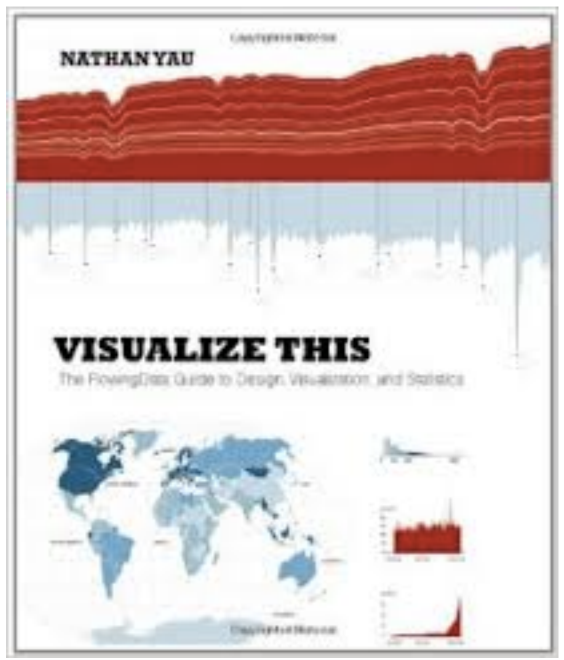
[Biological Data Visualization Workshop] Introduction to **Visual Data Analysis**

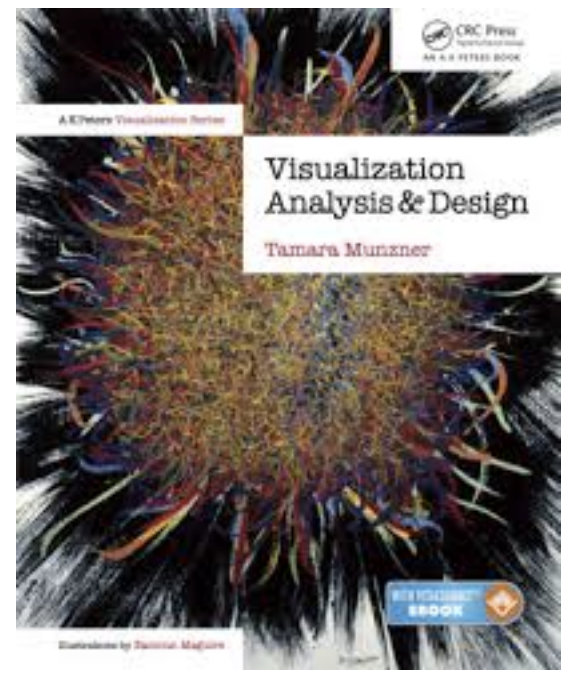
Prof Jan Aerts
Visual Data Analysis lab, ESAT/STADIUS
Faculty of Engineering
KU Leuven

@jandot - jan.aerts@kuleuven.be - http://vda-lab.be

TAs: Jansi Thiyagarajan, Daniel Alcaide

Books





basis for many slides

Overview of lecture

A. **Why visual analytics?**

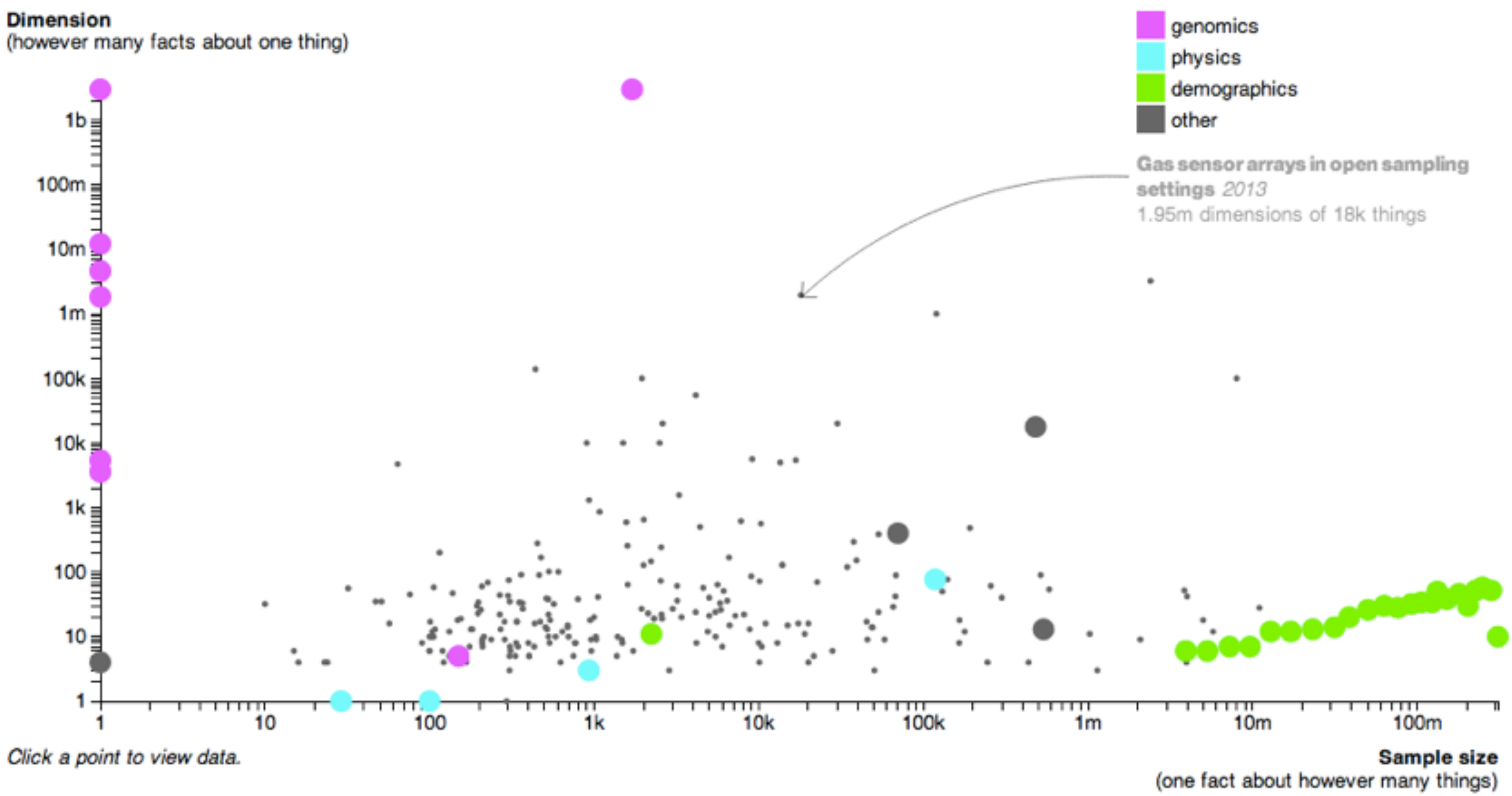
B. Data visualization

- Data foundations
- Human perception foundations
- Visualization foundations and examples

C. Visualization evaluation

D. Tools of the trade

A. What’s the problem?
(or: why are we looking at this in a MSc of Statistics?)
(or: about unknown unknowns and black boxes)


Dimension
(however many facts about one thing)
genomics
physics
demographics
other
Gas sensor arrays in open sampling settings 2013
1.95m dimensions of 18k things
1b
100m
10m
1m
100k
10k
1k
100
10
1
1
10
100
1k
10k
100k
1m
10m
100m
Click a point to view data.
Sample size
(one fact about however many things)

hypothesis-driven -> data-driven
hunting down unknown unknowns

Scientific Research Paradigms (Jim Gray, Microsoft)

1st	1,000s years ago	empirical
2nd	100s years ago	theoretical
3rd	last few decades	computational
4rd	today	data exploration

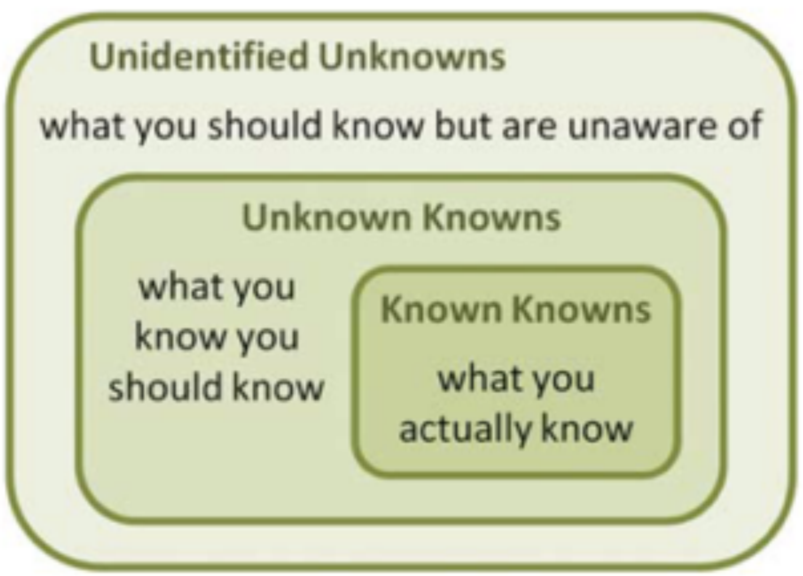
then: hypothesis → devise experiment → generate data

prove/disprove

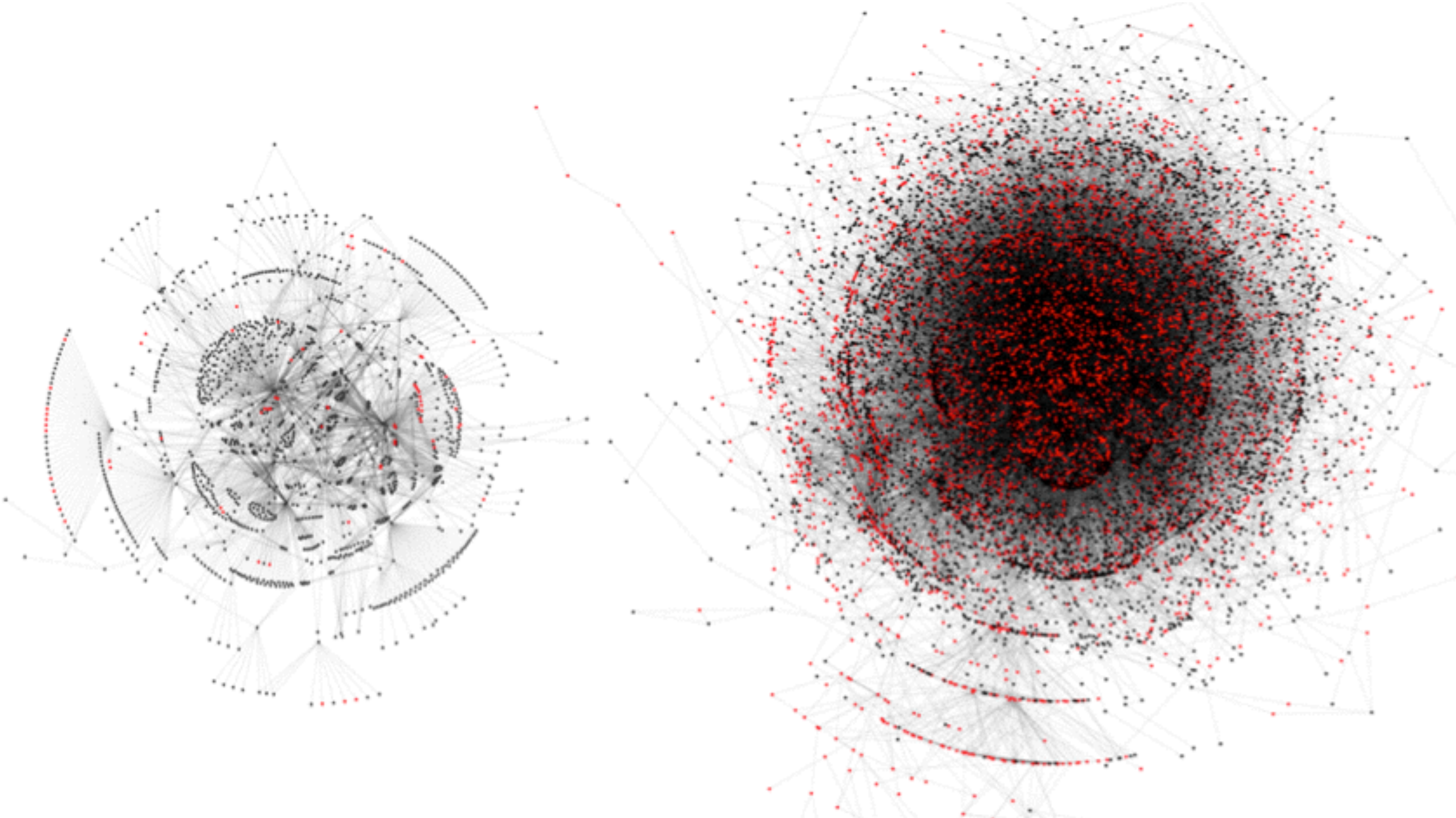
now: have data → what's my hypothesis?

Data analysis
moving from
hypothesis-first
to
data-first

Challenge
moving from
finding the right answer to a question
to
finding the right question given the data



known unknowns	unknown unknowns
automated analysis (machine learning, statistics, …)	visual data analysis
replace the human	augment the human



Martin Krzywinski

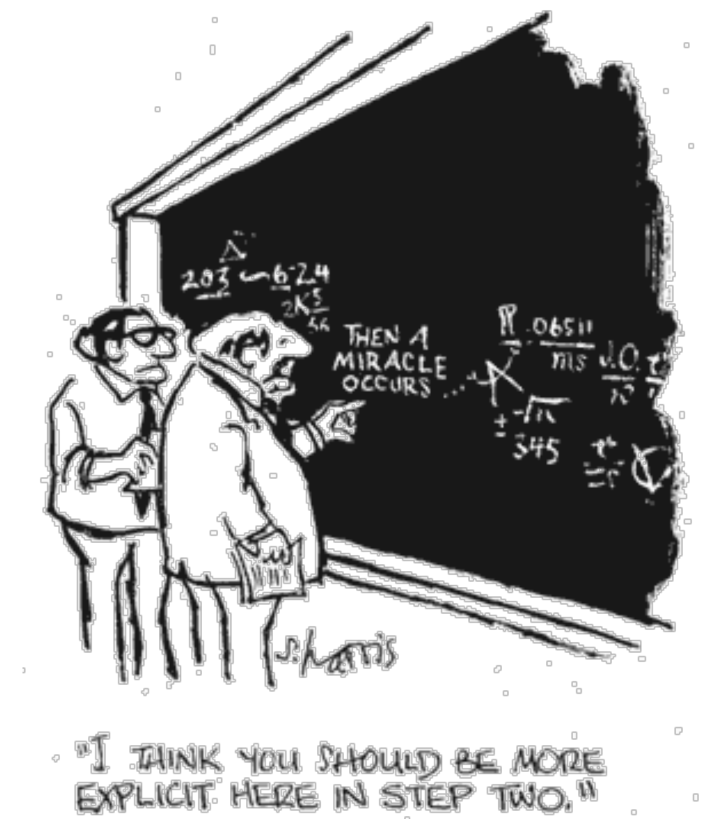
Opening the black box

complex algorithms
in data analysis

obvious **link between input and output** difficult to see

user needs to **blindly trust** data analyst (and data analyst needs to trust their own skills)

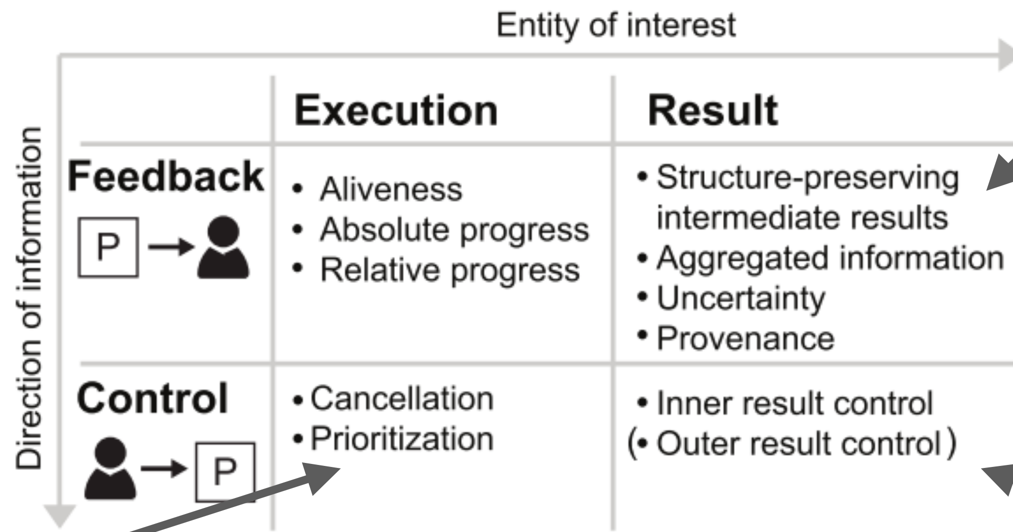
result:
algorithm acts as a **black box**



insight
in data
in algorithms

high-dimensional data
cell line + treatment + target + compound + …

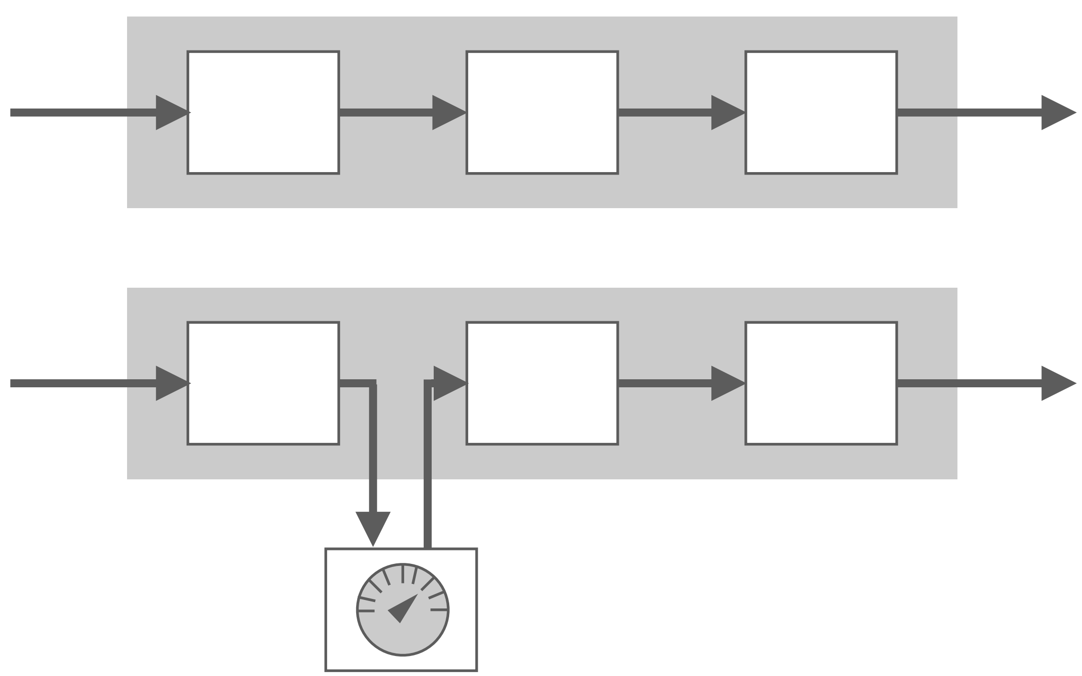
intermediate feedback: e.g. intermediate result, goodness-of-fit for best subset so far



Muhlbacher et al, 2014

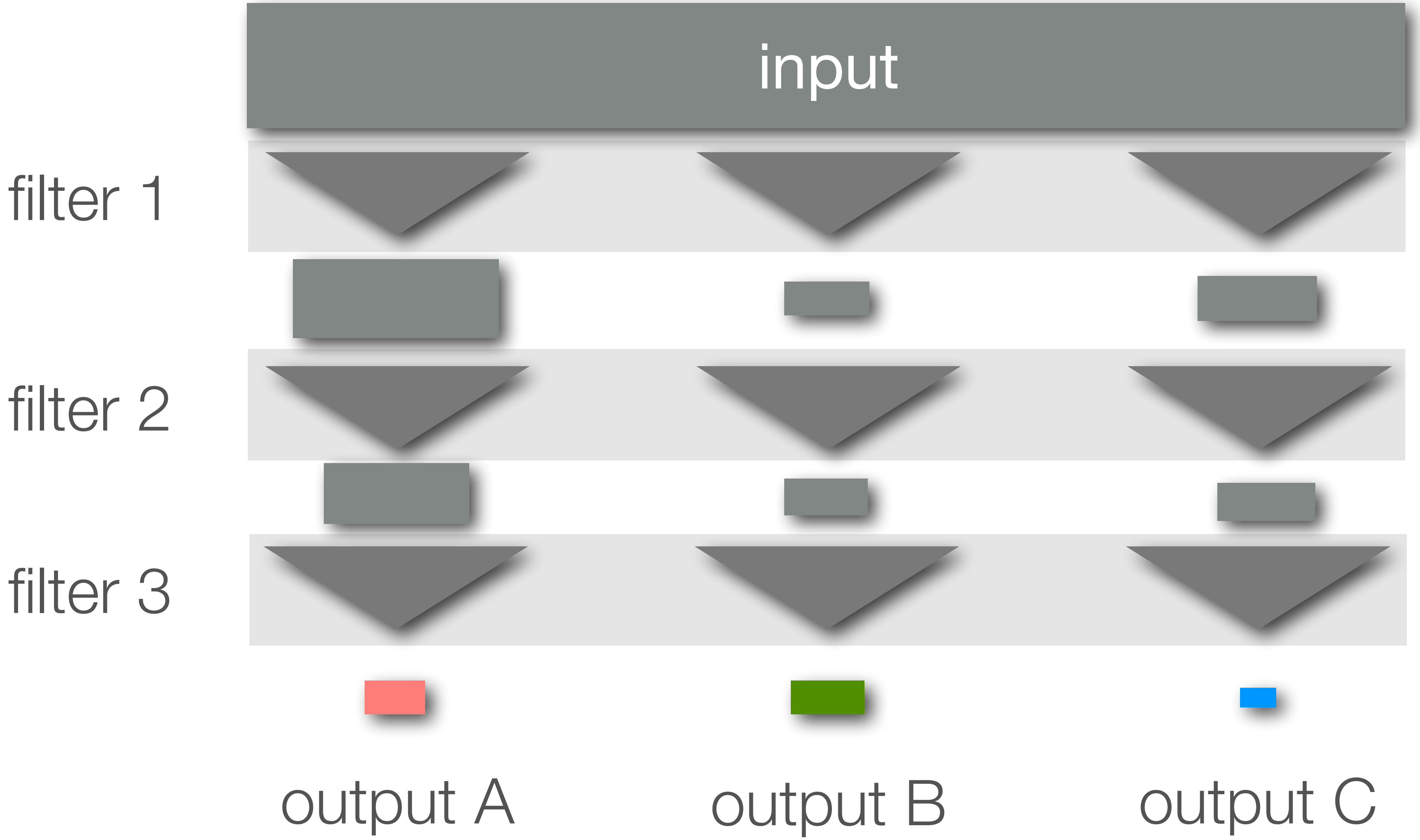
steer the final result: early validation of intermediate results, guided feature selection, weighting, avoiding being stuck in local minima, …

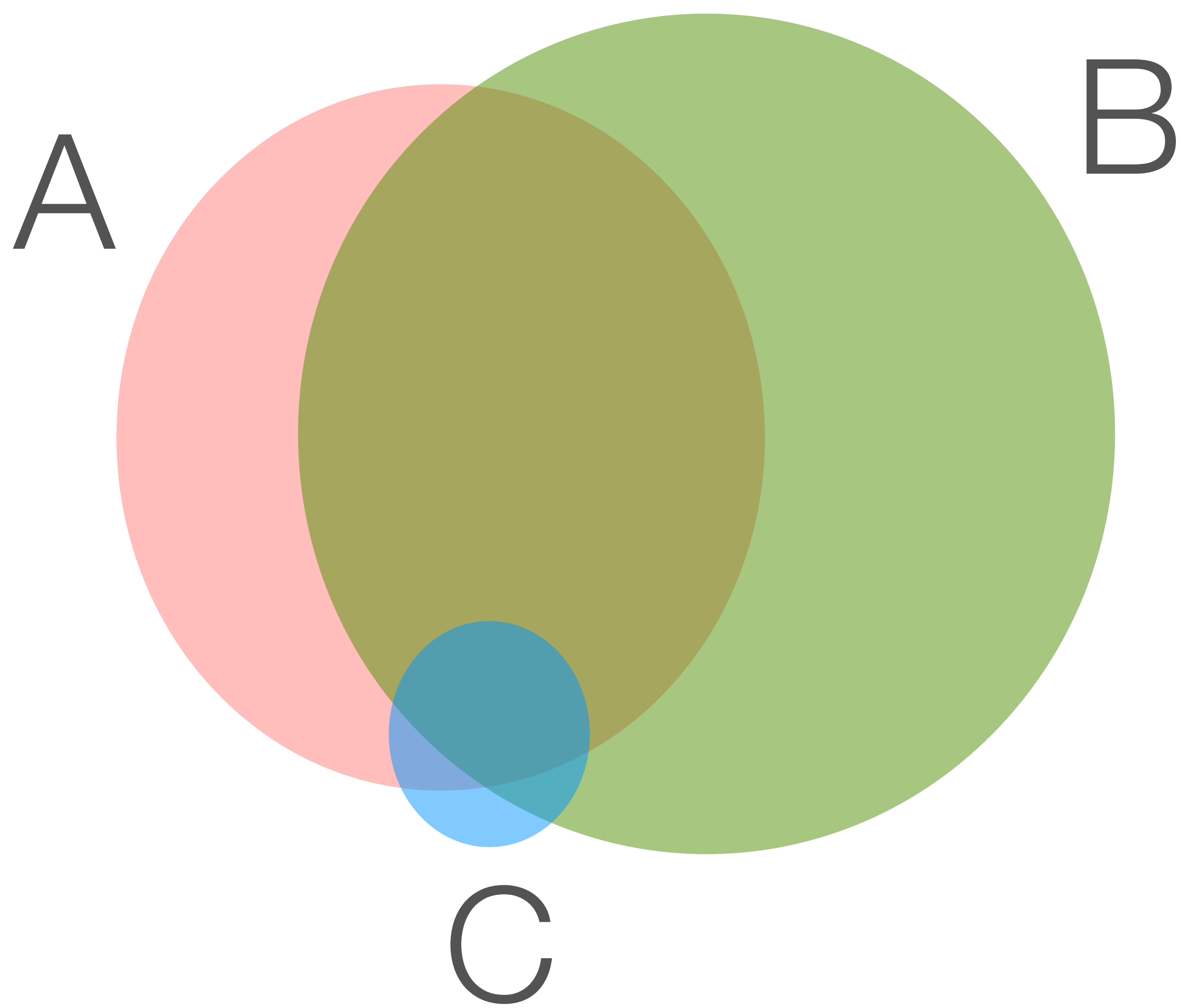
prioritization of remaining work: alter sequence of intermediate results to generate presumably more interesting ones earlier

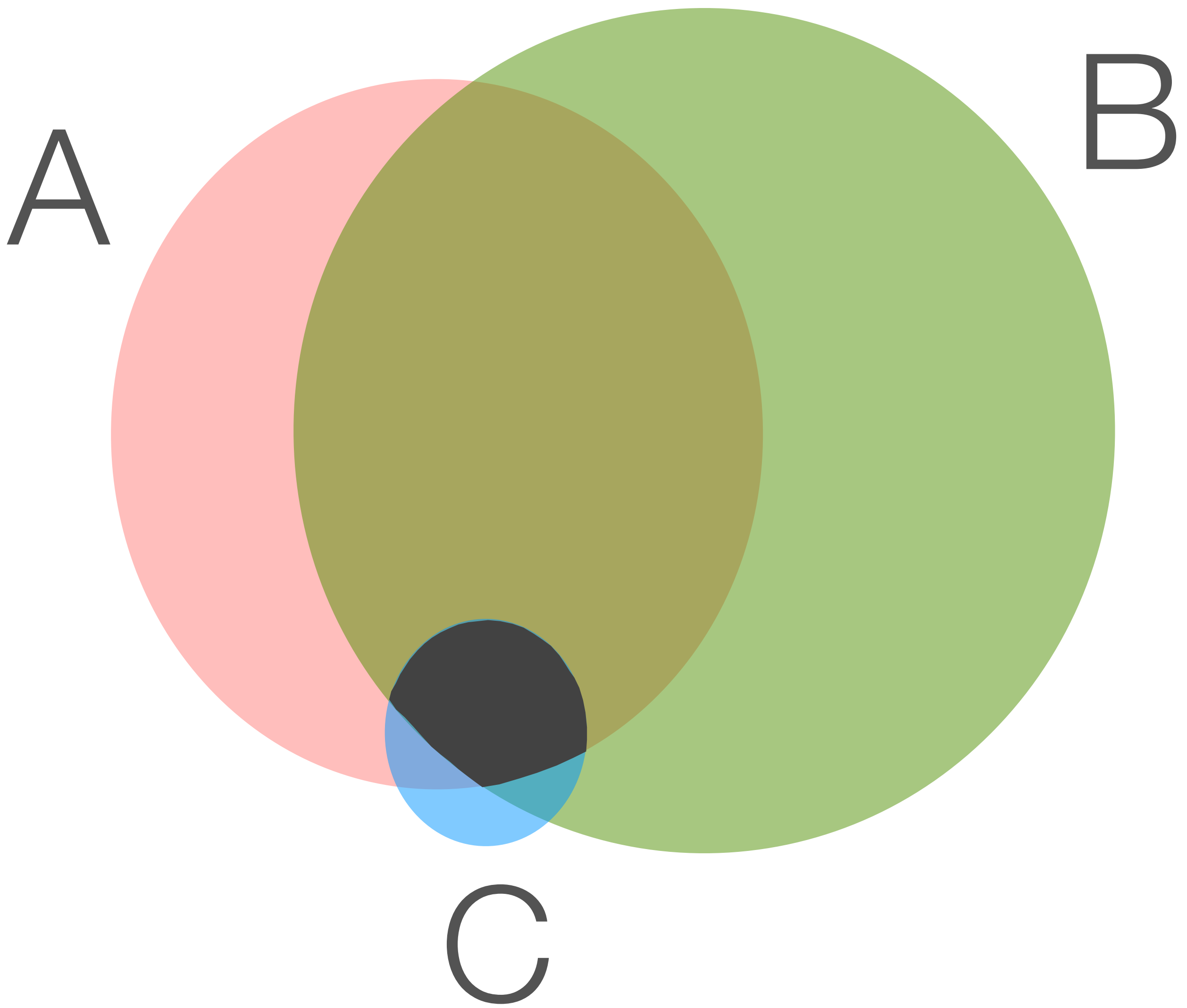


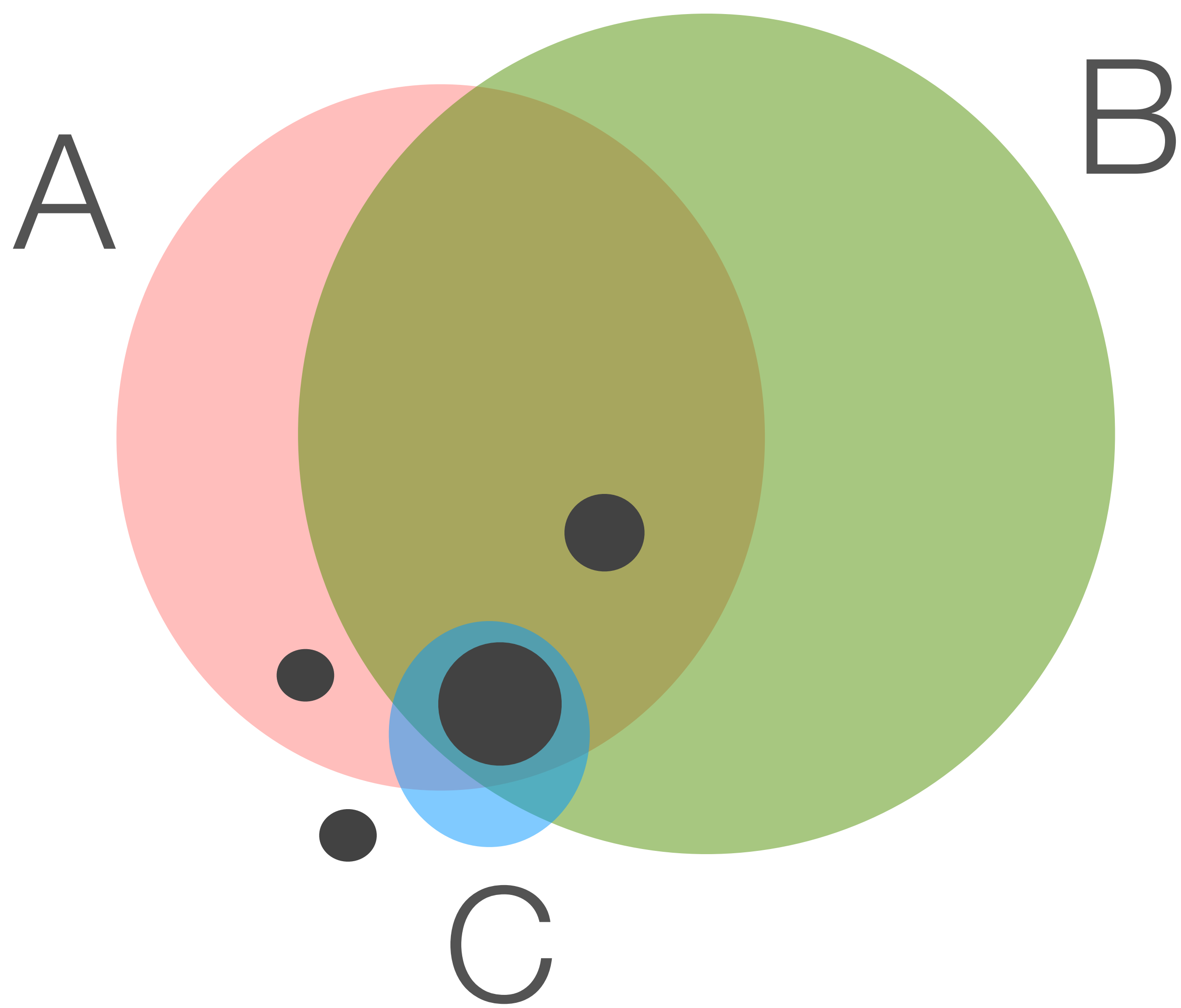
Example: mutation filtering

going from 5-6 million to 20,000

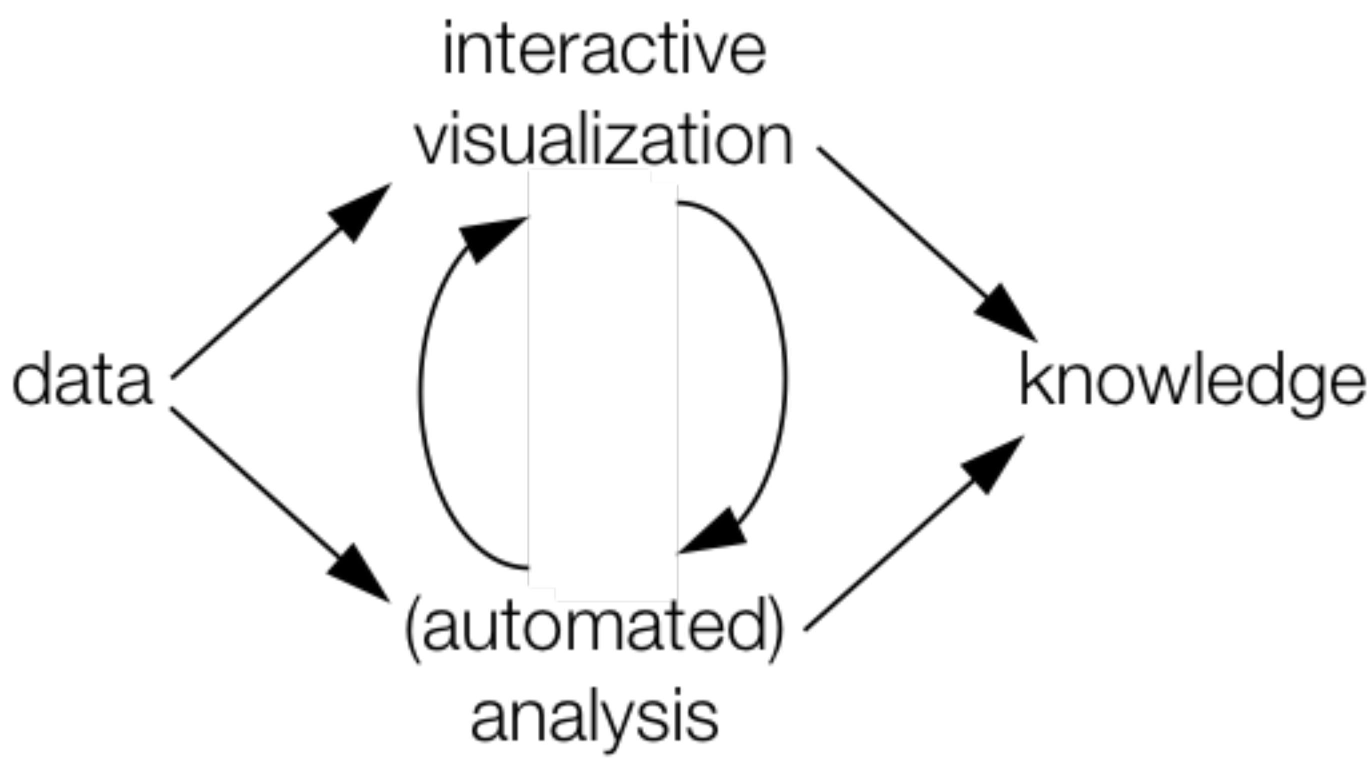



A
B
C


A
B
C


A
B
C

Visual Analytics to the rescue



Overview of lecture

A. Why visual analytics?

B. Data visualization

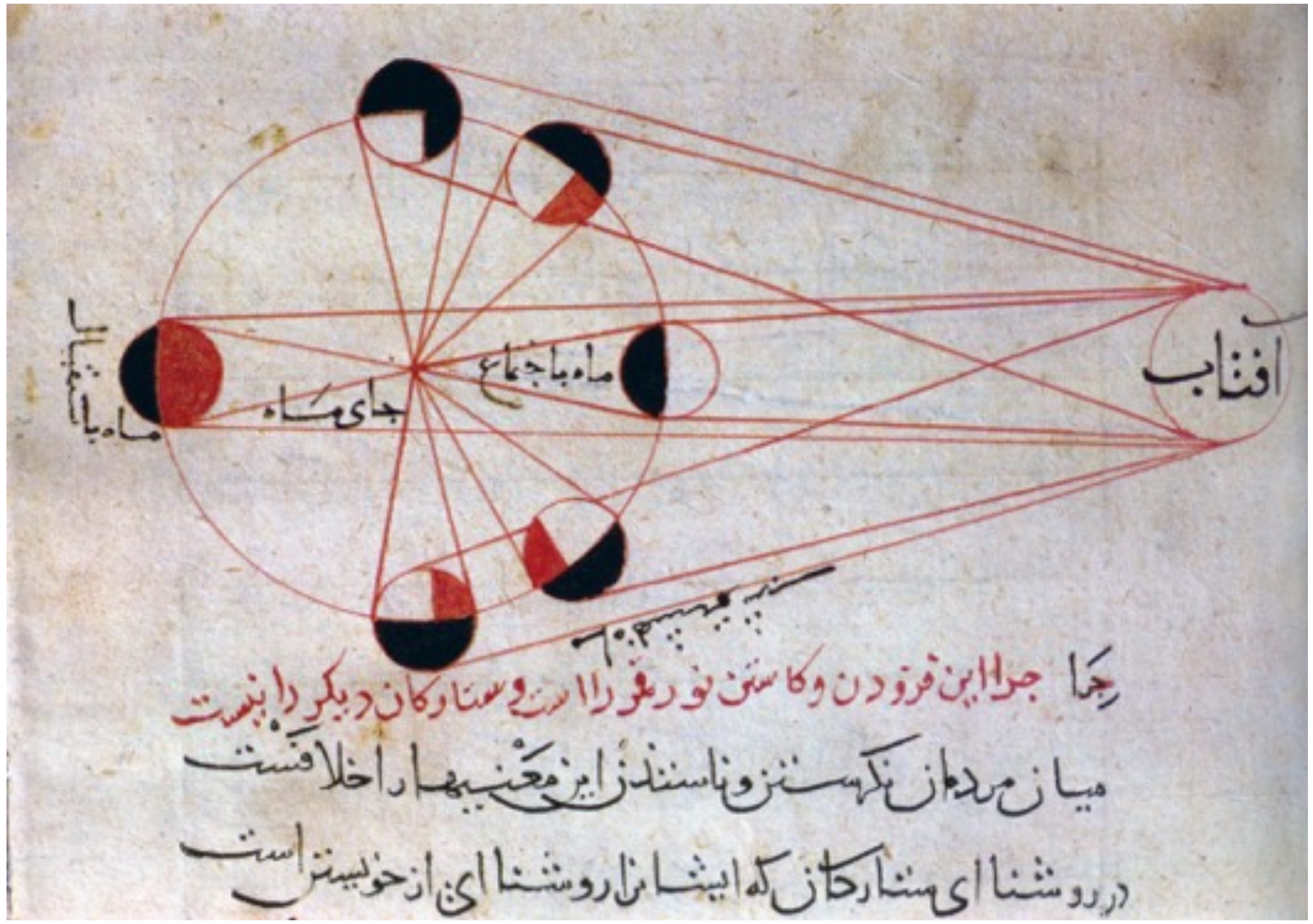
- Data foundations
- Human perception foundations
- Visualization foundations and examples

C. Visualization evaluation

D. Tools of the trade

B. Data Visualization

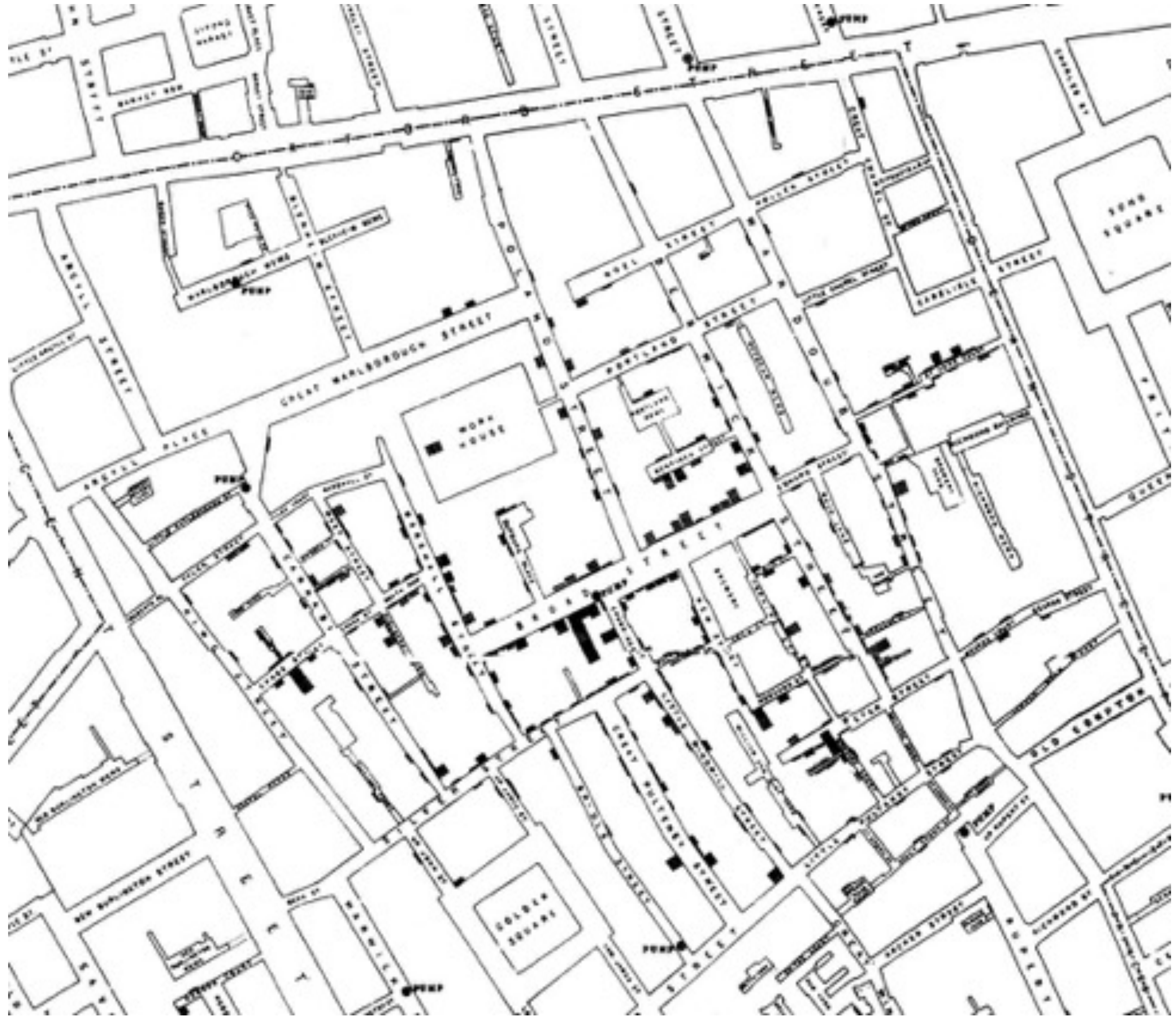
Historical perspective



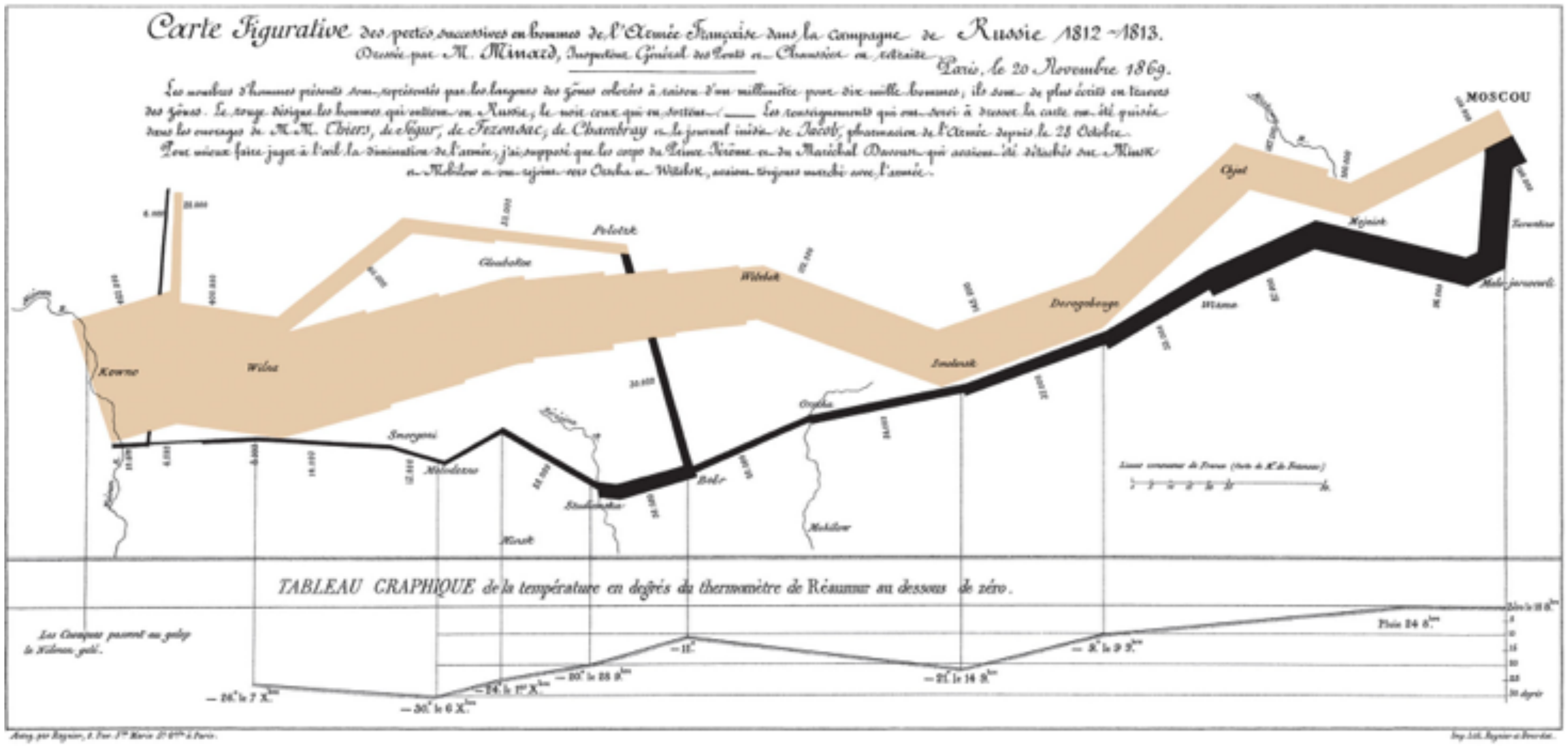
Al-Biruni - time series visualization: phases of the moon in orbit (circa 1030)

Hereford map - largest surviving map of the Middle Ages (1280s)

