Exploratory analysis

Jeff Leek

@jtleek

www.jtleek.com

Key ideas

- Visualization
- Summarization
- Showing the data
- Not being misled

Why explore data?

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To "debug" analyses
- To communicate results

Background

Perceptual tasks

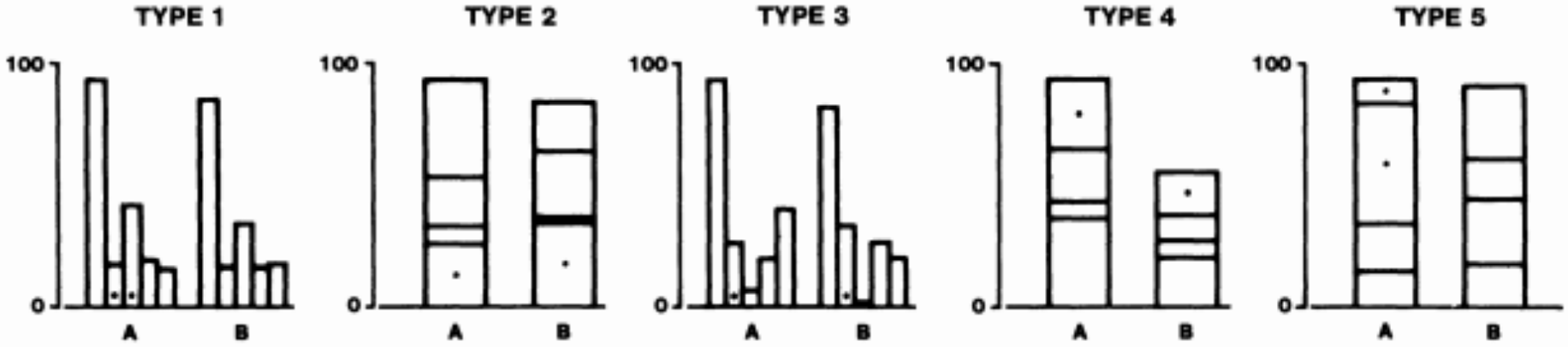
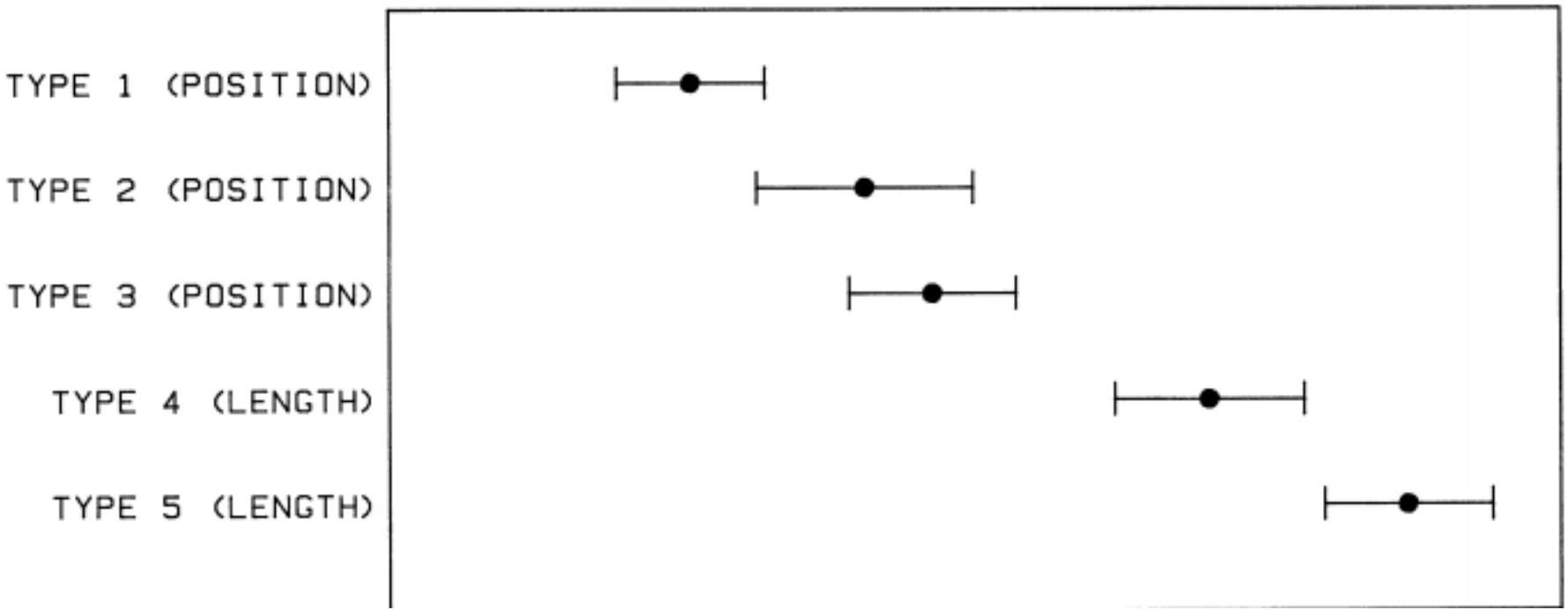


