

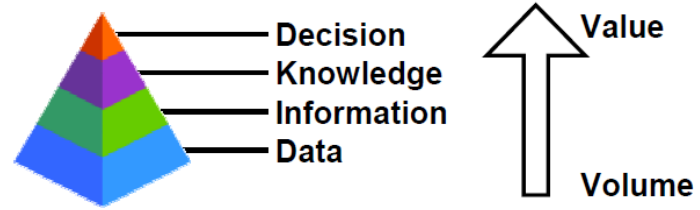
# What is Data Science

- **Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.**

- The inventor of the World Wide Web, Tim Berners-Lee, is often quoted as having said ,  
“Data is not information, information is not knowledge, knowledge is not understanding,  
understanding is not wisdom”
- This quotes a kind of Pyramid, where data and raw materials that make up the foundation at the bottom of the pile, and information, knowledge, understanding and wisdom represent higher and higher levels of the pyramid
- The major goal of a data scientist is to help people to turn data into information and onwards up the pyramid
- Data science is different from other areas such as mathematics of statistics. Data science is an applied activity and data scientists serve the needs and solve the problems of data users
- Before you can solve problem , you need to identify it and this process is not always as obvious as it might seem

# Business intelligence

- “Business intelligence is the process of transforming data into
- information and through discovery transforming that information into knowledge.”



# What Is Business Intelligence?

- How are sales year-to-date and How do they compare to last year?
- Who is most likely to respond to me current marketing campaign and how will they impact revenue?
- What is the turnover in employees compared to the last five years?
- How is potential fraud cost being managed over time?
- What are my most profitable products by region, by year, and year-to-date?

This is a simple business question, but the actual query can be quite complex

“What was the percentage change in revenue for a grouping of our top 20% products from one year ago over a rolling three-month time, period compared to this year for each region of the world?”

Business users typically want to answer questions that include terms such as *what*, *where*, *who*, and *when*.

For example, you find the following essential questions embedded in the sample question:

- *What products are selling best?* (“...top 20%...”)
- *Where are they selling?* (“...each region of the world...”)
- *When have they performed the best?* (“...percentage change in revenue...”)

Few scenarios where data mining can be helpful:

A retailer wants to increase revenues by identifying all potentially high-value customers to offer incentives to them.

The retailer also wants guidance in store layout by determining the products most likely to be purchased together.

A government agency wants faster and more accurate methods of highlighting possible fraudulent activity for further investigation, and so on.

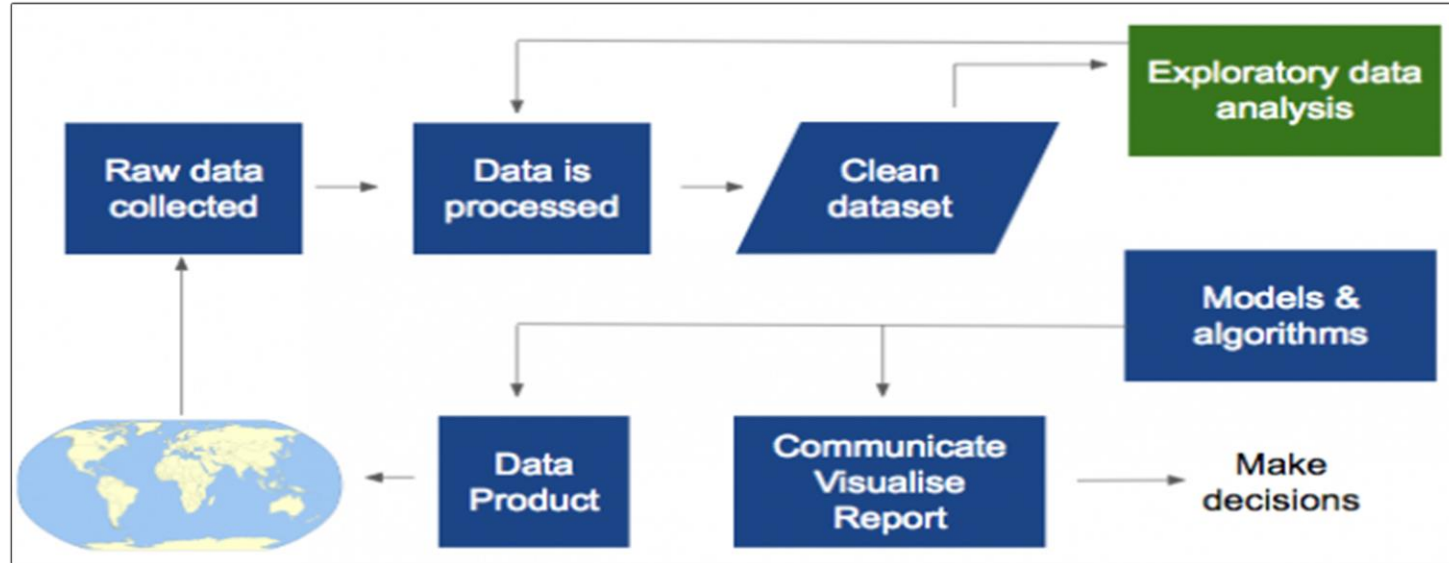
# Change Is the Motivation



- **Data was and has been always critical to organizations.**
- Data creating business value is not a new idea, however with times changing, volumes – variety –velocity increasing it has gained more importance.
- Change in approach is inevitable and probably much needed now than before.
- Success battles were fought on basis of data and its effective use is becoming the core basis of **competition & success.**



# Data Science Process



## **Most real-life projects that involve data can be broken down into several steps:**

1. Data Acquisition - we need to find (or collect) the data, and get some representation of it into the computer
2. Data Cleaning - Inevitably, there will be errors in the data, either because they were entered incorrectly, we misunderstood the nature of the data, records were duplicated or omitted. Many times, data is presented for viewing, and extracting the data in some other form becomes a challenge.
3. Data Organization - Depending on what you want to do, you may need to reorganize your data. This is especially true when you need to produce graphical representations of the data. Naturally, we need the appropriate tools to do these tasks.
4. Data Modelling and Presentation - We may fit a statistical model to our data, or we may just produce a graph that shows what we think is important. Often, a variety of models or graphs needs to be considered. It's important to know what techniques are available and whether they are accepted within a particular user community.

- Overview of Data Science

Data acquisition, profiling, preparation, and visualization.

Feature engineering.

Model training

Model evaluation, explanation, and interpretation

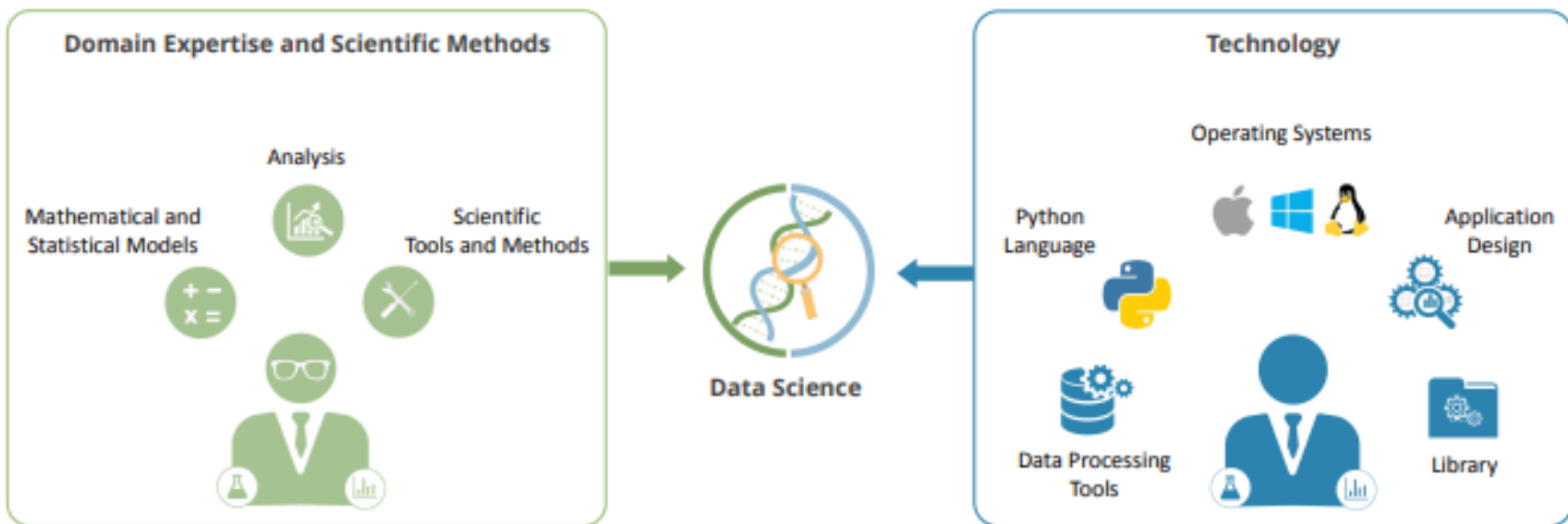
Model deployment.

