Biotechnology
Big Data
Artificial Intelligence

Promises, Dangers, Worries, Gotchas

Abhijit Dasgupta, PhD
Chief Data Scientist
Co-Founder

ZANSORS®
PERSONAL HEALTH ANALYTICS
Know Yourself

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


Biotechnology
DNA
genetic
genetically
desired
cells
organisms
bacteria
altered
biological
gene
also
diseases
grown
products
processed
manufacture
engineering
product
produce
cancer
treatment
living
field
process
extract
Much
plant
inserted
antibodies
recombinant
soybeans
pharmacology
quantities
substances
laboratory
category
qualities
genes
monoclonal
common
corn
shelf
made
industrial
substance
processes
see
problem
splicing
cultured
experimental
hybridomas
oil
techniques
similarly
drugs
attempts
normal
life
development
technological
example
therapy
hormones
chronic
material
pharming
mass
specially
waste
technique
way
systems
Hybrids
animal
general
medicine
growth
insulin
United
exploitation
prolonged
disorders
technology
many
bacterial
use
foods
medically
cloned
Livestock
used
acid
now
antibody-producing
treat
modified
increased
interferon
inserting
produces
called
includes
solution
slicks
fields
pests
involved
Plants
useful
created
States
resistance
nucleic
insert
studied
manipulate
large
manipulation
applications
break
agriculture


Biotechnology
DNA
genetic
desired
cells
organisms
altered
bacteria
genetically
diseases
biological
processed
products
gene
also
product
manufacture
engineering
prolonged disorders
technology
exploitation
United
insulin
many
bacterial
grown
medicine
growth
Hybrids
produce
cancer
animal
general
treatment
living
field
process
extract
antibodies
recombinant
soybeans
pharmacology
quantities
Much
plant
inserted
substances
laboratory
category
qualities
genes
monoclonal
common
corn
shelf
made
industrial
substance
processes
see
problem
splicing
cultured
hybridomas
oil
techniques
similarly
drugs
attempts
normal
life
development
technological
example
therapy
hormones
chronic
material
pharming
mass
experimental
waste
technique
way
systems
specially
acid
now
antibody-producing
use
foods
medically
cloned
Livestock
used
treat
modified
increased
interferon
inserting
produces
called
includes
solution
slicks
fields
pests
involved
Plants
useful
created
States
resistance
nucleic
insert
studied
manipulate
large
manipulation
applications
break
agriculture

Business needs

- Context (why we need it)
- Discovery (Finding it)
- Evidence of effectiveness
- Reproducibility
- Manufacturability
- Quality

Business needs

- Context (why we need it)
- Discovery (Finding it)
- Evidence of effectiveness
- Reproducibility
- Manufacturability
- Quality

Context (why we need it)

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

Context (why we need it)

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

Strengths Characteristics of a business which give it advantages over its competitors	**Weaknesses** Characteristics of a business which make it disadvantageous relative to competitors
Opportunities Elements in a company's external environment that allow it to formulate and implement strategies to increase profitability	**Threats** Elements in the external environment that could endanger the integrity and profitability of the business

Context (why we need it)

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

Context (why we need it)

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

Strengths Characteristics of a business which give it advantages over its competitors	**Weaknesses** Characteristics of a business which make it disadvantageous relative to competitors
Opportunities Elements in a company's external environment that allow it to formulate and implement strategies to increase profitability	**Threats** Elements in the external environment that could endanger the integrity and profitability of the business

Discovery (finding it)

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

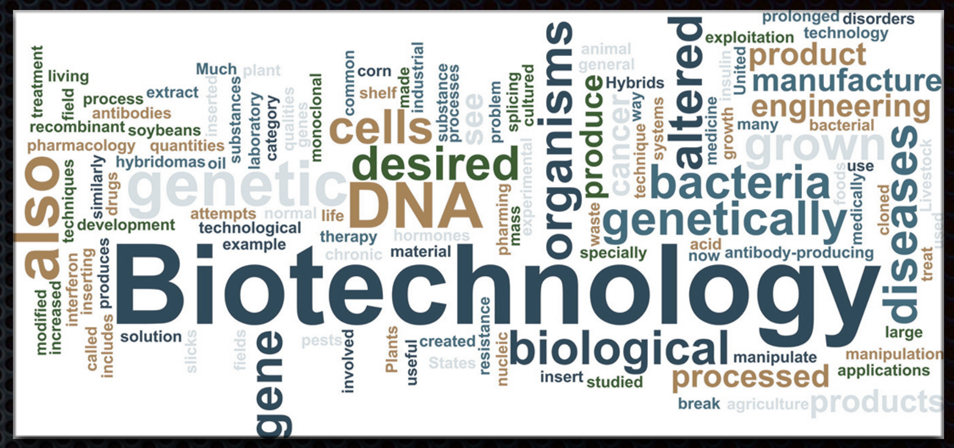
Discovery (finding it)

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

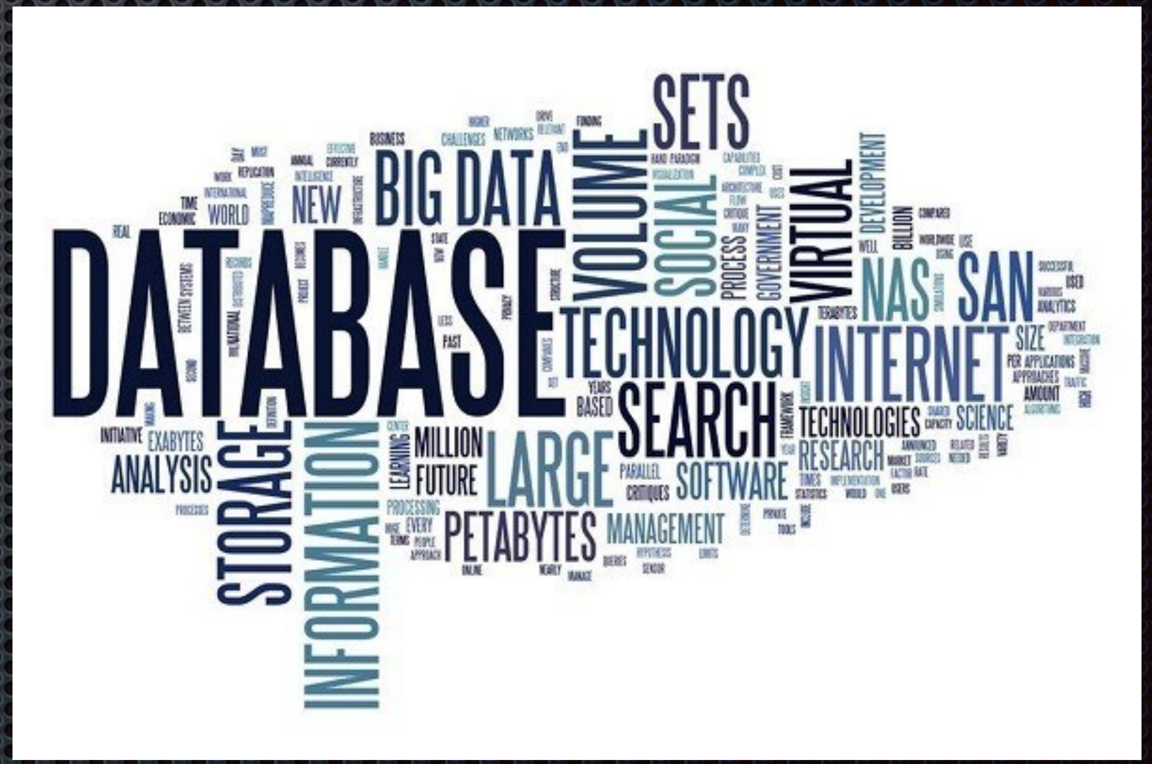
Discovery (finding it)

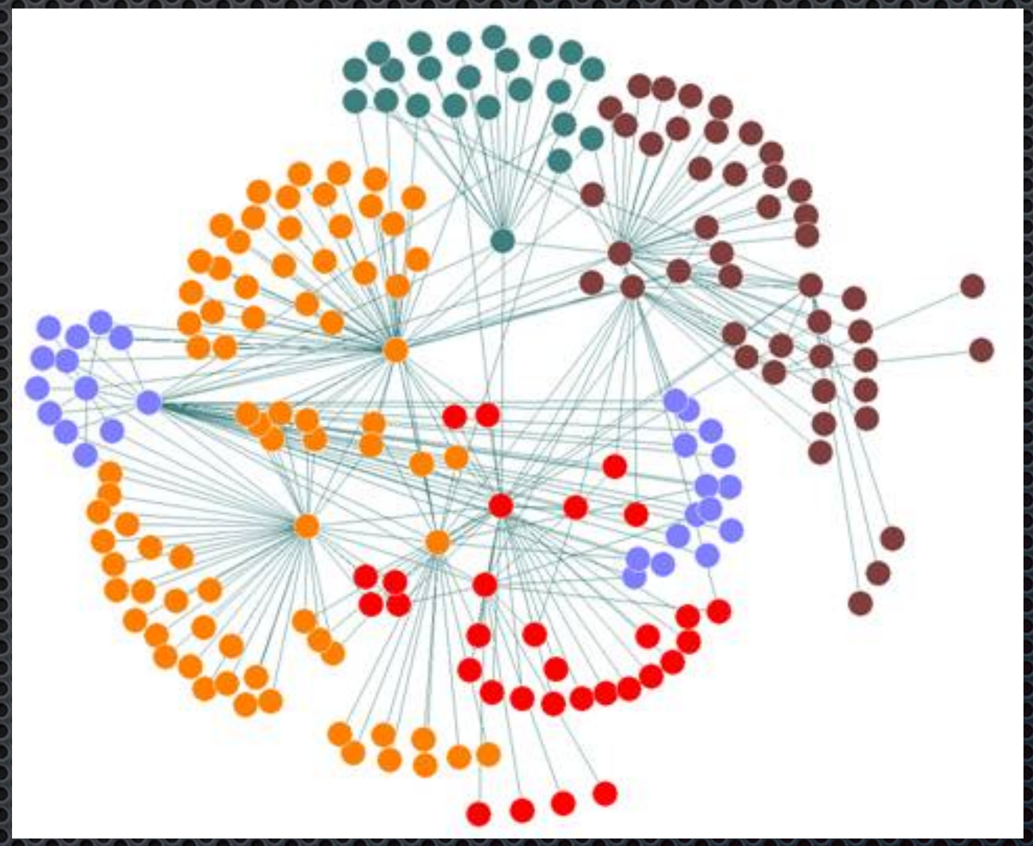
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates

Discovery (finding it)

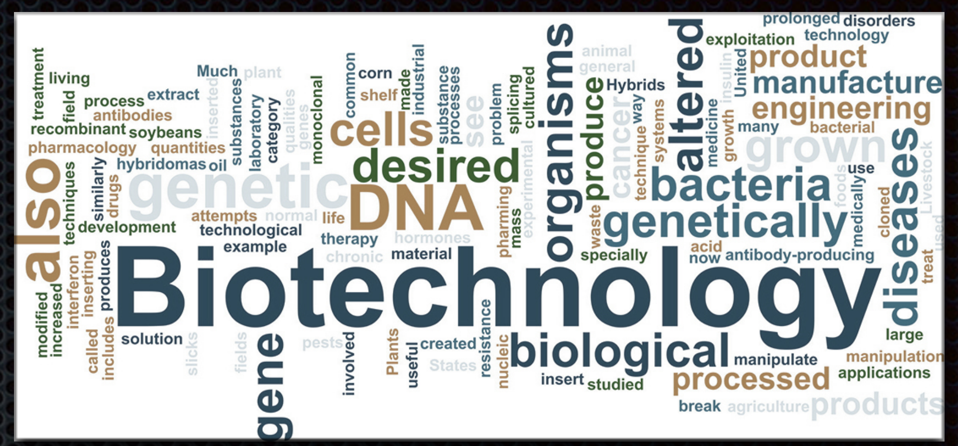


- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates

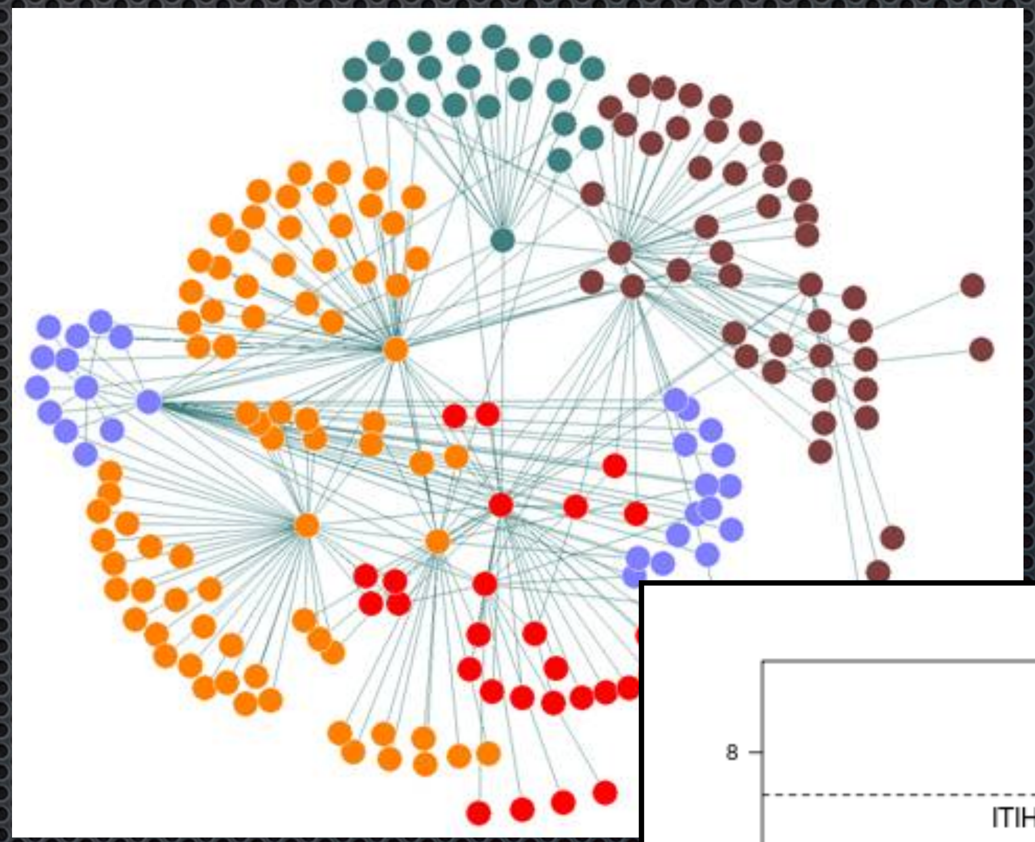


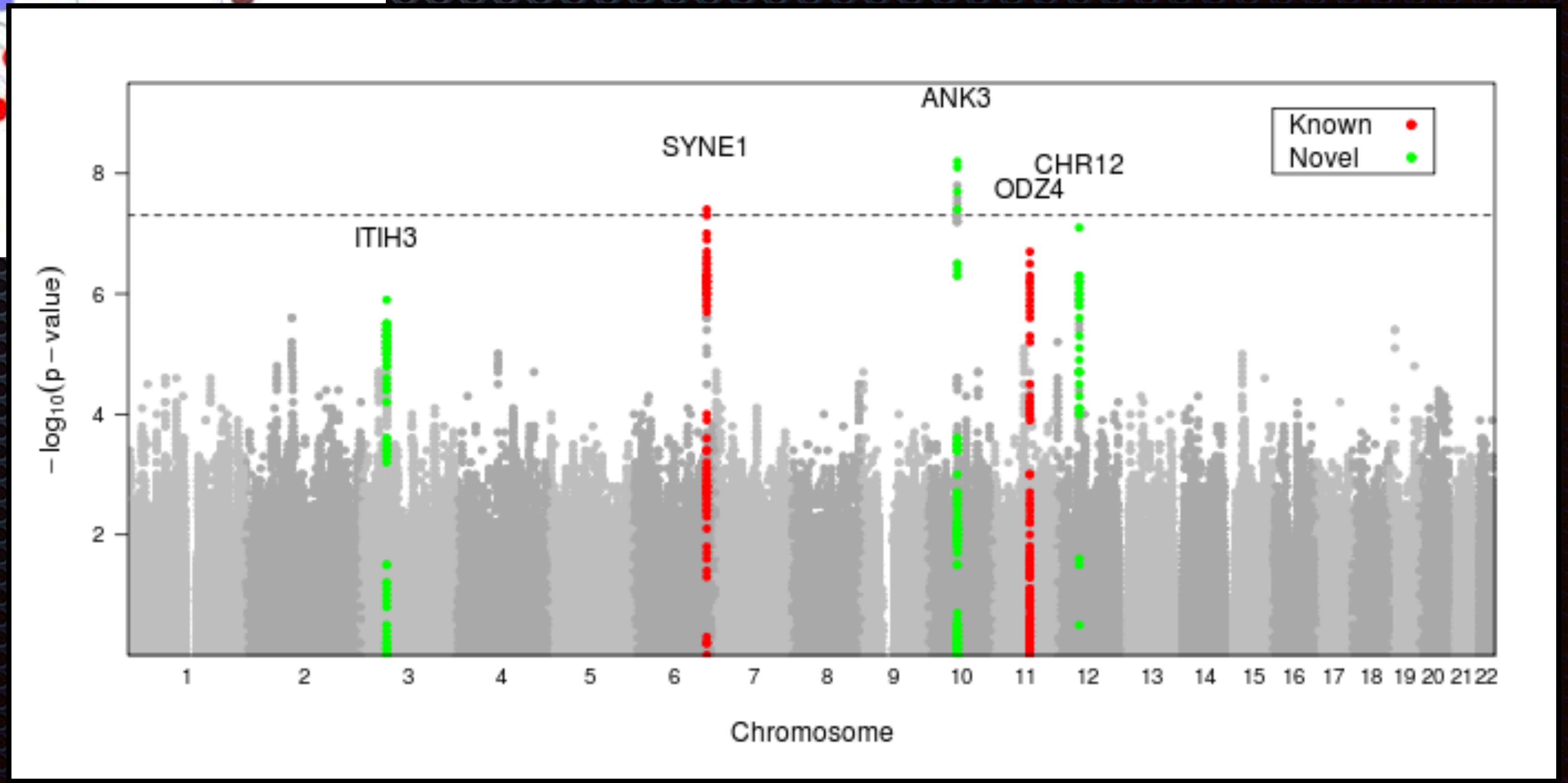


Discovery (finding it)



