

[I0U19A]

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 dataviz: Daniel Alcaide, Jansi Thiyagarajan, Houda Lamqaddam



Overview of lecture

A. Why visual analytics?

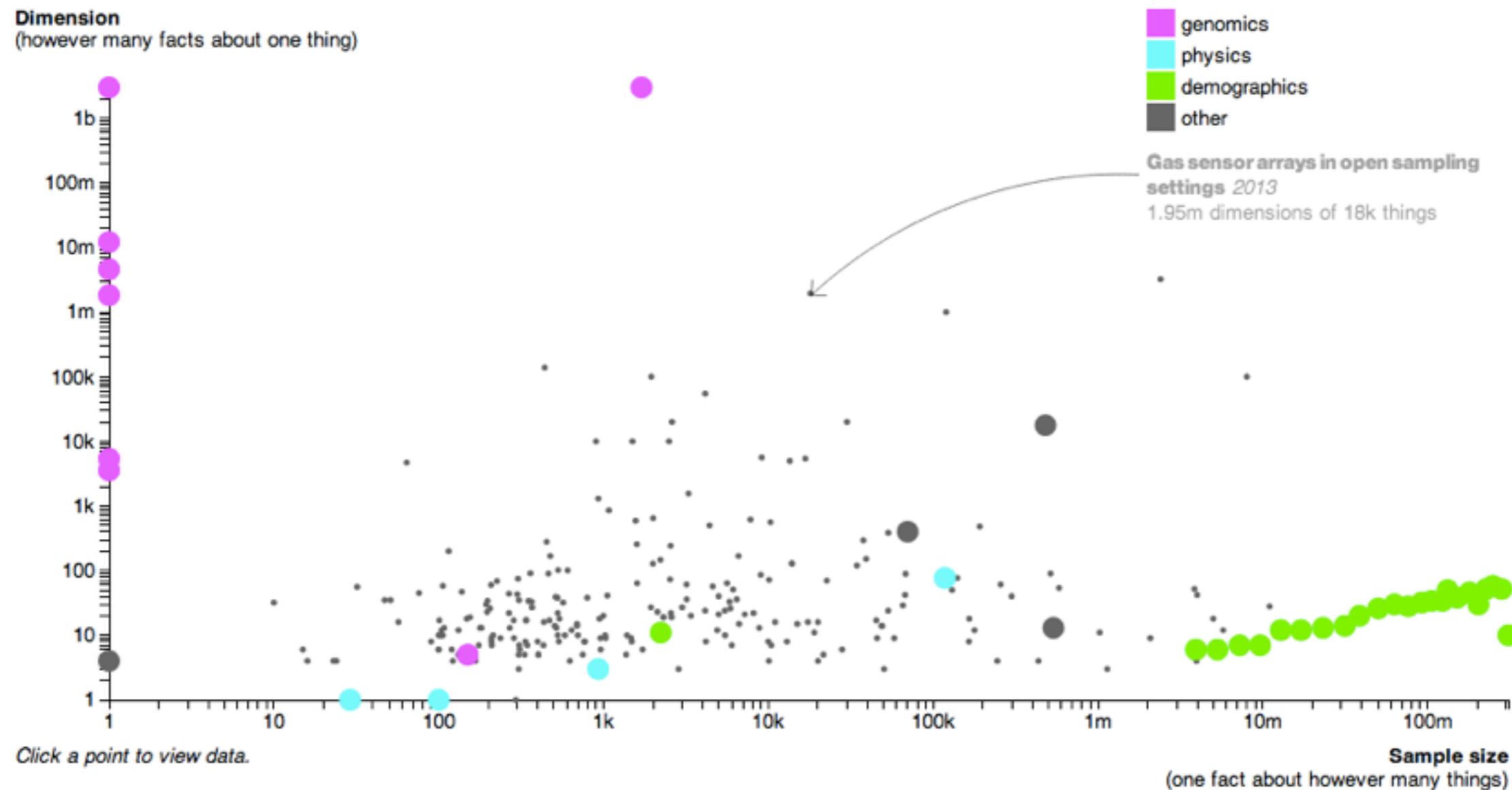
B. Data visualization

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

C. Visualization evaluation

D. Tools of the trade

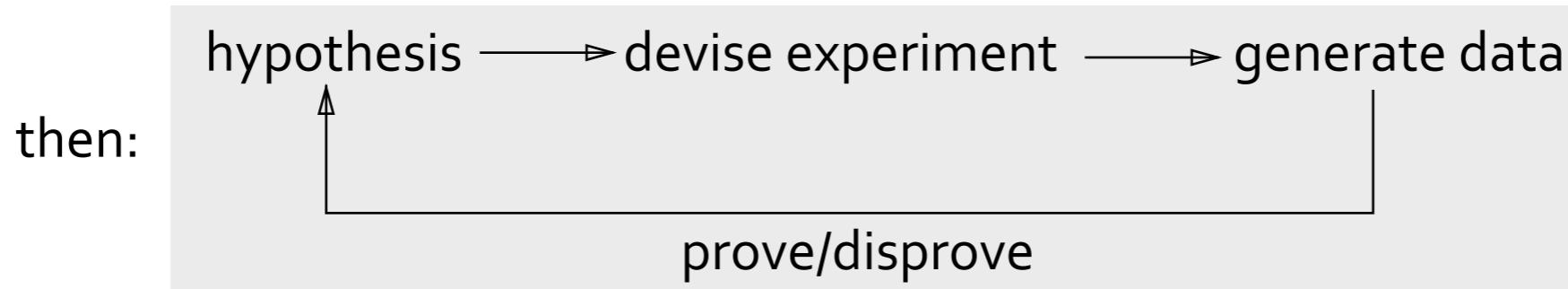
A. What's the problem?



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

Scientific Research Paradigms (Jim Gray, Microsoft)

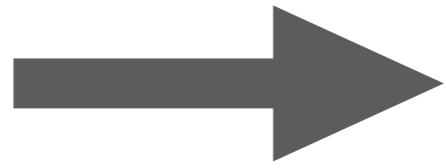
1st	1,000s years ago	empirical
2nd	100s years ago	theoretical
3rd	last few decades	computational
4rd	today	data exploration



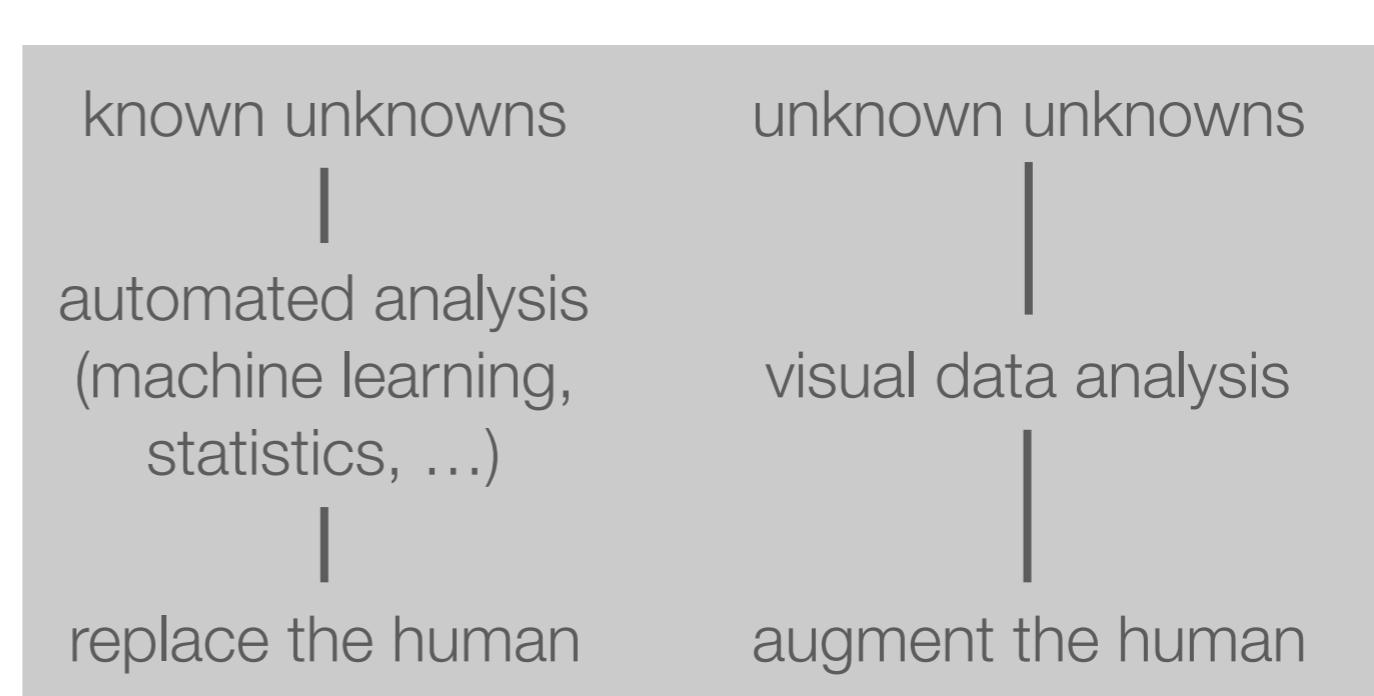
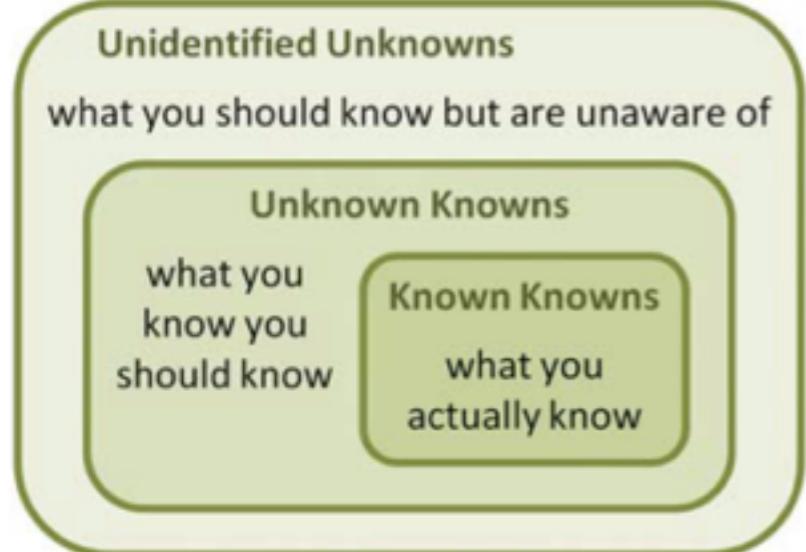
now:

```
graph LR; A[have data] --> B[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



Opening the black box

complex algorithms
in data analysis



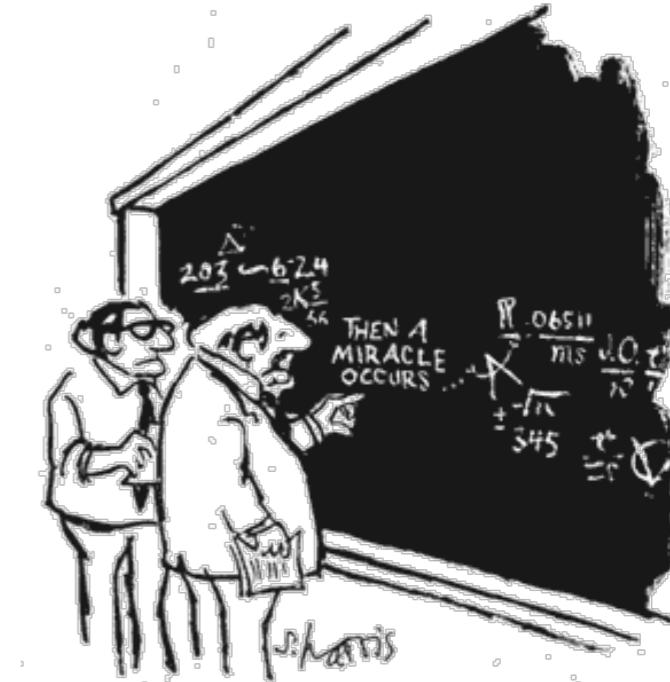
obvious **link between
input and output**
difficult to see



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

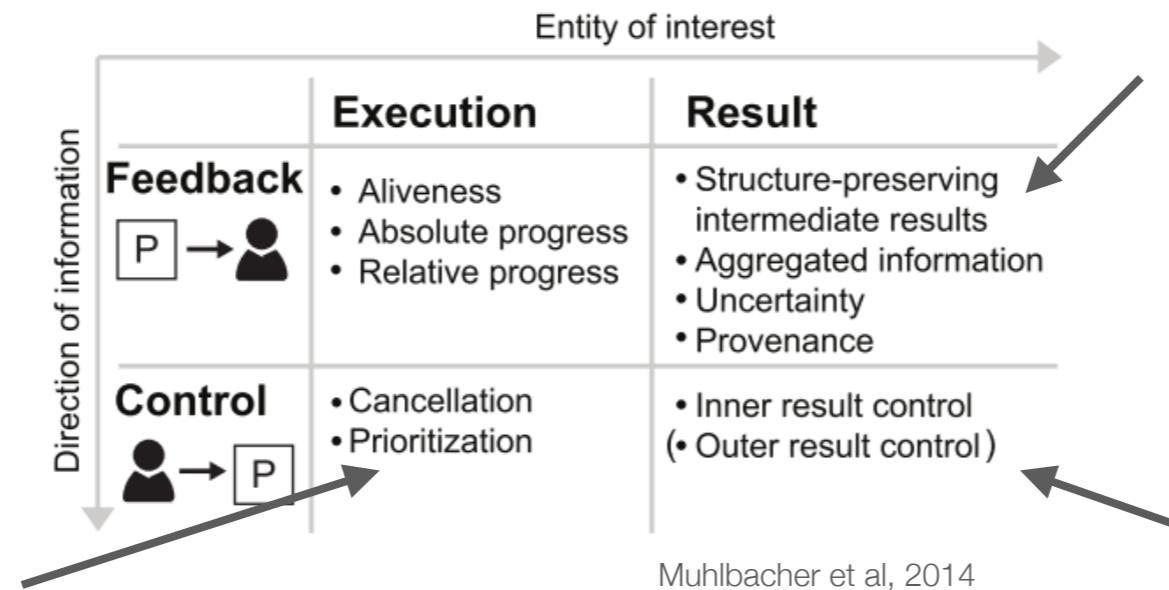
result:
algorithm acts as a
black box

insight
in data
in algorithms



• "I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

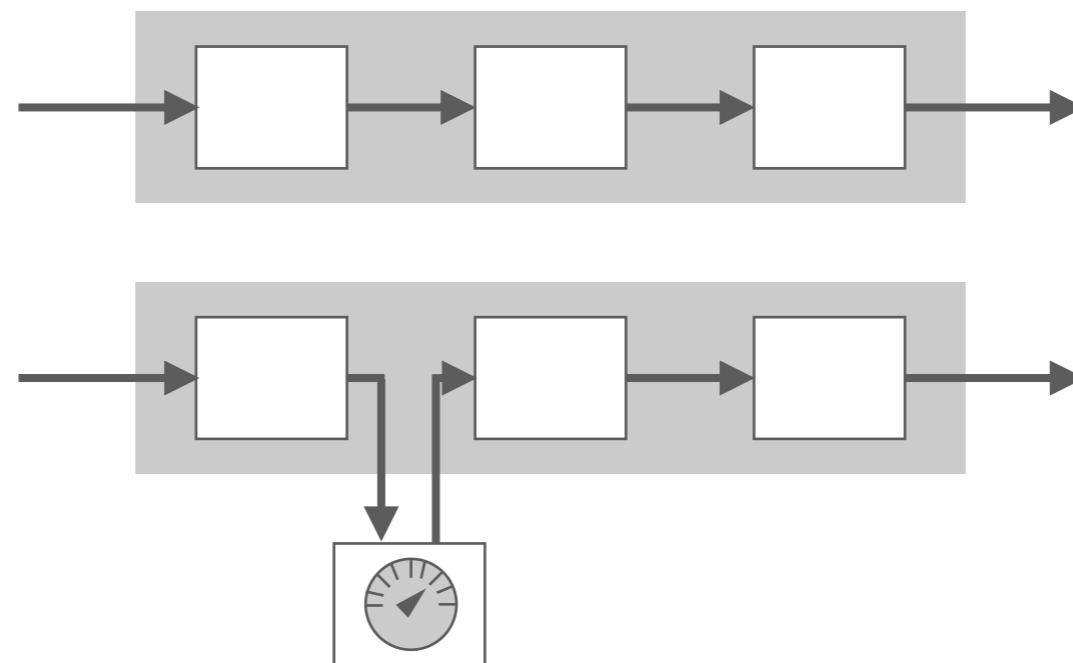
high-dimensional data
cell line + treatment +
target + compound + ...



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

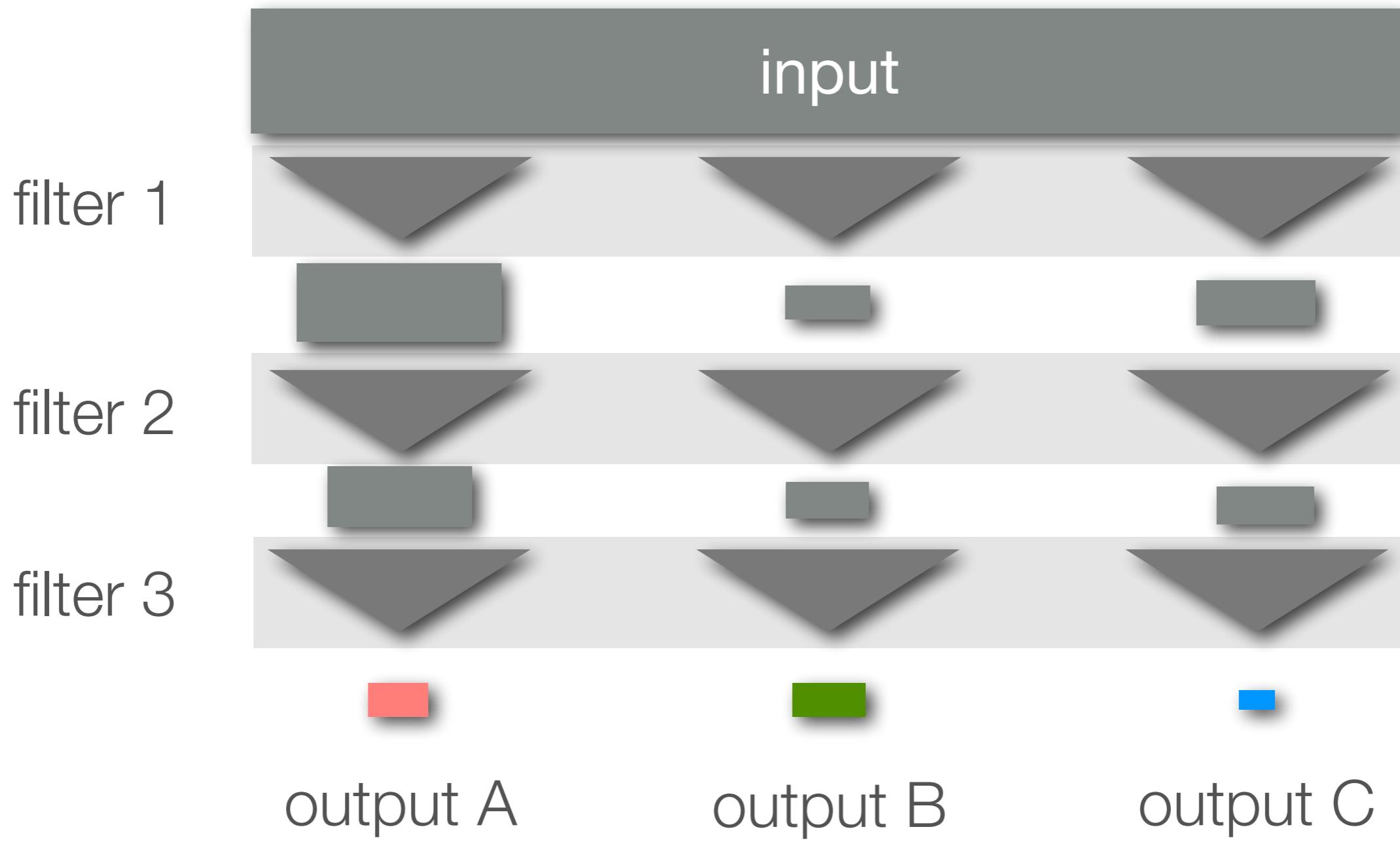
intermediate feedback: e.g. intermediate result, goodness-of-fit for best subset so far

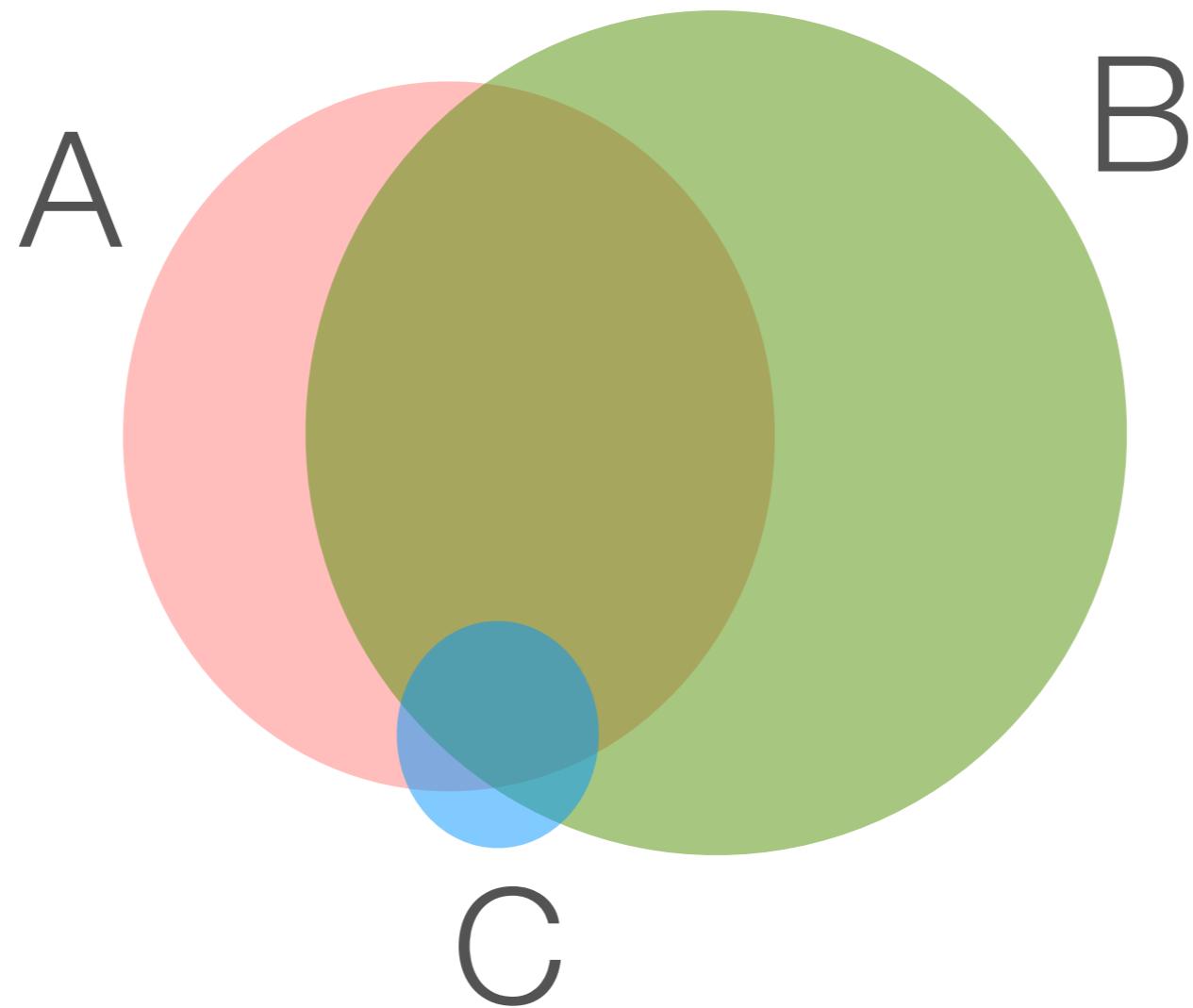
steer the final result: early validation of intermediate results, guided feature selection, weighting, avoiding being stuck in local minima, ...

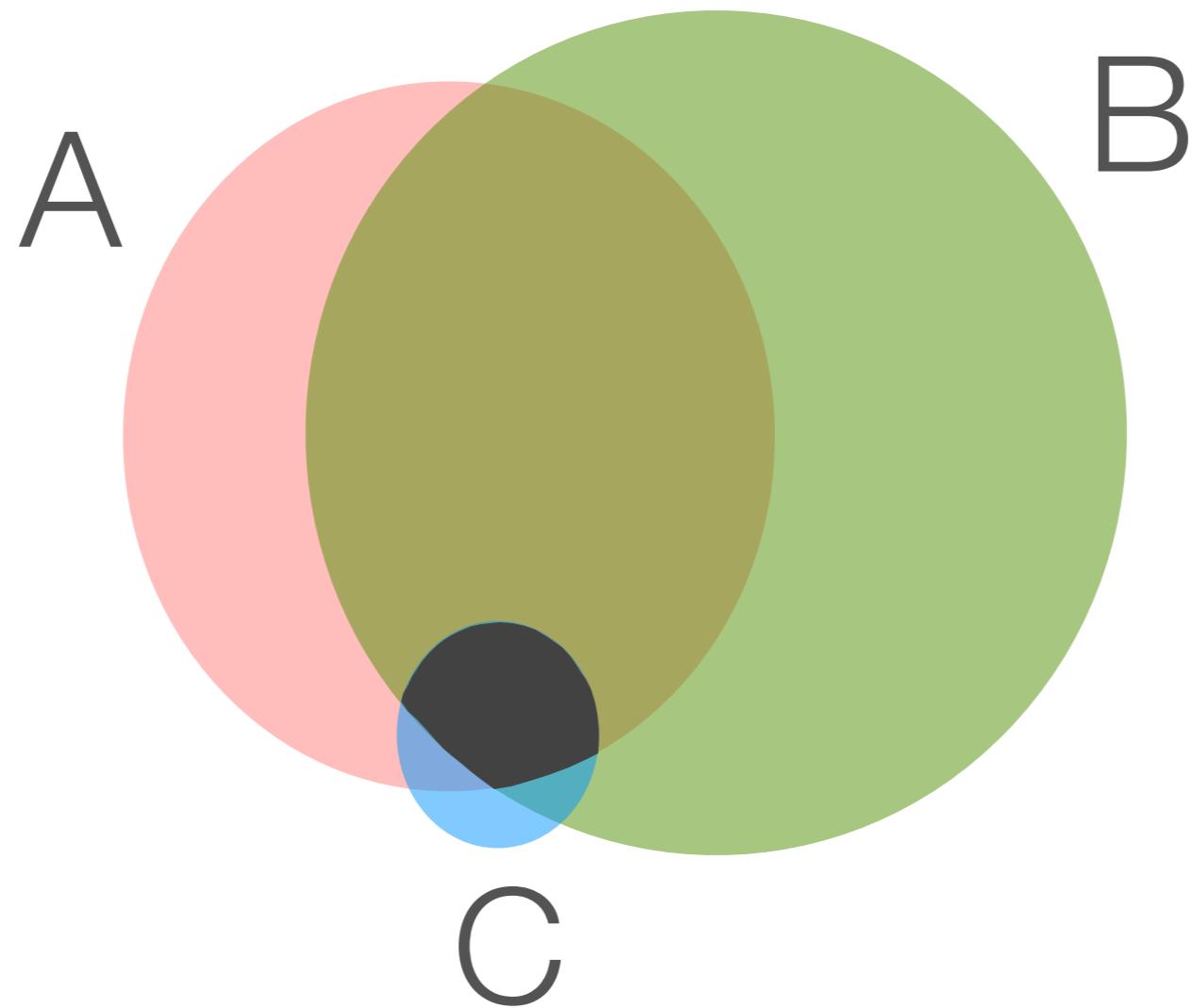


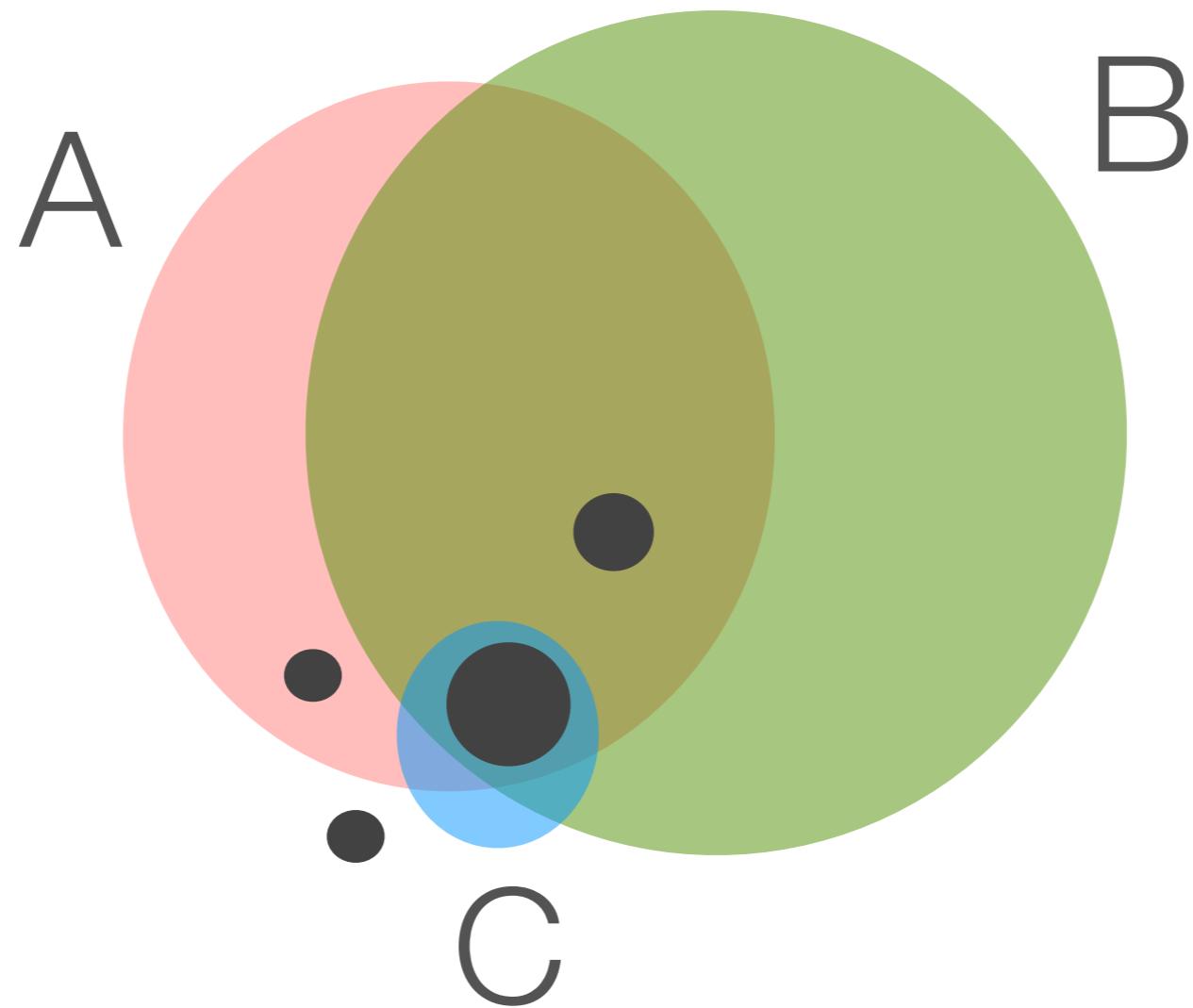
Example: mutation filtering

going from 5-6 million to 20,000

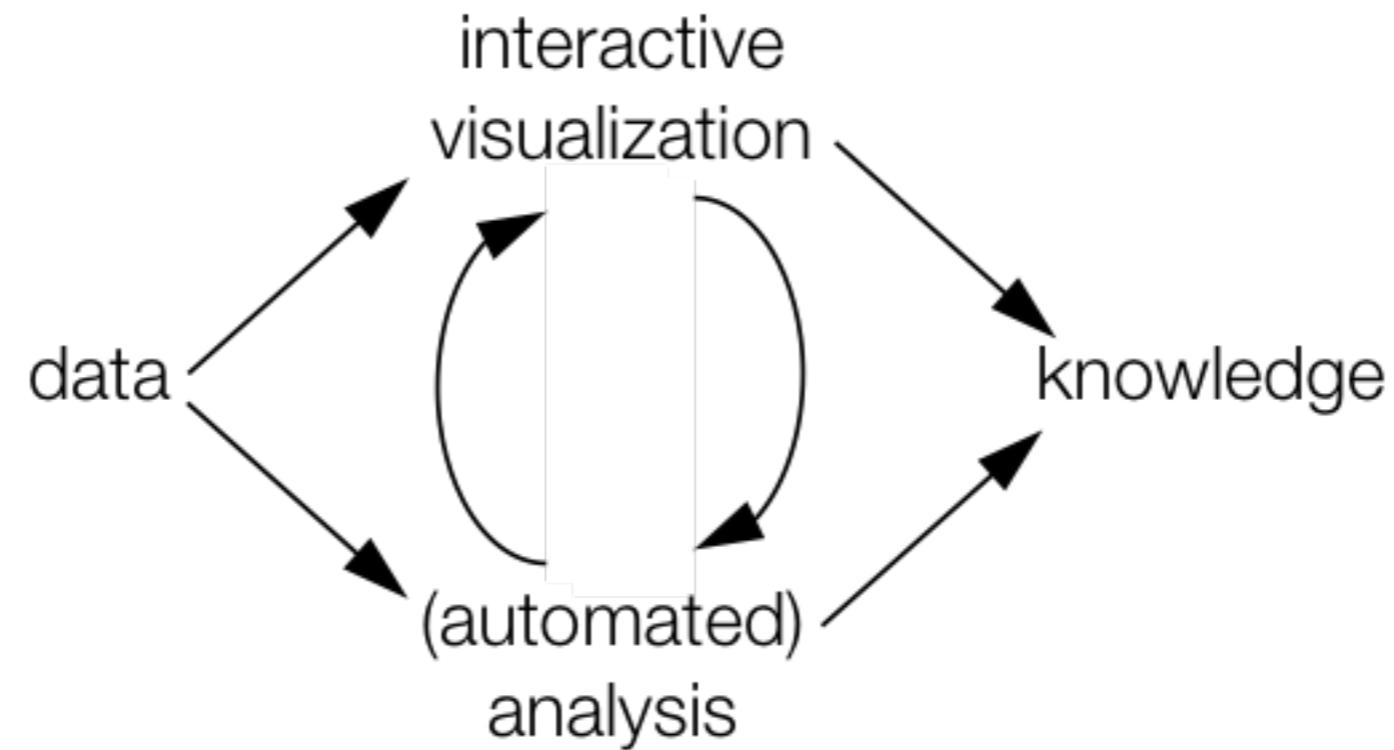