# **Who is a Data Scientist**

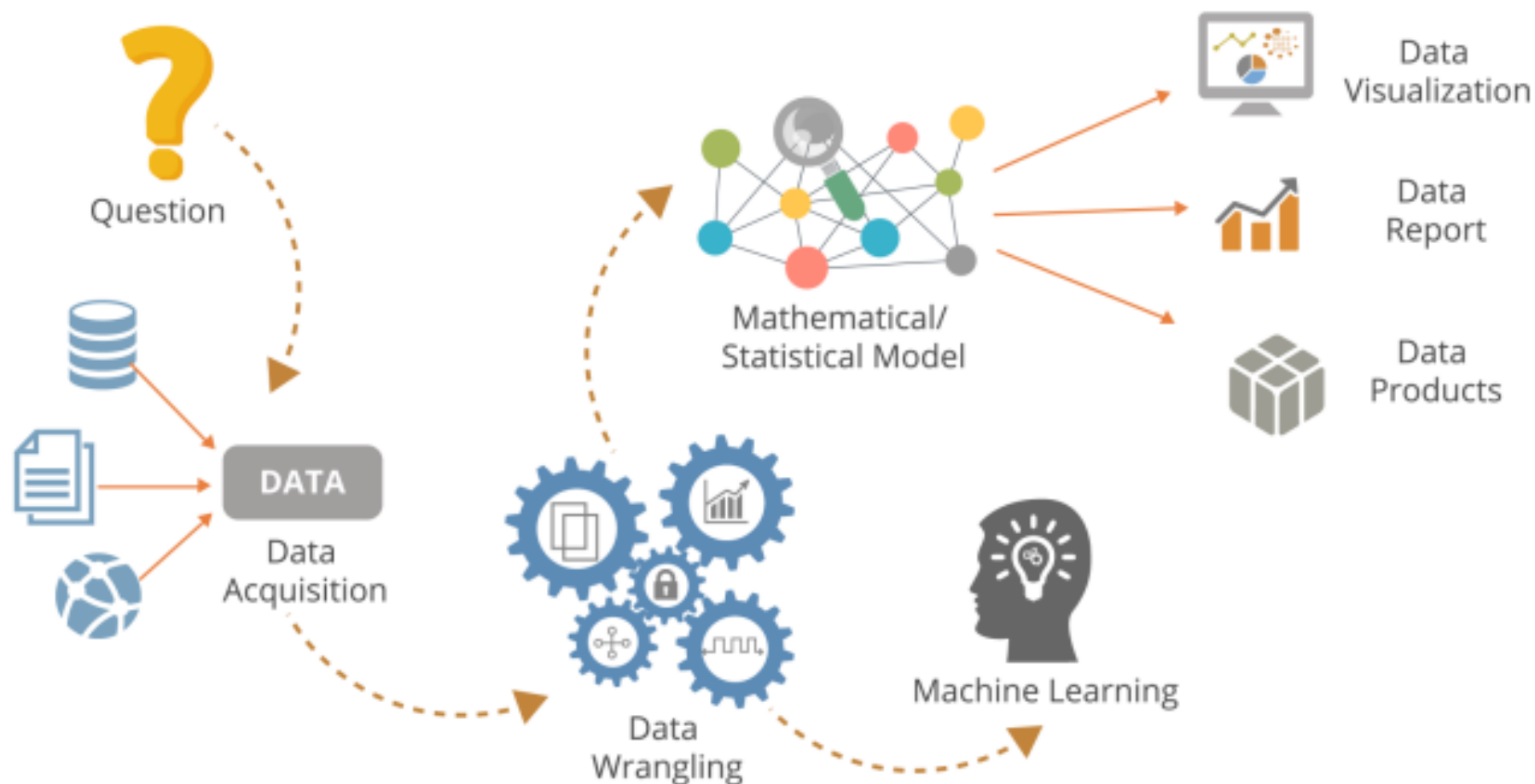
Person who is better at statistics than any software engineer and better at software engineering than any statistician.

# The Components of Data Science

When we combine domain expertise and scientific methods with technology, we get Data Science.



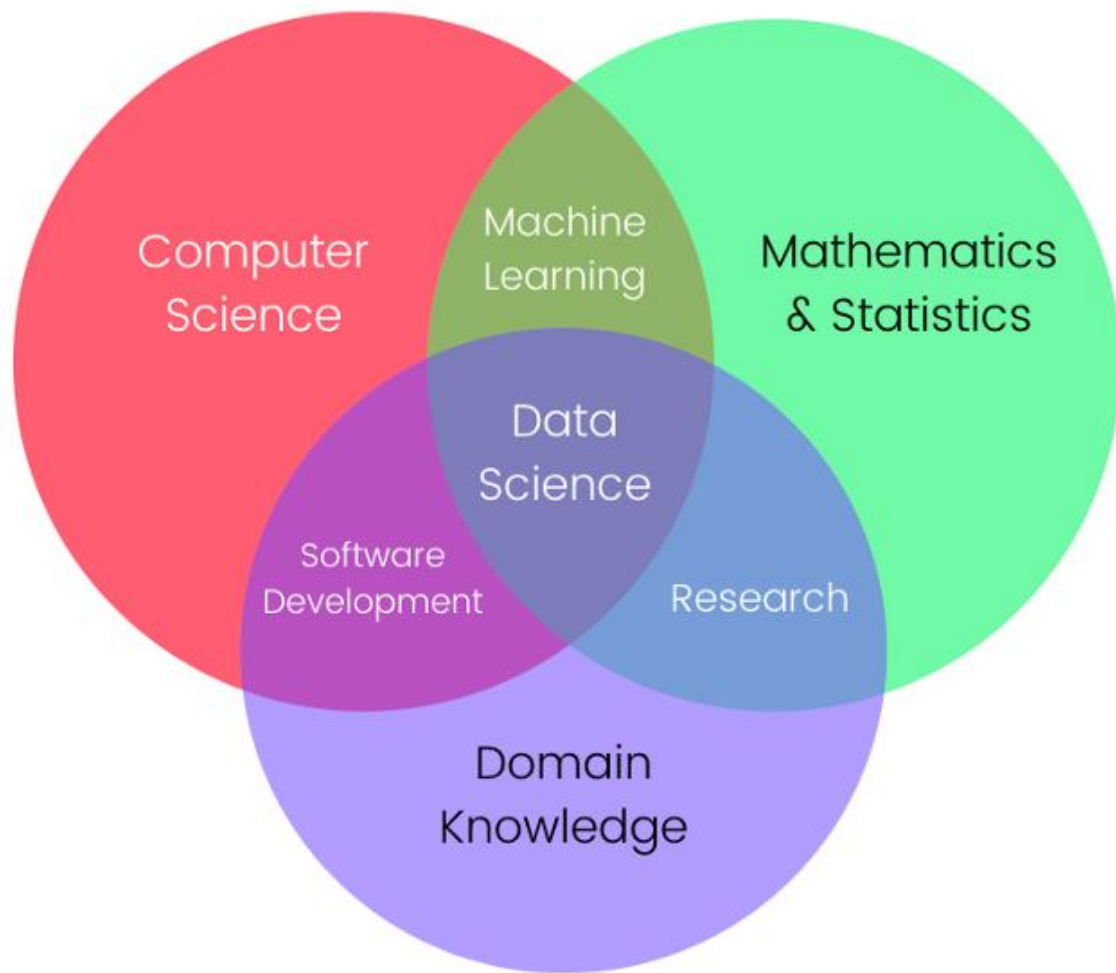
# A Day in a Data Scientist's Life



## Different Sectors Using Data Science

Various sectors use Data Science to extract the information they need to create different services and products.







# Technologies Used In Data Science



**Big Data**

**Data  
Cleansing**


**Web  
Scraping**

**Machine  
Learning**

**Data  
Visualization**

# Technologies Used In Data Science

## Big Data



Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

Big Data Analytics (or data science) can be studied at three levels.



```
graph TD; A[Big Data Analytics (or data science) can be studied at three levels.] --> B[A researcher level that focuses on the underlying mathematics and computing deeply]; B --> C[A business level that focuses on interpretation and business applications]; C --> D[An engineering level where the focus is on building working systems with known knowledge.]; D --> E[Think analytically, rigorously and systematically about a business problem and come up with a solution that leverages the available data];
```

A researcher level that focuses on the underlying mathematics and computing deeply

A business level that focuses on interpretation and business applications

An engineering level where the focus is on building working systems with known knowledge.

Think analytically, rigorously and systematically about a business problem and come up with a solution that leverages the available data



- Predictive maintenance or condition monitoring
- Warranty reserve estimation
- Propensity to buy
- Demand forecasting
- Process optimization
- Telematics

## Manufacturing



- Predictive inventory planning
- Recommendation engines
- Upsell and cross-channel marketing
- Market segmentation and targeting
- Customer ROI and lifetime value

## Retail



- Alerts and diagnostics from real-time patient data
- Disease identification and risk stratification
- Patient triage optimization
- Proactive health management
- Healthcare provider sentiment analysis

## Healthcare and Life Sciences



- Aircraft scheduling
- Dynamic pricing
- Social media – consumer feedback and interaction analysis
- Customer complaint resolution
- Traffic patterns and congestion management

## Travel and Hospitality



- Risk analytics and regulation
- Customer Segmentation
- Cross-selling and up-selling
- Sales and marketing campaign management
- Credit worthiness evaluation

## Financial Services



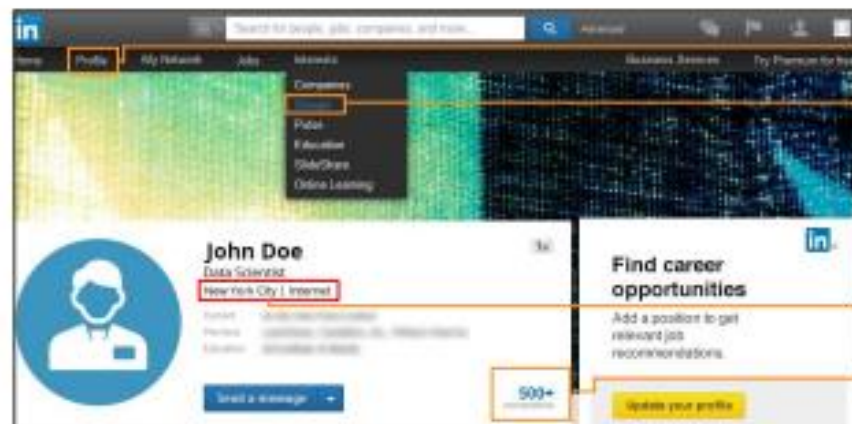
- Power usage analytics
- Seismic data processing
- Carbon emissions and trading
- Customer-specific pricing
- Smart grid management
- Energy demand and supply optimization

## Energy, Feedstock, and Utilities



# Using Data Science—Social Network Platforms

LinkedIn uses data points from its users to provide them with relevant digital services and data products.



Data Points



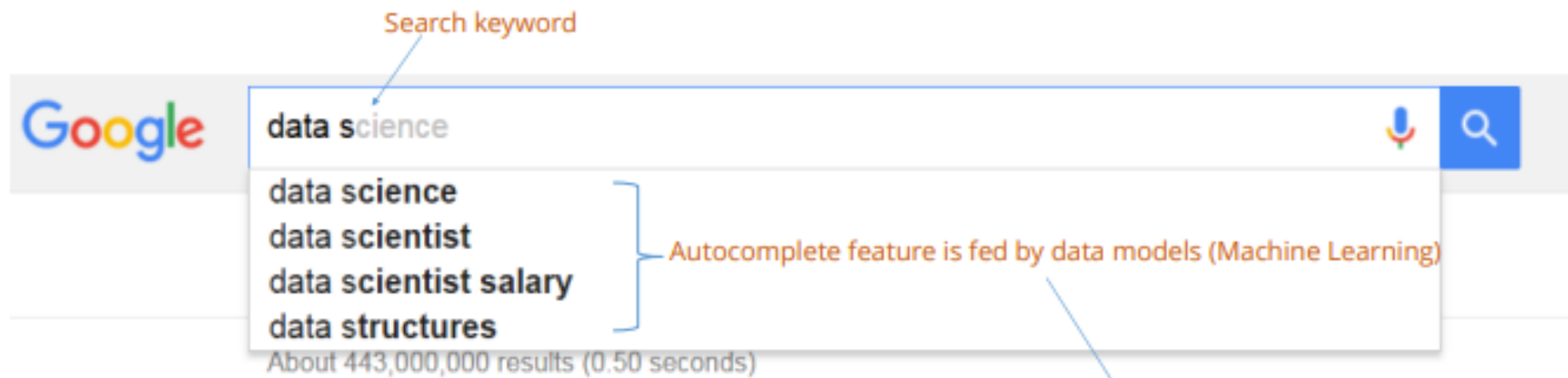
Information

Digital Services

Data Products

## Using Data Science—Search Engines

Google uses Data Science to provide relevant search recommendations as the user types a query.



