

Biotechnology Big Data Artificial Intelligence

Promises, Dangers, Worries, Gotchas

Slides available at https://webbedfeet.github.io/GMU_2019

Abhijit Dasgupta, PhD
Chief Data Scientist
Co-Founder



- Evidence of effective
- Interferon inserting Reduces
- called includes solution
- Quality

also **increased** **modified** **recombinant** **treatment** **living** **process** **ext** **antibodies** **Context** **pharmacology** **qua** **techniques** **hybrid** **similarly** **drugs** **development** **Evidenc** **Reprodu** **solution** **Manufac**

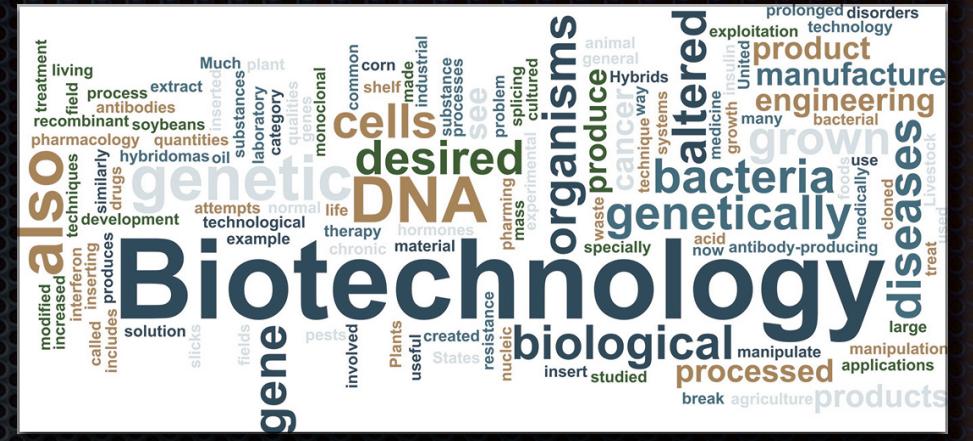
- **interferon** **inserting** **called** **incides**
- **produces**

- # • Quality

The word cloud includes the following words and their approximate bounding boxes:

- biotechnology (large, dark blue)
- bacterium (highlighted by a red oval, medium, dark blue)
- genetically (large, dark blue)
- processed (medium, brown)
- manipulate (small, dark blue)
- insert (small, dark blue)
- nucleic (small, brown)
- studied (small, green)
- resistance (small, dark blue)
- pharming (small, brown)
- mass (small, brown)
- experimental (small, light blue)
- waste (small, green)
- specially (small, green)
- cancer (small, light blue)
- technique (small, brown)
- way (small, light blue)
- systems (small, light blue)
- Hybrids (small, brown)
- animal (small, light blue)
- general (small, light blue)
- problem (small, dark blue)
- spliceing (small, green)
- cultured (small, brown)
- exploitation (small, light blue)
- United (small, light blue)
- insulin (small, light blue)
- medicine (small, brown)
- growth (small, brown)
- many (small, light blue)
- bacterial (small, brown)
- engineering (small, brown)
- manufacture (small, dark blue)
- product (small, brown)
- technology (small, light blue)
- large (small, green)
- manipulation (small, brown)
- applications (small, green)
- treat (small, brown)
- use (small, light blue)
- medically (small, light blue)
- cloned (small, brown)
- Livestock (small, light blue)
- used (small, light blue)

Context (why we need it)

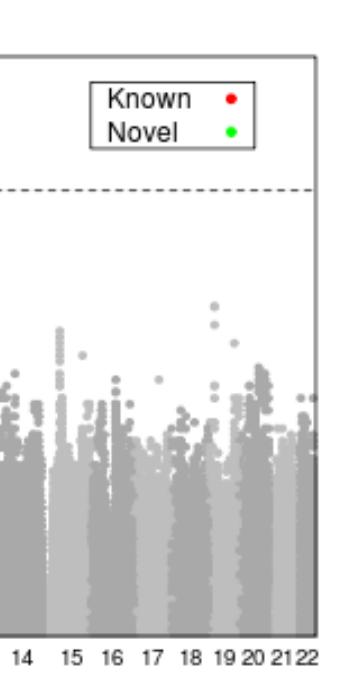
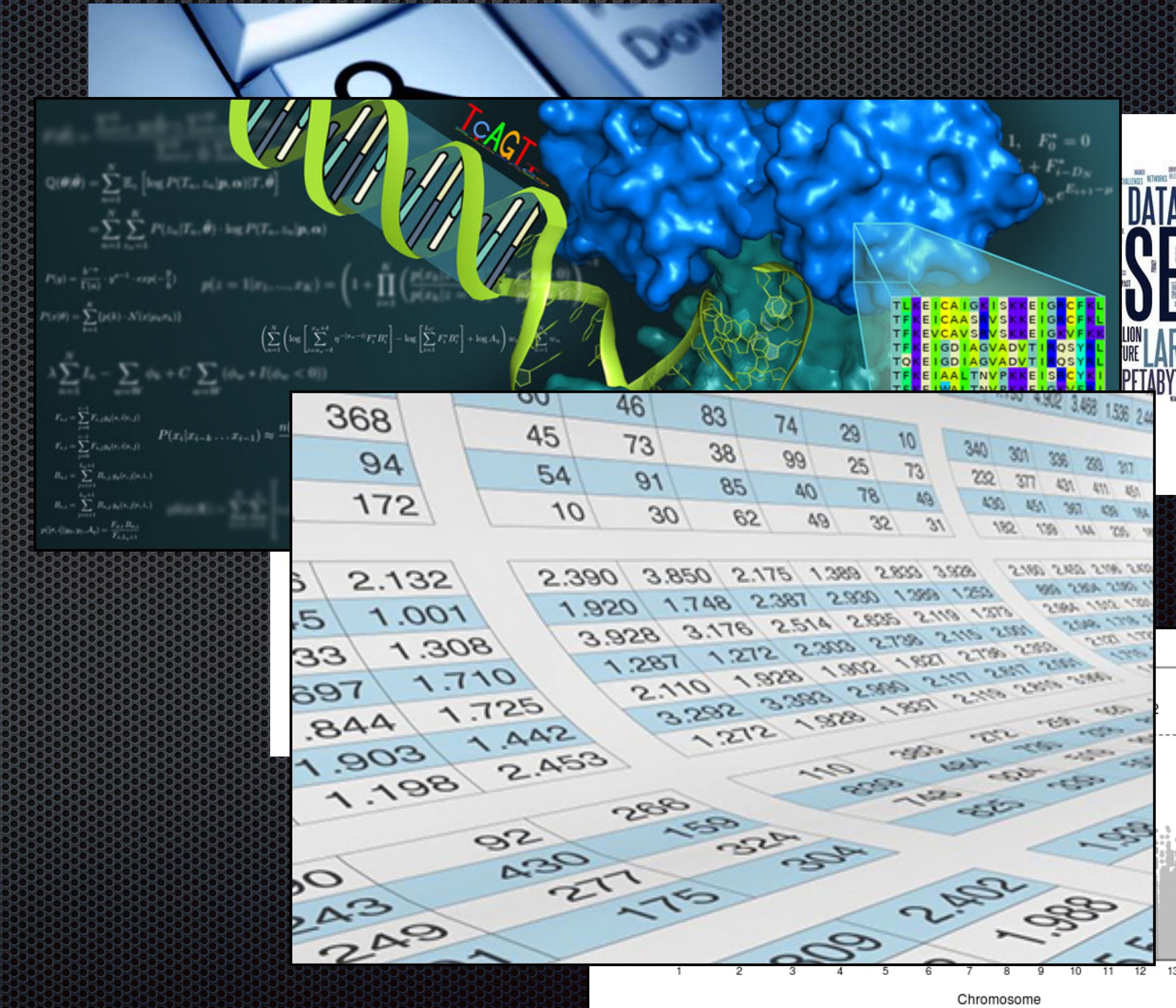


- Market research
 - Competition research
 - Focus groups
 - Gaps in the market
 - Market potential



Discovery (finding it)

- Identifying potential
- Searching the literature
- Searching databases
- Analyzing relationships
- Culling candidates



