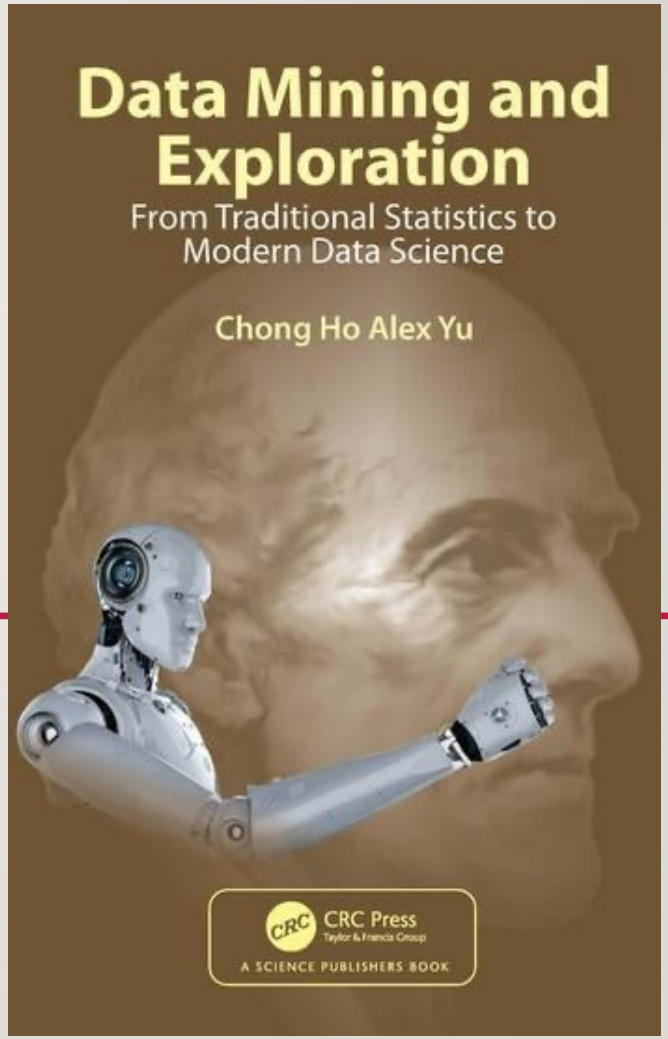
FROM TRADITIONAL STATISTICS TO MODERN DATA SCIENCE AND MACHINE LEARNING

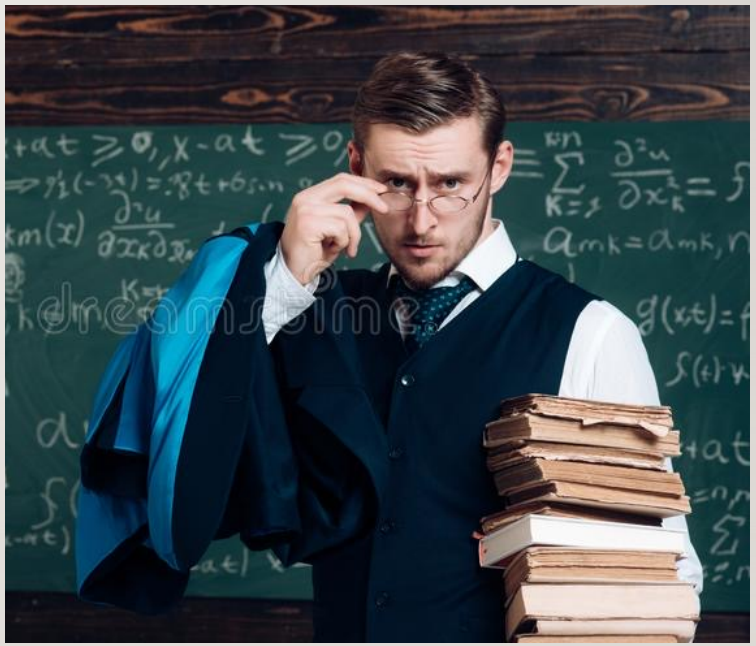
CHONG HO YU (ALEX)

SEMINAR AT 2023 SOUTH CALIFORNIA CHAPTER OF AMERICAN STATISTICAL ASSOCIATION MEETING



A HYPOTHETICAL STORY

- Karl Fisher earned his Ph.D. in statistics in 1981. After graduation he became a professor of statistics and research methods at a prestigious university.
- His course covered all essential concepts and procedures in the field, including t-test, ANOVA, Pearson's correlation, ordinal least square (OLS) regression, Chi-square analysis, multivariate procedures, and others.

