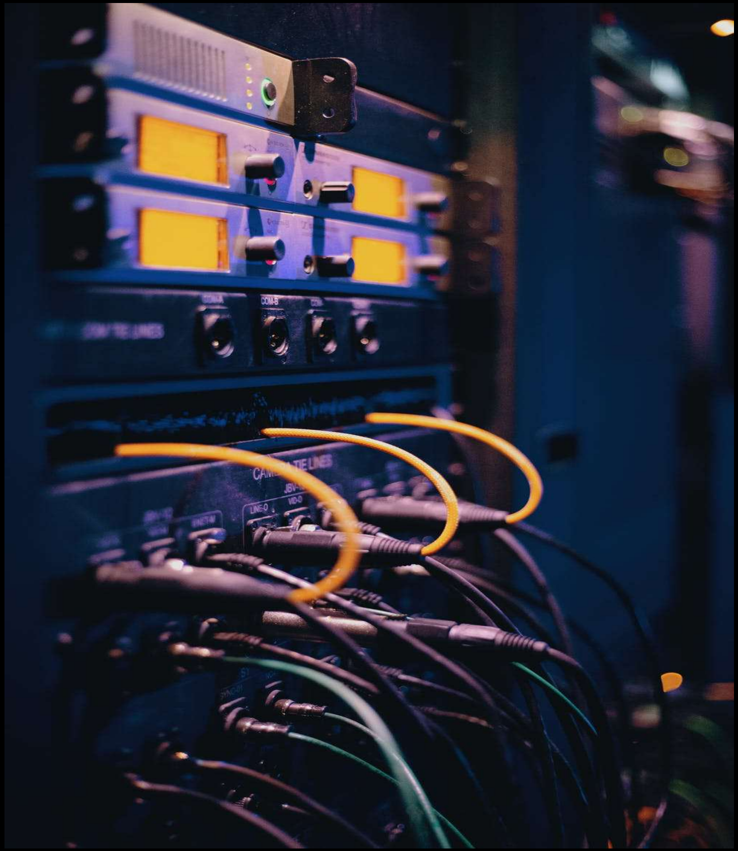
The Present in Data Science and Big Data

JEFF GILL
Distinguished Professor
Department of Government, Department of Mathematics & Statistics
Member, Center for Neuroscience and Behavior
Founding Director, Center for Data Science
American University

https://github.com/jgill22/EIM

Macro Data Forces



- We live in the data century, *whether we like it or not*

- Our personal lives, our careers, our finances, our social activities, our children's' lives, and our future prospects are all intertwined and affected by data collection, data storage, and data analysis by others (humans and machines), *whether we like it or not*

- Governments have mostly lost control over this process, *whether we like it or not*

- Personal education in data science, big data, statistical analysis, and data privacy is essential for people to exert some control and influence over their data future, *whether we like it or not*

Perspectives on Human History

- Homo sapiens are only about 200,000 years old, whereas the earth is 4.54 billion years old

- Humans now have more time to "do stuff" since 30+ years were added to average life expectancy in the 20th century

- We are now in the early-middle part of the fifth major revolution in human history: the Upper Paleolithic revolution (about 40,000 years ago) ⟶ the first agricultural/Neolithic revolution (about 12,000 years ago) ⟶ the second agricultural revolution (18th century) ⟶ the industrial revolution (1712 to early 20th century) ⟶ the information revolution (early 21st century onwards) ⟶ ????

- But people are typically not aware of being in a current ongoing revolution

- We are changing our environments, structures, institutions, and work-lives faster than ever before

Macro Technical and Social Forces

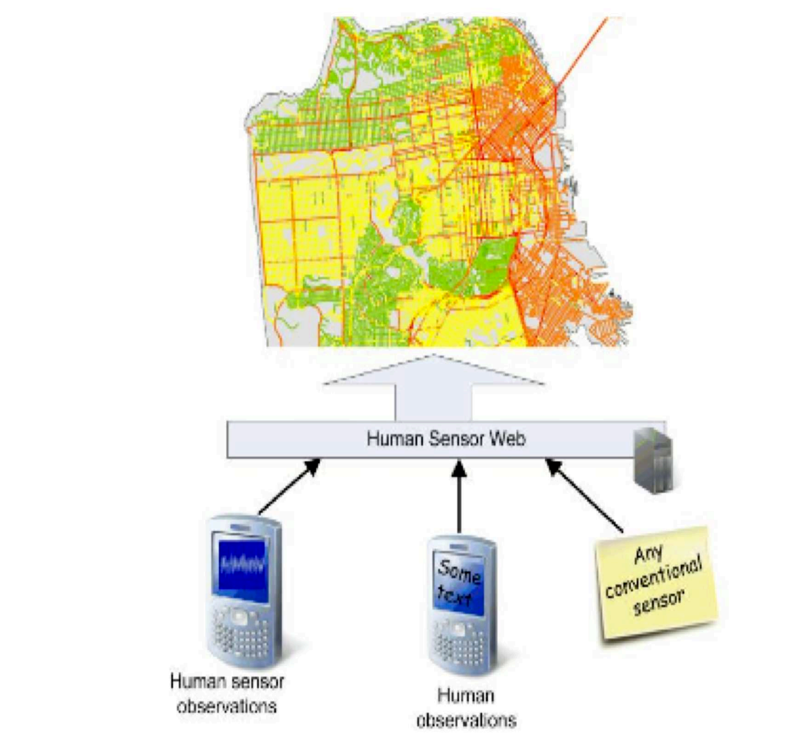
- The rest of the 21st Century will be the era of monumental intellectual progress in the social and biomedical sciences

- The key to research will be: digital computation, data analysis, infrastructure supporting the entire life-cycle of collecting and processing gigantic amounts of information, and the use of networked connections of information from diverse sources

- Data access and data analysis will play an indispensable part in progress to understand social, psychological, and physiological characteristics of what it means to be human

- Integration of disparate data resources will be essential to research and commercialization

- Long term preservation of data involves technical challenges and new business models



"A Change Is Gonna Come," Sam Cooke (1964)

- The future of the social and biomedical sciences data is not going be strictly in rectangular data files, data dictionaries, and PDF codebooks

