Biotechnology
Big Data
Artificial Intelligence

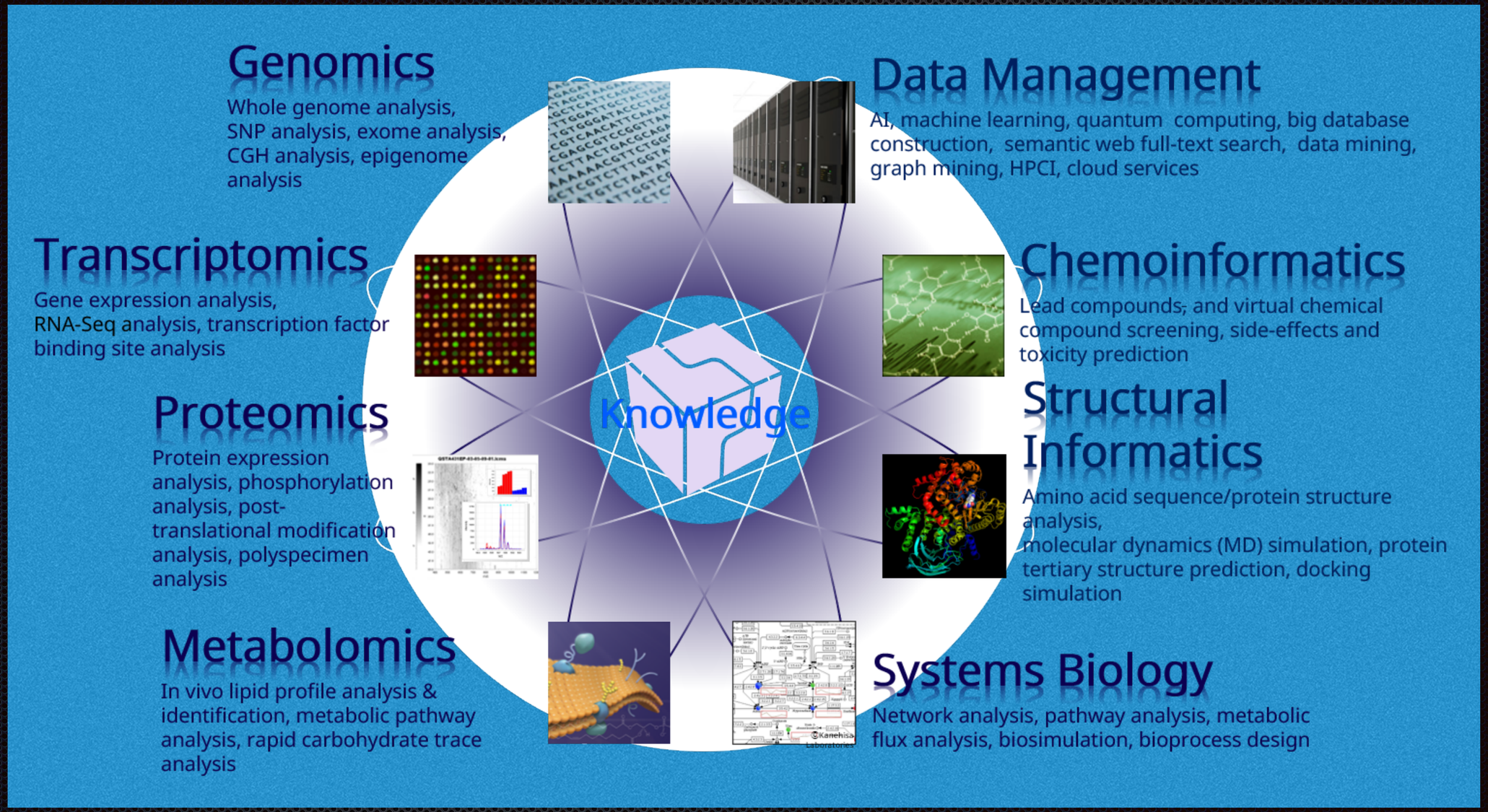
Promises, Dangers, Worries, Gotchas

Abhijit Dasgupta, PhD
Chief Data Scientist
Co-Founder

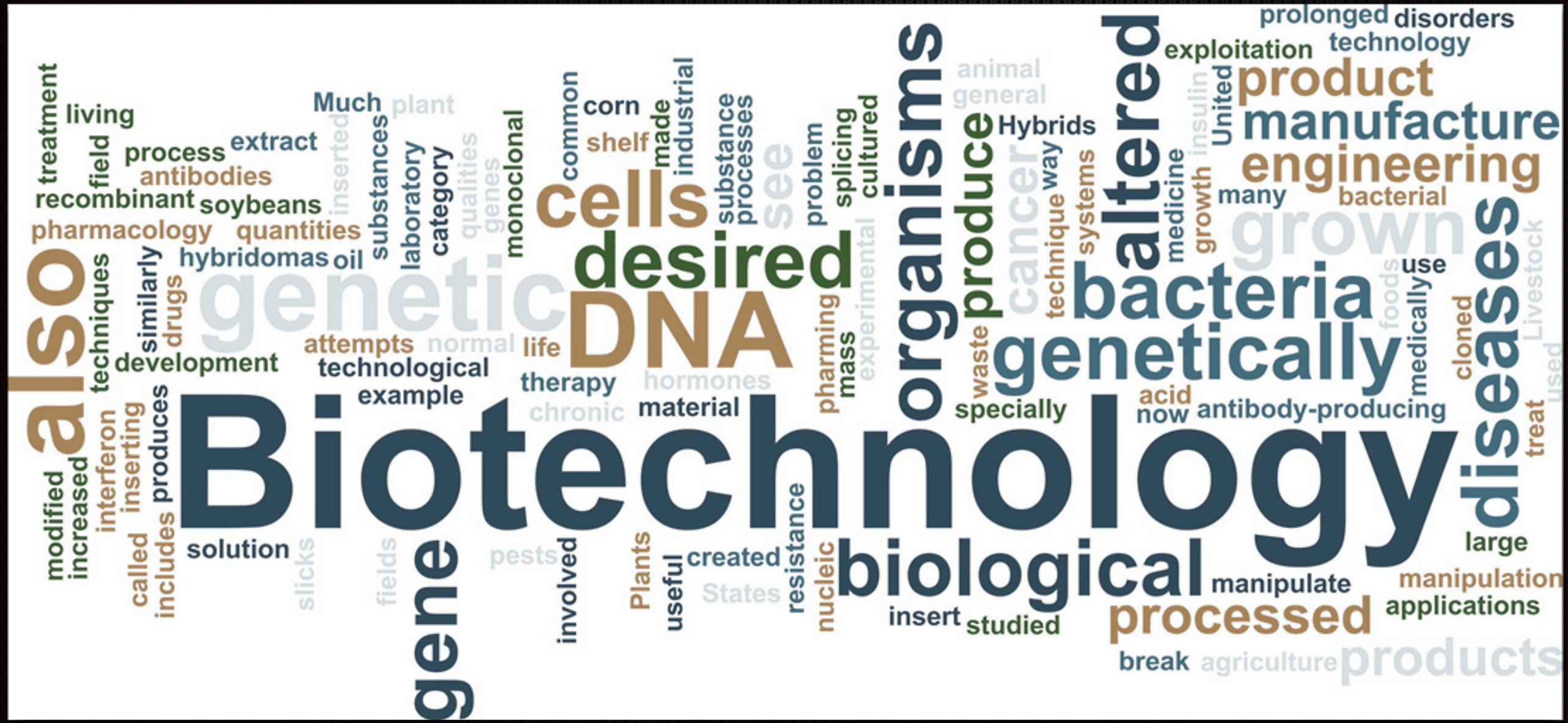
Slides available at https://webbedfeet.github.io/GMU_class

Hello!!

- Chief Data Science Officer @ Zansors
- Career spanning academics, government and industry research as data science collaborator and researcher
- Adjunct faculty at Georgetown
  - Data visualization, Big Data, Biostats/Bioinformatics
- Adjunct faculty at GMU & Georgetown



https://www.mki.co.jp/english/bioinformatics.html


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

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

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)

Biotechnology
also
genetic
DNA
cells
desired
organisms
altered
product
manufacture
engineering
bacteria
genetically
diseases
biological
processed
products
gene
produce
grows
genetically
Much
extract
antibodies
soybeans
hybridomas
development
technological
example
therapy
material
monoclonal
common
corn
sheff
industrial
processes
splicing
cultured
Hybrids
exploitation
technology
prolonged
disorders
animal
living
treatment
process
recombinant
pharmacology
quantities
oil
laboratory
category
substances
plant
involved
Plants
used
created
nucleic
genes
insert
studied
break
agriculture
manipulate
applications
manipulation
large
antibody-producing
specially
now
acid
many
growth
bacterial
use
medicinal
modified
increased
called
inserting
includes
produces
solution
similar
techniques
drugs
hybridomas
oil
attempts
normal
life
Hybrid
very
techniques
genes
Uses
exploitation
technology
prolonged
disorders
large
manipulation
applications
manipulate
studied
insert
break
agriculture
products
genetically
specially
now
antibody-producing
medicinal
use
bacterial
many
growth
genes
techniques
Hybrids
very
Uses
exploitation
technology
prolonged
disorders
animal
corn
sheff
industrial
processes
splicing
cultured
monoclonal
common
Much
plant
extract
antibodies
soybeans
hybridomas
oil
laboratory
category
substances
plant
involved
Plants
used
created
nucleic
genes
insert
studied
break
agriculture
manipulate
applications
manipulation
large
antibody-producing
specially
now
acid
many
growth
bacterial
use
medicinal
modified
increased
called
inserting
includes
produces
solution
similar
techniques
drugs
hybridomas
oil
attempts
normal
life
Hybrid
very
techniques
genes
Uses
exploitation
technology
prolonged
disorders
animal
living
treatment
process
recombinant
pharmacology
quantities

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

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

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)

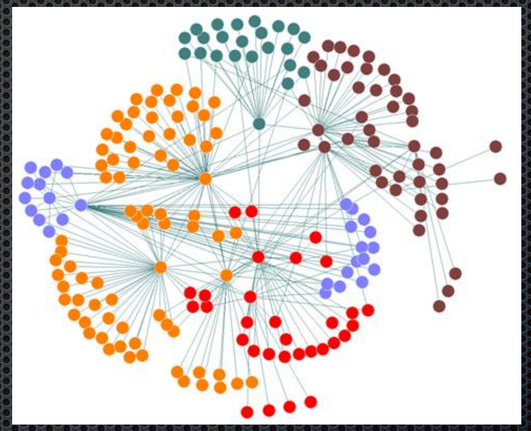
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates

Discovery (finding it)

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

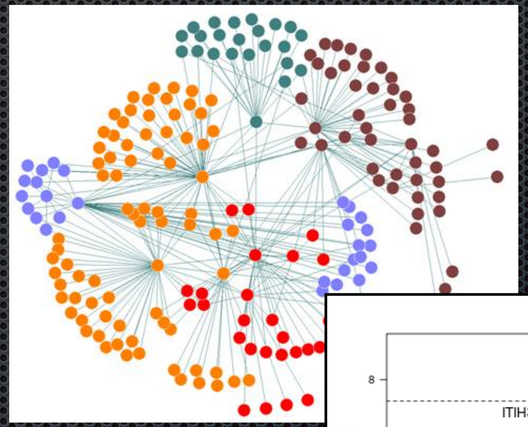
Discovery (finding it)

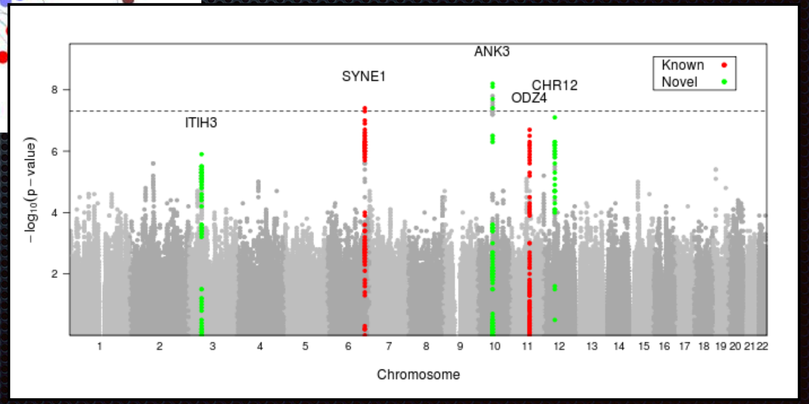
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates



Discovery (finding it)

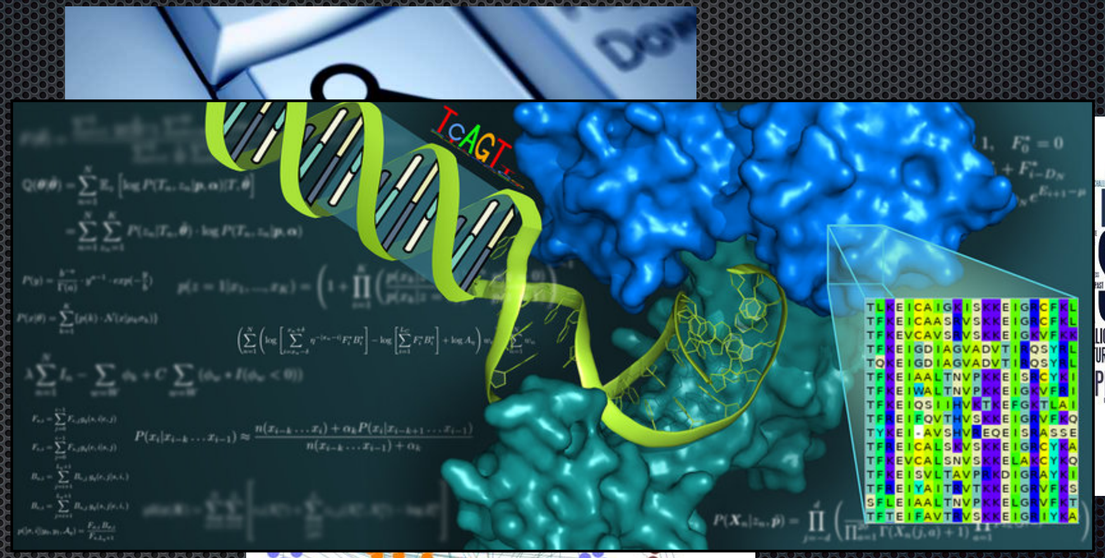
- Identifying potential
  - Searching the literature
  - Searching databases
  - Analyzing relationships
  - Culling candidates

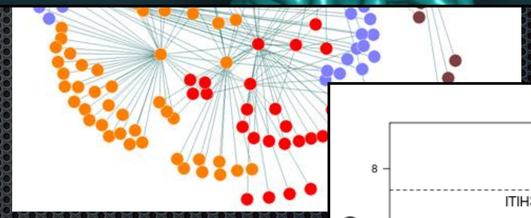


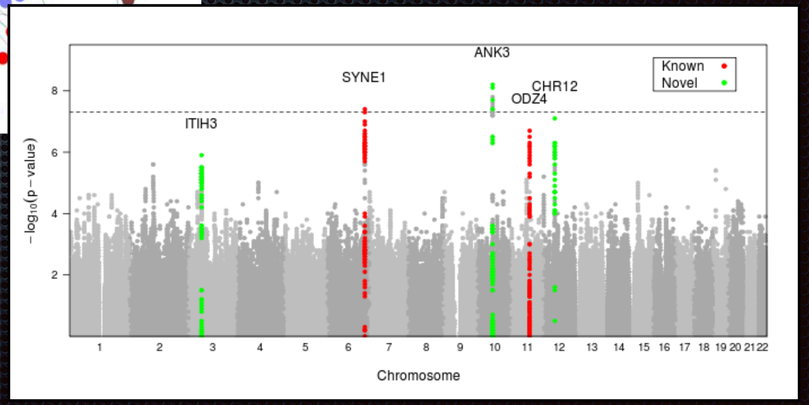


Discovery (finding it)

