The Big Data to Knowledge (BD2K) Guide to the Fundamentals of Data Science

SECTION 4:

DATA VISUALIZATION

TOOLS & COMMUNICATION

NILS GEHLENBORG

Assistant Professor

HARVARD MEDICAL SCHOOL

MARCH 31, 2017

TCC BD2K Training Coordinating Center

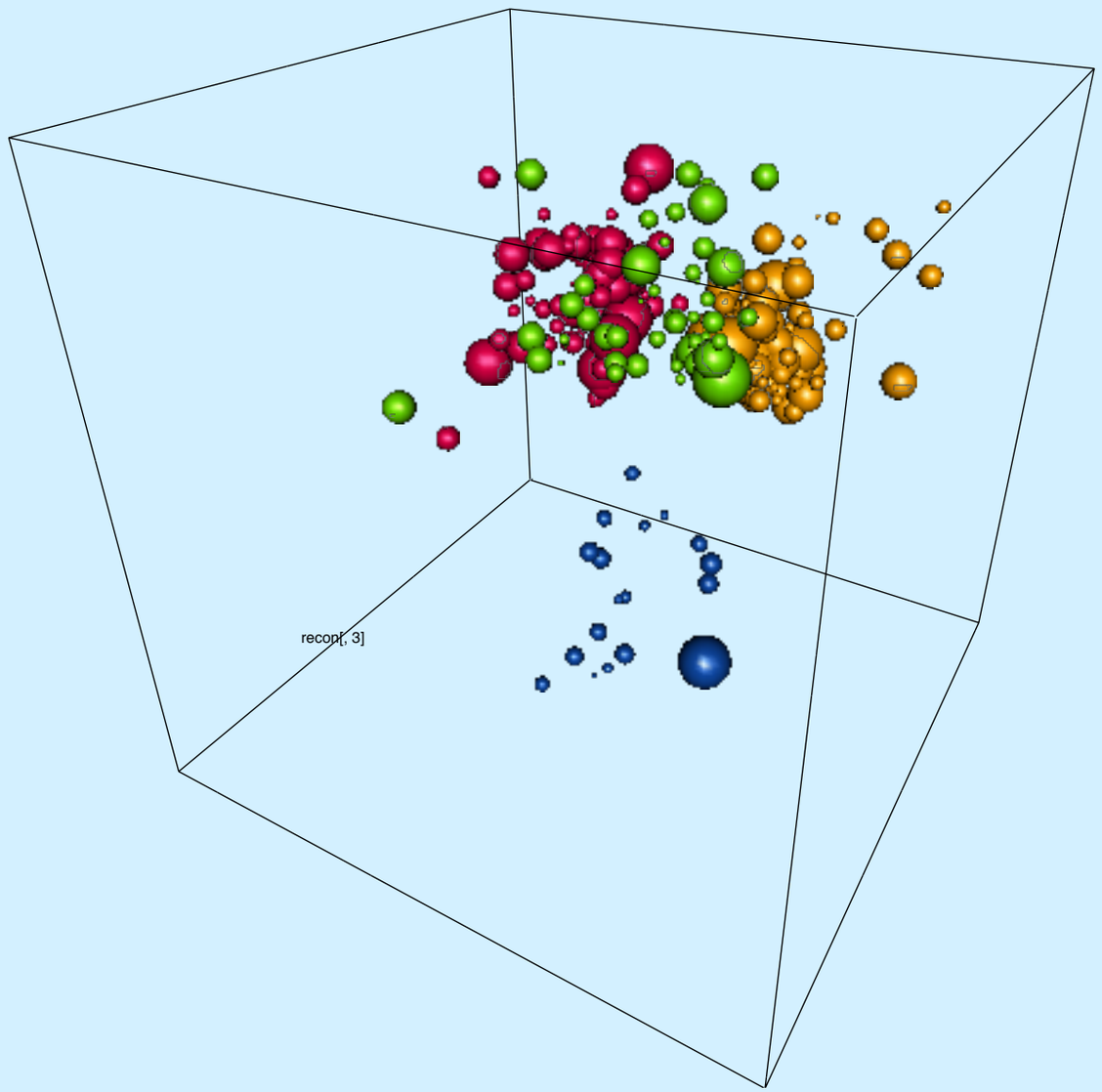
NIH Data Science at NIH

BD2K CCC

NILS GEHLENBORG

- Assistant Professor in the Department of Biomedical Informatics at Harvard Medical School.
- Research & develops novel tools to visualize 3D genome conformation data as well as heterogeneous data from large-scale cancer genomics studies.
- Co-Investigator for the 4D Nucleome Network Data Coordination & Integration Center hosted at Harvard Medical School.
- Co-founder, former general chair, & current steering committee chair of BioVis, the Symposium on Biological Data Visualization
- Co-founder of VIZBI, the annual workshop on Visualizing Biological Data.
- Co-chairs the Policy Working Group for the 4D Nucleome Network, an NIH Common Fund project.
- Served on the program committees of several international bioinformatics & data visualization conferences & held multiple editorial roles, including his current role as associate editor of BMC Bioinformatics.
- Contributed to the “Points of View” data visualization column in Nature Methods.

WARNING!



Principal Components Analysis
3D Interactive Visualization

Overview

Use Cases for Data Visualization

Encoding Data for Visual Representation

Common Mistakes

Visualization for Communication

Use Cases for Data Visualization

How does data visualization work?

Visualization uses perception to free up cognition.

http://www.uniprot.org/uniprot/P01308

How does data visualization work?

MALWMRLLPLLALLALW
GPDPAAAFVNQHLCGSHL
VEALYLVCGERGFFYTPKT
RREAEDLQVGQVELGGGP
GAGSLQPLALEGSLQKRGI
VEQCCTSICSLYQLENYCN

http://www.uniprot.org/uniprot/P01308

How does data visualization work?

MALWMRLLPLLALLALW
GPDPAAAAFVNQHLCGSHL
VEALYLVCGERGFFYTPKT
RREAEDLQVGQVELGGGP
GAGSLQPLALEGSLQKRGI
VEQCCTSICSLYQLENYCN

How does data visualization work?

Visualization uses perception to free up cognition.

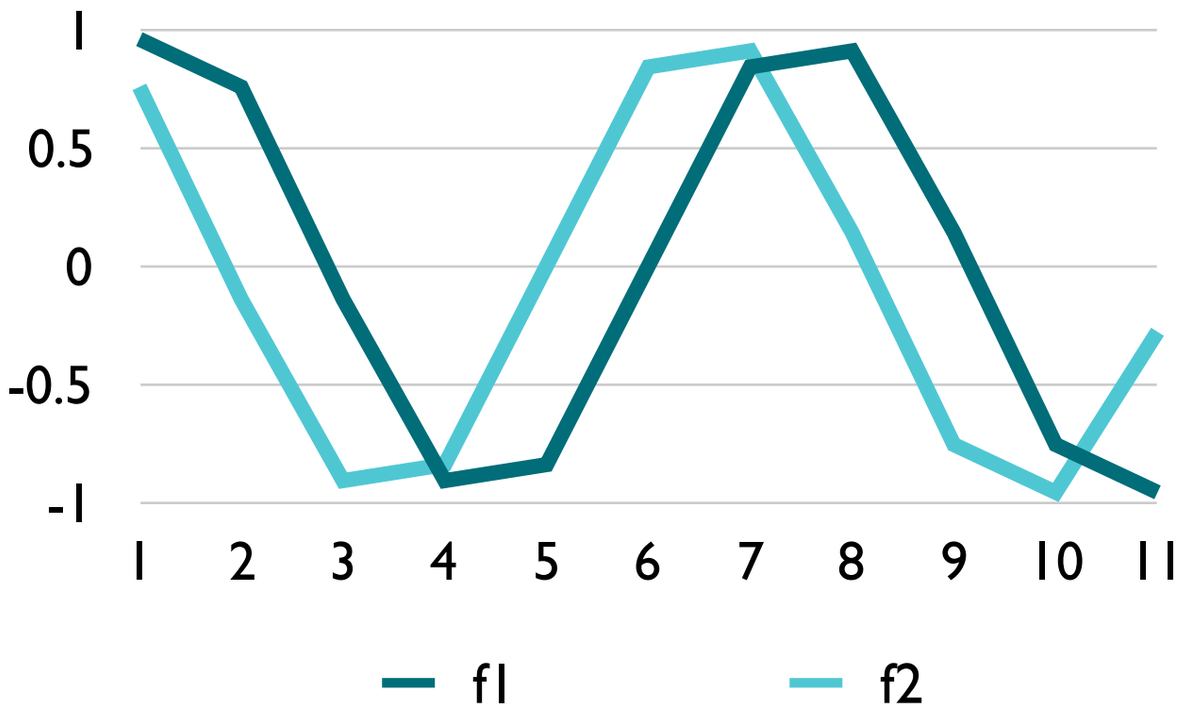
Visualization is an external cognitive aid
and augments working memory.

How does data visualization work?

$$
\begin{array}{r}
\mathbf{453 \times 862 =} \\
\hline
906 \\
+\ 27{,}180 \\
+\ 362{,}400 \\
\hline
390{,}486
\end{array}
$$

How does data visualization work?

x	f1	f2
1.00	0.96	0.76
2.00	0.76	-0.14
3.00	-0.14	-0.91
4.00	-0.91	-0.84
5.00	-0.84	0.00
6.00	0.00	0.84
7.00	0.84	0.91
8.00	0.91	0.14
9.00	0.14	-0.76
10.00	-0.76	-0.96
11.00	-0.96	-0.28



How does data visualization work?

Visualization uses perception to free up cognition.

Visualization is an external cognitive aid
and augments working memory.

Why data visualization?

I believe it when I see it.

— *Unknown*

Anscombe 1973, The American Statistician

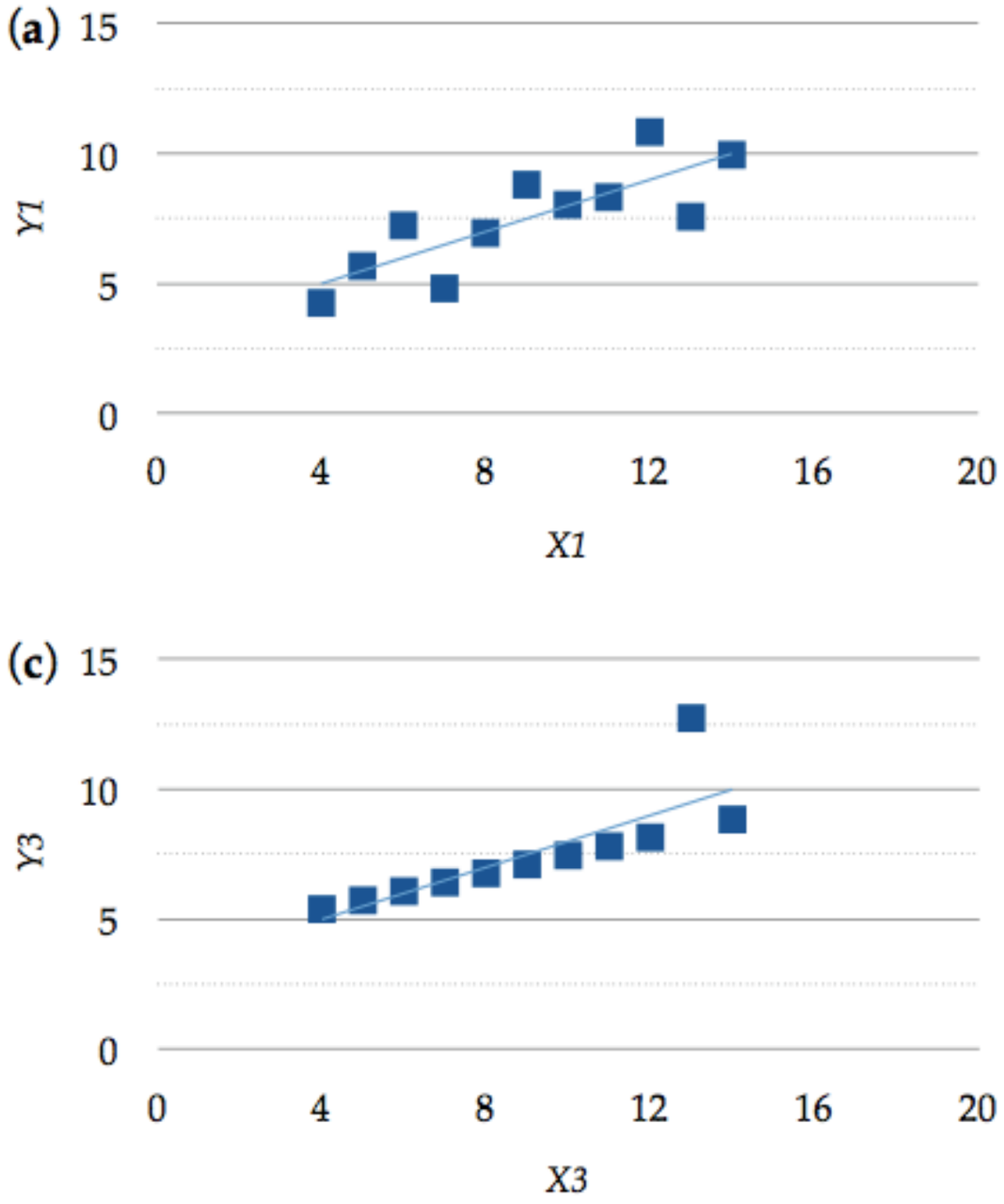
Anscombe's Quartet

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
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	6.89

mean(X) = **9**, var(X) = **11**, mean(Y) = **7.5**, var(Y) = **4.12**,
cor(X,Y) = **0.816**, linear regression line Y = **3 + 0.5*X**

Anscombe 1973, The American Statistician

Anscombe's Quartet

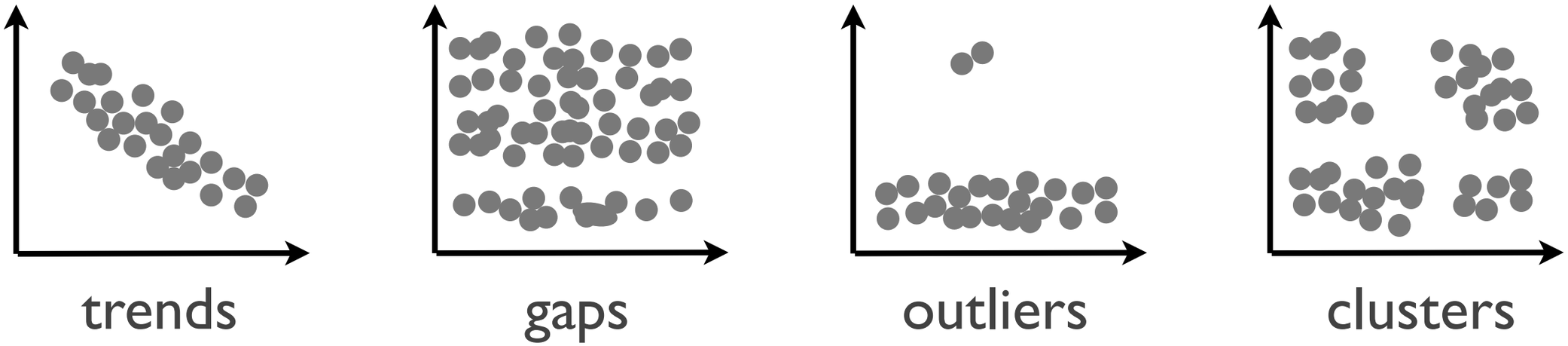


Why data visualization?

I'm wondering if there are any interesting patterns in my data.

— *Almost Everyone*

Exploration: Hypothesis Generation

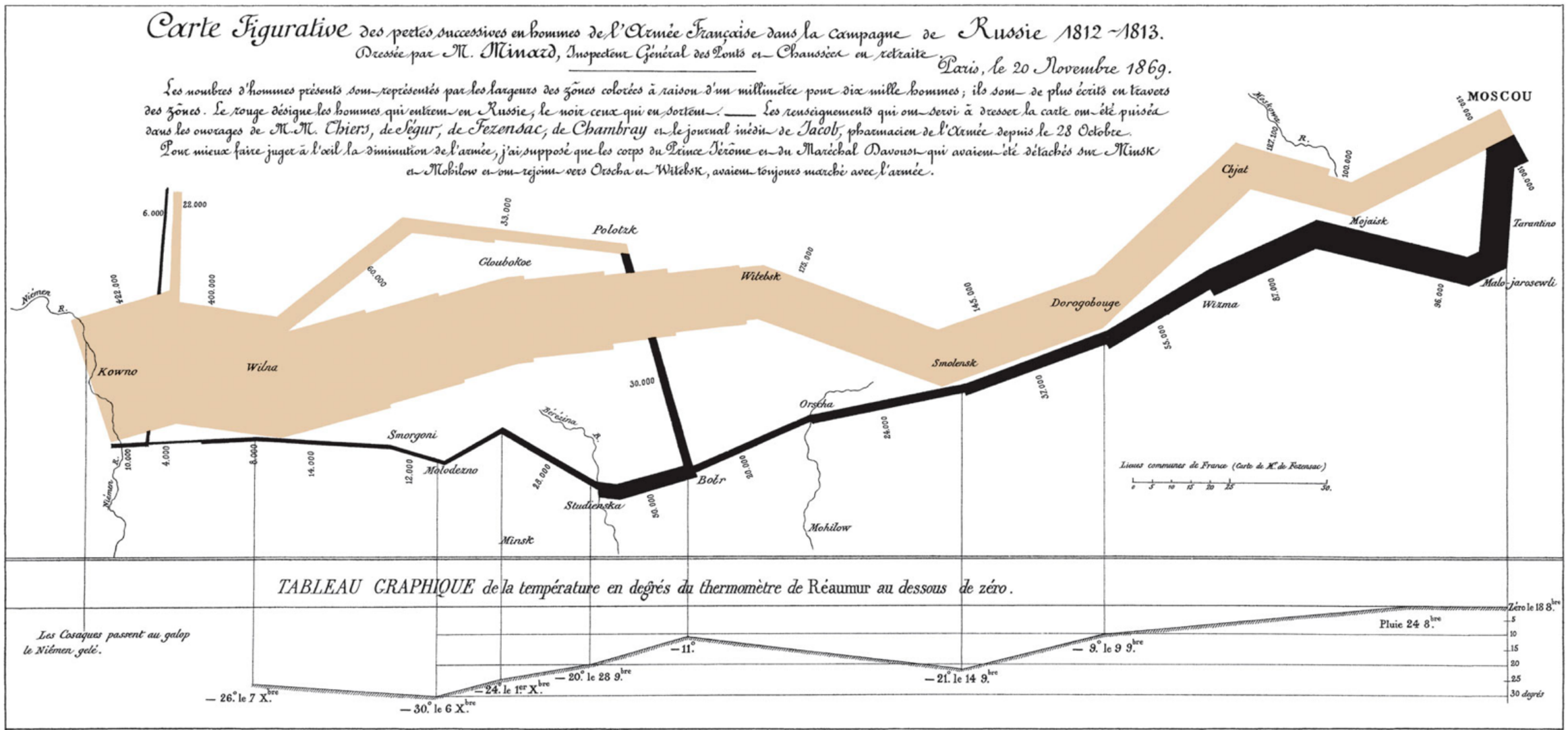


Why data visualization?

A good sketch is better than a long speech.

— *Napoleon Bonaparte*

Napoleon's March on Moscow



Visualization Use Cases

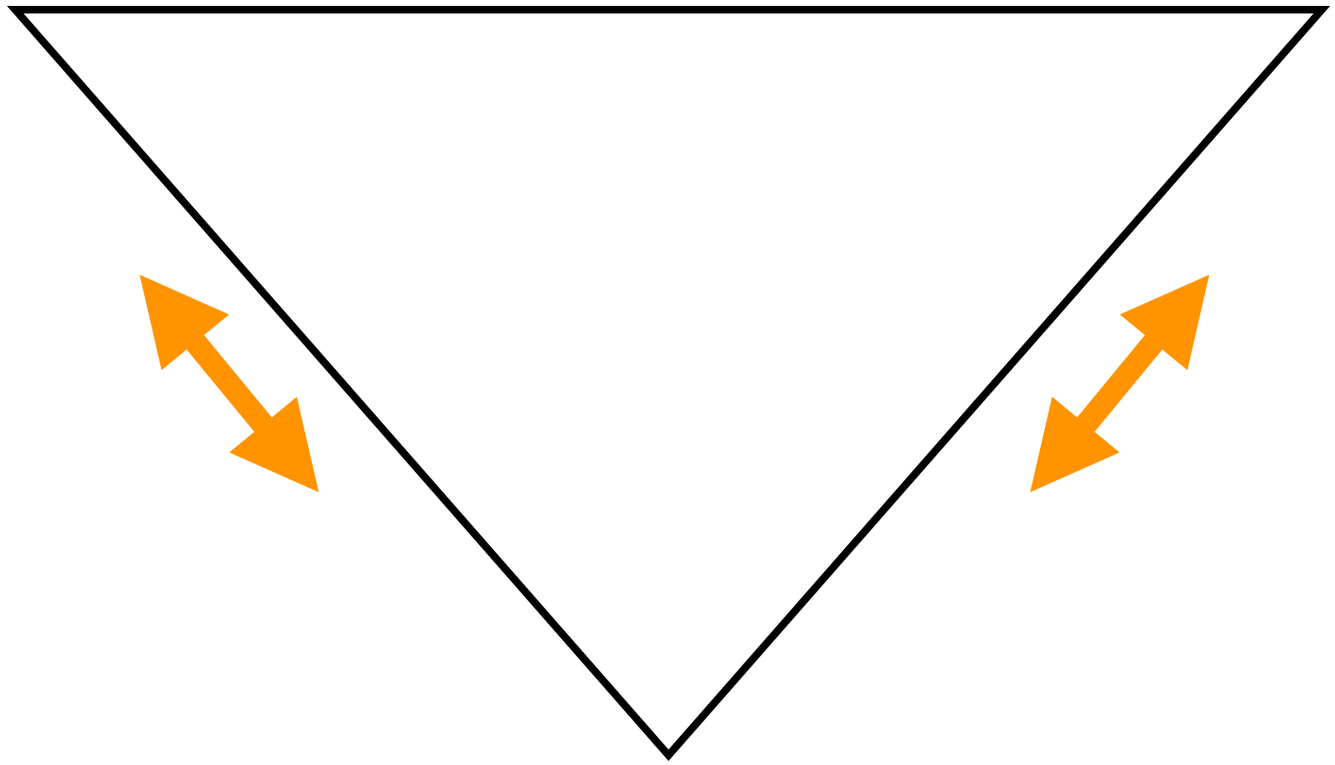
Confirmation

Exploration

Presentation

dialogue between computer & analyst

Confirmation ⟷ Exploration



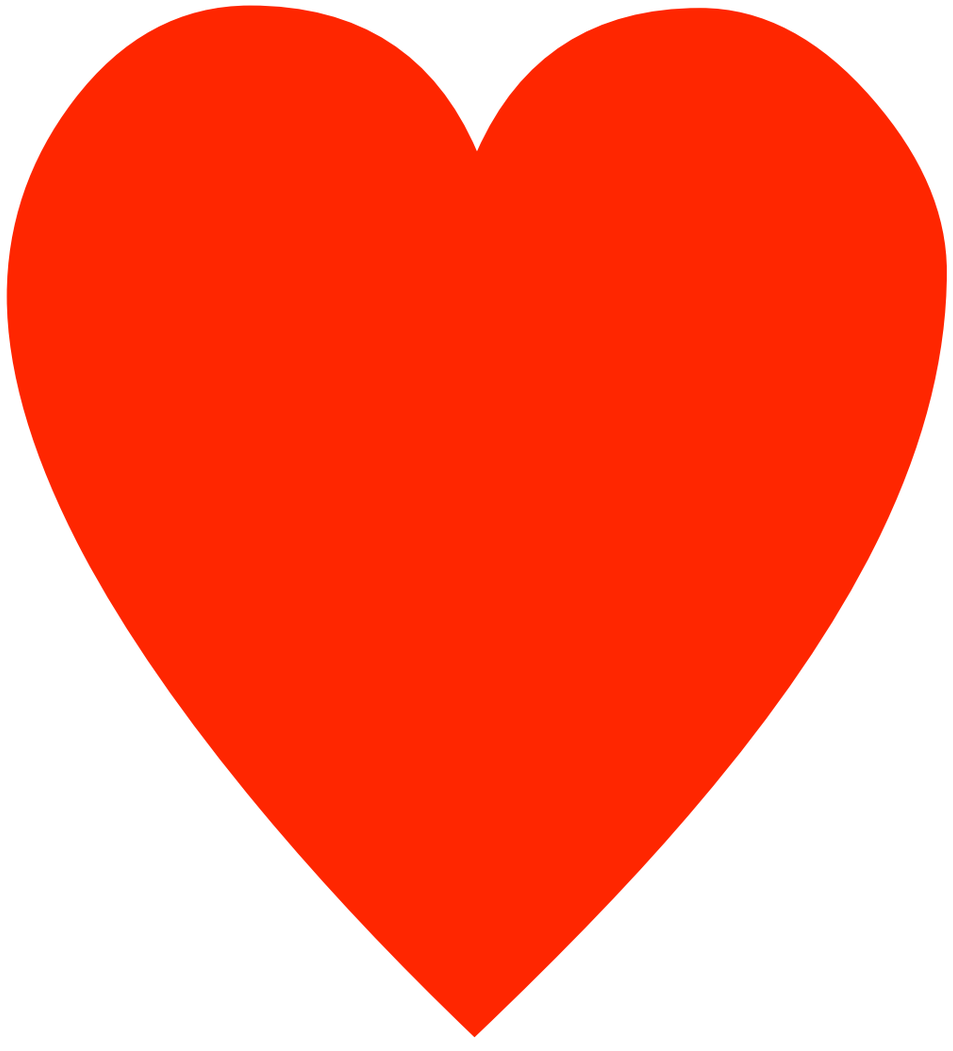
Presentation

dialogue between analyst & audience

Encoding Data for Visual Representation

Borkin et al., Evaluation of Artery Visualizations for Heart Disease Diagnosis, 2011