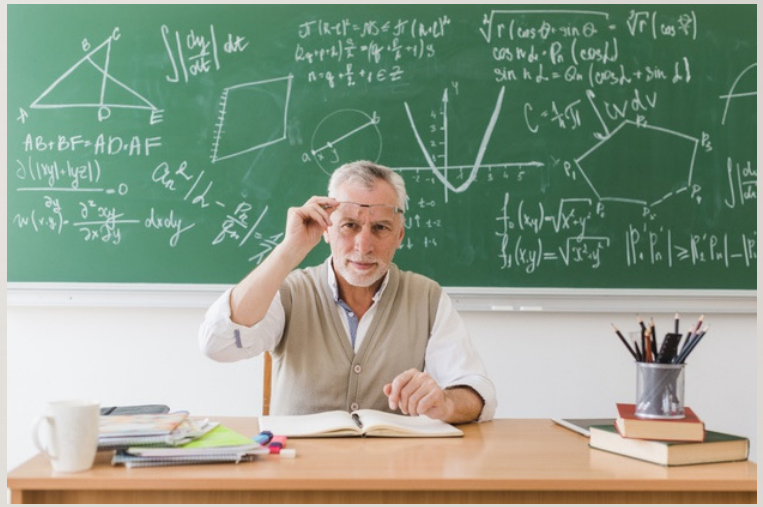


A HYPOTHETICAL STORY

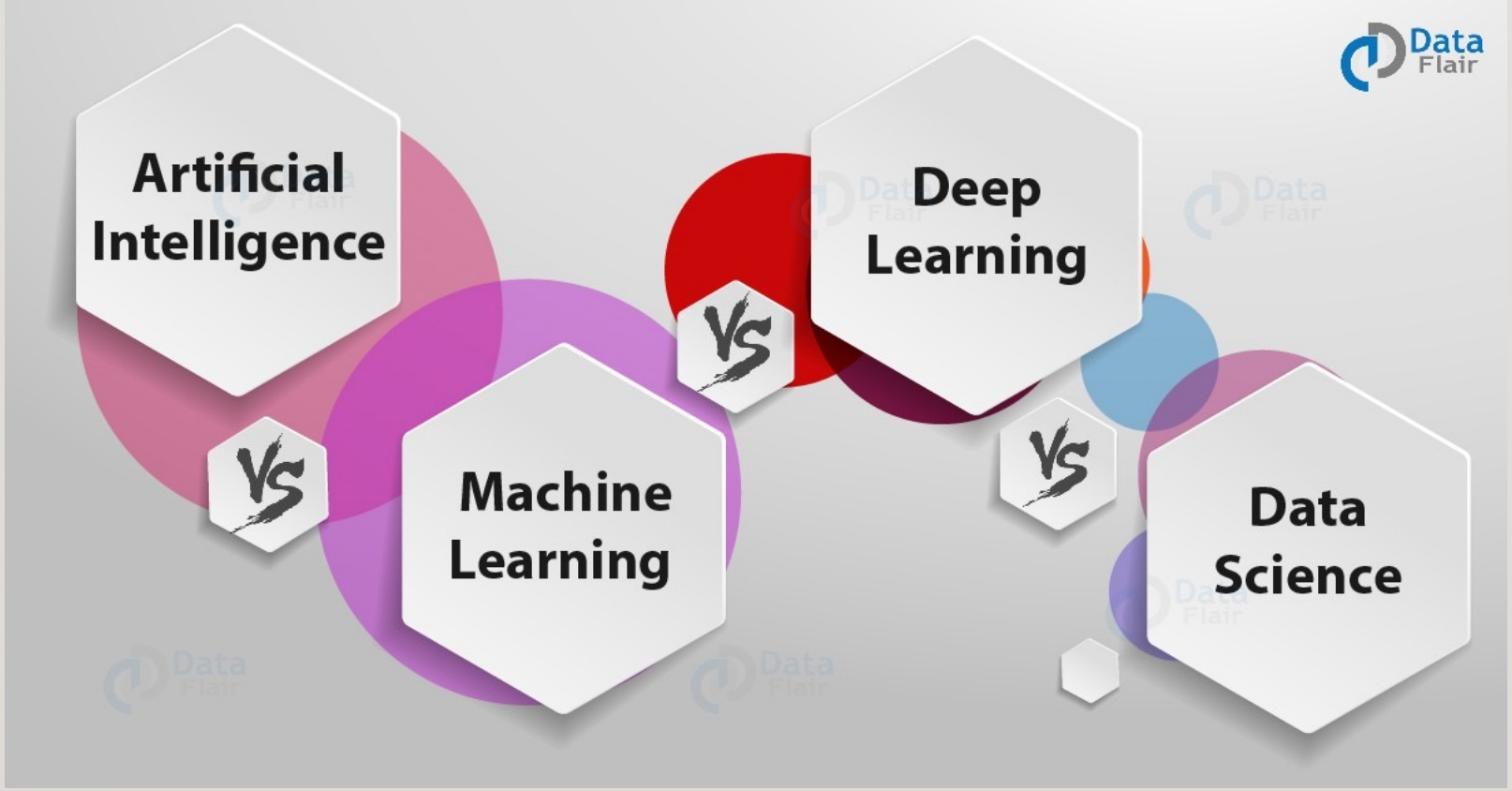
- In 1993 tragedy struck the young professor when he was severely injured in a car accident. He fell into a deep coma for three decades.
- In 2023, he woke up and miraculously regained all his cognitive abilities, but he was afraid that he would be unable to find a job due to a 30-year gap in familiarity with advancements in the field.

A HYPOTHETICAL STORY

- After reviewing several syllabi and textbooks, his confidence was soon restored because he found that the contents of statistical education remained unchanged.
- The introductory course attends to t-test, ANOVA, Pearson's correlation, OLS regression, Chi-square analysis, whereas the advanced course covers stepwise regression, logistic regression, and exploratory factor analysis.



WHAT IS HAPPENING IN THE WORLD (INDUSTRY)?



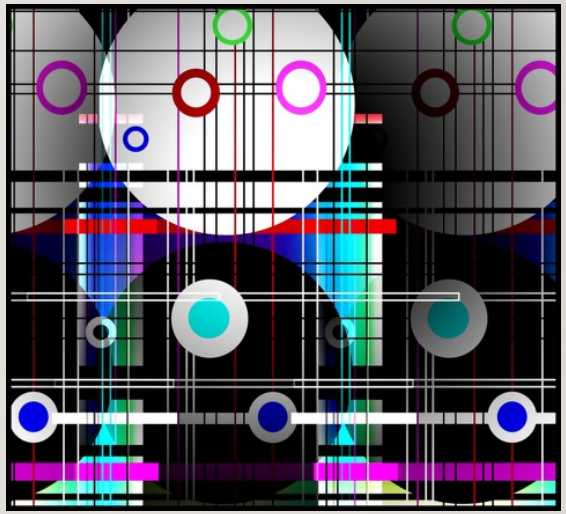
Credit: Rinu Gour

CLASSICAL METHODS

- 1805: The least square criterion for regression modeling
- 1892: Pearson's coefficient
- 1900: Chi-square analysis
- 1908: t-test
- 1930: ANOVA

MODERN DATA SCIENCE AND MACHINE LEARNING

- 1980s: Convolutional neural networks
- 1984: Classification (partition) tree
- 1993: C4.5 algorithm
- 1997: Gradient boosted tree
- 2001: Random forest
- 2005: Elastic net
- 2006: Adaptive LASSO
- 2014: XGBoost

