Introduction to Data Science

GENERAL ASSEMBLY

Agenda

01 What is Data Science

02 The Data Science Process

03 Machine Learning and (some) coding

04 Applications of Data Science

05 How to get started in Data Science

06 QnA

Meet Your Instructor

{Raghav Garg}

Analytics Manager II, Grab

- Masters in Business Analytics (MSBA) from NUS
- 7+ years of experience in data science and experimentation
- Linkedin - https://www.linkedin.com/in/raghavgarg91/

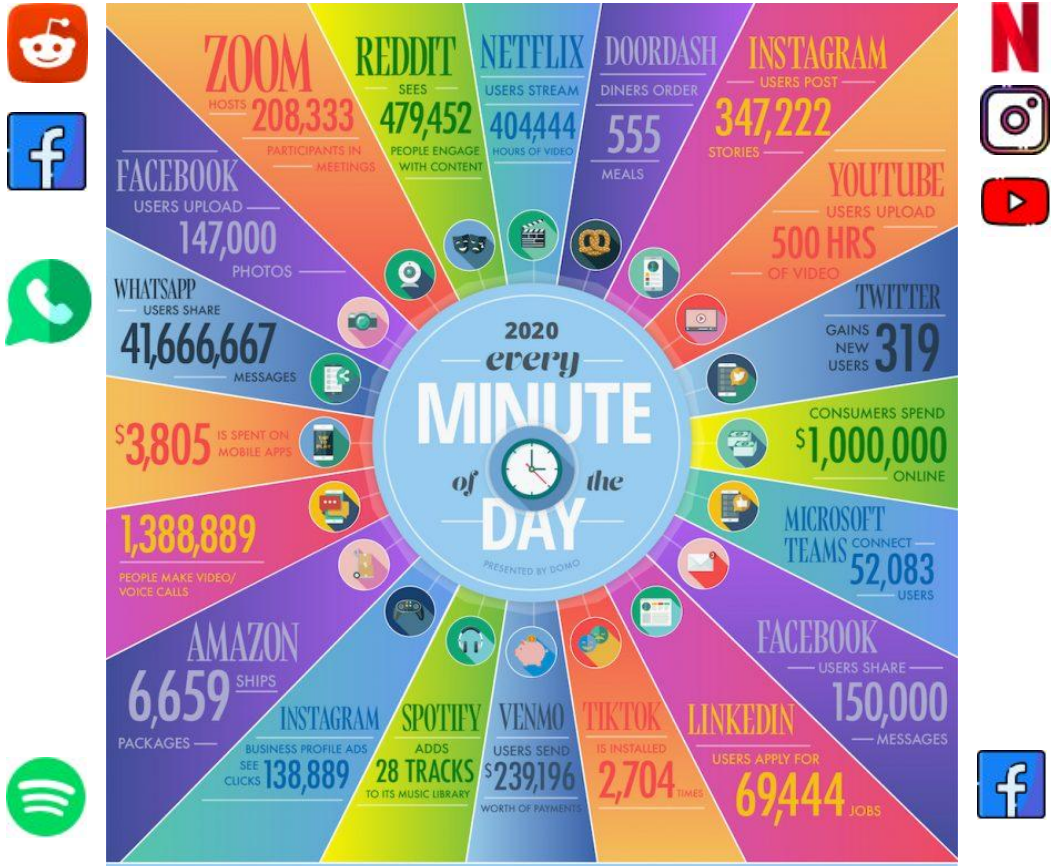
UnitedHealth Group

NUS National University of Singapore

DBS

Grab

2020 every minute of the day



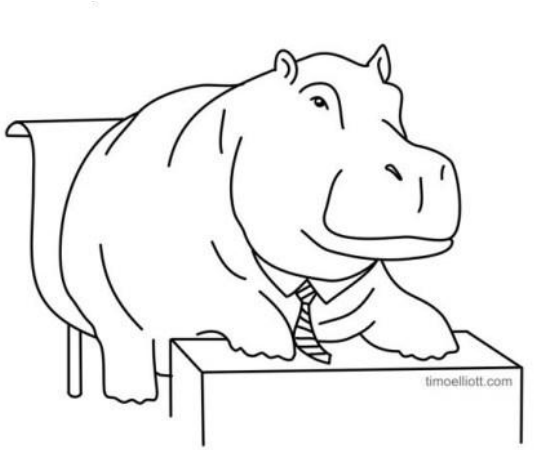
Data Science is taking data-driven way to solve a problem

Analytics uses **data, math, computer science and domain knowledge** to answer business questions, discover relationships, predict unknown outcomes and automate decisions.

The goal of **data science** is to gain insights and knowledge from any type of **data** (both structured and unstructured

