Intro to Data Science

January 2020

Avenga

OLENA DOMANSKA

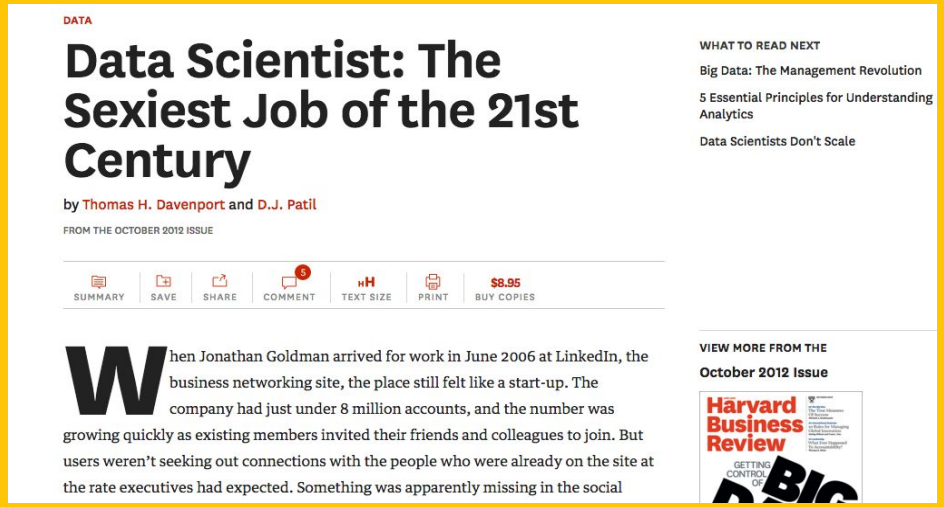
Head of Data Science,
Avenga

PhD in Mathematics
Assistant Professor,
LNU

What is the sexiest job in 21 century?

Data Scientist

Harvard Business Review (2012)



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT 5 | TEXT SIZE | PRINT | $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social

WHAT TO READ NEXT

Big Data: The Management Revolution

5 Essential Principles for Understanding Analytics

Data Scientists Don't Scale

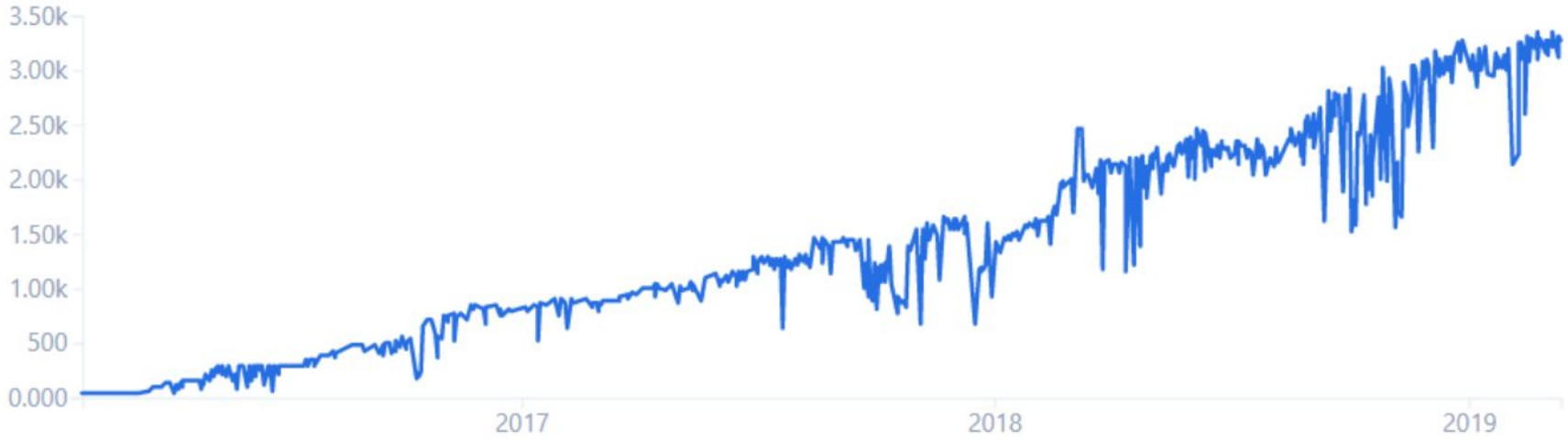
VIEW MORE FROM THE

October 2012 Issue

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

Avenga

Data Scientist job openings at the world's top companies



*Data from Thinknum - **Open dataset***

● Title (Count)

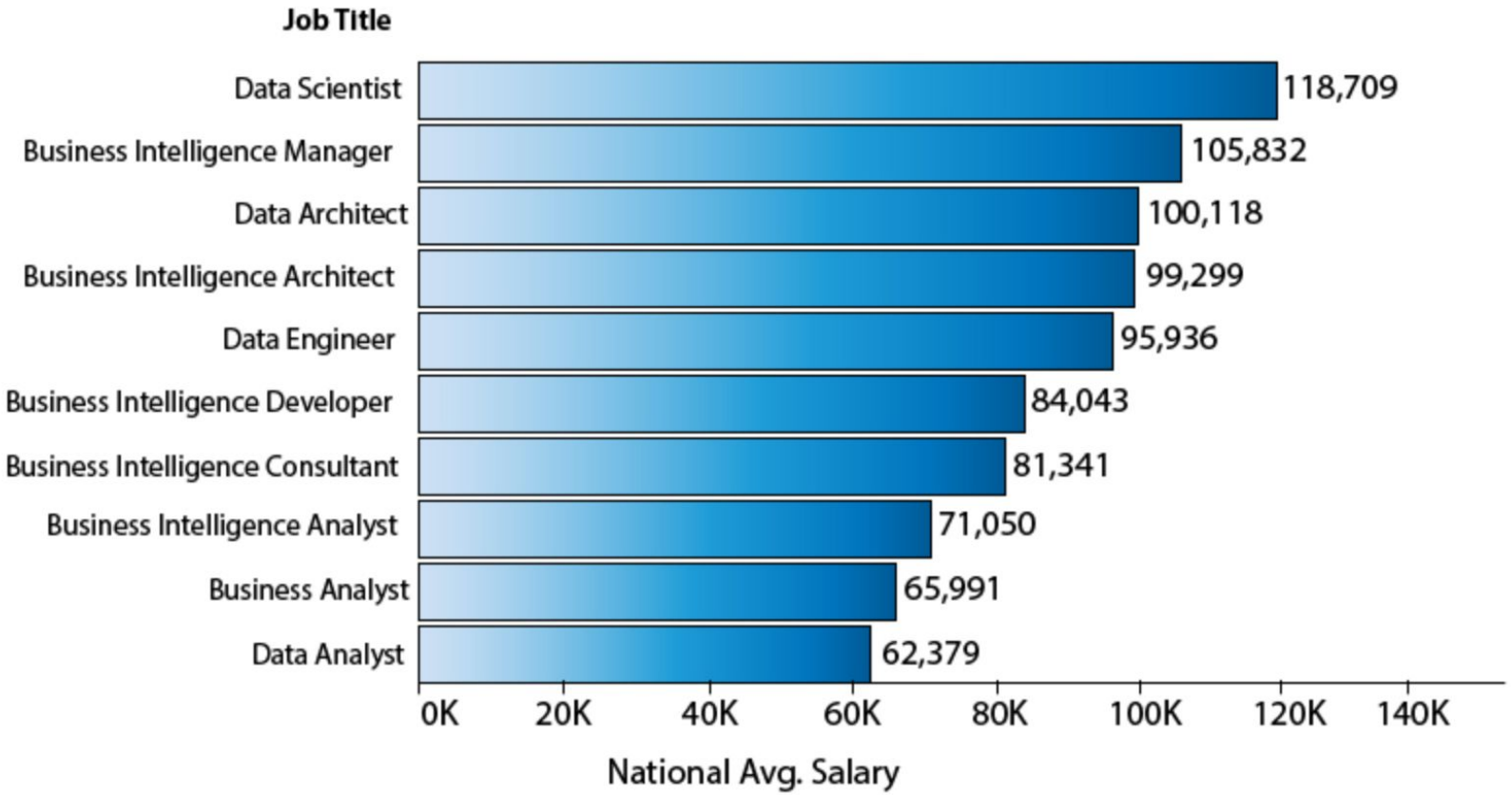
The world needs 28% more data scientists worldwide by 2020 to meet growing demand

