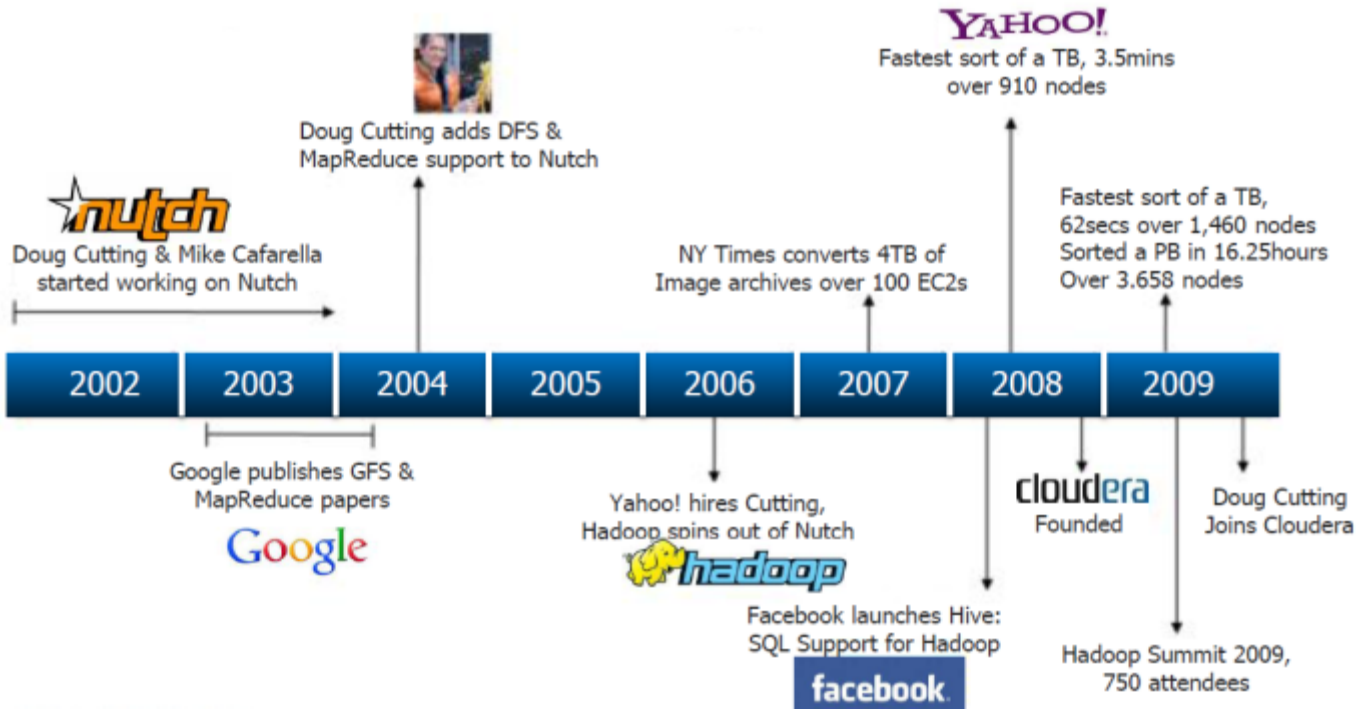
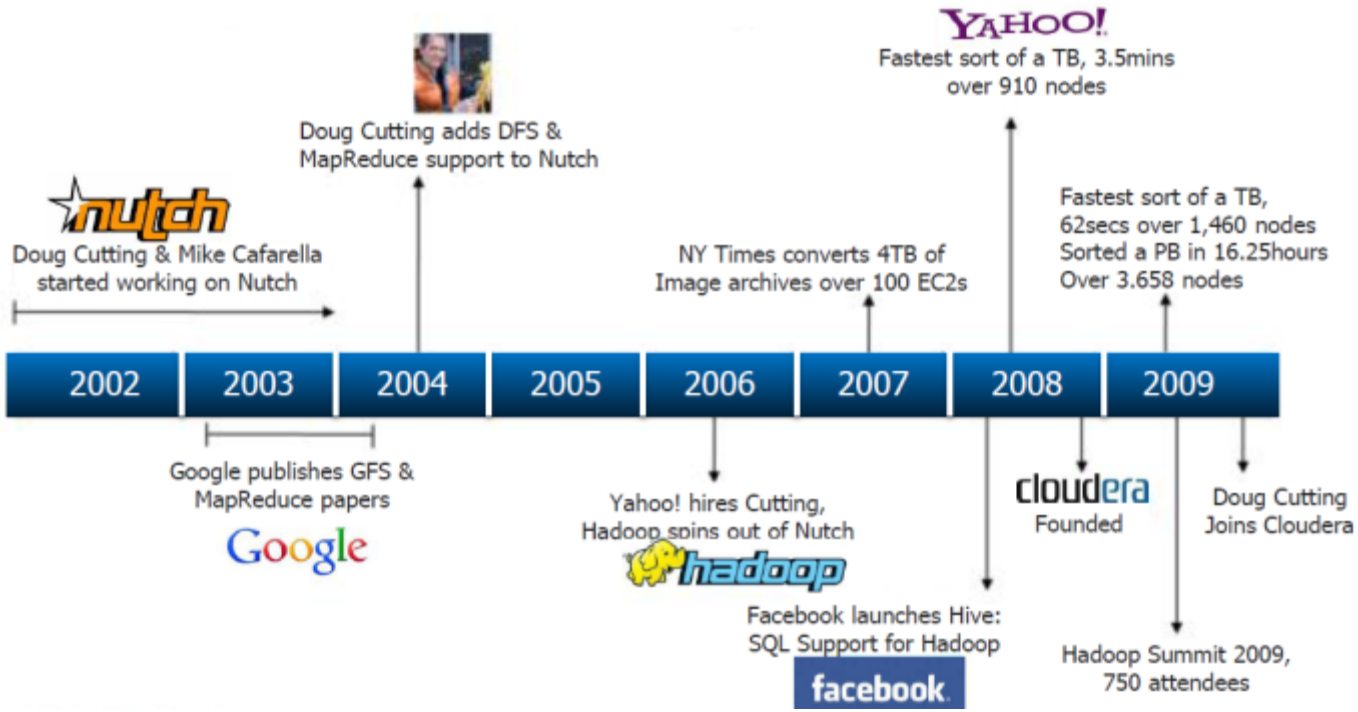
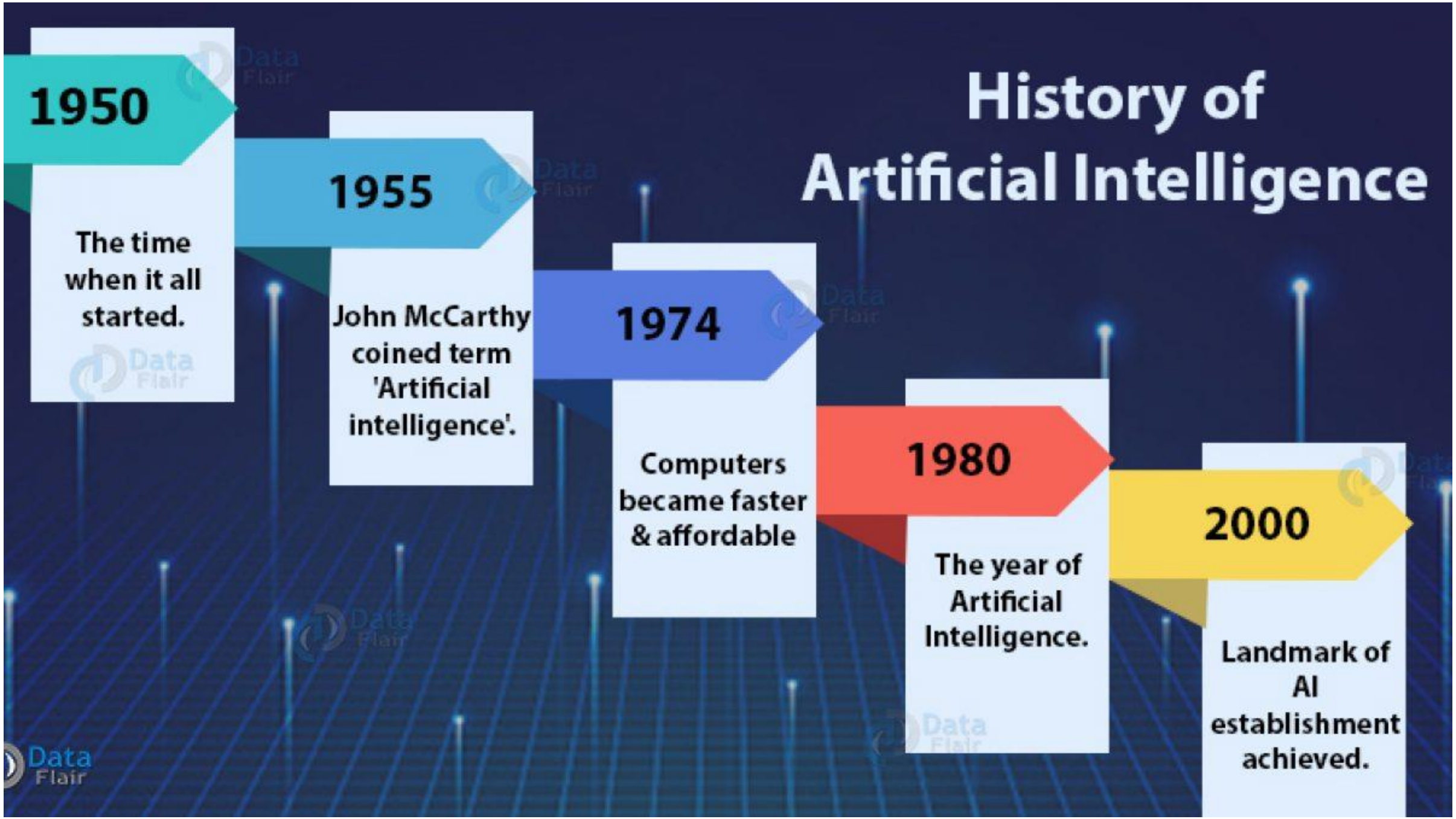
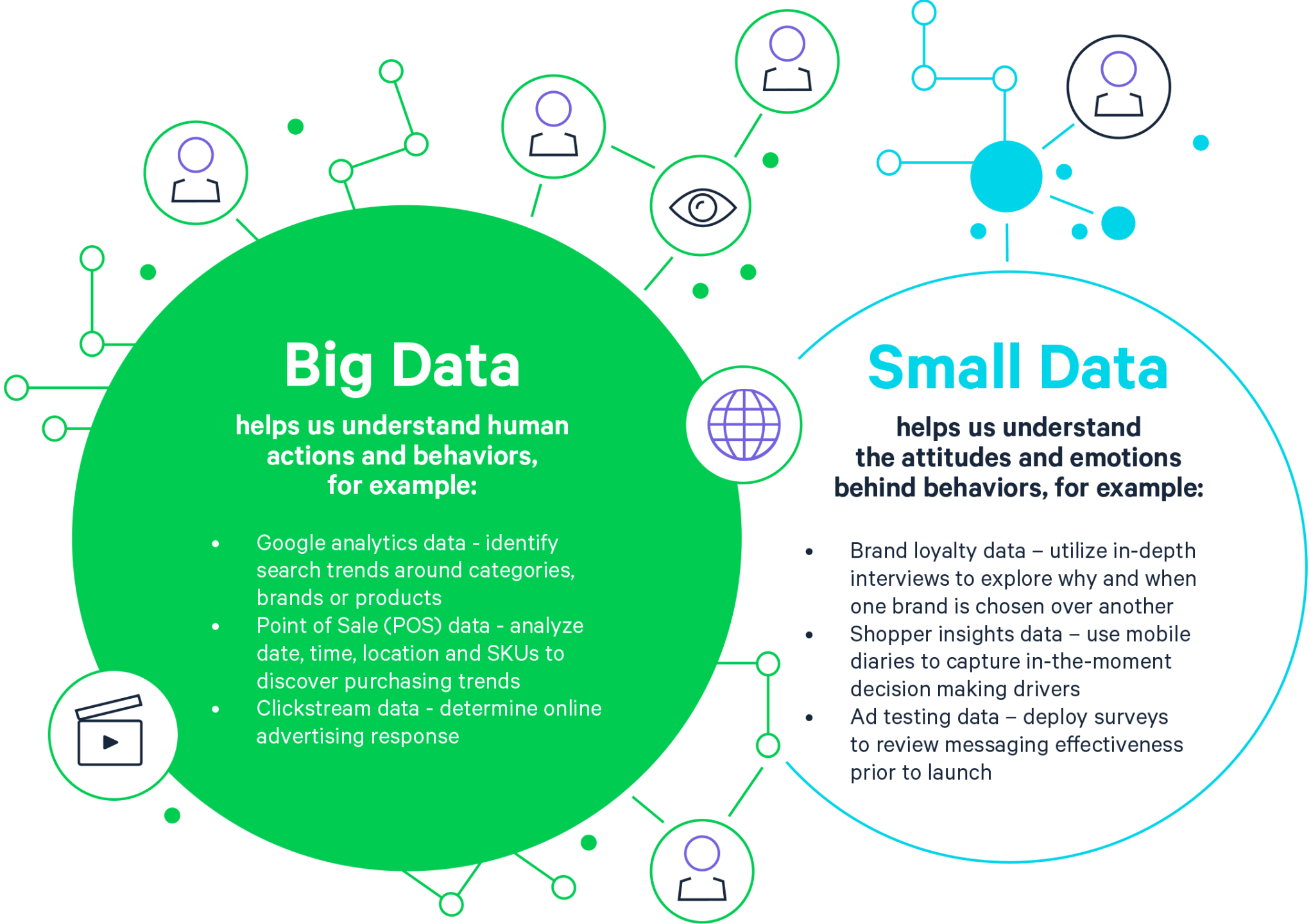
Introduction to Machine Learning