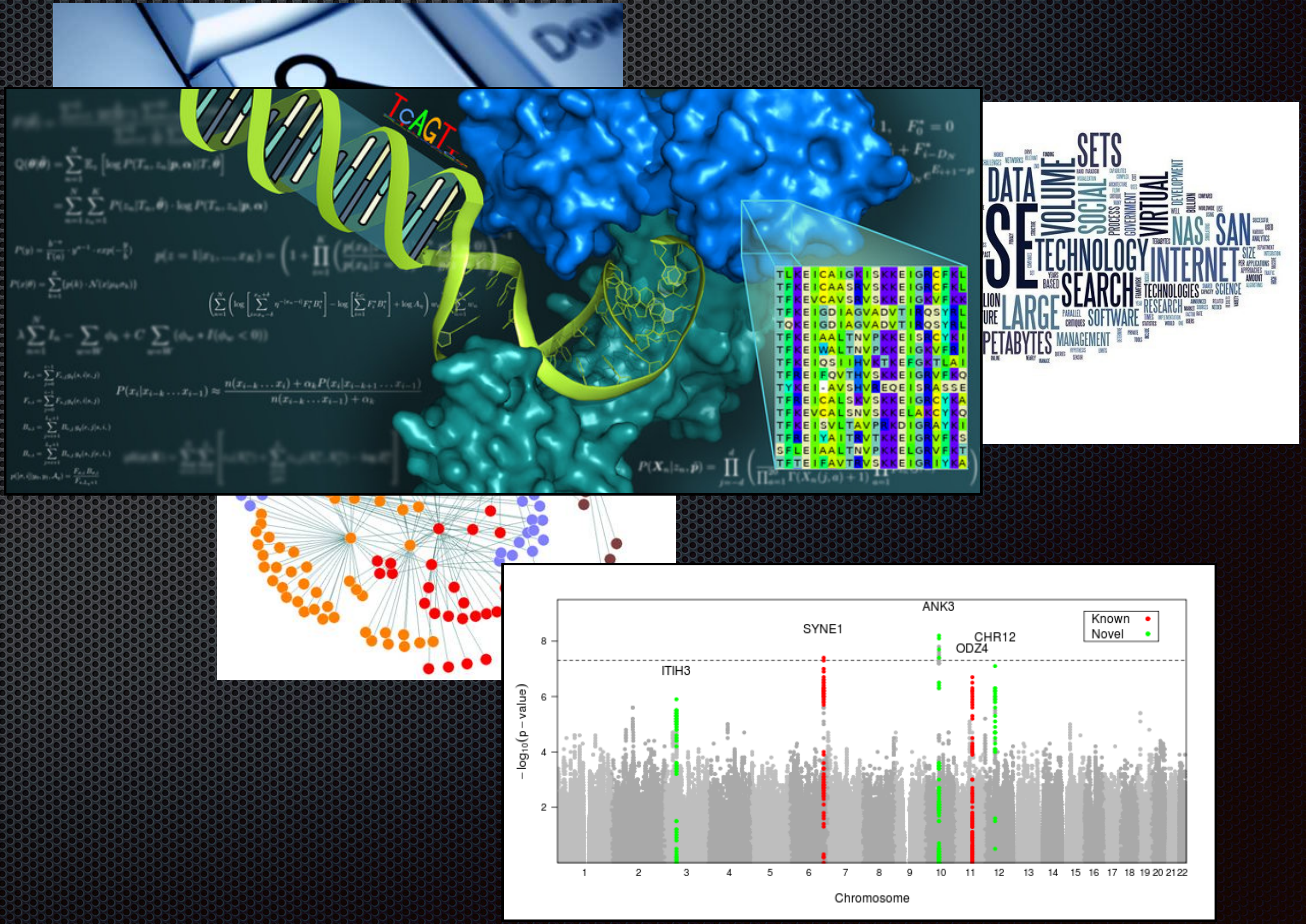
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates



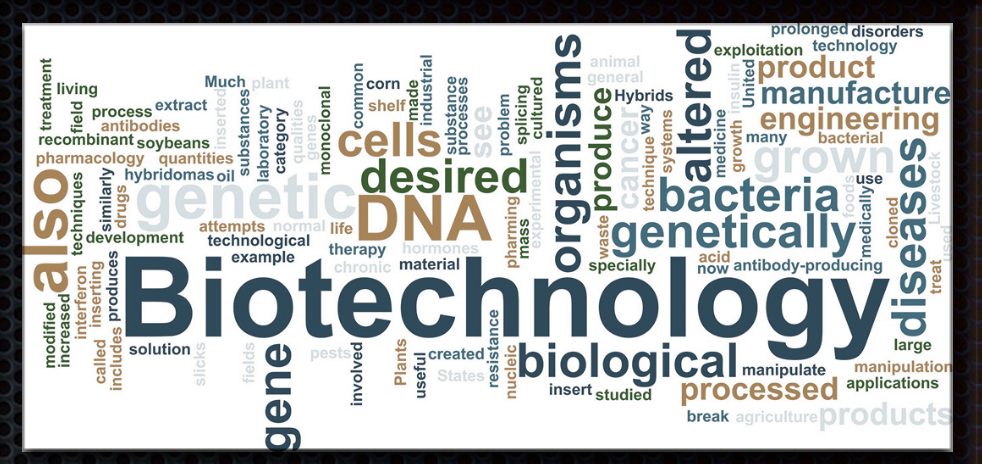


Discovery (finding it)

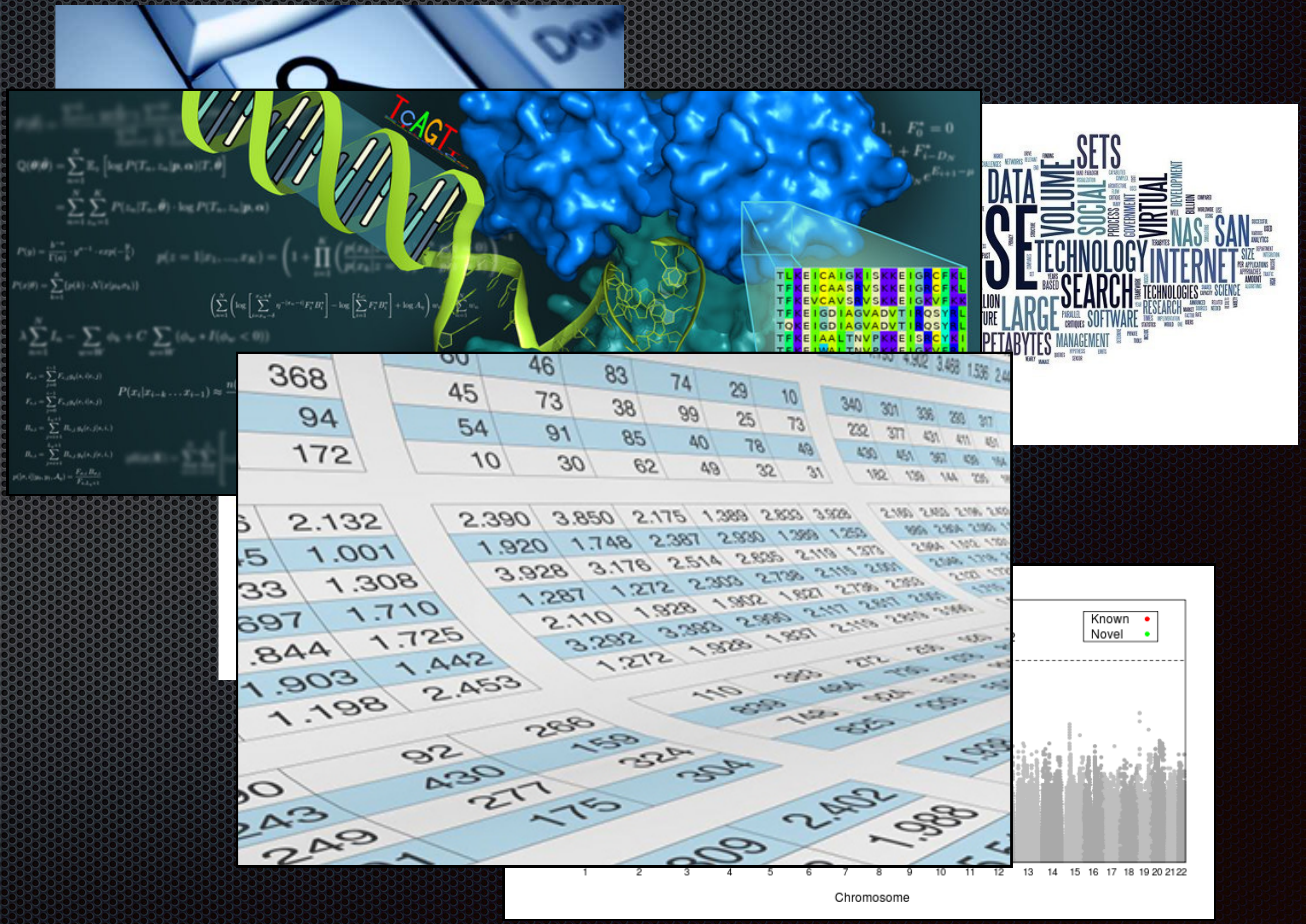
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates



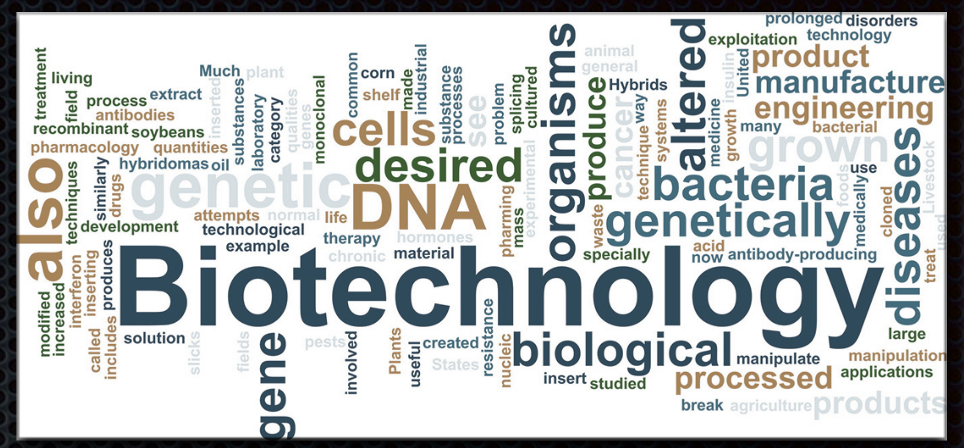
Discovery (finding it)



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



Evidence of effectiveness

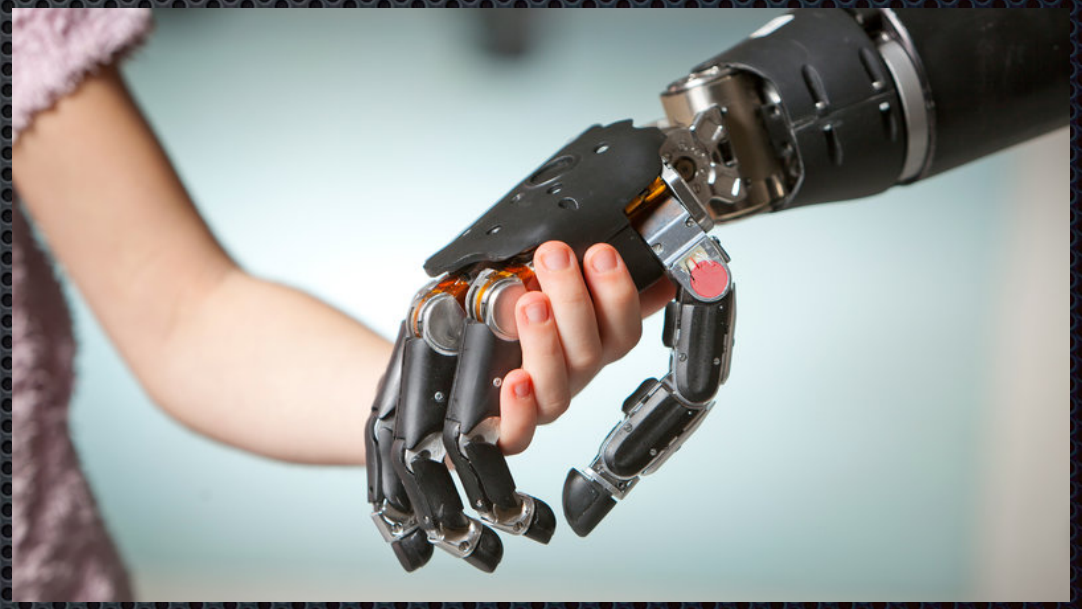


- Laboratory experiments
- Field experiments
- Observational studies
- Randomized trials

Evidence of effectiveness

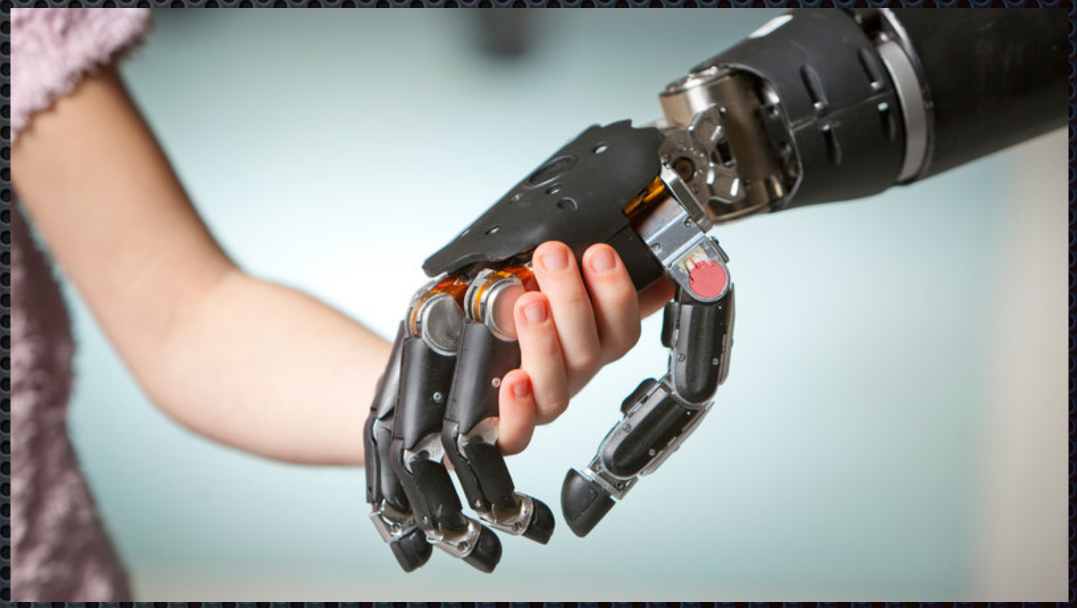
- Laboratory experiments
- Field experiments
- Observational studies
- Randomized trials

Evidence of effectiveness



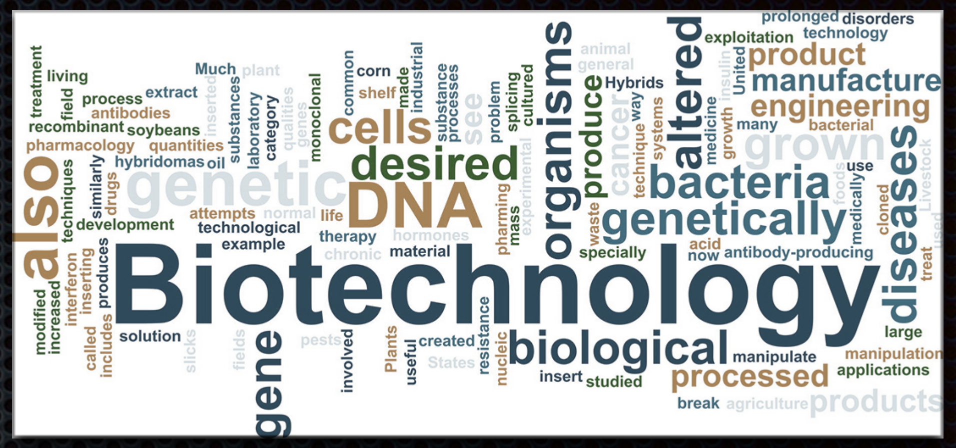
- Laboratory experiments
- Field experiments
- Observational studies
- Randomized trials

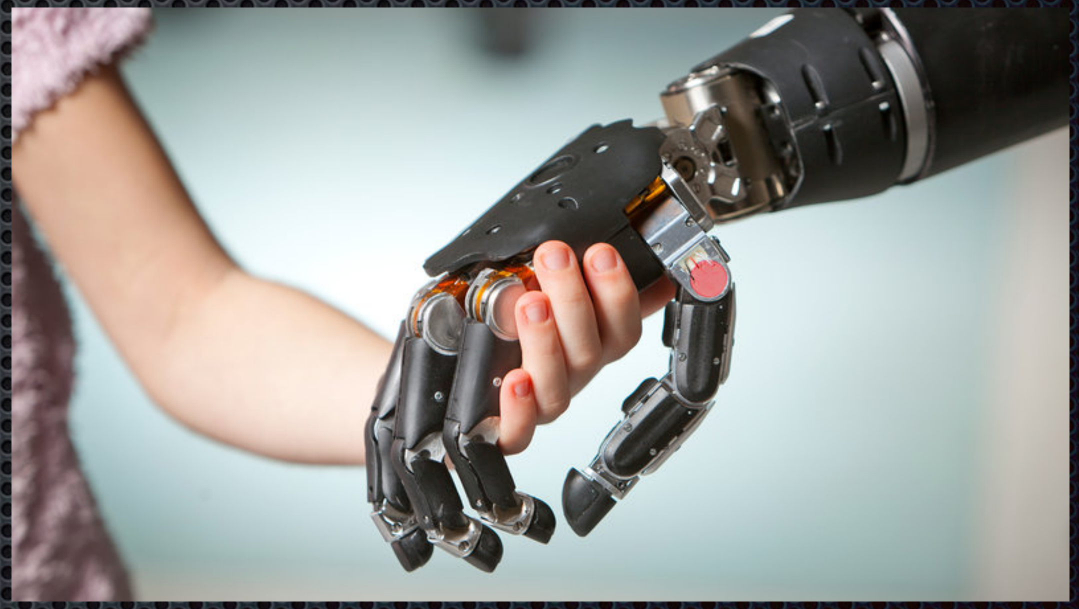
Evidence of effectiveness



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

Evidence of effectiveness



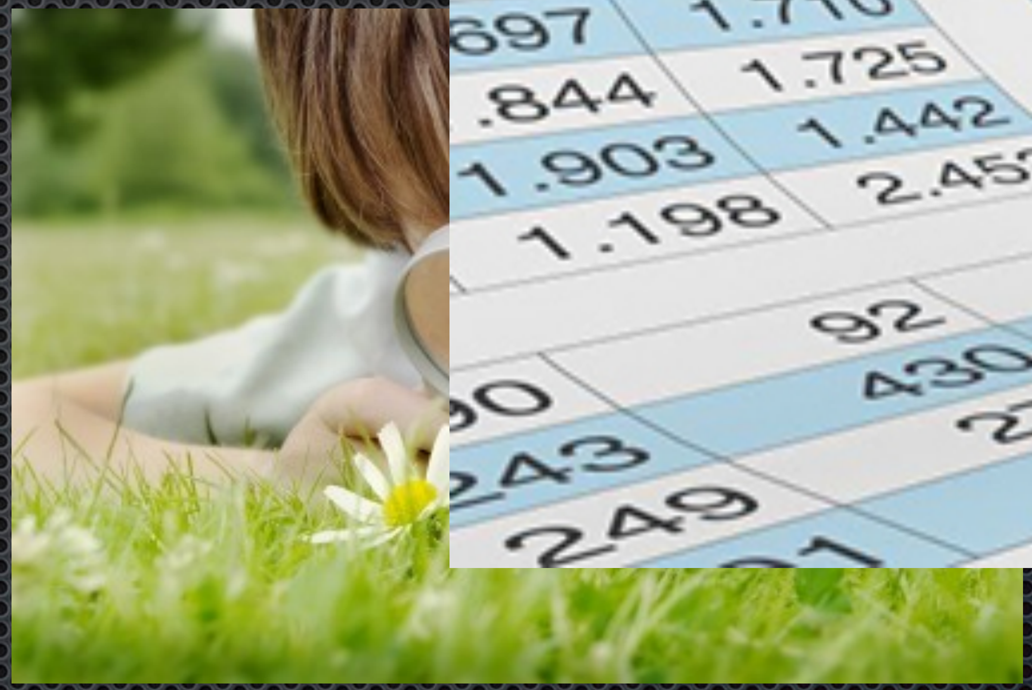


- Laboratory experiments
- Field experiments
- Observational studies
- Randomized trials

