[CODATA-RDA] Introduction to **Visual Data Analysis**

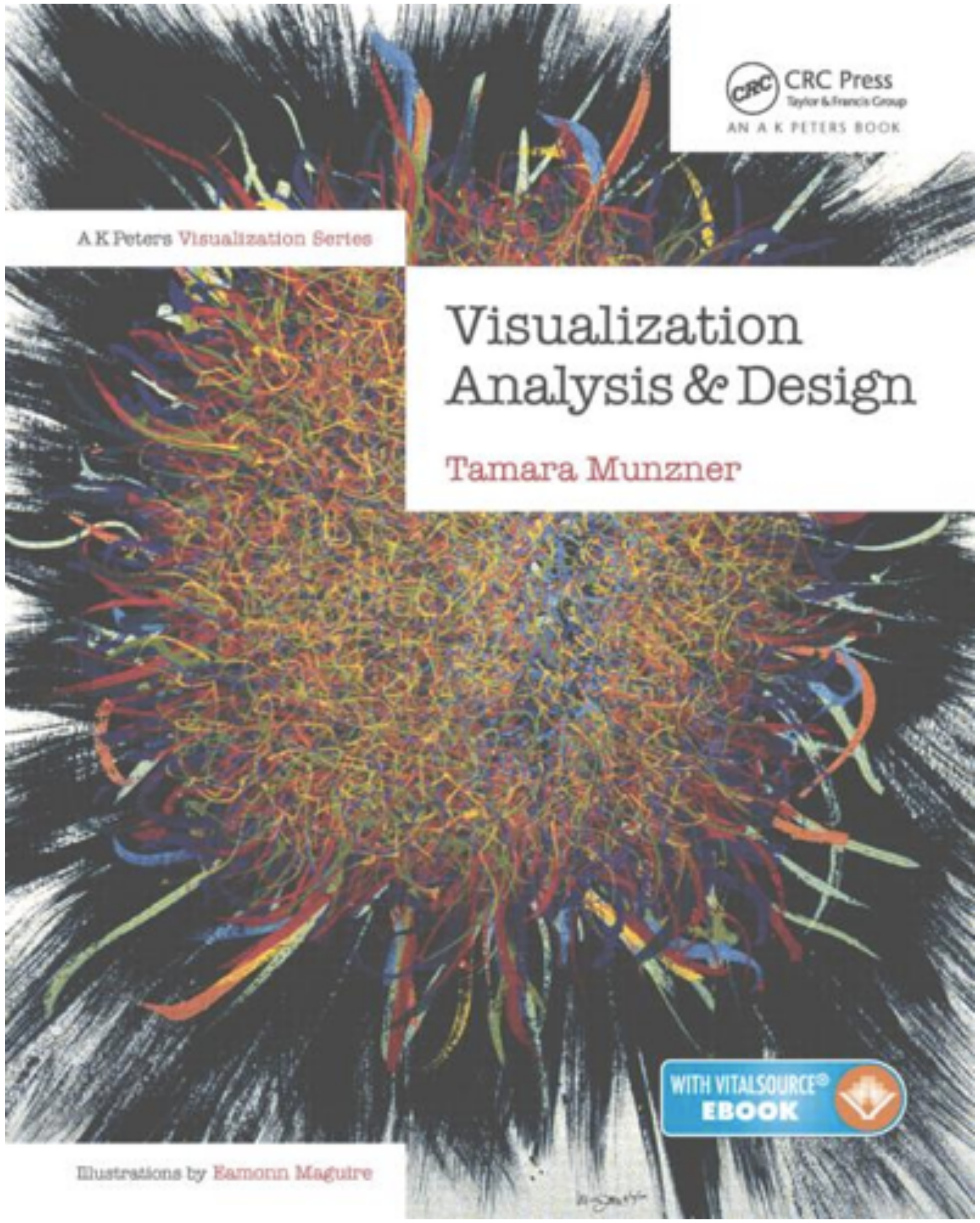
Prof Jan Aerts
Visual Data Analysis lab, ESAT/STADIUS
Faculty of Engineering
KU Leuven

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

KU LEUVEN

T Munzner - Visualization Analysis & Design

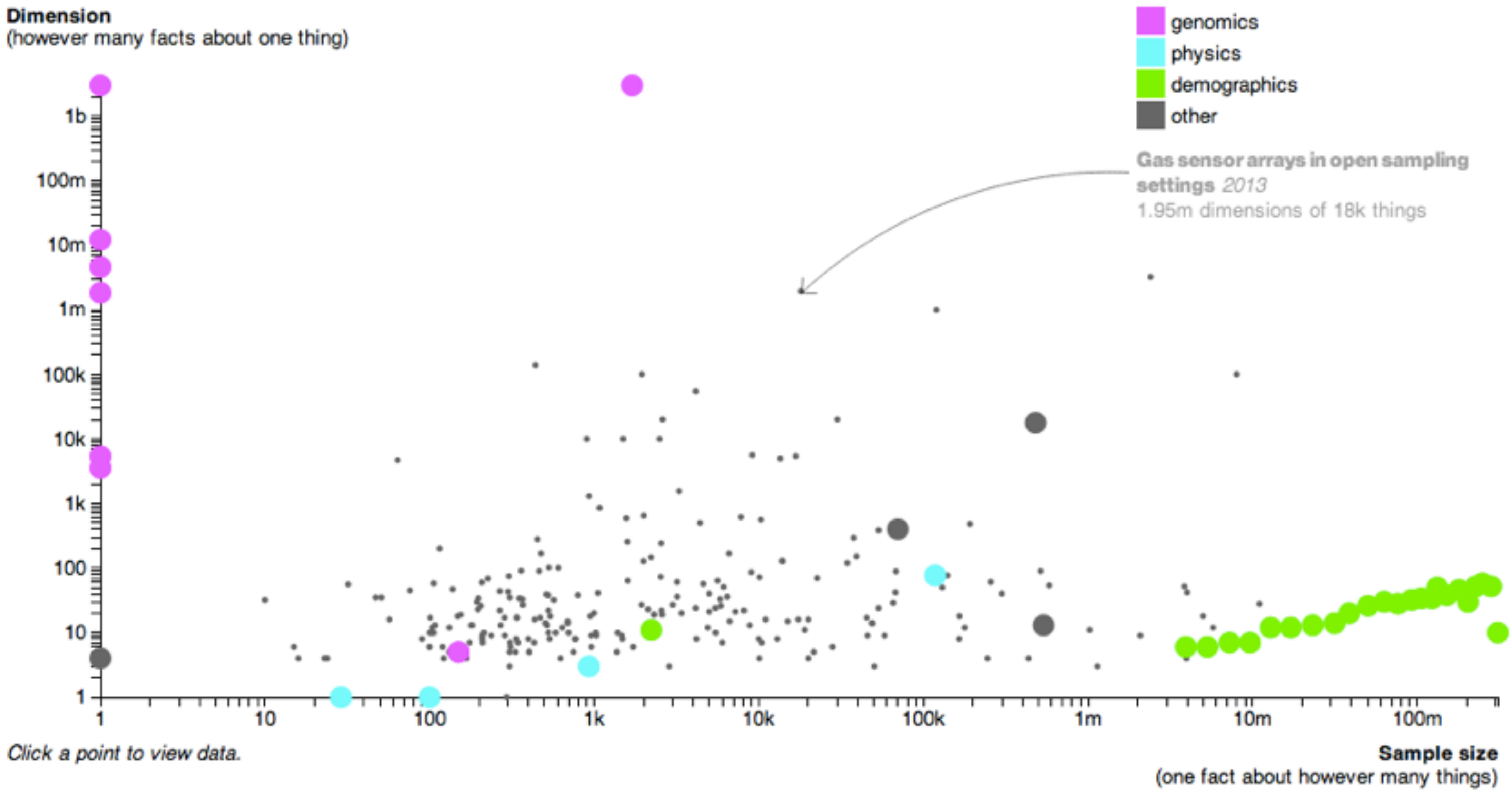
The “bible” for data visualization => go and buy it



What we'll talk about

A. What's the problem?
B. What is data visualization? (incl perception/cognition)
C. Models for dataviz
    1. What
    2. Why
    3. How
D. Bad graphics
E. Design process
F. Examples

[A] What's the problem?

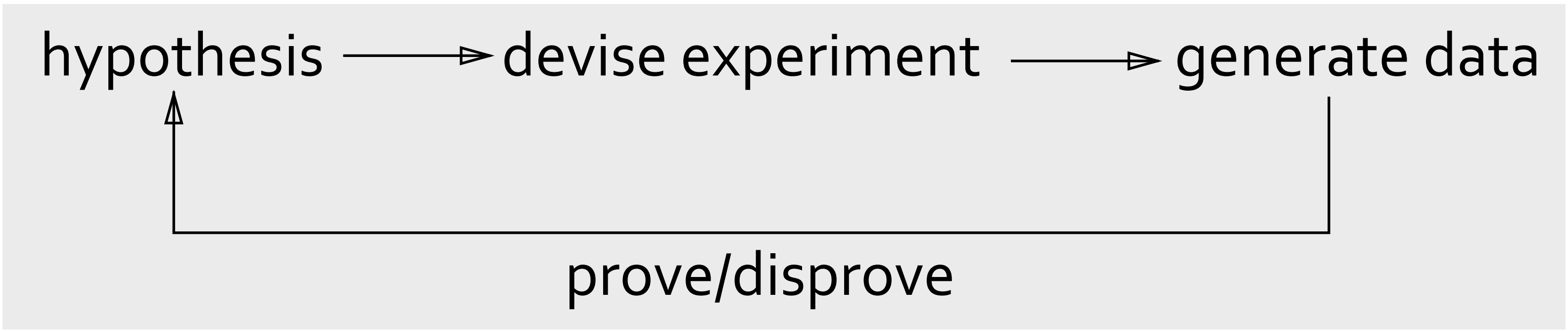

Dimension
(however many facts about one thing)
genomics
physics
demographics
other
Gas sensor arrays in open sampling settings 2013
1.95m dimensions of 18k things
1b
100m
10m
1m
100k
10k
1k
100
10
1
1
10
100
1k
10k
100k
1m
10m
100m
Click a point to view data.
Sample size
(one fact about however many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

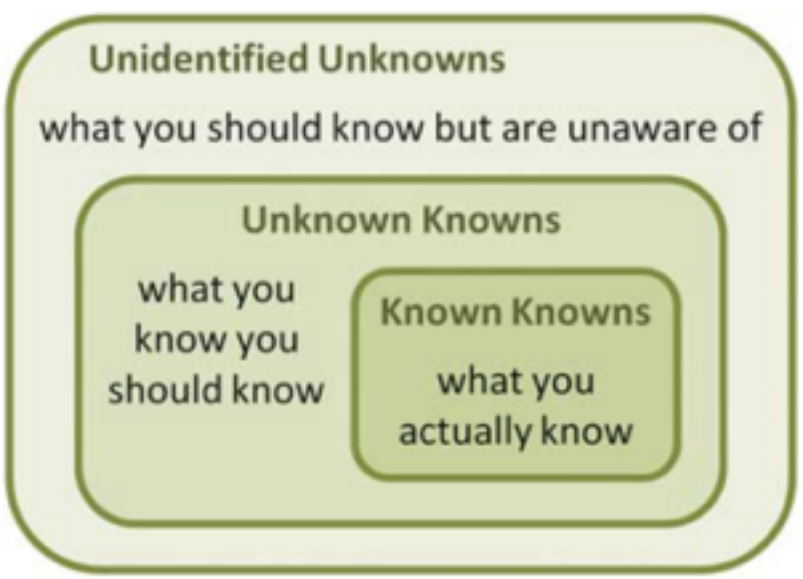
then:



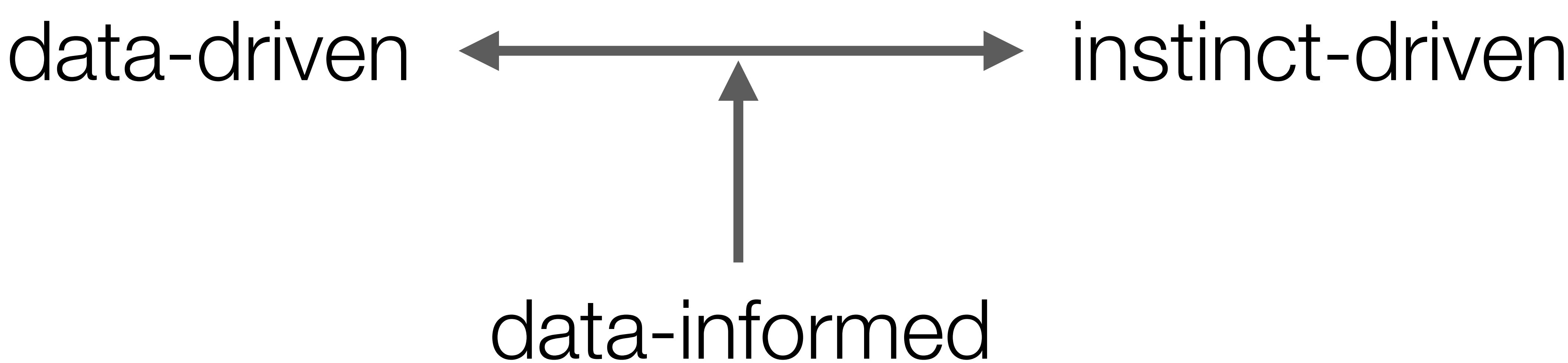
now:

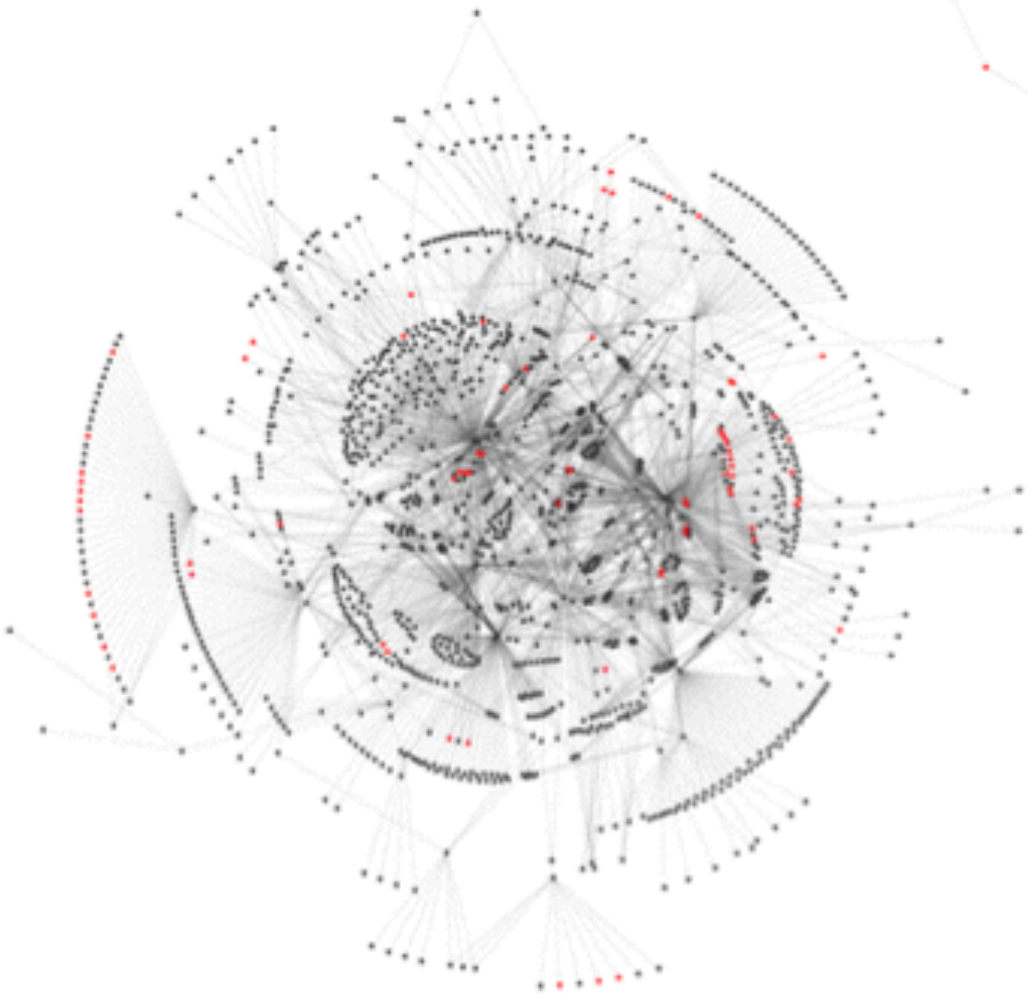
<u>Data analysis</u>
moving from
hypothesis-first
to
data-first

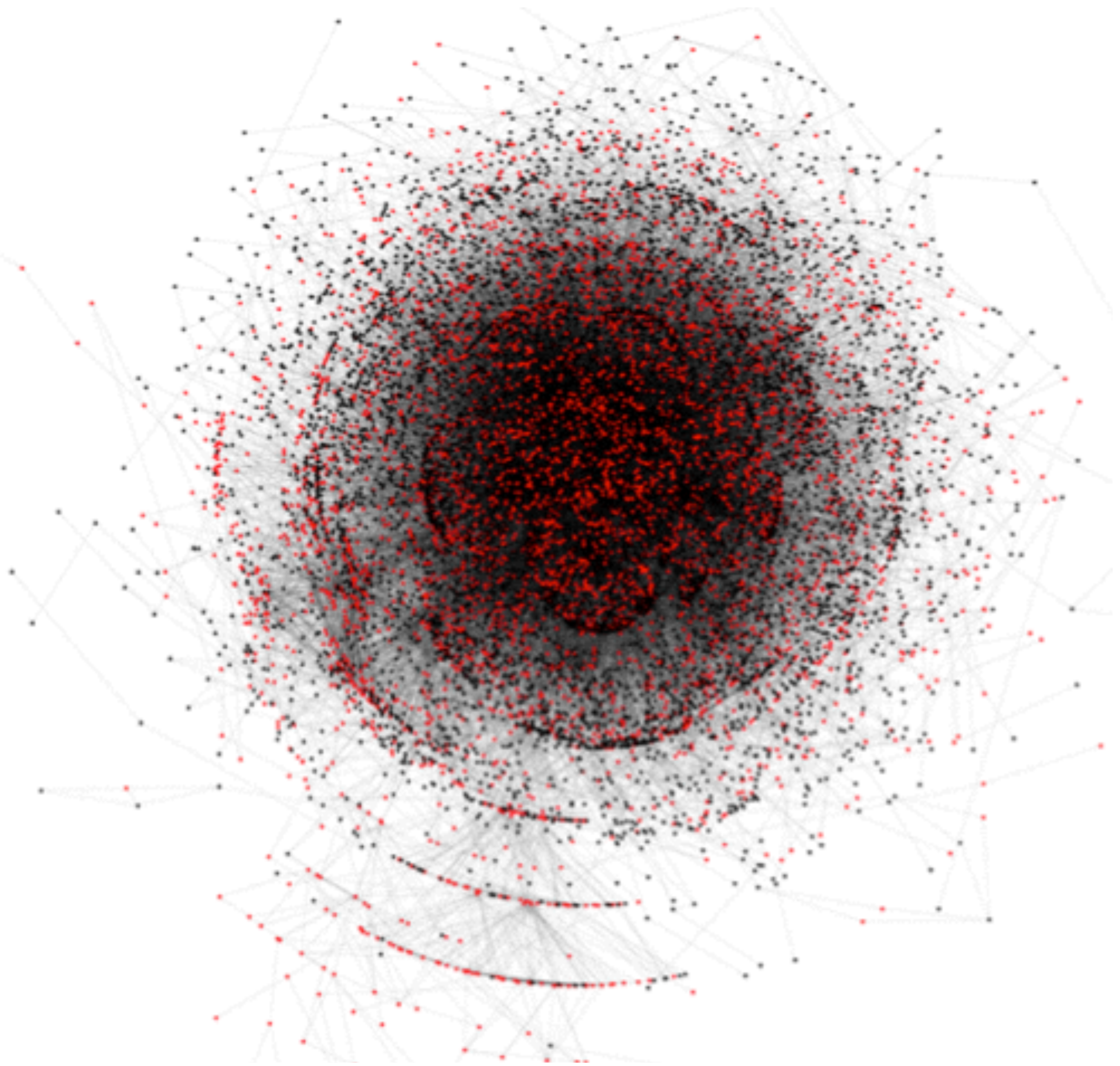
<u>Challenge</u>
moving from
finding the right answer to a question
to
finding the right question given the data



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


data-driven
instinct-driven
data-informed





Martin Krzywinski

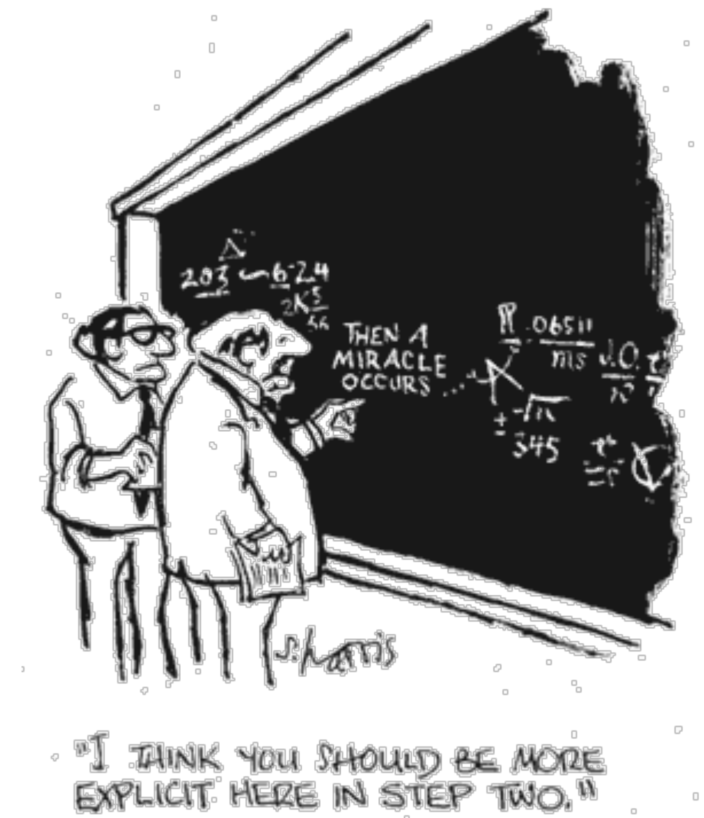
Opening the black box

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)

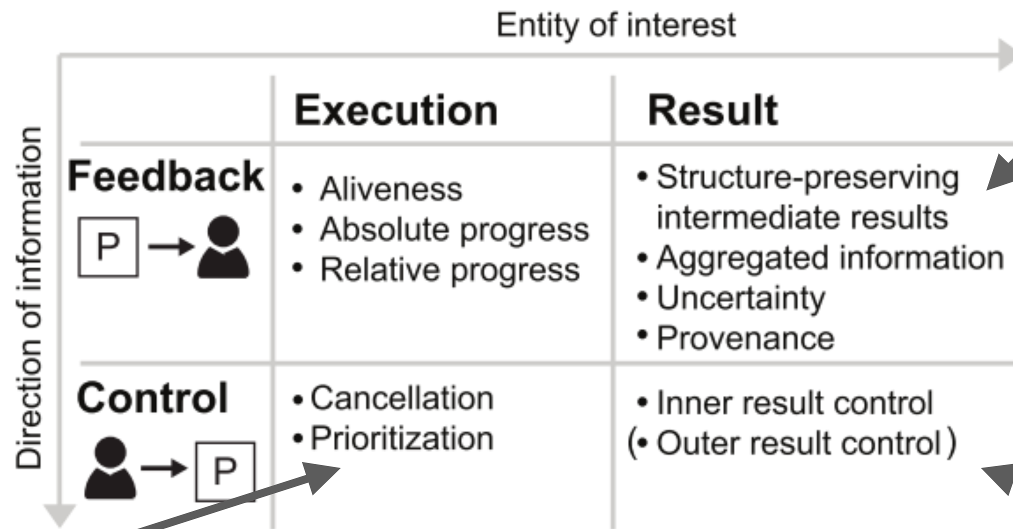
result:
algorithm acts as a
black box



insight
in data
in algorithms

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

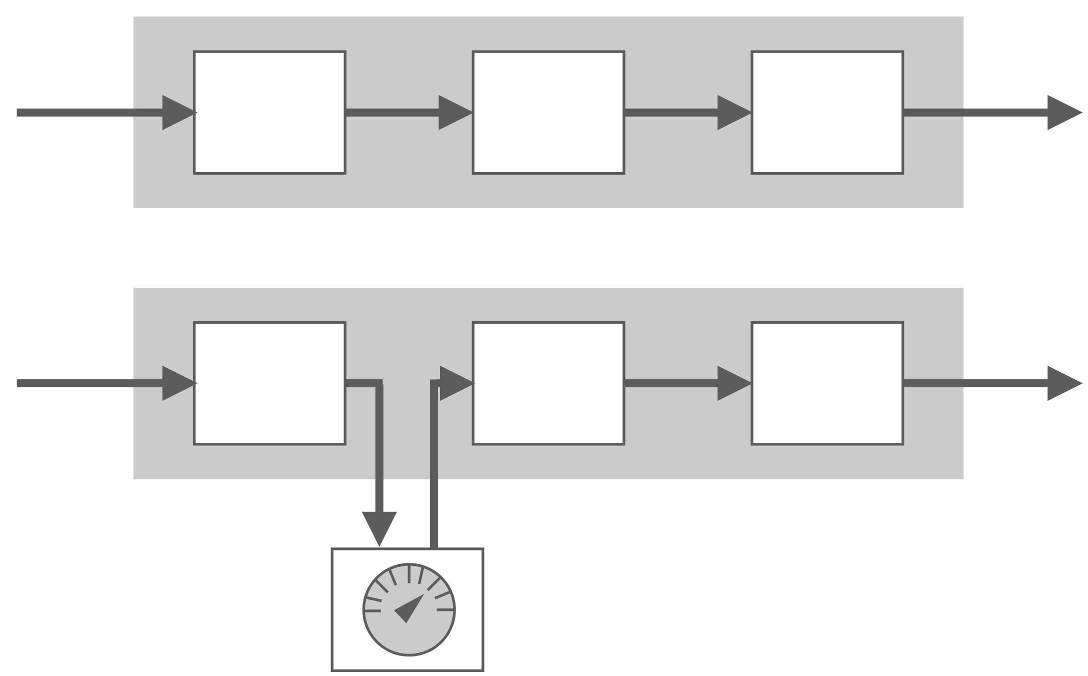
intermediate feedback: e.g. intermediate result, goodness-of-fit for best subset so far



Muhlbacher et al, 2014

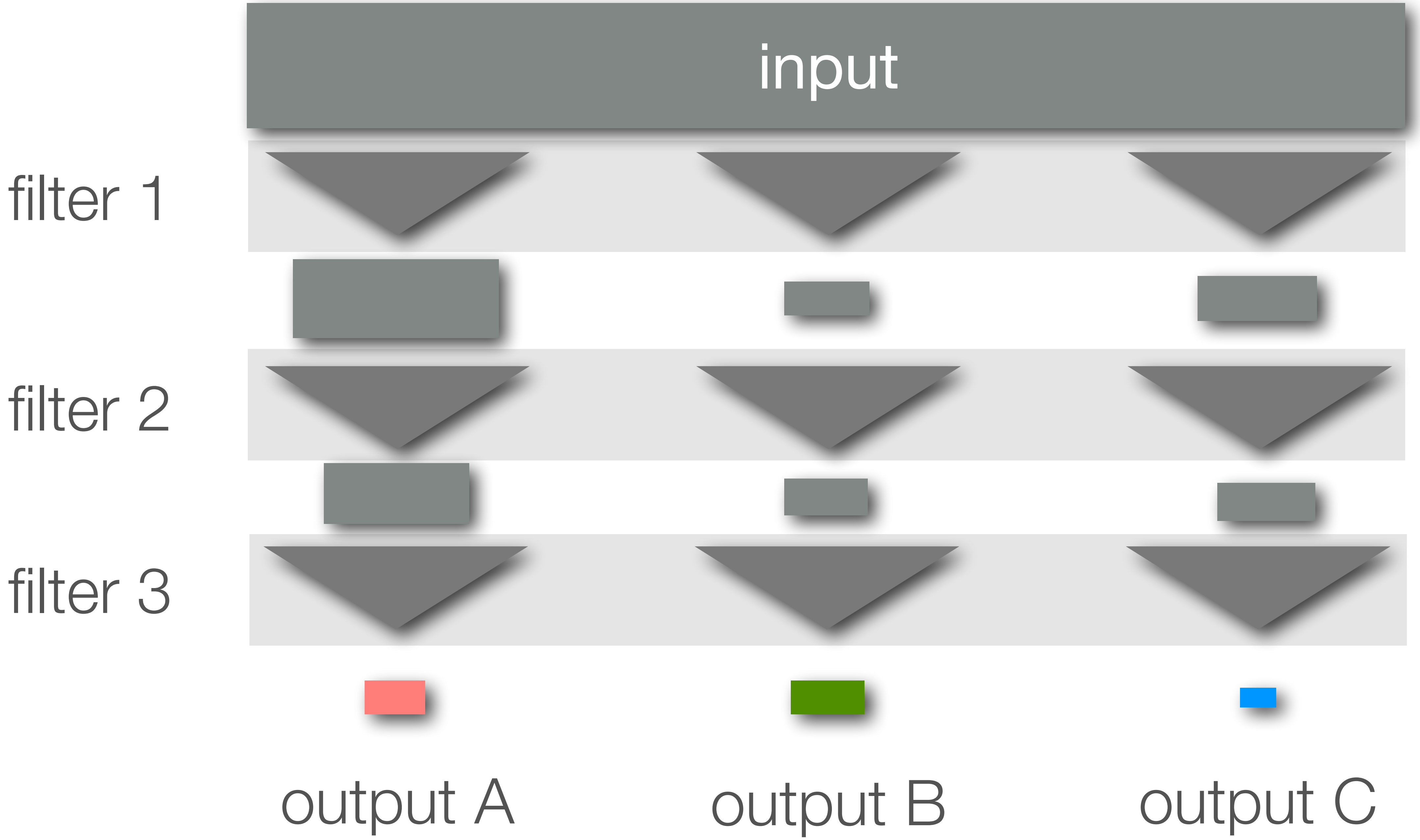
steer the final result: early validation of intermediate results, guided feature selection, weighting, avoiding being stuck in local minima, …

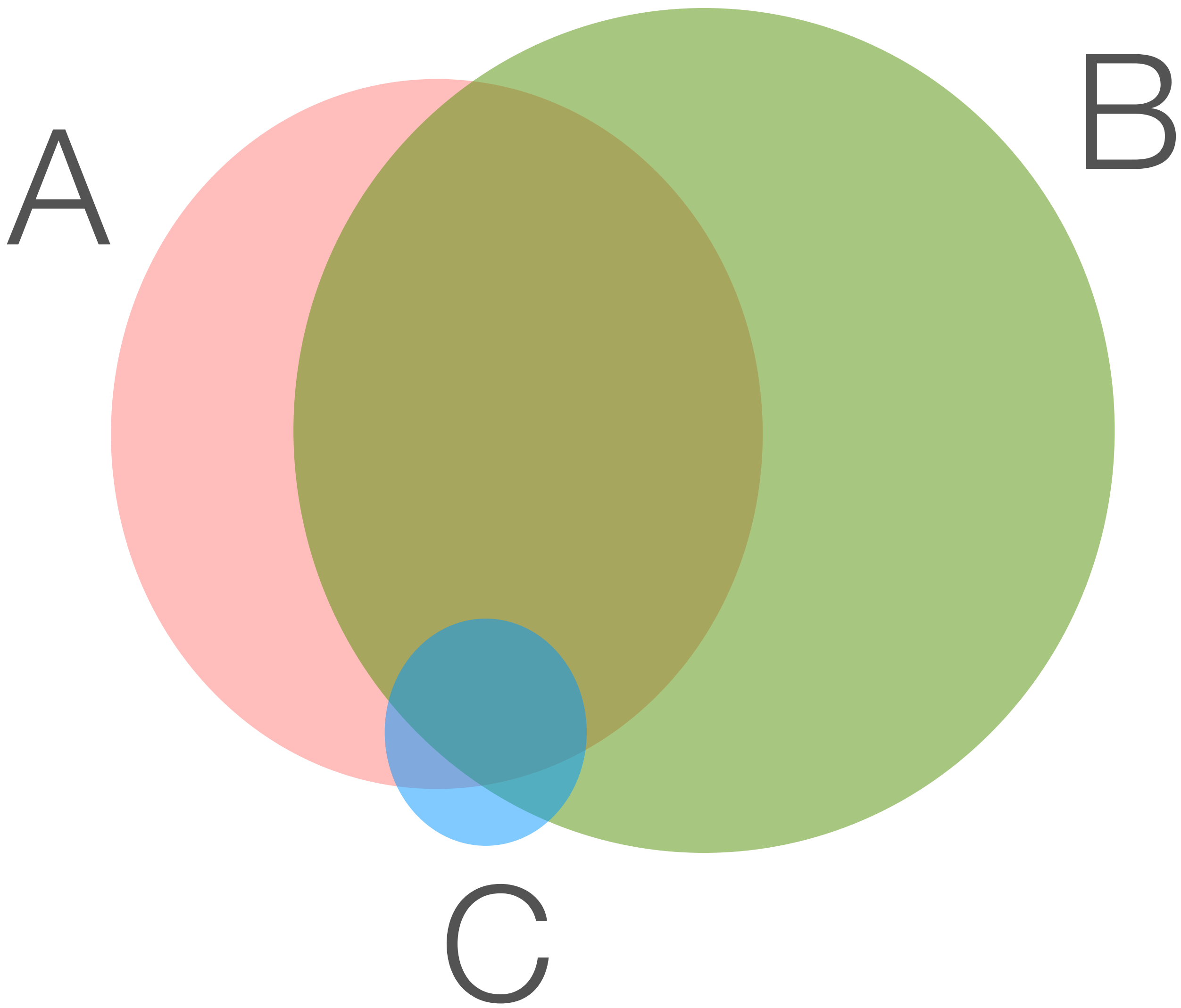
prioritization of remaining work: alter sequence of intermediate results to generate presumably more interesting ones earlier

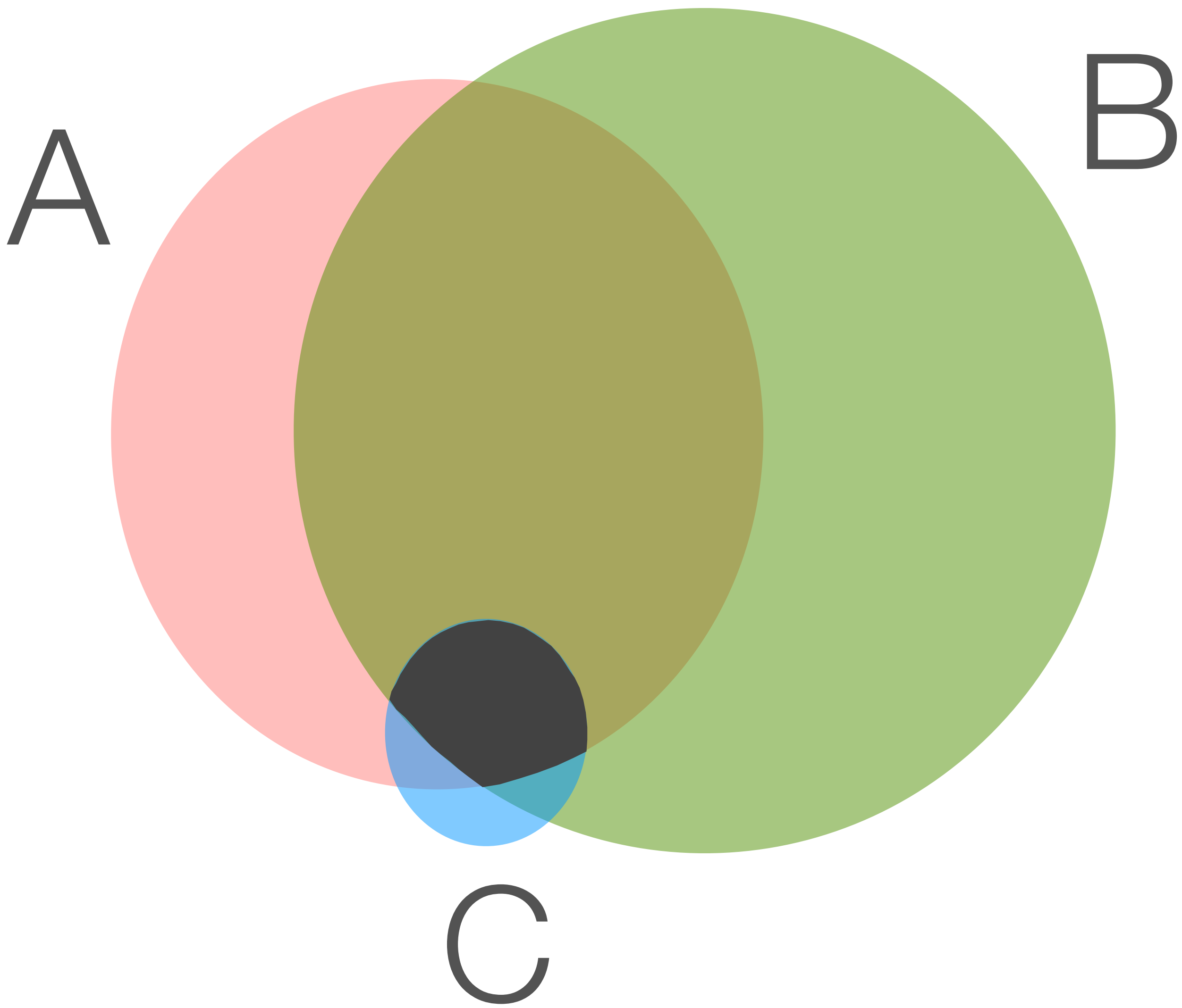


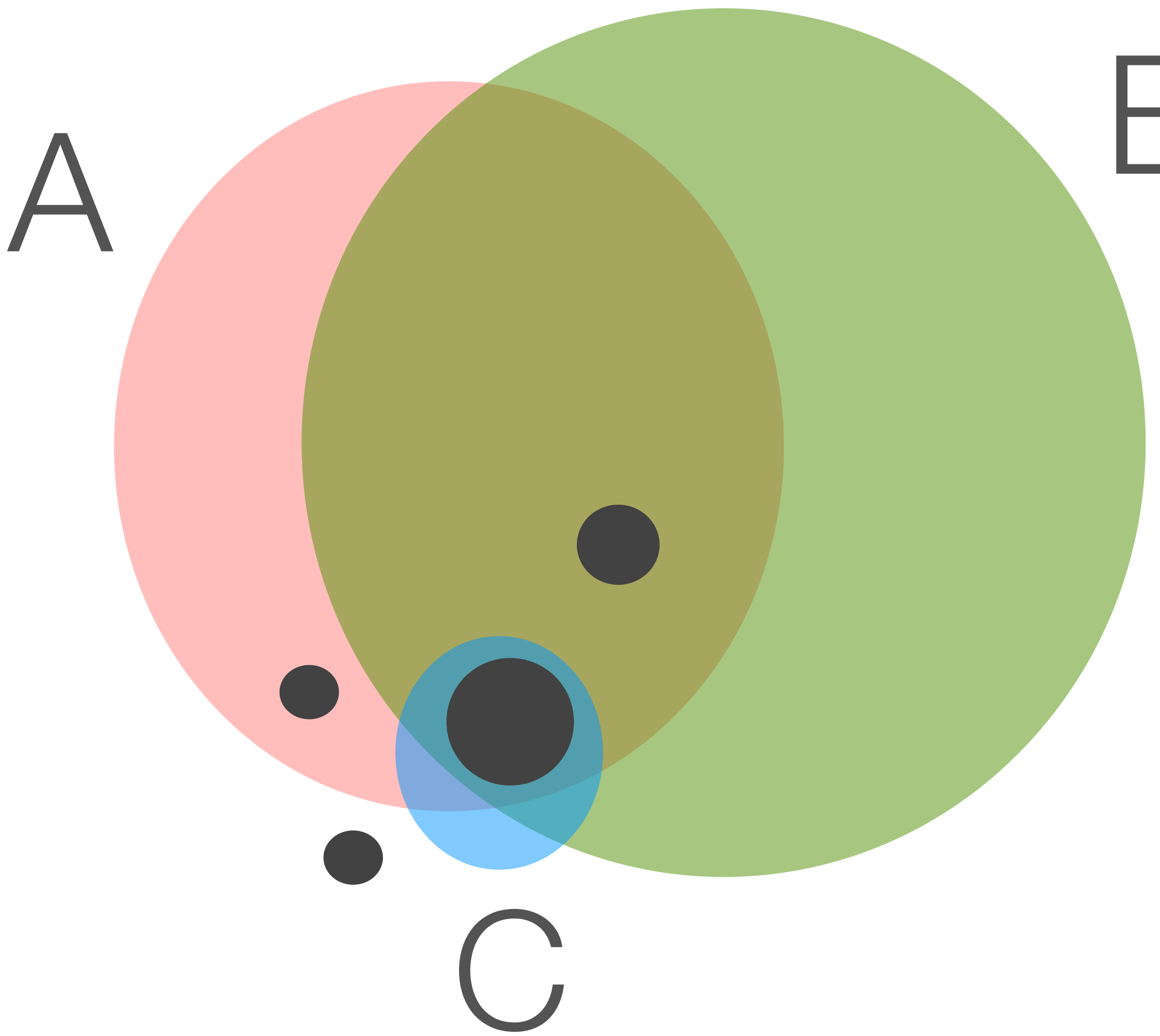
Example: mutation filtering

going from 5-6 million to 20,000




A
B
C

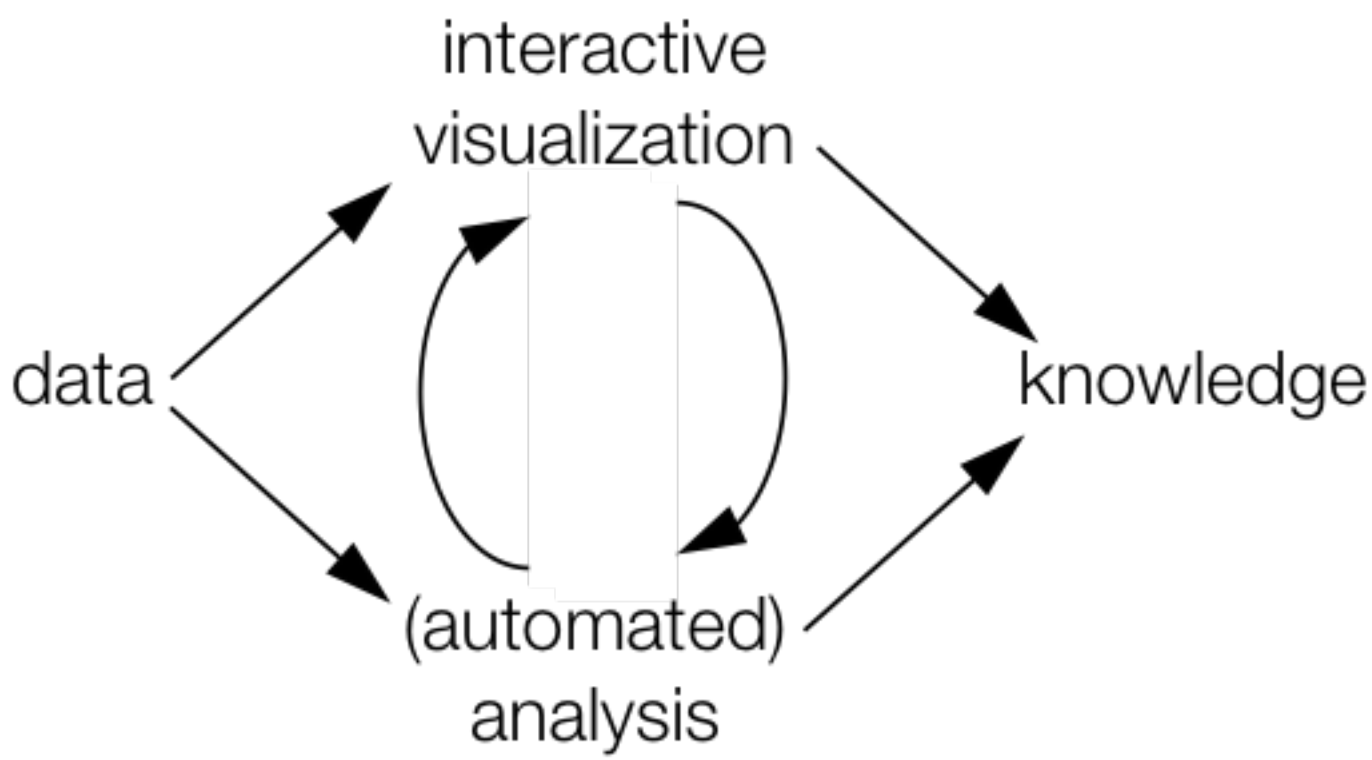

A
B
C


A
B
C

Visualization to the rescue

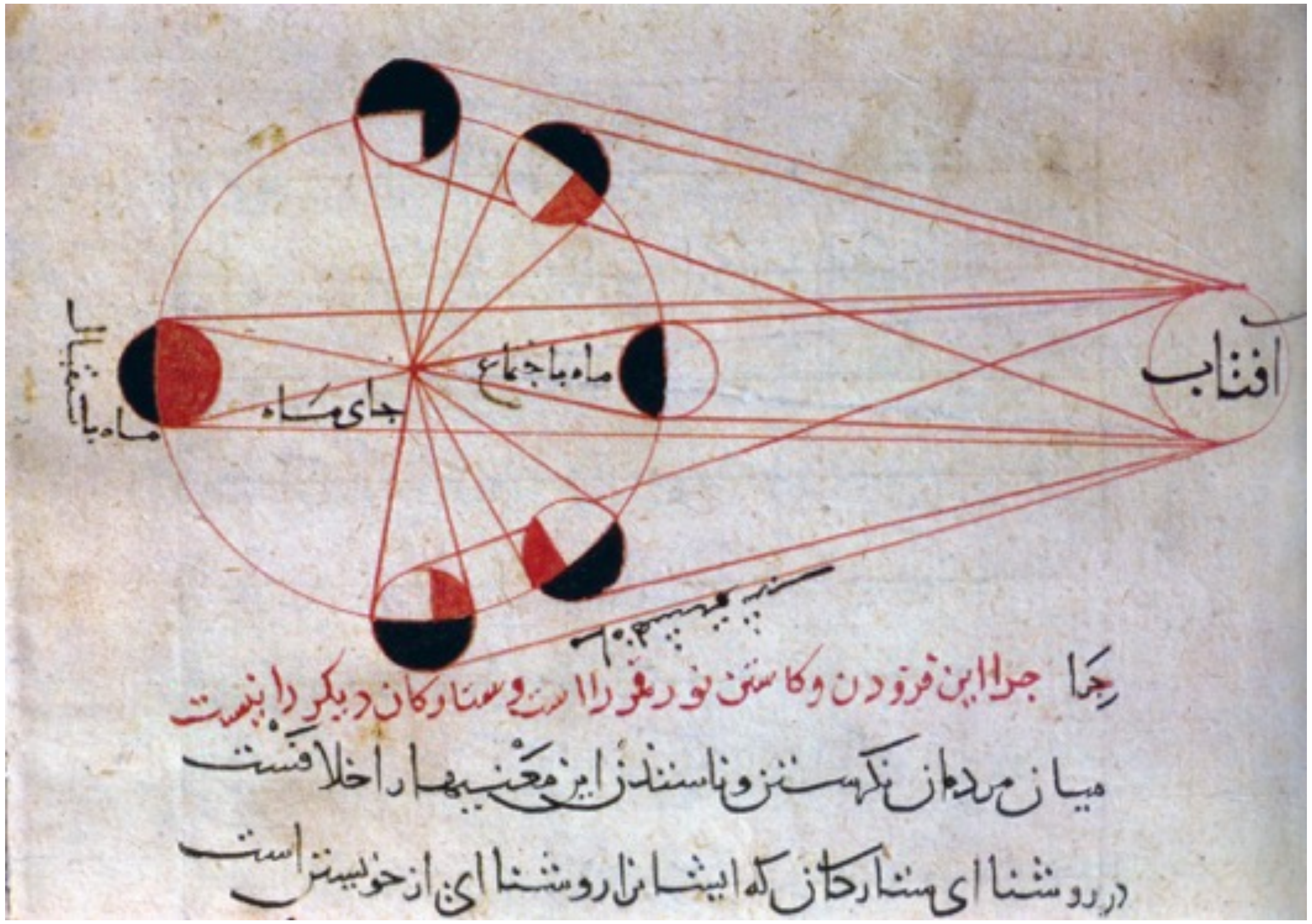
- visualization = **“trading zone”** (Peter Galison, Harvard); similar to how different cultures are able to exchange goods, despite differences in language and culture
- Might need the help of an **agent** = someone with enough knowledge of both languages/cultures to be able to translate and facilitate this trade
- trading zone between:
  - data and scientist
  - scientist (biologist) and scientist (engineer)
  - scientist and public

Visual Analytics for exploration



[B] Data Visualization

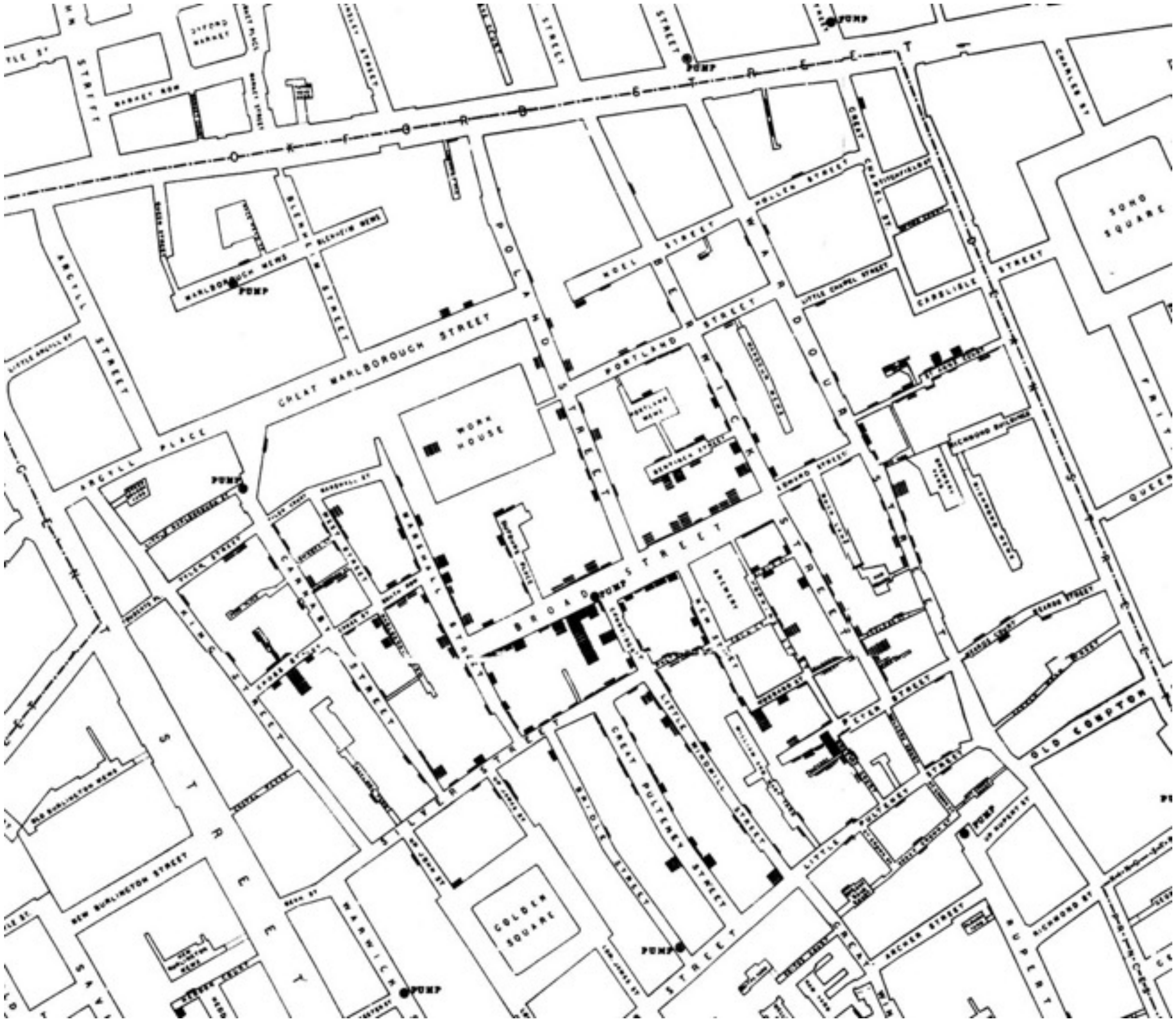
Visualisation is not new…



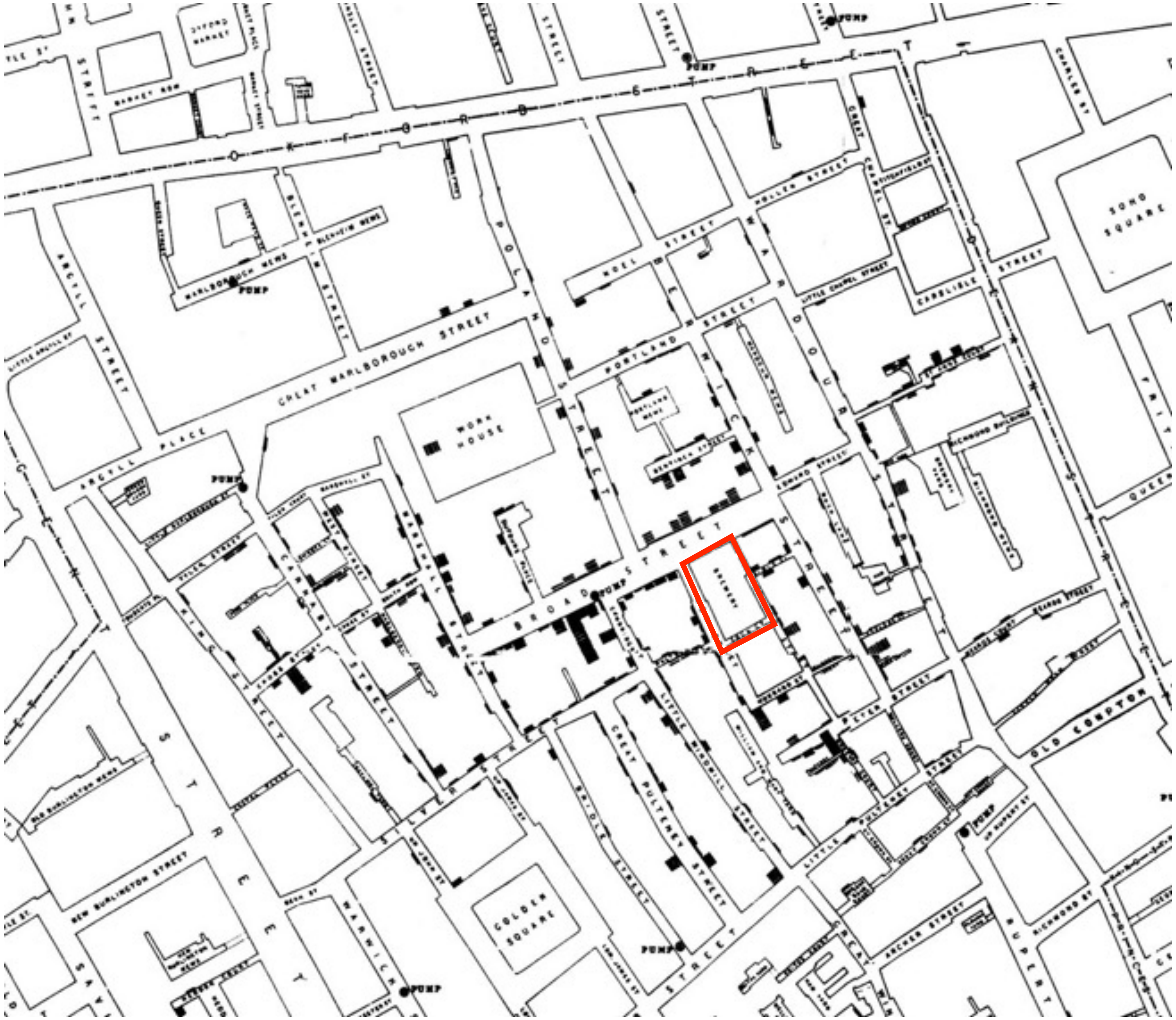
Al-Biruni - time series visualization: phases of the moon in orbit (circa 1030)

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

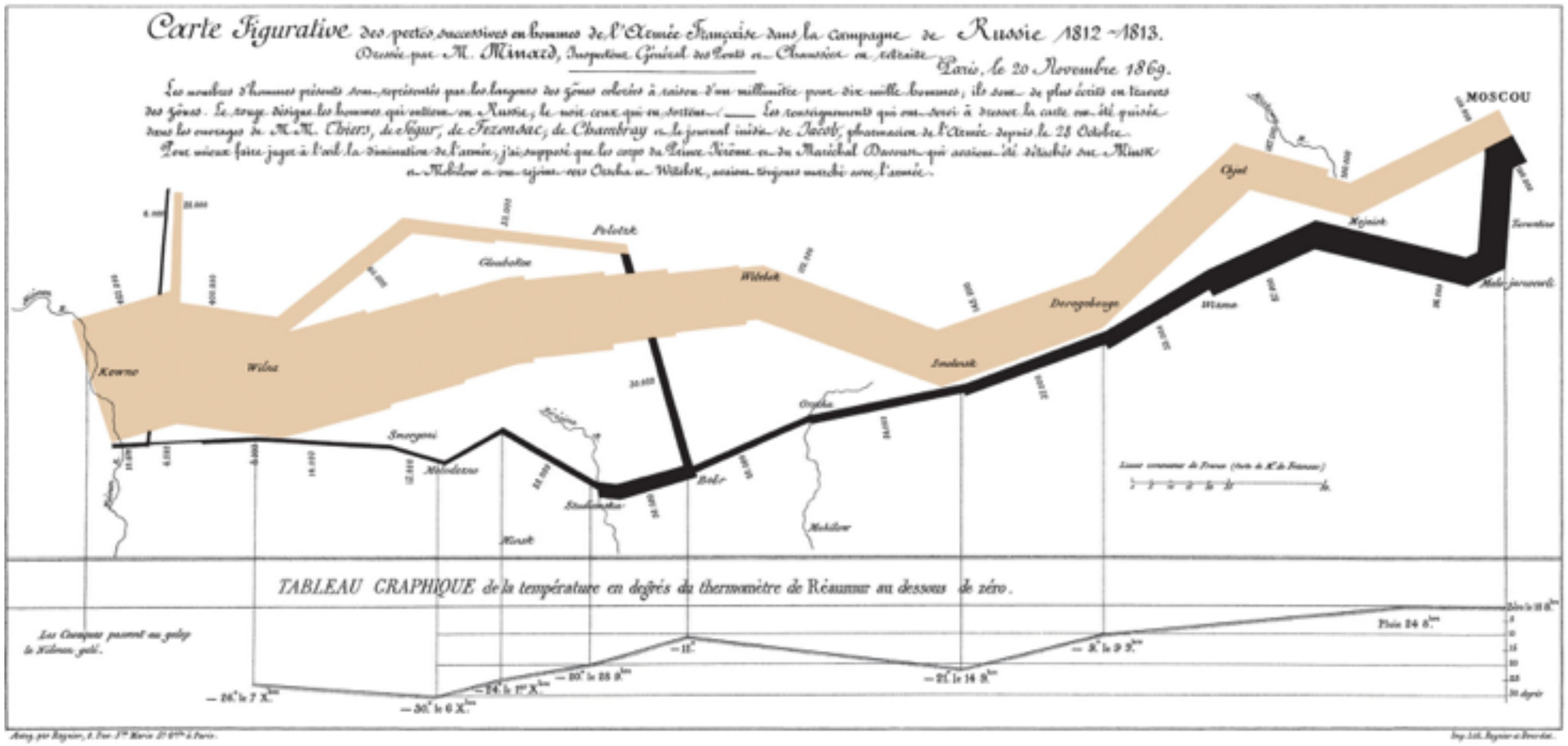
Peutinger Map - roads in Roman Empire, 1570
Abraham Ortelius (Flemish cartographer)



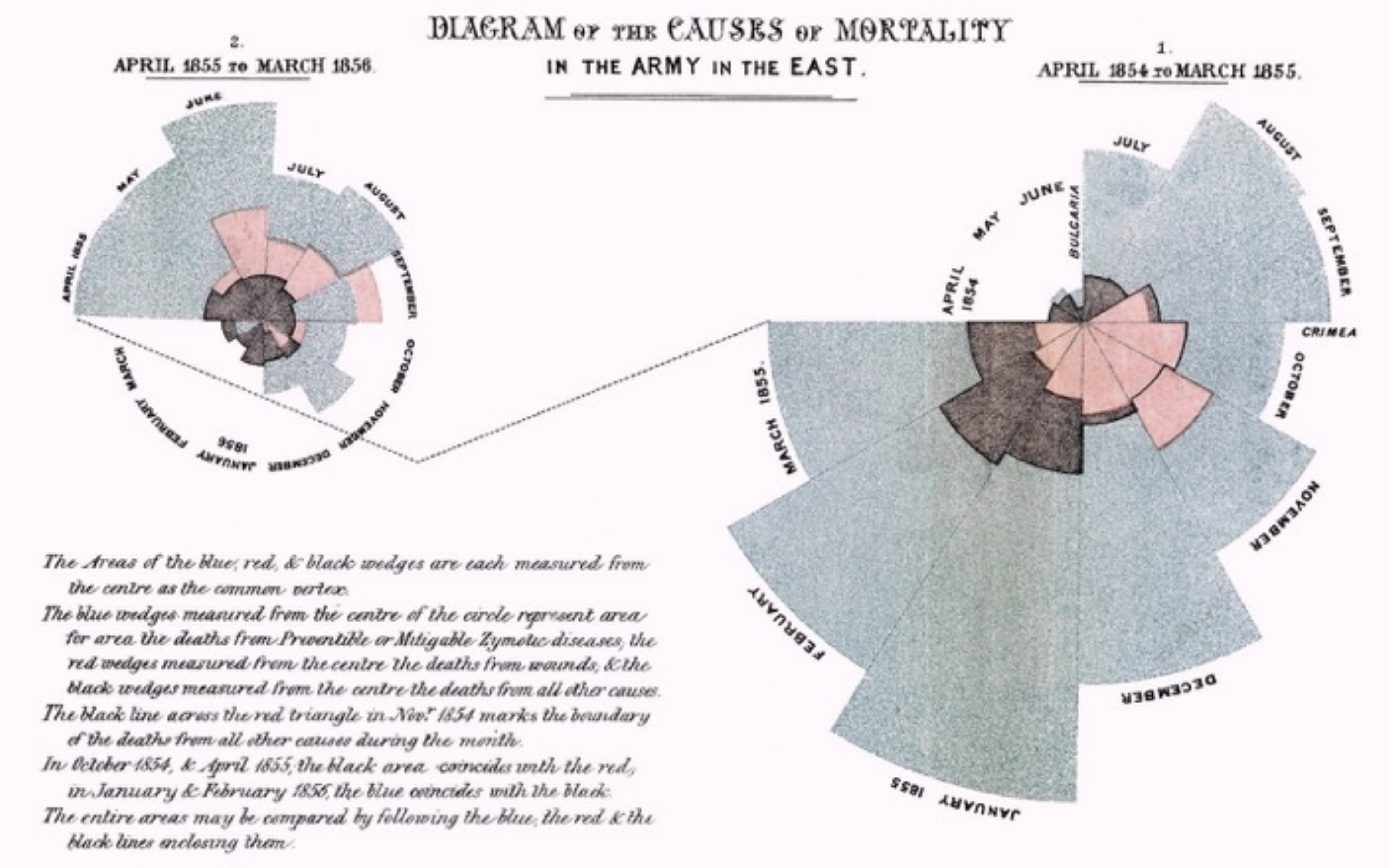
John Snow - cases of cholera in London (1663)