Figure 4. Graphs from position–length experiment.

http://www.jstor.org/stable/2288400



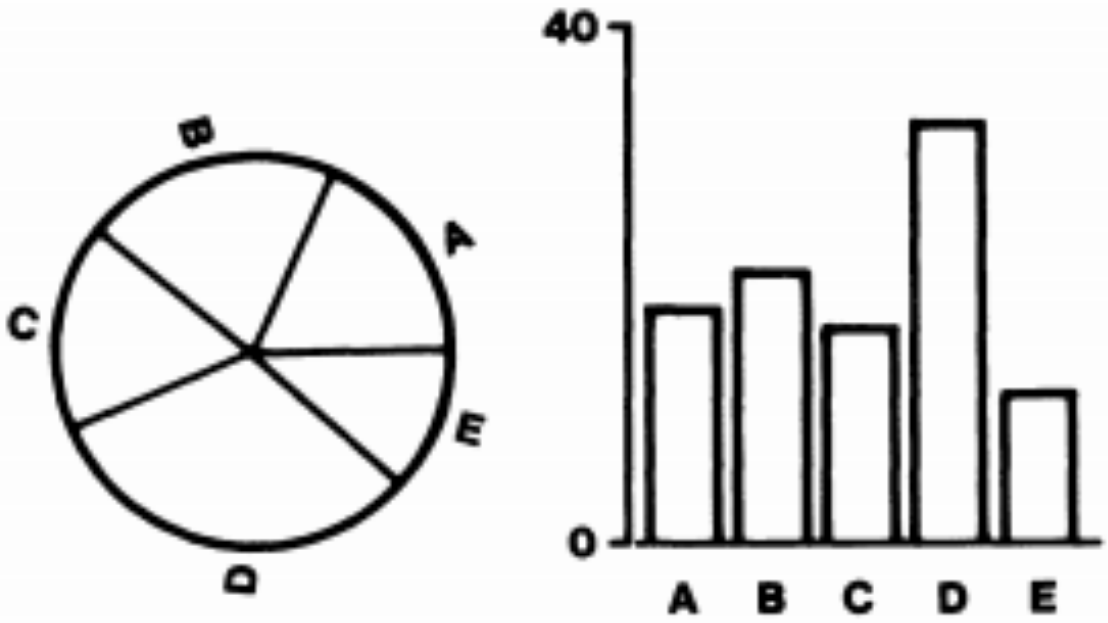
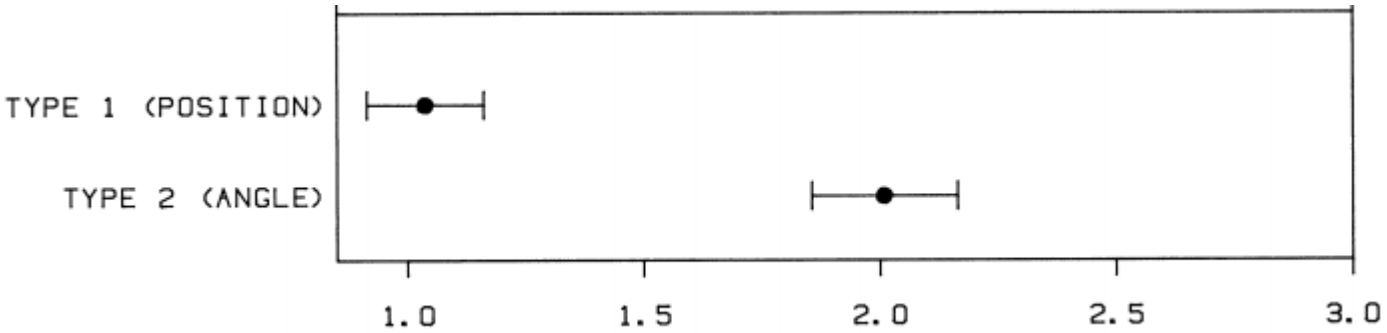