- These corresponding fields are moving to new and diverse data-types: genetic/genomic, digital video, geocoding/GIS, high-resolution still imaging, high-frequency sensor data, Internet traffic, mobile phone tracing, detailed personal information, and unstructured text

- These fields are moving to new sources of data: social networking and media, human physically generated, government administrative records, transactional financial information, and electronic human monitoring data

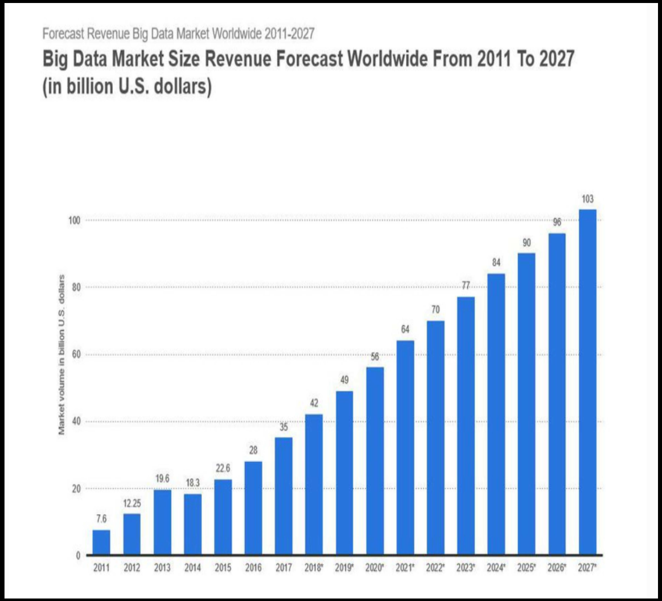
- Note that these are both qualitative and quantitative forms

- Such data require completely new documentation and archiving standards

- There are important privacy/confidentiality, anonymity, government, civil law, and regulatory issues

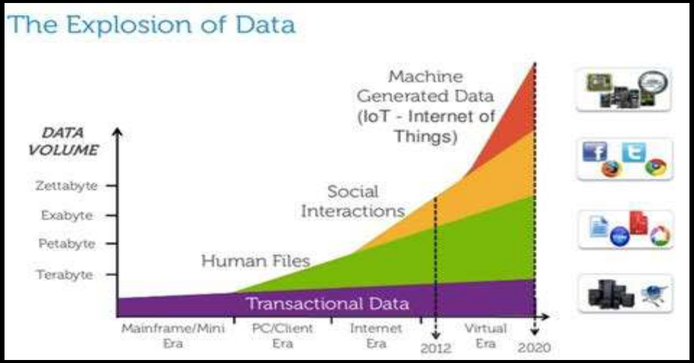
Data by the Numbers: *Every Single Day*. . .

- 23 billion text messages are sent
- 5.5 billion searches are made (64,000 per second on Google alone)
- 500 million tweets are sent
- 333 billion emails are sent
- 4 petabytes of data are created on Facebook, including 49 million GIFs
- 4 terabytes of data are created from each connected car



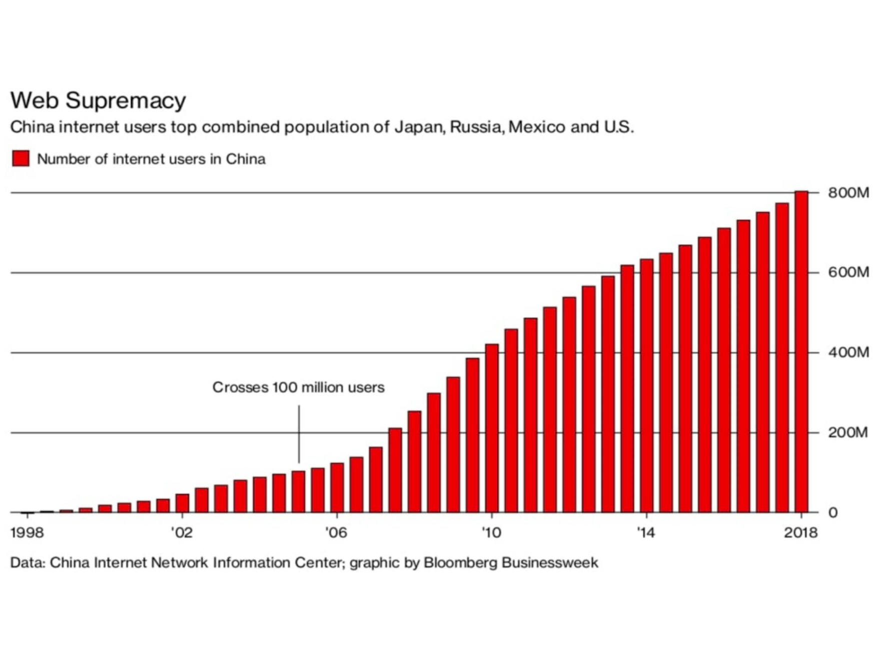
Data by the Numbers: *Every Single Day*. . .

- 65 billion messages are sent on WhatsApp
- 360 terabytes are uploaded to YouTube
- 4 terabytes of data per hour produced from autos
- More than 100 billion messages sent on Whatsapp
- Every minute 100 hours of video are uploaded to YouTube, equaling 0.023 Petabytes per day



Data by the Numbers: *Every Single Day*...

- 21.6 million GIFs are sent via Facebook messenger
- 282 billion spam emails are sent
- 222 million calls placed on Skype
- Venmo processes \$75M peer-to-peer transactions
- The Weather Channel receives 4×10^{10} forecast requests
- 65M Uber bookings
- The average online person generates 10^{18} bytes of data
- The CERN Large Hadron Collider generates 864 zettabytes of data



Data: China Internet Network Information Center; graphic by Bloomberg Businessweek

Data by the Numbers: Scale...