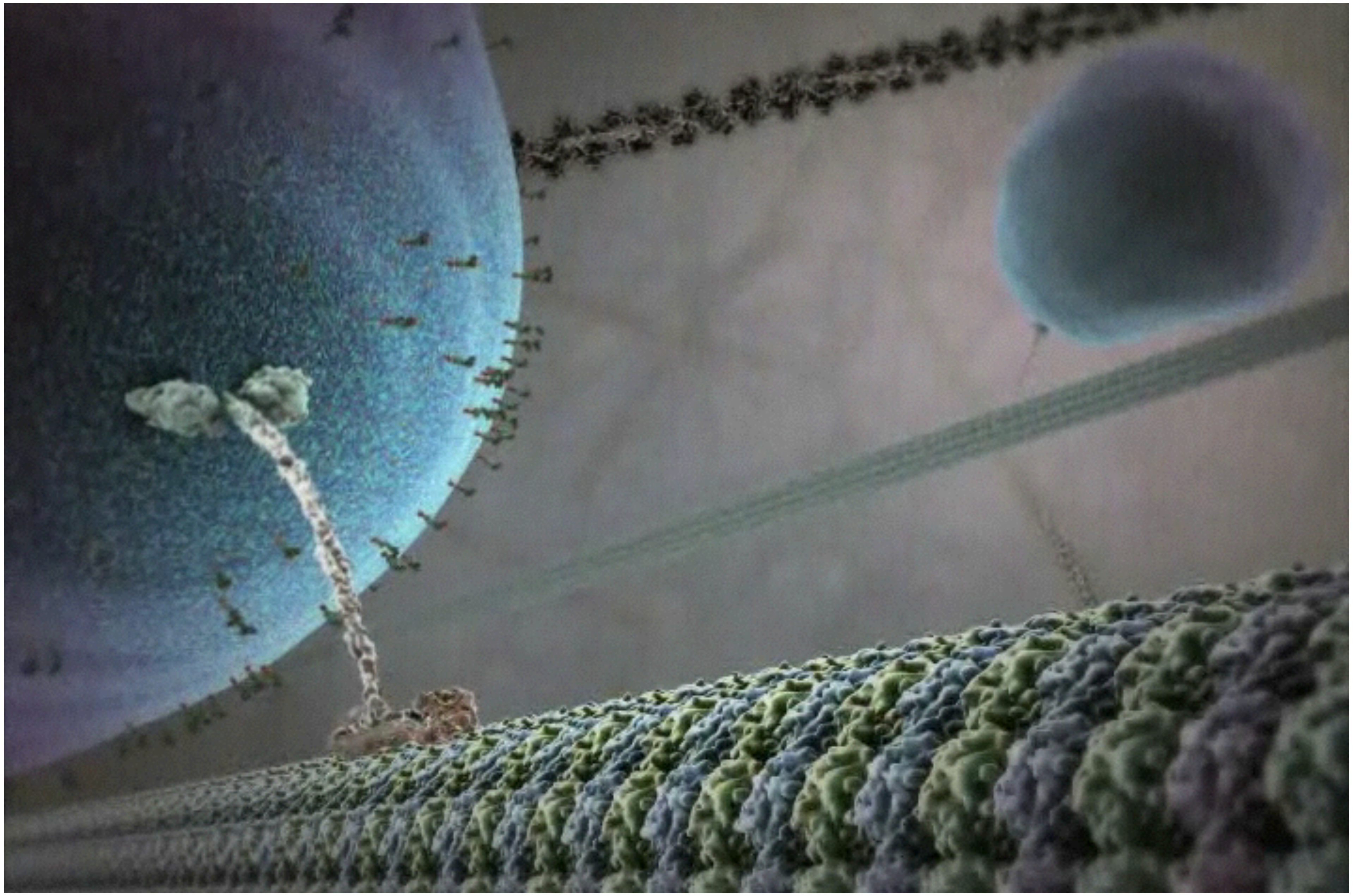
John Snow - cases of cholera in London (1663)

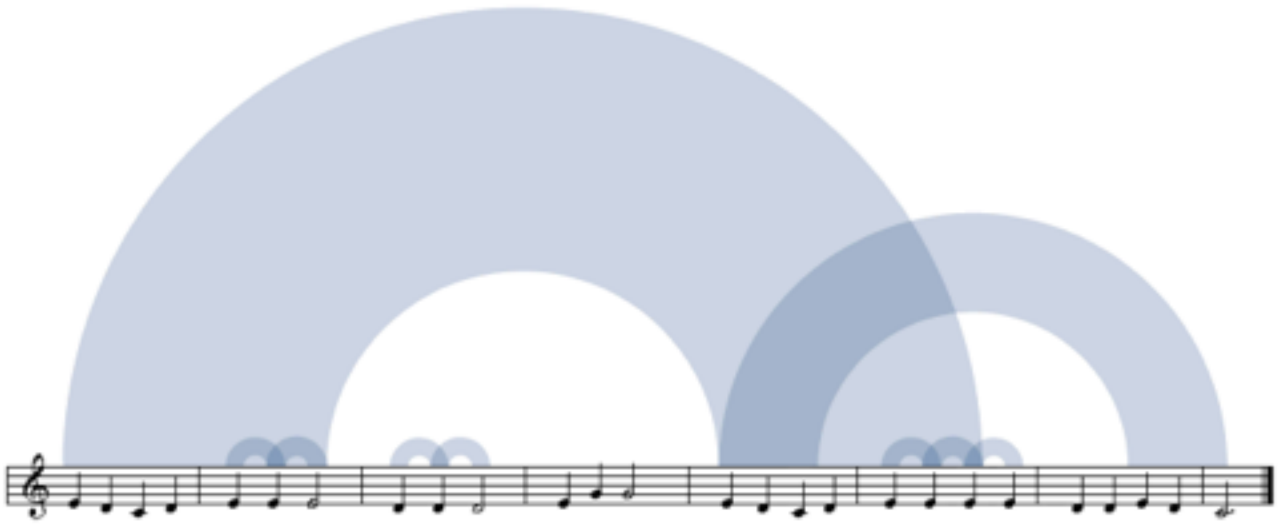


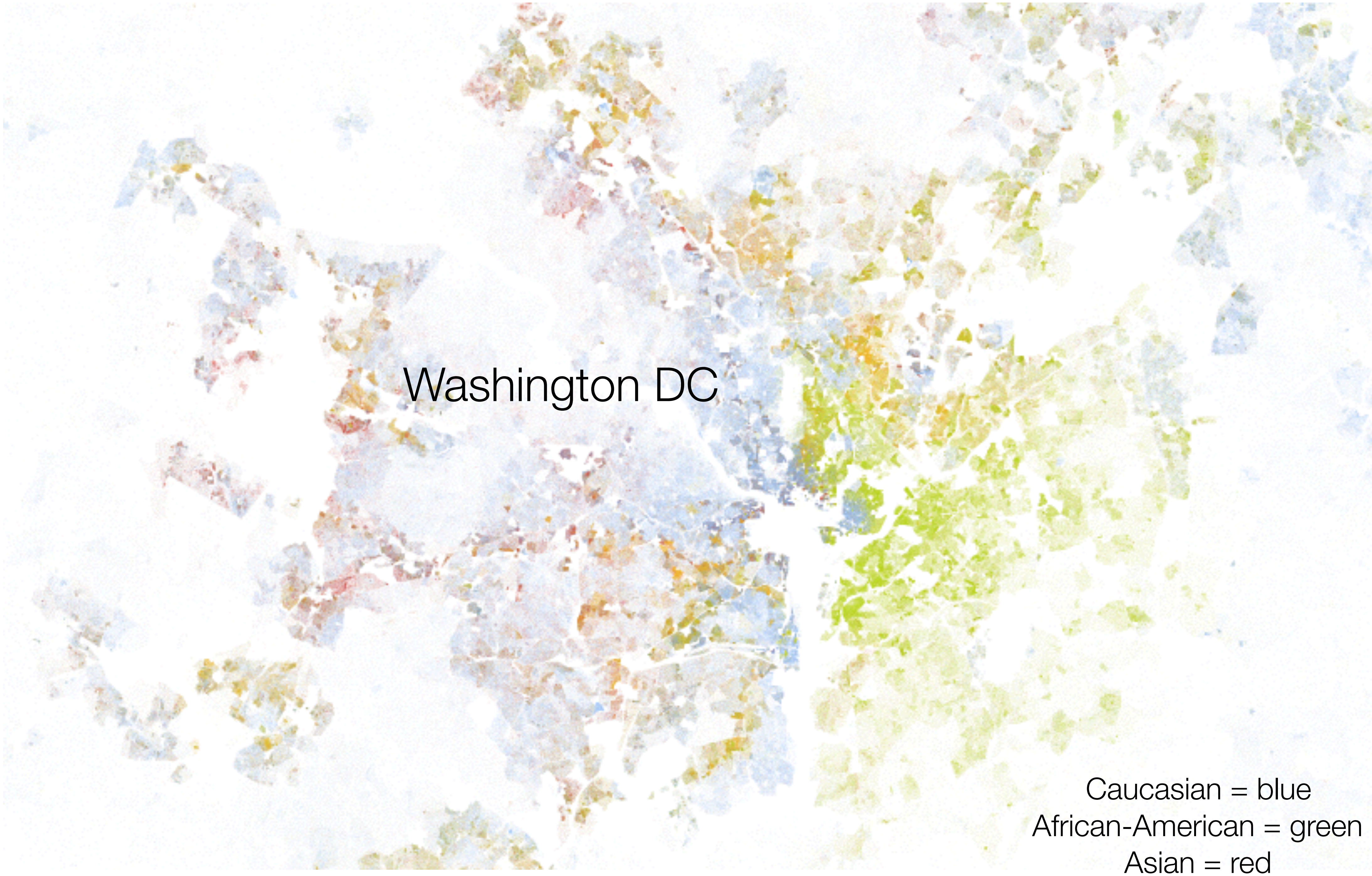
Charles Joseph Minard (1781-1870) - Napoleon’s march on Moscow



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 visualisation?

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 <=> perception
cognitive task => perceptive task

identify anomalies, clusters, trends

T. Munzner

Human perception is very strong

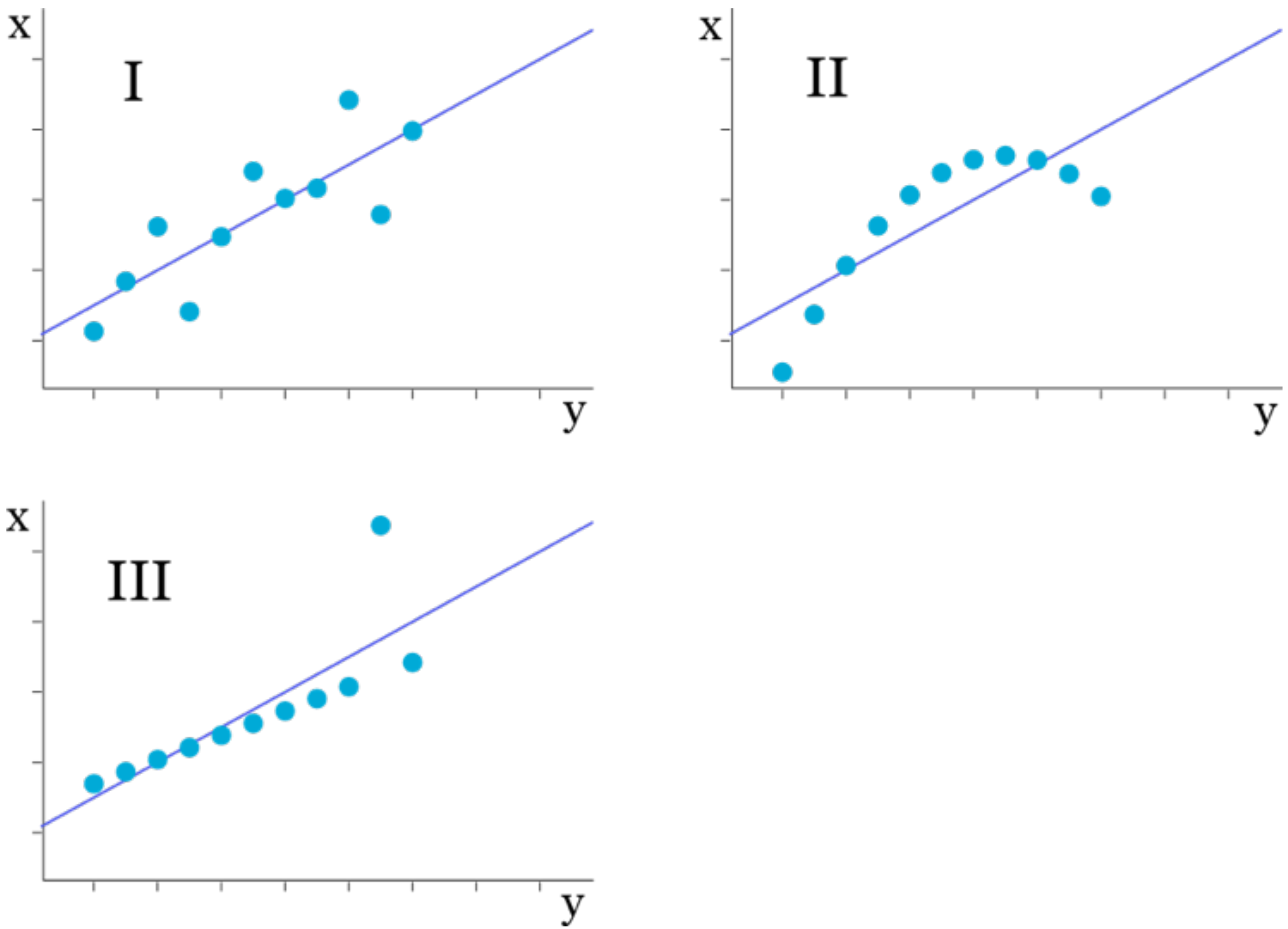
I		II		III	
x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46
8.0	6.95	8.0	8.14	8.0	6.77
13.0	7.58	13.0	8.74	13.0	12.74
9.0	8.81	9.0	8.77	9.0	7.11
11.0	8.33	11.0	9.26	11.0	7.81
14.0	9.96	14.0	8.10	14.0	8.84
6.0	7.24	6.0	6.13	6.0	6.08
4.0	4.26	4.0	3.10	4.0	5.39
12.0	10.84	12.0	9.13	12.0	8.15
7.0	4.82	7.0	7.26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


I
II
III
x
y

Pre-attentive vision

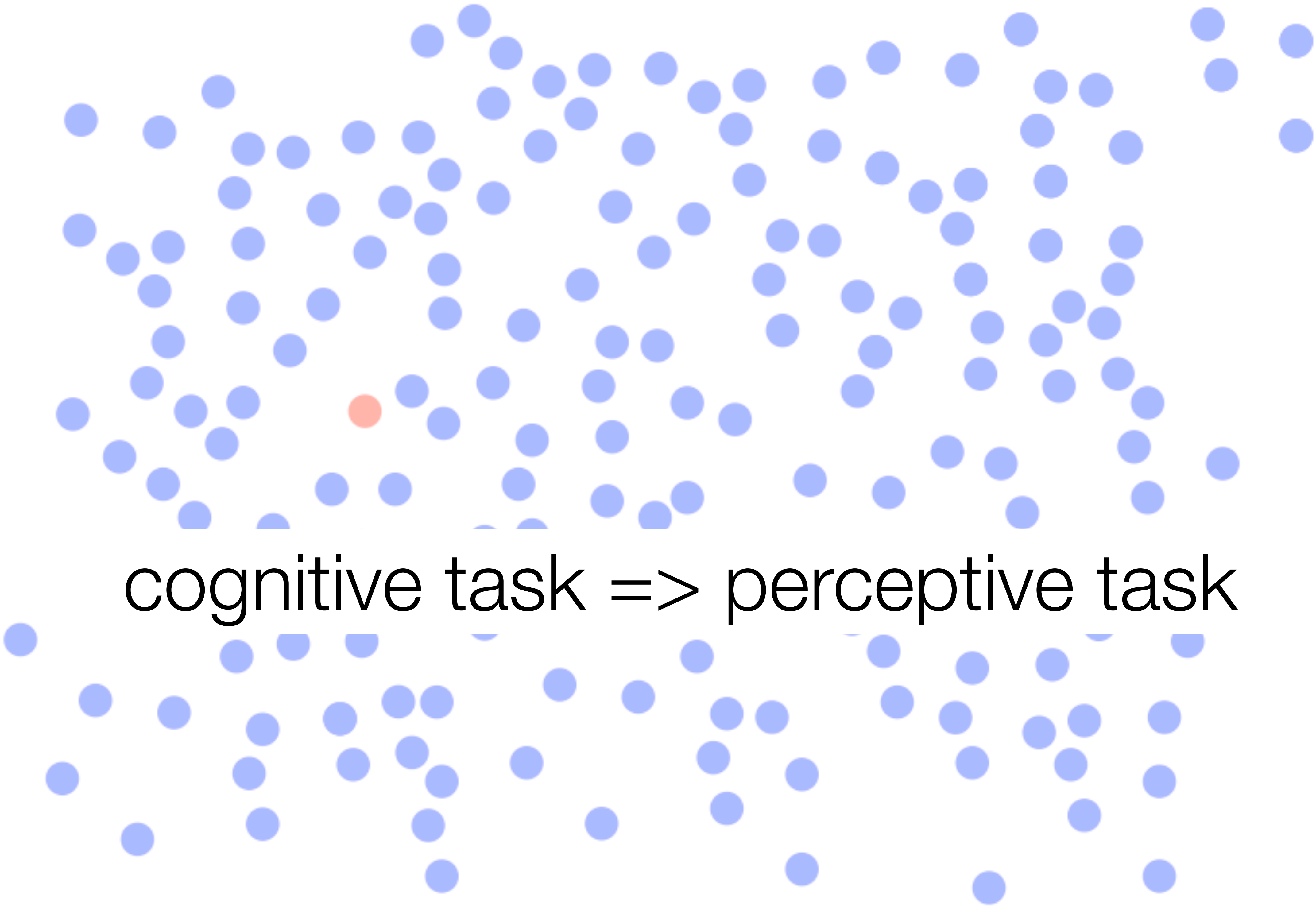
= ability of low-level human visual system to rapidly identify certain basic visual properties

- 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

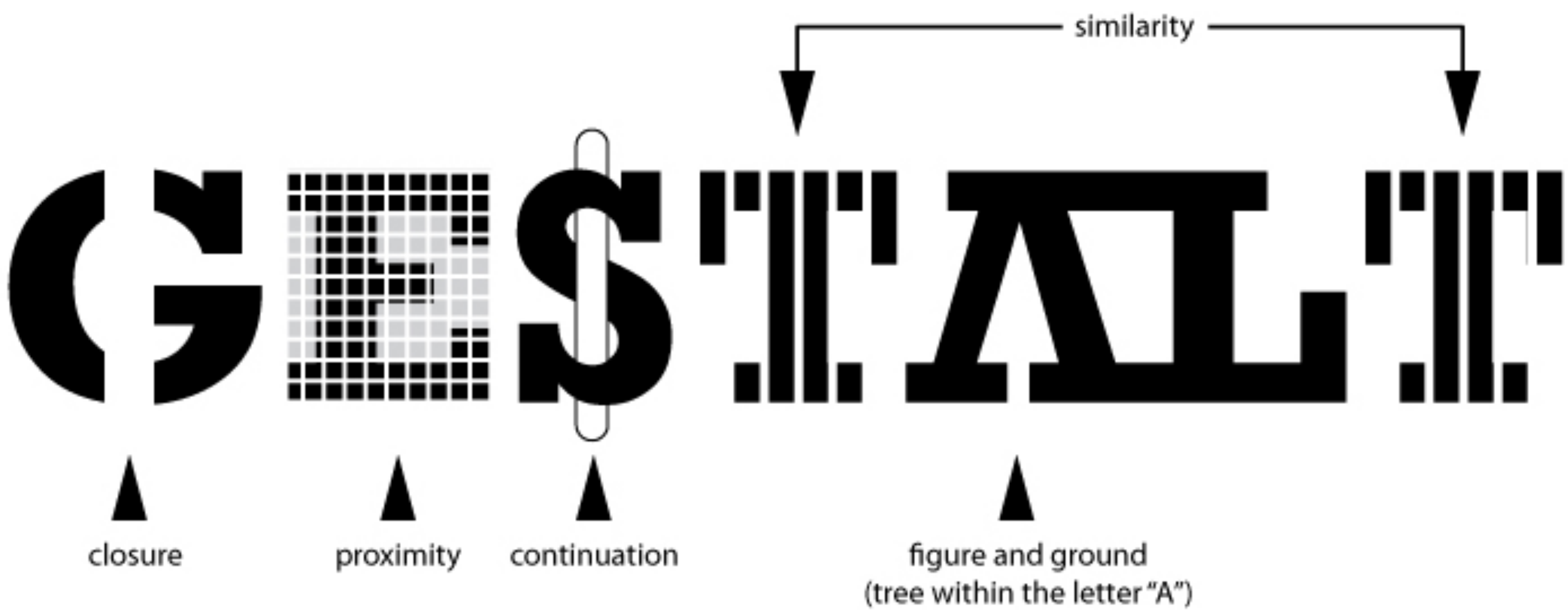
5



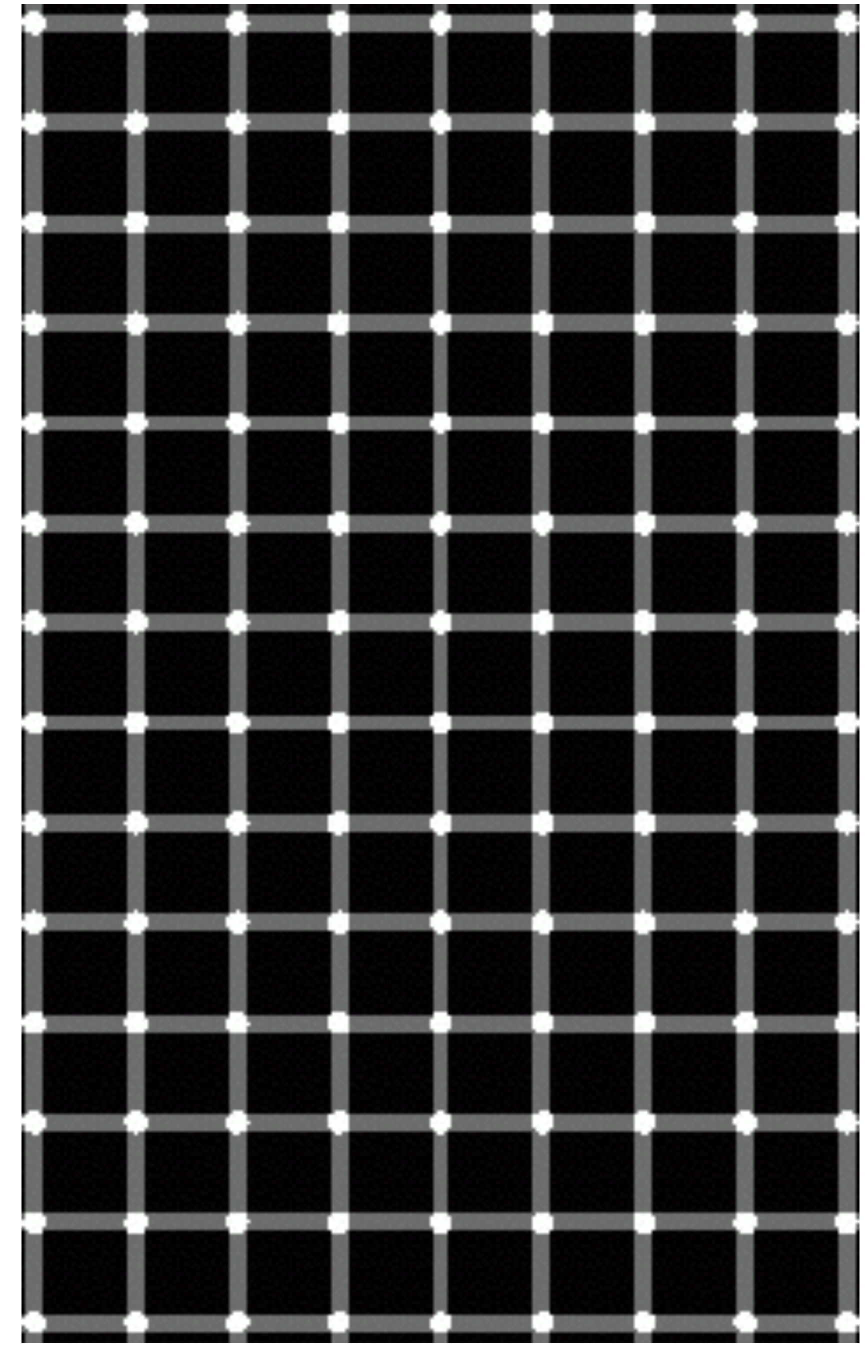
cognitive task => perceptive task

Human perception can be deceiving

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

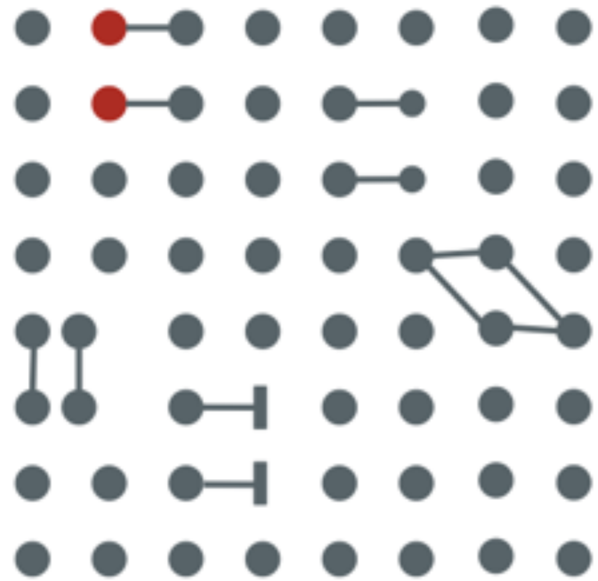


Hermann grid illusion

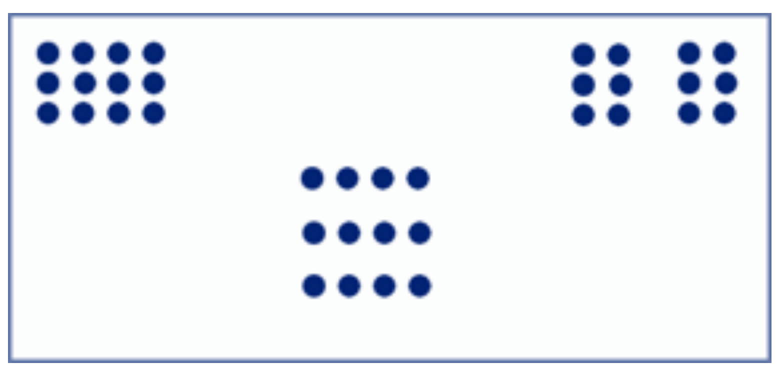


Simplicity

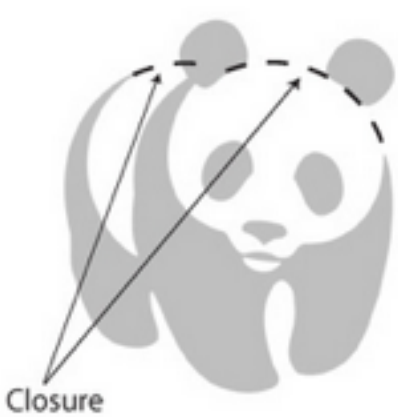
Connectedness



Proximity



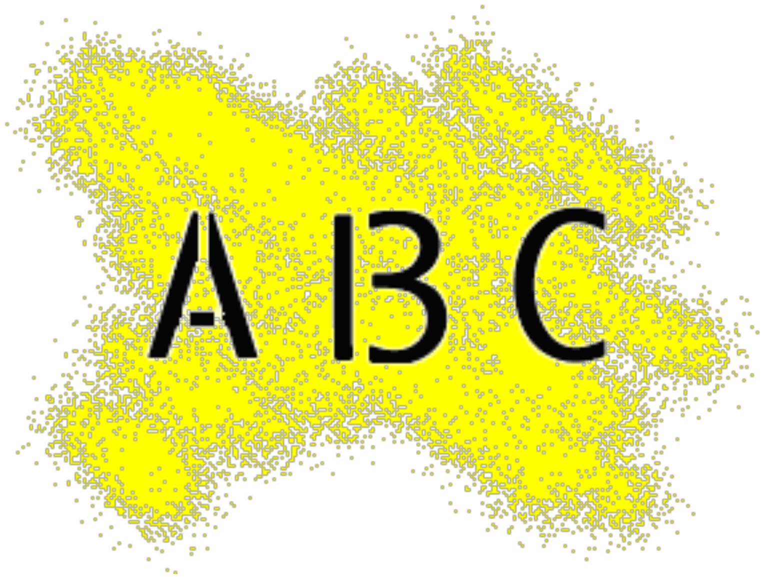
Closure



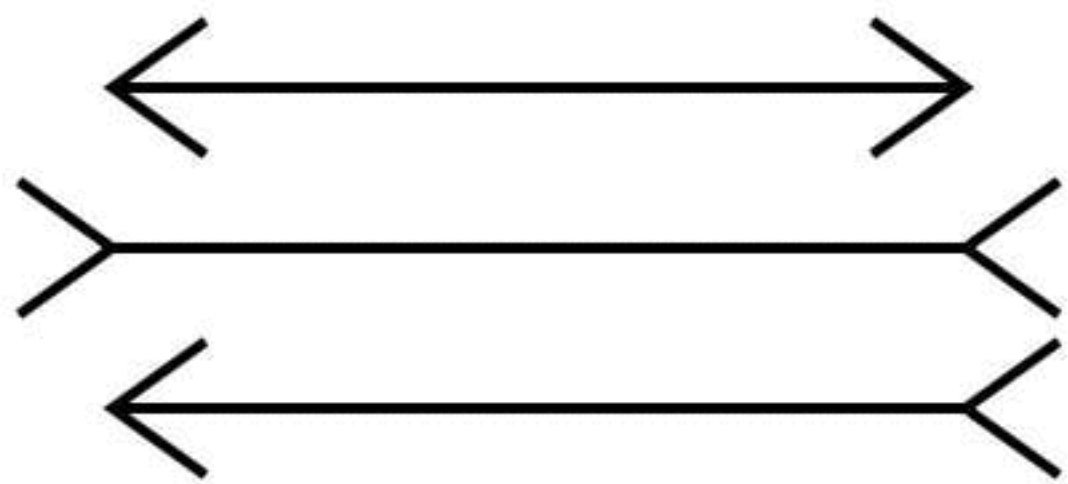
Kanizsa illusion

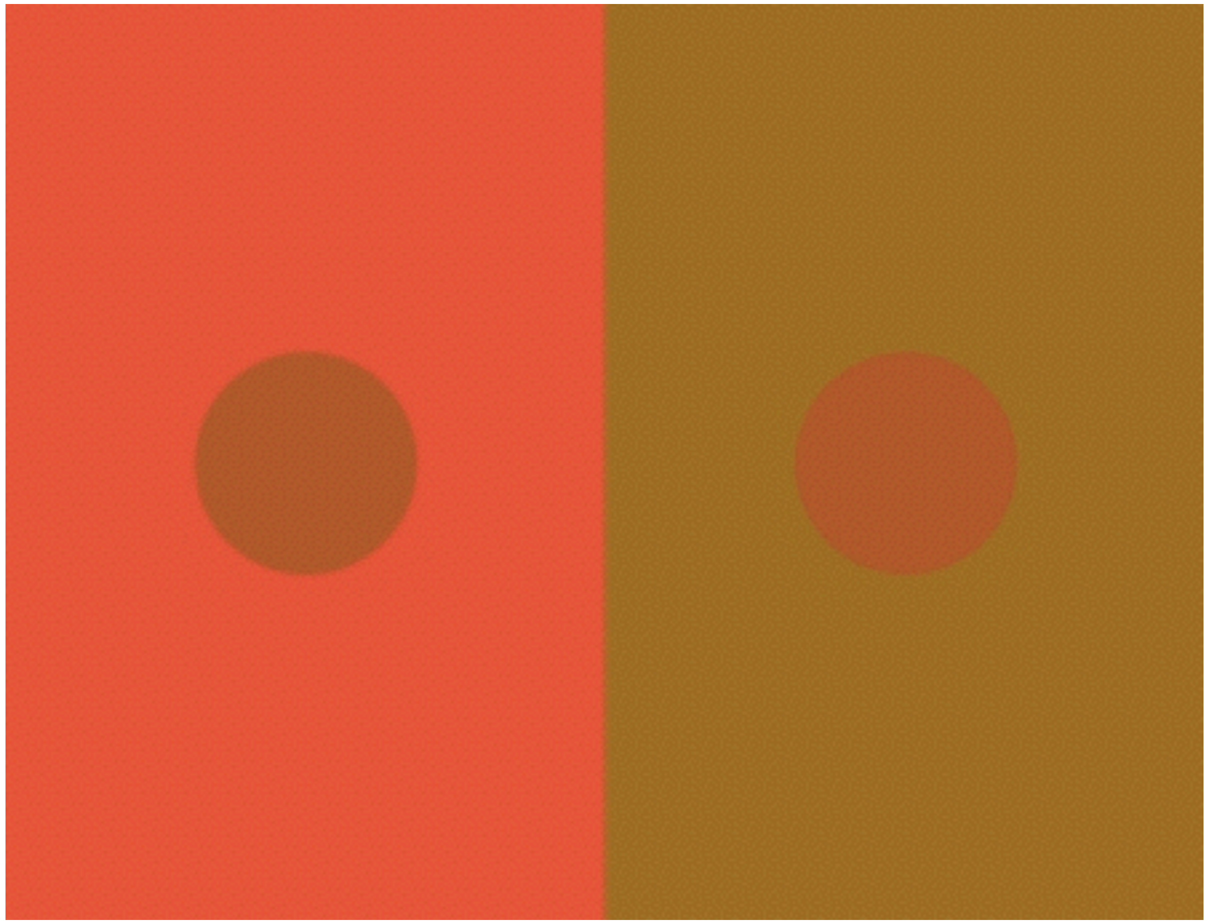
Symmetry,
Familiarity,
Continuity
Figure and Ground

Similarity


ABC


12 13 14





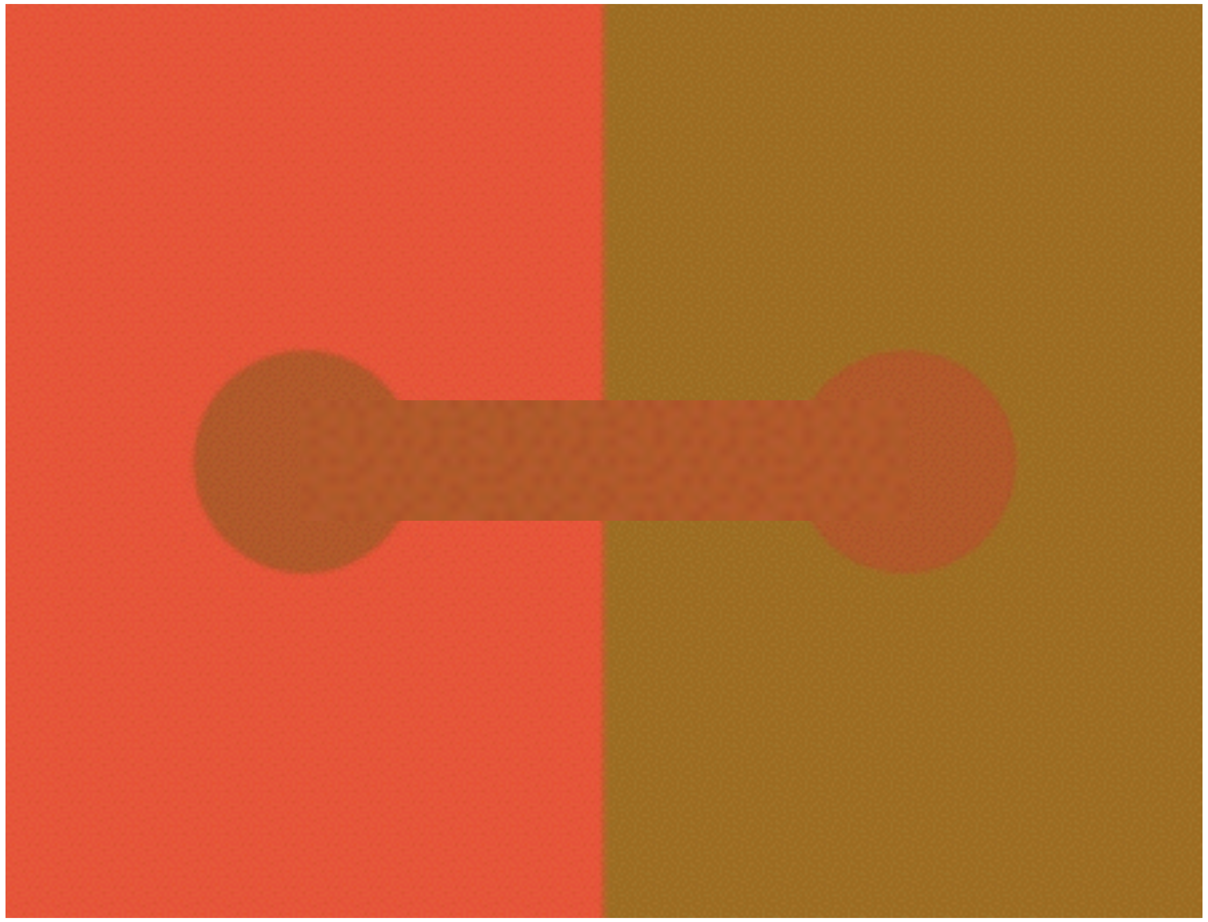
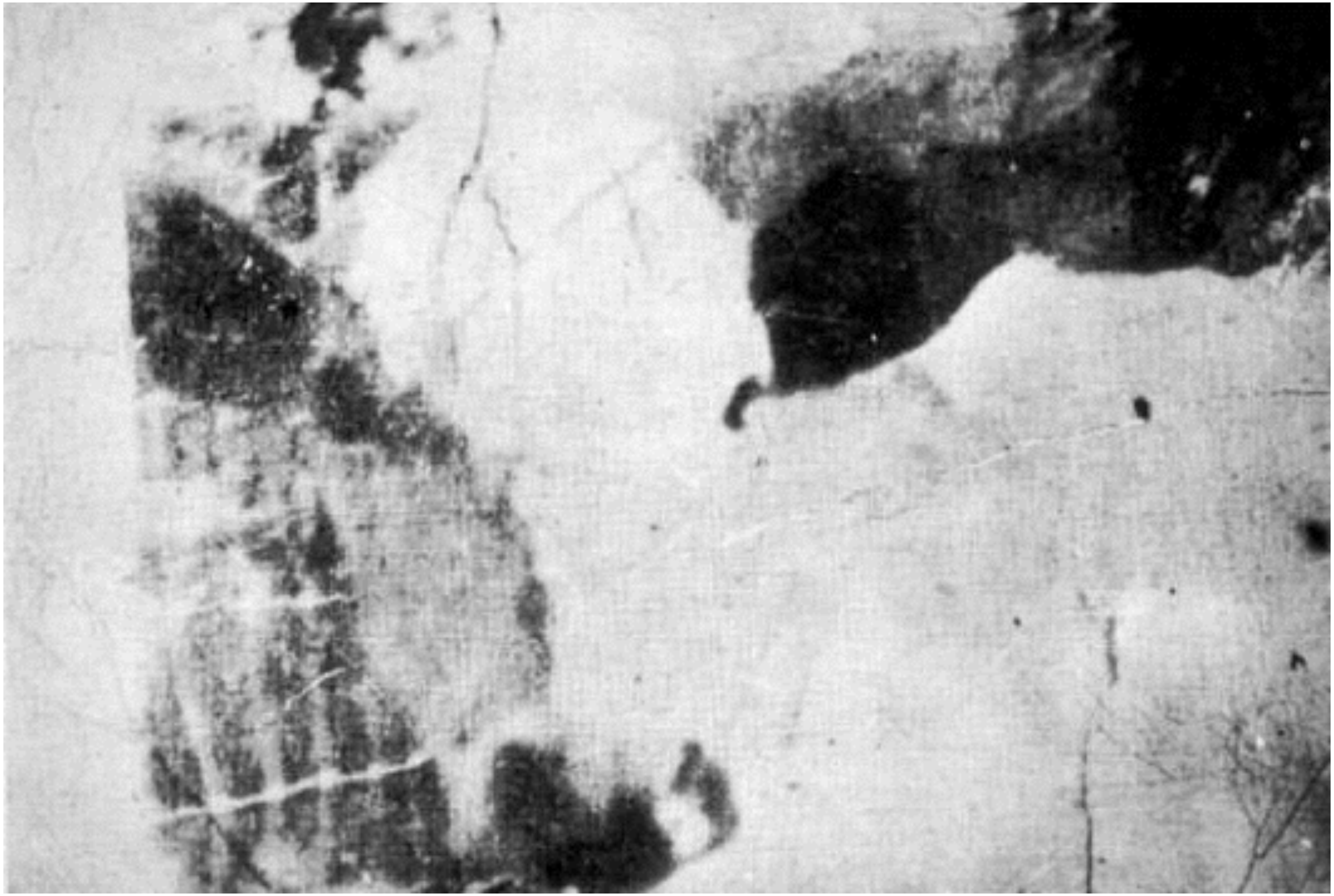
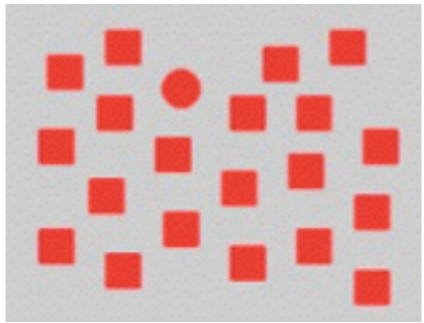


Figure and Ground



Limitations of preattentive vision

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



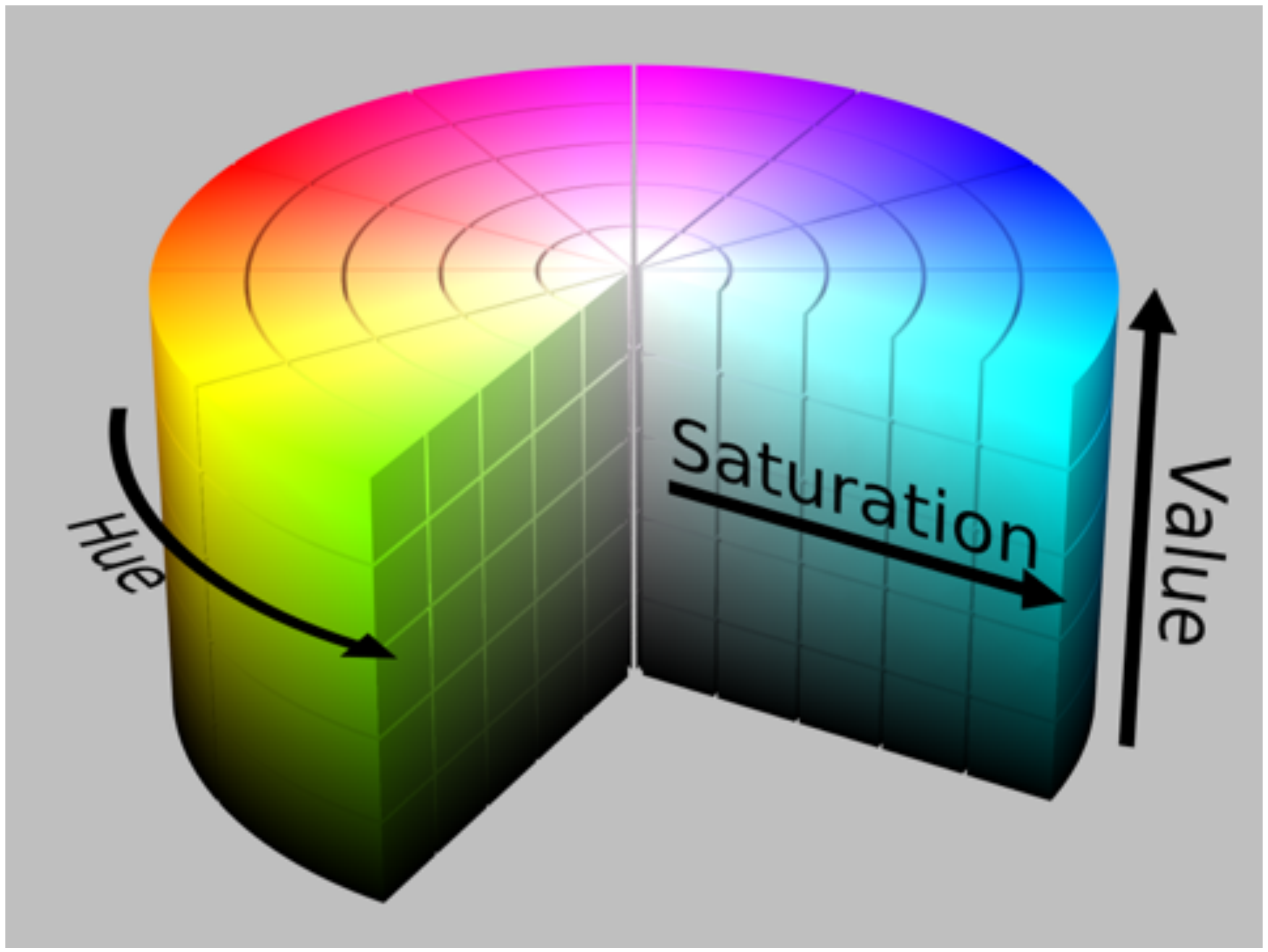
2. **Combining** pre-attentive features does *not* always work => would need to resort to “**serial search**” (most channel pairs; all channel triplets)

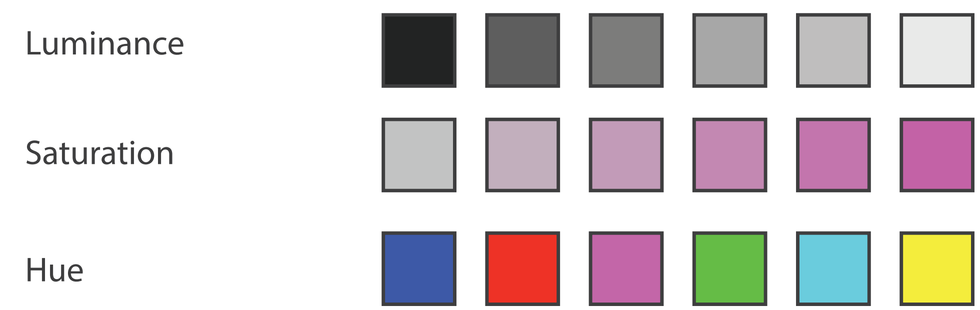
e.g. is there a red square in this picture

Selective attention

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

About colour

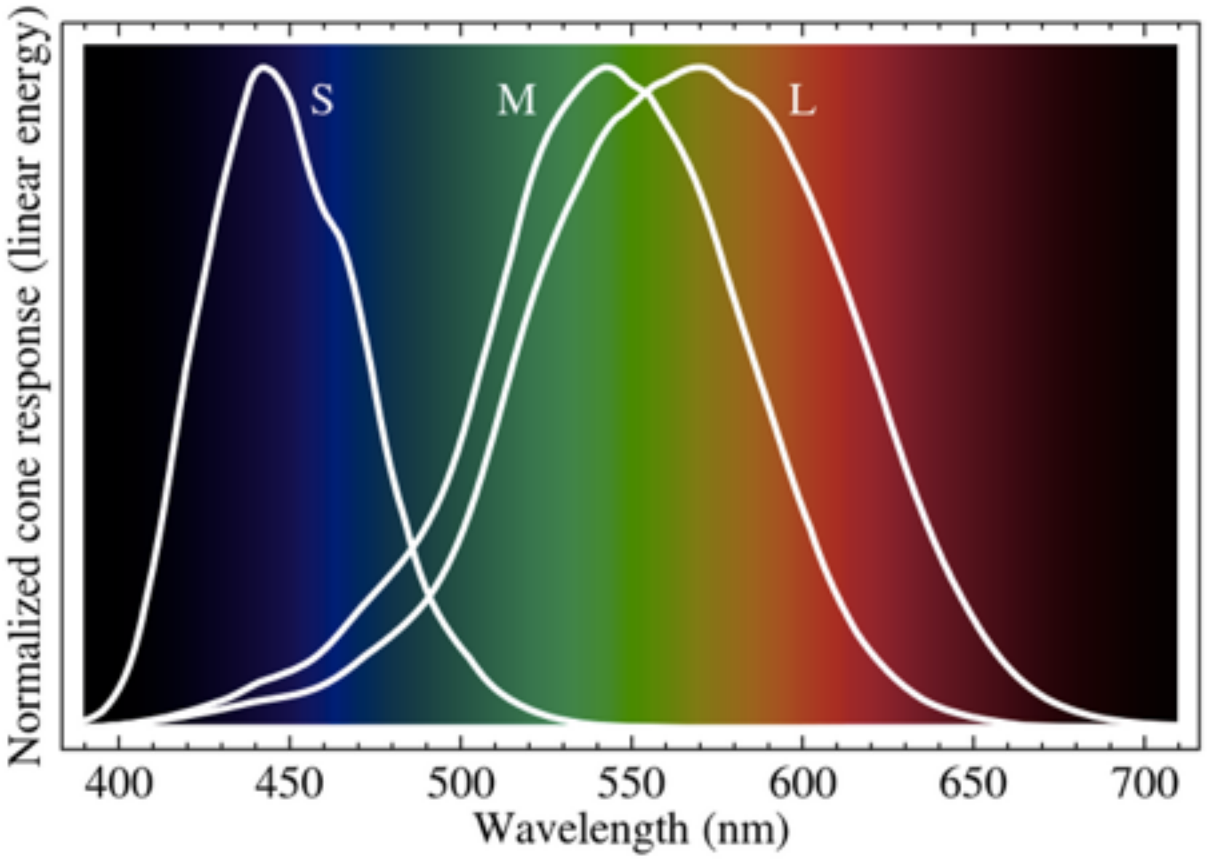




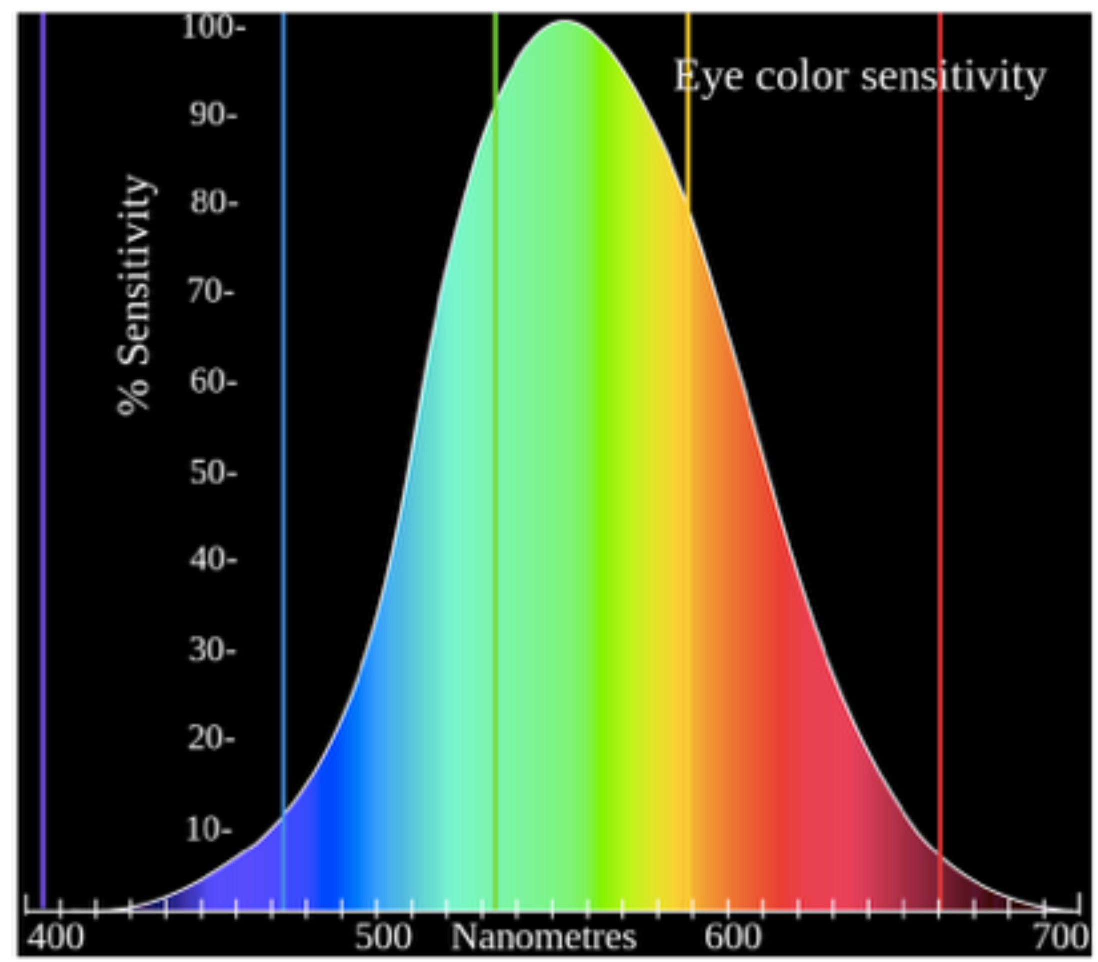
Use **HSV** (hue - saturation - value) instead of RGB

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

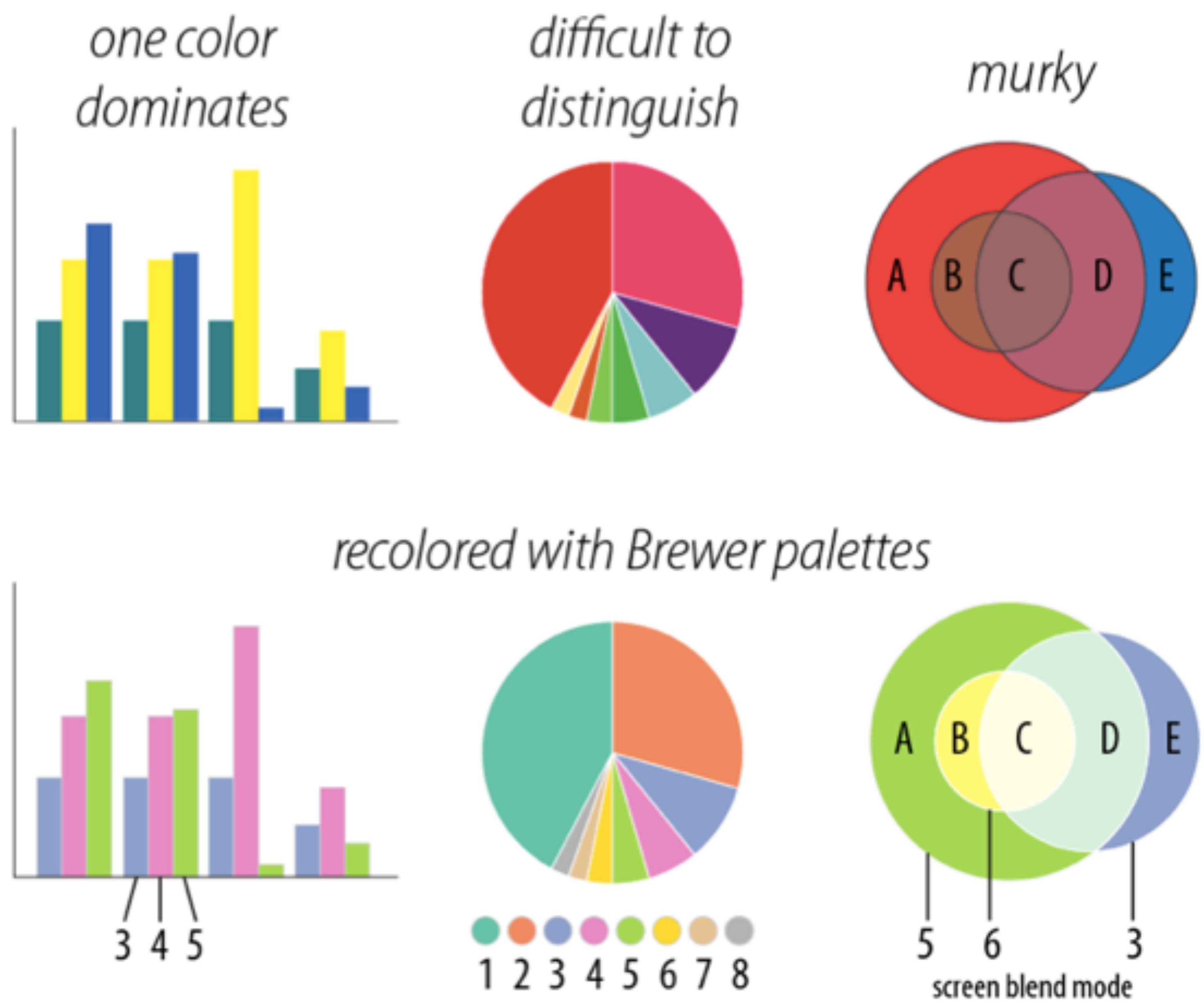
RGB has to do with *implementation*; HSV has to do with *perception*



sensitivity of S/M/L cones to colour wavelengths



colorbrewer2.org - Never pick your own colour



in R: please use RColorBrewer!

