GE 461
Introduction to Data Science

Spring 2022

Course Website

All course related material will be provided in the course website

- http://www.cs.bilkent.edu.tr/~ge461/2022Spring
- Check regularly for announcements!

Weekly topics, instructors are stated.

Slides will be provided here.

Assignments released on Moodle.

Various external links to other similar courses and online textbooks.

Instructors

Cross-department Course with Multiple Instructors.

CS Department

S. Aksoy,
C. Alkan,
S. Arashloo,
F. Can,
A.E. Cicek,
H. Dibeklioglu,
A. Dundar,
I. Korpeoglu,
E. Tuzun

EE Department

- T. Cukur,
- C. Tekin

IE Department

- S. Dayanik

TAs will be announced on the Course Website. They will be from all 3 departments.

Location & Time

When: Tue 10:30 – 12:20 and Thursday 15:30 – 17:30.

Where: B-204.

What: A lot! Introduction to data science fundamentals, techniques and applications; data collection, preparation, storage and querying; parametric models for data; models and methods for fitting, analysis, evaluation, and validation; dimensionality reduction, visualization; various learning methods, classifiers, clustering, data and text mining; applications in diverse domains such as business, medicine, social networks, computer vision; breadth knowledge on topics and hands-on experience through projects and computer assignments.

See weekly coverage.

Grading Policy

Final: 40%

Project: 60%

Multiple computer/programming/exercise assignments of various sizes.

A project can be assigned earlier than the indicated date on the weekly plan.

Projects can be individual or group based. Instructors will decide.

Projects will be uploaded to Moodle.

Piazza will be used as the forum to discuss.

Attendance:

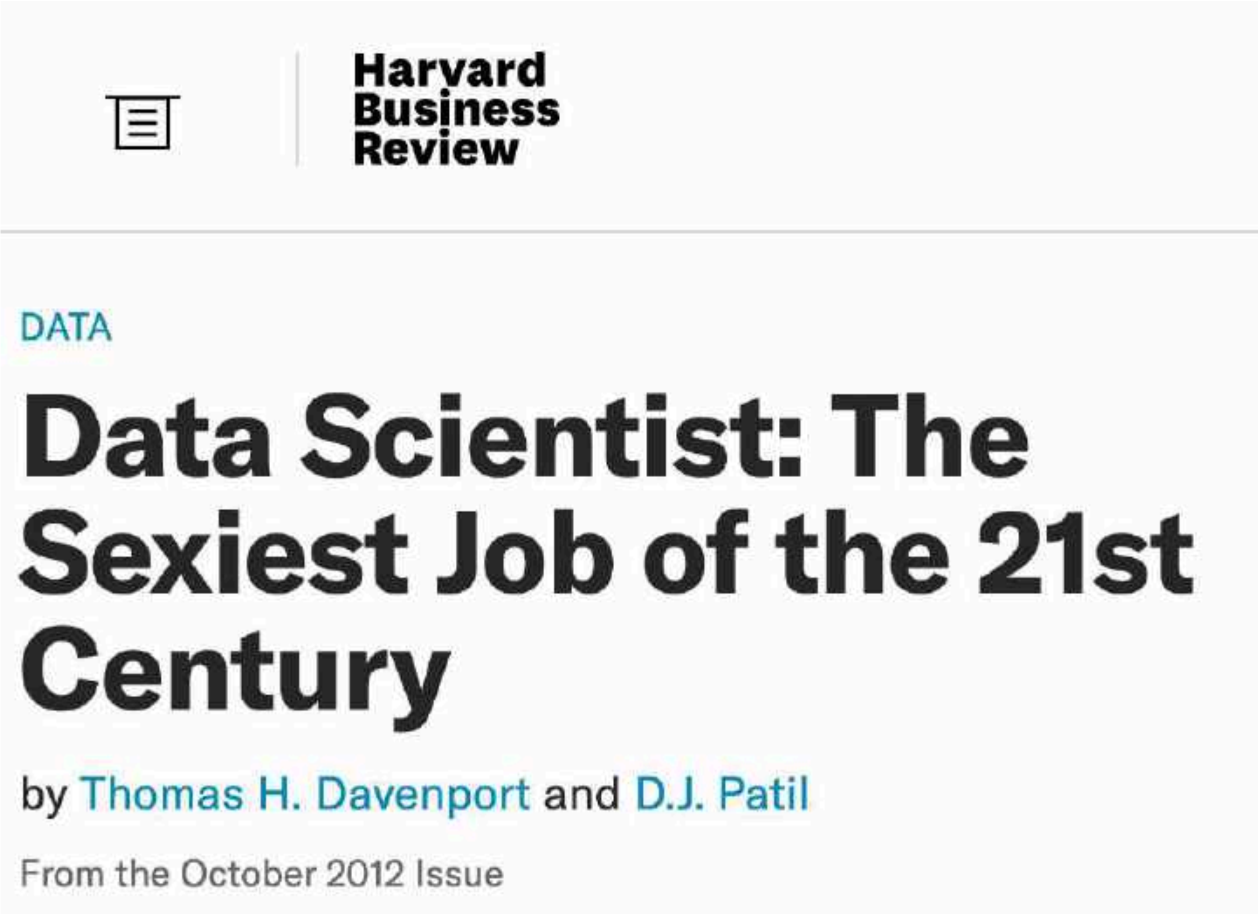
A student who misses more than **9 hours** will fail the course.

What is Data Science?

The field of study that uses various **methods** to extract useful insights and knowledge from the **data** to make data-driven decisions.

Methods can include/require, domain expertise, programming skills (i.e., scripting to process data), statistical modeling (i.e., machine learning algorithms), visualization techniques.

Usually performed on big data.



Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

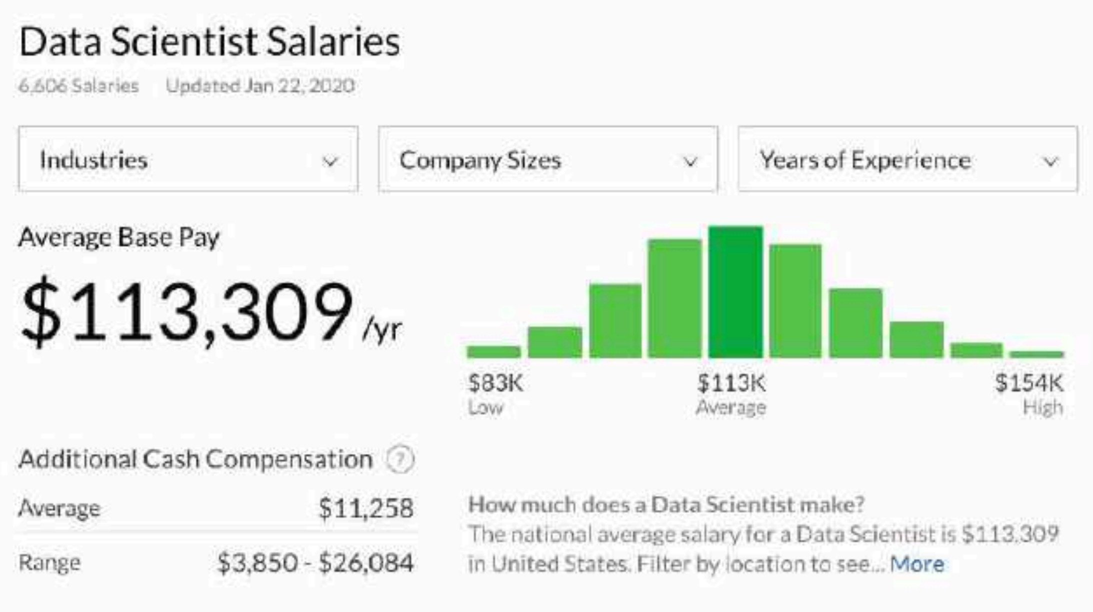
by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Recommended readings:

http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



Data Scientist Salaries

6,606 Salaries Updated Jan 22, 2020

Industries ⌄ | Company Sizes ⌄ | Years of Experience ⌄

Average Base Pay

$113,309 /yr

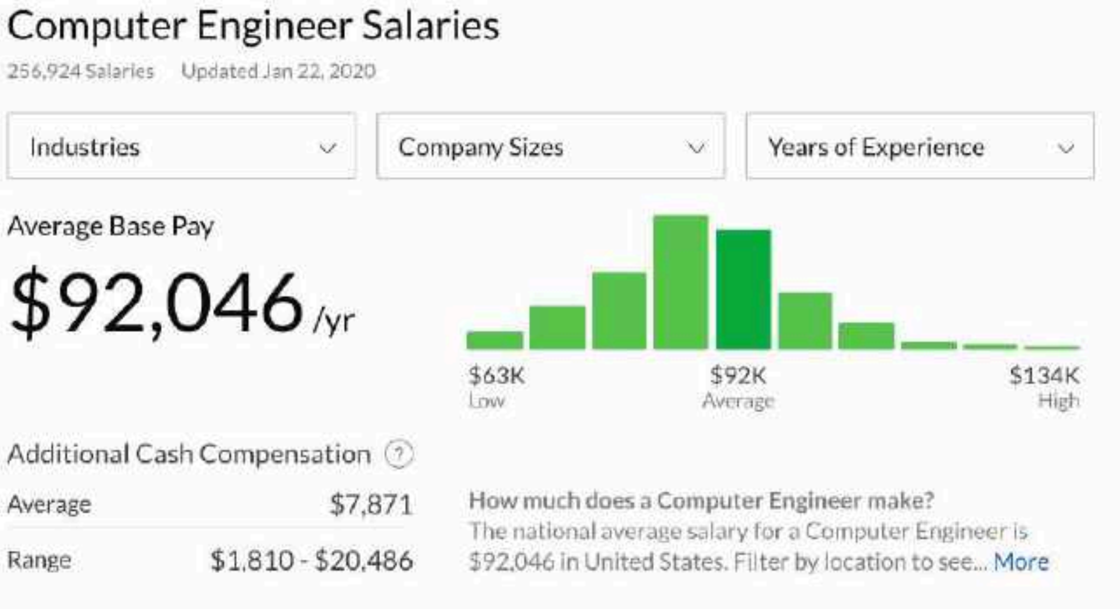
Additional Cash Compensation ?

Average	$11,258
Range	$3,850 - $26,084

How much does a Data Scientist make?

The national average salary for a Data Scientist is $113,309 in United States. Filter by location to see... More

vs



Computer Engineer Salaries

256,924 Salaries Updated Jan 22, 2020

Industries ⌄ | Company Sizes ⌄ | Years of Experience ⌄

Average Base Pay

$92,046 /yr

Additional Cash Compensation ?

Average	$7,871
Range	$1,810 - $20,486

How much does a Computer Engineer make?

The national average salary for a Computer Engineer is $92,046 in United States. Filter by location to see... More

What is NOT Data Science?

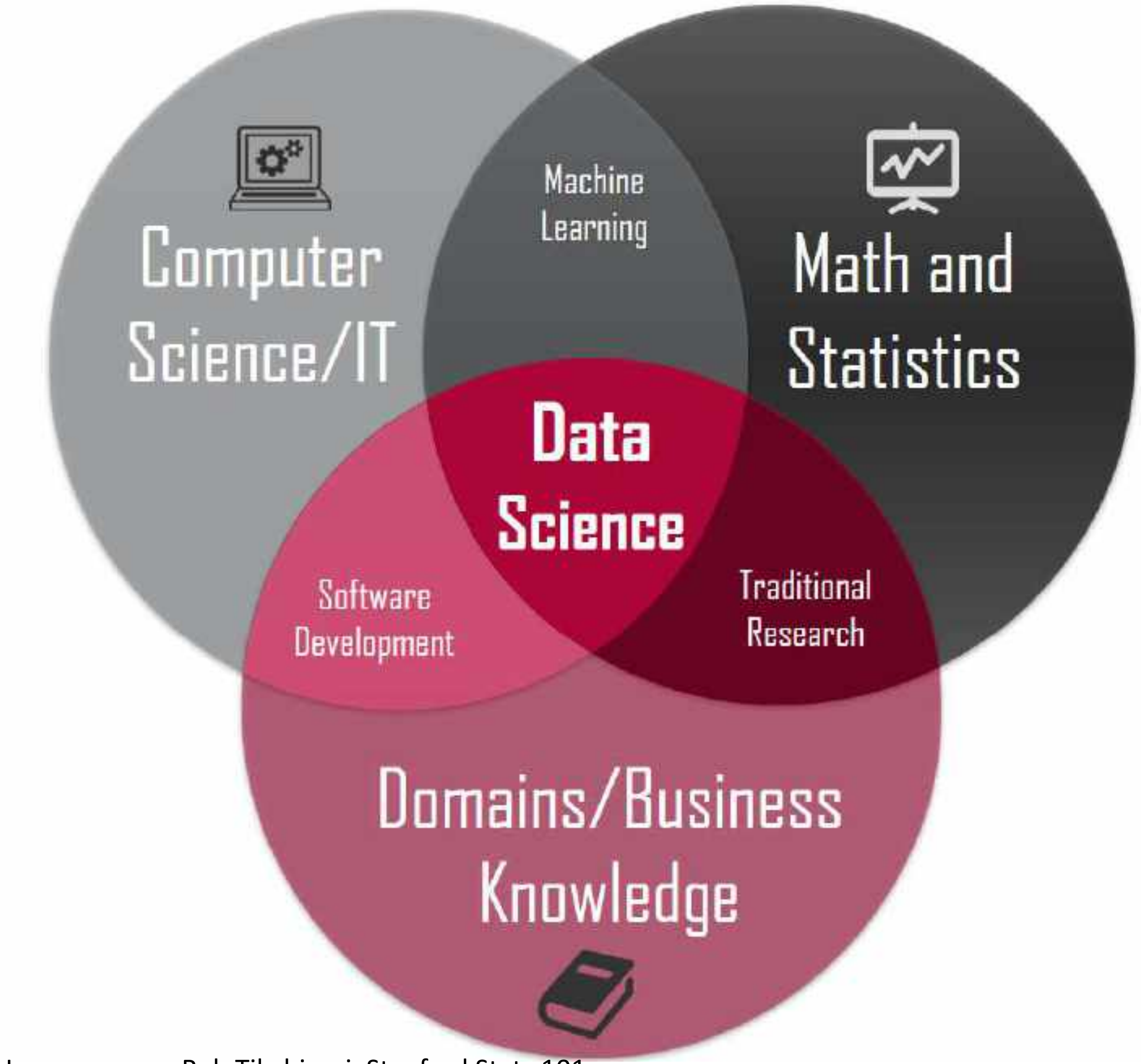
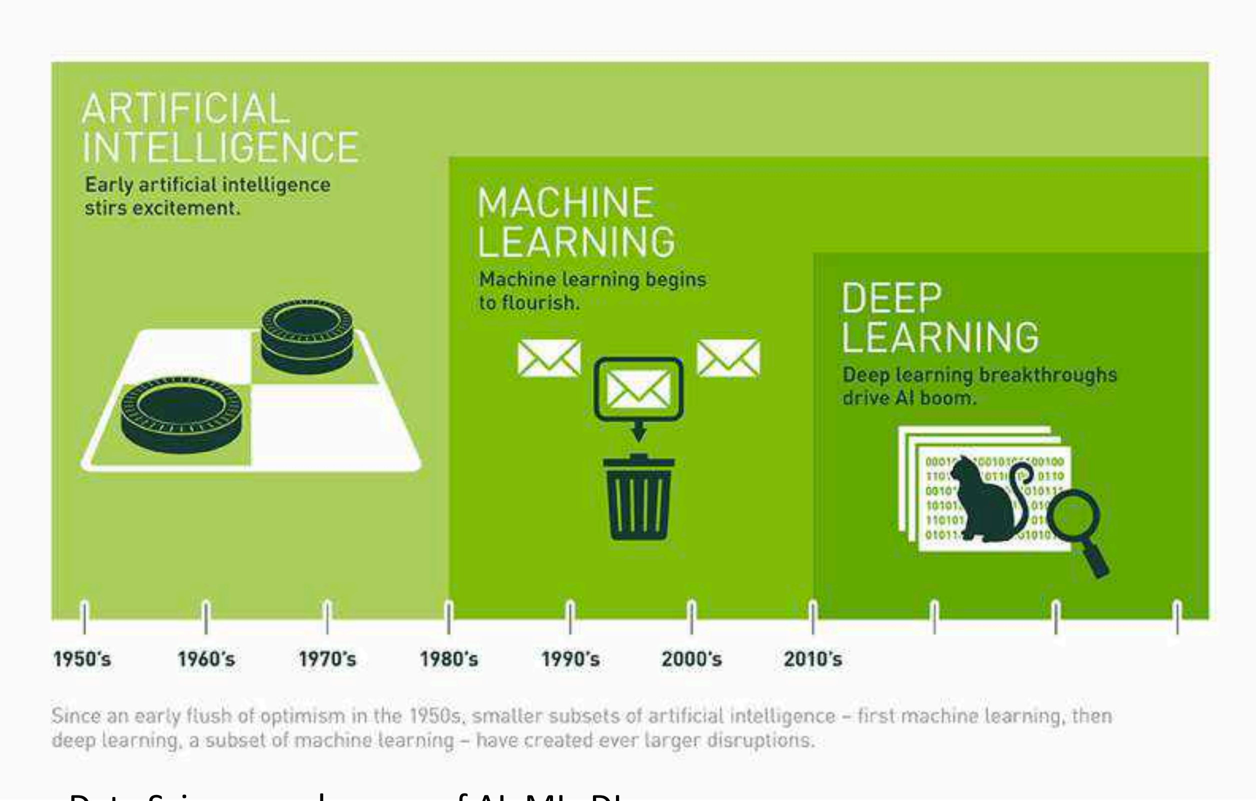


