Big Ideas to Help Statistics Students Learn to Think with Data

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

Statistical Society of Canada, June 6, 2018

nhorton@amherst.edu
http://nhorton.people.amherst.edu

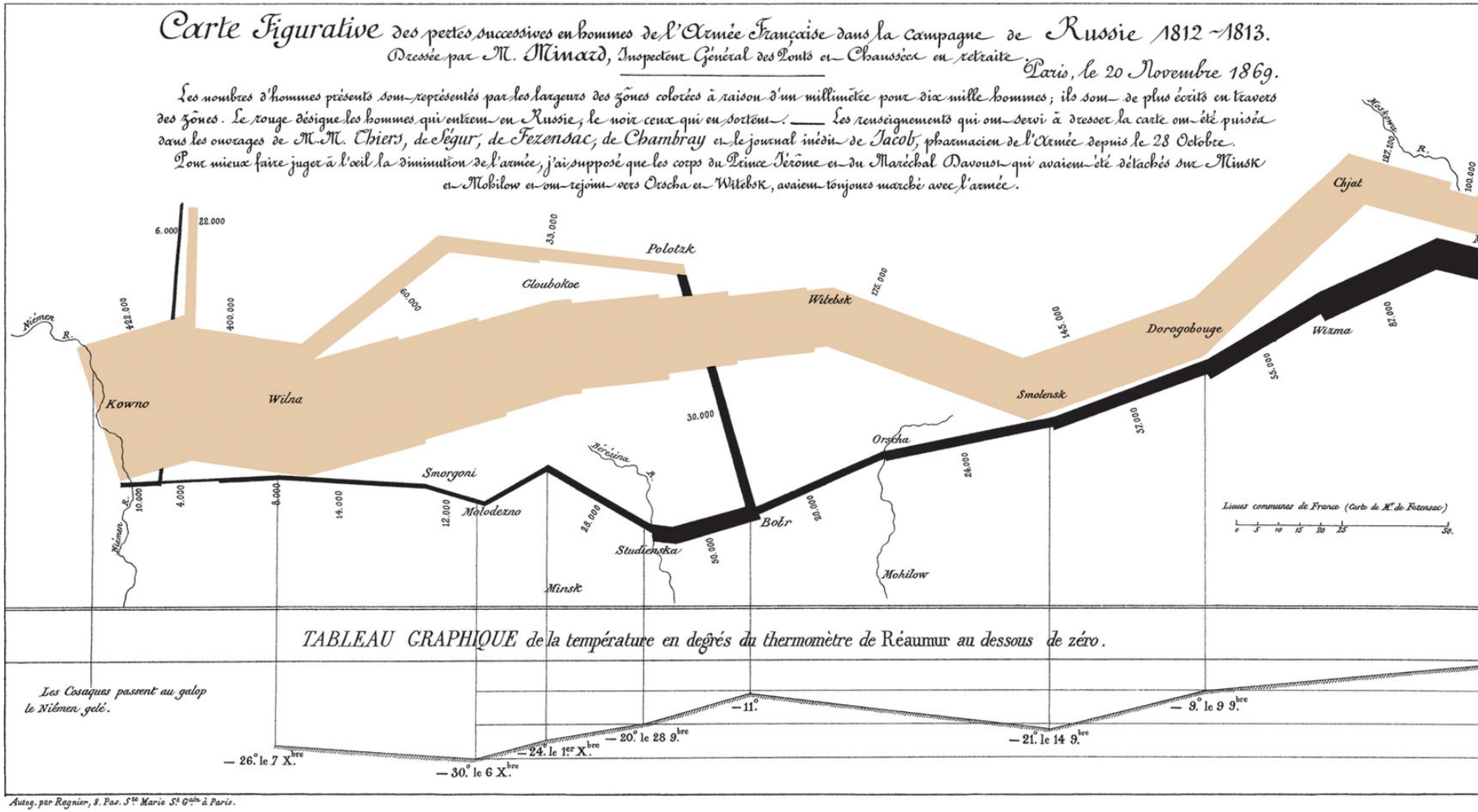
Acknowledgements

- Project MOSAIC: Danny Kaplan (Macalester College), Randy Pruim (Calvin College), and Ben Baumer (Smith College)
- Johanna Hardin (Pomona College) and the undergraduate guidelines working group
- The ASA revised GAISE College report group
- those listed in the annotated bibliography in Table 2 of Horton and Hardin (TAS 2015)

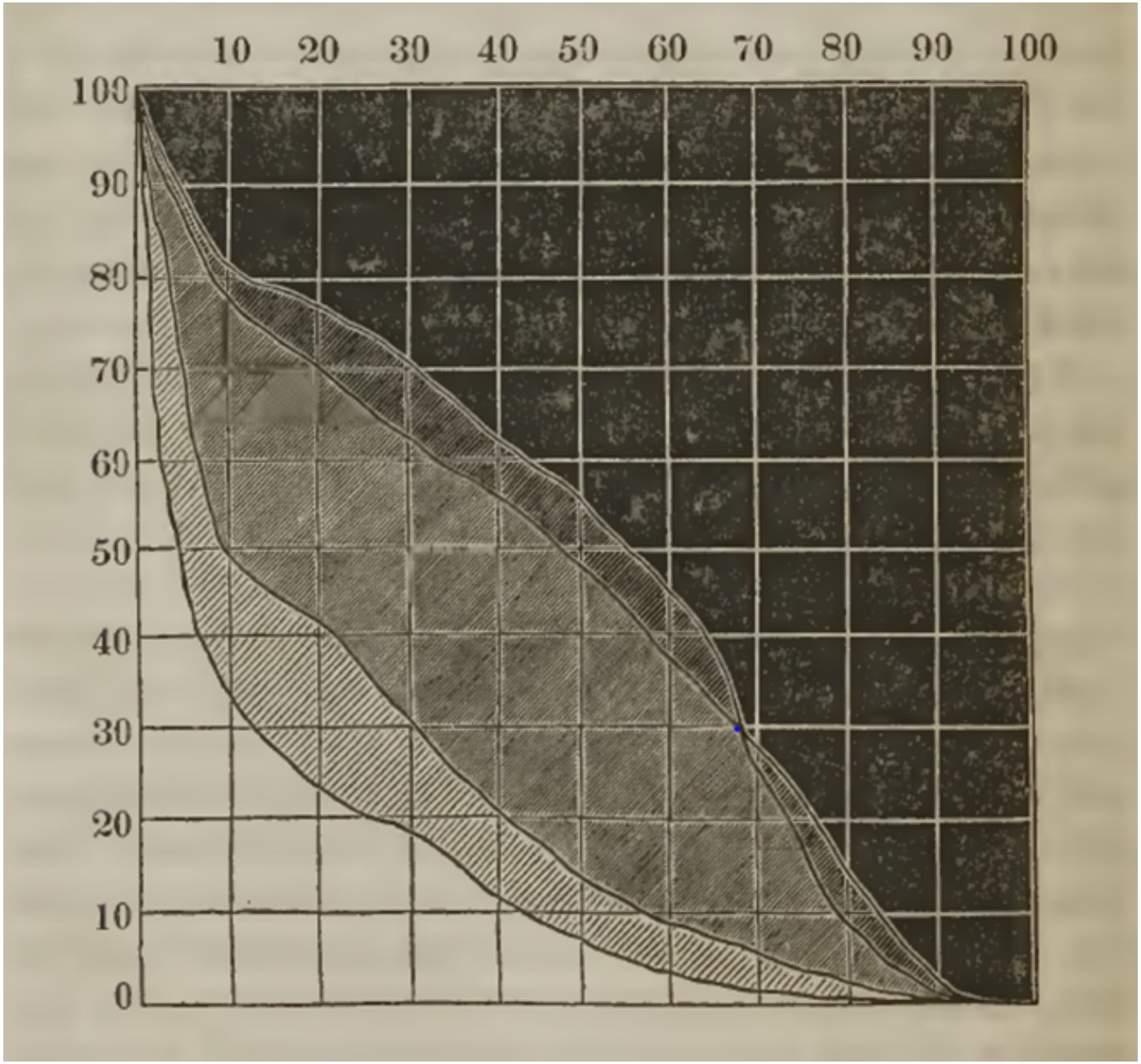
Plan

- a glimpse into the past
- a vision for statistics and data science education
- some big ideas
- closing thoughts

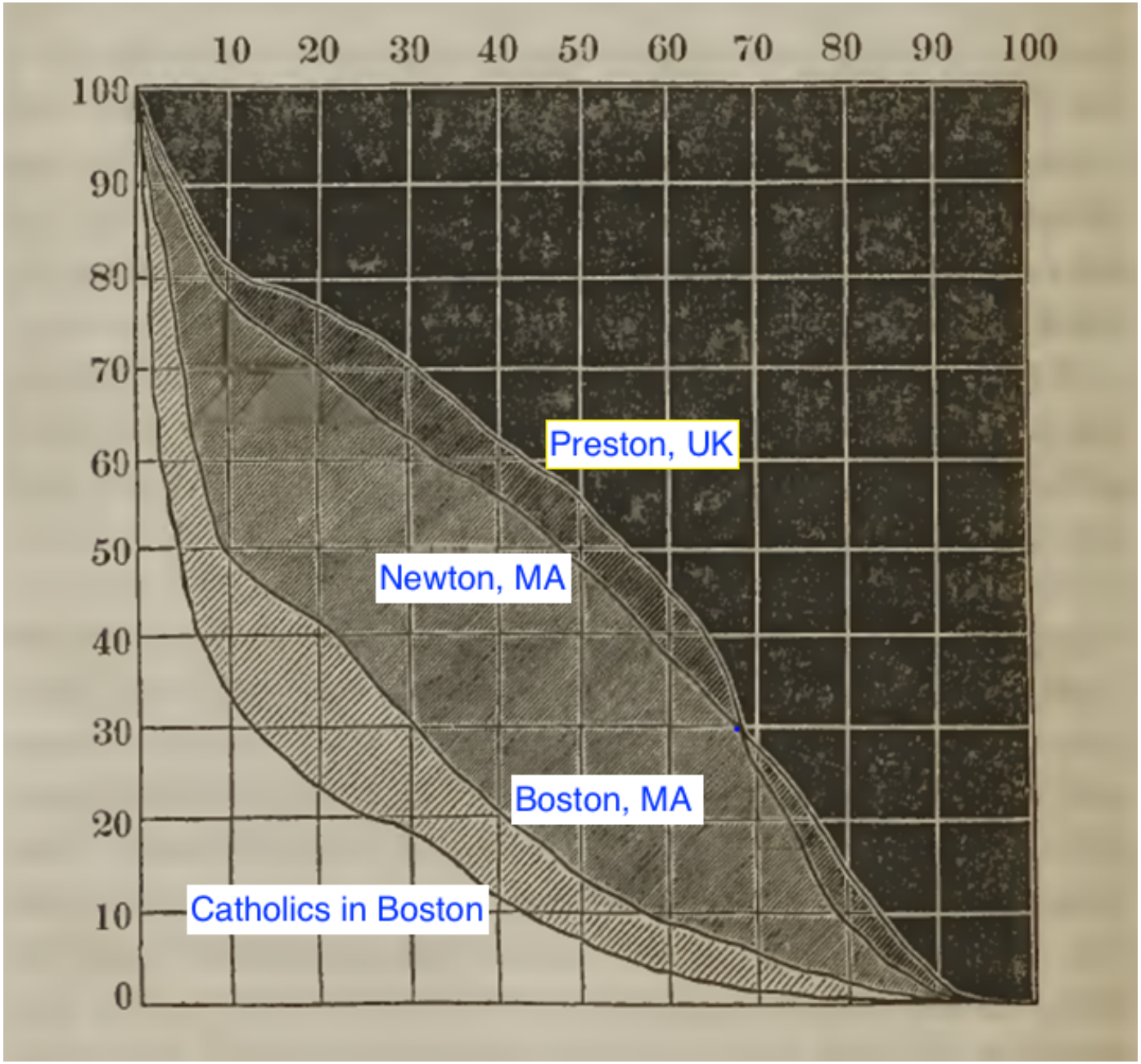
Minard and Napoleon's campaign (1812)



Shattuck and mortality report (1850)



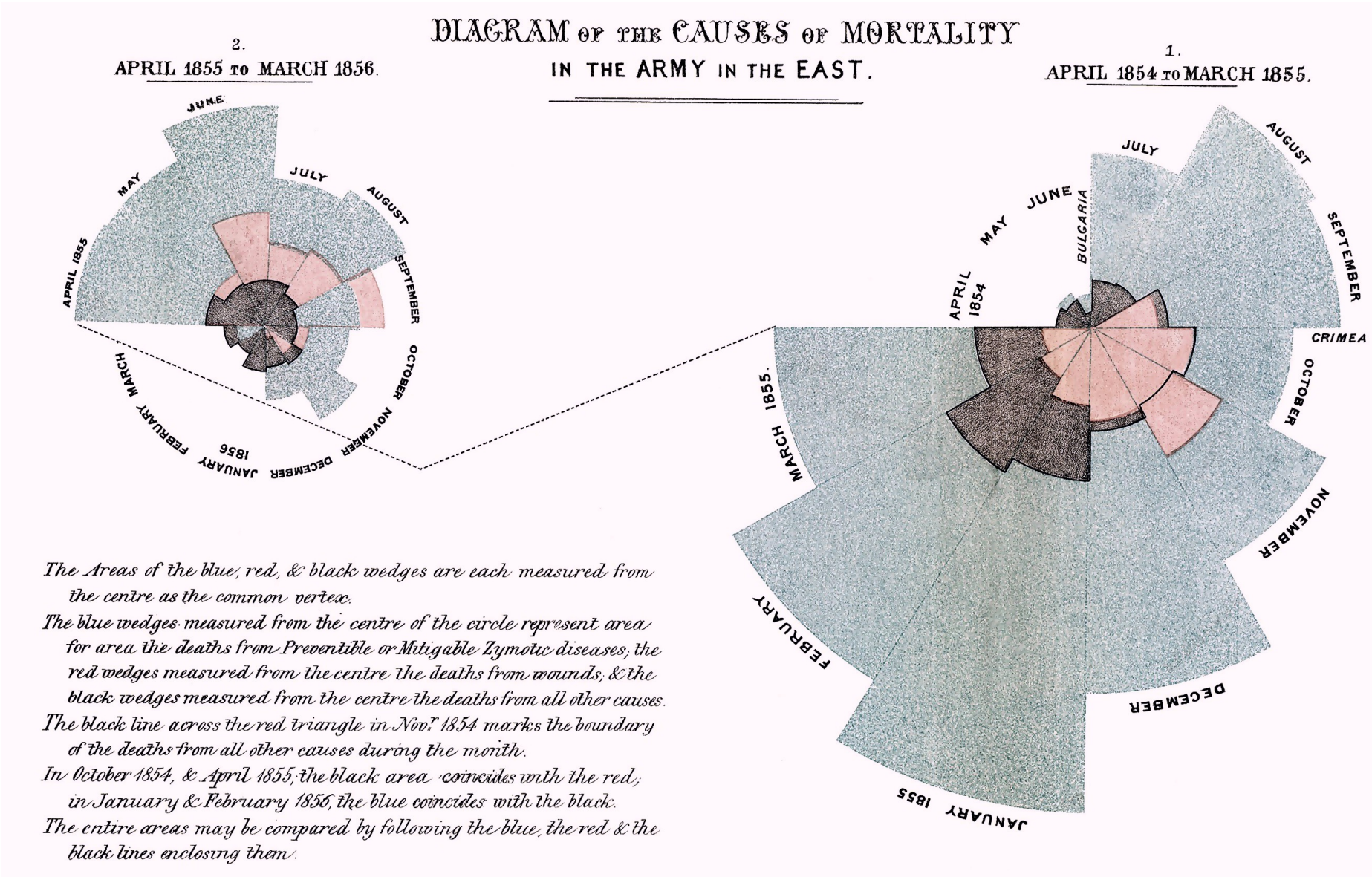
Shattuck and mortality report (1850)