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

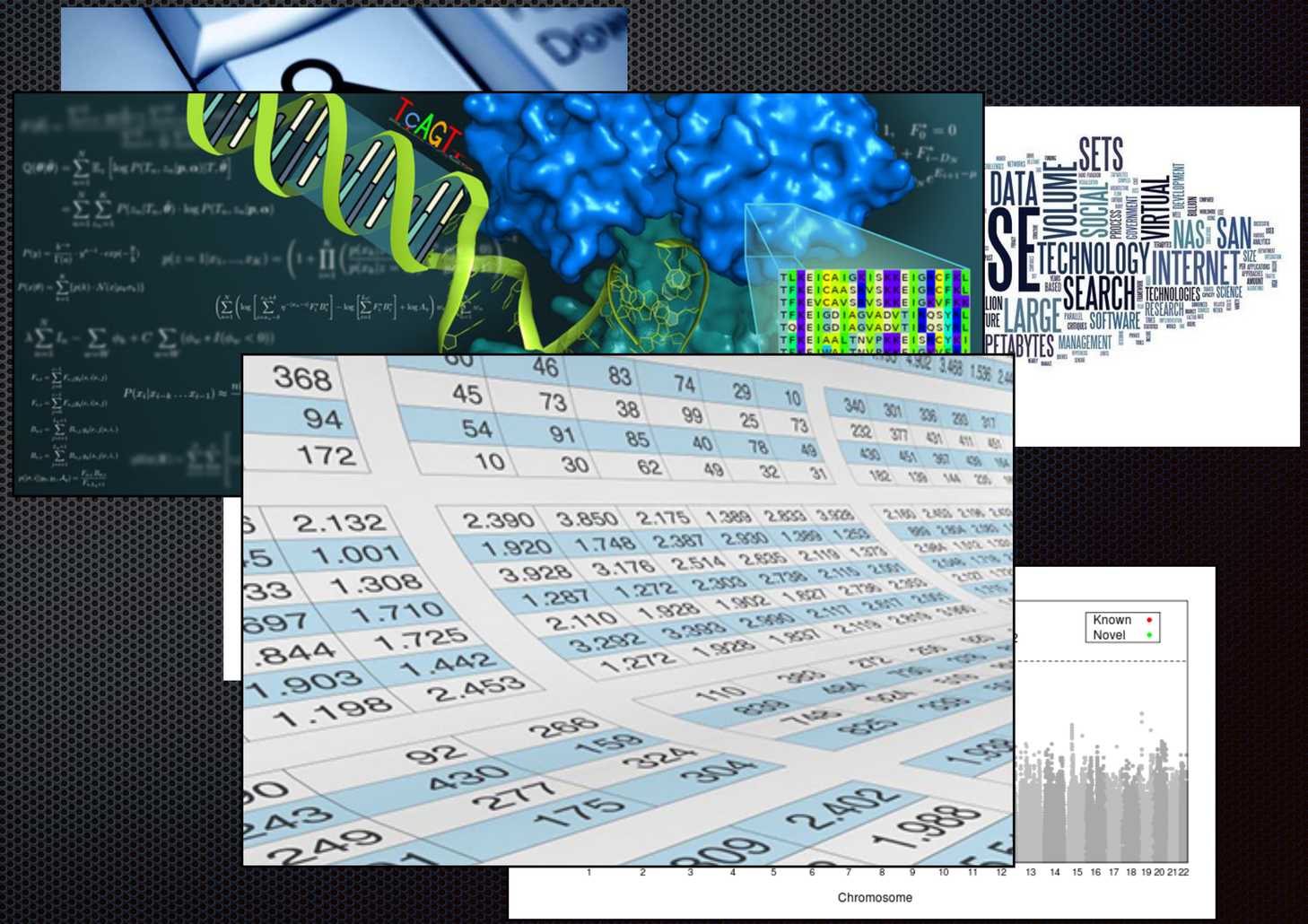




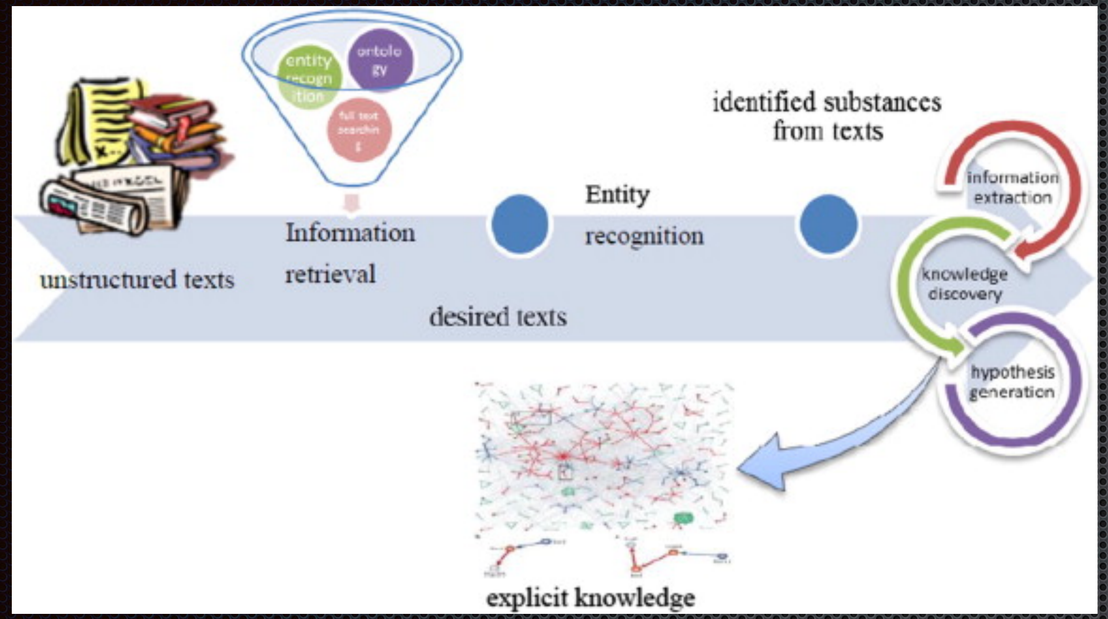


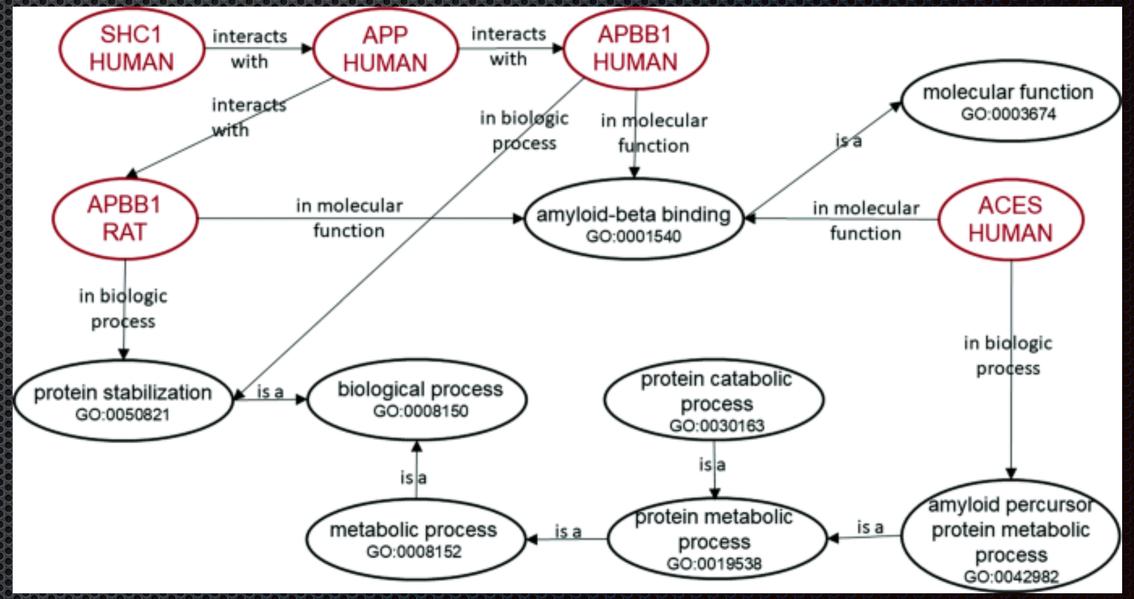
Discovery (finding it)

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



Discovery (finding it)





Natural language processing, knowledge graphs and machine-driven literature search

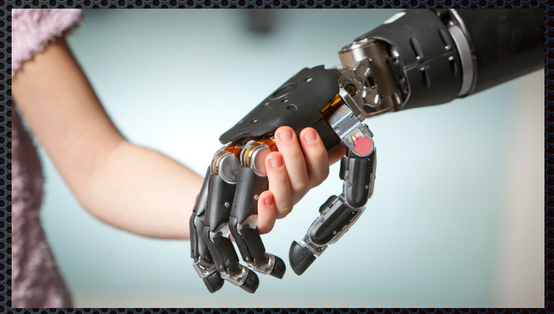
Evidence of effectiveness

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

Evidence of effectiveness

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

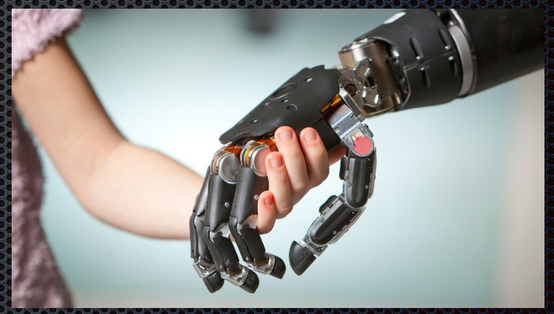
Evidence of effectiveness



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

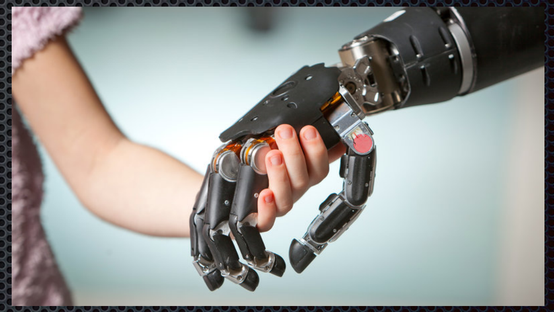
Evidence of effectiveness

- 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

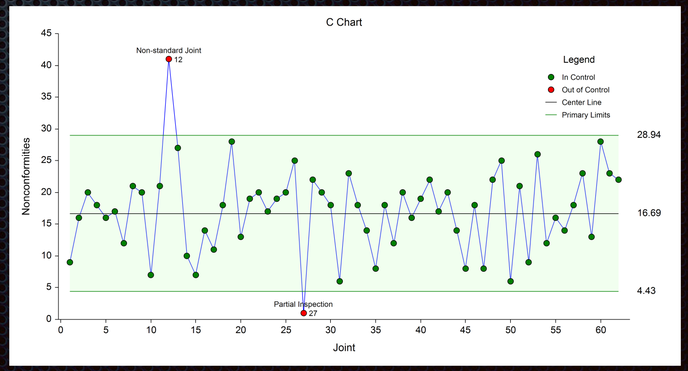
Manufacturability & Quality

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

International Organization for Standardization
ISO
13485

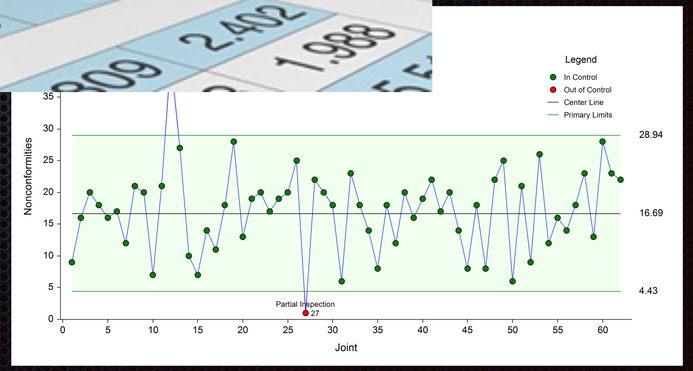
Manufacturability & Quality

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



Manufacturability & Quality

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



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

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

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

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

Data drives

Research

Development

Production

Marketing

Sales

Revisions

Central issue for data-driven business

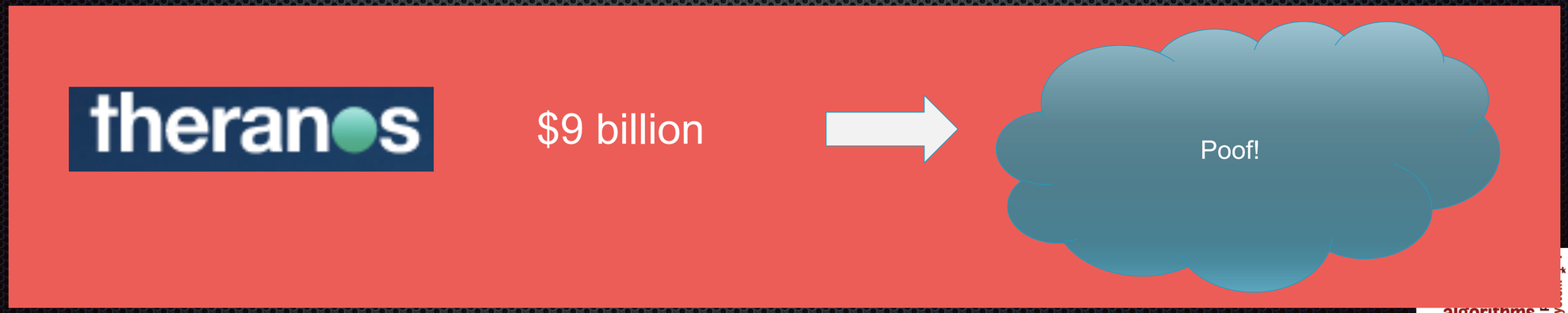
- The data <u>must</u> be reliable and valid
- Repeating the experiment must yield similar results (reliability)
- We’re measuring what we think we’re measuring (internal validity)
- The results can be generalized to larger populations (external validity)