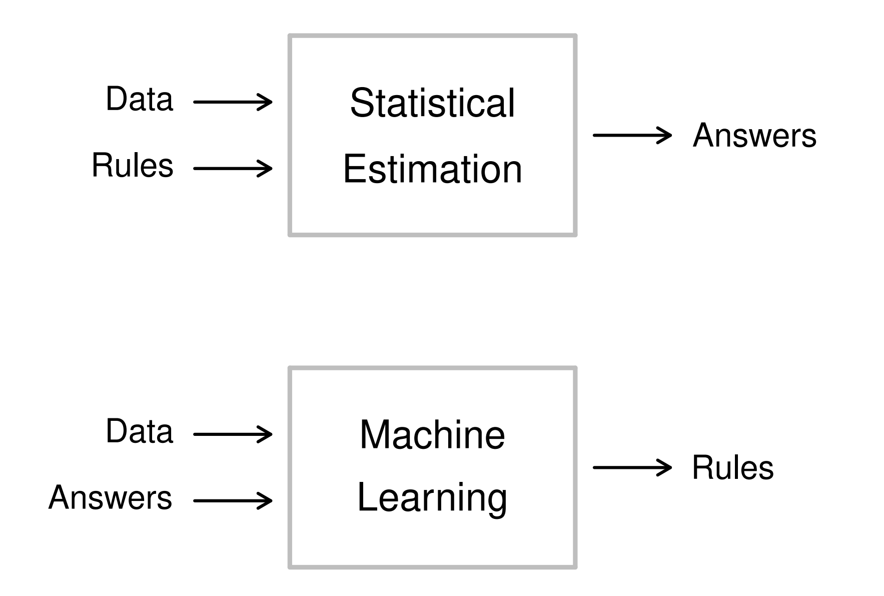
Abbrev.	Unit	Value	Byte Size
b	bit	$0/1$	$1/8$ of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	$1,000$ bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes
BB	brontobyte	$1,000^9$ bytes	1,000,000,000,000,000,000,000,000,000 bytes
gB	geopbyte	$1,000^{10}$ bytes	1,000,000,000,000,000,000,000,000,000,000 bytes
ZB	zotzabyte	$1,000^{11}$ bytes	1,000,000,000,000,000,000,000,000,000,000,000 bytes
CB	chamsbyte	$1,000^{12}$ bytes	1,000,000,000,000,000,000,000,000,000,000,000,000 bytes

What Is *Big Data*

- Basically what anyone wants it to be
- Classic definition: volume, variety, velocity, value, and veracity
- My definition: large enough to challenge available computational resources
- By this definition self-aware humans have always been in a "big data era"
- The current digital universe stored is at least 44 zettabytes ($1,000^7$)
- Sometime before 2025 463 exabytes ($1,000^6$ bytes) of stored data will be created every day
- So what are some tools to deal with data-size challenges?

Relatedly, What is Machine Learning?

- One answer is that it is a simple classifier
- It is actually just statistics with an emphasis on prediction and accuracy
- Basically five tools: Random Forests, Support Vector Machines, Neural Networks (in countless variations now, where the name comes from resembling how the neuro-cranial system works), and Regularization (LASSOs, elastic nets, ridge,. . .) Logit(!)
- ML is most effective when automated with *many* hopefully reliable examples to adapt to tasks independently, which is not how social scientists typically use it due to data limitations



What is Deep Learning

- One view: AI $\supset$ Machine Learning $\supset$ Deep Learning
- Deep learning algorithms establish initial parameters from the data and then train the computer to learn independently by recognizing data patterns using multiple layers of processing.
- These multiple layers can be in the single digits or the millions and each is a form of a neural network that are connected together and jointly estimated with "backpropogation"
- The goal is to establish an optimal set of weights for each connection between each layer in total
- Using a training dataset the key is minimizing the classification difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$.
- Achievements: near-human image classification, near-human speech transcription, near-human handwriting transcription, high quality text to speech, successful commercialization (Assistant and Alexa), autonomous driving, better search results, superhuman GO competing.

Privacy (or lack thereof)

- The explosion of digital sensors, Internet of Things (IoT), smartphone apps, has serious and long-lasting consequences

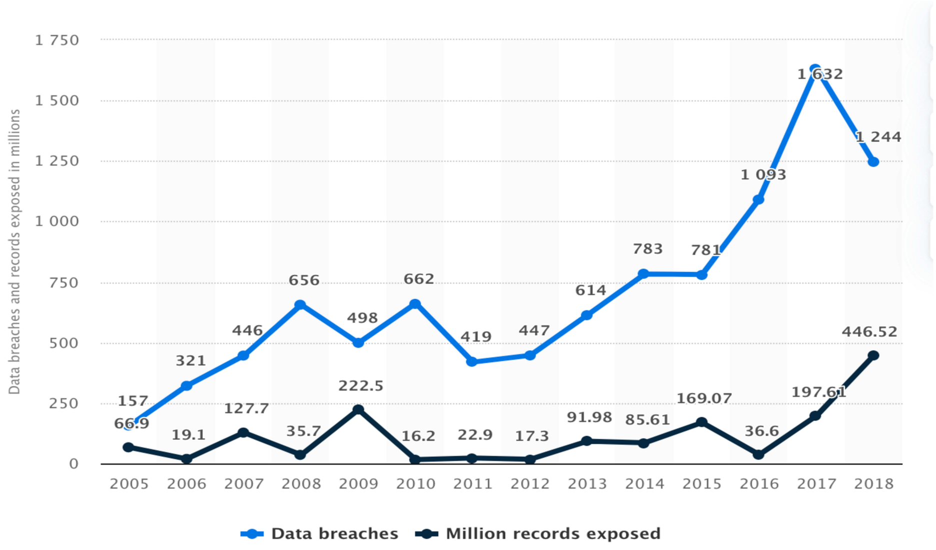
- Alexa is spying on you. Google is spying on you. The US government is spying on you (fingerprinting, etc.). Your phone is spying on you. If your car is recently manufactured it is spying on you. Your rental car company is spying on you. Your hotel is spying on you. Airbnb hosts are spying on you. And even more organizations are spying on you!