There’s a shortage of 151,717 data scientists in the U.S. alone

IBM estimates that in 2020, there will be 2,720,000 job openings available for data scientists.

www.tecla.io/blog/the-high-demand-for-data-scientists-and-how-to-hire-for-them/

Avenga



https://data-flair.training/blogs/why-learn-data-science/

Avenga

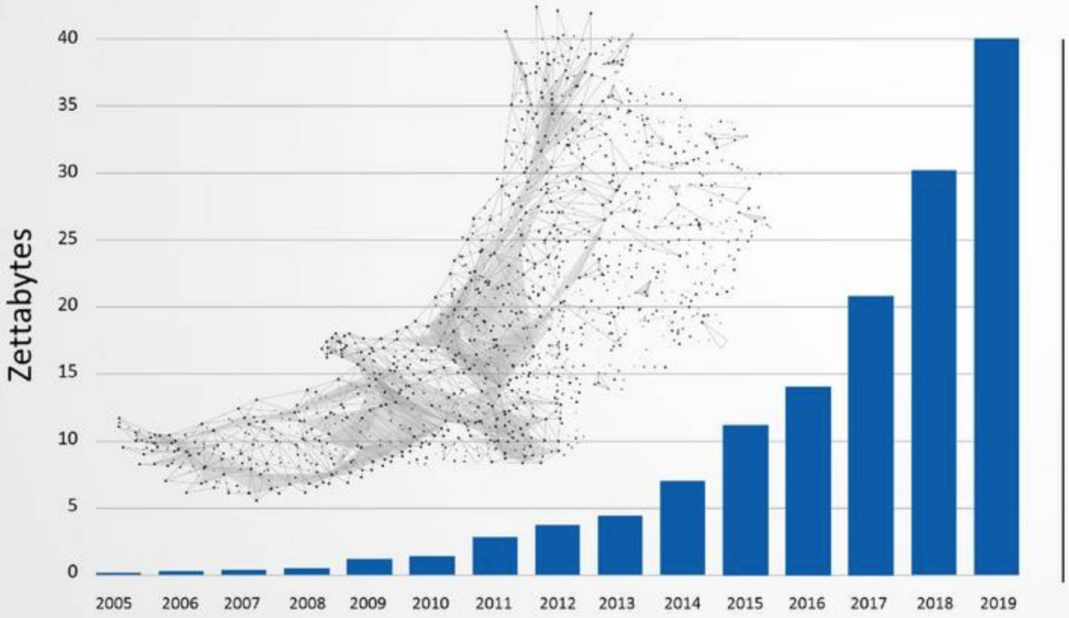
What do you mean "clean all this data"?

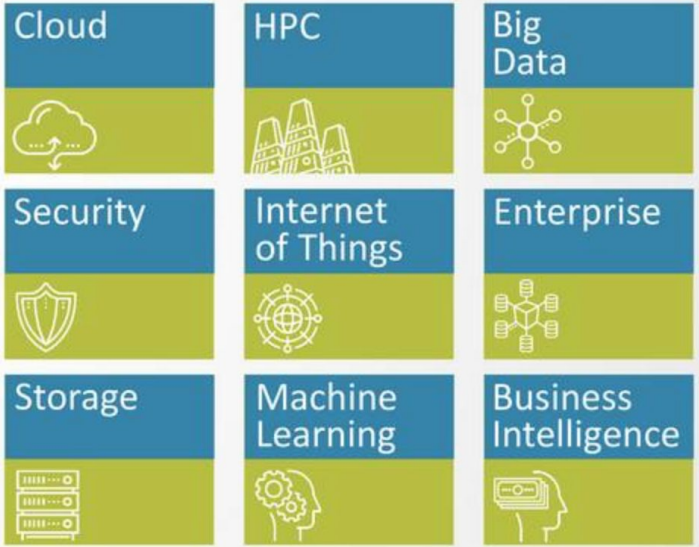
This was sold to me as the 'sexiest job of the 21st Century'.

We Live in a World of Data

- More Devices
- More Applications
- More Data

Exponential Data Growth Everywhere





flightradar24
LIVE AIR TRAFFIC
Greenland
Finland
Sweden
Russia
Ukraine
Kazakhstan
Mongolia
Libya
Mali
Niger
Chad
Ethiopia
Angola
Namibia
Madagascar
Australia
Peru
Brazil
Bolivia
Argentina
South Pacific Ocean
South Atlantic Ocean
Indian Ocean

TONS OF DATA

BUSINESS ON DATA EXCHANGE

MAKING SENSE OF DATA

LinkedIn

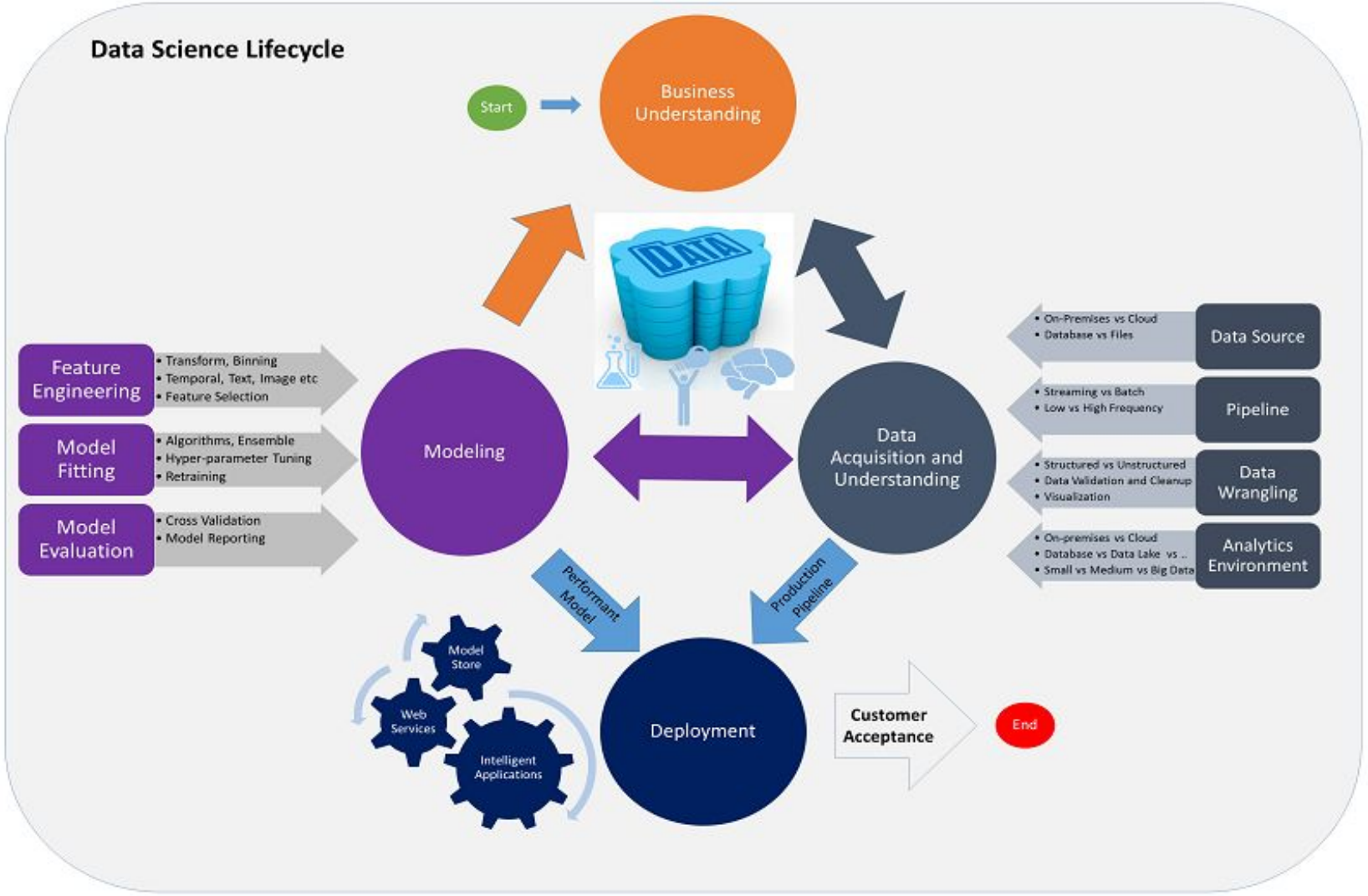
Data Science
create value from data through analytics

What Is Data Science?

