Principles of Data Visualization 2

Stefan Wyder

July 2016

Universität Zürich[UZH]

URPP Evolution in Action

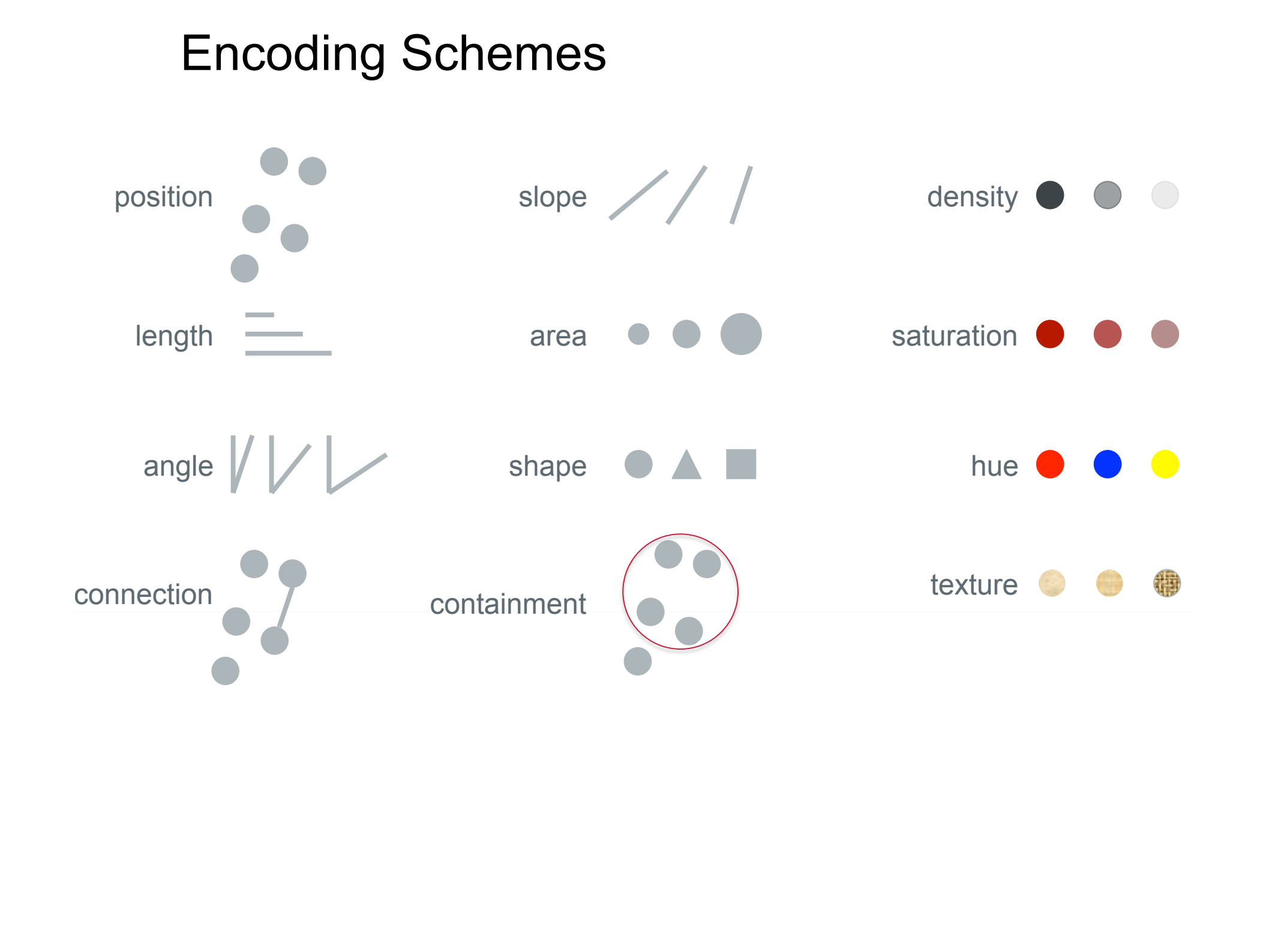
Visualization Goals

- **record** information
- **analyze** data to support reasoning
- **confirm** hypotheses
- **communicate** ideas to others

Outline

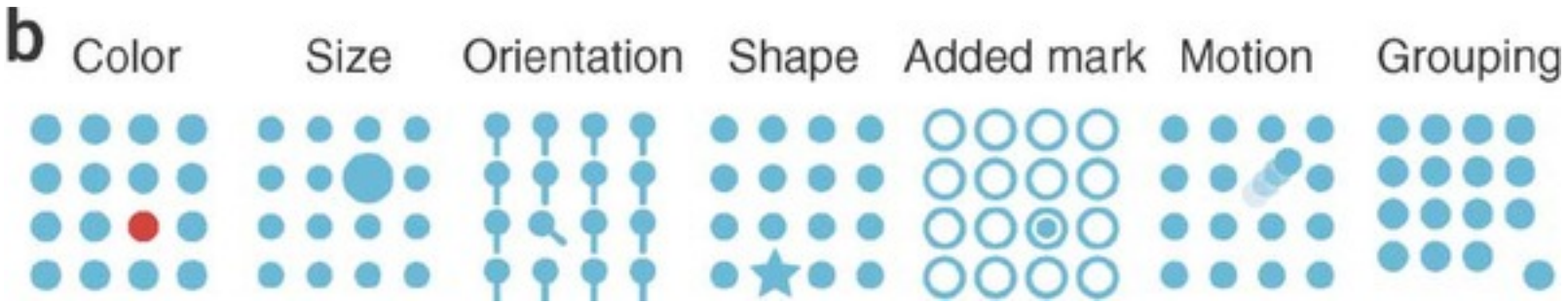
1. The properties of the data or information (HTL)
2. Use of salience, colors, consistency and layout (HTL)
3. The rules mapping data to images (SW)
4. Examples of effective visualizations in biology (SW)
5. Presentation and discussion of “good” and “bad” graphics (HTL & SW)

Encoding Schemes



Don't forget salience ("Hervorstechen")

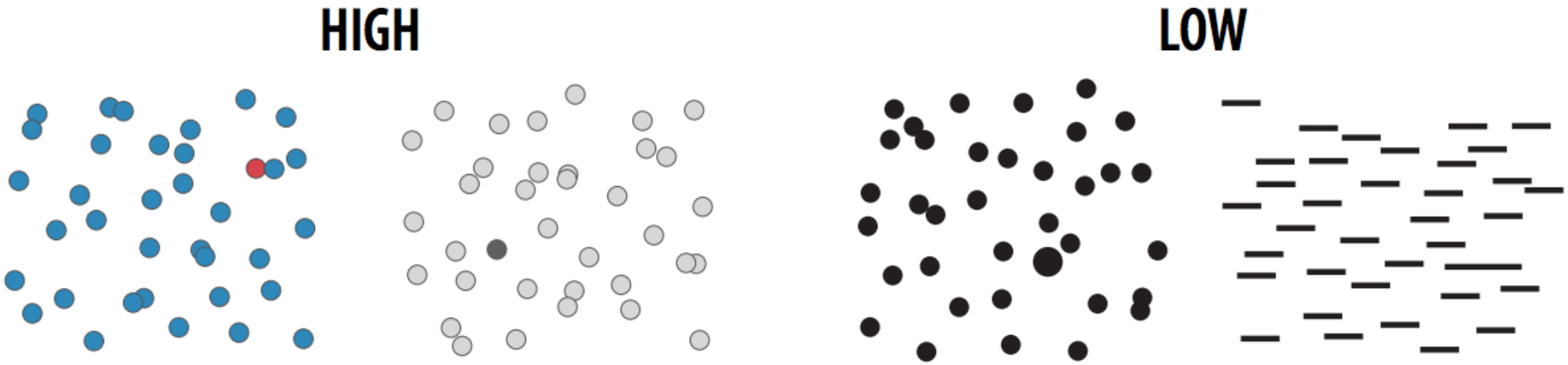
- Distinct features have high salience
- Choose salient encodings for primary navigation



- Focus attention by increasing salience of interesting patterns
  The reader will use salience to suggest what is important
- Context affects salience
- Color is good for categories - salience decreases with more hues/colors

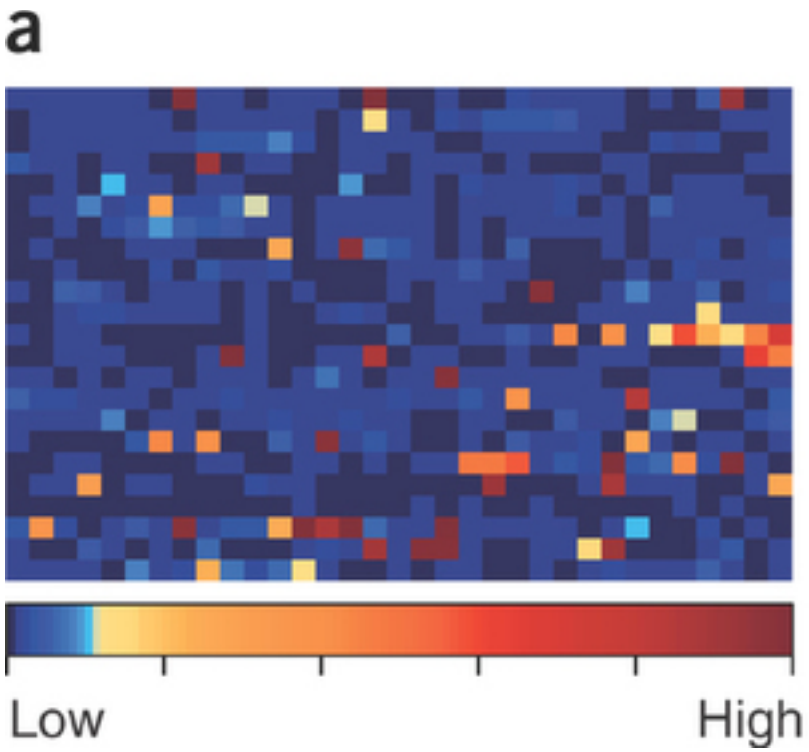
http://www.nature.com/nmeth/journal/v7/n10/full/nmeth1010-773.html

Salience



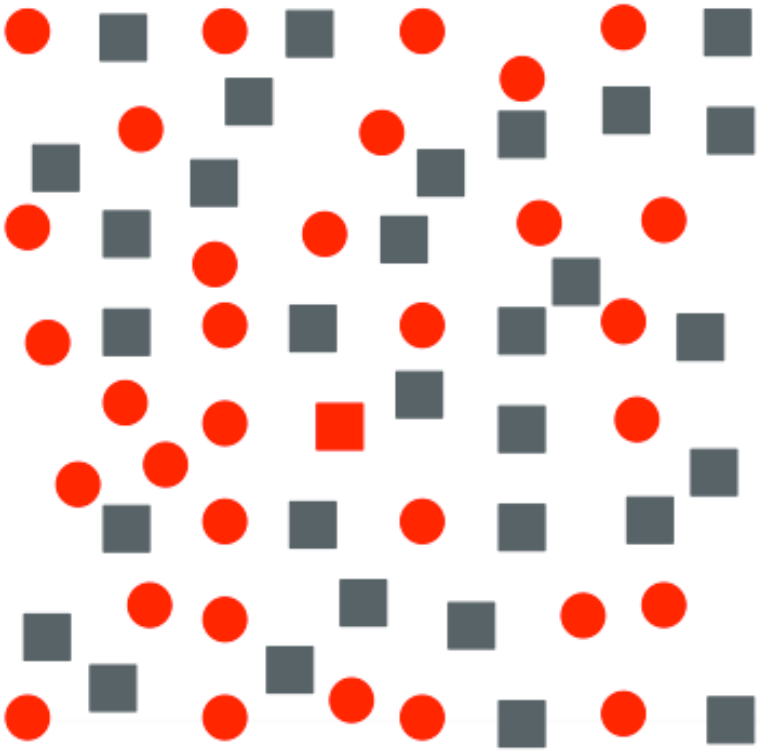
Fecteau JH, Munoz DP (2006) Salience, relevance, and firing: a priority map for target selection. Trends Cogn Sci 10: 382-390.
Yantis S (2005) How visual salience wins the battle for awareness. Nat Neurosci 8: 975-977.