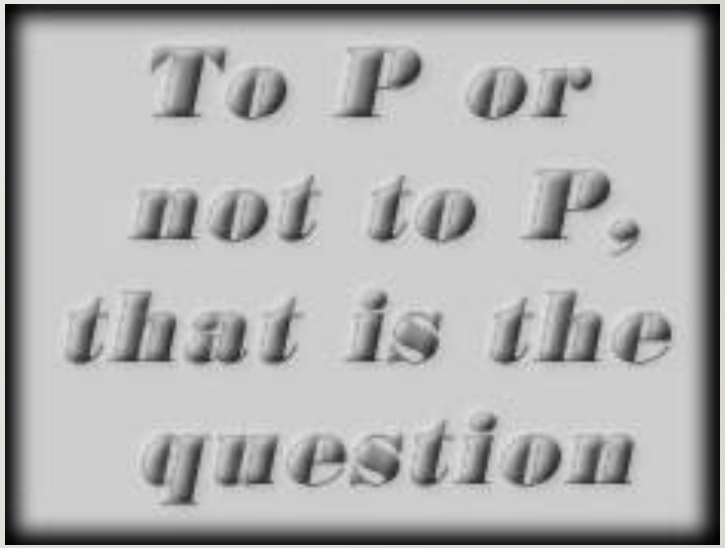


THE EARTH IS ROUND!

- Hypothesis testing is also known as confirmatory data analysis (CDA), which has been the norm in the academic world for about a century.
- In 1989, when Kenneth Rothman started the *Journal of Epidemiology*, he discouraged over–reliance on *p* values. However, the earth is round.
- When he left his position in 2001, the journal reverted to the *p*–value tradition (Nuzzo, 2014).

P(D|H) VS. P(D|H)

- Hypothesis-driven
- The logic of hypothesis testing is: Given that the null hypothesis is true, what is the chance of observing the data in the **long run**? P(D|H)?
- What we really want to know is: Given the data what is the best theory to explain the data **no matter whether the event can be repeated** : P(H|D)?



LOGICAL FALLACY

- Affirming the consequent
- P(D|H) <> P(H|D): “If H then D" does not logically imply "if D then H".
- If the theory/model/hypothesis is correct, it implies that we could observe Phenomenon X or Data X.
- X is observed.
- Hence, the theory is correct.

SHORTCOMING 1: DO NOT ADDRESS TREATMENT EFFECTIVENESS

- If George Washington was assassinated, then he is dead.
- George Washington is dead.
- Therefore, George Washington was assassinated.

- If it rains, the ground is wet.
- The ground is wet.
- It must rain.

EASY TO REJECT THE NULL

- When the sample size is large enough, even a trivial effect seems to be significant, resulting in rejecting the null hypothesis.
- In 2010-11 Daryl Bem conducted a series of experiment to assert that extrasensory perception (ESP) is real.
- *N* = 1,000
- In one study two curtains were presented to the participants and behind one of them is an erotic picture.

SHORTCOMING 2: EASY TO REJECT THE NULL

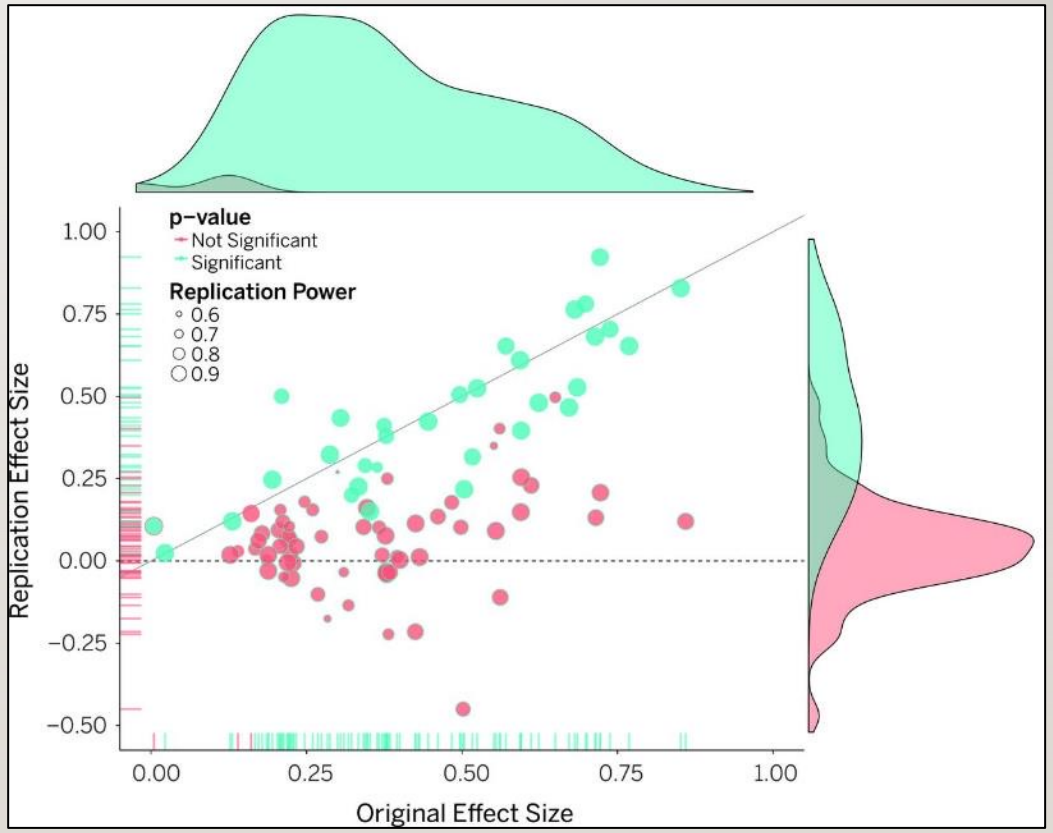
- By chance alone the participants should be correctly 50% but the actual hit rate is 53.1%. The *p* value is significant!
- The explanation is: Our erotic desire enables us to predict the location of the erotic image.
- 3.1% better. Is it a big deal?
- He reported the one-sided *p* value instead of the two-sided.

