

Principles of Data Visualization 2

Stefan Wyder

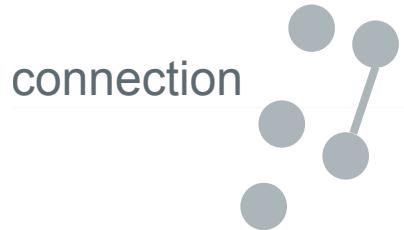
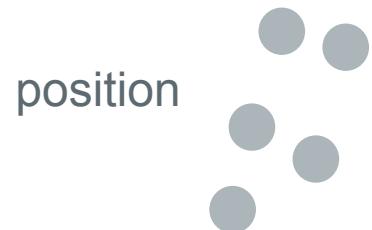
July 2017



Outline

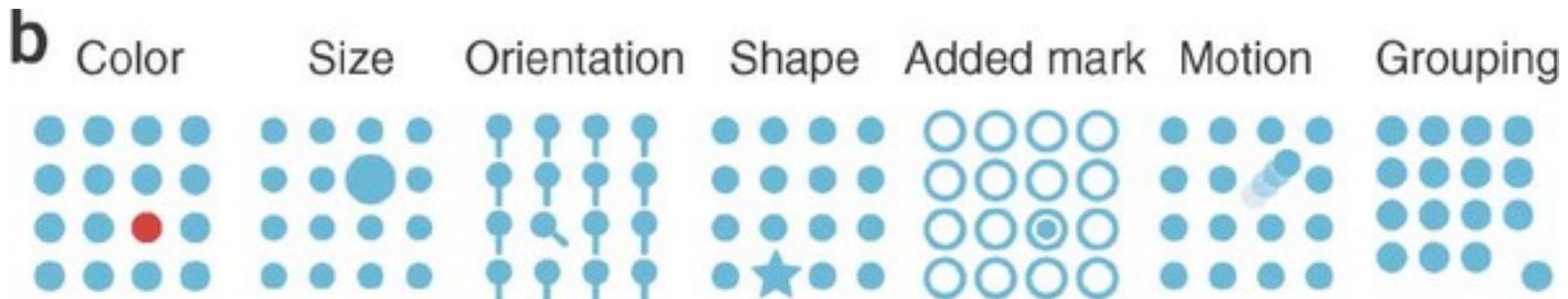
1. The properties of the data or information (HTL)
2. Use of salience, colors, consistency and layout (HTL)
3. The rules mapping data to images (SW)
4. Examples of effective visualizations in biology (SW)
5. Presentation and discussion of “good” and “bad” graphics (HTL & SW)

Encoding Schemes



Salience ("Hervorstechen")

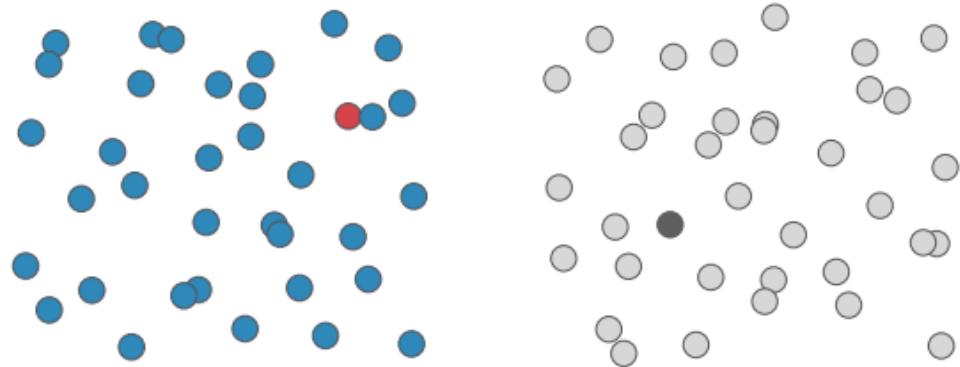
- Distinct features have high salience
- Choose salient encodings for primary navigation



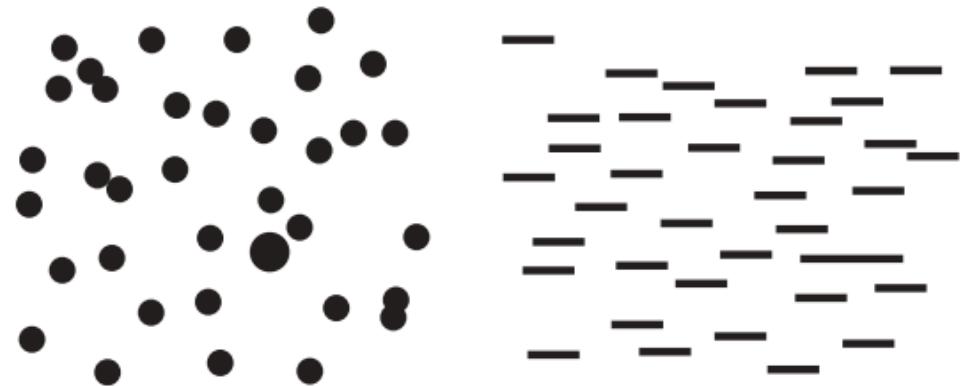
- Focus attention by increasing salience of interesting patterns
The reader will use salience to suggest what is important
- Context affects salience
- Color is good for categories - salience decreases with more hues/colors

Salience

HIGH



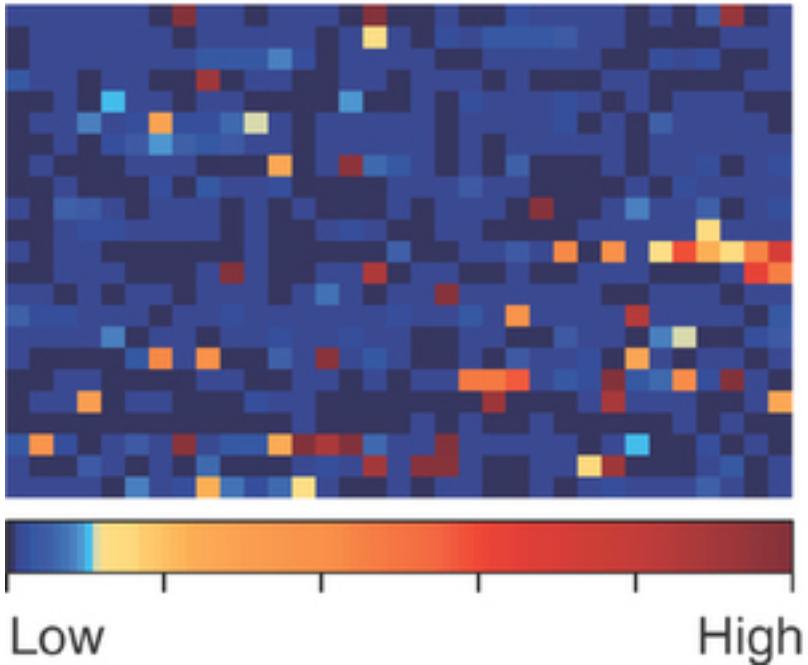
LOW



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.

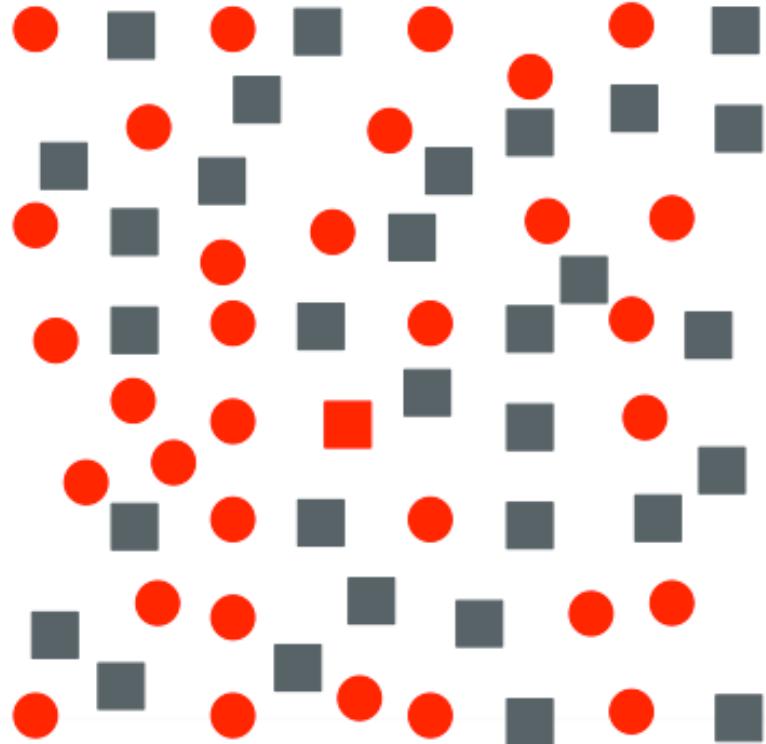
Discordances between salience and relevance

a



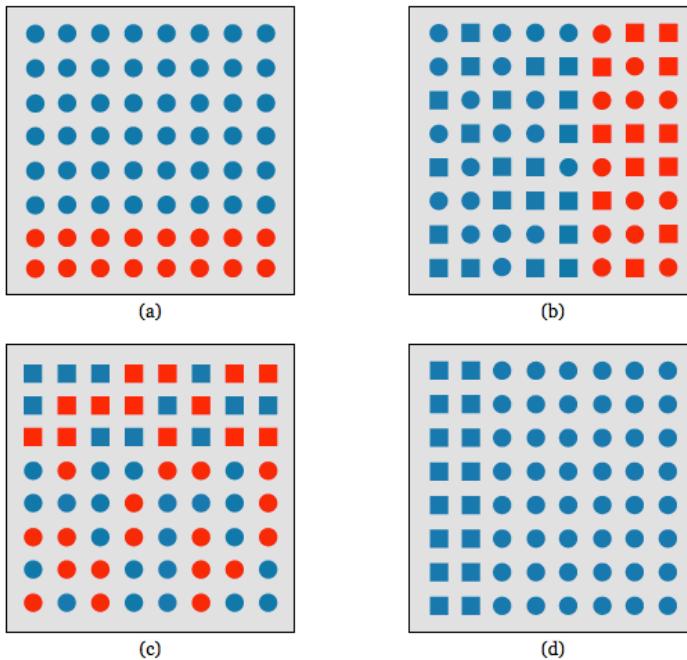
- a color scale that makes common sense
- lower values are actually more salient than higher ones because deep red is hard to see against the deep blue background of the lowest values