HIPPO vs Data Science

No Analytics? Welcome to HIPPO

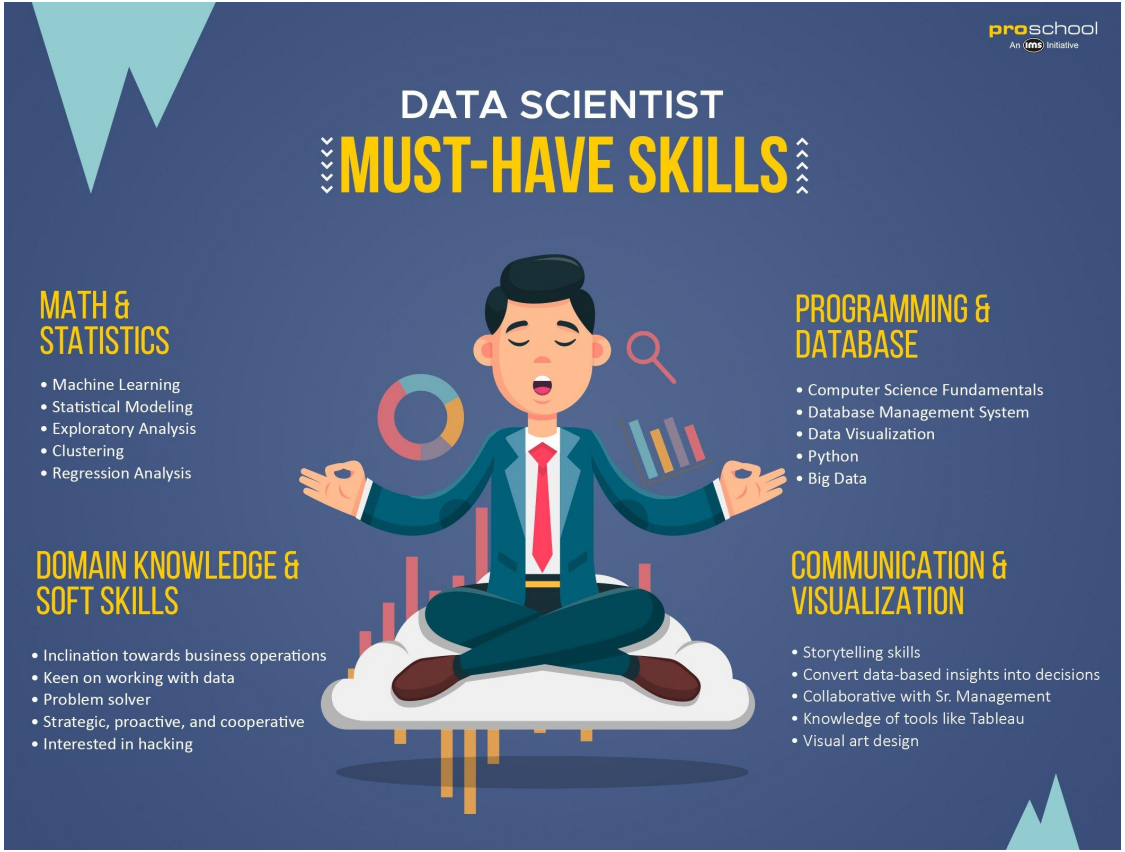


*Highest Paid Person's Opinion

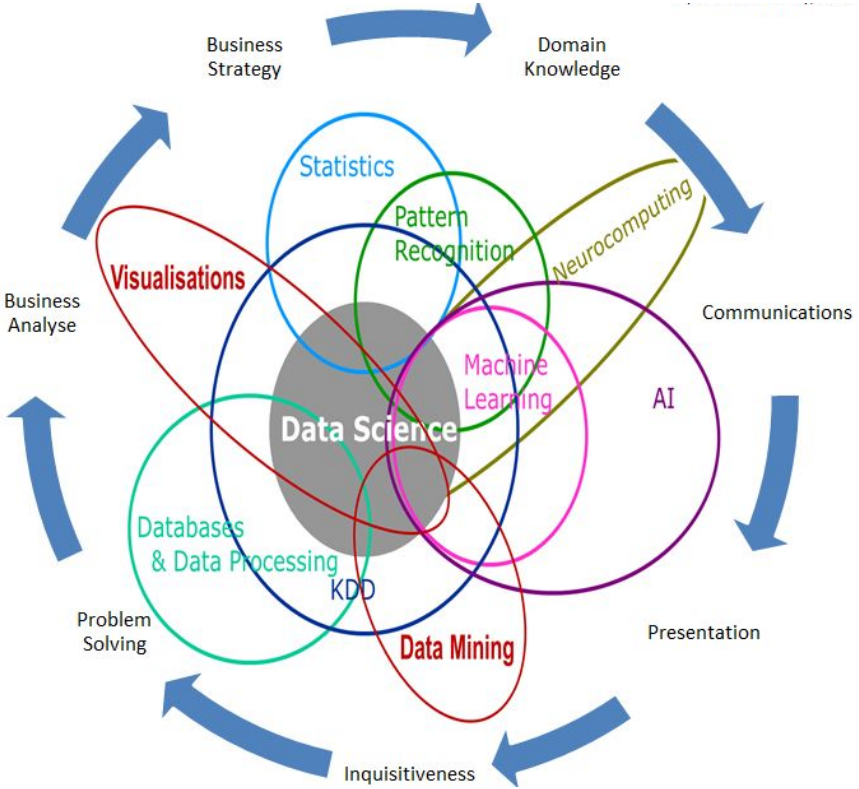
Opinions are good. **Data is better**

"When you two have finished arguing your opinions, I actually have data!"

Skills needed to become a Data Scientist!



Data Science is Multidisciplinary!



There is a saying ‘A jack of all trades and a master of none’.

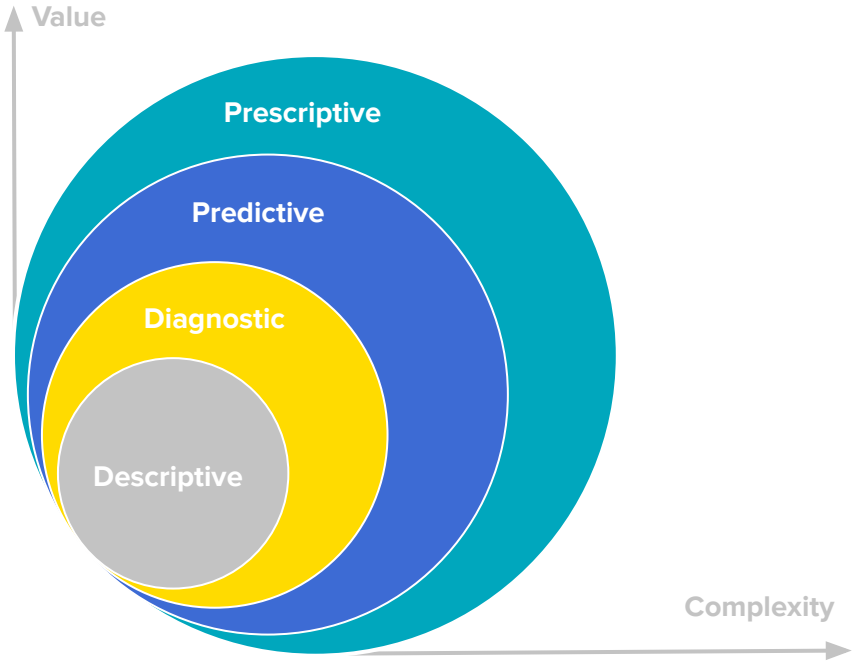
When it comes to being a data scientist you need to be a bit like this but perhaps a better saying would be **‘A jack of all trades and a master of some’.**

“

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

- Anonymous

Different degree of analytics



Descriptive: What's happening in my business?
- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: Why is it happening?
- Ability to drill down to the root-cause
- Ability to isolate all confounding factors

Predictive: What's likely to happen?
- Historical patterns being used to predict specific outcomes using algorithms

Prescriptive: What do I need to do?
- Recommended actions and strategies based on champion strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Source: principa.co.za

The 6-Step Data Science Framework

1. Identify the Problem

Ask relevant questions, identify opportunity areas and define impactful problems to be solved

2. Collect Raw Data

Collect and gather the data from multiple sources

3. Process the Data

Examine the data at high level, clean the data by combining multiple data, missing value imputation, outlier treatment, etc.

