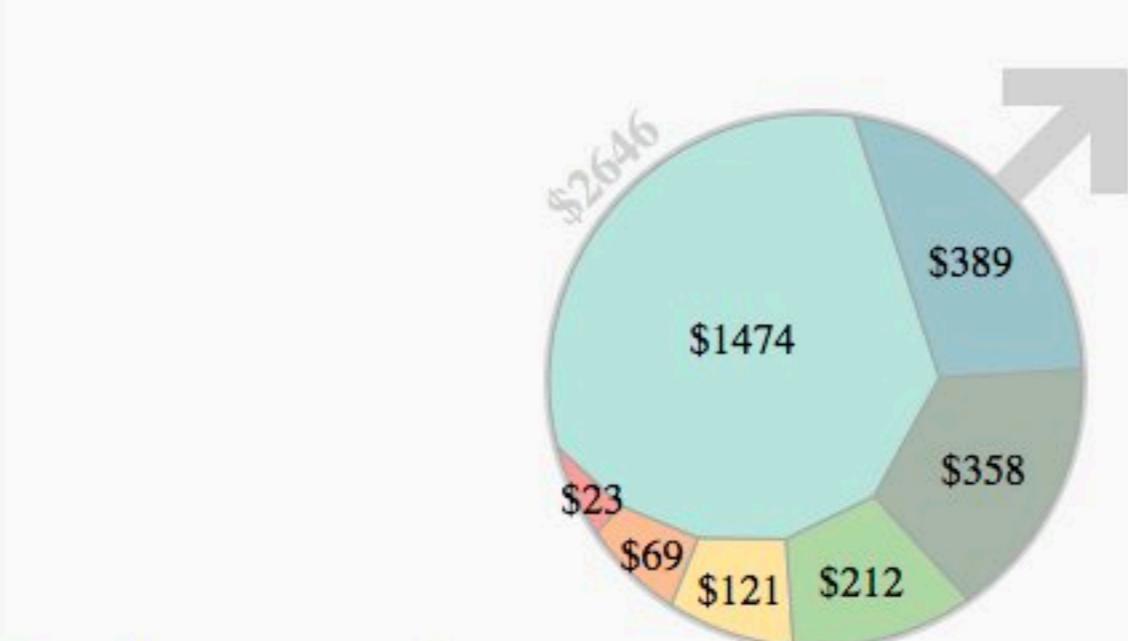
10:44 AM - 25 Apr 2018

5 Retweets 11 Likes

1 5 11

Redesign Exercise

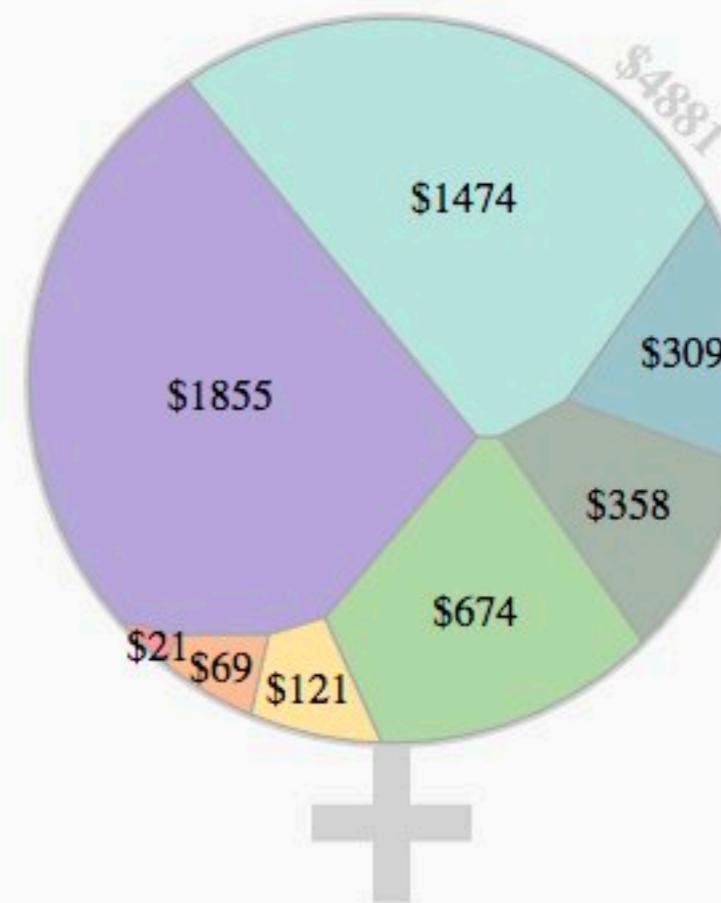
The Individual Costs of Being Obese in the U.S. (2010)



Annual costs of being obese

- Wage Discrimination
- Direct Medical
- Short-term Disability
- Productivity (Presenteeism)
- Sick Leave (Absenteeism)
- Life Insurance
- Disability Pension Insurance
- Gasoline for cars

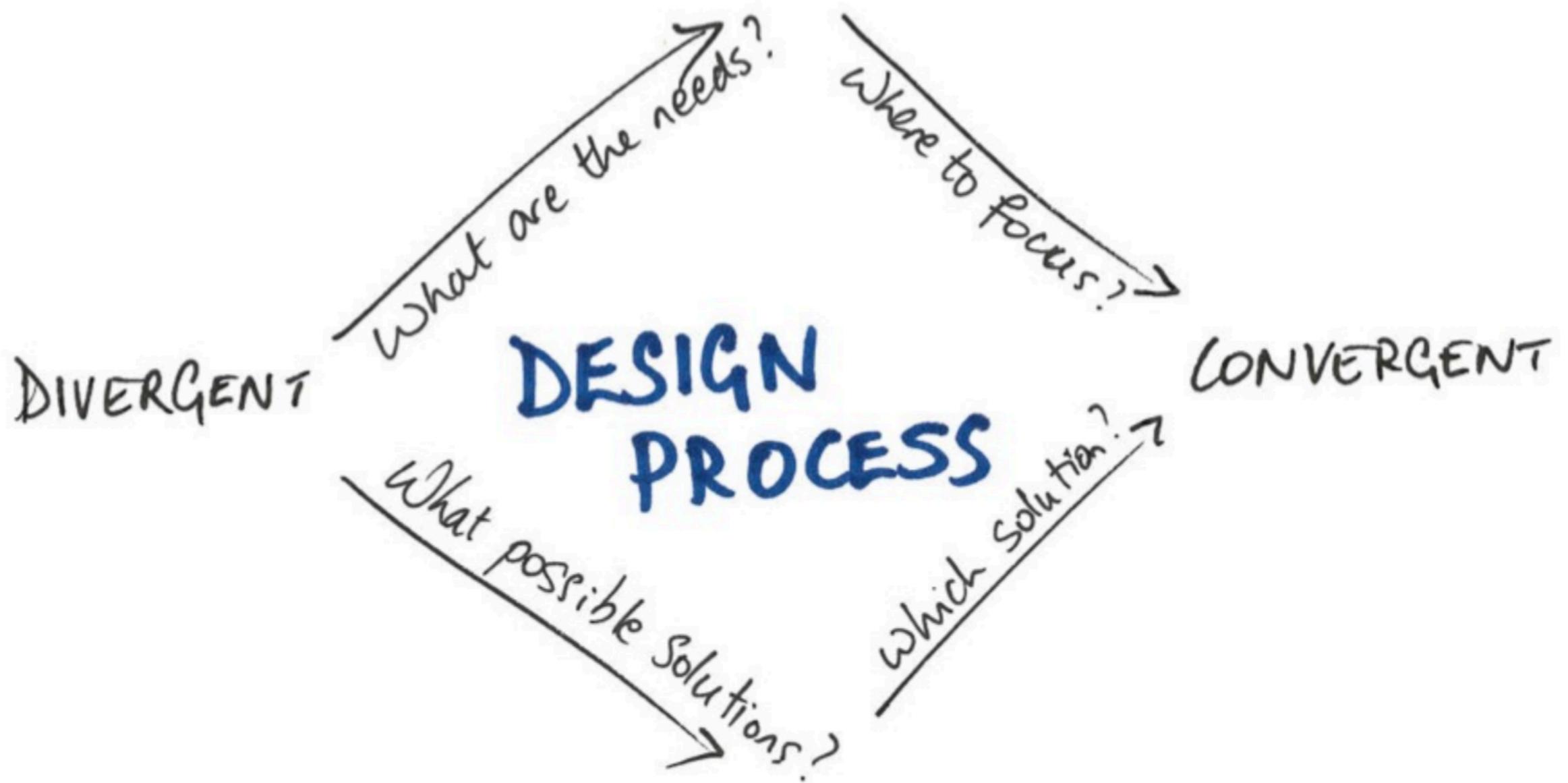
Remake of HowMuch.net's post 'The Costs of Being Fat, in Actual Dollars'



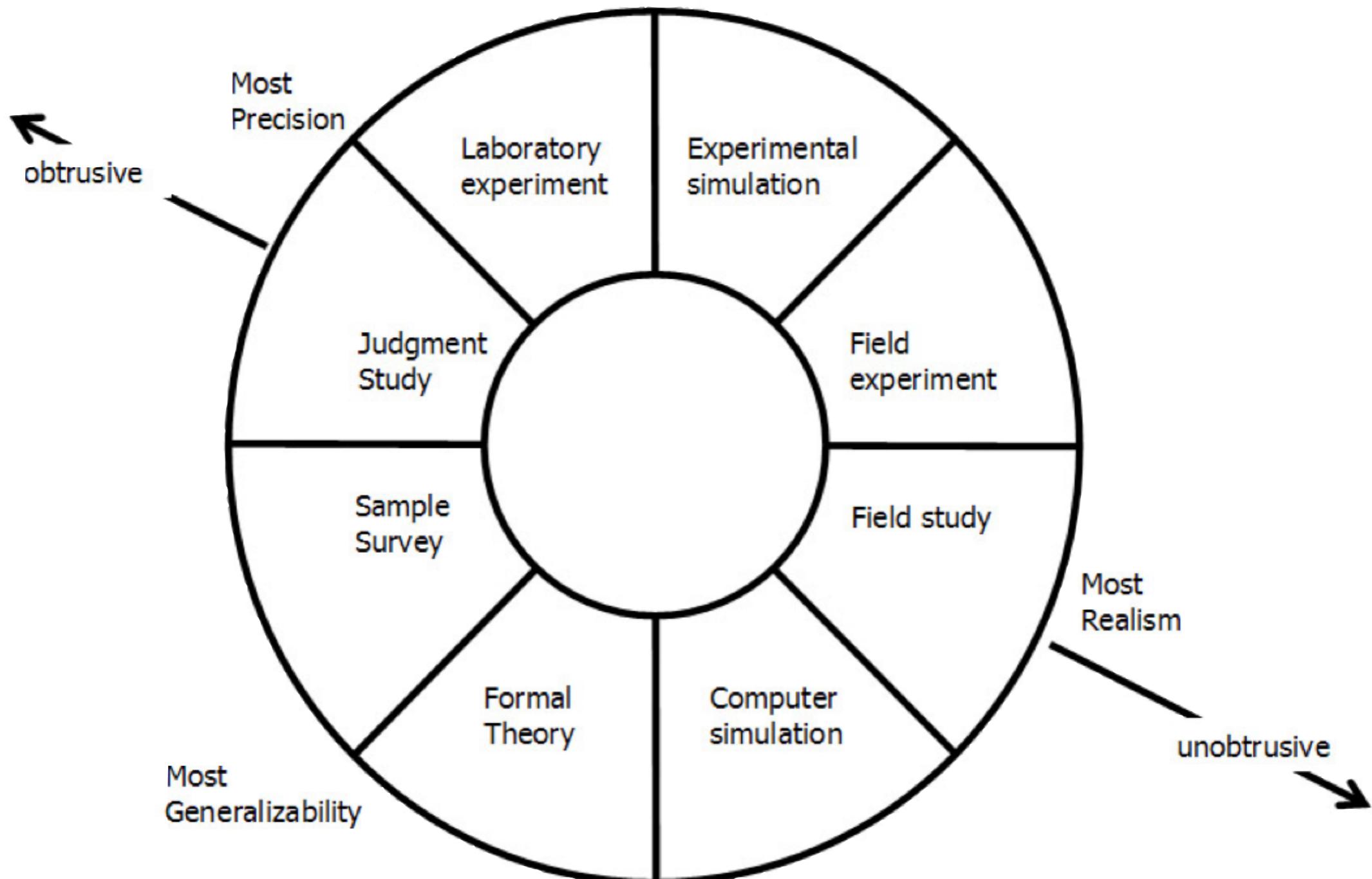
by @_Kcnarf

bl.ocks.org/Kcnarf/e649c8723eff3fd64a23f75901910930

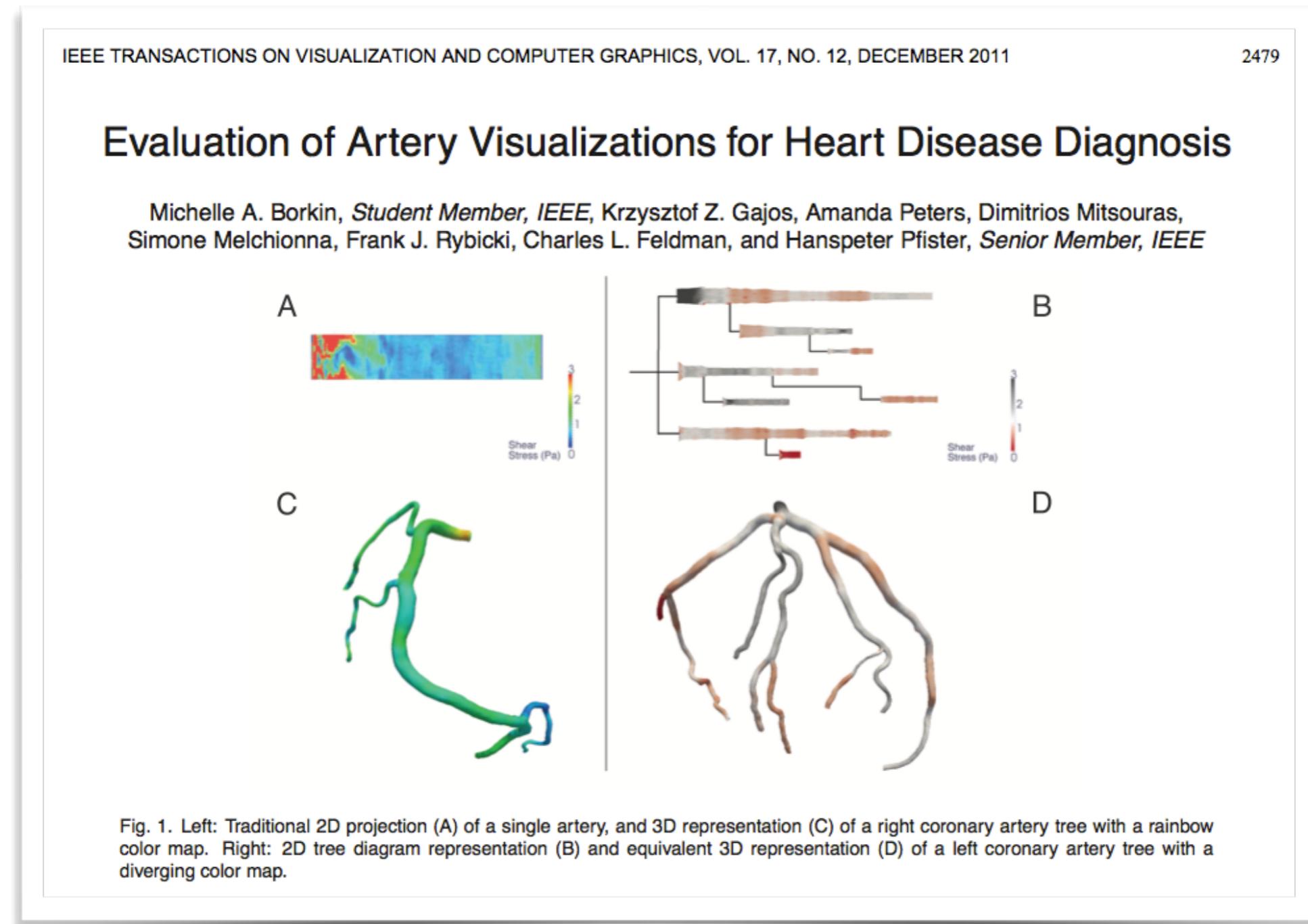
Design Process



Validation Techniques



Why we should ❤️ good visualizations



Single View Interactions

Manipulate

⌚ Change over Time



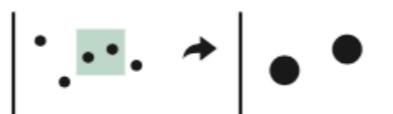
🔍 Select



ניווט (Navigate)

→ Item Reduction

→ Zoom
Geometric or *Semantic*



→ Pan/Translate



→ Constrained



→ Attribute Reduction

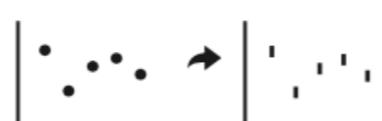
→ Slice



→ Cut



→ Project



Multi-View Interactions

Facet

④ Juxtapose and Coordinate Multiple Side-by-Side Views

→ Share Encoding: Same/Different

→ *Linked Highlighting*



→ Share Data: All/Subset/None



→ Share Navigation



		Data		
		All	Subset	None
Encoding	Same	Redundant	Overview/Detail	Small Multiples
	Different	Multiform	Multiform, Overview/Detail	No Linkage

⑤ Partition into Side-by-Side Views



⑥ Superimpose Layers



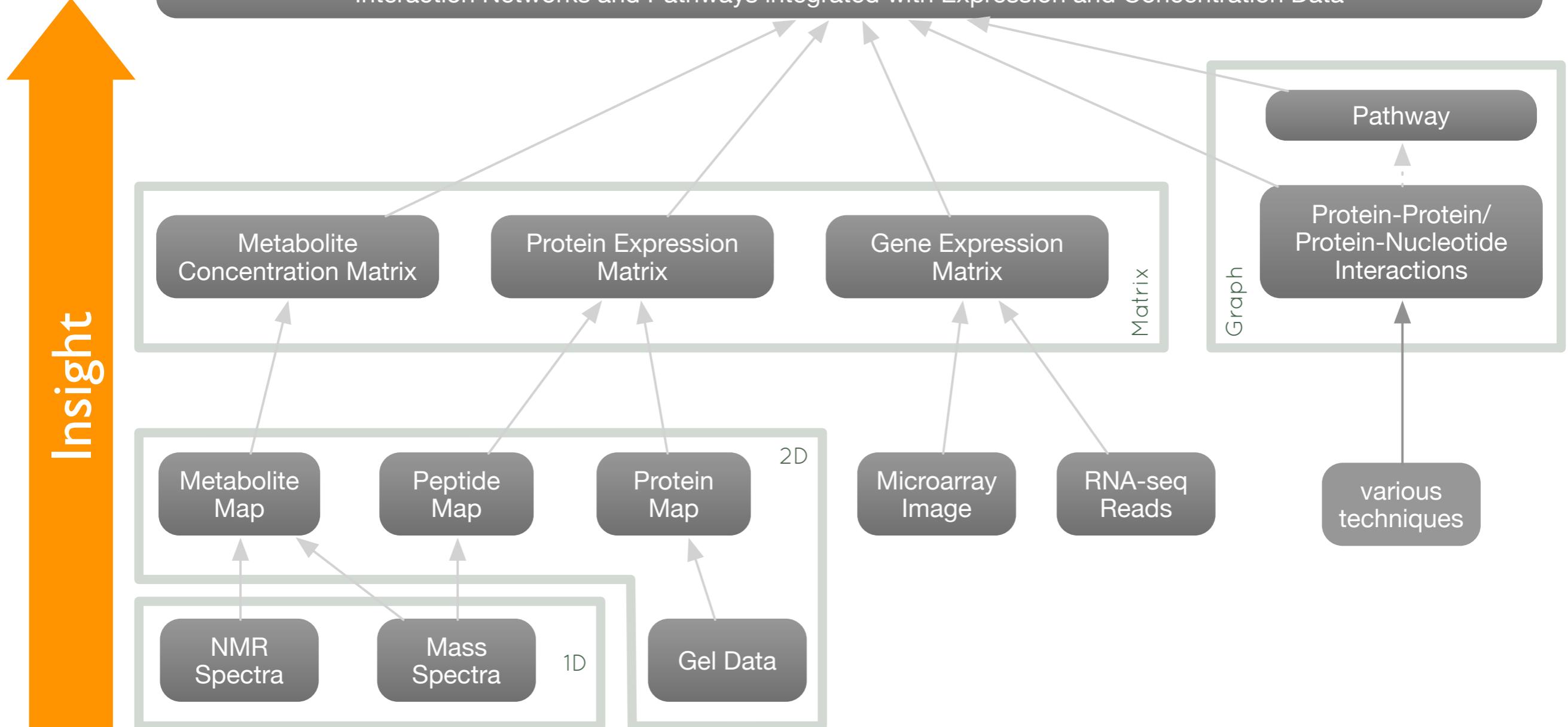
High-Dimensional Multivariate Data Homogeneous Tables

Multivariate Data

- typical “omics” data: transcriptomics, proteomics, metabolomics
- expression/concentration levels of many biological entities (transcripts, proteins, etc.) across many different conditions/time points
- entity levels measured per sample on a “genome-wide” scale
- often entities are not measured directly



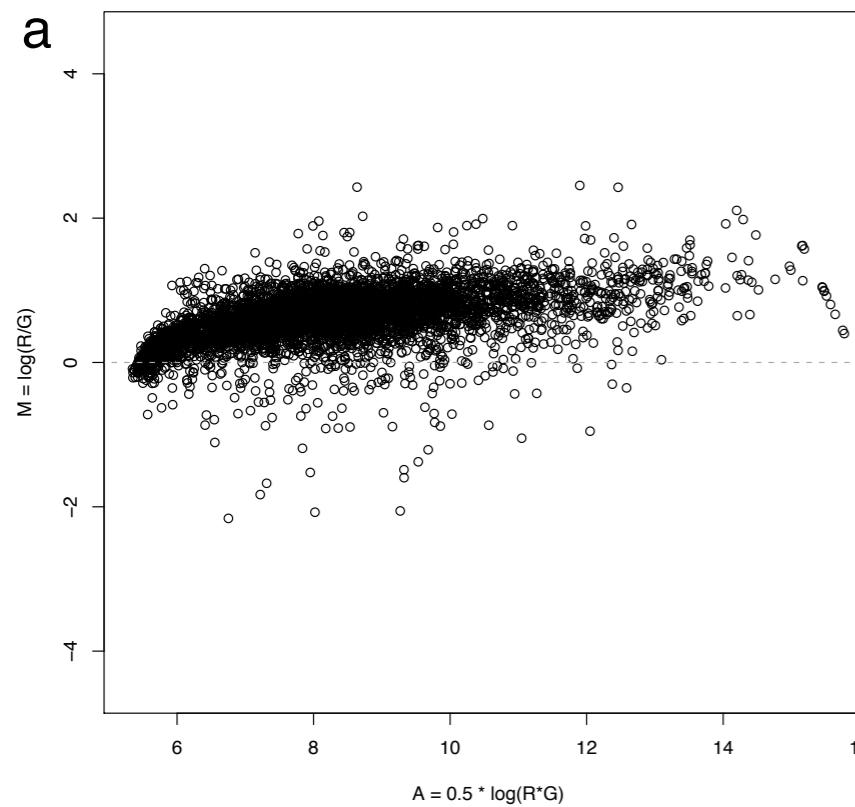
Multivariate Data



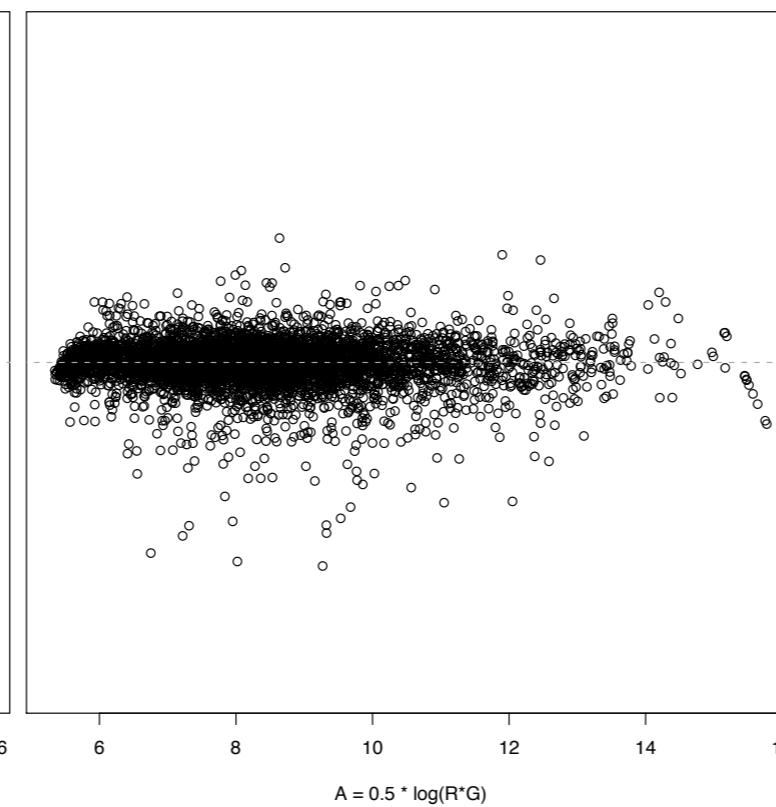
Multivariate Data: Transcriptomics

MA Plot: 1 array

before normalization

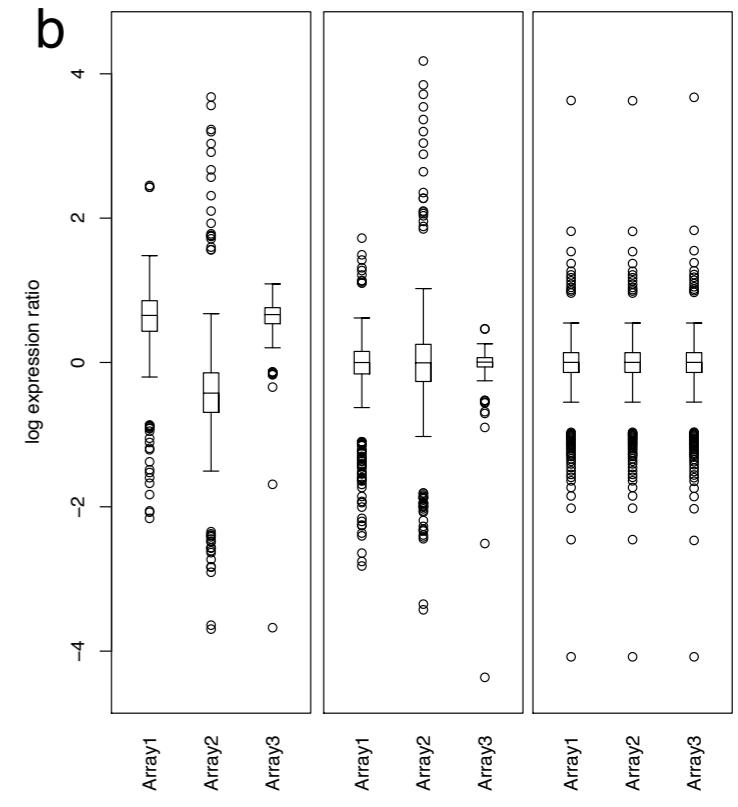


after normalization

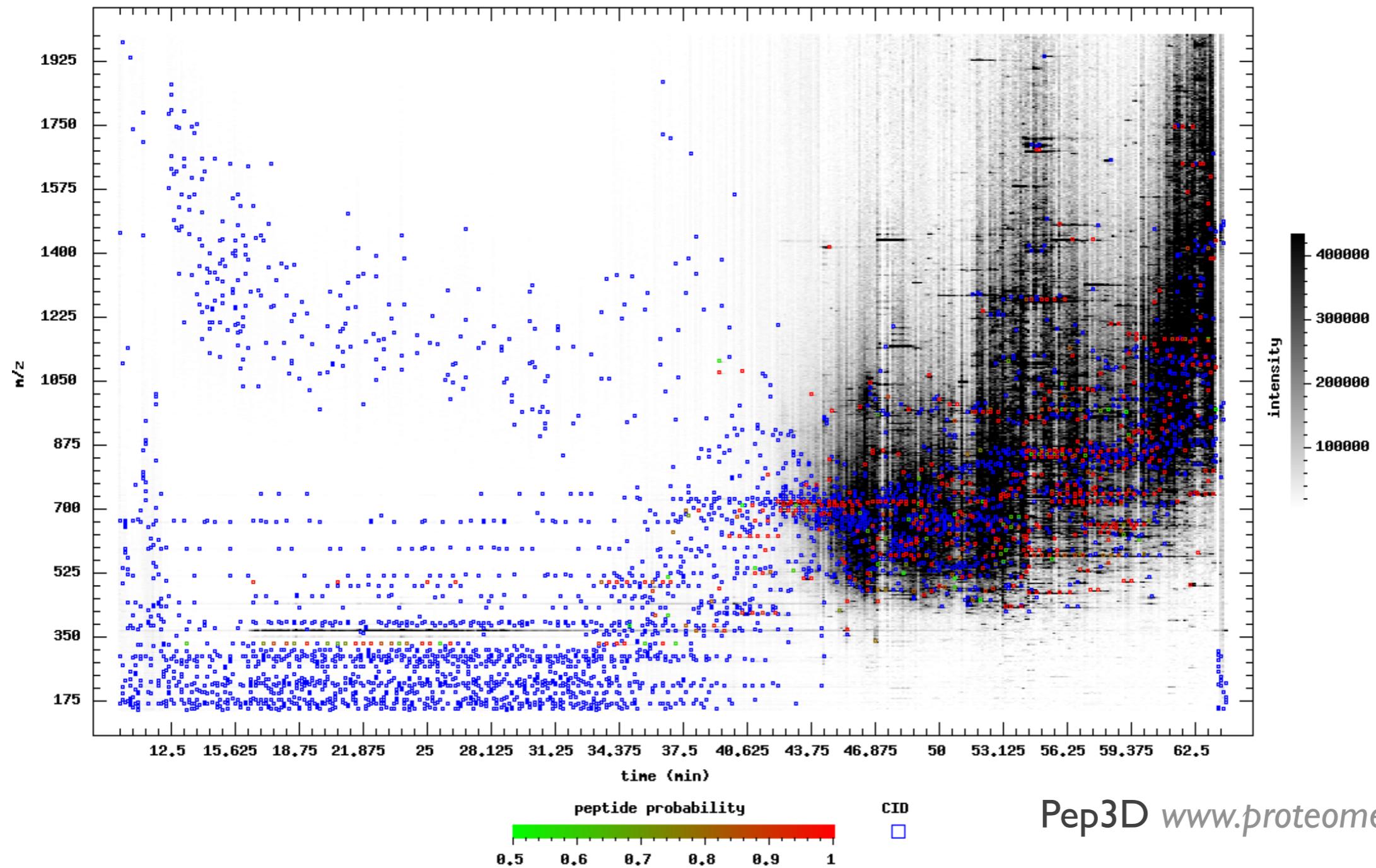


Box Plot: 3 arrays

1 2 3

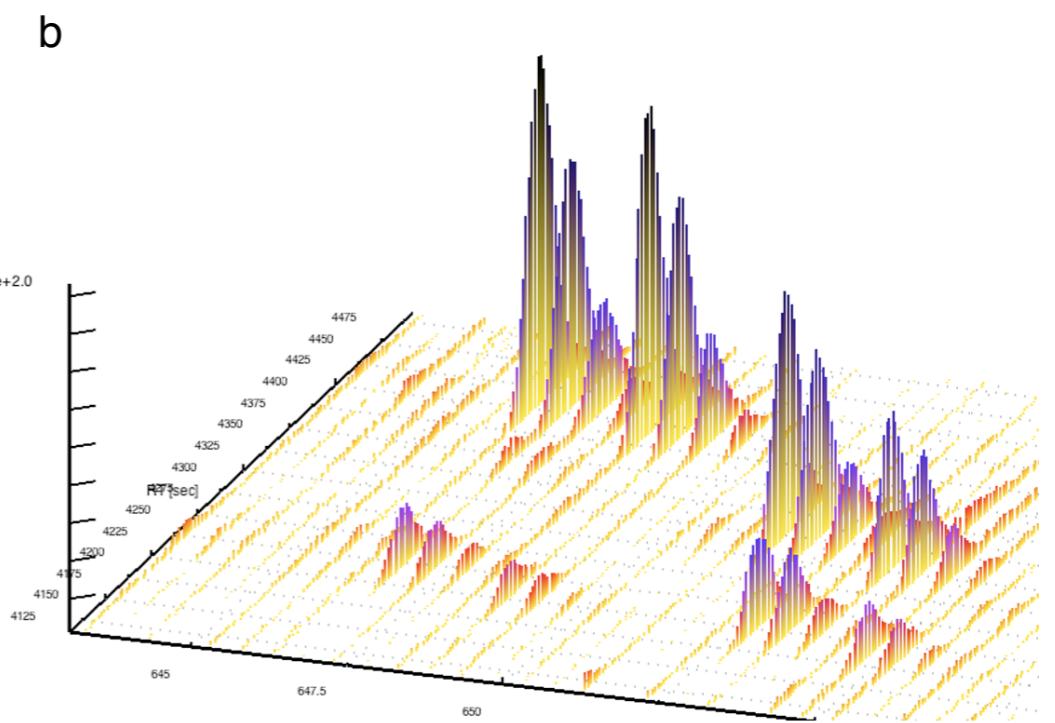
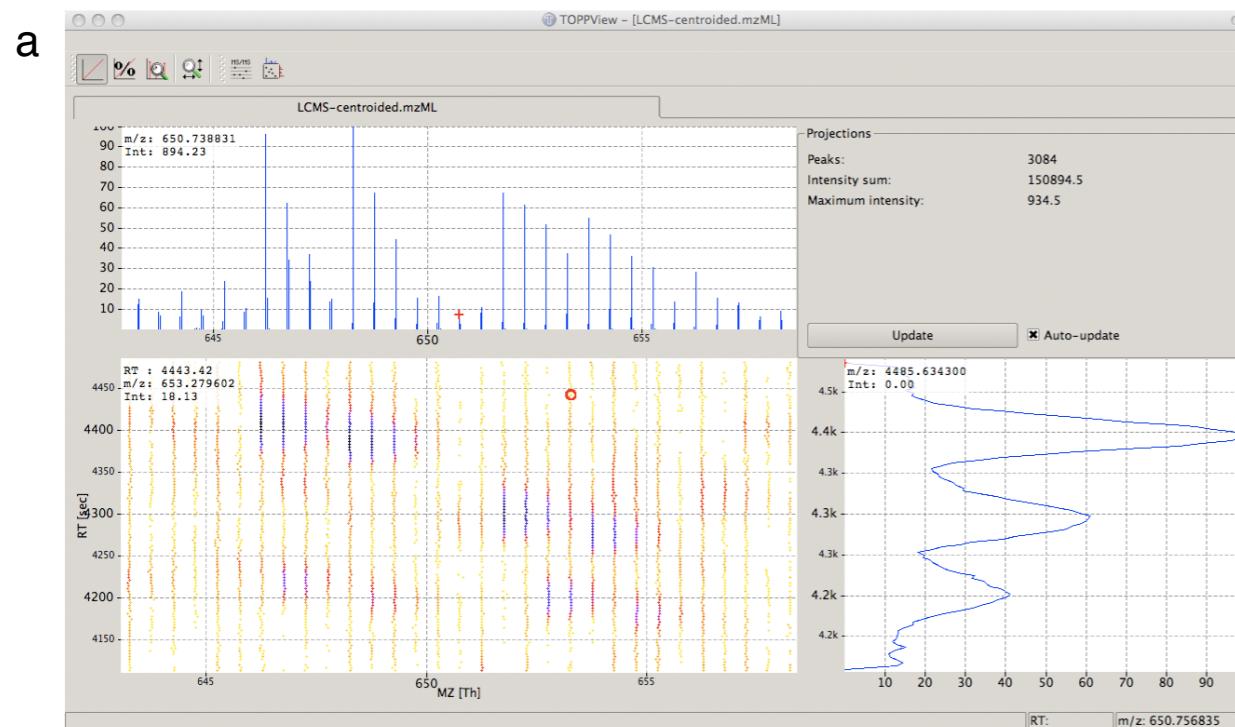