Figure 3. Graphs from position–angle experiment.

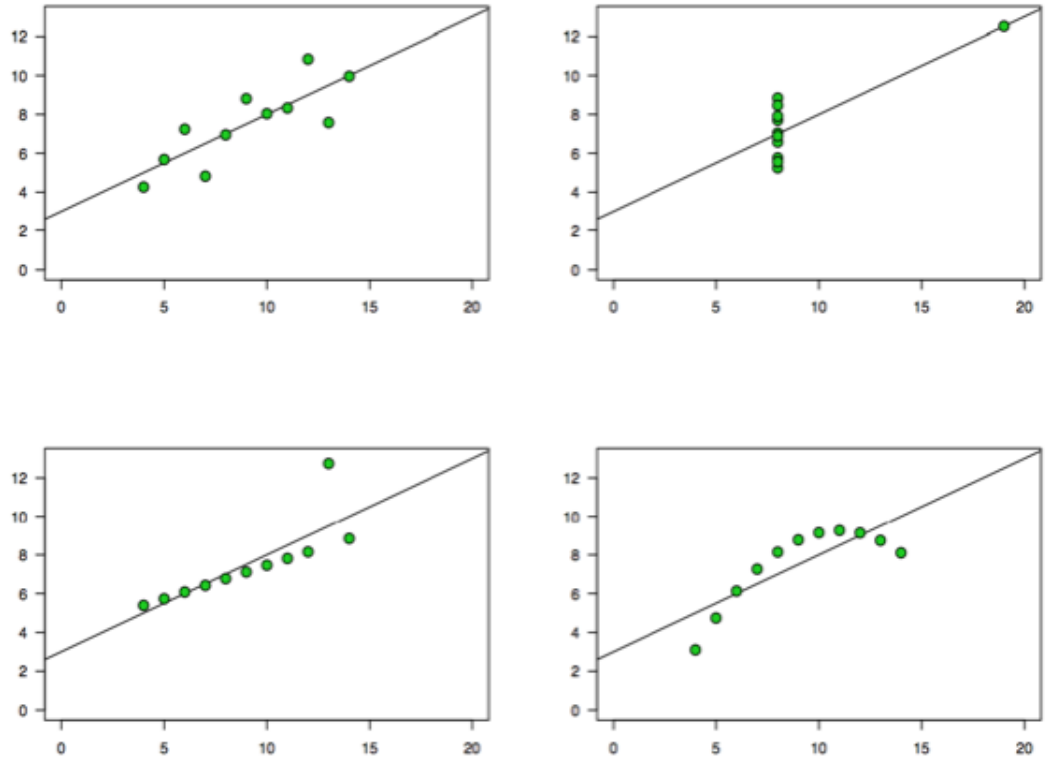


Background

Graphs reveal structure summaries don’t

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