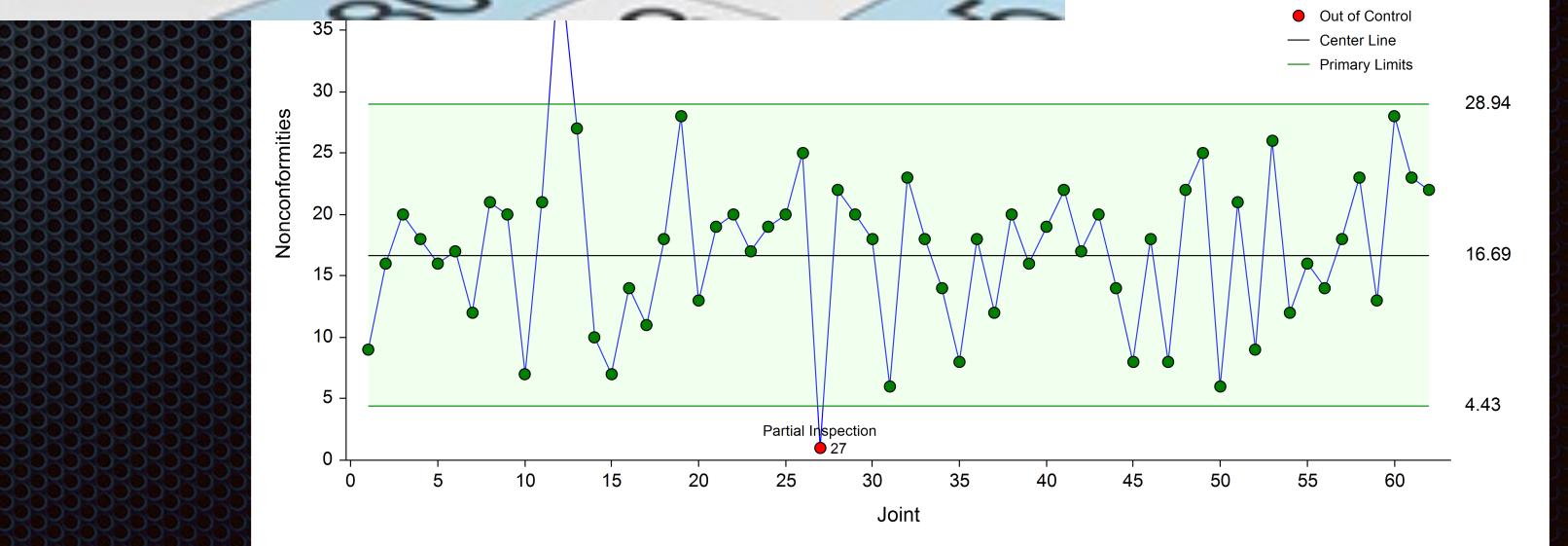
Evidence of effectiveness

- Laboratory experiments
- Field experiments
- Observational studies
- Randomized trials



Manufacturability & Quality

- Translation to a process
- Establishing standards
- Reproducibility
- Quality control / Six Sigma



A word cloud diagram centered around the word 'Biotechnology'. Other prominent words include 'genetic', 'DNA', 'cells', 'desired', 'organisms', 'biological', 'process', 'manufacture', and 'engineering'. The words are colored in shades of blue, orange, and grey, and their size corresponds to their frequency or importance.

Data drives

Research

Development

Production

Marketing

Sales

Revisions



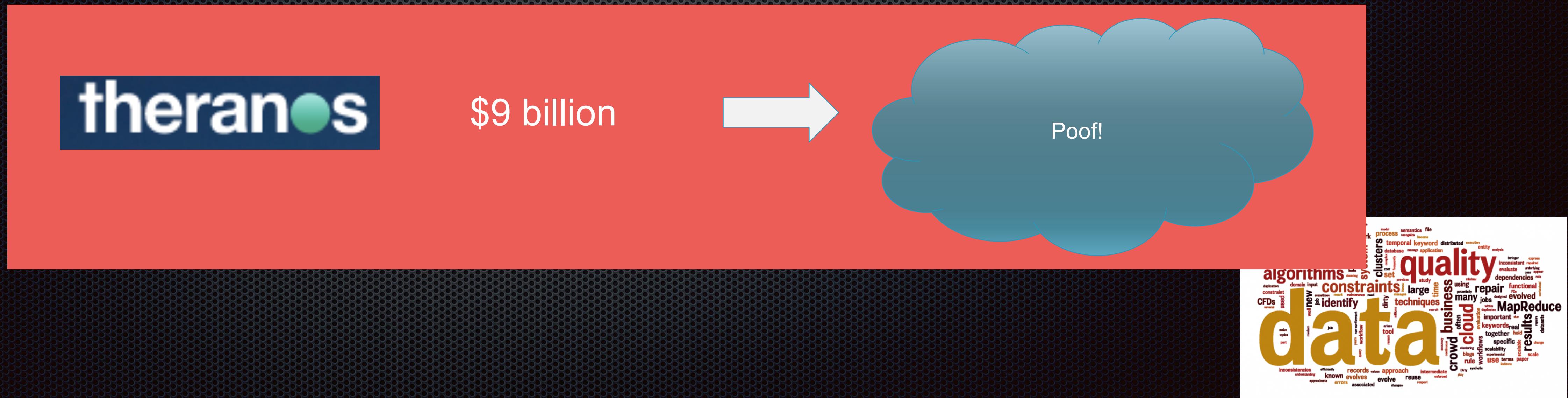
Central issue for data-driven business

- The data must be reliable and valid
 - Repeating the experiment must yield similar results (reliability)
 - We're measuring what we think we're measuring (internal validity)
 - The results can be generalized to larger populations (external validity)



Central issue for data-driven business

- This is a matter of trust
 - If the foundational data on which the business is built isn't trustworthy, you can't sell anything



Central issue for data-driven business

- It doesn't matter how much data you can get
 - It matters how you got it
 - Experimental design
 - Ethics
 - Provenance



Central issue for data-driven business

- It doesn't matter how much data you can get
 - It matters how you use it

Analytics

- Translation to business practice
 - Follow-up



system automation engineering brain interface mechanical communication

internet relations business machine robotic

cybernetics cyber modern scifi cyborg human

technology science design

artificial robot concept tech

intelligence

information future computer virtual digital

cooperation futuristic people

mind man connection

internet business line
cyber modern cyborg human
scifi electronic creation
technology science design

system automation
engineering brain
mechanical commun
ar
in'
information f
coo

The word cloud is centered around the word 'learning'. Other prominent words include 'data', 'analysis', 'research', 'processing', 'mining', 'statistics', 'machine', 'intelligence', 'systems', 'modelling', 'deep-learning', 'bigdata', and 'data'. The words are in various sizes and colors (black, red, orange, brown), and smaller text provides additional context or related terms.

learning

data

analysis

research

processing

mining

statistics

intelligence

systems

modelling

deep-learning

bigdata

data

intelligence

systems

modelling

deep-learning

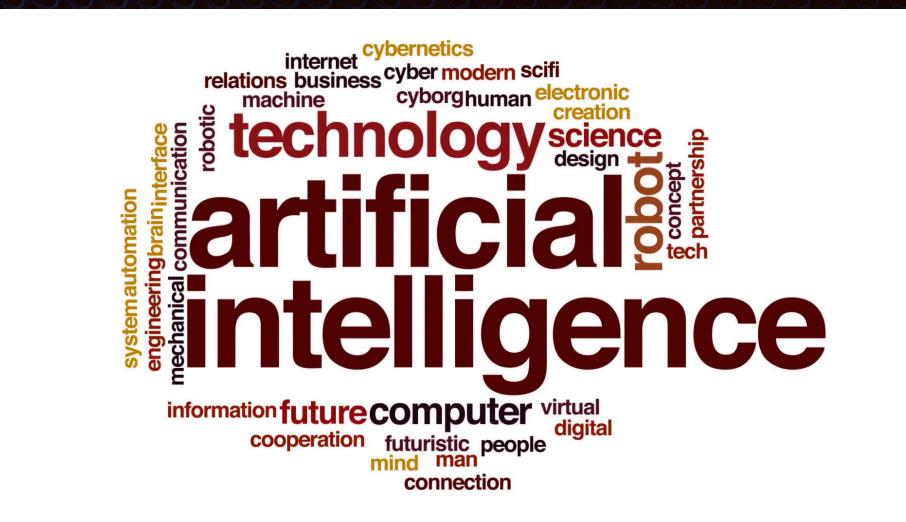
bigdata

data

total **status** **collection** **interpretation**
data **probability** **statistica**
prediction **population** **Inference**
experiments **science**
theory

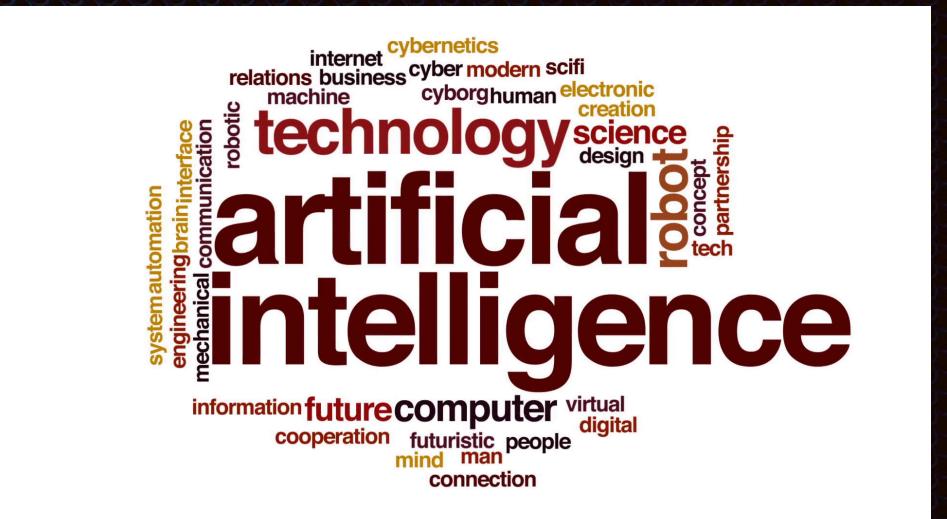
What is AI?