1 = before normalization
 2 = after within-array normalization
 3 = after between-array normalization



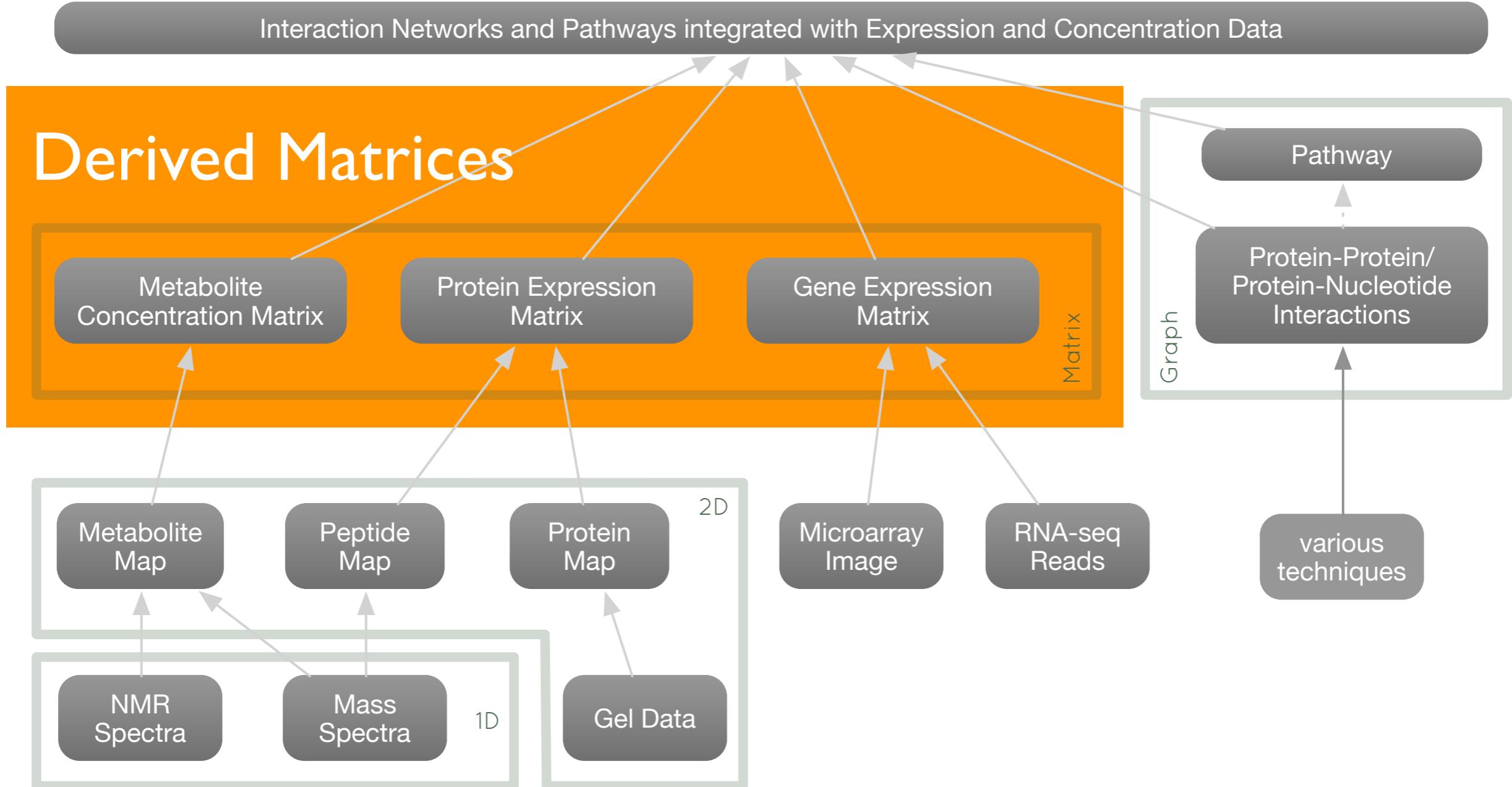
Pep3D www.proteomecenter.org

Multivariate Data: Proteomics



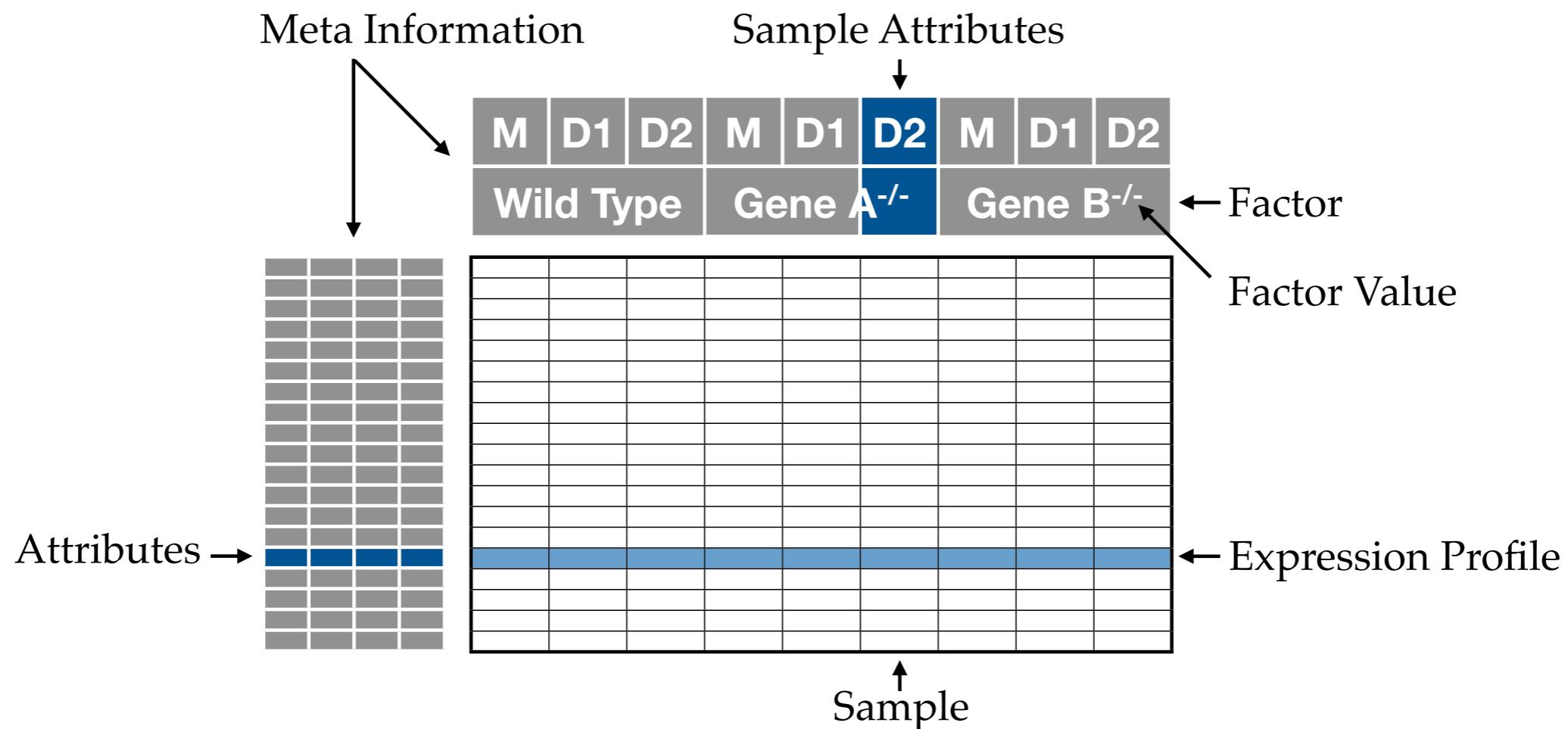
TOPPView www.open-ms.de

Multivariate Data: Derived Data



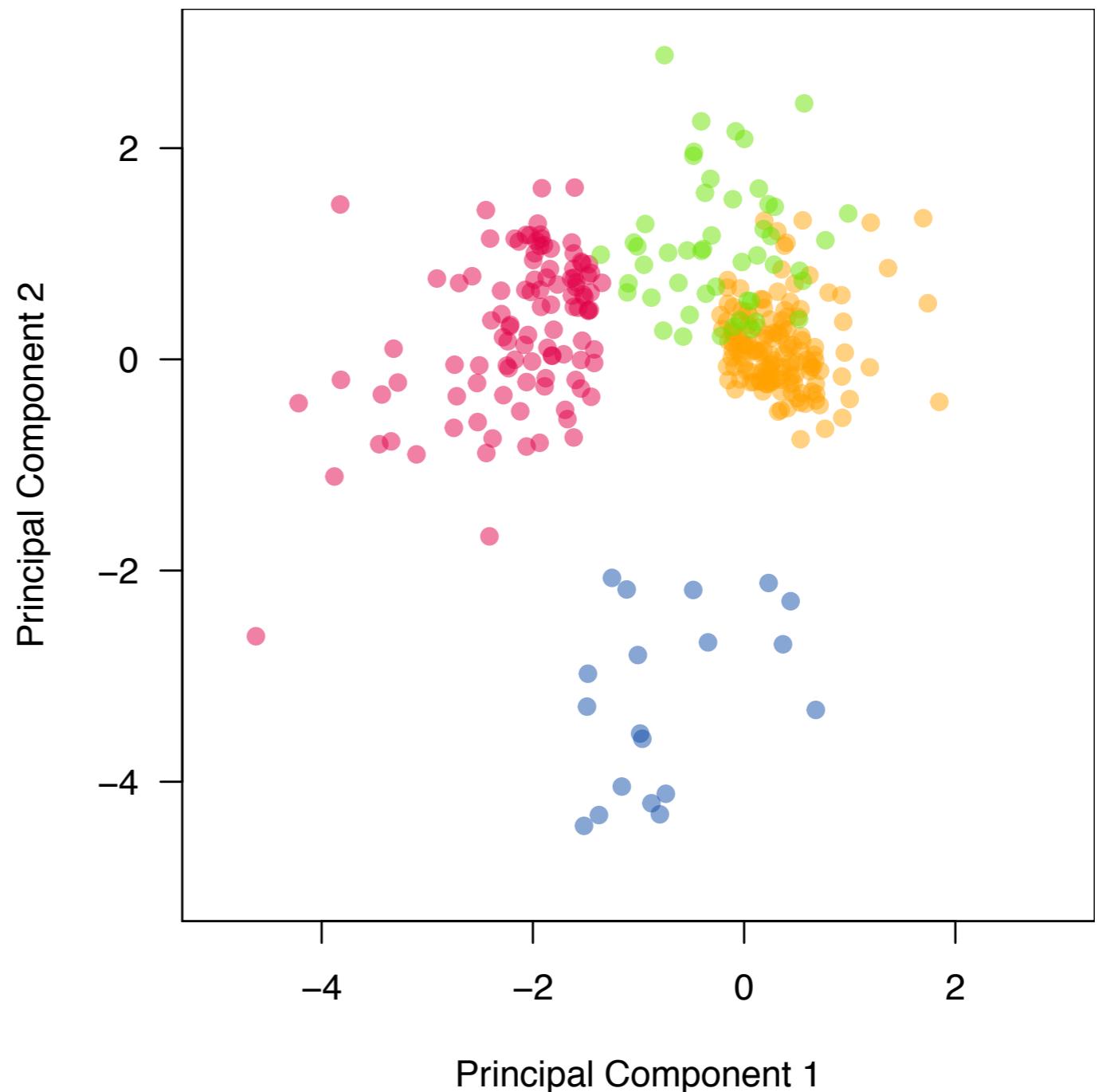
Multivariate Data: Derived Matrices

- matrices of multi-dimensional vectors
- usually abundance profiles, e.g. transcript or protein levels, metabolite concentrations



Multivariate Data: Derived Matrices

Scatter Plot

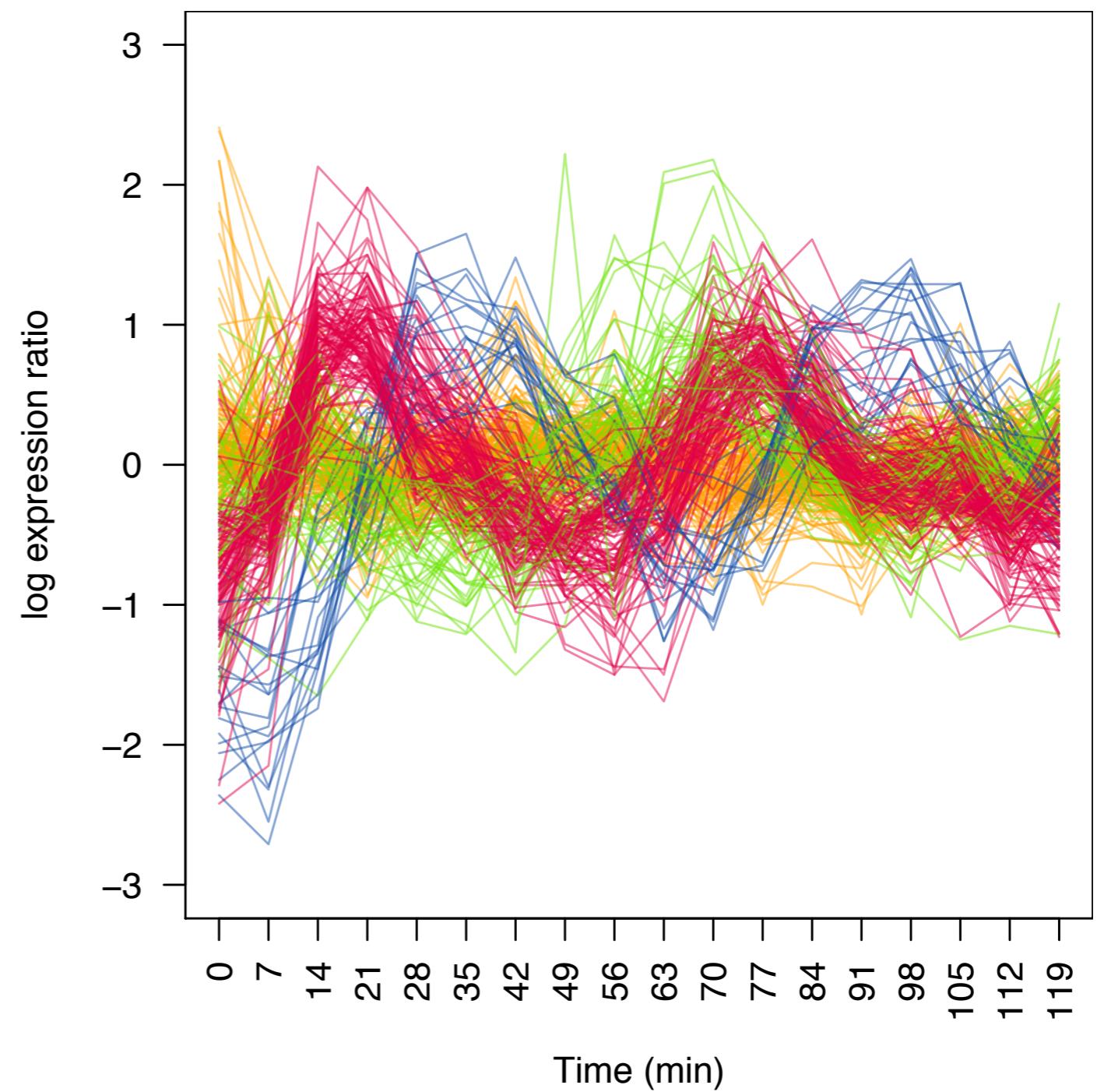


Multivariate Data: Derived Matrices

- **Scatter Plots and Dimensionality Reduction**
 - used to visualize high-dimensional profiles as projections in lower-dimensional spaces (usually 2D, sometimes also 3D ...)
 - there is always a loss of information in the process, goal is to minimize the loss of information
 - many different algorithms: Principal Components Analysis (PCA), Multi-Dimensional Scaling (MDS), Isomap, etc.
 - **Pros** - good choice to get an idea about the overall structure of the whole data set: clusters, outliers, gaps in the data
 - **Cons** - because of the dimensionality reduction the original profiles are not accessible in the visualization

Multivariate Data: Derived Matrices

Profile Plot a.k.a.
Parallel Coordinates



Multivariate Data: Derived Matrices

- **Profile Plot/Parallel Coordinate Plots**

- **Pros**

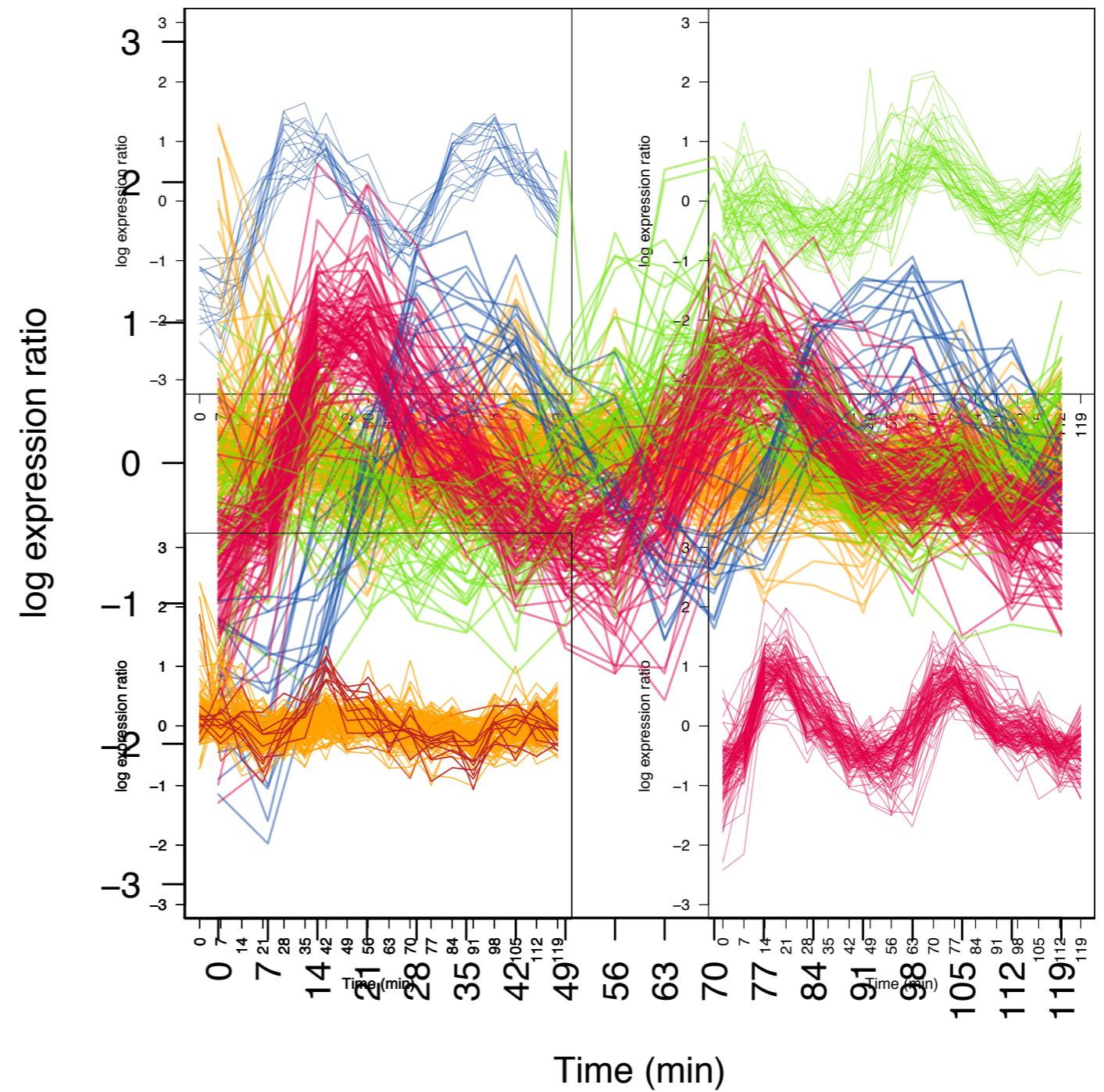
- encoding by position: profiles easy to read
 - color-coding of expression profiles (groups) very efficient

- **Cons**

- overplotting
 - grows horizontally with every additional sample

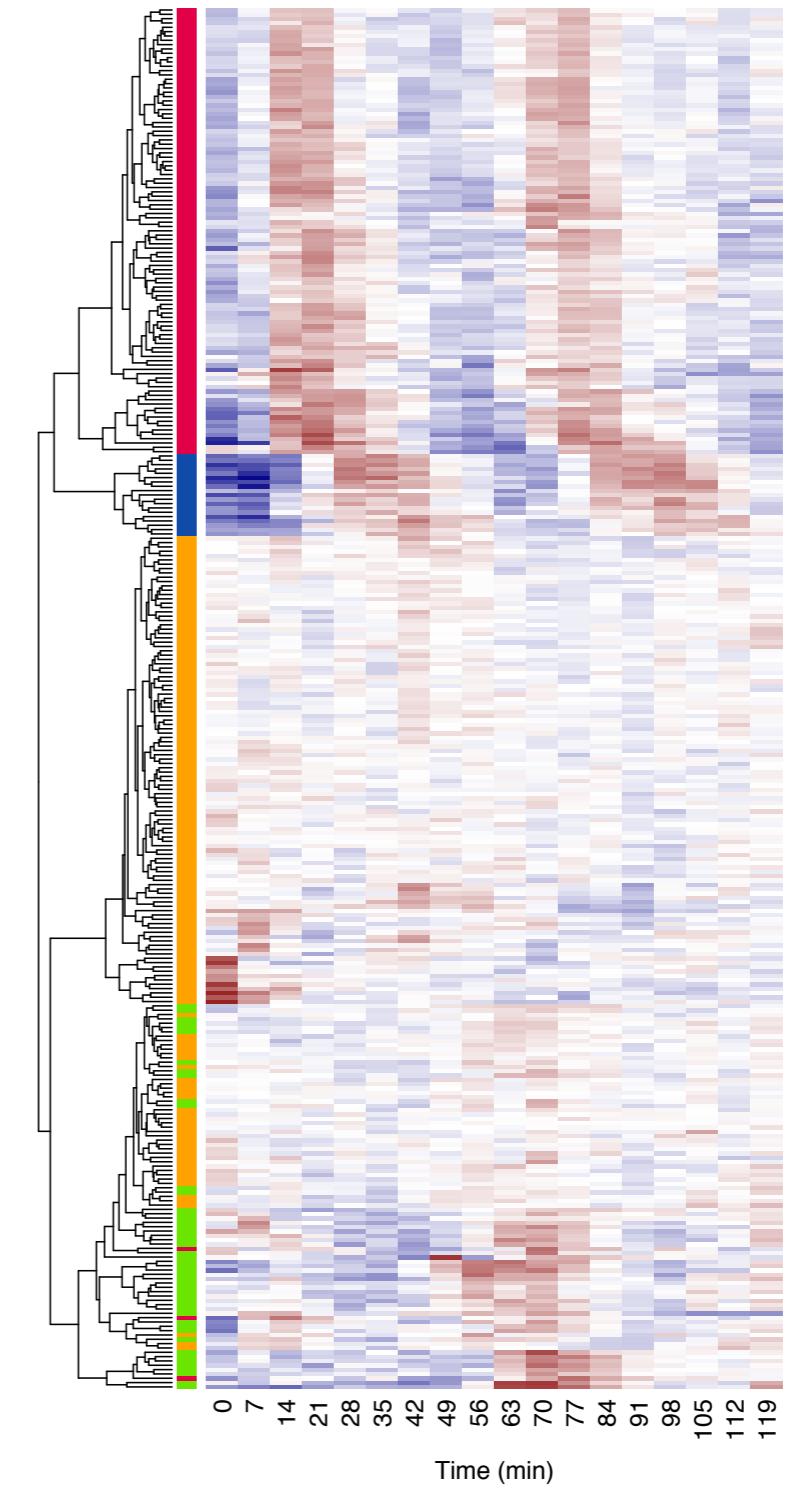
Multivariate Data: Derived Matrices

Profile Plot a.k.a.
Parallel Coordinates



Multivariate Data: Derived Matrices

Heat Map with Dendrogram



Multivariate Data: Derived Matrices

- **Heatmap**

- **Pros**

- no overplotting, yet a very dense information display
 - can be combined with dendrogram and additional information can be encoded in further columns or in the height of rows

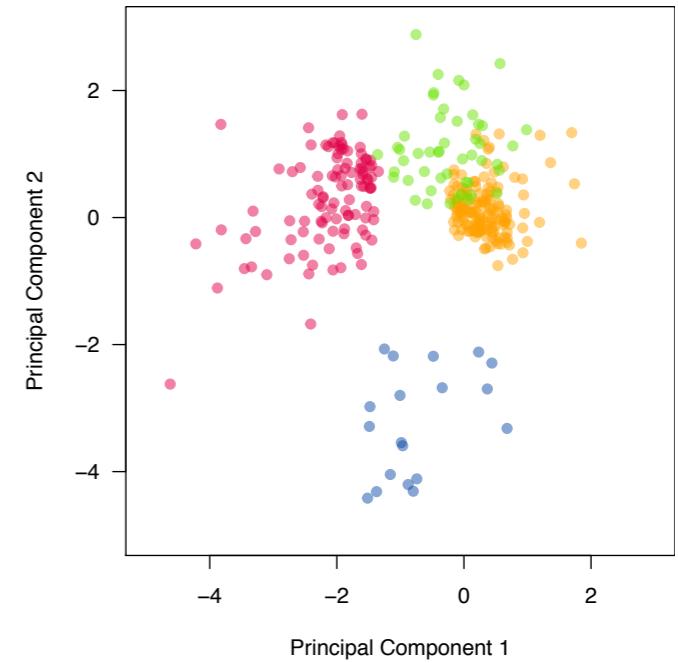
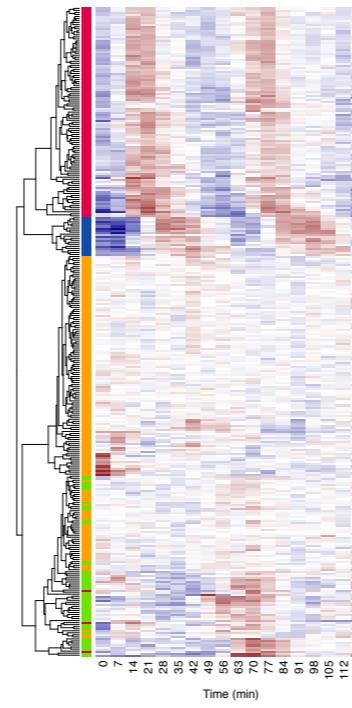
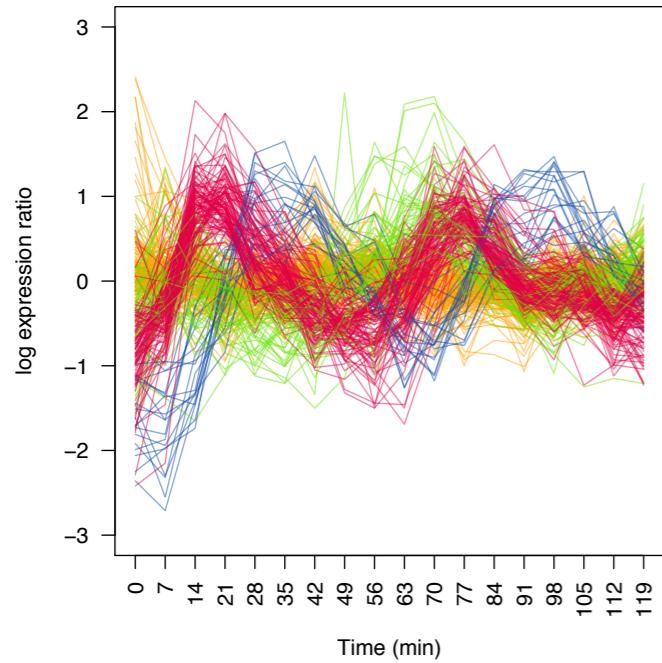
- **Cons**

- only qualitative interpretation possible due to color coding
 - grows horizontally with every additional sample and grows vertically with every additional profile

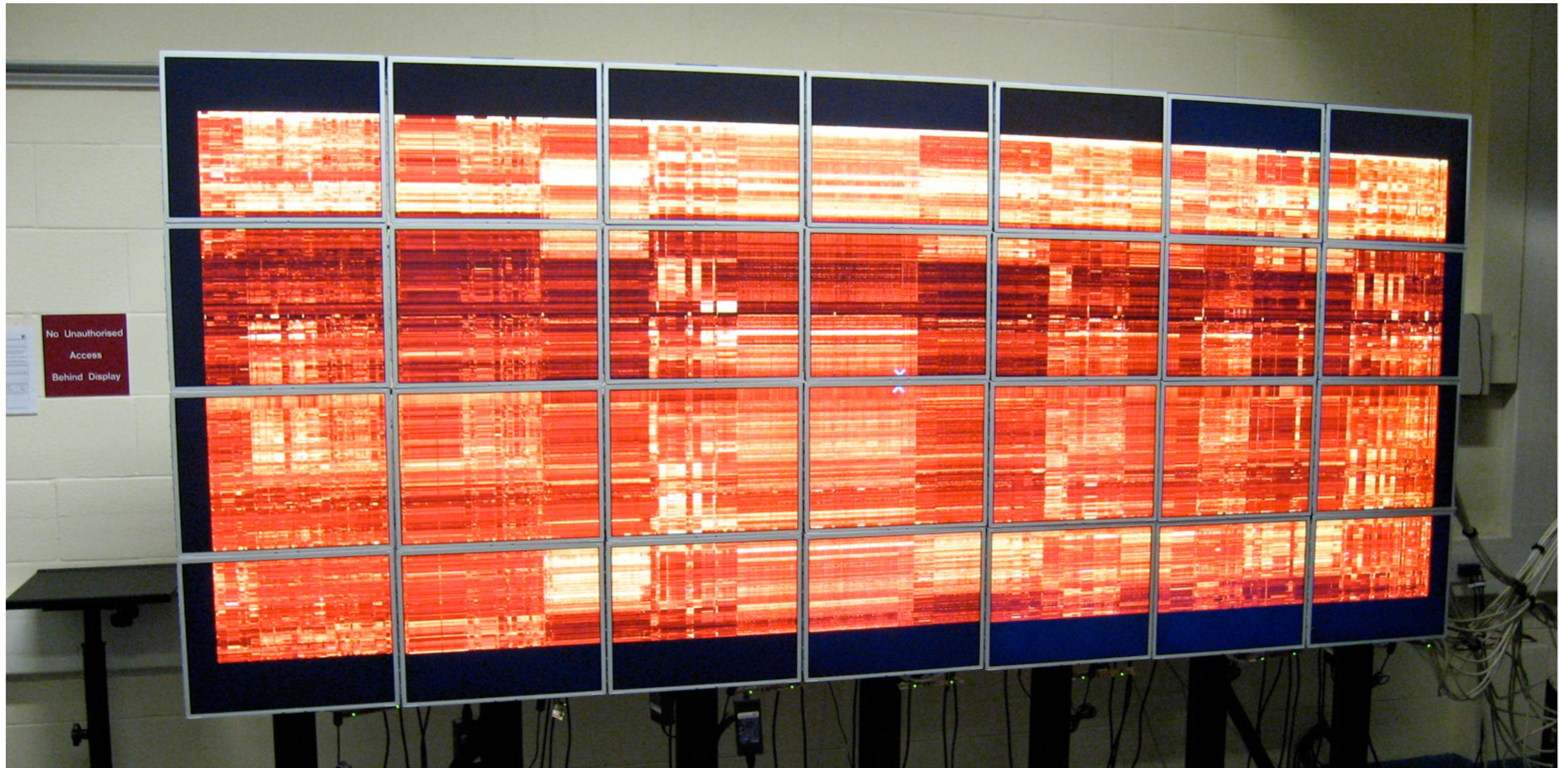
Multivariate Data: Summary

few, high-res

many, low-res



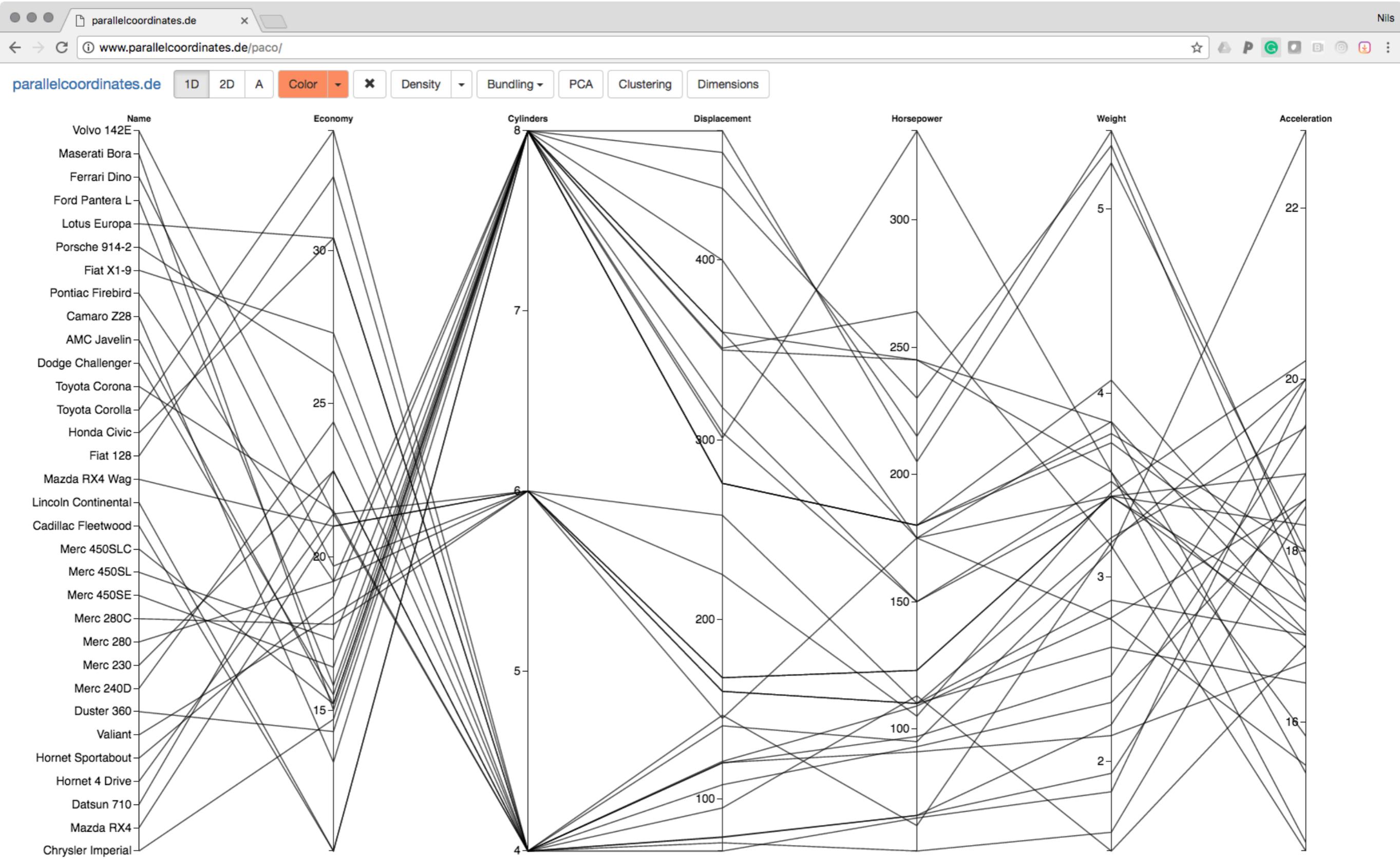
Problem: Very Large Expression Matrices



Power Wall (7x4 screens = 11,200x4,800), University of Leeds

1000 transcripts, 5372 samples

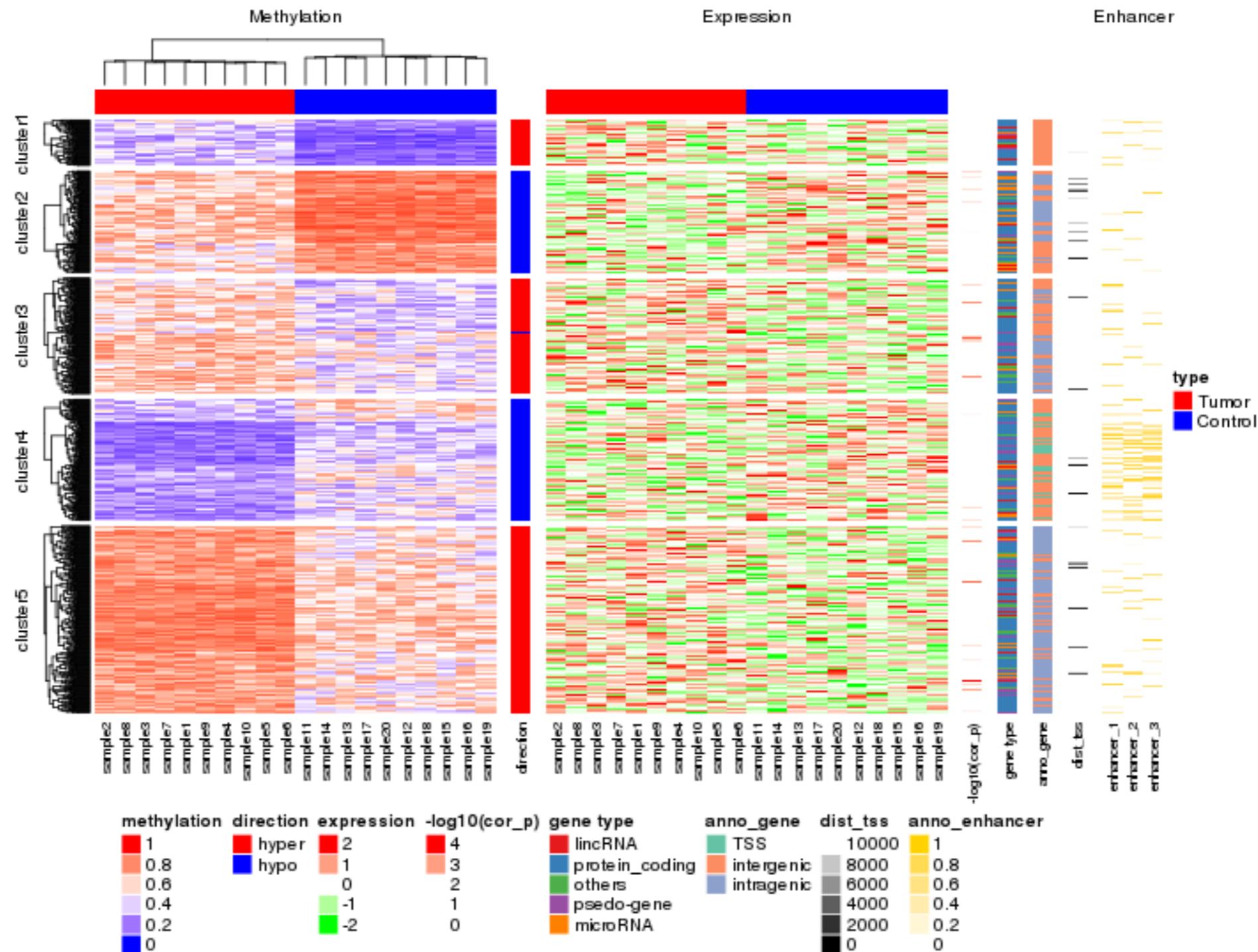
High-Dimensional Multivariate Data Heterogeneous Tables