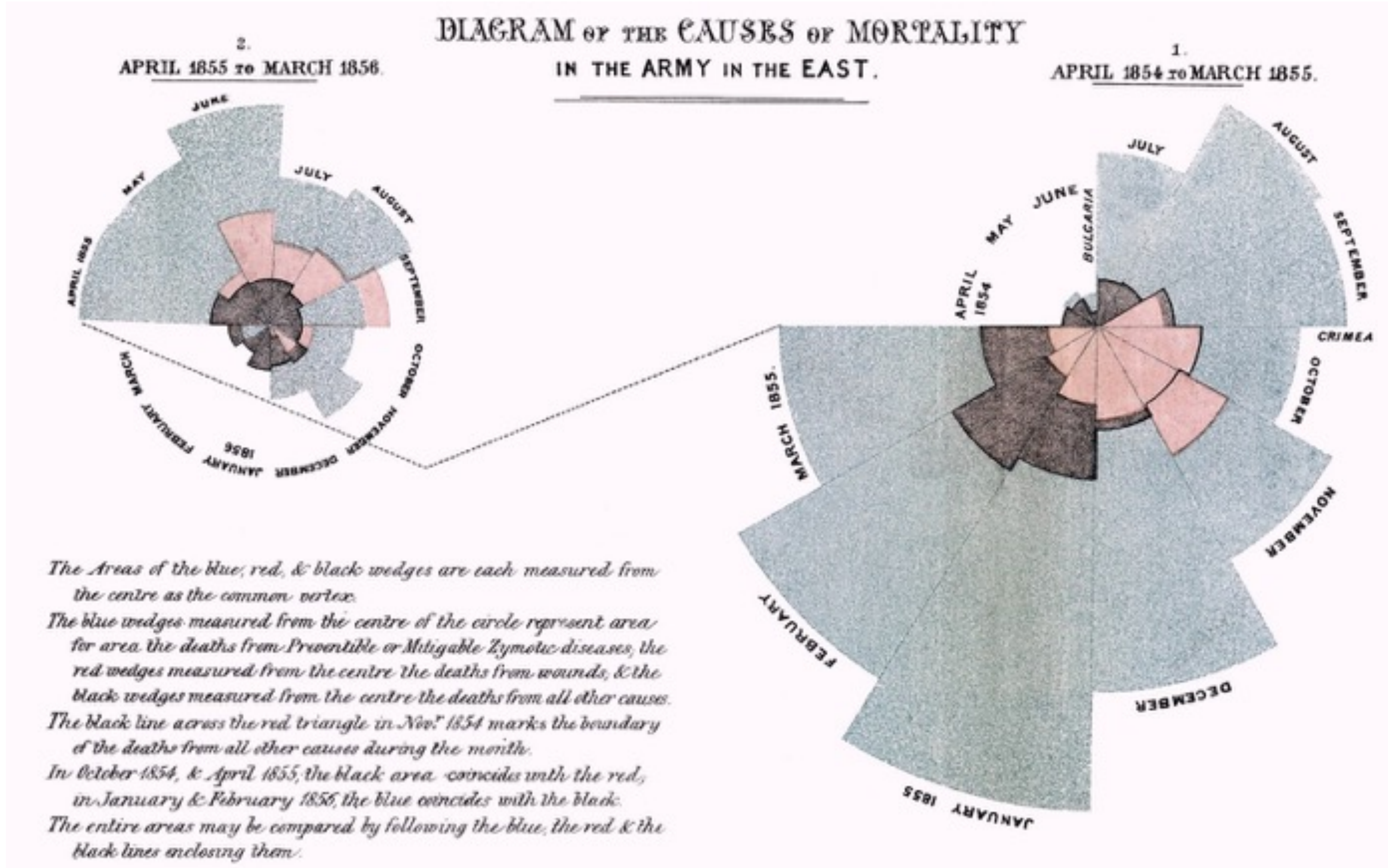
Peutinger Map - roads in Roman Empire, 1570
Abraham Ortelius (Flemish cartographer)



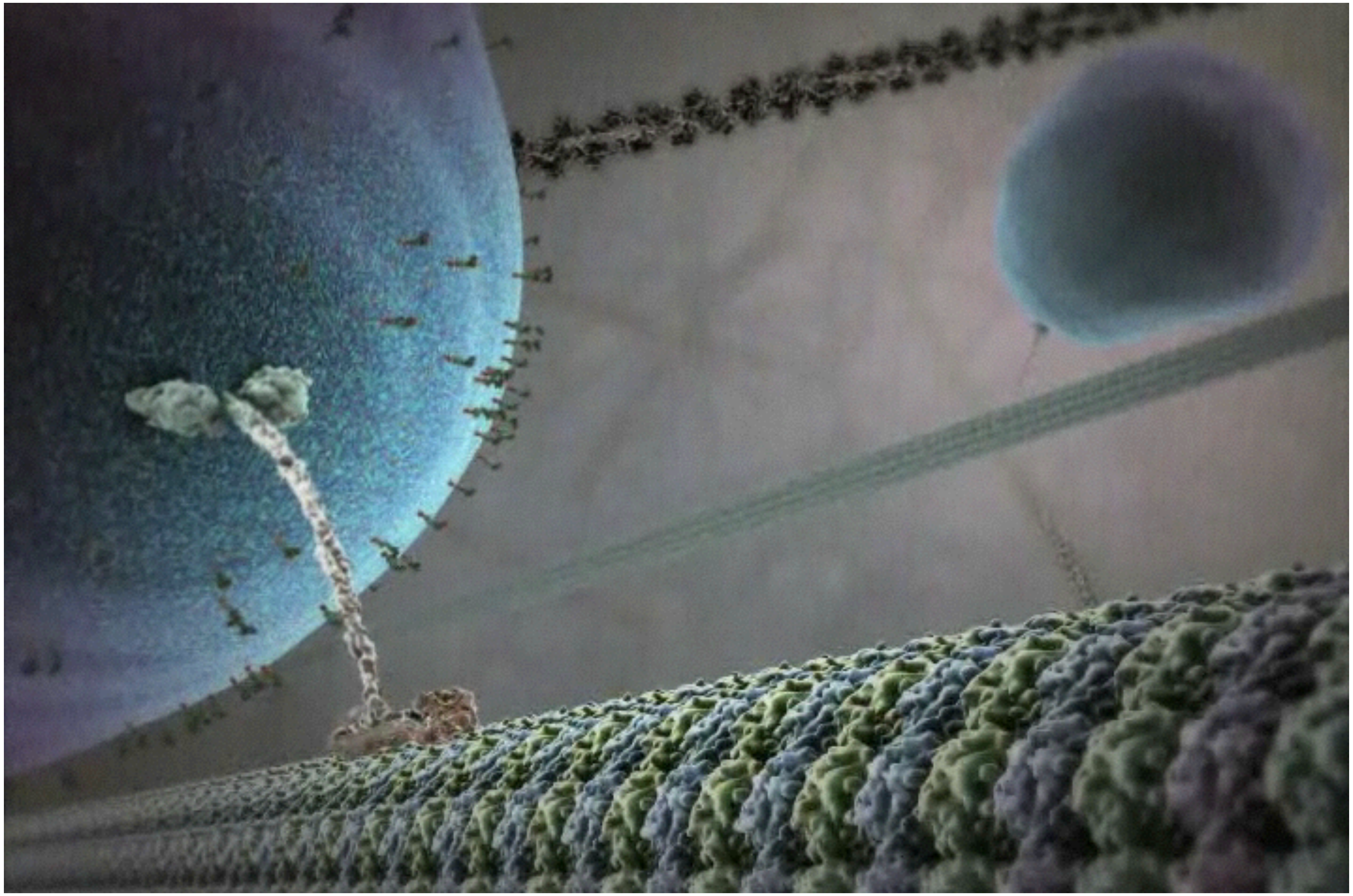
John Snow - cases of cholera in London (1663)



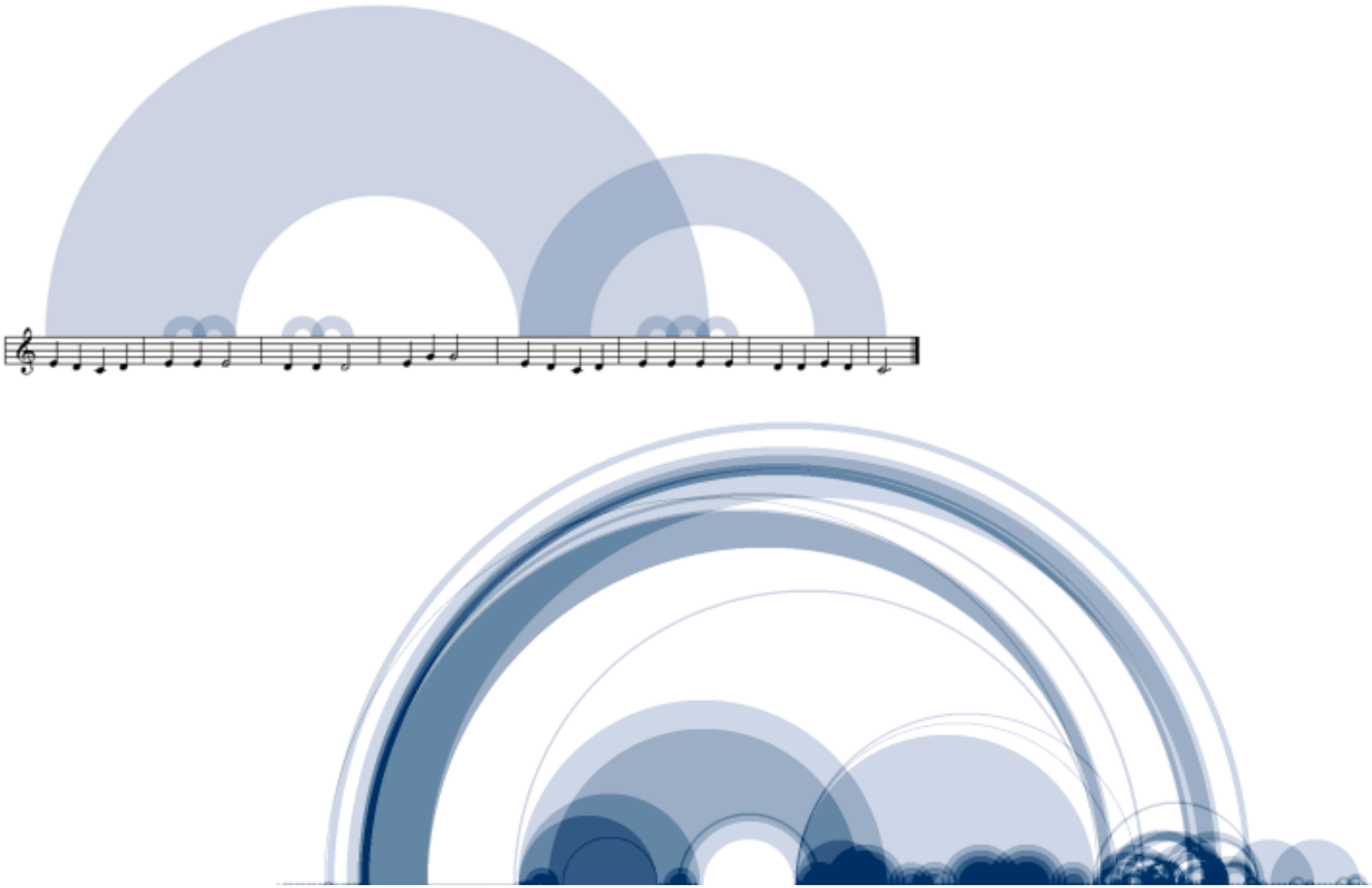
Charles Joseph Minard (1781-1870) - Napoleon’s march on Moscow



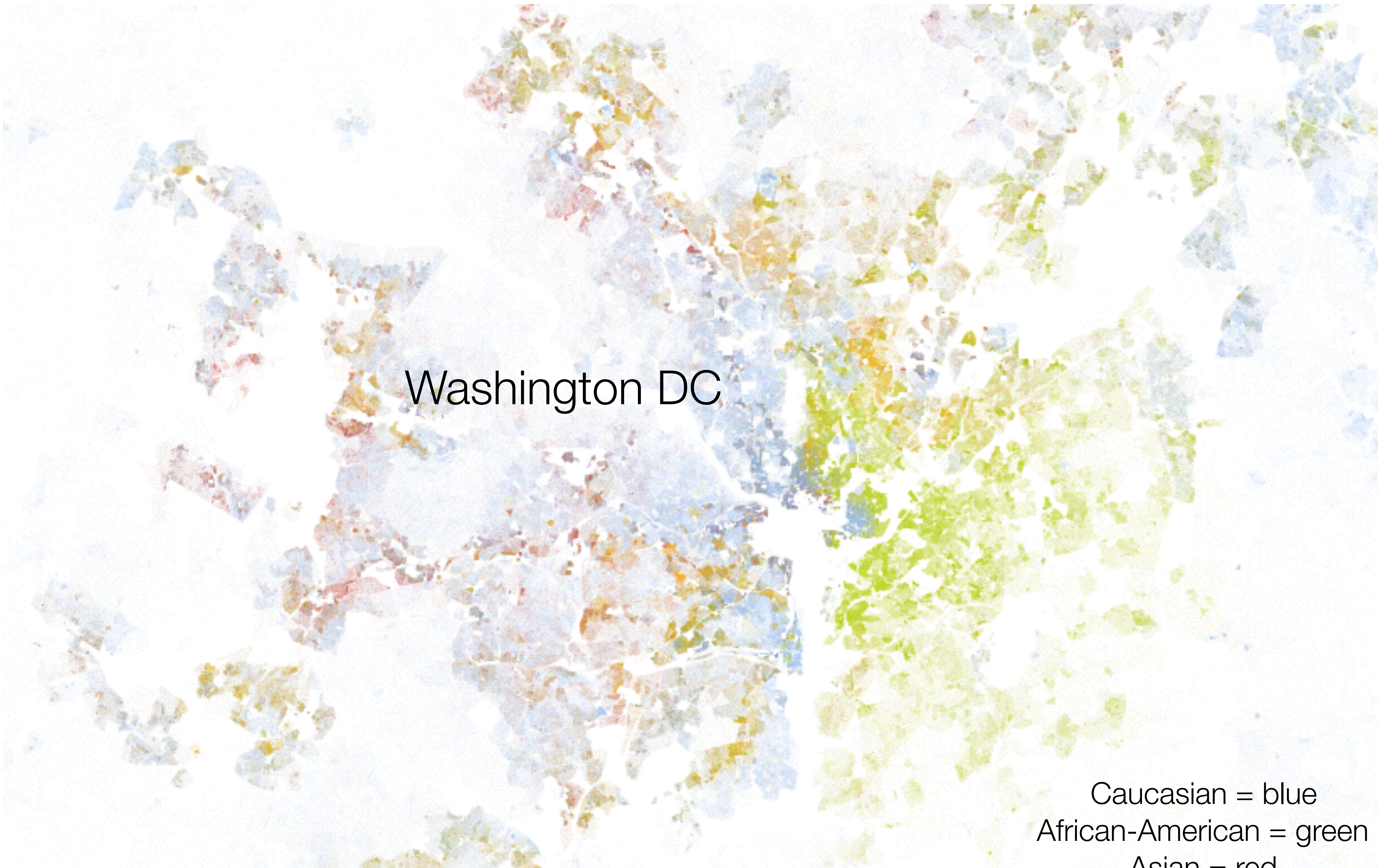
Florence Nightingale (1820-1910)
coxcomb chart monthly deaths from battle and other causes



http://multimedia.mcb.harvard.edu/anim_innerlife.html



Shape of Songs: “Like a Prayer” (Madonna)
Martin Wattenberg



Racial dot map: one dot per person in the US, coloured by ethnicity

source:http://www.coopercenter.org/demographics/Racial-Dot-Map

What is data visualization?

perception vs cognition

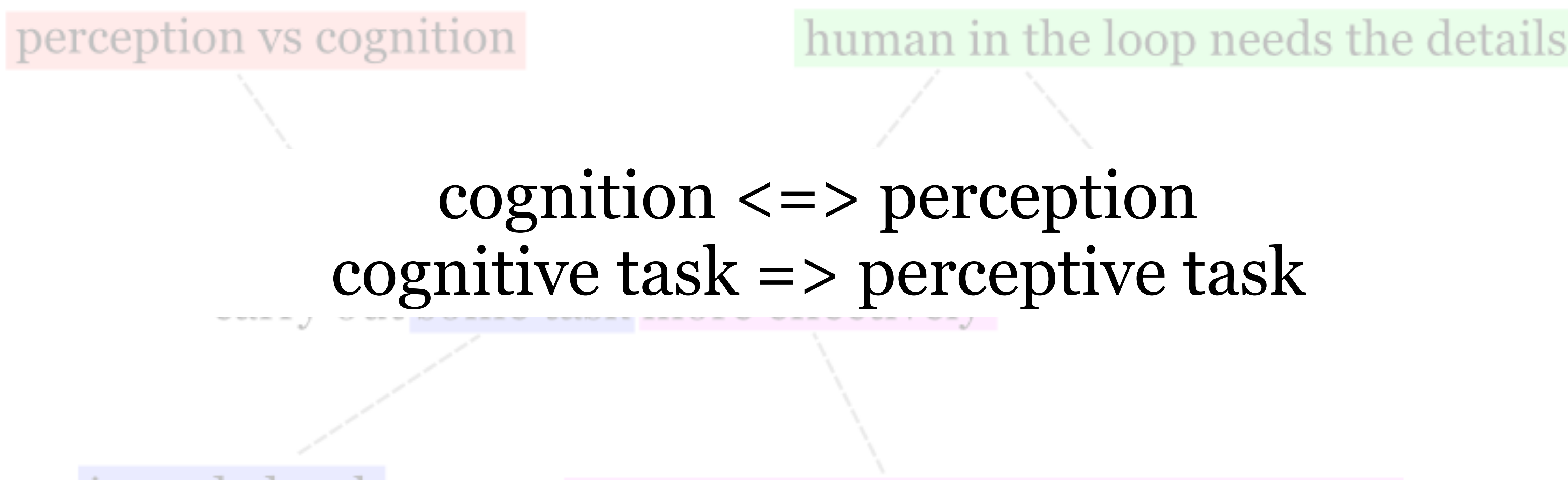
human in the loop needs the details

computer-based visualization systems providing
visual representations of datasets to help people
carry out some task more effectively

intended task

measurable definitions of effectiveness

T. Munzner



cognition <=> perception
cognitive task => perceptive task

identify anomalies, clusters, trends

T. Munzner

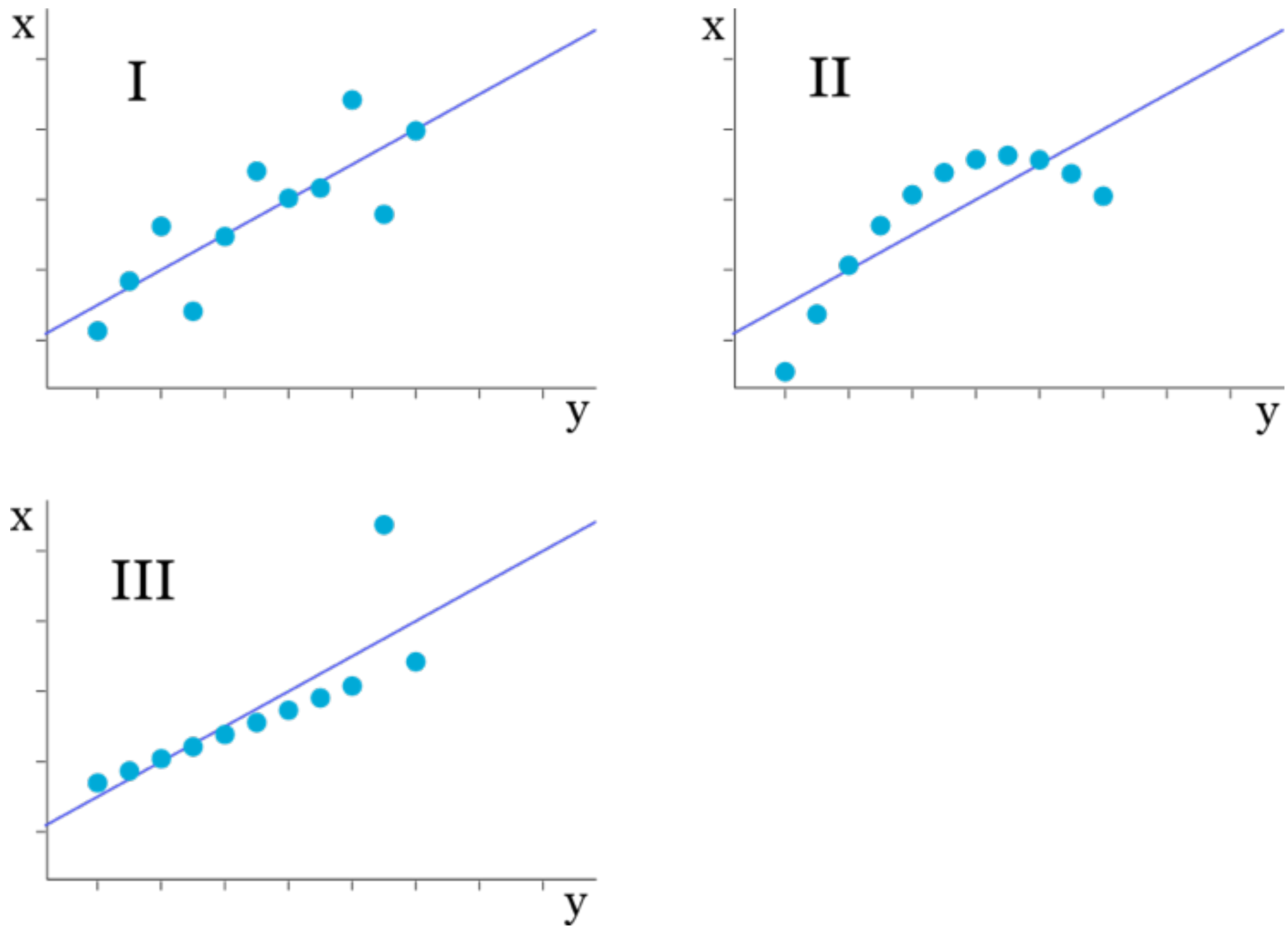
I		II		III	
x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46
8.0	6.95	8.0	8.14	8.0	6.77
13.0	7.58	13.0	8.74	13.0	12.74
9.0	8.81	9.0	8.77	9.0	7.11
11.0	8.33	11.0	9.26	11.0	7.81
14.0	9.96	14.0	8.10	14.0	8.84
6.0	7.24	6.0	6.13	6.0	6.08
4.0	4.26	4.0	3.10	4.0	5.39
12.0	10.84	12.0	9.13	12.0	8.15
7.0	4.82	7.0	7.26	7.0	6.42
5.0	5.68	5.0	4.74	5.0	5.73

n = 11

mean x = 9.0
mean y = 7.5

variance x = 11.0
variance y = 4.12

correlation x & y = 0.816
regression line: y = 3+0.5x



x
I
y
x
II
y
x
III
y

Misunderstandings visual design

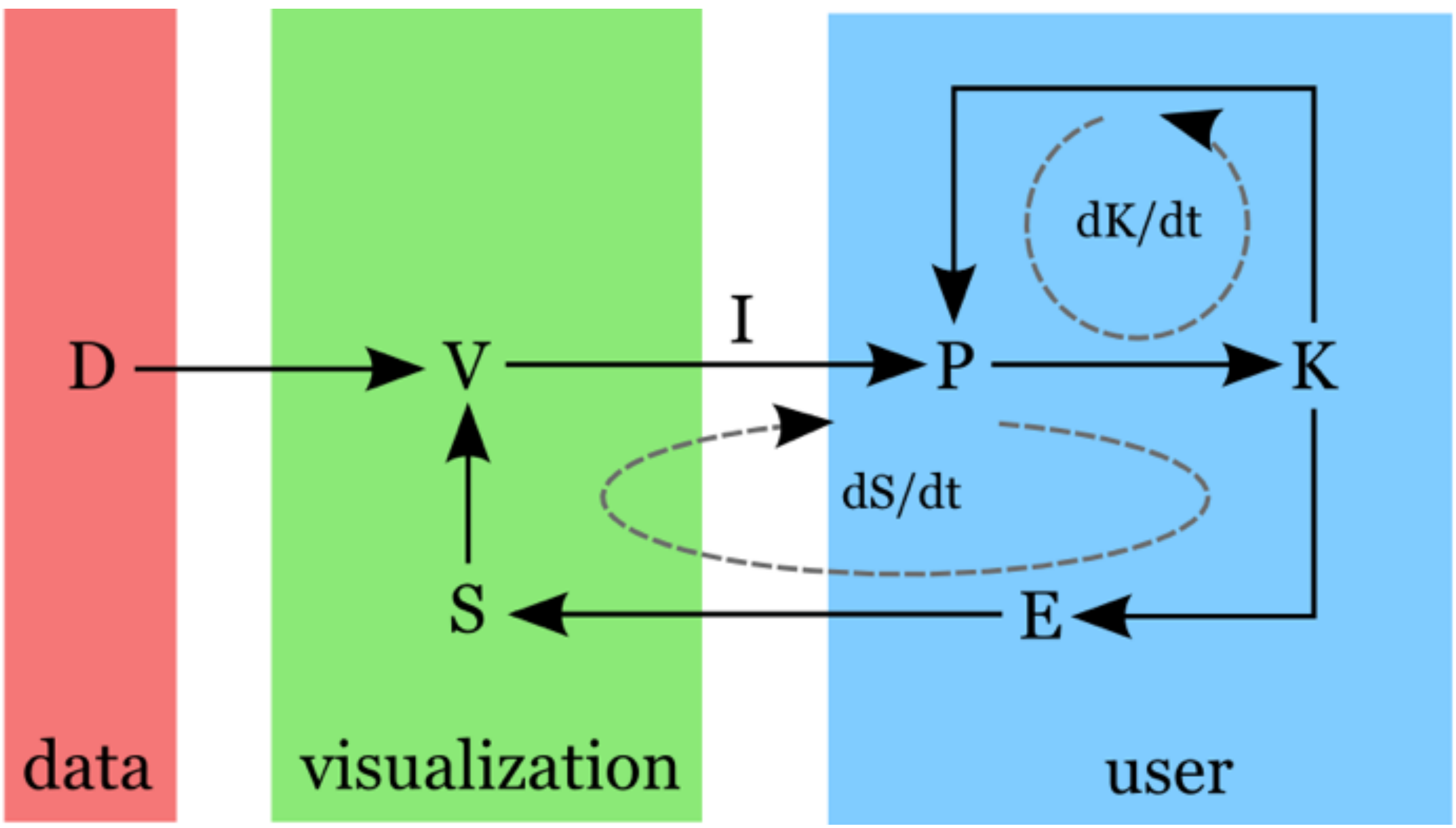
Is it a matter of talent? creativity? inspiration? aesthetics? taste?

practice => anyone can learn (to some extent)

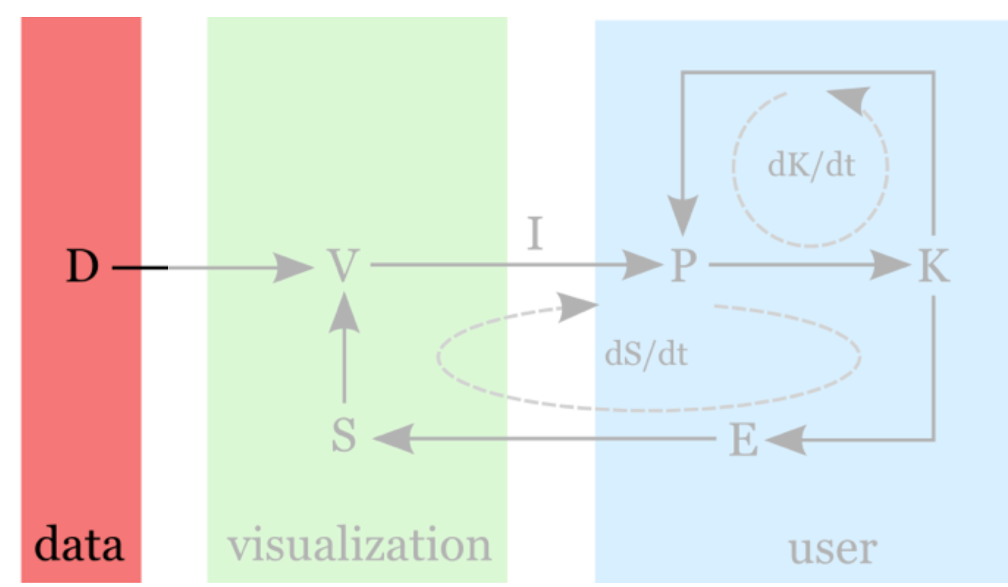
be rational

aim at usability => beauty/aesthetics often follows

Components

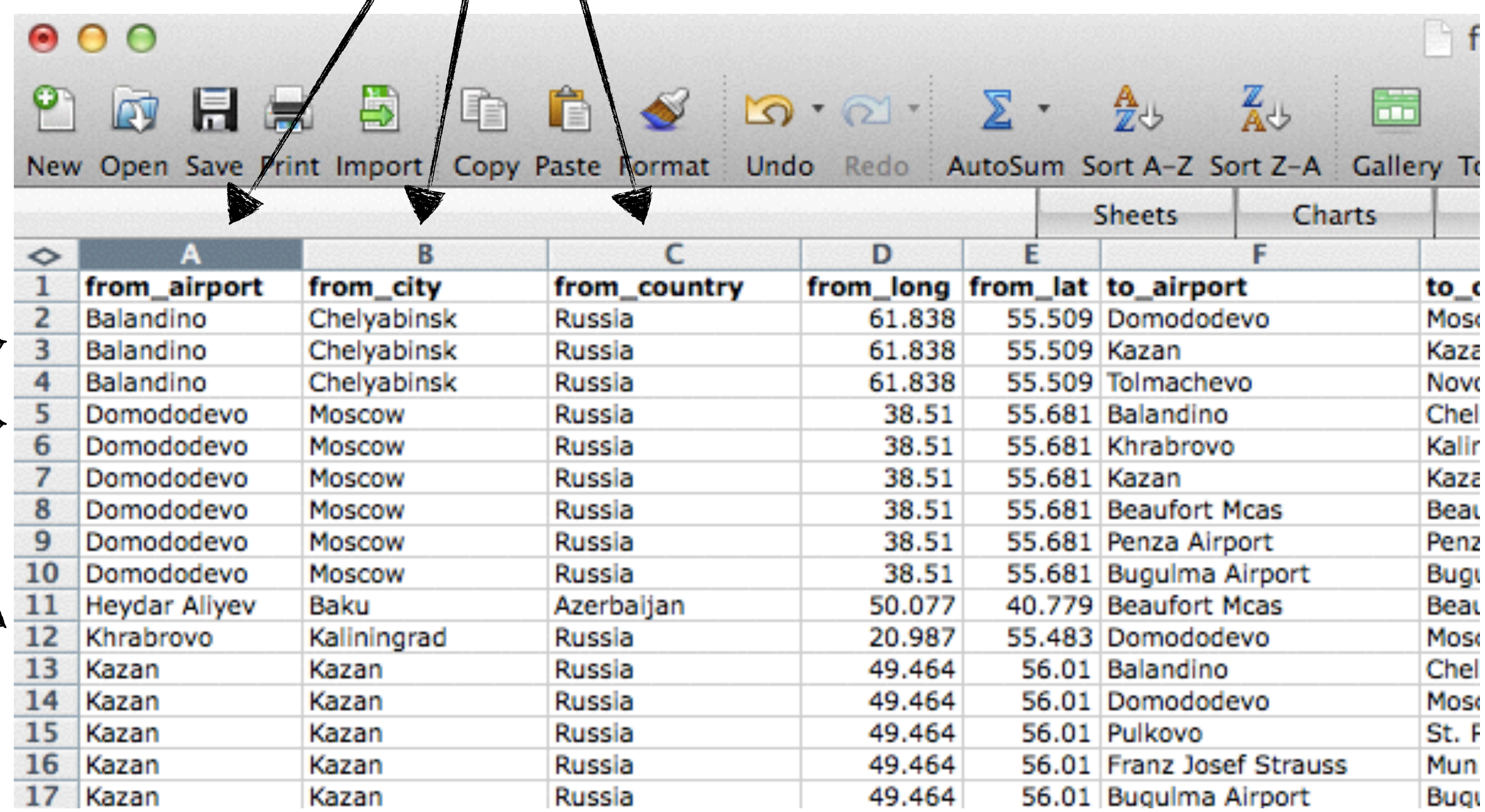


Data foundations

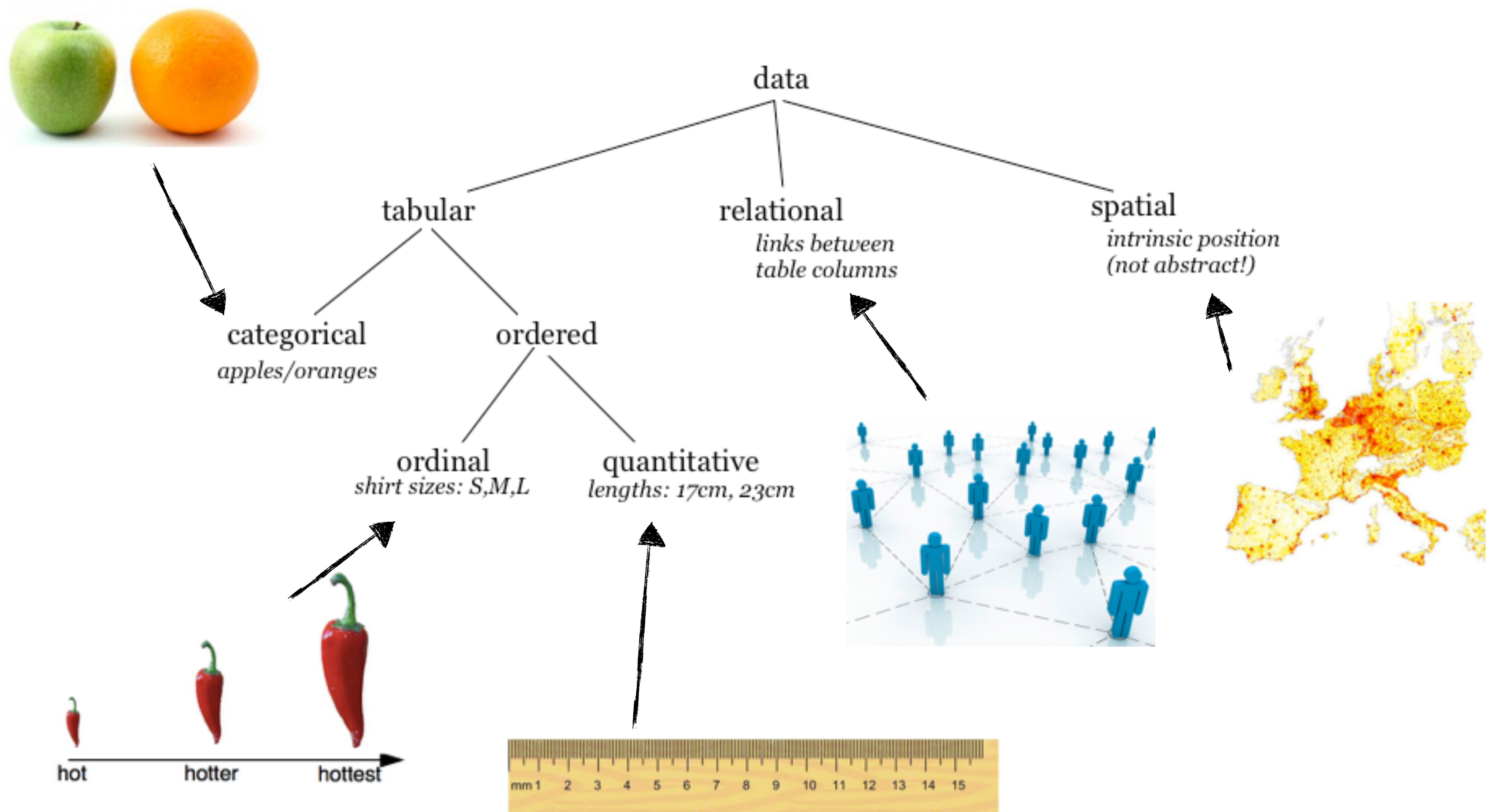


data attributes
dimensions
features
properties

data items
datapoints

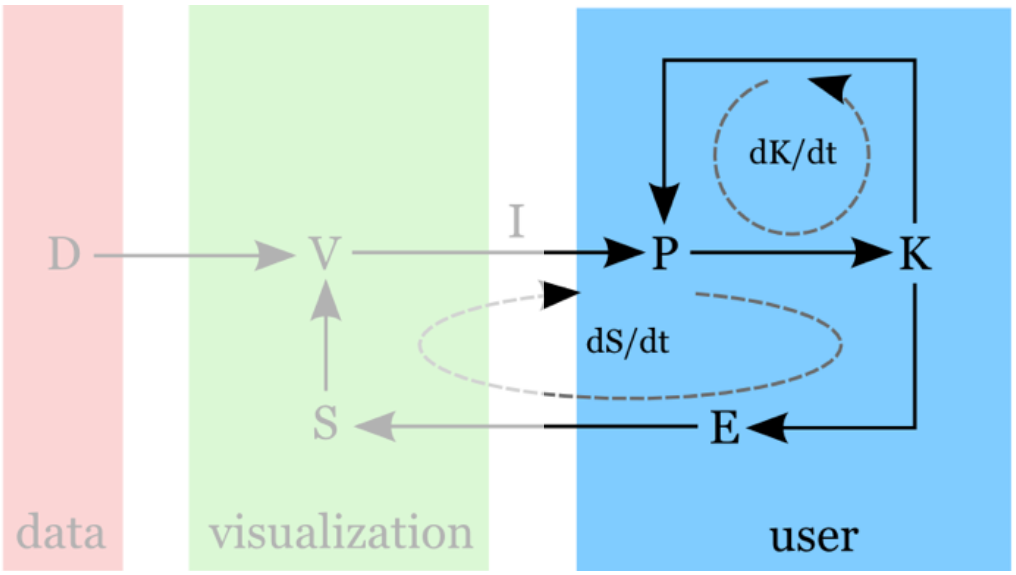


	A	B	C	D	E	F	
1	from_airport	from_city	from_country	from_long	from_lat	to_airport	to_c
2	Balandino	Chelyabinsk	Russia	61.838	55.509	Domododevo	Mosc
3	Balandino	Chelyabinsk	Russia	61.838	55.509	Kazan	Kaza
4	Balandino	Chelyabinsk	Russia	61.838	55.509	Tolmachevo	Novo
5	Domododevo	Moscow	Russia	38.51	55.681	Balandino	Chel
6	Domododevo	Moscow	Russia	38.51	55.681	Khrabrovo	Kalir
7	Domododevo	Moscow	Russia	38.51	55.681	Kazan	Kaza
8	Domododevo	Moscow	Russia	38.51	55.681	Beaufort Mcas	Beau
9	Domododevo	Moscow	Russia	38.51	55.681	Penza Airport	Penz
10	Domododevo	Moscow	Russia	38.51	55.681	Bugulma Airport	Bugu
11	Heydar Aliyev	Baku	Azerbaijan	50.077	40.779	Beaufort Mcas	Beau
12	Khrabrovo	Kaliningrad	Russia	20.987	55.483	Domododevo	Mosc
13	Kazan	Kazan	Russia	49.464	56.01	Balandino	Chel
14	Kazan	Kazan	Russia	49.464	56.01	Domododevo	Mosc
15	Kazan	Kazan	Russia	49.464	56.01	Pulkovo	St. P
16	Kazan	Kazan	Russia	49.464	56.01	Franz Josef Strauss	Mun
17	Kazan	Kazan	Russia	49.464	56.01	Bugulma Airport	Bugu



S Stevens "On the theory of scales and measurements" (1946)

Human perception foundations

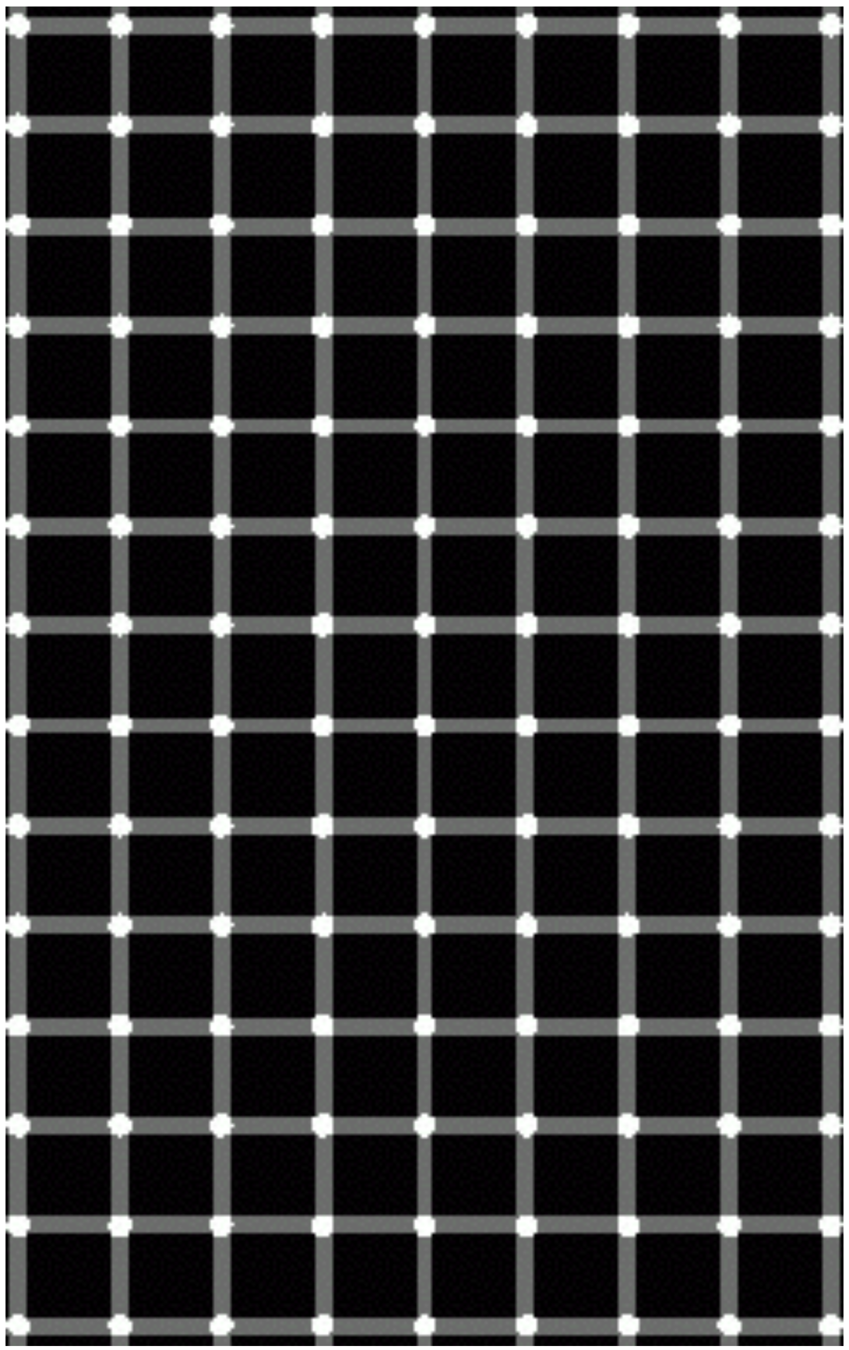


Human perception

- Gestalt laws
- pre-attentive vision
- selective attention
- colour

- What does this mean for visual design?

Gestalt laws - interplay between parts and the whole (Kurt Koffka)



series of principles

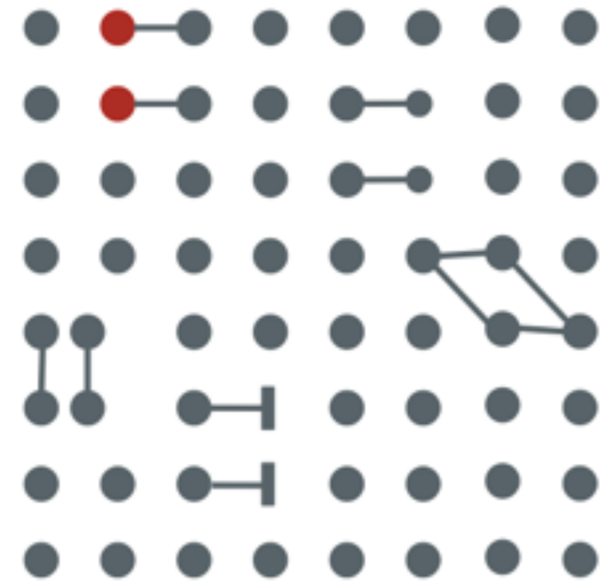
Election results Florida:

- black = Bush
- white = Gore

Hermann grid illusion

Simplicity

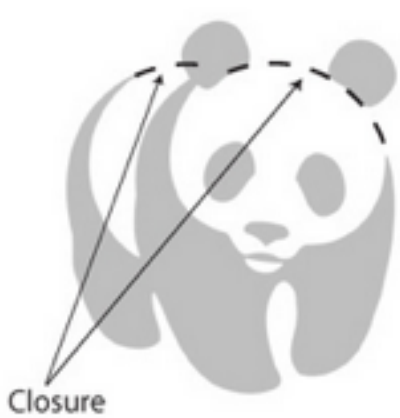
Connectedness



Proximity

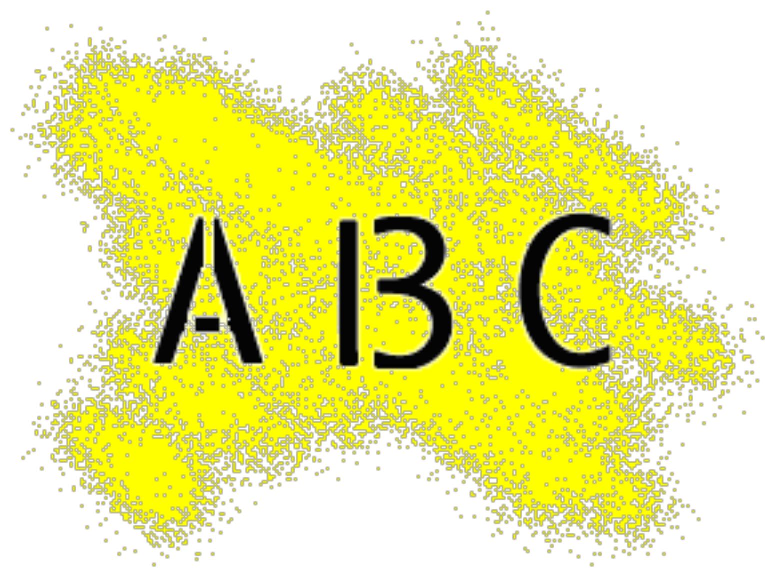
Similarity

Closure

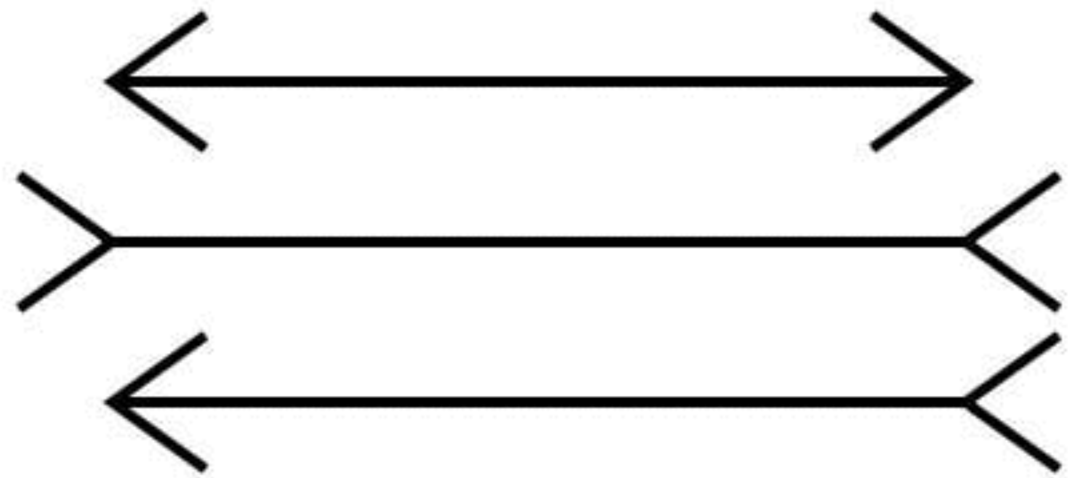


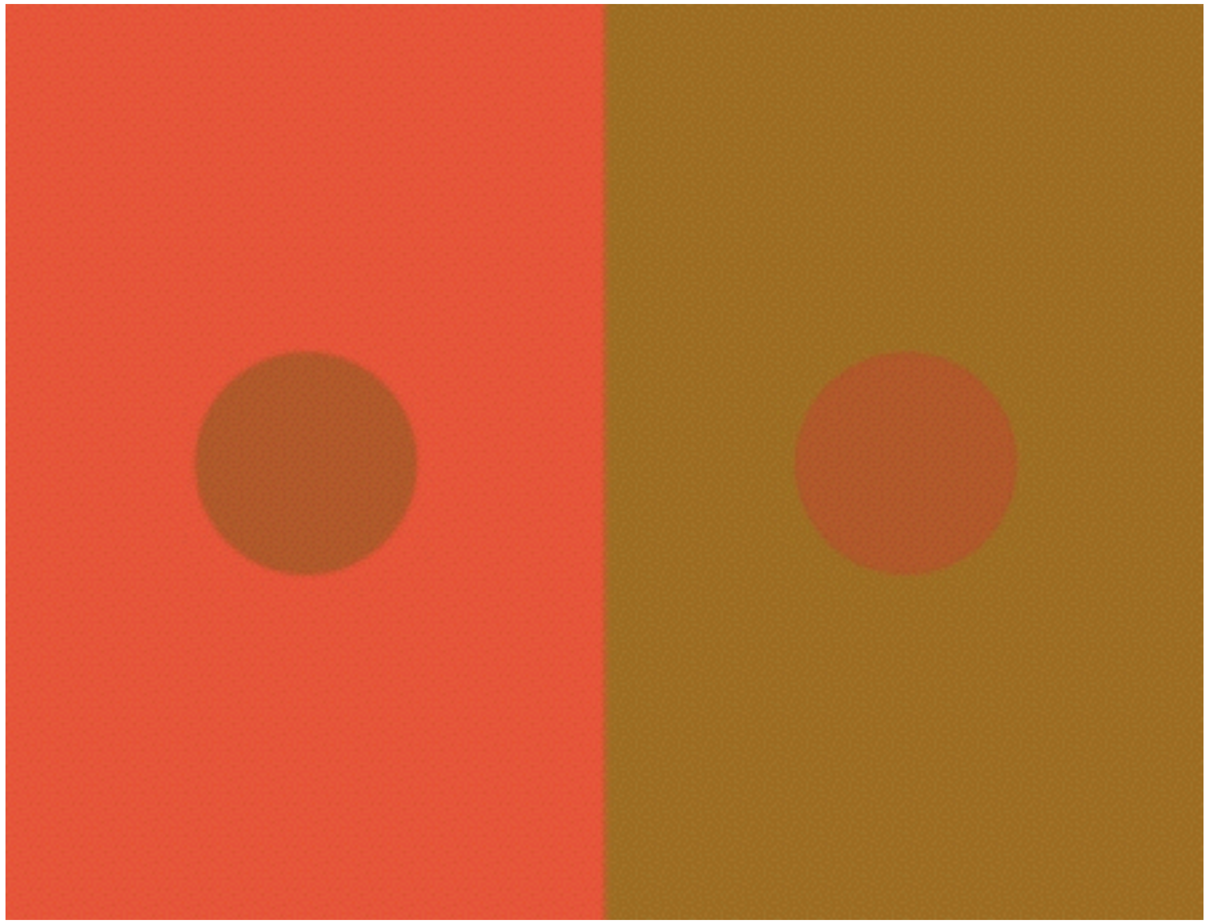
Kanizsa illusion

Symmetry, Familiarity, Continuity
Figure and Ground


ABC

12 13 14





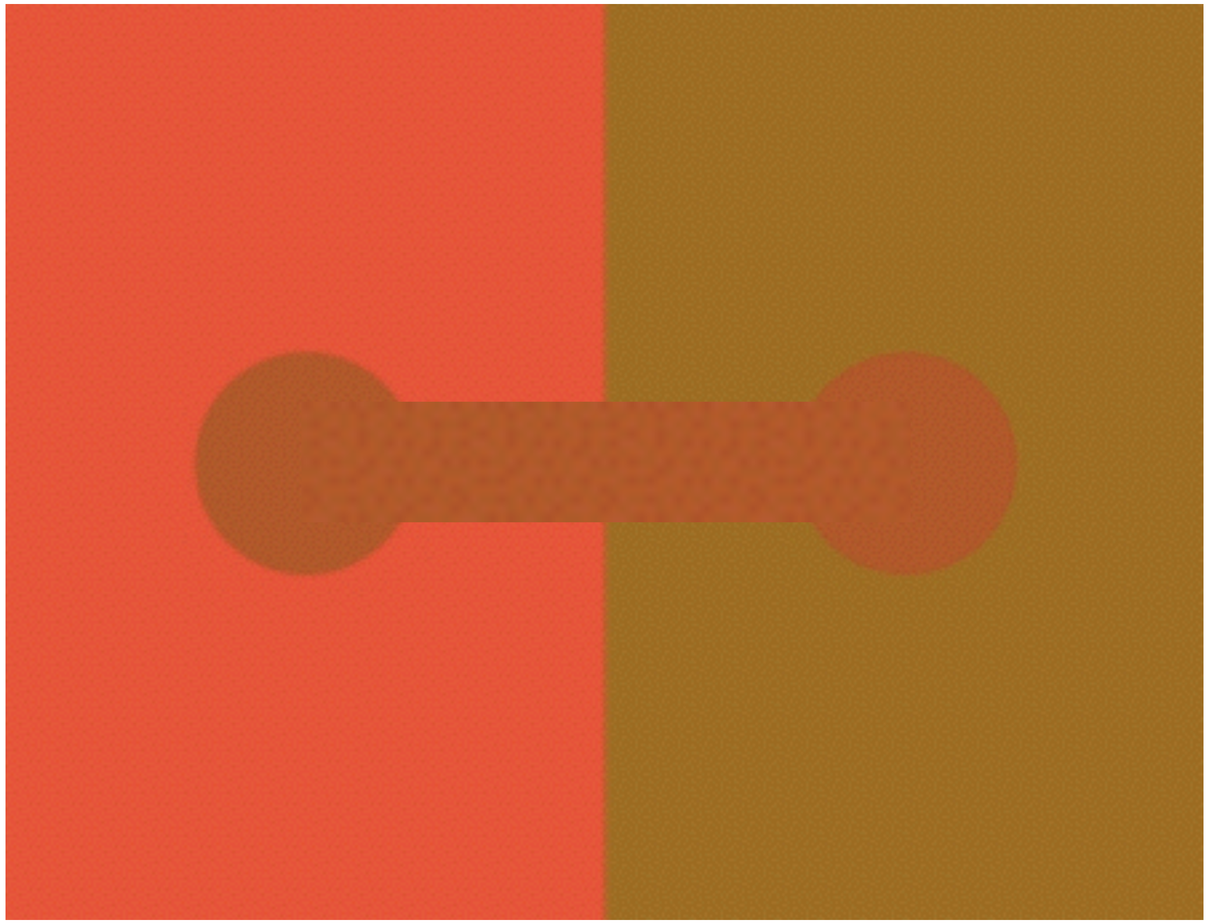
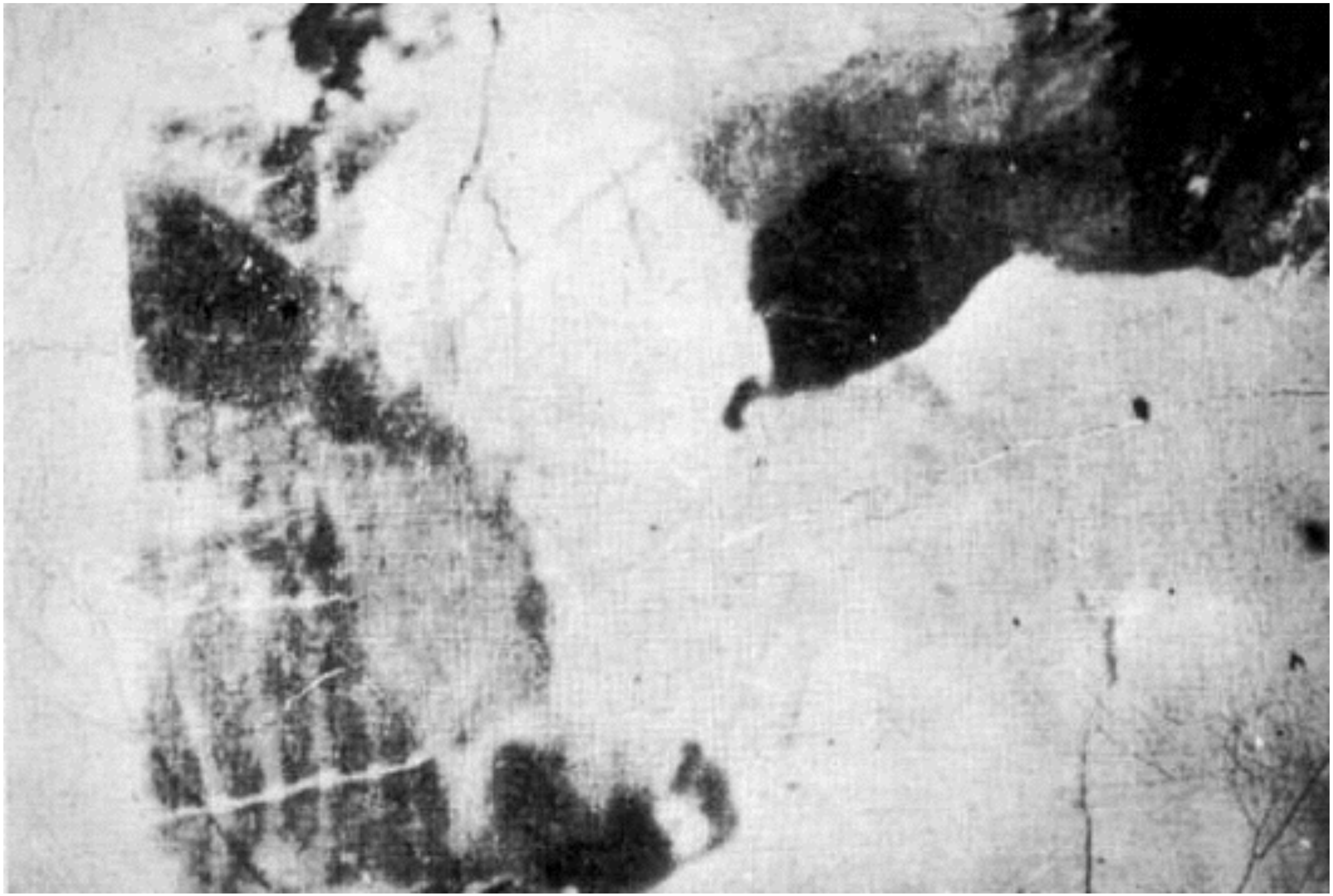