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

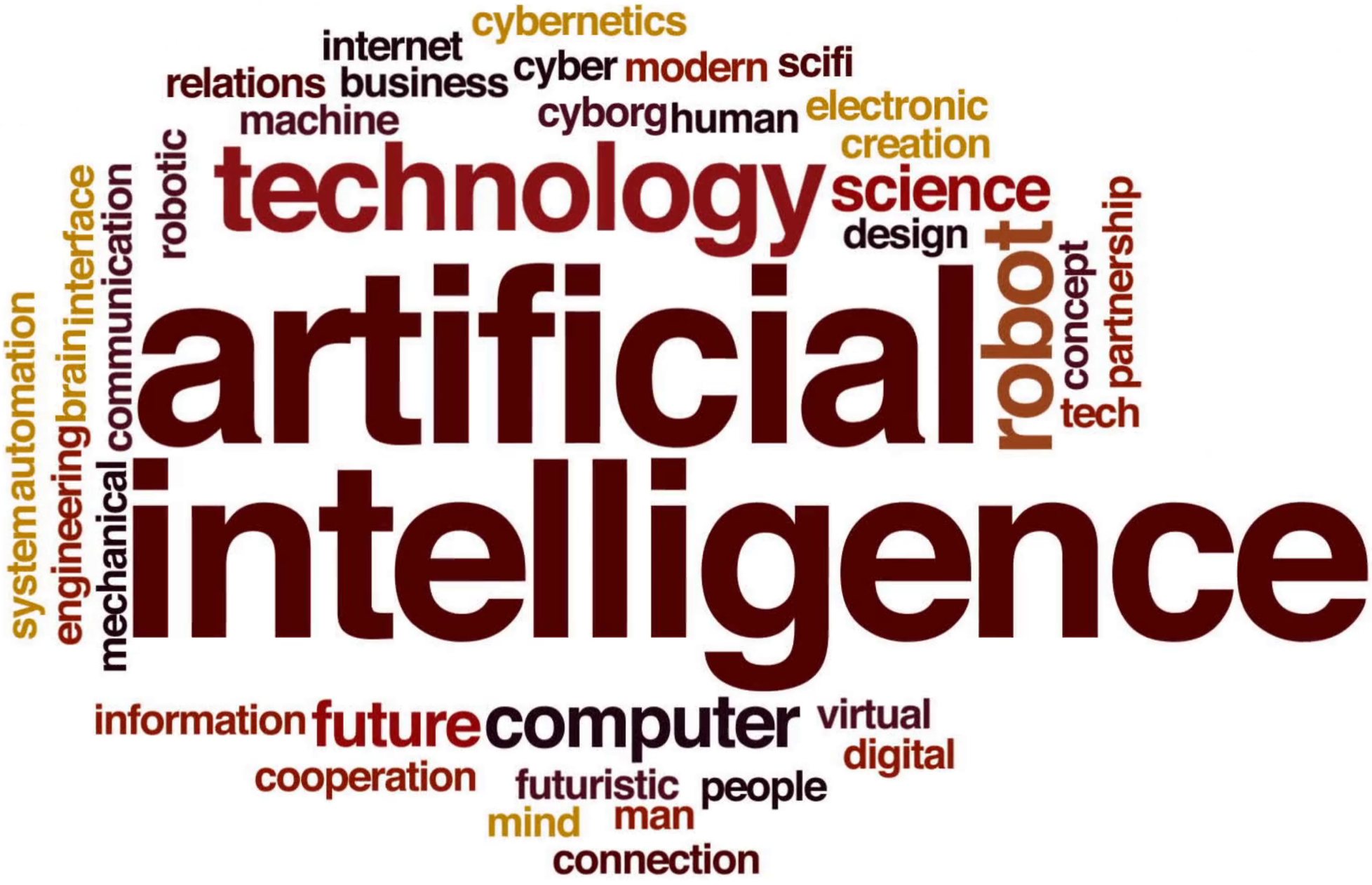
- It doesn’t matter how much data you can get
- It matters how you got it
  - Experimental design
  - Ethics
  - Provenance

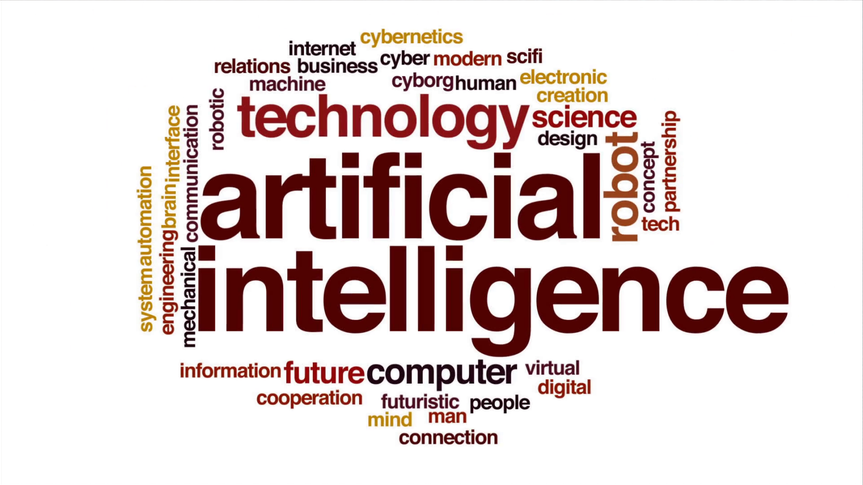
Central issue for data-driven business

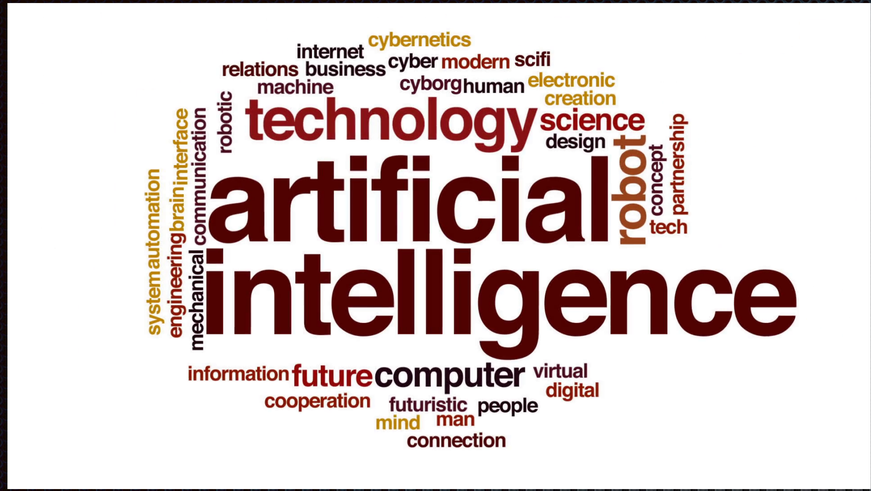
- It doesn’t matter how much data you can get
- It matters how you use it
  - Analytics
  - Translation to business practice
  - Follow-up

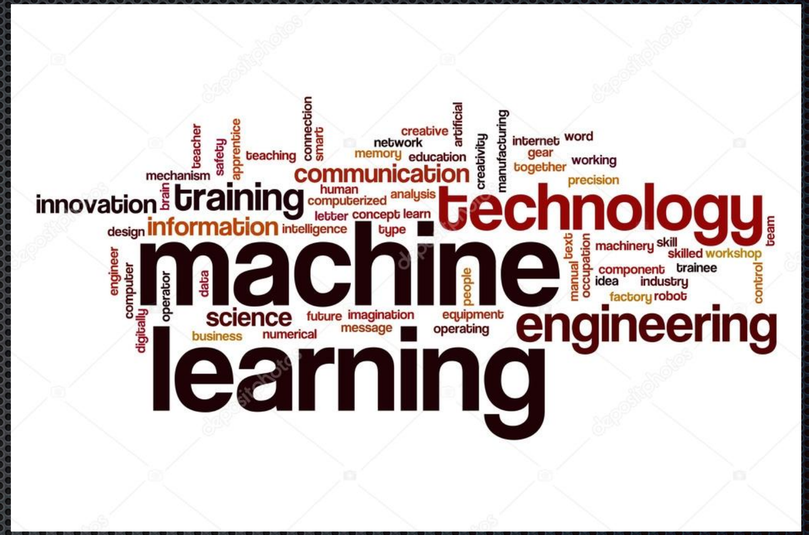
Central issue for data-driven business

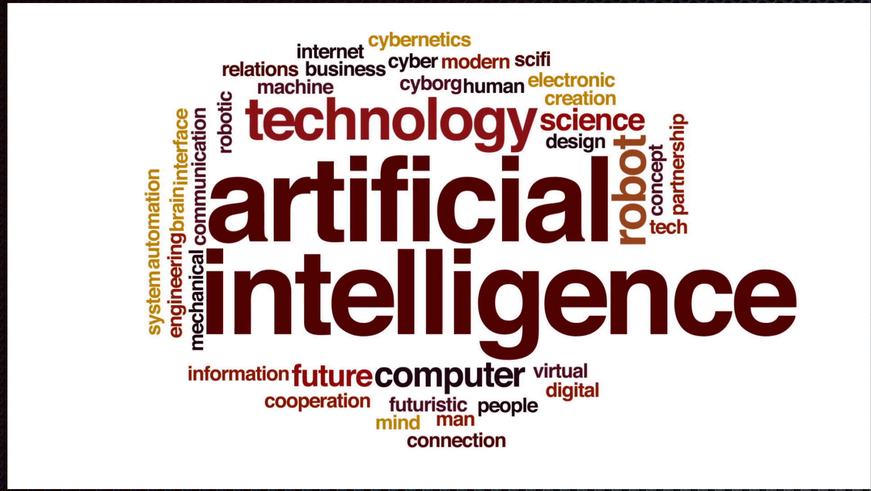
- It doesn’t matter how much data you can get
- It matters how you use it
  - Analytics
  - Translation to business practice
  - Follow-up

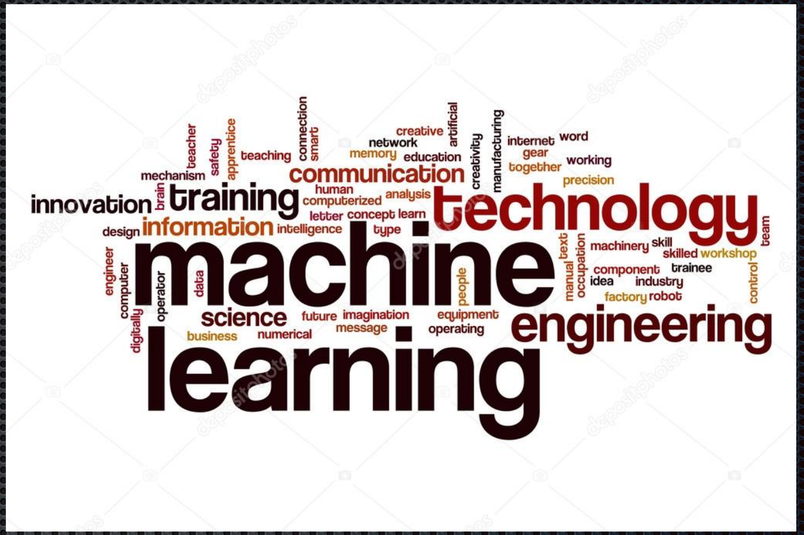

cybernetics
internet
cyber
modern
scifi
relations
business
machine
cyborg
human
electronic
creation
technology
science
design
robotic
interface
communication
brain
engineering
system
automation
artificial
robot
concept
partnership
tech
intelligence
mechanical
information
future
computer
virtual
digital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