- The ability of machines (computers) to learn, reason and act on their own
- The ability of machines to demonstrate intelligence
- The ability of machines to be autonomous, independent, and make decisions based on their environment



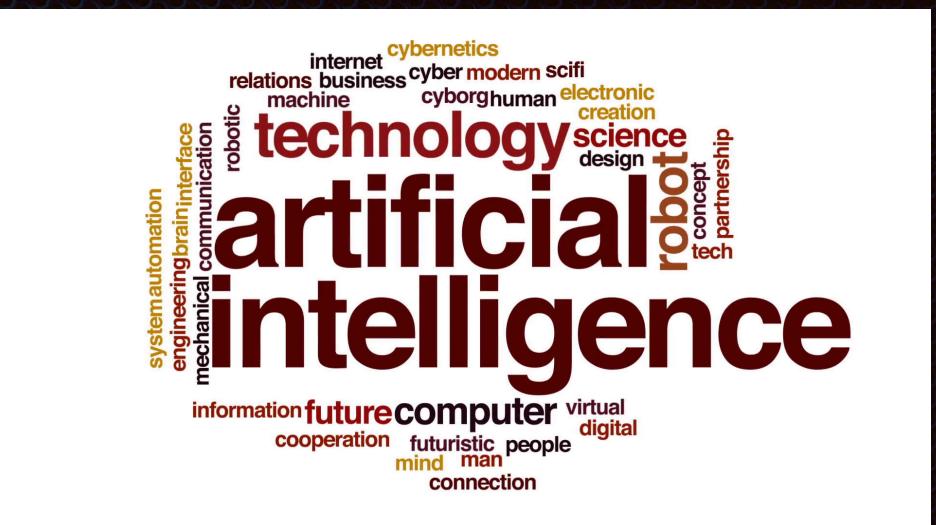
All in Biotech

- Drug discovery
 - Culling through millions of candidates to identify the next molecule for drug development
 - Patient management and decision support
 - Fast analytics to provide likely causes, drug effects and interactions, risks and side-effects based on patient history, EMR mining and literature mining



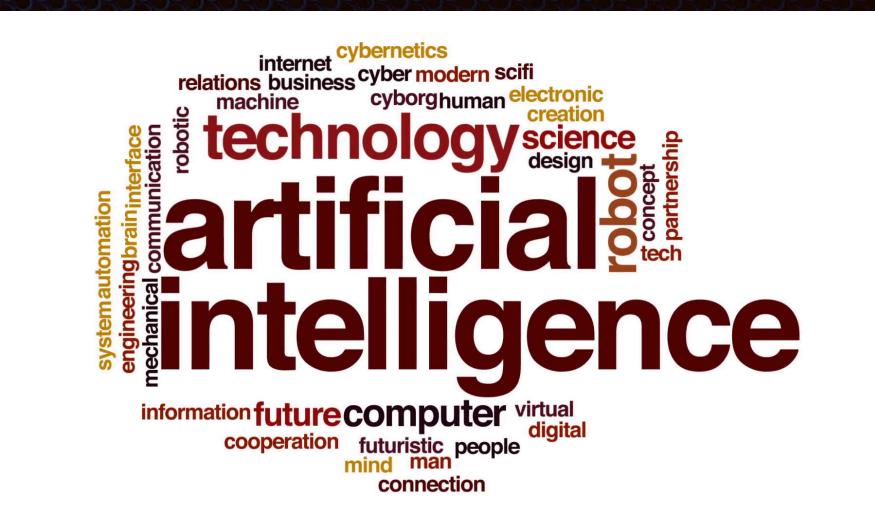
AI in Biotech

- ❖ Prosthetics/exoskeletons and predictive movement
- ❖ Environmental assessment and closed-loop therapeutic deployment
- ❖ Performance optimization through sensing, adaptation and intelligent coaching
- ❖ Precision medicine
 - ❖ Identifying therapies better targeted at the individual



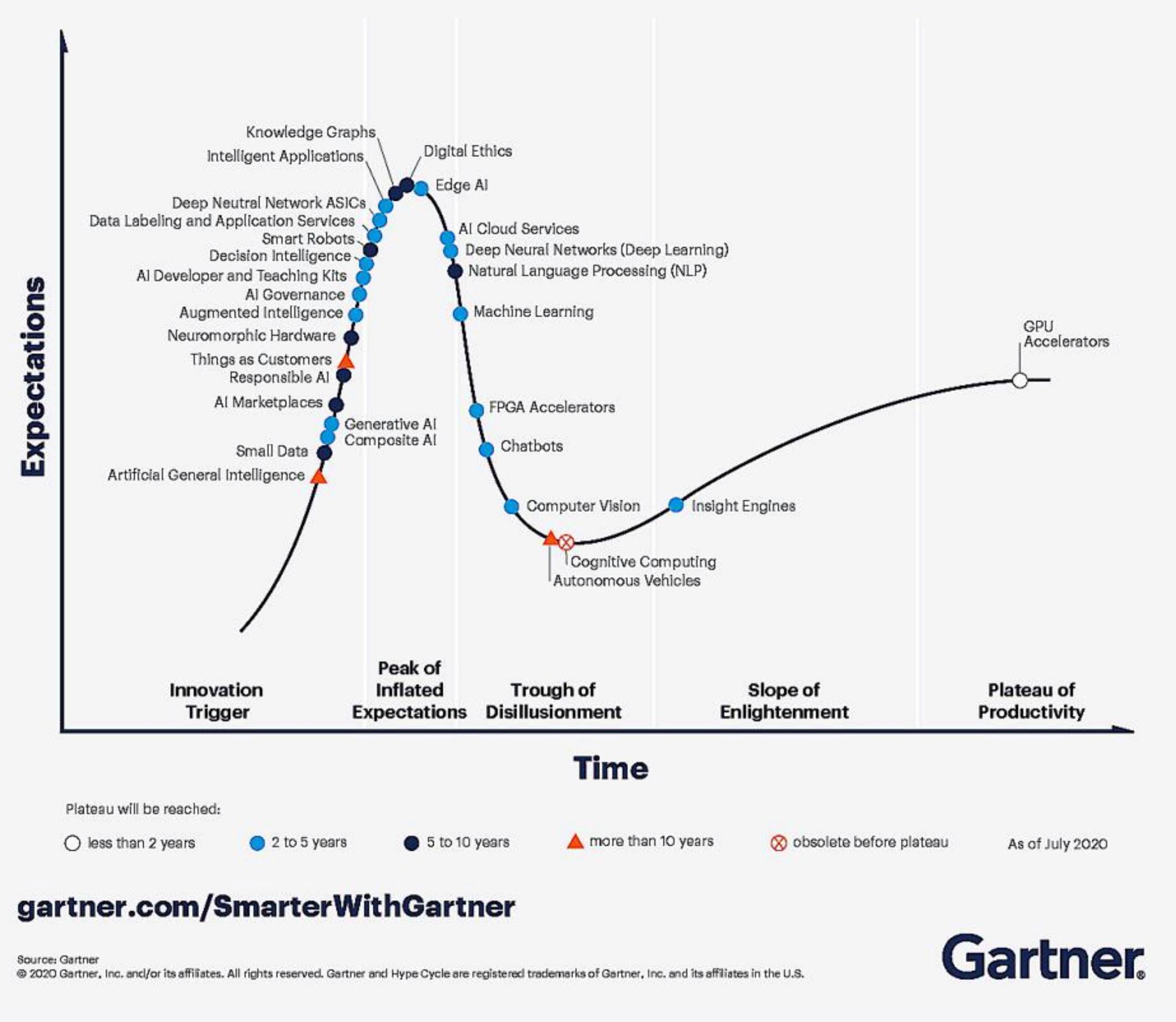
Are we there yet?