Figure and Ground



Pre-attentive vision

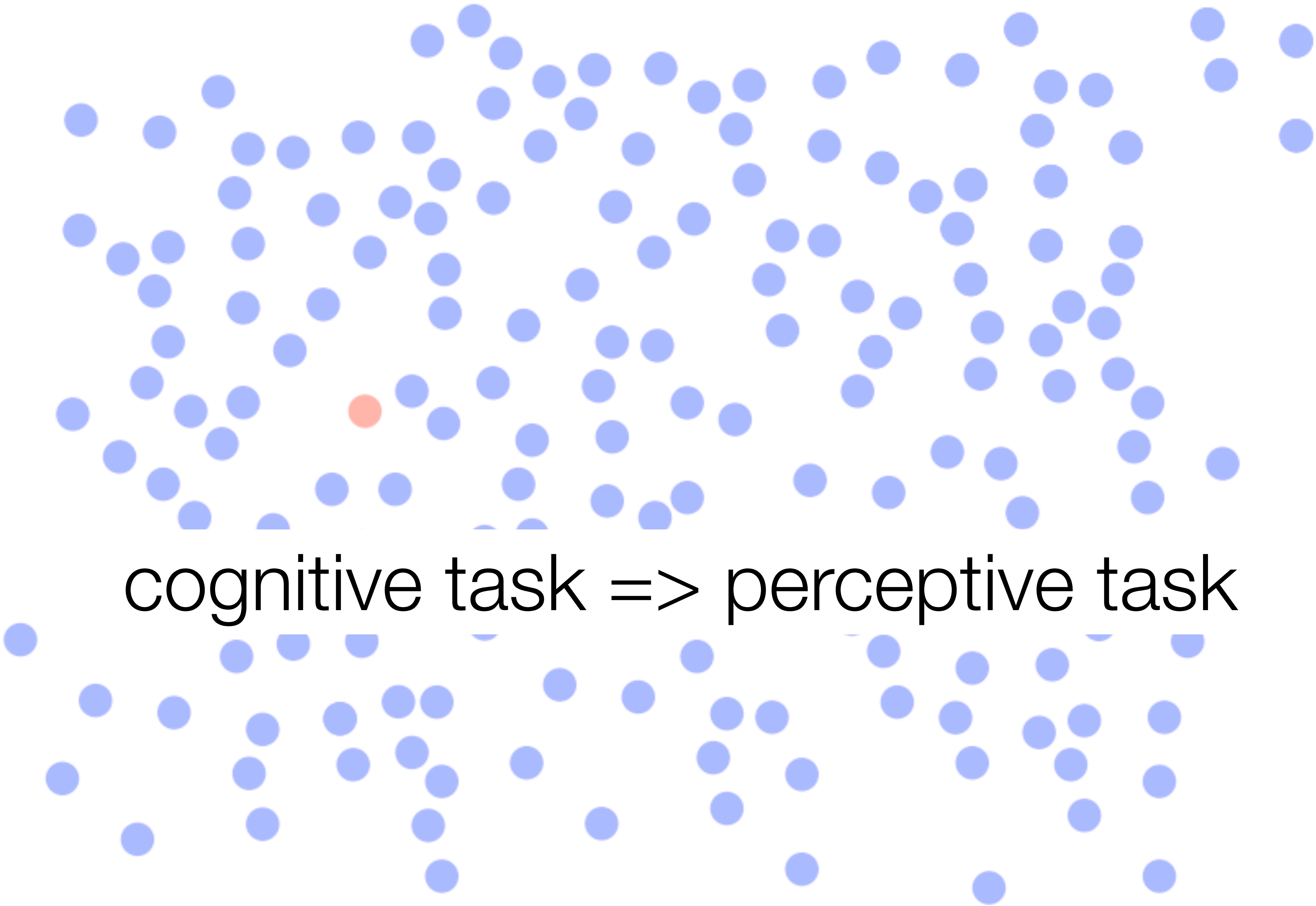
= ability of low-level human visual system to rapidly identify certain basic visual properties

- some features “pop out”
- used for:
  - target detection
  - boundary detection
  - counting/estimation
  - ...
- visual system takes over => all cognitive power available for interpreting the figure, rather than needing part of it for processing the figure

5

5

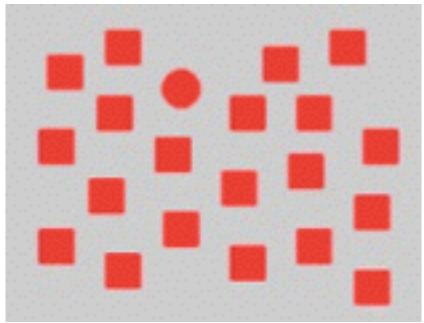
5



cognitive task => perceptive task

Limitations of preattentive vision

1. Speed depends on **which channel** (use one that is good for categorical)



2. **Combining** pre-attentive features does *not* always work => would need to resort to "**serial search**" (most channel pairs; all channel triplets)

e.g. is there a red square in this picture

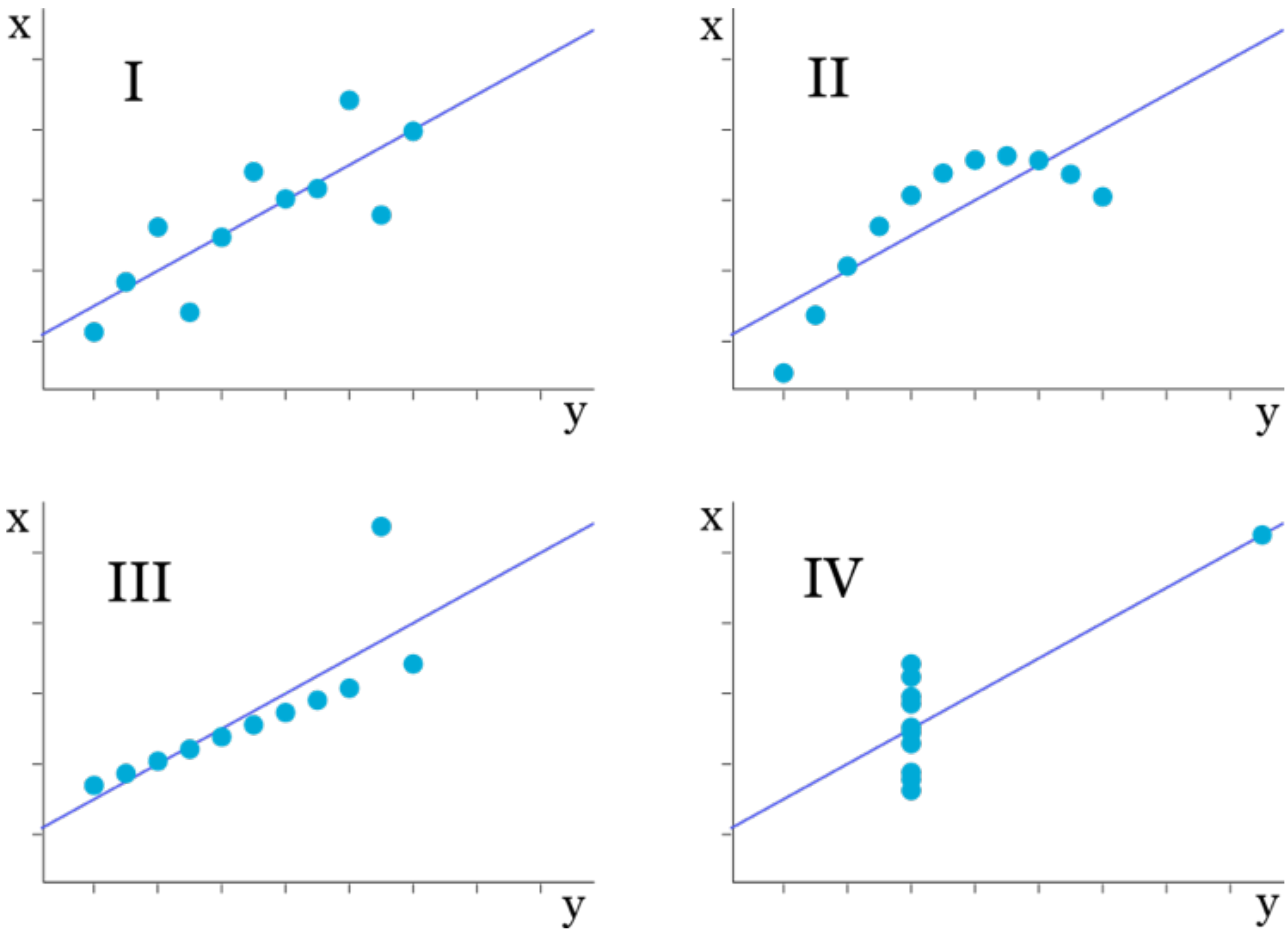
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.80

n = 11

mean x = 9.0
mean y = 7.5

variance x = 11.0
variance y = 4.12

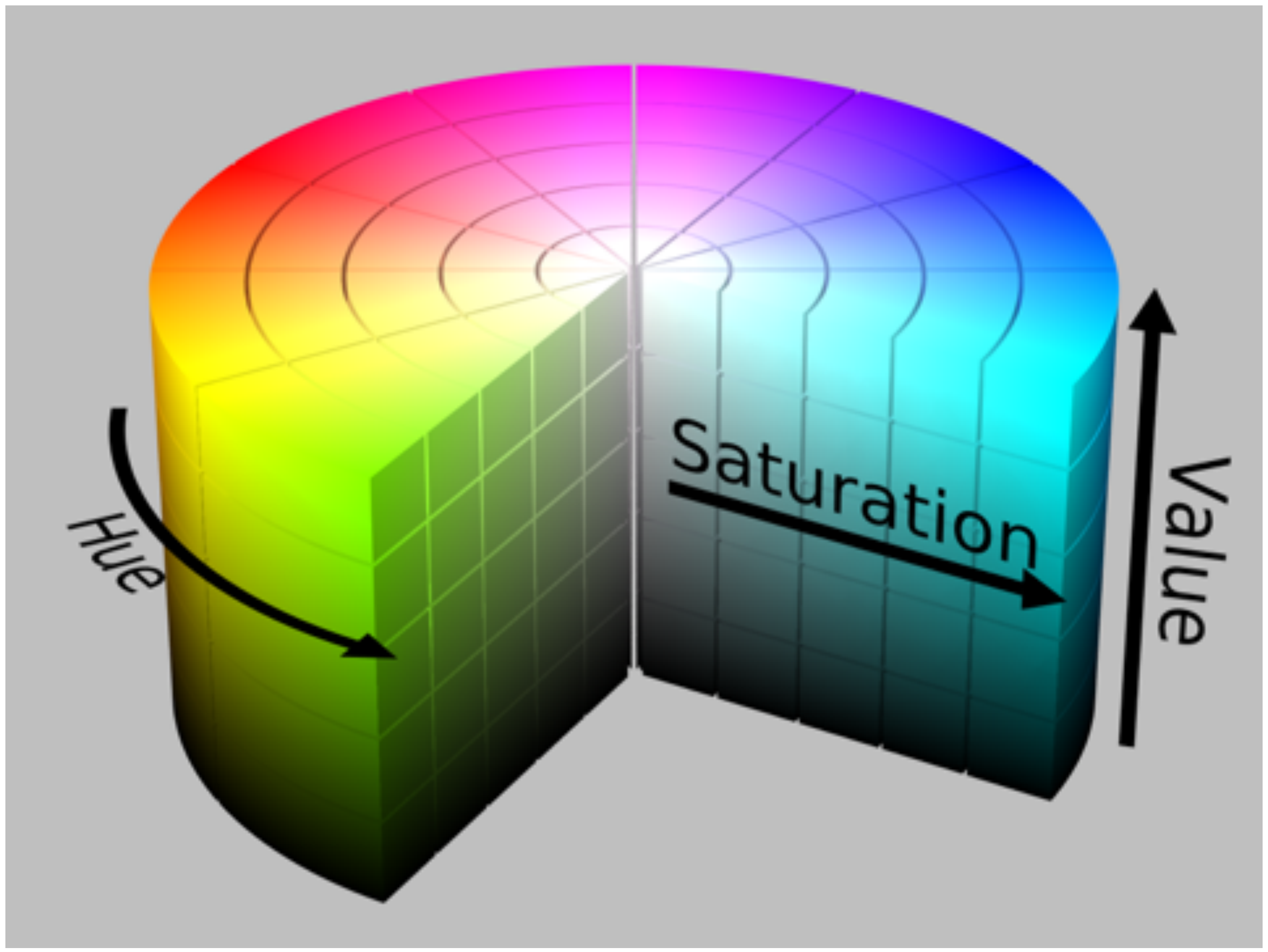
correlation x & y = 0.816
regression line: y = 3+0.5x

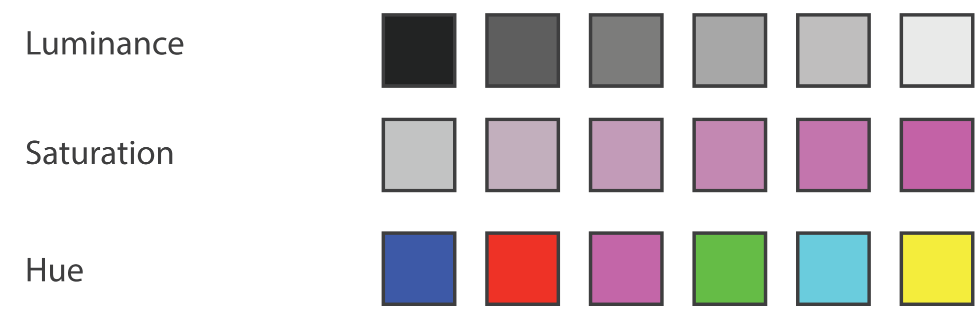

x
I
y
x
II
y
x
III
y
x
IV
y

Selective attention

https://www.youtube.com/watch?v=vJG698U2Mvo

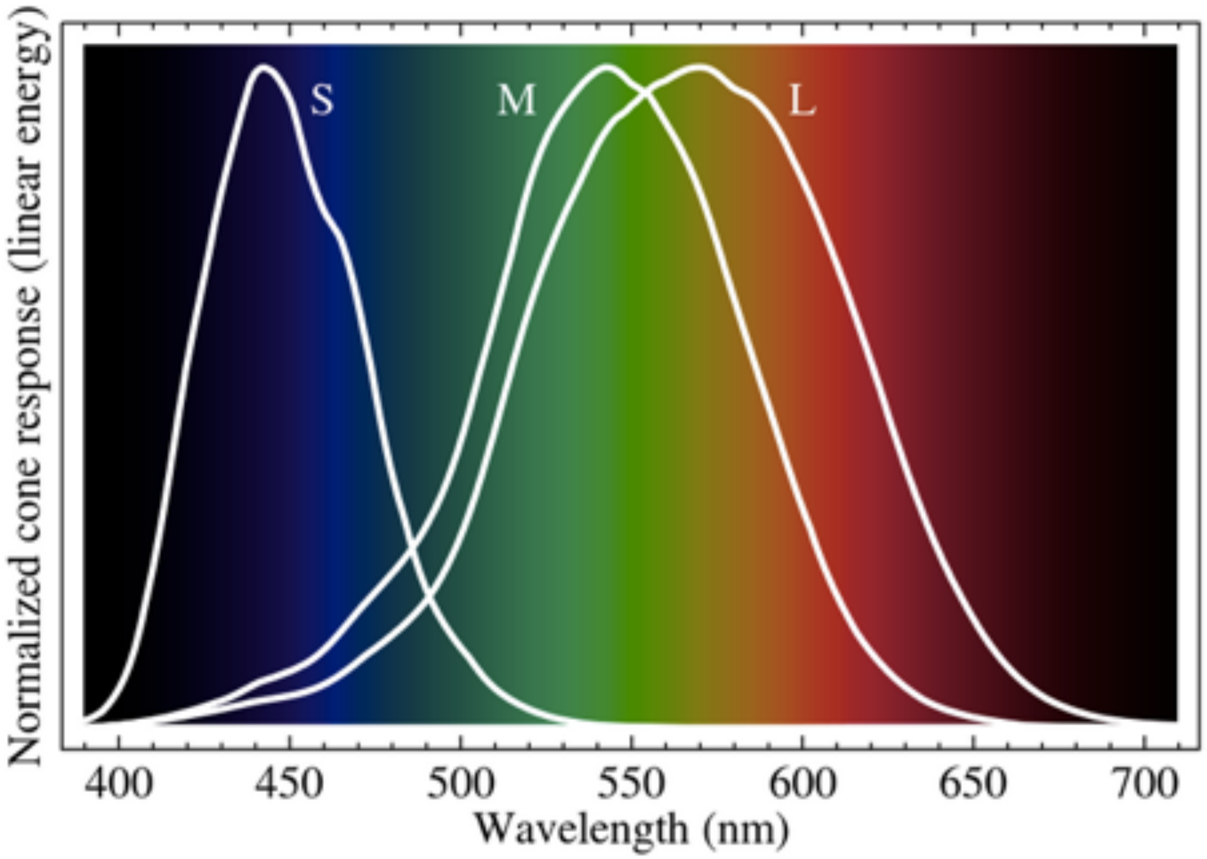
About colour



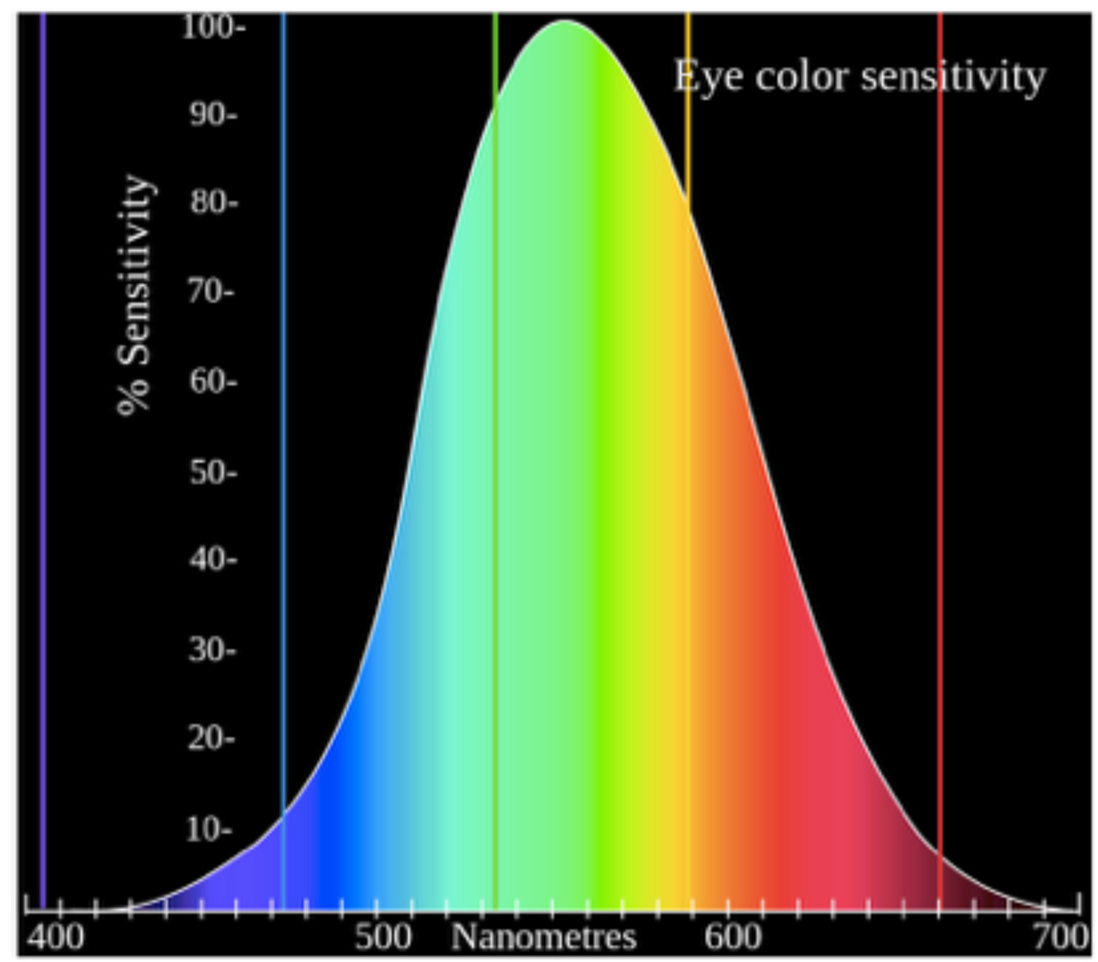


Use **HSV** (hue - saturation - value) instead of RGB

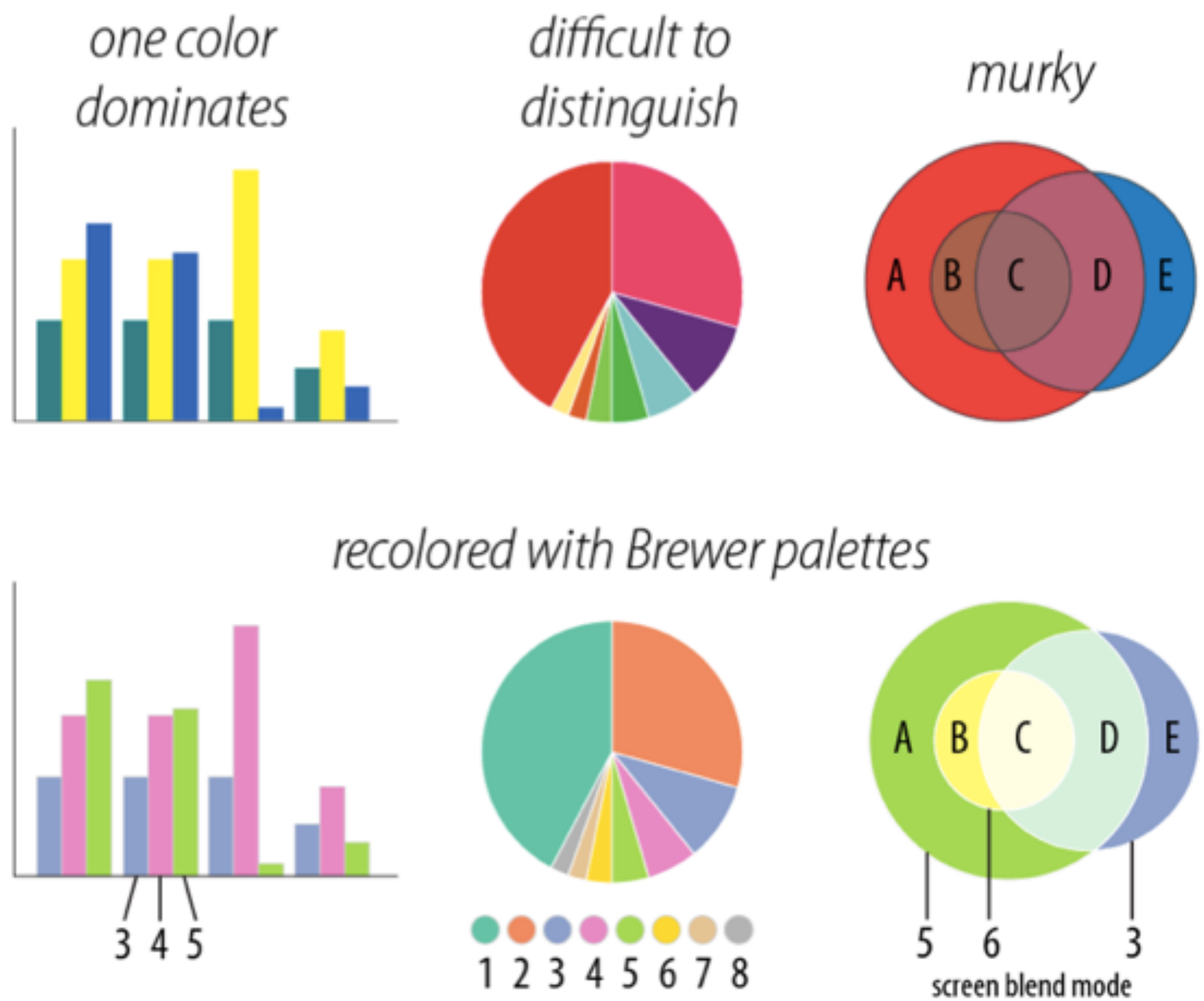
- HSV: separates *luma* (intensity) from *chroma* (colour)
- RGB has to do with *implementation*; HSV has to do with *perception*



sensitivity of S/M/L cones to colour wavelengths



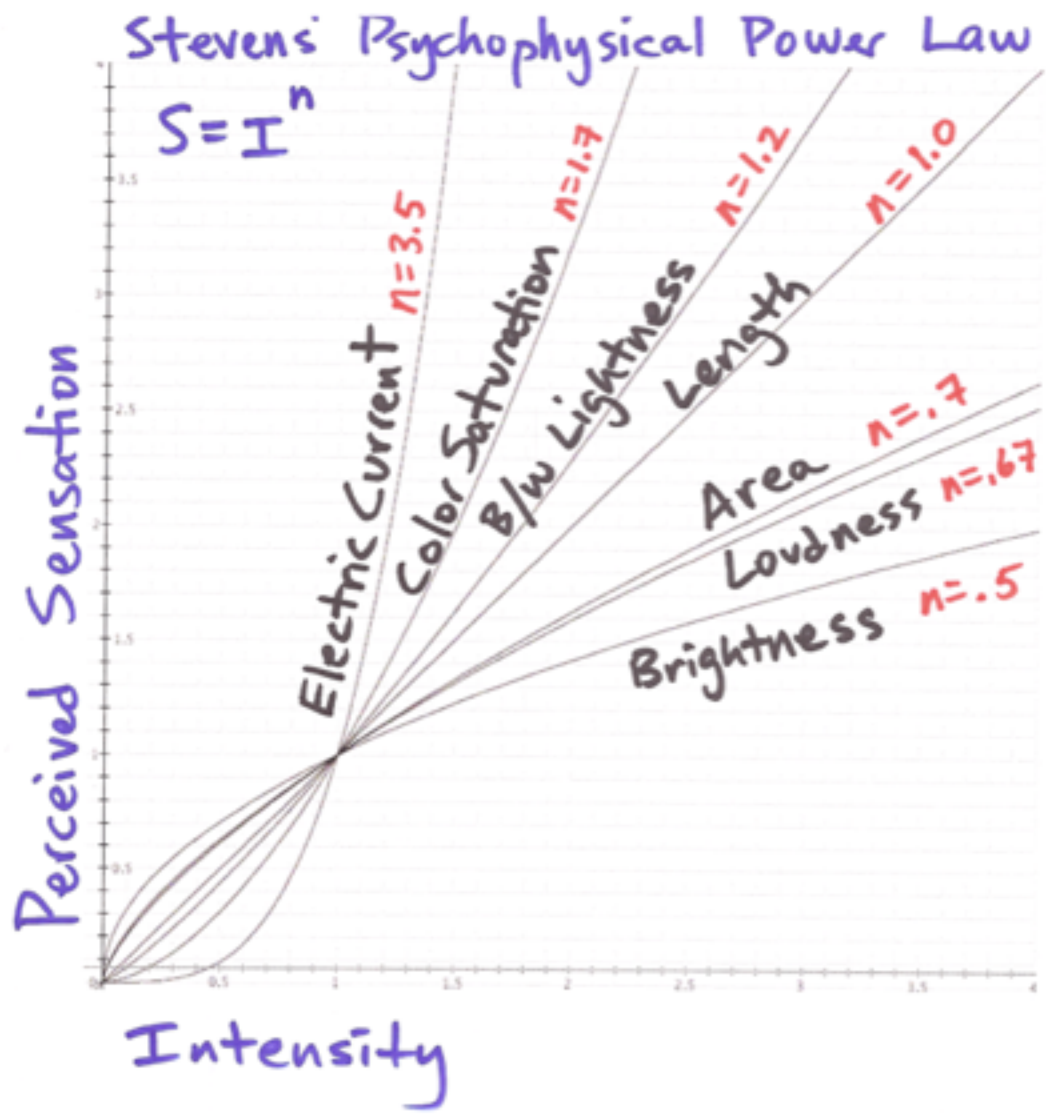
colorbrewer2.org - Never pick your own colour



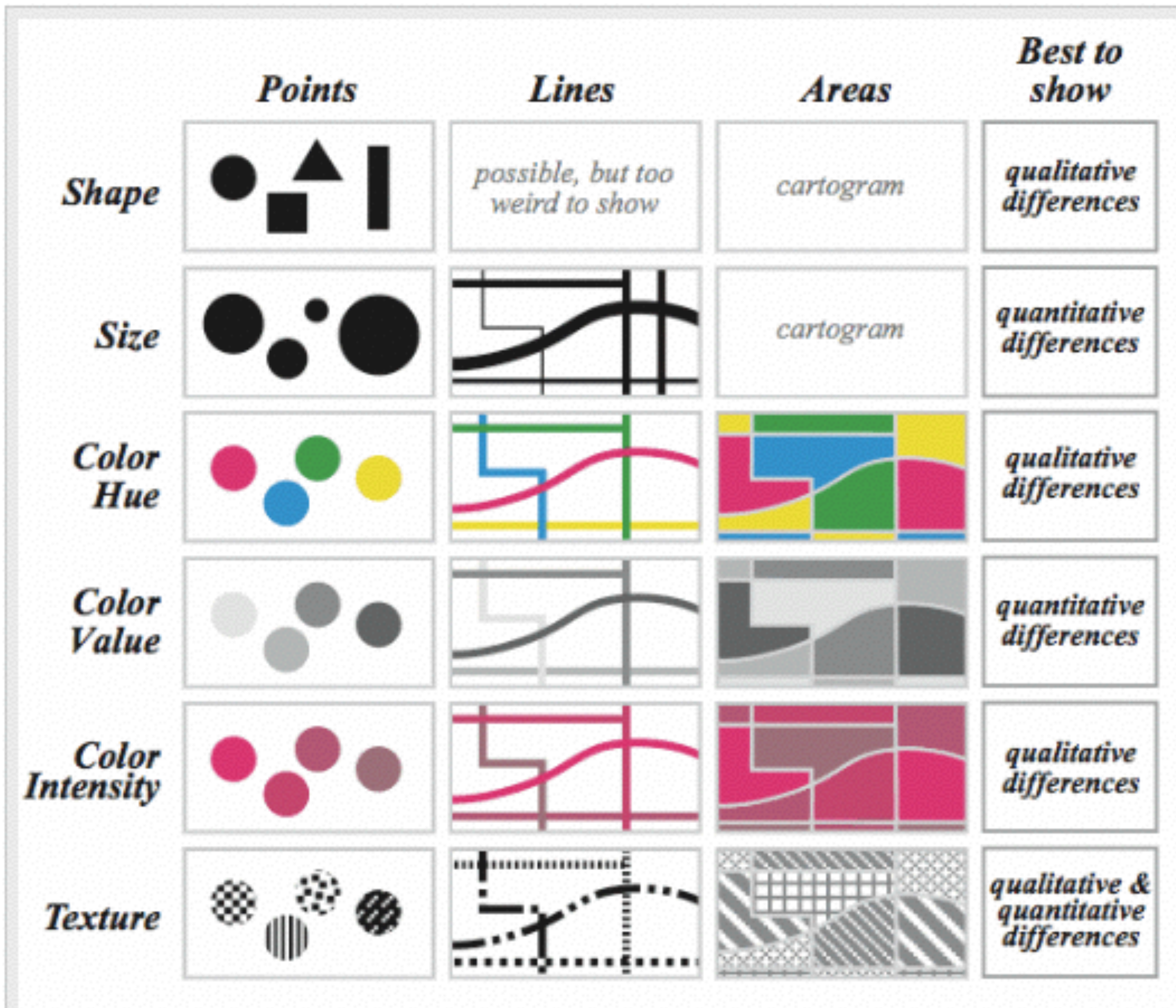
in R: please use RColorBrewer!

Steven's psychophysical law

= proposed relationship between the magnitude of a physical stimulus and its perceived intensity or strength

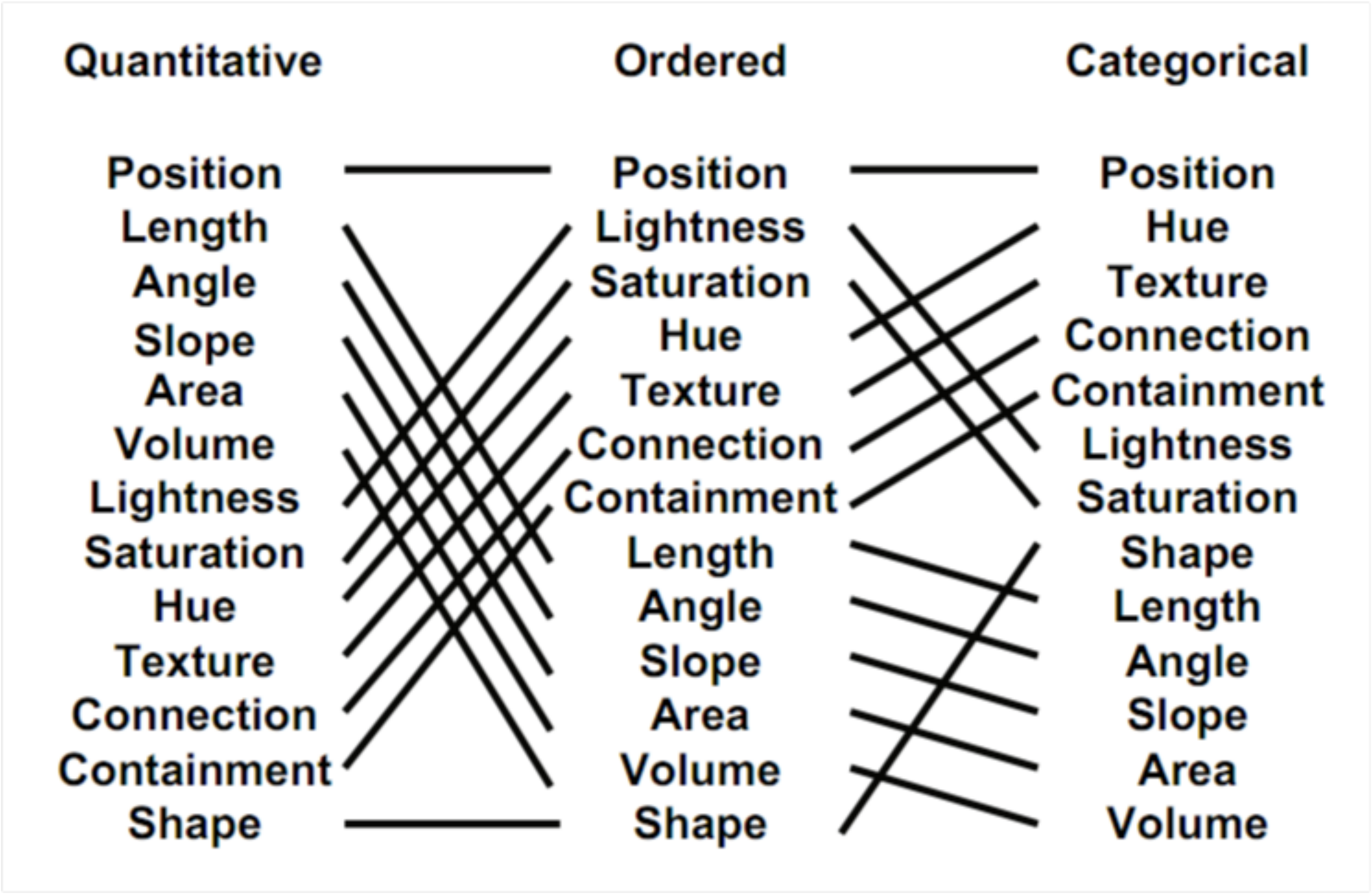


Semiology of graphics



Accuracy of quantitative perceptual tasks

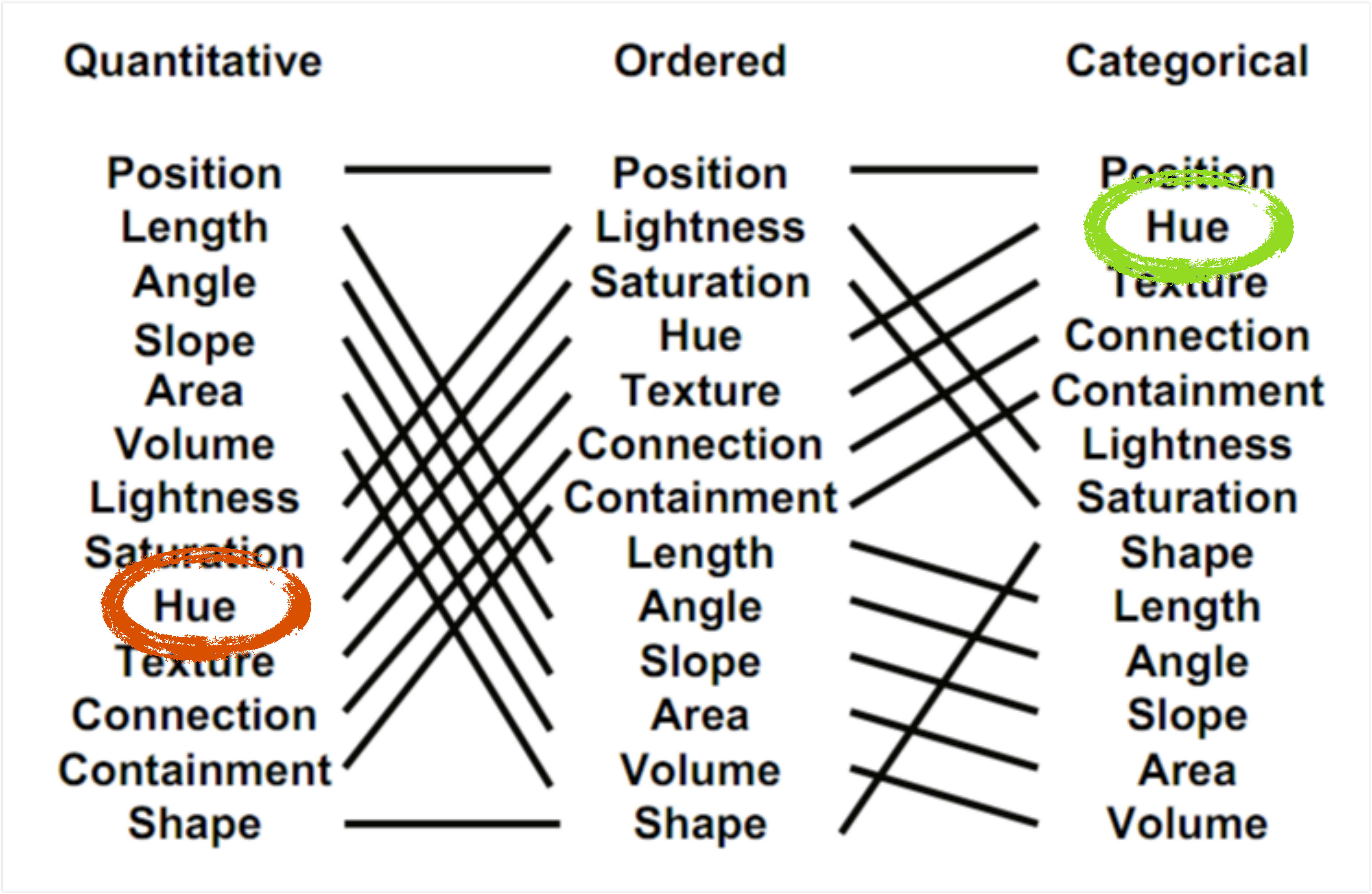
how much (quantitative) what/where (qualitative)



McKinlay

Accuracy of quantitative perceptual tasks

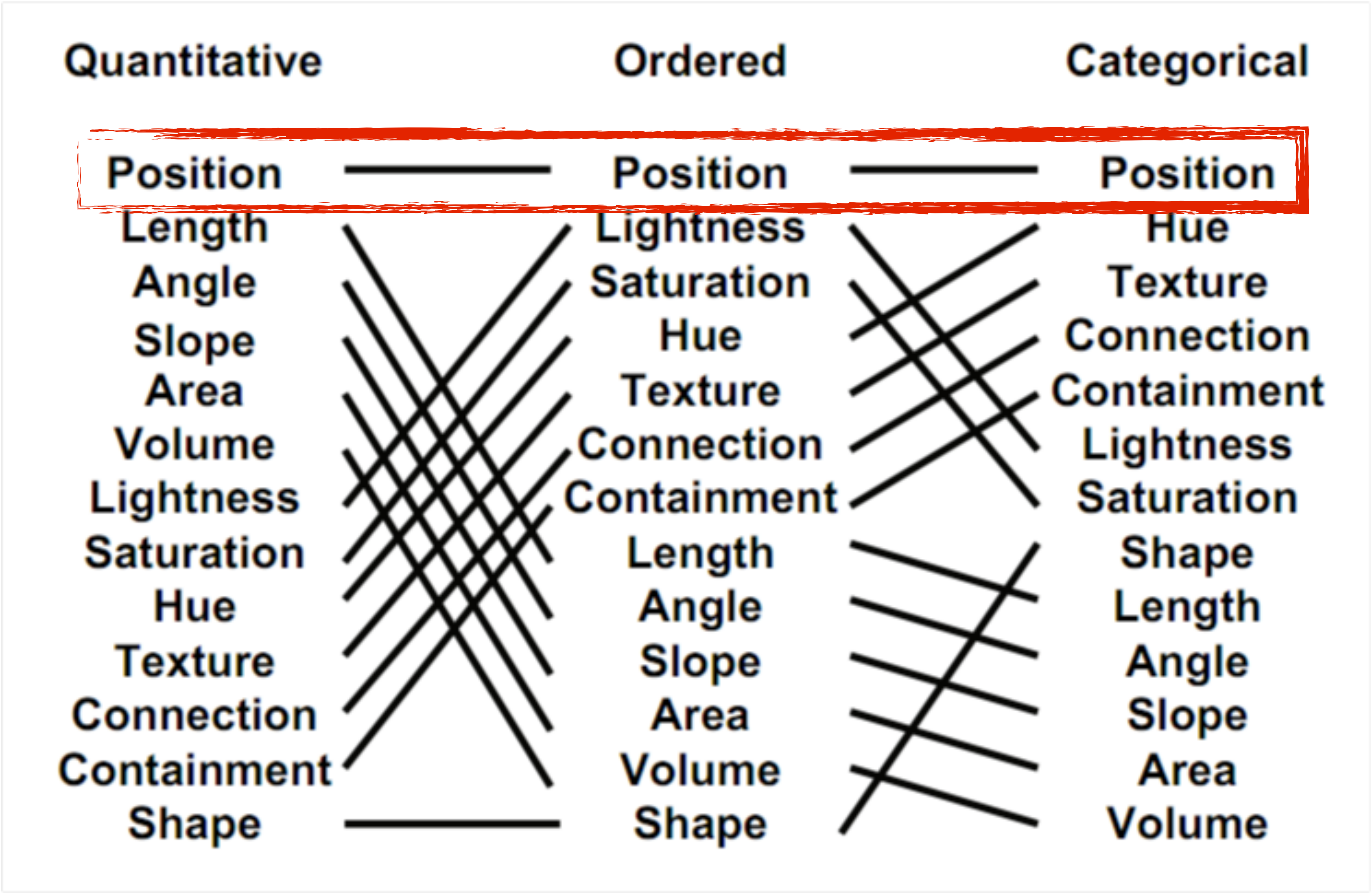
how much (quantitative) what/where (qualitative)