Visual interference



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

Feature Hierarchy in the visual system



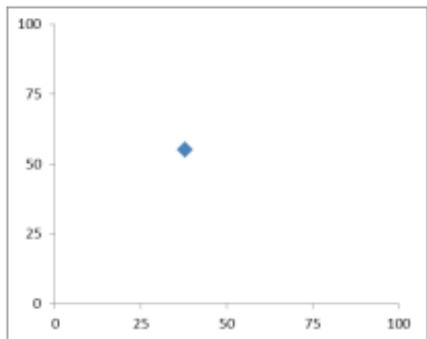
- b) random variations in shape have no effect on a viewer's ability to see colour patterns
- c) random variations in color have a strong effect on a viewer's ability to see shape patterns
- Color > Shape
- Interactions between different visual features hide or mask information in a display
- We want to choose a data-feature mapping that does not produce visual interference

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



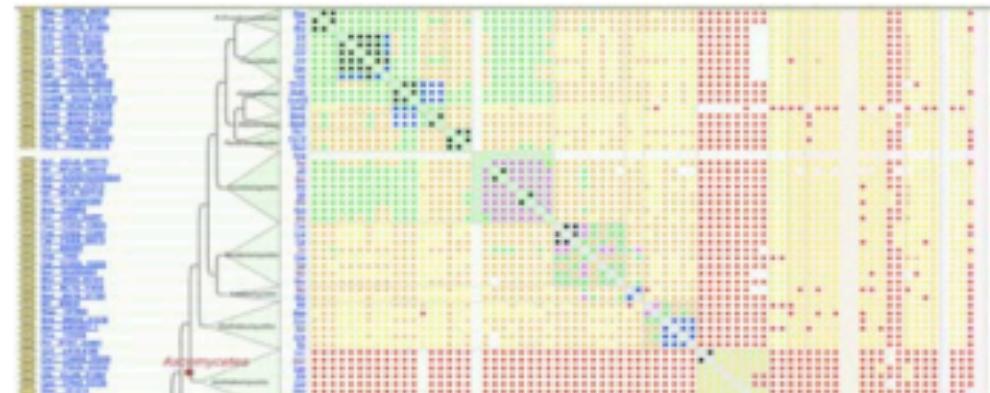
- We can not effectively visualize 4 or more dimensions on a 2-d display

Solution 1

Small Multiples



Multiple (coupled) windows



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

Solution 2: Interaction / Operations on the data

Search, filter, select

Zoom, Pan

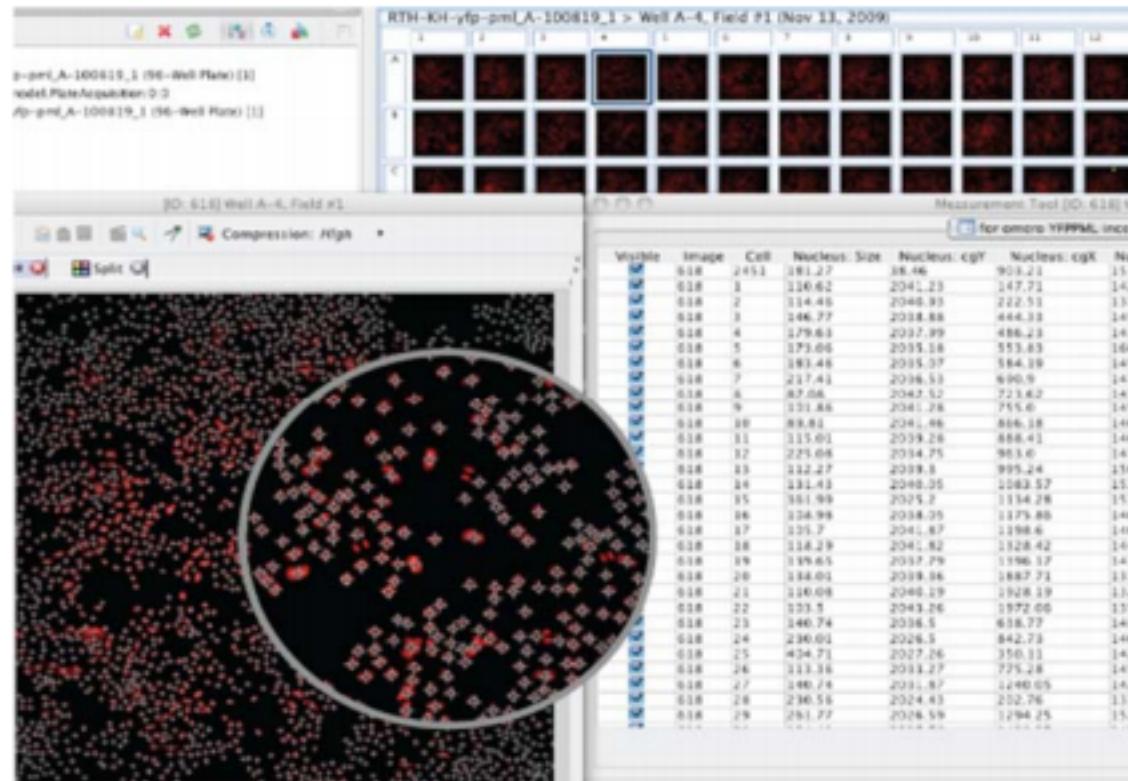
Pruning

Brushing

Details on demand

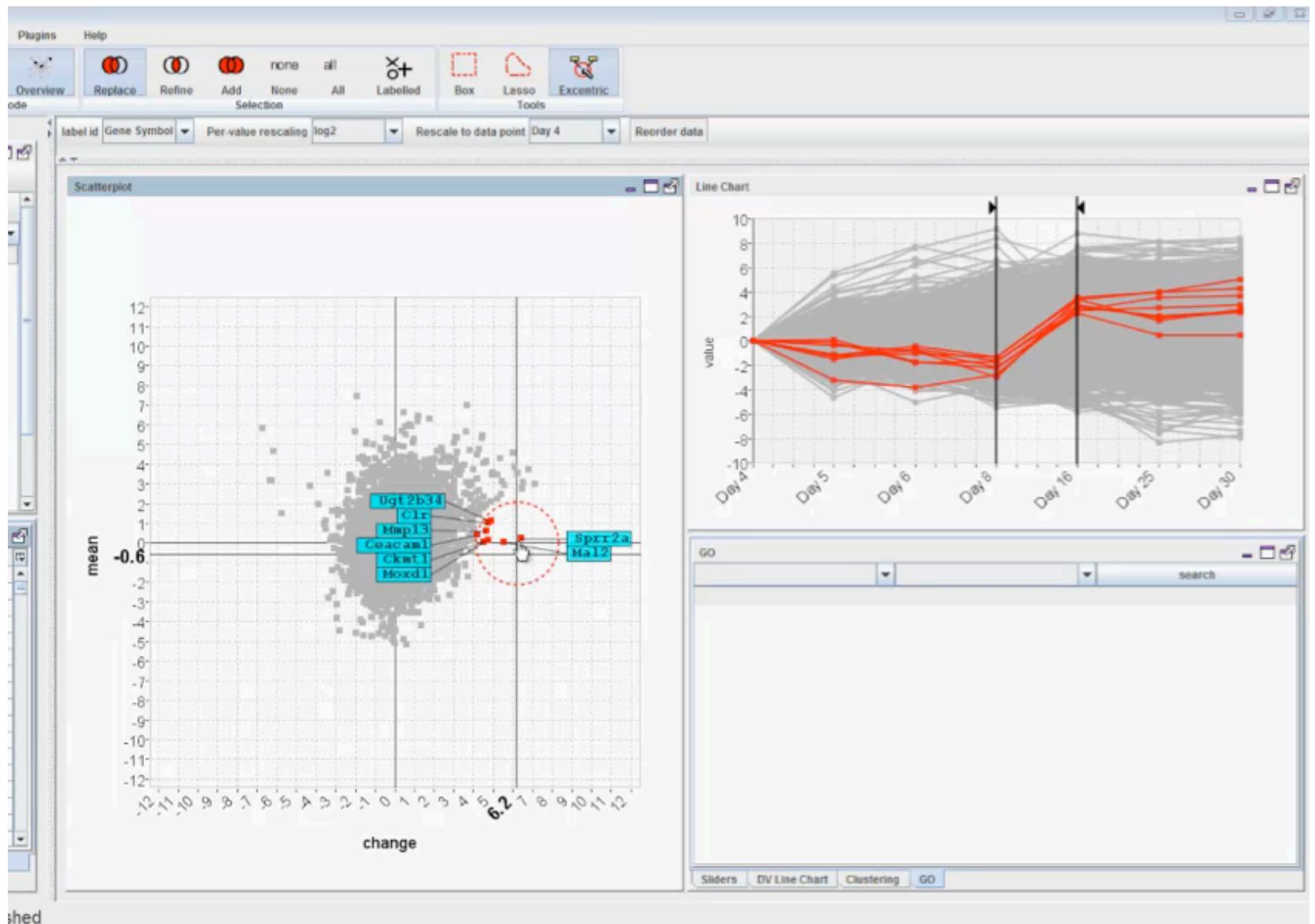
Focus & context

User status



Solution 3: Interaction / Linked views

Time-series
microarray
data



Katy Borner

Examples of effective visualizations in biology

Is a graphical representation really necessary?