Image source: Rob Tibshirani, Stanford Stats 101

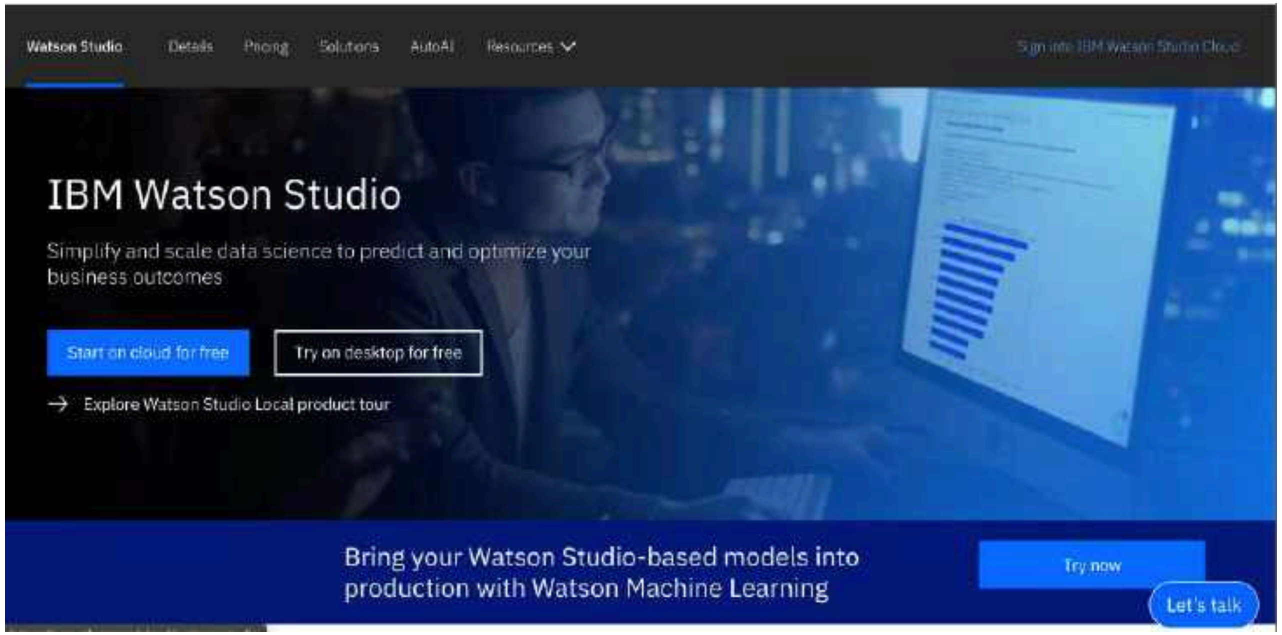

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Data Science makes use of AI, ML, DL

https://blogs.nvidia.com/blog/2016/07/29/whatsdifference-artificial-intelligence-machine-learningdeep-learning-ai/

What is NOT Data Science? Example

An AI breakthrough in 2011, now empowers Data Science.



IBM Watson Studio

Simplify and scale data science to predict and optimize your business outcomes

Bring your Watson Studio-based models into production with Watson Machine Learning

Data Science vs Other Related Terms

Many terms are used interchangeably; vague definitions.

Data Science aims at finding the right questions, more predictive analysis. Somewhat involves creativity.

On the other hand, **Business Intelligence** aims helping in the decision making of a business based on past data.

Data mining is a technique that searches for patterns in the data and can be considered as a tool of Data Science.

For example: Baby diapers and beer are frequently bought together.

Data analytics aims at analyzing data to find answers to concrete questions.

For instance, optimizing the teller processes at the bank to serve more customers.

It is a tool for **Business Intelligence.**

Why Now? Some advances

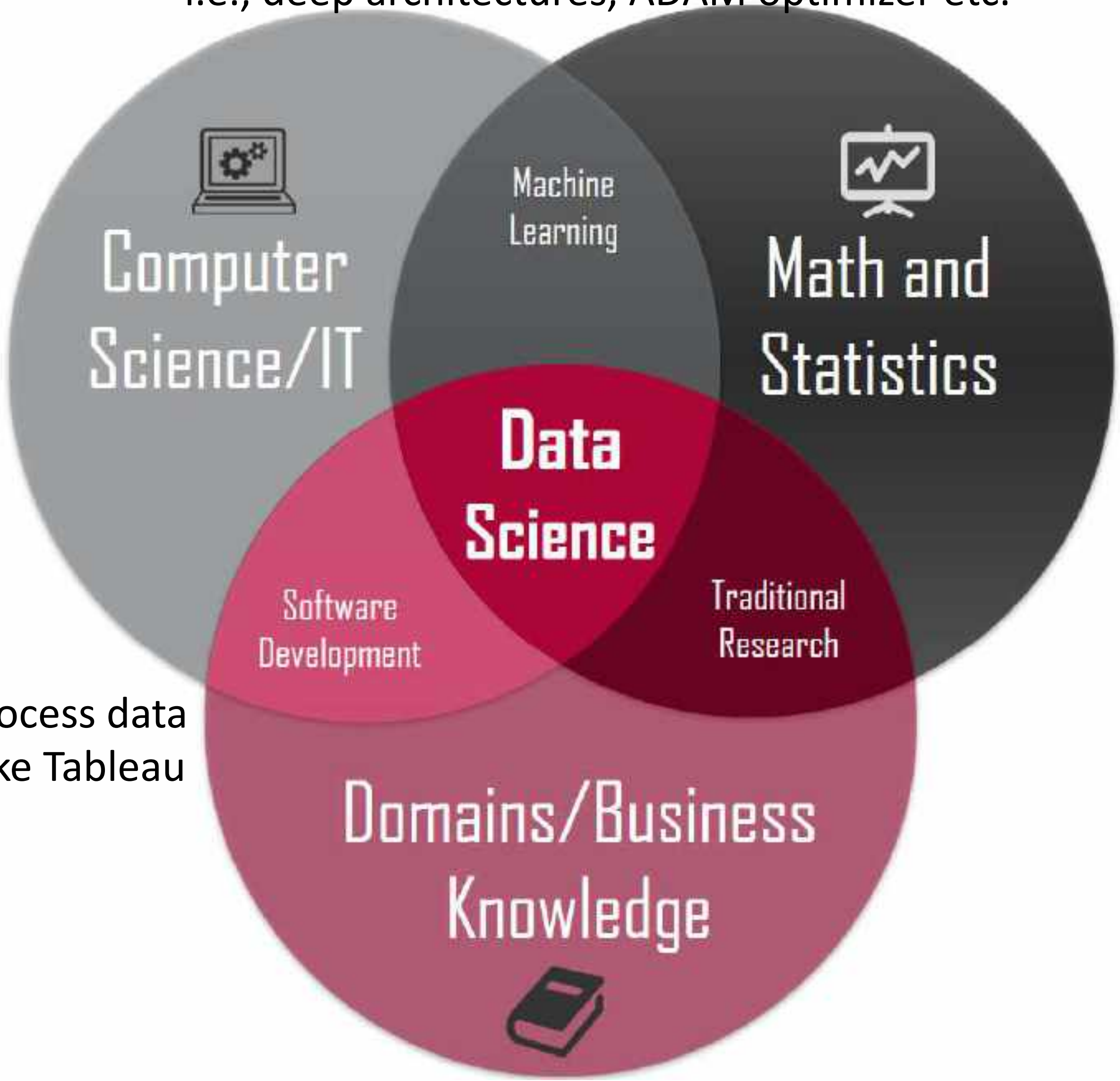
Better machine learning algorithms
i.e., deep architectures, ADAM optimizer etc.

Faster Computers

GPU power to crunch large datasets

Better ways (NoSQL) to manage
Data (Hadoop, Hive, HBase)

Python and R vs SAS and SPSS to process data
Advanced data visualization tools like Tableau



+ big data

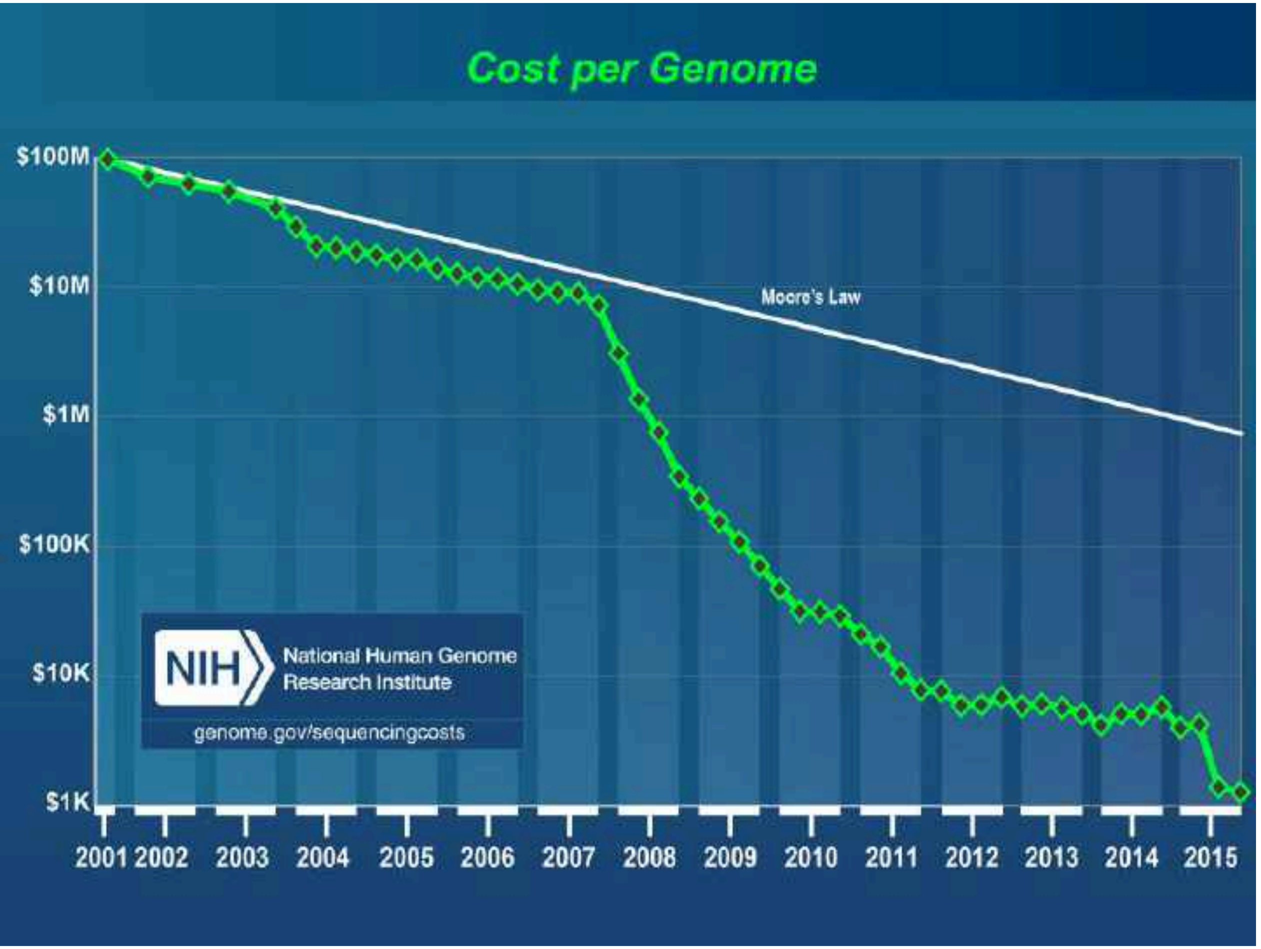
Data is ubiquitous
Cheap to produce and store