Citing the Wikipedia

Data science is, in general terms, the extraction of knowledge from data.

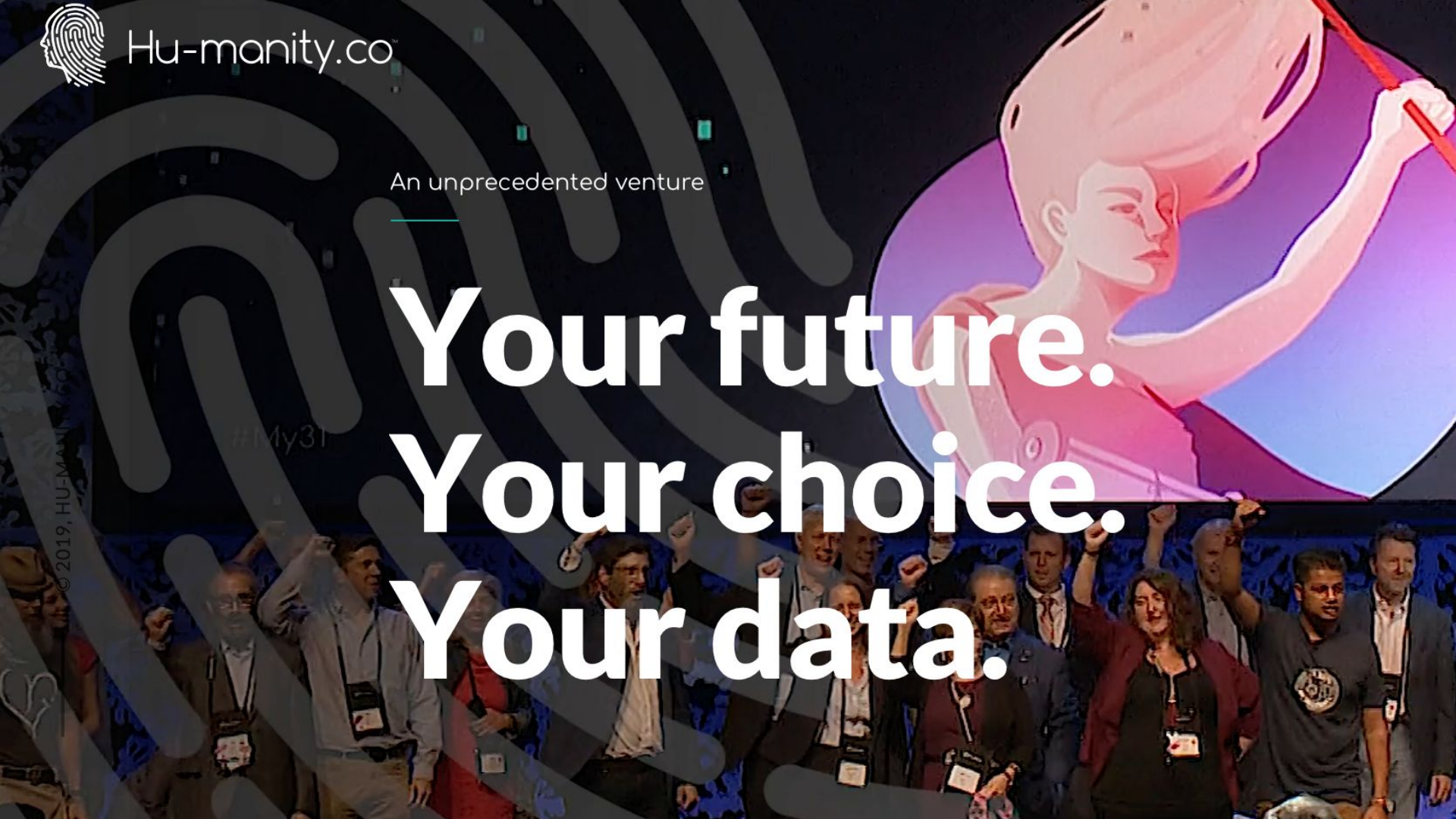
The key word in this job title is "science," with the main goals being to extract meaning from data and to produce data products.


Data Science Lifecycle
Start
Business Understanding
Data Acquisition and Understanding
Modeling
Deployment
Feature Engineering
• Transform, Binning
• Temporal, Text, Image etc
• Feature Selection
Model Fitting
• Algorithms, Ensemble
• Hyper-parameter Tuning
• Retraining
Model Evaluation
• Cross Validation
• Model Reporting
• On-Premises vs Cloud
• Database vs Files
Data Source
• Streaming vs Batch
• Low vs High Frequency
Pipeline
• Structured vs Unstructured
• Data Validation and Cleanup
• Visualization
Data Wrangling
• On-premises vs Cloud
• Database vs Data Lake vs ..
• Small vs Medium vs Big Data
Analytics Environment
Performant Model
Production Pipeline
Model Store
Web Services
Intelligent Applications
Customer Acceptance
End

