Data Visualization for Biomedical Applications

Lecture 3

BMI706 - 5 April 2018

Nils Gehlenborg, PhD

Administrative

- Submission deadline for Homework 2 assignment: Sunday, April 9 at 11:59 pm Eastern

- Next assignment is part of the project, it doesn't count towards your homework grade

- Nils will be traveling on April 18: Peter Kerpedjiev and remote lecture

Talk Announcements

- Friday, April 6, 1:00pm to 6:00pm - Data Science in Biomedicine Symposium

  Fernanda Viegas & Martin Wattenberg (Google)

  Visualization: The secret weapon for machine learning

  Dana-Farber Cancer Institute

- http://bcb.dfci.harvard.edu/index.php/calendar/dfci-fstrf-marvin-zelen-symposium

Talk Announcements

- Friday, April 20, 12:30pm to 2:00pm

  Ben Shneiderman (University of Maryland)

  Interactive Visual Discovery in Event Analytics: Electronic Health Records and Other Applications

  Harvard Geological Museum (Room 100)
  24 Oxford Street, Cambridge MA 02138

- https://www.seas.harvard.edu/calendar/event/111371

Review of Session 2

Visualization Pitfalls

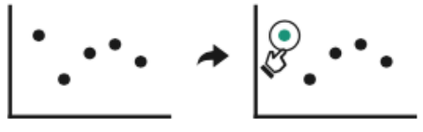
- Color
- Interactions between encodings
- 3D Visualization
- Animation

Single View Interactions

Manipulate

Change over Time

Select

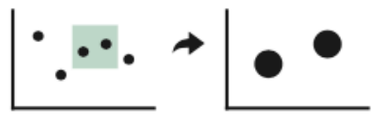


Navigate

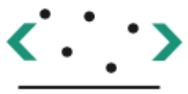
→ Item Reduction

→ *Zoom*

Geometric or *Semantic*



→ *Pan/Translate*

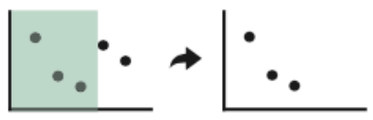


→ *Constrained*

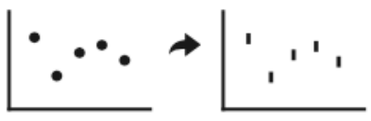
→ Attribute Reduction

→ *Slice*

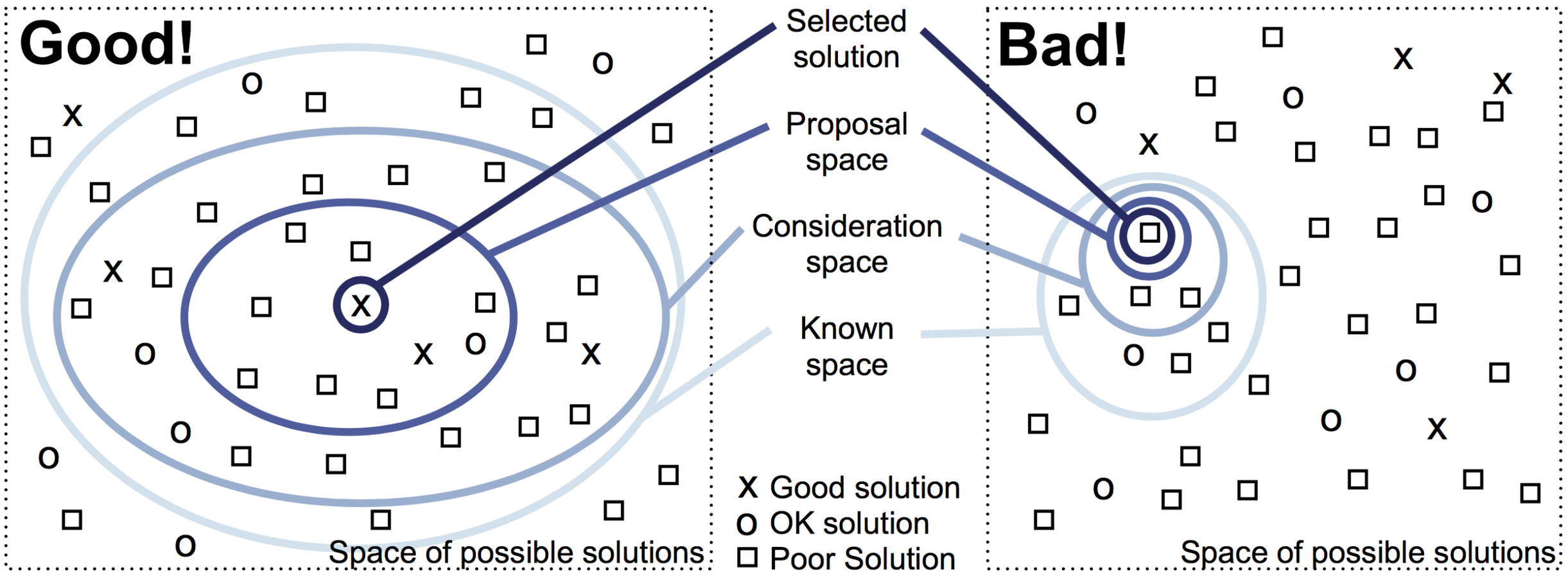
→ *Cut*



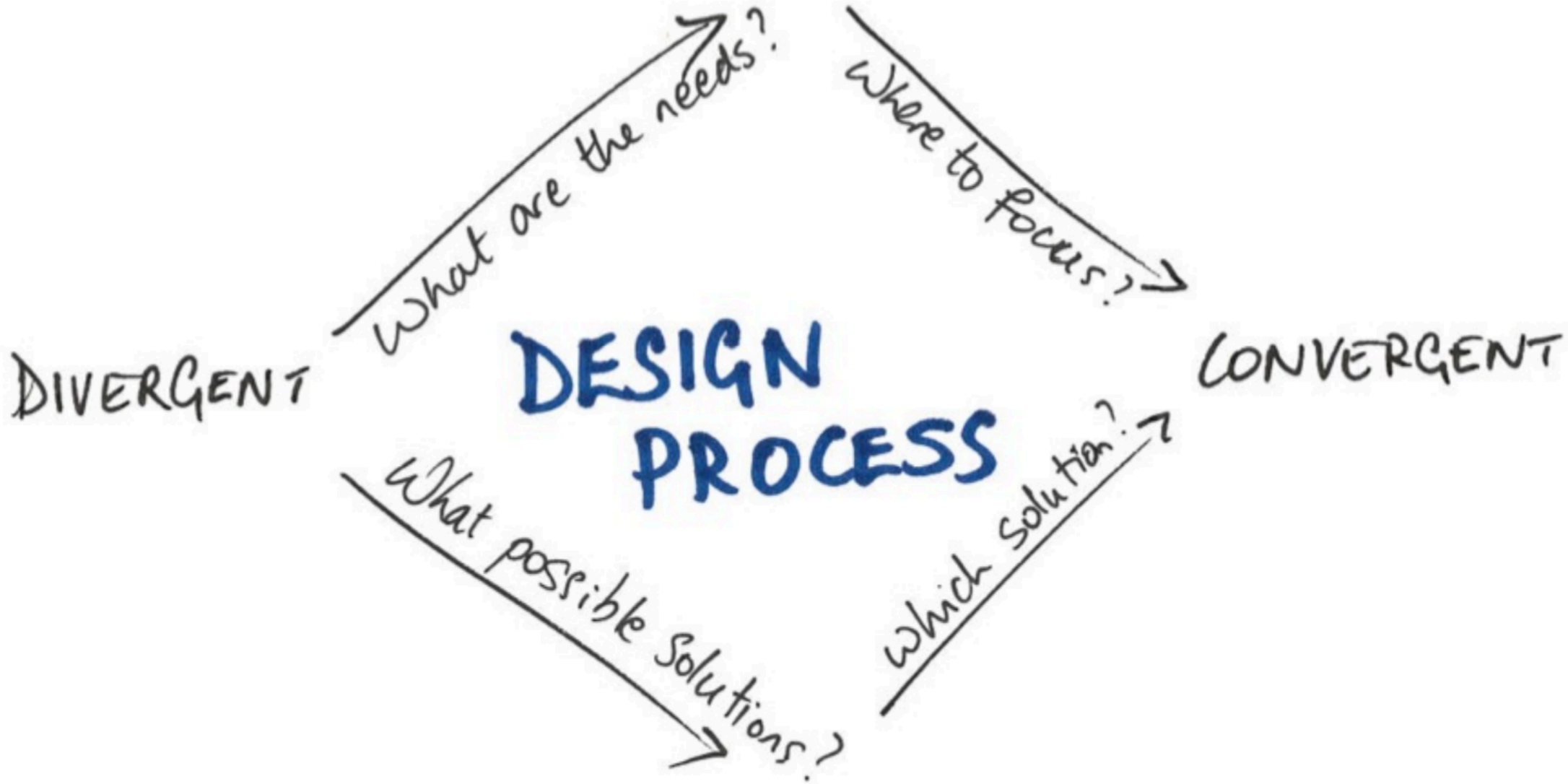
→ *Project*



Munzner, Visualization Analysis and Design, 2014 (Chapter 11)

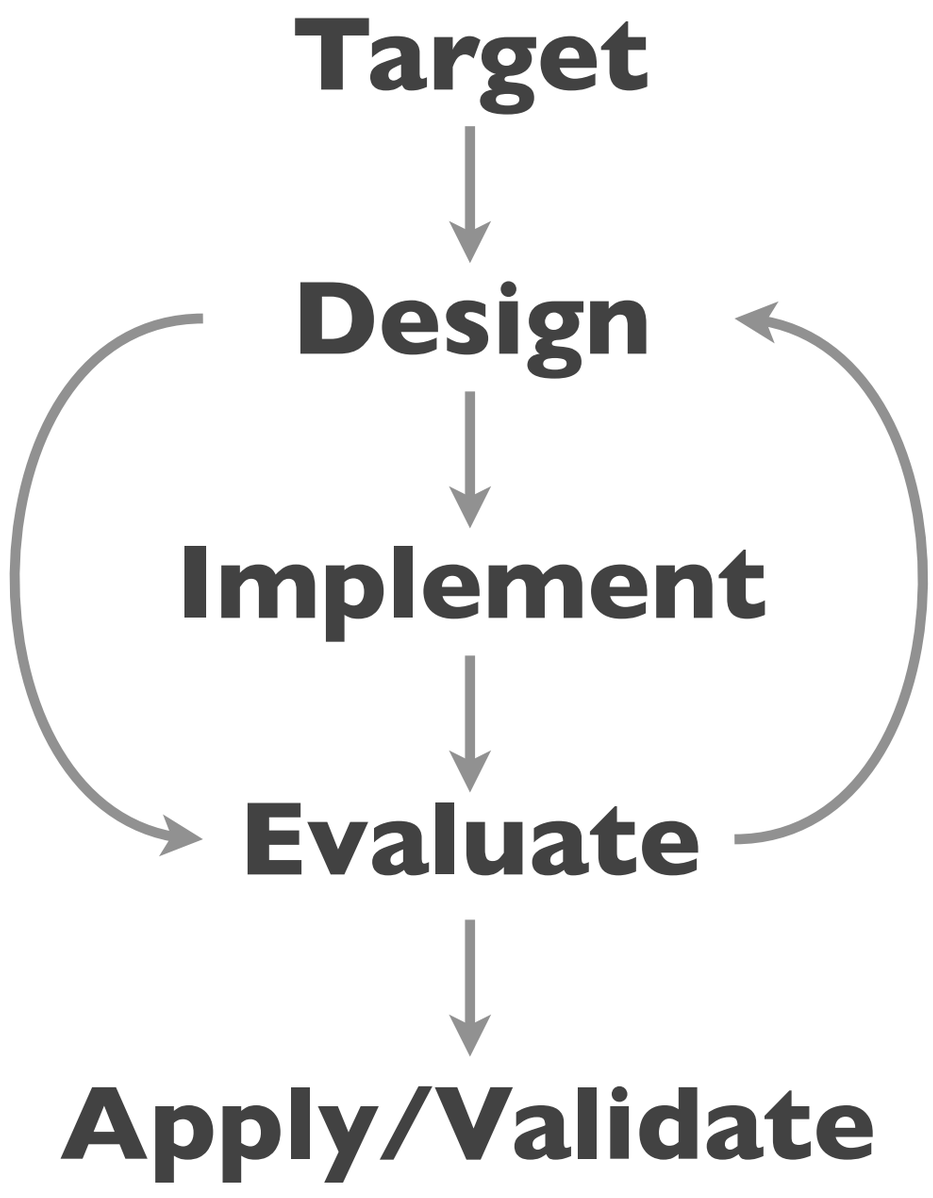

Good!
Bad!
Selected solution
Proposal space
Consideration space
Known space
X Good solution
O OK solution
Poor Solution
Space of possible solutions
Space of possible solutions

Five Design-Sheet Methodology



http://fds.design

User-Centered Participatory Design



user-centered design

usability engineering

participatory design

Tasks?

Evaluate != Validate?

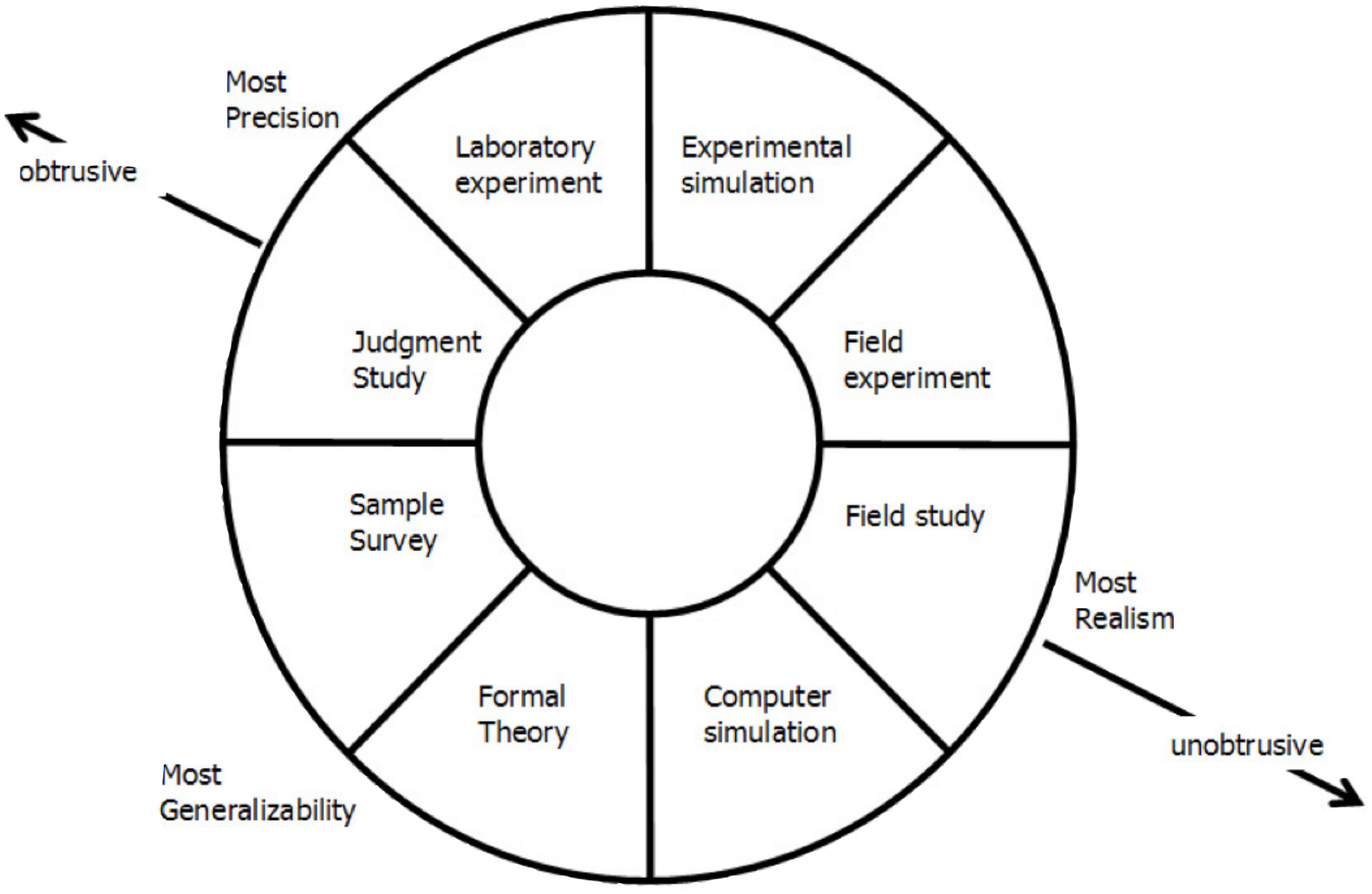
adapted from Miriah Meyer, University of Utah and Cydney Nielsen, University of British Columbia

Lecture 3: Learning Goals

- How can we evaluate a visualization method or tool?
- What are common techniques for multivariate data visualization?
- What is the long tail of data visualization?