- For example, every time Amazon's Alexa AI activates on your wake command it keeps a recording of everything said in the room during operation "to improve our algorithms" (read the fine print sometime; it's scary)

- Substantially reduced costs for storage drives means that corporations and governments save more process traces, network logs, domain specific data, and geospatial data than ever before

Privacy (or lack thereof)

- This means that machine learning algorithms (generally speaking) can associate individual data across disparate data sources to search for particular behavior
- NYT Magazine article series the week of December 16 showed how we are all tracked by our phones and these go into commercial and government databases forever
- At right: annual number of data breaches and exposed personal records in the US, 2005-2018



Data Science for Global Mischief

- I will not comment much on this since everybody here reads the news
- Except to say that it is naïve to believe that there are governments who do *not* practice it
- And never mind the tens of thousands of non-governmental nefarious organizations involved
- This is where it is unfortunate that most data science tools are free or easily purchased

Specific Trends to Pay Attention To

- Blockchain. A highly secured ledger that tracks and archives P2P transactions including bitcoin, but is also widely used by the US government and others
- Regulatory Issues. These are highly mixed from the European General Data Protection Regulation (GDPR), to a seemingly lax US approach
- AI and Intelligent/Invasive Apps. They know more about you than you know
- Augmented reality (AR) and virtual reality (VR). More than just about games
- Edge Computing. IoT that watches you all the time
- Usage. Less than 1% of all generated and stored data are being analyzed and this number is actually going *down*
- Commercialization. The big data analytics market is currently worth over $500B in business (this is probably a low estimate)
- Ethics. Big data and big data analytics (AI, etc.) provide governments, corporations, and others with powerful tools to harm people in different ways

Is Data Science a Field?

- Yes! The parents: statistics, machine learning (CS), mathematics, and the social sciences

- The last one is the most important because the huge majority of data science work is done to understand *people*, socially, politically, biomedically, and commercially

- Yet there is a shortage of data scientists in academia, government, and industry

- Recent (and typical) ad:

  > *1. Data Scientist*
  >
  > *Median Base Salary: $130,000*
  >
  > *Job Openings (YoY Growth): 4,000+ (56%)*
  >
  > *Career Advancement Score (out of 10): 9*
  >
  > *Required Skills: Data Science, Data Mining, Data Analysis, R, Python, Machine Learning*

- The *Harvard Business Review* named Data Science "the sexist job of the 21st century" in 2012.

Is Data Science a Field?

- A recent University of Wisconsin study: the average *starting* salary for MS DS degrees is $90K
- US Bureau of Labor Statistics: "Employment of data scientists is projected to grow 36 percent from 2021 to 2031, much faster than the average for all occupations."
- There were about 30M job ads for data scientists in the US alone in 2020 according to IBM
- The recruiter Glassdoor evaluated the 50 in 2022 best jobs in America that pay over $100,000 and Data Scientist is ranked No. 3 (always in the top 5 over the last 15 years)
- There were about 30M job ads for data scientists in the US alone in 2020 according to IBM
- A strong trend exists now for social science PhDs with technical training in data science to go to industry. Where? Why?

How the Data Century Affects Us, General Research

- Interesting and important forms of social science data are bigger and more complex than ever in the way that I have described and in additional ways
- We now have more analytical tools than ever, with huge progress in *qualitative* analysis
- But we need more!
- And yet social science departments do not typically have the large and expensive infrastructure for existing and future big data challenges
- Does this increase the Gini Index of social science researcher resources? I think so
- This is where an insightful Center for Data Science in a university can be most helpful
- The role of such a center is going is critical to university success in the 21st century

Anecdotal Story

- I was asked to answer two questions for this workshop ("An Introduction to Machine Learning and Big Data") here at the Universidad Católica del Uruguay...

- Why is this course important for academics?

  This course is very important to researchers who use empirical data analysis in their research in the 21st century. Data science problems in academia now often involve large data sets which provide challenges related to variable selection, clustering among a large number of cases, missing data issues, and prediction classification. New tools in this area such as machine learning algorithms, neural networks, nonparametric clustering, penalized regression, imputation methods, and more will be covered.

- Why is it important for the labor market?

  This course provides the most important set of skills available today. There is no more valued expertise in the global labor market than machine learning and big data analysis, and these are in high demand by corporations, government, and academia. The labor market for data scientists in every modern country in the world exceeds the number of job candidates.

How the Data Century Affects Us, Journal Scholarship

- The conventional model of journal publishing is becoming increasing outdated in this age of rapid knowledge transfer

- Academic journals were created in the 17th century to decrease the time of dissemination of knowledge since books at the time took a very long time to be physically printed and bound

- There is a pressing need to get new knowledge out in social science and a journal review time-span that can take well over a year from submission to publication belongs in the Triassic Era

- The traditional journal model where we give commercial entities product for free so that they can sell it back to our university libraries is increasingly obsolete, save for tenure/promotion time

- So the state of scholarly publishing is about to change fundamentally, and already has, `arXiv`, etc.

- We also live in a time in the social sciences when the *achievement* of a publication often means more than the *actual content* of a publication

How the Data Century Affects Us, Teaching

- The freshman you will be teaching this Fall were born after: the creation of the Internet, ubiquitous sophisticated mobile technology, 9/11, the end of the first Cold War, and the advent of 24 hour constant delivery of the news

- Students sit in the classroom wired into their regular social environment every second of the lecture

- They can immediately fact-check anything you say in class, and yet some of what they will get from that search are not actually "facts"

- They also increasingly want "value" out of the experience in literally the vocational sense

- On the positive side, surveys show that students very much miss the on-campus experience during the pandemic

- Universities are increasing tracking everything that undergraduates do through their phones: when do they attend class, when are they in their dormitories, where do they go off campus, when they visit the campus health clinic, when do they eat, and more (Orwell was an unimaginative by comparison)

How the Data Century Affects Governments, Research and Development

- Interesting and important forms of human-centered data are bigger and more complex than ever in the way that I have described and in additional ways
- We now have more analytical tools than ever, with huge progress in *qualitative* analysis
- But we need more!
- And yet governments do not always have the large and expensive infrastructure for existing and future big data challenges
- Does this increase the Gini Index of government resources? I think so
- It is an obvious fact that governments need to spend significant resources building data science infrastructure
- This is where an insightful Center for Data Science in a university setting can be most helpful as well as degree programs
- The role of such a center is going is critical to governmental success in the 21st century

Ongoing Big Data Challenges for Individual Policy-Makers in Government

Challenge #1: For those with official administrative and political duties there simply is not enough time in the day to stay abreast of the machine learning/AI/big data literatures

Challenge #2: This means that building a team of technical experts (data scientists, computer scientists, statisticians, software engineers) and technical managers who understand the important issues is critical in any government agency in any nation in the world

Challenge #3: There simply are not enough of these people being trained in the pipeline