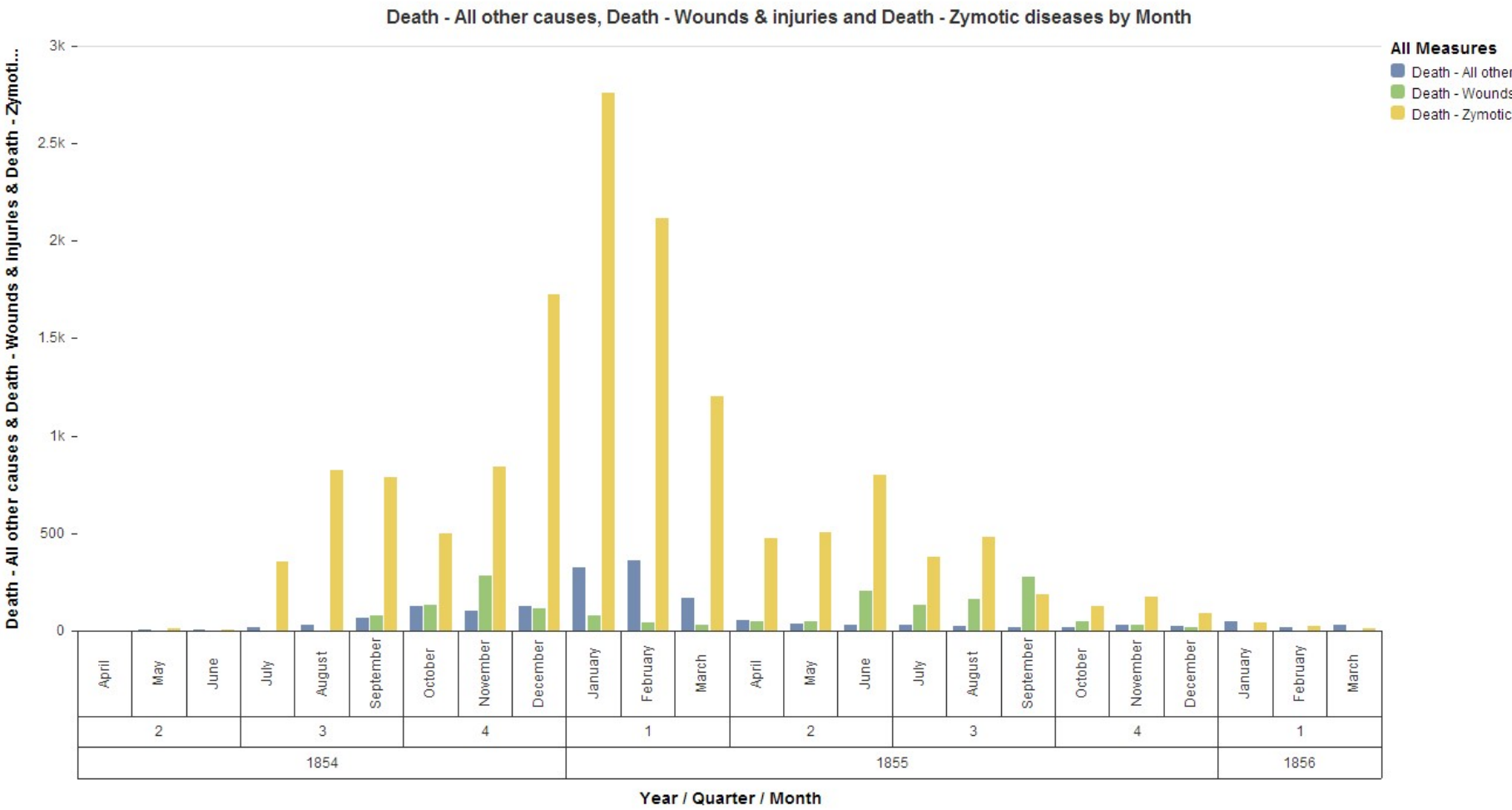
Florence Nightingale (see Utts, *Amstat News* June 2016)

Fast forward a few years across the pond...

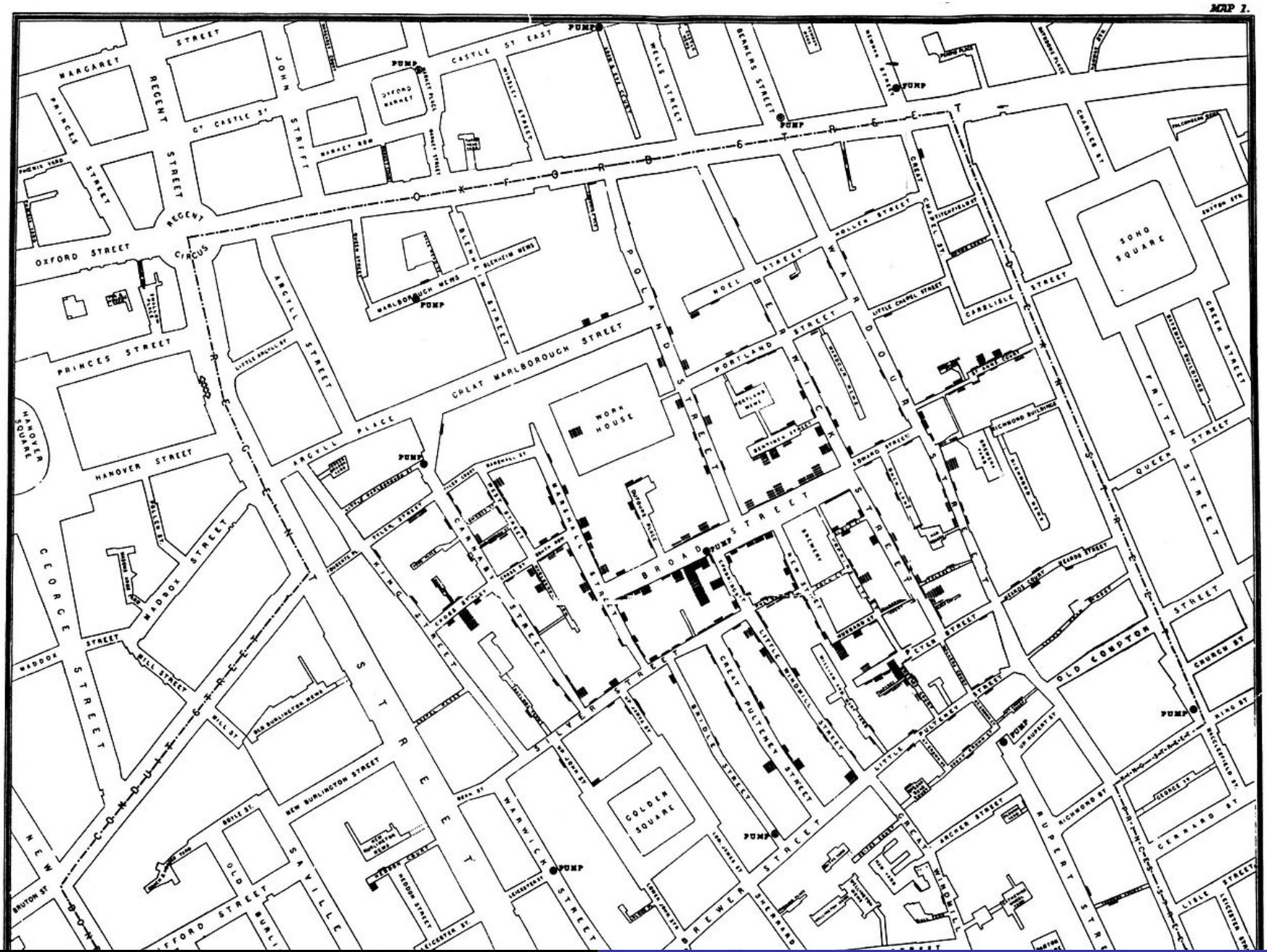
Florence Nightingale (see Utts, *Amstat News* June 2016)



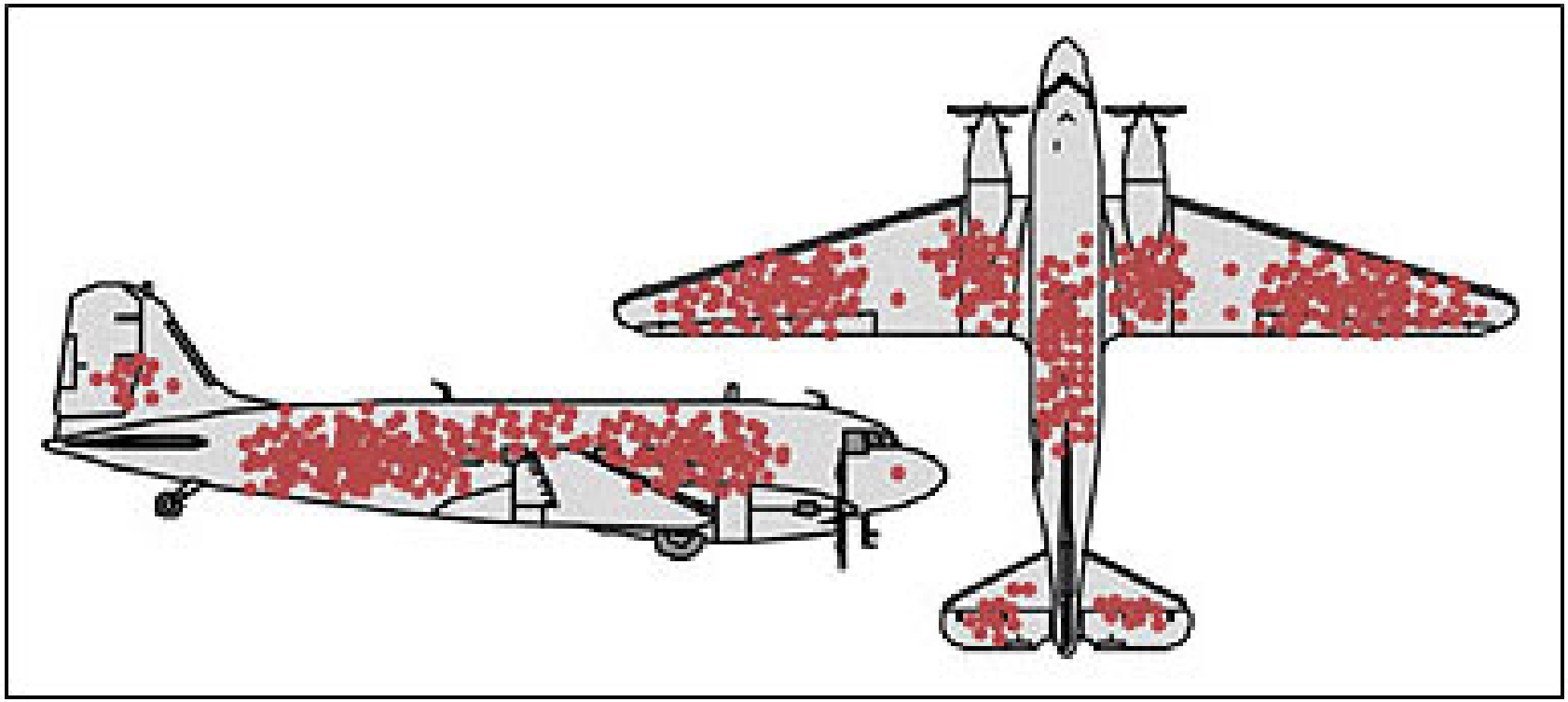
reworked graph: (Source: itelligencegroup.com)



And in London: Snow and the cholera epidemic



Wald and World War II airplanes (credit to Cameron Moll)



Credit: Cameron Moll

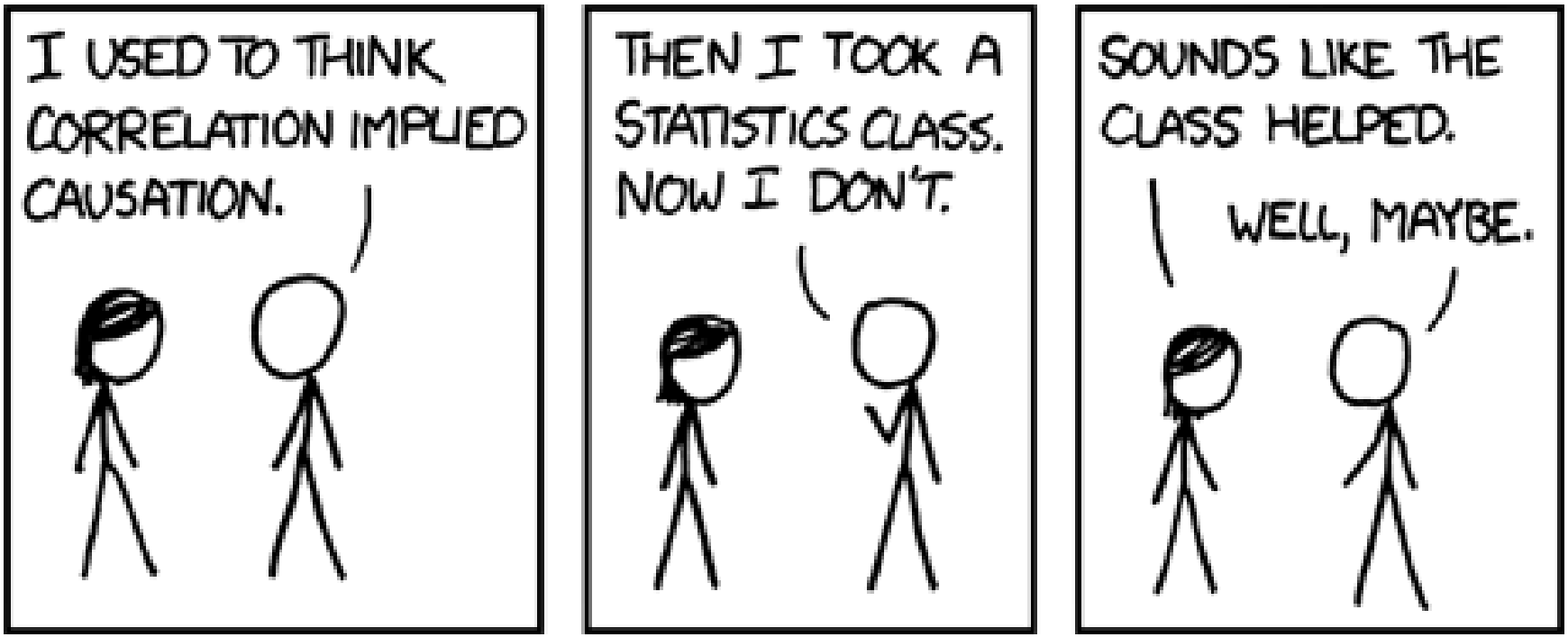
Cobb (1982)

Let students do what statisticians do: analyze non-trivial datasets by considering a variety of models, **using their imagination and developing their judgment** *in the process.*

Cobb (1982)

I believe it is the use of imagination and judgment that makes our subject appealing. **We owe it to our students not to keep that a secret**.

Is statistics a dirty word? (Source: xkcd.com)



What are we teaching our students?

What is data science?

Obama video keynote for Hadoop/Strata conference:
`https://www.youtube.com/watch?v=vbb-AjiXyh0`

What is data science?

Obama video keynote for Hadoop/Strata conference:
`https://www.youtube.com/watch?v=vbb-AjiXyh0`

Implications?

What are we teaching our students?

- much of my talk will focus on undergraduate programs
- there are important implications for introductory courses
- even more for our courses that follow intro
- take home message: we need to make **many** changes across the board to be relevant in a new era of data science

a vision for statistics education (John Bailar)

As academic statisticians, we are missing the boat. We are barking up the wrong tree. ... The kinds of statistics that we teach in undergraduate and especially in graduate programs have almost **nothing to contribute to anything that matters**. *... Then we wonder why the world passes us by.*

a vision for statistics education

The Committee on Applied and Theoretical Statistics (CATS) noted widespread sentiment in the statistical community that upper-level undergraduate and graduate curricula for statistics majors ... are currently structured in ways that **do not provide sufficient exposure to modern statistical analysis, computational and graphical tools, communication skills**, *and the ever growing interdisciplinary uses of statistics.*

more from CATS workshop

The growth that statistics has undergone is often not reflected in the education that future statisticians receive. There is a need to incorporate more meaningfully into the curriculum the **computational and graphical tools** *that are today so important to many professional statisticians. There is a need for improved training of statistics students in* **written and oral communication skills**, *which are crucial for effective interaction with scientists and policy makers.*

more from CATS workshop

The current curriculum in most statistics departments is, however, entirely too focused on hypothesis testing (Ed Rothman).