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

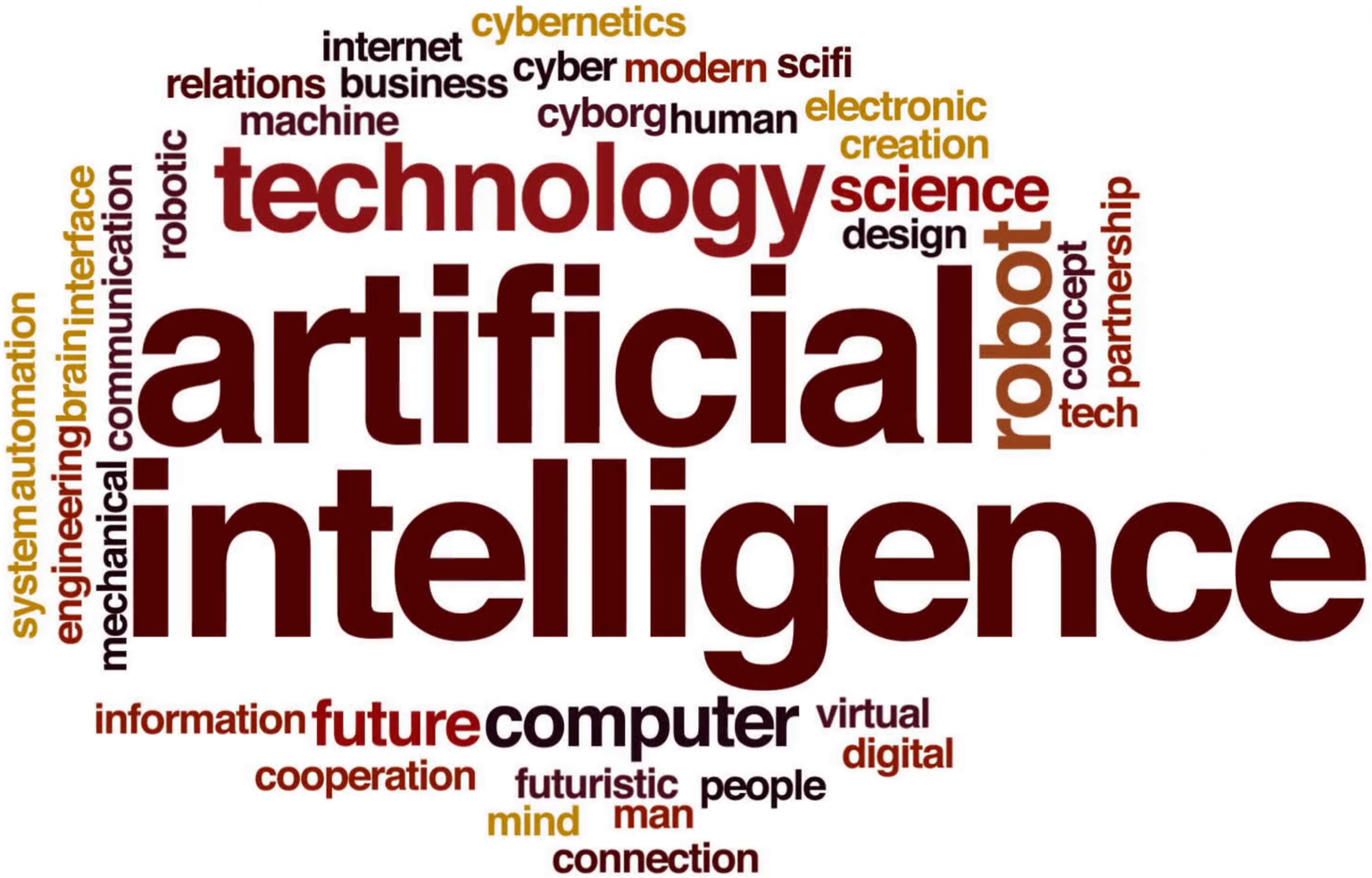

teacher
safety
apprentice
connection
smart
creative
artificial
network
memory
education
teaching
mechanism
brain
communication
human
computerized
analysis
innovation
training
letter
concept
learn
type
design
information
intelligence
creativity
manufacturing
internet
word
gear
together
working
precision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


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


innovation training
communication
technology
information
machine
engineering
science
learning

organization
scientific
descriptive
Statistics
collection
interpretation
data
statistical
Inference
theory
experiments
mathematical
probability
discipline

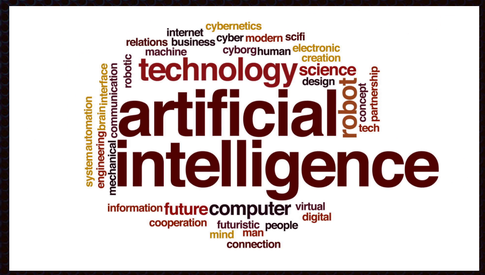
cybernetics
internet
relations
business
cyber
modern
scifi
machine
cyborg
human
electronic
creation
robotic
technology
science
system
automation
engineering
brain
interface
mechanical
communication
information
data
analysis
learning
machine
bigdata
deep-learning
modelling
information
reasoning
phenomena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


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

probability
prediction
experime
theo
discipline
Inference

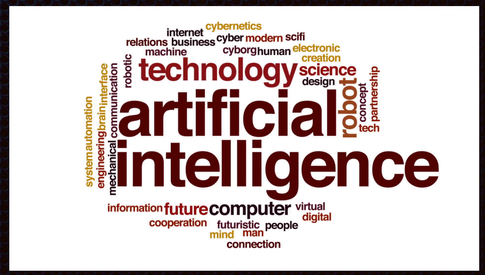
What is AI?

- 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



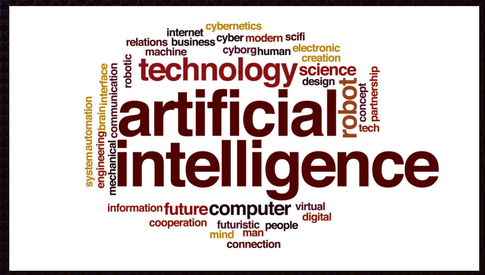
AI in Biotech

- 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



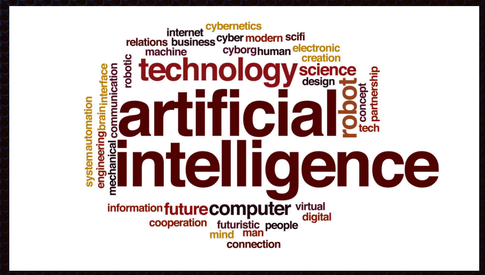
AI in Biotech

- Image analysis
  - Culling through radiographic data (X-rays, MRI, CT scans)
  - Better evaluation of drug benefit
- Natural language processing
  - Knowledge discovery and curation
  - Records-mining
- Bioinformatics
  - Processing, predicting, scaling up



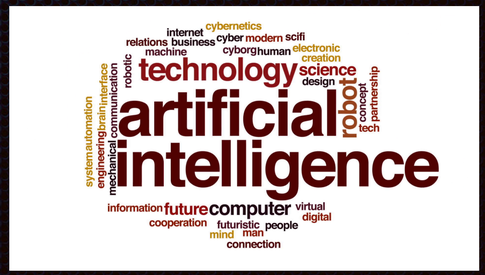
AI in Biotech

- 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

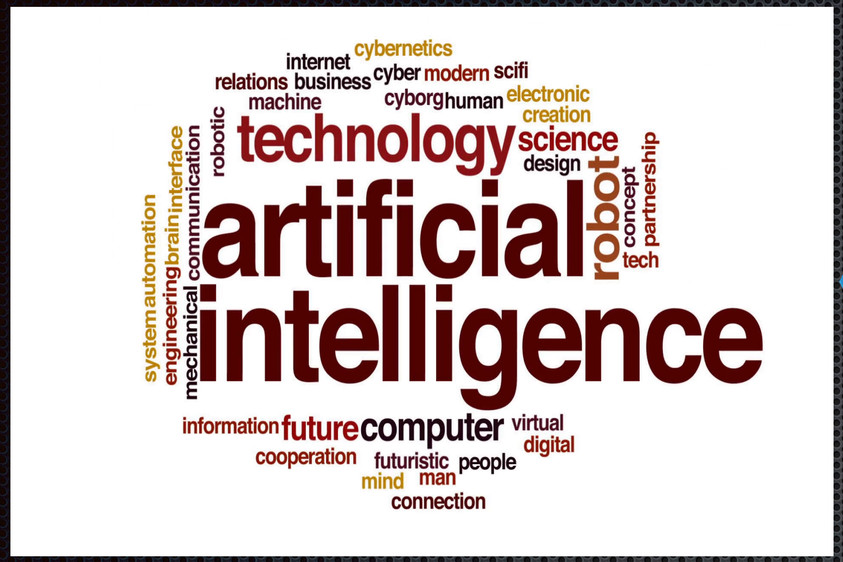


Are we there yet?

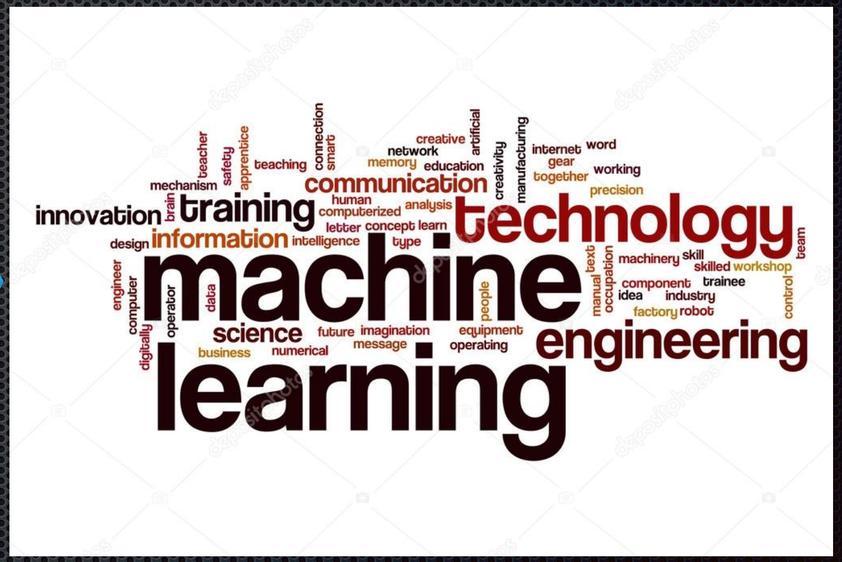
- Not really
  - Self-driving cars and planes on auto-pilot
  - Experimental situations
  - Some medical decision support