YAHOO!
Fastest sort of a TB, 3.5mins
over 910 nodes
Doug Cutting adds DFS &
MapReduce support to Nutch
nutch
Doug Cutting & Mike Cafarella
started working on Nutch
Fastest sort of a TB,
62secs over 1,460 nodes
Sorted a PB in 16.25hours
Over 3.658 nodes
NY Times converts 4TB of
Image archives over 100 EC2s
2002
2003
2004
2005
2006
2007
2008
2009
Google publishes GFS &
MapReduce papers
Google
Yahoo! hires Cutting,
Hadoop spins out of Nutch
hadoop
cloudera
Founded
Doug Cutting
Joins Cloudera
Facebook launches Hive:
SQL Support for Hadoop
facebook
Hadoop Summit 2009,
750 attendees


YAHOO!
Fastest sort of a TB, 3.5mins
over 910 nodes
Doug Cutting adds DFS &
MapReduce support to Nutch
nutch
Doug Cutting & Mike Cafarella
started working on Nutch
Fastest sort of a TB,
62secs over 1,460 nodes
Sorted a PB in 16.25hours
Over 3.658 nodes
NY Times converts 4TB of
Image archives over 100 EC2s
2002
2003
2004
2005
2006
2007
2008
2009
Google publishes GFS &
MapReduce papers
Google
Yahoo! hires Cutting,
Hadoop spins out of Nutch
hadoop
cloudera
Founded
Doug Cutting
Joins Cloudera
Facebook launches Hive:
SQL Support for Hadoop
facebook
Hadoop Summit 2009,
750 attendees


History of Artificial Intelligence
1950
The time when it all started.
1955
John McCarthy coined term 'Artificial intelligence'.
1974
Computers became faster & affordable
1980
The year of Artificial Intelligence.
2000
Landmark of AI establishment achieved.
Data Flair


Big Data
helps us understand human actions and behaviors, for example:
• Google analytics data - identify search trends around categories, brands or products
• Point of Sale (POS) data - analyze date, time, location and SKUs to discover purchasing trends
• Clickstream data - determine online advertising response
Small Data
helps us understand the attitudes and emotions behind behaviors, for example:
• Brand loyalty data – utilize in-depth interviews to explore why and when one brand is chosen over another
• Shopper insights data – use mobile diaries to capture in-the-moment decision making drivers
• Ad testing data – deploy surveys to review messaging effectiveness prior to launch

	Big Data	Small Data
Data Condition	Always unstructured, not ready for analysis, many relational database tables that need merged	Ready for analysis, flat file, no need for merging tables.
Location	Cloud, Offshore, SQL Server, etc.	Database, local PC
Data Size	Over 50K Variables, over 50K individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper
Data Purpose	No intended purpose	Intended purpose for Data Collection

Big data and small data

Analysis

01

Big data: Hard to get the information

Small data: Easy to get the information

Source

02

Big data: Outside the enterprise

Small data: Trad. enterprise data

Information

03

Big data: Big picture, hidden correlations

Small data: Specific, tergeted

Size

04

Big data > Terabytes (10^12)

Small data < Terabytes (10^12)

Use

05

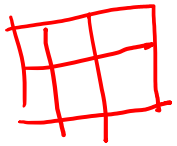
Big data: Complex, BI, predictive, insights

Small data: BI, analysis, reporting

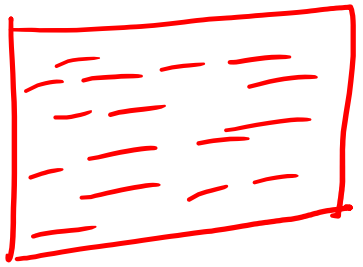
Small Data vs Big Data

Big Data
- structured
- semistructured
- unstructured

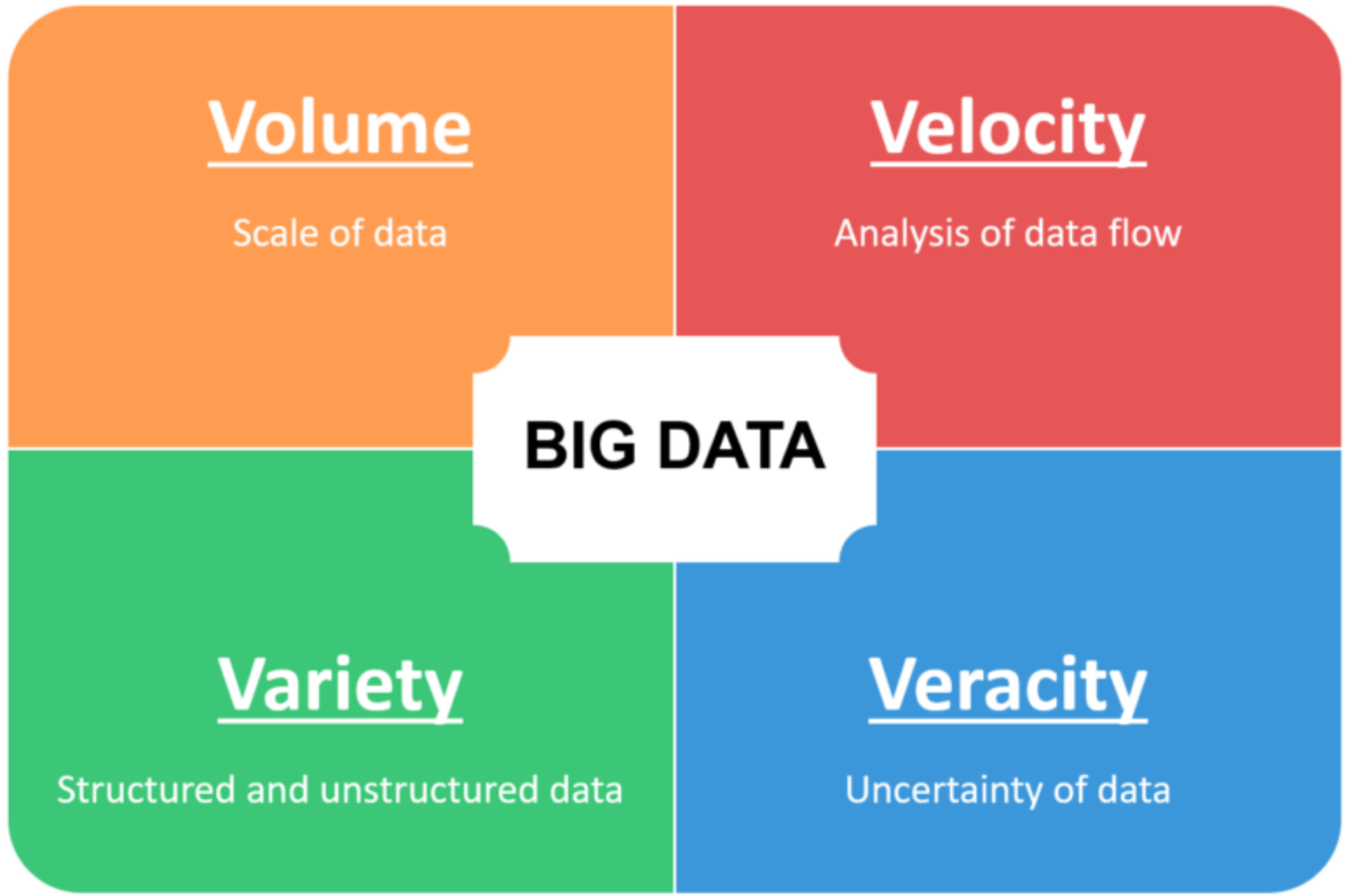
examples:—
- Employee Data
- Online Transactions
- Custome Feedback
- Student Data
- Sales
- Railway reservations