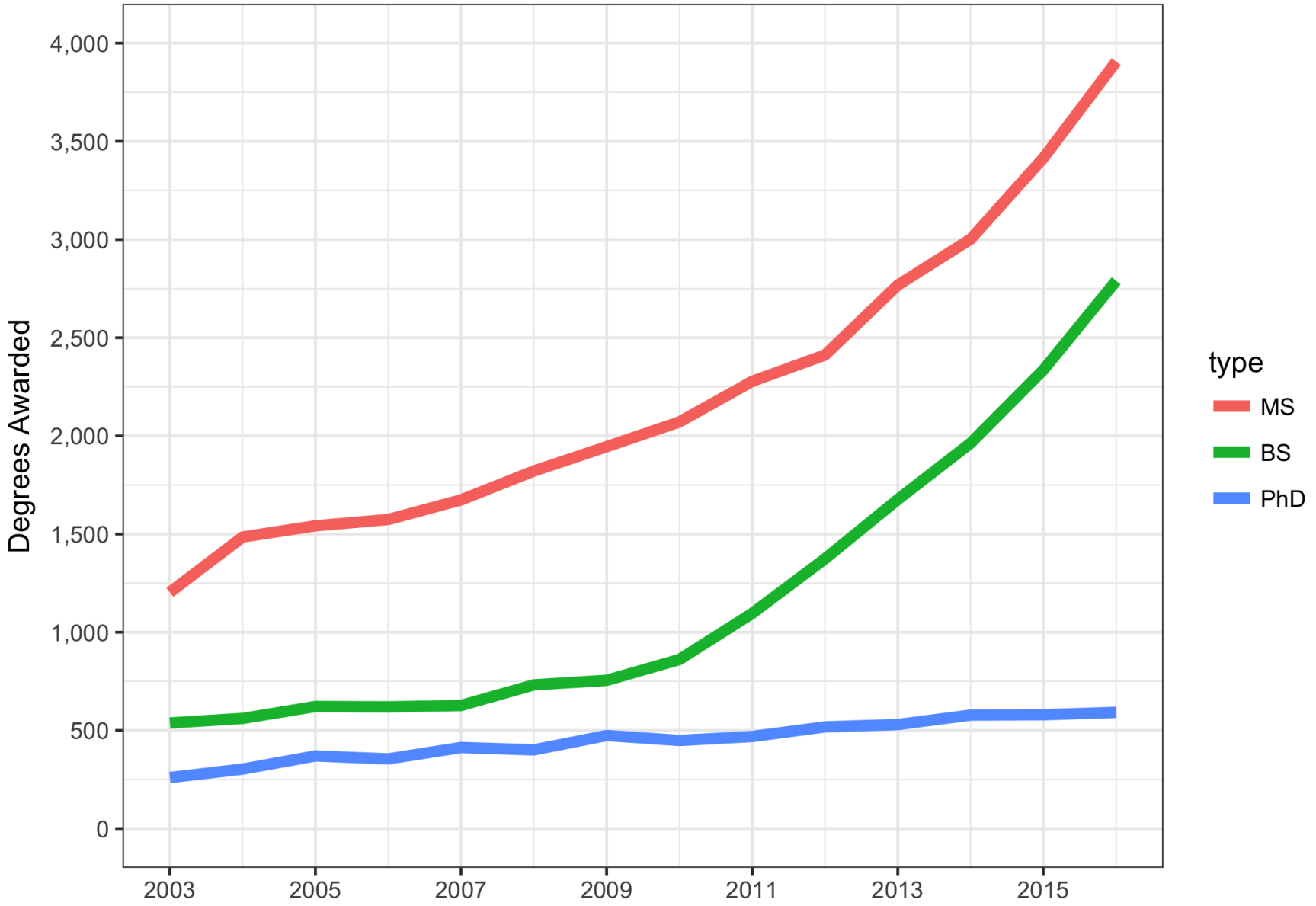
We risk being ignored if we do not stay relevant. (Carl Morris)

more from CATS workshop

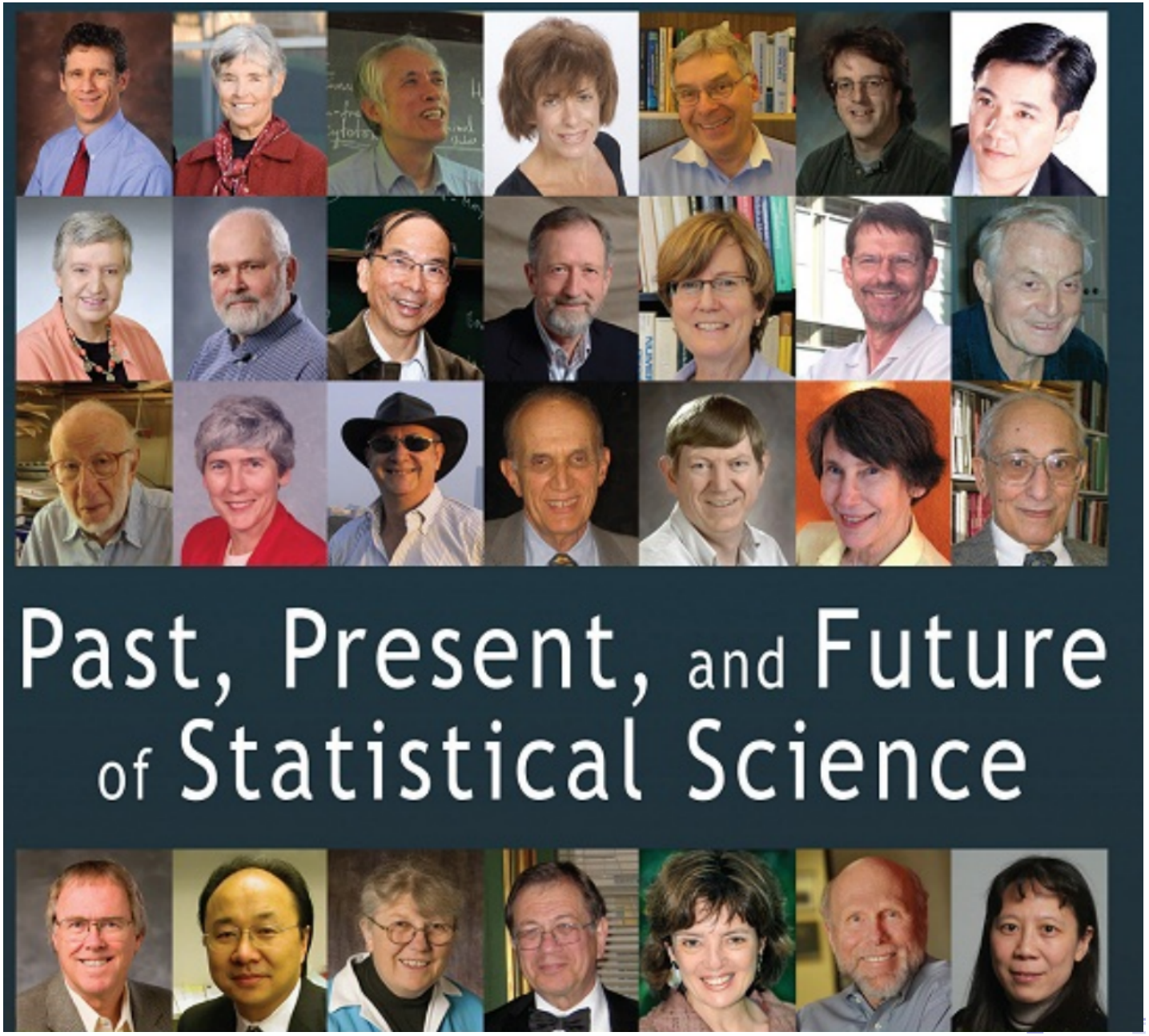
> *The current curriculum in most statistics departments is, however, entirely too focused on hypothesis testing (Ed Rothman).*
>
> *We risk being ignored if we do not stay relevant. (Carl Morris)*

All are quotes from 1992.

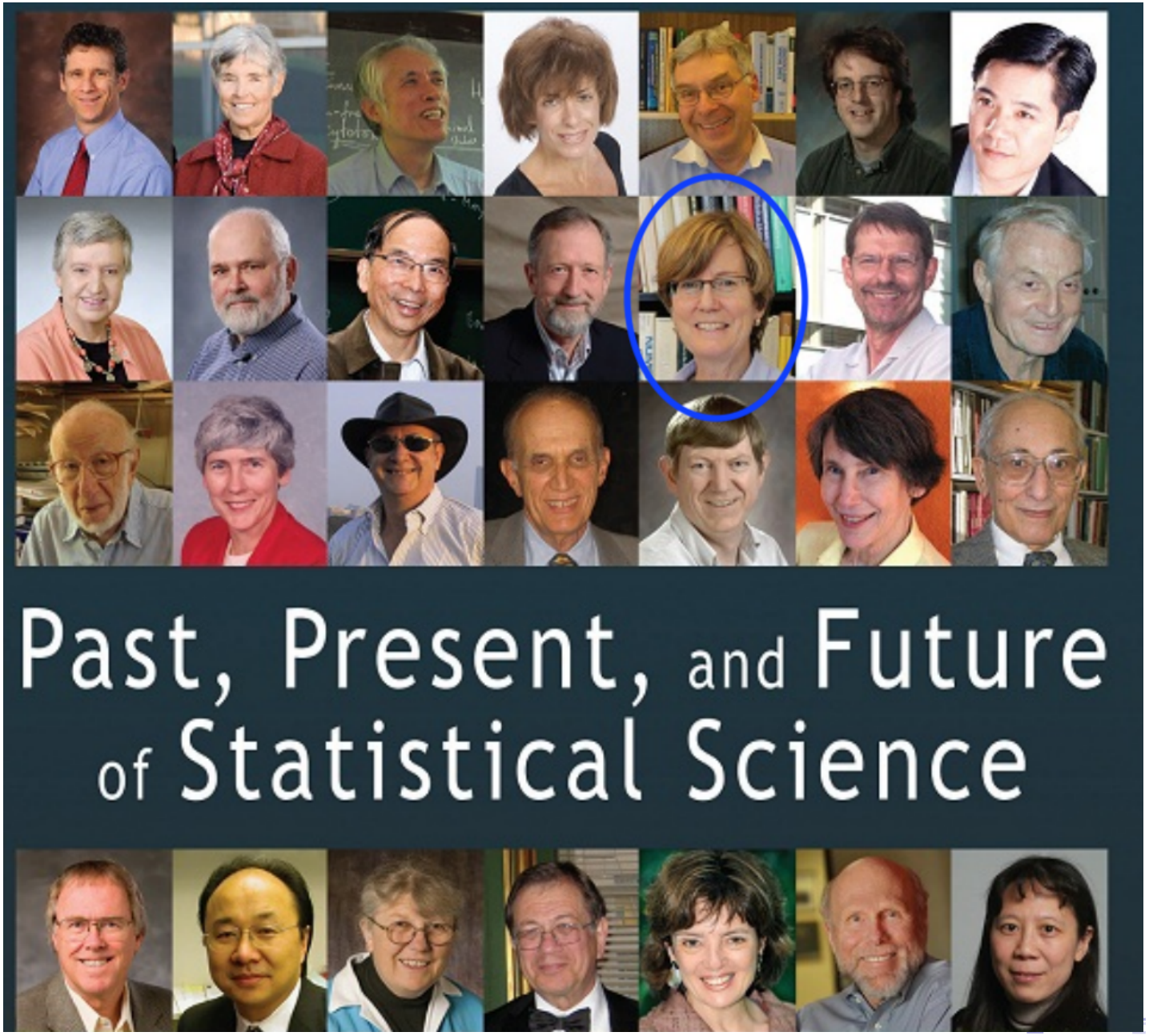
Progress: US degrees in statistics over time (Source: IPEDS)



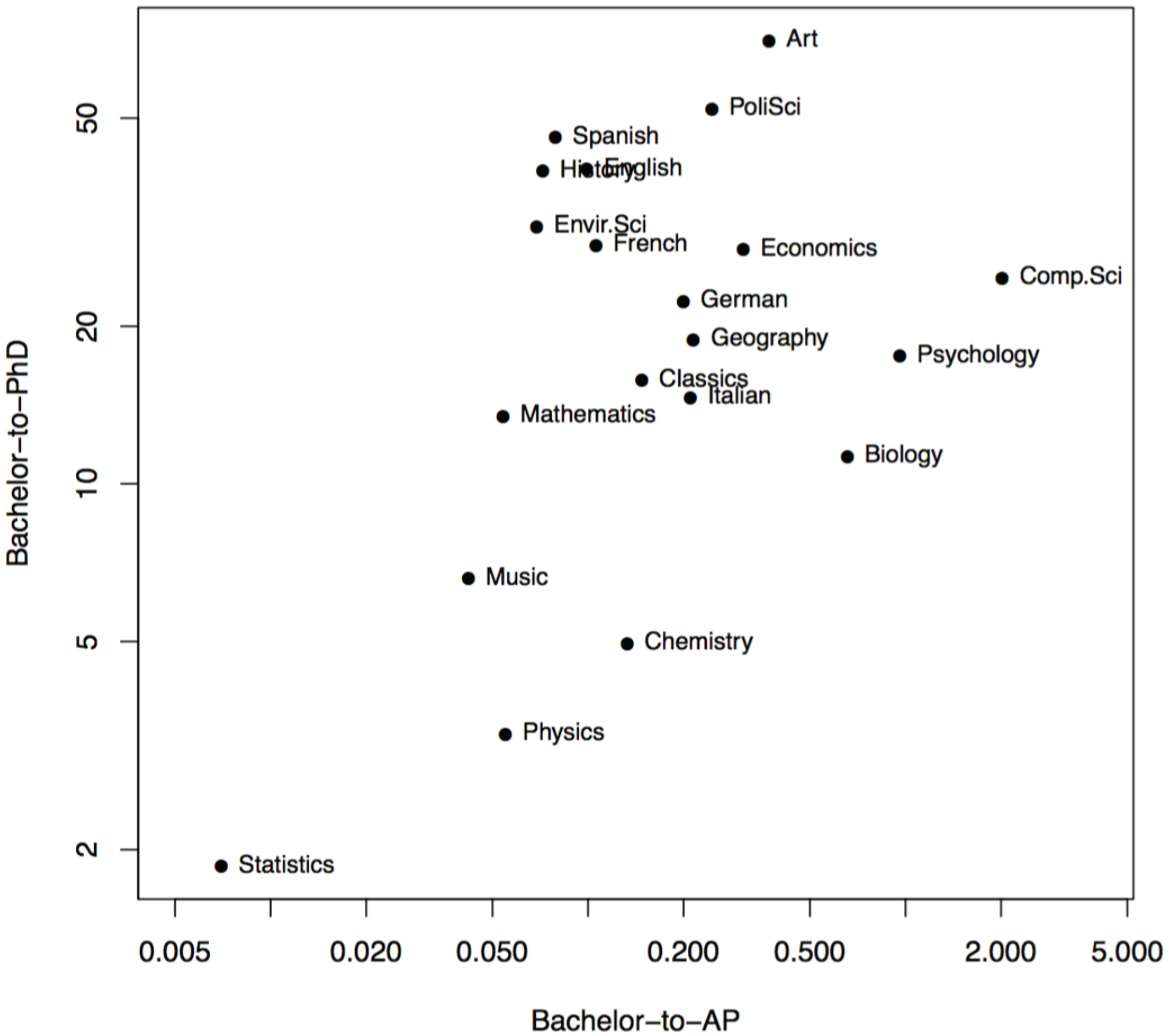
COPSS book: *Past, Present, and Future*



Class use of *Past, Present, and Future*



But where are the majors? (Johnstone, COPSS PPF)



ASA Undergrad Guidelines, 2014

- This is an exciting time to be a statistician.
- The contribution of the discipline of statistics to scientific knowledge is widely recognized with increasingly positive public perception.
- How do we prepare (undergraduate?) students?