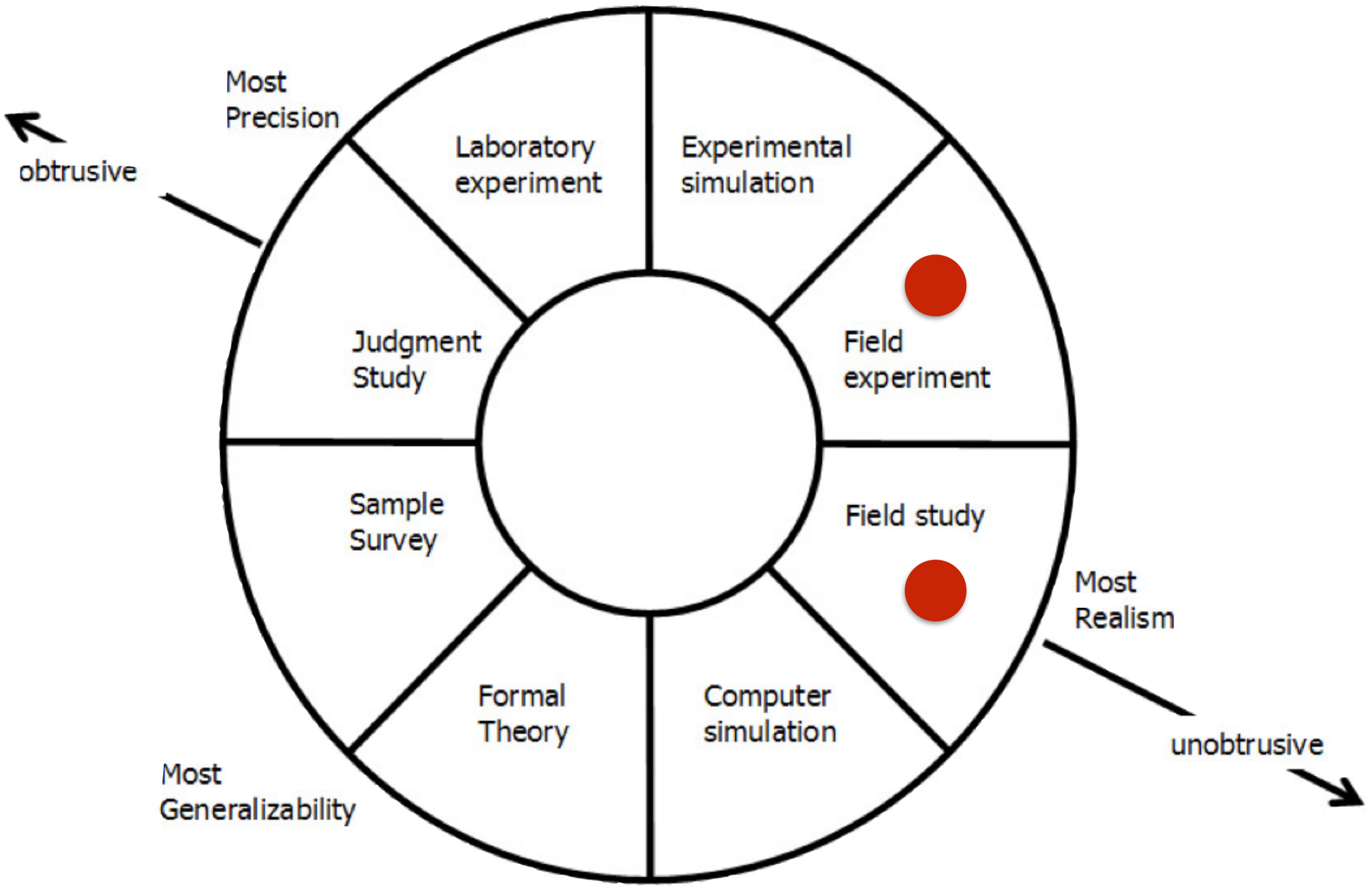
Evaluation

Validation Techniques



Carpendale 2008, Evaluating Information Visualizations

Validation Techniques



Carpendale 2008, Evaluating Information Visualizations

Qualitative Usability Testing

- **Field study**
  - Intended to be unobtrusive
  - Conducted in the actual situation (a user with their own analysis tasks and data)
  - Watching someone use your tool can be very informative; you may be surprised!

Carpendale 2008, Evaluating Information Visualizations

Qualitative Usability Testing

- **Field experiment**
  - Design user tasks to simulate real analyses
  - Recruit group of users and arrange one-on-one sessions
  - Encourage thinking aloud and note top usability issues

Carpendale 2008, Evaluating Information Visualizations

Qualitative Usability Testing

- Learn through observation
- Learn about behavior, not just opinions
- Not about proof, but about insight and context
- Test early and test often
- **Important that your users are assured that you are testing the software, not them**

collected by Francis Rowland, EMBL-EBI

Quantitative User Testing

- Error Rate
- Time (see LineUp paper next week)
- Insights
  - **Oh boy!**

See required reading for Week 5

Gremlin: an interactive visualization model for analyzing genomic rearrangements

O'Brien TM, Ritz AM, Raphael BJ, Laidlaw DH

IEEE Trans. Vis. Comput. Graph., 2010

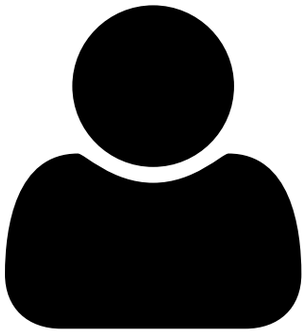
Evaluation

What is an effective visualization and what is not?

Considerations - e.g. measuring effectiveness quantitatively is extremely problematic

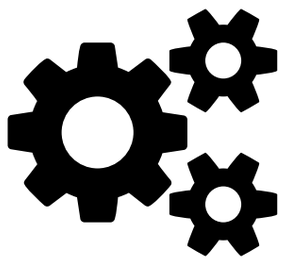
Evaluation

1



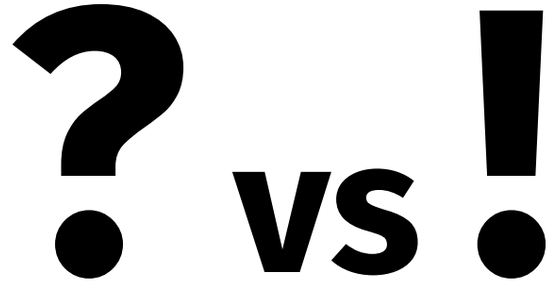
human aspect

2



system complexity

3



outcome is question, not answer

Evaluation

What is an effective visualization and what is not?

Considerations - e.g. measuring effectiveness quantitatively is extremely problematic

Impact - e.g. on developers (are we doing good work?), users (is this a good tool?), and reviewers (should this be published/funded?), e.g. in evaluating visual against analytical approaches

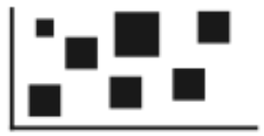
Tabular Data

Arrange Tables

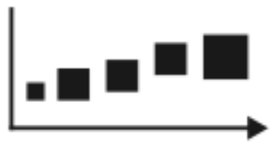
Express Values

Separate, Order, Align Regions

→ Separate



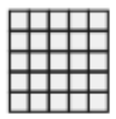
→ Order



→ Align

→ *1 Key*
List

→ *2 Keys*
Matrix



→ *3 Keys*
Volume

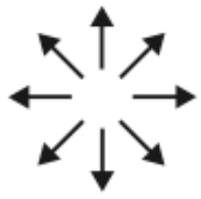
→ *Many Keys*
Recursive Subdivision

Axis Orientation

→ Rectilinear

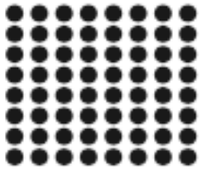
→ Parallel

→ Radial



Layout Density

→ Dense



→ Space-Filling

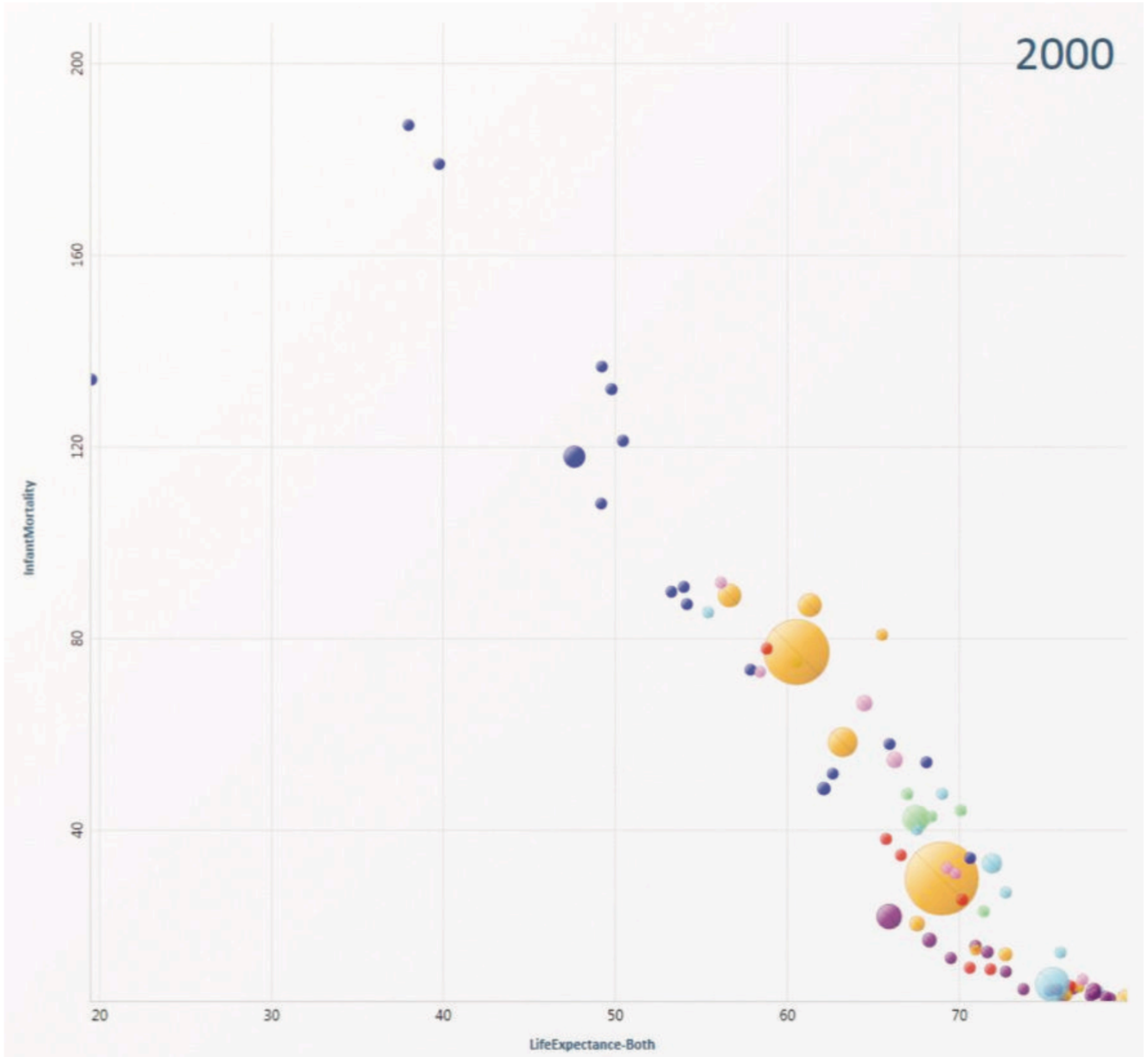


Figure 7.2. Scatterplot. Each point mark represents a country, with horizontal and vertical spatial position encoding the primary quantitative attributes of life expectancy and infant mortality. The color channel is used for the categorical country attribute and the size channel for quantitative population attribute. From [Robertson et al. 08, Figure 1c].

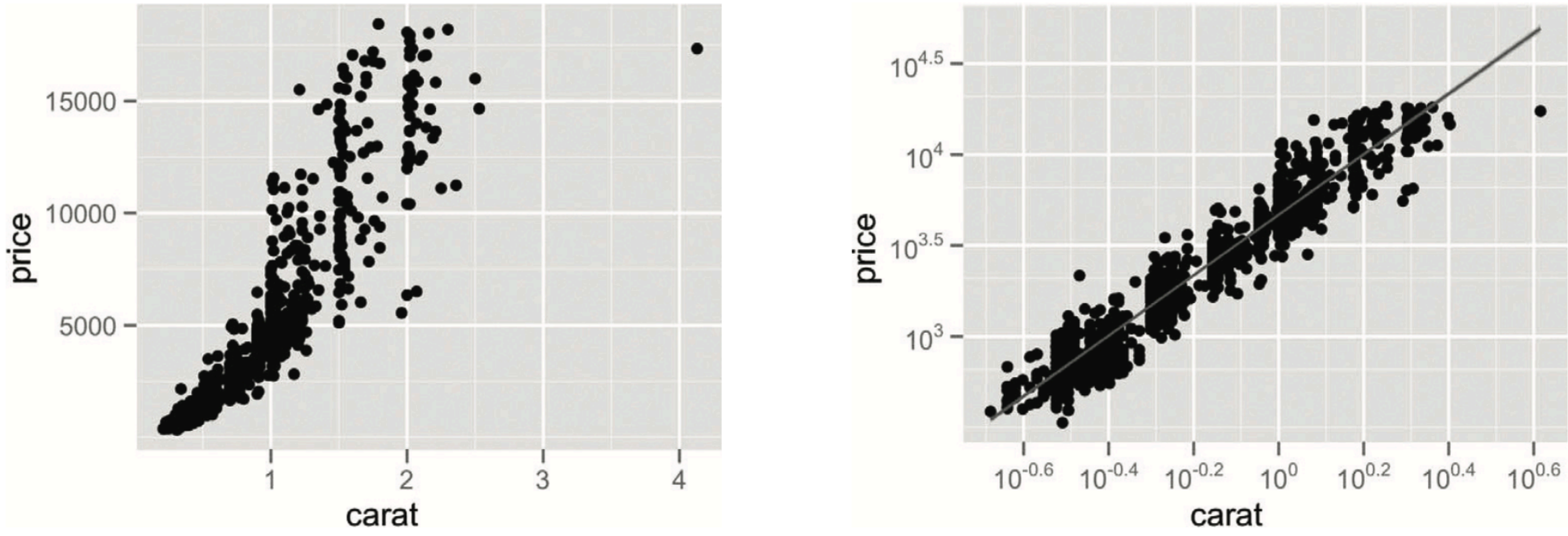
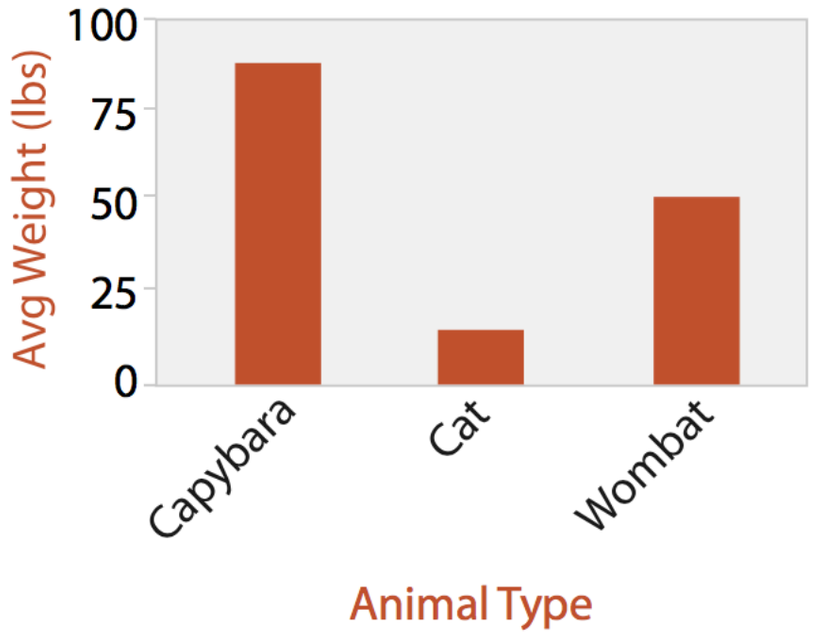
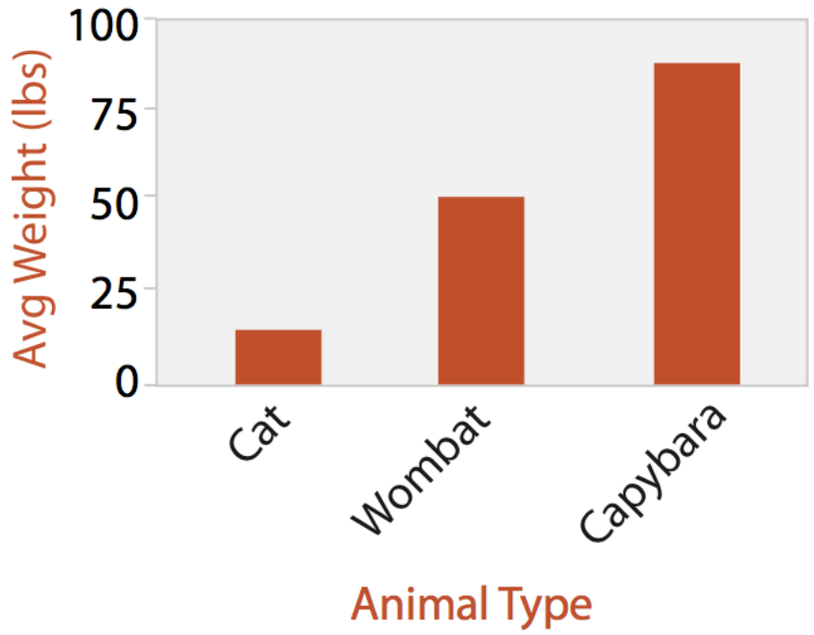


Figure 7.3. Scatterplots. (a) Original diamond price/carat data. (b) Derived log-scale attributes are highly positively correlated. From [Wickham 10, Figure 10].


100
75
50
25
0
Avg Weight (lbs)
Capybara
Cat
Wombat
Animal Type


100
75
50
25
0
Avg Weight (lbs)
Cat
Wombat
Capybara
Animal Type

Bar Charts for Items & Categories

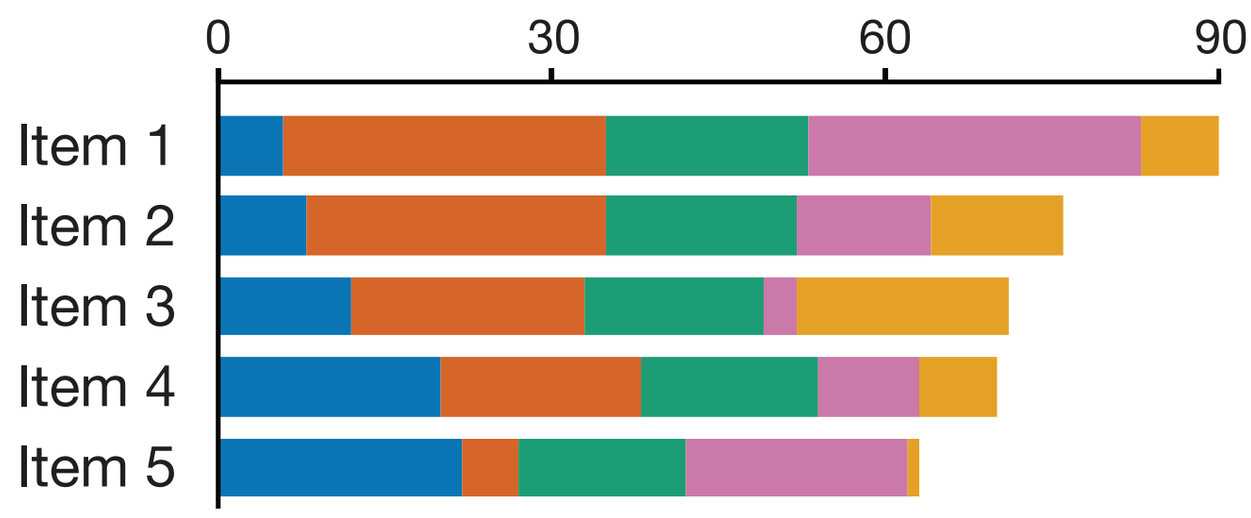
	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

