Cheap, Fast, or Good - Pick Two

Data Instruction in the Age of Data Science

Joel Herndon, Duke University
Justin Joque, University of Michigan
Angela Zoss, Duke University

http://bit.ly/IASSIST-DataInstruction

Group notes at:
http://bit.ly/IASSIST-DataInstruction

Introduction

Methodology

John W. Tukey

EXPLORATORY DATA ANALYSIS

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

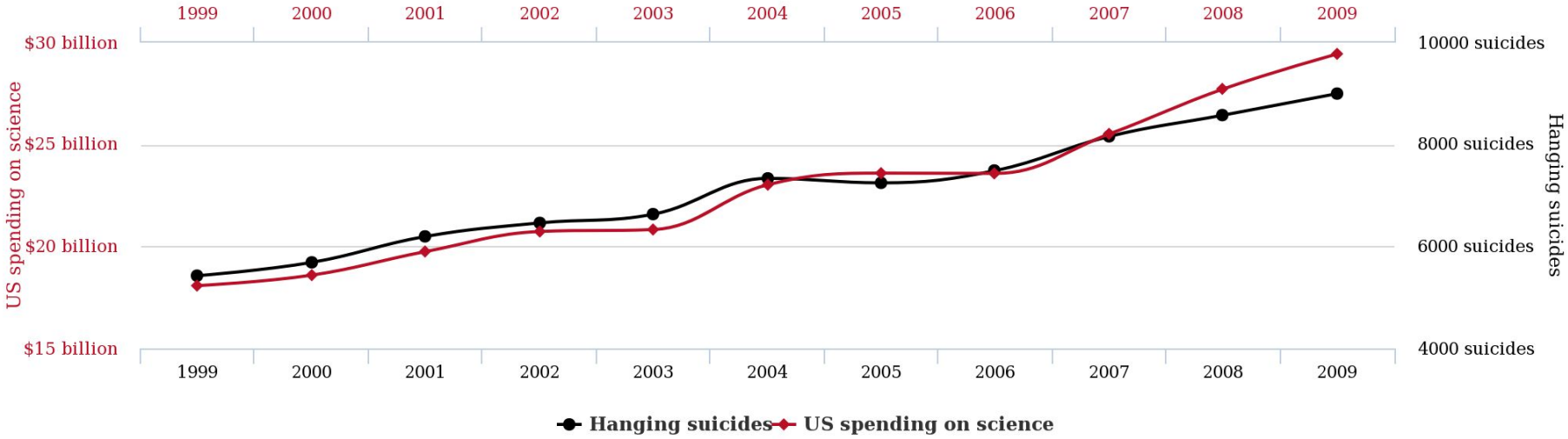
Illustration: Marian Bantjes

"All models are wrong, but some are useful."

US spending on science, space, and technology

correlates with

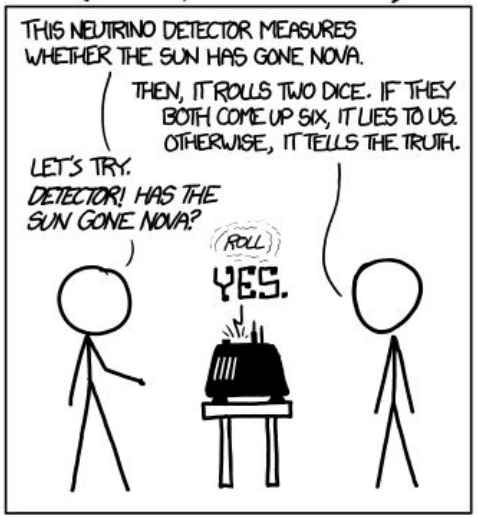
Suicides by hanging, strangulation and suffocation

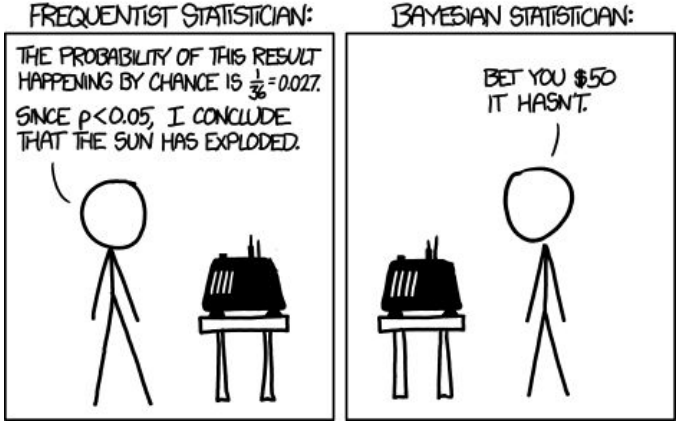


tylervigen.com

DID THE SUN JUST EXPLODE?