When good visualization really matters



Borkin et al., Evaluation of Artery Visualizations for Heart Disease Diagnosis, 2011

Why we should ♥ good visualizations

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 17, NO. 12, DECEMBER 2011

Evaluation of Artery Visualizations for Heart Disease Diagnosis

Michelle A. Borkin, *Student Member, IEEE*, Krzysztof Z. Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank J. Rybicki, Charles L. Feldman, and Hanspeter Pfister, *Senior Member, IEEE*

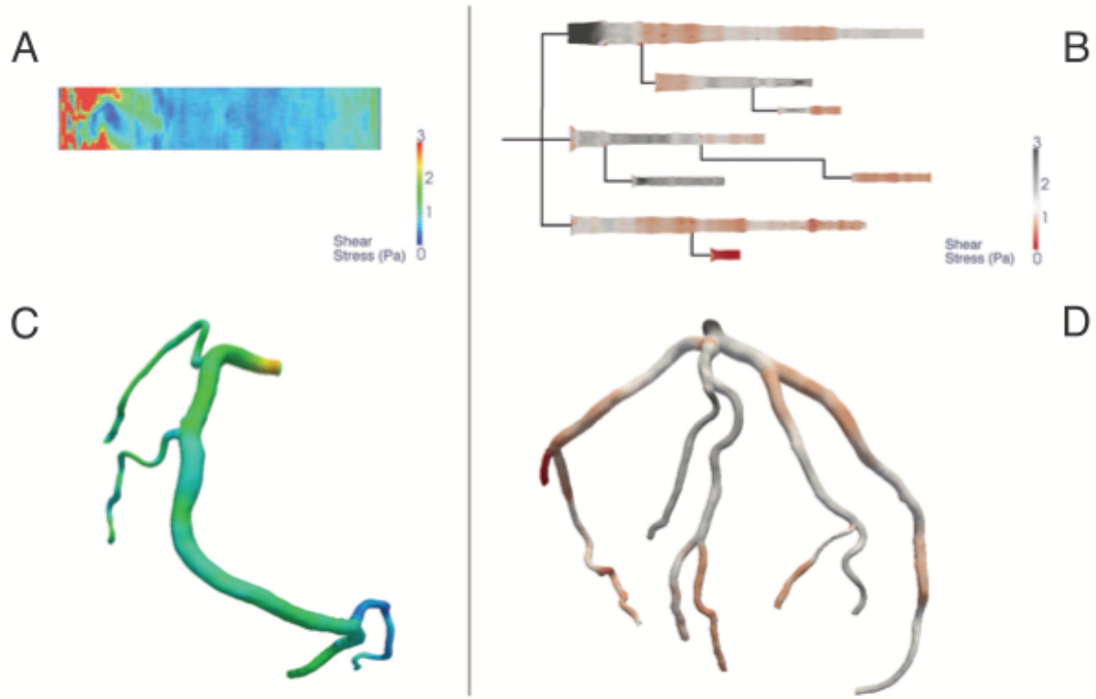


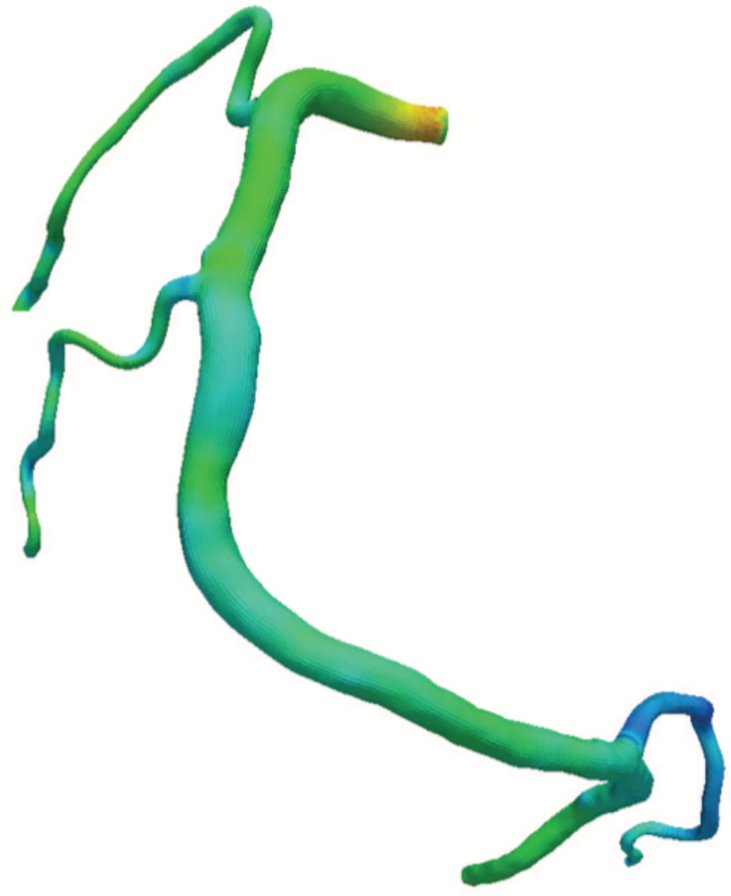
Fig. 1. Left: Traditional 2D projection (A) of a single artery, and 3D representation (C) of a right coronary artery tree with a rainbow color map. Right: 2D tree diagram representation (B) and equivalent 3D representation (D) of a left coronary artery tree with a diverging color map.

Borkin et al., Evaluation of Artery Visualizations for Heart Disease Diagnosis, 2011

Why we should ♥ good visualizations

Traditional 2D projection of a single artery

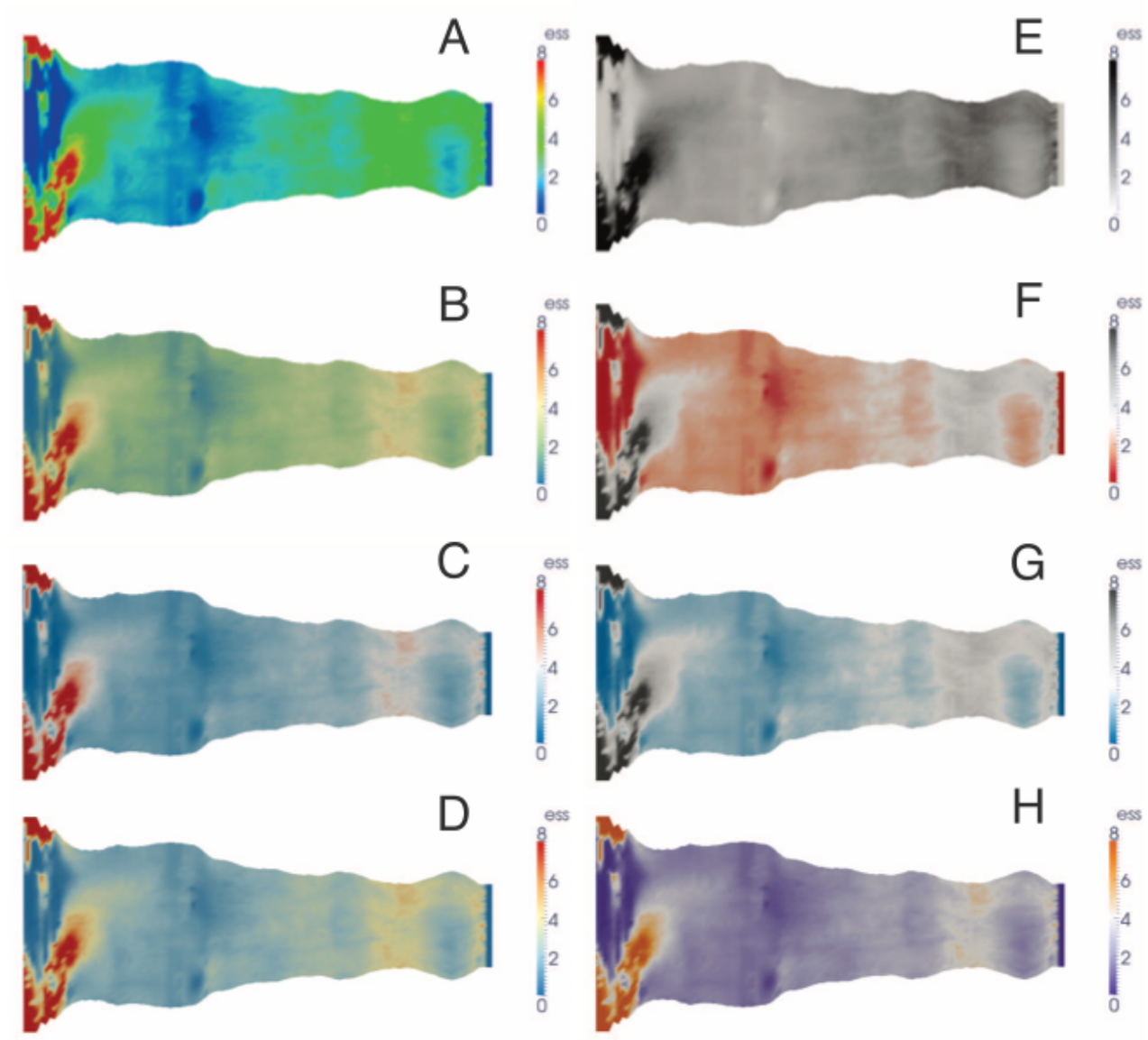
Traditional 3D projection of right artery tree



Borkin et al., Evaluation of Artery Visualizations for Heart Disease Diagnosis, 2011