Visual Analytics to the rescue



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

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

perception vs cognition

human in the loop needs the details

cognition \Leftrightarrow perception
cognitive task \Rightarrow perceptive task

identify anomalies, clusters, trends

T. Munzner

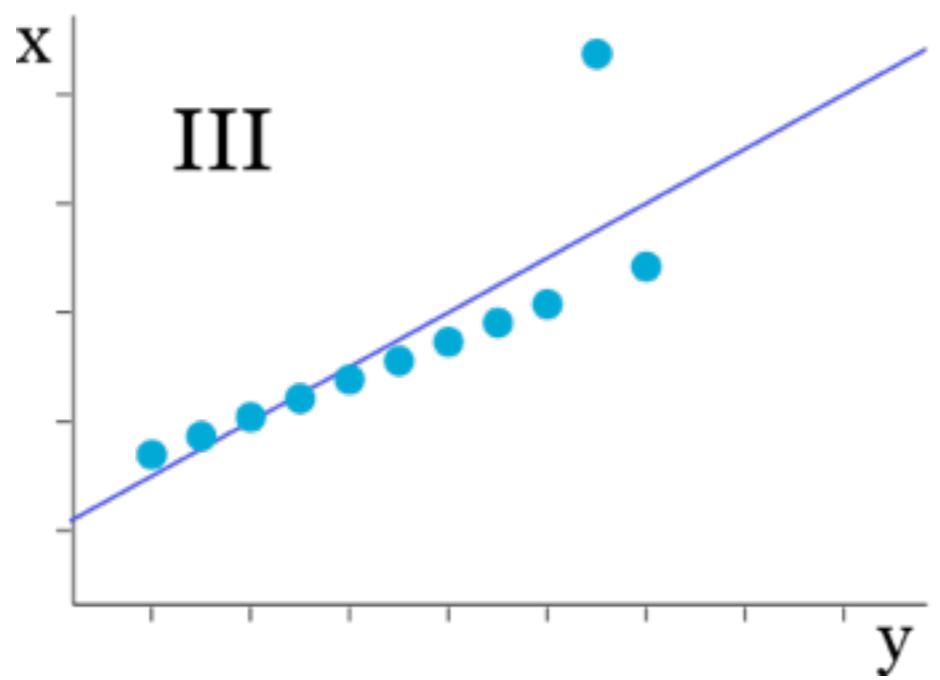
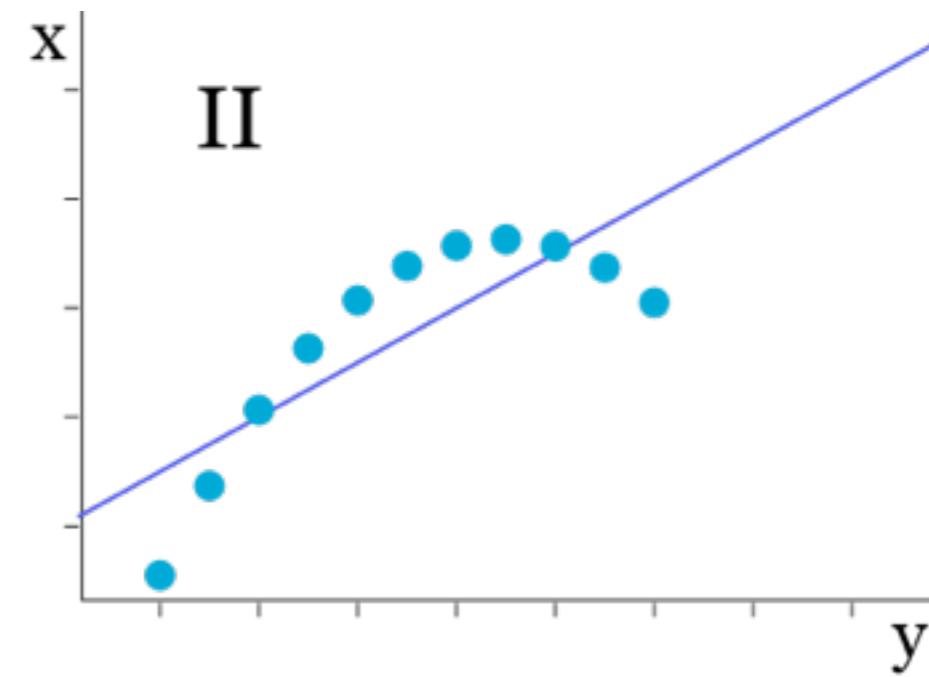
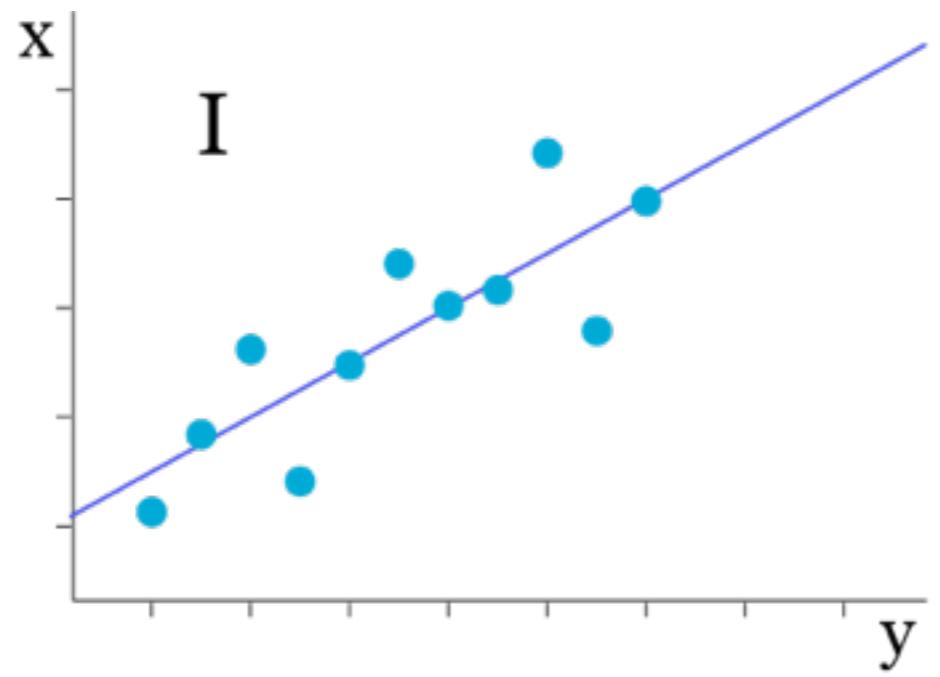
I		II		III	
x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46
8.0	6.95	8.0	8.14	8.0	6.77
13.0	7.58	13.0	8.74	13.0	12.74
9.0	8.81	9.0	8.77	9.0	7.11
11.0	8.33	11.0	9.26	11.0	7.81
14.0	9.96	14.0	8.10	14.0	8.84
6.0	7.24	6.0	6.13	6.0	6.08
4.0	4.26	4.0	3.10	4.0	5.39
12.0	10.84	12.0	9.13	12.0	8.15
7.0	4.82	7.0	7.26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

n = 11

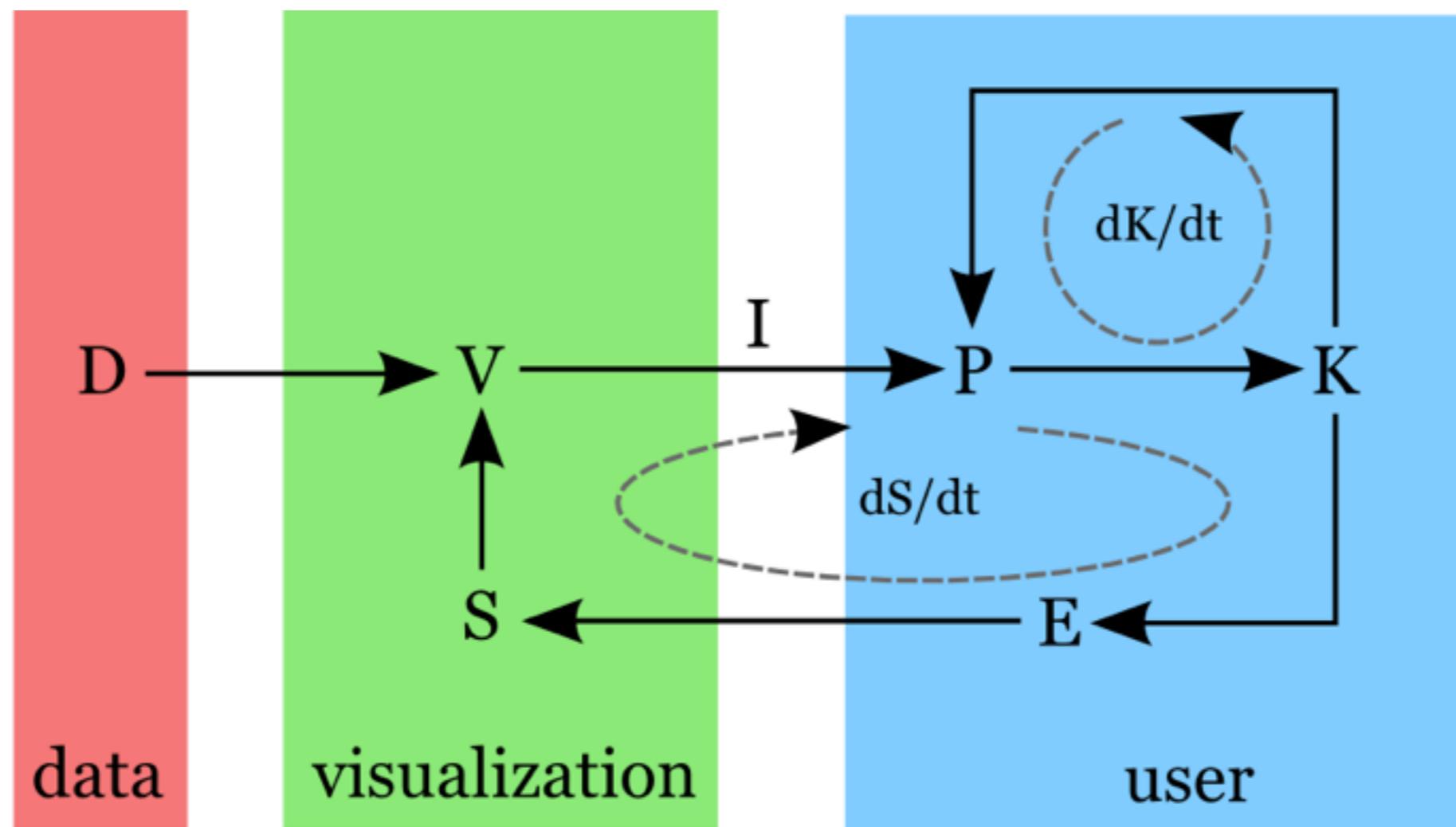
mean x = 9.0
mean y = 7.5

variance x = 11.0
variance y = 4.12

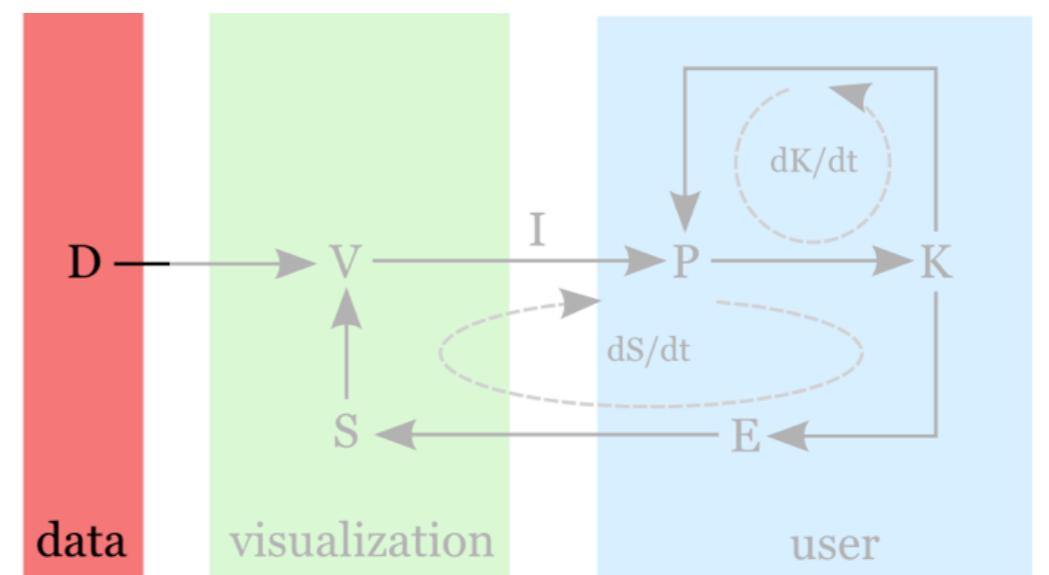
correlation x & y = 0.816
regression line: y = 3+0.5x



Components



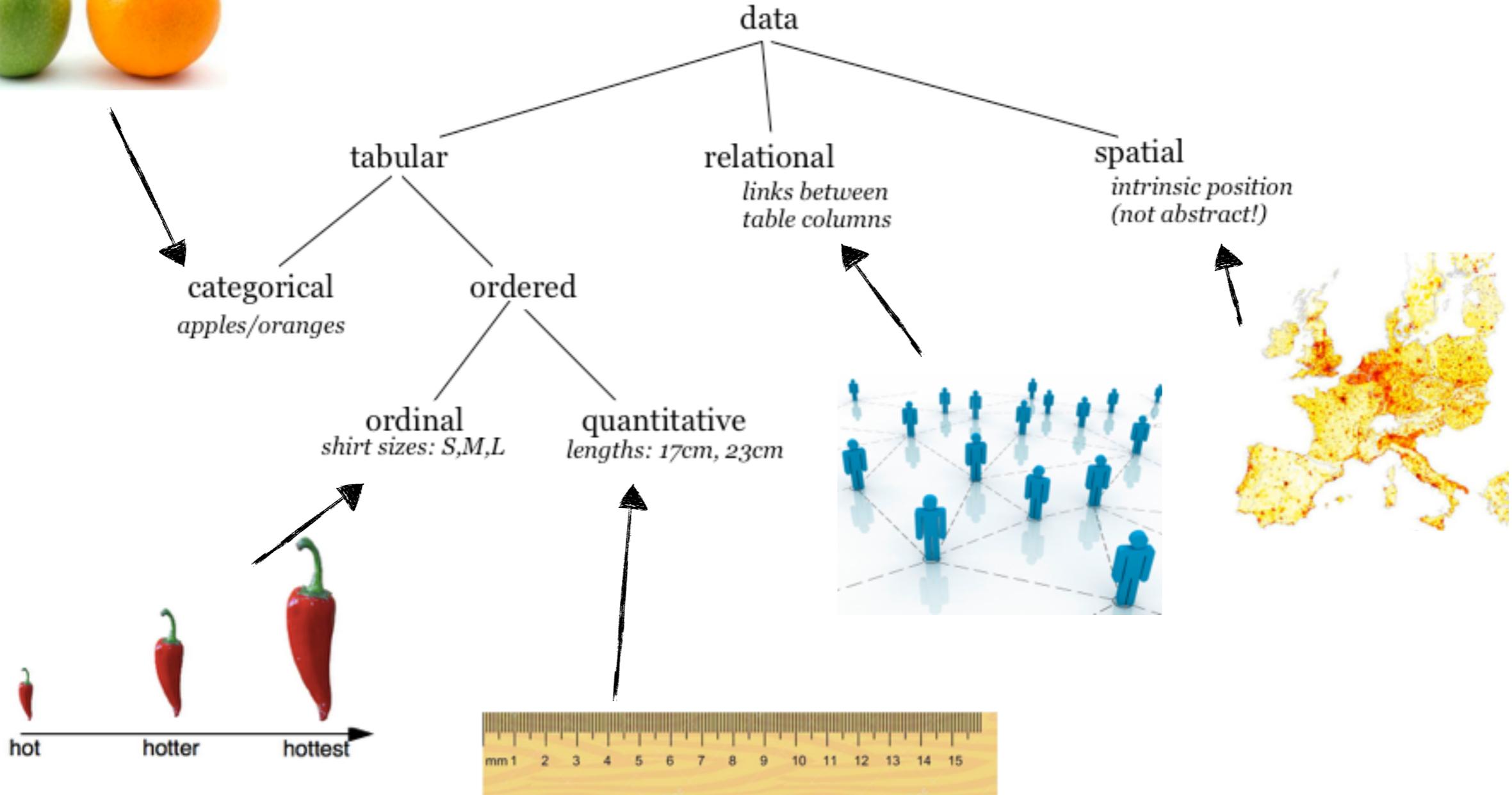
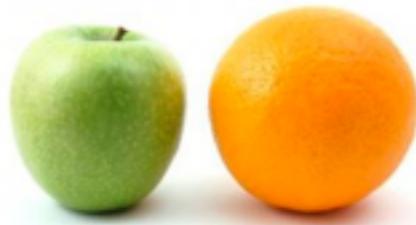
Data foundations



data attributes
dimensions
features
properties

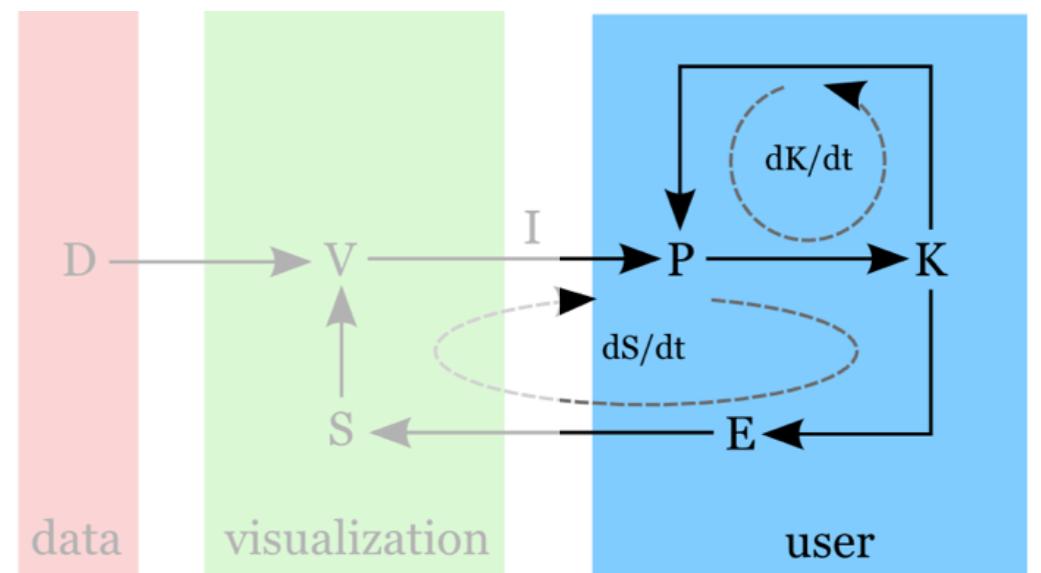
	A	B	C	D	E	F	G
1	from_airport	from_city	from_country	from_long	from_lat	to_airport	to_c
2	Balandino	Chelyabinsk	Russia	61.838	55.509	Domododevo	Mos
3	Balandino	Chelyabinsk	Russia	61.838	55.509	Kazan	Kaza
4	Balandino	Chelyabinsk	Russia	61.838	55.509	Tolmachevo	Nov
5	Domododevo	Moscow	Russia	38.51	55.681	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	Mos
13	Kazan	Kazan	Russia	49.464	56.01	Balandino	Chel
14	Kazan	Kazan	Russia	49.464	56.01	Domododevo	Mos
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	Buqulma Airport	Buqi

data items
datapoints



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

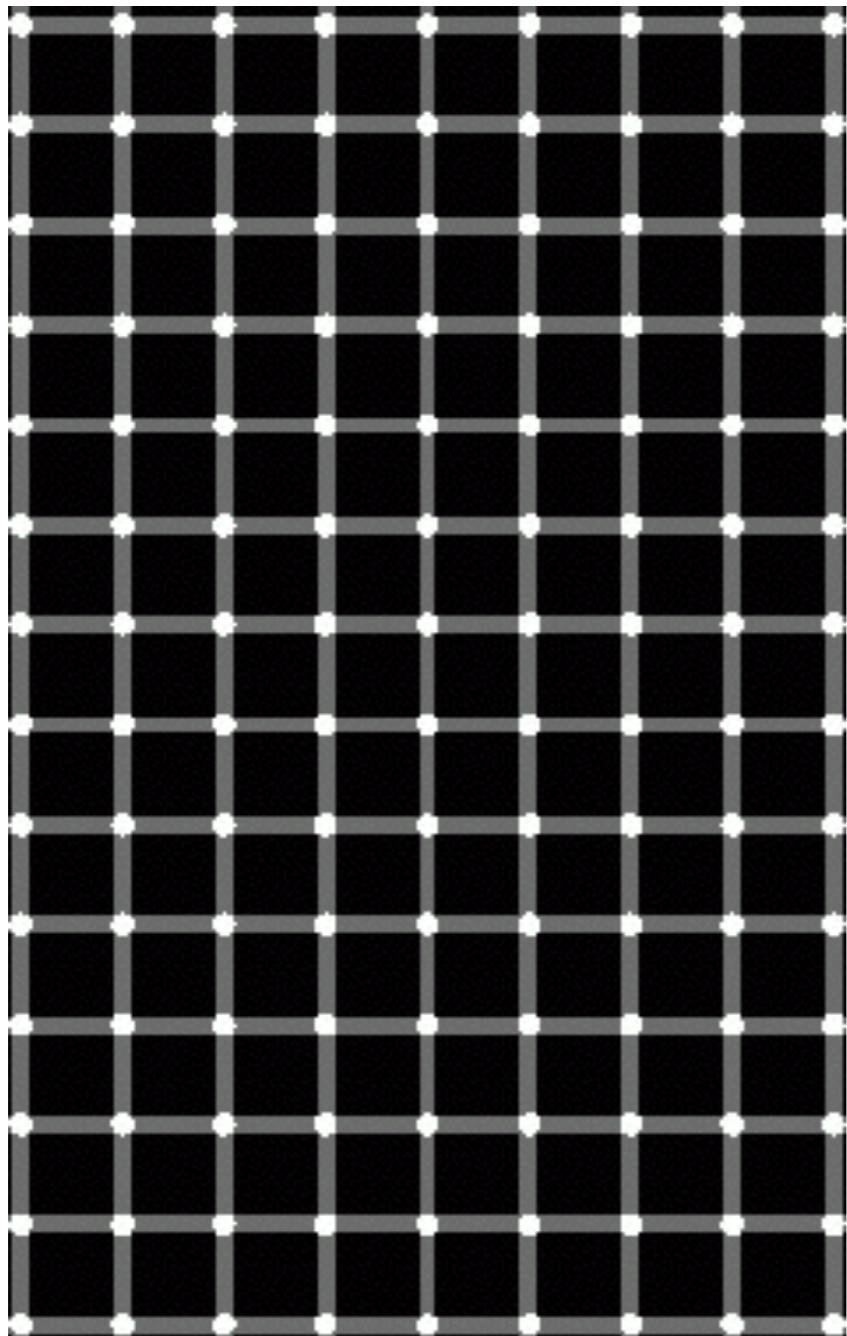
Human perception foundations