`https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statis.aspx`

ASA Undergrad Guidelines, 2014

We are concerned that many of our graduates do not have sufficient skills to be effective in the modern workforce. Thomas Lumley (personal communication) has stated that our students know how to deal with $n \to \infty$, but cannot deal with a million observations.

If statistics is the science of learning from data, then our students need to be able to "think with data" (as Diane Lambert of Google has so elegantly described).
- Horton and Hardin (TAS, 2015)

NAS Undergraduate Data Science report, 2018

> *Data science is emerging as a field that is revolutionizing science and industries alike. Work across nearly all domains is becoming more data driven, affecting both the jobs that are available and the skills that are required. As more data and ways of analyzing them become available, more aspects of the economy, society, and daily life will become dependent on data.*

`https://nas.edu/envisioningds`

Finding 2.3: A critical task in the education of future data scientists is to
requires exposure to key concepts in data science, real-world data and pr
the limitations of tools, and ethical considerations that permeate many ap
involved in developing data acumen include the following:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,
- Data description and visualization,
- Data modeling and assessment,
- Workflow and reproducibility,
- Communication and teamwork,
- Domain-specific considerations, and
- Ethical problem solving.

Some big ideas to ensure that we stay relevant

1. Increasing importance of data analysis, computation, and data science

Big idea #1: Data science and computation

ASA undergrad guidelines for statistics programs:

- Working with data requires **extensive computing skills**.
- To be prepared for statistics and data science careers, students need the ability to **access and wrangle** data in various ways, and the ability to perform **algorithmic problem-solving**.
- In addition to more traditional mathematical and statistical skills, students should be fluent in higher-level programming languages and **facile with database systems**.
- Good news: tools are now simpler, cheaper, and more powerful!

24 years of R: one of many solutions

The New York Times

Business Computing

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Data Analysts Captivated by R's Power

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

FACEBOOK
TWITTER

Motivating example

An analyst wants to calculate the mean `pH` of assays from two treatments. What's the simplest way to do this in base R? Using other packages?

Possible answer

```
> with(chem, aggregate(pH, by=list(treat),
  FUN=mean, na.rm=TRUE, simplify=TRUE))
  Group.1        x
1    grpA 1.904762
2    grpB 1.756757
```

Possible answer

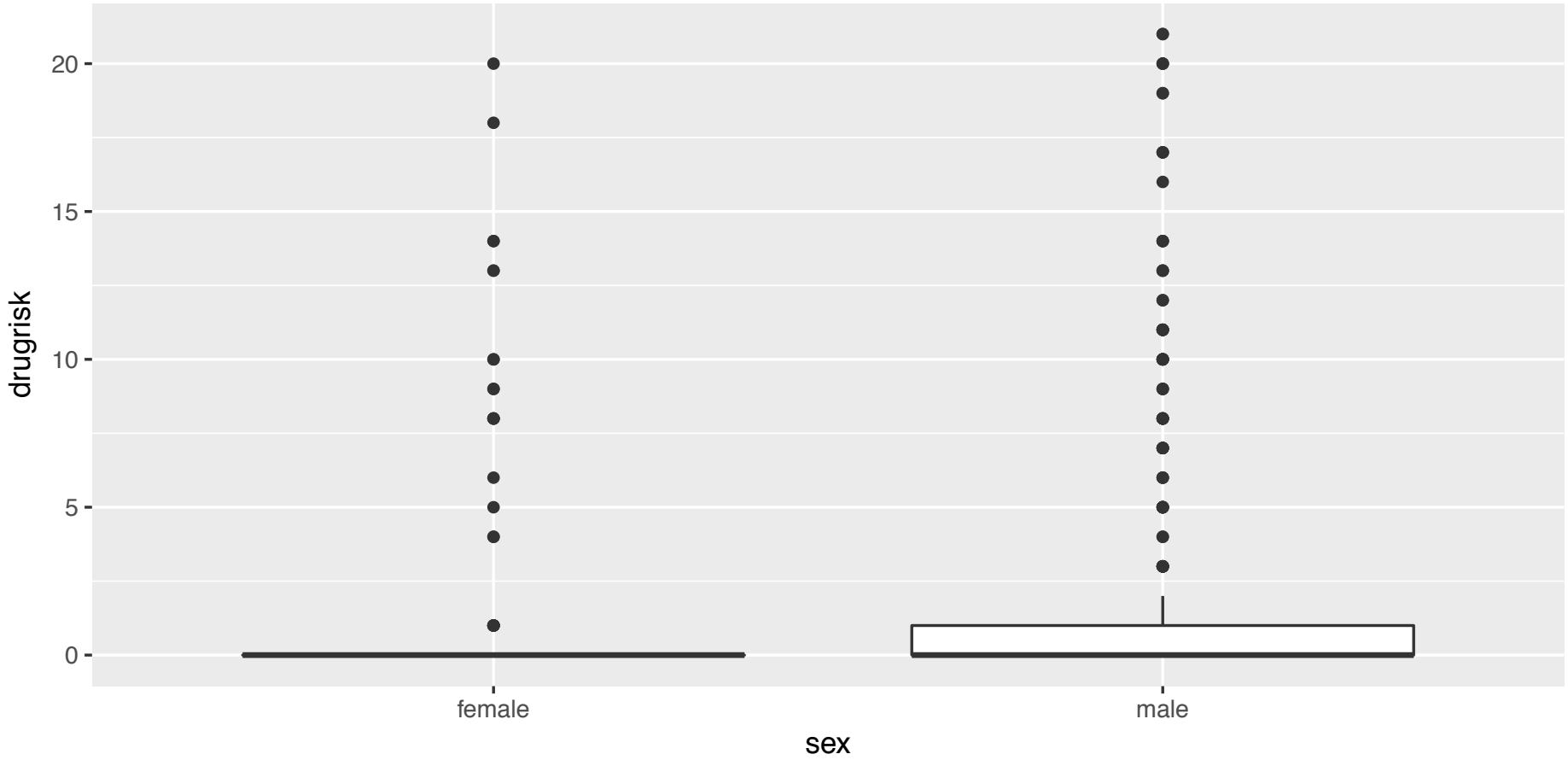
```
> with(chem, tapply(pH, treat, mean, na.rm=TRUE))
    grpA     grpB
1.904762 1.756757
```

A better answer (suitable for intro)

```
> library(mosaic)
> favstats(pH ~ treat, data = chem)
     sex min Q1 median Q3 max mean   sd   n missing
1   grpA   0  0      0  1  11 1.90 4.37 357       2
2   grpB   0  0      0  0  10 1.76 4.15 111       0
```

mosaic modeling language (Y ~ X)

```
> gf_boxplot(pH ~ treat, data = chem)
```



mosaic modeling language (Y ~ X)

```
> lm(pH ~ treat, data = chem)

Coefficients:
(Intercept)          grpB
      1.905        -0.148
```

One simple approach to:

- generate descriptive statistics
- create graphical displays
- fit regression models

See *R Journal* paper (journal.r-project.org/archive/2017/RJ-2017-024) and Little Books (www.github.com/ProjectMOSAIC/LittleBooks)

Enough R for Intro Stats

Numerical Summaries

These functions have a formula interface to match plotting.

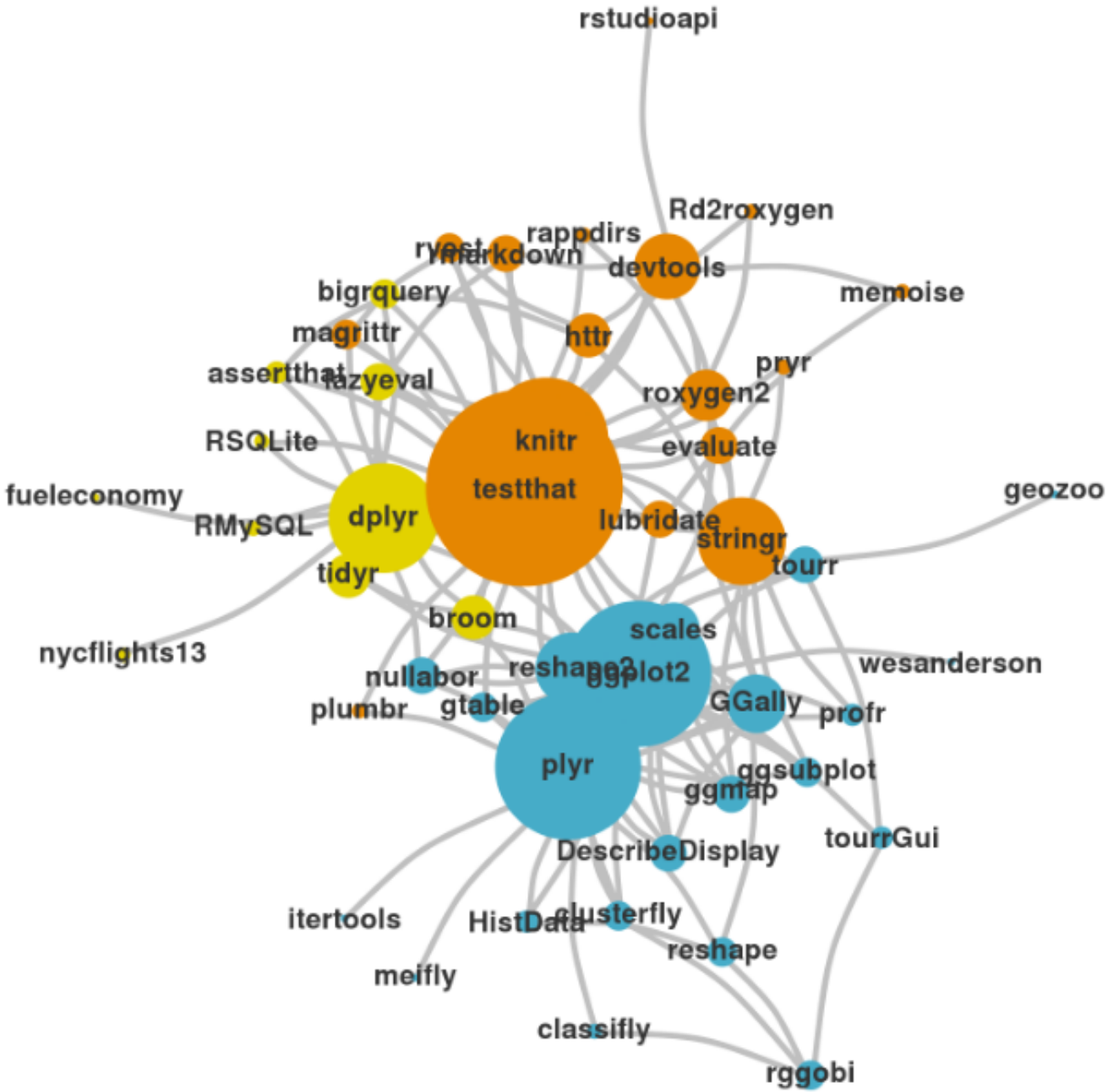
```
favstats()    # mosaic
tally()       # mosaic
mean()        # mosaic augmented
median()      # mosaic augmented
sd()          # mosaic augmented
var()         # mosaic augmented
diffmean()    # mosaic
```

Randomization/Simulation

```
rflip()       # mosaic
do()          # mosaic
sample()      # mosaic augmented
resample()    # with replacement
shuffle()     # mosaic
rbinom()
rnorm()       # etc, if needed
```

Distributions

More good news: the advent of the 'tidyverse'



Design goals of tidyverse

- tools that work well together, each one designed for a particular task
- if you don't succeed at first, try, try again (CS prototyping)
  1. `stats::reshape()`
  2. `reshape` package
  3. `reshape2` package
  4. `tidyr` package

Impact of mosaic and the tidyverse

- Small number of simple idioms
- Combine to do powerful operations
- Round off rough edges of R

More big ideas to ensure that we stay relevant

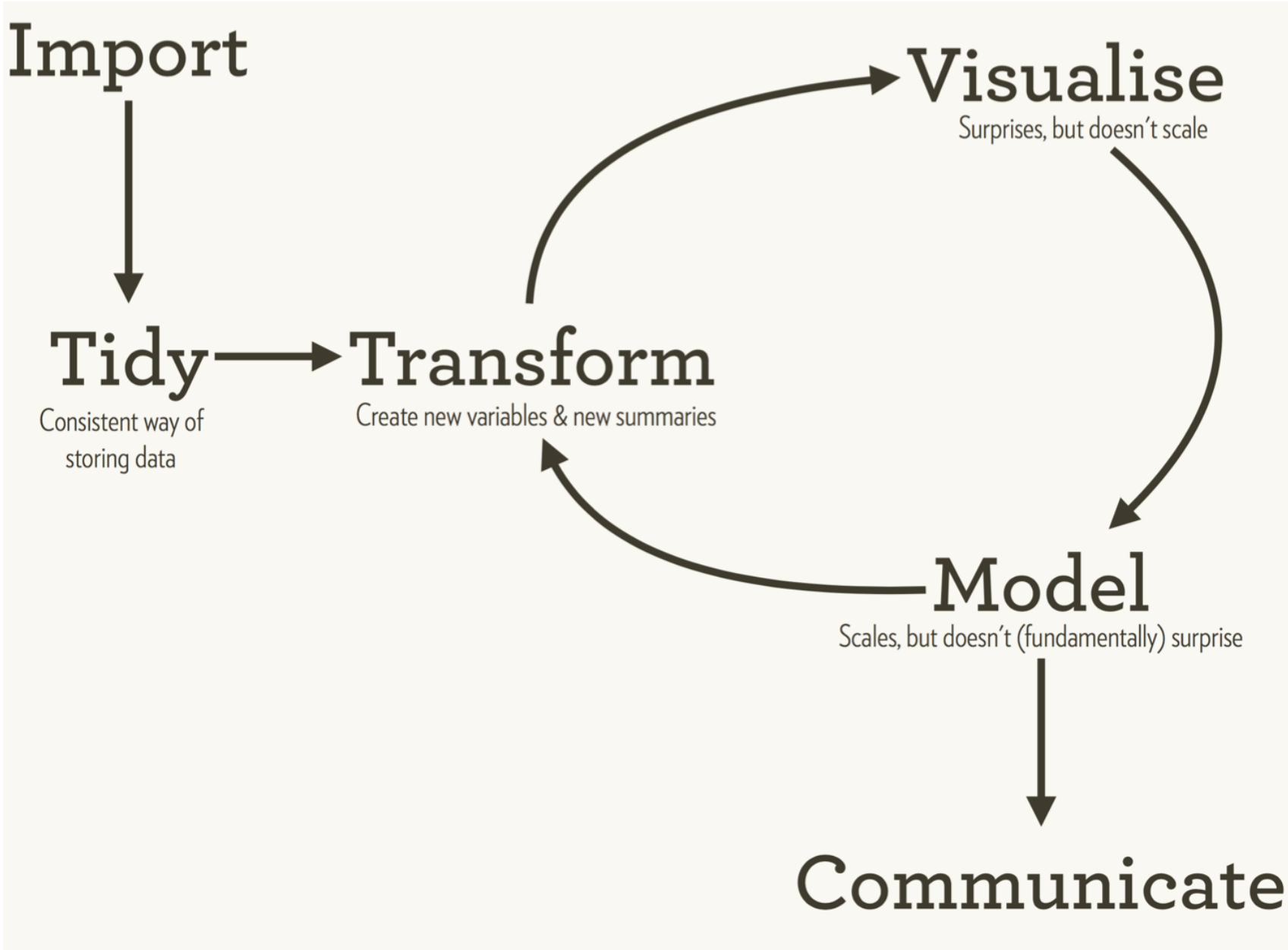
1. Increasing importance of data science and computation
2. Real applications and data

Big idea #2: Real applications and data

ASA undergrad guidelines for statistics programs:

- **Data should be a major component** of statistics courses.
- Programs should emphasize concepts and approaches for working with complex data and **provide experiences in designing studies and analyzing non-textbook data**.