- Not really
 - Self-driving cars and planes on auto-pilot
 - Experimental situations
 - Some medical decision support

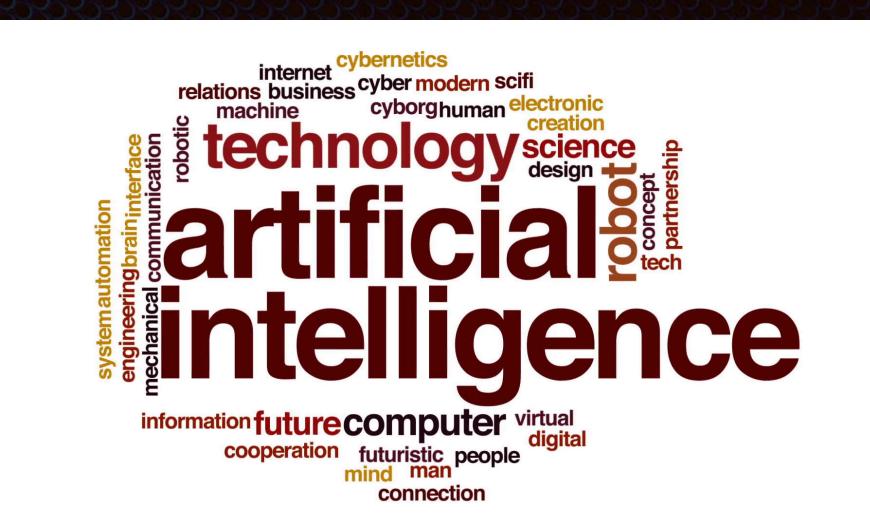


Are we there yet?

Hype Cycle for Artificial Intelligence, 2020



Gartner Hype Cycle
2020



So where are we?

system automation
engineering brain interface
mechanical communication

internet relations business machine robotic

cybernetics cyber modern scifi cyborg human

electronic creation science design

technology

artificial robot concept tech partnership

intelligence

information future computer virtual digital

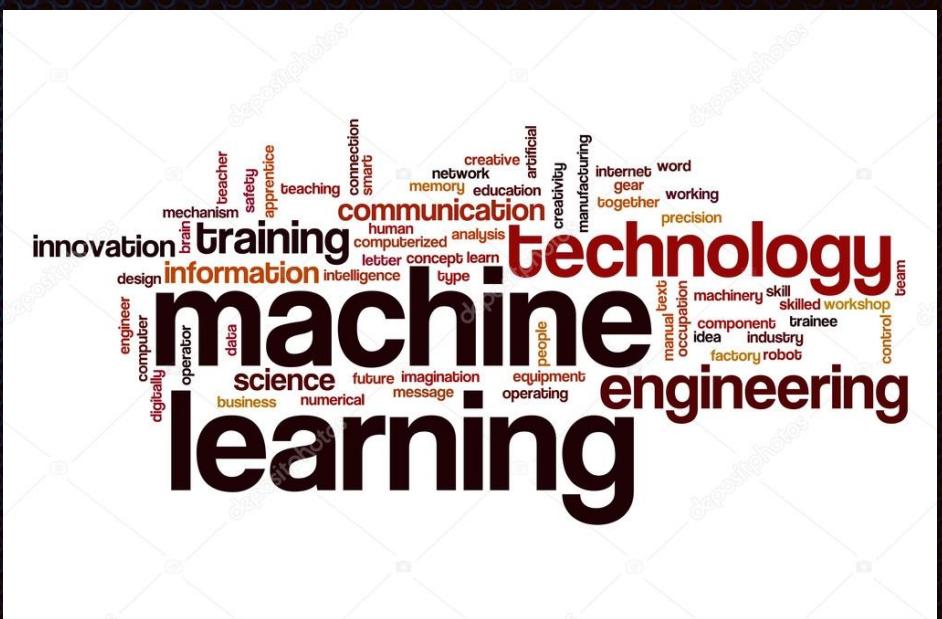
cooperation futuristic people

mind man connection

Common use

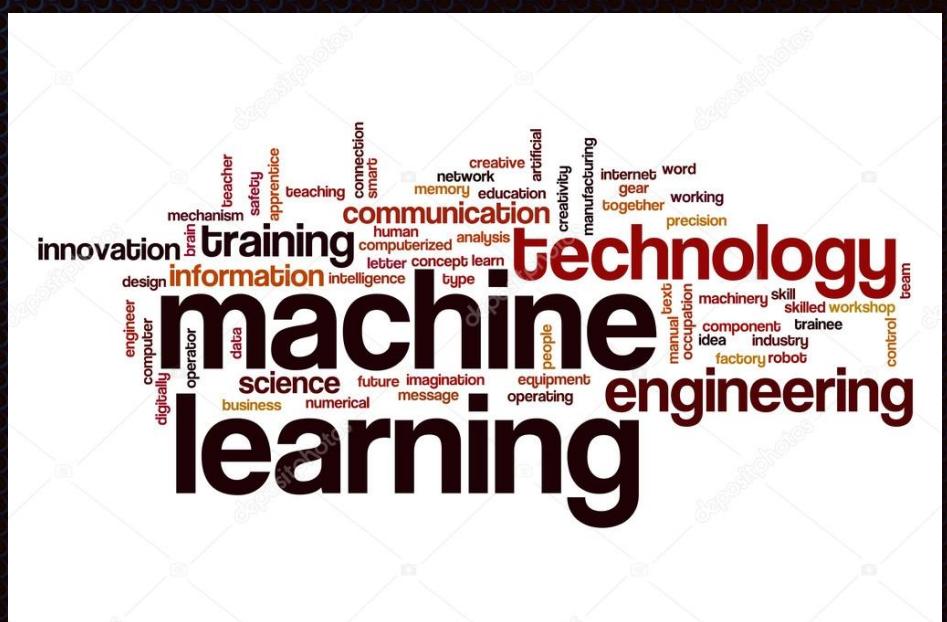
Machine learning

- Artificial intelligence is about learning, reasoning and acting
 - We are still trying to get the learning part down well
 - Learning innate patterns in data
 - Learning how an outcome is influenced by different factors
 - Learning to predict the future by observing the past
 - Learning to discern what is real and what is artifact



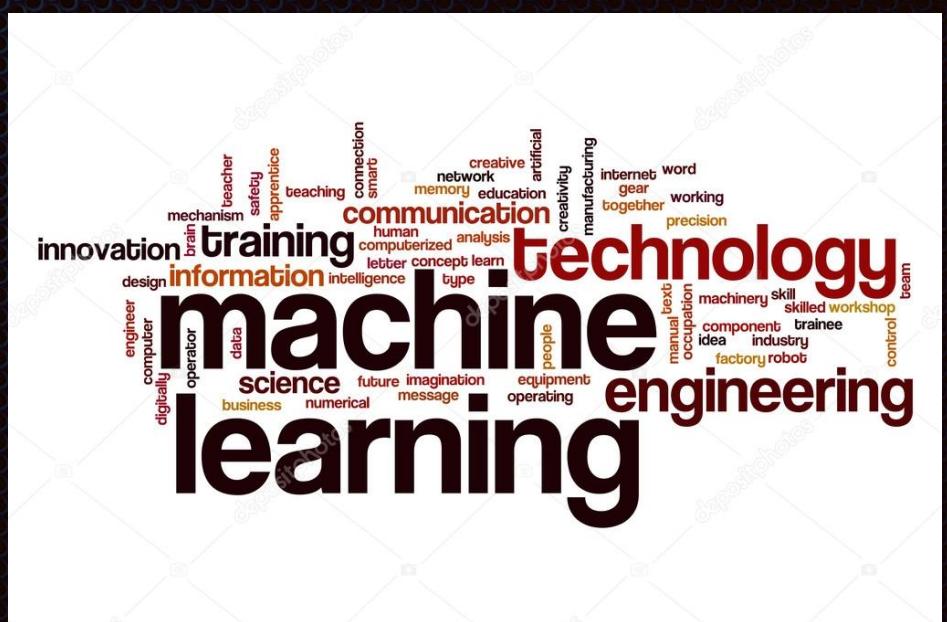
Machine learning success stories

- Breast cancer therapeutics
 - Bioinformatics used to understand innate differences in different types of BrCa, leading to very successful targeted therapeutics
 - Tamoxifen, Herceptin, Fluoracil, Paclitaxel