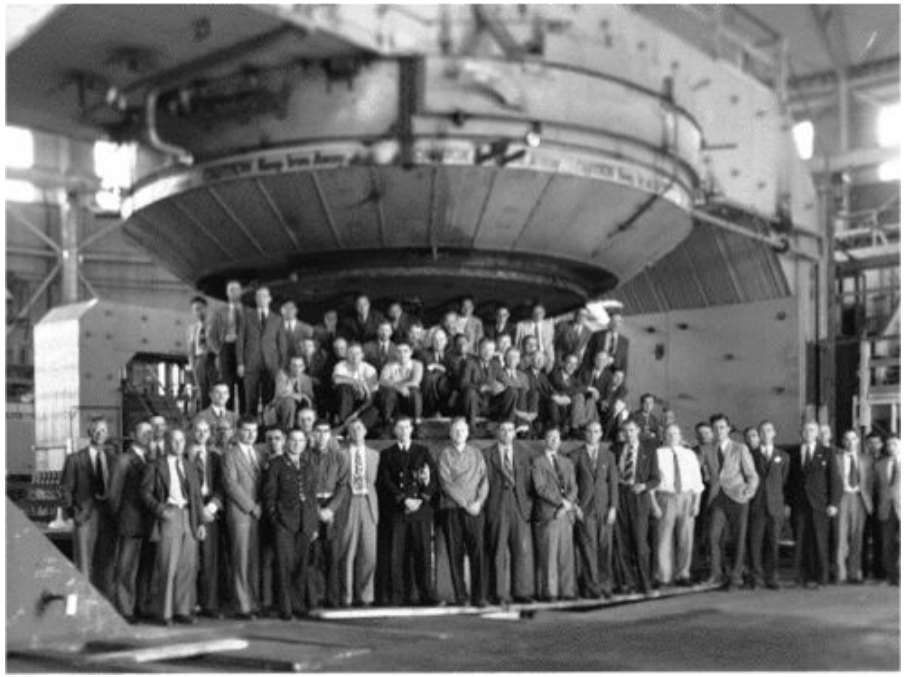
(IT'S NIGHT, SO WE'RE NOT SURE.)