DO NOT ADDRESS TREATMENT EFFECTIVENESS

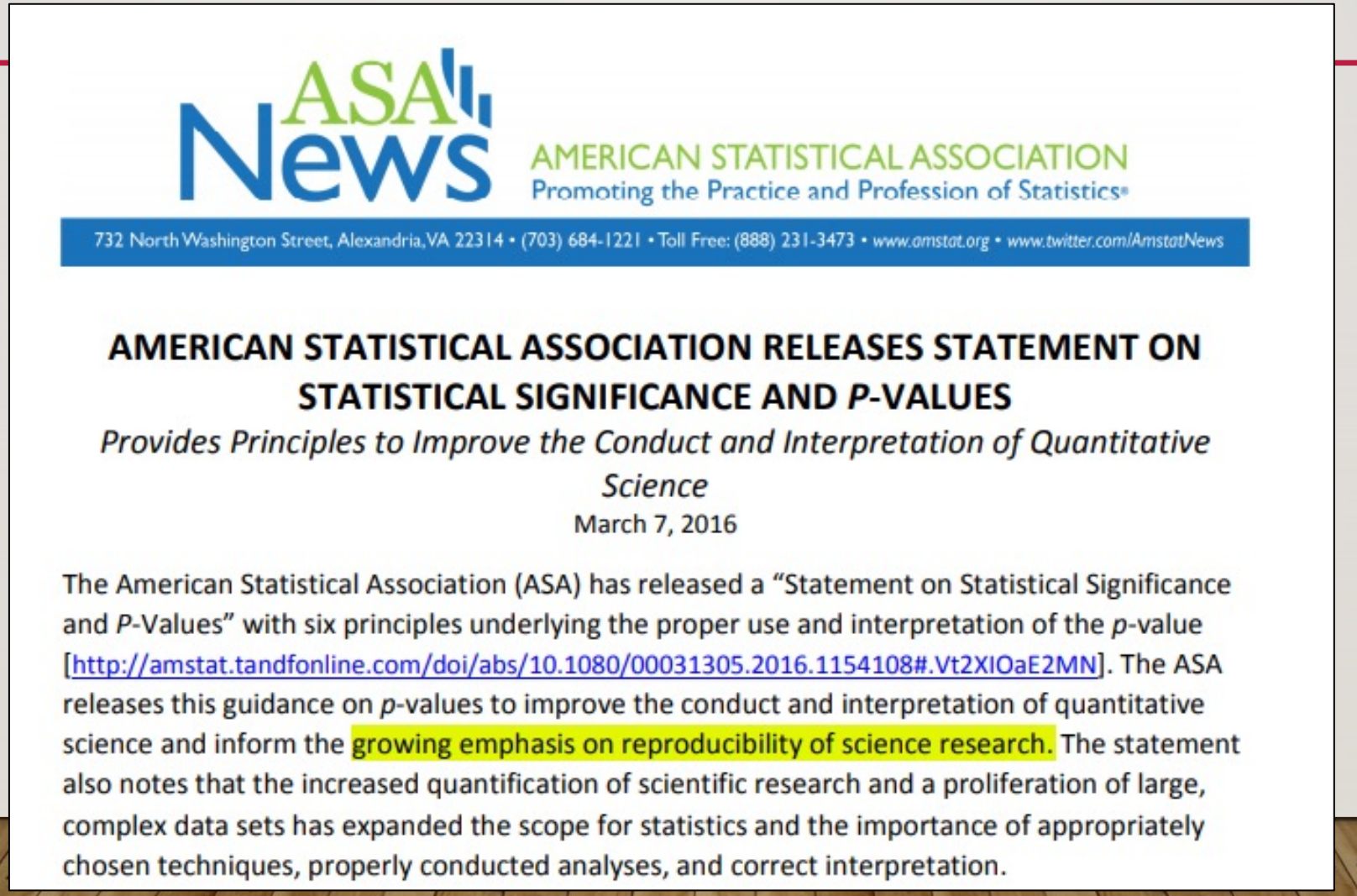
- Study of aspirin for preventing myocardial infarction (MI).
- n = 22,000
- p < .00001. Yeah!!!
- The study was stopped early due to the "conclusive evidence," and aspirin was highly recommended.
- Effect size: 0.77% only
- Further research found even smaller effects, and the recommendation was modified.

LACK REPRODUCIBILITY

- Open Science Collaboration (OSC) (2015) found that a large portion of the replicated results were not as strong as the original reports in terms of significance (*p* values) and magnitude (effect sizes).
- 97% of the original studies reported significant results ($p < .05$), but only 36% of the replicated studies yielded significant findings.
- The average effect size of the replicated studies was only half of the initial studies (ES = 0.197 vs. ES = 0.403).



LACK REPRODUCIBILITY



ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

COUNT ON THEORETICAL DISTRIBUTION

- The test statistics is compared against a "known" distribution but indeed we don't know!
- What researchers did is to **make inferences from the population to the sample,** but not the sample to the population.
- Tukey (1986) questioned our reliance on supposedly known distributions; he asserted that progress of statistics can only be made when we move away from certainty.

POINT ESTIMATE

- Unlike confidence interval (CI) that indicates a possible range of the population parameter, hypothesis testing yields a point estimate.
- Based on the conviction that in the population there is one and only one fixed constant that can represent the **true** parameter.
- This belief is challenged by **Bayesians** because in reality the population body is ever changing and thus there is no such thing as a fixed parameter.

ARBITRARY CUTOFF

- Nothing "magical" about .05.
- Fisher (1956) stated, "No scientific worker has a fixed level of significance from year to year, and in all circumstances, he rejects hypothesis; he rather gives his mind to each particular case in the light of his evidence of ideas" (p.41).
- Rosnow and Rosenthal (1989) stated his objection against blindly using .05 in a humorous way: "God loves the .06 nearly as much as .05" (p.1277).