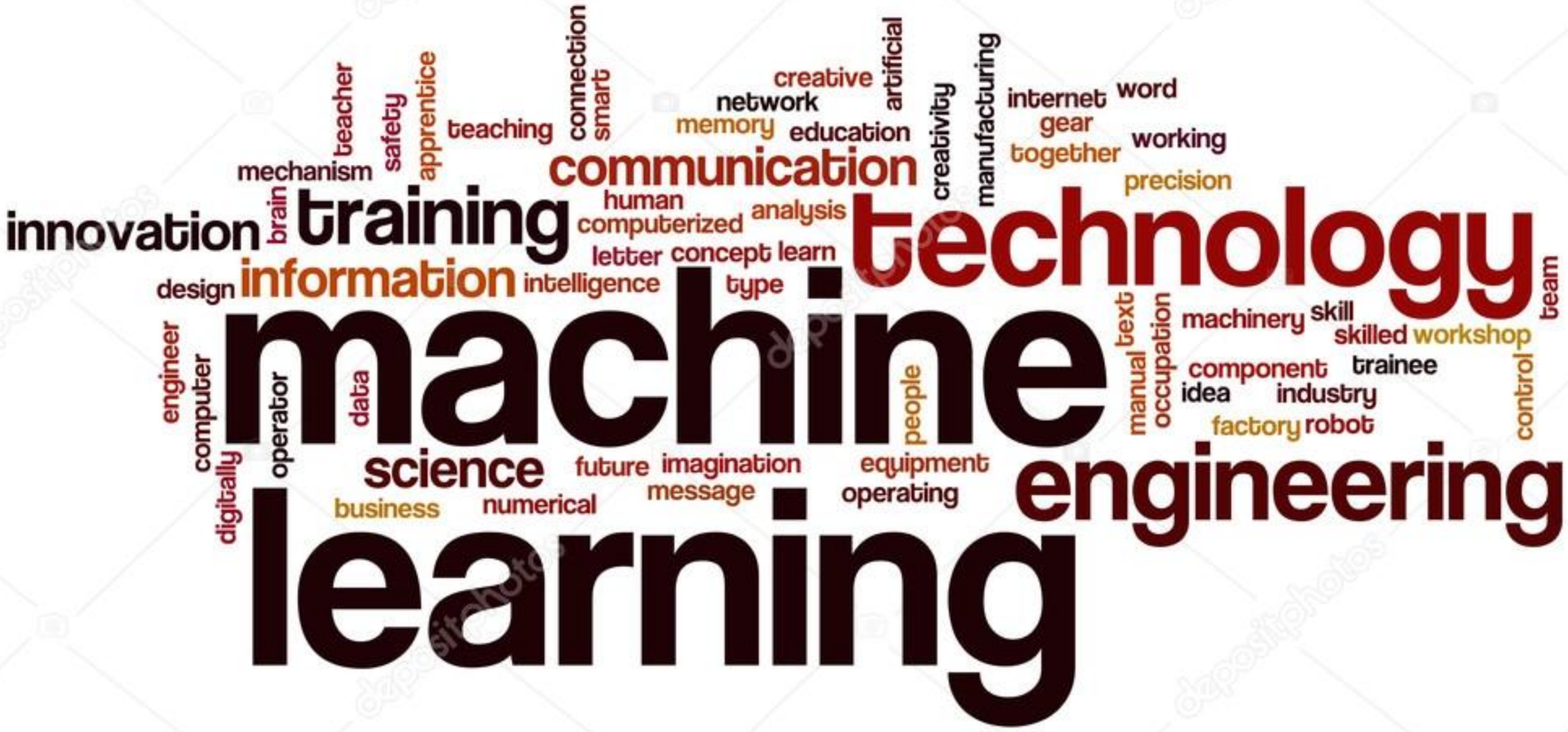


So where are we?



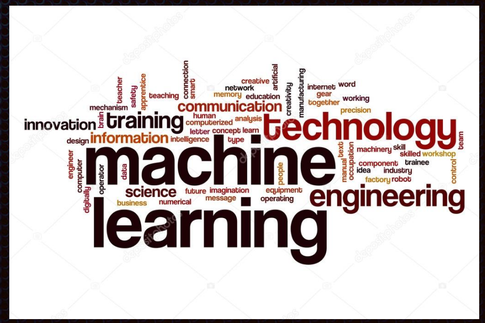
Common use




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

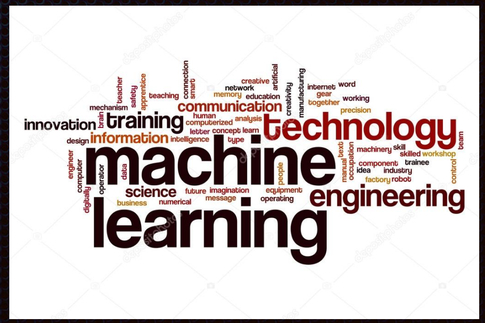
Machine learning

- 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



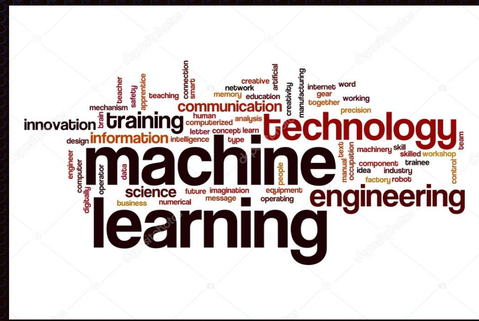
Machine learning success stories

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



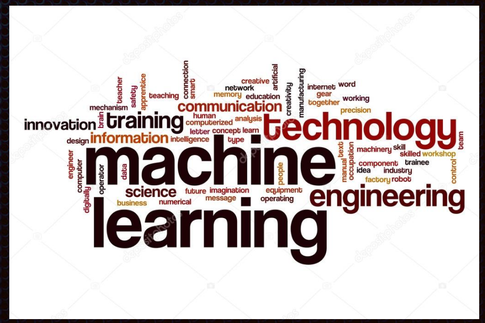
Machine learning success stories

- Identification of subgroups in various diseases, leading to more targeted interventions
  - EGFR & Lung Cancer
  - APoE & Alzheimer’s prevention
  - BRCA & Breast cancer prophylaxis
  - Immunotherapies
  - Biomarker generation for non-invasive assessments (ctDNA and other assays))
  - Gene editing (CRISP-R)



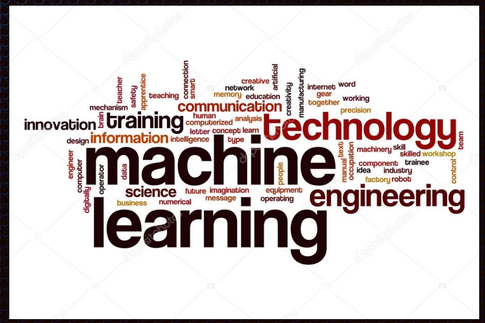
Machine learning success stories

- 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



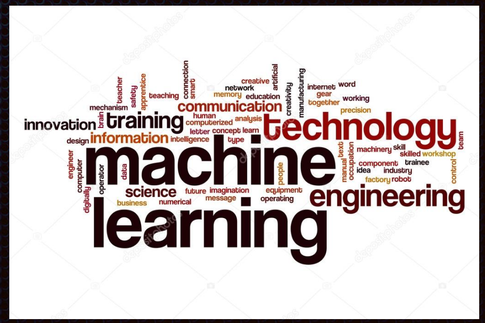
What is ML?

- 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?

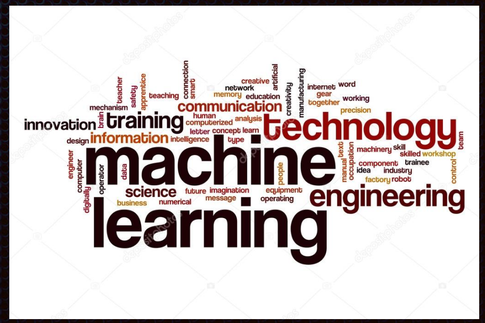
- Algorithms that are mainly meant for prediction
  - Supervised learning
- They cannot often express why the predictions happen
  - What factors influenced the prediction in what way?
  - It's a function of the complexity of the model and our inability to express or comprehend that complexity easily
- Increasing research, emphasis on "explainable AI" or XAI
  - LIME, SHAP, others
  - Intrinsically explainable models (Microsoft)



ML and Big Data

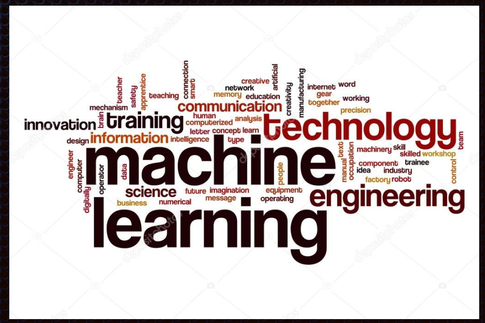
Machine learning is a greedy, hungry, voracious animal