McKinlay

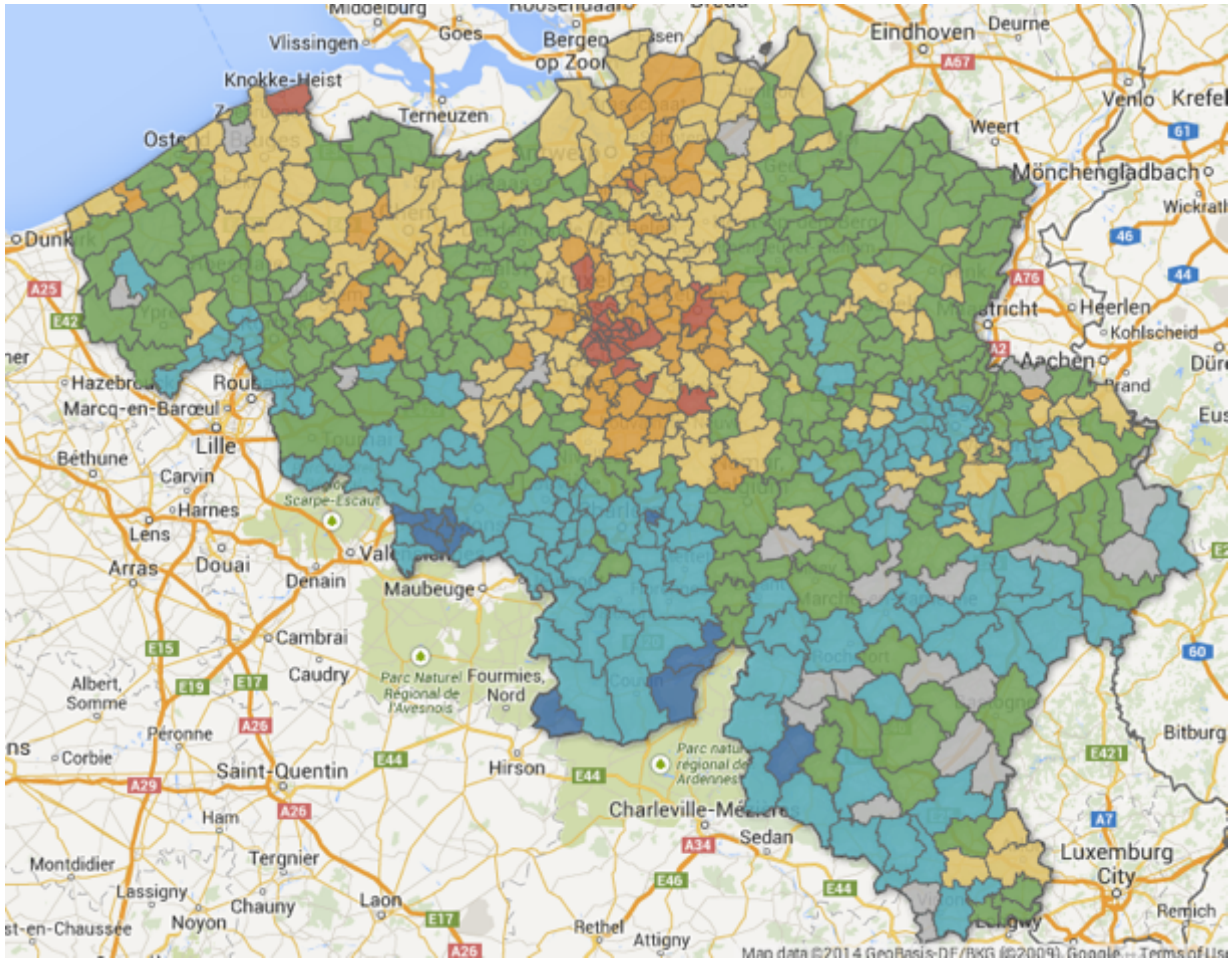
Accuracy of quantitative perceptual tasks

how much (quantitative) what/where (qualitative)



"power of the plane"

McKinlay



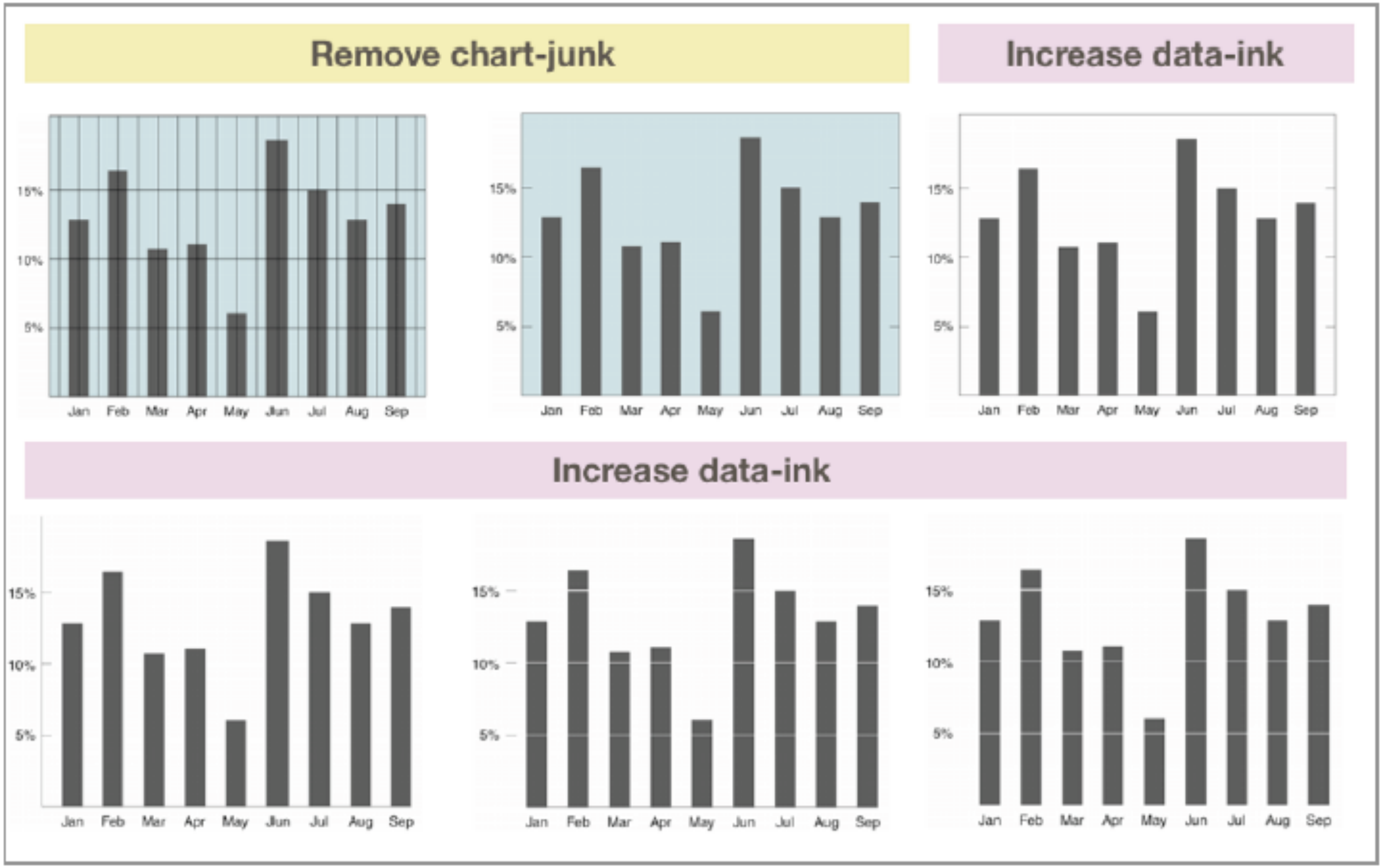
average prices per
municipality in Belgium
(De Standaard, 24/8/2014)

Rules of thumb

(partly based on book Munzner)

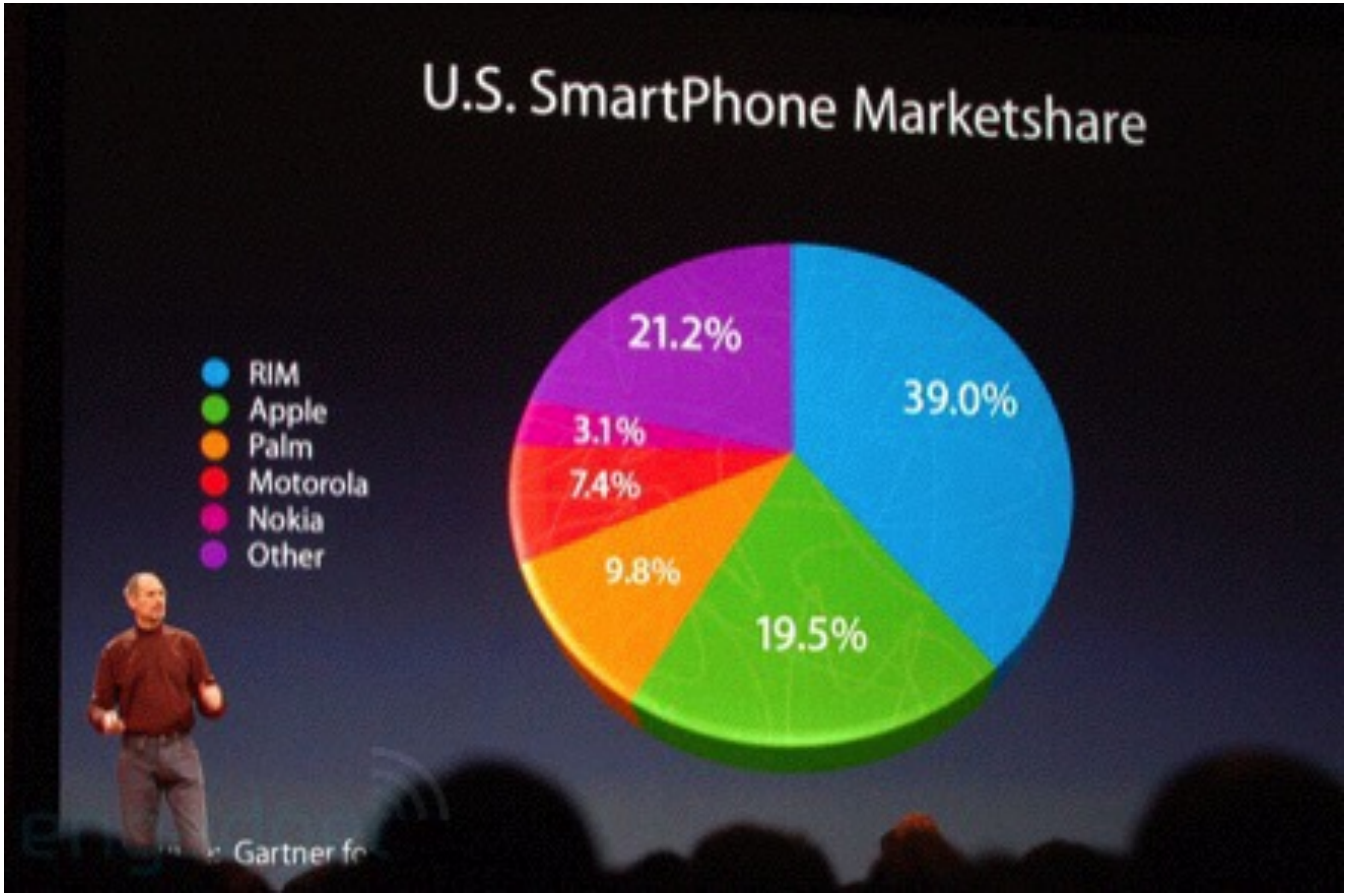
1. maximize data-to-ink ratio (Edward Tufte)
2. beware of the lie-factor
3. no unjustified 3D
4. eyes beat memory
5. overview first, zoom & filter, details on demand
6. don’t overengineer

1. maximize data-to-ink ratio

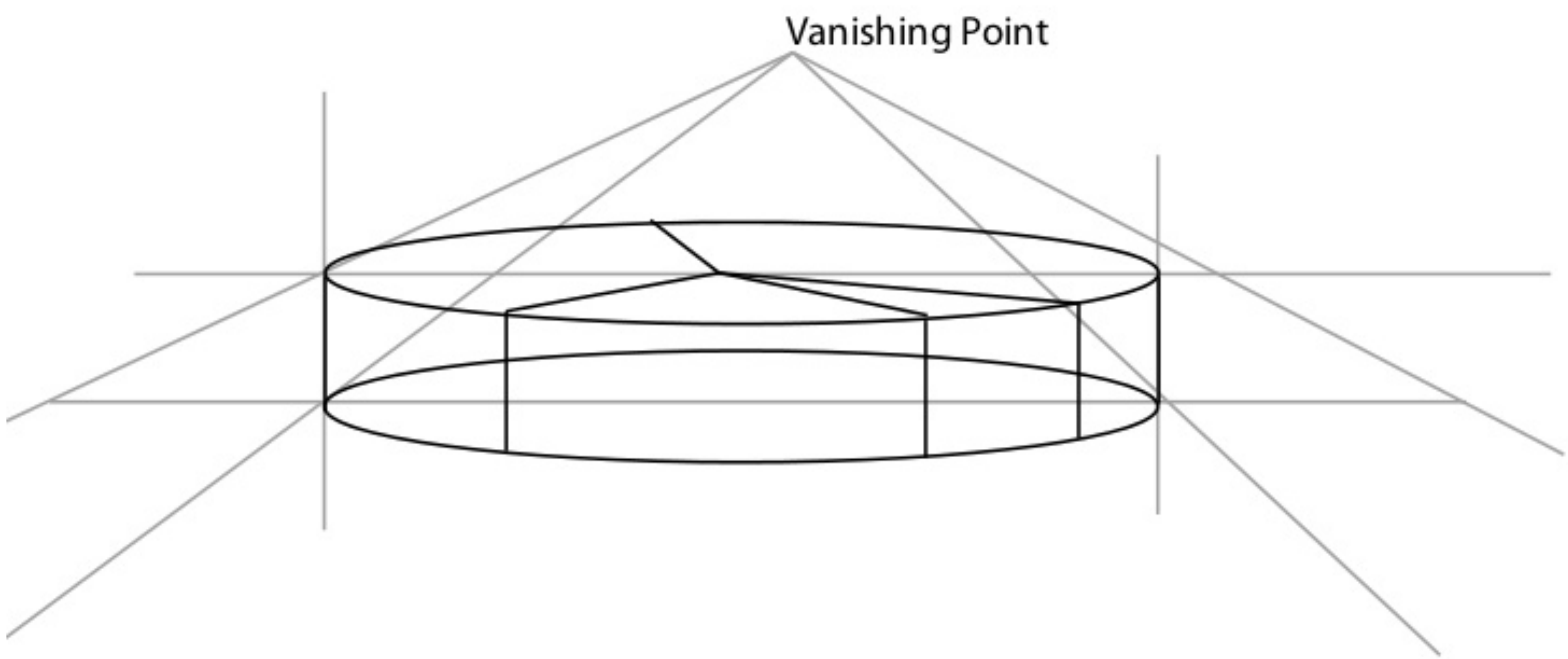


$$\textbf{“data-to-ink ratio”} = \frac{\text{data-ink}}{\text{total ink}} = 1 - \text{proportion of graphic that can be erased}$$

2. beware of the lie-factor



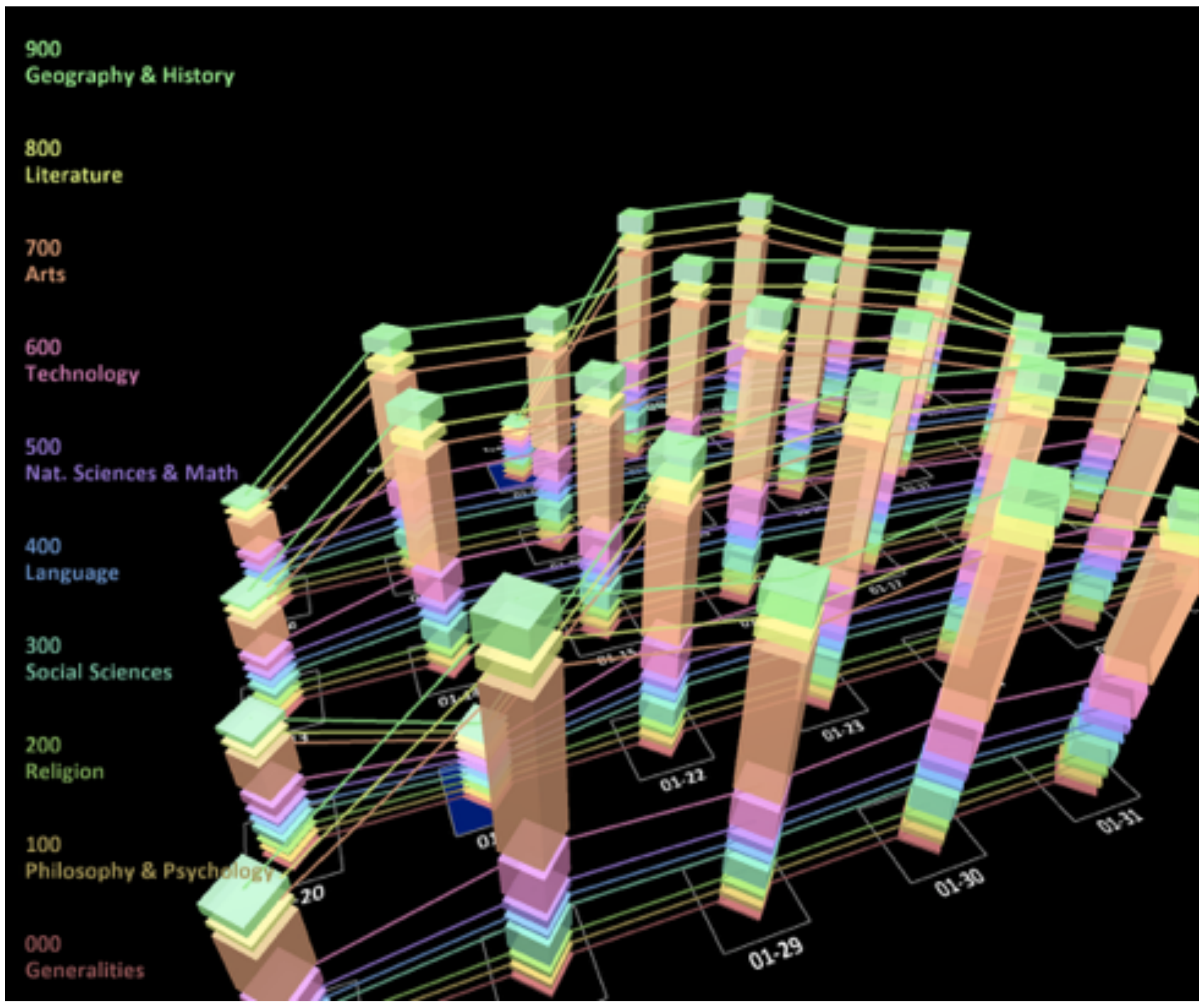
3D Charts!

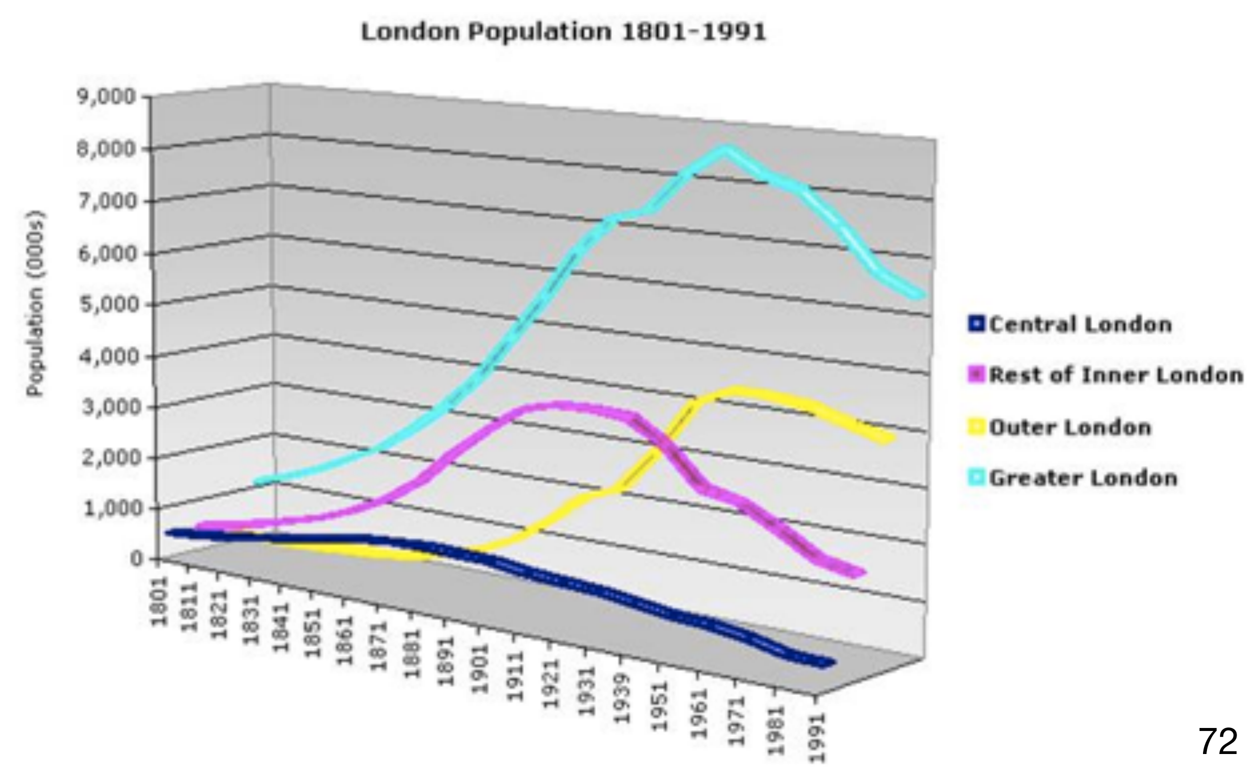


$$\textbf{“lie factor”} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

3. no unjustified 3D

issues with occlusion, perspective distortion, text legibility, …





4. eyes beat memory

animation vs side-by-side views

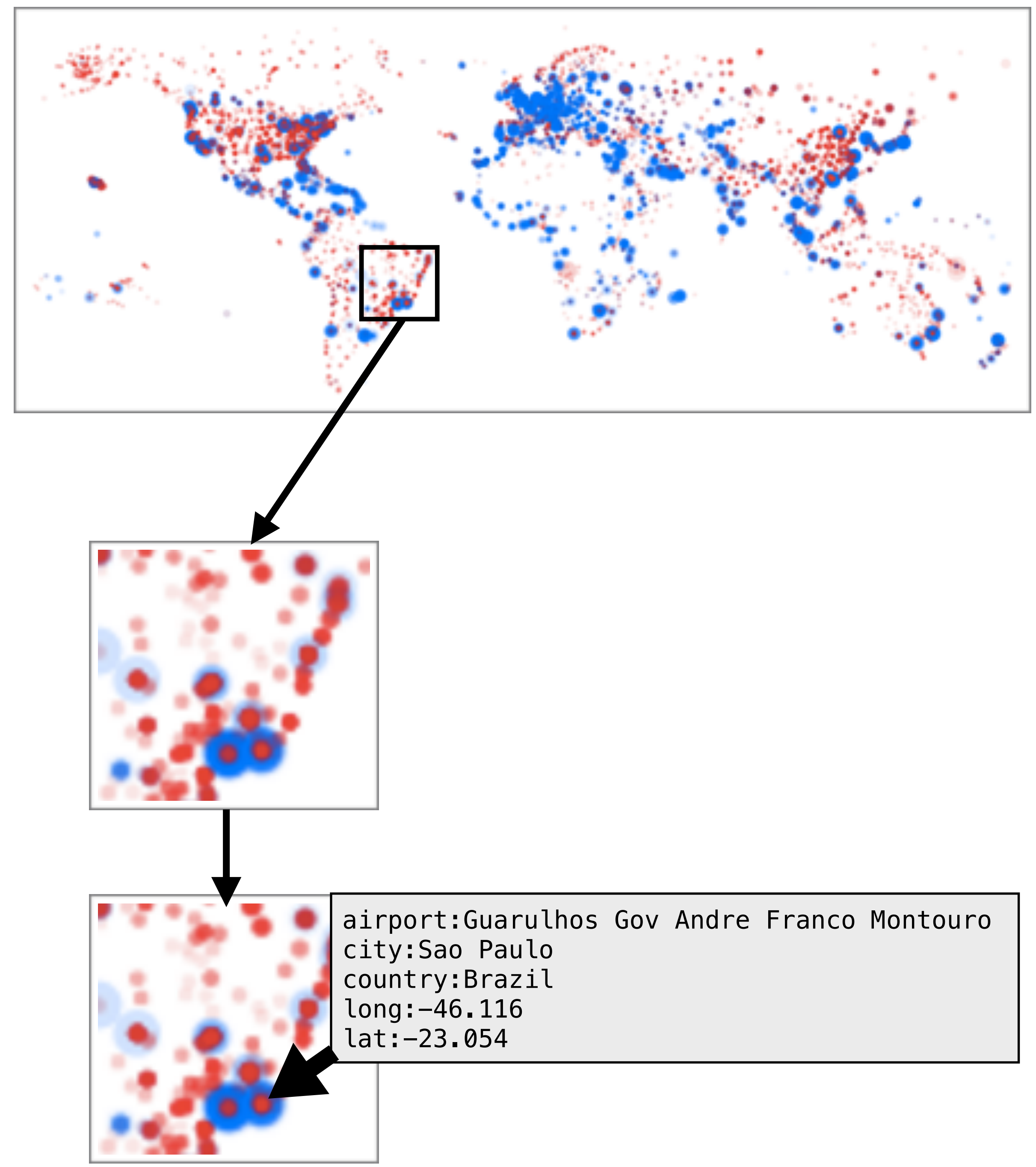
switch between different views that are visible at same time = lower cognitive load than consulting memory to compare current view with what was seen before

=> try to represent dynamic processes in a dynamic way

5. overview first, zoom & filter, details on demand

task taxonomy Ben Schneiderman:

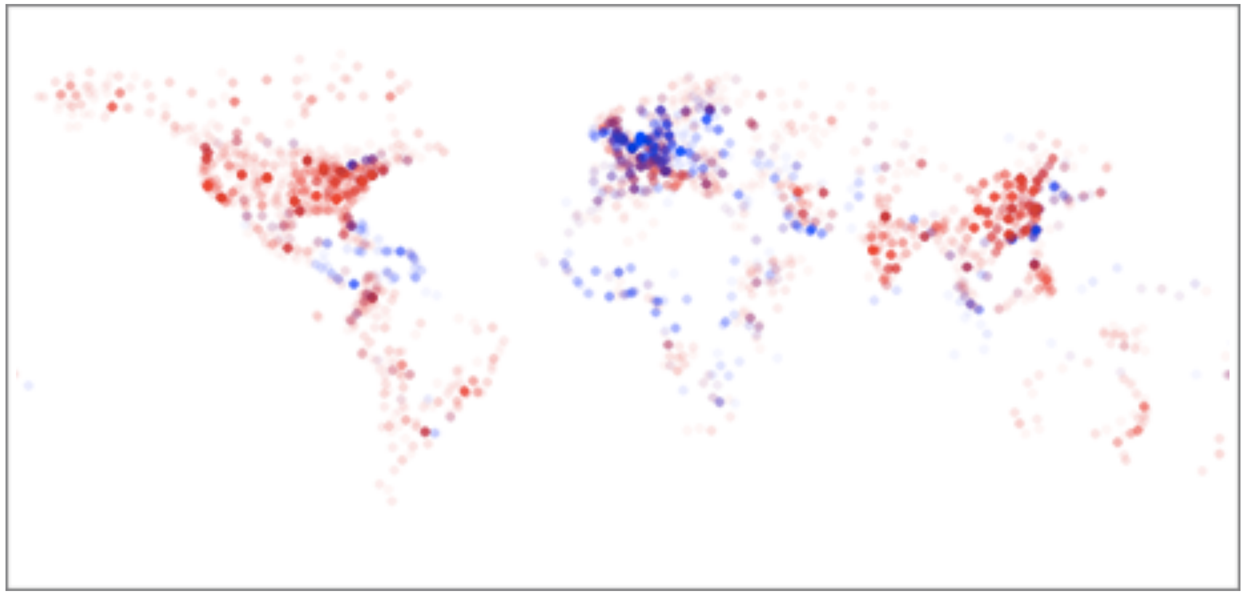
- *Overview*: see overall patterns in data
- *Zoom*: see a subset of data
- *Filter*: see a subset based on values
- *Detail on demand*: see values of items
- *Relate*: compare values
- *History*: keep track of actions
- *Extract*: mark and captur



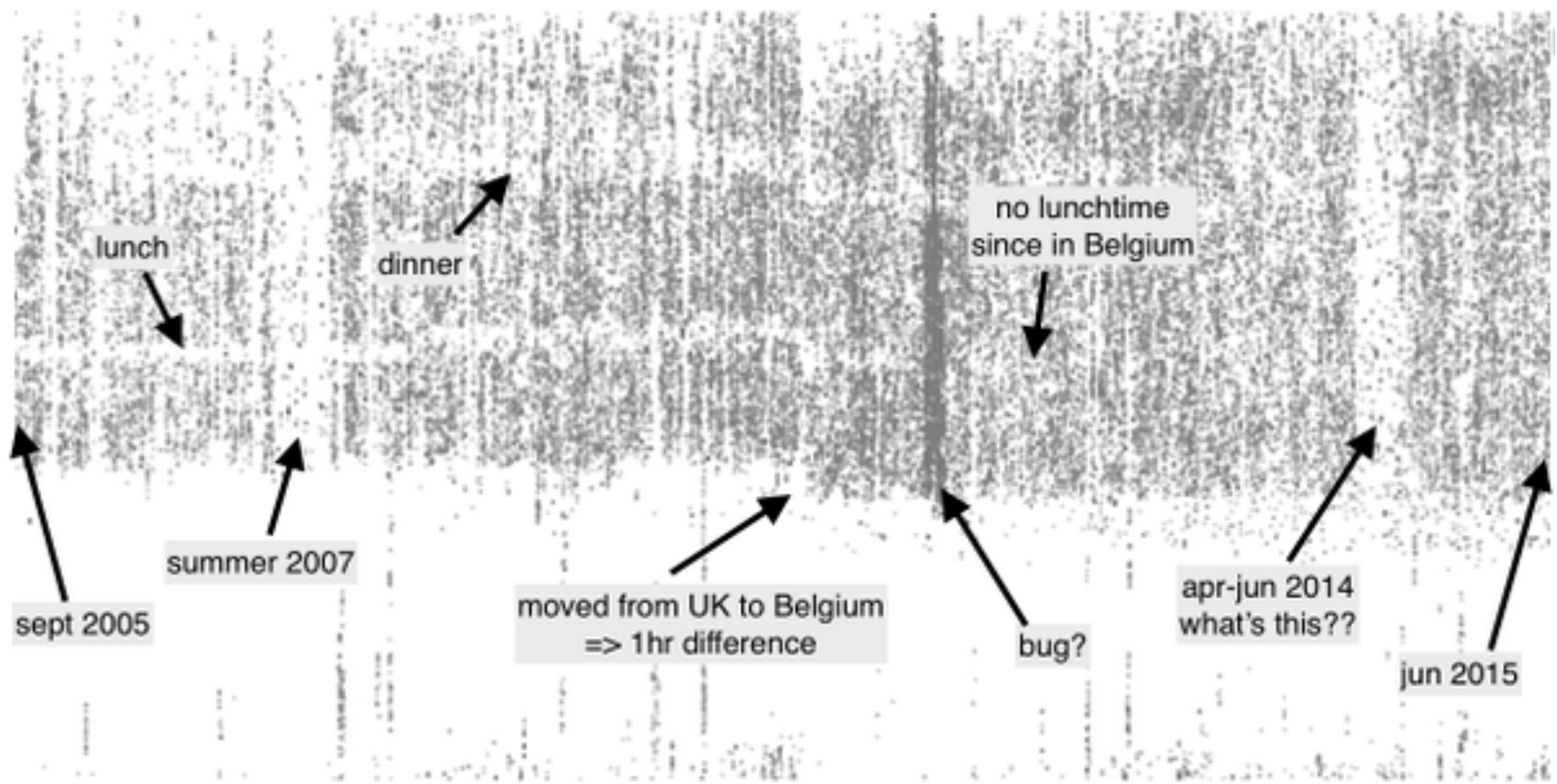
6. “underengineer”, if possible

important advantage of human in the loop vs algorithm: you can take shortcuts

- keep interaction simple (see mouse position & data filter flight patterns; vda-lab.be blog - hands-on visualisation using p5)
- might be OK to not handle edge cases in fast prototyping
- simple raw data visualization can have emergent properties

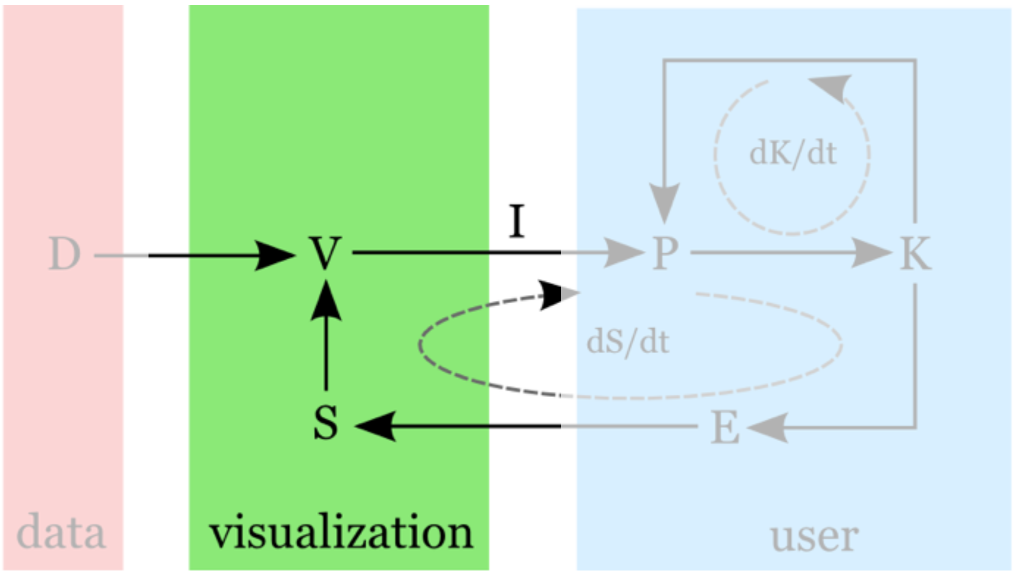


what is drawn: roads
what we see: cities, mountain ranges, population density, …
=> we observe higher-level patterns; not the raw data

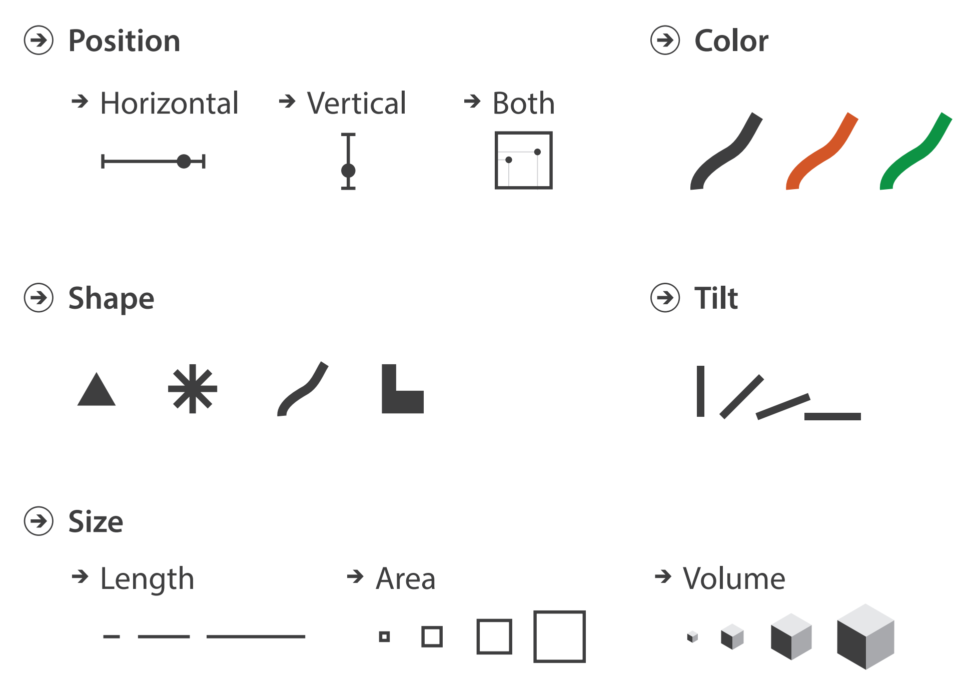


“send” date & time for all my
outgoing email since 2005

Visualization foundations and techniques

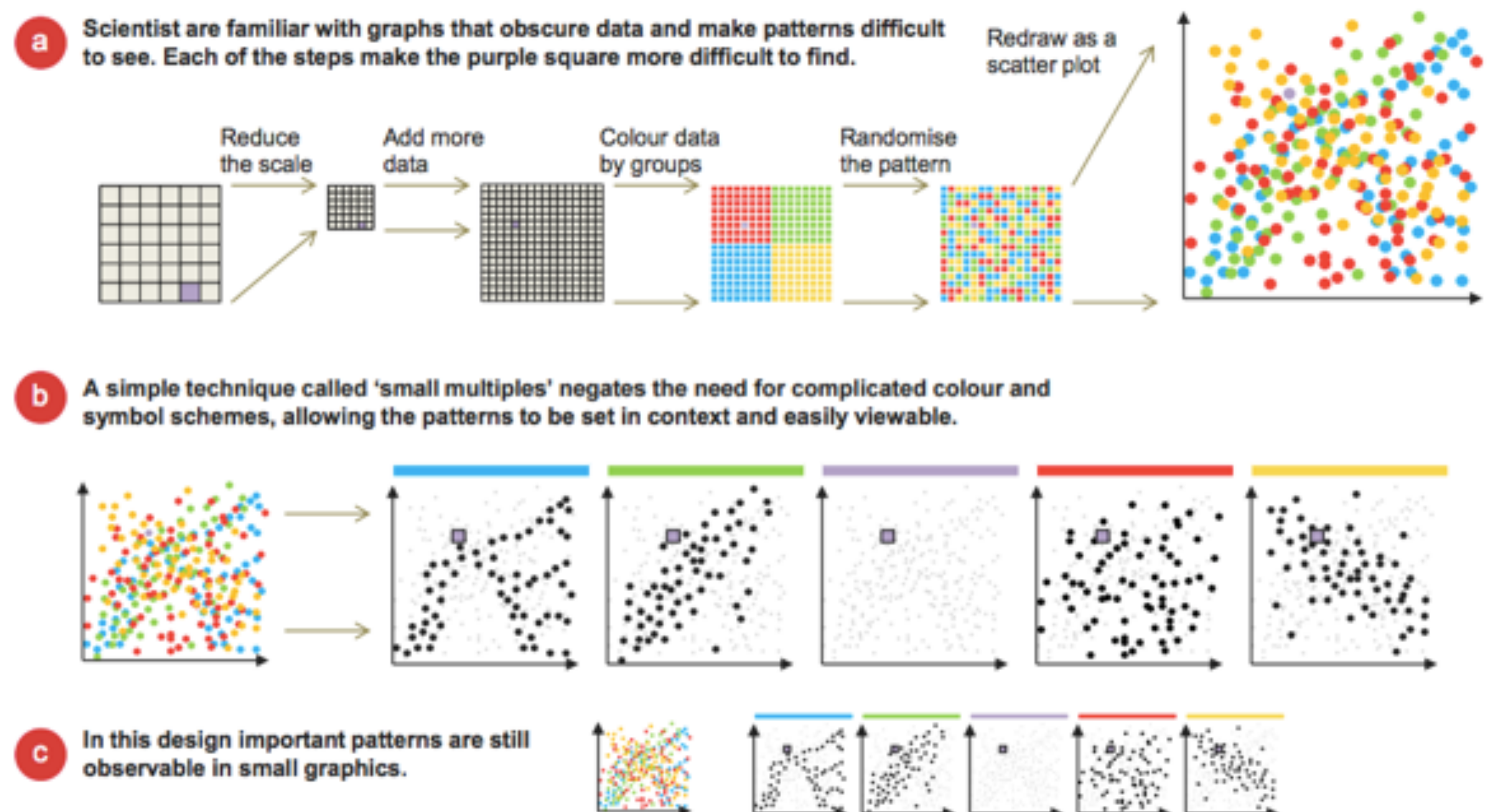


Mapping data to marks with properties



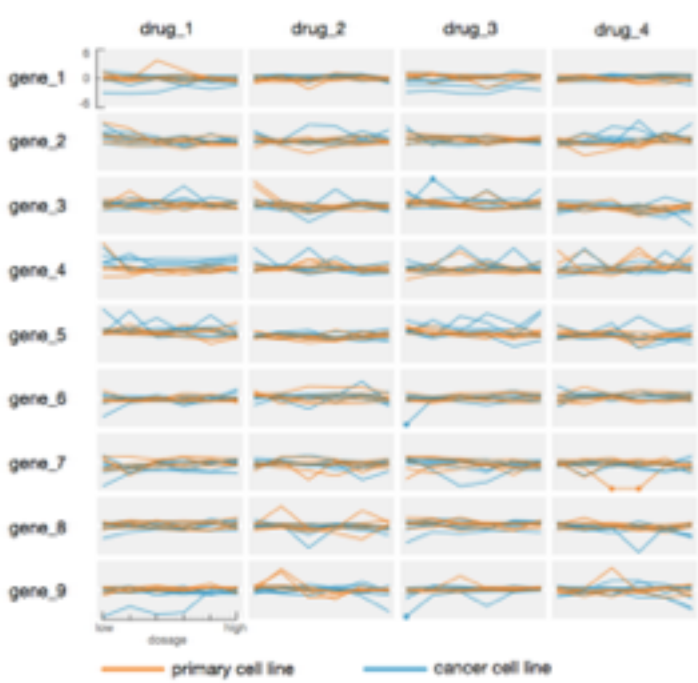
Combining visuals

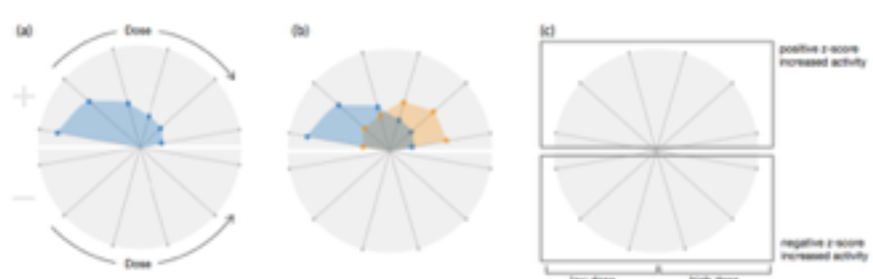
- **juxtapose**/facet,
  e.g. *small multiples*
  - easier to see trends
  - pop-out effect
  - very powerful!



reference?

L1000 data

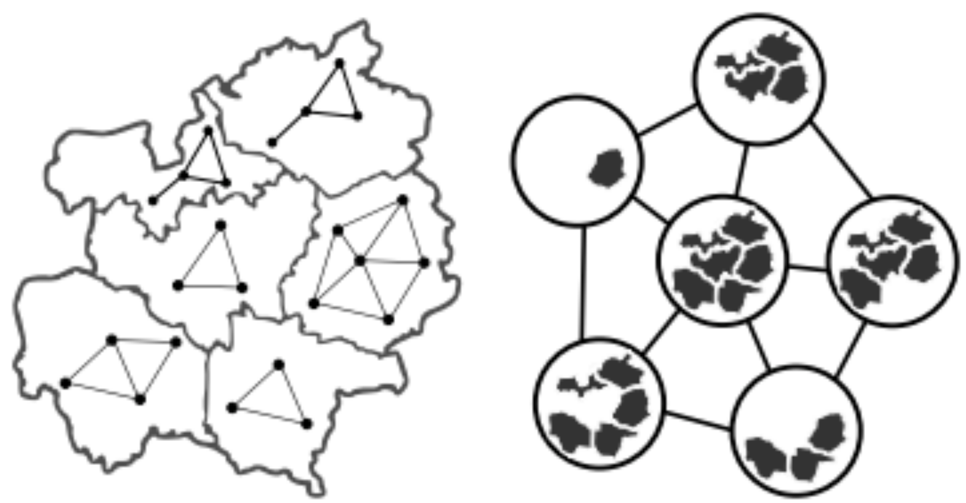




- **superimpose**

Hadlak et al, 2015

- **embed**/nest
  - e.g. network + map: each can be “base”



Hadlak et al, 2015

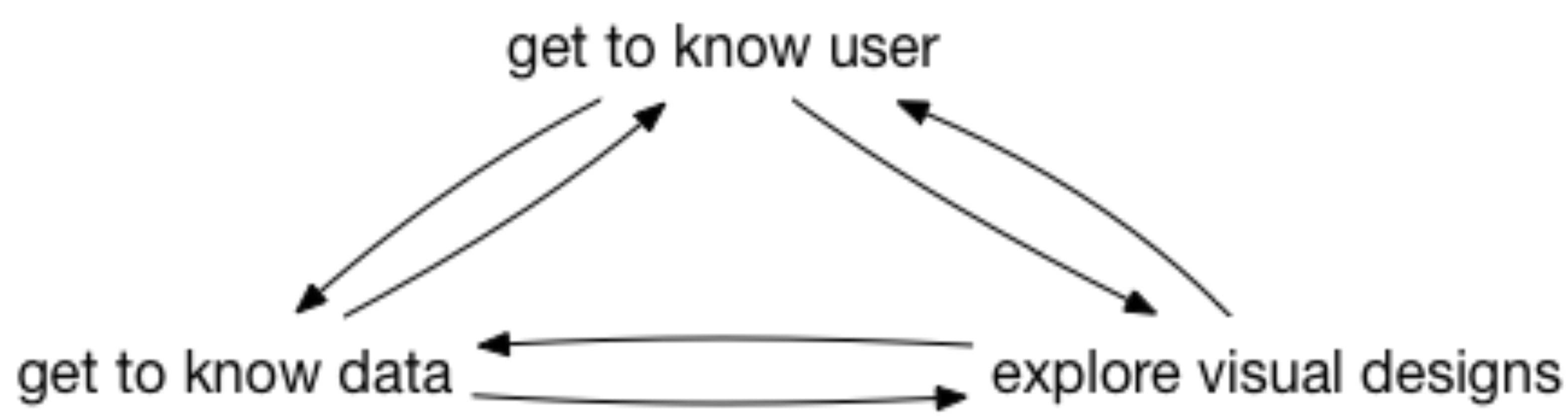
Creating data visualisations - the process

1. Get to know the user

2. Get to know the data

3. Create visual designs

4. Iterate



Design decision styles

1. **Unintended design** - Design decisions based on what's easiest to implement. Developer focuses on development and deployment without any consideration of what will happen when people use the tool.

2. **Self design** - Design decisions based on by developer's own use.

3. **Genius design** - Developer still does not look beyond own experience, but that experience is extensive.

4. **Task-focused design** - Developer investigates which actions the user wants/needs to perform.

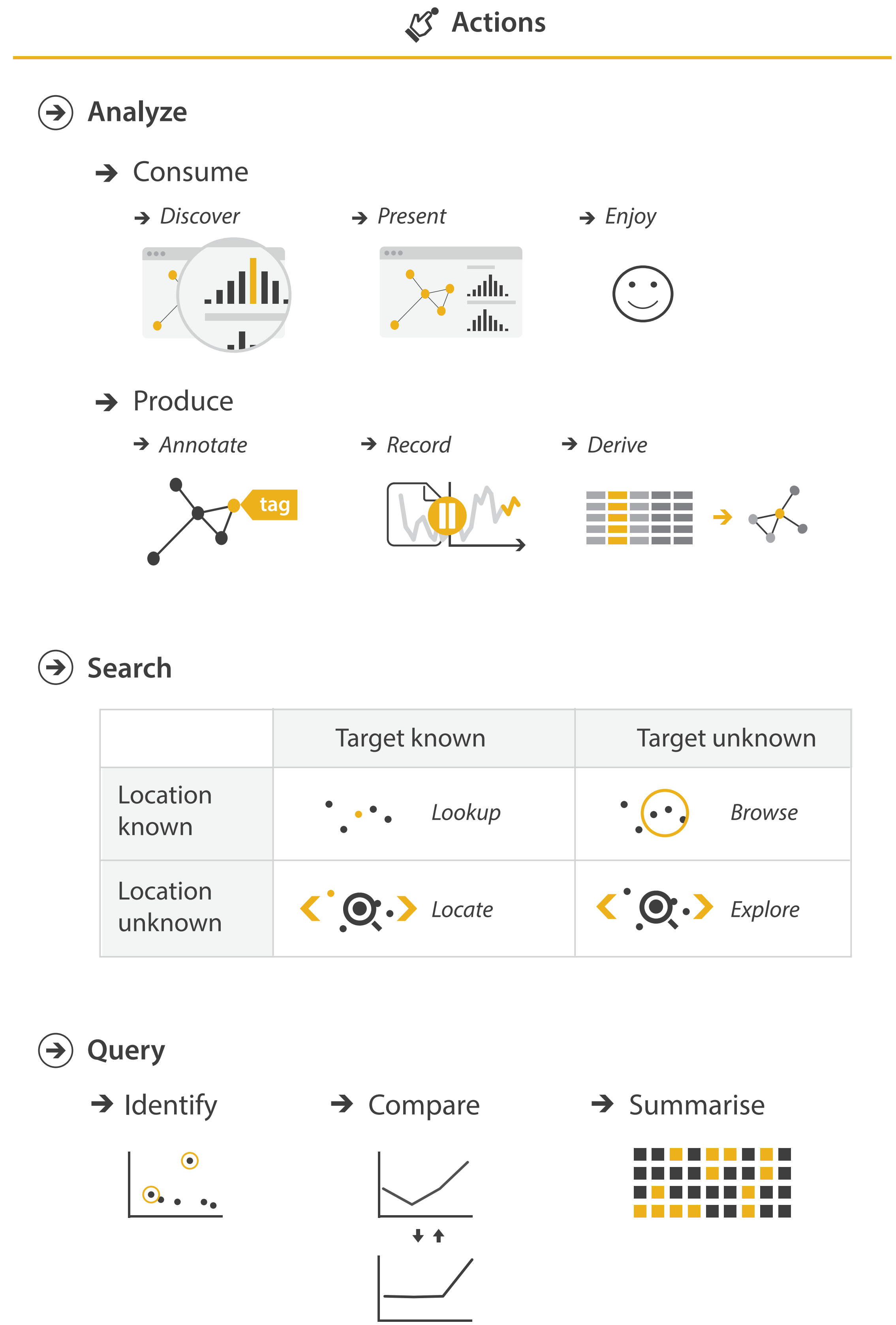
5. **Goal-focused design** - Developer goes further than activities and investigates goals, needs and contexts of the user.