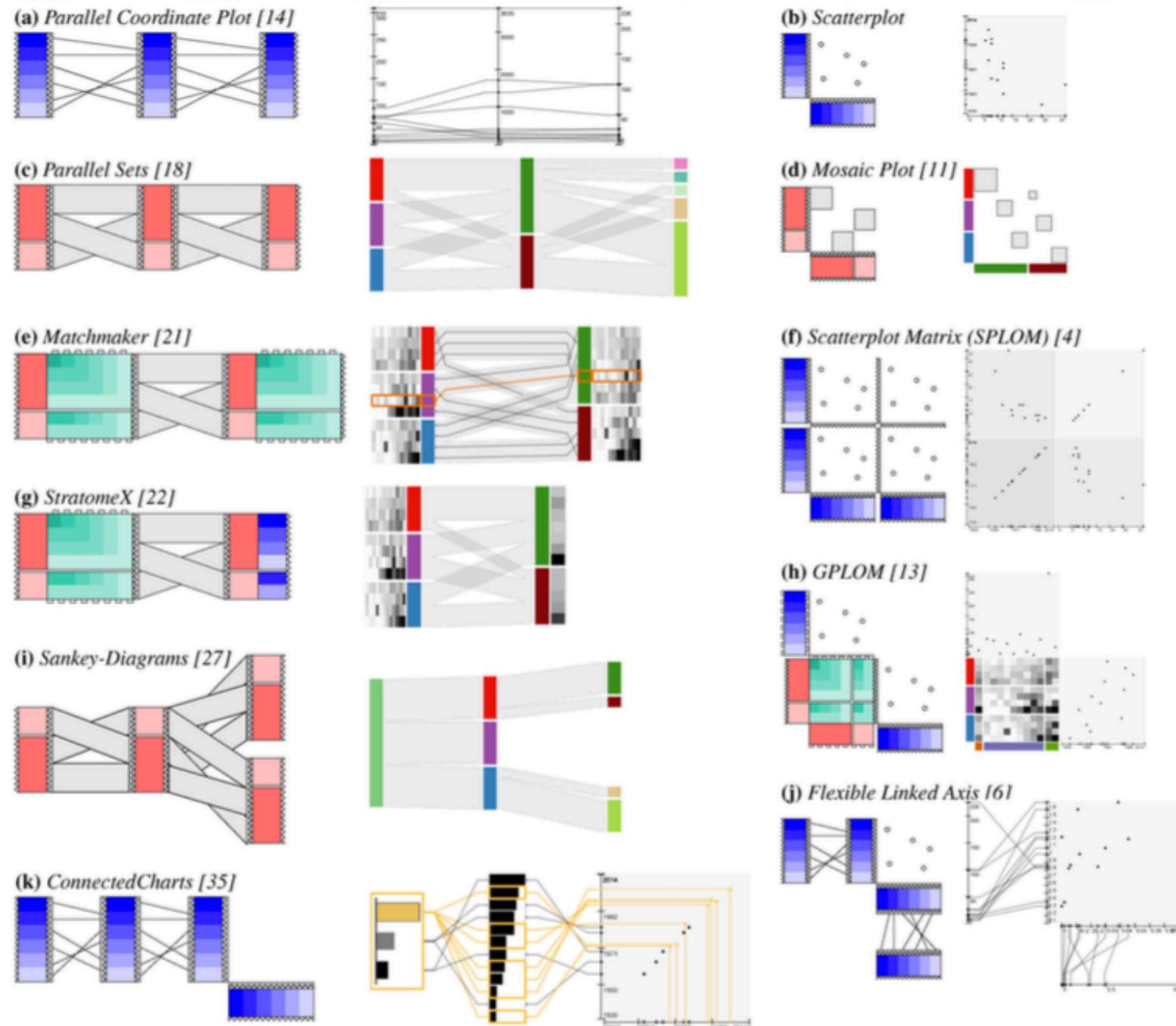
drop a csv-formatted file to load your own data. Note that the first line must describe the data scheme. See [this dataset](#) for an example.

<http://www.parallelcoordinates.de/paco/>

Correspondence between methylation, expression and other genomic features



Domino



Domino

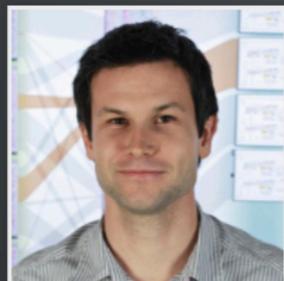
Extracting, Comparing, and Manipulating Subsets
across Multiple Tabular Datasets



JKU
JOHANNES KEPLER
UNIVERSITY LINZ



Samuel Gratzl



Marc Streit



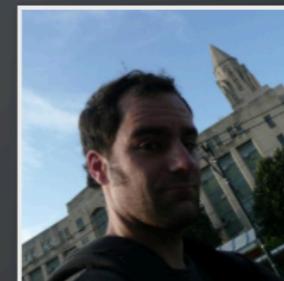
HARVARD
MEDICAL SCHOOL



Nils Gehlenborg



HARVARD
School of Engineering
and Applied Sciences



Alexander Lex



Hanspeter Pfister

https://sgratzl.github.io/domino_vis2014/#/title

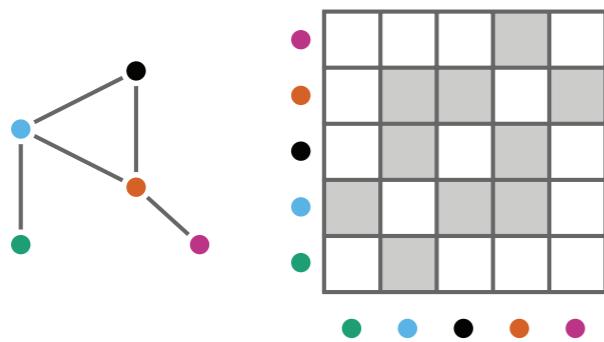
Network Visualization

Refresher

- What is a graph? **This is a data structure not a visual representation!**
 - Set **V** of vertices (= nodes) and set **E** of edges
 - Directed graph vs undirected graph
 - Directed acyclic graph (DAG)
- What is a tree?
 - Binary tree: two children per vertex
 - General tree: any number of children

Storing Graphs and Trees

- Use a matrix for adjacency (edges explicit)
 - edges are matrix cells



- Use a list with two fields per vertex (edges implicit)
 - incoming edges
 - outgoing edges

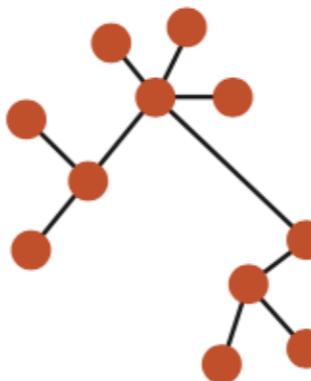
Arrange Networks and Trees

→ Node–Link Diagrams

Connection Marks

NETWORKS

TREES



→ Adjacency Matrix

Derived Table

NETWORKS

TREES

■	■	■	■	■
■	■	■	■	■
■	■	■	■	■
■	■	■	■	■
■	■	■	■	■

→ Enclosure

Containment Marks

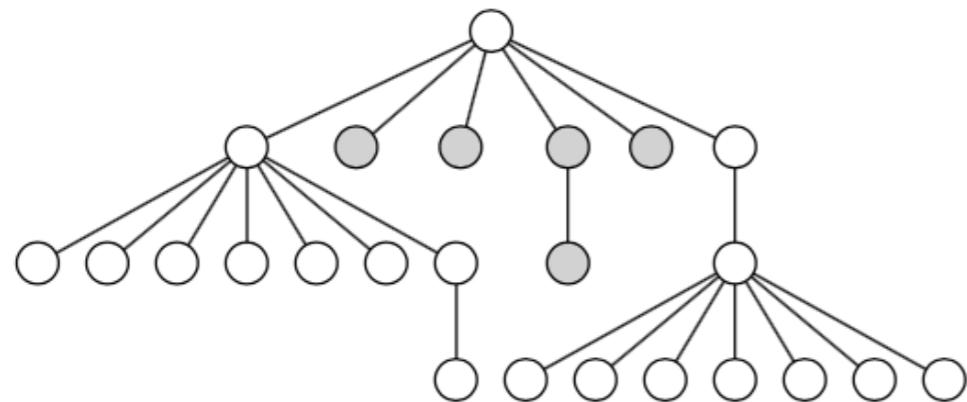
NETWORKS

TREES

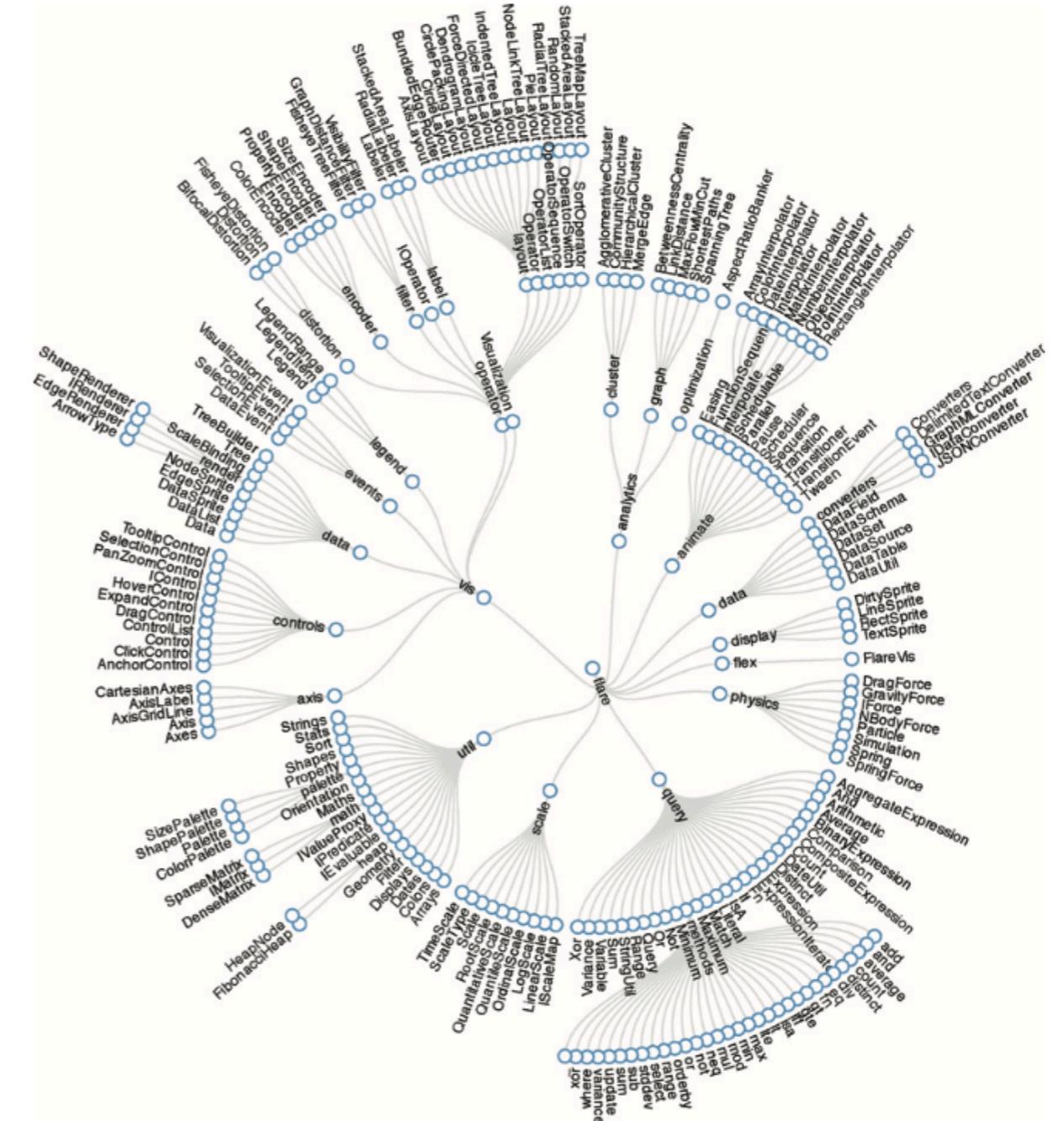


Node-Link Diagram Approaches

Trees



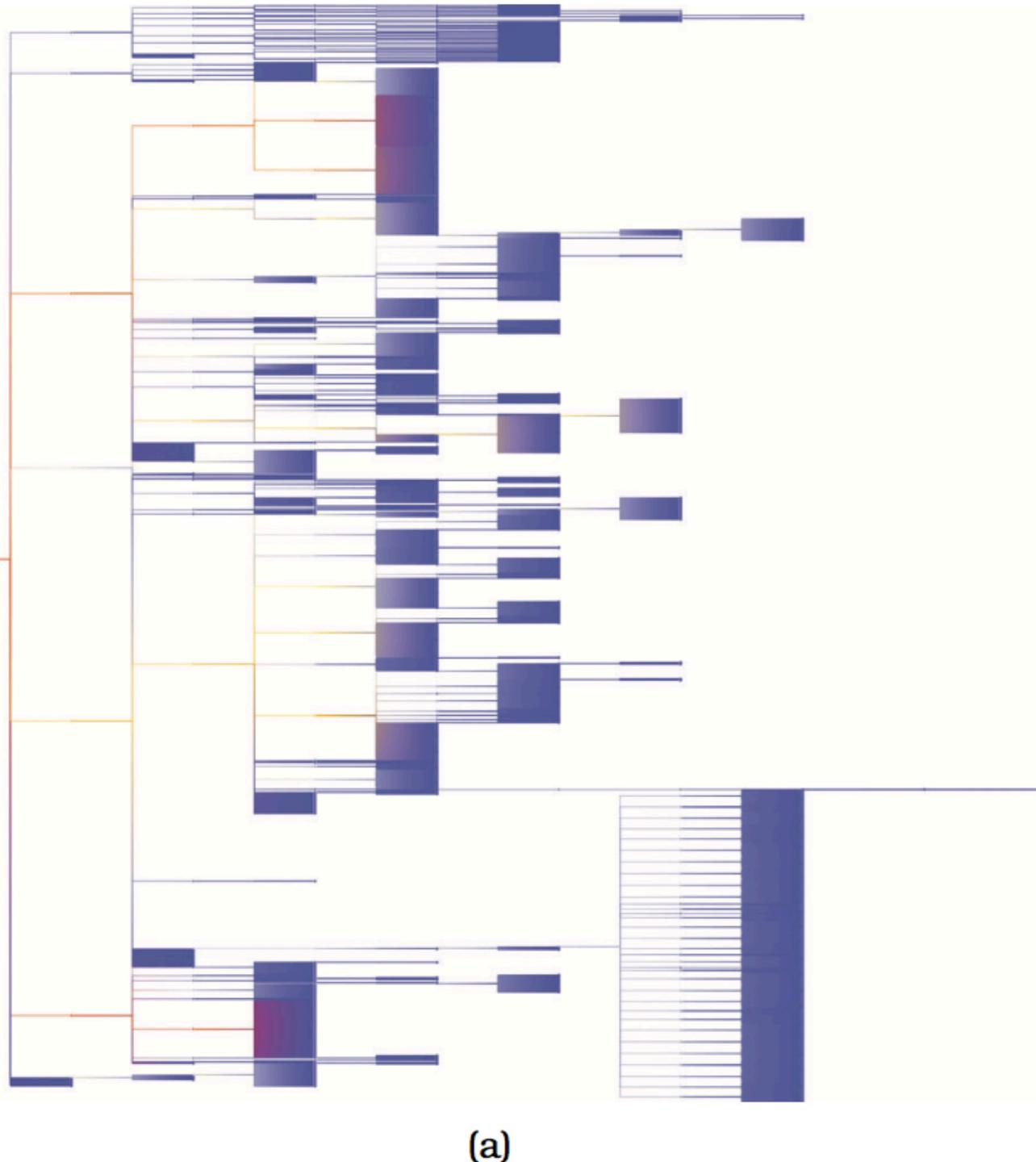
(a)



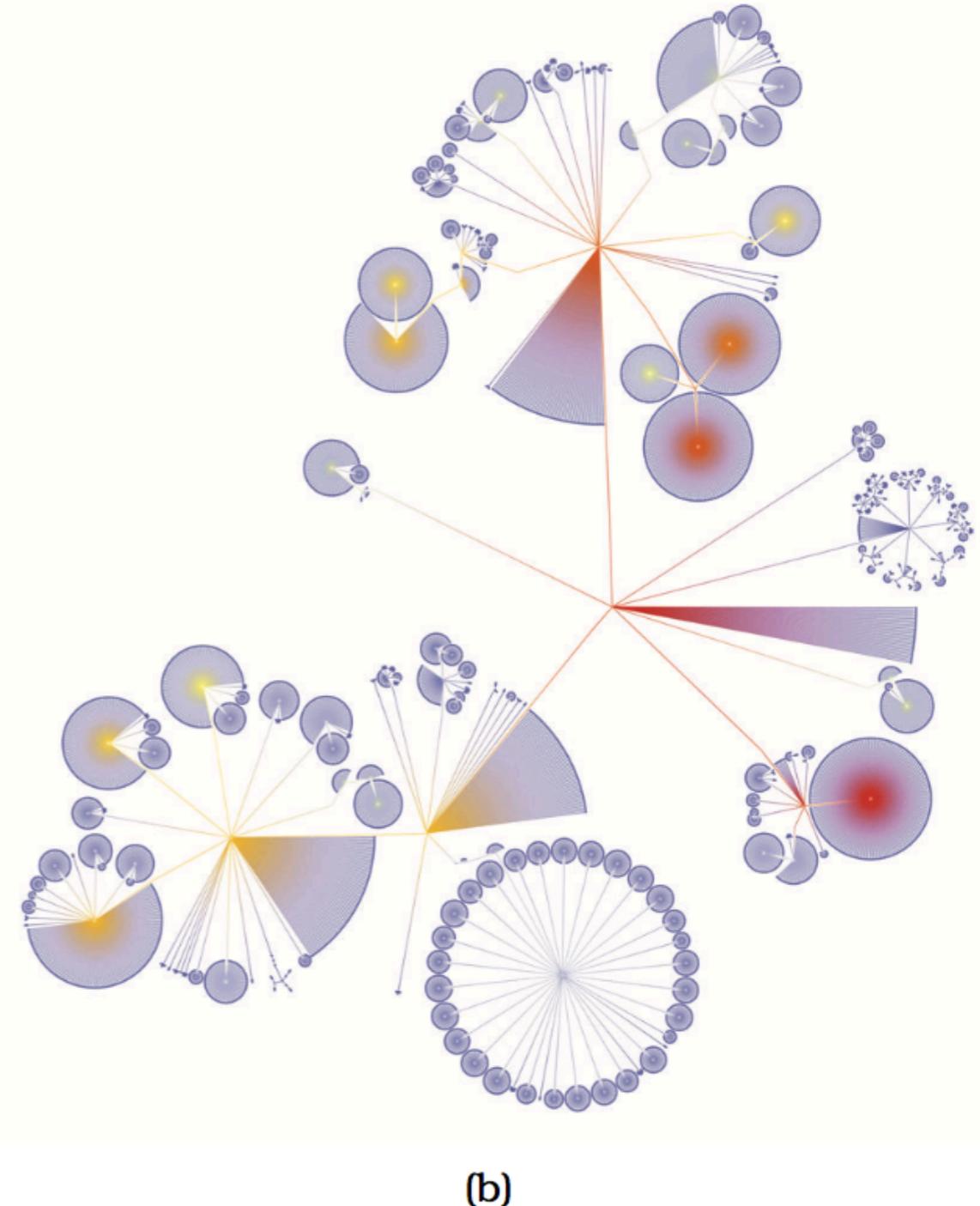
(b)

Figure 9.2. Node-link layouts of small trees. (a) Triangular vertical for tiny tree. From [Buchheim et al. 02, Figure 2d]. (b) Spline radial layout for small tree. From <http://mbostock.github.com/d3/ex/tree.html>.

Trees



(a)



(b)

Figure 9.3. Two layouts of a 5161-node tree. (a) Rectangular horizontal node-link layout. (b) BubbleTree node-link layout.

Graphs

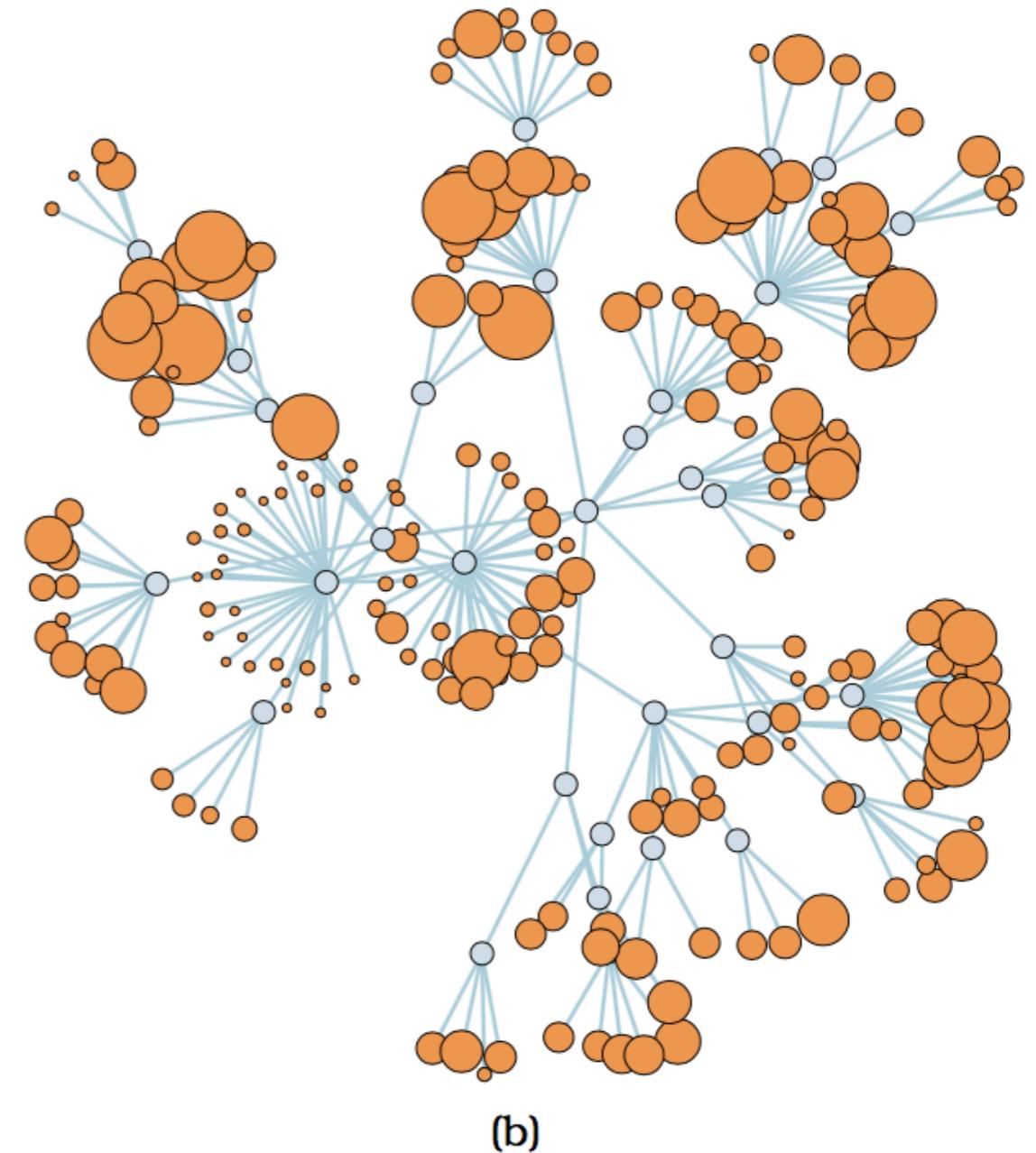
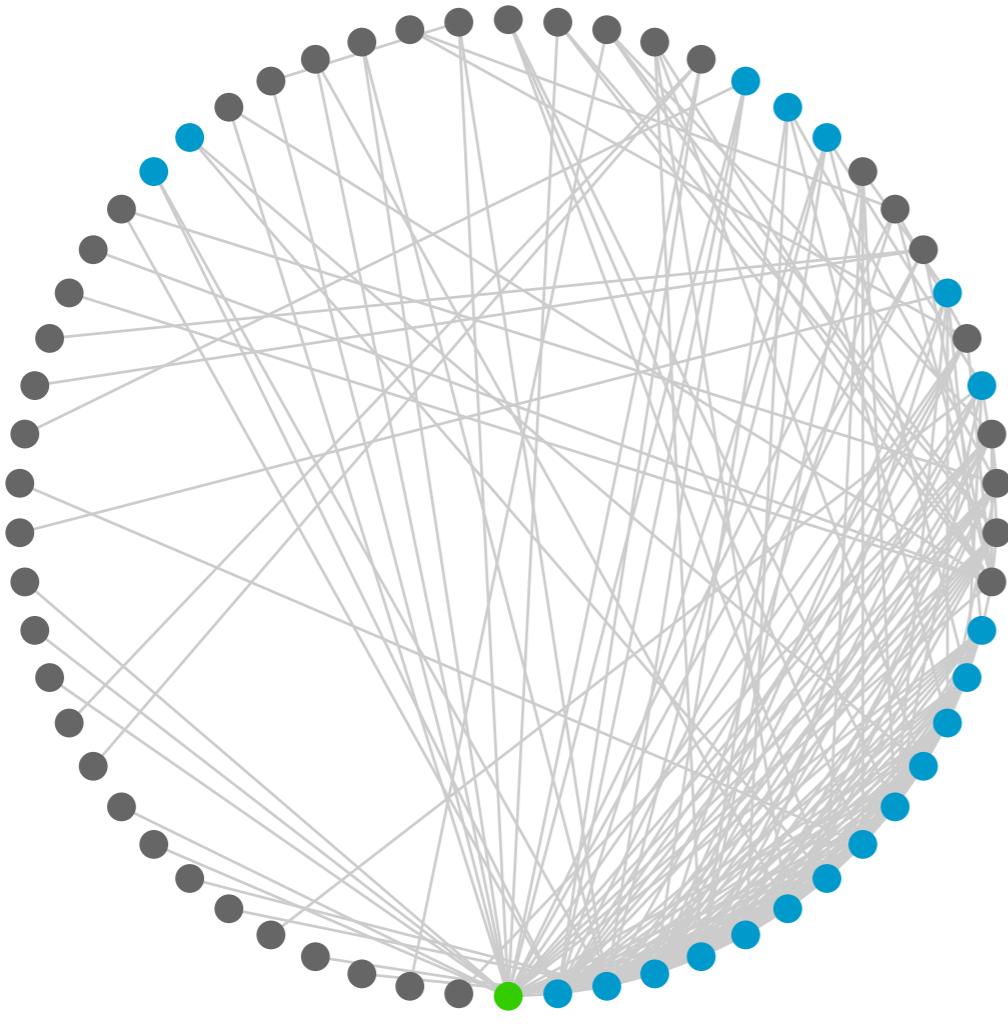
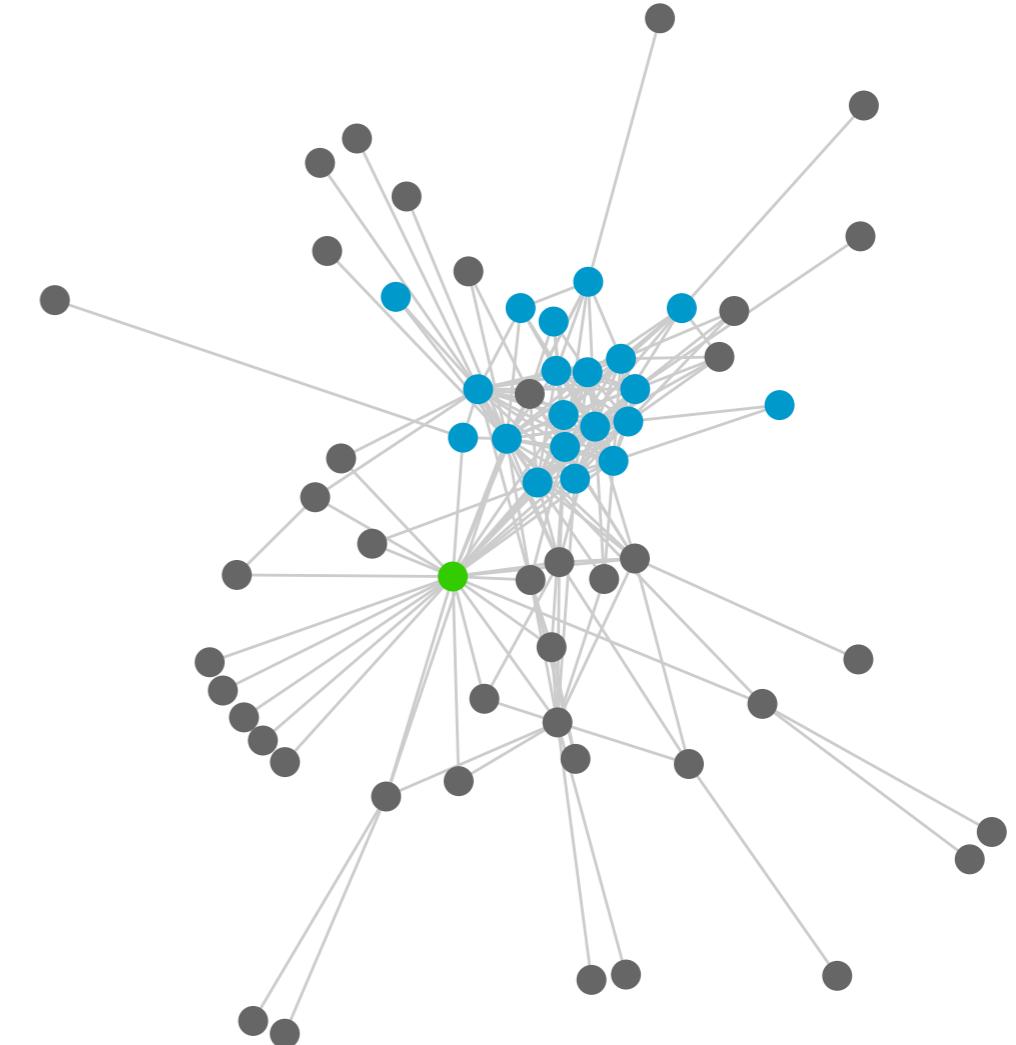


Figure 9.4. Node-link layouts of small networks. (a) Force-directed placement of small network of 75 nodes, with size coding for link attributes. (b) Larger network, with size coding for node attributes. From <http://bl.ocks.org/mbostock/4062045> and <http://bl.ocks.org/1062288>.