Color Maps

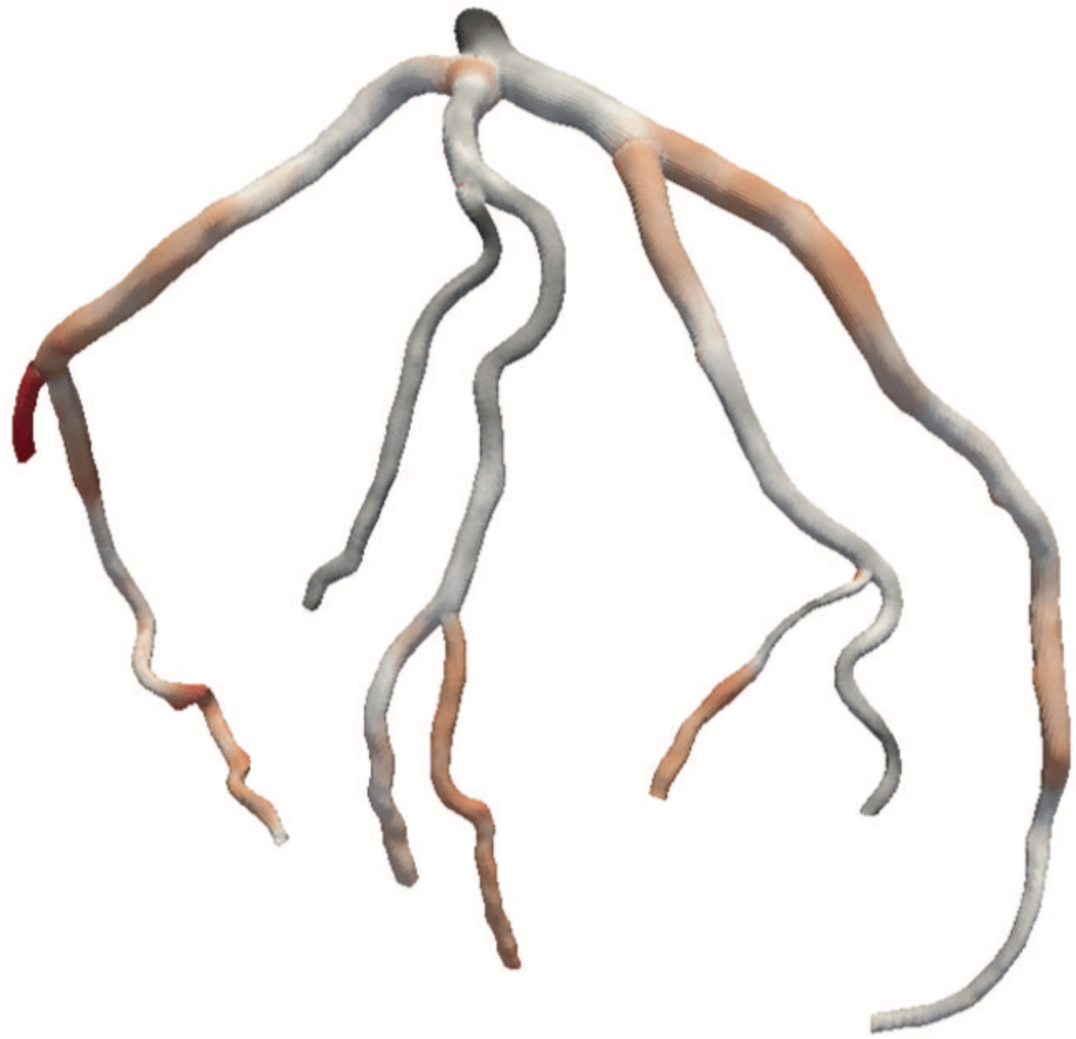
1st choice



2nd choice

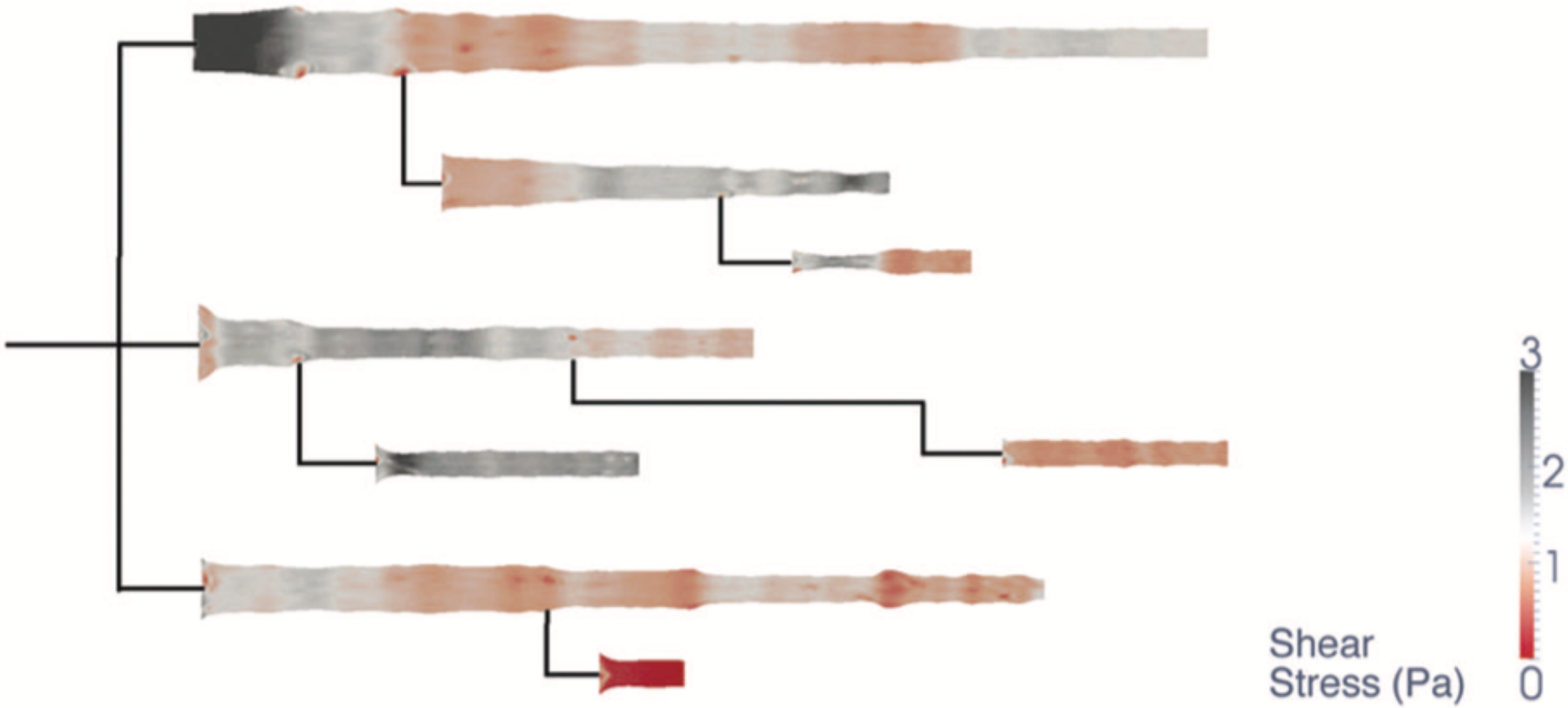
Projections

3D projection of left artery tree with diverging color map

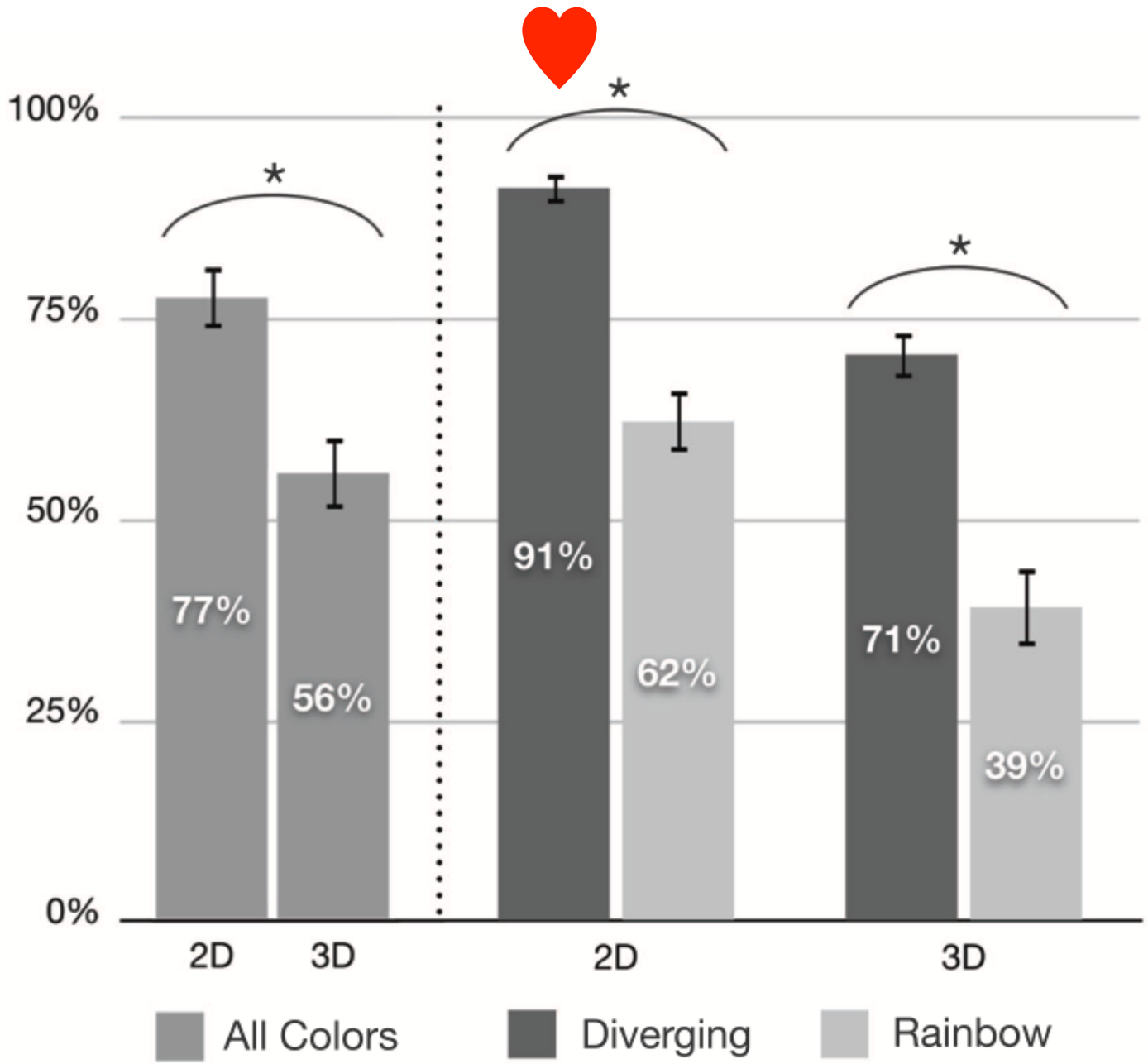


Projections

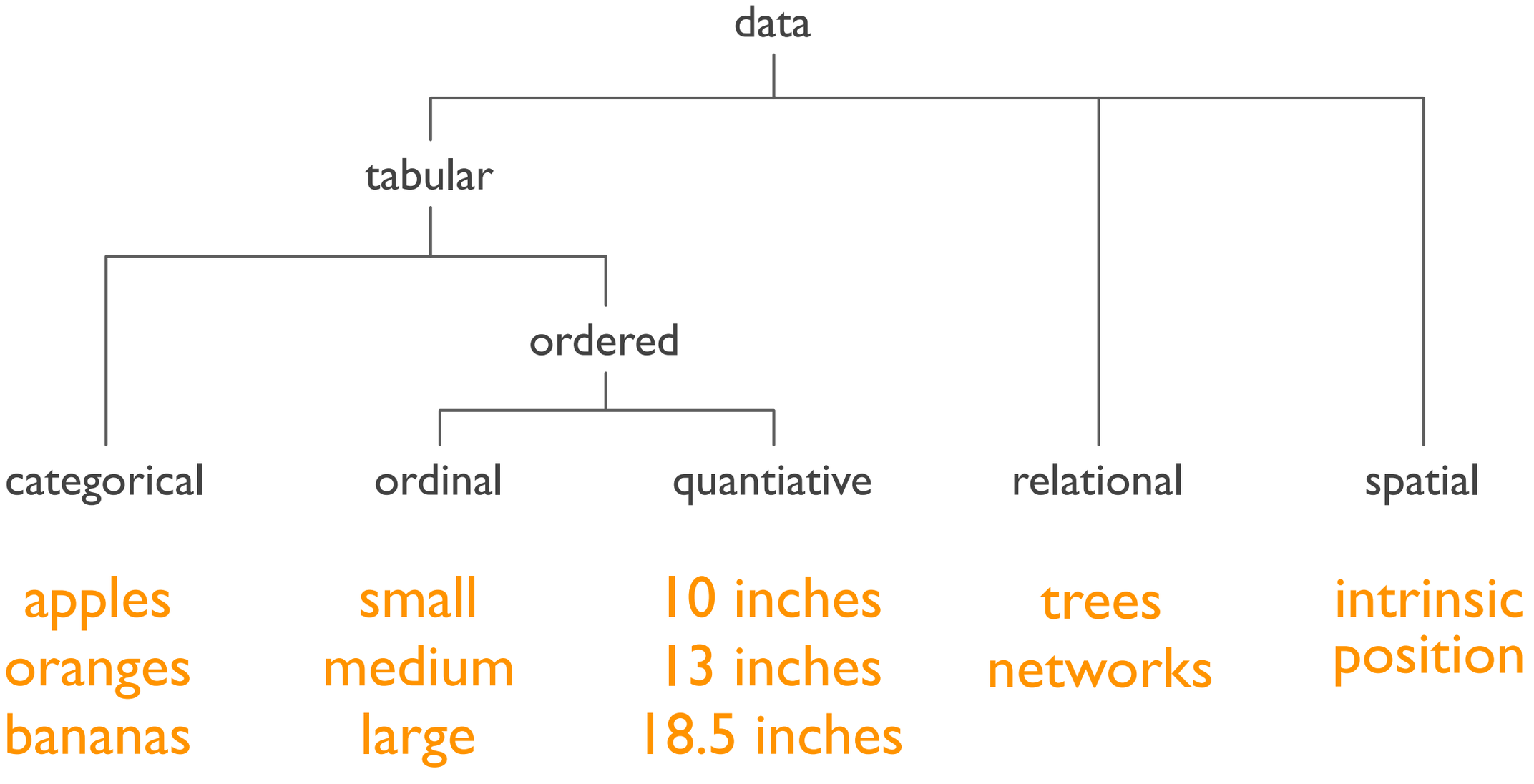
Novel 2D projection of left artery tree



Findings: Percentage of low ESS areas identified



Visual Encoding of Data



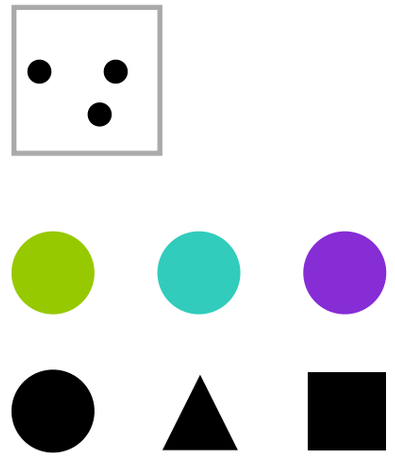
Visual Channels: Rankings

Categorical
What? Where?

position*
planar

color hue

shape



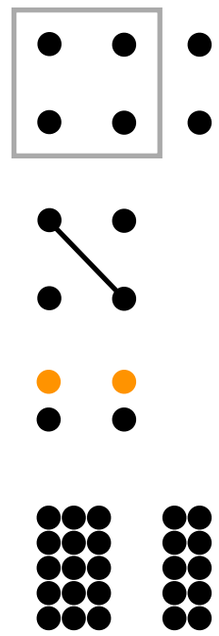
Relational
With whom?

containment

connection

similarity

position*
proximity



Visual Channels: Rankings

Ordinal & Quantitative

How much?

position*
common scale

position*
unaligned scale

length (1D)

angle/tilt

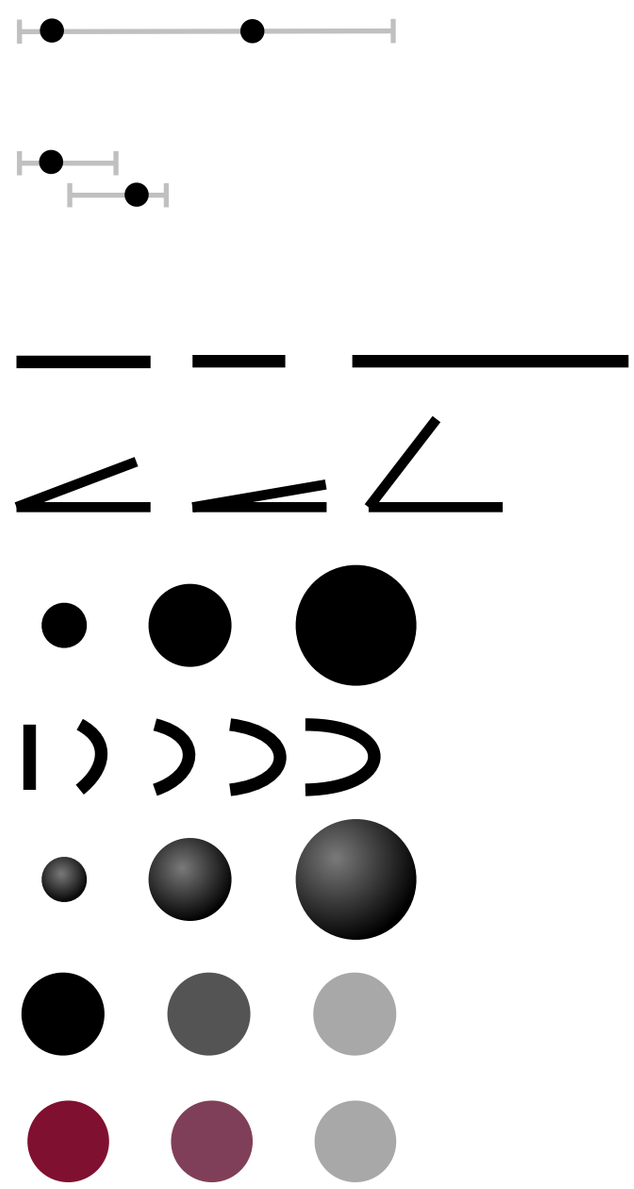
area (2D)

curvature

volume (3D)

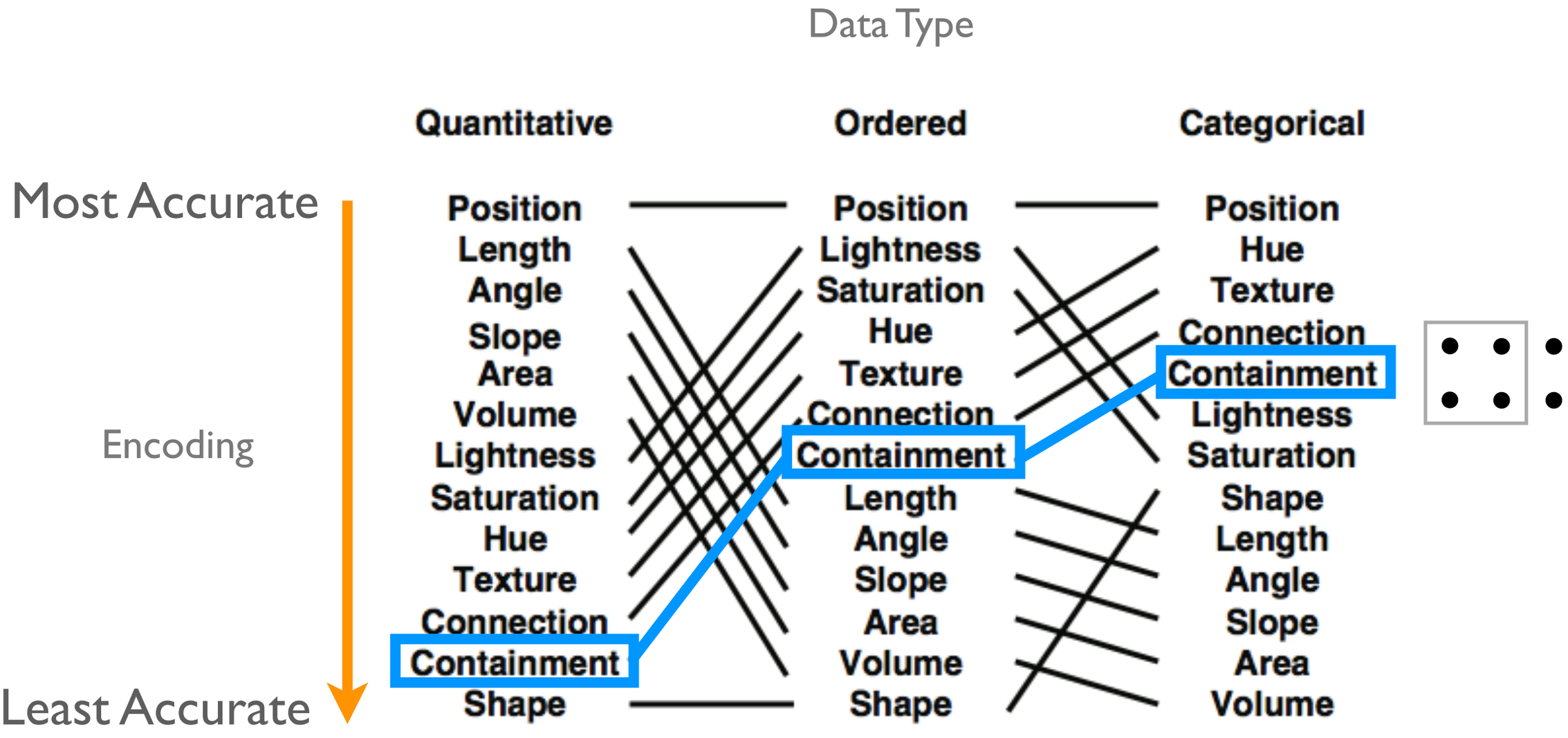
lightness

color saturation



Munzner 2009, in *Fundamentals of Computer Graphics (redrawn from Mackinlay 1986)*

Ranking of Encodings



F Xu et al., Nature Communications 3, 1258 doi:10.1038/ncomms2256

What's wrong? What's right?

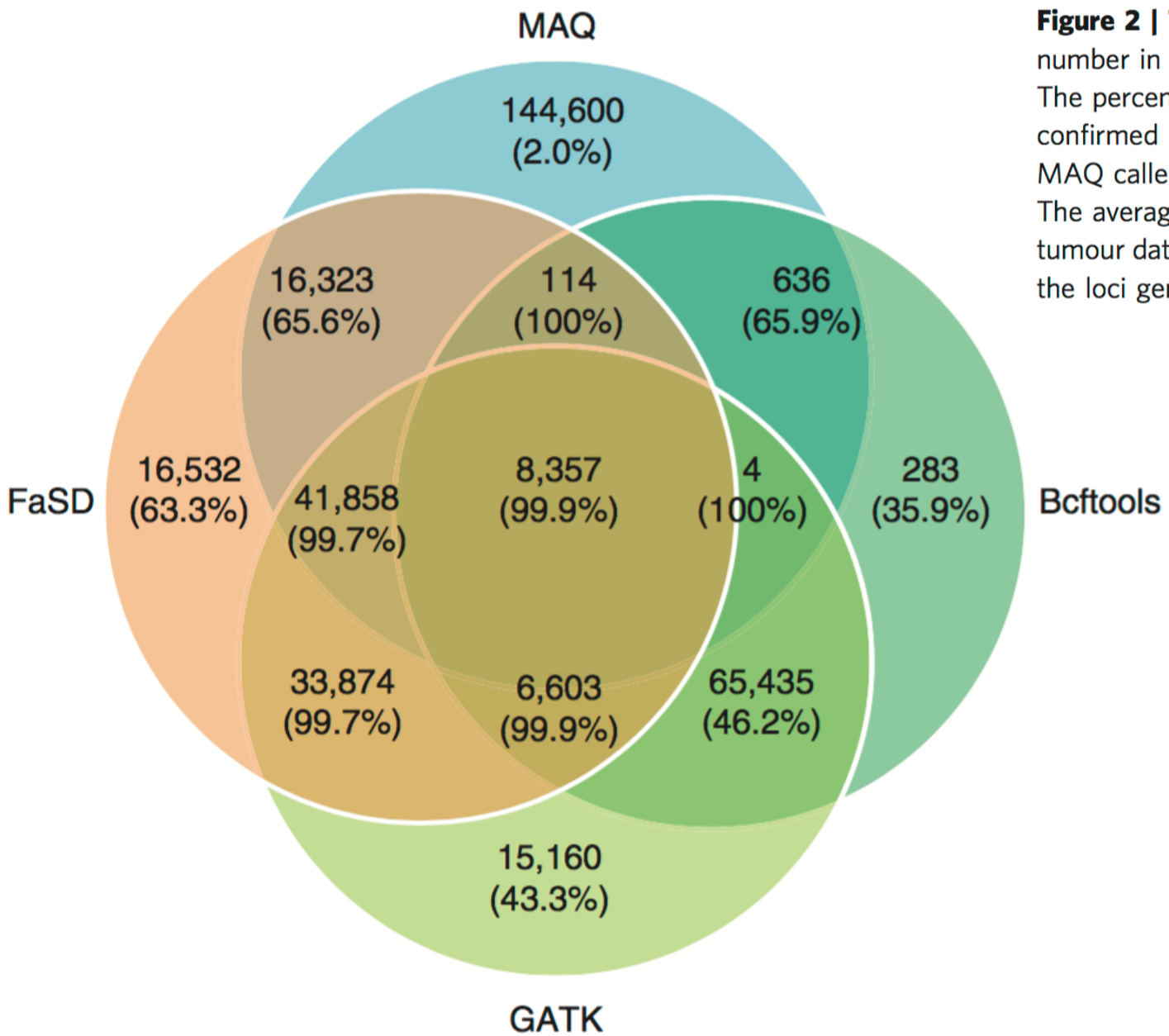


Figure 2 | The Venn diagram of SNPs detected by different tools. The number in each cell is the number of SNPs in the corresponding category. The percentage under the number is the proportion of SNPs that were confirmed by the Affymetrix SNP array. The FaSD, GATK, Bcftools and MAQ called 123661, 171291, 81432 and 211892 SNPs in total, respectively. The average depth of this data set was 10 ×. The figure is based on the tumour data set and Bowtie was used as aligner, and statistics are based on the loci genotyped by the Affymetrix SNP array.

F Xu et al., Nature Communications 3, 1258 doi:10.1038/ncomms2256

What's wrong? What's right?

Euler Diagram

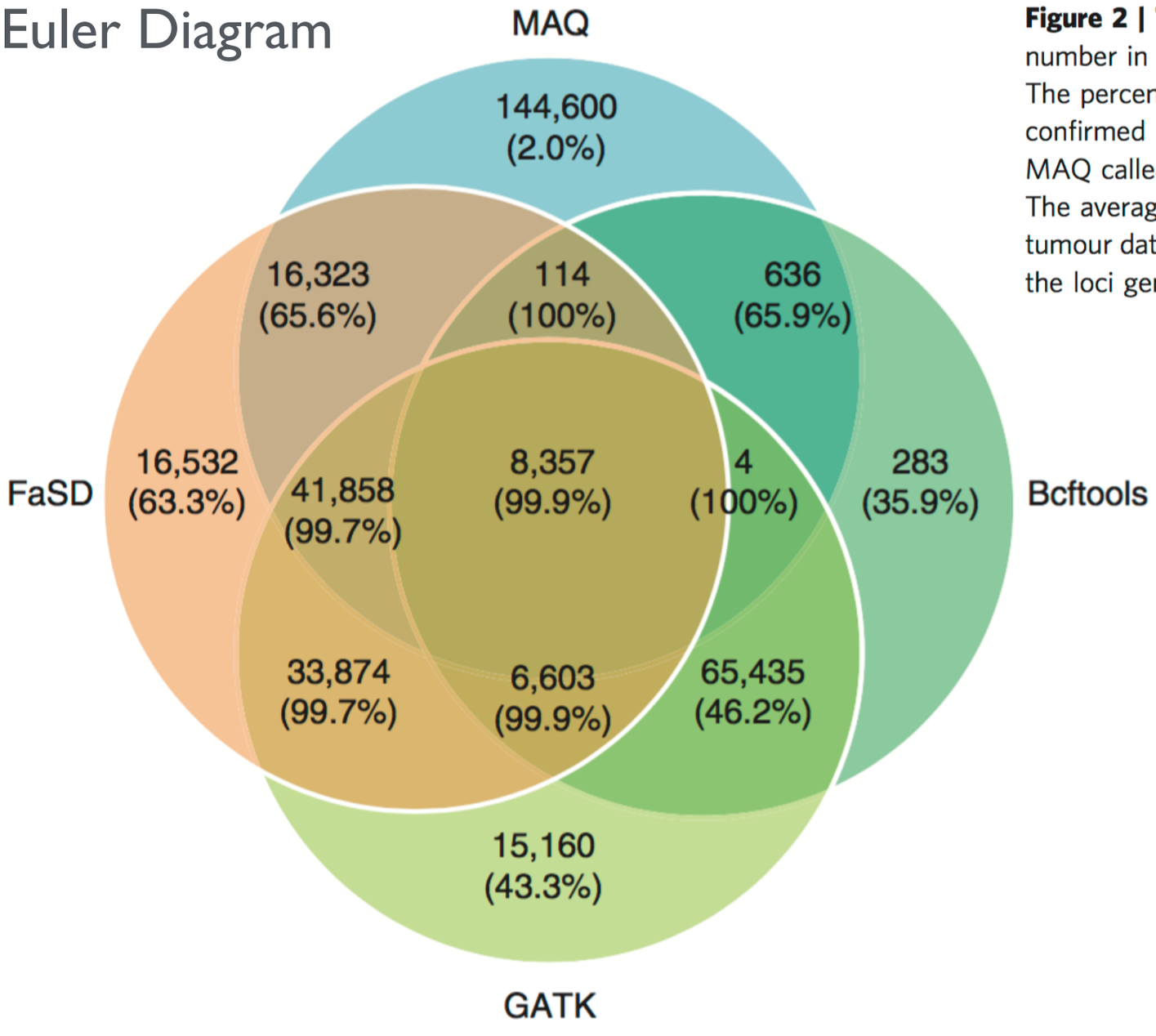
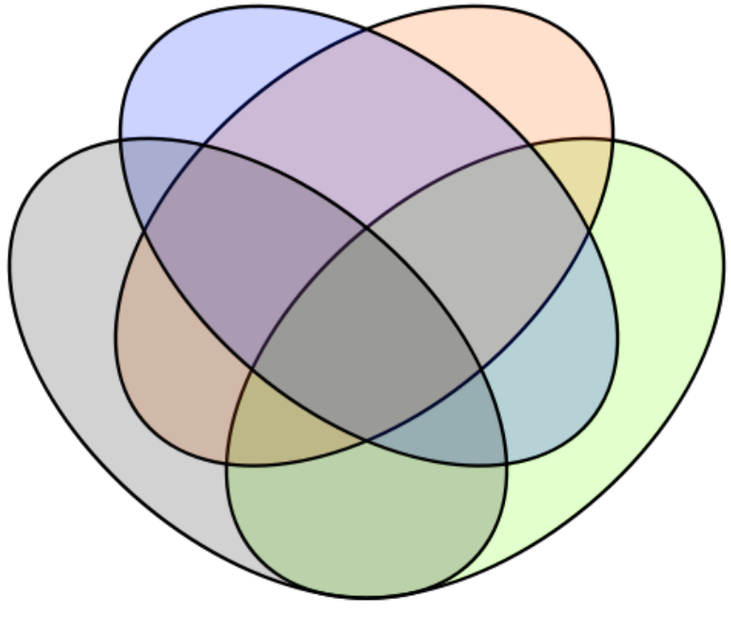


Figure 2 | The Venn diagram of SNPs detected by different tools. The number in each cell is the number of SNPs in the corresponding category. The percentage under the number is the proportion of SNPs that were confirmed by the Affymetrix SNP array. The FaSD, GATK, Bcftools and MAQ called 123661, 171291, 81432 and 211892 SNPs in total, respectively. The average depth of this data set was 10 ×. The figure is based on the tumour data set and Bowtie was used as aligner, and statistics are based on the loci genotyped by the Affymetrix SNP array.