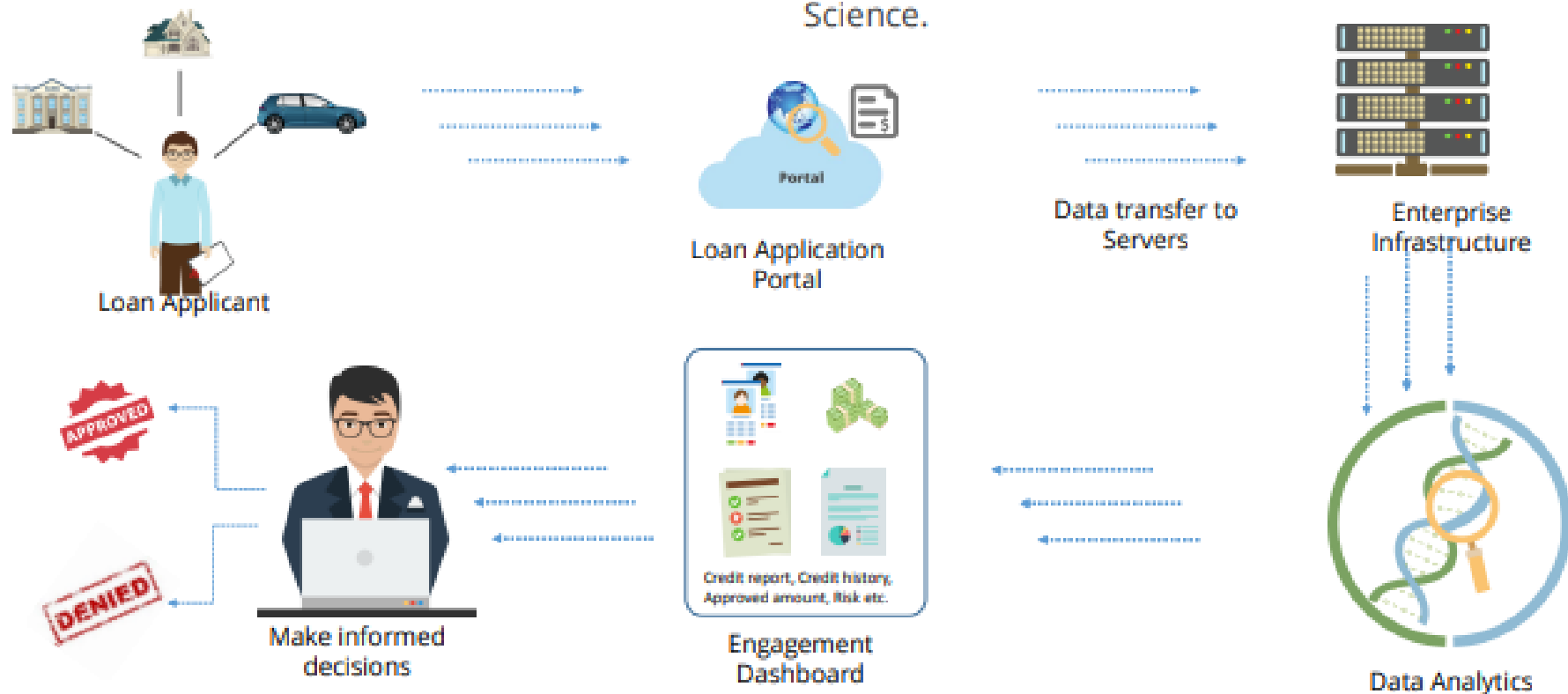
Fast and real-time analytics is made possible by modern and advanced infrastructure, tools, and technologies

### Influencing Factors

1. Query Volume – Unique and verifiable users
2. Geographical locations
3. Keyword/phrase matches on the web
4. Some scrubbing for inappropriate content

# Using Data Science—Finance

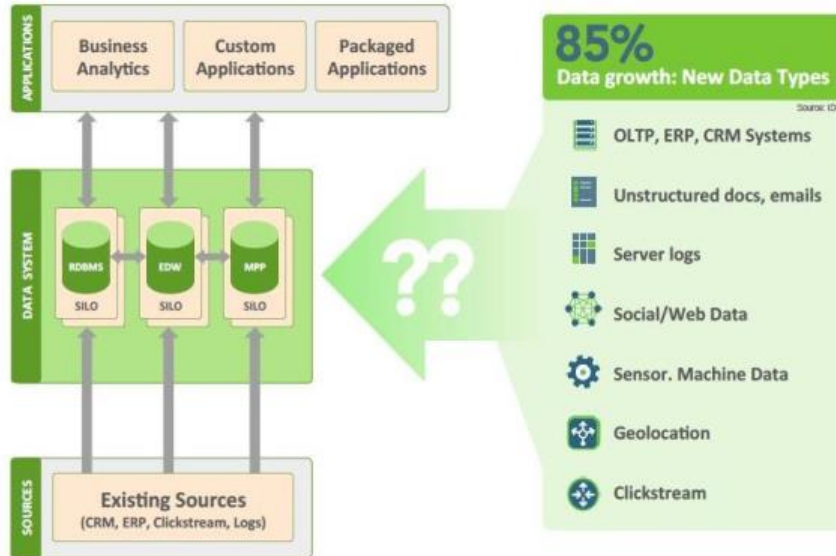
A Loan Manager can easily access and sift through a loan applicant's financial details using Data Science.



# Traditional Data Architecture

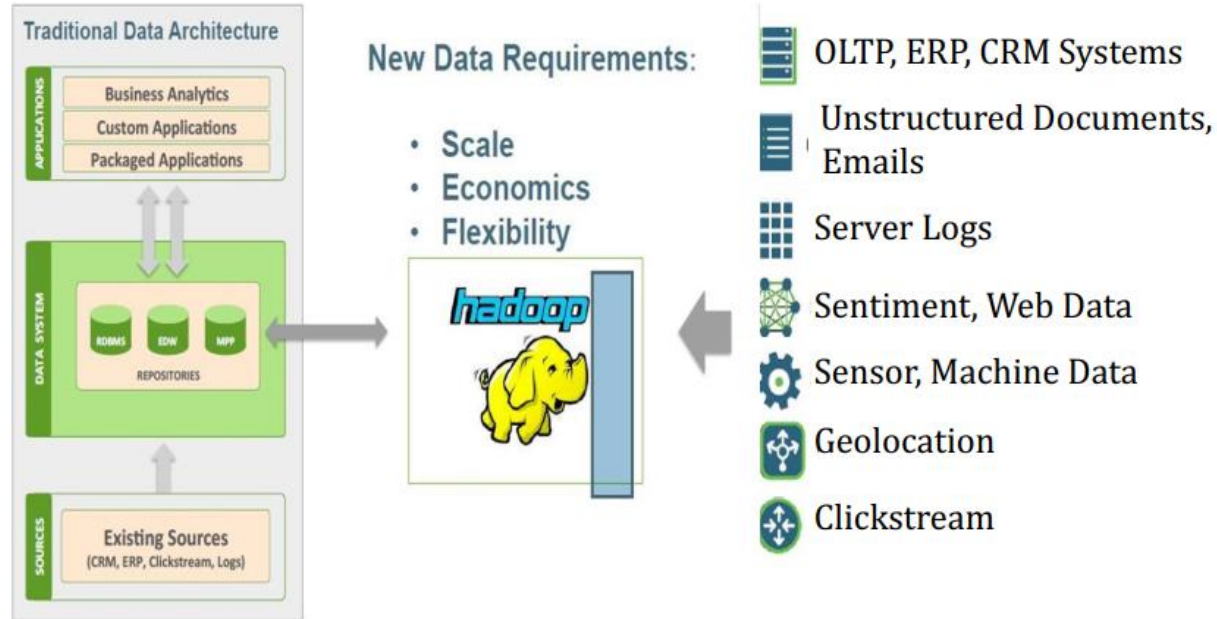
Disadvantages of traditional data architecture under pressure:

- Can't manage new data paradigm
- Constrains data to specific schema
- Siloed data
- Limited scalability
- Economically unfeasible
- Limited analytics





# Modern Data Architecture



# Technologies Used In Data Science

**Big Data  
Technologies**

**Apache  
Hadoop.**

Microsoft  
HDInsight.

**NoSQL**

Hive

Sqoop

PolyBase

Big data in  
EXCEL

Presto.