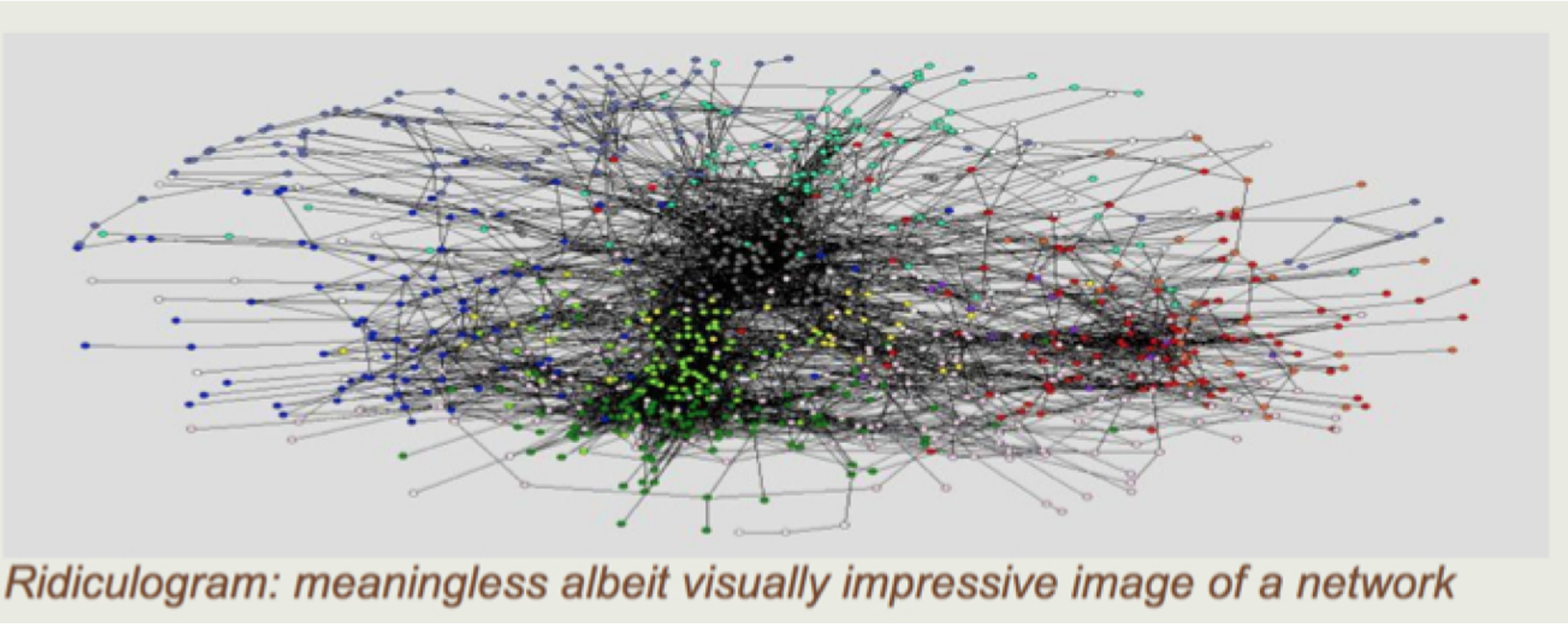
$\hat{\beta}_0 = 3.0, \quad \hat{\beta}_1 = 0.5, \quad$ p-value (slope) = 0.002, $\quad R^2 = 0.67.$