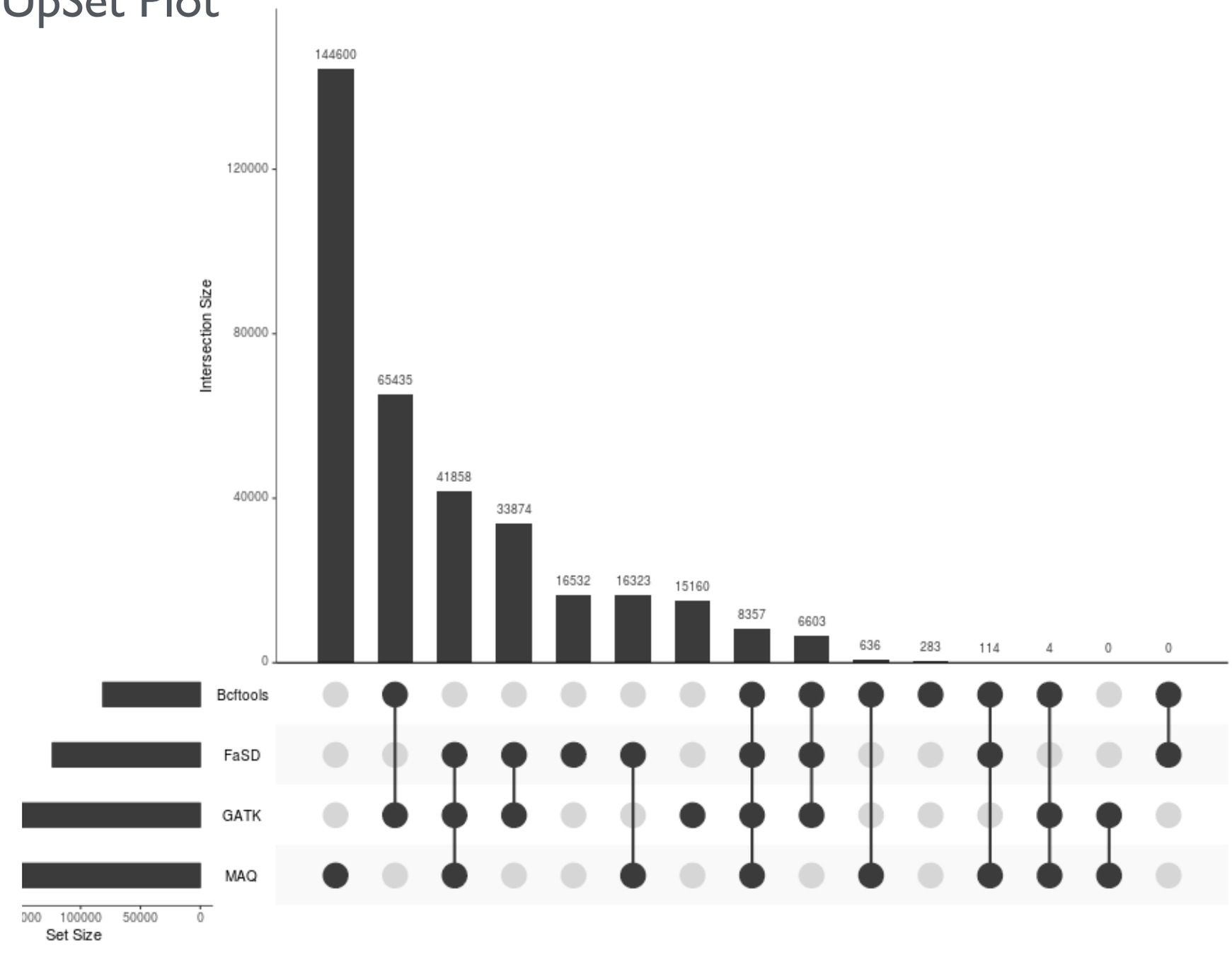
Venn Diagram



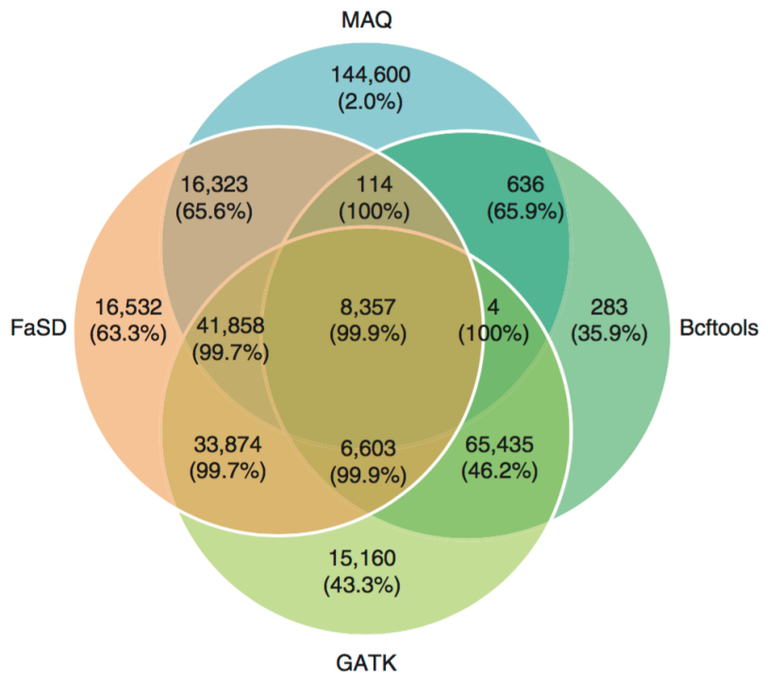
Conway et al., bioRxiv 120600; doi: https://doi.org/10.1101/120600

What's wrong? What's right?

UpSet Plot



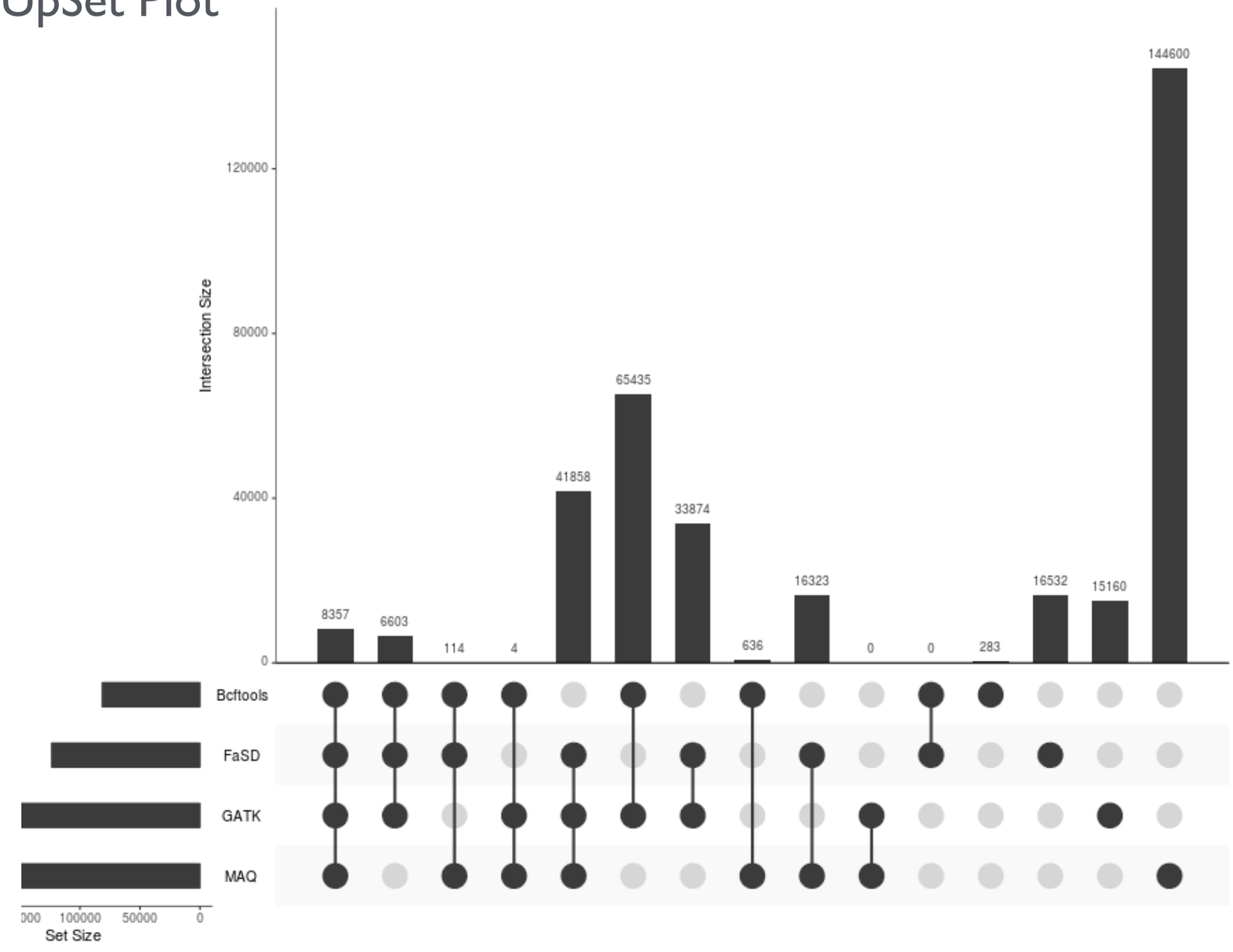
Euler Diagram



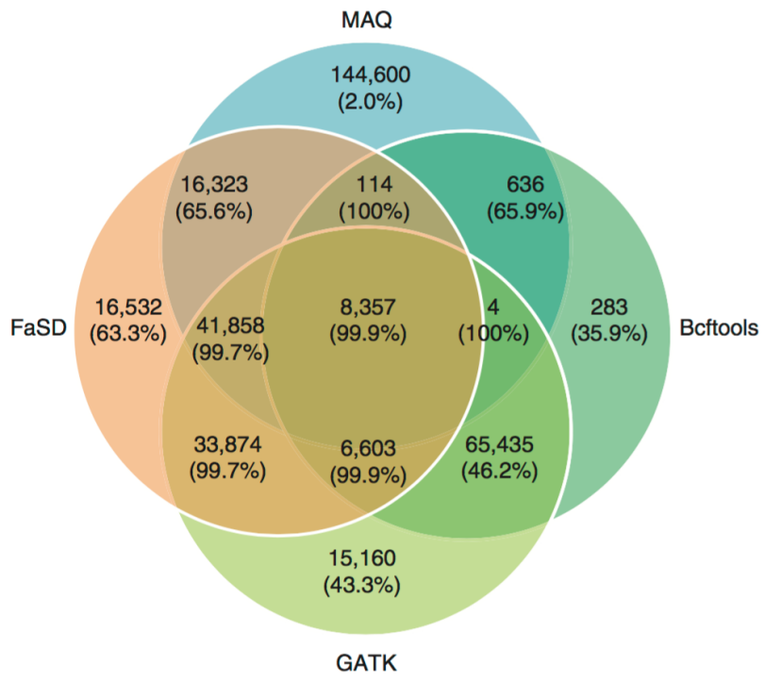
Conway et al., bioRxiv 120600; doi: https://doi.org/10.1101/120600

What's wrong? What's right?

UpSet Plot



Euler Diagram



Ranking of Encodings

Principle of Importance Ordering (Mackinaly 1986):

Encode more important information more effectively.

Ranking of Encodings

- How accurately can the data be read from the visualization?

Streit and Gehlenborg, Nature Methods, 2014

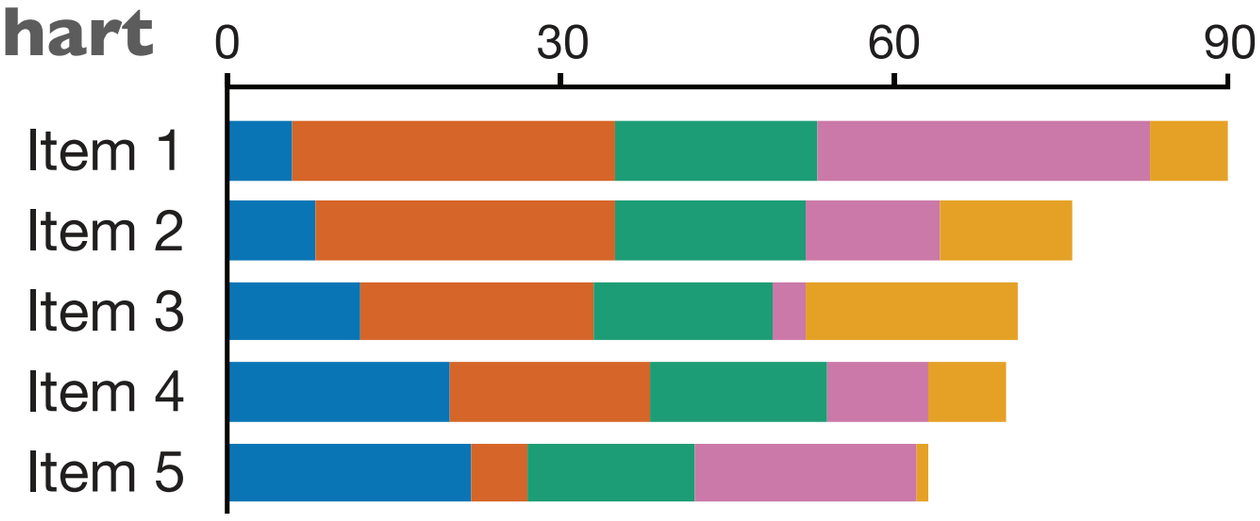
Bar Charts for Multiple Items and Categories

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

Bar Charts for Multiple Items and Categories

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

Stacked Bar Chart

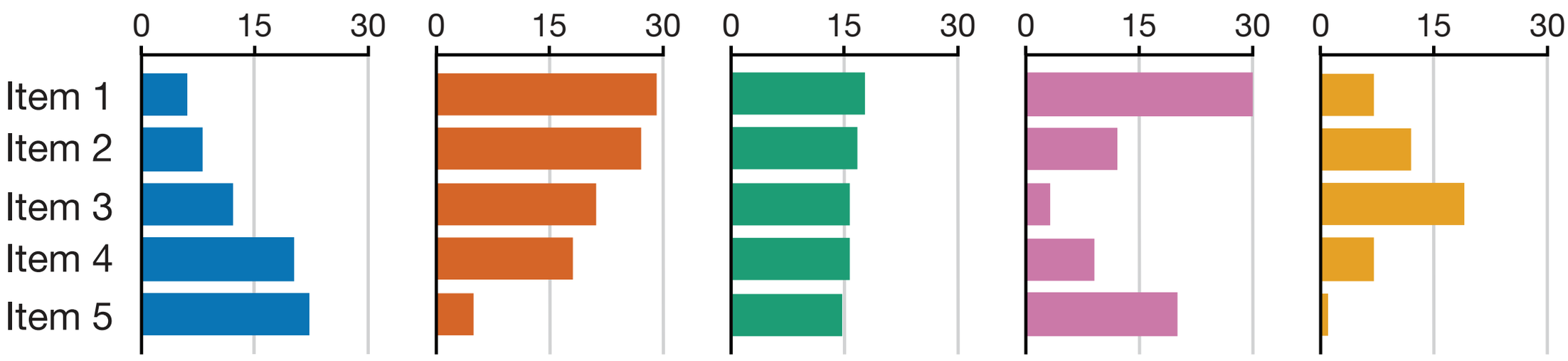


Streit and Gehlenborg, Nature Methods, 2014

Bar Charts for Multiple Items and Categories

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

Layered Bar Chart

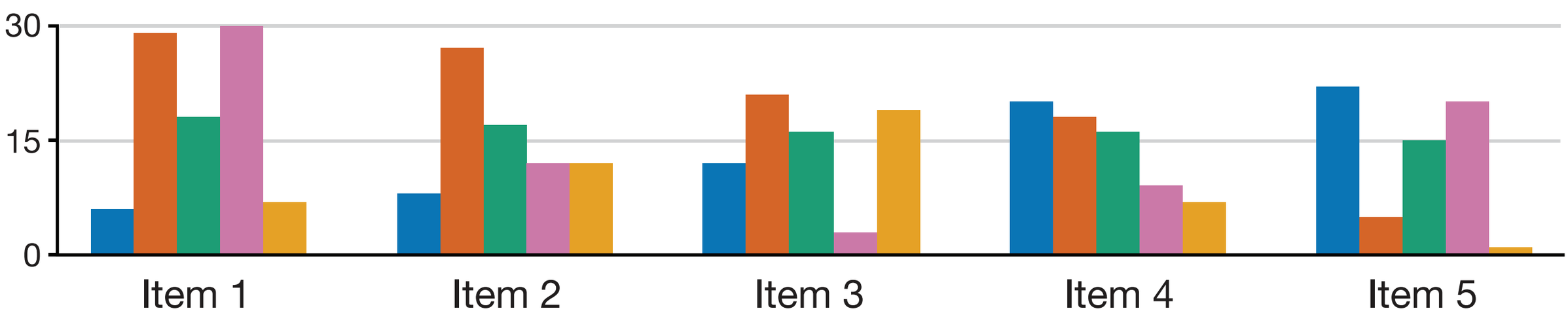


Streit and Gehlenborg, Nature Methods, 2014

Bar Charts for Multiple Items and Categories

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

Grouped Bar Chart



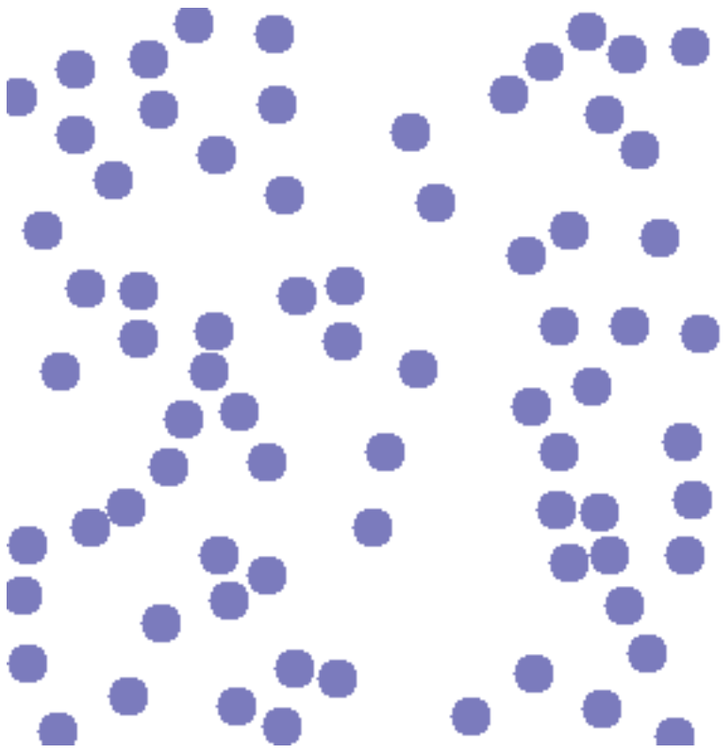
Ranking of Encodings

- How accurately can the data be read from the visualization?

- Which channels are processed preattentively?

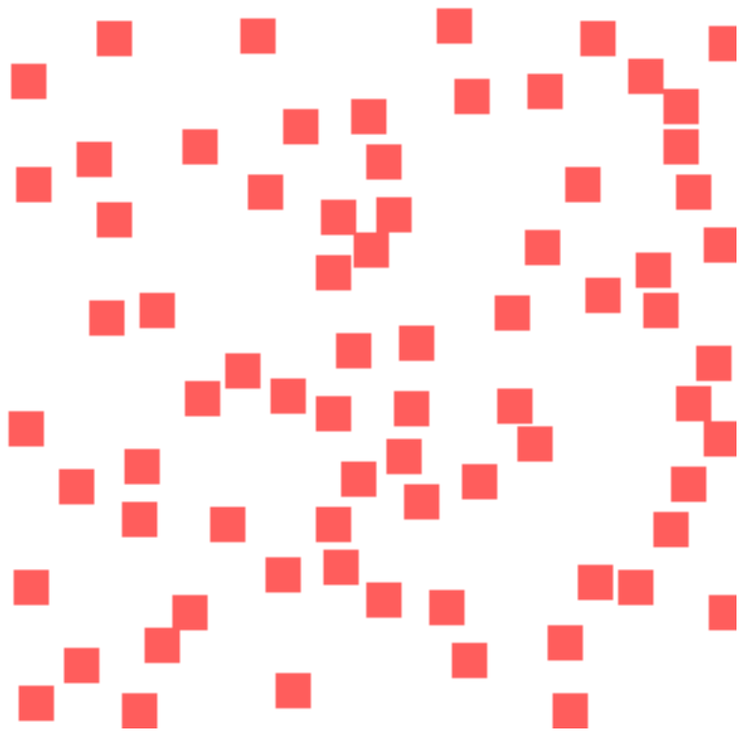
Preattentive Processing: Color

Can you spot this: ●?



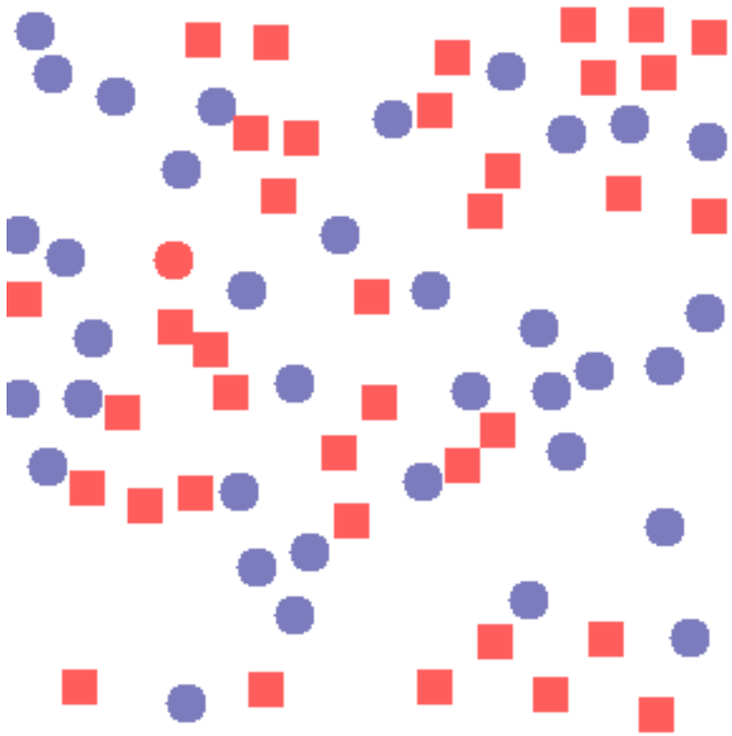
Preattentive Processing: Shape

Can you spot this: ●?



Preattentive Processing: Shape and Color

Can you spot this: ●?



Ranking of Encodings

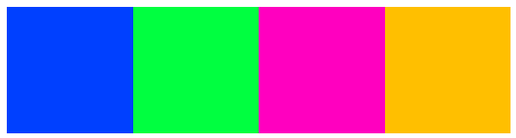
- How accurately can the data be read from the visualization?
- Which channels are processed preattentively?
- How many classes can be distinguished?
- Can the channels be separated from each other?