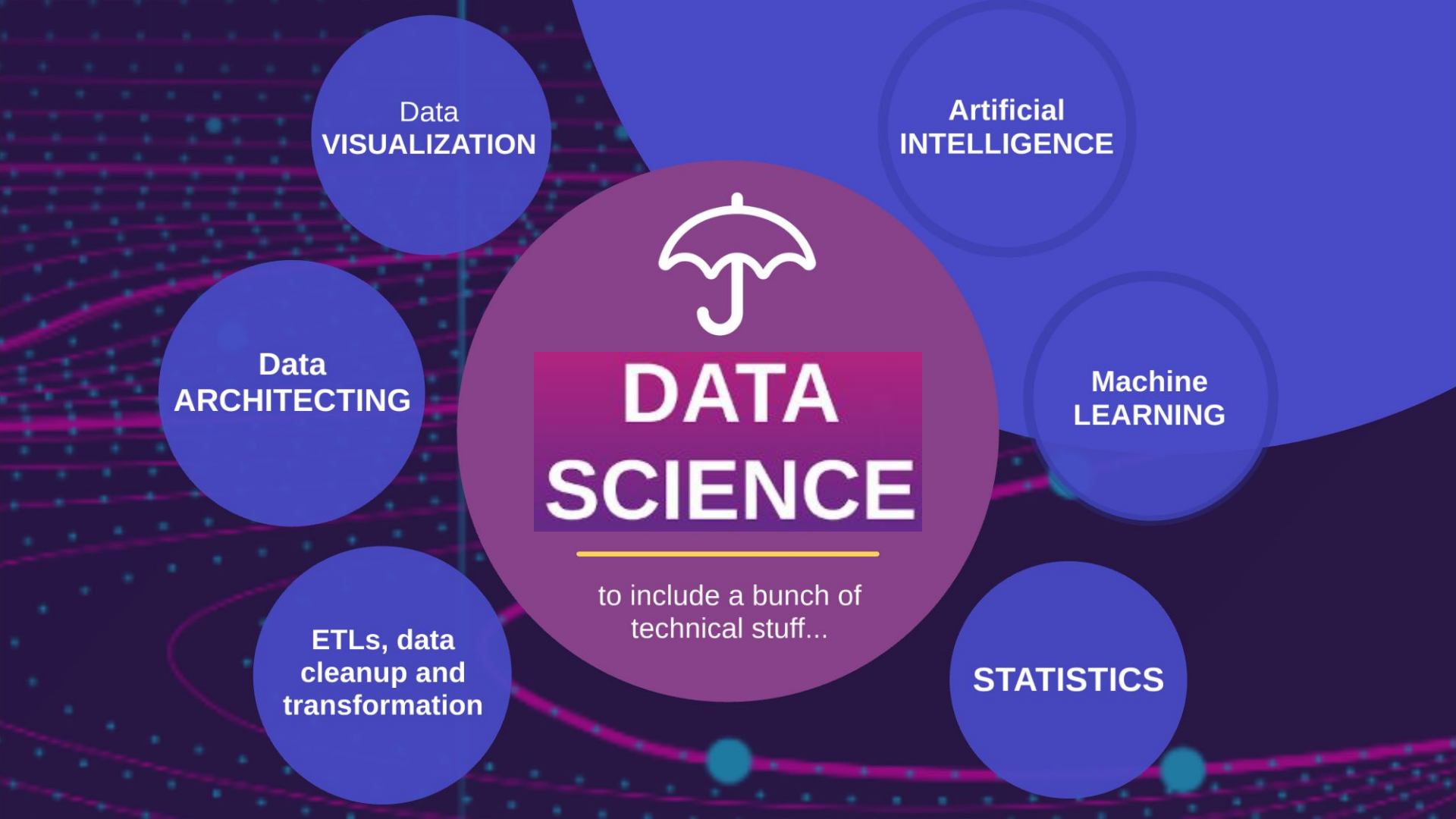
Hu-manity.co™

An unprecedented venture

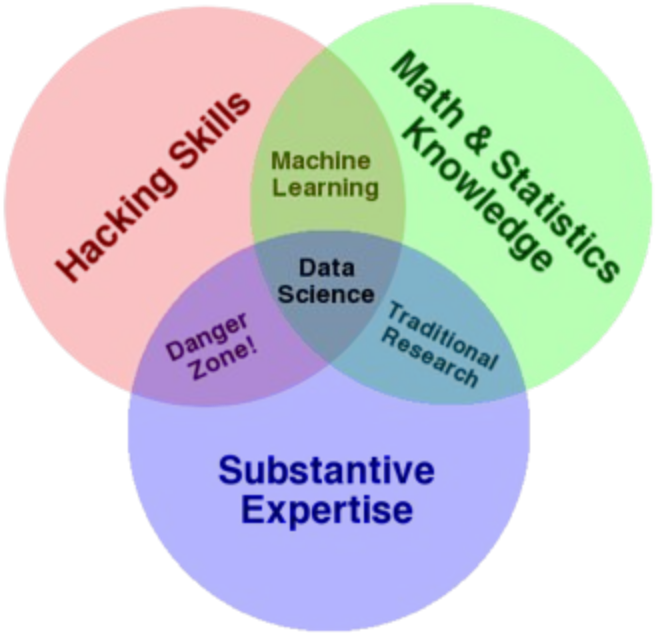
Your future. Your choice. Your data.



© 2019, HU-MANITY.CO


Data
VISUALIZATION
Artificial
INTELLIGENCE
Data
ARCHITECTING
DATA
SCIENCE
Machine
LEARNING
to include a bunch of
technical stuff...
ETLs, data
cleanup and
transformation
STATISTICS

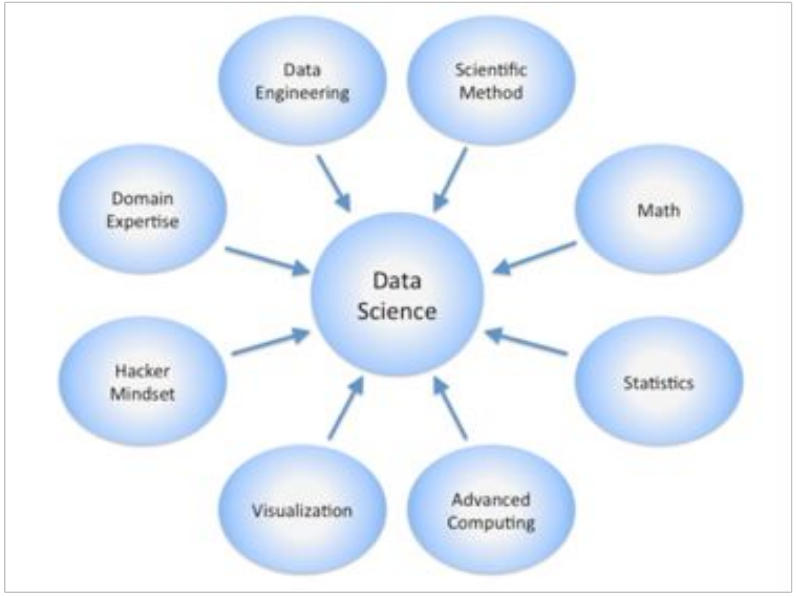
Famous Data Science Venn diagram



What skills do you need to become a consummate data scientist?

https://www.houseofbots.com/news-detail/3994-1-what-skills-do-you-need-to-become-a-consummate-data-scientist

What is Data Scientist?



http://juangabrielgomila.com/en/what-is-a-data-scientist/

AI

AI in Context of Human History

Perspective:

- Universe created
  13.8 billion years ago
- Earth created
  4.54 billion years ago
- Modern humans
  300,000 years ago
- Civilization
  12,000 years ago
- Written record
  5,000 years ago

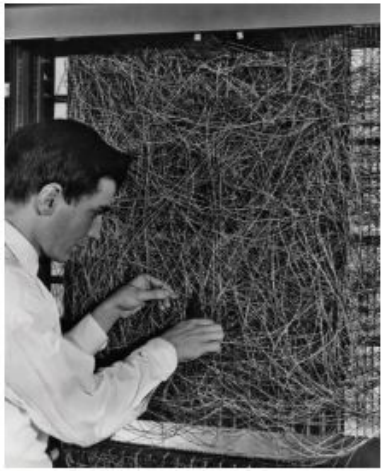
1700s and beyond: Industrial revolution, steam engine, mechanized factory systems, machine tools

AI in Context of Human History

Dreams, mathematical foundations, and engineering in reality.

Alan Turing, 1951: "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control."

AI in Context of Human History



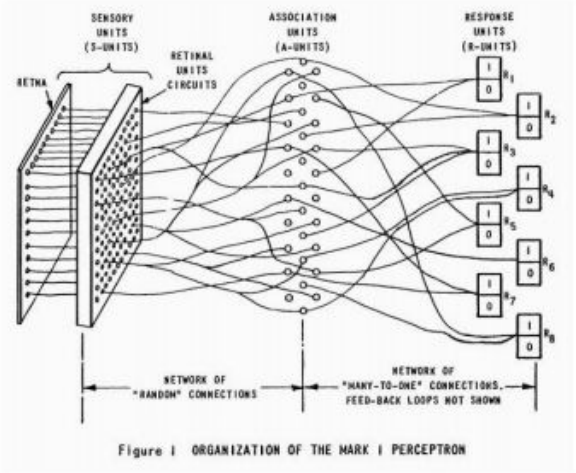


Figure 1 ORGANIZATION OF THE MARK I PERCEPTRON

Dreams, mathematical foundations, and engineering in reality.

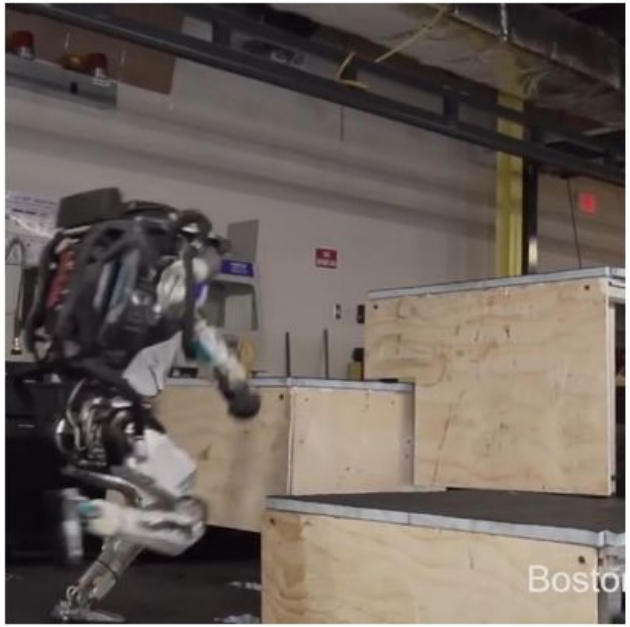
Frank Rosenblatt, Perceptron (1957, 1962): Early description and engineering of single-layer and multi-layer artificial neural networks.