Does the legend obviate the figure?

What is my message?

Does my figure communicate it clearly?

Are there extraneous or ornamental elements?

What can I remove without changing the overall story?

Have I left the reader wanting more, or less?

The reader does not know what they need to know.

You must tell them.

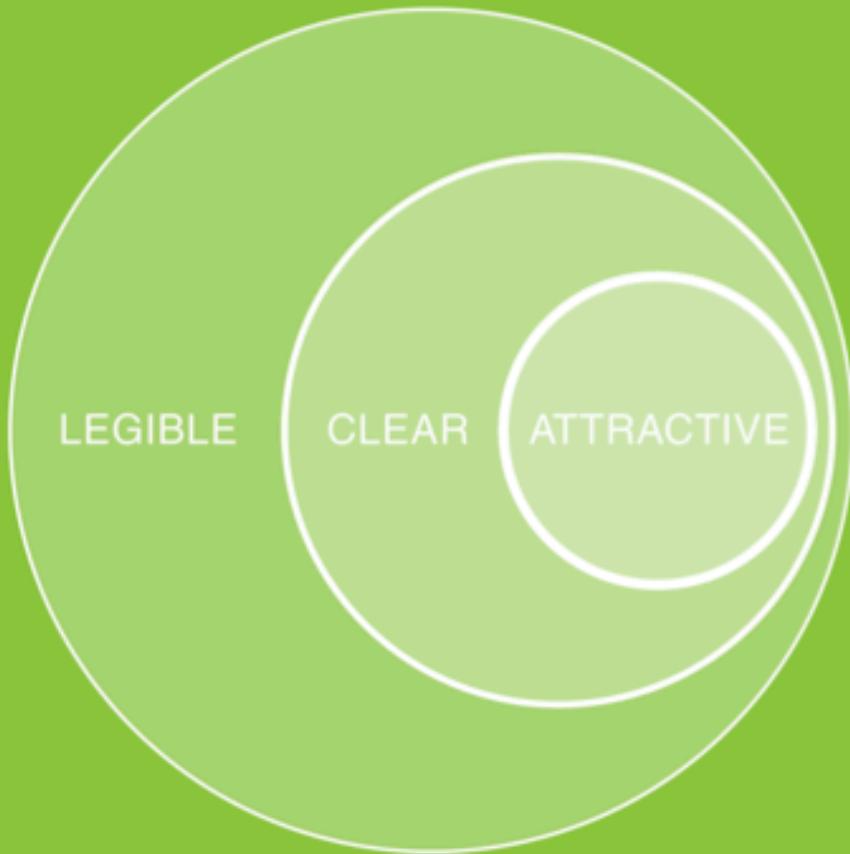
The reader does not know what is important.

You must show them.

The reader's cognitive and visual acuity are limited.

The reader may have time for a good presentation.

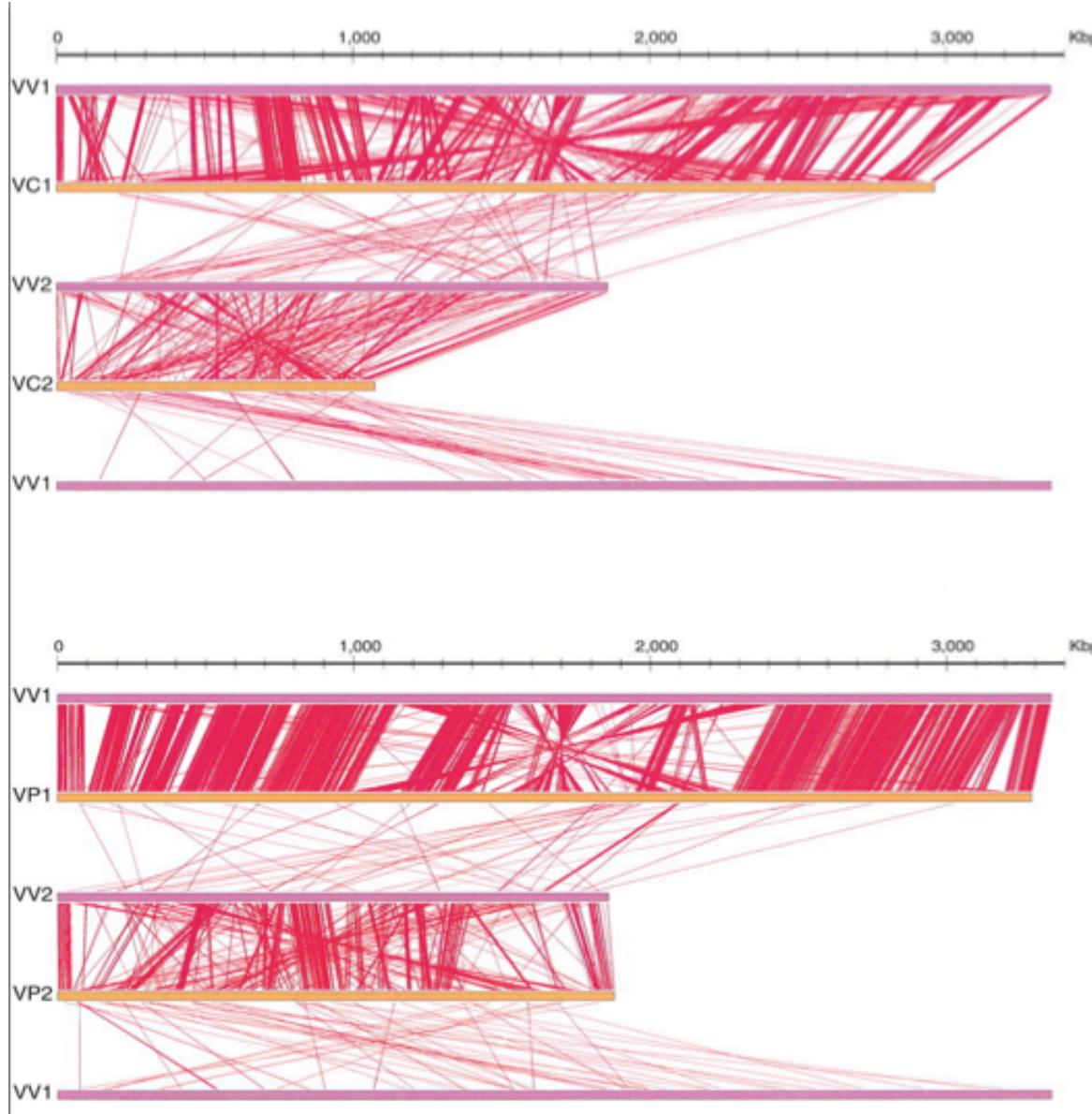
The reader never has time for a bad presentation.



quality of communication

- GOOD
- BETTER
- BEST

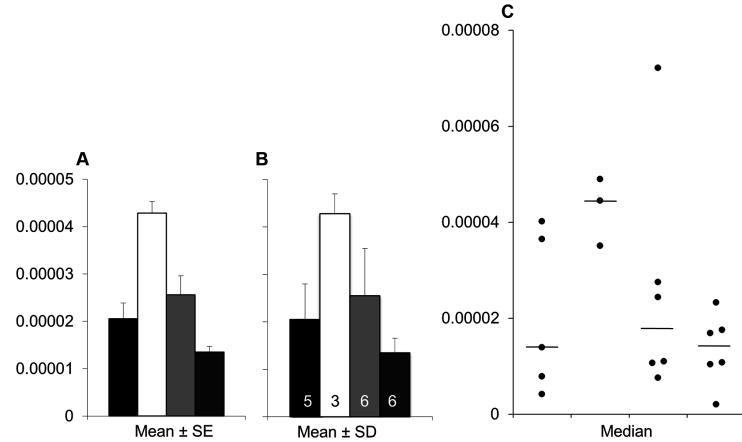
Help the reader identify meaningful patterns



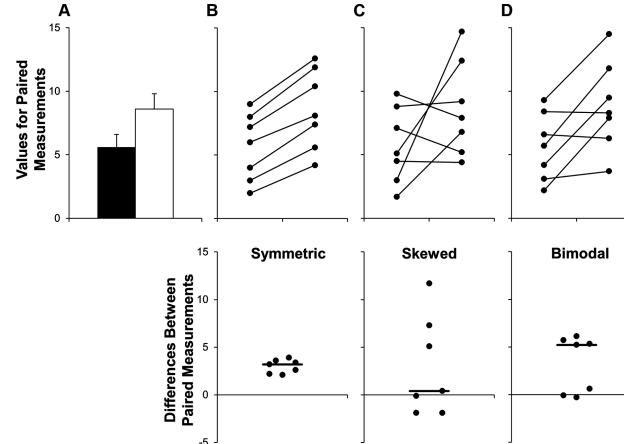
Intra- and interchromosomal shuffling of *Vibrio* genes