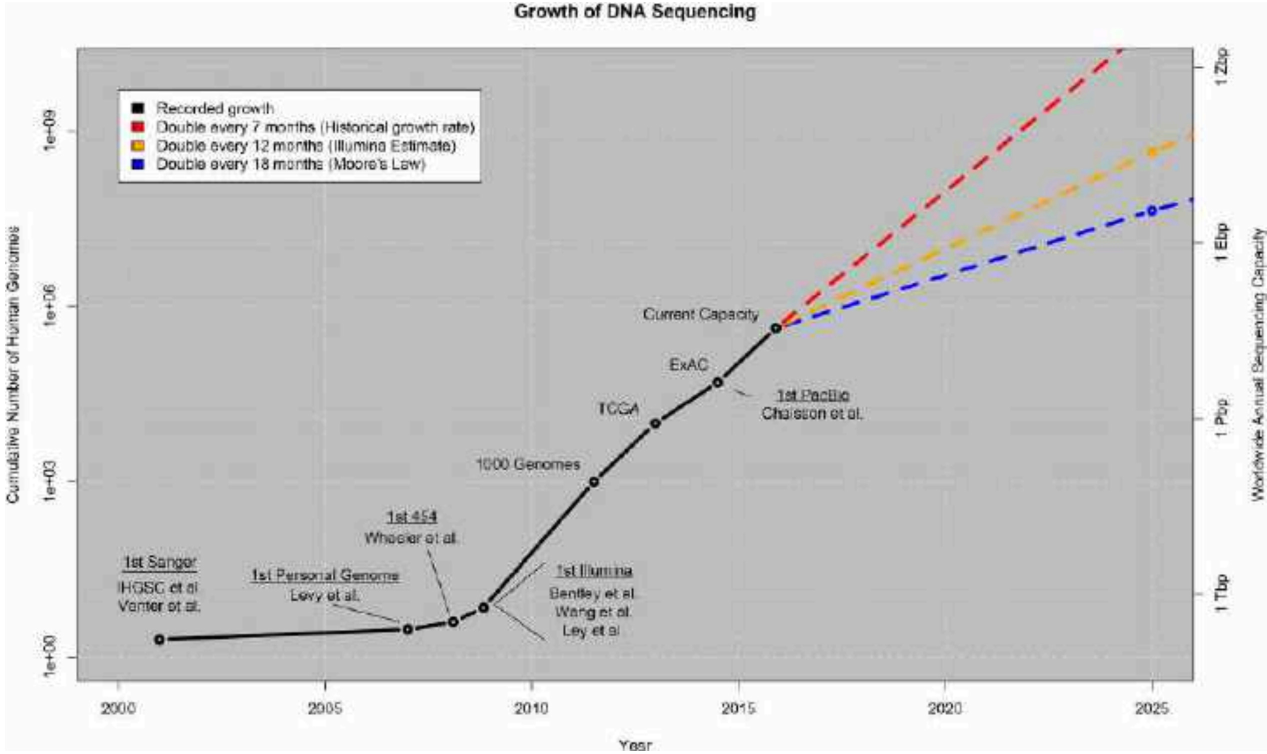
amazon S3

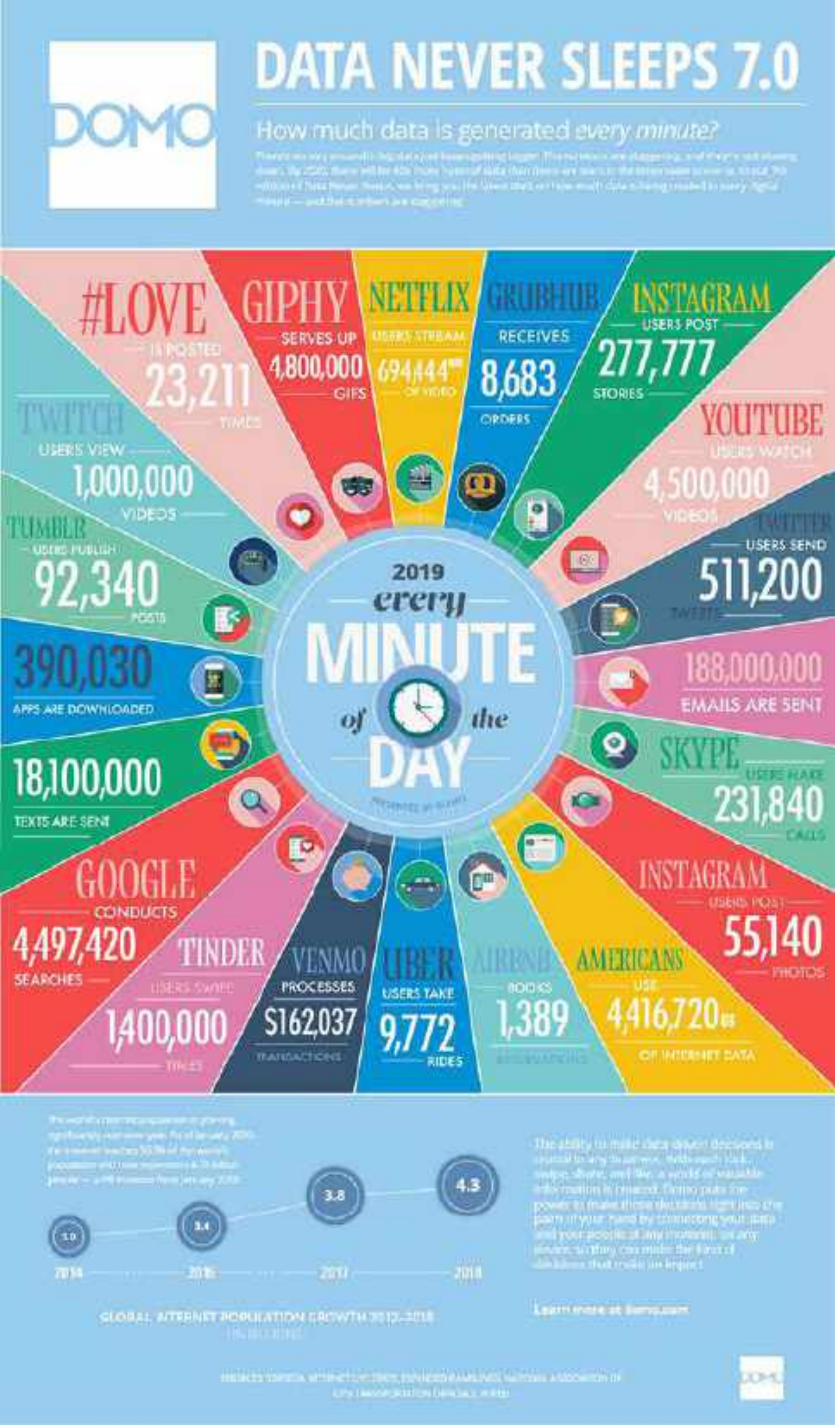
Image source: Rob Tibshirani, Stanford Stats 101

Big Data

Data is easy to produce, cheap to store. One example from genomics.







DATA NEVER SLEEPS 7.0

How much data is generated *every minute?*

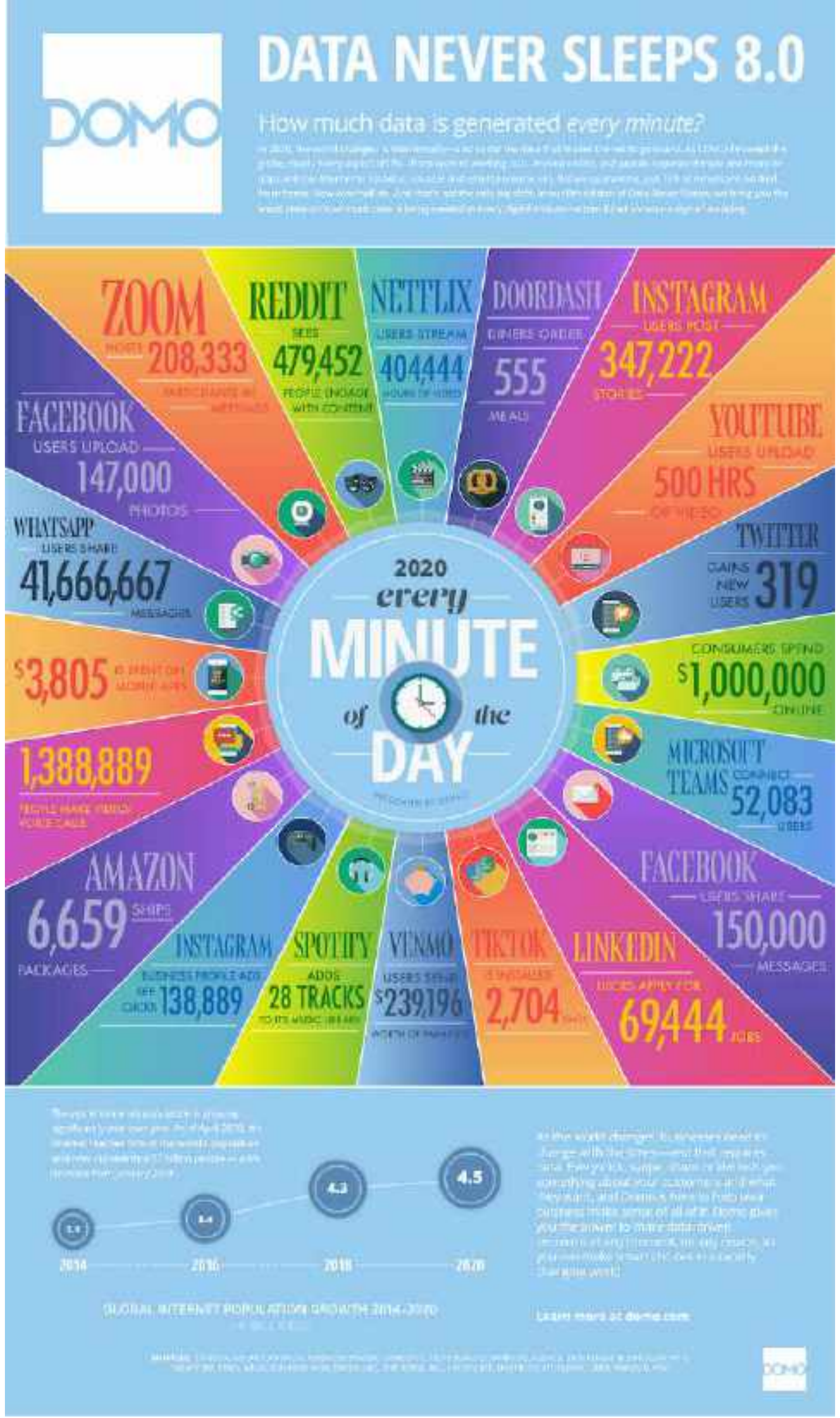
[illegible]

The world's internet population is growing significantly year-over-year. As of January 2019, the internet reaches 56.1% of the world's population and now represents 4.39 billion people — a 9% increase from January 2018.

The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at domo.com

SOURCES: [illegible]



DATA NEVER SLEEPS 8.0

How much data is generated *every minute?*

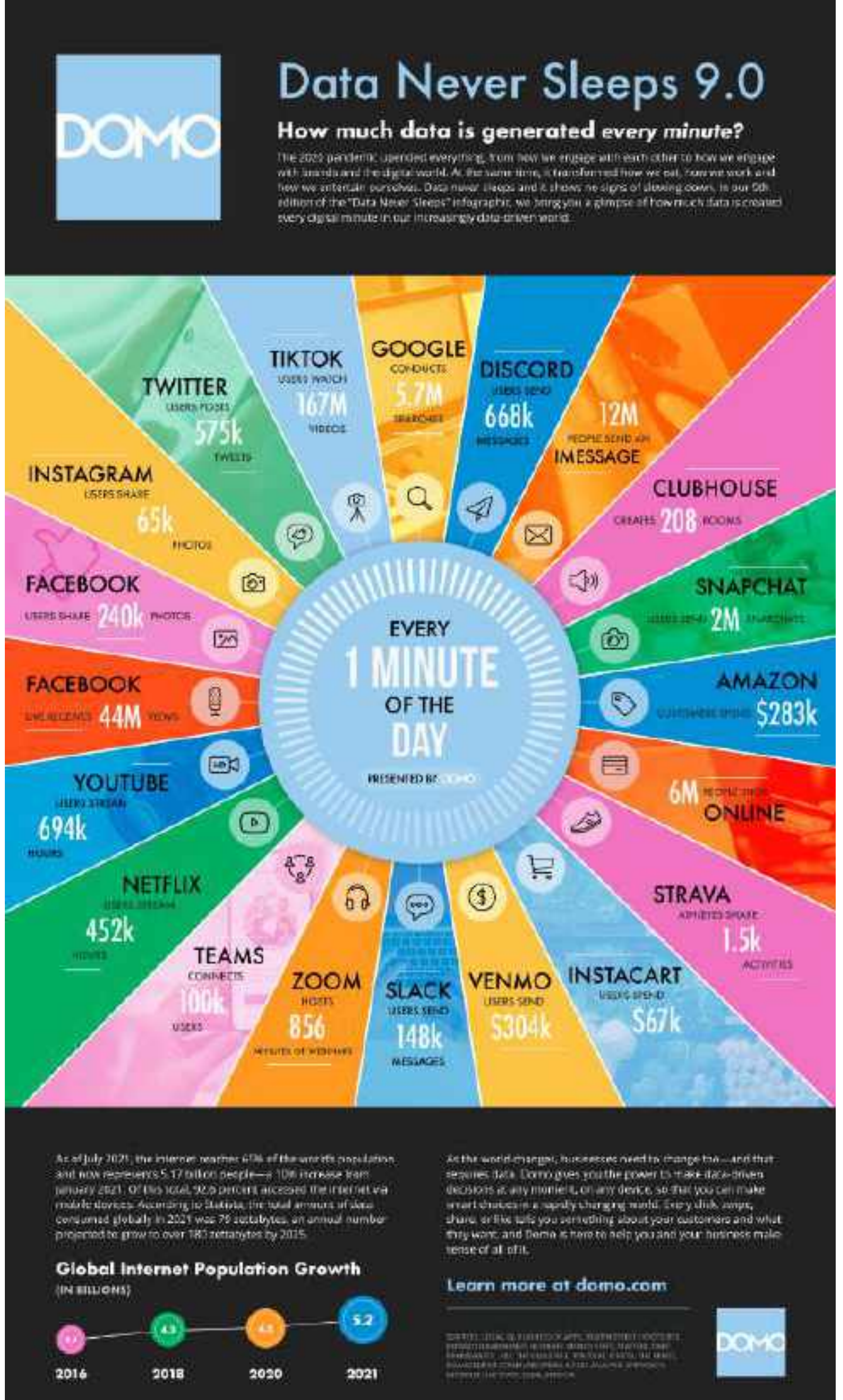
[illegible]

The world's internet population is growing significantly year-over-year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people—a 3% increase from January 2019.

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

Learn more at domo.com

SOURCES: [illegible]



Data Never Sleeps 9.0

How much data is generated every minute?

The 2020 pandemic upended everything, from how we engage with each other to how we engage with brands and the digital world. At the same time, it transformed how we eat, how we work and how we entertain ourselves. Data never sleeps and it shows no signs of slowing down. In our 9th edition of the "Data Never Sleeps" infographic, we bring you a glimpse of how much data is created every digital minute in our increasingly data-driven world.

As of July 2021, the internet reaches 65% of the world's population and now represents 5.17 billion people—a 10% increase from January 2021. Of this total, 92.6 percent accessed the internet via mobile devices. According to Statista, the total amount of data consumed globally in 2021 was 79 zettabytes, an annual number projected to grow to over 180 zettabytes by 2025.

Global Internet Population Growth

(IN BILLIONS)

As the world changes, businesses need to change too—and that requires data. Domo gives you the power to make data-driven decisions at any moment, on any device, so that you can make smart choices in a rapidly changing world. Every click, swipe, share, or like tells you something about your customers and what they want, and Domo is here to help you and your business make sense of all of it.

Learn more at domo.com

SOURCES: [illegible]

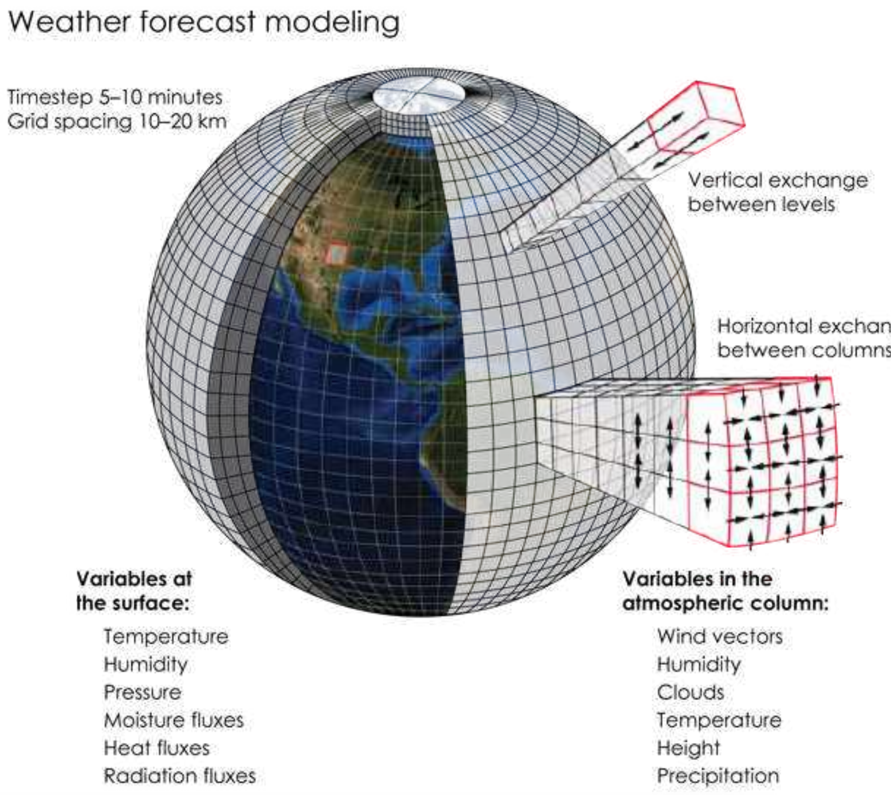
Database (old) vs Data Science (new)

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,…

ACID = Atomicity, Consistency, Isolation and Durability   CAP = Consistency, Availability, Partition Tolerance

Slide by John Canny

Modelling vs Data-Driven Solutions



CIMMS, U of Wisconsin

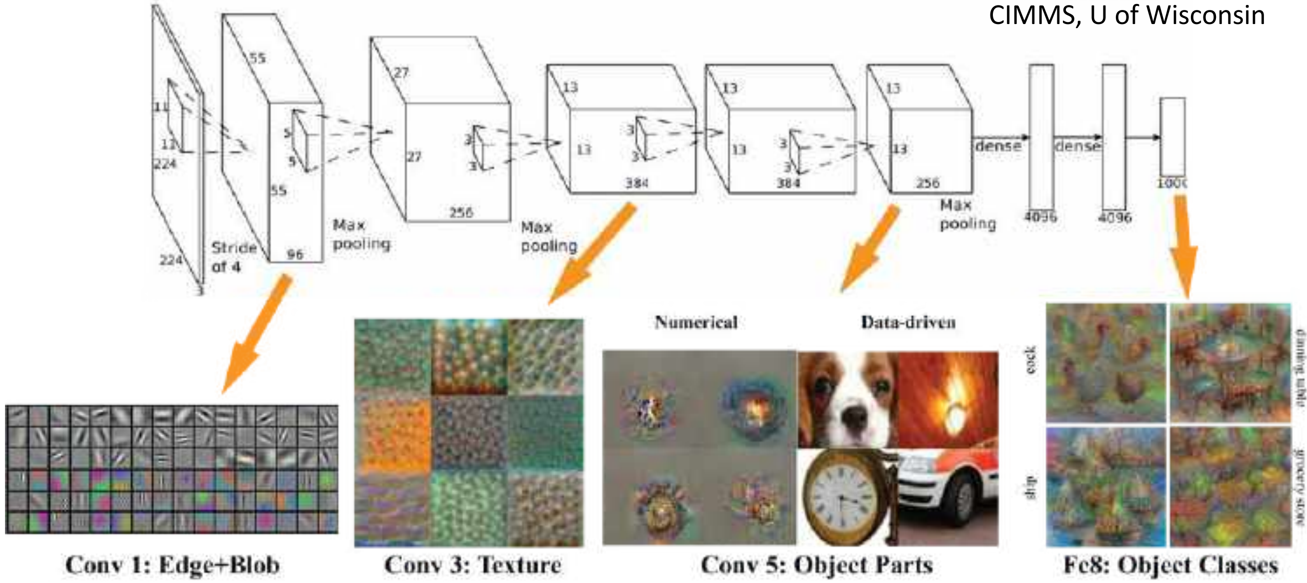
Scientific modelling

Background knowledge, set of rules, principles, representations etc.
Example: Weather forecasting.