Streit and Gehlenborg, Nature Methods, 2014

Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

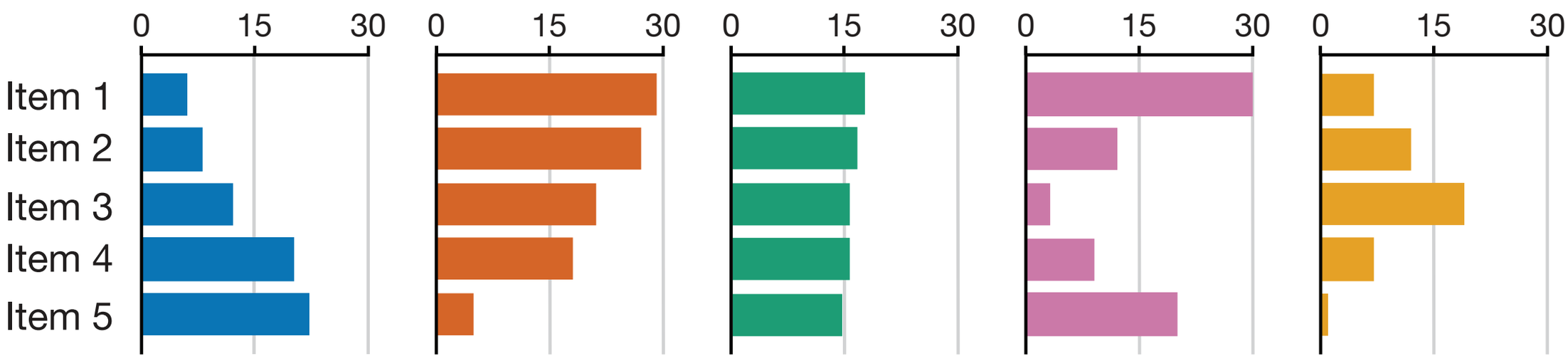
Stacked Bar Chart



Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

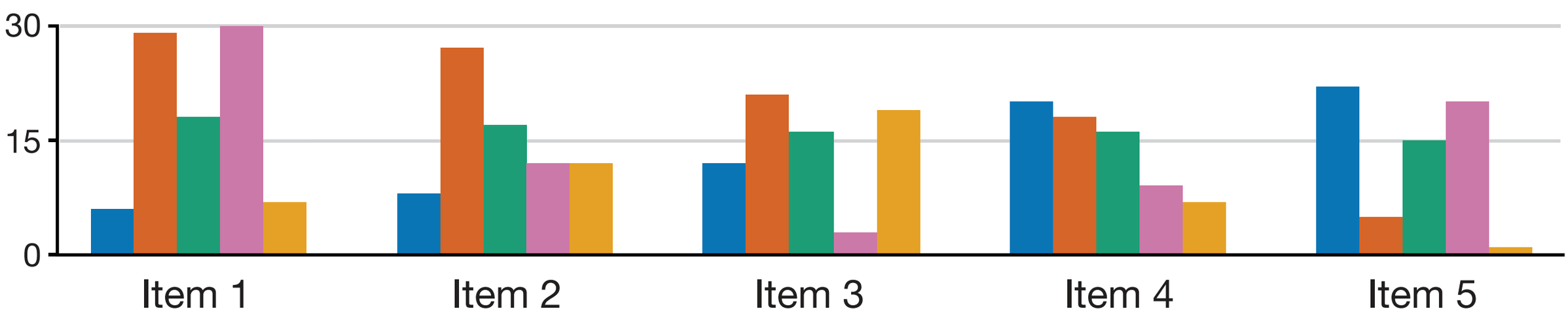
Layered Bar Chart



Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

Grouped Bar Chart



Streit and Gehlenborg, Nature Methods, 2014

Table Lens

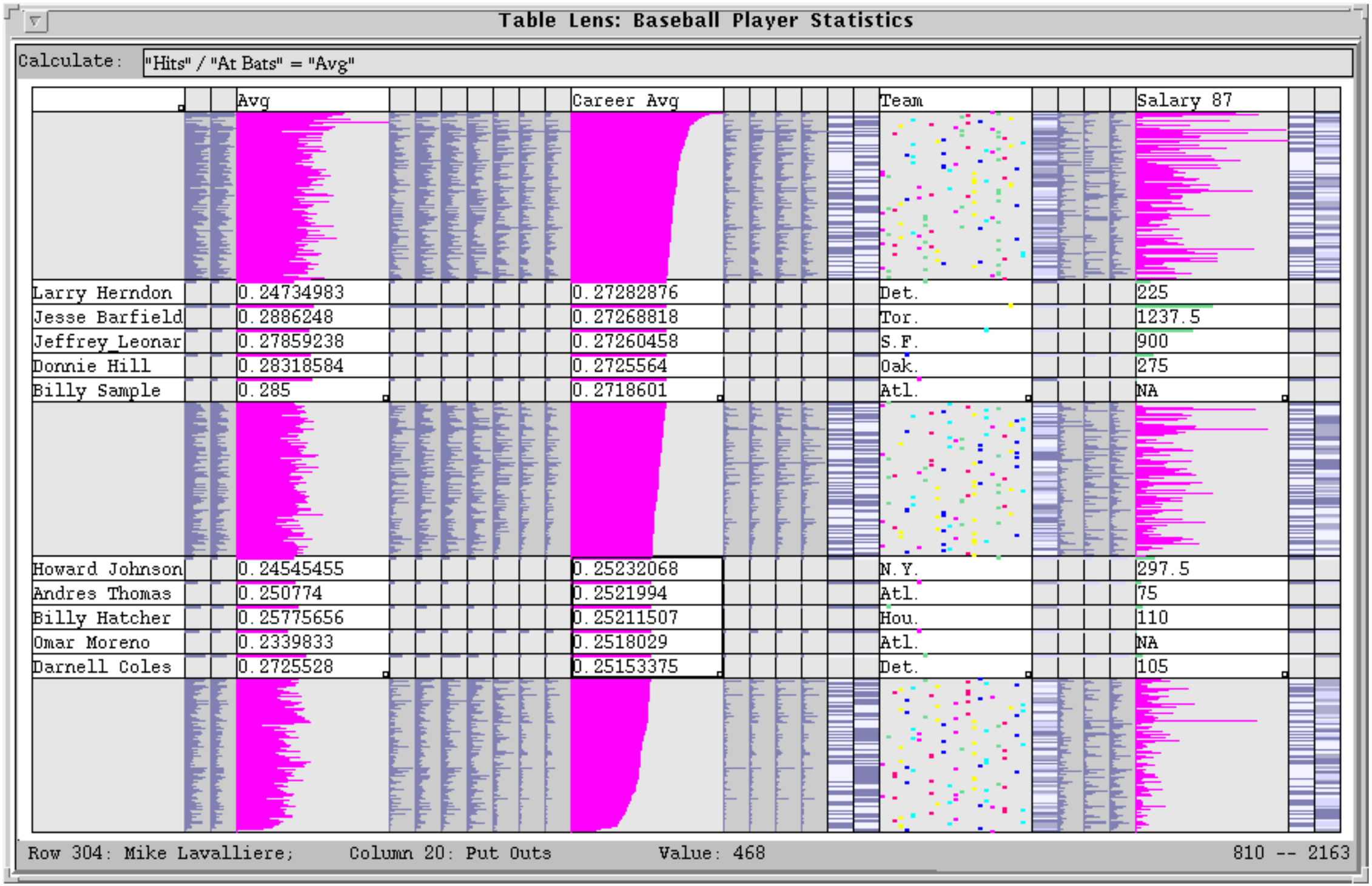


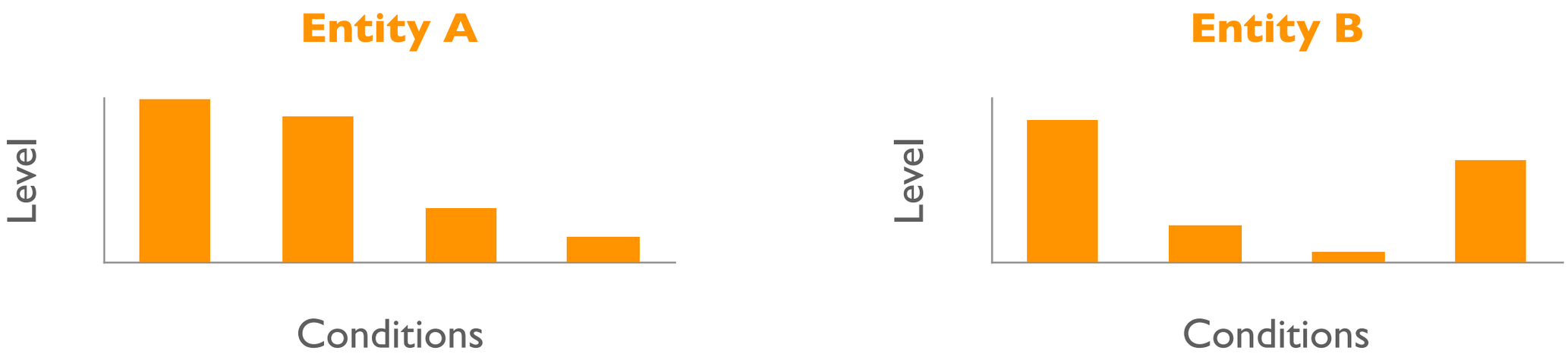
Table Lens (Rao & Card), 1994

Multivariate Data

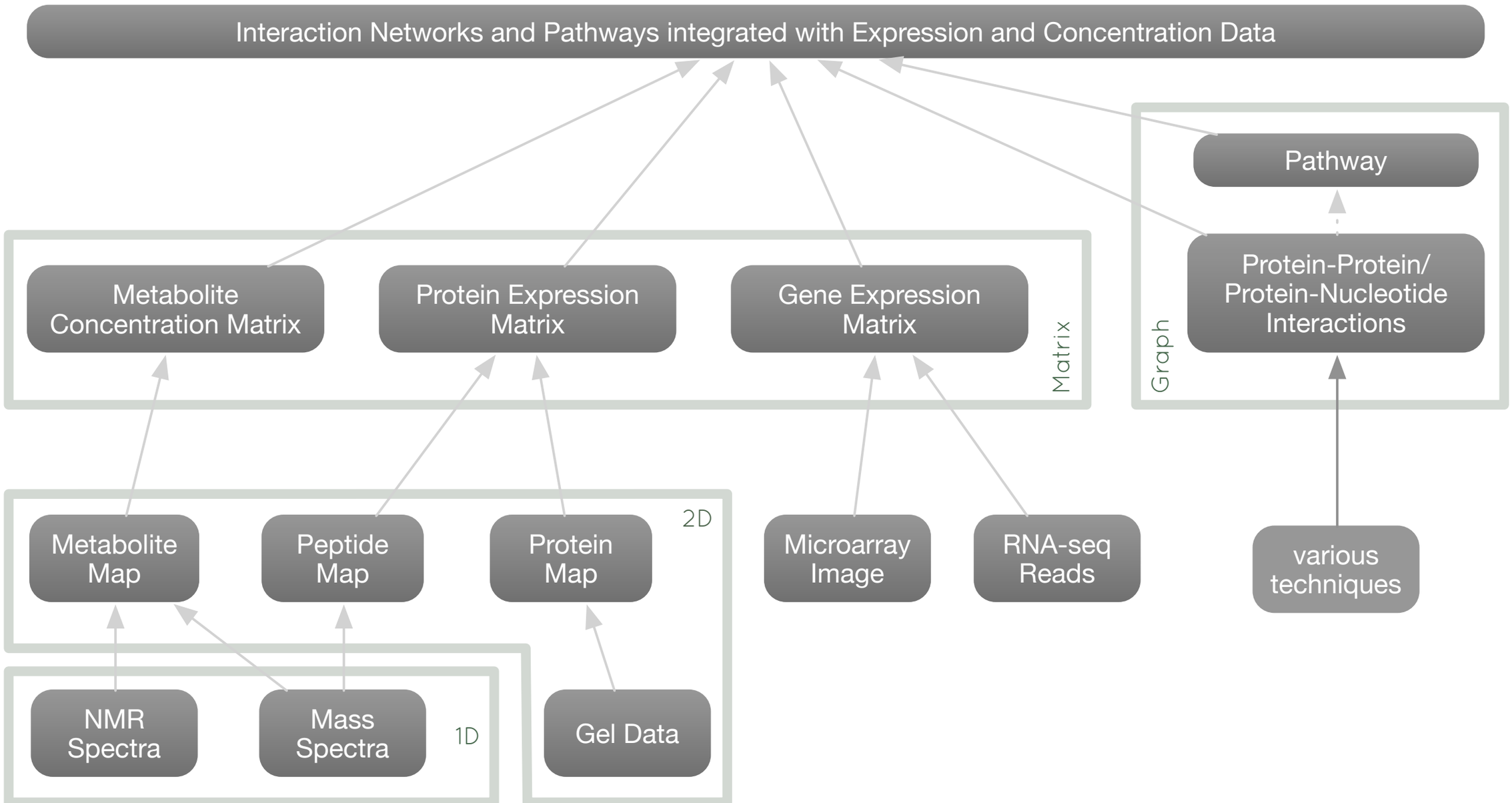
Homogeneous Tables

Multivariate Data

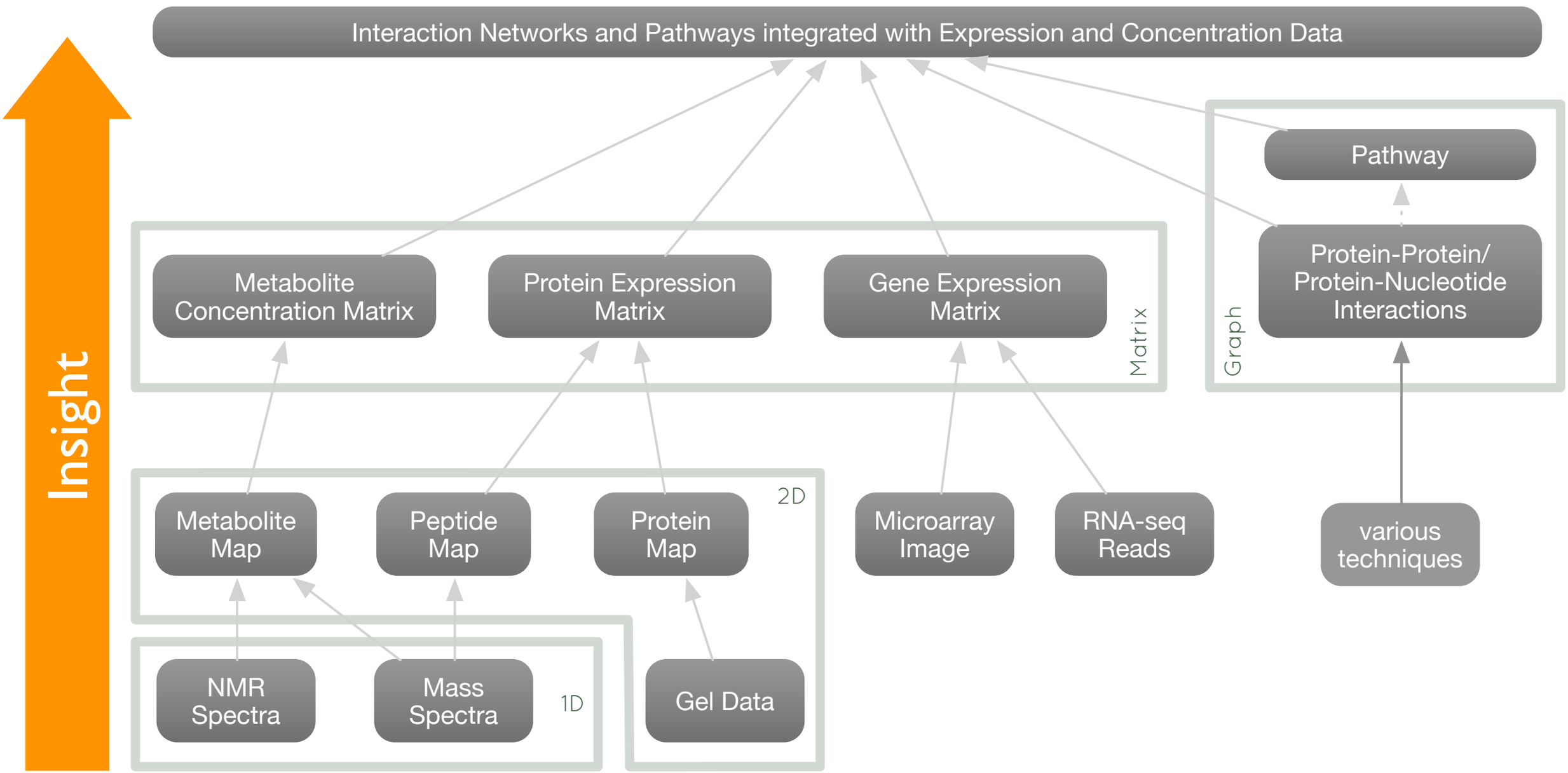
- typical “omics” data: transcriptomics, proteomics, metabolomics
- expression/concentration levels of many biological entities (transcripts, proteins, etc.) across many different conditions/time points
- entity levels measured per sample on a “genome-wide” scale
- often entities are not measured directly