Graph Layouts

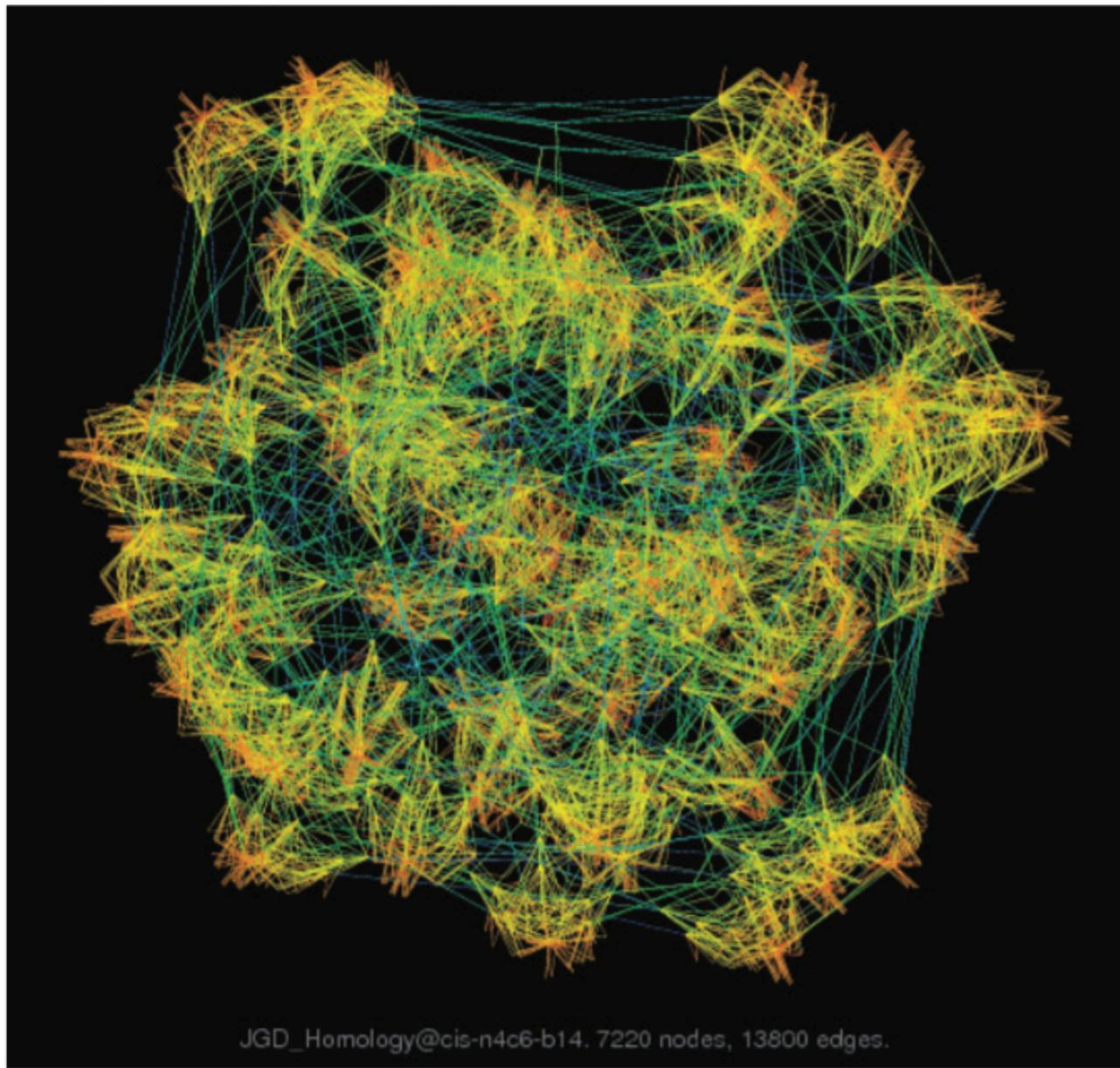


Circular Layout

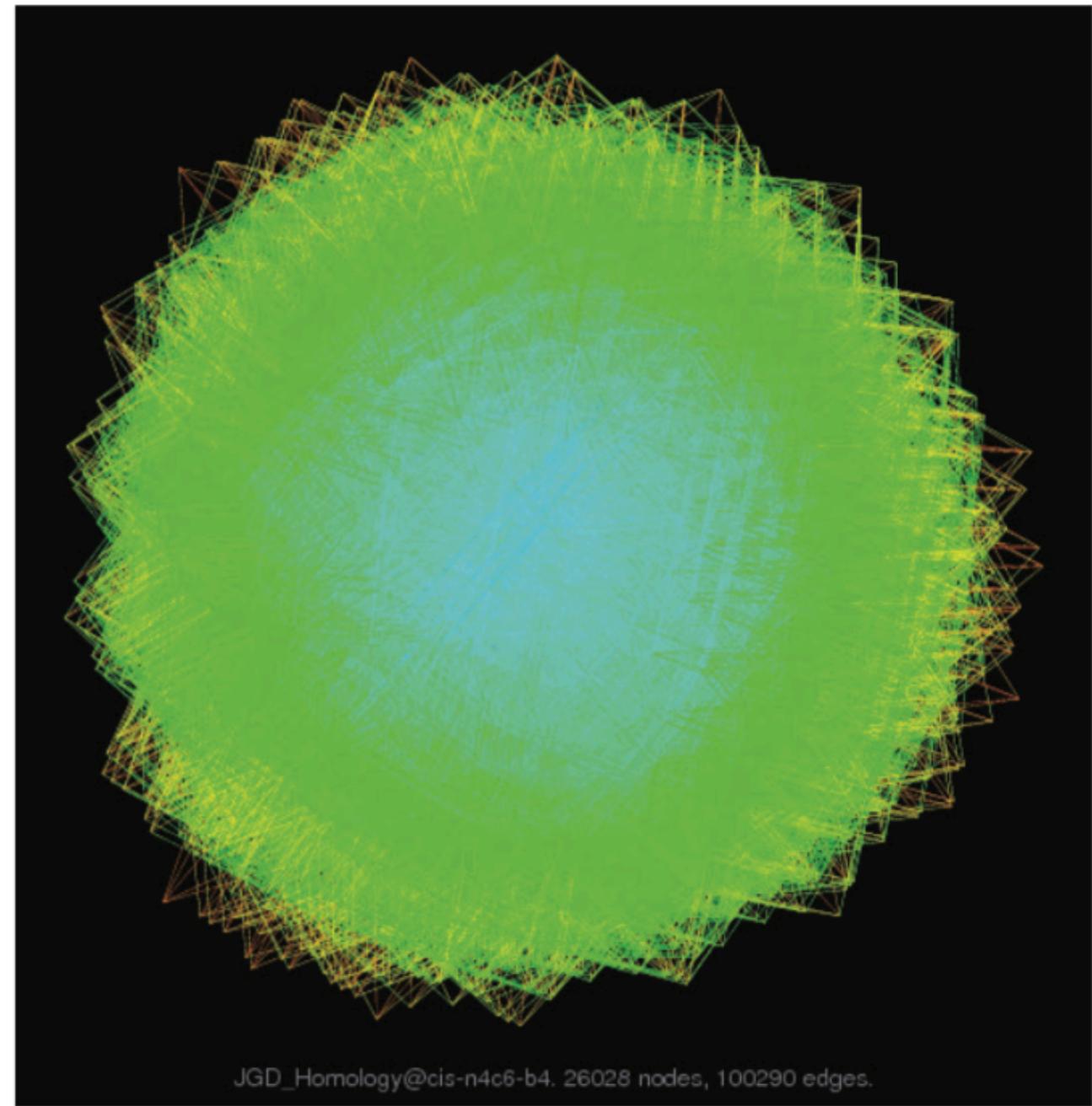


Force-directed Layout

Graph Layouts



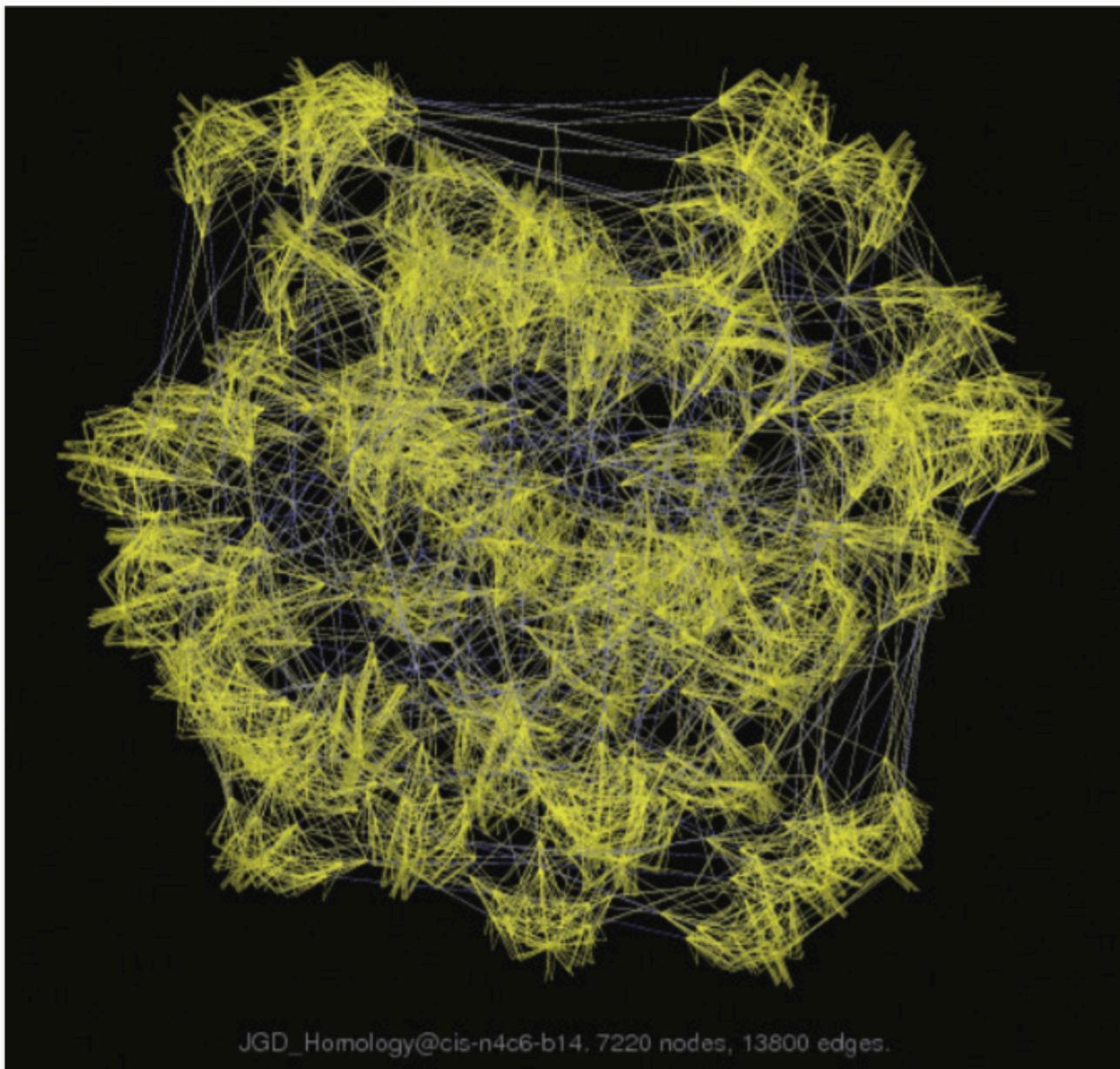
(a)



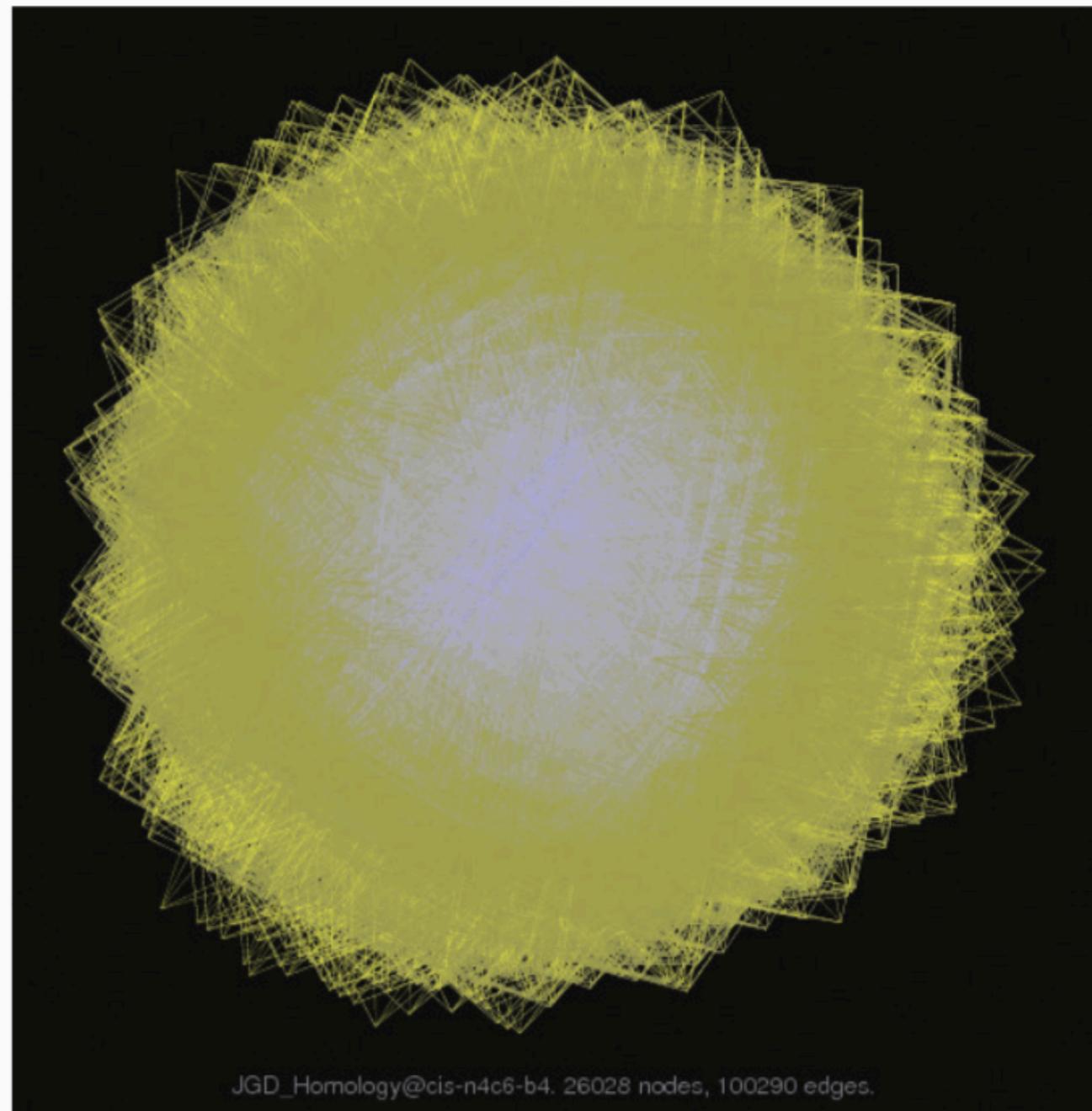
(b)

Figure 9.5. Multilevel graph drawing with sfdp [Hu 05]. (a) Cluster structure is visible for a large network of 7220 nodes and 13,800 edges. (b) A huge graph of 26,028 nodes and 100,290 edges is a “hairball” without much visible structure. From [Hu 14].

Reminder: Red-Green Blindness!

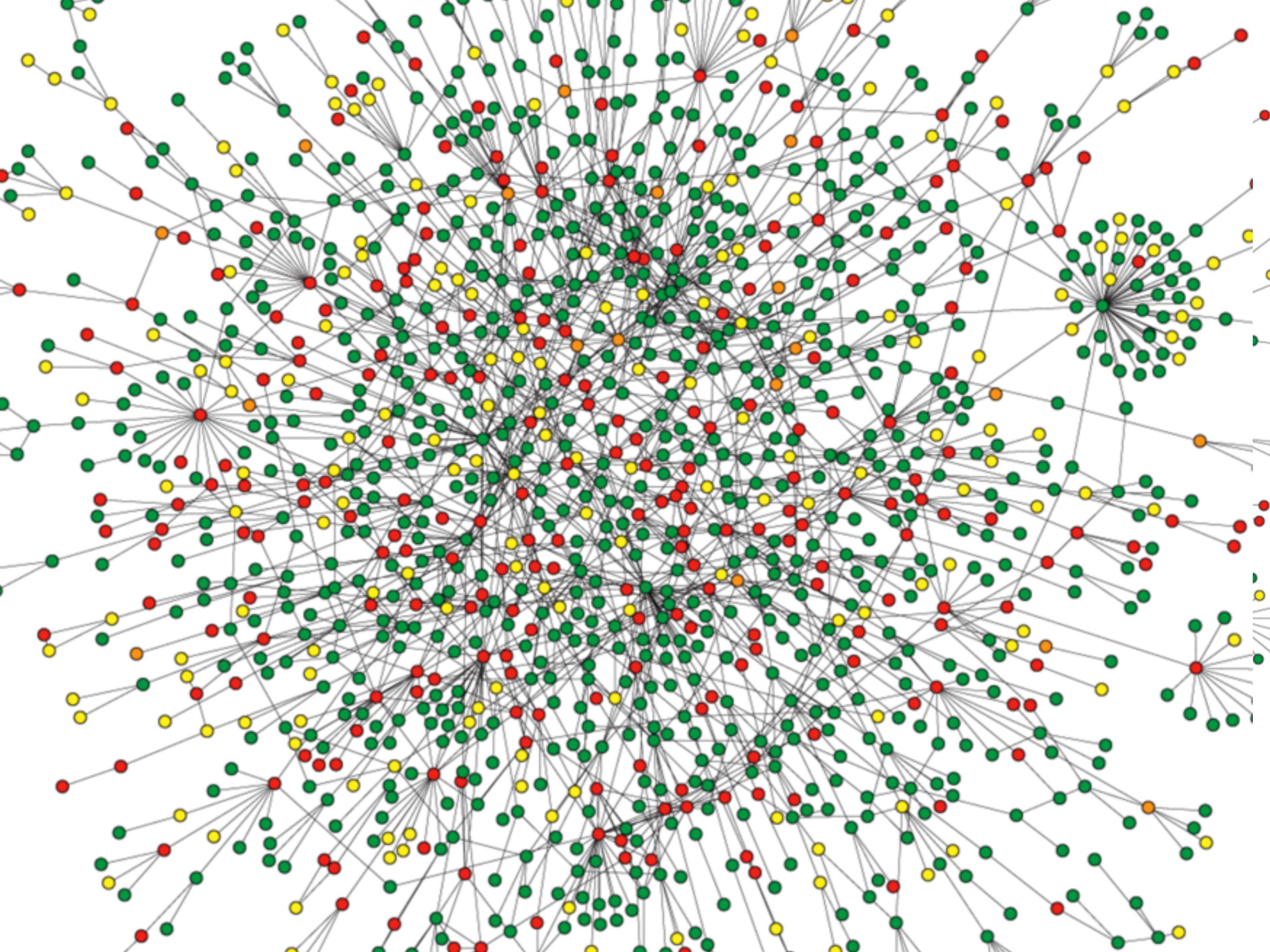


(a)

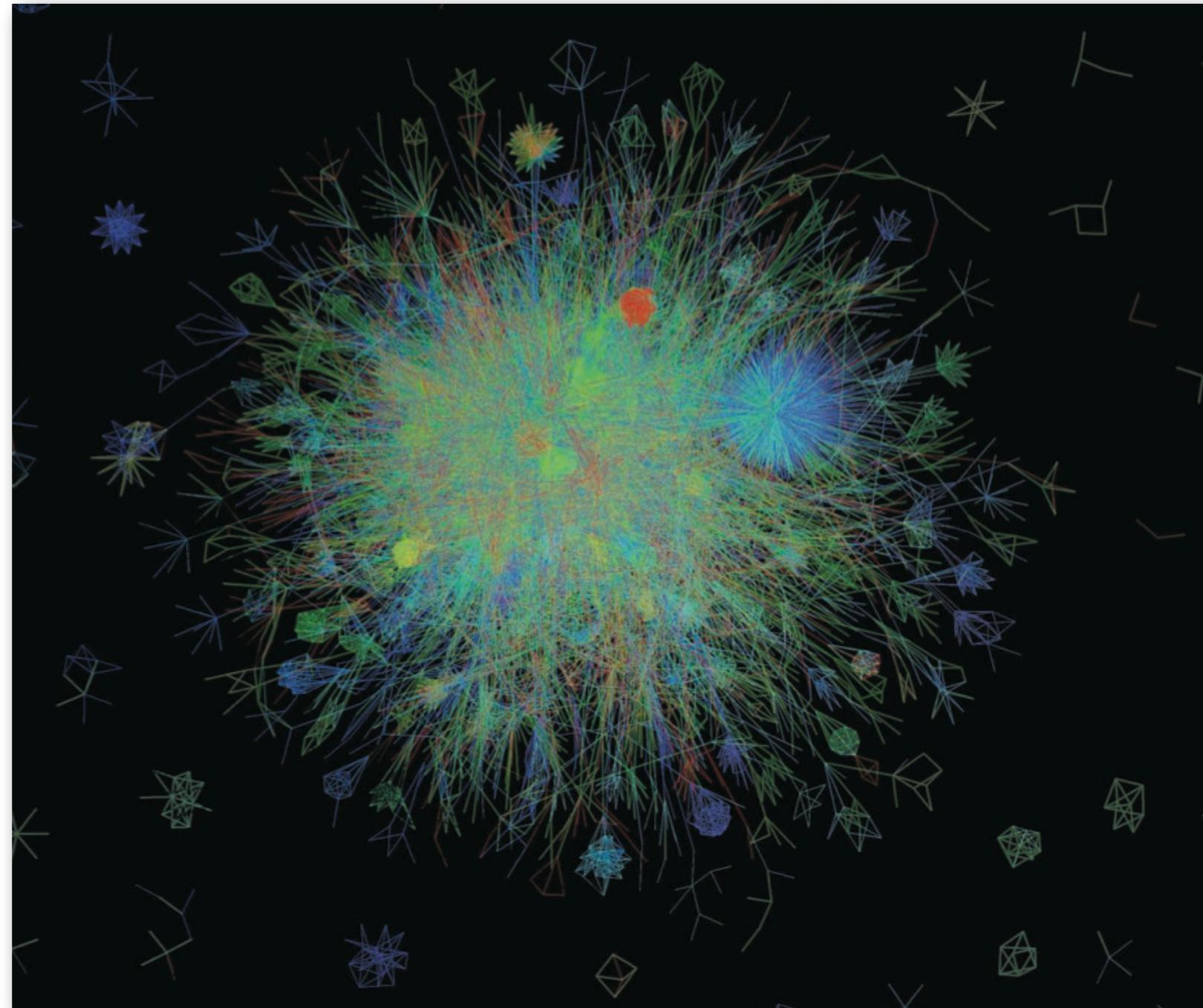


(b)

Figure 9.5. Multilevel graph drawing with sfdp [Hu 05]. (a) Cluster structure is visible for a large network of 7220 nodes and 13,800 edges. (b) A huge graph of 26,028 nodes and 100,290 edges is a “hairball” without much visible structure. From [Hu 14].

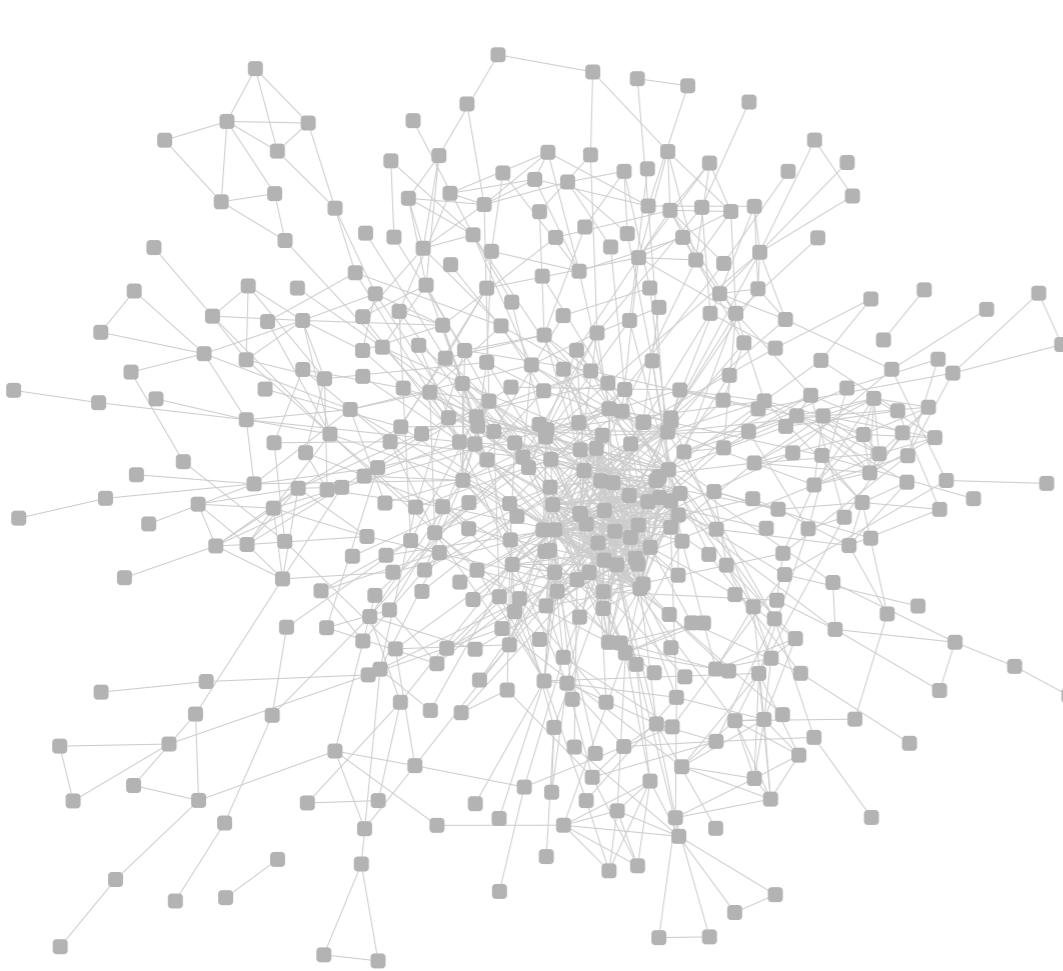


Interactions: Hairballs, Ridiculograms et al.

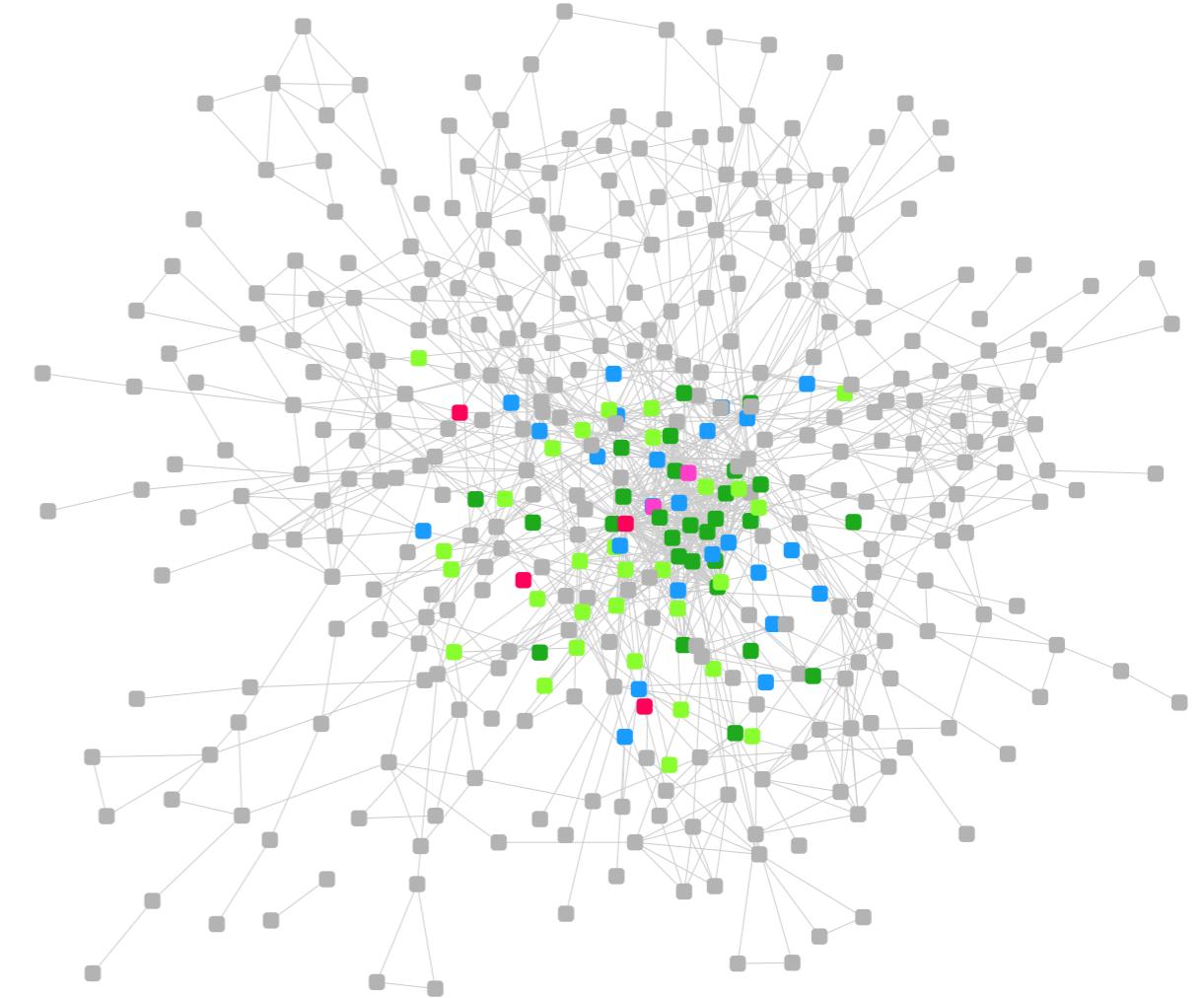


Interactions: Manual Layout

Step 1



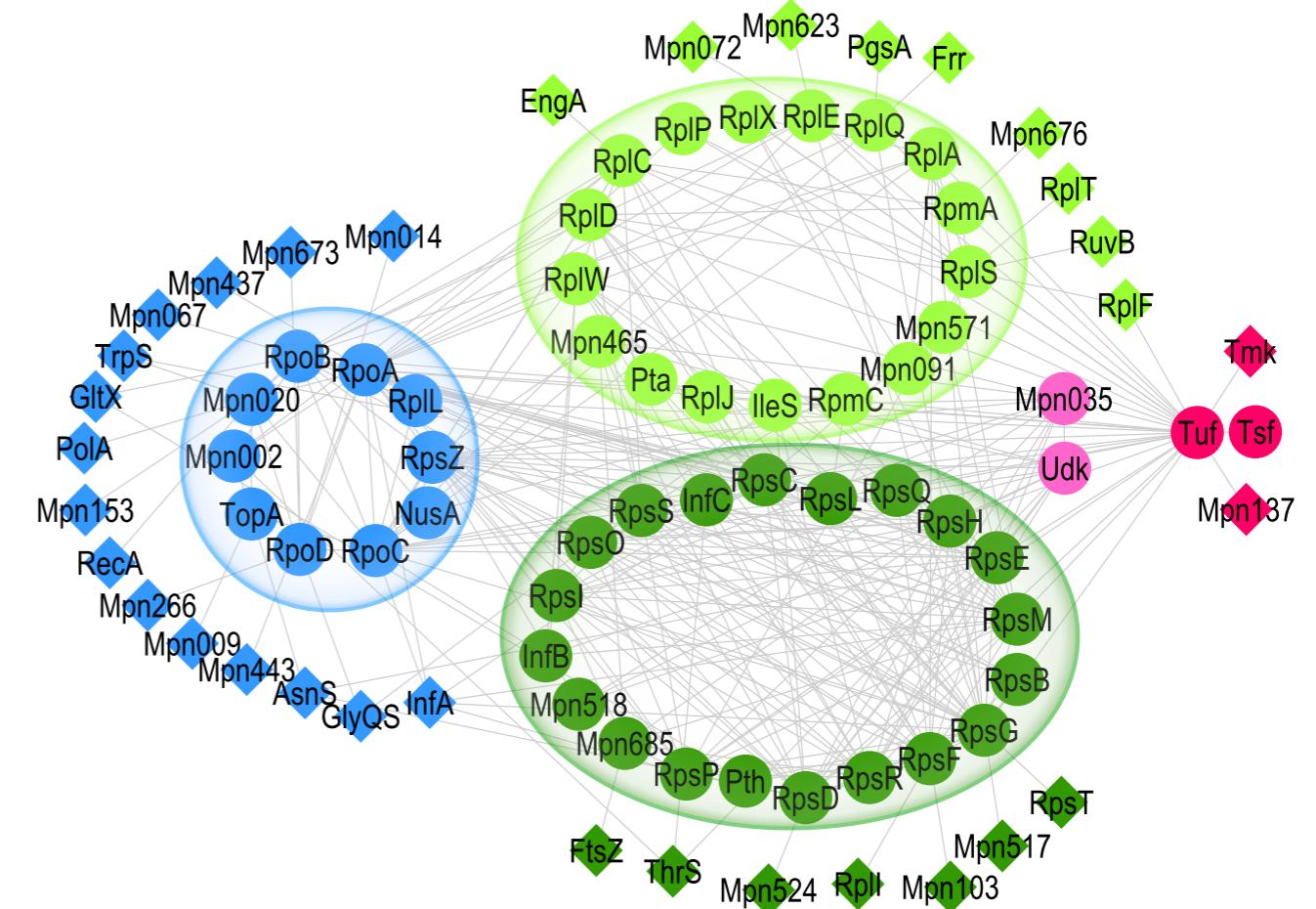
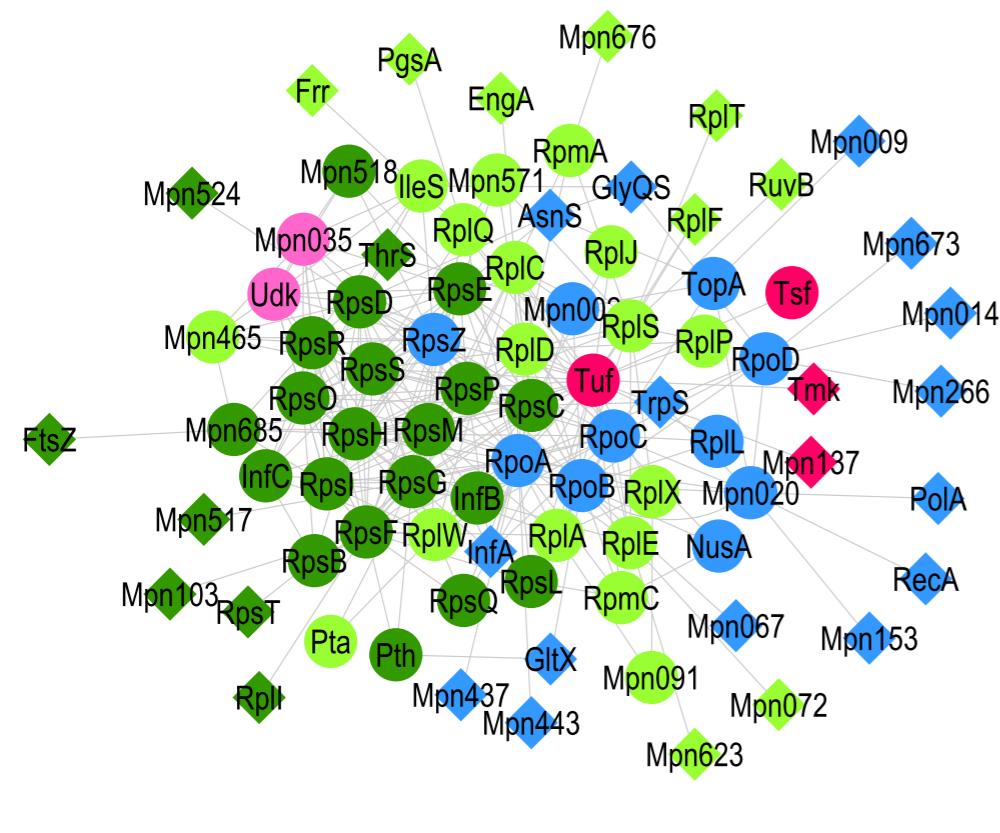
Step 2