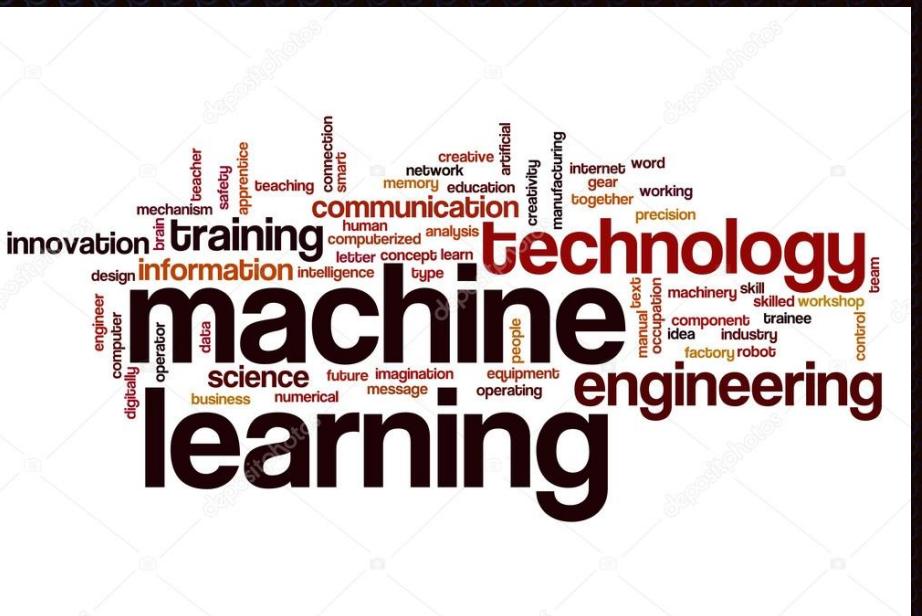
Machine learning success stories

- Identification of subgroups in various diseases, leading to more targeted interventions
 - EGFR & Lung Cancer
 - APoE & Alzheimer's prevention
 - BRCA & Breast cancer prophylaxis
 - Immunotherapies
 - Gene editing (CRISP-R)



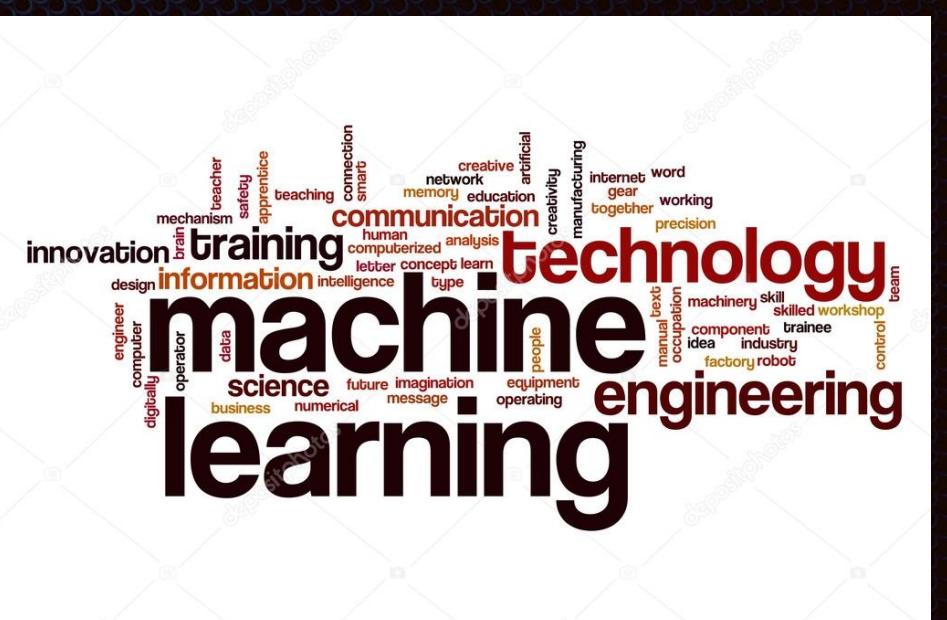
Machine learning success stories

- Risk assessment, epidemiology and causality
 - Cancer risk prediction
 - Gail model for Breast Cancer
 - Impact of inflammation on health
 - Identification of genetic and proteomic risk factors in complex disease



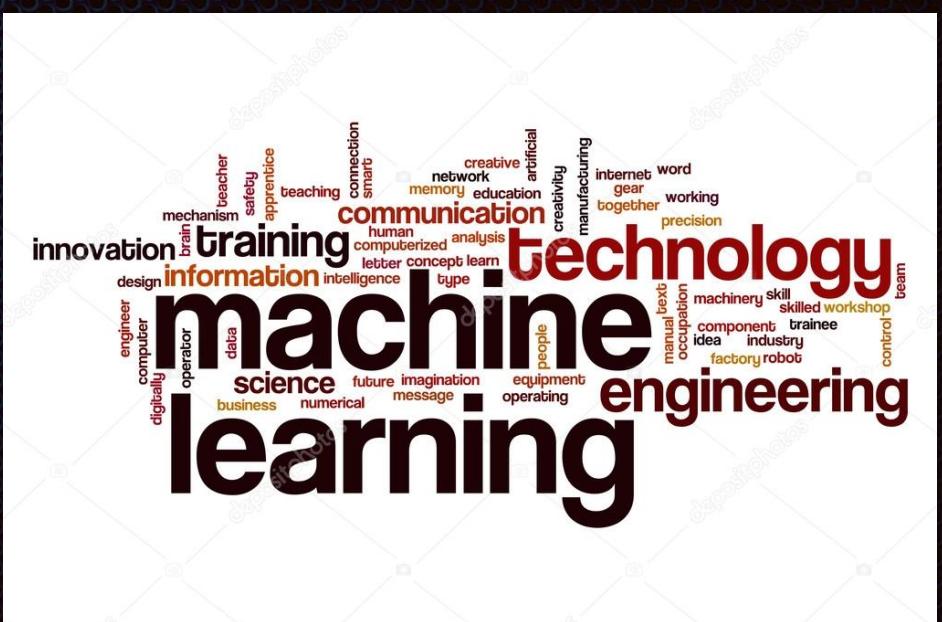
What is ML?

- Algorithms that ingest data and adapt to mimic and express inherent data patterns
 - Learns from data
 - Can provide predictions
 - Algorithms can be
 - Simple (linear regression)
 - Complex (decision trees)
 - Indecipherable or magical (neural networks / deep learning)

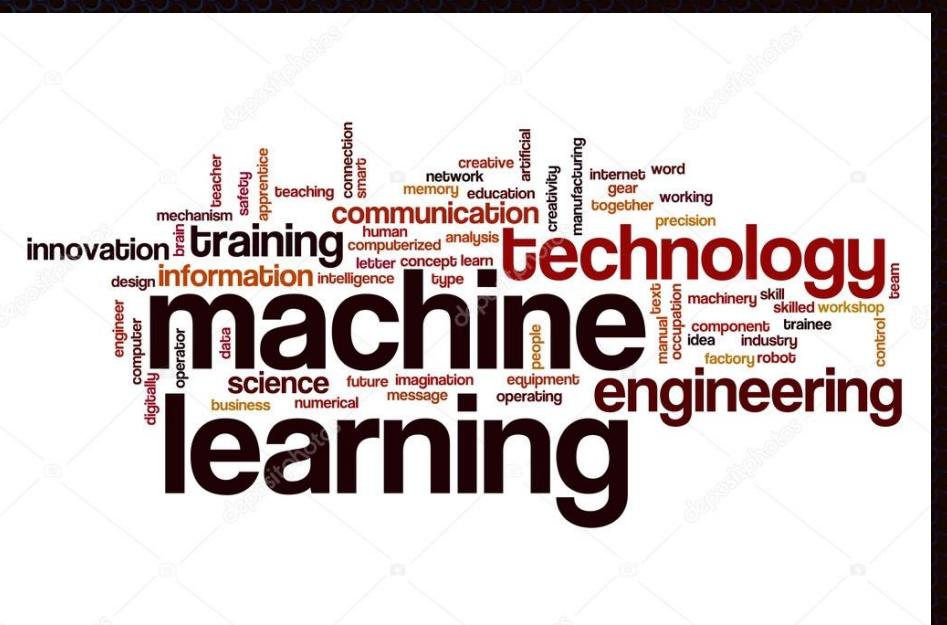


What is ML?