Discordances between salience and relevance



- a color scale that makes common sense
- lower values are actually more salient than higher ones because deep red is hard to see against the deep blue background of the lowest values

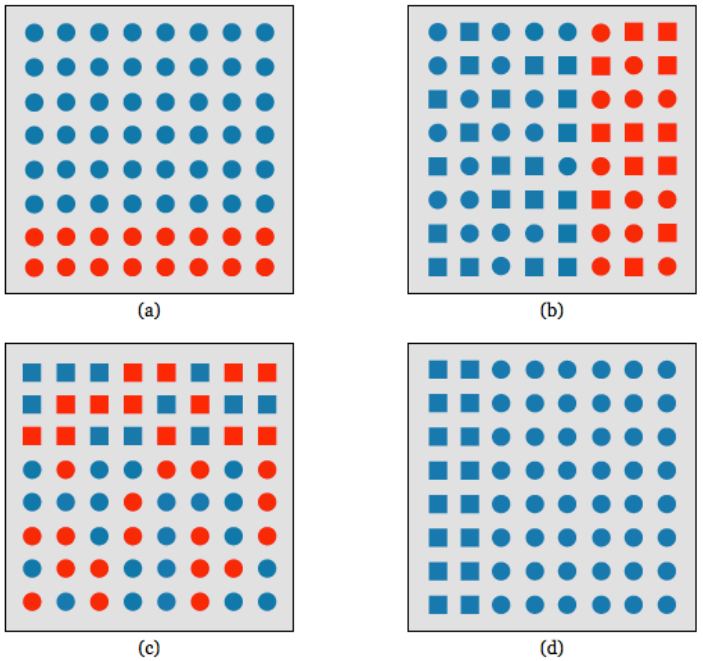
Visual interference



- Spot the red square
- difficult to detect
- serial search required

Katy Borner

Feature Hierarchy in the visual system



- b) random variations in shape have no effect on a viewer's ability to see colour patterns
- c) random variations in color have a strong effect on a viewer's ability to see shape patterns

- Color > Shape
- Interactions between different visual features hide or mask information in a display
- We want to choose a data-feature mapping that does not produce visual interference

https://www.csc.ncsu.edu/faculty/healey/PP/

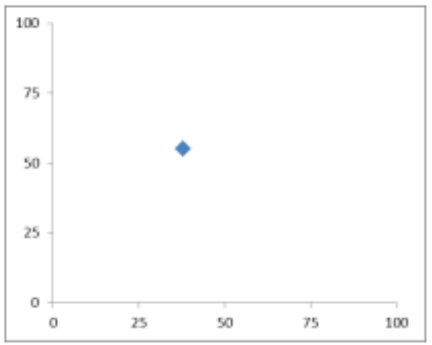
Bertin's Image Theory

- We can only perceive 3 variables (2 planar and 1 retinal) "efficiently" (preattentive, without additional attention)

PLANAR

Spatial dimension 1

Spatial dimension 2



RETINAL

Texture

Color

Shape

Orientation

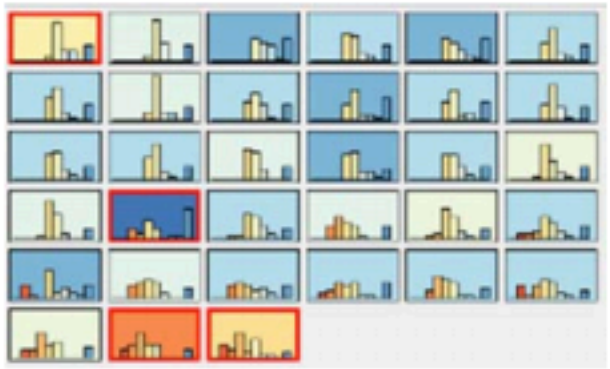
Size

Brightness

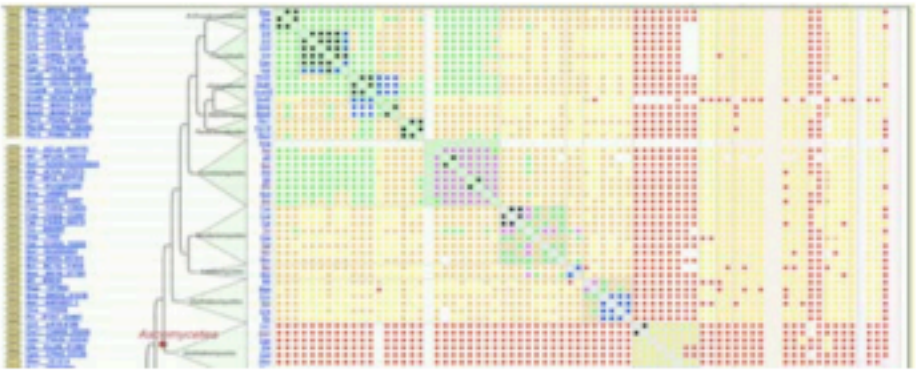
- We can not effectively visualize 4 or more dimensions on a 2-d display