1. Get to know the user

1. Talk to the user
    - what they *want* != what the *need* => need to find **underlying goals**
    - e.g. let them **imagine** what they could do if some technologies were available that are (still) science-fiction (e.g. nanobots in blood; Gaviscon commercial http://m.youtube.com/watch?v=_skKmcLdyVQ)
    - additional methods, e.g. **card sorting**
    - if possible: tape the discussion (w/ agreement)

2. Post-hoc analysis of interview (e.g. using Munzner taxonomies)

- data abstraction
- task abstraction

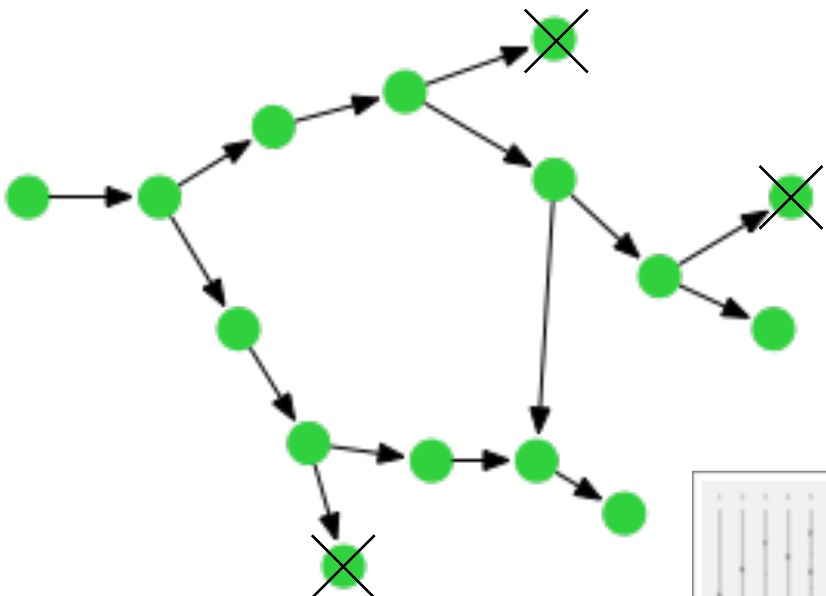


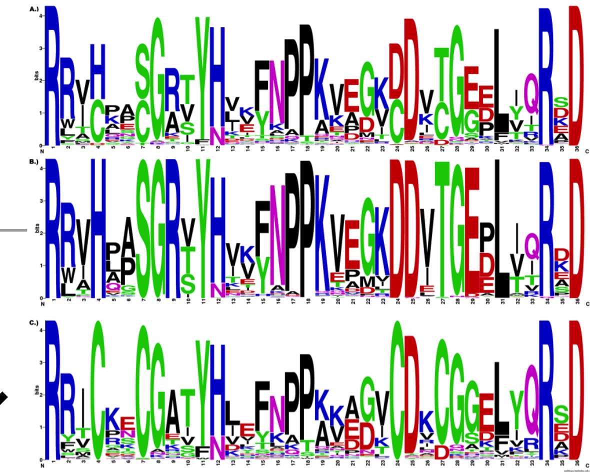
2. Get to know the data

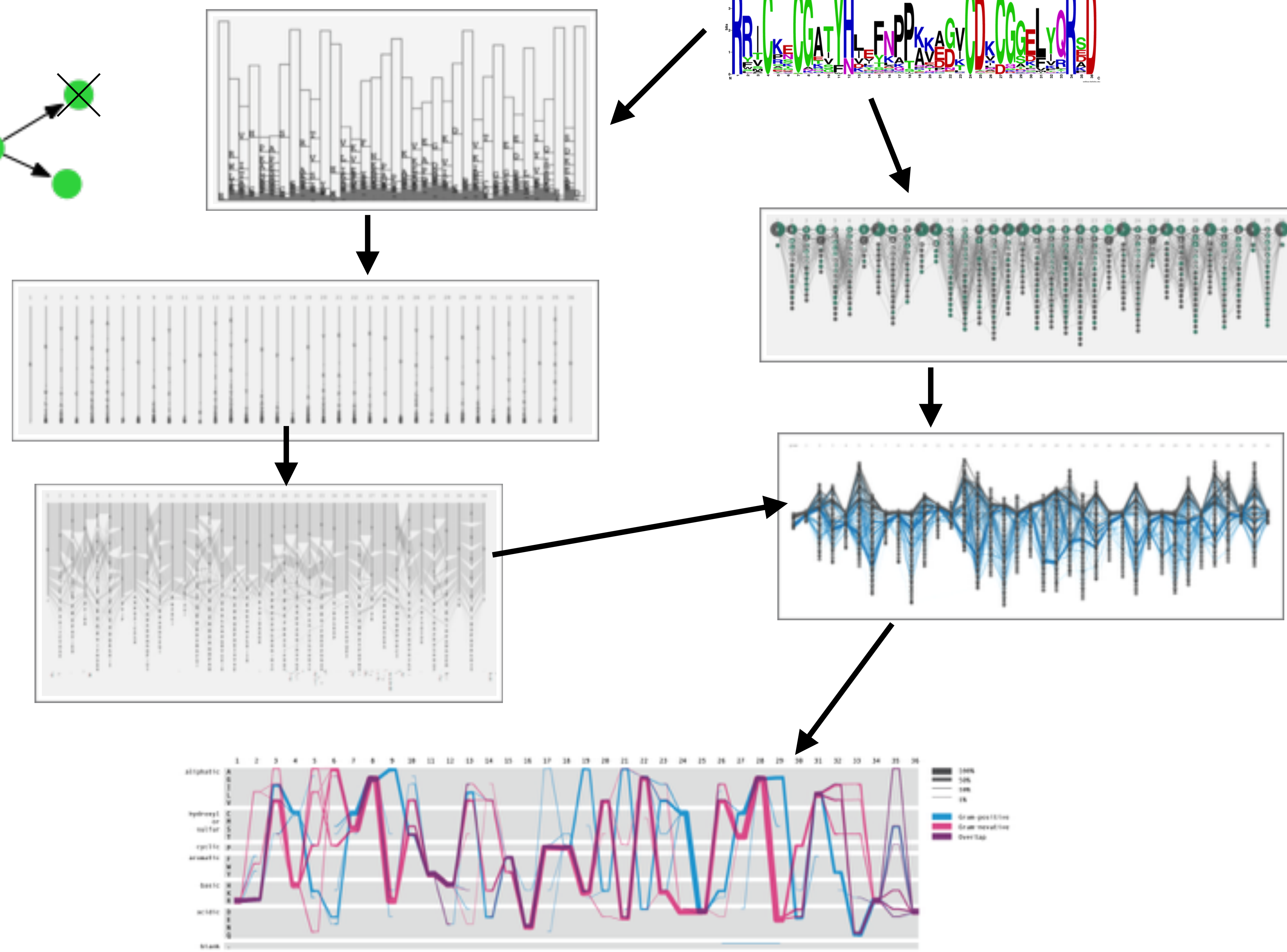
Questions to answer:

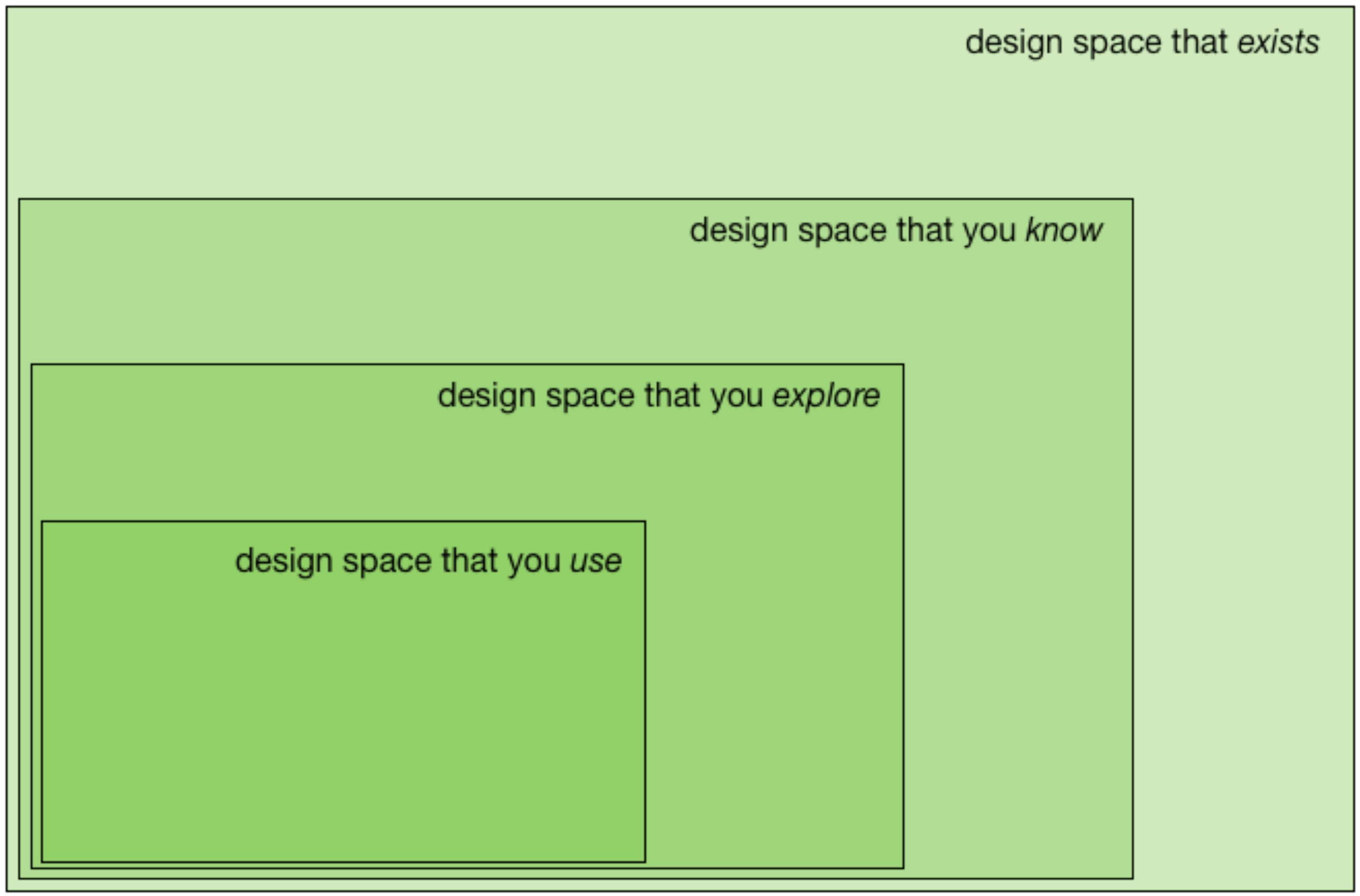
- how many **dimensions**? What are the **types** of dimensions (categorical, numerical, geo-spatial, …)?
- For each dimension: what is **distribution** of datapoints?
- Are there any **correlations** between dimensions?
- What does **principal component analysis** or **singular value decomposition** reveal?
- What does **hierarchical clustering** show?
- Are there any **local clusters**? E.g. use **topological data analysis**

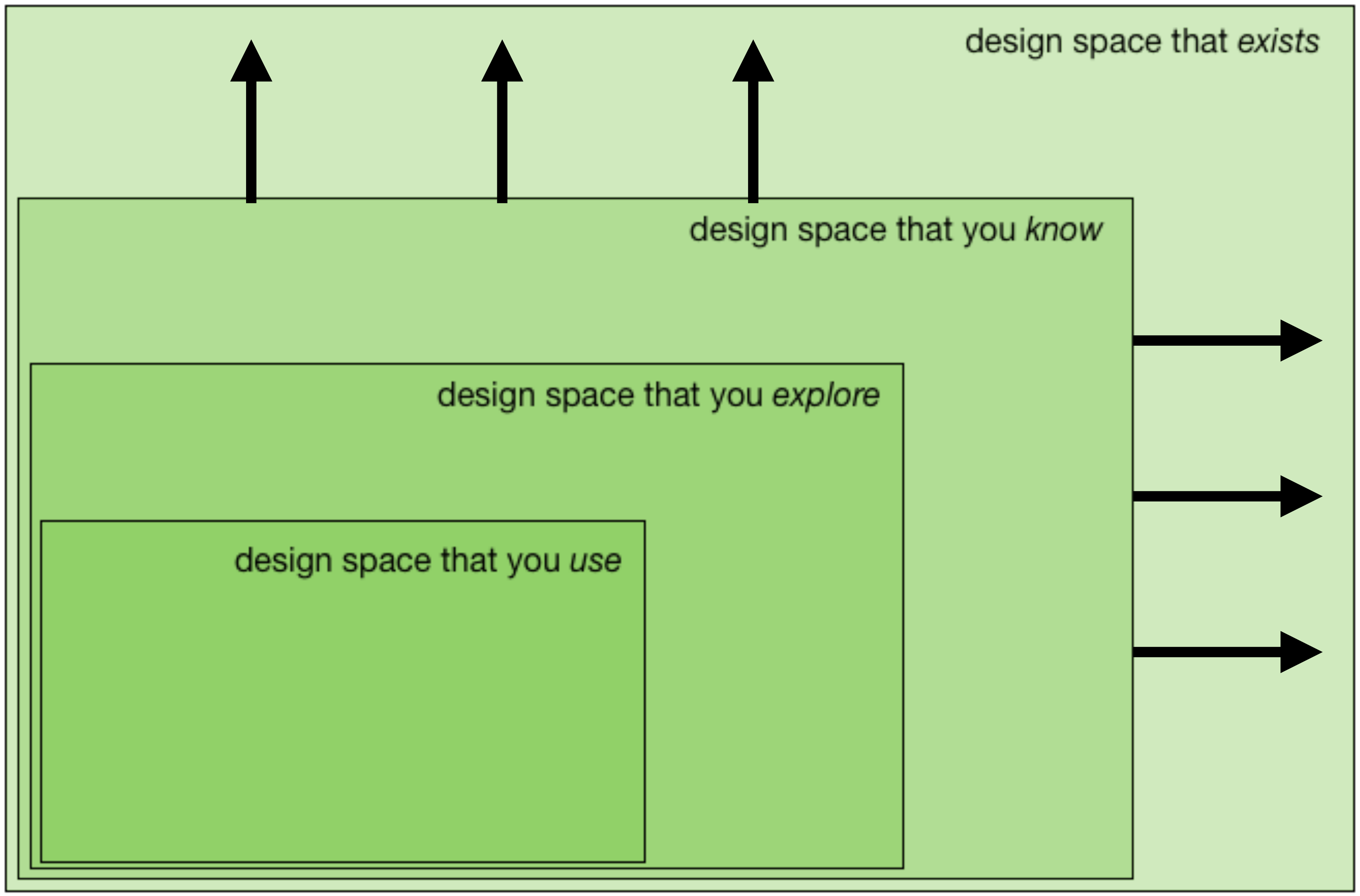
3. Create visual designs








design space that *exists*
design space that you *know*
design space that you *explore*
design space that you *use*

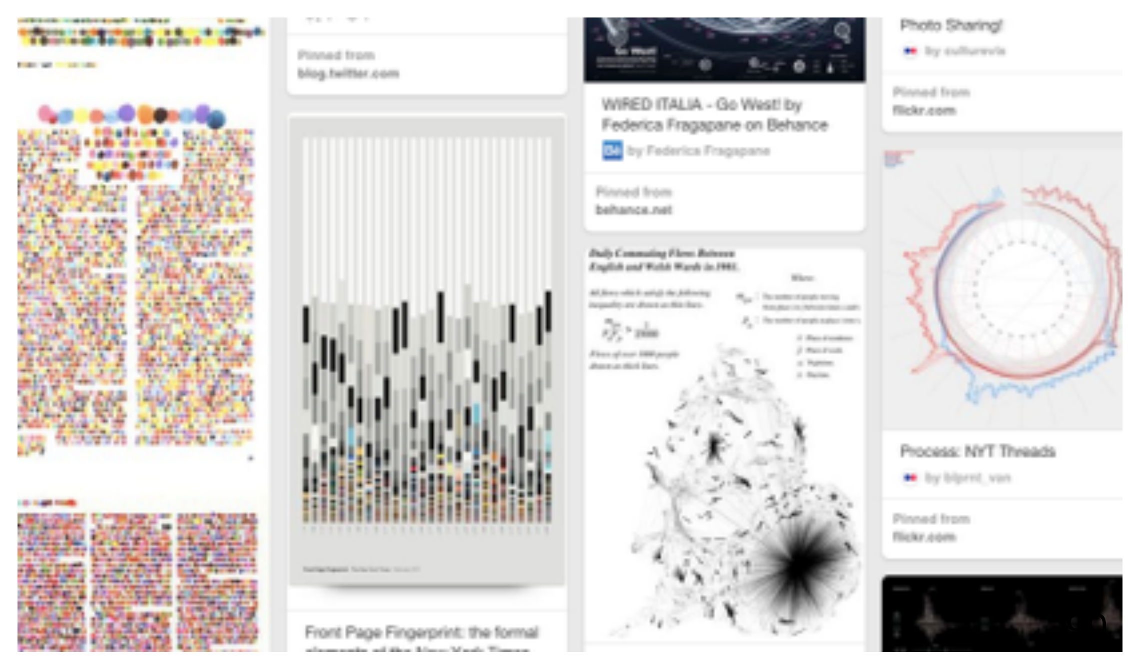


problem: initial design space that you know is small => how to expand?

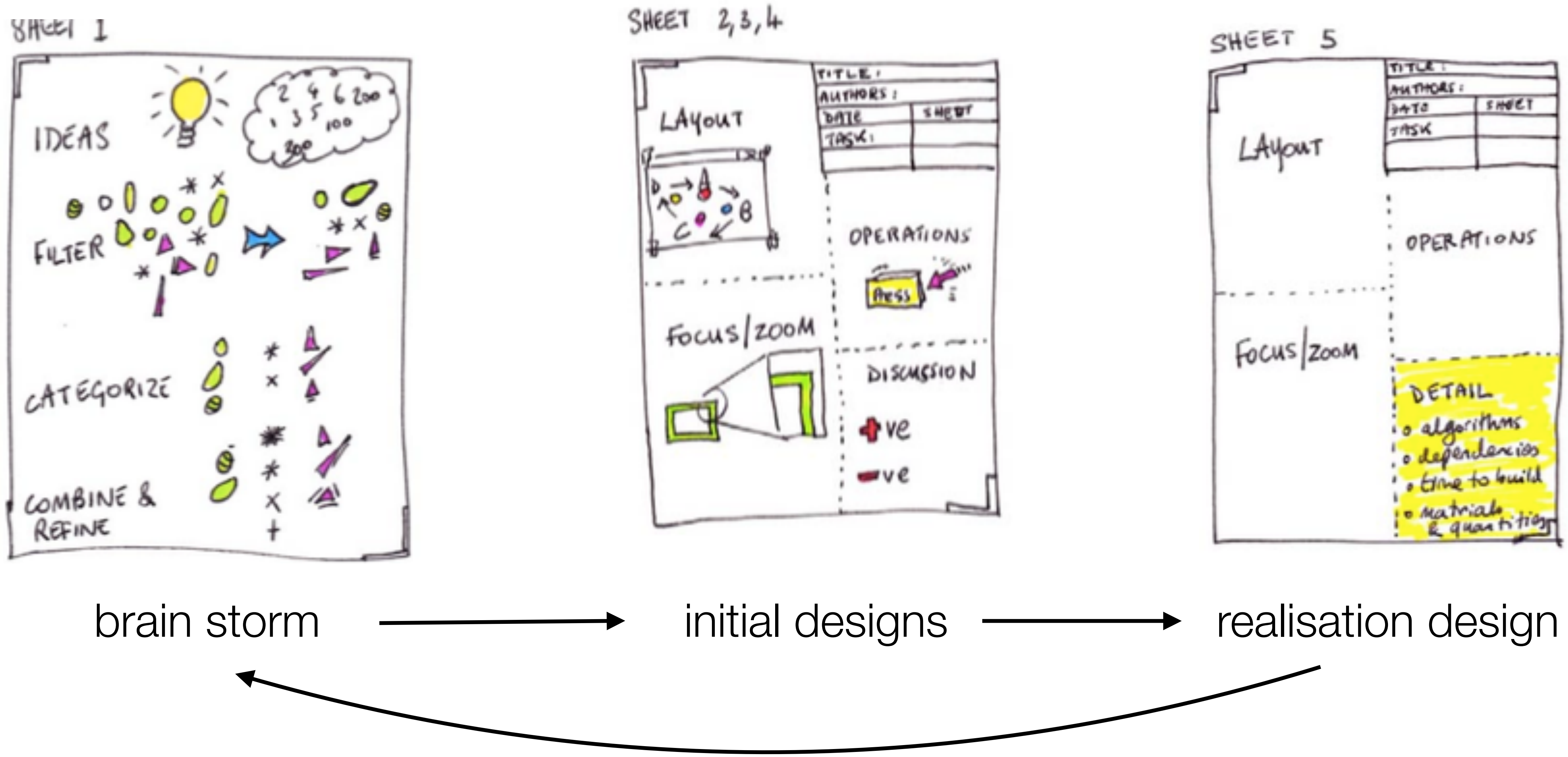
= “design fixation”

Tips to explore design space:

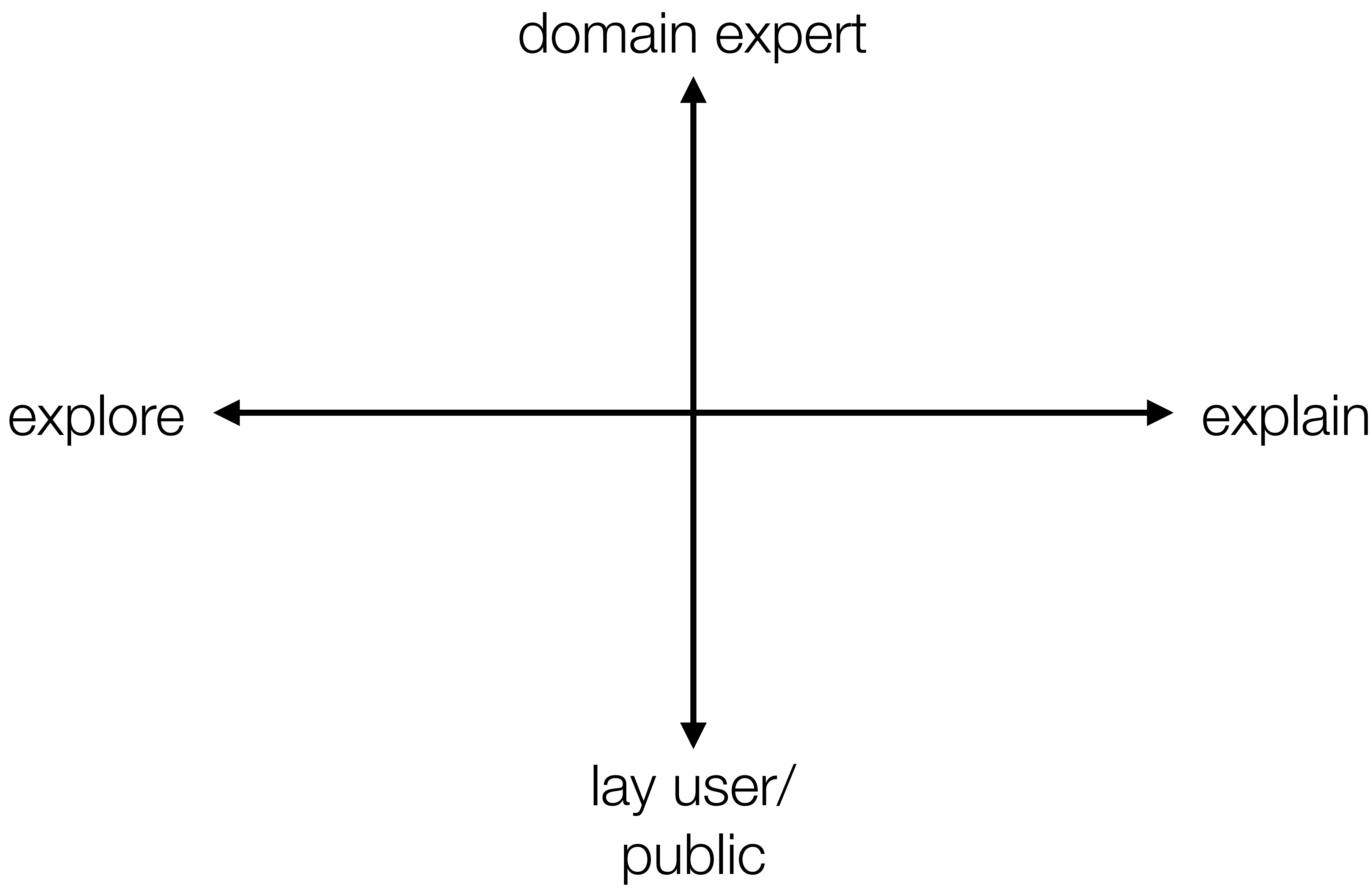
- Use systematic approach such as 5-design-sheet (see paper)
- Make very “cheap” sketches: pen & paper => don’t get attached to a design because you spent a lot of time on it
- Extend your visual library: collect interesting visuals (useful *and* useless), e.g. https://www.pinterest.com/aertsjan/data-visualizationart/
- Can you go to a higher abstraction level?



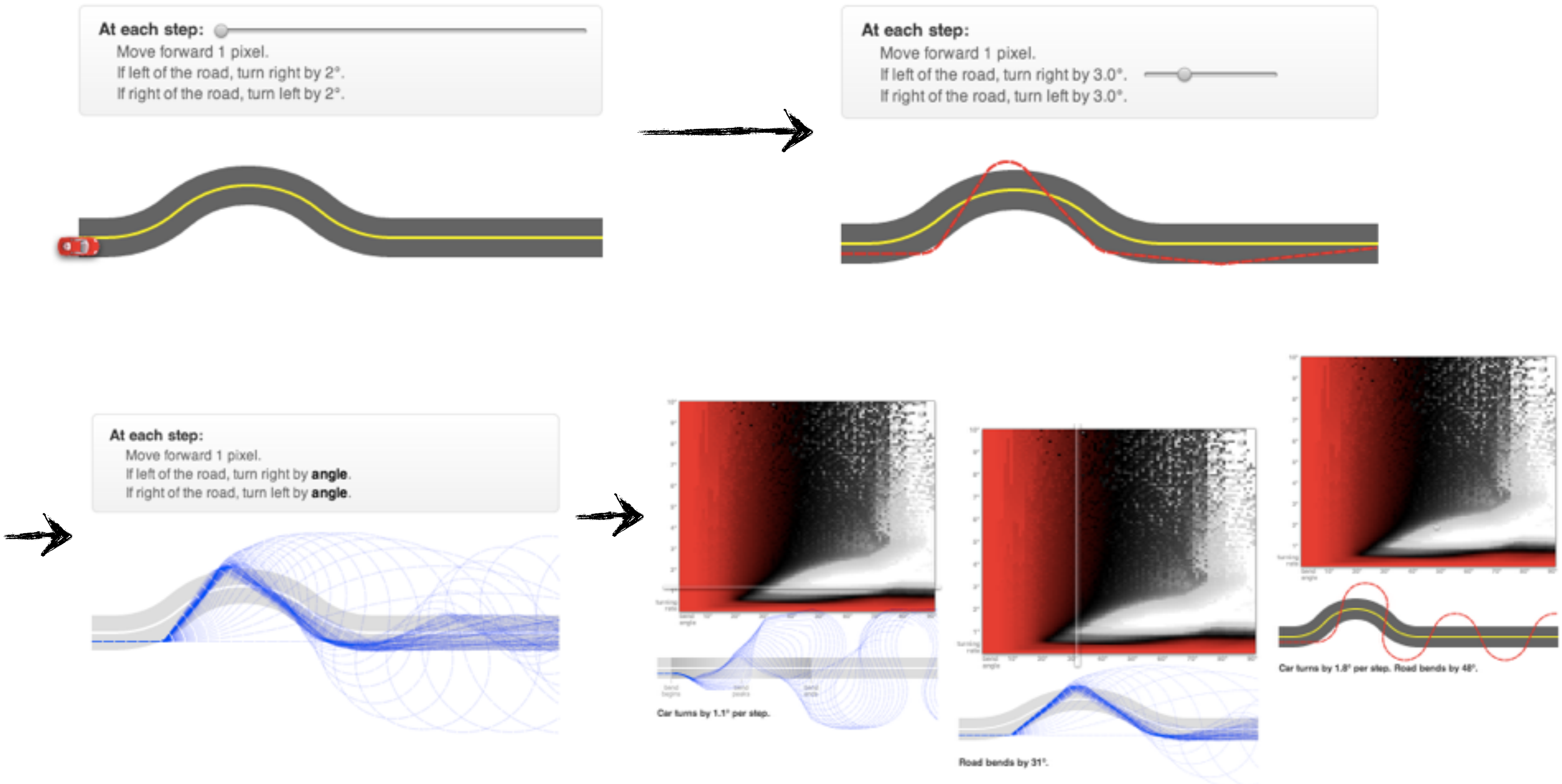
5-design sheet methodology



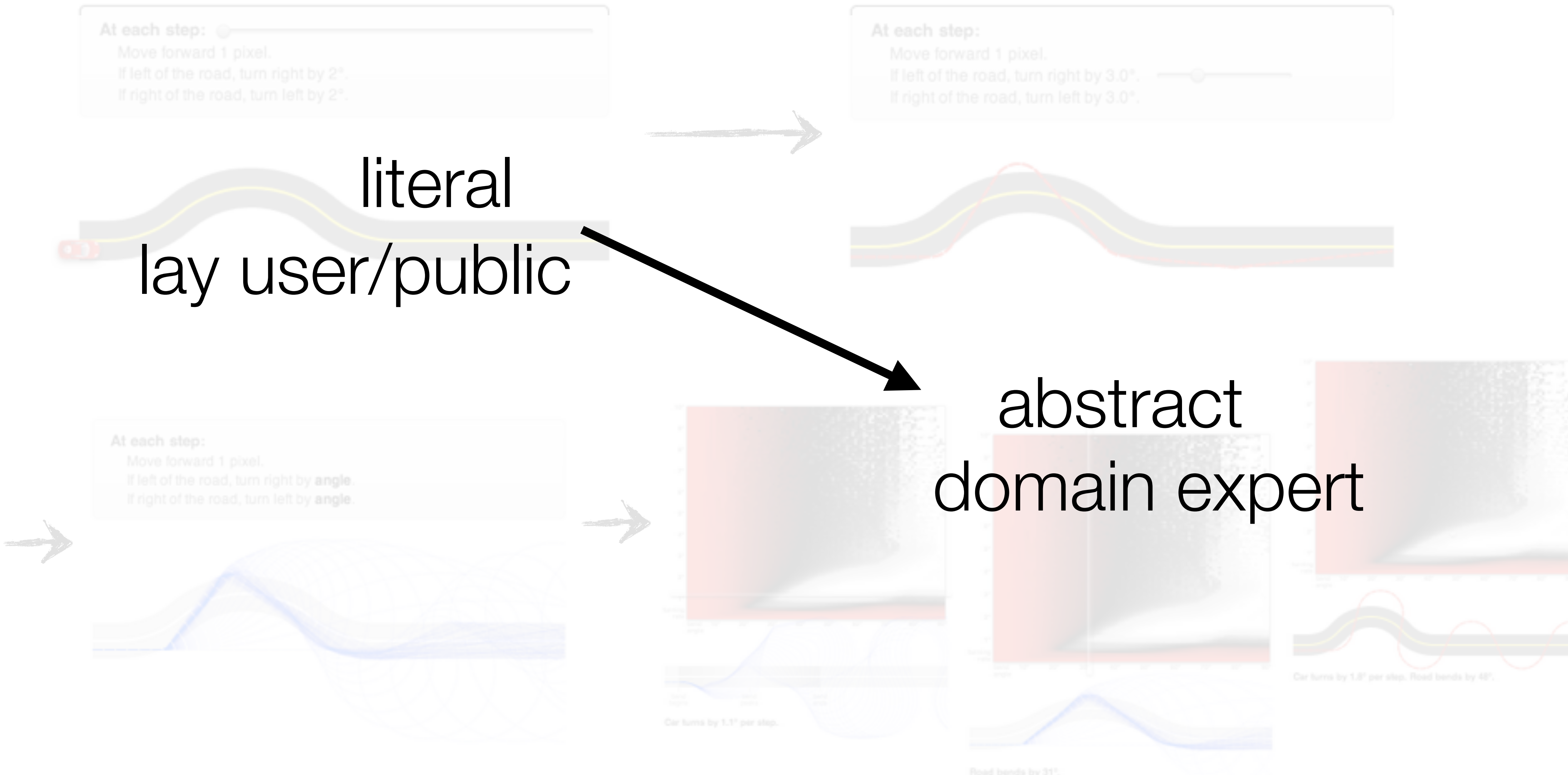
Data visualization framework



Bret Victor - Ladder of abstraction



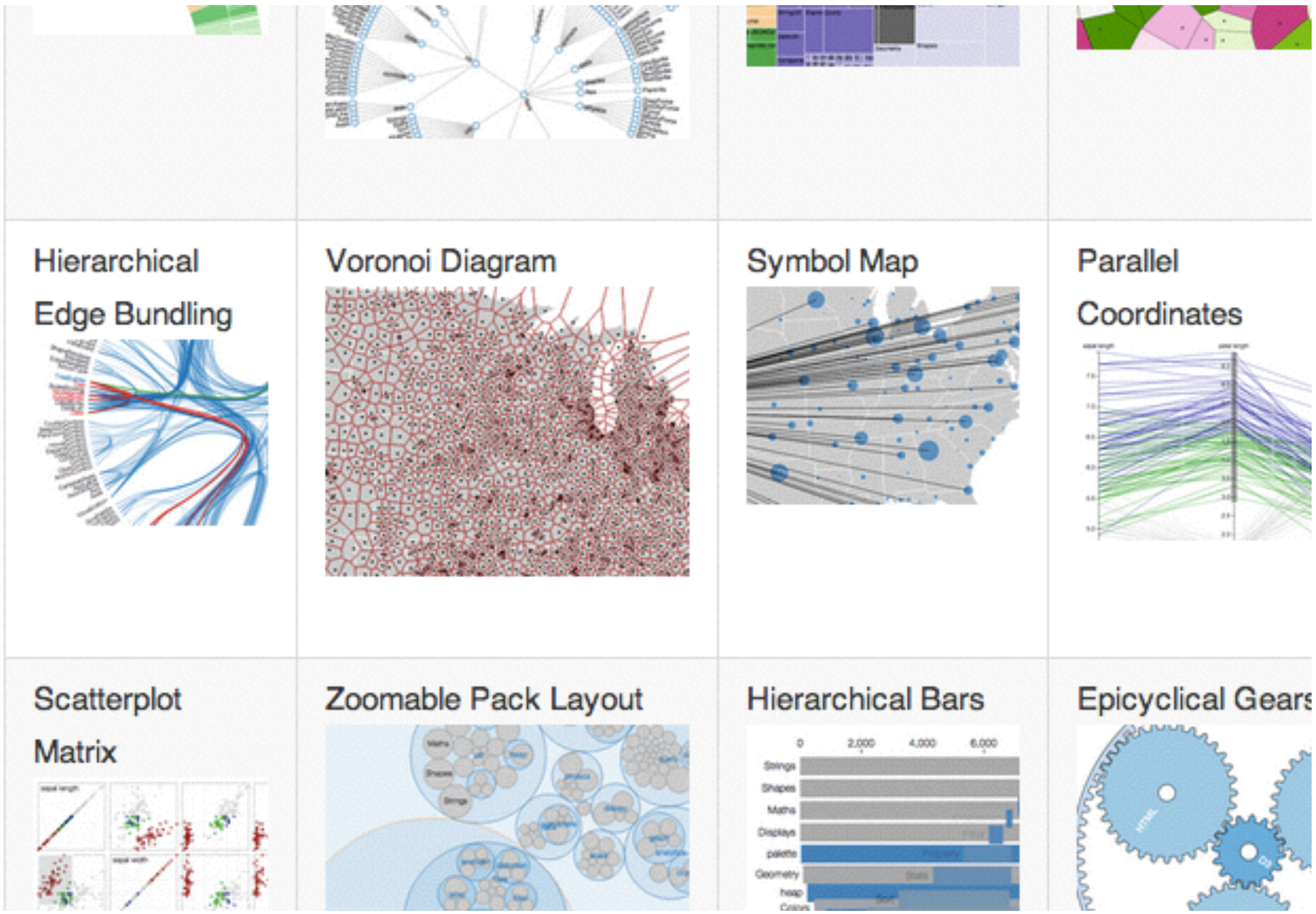
Bret Victor - Ladder of abstraction



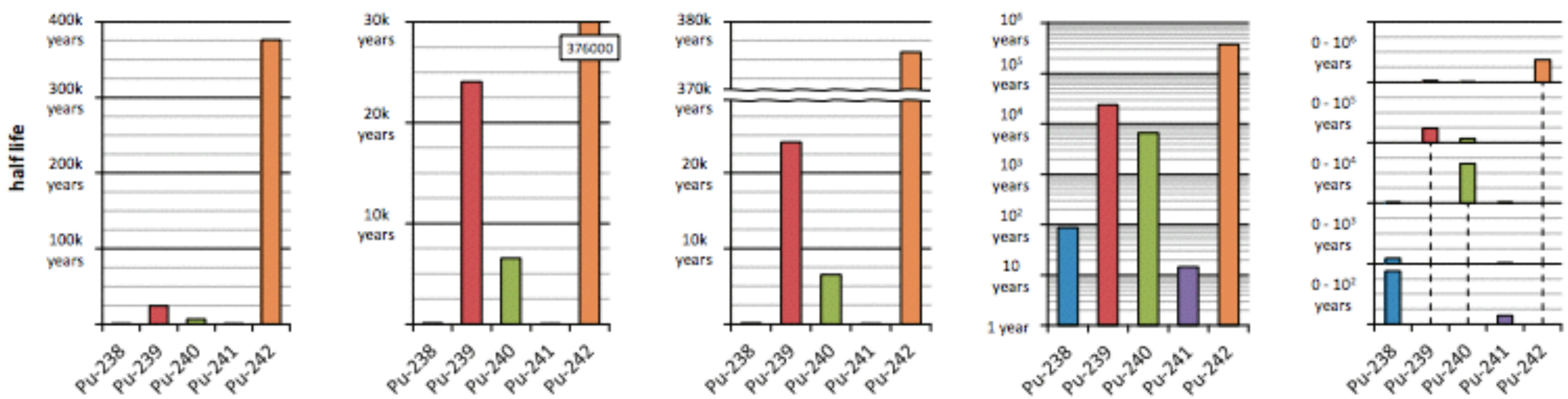
Some examples

very (!) limited scattering of visuals, only to indicate the breadth of possibilities

See D3.js examples (https://github.com/mbostock/d3/wiki/Gallery)



Scale-stack bar chart



(a) (b) (c) (d) (e)

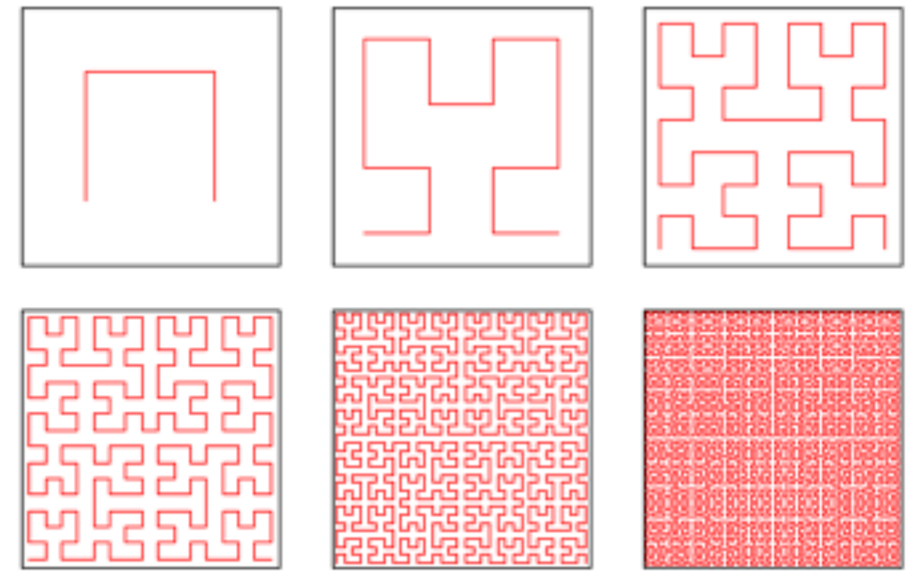
Figure 1: *Different bar charts for visualizing data with large value range. Half lives of different plutonium isotopes (see Sect. 6.1) are visualized with: (a) classic linear bar chart, (b) linear bar chart with cut-off bars, (c) linear bar chart with scale break, (d) logarithmic bar chart, and (e) our novel scale-stack bar chart.*

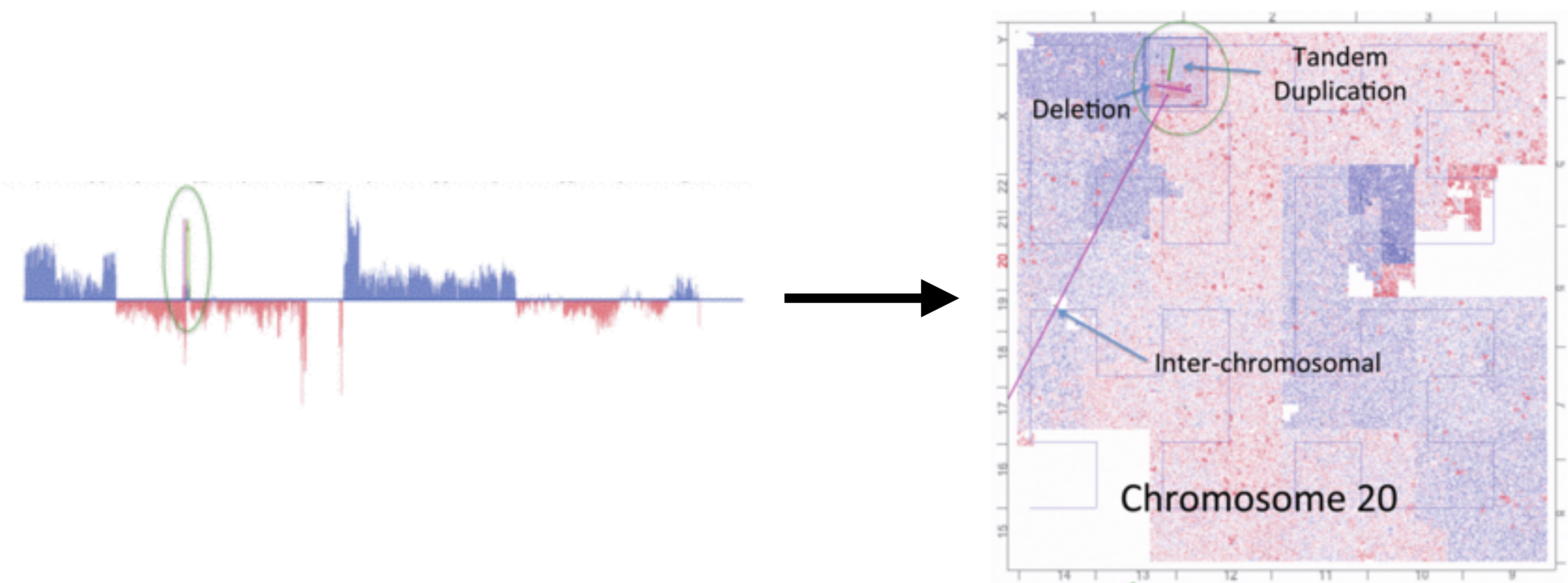
http://www.vis.uni-stuttgart.de/~sadlo/download/hlawatsch2013scaleStack.pdf

1D, 2D

space-filling curves: 1D -> 2D

e.g. Hilbert curve

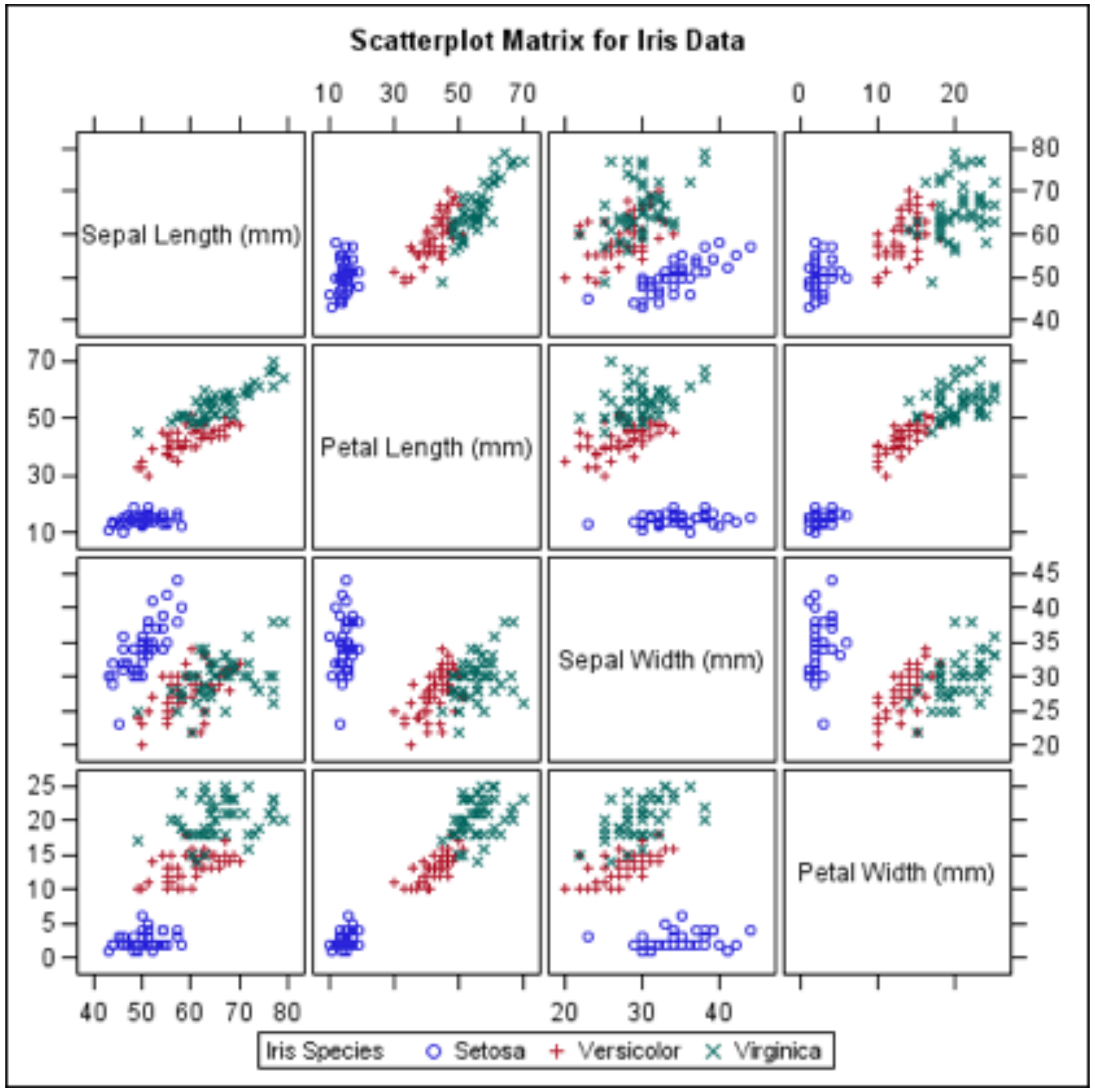




Iris setosa

Iris versicolor

Iris virginica



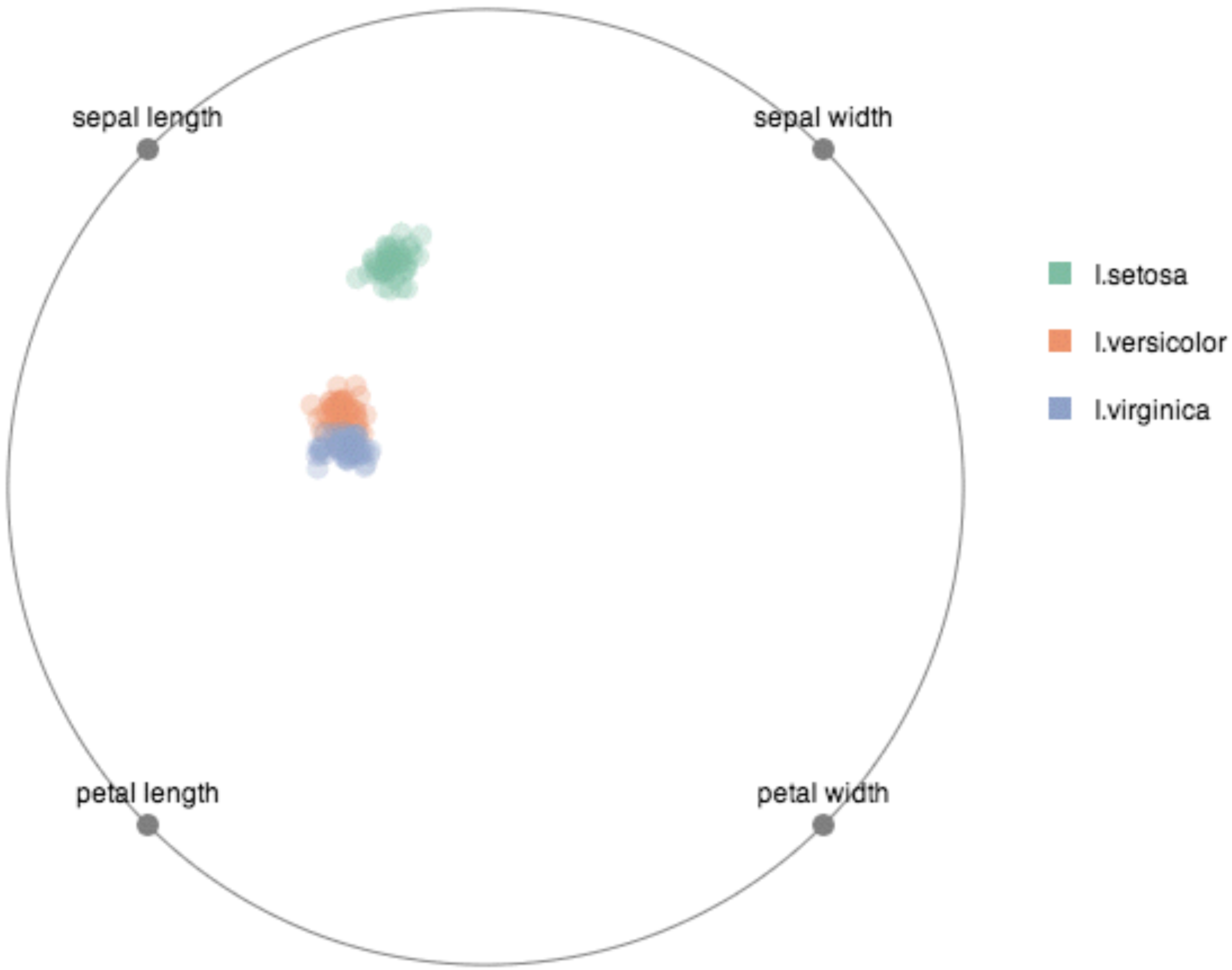
scatterplot matrix

(source: support.sas.com)

