Exploratory data analysis

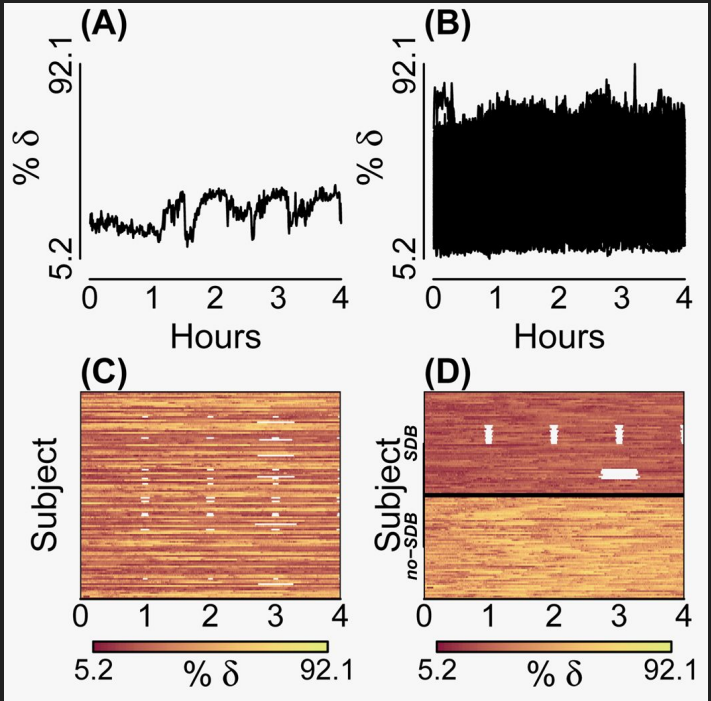
Brian Caffo (@bcaffo)

Contact info at: www.bcaffo.com

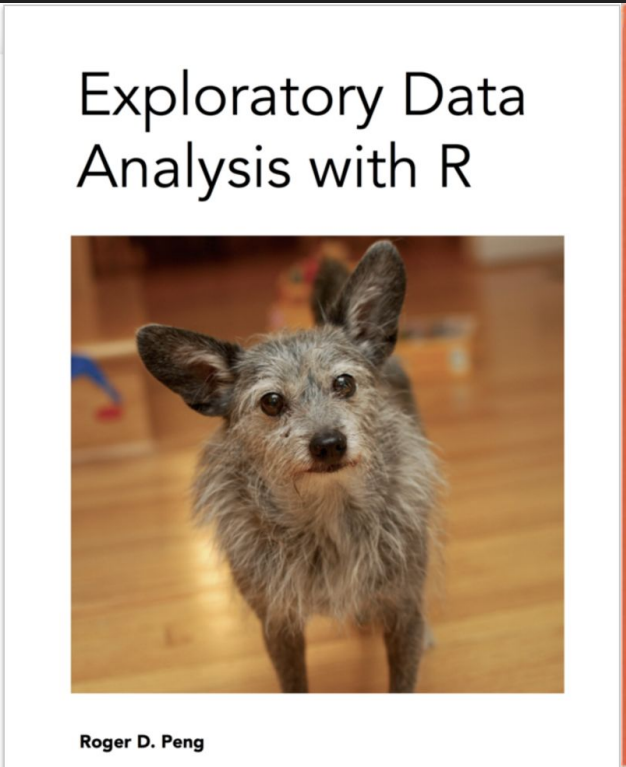
Lasagna plots: A saucy alternative to spaghetti plots

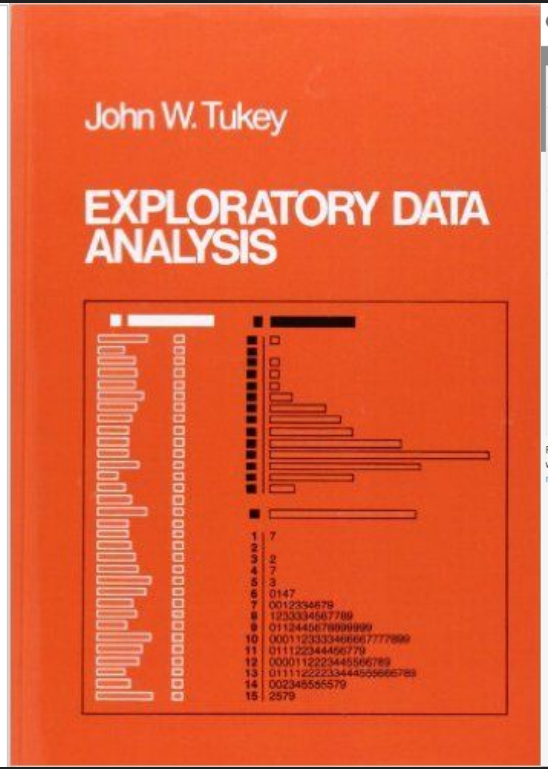
Bruce J. Swihart,[1] Brian Caffo,[1] Bryan D. James,[2] Matthew Strand,[3] Brian S. Schwartz,[4] and Naresh M. Punjabi[5]

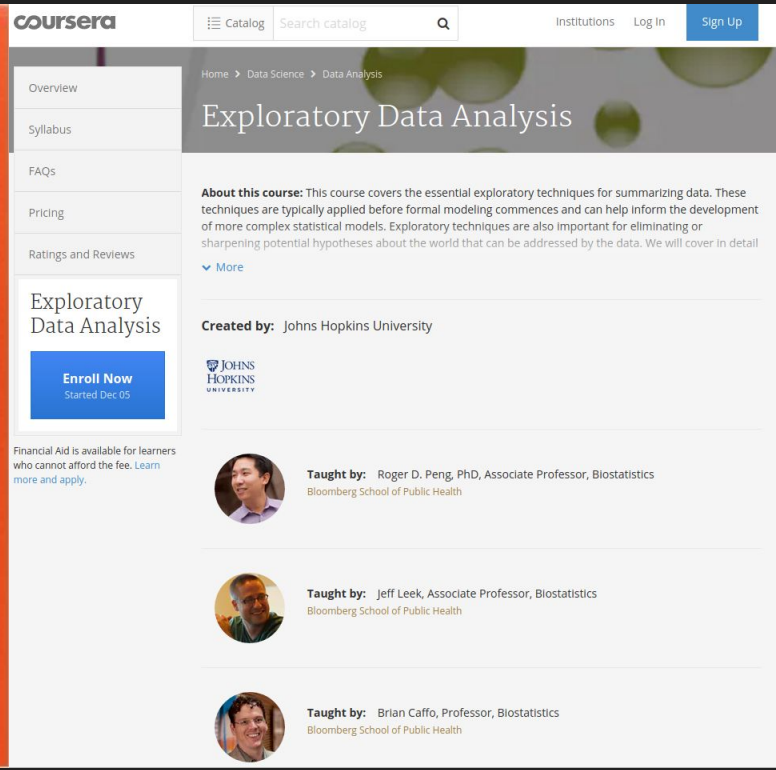
Author information ► Copyright and License information ►



Resources


Exploratory Data Analysis with R
Roger D. Peng


John W. Tukey
EXPLORATORY DATA ANALYSIS


coursera
Catalog
Search catalog
Institutions
Log In
Sign Up
Overview
Syllabus
FAQs
Pricing
Ratings and Reviews
Home > Data Science > Data Analysis
Exploratory Data Analysis
About this course: This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data. We will cover in detail
More
Exploratory Data Analysis
Enroll Now
Started Dec 05
Financial Aid is available for learners who cannot afford the fee. Learn more and apply.
Created by: Johns Hopkins University
Taught by: Roger D. Peng, PhD, Associate Professor, Biostatistics
Bloomberg School of Public Health
Taught by: Jeff Leek, Associate Professor, Biostatistics
Bloomberg School of Public Health
Taught by: Brian Caffo, Professor, Biostatistics
Bloomberg School of Public Health

EDA

- Focus on discovery and hypothesis generation
- Free form and less structured (improv jazz)
- Controls error rates and performs uncertainty quantification loosely
- Can use graphs, models, prediction, hypothesis tests, ...

CDA

- Focus on hypothesis confirmation
- Prescriptive, protocolized & planned (classical)
- Attempts to strictly control error rates or uncertainty quantification
- Tends to focus on formal inferential or prediction techniques (though can employ graphs …)

EDA versus CDA

Data analysis generally falls on a spectrum from the high prescriptive and formal setting of regulated clinical trials for drug development to more exploratory data analysis found in high throughput measurement technologies, the EDA/CDA division is more useful conceptually than practically

Alternate dichotomy: hypothesis driven versus purely empirical studies

Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era

Douglas B. Kell[1]* and Stephen G. Oliver[2]

Simply Statistics A statistics blog by Rafa Irizarry, Roger Peng, and Jeff Leek

The key word in "Data Science" is not Data, it is Science

12 Dec 2013

Warning: the more you use your data for hypothesis generation and exploration, the harder it gets to control error rates on the same data

estimates vary on the extent and consequences of this problem

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

The Extent and Consequences of P-Hacking in Science

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions

Published: March 13, 2015 • http://dx.doi.org/10.1371/journal.pbio.1002106

An estimate of the science-wise false discovery rate and application to the top medical literature

Leah R. Jager

Jeffrey T. Leek*

+ Author Affiliations

*To whom correspondence should be addressed. jleek@jhsph.edu

EDA

Steps in an EDA

Read in data

Figure out what it is

Pre-process it

Look at dimensions

Look at values

Make tables

Hunt for messed up values

Hunt for NAs

Plot it

Don’t fool yourself

slide courtesy of J Leek

Steps in an EDA

Read in data

Figure out what it is

Pre-process it

Look at dimensions

Look at values

Make tables

Hunt for messed up values

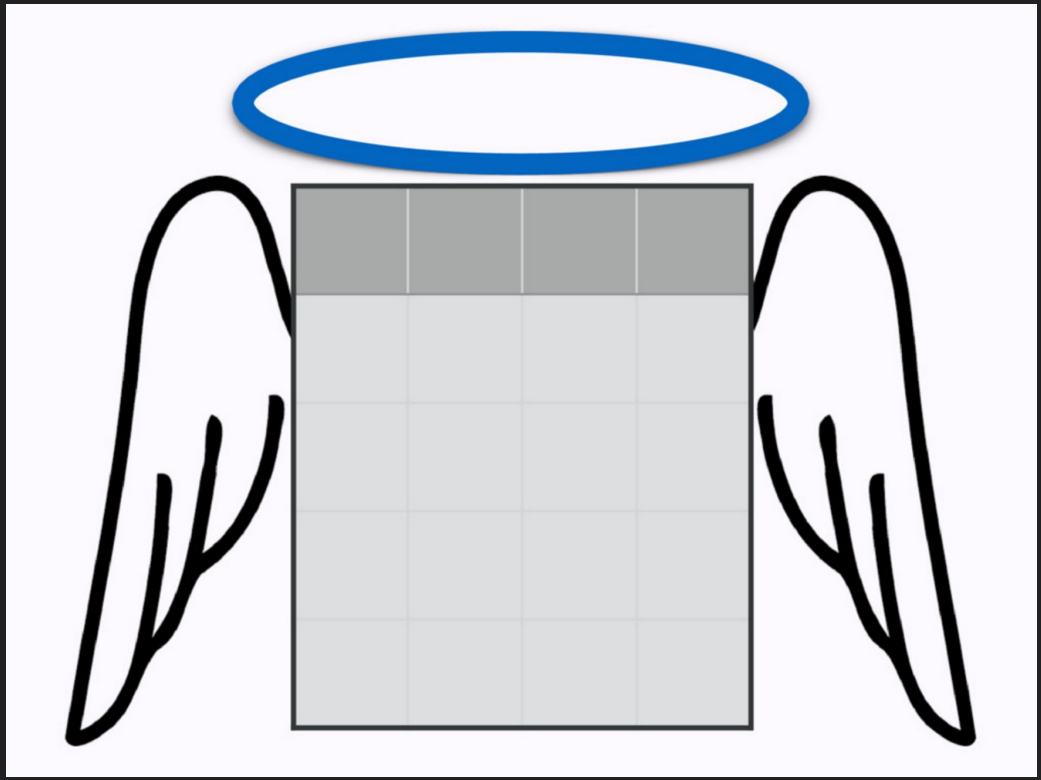
Hunt for NAs

Plot it

Don’t fool yourself

Preprocess it

Remember rectangles (Jenny Bryan)