Youtube video
Facebook Posts
Customer Reviews
Amazon Reviews
⭐⭐

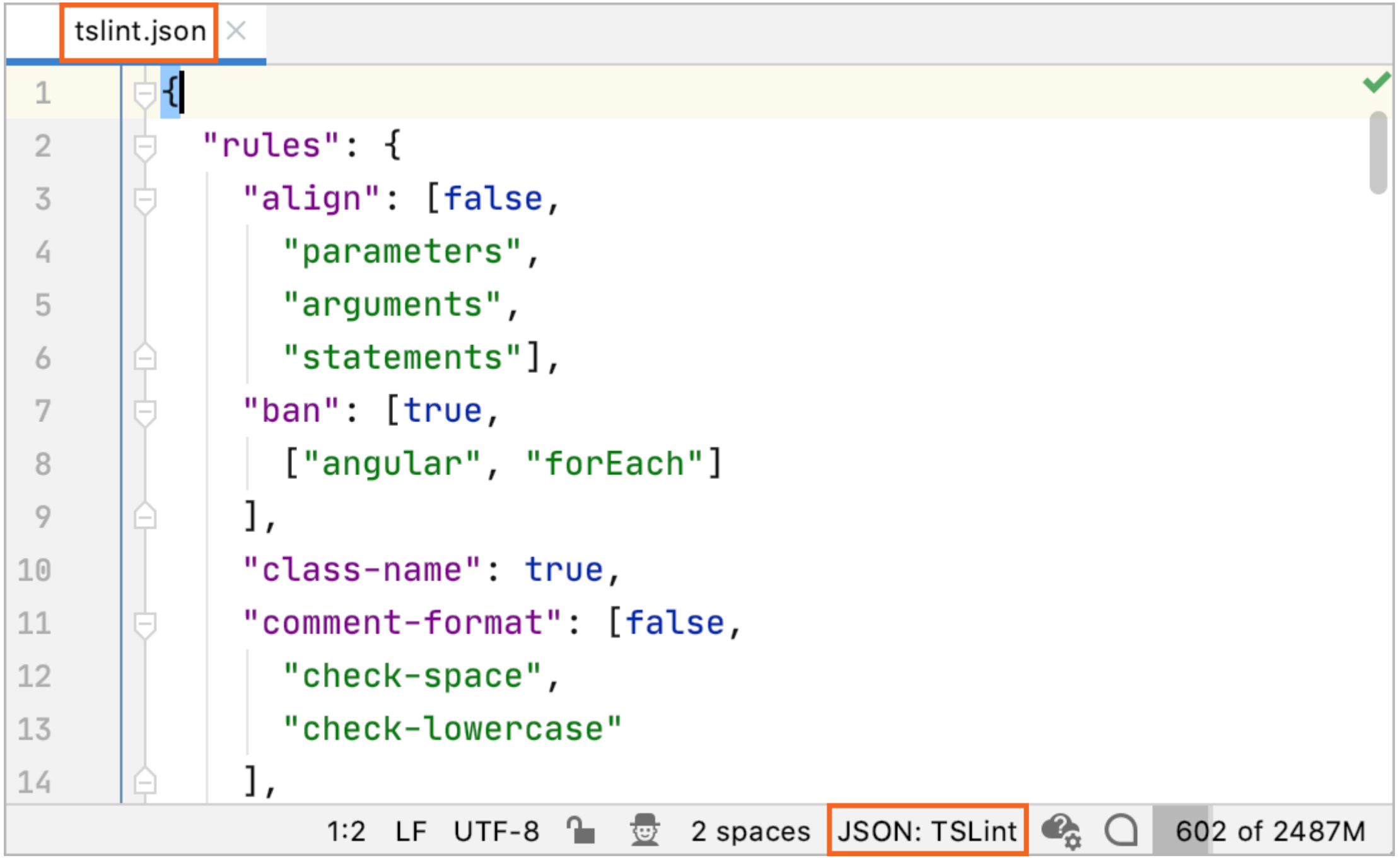


IOT machine Data


Volume
Scale of data
Velocity
Analysis of data flow
BIG DATA
Variety
Structured and unstructured data
Veracity
Uncertainty of data

Category	Big Data	Small Data
Data Sources	Data generated outside the enterprise from nontraditional data sources, Include: • Social media • Sensor data • Log data • Device data • Video, Images, ect.	Traditional enterprise data. Includes: • Enterprise Resource Planning transactional data • Customer Relationship Management (CRM) system • Web transactions • Financial data e.g. general ledger data
Volume	• Terrabytes (10ˇ12) • Petabytes (10ˇ15) • Exabytes (10ˇ18) • Zettabytes(10ˇ21)	• Gigabytes (10ˇ9) • Terabytes (10ˇ12)
Velocity	• Often real-time • Requires immediate response	• Batch or near real-time • Does not always require immediate response
Variety	• Srtuctured • Unstructured • Multi-structured	• Structured • Unstructured
Value	• Complex, advanced, predictive business analysis and insights	• Business Intelligence, analysis and reporting

FEATURES	STRUCTURED	SEMI STRUCTURED	UNSTRUCTURED
Format Type	Relational Database	HTML, XML, JSON	Binary, Character
Version Management	Rows, columns, tuples	Not as common – graph is possible	Whole data
Implementation	SQL	Anonymous nodes	-
Robustness	Robust	Limited robustness	-
Storage Requirement	Less	Significant	Large
Applications	DBMS, RDF, ERP system, Data Warehouse, Apache Parquet, Financial Data, Relational Table	Server Logs, Sensor Output	No SQL, Video, Audio, Social Media, Online Forums, MRI, Ultrasound

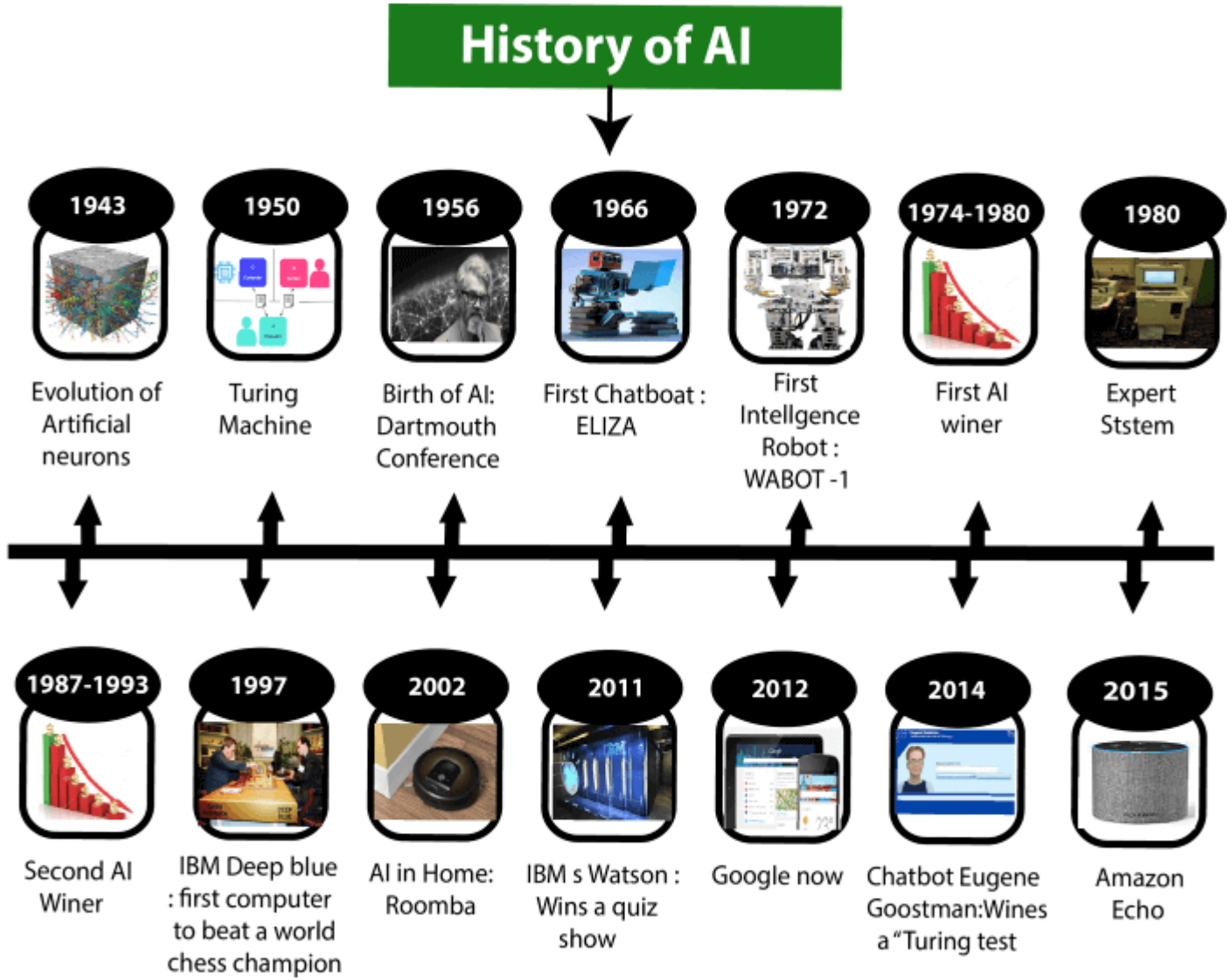

tslint.json
{
"rules": {
"align": [false,
"parameters",
"arguments",
"statements"],
"ban": [true,
["angular", "forEach"]
],
"class-name": true,
"comment-format": [false,
"check-space",
"check-lowercase"
],
1:2 LF UTF-8 2 spaces JSON: TSLint 602 of 2487M

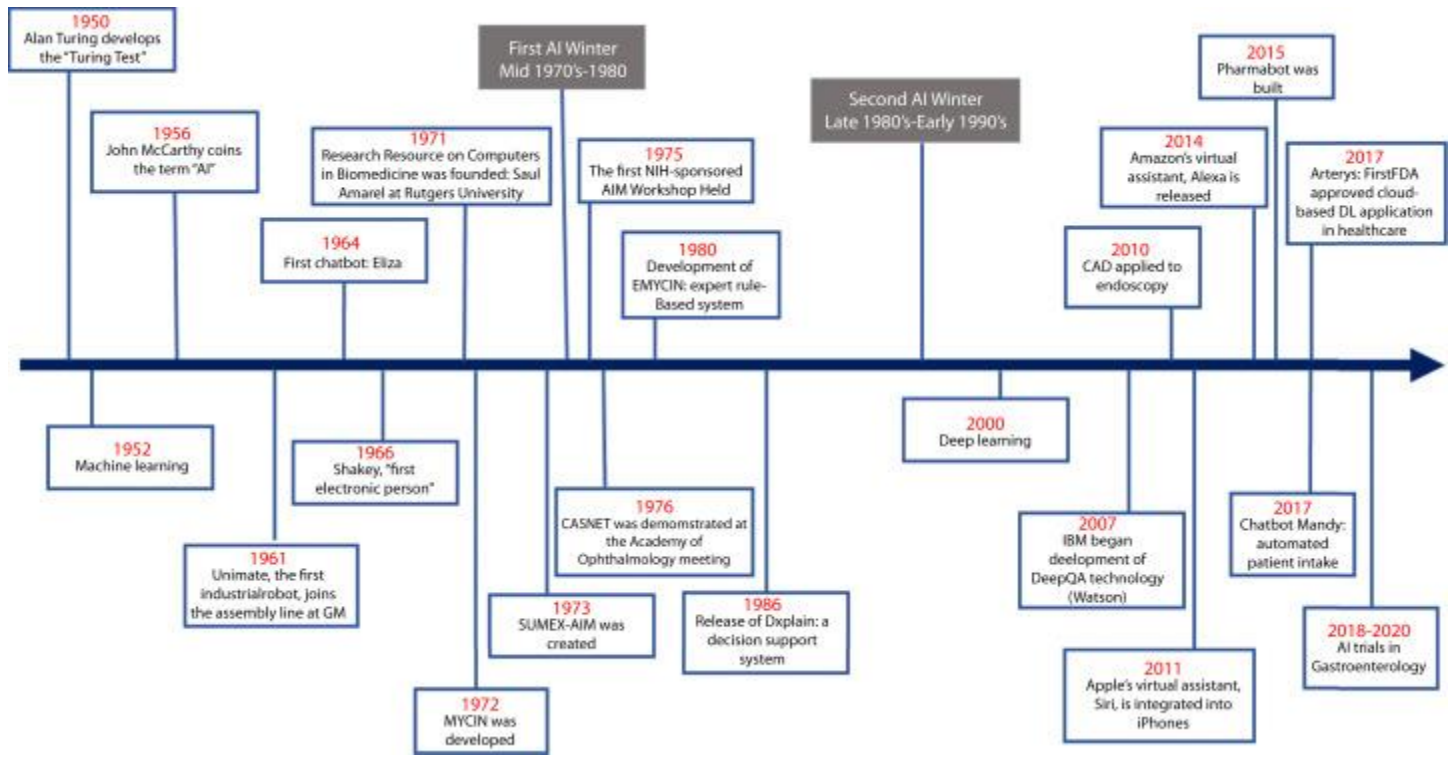
FirstName	LastName	isAlive	Age	Address
John	Smith	True	27	21 2nd