Data-Driven Solutions

No or little apriori model, which is replaced by an inference algorithm (e.g., Neural Network, SVM etc.).

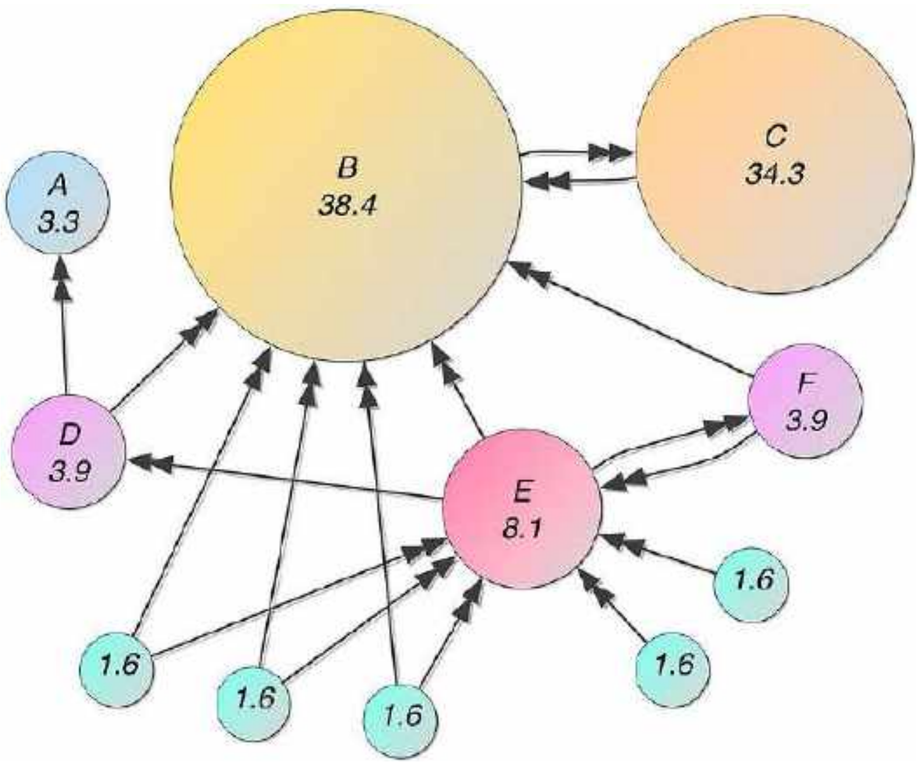
Example: Image classification.

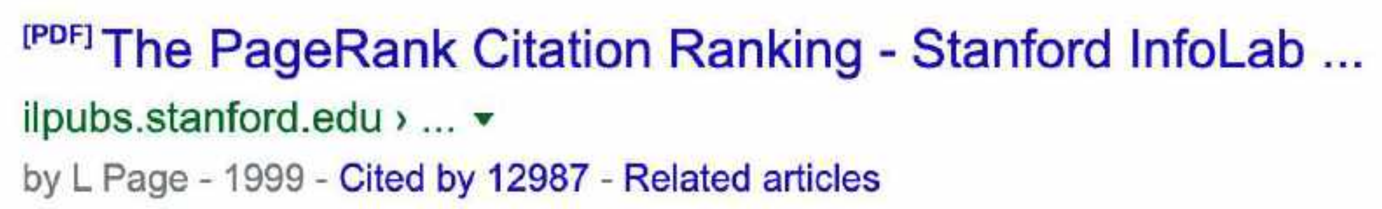


AlexNet/VGG-F visualization from Brown CSCI1430

Some examples - Search

Google PageRank Algorithm

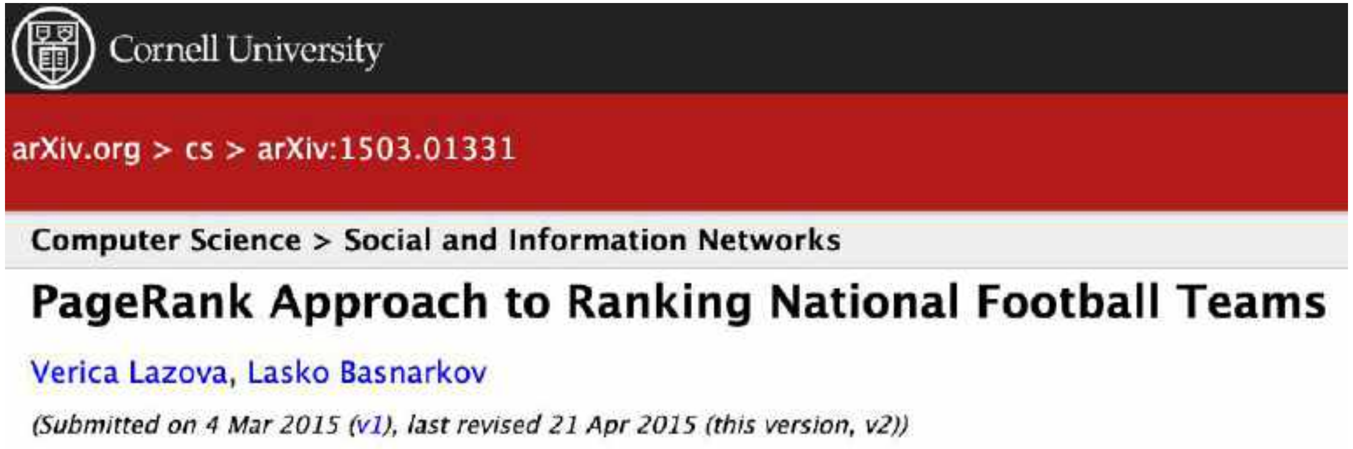



[PDF] The PageRank Citation Ranking - Stanford InfoLab ...

ilpubs.stanford.edu › ... ▾

by L Page - 1999 - Cited by 12987 - Related articles

Original Paper


Cornell University

arXiv.org > cs > arXiv:1503.01331

Computer Science > Social and Information Networks

PageRank Approach to Ranking National Football Teams

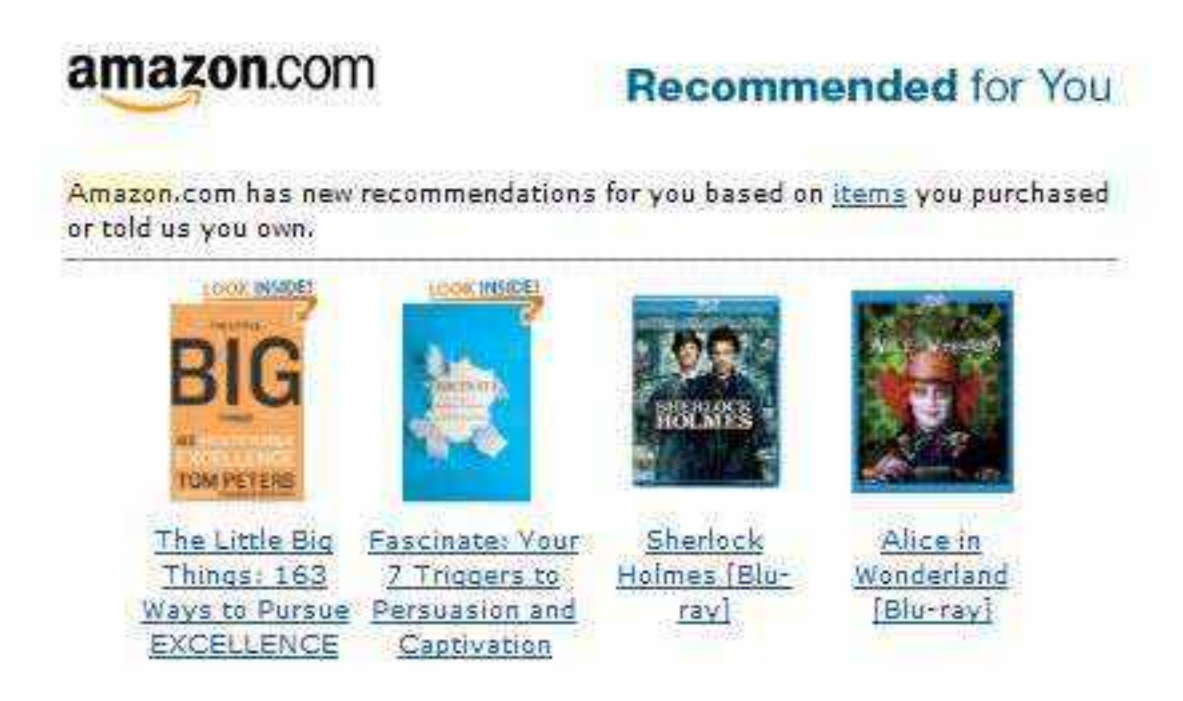
Verica Lazova, Lasko Basnarkov

(Submitted on 4 Mar 2015 (v1), last revised 21 Apr 2015 (this version, v2))

Used in many applications to
have data driven answers to various problems

Image source: Wikipedia

Some examples – Recommendation Systems

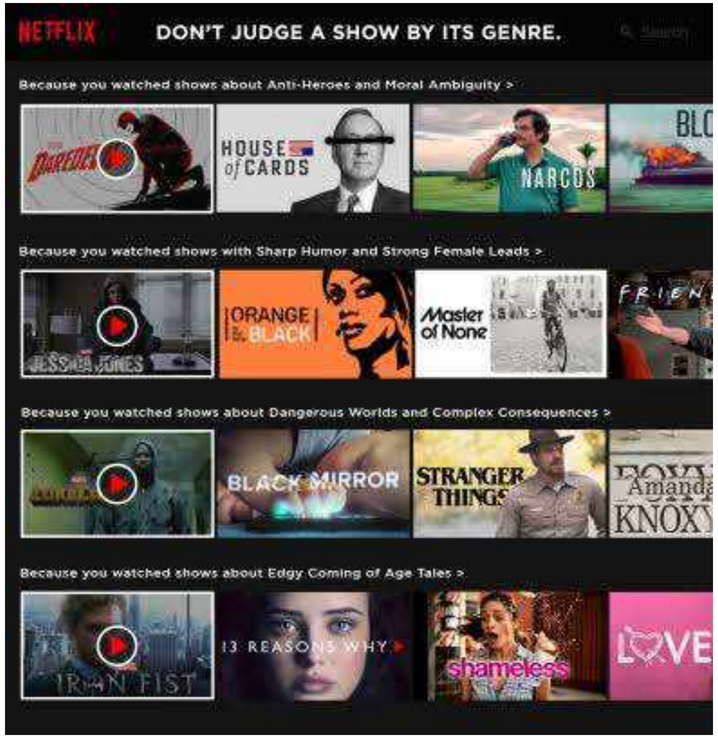


amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

- The Little Big Things: 163 Ways to Pursue EXCELLENCE
- Fascinate: Your 7 Triggers to Persuasion and Captivation
- Sherlock Holmes [Blu-ray]
- Alice in Wonderland [Blu-ray]



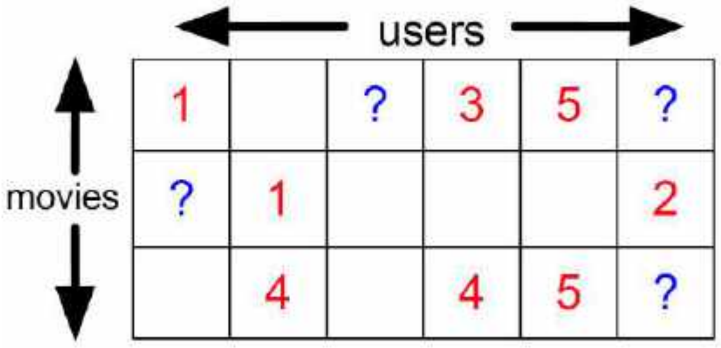
NETFLIX DON'T JUDGE A SHOW BY ITS GENRE.

Because you watched shows about Anti-Heroes and Moral Ambiguity >

Because you watched shows with Sharp Humor and Strong Female Leads >

Because you watched shows about Dangerous Worlds and Complex Consequences >

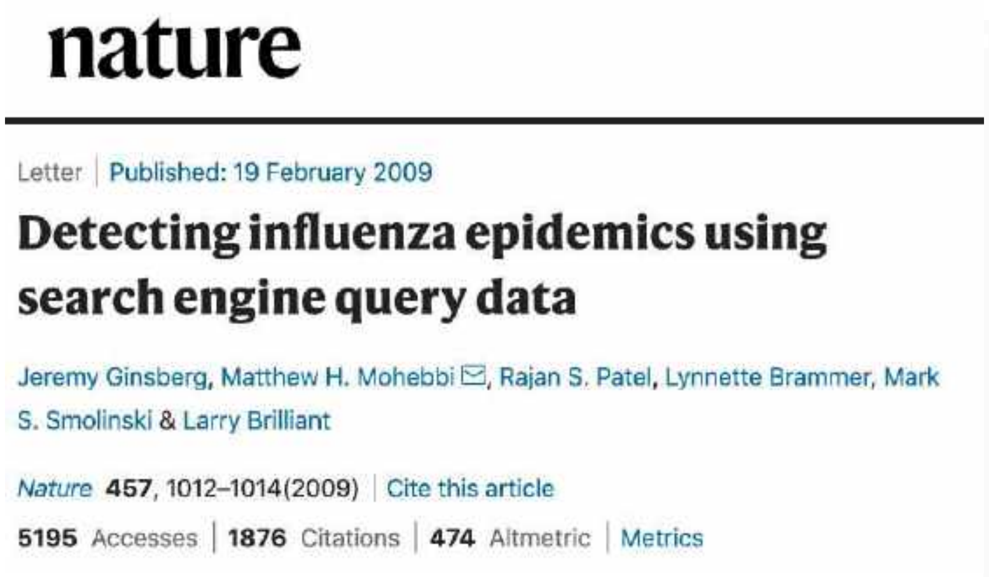
Because you watched shows about Edgy Coming of Age Tales >



1		?	3	5	?
?	1				2
	4		4	5	?

Some examples – Flu Trends

Google Flu Trends



nature

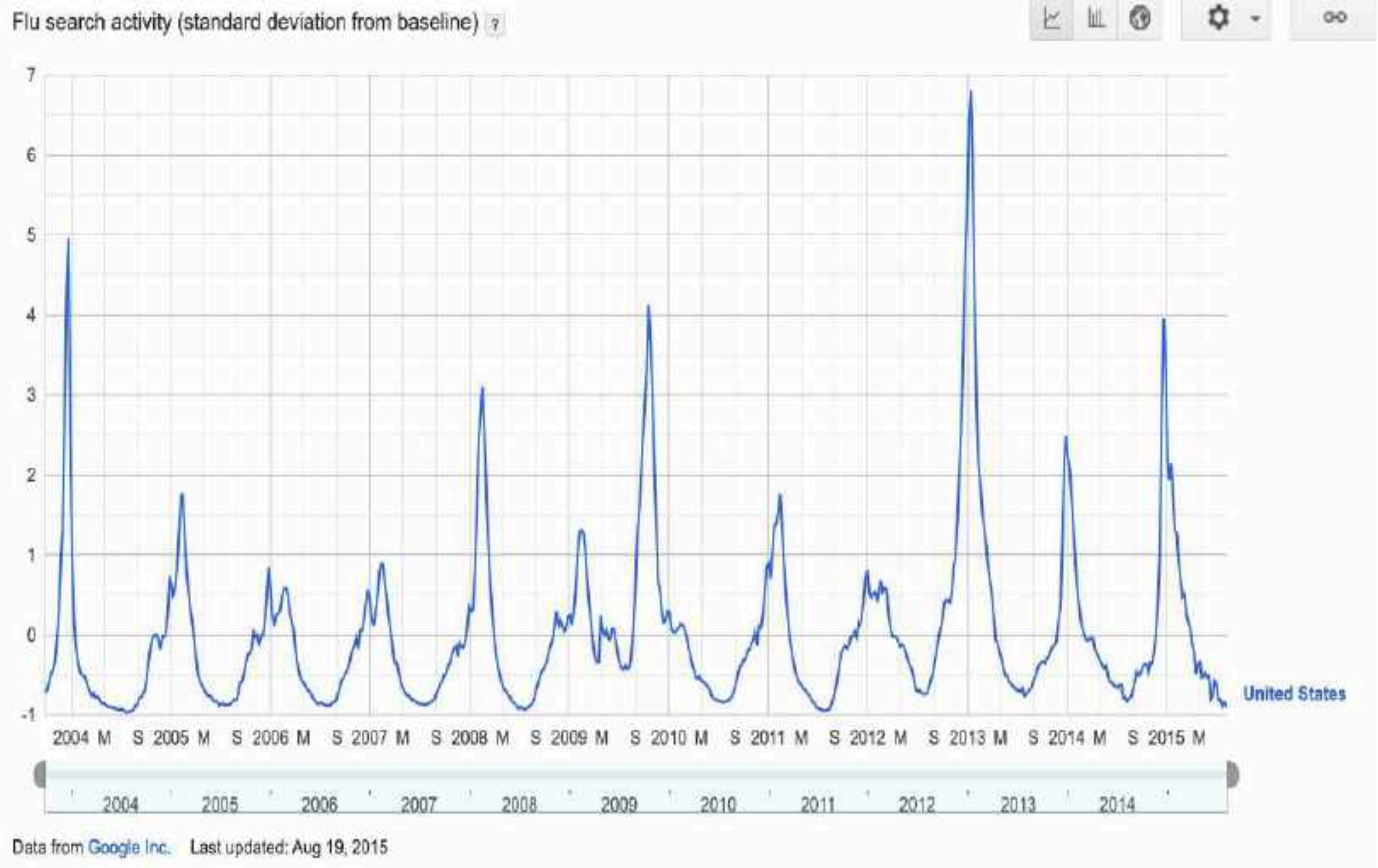
Letter | Published: 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