Solution 1

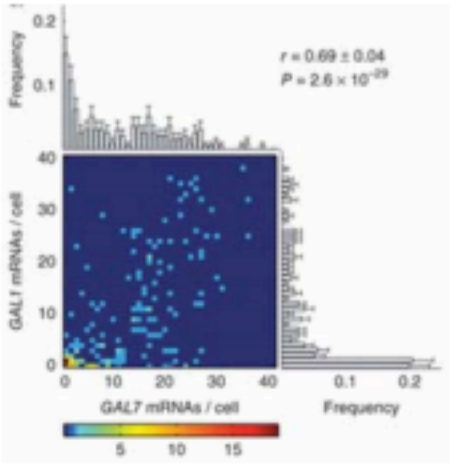
Small Multiples



Multiple (coupled) windows



each view uses the same
visual encodings
but shows a different data set



Katy Borner

Solution 2: Interaction / Operations on the data

Search, filter, select

Zoom, Pan

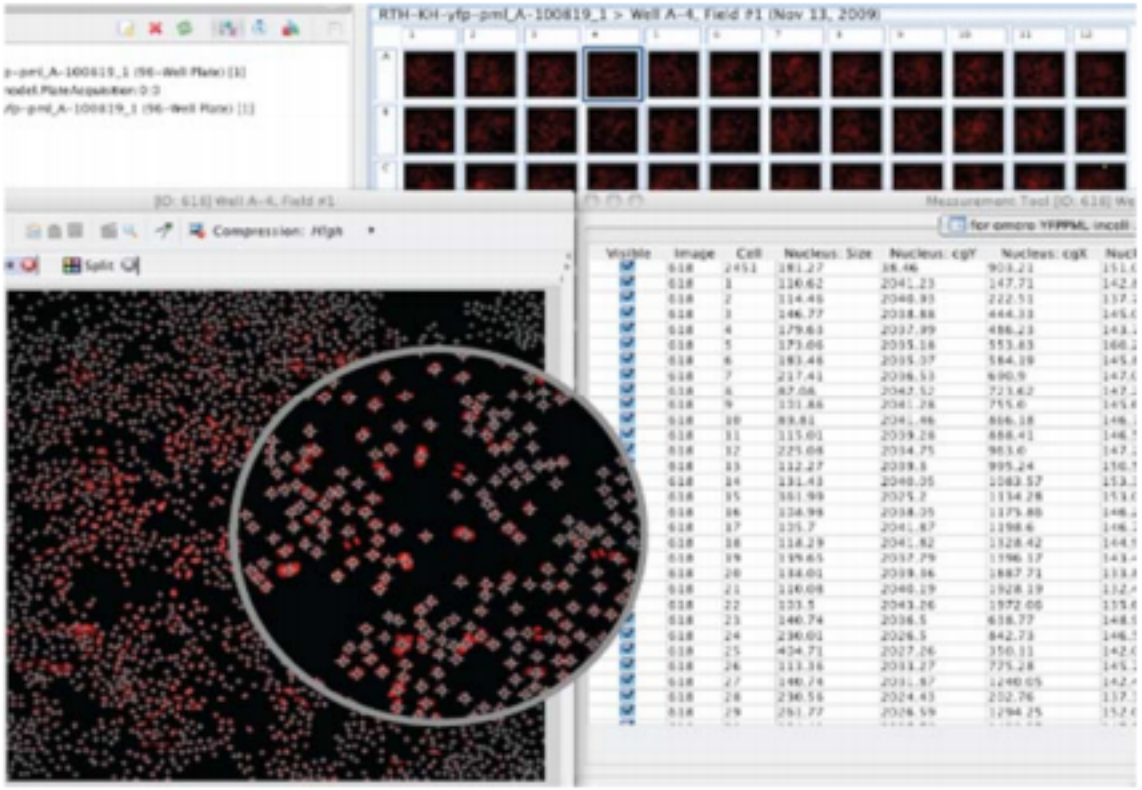
Pruning

Brushing

Details on demand

Focus & context

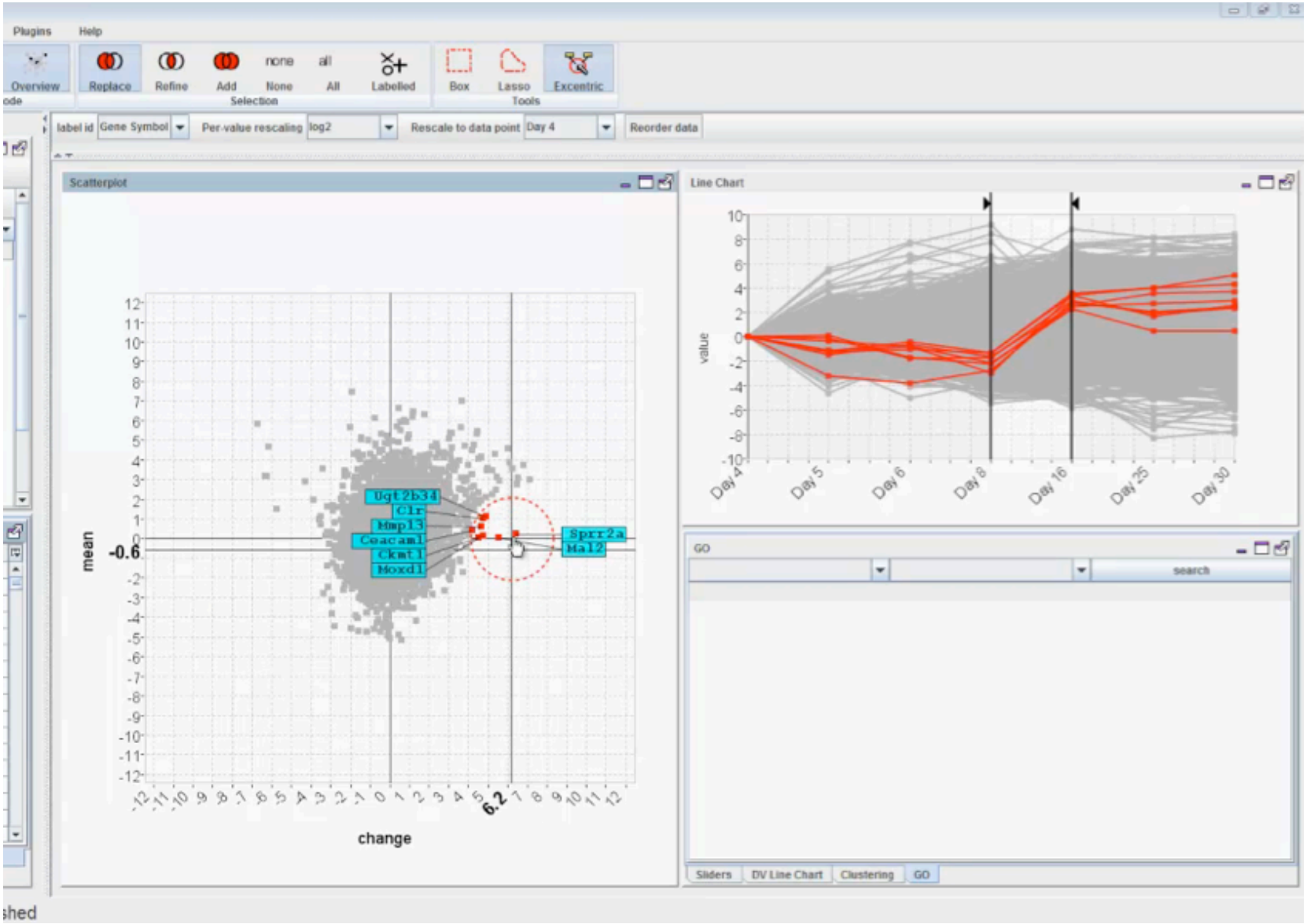
User status



Katy Borner

Solution 3: Interaction / Linked views

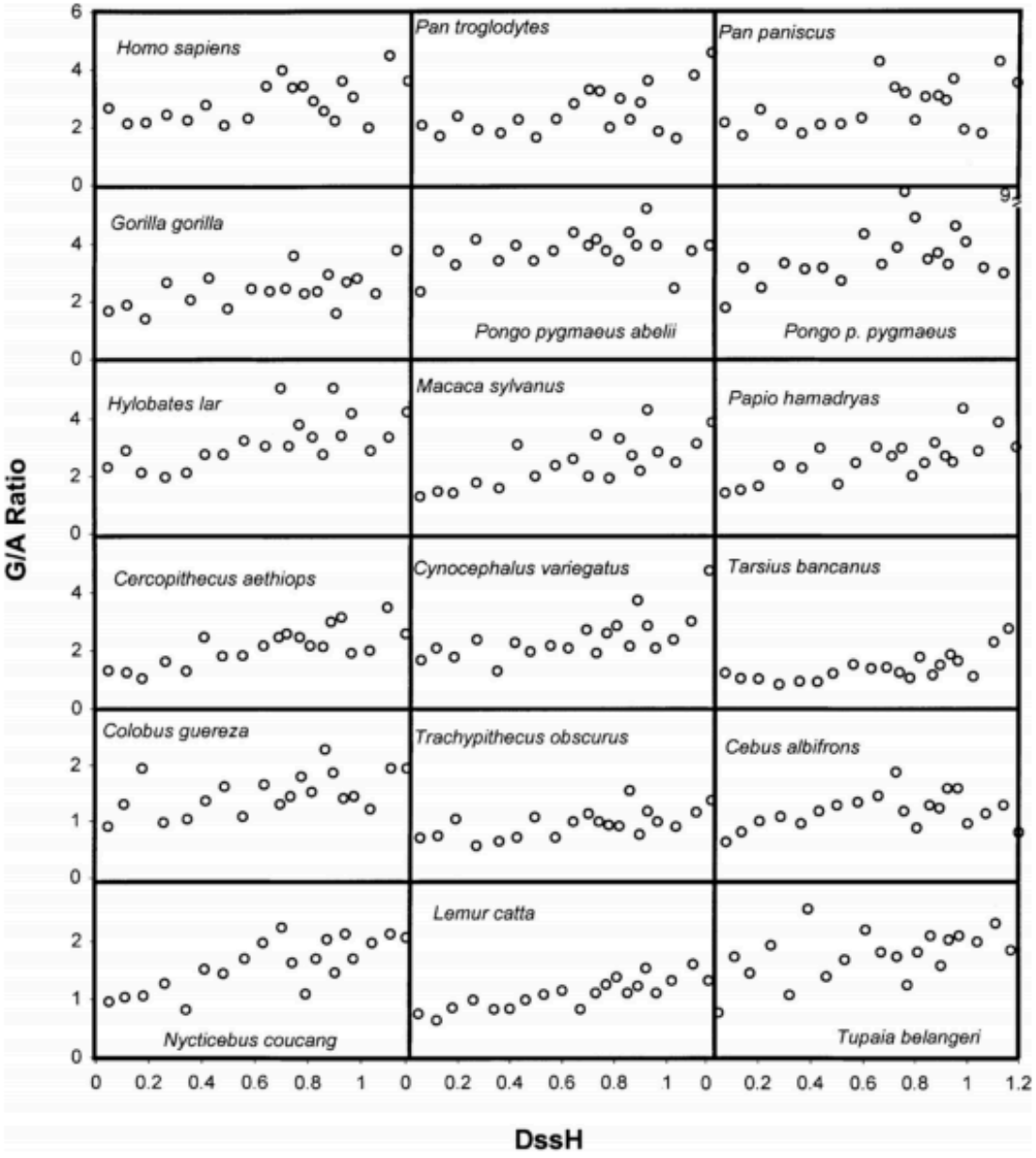
Time-series microarray data



Katy Borner

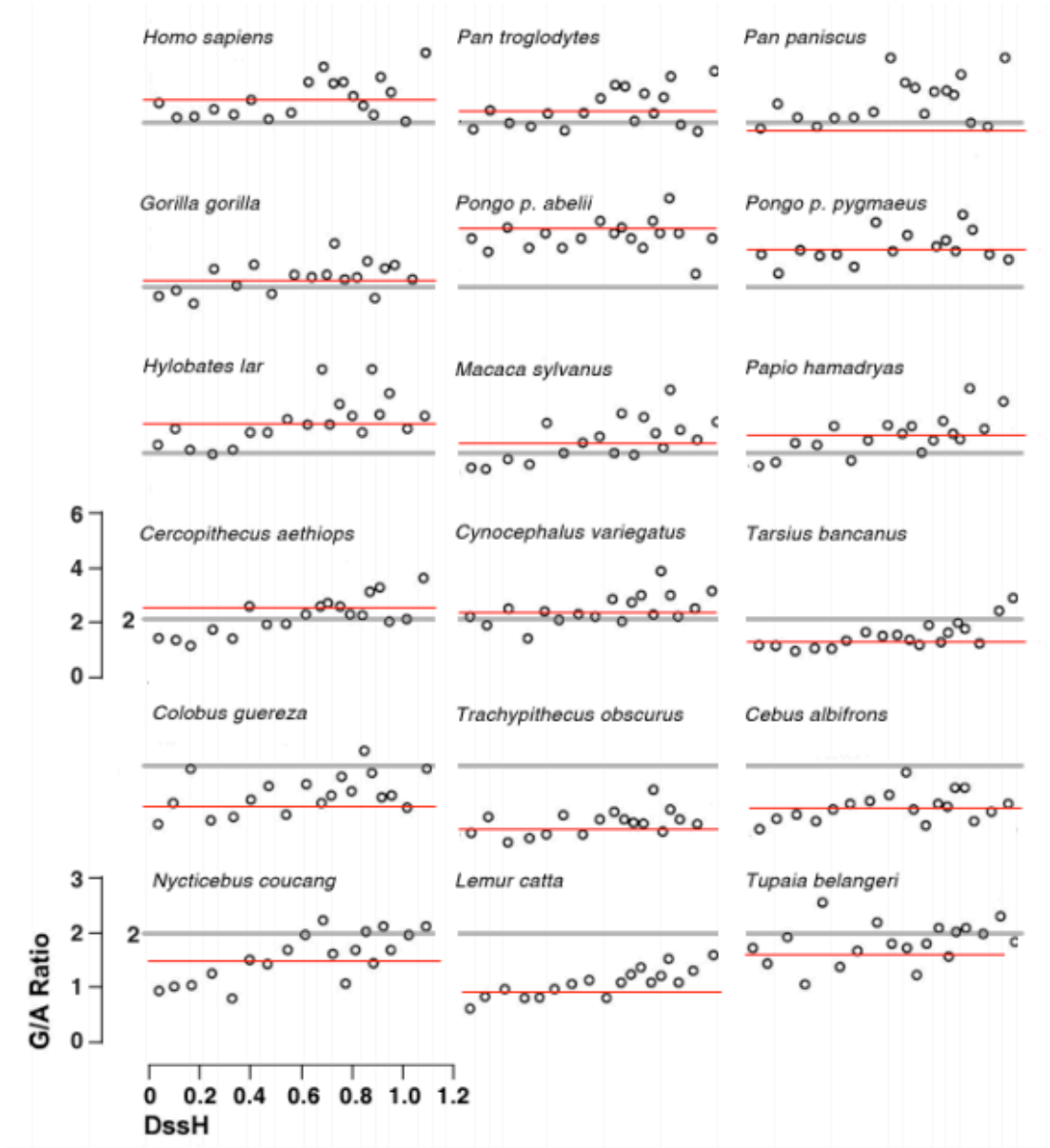
Examples of effective visualizations in biology