Evidence of effectiveness

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




Manufacturability & Quality

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

Manufacturability & Quality

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

Manufacturability & Quality

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

Manufacturability & Quality

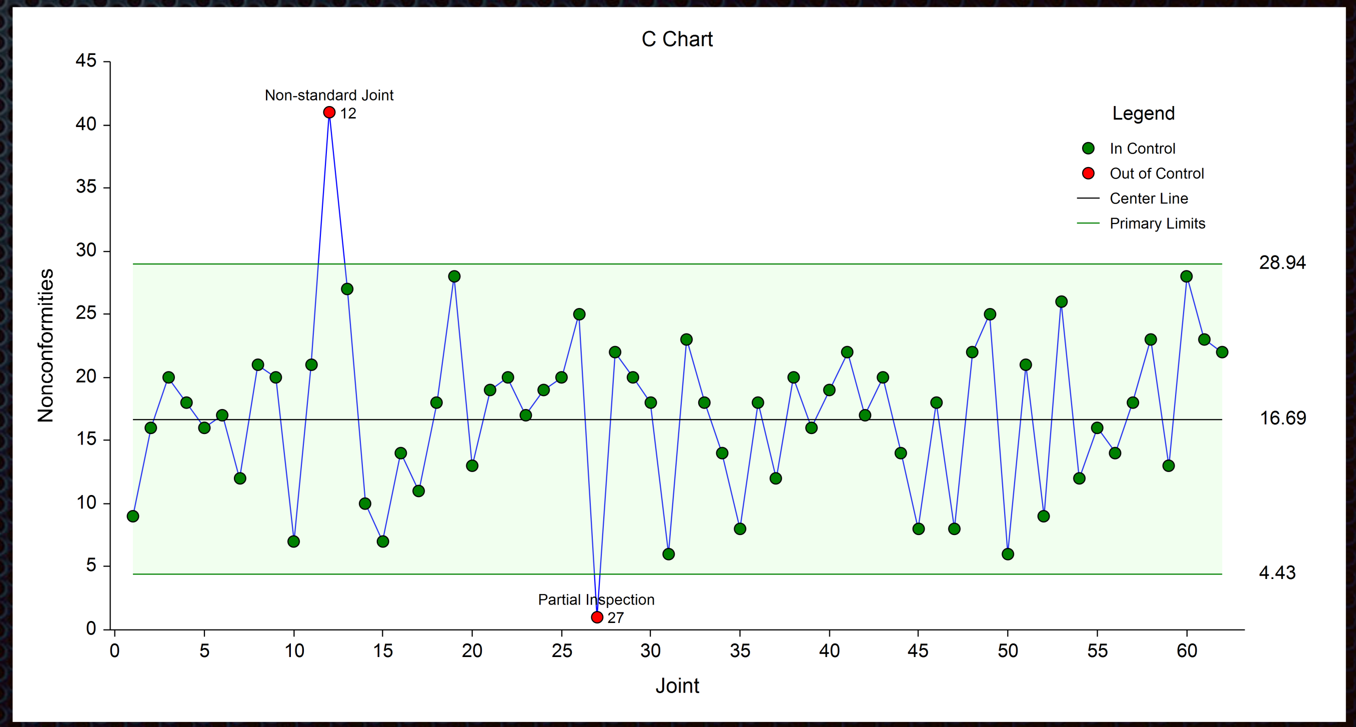
- Translation to a process
- Establishing standards
- Reproducibility
- Quality control / Six Sigma

International Organization for Standardization ISO 13485

Manufacturability & Quality

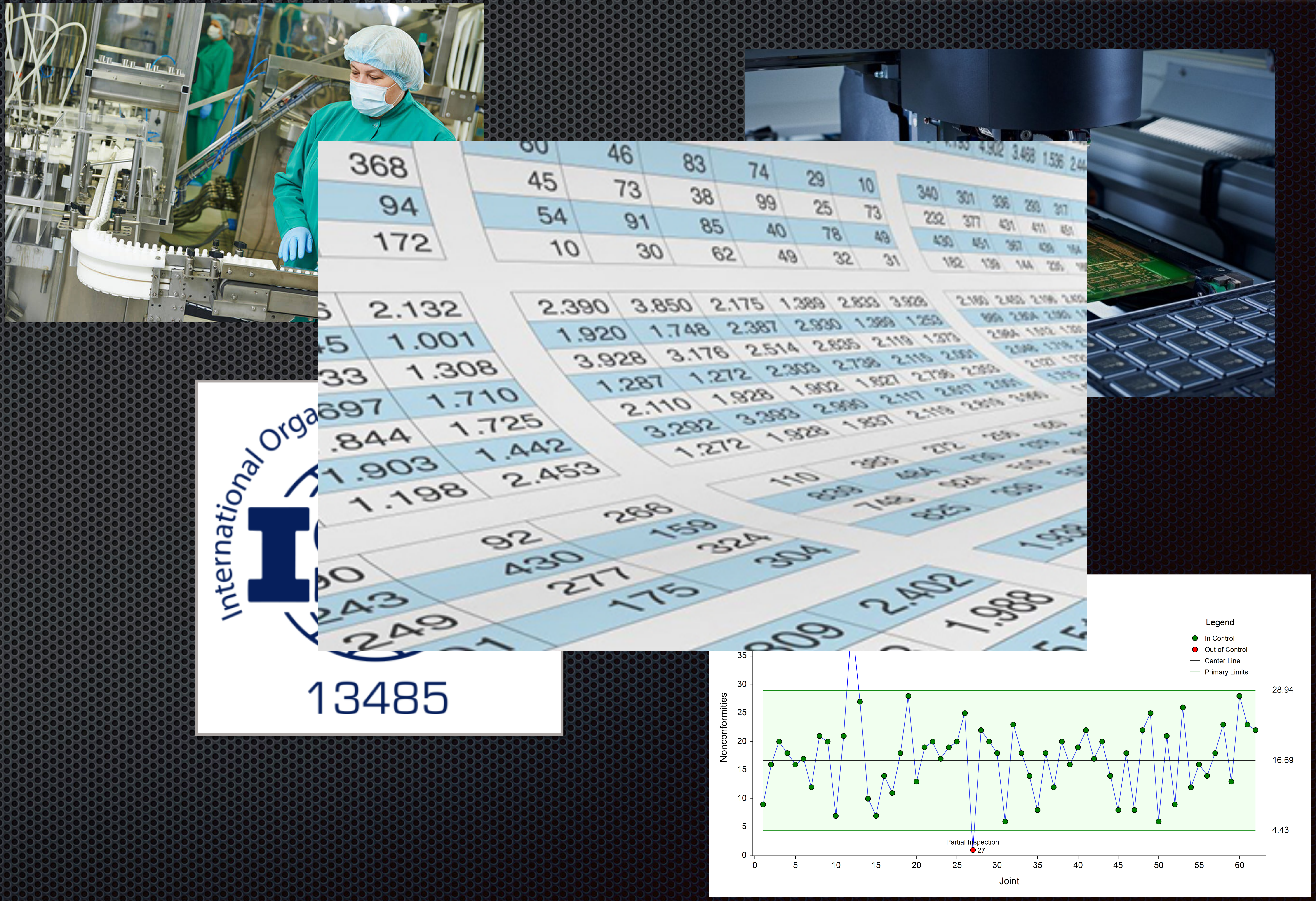
- Translation to a process
- Establishing standards
- Reproducibility
- Quality control / Six Sigma

International Organization for Standardization
ISO
13485



Manufacturability & Quality

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




data
quality
rules
algorithms
constraints
MapReduce
repair
cloud
business
crowd
results
performance
system
clusters
identify
techniques
large
many
organizations
dependencies
functional
evolved
inconsistent
evaluate
important
keywords
real
together
specific
scalability
workflows
workflow
temporal
keyword
distributed
process
semantics
file
output
work
detection
accurately
discovery
improved
languages
consider
model
recognize
become
execution
entity
analysis
database
manage
application
express
repaired
Stringer
underlying
case
appear
role
minimal
study
set
cost
cleaning
provides
domain
input
duplication
constraint
CFDs
several
used
new
job
sometimes
meant
maintenance
need
manages
time
using
potentially
jobs
designed
FDs
contextual
within
duplicates
due
repairs
datasets
dirty
well
millions
non-conformant
search
evaluation
often
arises
tool
make
join
topics
modern
part
years
query
remain
accuracy
continuous
clustering
blogs
rule
experimental
use
terms
paper
scale
scalable
hold
change
ReStore
inconsistencies
efficiently
records
values
approach
intermediate
Dirty
synthetic
understanding
known
evolves
evolve
reuse
enforced
play
approximate
errors
associated
changes
respect
number
complex
frequently
capture


data
quality
rules
algorithms
constraints
MapReduce
repair
business
cloud
crowd
results
clusters
system
performance
identify
techniques
large
many
dependencies
functional
evolved
important
keywords
real
together
specific
scalability
workflows
rule
blogs
use
terms
paper
scale
scalable
hold
time
using
jobs
often
evaluation
temporal
keyword
distributed
database
application
process
semantics
file
work
output
detection
organizations
new
set
study
tool
dirty
used
CFDs
constraint
domain
input
inconsistent
evaluate
entity
records
approach
intermediate
reuse
evolve
known
evolves
errors
associated
inconsistencies
datasets
workflow


data
quality
rules
algorithms
constraints
MapReduce
business
cloud
repair
crowd
results
clusters
system
performance
organizations
identify
techniques
dependencies
functional
evolved
many
large
using
time
new
CFDs
used
identify
dirty
tool
records
approach
evolves
evolve
reuse
known
associated
errors
inconsistencies
intermediate
workflows
workflow
rule
blogs
use
terms
paper
scale
scalable
scalability
specific
together
real
keywords
important
jobs
evaluation
often
inconsistent
evaluate
temporal
keyword
distributed
database
application
entity
process
output
detection
work
semantics
file
domain
input
constraint
study
set
datasets
hold
capture
model
consider
languages
recognize
become
execution
analysis
discovery
Stringer
express
repaired
underlying
case
appear
role
minimal
designed
duplicates
within
due
repairs
change
experimental
clustering
synthetic
Dirty
play
enforced
respect
changes
values
efficiently
understanding
approximate
search
millions
arises
remain
join
make
topics
part
modern
non-conformant
years
query
well
job
several
sometimes
meant
maintenance
need
manages
potentially
cost
number
accuracy
continuous

data
quality
rules
algorithms
constraints
performance
organizations
system
clusters
detection
output
process
semantics
file
work
temporal
keyword
distributed
database
application
entity
inconsistent
evaluate
dependencies
functional
repair
evolved
MapReduce
large
using
study
business
cloud
crowd
many
jobs
identify
techniques
time
new
domain
input
constraint
CFDs
used
dirty
tool
important
keywords
real
results
together
specific
scalability
scalable
workflows
workflow
blogs
rule
use
terms
paper
scale
hold
datasets
inconsistencies
records
approach
intermediate
known
evolves
evolve
reuse
errors
associated
set
well
often
evaluation
Stringer
express
repaired
underlying
case
appear
role
minimal
maintenance
need
sometimes
meant
manages
search
within
duplicates
due
join
make
topics
part
modern
non-conformant
query
arises
remain
millions
accuracy
continuous
clustering
experimental
synthetic
Dirty
play
enforced
respect
changes
values
efficiently
understanding
approximate
consider
languages
discovery
model
recognize
become
execution
analysis
capture
number
cost
complex
several
potentially
designed
contextual
FDs
change
repairs
ReStore
job
Hence

Data drives

Research

Development

Production

Marketing

Sales

Revisions

Central issue for data-driven business

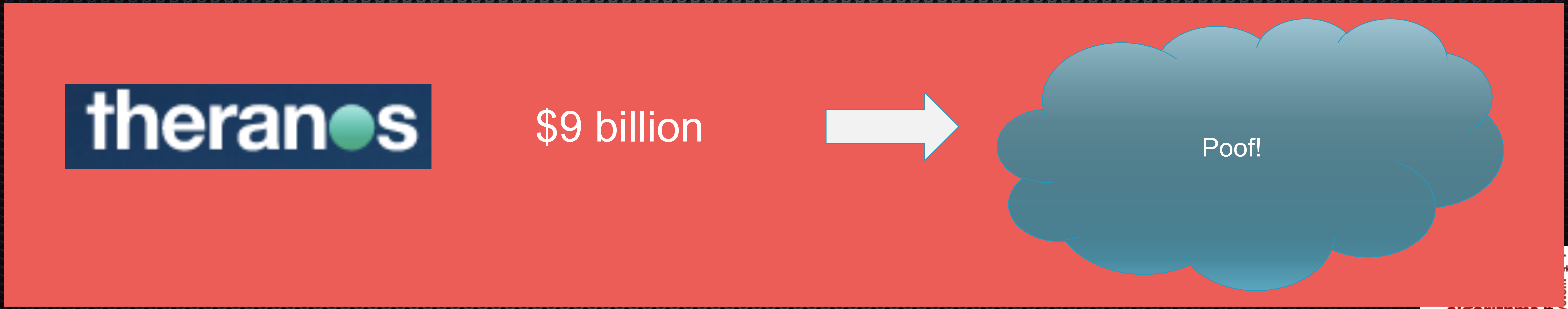
- The data must be reliable and valid
- Repeating the experiment must yield similar results (reliability)
- We're measuring what we think we're measuring (internal validity)
- The results can be generalized to larger populations (external validity)

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

- 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

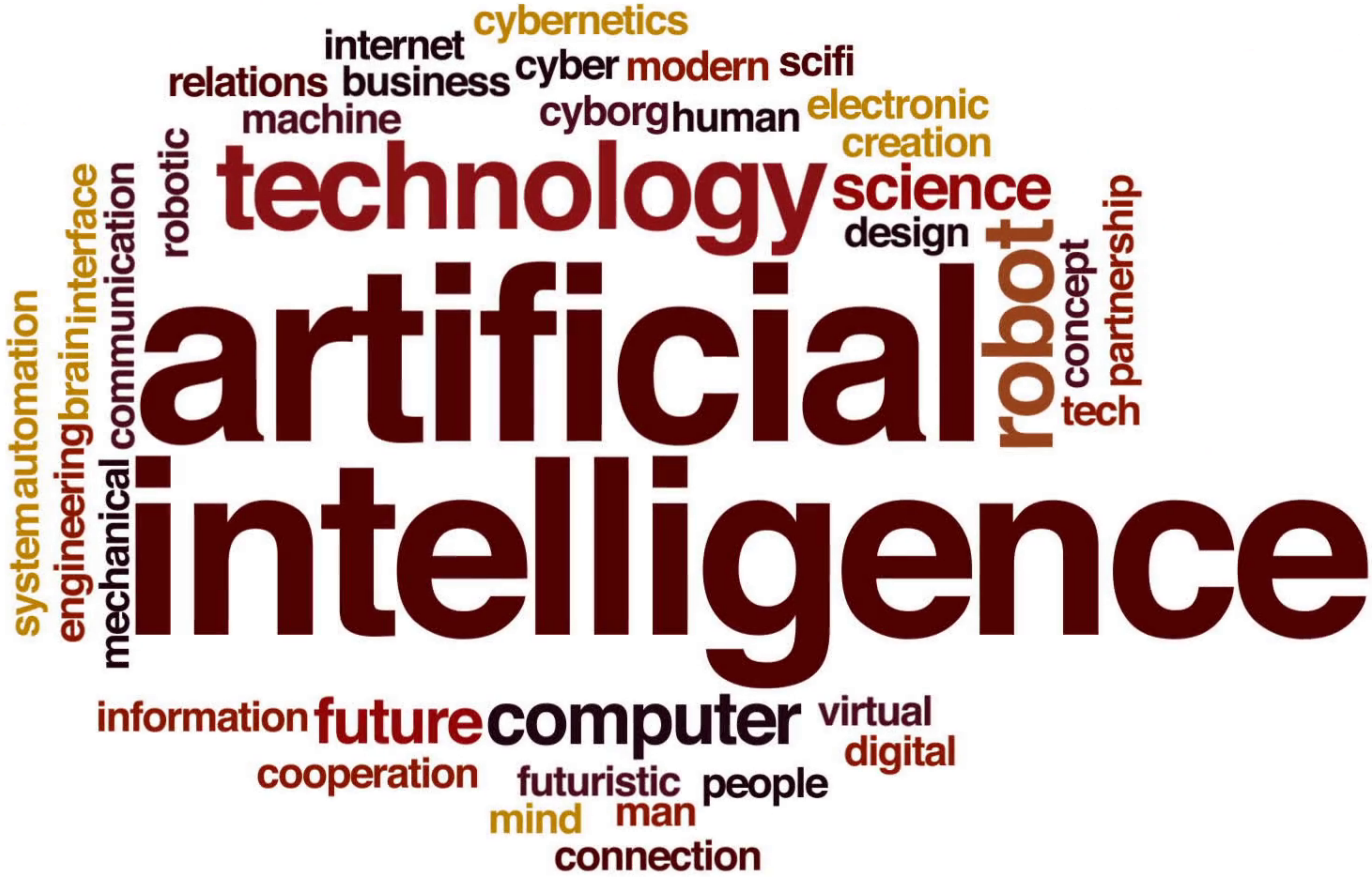
- It doesn't matter how much data you can get
- It matters how you got it
  - Experimental design
  - Ethics
  - Provenance