Basic principles

Avoid ridiculograms

http://linxe-eye.blogspot.com/2008/07/when-visuals-gone-wrong.html

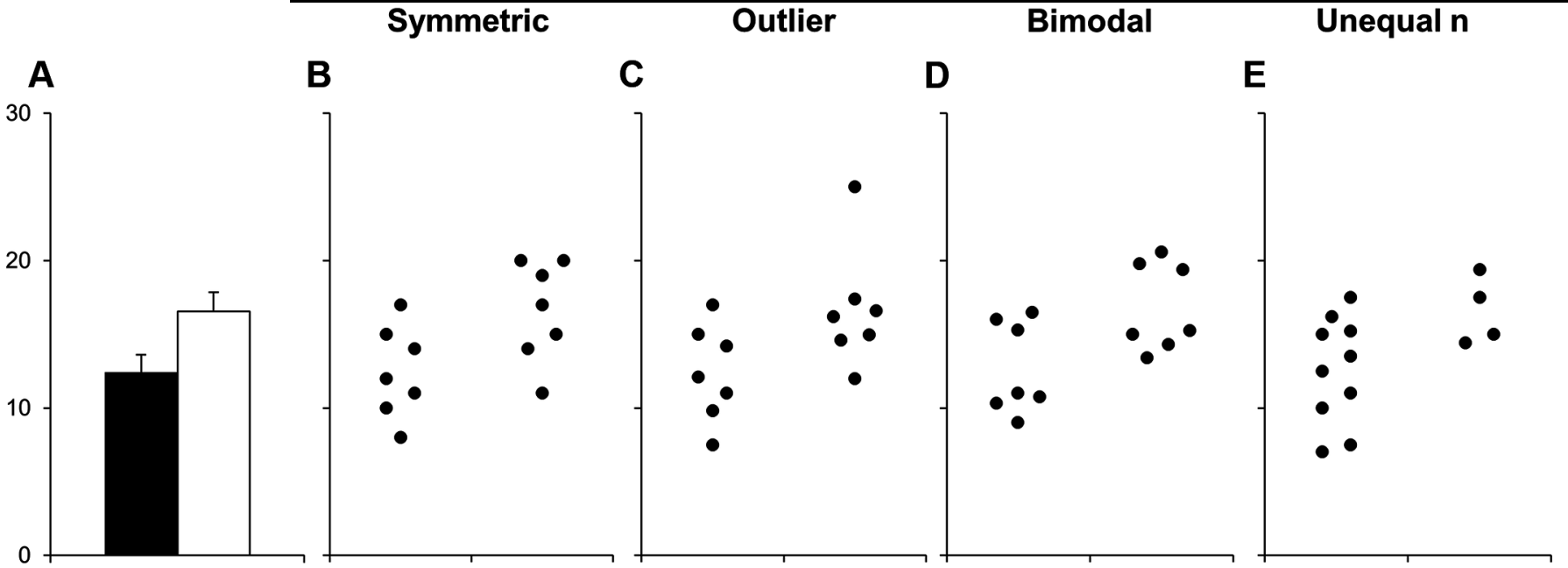


Ridiculogram: meaningless albeit visually impressive image of a network

Basic principles

Show the data

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