General advice from Jenny Bryan

- Data wrangling is work!
- No one ever said “I really regret getting my data into such a well organized and thought out format”
- Save your steps / use version control and reproducible research!
- Try to get your data into a rectangle
  - Name your columns with a sensible naming convention
  - Use names that are amenable to software packages
    - No spaces, special characters, use capitalization like a coder
  - No special features if you’re using a spreadsheet (like embedded graphs)
  - Don’t use a number for missing values (888, 9999)

Evolving R tools grammar of data wrangling “tidyverse”

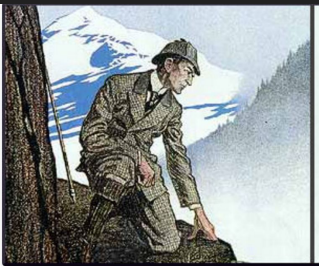
Don’t fool yourself

https://en.wikipedia.org/wiki/Blind_men_and_an_elephant#/media/File:Blind_monks_examining_an_elephant.jpg

slide courtesy of J Leek

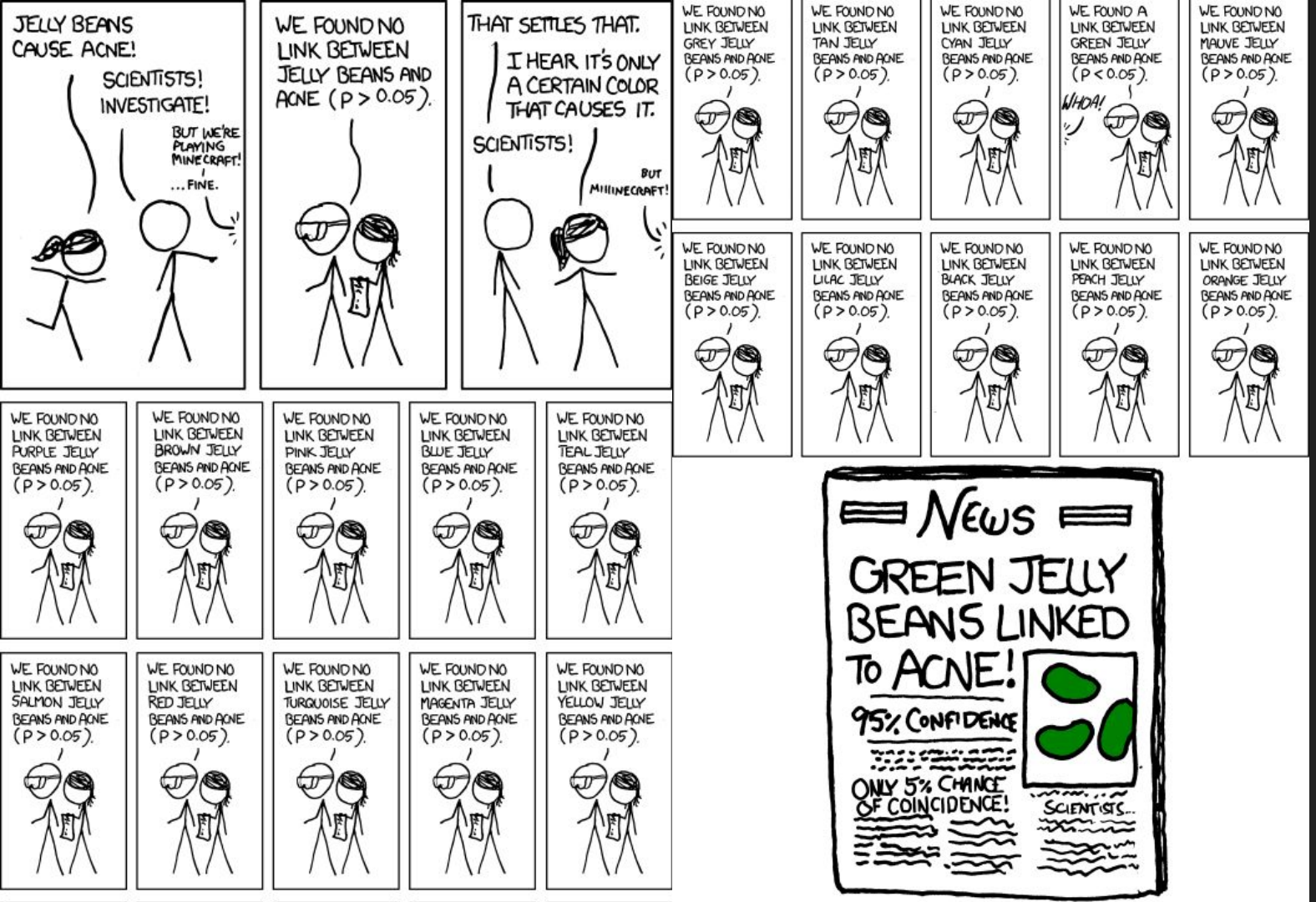
BULLSEYE PAINTING COMPANY, LLC

Contractor: BULLSEYE PAINTING COMPANY, LLC



Quote Investigator

Exploring the Origins of Quotations


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

XKCD 882

Example common ways you can fool yourself

- Issues with the data that you have (elephant, drunk)
- True things may not paint a complete picture (elephant)
- Confirmation bias (bullseye)
- False findings (bullseye, multiplicity)
- Repeatedly looking for things until you find something (multiplicity)

Interocular content

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

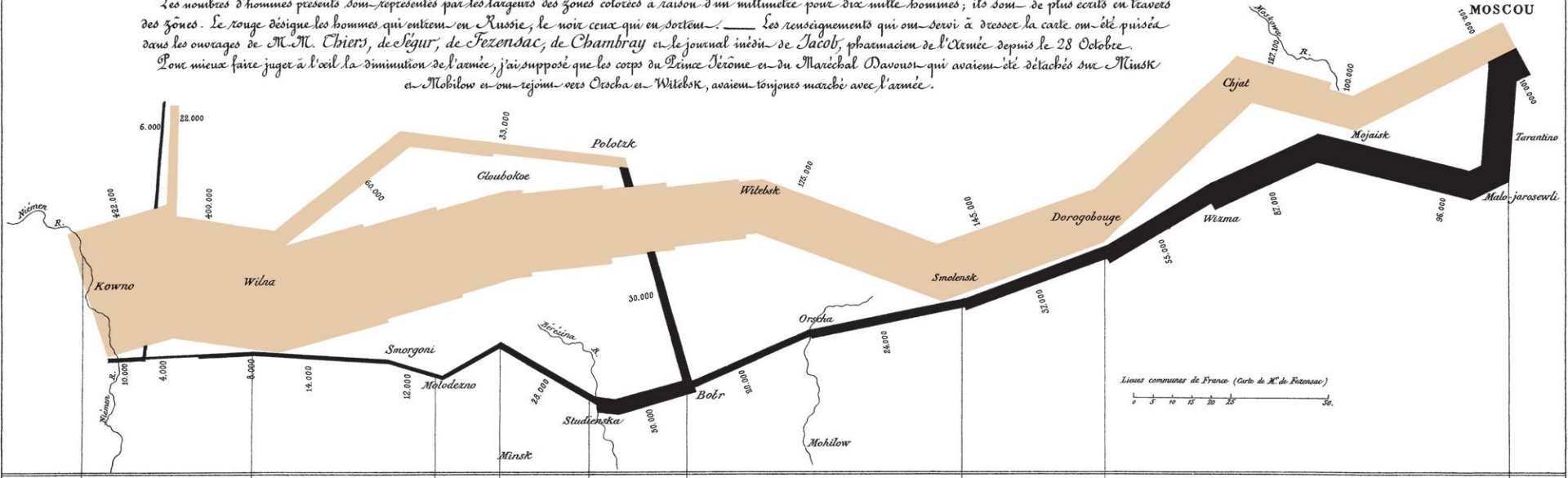
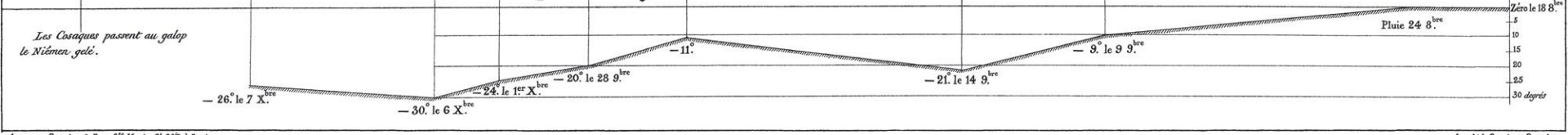
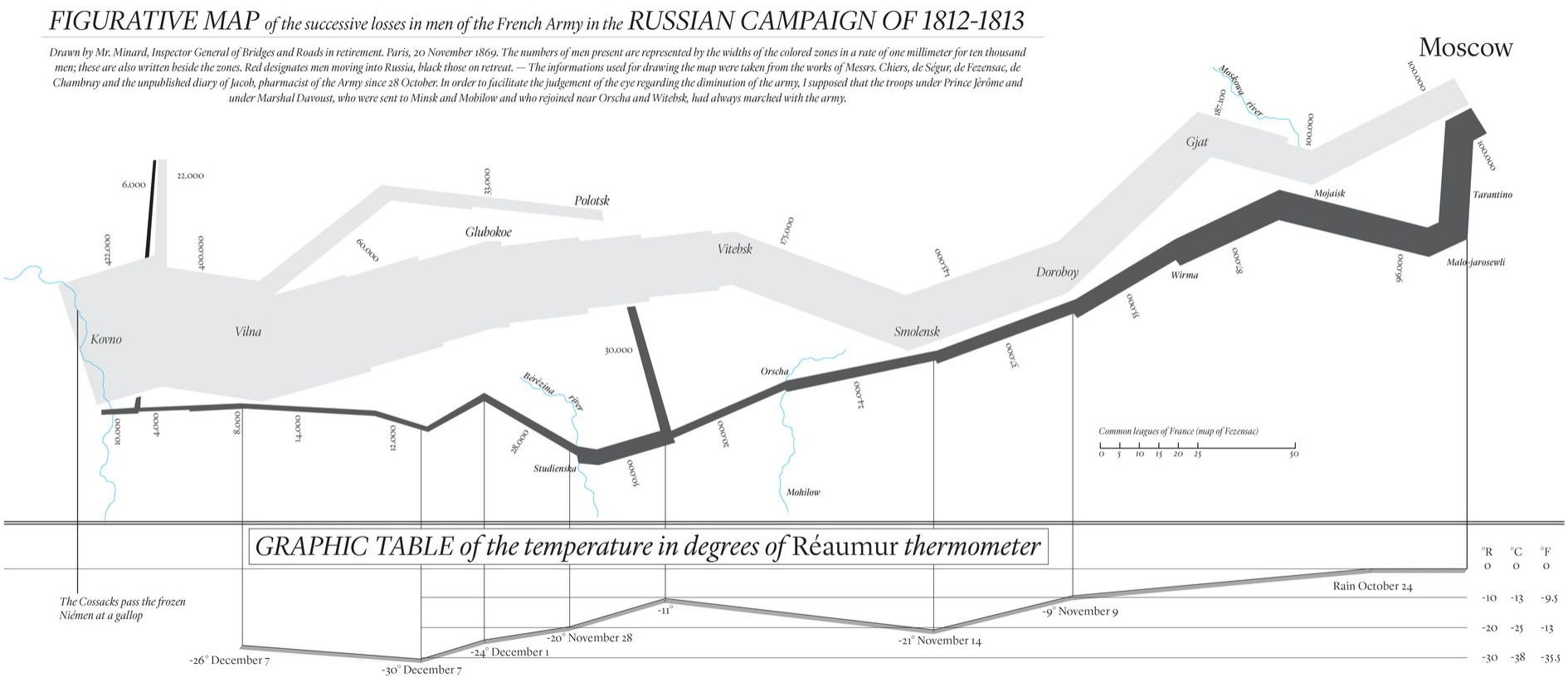