Interactions: Manual Layout

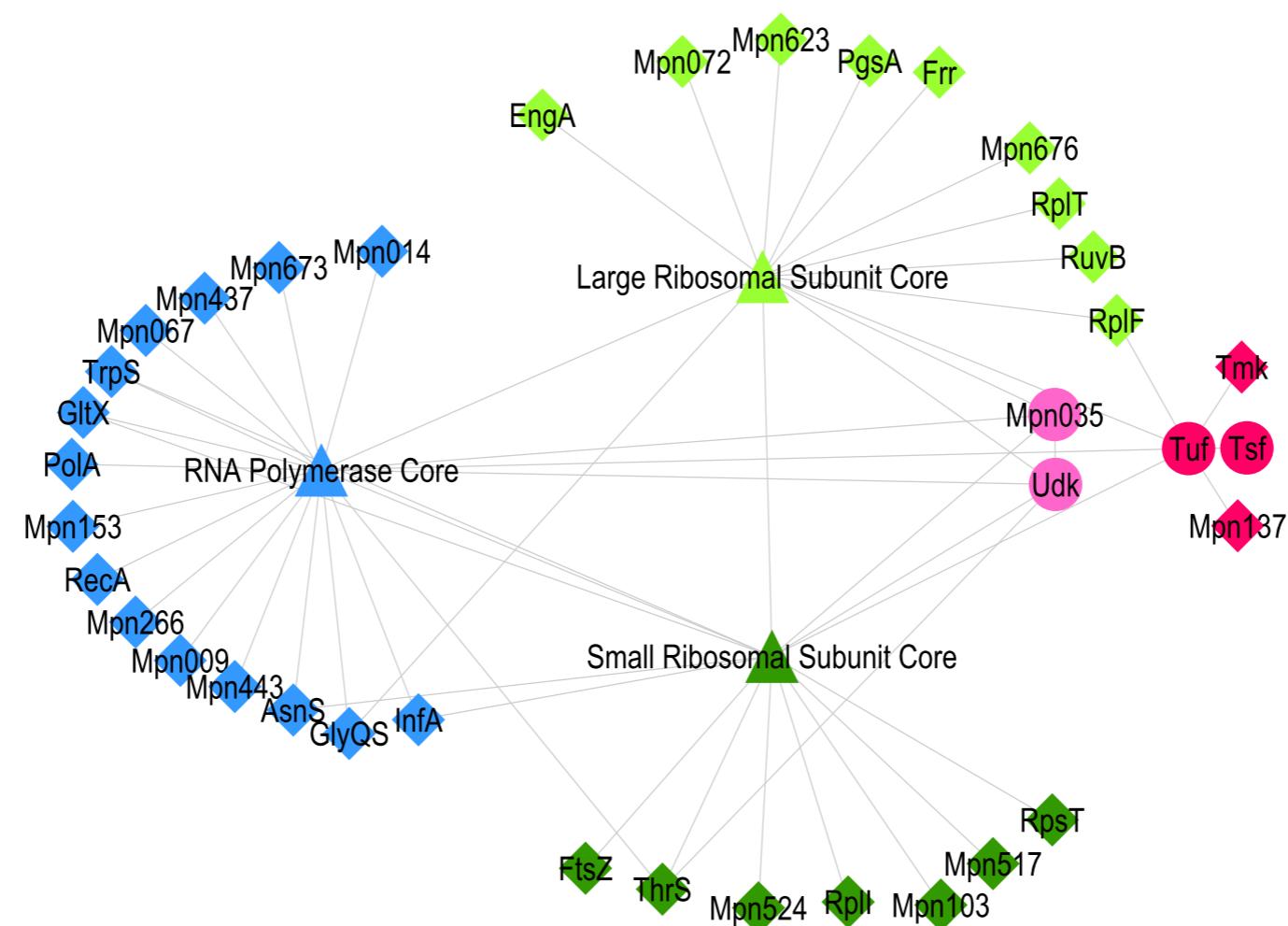
Step 3

Step 4

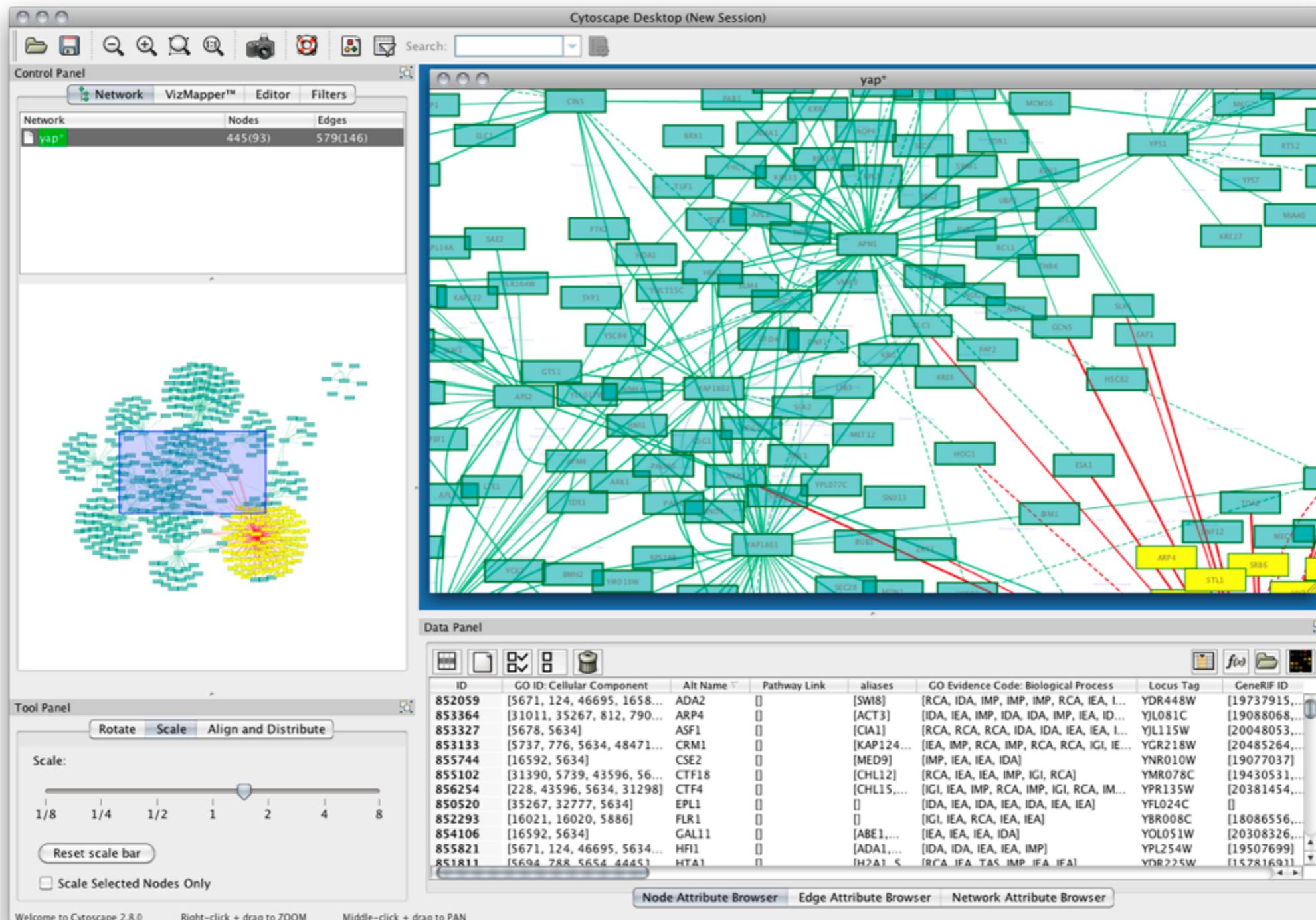


Interactions: Manual Layout

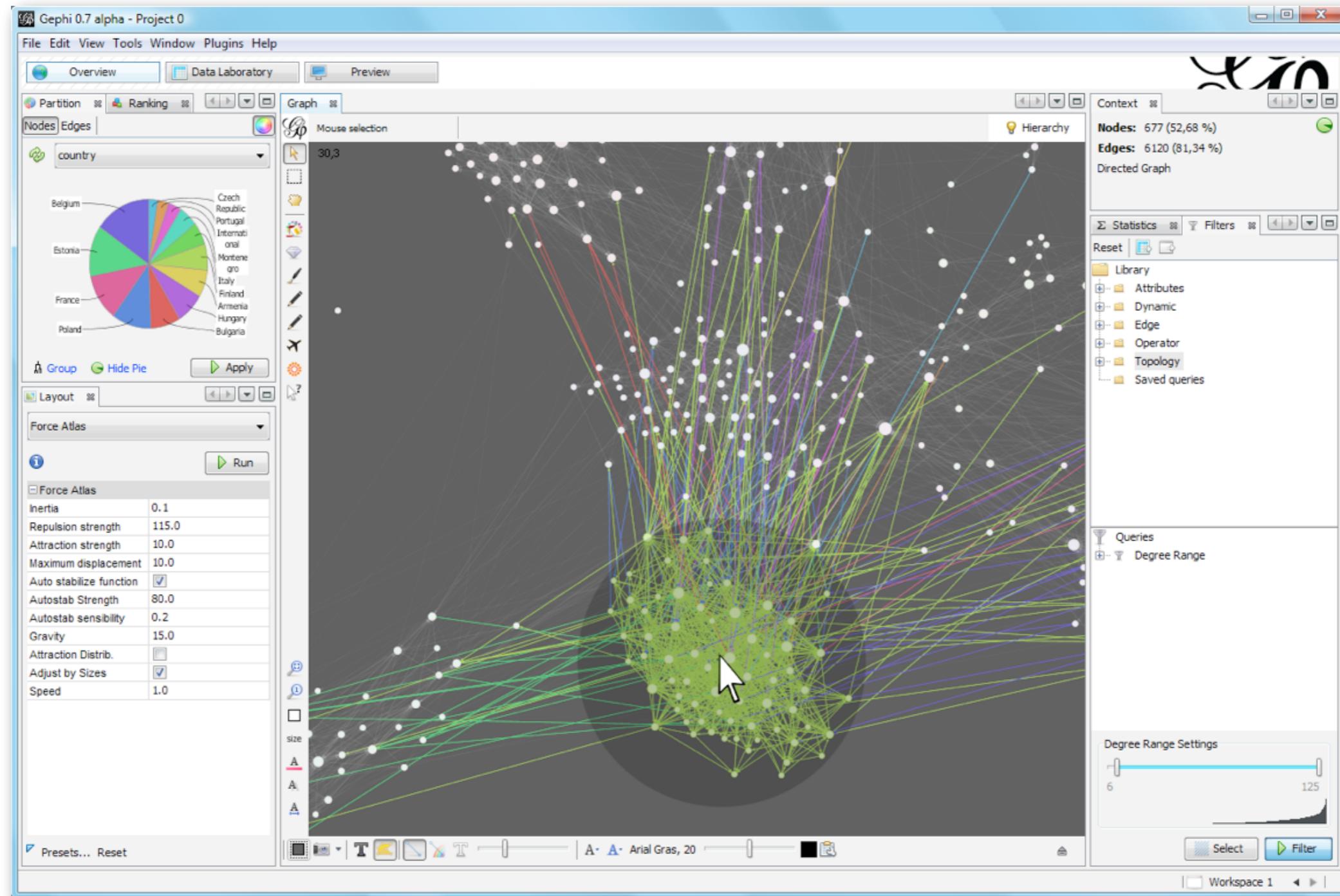
Step 5



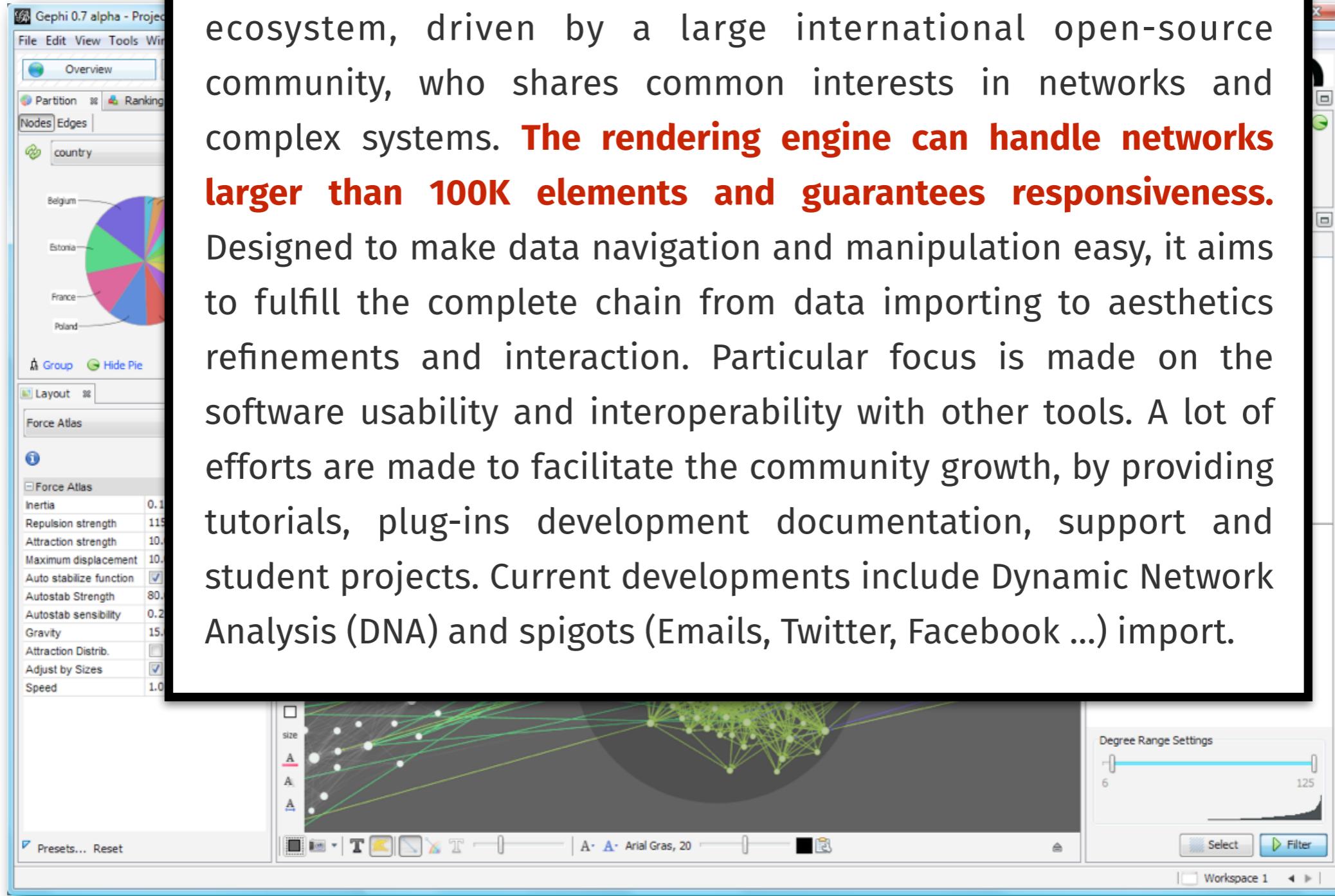
Cytoscape



Gephi



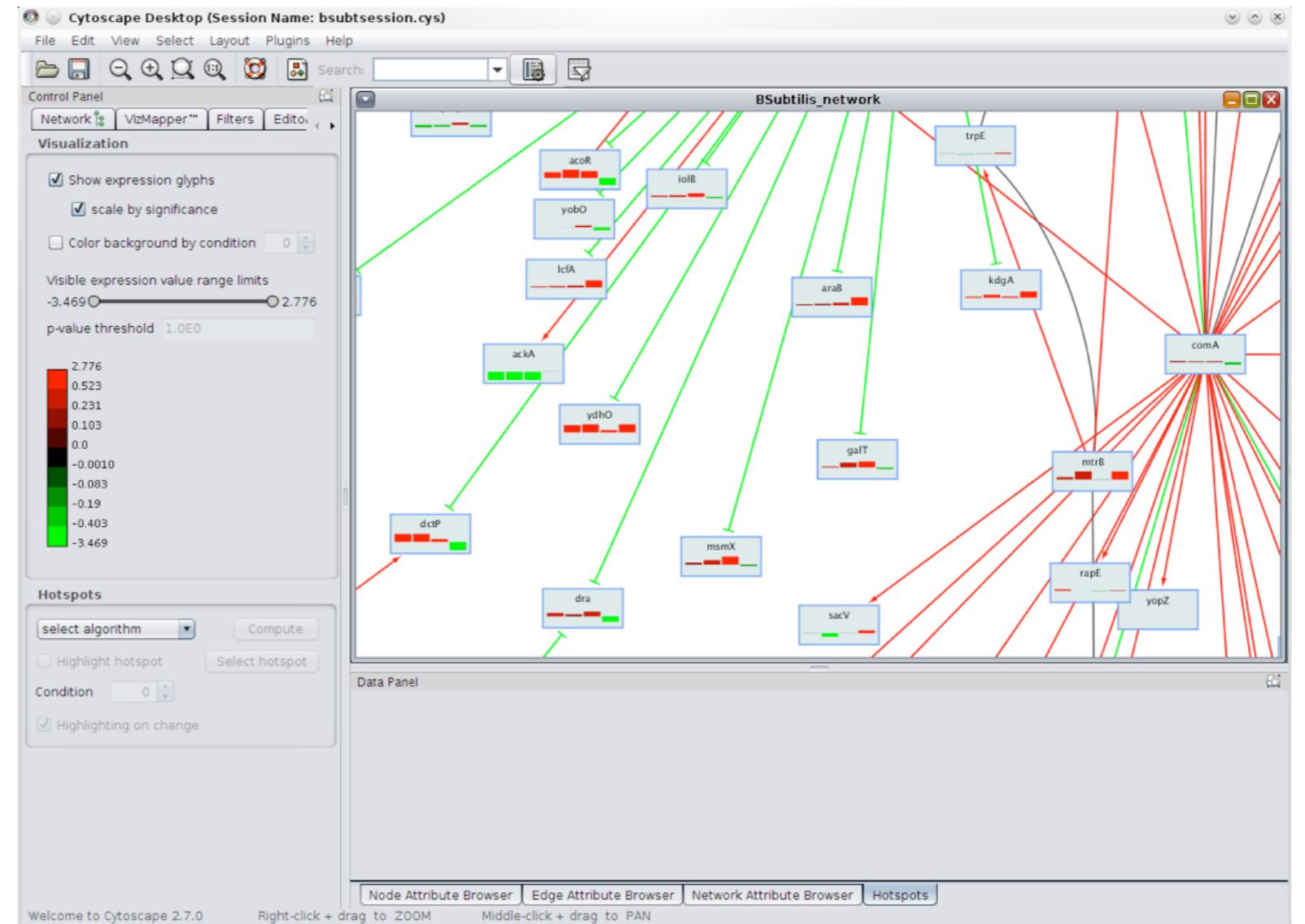
Gephi



Interactions: Gene Regulatory Networks

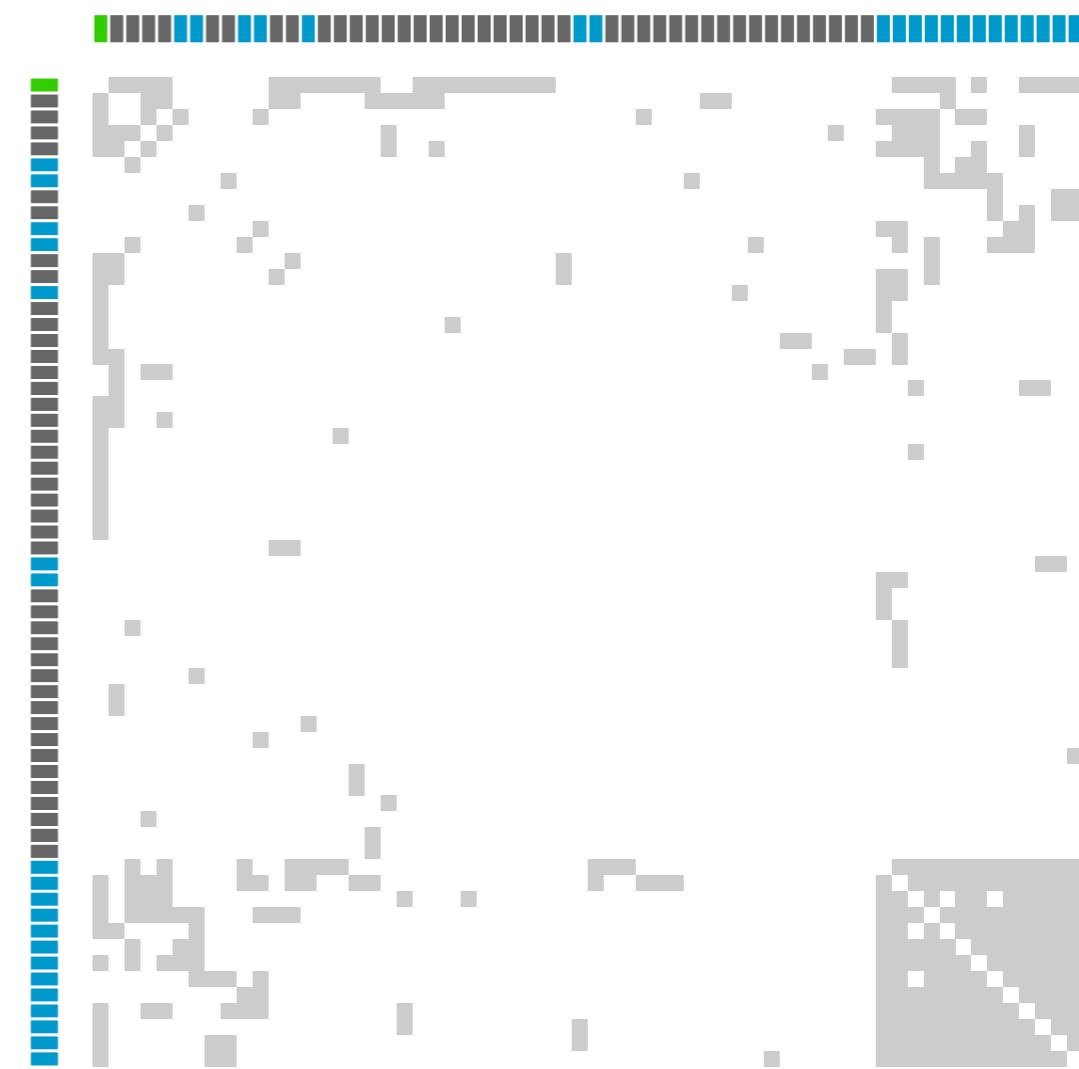
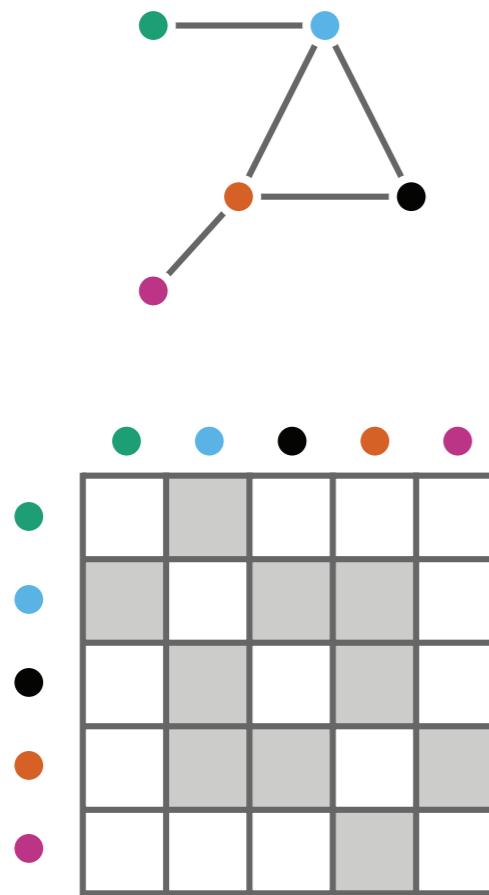
SpotXplore

- maps expression profiles onto regulatory network
- statistics can be visualized
- interaction
- highlight subnetworks
- Cytoscape plugin



Adjacency Matrix Approaches

Adjacency Matrix



Adjacency Matrix

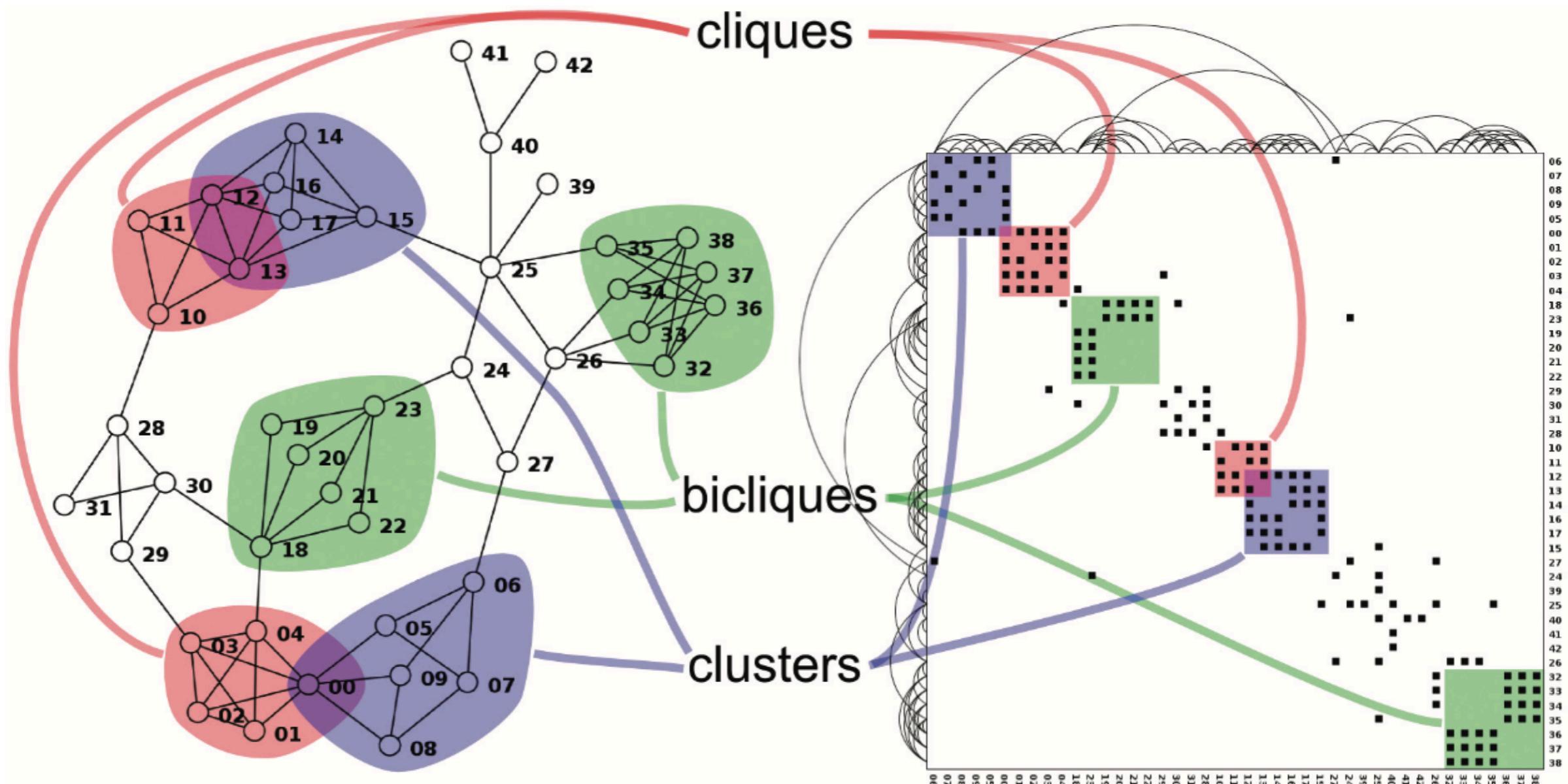
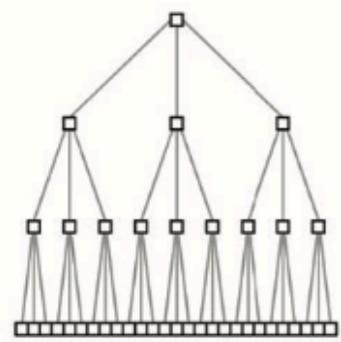
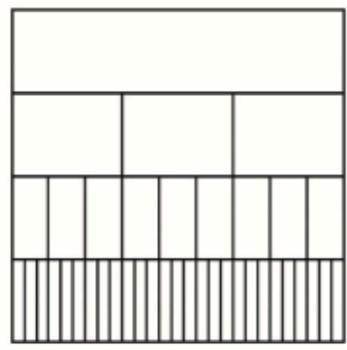


Figure 9.7. Characteristic patterns in matrix views and node-link views: both can show cliques and clusters clearly. From [McGuffin 12, Figure 6].

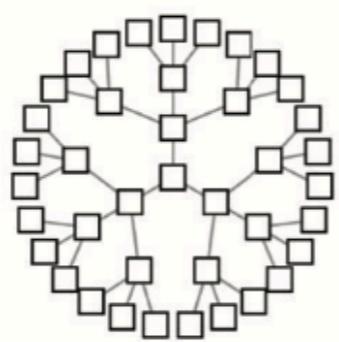
Nested Representations



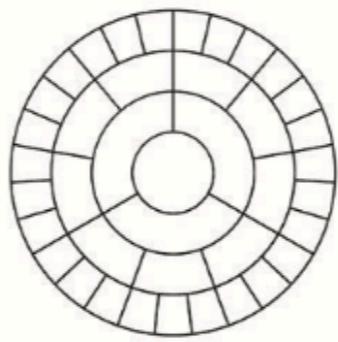
(a)



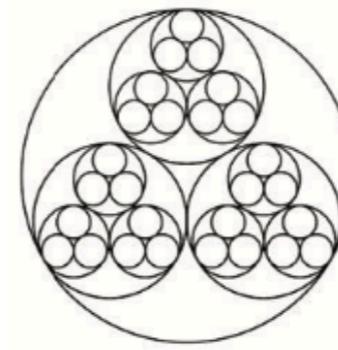
(b)



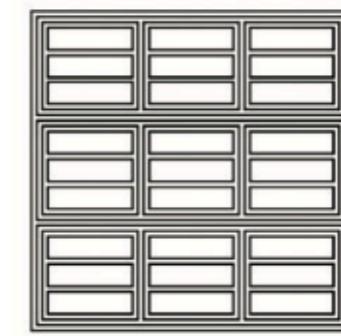
(c)



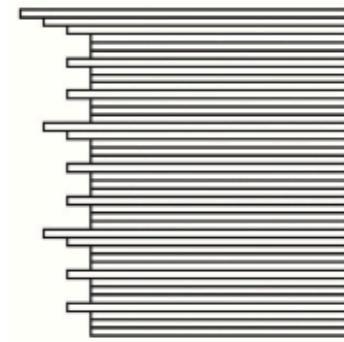
(d)



(e)



(f)



(g)

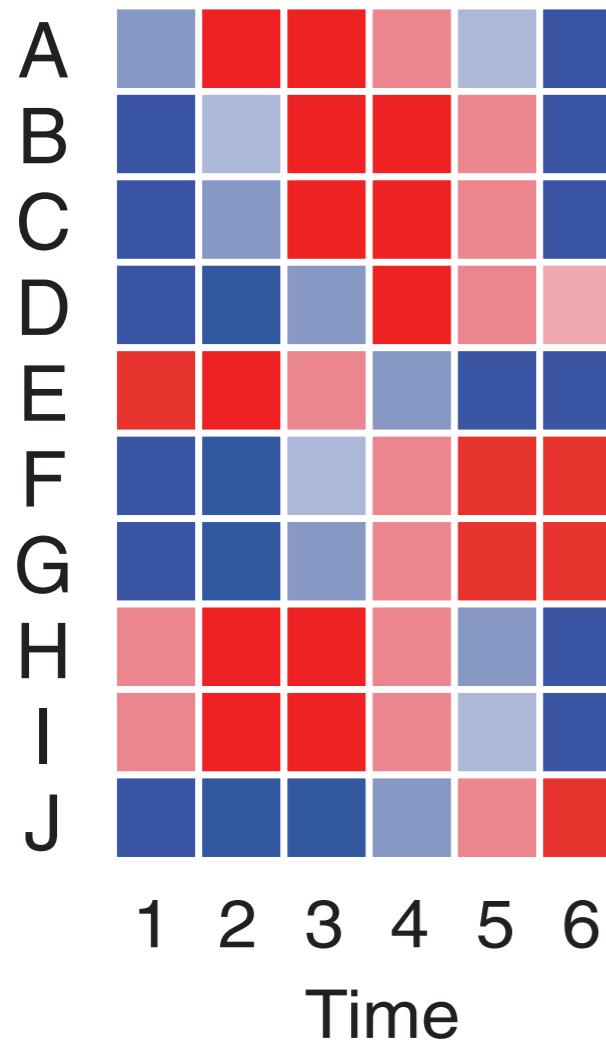


i Use mouse wheel to zoom in and out. Drag zoomed map to pan it. Double-click a ticker to display detailed information in a new window. Hover mouse cursor over a ticker to see its main competitors in a stacked view with a 3-month history graph.

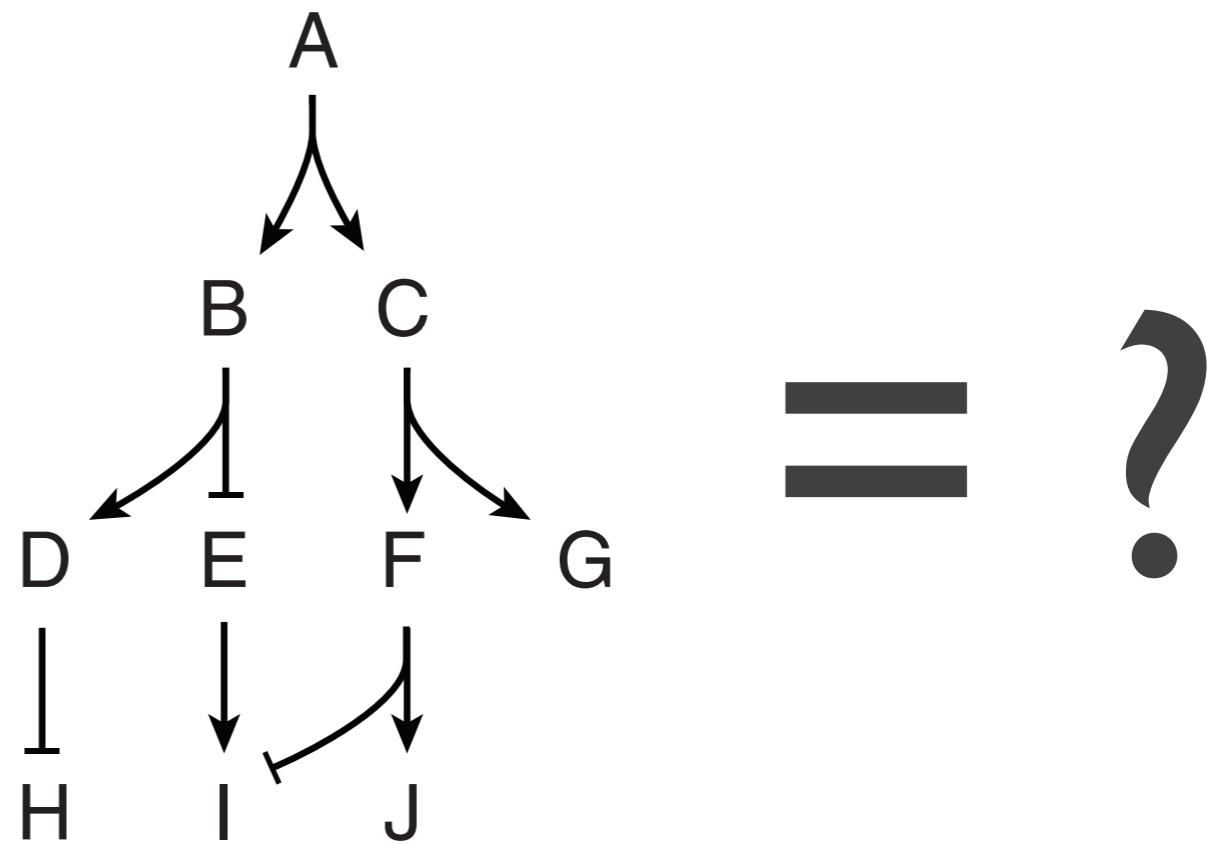
-10%	-6.7%	-3.3%	0%	+3.3%	+6.7%	+10%
------	-------	-------	----	-------	-------	------

Graphs and Multivariate Data

Interactions: And Multivariate Data?