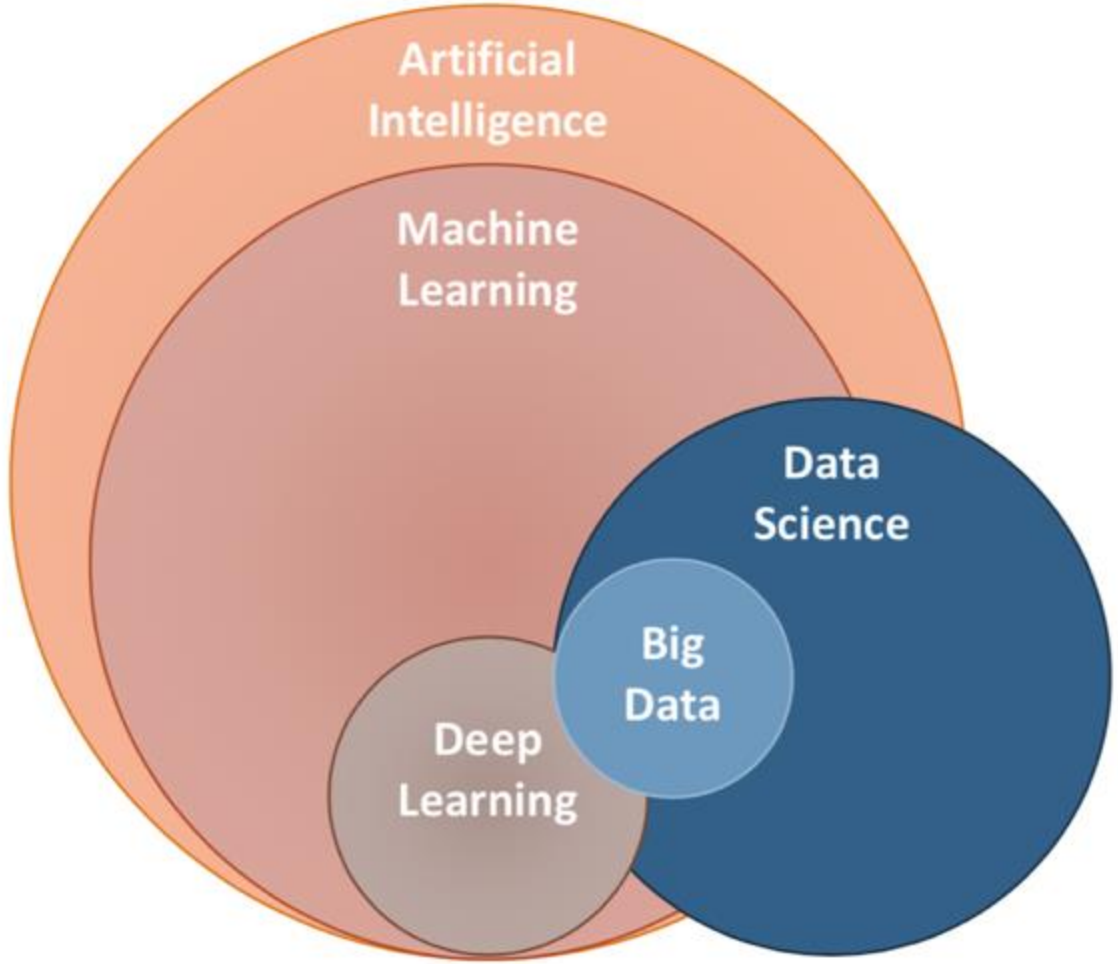
FirstName-John, LastName - Smith, isALive-True, Age -27, Address 21 2nd