- Algorithms that are mainly meant for prediction
 - Supervised learning
 - They cannot often express why the predictions happen
 - What factors influenced the prediction in what way?
 - It's a function of the complexity of the model and our inability to express or comprehend that complexity easily
 - Some advances in interpretable ML with Shapley values

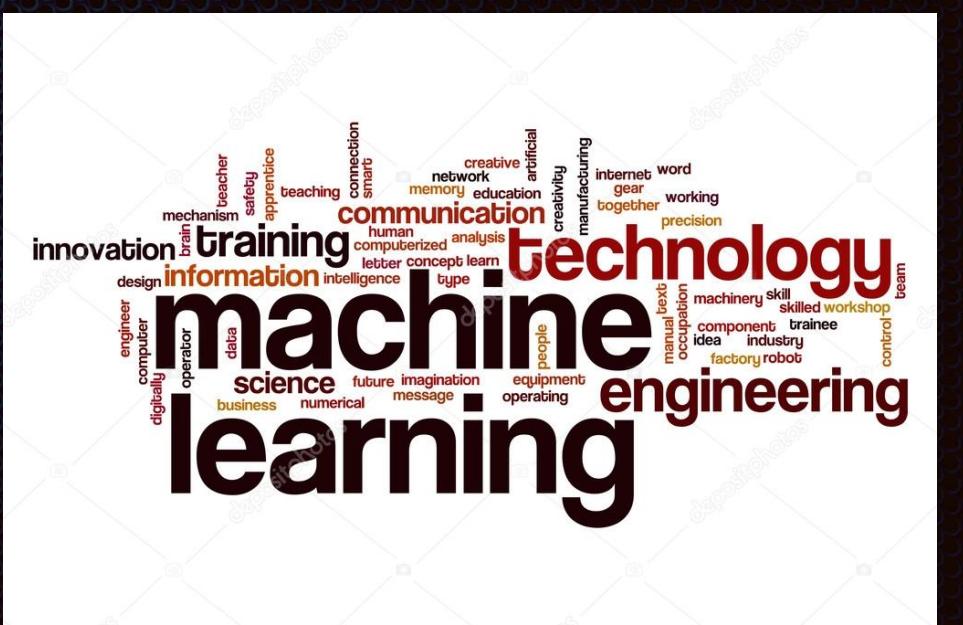


ML and Big Data



ML and Big Data

- Complex models require a lot of data to work well
 - Deep learning works for Google and Facebook because of their data lakes
 - So we have been collecting lots of data in the hope (and prayer) that we can learn from it
 - And, by lots, I mean....

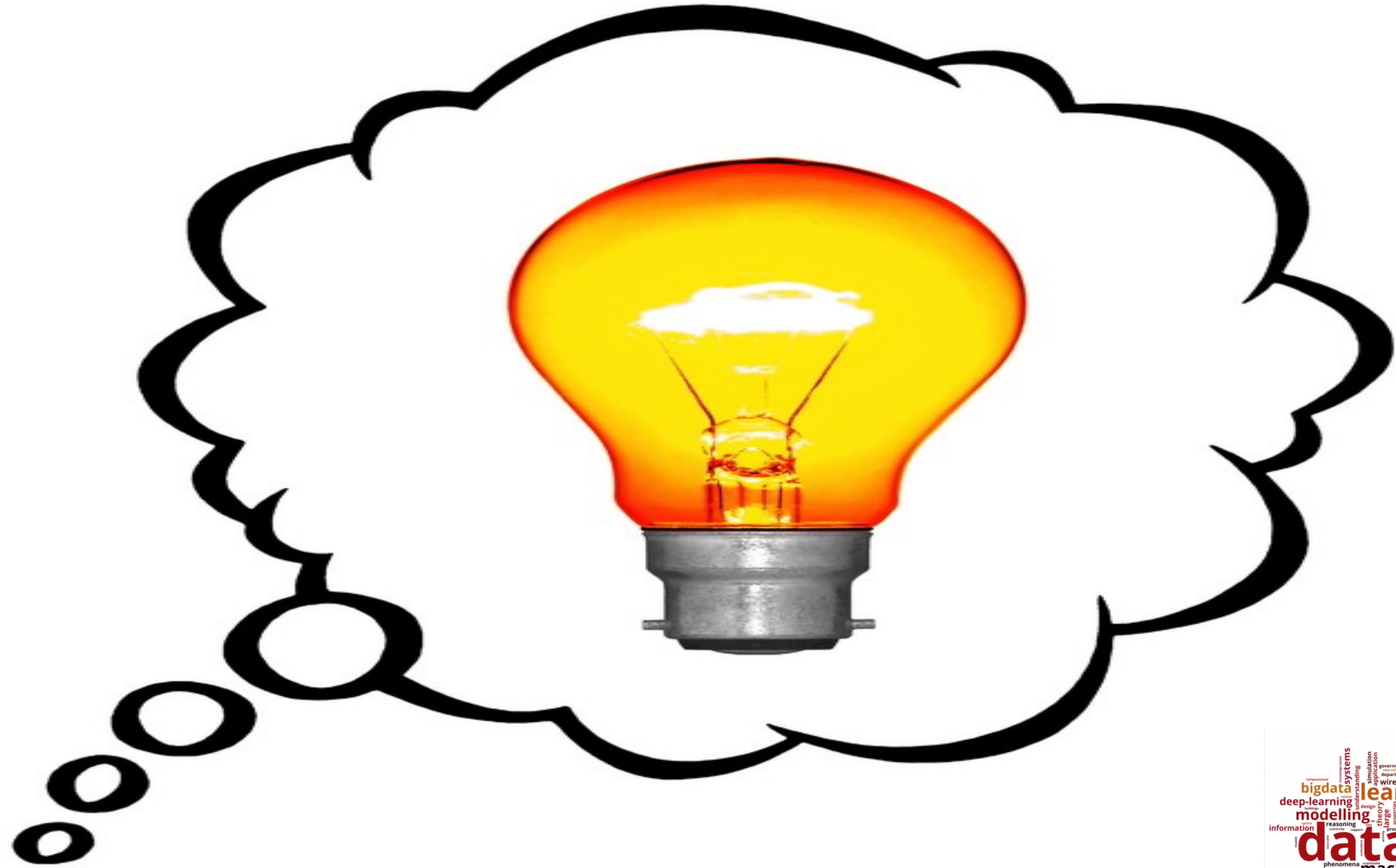




Big Data in Drug Discovery

- Pharma is losing patent control
- Getting a drug to market requires many years and billions of dollars
- However, we have plenty of clinical, bioinformatic, molecular and trial data





A bright idea

- Let's mine this data for good candidate molecules
 - No assays, just data
 - To confirm, outsource to do targeted assays



Grab public data.
Start a company.
Sell the company.
Repeat

Dr. Atul Butte, UCSF, 2015

Data sitting still is useless
Data needs to stay in motion
Data needs to get us someplace



Where's the data



Where's the data

- Our bodies
 - Genetic/proteomic/molecular code (1.5 Gb per cell)
 - Movement / behavior
 - Sleep
 - Gait
- Sounds



Where's the data

- Interactions with the world
 - Health records and Insurance
 - Government data
 - Telecommunication
 - Social media
 - Economic behavior



Where's the data

- Data we voluntarily give, maybe inadvertently
 - Sensors we wear and carry
 - Movement through GPS (Google Maps, anyone)
 - Electronic communication
 - Economic activity



Where's the data

- Data we voluntarily give, by choice
 - Medical and health data
 - Surveys



All told, 2 Zb* per year

**A Zb is 10^{21} bytes*



Can we trust scientific discoveries made using machine learning?

Rice U. expert: Key is creating ML systems that question their own predictions

RICE UNIVERSITY

Statistician: Machine Learning Is Causing A “Crisis in Science”

Many researchers now use machine learning to analyze data. There's just one glaring problem.