+

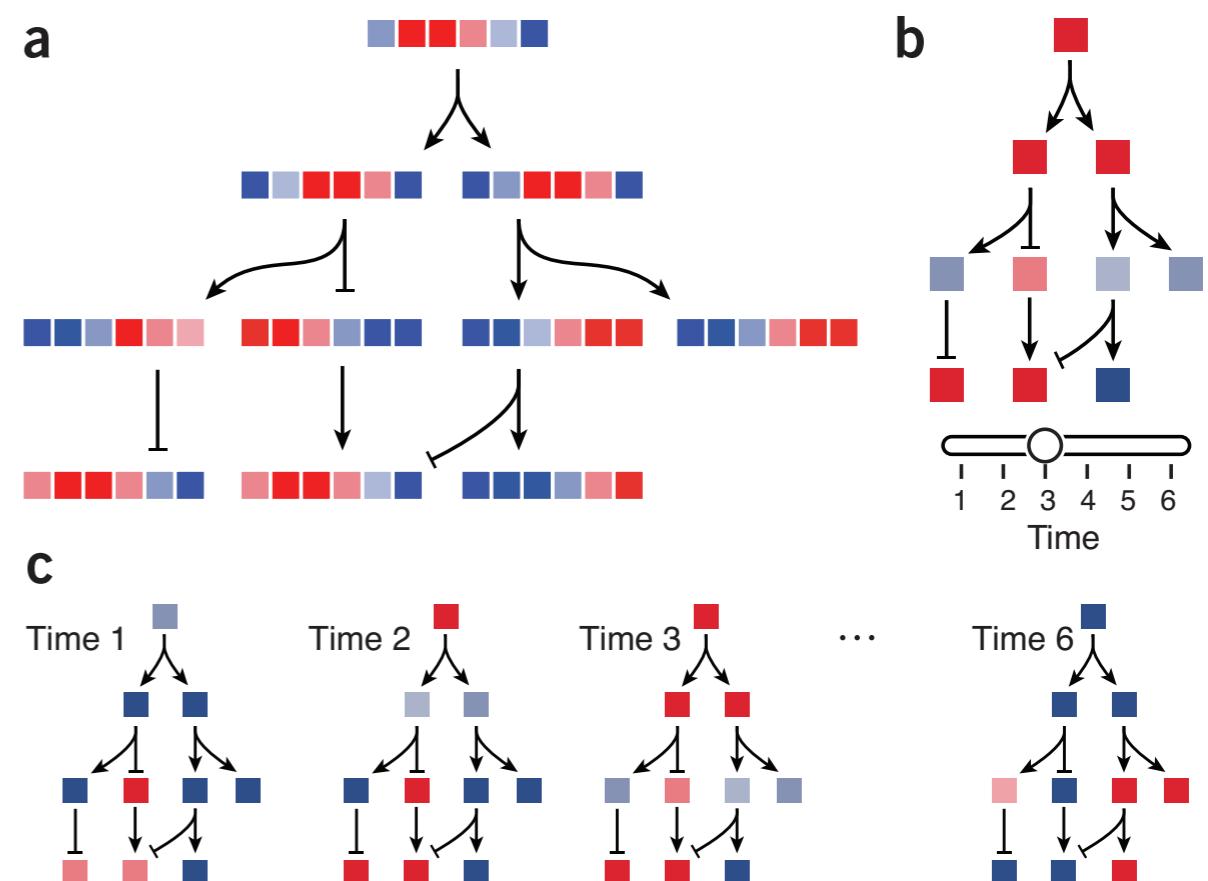


=

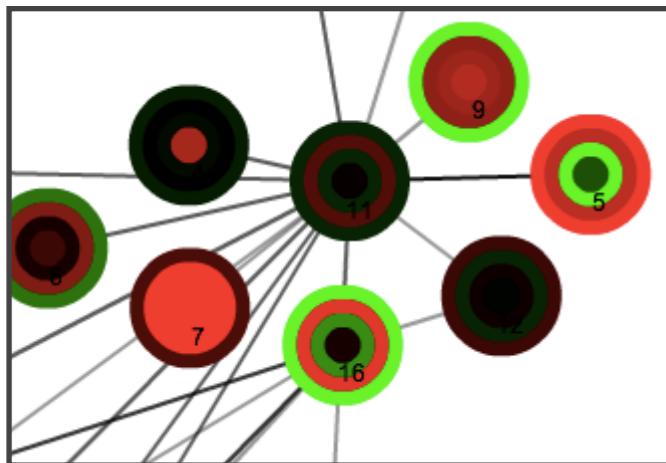
?

Interactions: And Multivariate Data!?

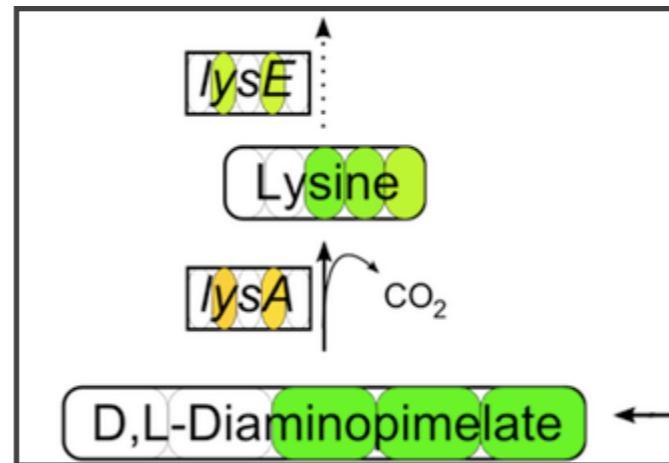
- (a) **Complex glyphs** = multiple values per node?
- (b) **Animation** = one value per node, one network shown at a time?
- (c) **Small multiples** = one value per node, all networks shown simultaneously?



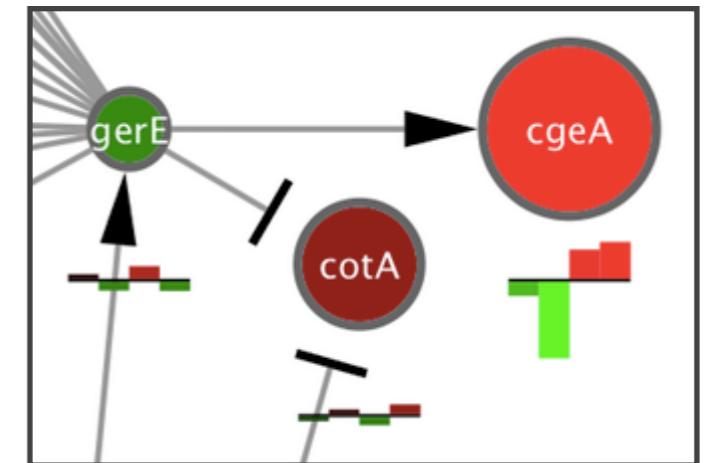
Interactions: And Multivariate Data!



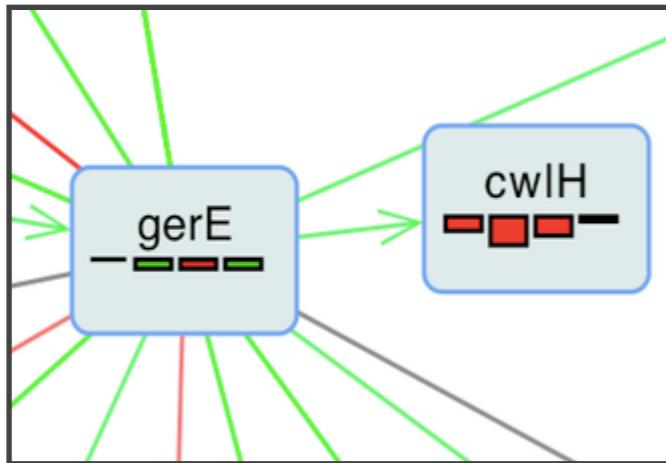
Lichen



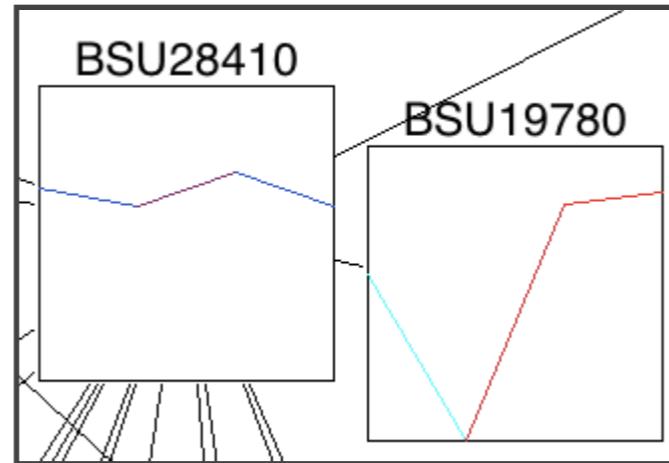
Prometra



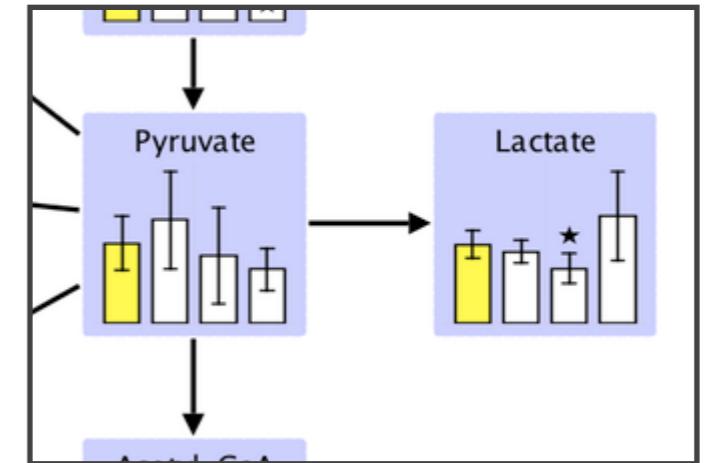
VistaClara (Cytoscape)



GENeVis

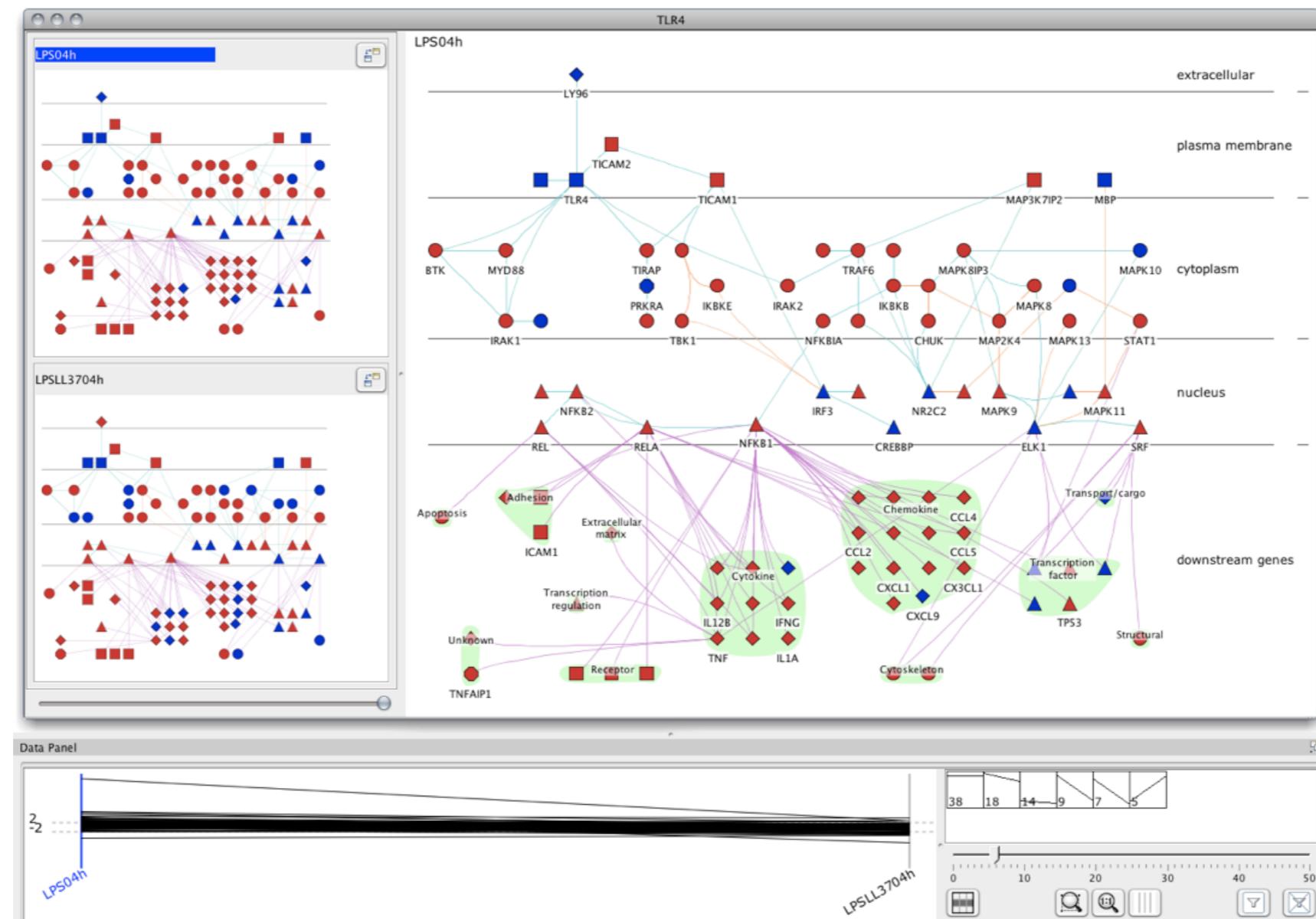


VisANT



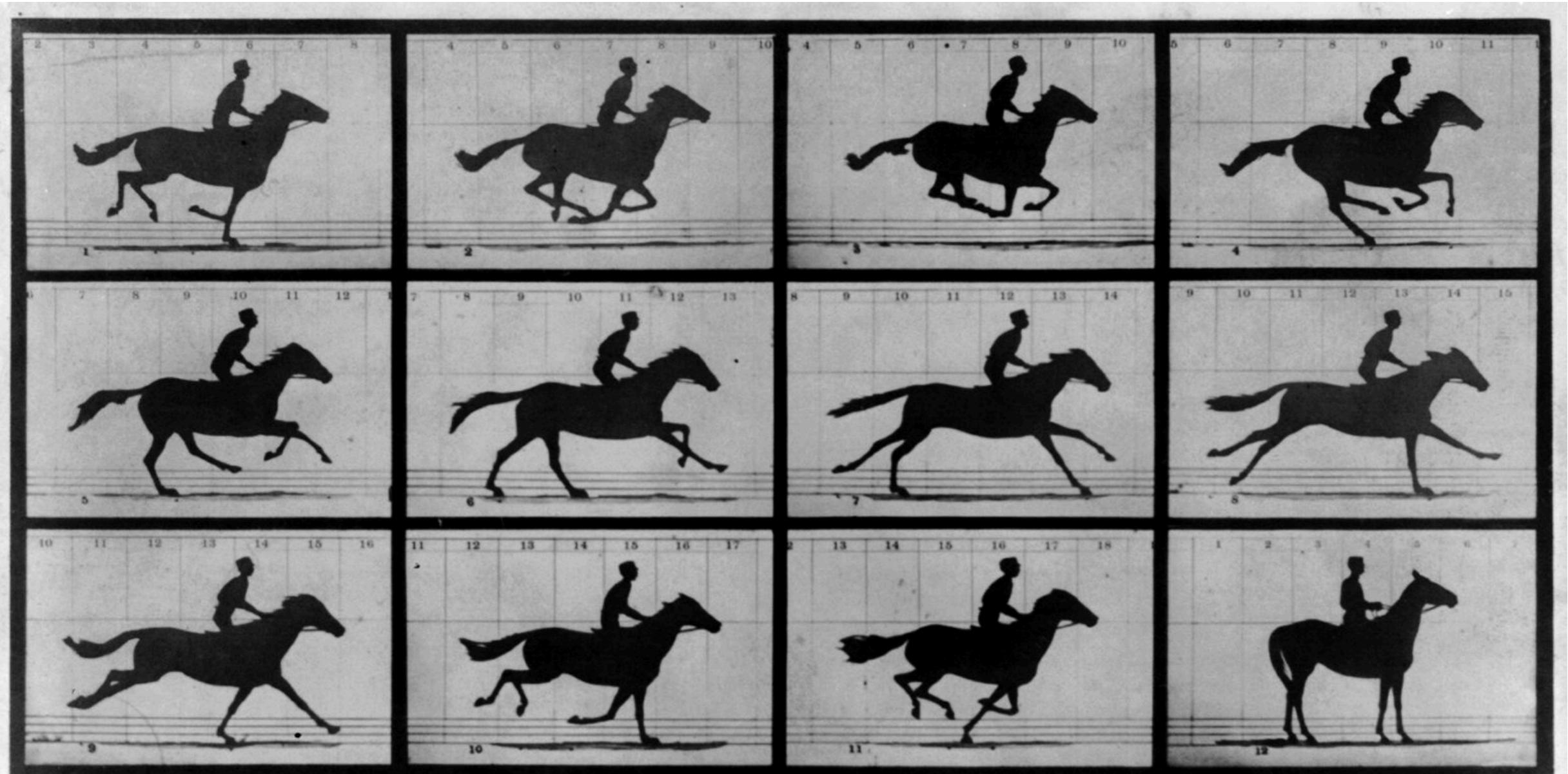
VANTED

Interactions: And Multivariate Data!



Cerebral (Cytoscape plugin)

Small Multiples



Copyright, 1878, by MUYBRIDGE.

MORSE'S Gallery, 417 Montgomery St., San Francisco.

THE HORSE IN MOTION.

Illustrated by
MUYBRIDGE.

AUTOMATIC ELECTRO-PHOTOGRAPH.

"SALLIE GARDNER," owned by LELAND STANFORD; running at a 1.40 gait over the Palo Alto track, 19th June, 1878.

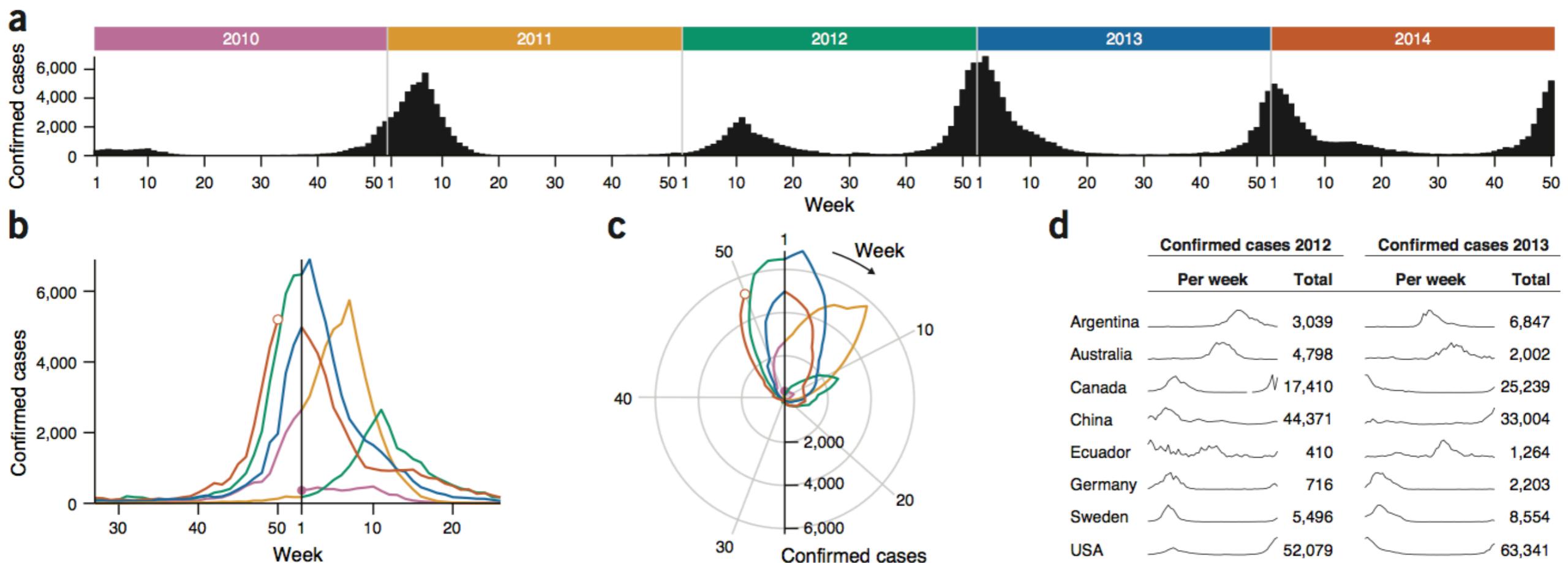
The negatives of these photographs were made at intervals of twenty-seven inches of distance, and about the twenty-fifth part of a second of time; they illustrate consecutive positions assumed in each twenty-seven inches of progress during a single stride of the mare. The vertical lines were twenty-seven inches apart; the horizontal lines represent elevations of four inches each. The exposure of each negative was less than the two-thousandth part of a second.

Small Multiples

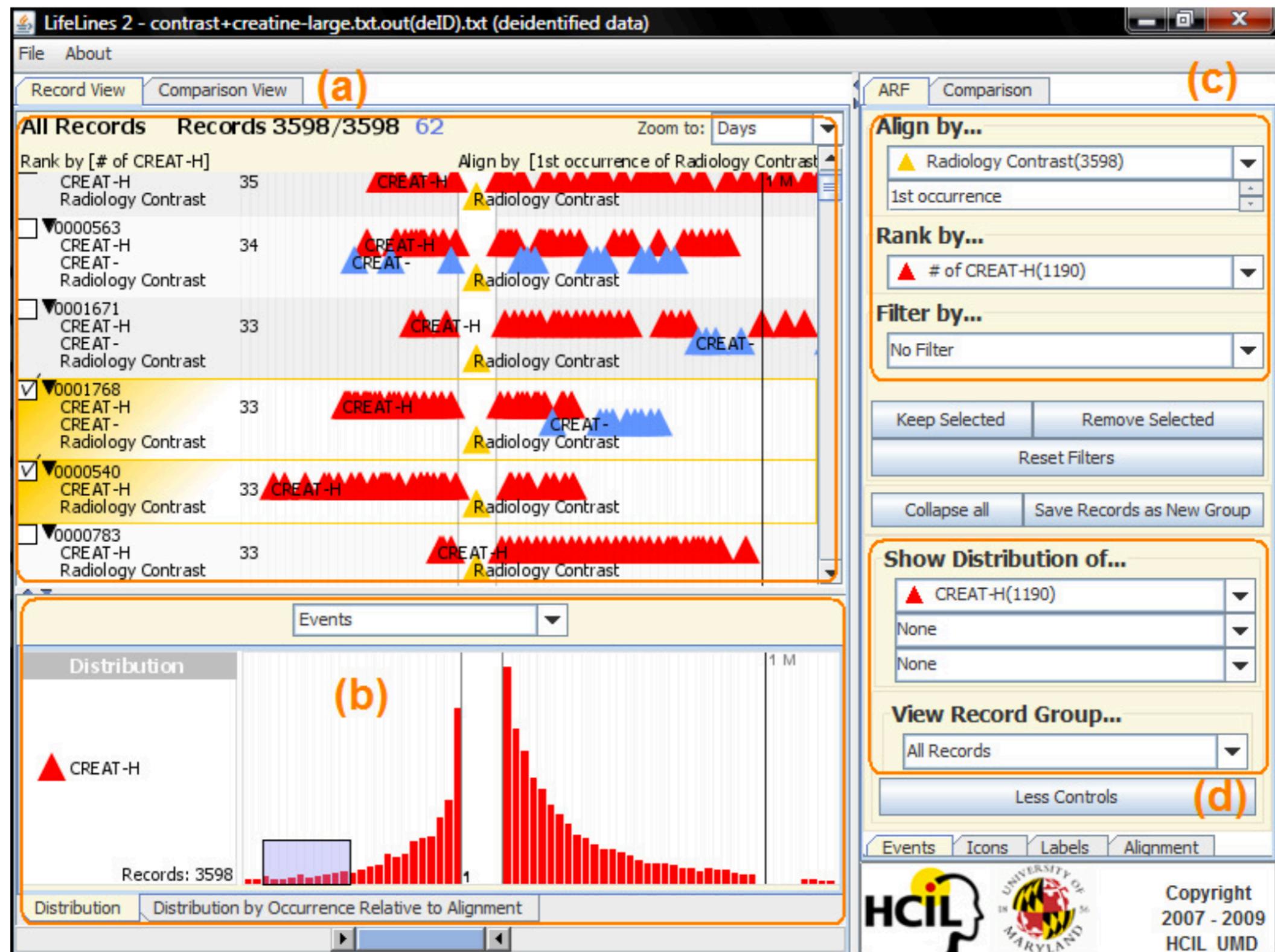


Temporal Data

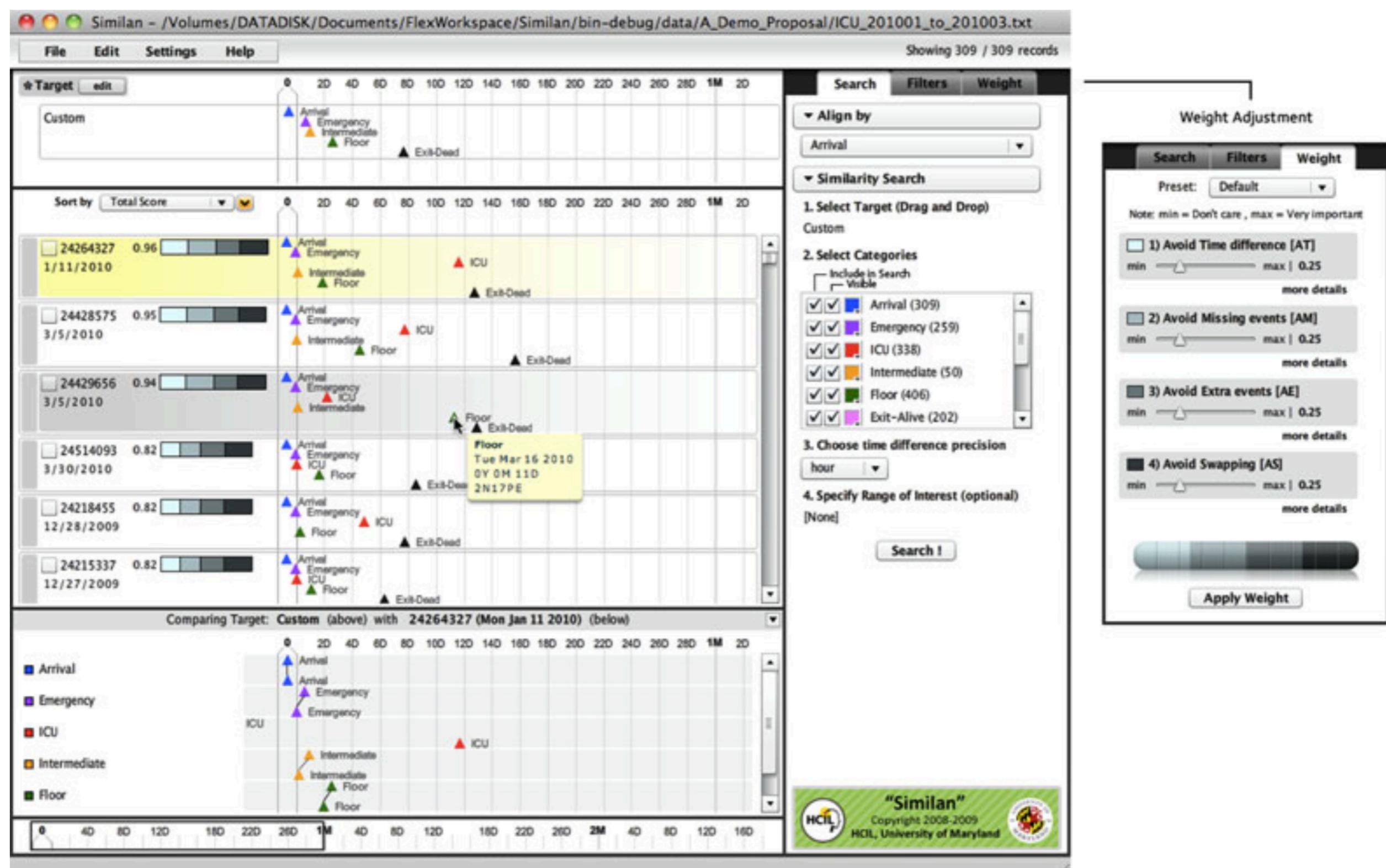
Temporal Data



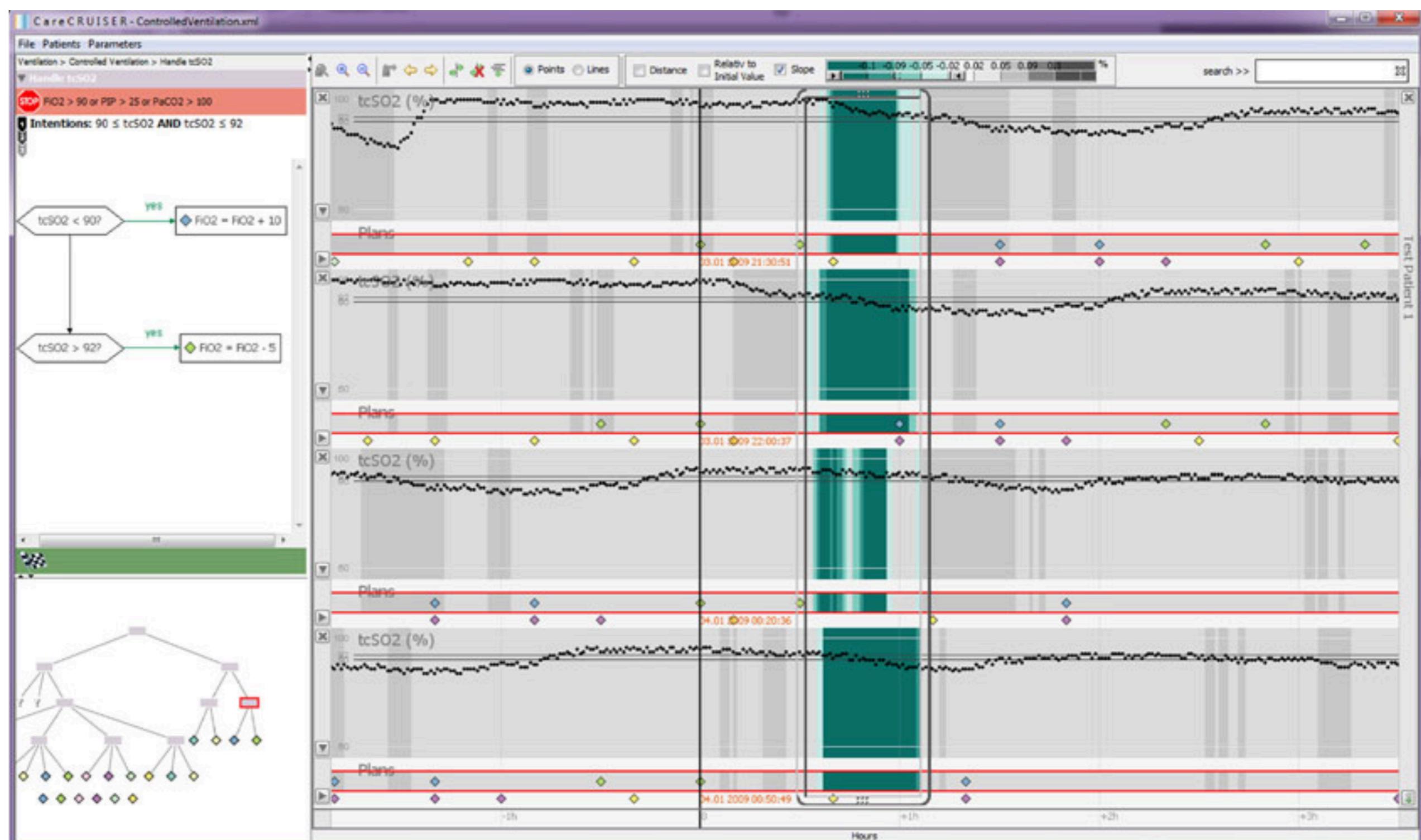
LifeLines 2



Similan



CareCruiser



Temporal Data: Properties

- Scale
- Scope
- Arrangement
- Viewpoint

Scales

Fig. 3.1 Ordinal scale. Only relative order relations are present. At this level it is not possible to discern whether Valentina woke up before or after Arvid arrived.

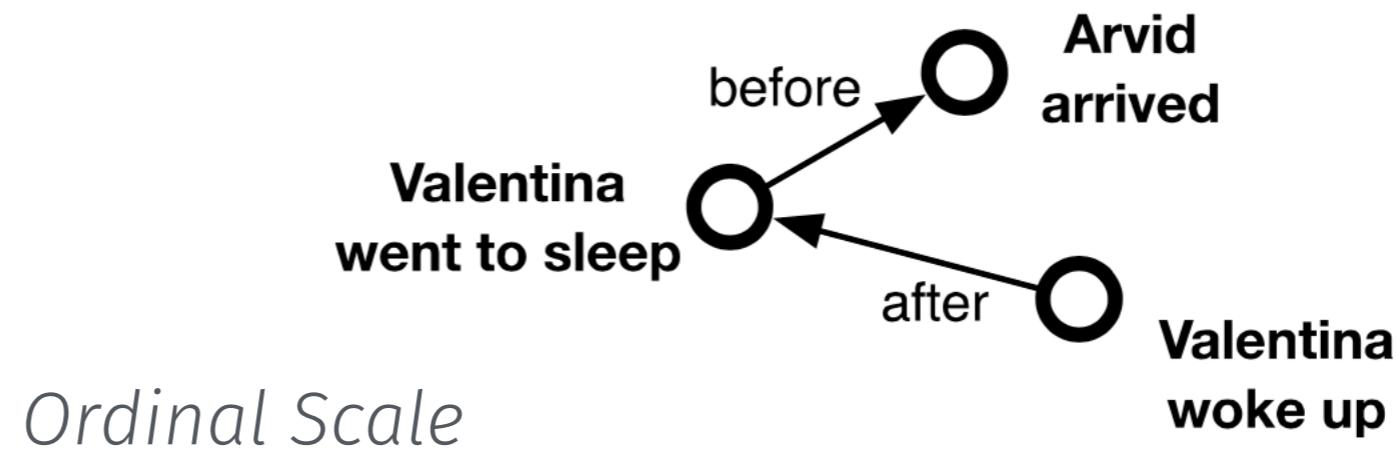


Fig. 3.2 Discrete scale. Smallest possible unit is minutes. Although Arvid arrived and Valentina woke up within the same minute, it is not possible to model the exact order of events.