Nature **457**, 1012–1014(2009) | Cite this article

5195 Accesses | **1876** Citations | **474** Altmetric | Metrics



Flu search activity (standard deviation from baseline)

Data from Google Inc. Last updated: Aug 19, 2015

AlexNet/VGG-F visualization from Brown CSCI1430

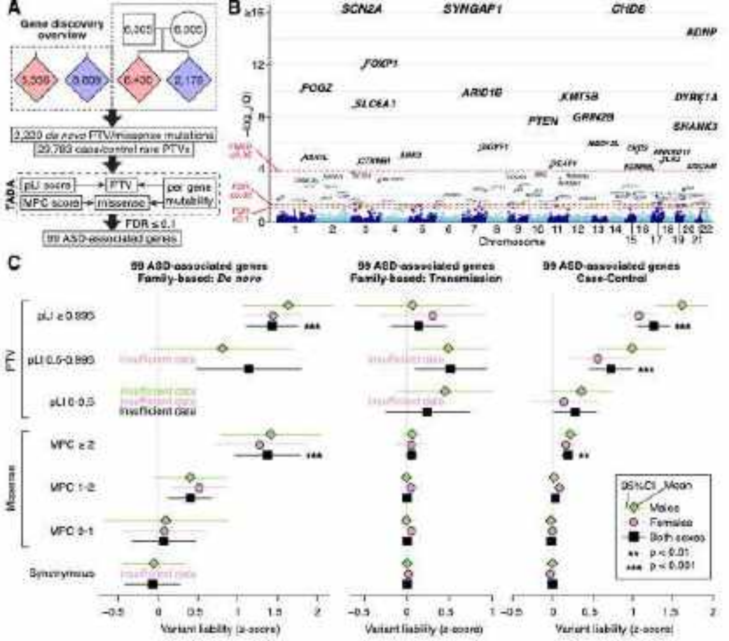
Some examples – Comp. Biology

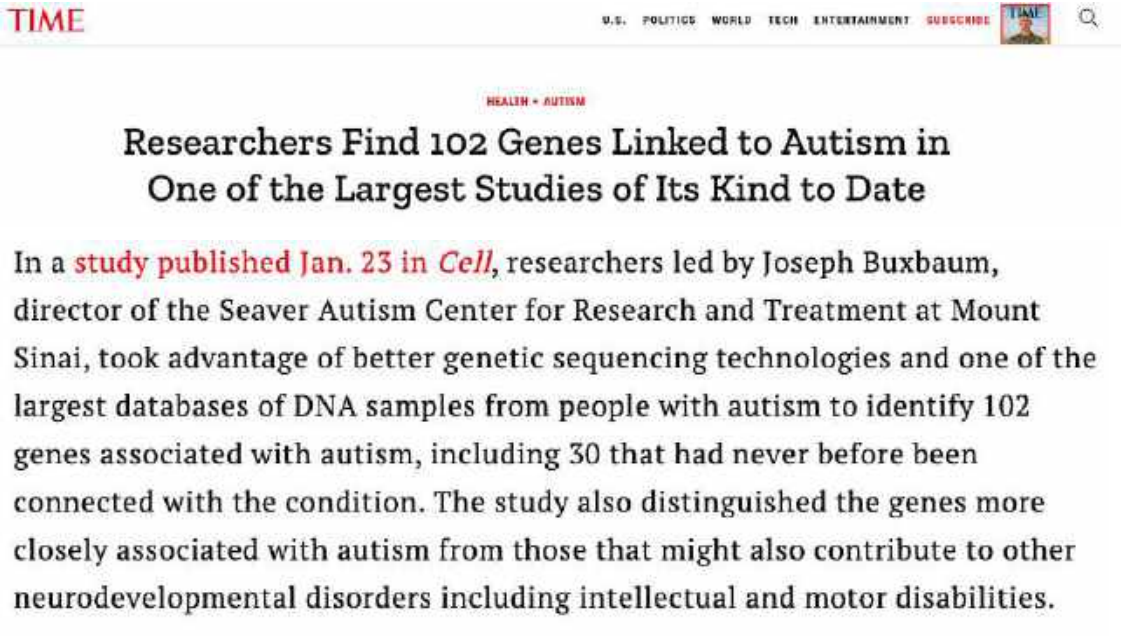
Data Science for Gene Risk Prediction

It is not enough to collect the data.

What does the data tell us?

Use methods to analyze the it.

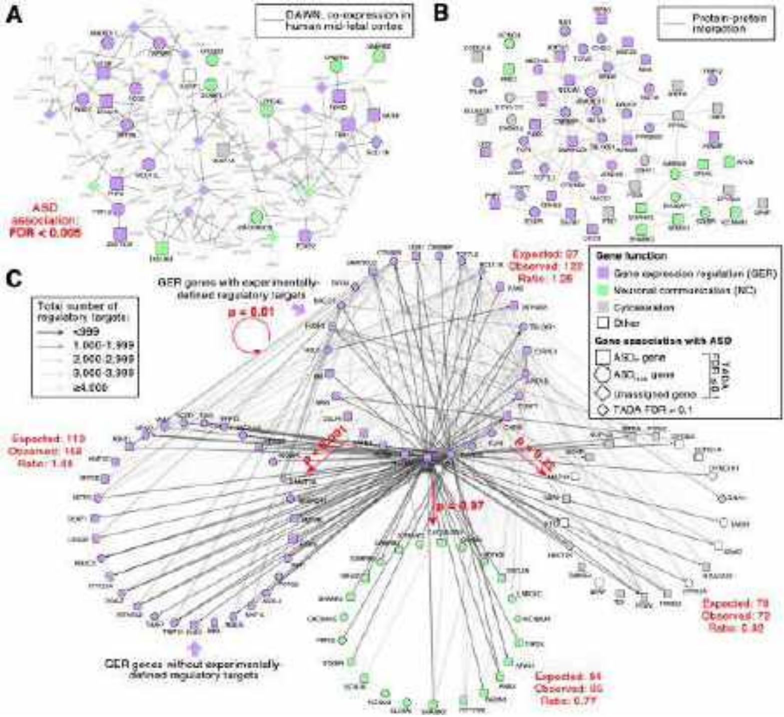




HEALTH • AUTISM

Researchers Find 102 Genes Linked to Autism in One of the Largest Studies of Its Kind to Date

In a study published Jan. 23 in *Cell*, researchers led by Joseph Buxbaum, director of the Seaver Autism Center for Research and Treatment at Mount Sinai, took advantage of better genetic sequencing technologies and one of the largest databases of DNA samples from people with autism to identify 102 genes associated with autism, including 30 that had never before been connected with the condition. The study also distinguished the genes more closely associated with autism from those that might also contribute to other neurodevelopmental disorders including intellectual and motor disabilities.



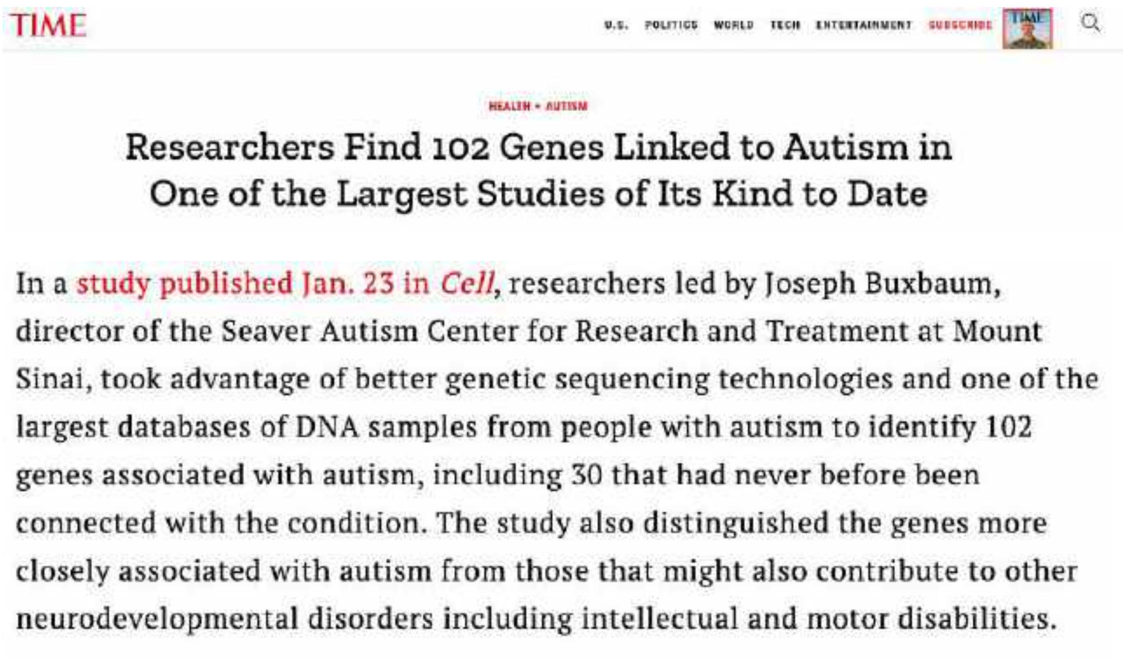
A
B
C
ASD association FDR < 0.005
Protein-protein interaction
Gene function
Gene association with ASD
p < 0.01

Satterstrom *et al.*, CELL 2020

Some examples – Comp. Biology

Machine Learning for Gene Risk Prediction

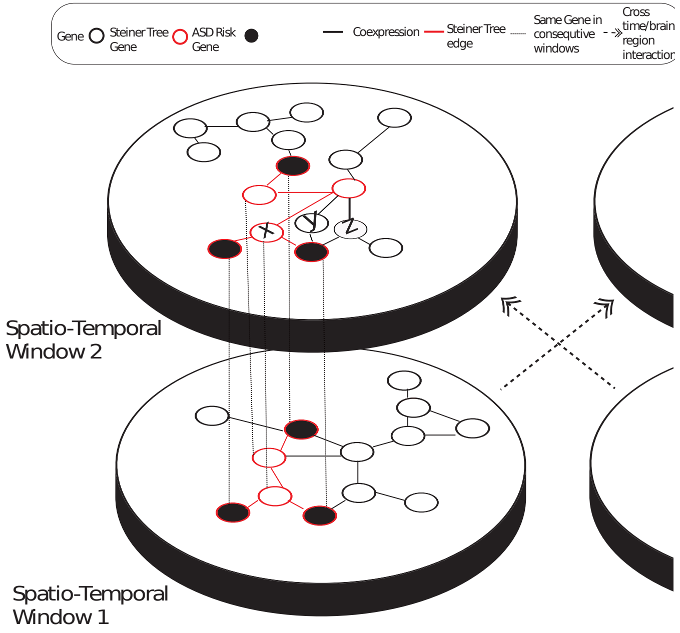
Build algorithms to predict the risk



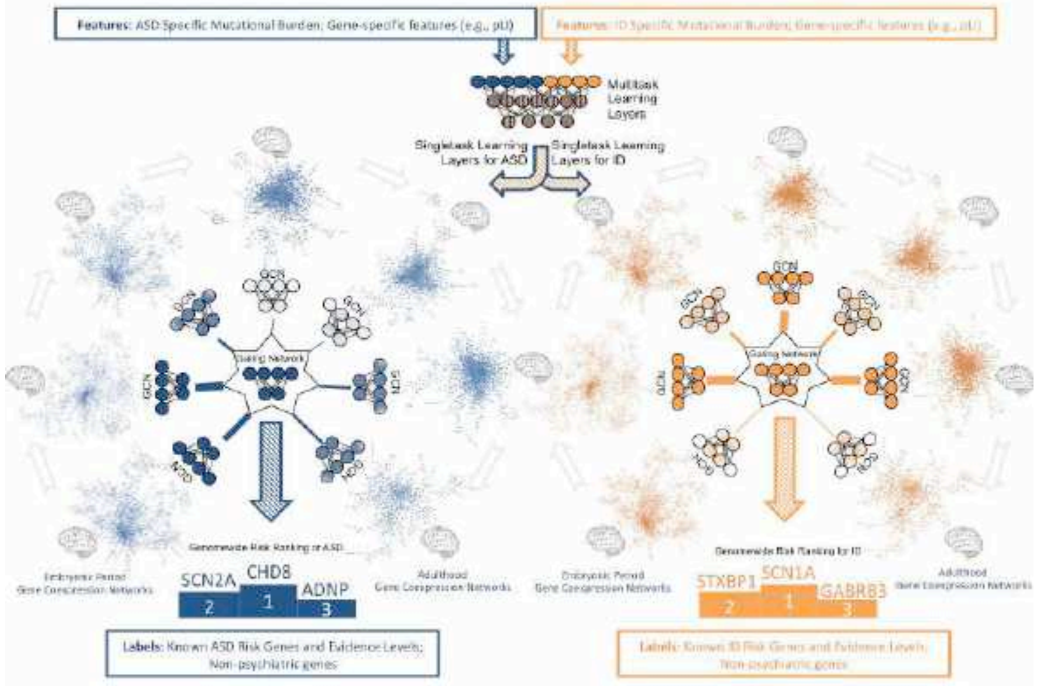
Researchers Find 102 Genes Linked to Autism in One of the Largest Studies of Its Kind to Date

In a study published Jan. 23 in *Cell*, researchers led by Joseph Buxbaum, director of the Seaver Autism Center for Research and Treatment at Mount Sinai, took advantage of better genetic sequencing technologies and one of the largest databases of DNA samples from people with autism to identify 102 genes associated with autism, including 30 that had never before been connected with the condition. The study also distinguished the genes more closely associated with autism from those that might also contribute to other neurodevelopmental disorders including intellectual and motor disabilities.

Satterstrom *et al.*, CELL 2020



Spatio-temporal Network-based Analysis. Norman and Cicek, Bioinformatics 2019.