ARBITRARY CUTOFF

- This cut-off leads to a mechanistic and dichotomous/binary conclusion (Yes or no? Let me go).
- Defense: "Even if the finding is a distribution, like what Bayesians offer, eventually you need to make a dichotomous decision. For example, either you adopt the policy recommendation or not."
- True! In most cases we make a Y/N decision. The problem is: Not only is the decision dichotomous, but also the evidence is dichotomous (is it below or above .05?)

INCAPABLE OF PERFORMING BIG DATA ANALYTICS

- Experts on data science predict that the size of digital data will double every two years; this indicates a 50-fold growth from 2010 to 2020 (Ffoulkes, 2017).
- Traditional statistical procedures were designed for small-sample studies only. If the *n* is too large, virtually any trivial effect would be mistakenly detected as significance.

LETTER

doi:10.1038/nature11421

A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

Human behaviour is thought to spread through face-to-face social networks, but it is difficult to identify social influence effects in observational studies[9–13], and it is unknown whether online social networks operate in the same way[14–19]. Here we report results from a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during the 2010 US congressional elections. The results show that the messages directly influenced political self-expression, information seeking and real-world voting behaviour of millions of people. Furthermore, the messages not only influenced the users who received them but also the users' friends, and friends of friends. The effect of social transmission on real-world voting was greater than the direct effect of the messages themselves, and nearly all the transmission occurred between 'close friends' who were more likely to have a face-to-face relationship. These results suggest that strong ties are instrumental for spreading both online and real-world behaviour in human social networks.

with all users of at least 18 years of age in the United States who accessed the Facebook website on 2 November 2010, the day of the US congressional elections. Users were randomly assigned to a 'social message' group, an 'informational message' group or a control group. The social message group ($n = 60{,}055{,}176$) was shown a statement at the top of their 'News Feed'. This message encouraged the user to vote, provided a link to find local polling places, showed a clickable button reading 'I Voted', showed a counter indicating how many other Facebook users had previously reported voting, and displayed up to six small randomly selected 'profile pictures' of the user's Facebook friends who had already clicked the I Voted button (Fig. 1). The informational message group ($n = 611{,}044$) was shown the message, poll information, counter and button, but they were not shown any faces of friends. The control group ($n = 613{,}096$) did not receive any message at the top of their News Feed.

The design of the experiment allowed us to assess the impact that the treatments had on three user actions; clicking the I Voted button,

INCAPABLE OF PERFORMING BIG DATA ANALYTICS

- Running t-tests with 61M obs!
- Are the p values meaningful here?

We first analyse direct effects. We cannot compare the treatment groups with the control group to assess the effect of the treatment on self-expression and information seeking, because the control group did not have the option to click an I Voted button or click on a polling-place link. However, we can compare the proportion of users between the two treatment groups to estimate the causal effect of seeing the faces of friends who have identified themselves as voters (Fig. 1). Users who received the social message were 2.08% (s.e.m., 0.05%; *t*-test, $P < 0.01$) more likely to click on the I Voted button than those who received the informational message (20.04% in the social message group versus 17.96% in the informational message group). Users who received the social message were also 0.26% (s.e.m., 0.02%; $P < 0.01$) more likely to click the polling-place information link than users who received the informational message (Fig. 1).

Although acts of political self-expression and information seeking are important in their own right, they do not necessarily guarantee that a particular user will actually vote. As such, we also measured the effect that the experimental treatment had on validated voting, through examination of public voting records. The results show that users

INCAPABLE OF PERFORMING BIG DATA ANALYTICS

ordinary. For example, in a now famous experiment on social influence on voting carried out on Facebook (Bond *et al.* 2012), a dataset that was millions of lines long was analysed with the 100-year-old t-test, which had been created primarily to deal with a problem of small sample size (Box 1987). The description of the types of test available that could be used is well beyond the scope of this chapter (readers unfamiliar with the options in this area should consult Agresti and Finlay 2009). In this section, I will simply make some points about their application in the context of very large datasets.

- Bright, J. (2017). Big social science: Doing big data in the social sciences. In Nigel G. Fielding, Raymond M. Lee & Grant Blank (Eds.). *The SAGE handbook of online research methods* (pp. 125-139). Sage. DOI: https://dx.doi.org/10.4135/9781473957992

INCAPABLE OF PERFORMING BIG DATA ANALYTICS

- In 2014 a Ph.D. student at Cornell University and a Facebook employee jointly published a journal article about how media input affected emotion and language use.
- In this study 689,003 Facebook users were randomly assigned into four groups: One group received fewer negative news feed whereas one group received fewer positive news feed. Two control groups had positive or negative news feed randomly deleted.

INCAPABLE OF PERFORMING BIG DATA ANALYTICS

- After the experiment, it was found that "people who had positive content experimentally reduced on their Facebook news feed for one week used more negative words in their status…when news feed negatively was reduced the opposite pattern occurred… Significantly more positive words were used in peoples' status updated."
- This study was a big hit as it was mentioned by 337 news outlets.
- Actually, the percentage of positive words…decreased by 0.1% compared with control, $p < .0001$, Cohen's d = 0.02, whereas the percentage of words that were negative increased by 0.04%, $p = .0007$, d = .0001.