Multivariate Data

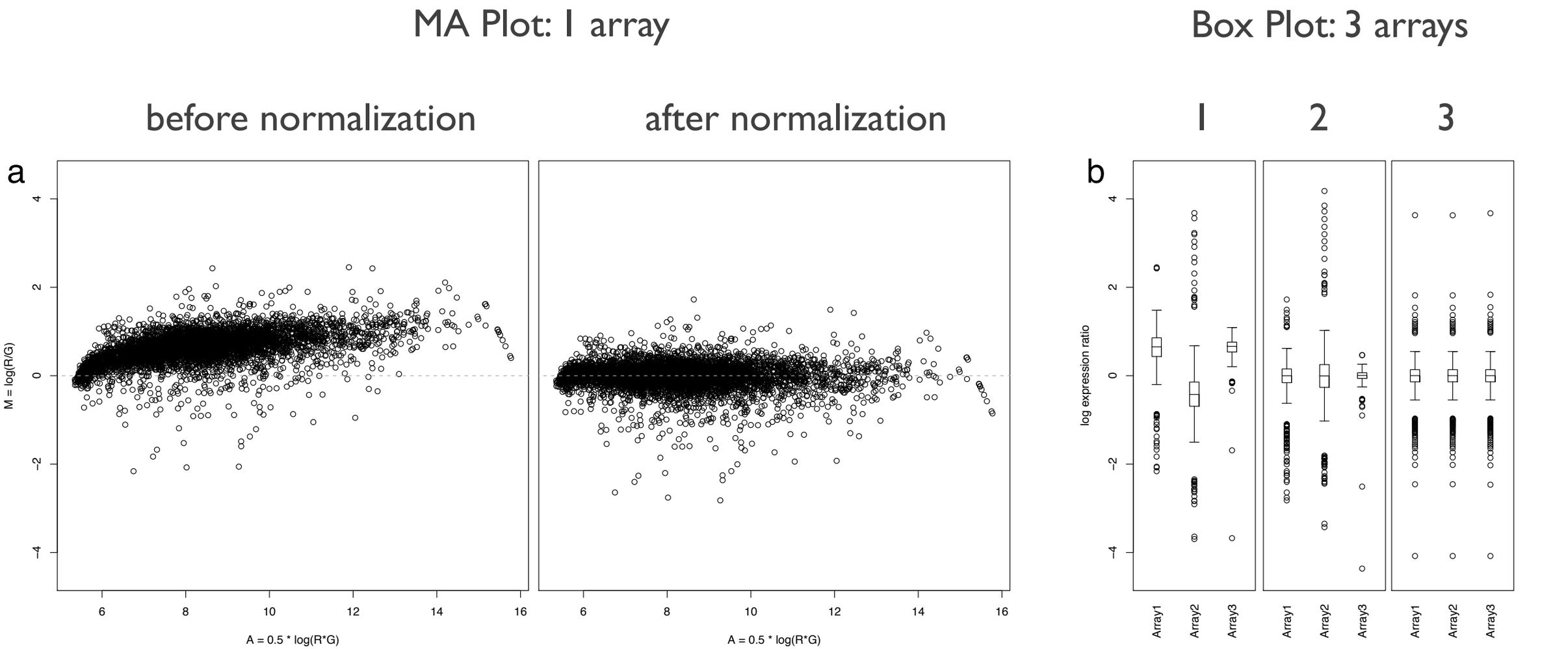


Multivariate Data

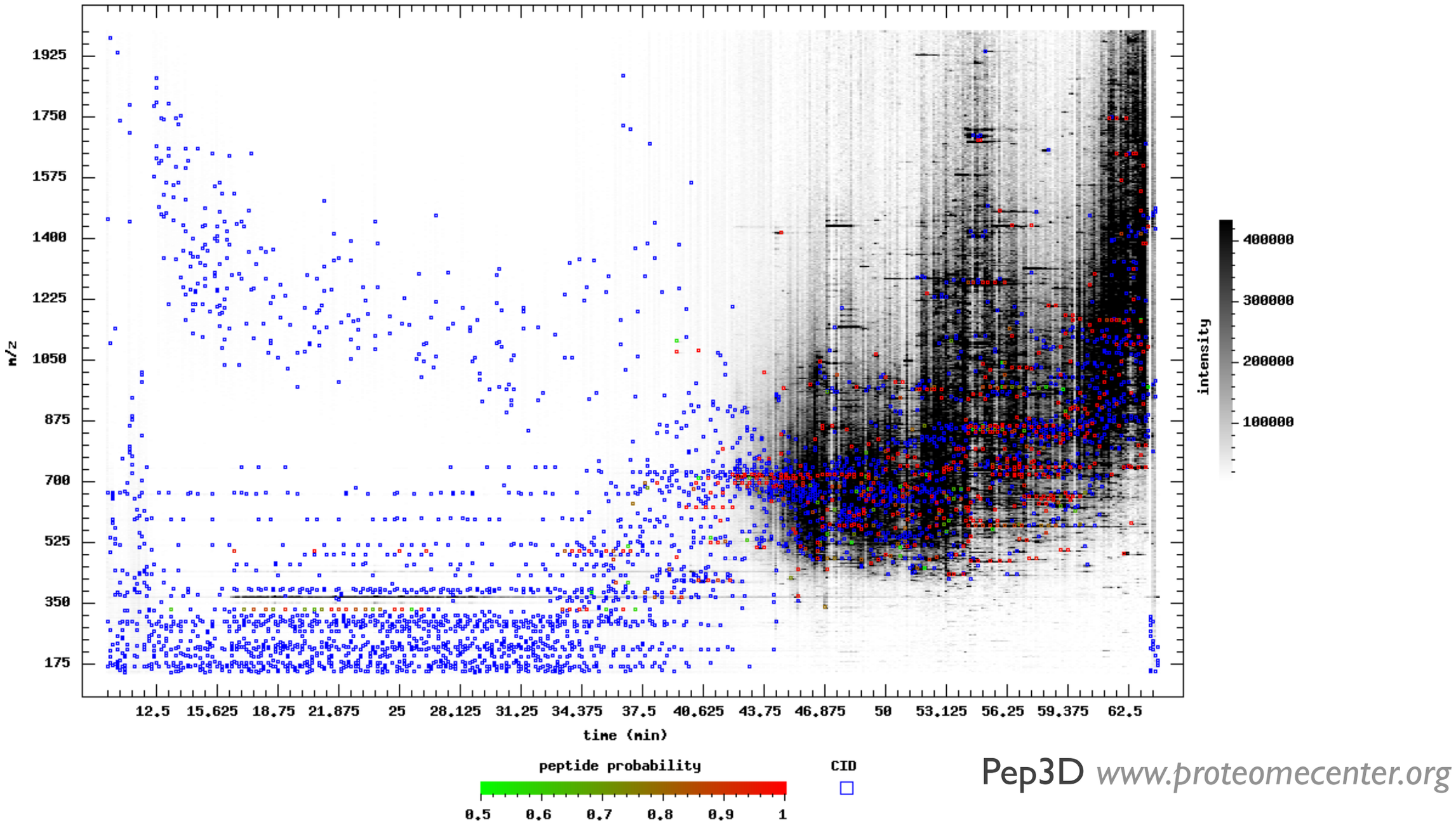


Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Transcriptomics



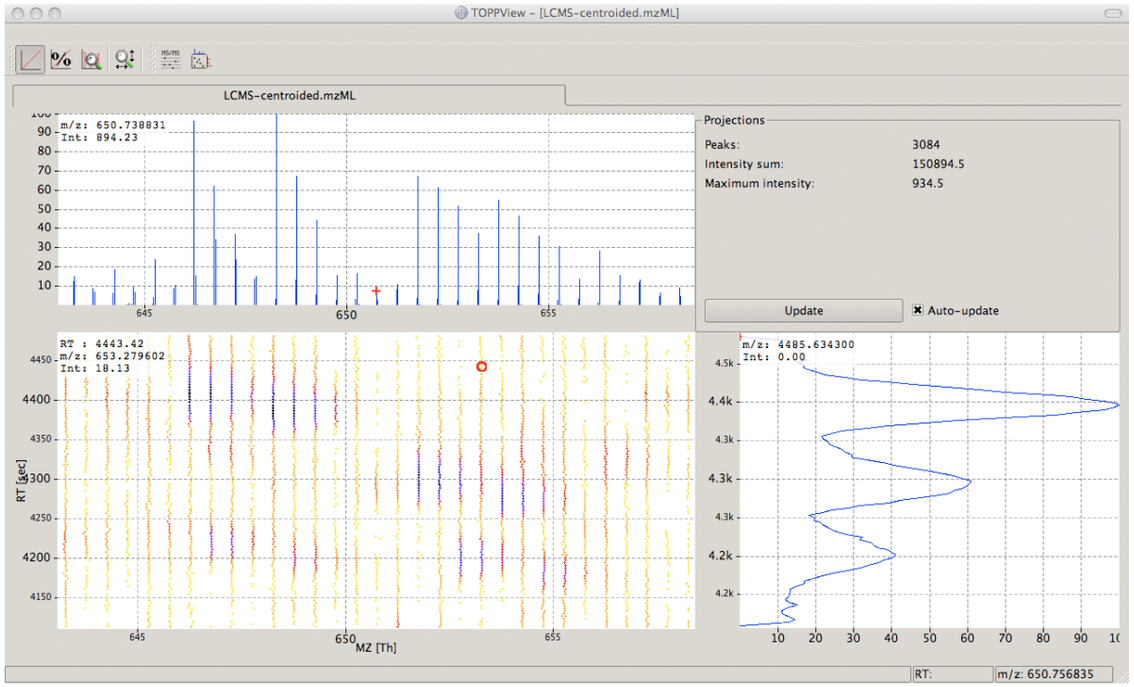
1 = before normalization
2 = after within-array normalization
3 = after between-array normalization



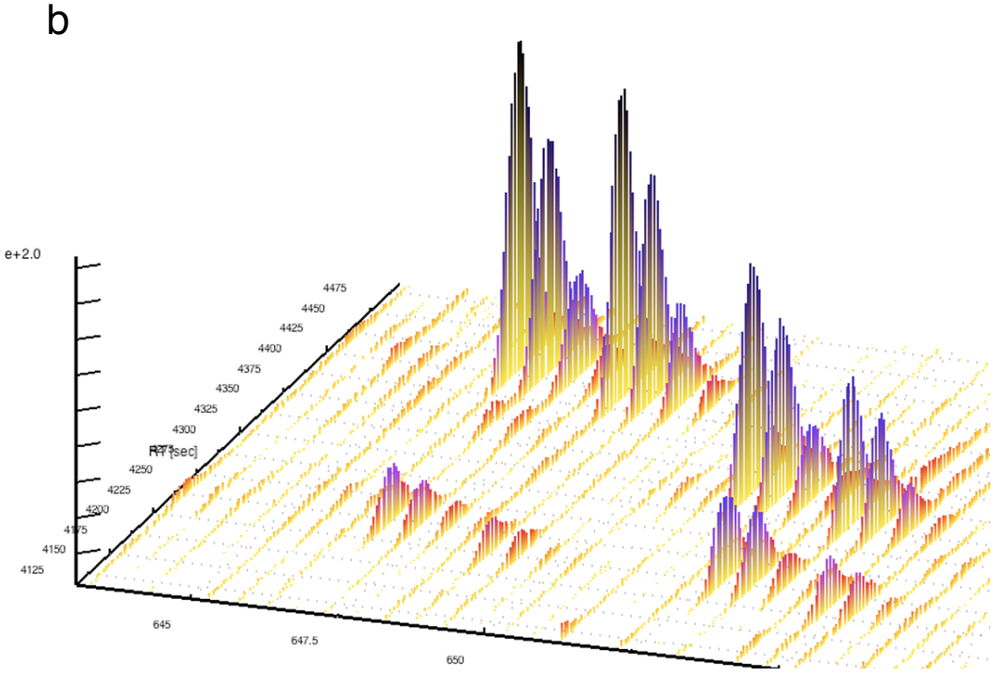
Pep3D *www.proteomecenter.org*

Multivariate Data: Proteomics

a

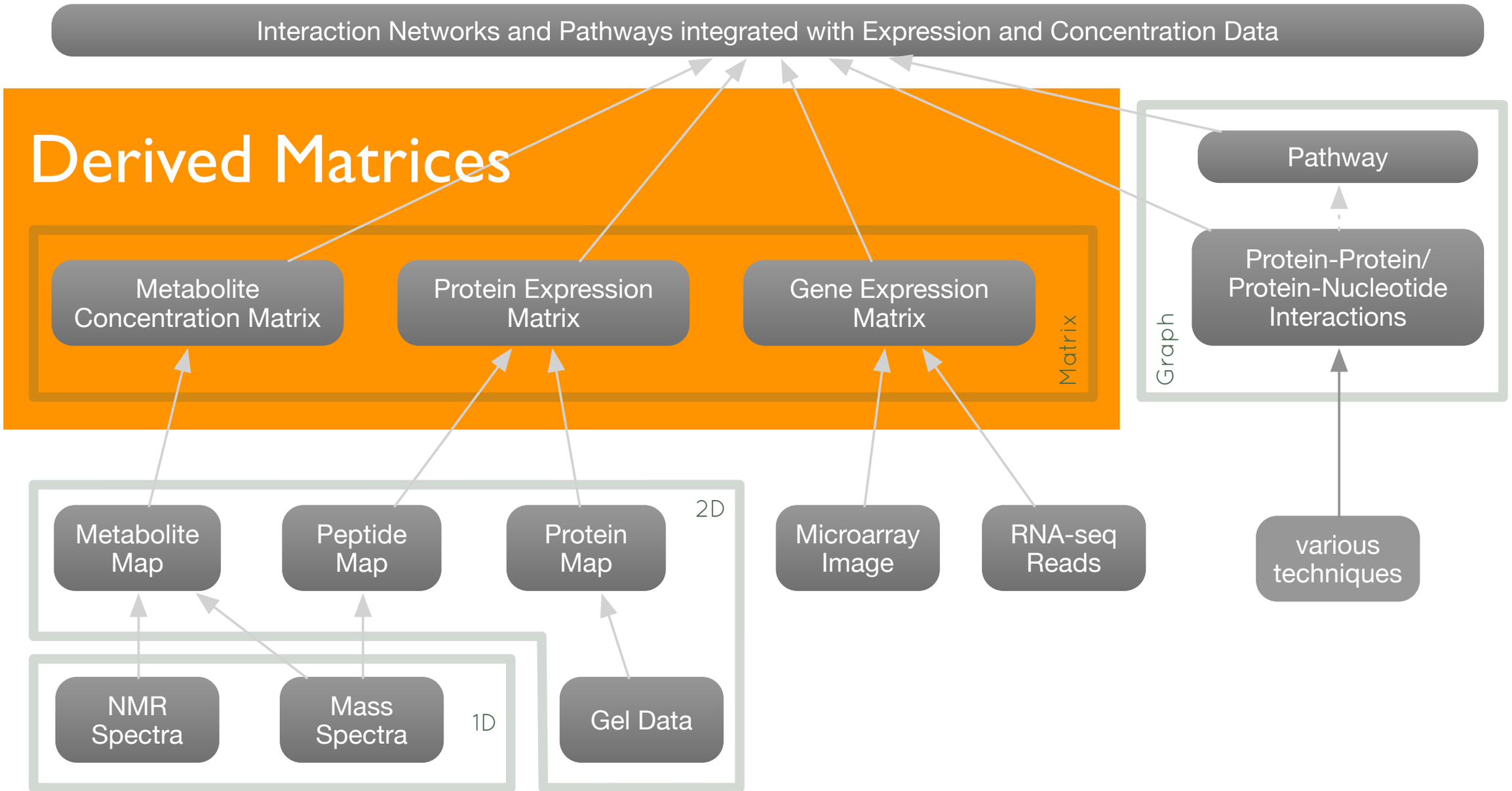


b



TOPPView *www.open-ms.de*

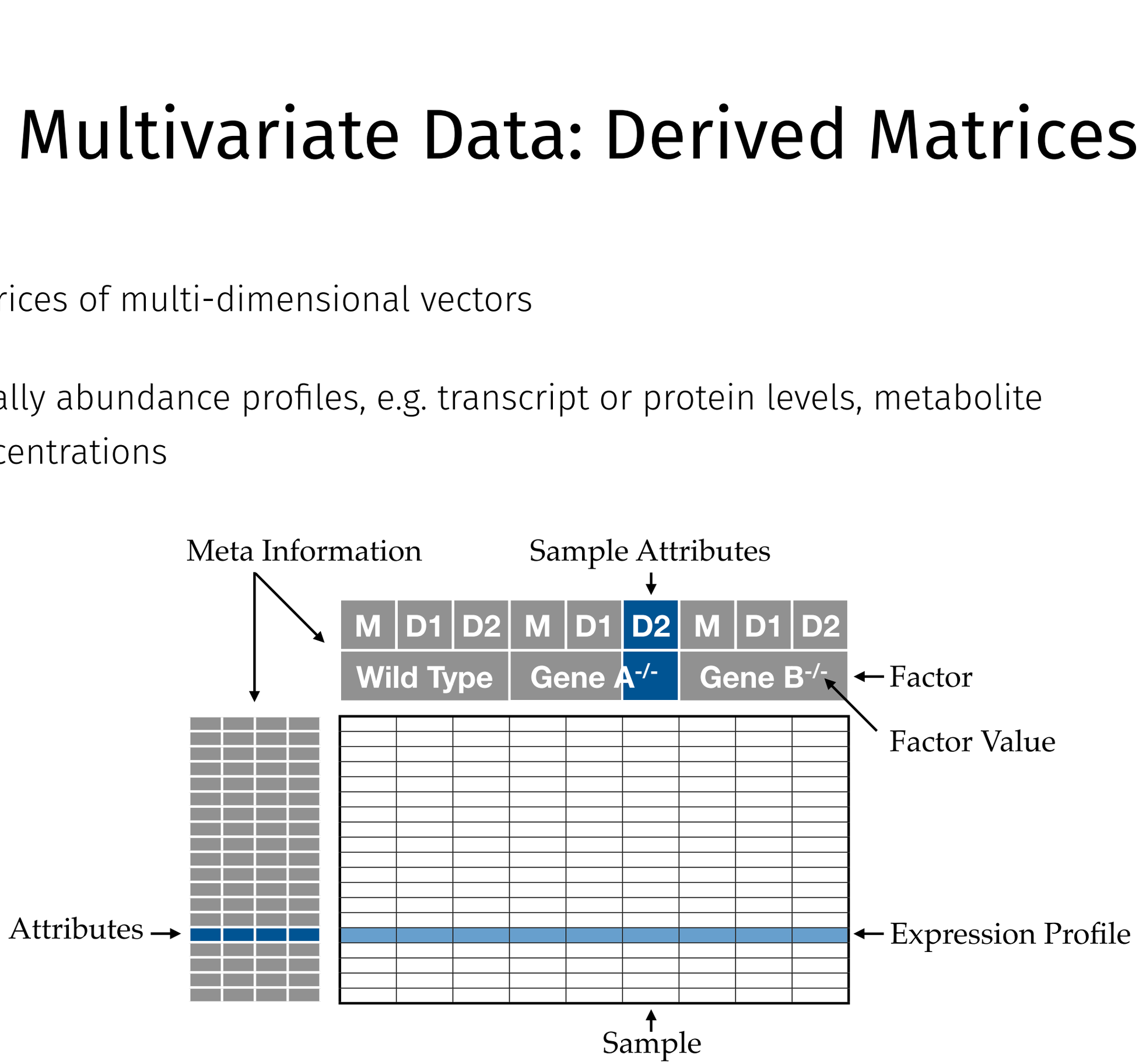
Multivariate Data: Derived Data



Multivariate Data: Derived Matrices

- matrices of multi-dimensional vectors

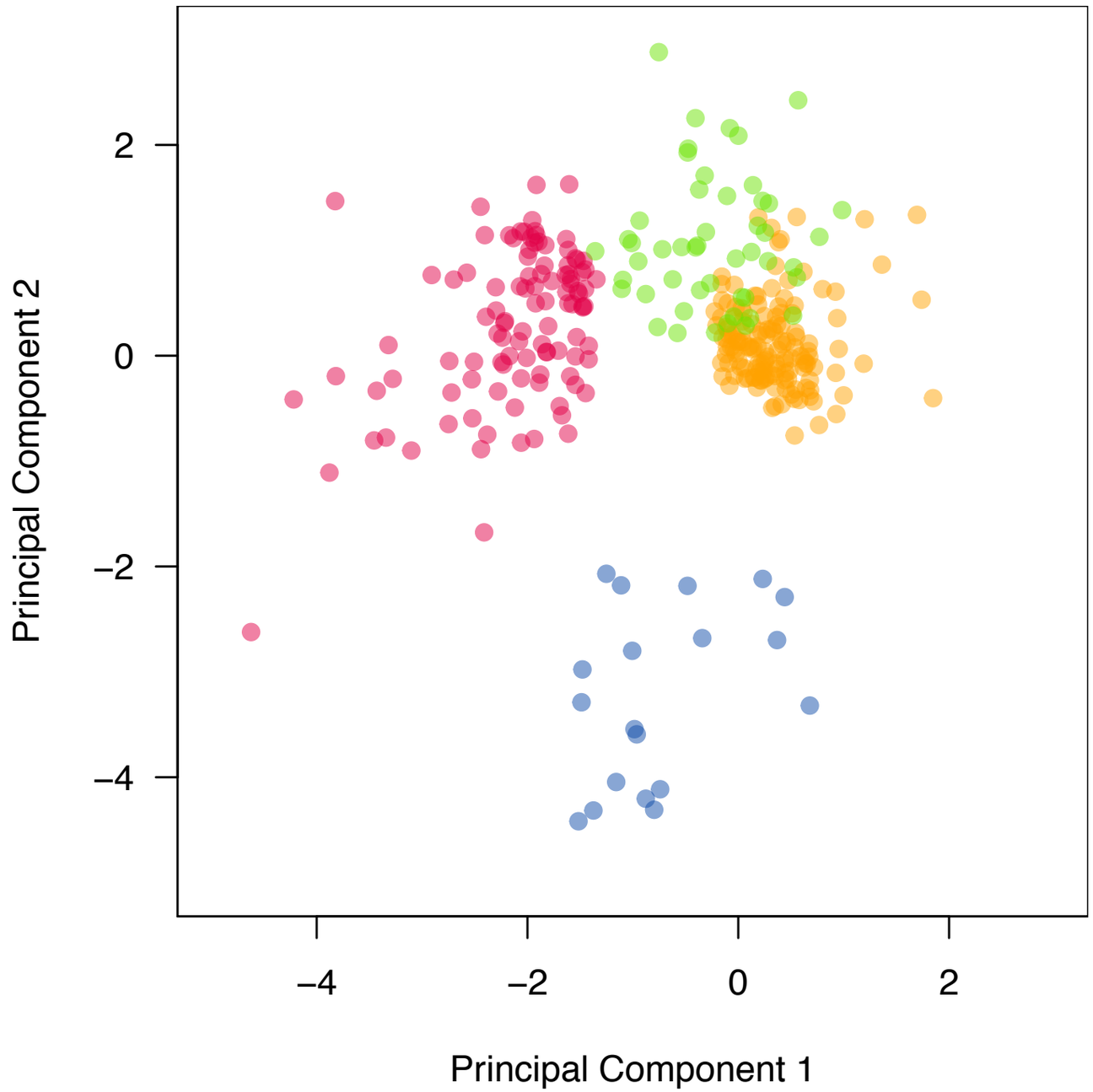
- usually abundance profiles, e.g. transcript or protein levels, metabolite concentrations



Gehlenborg et al. 2010, Nature Methods

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.