Statistics and data analysis cycle (due to Wickham)



Key idioms for dealing with big(ger) data

select:	subset variables
filter:	subset rows
mutate:	add new columns
summarize:	reduce to a single row
group-by:	aggregate
join:	merge tables
gather/spread:	transpose (e.g., wide to tall)

Hadley Wickham, `bit.ly/bigrdata4` and "Building precursors to data science" (CHANCE, 2015, `https://nhorton.people.amherst.edu/precursors`)

More 'Variety' of data

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

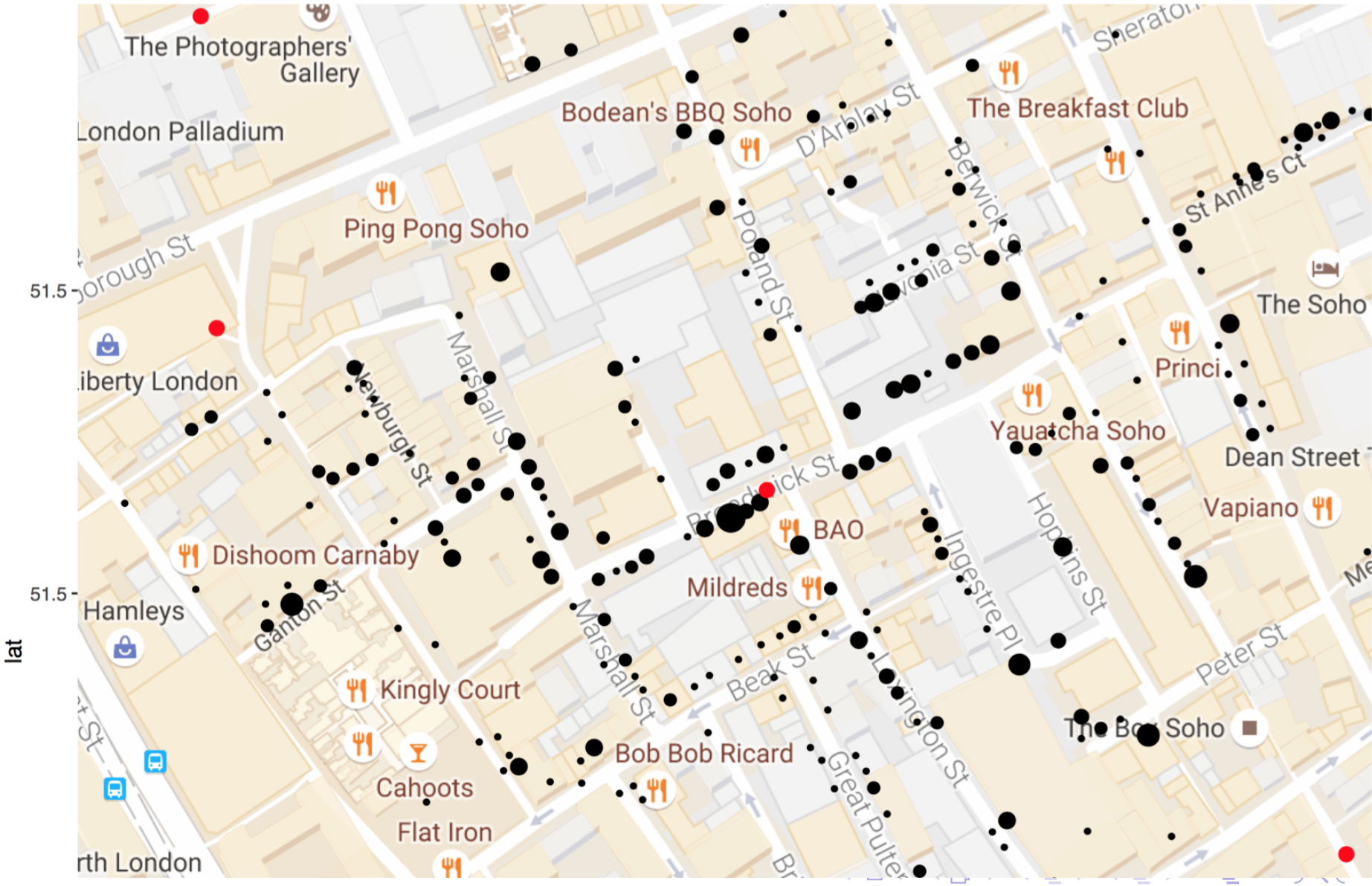
FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

Bigger (medium?) data (more 'Volume')

- use SQL (structured query language) to access databases within dplyr
- Climate change data
- Genomics data (e.g., https://www.ensembl.org)
- Smart cities (e.g., NYC Taxis, 1.1 billion rides)
- (US) Medicare health insurance data
- (US) airline delays

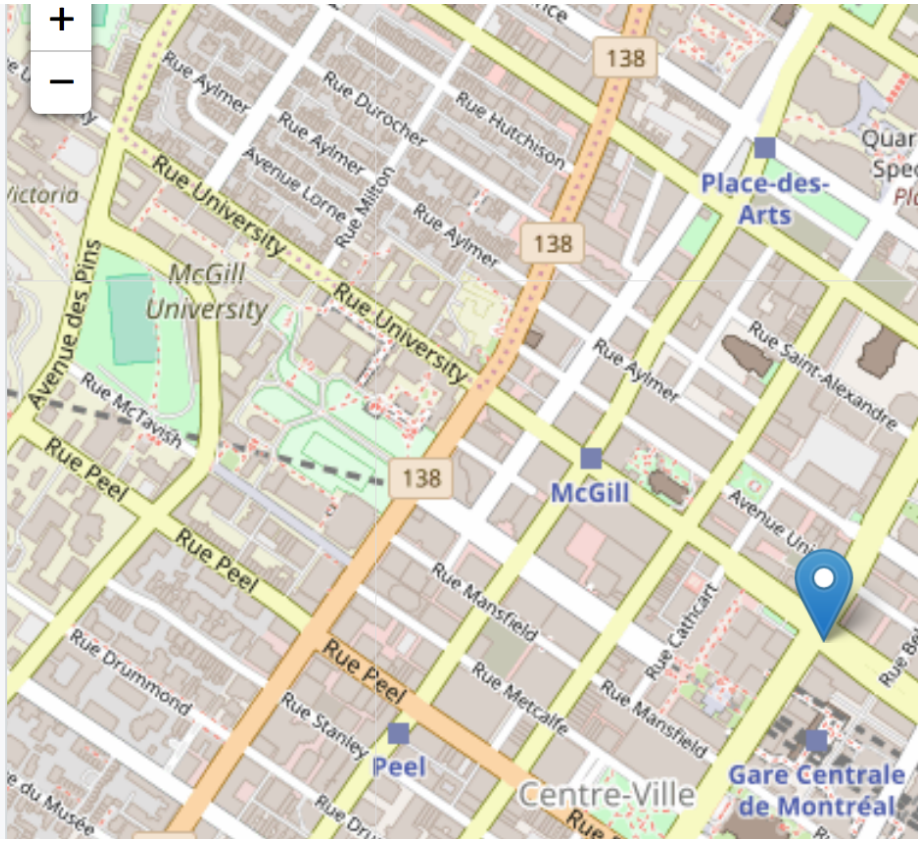
‘Variety’: maps as data

Snow and the cholera epidemic in London...

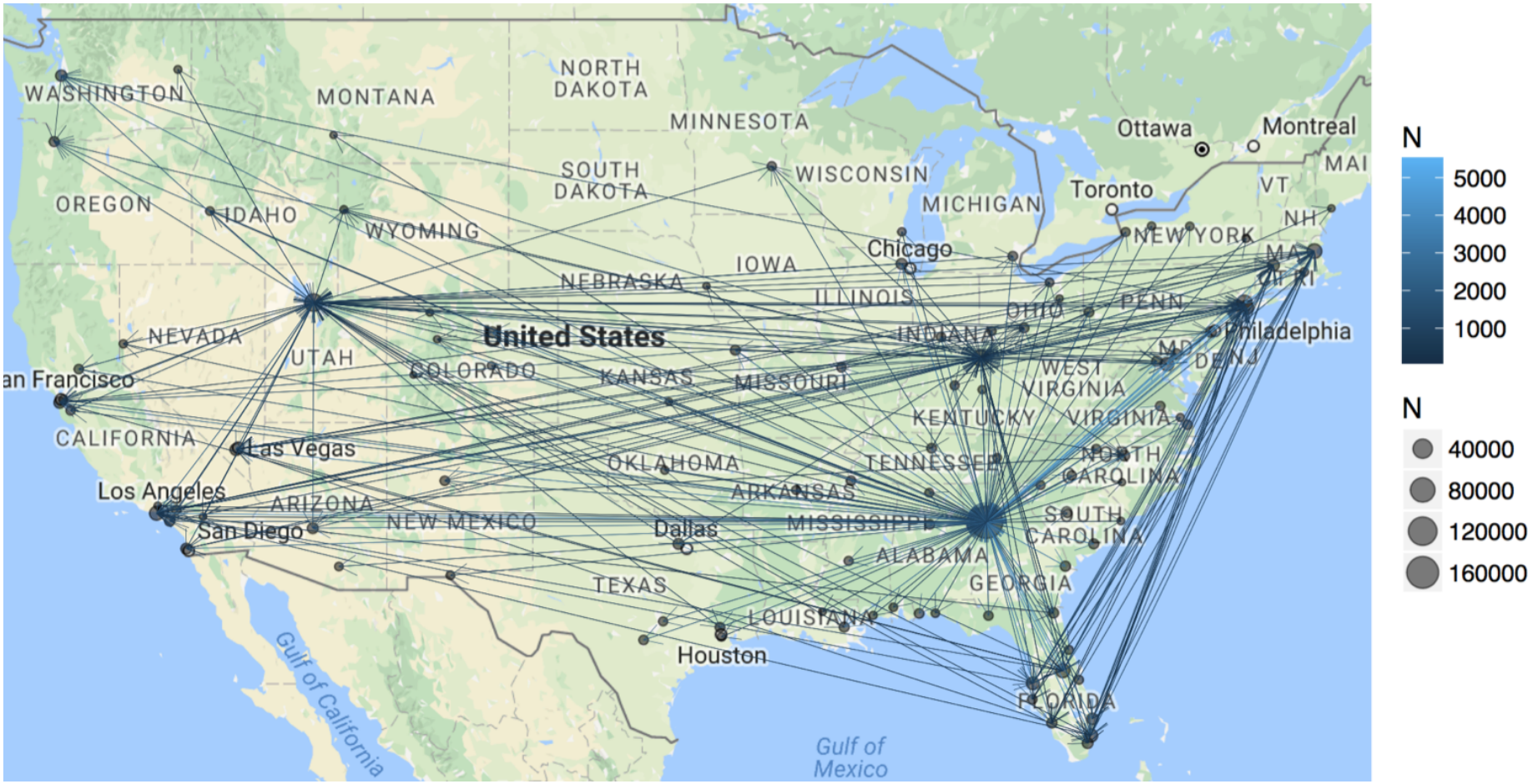


Dynamic data: ‘Velocity’ and data streams

```
library(leaflet)
m <- leaflet() %>%
  addTiles() %>% # Add OpenStreetMap tiles
  addMarkers(lng = -73.5673, lat = 45.5017,
             popup = "Montreal")
m
```



Network science and bigger data



More big ideas to ensure that we stay relevant

1. Increasing importance of data science and computation
2. Real applications and data
3. Statistical methods and foundations

Big idea #3: Statistical methods and foundations

ASA undergrad guidelines for statistics programs:

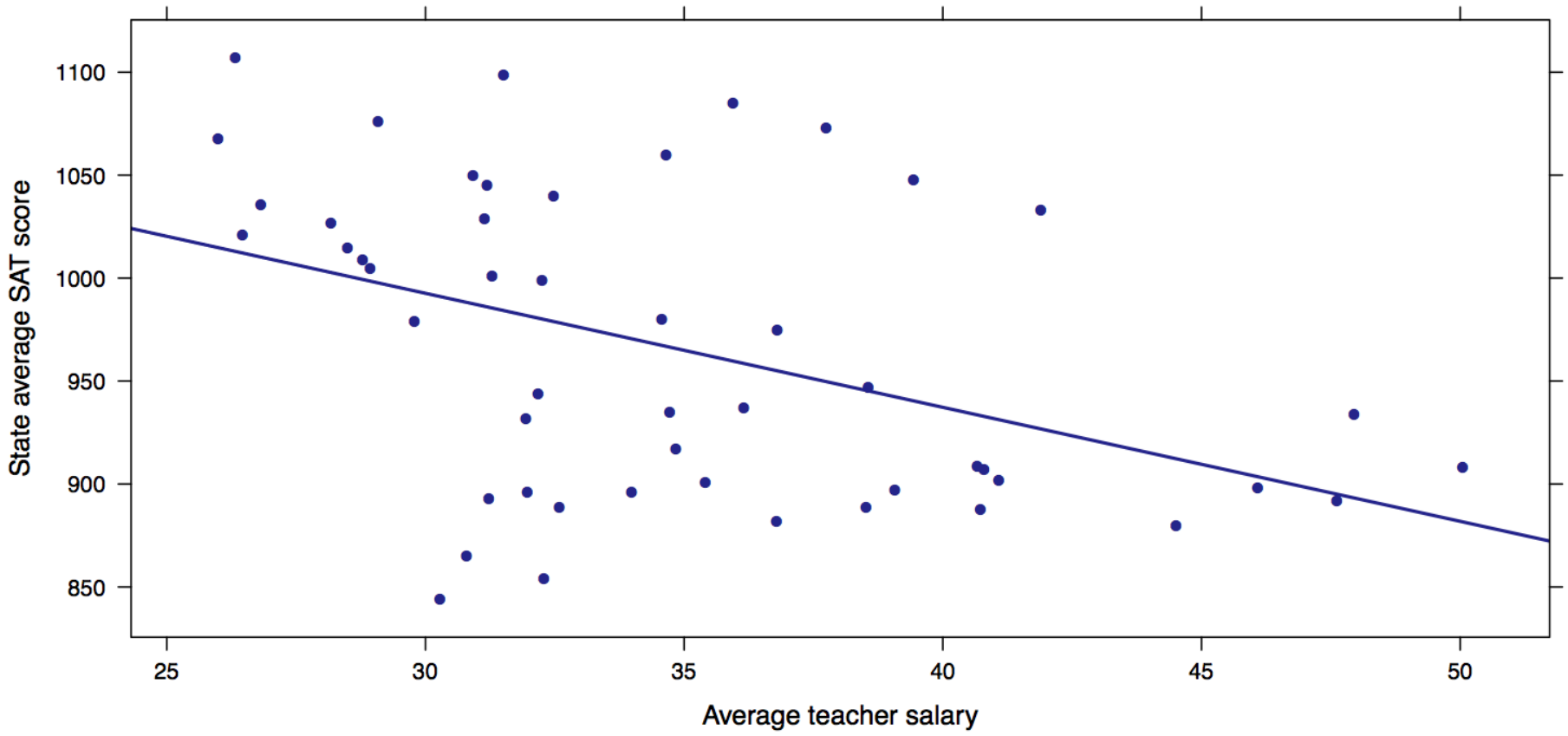
- Students require exposure to and practice with a variety of **predictive and explanatory models** in addition to methods for model building and assessment.
- They must be able to understand issues of **design, confounding, and bias**.
- They need to know how to apply their knowledge of **theoretical foundations** to the sound analysis of data.