Jon Christian

February 18th 2019

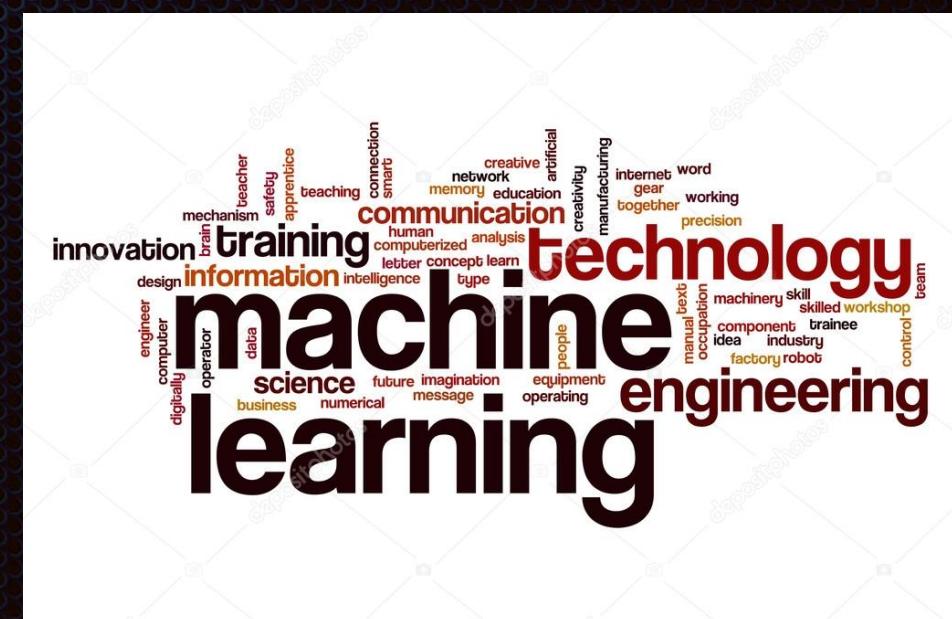
February 16, 2019

Science & Environment

AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh

Science correspondent, BBC News, Washington



open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance.

Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p -values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful. “Negative” is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim

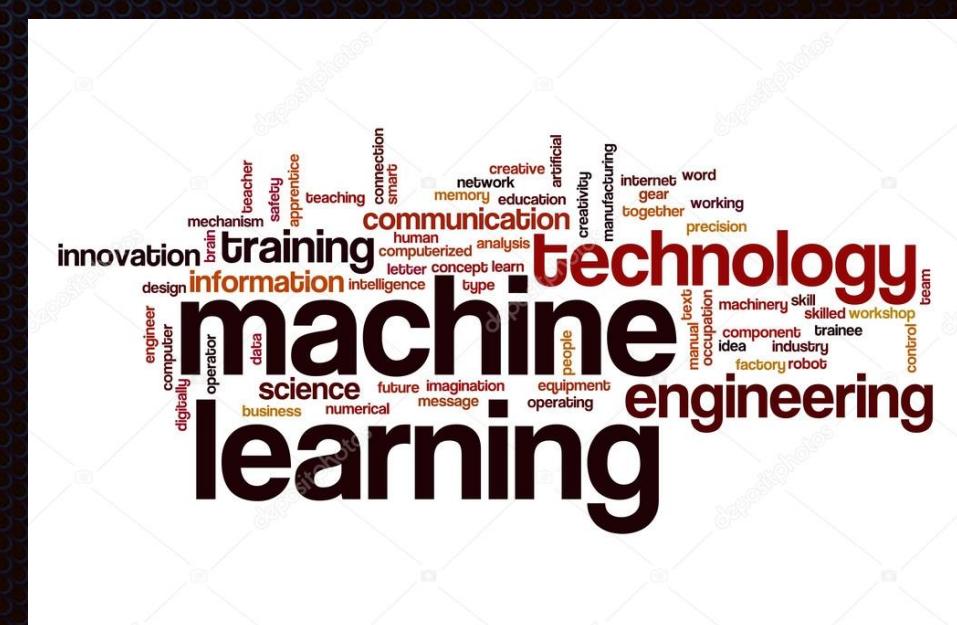
characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R + \beta R + \alpha)$. A research finding is thus

Starkey, Leannick, & P. (2005) Why most published

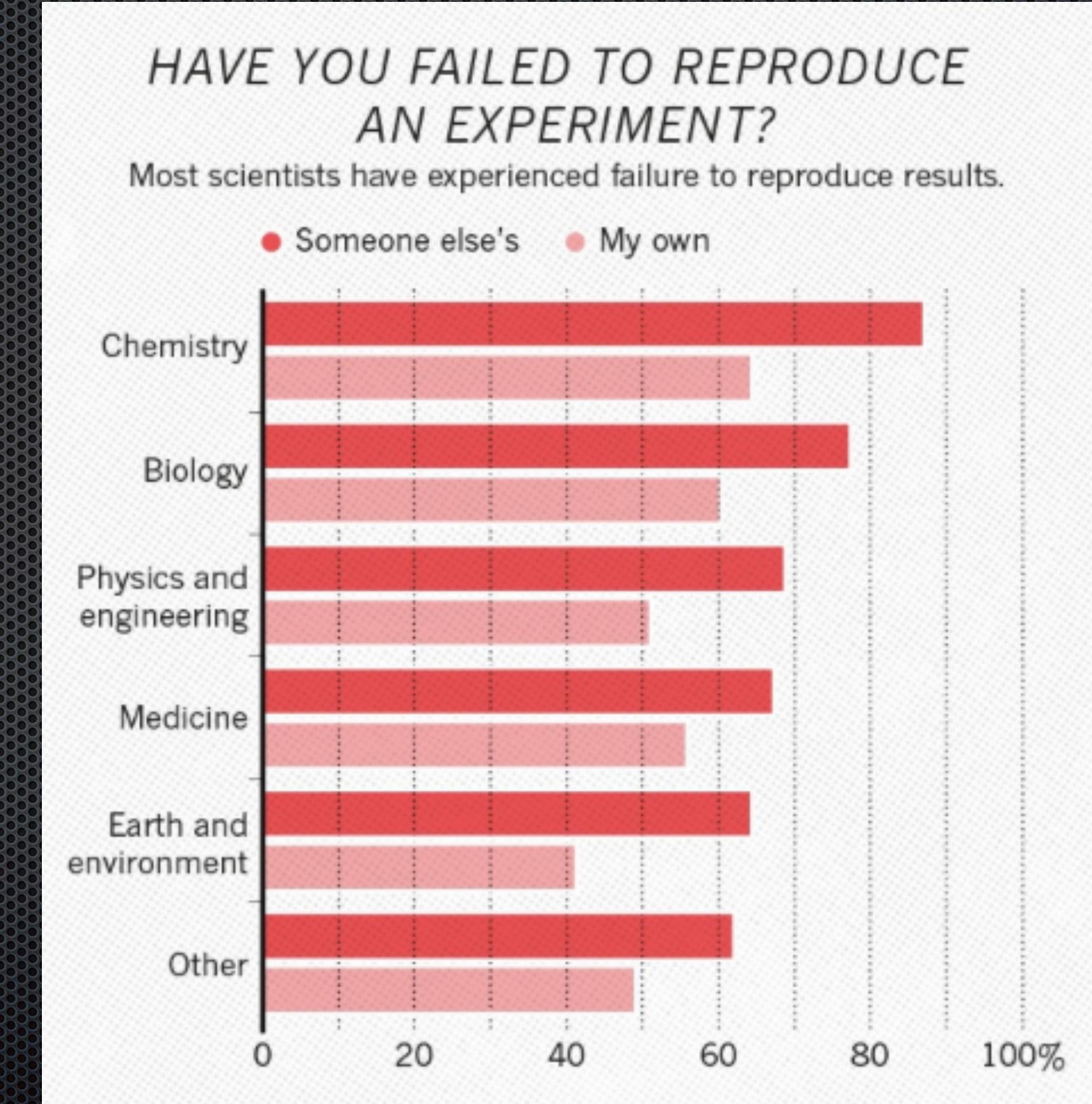
Pharma report:

89% of research findings
are not reproducible

This has huge implications
for biotech and pharma
development, right?!!



IS THERE A REPRODUCIBILITY CRISIS?



Issues with Big Data

- Variability – Inconsistencies in the data
 - Veracity – Poor data quality
 - Complexity – Data management challenges

These are really the more pressing challenges

This is behind most of the problems!



Data is NOT information

- Data is just that, data
 - It only contributes to information when given a context (your question)
 - In that context, the data-derived information can be translated into action
 - Products
 - Services



The question drives the learning

Not the other way around



My personal experience

- You've got data
- I need data and a question
- 80% of my time will be to help you figure out the question
- 80% of the remaining time will be spent getting your data into a usable state for your question

Need to find the right Big Data
for your question

Need to see if your question is
answerable given available data

Biases and ethics

- Our learning is only as good as the training data that is used
 - Training data is often biased
 - Funding biases
 - Historical biases
 - Incompatibilities



Biases and ethics

- Genomic and pathway databases ingest the scientific literature
 - The relative volume of literature in various fields is driven by funding
 - Databases are biased to cancer, heart disease and diabetes
 - Good luck if you have a rarer disease to treat



Biases and ethics

- Historically, minorities had lower chances of getting loans, mortgages, etc.
 - This is reflected in the historical data
 - This data is used to train ML models
 - Now the historical racial bias is codified, since the models will make predictions based on patterns it saw in the training data
 - Similar issue with clinical trials
 - Most studies were done on middle aged white men
 - Do these data apply to women, children, other ethnicities?
 - Huge issue recognized in data ethics



Biases and ethics

- aka, we like shiny new things
 - Historical bioinformatic data incompatible with currently produced data
 - We have lost the opportunity to use this large corpus of data
 - Billions of dollars in international funding
 - We can't tell how these data relate to current findings easily



Ethical use of AI



Responsible AI

We are committed to the advancement of AI driven by ethical principles that put people first.