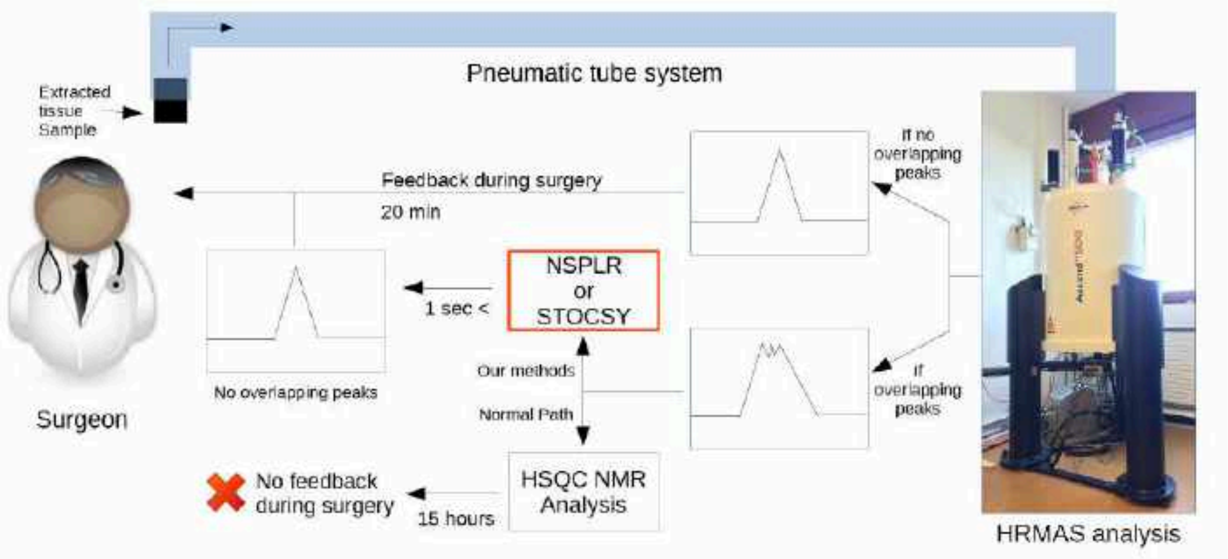

Features: ASD Specific Mutational Burden; Gene-specific features (e.g., pLI)
Features: ID Specific Mutational Burden; Gene-specific features (e.g., pLI)
Multitask Learning Layers
Singletask Learning Layers for ASD
Singletask Learning Layers for ID
Gating Network
Gating Network
GCN
GCN
Genomewide Risk Ranking of ASD
Genomewide Risk Ranking for ID
Embryonic Period Gene Coexpression Networks
Adulthood Gene Coexpression Networks
Embryonic Period Gene Coexpression Networks
Adulthood Gene Coexpression Networks
SCN2A
CHD8
ADNP
2
1
3
STXBP1
SCN1A
GABRB3
2
1
3
Labels: Known ASD Risk Genes and Evidence Levels; Non-psychiatric genes
Labels: Known ID Risk Genes and Evidence Levels; Non-psychiatric genes

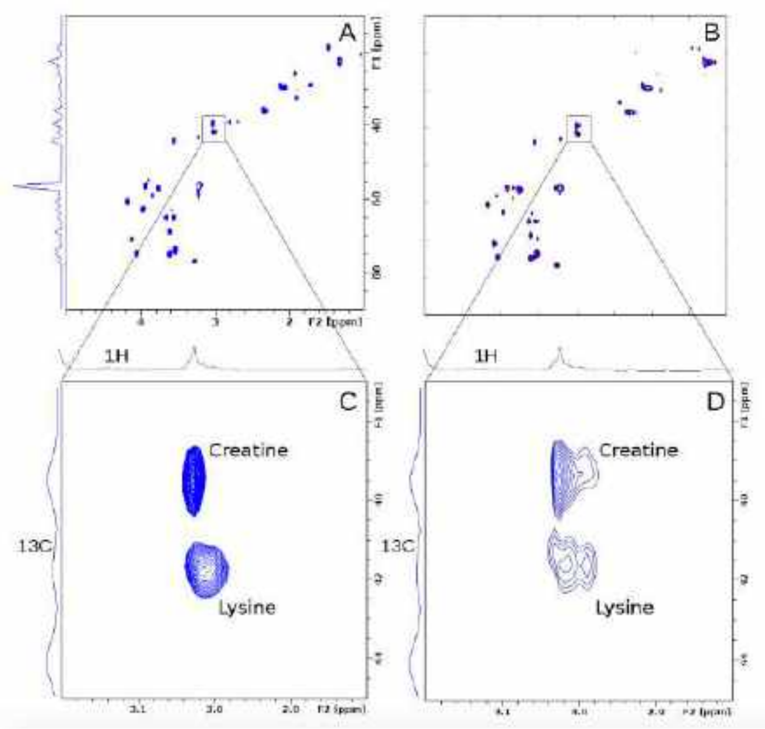
Multi-Task Learning for Autism Gene Risk Prediction. Karakahya *et al.*, in prep.

Some examples – Comp. Biology

Data Science for Online Feedback to Surgeons

Use Multiple Multivariate Regression to predict the result of a test that is infeasible to perform during surgery due to time requirement.



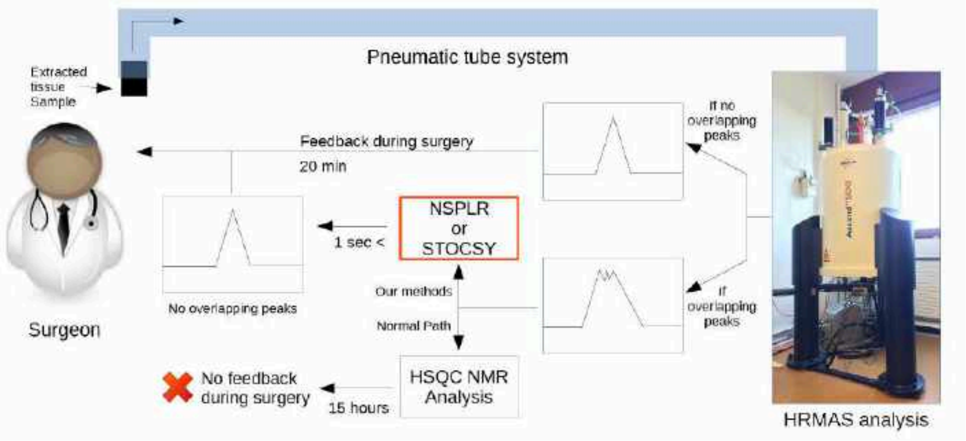


Karakaslar *et al.*, IEEE/ACM TCBB 2019, in press.

Some examples – Comp. Biology

Machine Learning for Online Feedback to Surgeons

Design a neural network that learns important parts of to classify tumors.



Extracted tissue Sample

Pneumatic tube system

Surgeon

Feedback during surgery

20 min

If no overlapping peaks

No overlapping peaks

1 sec <

NSPLR or STOCSY

Our methods

Normal Path

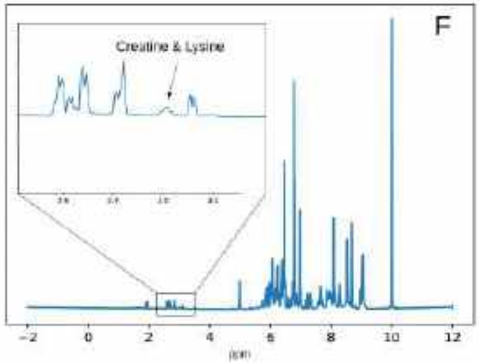
If overlapping peaks

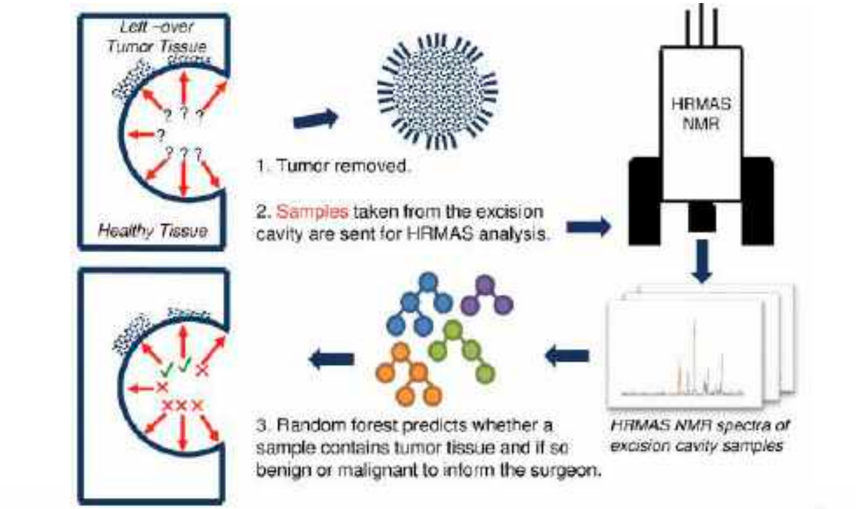
No feedback during surgery

15 hours

HSQC NMR Analysis

HRMAS analysis





Left –over Tumor Tissue

Healthy Tissue

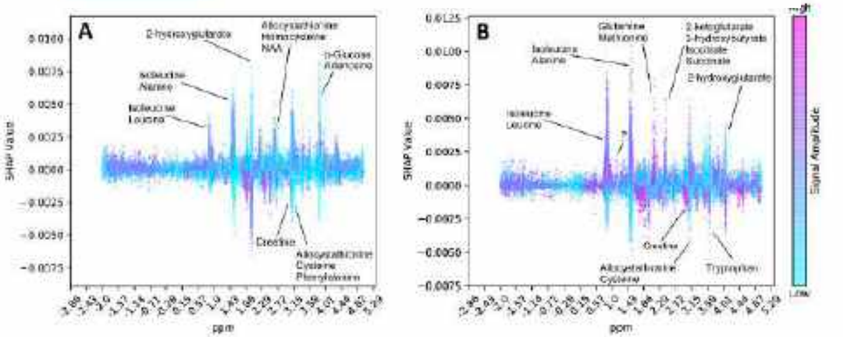
1. Tumor removed.

2. Samples taken from the excision cavity are sent for HRMAS analysis.

HRMAS NMR

HRMAS NMR spectra of excision cavity samples

3. Random forest predicts whether a sample contains tumor tissue and if so benign or malignant to inform the surgeon.



Cakmakci *et al.,* PLoS Comp. Bio 2020

Data Science Pipeline

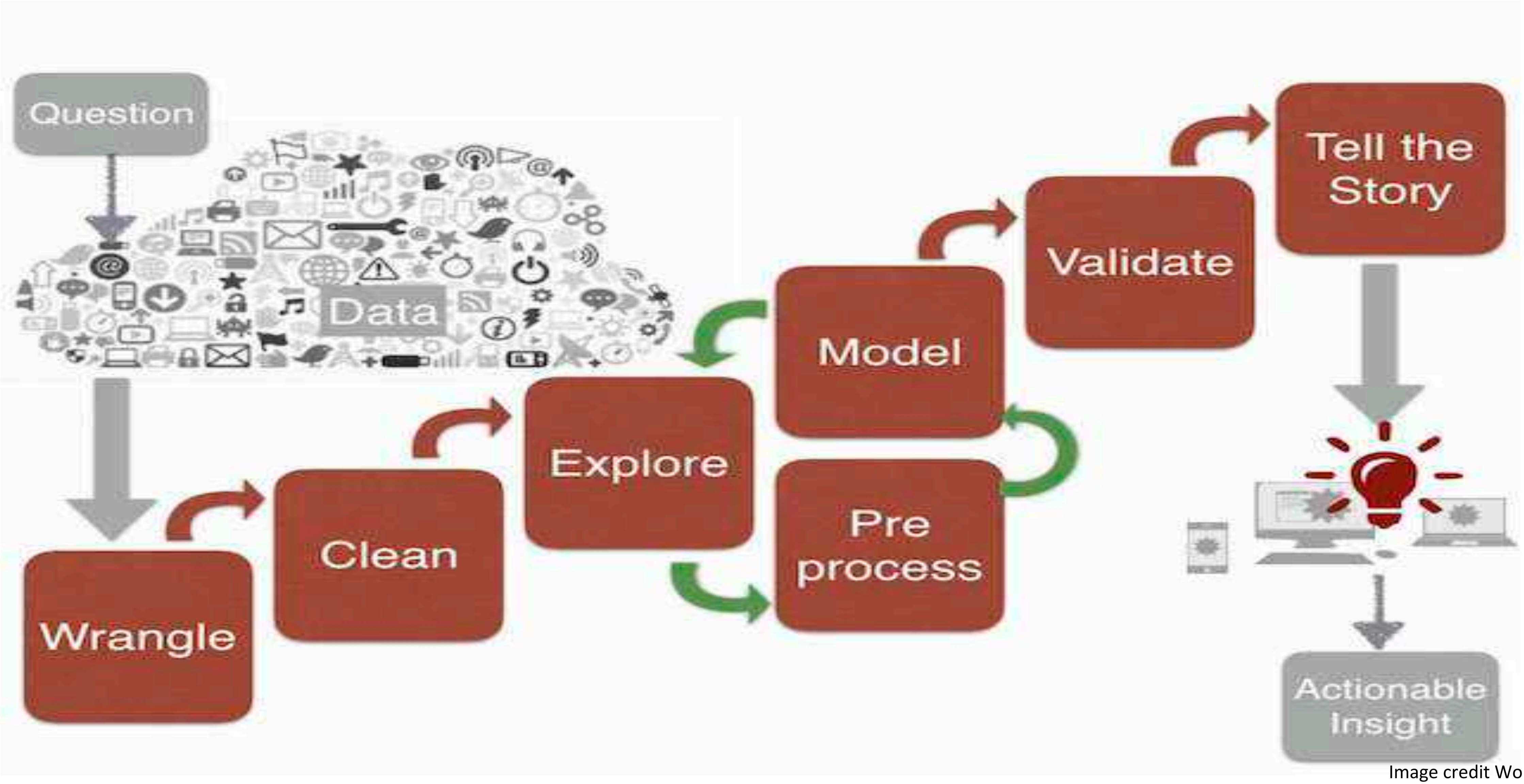


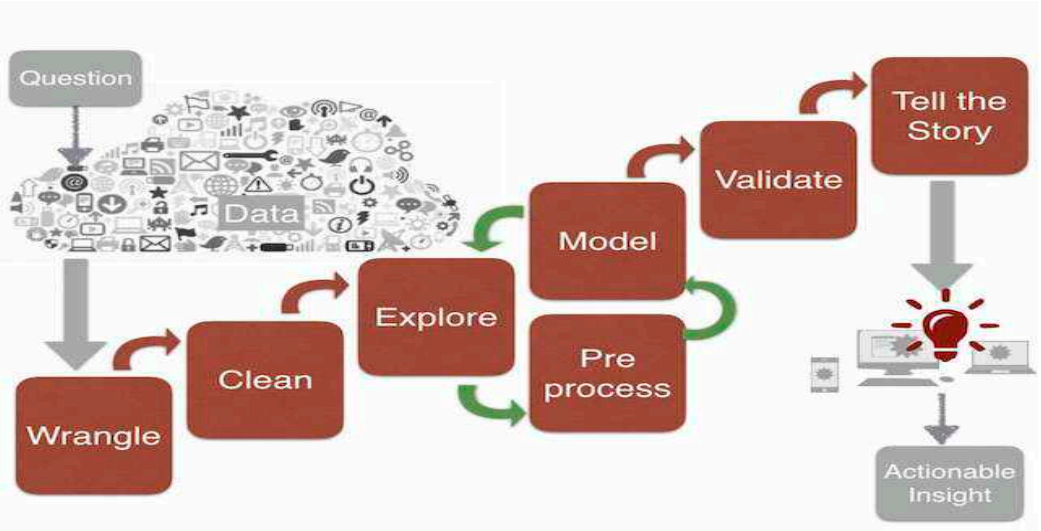
Image credit Wolfram Research

Data Science Pipeline - Data Collection

Many data types, many ways

- Sensors
- Crowdsourcing, putting humans at work once computers fail:
  - Mechanical Turk
- Crawling
- Questionnaires..

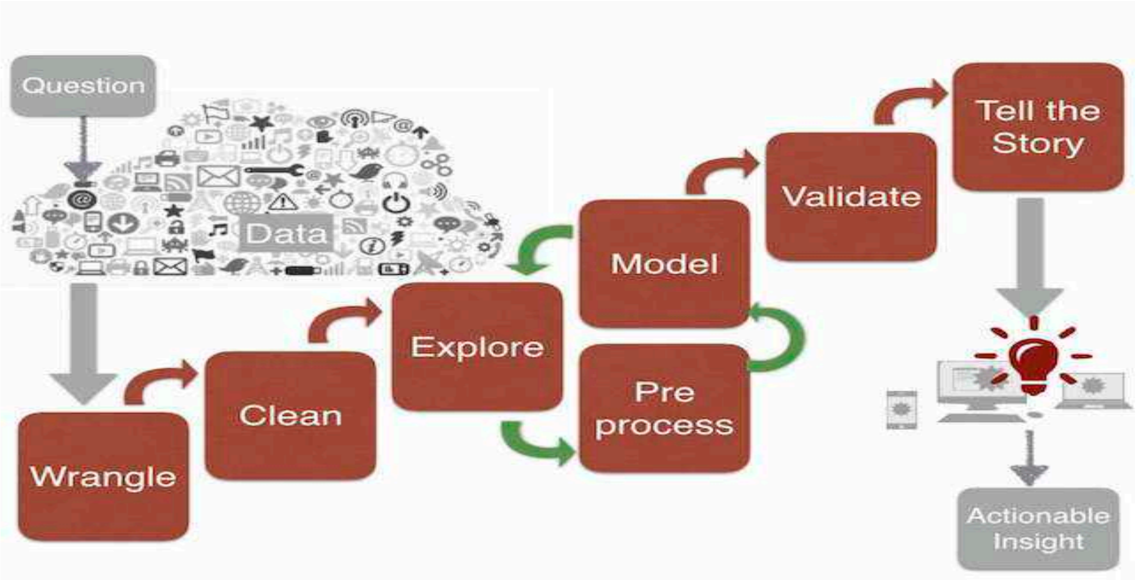
The Turk



Data Science Pipeline - Data Wrangling

After you obtain the raw data converting it into a more useful format

- Gather multiple files into single, standardized format
- For example: Unite multiple crawled files into one, get rid of html tags etc.