NAS Undergraduate Data Science Education (https://nas.edu/envisioningds) report 2018

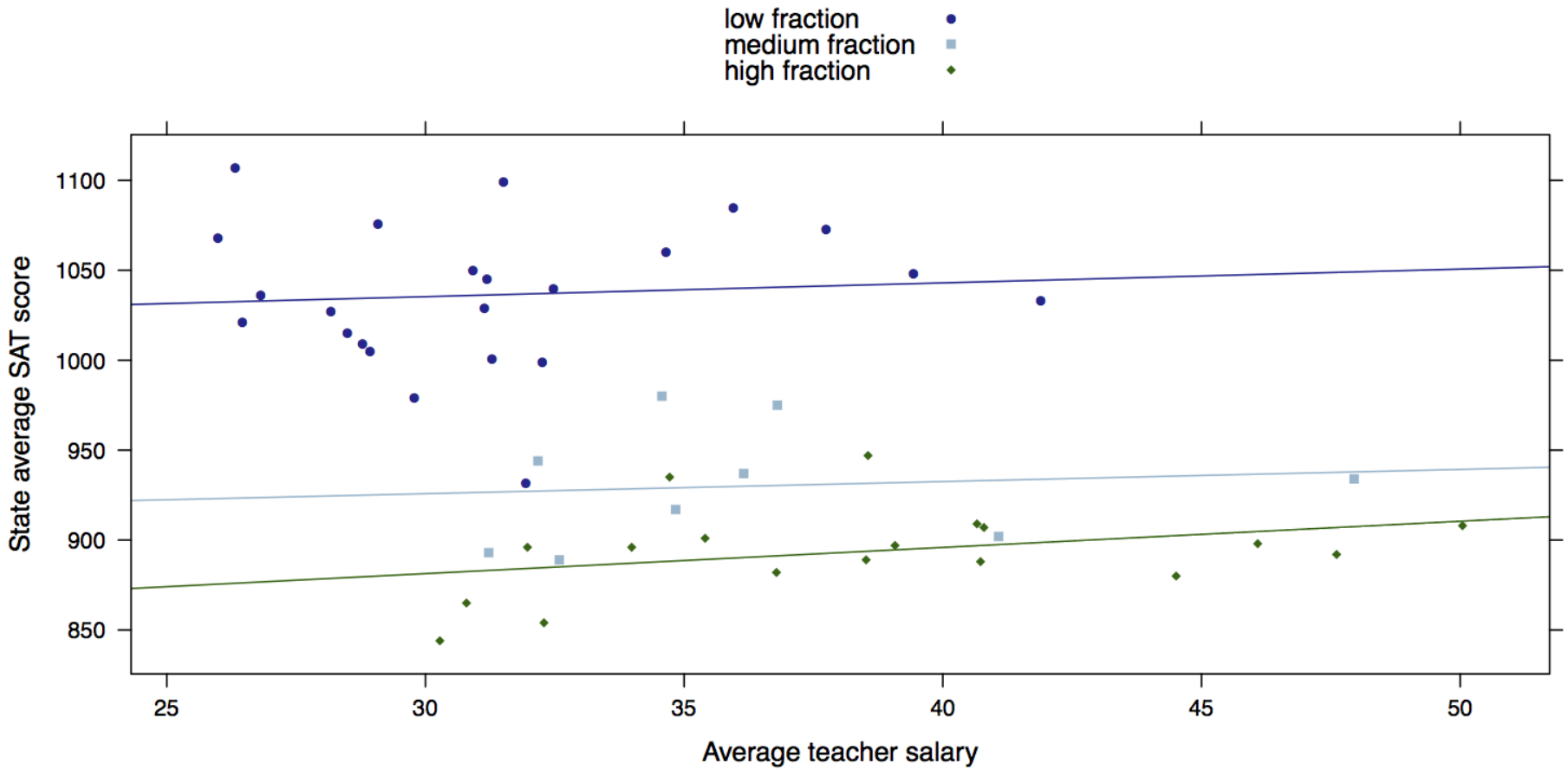
Important statistical foundations might include the following:

- Variability, uncertainty, sampling error, and inference
- Multivariate thinking,
- Nonsampling error, design, experiments (e.g., A/B testing), biases, confounding, and causal inference,
- Exploratory data analysis,
- Statistical modeling and model assessment, and
- Simulations and experiments

College entrance scores and teacher salaries (US state data from 2010)



Multivariate thinking and confounding



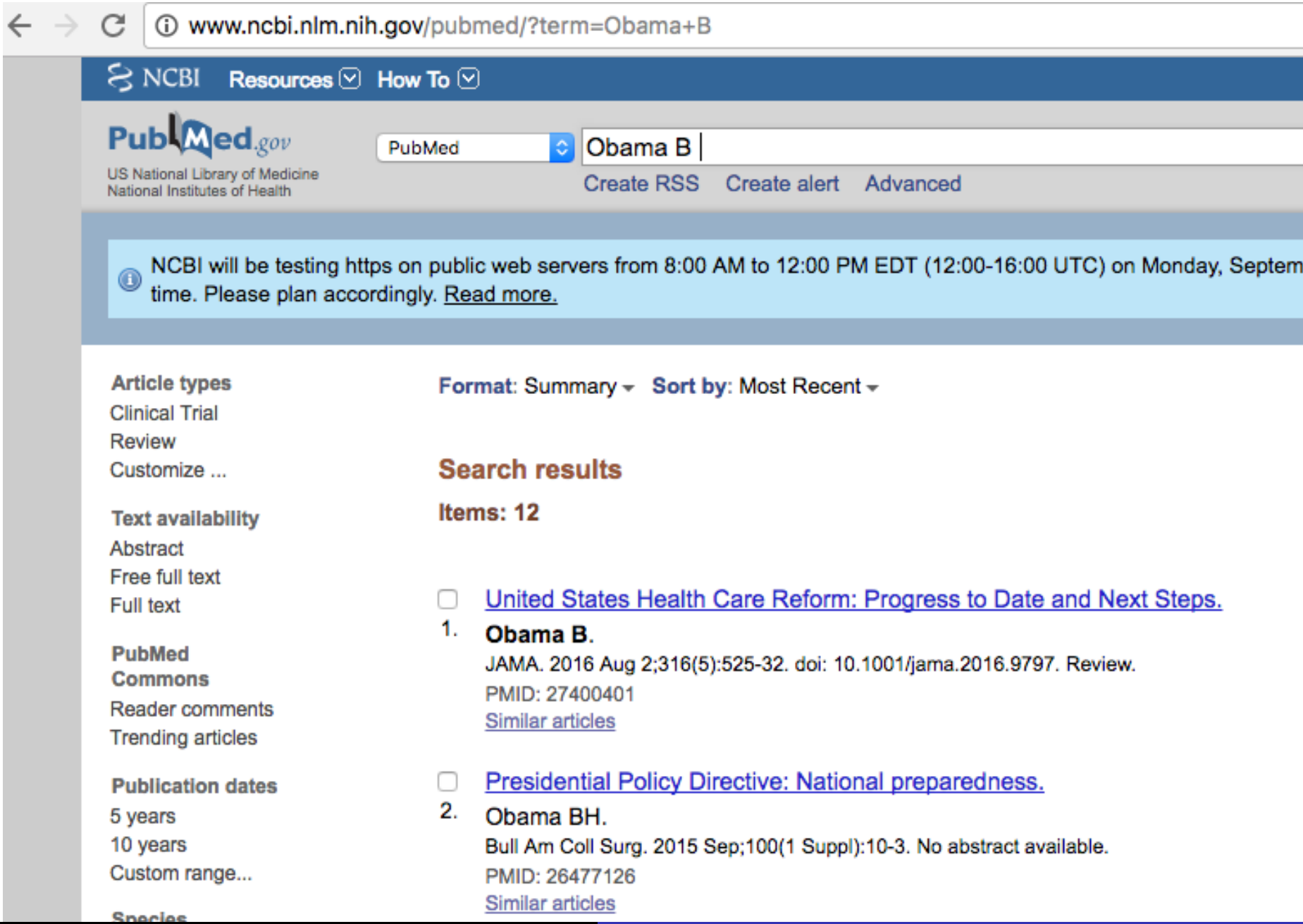
AP Statistics Vocabulary

Both Sides

confounding

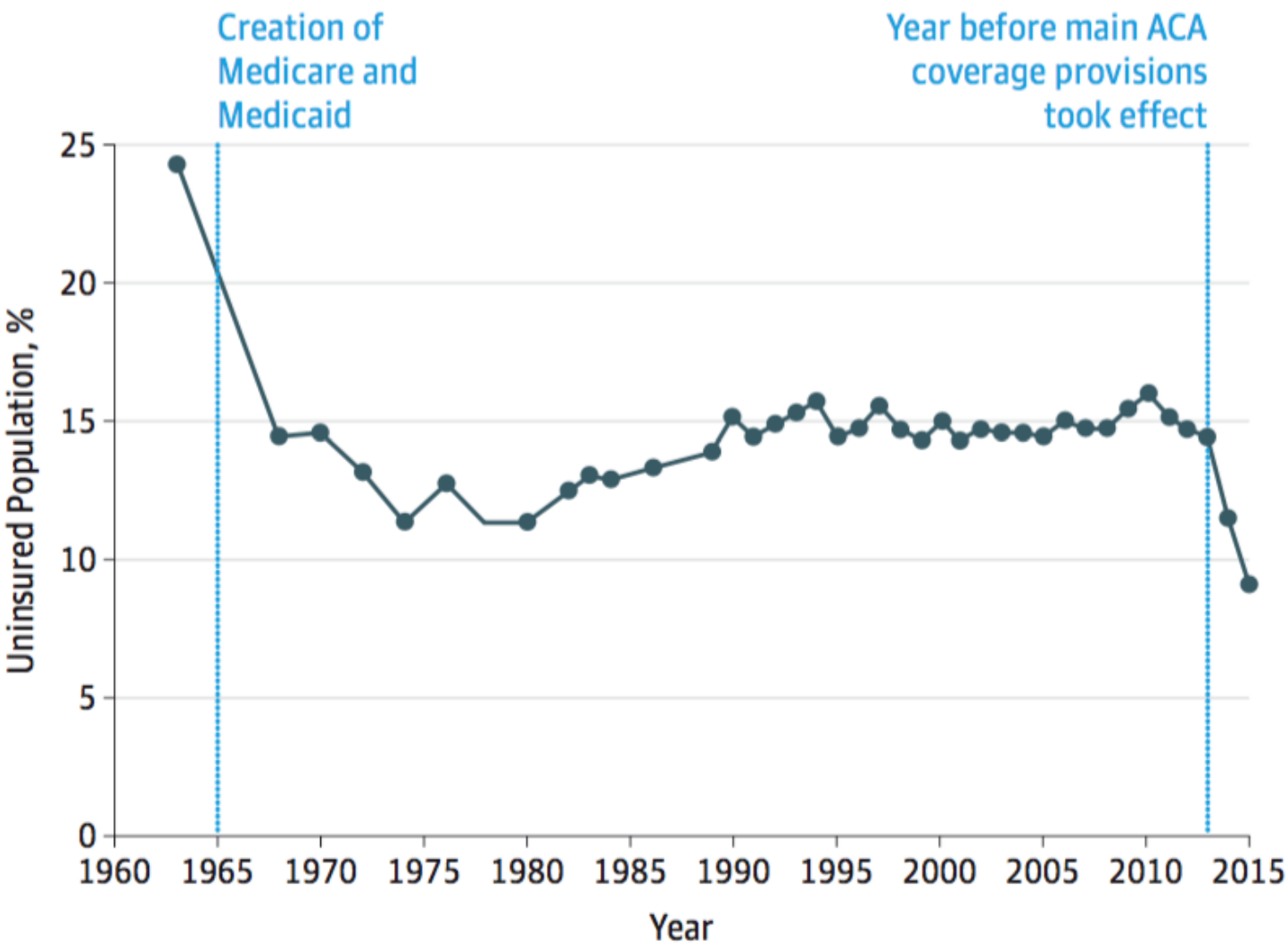
when the levels of one factor are associated with the levels of another factor so their effects cannot be separated

Design and confounding: President Obama's publications



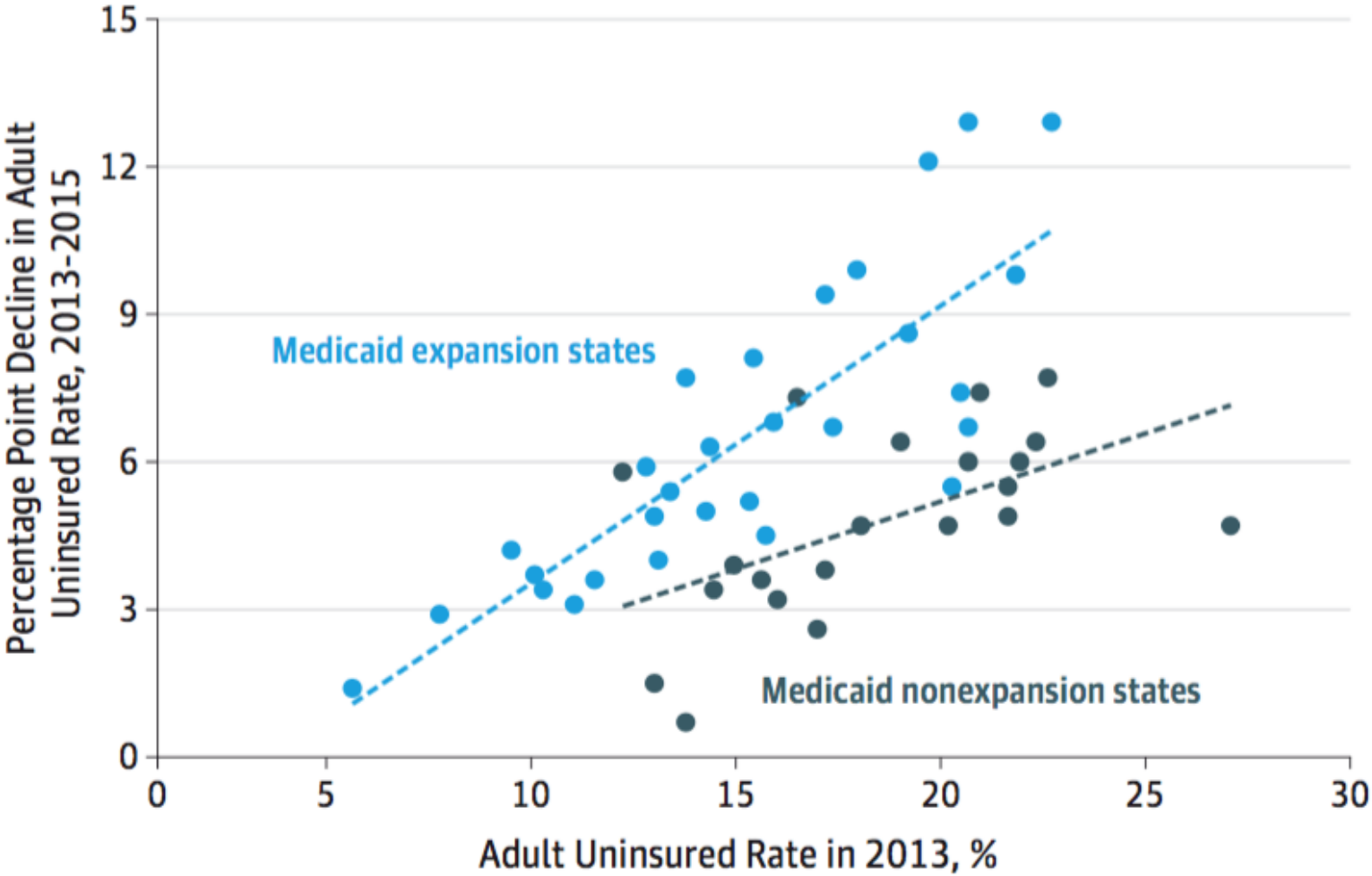
Obama's single author JAMA paper

Figure 1. Percentage of Individuals in the United States Without Health Insurance, 1963-2015



Obama's single author JAMA paper

Figure 2. Decline in Adult Uninsured Rate From 2013 to 2015 vs 2013 Uninsured Rate by State



Empirical problem-solving

Setting: Let A, B, and C be independent random variables each independently distributed uniformly in the interval [0,1].

Question: What is the probability that the roots of the quadratic equation given by $Ax^2 + Bx + C = 0$ are real?

Source: Rice Mathematical Statistics and Data Analysis third edition exercise 3.11 (also in first and second editions)

Empirical problem-solving

Setting: Let A, B, and C be independent random variables each independently distributed uniformly in the interval [0,1].

Question: What is the probability that the roots of the quadratic equation given by $Ax^2 + Bx + C = 0$ are real?