# Big Data Use-Cases

- **Web and e-tailing**

- Recommendation Engines
- Ad Targeting
- Search Quality
- Abuse and Click Fraud Detection



- **Telecommunications**

- Customer Churn Prevention
- Network Performance Optimization
- Calling Data Record (CDR) Analysis
- Analysing Network to Predict Failure



- **Government**

- Fraud Detection and Cyber Security
- Welfare Schemes
- Justice



- **Healthcare and Life Sciences**

- Health Information Exchange
- Gene Sequencing
- Serialization
- Healthcare Service Quality Improvements
- Drug Safety



# Big Data Use-Cases

- **Banks and Financial services**

- Modeling True Risk
- Threat Analysis
- Fraud Detection
- Trade Surveillance
- Credit Scoring and Analysis



- **Retail**

- Point of Sales Transaction Analysis
- Customer Churn Analysis
- Sentiment Analysis



- **Transportation Services**

- Data from Location based social network
- High speed data from telecom
- Transport demand models
- Route Planning



- **Hotels and Food Delivery Services**

- Customer Demands
- Details of Customers
- Availability and Seasonal Data Changes



# Technologies Used In Data Science

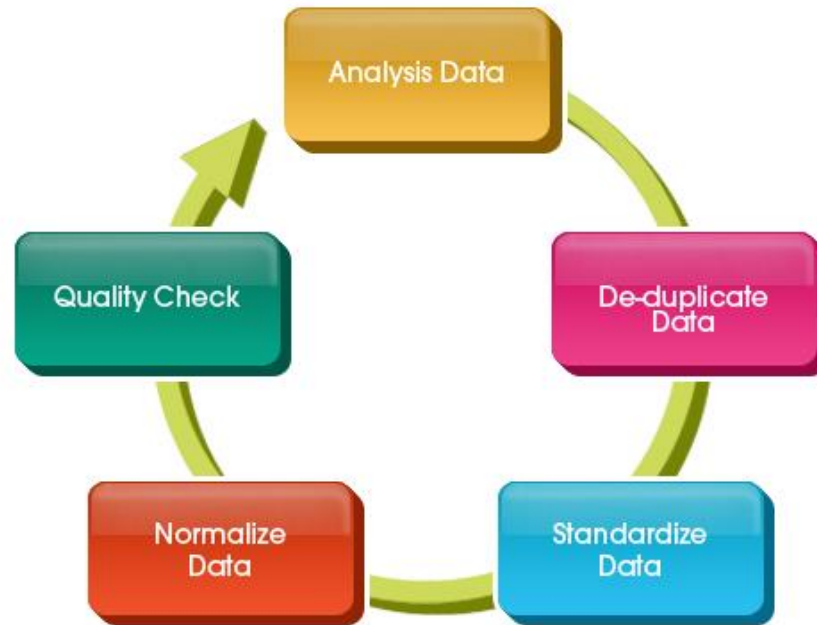
## Data Cleansing

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

# Data Cleaning Techniques

- Data cleaning techniques are used to correct, transform, and organize data to improve its quality and accuracy. Here are some of the most common data-cleaning techniques:
- Data Normalization: Normalization is the process of transforming data into a standard format, making it easier to process and clean.
- Data Transformation: Data transformation is the process of converting data from one format to another, making it easier to use and analyze.
- Data Integration: Data integration is the process of combining data from multiple sources into a single, consistent format.
- Data Reduction: Data reduction is the process of removing unnecessary data, such as duplicates or irrelevant information, to simplify and improve data quality.
- Data Imputation: Data imputation is the process of filling in missing data with estimates or values derived from other data.
- Data Deduplication: Data deduplication is the process of removing duplicate data entries to ensure data accuracy and consistency.
- Data Enrichment: Data enrichment is the process of adding additional information to data, such as geolocation data or demographic information, to enhance its value.

# Data Cleansing



# Technologies Used In Data Science



## Web Scrapping



**Web Scraping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer .**



- Web scraping is all about collecting content from websites. Scrapers come in many shapes and forms and the exact details of what a scraper will collect will vary greatly, depending on the use cases.
- A very common example is search engines, of course. They continuously crawl and scrape the web for new and updated content, to include in their search index.

# Technologies Used In Data Science

ScrapingBee

ScrapeBox

ScreamingFrog

Scrapy

pyspider

Beautiful Soup

Diffbot

Common Crawl

# Technologies Used In Data Science

## **Data Visualization**

**Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.**

# Technologies Used In Data Science

Machine Learning Software's

Tensor Flow

PyTorch

Scikit-Learn

Keras

XGBoost

Apache Spark Mllib

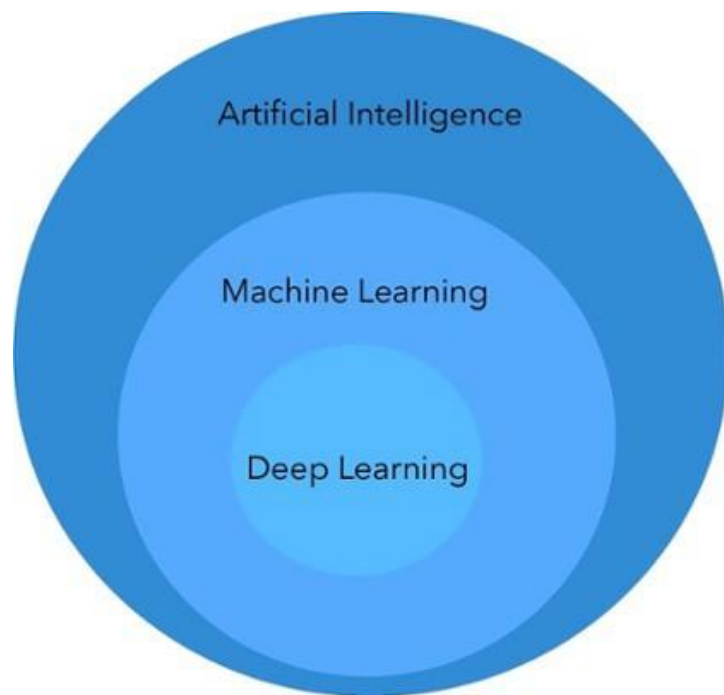
Microsoft Azure Machine Learning

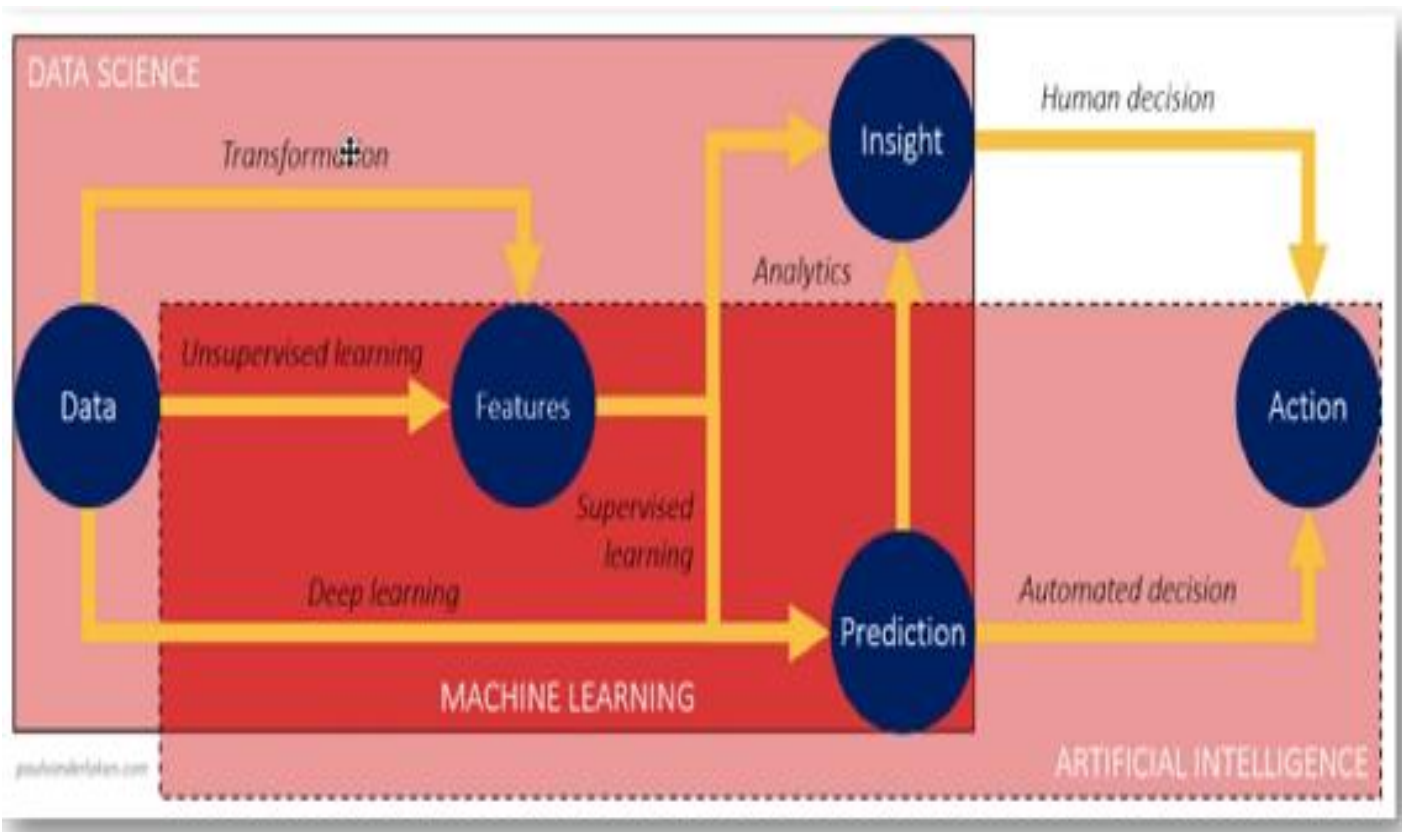
RapidMiner

## Technologies Used In Data Science

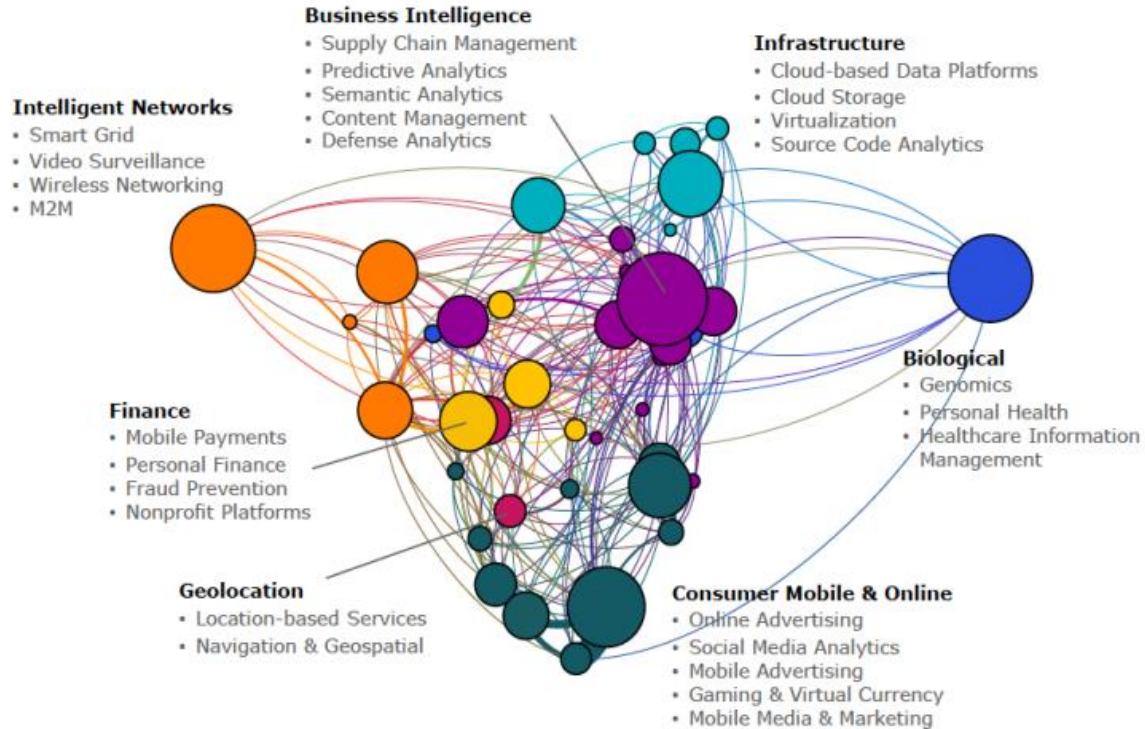
### **Machine Learning**

**Machine learning is the subfield of computer science that, according to Arthur Samuel, gives "computers the ability to learn without being explicitly programmed." Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "machine learning" in 1959 while at IBM.**





