The Present and Future of Data Science

Profesional Development Workshop, July 2021

JEFF GILL

Distinguished Professor

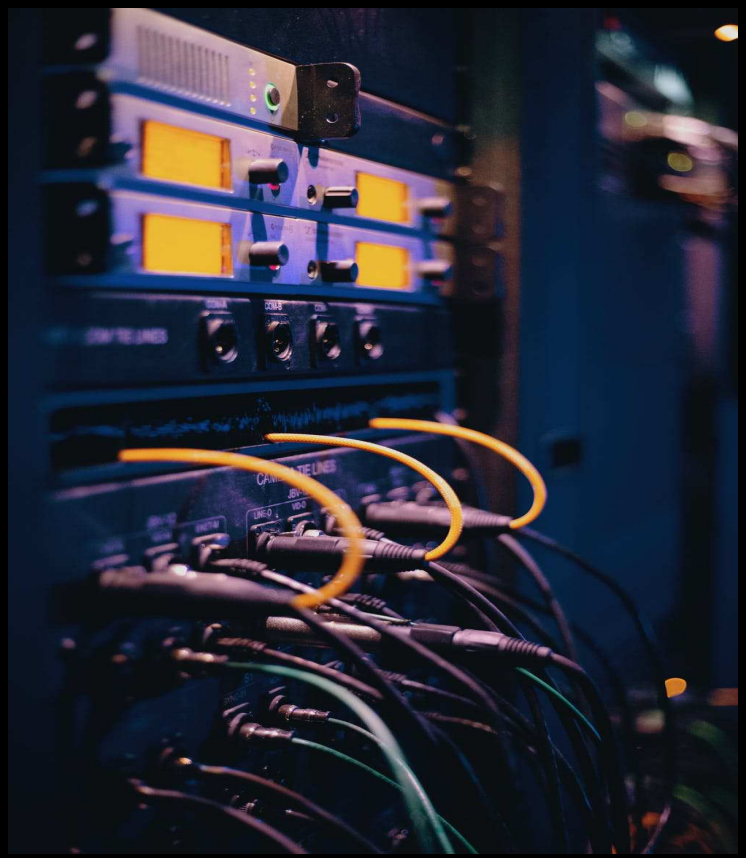
Department of Government, Department of Mathematics & Statistics

Member, Center for Neuroscience and Behavior

Founding Director, Center for Data Science

American University

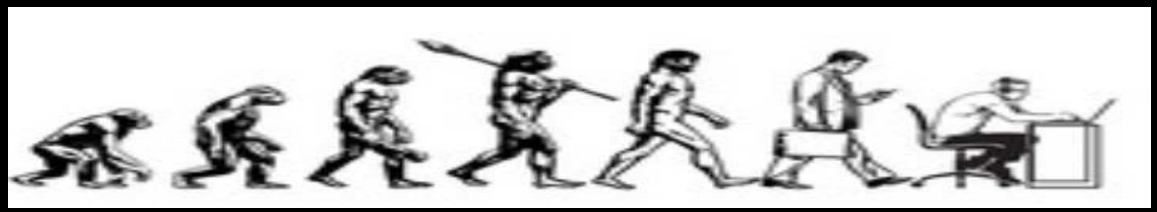
Macro Data Forces



- We live in the data century, *whether we like it or not*
- Our personal lives, our careers, our finances, our social activities, our childrens' 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*