Challenge #4: So inter-agency, inter-government cooperation, and developing relationships with national universities are critical

Ongoing Big Data Challenges for the World

- Often poor understanding and acceptance of general data challenges
- Difficulty in determining data quality in large data streams
- Confusing array of big data technology (hardware, software, transmission, etc.)
- Misuse of readily available, and often free, software tools
- Dangerous security holes and dangerous people
- The process of converting sources into actual insights and results
- Communication of results, including measures of uncertainty, to general audiences

These challenges require big steps
forward in human-machine interaction

Models: The Necessity of Simplification

- We do not learn without simplification of natural phenomenon
- Every model is a simplification/approximation and thus actually *wrong* Therefore models are never "true," but good ones extract important features
- KKV (p.43):

> . . . the difference between the amount of complexity in the world and that in the thickest of descriptions is till vastly larger than the difference between this thickest of descriptions and the most abstract quantitative or formal analysis

On Models

- Formal/Mathematical Model: a mathematical and logical construct.
- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

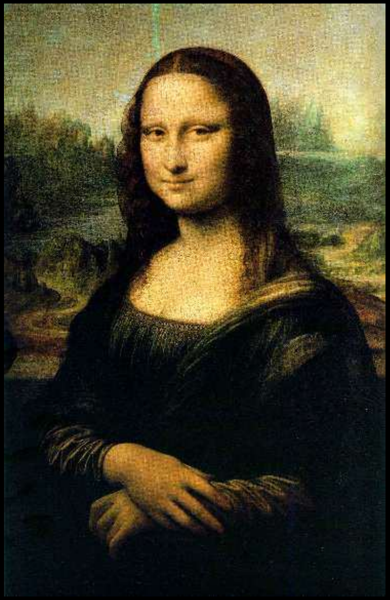
- Two models of humans...

On Models

- Formal/Mathematical Model: a mathematical and logical construct.
- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

- Two models of humans...

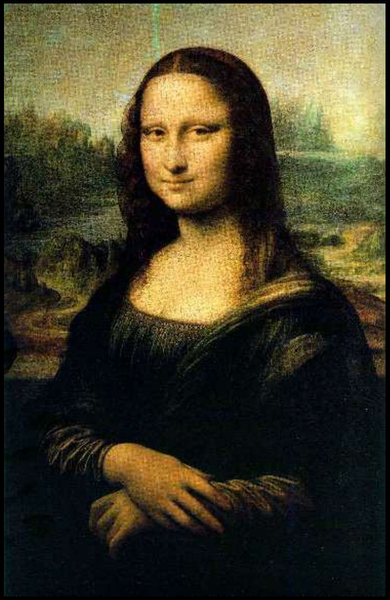


On Models

- Formal/Mathematical Model: a mathematical and logical construct.
- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

- Two models of humans...



Models: Scope and Assumptions

- Advantages of restrictive models:
  - clear
  - parsimonious, easy to understand and explain
  - abstract
- Advantages of non-restrictive models:
  - detailed
  - contextual
  - realistic

Models: Characteristics

- Model: a necessarily unrealistic picture of nature, a formal representation and simplification using symbology and assumptions

- Characteristics of quantitative models:
  - looking at underlying trends and principles
  - usually symbolic and abstract
  - note: the quantification process produces precision but not necessarily accuracy since there is always measurement error.

- Characteristics of qualitative models:
  - good at seeing causality, but often not generalizable
  - complements description
  - provides nuance and detail otherwise unobservable.

Models: Some Definitions

- Descriptive Model: a narrative simplification describing key causal factors

- Statistical Model: has a systematic component (replicative) and a non-systematic component (varying)

- Formal Model: a purely mathematical representation of reality with no non-systematic component

- Key Distinction: do we think that it is a probabilistic world (there always exists variation), or a deterministic world (variation is just science attributable to what we have not yet measured)

- Causal Inference is concerned with what would have happened to case i's outcome variable, y_i, if it had received a different treatment level

- Causal inference can also be considered as a special case of *prediction* under varying circumstances, only with much stricter assumptions than usual

- A huge amount of work with big data is done to make predictions rather than classical inference

Thoughts On Science

- It turns out that we are the "hard" science (Bob Keohane is wrong).

- Simon DeDeo, a research fellow in applied mathematics and complex systems at the Santa Fe Institute:

  > "In physics, you typically have one kind of data and you know the system really well," said DeDeo. "Now we have this new multimodal data [gleaned] from biological systems and human social systems, and the data is gathered before we even have a hypothesis." The data is there in all its messy, multi-dimensional glory, waiting to be queried, but how does one know which questions to ask when the scientific method has been turned on its head?

  `http://www.wired.com/wiredscience/2013/10/topology-data-sets/`

- New trends: big data, data engineering, biometric measurement, machine learning, sophisticated marketing research, google research, non-random samples from online surveys, text as data.

Terminology

Inference: using sample data and a model to make claims (estimates) about population parameters.

Prediction: using sample data and a model to make claims about future observations.

Statistic: a descriptive measure based on a population or sample data that does not depend on a parameter.

Model: a necessarily unrealistic picture of nature, a formal representation and simplification using symbology and assumptions.

Statistical Taxonomy

- Frequentists: From Neymann/Pearson/Wald setup. An orthodox view that sampling is infinite and decision rules can be sharp.

- Bayesians: From Bayes/Laplace/de Finetti tradition. Unknown quantities are treated probabilistically and the state of the world can always be updated.

- Likelihoodists: From Fisher. Single sample inference based on maximizing the likelihood function and relying on the Birnbaum (1962) Theorem. Bayesians that don't know that they are.