AI in Context of Human History

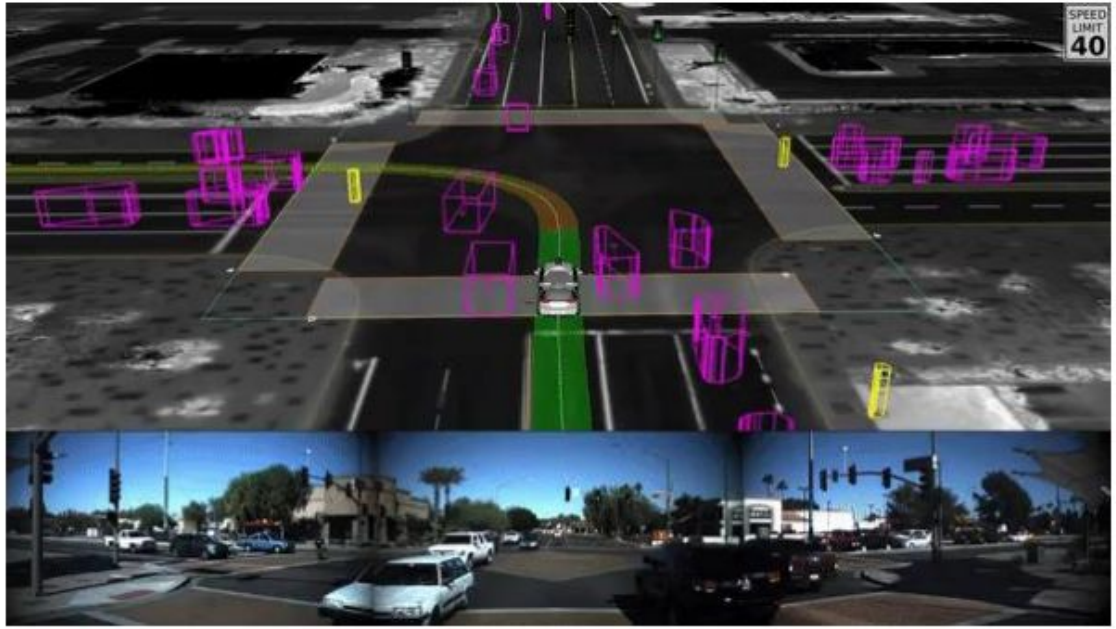
Kasparov vs Deep Blue, 1997

Lee Sedol vs AlphaGo, 2016

AI in Context of Human History



Robots on two legs.



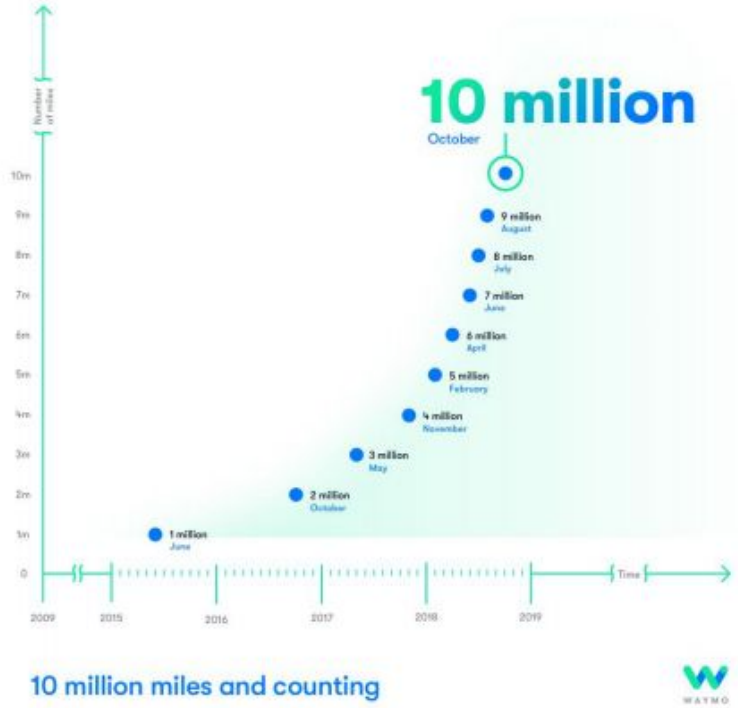
Robots on four wheels.

Human is Responsible

Machine is Responsible

Waymo

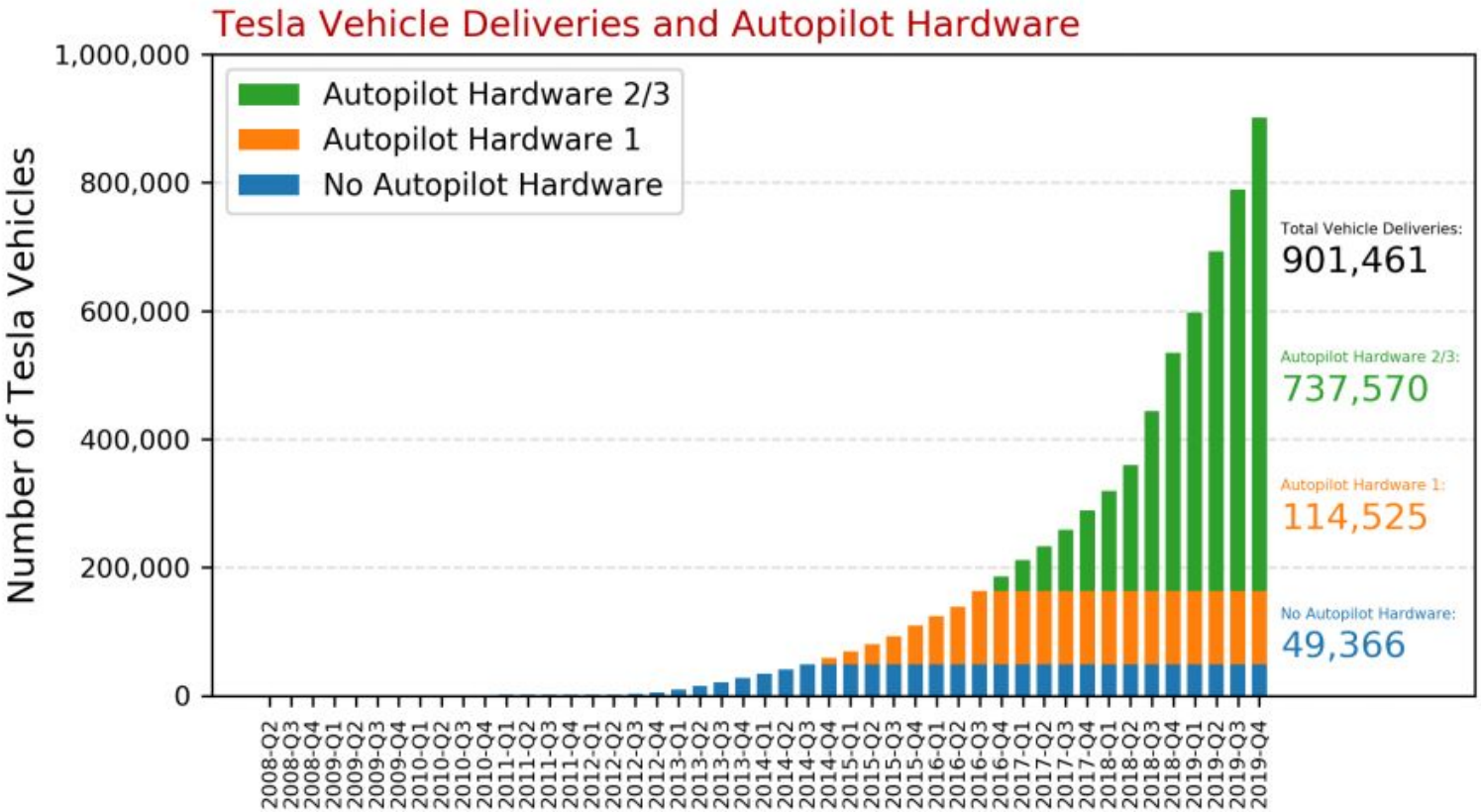
October, 2018:



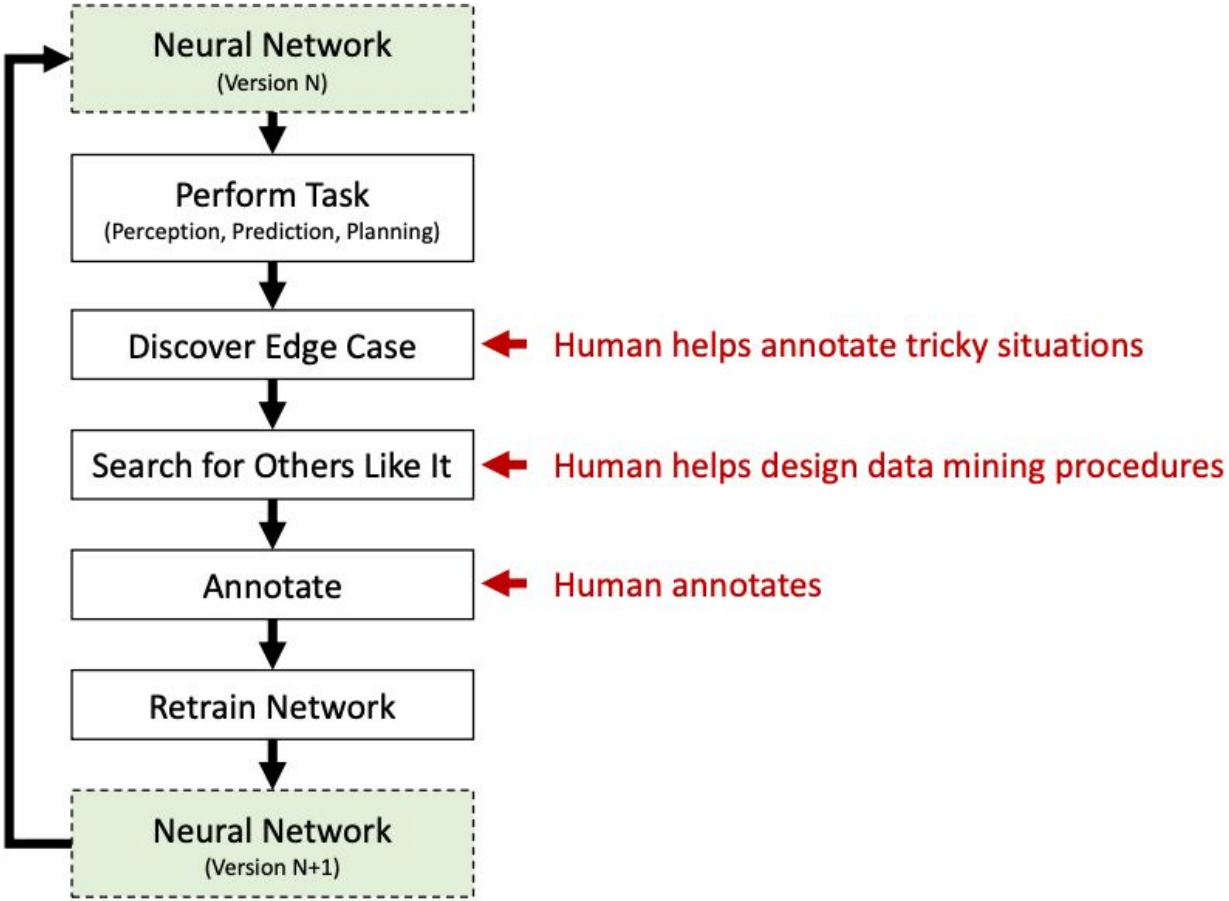
January, 2020:

- **On-road:** 20 million miles
- **Simulation:** 10 billion miles
- **Testing & Validation:** 20,000 classes of structured tests
- Initiated testing without a safety driver

Tesla Autopilot

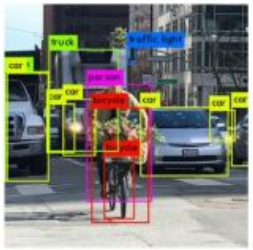


Active Learning Pipeline
(aka Data Engine)



Single-Task Learning

Task 1: Object Detection



- Neural Network 1 (Version N)
- Perform Task 1 (Object Detection)
- Discover Edge Case
- Search for Others Like It
- Annotate
- Retrain Network
- Neural Network 1 (Version N+1)

Task 2: Lane Markings

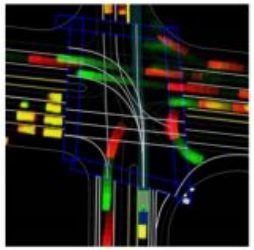
- Neural Network 2 (Version N)
- Perform Task 2 (Lane Marking Detection)
- Discover Edge Case
- Search for Others Like It
- Annotate
- Retrain Network
- Neural Network 2 (Version N+1)

Task 3: Drivable Area

- Neural Network 3 (Version N)
- Perform Task 3 (Drivable Area)
- Discover Edge Case
- Search for Others Like It
- Annotate
- Retrain Network
- Neural Network 3 (Version N+1)

...

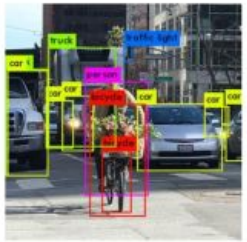
Task 100: Trajectory Generation



- Neural Network 100 (Version N)
- Perform Task 100 (Trajectory Generation)
- Discover Edge Case
- Search for Others Like It
- Annotate
- Retrain Network
- Neural Network 100 (Version N+1)

Multi-Task Learning

Task 1:
Object Detection

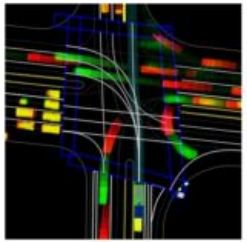


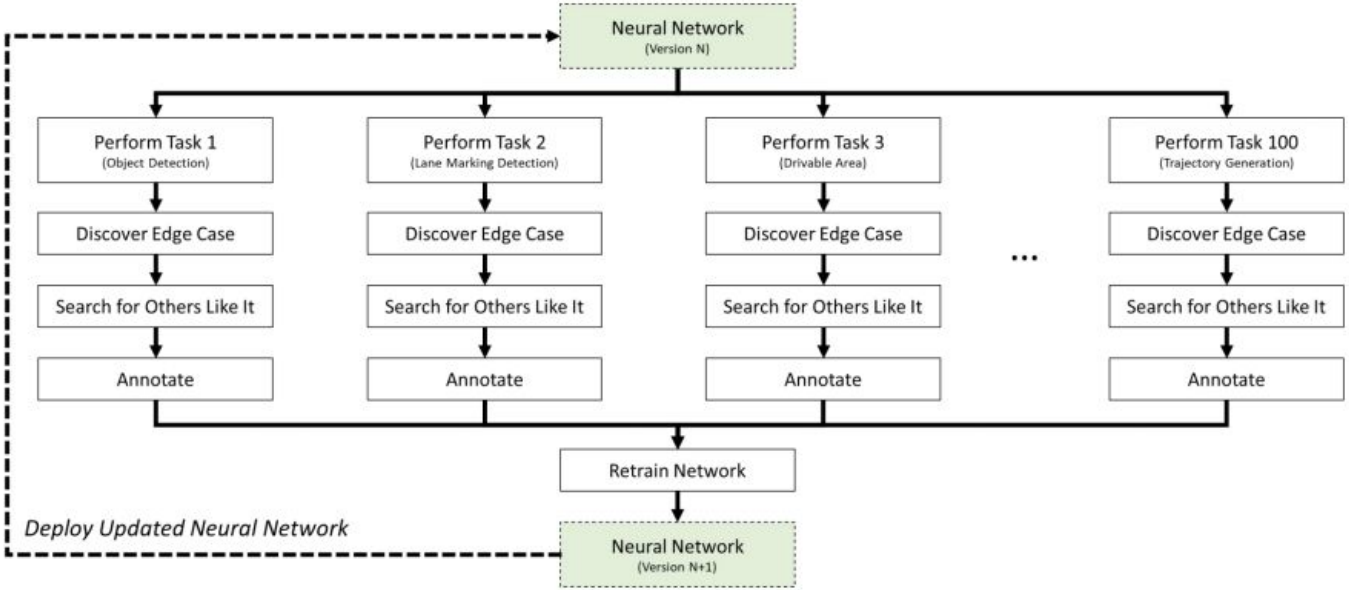
Task 2:
Lane Markings

Task 3:
Drivable Area

...

Task 100:
Trajectory Generation





Types of Artificial Intelligence

- **ANI – Artificial Narrow Intelligence** – It comprises of basic/role tasks such as those performed by chatbots, personal assistants like SIRI by Apple and Alexa by Amazon.

- **AGI – Artificial General Intelligence** – Artificial General Intelligence comprises of human-level tasks such as performed by self-driving cars by Uber, Autopilot by Tesla. It involves continual learning by the machines.

- **ASI – Artificial Super Intelligence** – Artificial Super Intelligence refers to intelligence way smarter than humans.

https://www.upwork.com/hiring/for-clients/artificial-intelligence-and-natural-language-processing-in-big-data/

See: still and video image recognition

Hear: receive input via text or spoken language.