Show the raw data (<100 points)

better

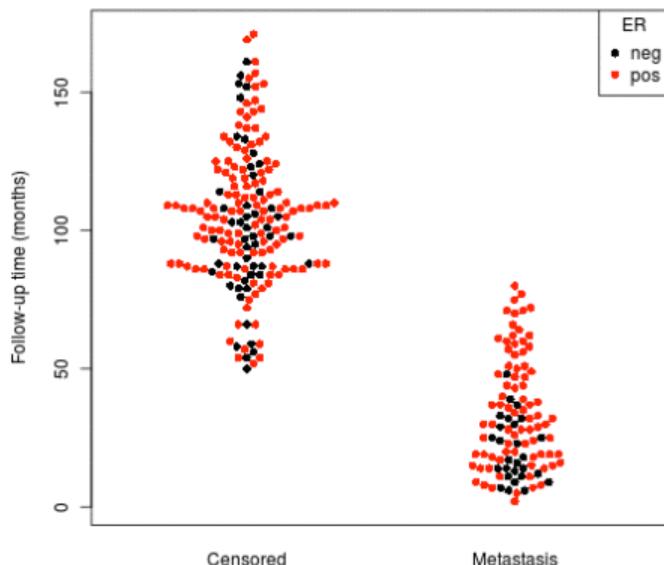


Paired Measurements

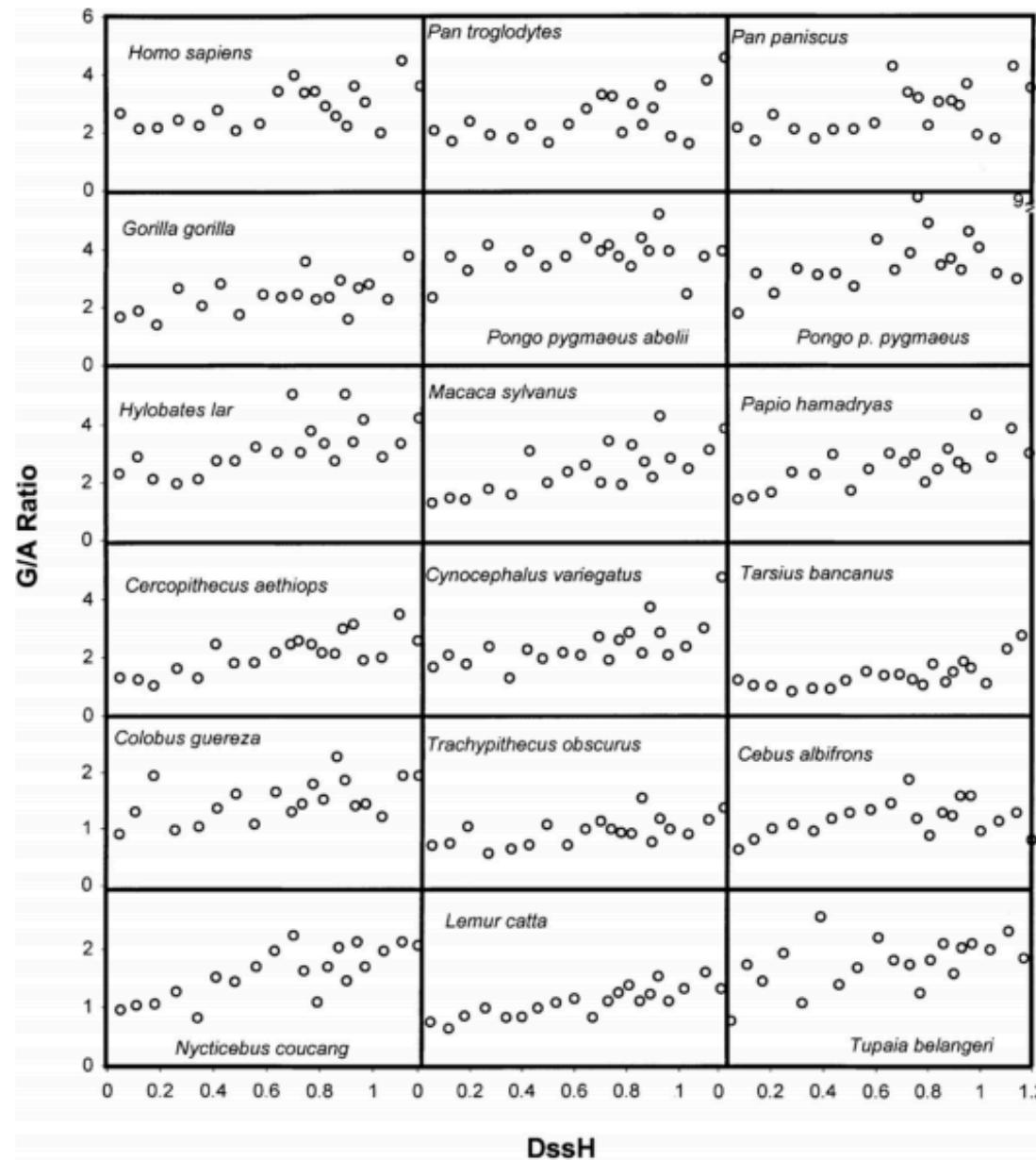


Boxplots?