# Applications Of Datascience





# Top Data Science Trends For 2024

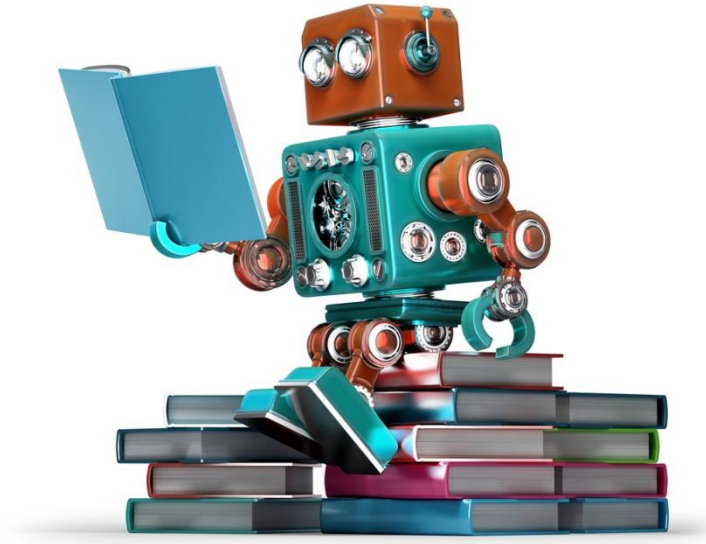
- Augmented Analytics
- Responsible AI
- Edge Computing for Data Science
- Quantum Computing Integration
- Continuous Learning Models
- Natural Language Processing (NLP) Advancements
- Federated Learning
- Blockchain in Data Science

# Introduction to Machine Learning

# What is Machine Learning ?

---

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed



# What is Machine Learning?

---

The ability to perform a task in a situation which has never been encountered before (Learning = Generalization)

Automating automation

Getting computers to program themselves

Writing software is the bottleneck

Let the data do the work instead!

# A short story

Samuel's claim to fame was that back in the 1950, he wrote a checkers playing program and the amazing thing about this checkers playing program was that Arthur Samuel himself wasn't a very good checkers player.

But what he did was he had to programmed maybe tens of thousands of games against himself, and by watching what sorts of board positions tended to lead to wins and what sort of board positions tended to lead to losses, the checkers playing program learned over time what are good board positions and what are bad board positions.

And eventually learn to play checkers better than the Arthur Samuel himself was able to.

This was a remarkable result.

Arthur Samuel himself turns out not to be a very good checkers player.

But because a computer has the patience to play tens of thousands of games against itself, no human has the patience to play that many games.

By doing this, a computer was able to get so much checkers playing experience that it eventually became a better checkers player than Arthur himself.

# What is Machine Learning

- “A computer program is said to learn from experience  $E$  with some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .” -Tom M. Mitchell Consider **playing checkers**.
- $E$  = the experience of playing many games of checkers
- $T$  = the task of playing checkers.
- $P$  = the probability that the program will win the next game

- classification problems where the goal is to categorize objects into a fixed set of categories.
- **Face detection:** Identify faces in images (or indicate if a face is present).
- **Email filtering:** Classify emails into spam and not-spam.
- **Medical diagnosis:** Diagnose a patient as a sufferer or non-sufferer of some disease.
- **Weather prediction:** Predict, for instance, whether or not it will rain tomorrow.

- Facial recognition technology allows social media platforms to help users tag and share photos of friends.
- Optical character recognition (OCR) technology converts images of text into movable type
- Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences.
- Self-driving cars that rely on machine learning to navigate may soon be available to consumers.



# Traditional Programming Vs Machine Learning

---

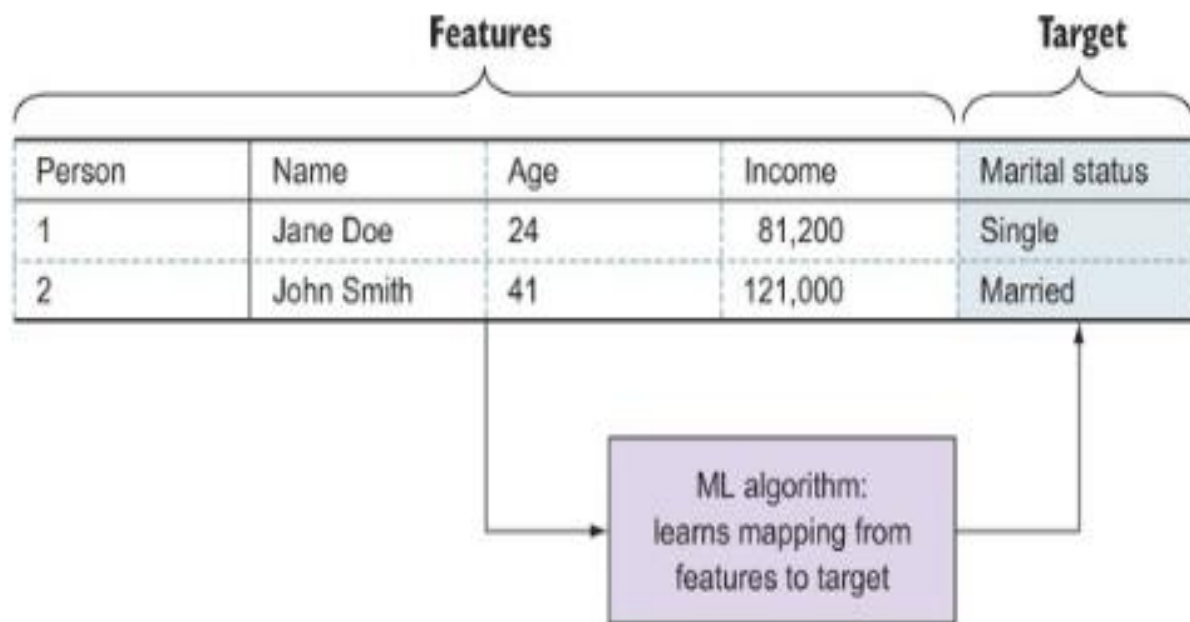


- Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it.

**Features in columns**

Person	Name	Age	Income	Marital status
1	Jane Doe	24	81,200	Single
2	John Smith	41	121,000	Married

**Examples  
in rows**



**New data with no target**

Person	Name	Age	Income	Marital status
1	Trent Mosley	26	67,500	
2	Lilly Peters	52	140,000	

ML model:  
predicts the target  
variable on new data

Marital status
Single
Married

### Testing data with target

Person	Name	Age	Income	Marital status
1	Trent Mosley	26	67,500	Single
2	Lilly Peters	52	140,000	Married

ML model:  
predicts the target  
variable on new data

Marital status
Single
Married

Predictions  
compared to  
true values

# Types of Learning

---



**Supervised  
Learning**

**Unsupervised  
Learning**

**Reinforcement  
Learning**

# Types of Learning

---

## Supervised Learning

- ◆ Makes machine learn explicitly
- ◆ Data with clearly defined output is given
- ◆ Direct feedback is given
- ◆ Predicts outcome/ future
- ◆ Resolves classification & regression problems



## Unsupervised Learning

- ◆ Machine understands the data (Identifies patterns/ structures)
- ◆ Evaluation is qualitative or indirect
- ◆ Does not predict / find anything specific



## Reinforcement Learning

- ◆ An approach to AI
- ◆ Reward based learning
- ◆ Learning from +ve & -ve reinforcement
- ◆ Machine learns how to act in a certain environment
- ◆ To maximize rewards





# Algorithms

- **Supervised learning**
  - Prediction
  - Classification (discrete labels), Regression (real values)
- **Unsupervised learning**
  - Clustering
  - Probability distribution estimation
  - Finding association (in features)
  - Dimension reduction

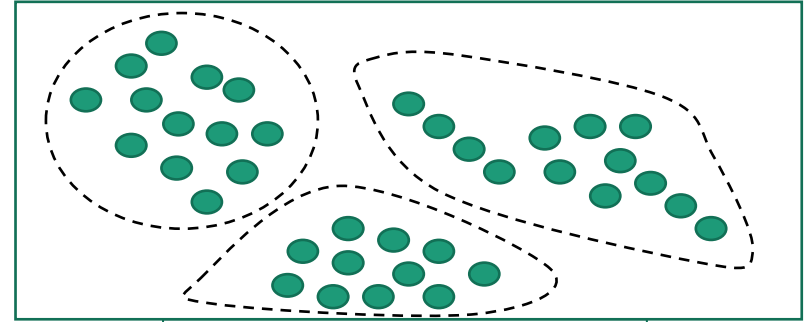
# Algorithms

- The success of machine learning system also depends on the algorithms.
- The algorithms control the search to find and build the knowledge structures.
- The learning algorithms should extract useful information from training examples.