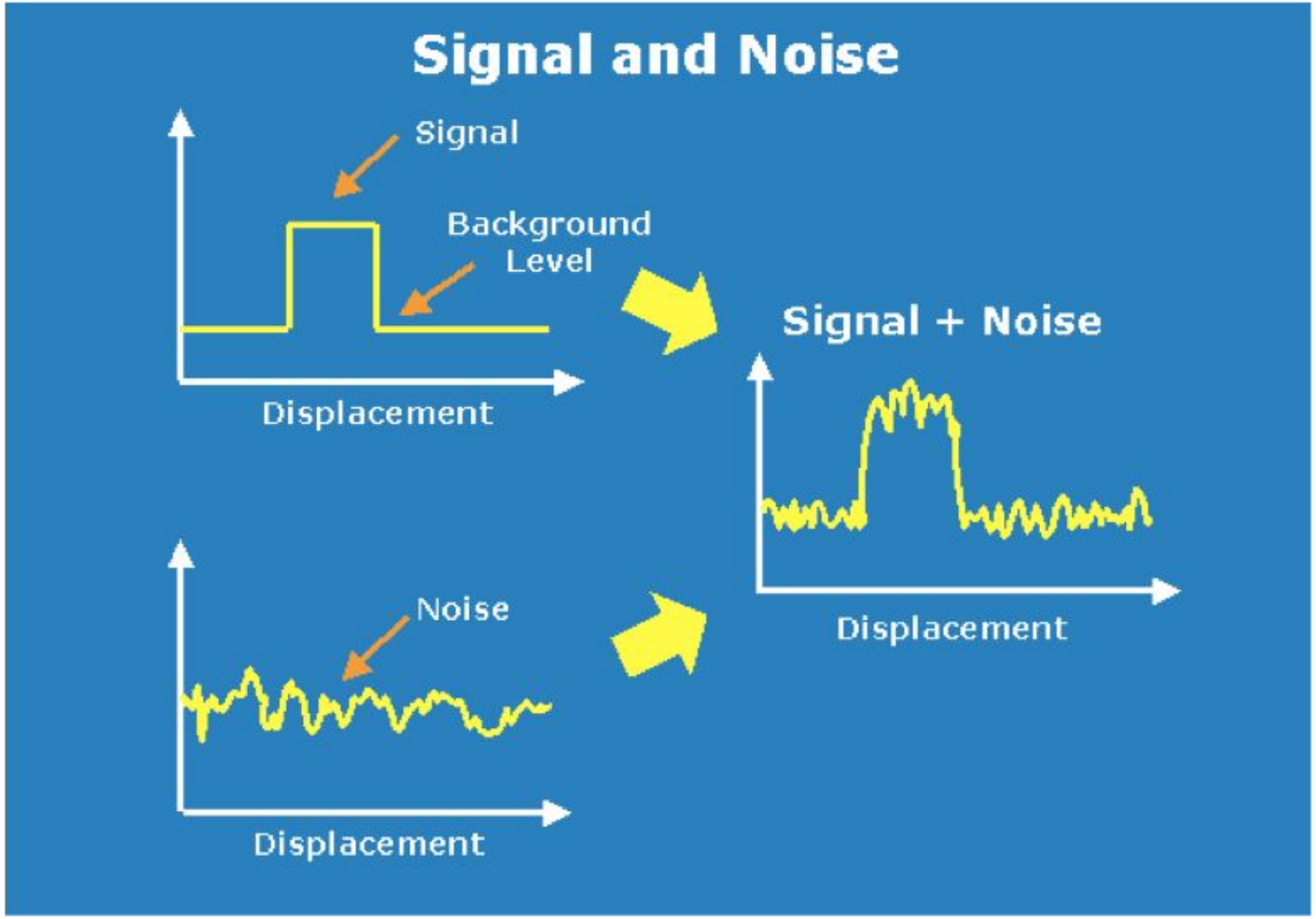




The first to disintegrate a nucleus was Rutherford, and there is a picture of him holding the apparatus in his lap. I then always remember the later picture when one of the famous cyclotrons was built at Berkeley, and all of the people were sitting in the lap of the cyclotron.

-Maurice Goldhaber, former director of Brookhaven Laboratory




Signal and Noise
Signal
Background Level
Displacement
Signal + Noise
Displacement
Noise
Displacement

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

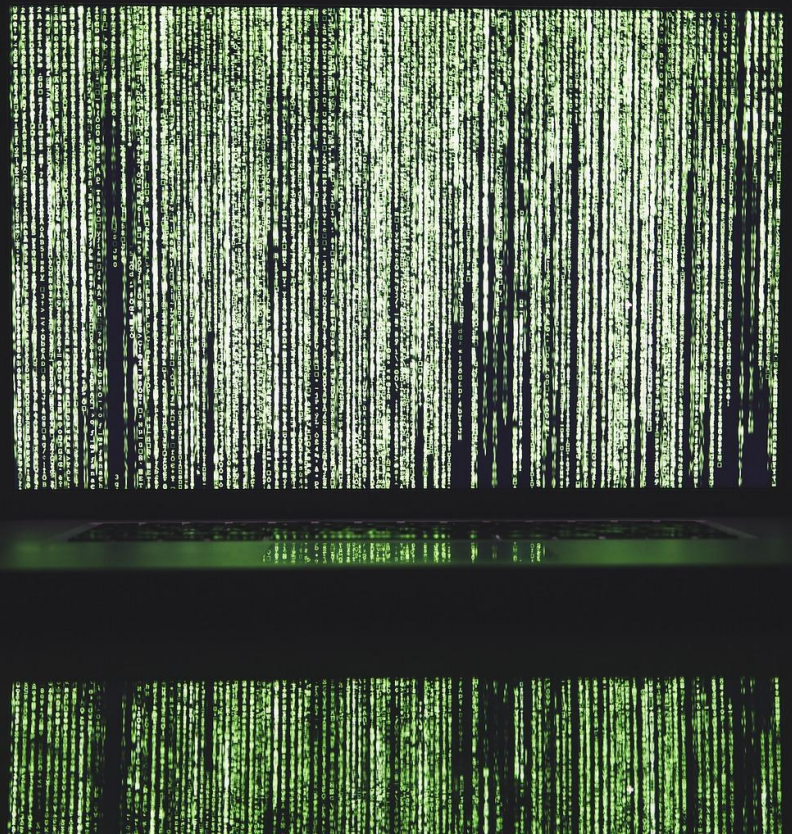
should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1 - \beta)R/(R$


Non est potestas Super Terram quæ Comparetur ei Iob. 41. 24.