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:

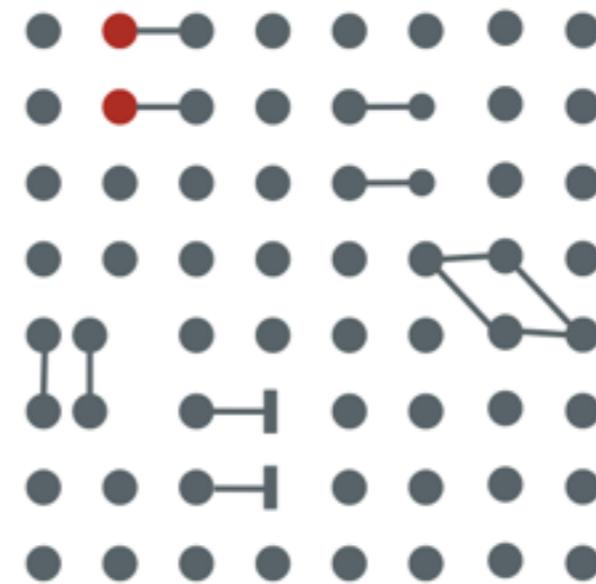
- black = Bush
- white = Gore

Hermann grid illusion

Simplicity



Connectedness



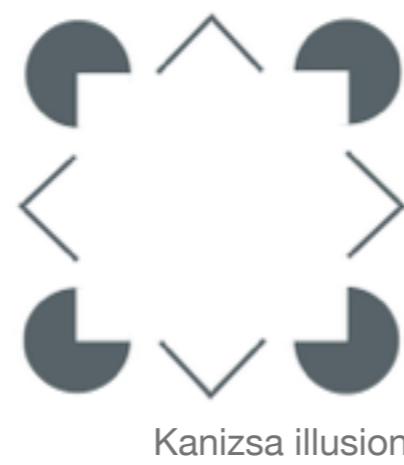
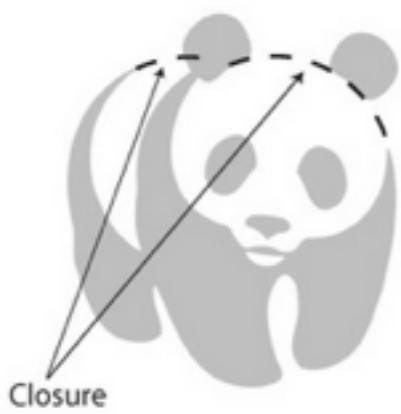
Proximity



Similarity



Closure

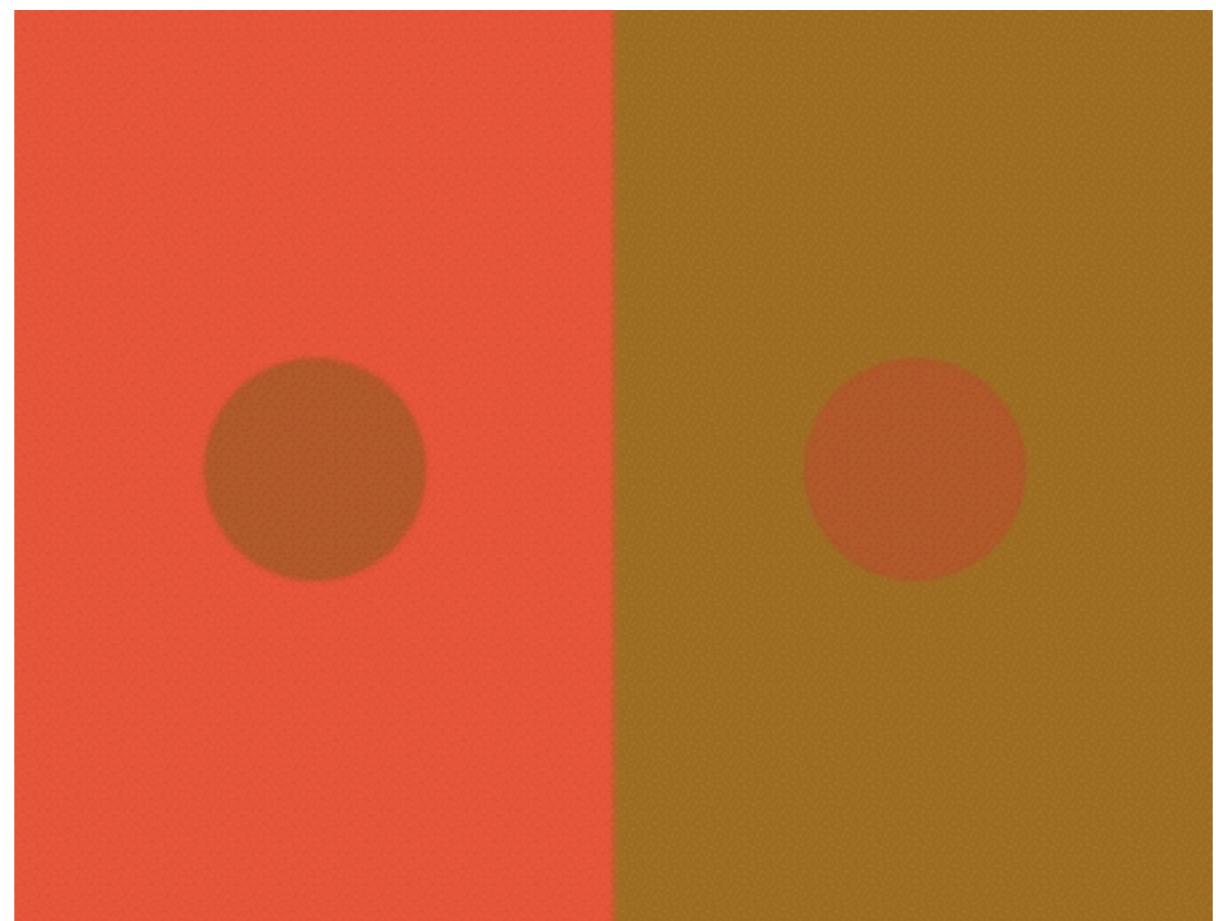


Kanizsa illusion

Symmetry, Familiarity, Continuity
Figure and Ground

A B C

I2 I3 I4



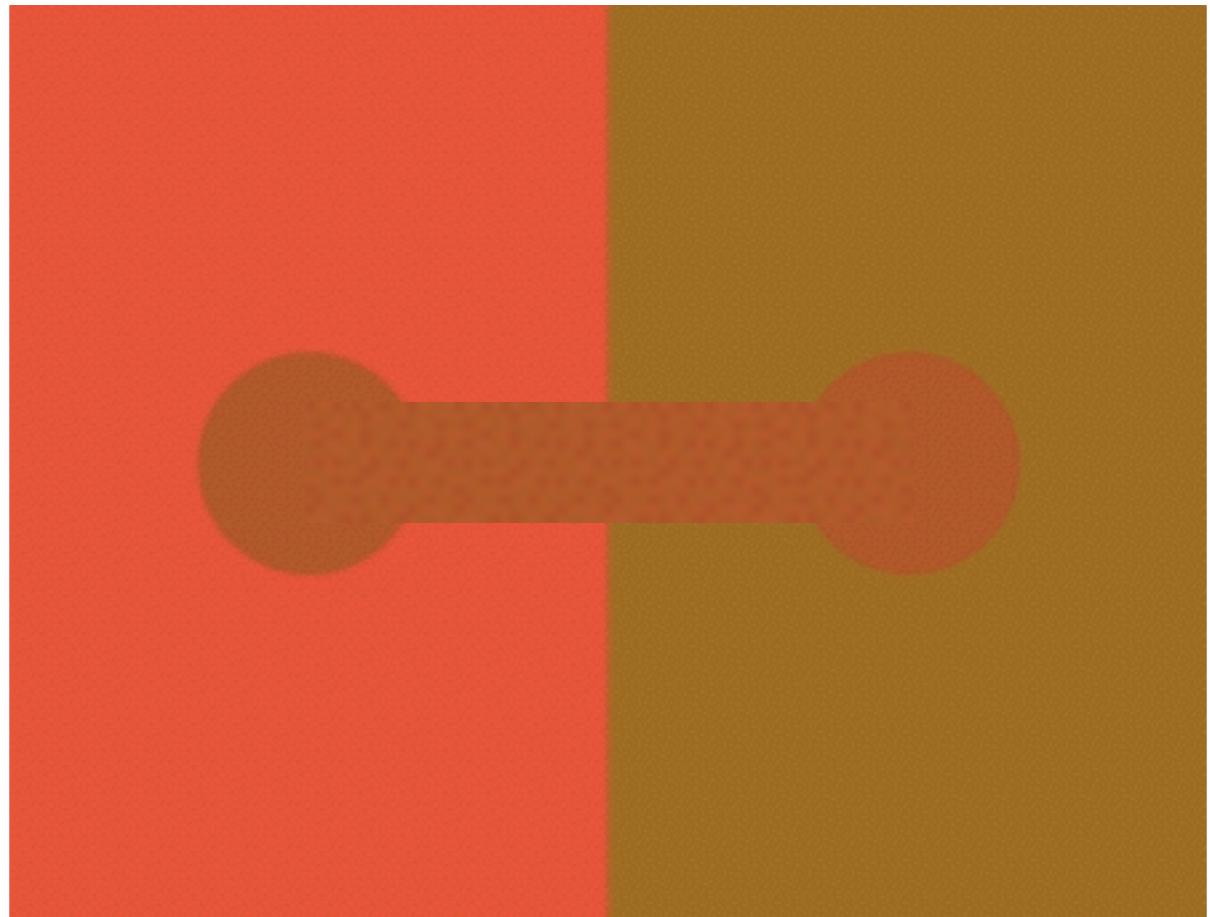




Figure and Ground



Pre-attentive vision

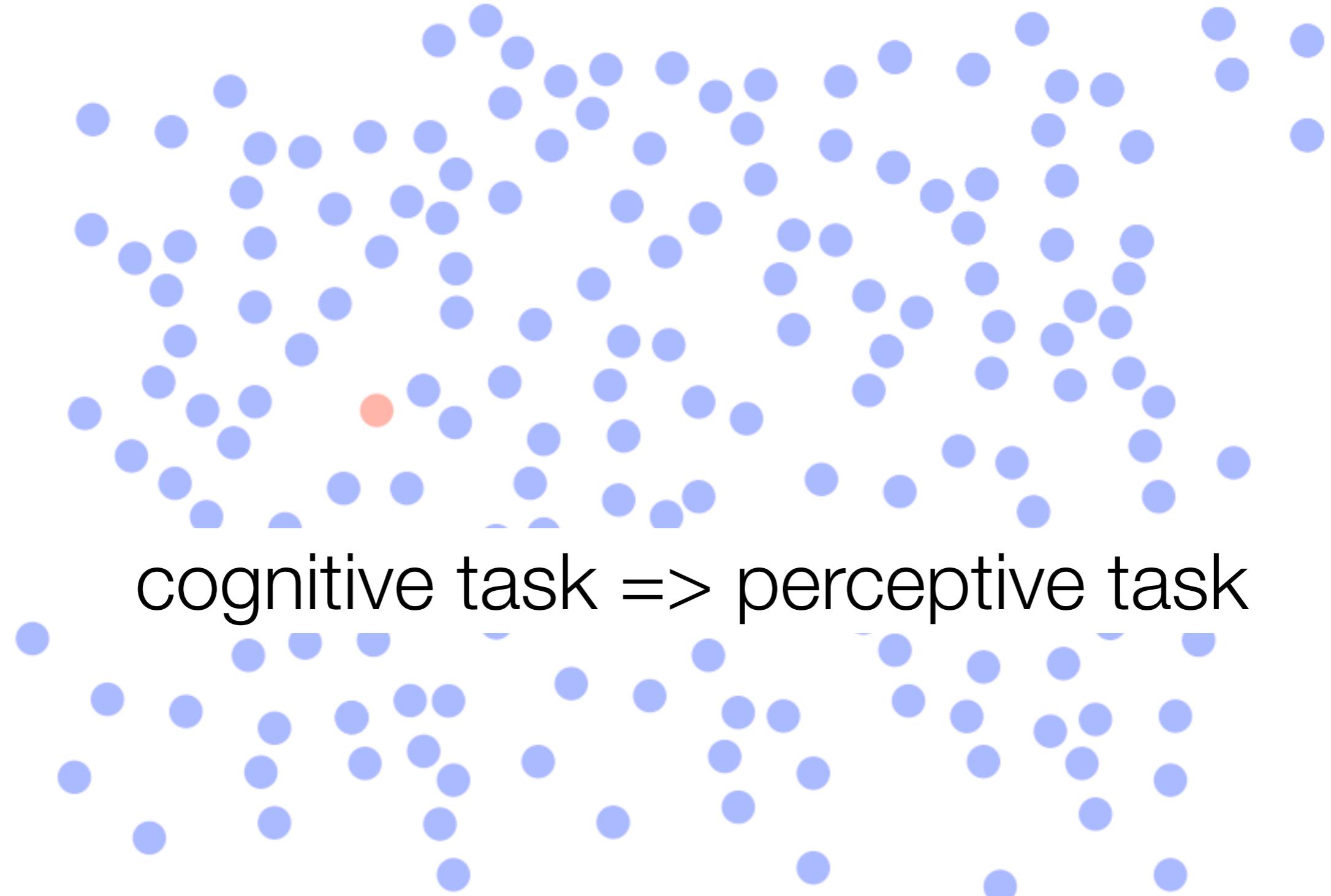
= ability of low-level human visual system to rapidly identify certain basic visual properties

- 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

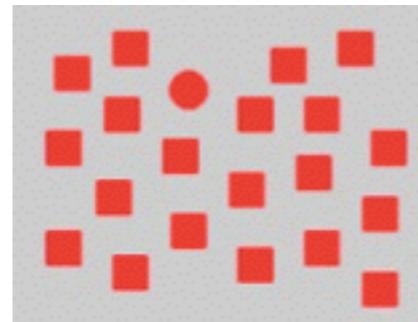
5

5



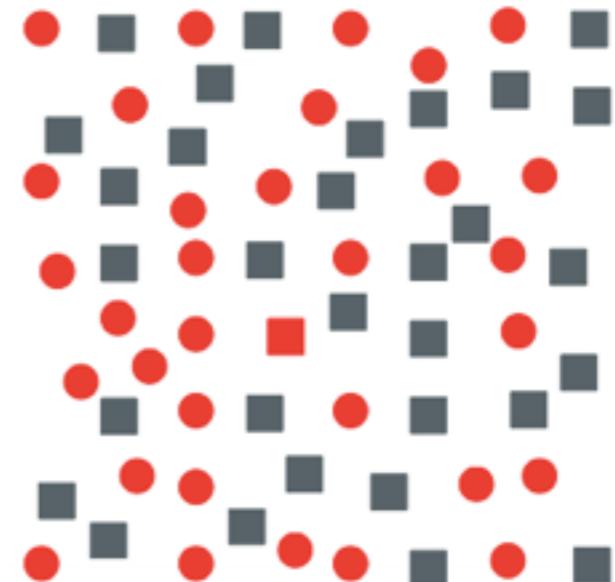
Limitations of preattentive vision

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



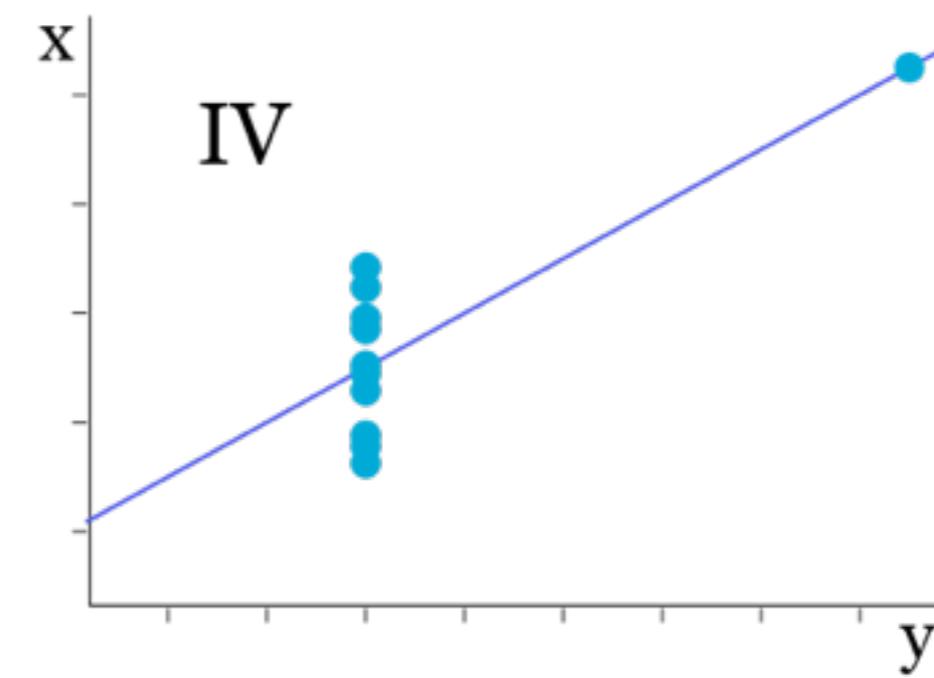
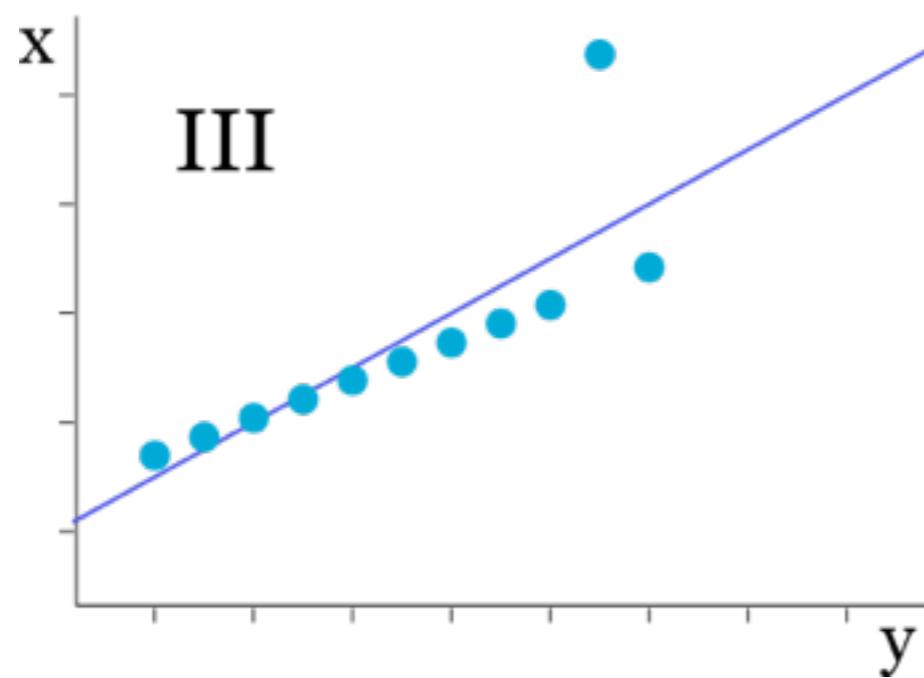
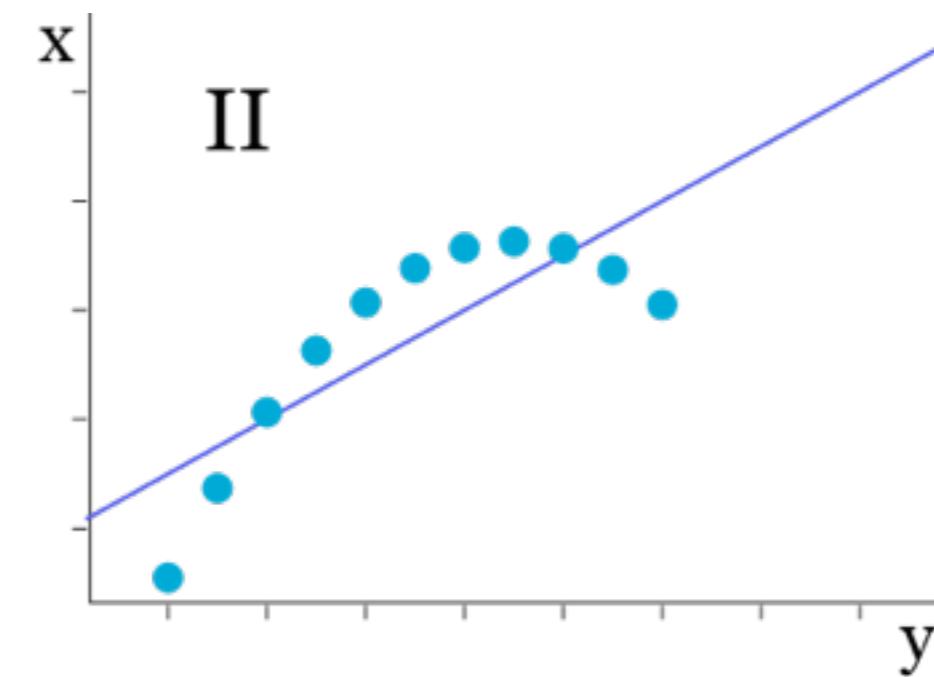
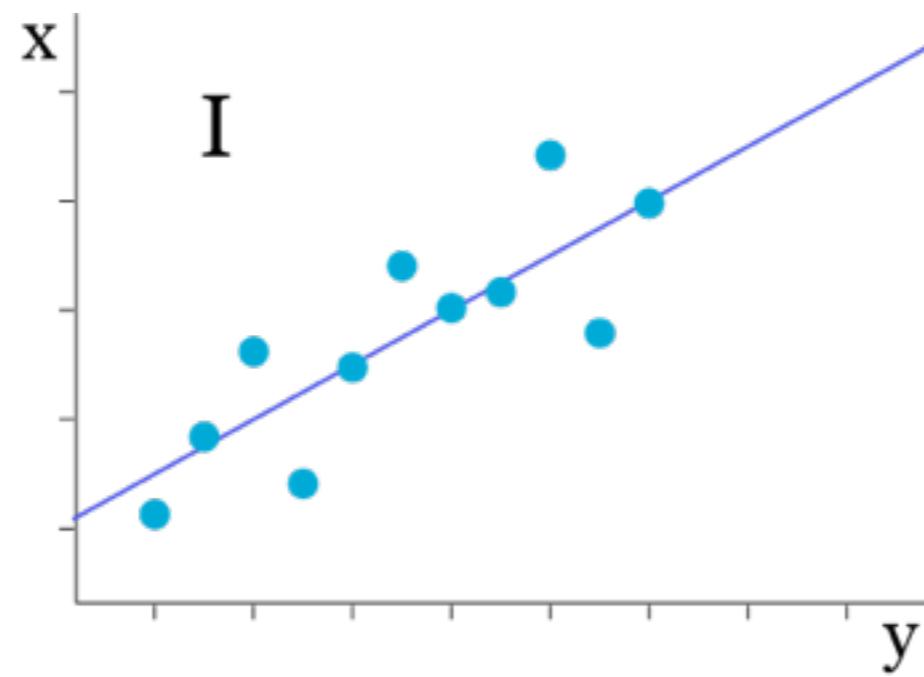
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

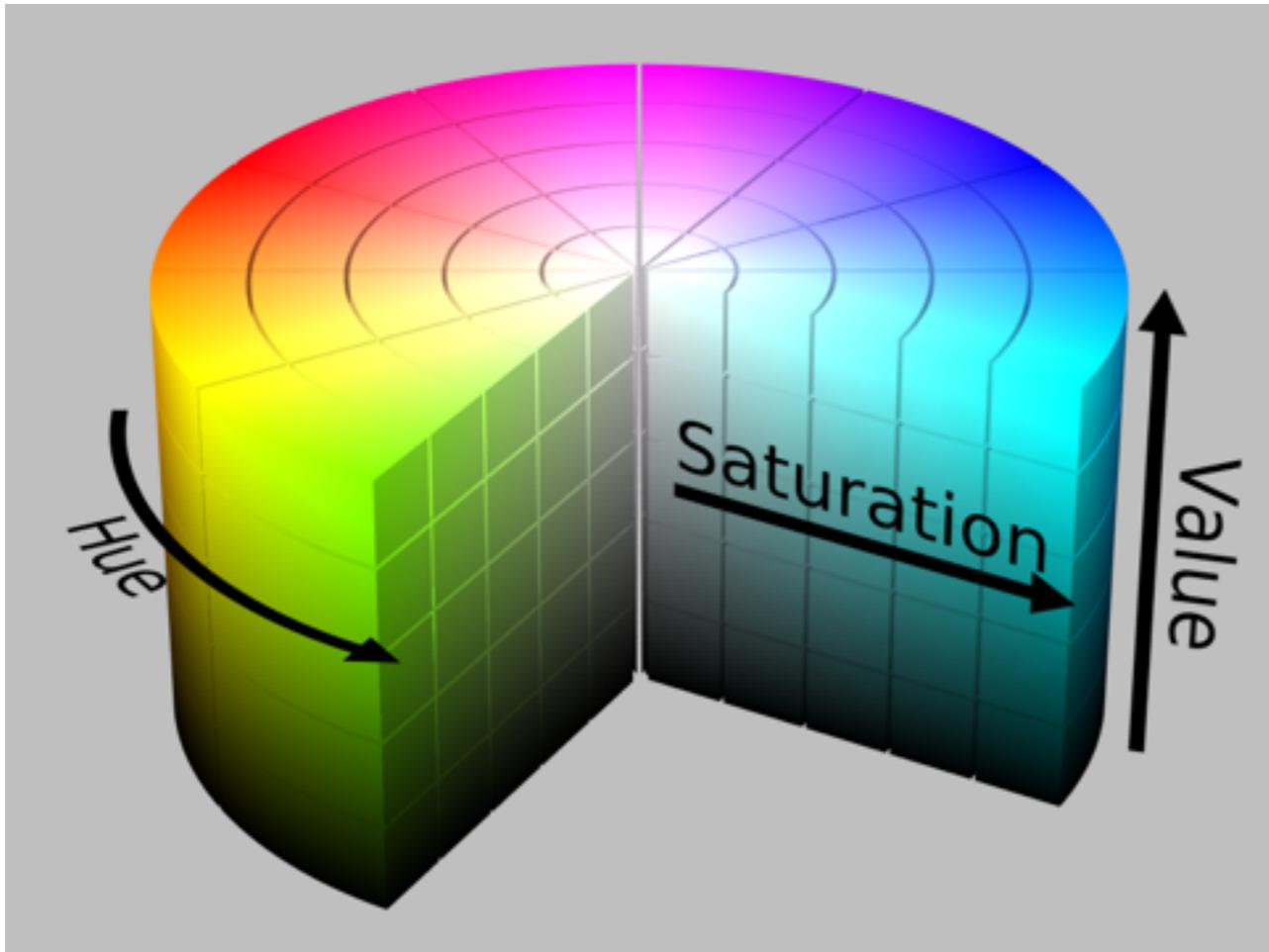
mean x = 9.0
mean y = 7.5

variance x = 11.0
variance y = 4.12

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



About colour



Luminance



Saturation



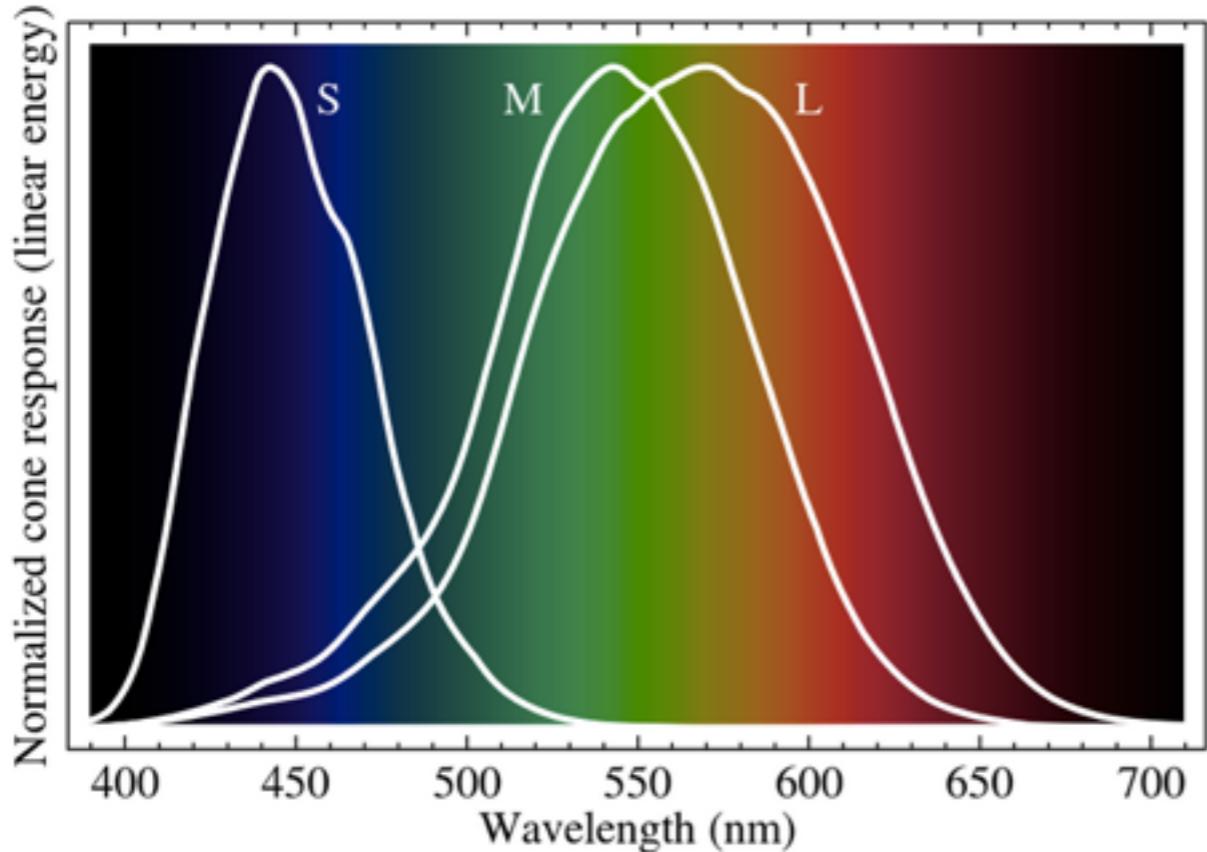
Hue



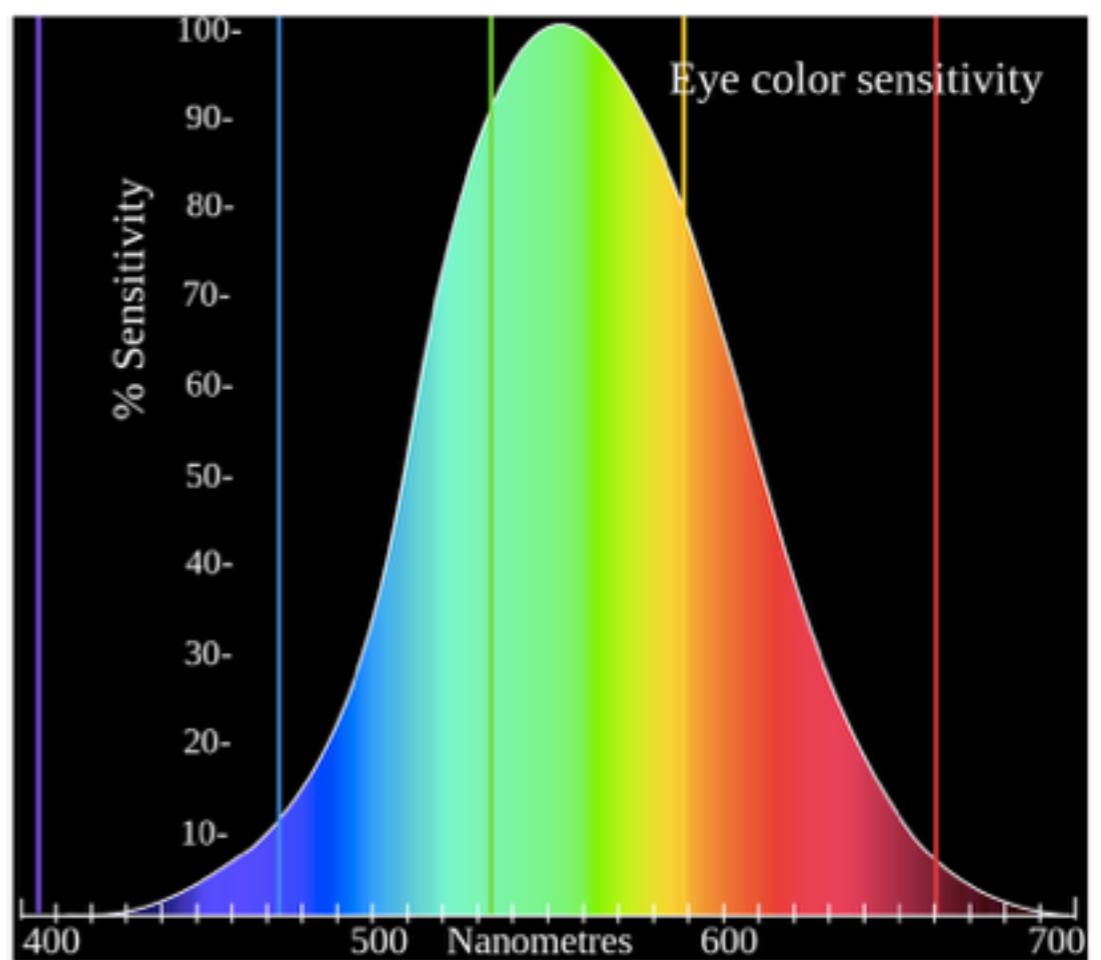
Use **HSV** (hue - saturation - value) instead of RGB

HSV: separates *luma* (intensity) from *chroma* (colour)

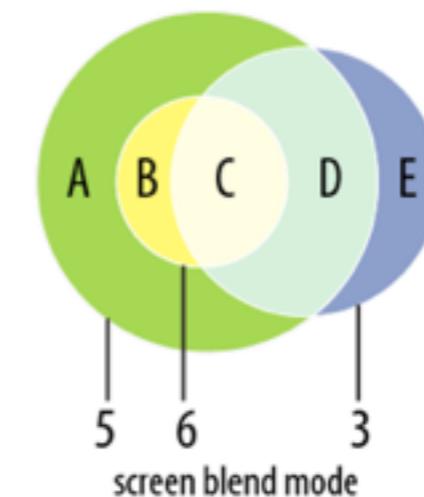
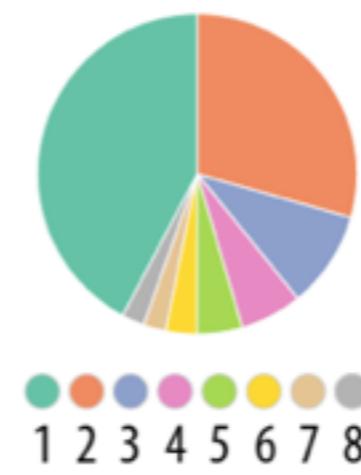
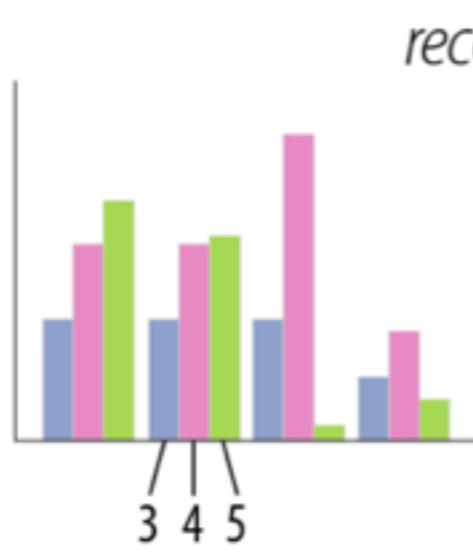
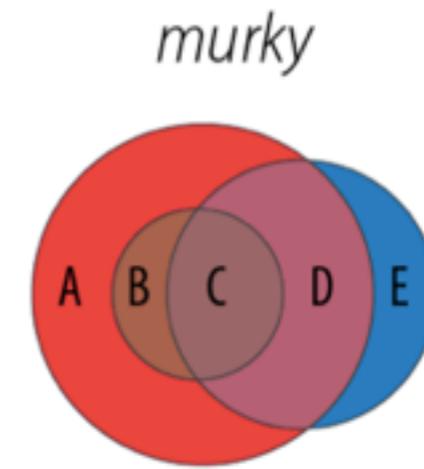
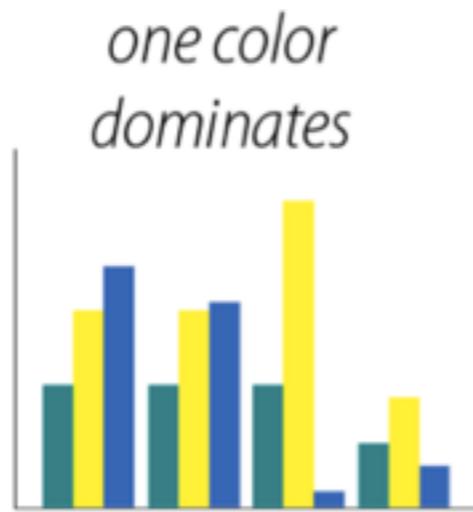
RGB has to do with *implementation*; HSV has to do with *perception*



sensitivity of S/M/L cones to colour wavelengths



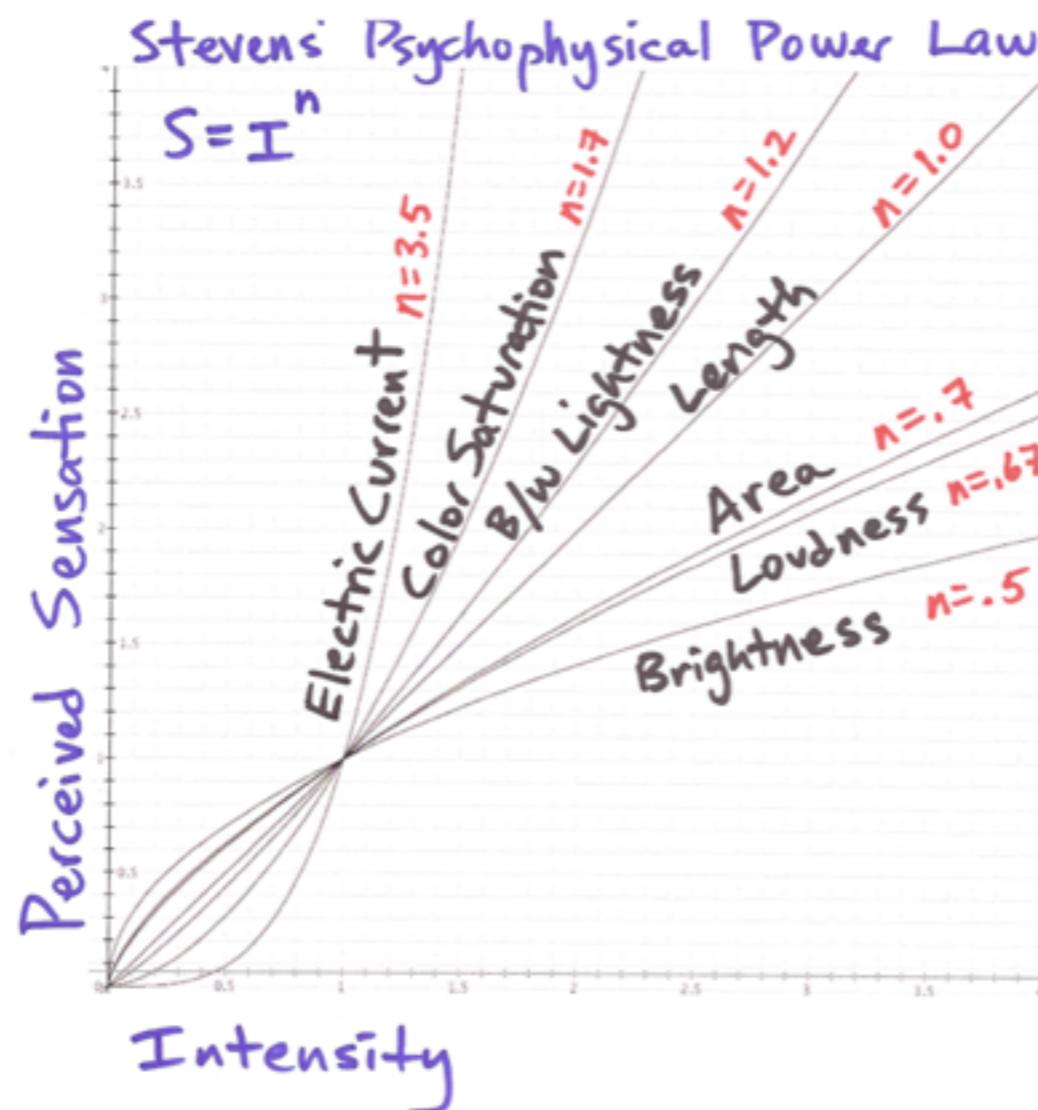
colorbrewer2.org - Never pick your own colour



in R: please use RColorBrewer!