Focus on data



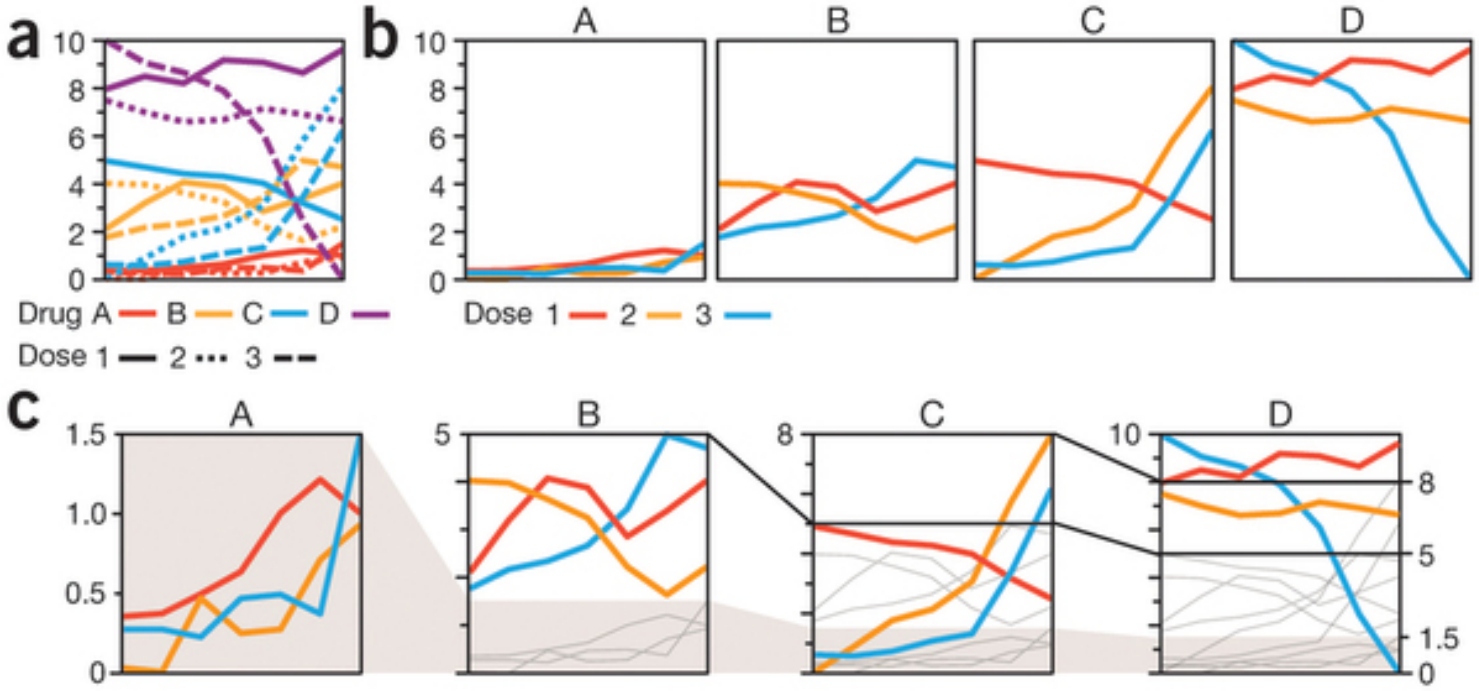
Raina et al. (2005)

Focus on data 2



Better data-to-ink-ratio removal of unnecessary elements

Variation in data range

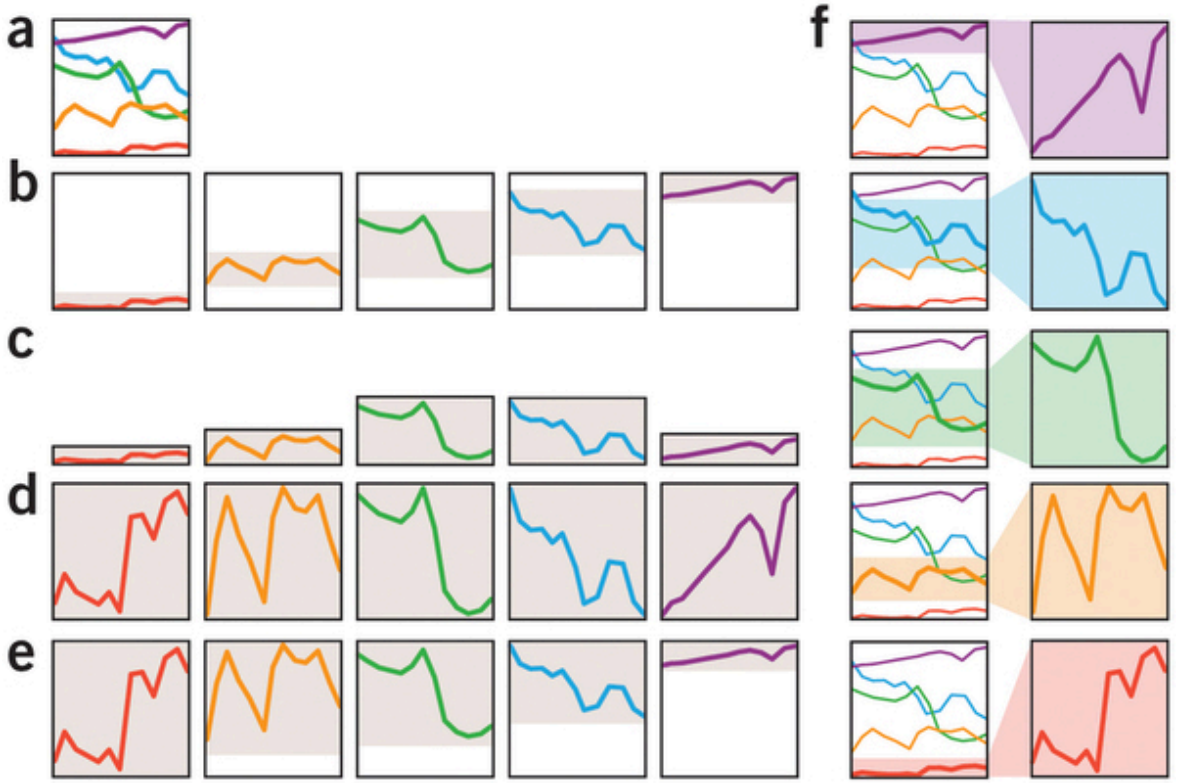


Small multiples / Subplots of time-series data

a) Small-multiple plots isolate and untangle the categories but lose context as categories are separated
b) Subtle scale annotations provide context while maintaining clarity

http://www.nature.com/nmeth/journal/v12/n7/full/nmeth.3451.html

Variation in data range 2

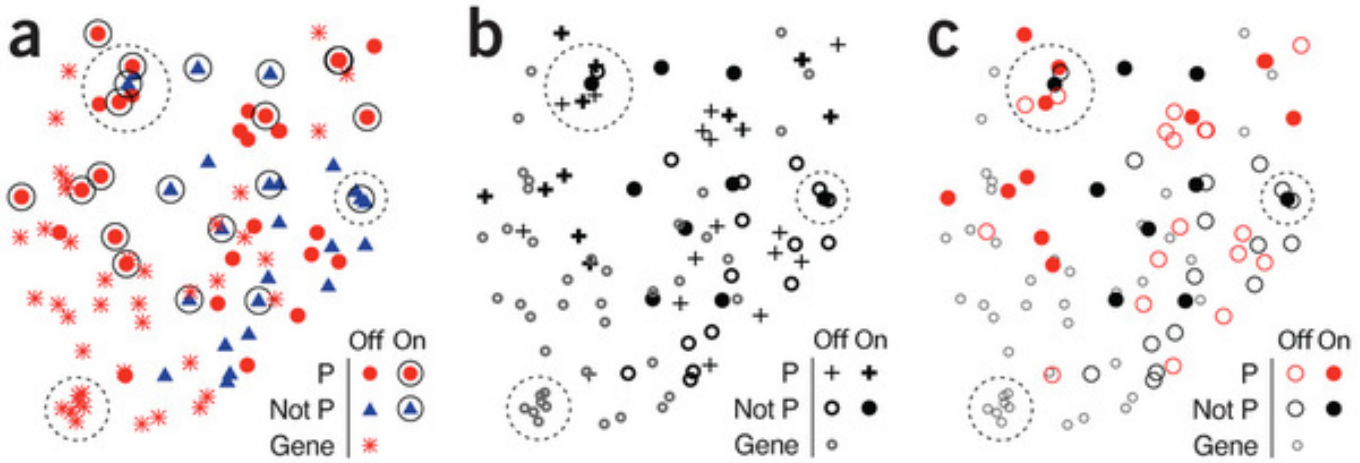


(f) Use an overview and scaled detail to contextualize, highlight and examine each category. Colored backgrounds emphasize differences in scale expansion

http://www.nature.com/nmeth/journal/v12/n7/full/nmeth.3451.html

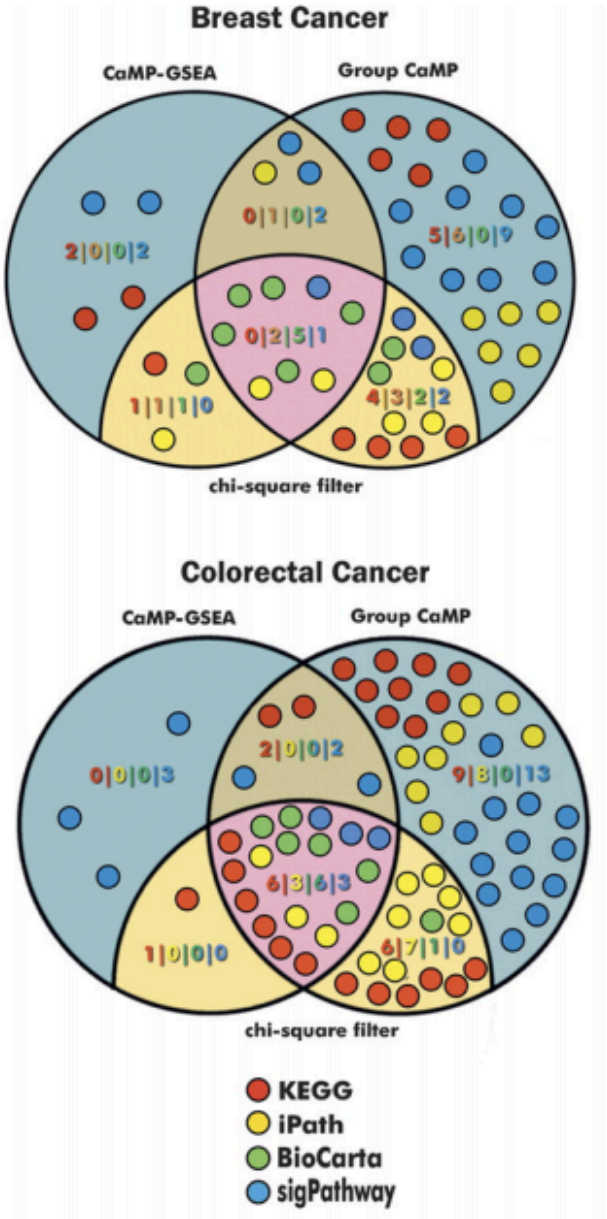
Good legends

Natural hierarchy. By varying shape and color meaningfully, the encoding becomes more memorable