THE BOTTOM LINE IS…

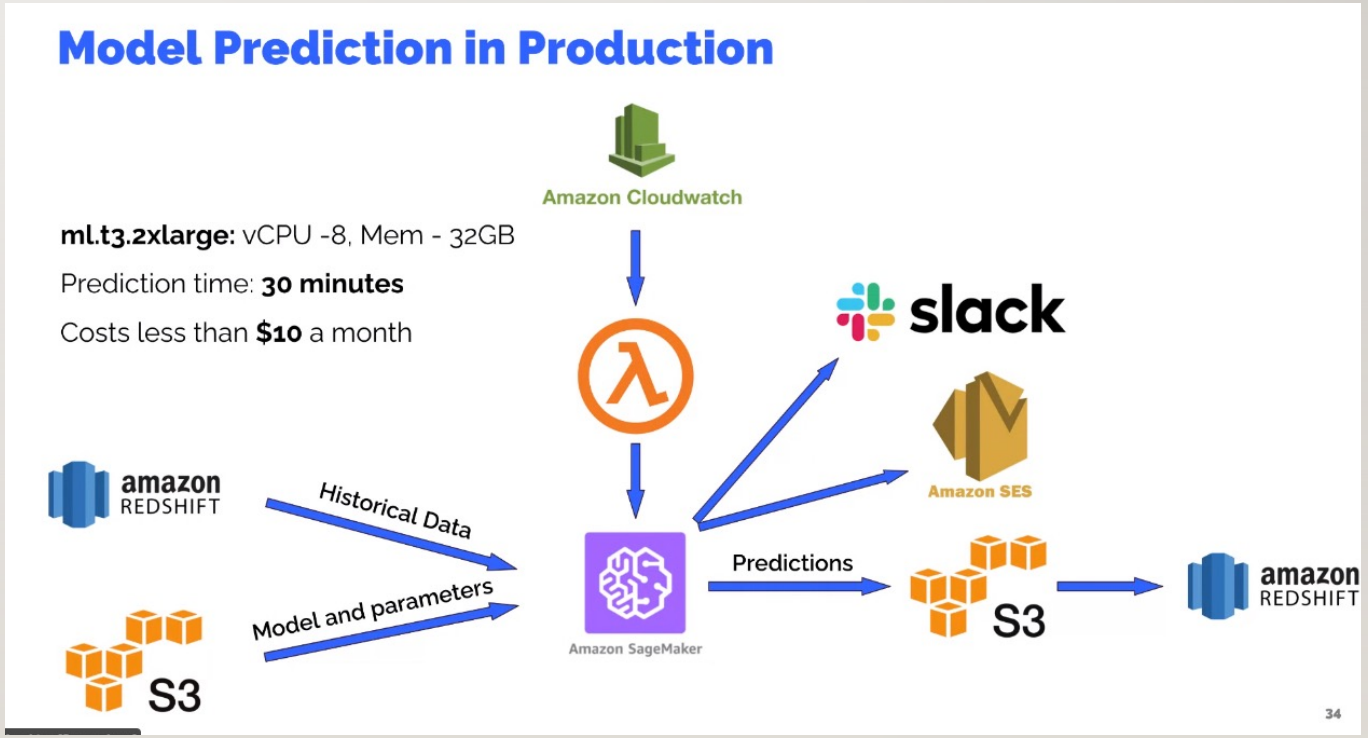
- I do not suggest totally eliminating traditional statistics. We need different tool sets.
- The world is changing. The only constant is change.
- Do we want ourselves to be like Kodak, Blockbuster, and Sears, or do we want to be like Sony, Netflix, and Amazon?

WHAT ARE THE DIFFERENCES?

- The algorithms (e.g., machine learning) developed in the late 20th and early 21st centuries are more efficient than those created in the late 19th and early 20th centuries.
- AI algorithms used for recommendation systems by YouTube, Netflix, Amazon, and Google process trillions of observations in a second, with highly accurate results.
- Even though using a laptop of 16 GB RAM only, data analyst Tavish Stave (2015) can build an entire model in less than 30 minutes based on data sets of millions of observations with thousands of parameters.
- If the analyst were to use traditional statistical models, a supercomputer would be required to perform the same task.

BETTER EFFICIENCY AND ACCURACY

- Amazon SageMaker is utilized to detect online fraud. Even though the data size is gigantic, the prediction time is cut down to 30 minutes and the cost is less than $10 a month. Moreover, the predictive accuracy improves 25%, compared with previous models.
- You don't need a supercomputer to run the program. Rather, what it takes is an 8-core CPU and 32GB of RAM!



WHAT ARE THESE BUZZWORDS?

- Data science
- Data mining
- Big data analytics
- Artificial intelligence
- Machine learning/Deep learning
- Business intelligence