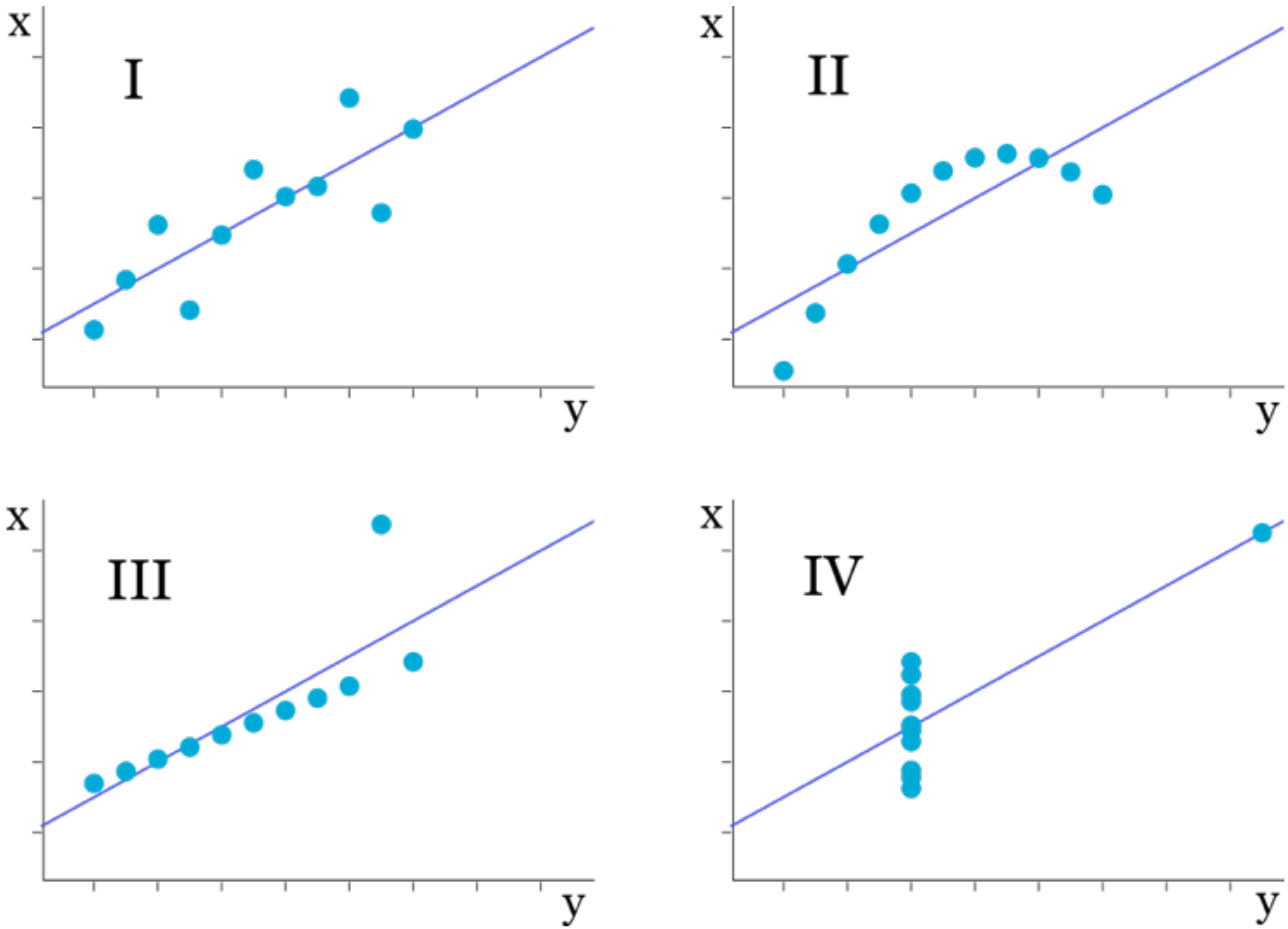
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0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

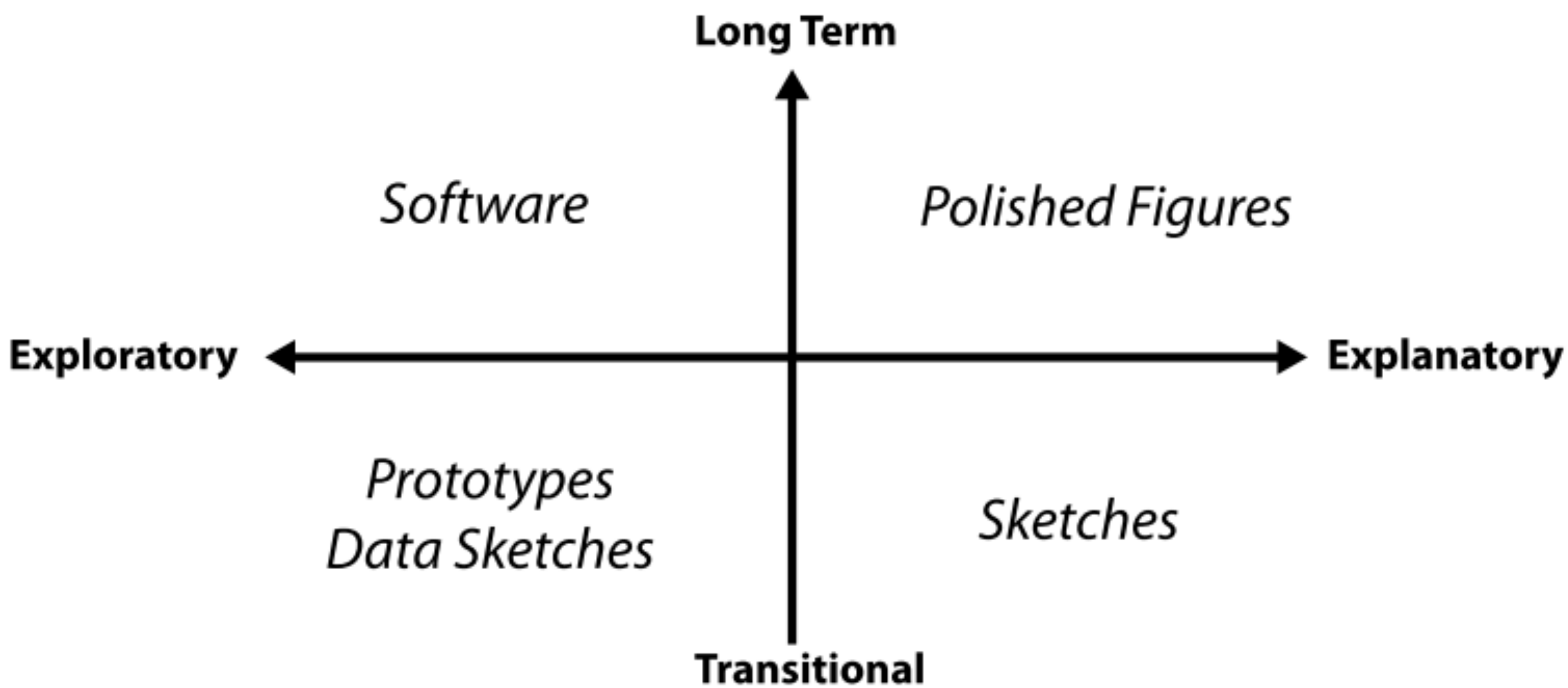
[C] Models for data visualization

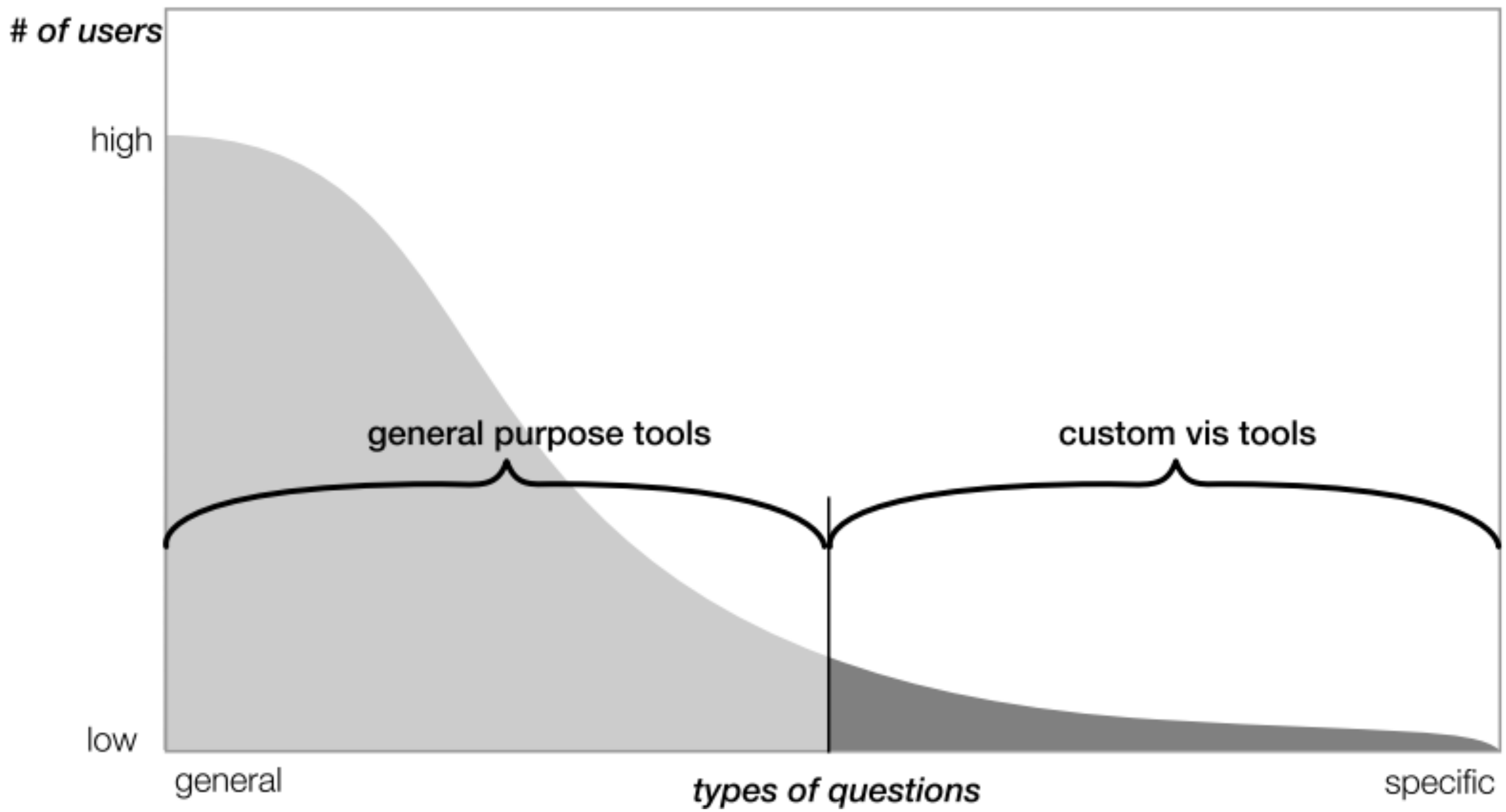
Why do we need to analyze?

design space = infinite

=> frameworks/models impose structure:

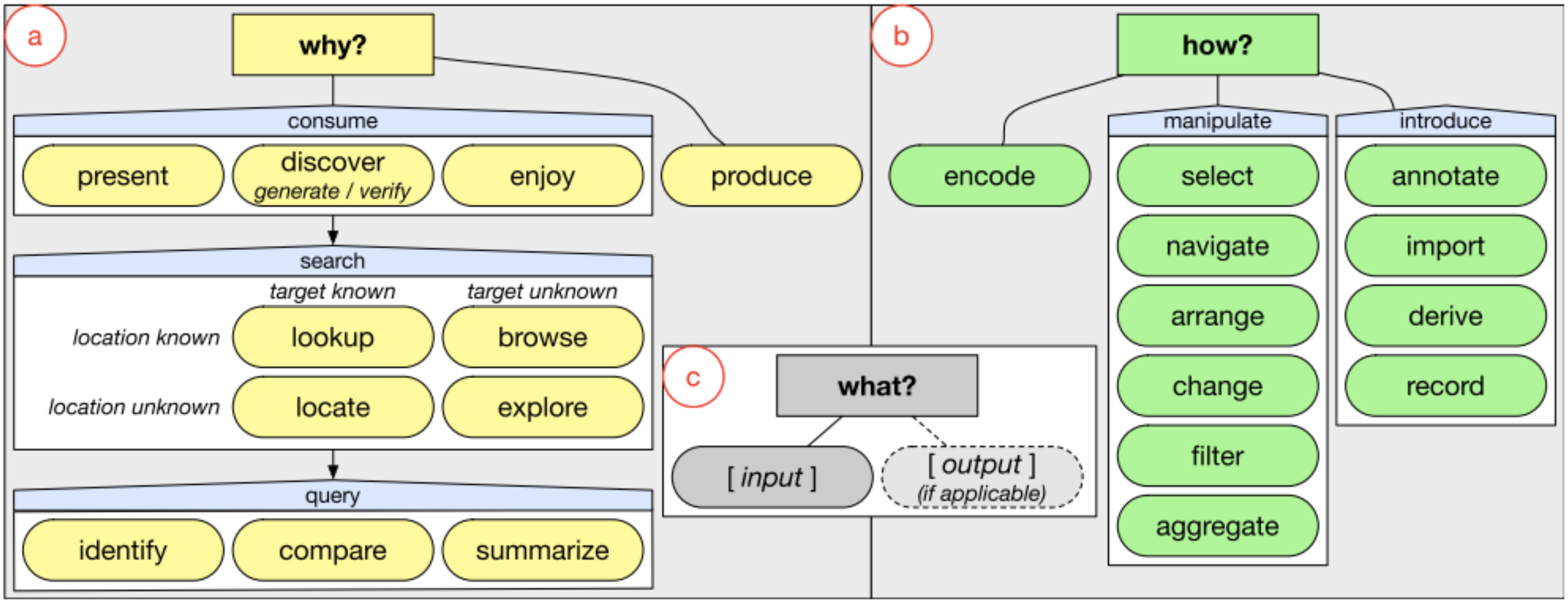
- scaffold to help think systematically about choices
- analyse existing visualisations as stepping stone to designing new ones


Long Term
Software
Polished Figures
Exploratory
Explanatory
Prototypes
Data Sketches
Sketches
Transitional



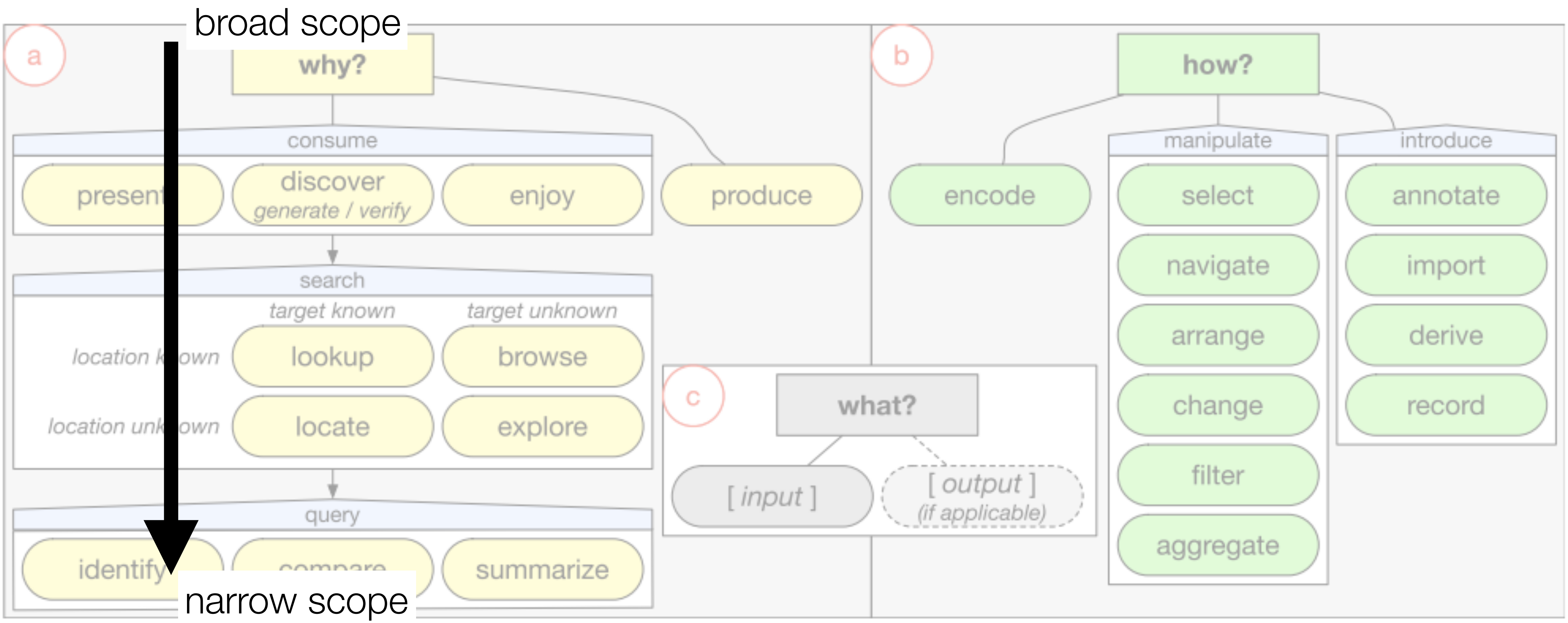
# of users
high
low
general purpose tools
custom vis tools
general
types of questions
specific

3 questions



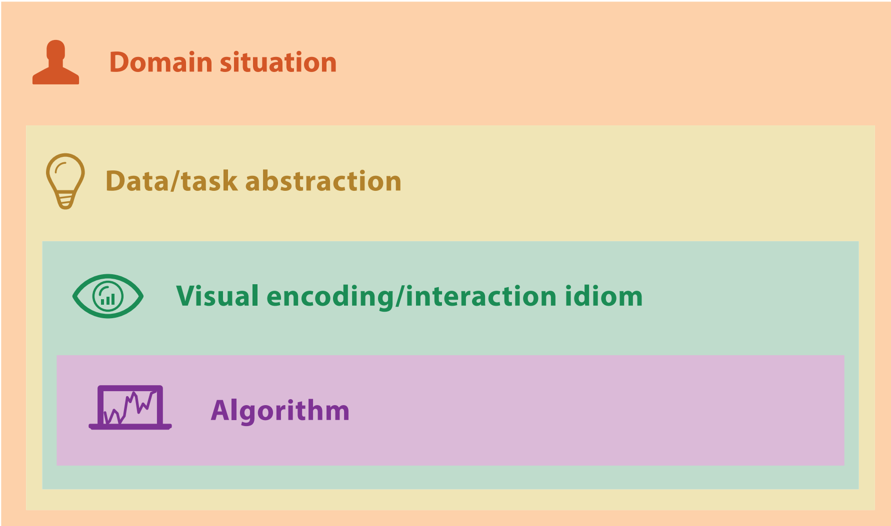
Brehmer & Munzner, 2013

"Visualization practitioners work with users to identify **why** and **what**, subsequently drawing from their specialized knowledge of visual encoding and interaction techniques to design **how** that task is to be supported"



Brehmer & Munzner, 2013

"Visualization practitioners work with users to identify **why** and **what**, subsequently drawing from their specialized knowledge of visual encoding and interaction techniques to design **how** that task is to be supported"



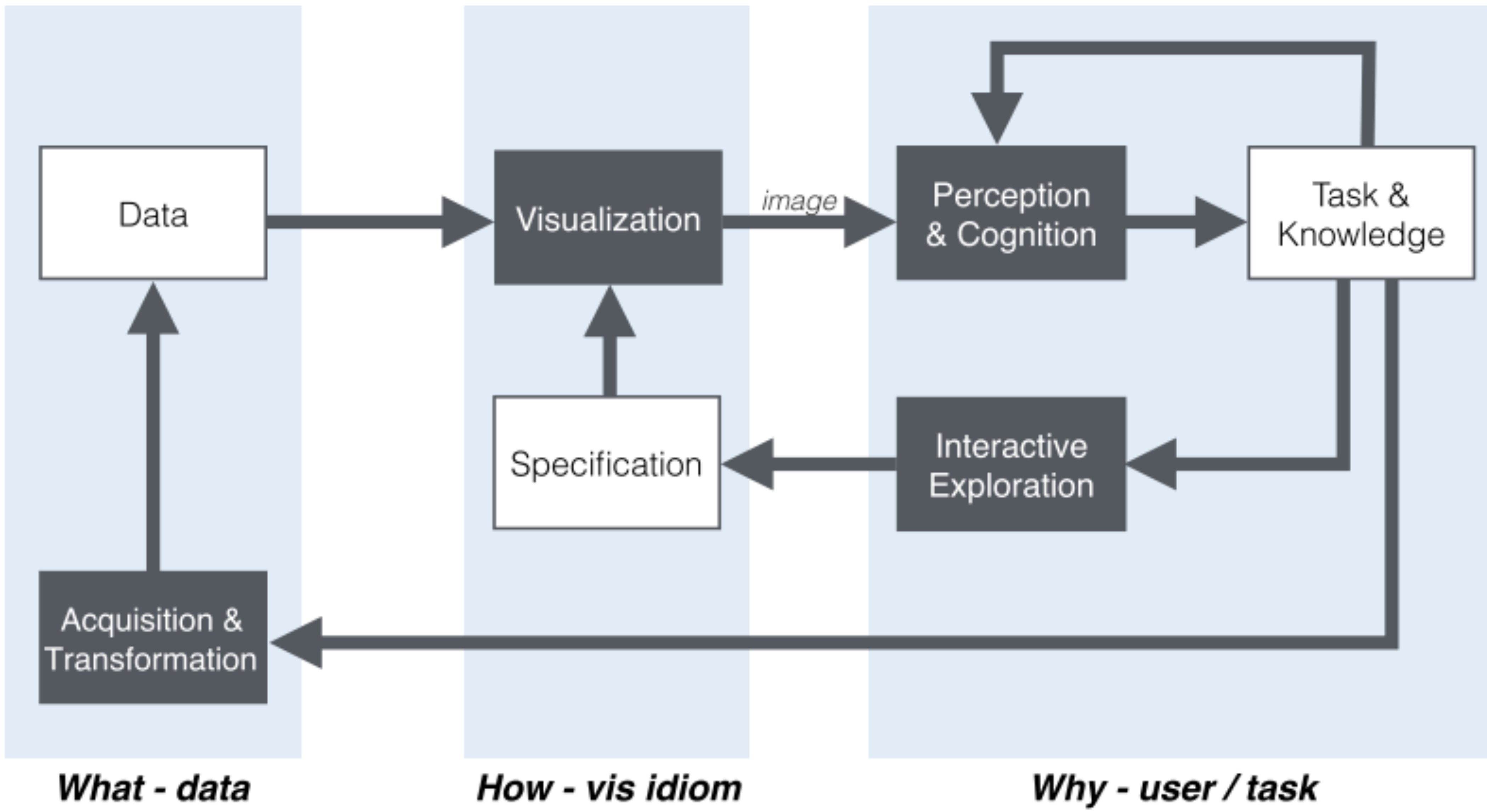
4 levels

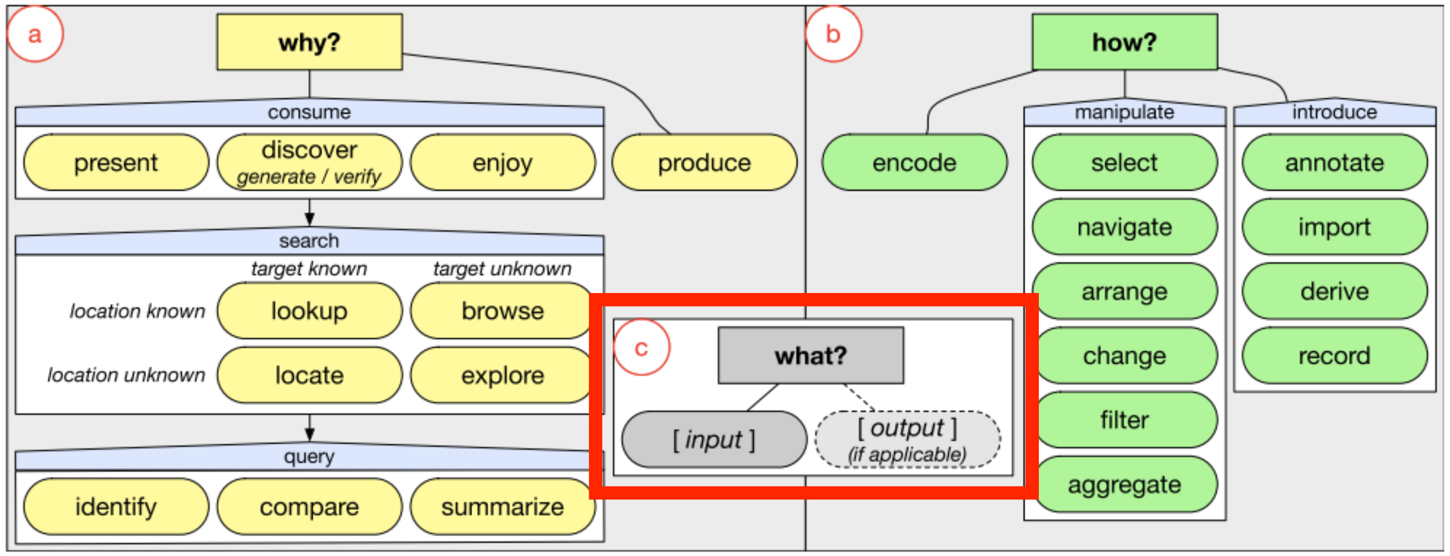
problem-driven

technique-driven

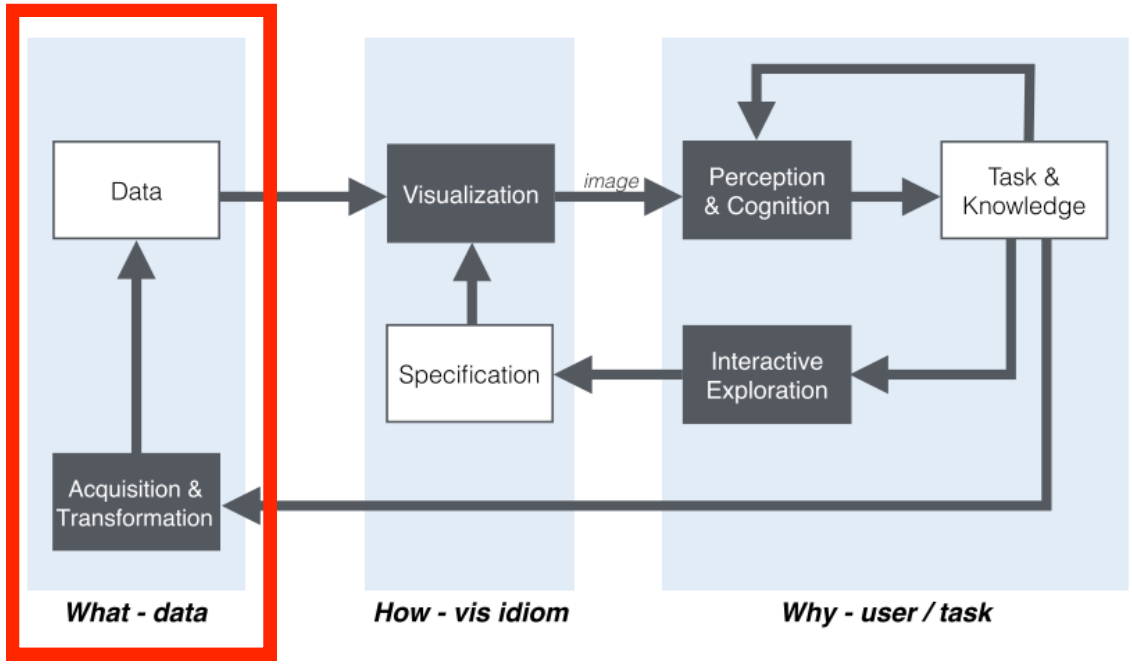
- **domain situation**: who are the target users?
- **abstraction**: translate from domain to vocabulary of dataviz
  - *what* is shown? => **data abstraction** (often: need data transformation)
  - *why* is the user looking at it? => **task abstraction**
- **idiom**: *how* is it shown?
  - visual encoding idiom: how to draw
  - interaction idiom: how to manipulate
- **algorithm**: efficient computation

Components


Data
Visualization
image
Perception & Cognition
Task & Knowledge
Specification
Interactive Exploration
Acquisition & Transformation
What - data
How - vis idiom
Why - user / task

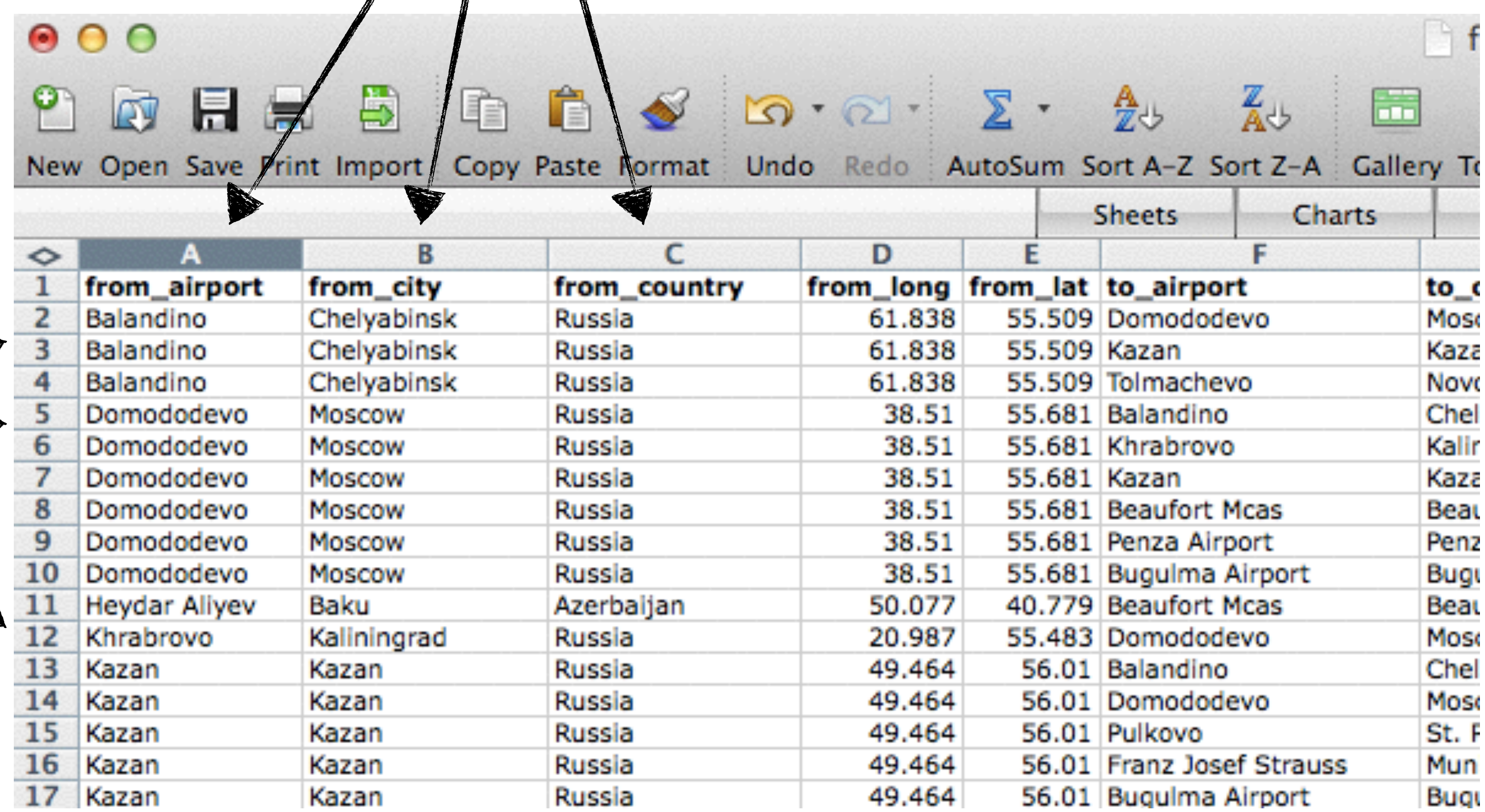


What - Data foundations

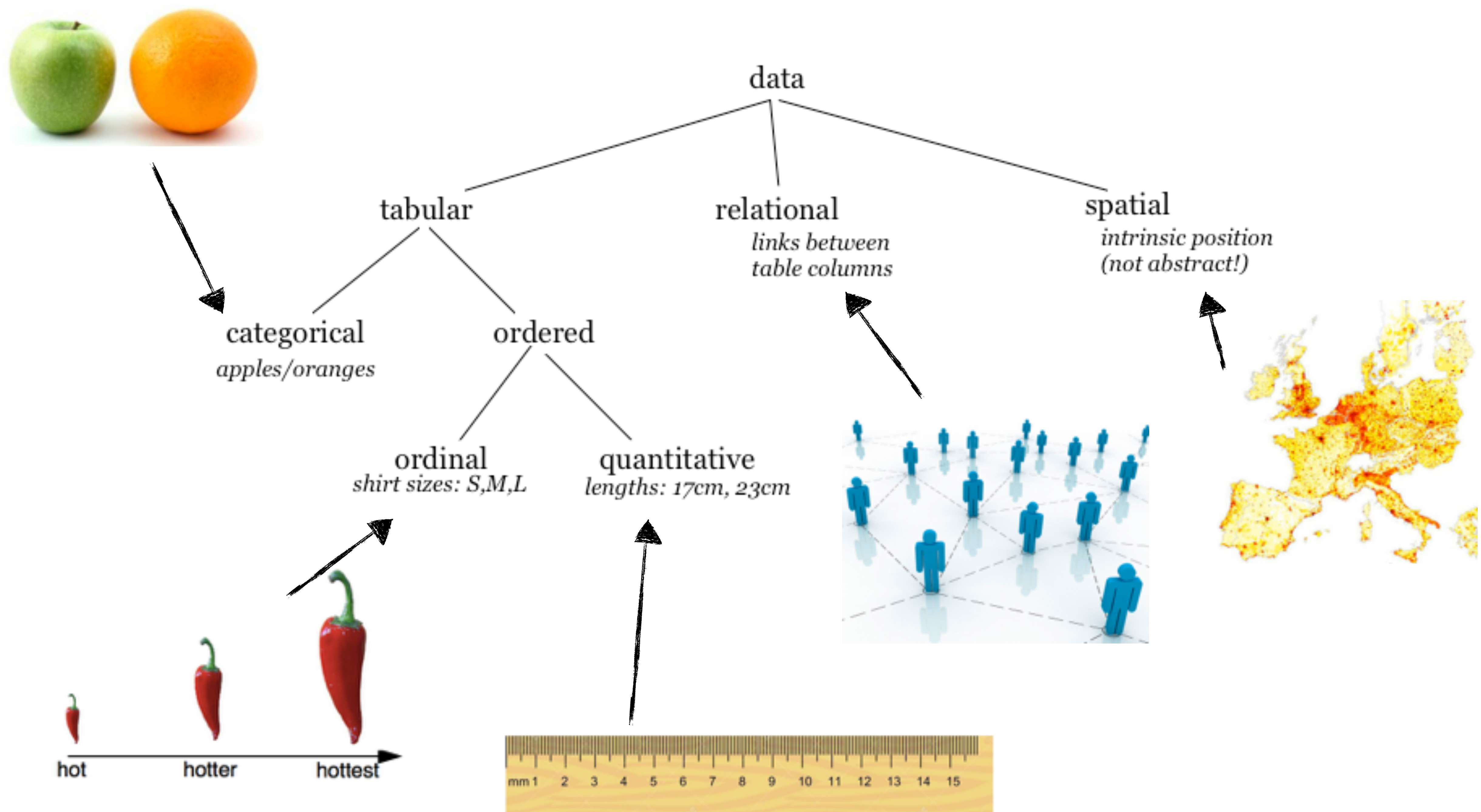


data attributes
dimensions
features
properties

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



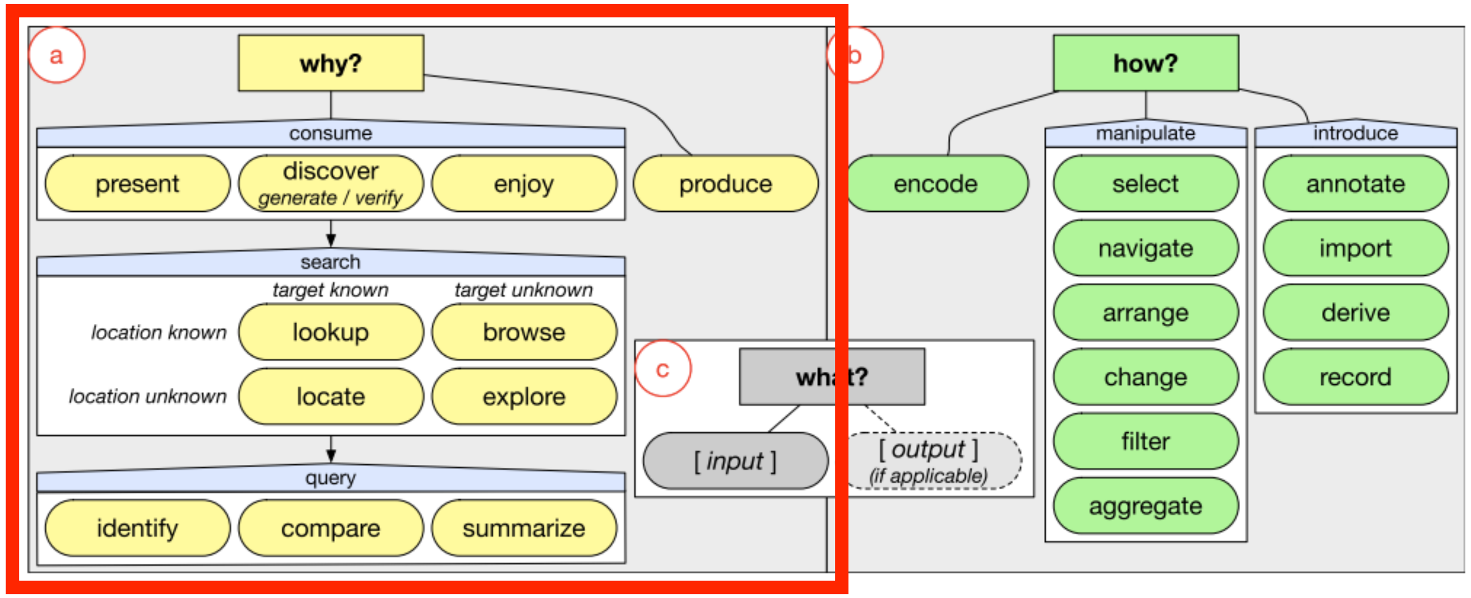
S Stevens “On the theory of scales and measurements” (1946)


What?
Datasets
Attributes
Data Types
Items
Attributes
Links
Positions
Grids
Data and Dataset Types
Tables
Items
Attributes
Networks & Trees
Items (nodes)
Links
Attributes
Fields
Grids
Positions
Attributes
Geometry
Items
Positions
Clusters, Sets, Lists
Items
Dataset Types
Tables
Attributes (columns)
Items (rows)
Cell containing value
Networks
Link
Node (item)
Fields (Continuous)
Grid of positions
Cell
Attributes (columns)
Value in cell
Multidimensional Table
Key 1
Key 2
Attributes
Value in cell
Trees
Geometry (Spatial)
Position
Dataset Availability
Static
Dynamic
Attribute Types
Categorical
Ordered
Ordinal
Quantitative
Ordering Direction
Sequential
Diverging
Cyclic
What?
Why?
How?

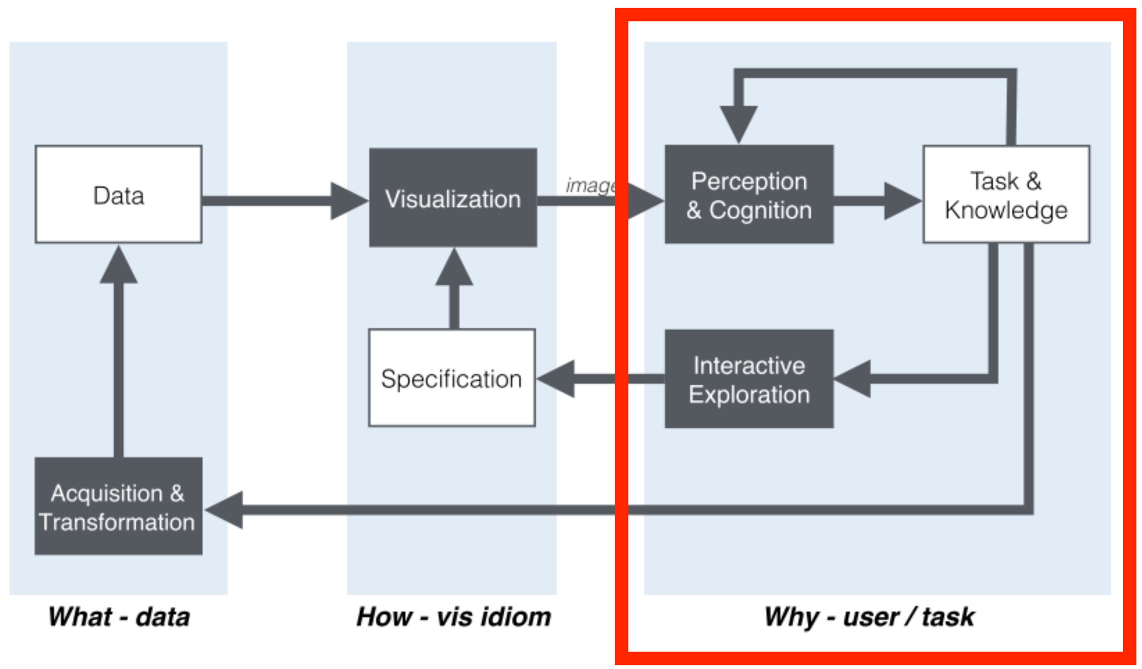
Get to know the data: explore in R

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**



Why - Understand the task/goal



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.

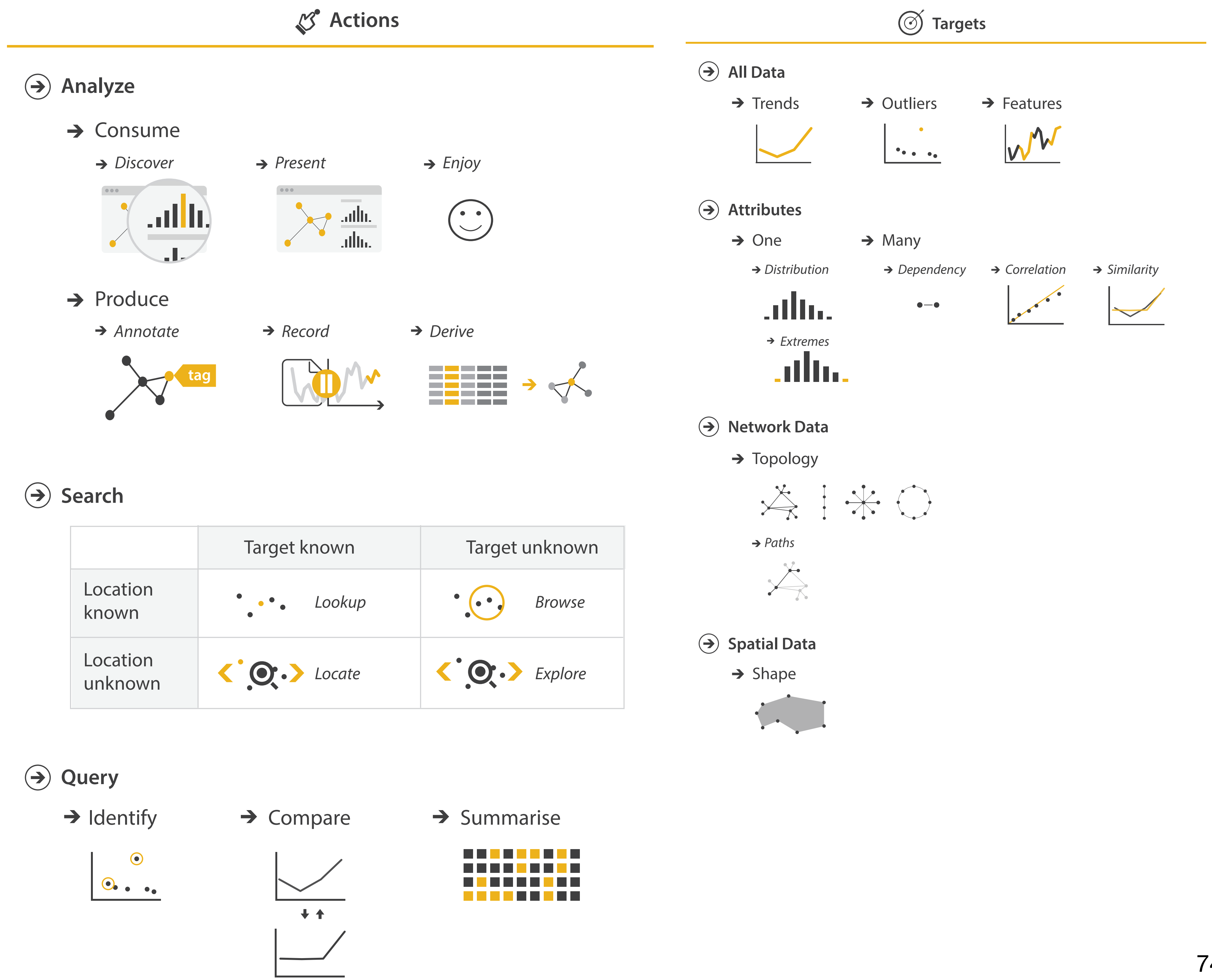
3. **Genius design** - Developer still does not look beyond own experience, but that experience is extensive.

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

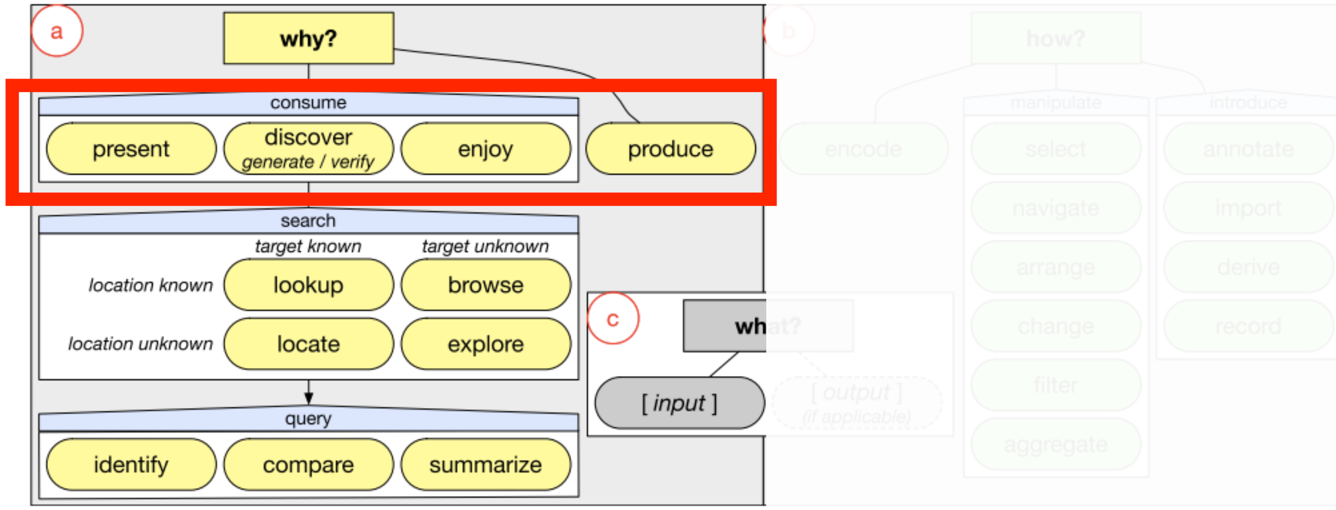
5. **Goal-focused design** - Developer goes further than activities and investigates goals, needs and contexts of the user.

why = pair **of action + target**

e.g. “discover distribution”, “compare trends”, “locate outliers”, “browse topology”, …

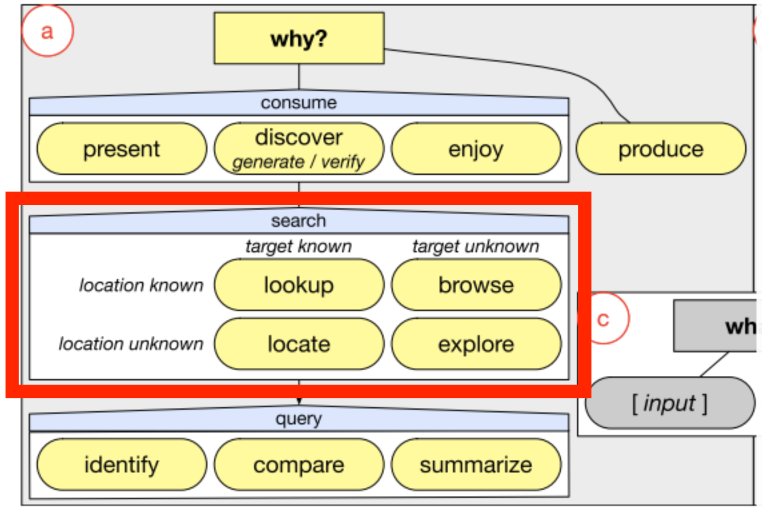


[why] Consume



- **present**: communicate message; guide audience through series of cognitive operations
  - different settings: teaching, decision making, planning, …; collaborative, live, pre-recorded, …
- **discover**: generation and verification of hypothesis
- **enjoy**: casual encounters with dataviz; novelty stimulates curiosity and therefore exploration
- **produce**: generate new artefacts (e.g. transformed/derived data, annotations, …)

[why] Search

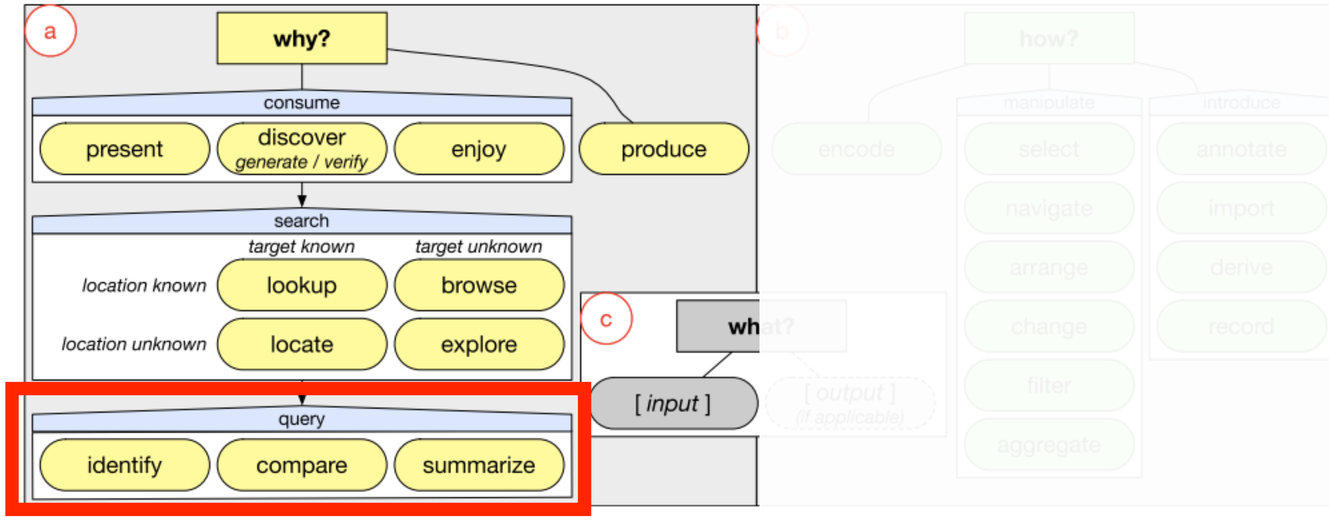


find elements of interest

categorized by *what* is being searched

- **lookup**: target known + location known; e.g. searching for Hoegaarden on a map of Belgium if you know where Hoegaarden is
- **locate**: target known + location unknown; e.g. searching for Hoegaarden on a map of Belgium if you do *not* know where Hoegaarden is
- **browse**: target unknown + location known; e.g. searching for leaf nodes with few siblings in tree visualisation
- **explore**: target unknown + location unknown, often starting from *overview* of the data; e.g. searching for outliers in scatterplot, anomalous spikes or periodic patterns in line plot of time-series data

[why] Query

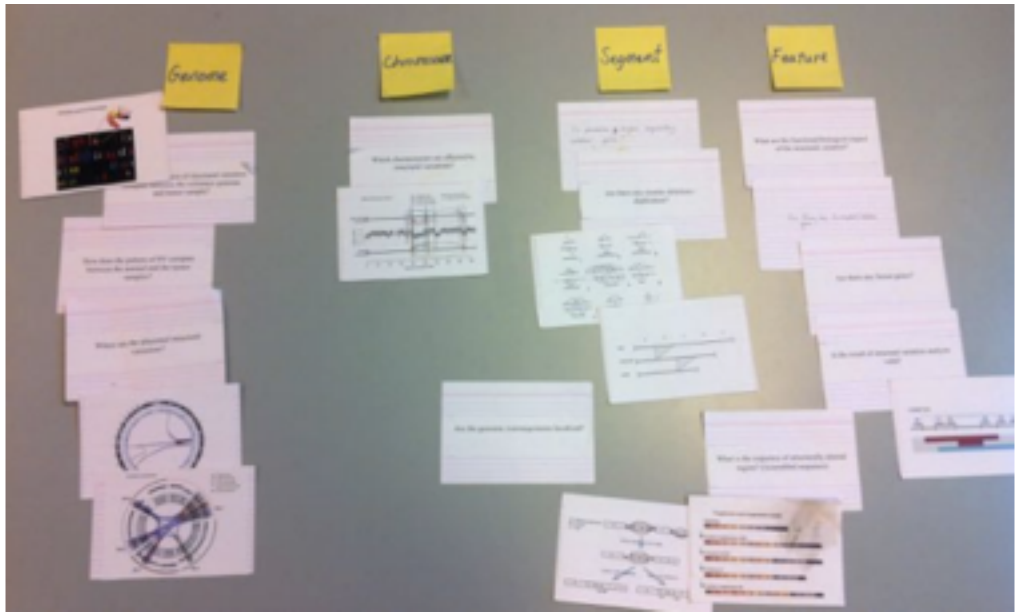


- **identify**: 1 search target
- **compare**: several search targets
- **summarize**: all search targets; often associated with overviews of data

How do you find out? Get to know the user

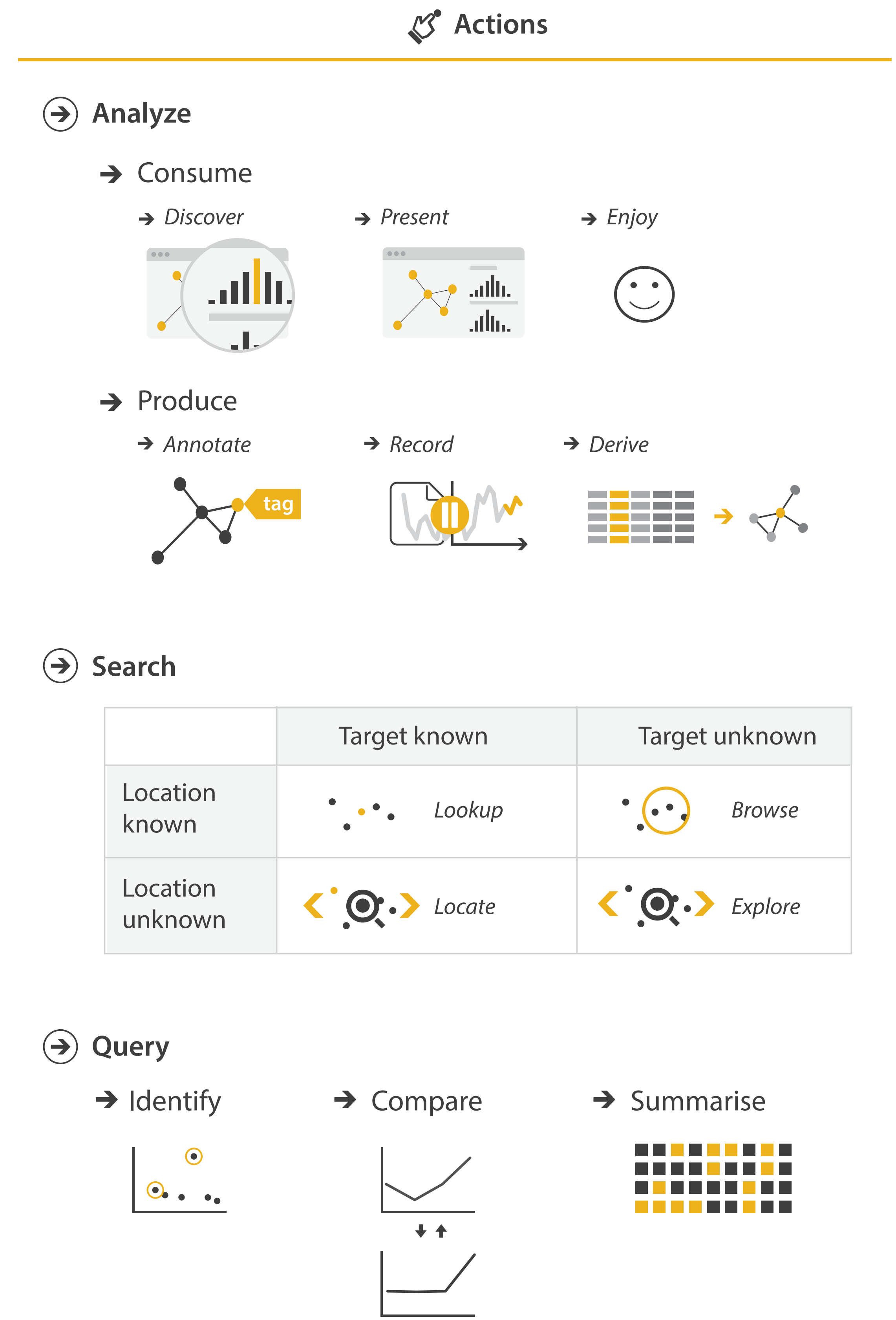
talk to the user (might or might not be yourself)

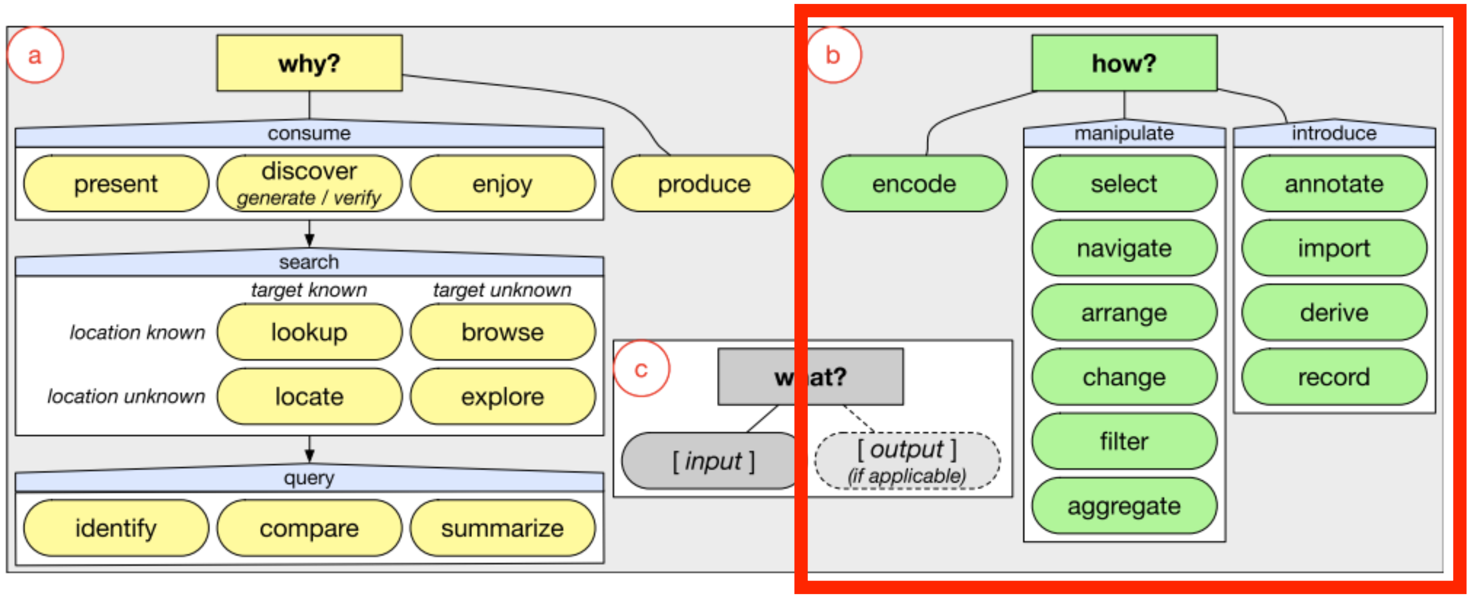
- what they *want* != what the *need* => need to find **underlying goals**
- e.g. let them **imagine** what they could do if some technologies were available that are (still) science-fiction (e.g. nanobots in blood; Gaviscon commercial http://m.youtube.com/watch?v=_skKmcLdyVQ)
- additional methods, e.g. **card sorting**
- if possible: tape the discussion (w/ agreement)
- ask "why?" 3 times



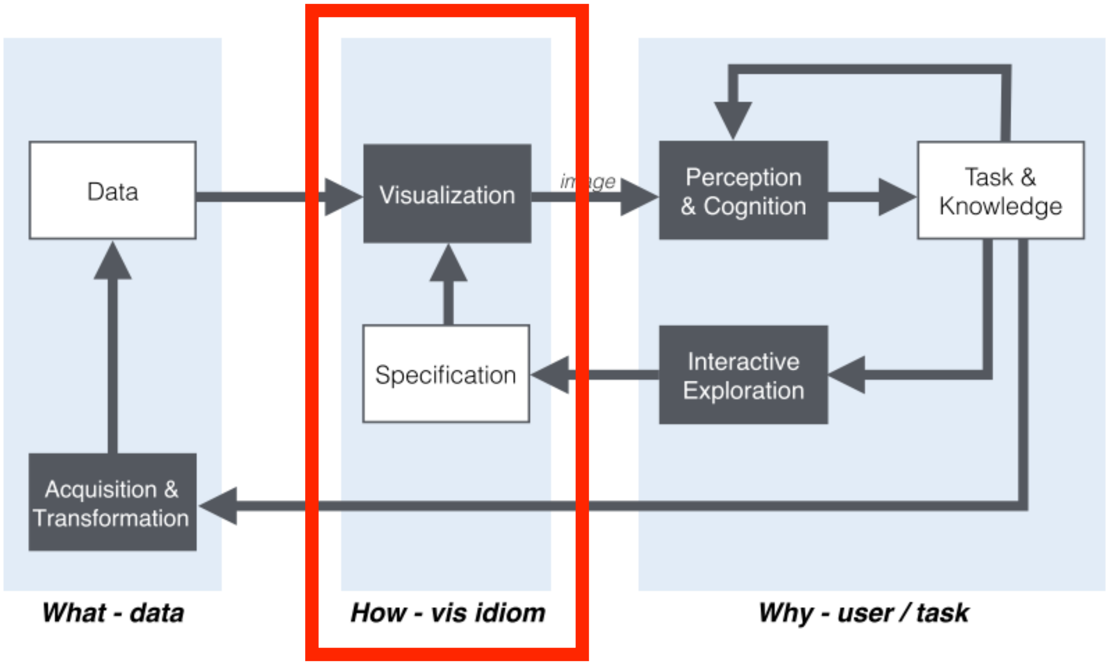
2. Post-hoc analysis of interview (e.g. using Munzner taxonomies)

- task abstraction
- see examples later

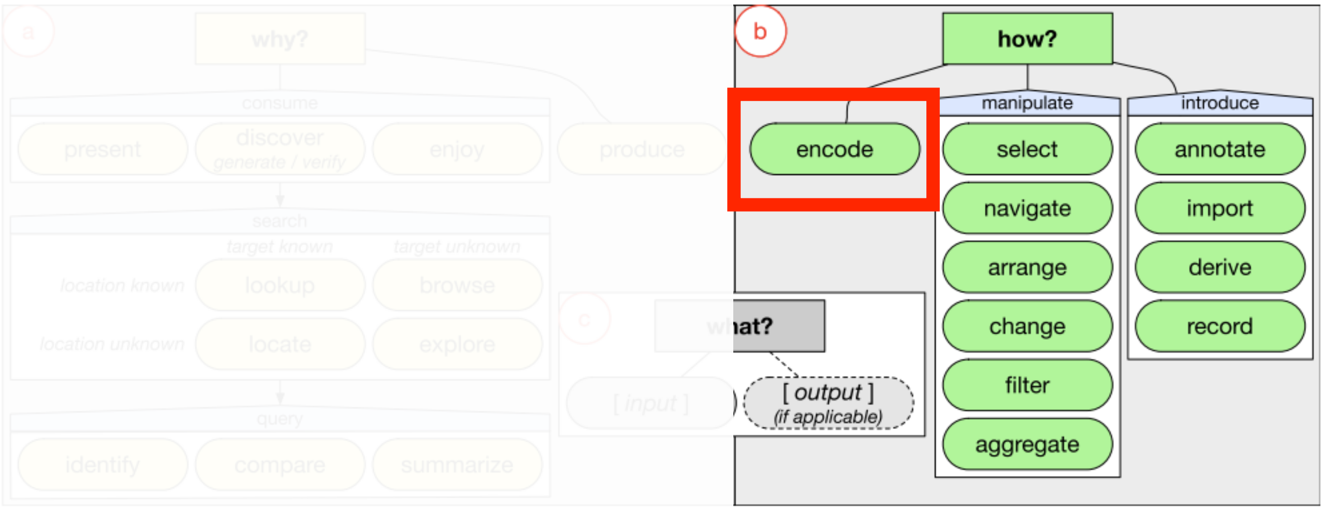




How - Building visualizations



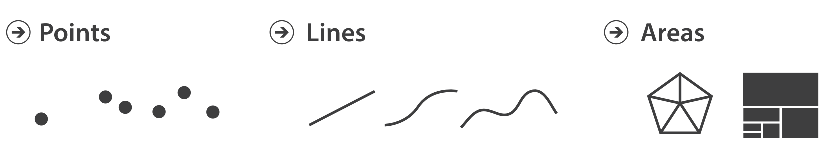
[how] Encode

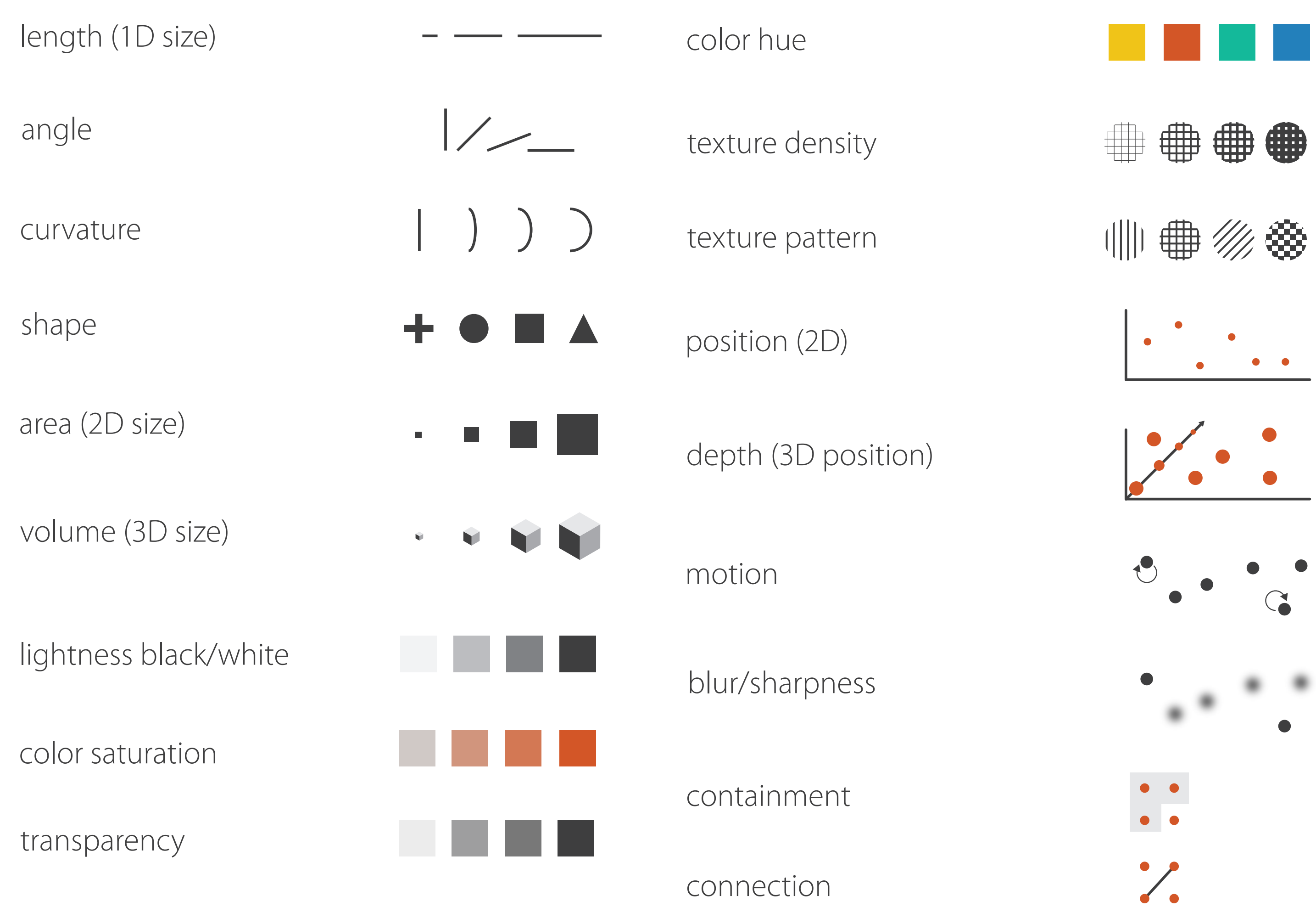


- **encode**: visual encoding = matching marks to channels

“channel”

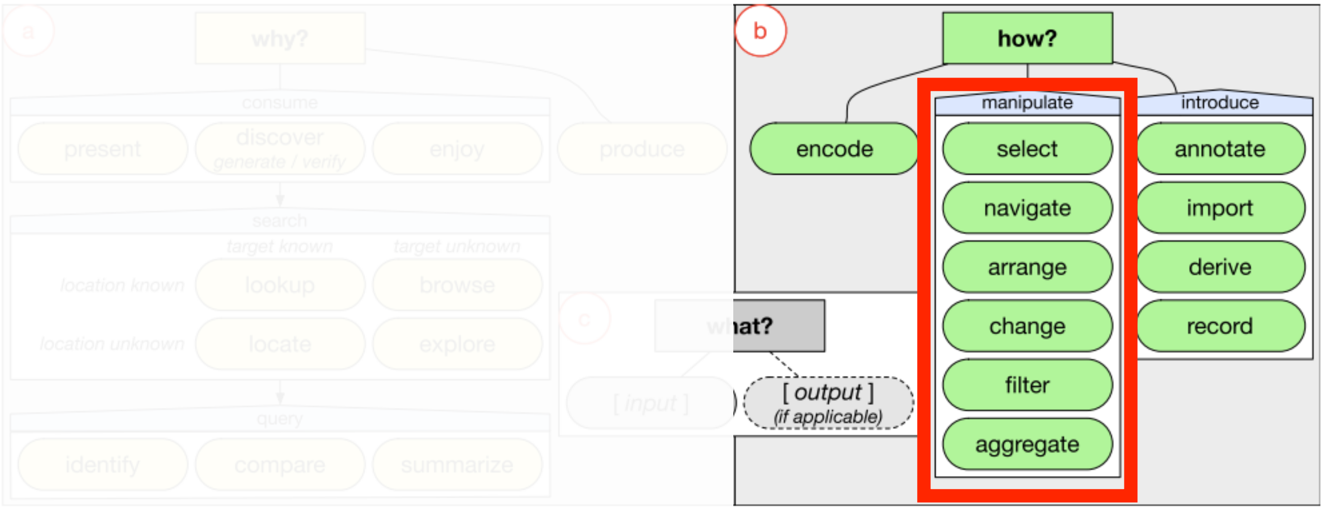
“mark”





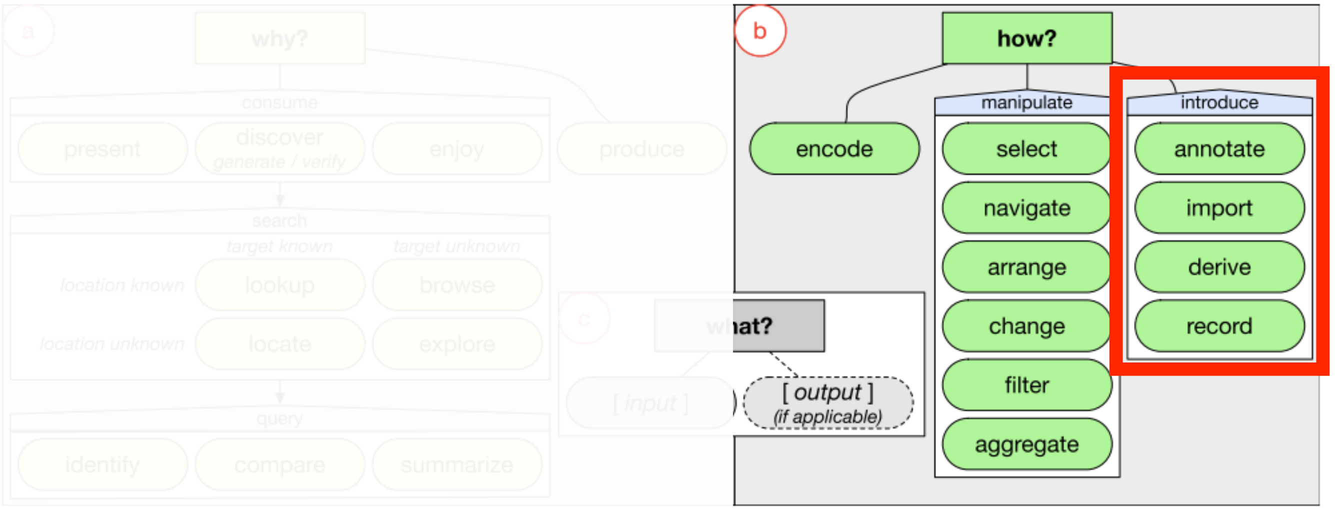
See further “What visual encoding should I use?”