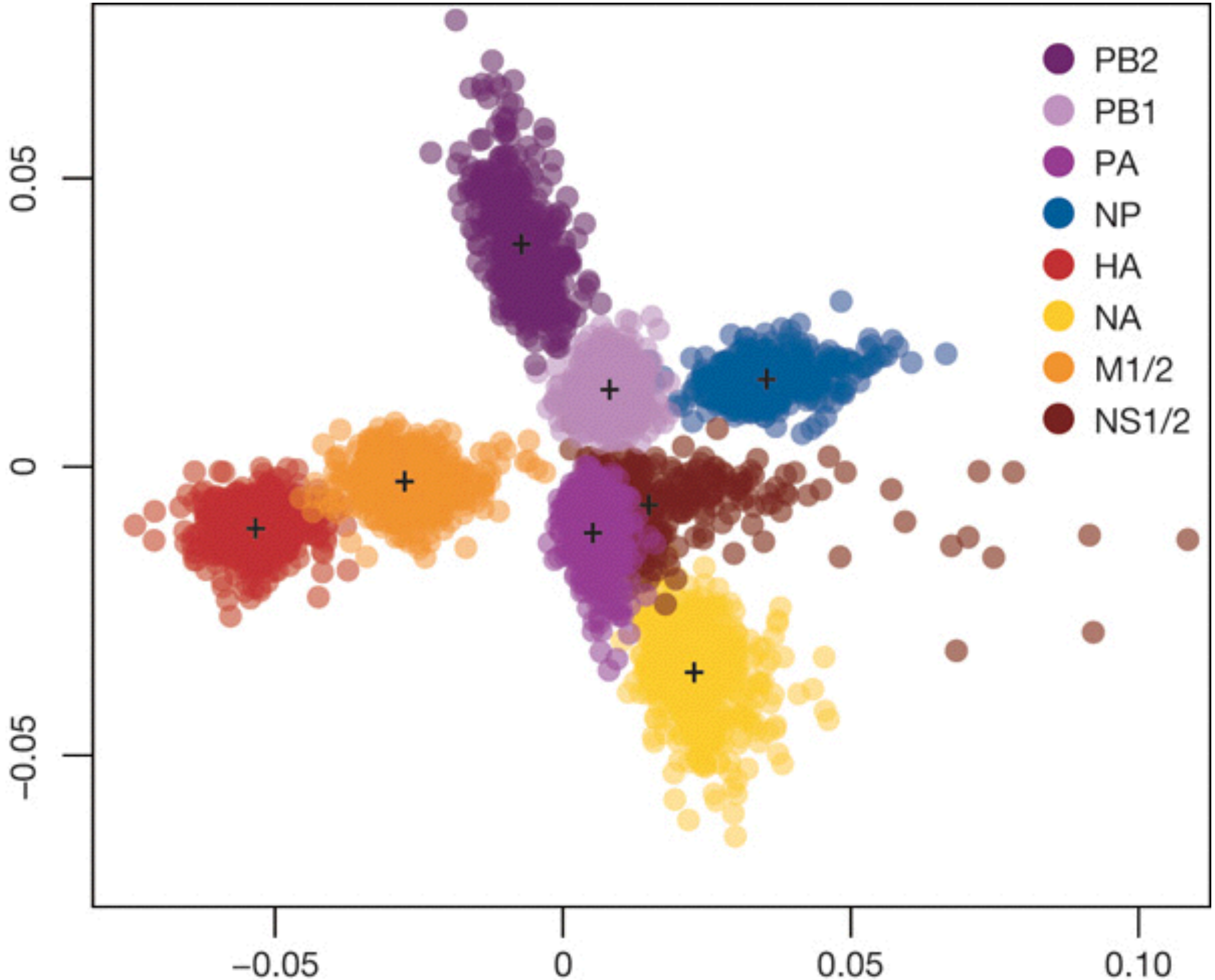
Iris setosa

Iris versicolor

Iris virginica

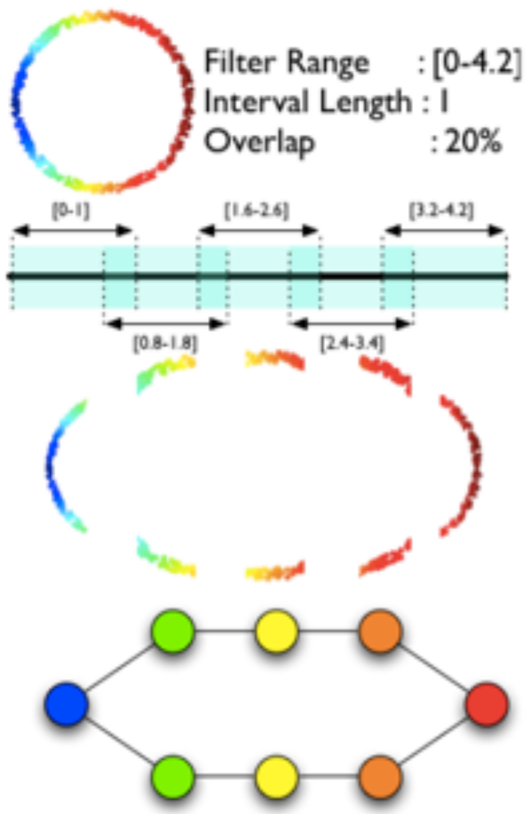


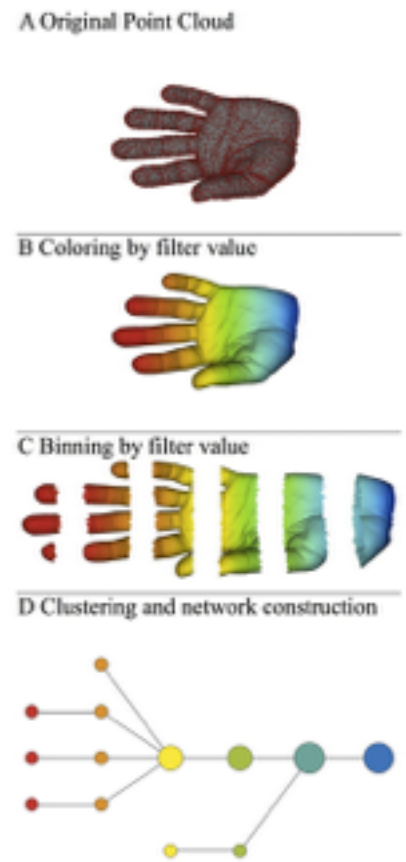
radviz



multi-dimensional scaling plot of distances between samples trees

(source:doi:10.1038/nature06945)

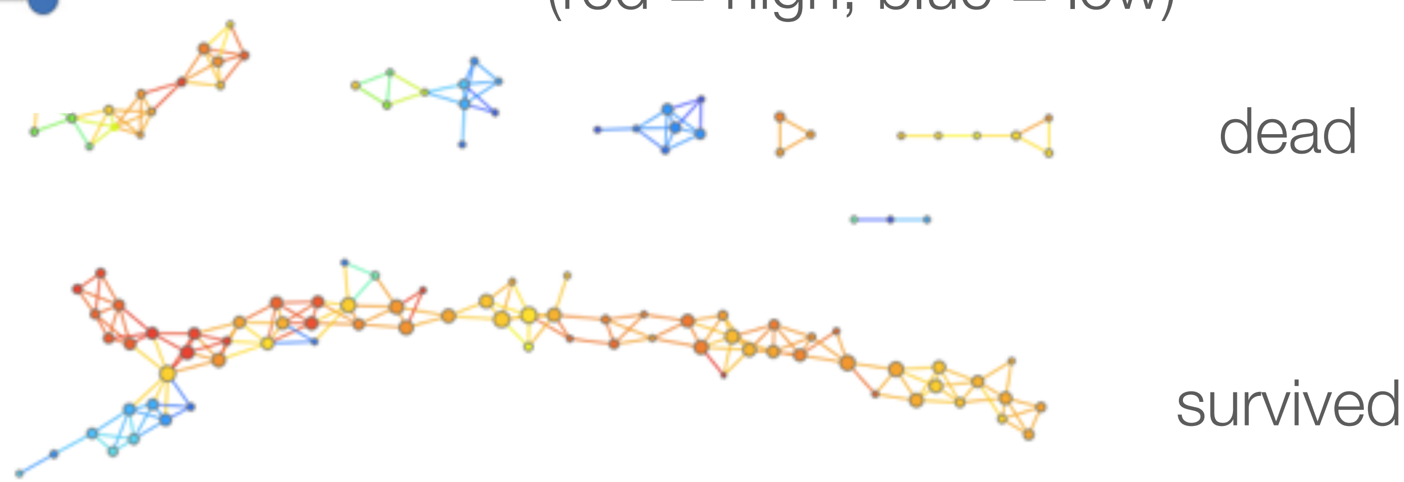




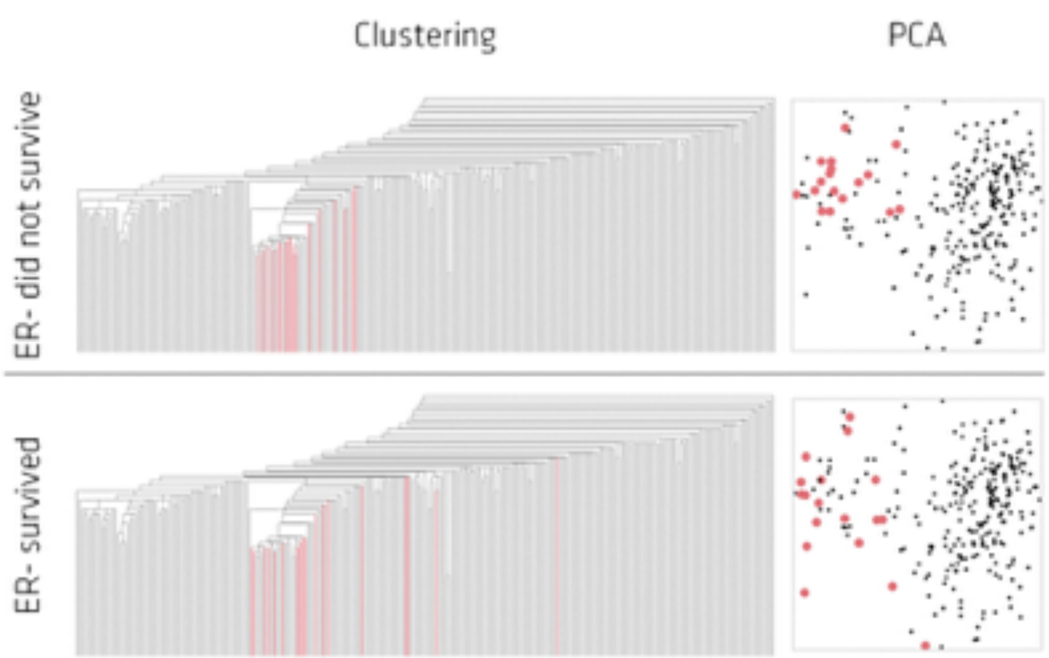
topological data analysis

local clustering in global context

colour = ESR1 expression (red = high; blue = low)



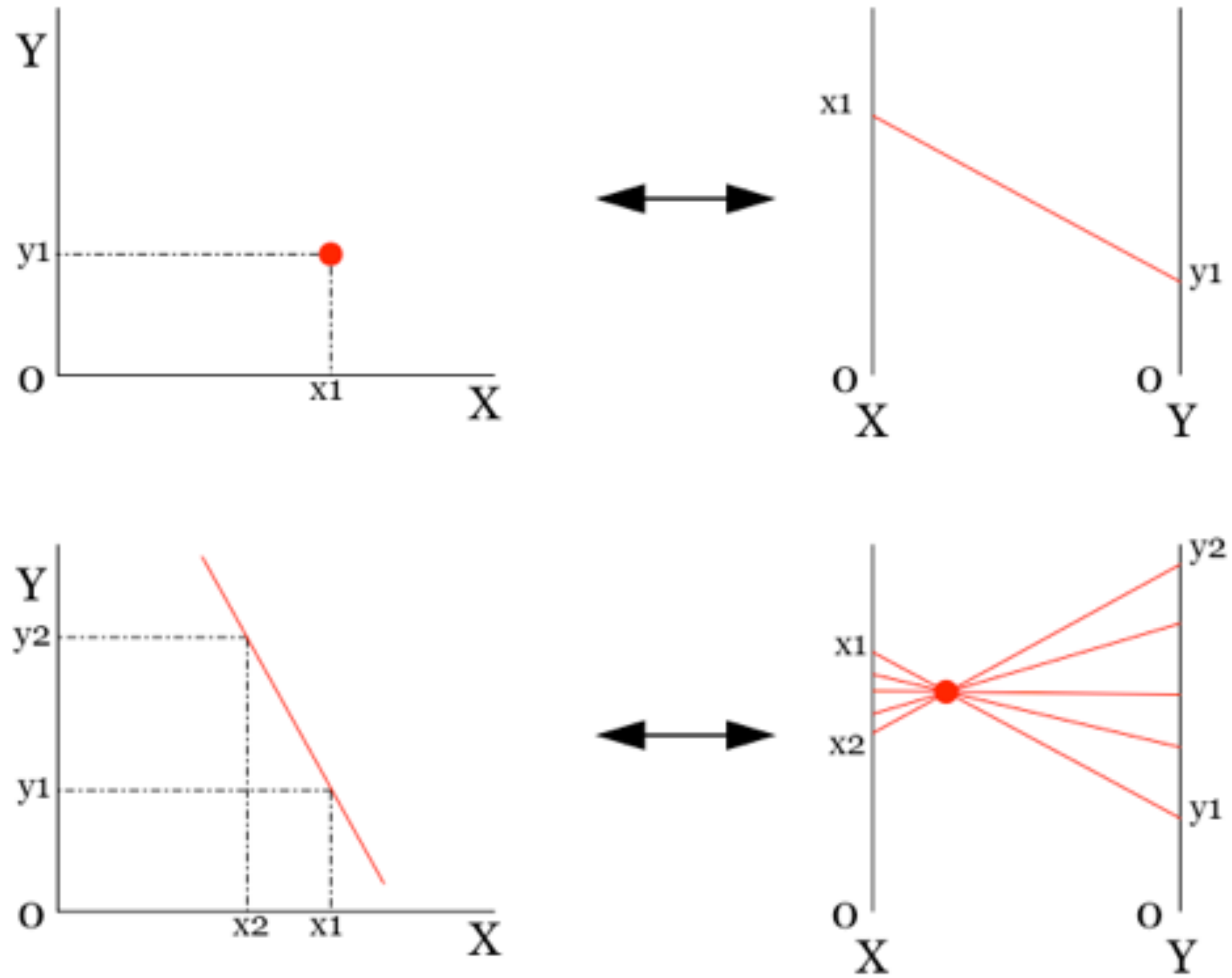
Could they have found this with just clustering or PCA? No.



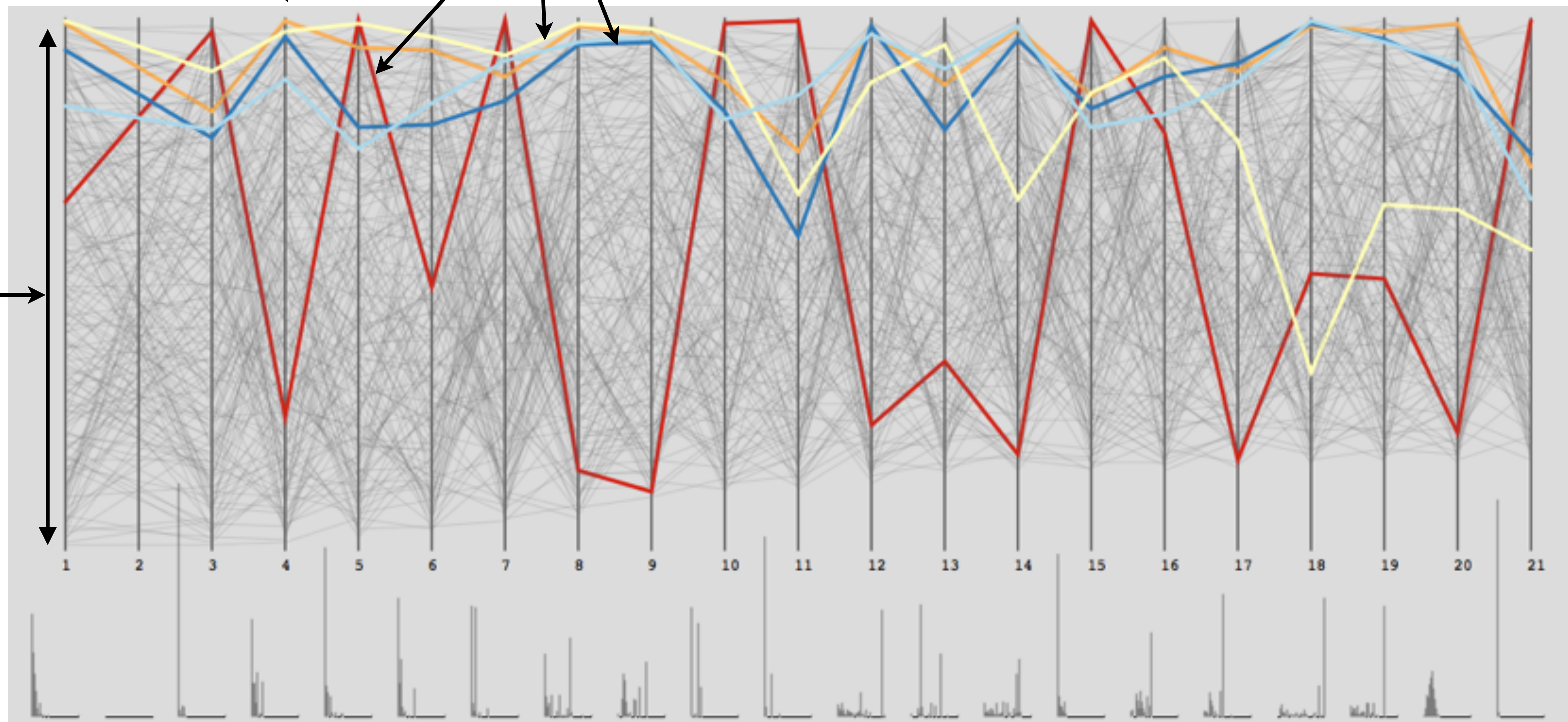
lens = L-infinity centrality + event death

Ayasdi, Inc

- parallel coordinates: point in Euclidean space = line in parallel coordinate space and vice versa

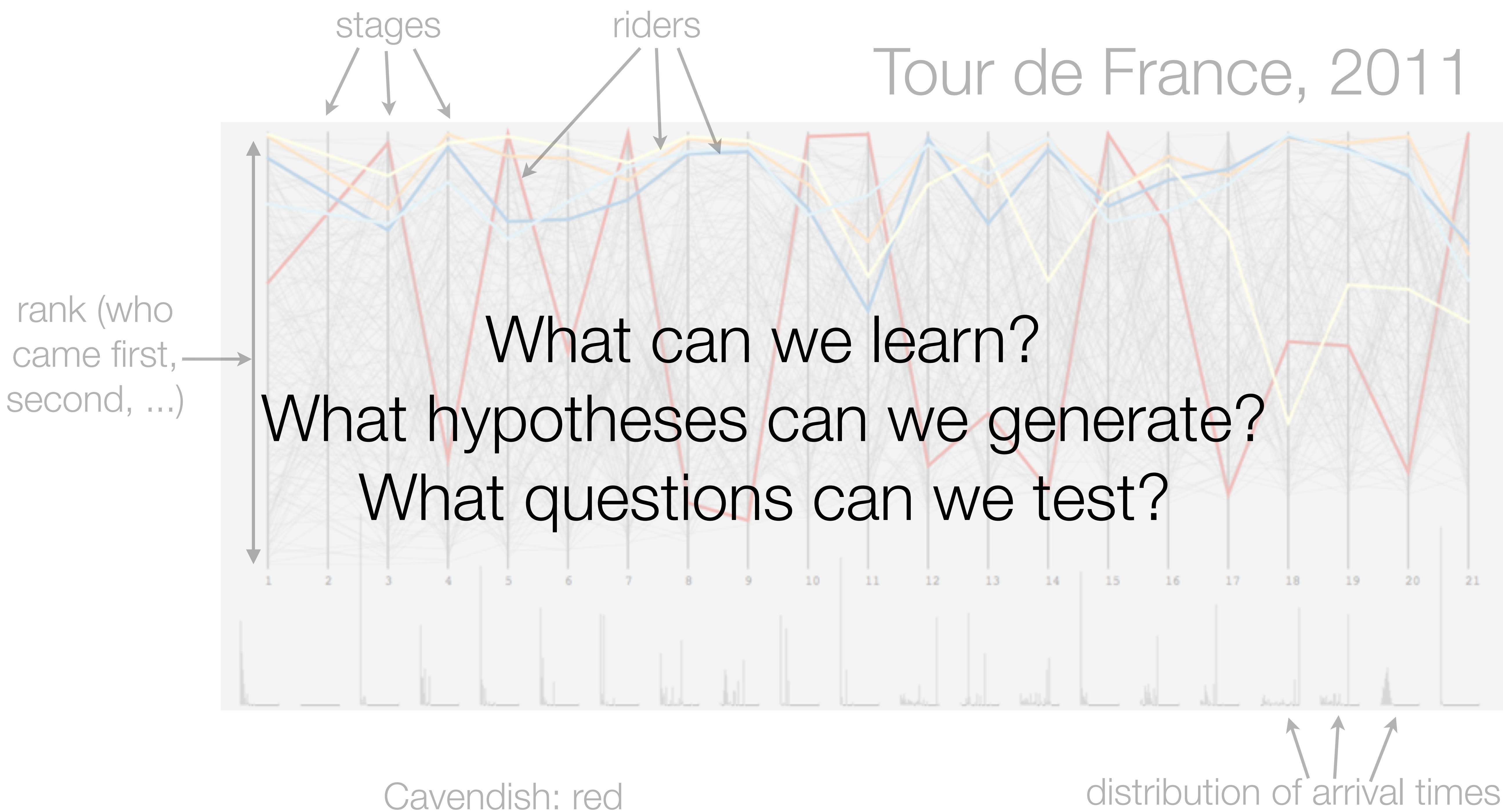


Tour de France, 2011



Cavendish: red
Evans: orange
Gilbert: yellow
Andy Schleck: light blue
Frank Schleck: dark blue

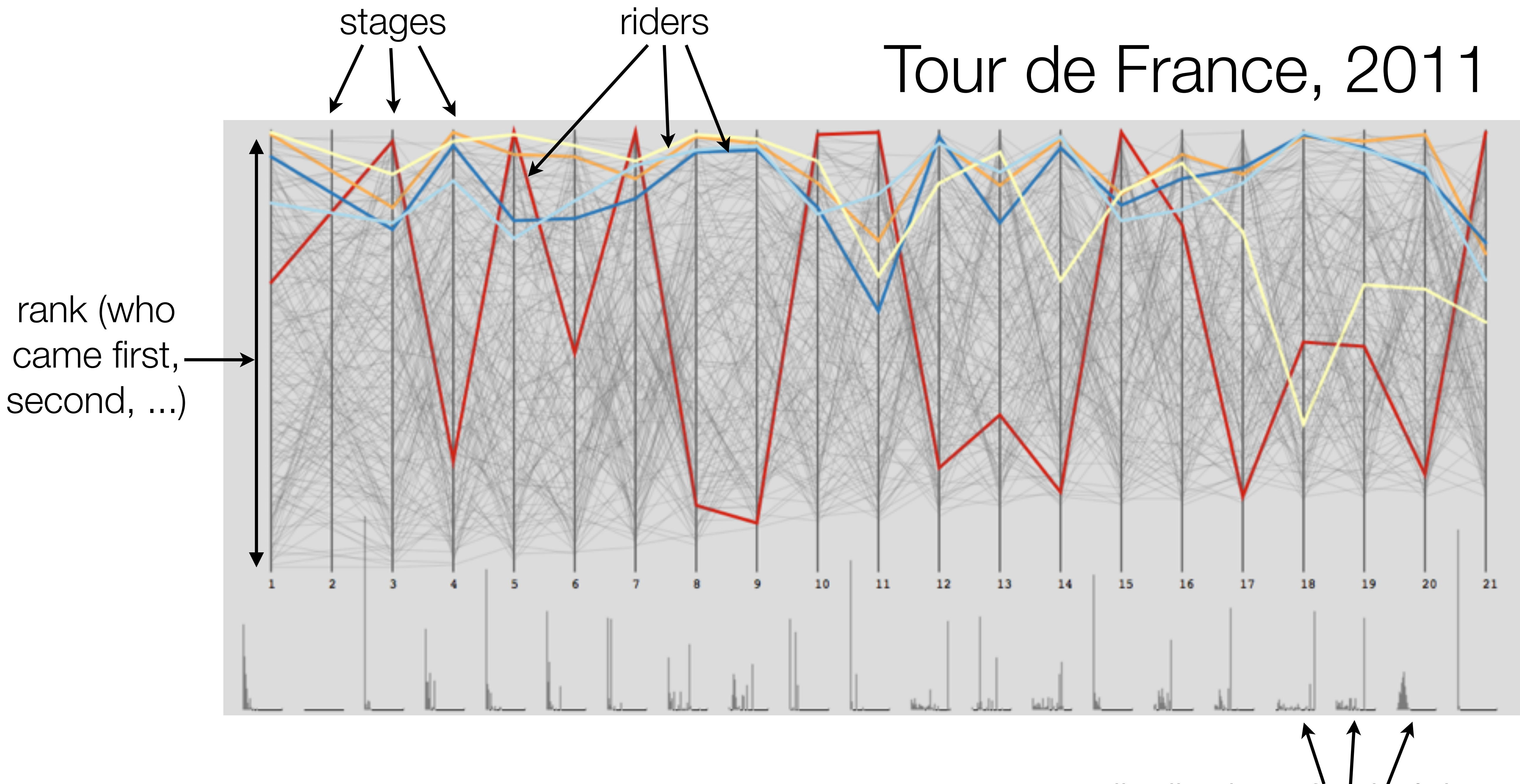
Tour de France, 2011



What can we learn?
What hypotheses can we generate?
What questions can we test?

Cavendish: red
Evans: orange
Gilbert: yellow
Andy Schleck: light blue
Frank Schleck: dark blue

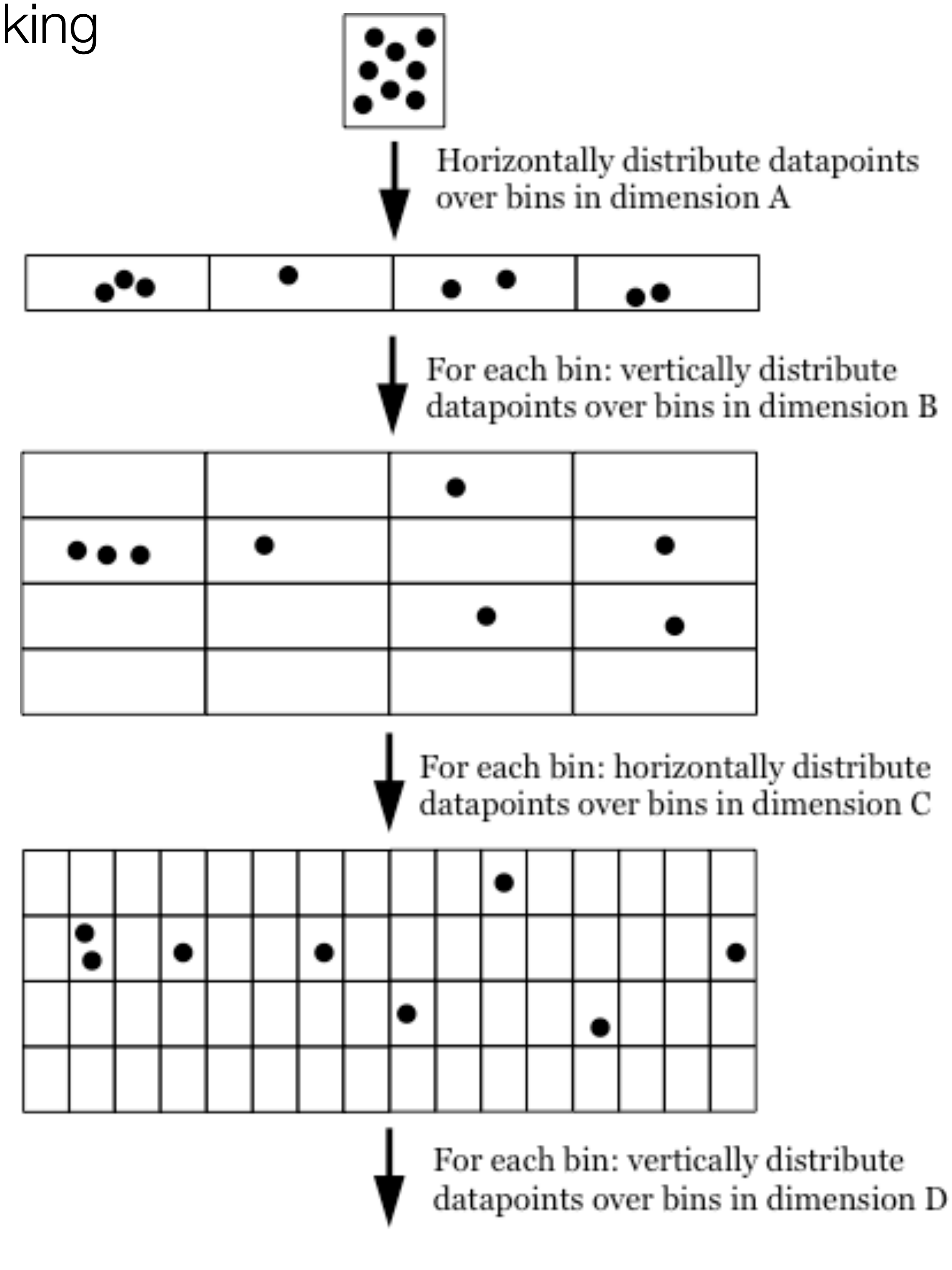
Tour de France, 2011

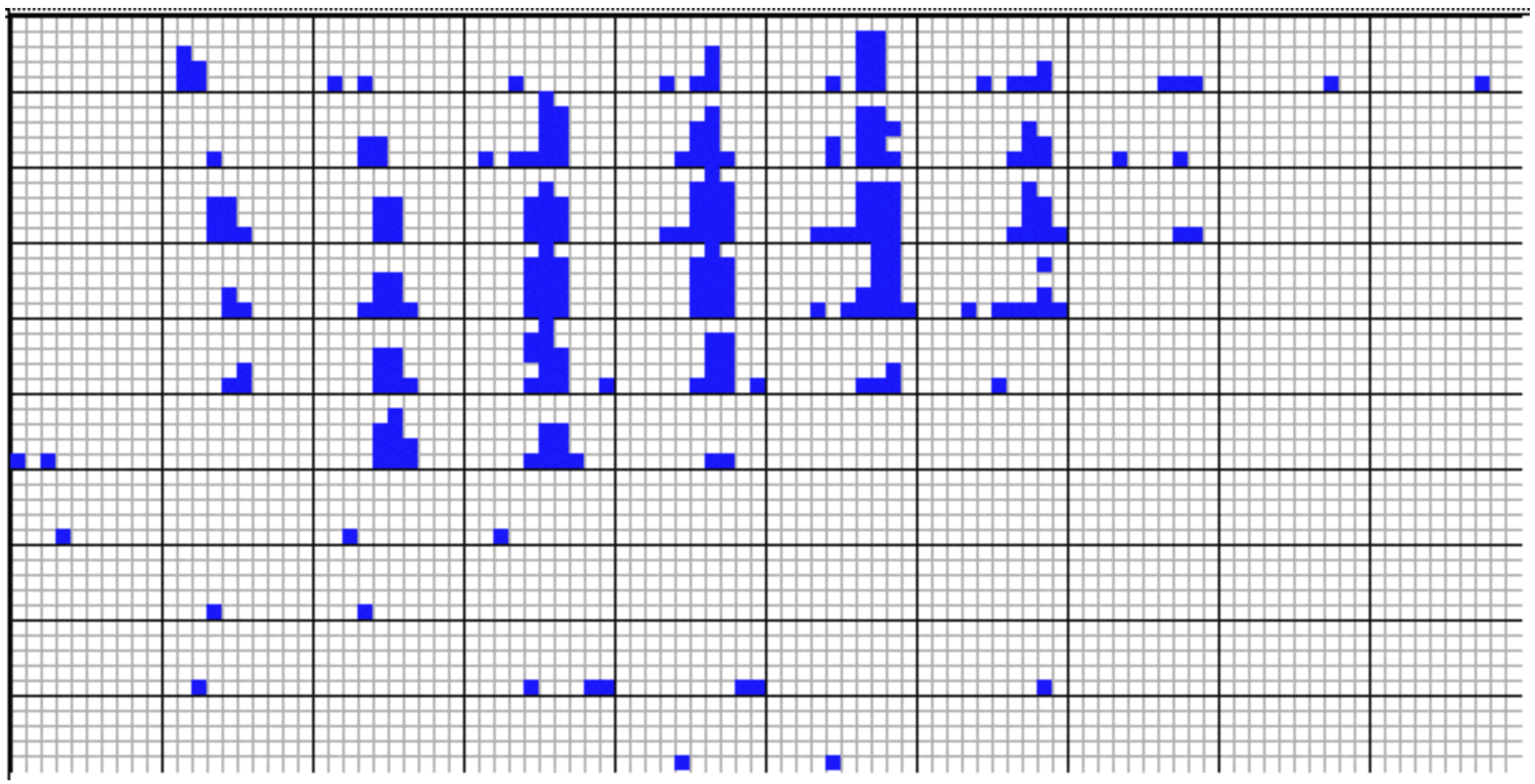


Cavendish: red
Evans: orange
Gilbert: yellow
Andy Schleck: light blue
Frank Schleck: dark blue

- Which trips are in mountains?
- When do riders start giving up and leaving the Tour?
- Is Cavendish an expert for the mountains or for flats?
- How does Philippe Gilbert perform across the Tour?
- What's with the group of riders performing bad in stages 18 and 20, but good in stage 19?

dimensional stacking



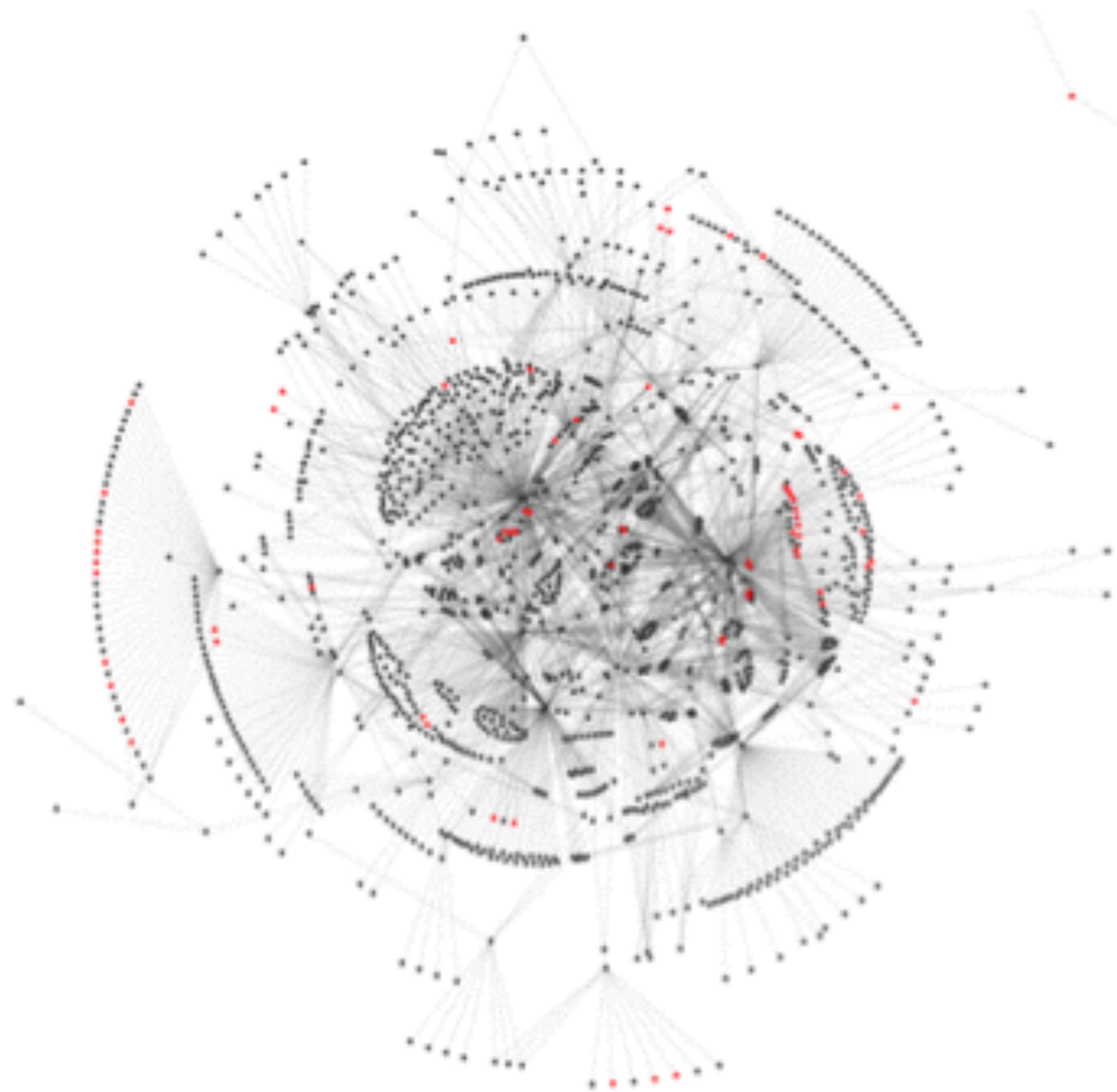


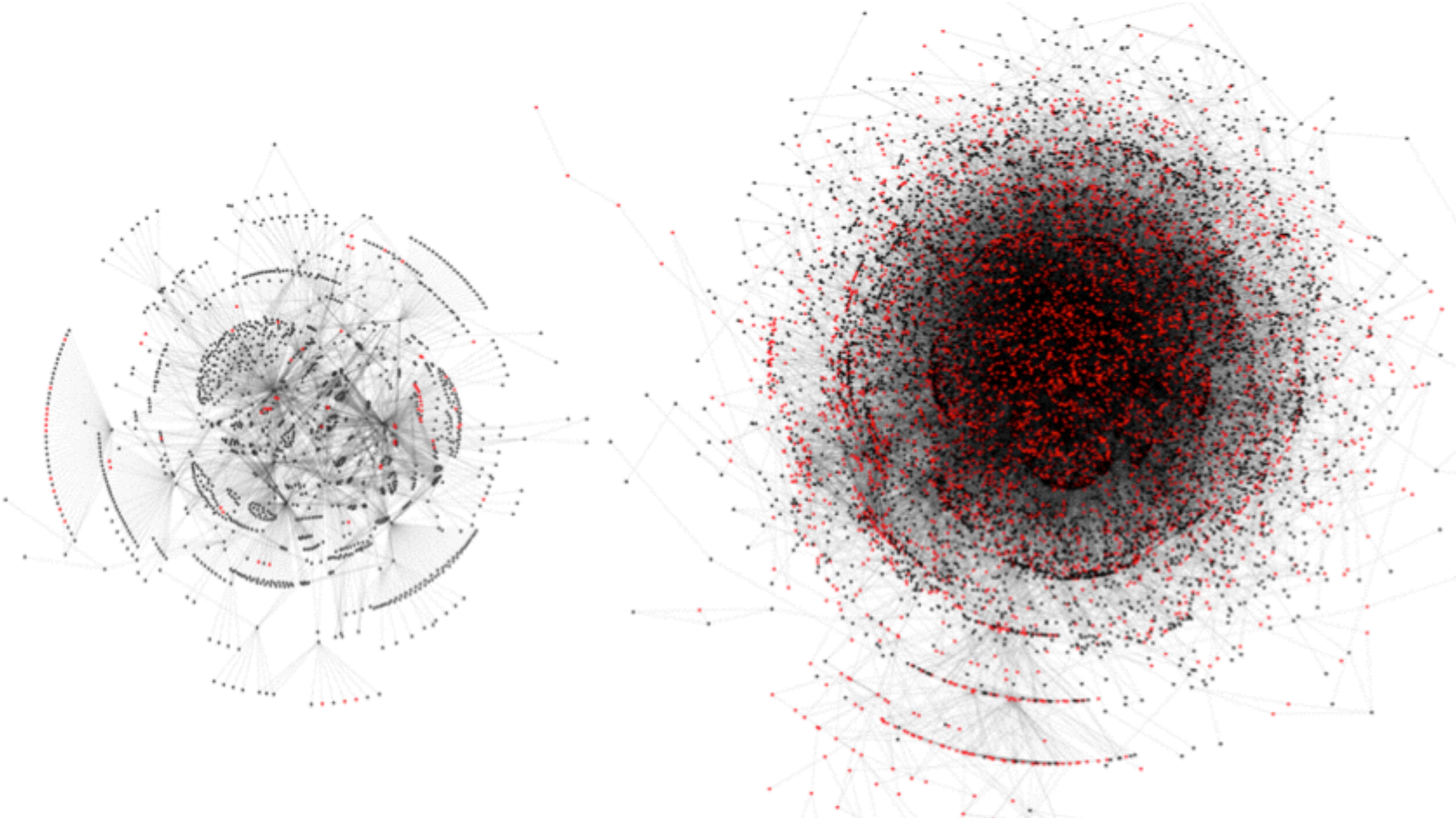
4D drill-hole data: 3 spatial dimensions + ore grade as the fourth dimension

(source: http://web.cs.wpi.edu/)

networks: node-link diagrams

max 20 nodes…





Martin Krzywinski

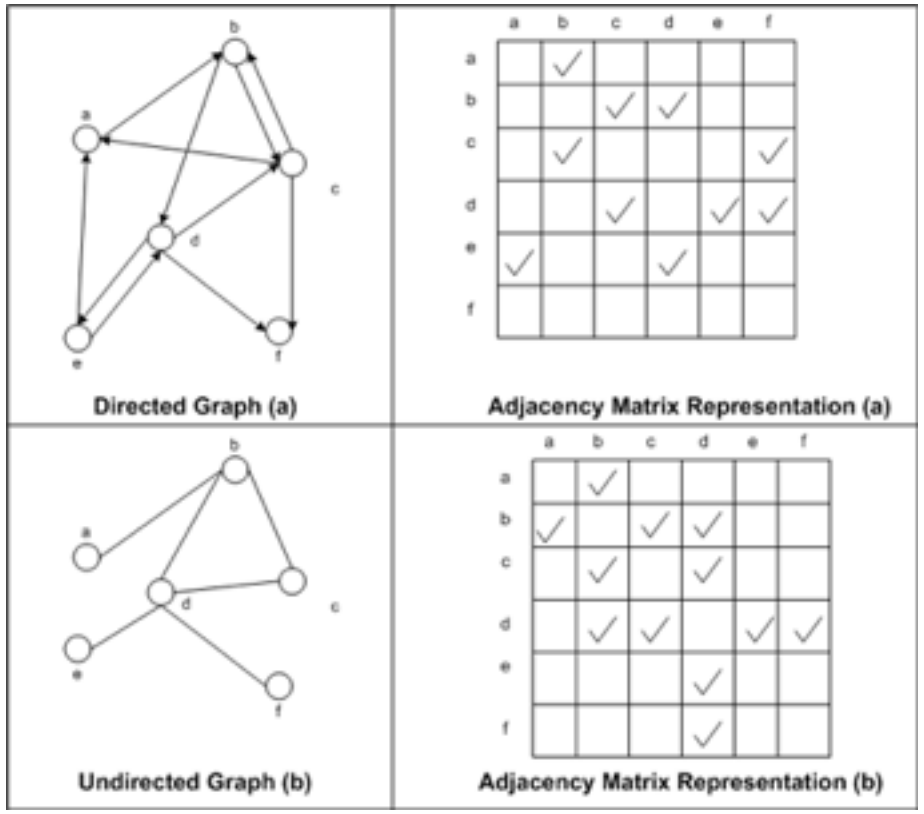
networks

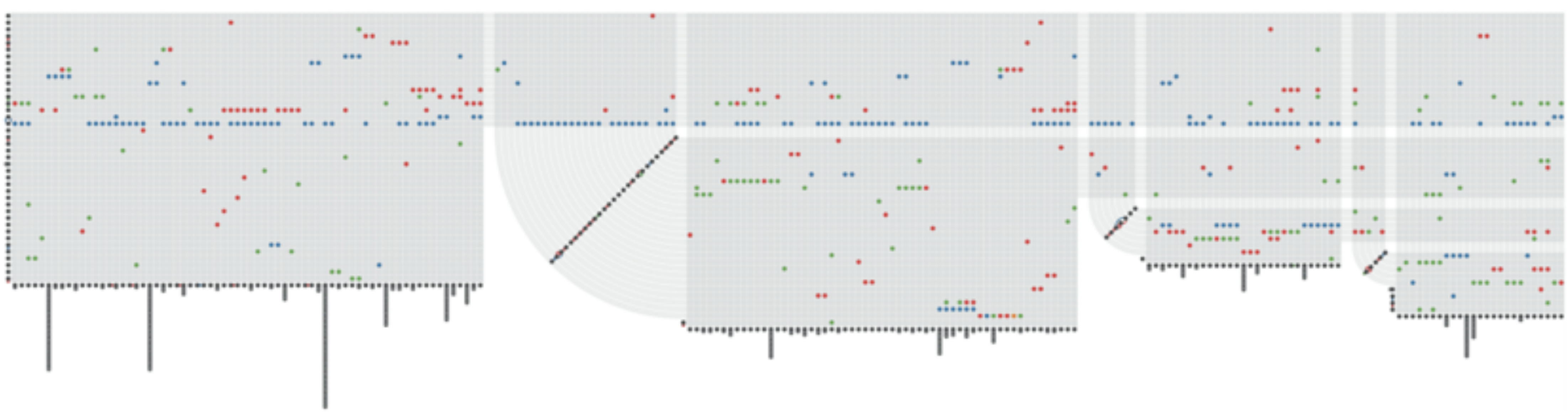
same network

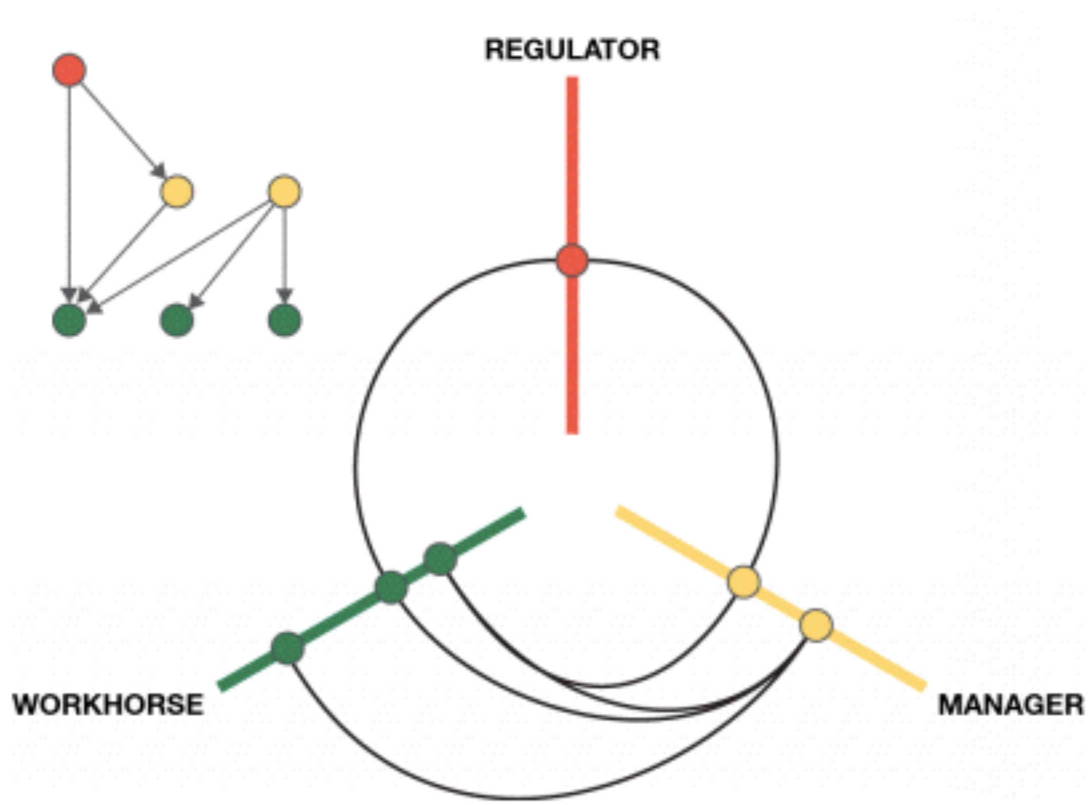
Martin Krzewinsky

- adjacency matrix

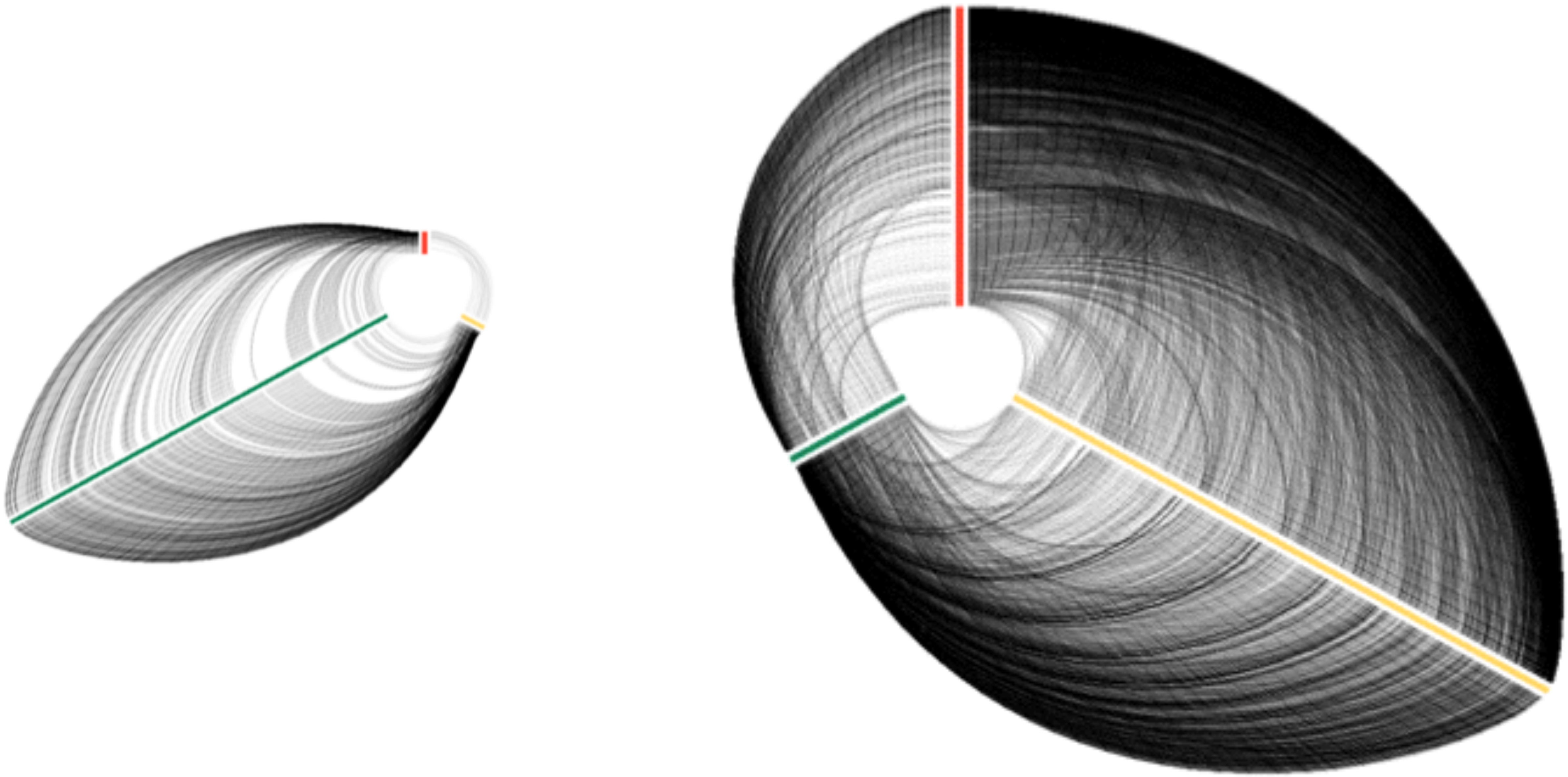
- compressed adjacency matrix (Dinkla *et al*, 2012)