4. Explore the Data

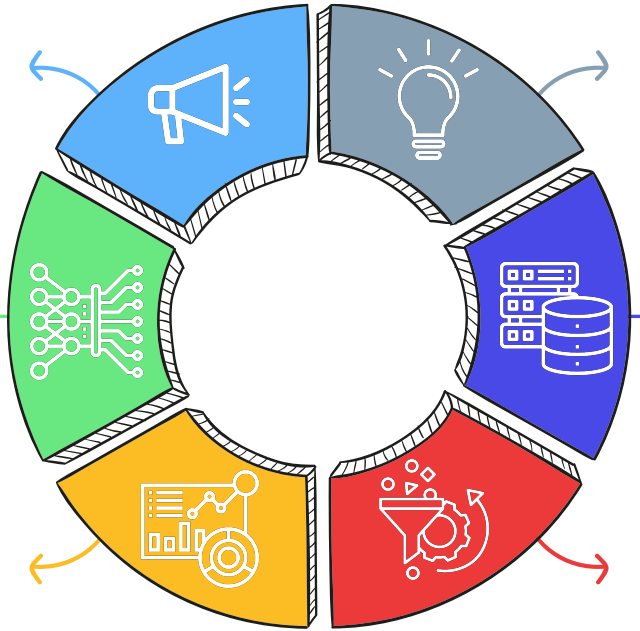
Hypothesis formation, descriptive analytics and visual inspection of data

5. Perform in-depth analysis

Create predictive models, evaluate and refine models

6. Communicate Results

Tell business insights in a clear and actionable story. Effectively communicate to non-technical audiences



The 6-Step Data Science Framework

01

Identify the Problem

- **Domain Knowledge** (needs)

- **Product Intuition** (metrics)

- **Business Strategy** (priorities)

- **Teamwork** (people and resources)