Brief comments?

joque@umich.edu

Technology
for data
instruction

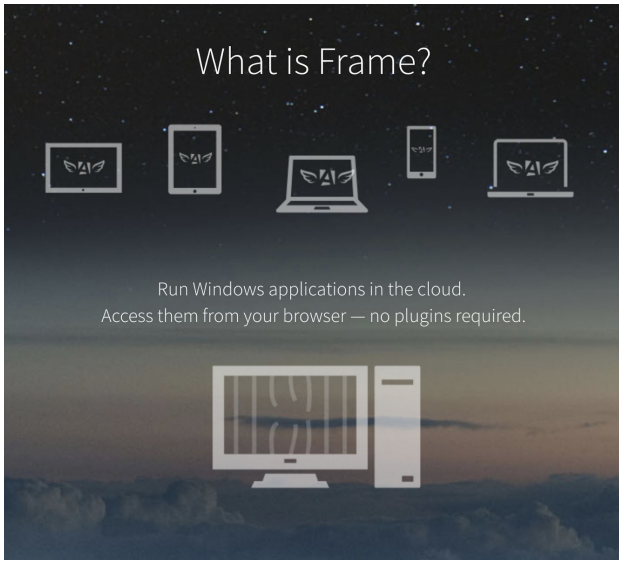


Instruction
of data
technology

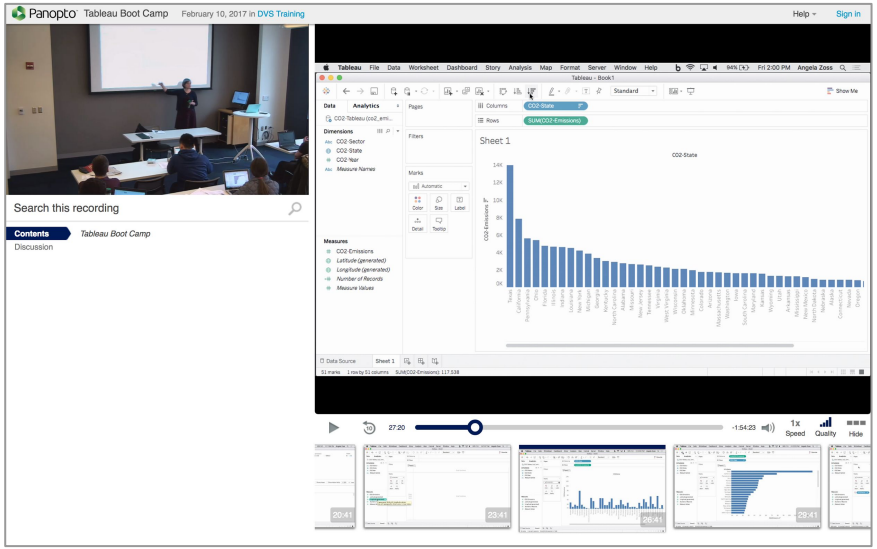
What technology should we be using to teach?

Hardware?

Software installation?



Physical or virtual?



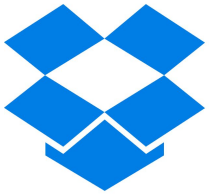
Sharing files?

SlideShare

figshare

box

GitHub



Dropbox

Atlassian

Bitbucket

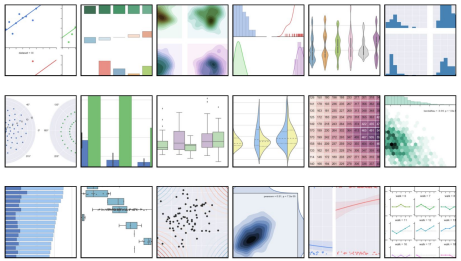
What technology should we be teaching?

Open access?

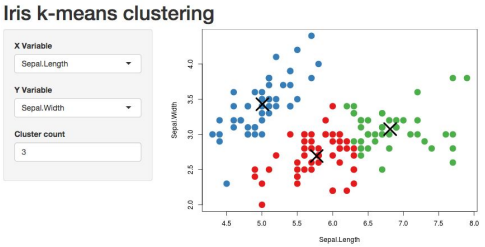
Free, but potentially unreliable

Scripting?