Source: Rice Mathematical Statistics and Data Analysis third edition exercise 3.11 (also in first and second editions)

Note: I continue to use this excellent book for my probability and statistical foundations courses

The distribution of $Y = B^2$ is given by:

$$f(y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The distribution of $W = 4AC$ is given by:

$$f(w) = \begin{cases} -\log(w/4)/4 & \text{if } 0 \leq w \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Since Y and W are independent, the joint distribution is given by:

$$f(y, w) = \begin{cases} \frac{-\log(w/4)}{8\sqrt{y}} & \text{if } 0 \leq y \leq 1 \text{ and } 0 \leq w \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

The discriminant $B^2 - 4AC$ is non-negative when $Y > W$.

$$\begin{aligned} P(Y > W) &= \int_0^1 \int_0^y f(y, w)\, dw\, dy \\ &= \int_0^1 \int_0^y \frac{-\log(w/4)}{8\sqrt{y}}\, dw\, dy \\ &= \int_0^1 \frac{\sqrt{y}(-\log(y) + 1 + \log(4))}{8}\, dy \\ &= \frac{5 + \log(64)}{36} \approx 0.254413. \end{aligned}$$

Empirical problem-solving

- Answer in the back of the book: 1/9

Empirical problem-solving

- Answer in the back of the book: 1/9
- Straightforward to code in R (or other environments):

```
> numsim <- 1000000
> A <- runif(numsim)
> B <- runif(numsim)
> C <- runif(numsim)
> discrim <- B^2 - 4 * A * C
> realroot <- discrim >= 0
```

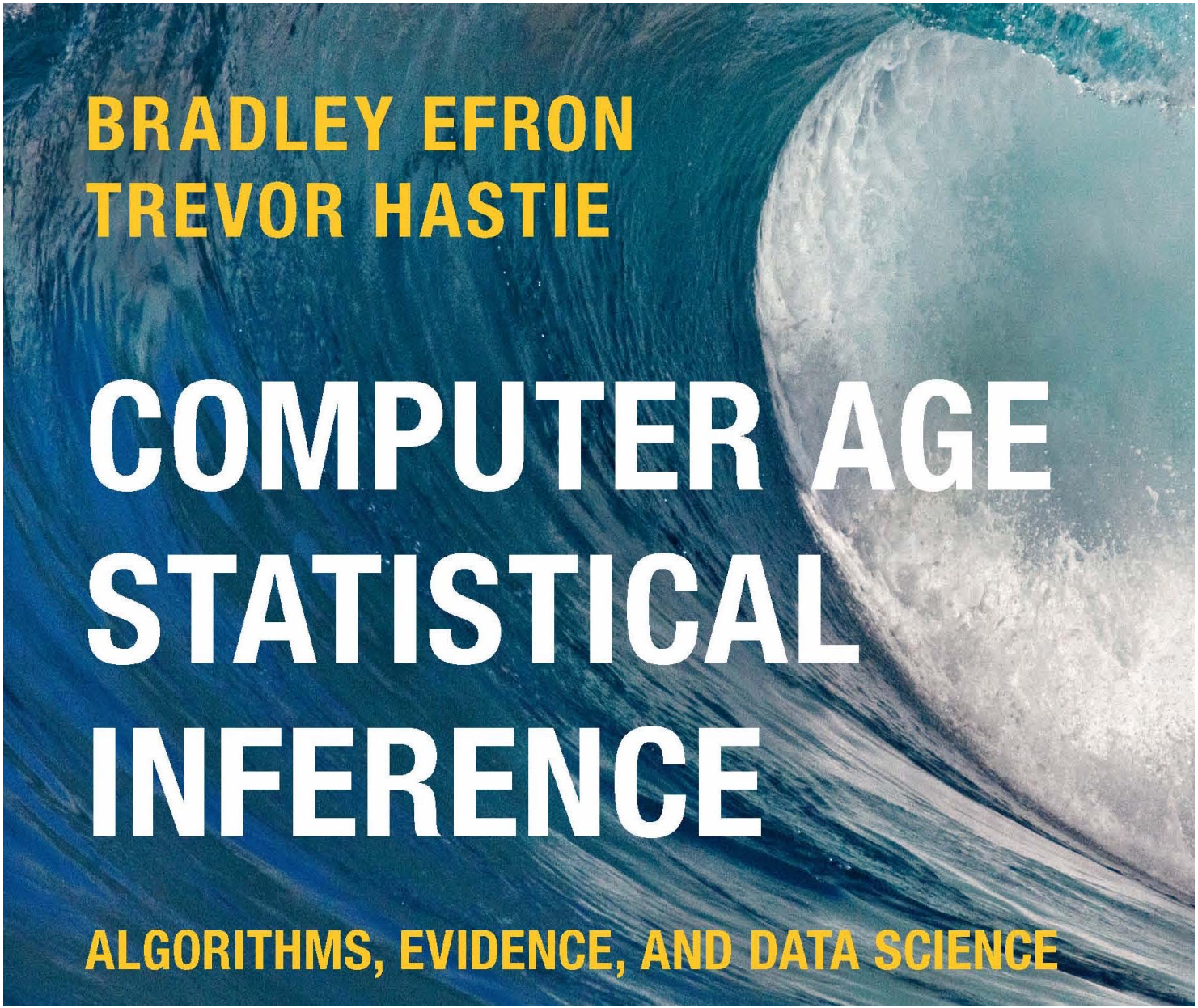
Empirical problem-solving

- Answer in the back of the book: 1/9
- Straightforward to code in R (or other environments):

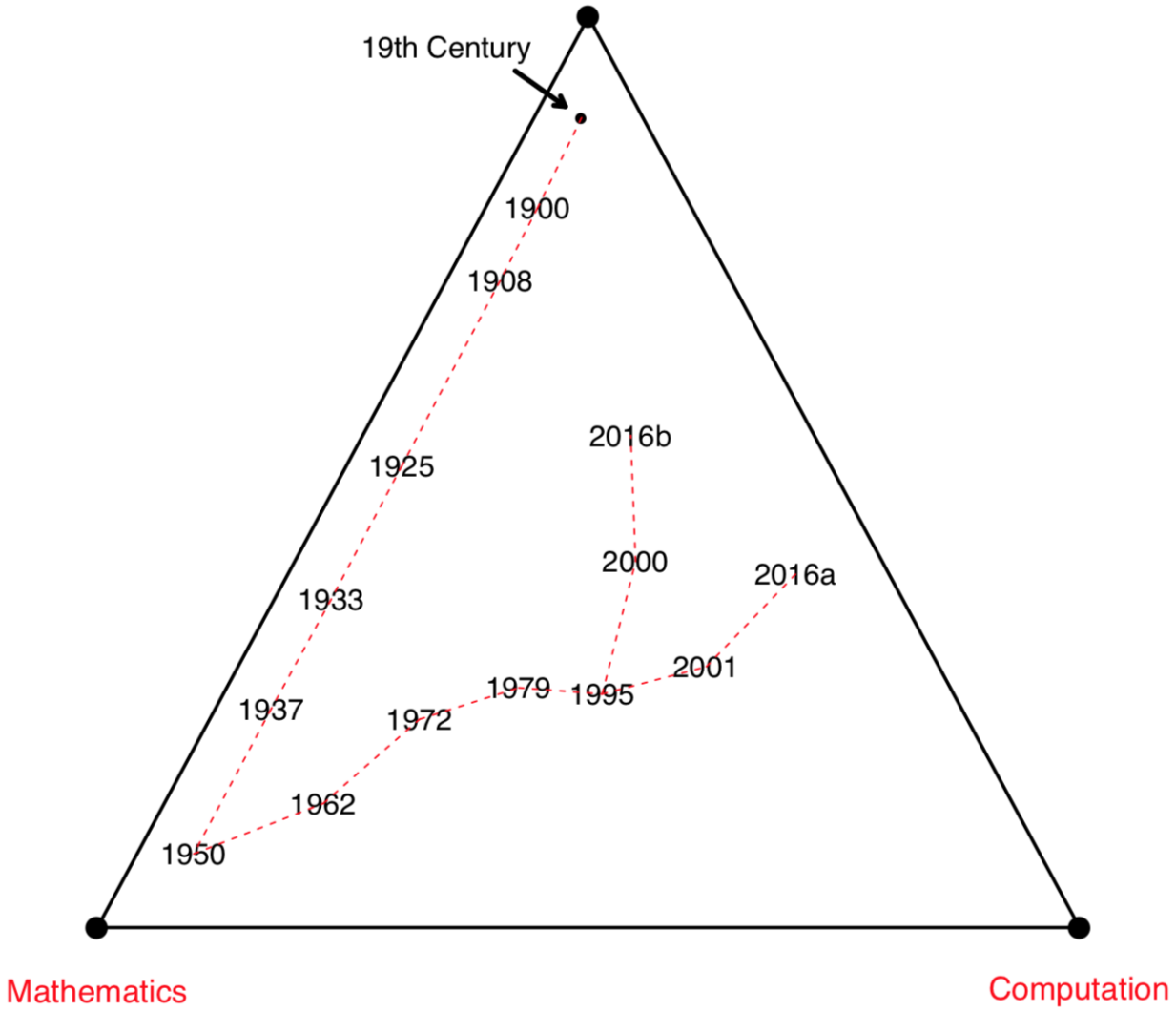
```
> numsim <- 1000000
> A <- runif(numsim)
> B <- runif(numsim)
> C <- runif(numsim)
> discrim <- B^2 - 4 * A * C
> realroot <- discrim >= 0

> binom.test(realroot==TRUE, numsim)
95 percent confidence interval:
 0.2537 0.2554
sample estimates:
probability of success
                0.2545
```

IMPLICATION?

The continuing importance of theory

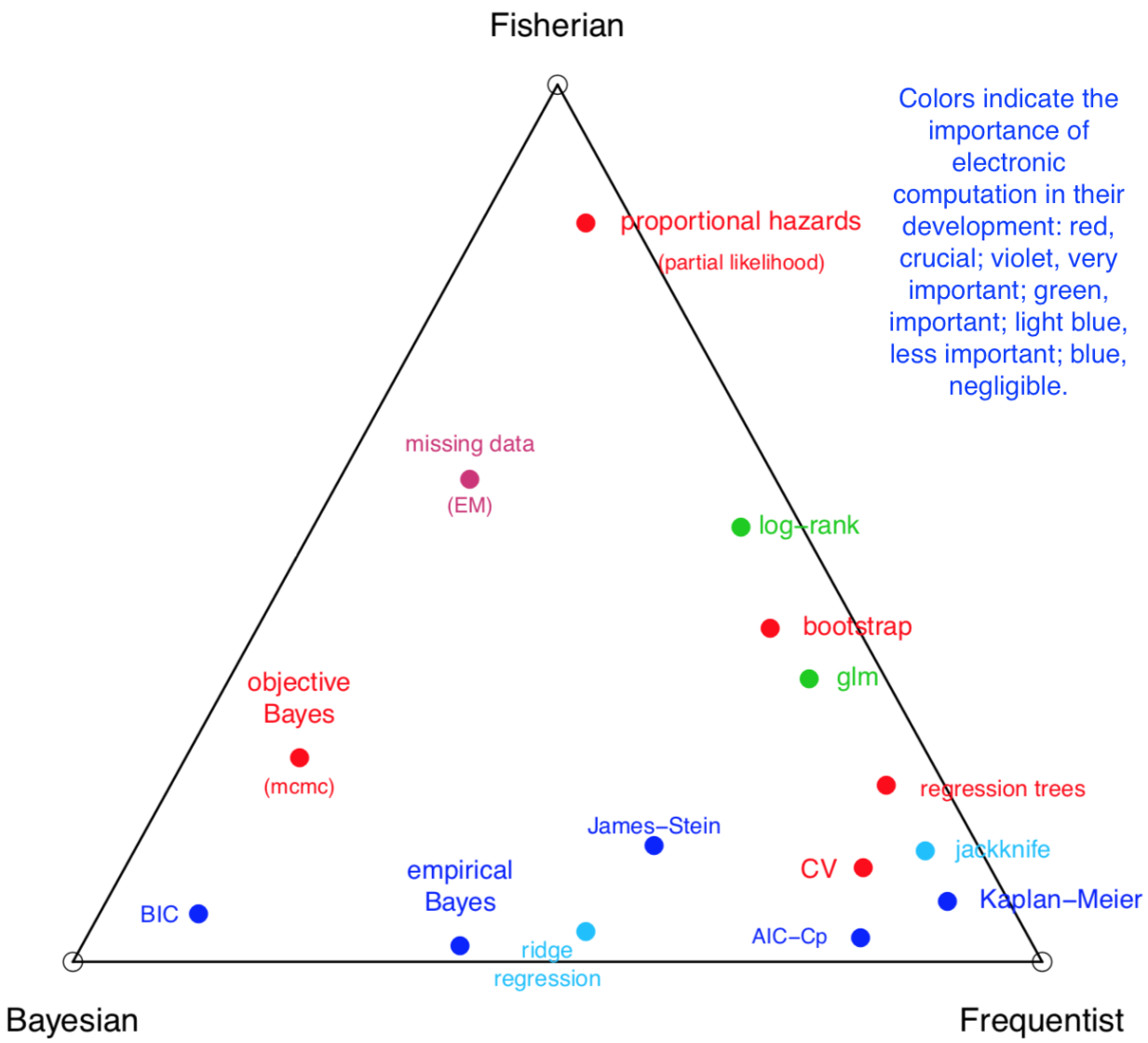


The continuing importance of theory (CASI)



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

The continuing importance of theory (CASI)



Still need theory

1. but perhaps not the same focus on p-values
2. needs to be introduced and reinforced along with computation
3. what are the key ideas and approaches that statistics brings to data science?
4. what to teach in theoretical statistics? (See Horton *TAS*, 2013.)

Last of the big ideas to ensure that we stay relevant

1. Increasing importance of data science and computation
2. Real applications and data
3. Statistical methods and foundations
4. Communication and knowledge transference

Big idea #4: Communication and knowledge transference

ASA undergrad guidelines for statistics programs:

- Students need to be able to **communicate** complex statistical methods in basic terms to managers and other audiences and to visualize results in an accessible manner.
- They must have a clear understanding of **ethical standards**.
- Programs should provide **multiple opportunities to practice and refine** these statistical practice skills and use of analysis cycle.

R Markdown and reproducible analysis

The ability to express statistical computations is an essential skill (Nolan and Temple Lang, TAS 2010)

- R Markdown used as first workflow for introductory statistics students at colleges and universities all over the country (Baumer et al, *TISE*, 2014)
- easy to deploy as a cloud application (fewer barriers for students)
- forms a 'necessary but not sufficient' component of reproducible research
- tightly integrated into RStudio (designed for experts, useful for newbies)
- also available in environments such as Jupyterhub

Dynamic visualization and Shiny

https://r.amherst.edu/apps/nhorton/sat/

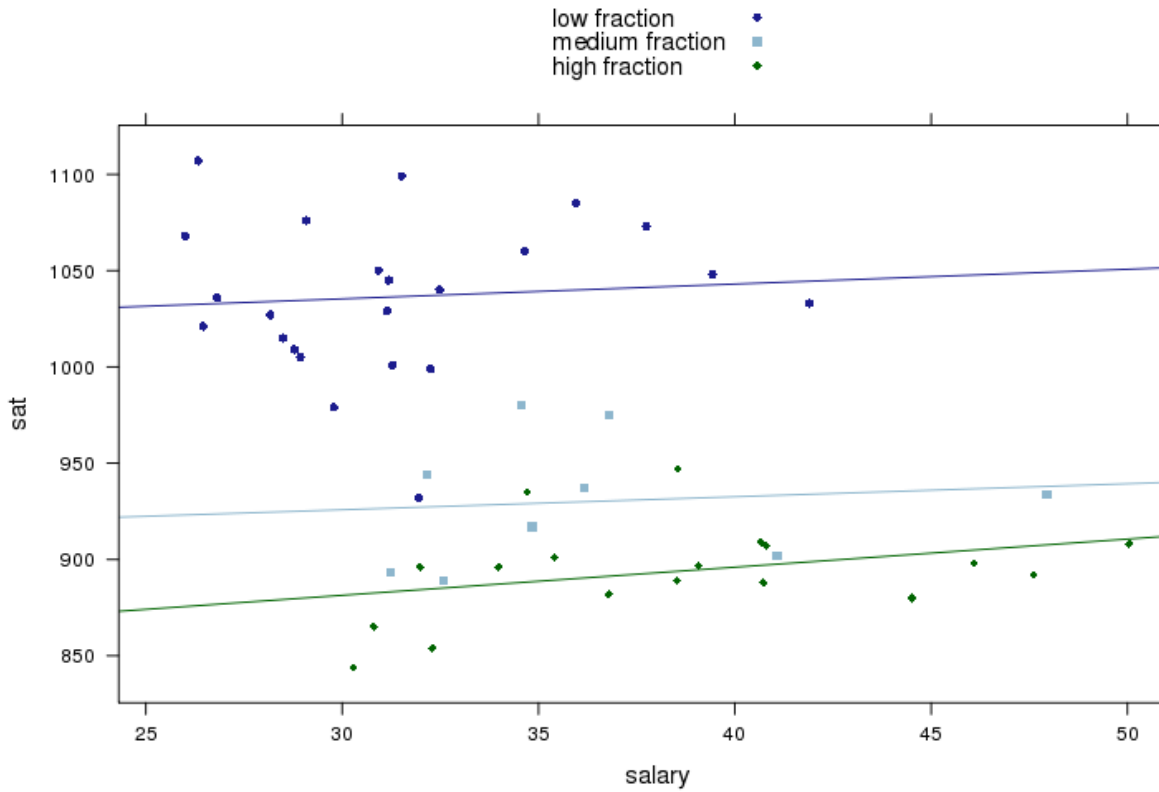
SAT scores and teacher salaries

Stratify by percent taking SAT?