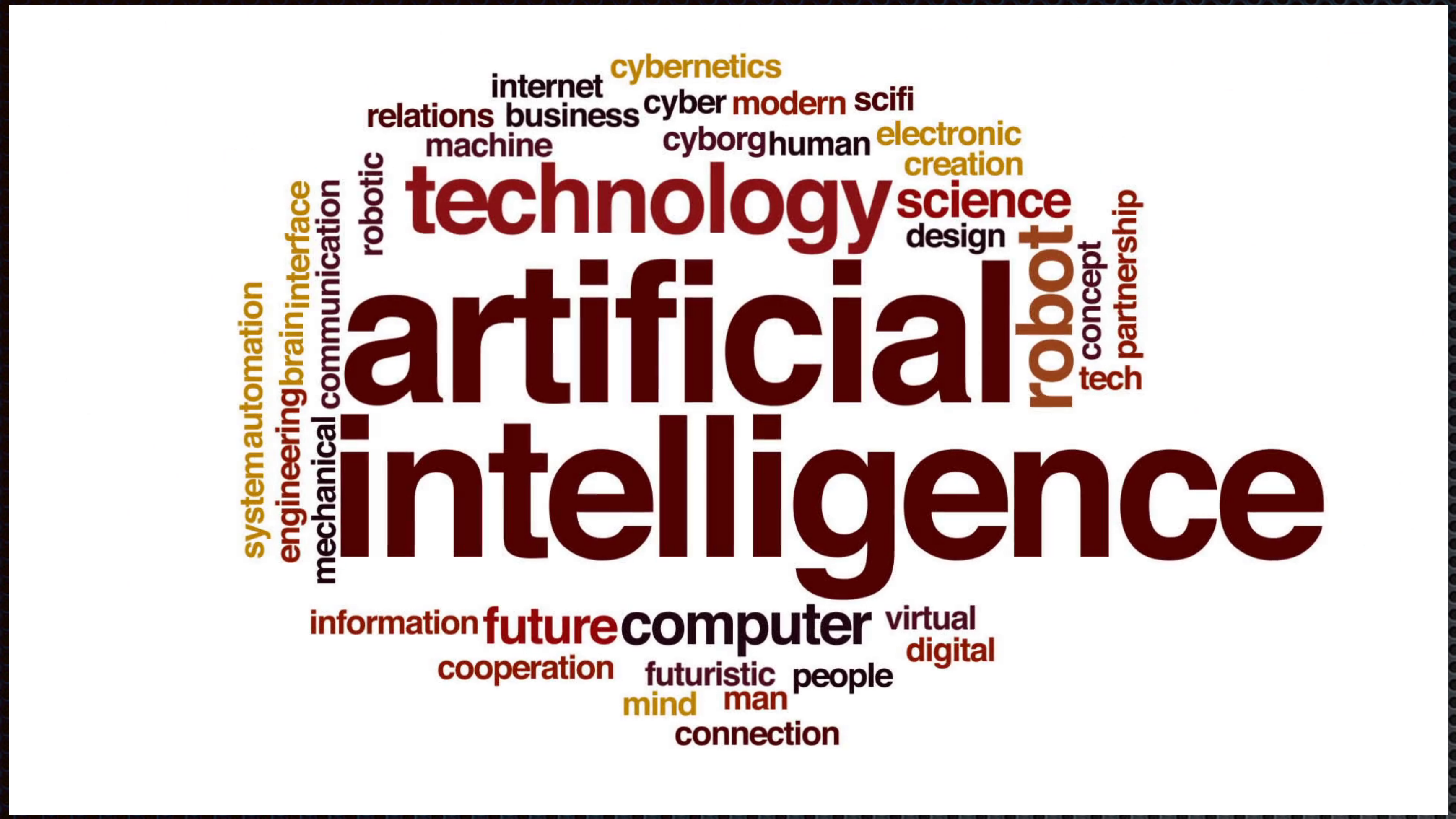
Central issue for data-driven business

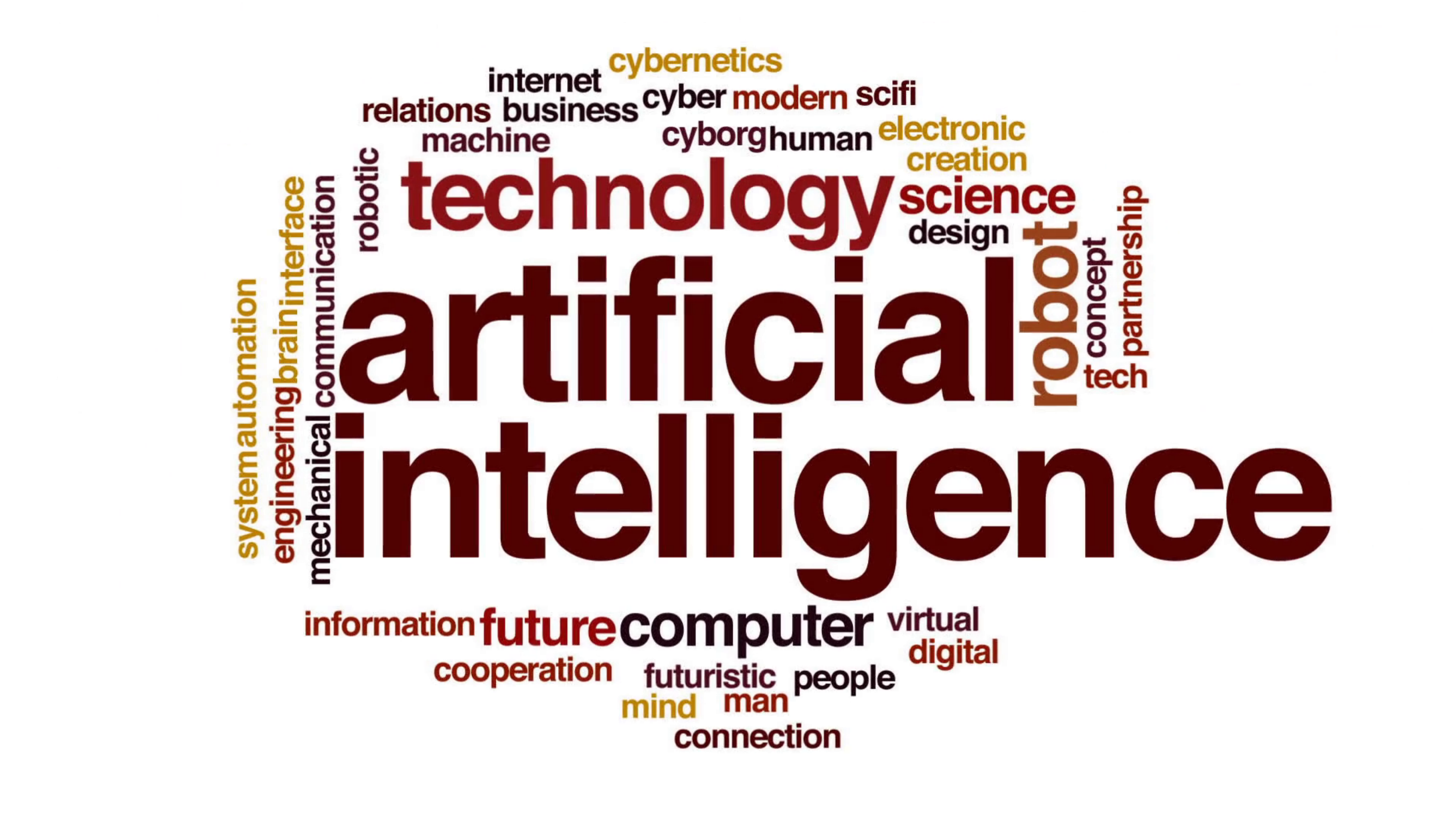
- It doesn't matter how much data you can get
- It matters how you use it
  - Analytics
  - Translation to business practice
  - Follow-up

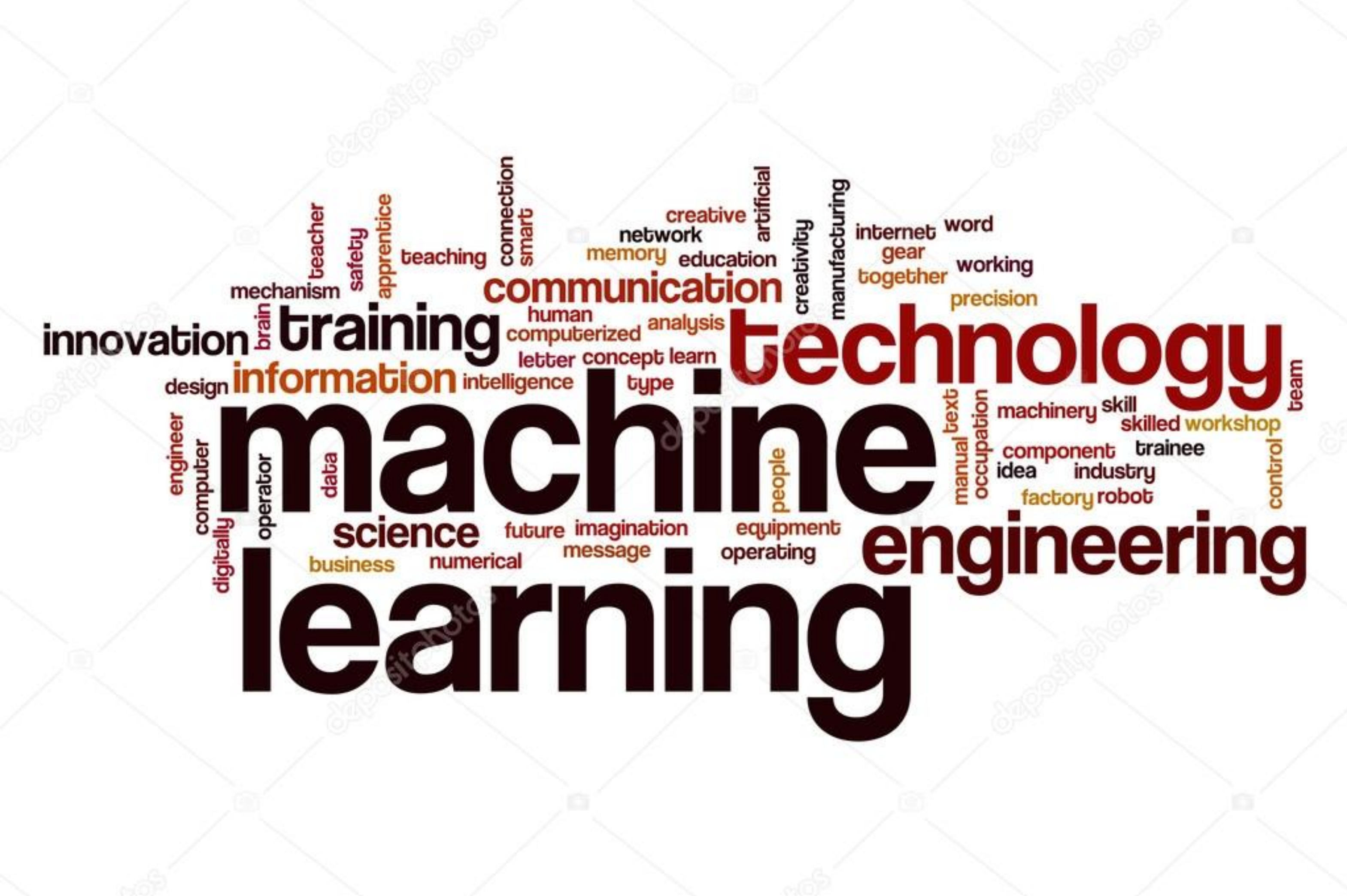
Central issue for data-driven business

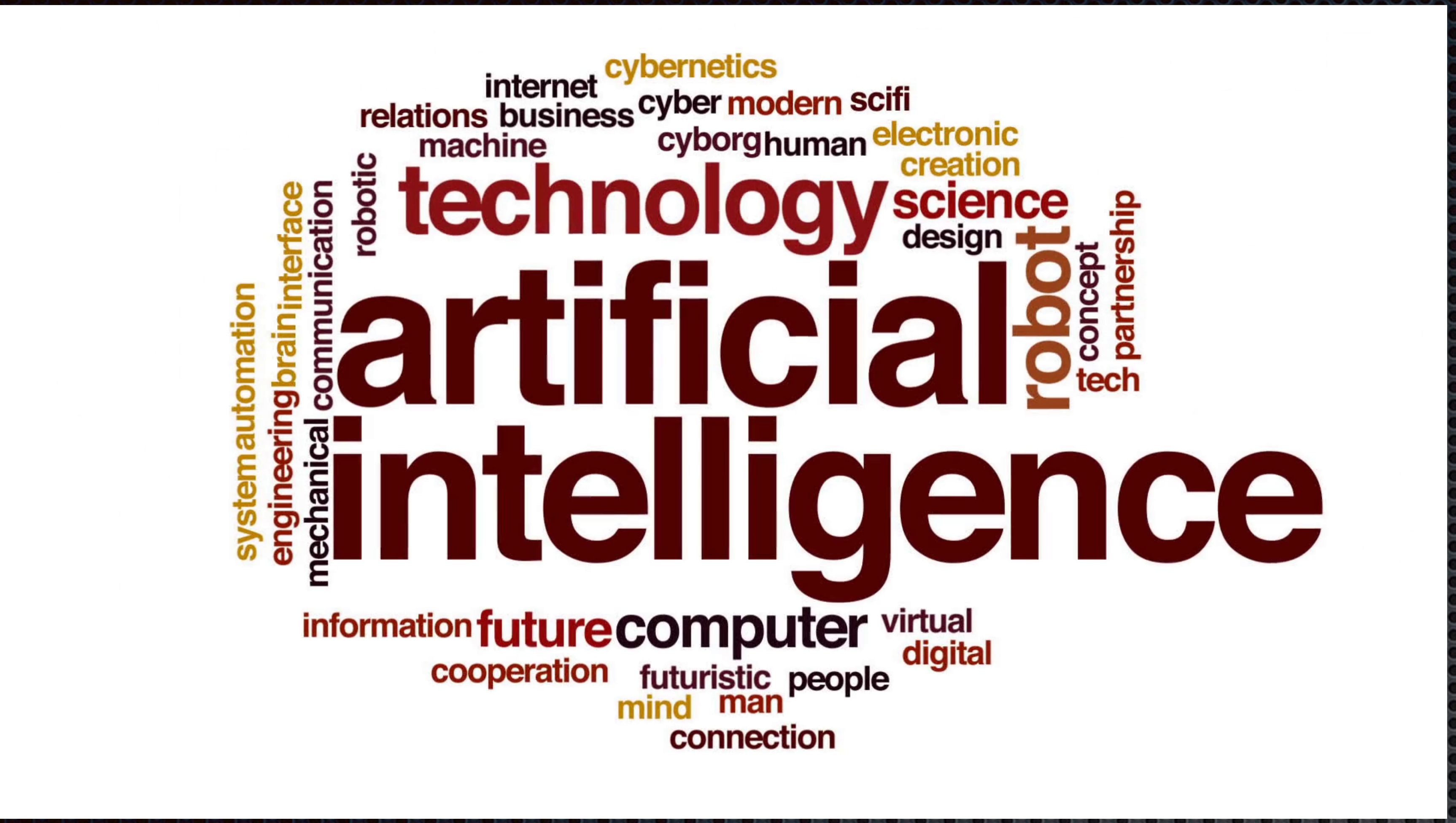
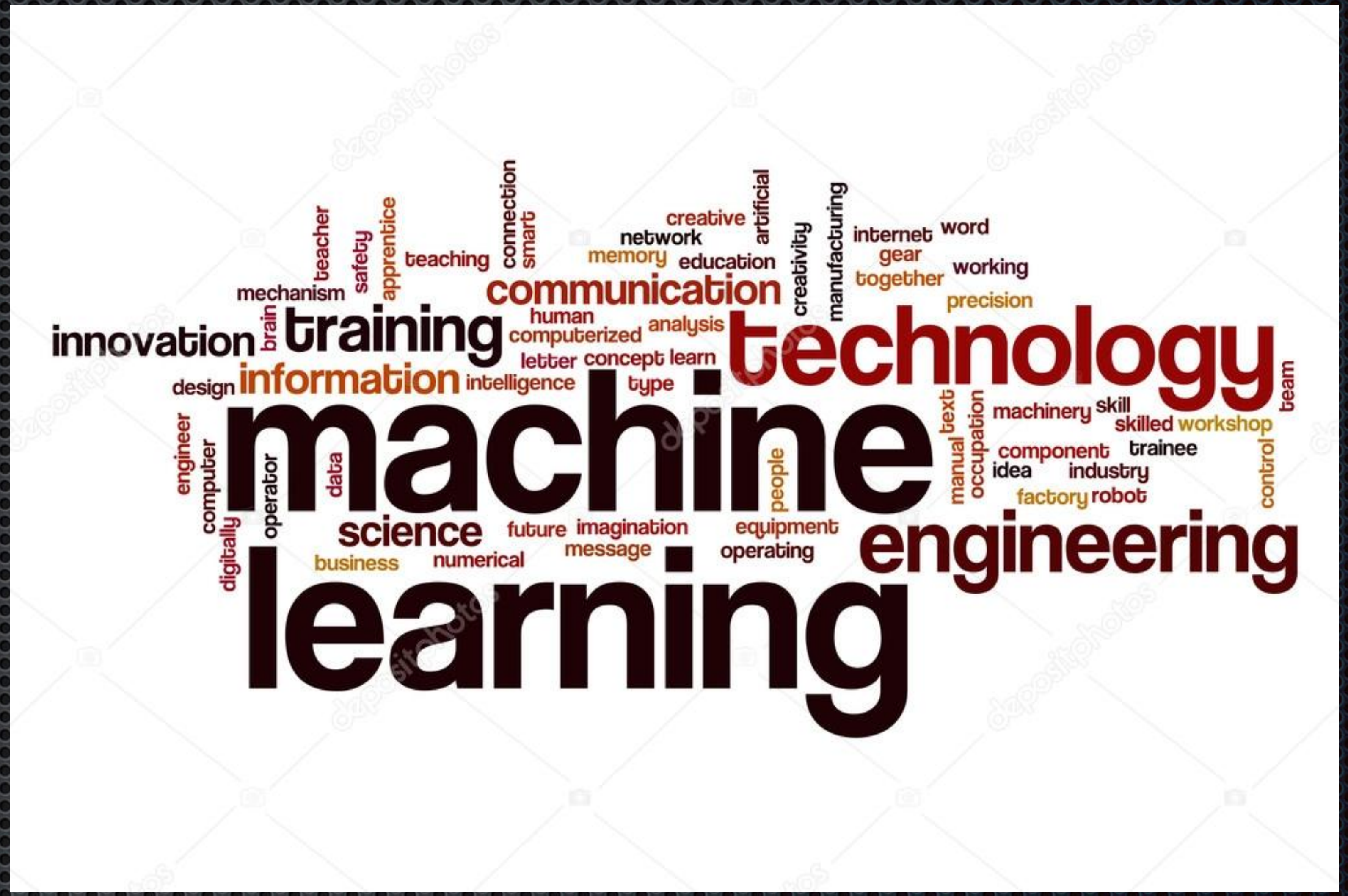
- It doesn't matter how much data you can get
- It matters how you use it
  - Analytics
  - Translation to business practice
  - Follow-up


cybernetics
internet
cyber
modern
scifi
relations
business
machine
cyborg
human
electronic
creation
robotic
technology
science
design
system
automation
engineering
brain
interface
communication
mechanical
artificial
intelligence
robot
concept
partnership
tech
information
future
computer
virtual
digital
cooperation
futuristic
people
mind
man
connection


cybernetics
internet
relations
business
cyber
modern
scifi
machine
cyborg
human
electronic
creation
robotic
technology
science
design
interface
communication
brain
automation
system
engineering
mechanical
artificial
robot
concept
partnership
tech
intelligence
information
future
computer
virtual
digital
cooperation
futuristic
people
mind
man
connection