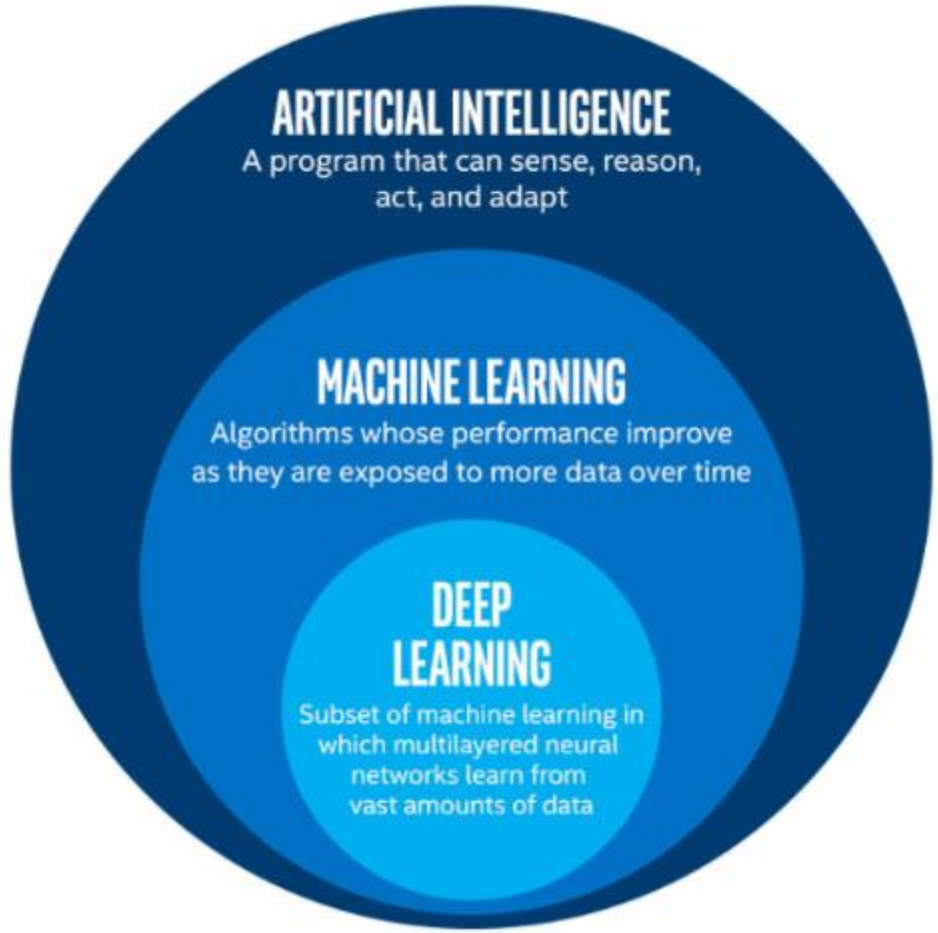

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

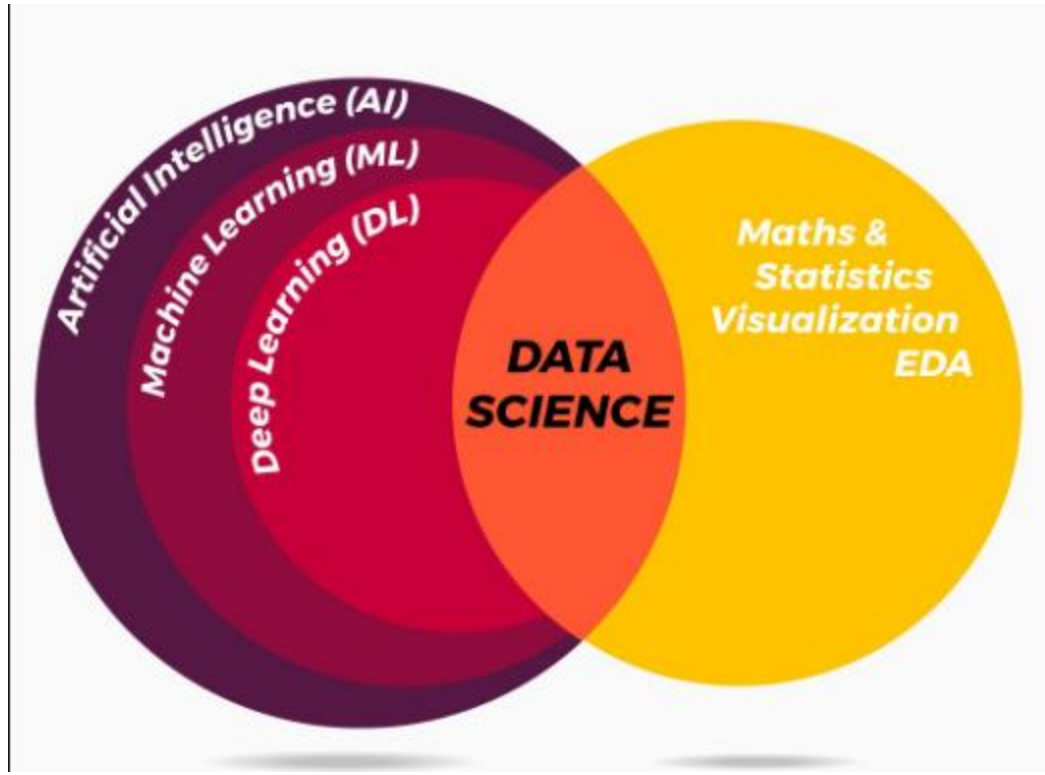
History of AI



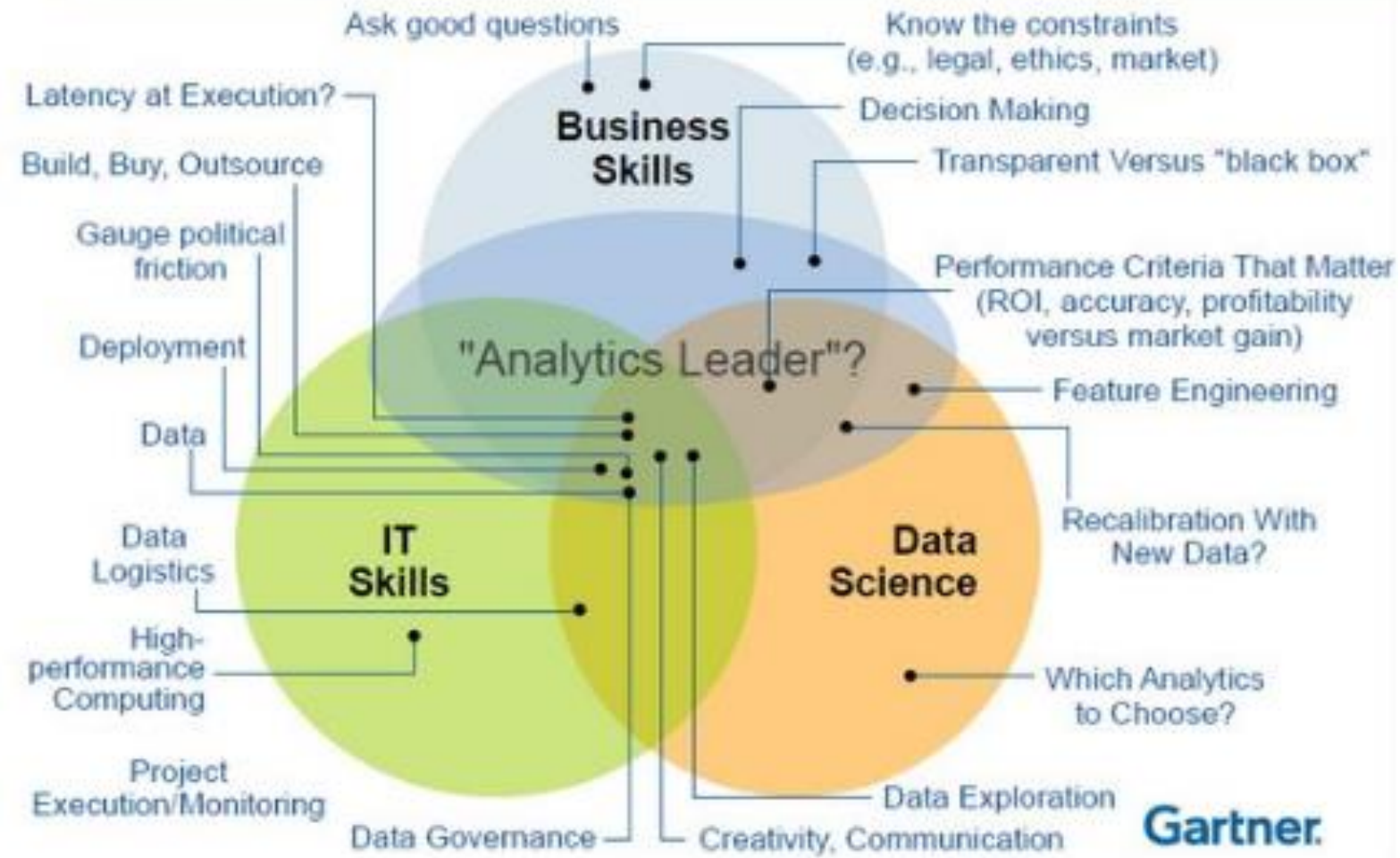

1950
Alan Turing develops the "Turing Test"
1956
John McCarthy coins the term "AI"
1971
Research Resource on Computers in Biomedicine was founded: Saul Amarel at Rutgers University
First AI Winter
Mid 1970's-1980
1975
The first NIH-sponsored AIM Workshop Held
Second AI Winter
Late 1980's-Early 1990's
2015
Pharmabot was built
2014
Amazon's virtual assistant, Alexa is released
2017
Arterys: FirstFDA approved cloud-based DL application in healthcare
1964
First chatbot: Eliza
1980
Development of EMYCIN: expert rule-Based system
2010
CAD applied to endoscopy
1952
Machine learning
1966
Shakey, "first electronic person"
2000
Deep learning
1961
Unimate, the first industrialrobot, joins the assembly line at GM
1976
CASNET was demomstrated at the Academy of Ophthalmology meeting
2007
IBM began deelopment of DeepQA technology (Watson)
2017
Chatbot Mandy: automated patient intake
1973
SUMEX-AIM was created
1986
Release of Dxplain: a decision support system
2018-2020
AI trials in Gastroenterology
1972
MYCIN was developed
2011
Apple's virtual assistant, Siri, is integrated into iPhones

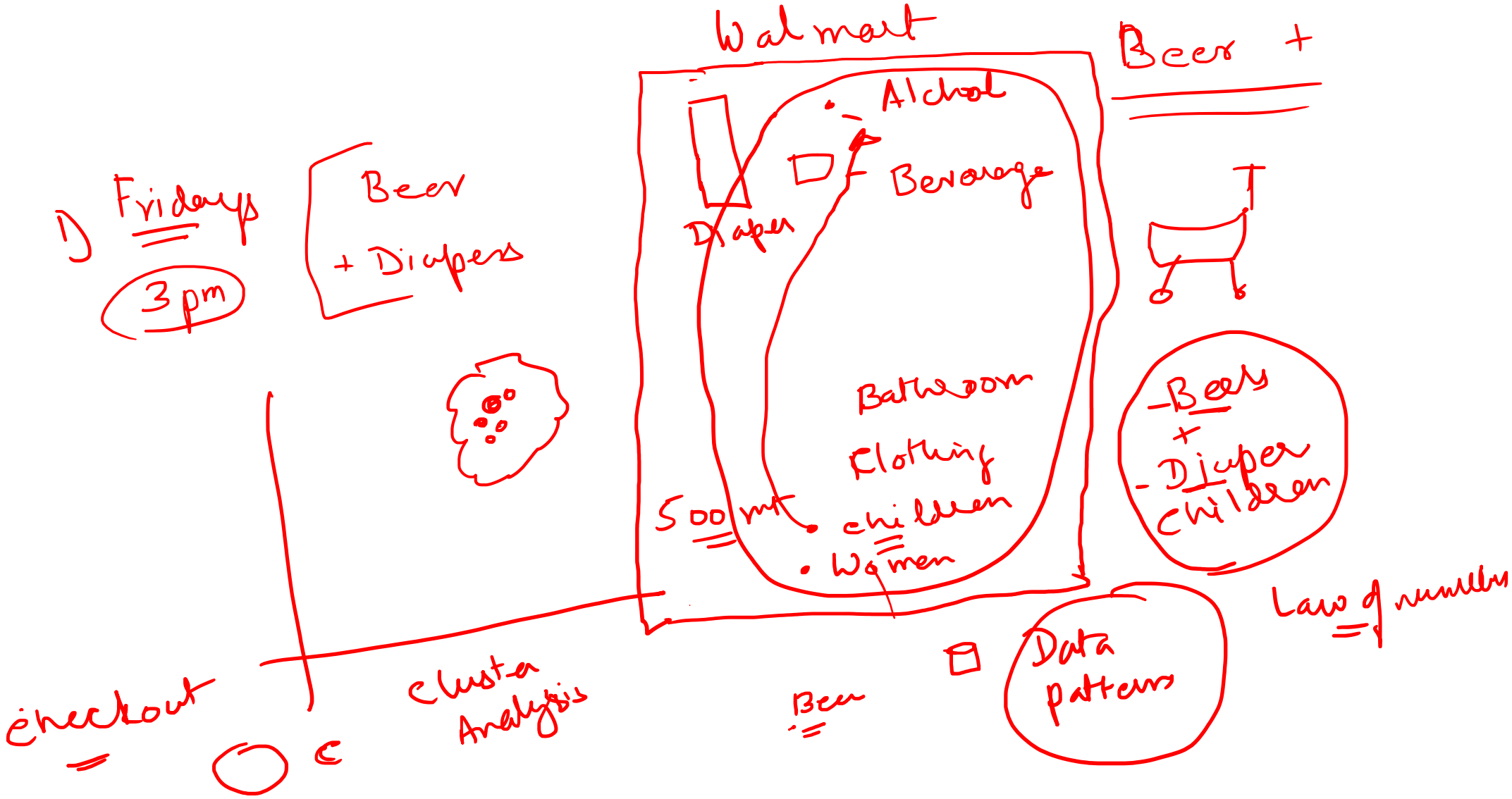

Artificial
Intelligence
Machine
Learning
Data
Science
Big
Data
Deep
Learning


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


Artificial Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)
DATA SCIENCE
Maths &
Statistics
Visualization
EDA

Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...




Walmart
Beer +
Alchol
Beverage
Diaper
Fridays
3pm
Beer
+ Diapers
Bathroom
Clothing
children
Women
500 mt
-Beer
+
-Diaper
children
checkout
cluster Analysis
Beer
Data patterns
Law of numbers

→ Young Couples

Husband is working

1 child — Wife is homemaker

5-day week

week

Mo - Frid → every

Babysitter

Husband

Fri

Sat

Beer

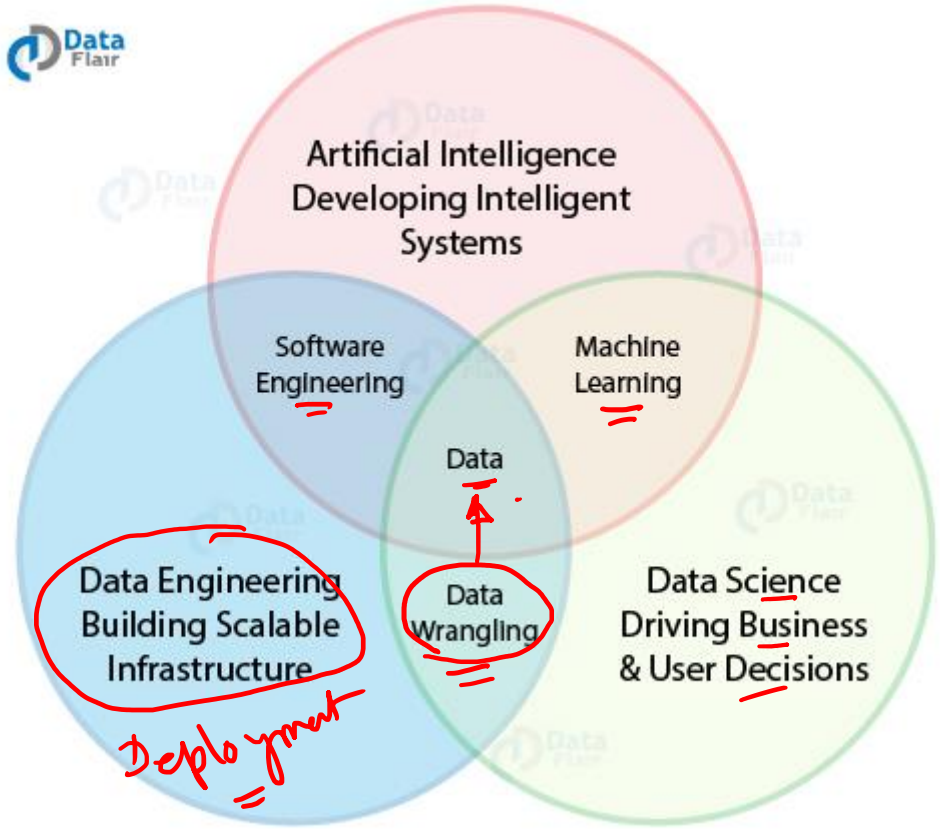
Fri

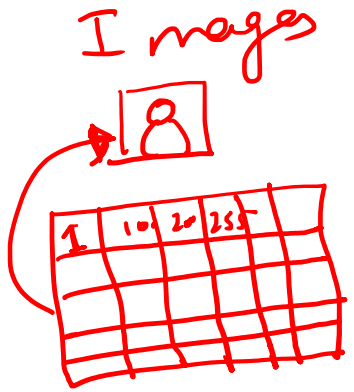
S.

Sa

children

+ Diaper


Data Flair
Artificial Intelligence
Developing Intelligent
Systems
Software
Engineering
Machine
Learning
Data
Data Engineering
Building Scalable
Infrastructure
Data
Wrangling
Data Science
Driving Business
& User Decisions
Deployment


Images
1.
10
20
255

Data Science vs Artificial Intelligence

Factors	Data Science	Artificial Intelligence
Scope	Involves various underlying data operations	Limited to the implementation of ML algorithms
Type of Data	Structured and unstructured	Standardized in the form of embeddings and vectors
Tools	R, Python, SAS, SPSS, TensorFlow, Keras, Scikit-learn	Scikit-learn, Kaffe, PyTorch, TensorFlow, Shogun, Mahout
Applications	Advertising, Marketing, Internet Search Engines	Manufacturing, Automation, Robotics, Transport, Healthcare

Data Science vs Data Analytics

	Data Science	Data Analytics
SKILLSET	• Data Modelling • Predictive Analytics • Advanced Statistics • Engineering/Programming	• BI Tools • Intermediate Statistics • Solid Programming Skills • Regular Expression (SQL)
SCOPE	Macro	Micro
EXPLORATION	• Search Engine Exploration • Machine Learning • Artificial Intelligence • Big data - Often Unstructured	• Data Visualization Techniques • Designing Principles • Big Data - Mostly Structured
GOALS	Discover New Questions to Drive Innovation	Use Existing Information to Uncover Actionable Data