Yes

Yes

No



Dynamic visualization and Shiny

server.R | ui.R

```
shinyServer(function(input, output) {
  output$distPlot <- renderPlot({

    # mosaic setup
    require(mosaic); require(mosaicData)
    trellis.par.set(theme=theme.mosaic())

    # create new variable
    SAT = mutate(SAT, fracgrp = cut(frac,
      breaks=c(0, 22, 49, 81),
      labels=c("low fraction", "medium fraction", "high fraction")))

    # generate the desired plot
    if (input$stratify == "No") {
      xyplot(sat ~ salary, type=c("p", "r"), data=SAT)
    } else {
      xyplot(sat ~ salary, groups=fracgrp, auto.key=TRUE,
        type=c("p", "r"), data=SAT)
    }

    })
  }
)
```

Dynamic visualization and Shiny

server.R | ui.R

```
library(shiny)

shinyUI(fluidPage(
  # Application title
  titlePanel("SAT scores and teacher salaries"),

  sidebarLayout(
    sidebarPanel(
      selectInput("stratify", "Stratify by percent taking SAT?",
                  choices = c("Yes", "No"), selected="No")),
    # Show a plot of the generated distribution
    mainPanel(plotOutput("distPlot"))
  )
))
```

Revision control systems

> *Version control is the only reasonable way to keep track of changes in code, manuscripts, presentations, and data analysis projects.*

Karl Broman,
`http://kbroman.org/github_tutorial/pages/why.html`

Revision control systems

> *Version control is the only reasonable way to keep track of changes in code, manuscripts, presentations, and data analysis projects.*

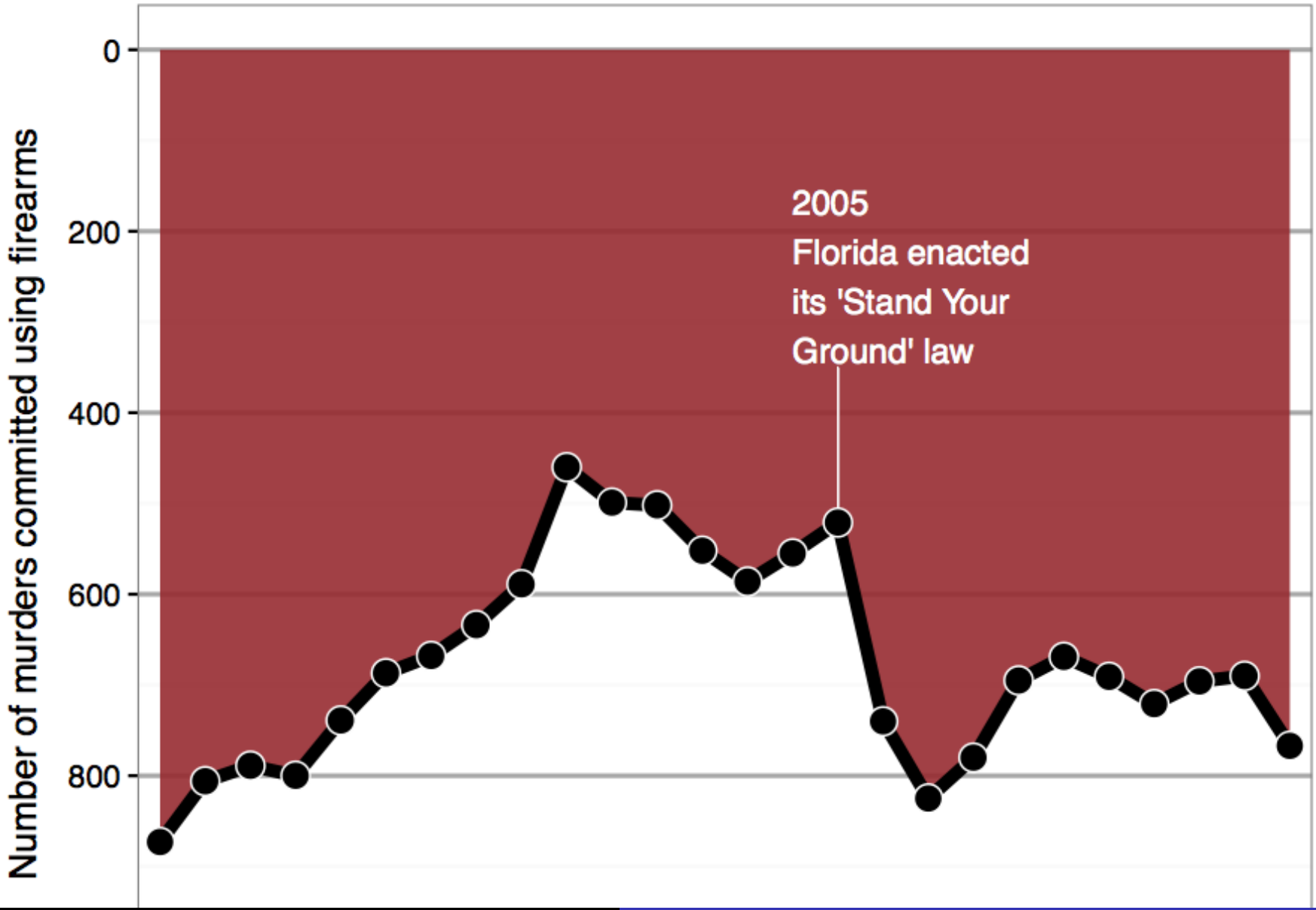
Karl Broman,
`http://kbroman.org/github_tutorial/pages/why.html`

> *If you need to collaborate on data analysis or code development, then all involved should use Git.*

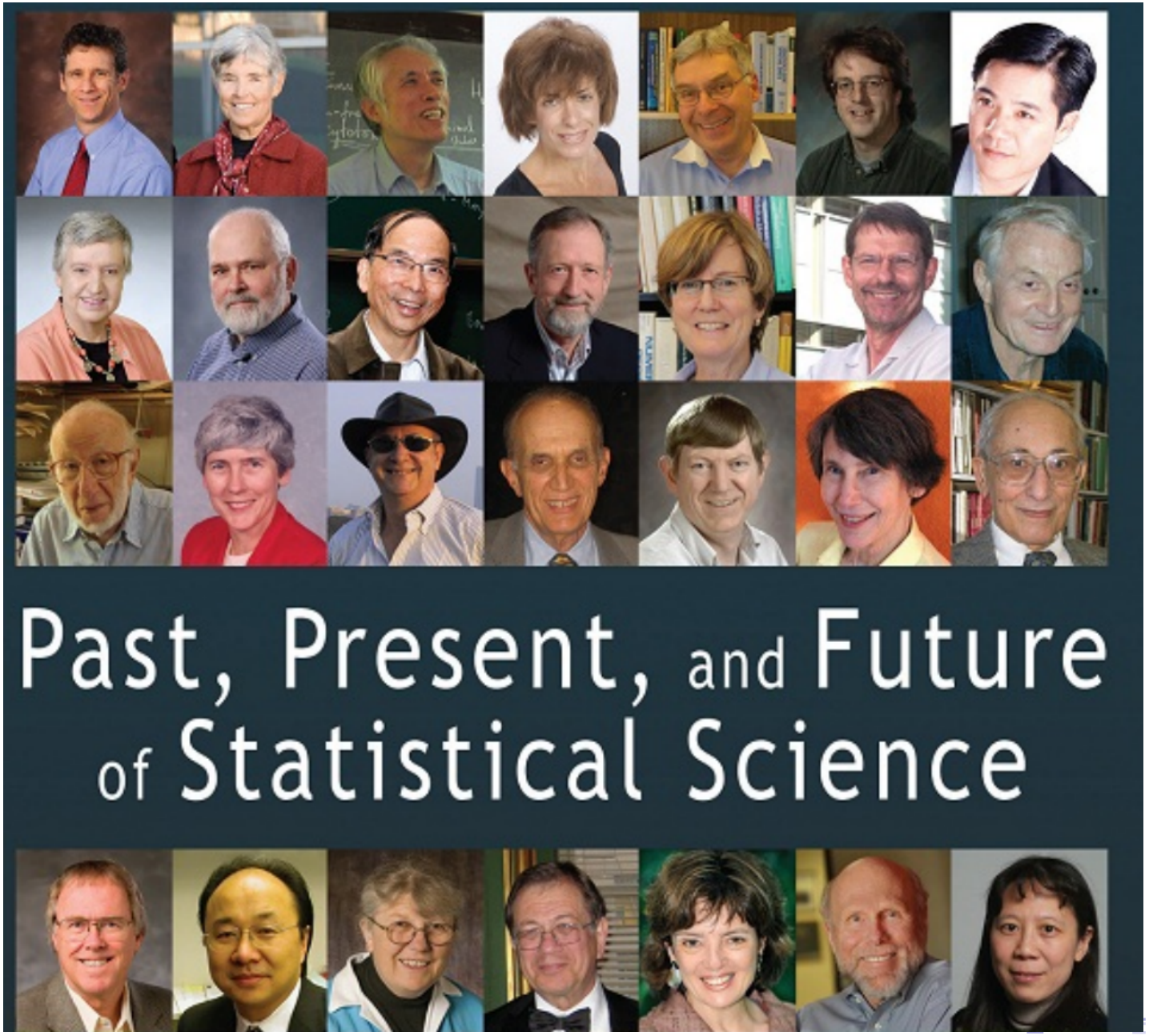
Jenny Bryan, `http://happygitwithr.com`

Ethics

If not now, when?



Class use of *Past, Present, and Future*



Challenges and opportunities for statistics

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

revised GAISE College Report (2016)

1. Teach statistical thinking.
   - Teach statistics as an **investigative process of problem-solving and decision-making**.
   - Give students experience with **multivariable thinking**.
2. Focus on conceptual understanding.
3. Integrate **real data** with a context and purpose.
4. Foster **active learning**.
5. Use **technology to explore concepts and analyze data**.
6. Use assessments to improve and evaluate student learning.

Key learning outcomes from Berkeley's Data8.org course

- Calculate specified statistics of a given dataset.
- Identify the sources of randomness in an experiment.
- Formulate a null hypothesis that relates to a given question, which can be assessed using a statistical test.
- Carry out statistical analyses including computing confidence intervals and performing hypothesis tests in a variety of data settings.
- Given the result of a statistical analysis from the course, form correct conclusions about a question based on its meaning.

Key learning outcomes from Berkeley's Data8.org course (cont.)

- Given a question and an analysis, explain whether the analysis addresses the question and how the analysis could change and still address the question.
- Articulate the benefits and limits of computing technology for analyzing data and answering questions.
- Correctly generate and interpret histograms, bar charts, and box plots.
- **Correctly make predictions using regression and classification techniques.**
- **Assess the accuracy and variability of a prediction.**
- (Plus how to write a function!)

Challenges and opportunities

Read the Introducing AP Computer Science Principles video transcript

Computer Science: The New Literacy

Whether it's 3-D animation, engineering, music, app development, medicine, visual design, robotics, or political analysis, computer science is the engine that powers the technology, productivity, and innovation that drive the world. Computer science experience has become an imperative for today's students and the workforce of tomorrow.

The AP Program designed AP Computer Science Principles with the goal of creating leaders in computer science fields and attracting and engaging those who are traditionally underrepresented with essential computing tools and multidisciplinary opportunities.

Largest first year AP exam ever (45,000 students took the exam)

Challenges and opportunities

Read the Introducing AP Computer Science Principles video transcript

Computer Science: The New Literacy

Whether it's 3-D animation, engineering, music, app development, medicine, visual design, robotics, or political analysis, computer science is the engine that powers the technology, productivity, and innovation that drive the world. Computer science experience has become an imperative for today's students and the workforce of tomorrow.

The AP Program designed AP Computer Science Principles with the goal of creating leaders in computer science fields and attracting and engaging those who are traditionally underrepresented with essential computing tools and multidisciplinary opportunities.

Largest first year AP exam ever (45,000 students took the exam)
Second year (numbers still rough) more than 83,000 students took the exam

Big Idea 3: Data and Information

Data and information facilitate the creation of knowledge. Computing enables and empowers new methods of information processing, driving monumental change across many disciplines — from art to business to science. Managing and interpreting an overwhelming amount of raw data is part of the foundation of our information society and economy. People use computers and computation to translate, process, and visualize raw data and to create information. Computation and computer science facilitate and enable new understanding of data and information that contributes knowledge to the world. Students in this course work with data using a variety of computational tools and techniques to better understand the many ways in which data is transformed into information and knowledge.

Challenges and opportunities

Enduring Understandings (Students will understand that …)	**Learning Objectives** (Students will be able to …)
EU 3.1 People use computer programs to process information to gain insight and knowledge.	**LO 3.1.1** Find patterns and test hypotheses about digitally processed information to gain insight and knowledge. [P4]

Challenges and opportunities

LO 3.1.3 Explain the insight and knowledge gained from digitally processed data by using appropriate visualizations, notations, and precise language. [P5]	**EK 3.1.3A** Visualization tools and software can communicate information about data.
	EK 3.1.3B Tables, diagrams, and textual displays can be used in communicating insight and knowledge gained from data.
	EK 3.1.3C Summaries of data analyzed computationally can be effective in communicating insight and knowledge gained from digitally represented information.
	EK 3.1.3D Transforming information can be effective in communicating knowledge gained from data.
	EK 3.1.3E Interactivity with data is an aspect of communicating.

Challenges and opportunities

EU 3.2 Computing facilitates exploration and the discovery of connections in information.

LO 3.2.1 Extract information from data to discover and explain connections or trends. [P1]

Closing thoughts

> *Curriculum unavoidably involves decisions about scarce resources, so curricular innovation cannot escape being political, and of course "all politics is local" (ONeill and Hymel, 1995).*
>
> *Curriculum is political for economic reasons because, averaged over the long term, faculty FTEs and course offerings are at best a zero-sum game. Thus* **changing curriculum, like moving a graveyard, depends on local conditions: Whose cherished ancestry is uprooted by the change?**

(Cobb 'Mere renovation is too little, too late: we need to rethink our undergraduate curriculum from the ground up' arXiv 2015)

Closing thoughts

- it's never been easier to extract meaning from data (improved tools)
- era of (cheap) cloud computing: transformative opportunities to simplify access for students
- include multivariate thinking (and multiple regression) in intro stats (MLR now 20% of our intro course)
- bolster computing early and often in statistics courses (beginning with intro and continuing in probability/theoretical stat)
- use project-based learning to teach statistics and data science analysis cycle and reproducible workflows
- perhaps less focus on p-values?

Big Ideas to Help Statistics Students Learn to Think with Data

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

Statistical Society of Canada, June 6, 2018

nhorton@amherst.edu
http://nhorton.people.amherst.edu