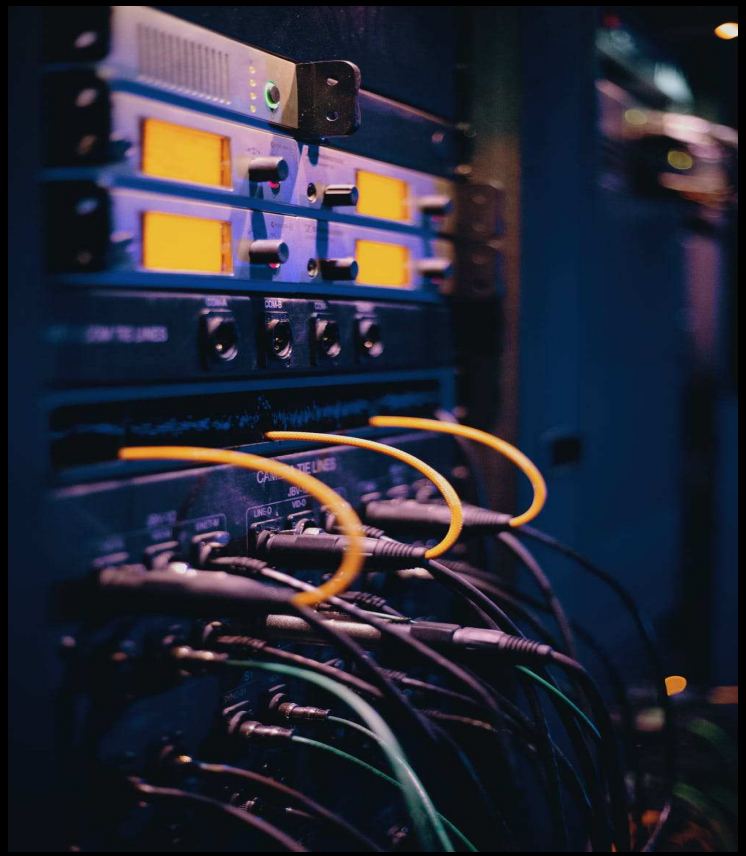
The Present in Data Science and Big Data