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



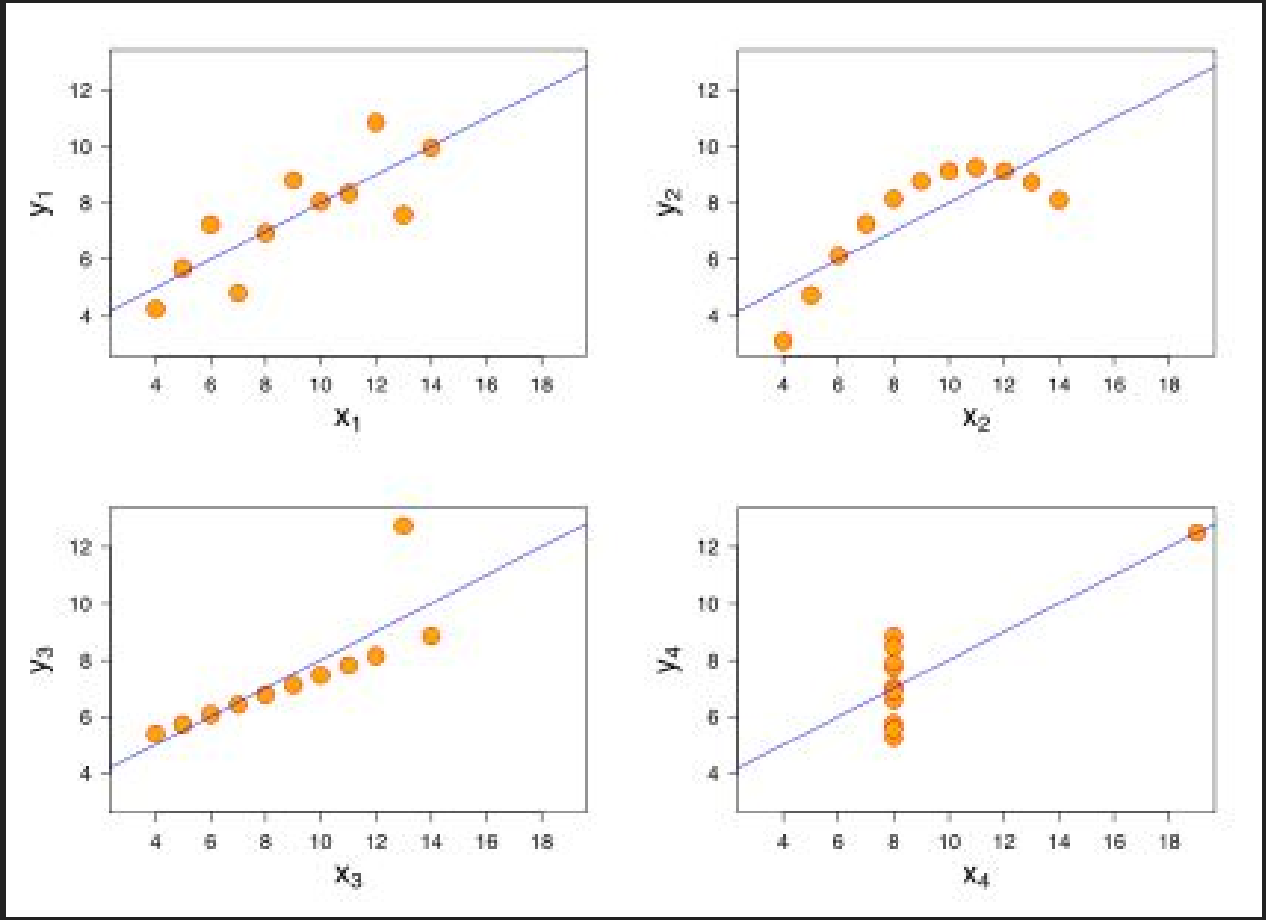
Autog. par Regnier, 8. Pas. Ste Marie St Gdin à Paris.

Imp. Lith. Regnier et Dourdet.



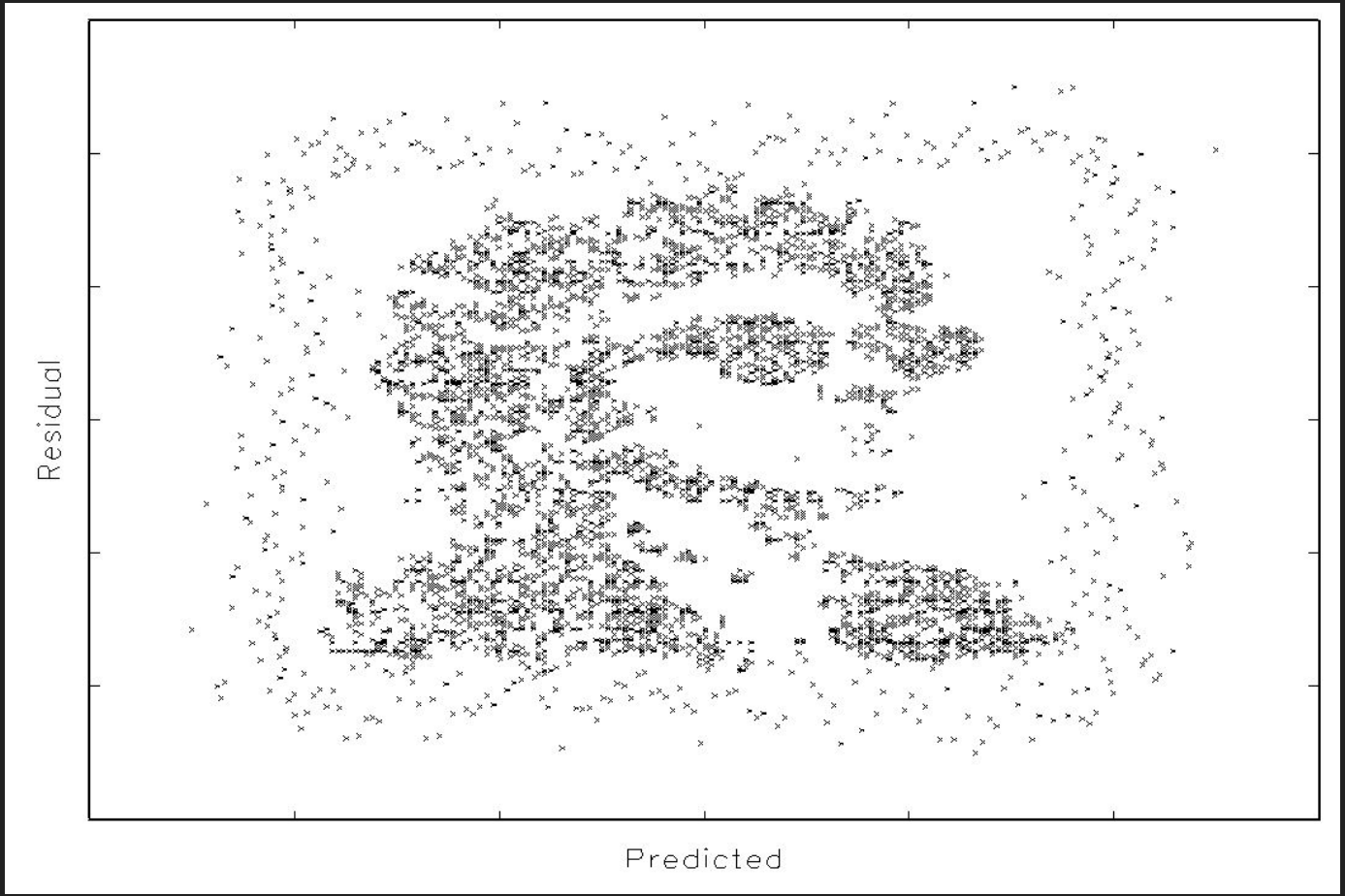
https://upload.wikimedia.org/wikipedia/commons/5/5d/Minard_map_of_napoleon.png

Why plot



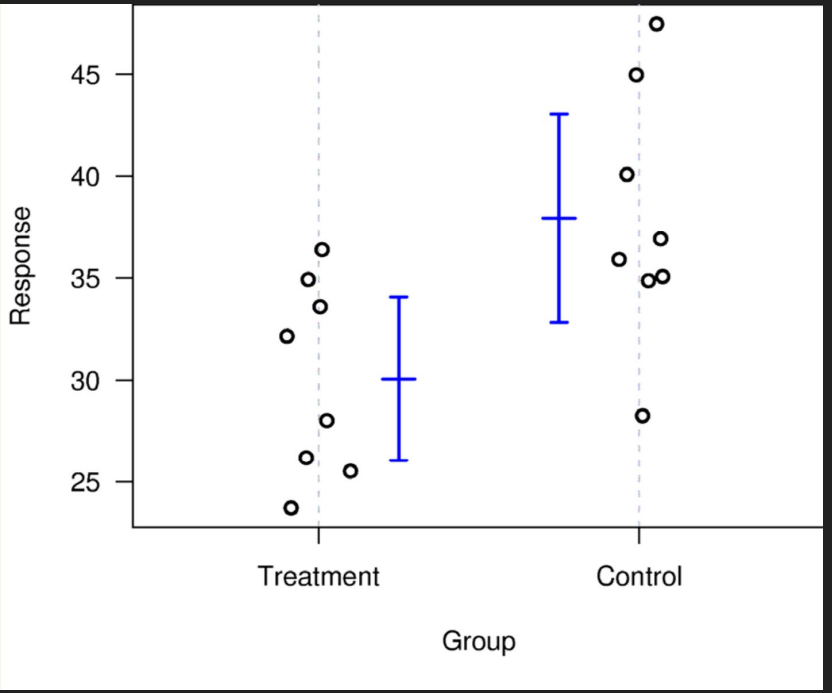
http://en.wikipedia.org/wiki/Anscombe's_quartet