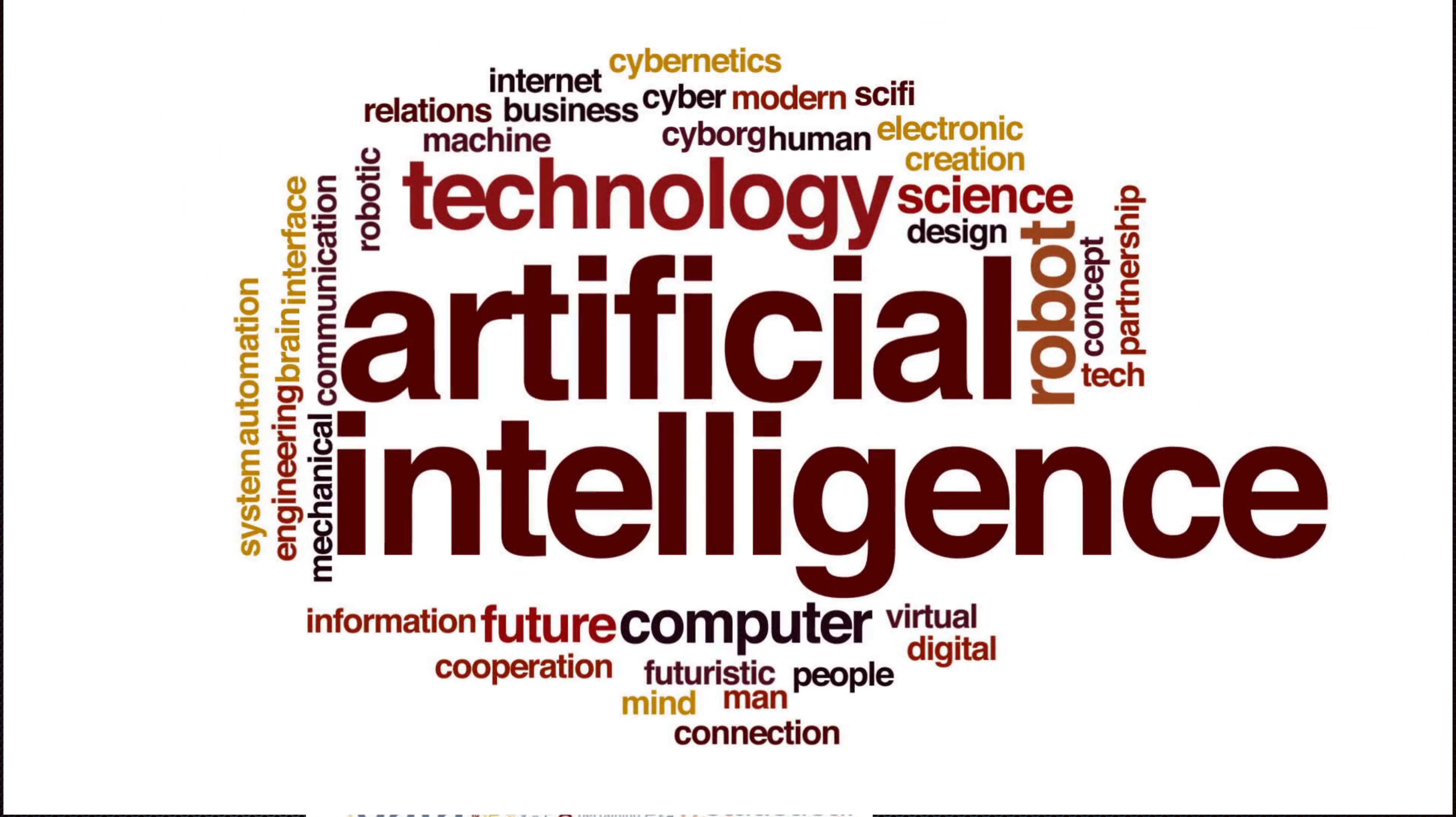

cybernetics
internet
relations
business
cyber
modern
scifi
machine
cyborg
human
electronic
creation
technology
science
design
robotic
communication
interface
brain
automation
system
engineering
mechanical
artificial
robot
concept
partnership
tech
intelligence
information
future
computer
virtual
digital
cooperation
futuristic
people
mind
man
connection


mechanism
teaching
network
creative
memory
education
internet
word
gear
together
working
precision
innovation
training
communication
human
analysis
computerized
letter
concept
learn
technology
design
information
intelligence
type
machinery
skill
skilled
workshop
machine
component
trainee
idea
industry
factory
robot
science
business
numerical
future
imagination
message
equipment
operating
engineering
learning


artificial
intelligence
technology
cybernetics
internet
relations business
cyber modern scifi
machine
cyborg human
electronic
creation
science
design
robot
concept
partnership
tech
robotic
interface
communication
brain
engineering
automation
system
mechanical
information
future
computer
virtual
digital
cooperation
futuristic
people
mind
man
connection

machine
learning
technology
training
innovation
communication
information
engineering
science
intelligence
design
mechanism
teaching
network
creative
internet
word
working
precision
memory
education
human
computerized
analysis
concept
type
machinery
skill
skilled
workshop
component
trainee
idea
industry
factory
robot
future
imagination
message
equipment
operating
business
numerical
data
operator
computer
engineer
digitally

Statistics
organization
scientific
descriptive
samples
analysis
collection
data
interpretation
statistical
Inference
parameters
mathematics
provides
also
word
wide
inferential
consider
probability
population
science
experiments
theory
referring
pertaining
discipline
mathematical
branch
called
design
prediction
total
statistic

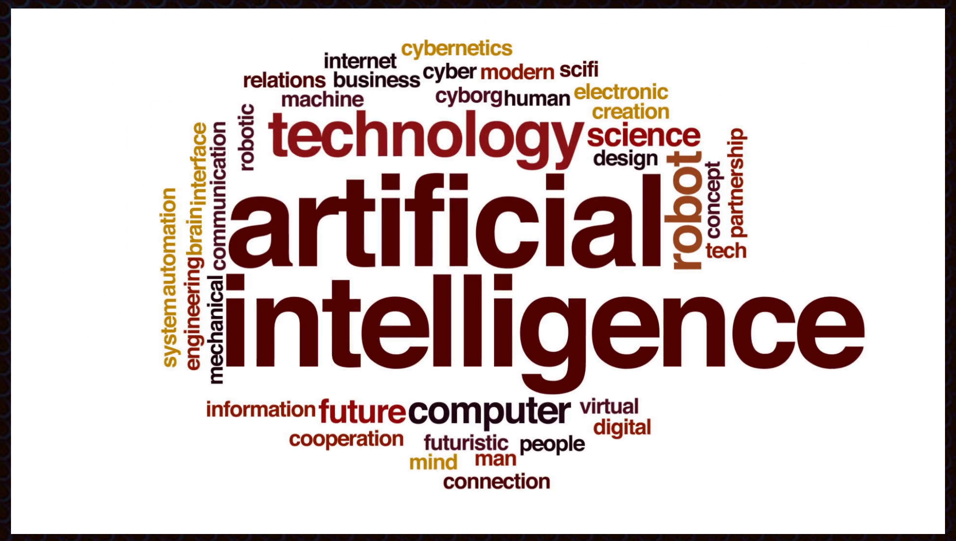

data
analysis
learning
machine
processing
research
networks
bigdata
modelling
deep-learning
statistics
materials
mining
energy
performance
science
information
systems
signal
applications
computer
technology
human
vision
engineering
smart recognition
statistical
development
reasoning
wireless
topological
phenomena
social
sports
wave
football
coastal
mathematical
intelligence
government
department
simulation
application
understanding
theory
large
design
process
security
allocation
space
methods
geometric
analytics
machine-learning
management
pattern
collection
interpretation
Inference
probability
prediction
population
sciences
experiments
discipline
pertaining
mathematica
cybernetics
internet
relations
business
cyber
modern
scifi
cyborg
electronic
creation
robotic
interface
communication
brain
automation
system
engineering
mechanical


artificial
intelligence
technology
robot
science
computer
future
cybernetics
internet
relations
business
cyber
modern
scifi
machine
cyborg
human
electronic
creation
design
robotic
interface
communication
brain
automation
system
engineering
mechanical
concept
partnership
tech
information
virtual
digital
cooperation
futuristic
people
mind
man
connection

Inference
probability
discipline
prediction

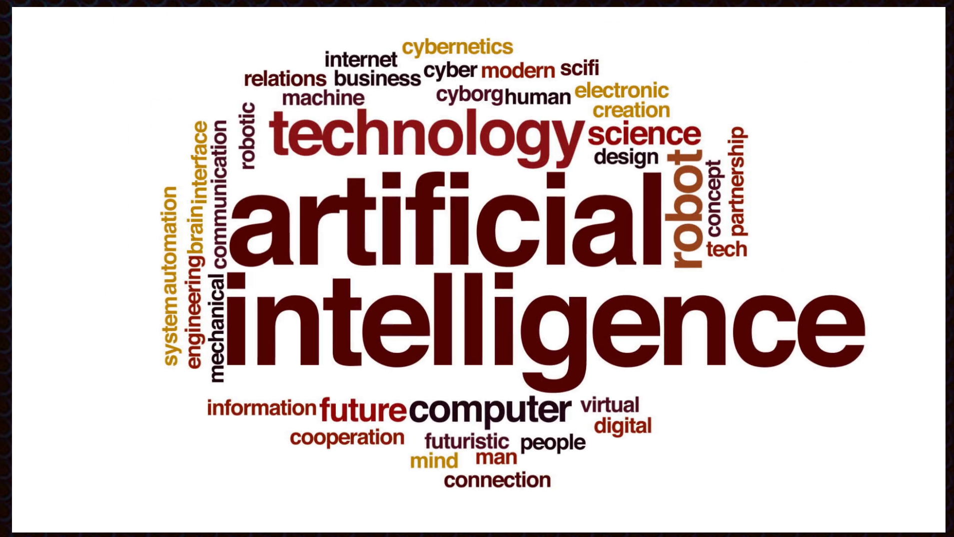
What is AI?

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



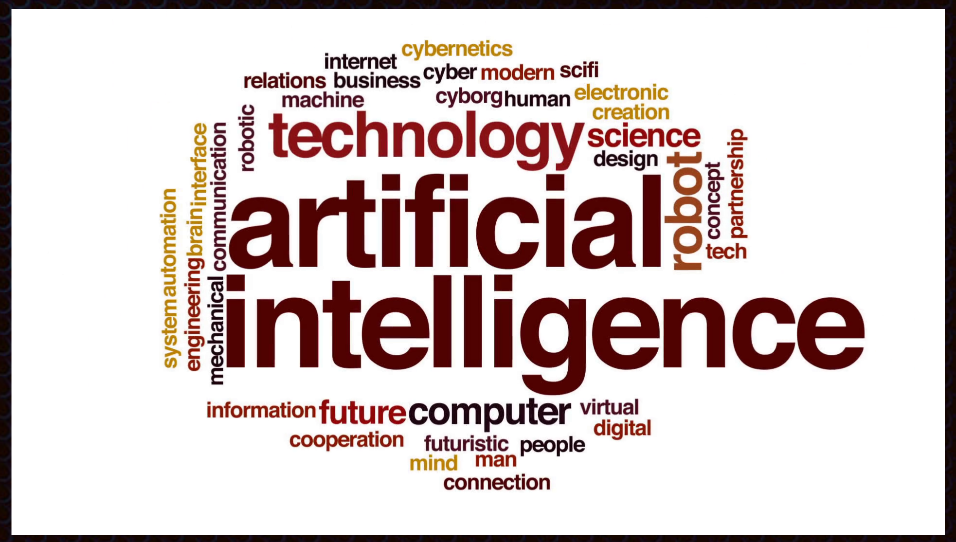
AI in Biotech

- 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



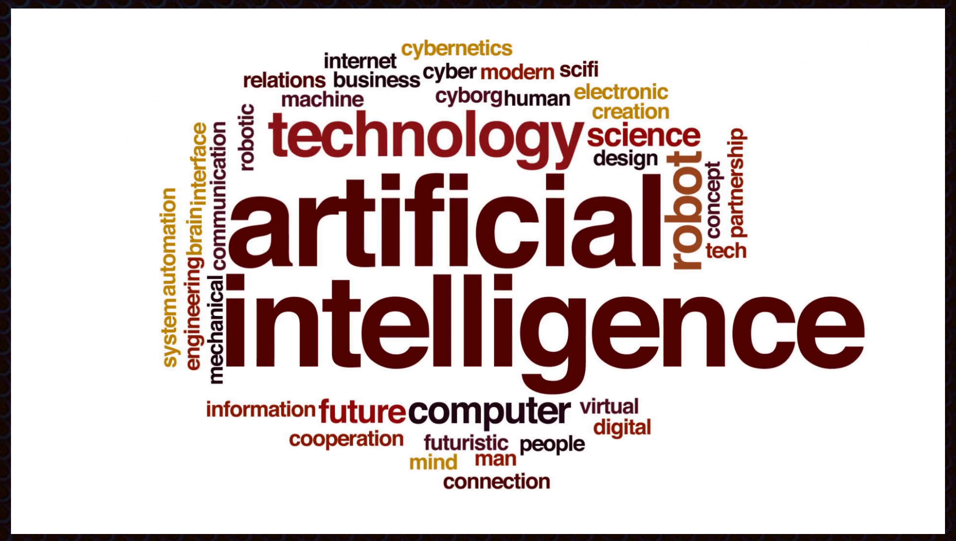
AI in Biotech

- 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

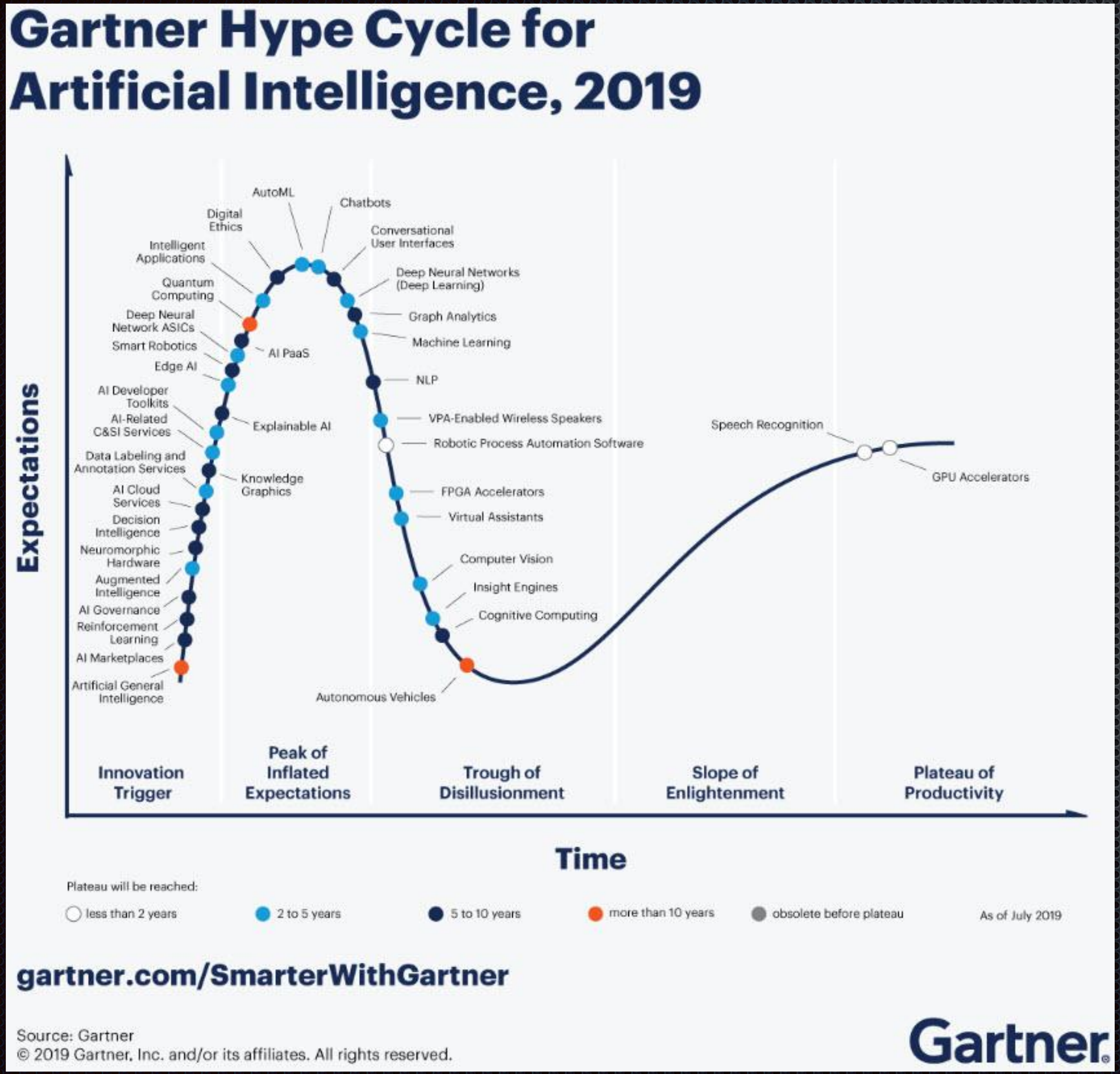


Are we there yet?

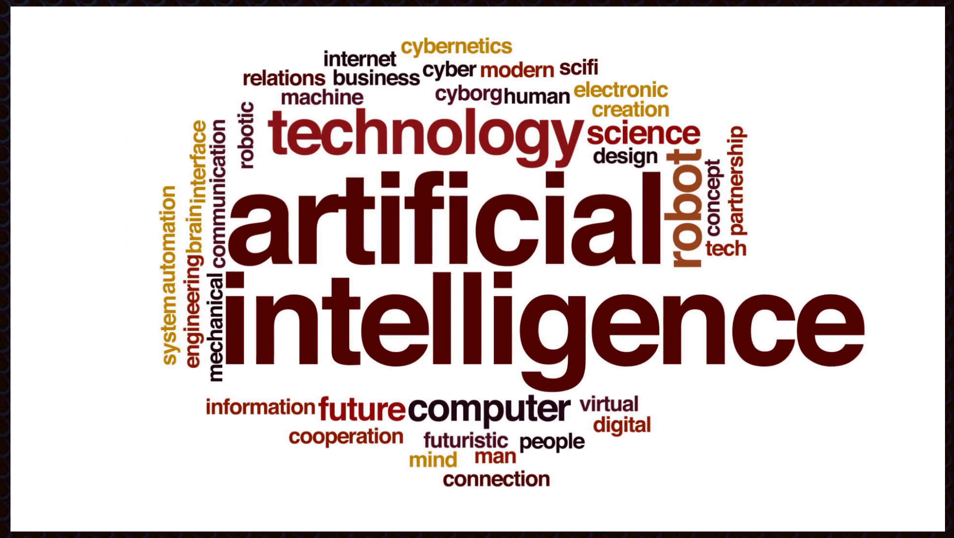
- Not really
  - Self-driving cars and planes on auto-pilot
  - Experimental situations
  - Some medical decision support



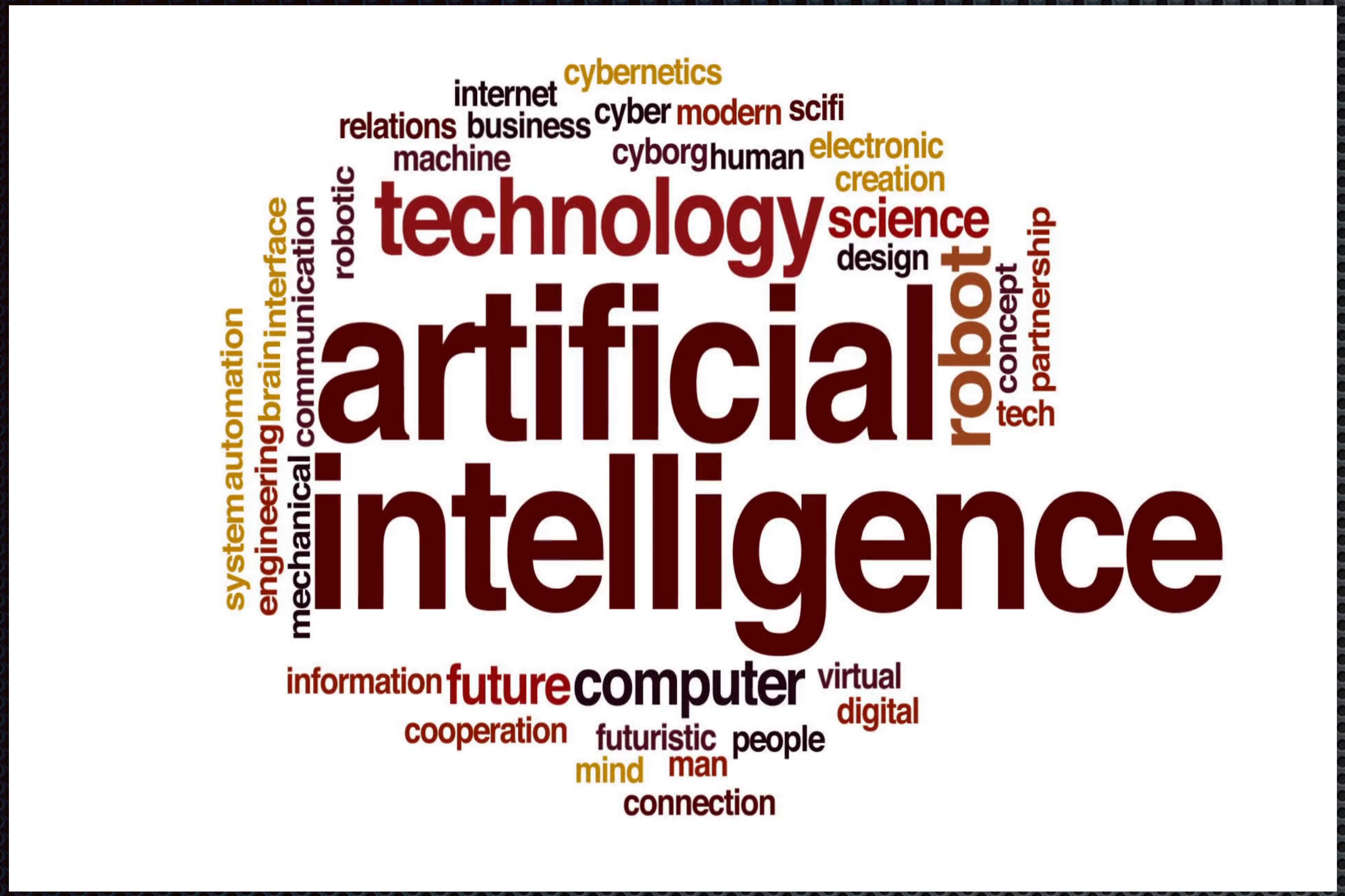
Are we there yet?