02

Collect Raw Data

- **Querying Structured Databases:** SQL

- **Retrieving unstructured Info:** Informational retrieval/text mining

- **Distributed Storage:** Hadoop HDFS, Spark

03

Process the Data

- **Scripting Language:** Python or R

- **Data Wrangling & Cleaning:** Pandas library

- **Distributed processing:** Hadoop mapreduce/Spark

04

Explore the Data

- **Scientific Computing:** Python: numpy, matplotlib, scipy, pandas

- **Inferential Statistics:** Hypothesis testing, Correlation vs Causation

- **Experimental Design:** A/B testing

05

Perform in-depth Analysis

- **Machine Learning:** Supervised/Unsupervised algorithms

- **ML Tools Library:** Python: scikit-learn

- **Advanced Maths:** Linear Algebra and Multivariate Calculus

06

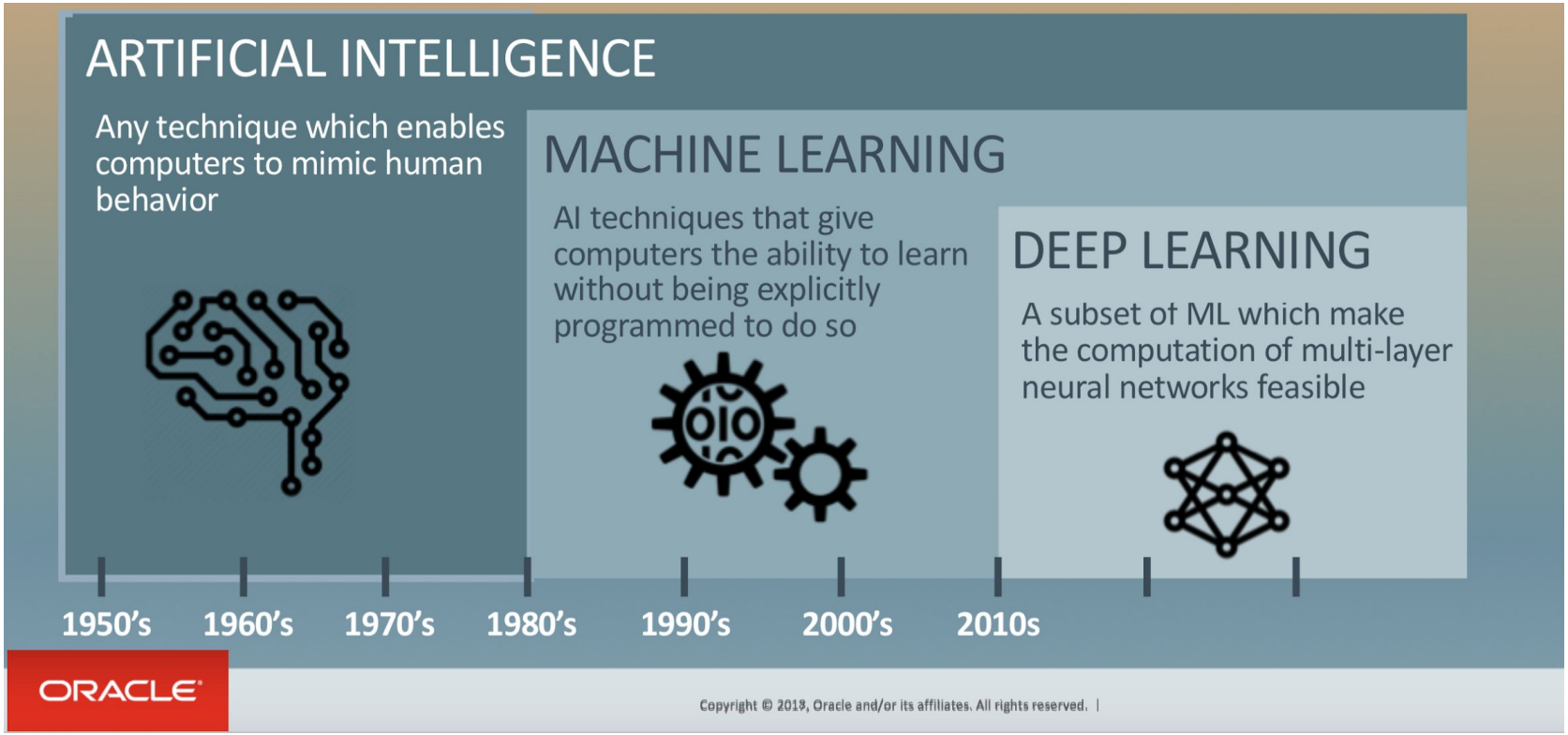
Communicate Results

- **Business Acumen :** Non technical terminology

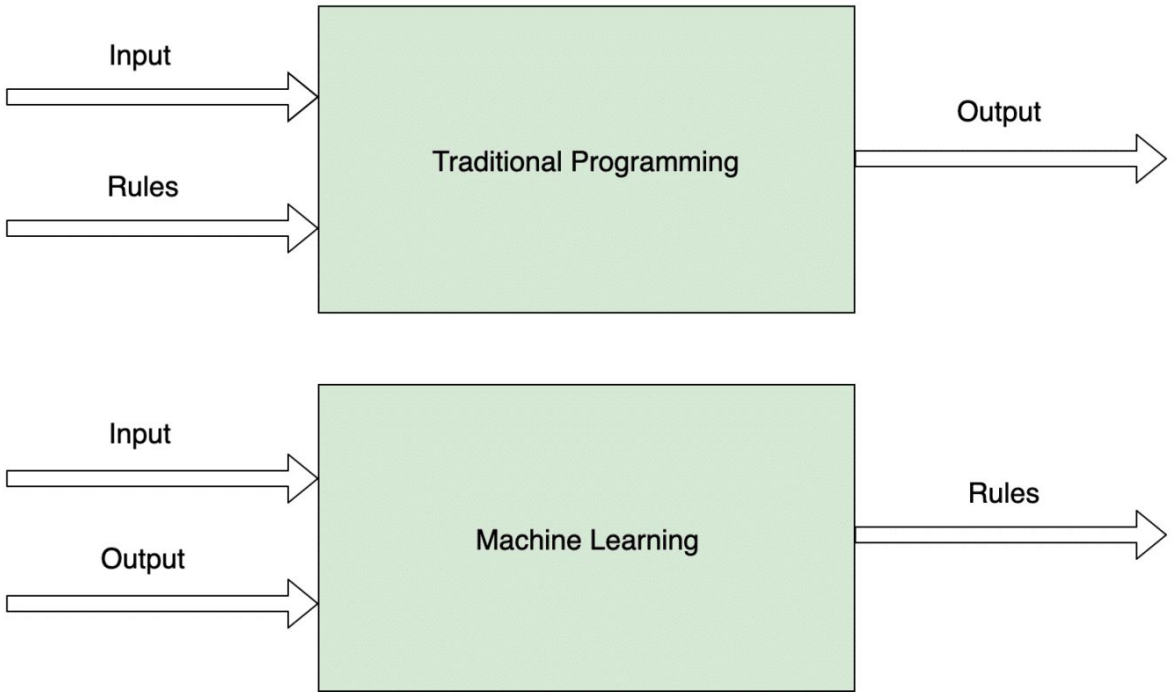
- **Data Visualization Tool(s) :** Tableau, D3.js, PowerBI, etc.

- **Data Storytelling :** presentation skills with clear actionables

AI vs Machine Learning vs Deep Learning



Traditional programming vs Machine Learning

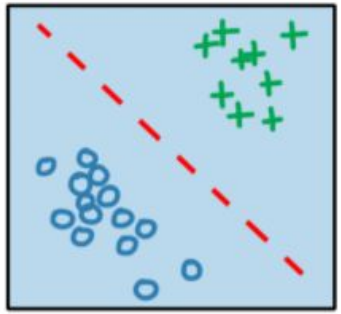


Classification of Predictive Analytics Algorithms

Supervised Learning

Machine learns by using labelled data to achieve certain pre-defined outcome

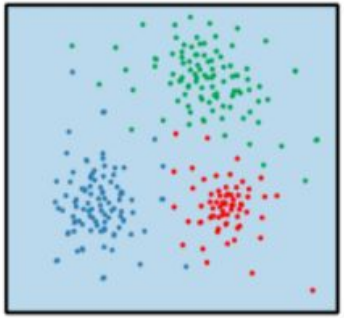
Eg. Customer segmentation, product recommendation, etc.



UnSupervised Learning

Machine learns by using unlabelled data to discover underlying hidden pattern

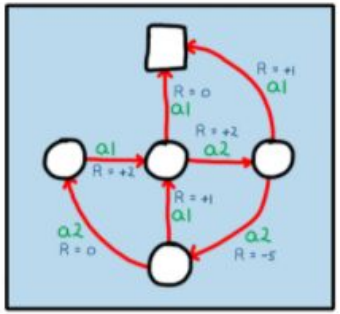
Eg.Predict sales,Churn prediction, Image classification, etc.



Reinforcement Learning

Agent interacts with the environment by producing actions and discovers errors and rewards

Eg. Self driving cars, gaming, etc.