<http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128>

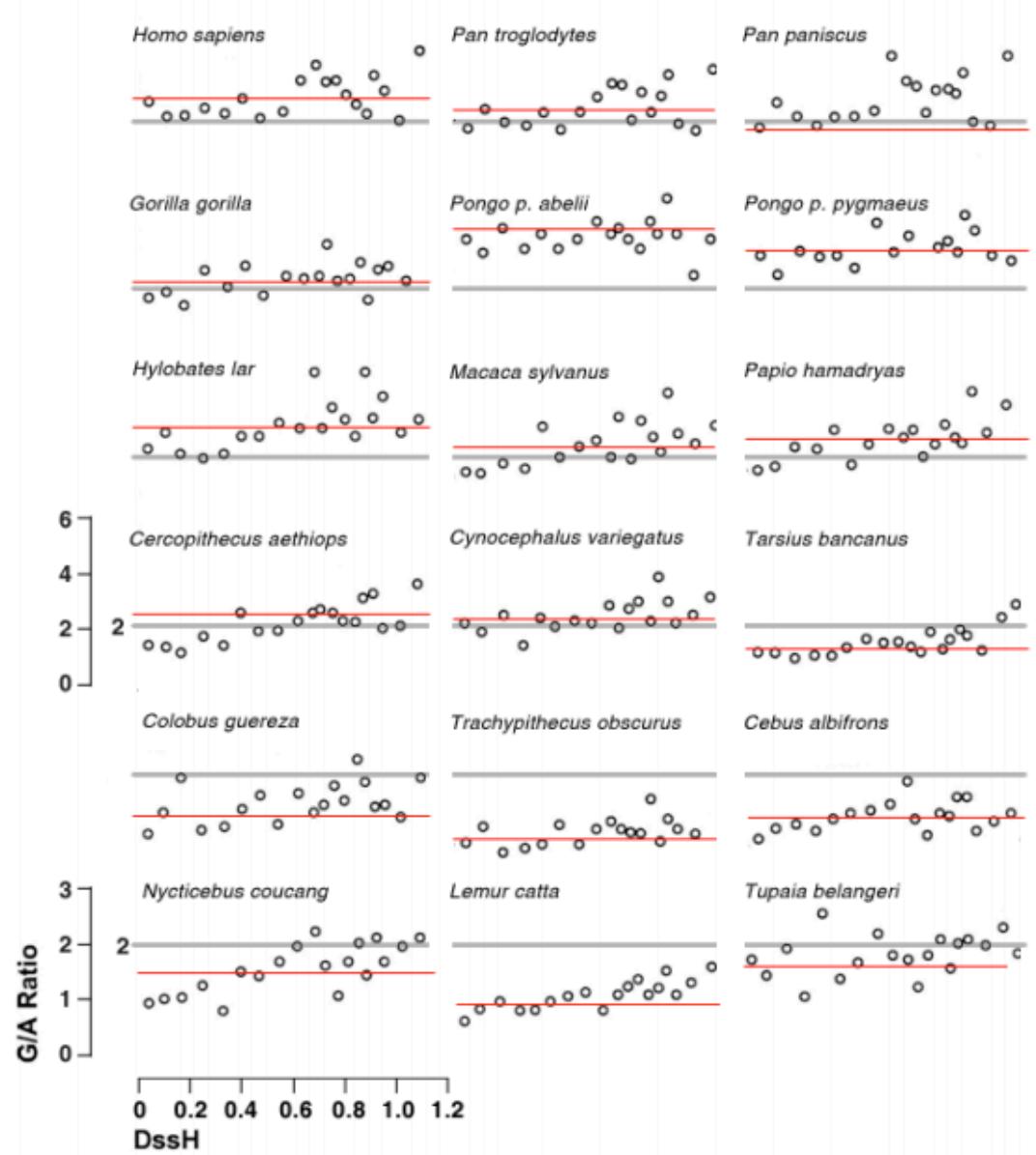


Focus on data



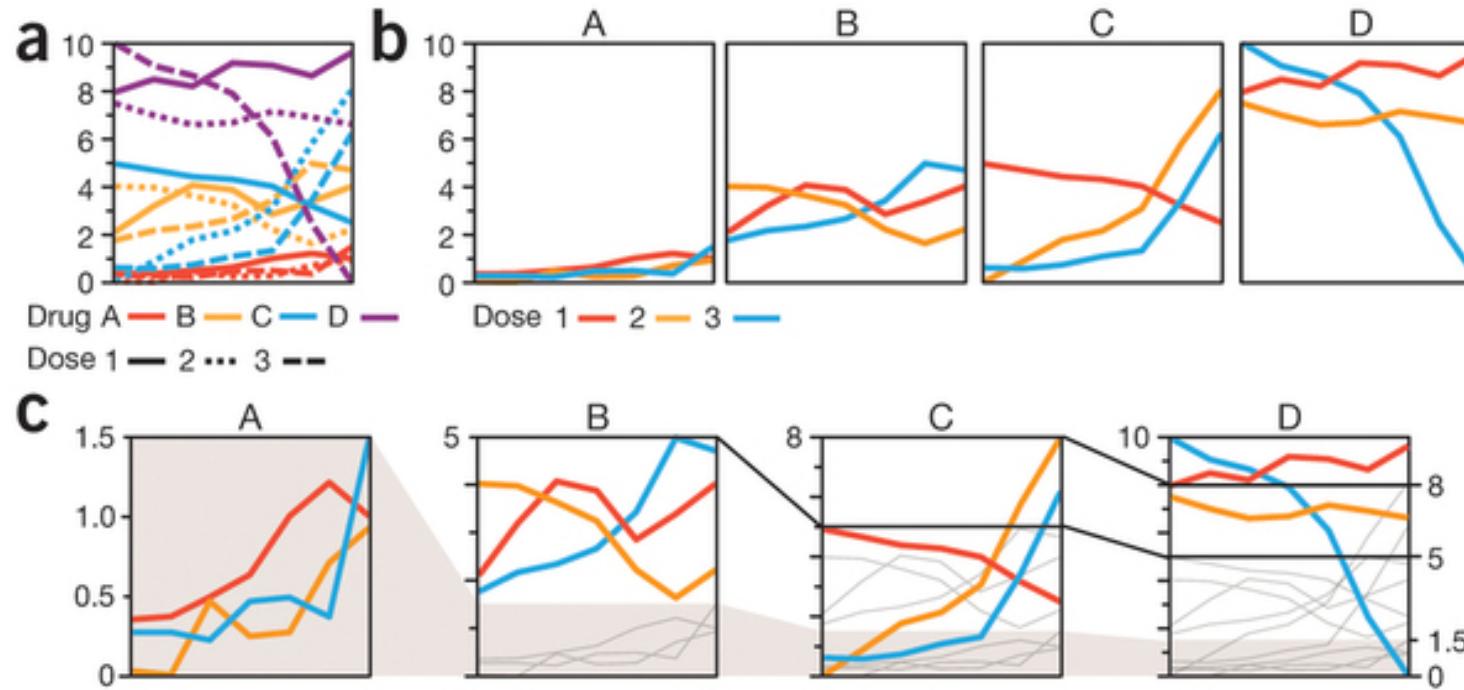
Raina et al. (2005)

Focus on data 2



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

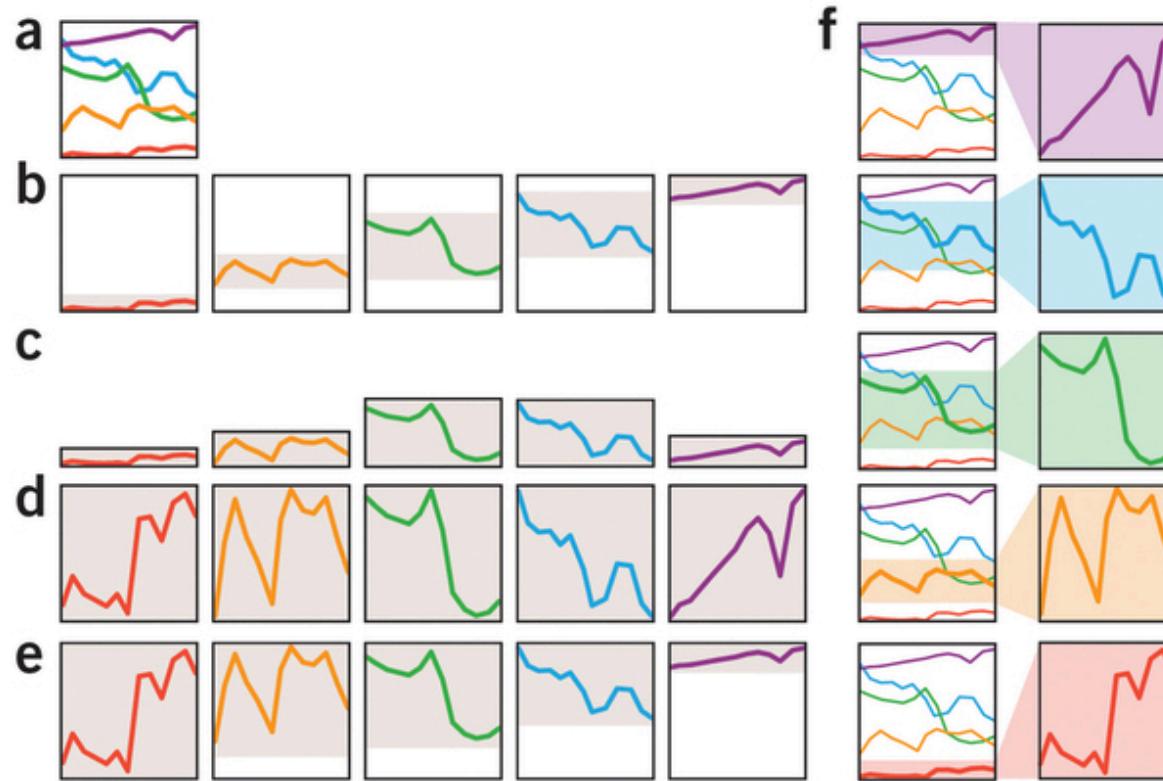
Variation in data range



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

Variation in data range 2

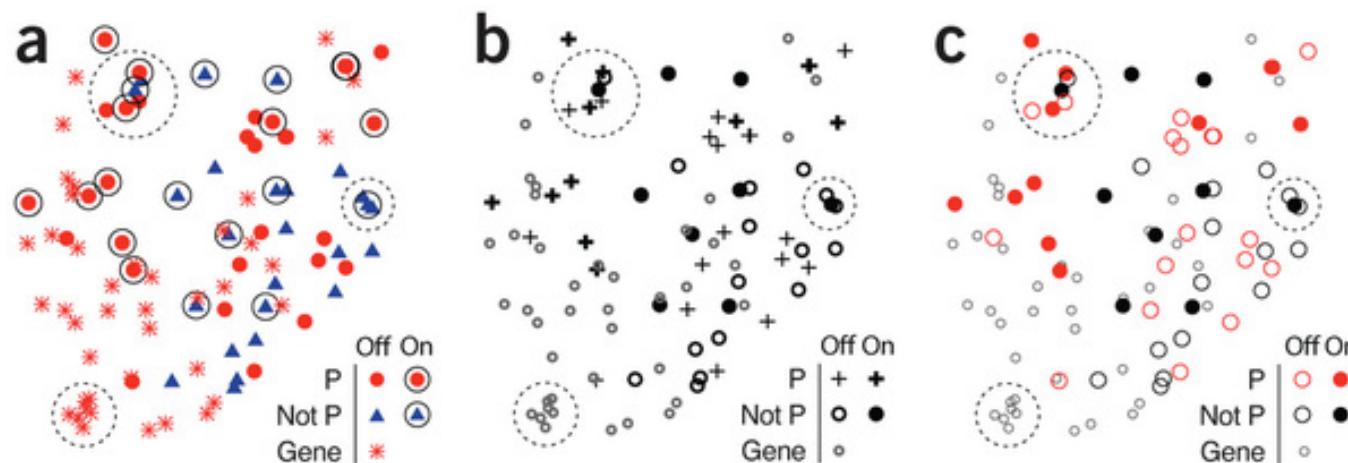


(f) Use an overview and scaled detail to contextualize, highlight and examine each category. Colored backgrounds emphasize differences in scale expansion

Good legends

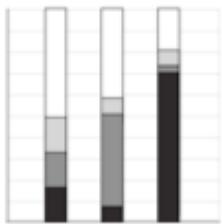


Natural hierarchy. By varying shape and color meaningfully, the encoding becomes more memorable

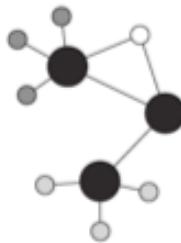
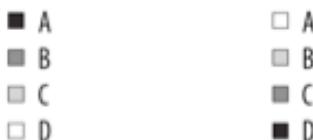
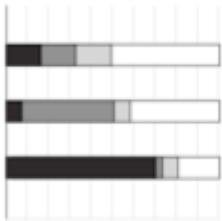