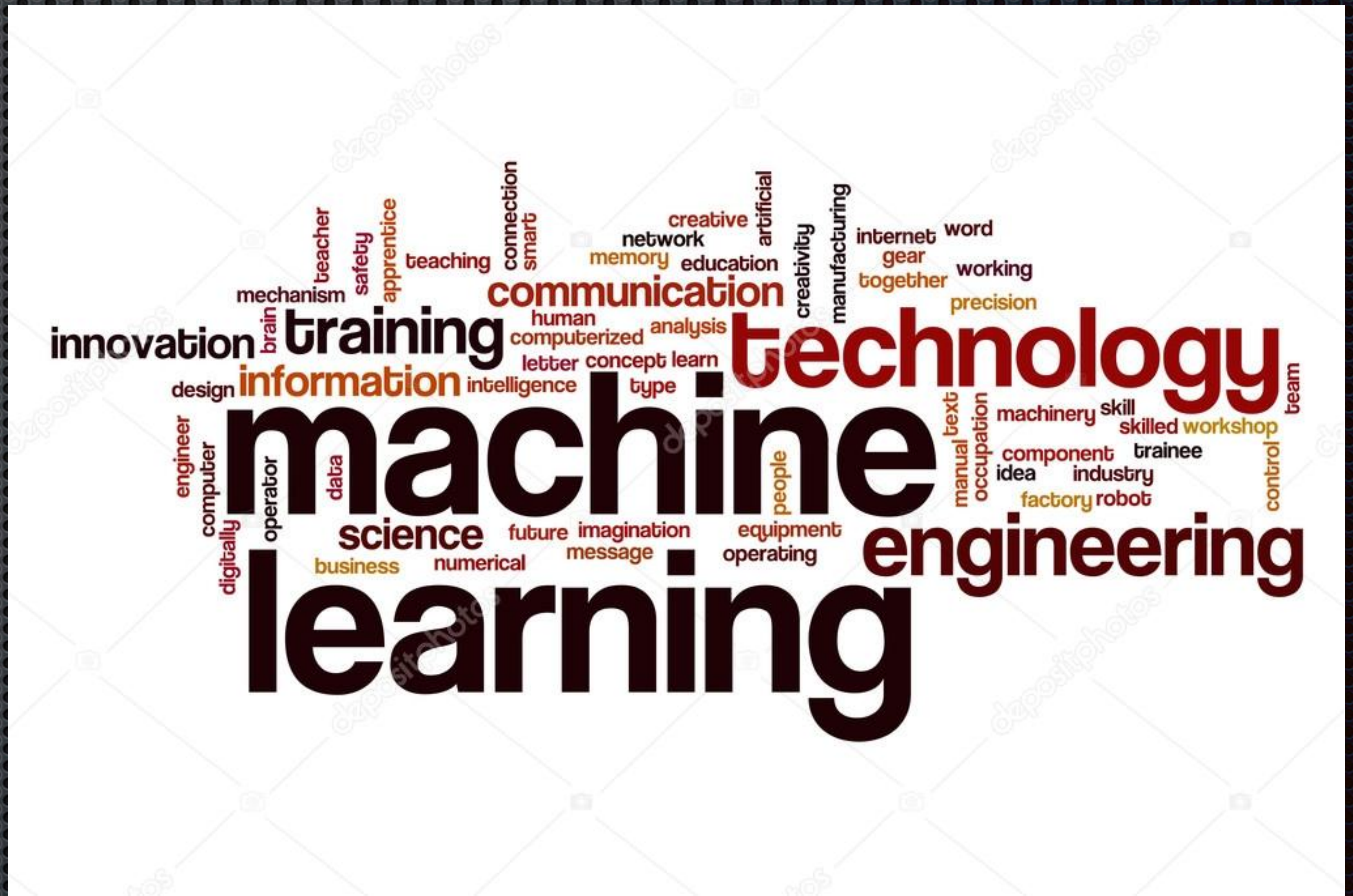
Gartner Hype Cycle 2019

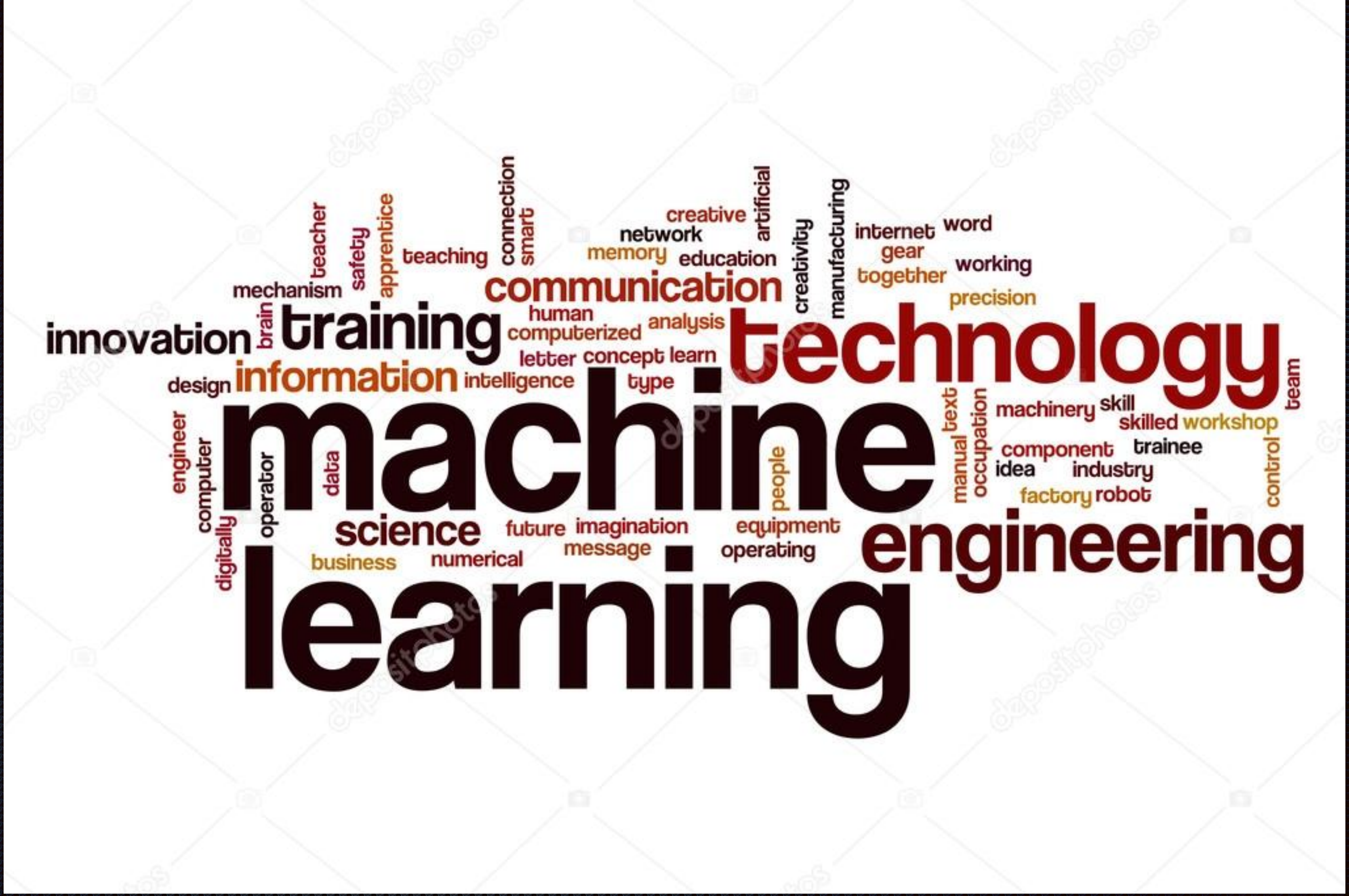


So where are we?



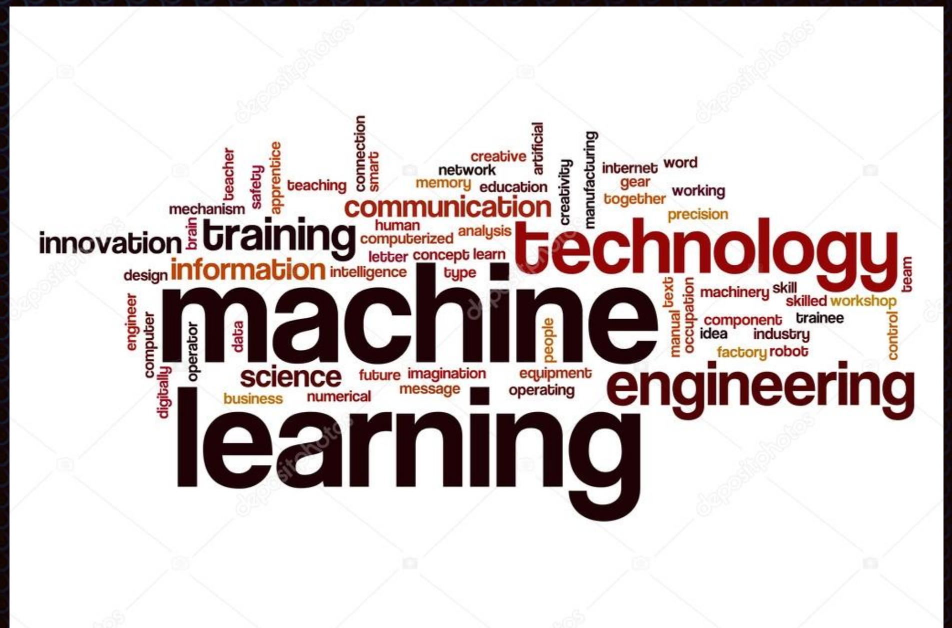
Common use




machine
learning
technology
engineering
training
innovation
communication
information
science
teacher
safety
apprentice
teaching
connection
smart
creative
network
memory
education
artificial
creativity
manufacturing
internet
word
gear
working
together
precision
mechanism
brain
human
analysis
computerized
letter
concept
learn
design
intelligence
type
team
text
occupation
machinery
skill
skilled
workshop
component
trainee
manual
idea
industry
control
factory
robot
engineer
computer
operator
data
people
digitally
future
imagination
equipment
business
numerical
message
operating

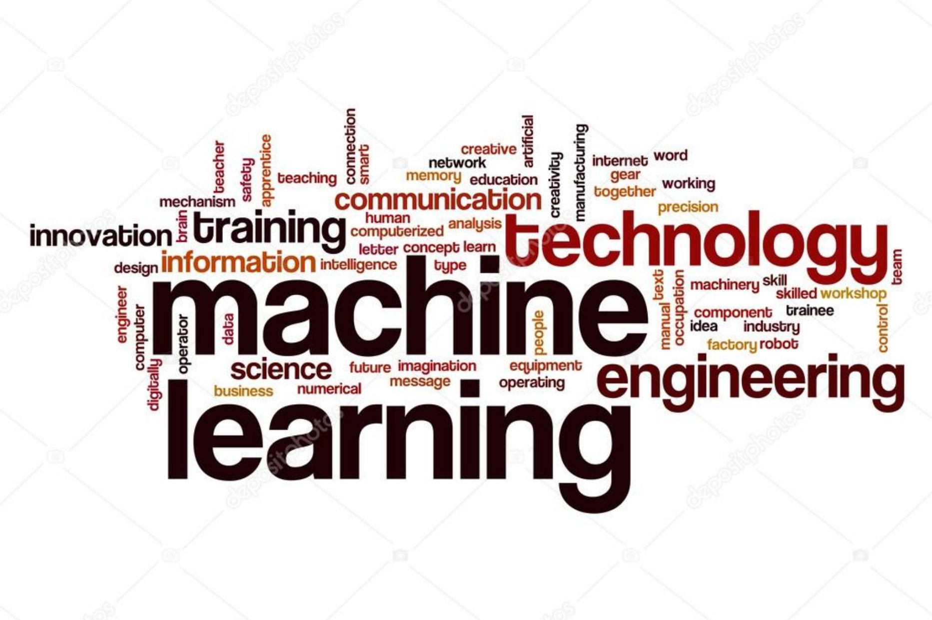
Machine learning

- 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



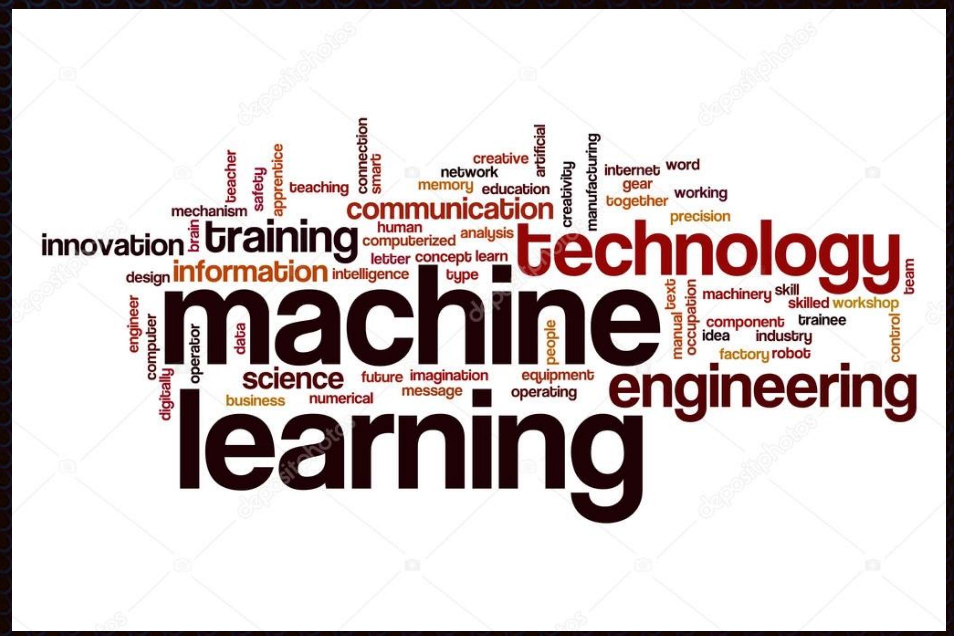
Machine learning success stories

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



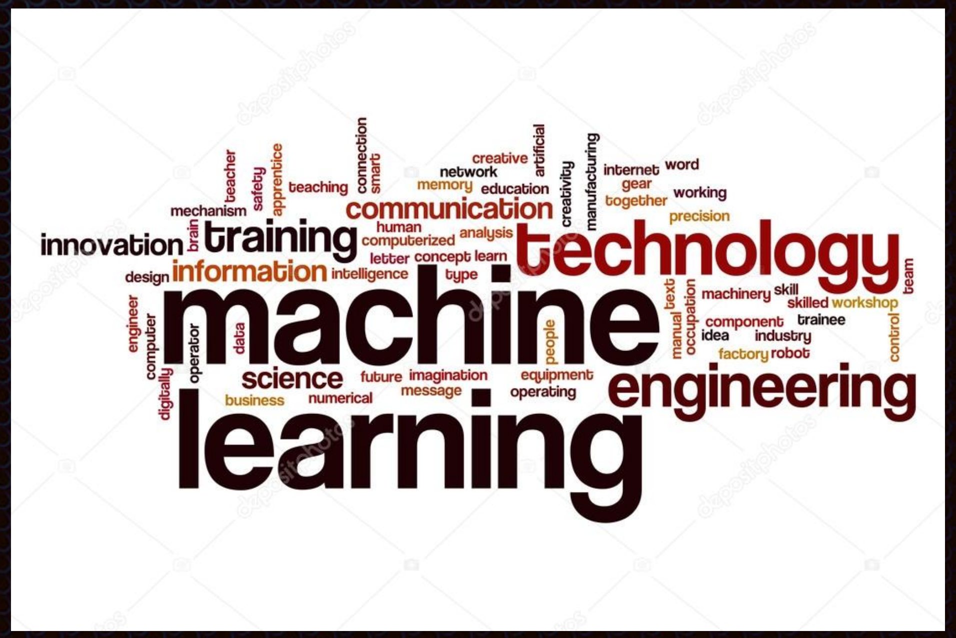
Machine learning success stories

- 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)