The pseudo-Frequentist NHST is wrong

- A few authors have noted this (just a sample): *Barnett 1973, Berger, Boukai, and Wang 1997, Berger Thomas Sellke 1987, Berkhardt and Schoenfeld 2003, Bernardo 1984, Brandstätter 1999, Carver 1978, 1993, Dar, Serlin and Omar 1994, Cohen 1988, 1994, 1992, 1977, 1962, Denis 2005, Falk and Greenbaum 1995, Gelman, Carlin, Stern, and Rubin 1995, Gigerenzer 1987, 1993, 1998, Gigerenzer and Murray 1987, Gill 1999, 2005, Gliner, Leech and Morgan 2002, Grayson 1998, Greenwald 1975, Greenwald, Gonzalez, Harris and Guthrie 1996, Hager 2000, Howson and Urbach 1993, Hunter 1997, Hunter and Schmidt 1990, Jeffreys 1961, Kirk 1996, Krueger 1999, 2001, Lindsay 1995, Loftus 1991, 1993a, 1993b, 1994, 1996, Loftus and Bamber 1990, Macdonald 1997, Meehl 1967, 1978, 1990, 1978, Nickerson 2000, Oakes 1986, Pollard 1993, Pollard and Richardson 1987, Robinson and Levin 1997, Rosnow and Rosenthal 1989, Rozeboom 1960, 1997, Schmidt 1996, Schmidt and Hunter 1977, Sedlmeier and Gigerenzer 1989, Thompson 2002, Wilkinson 1999.*

- Why?

  1. Artificial Model Selection Criteria
  2. The Arbitrariness of Alpha
  3. Replication Fallacy
  4. Asymmetry and Accepting the Null Hypothesis
  5. Probabilistic Modus Tollens
  6. Inverse Probability Problem

Fisher Test of Significance

1. Identify the null hypothesis.

2. Determine the appropriate test statistic and its distribution under the the assumption that the null hypothesis is true.

3. Calculate the test statistic from the data.

4. Determine the achieved significance level that corresponds to the test statistic using the distribution under the assumption that the null is true.

5. Reject H_0 if the achieved significance level is sufficiently small. Otherwise reach no conclusion.

Neyman-Pearson Hypothesis Testing

1. Identify an hypothesis of interest, Θ_B, and a complementary hypothesis, Θ_A.

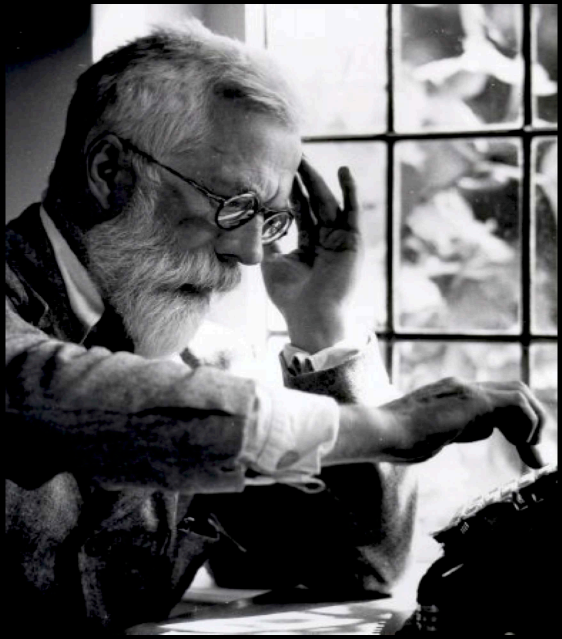
2. Determine the appropriate test statistic and its distribution under the assumption that Θ_A is true.

3. Specify a significance level (α), and determine the corresponding critical value of the test statistic under the assumption that Θ_A is true.

4. Calculate the test statistic from the data.

5. Reject Θ_A and accept Θ_B if the test statistic is further than the critical value from the expected value of the test statistic (calculated under the assumption that Θ_A is true). Otherwise accept Θ_A.

On Fisher



- Fred Hoyle (Astronomer and Mathematician) on Fisher: "So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from extreme urbanity into a boiling cauldron of wrath, you got by with little worse than a thick head from the port which he, like the Cambridge mathematician J.E.Littlewood, loved to drink in the evening."

- Fisher (1955, JRSS-B) on N&P testing: "The phrase 'errors of the second kind' although apparently a harmless piece of technical jargon, is useful as indicating the type of mental confusion in which it was coined. . . "

- Fisher (Annals of Eugenics, 1937): "His example on this point is valuable; whereas he was a clumsy mathematician. Had it not been for his arrogant temper, his taste for numerical example might well have saved him from serious theoretical mistakes."

The Current Paradigm: Null Hypothesis Significance Testing

1. two hypotheses are posited: a null or restricted hypothesis (H_0) which competes with an alternative or research hypothesis (H_1) describing two complementary notions about some phenomenon.

2. the test evaluates a parameter vector: $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, and the null hypothesis places restrictions on some subset ($\ell \leq m$) of the theta vector such as: $\theta_i = k_1\theta_j + k_2$ with constants k_1 and k_2.

3. a test statistic (T), some function of θ and the data, is calculated and compared with its known distribution under the assumption that H_0 is true.

4. the test procedure assigns one of two decisions (D_0, D_1) to all possible values in the sample space of T, which correspond to supporting either H_0 or H_1 respectively.