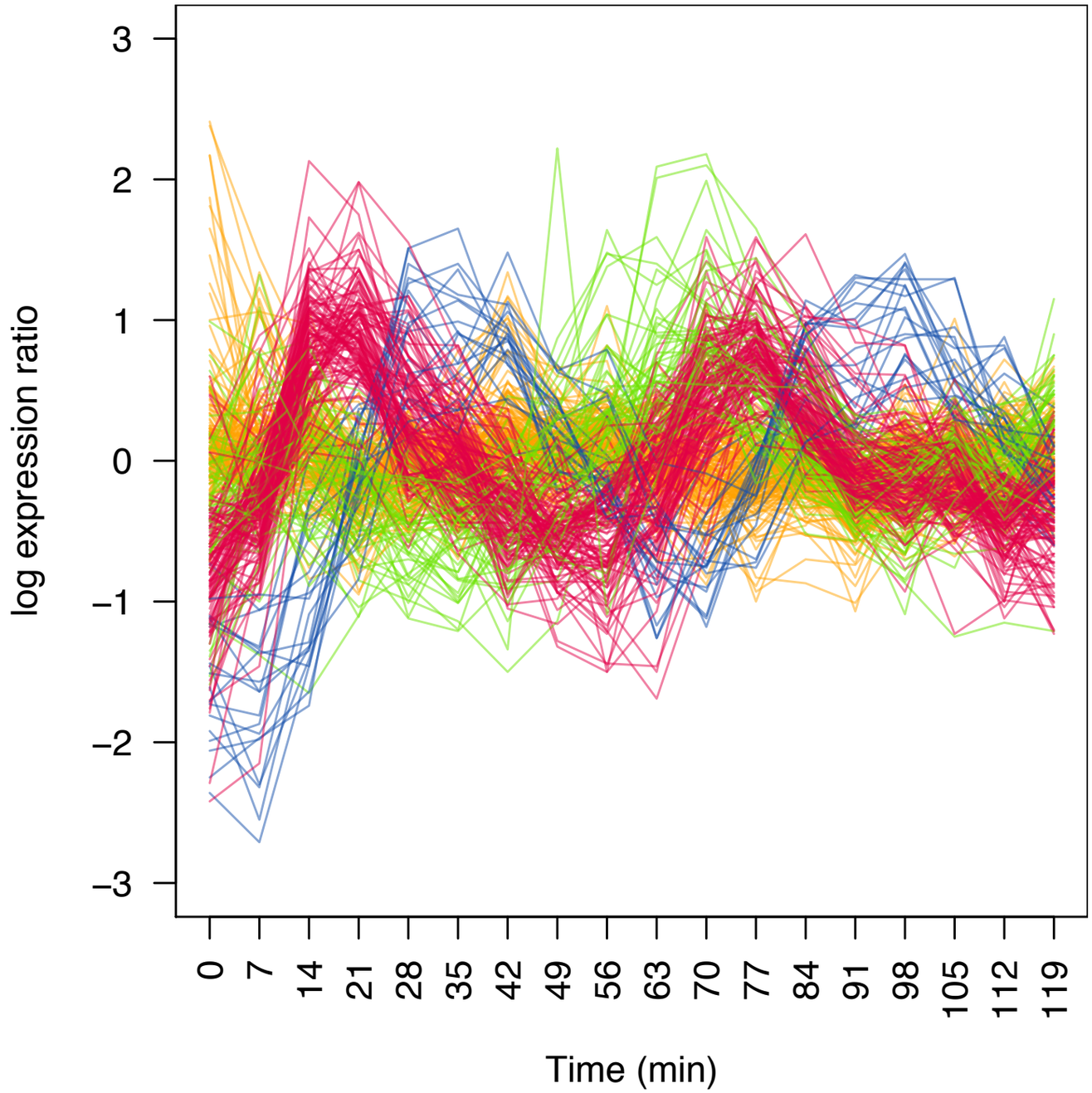
- **Pros** - good choice to get an idea about the overall structure of the whole data set: clusters, outliers, gaps in the data

- **Cons** - because of the dimensionality reduction the original profiles are not accessible in the visualization

Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

Profile Plot *a.k.a.* Parallel Coordinates



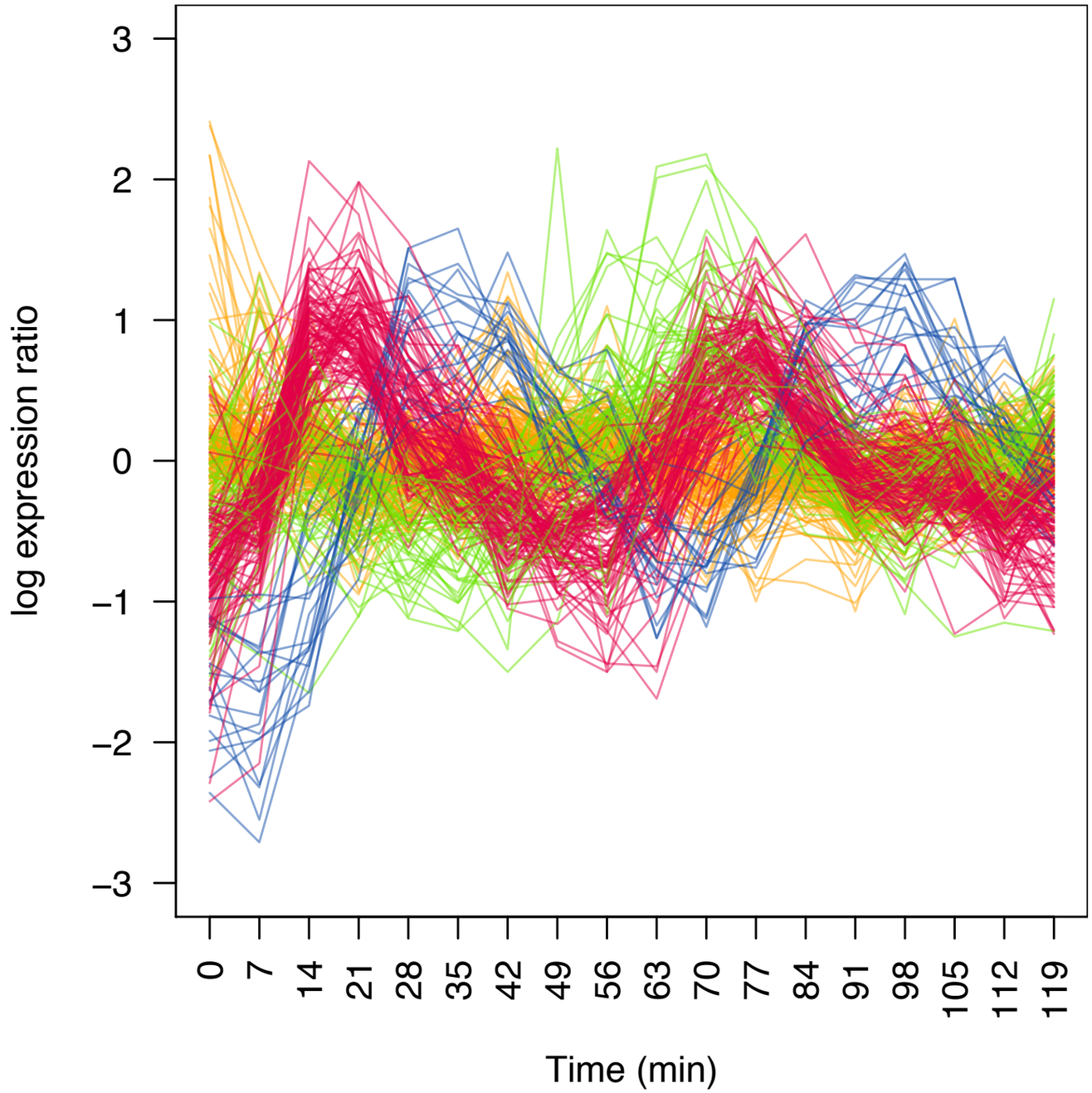
Multivariate Data: Derived Matrices

- **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

Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

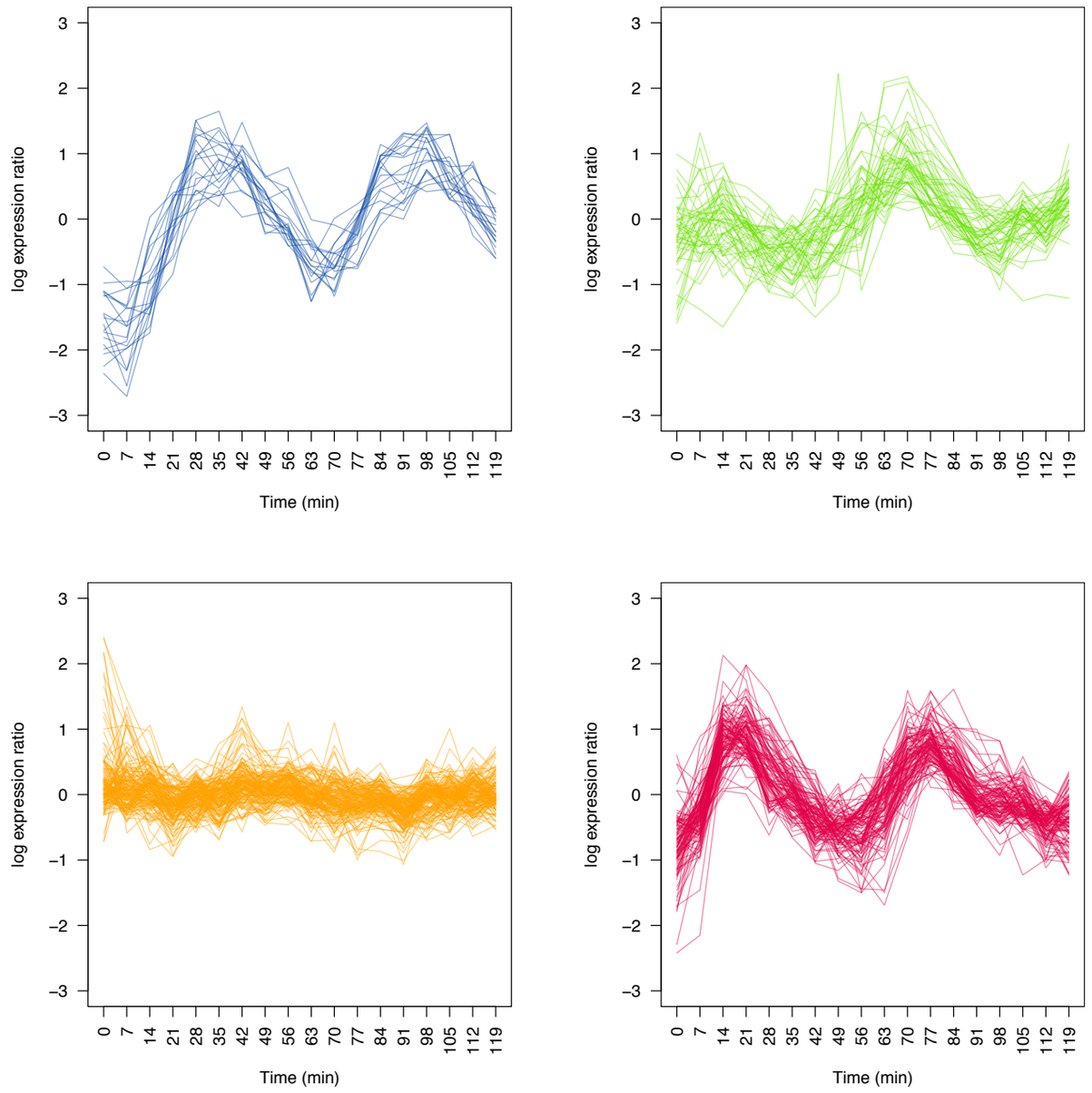
Profile Plot *a.k.a.* Parallel Coordinates



Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

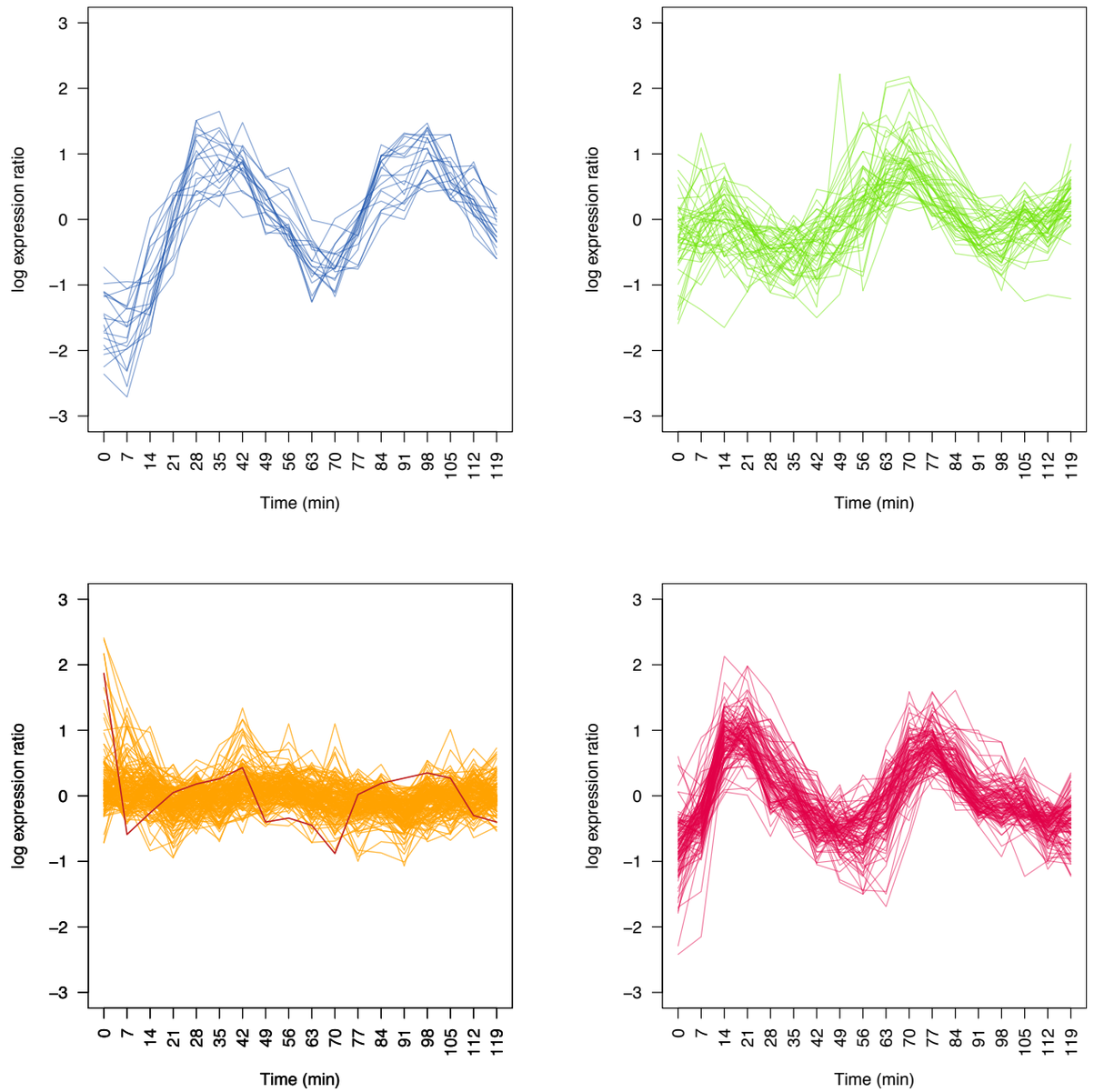
Profile Plot *a.k.a.* Parallel Coordinates



Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

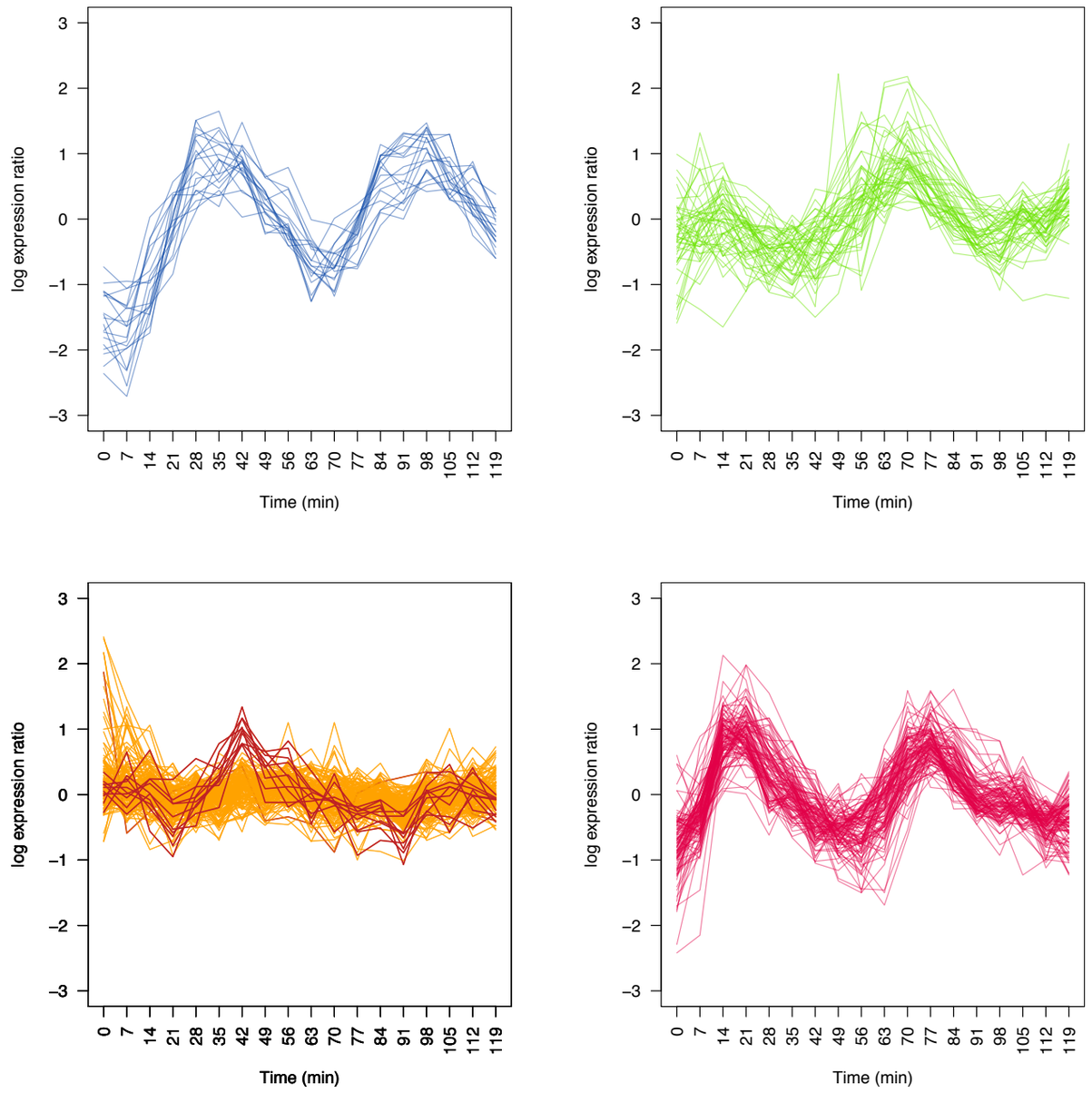
Profile Plot *a.k.a.* Parallel Coordinates



Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

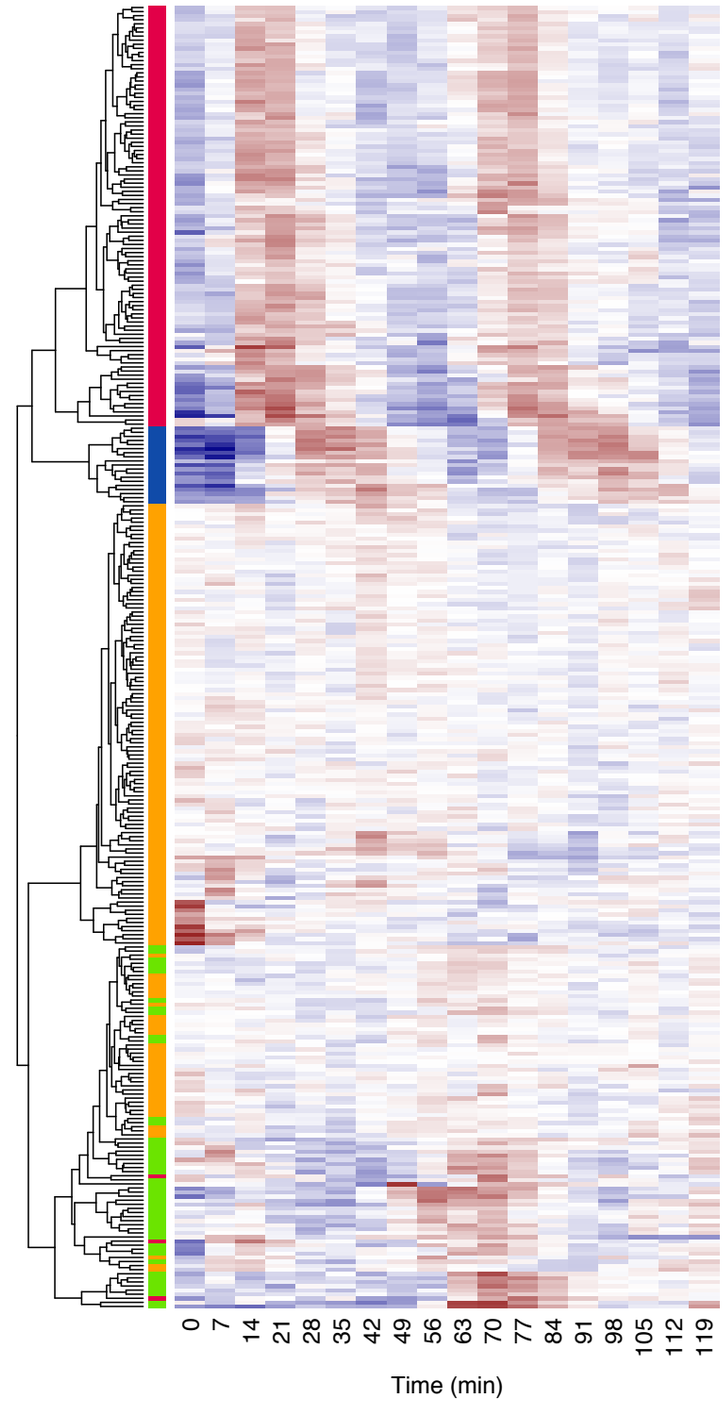
Profile Plot *a.k.a.* Parallel Coordinates



Gehlenborg et al. 2010, Nature Methods

Multivariate Data: Derived Matrices

Heat Map with Dendrogram



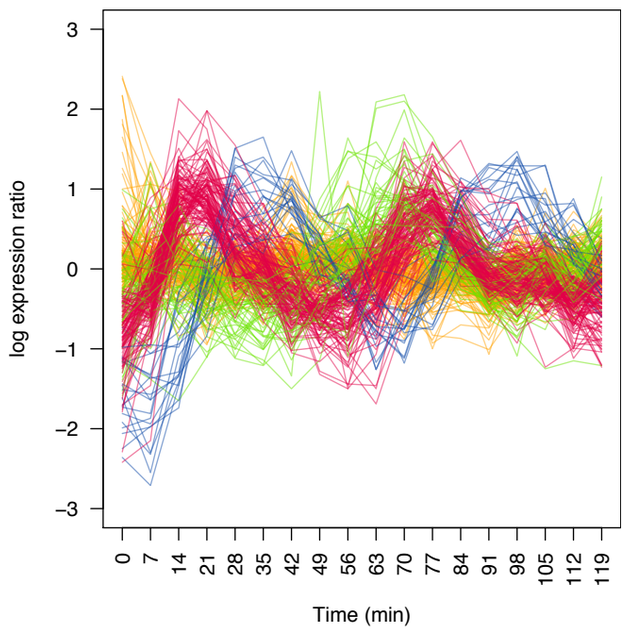
Multivariate Data: Derived Matrices

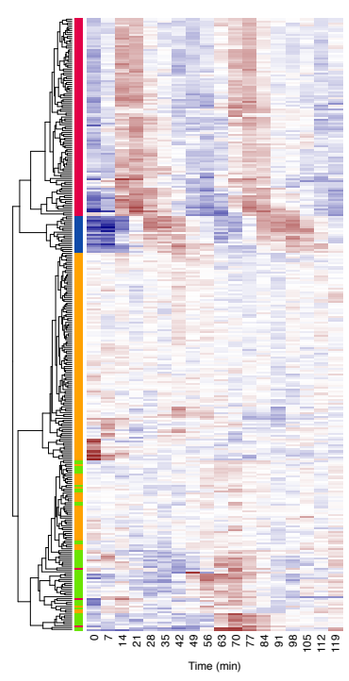
- **Heatmap**
  - **Pros**
    - no overplotting, yet a very dense information display
    - can be combined with dendrogram and additional information can be encoded in further columns or in the height of rows
- **Cons**
  - only qualitative interpretation possible due to color coding
  - grows horizontally with every additional sample and grows vertically with every additional profile

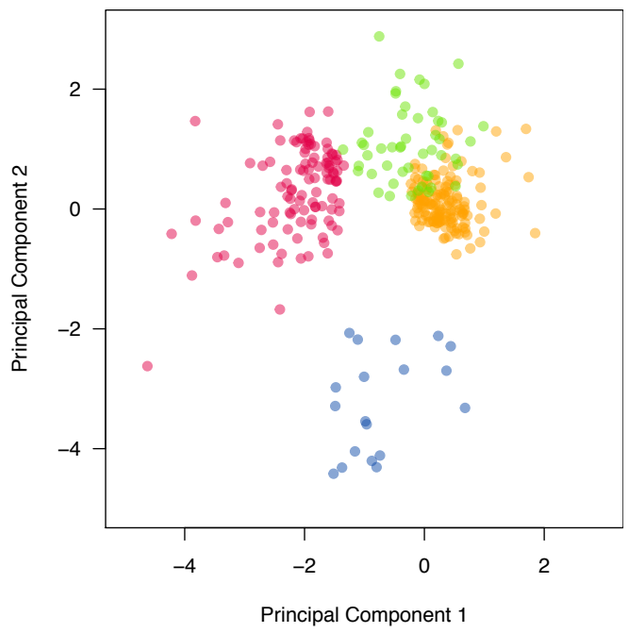
Multivariate Data: Summary

few, high-res

many, low-res

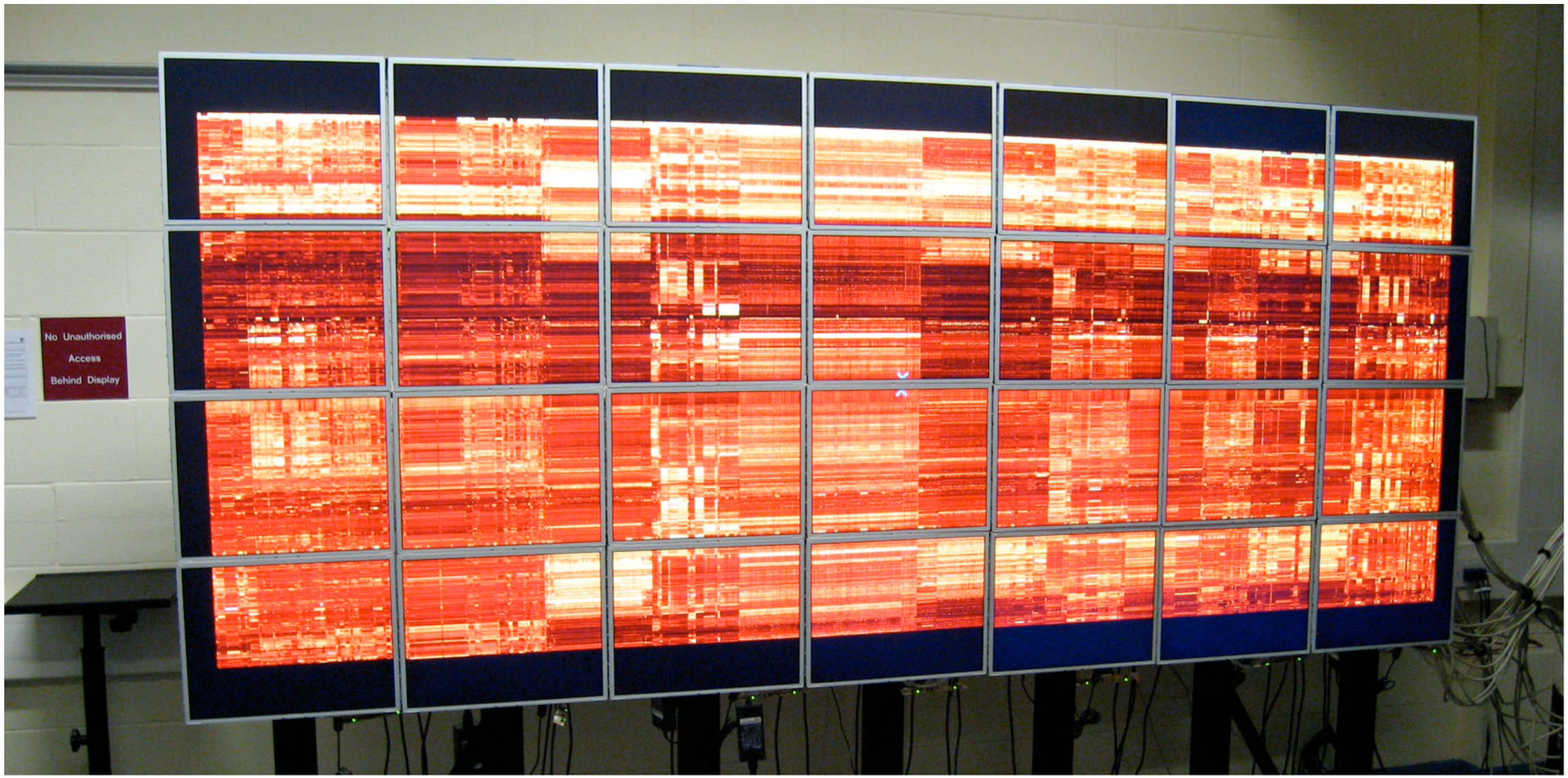






data: Lukk et al., *2010, Nature Biotechnology*

Problem: Very Large Expression Matrices



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

1000 transcripts, 5372 samples

Multivariate Data

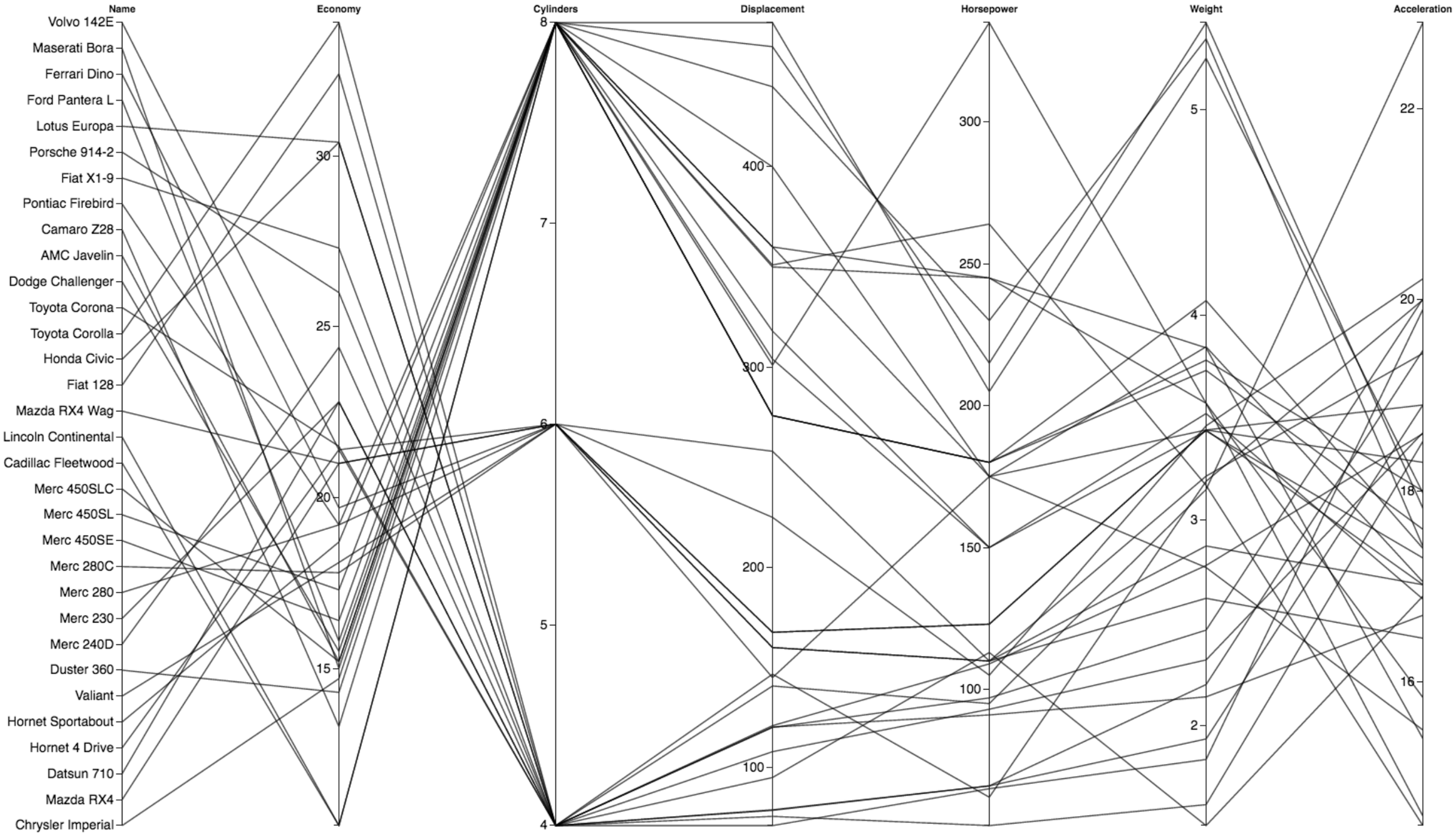
Heterogeneous Tables

parallelcoordinates.de

Nils

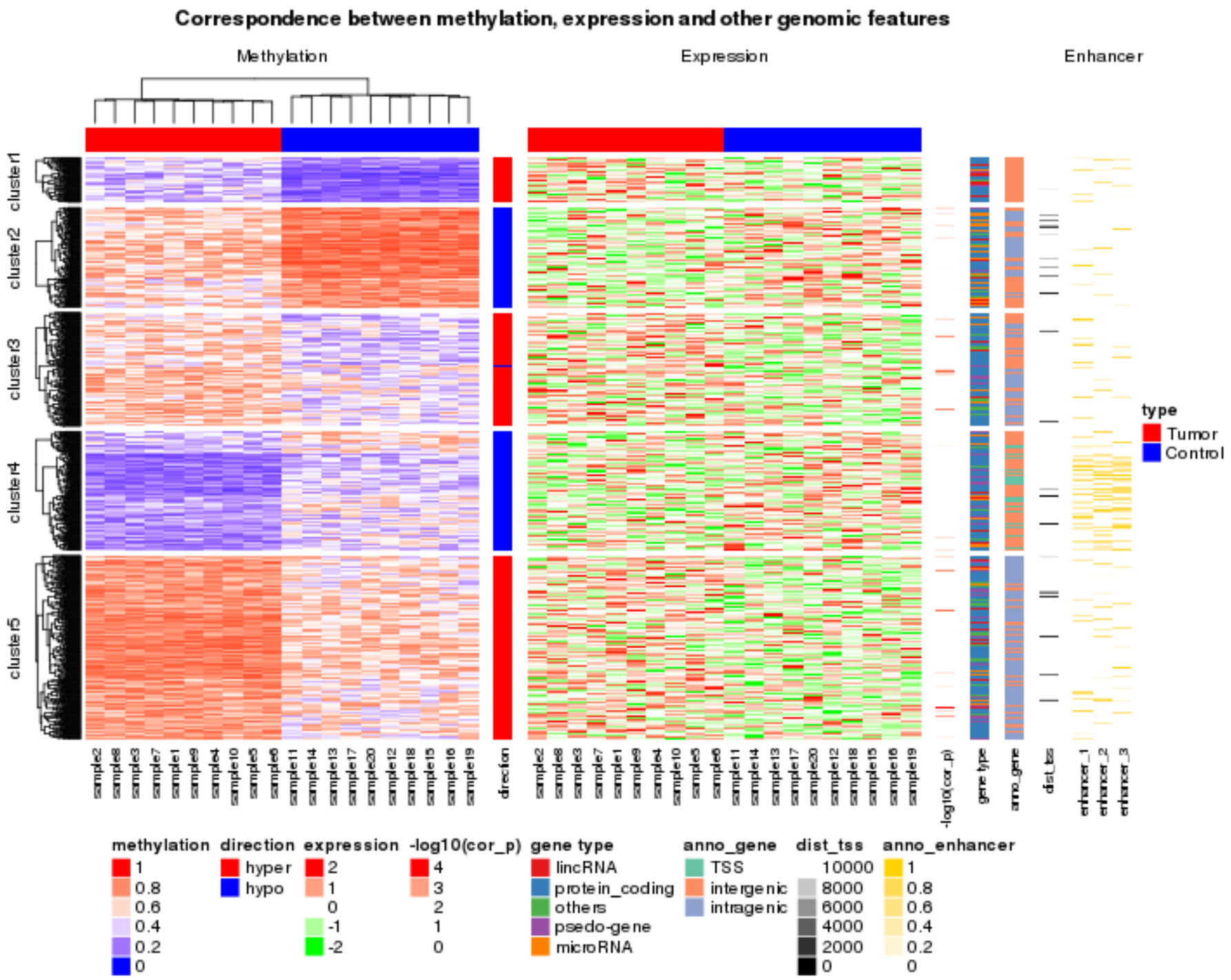
www.parallelcoordinates.de/paco/

parallelcoordinates.de | 1D | 2D | A | Color | ✖ | Density | Bundling | PCA | Clustering | Dimensions