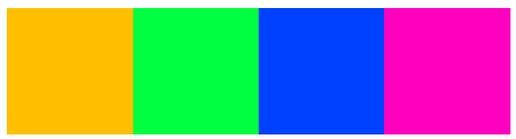
Common Mistakes

after Borland & Taylor 2007, IEEE CG&A

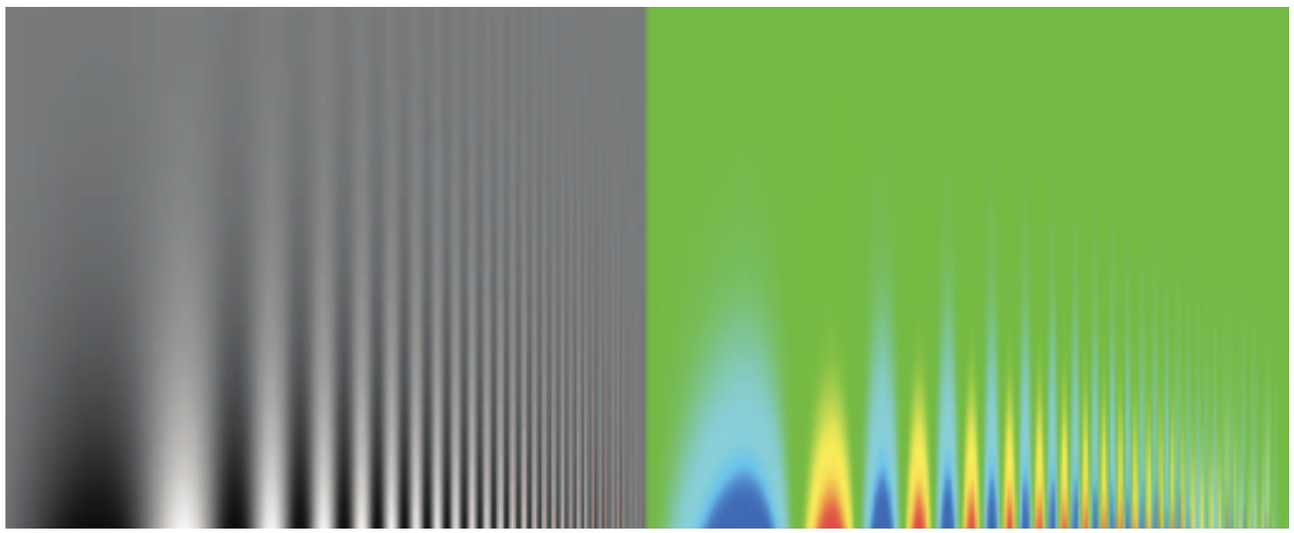
Color Pitfalls: Rainbow Color Map

hard to order



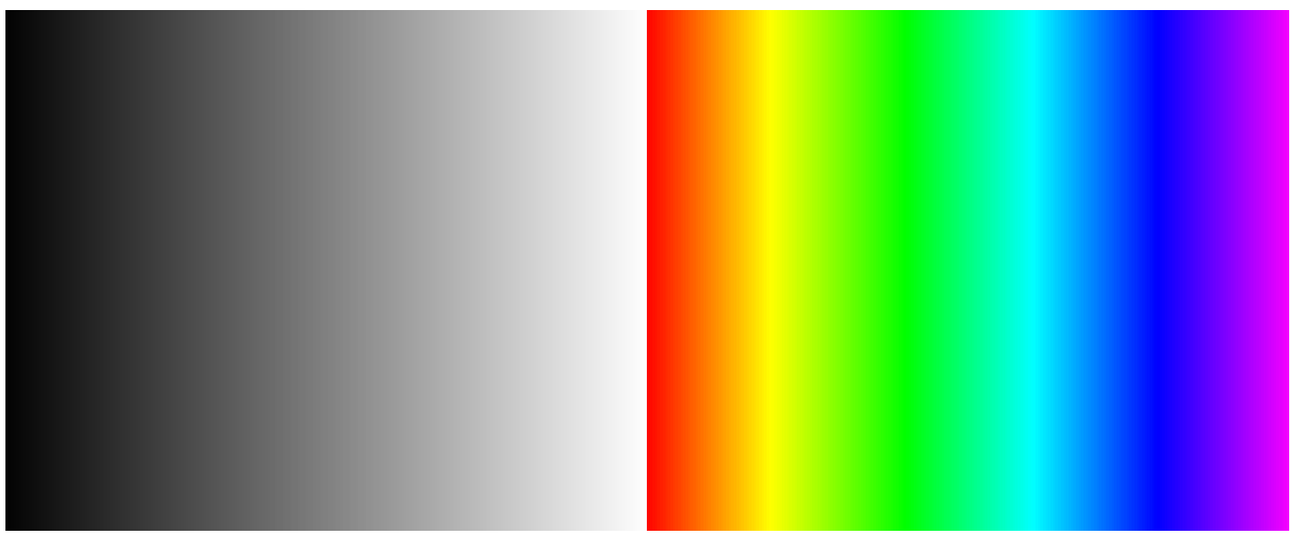


lower resolution



easy to order

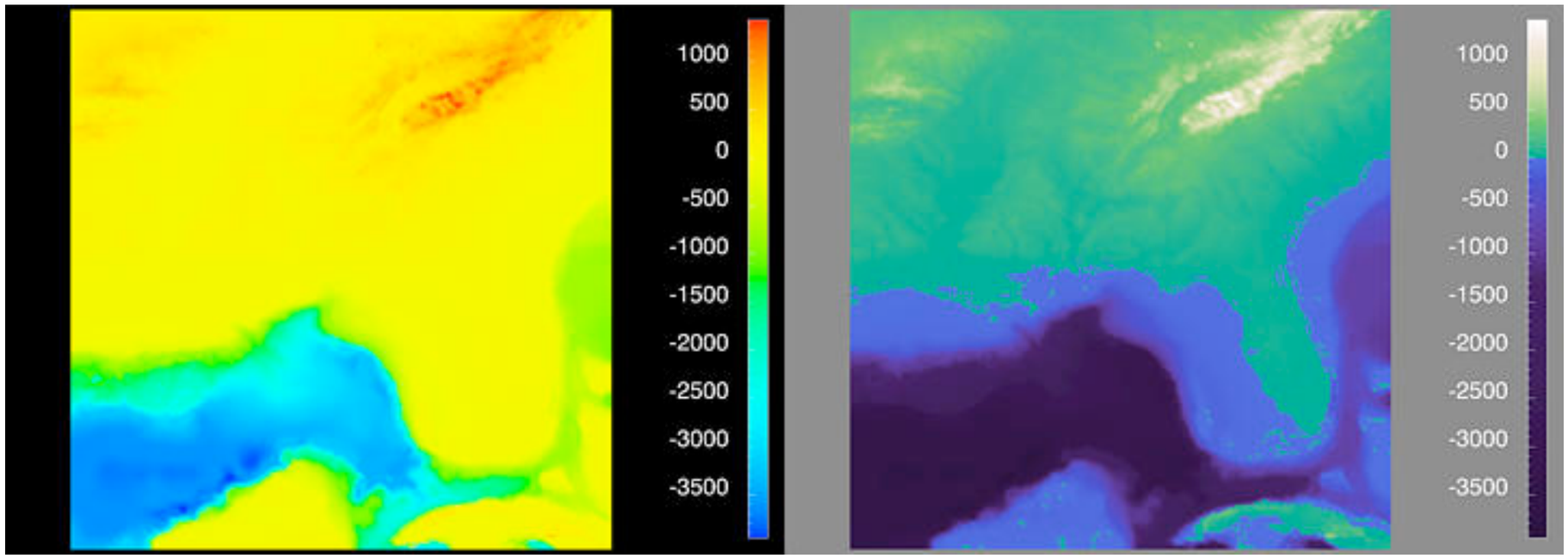
creates artifacts



Rogowitz & Treinish 1996, http://researchweb.watson.ibm.com/people/l/lloydt/color/color.HTM

Color Pitfalls: Rainbow Color Map

Southeastern United States and Gulf of Mexico

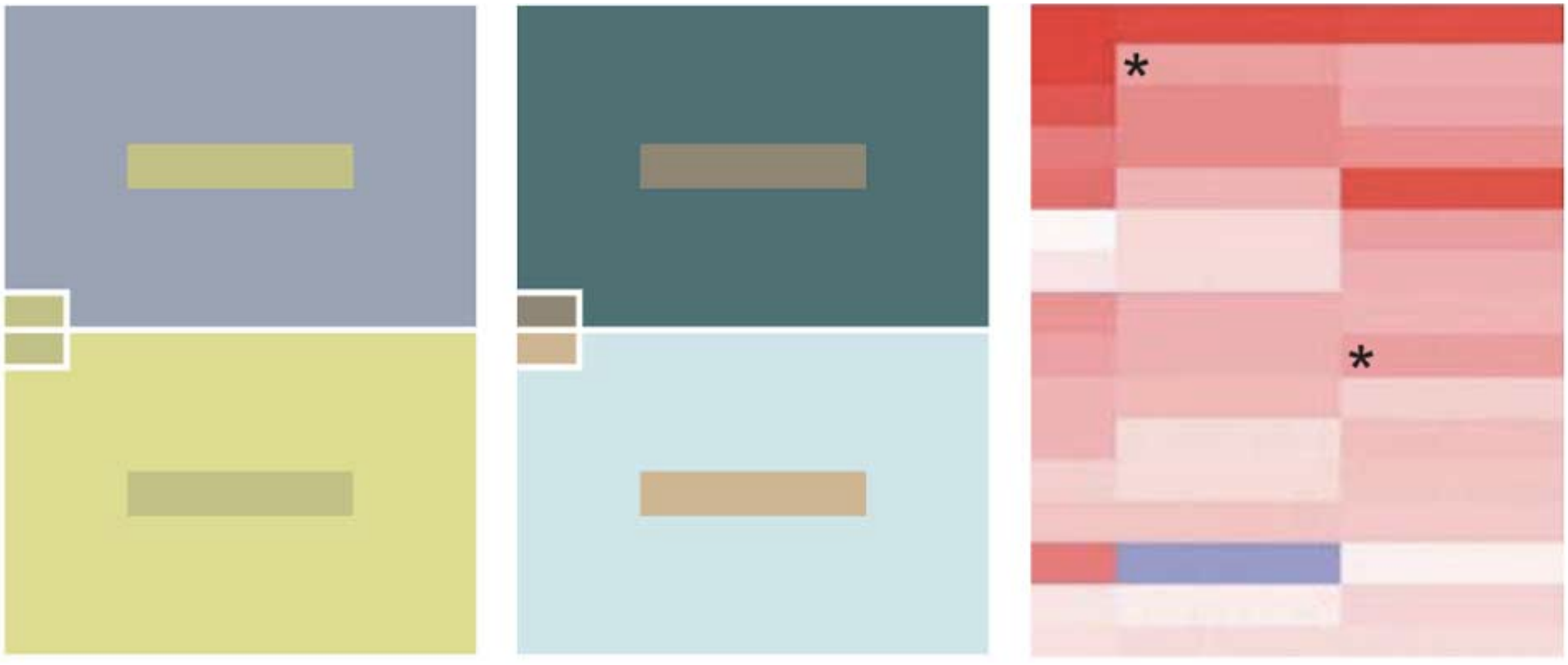


Problems:

- zero crossing not explicit
- lack of ordering of colors makes it hard to interpret the map

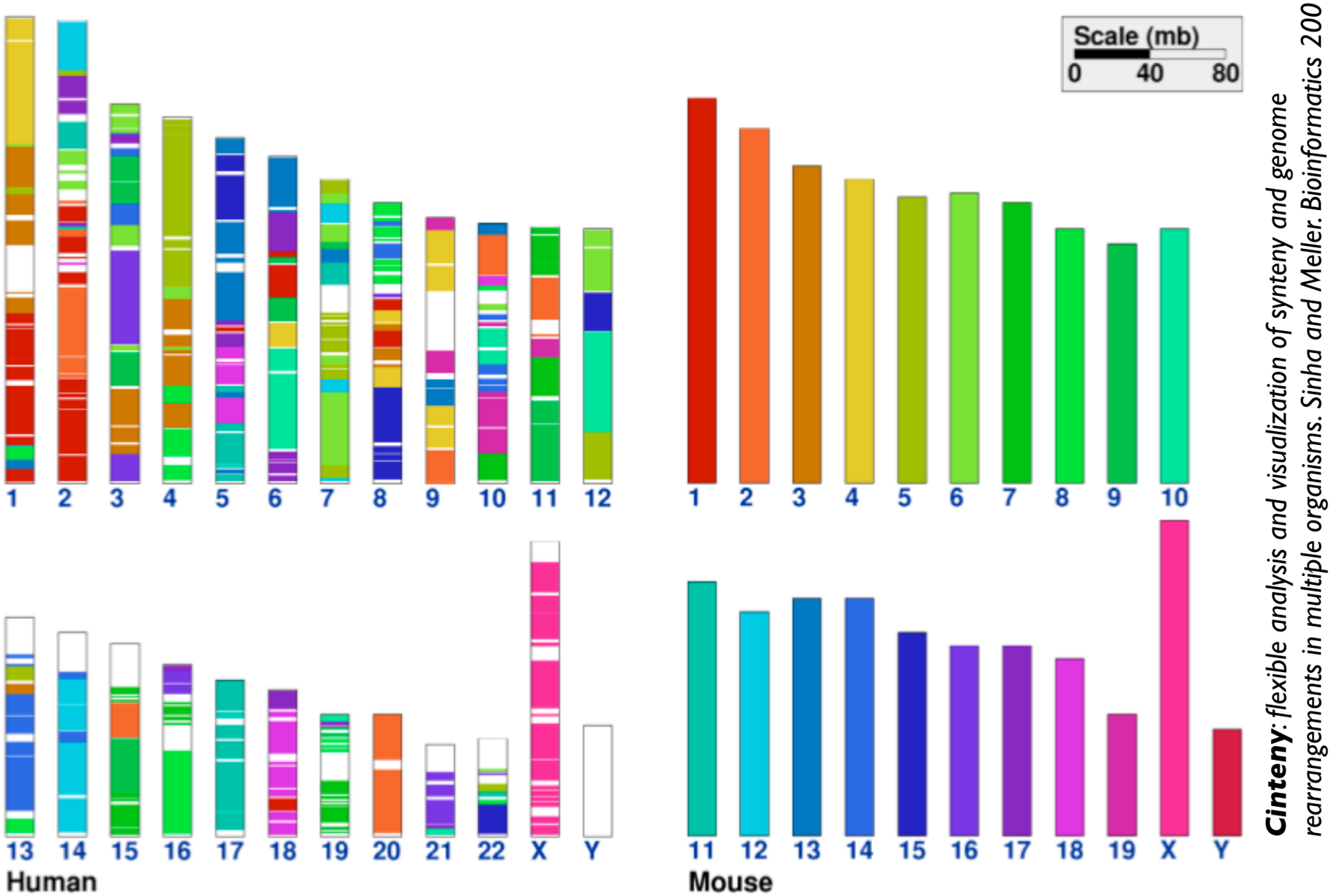
Color Pitfalls: Relativity

Color is a relative medium and context matters



Color Pitfalls: Discriminability

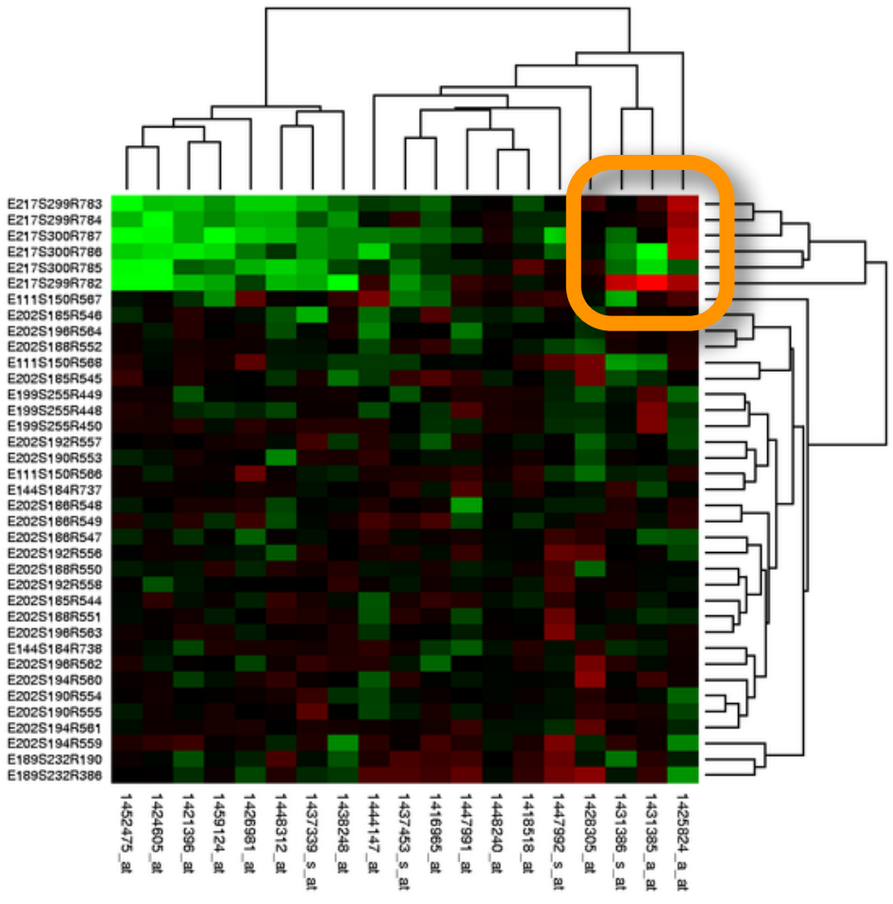
Only 6-12 colors are visually discernible



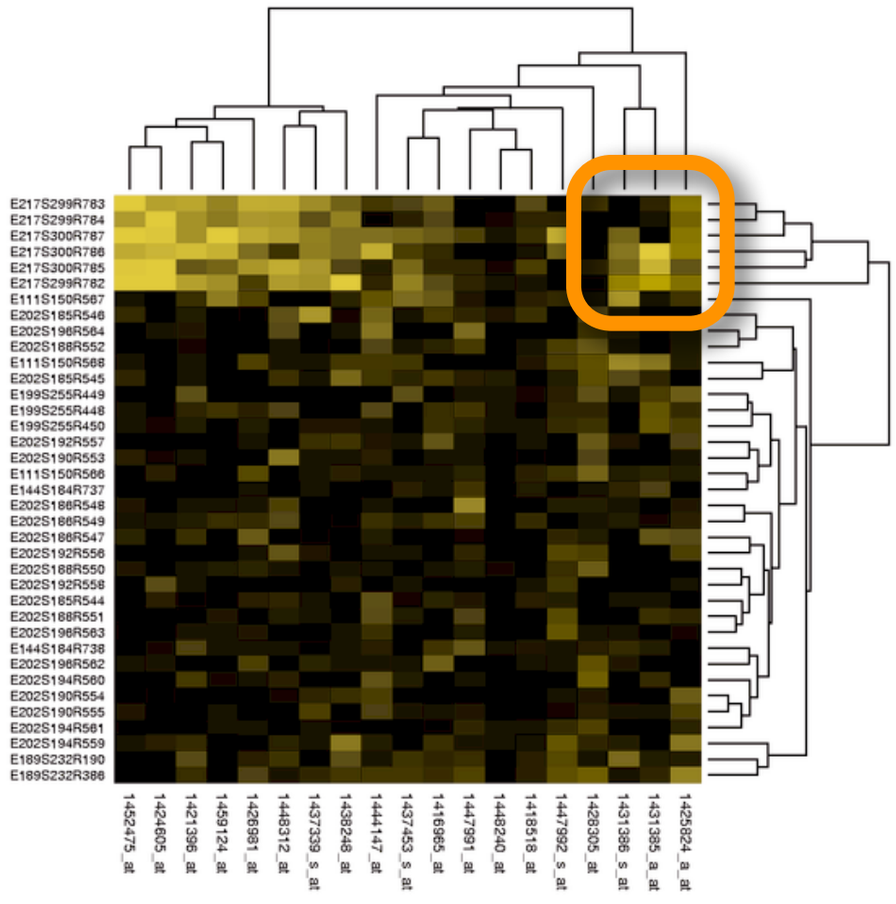
***Cinteny**: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. Sinha and Meller. Bioinformatics 2007*

estimate: Howard Hughes Medical Institute, http://www.hhmi.org/senses/b130.html

Color Pitfalls: Color Blindness



Normal Vision

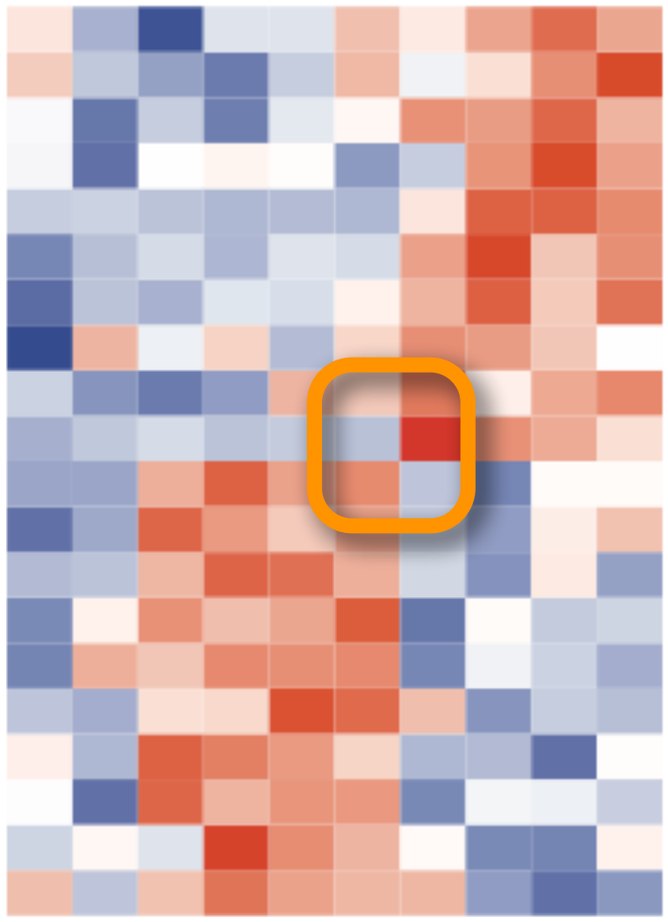


Deuteranope Vision
("Red-Green Blindness")