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

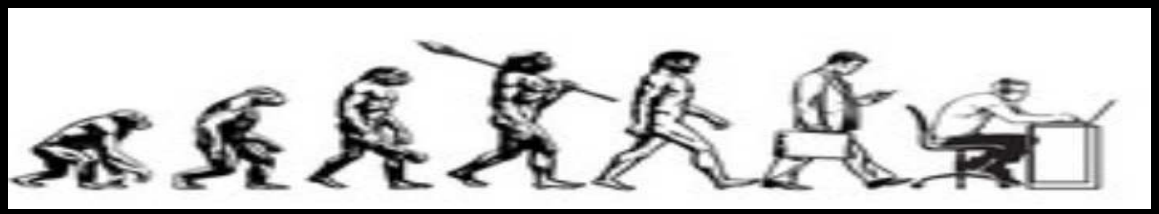
Macro Data Forces



- 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*

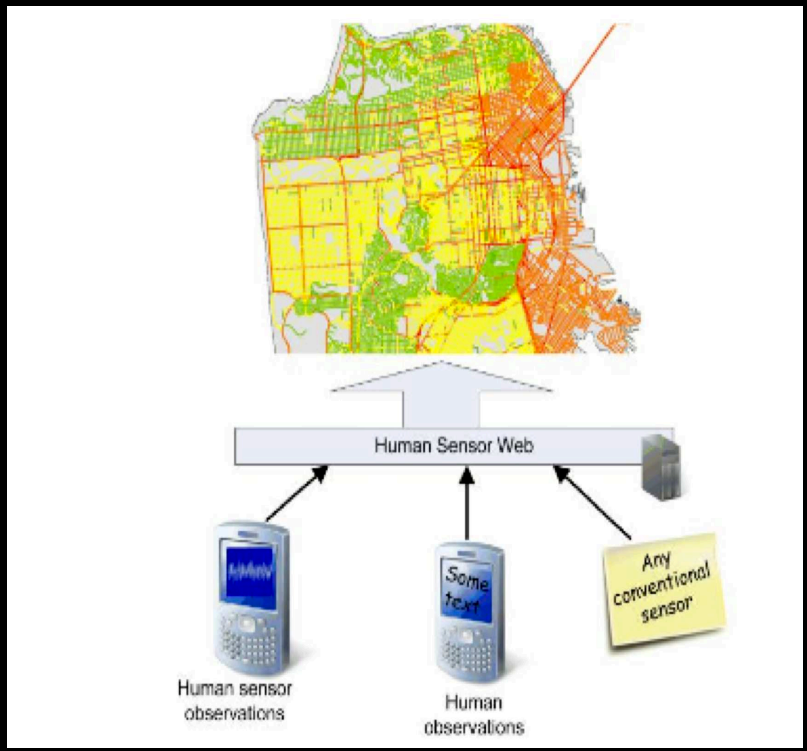
Perspectives on Human History

- 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) $\longrightarrow$ the first agricultural/Neolithic revolution (about 12,000 years ago) $\longrightarrow$ the second agricultural revolution (18th century) $\longrightarrow$ the industrial revolution (1712 to early 20th century) $\longrightarrow$ the information revolution (early 21st century onwards) $\longrightarrow$????
- 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

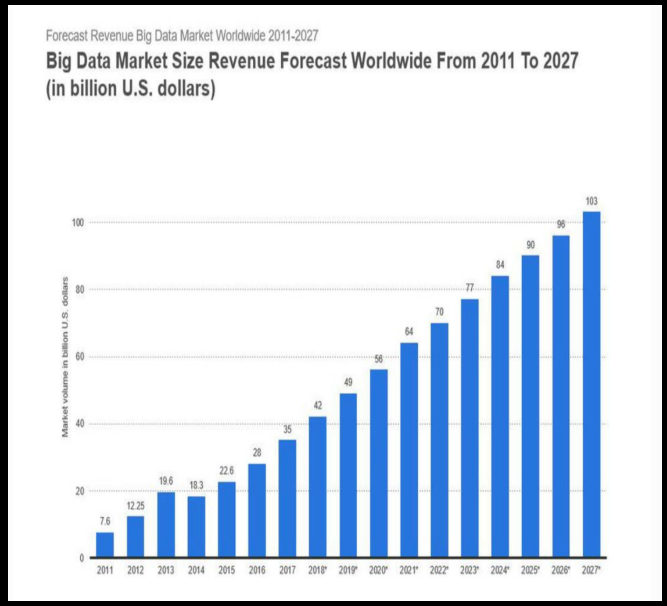


"A Change Is Gonna Come," Sam Cooke (1964)

- 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

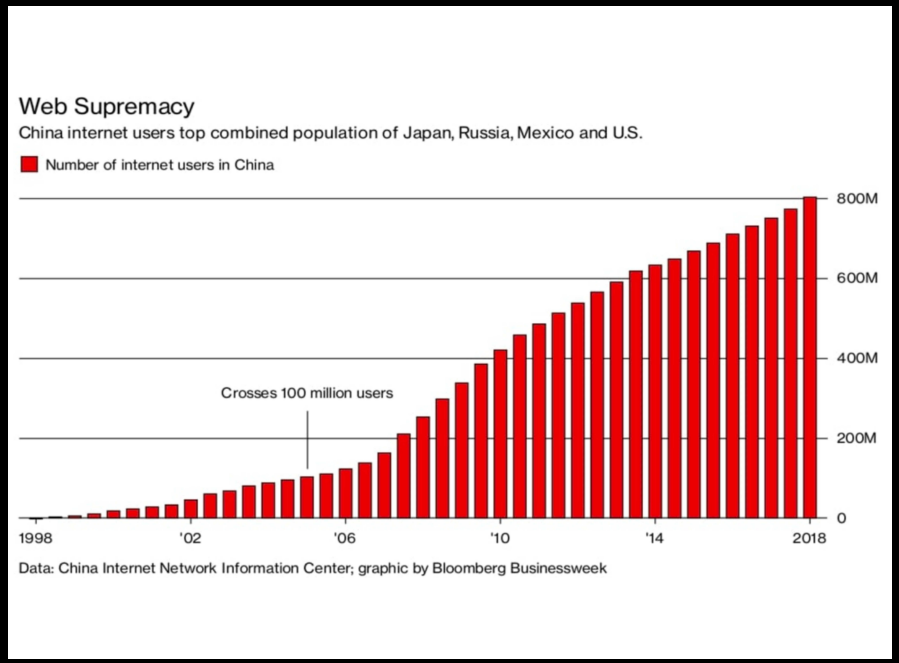


Data by the Numbers: *Every Single Day*...

- 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 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?