[how] Manipulate



- **select**: demarcate one or more elements => distinction between selected and non-selected; e.g. by clicking or dragging
- **navigate**: alter a user's viewpoint (zoom, pan, rotate)
- **arrange**: organizing visual elements spatially (e.g. reordering axes in parallel coordinates plot)
- **change**: change in visual encoding (small: transparency or size of points in scatterplot; large: switch from linear to radial layout). Ideally: using animated transitions
- **filter**: adjust exclusion/inclusion criteria for elements. Temporary vs permanent
- **aggregate/segregate**: change granularity of visual elements. E.g. aggregate clique of nodes in network into single node.

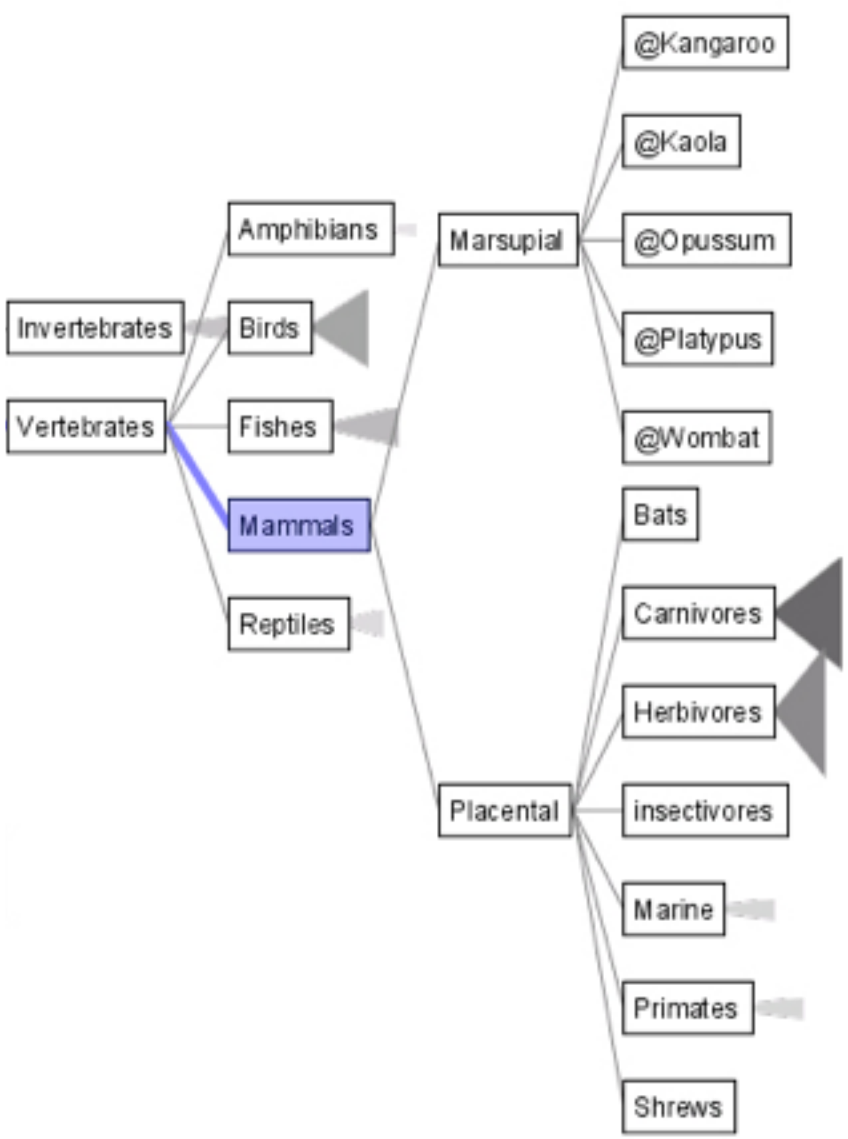
[how] Introduce



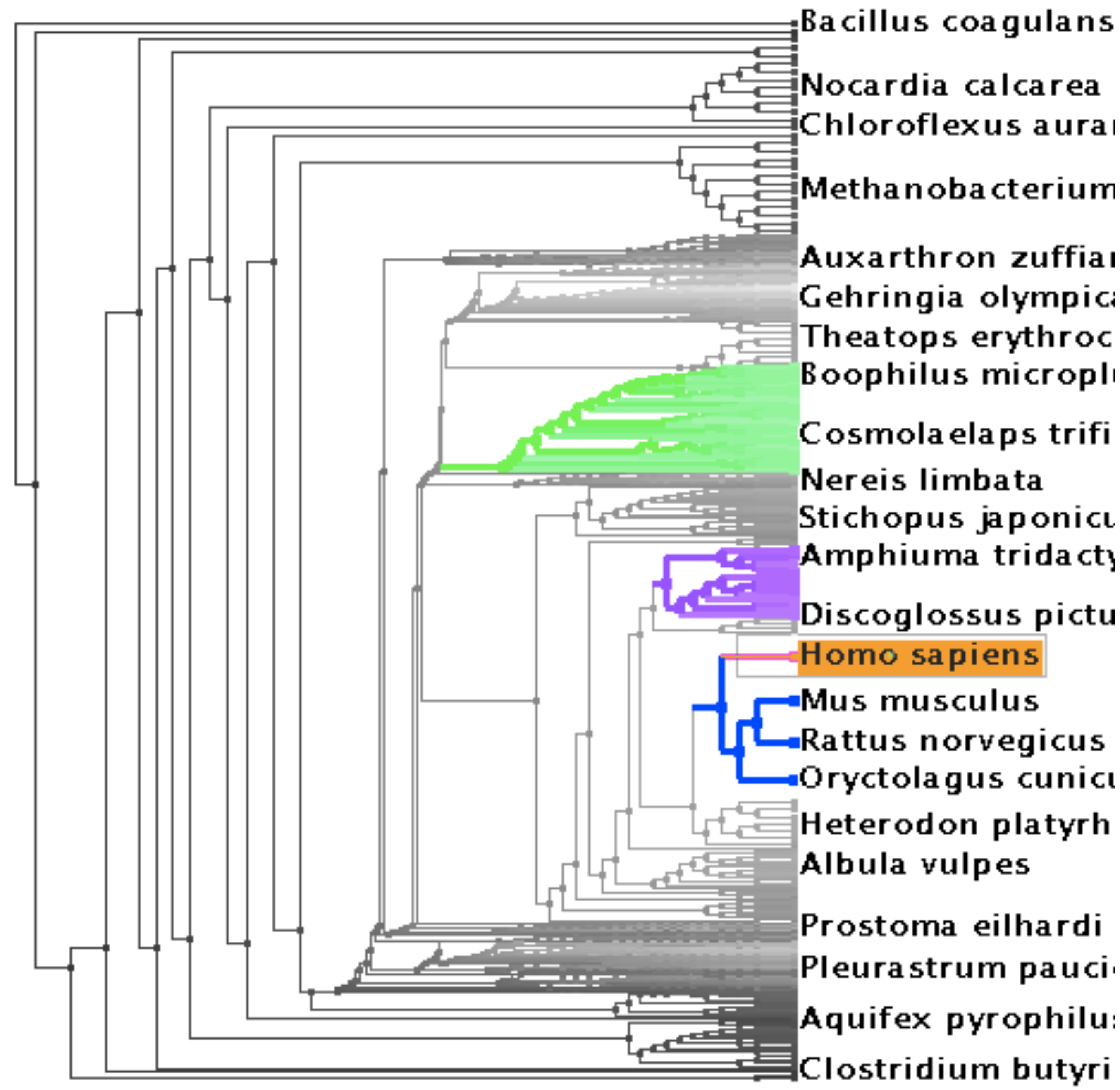
- **annotate**: if associated with particular data element: = new feature of that element. E.g. add label to point in scatterplot
- **import**: add new data, either from external source or by hand
- **derive**: Don't just draw what you're given => data transformations (e.g. dimensionality reduction, TDA, deltas, …; exports and imports => trade balance)
- **record**: save visualization artefacts (e.g. screenshots)

Example of analysis

Compare SpaceTree vs TreeJustaposer



Plaisant et al, 2002



Munzner et al, 2003

What?

- Tree

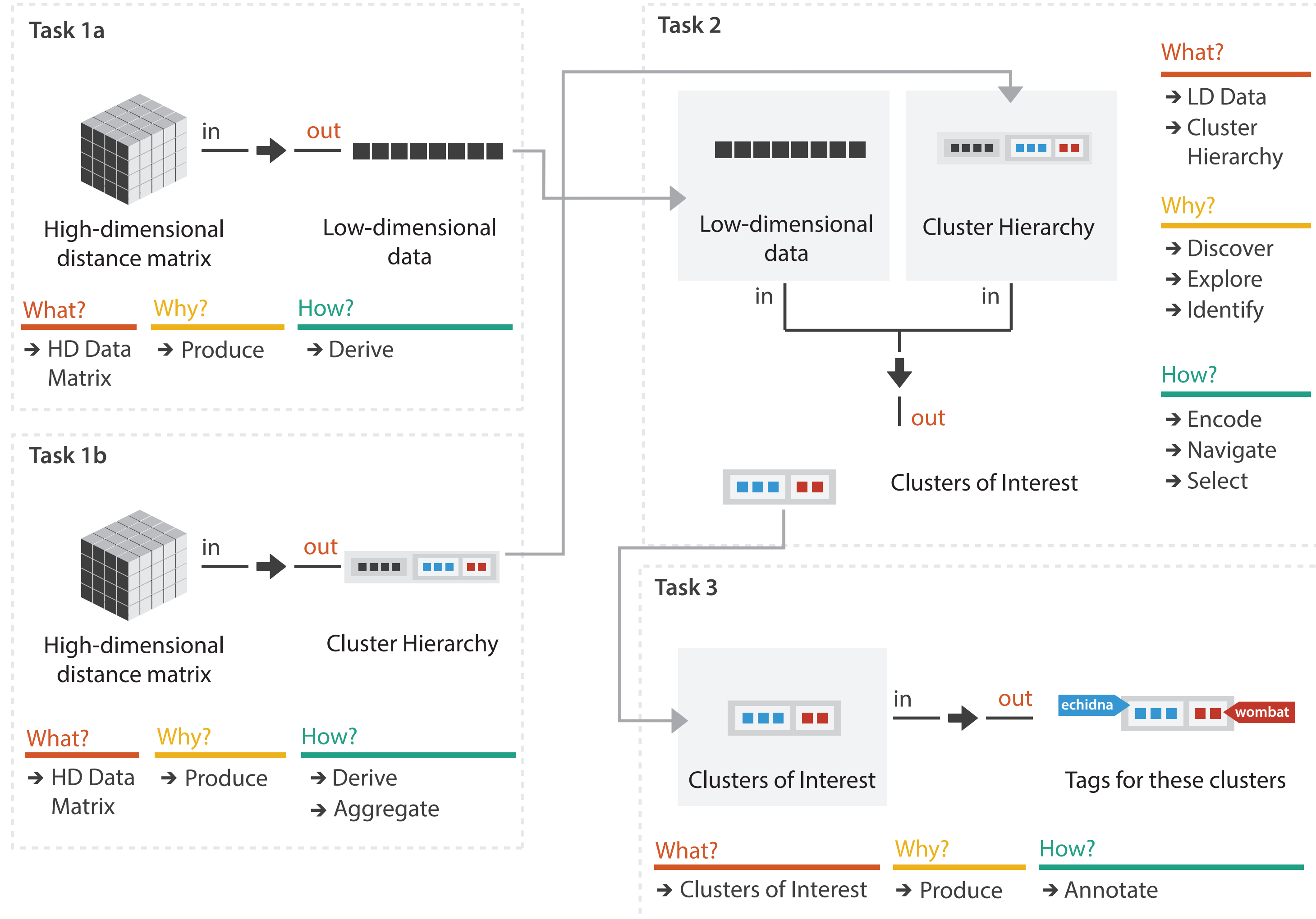
Why?

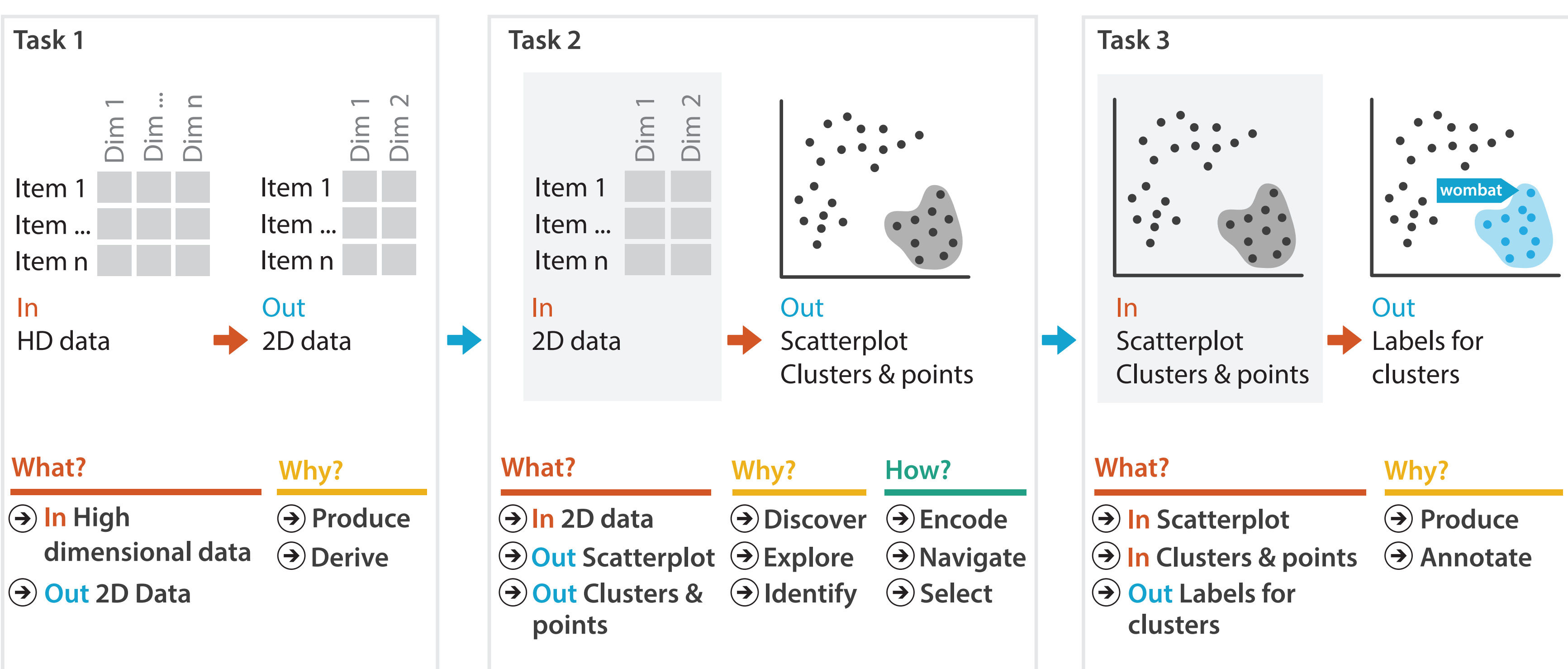
- Actions
  - Present
  - Locate
  - Identify
- Targets
  - Path between two nodes

How?

- SpaceTree
  - Encode
  - Navigate
  - Select
  - Filter
  - Aggregate
- TreeJuxtaposer
  - Encode
  - Navigate
  - Select
  - Arrange

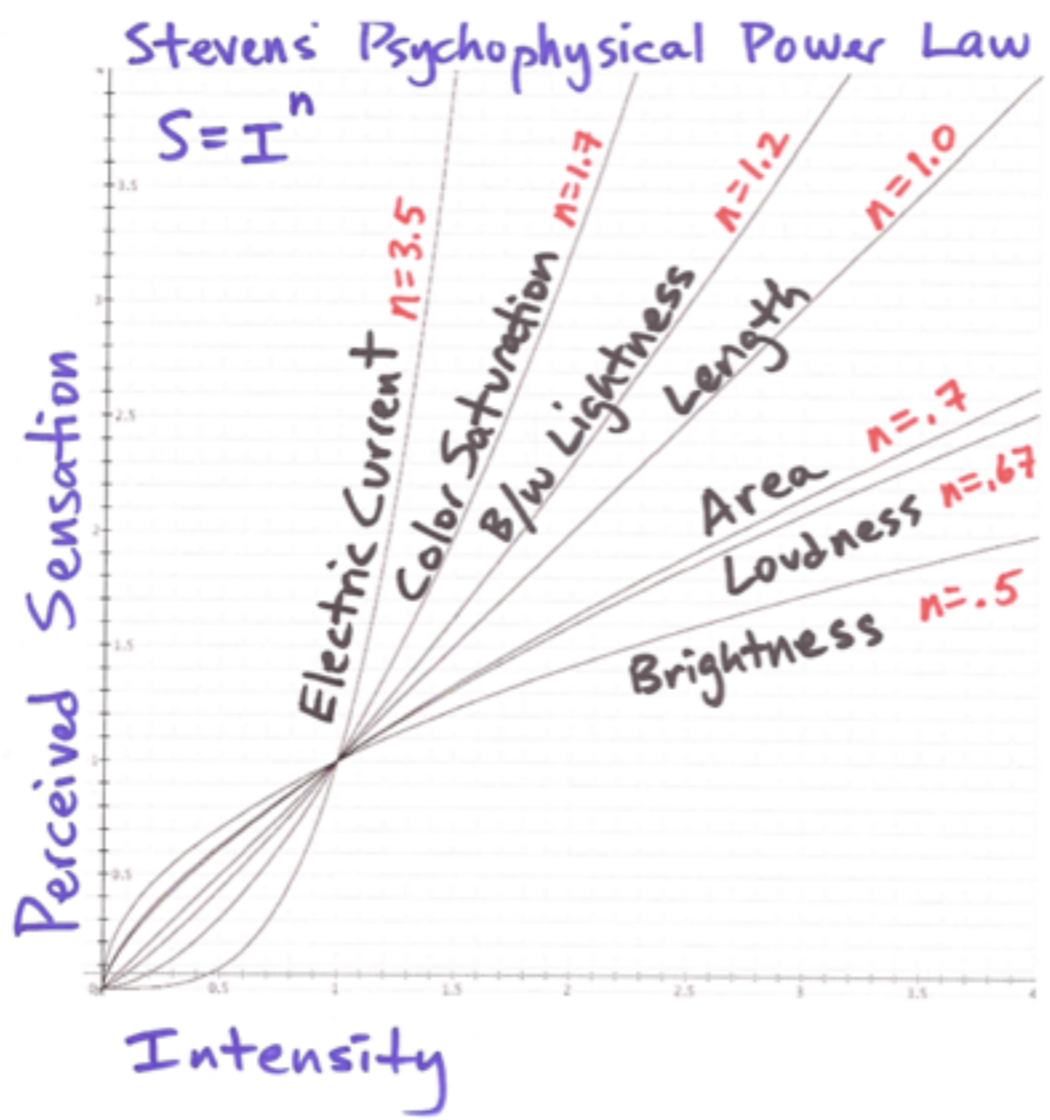
More involved examples…



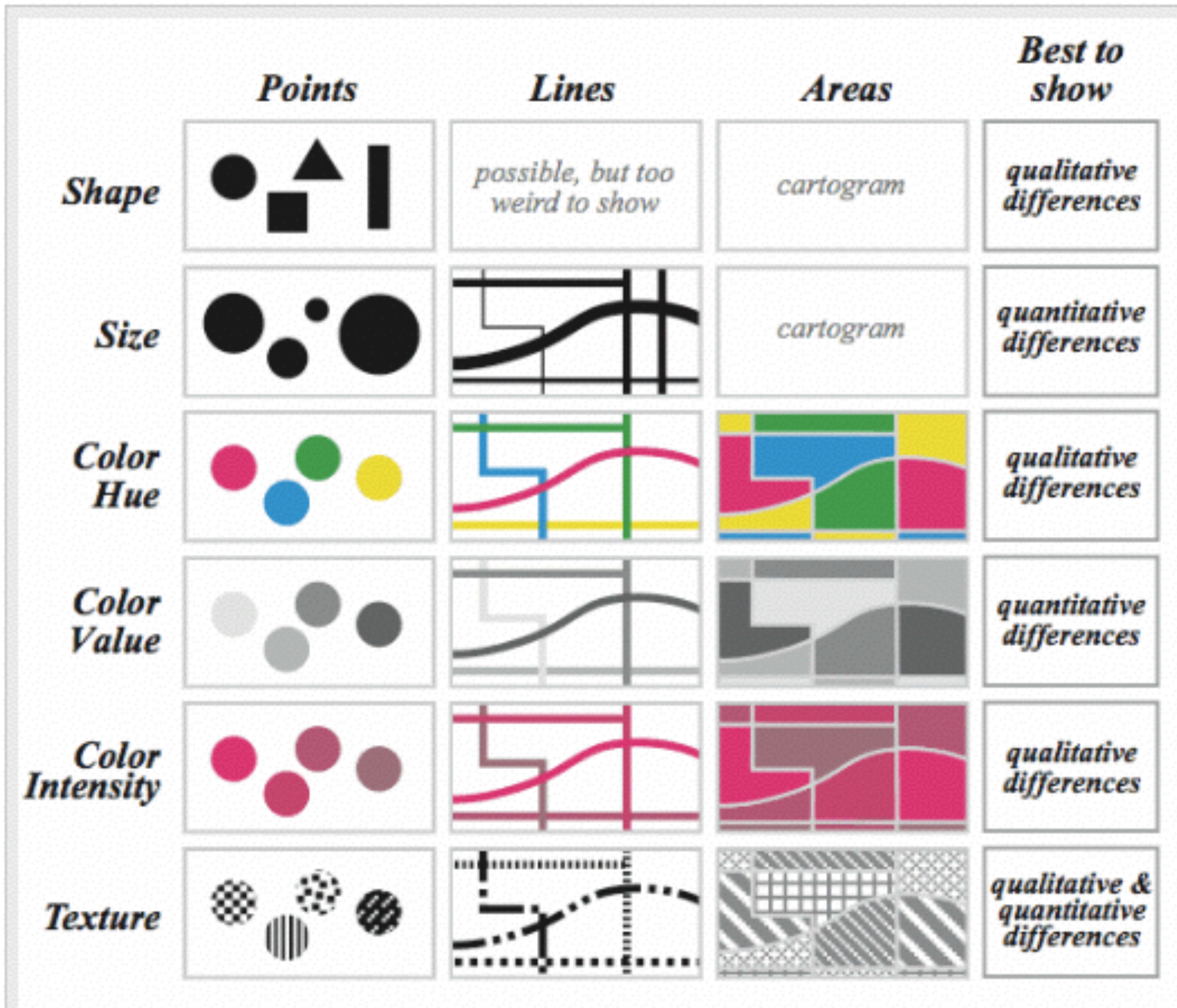


What visual encoding should I use?

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

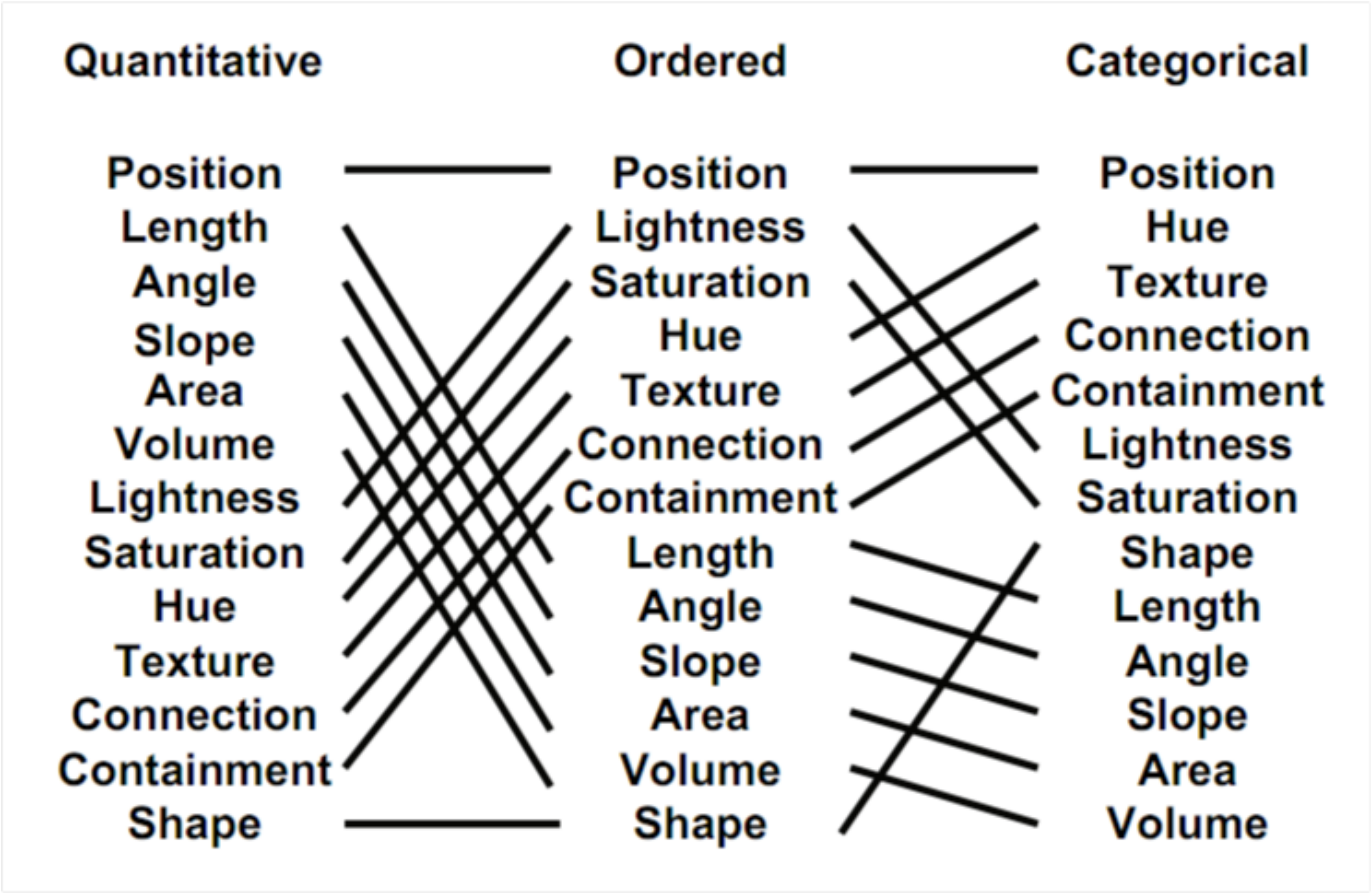


Semiology of graphics



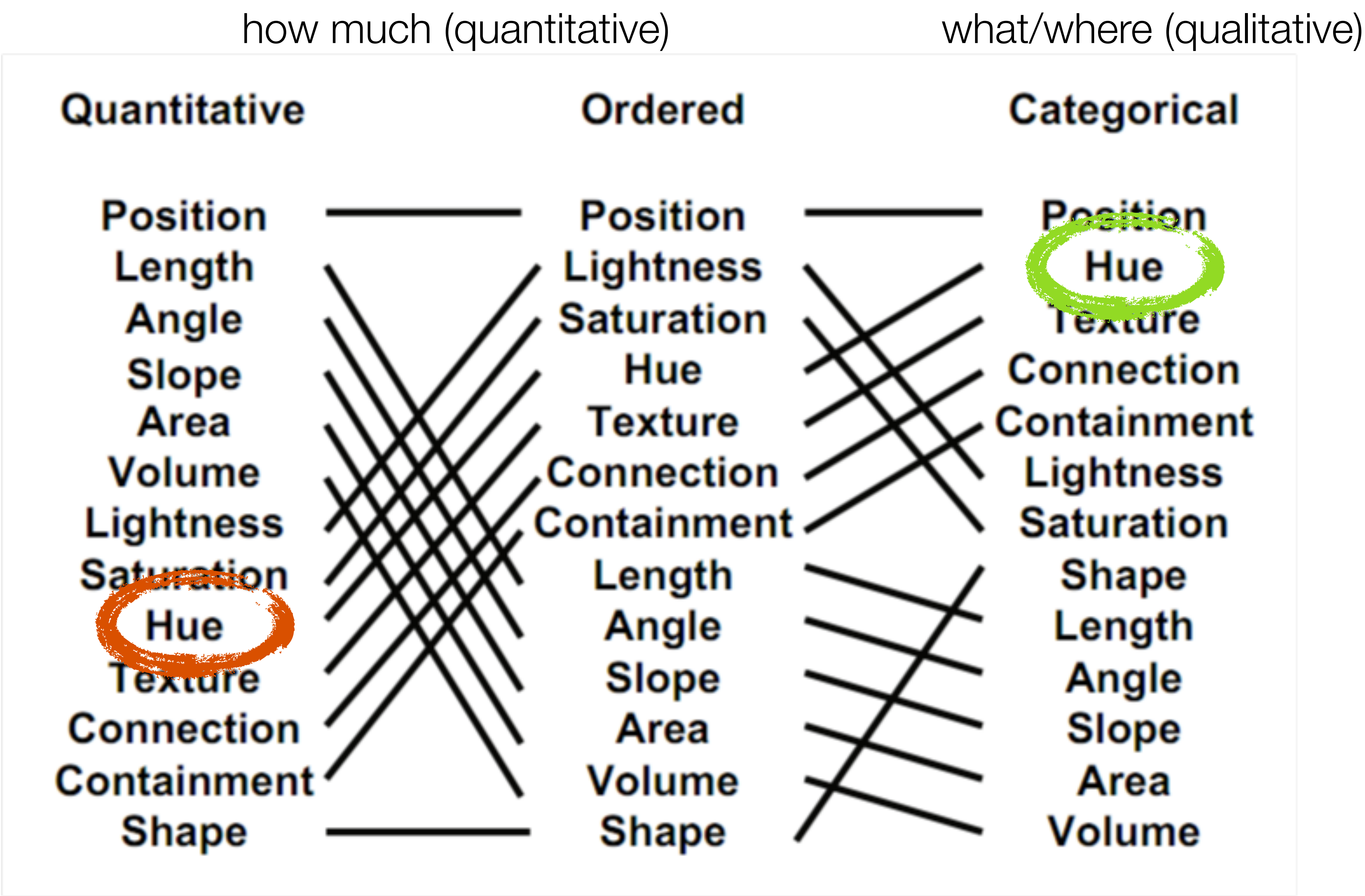
Accuracy of quantitative perceptual tasks

how much (quantitative) what/where (qualitative)



McKinlay

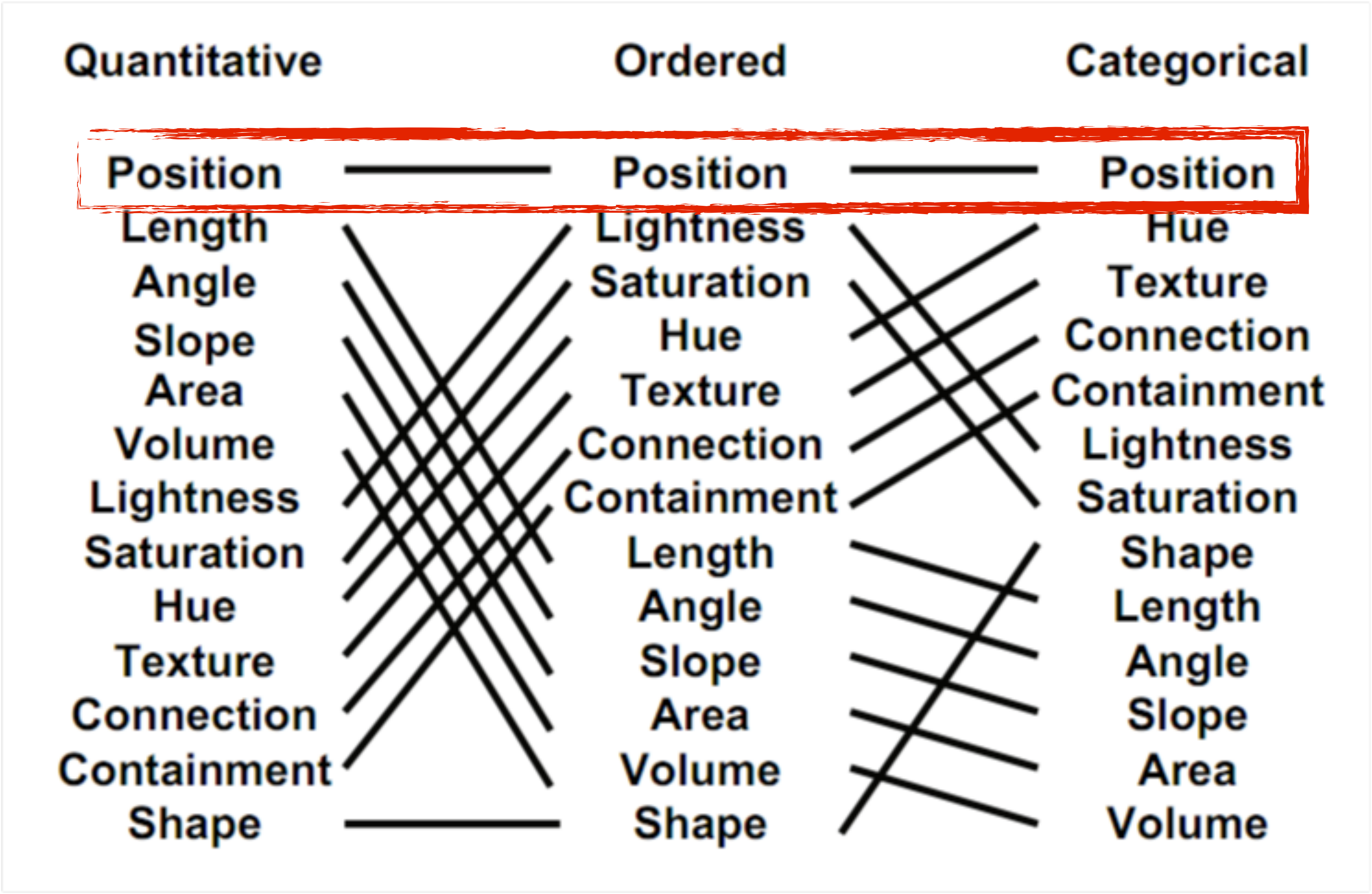
Accuracy of quantitative perceptual tasks



McKinlay

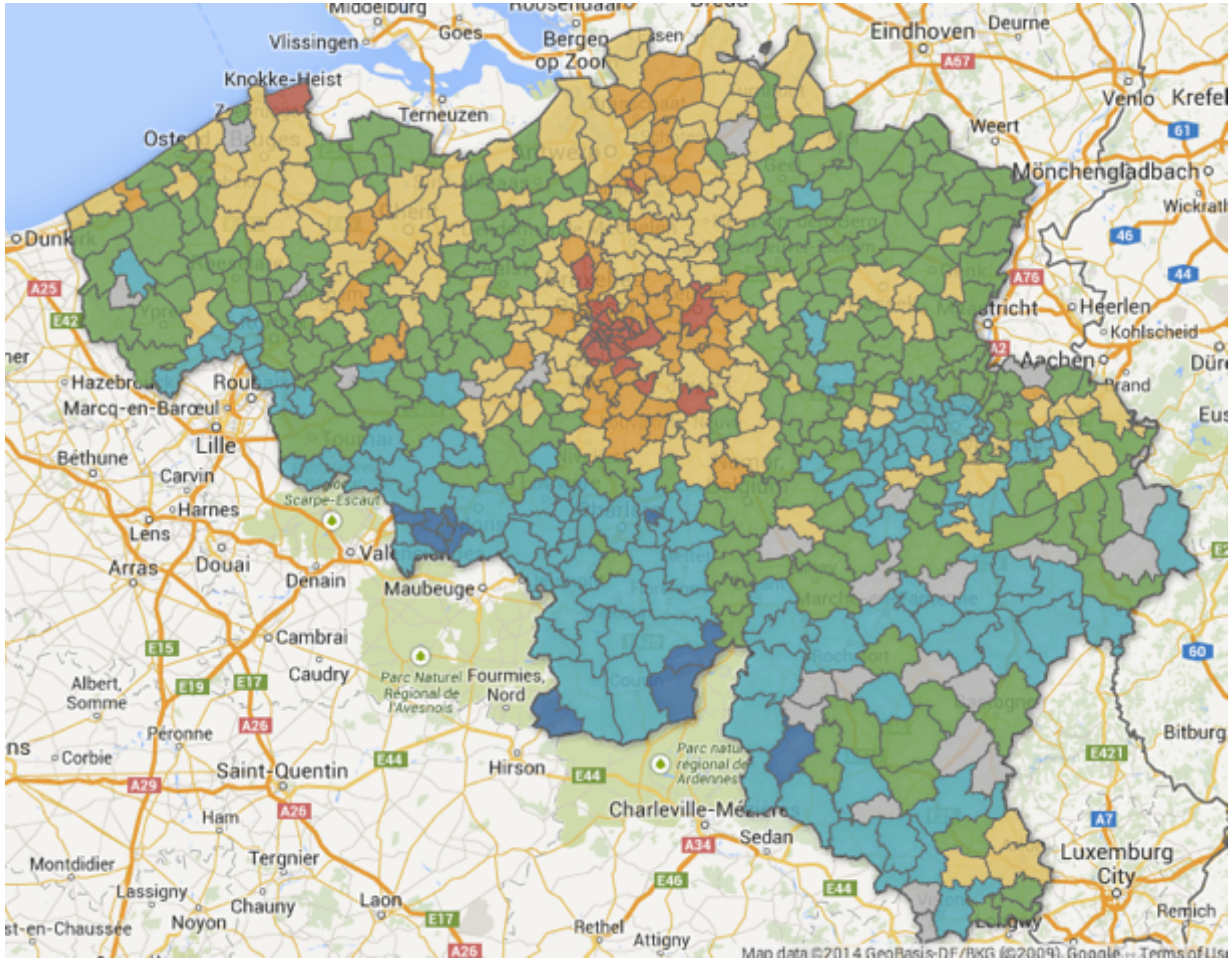
Accuracy of quantitative perceptual tasks

how much (quantitative) what/where (qualitative)

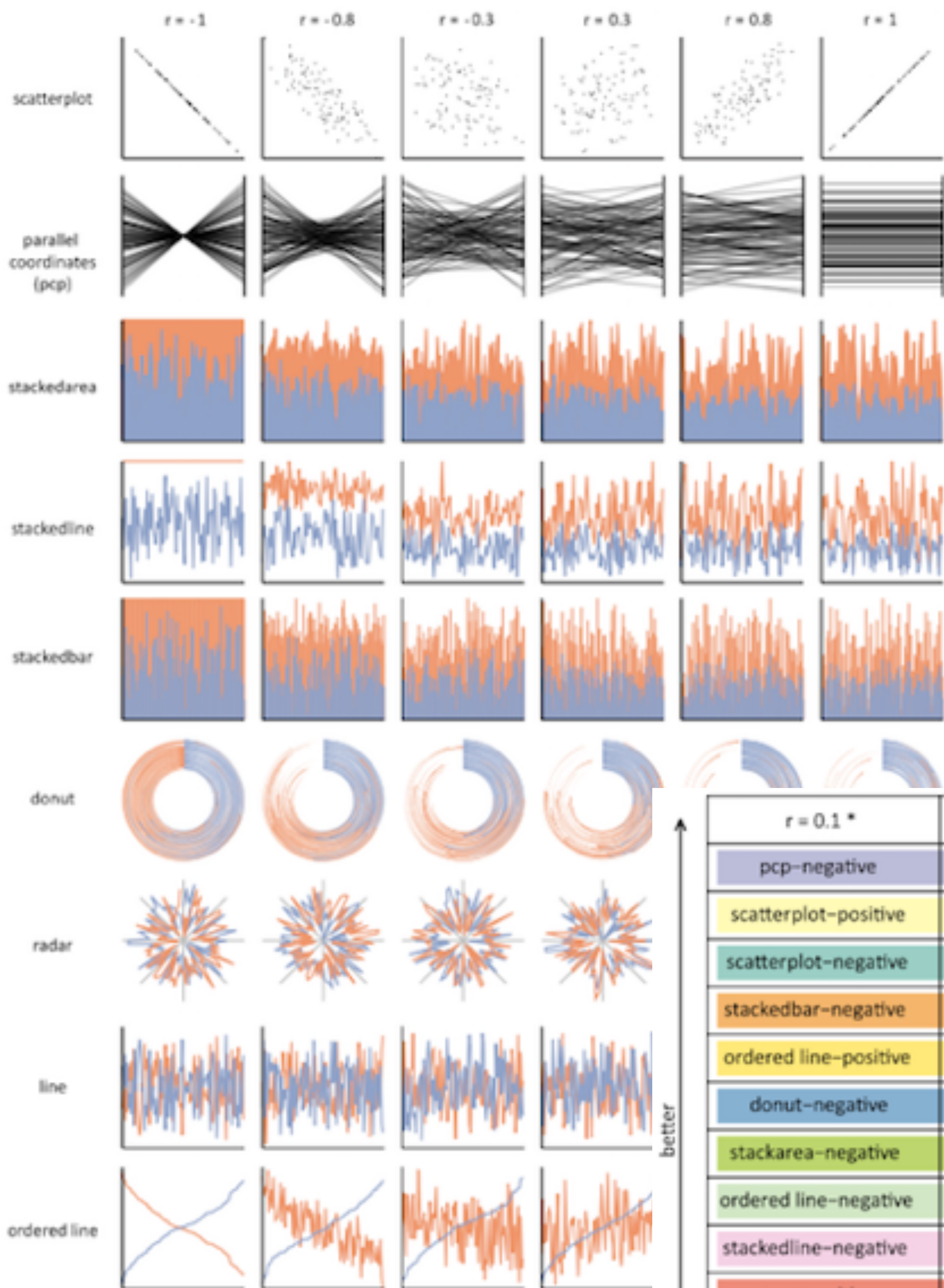


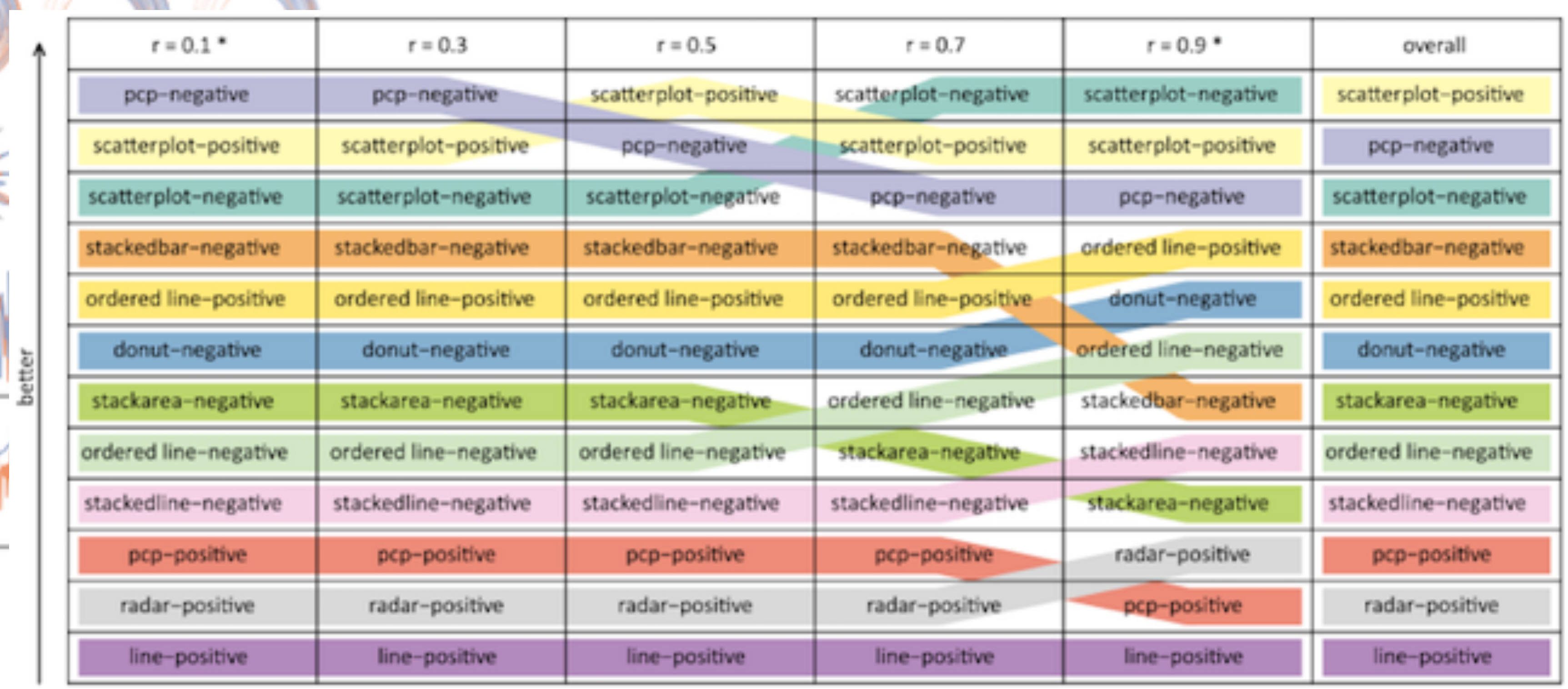
McKinlay

"power of the plane"



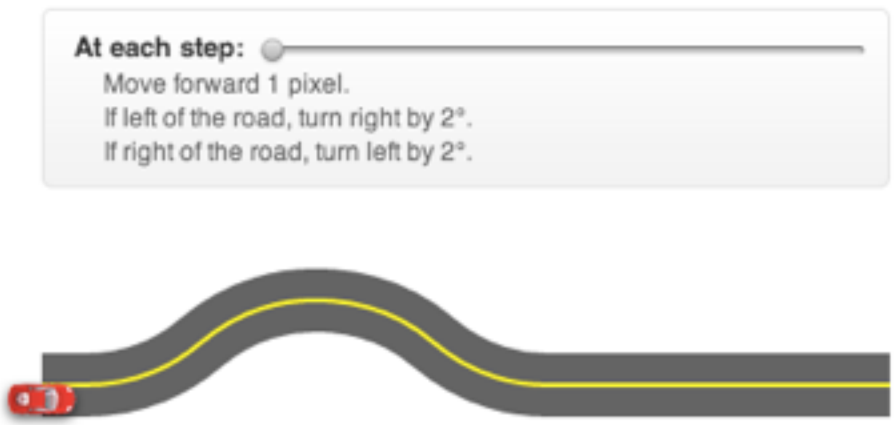
average prices per
municipality in Belgium
(De Standaard, 24/8/2014)

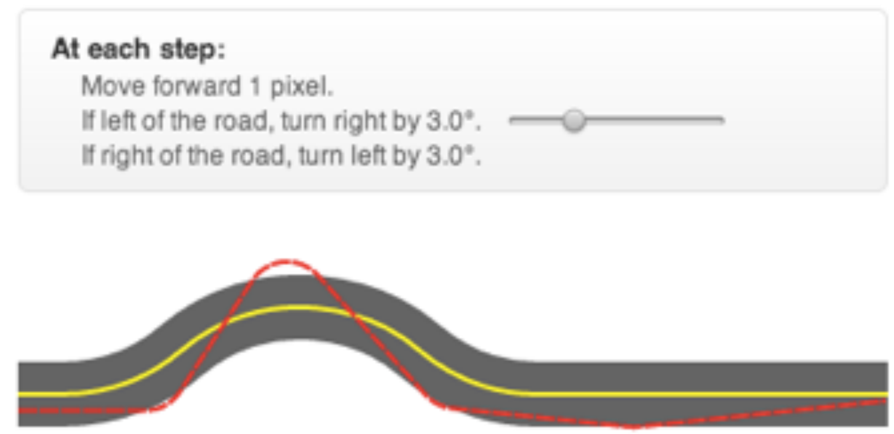


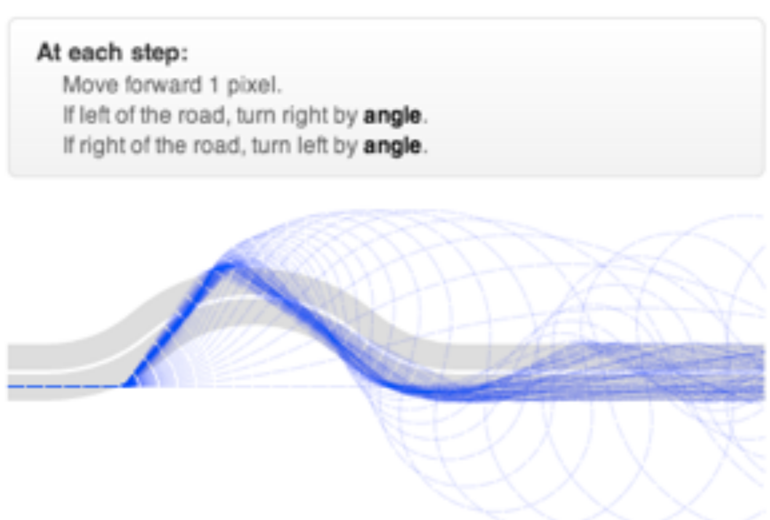


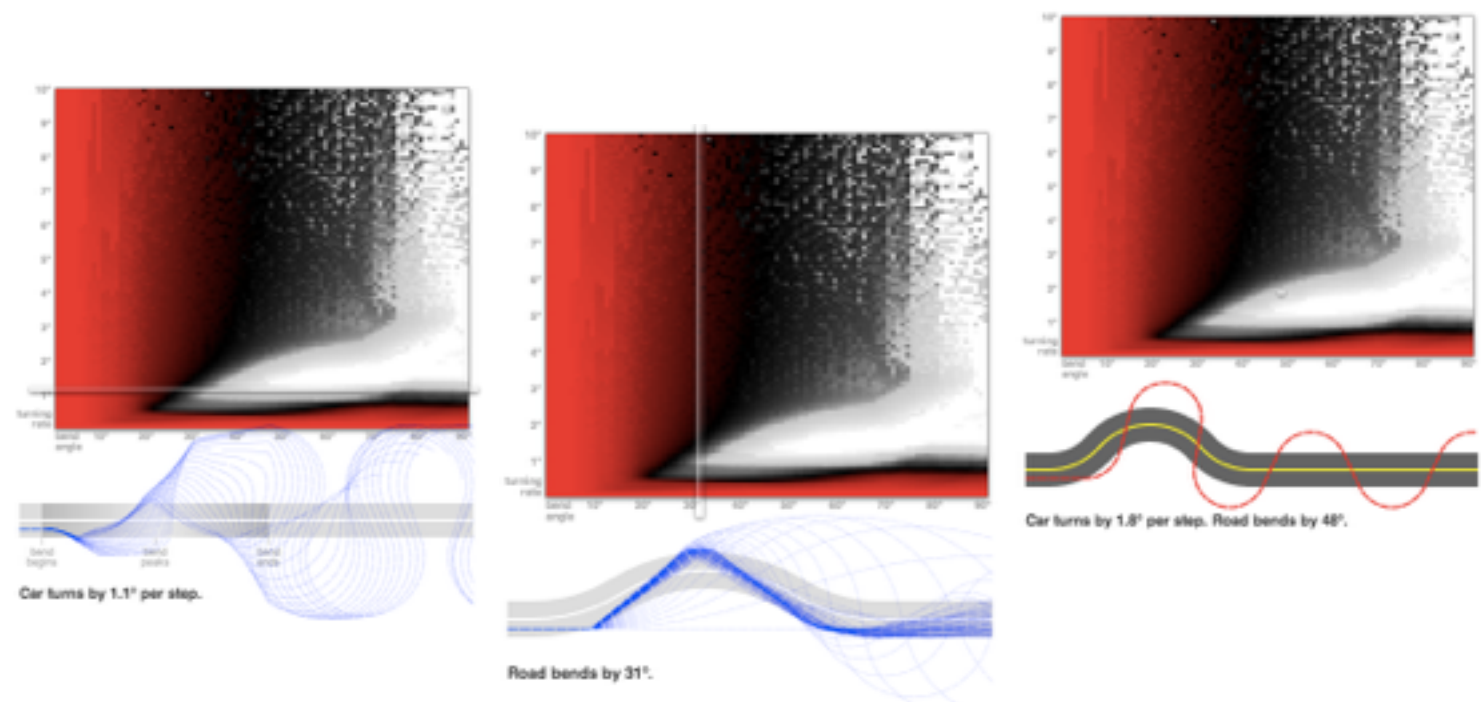
Harrison et al, 2014

Bret Victor - Ladder of abstraction

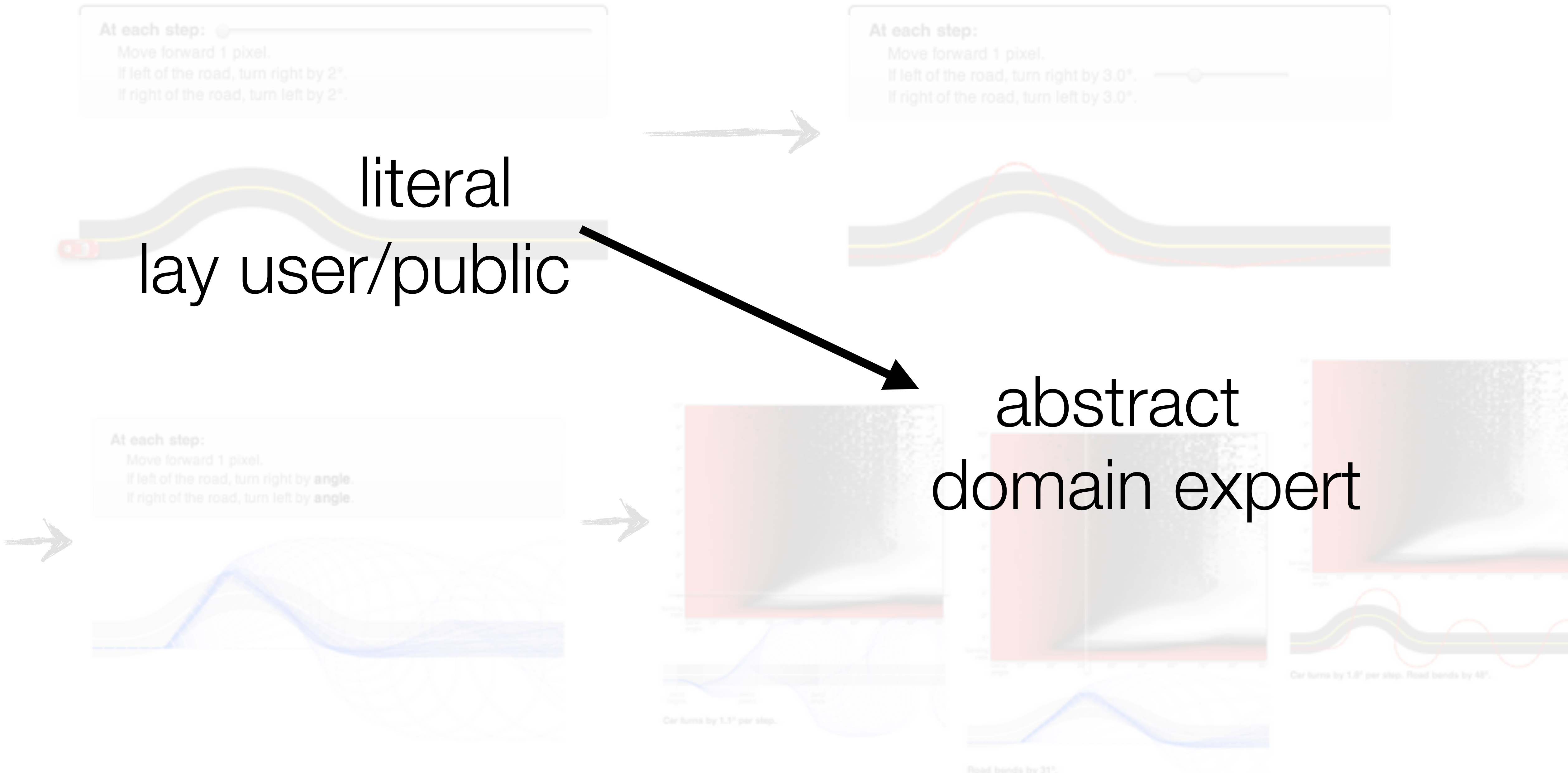








Bret Victor - Ladder of abstraction

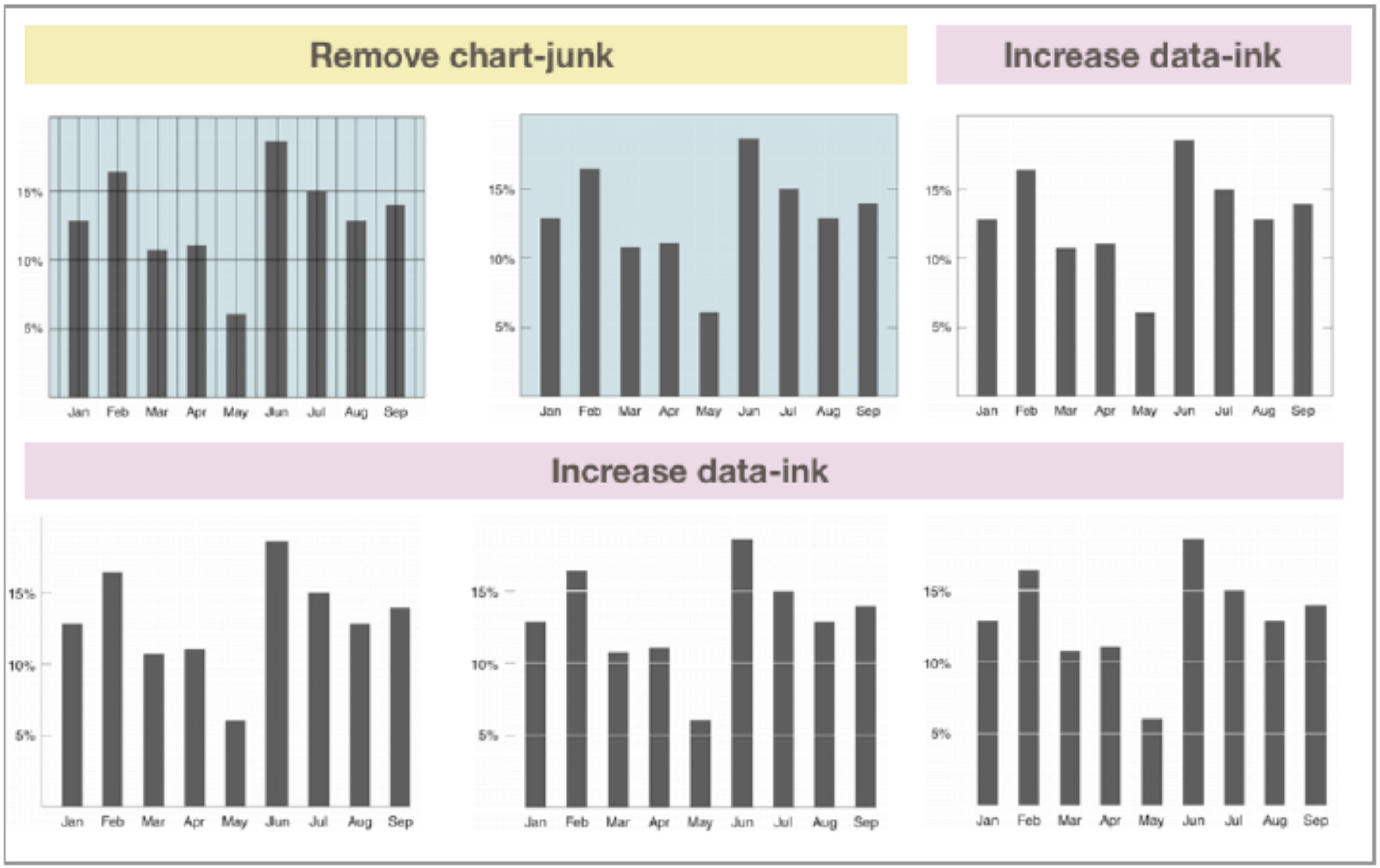


Rules of thumb

(partly based on book Munzner)