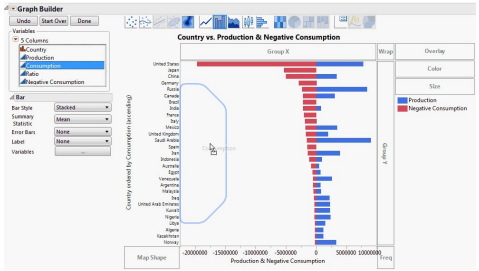
Powerful and reproducible, but complicated



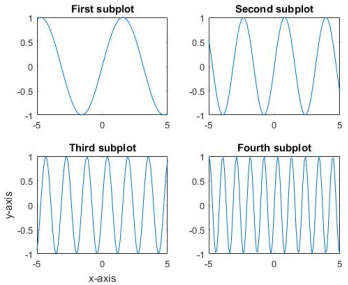
Python (fully open)



R (fully open)

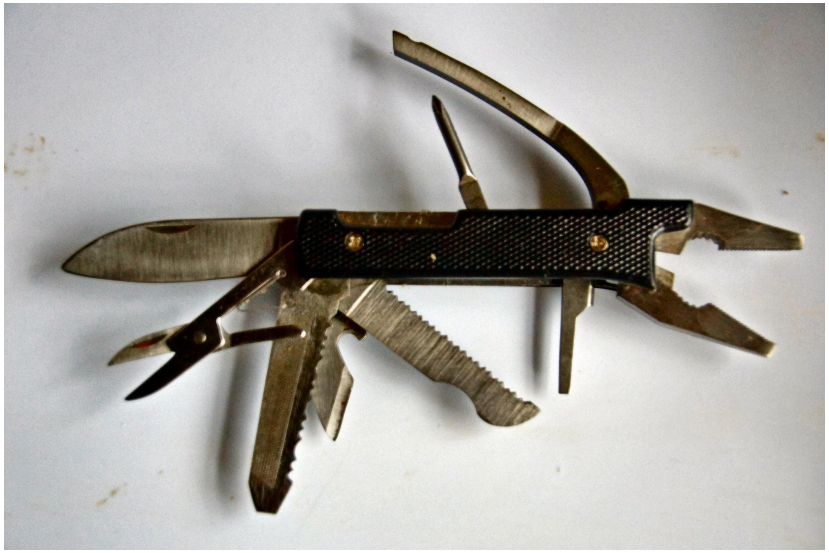


JMP Pro (proprietary)



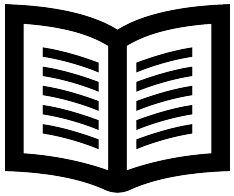
MATLAB (proprietary)

One-stop-shop or highly specialized tool?



What does it mean to teach a technology?