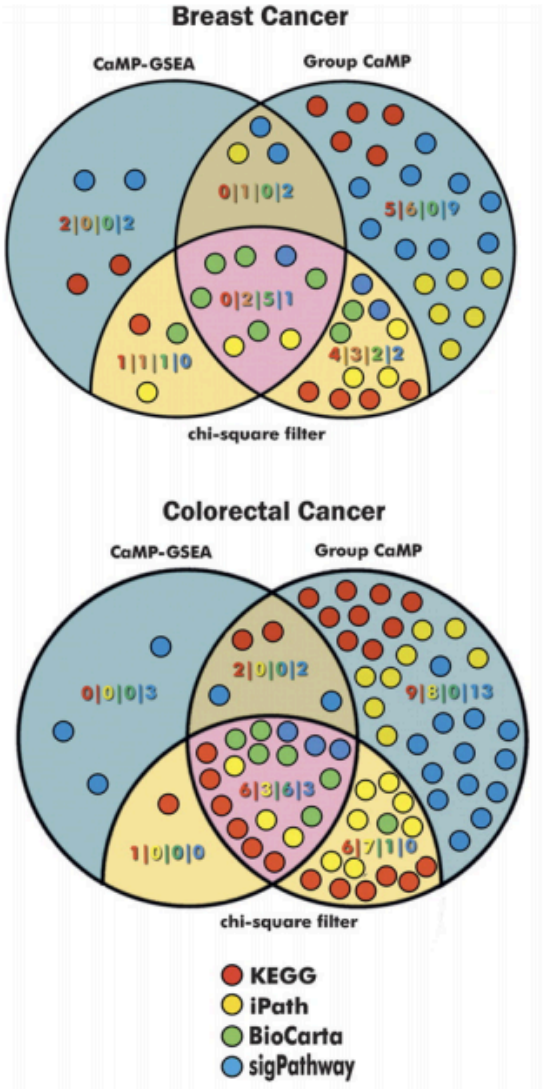


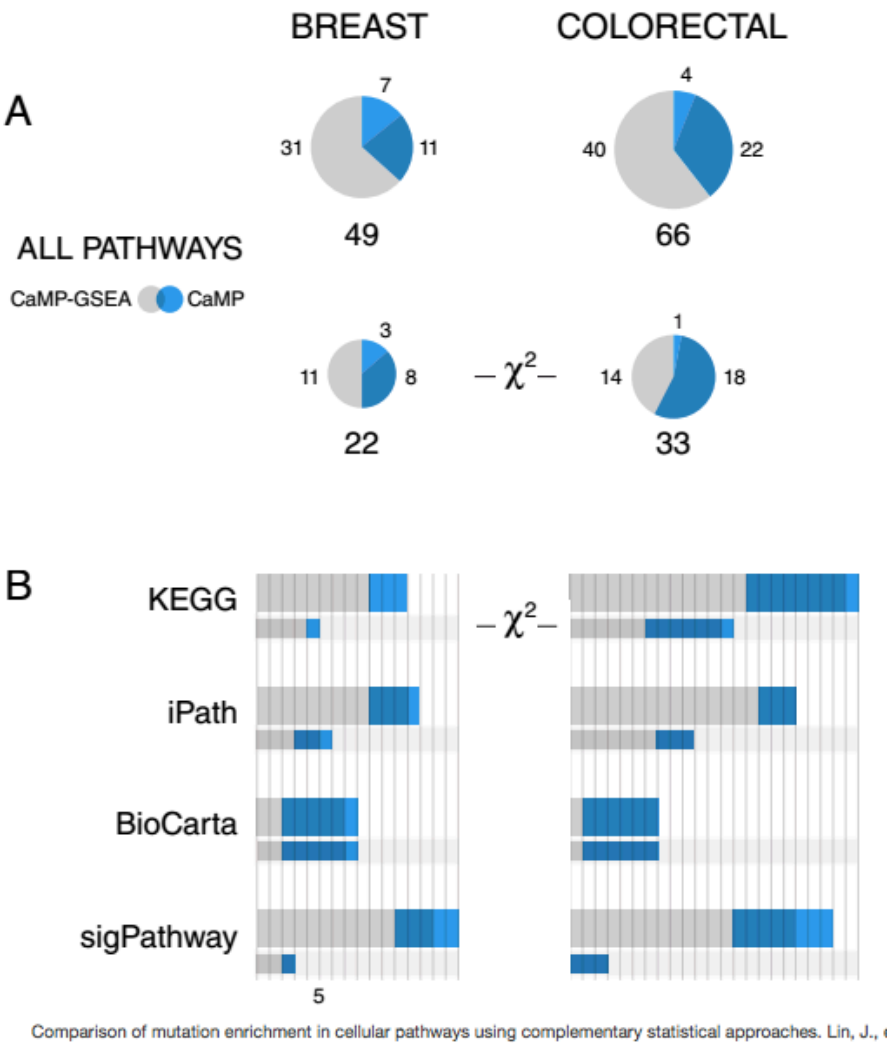
http://www.nature.com/nmeth/journal/v10/n6/full/nmeth.2490.html

Refactoring Complexity



Refactoring Complexity 2

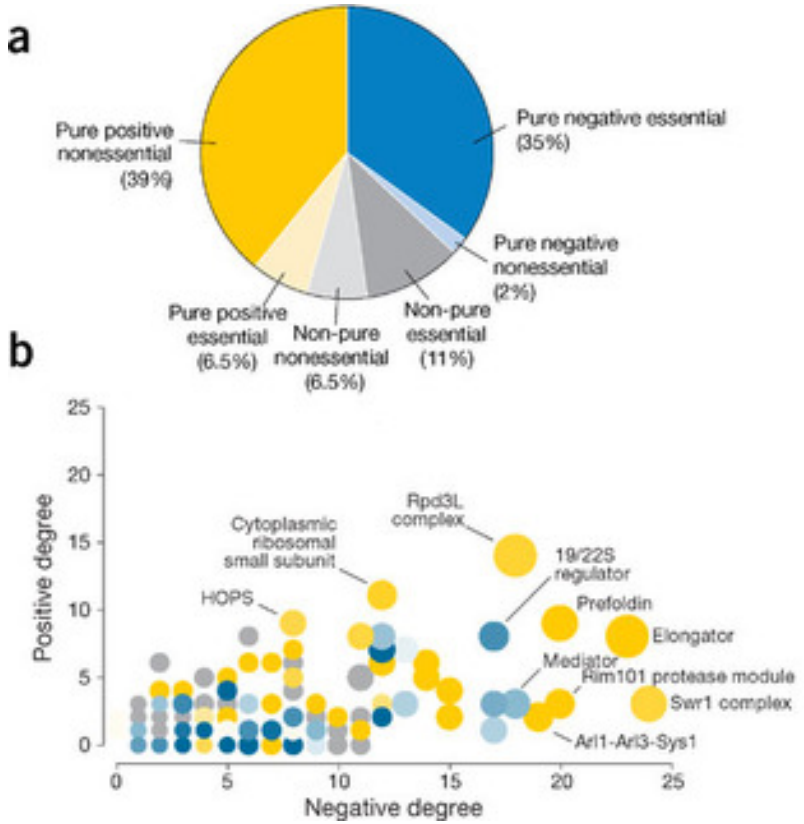


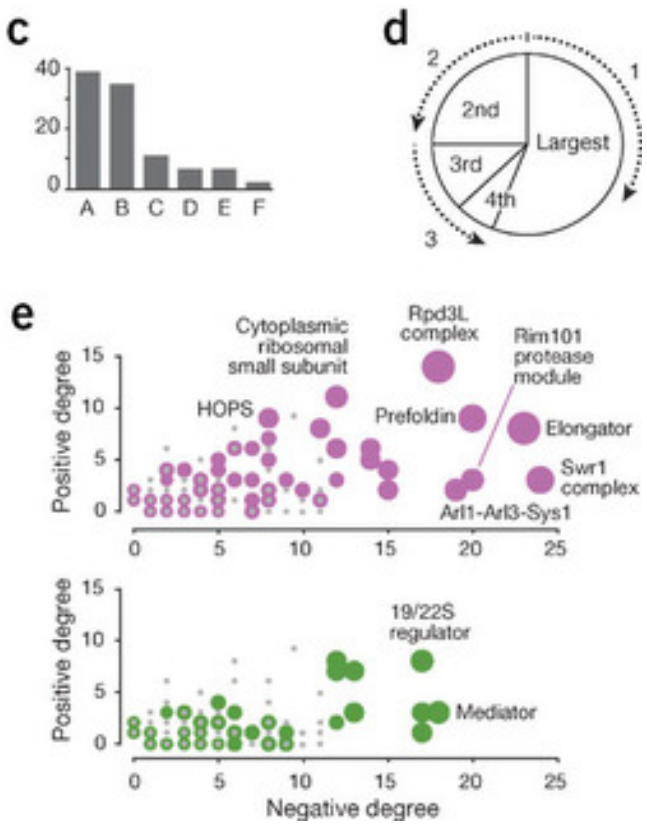


Comparison of mutation enrichment in cellular pathways using complementary statistical approaches. Lin, J., (

Message is clearer
(Breast > Colorectal, KEGG > sigPathway > iPath > BioCarta)