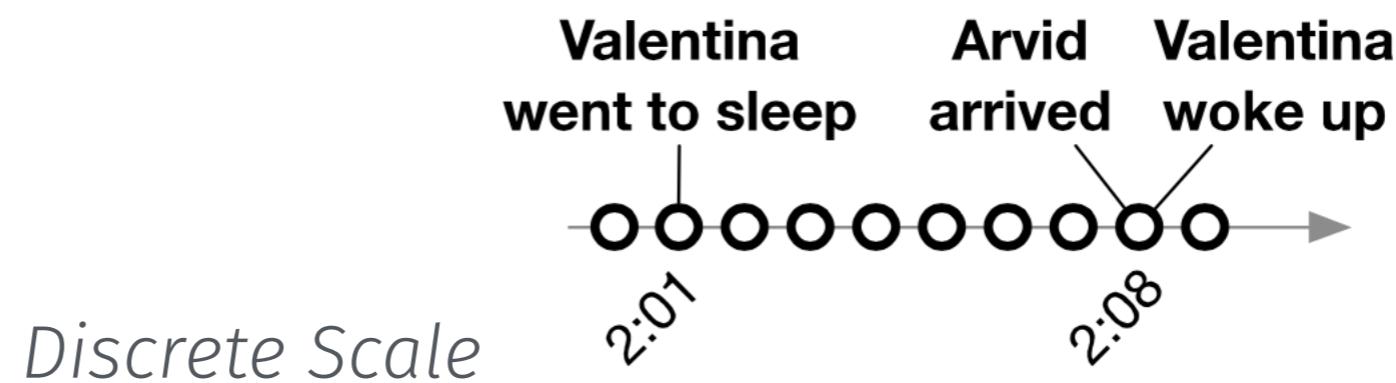
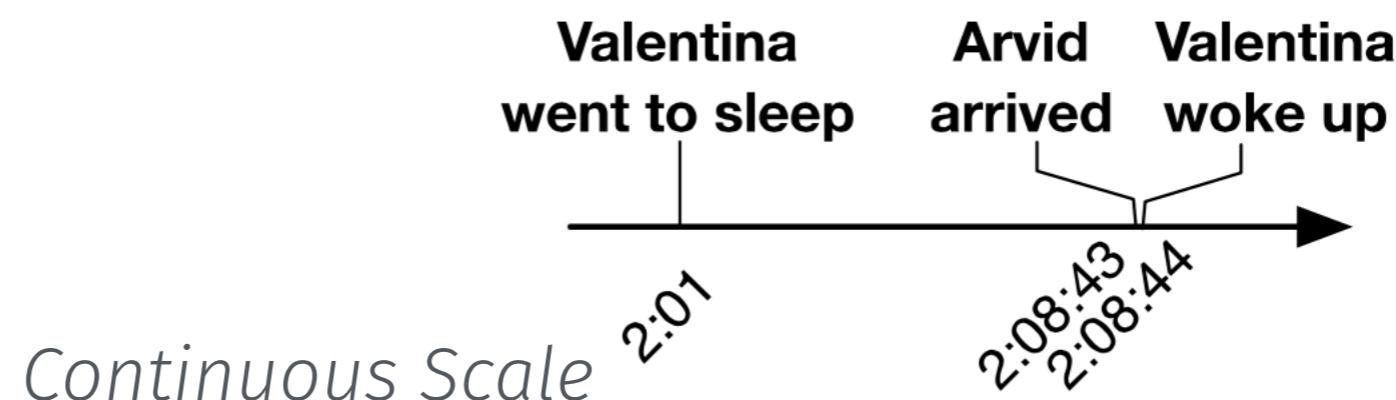


Fig. 3.3 Continuous scale. Between any two points in time, another point in time exists. Here, it is possible to model that Arvid arrived shortly before Valentina woke up.



Scope: Point vs Interval

Fig. 3.7 Time value “August 1, 2008” in a point-based domain. No information is given in between two time points.

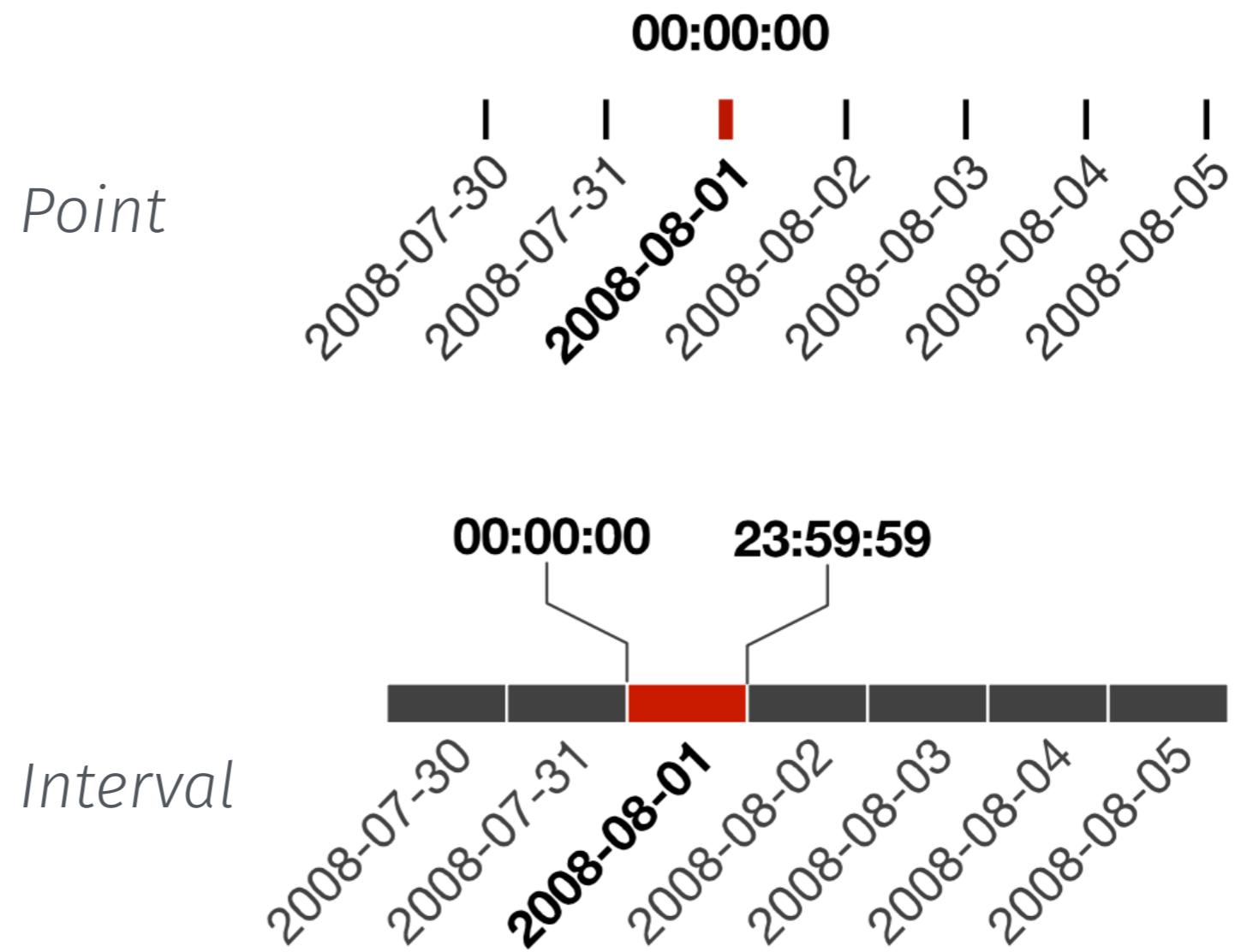


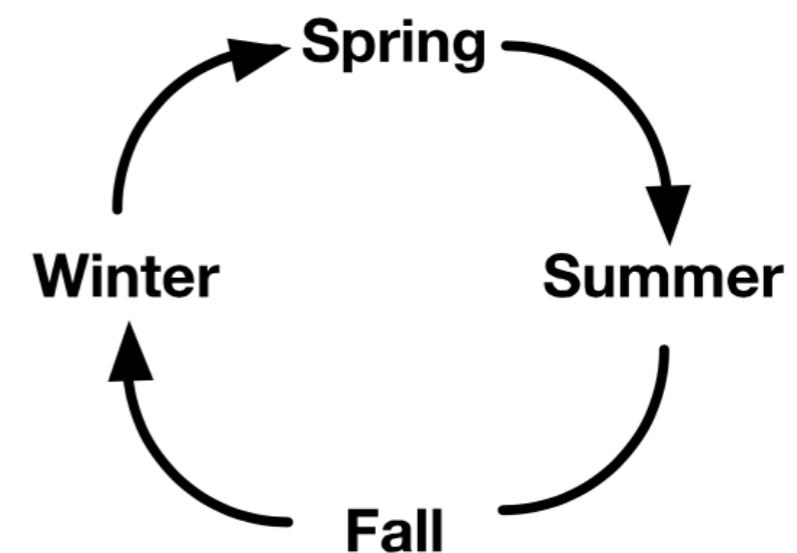
Fig. 3.8 Time value “August 1, 2008” in an interval-based domain. Each element covers a subsection of the time domain greater than zero.

Arrangement: Linear vs Cyclic

Fig. 3.10 Linear time. Time proceeds linearly from past to future.



Fig. 3.11 Cyclic time. Set of recurring time values such as the seasons of the year.

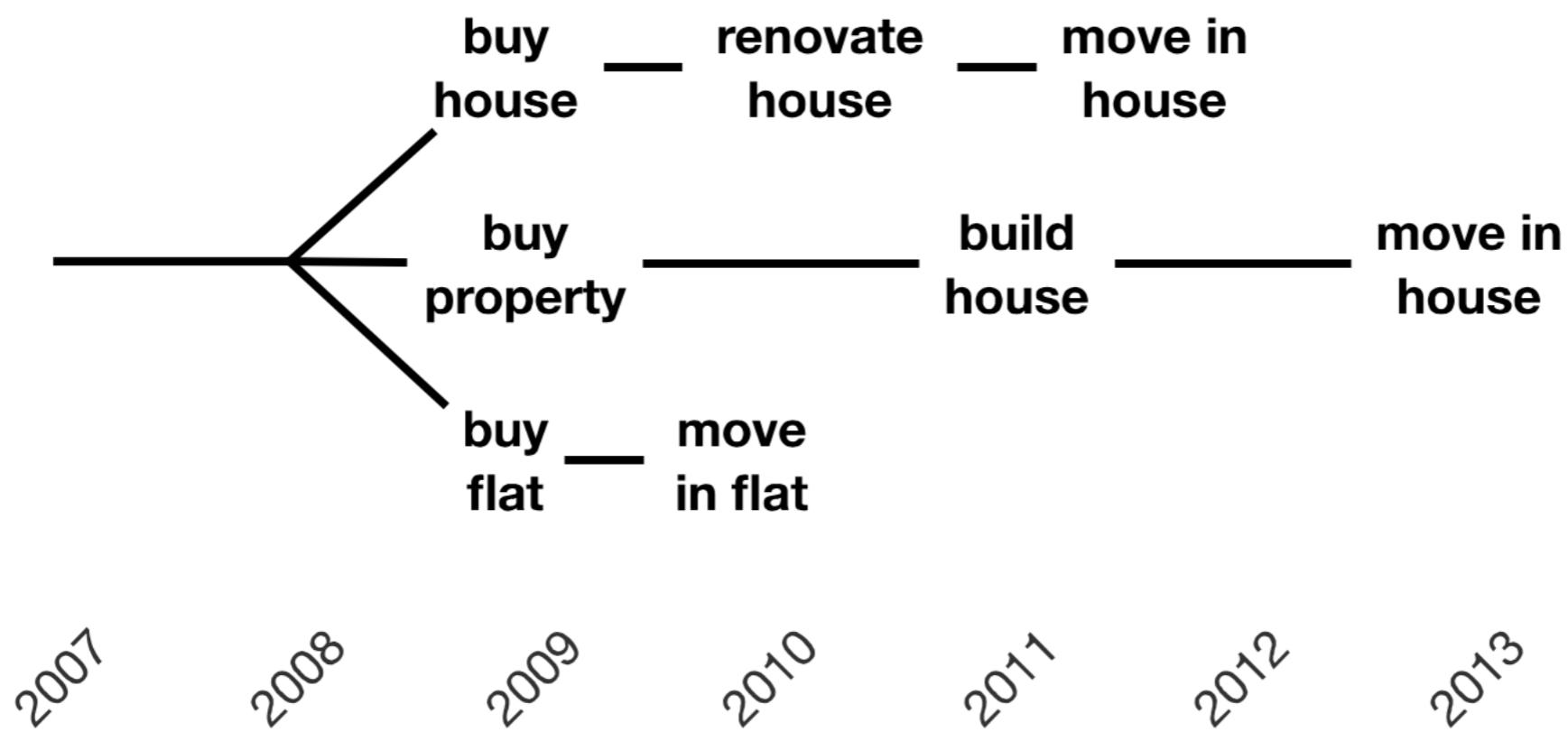


Viewpoint: Ordered

totally ordered: only one event can happen at a time

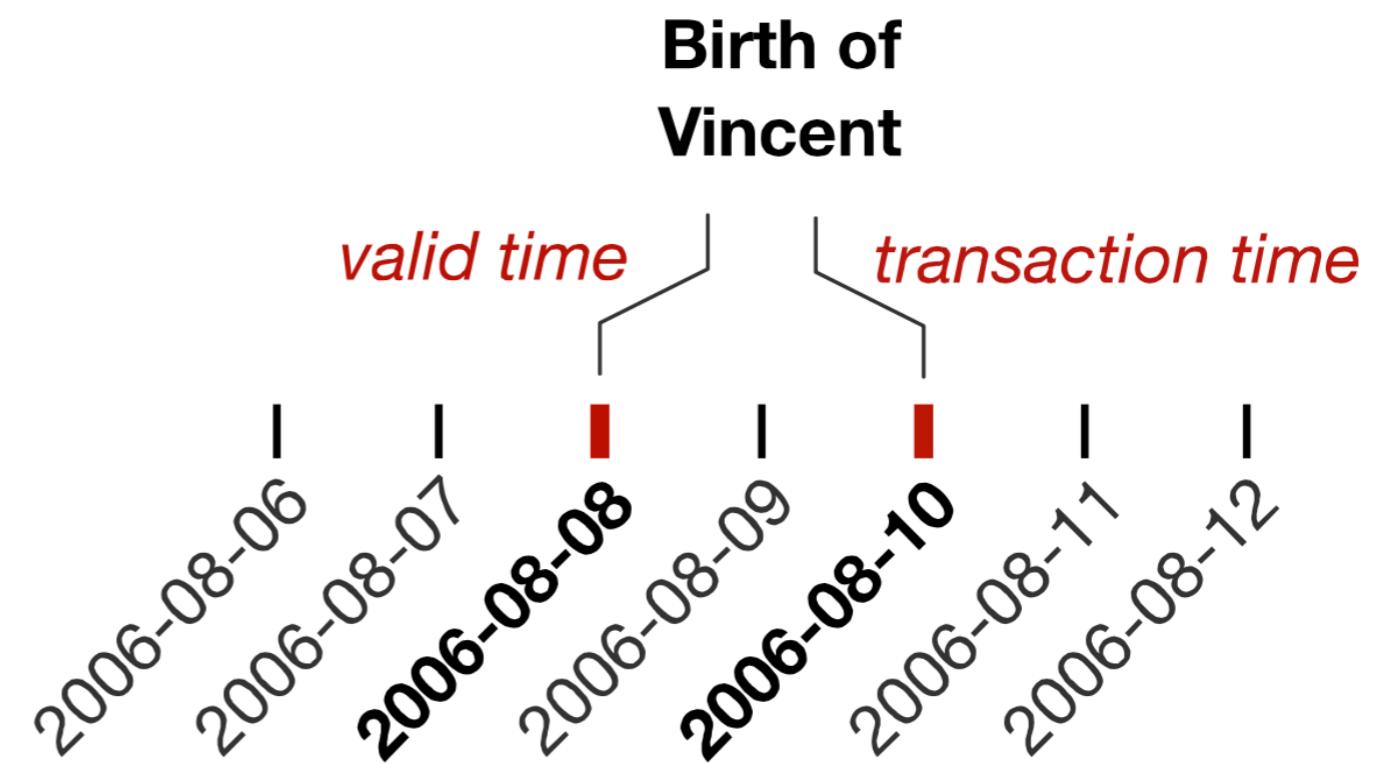
partially ordered: events can overlap

Viewpoint: Branching

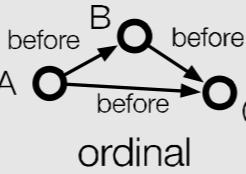
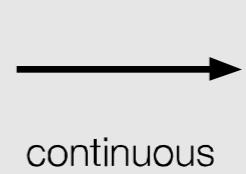
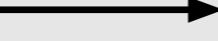
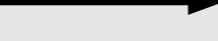
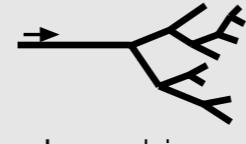
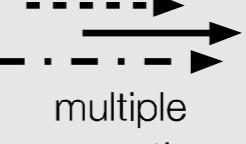


Viewpoint: Multiple Perspectives

Fig. 3.13 Multiple perspectives. Vincent was born on August 8, 2006 (valid time) and this fact was stored in the register of residents two days later on August 10, 2006 (transaction time).



Temporal Data: Properties

scale	 ordinal	 discrete	 continuous
scope	 point-based	 interval-based	
arrangement	 linear	 cyclic	
viewpoint	 ordered	 branching	 multiple perspectives
Abstractions			
granularity & calendars	 none	 single	 multiple
time primitives	 instant	 interval	 span
determinacy	 determinate	 indeterminate	

Temporal Data: Design Space

What?

time

scale, scope, arrangement, viewpoint
granularity & calendars, time primitives,
determinacy
see Chapter 3

data

scale, frame of reference, kind of data,
number of variables
see Chapter 3

time & data

internal time, external time
see Chapter 3

Why?

1st level



individual values



sets

2nd level



lookup



comparison

3rd level



identification



localization

How?

mapping



static



dynamic

dimensionality



2D

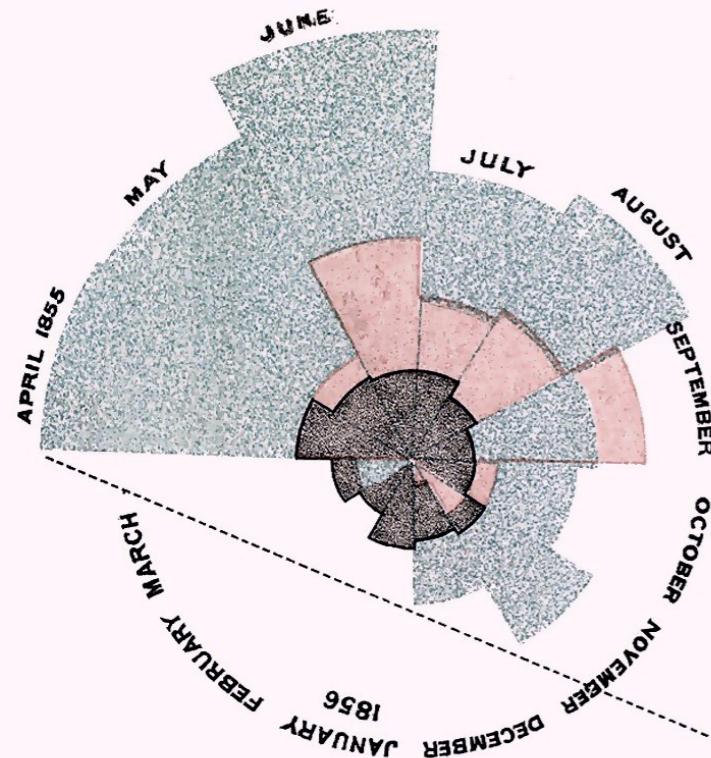


3D

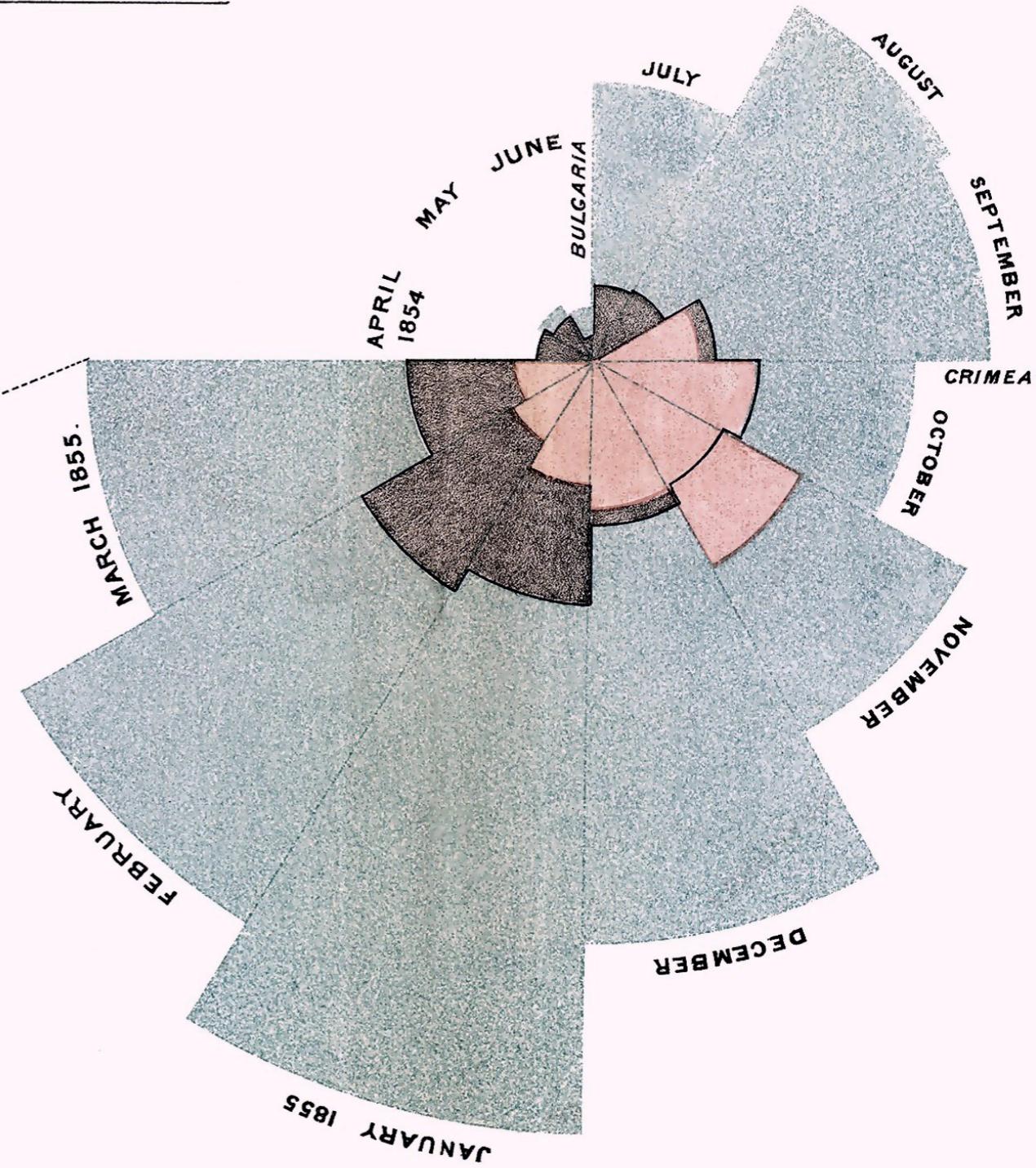
Nightingale: Rose Charts (1858)

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.



1.
APRIL 1854 TO MARCH 1855.



The areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

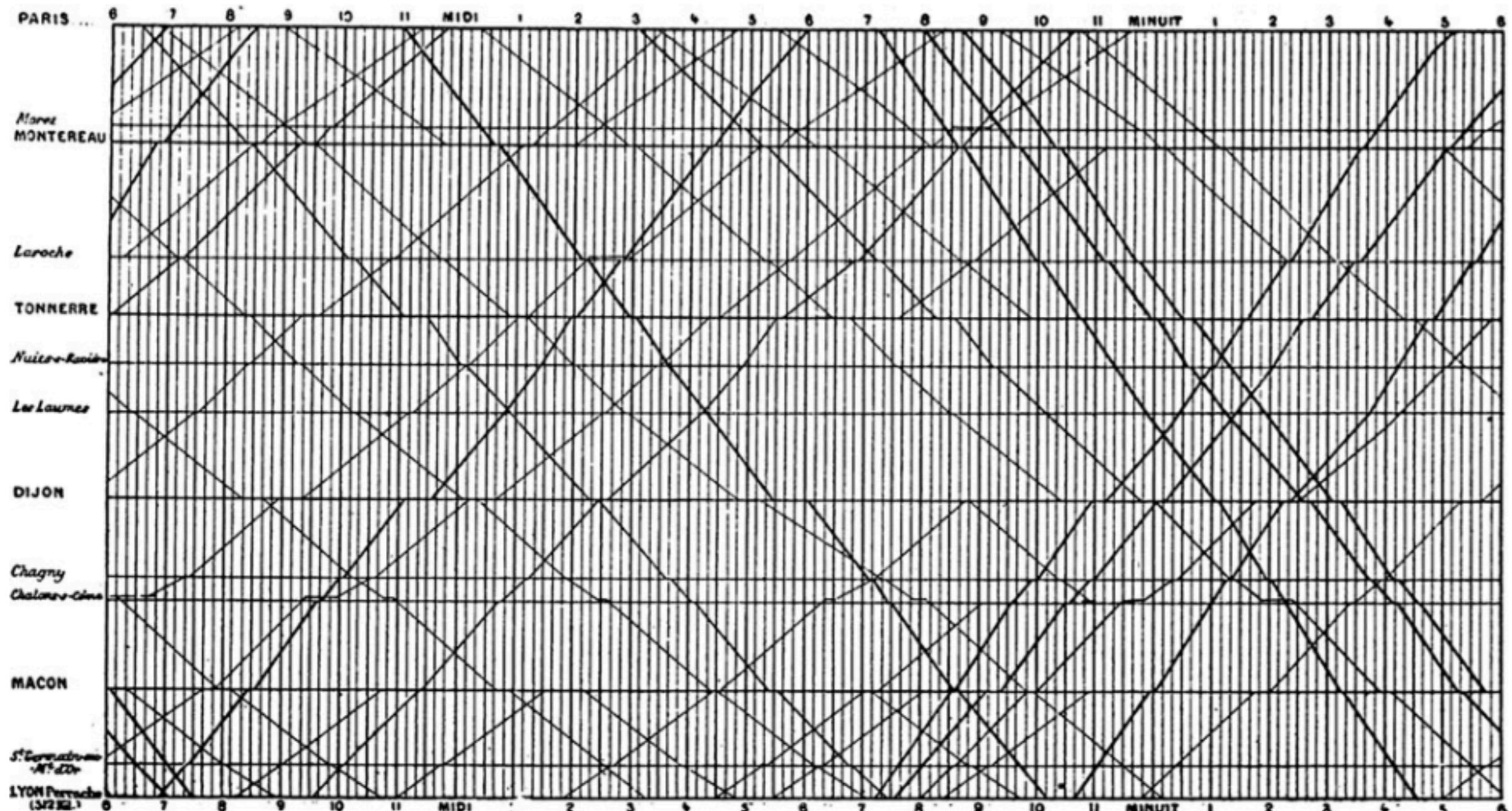
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.^r 1854 marks the boundary of the deaths from all other causes during the month.

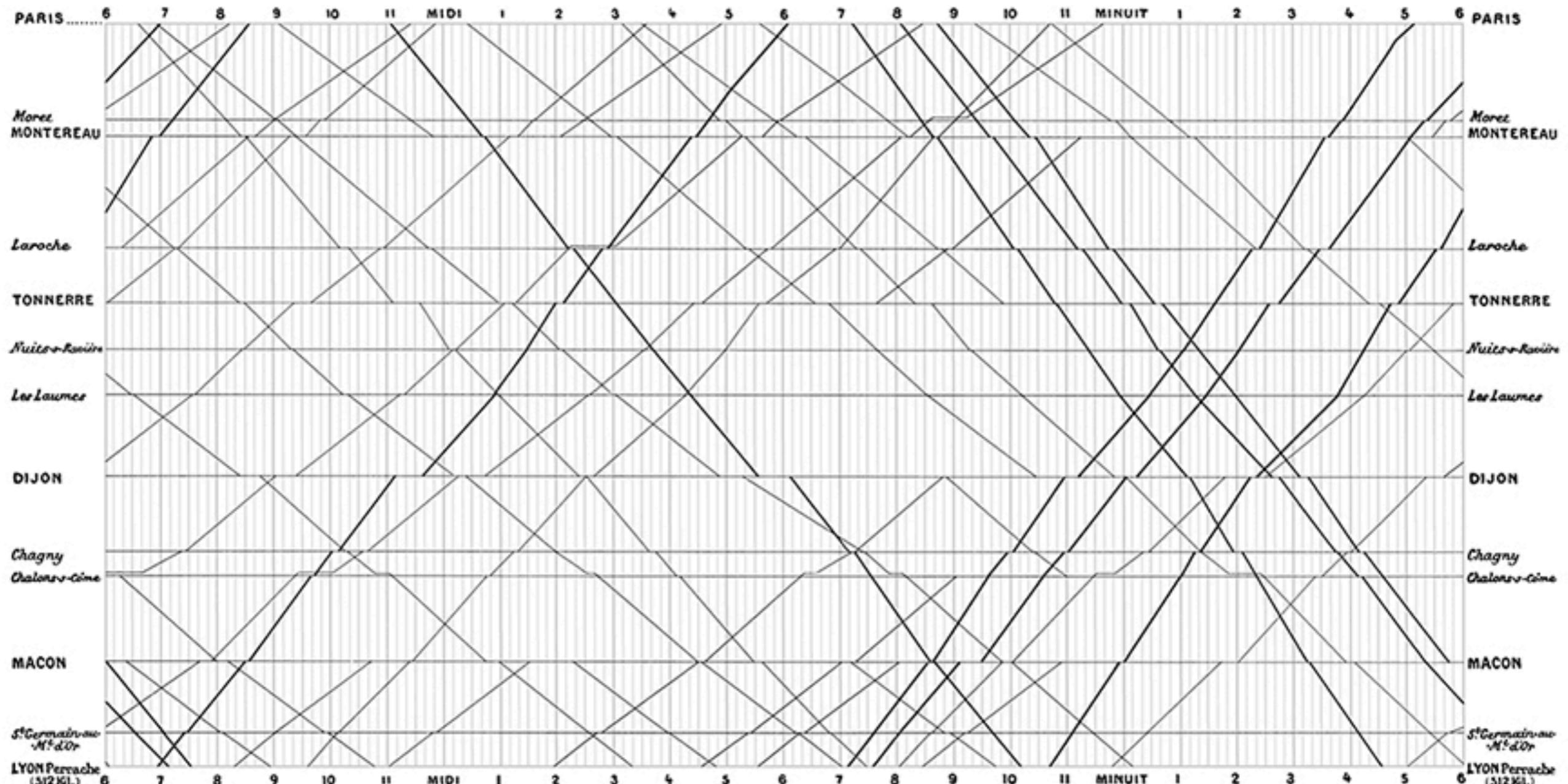
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

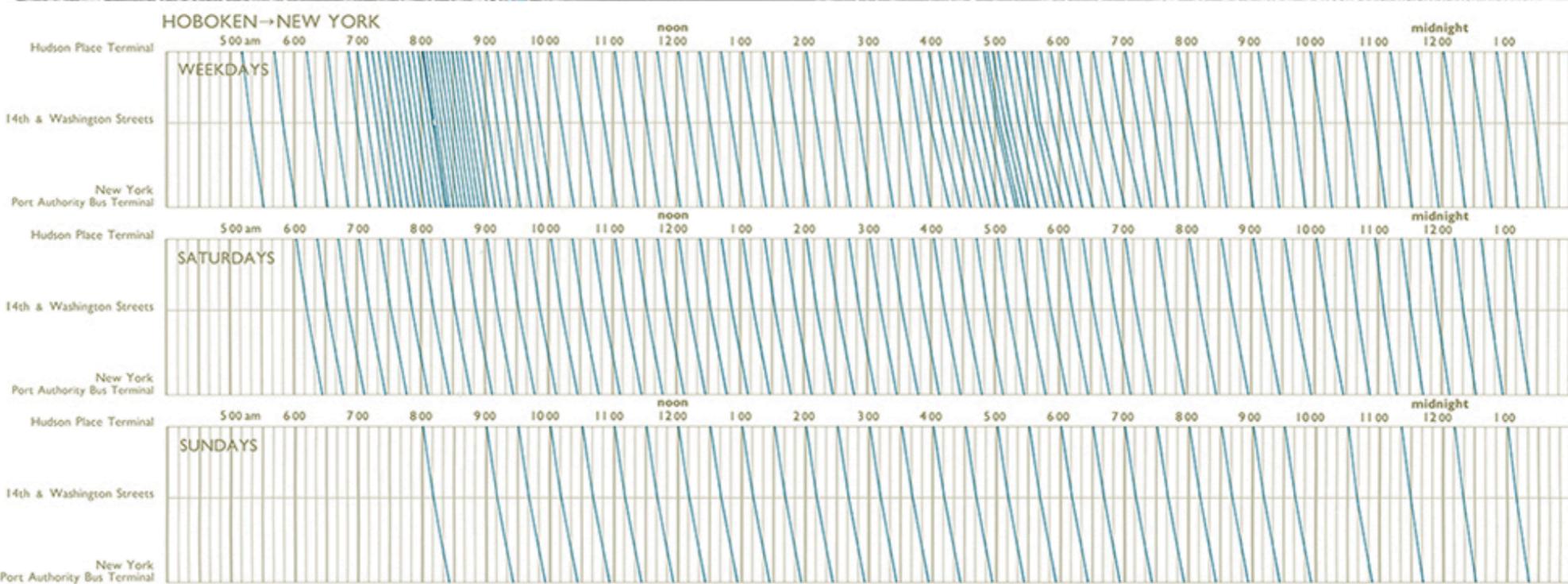
Marey: Train Schedule (1875)



Marey: Train Schedule (1875) - Redesigned by Tufte (1983)

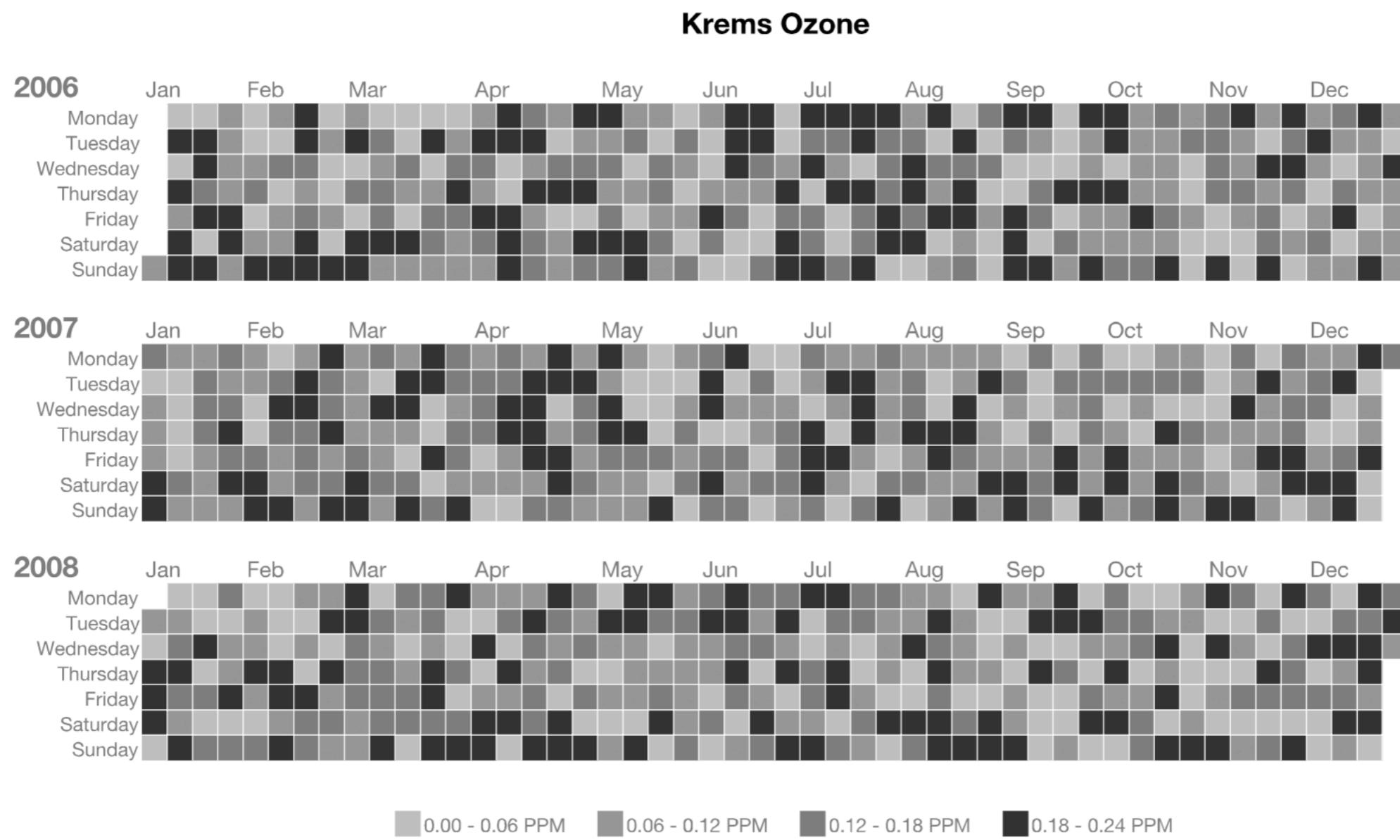


Tufte: Bus Schedule (1990)



from: Tufte, Envisioning Information (1990)