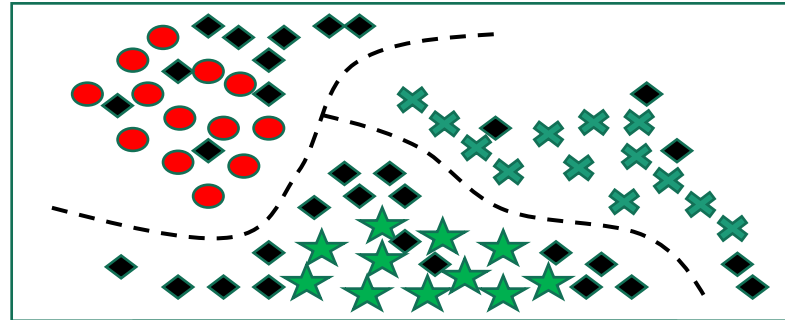
# Algorithms



Supervised learning



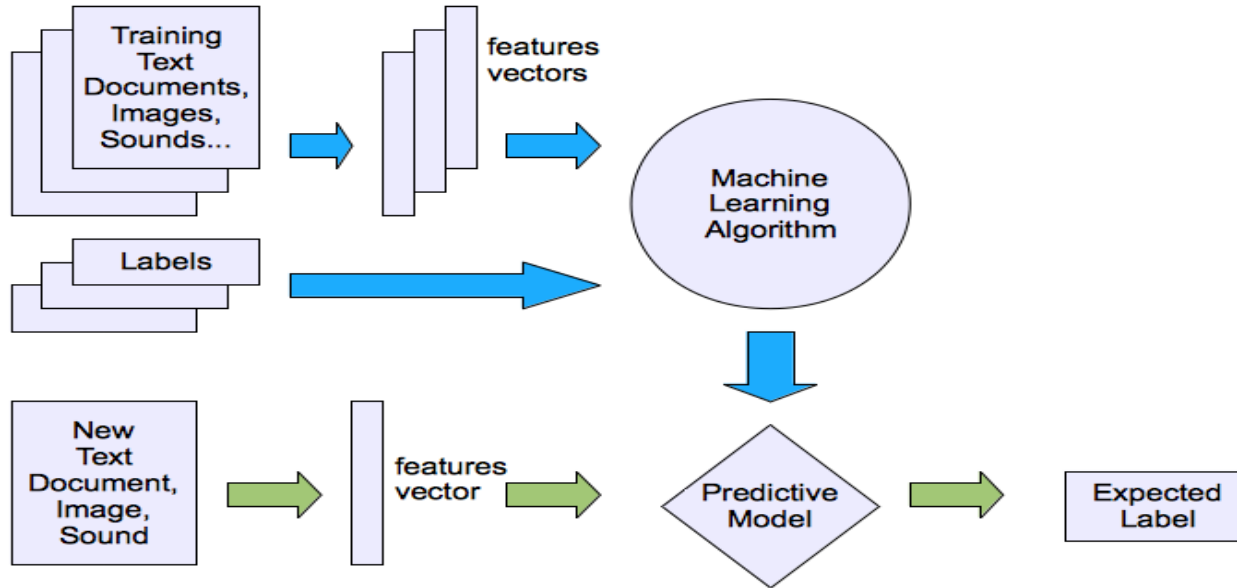
Unsupervised learning



Semi-supervised learning

# Machine learning structure

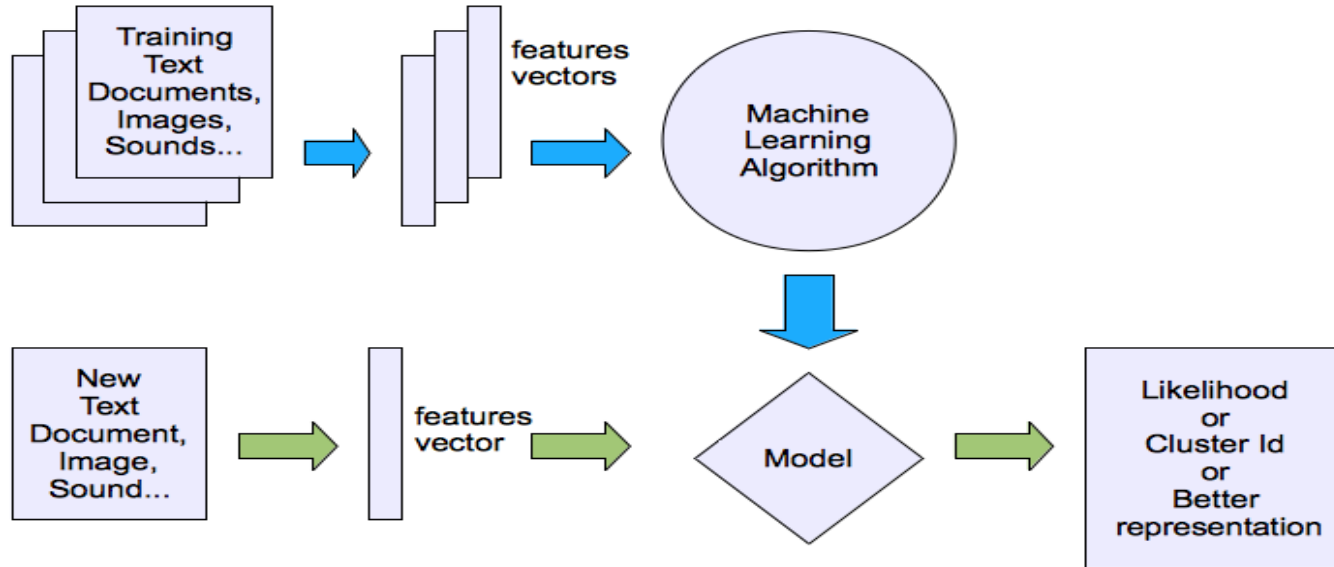
- Supervised learning



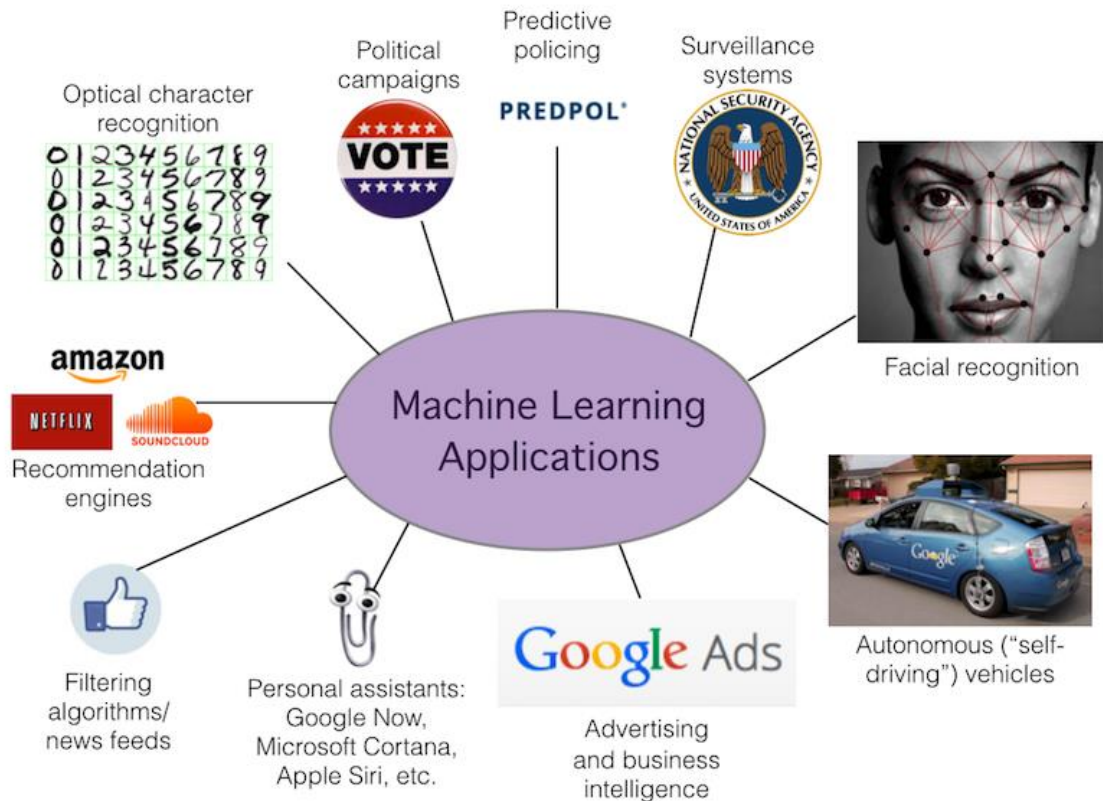
- Regression analysis is a statistical technique used to find the relations between two or more variables. In regression analysis one variable is independent and its impact on the other dependent variables is measured. When there is only one dependent and independent variable we call it simple regression. On the other hand, when there are many independent variables influencing one dependent variable we call it multiple regression

# Machine learning structure

- Unsupervised learning



# Machine Learning Applications



# Machine Learning: Problem Types

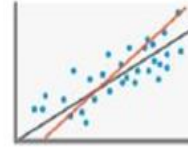
---



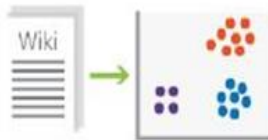
Classification  
(supervised – predictive)



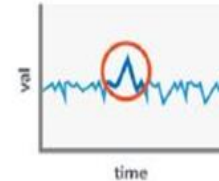
Game Playing  
(Reinforcement  
Learning)



Regression  
(supervised – predictive)



Clustering  
(unsupervised – descriptive)



Anomaly Detection  
(unsupervised – descriptive)



# ML in Big Data

---

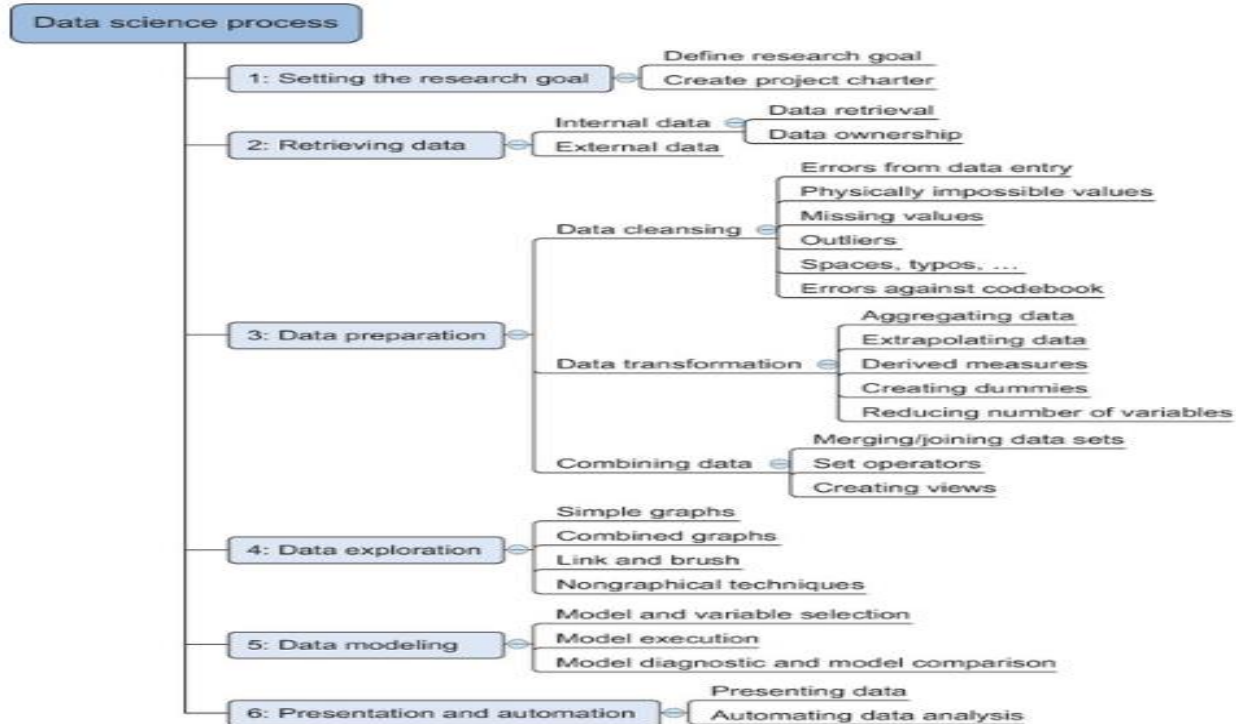
Ability to learn on large corpus of data is a real boon for ML.