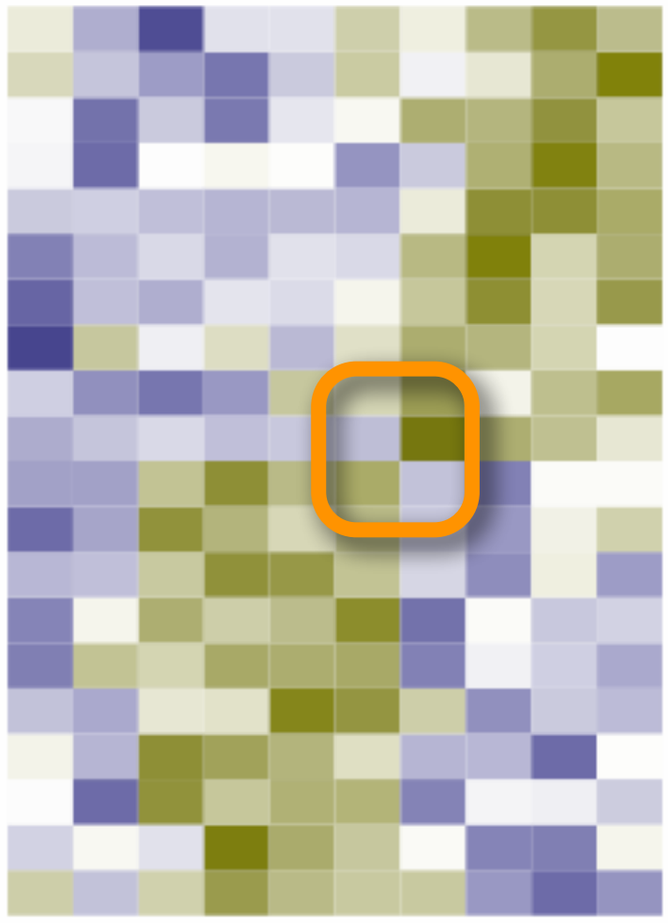
~ 7% of male population affected

estimate: Howard Hughes Medical Institute, http://www.hhmi.org/senses/b130.html

Color Pitfalls: Color Blindness



Normal Vision

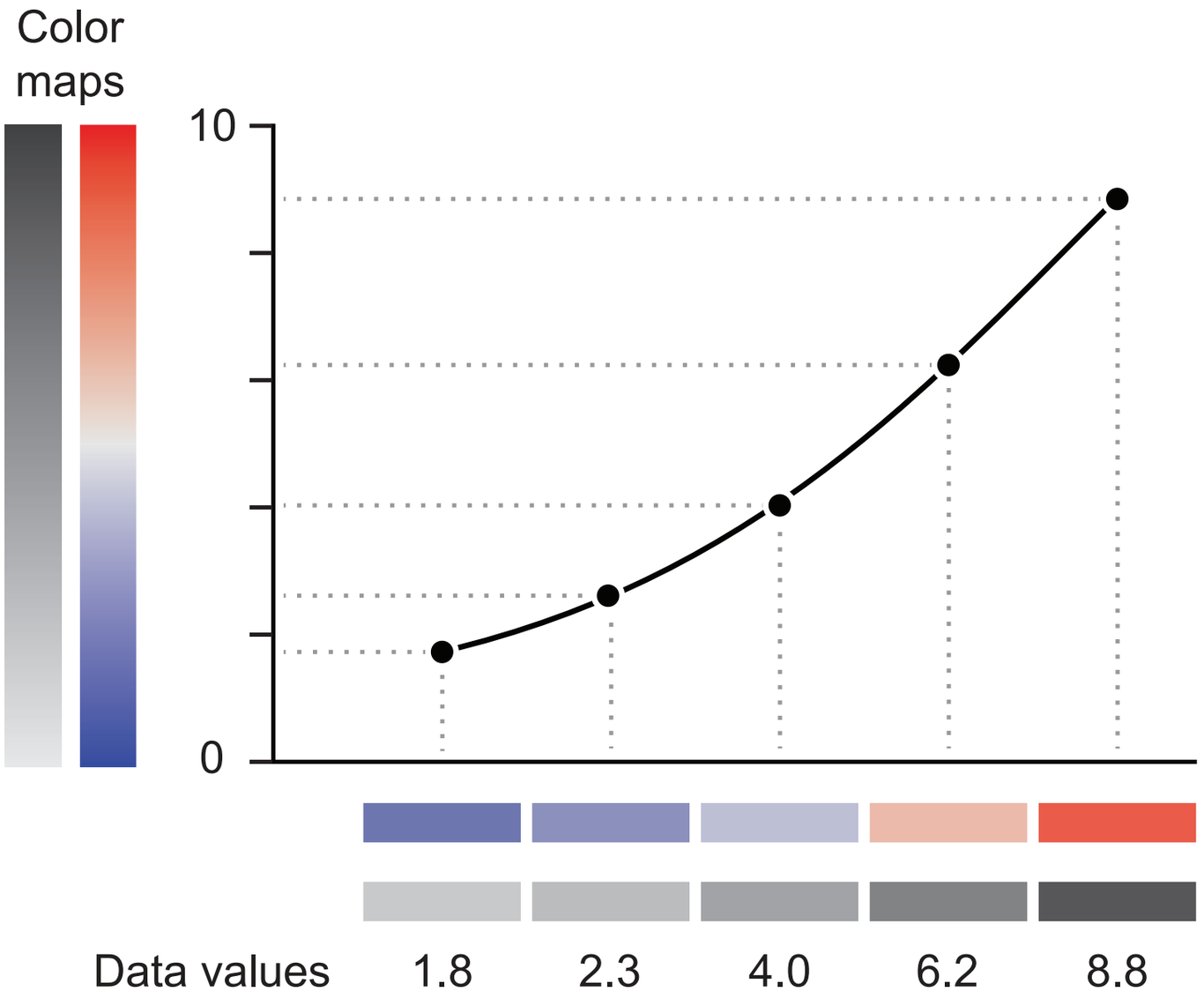


Deuteranope Vision
("Red-Green Blindness")

~ 7% of male population affected

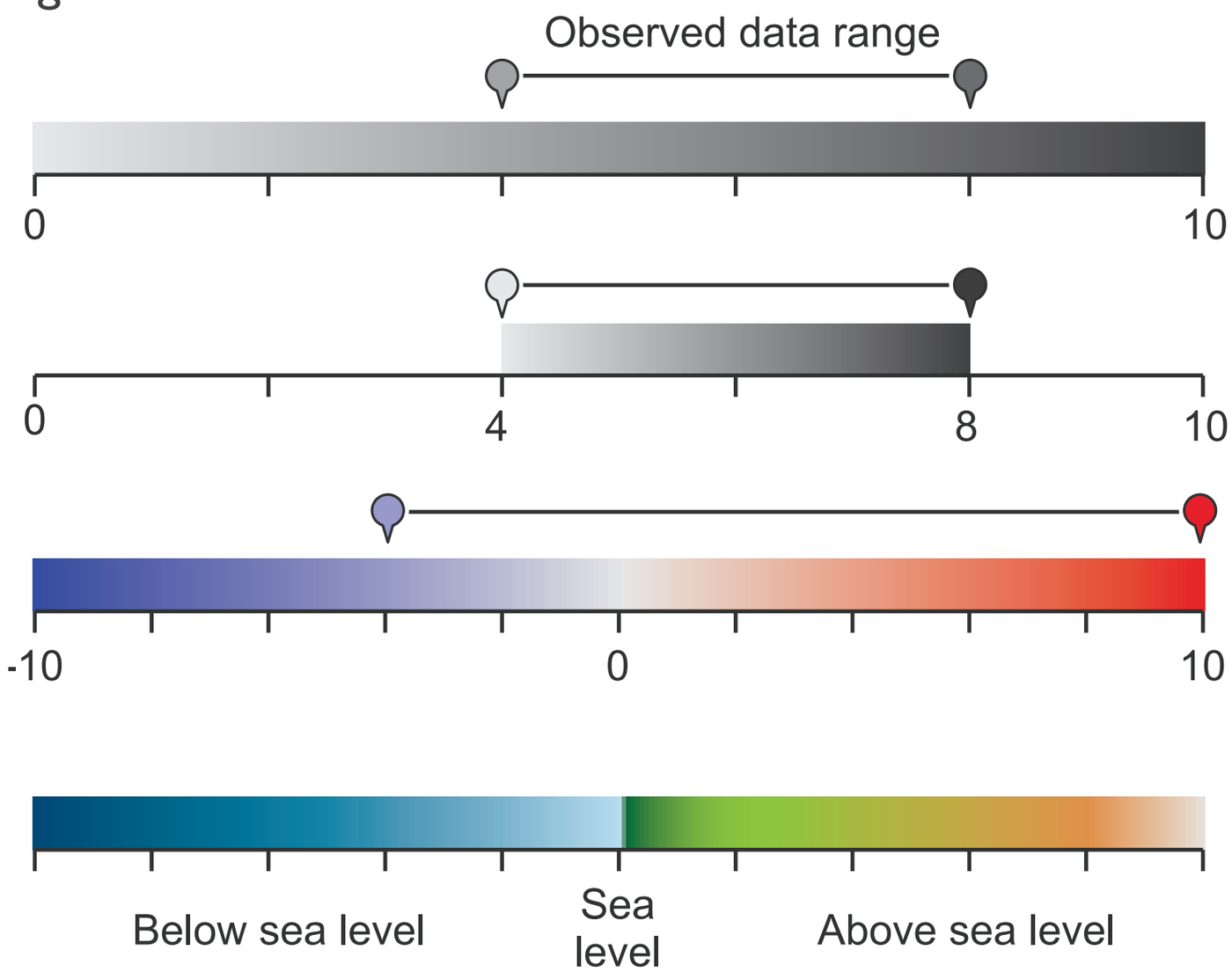
Color Pitfalls: Color Mapping

Bad color mapping!



Color Pitfalls: Color Mapping

Good color mapping!



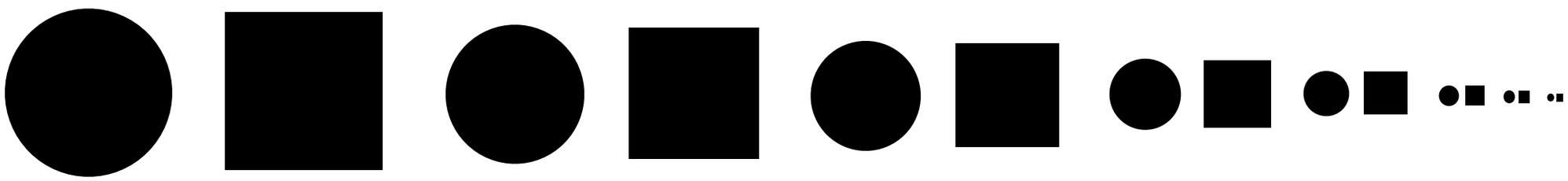
Remember!

Color used poorly is worse than no color at all.

— *Edward Tufte*

after Mackinlay, ACM Transactions on Graphics 5, 110-141, 1986

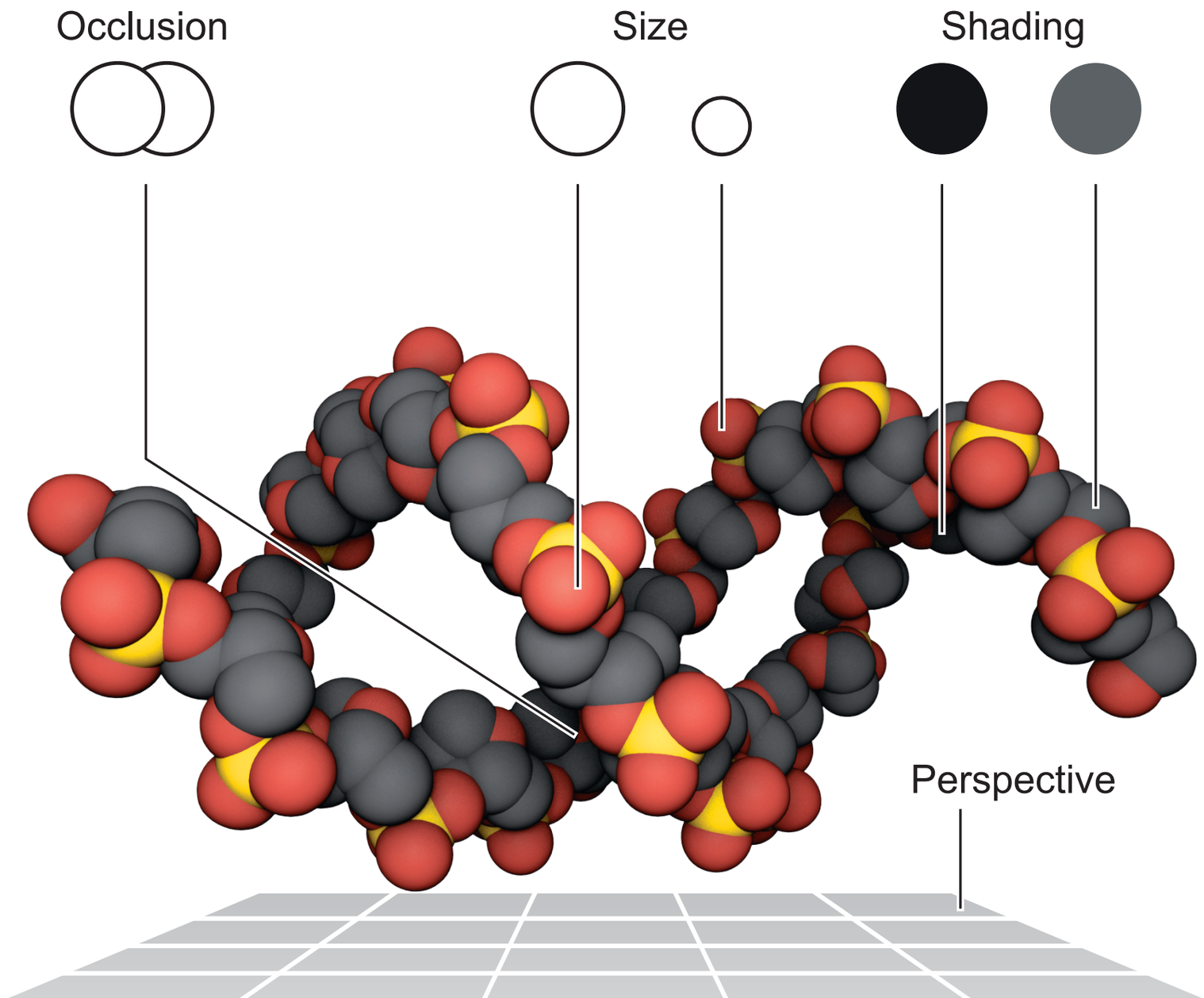
Encoding Pitfalls: Interference between Channels



shape and size

Gehlenborg and Wong, Nature Methods, 2012

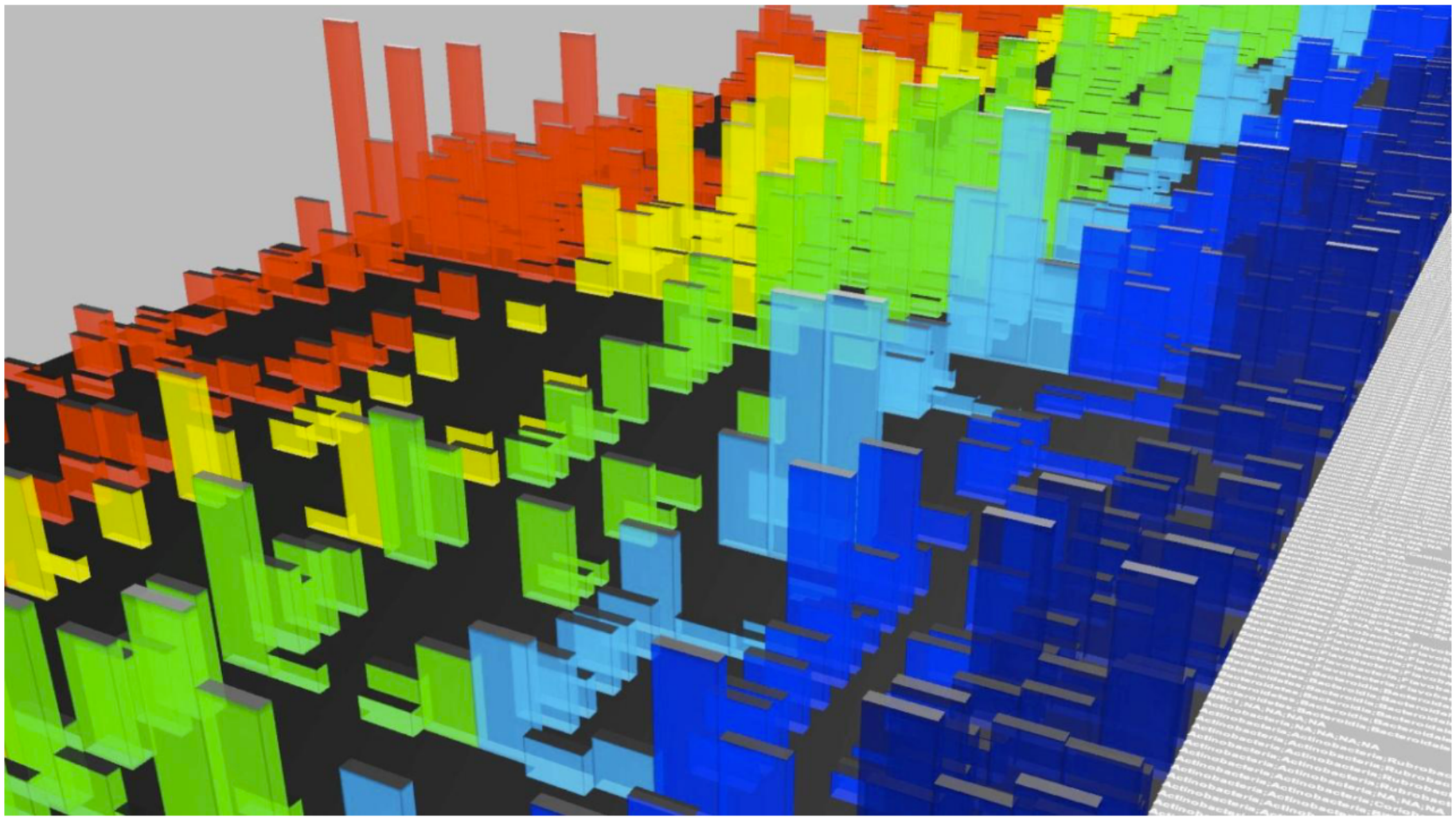
3D: Depth Cues



3D Pitfalls: Perspective

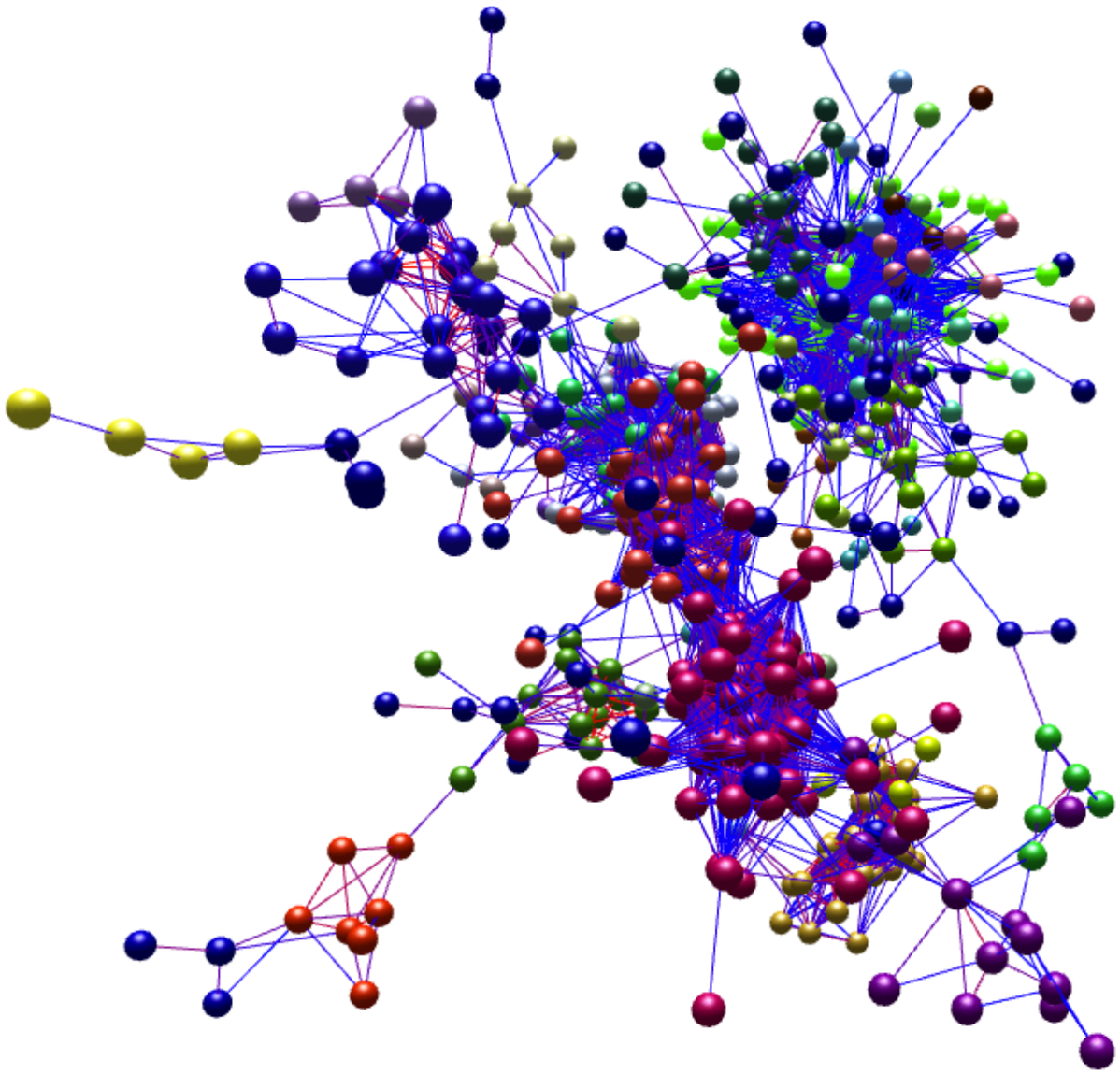
- **perspective distortion:** interferes with size channel encoding

- **shading:** interferes with color, lightness, and saturation channel encodings