Learning/experience goals for data instruction

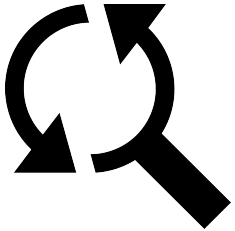


Vocabulary

Enjoyment

Mastery

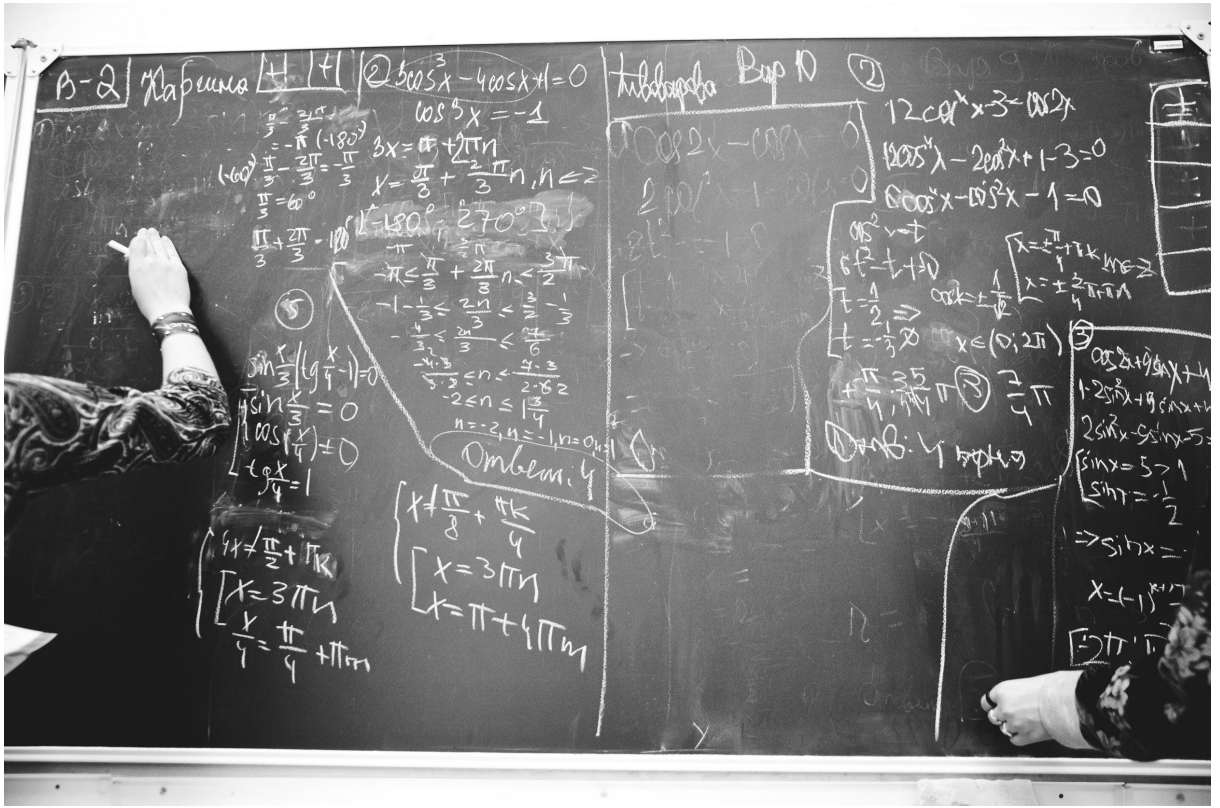
Exposure



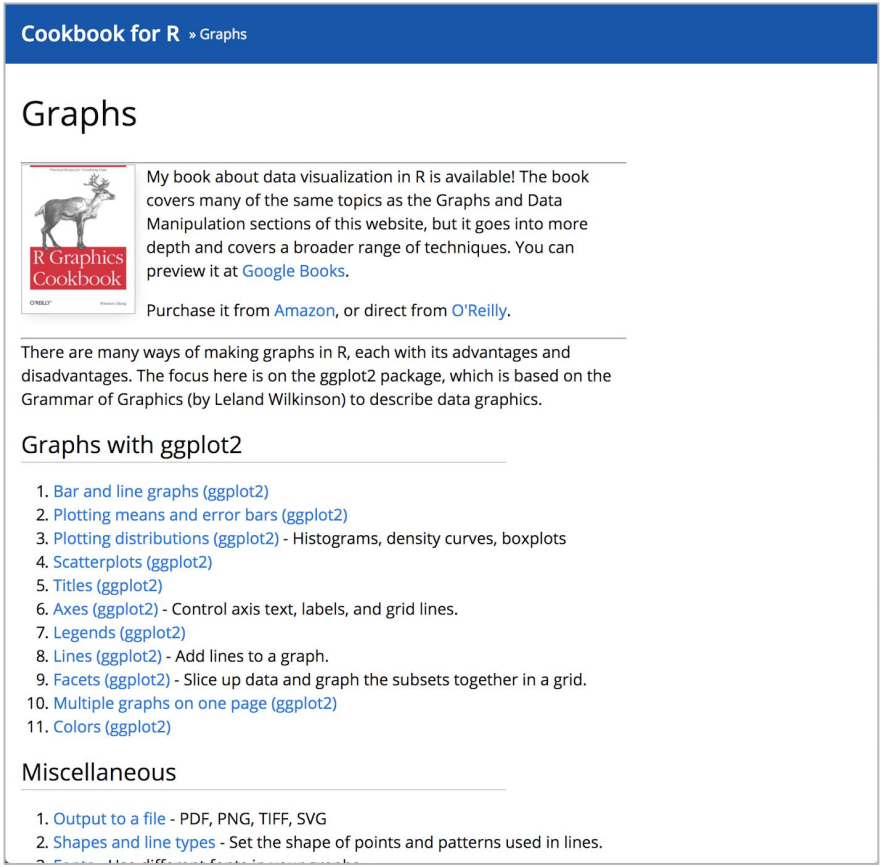
Problem solving

Confidence

Conceptual vs. applied



Walk throughs



Cookbook for R » Graphs

Graphs

My book about data visualization in R is available! The book covers many of the same topics as the Graphs and Data Manipulation sections of this website, but it goes into more depth and covers a broader range of techniques. You can preview it at Google Books.

Purchase it from Amazon, or direct from O'Reilly.