DATA SCIENCE
001011
101110
ANALYSIS
STRUCTURE
ALGORITHM
PROCESS
PROGRAMMING
SOLVING
KNOWLEDGE


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

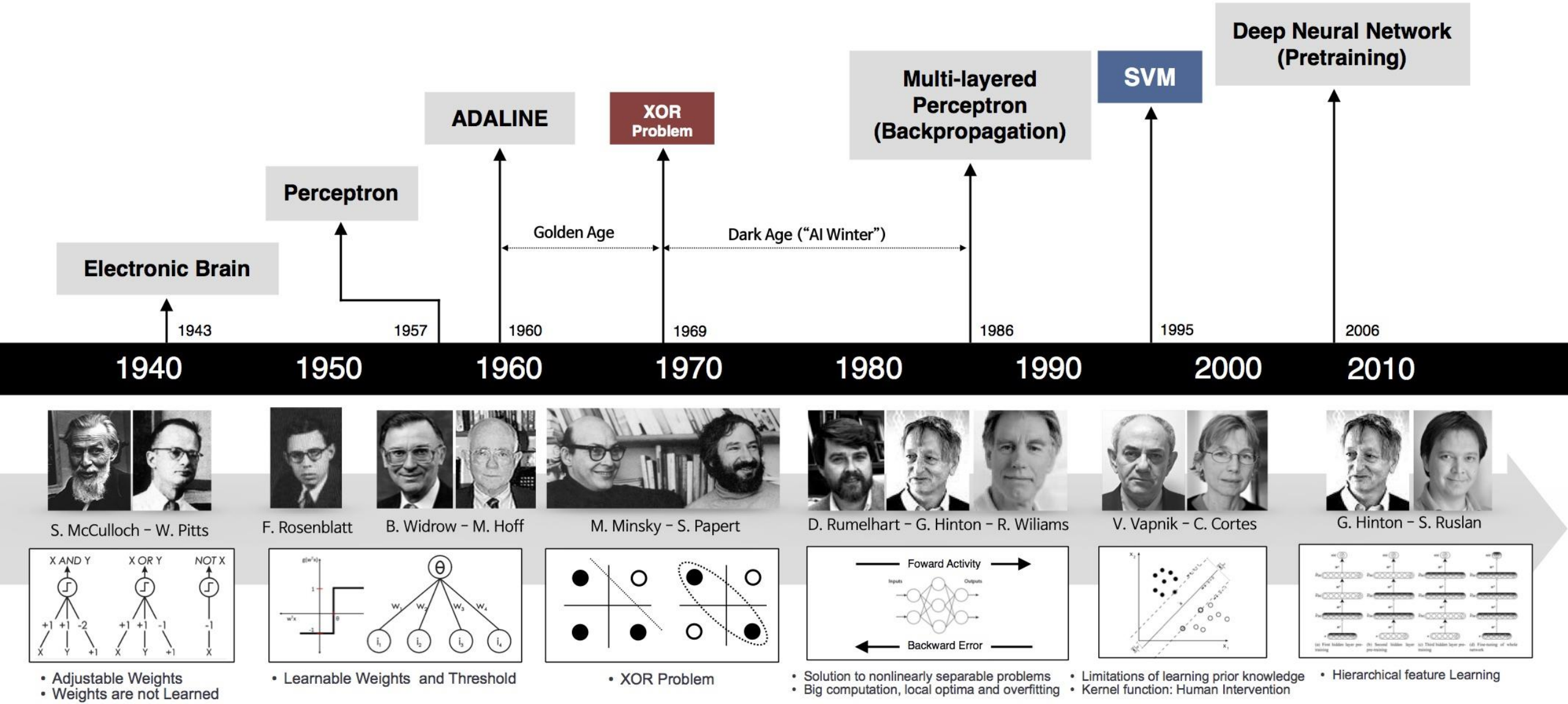
2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

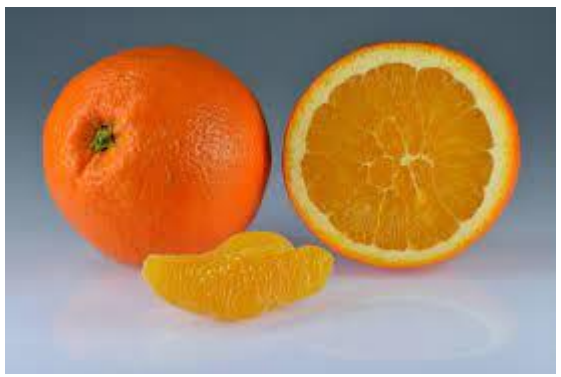
2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

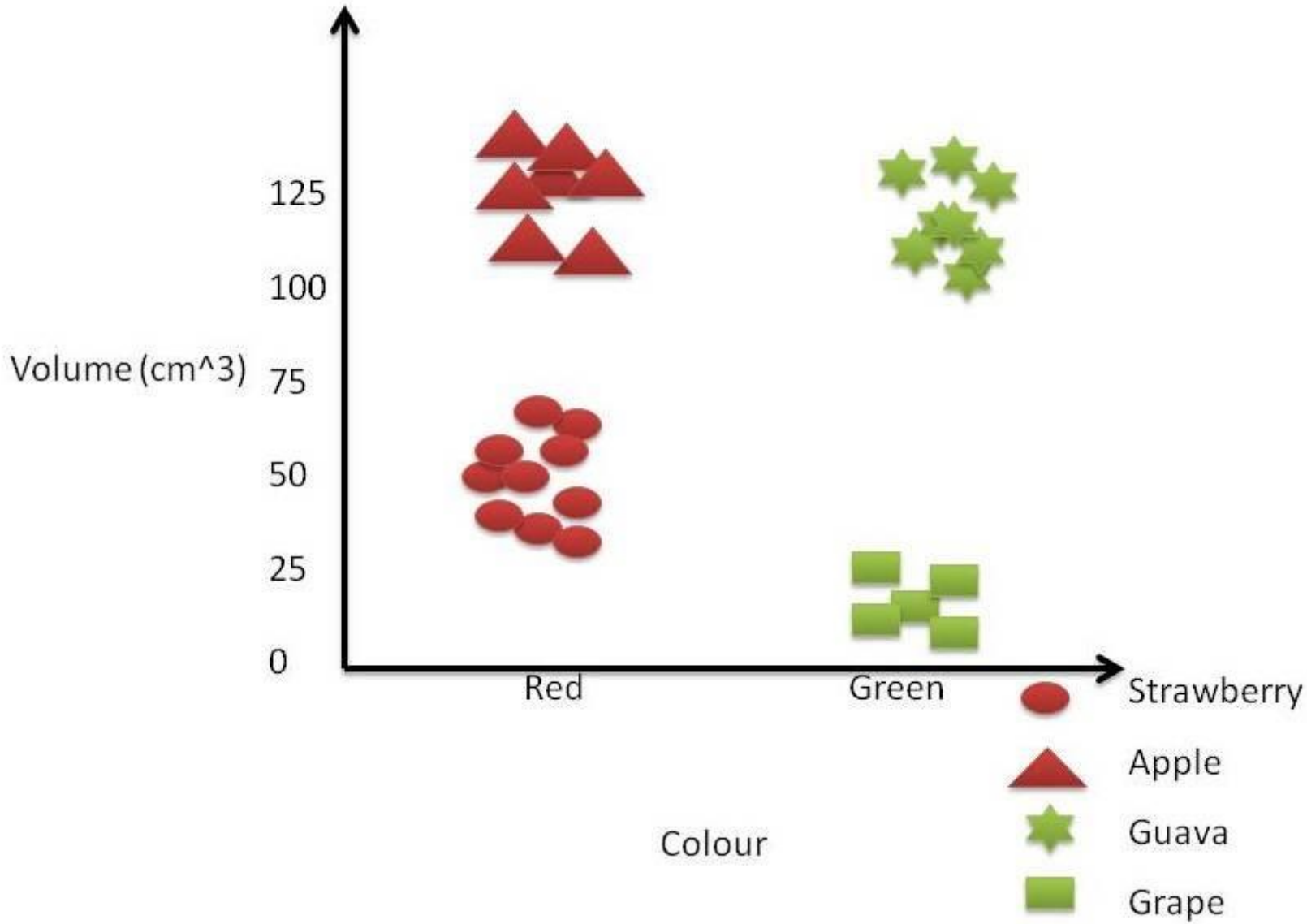
2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

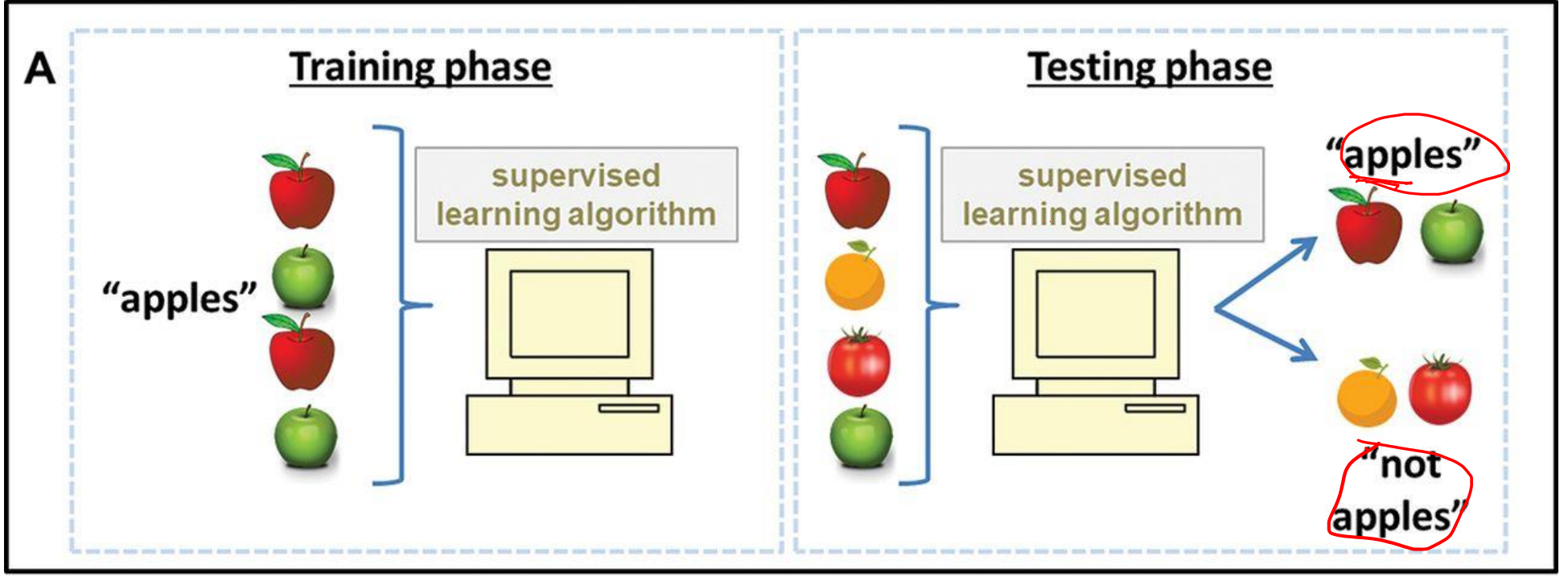
2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