Legend order

consistent

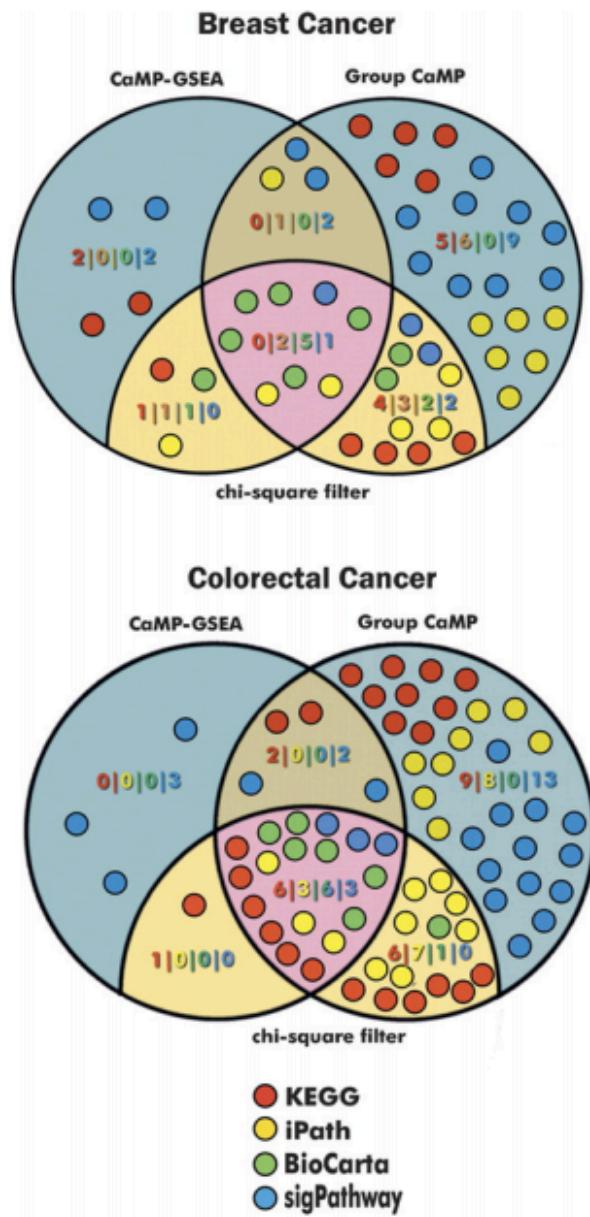


inconsistent

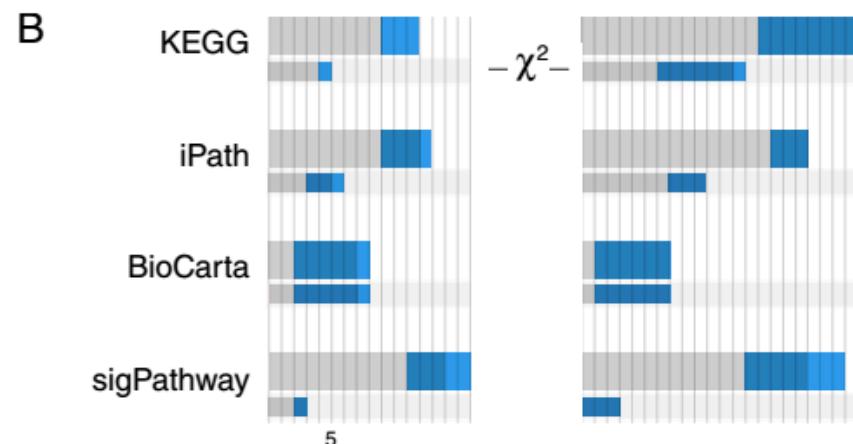
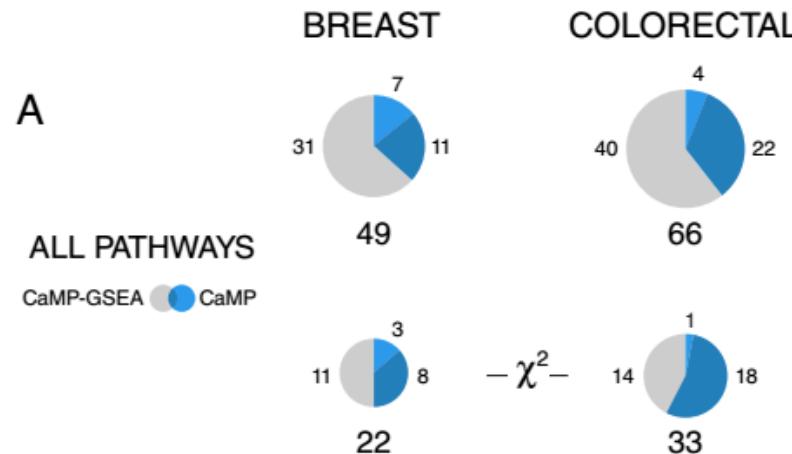
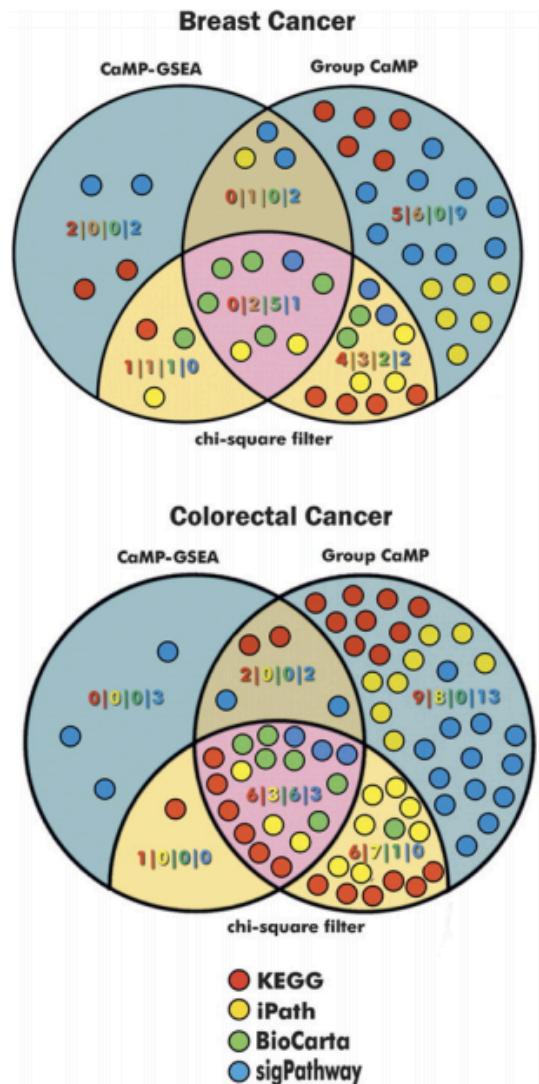


- Order elements in the legend consistently with their appearance in the figure
- more visually balanced when darker tones are at the bottom

Refactoring Complexity

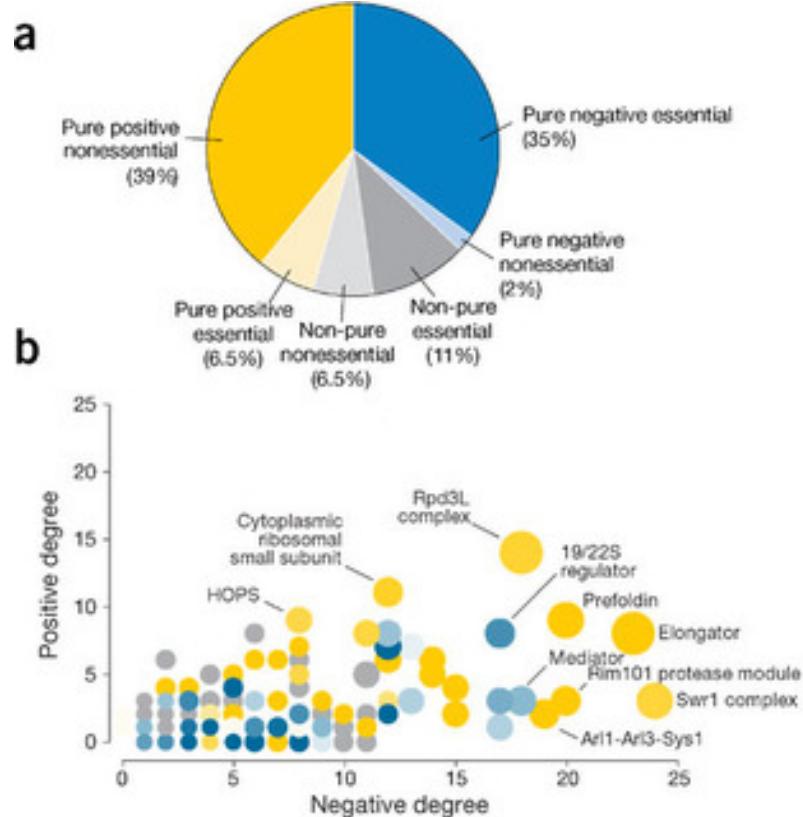


Refactoring Complexity 2

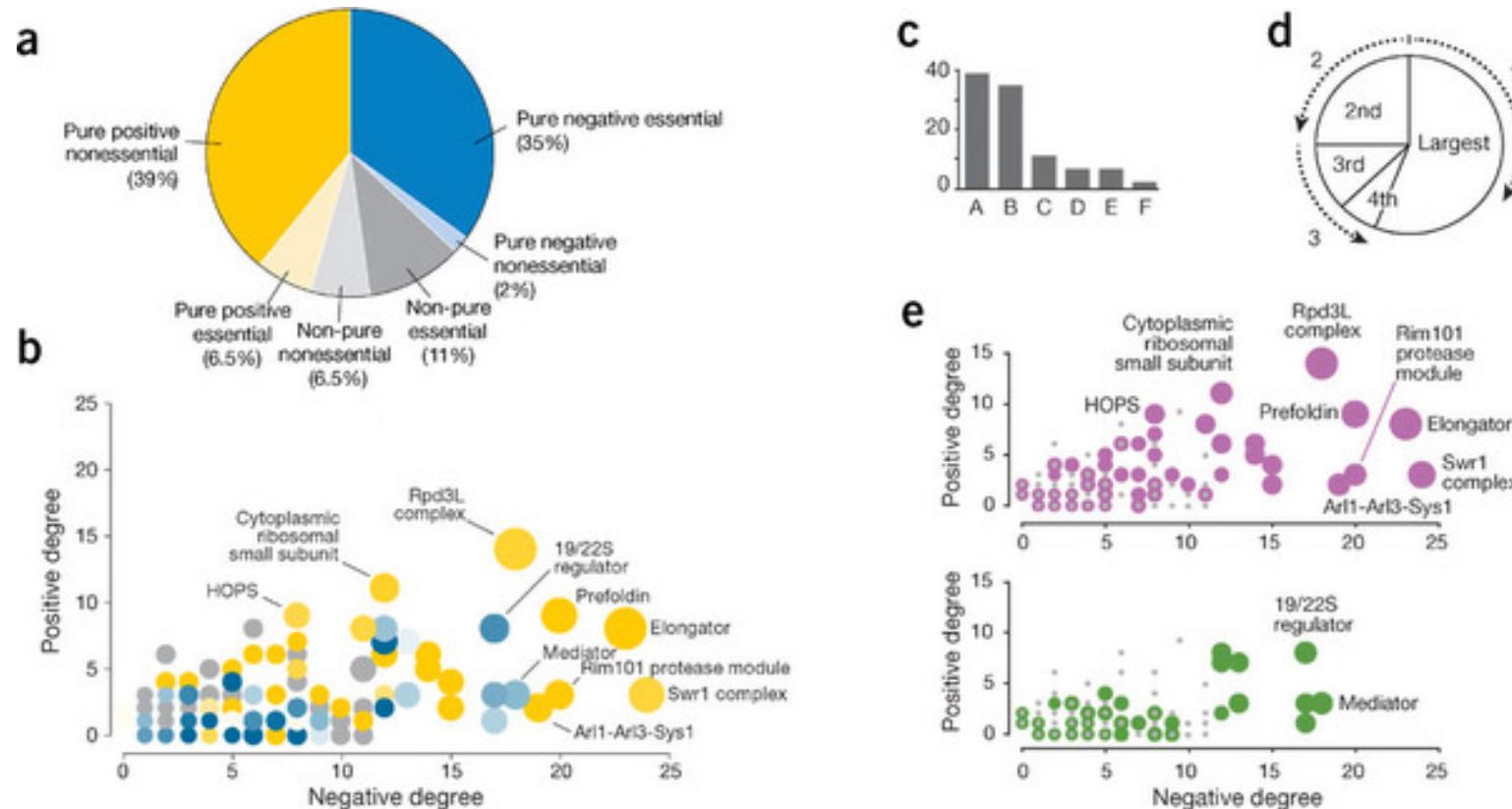


Message is clearer
(Breast > Colorectal, KEGG > sigPathway > iPath > BioCarta)