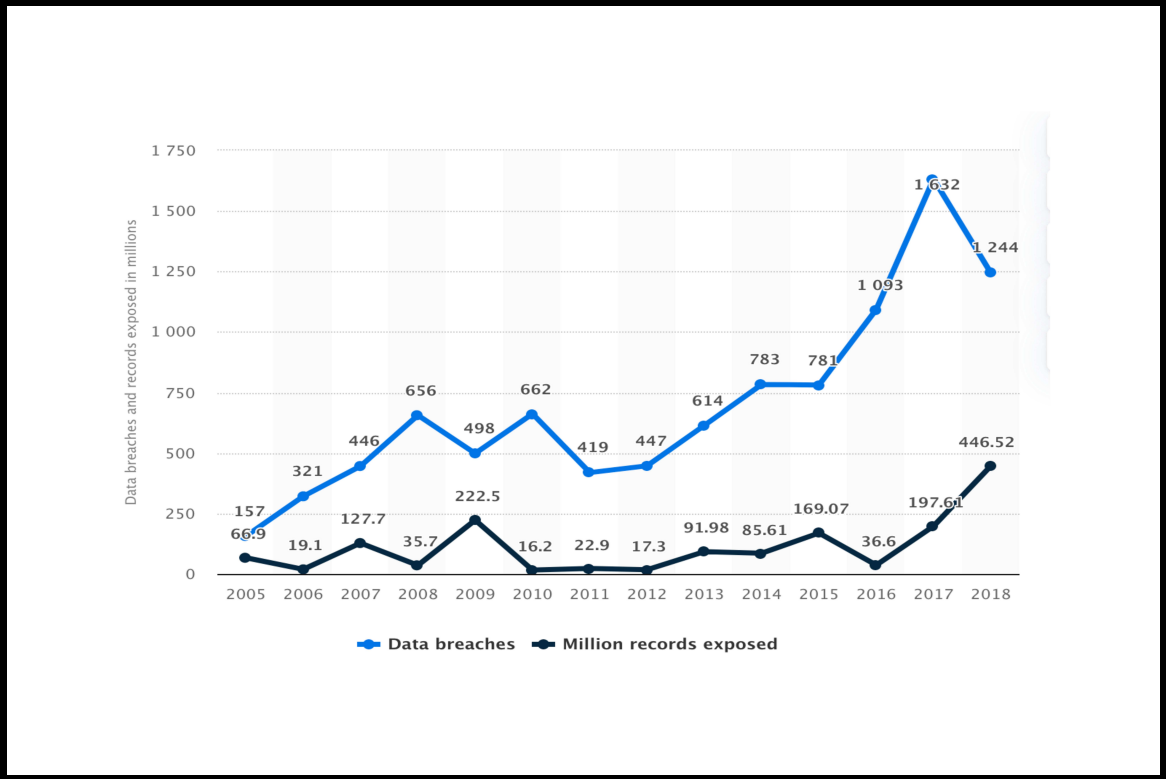
- One answer is that it is a simple classifier
- It is actually just statistics with an emphasis on prediction and accuracy
- Basically 5 tools: Support Vector Machines, Random Forests, Neural Networks (in countless variations now, where the name comes from resembling how the neuro-cranial system works), 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
- Deep learning algorithms (a subset of ML) establish initial parameters from the data and then train the computer to learn independently by recognizing data patterns using multiple layers of processing.
- Currently social and biomedical scientists are widely using these tools: we *are already* data scientists

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.
- The recruiter Glassdoor ranks the 50 in 2021 best jobs in America that pay over $100,000. Data Scientist is ranked No. 2.
- There were about 30M job ads for data scientists in the US alone in 2020 according to IBM
- There is also a 15 year increasing trend for PhD *social scientists* to succeed in this labor market

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

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)

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).

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

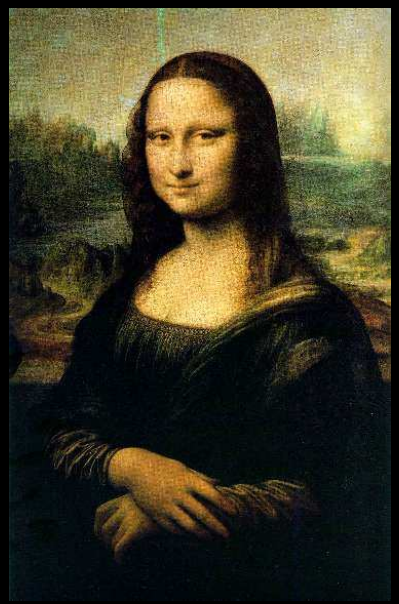
- Two models of humans...