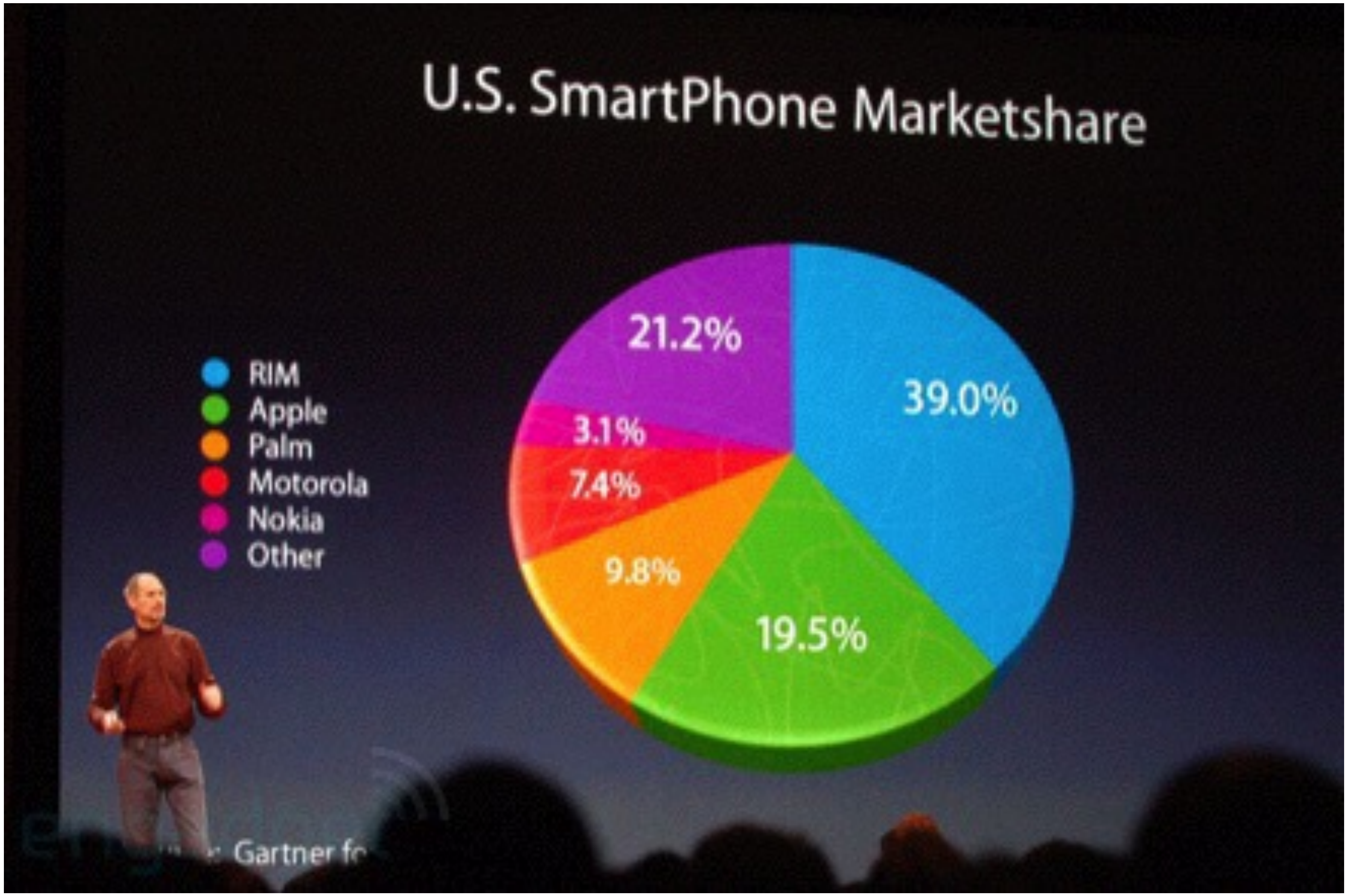
1. maximize data-to-ink ratio (Edward Tufte)
2. beware of the lie-factor
3. no unjustified 3D
4. eyes beat memory
5. focus + context
6. overview first, zoom & filter, details on demand
7. don't overengineer

1. maximize data-to-ink ratio

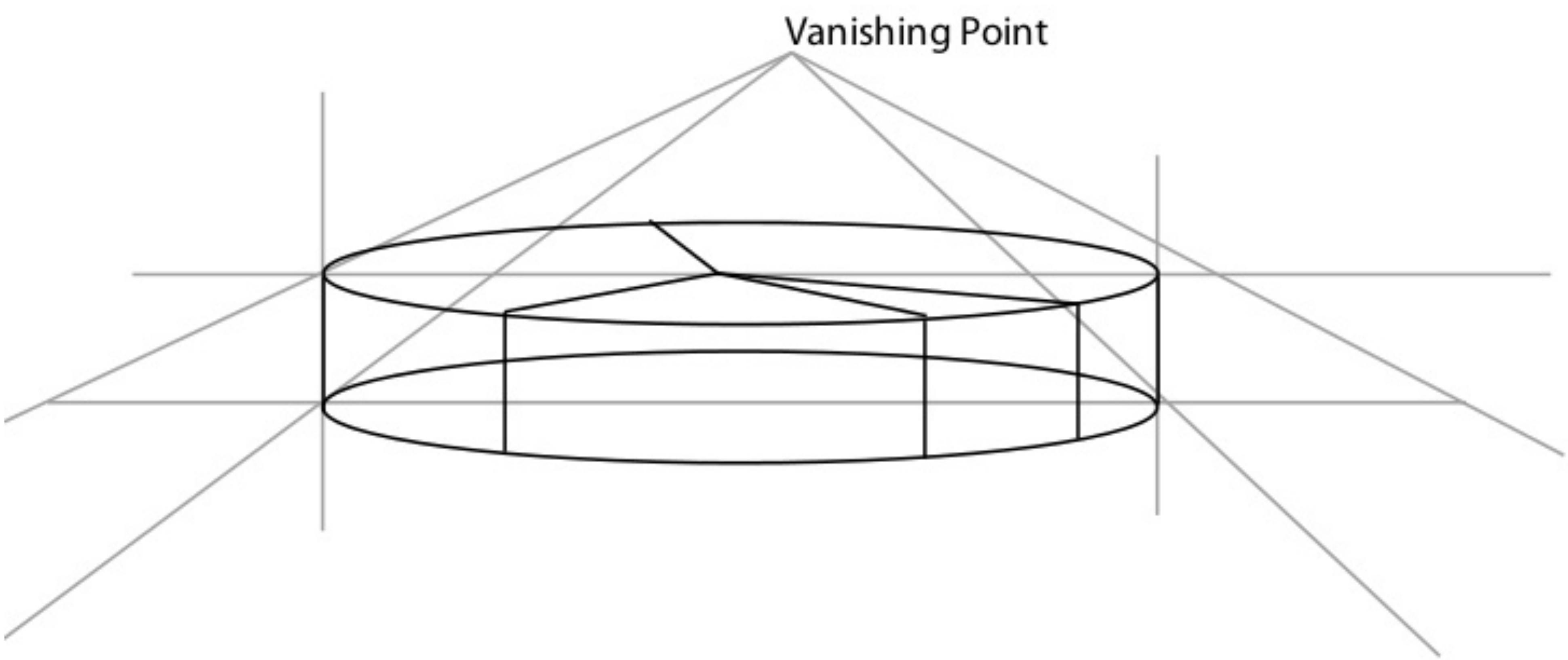


$$\textbf{“data-to-ink ratio”} = \frac{\text{data-ink}}{\text{total ink}} = 1 - \text{proportion of graphic that can be erased}$$

2. beware of the lie-factor



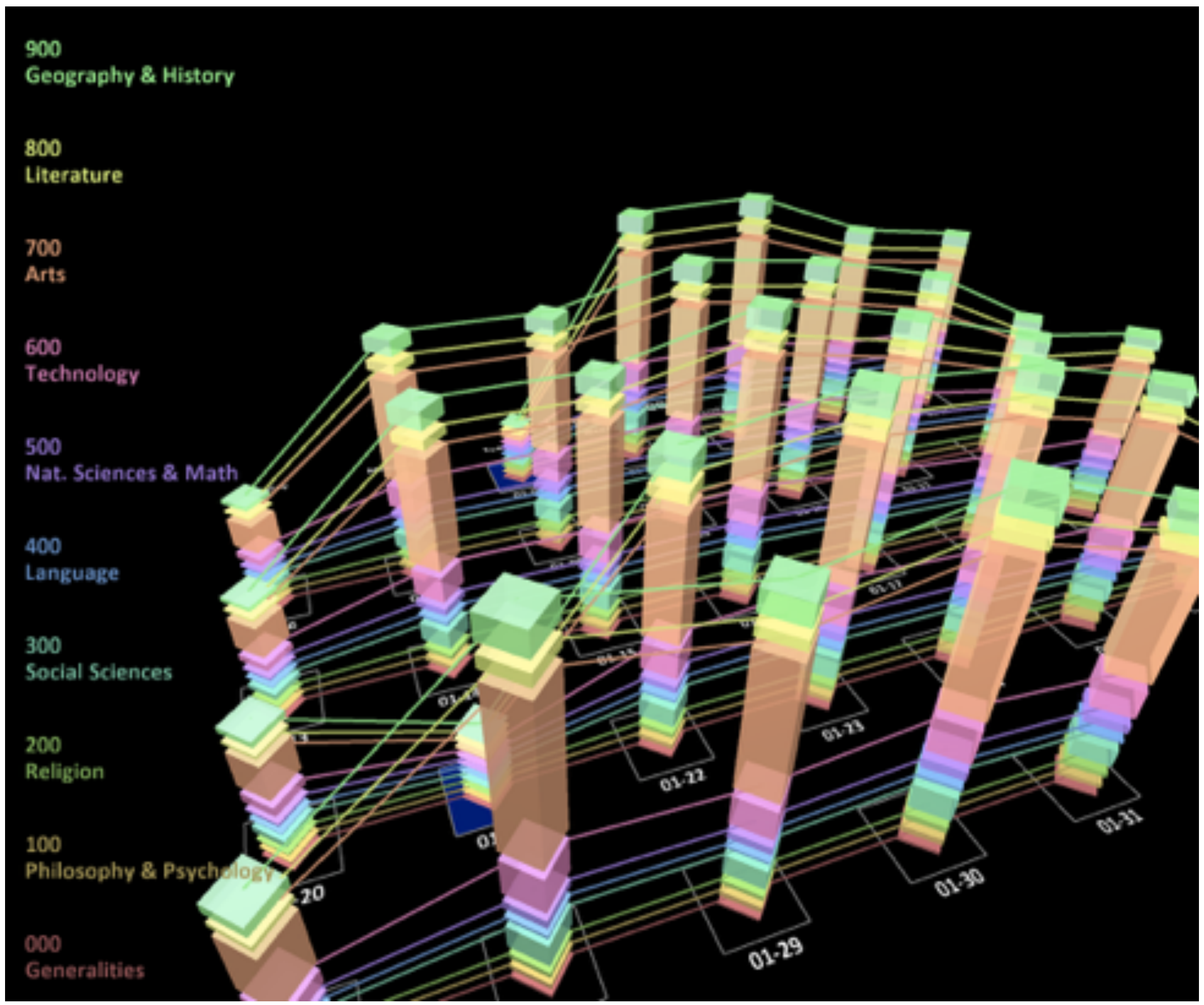
3D Charts!

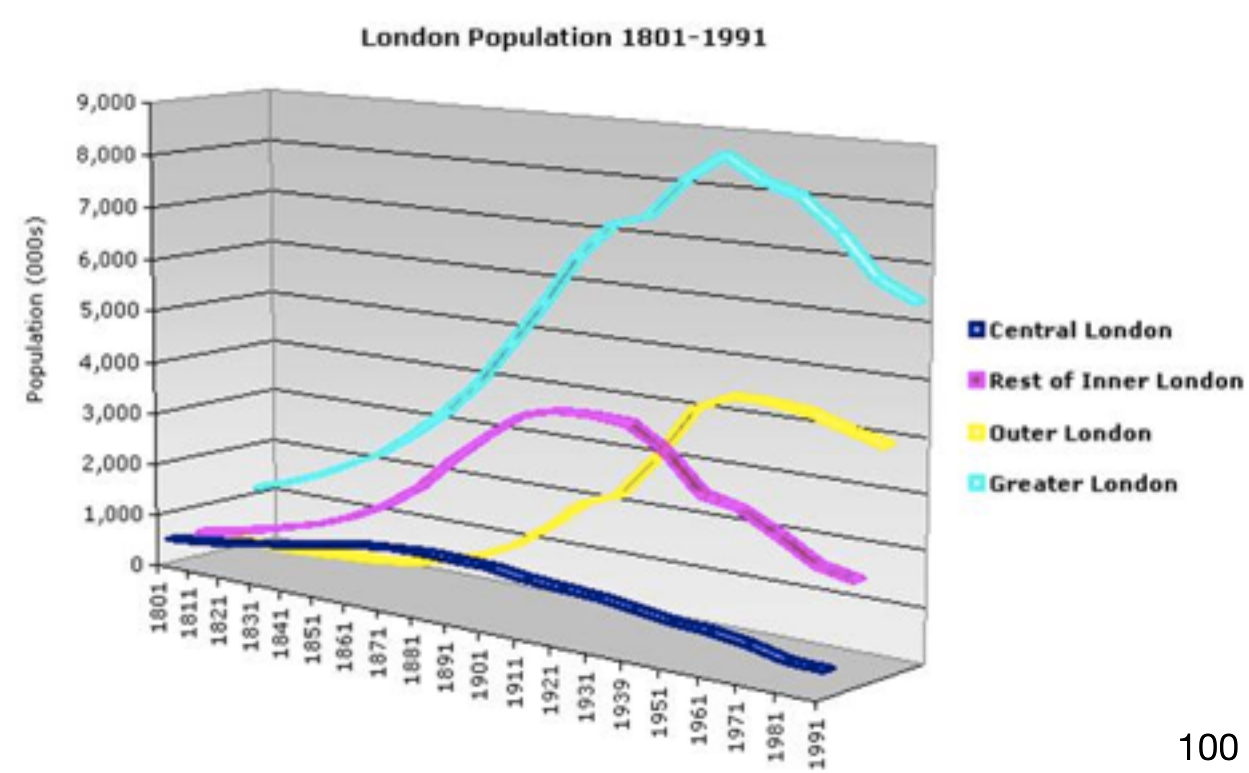


$$\textbf{“lie factor”} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

3. no unjustified 3D

issues with occlusion, perspective distortion, text legibility, …



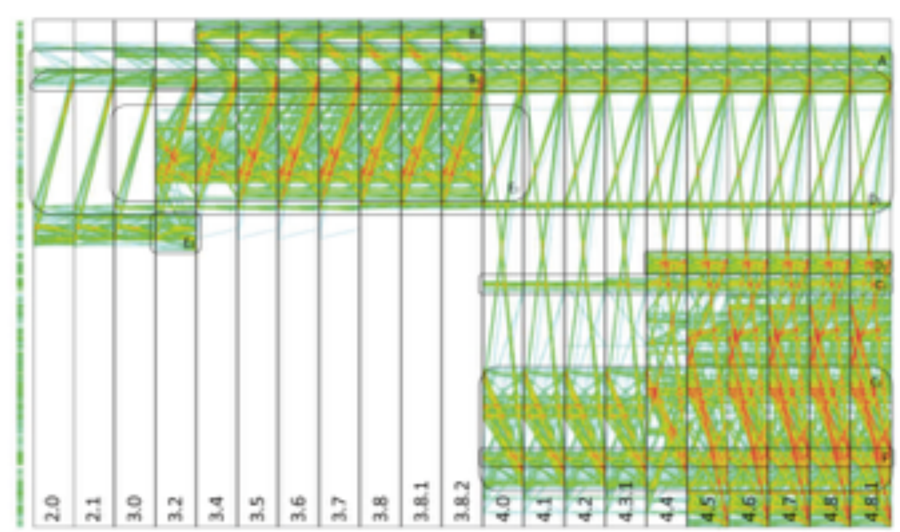


4. eyes beat memory

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 static way



5. focus & context

Show selected regions in greater detail (focus)

Preserve global view at reduced detail (context)

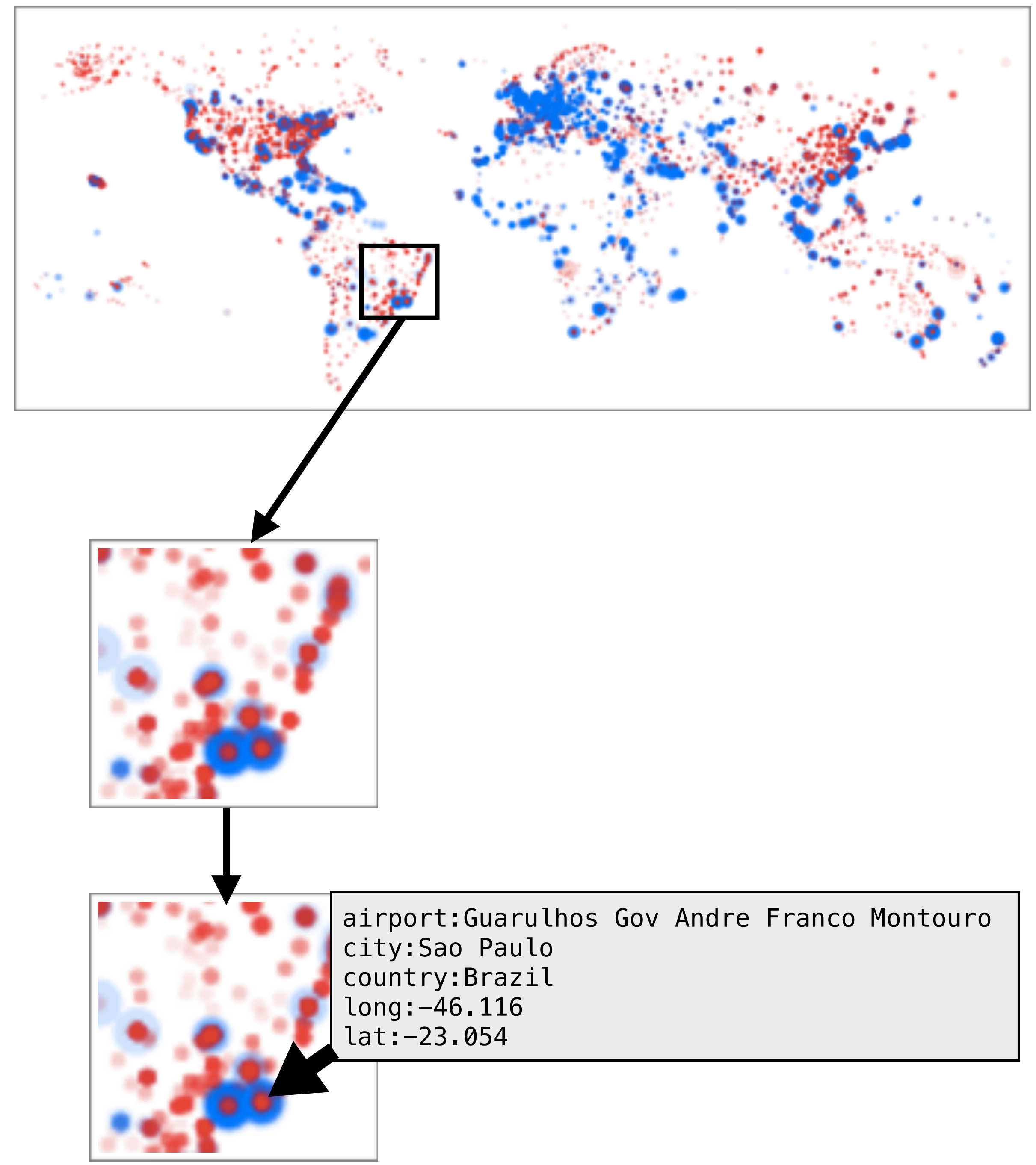
No occlusion (all information is visible simultaneously)

(Keahey, 2003)

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

task taxonomy Ben Schneiderman:

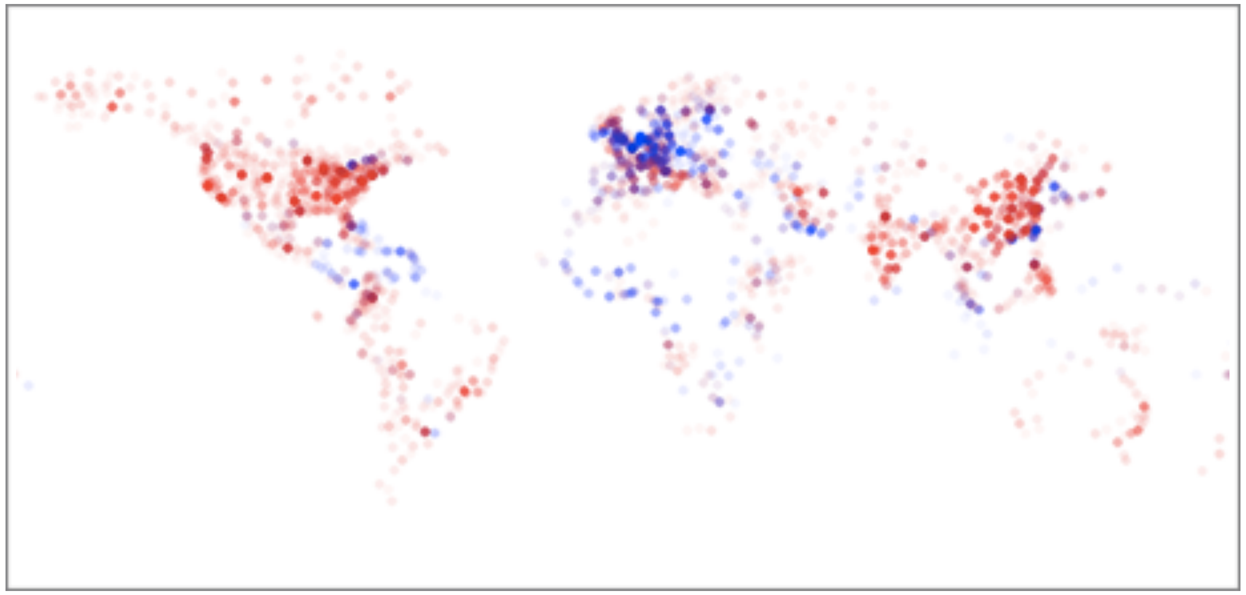
- *Overview*: see overall patterns in data
- *Zoom*: see a subset of data
- *Filter*: see a subset based on values
- *Detail on demand*: see values of items
- *Relate*: compare values
- *History*: keep track of actions
- *Extract*: mark and captur



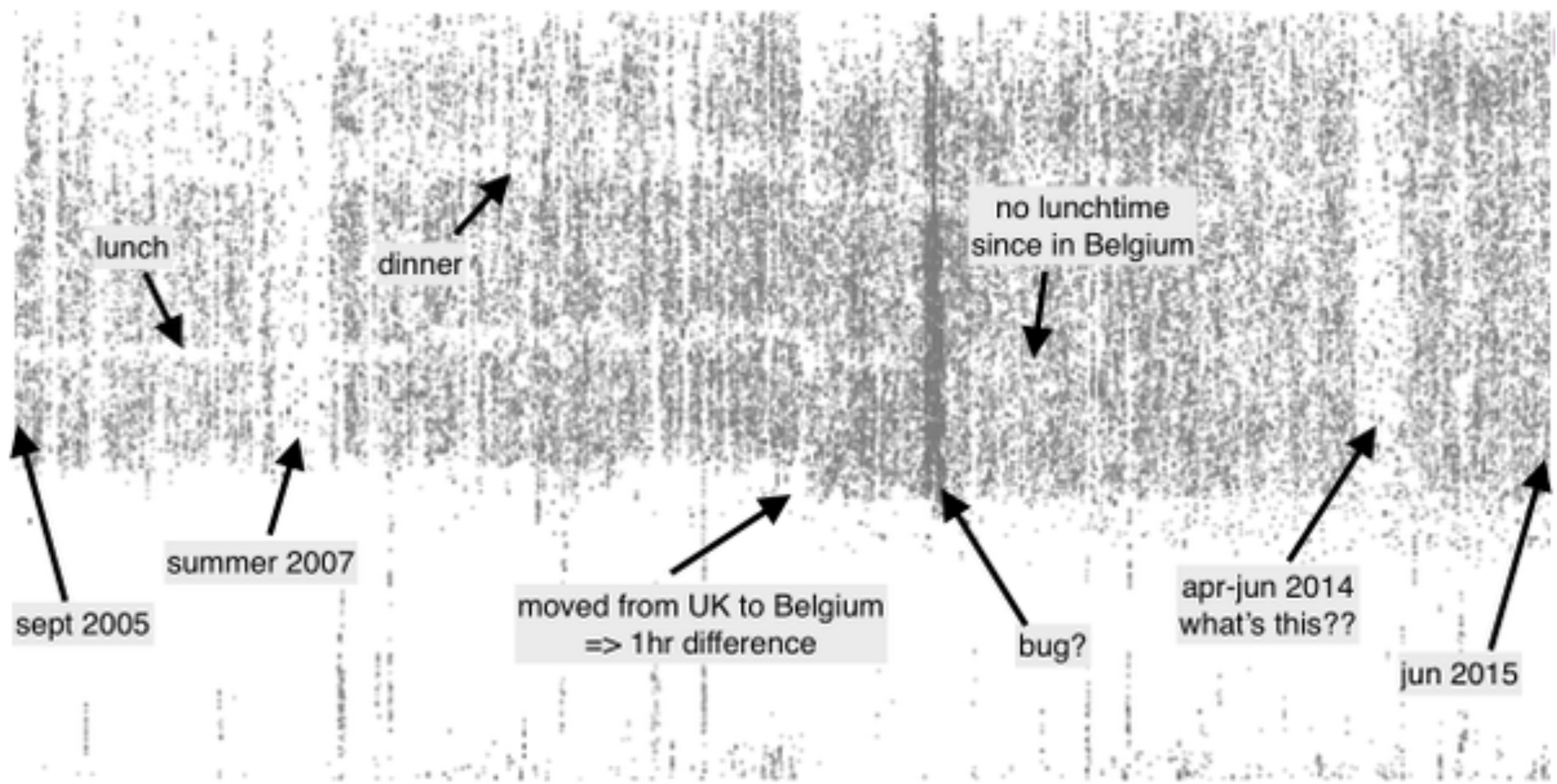
7. “underengineer”, if possible

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



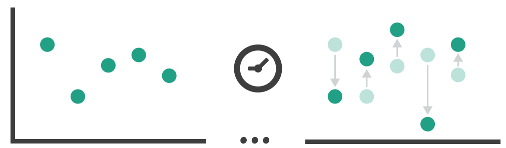
“send” date & time for all my outgoing email since 2005

How to handle complexity?

- manipulate
- facet
- reduce
- derive (see earlier)

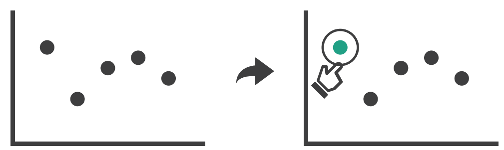
Manipulate

Change View Over Time



e.g. animated transitions

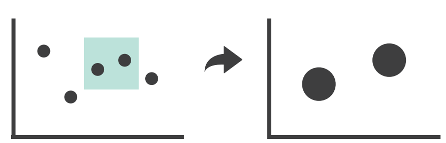
Select



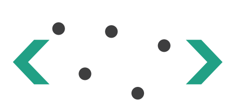
Navigate

Item Reduction

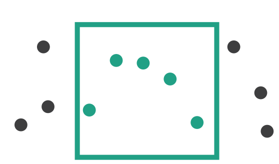
Zoom
Geometric or *Semantic*



Pan/Translate

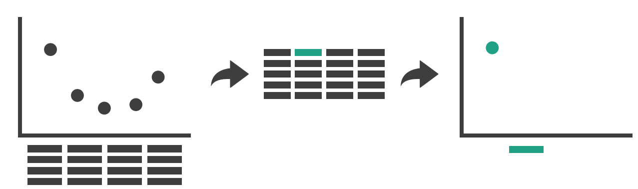


Constrained

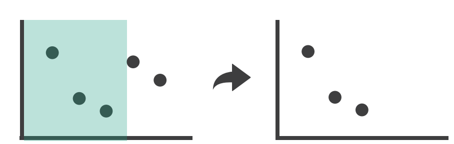


Attribute Reduction

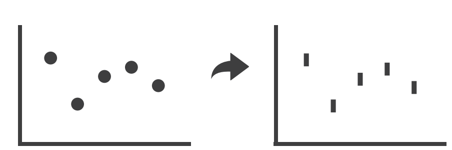
Slice

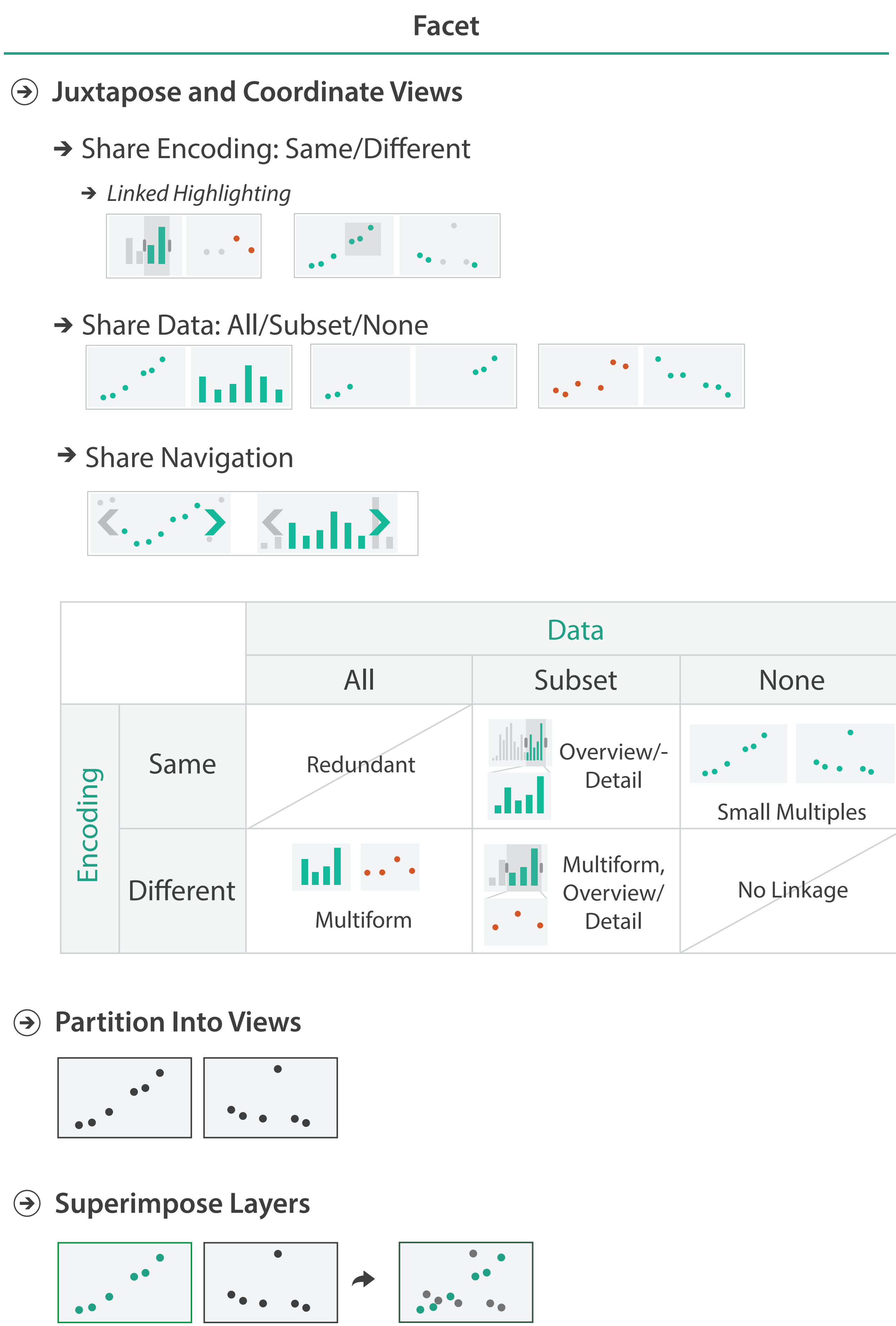


Cut

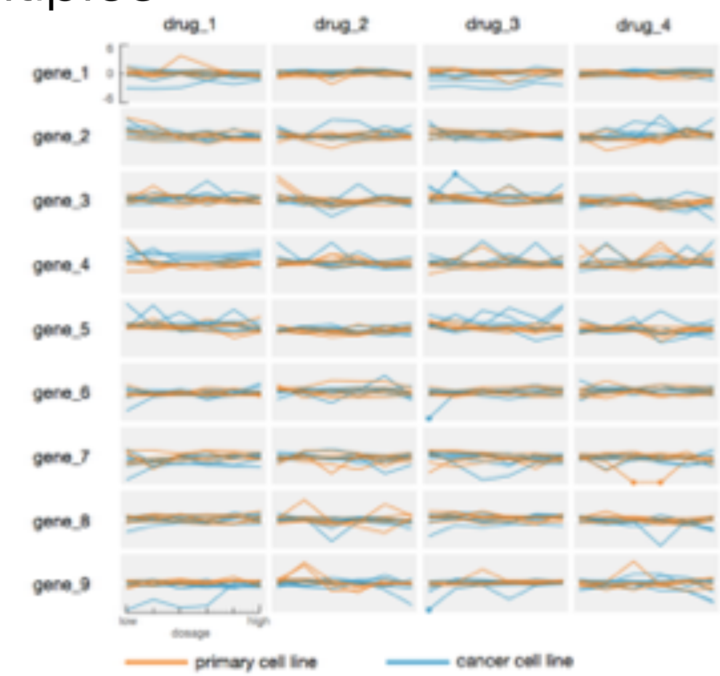


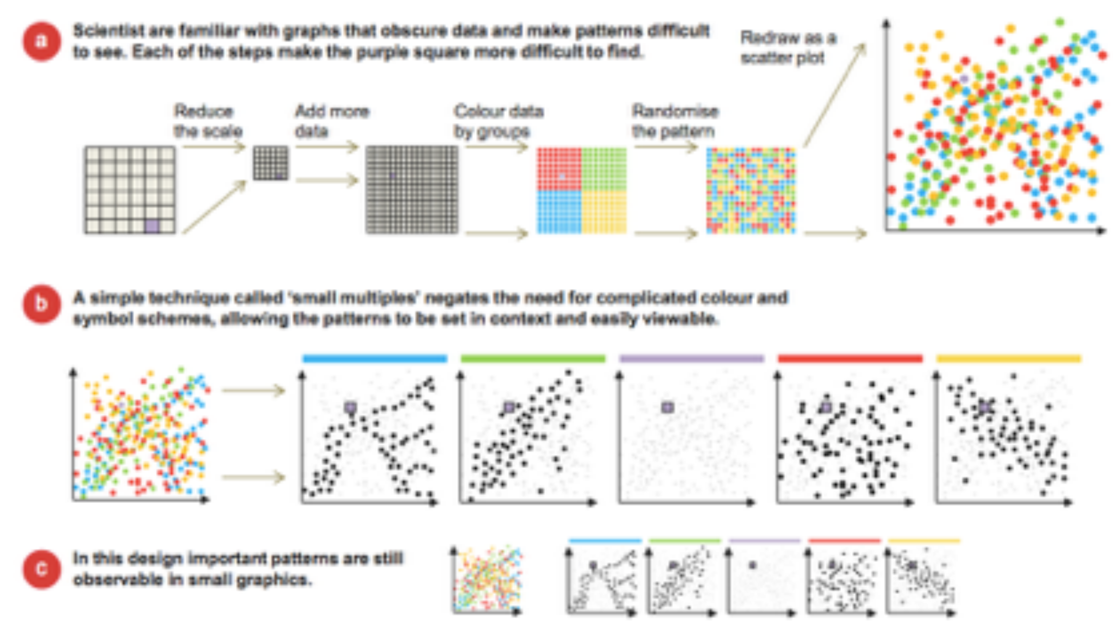
Project





e.g. small multiples

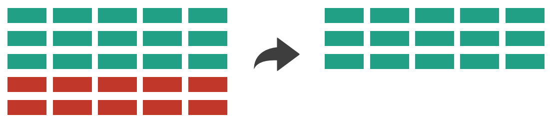




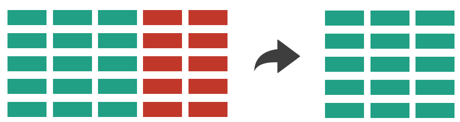
Reducing Items and Attributes

Filter

→ Items



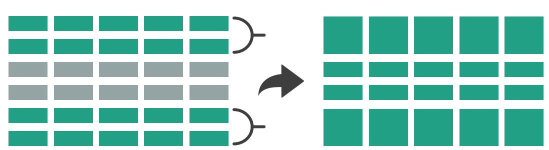
→ Attributes



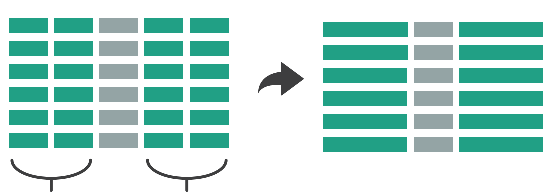
e.g. box plot

Aggregate

→ Items



→ Attributes

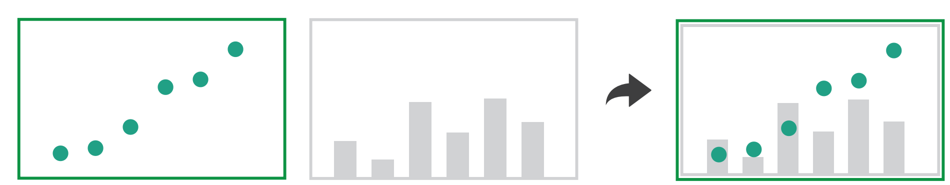


e.g. dimensionality reduction

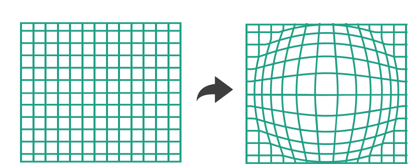
Embed

→ Elide Data

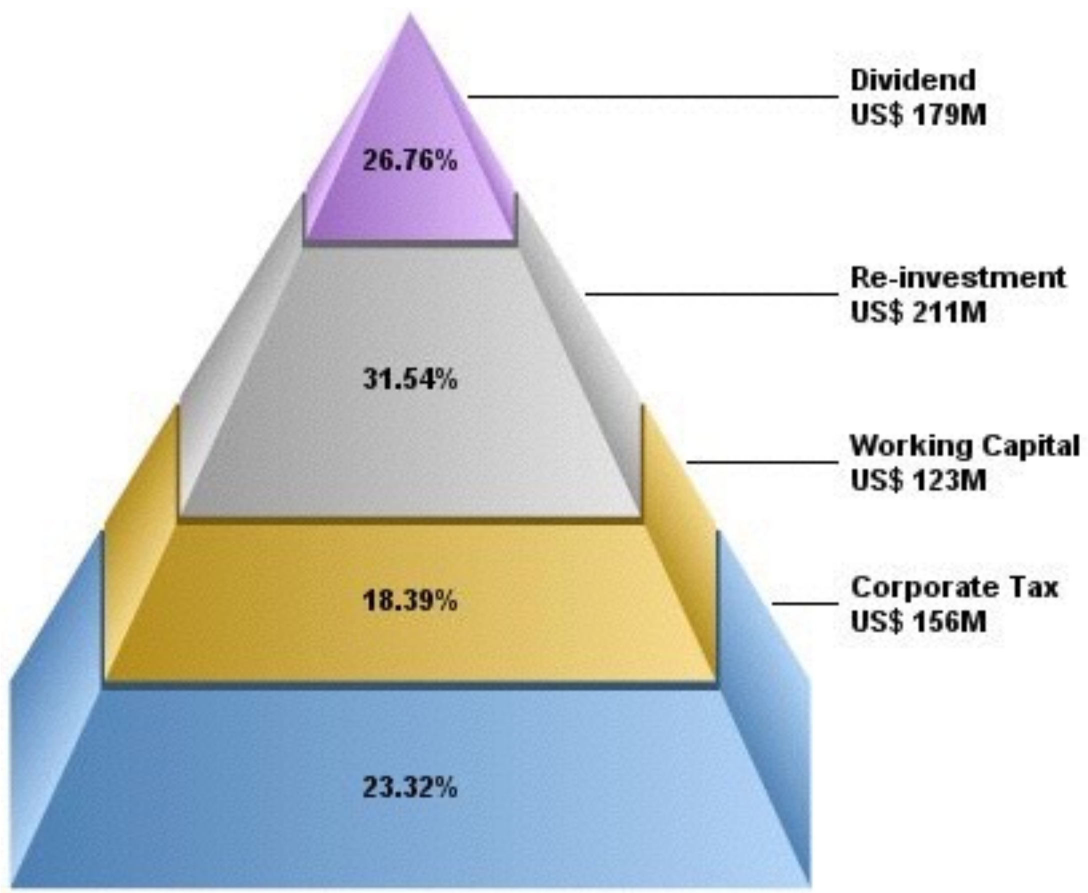
→ Superimpose Layer

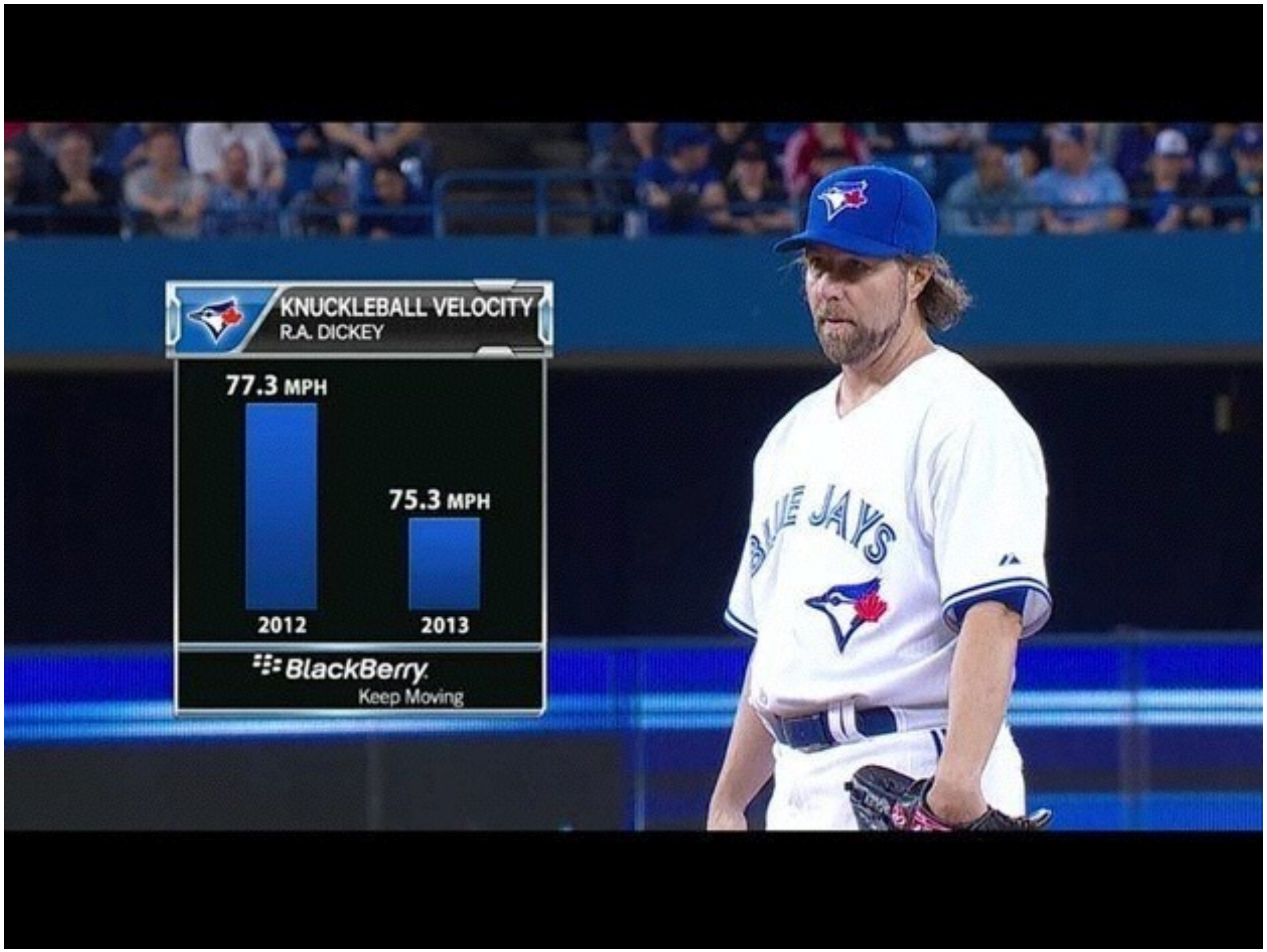


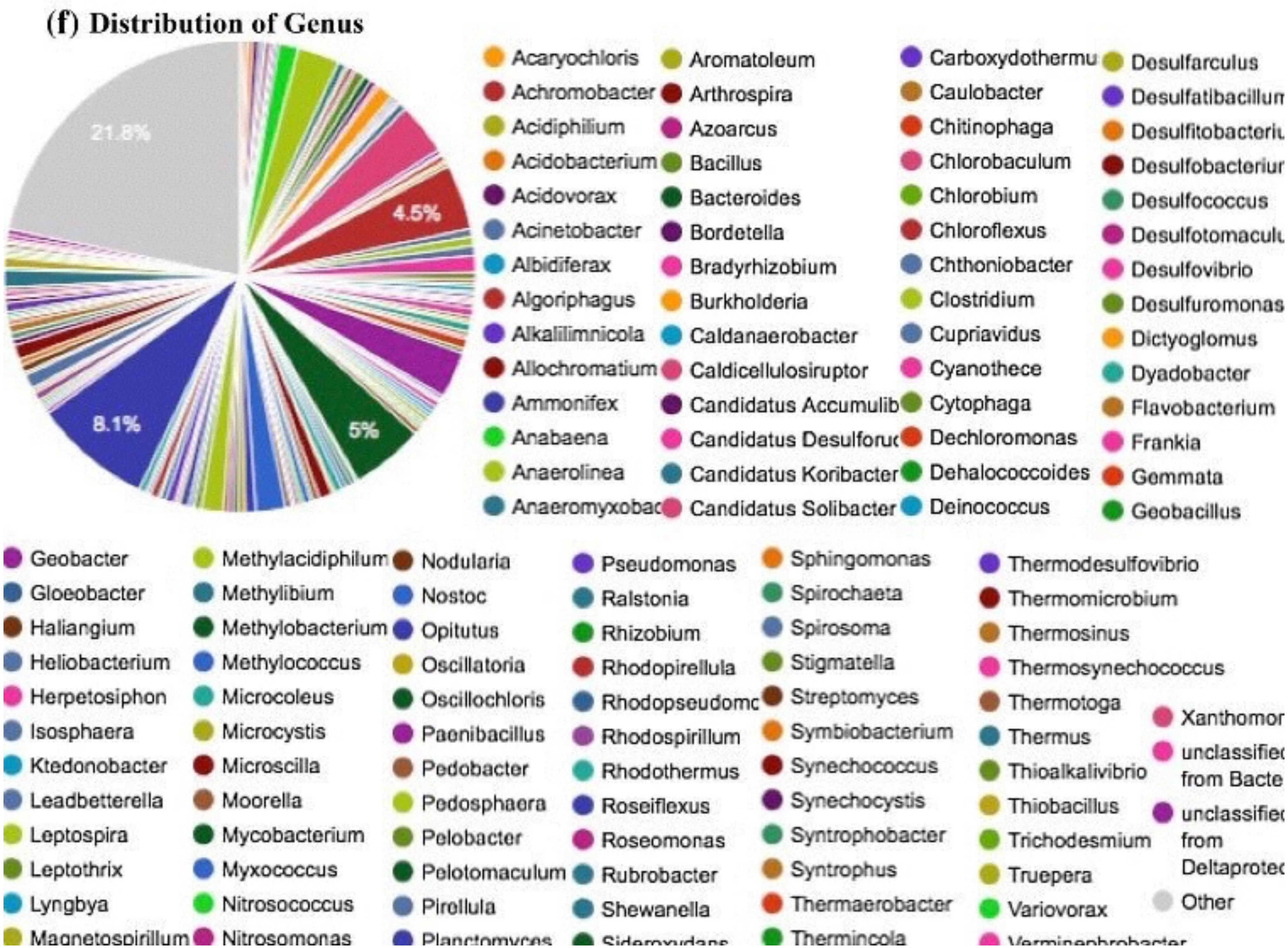
→ Distort Geometry

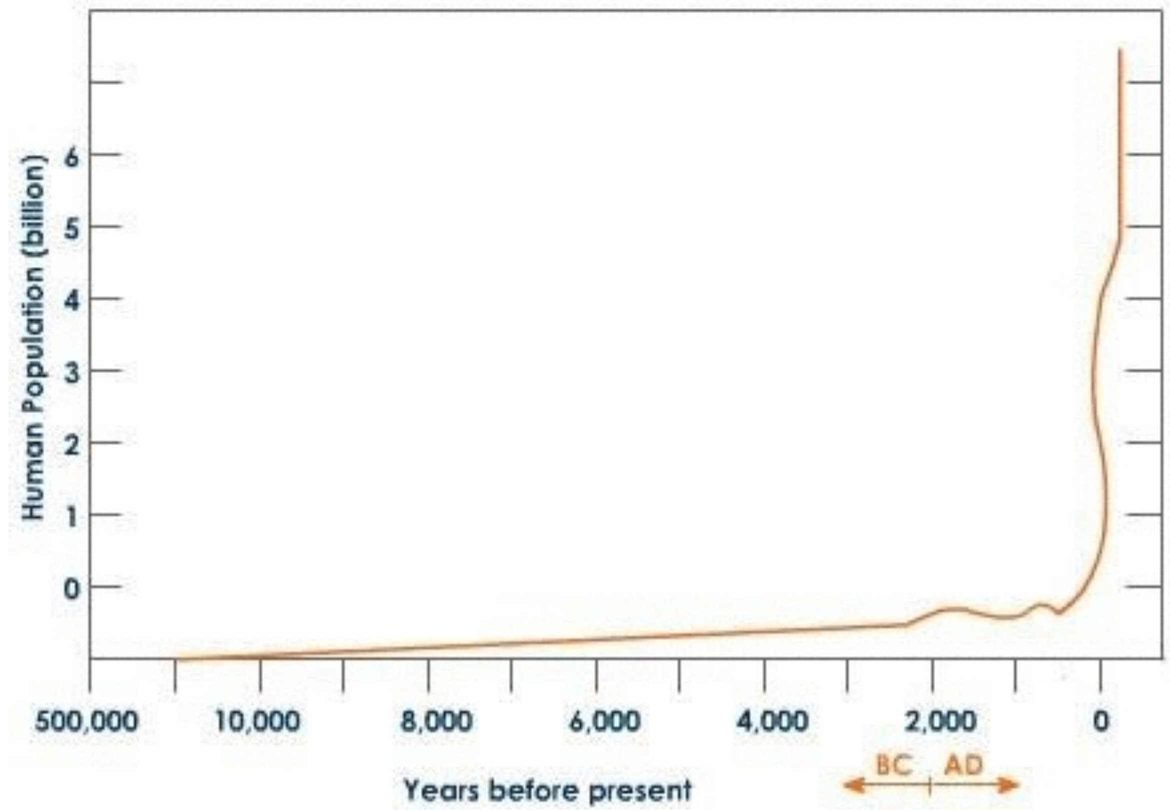


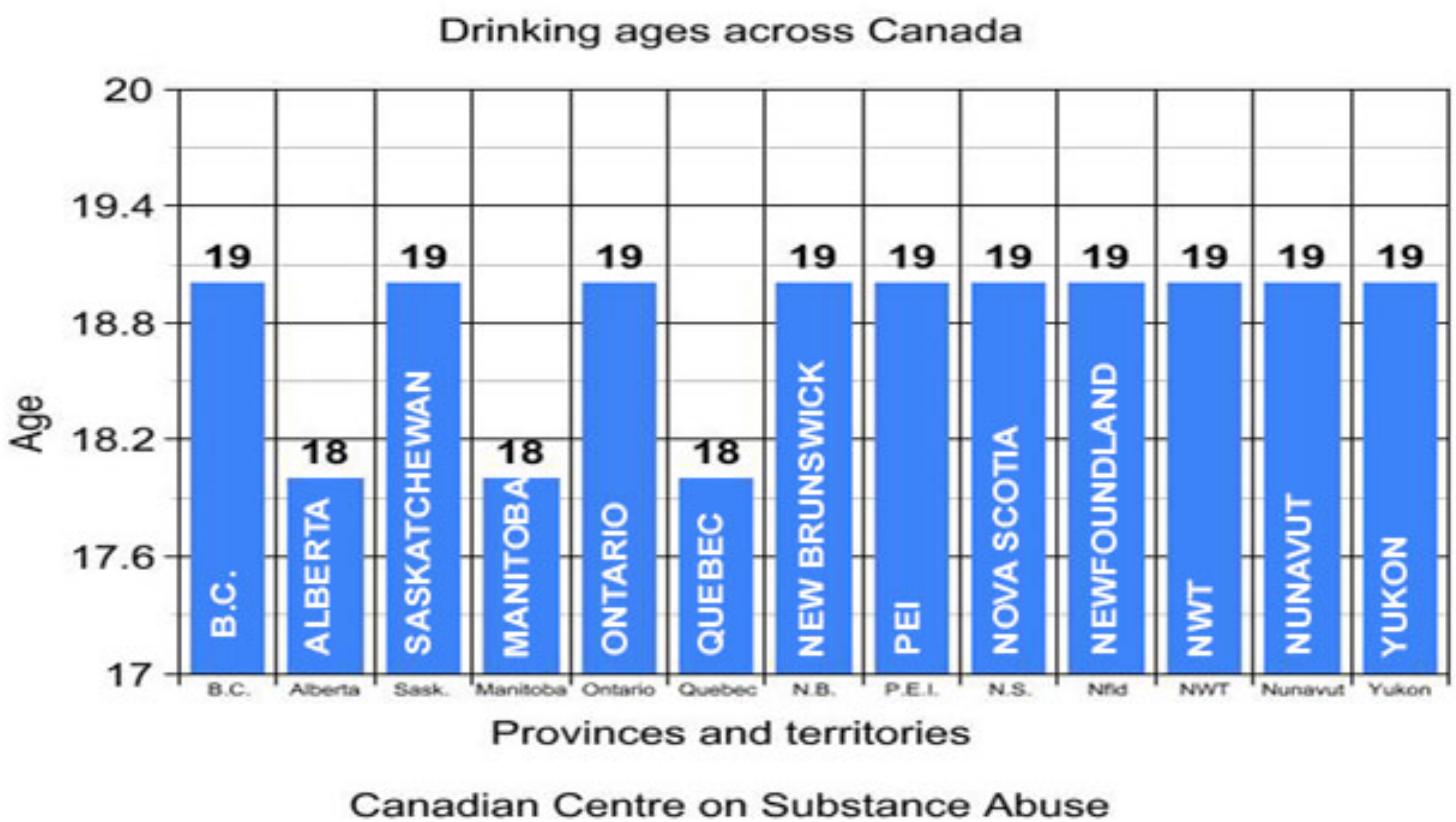
[D] #badgraphoftheday

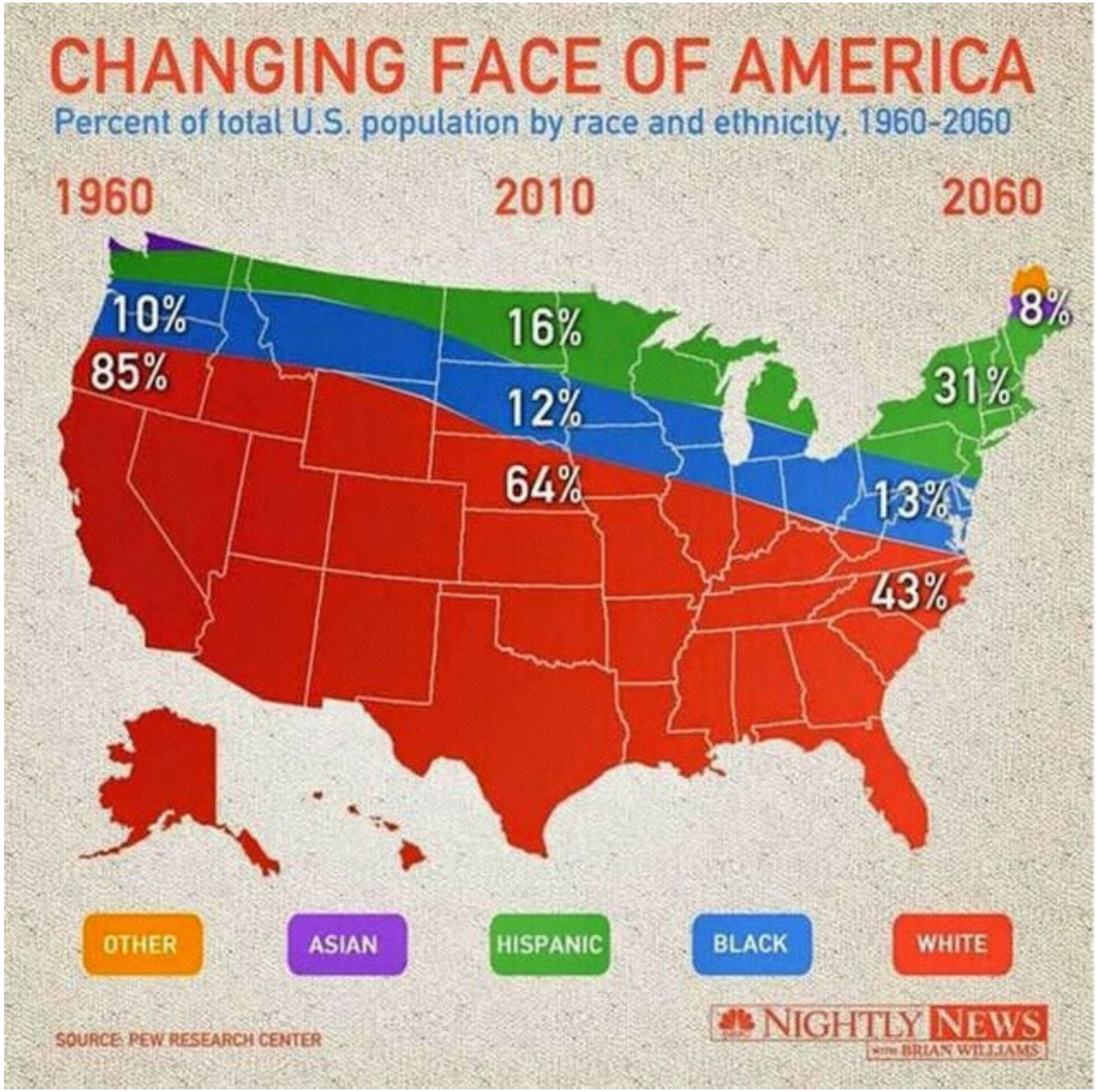

Dividend
US$ 179M
Re-investment
US$ 211M
Working Capital
US$ 123M
Corporate Tax
US$ 156M
26.76%
31.54%
18.39%
23.32%