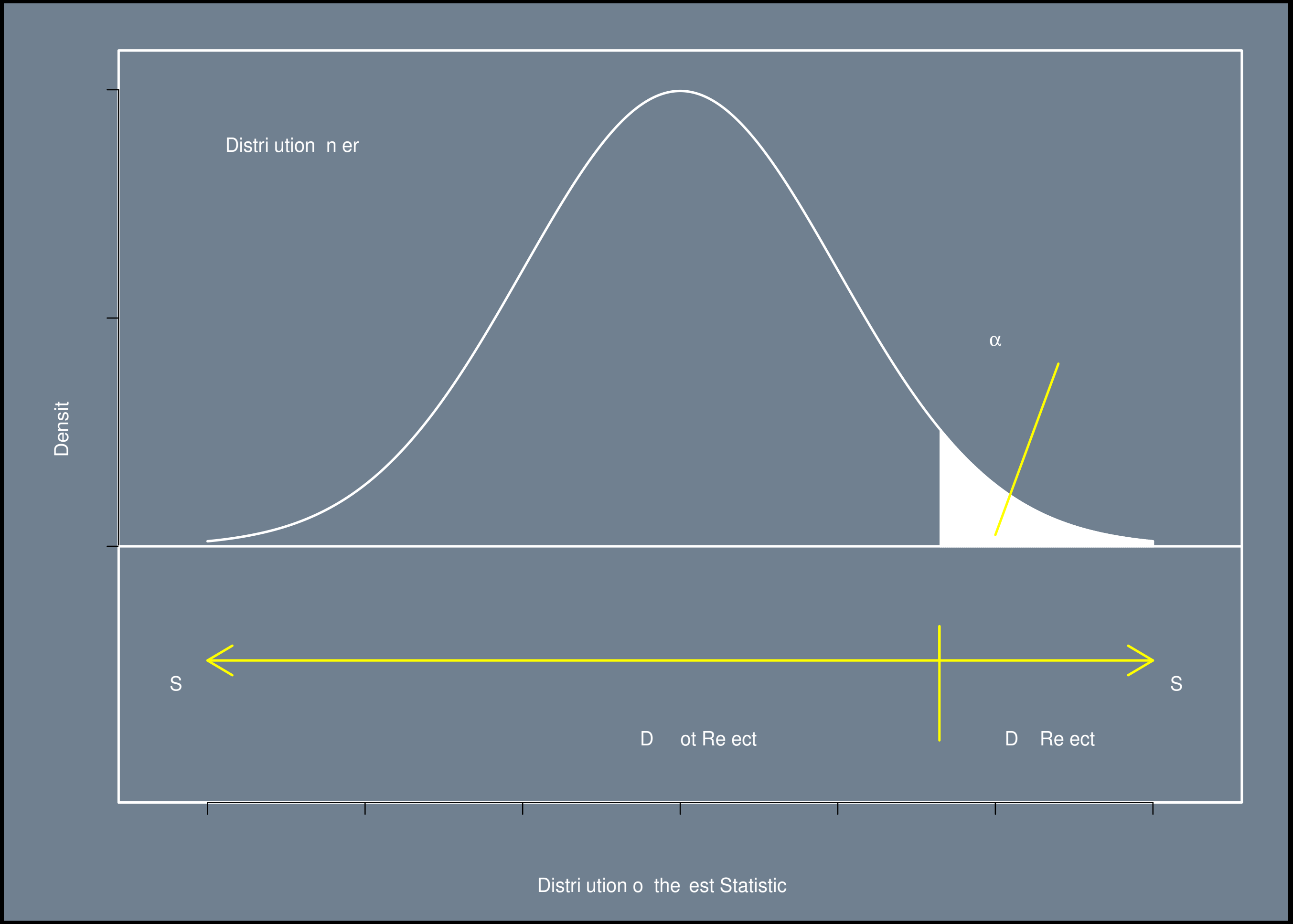
5. the p-value ("associated probability") is equal to the area in the tail (or tails) of the assumed distribution under H_0 which starts at the point designated by the placement of T on the horizontal axis and continues to infinity.

The Current Paradigm: Null Hypothesis Significance Testing

6. the sample space of T is segmented into two complementary regions (S_0, S_1) whereby the probability that T falls in S_1, causing decision D_1, is either a predetermined null hypothesis cumulative distribution function (CDF) level: the probability of getting this or some lower value given a specified parametric form such as normal, F, t, etc. ($\alpha =$ size of the test, Neyman and Pearson), or the cumulative distribution function level corresponding to the value of the test statistic under H_0 is reported (p-value $= \int\limits_{S_1} P_{H_0}(T = t)dt$, Fisher).

7. thus decision D_1 is made if the test statistic is sufficiently atypical given the distribution under H_0. This process is illustrated for a one tail test at $\alpha = 0.05$ in the figure.

The Current Paradigm: Null Hypothesis Significance Testing



Modus Tollens

If A then B	If H_0 is true then the data will follow an expected pattern
Not B observed	The data do not follow the expected pattern
Therefore not A	Therefore H_0 is false.

Probabilistic Modus Tollens

If A then B is highly likely	If H_0 is true then the data are highly likely to follow an expected pattern
Not B observed	The data do not follow the expected pattern
Therefore A is highly unlikely	Therefore H_0 is highly unlikely.

Probabilistic Modus Tollens

If A then B is highly likely	If a person is an American, then it is highly unlikely she is a member of Congress.
Not B observed	The person is a member of Congress
Therefore A is highly unlikely	Therefore it is highly unlikely she is an American.

Misconceptions about Inverse Probability

- The order of conditionality can be really important.

- suspected probability of AIDS in risk group: $P(A) = 0.02$

  probability of correct positive classification: $P(C|A) = 0.95$

  probability of correct negative classification: $P(C^c|A^c) = 0.97$

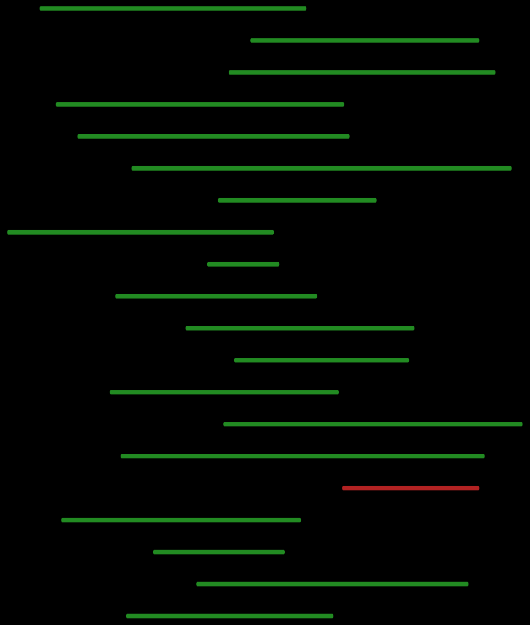
- Suppose we want $P(A|C)$, from:
$$P(A|C) = \frac{P(A)}{P(C)}P(C|A)$$

- Getting the unconditional:
$$\begin{aligned} P(C) &= P(C \cap A) + P(C \cap A^c) \\ &= P(C|A)P(A) + P(C|A^c)]P(A^c) \\ &= P(C|A)P(A) + [1 - P(C^c|A^c)]P(A^c) \\ &= (0.95)(0.02) + (1 - 0.97)(0.98) \cong 0.05 \end{aligned}$$

- So now we can calculate:

$$P(A|C) = \frac{P(A)}{P(C)}P(C|A) = \frac{0.02}{0.05}(0.95) = 0.38$$

Confidence



- Which of these is the correct interpretation of a $(1-\alpha)$ confidence interval?
  - An interval that has a $1-\alpha\%$ chance of containing the true value of the parameter.
  - An interval that over $1-\alpha\%$ of replications contains the true value of the parameter, *on average*.