Steven's psychophysical law

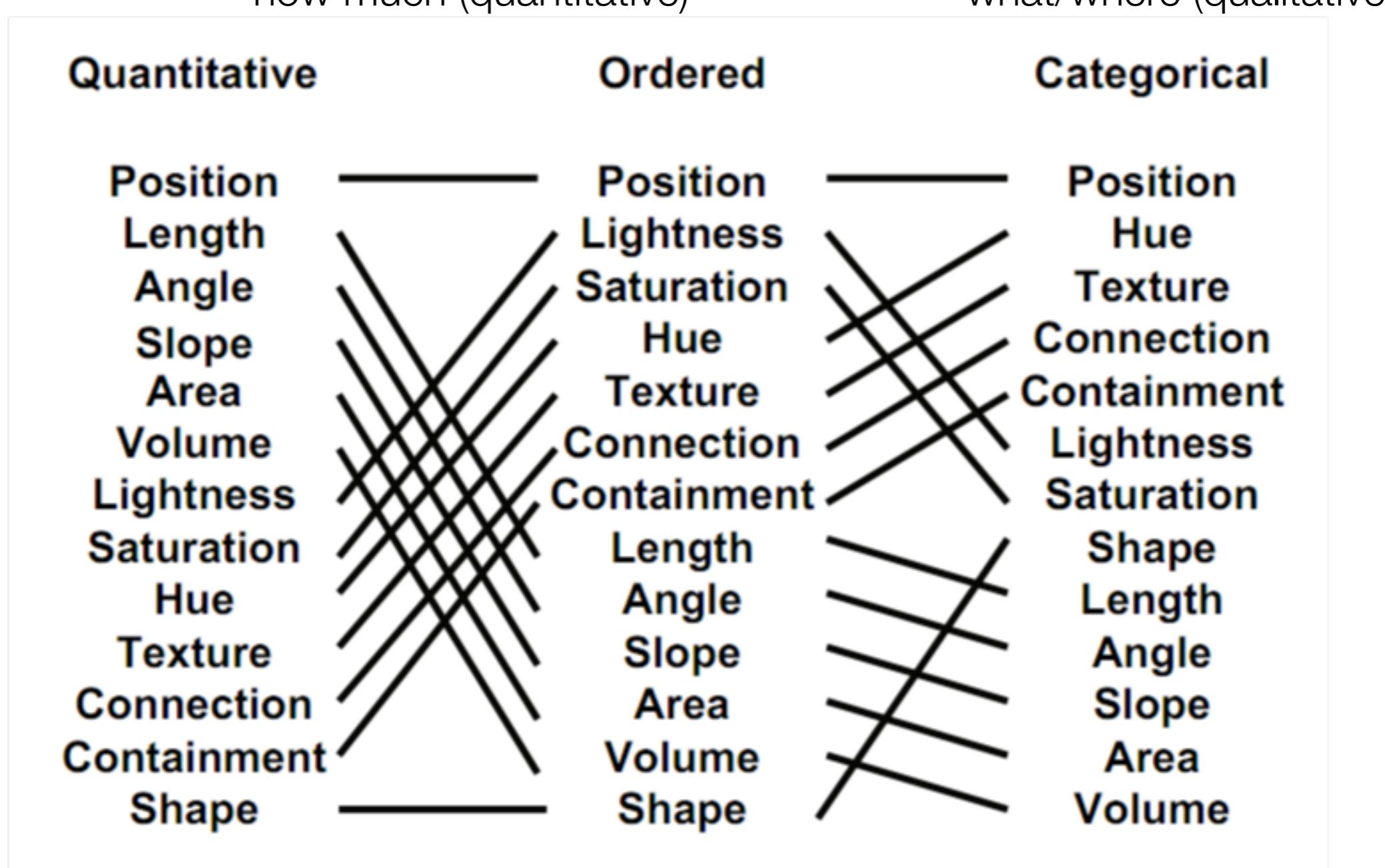
= proposed relationship between the magnitude of a physical stimulus and its perceived intensity or strength



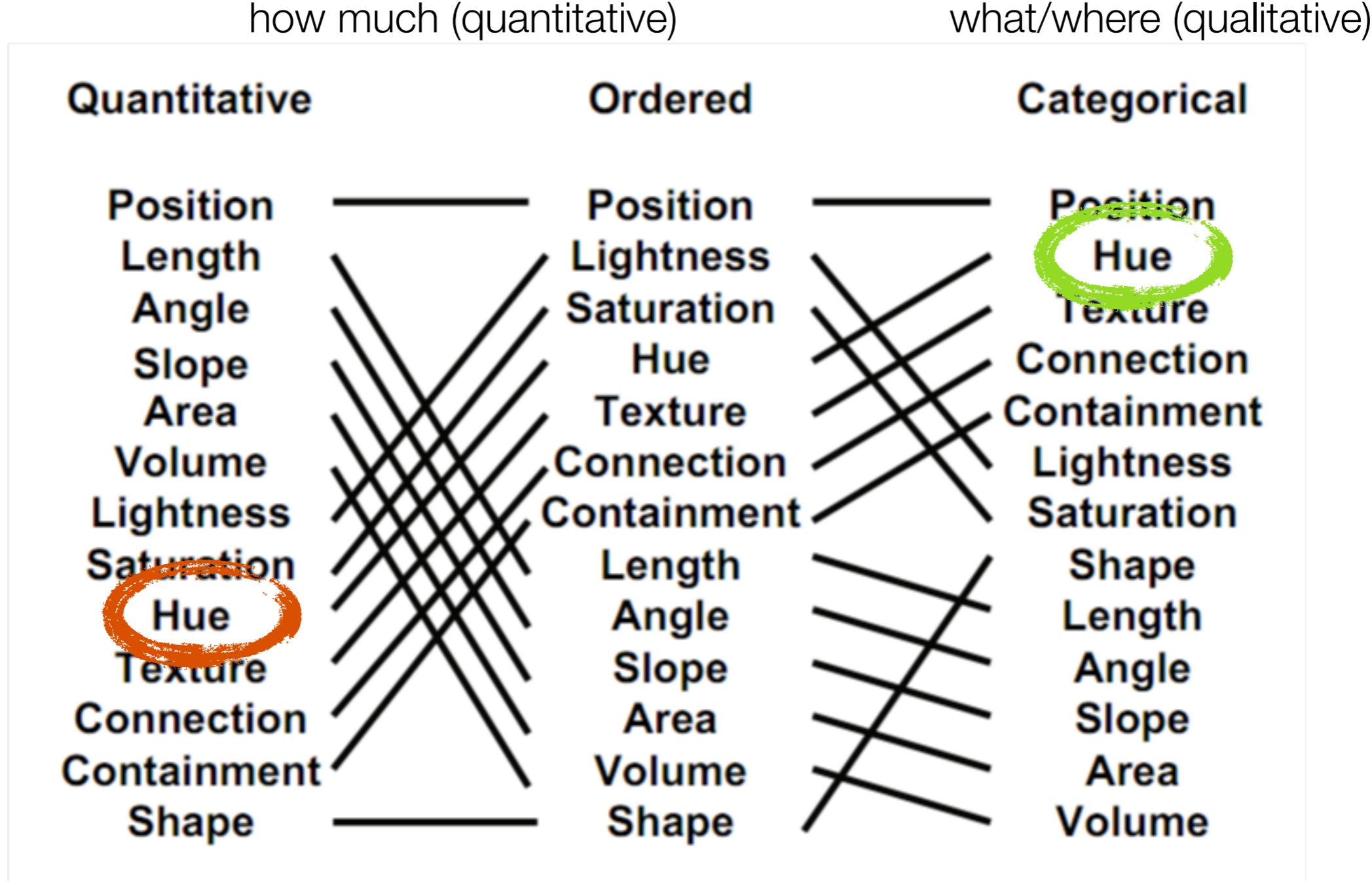
Semiology of graphics

	Points	Lines	Areas	Best to show
Shape		possible, but too weird to show	cartogram	qualitative differences
Size			cartogram	quantitative differences
Color Hue				qualitative differences
Color Value				quantitative differences
Color Intensity				qualitative differences
Texture				qualitative & quantitative differences

Accuracy of quantitative perceptual tasks

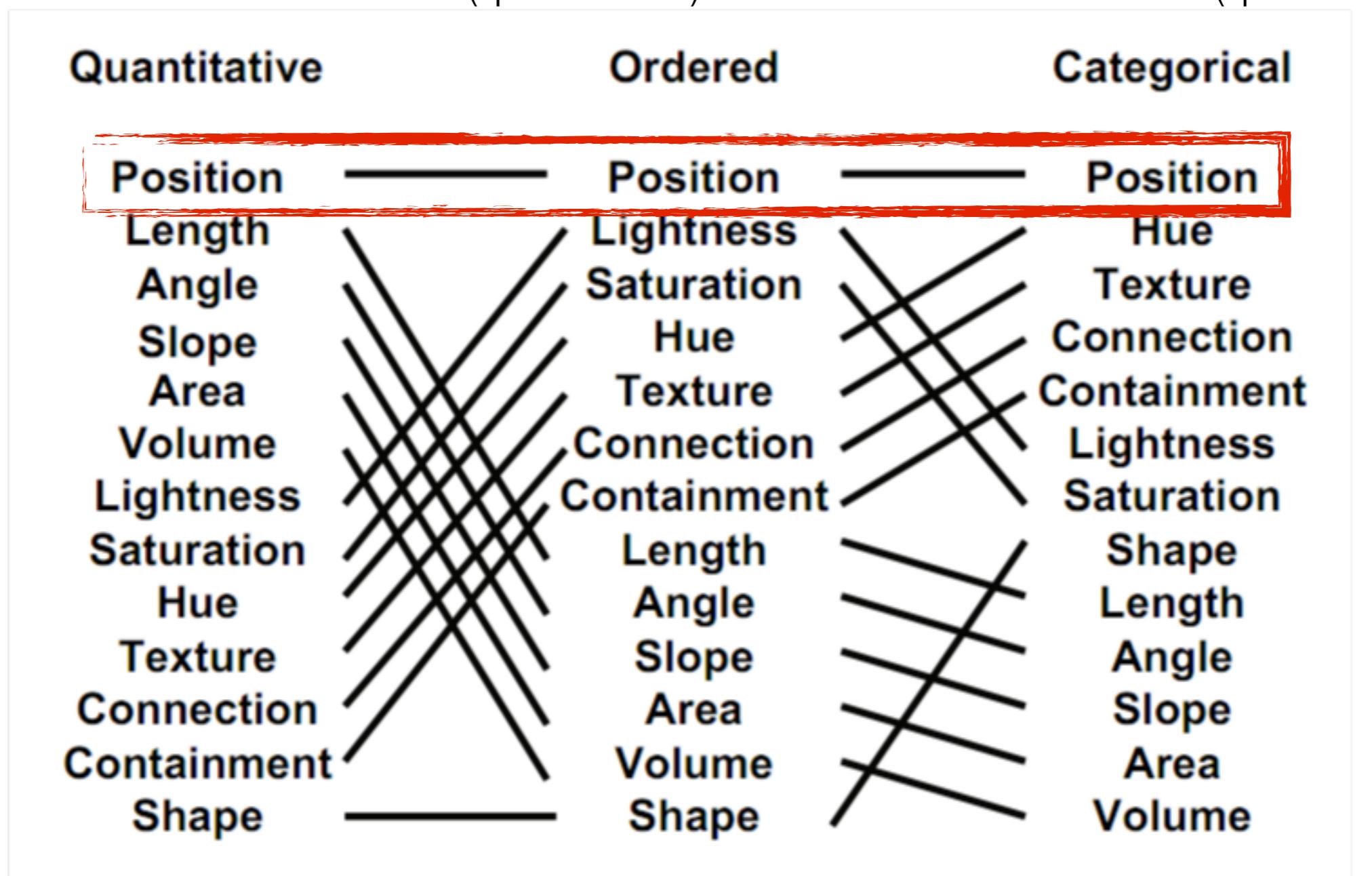


Accuracy of quantitative perceptual tasks



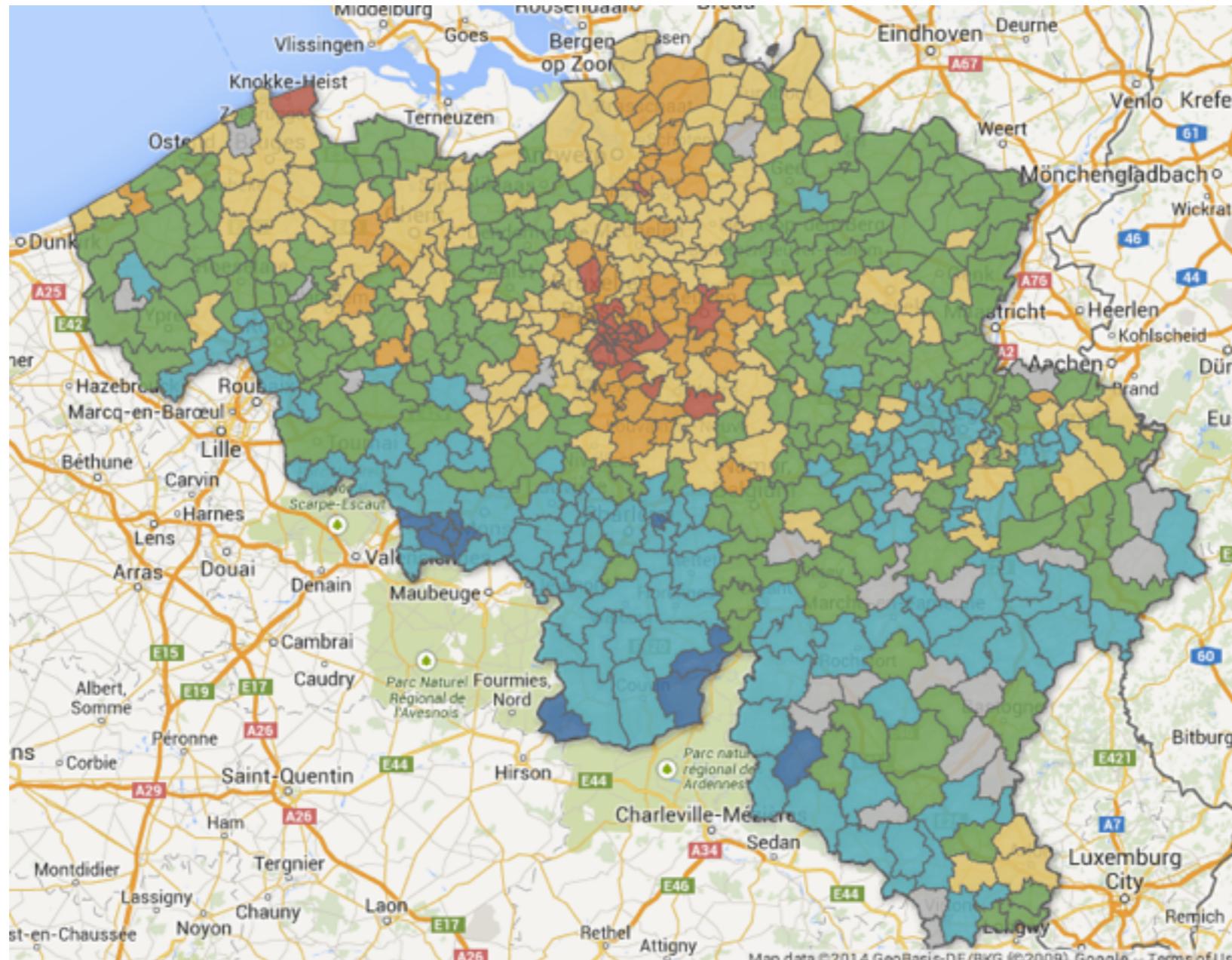
McKinlay

Accuracy of quantitative perceptual tasks



“power of the plane”

McKinlay



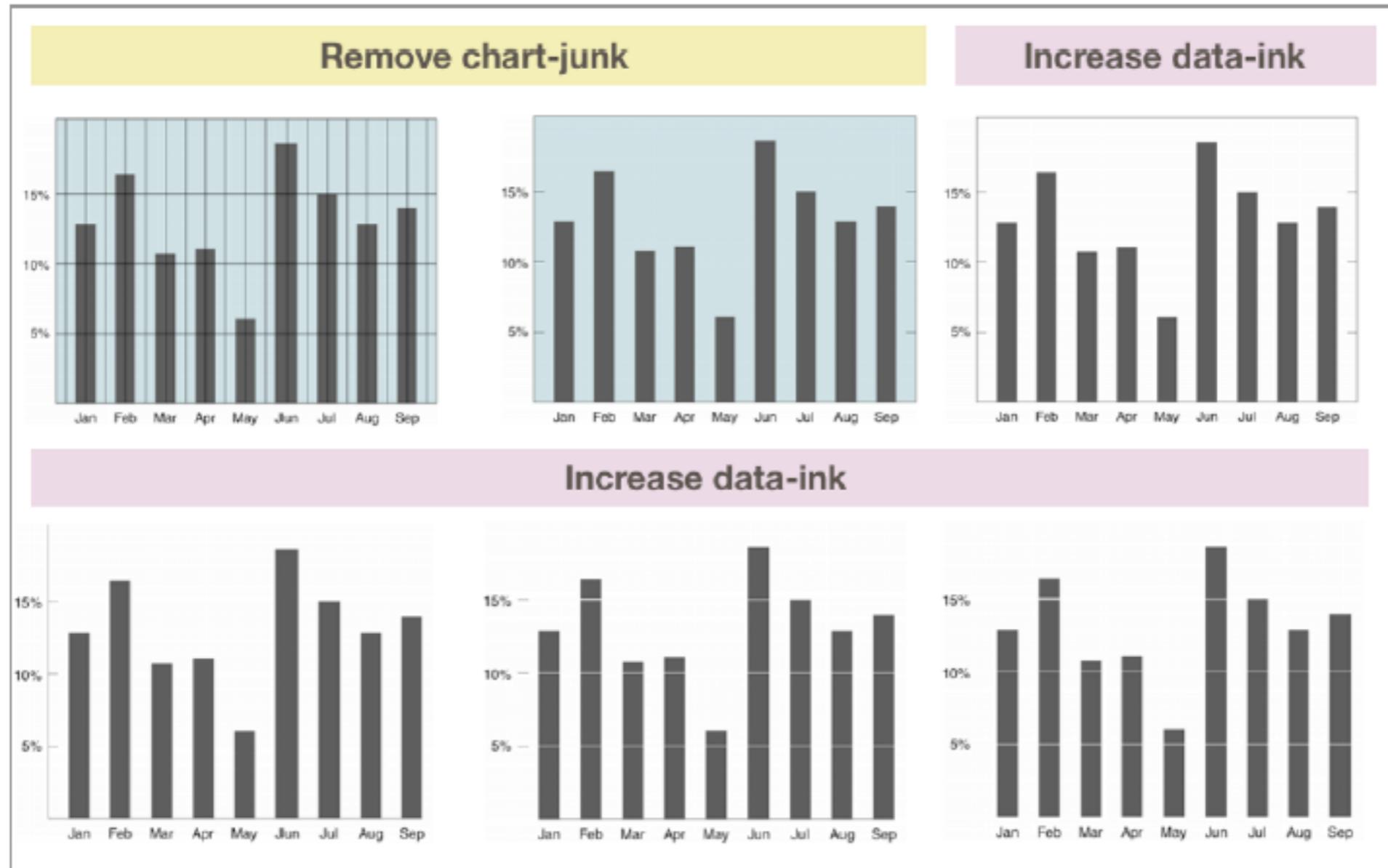
average prices per
municipality in Belgium
(De Standaard, 24/8/2014)

Rules of thumb

(partly based on book Munzner)

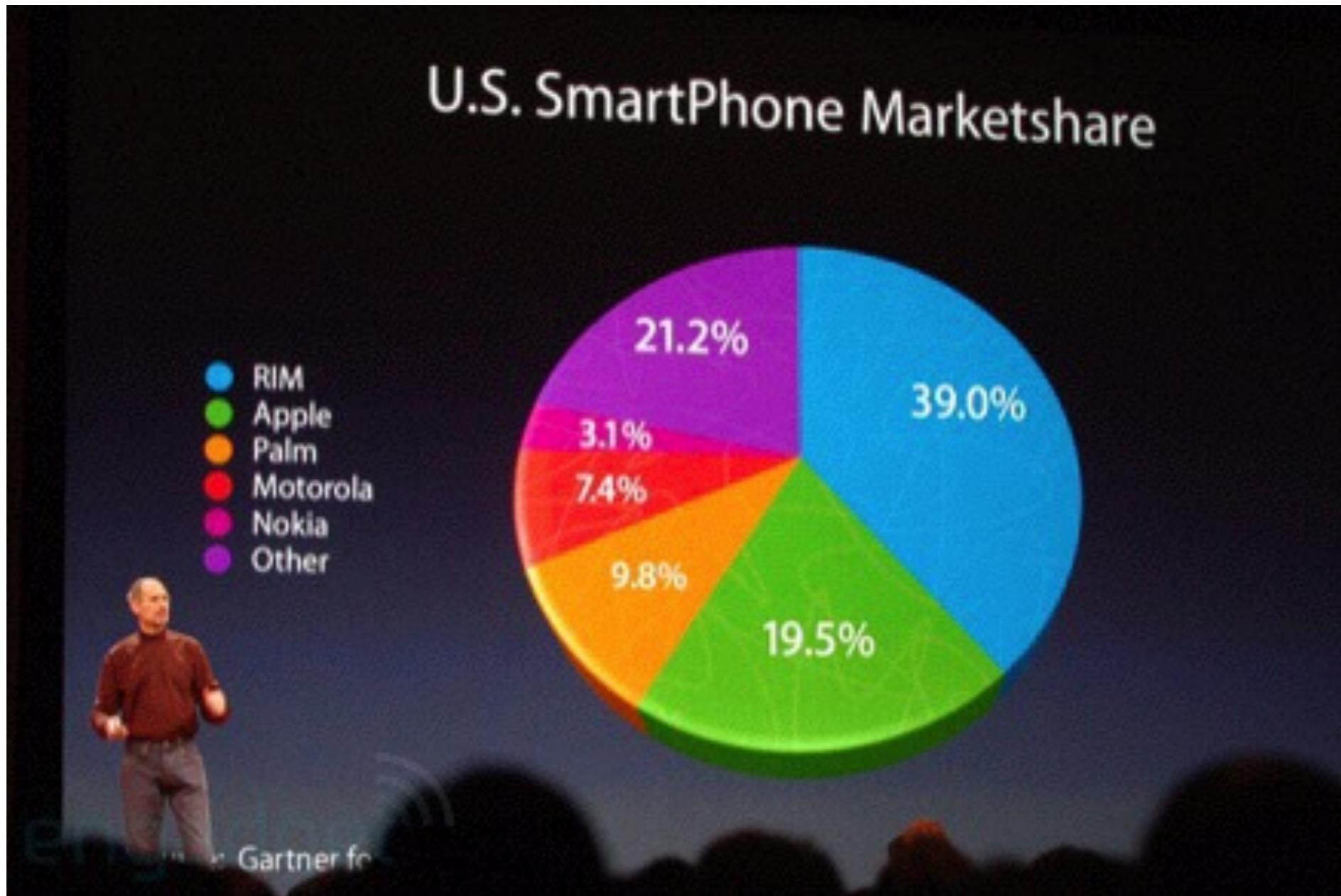
1. maximize data-to-ink ratio (Edward Tufte)
2. beware of the lie-factor
3. no unjustified 3D
4. eyes beat memory
5. focus + context
6. overview first, zoom & filter, details on demand
7. don't overengineer

1. maximize data-to-ink ratio

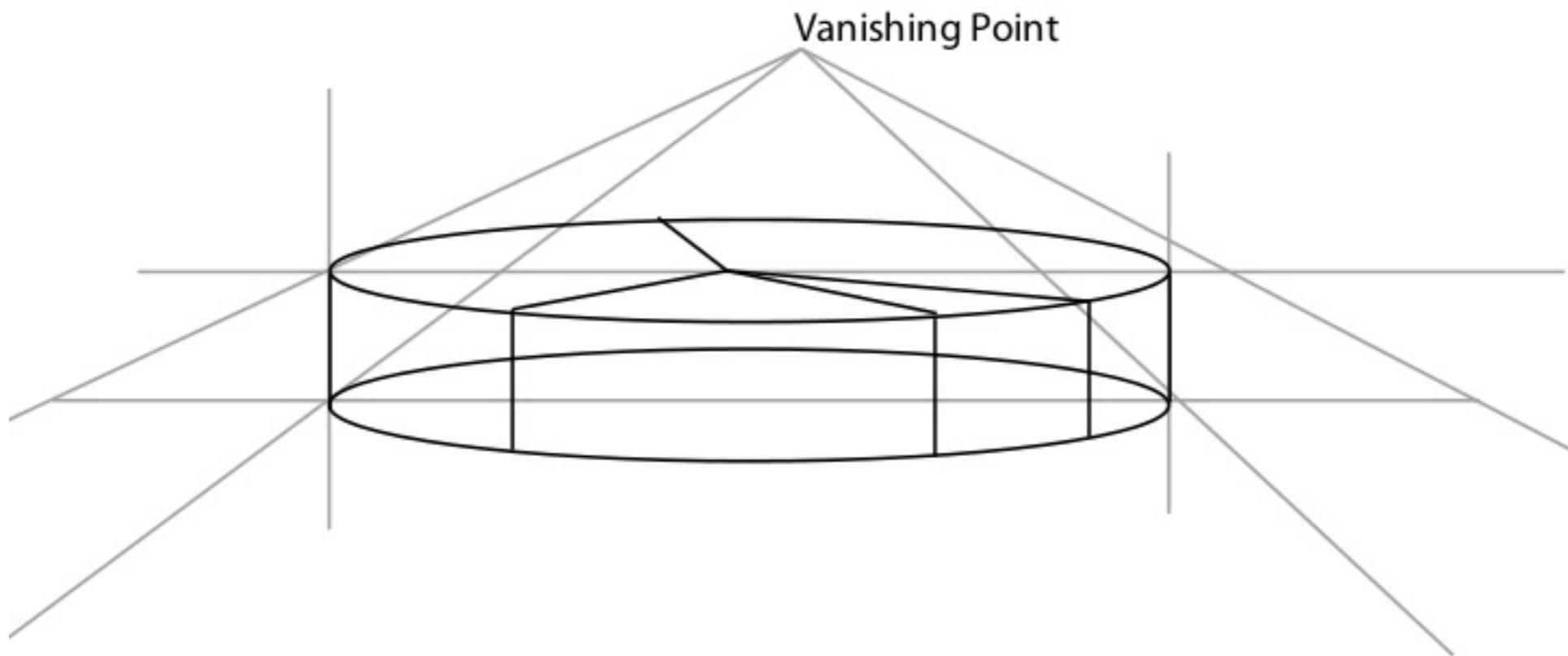


“data-to-ink ratio” = $\frac{\text{data-ink}}{\text{total ink}}$ = 1 - proportion of graphic that can be erased

2. beware of the lie-factor



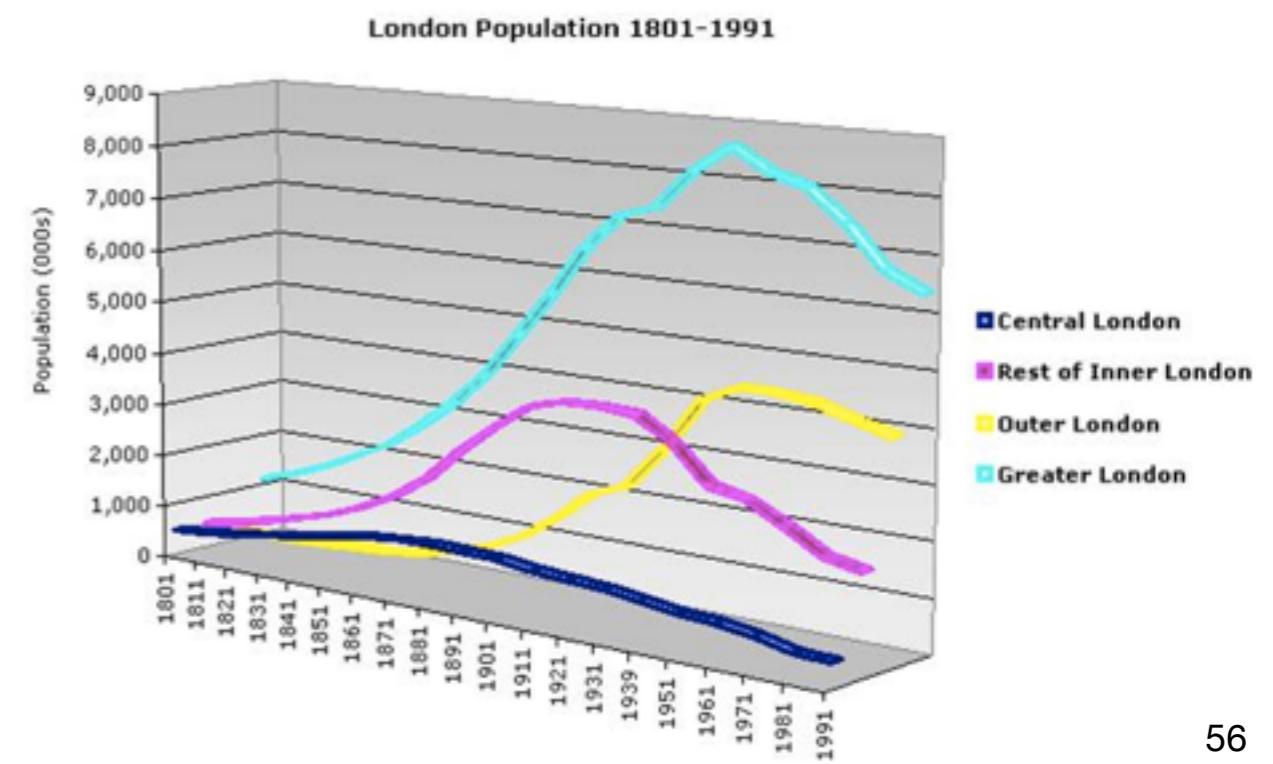
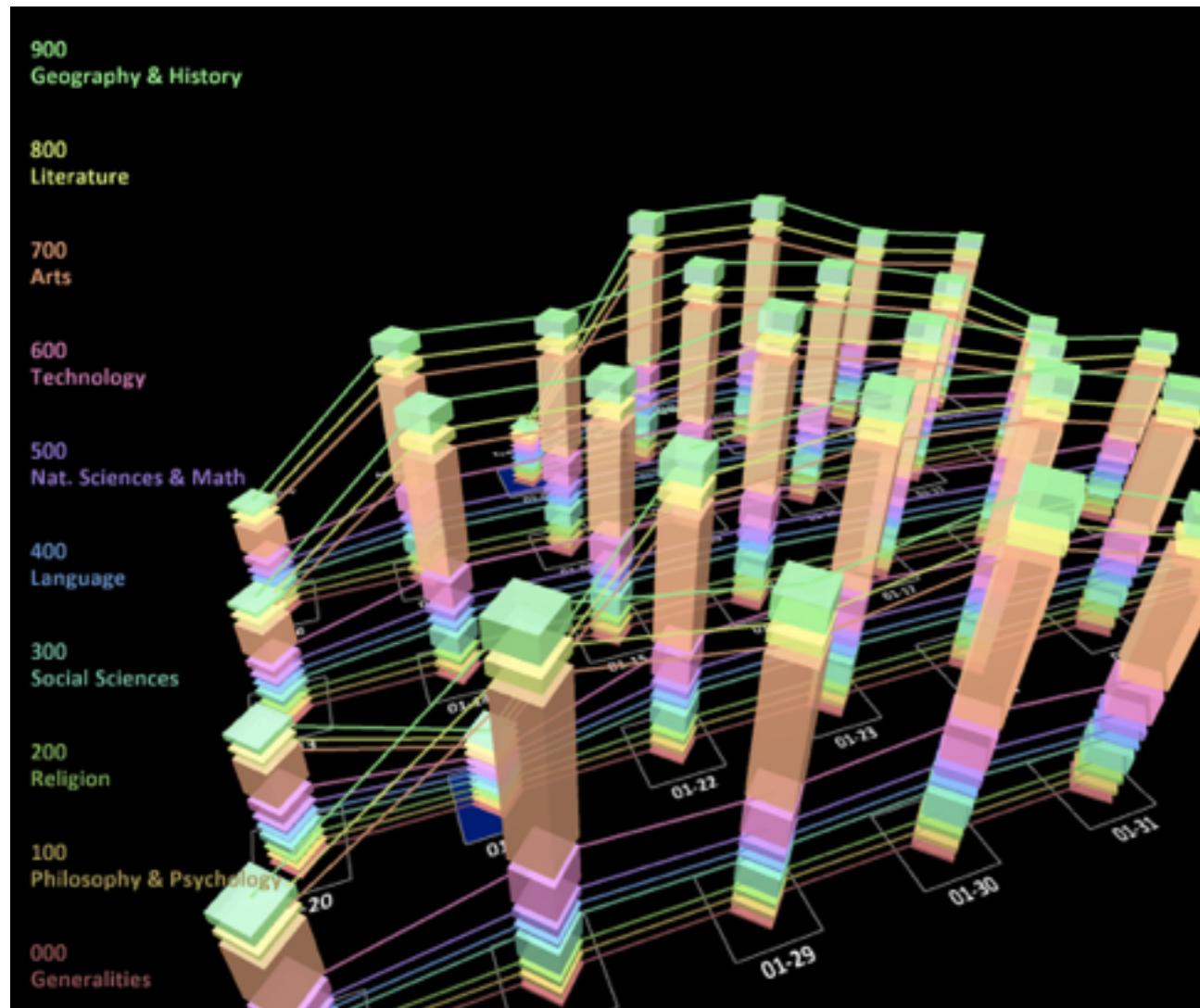
3D Charts!



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

3. no unjustified 3D

issues with occlusion, perspective distortion, text legibility, ...

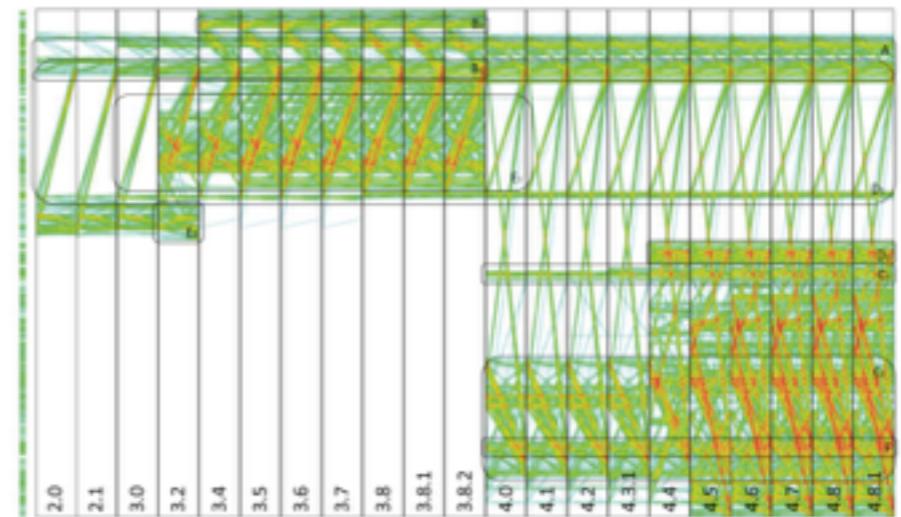


4. eyes beat memory

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 static way



5. focus & context

Show selected regions in greater detail (focus)

Preserve global view at reduced detail (context)

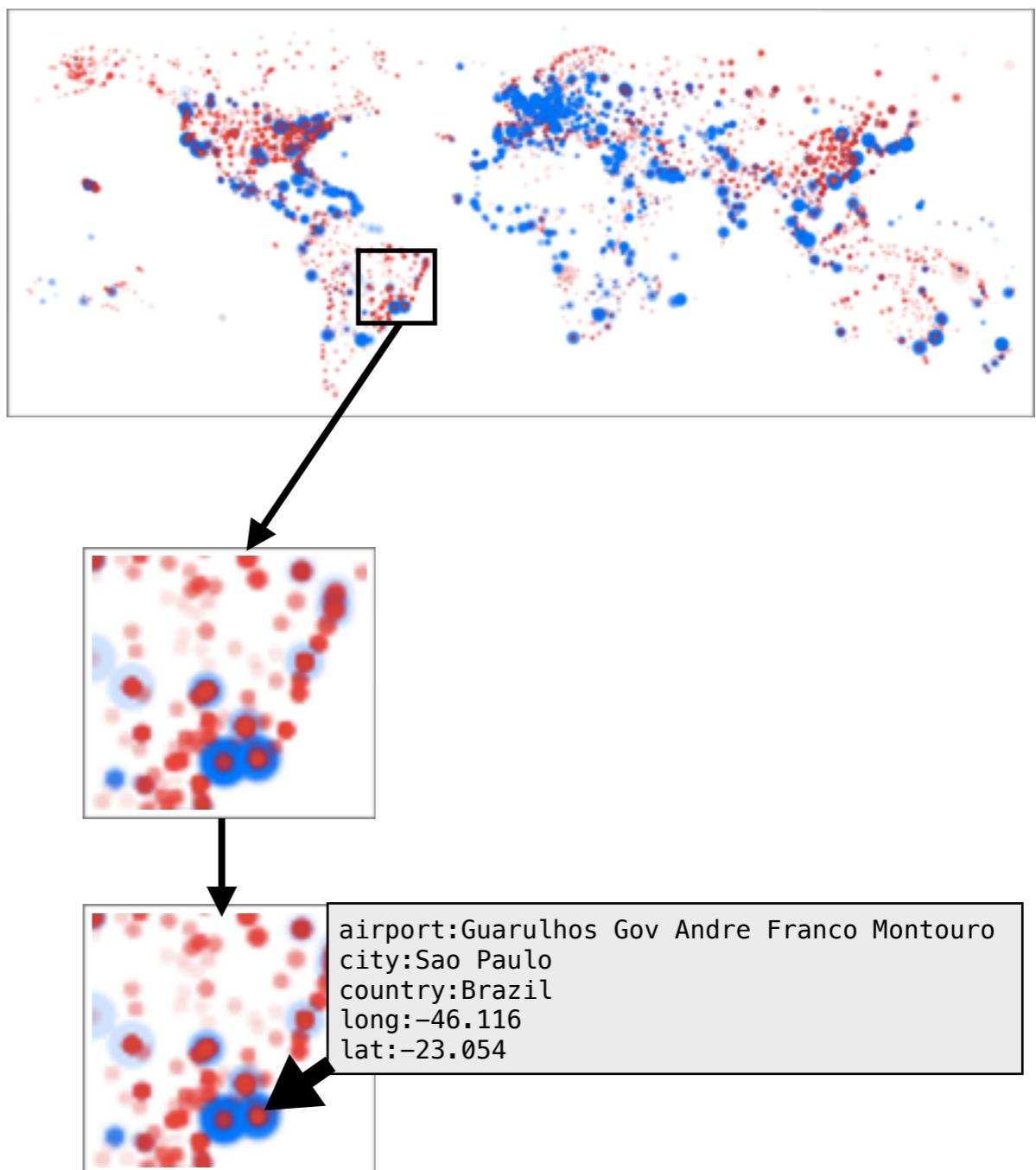
No occlusion (all information is visible simultaneously)

(Keahey, 2003)

6. 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 capture

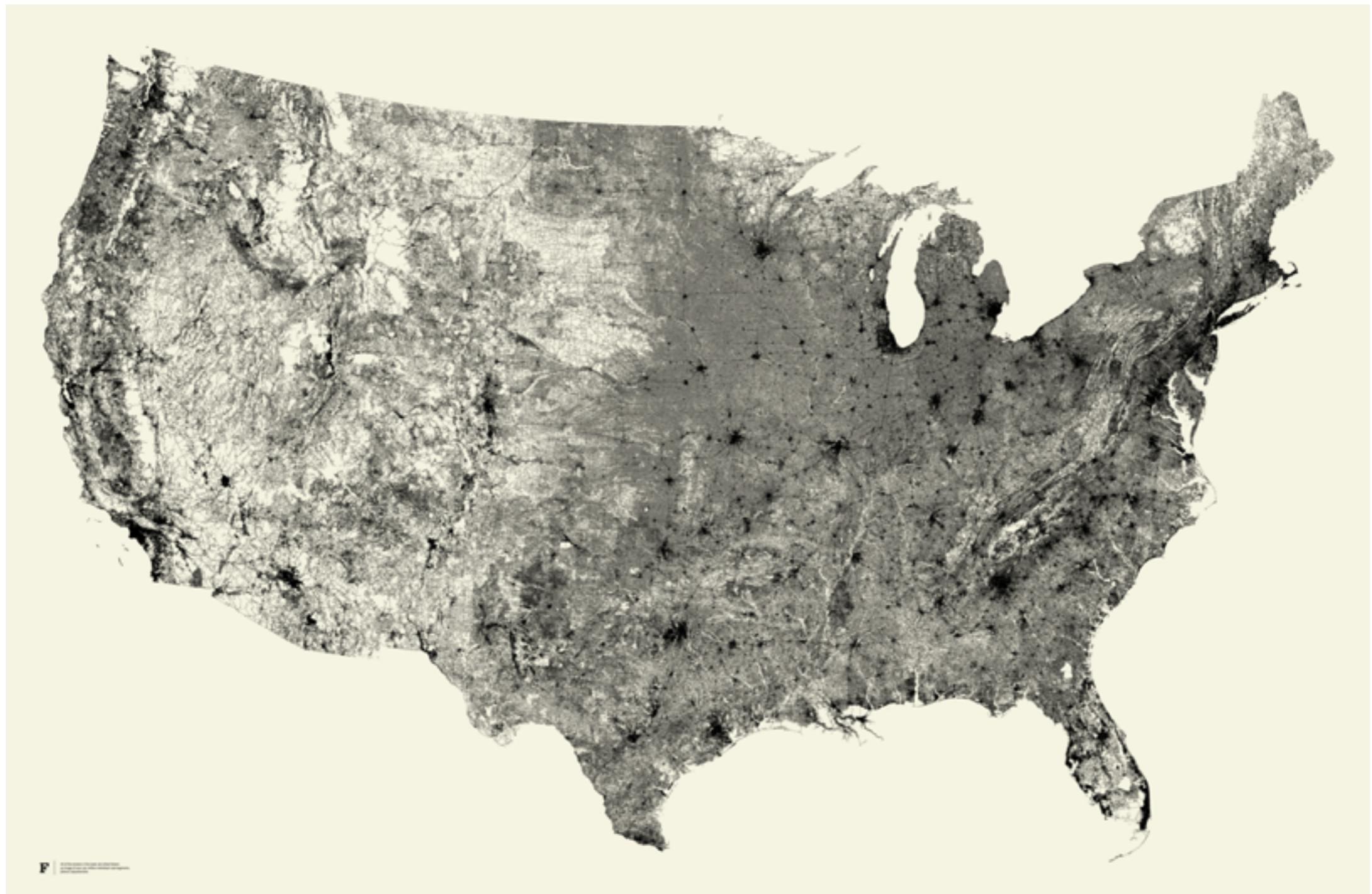


7. “underengineer”, if possible

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

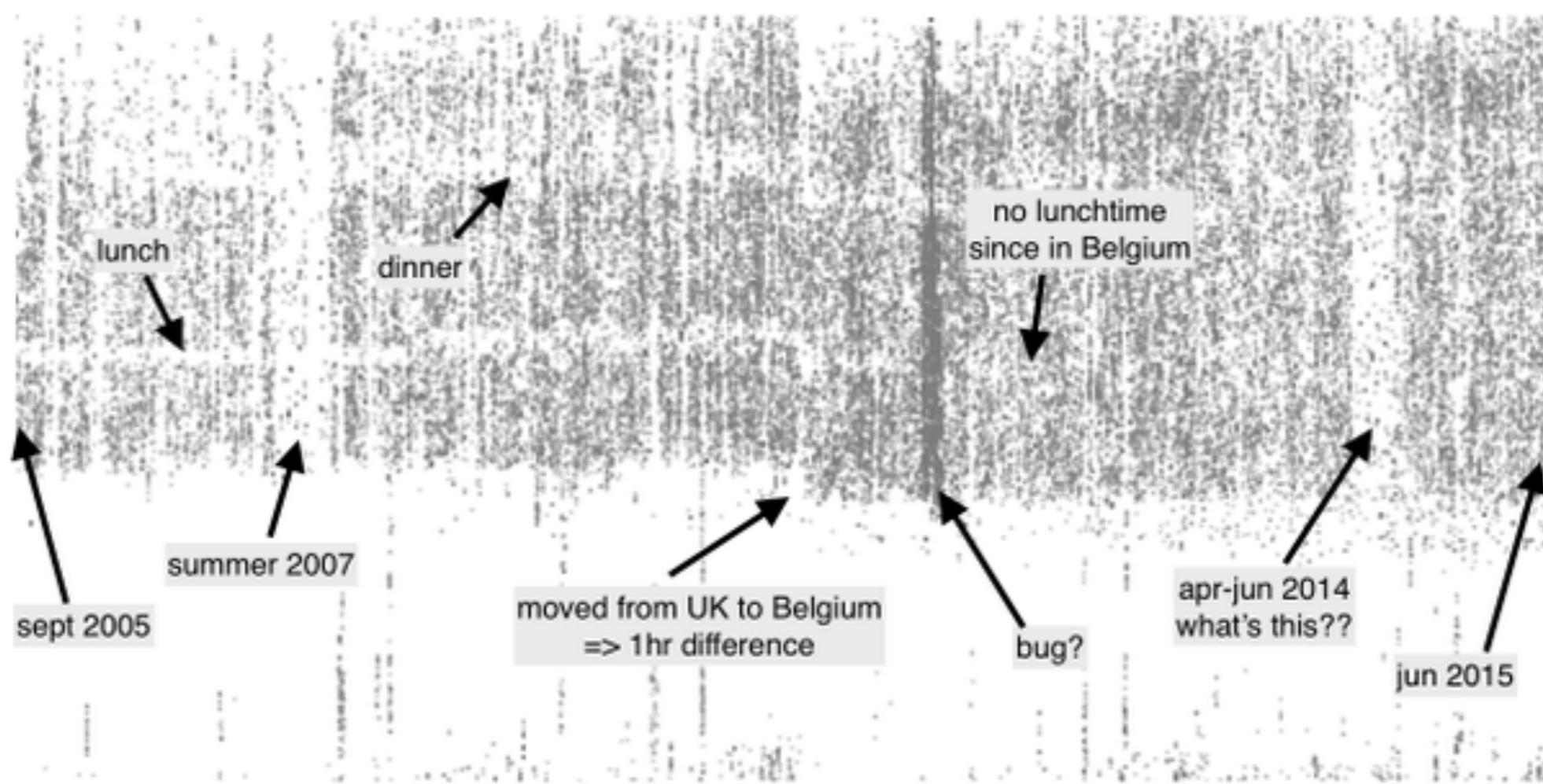




what is drawn: roads

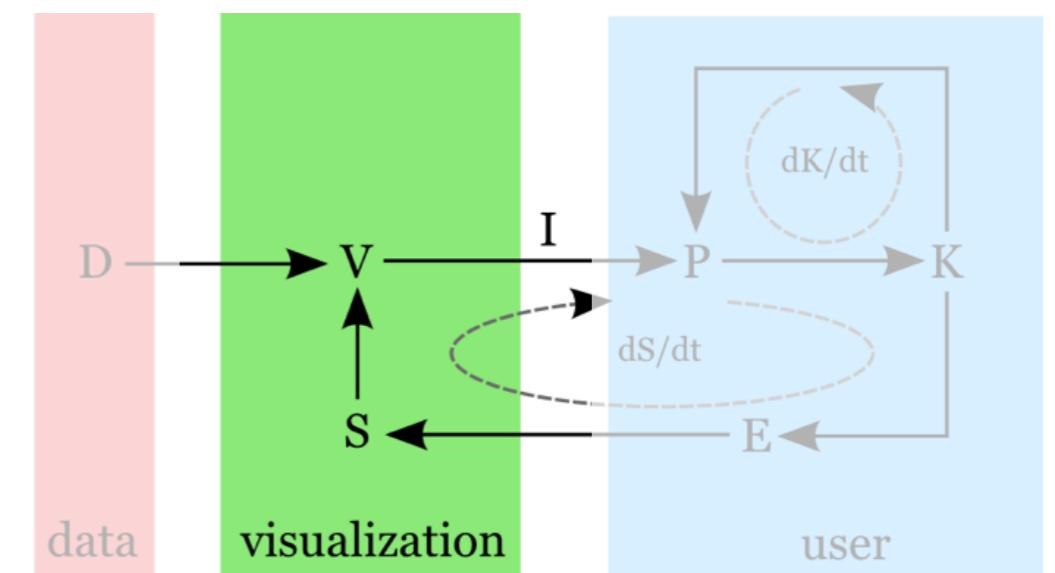
what we see: cities, mountain ranges, population density, ...