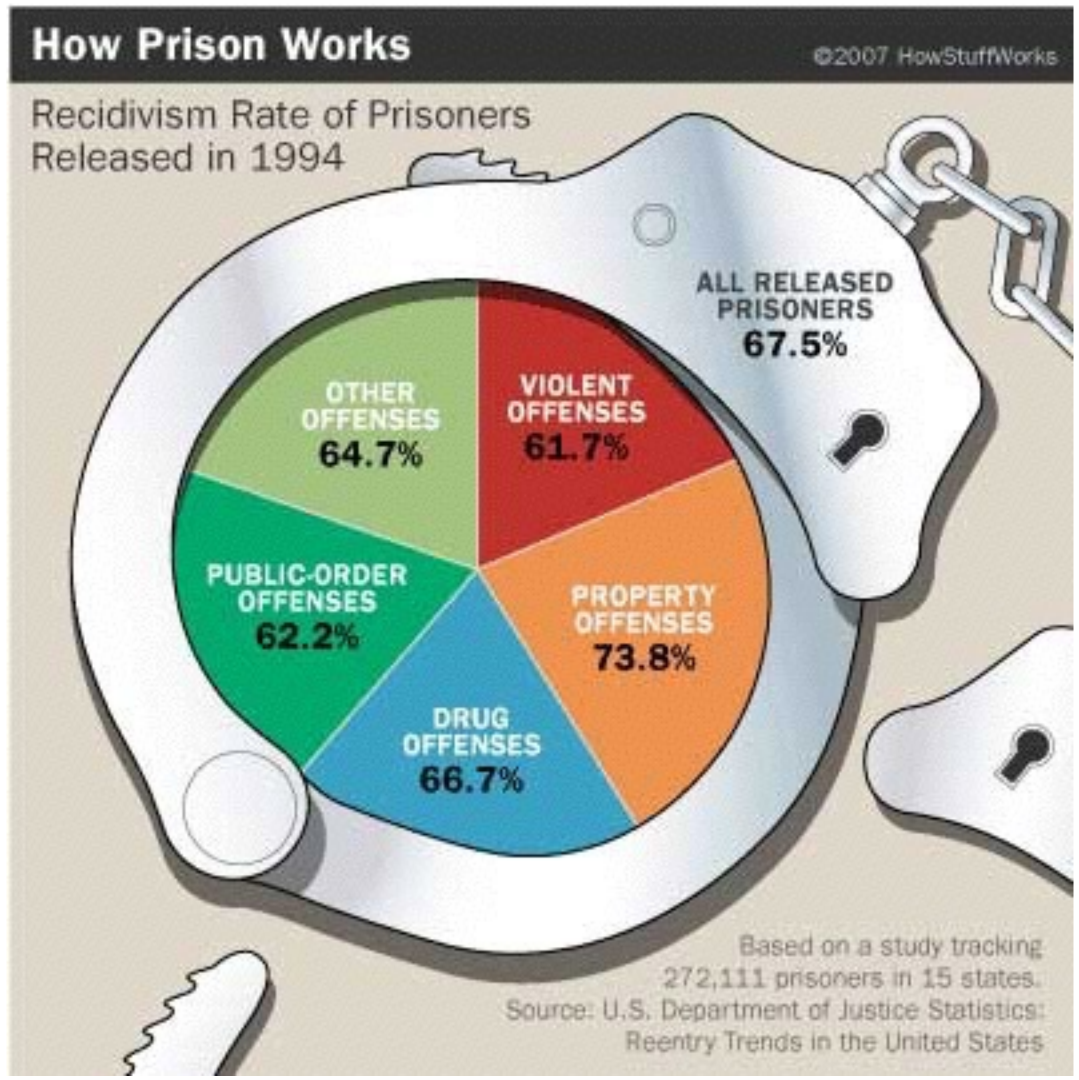

KNUCKLEBALL VELOCITY
R.A. DICKEY
77.3 MPH
75.3 MPH
2012
2013
BlackBerry
Keep Moving
BLUE JAYS

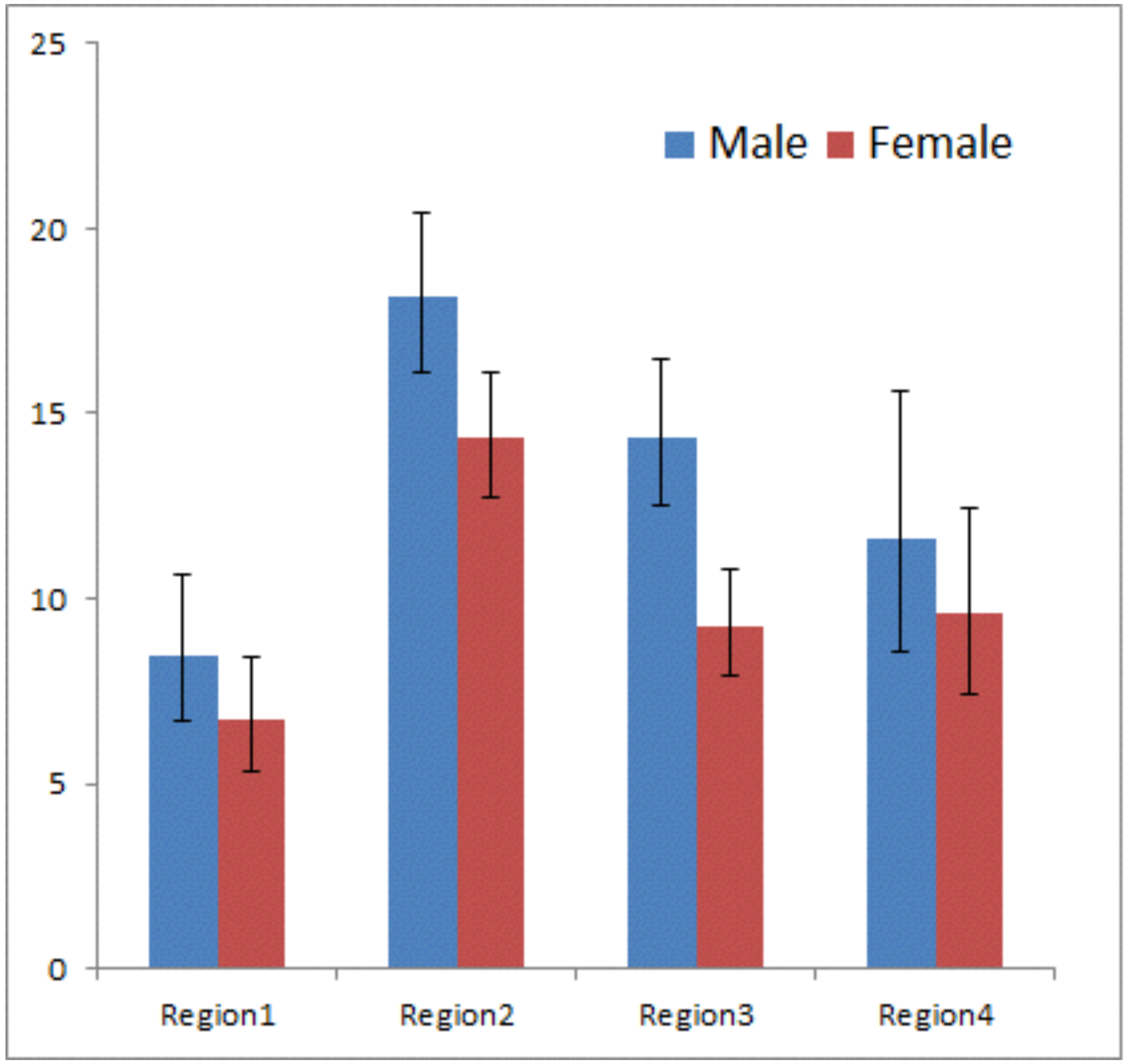

(f) Distribution of Genus
21.8%
4.5%
8.1%
5%
Acaryochloris
Achromobacter
Acidiphilium
Acidobacterium
Acidovorax
Acinetobacter
Albidiferax
Algoriphagus
Alkalilimnicola
Allochromatium
Ammonifex
Anabaena
Anaerolinea
Anaeromyxobac
Aromatoleum
Arthrospira
Azoarcus
Bacillus
Bacteroides
Bordetella
Bradyrhizobium
Burkholderia
Caldanaerobacter
Caldicellulosiruptor
Candidatus Accumulib
Candidatus Desulforu
Candidatus Koribacter
Candidatus Solibacter
Carboxydothermu
Caulobacter
Chitinophaga
Chlorobaculum
Chlorobium
Chloroflexus
Chthoniobacter
Clostridium
Cupriavidus
Cyanothece
Cytophaga
Dechloromonas
Dehalococcoides
Deinococcus
Desulfarculus
Desulfatibacillum
Desulfitobacteriu
Desulfobacteriur
Desulfococcus
Desulfotomacul
Desulfovibrio
Desulfuromonas
Dictyoglomus
Dyadobacter
Flavobacterium
Frankia
Gemmata
Geobacillus
Geobacter
Gloeobacter
Haliangium
Heliobacterium
Herpetosiphon
Isosphaera
Ktedonobacter
Leadbetterella
Leptospira
Leptothrix
Lyngbya
Magnetospirillum
Methylacidiphilum
Methylibium
Methylobacterium
Methylococcus
Microcoleus
Microcystis
Microscilla
Moorella
Mycobacterium
Myxococcus
Nitrosococcus
Nitrosomonas
Nodularia
Nostoc
Opitutus
Oscillatoria
Oscillochloris
Paenibacillus
Pedobacter
Pedosphaera
Pelobacter
Pelotomaculum
Pirellula
Planctomyces
Pseudomonas
Ralstonia
Rhizobium
Rhodopirellula
Rhodopseudomo
Rhodospirillum
Rhodothermus
Roseiflexus
Roseomonas
Rubrobacter
Shewanella
Sideroxydans
Sphingomonas
Spirochaeta
Spirosoma
Stigmatella
Streptomyces
Symbiobacterium
Synechococcus
Synechocystis
Syntrophobacter
Syntrophus
Thermaerobacter
Thermincola
Thermodesulfovibrio
Thermomicrobium
Thermosinus
Thermosynechococcus
Thermotoga
Thermus
Thioalkalivibrio
Thiobacillus
Trichodesmium
Truepera
Variovorax
Verminephrobacter
Xanthomor
unclassified from Bacte
unclassified from Deltaprote
Other

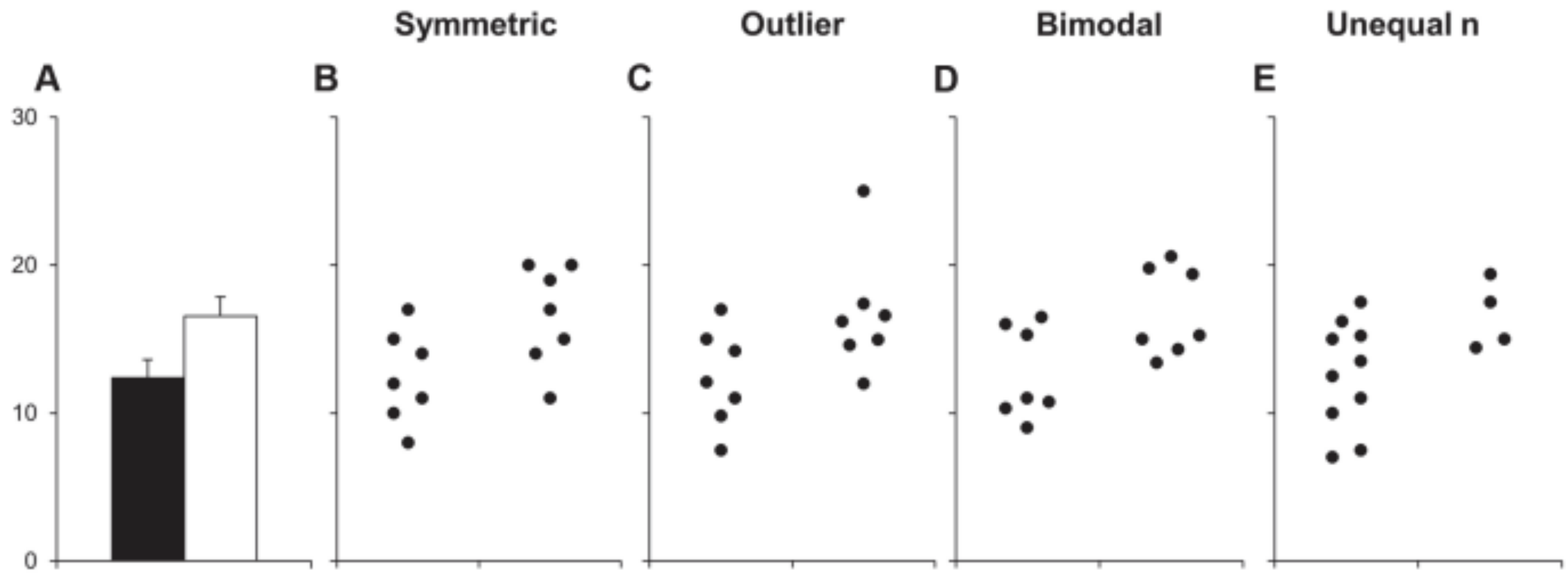

Human Population (billion)
0
1
2
3
4
5
6
500,000
10,000
8,000
6,000
4,000
2,000
0
Years before present
BC
AD


Drinking ages across Canada
Age
20
19.4
18.8
18.2
17.6
17
19
18
19
18
19
18
19
19
19
19
19
19
19
B.C.
ALBERTA
SASKATCHEWAN
MANITOBA
ONTARIO
QUEBEC
NEW BRUNSWICK
PEI
NOVA SCOTIA
NEWFOUNDLAND
NWT
NUNAVUT
YUKON
B.C.
Alberta
Sask.
Manitoba
Ontario
Quebec
N.B.
P.E.I.
N.S.
Nfld
NWT
Nunavut
Yukon
Provinces and territories
Canadian Centre on Substance Abuse

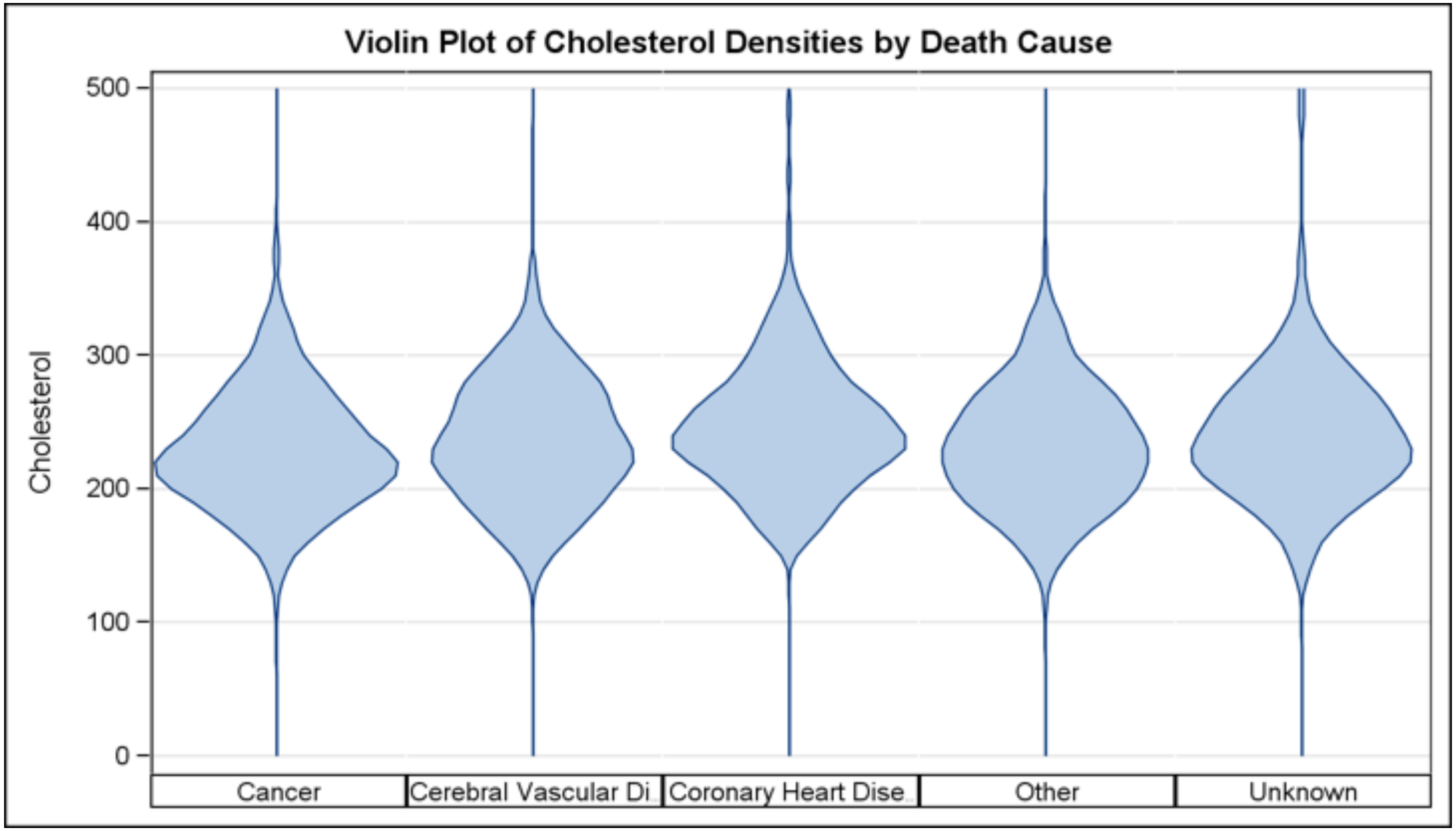

CHANGING FACE OF AMERICA
Percent of total U.S. population by race and ethnicity, 1960-2060
1960
2010
2060
10%
85%
16%
12%
64%
8%
31%
13%
43%
OTHER
ASIAN
HISPANIC
BLACK
WHITE
SOURCE: PEW RESEARCH CENTER
NIGHTLY NEWS
WITH BRIAN WILLIAMS

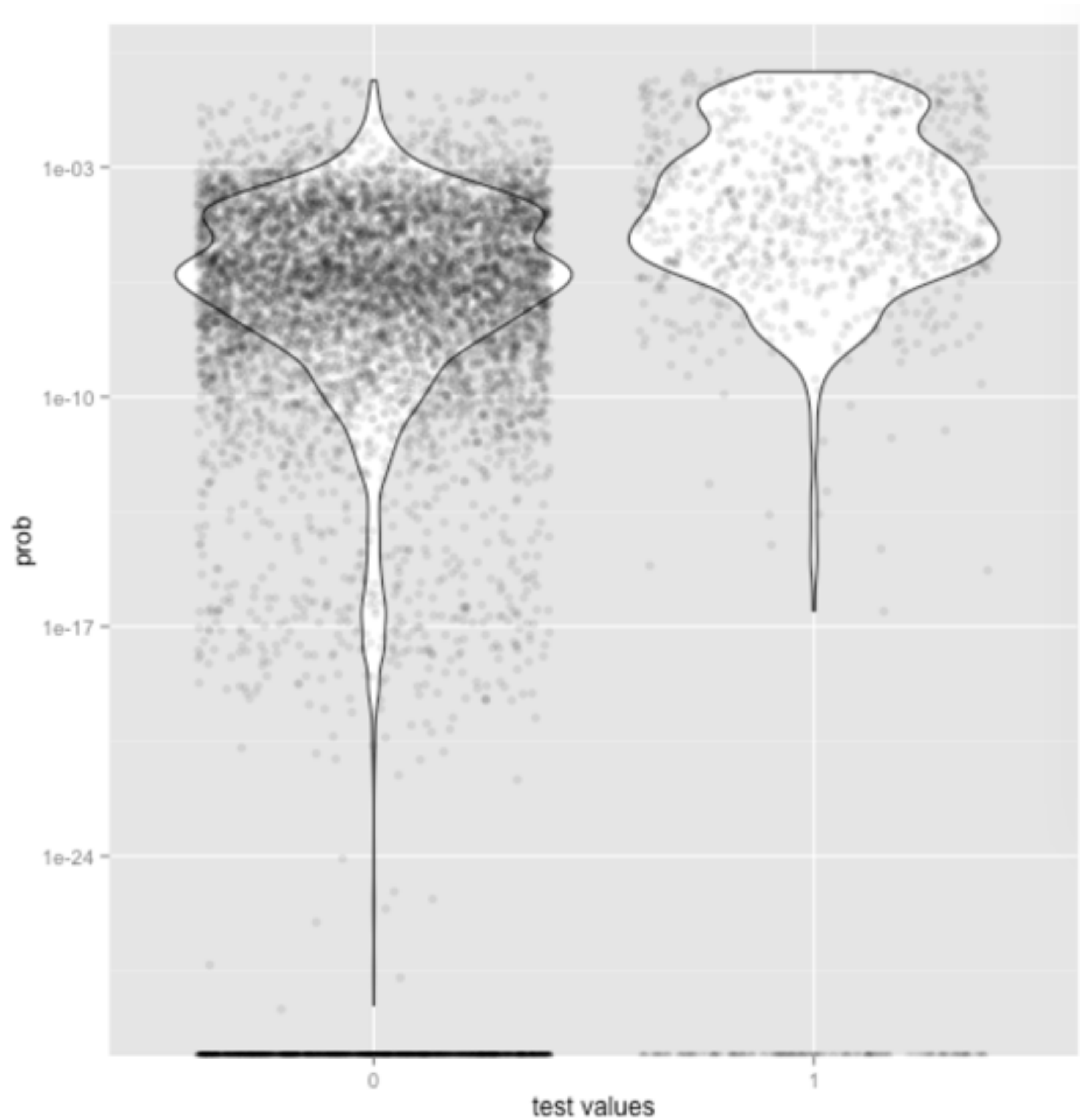

How Prison Works
©2007 HowStuffWorks
Recidivism Rate of Prisoners
Released in 1994
ALL RELEASED PRISONERS
67.5%
OTHER OFFENSES
64.7%
VIOLENT OFFENSES
61.7%
PUBLIC-ORDER OFFENSES
62.2%
PROPERTY OFFENSES
73.8%
DRUG OFFENSES
66.7%
Based on a study tracking
272,111 prisoners in 15 states.
Source: U.S. Department of Justice Statistics:
Reentry Trends in the United States


Male
Female
25
20
15
10
5
0
Region1
Region2
Region3
Region4



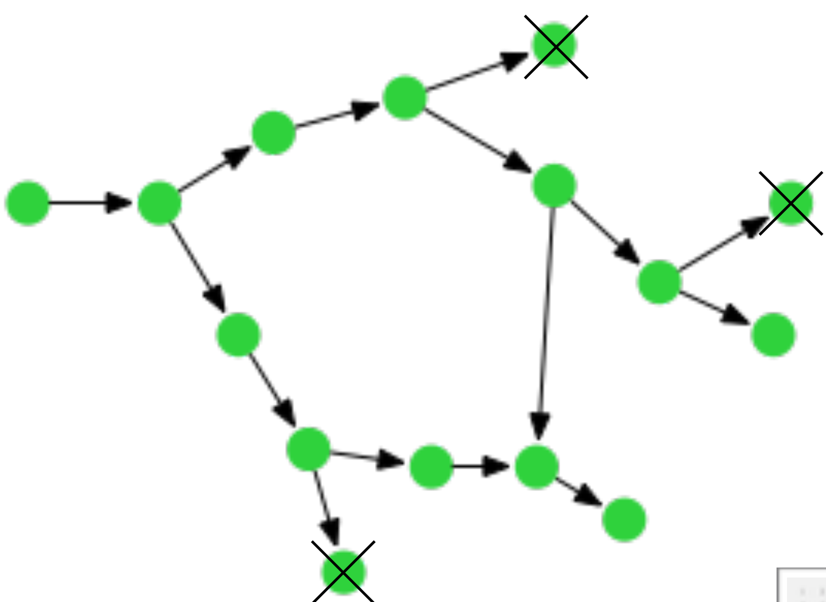
Weissgerber et al, 2015

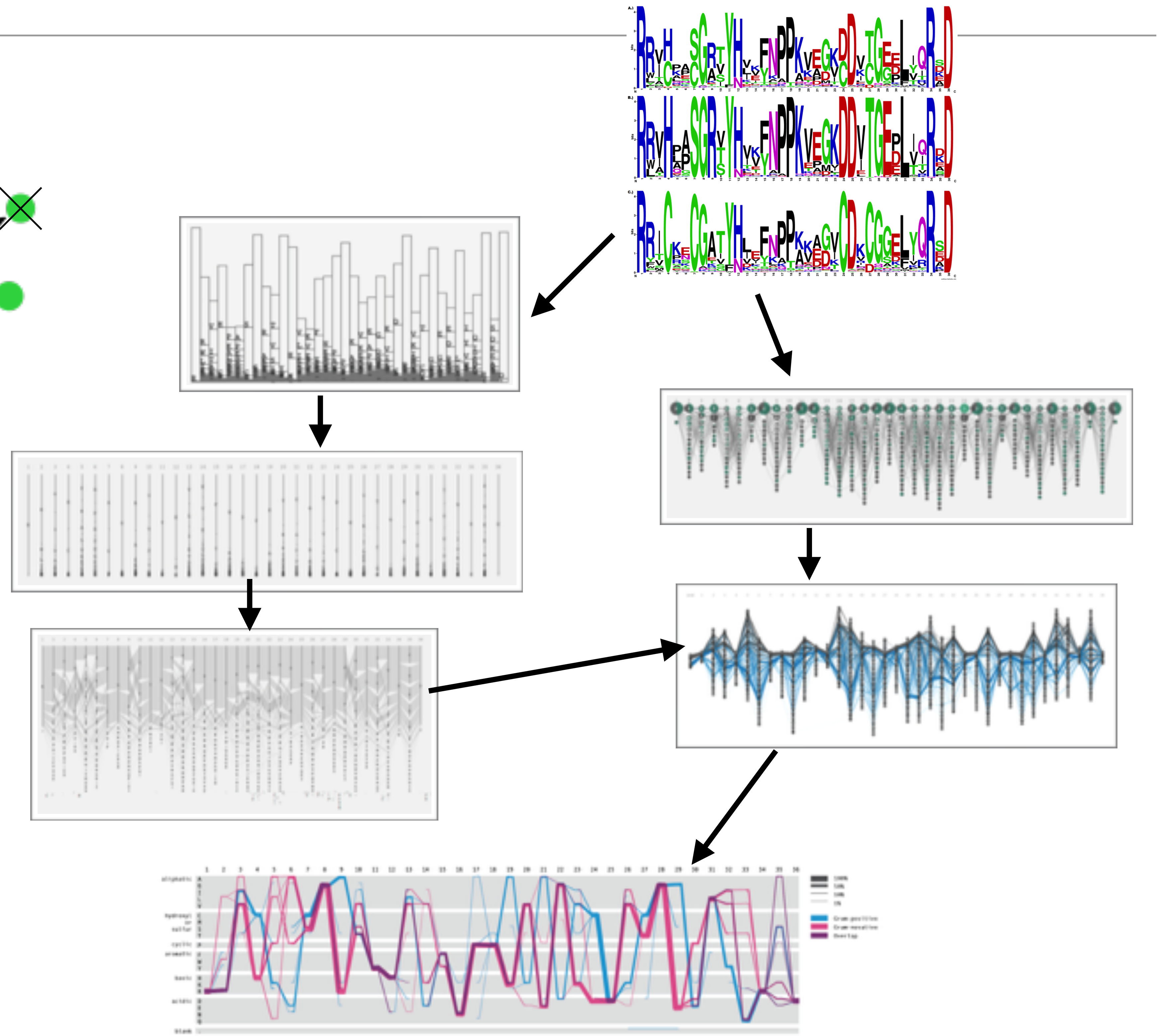

Violin Plot of Cholesterol Densities by Death Cause
Cholesterol
500
400
300
200
100
0
Cancer
Cerebral Vascular Di
Coronary Heart Dise
Other
Unknown


1e-03
1e-10
1e-17
1e-24
prob
0
1
test values

[E] Design process

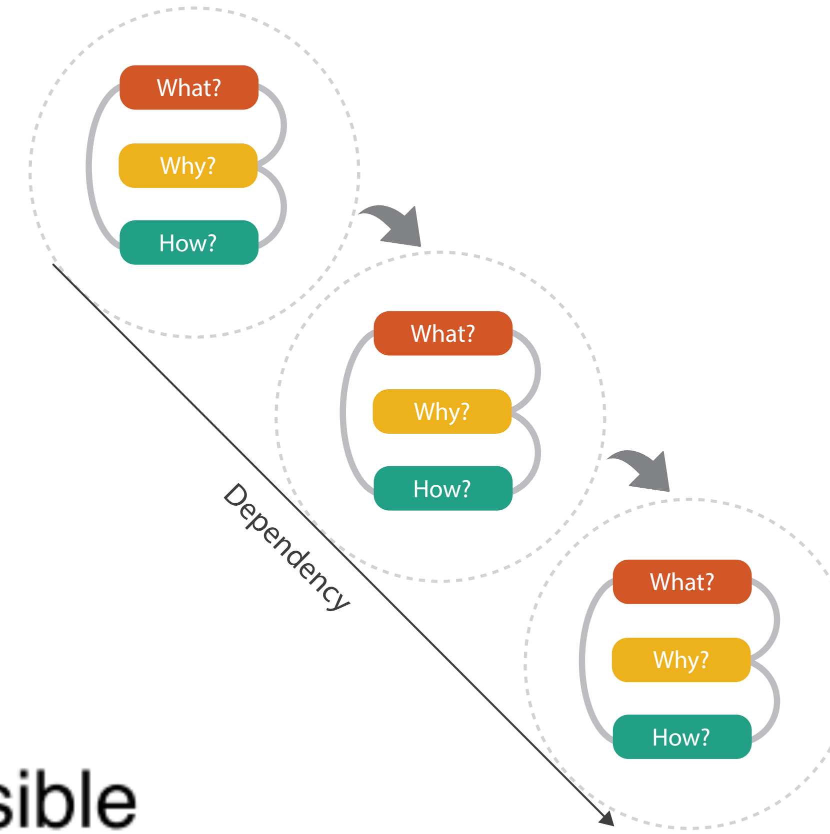
“Design = search problem” (M Bostock)

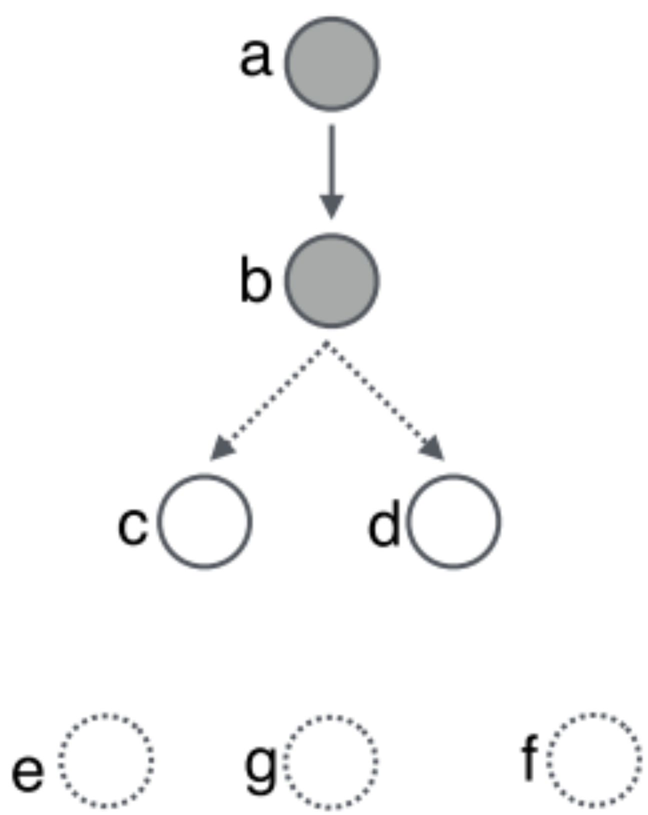




Sequence Diversity Diagram

The adjacent possible





Realised idea

Adjacent possible

Unattainable / unknown idea

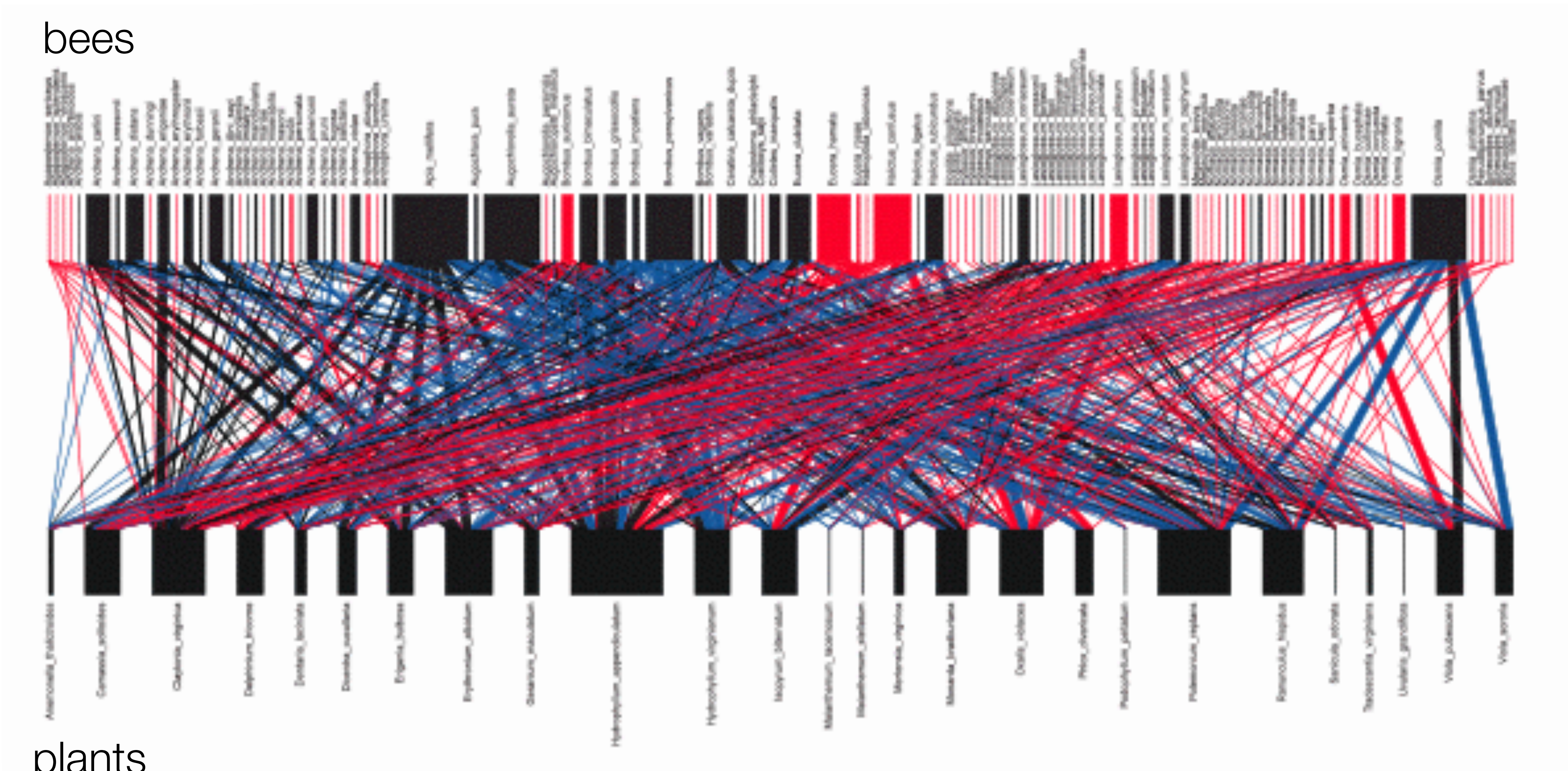
An example design process

Taken from work and blog post by Moritz Stefaner: “Where the Wild Bees Are” (http://well-formed-data.net/archives/972/where-the-wild-bees-are)

- Created by well-know data visualisation expert (“truth and beauty operator”)
- Good example of thinking process
- (-) Infographic => more emphasis on metaphor; none on interaction

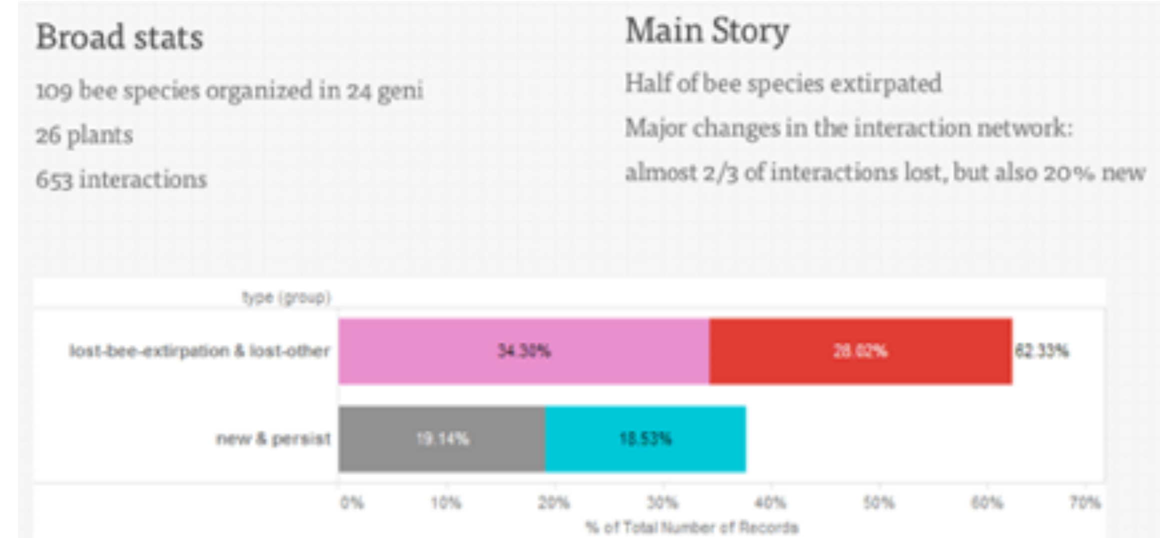
[1] Understand the research; doubt the data

Question: how do bee/plant-relationships change over time? Data: which bees flew to which plants in late 19th century, and which today?



(rotated) picture in original Science paper

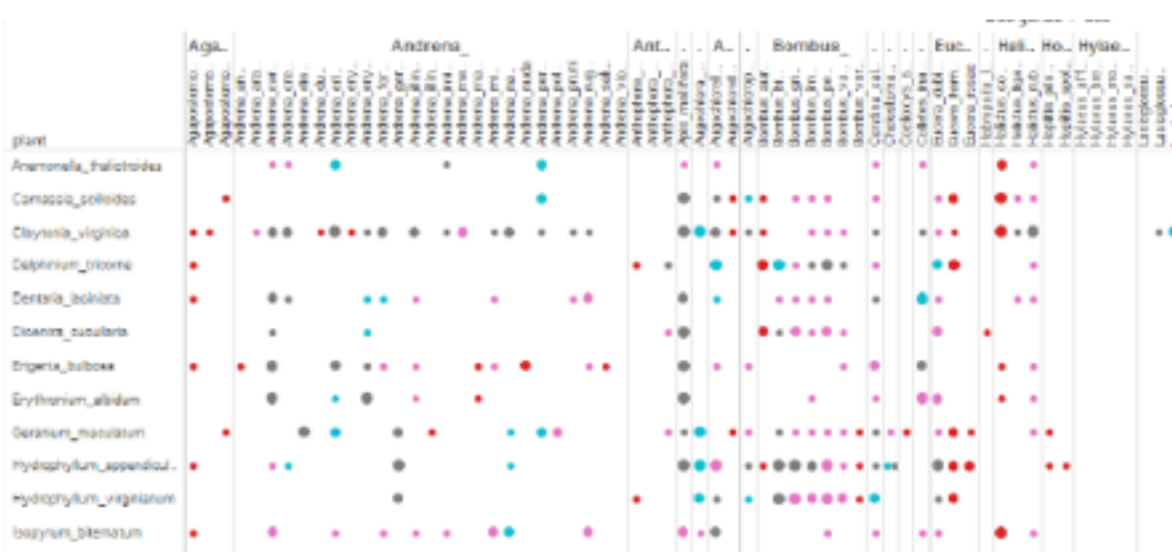
[2] Putting stakes in the ground:



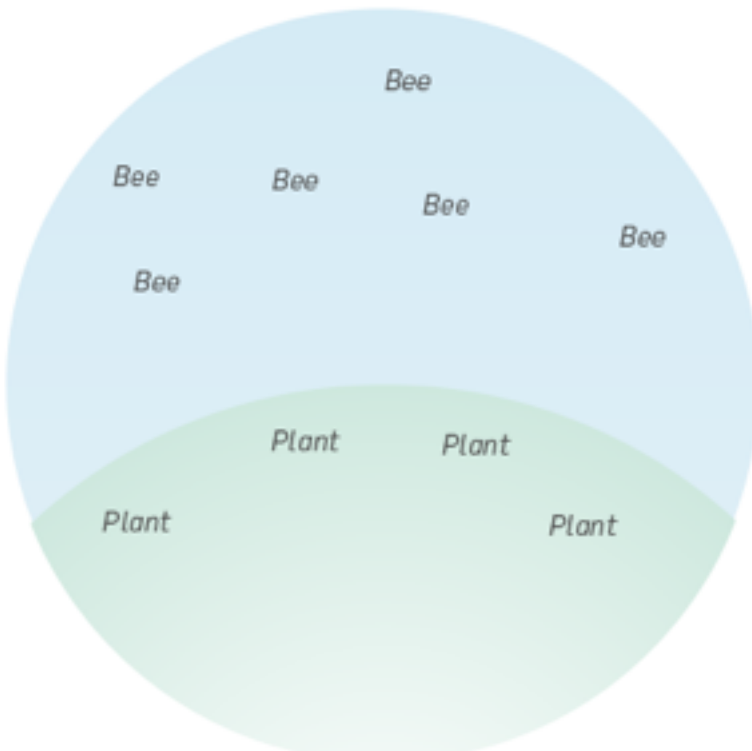
- What are the key statistics?
    - 109 bee species organised in 24 geni
    - 26 plants
    - 653 interactions between them
    - >50% of bee species from 19th century not found today
    - major changes in interaction network
        - 2/3 of interactions lost
        - 20% new interactions
- What is the main message?

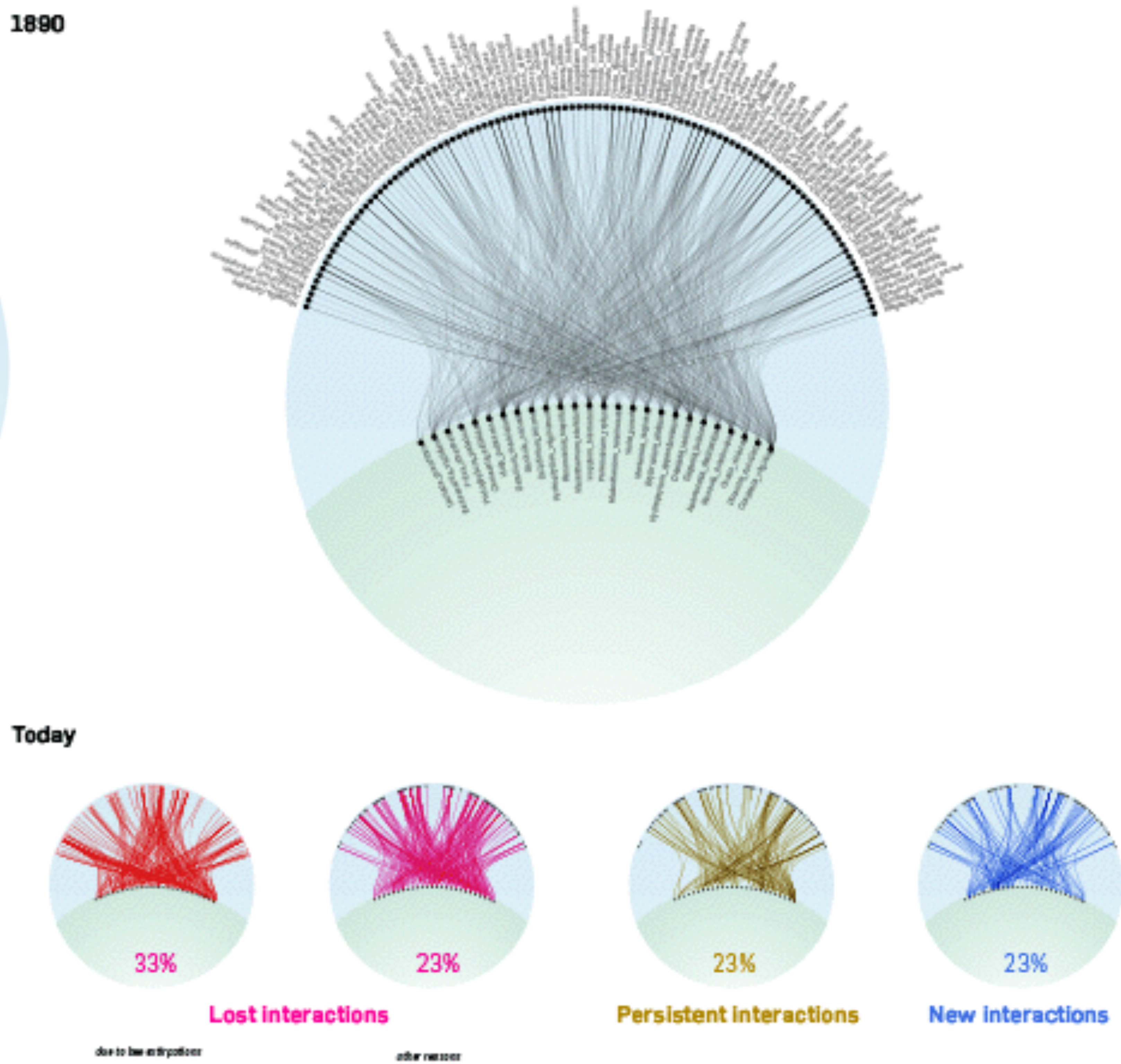
[3] Get a feel for your data

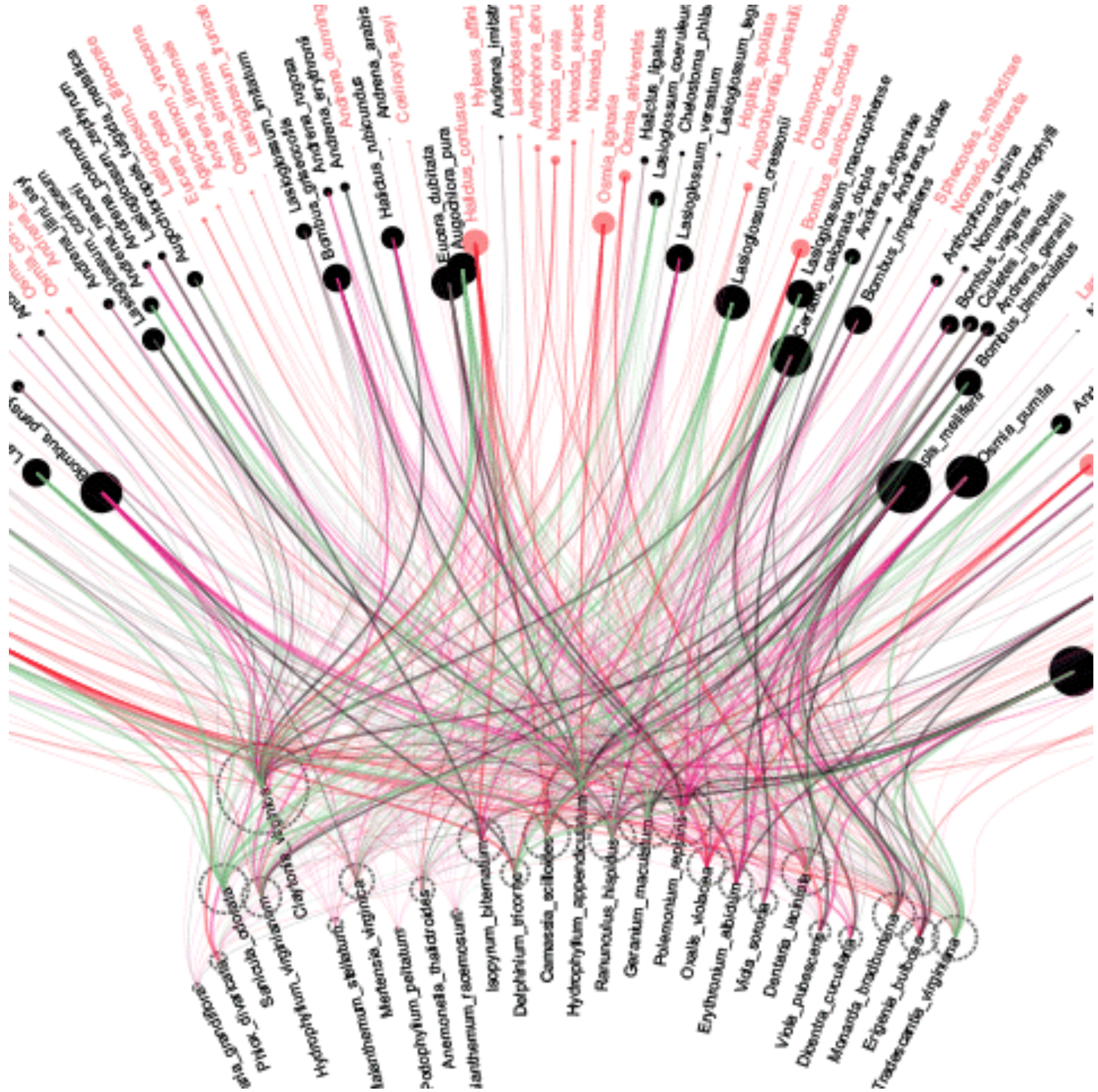
- Using Tableau, R, …
- What are distributions of dimensions? Interactions? …



[4] Finding a metaphor





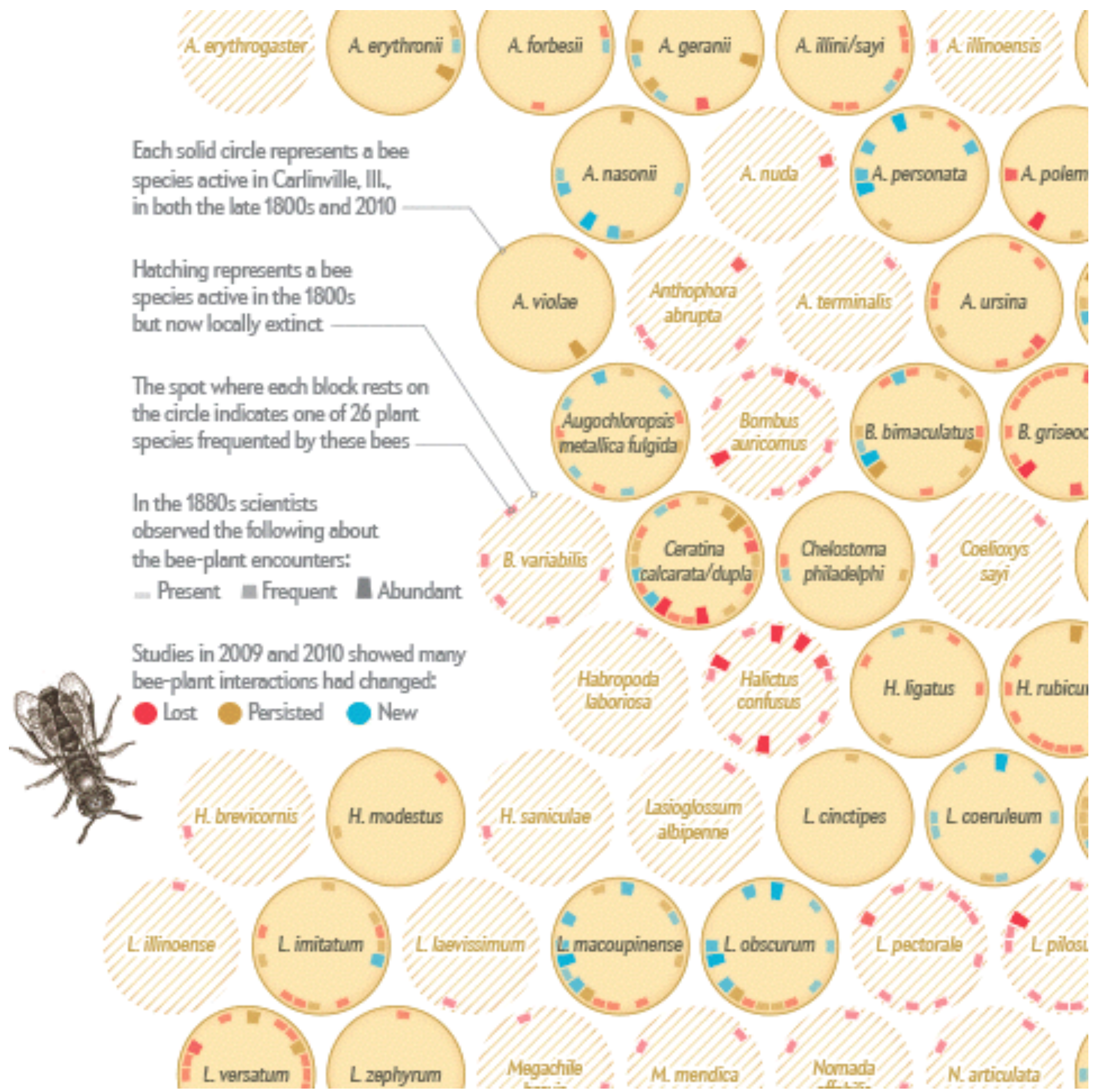


Andrena_arabis
Halictus_rubicundus
Eucera_dubitata
Augochlora_pura
Halictus_confusus
Osmia_lignaria
Osmia_atriventris
Halictus_ligatus
Lasioglossum_coeruleum
Lasioglossum_versatum
Lasioglossum_cressonii
Lasioglossum_macoupinense
Bombus_auricomus
Andrena_erigeniae
Andrena_violae
Bombus_impatiens
Anthophora_ursina
Bombus_vagans
Colletes_inaequalis
Andrena_geranii
Bombus_bimaculatus
Osmia_pumila
Andrena_erythronii
Bombus_griseocollis
Lasioglossum_imitatum
Claytonia_virginica
Phlox_divaricata
Sanicula_odorata
Smilacina_virginianum
Hydrophyllum_virginianum
Mertensia_virginica
Podophyllum_peltatum
Anemonella_thalictroides
Isopyrum_biternatum
Delphinium_tricorne
Camassia_scilloides
Hydrophyllum_appendiculatum
Ranunculus_hispidus
Geranium_maculatum
Polemonium_reptans
Oxalis_violacea
Erythronium_albidum
Viola_sororia
Dentaria_laciniata
Viola_pubescens
Dicentra_cucullaria
Monarda_bradburiana
Erigenia_bulbosa
Tradescantia_virginiana

Small multiples

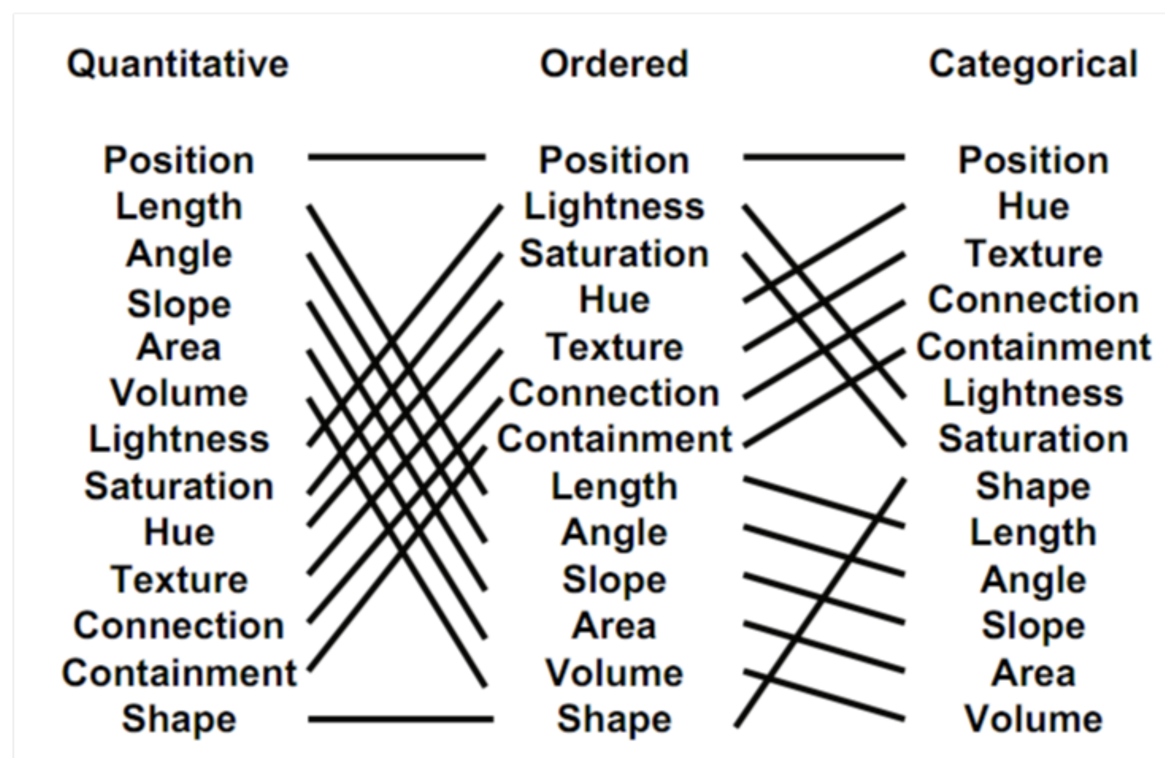
1 circle = 1 bee species
1 bar = 1 plant species

Finished infographic



A possible design workflow

1. Are there too many dimensions? => dimensionality reduction (PCA, SVD, …)
2. Are there too many datapoints? => aggregate/summarize or filter
3. Which is the most important (derived) dimension? Of what type? => look up in table McKinlay and pick top one (by definition: “position”)
4. Which is the second most important? => work your way down
5. Draw and store.
6. Start from scratch again.



Questions that a vis designer should ask when designing a vis idiom:

	Pop-out Effect	***Effectiveness Principle***	***Pattern Expressiveness***	***Interactive Exploration***
Considerations	· Pre-attentive · Gestalt psychology	· Stevens' power law · Cleveland and McGill · Mackinlay	· Visual learning	· Interaction techniques
Domain Knowledge	Independent	Independent	Dependent	Dependent
Goal	Speed	Accuracy	Comprehension	Exploration

1. **Pop-out effect**: What are the elements that pop-out when you first look at the visualization?