Even simplistic ML models shine when they are trained on huge amount of data.

With big data toolsets, a wide variety of ML application have started to emerge other than academic specifics one

Big data democratising ML for general public

# The data science process



# Machine Learning

## **K-Nearest Neighbours**

# Different Learning Methods

## **Eager Learning**

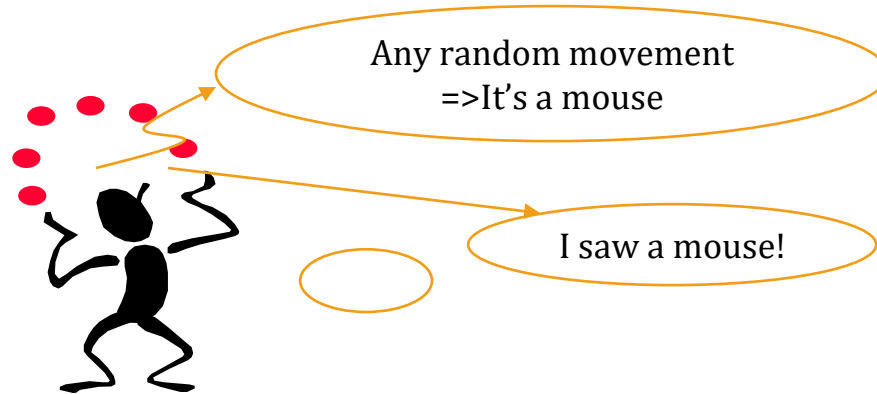
- Explicit description of target function on the whole training set

## **Instance-based Learning**

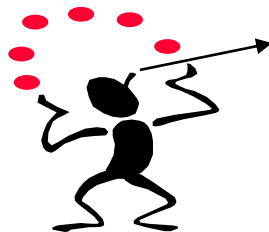
- Learning=storing all training instances
- Classification=assigning target function to a new instance
- Referred to as “Lazy” learning

# Different Learning Methods

## Eager Learning



# Instance Based Learning



Its very similar to a  
Desktop!!



# Classification

- Given: dataset of instances with known categories
- Goal: using the “knowledge” in the dataset, classify a given instance
  - predict the category of the given instance that is rationally consistent with the dataset

# Instance Based Learning

## K-Nearest Neighbor Algorithm

- Weighted Regression
- Case-based reasoning



## **K\_Nearest Neighbours**

- For a given instance T, get the top k dataset instances that are “nearest” to T
  - Select a reasonable distance measure
- Inspect the category of these k instances, choose the category C that represent the most instances
- Conclude that T belongs to category C

# K\_Nearest Neighbours

## Features

- All instances correspond to points in an  $n$ -dimensional Euclidean space
- Classification is delayed till a new instance arrives
- Classification done by comparing feature vectors of the different points
- Target function may be discrete or real-valued

# K\_Nearest Neighbour Classifier

## Learning by Analogy

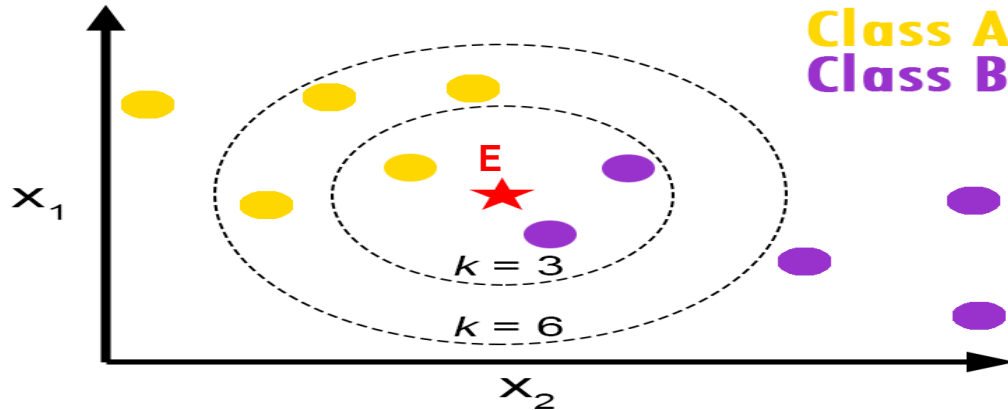
- Tell me who your friends are and I'll tell you who you are?
- A new example is assigned to the most common class among the (K) examples that are most similar to it



# K\_Nearest Neighbour Algorithm

To determine the class of a new example E:

- Calculate the distance between E and all examples in the training set
- Select K-nearest examples to E in the training set
- Assign E to the most common class among its K-nearest neighbors



# Distance Between Neighbors

Each example is represented with a set of numerical attributes



**Jay:**  
**Age=35**  
**Income=95K**  
**No. of credit cards=3**



**Rina:**  
**Age=41**  
**Income=215K**  
**No. of credit cards=2**

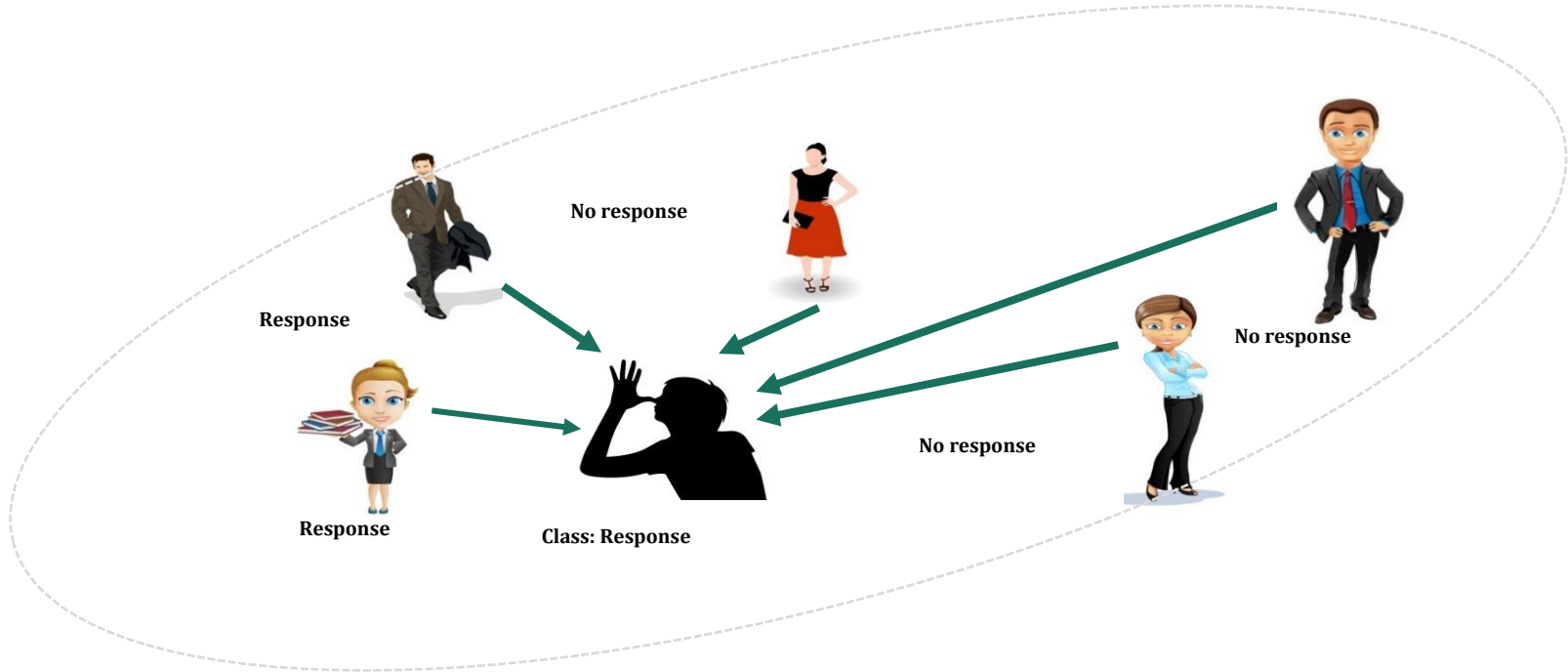
- “Closeness” is defined in terms of the Euclidean distance between two examples
- The Euclidean distance between  $X=(x_1, x_2, x_3, \dots, x_n)$  and  $Y=(y_1, y_2, y_3, \dots, y_n)$  is defined as:

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

$$\text{Distance (Jay, Rina)} = \sqrt{(35-41)^2 + (95,000-215,000)^2 + (3-2)^2}$$

# K\_Nearest Neighbours: Instance Based Learning

- No model is built: Store all training examples
- Any processing is delayed until a new instance must be classified



## K\_Nearest Neighbours: Example

Customer	Age	Income	No. credit cards	Response
Jay	35	35K	3	No
Rina	22	50K	2	Yes
Hema	63	200K	1	No
Tommy	59	170K	1	No
Neil	25	40K	4	Yes
Dravid	37	50K	2	?