=> we observe higher-level patterns; not the raw data

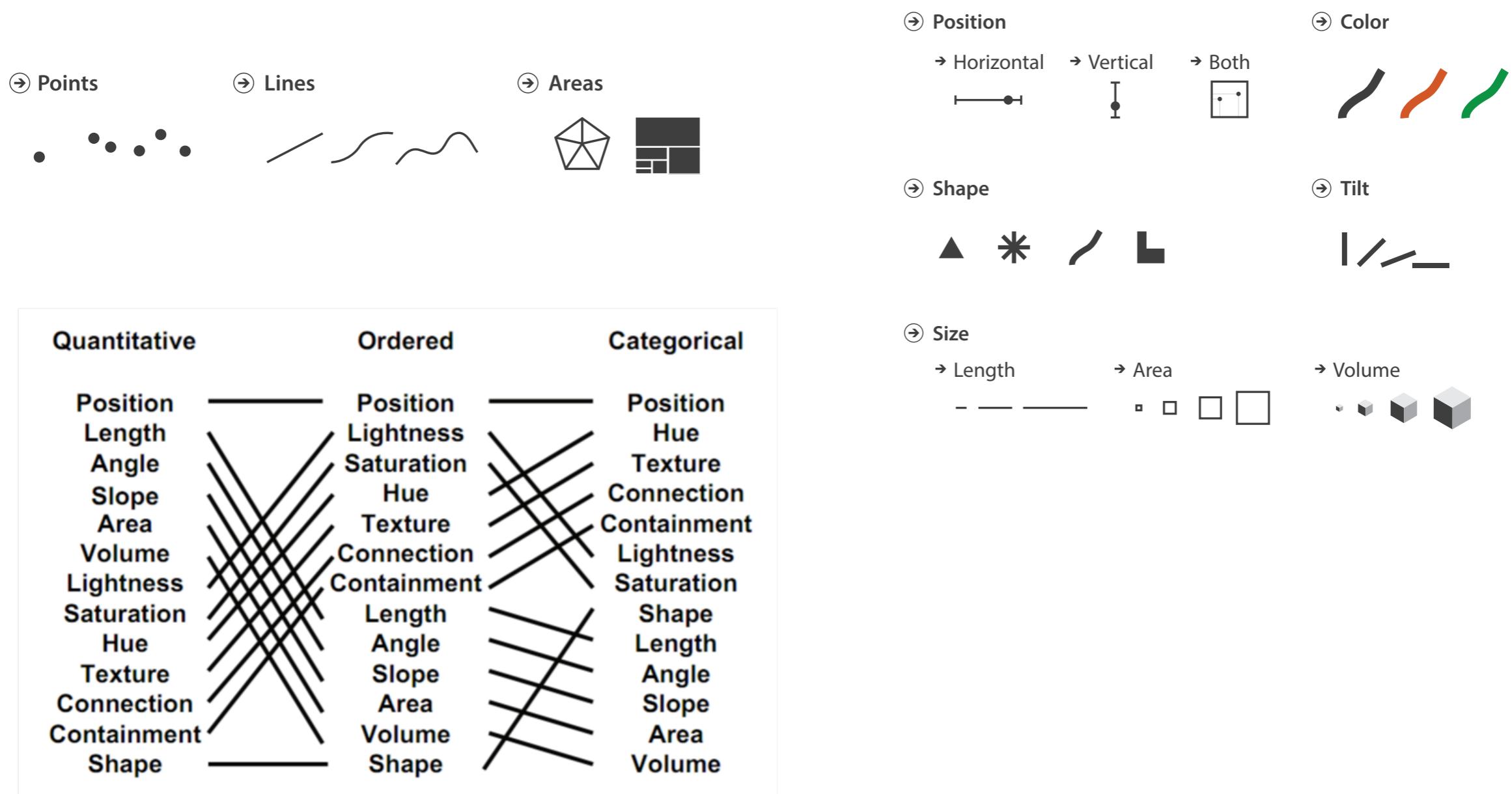


“send” date & time for all my outgoing email since 2005

Visualization foundations and techniques

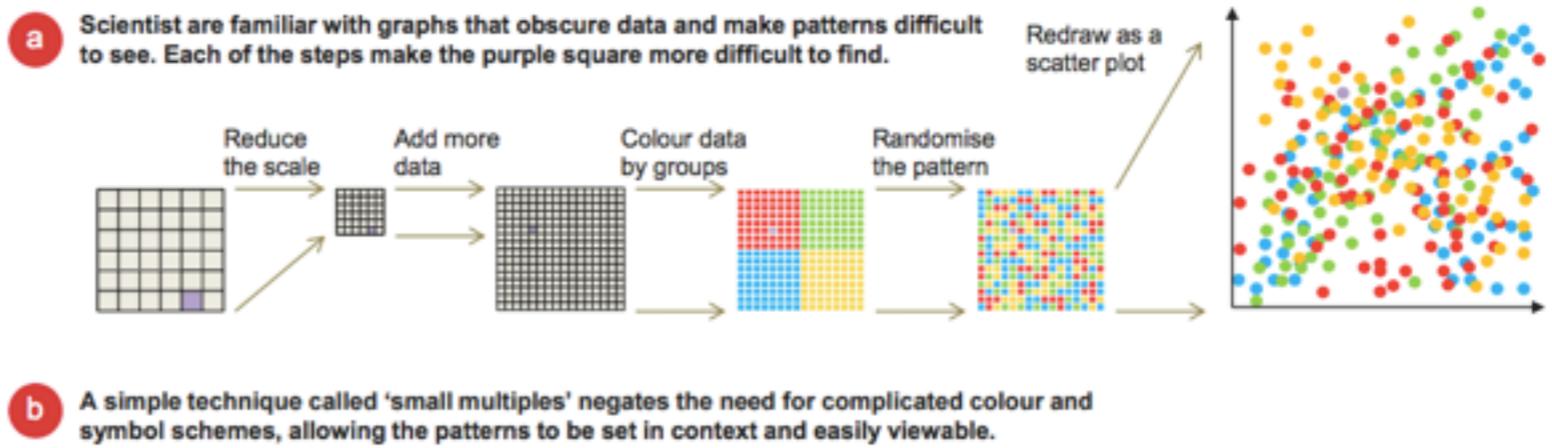
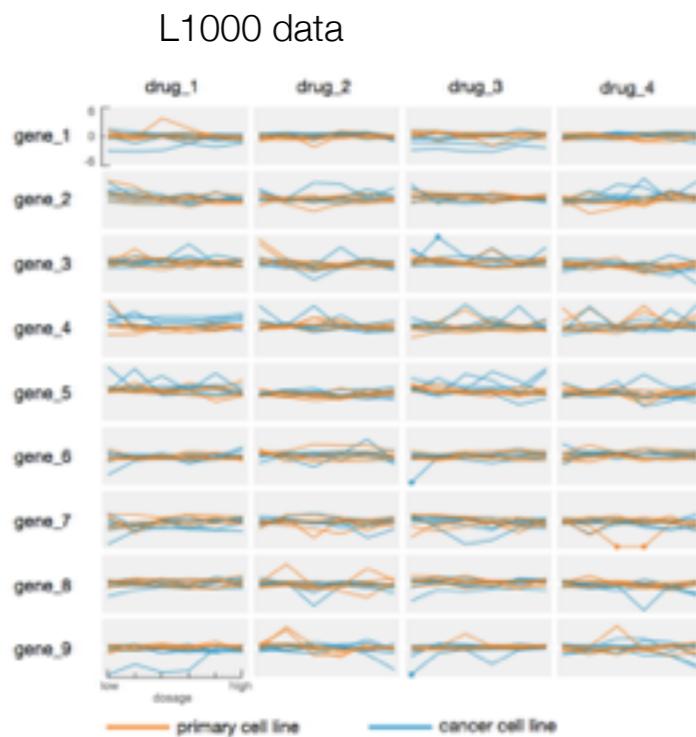


Mapping data to marks with properties

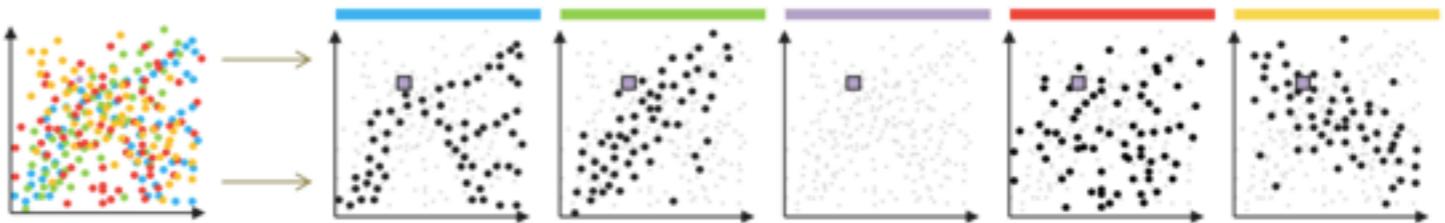


Combining visuals

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

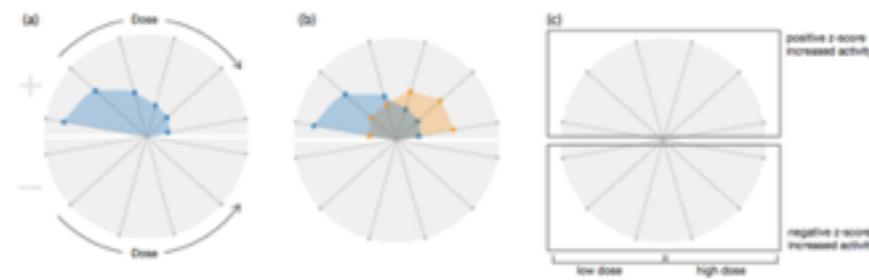
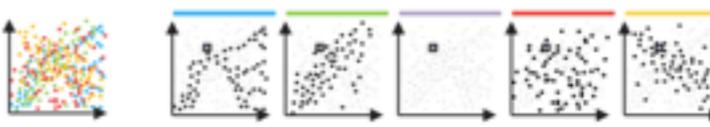


b A simple technique called 'small multiples' negates the need for complicated colour and symbol schemes, allowing the patterns to be set in context and easily viewable.

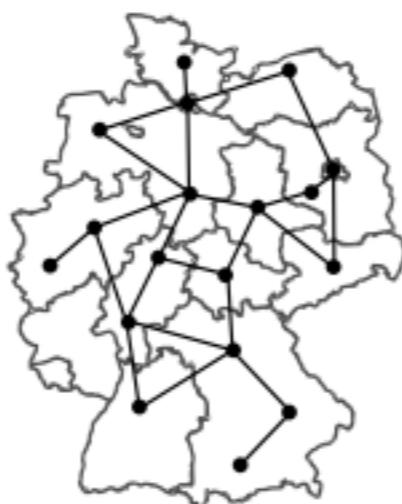


c In this design important patterns are still observable in small graphics.

reference?



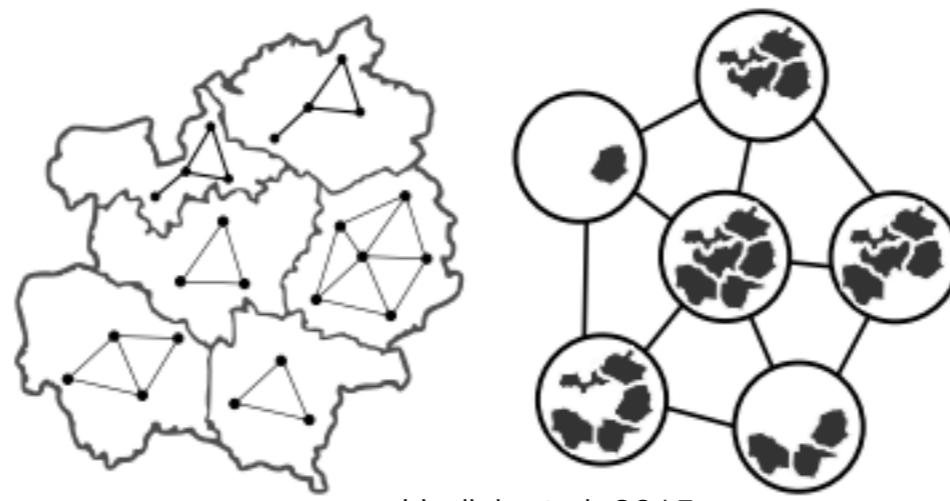
- **superimpose**



Hadlak et al, 2015

- **embed/nest**

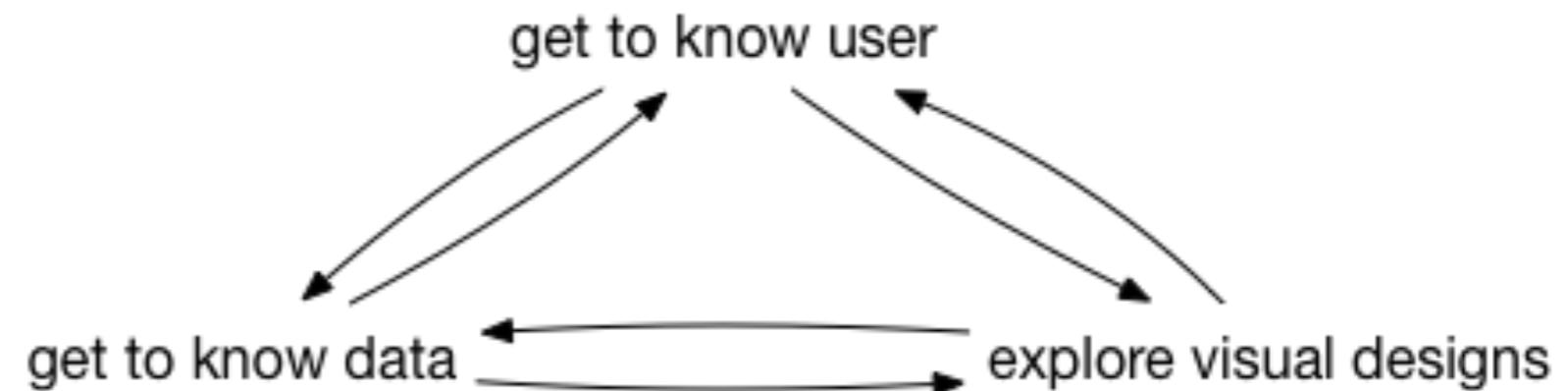
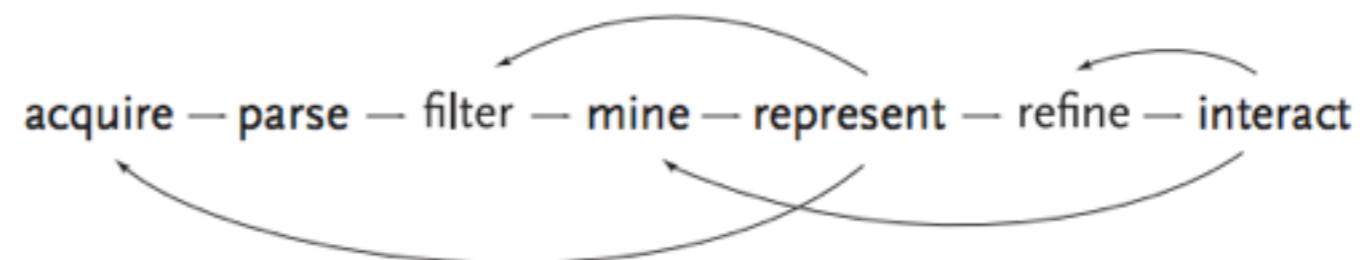
- e.g. network + map: each can be “base”



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

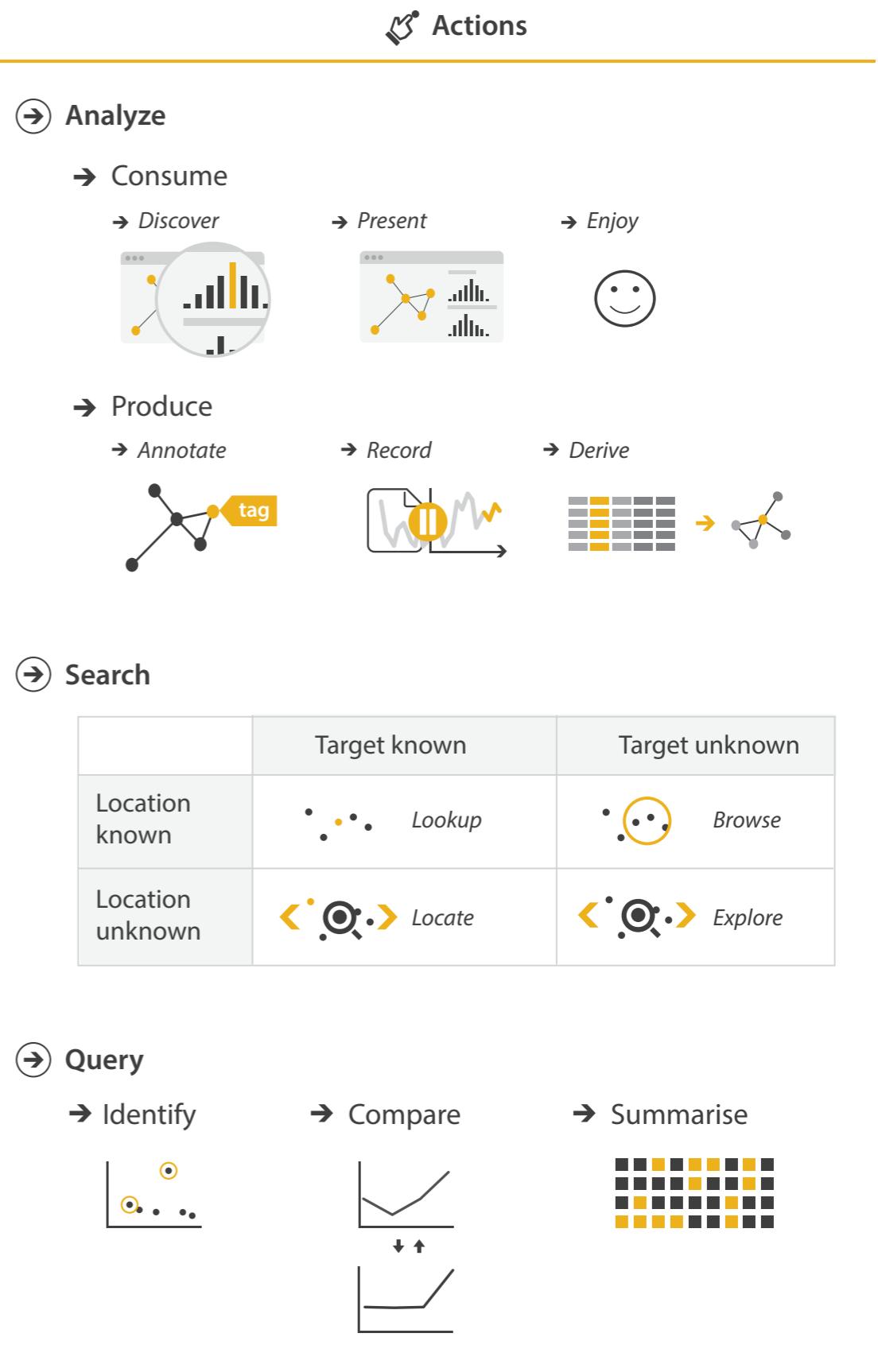
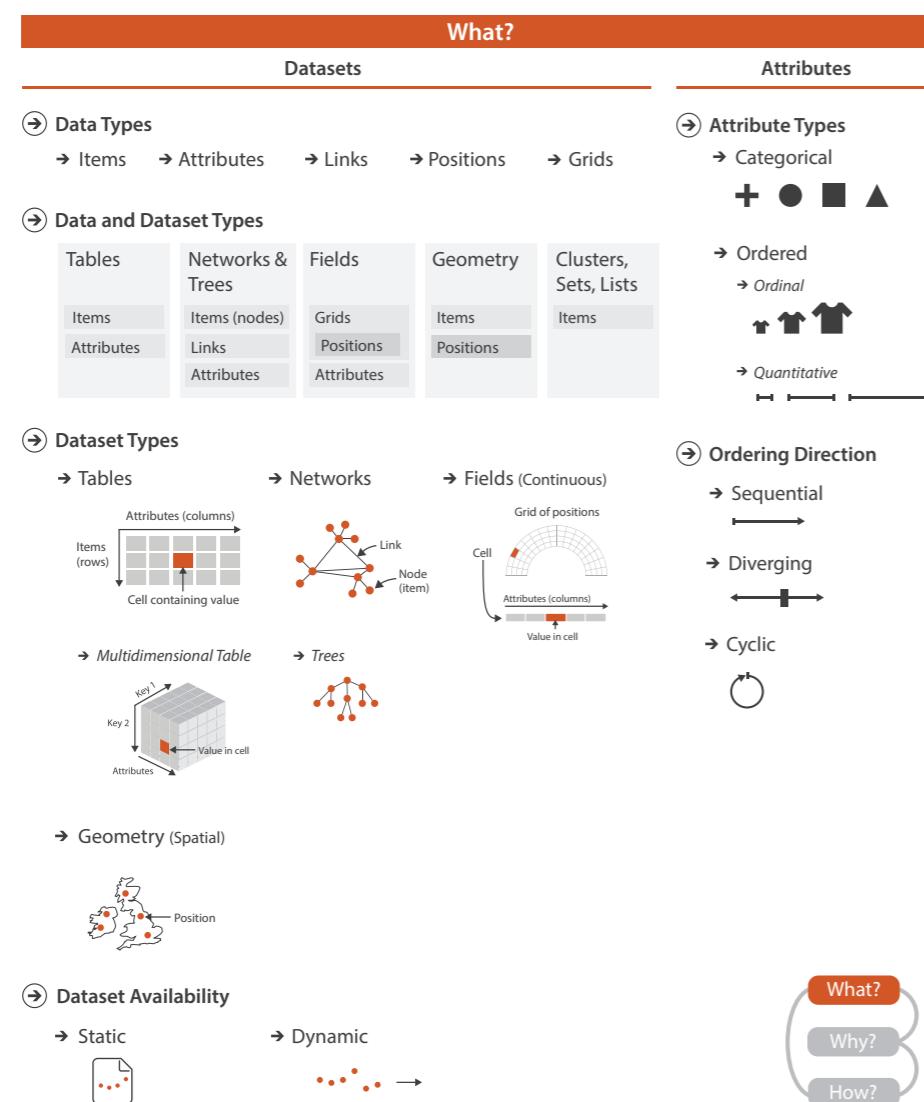
1. Talk to the user

- 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)
- ask “why?” 3 times



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

- data abstraction

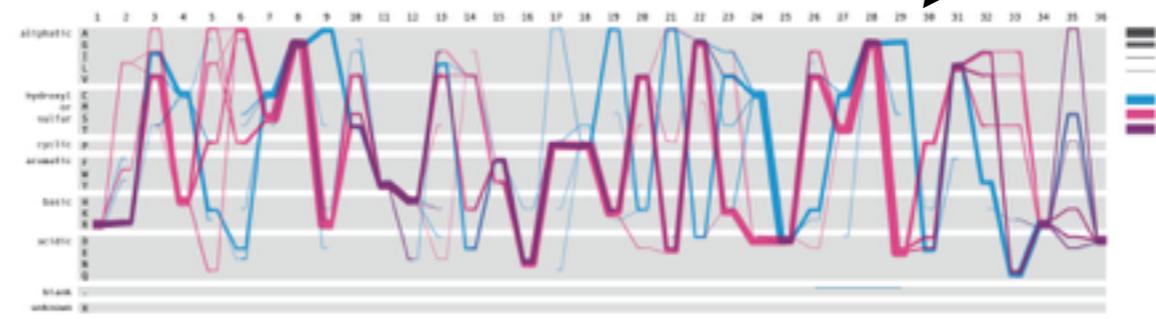
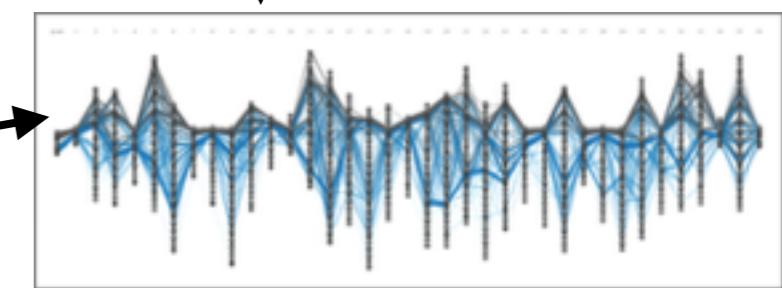
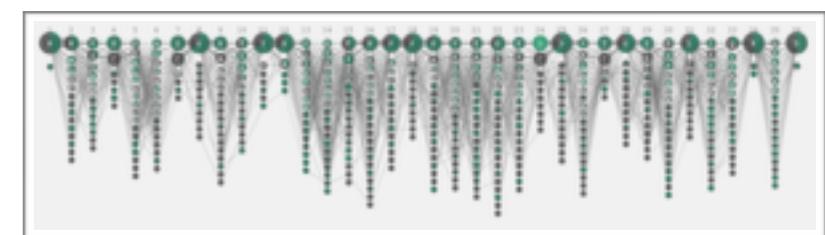
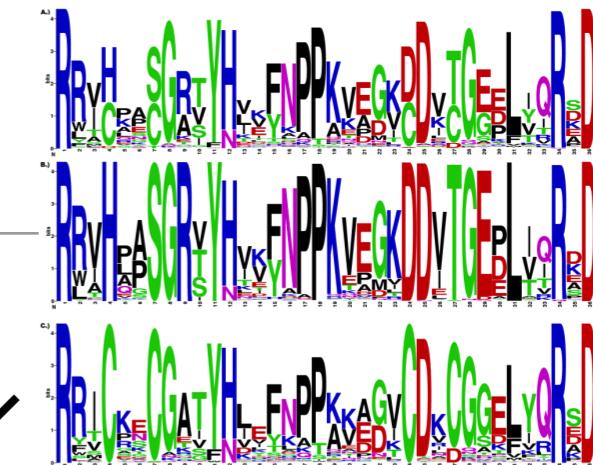
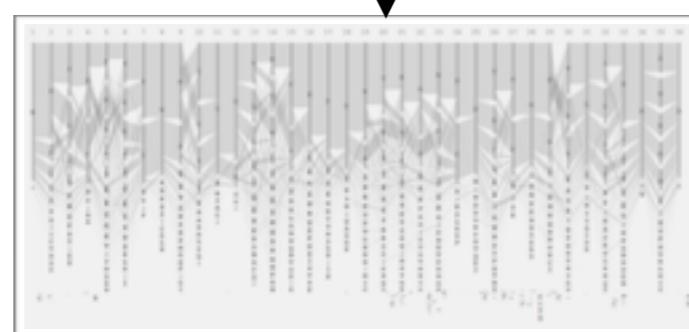
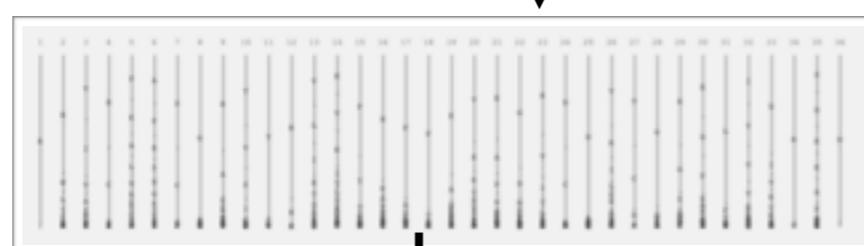
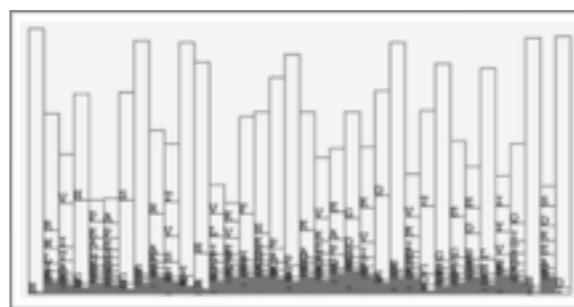
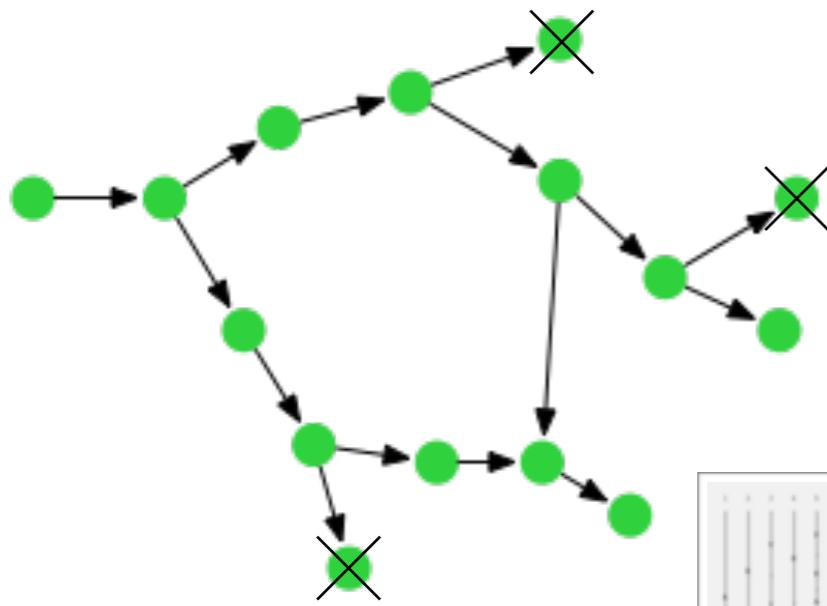


2. Get to know the data

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

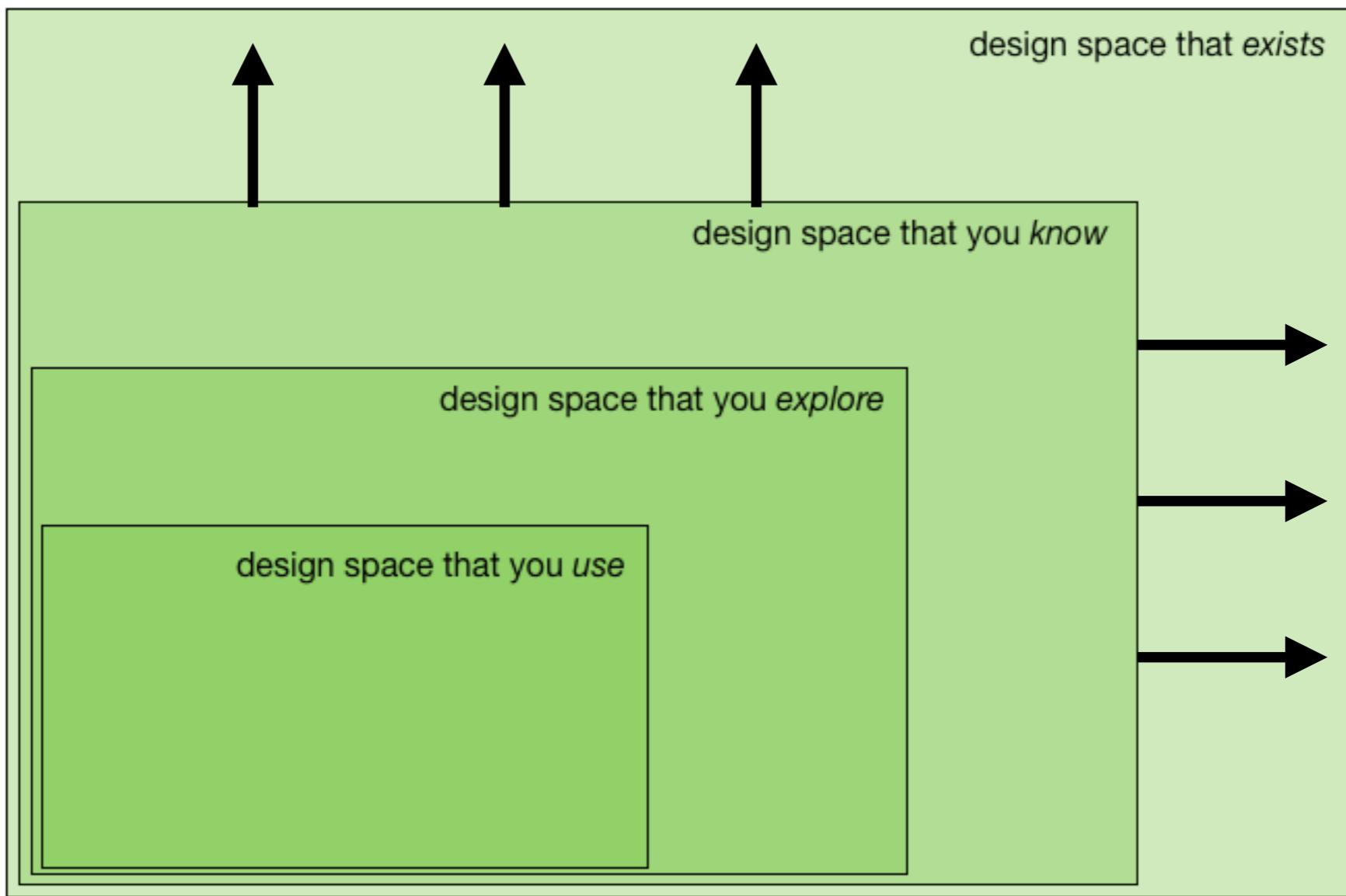


design space that exists

design space that you *know*

design space that you *explore*

design space that you *use*

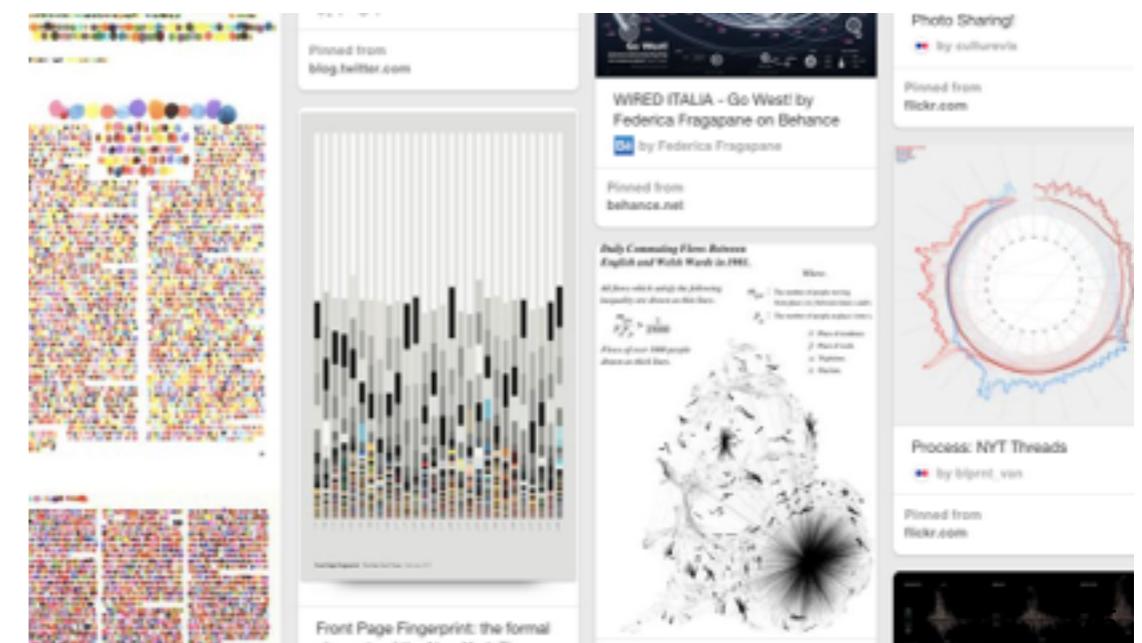


problem: initial design space that you know is small => how to expand?