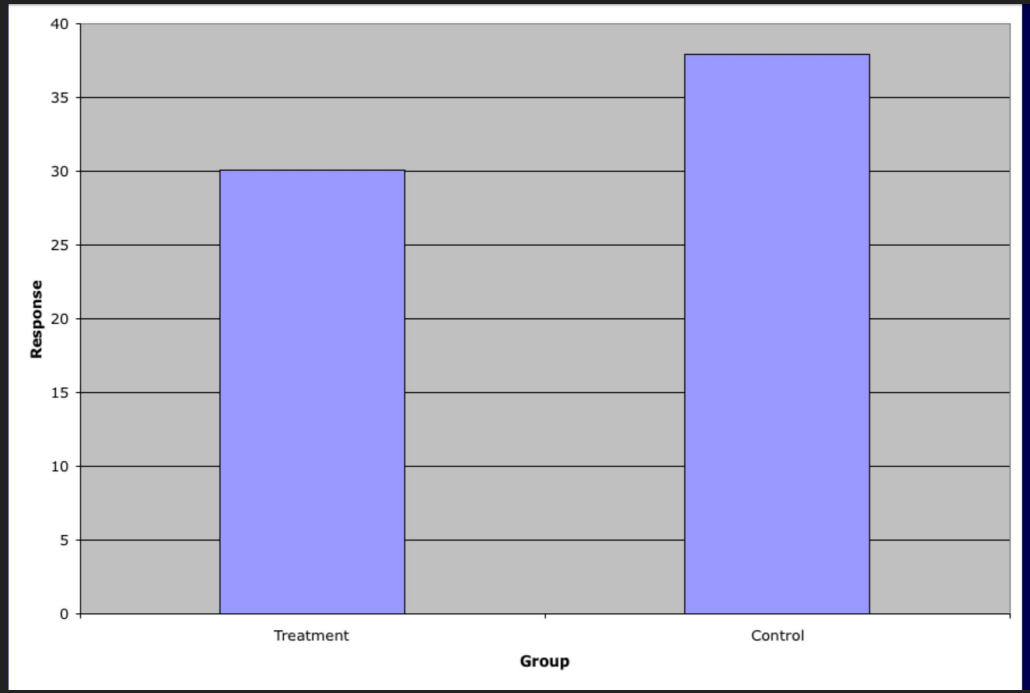
Stefanski's residual (sur)realism



Some general principles

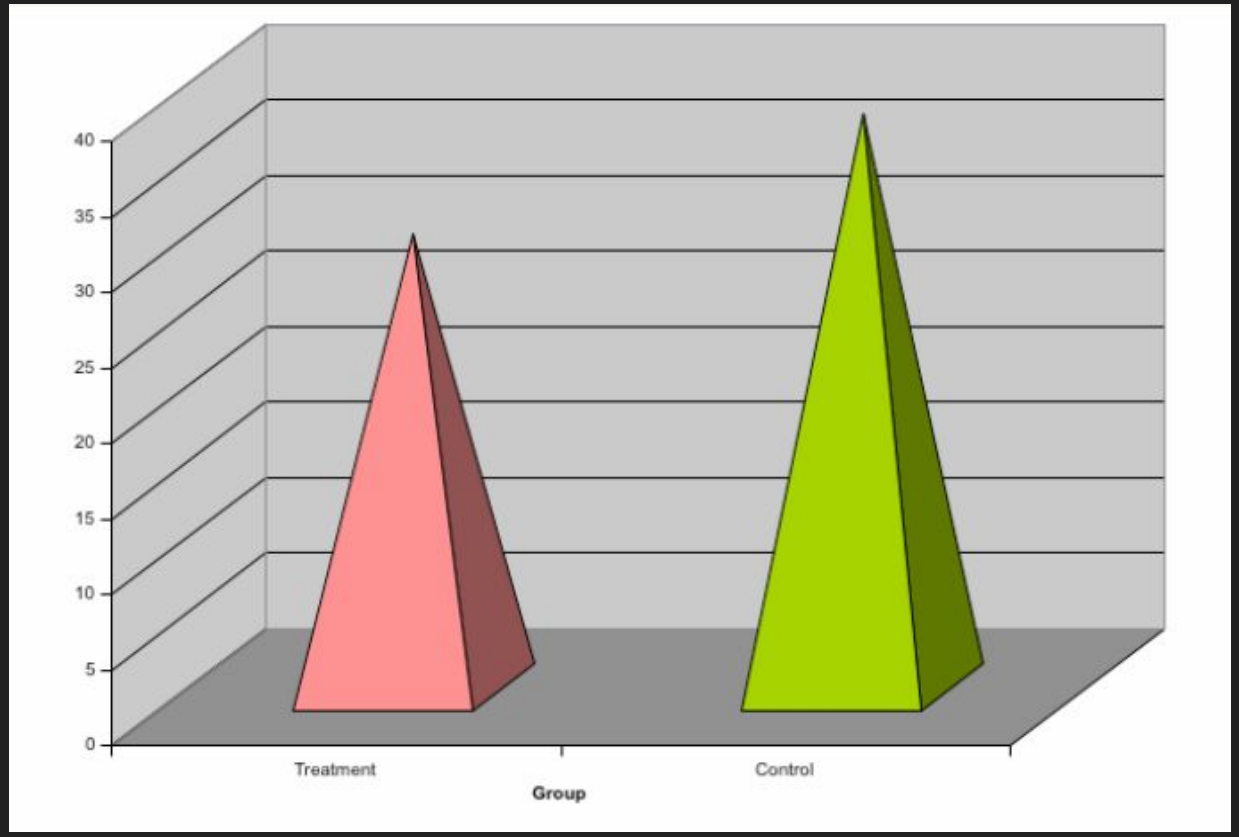
Maximize data / ink ratio (Tufte)