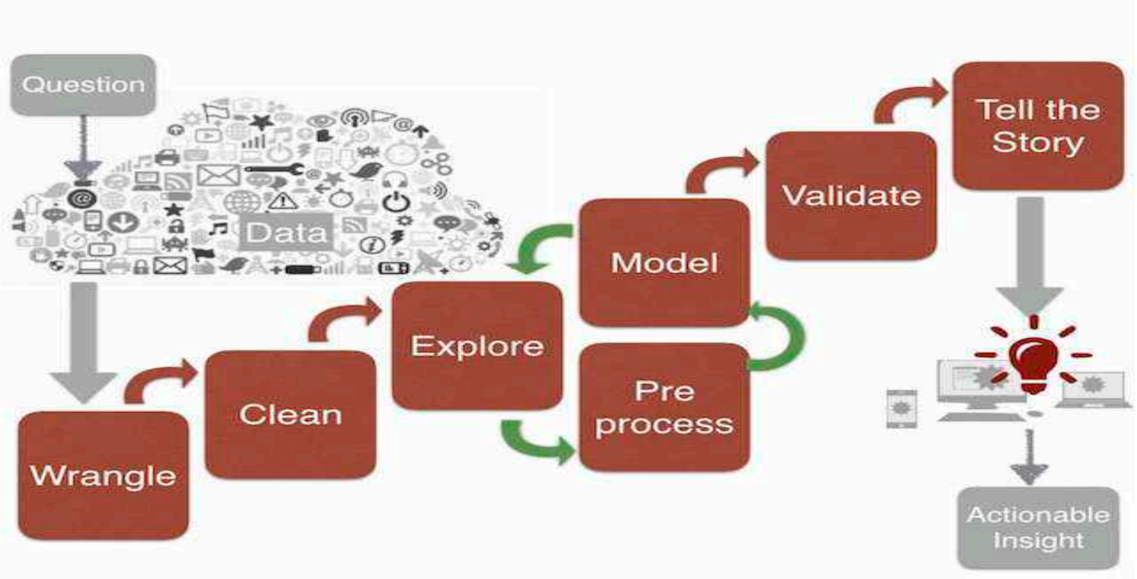
Data Science Pipeline - Data Cleaning

Dig deeper into the data after standardization and detect problems.

Inconsistencies

Outliers

Missing values



Data Science Pipeline
Explore – Preprocess – Model Cycle

1. Explore the structure of the data and decide on the appropriate model to analyze.

For instance: sequence data, maybe LSTM?

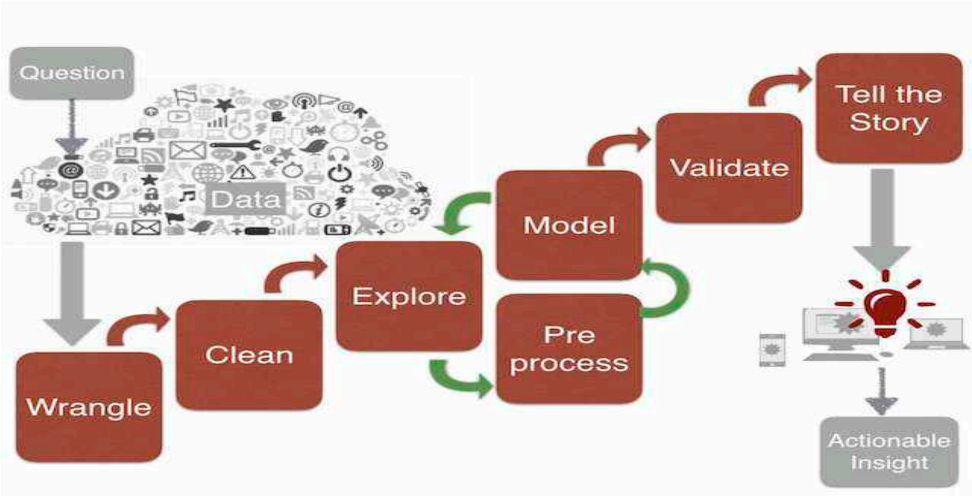
image data, maybe Convolutional Neural Networks.

2. Preprocess the data to be fit into the model

For instance, RGB -> Grayscale

3. Apply the model and analyze results

4. Go to 1.



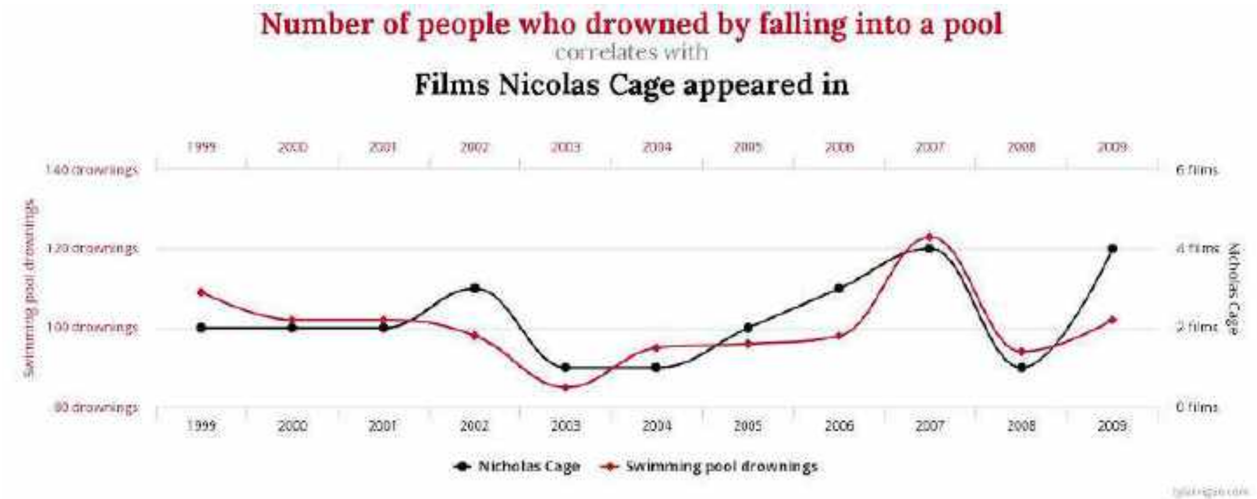
Data Science Pipeline - Validation

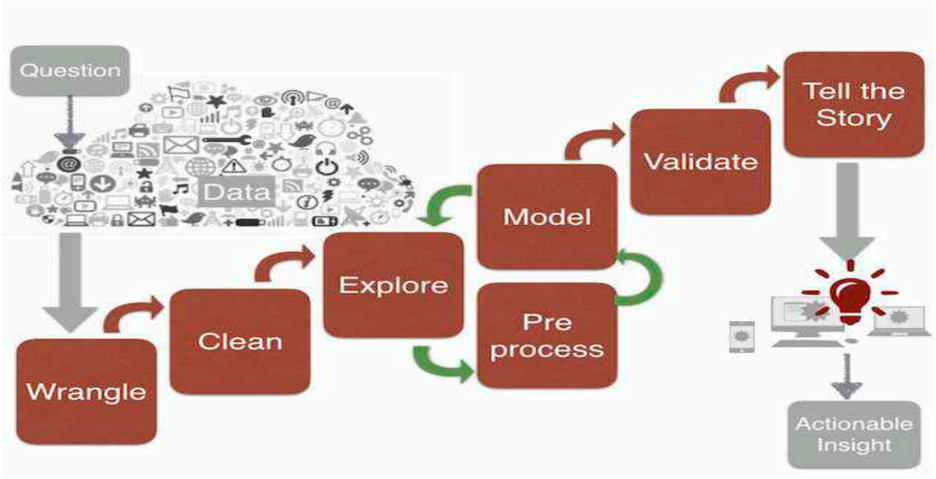
After you fine-tuned your model in the previous cycle validate your data on a data that has not been seen by the model.

Validate that your claim is not just random finding.

Multiple hypothesis correction

Correlation is not causation.

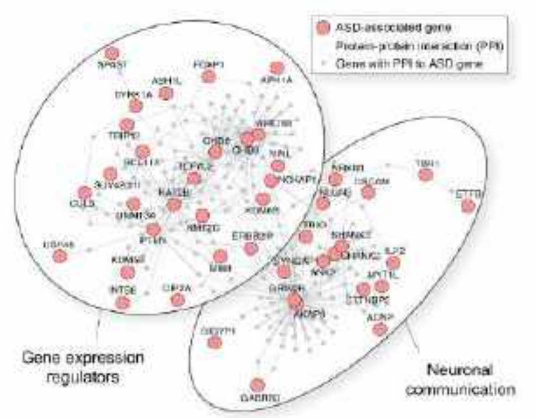


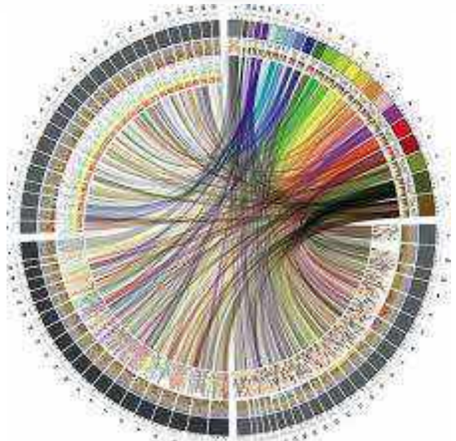


Data Science Pipeline – Story Telling

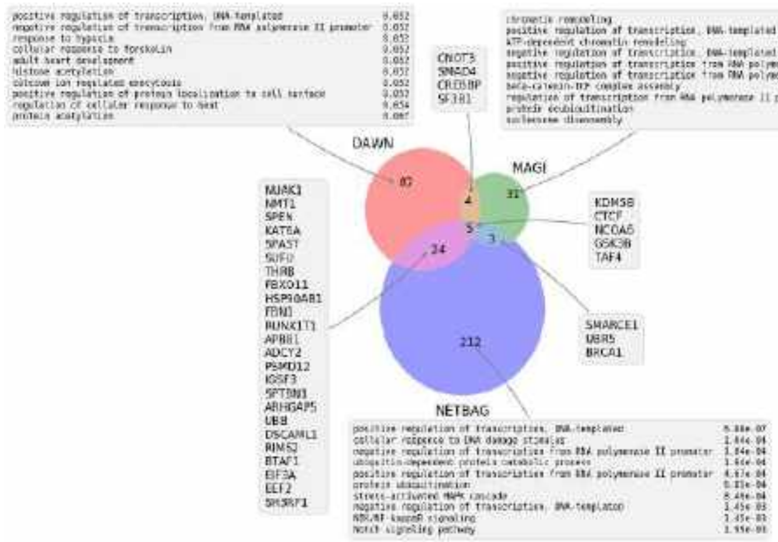
A data scientist also needs to communicate well.

Infographics and how you convey the story is important.





vs



positive regulation of transcription, DNA-templated	0.052
negative regulation of transcription from RNA polymerase II promoter	0.052
response to hypoxia	0.052
cellular response to forskolin	0.052
adult heart development	0.052
histone acetylation	0.052
calcium ion regulated exocytosis	0.052
positive regulation of protein localization to cell surface	0.052
regulation of cellular response to heat	0.054
protein acetylation	0.067

CNOT3
SMAD4
CREBBP
SF3B1

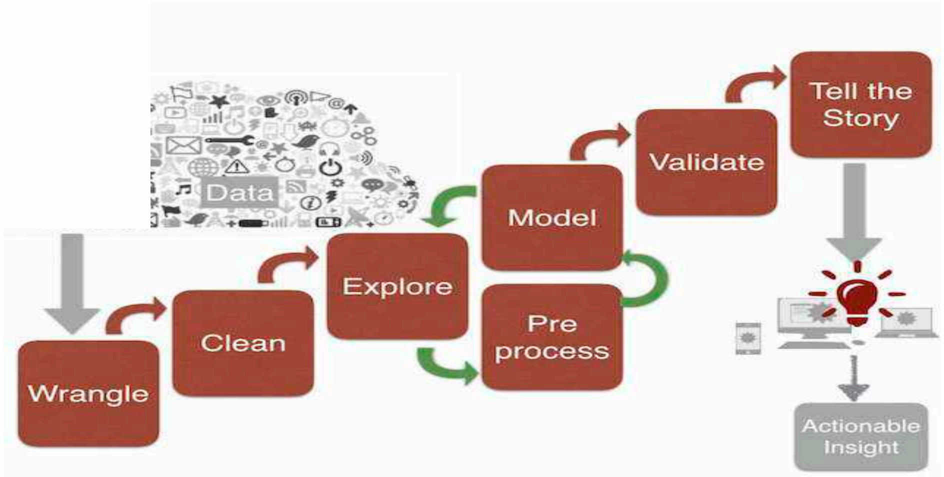
chromatin remodeling	2.30e-11
positive regulation of transcription, DNA-templated	2.30e-11
ATP-dependent chromatin remodeling	2.30e-11
negative regulation of transcription, DNA-templated	6.40e-11
positive regulation of transcription from RNA polymerase II promoter	9.00e-09
negative regulation of transcription from RNA polymerase II promoter	1.14e-08
beta-catenin-TCF complex assembly	1.59e-07
regulation of transcription from RNA polymerase II promoter	5.10e-07
protein deubiquitination	1.25e-06
nucleosome disassembly	1.81e-06

MUAK1
NMT1
SPEN
KAT6A
SPAST
SUFU
THRB
FBXO11
HSP90AB1
FBN1
RUNX1T1
APBB1
ADCY2
PSMD12
IGSF3
SPTBN1
ARHGAP5
UBB
DSCAML1
RIMS2
BTAF1
EIF3A
EEF2
SRSRF1

KDM5B
CTCF
NCOA6
GSK3B
TAF4

SMARCE1
UBR5
BRCA1

positive regulation of transcription, DNA-templated	5.88e-07
cellular response to DNA damage stimulus	1.84e-04
negative regulation of transcription from RNA polymerase II promoter	1.84e-04
ubiquitin-dependent protein catabolic process	1.66e-04
positive regulation of transcription from RNA polymerase II promoter	4.67e-04
protein ubiquitination	5.81e-04
stress-activated MAPK cascade	8.49e-04
negative regulation of transcription, DNA-templated	1.45e-03
NIK/NF-kappaB signaling	1.45e-03
Notch signaling pathway	1.95e-03



Data Storage and Cloud

Database Systems

- Relational databases, organized around tables, SQL
- NoSQL databases for online distributed databases, eventual consistency: Cassandra, HBase

Cloud Storage

- Ubiquitous computing, data access from everywhere
- No worries on losing data

Cloud Computing

- Distributed computing on large scale data
- Map Reduce, Hadoop

Statistical Modeling

Parametric Models

Family of probability distributions with a finite number of parameters

For example: Binomial distribution has 2 (n,p)

Non-parametric Models

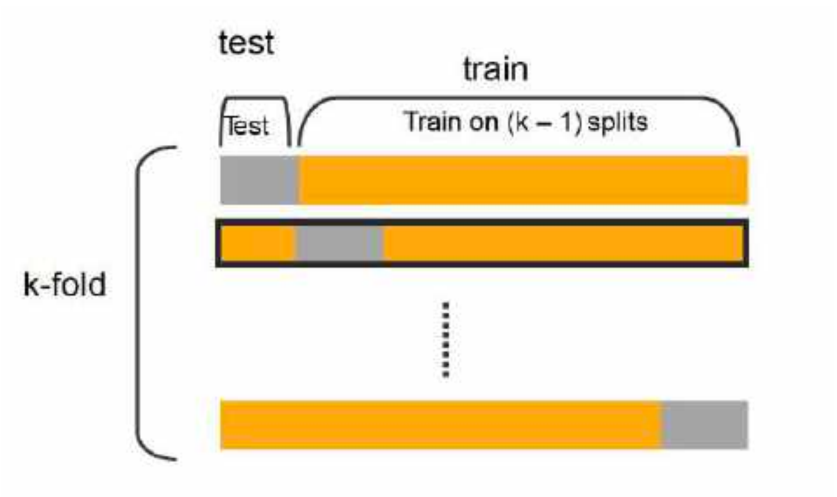
Parameter set is infinite dimensional i.e., grows with the data size. For example: k nearest neighbors classification.