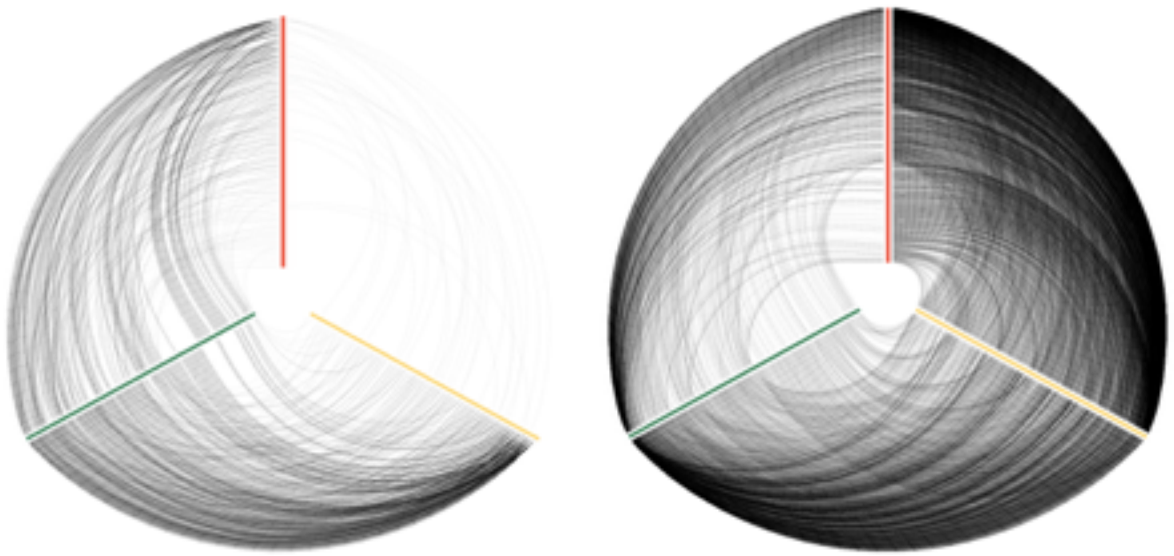

REGULATOR
WORKHORSE
MANAGER

- hive plots

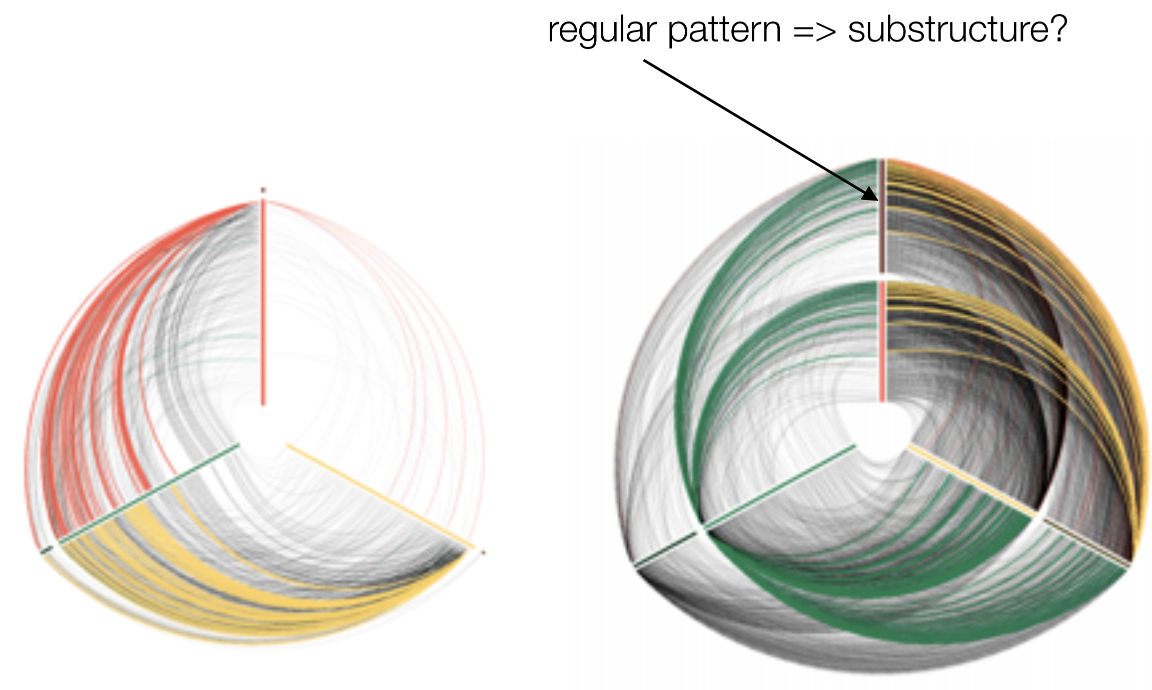


Martin Krzywinski

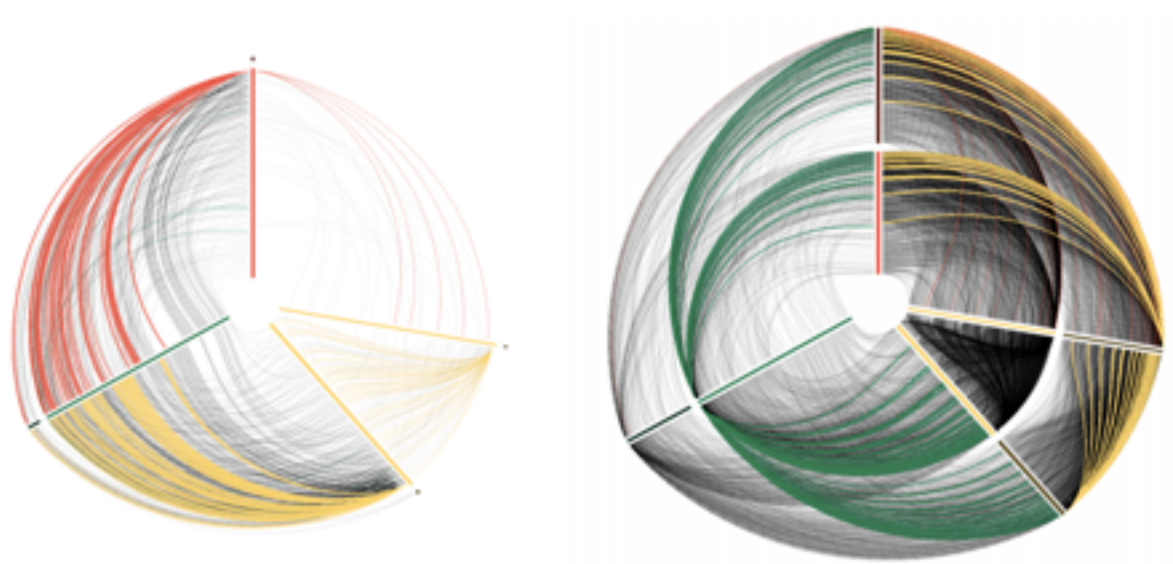
networks



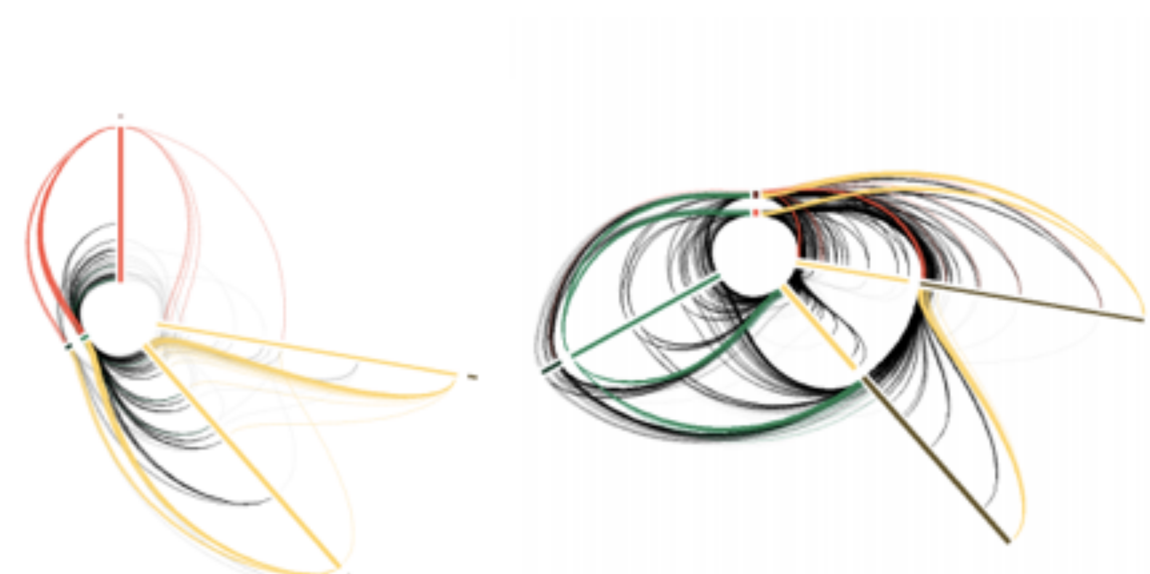
scaled to 100%



segmentation: non-persistent vs persistent

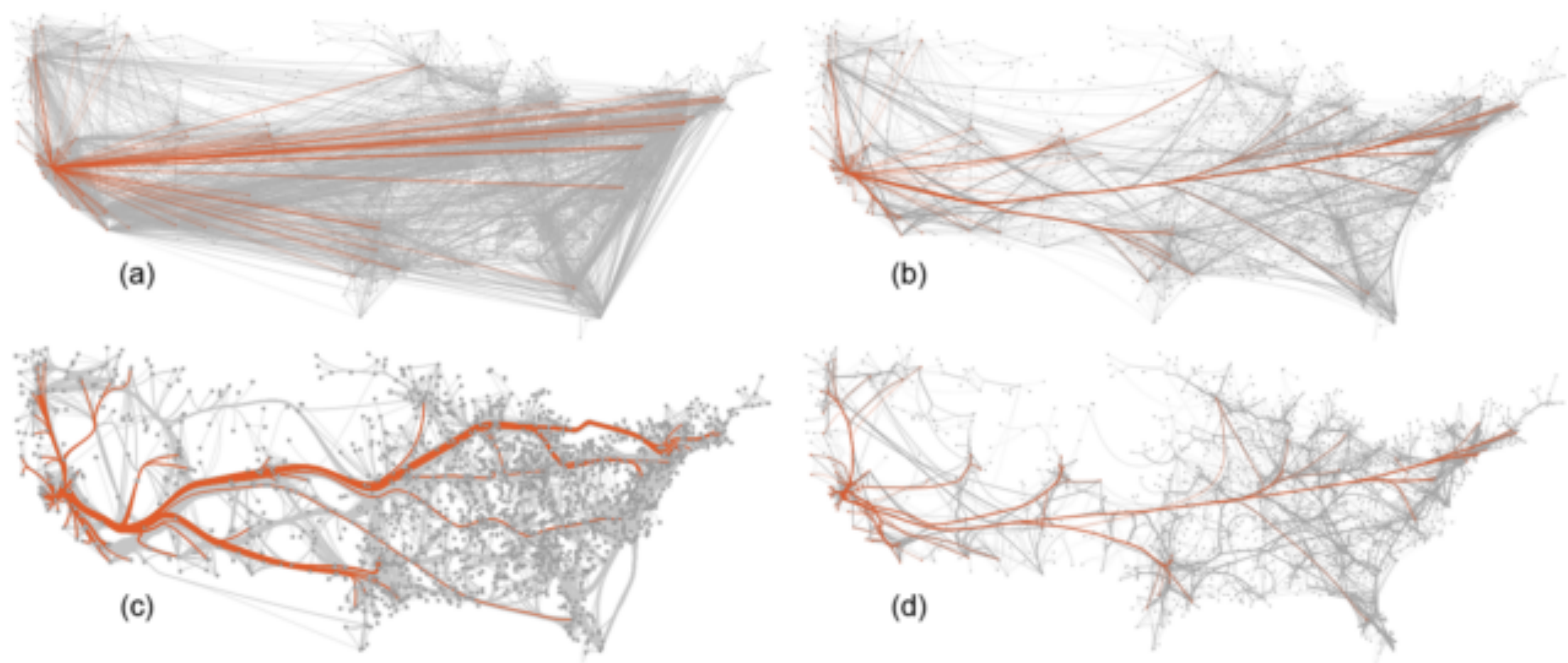


intra-axis connections



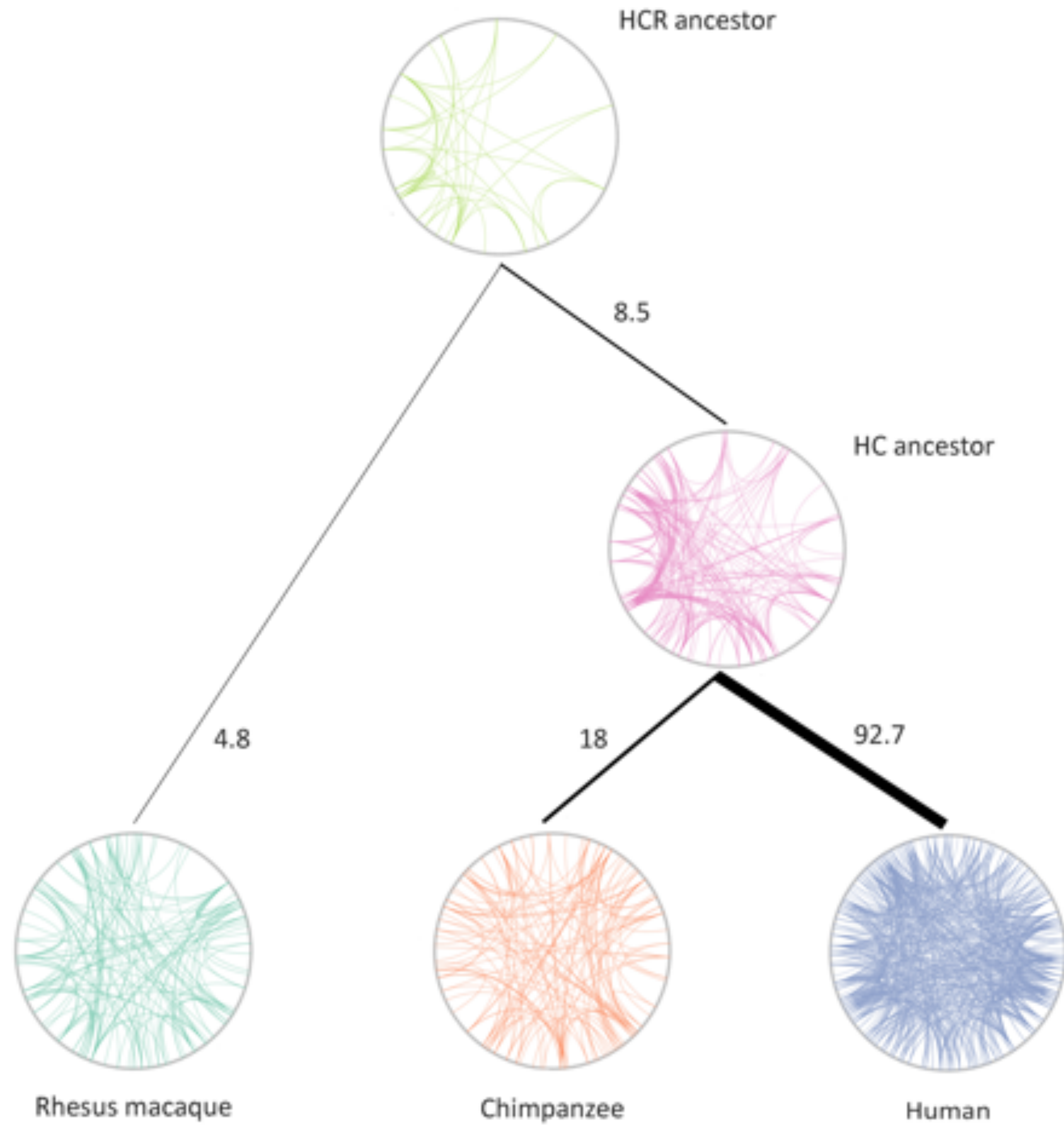
absolute connectivity

edge bundling

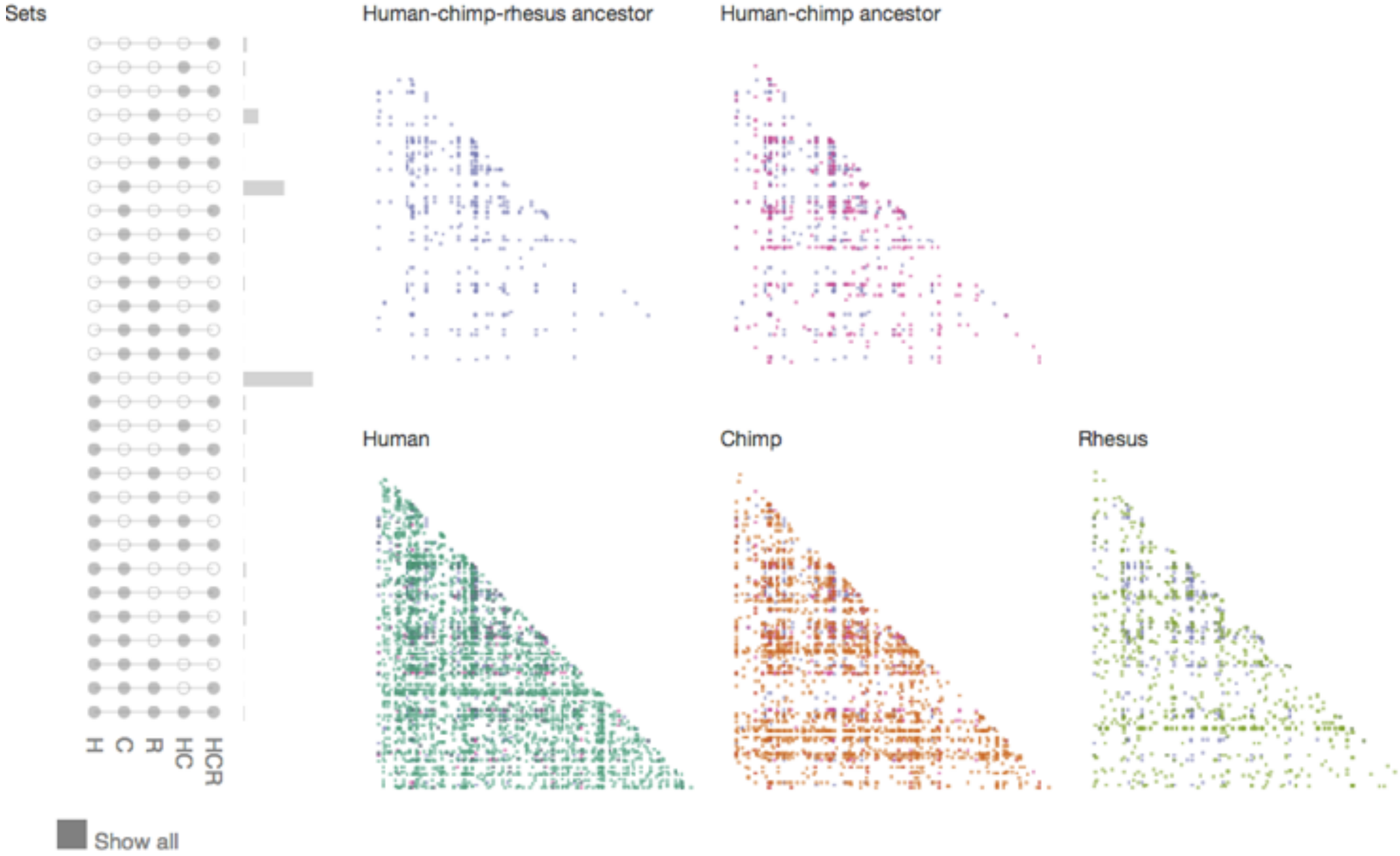


Holten & van Wijk, 2009

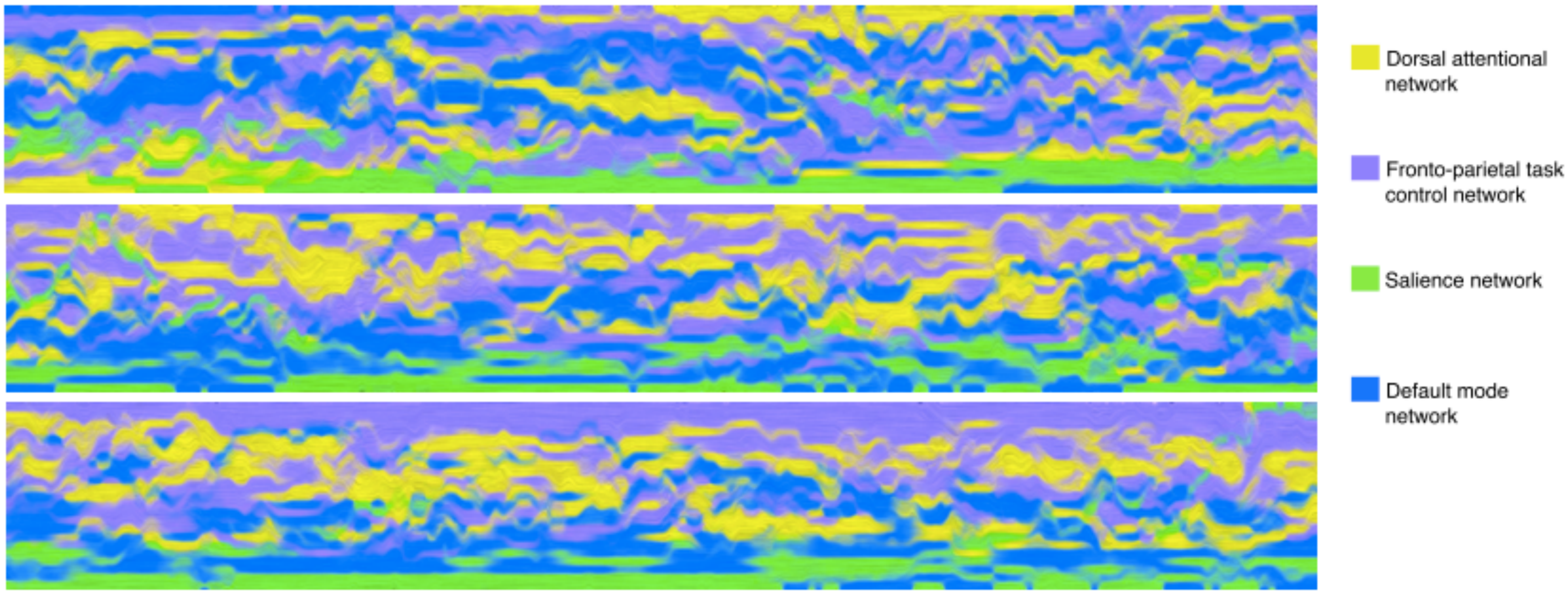
gene interaction networks in different species



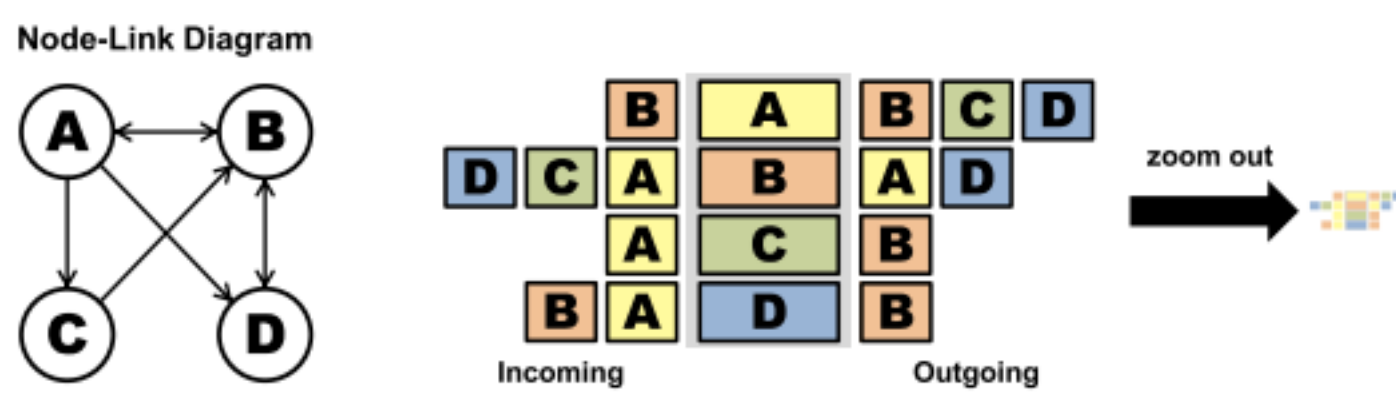
Network comparison gene interaction



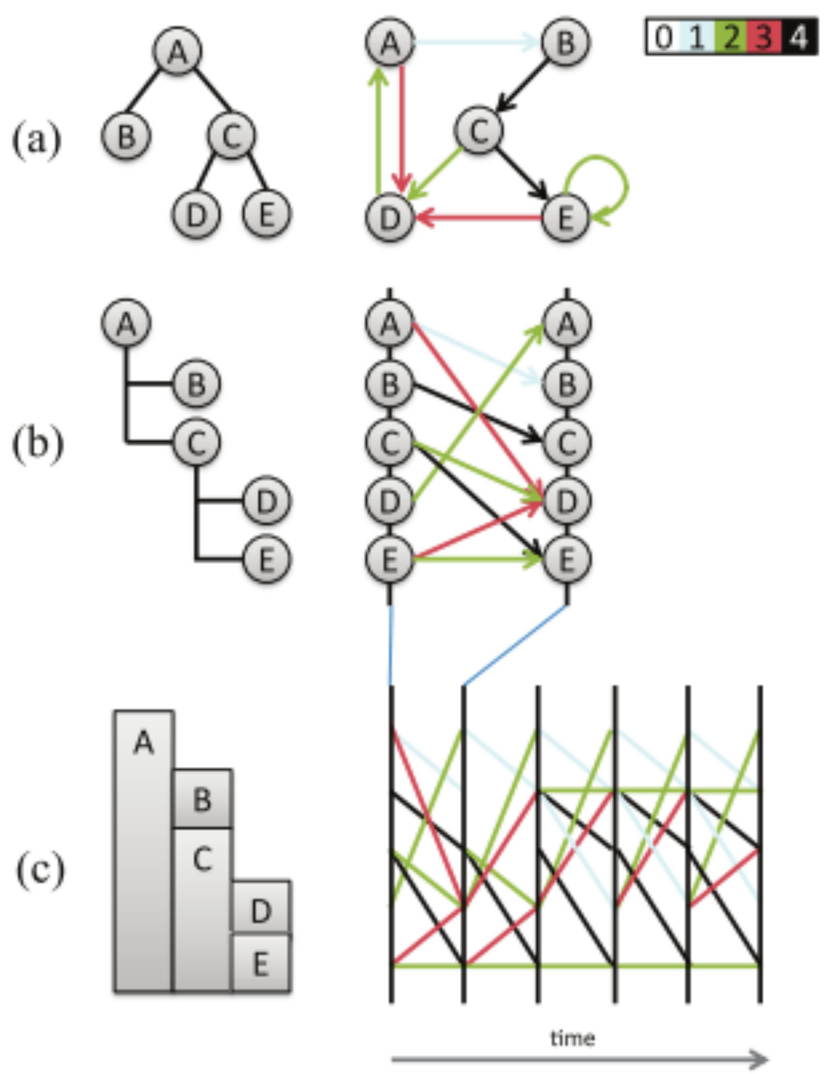
Let It Flow (Cui et al, 2011)

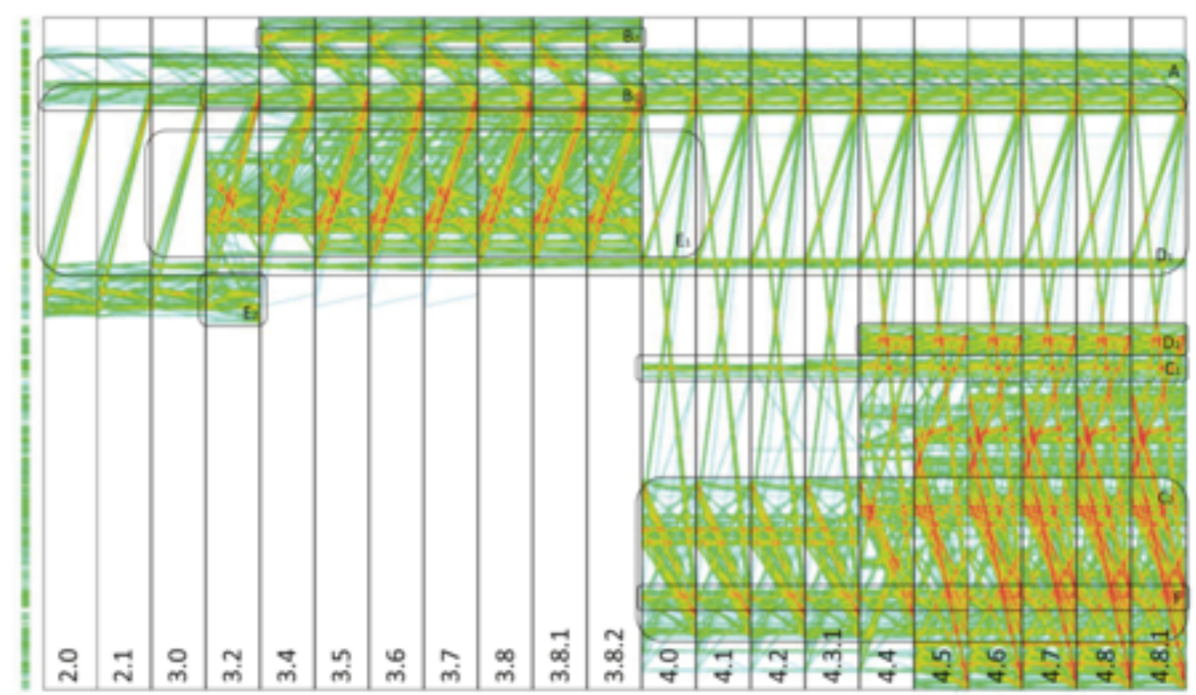


Visual Adjacency List (Hlawatsch et al, 2014)

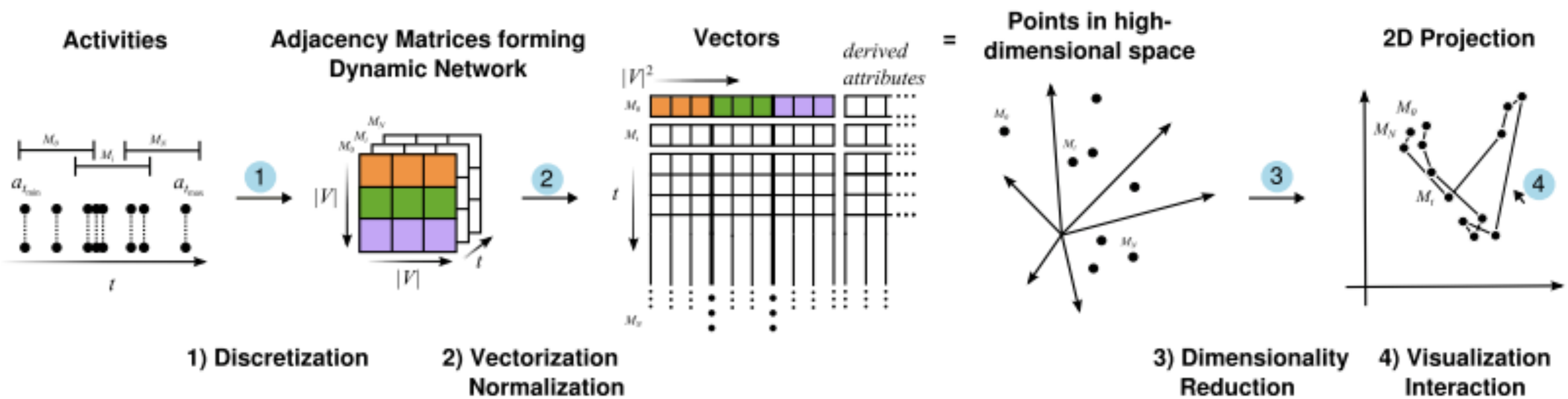


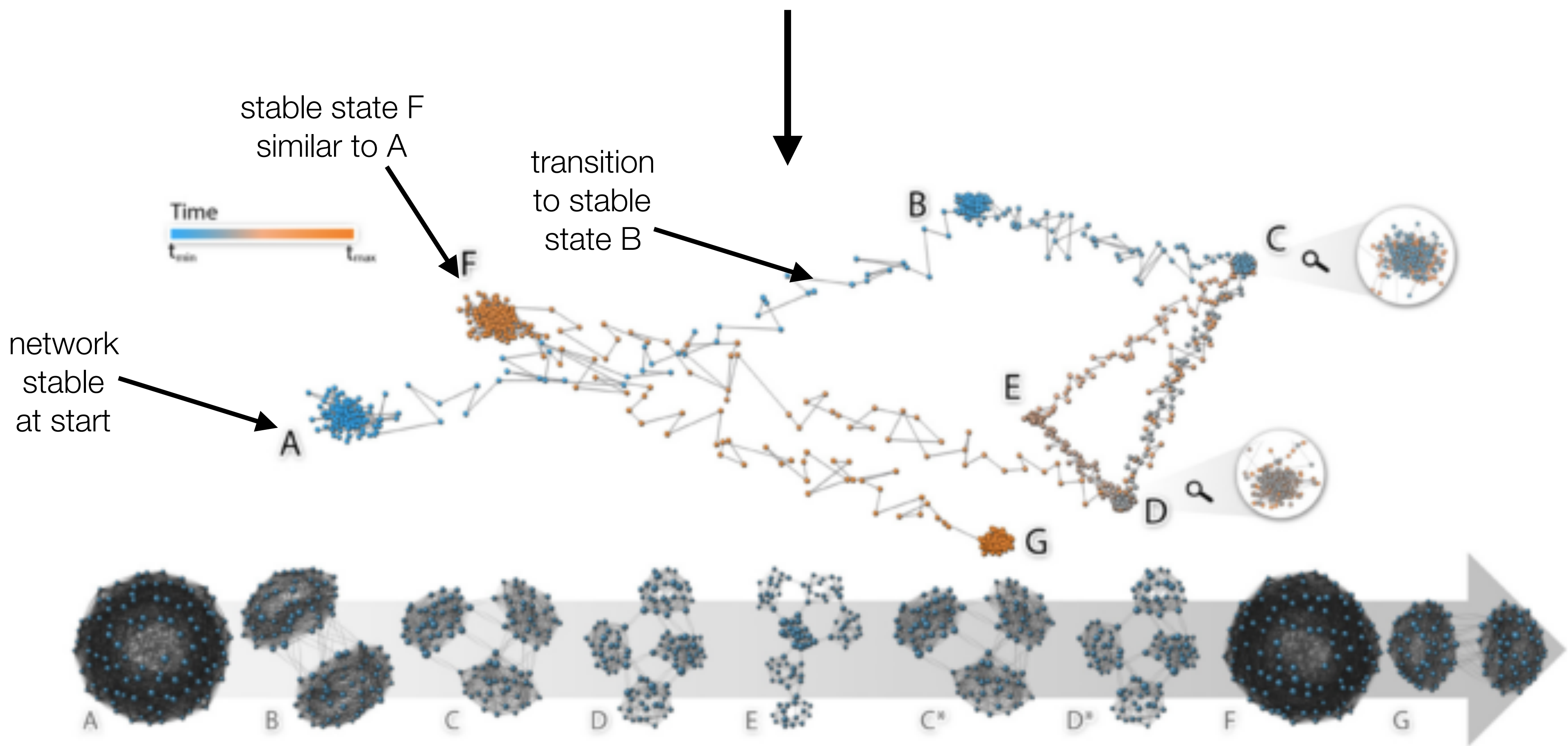
Parallel Edge Splatting (Burch et al, 2011)





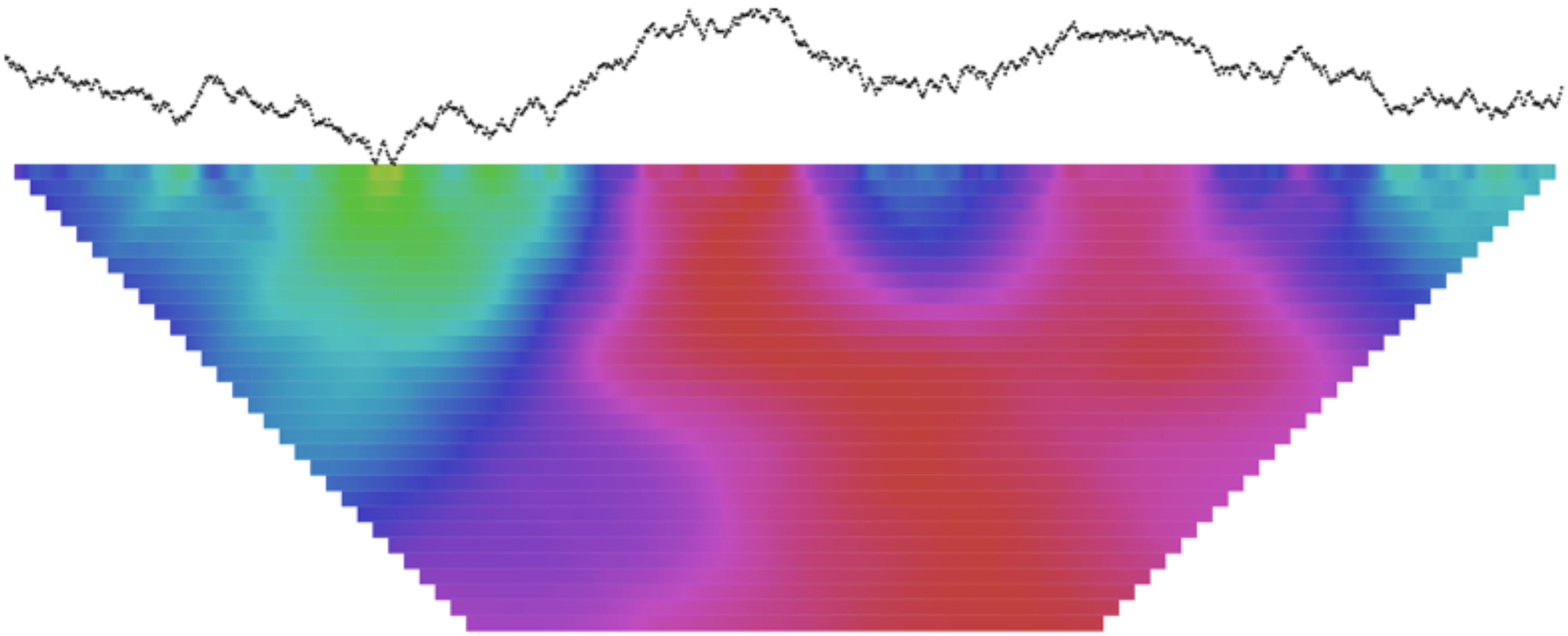
dynamic networks



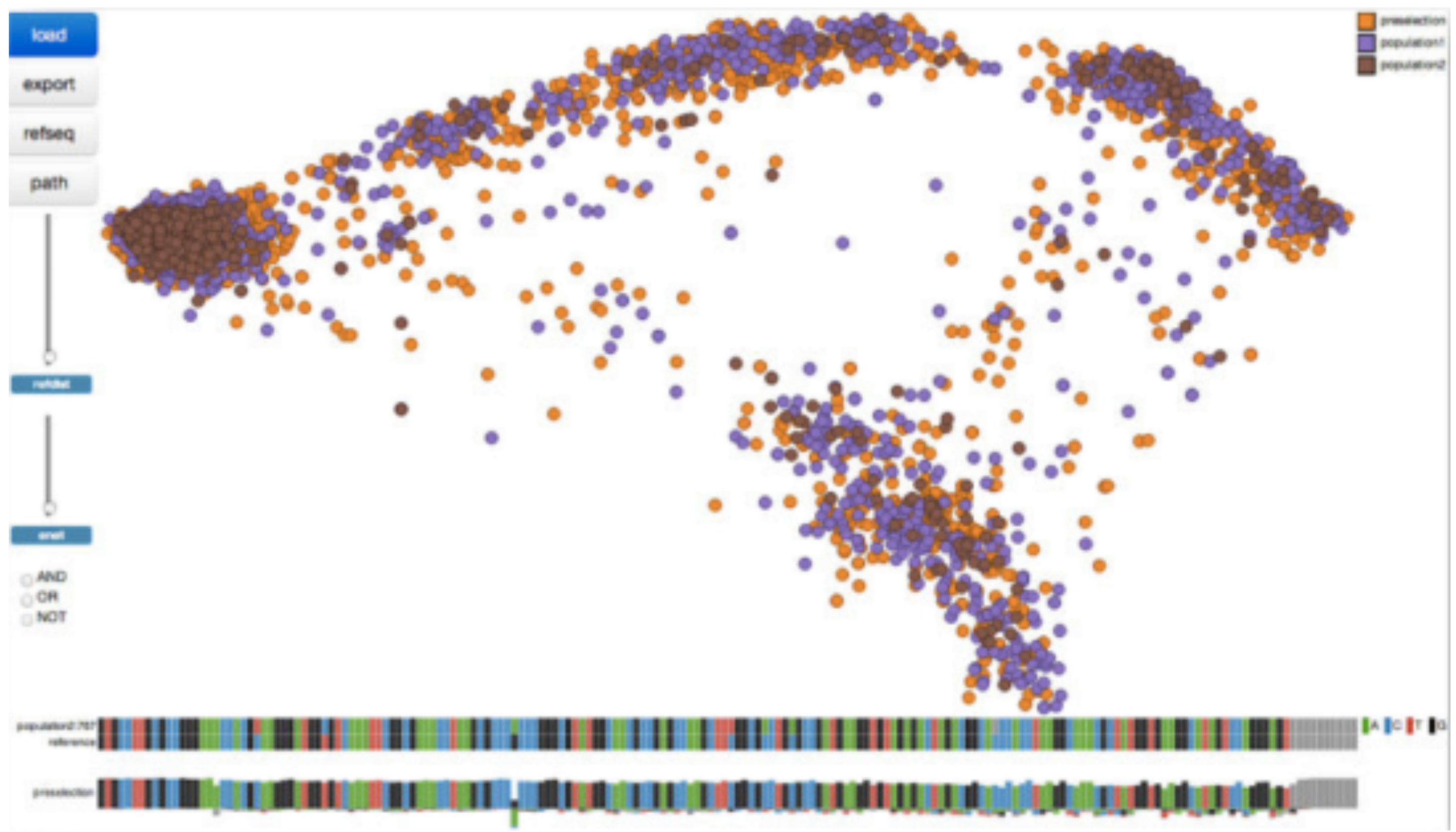


van den Elzen et al, 2015

continuous signal -> what should my histogram bin-sizes be?



invis - exploring high-dimensional RNA sequences from in vitro selection



Demiralp et al, 2013

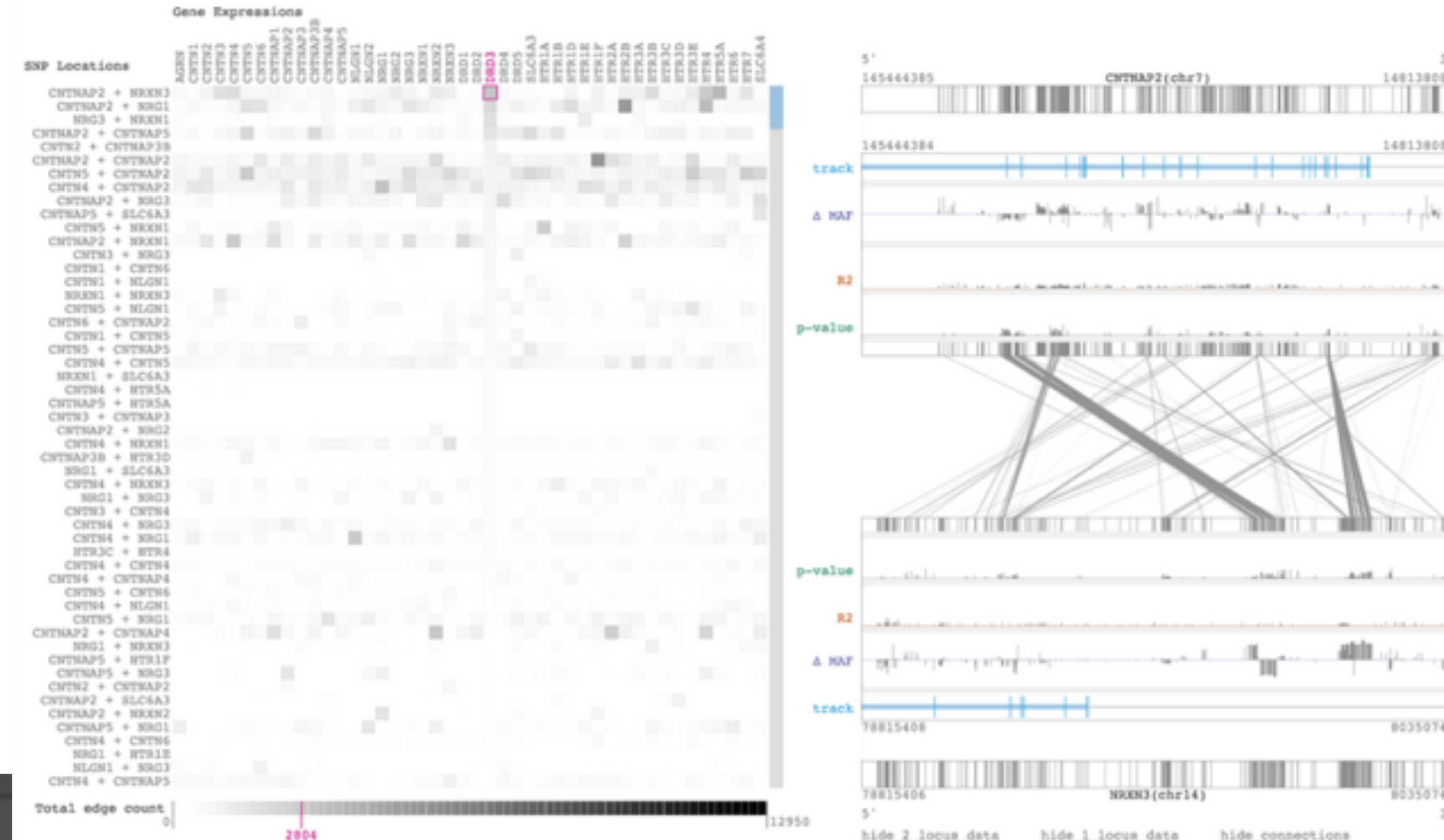
abyssExplorer - making sense of sequence assemblies

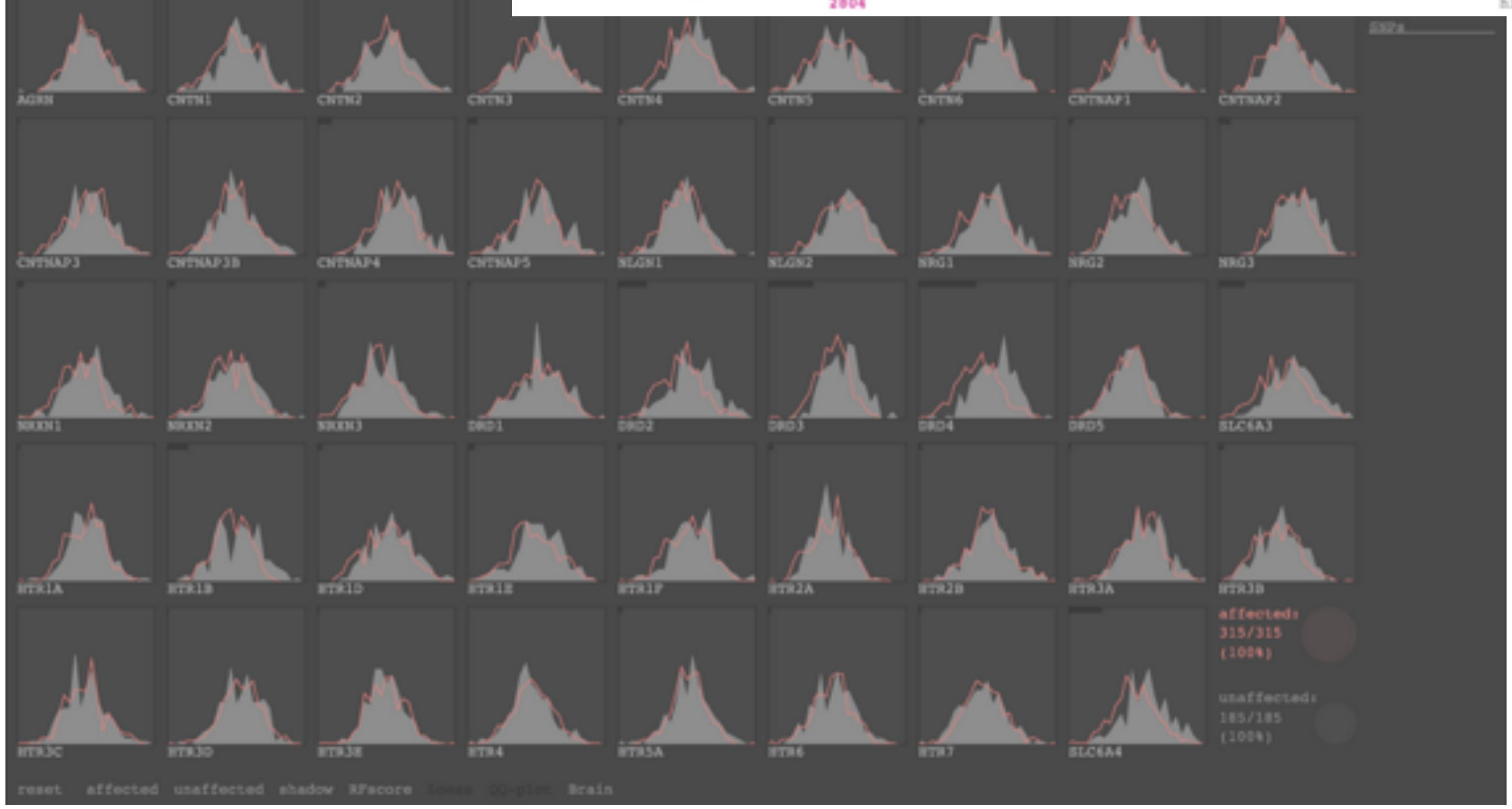
Nielsen et al, 2009

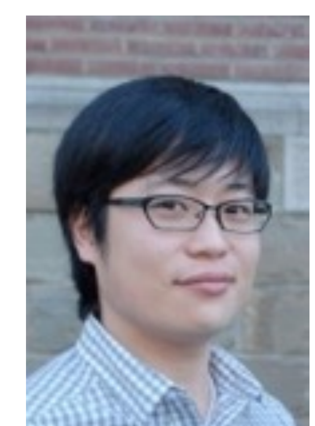
Aracari

Bartlett C et al. BMC Bioinformatics (2012)