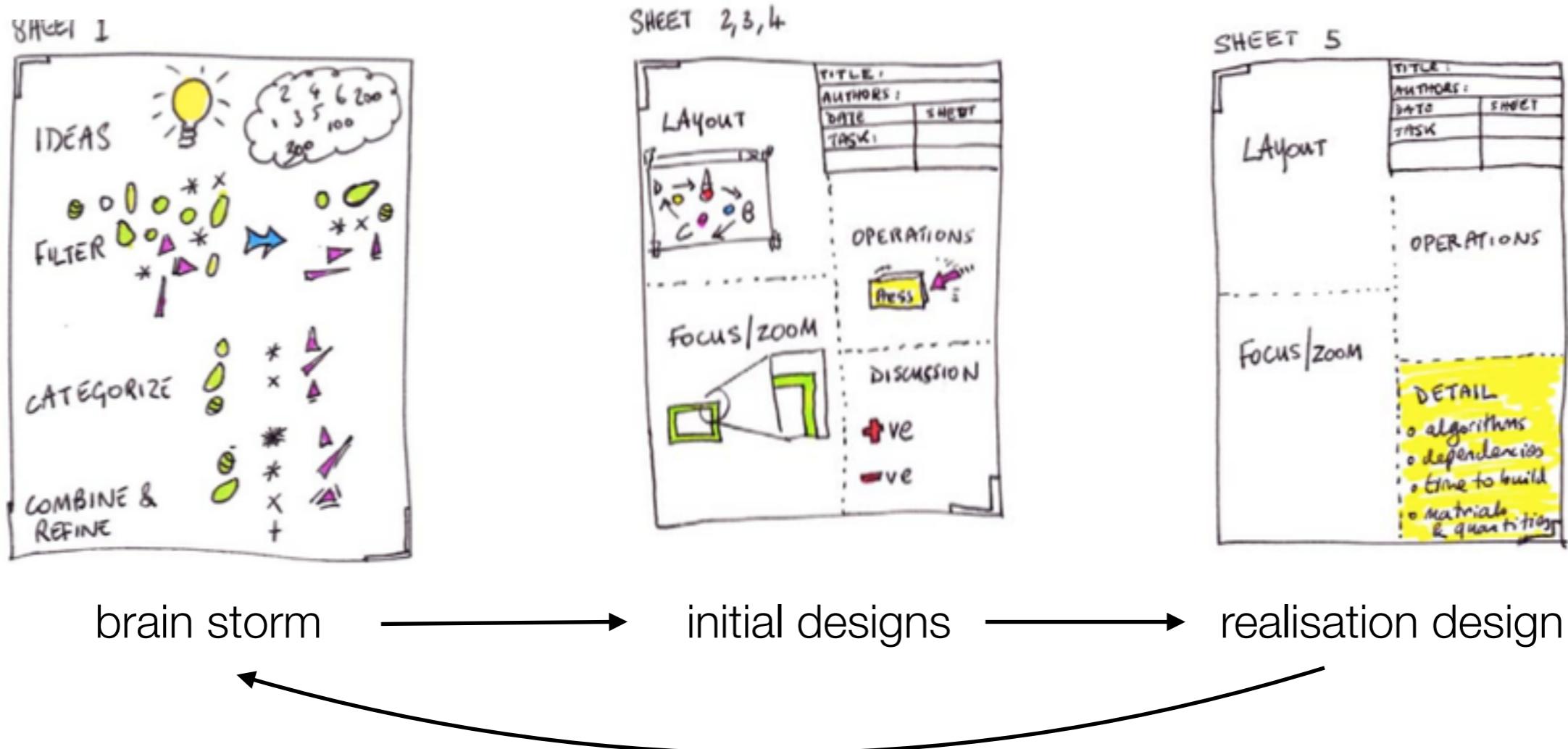
= “design fixation”

Tips to explore design space:

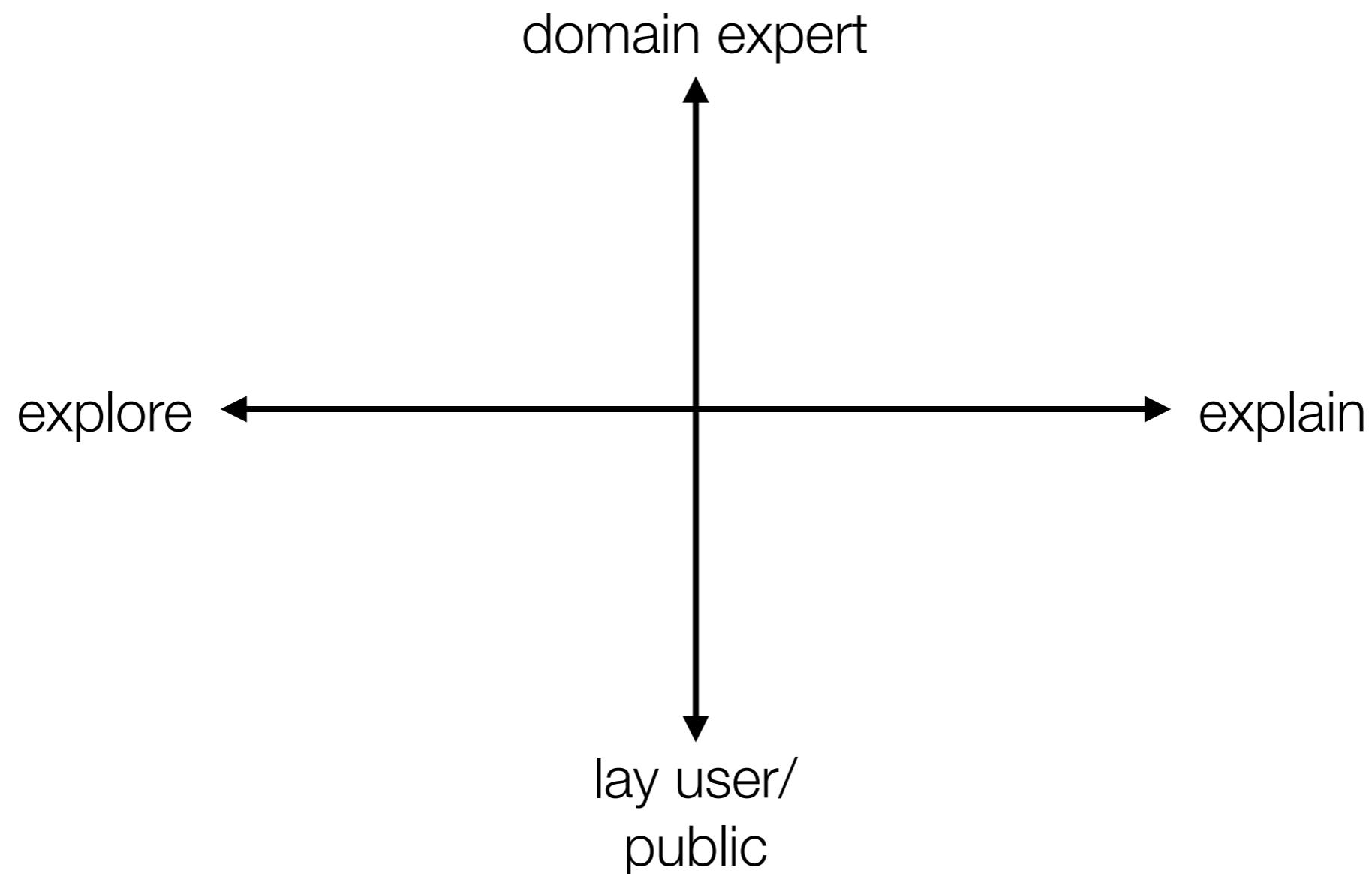
- Use systematic approach such as 5-design-sheet (see paper)
- Make very “cheap” sketches: pen & paper => don’t get attached to a design because you spent a lot of time on it
- Extend your visual library: collect interesting visuals (useful *and* useless), e.g. <https://www.pinterest.com/aertsjan/data-visualizationart/>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



5-design sheet methodology



Data visualization framework



Bret Victor - Ladder of abstraction

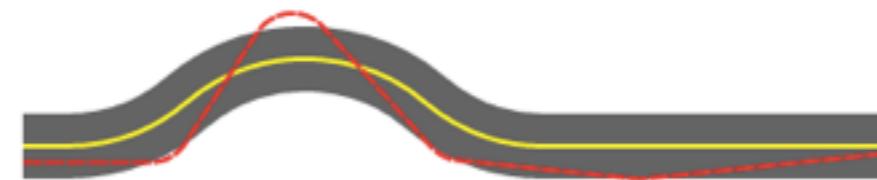
At each step:

Move forward 1 pixel.
If left of the road, turn right by 2° .
If right of the road, turn left by 2° .



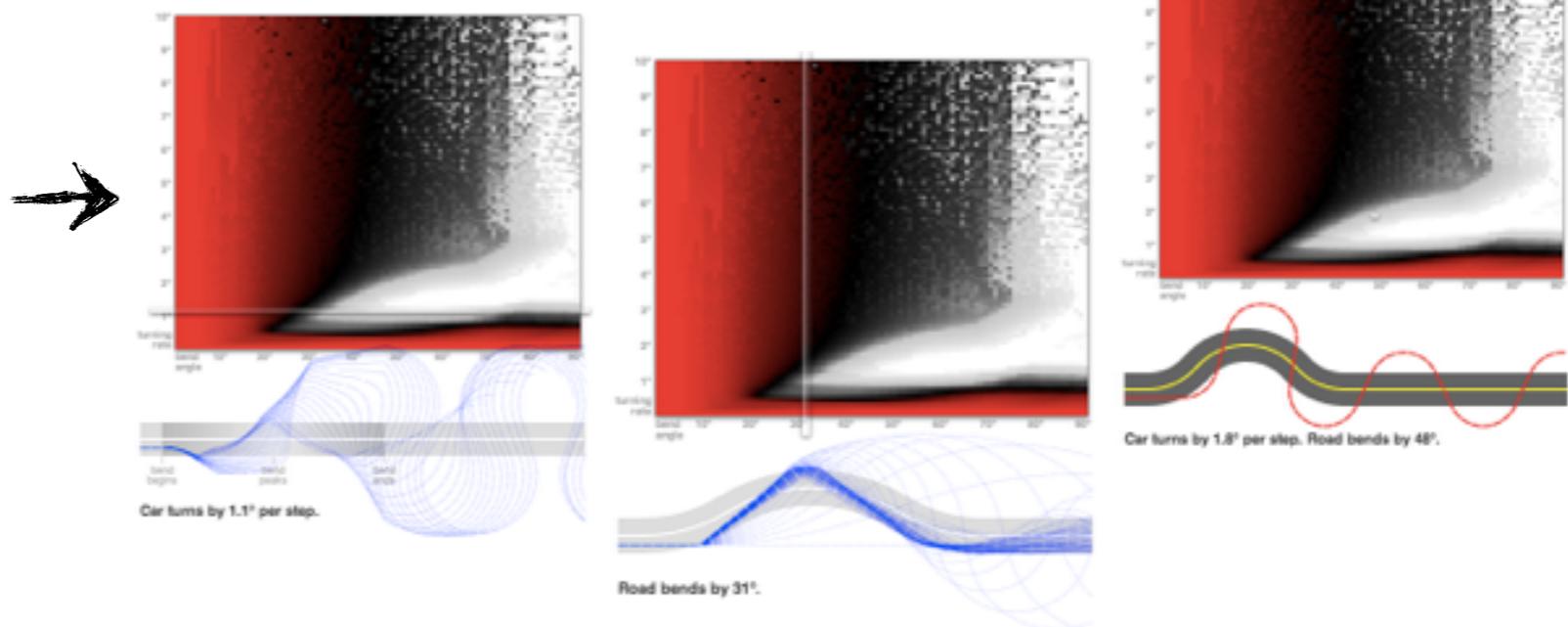
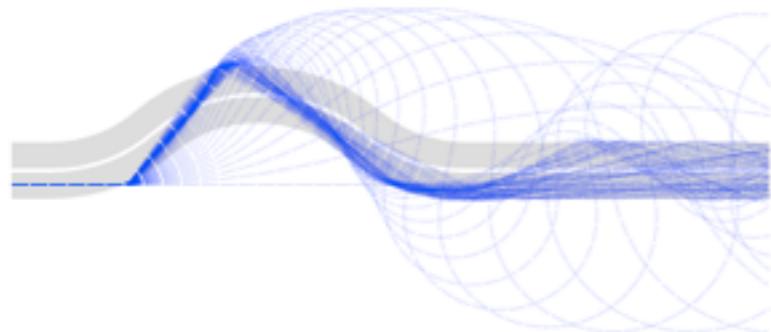
At each step:

Move forward 1 pixel.
If left of the road, turn right by 3.0° .
If right of the road, turn left by 3.0° .

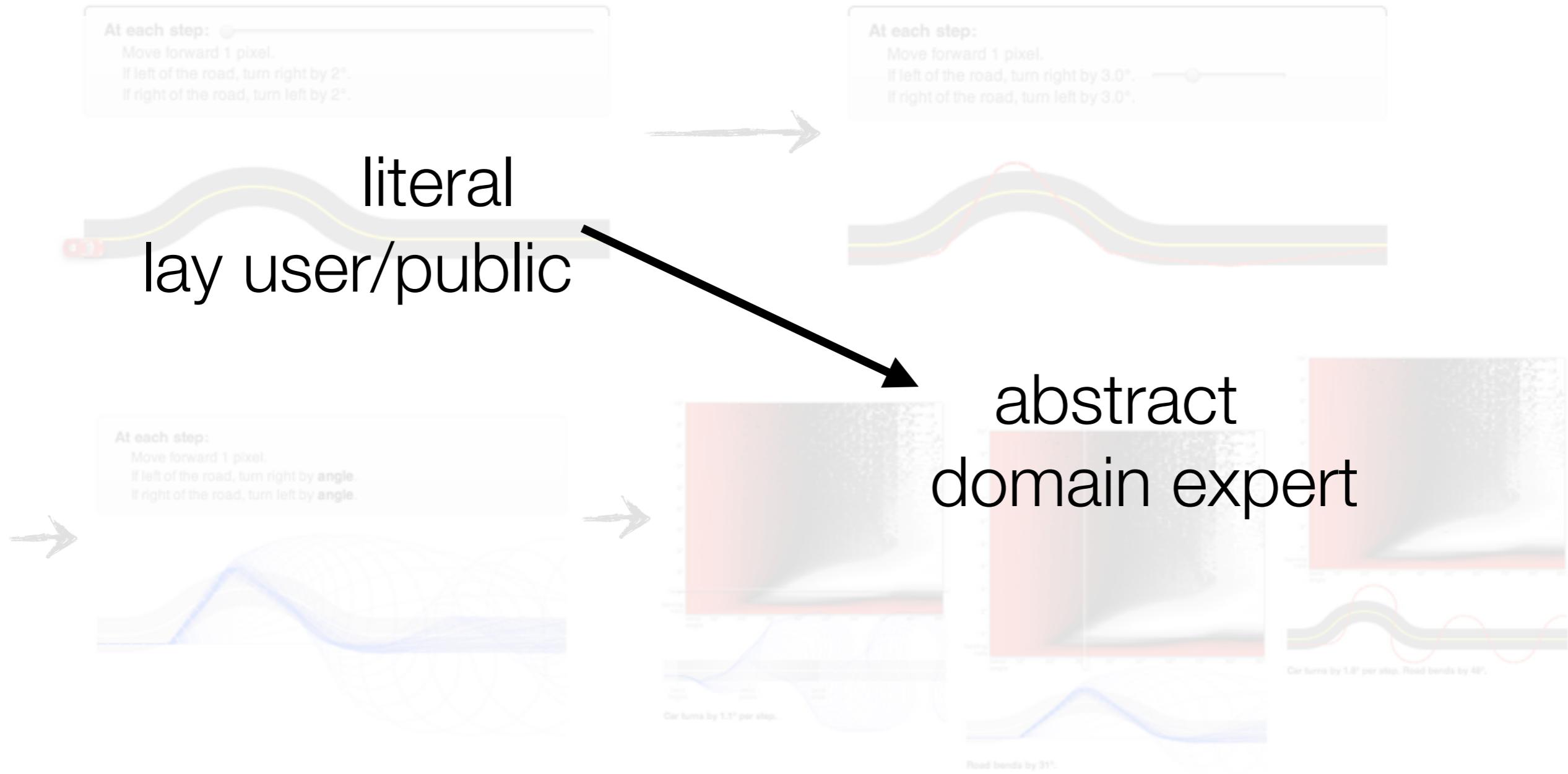


At each step:

Move forward 1 pixel.
If left of the road, turn right by **angle**.
If right of the road, turn left by **angle**.



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>)





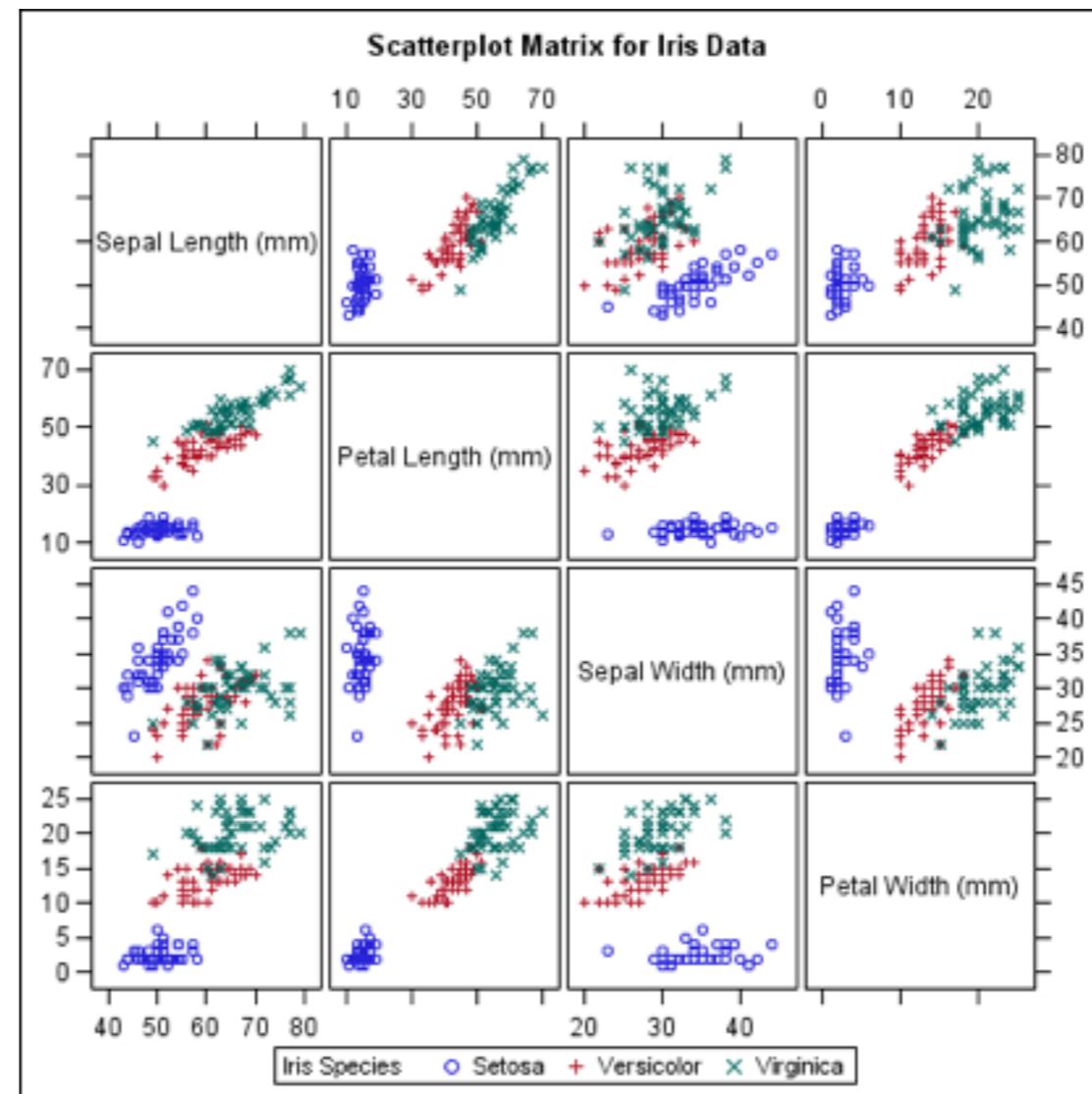
Iris setosa



Iris versicolor



Iris virginica



scatterplot matrix

(source: support.sas.com)



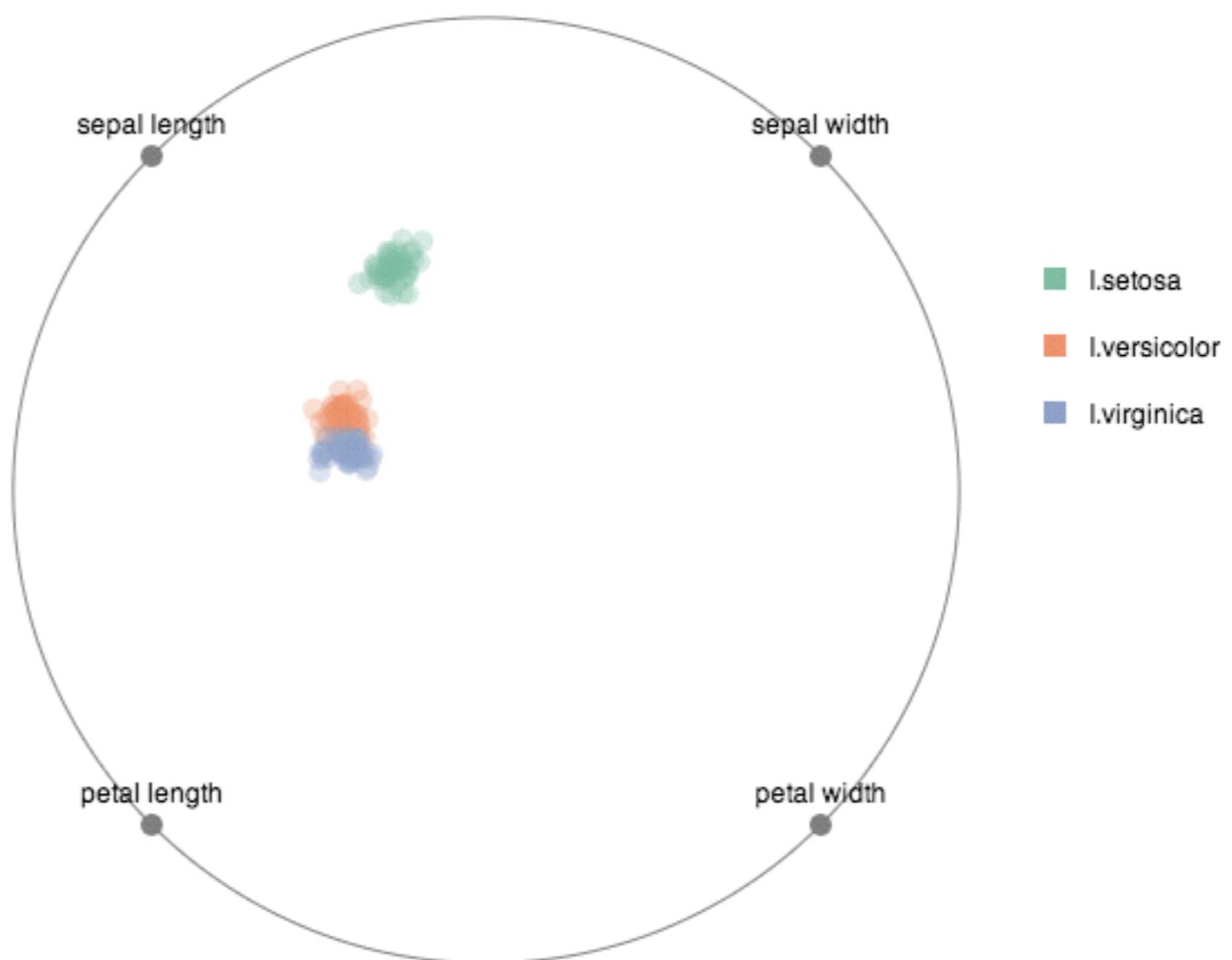
Iris setosa



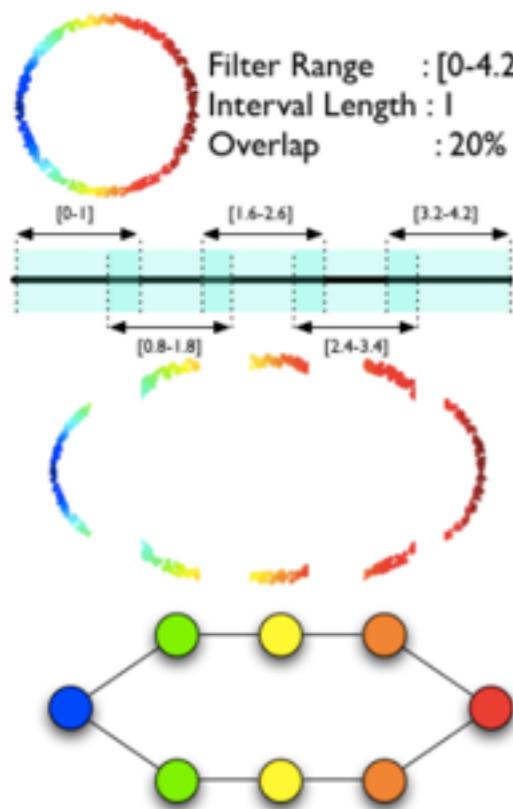
Iris versicolor



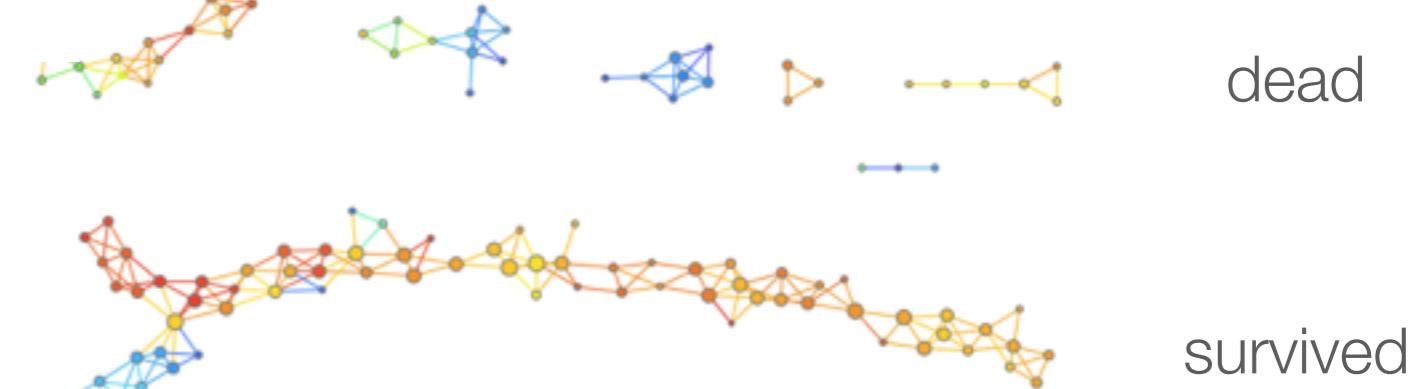
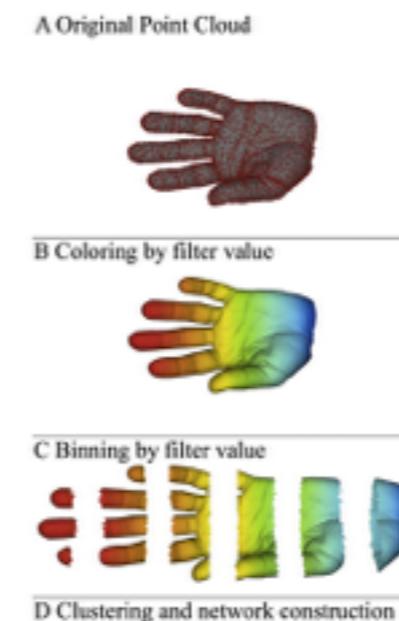
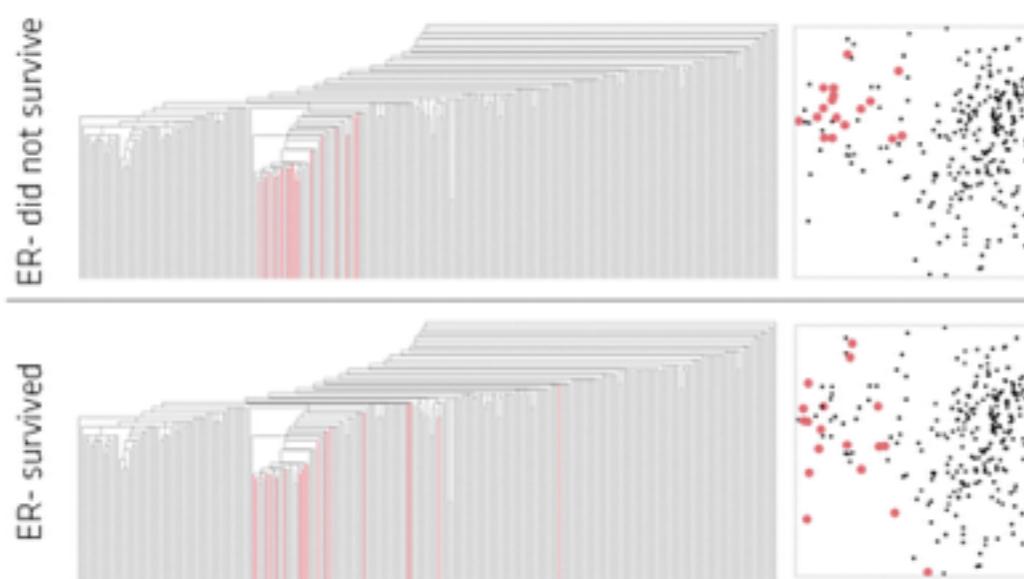