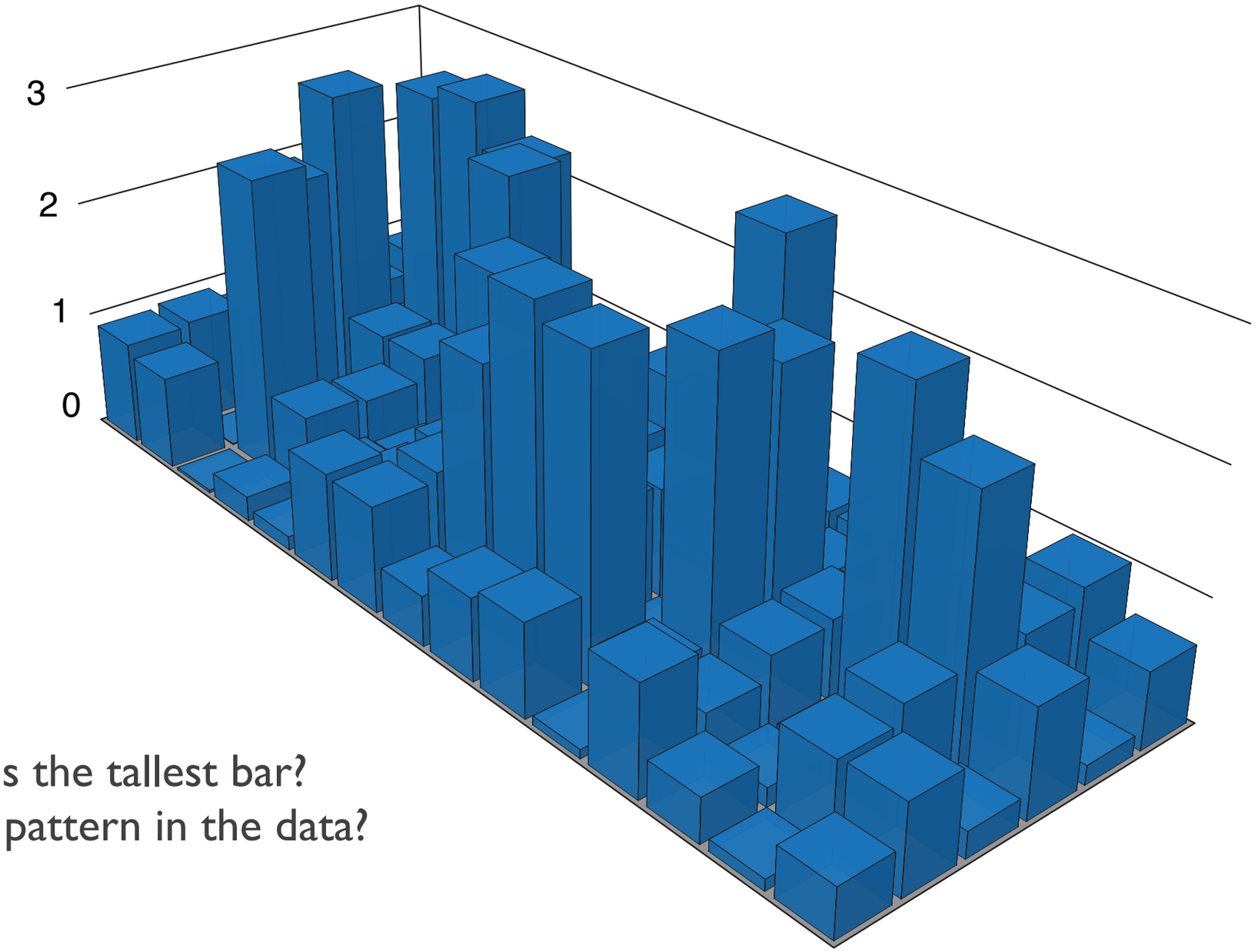


BioLayout 3D 2.0 sample dataset, http://www.biolayout.org

3D Pitfalls: Occlusion

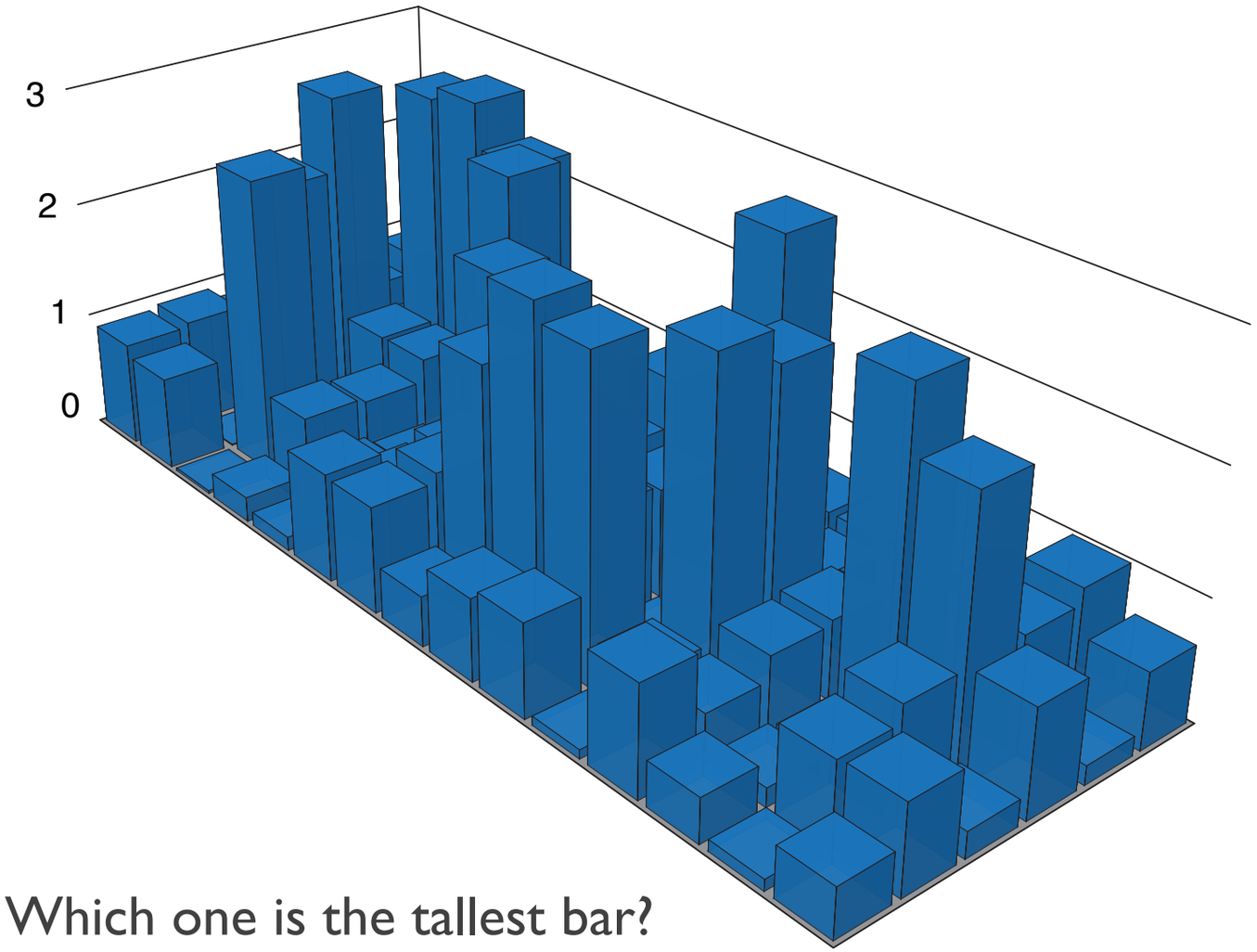


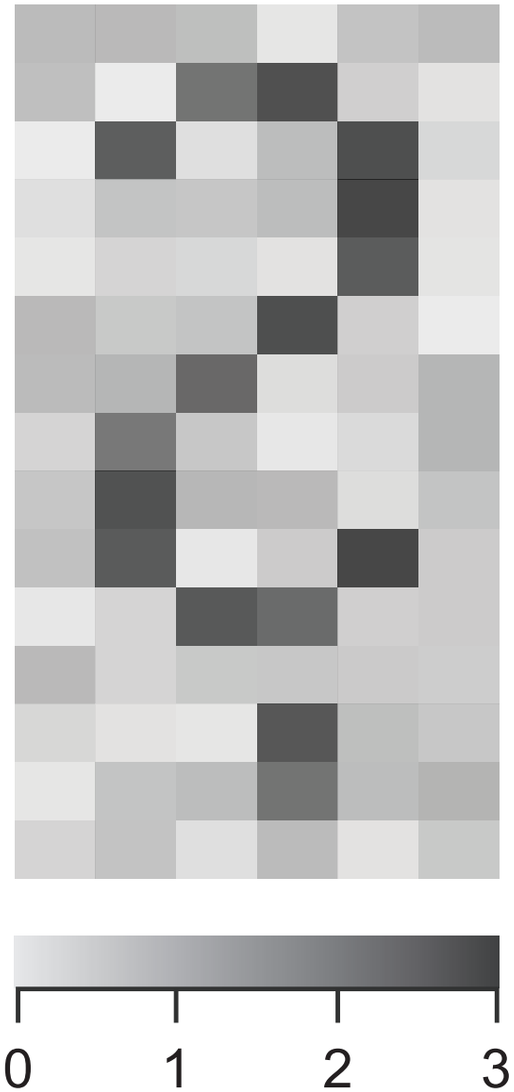
3D Pitfalls: Occlusion and Perspective



Which one is the tallest bar?
What is the pattern in the data?

3D Pitfalls: Occlusion and Perspective

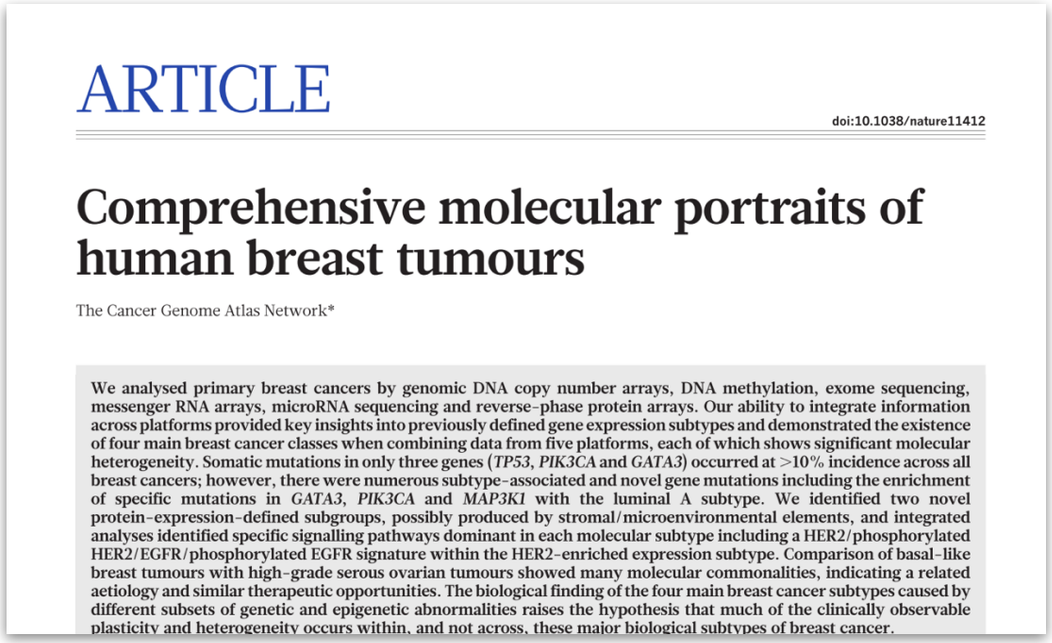




Which one is the tallest bar?
What is the pattern in the data?

Visualization for Communication

Insights into the Paper, Data into the Database



ARTICLE

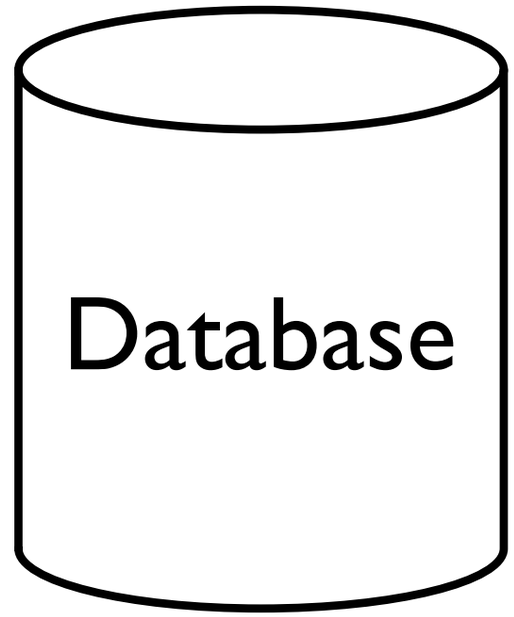
doi:10.1038/nature11412

Comprehensive molecular portraits of human breast tumours

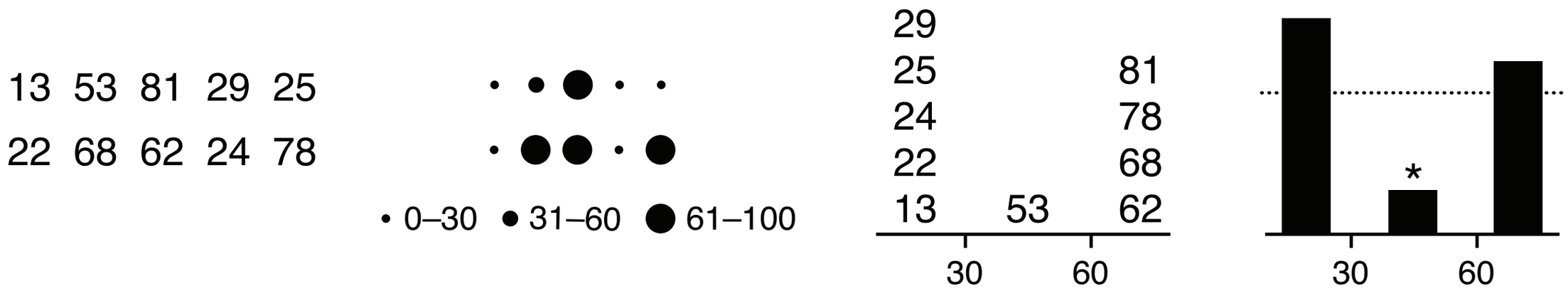
The Cancer Genome Atlas Network*

We analysed primary breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays. Our ability to integrate information across platforms provided key insights into previously defined gene expression subtypes and demonstrated the existence of four main breast cancer classes when combining data from five platforms, each of which shows significant molecular heterogeneity. Somatic mutations in only three genes (*TP53*, *PIK3CA* and *GATA3*) occurred at >10% incidence across all breast cancers; however, there were numerous subtype-associated and novel gene mutations including the enrichment of specific mutations in *GATA3*, *PIK3CA* and *MAP3K1* with the luminal A subtype. We identified two novel protein-expression-defined subgroups, possibly produced by stromal/microenvironmental elements, and integrated analyses identified specific signalling pathways dominant in each molecular subtype including a HER2/phosphorylated HER2/EGFR/phosphorylated EGFR signature within the HER2-enriched expression subtype. Comparison of basal-like breast tumours with high-grade serous ovarian tumours showed many molecular commonalities, indicating a related aetiology and similar therapeutic opportunities. The biological finding of the four main breast cancer subtypes caused by different subsets of genetic and epigenetic abnormalities raises the hypothesis that much of the clinically observable plasticity and heterogeneity occurs within, and not across, these major biological subtypes of breast cancer.

≠

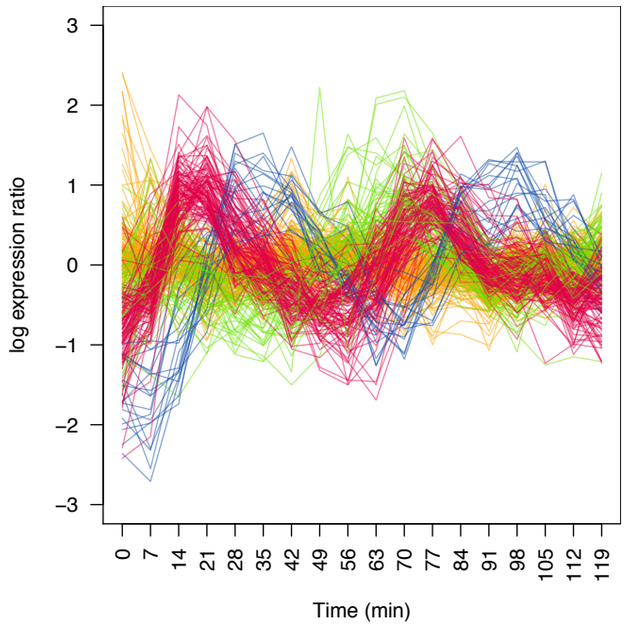


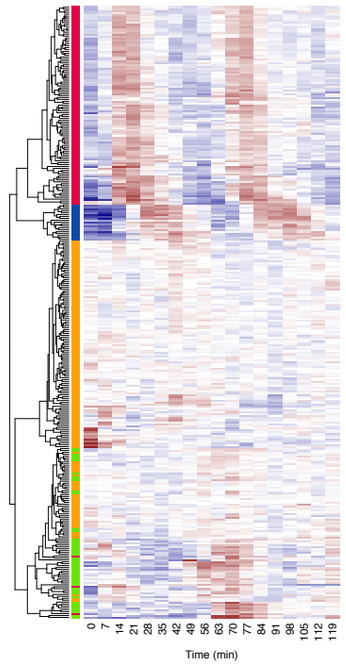
Focus on the Message: Reduce Data Detail