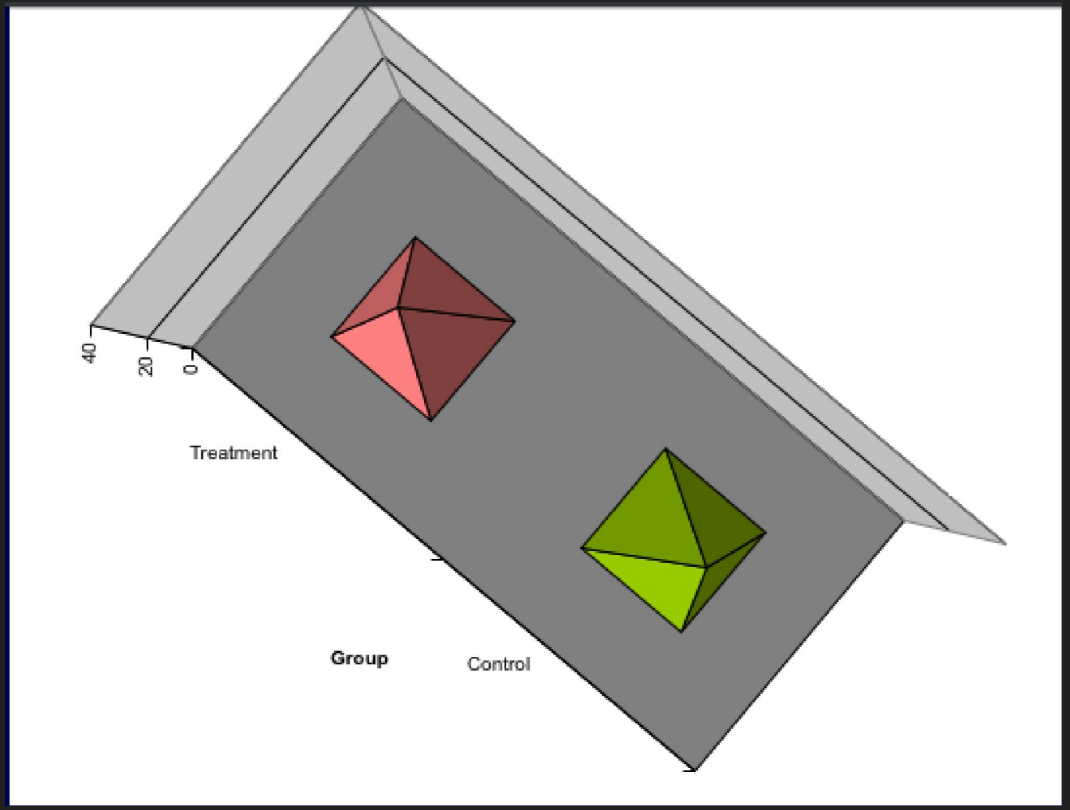
slide courtesy of K Broman

Generally, don’t use 3D



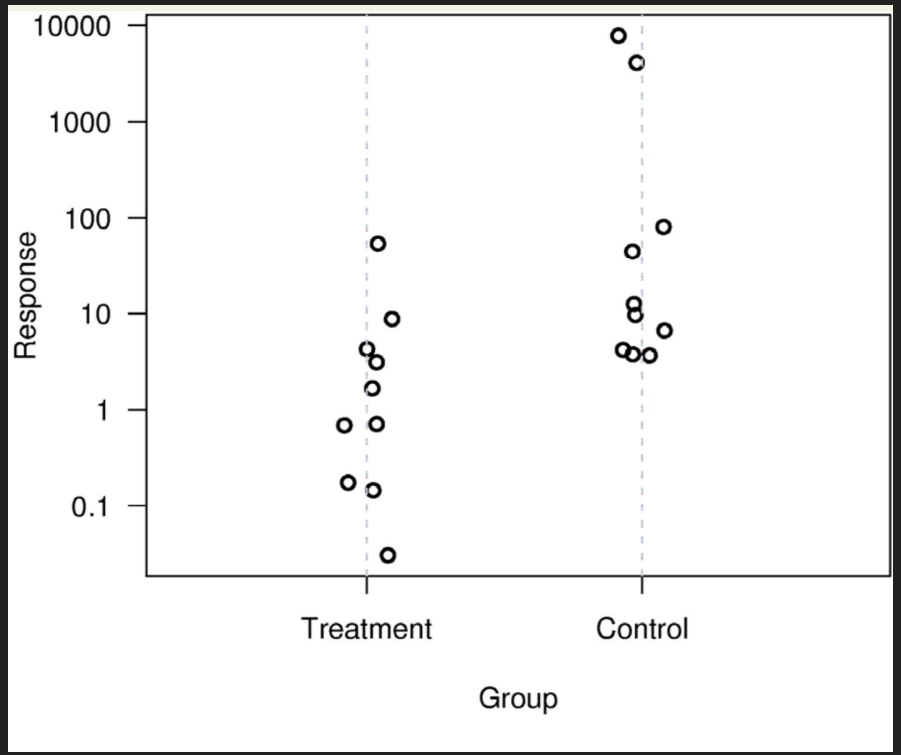
slide courtesy of K Broman

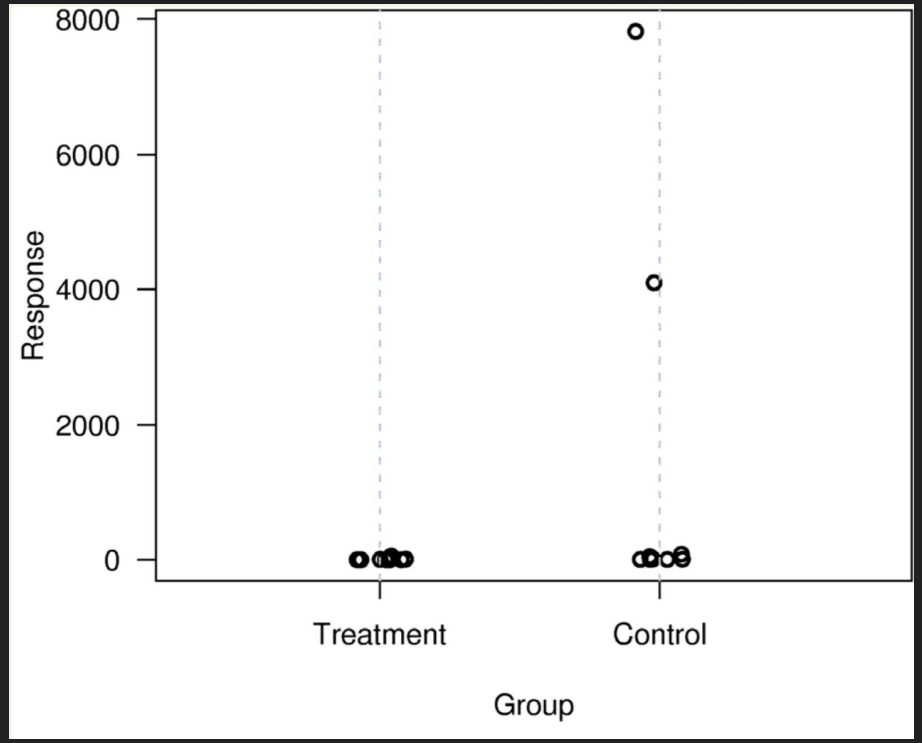
Generally, don’t use 3D



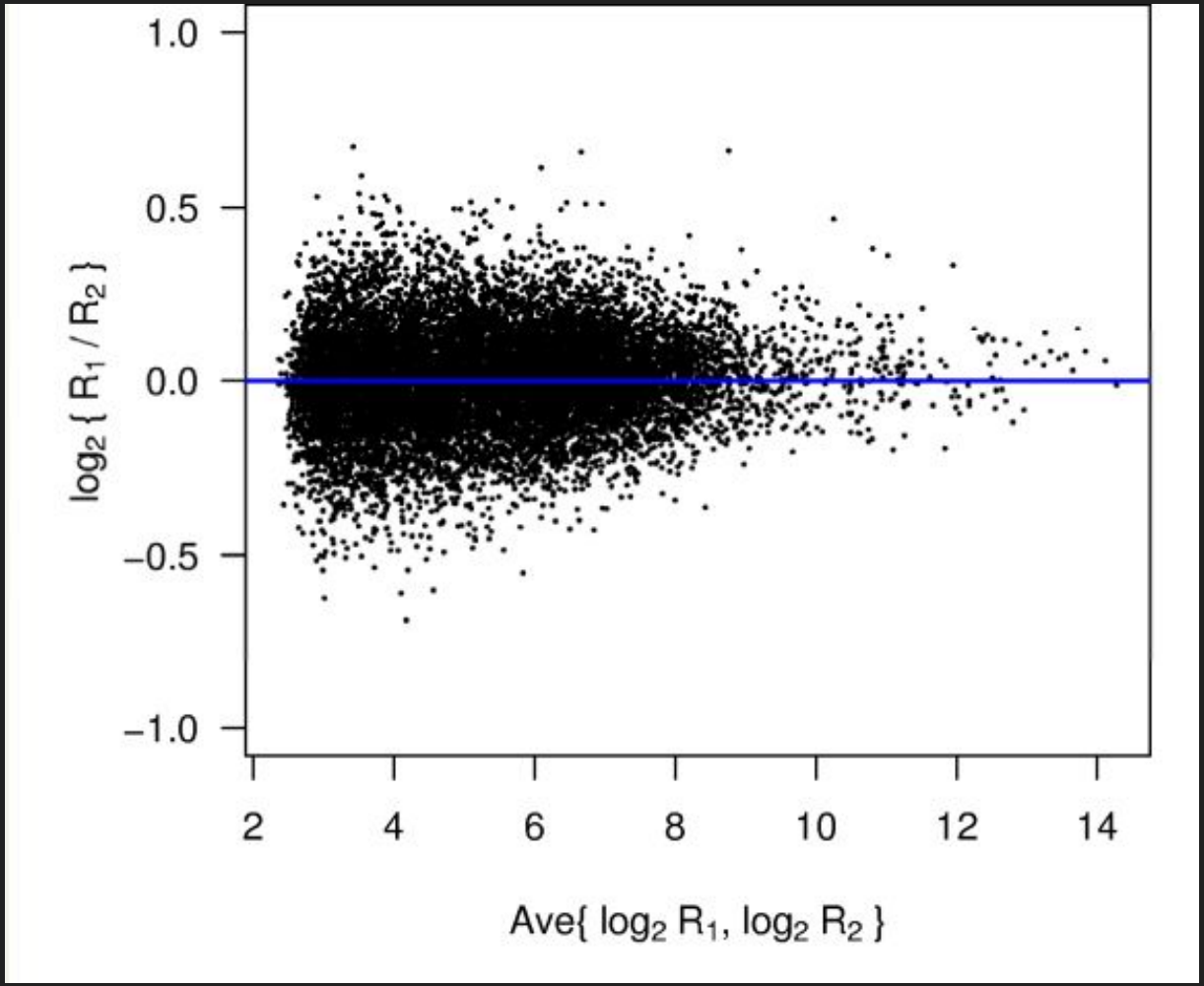
slide courtesy of K Broman

Logging

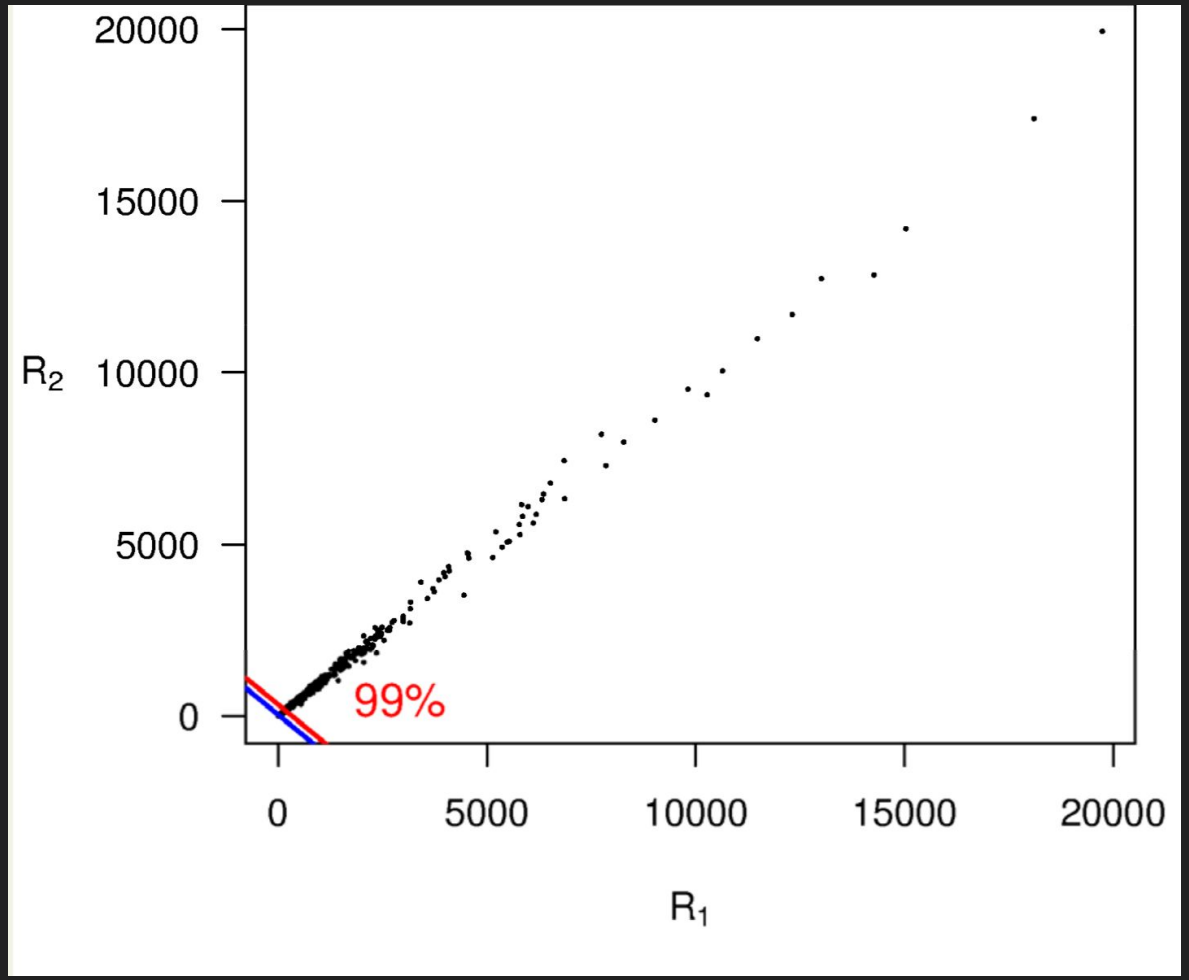




slide courtesy of K Broman



slide courtesy of K Broman



slide courtesy of K Broman

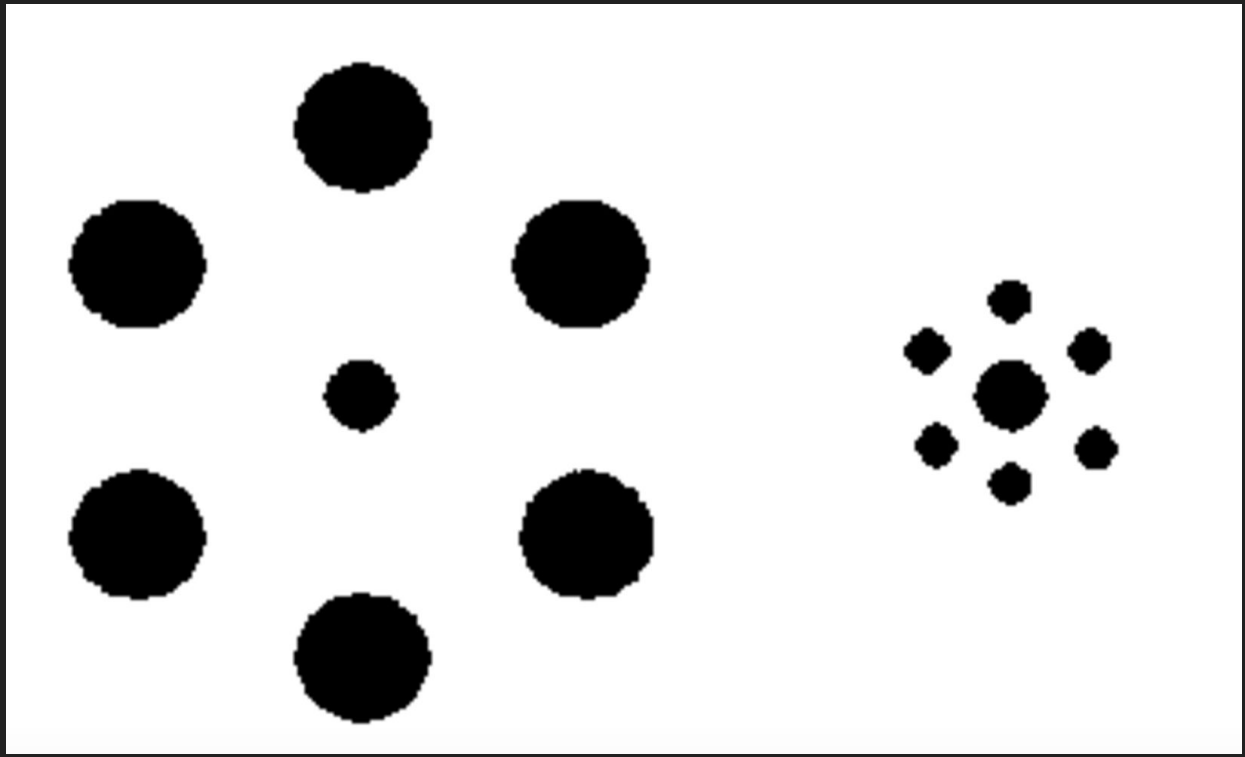
Characteristics of exploratory plots

- They are made quickly
- A large number are made
- The goal is for personal understanding
- Axes/legends are generally cleaned up
- Color/size are primarily used for information

Theory of EDA

- EDA is part statistics, part psychology