Iris virginica



radviz



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



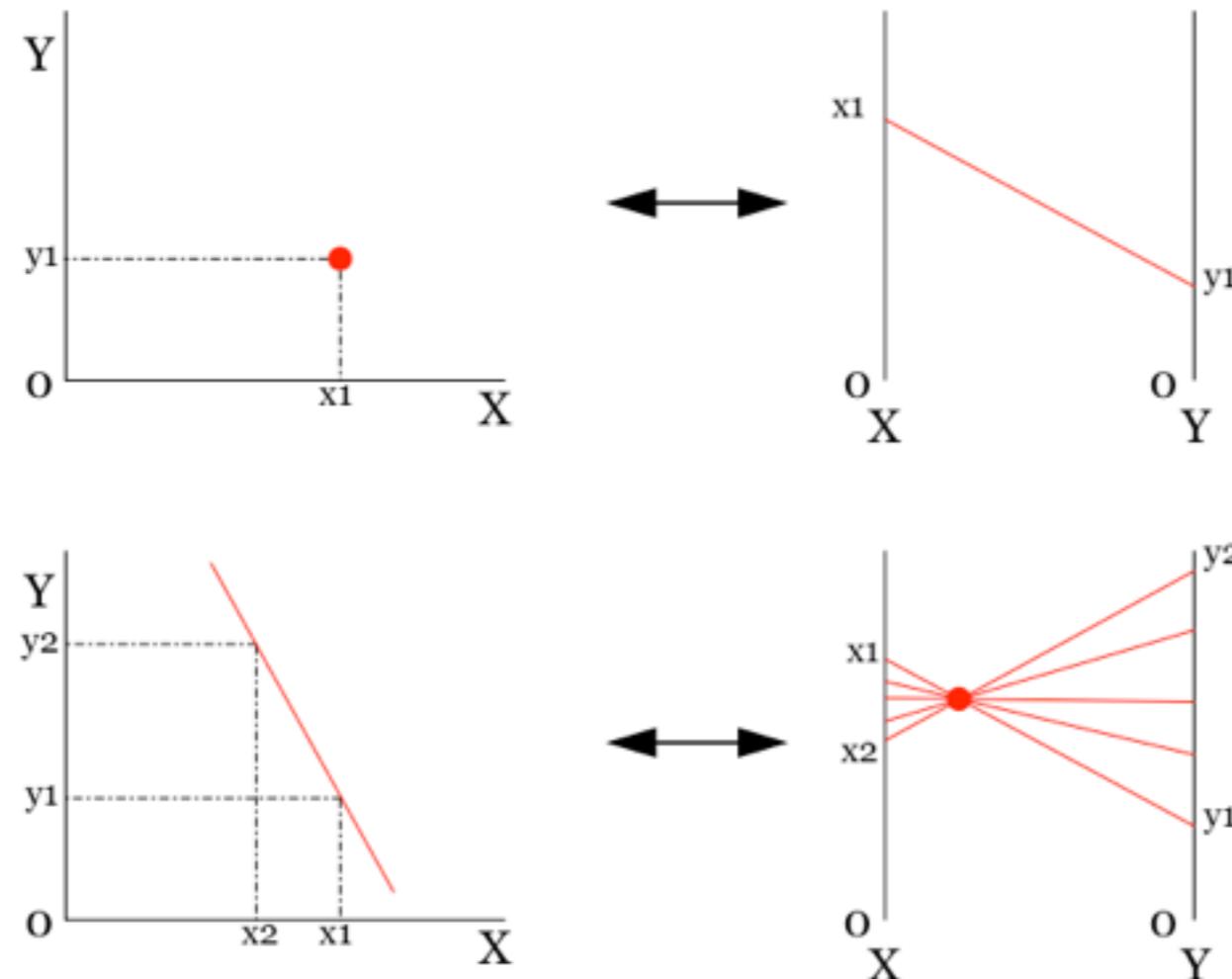
Ayasdi, Inc

topological data analysis

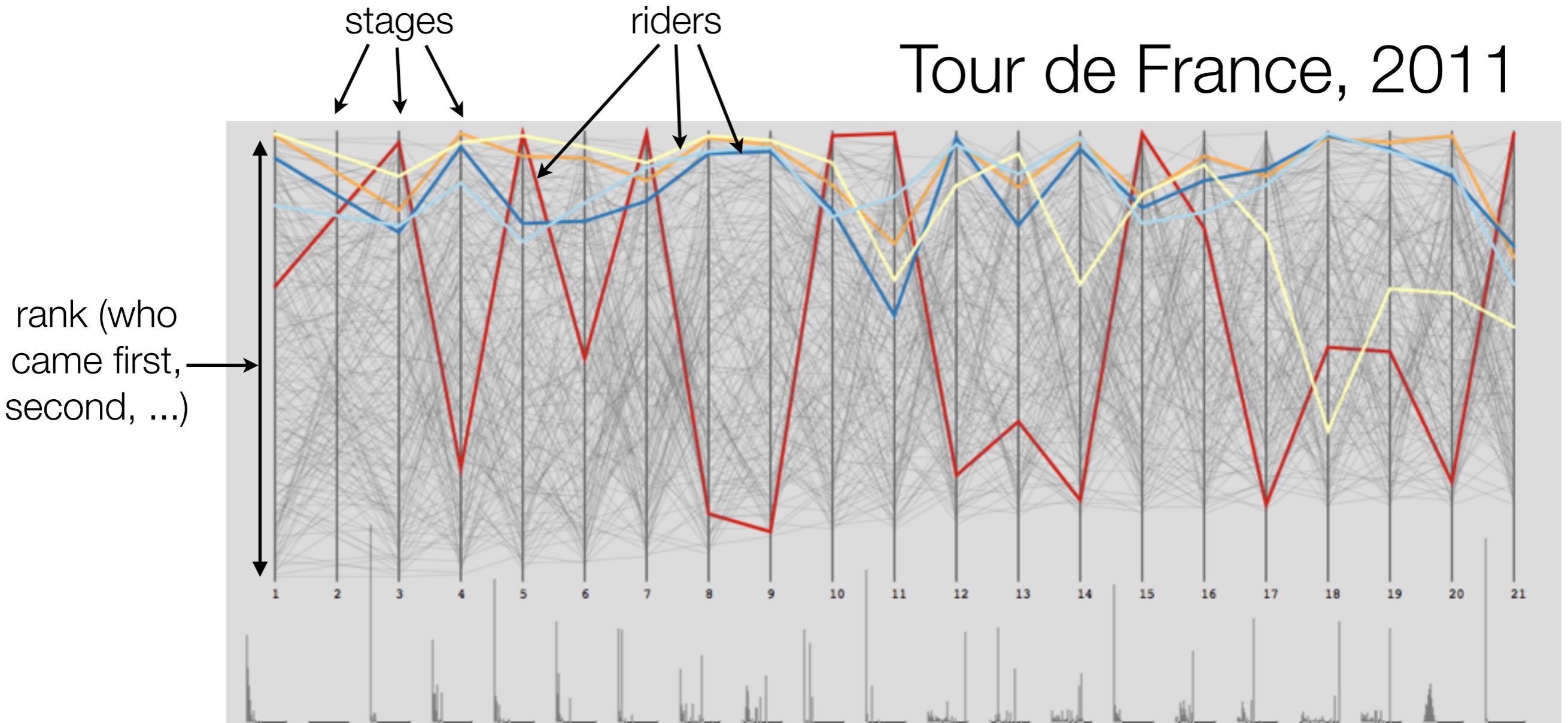
local clustering in global context

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

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

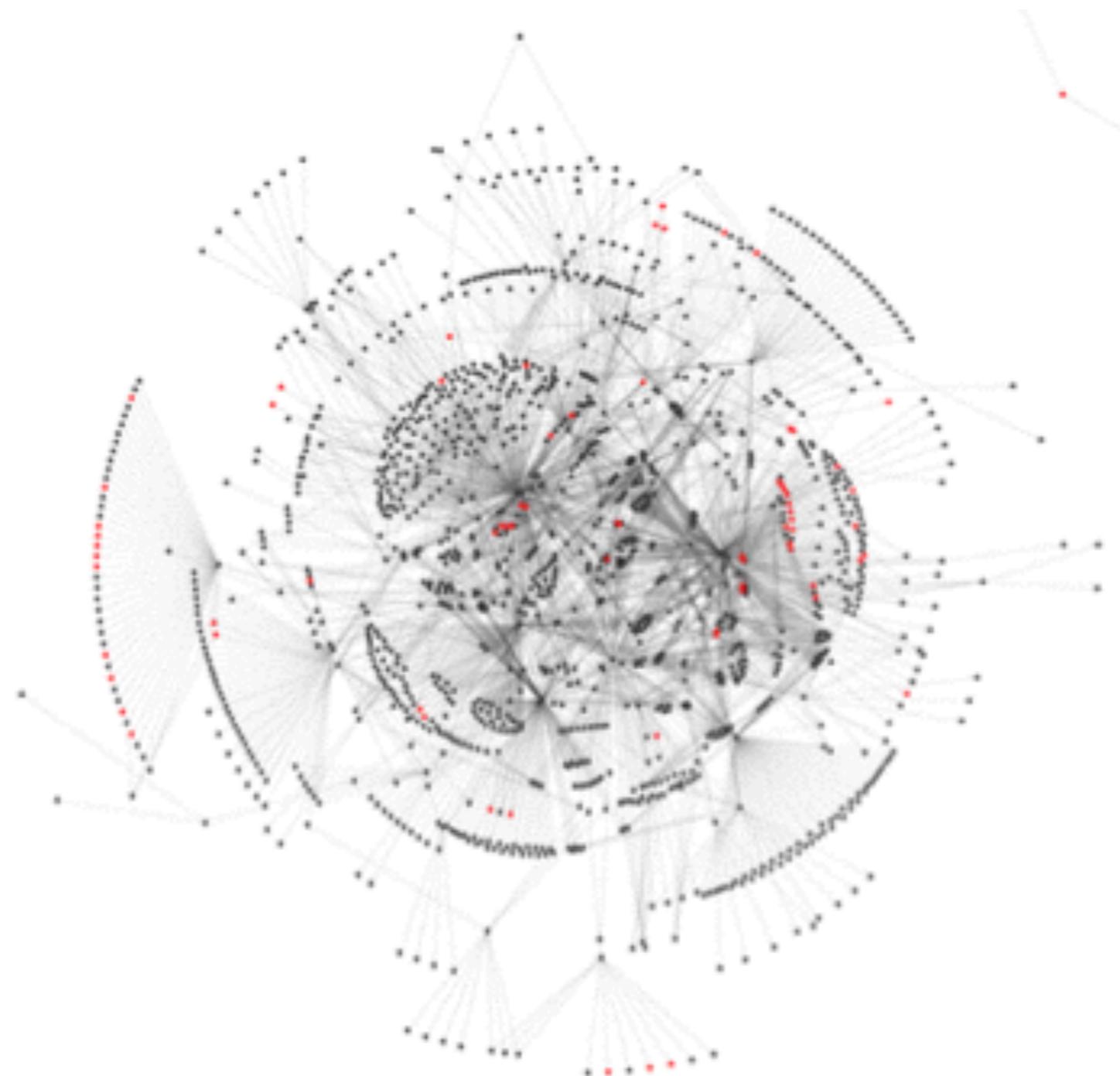


Tour de France, 2011



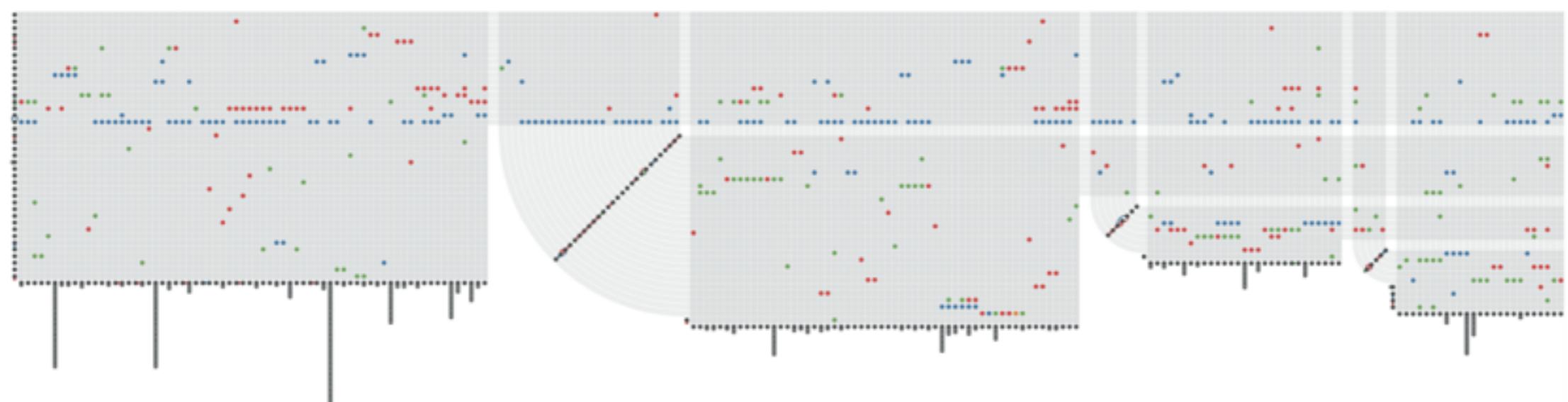
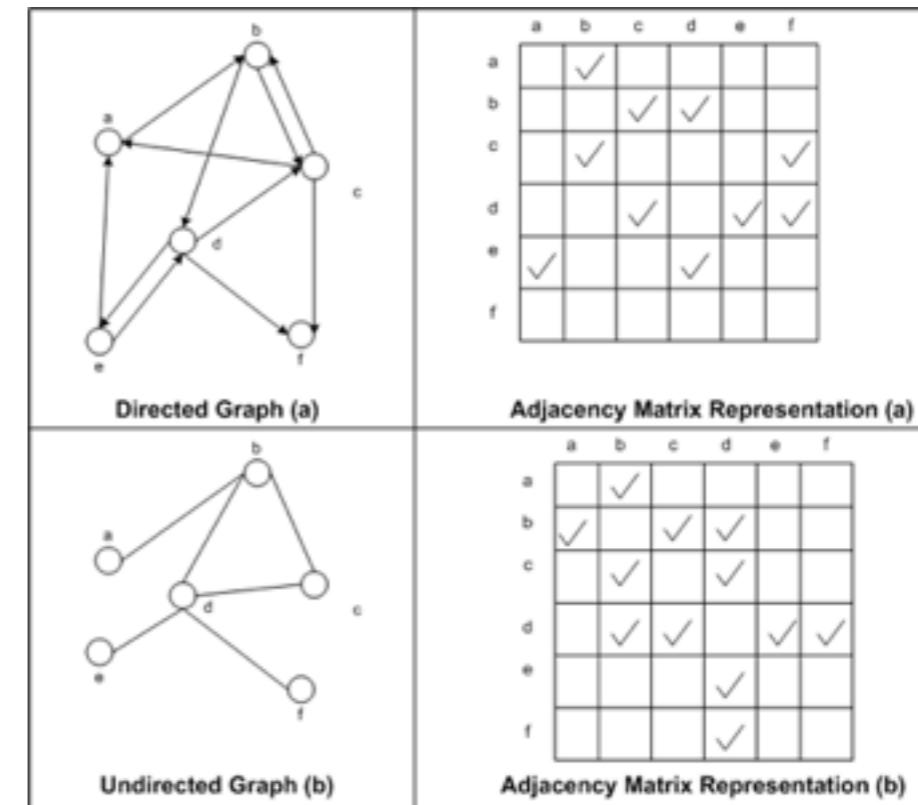
networks: node-link diagrams

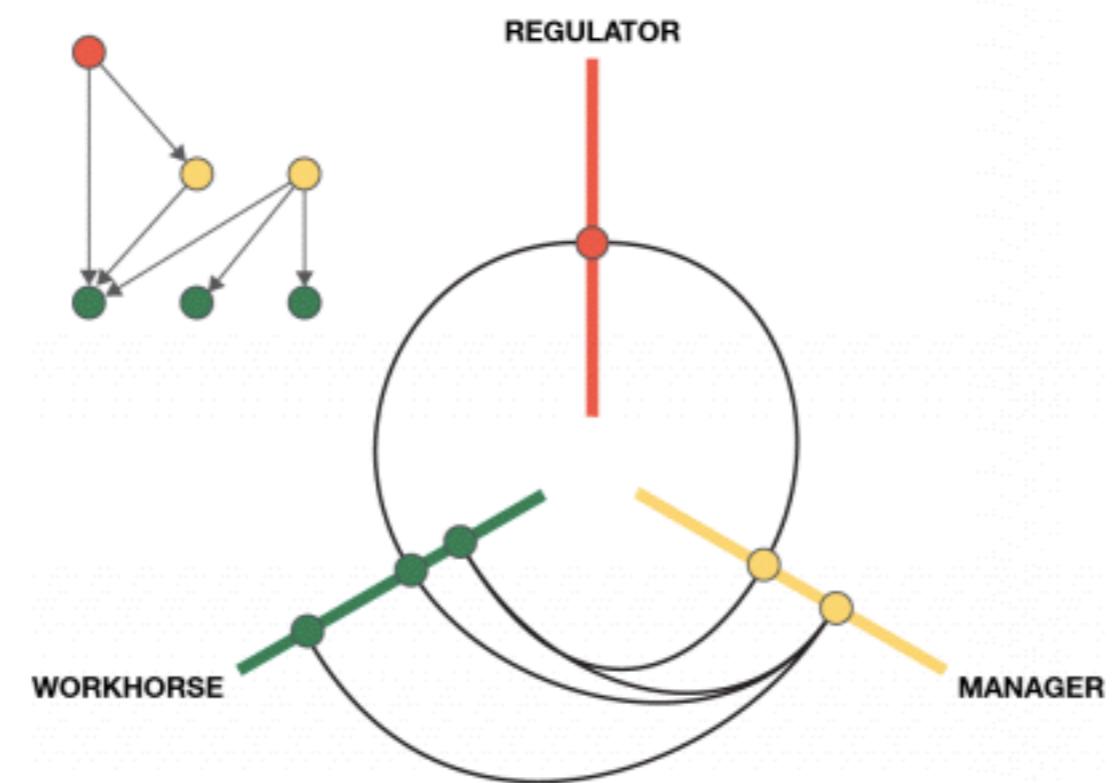
max 20 nodes...



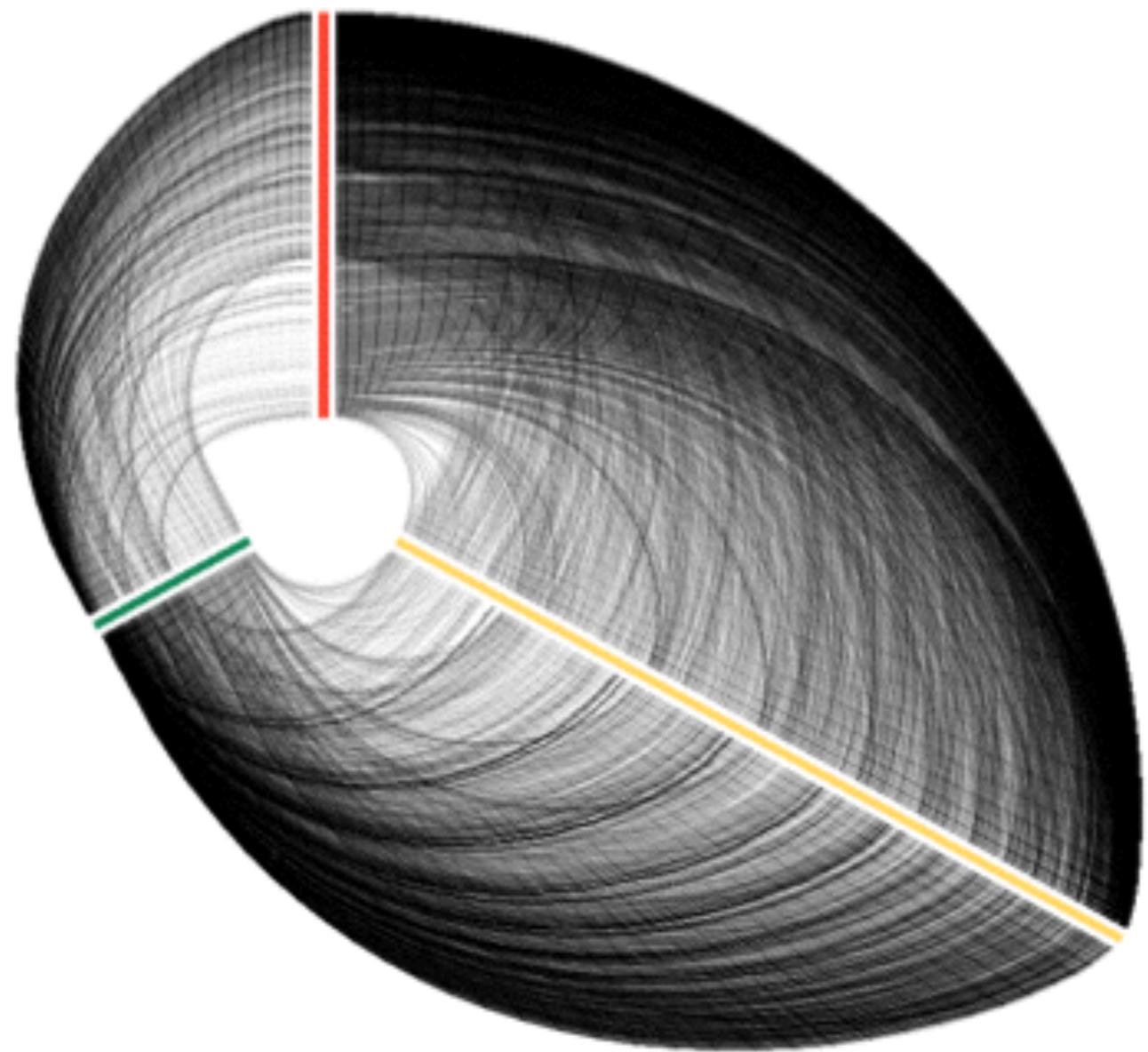
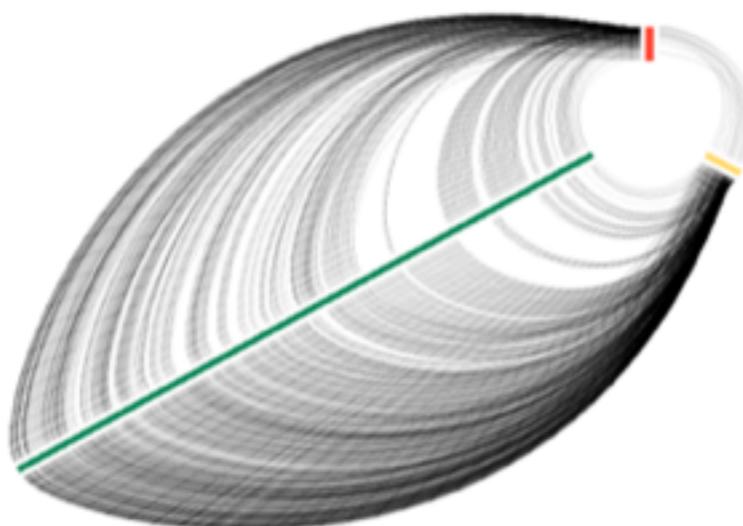
- adjacency matrix

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



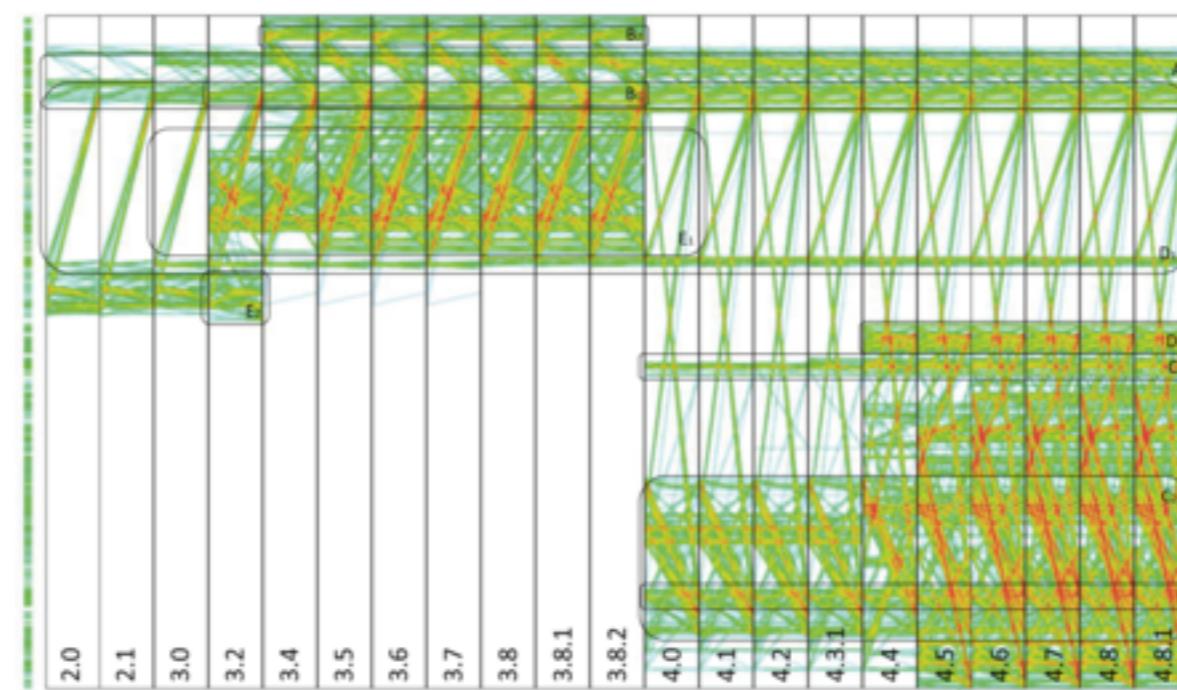
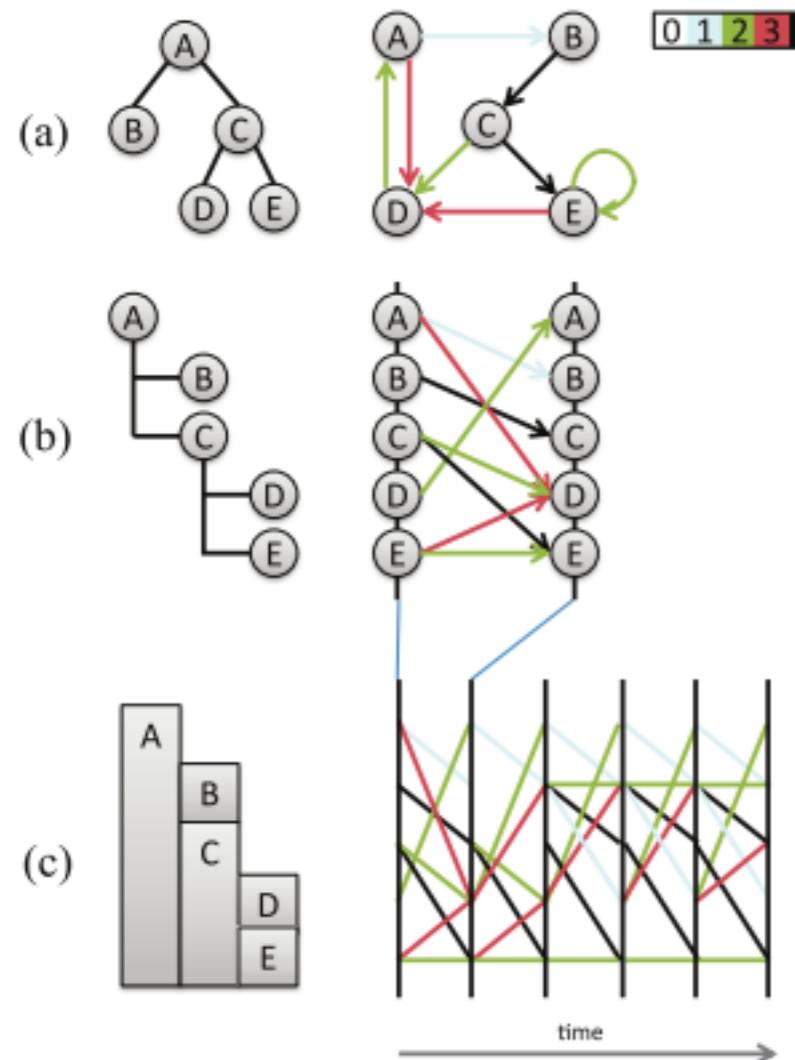


- hive plots

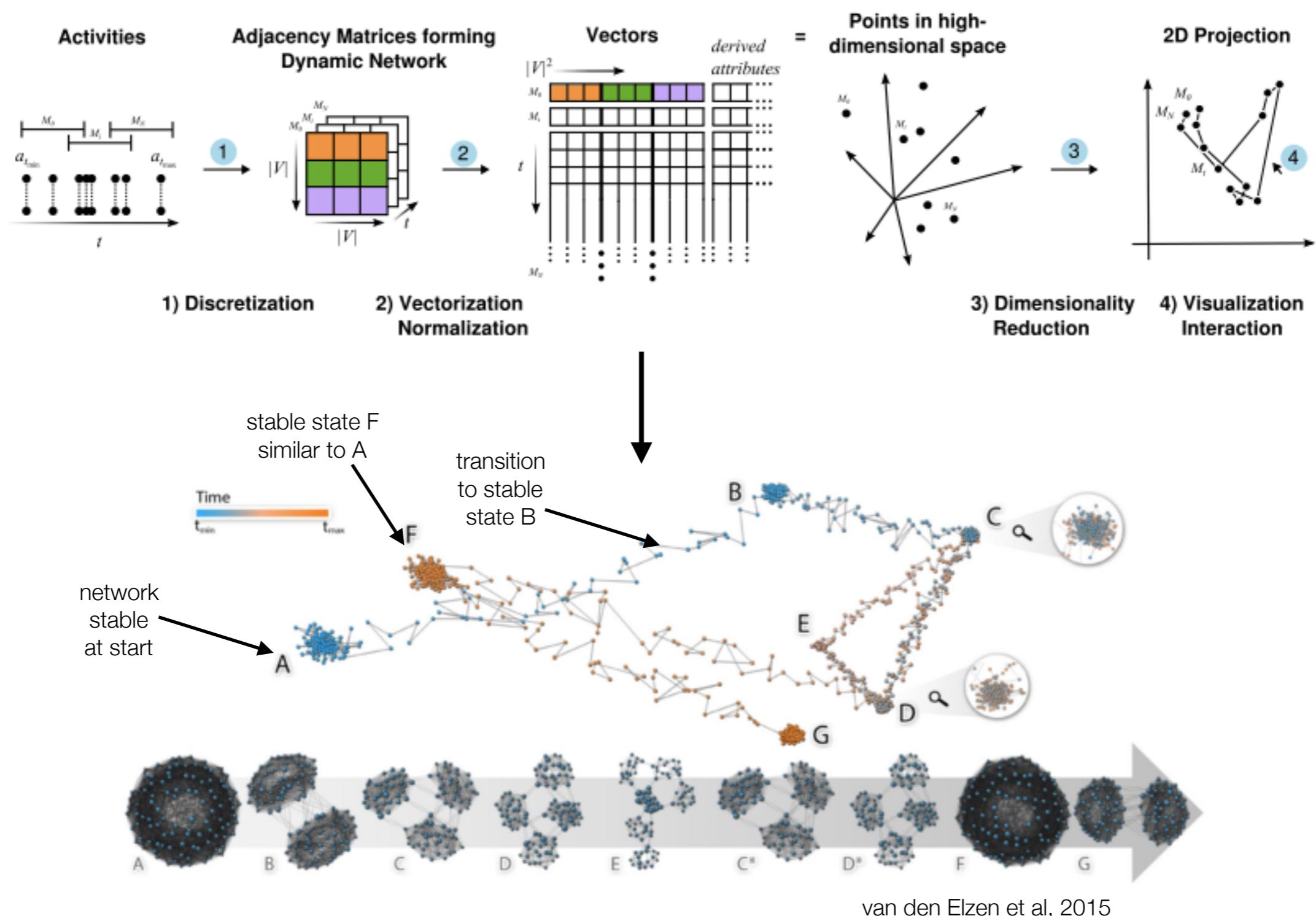


Martin Krzywinski

Parallel Edge Splatting (Burch et al, 2011)

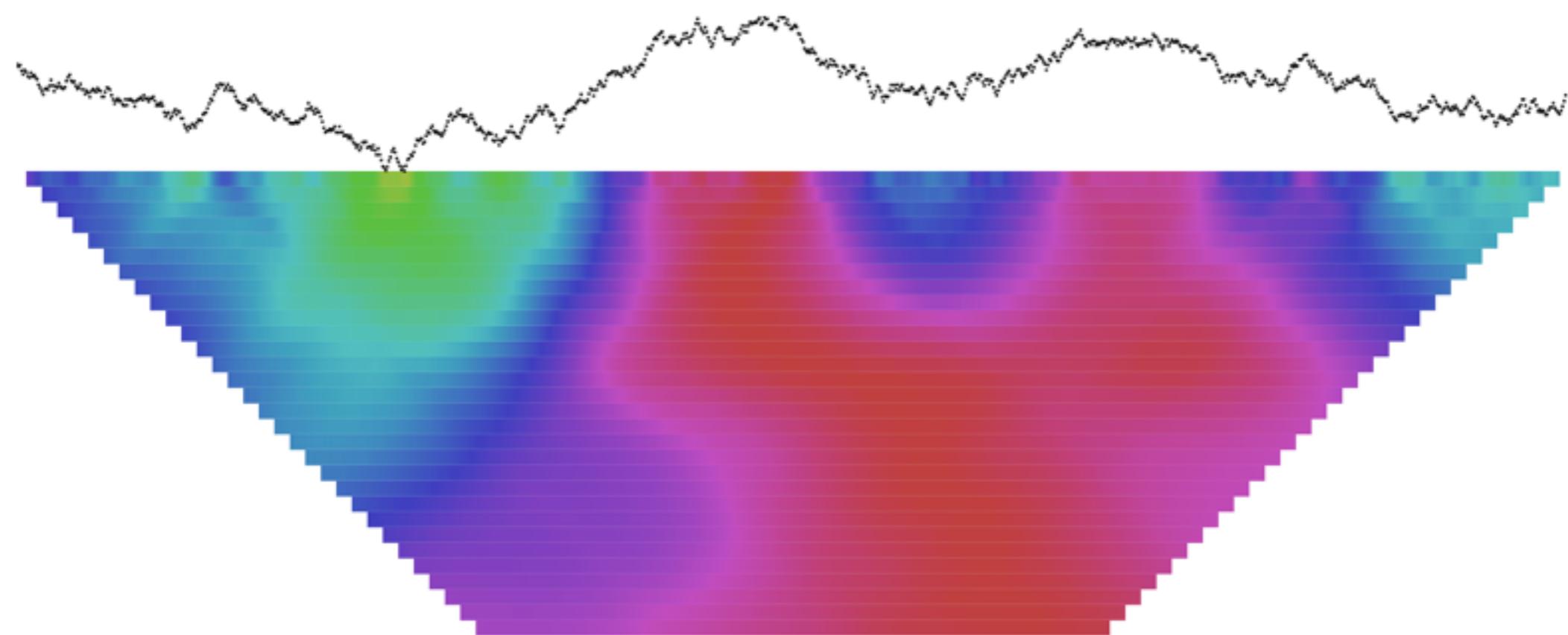


dynamic networks



van den Elzen et al, 2015

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

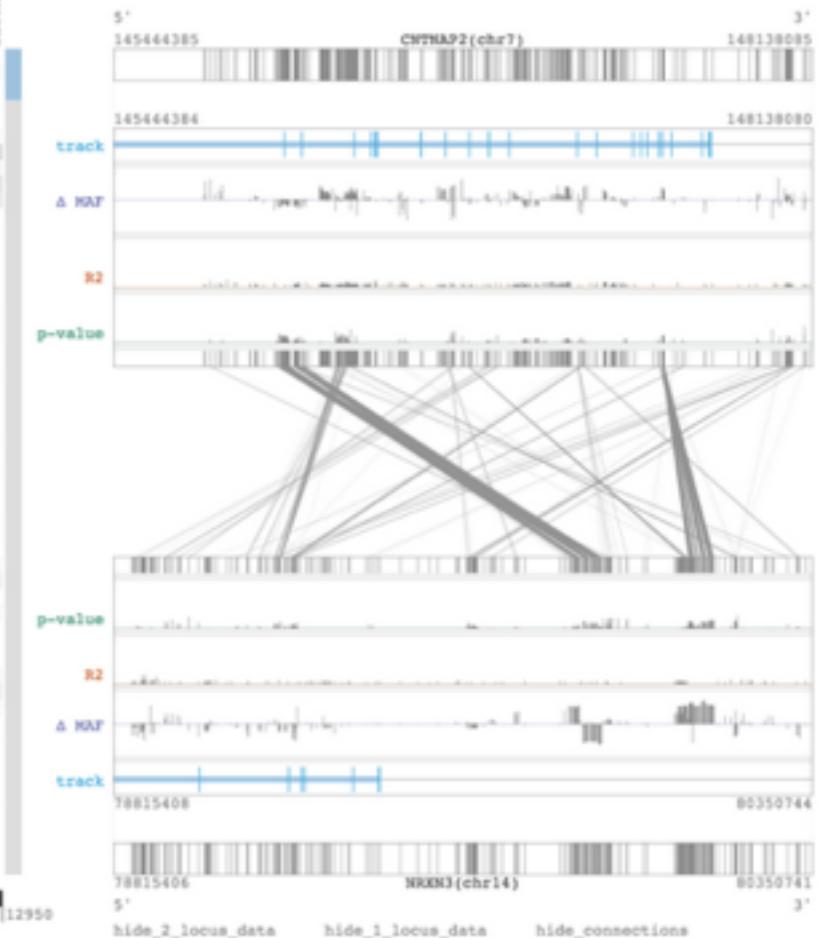
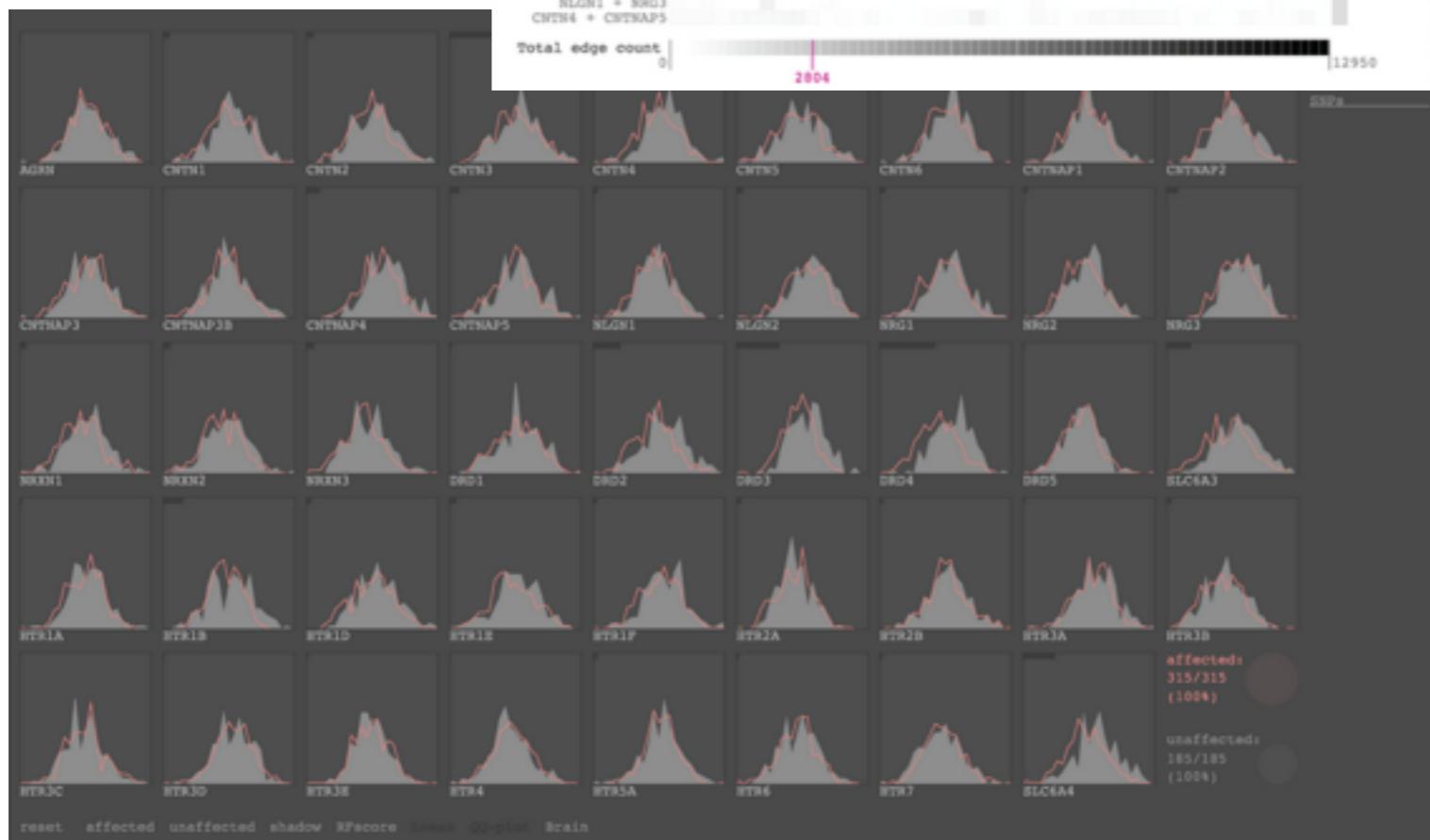




Aracari

Bartlett C et al. BMC Bioinformatics (2012)

2-locus eQTL data

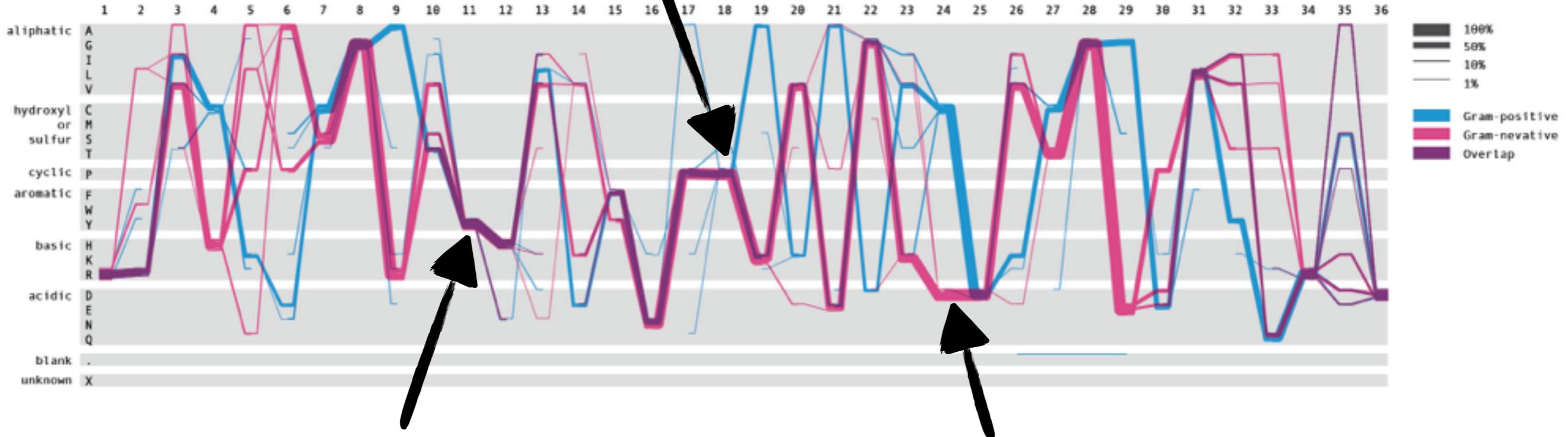
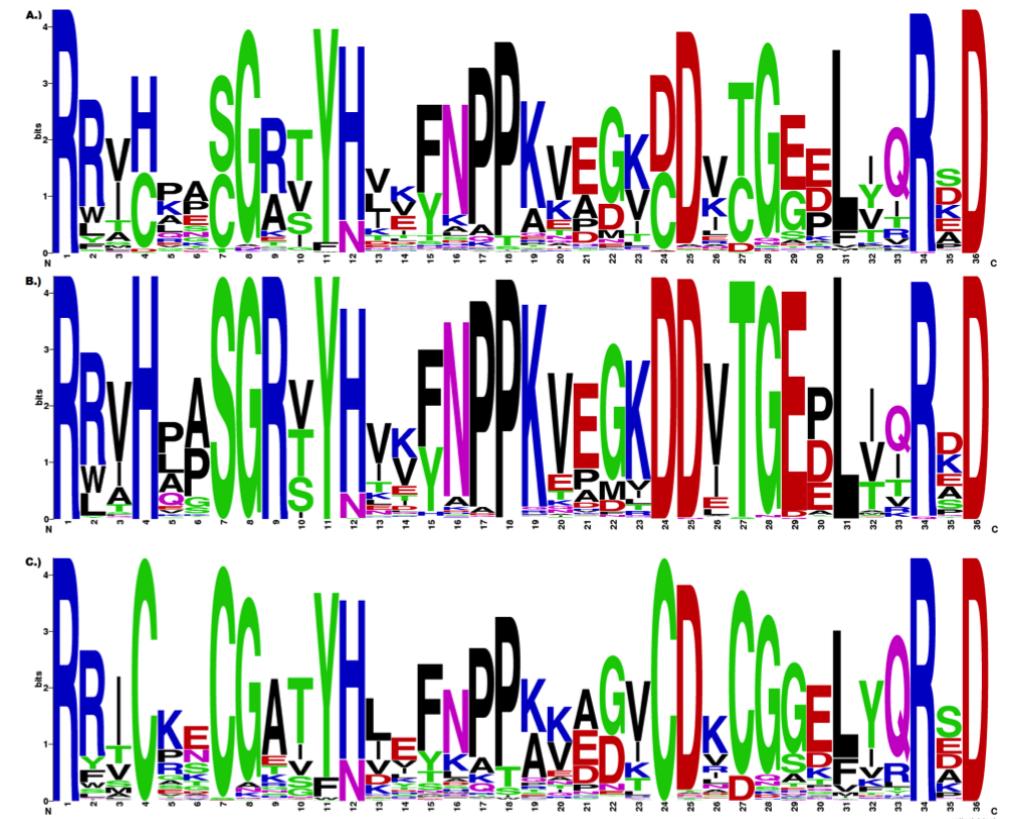


Ryo Sakai
94

Sequence Diversity Diagram



subgroup

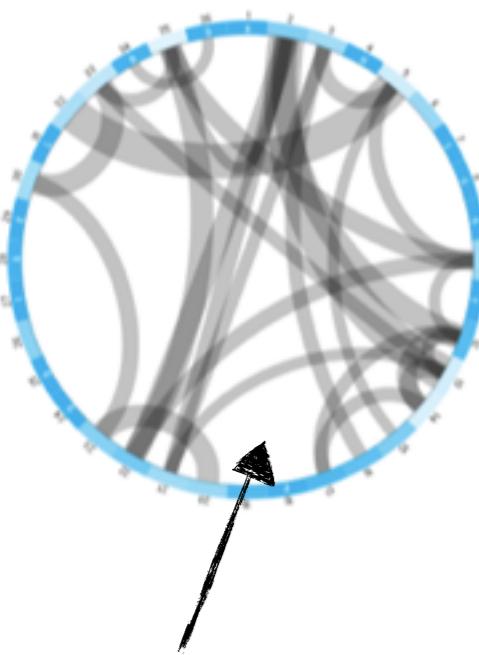
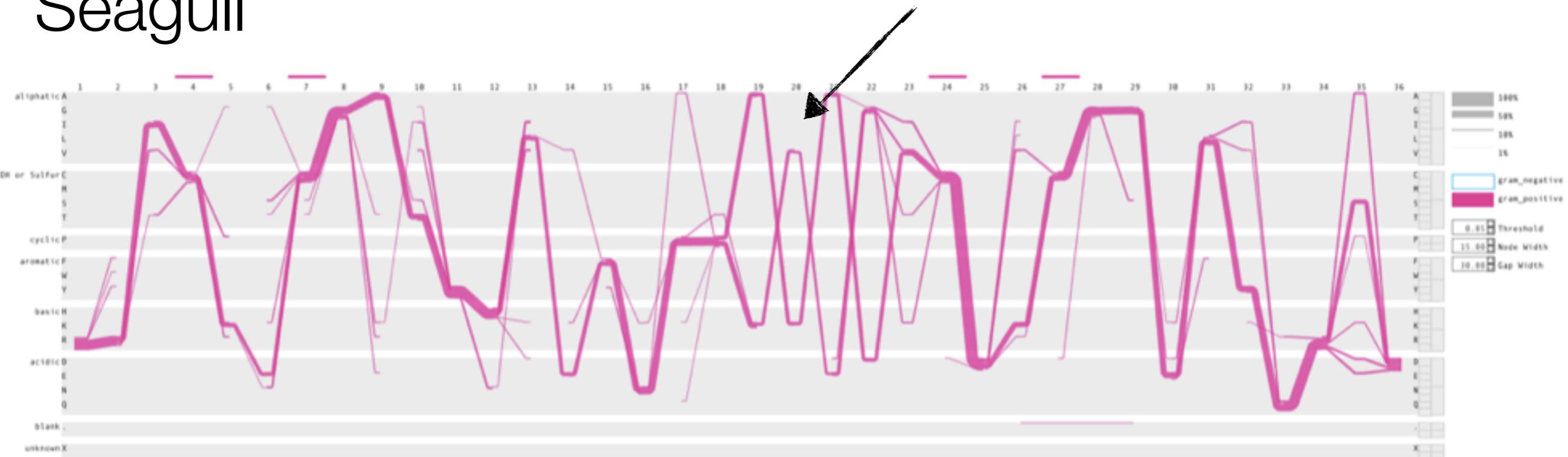


similarity

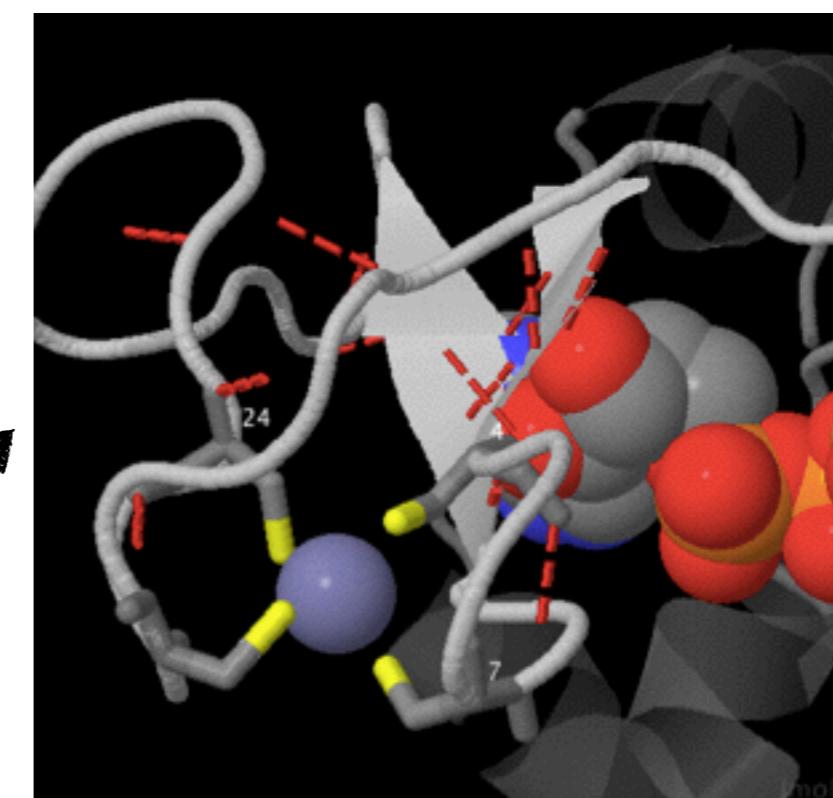
difference

Seagull

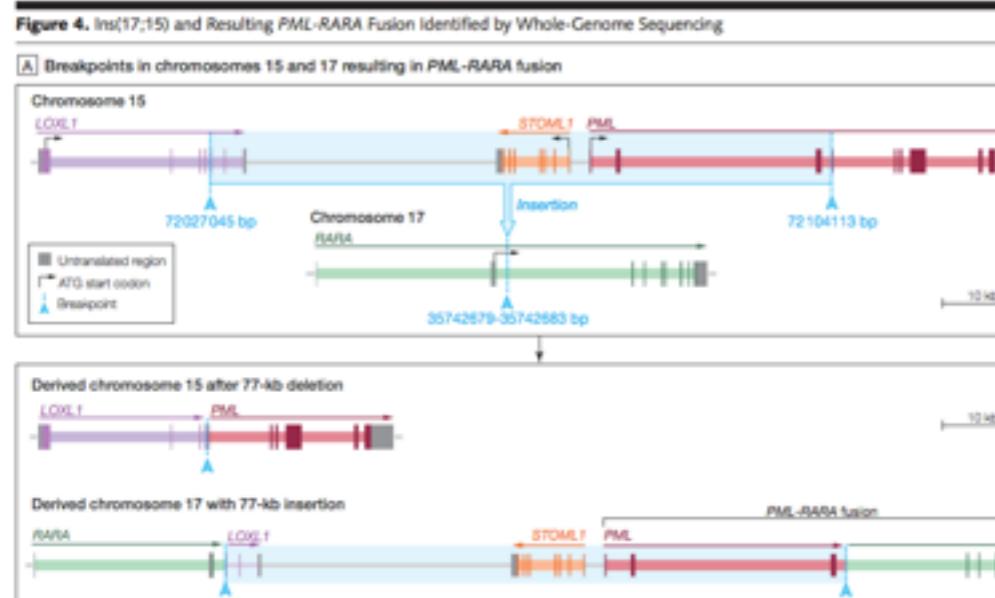