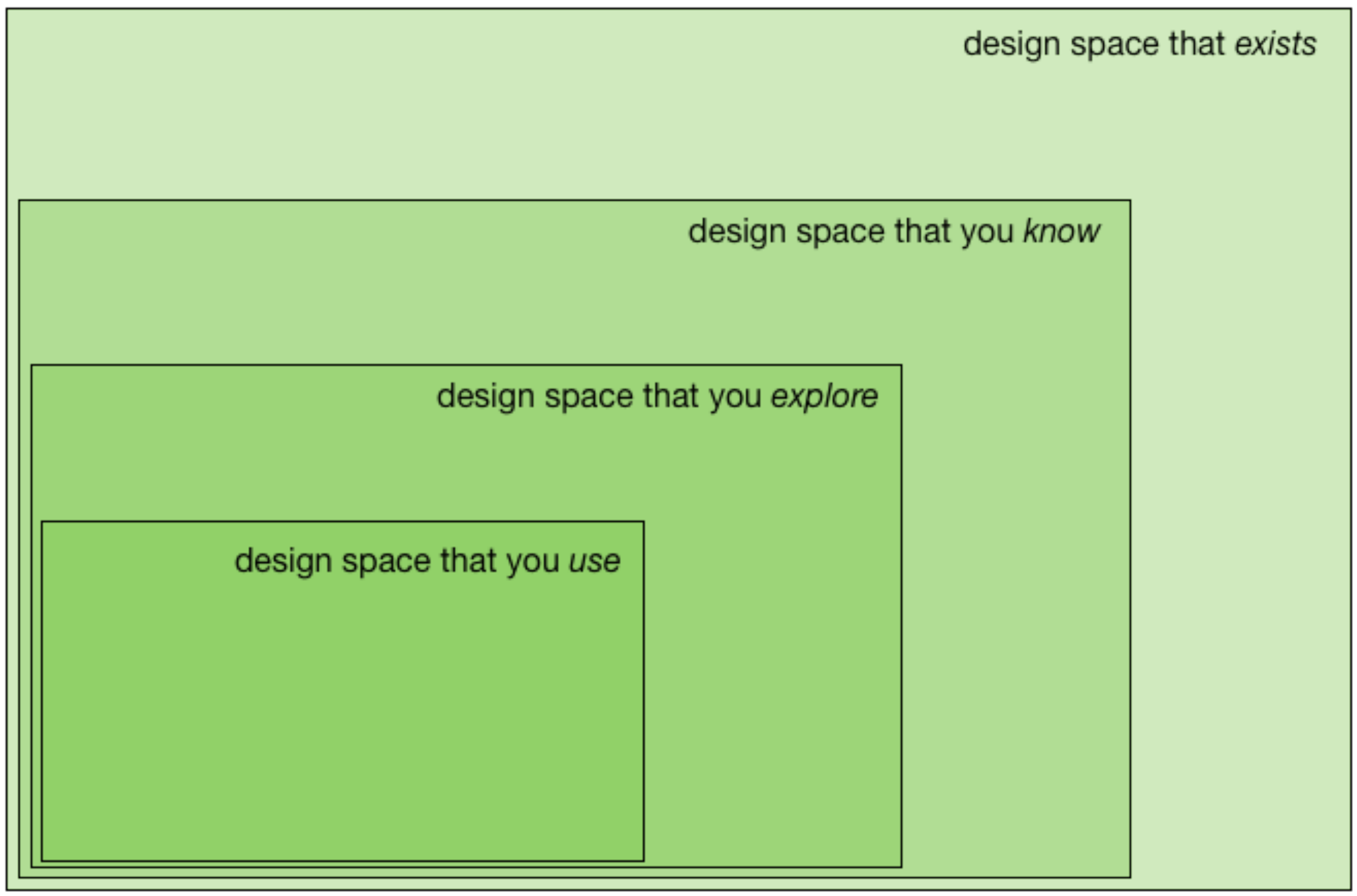
2. **Effectiveness**: Are there more perceptually accurate visual variables you can use? (see Mackinlay)

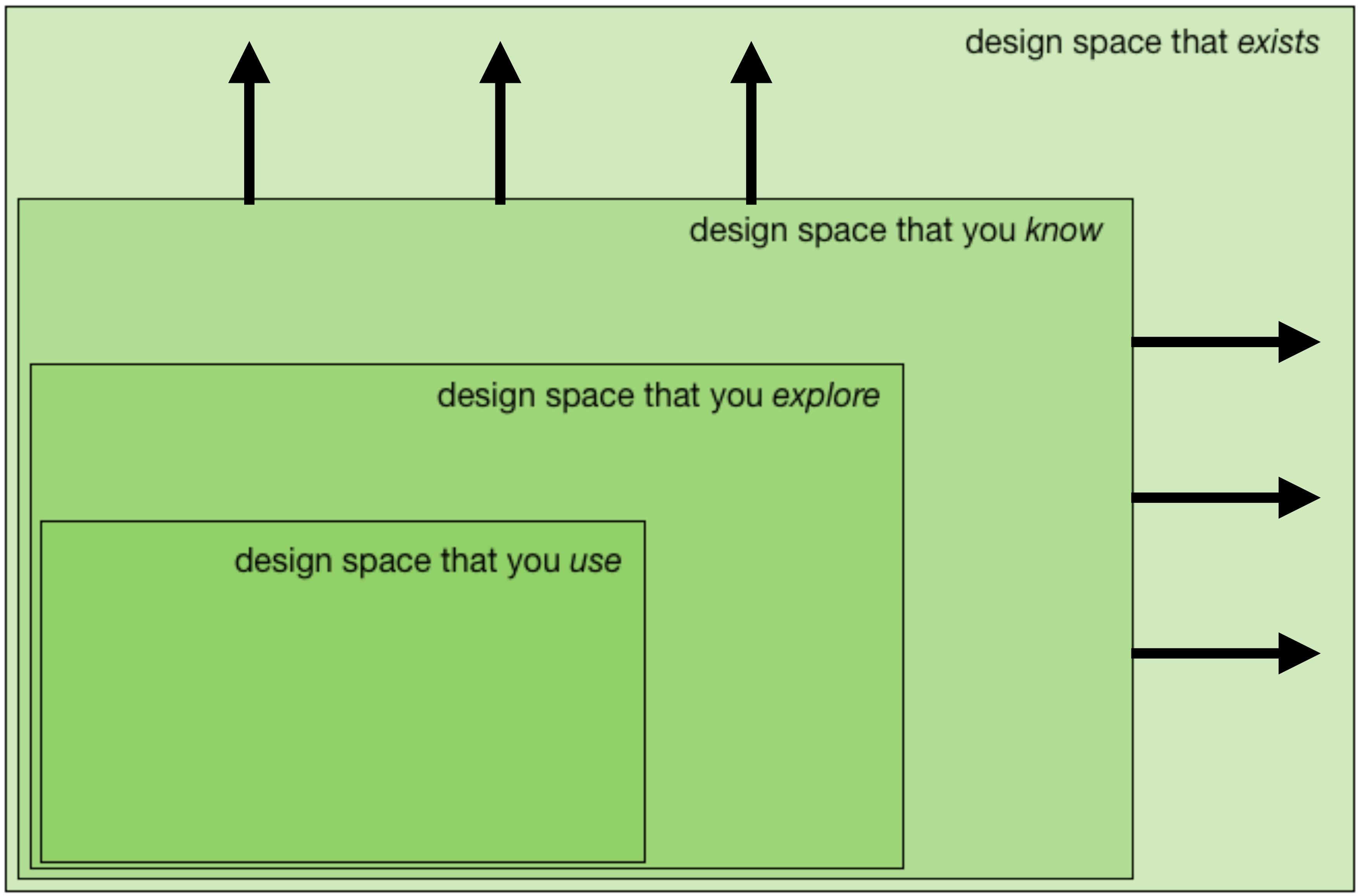
3. **Patterns**: Are the patterns easy to decode?

4. **Exploration**: How does the interactivity augment our abilities in visual data analysis?

Questions 1-2: relate to the **perception and cognition**; questions 3-4: involve the **user's knowledge and tasks**.

Design fixation

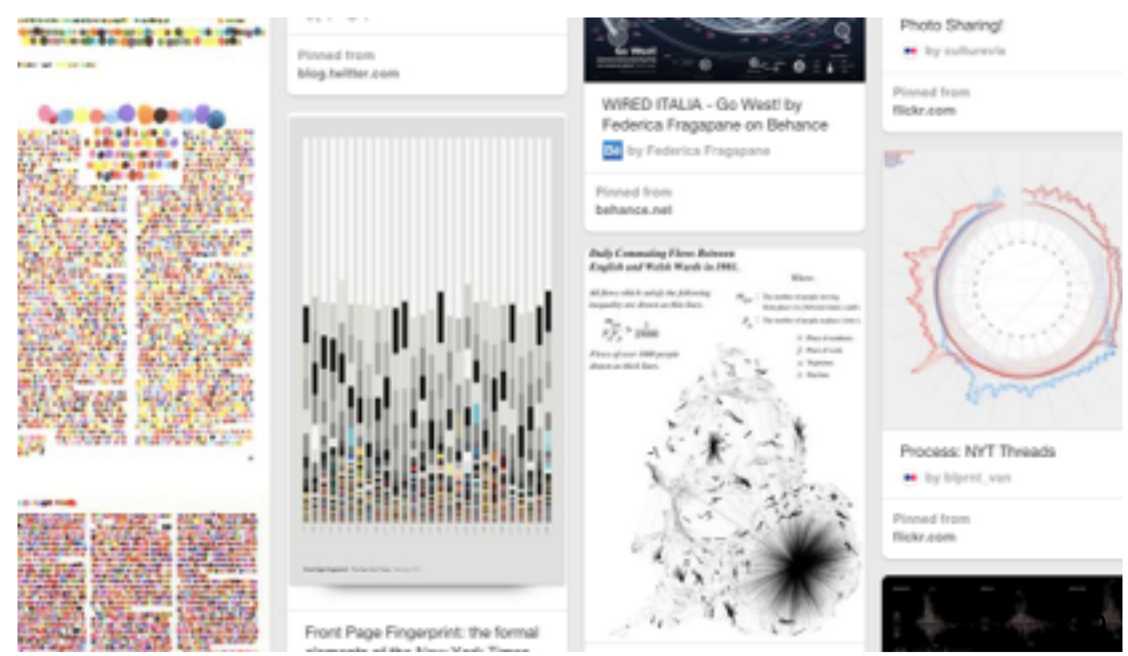




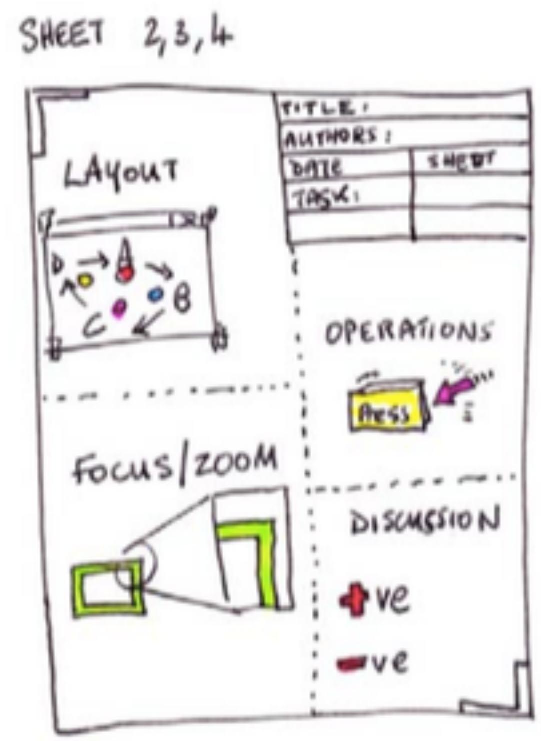
problem: initial design space that you know is small => how to expand?

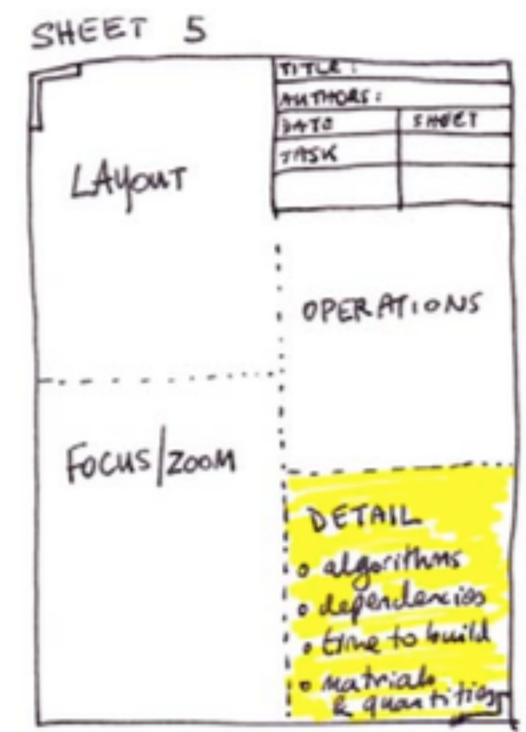
Tips to explore design space:

- 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



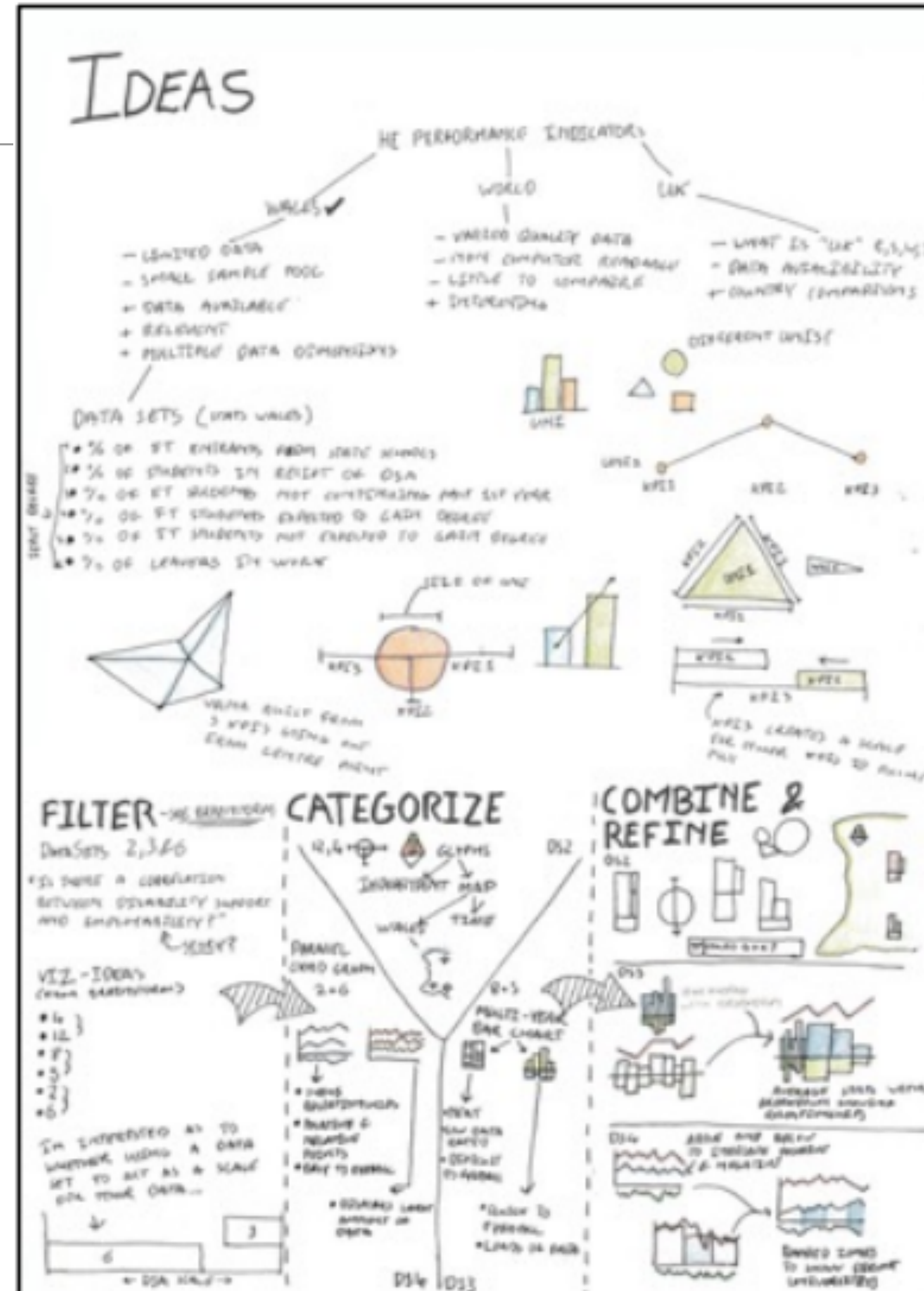


brain storm → initial designs → realisation design

Roberts et al (2015).Sketching Designs Using the Five Design-Sheet Methodology. See http://fds.design

Sheet 1: brainstorm

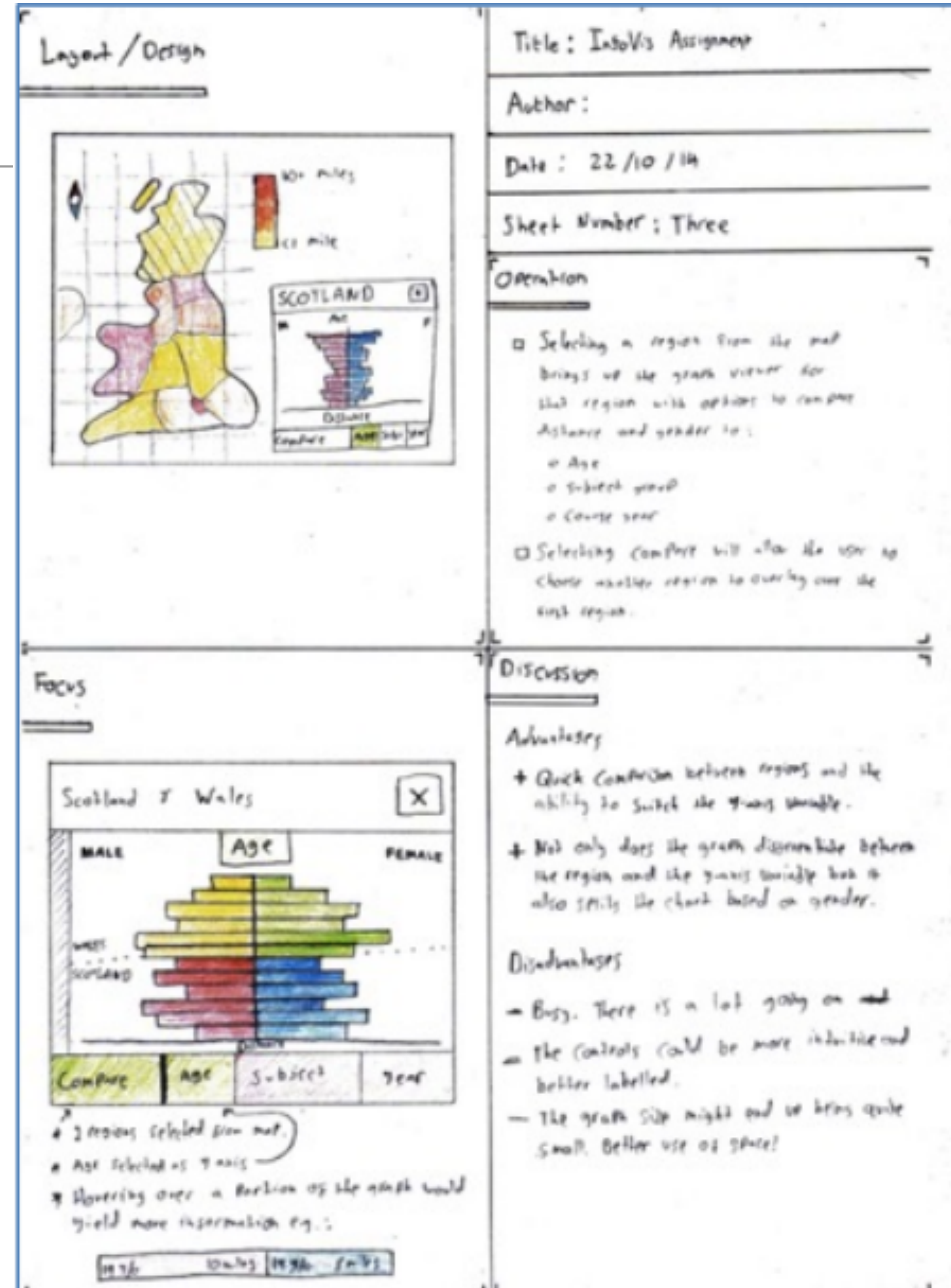
- list many **ideas** (including bad ones!)
- **filter** out the ones that are not non-sense
- **categorize** these filtered ideas: how can you group them; how would you describe them?
- **combine & refine** these ideas: interesting things often happen when combining visuals



Sheet 2-4: initial designs

Pick 3 ***meaningfully-different!*** (not aesthetically-different) ideas from sheet 1

- mention **meta-info**: date, title, number to refer to later, …
- draw an **overview** of the visual
- describe the **parts** of the visual
- **focus/parti**: explain/annotate the core of the visual: how does the most important part of the visual work?
- **discuss**

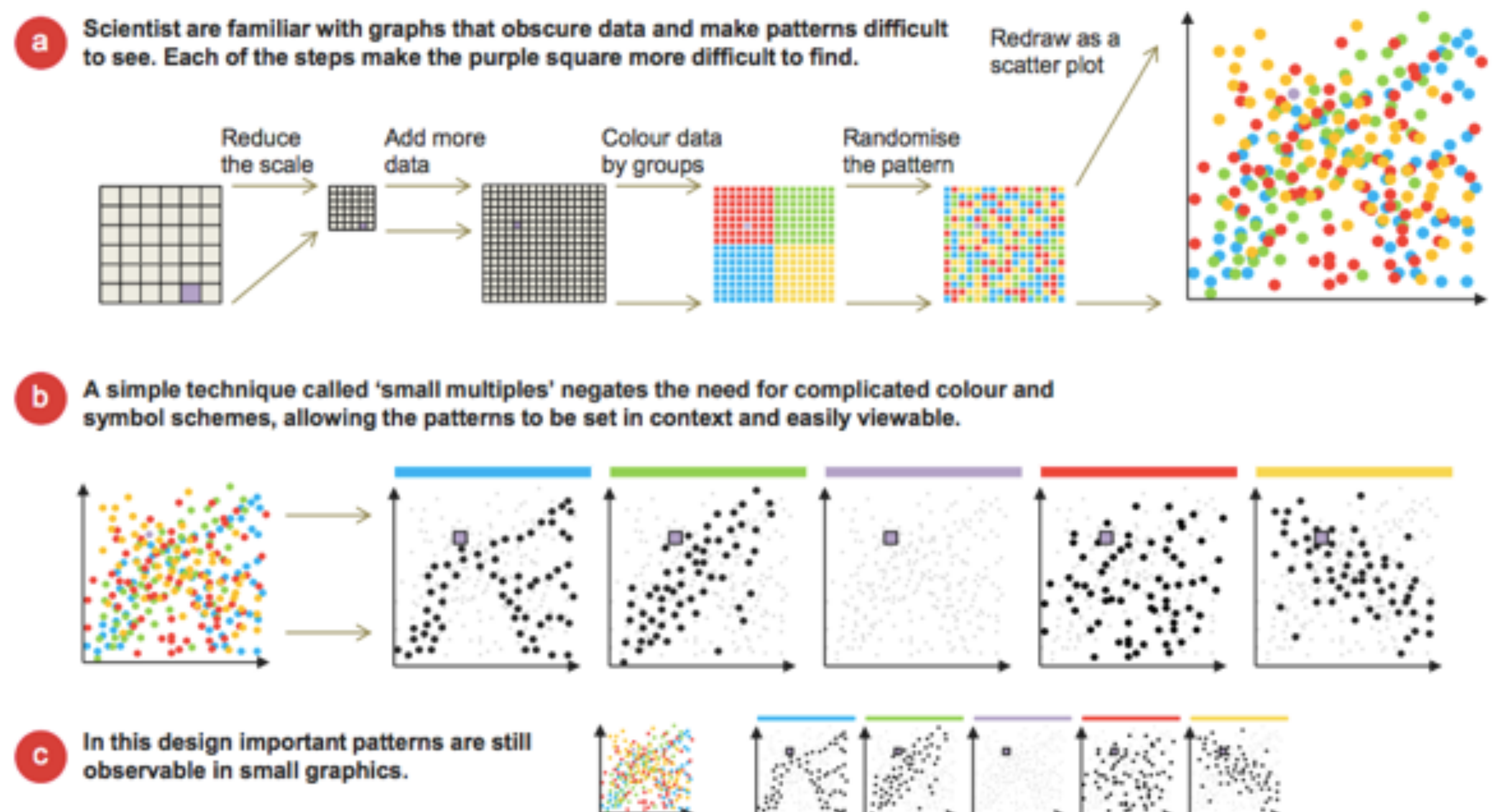


Sheet 5: realization sheet

Pick one or combine several of sheets 2-4, and work out in more detail (same parts as sheets 2-4)

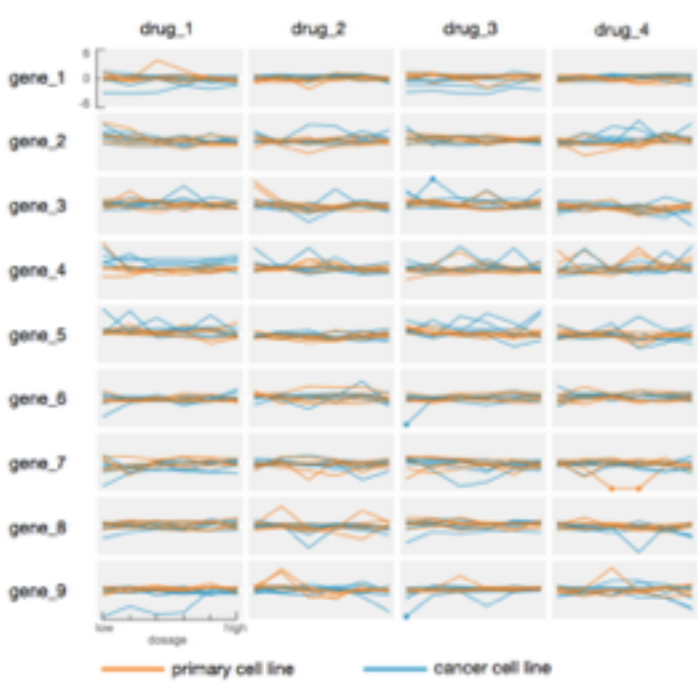
Combining visuals

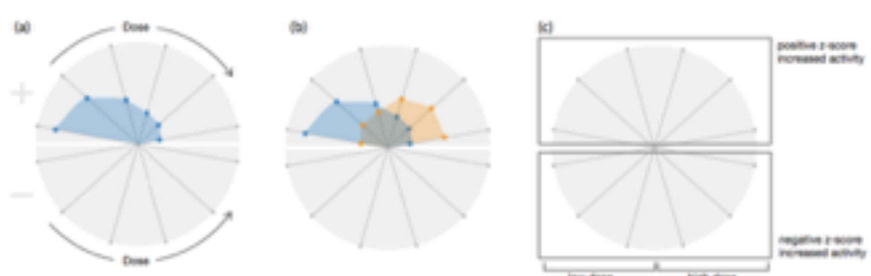
- **juxtapose**/facet,
  e.g. *small multiples*
  - easier to see trends
  - pop-out effect
  - very powerful!



reference?

L1000 data



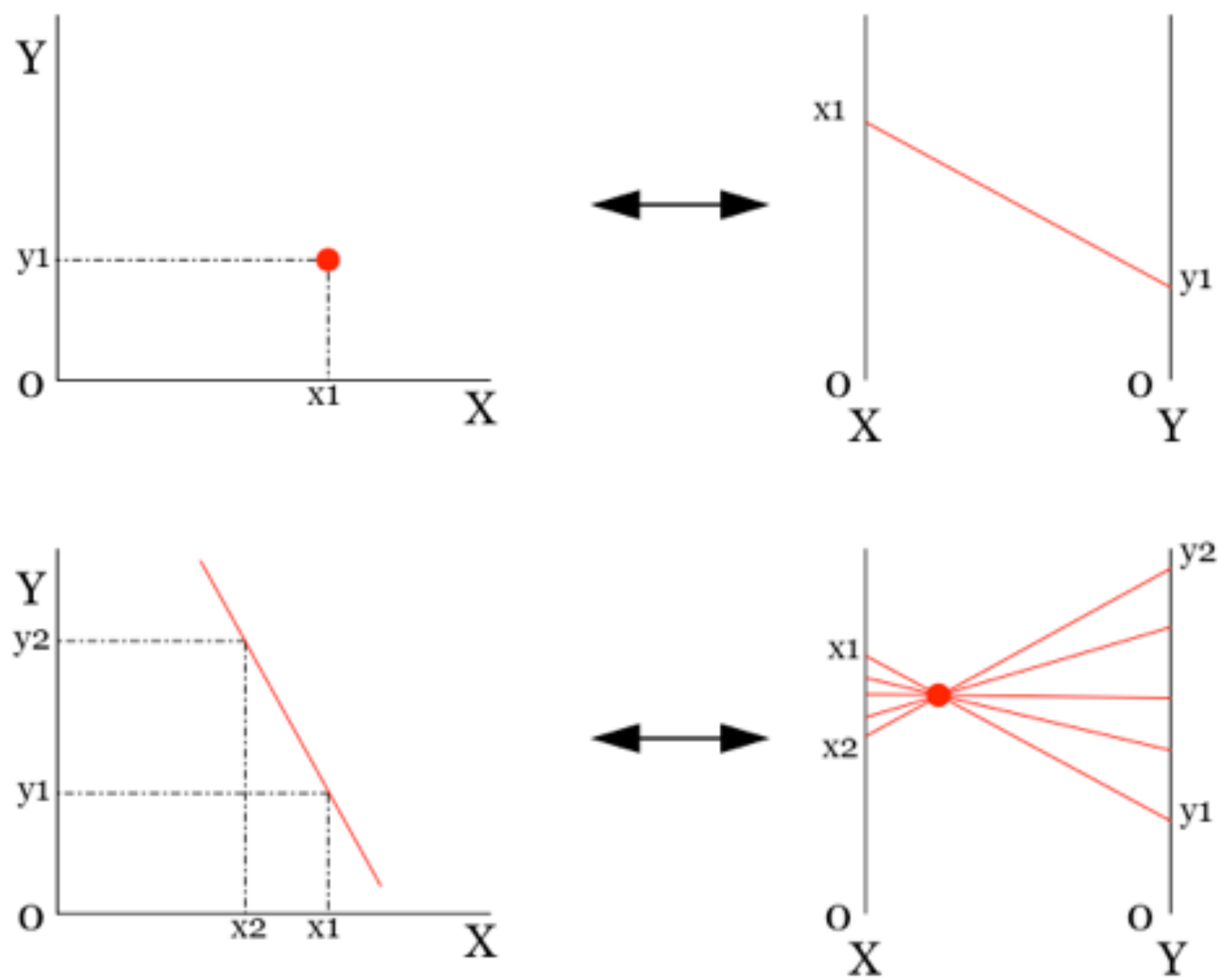


[F] Examples

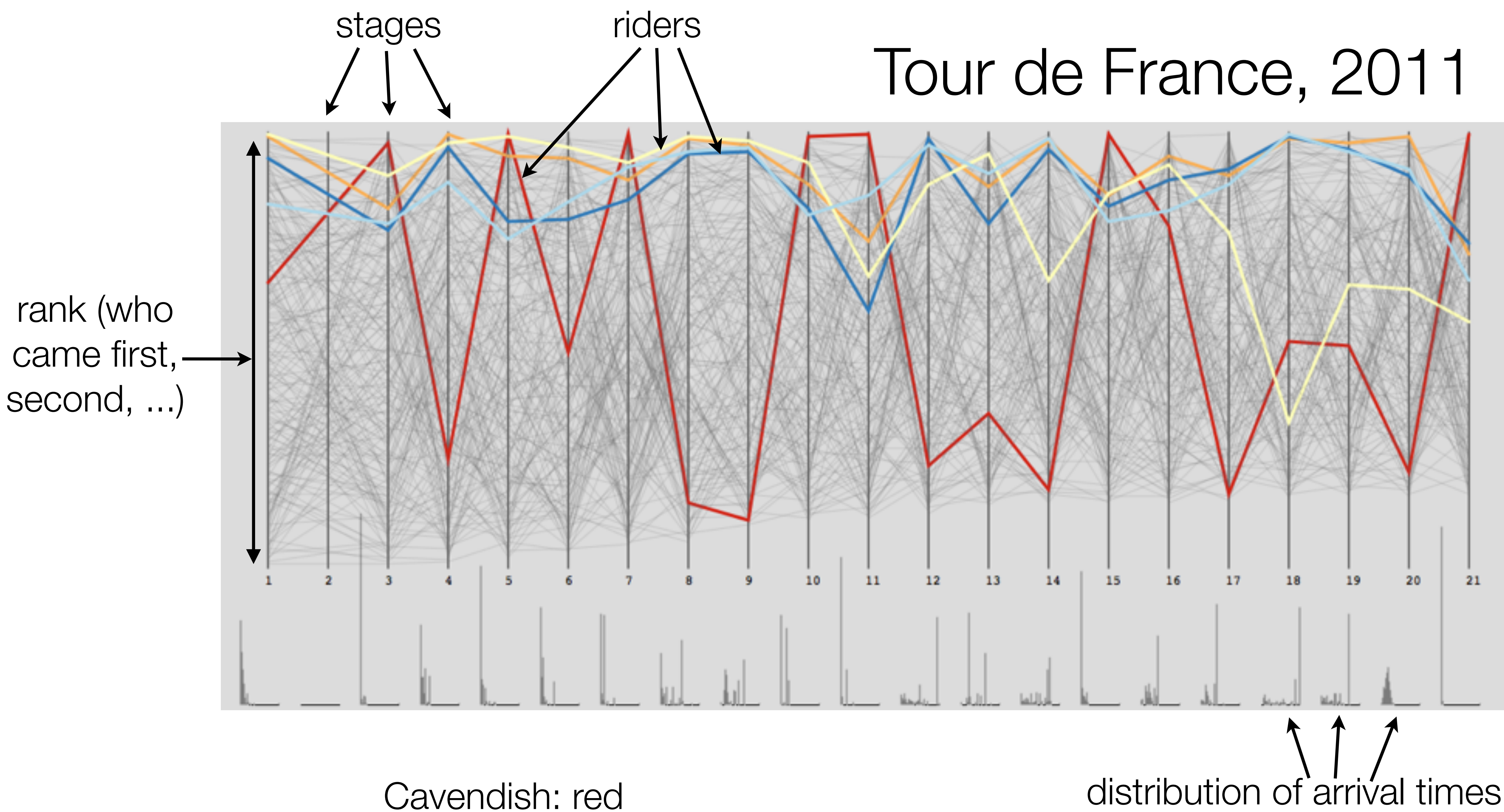
very (!) limited scattering of visuals, only to indicate the breadth of possibilities

High-dimensional data

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



Tour de France, 2011

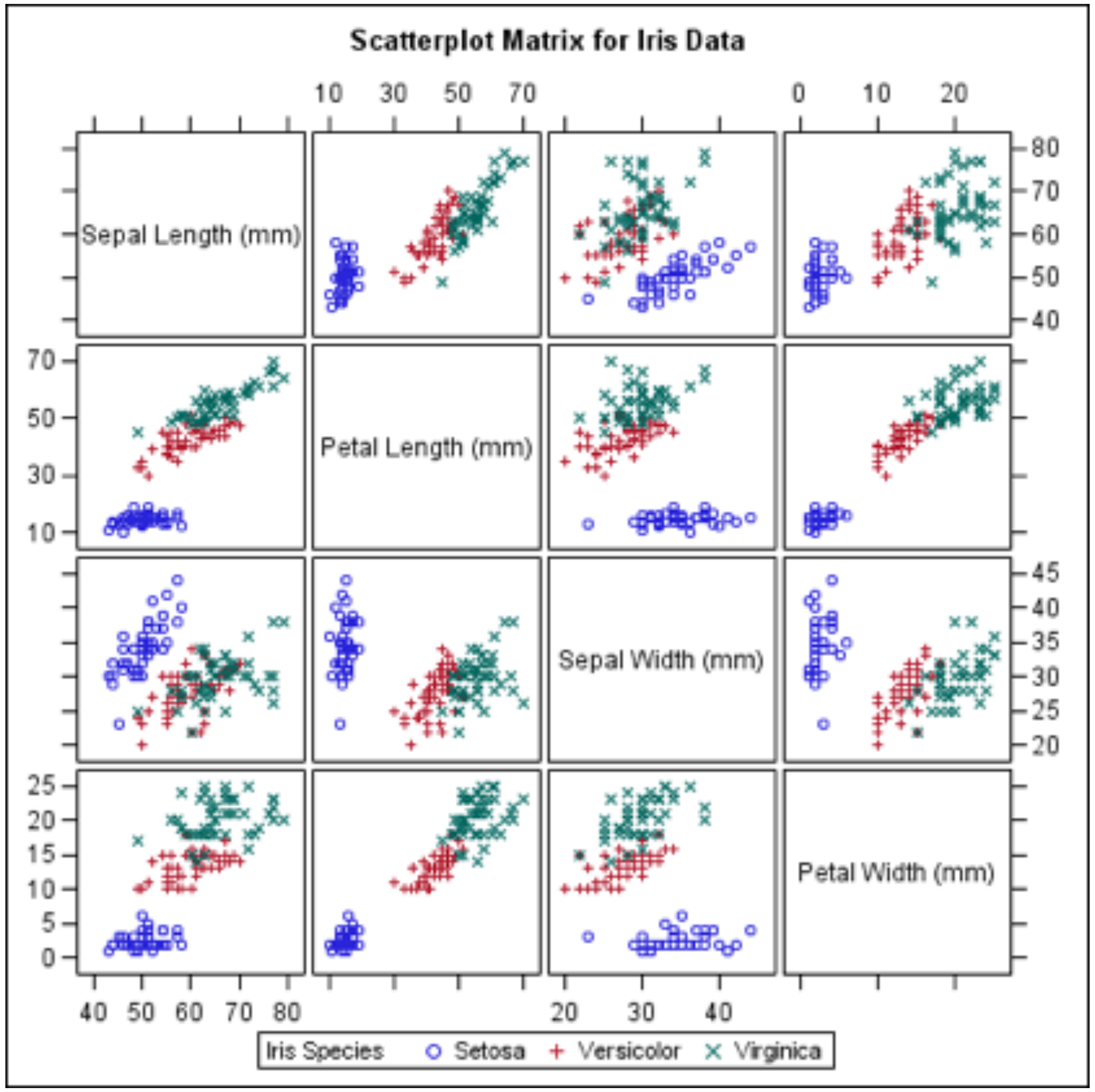


Cavendish: red
Evans: orange
Gilbert: yellow
Andy Schleck: light blue
Frank Schleck: dark blue

Iris setosa

Iris versicolor

Iris virginica



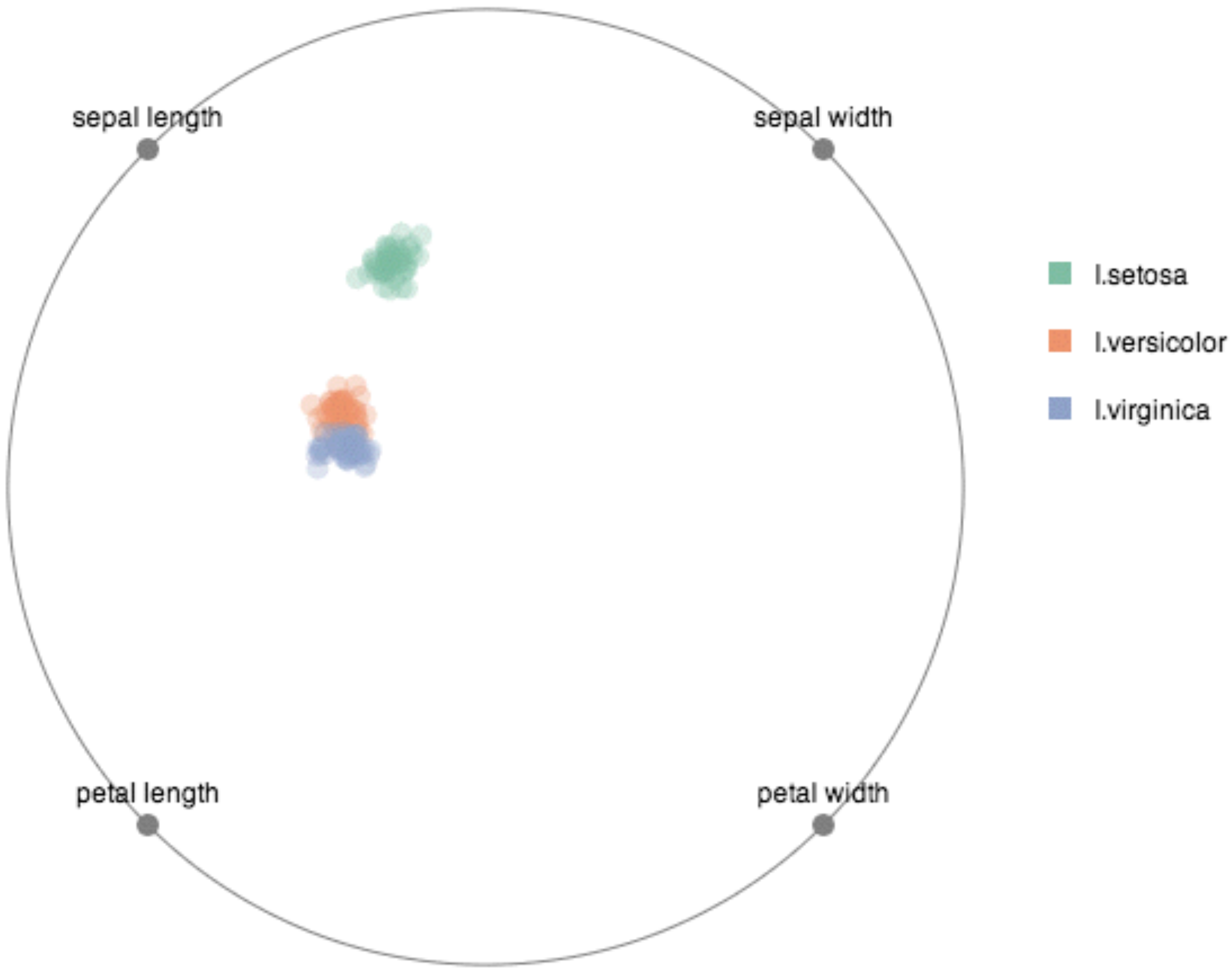
scatterplot matrix

(source: support.sas.com)

Iris setosa

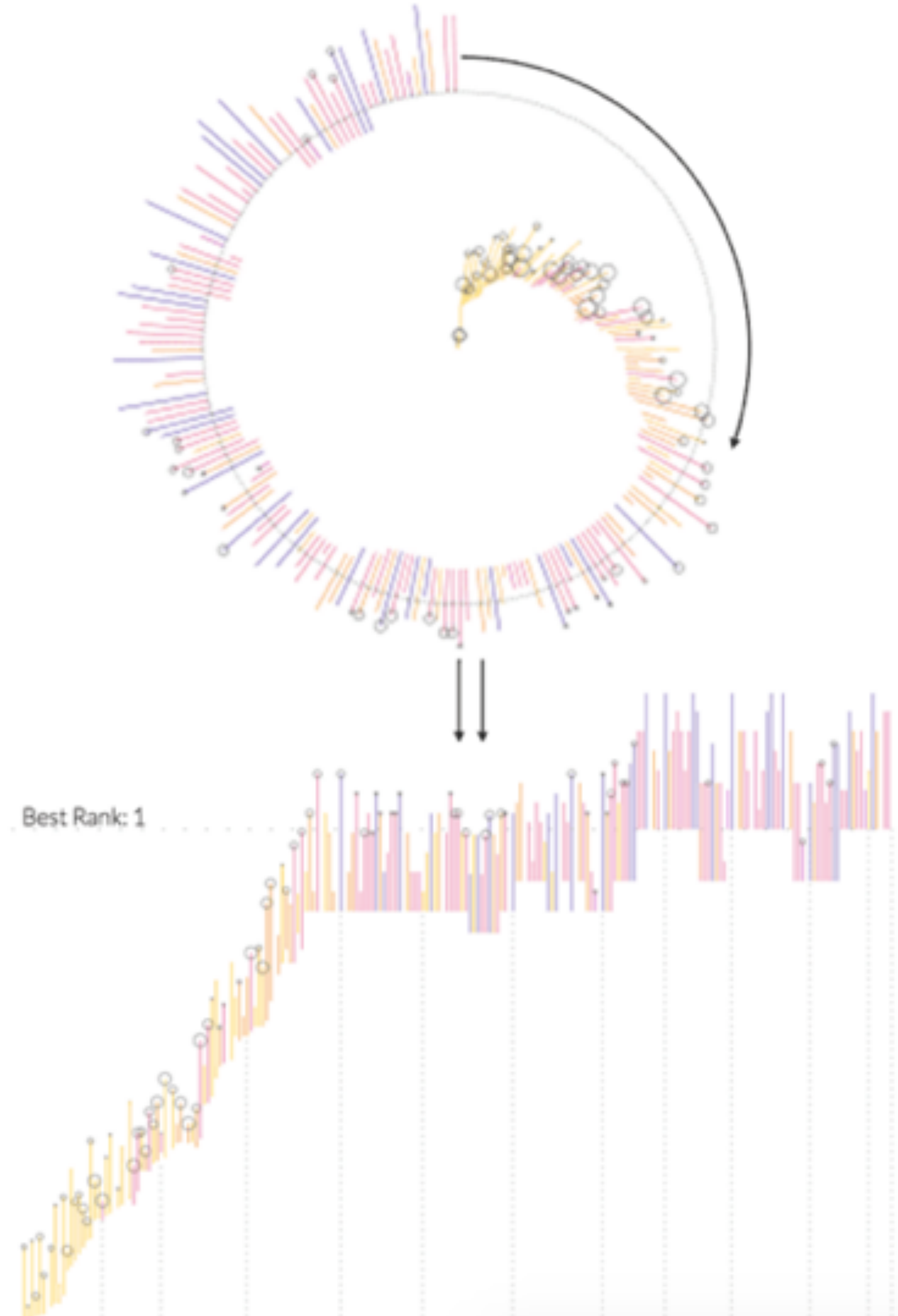
Iris versicolor

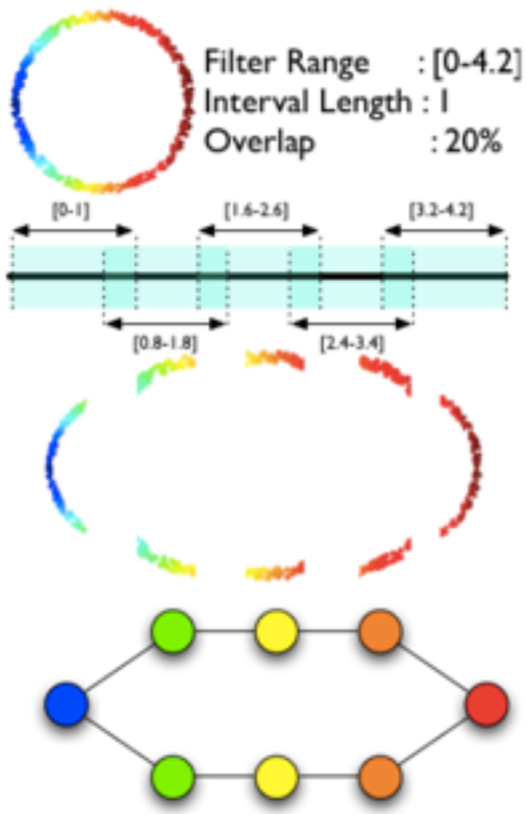
Iris virginica

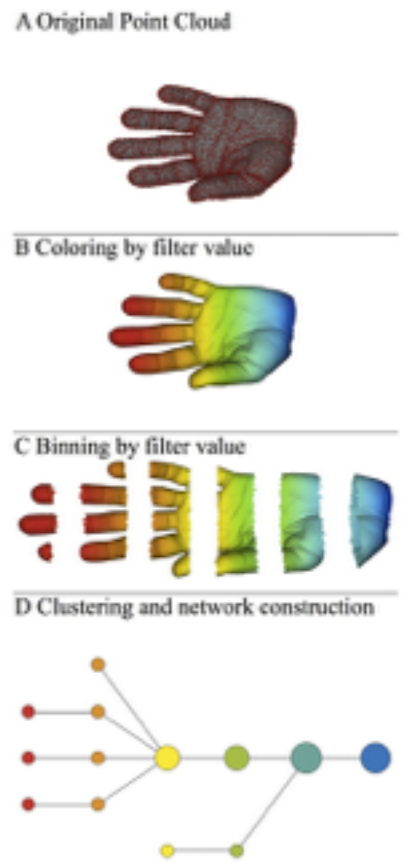


radviz

Untangling tennis (Kim Albrecht; untangling-tennis.net)







topological data analysis

local clustering in global context

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

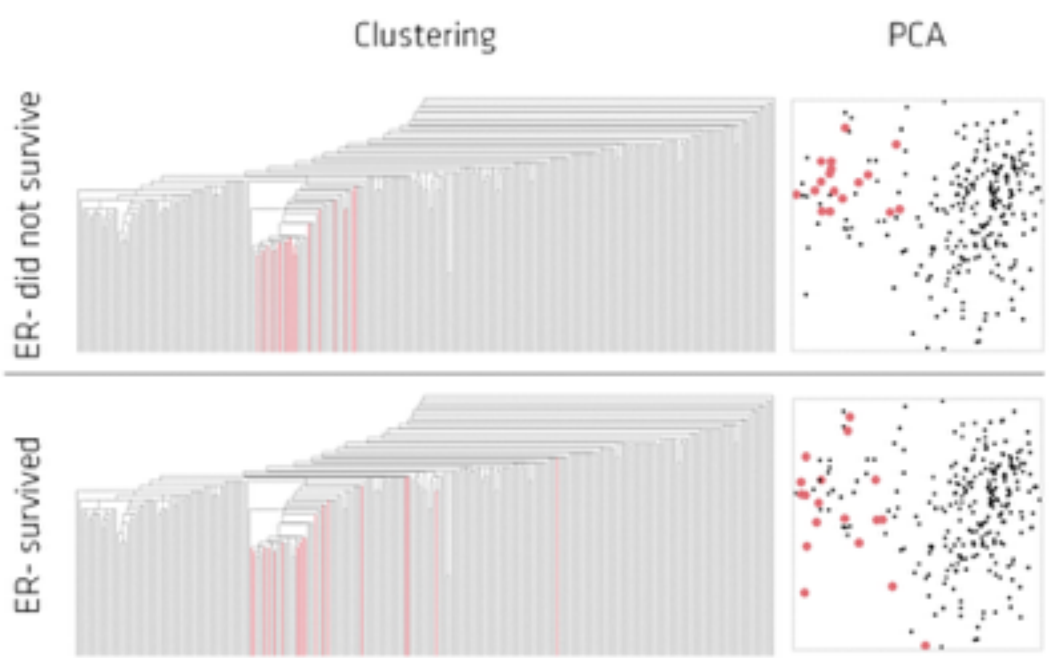
dead

survived

lens = L-infinity centrality + event death

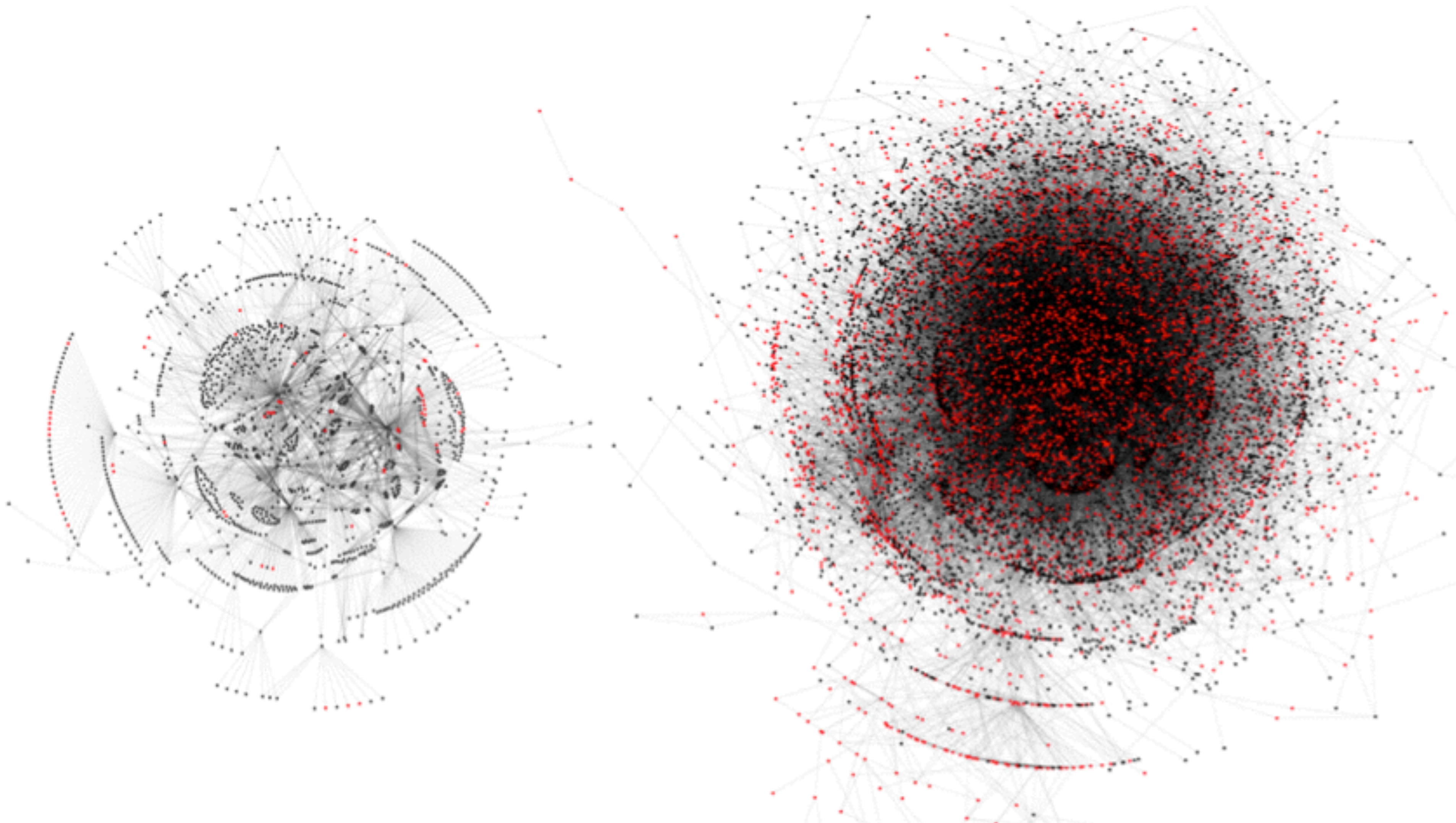
Ayasdi, Inc

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

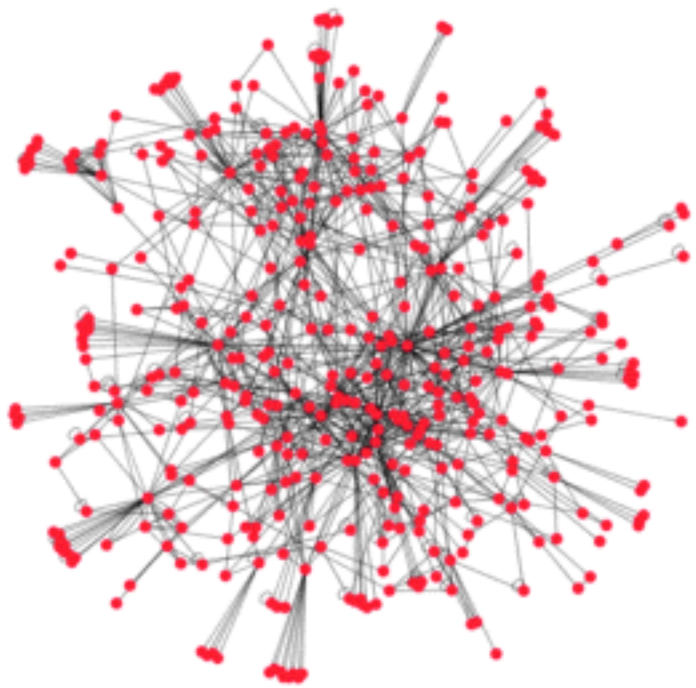


Networks

node-link diagrams



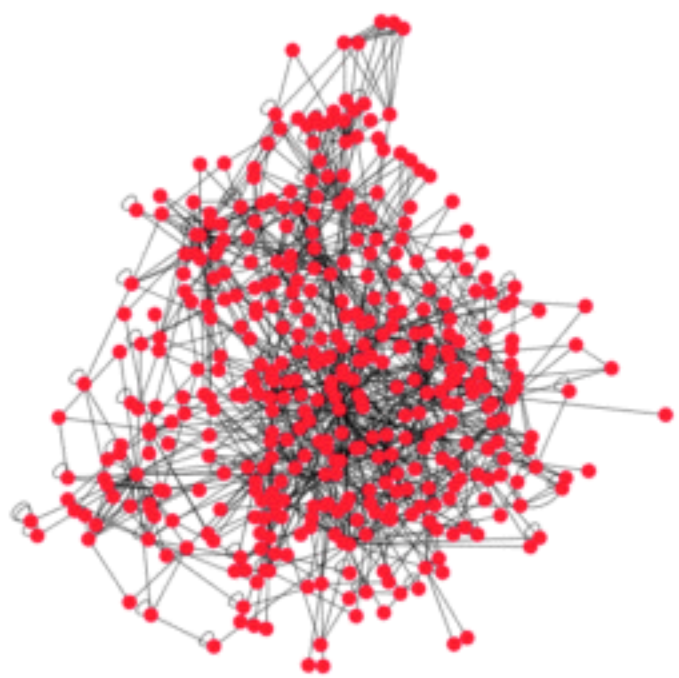
Martin Krzywinski



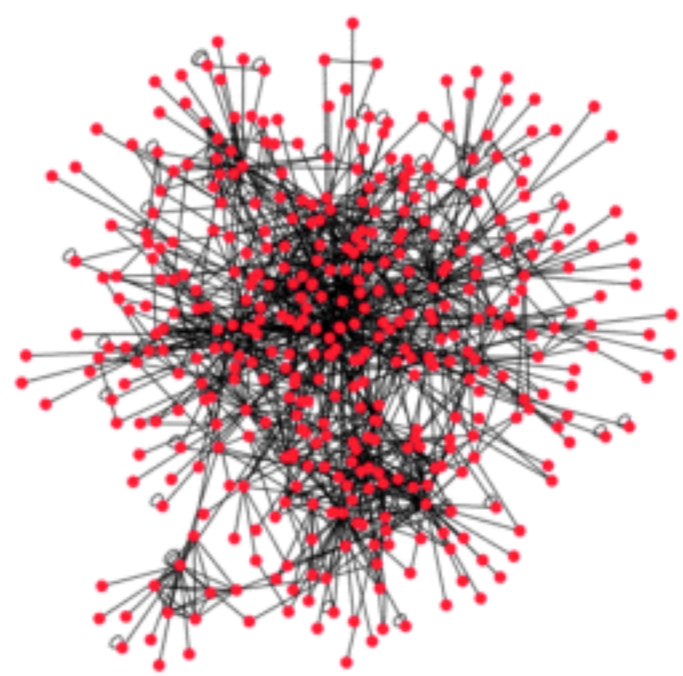
EDGE WEIGHTED
SPRING EMBEDDED

SPRING EMBEDDED

same network

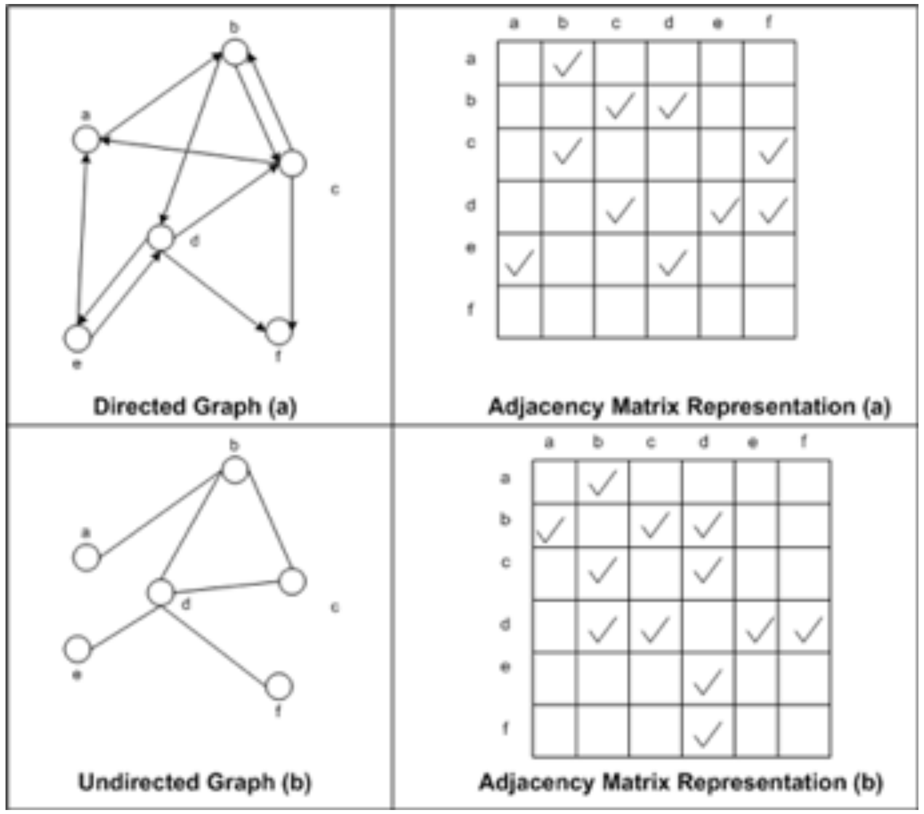


ORGANIC

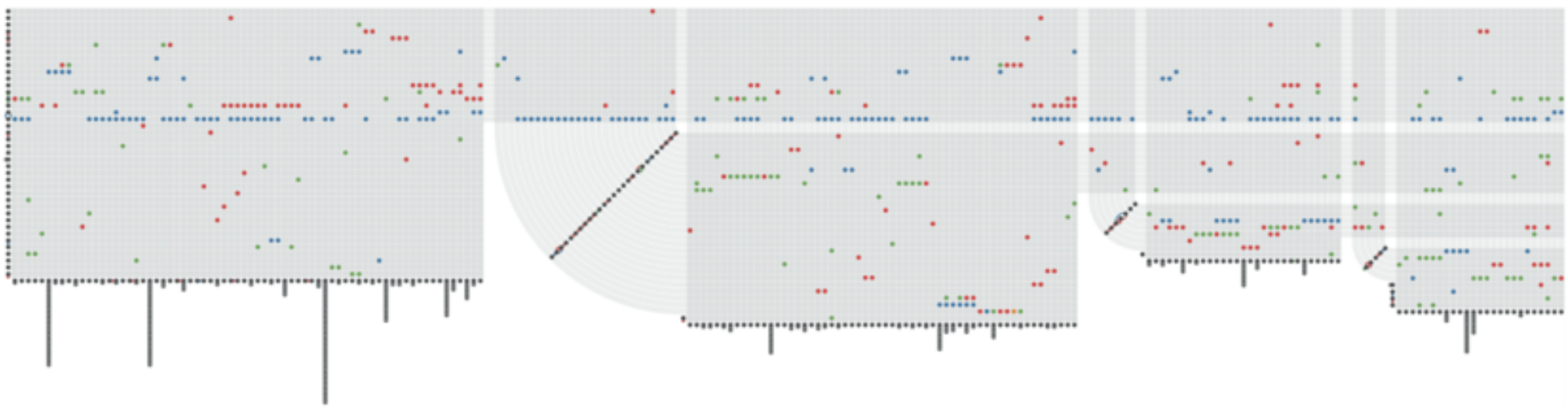


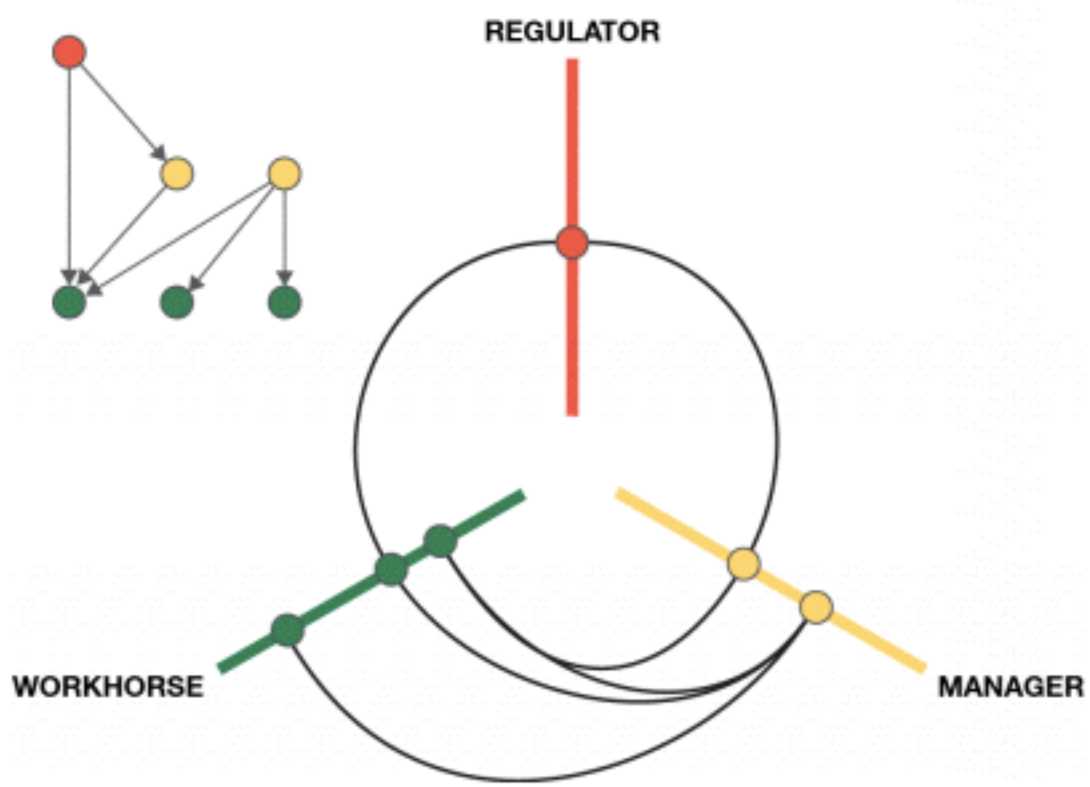
FORCE DIRECTED

- adjacency matrix

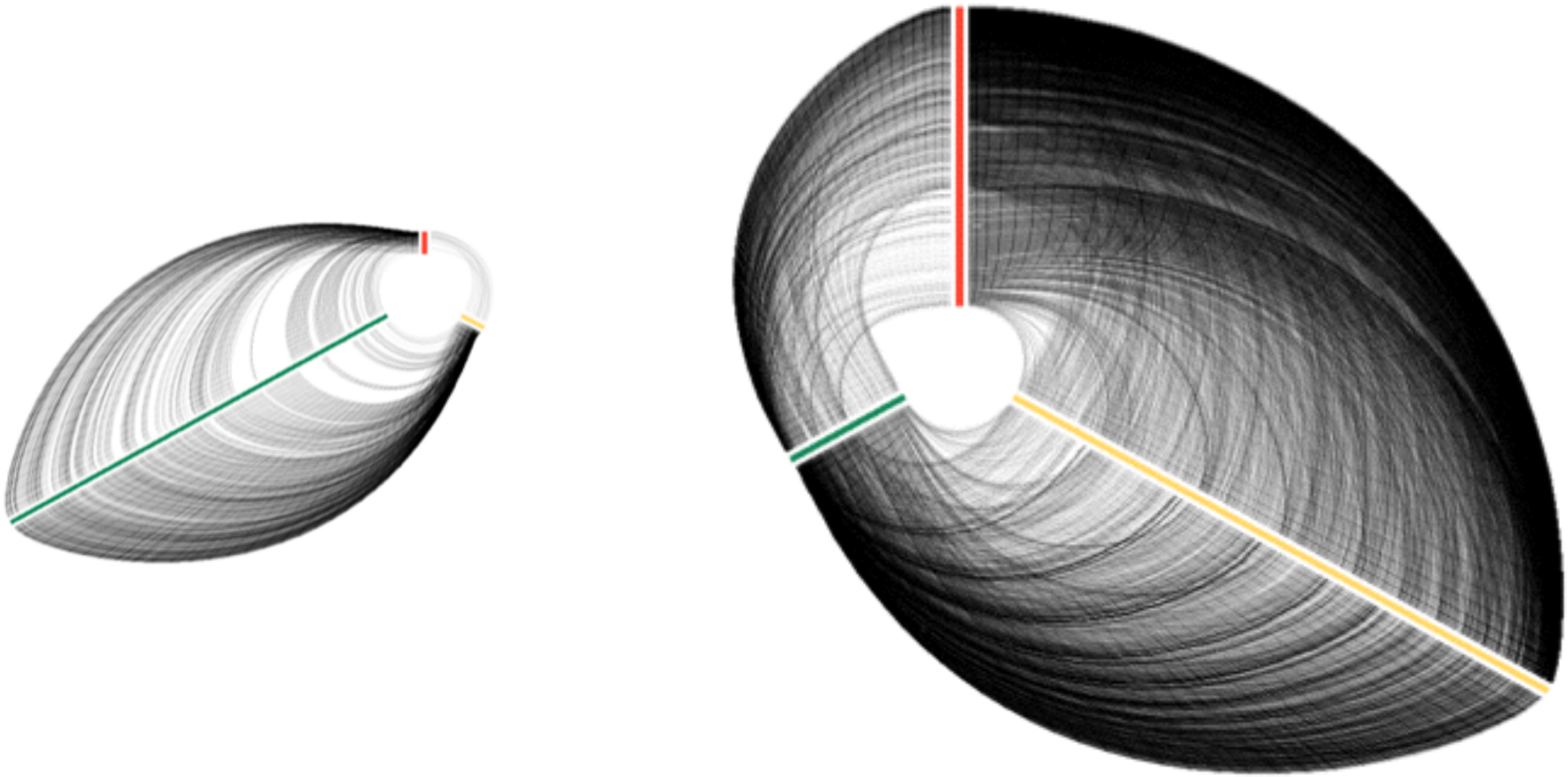


- compressed adjacency matrix (Dinkla *et al*, 2012)