There are many ways of making graphs in R, each with its advantages and disadvantages. The focus here is on the ggplot2 package, which is based on the Grammar of Graphics (by Leland Wilkinson) to describe data graphics.

Graphs with ggplot2

1. Bar and line graphs (ggplot2)
2. Plotting means and error bars (ggplot2)
3. Plotting distributions (ggplot2) - Histograms, density curves, boxplots
4. Scatterplots (ggplot2)
5. Titles (ggplot2)
6. Axes (ggplot2) - Control axis text, labels, and grid lines.
7. Legends (ggplot2)
8. Lines (ggplot2) - Add lines to a graph.
9. Facets (ggplot2) - Slice up data and graph the subsets together in a grid.
10. Multiple graphs on one page (ggplot2)
11. Colors (ggplot2)

Miscellaneous

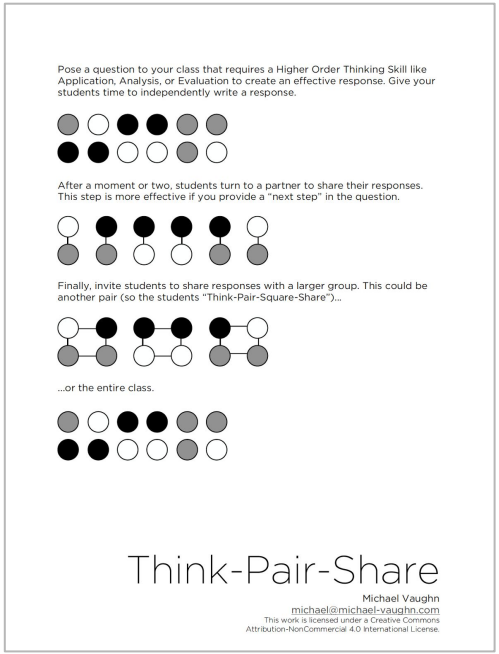
1. Output to a file - PDF, PNG, TIFF, SVG
2. Shapes and line types - Set the shape of points and patterns used in lines.

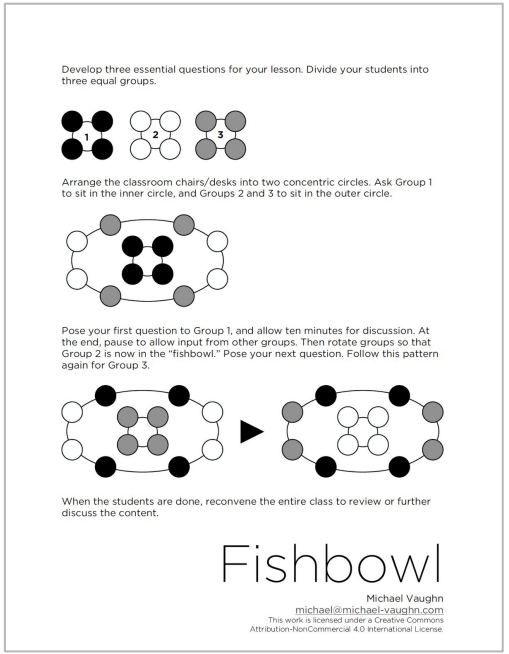
http://www.cookbook-r.com/Graphs

Open work time

http://miriamposner.com/blog/a-better-way-to-teach-technical-skills-to-a-group

Active learning





http://michael-vaughn.com/presentation-design-resources

Brief comments?

angela.zoss@duke.edu

Source: http://bit.ly/2pTZ3Pv

Data Instruction in the Age of Data Science:

Environment

DUKE UNIVERSITY
LIBRARIES

Spring 2017 Workshop Series

Registration and locations are online at
http://library.duke.edu/data/news

Workshop	Date	Time
Intro to R - Data Transformations, Analysis, and Data Structures	Jan 17	1:00pm – 3:00pm
OpenRefine - Data Cleaning, Faceting, and Transformations	Jan 23	1:30pm – 3:30pm
Graphic Design for Effective Diagrams	Jan 24	9:30am – 11:30am
Data Visualization with Excel	Jan 25	1:30pm – 3:30pm
Introduction to ArcGIS	Jan 26	1:00pm – 3:00pm
Adobe Illustrator for Diagrams and Visualizations	Jan 31	9:30am – 11:30am
Regular Expressions - Pattern Matching	Jan 31	10:00am – 12:00pm
QGIS	Feb 1	1:00pm – 3:00pm
Data Management Fundamentals	Feb 6	1:00pm – 2:30pm

The Evolution of (Library) Data Instruction

The Deep (Data) Past

- Largely class-based instruction
- Instruction based on data sources, data selection, data cleaning, data use
- Some coverage of popular statistical software and occasionally GIS software

Rise of Workshops

- Ability to reach a broad audience
- More efficient than consults
- Marketing benefits
- Opportunities to develop new services and content

Age of “Data Science”

- Increasingly diverse audience
- Broader range of data tools
- Wider domain of potential workshop topics
- Yet, data professionals have more demands on their time...