Refactoring Complexity 2



Refactoring Complexity 2

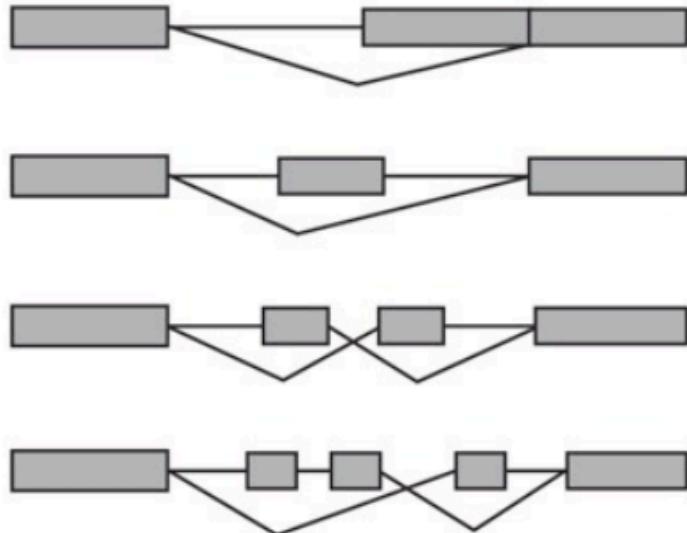


Busy / difficult to read graph
8 point sizes, 11 shades of yellow/
13 blue

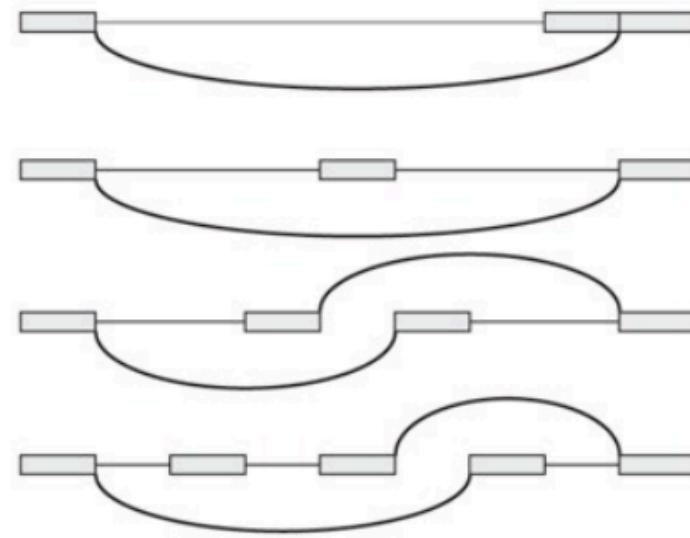
Reduce visual complexity
Limit the color value and size
scales (0–3, 4–7 and others)

Uniform spacing and sizing

spacing variation is implied



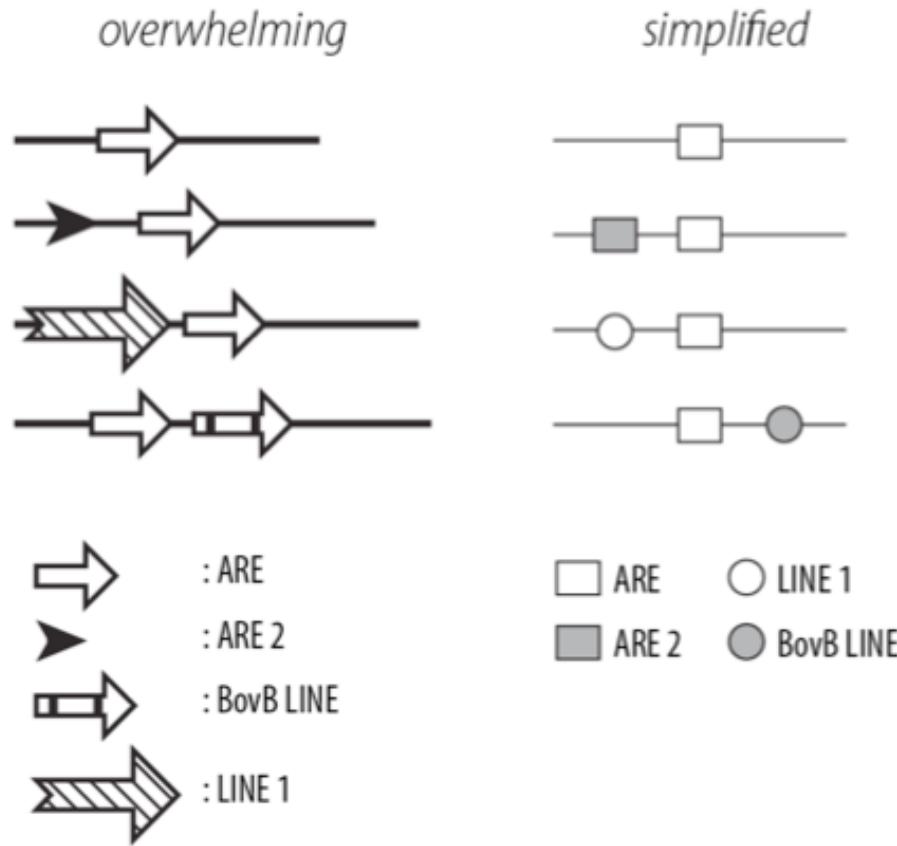
variation refactored



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

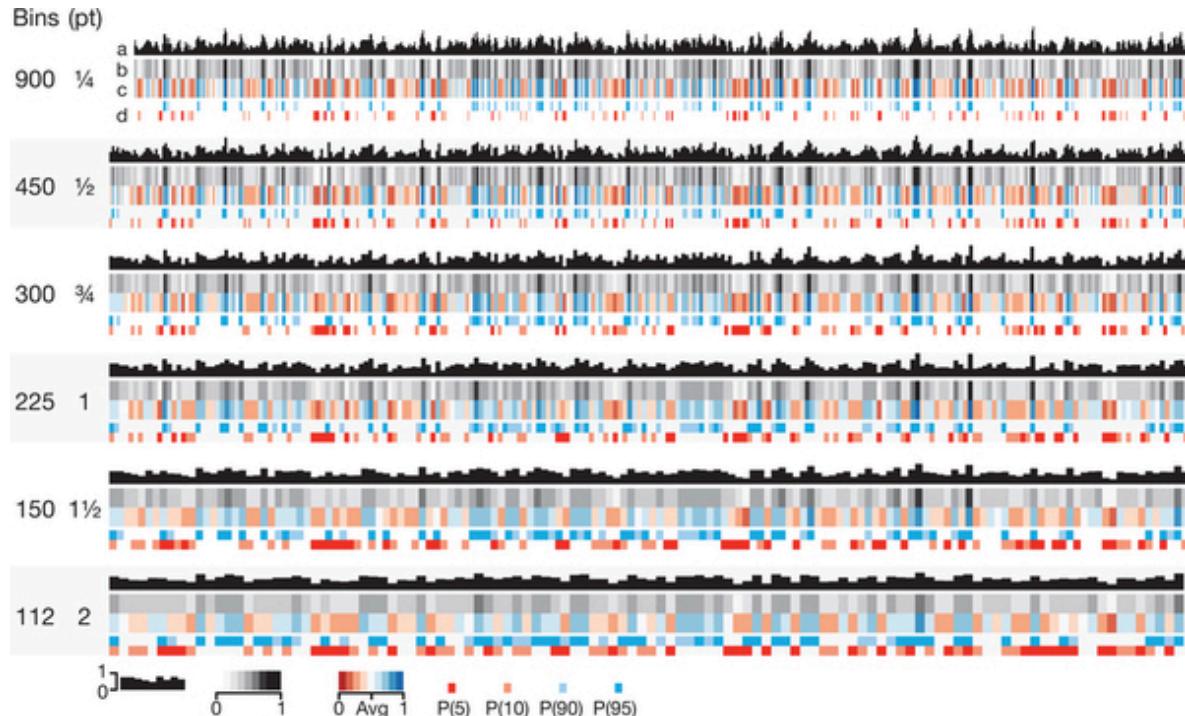
Remove redundancies



Nikaido et al. (1999)

- unclear whether the arrows' size and distance is meaningful
- All arrows point in the same direction
- Any variation in the figure will be interpreted as important to its message