How Supervised Machine Learning works

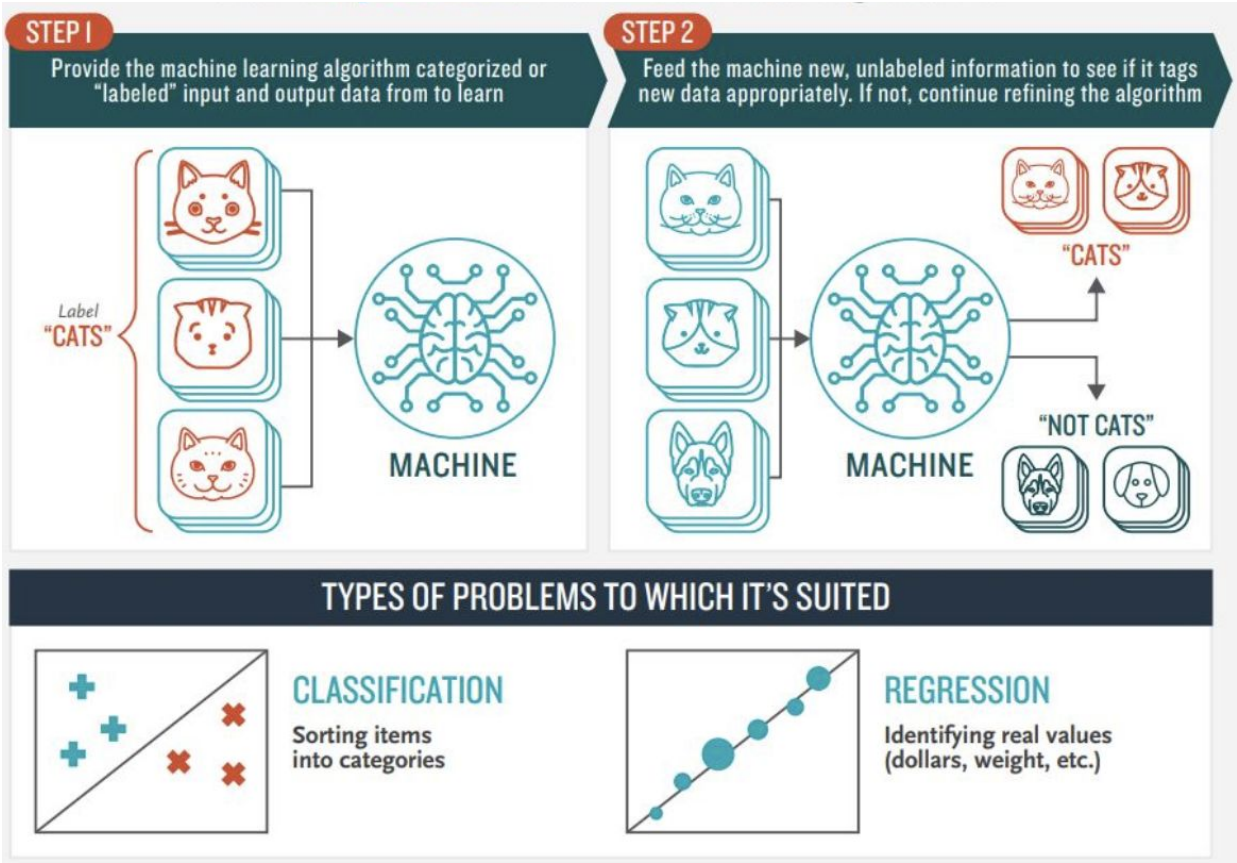
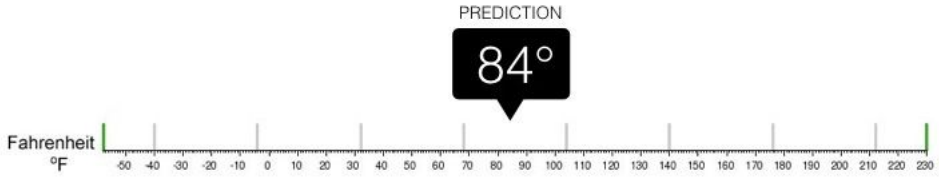


Image Source - *boozallen*

Supervised Learning - Regression vs Classification

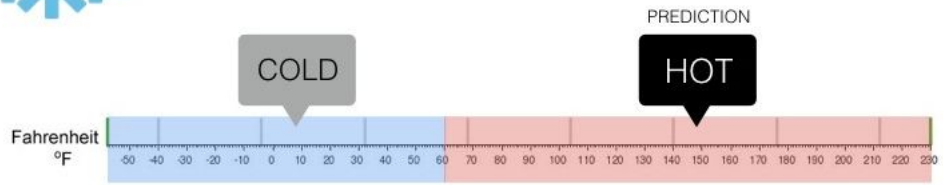
Regression

What is the temperature going to be tomorrow?



Classification

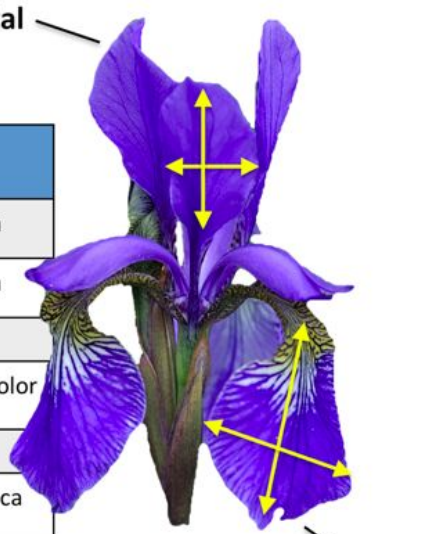
Will it be Cold or Hot tomorrow?