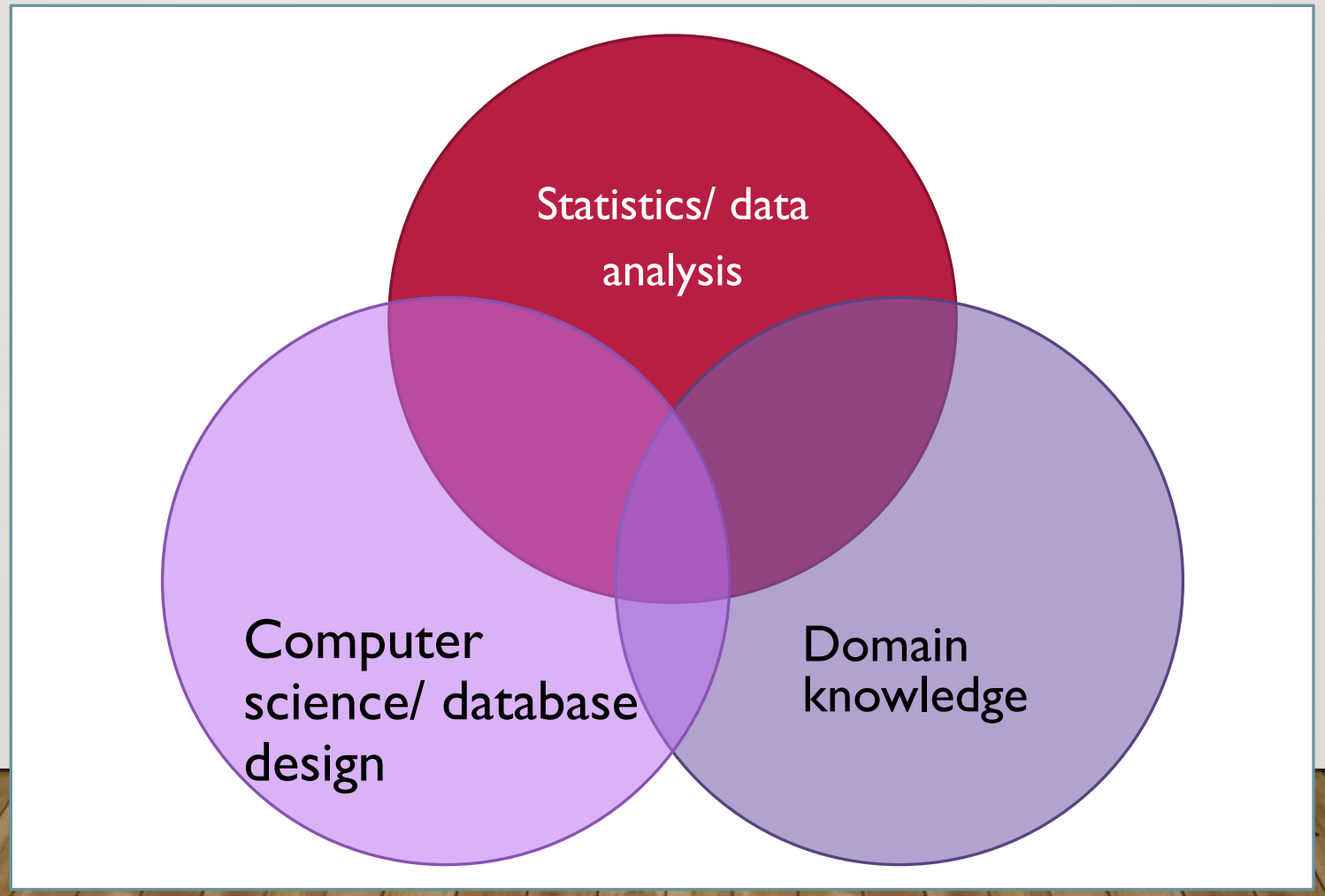
DATA SCIENCE IS THE UMBRELLA TERM

- Data science (DS) is the fusion of statistics/data analysis, computer science, and domain knowledge.
- Data mining analyzes structured data by extracting insight (mining) from a mountain of data. It is a specific application of data science. DS includes both data mining and text mining (for unstructured data).
- Big data analytics deals with big data. DS includes both big, medium, and small data.
- Business intelligence (BI) is the application of data mining in business settings.
- Machine learning, which is based on artificial intelligence (AI) focuses on the design and evaluation of algorithms for extracting patterns from data.

WHAT IS DATA SCIENCE? A SIMPLER APPROACH

- DS consists of 4As (Saltz & Stanton, 2018):
  - Data architecture: Database design
  - Data acquisition: Data collection
  - Data analysis: Machine learning, pattern recognition
  - Data archiving: Make the data reusable.

WHAT IS DATA SCIENCE? THE SIMPLEST APPROACH



THE ROLE OF COMPUTER SCIENCE

In traditional research methods, statistical procedures are incapable of learning from the data.	**Today computer scientists can utilize AI algorithms to make an adaptive system i.e. the model can learn from the new data to improve accuracy.** **For this reason, Google predicts that China might overtake USA in machine learning applications because China, as a more populated nation, has more data.**

THE ROLE OF DATABASE/DATA WAREHOUSE DESIGN

Data are not well-structured and usually data analysts do not receive training in database design.	The data are cleaned, well-structured, and can be accessed through structural query language (SQL). They are ready for reporting and analysis
Data are "isolated" and noncumulative. In a typical IRB application, the researcher states that the data would be destroyed after 7 years in order to protect confidentiality of human subjects.	The system grows as data are appended continuously. Data are retained for long-term trend analyses

BIG DATA ANALYTICS: DEAL WITH BIG DATA

- **High volume:**
  - No definite cut-off, may carry thousands of rows or columns
  - Challenges to data storage, data management, and data analysis.
- **High velocity:**
  - Data stream is ongoing
  - Needs real-time analysis (e.g. credit card fraud detection)
- **High variety:**
  - Contains different types of data (e.g. numbers, texts, images, audio files, video clips…etc.).
  - Challenges to traditional data analysts, who are accustomed to the analysis of structured data.

COMMON GROUND

- Terms
  - Data science, data mining, big data analytics, machine learning, business intelligence
- Common ground
  - Data-driven, not hypothesis-driven
  - Pattern seeking, not cut-off thinking
  - Multiple approaches (e.g. ensemble methods), not a single one.
  - Utilize artificial intelligence, not just human judgment.