- Unfortunately we (humans) are designed to find patterns even when there aren't any
- Visual perception is biased by your humanness.
- The key goal in EDA is to not trick yourself

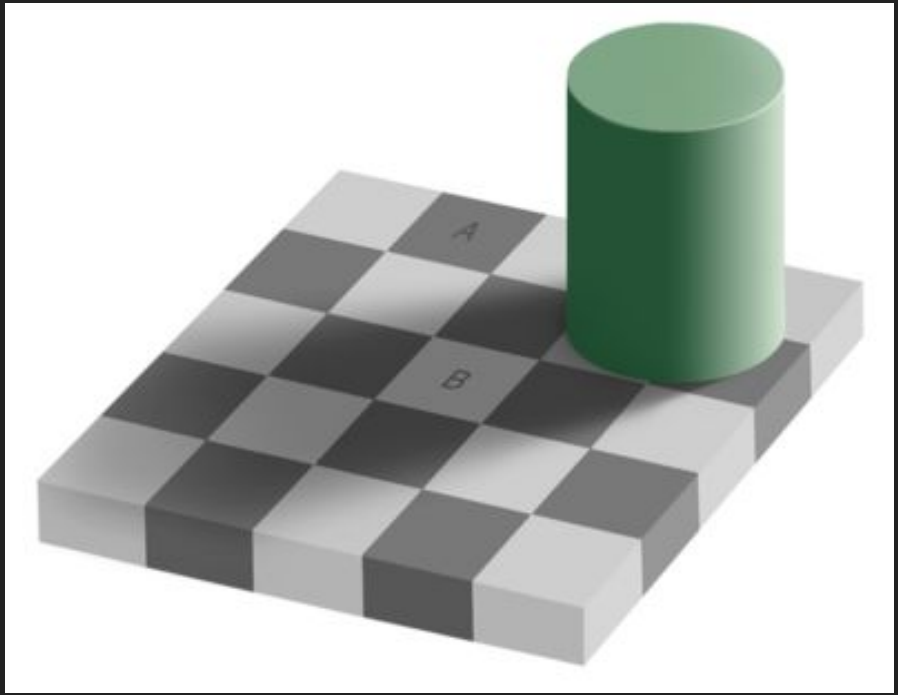
What optical illusions teach us about plotting



http://brainden.com/visual-illusions.htm

slide courtesy of J Leek

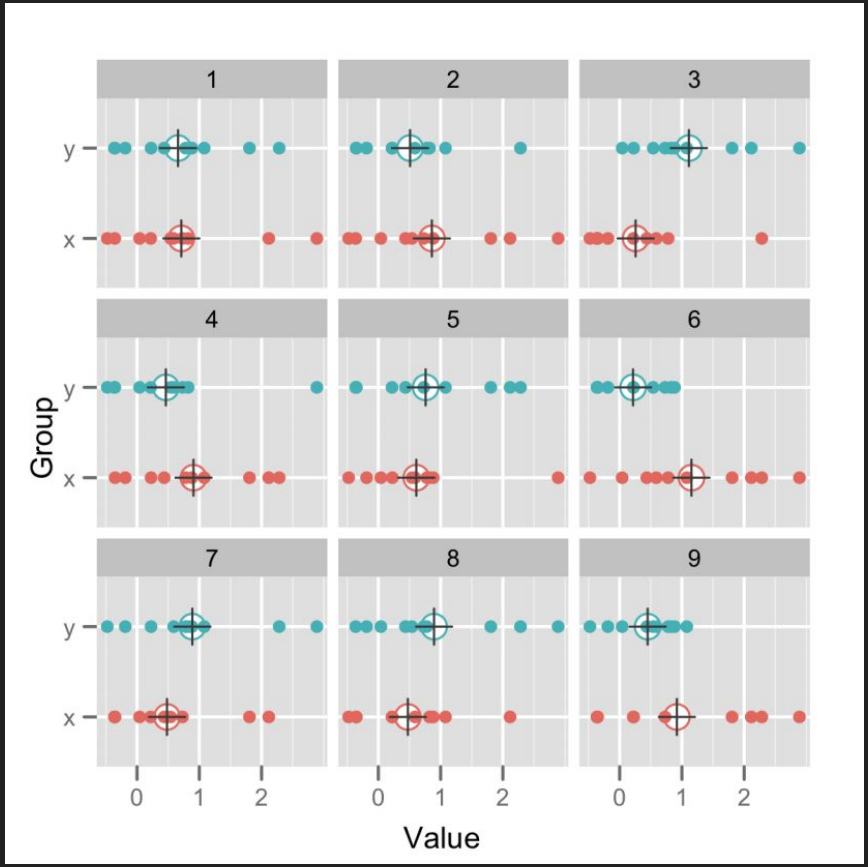
What optical illusions teach us about plotting



http://blog.revolutionanalytics.com/2012/12/create-optical-illusions-with-r.html

slide courtesy of J Leek

Plots can be thought of as test statistics



http://vita.had.co.nz/papers/inference-infovis.pdf

slide courtesy of J Leek

Background perceptual tasks

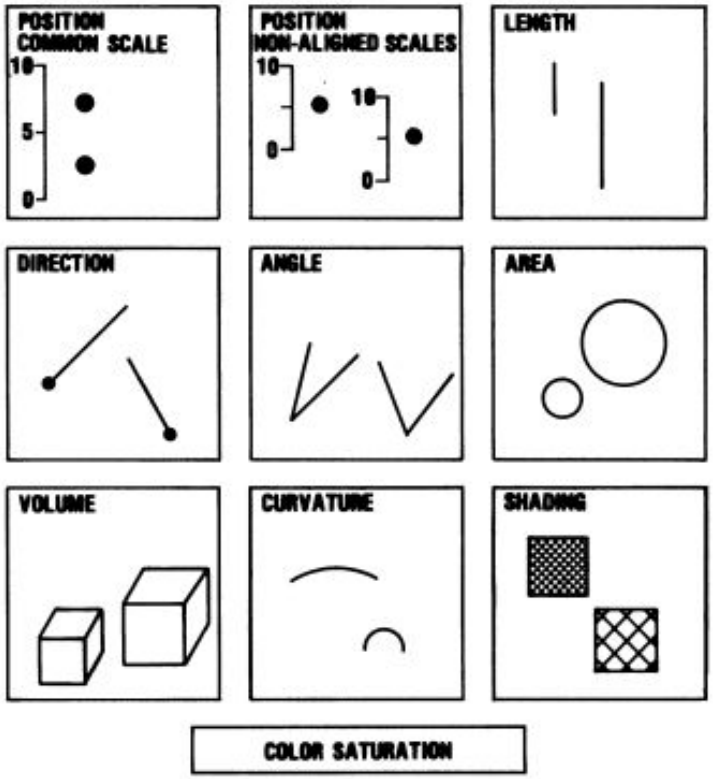


Figure 1. Elementary perceptual tasks.