Supervised Learning - IRIS data

Samples
(instances, observations)

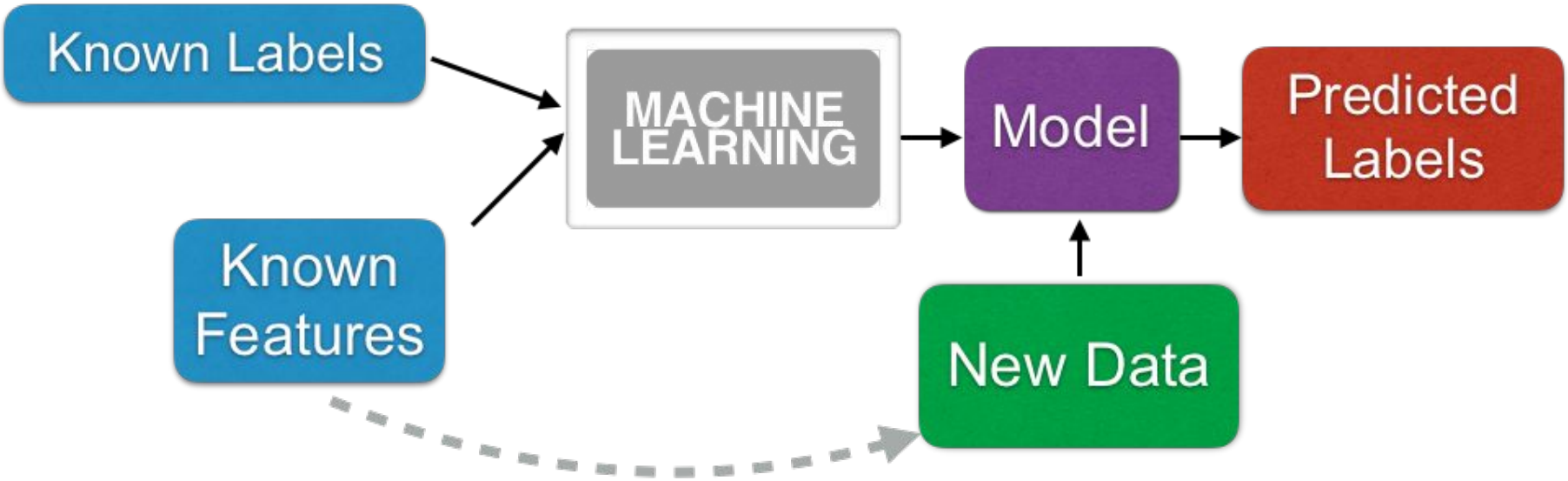
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Supervised Learning