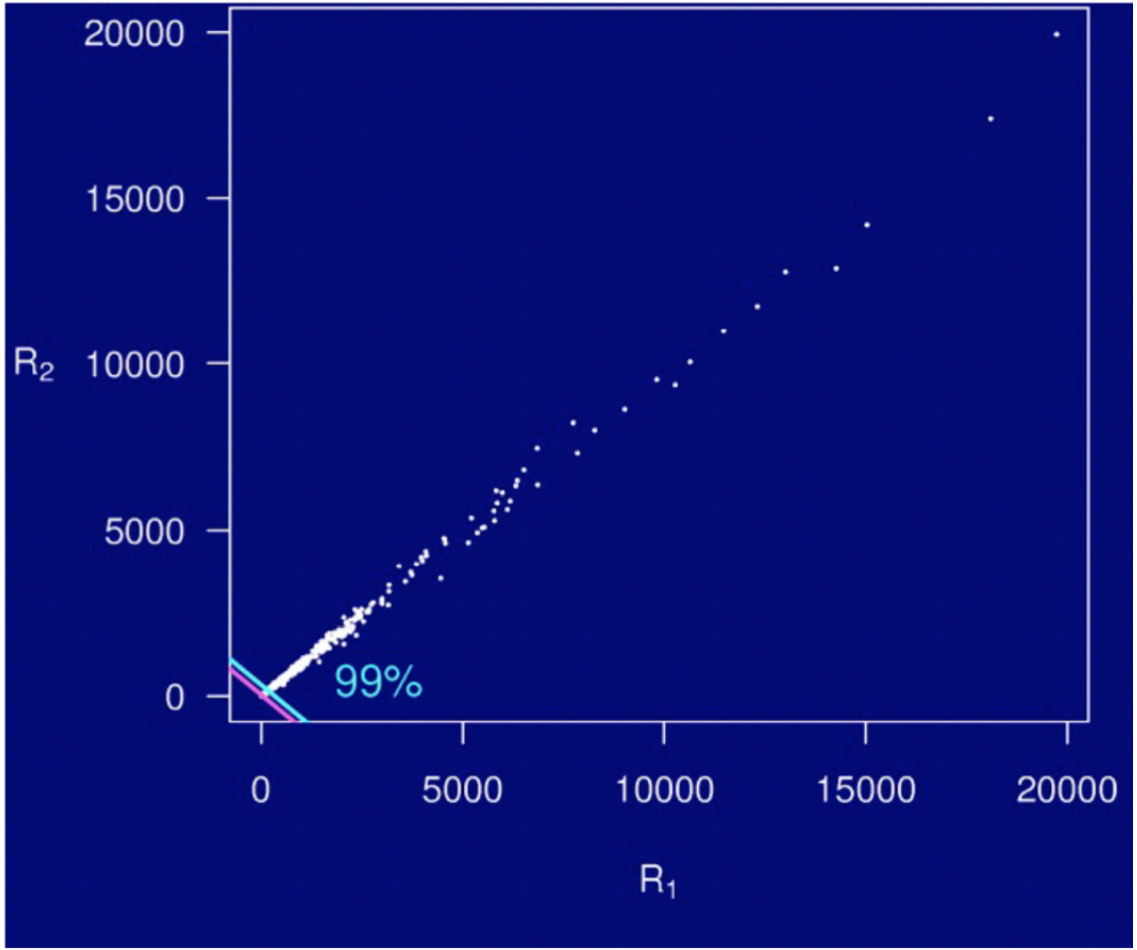


Test			p value	
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Basic principles

Be careful with scale

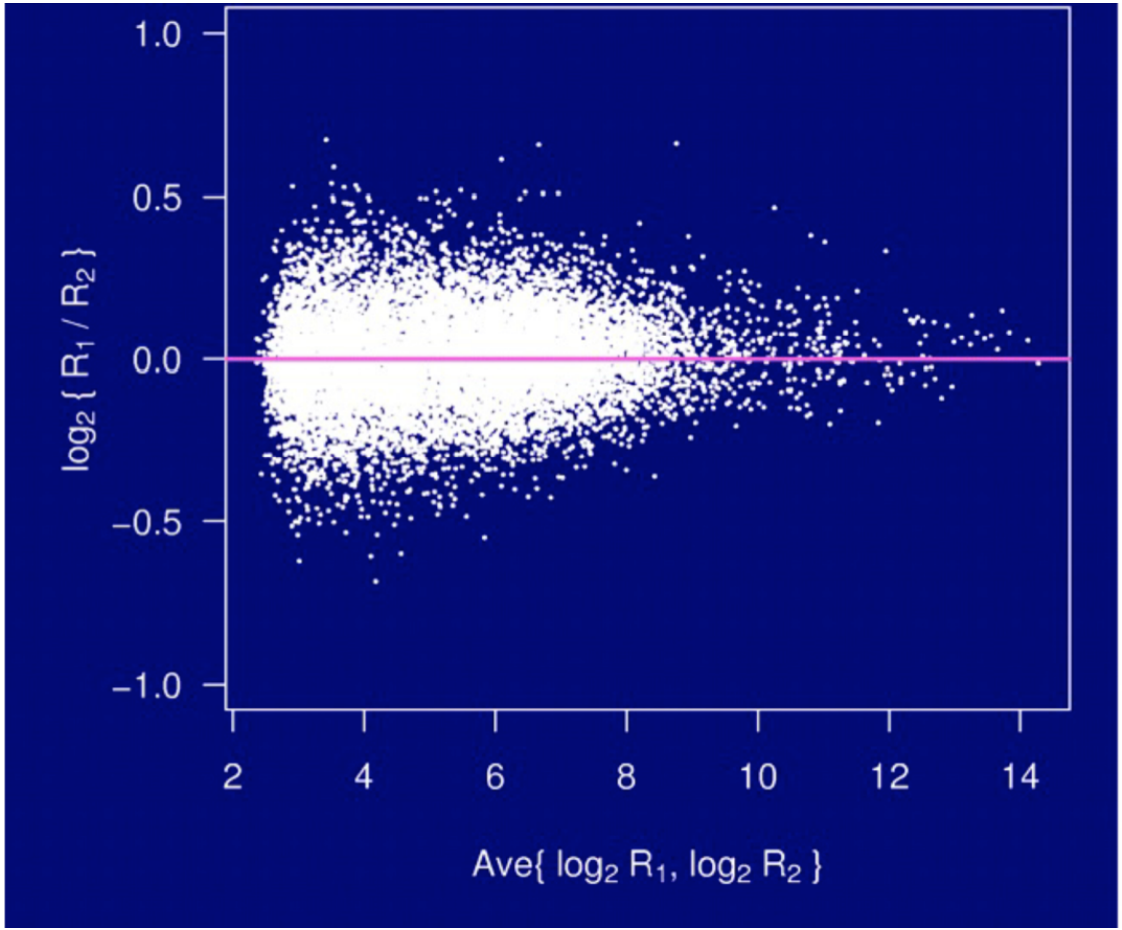
https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Basic principles