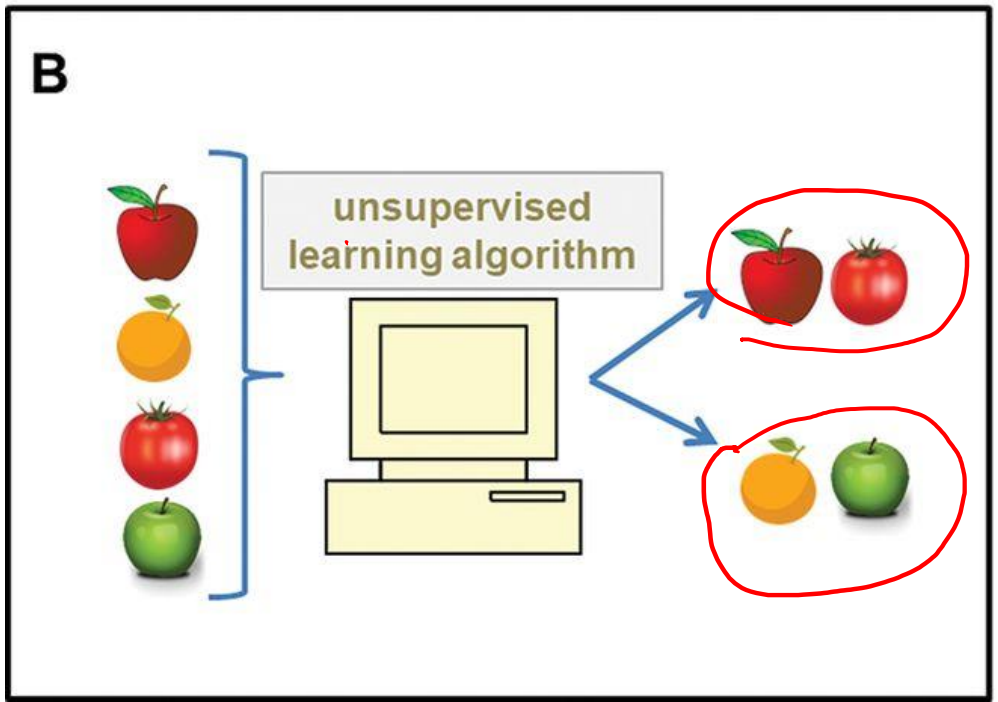
2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

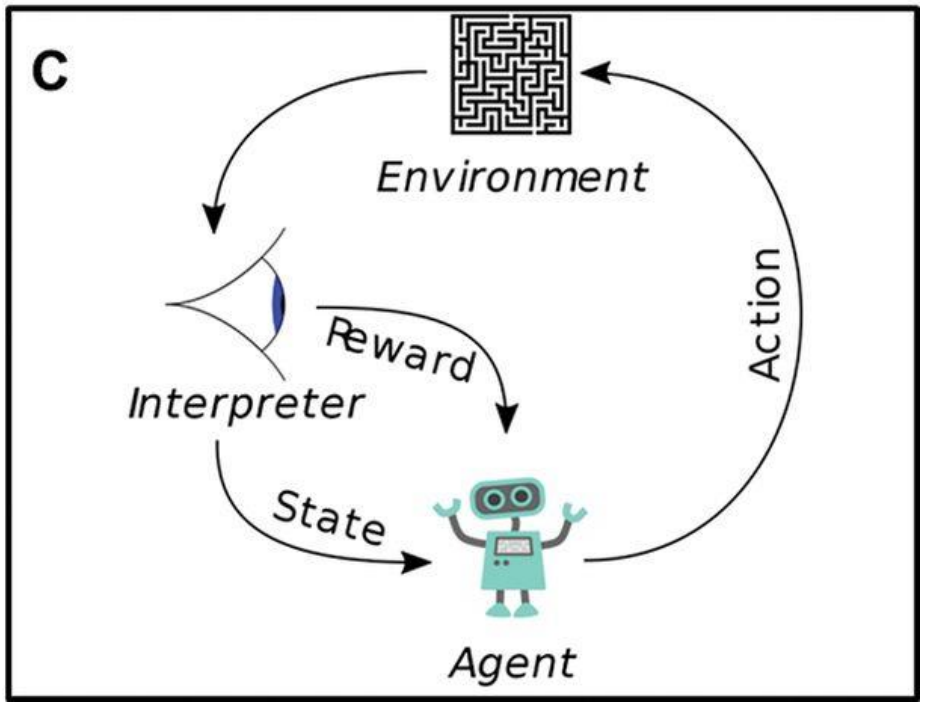

Deep Neural Network (Pretraining)
SVM
Multi-layered Perceptron (Backpropagation)
XOR Problem
ADALINE
Perceptron
Electronic Brain
Golden Age
Dark Age ("AI Winter")
1943
1957
1960
1969
1986
1995
2006
1940
1950
1960
1970
1980
1990
2000
2010
S. McCulloch - W. Pitts
F. Rosenblatt
B. Widrow - M. Hoff
M. Minsky - S. Papert
D. Rumelhart - G. Hinton - R. Wiliams
V. Vapnik - C. Cortes
G. Hinton - S. Ruslan
X AND Y
X OR Y
NOT X
Foward Activity
Backward Error
• Adjustable Weights
• Weights are not Learned
• Learnable Weights and Threshold
• XOR Problem
• Solution to nonlinearly separable problems
• Big computation, local optima and overfitting
• Limitations of learning prior knowledge
• Kernel function: Human Intervention
• Hierarchical feature Learning



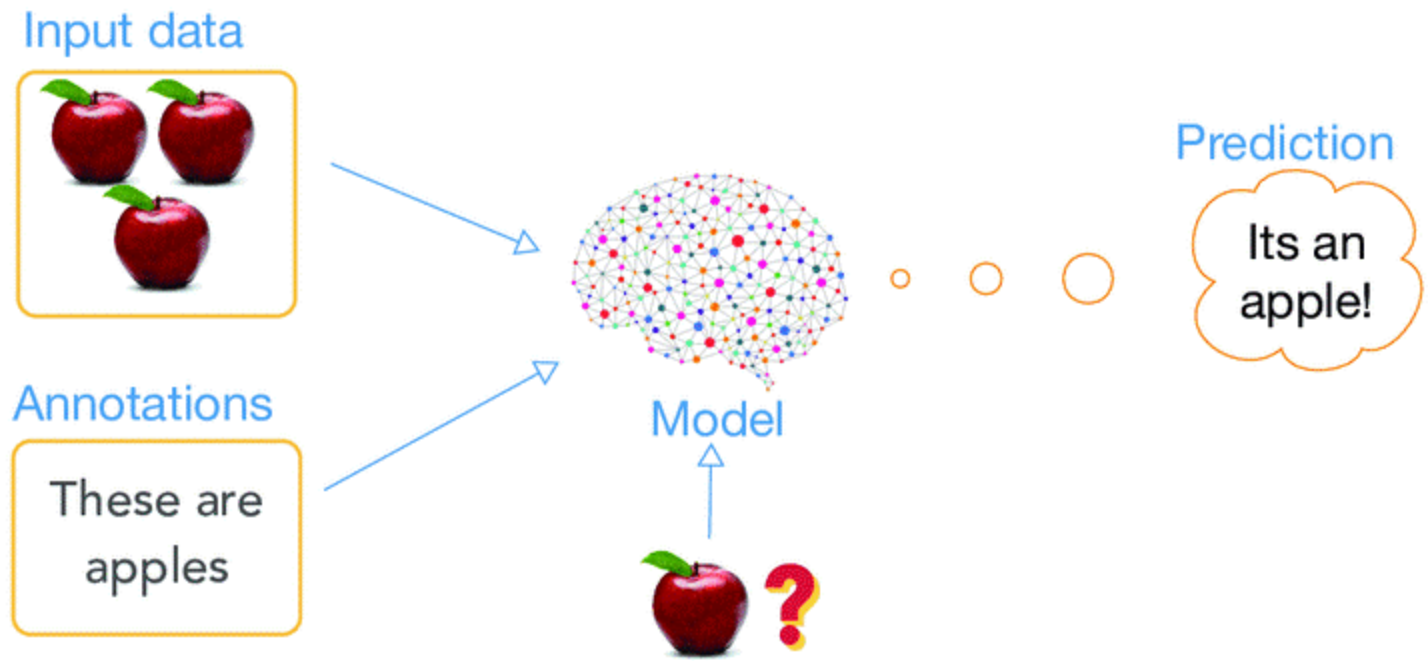

125
100
Volume (cm^3)
75
50
25
0
Red
Green
Colour
Strawberry
Apple
Guava
Grape


A
Training phase
"apples"
supervised learning algorithm
Testing phase
supervised learning algorithm
"apples"
"not apples"