## K\_Nearest Neighbours: Example

Customer	Age	Income	No. credit cards	Response	Distance from Dravid
Jay	35	35K	3	No	$\sqrt{(35 - 37)^2 + (35 - 50)^2 + (3 - 2)^2}$ = 15.16
Rina	22	50K	2	Yes	15
Hema	63	200K	1	No	152.23
Tommy	59	170K	1	No	122
Neil	25	40K	4	Yes	15.74
Dravid	37	50K	2	?	0



## K\_Nearest Neighbours: Strengths and Weaknesses



### Strengths

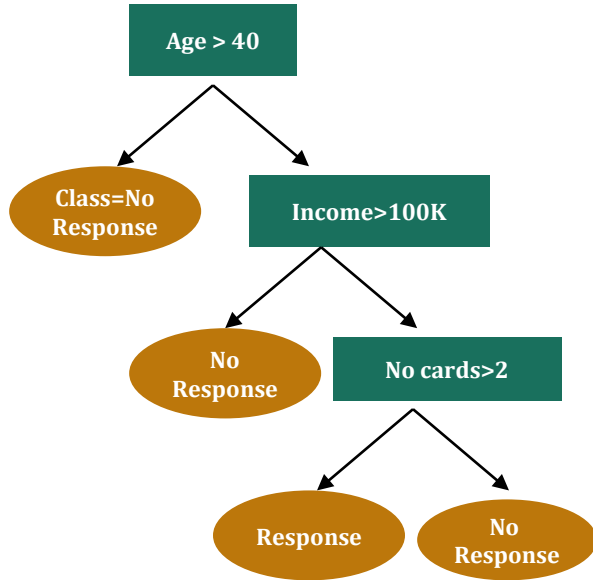
- Simple to implement and use
- Comprehensible: easy to explain prediction
- Robust to noisy data by averaging k-nearest neighbors
- Some appealing applications (will discuss next in personalization)

### Weaknesses

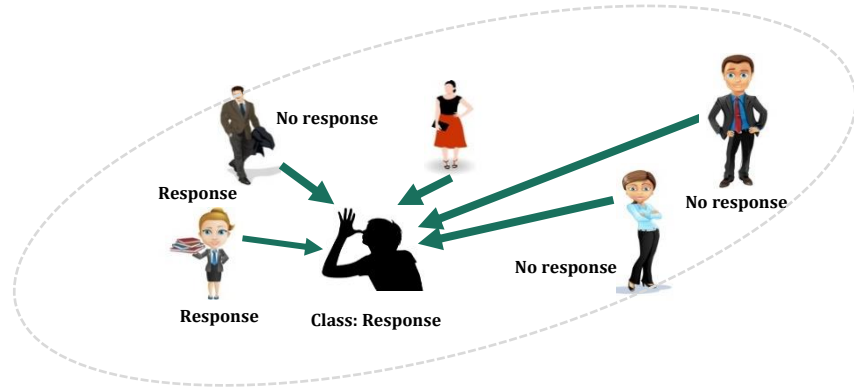
- Need a lot of space to store all examples
- Takes more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples)
- Each attribute is treated equally

# K\_Nearest Neighbours: Classifier

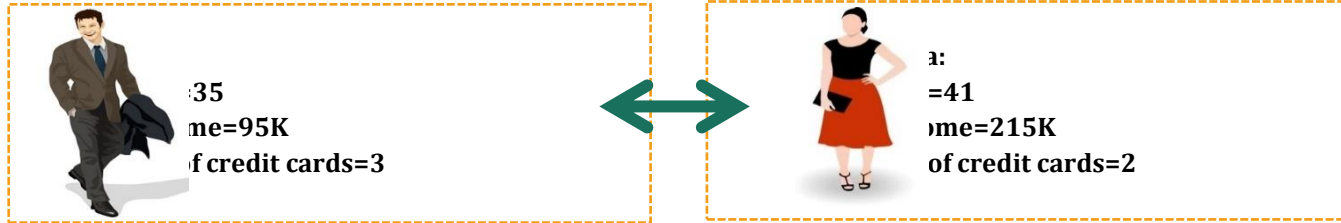
## Classification Tree Modes



## K-Nearest Neighbors



# K\_Nearest Neighbours: Strengths and Weaknesses



$$\text{Distance (Jay, Rina)} = \sqrt{[(35-41)^2 + (95,000-215,000)^2 + (3-2)^2]}$$

- Distance between neighbors could be dominated by some attributes with relatively large numbers (e.g., income in our example)
- **Important to normalize some features** (e.g., map numbers to numbers between 0-1)

**Example:** Income

Highest income = 500K

Davis's income is normalized to 95/500, Rina income is normalized to 215/500, etc.)

## K\_Nearest Neighbours: Strenghts and Weaknesses

Normalization of Variables				
Customer	Age	Income	No. credit cards	Response
Jay	$55/63 = 0.175$	$35/200 = 0.175$	$3/4 = 0.75$	No
Rina	$22/63 = 0.34$	$50/200 = 0.25$	$2/4 = 0.5$	Yes
Hema	$63/63 = 1$	$200/200 = 1$	$1/4 = 0.25$	No
Tommy	$59/63 = 0.93$	$170/200 = 0.175$	$1/4 = 0.25$	No
Neil	$25/63 = 0.39$	$40/200 = 0.2$	$4/4 = 1$	Yes
Dravid	$37/63 = 0.58$	$50/200 = 0.25$	$2/4 = 0.5$	Yes

## K-Nearest Neighbor: Strengths & Weaknesses

- Distance works naturally with numerical attributes  
 $d(\text{Rina}, \text{Johm}) = \sqrt{(35-37)^2 + (35-50)^2 + (3-2)^2} = 15.16$
- What if we have nominal attributes?

**Example:** Married

Customer	Married	Income	No. credit cards	Response
Jay	Yes	35K	3	No
Rina	No	50K	2	Yes
Hema	No	200K	1	No
Tommy	Yes	170K	1	No
Neil	No	40K	4	Yes
Dravid	Yes	50K	2	Yes

# Non-Numeric Data

- Feature values are not always numbers
- Example
  - Boolean values: Yes or no, presence or absence of an attribute
  - Categories: Colors, educational attainment, gender
- How do these values factor into the computation of distance?

# Dealing with Non-Neumeric Data

- Boolean values => convert to 0 or 1
  - Applies to yes-no/presence-absence attributes
- Non-binary characterizations
  - Use natural progression when applicable; e.g., educational attainment: GS, HS, College, MS, PHD => 1,2,3,4,5
  - Assign arbitrary numbers but be careful about distances; e.g., color: red, yellow, blue => 1,2,3
- How about unavailable data?  
(0 value not always the answer)

# Preprocessing Your Dataset

- Dataset may need to be preprocessed to ensure more reliable data mining results
- Conversion of non-numeric data to numeric data
- Calibration of numeric data to reduce effects of disparate ranges
  - Particularly when using the Euclidean distance metric



## k-NN Variations

- Value of  $k$ 
  - Larger  $k$  increases confidence in prediction
  - Note that if  $k$  is too large, decision may be skewed
- Weighted evaluation of nearest neighbors
  - Plain majority may unfairly skew decision
  - Revise algorithm so that closer neighbors have greater “vote weight”
- Other distance measures

## Other Distance Measures

- City-block distance (Manhattan dist)
  - Add absolute value of differences
- Cosine similarity
  - Measure angle formed by the two samples (with the origin)
- Jaccard distance
  - Determine percentage of exact matches between the samples (not including unavailable data)

$$A=(a,b,c,d) \quad B=(a,c,f,g) \quad J=A \cap B / A \cup B = 2/6 = 1/3$$

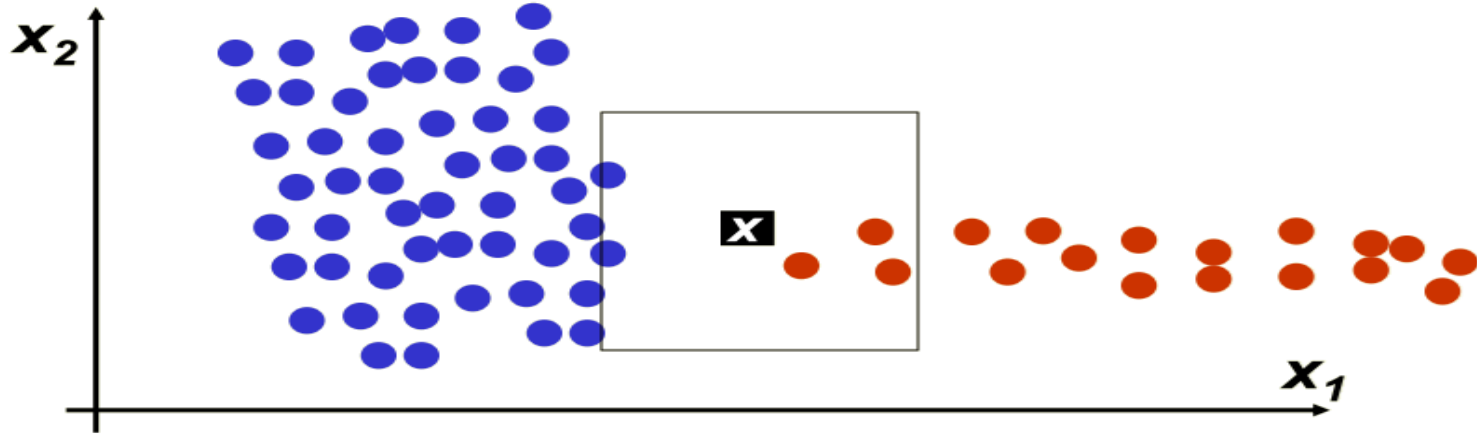
Mainly used in text mining

- Others

## Distance-Weighted Nearest Neighbor Algorithm

- Assign weights to the neighbors based on their 'distance' from the query point
  - Weight 'may' be inverse square of the distances (the farther away, the less weight the point has)
- All training points may influence a particular instance
  - Shepard's method

## How to Choose "K"?



- For  $k = 1, \dots, 5$  point  $x$  gets classified correctly
  - red class
- For larger  $k$  classification of  $x$  is wrong
  - blue class

# How to Find Optional Value of "K"?

## Cross-Validation

- Use P- fold cross validation
- Divide training data into p-parts.
- Select only (p-1) parts for training and remaining 1 part for testing.
  - There are (p-1) combinations of training and test set pairs
- For each of the (p-1) combination learn K-NN model with different K and compute prediction error for test set.
- Compute the average test error for different K
- Select K with minimum average test error.

LOOCV = Leave One Out Cross Validation

Build model on all the data set except one instance and then test the model on that one instance(row/sample)

Thus, for a dataset of m, train on m-1 instances and test on the one instance and this will be done for each instance...

## K-NN: Computational Complexity

- Basic k-NN algorithm stores all examples. Suppose
- we have  $n$  examples each of dimension  $d$ 
  - $O(d)$  to compute distance to one example
  - $O(nd)$  to find one nearest neighbor
  - $O(knd)$  to find  $k$  closest examples
  - **Thus complexity is  $O(knd)$**
- This is prohibitively expensive for large number of samples
- But we need large number of samples for k-NN to work well!