Speak: respond meaningfully to our input – same language or any language.

Make human-like decisions. Offer advice or new knowledge.

Learn: Change its behavior based on changes in its environment.

Move and manipulate physical objects.

Image Processing

Convolutional Neural Nets

Natural Language Processing

Recurrent Neural nets

Question Answering Machines (e.g. Watson)

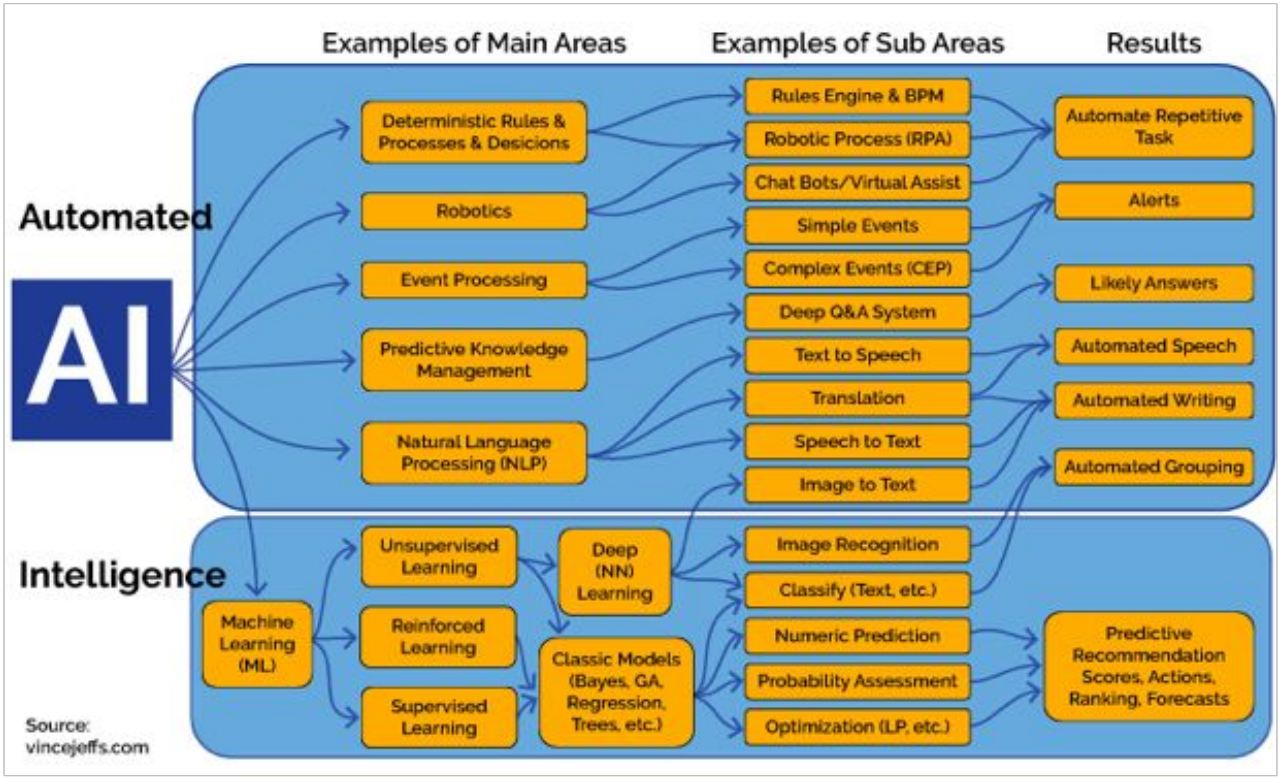
Generative Adversarial Neural Nets

Reinforcement Learning

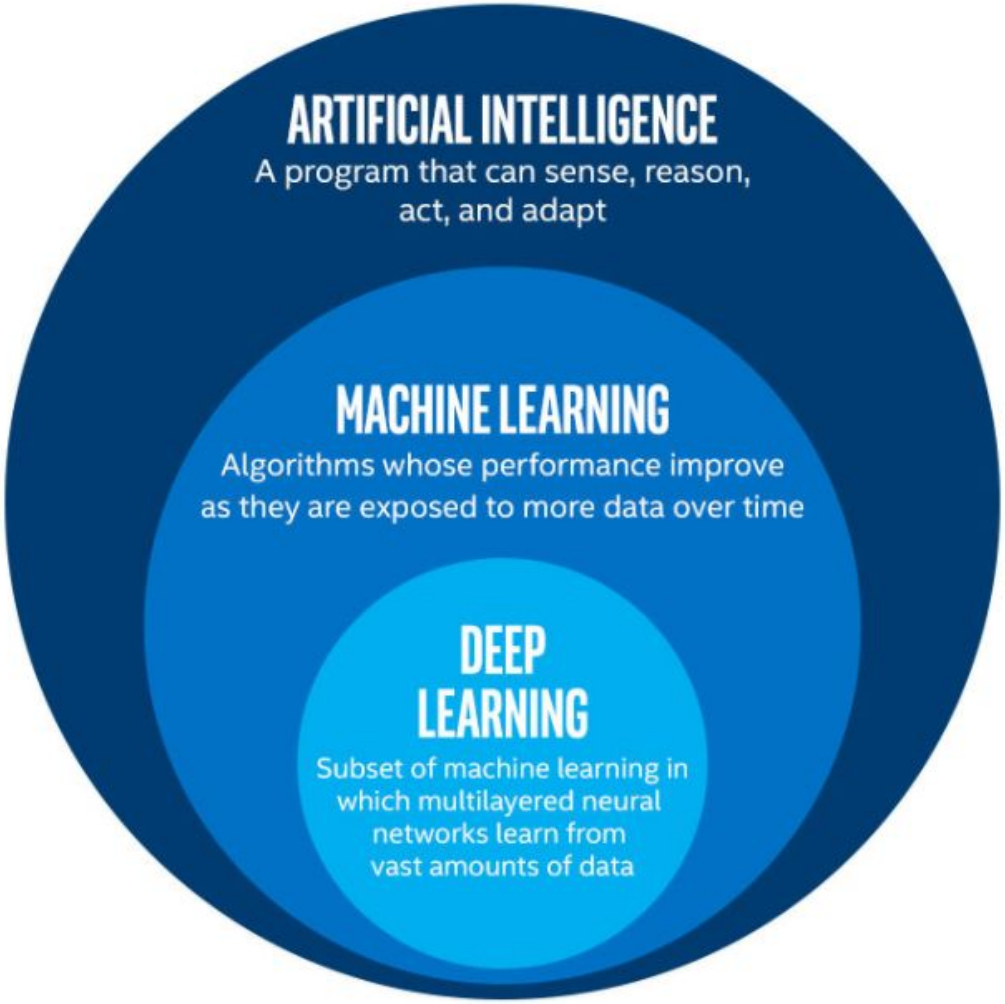
Robotics

Machine Learning
use learning algorithms that simulate human learning capabilities

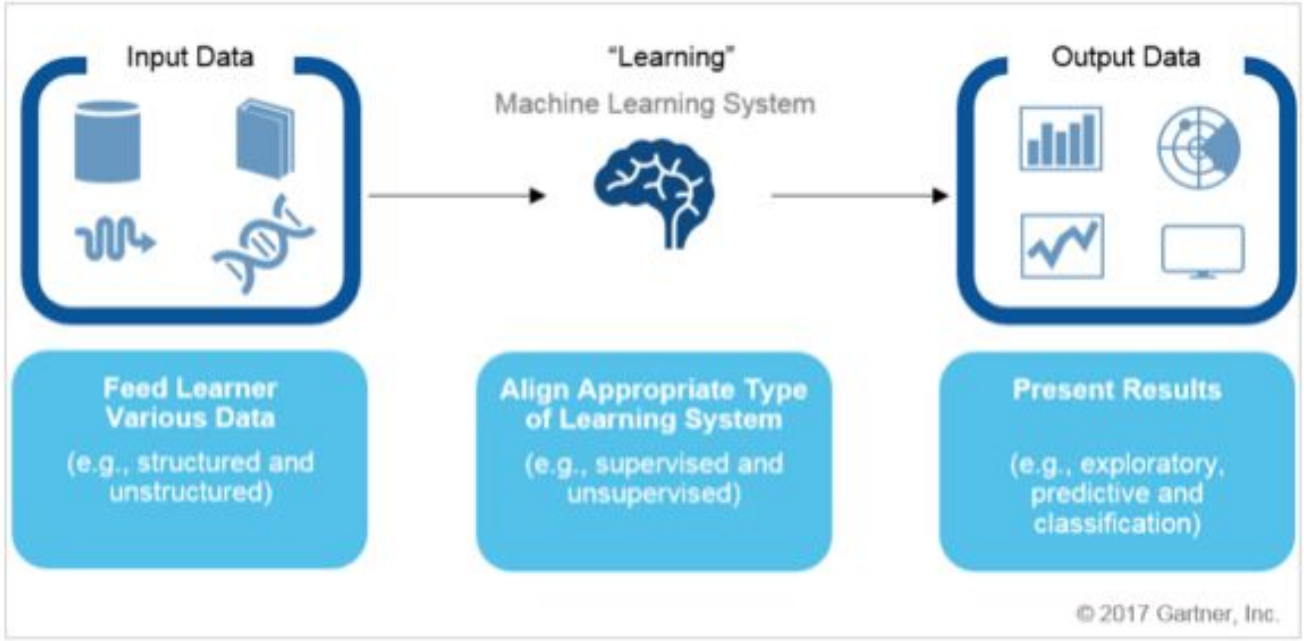
Difference between Automated & Intelligence AI