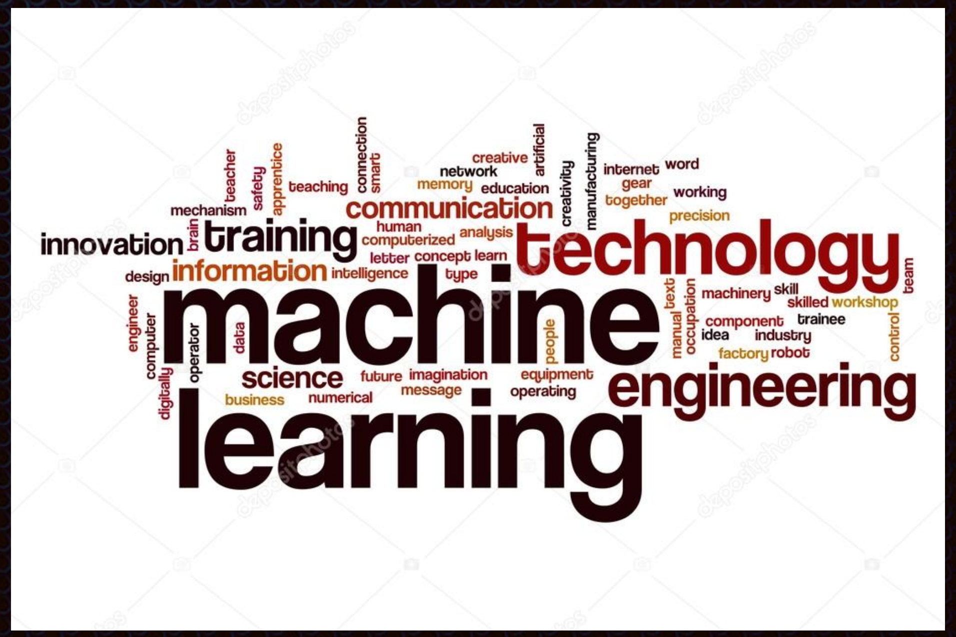
Machine learning success stories

- 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



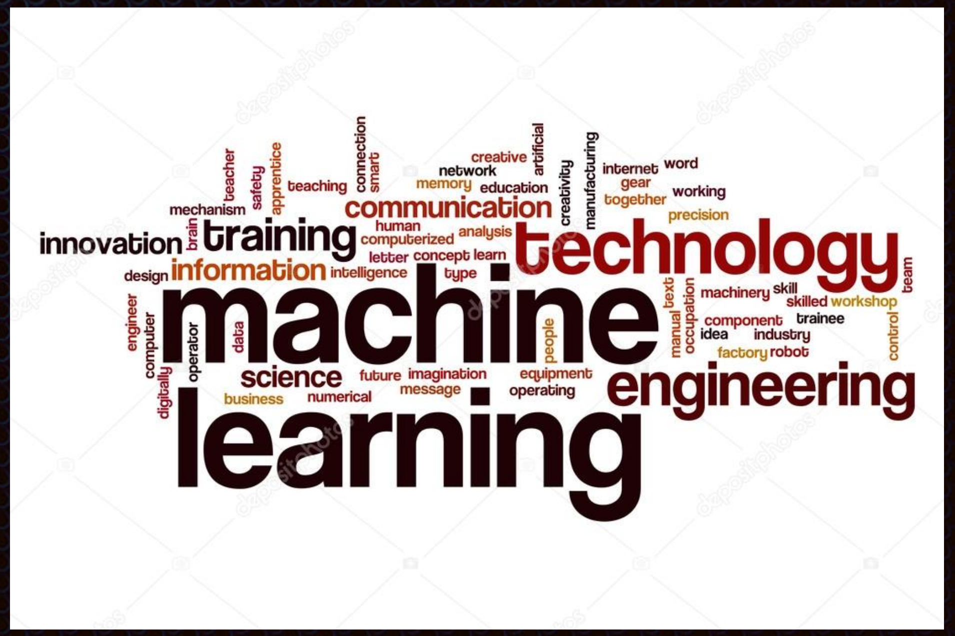
What is ML?

- 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?

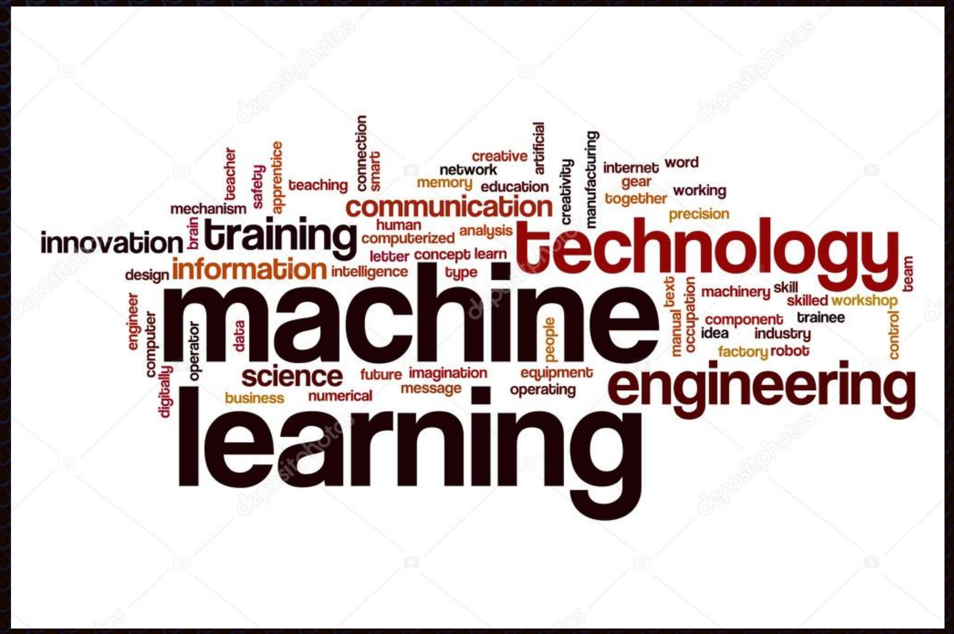
- Algorithms that are mainly meant for prediction
  - Supervised learning
- They cannot often express why the predictions happen
  - What factors influenced the prediction in what way?
  - It’s a function of the complexity of the model and our inability to express or comprehend that complexity easily



ML and Big Data

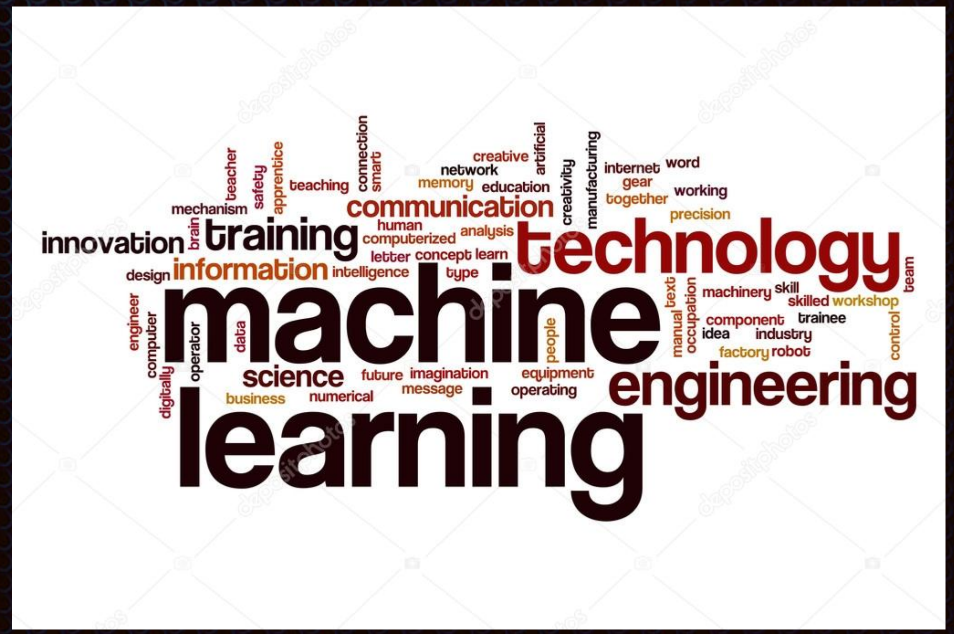
Machine learning is a greedy, hungry, voracious animal

with respect to 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 stores
- 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

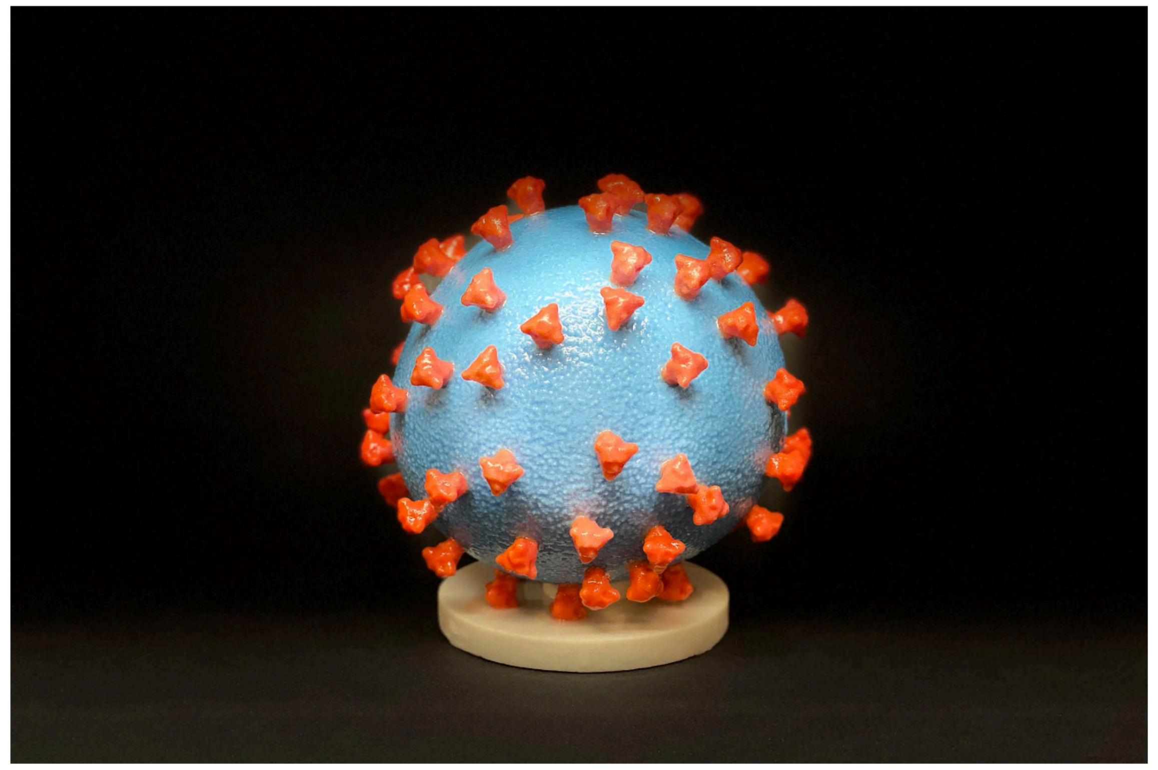
- Pharma is losing patent control
- Getting a drug to market requires many years and billions of dollars
- However, we have plenty of clinical, bioinformatic, molecular and trial data

AI and coronavirus

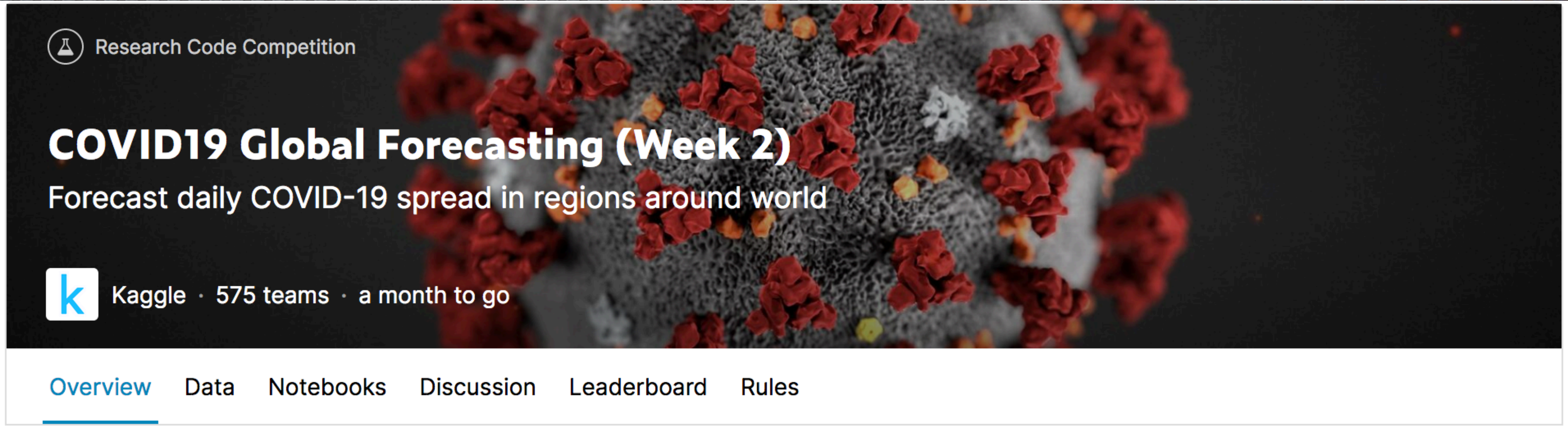
WILL KNIGHT BUSINESS 03.17.2020 08:00 AM

Researchers Will Deploy AI to Better Understand Coronavirus

More than 2,000 papers have been published about the virus since December. It will take some smart algorithms to mine insights from them.



PHOTOGRAPH: SCIENCE SOURCE



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate

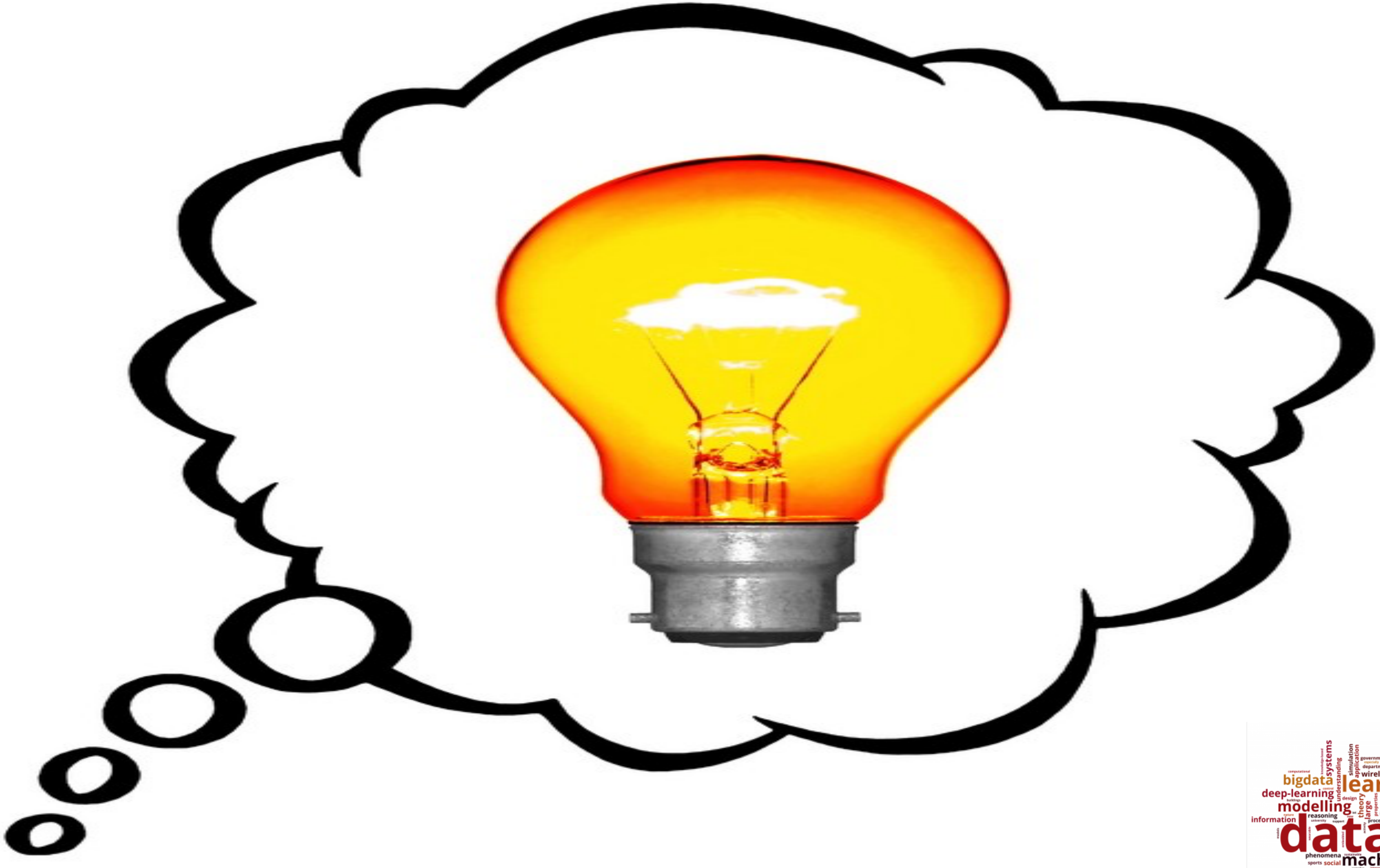
ARTIFICIAL INTELLIGENCE

Hospitals Tap AI to Help Manage Coronavirus Outbreak

Health-care providers are enlisting the technology to monitor patients, screen visitors

By Jared Council

Updated March 20, 2020 8:25 am ET




data
analysis
learning
machine
modelling
processing
networks
research
bigdata
deep-learning
systems
mining
statistics
materials
energy
information
performance
science
wireless
applications

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

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

Science & Environment

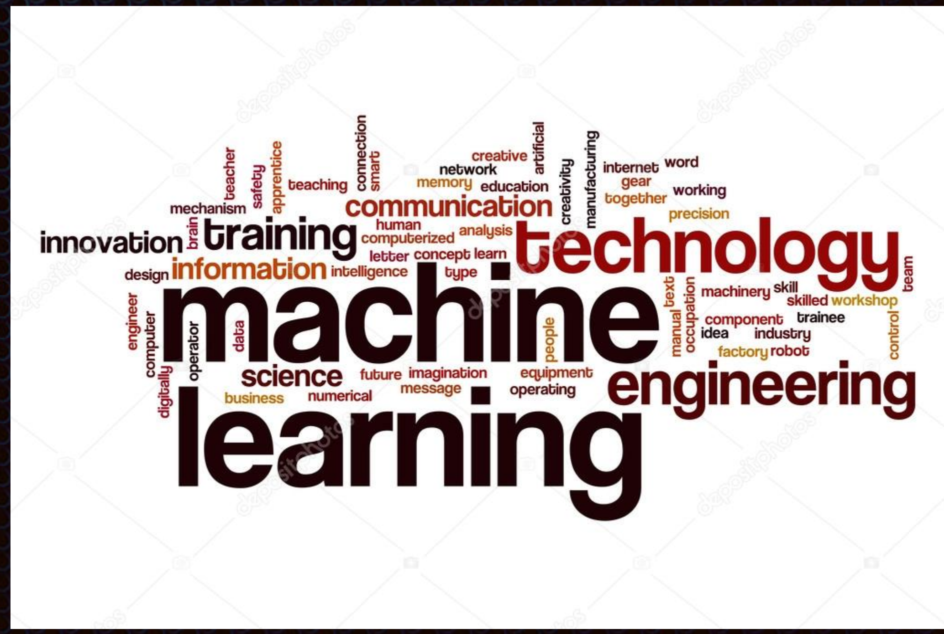
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

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



Can we trust scientific discoveries made using machine learning?