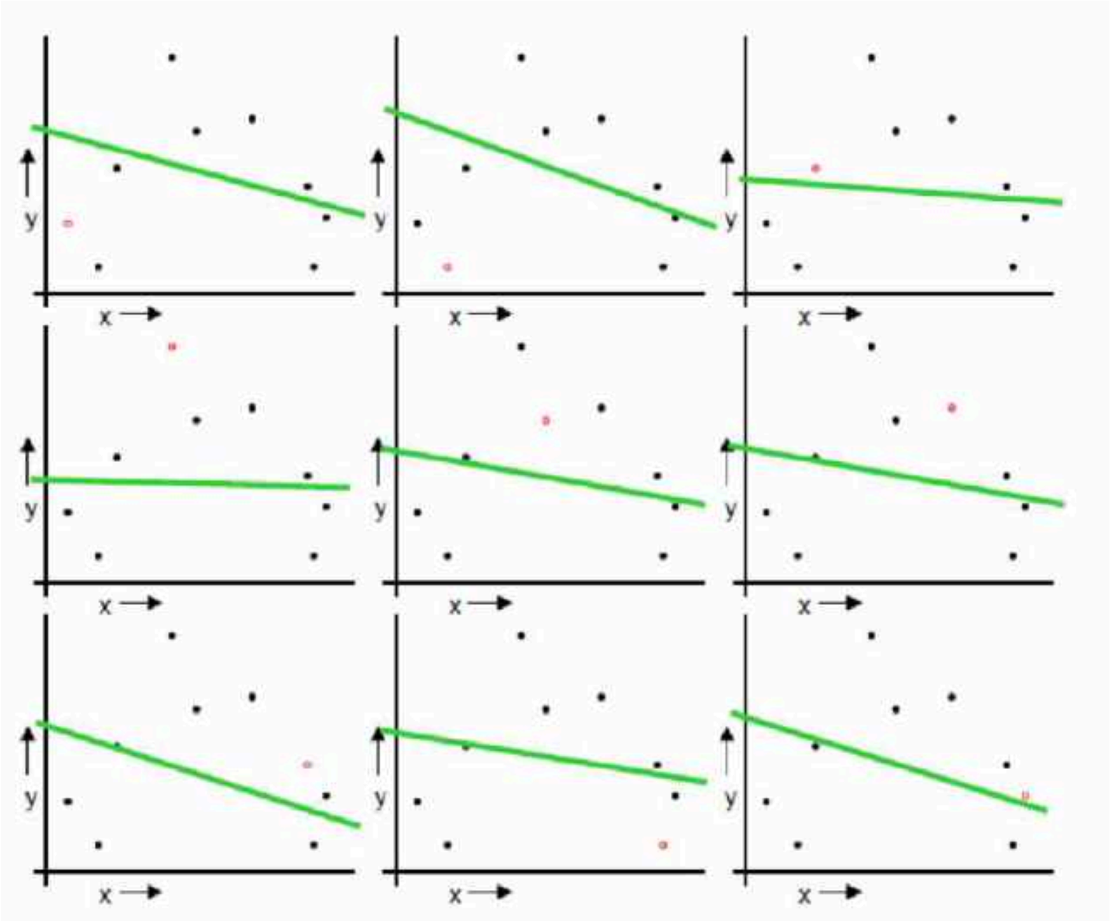
Model Validation

Experimental Design

Cross Validation

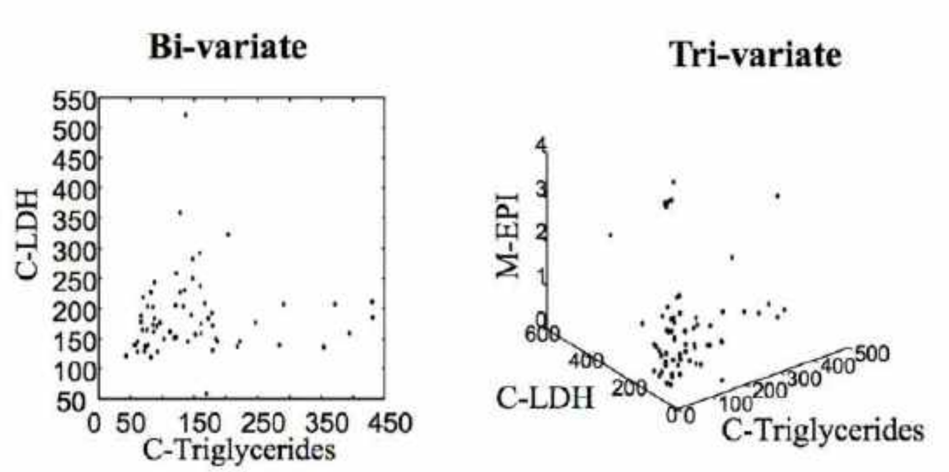
Statistical Tests for validation





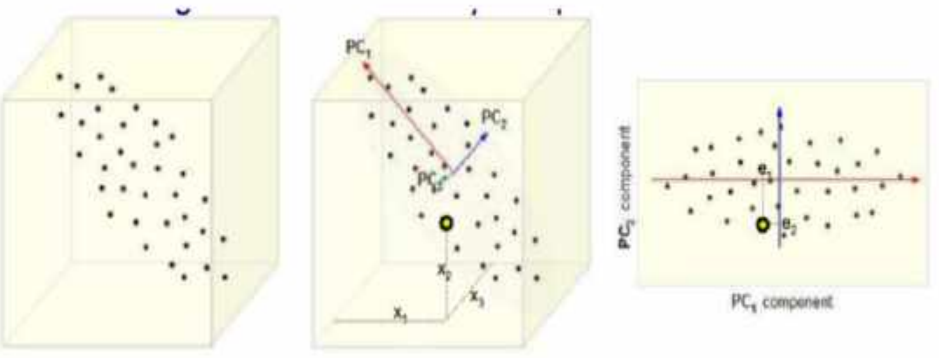
Unsupervised Learning

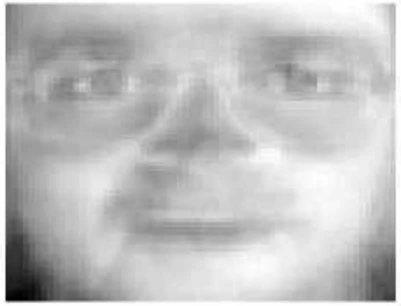
Feature extraction: Principal Component Analysis, t-SNE etc.



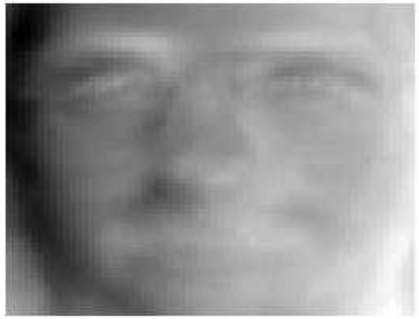
How can we visualize the other variables???

... difficult to see in 4 or higher dimensional spaces...





PC1



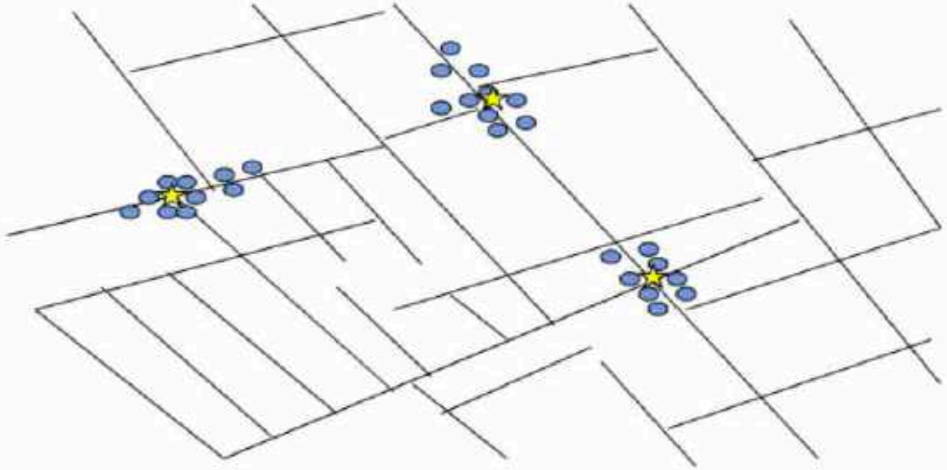
PC2

Unsupervised Learning – cont’d

Clustering: Finding groups of data points which are similar to each other.

John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.

The locations indicated that cases were clustered around certain intersections where there were polluted wells – thus exposing both the problem and the solution



From: Nina Mishra HP Labs

Unsupervised Learning – cont'd

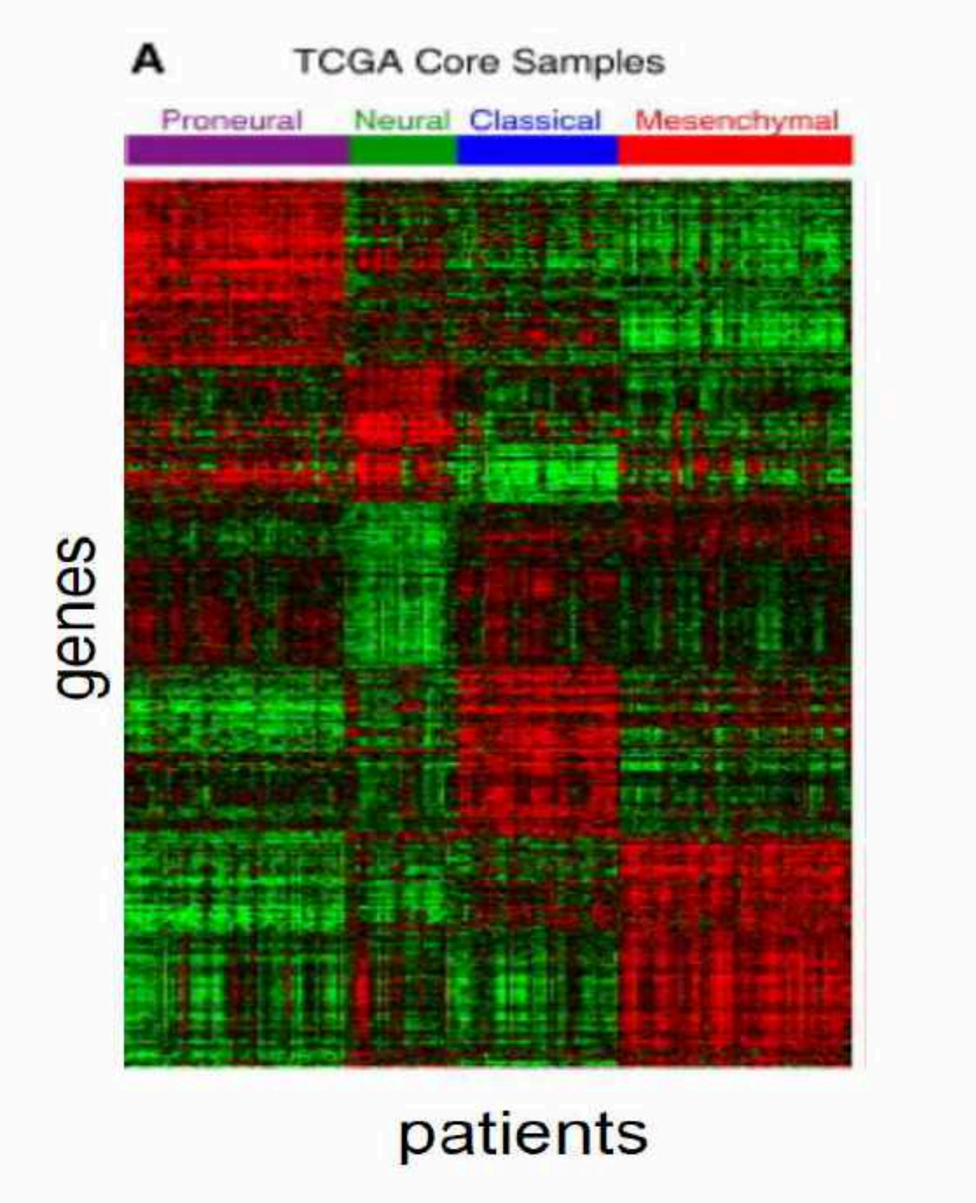
Clustering: Finding groups of data points which are similar to each other.

Given a sample of breast cancer patients and their gene activity level measurements. Can you find subgroups? (e.g., aggressive, mild etc.)

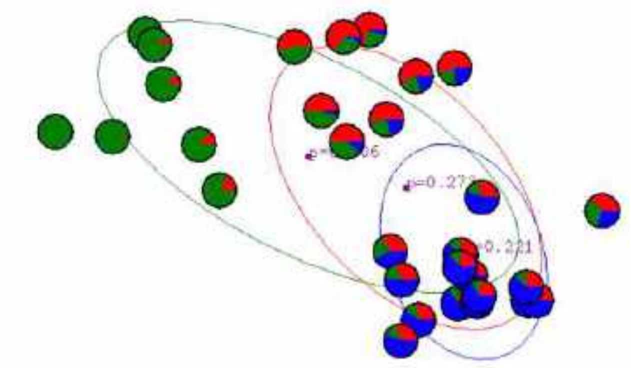
So many other applications:

Targeted advertising

LinkedIn contact suggestion



Unsupervised Learning – cont’d



GMM example

Winner take all rule, competitive learning

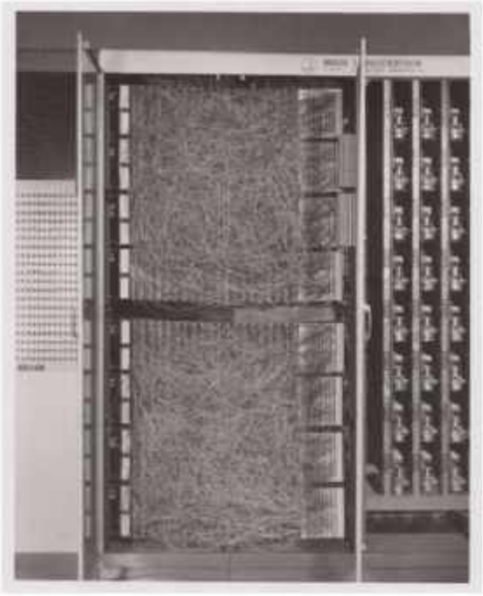
Several algorithm examples

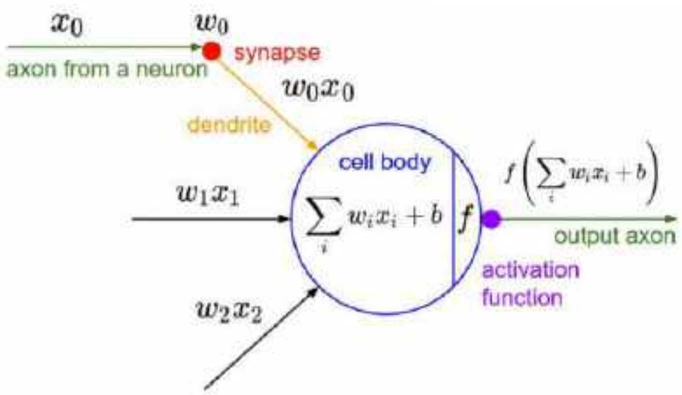
- k-means
  - k cluster centers as means of assigned data points
- Gaussian Mixture Models
  - assumes k Gaussian processes generate data
- Spectral Clustering
  - Generate eigenvalues/eigenvectors of the Laplacian of the similarity matrix
  - Use smallest eigenvalue and corresponding eigenvectors
  - for dimension reduction

Supervised Learning

When the data has labels learn a predictive model using features.

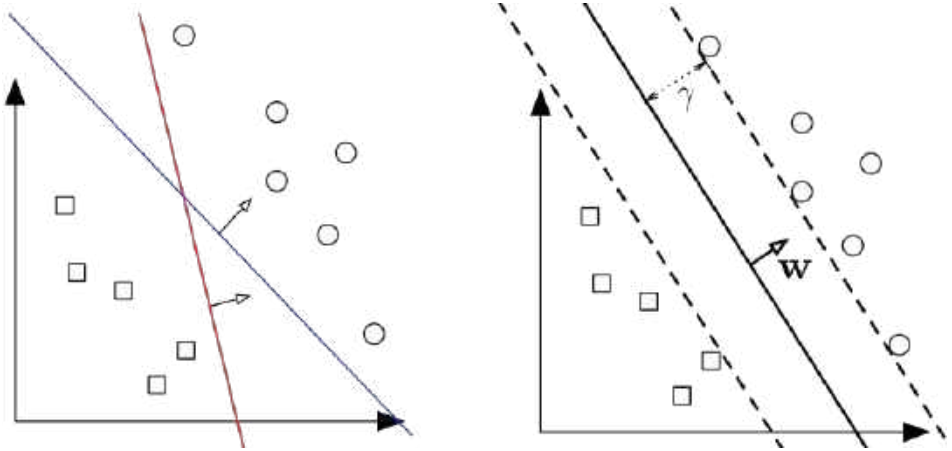
- Neural Network Architectures
  - Perceptron
  - Multi Layer Perceptron
  - Convolutional Networks
  - Recurrent Neural Networks
- Neural Network Training
  - Backpropagation
  - Optimizers
- Support Vector Machines
- Decision Trees
- Ensemble Learning
  - Random Forest
  - XGBoost, AdaBoost





'Mark I Perceptron at the Cornell Aeronautical Laboratory', hardware implementation of the first Perceptron (Source: Wikipedia / Cornell Library)

Neural Networks



SVM example – image source Cornell cs4780

Reinforcement Learning

Learning a policy by experience, reward, penalty like humans.

Q-Learning

Deep Q-Network

AlphaGo beats a 9-dan (professional) 4-1, gets 9-dan
Later AlphaZero is developed for GO, Shogi and Chess

AlphaZero beats a top professional player. First, time in a RTS game.
Again, by DeepMind.