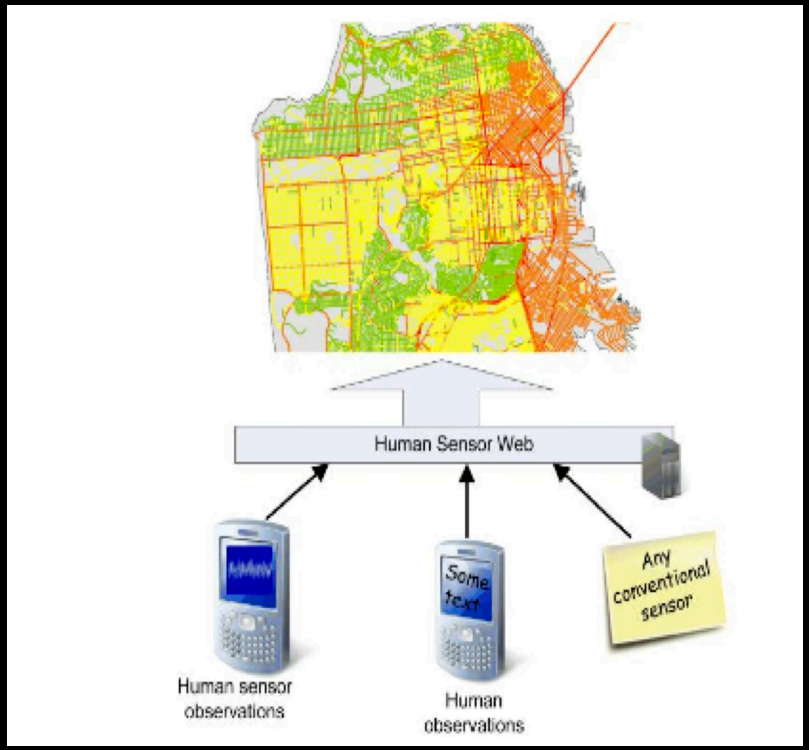
Perspectives on Human History

- 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 Palaeolithic 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, hence this talk
- 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

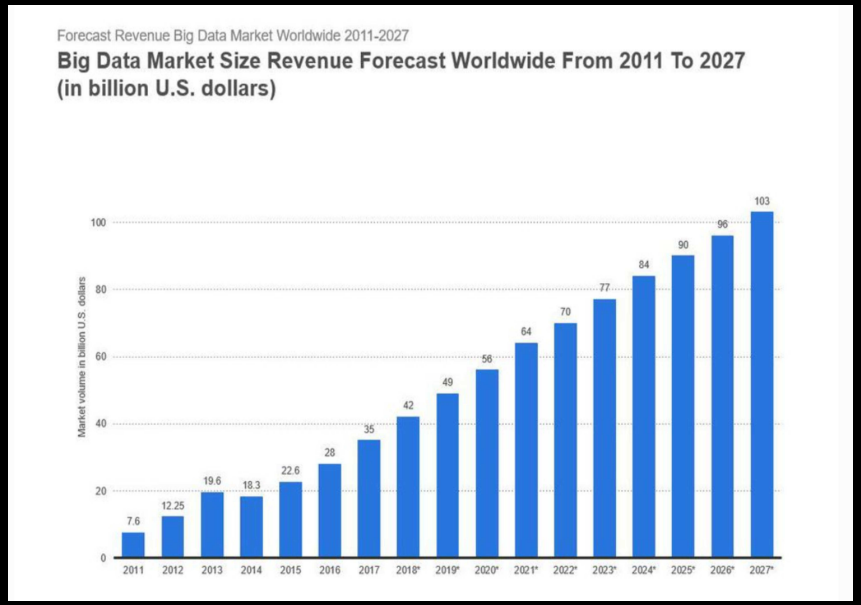


"A Change Is Gonna Come," Sam Cooke (1964)

- 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

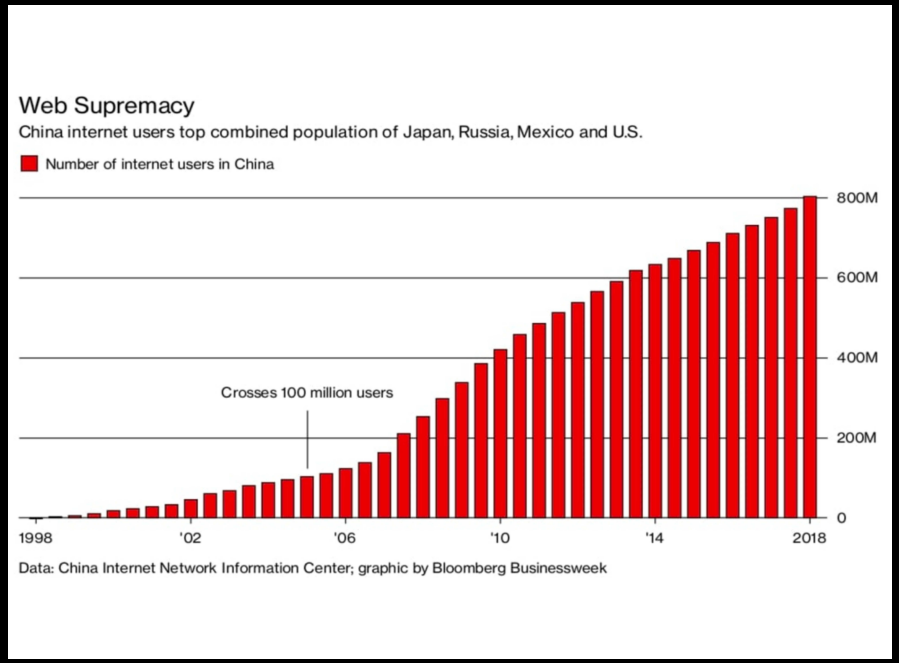
Data by the Numbers: *Every Single Day*...