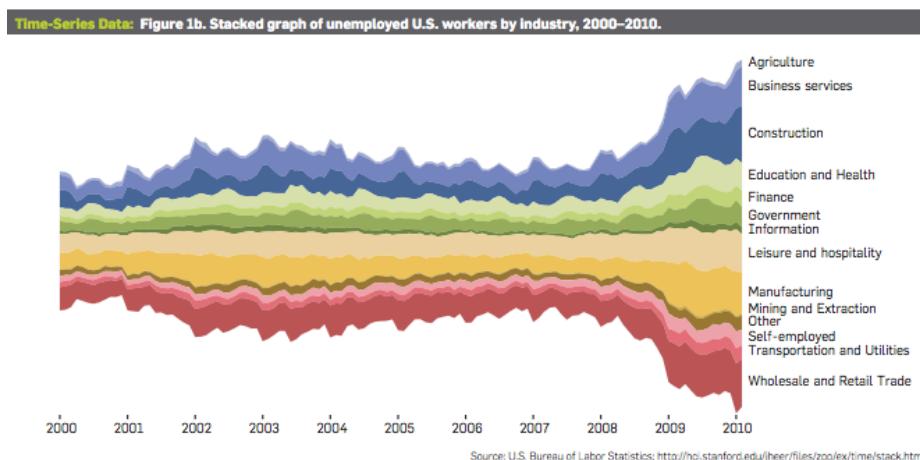
Binning high-resolution data



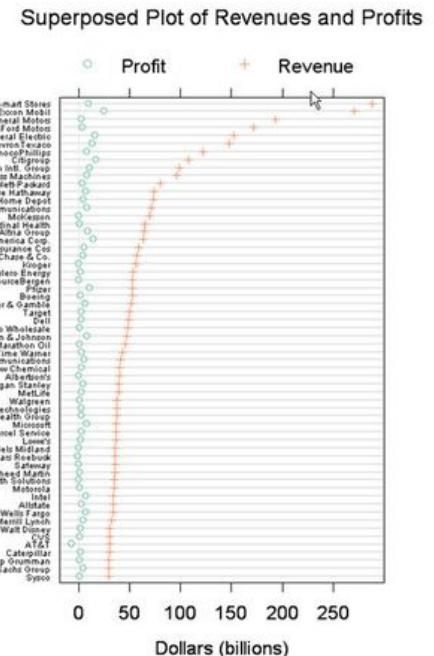
Read coverage of simulated sequencing
(a) histogram, (b) heatmap
(c) Coverage relative to the average
(d) Bins with values at least as extreme as the 5th, 10th, 90th or 95th percentile are marked

- 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)

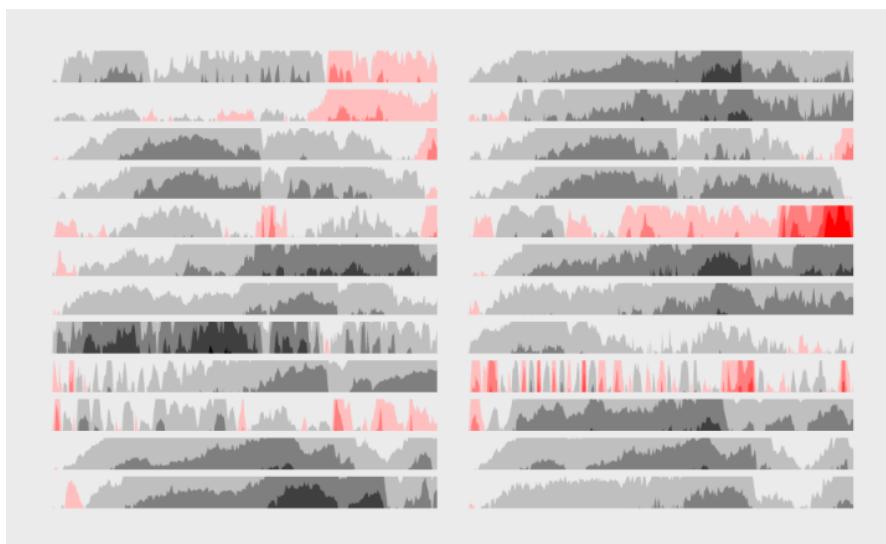
Many other types of charts exist



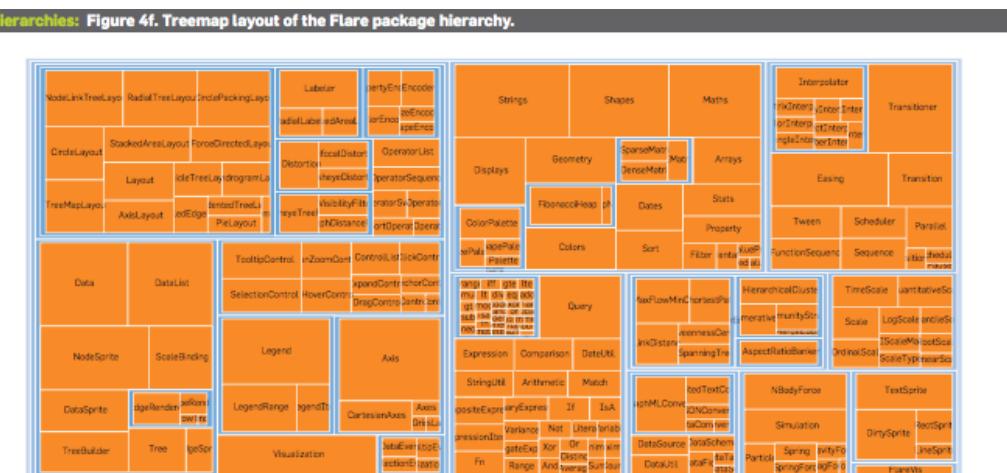
Stacked graph (controversial)



Dotplot



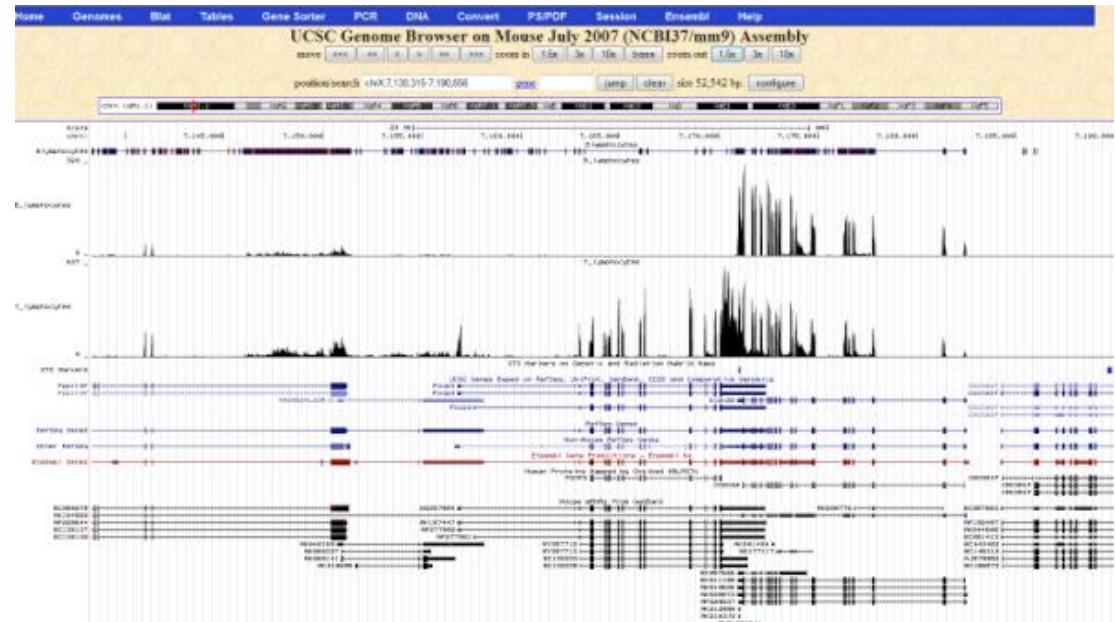
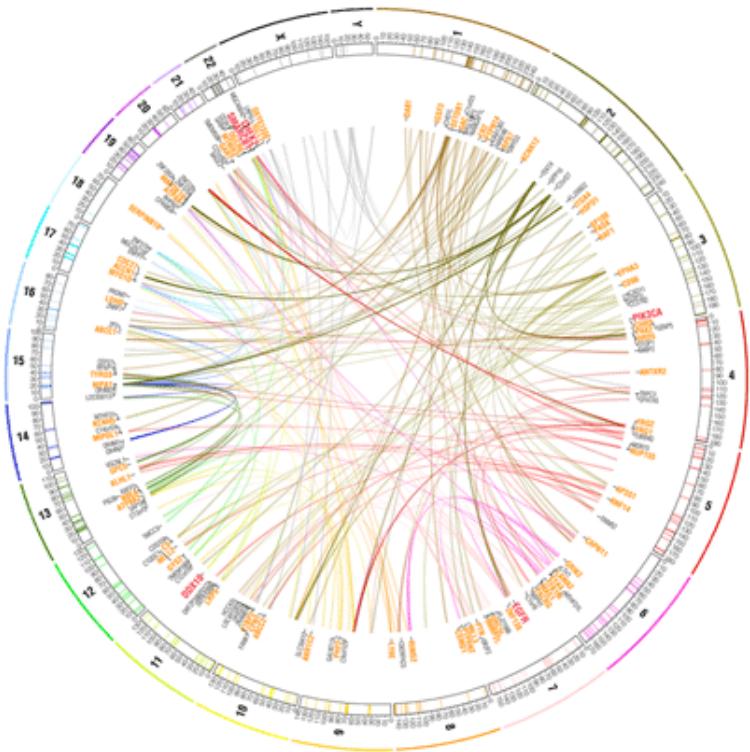
Horizon graph (flowing data)



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/mimolch/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
- Remove unnecessary variation - Any variation in the figure will be interpreted as important to its message
- Display uncertainty (e.g. confidence intervals)
- 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

Visualization Goals

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