Compare things directly

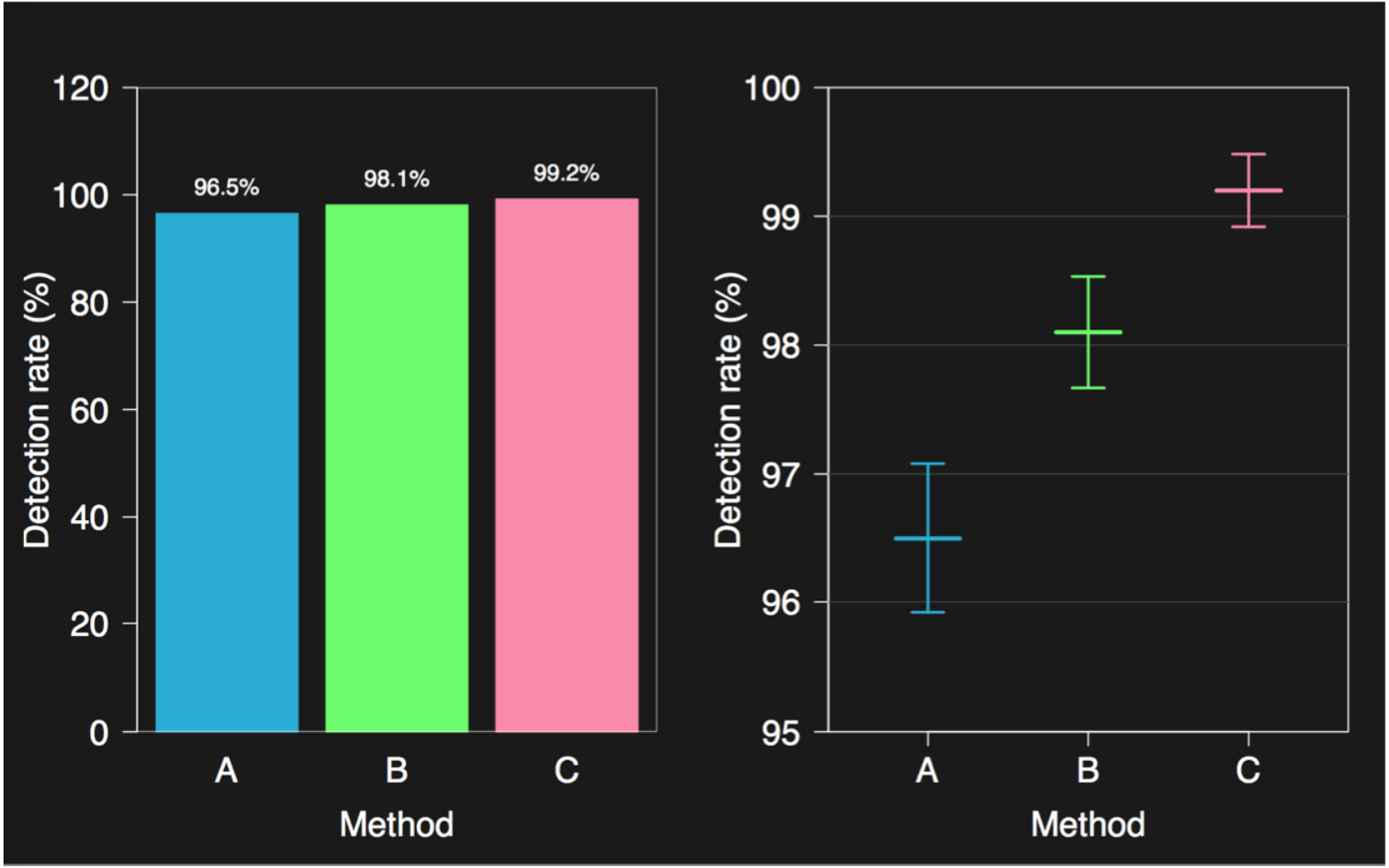
https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Basic principles

- Use common scales
- Start at zero

https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Further resources

- Karl Broman's guide to displaying data
  - https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf
- Data visualization at Nature
  - http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html