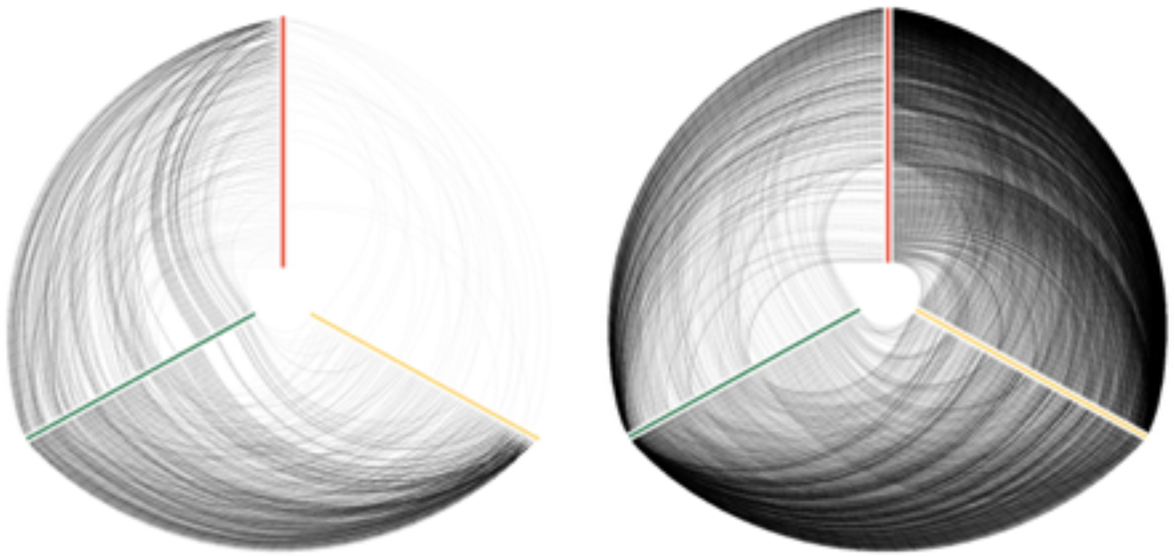



REGULATOR
WORKHORSE
MANAGER

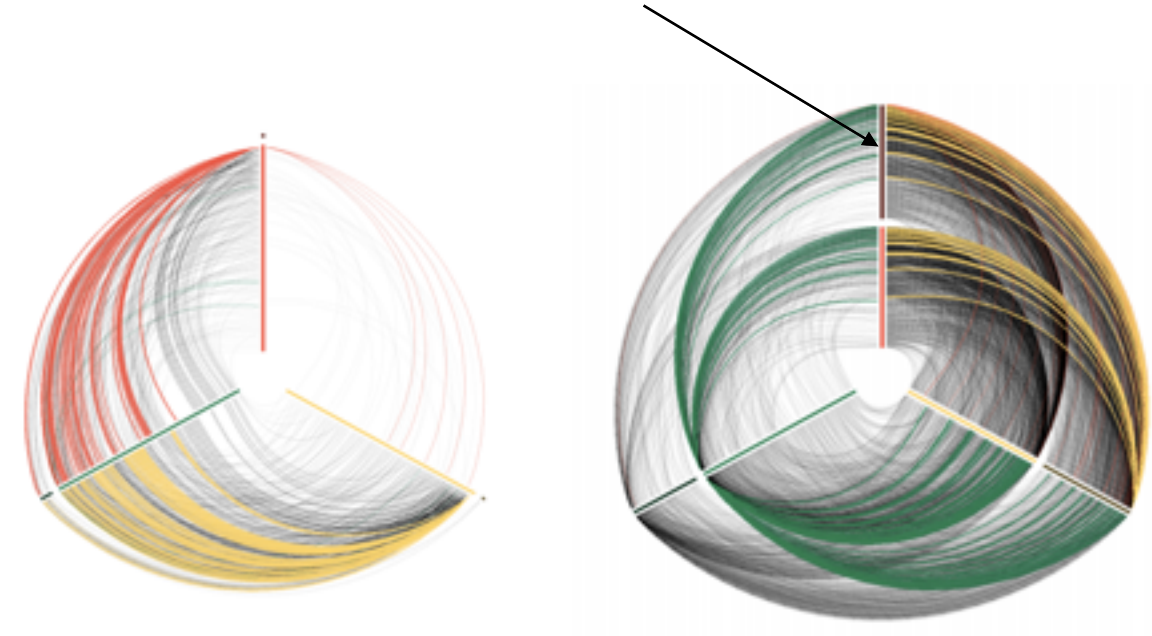
- hive plots



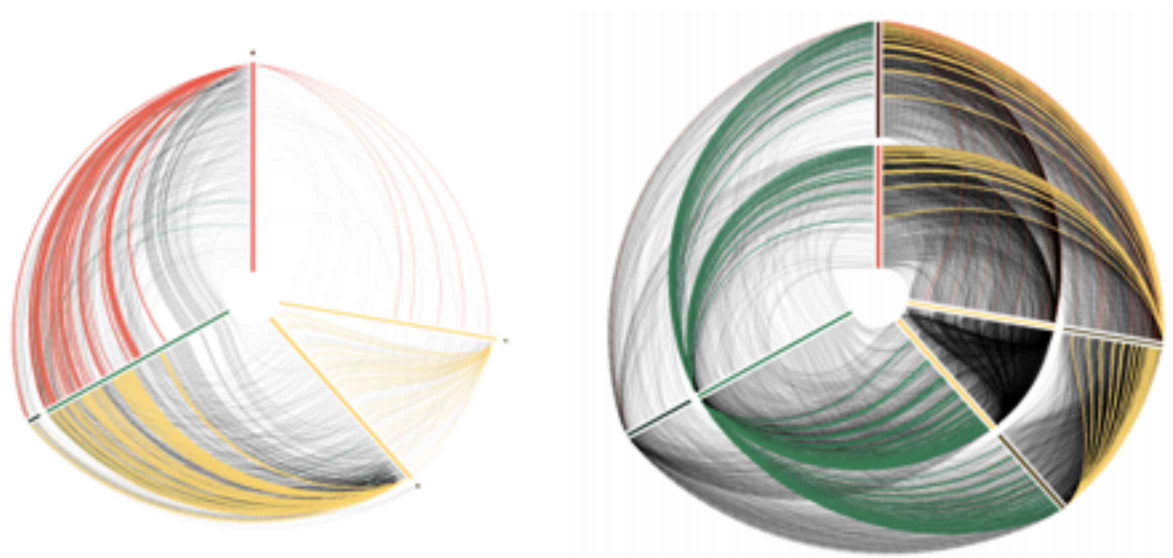
Martin Krzywinski



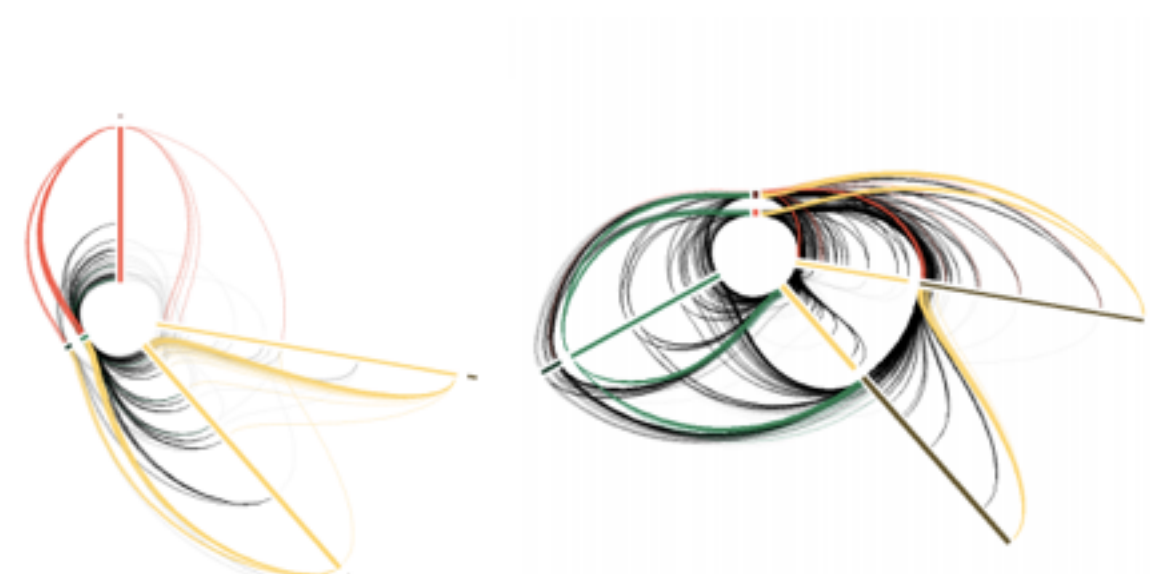
scaled to 100%



segmentation: non-persistent vs persistent



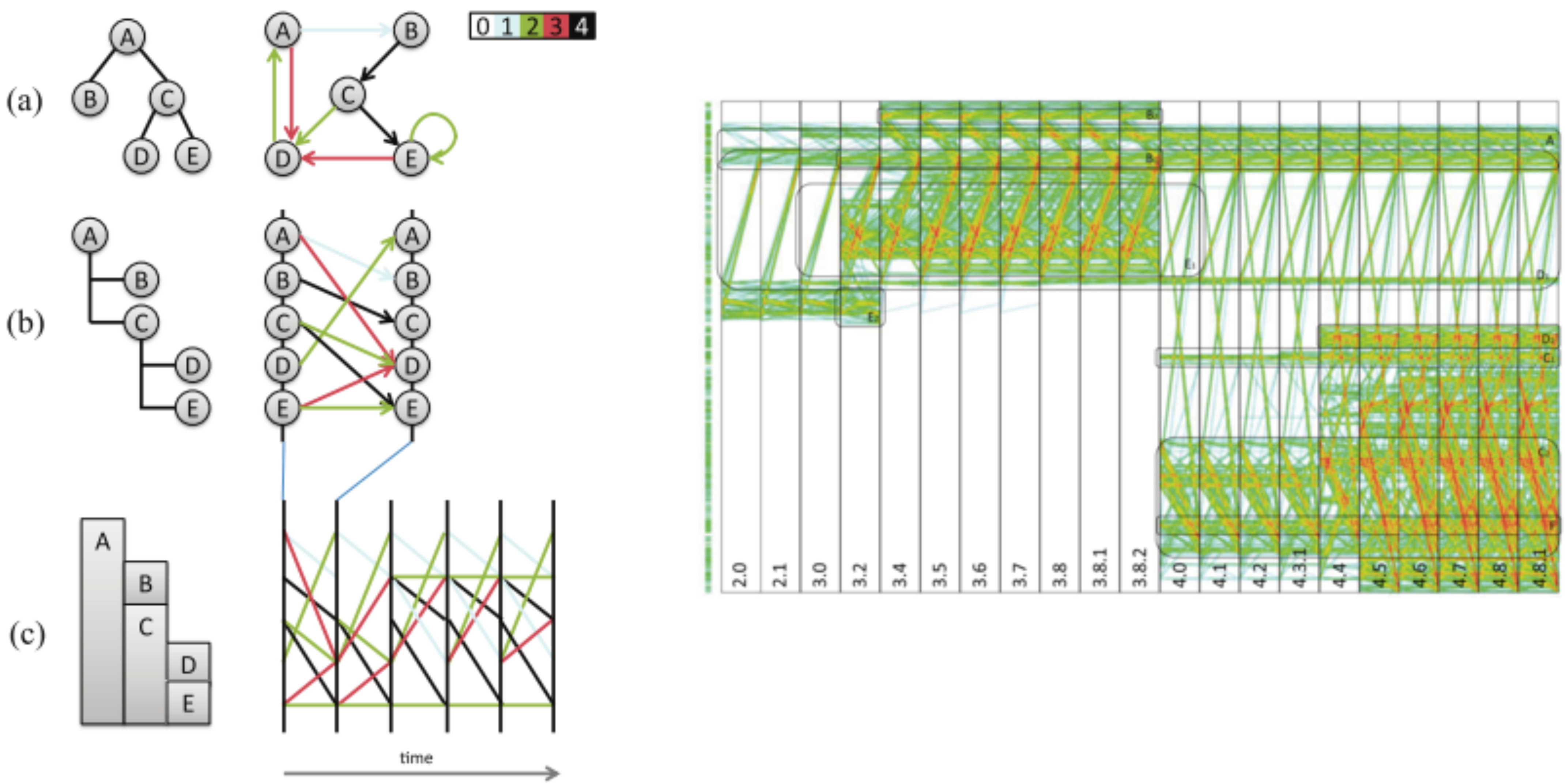
intra-axis connections

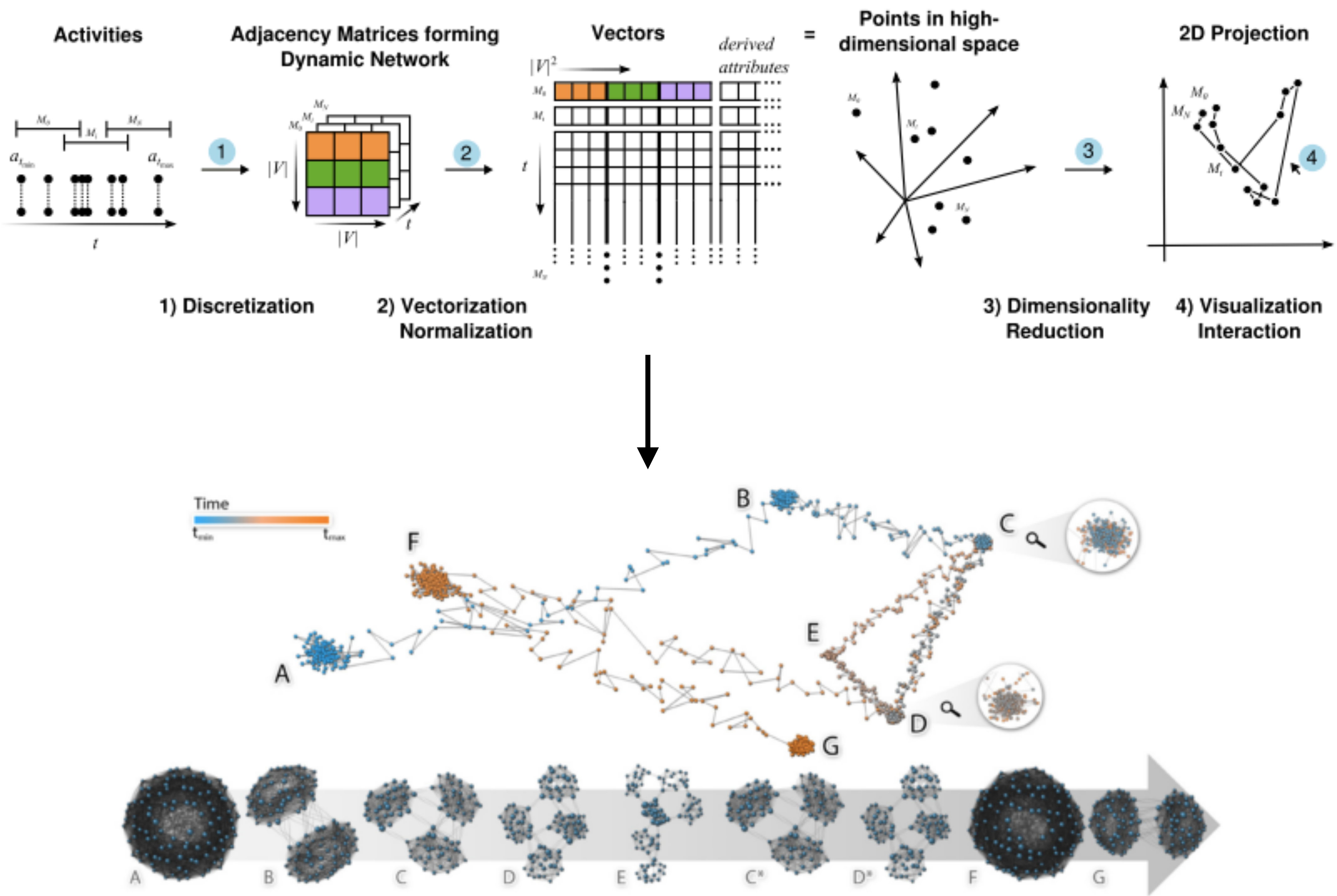


absolute connectivity

Dynamic networks

Parallel Edge Splatting (Burch et al, 2011)

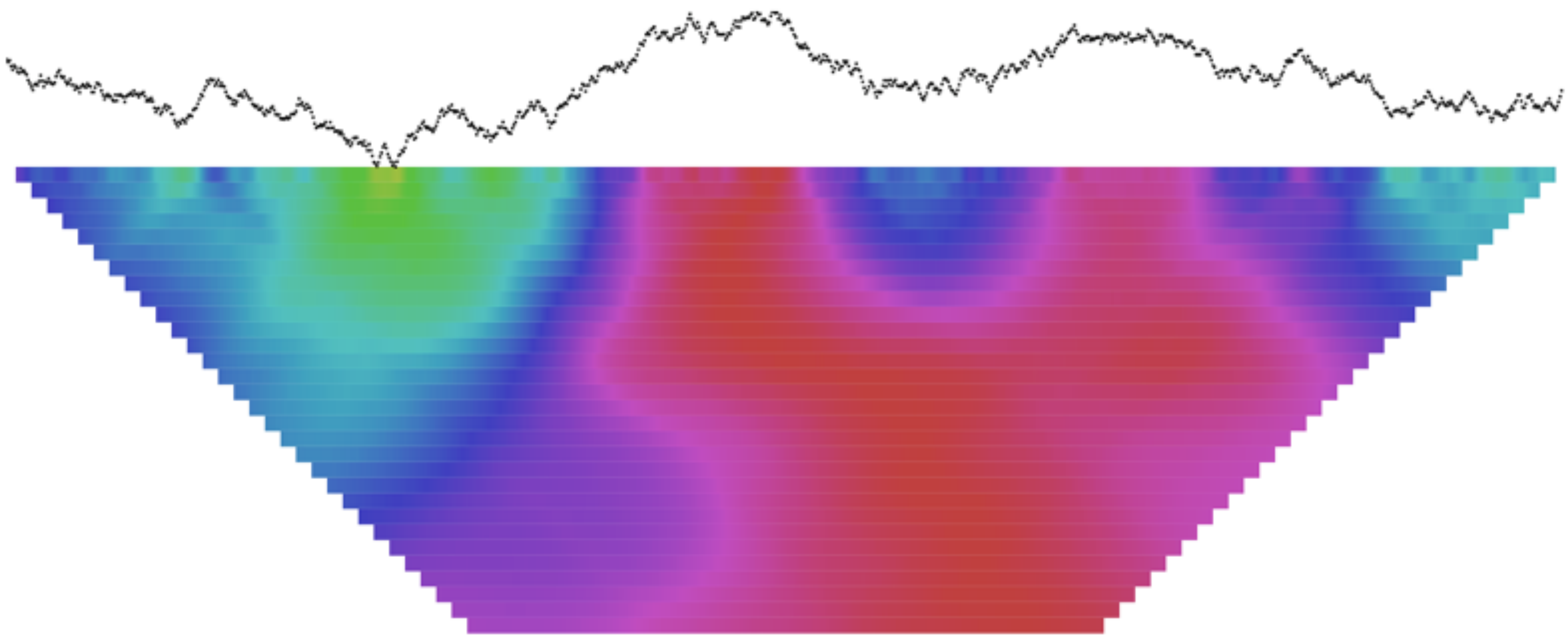




van den Elzen et al, 2015

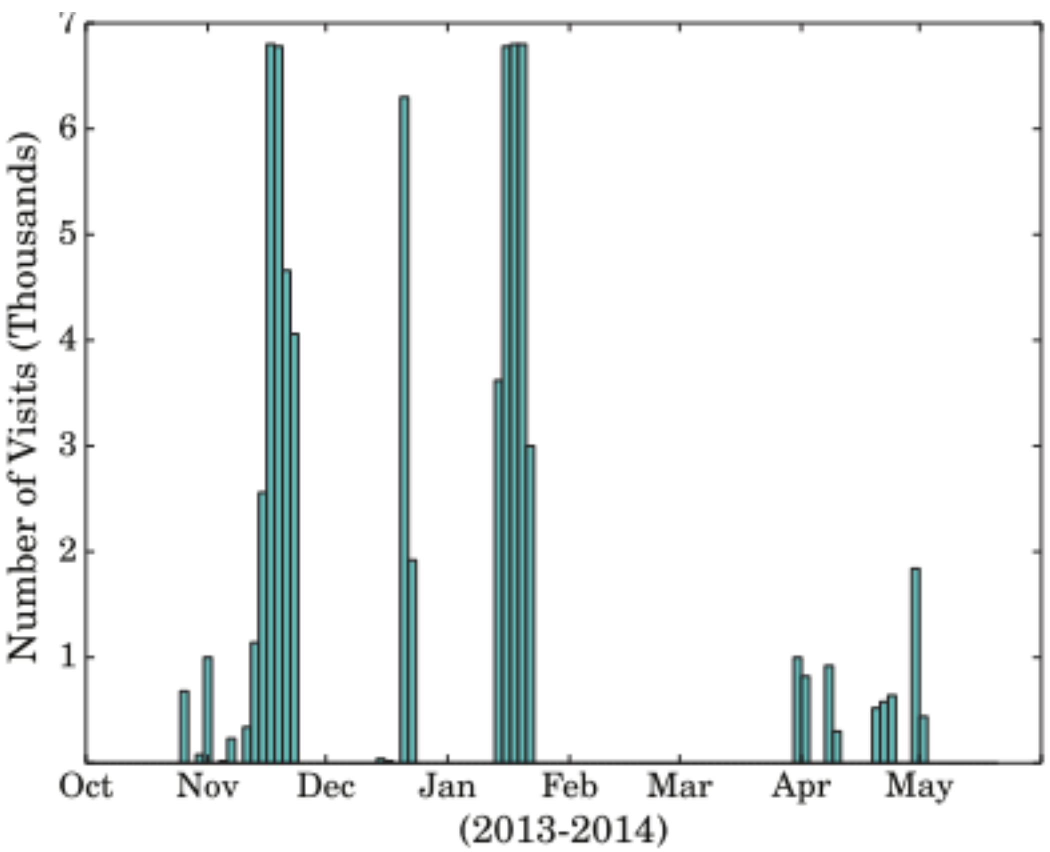
Parameter space

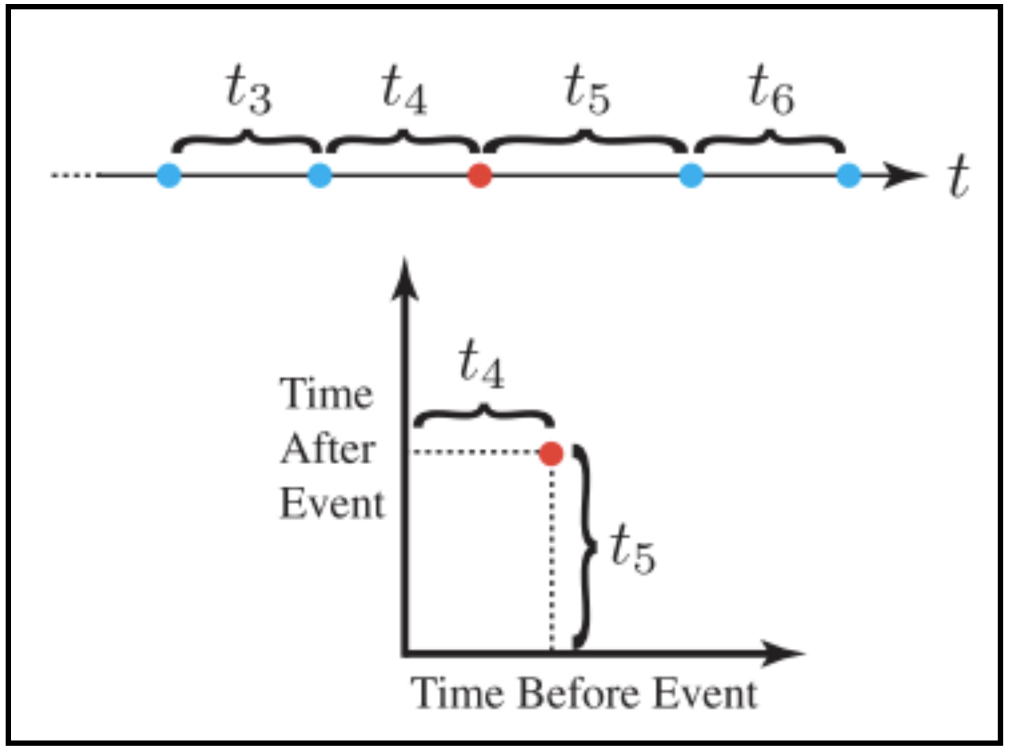
continuous signal -> what should my histogram bin-sizes be?

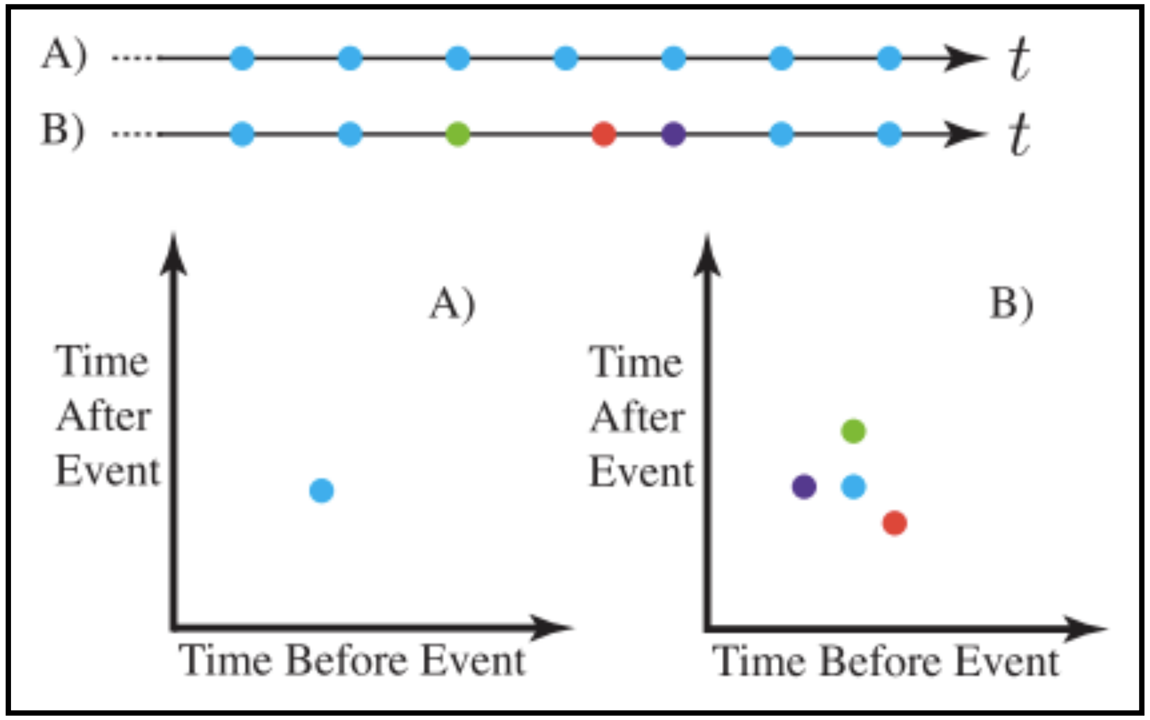


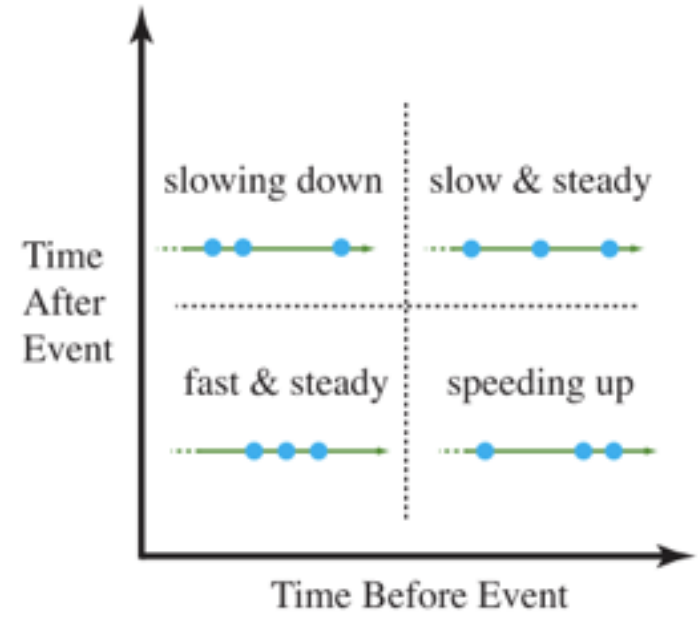
Time

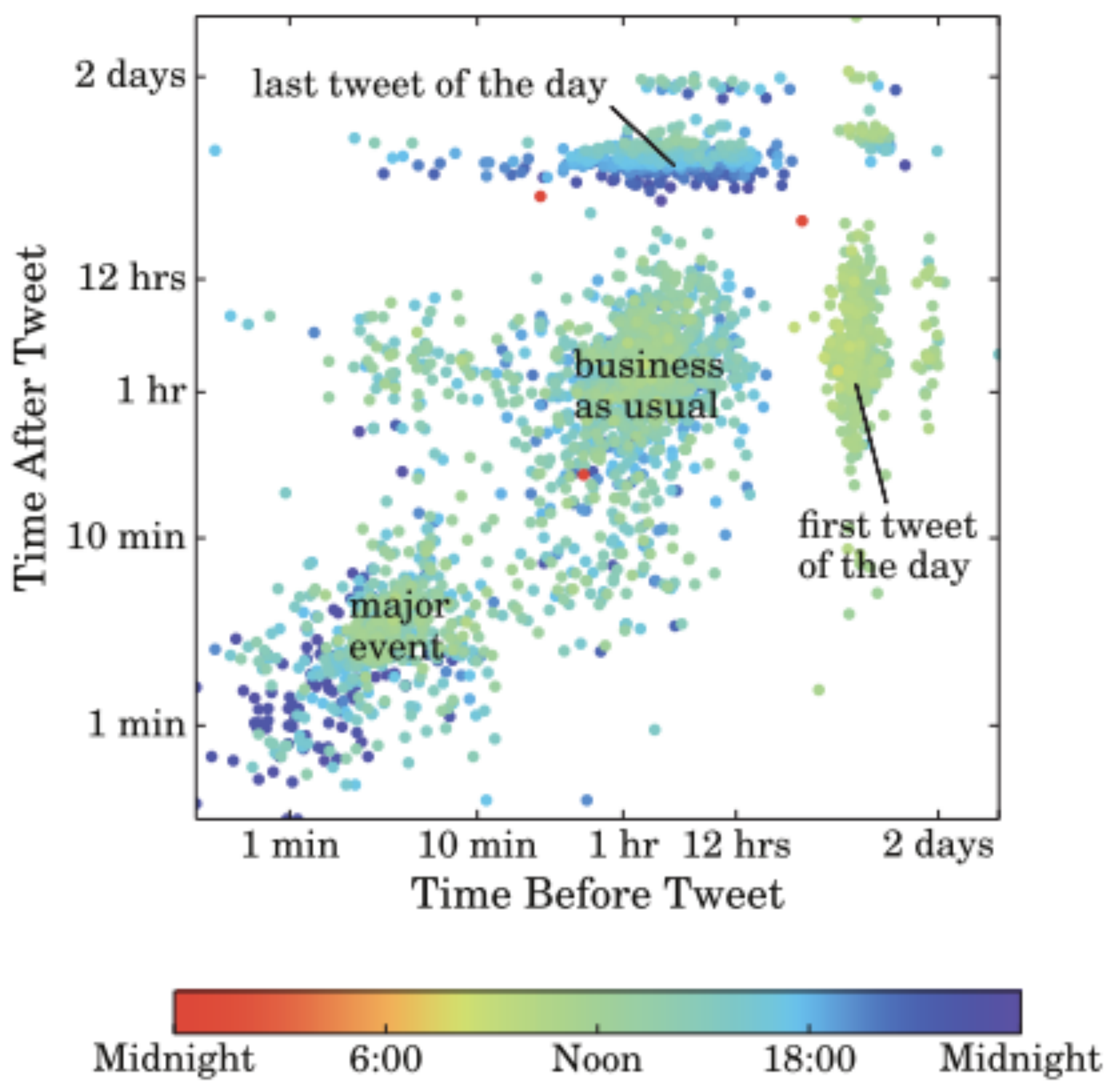
- Time map (Watson, 2015): events



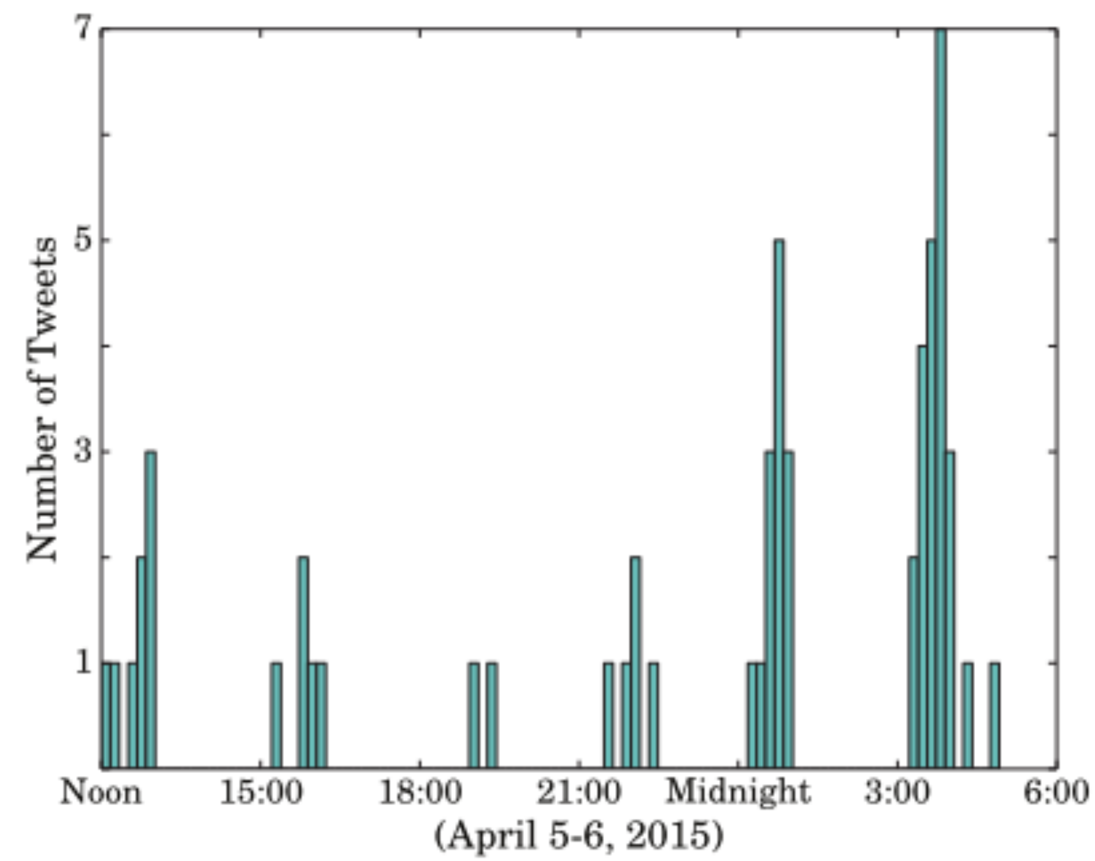




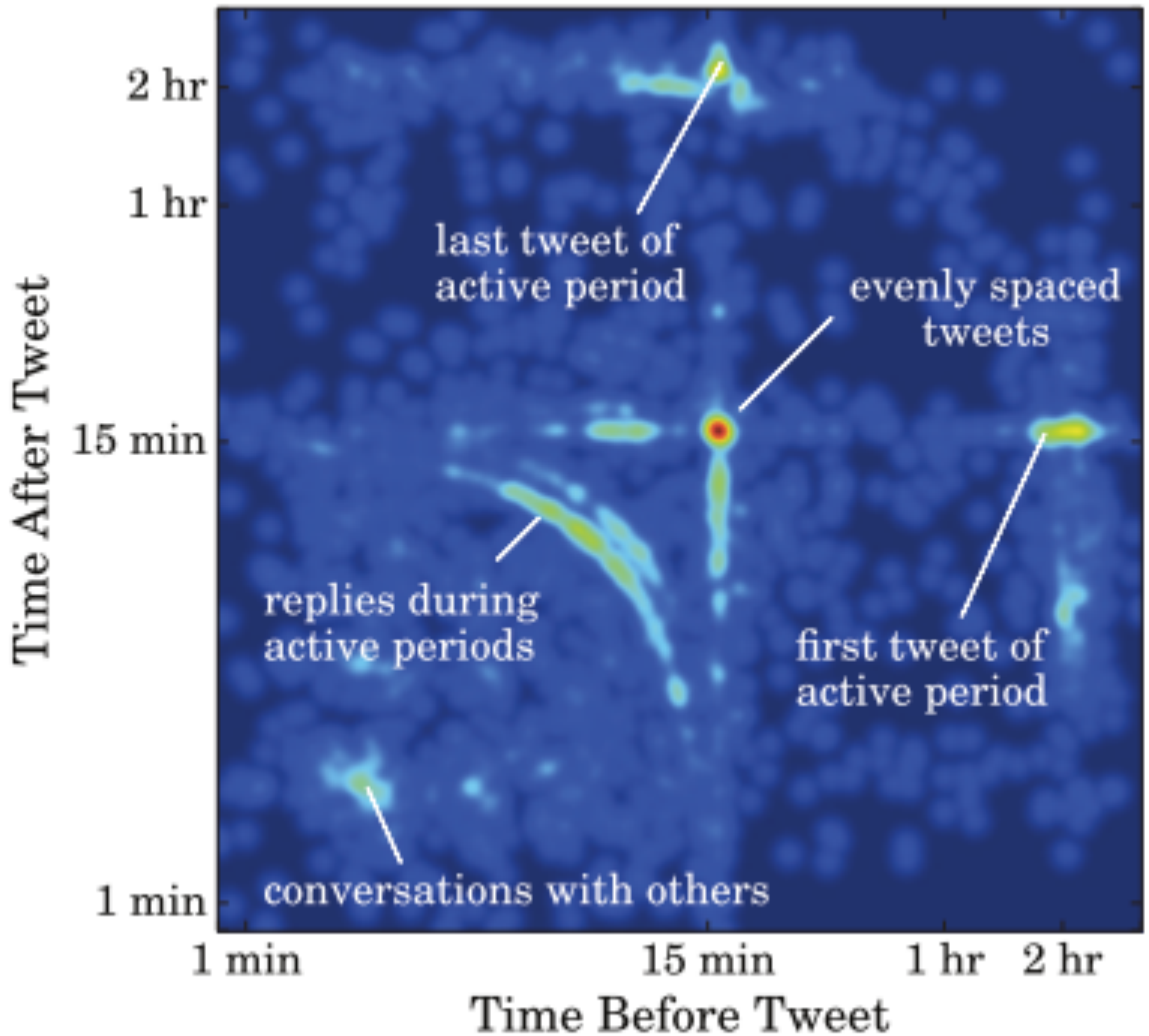




tweets by @barackobama



tweets by @oliviataters



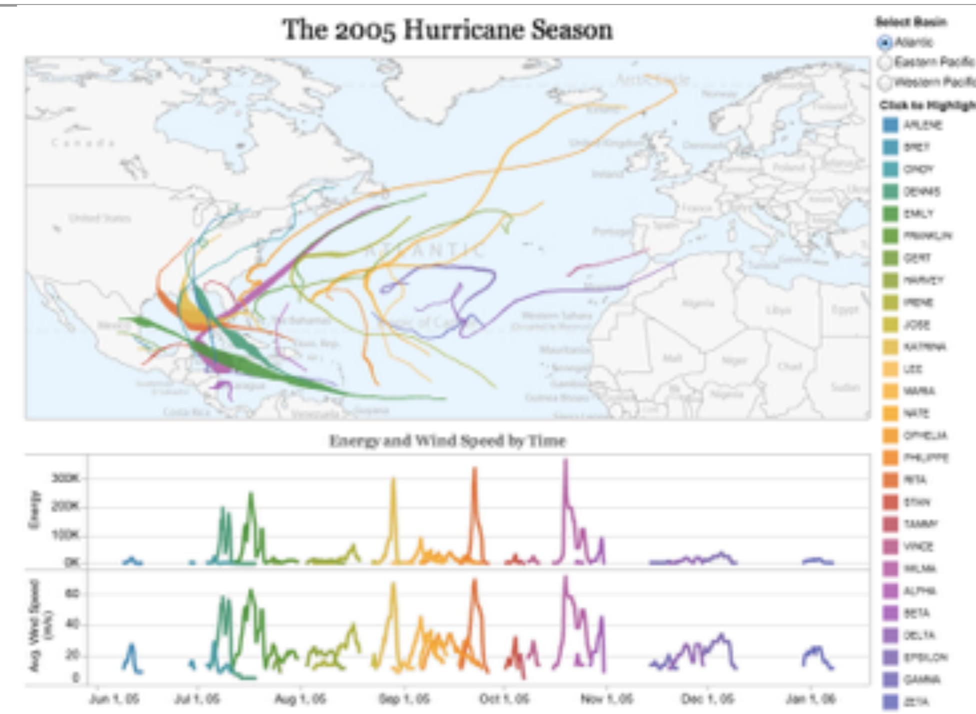
@oliviataters = bot

G. Tools of the trade

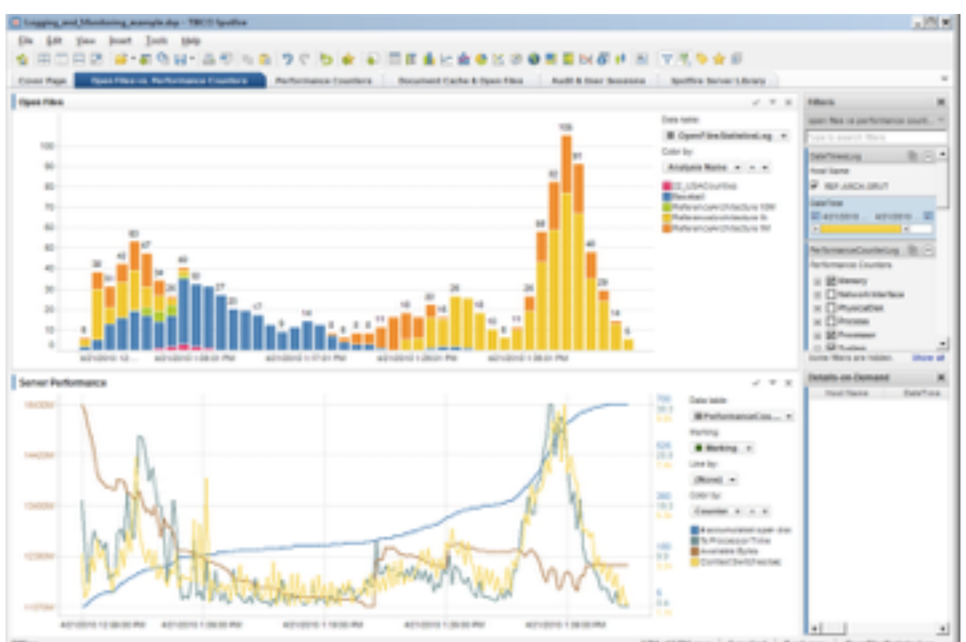
using
drawing
coding

Using data visualizations

Microsoft Excel



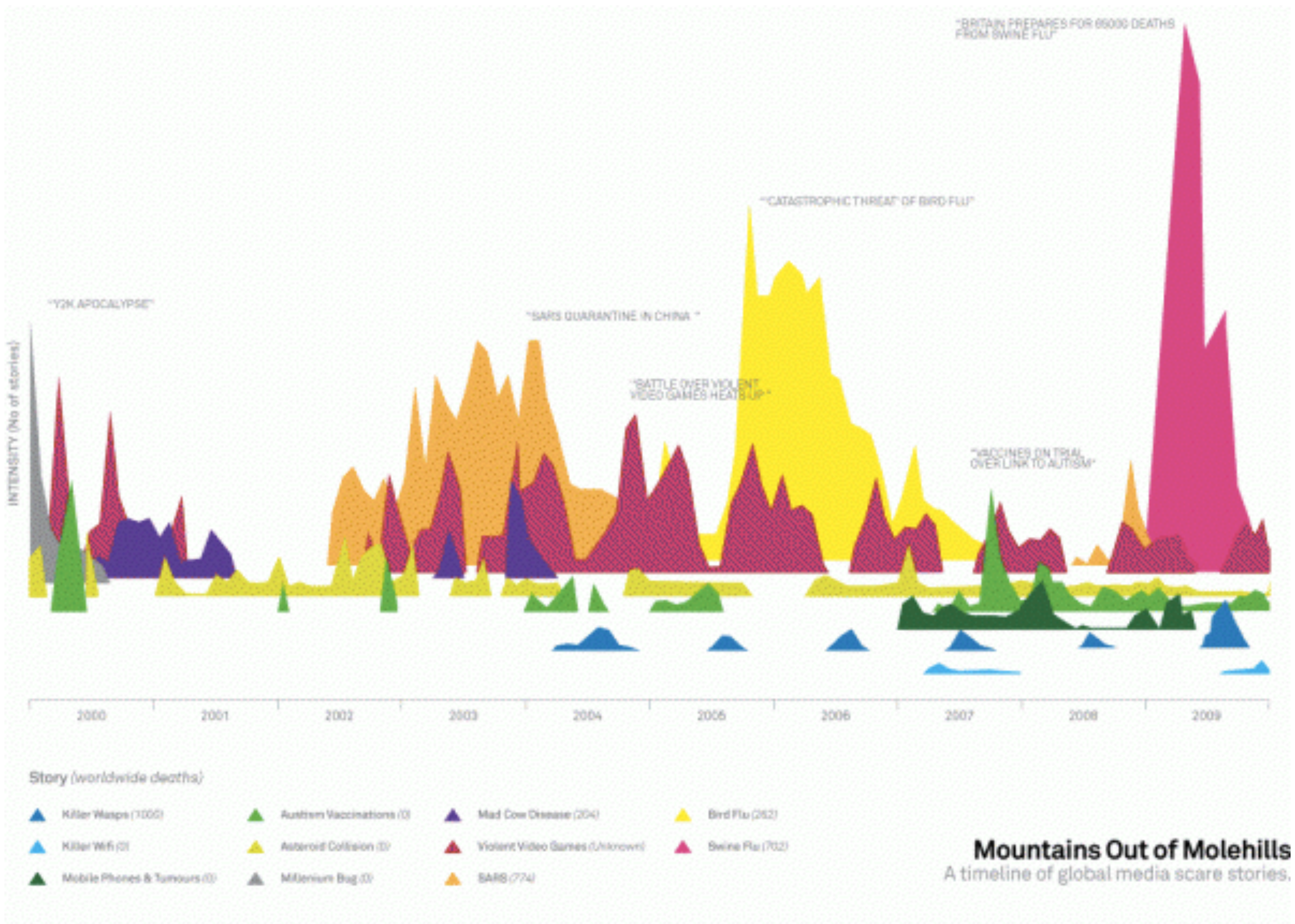
Tableau



Tibco Spotfire

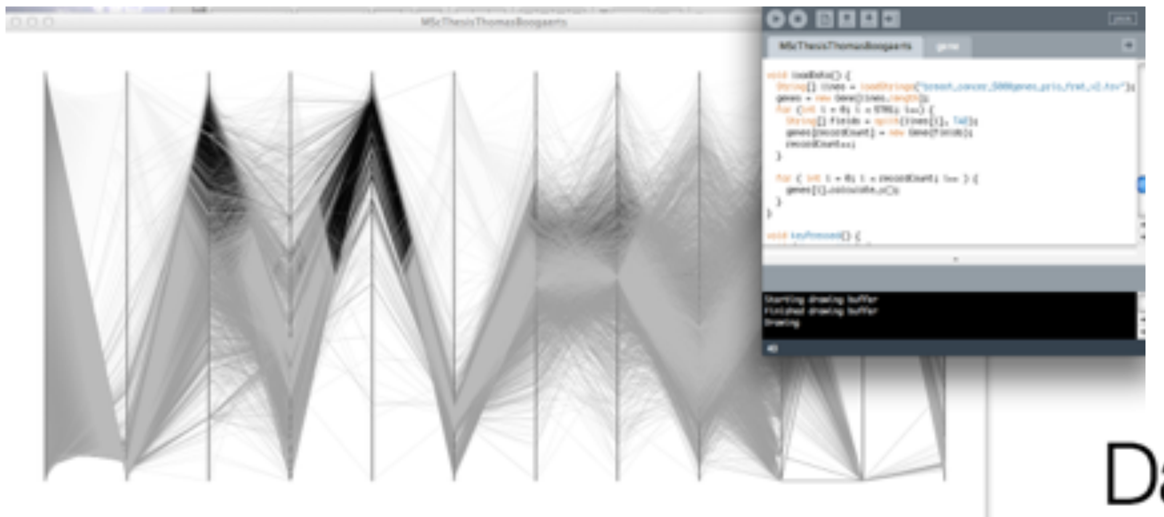
ManyEyes

Drawing data visualizations



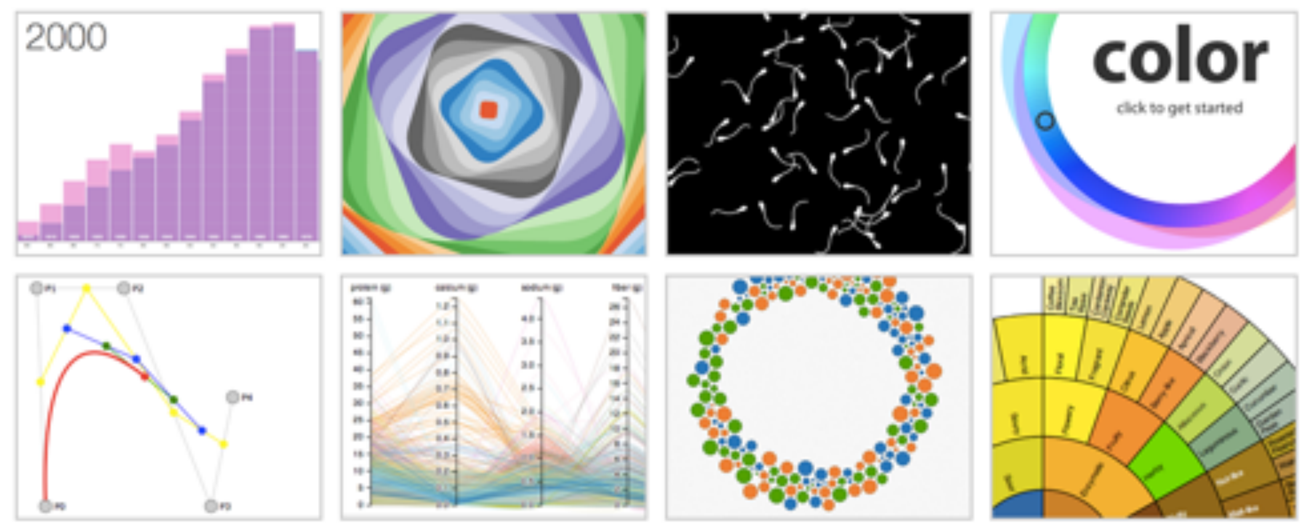
Adobe Illustrator

Coding data visualizations

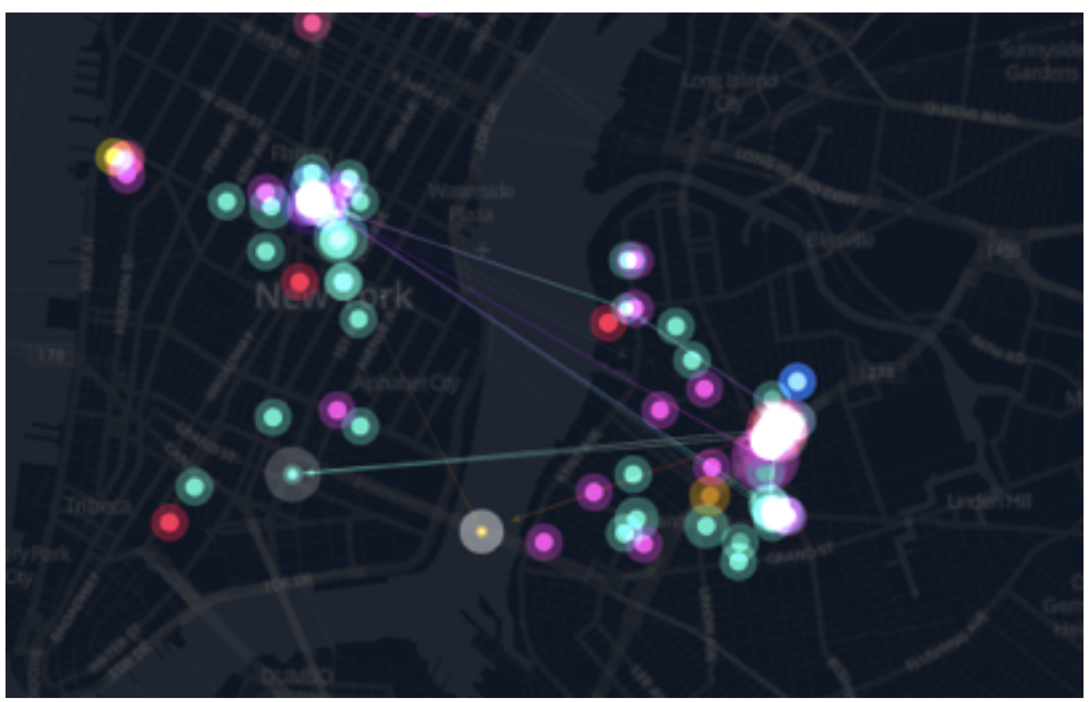


processing.org

Data-Driven Documents



d3.js



paper.js

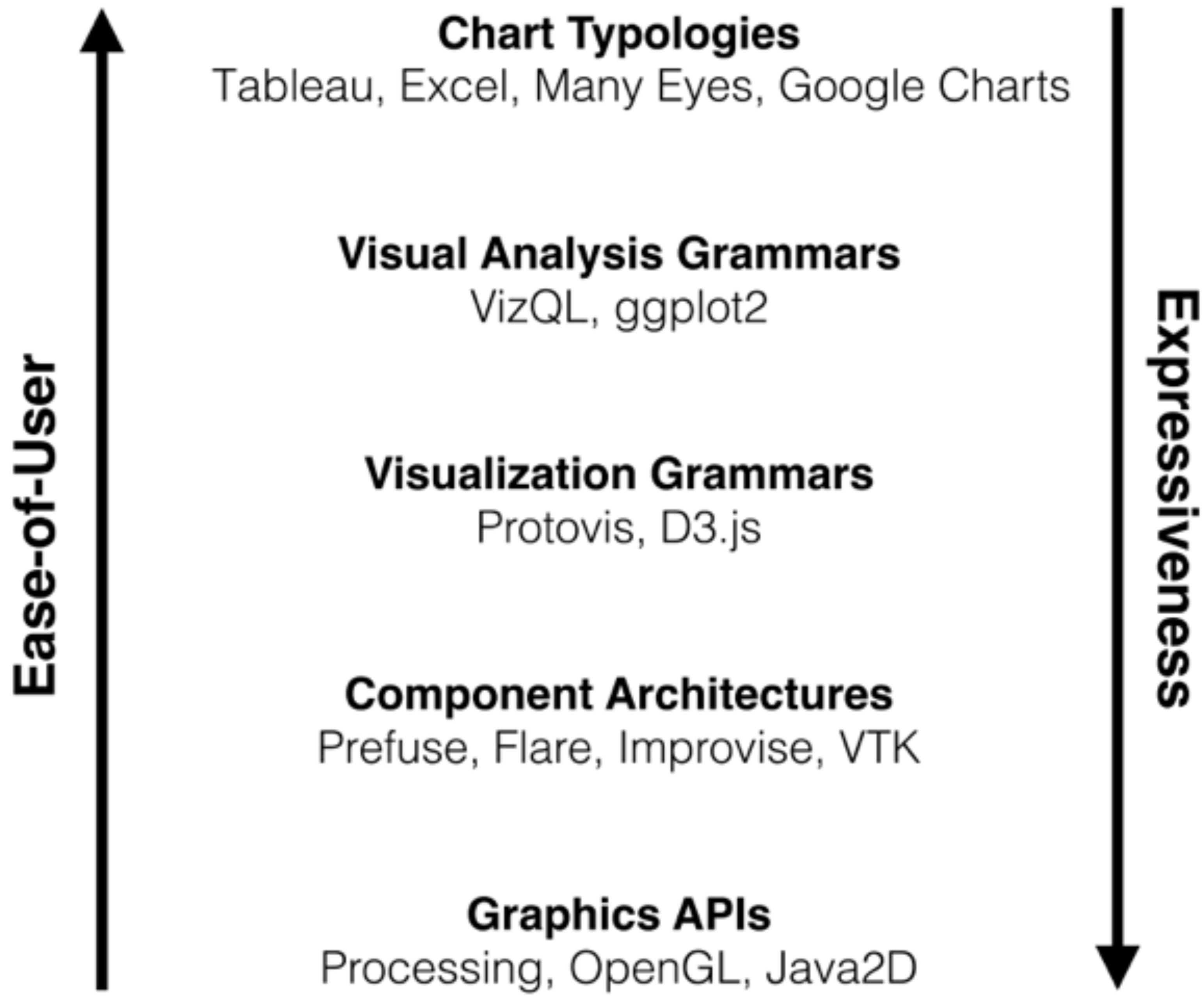


Chart Typologies
Tableau, Excel, Many Eyes, Google Charts
Visual Analysis Grammars
VizQL, ggplot2
Visualization Grammars
Protovis, D3.js
Component Architectures
Prefuse, Flare, Improvise, VTK
Graphics APIs
Processing, OpenGL, Java2D
Ease-of-User
Expressiveness


LOONEY TUNES
That's all Folks!

(…actually, we only scratched the surface…)