sequence diversity diagram



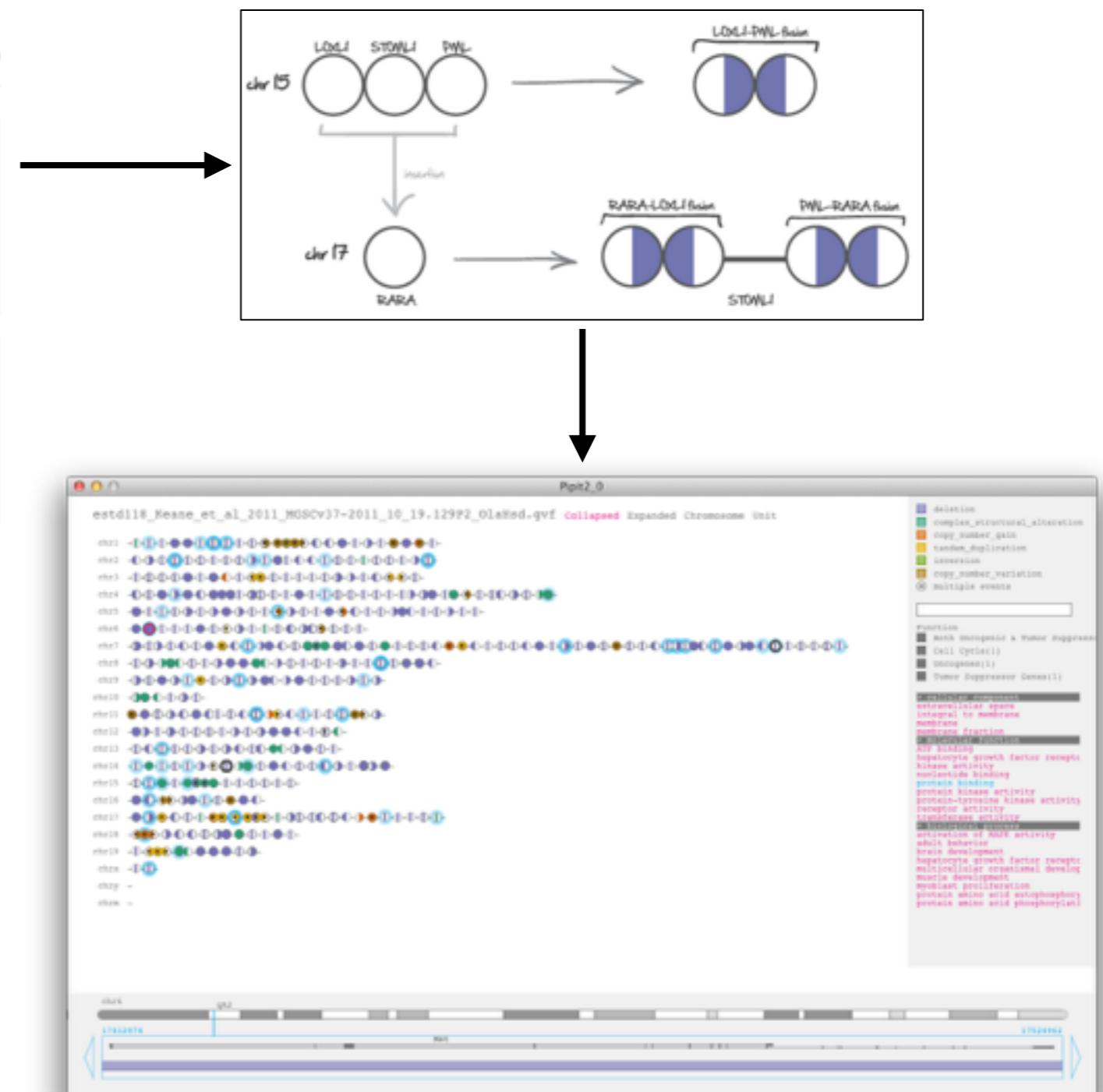
3D view



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)

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

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”

- categorize

- cluster

- rank

problem: small sample size

- user tasks:

- compare

- associate

- correlate

- identify

- locate

- distinguish

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

problem: you misunderstood their needs

abstraction: you're showing them the wrong thing

encoding: the way you show it doesn't work

algorithm: your code is too slow

Overview of lecture

A. Why visual analytics?

B. Data visualization

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

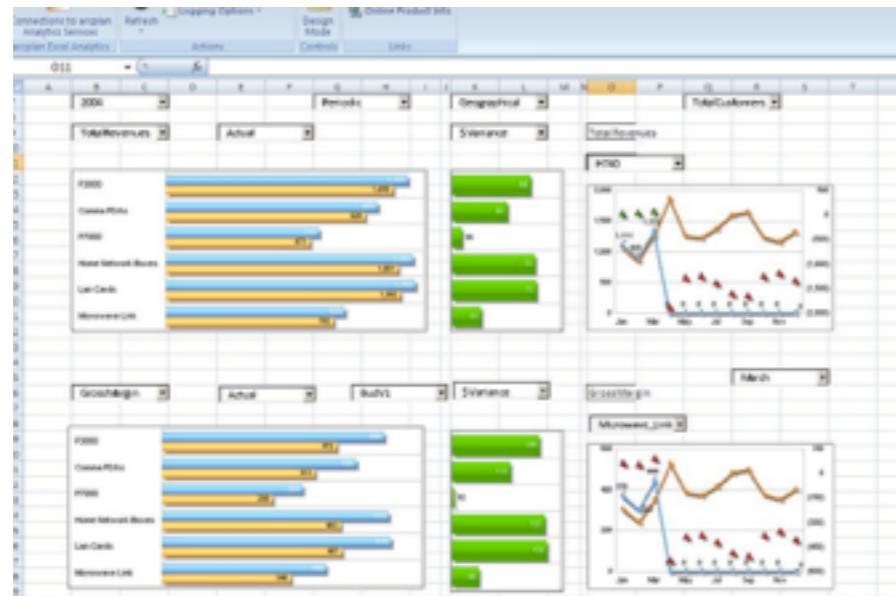
C. Visualization evaluation

D. Tools of the trade

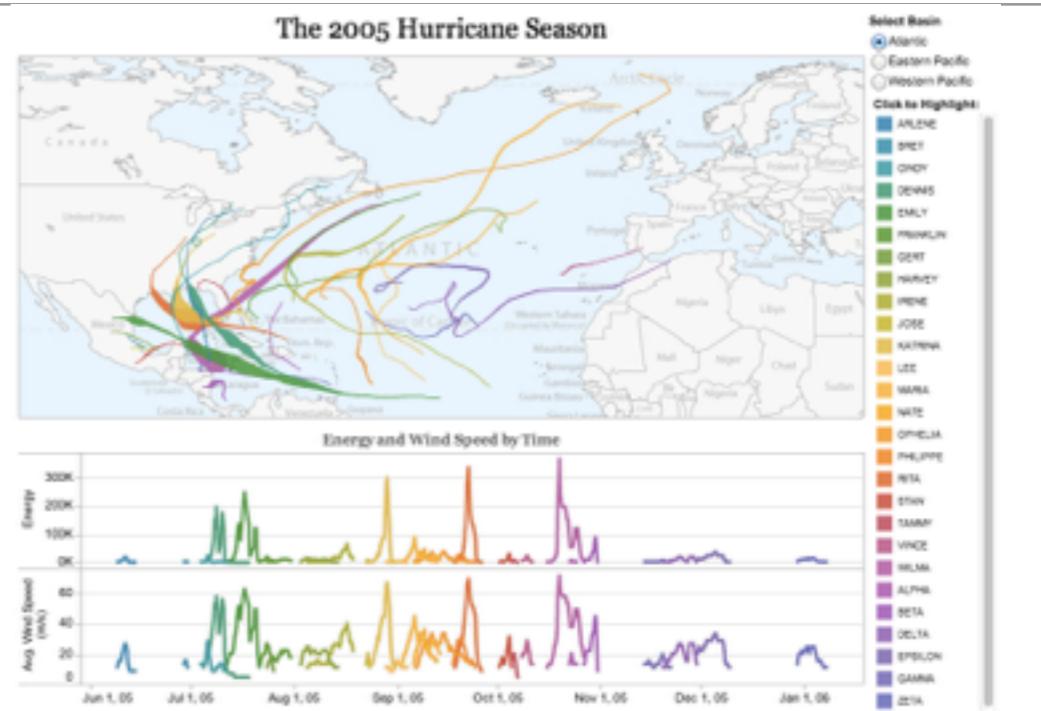
D. Tools of the trade

using
drawing
coding

Using data visualizations



Microsoft Excel



Tableau

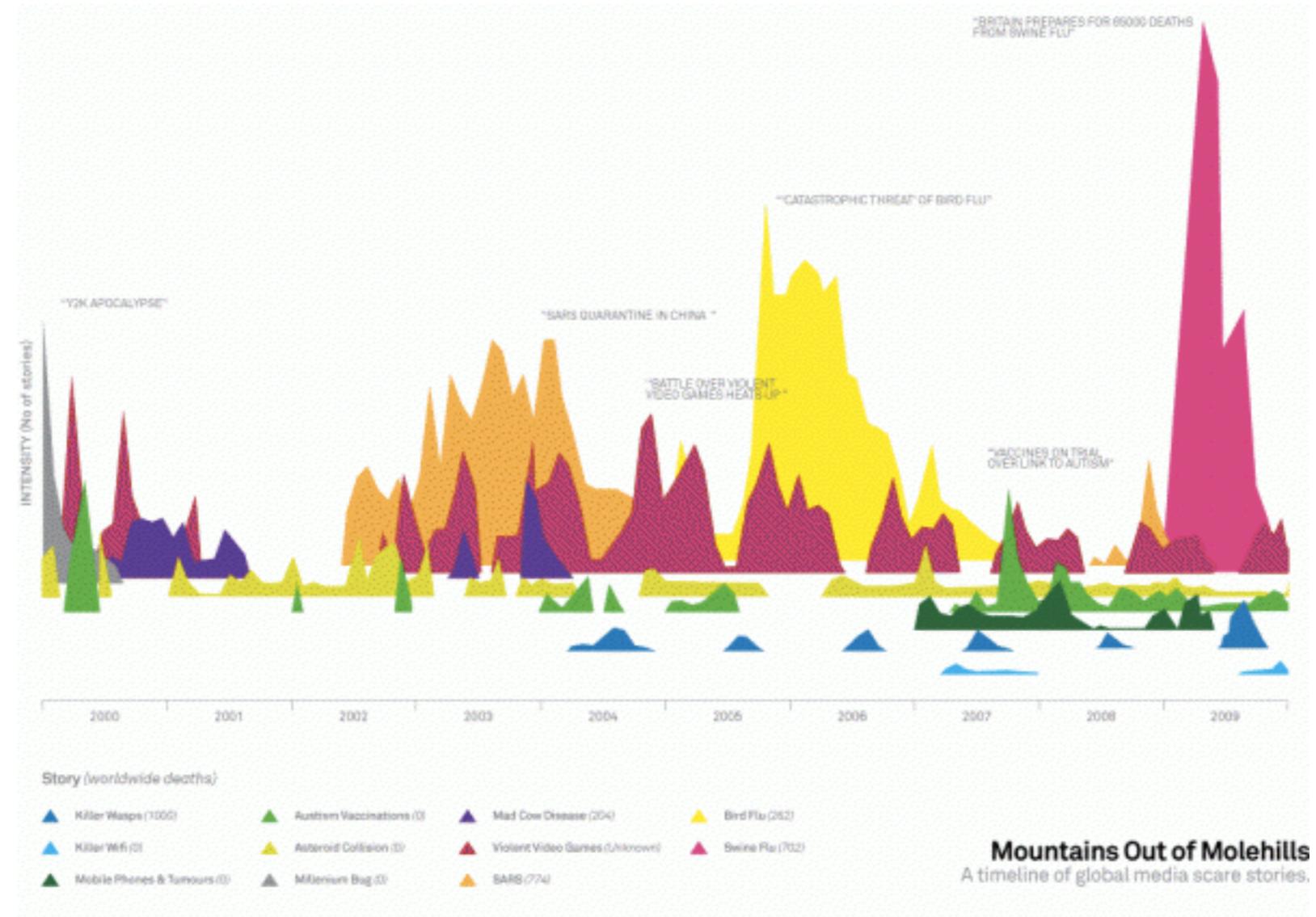


Tibco Spotfire



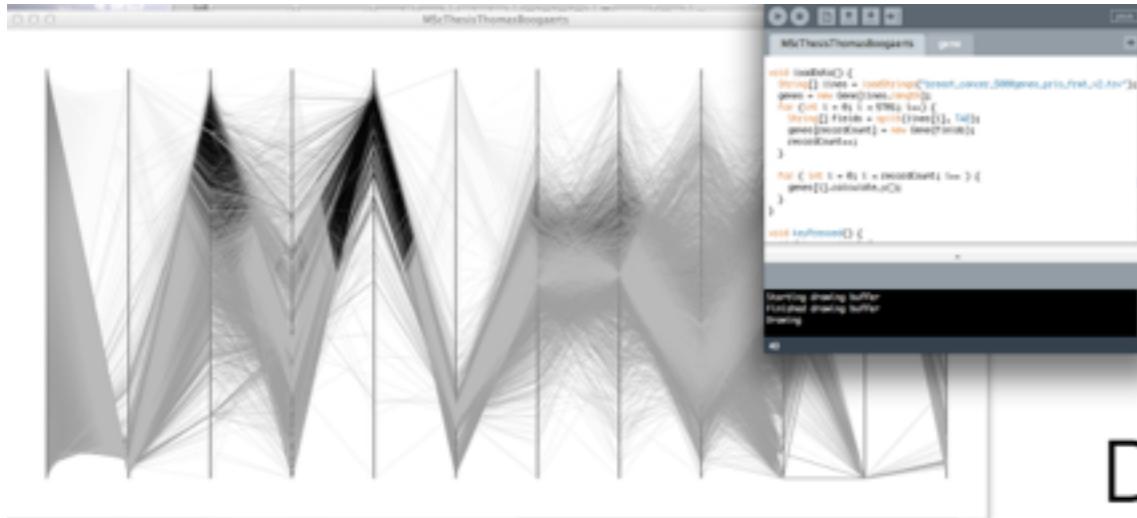
ManyEyes

Drawing data visualizations

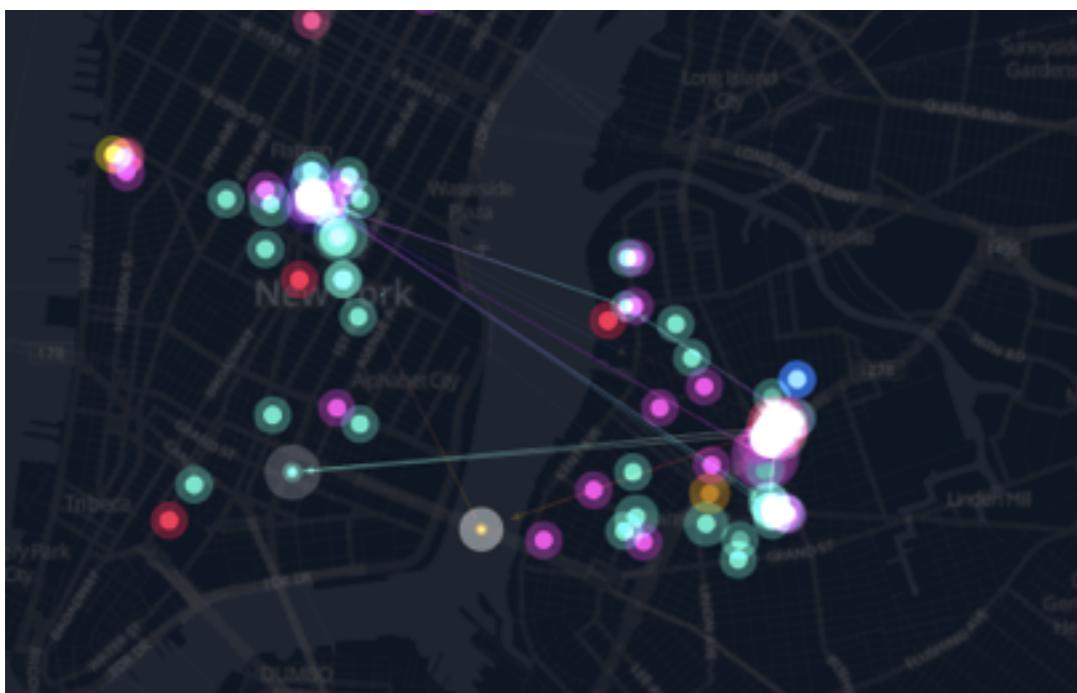


Adobe Illustrator

Coding data visualizations

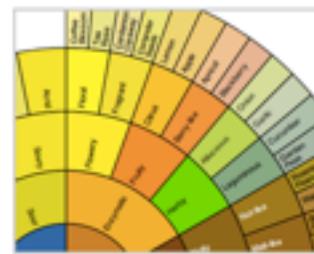
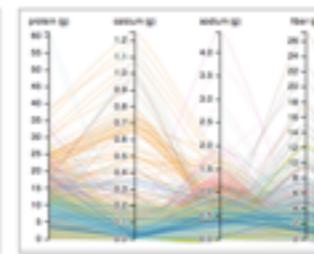
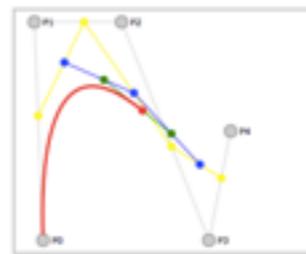
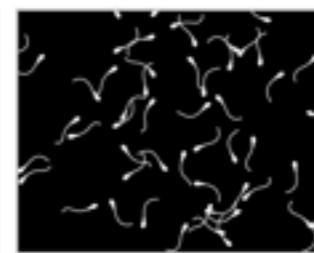
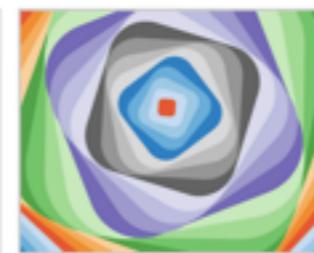
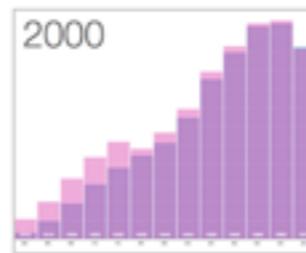


processing.org



paper.js

Data-Driven Documents



d3.js

LOONEY TUNES



That's all Folks.

(...actually, we only scratched the surface...)