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

slide courtesy of J Leek

Position vs. length

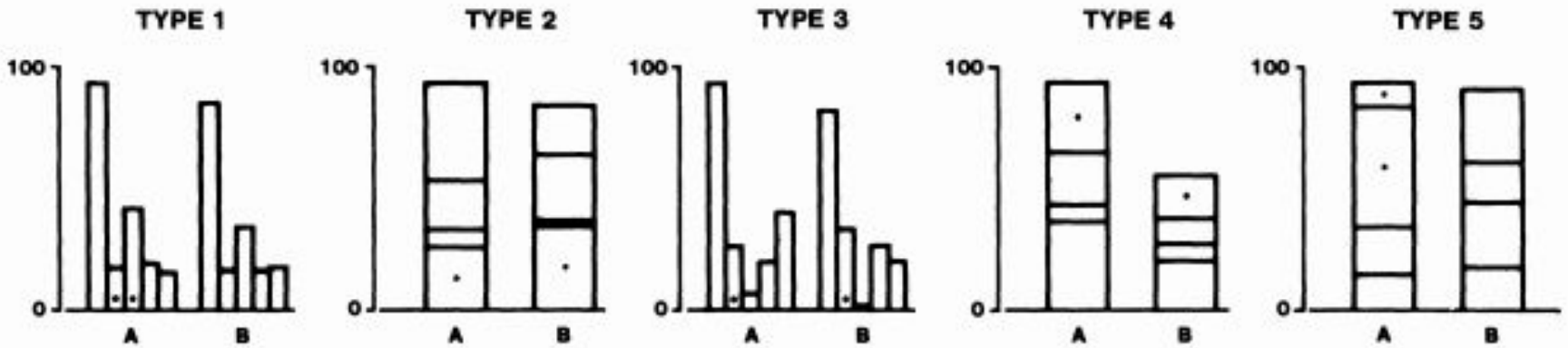
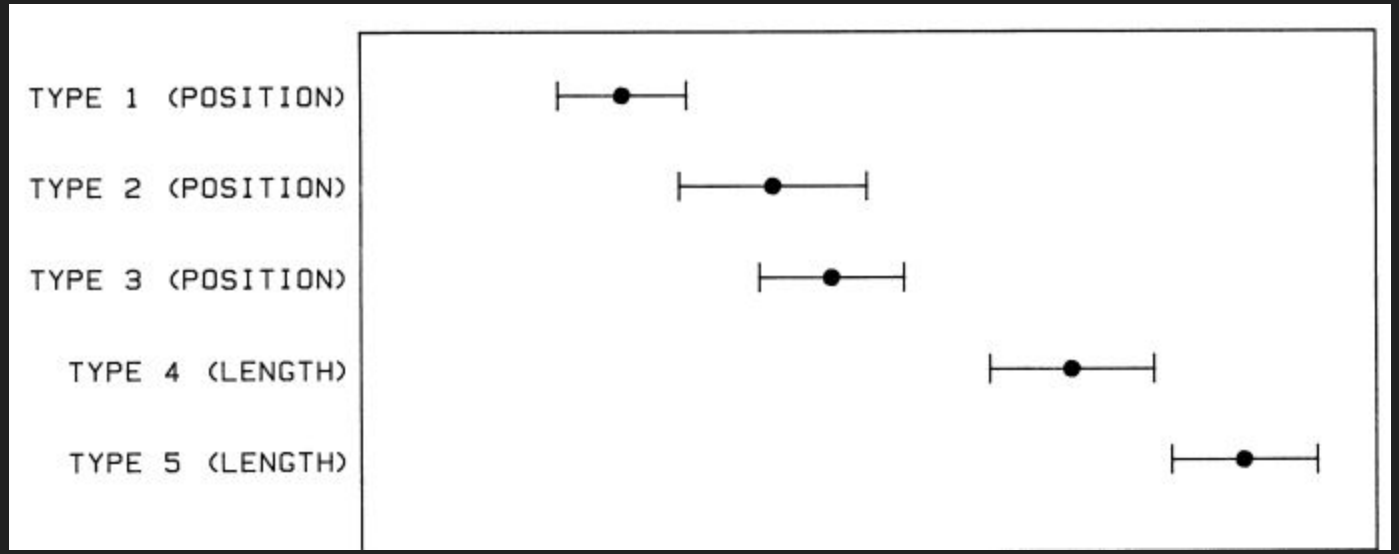


Figure 4. Graphs from position–length experiment.

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

slide courtesy of J Leek

Position vs. length results (log abs difference)



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

slide courtesy of J Leek

Position vs. angle

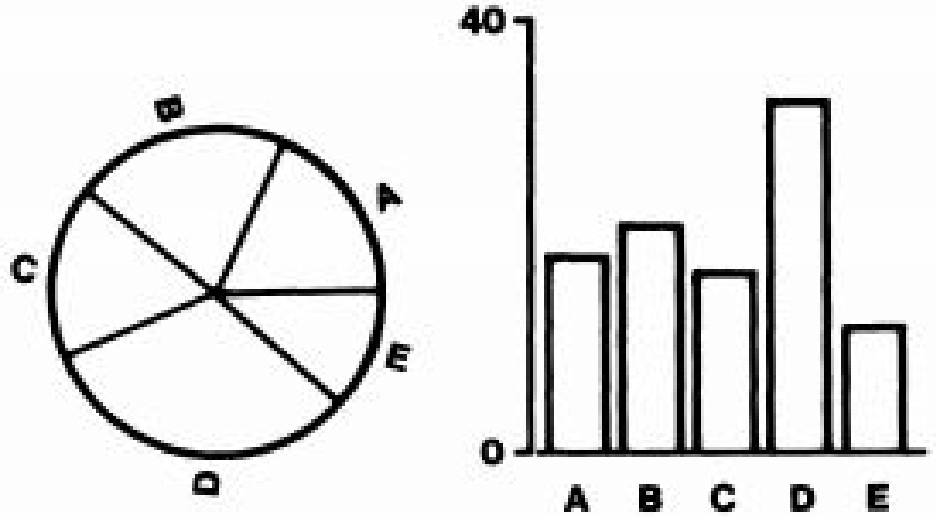
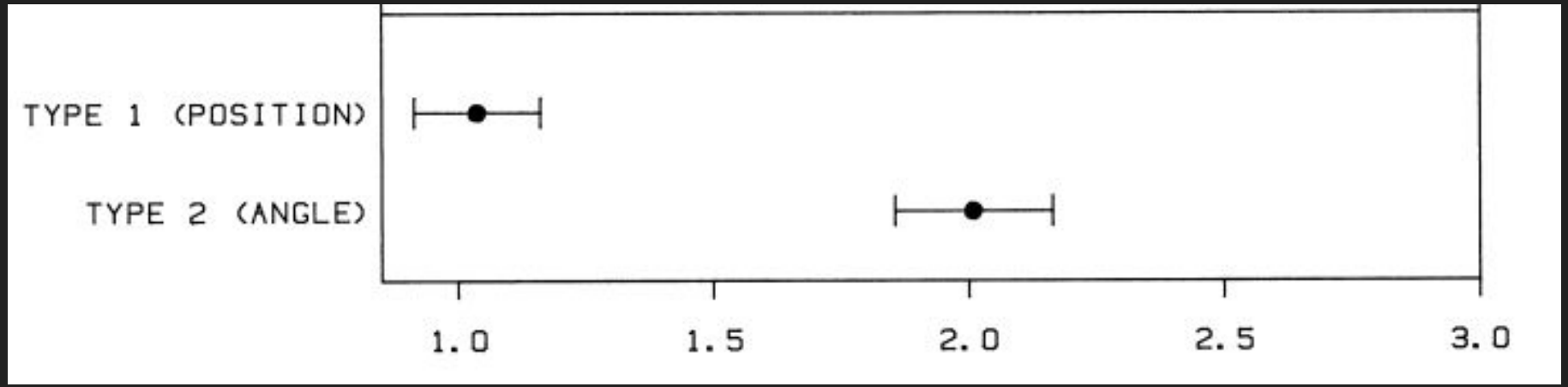


Figure 3. Graphs from position–angle experiment.

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

slide courtesy of J Leek

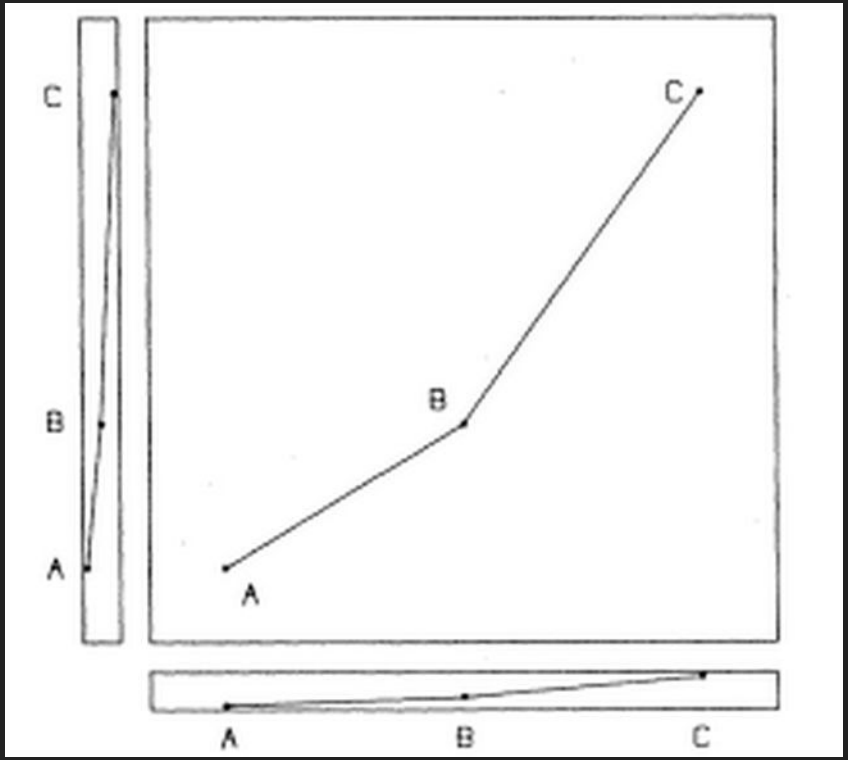
Position vs. angle - results (log abs difference)



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

slide courtesy of J Leek

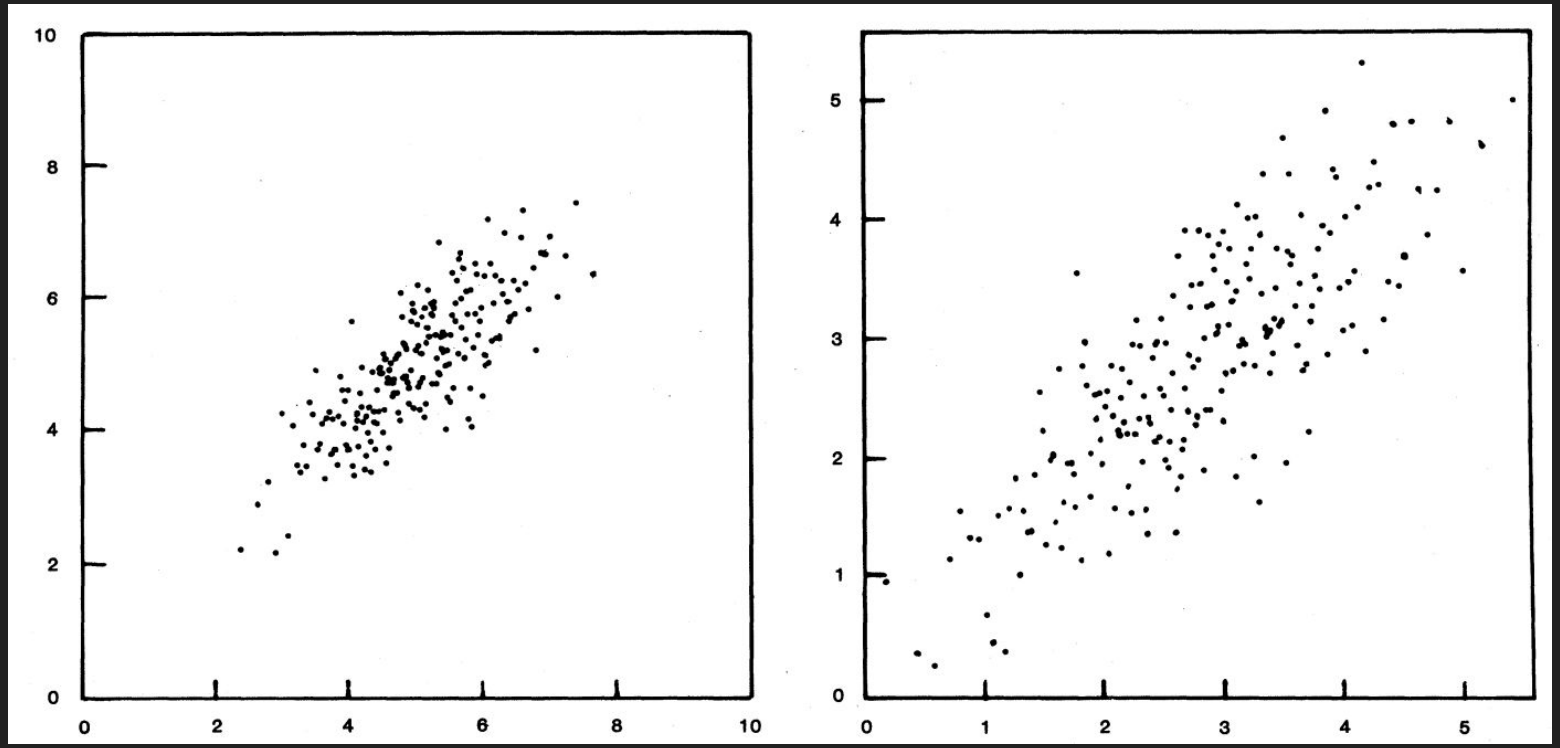
The worst - maybe slopes?