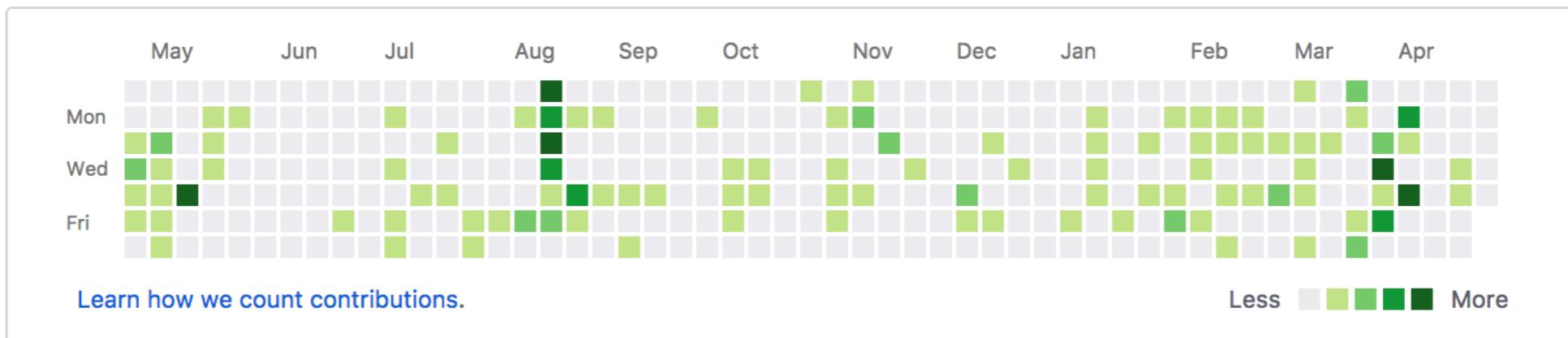
Tile Maps



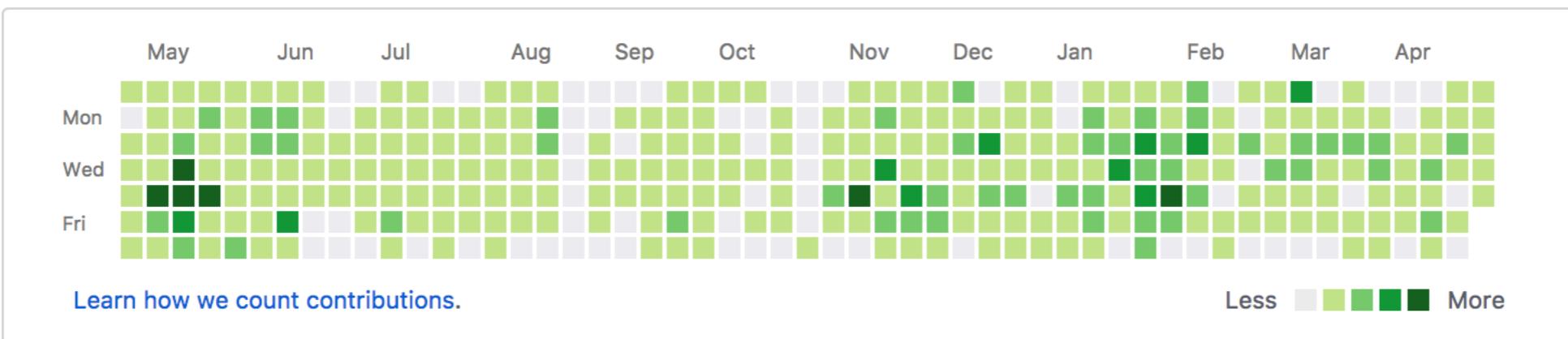
Tile Maps

492 contributions in the last year

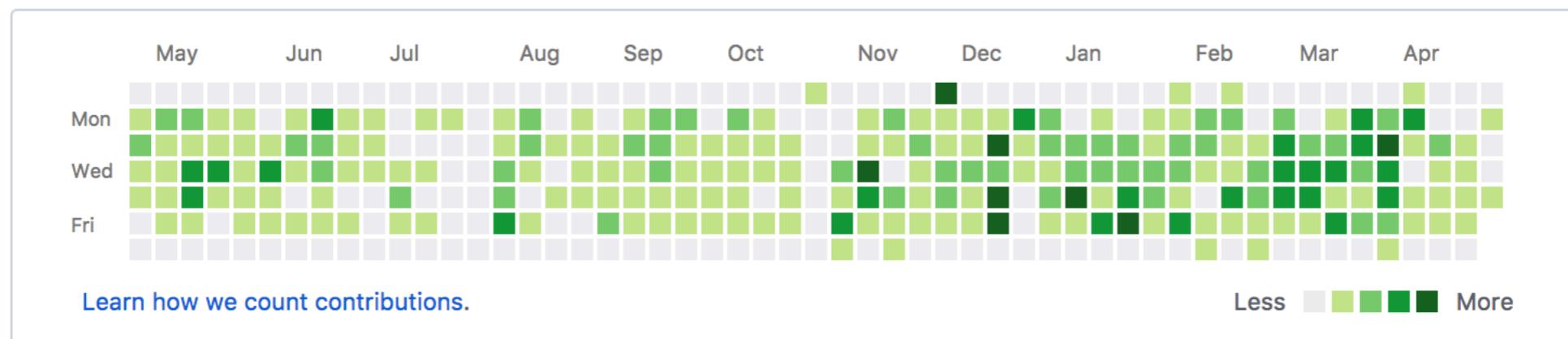
Contribution settings ▾



3,259 contributions in the last year



1,884 contributions in the last year



Calendar-based Views

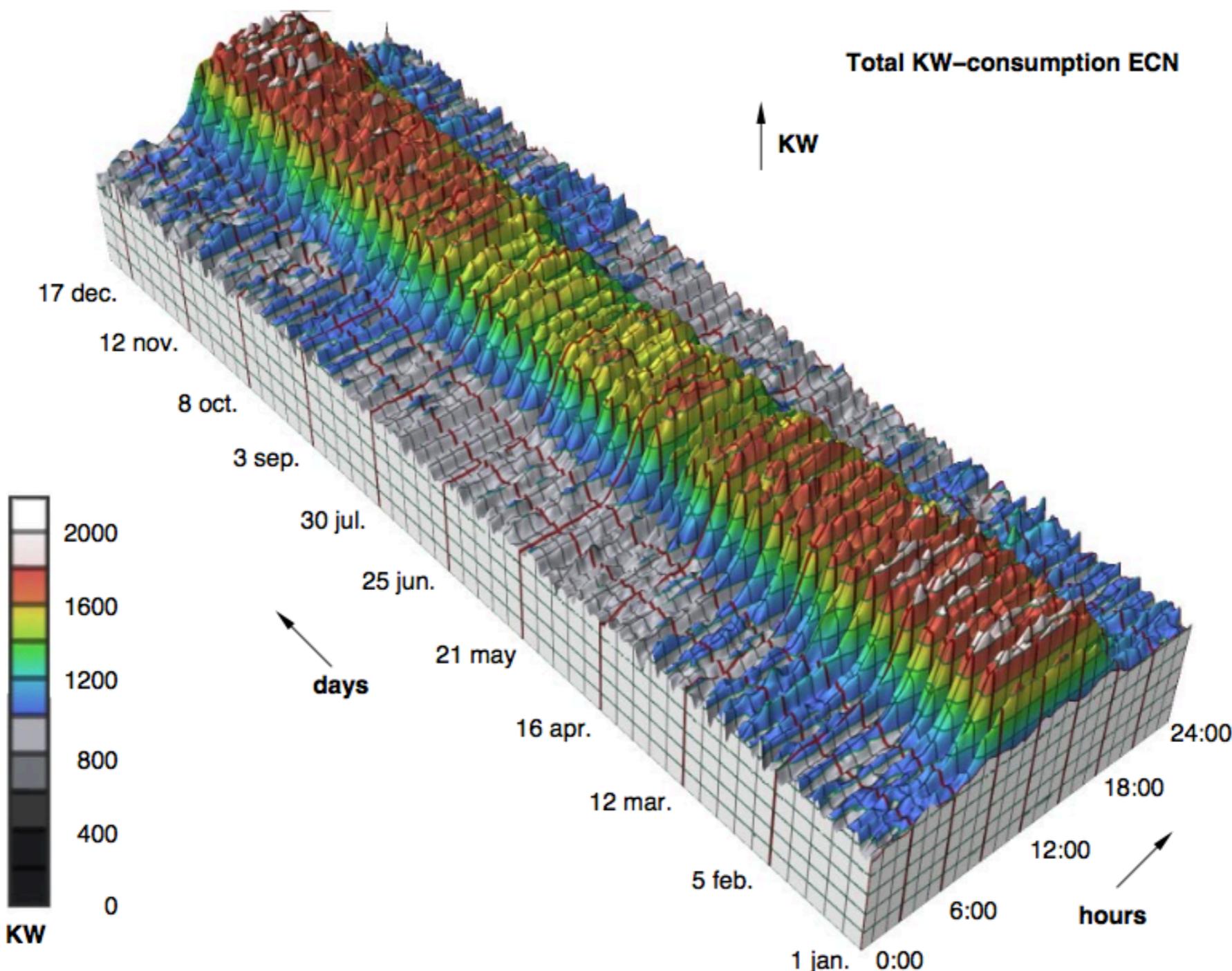


Figure 1. Power demand by ECN, displayed as a function of hours and days

Calendar-based Views

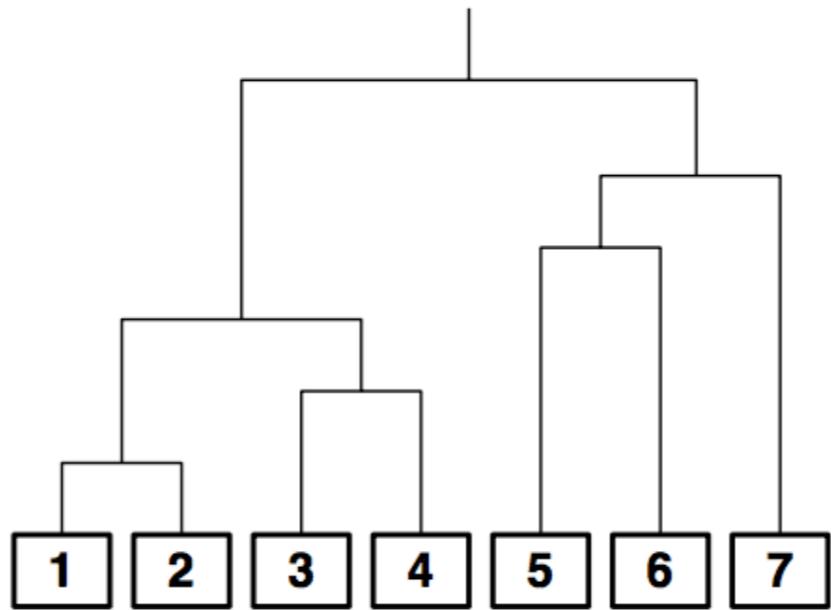


Figure 2. Dendrogram

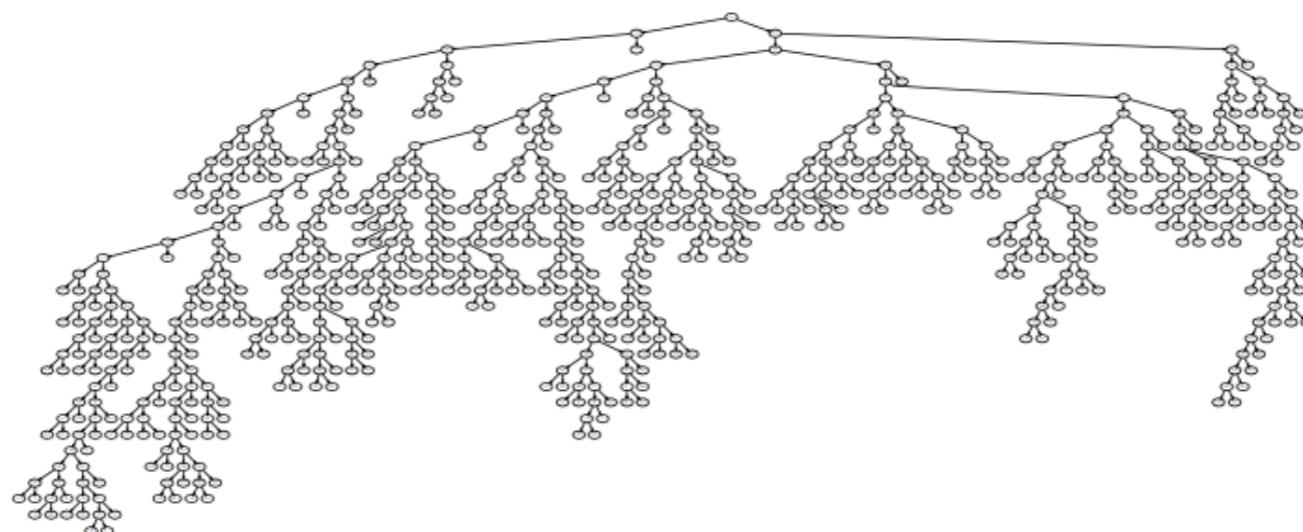


Figure 3. Full clustering tree

Calendar-based Views

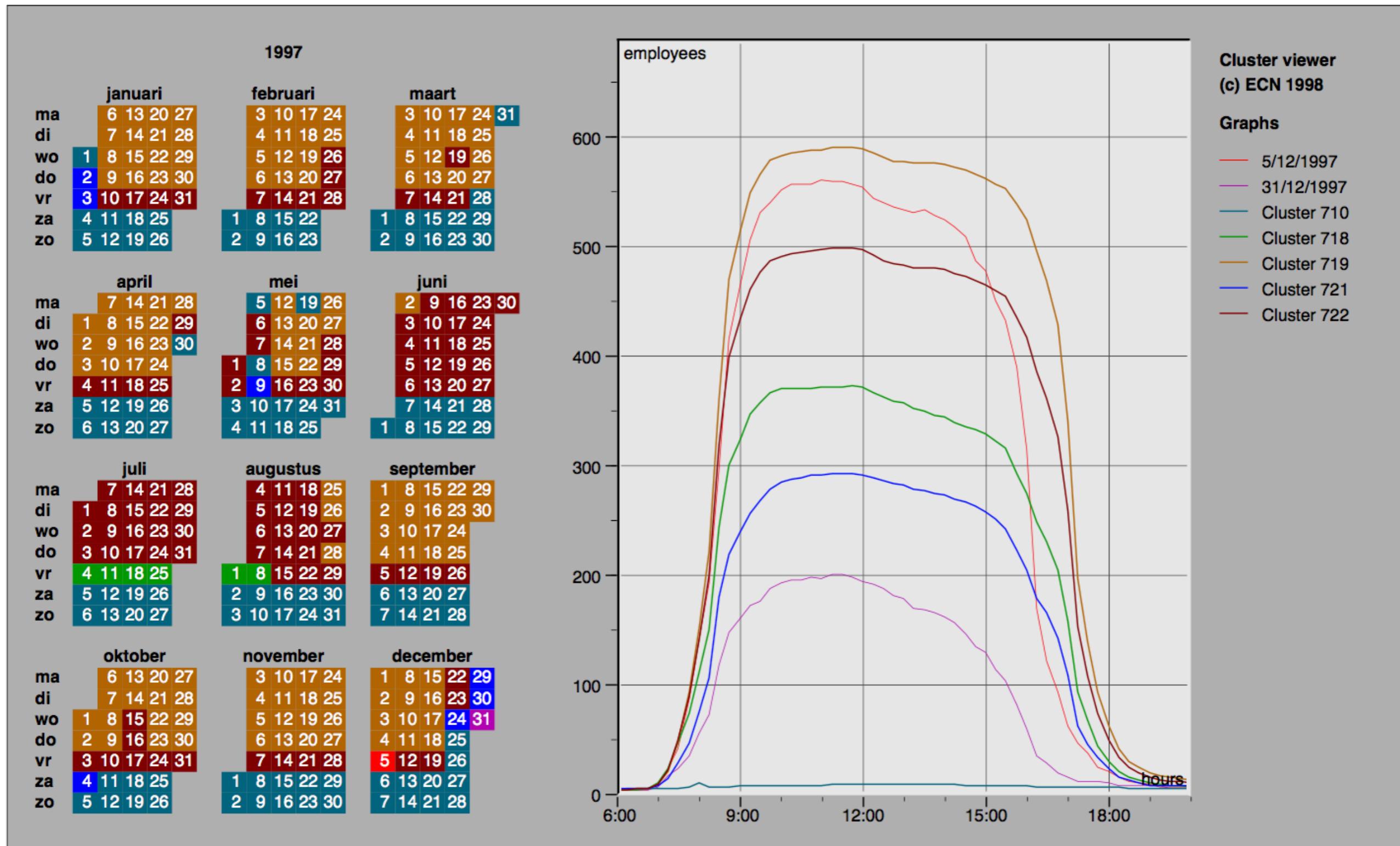
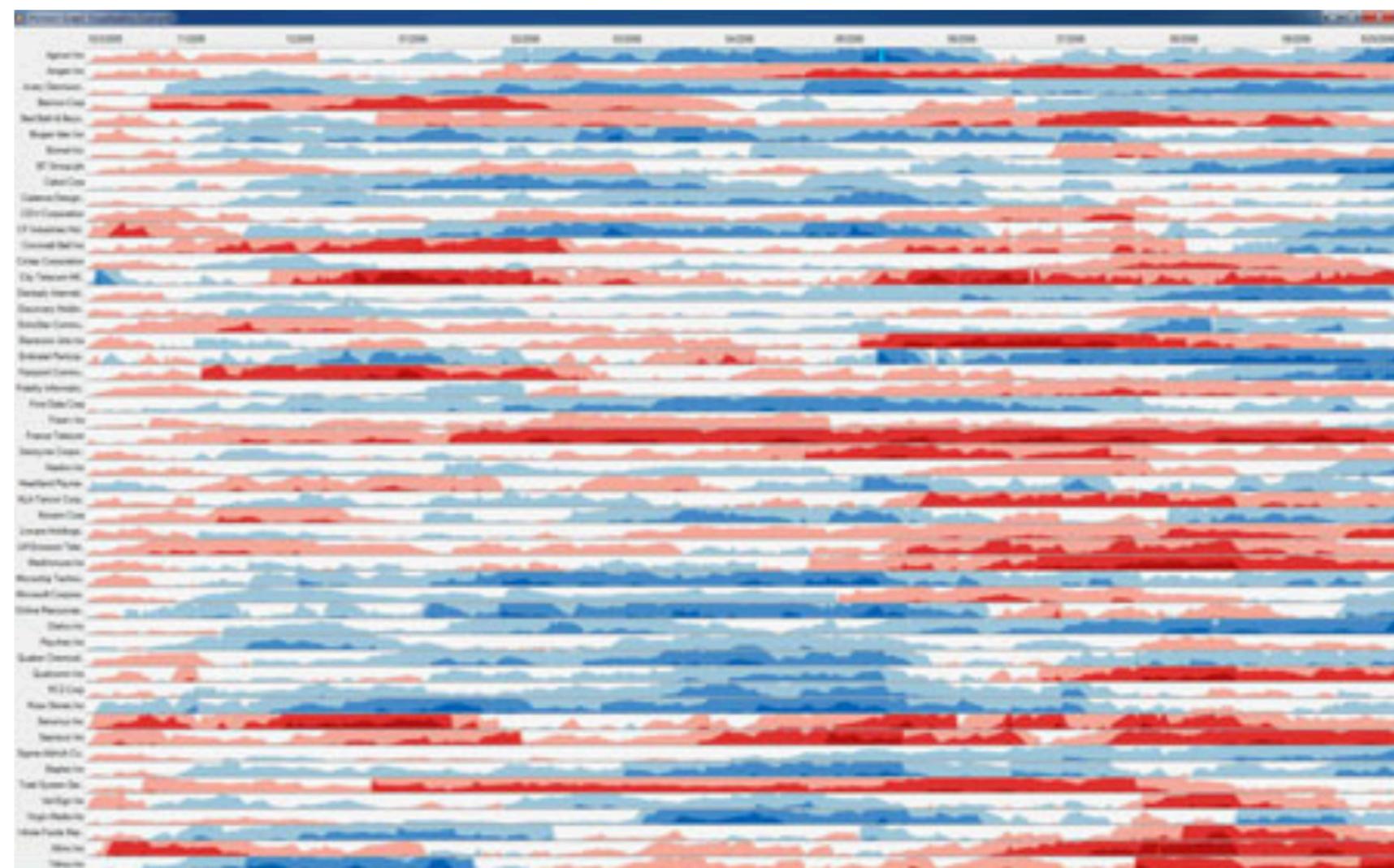
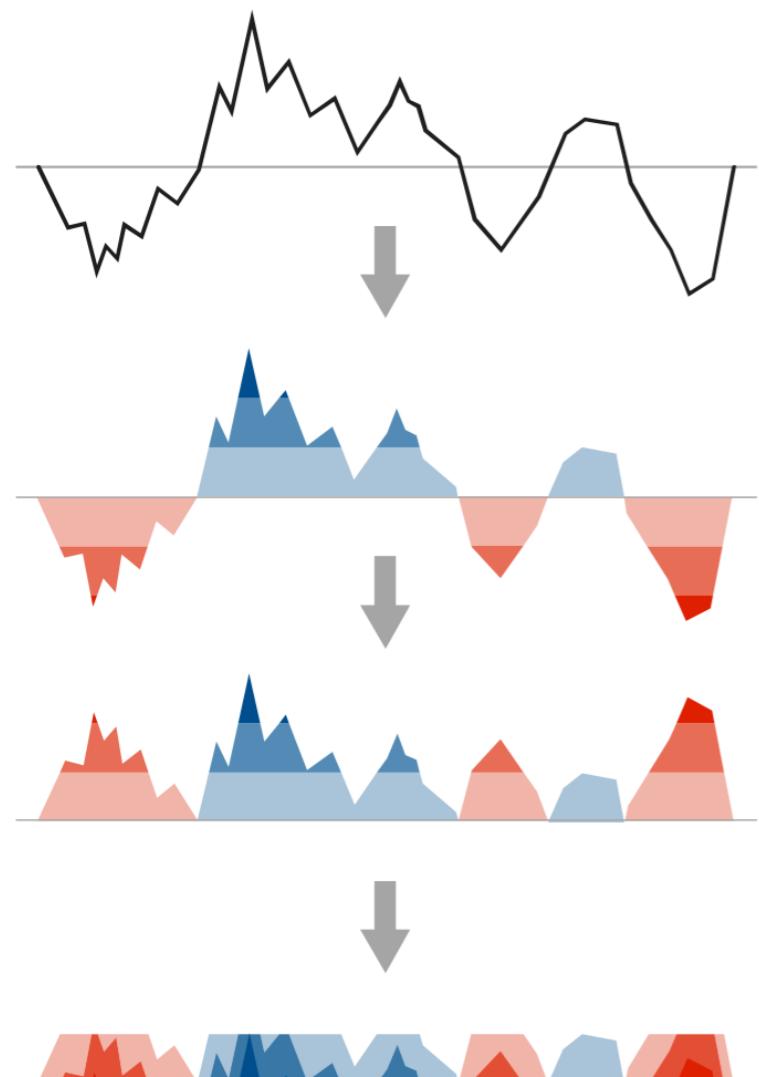


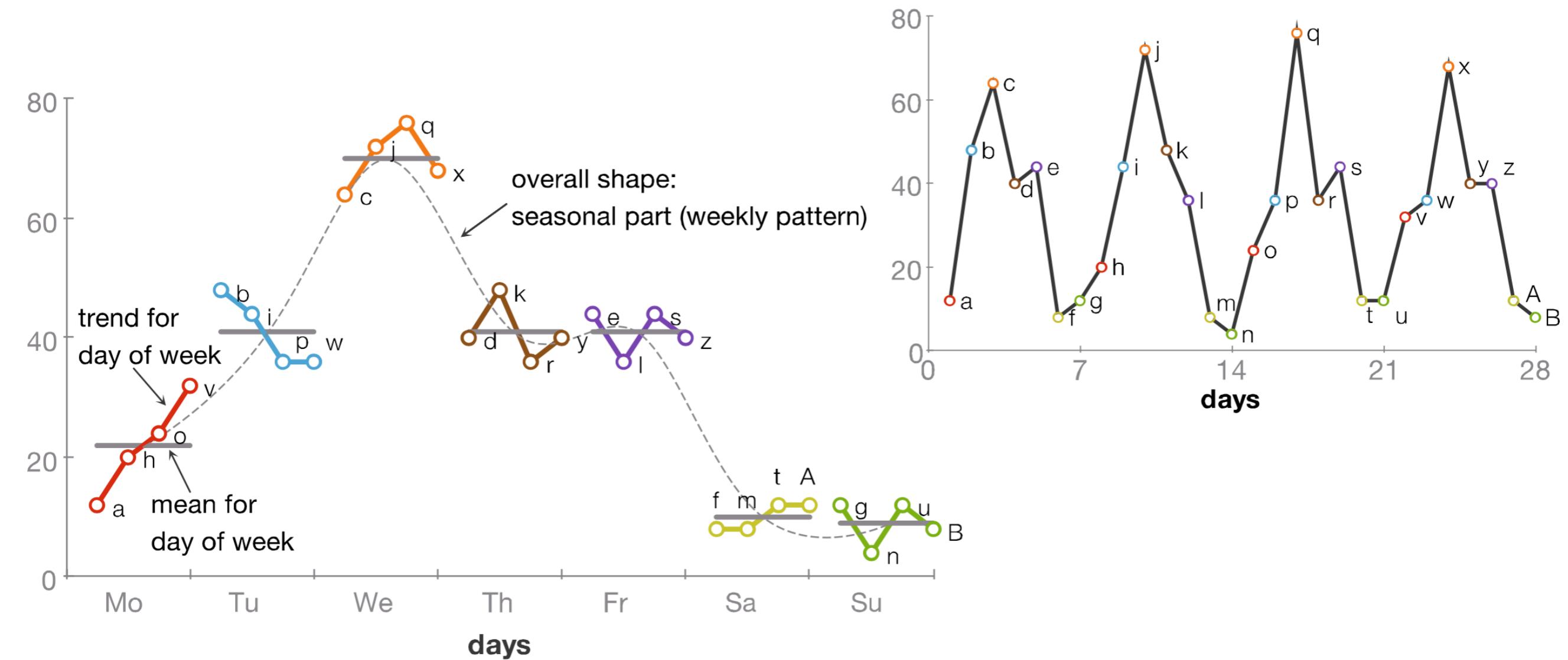
Figure 4. Calendar view of the number of employees

Horizon Charts

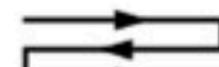
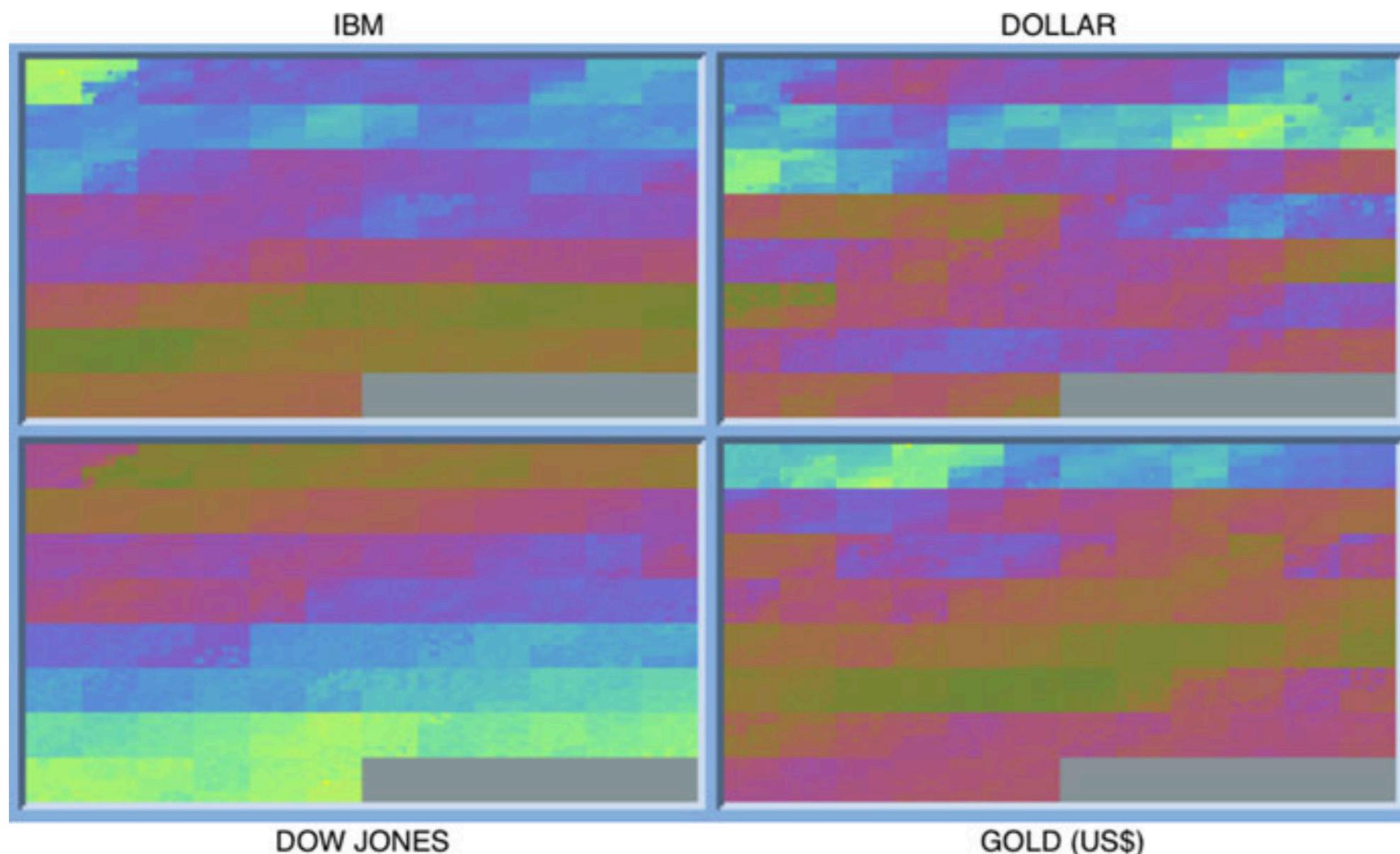


from: Aigner et al., Visualization of Time-oriented Data (2011)

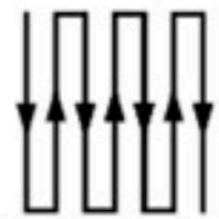
Cycle Plot



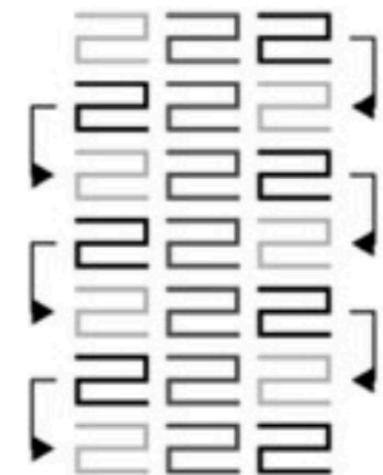
Pixel Maps



a. left-right

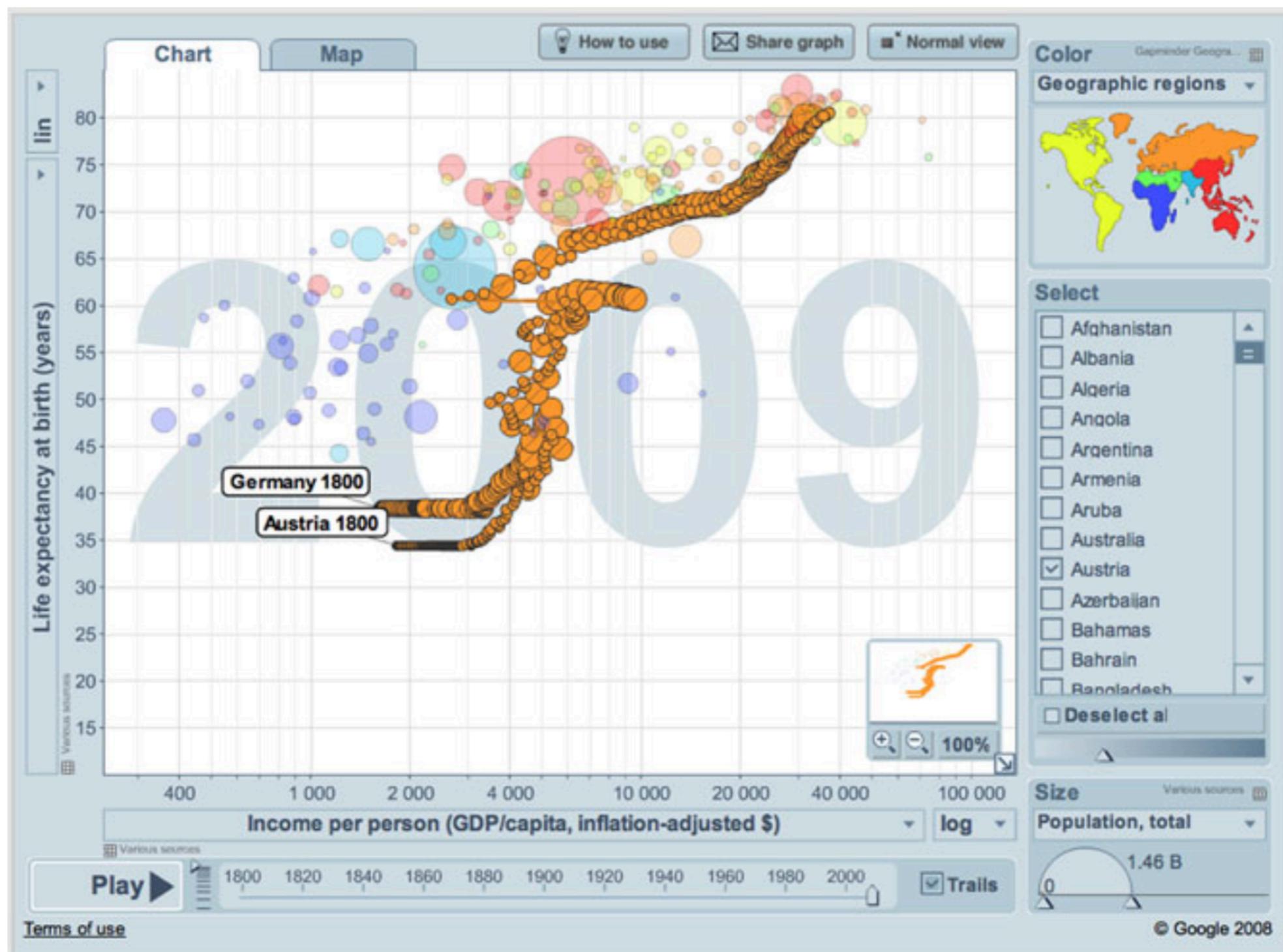


b. top-down



c. back-and-forth loop

Trendalyzer/Gap Minder



from: Aigner et al., Visualization of Time-oriented Data (2011)