- What interpretation do people really *want*.

Contrived Ignorance, Buried Assumptions

- Models with uniform priors.
- Normality.
- Correlation coefficient.
- Only two models tested.
- No such thing as specification searches.

Model Selection

1. "Illusion of theory confirmation"
   Leamer (1978): "Believers report the summary statistics from the n^{th} equation* as if the other $n-1$ were not tried, as if the n^{th} equation* defined a controlled experiment.

2. Rozeboom (1960): "the null hypothesis significance test introduces a strong bias in favor of one out of what may be a large number of reasonable alternatives."

3. Raftery (1995): two entirely plausible and statistically significant competing models can lead to substantively different conclusions using the exact same data.

4. The significance levels reported in the final work have different interpretations than the significance levels in intermediate models

5. V independent variables, pick 2^V possible models, for instance: $V = 20$ produces $1,048,576$.

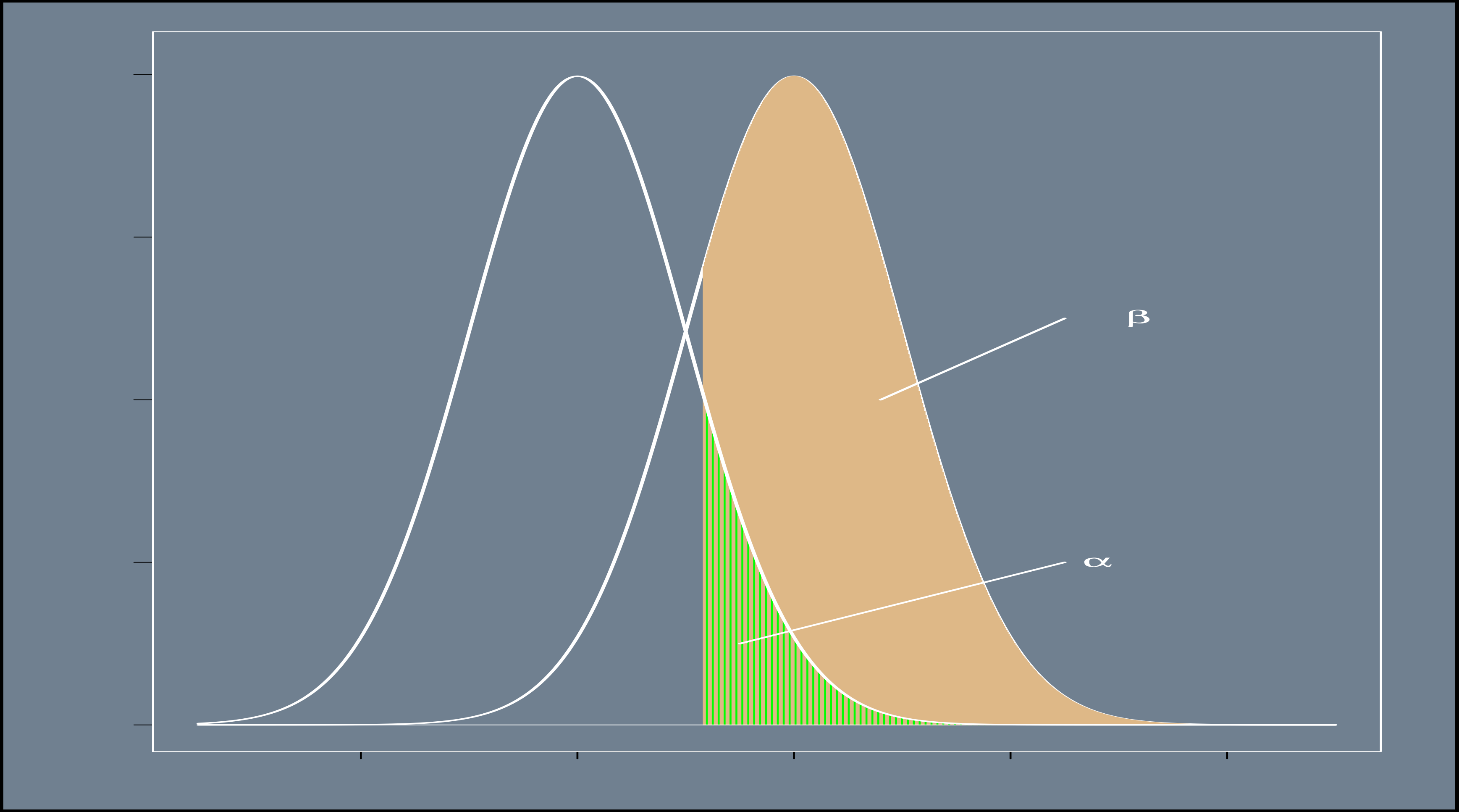
6. lottery paradox

7. file drawer problem

Effect Size

- Generally it is better to design studies to maximize potential effect size, labeled θ.
- This is why many drug studies give the lowest possible level to the control group and the highest safe level to the treatment group.
- Generally the desired minimal effect size is assumed in sample size calculations.
- Given this effect size, what sample size is necessary to get:
  - a specific standard error,
  - a desired significance level,
  - a desired power level.
- *For observational studies we have to live with the sample size that is provided.*

0.8 Power Illustration



Power and Sample Size

- Wilkerson and Olson (1997) asked 52 psychology graduate students to evaluate two tests which report p-values of 0.05 and are identical in every way except that one has a sample size of 25 and the other has a sample size of 250.
- The graduate students were asked which test had the greatest probability of making a Type I error.
- Only 6 out of the 52 correctly observed that the two tests have an identical probability of falsely rejecting the null hypothesis.

The Arbitrariness of Alpha

1. There is absolutely no theoretical justification for the standard significance thresholds.
2. *Level of significance* is therefore arbitrary and subject to custom by disciplines.
3. These values and the subsequent tables come from Fisher, whose work emphasized experimentation.
4. Fisher (1933):

   > "...the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 percent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials."

5. And:

   > "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point) or one in a hundred (the 1 percent point). Personally, the writer prefers to set the low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level."

Plan for the Rest of the Workshop

- Missing data
- Survey of machine learning
- Clustering
- Regularization
- MCMC