- 23 billion text messages are sent
- 5 billion searches are made (40,000 per second on google alone)
- 500 million tweets are sent
- 294 billion emails are sent
- 4 petabytes of data are created on Facebook
- 4 terabytes of data are created from each connected car
- 65 billion messages are sent on WhatsApp
- 360 terabytes are uploaded to YouTube



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

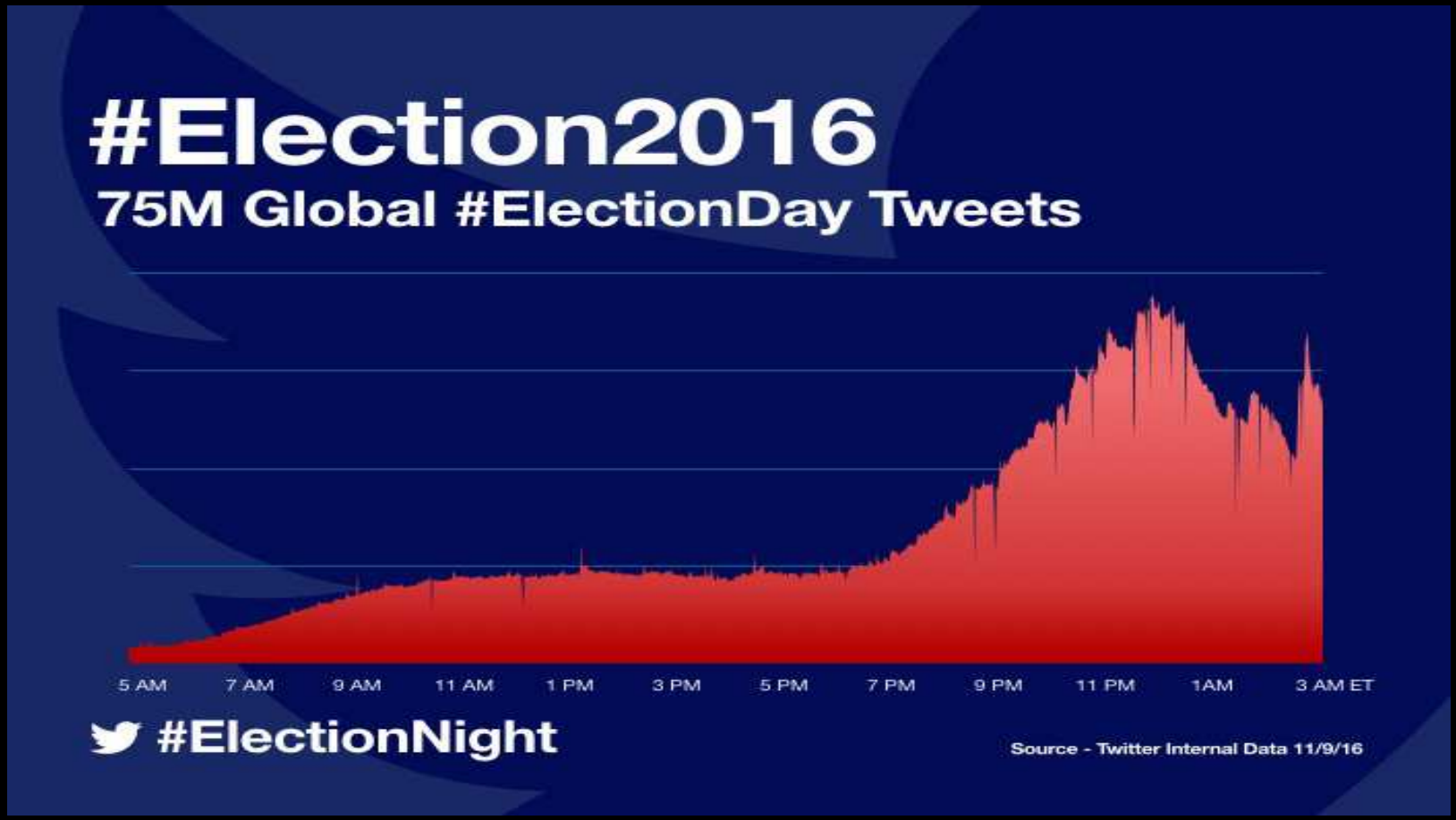
- 21.6 million GIFs are sent via Facebook messenger
- 149 billion spam emails are sent
- 222 million calls placed on Skype
- Venmo processes \$75M peer-to-peer transactions
- The Weather Channel receives 2.6×10^{10} forecast requests
- 65M Uber bookings
- The average online person generates 10^{18} bytes 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

November 8th, Evening 2016: Over 75M Tweets (still a record for one topic)



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
- Sometime before 2025 463 exabytes of stored data will be created every day
- So what are some tools to deal with data-size challenges?