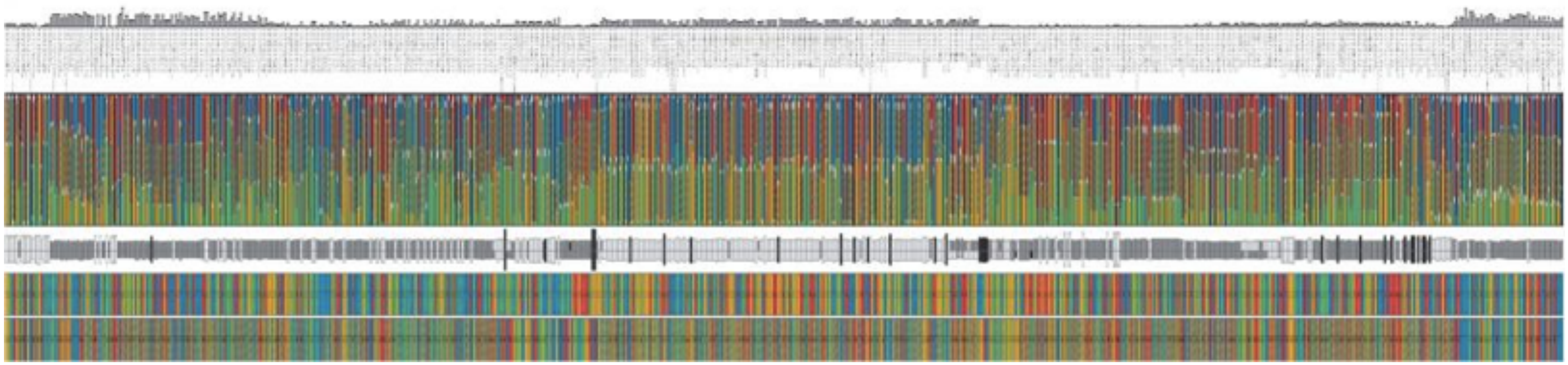
2-locus eQTL data





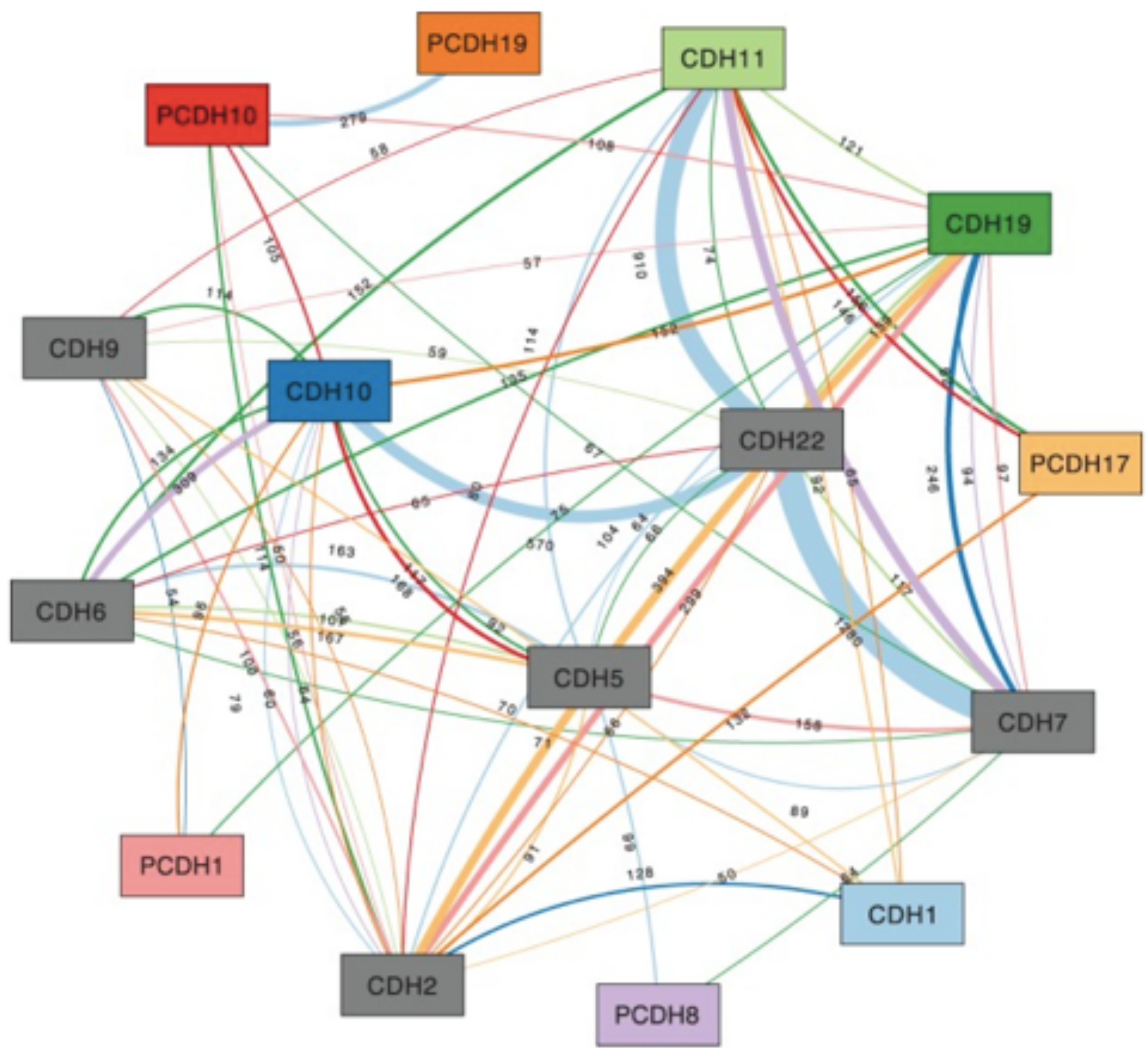


Ryo Sakai

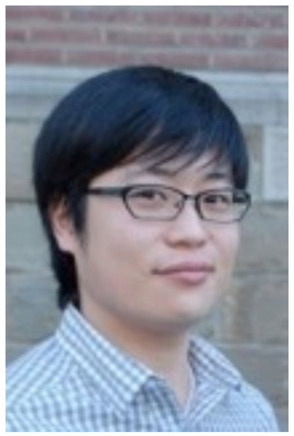


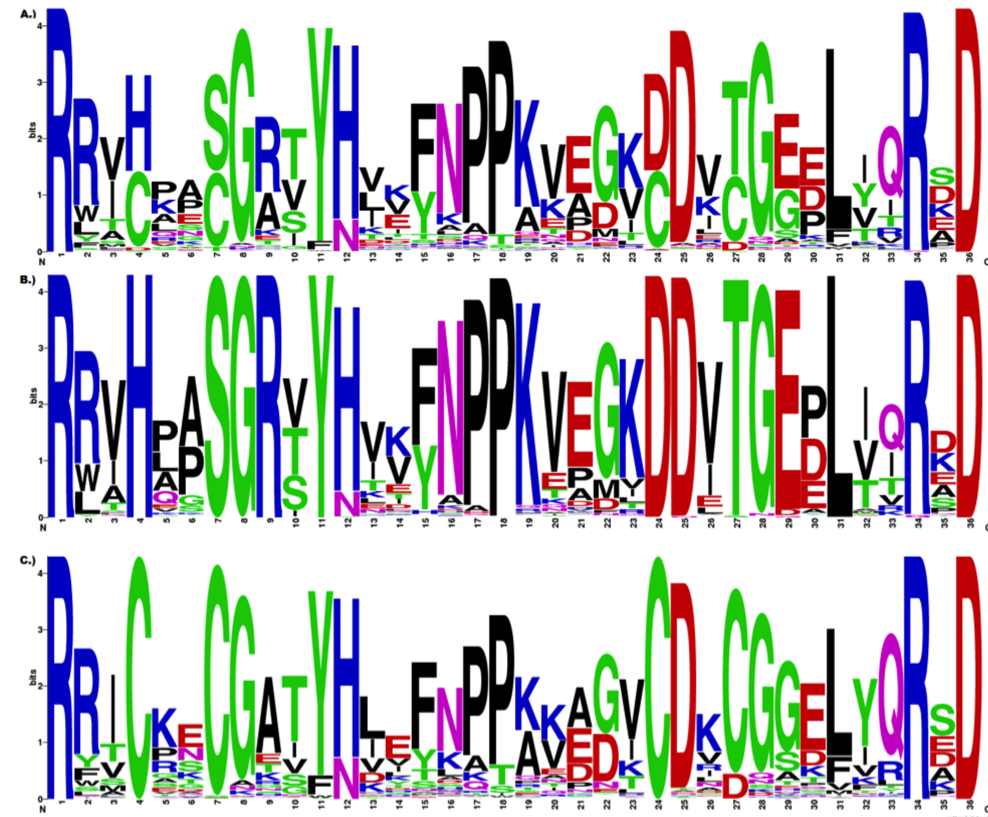
Reveal

Jäger, G et al. Bioinformatics (2012)

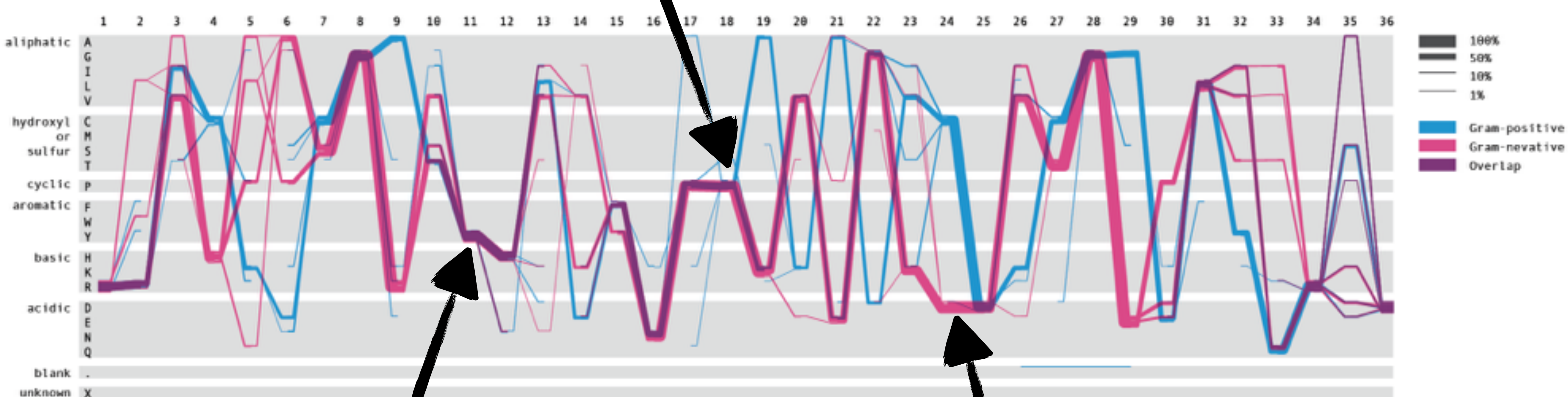


Sequence Diversity Diagram





subgroup

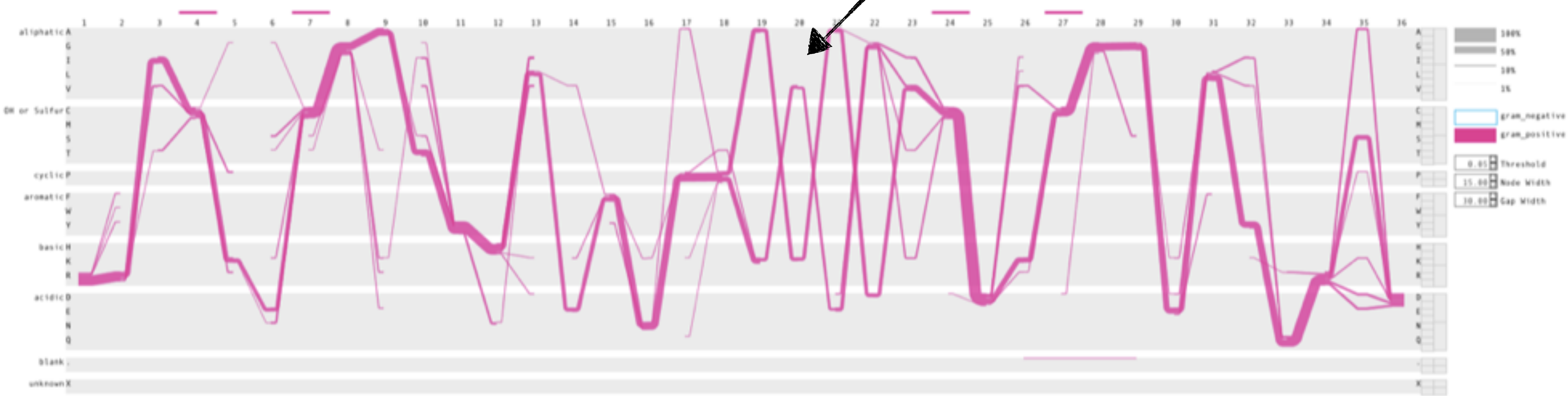


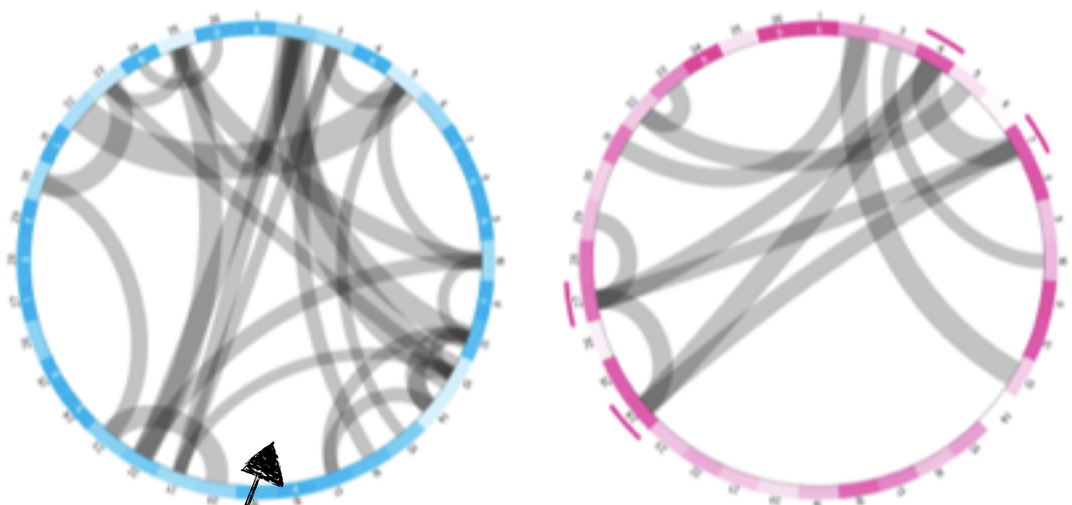
similarity

difference

Seagull

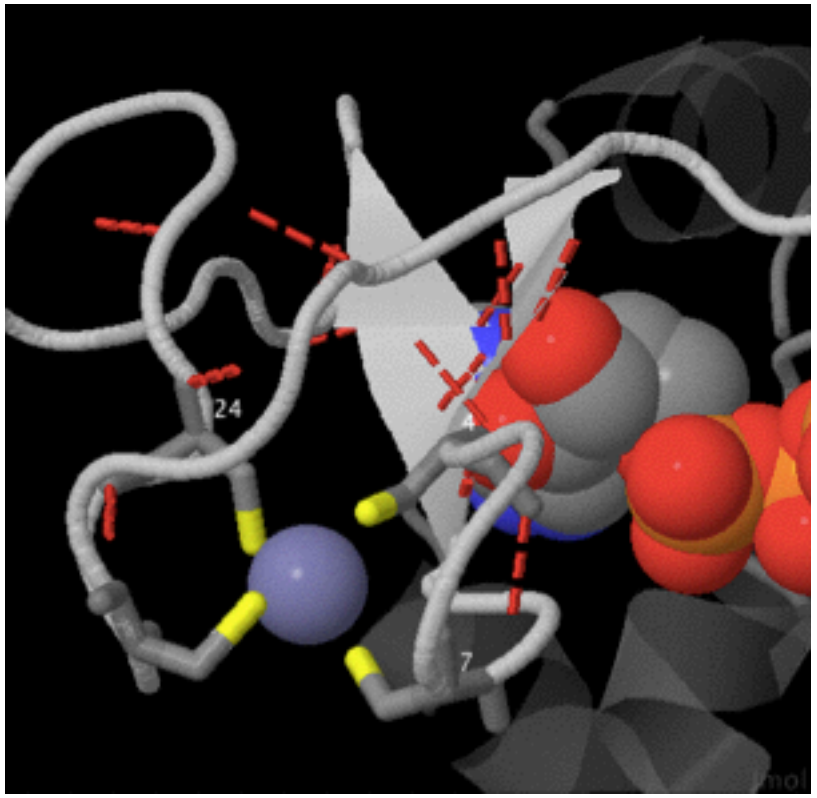
sequence diversity diagram



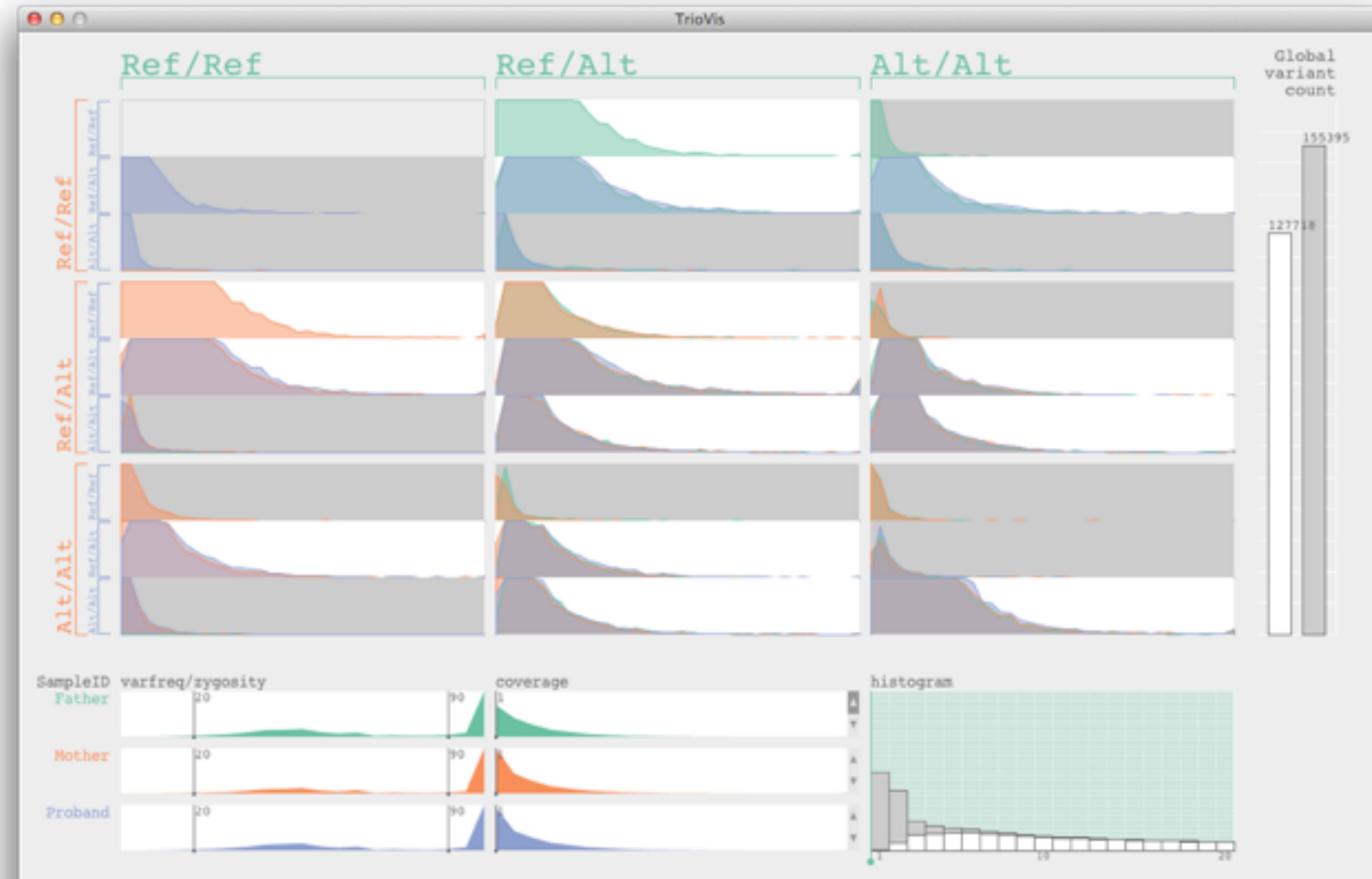
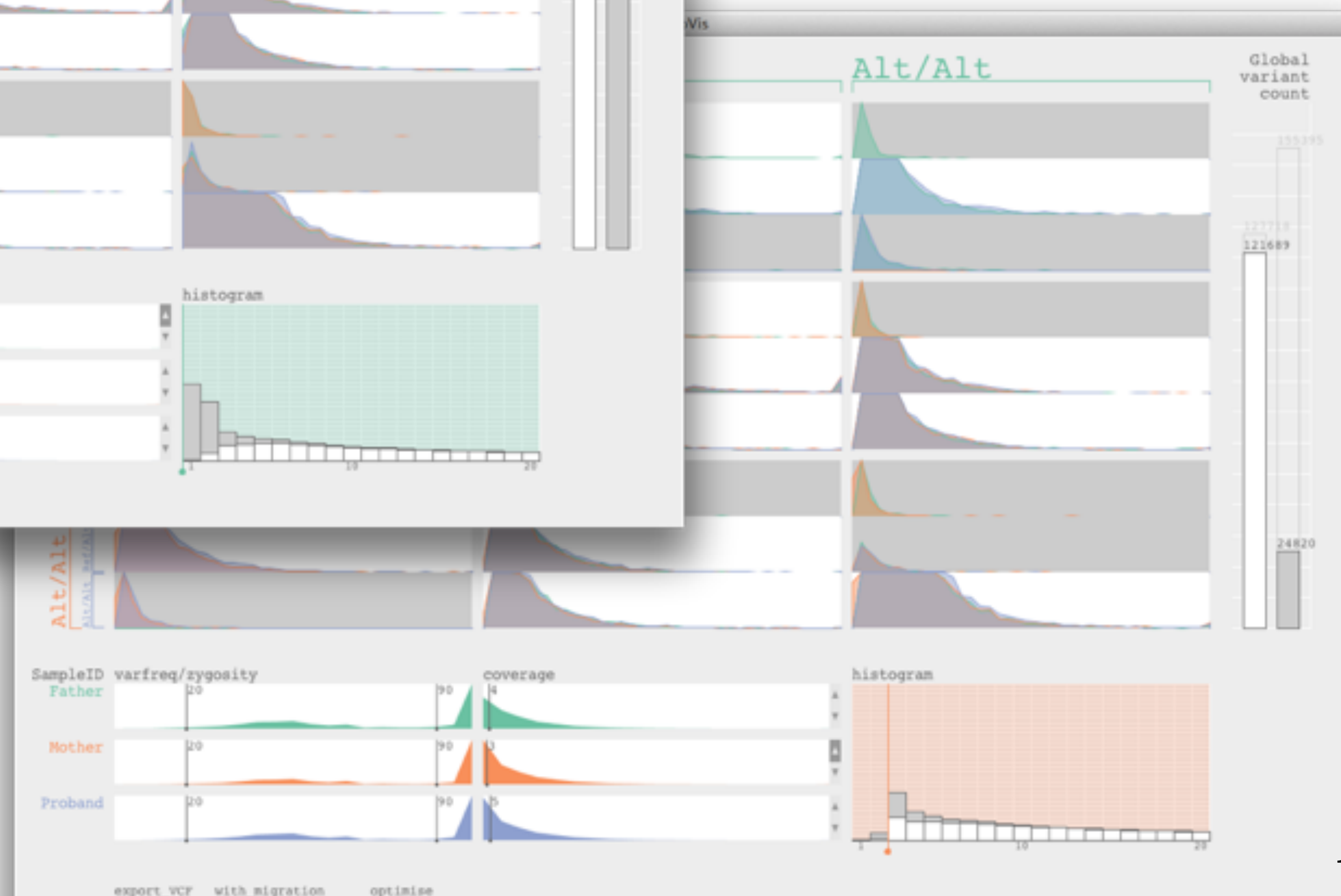


mutual information

3D view

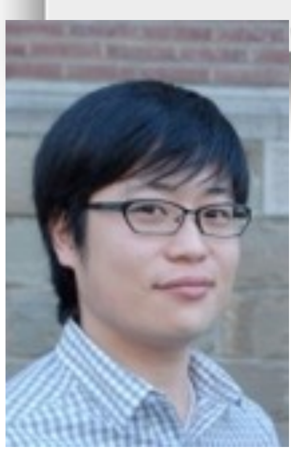


Data filtering (visual parameter setting)

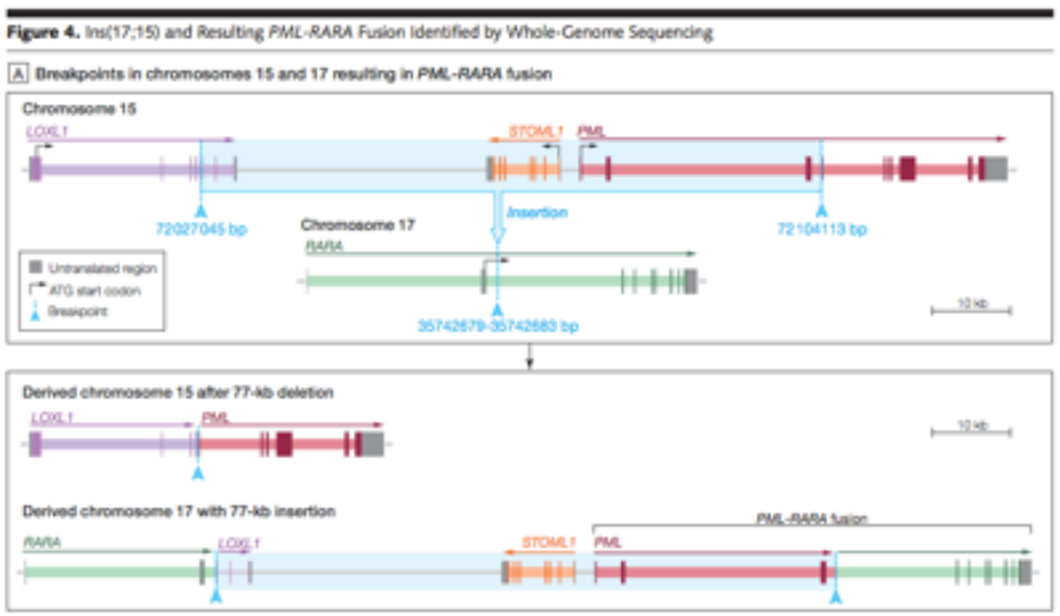
TrioVis

Sakai R et al. Bioinformatics (2013)

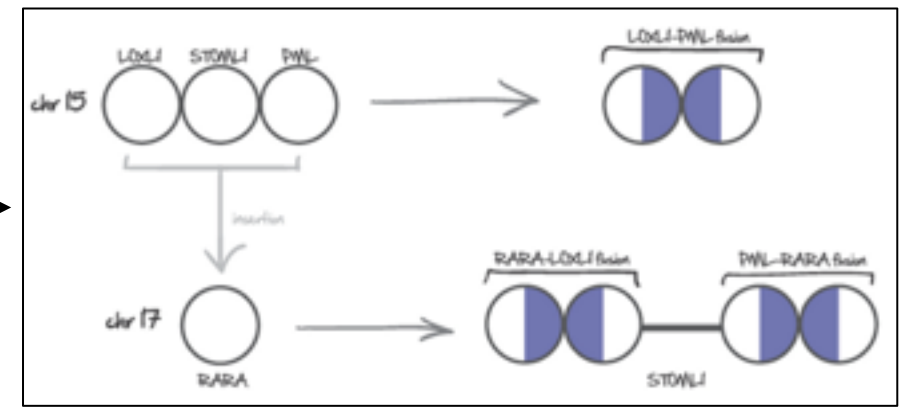


Ryo Sakai

Effect of structural genomic variation



Welch et al, 2011

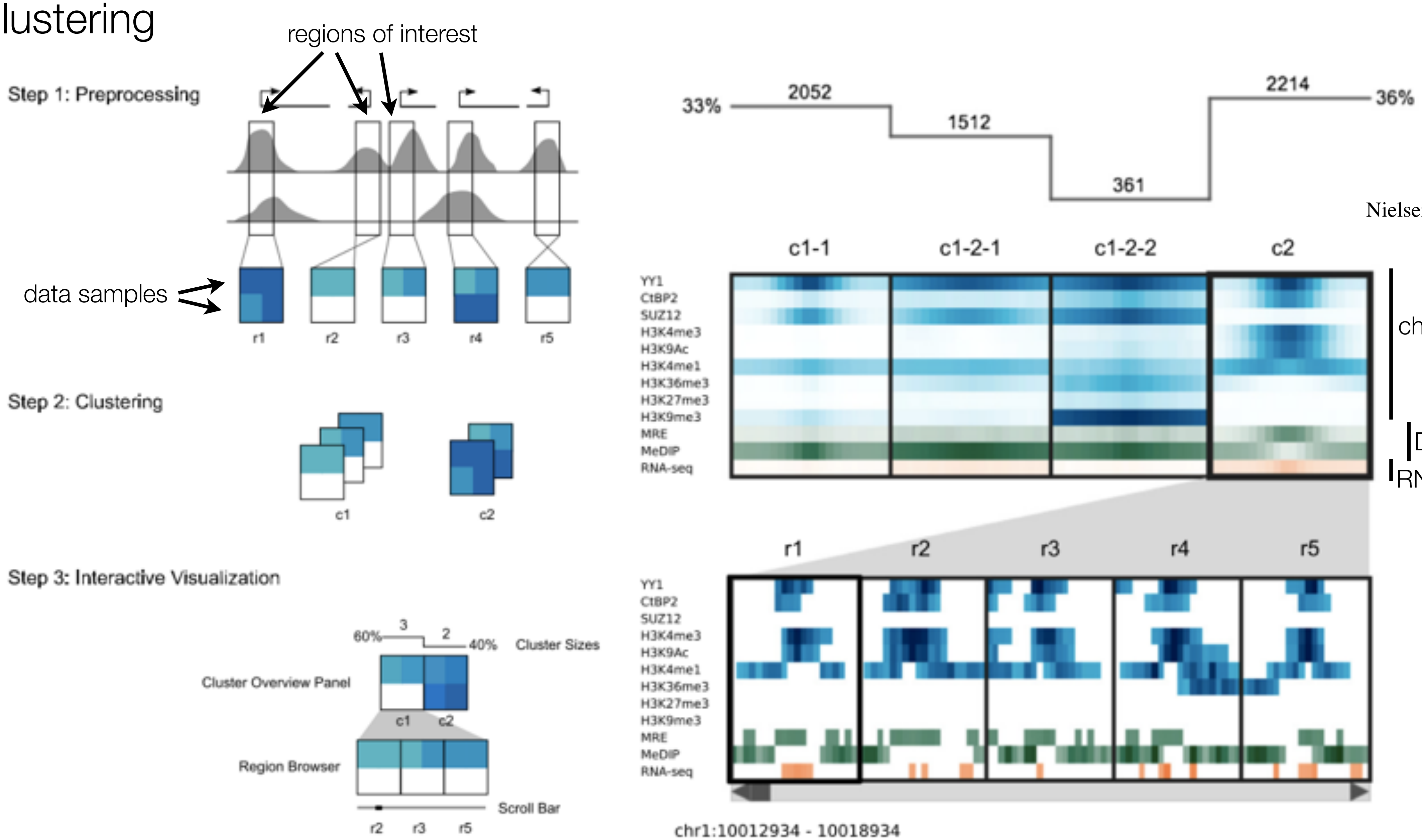


interested in *effect*, not *proof*

collab. Matthieu Moisse & Joke Reumers
University Hospital Leuven
Janssen Pharmaceuticals (J&J)

User-guided analysis

clustering



Spark

Nielsen et al. Genome Research (2012)

Overview of lecture

A. Why visual analytics?

B. Data visualization

- Data foundations
- Human perception foundations
- Visualization foundations and examples

C. Visualization evaluation

D. Tools of the trade

C. Visualization evaluation

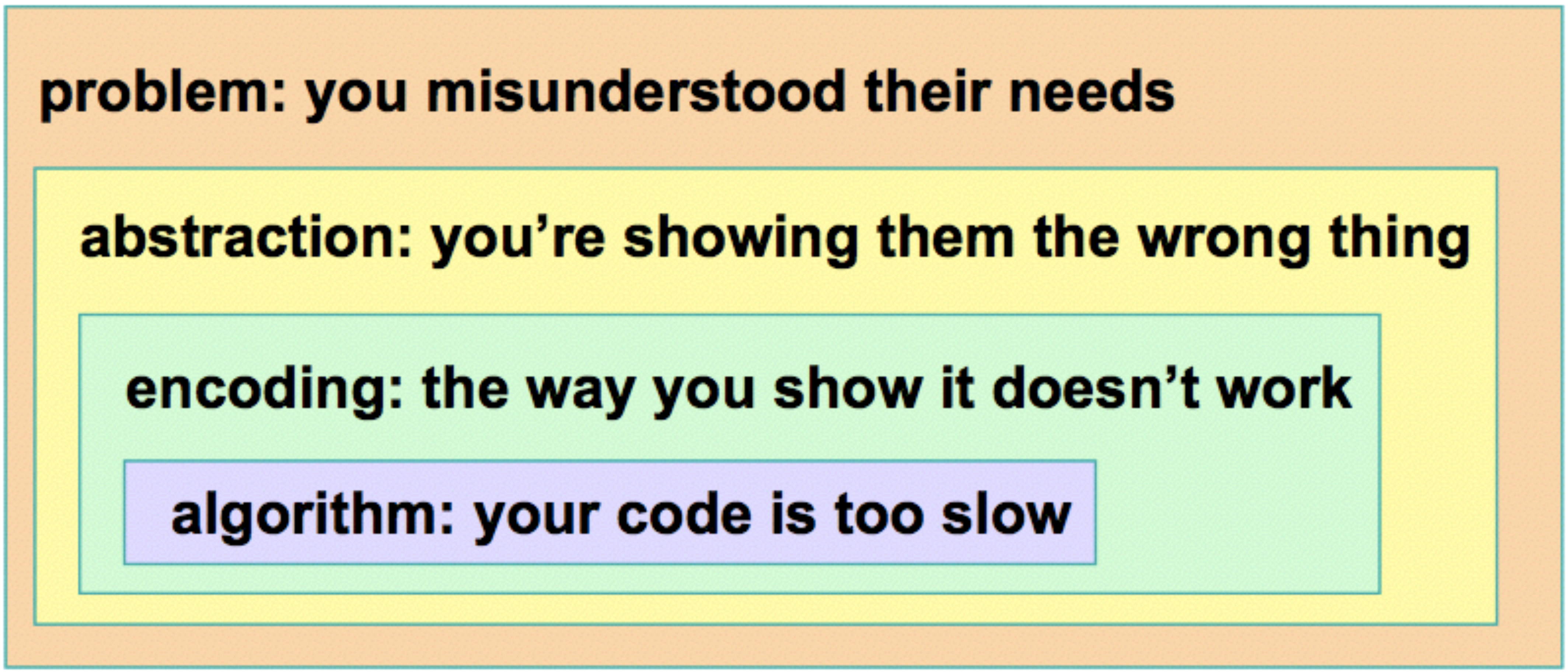
Quantitative evaluation

- spike data with known signal, and record time that it takes for user to find that signal => measure => run statistics “this visualization is better than that one”
- user tasks:
  - identify
  - locate
  - distinguish
  - categorize
  - cluster
  - rank
  - compare
  - associate
  - correlate

Qualitative evaluation

- very close interaction with domain expert
- let expert use the interactive visualization and try to find out what insights he/she gained from the visualization
  - experimenter observation
  - think-aloud protocol
  - collecting participant opinions

Make sure you measure the right thing



Overview of lecture

A. Why visual analytics?

B. Data visualization

- Data foundations
- Human perception foundations
- Visualization foundations and examples

C. Visualization evaluation

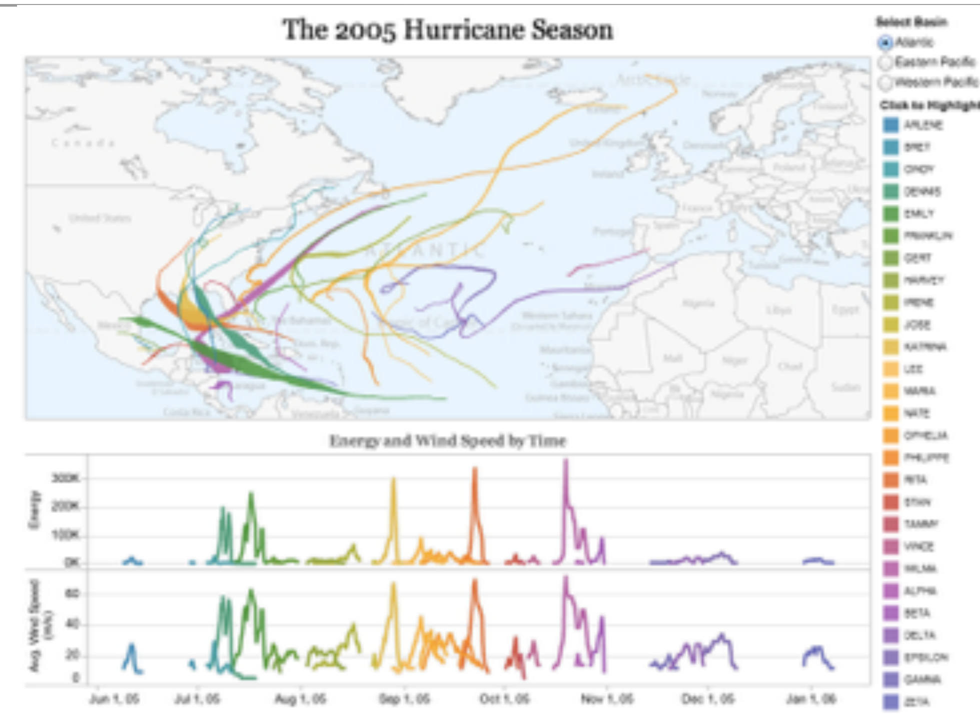
D. Tools of the trade

D. Tools of the trade

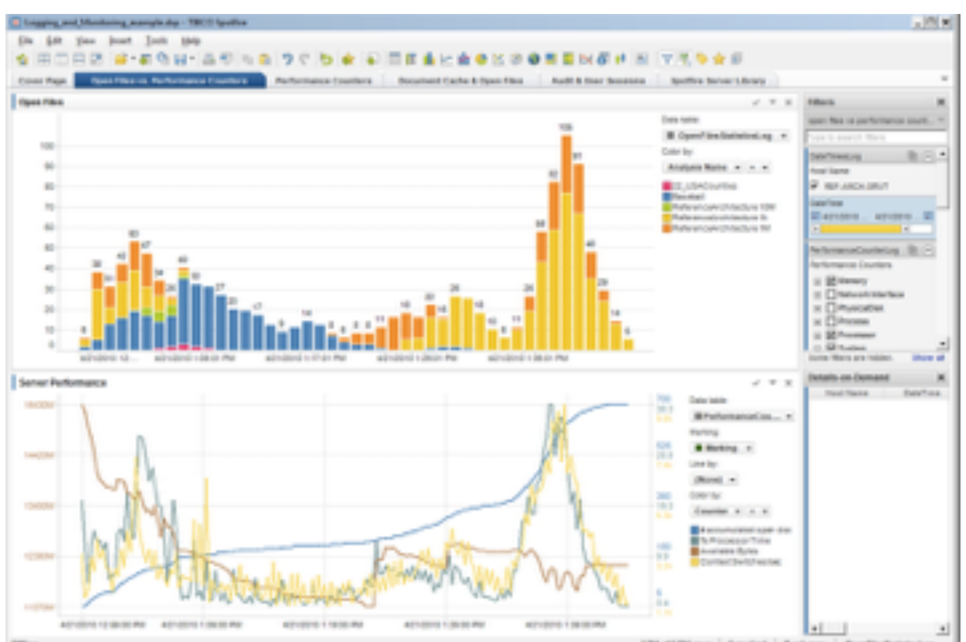
using
drawing
coding

Using data visualizations

Microsoft Excel



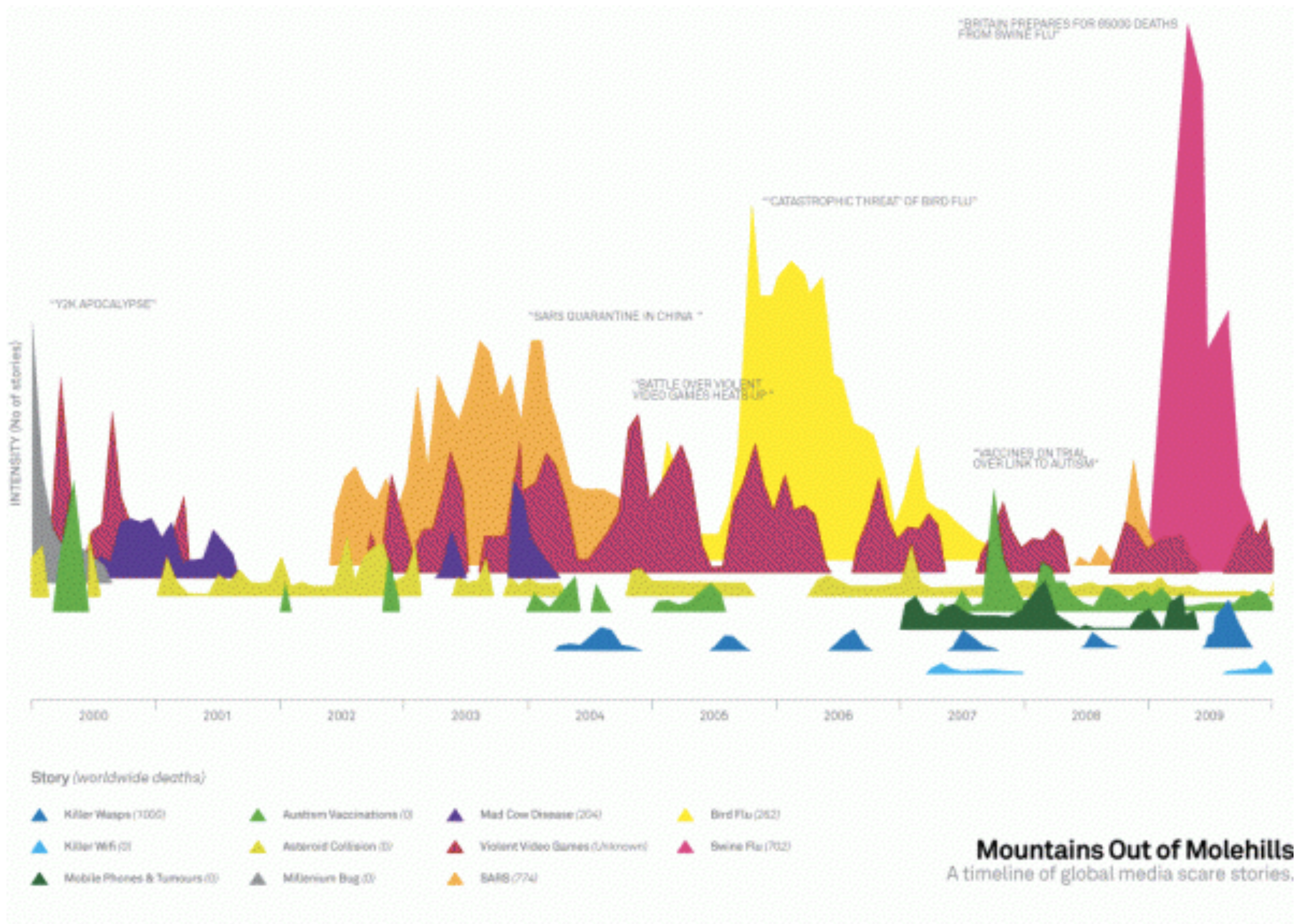
Tableau



Tibco Spotfire

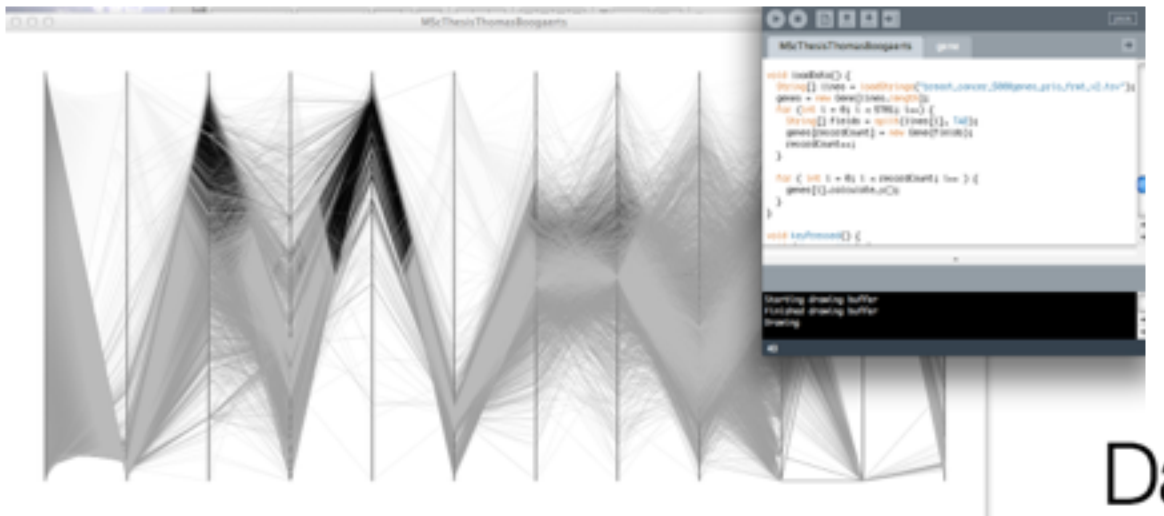
ManyEyes

Drawing data visualizations

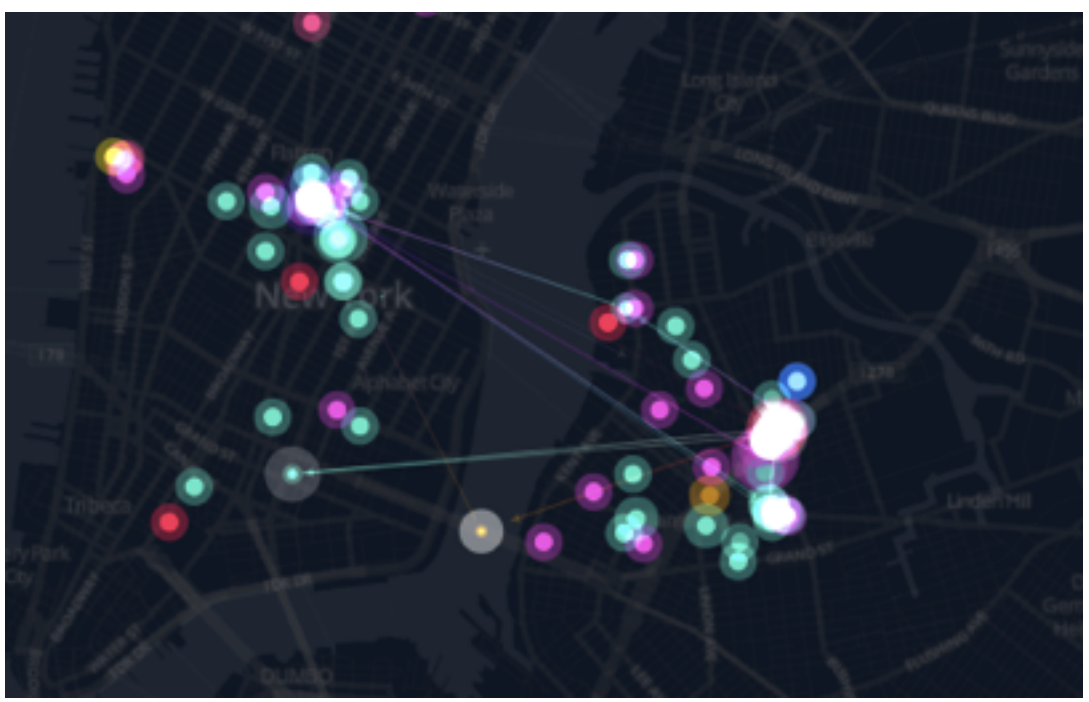


Adobe Illustrator

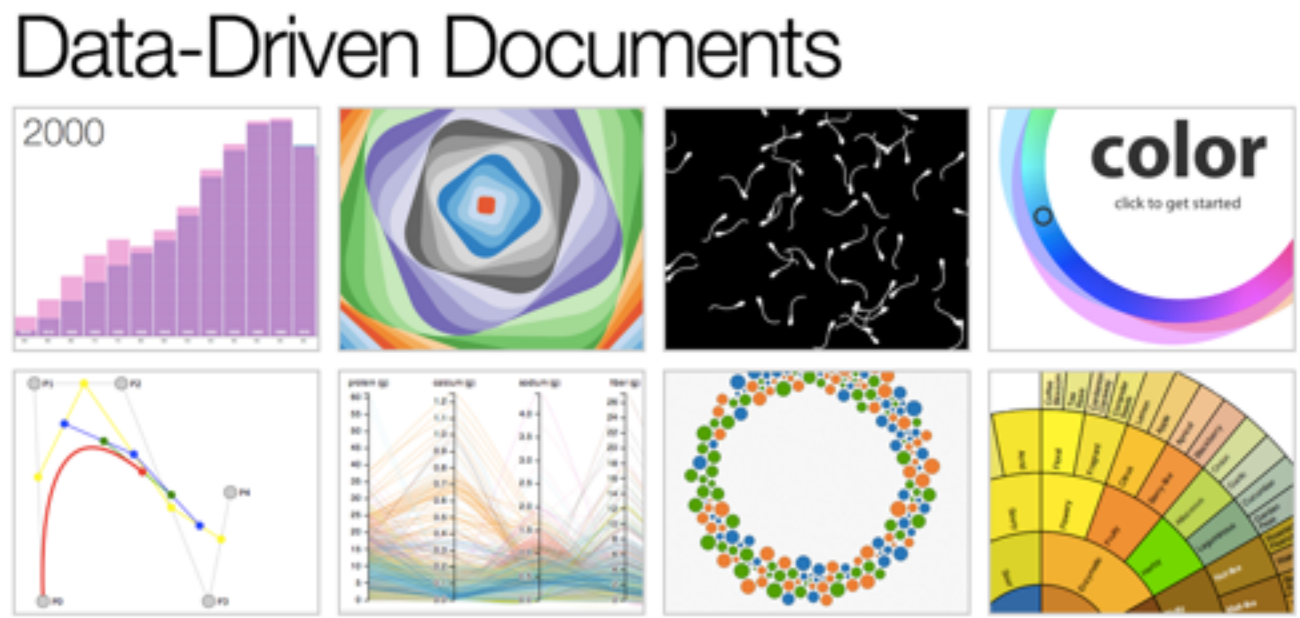
Coding data visualizations



processing.org



paper.js



d3.js


LOONEY TUNES
That's all Folks!

(…actually, we only scratched the surface…)