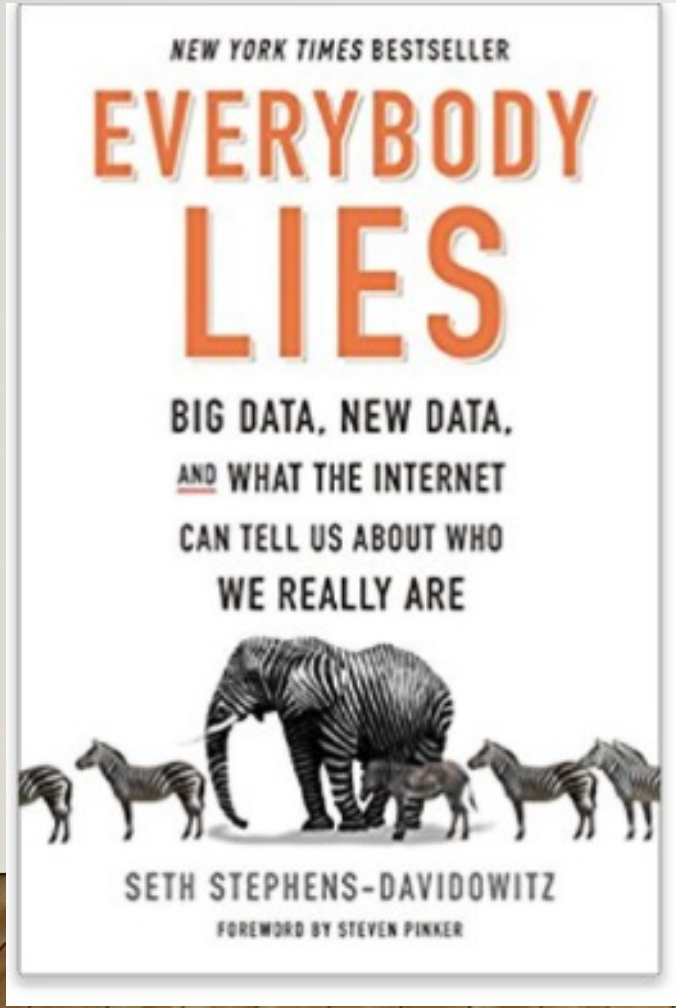
ADVANTAGES OF BIG DATA

- The behavioral data collected in naturalistic settings (e.g. Google, Facebook) may be more accurate than experimental or survey data.
- Opinion polls indicated that more than 40 percent of Americans attend church every week. However, by examining church attendance records, Hadaway and Marlar (2005) concluded that the actual attendance was fewer than 22 percent.

ADVANTAGES OF BIG DATA

- In 2016 election all polls indicate that more voters prefer Clinton to Trump.
- The result is opposite. Why? What happened?
- Many people did not want to say that they support Trump, especially after the Access Hollywood tape was released.

EVERYBODY LIES!



- In survey most voters said that the race of the candidate doesn't matter. Google search data show the otherwise!
- https://www.youtube.com/watch?v=g0m4UQ3frws
- https://trends.google.com/trends/

DICHOTOMOUS THINKING VS. PATTERN RECOGNITION

- Classical method
  - Reject or not to reject the null hypothesis based on an absolute criterion
  - Alpha = .05
  - P-hacking

- New method
  - Pattern recognition
  - The decision is not based on an absolute cut-off
  - AIC, BIC for model comparison

CONFIRMATION VS. EXPLORATION

- Classical method
  - Start with a pre-formulated hypothesis or a strong theory/model
  - Collect data to confirm the hypothesis, theory, or model.

- New method
  - Data-driven
  - No strong theory or hypothesis
  - May examine 100-1000 potential predictors.

SINGLE METHOD VS. TRIANGULATION

- Classical method
  - Use a single statistical procedure e.g. ANOVA, regression, Chi-square…etc,
  - No variants of OLS regression, ANOVA…etc.

- New method
  - Triangulation
  - Run multiple models for comparison and screening e.g. SVM, random forest, gradient boosting, neural networks…etc.

SINGLE SAMPLE VS. CROSS-VALIDATION

- Classical method
  - Use all observations in the sample.
  - Overfitting: good for one sample but not others.
  - Replication crisis

- New method
  - Partition the data into subsets.
  - Repeat the same analysis across all sub-samples
  - Avoid overfitting

SMALL DATA VS. BIG DATA

- Classical method
  - Small sample size
  - Self-report data
  - Structured data (e.g. numeric)

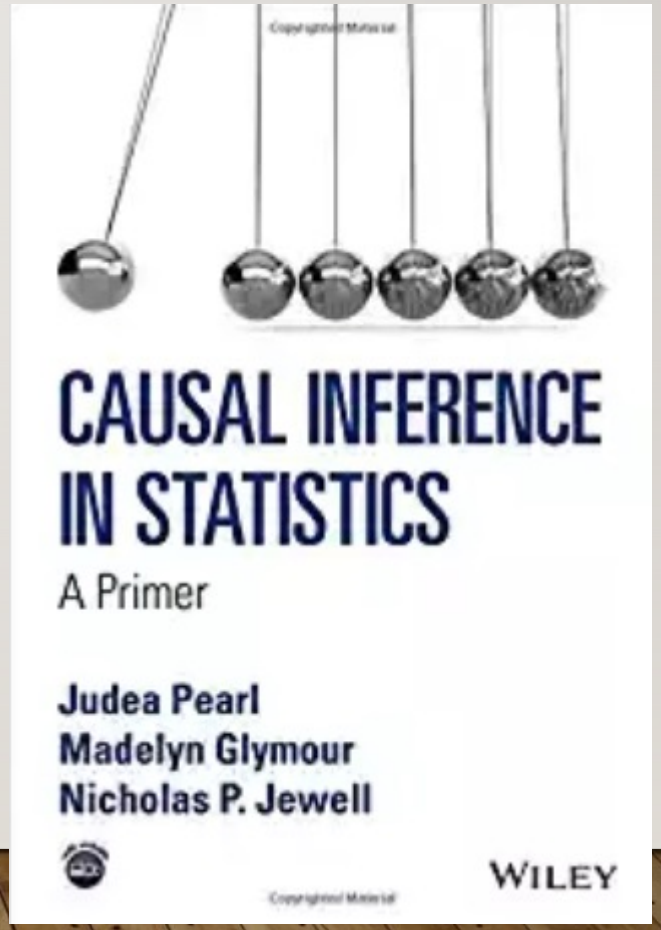
- New method
  - Big data (e.g. Google)
  - Behavioral data
  - Structured, semi-structured, and unstructured data (e.g. blogs, social media)

WHAT ARE THEIR COMMON GROUNDS?

- LASSO, Ridge, elastic net in penalized or generalized Regression are built on the framework of regression.
- The criterion of LogWorth statistics in partition tree is based on the p value.
- Random forest is an extension of bootstraping in resampling.
- Cross-validation is also a concept of resampling.
- Cluster Analysis and PCA are inherited from traditional multivariate statistical analysis.
- Bottom line: They are not totally contradictory! There could be a natural and logical transition!

CAUSAL INFERENCE

- DS focuses on prediction while traditional state emphasizes inference and explanation.
- You need the right data, not necessarily big data.
- Judea Pearl: Data are dumb…you are smarter than the data, especially in causal inference.
- Big data analytics is not the answer to causal inferences.



SMALL DATA AND NESTED DATA

- Not all research problems need powerful tools. e.g., Using neural networks on a data set with 50 cases and three variables is overkill.
- For hierarchical, nested data, multi-level models (e.g., HLM) are preferable to handle the data structure properly.