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...

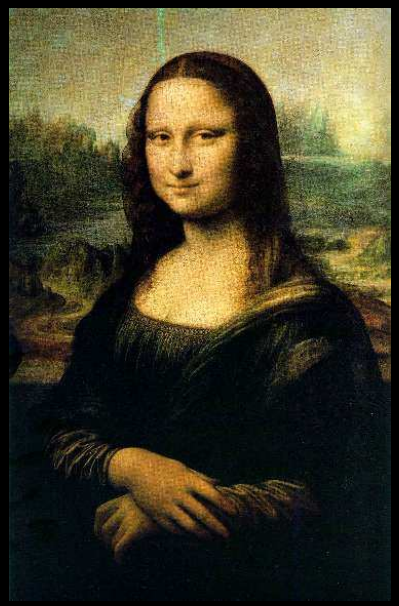


On Models

- 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

Models: Machine Learning/Big Data Context

- Describe the Objective:
  - Formulate an exact statement of the problem to be solved: classifying types, searching for patterns, sorting scores, scoping, etc.
  - Moving from a vague goal ("understanding credit card transactions"), to a clear question ("why do consumers spend more with prestige cards?"), to specific tasks ("we want to use FRED credit card data to model spending by card type looking for important features") [`https://fred.stlouisfed.o`
  - This process sounds obvious but almost all big data work is done in teams (social scientists, computer scientists, data scientists, managers), so agreement is essential

- CMU Data Science Project Scoping Guide Initial Screening Criteria
  - Impactful: The problem we're solving is real, important, and has social impact
  - Solvable: Data can play a role in solving the problem, and the organization has access to the right data
  - Actionable: The organization has prioritized this problem, is ready to take actions based on the work, and is willing to commit resources to validate and implement it

Models: Machine Learning/Big Data Context

- Formulate the Method:
  - What type of machine learning task is needed?
  - Regression? Classification? Outlier detection in new data? Risk determination? Path analysis?
  - This will narrow the set of tools down to a manageable set of alternatives
  - SVM, random forests, neural networks, regularization, categorical outcomes regression, regularization, etc.

Models: Machine Learning/Big Data Context

- Data Preparation and Exploration:
  - Data acquisition: government, academic, or corporate sourced? Web scraping? Experimentation?
  - The usual data cleaning, labeling, recoding, dealing with missingness, and documenting
  - Visual and descriptive exploration
  - Identification of possibly important variables
  - Determining the levels of measurement for variables to be included in the analysis
  - Data storage and preservation are often a challenge for very big data

Models: Machine Learning/Big Data Context

- Feature Engineering
  - What are the outcome variables of interest?
  - What are the features of interest?
  - Are temporal effects important?
  - Are spatial effects important?
  - Are interactions of possible importance?
  - Are hierarchies (levels of agreggation) important?

Models: Machine Learning/Big Data Context

- Multiple Model Specification
  - Determine a set of competing model approaches since it is not known in advance which will perform the best with this specific dataset
  - Apply this suite of models to these data at hand
  - This includes evaluation methods to judge fit, prediction accuracy, reliability,...
  - Subsetting, recoding, or combining data units may be needed after this step

Models: Machine Learning/Big Data Context

- Selection and Optimization
  - Which models, model features, and model tunings are best?
  - What are the implications of specific features?
  - Robustness and Resistance evaluation
  - Determination of errors and risks
  - What are the tradeoffs?
  - This step is also called interpretation becase we are interpreting the implications of approaches and features

Models: Machine Learning/Big Data Context

- Validation
  - After picking a model (or possibly several) on historical/validation/test data (more on this later), then validate it on out-of-sample data (either new data or a subset of the current data)
  - Part of this approach can be simulation, experiments, or new data acquisition
  - There are lots of approaches here and lots of definitions of valid

Models: Machine Learning/Big Data Context

- Fielding
  - Suppose now that there is a selection of the best model and it has been validated on out-of-sample data
  - Usually the process is not a one-off endeavor in the corporate or governmental setting
  - Now the model is applied to new data that comes in over time or by broadening the scope of the question
  - Putting the model into practice can produce huge non-human analytical feats
  - But it is important to realize that the data generation process can change over time

Plan for the Rest of the Workshop

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