Relatedly, What is Machine Learning?

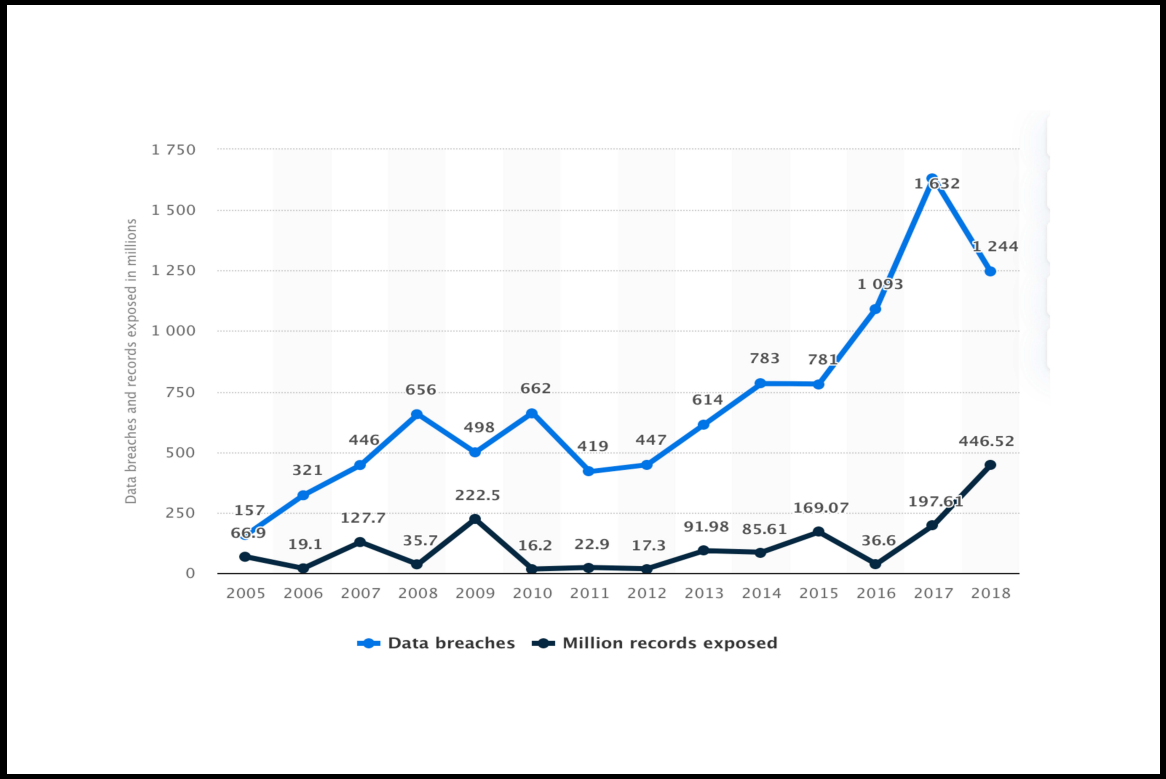
- One answer is that it is a set of simple classifiers
- It is actually just statistics with an emphasis on prediction and accuracy
- Basically 4 tools: Support Vector Machines, Random Forests, Neural Networks (in countless variations now, where the name comes from resembling how the neuro-cranial system works), and Logit Regression(!)
- ML is most effective when automated with *many* hopefully reliable examples to adapt to tasks independently, which is not how all 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.
- The biggest users of machine learning by far are large mega-corporations around the world with ridiculously large and pliable datasets.

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. Your 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 Amazons 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)
- Ridiculously 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, 2019 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 nevermind 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 $100B in business (this is probably a low estimate)

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 just released its annual ranking of the 50 in 2011 best jobs in America that pay over $100,000. Data Scientist is ranked No. 2 (behind Javascript Programmer).
- There were about 30M job ads for data scientists in the US alone in 2020

Ongoing Data Science 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

Some Points about R

- The environment and language R is the lingua franca of modern statistics and data analysis.
- This professional development workshop is centered around doing work in R for this reason.
- It can be challenging on the front-end but becomes intuitive shortly there after.
- It is literally the most powerful statistical language and yet it is free.
- R was created by and is maintained by academics so it will never be commercialized.
- Despite this every major corporation and government agency uses R.

Wrapping Up...

- Human life is more complex, data-oriented, and technical than ever before
- Every field, including mine, needs to understand that they are also a data science field
- By "data" I mean qualitative and quantitative data (e.g. "evidence"), so this discussion did not exclude anyone
- THANK YOU