Supervised Learning

Known Labels

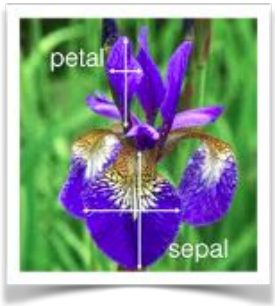
Model-building

MACHINE LEARNING

Model

Setosa
Versicolor
Virginica

Predicted Labels



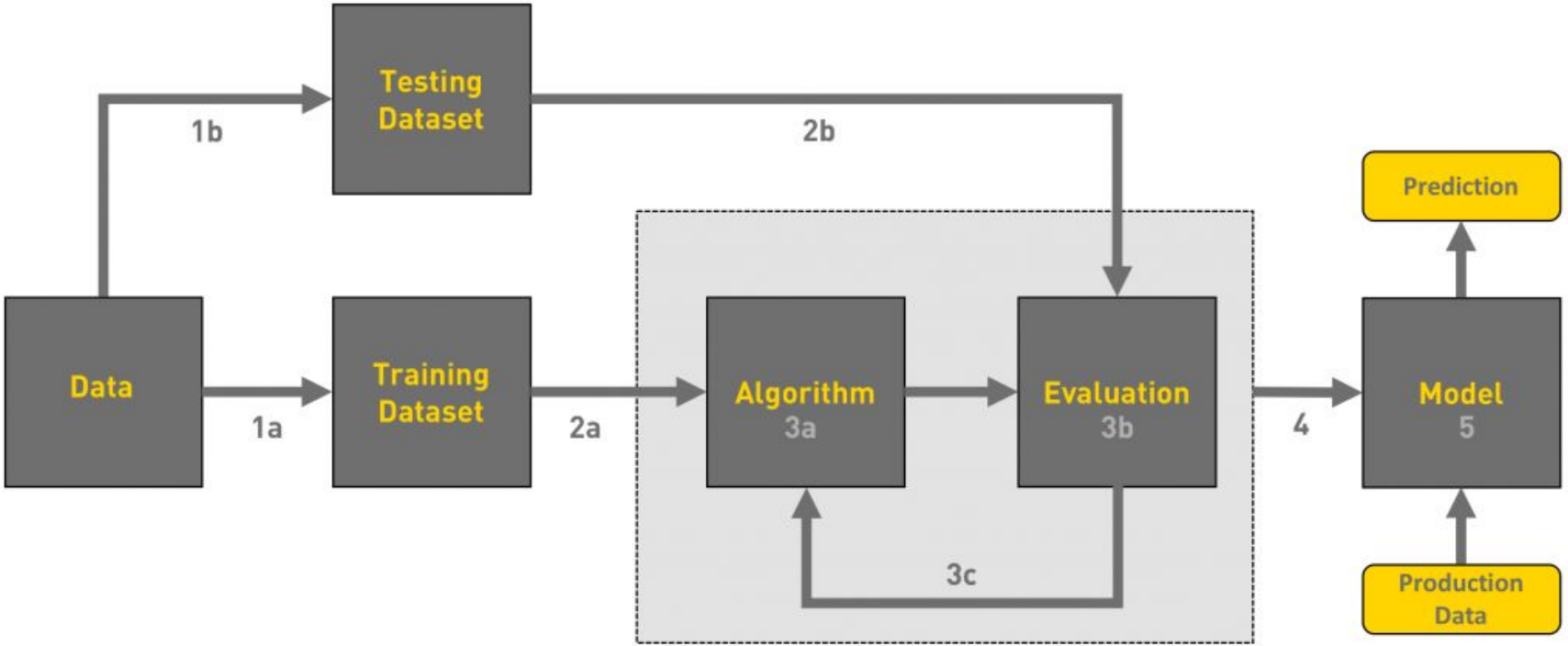
Known Features

Sepal Length
Sepal Width
Petal Length
Petal Width

New Data

Sepal Length
Sepal Width
Petal Length
Petal Width

Supervised Learning - Train Test Split



How Unsupervised Machine Learning works

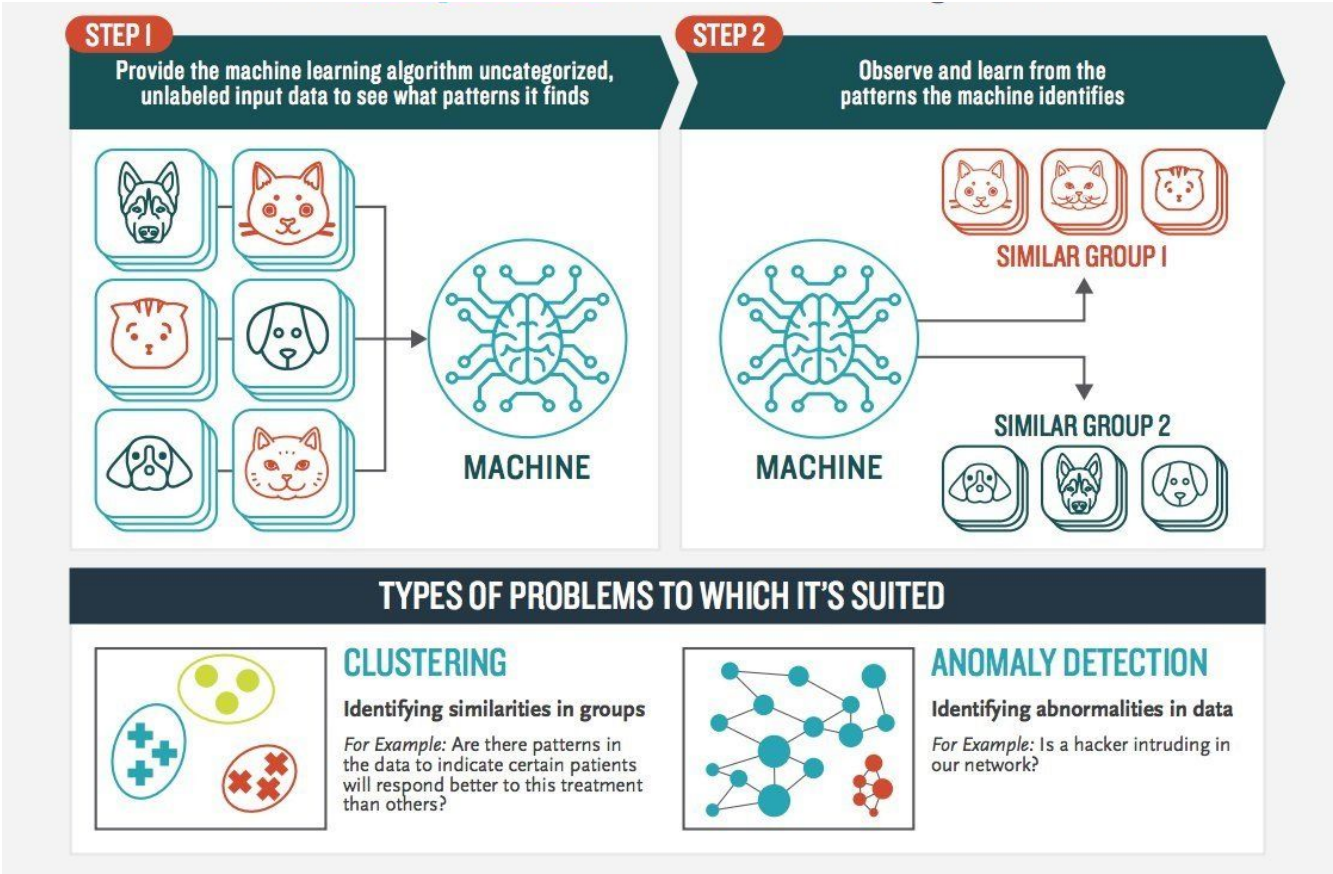
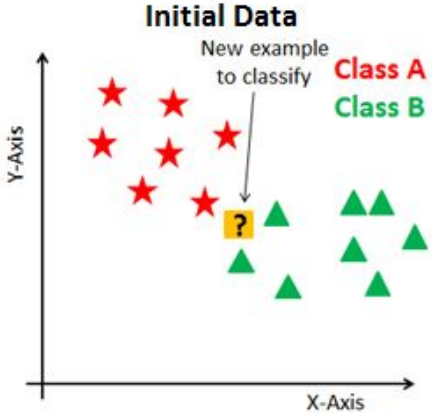
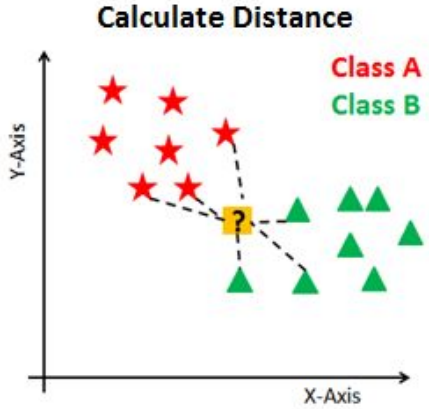
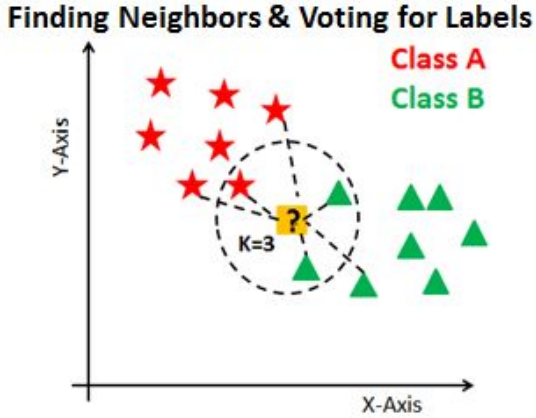