Considerations

Internal Considerations

- Fixed number of data professionals
- Data professionals' areas of expertise
- Limited ability to master multiple tools
- What is/are data anyway? How expansive should data workshops be?

Increasing Audience Diversity and Size

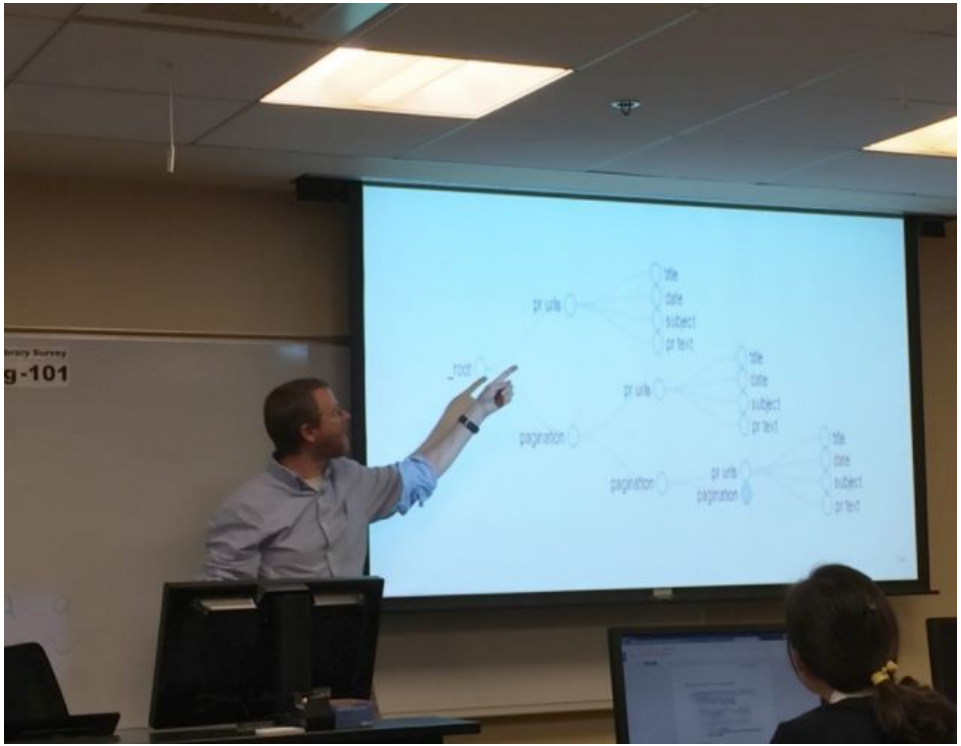
- Which topics to cover?
- How much ‘scaffolding’?
- Needs of advanced users?
- How many audiences can one please in a session?

Expanded Tools

- We have more tools…
- Audience for the tool
- Matching tools and workshops?

Palladio. Visualize complex historical data with ease.

Where does data instruction fit in the environment?



- Complementary instruction?
- Supplementary instruction?
- Replacing curricular content?
- How does the instruction relate to professional development? For students? For data staff?

Strategies

Source: http://bit.ly/2q1jarf

Utilize (or partner with) external resources

Try new forms of public programming

R we having fun yet? A learning series on [R].

Sponsored by Duke University Libraries, Data & Visualization Services

This Spring the Data & Visualization Services Department will host a series of campus/community-oriented, informal sessions on the R programming language. Our goals is to promote a friendly environment for exploring the extensible capabilities of the R software environment specifically supported through R and RStudio. Beginner's are welcome, experts will be encouraged to share topical expertise. How has R enabled your work? What else can R help you accomplish? Join us most Thursdays at noon in the Edge (Workshop Room -- map). Bring your lunch (or not); light refreshments will be provided.

When: Thursday's at noon

Who: R enthusiasts, beginngers through advanced
Sponsored by the Data & Visualization Services Department

Brief comments?

joel.herndon@duke.edu

Discussion