with respect to data



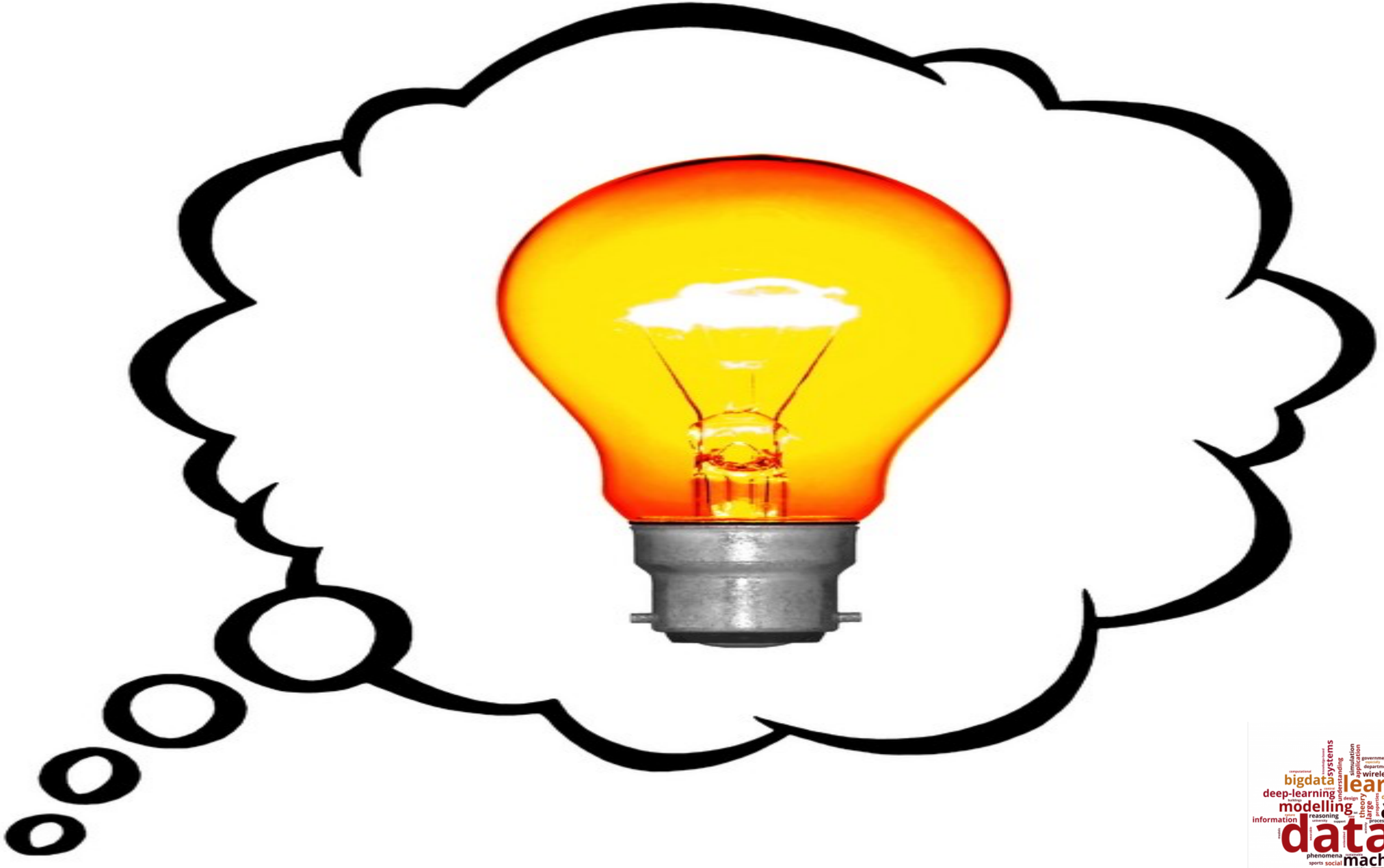
ML and Big Data

- 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



data
analysis
learning
machine
bigdata
modelling
deep-learning
materials
energy
research
networks
processing
statistics
performance
science
mining
signal
applications
information
wireless
coastal
topological
computer
systems
simulation
application
understanding
reasoning
technology
engineering
human
football
wave
vision
social
phenomena
sports
security
allocation
management
machine learning
statistical
smart
recognition
geometric
methods
space
analytics
development
mathematical
government
department
intelligence
robotics
working
stream
software
network
dynamics
linked
design
theory
large
process
properties
buildings
university
support
computational
environment
simulation
complex
related
networking
pattern
image
impact
dynamics
segmentation
language
transport
singular

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

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
research
networks
processing
statistics
mining
systems
energy
signal
computer
science
performance
vision
wave
football
social
phenomena
information
reasoning
wireless
coastal
topological
applications
statistical
recognition
smart
engineering
technology
human
security
space
geometric
methods
dynamics
segmentation
language
transport
singular
machine-learning
management
based
impact
allocation
pattern
image
imaging
discovery
sports
computational
computational-science
computational-methods
computational-mathematics
computational-modelling
computational-analysis
computational-data
computational-research
computational-science
computational-systems
computational-theory
computational-methods

Where’s the data

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
networks
processing
statistics
performance
science
mining
systems
signal
computer
information
reasoning
applications
wireless
vision
football
wave
human
technology
engineering
smart
recognition
statistical
social
phenomena
sports
development
coastal
mathematical
topological
geometric
methods
space
analytics
security
allocation
pattern
management
machine-learning
imaging
government
department
intelligence
simulation
application
understanding
theory
design
linked
process
properties
objects
robotics
software
stream
network
computational
sciences
dynamics
segmentation
language
transport
engineering
complex
image
impact
based
conversations
computational
computational
computational

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

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
processing
networks
statistics
performance
science
mining
signal
applications
information
systems
computer
wireless
coastal
topological
reasoning
technology
engineering
vision
football
wave
human
social
phenomena
statistical
smart
recognition
geometric
methods
space
computational
government
mathematical
department
intelligence
simulation
application
understanding
robotics
stream
software
network
sciences
analytics
dynamics
segmentation
language
transport
singular
machine-learning
management
based
security
allocation
pattern
impact
image
discovery
computational
computational
computational

Where’s the data

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

All told, 2 Zb* per year

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
networks
processing
statistics
mining
performance
science
signal
computer
applications
systems
information
vision
football
wave
social
phenomena
reasoning
technology
engineering
human
statistical
geometric
methods
space
wireless
coastal
mathematical
topological
government
department
intelligence
simulation
application
understanding
theory
design
development
linked
analytics
smart
recognition
machine-learning
management
security
allocation
sports
pattern
imaging
process
properties
objects
robotics
working
looking
stream
software
network
computational
computational
computational

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

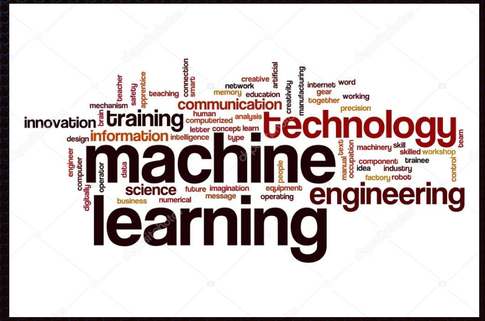
AAAS: Machine learning 'causing science crisis'

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

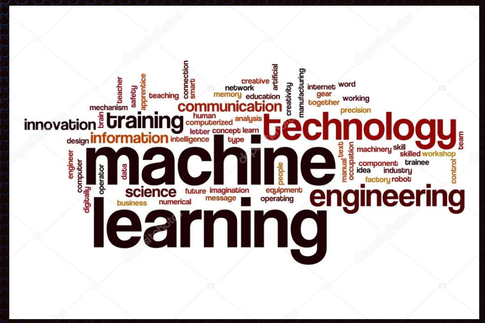
AAAS: Machine learning 'causing science crisis'

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

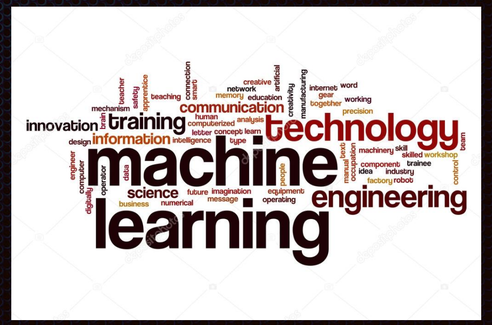
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed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

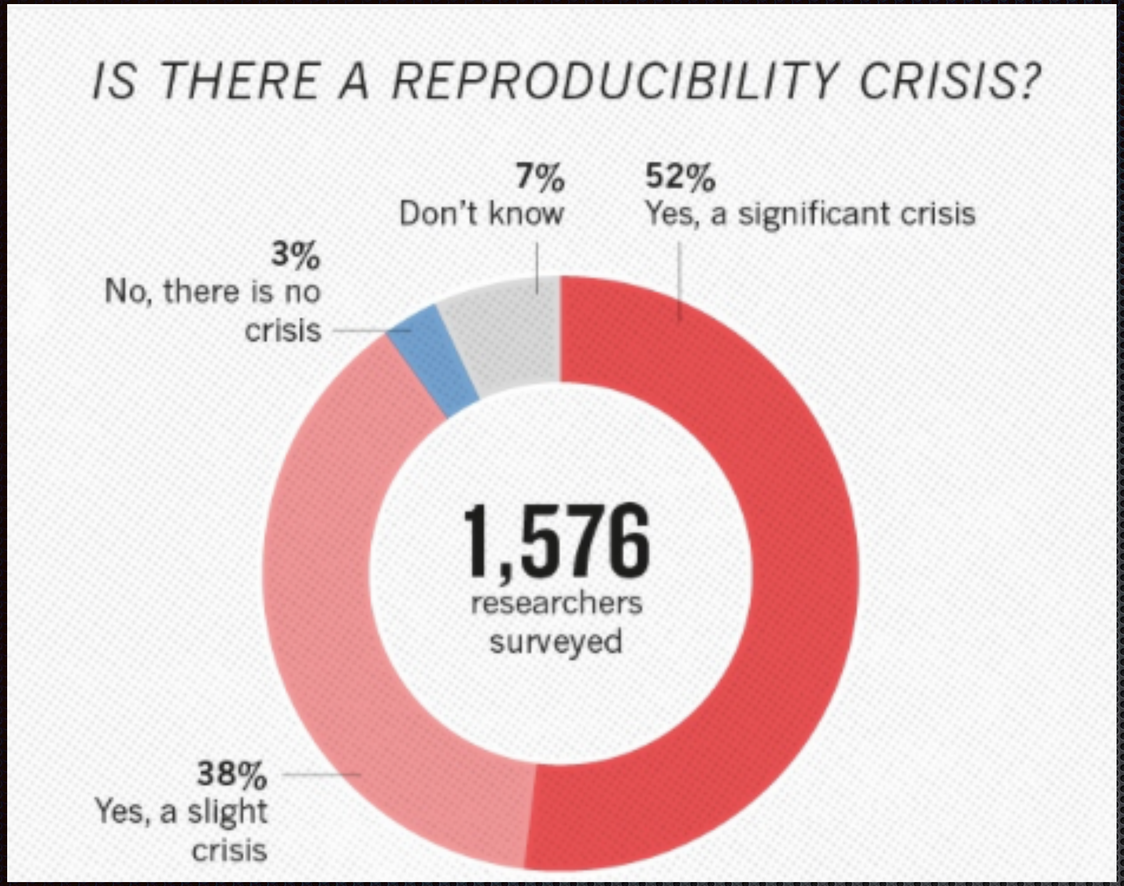
Citation: Ioannidis JPA (2005) Why most published