Refactoring Complexity 2



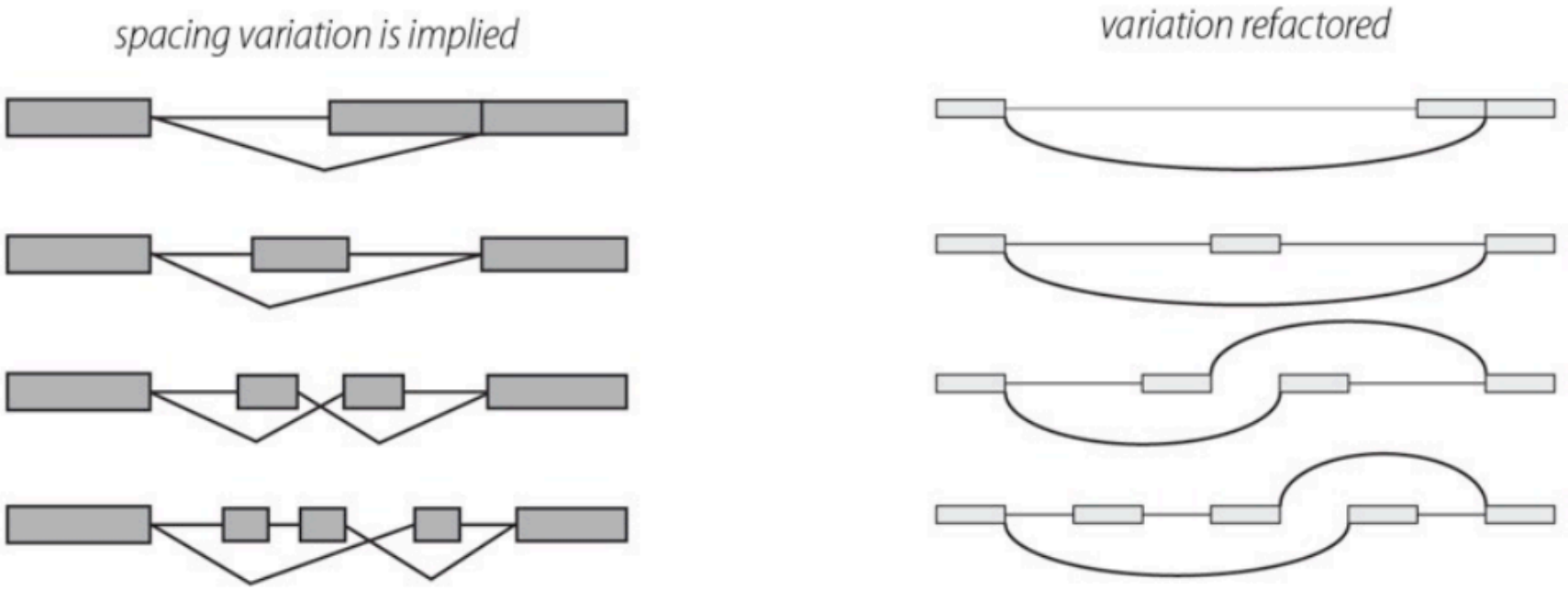


Busy / difficult to read graph
8 point sizes, 11 shades of yellow/ 13 blue

Reduce visual complexity
Limit the color value and size scales (0–3, 4–7 and others)

http://www.nature.com/nmeth/journal/v8/n3/full/nmeth0311-189.html

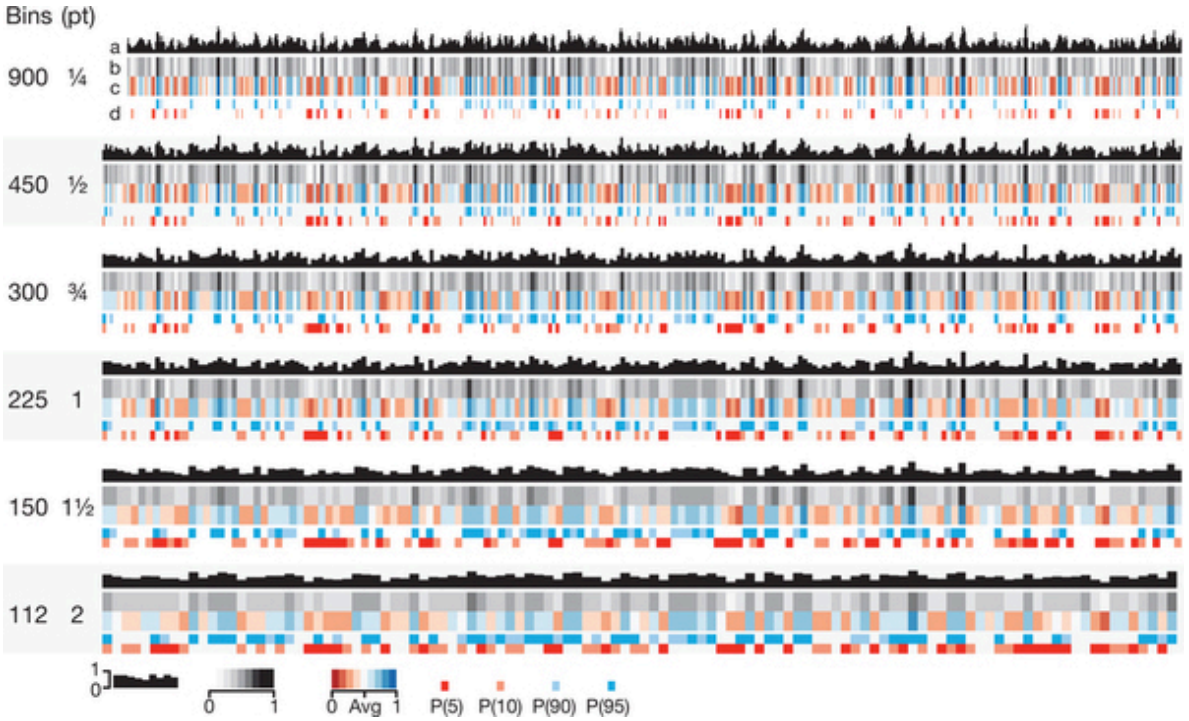
Uniform spacing and sizing



Sharov et al. (2005)

- Keep the size, spacing and alignment fixed of as many elements as possible
- Any variation in the figure will be interpreted as important to its message

Binning high-resolution data

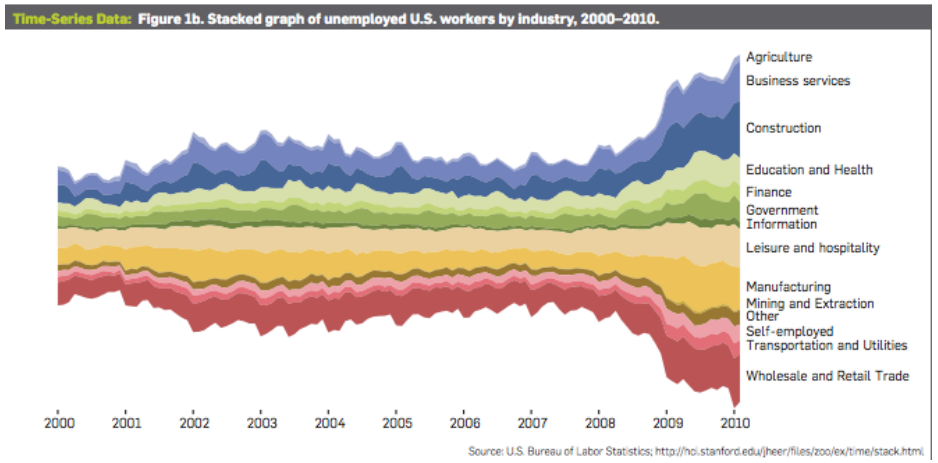


Read coverage of simulated sequencing
(a) histogram, (b) heatmap
(c) Coverage relative to the average
(d) Bins with values at least as extreme as the 5th, 10th, 90th or 95th percentile are marked

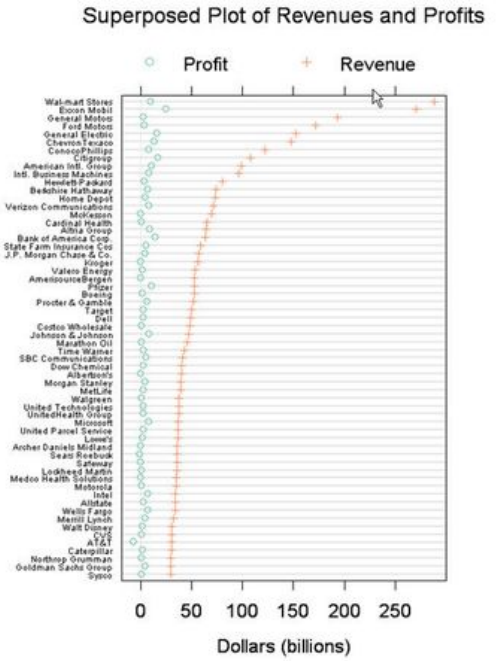
- Lines thinner than 1/2 pt cannot be comfortably resolved if less than 1/2 pt apart
- Finding local maxima is relatively easy even with 1/4-pt bins, but judging the average, assessing variability and discerning minima are difficult with bins smaller than 1 pt
- We suggest not binning data into more than ~250 intervals for one-column figures (3.5 inches wide) or ~500 intervals for two-column figures (7.2 inches)