https://www.upwork.com/hiring/for-clients/artificial-intelligence-and-natural-language-processing-in-big-data/


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

The Basics of Machine Learning Technology



Source: Gartner (January 2017)

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

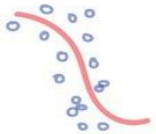
CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»



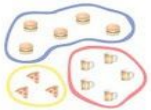
Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»



Find hidden dependencies

ASSOCIATION

«Find what clothes I often wear together»

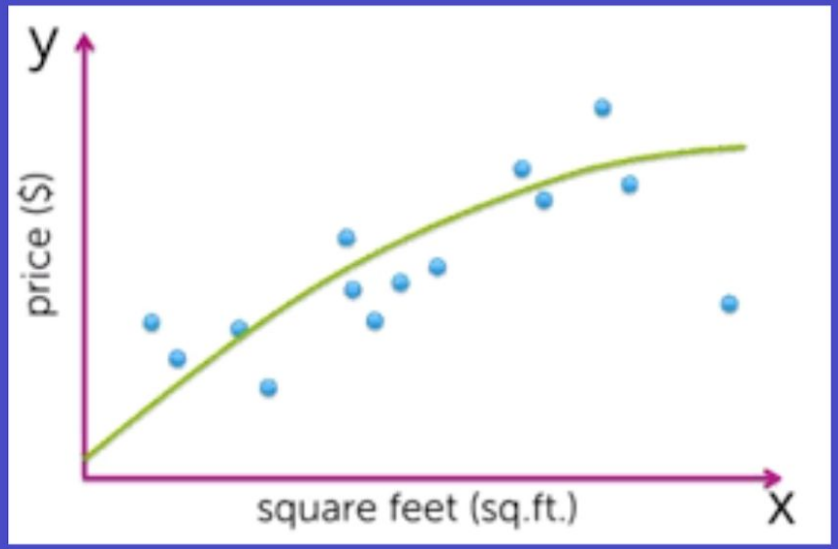
Identify sequences

DIMENSION REDUCTION (generalization)

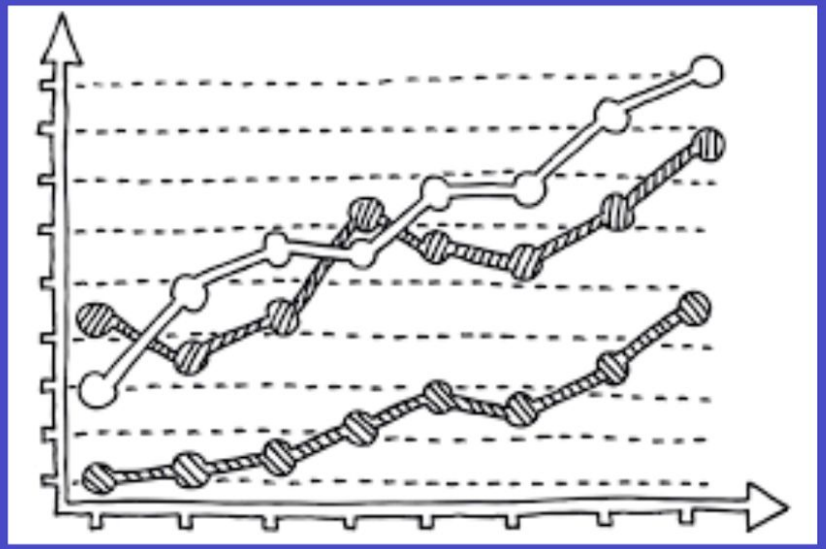
«Make the best outfits from the given clothes»

ESTIMATE A CONTINUOUS RESPONSE **(REGRESSION)**

Estimate price of a house from size and location and other characteristics

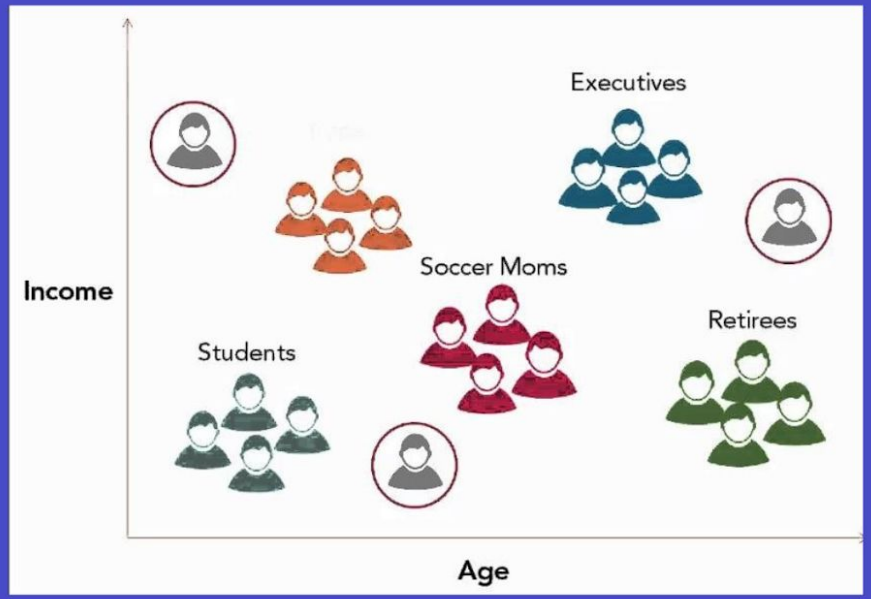
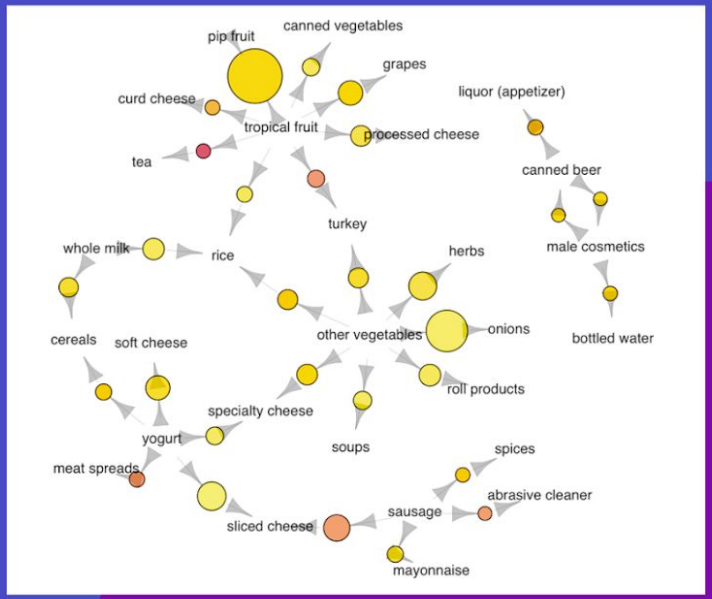


Estimate next month sales of product X at store Y



UNSPERVISED **LEARNING**

Customer segmentation (*clustering*)



What products are bought together (*basket analysis*)