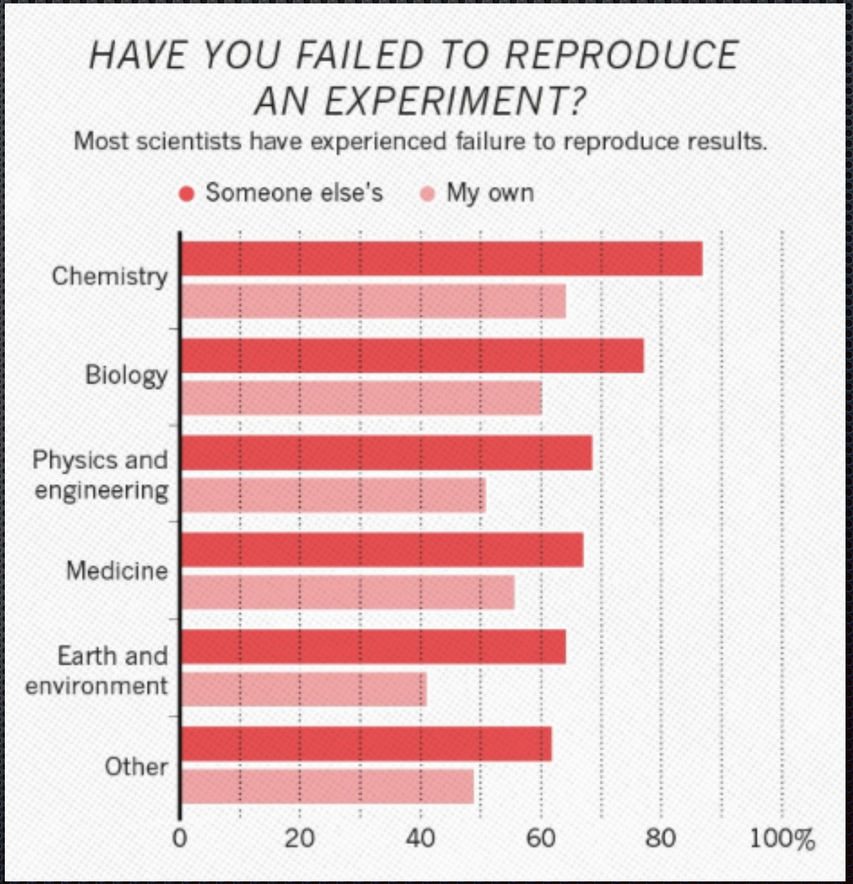
Pharma report:

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
materials
energy
research
networks
processing
statistics
mining
science
performance
signal
systems
computer
information
vision
football
wave
social
phenomena
human
technology
engineering
statistical
recognition
smart
wireless
coastal
topological
applications
reasoning
space
methods
geometric
dynamics
analytics
sciences
mathematical
intelligence
department
government
simulation
application
understanding
theory
large
process
development
design
linked
security
allocation
pattern
management
based
sports
machine-learning
image
impact
robotics
working
looking
stream
software
network
computational
university
support
complex
related
retrieval
networking
conversations
conversation
imaging
discovery
discover
semantic
semantics
segmentation
language
transport
singular

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 <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

Need to <u>find</u> the right Big Data for your question

⬍

Need to <u>see</u> 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

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
networks
processing
statistics
performance
science
mining
systems
signal
computer
applications
information
wireless
coastal
topological
statistical
vision
football
wave
social
sports
phenomena
human
technology
engineering
smart
recognition
geometric
methods
space
reasoning
security
allocation
mathematical
department
intelligence
simulation
application
understanding
theory
design
development
linked
analytics
dynamics
segmentation
language
transport
regular
machine-learning
management
based
impact
pattern
networking
method
robotics
working
software
stream
computational
government
especially
adaptive
objective
properties
process
image
imaging
discovery
semantics
semantic
conversations
complex
compute
relational
related
retrieval
network
cluster
image-processing
image-analysis
image-segmentation
image-classification
image-recognition
image-retrieval

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

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
processing
networks
mining
science
performance
statistics
signal
computer
applications
information
reasoning
phenomena
social
sports
human
vision
engineering
statistical
wireless
coastal
topological
systems
simulation
understanding
mathematical
department
intelligence
development
technology
football
wave
security
allocation
pattern
recognition
smart
machine-learning
management
imaging
process
space
geometric
methods
dynamics
segmentation
language
transport
engineer
analytics
sciences
design
theory
large
properties
linked
robotics
working
software
stream
computational
government
application
objects
object
model
models
network
networks
computer
signal
statistical
recognition
smart
machine-learning
management
based
impact
image
images
conversations
conversation
computational
computational
computational

Biases and ethics

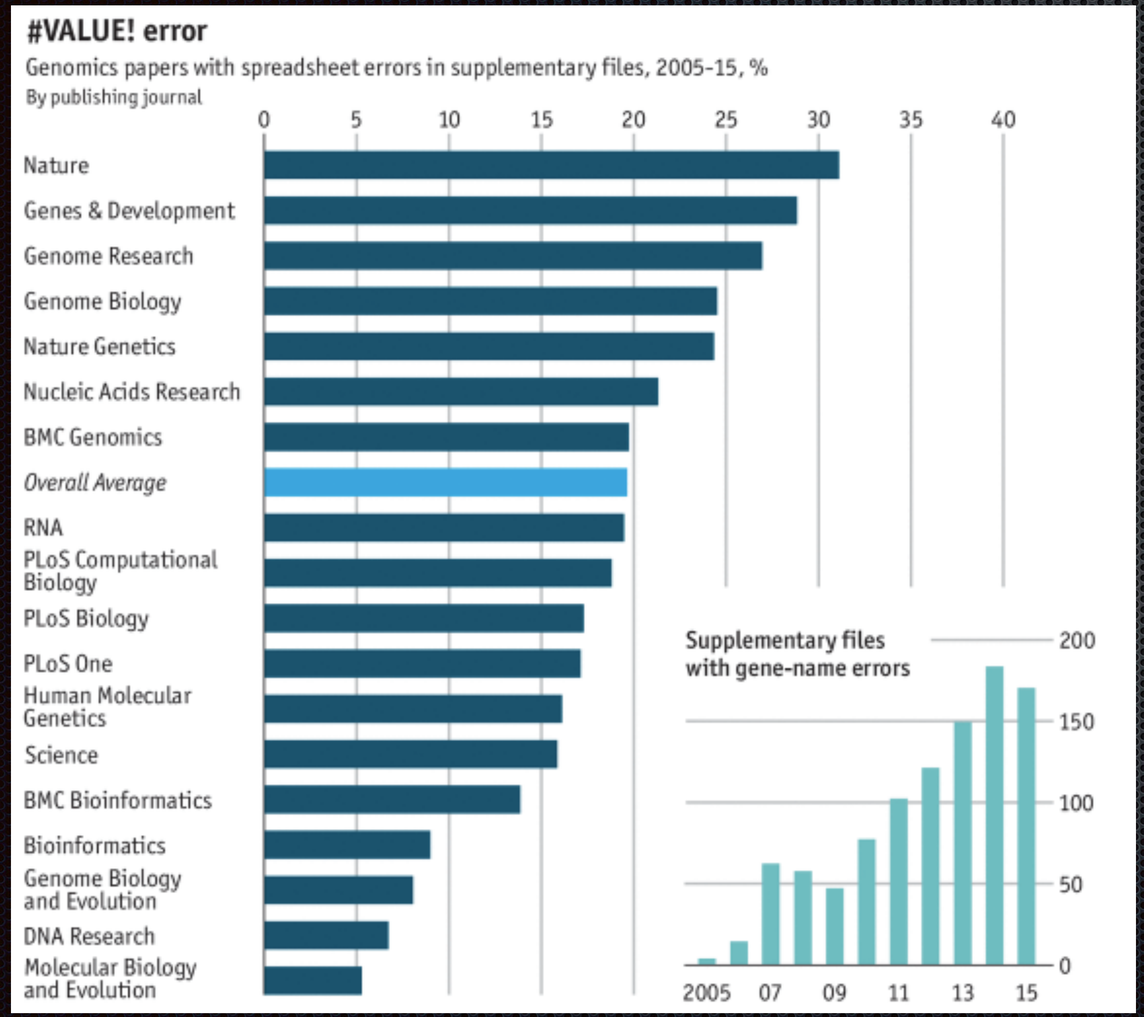
- aka, we like shiny new things

- 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



data
machine
learning
analysis
modelling
bigdata
deep-learning
materials
energy
research
networks
processing
statistics
performance
science
mining
vision
signal
computer
systems
information
reasoning
phenomena
social
human
wave
football
engineering
smart
recognition
statistical
machine-learning
topological
wireless
coastal
applications
theory
space
geometric
methods
dynamics
technology
security
allocation
pattern
management
based
sports
government
department
intelligence
mathematical
robotics
working
understanding
simulation
application
process
development
linked
analytics
sciences
design
computational
software
network
stream
language
segmentation
regular
impact
image
imaging
discovery
decision
machine-learning
conversations
complex
computer
retrieval
networking
related
automatic
adaptive
properties
objects
artificial
intelligence
robotics
working
coastal
mathematical
department
wireless
topological
applications

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

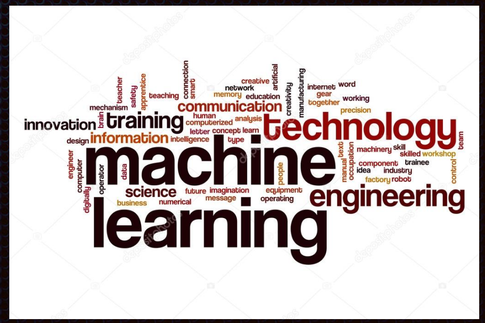
data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
processing
statistics
mining
networks
signal
computer
science
performance
applications
information
systems
wireless
coastal
topological
vision
football
wave
social
phenomena
reasoning
technology
engineering
statistical
smart
recognition
machine-learning
human
pattern
security
allocation
sports
management
mathematical
government
department
intelligence
simulation
application
understanding
computational
design
theory
large
process
properties
development
linked
space
analytics
methods
geometric
segmentation
language
transport
dynamic
singular
impact
based
image
imaging
discovery
networking
method
related
complex
conversations
complex
semantic
network
software
stream
robotics
working
looking
objects
objective
building
sciences
sensing
signal
image
pattern
impact
semantic
management
based
machine-learning
discovery
recognition
smart
statistical
engineering
performance
technology
science
mining
human
pattern
security
allocation

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

data
analysis
learning
machine
modelling
bigdata
deep-learning
materials
energy
research
networks
processing
statistics
performance
science
mining
vision
signal
computer
information
reasoning
applications
wireless
coastal
topological
technology
engineering
statistical
smart
recognition
human
football
wave
social
phenomena
sports
systems
simulation
application
understanding
government
department
intelligence
mathematical
theory
design
development
space
geometric
methods
dynamic
segmentation
language
transport
singular
machine-learning
management
based
allocation
security
networking
vertical
conversations
twitter
twitter
network
software
stream
robotics
working
looking
especially
objects
objective
linked
properties
process
analytics
sciences
image
impact
imaging
discovery
semantic
computational
building
urban
sensing
support
sensor
model
models
modeling
computational
computational
computational

“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.app

Abhijit Dasgupta

Director/Chief Data Scientist

ARAASTAT

Zansors