Forbes

Dec 29, 2019, 01:00am EST | 41,157 views

Ethical Concerns of AI



Kathleen Walch Contributor

COGNITIVE WORLD Contributor Group ⓘ

AI

MIT
Technology
Review

AI Ethics

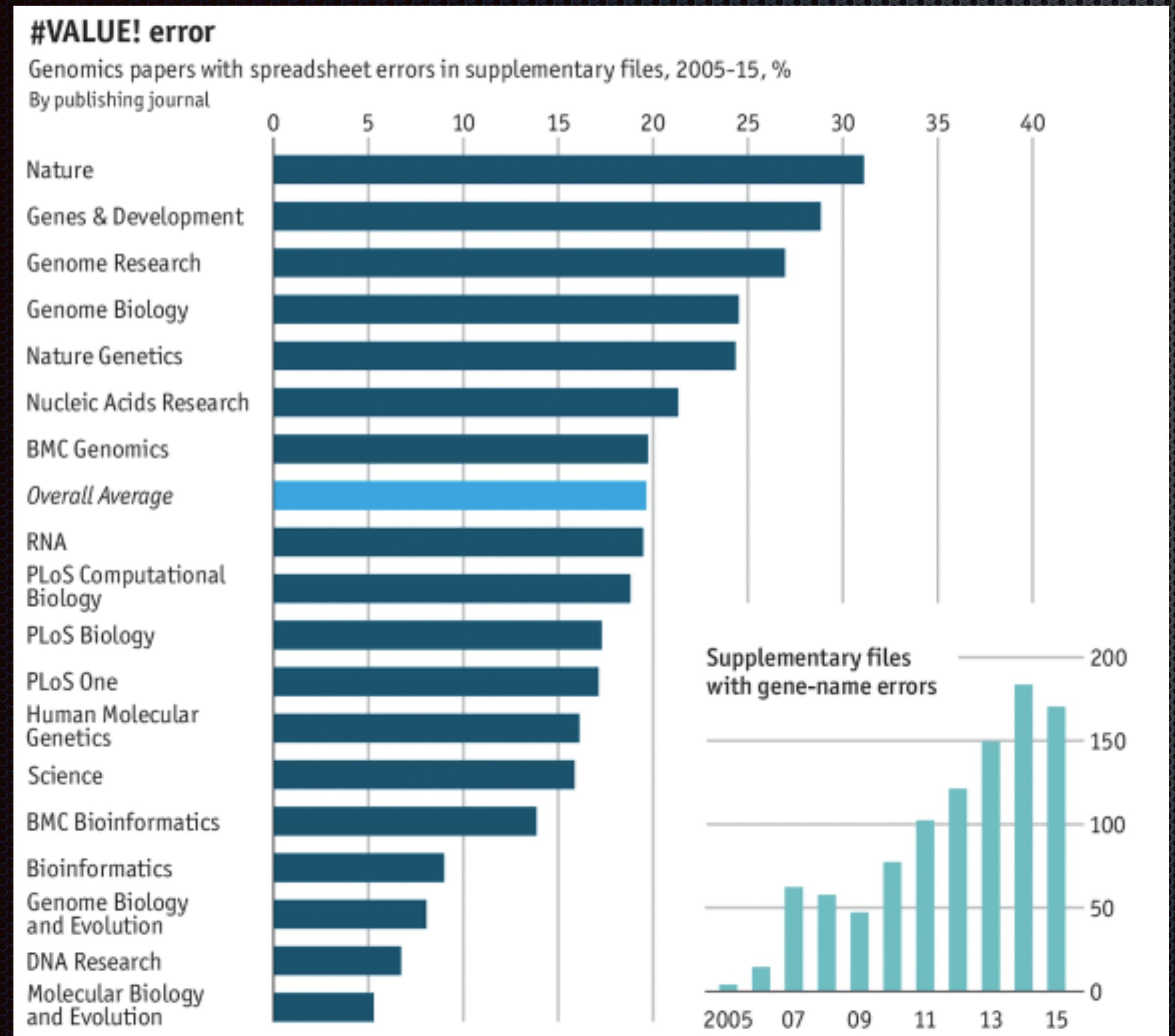
As algorithms play a bigger role in decisions that affect our lives, concerns about bias in artificial intelligence, explainability, and legal responsibility have fueled interest in AI ethics.

Garbage in, garbage out

- Make sure the data you're leveraging for your business and your career is of the highest quality
 - Provenance
 - How it was collected
 - How it is stored



Garbage in, garbage out



Garbage in, garbage out

- Make sure you use the data
 - Appropriately
 - Ethically
 - Really think through and understand your data, your question, and how they relate
 - Don't just dump things into AI just because it's the cool thing



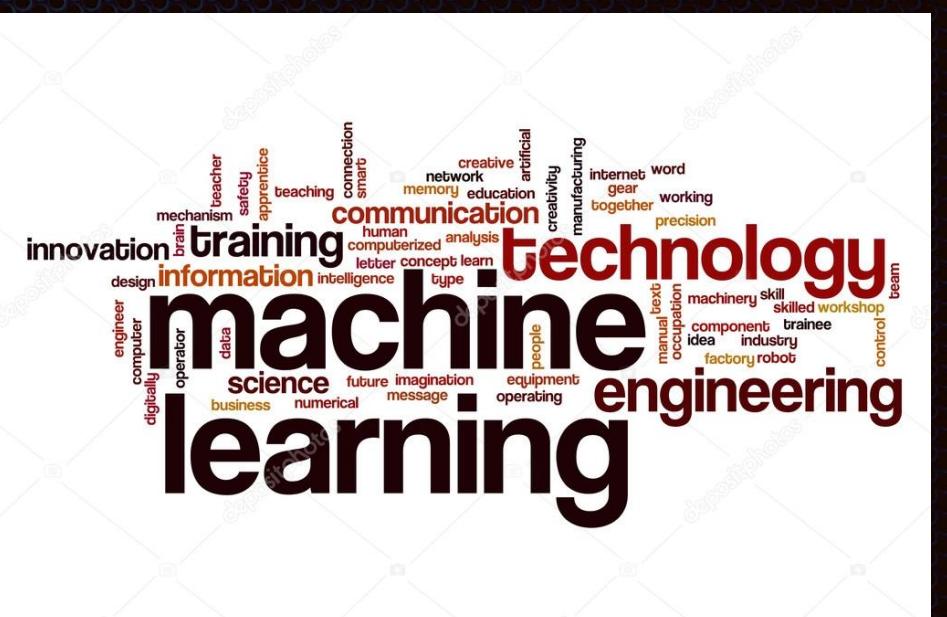
Big Data + ML/AI ≠ Sorceror's Stone

- Big Data can mean Big Noise
 - No signal to be had
 - ML will give spurious, irreproducible answers
 - Dietary studies are notorious for this, as are psych studies



Big Data + ML/AI ≠ Sorceror's Stone

- Finding and assessing rare events is VERY HARD
 - Suppose an event occurs in 1 out of 10,000 people
 - You fit a ML model to this huge database (think Medicare)
 - I use the amazing “Ostrich Method”: Just say no
 - I’ll have an accuracy of 99.99% in this scenario
 - Beat that, you fake intelligence.



The crystal ball

- Storage costs are a huge issue
 - Analytic methods need progress
 - Still leveraging methods from 20 or more years ago
 - A fast moving area



The crystal ball

- Lots of unusable data
 - Lack of compatibility
 - Lack of documentation
- More efficiency and focus needed
 - Data generation
 - Data analysis
 - Data products



Symbiosis



“Begin with the end in mind”

- Find a problem that needs solving
 - Find the tools, data and methods that will help you figure it out
 - Create a diverse team (in skills, talent and background)
 - Work as a team to develop a product that solves it
 - Use evidence-based best practices to convince the market that you've solved it



Contact info

- abhijit@zansors.com
- [@webbedfeet / @Zansors](https://twitter.com/webbedfeet)
- <https://webbedfeet.netlify.com>



Abhijit Dasgupta

Director/Chief Data Scientist

ARAASTAT

Zansors