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

RICE UNIVERSITY

Science & Environment

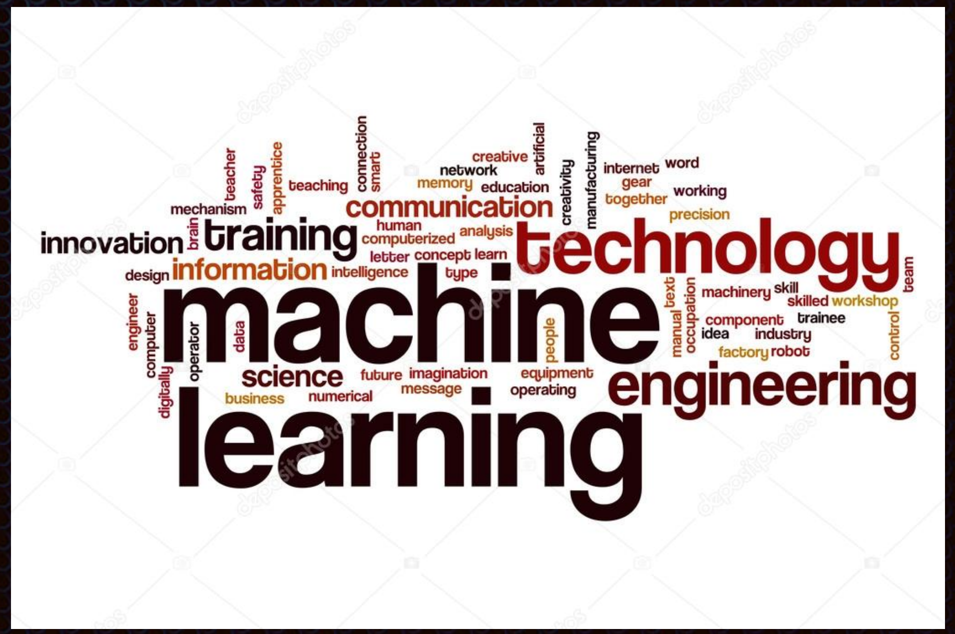
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

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

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.

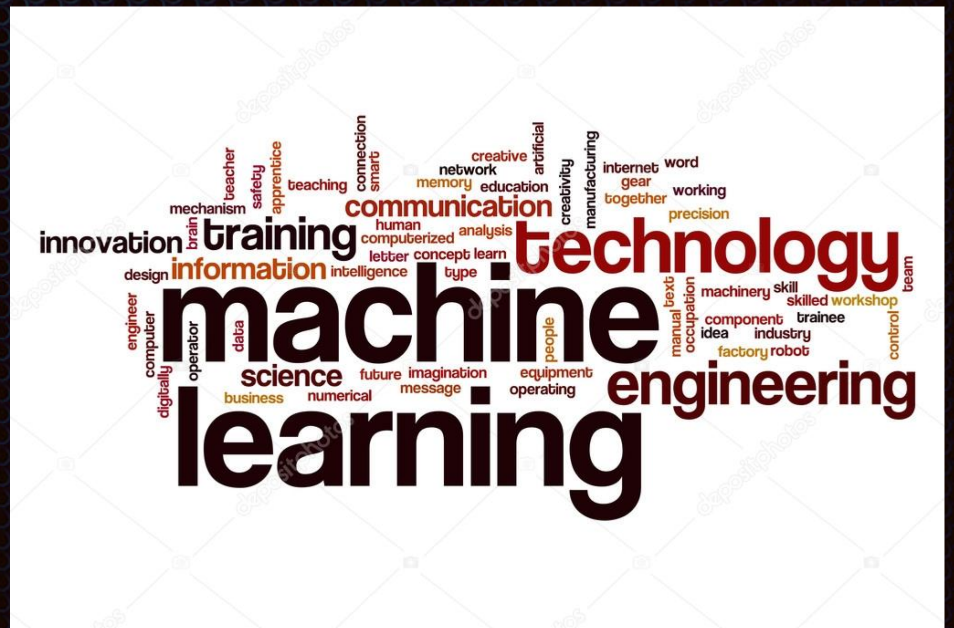
should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer,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

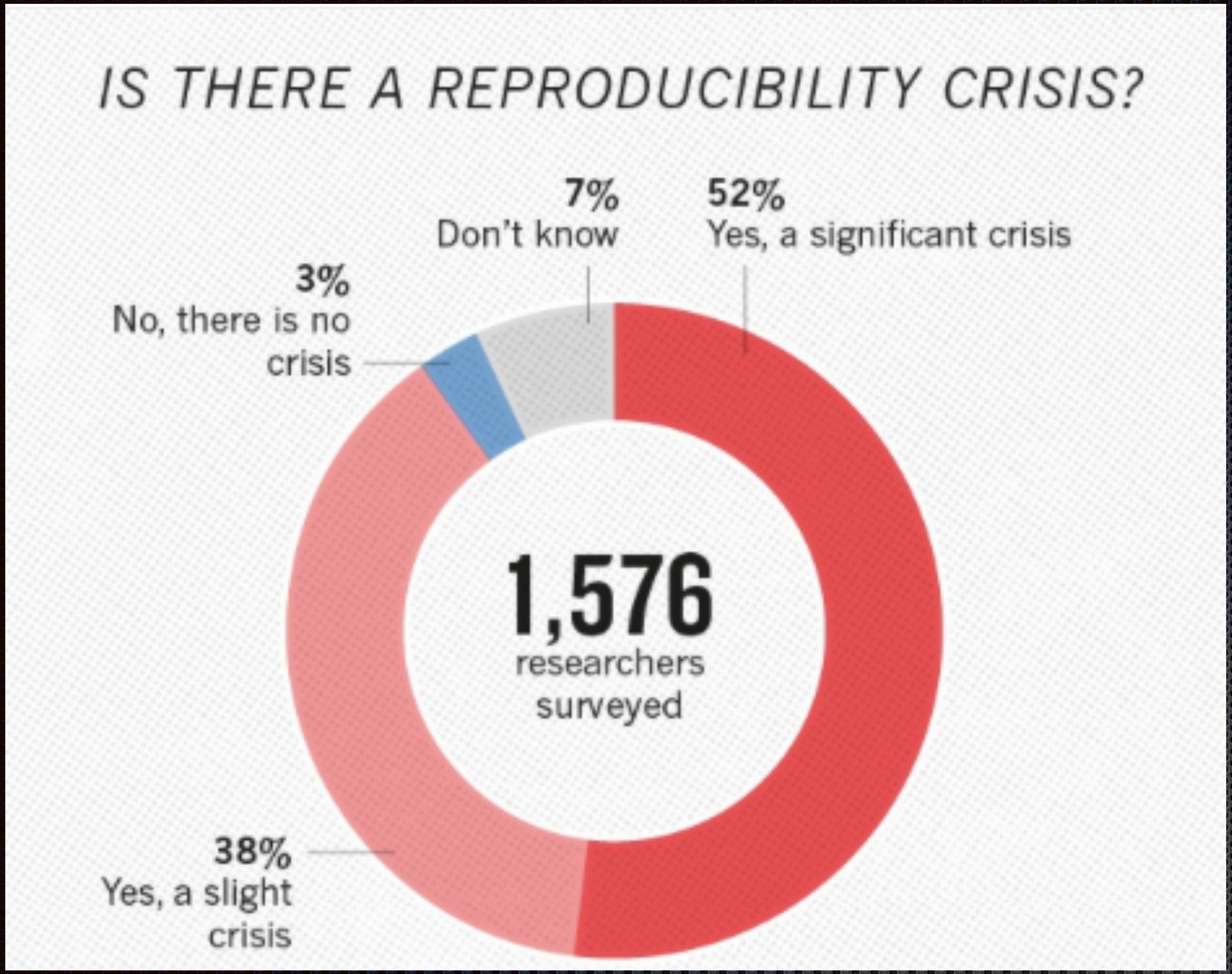
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed 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

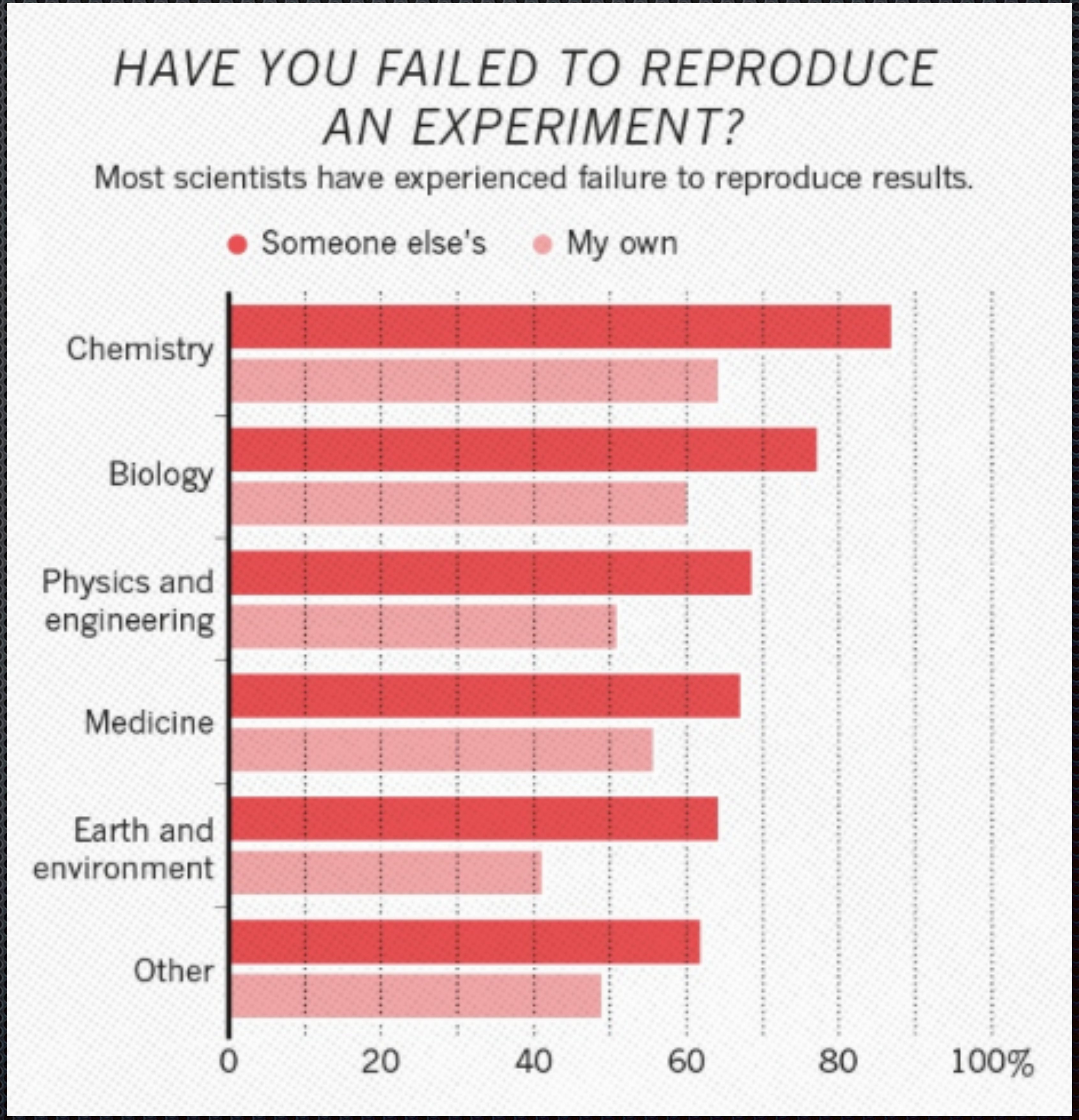
Pharma report:

89% of research findings are <u>not</u> reproducible

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

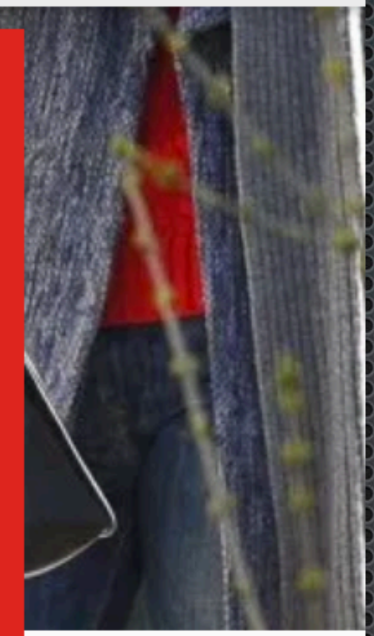



IS THERE A REPRODUCIBILITY CRISIS?
7%
Don't know
52%
Yes, a significant crisis
3%
No, there is no crisis
1,576
researchers surveyed
38%
Yes, a slight crisis


HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.
Someone else's
My own
Chemistry
Biology
Physics and engineering
Medicine
Earth and environment
Other
0
20
40
60
80
100%


data
analysis
learning
machine
modelling
bigdata
deep-learning
processing
statistics
mining
research
networks
materials
energy
information
performance

B A guide to healthy skepticism of artificial intelligence and coronavirus



REPORT

A guide to healthy skepticism of artificial intelligence and coronavirus

Alex Engler · Thursday, April 2, 2020

"When AI models leave development and start making real-world predictions, they nearly always degrade in performance."

"There is no value in AI without subject-matter expertise."

"[A]n inflated accuracy number can actually be an important sign that an AI model is not going to be effective out in the world."

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 <u>and a question</u>
- 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

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

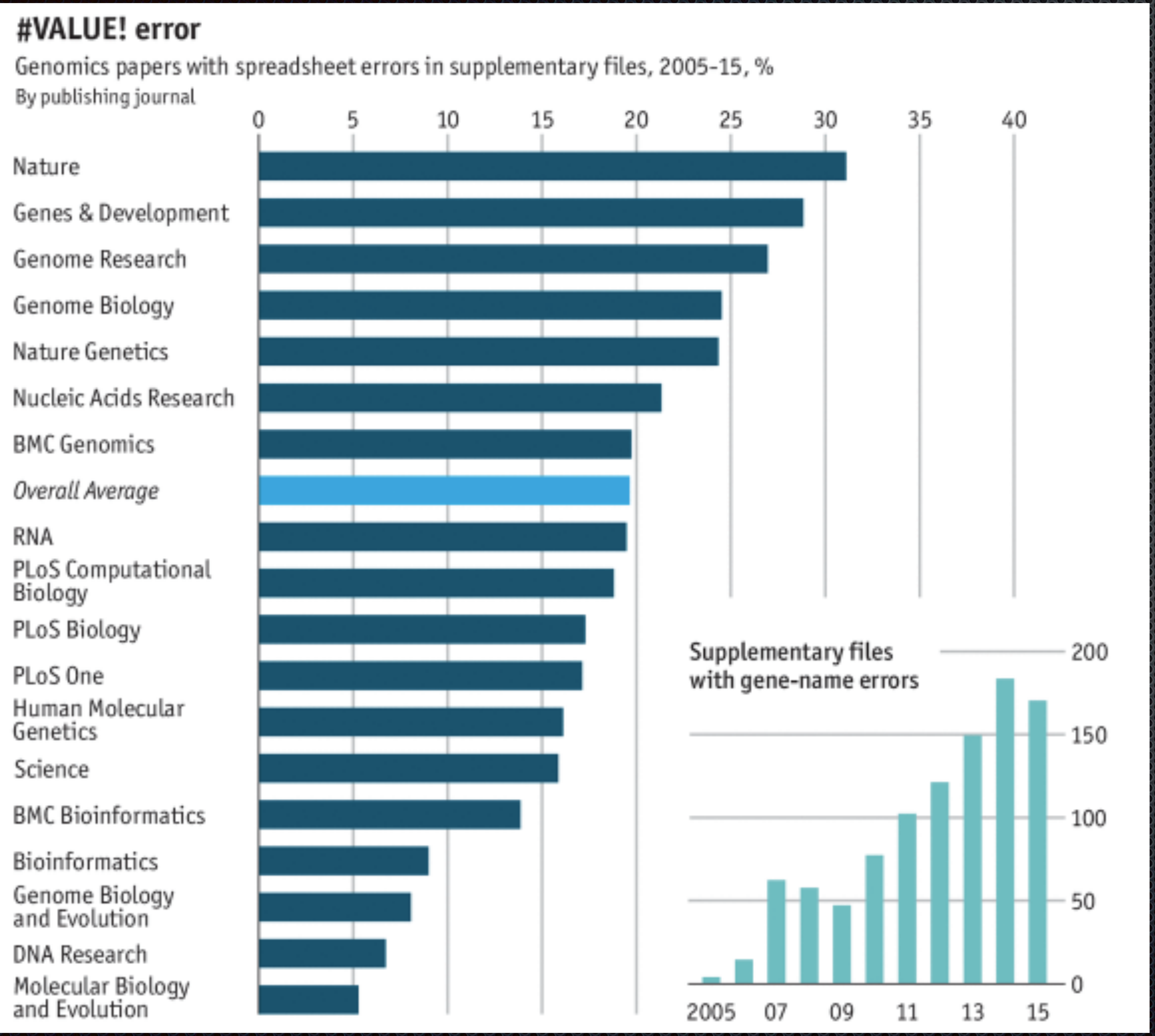
- aka, we like shiny new things

- 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

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

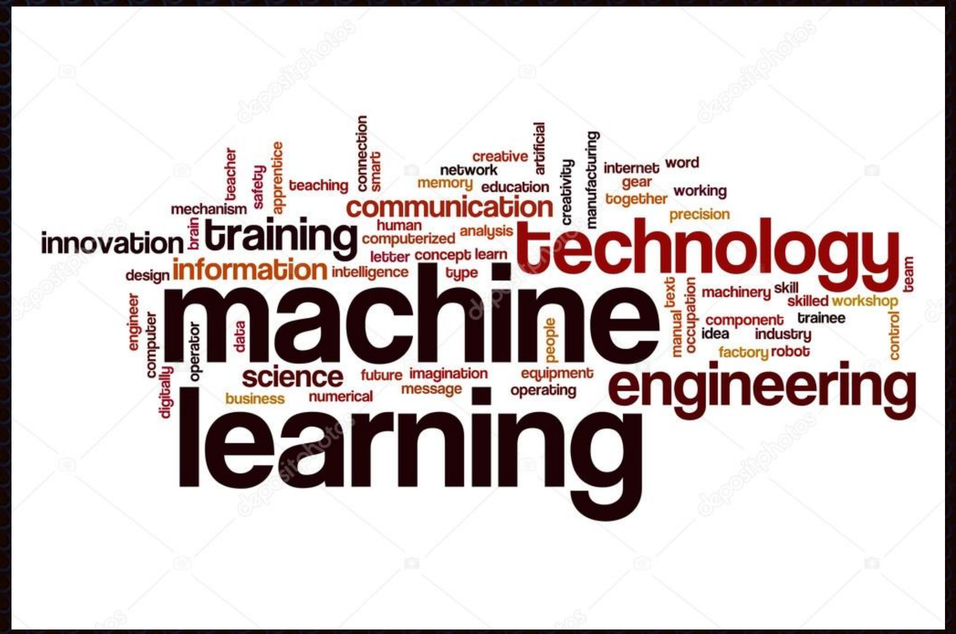
- Make sure you use the data
  - Appropriately
  - Ethically
  - Really think through and understand your data, your question, and how they relate
  - Don't just dump things into AI just because it's the cool thing

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 your’ve solved it

Contact info

- abhijit@zansors.com
- @webbedfeet / @Zansors
- https://webbedfeet.netlify.com