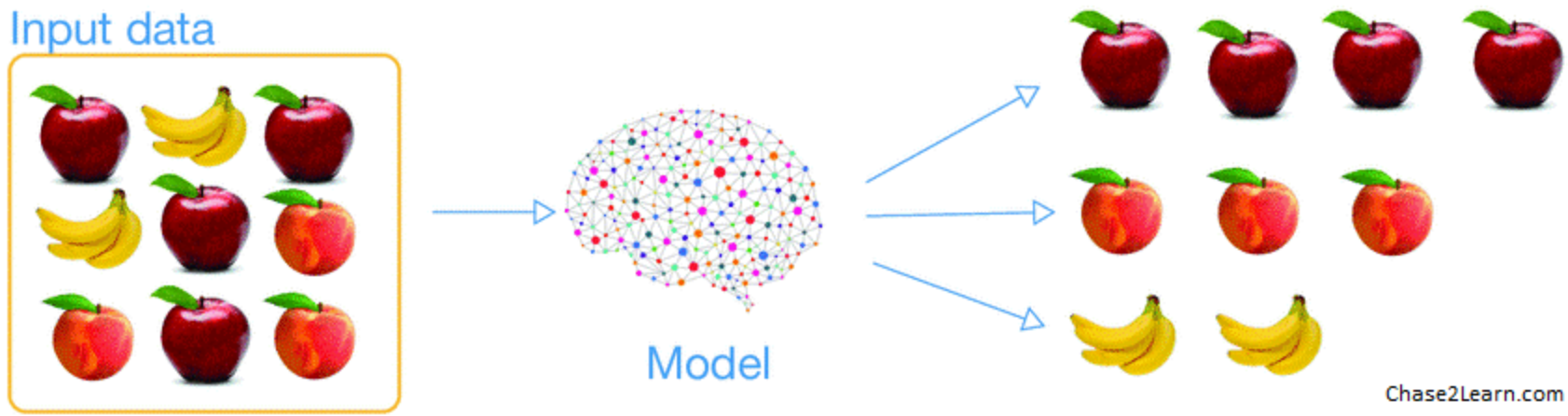

B
unsupervised learning algorithm


C
Environment
Reward
Interpreter
Action
State
Agent

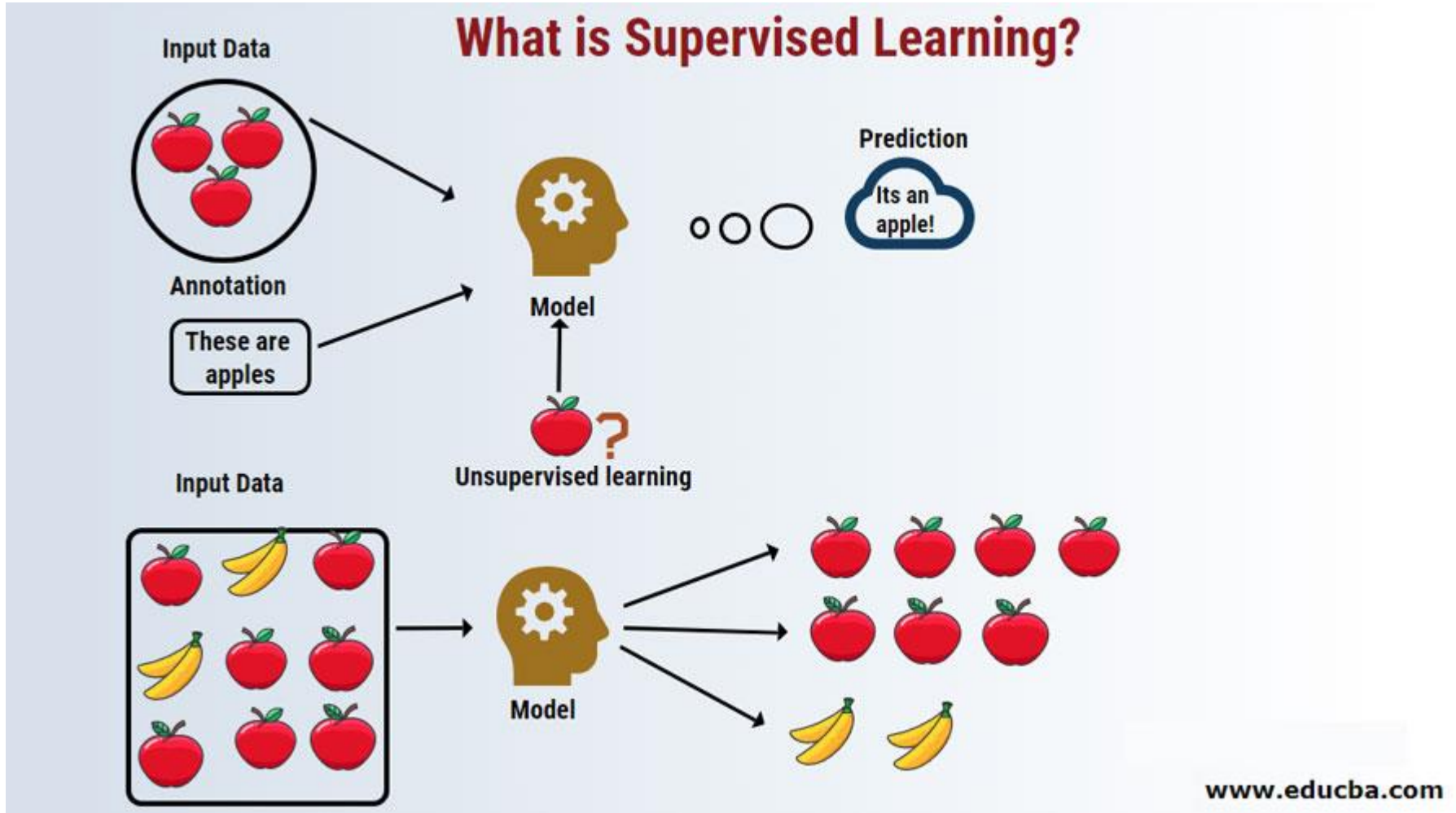
supervised learning



unsupervised learning

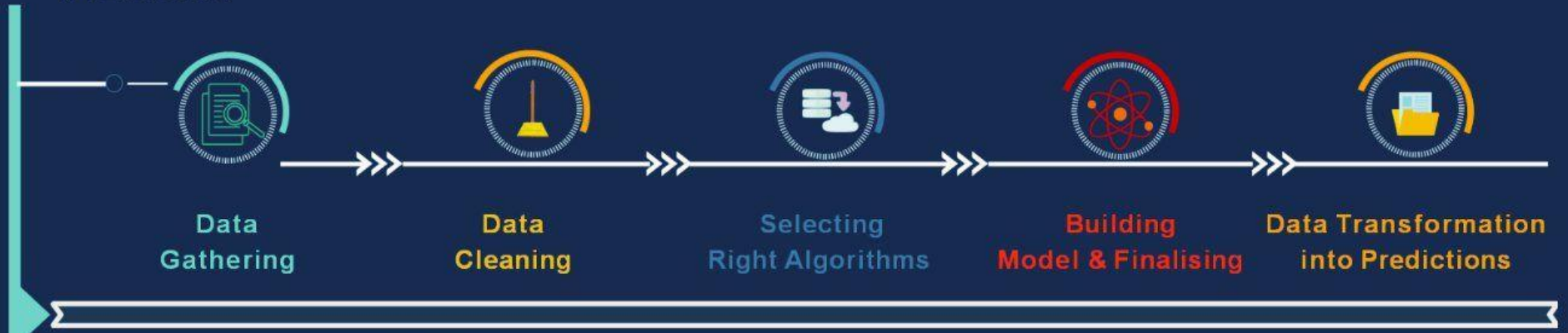


Chase2Learn.com


What is Supervised Learning?
Input Data
Annotation
These are apples
Model
Prediction
Its an apple!
Unsupervised learning
Input Data
Model
www.educba.com

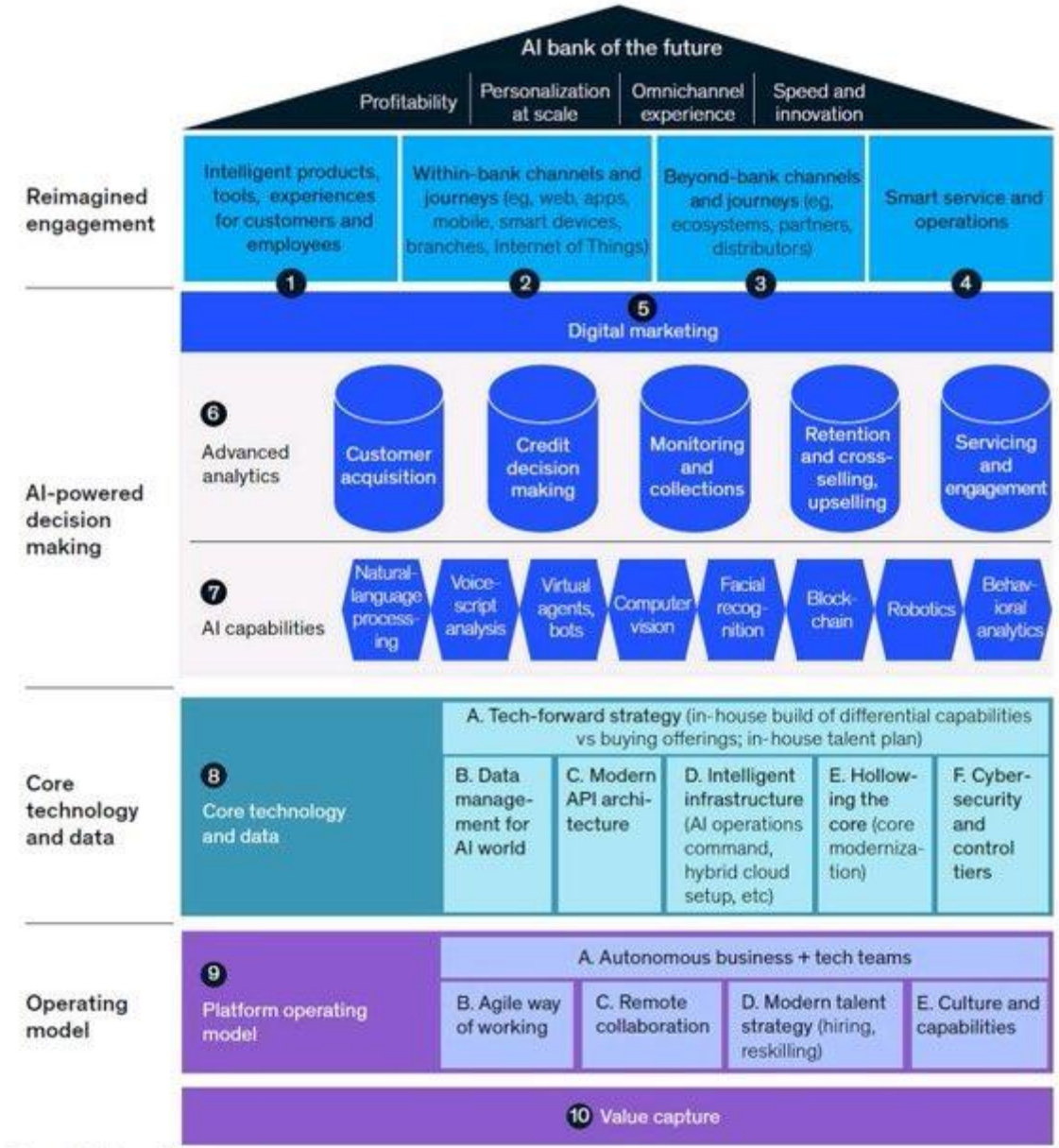
A simple Machine Learning Process

Machine Learning Process, is the first step in ML process to take the data from multiple sources and followed by a fine-tuned process of data, this data would be the feed for ML algorithms based on the problem statement, like predictive, classification and other models which are available in the space of ML world



Follow me @ingliguori

The AI Bank of the future



Source: McKinsey & Company

PYTHON LIBRARIES & FRAMEWORKS

Machine Learning

- Numpy
- Keras
- Theano
- Pandas
- PyTorch
- TensorFlow
- Scikit-Learn
- Matplotlib
- Scipy
- Seaborn

Web Development

- Django
- Flask
- Bottle
- CherryPy
- Pyramid
- Web2Py
- TurboGears
- CubicWeb
- Dash
- Falcon

Automation Testing

- Splinter
- Robot
- Behave
- PyUnit
- PyTest

Game Development

- PyGame
- PyGlet
- PyOpenGl
- Arcade
- Panda3D

Image Processing

- OpenCV
- Mahotas
- SimpleITK
- Pillow
- Scikit-image

Web Scrapping

- Requests
- Beautiful Soup
- Selenium
- Lxml
- Scrapy

LIKE TO SUPPORT

What do you want the machine learning system to do?

I want to see if there are natural clusters or dimensions in the data I have about different situations.

I want to learn what actions to take in different situations.

Do you want the ML system to be active or passive?

ACTIVE

The system's own actions will affect the situations it sees in the future.

PASSIVE

The system will learn from data I give it.

Do you have access to data that describes a lot of examples of situations and appropriate actions for each situation?

Yes

No

Could a knowledgeable human decide what actions to take based on the data you have about the situation?

Yes

No

Could there be patterns in these situations that humans haven't recognized before?

Yes

No

Will the system be able to gather a lot of data by trying sequences of actions in many different situations and seeing the results?

Yes

No

UNSUPERVISED LEARNING MAY BE APPROPRIATE

clustering
anomaly detection

SUPERVISED LEARNING MAY BE APPROPRIATE

neural nets
support vector machines
regression
recommender systems

MACHINE LEARNING IS NOT USEFUL

REINFORCEMENT LEARNING MAY BE APPROPRIATE

Credit: Thomas Malone, MIT Sloan | Design: Laura Wentzel

Major Applications of Machine Learning in Cybersecurity

1 WHERE IS ML APPLICABLE?

- Where we have lots of data either on the cloud or on the endpoint, IoT- IIoT, working on combination with big data and analytics
- To identify anomalies, suspicious or unusual behaviour
- Detect and correct known vulnerabilities and zero-day attacks
- When computer or machine time versus human time is a major requirement

2 INCIDENT RESPONSE & FORENSICS

- In the unfortunate case of an attack, an automated response in critical in order to minimize the impact, conduct forensics and to defend effectively
- From a defensive perspective we need to be able to respond in computer or machine time versus human time to stop some of the attacks
- Defense against intelligent cyber weapons can only be achieved by intelligent software
- The accuracy and effectiveness of the response to an attack could also be improved leveraging ML which is also quite import considering that cybersecurity has quite low fault tolerance as it only takes one vulnerability to be exploited in order to have a data breach

3 THREAT EXAMPLES

Specific threats that could be addressed with ML:

- Spear Phishing
- Ransomware
- DDoS
- Watering Hole
- Webshell
- DNS Poisoning
- Port Scanning
- Defense against intelligent cyber weapons

4 FRAUD DETECTION

- Machine Learning (ML) is increasingly being introduced to fight e-commerce fraudsters
- There is currently access to lots of information about suspect fraudsters, including their purchase activities and profile, online browsing activities, social networks and fake identification they submit to get tier orders approved
- The challenge is how we can make sense of this unstructured data and then make good approve / decline decisions for thousands of merchants in real-time

5 ENHANCE HUMAN ANALYSIS

- ML might help to address the acute problem of scarce and expensive expertise through resource optimization or increase in staff productivity
- Also a substantial reduction in false positive rates would positively impact cybersecurity operations and ML is very effective in achieving this goal
- We need to be cognizant that the widening cyber-security skills gap is seriously threatening companies and this serious issue needs to be addressed in terms of cyber risk

exploited in order to have a data breach