http://www.nature.com/nmeth/journal/v13/n6/full/nmeth.3873.html

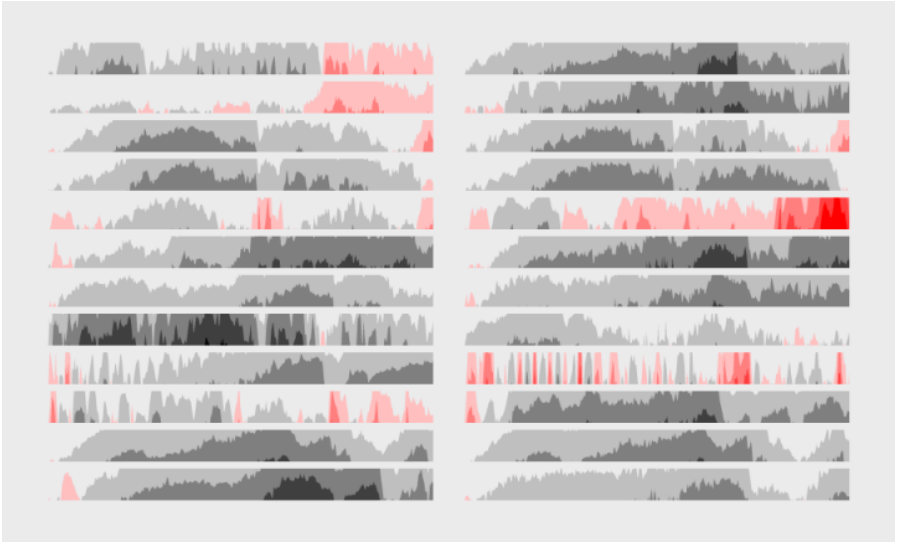
Many other types of charts exist



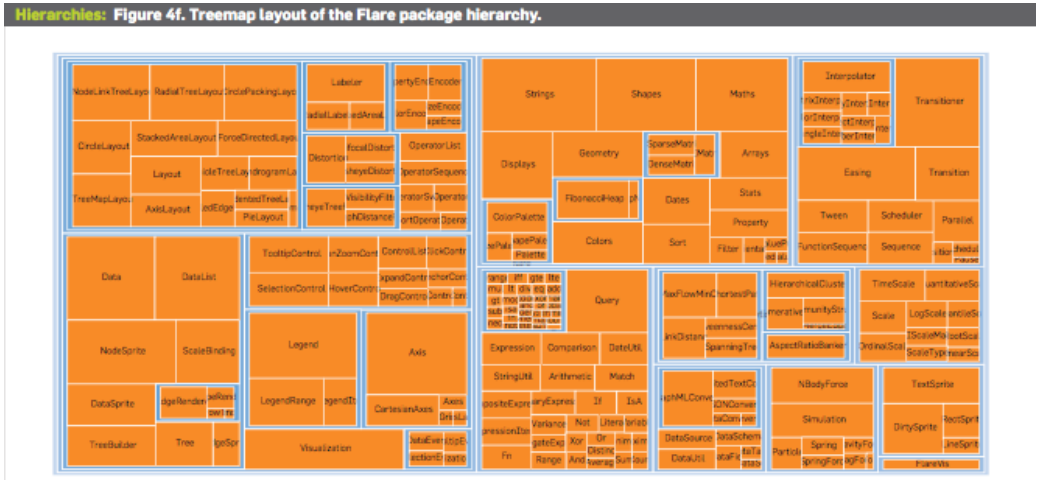
Stacked graph (controversial)



Dotplot



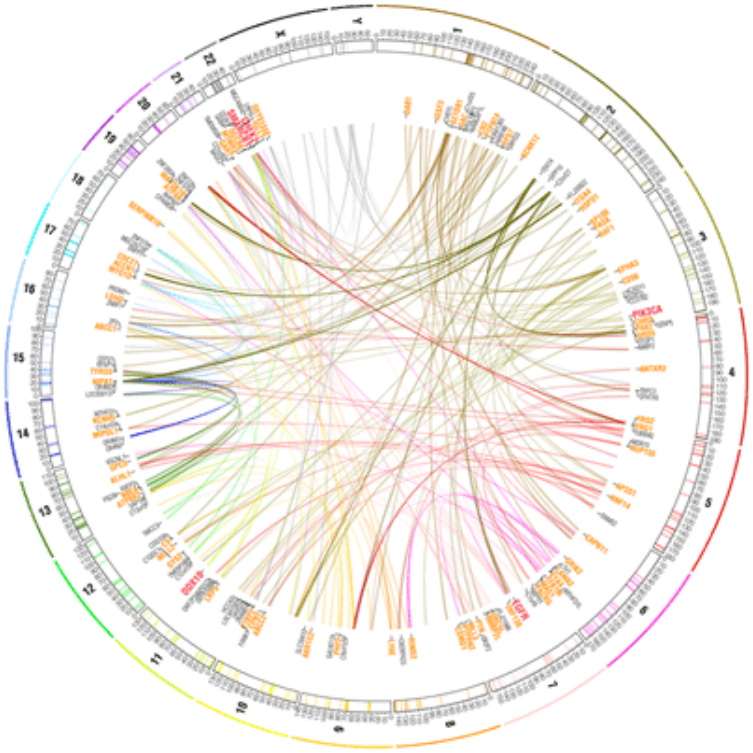
Horizon graph
(flowing data)

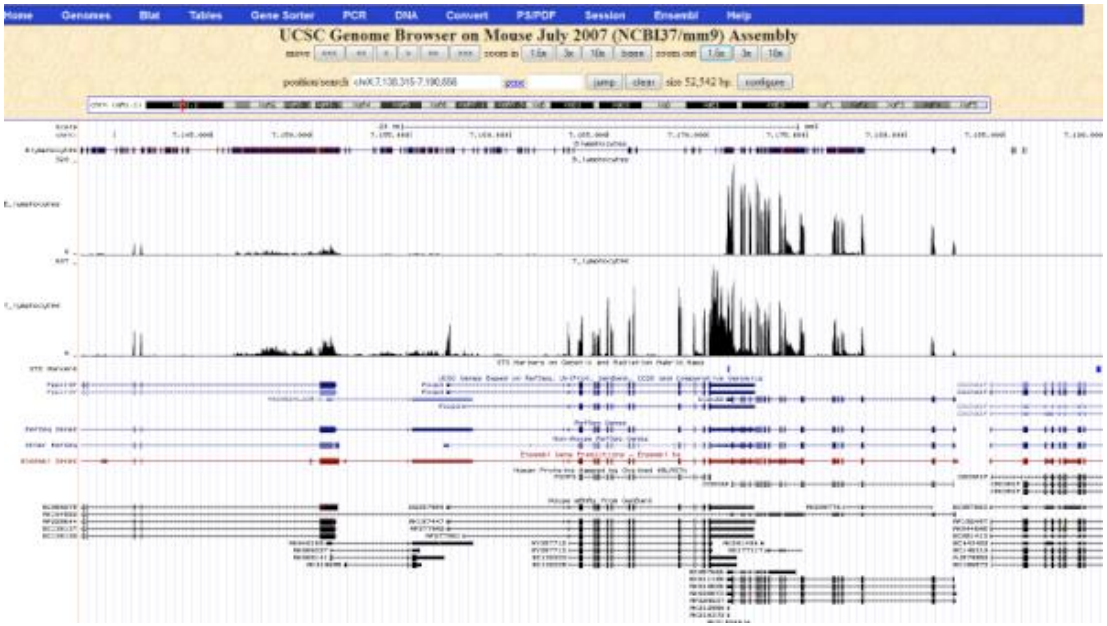


Treemap

Visualization zoo

Previous URPP tutorial





Some tools to visualize biological data (ensembl/UCSC genome browsers, IGV, circos) have been presented in a previous URPP tutorial https://github.com/milchmolch/Genomic_Visualization

Tufte's design principles

- maximize the data-ink ratio
- avoid chart junk (sometimes)
- use multifunctioning elements
- separate layers
- maximize the data density
  shrink the graphics
  maximize the amount of data shown (sometimes)

- Show data variation, not design variation

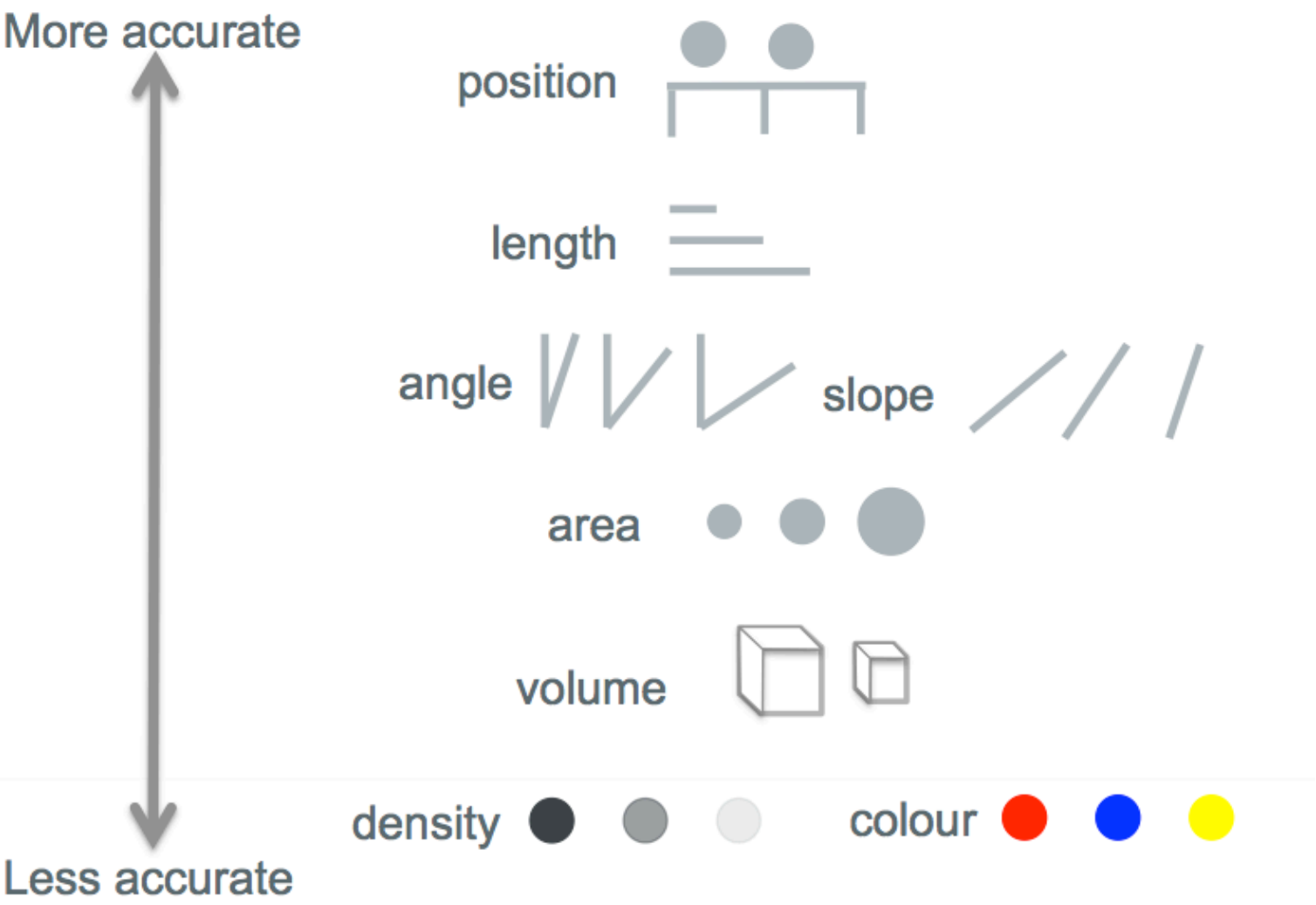
Take home message (I try)

- Maximize the data-to-ink-ratio
- Show the data - show individual data points if possible
- Reduce the complexity
  only 6-12 colors are visually discernable
  Use small multiples if more than 6-7 categories
- Display uncertainty
- Use transparency to improve clarity
- Do not trust the R defaults