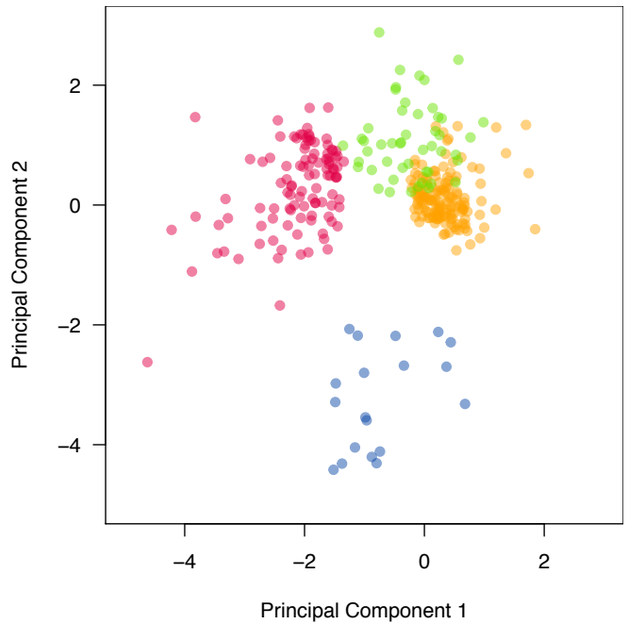


Focus

Choose a Plot Appropriate for your Message





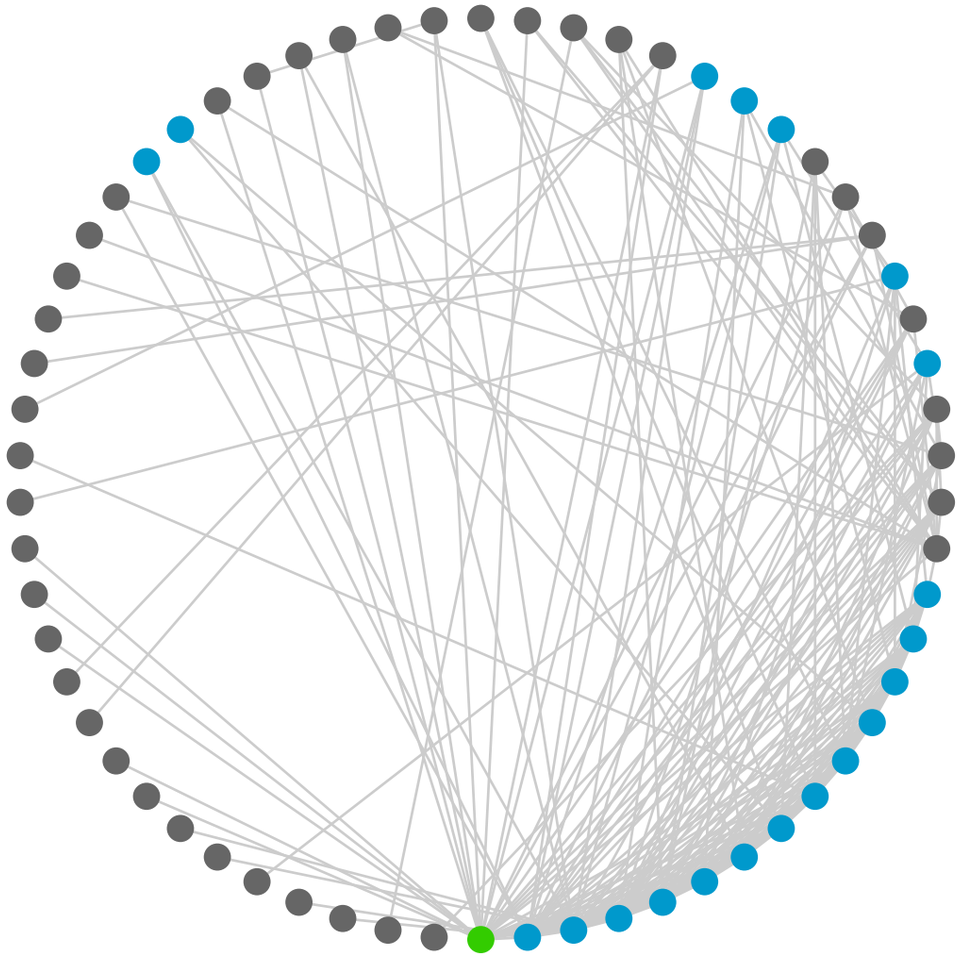


few, high-res

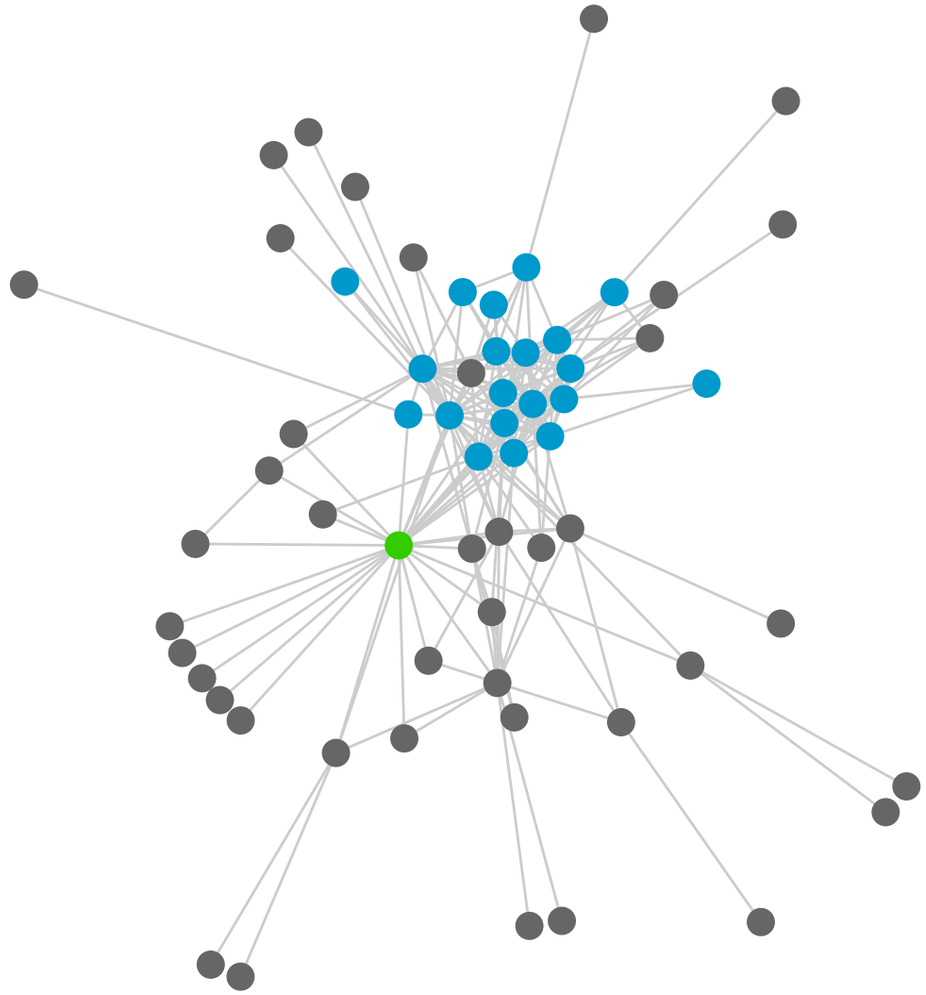
many, low-res

Gehlenborg and Wong, Nature Methods, 2012

Choose a Plot Appropriate for your Message



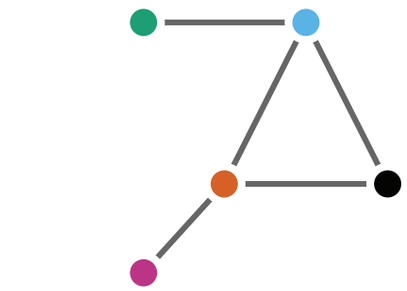
Circular Layout

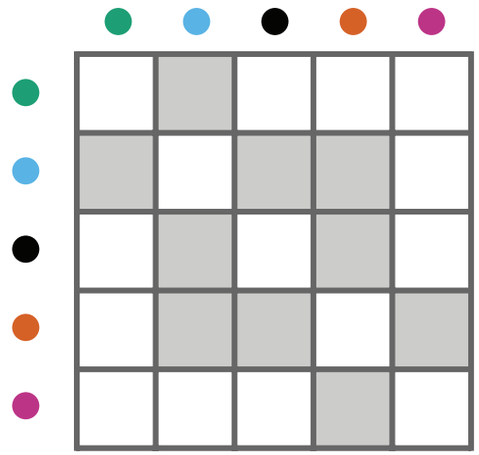


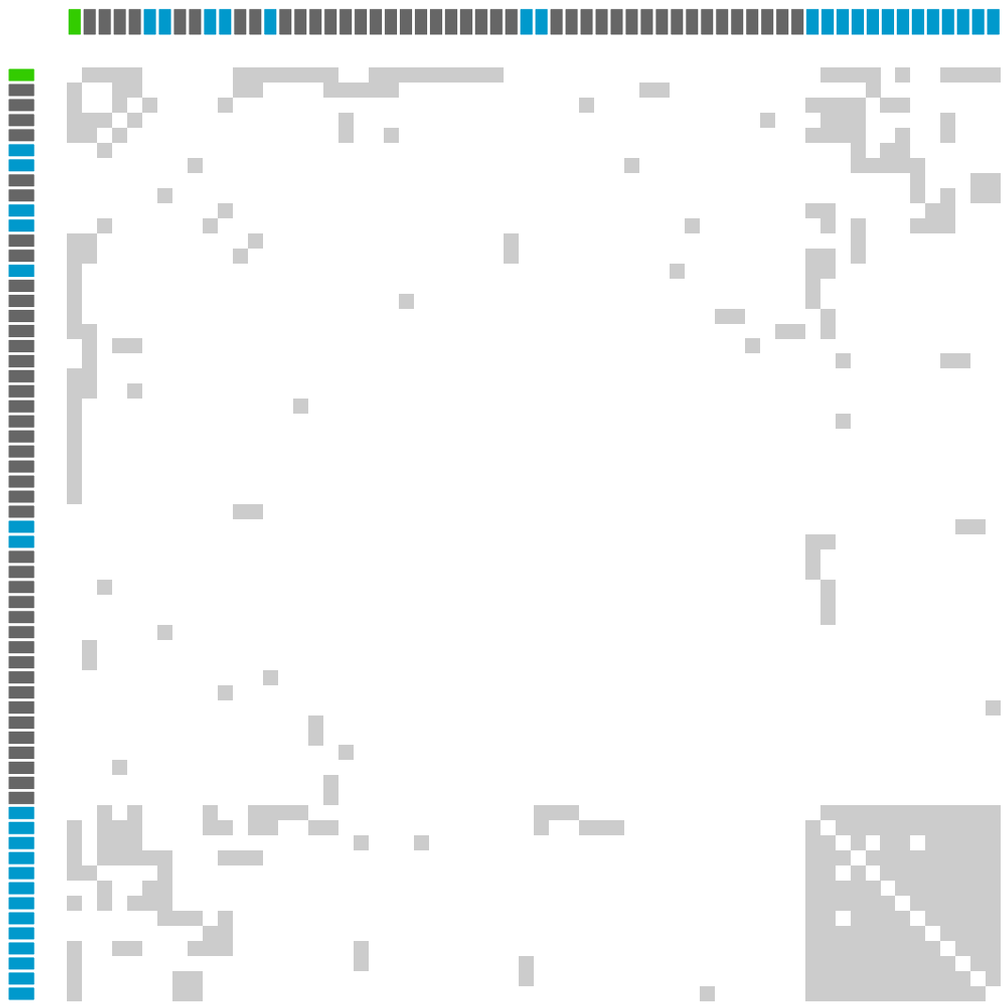
Force-directed Layout

Gehlenborg and Wong, Nature Methods, 2012

Choose a Plot Appropriate for your Message





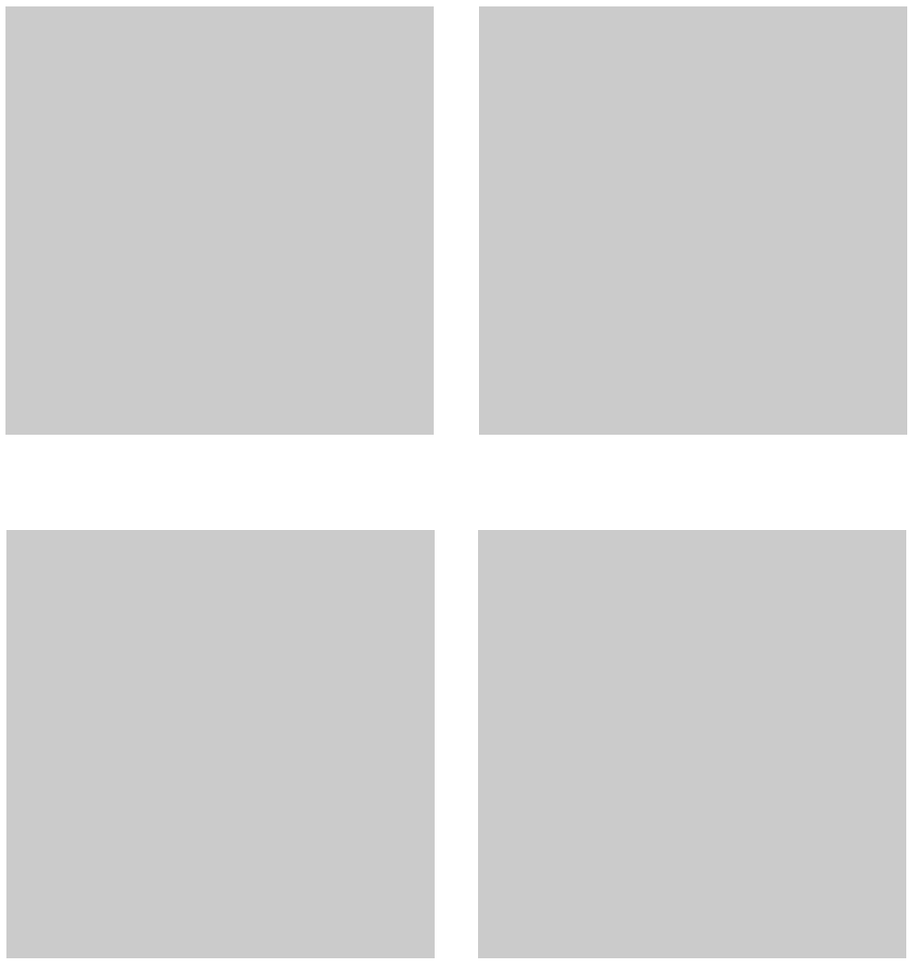


after Wong, Nature Methods, 2011

Use Layout to Convey Meaning

after Wong, Nature Methods, 2011

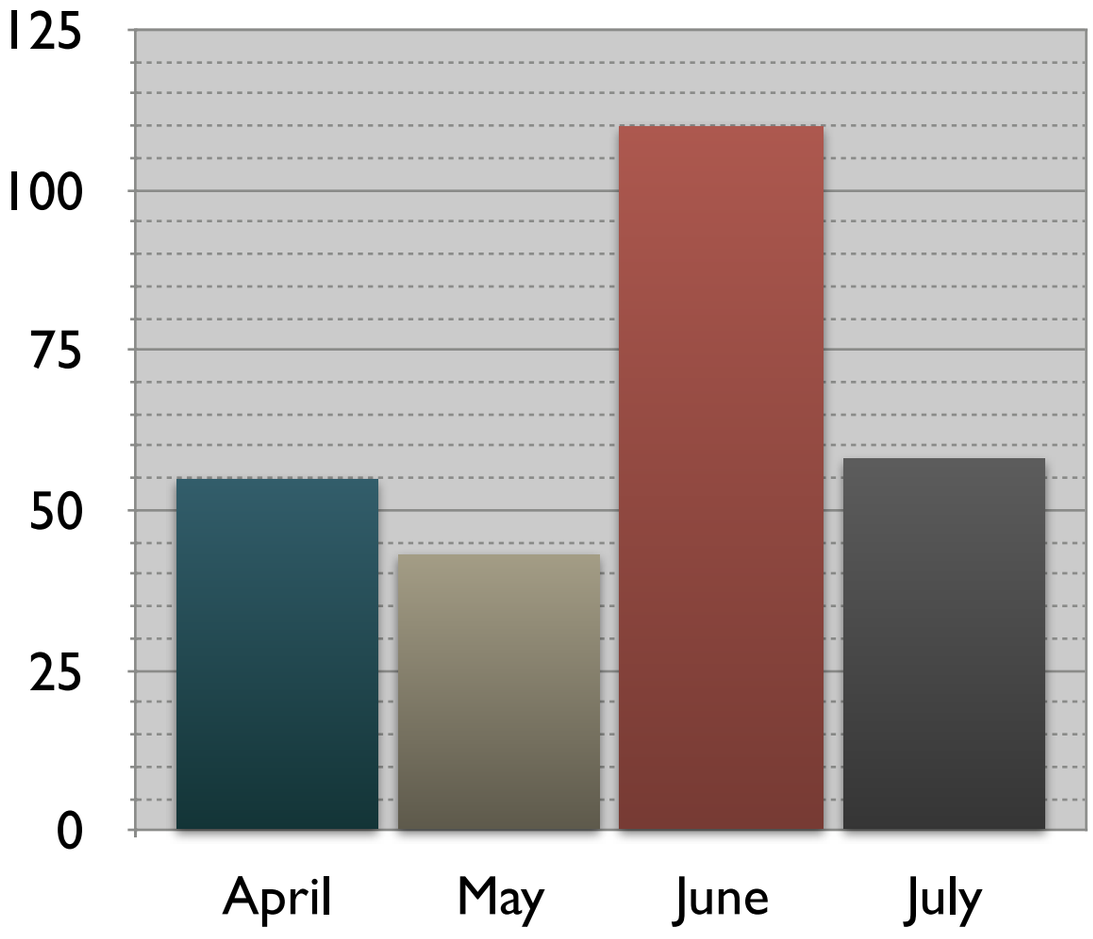
Use Layout to Convey Meaning

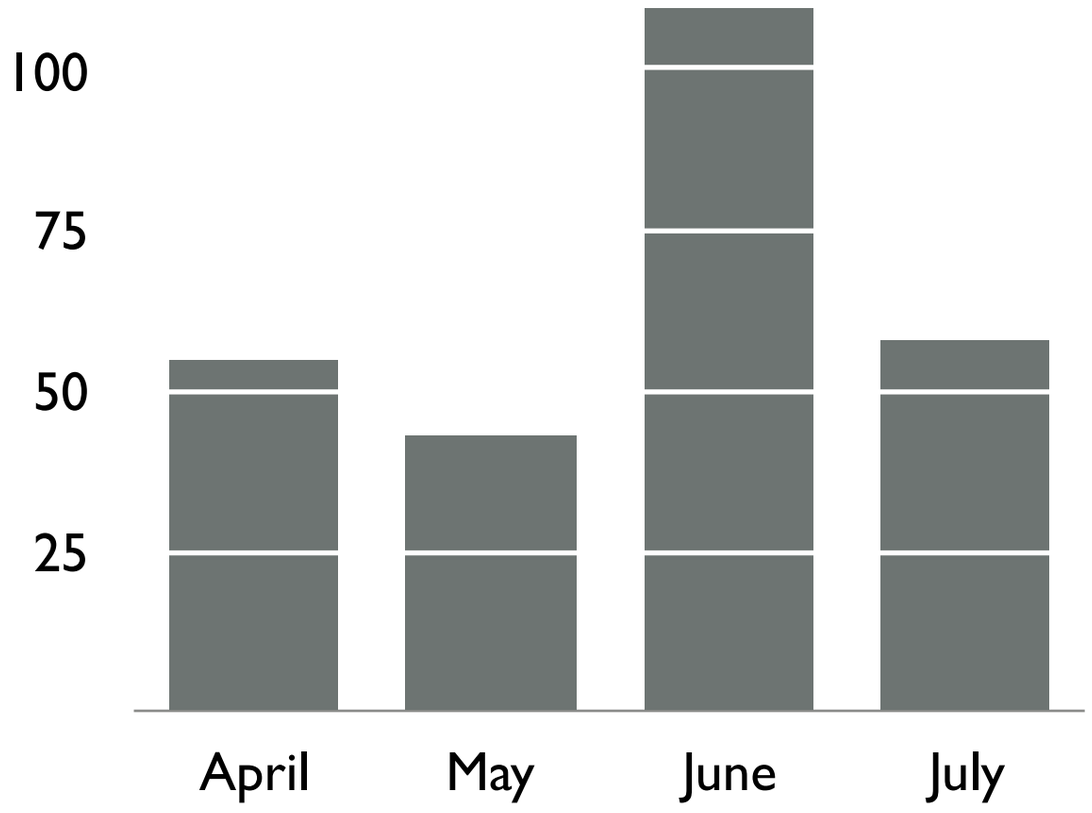


after Wong, Nature Methods, 2011

Use Layout to Convey Meaning

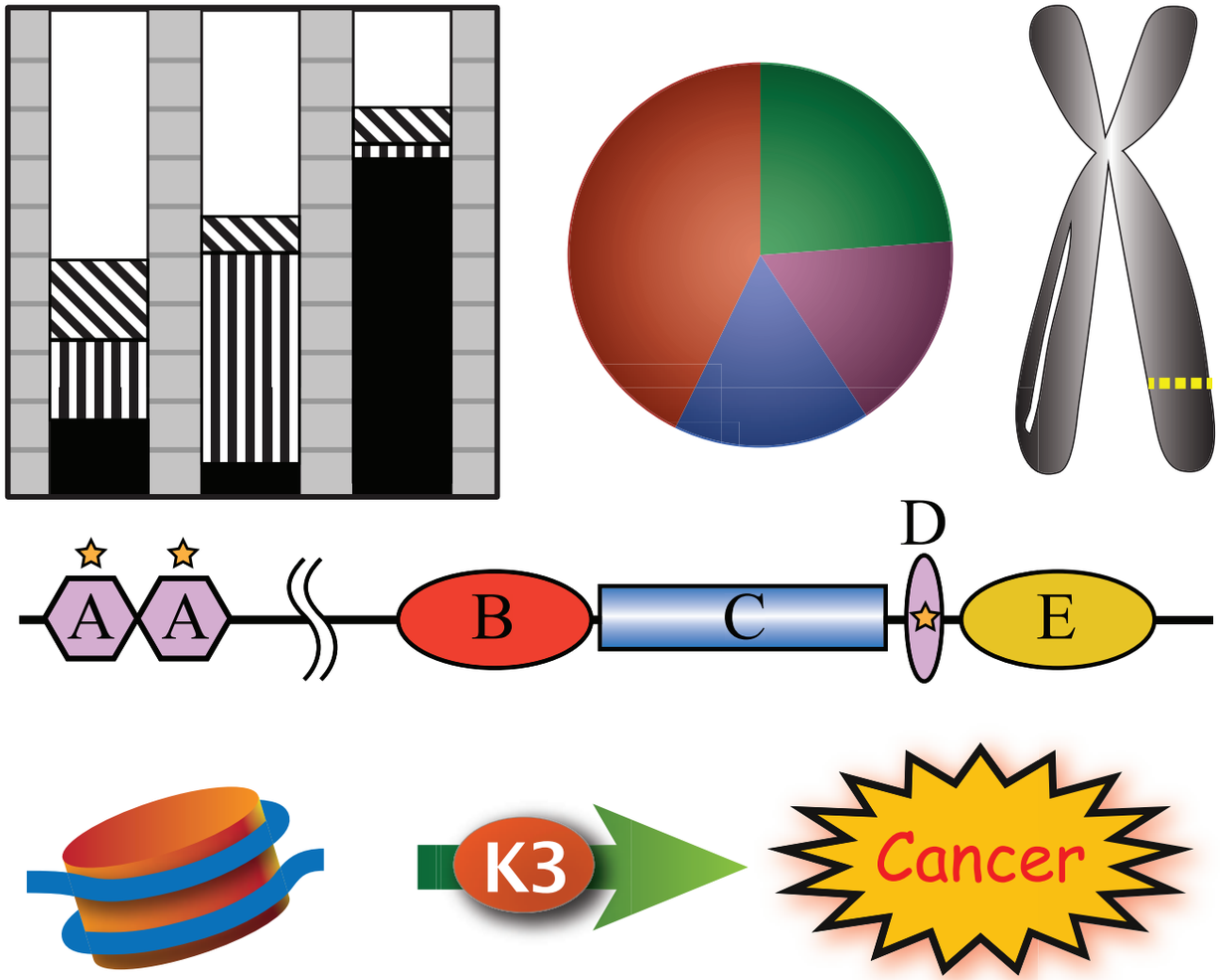
Focus on the Data: Maximize Data/Ink Ratio





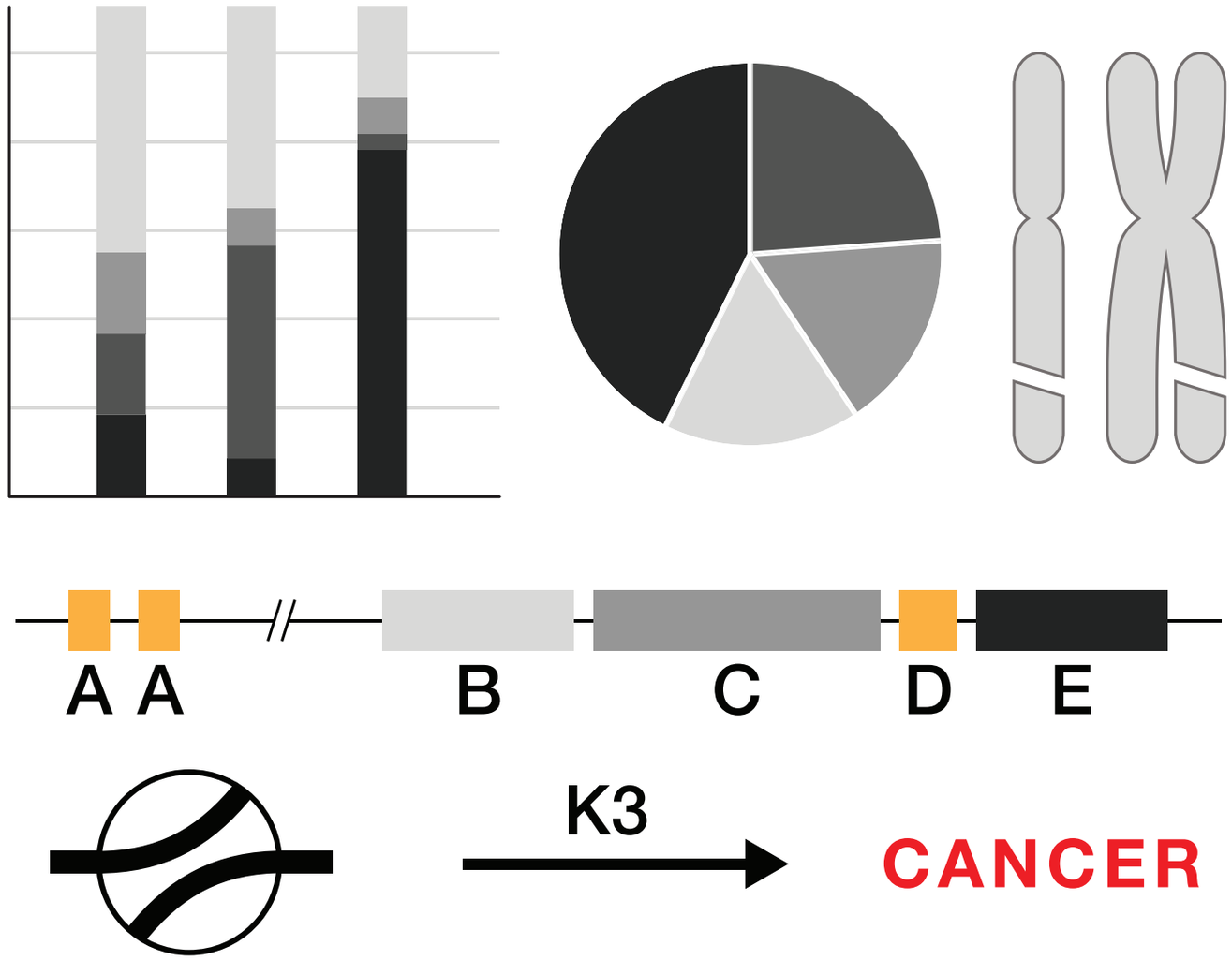
Krzywinski, Nature Methods, 2013

Don’t shout at your audience!



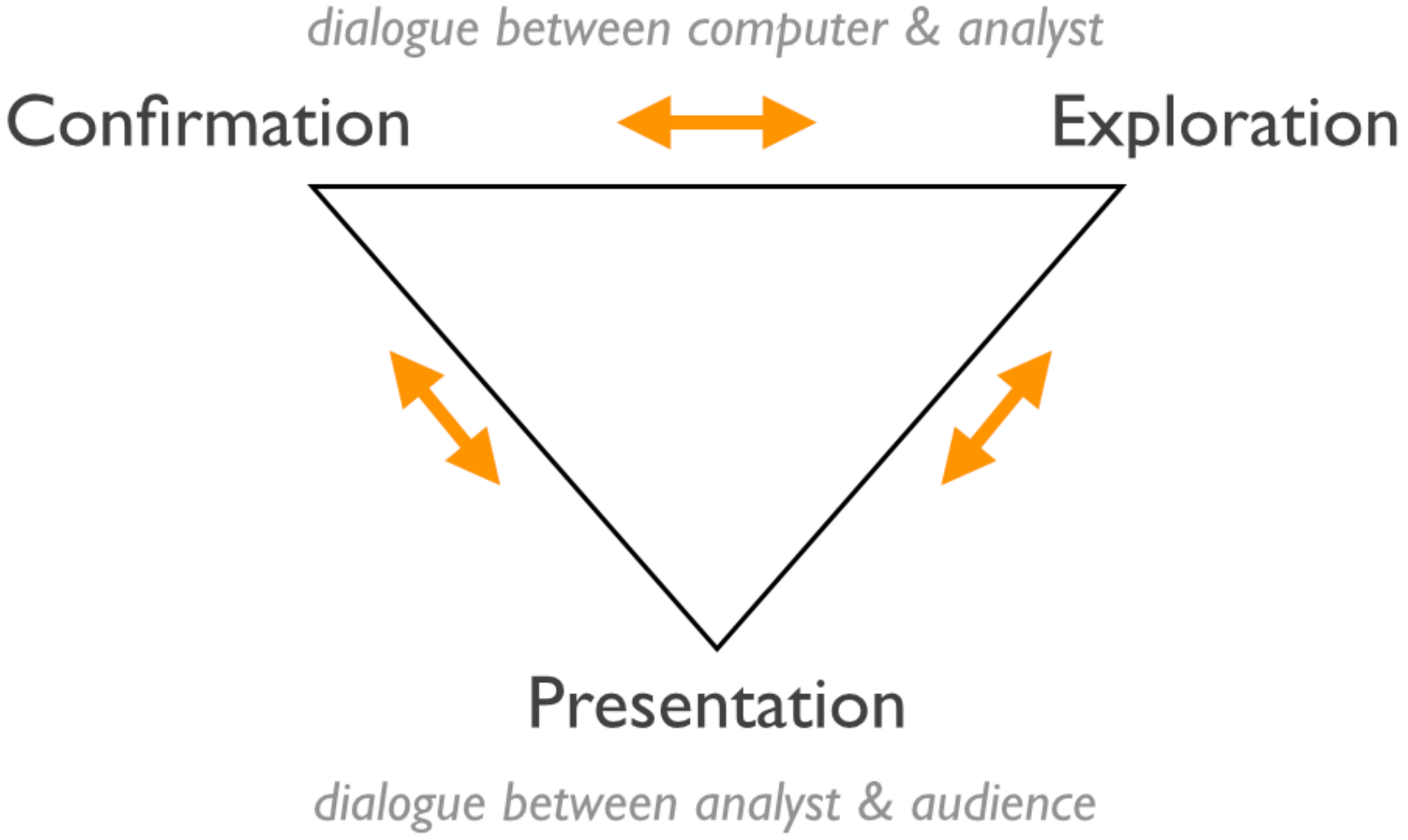
Krzywinski, Nature Methods, 2013

Don’t shout at your audience!



as much data as possible

Insight



as much data as necessary

Data Visualization Tools & Communication

The BD2K Guide to the Fundamentals of Data Science Series

Nils Gehlenborg, PhD
Department of Biomedical Informatics, Harvard Medical School
nils@hms.harvard.edu · http://gehlenborglab.org

BD2K Lecture - 31 March 2017