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

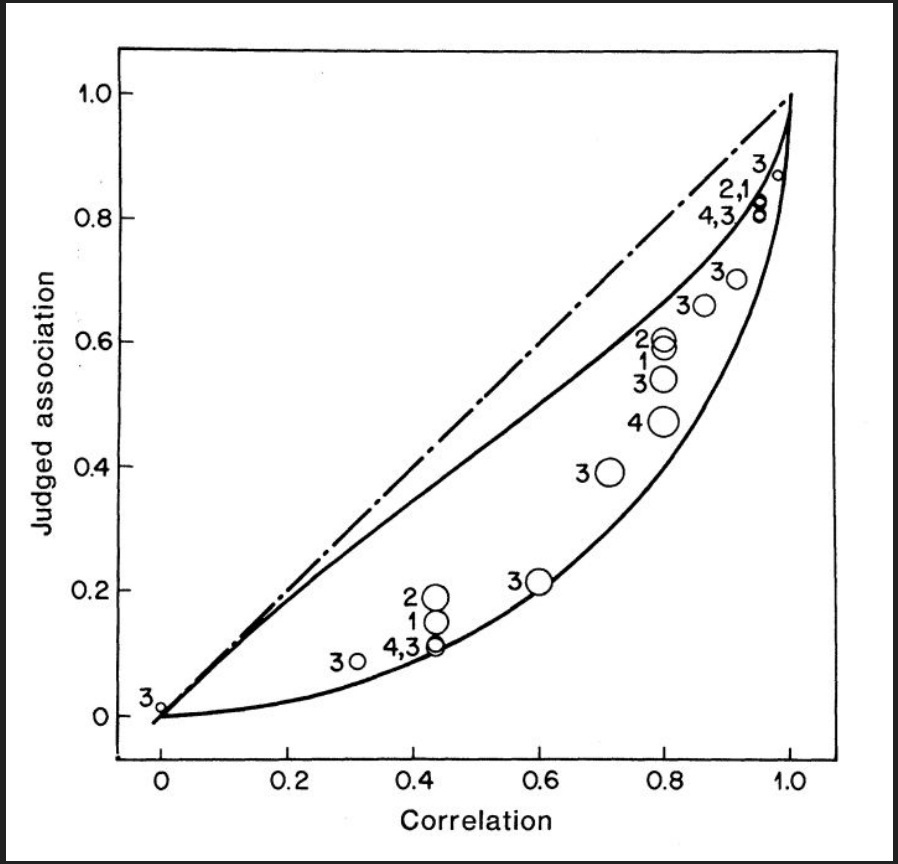
slide courtesy of J Leek

Scale matters



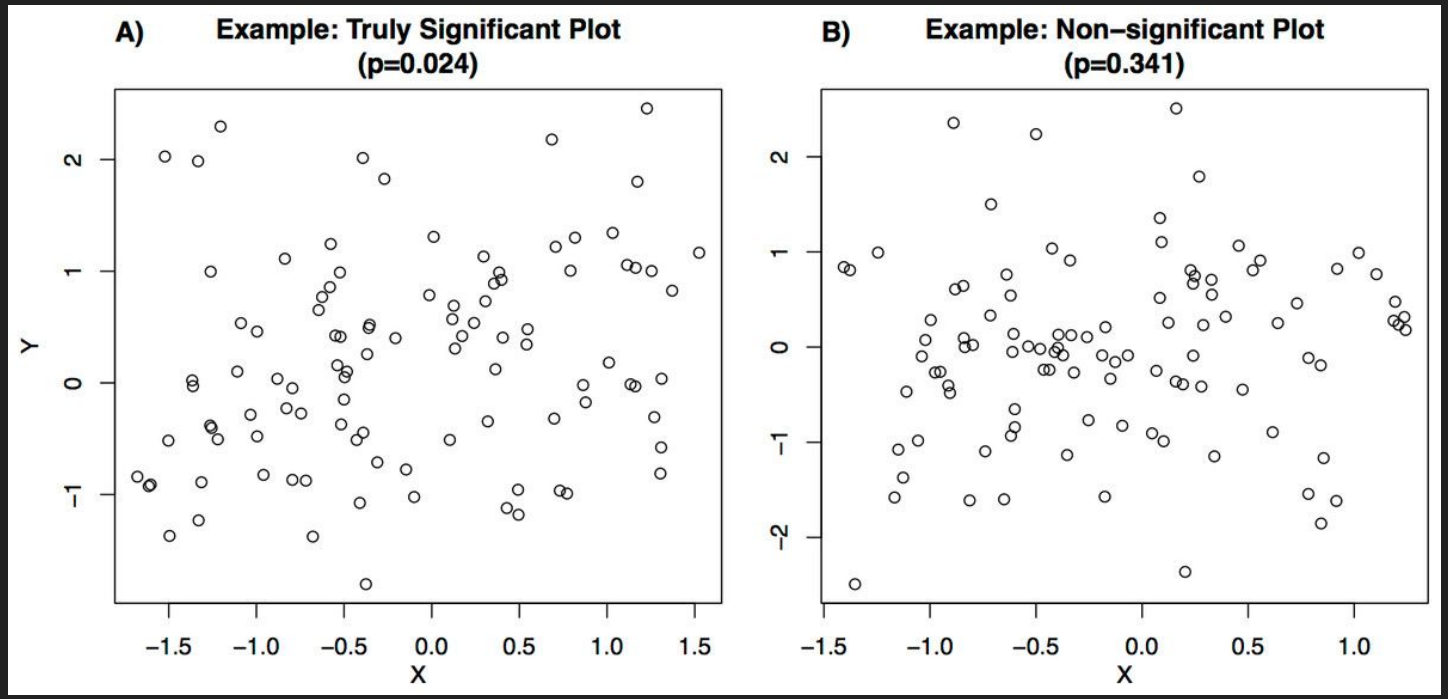
http://statweb.stanford.edu/~cgates/PERSI/papers/scatter82.pdf slide courtesy of J Leek

People perceive correlations weirdly



http://statweb.stanford.edu/~cgates/PERSI/papers/scatter82.pdf slide courtesy of J Leek

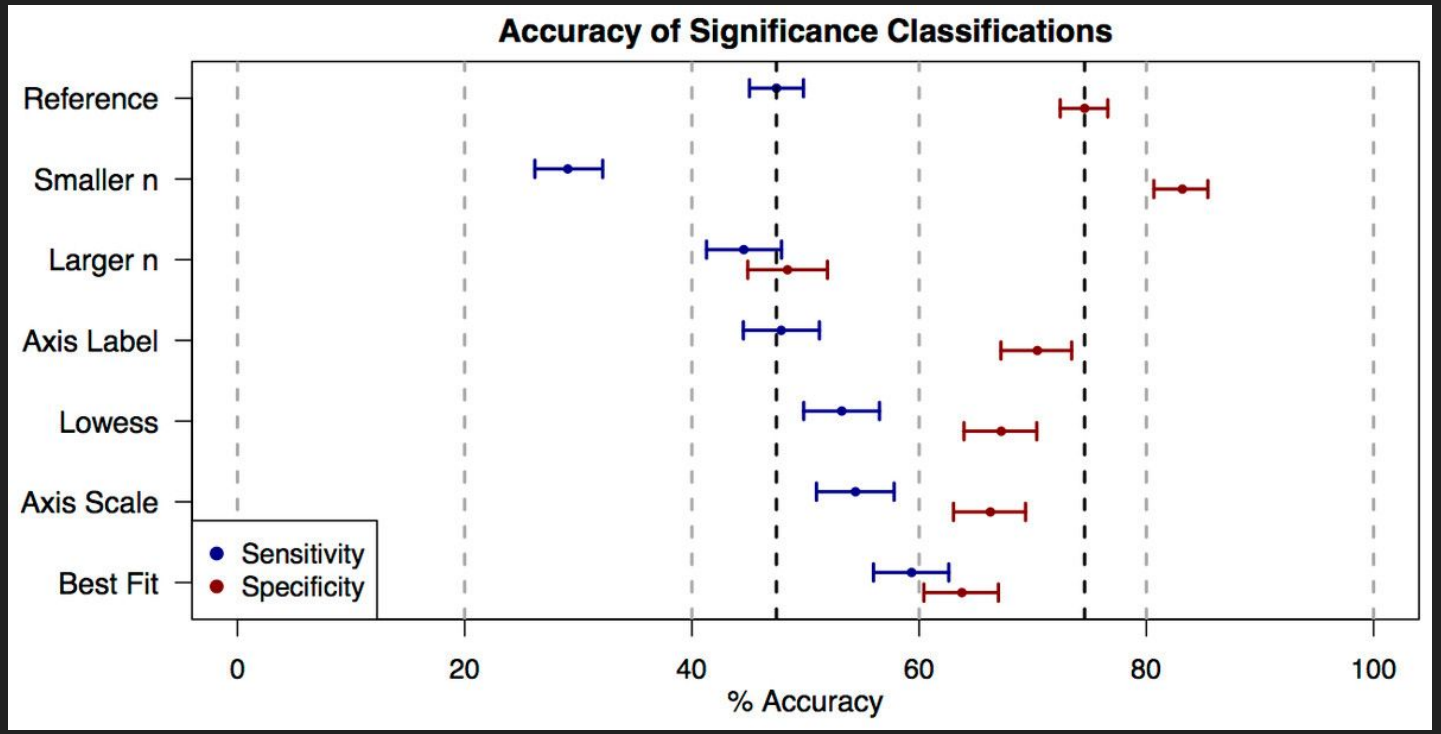
Detecting even linear relationships



https://peerj.com/articles/589/

slide courtesy of J Leek

People are bad at significance in plots



https://peerj.com/articles/589/

slide courtesy of J Leek

Summary

- Use common scales when possible
- When possible use position comparisons
- Angle comparisons are hard to interpret (no piecharts!)
- No 3-D barcharts
- Be careful not to "fool" yourself about significance (either way)

slide courtesy of J Leek

Acknowledgements

Jeff Leek

Karl Broman

Jenny Bryan

Genevera Allen

XKCD

Wikipedia

Leonard Stefanski

Rstudio

Some resources