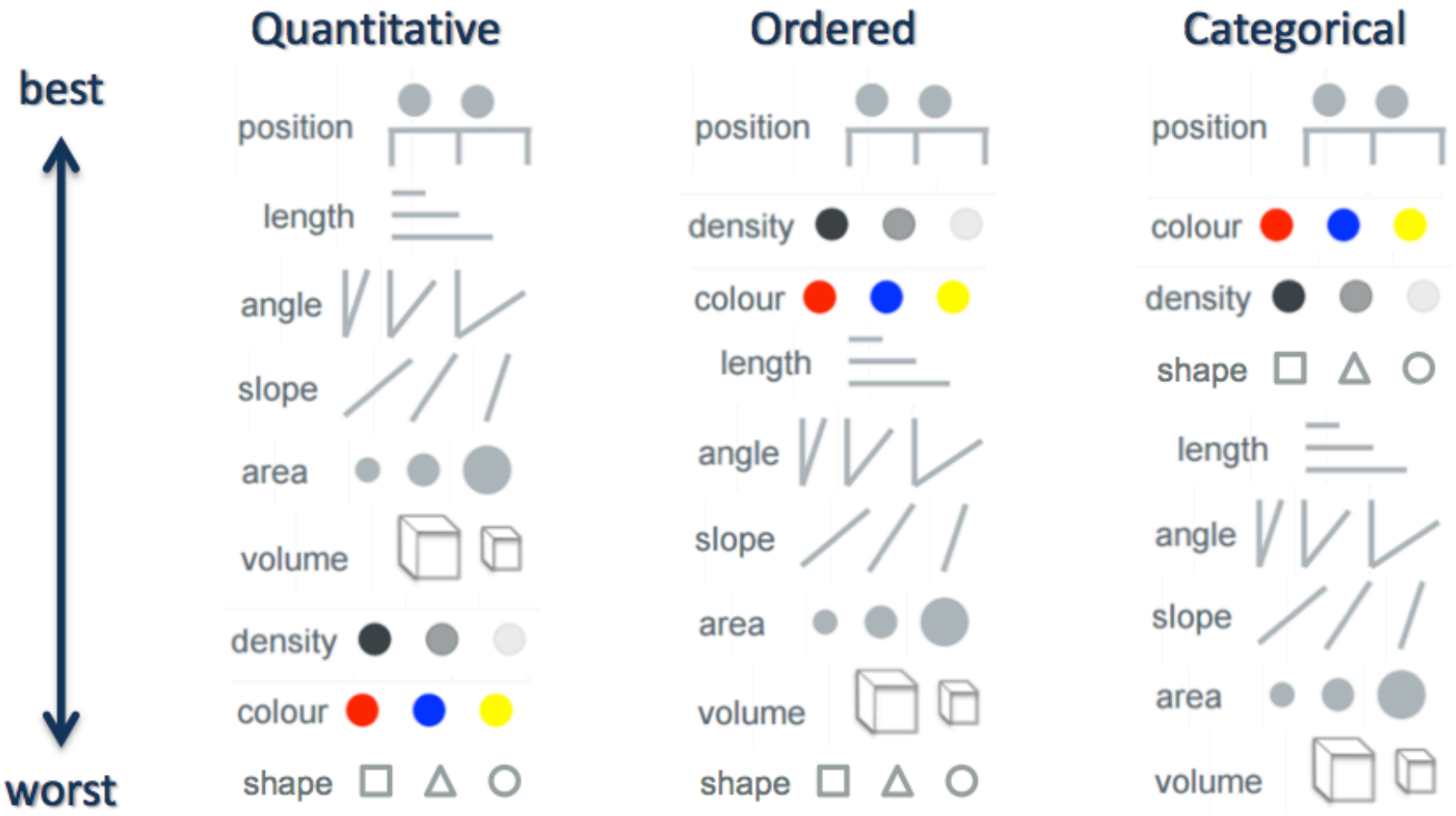
Sources

- http://mkweb.bcgsc.ca/vizbi/2012/principles.pdf
- Points of view http://clearscience.info/wp/?p=546 column on data visualization in Nature method

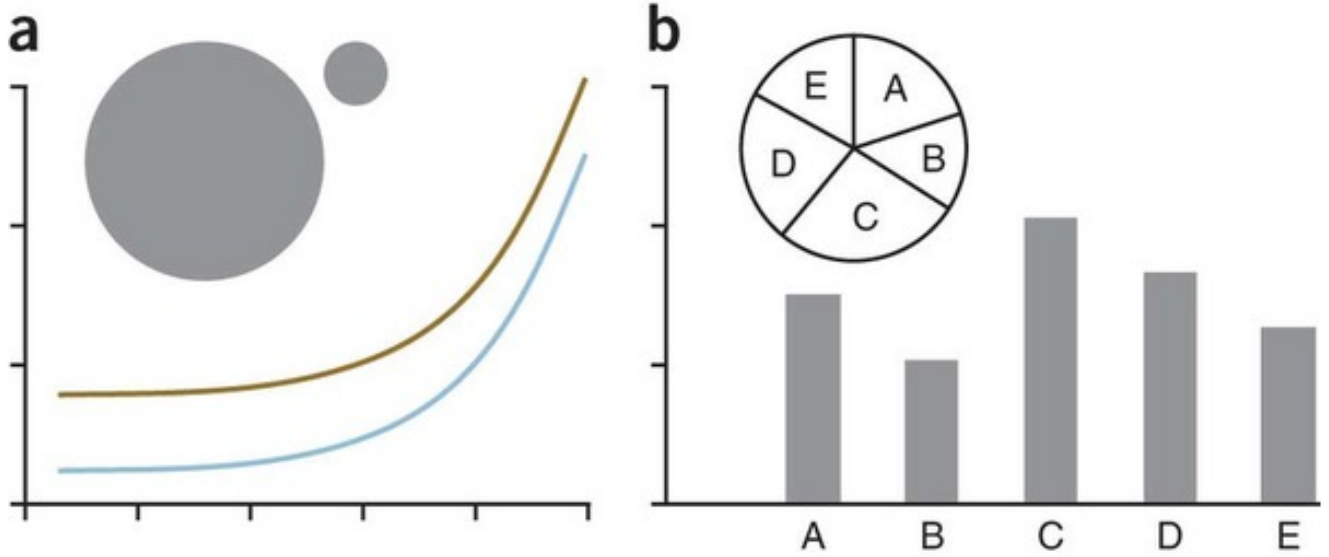
Accuracy of Quantitative Perceptual Tasks



Accuracy of Perceptual Tasks



Accuracy of visual tasks

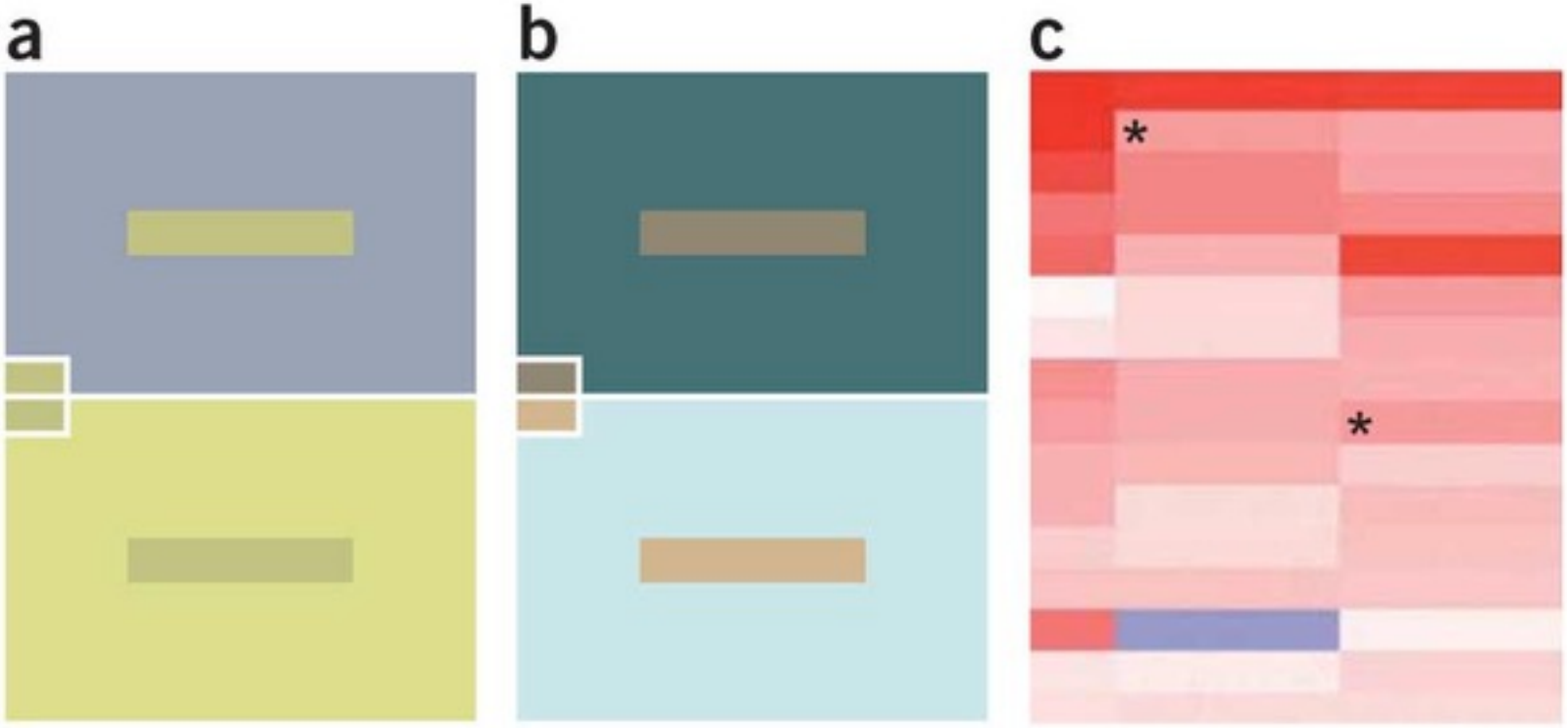


a) Judgeing relative areas is very difficult: 14x difference between circles Constant disparity between curves!
b) To read a pie diagram, we can read angle, areas or arc lengths

Each of these perceptual tasks ranks low in efficiency and accuracy

(Points of View)

Neighboring colors can affect visual perception



(a) The same color can look different
(b) Different colors can appear to be nearly the same by changing the background color
(c) The rectangles in the heat map indicated by the asterisks (*) are the same color but appear to be different

http://www.nature.com/nmeth/journal/v7/n8/full/nmeth0810-573.html

Tufte's Views on Graphical Excellence

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.