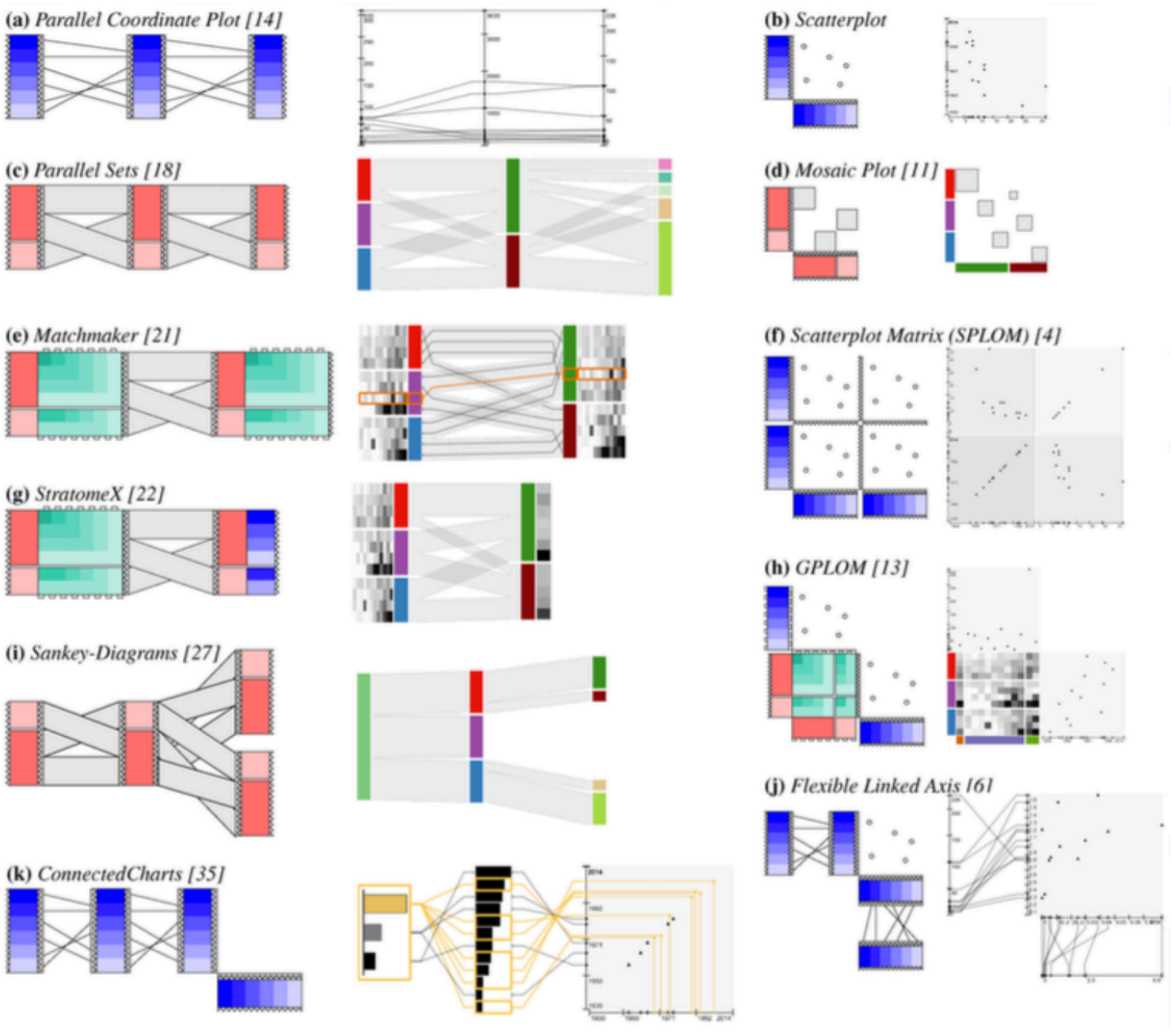


drop a csv-formatted file to load your own data. Note that the first line must describe the data scheme. See this dataset for an example.

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


Correspondence between methylation, expression and other genomic features
Methylation
Expression
Enhancer
cluster1
cluster2
cluster3
cluster4
cluster5
sample2
sample8
sample3
sample7
sample1
sample9
sample4
sample10
sample5
sample6
sample11
sample14
sample13
sample17
sample20
sample12
sample18
sample15
sample16
sample19
direction
-log10(cor_p)
gene type
anno_gene
dist_tss
enhancer_1
enhancer_2
enhancer_3
type
Tumor
Control
methylation
1
0.8
0.6
0.4
0.2
0
direction
hyper
hypo
expression
2
1
0
-1
-2
-log10(cor_p)
4
3
2
1
0
gene type
lincRNA
protein_coding
others
psedo-gene
microRNA
anno_gene
TSS
intergenic
intragenic
dist_tss
10000
8000
6000
4000
2000
0
anno_enhancer
1
0.8
0.6
0.4
0.2
0

Domino


(a) Parallel Coordinate Plot [14]
(b) Scatterplot
(c) Parallel Sets [18]
(d) Mosaic Plot [11]
(e) Matchmaker [21]
(f) Scatterplot Matrix (SPLOM) [4]
(g) StratomeX [22]
(h) GPLOM [13]
(i) Sankey-Diagrams [27]
(j) Flexible Linked Axis [61]
(k) ConnectedCharts [35]

Domino

Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets

CALEYDO

JKU JOHANNES KEPLER UNIVERSITY LINZ

HARVARD MEDICAL SCHOOL

HARVARD School of Engineering and Applied Sciences

Samuel Gratzl

Marc Streit

Nils Gehlenborg

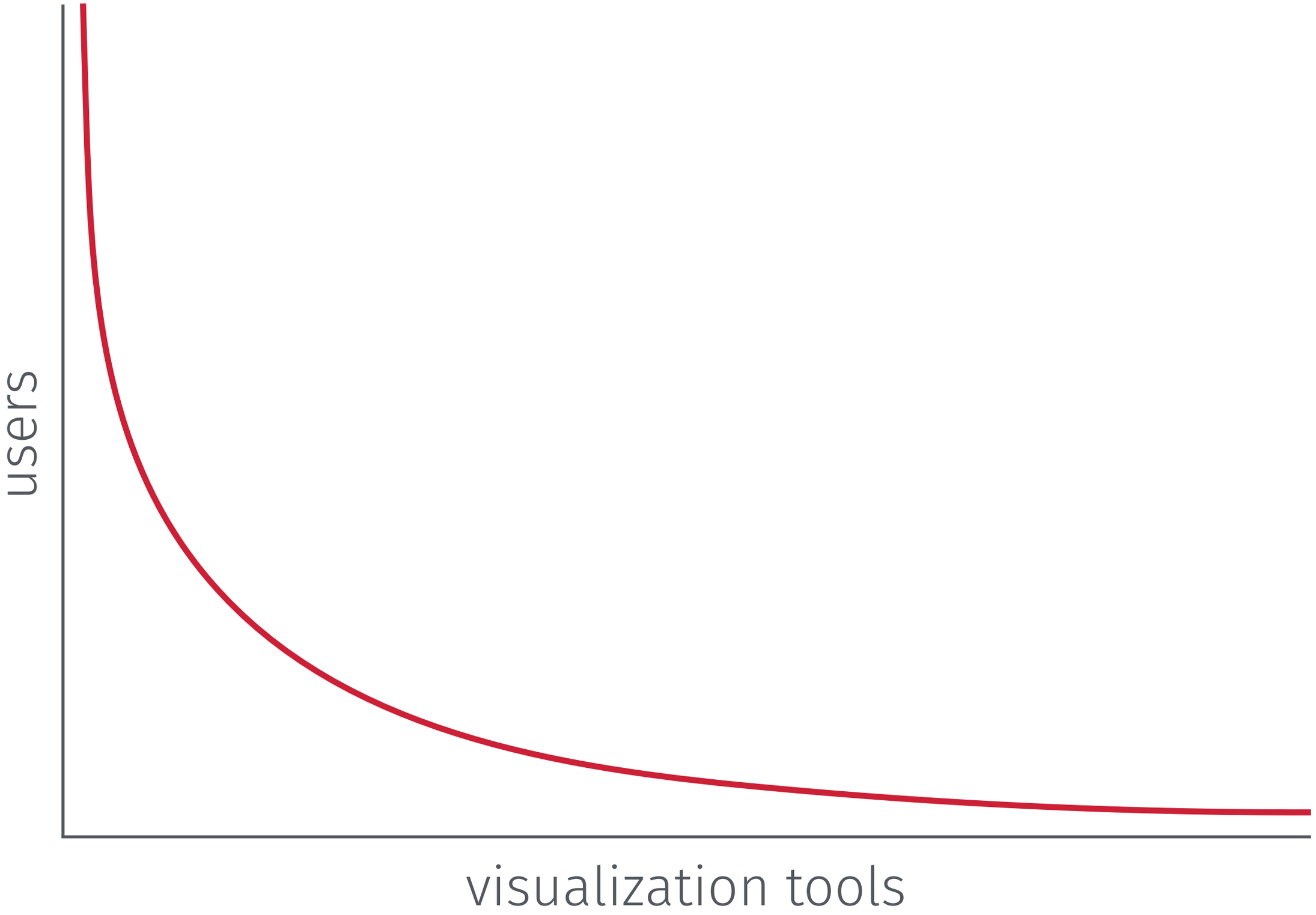
Alexander Lex

Hanspeter Pfister

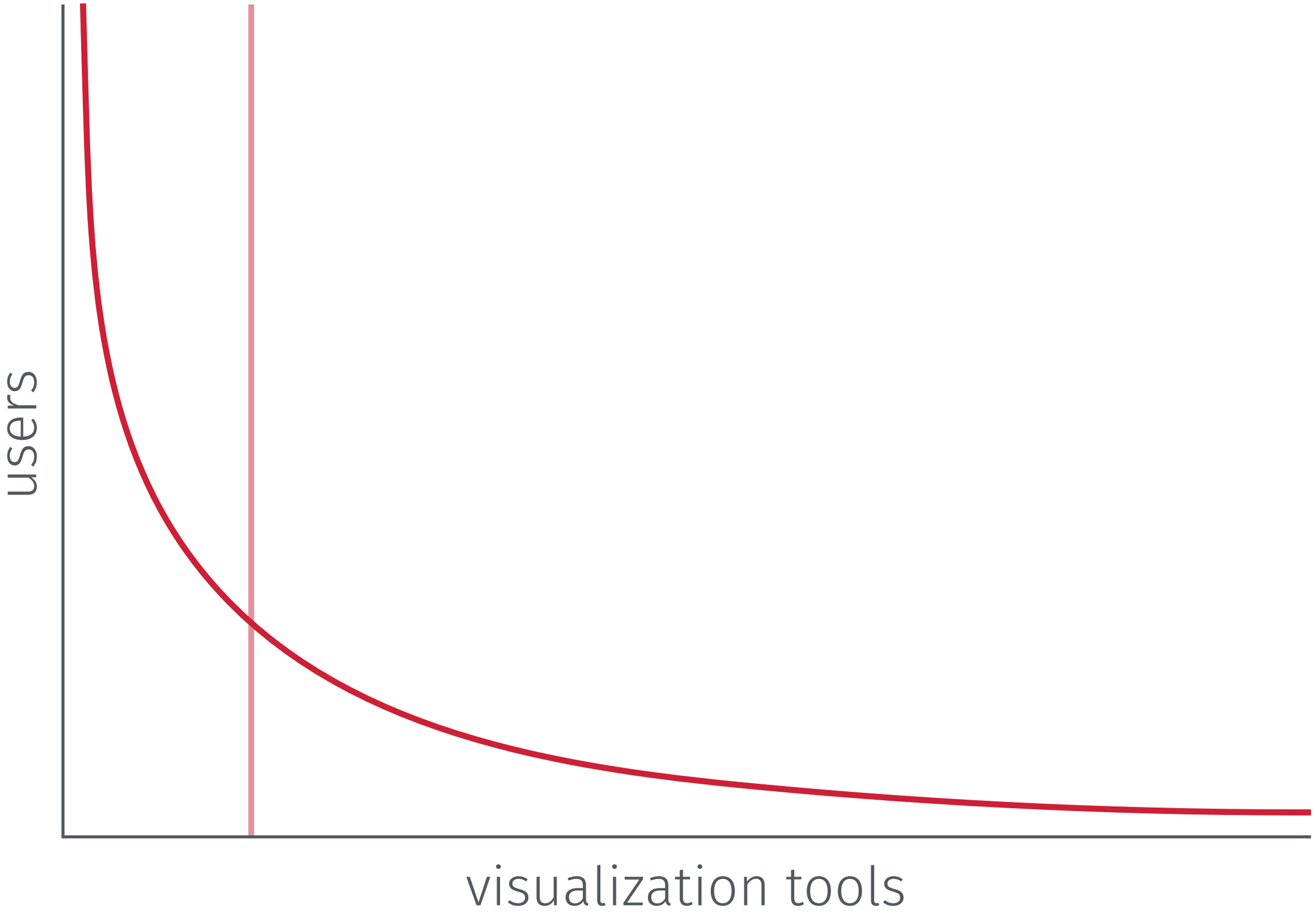
https://sgratzl.github.io/domino_vis2014/#/title

The Long Tail

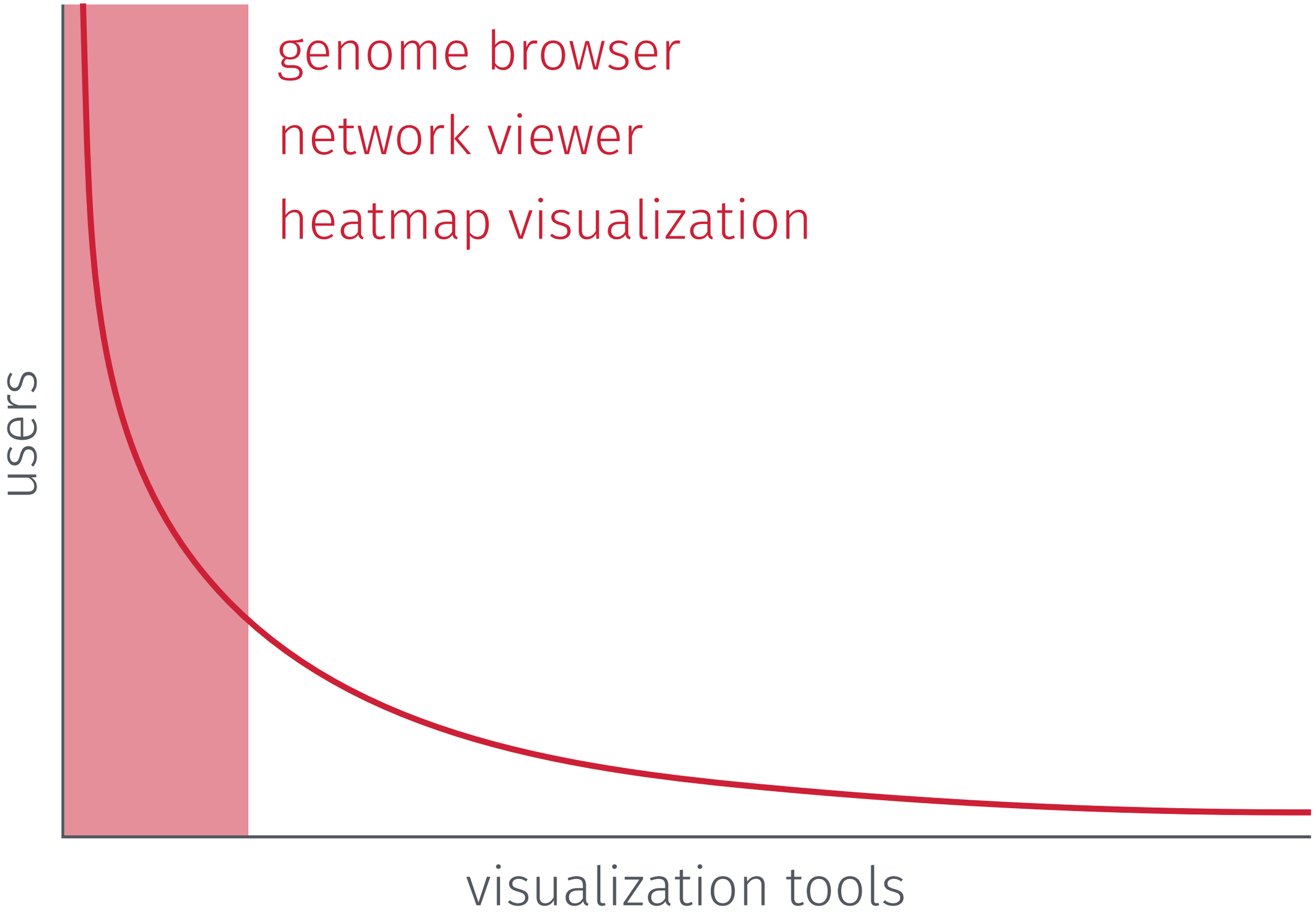
Data Visualization: The Long Tail



Data Visualization: The Long Tail



Data Visualization: The Long Tail



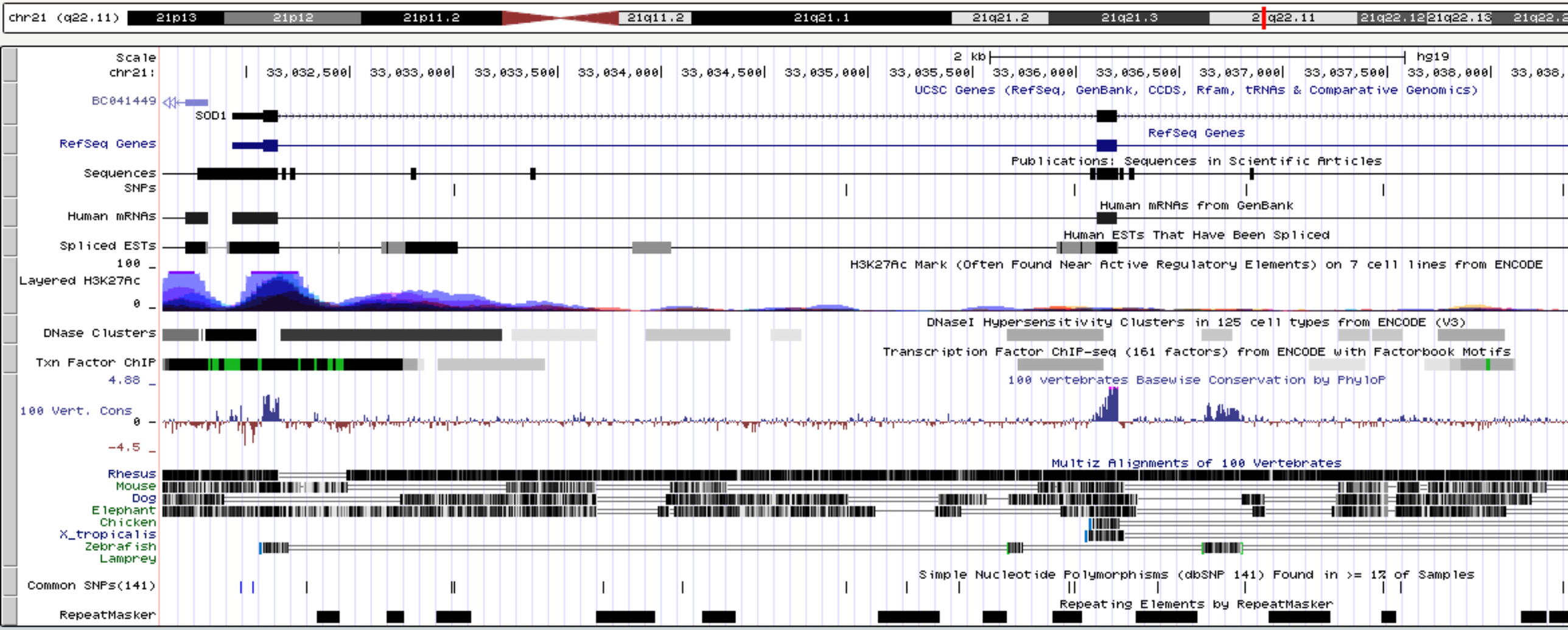
genome-euro.ucsc.edu

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | View | Help | About Us

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr21:33,031,597-33,041,570 9,974 bp. enter position, gene symbol or search terms go



move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or do new position.

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | reverse | resize | refresh

collapse all

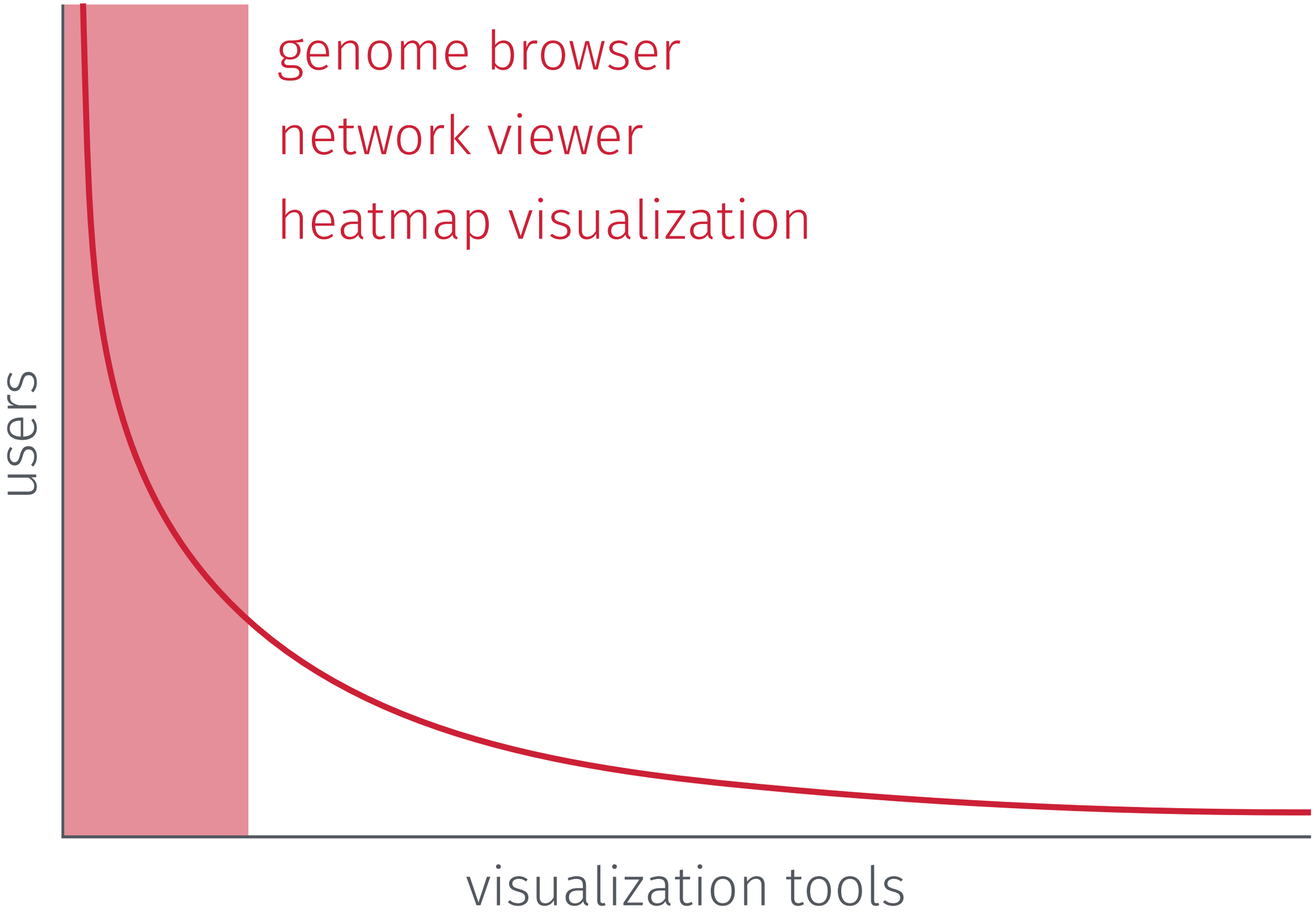
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

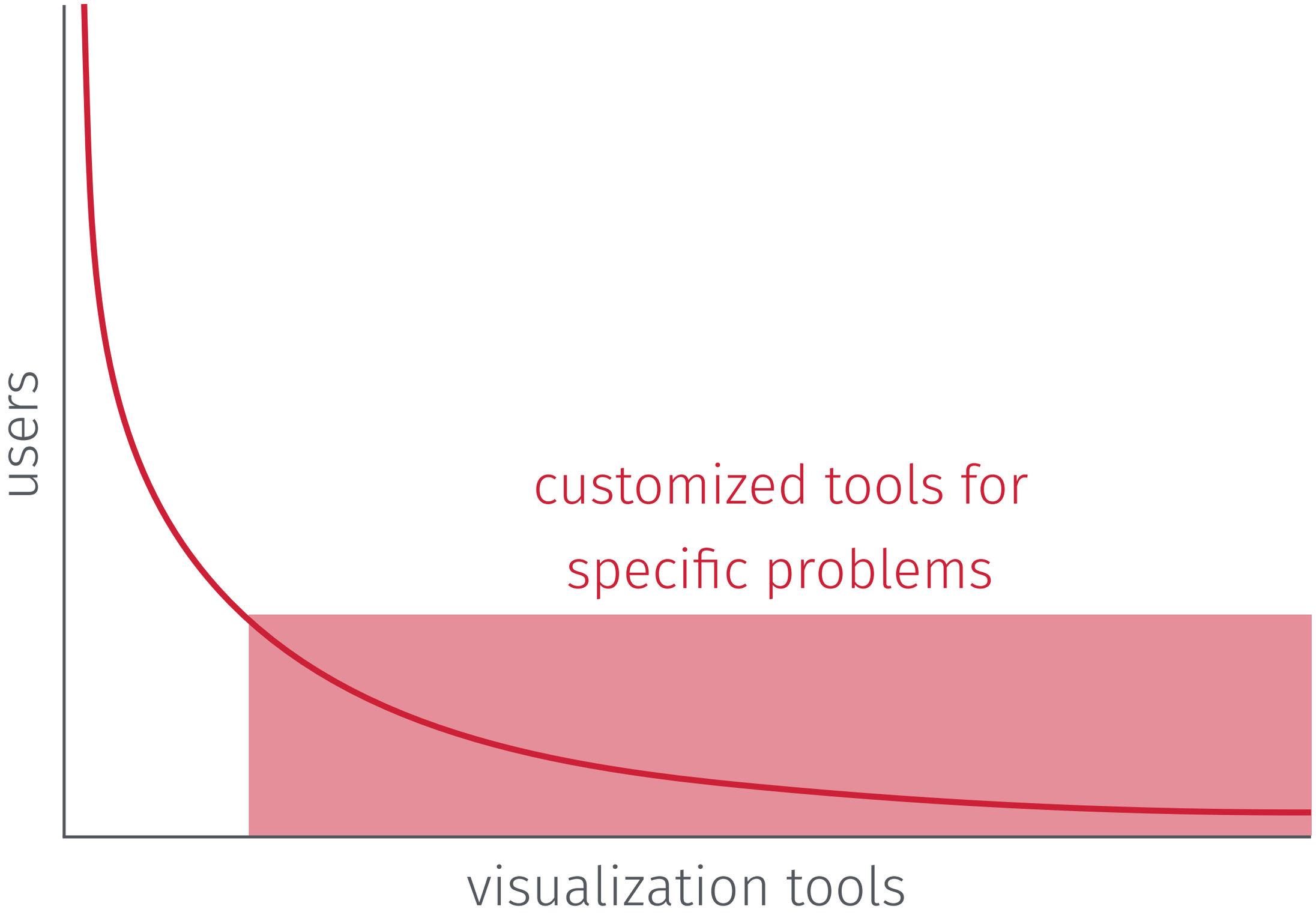
Mapping and Sequencing

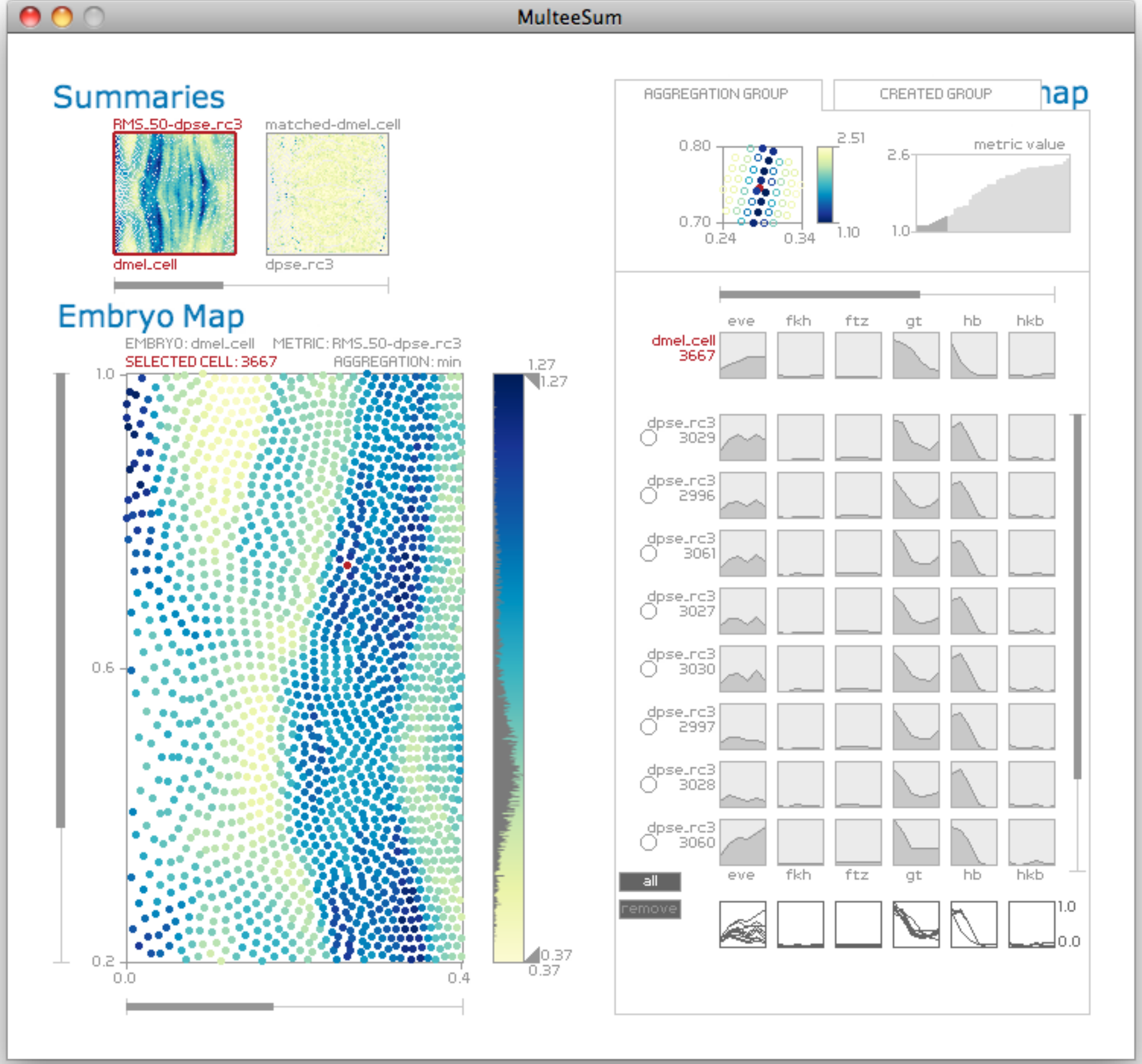
Display a menu for "http://genome-euro.ucsc.edu/cgi-bin/hgc?hgsid=200830848_Pr7vSRBAuNDdCe9FN6cl81oDcNRD&c=chr21&o=33031596&t=33041570&g=phyloP100wayAll&i=phyloP100wayAll"

Data Visualization: The Long Tail



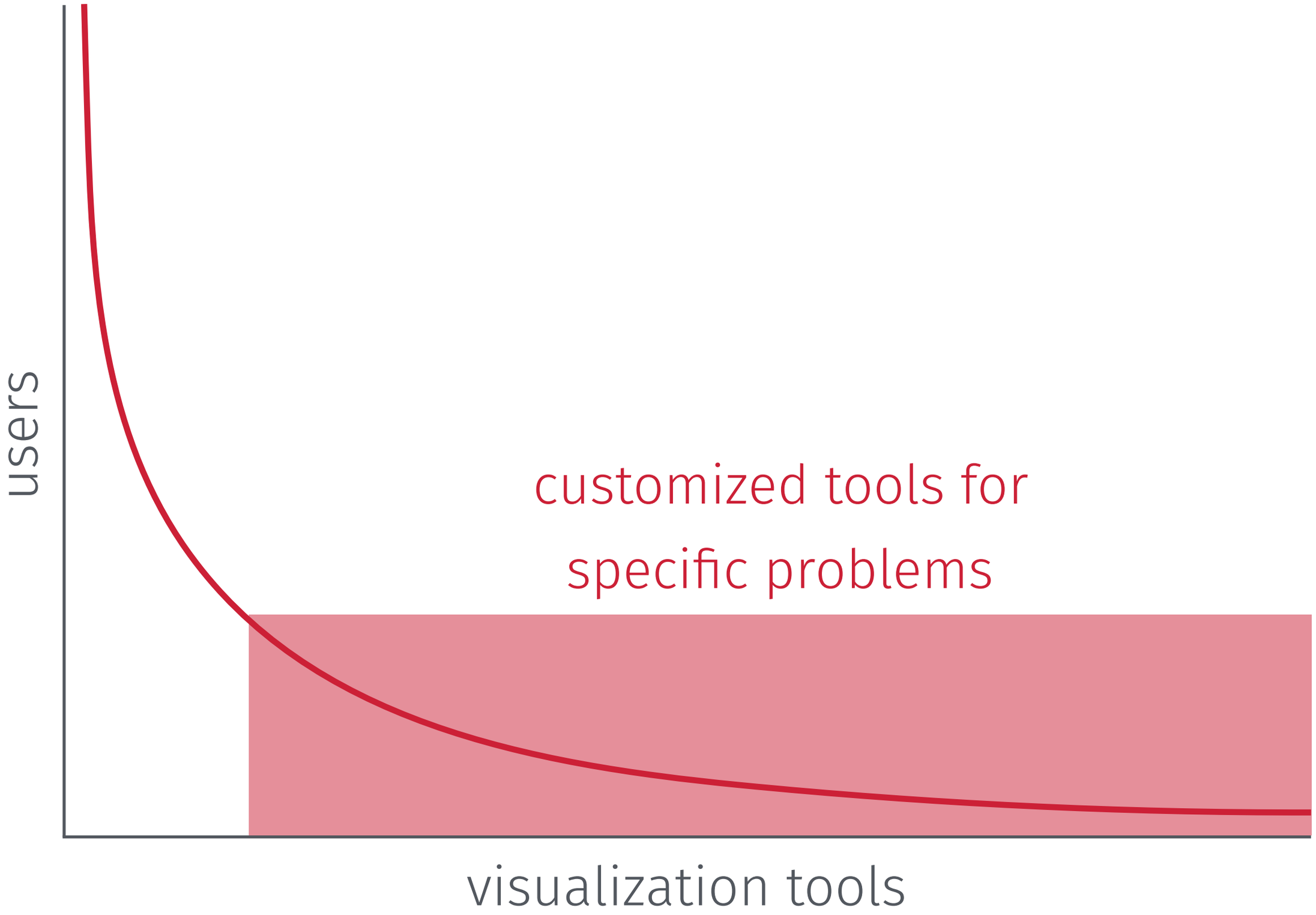
Data Visualization: The Long Tail





Meyer et al., “MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data”, 2010

Data Visualization: The Long Tail



Data Visualization: The Long Tail

Infrastructure is required to enable design and implementation of customized tools within reasonable timeframes!