Image Source - boozallen

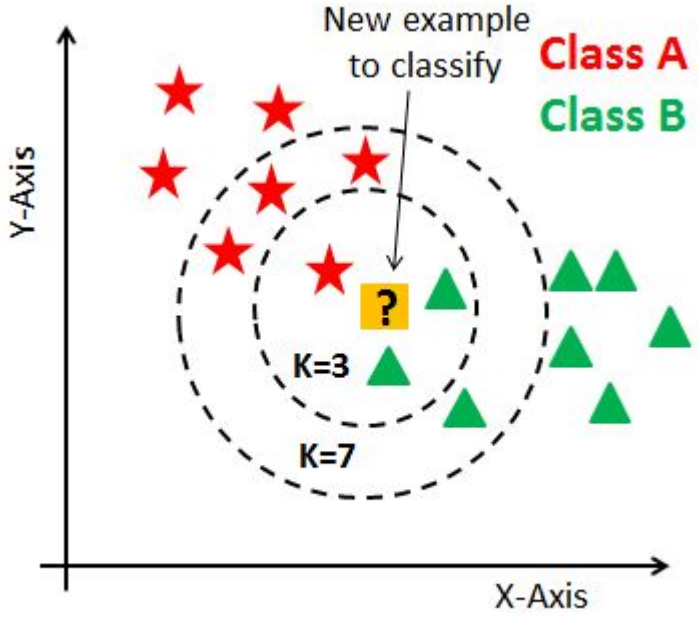
KNearestNeighbor (KNN) Classifier







KNearestNeighbor (KNN) Classifier



Let's code

1. Colab : Google's jupyter notebook on Cloud
    a. http://tiny.cc/DSCoding

2. Machine learning on Iris Dataset
    a. Supervised learning using KNearestNeighbourClassifier
    b. Unsupervised using KMeans Clustering

Applications of Data Science

Recommendation System

Customers Who Bought This Item Also Bought

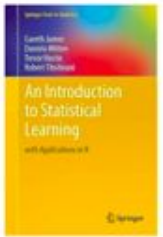
Page 1 of 15

Data Science from Scratch: First Principles with Python
› Joel Grus
54
#1 Best Seller in Data Mining
Paperback
$33.99 Prime

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and…
› Wes McKinney
118
Paperback
$27.68 Prime

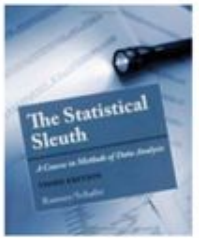
Data Science for Business: What You Need to Know about Data Mining and…
› Foster Provost
135
Paperback
$37.99 Prime

Reproducible Research with R and R Studio, Second Edition…
Christopher Gandrud
3
Paperback
$51.97 Prime



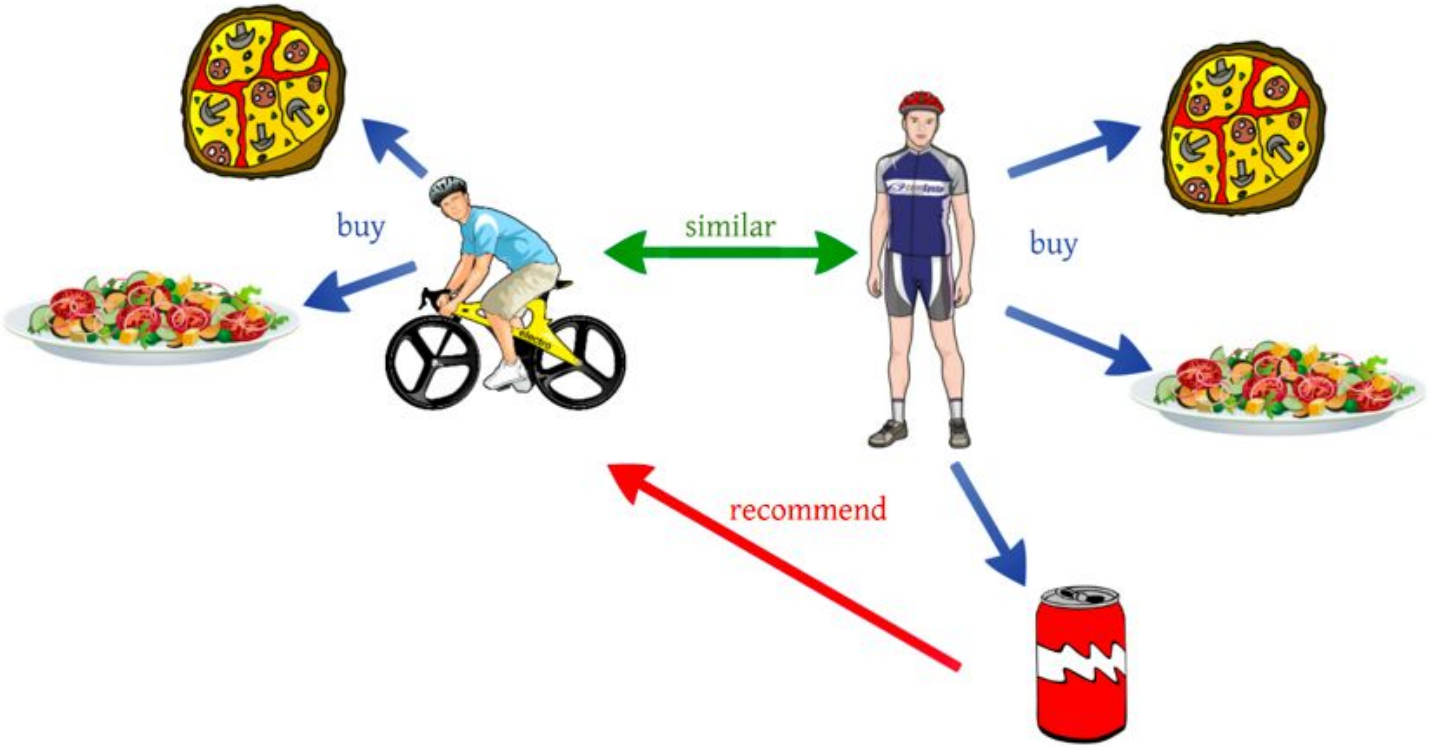
An Introduction to Statistical Learning: with Applications in R…
› Gareth James
105
Hardcover
$68.35 Prime

Data Smart: Using Data Science to Transform Information into Insight
› John W. Foreman
99
#1 Best Seller in Computer Simulation
Paperback
$28.16 Prime



The Statistical Sleuth: A Course in Methods of Data Analysis
Fred Ramsey
6
Hardcover
$284.42 Prime

Recommendation System



Recommendation System

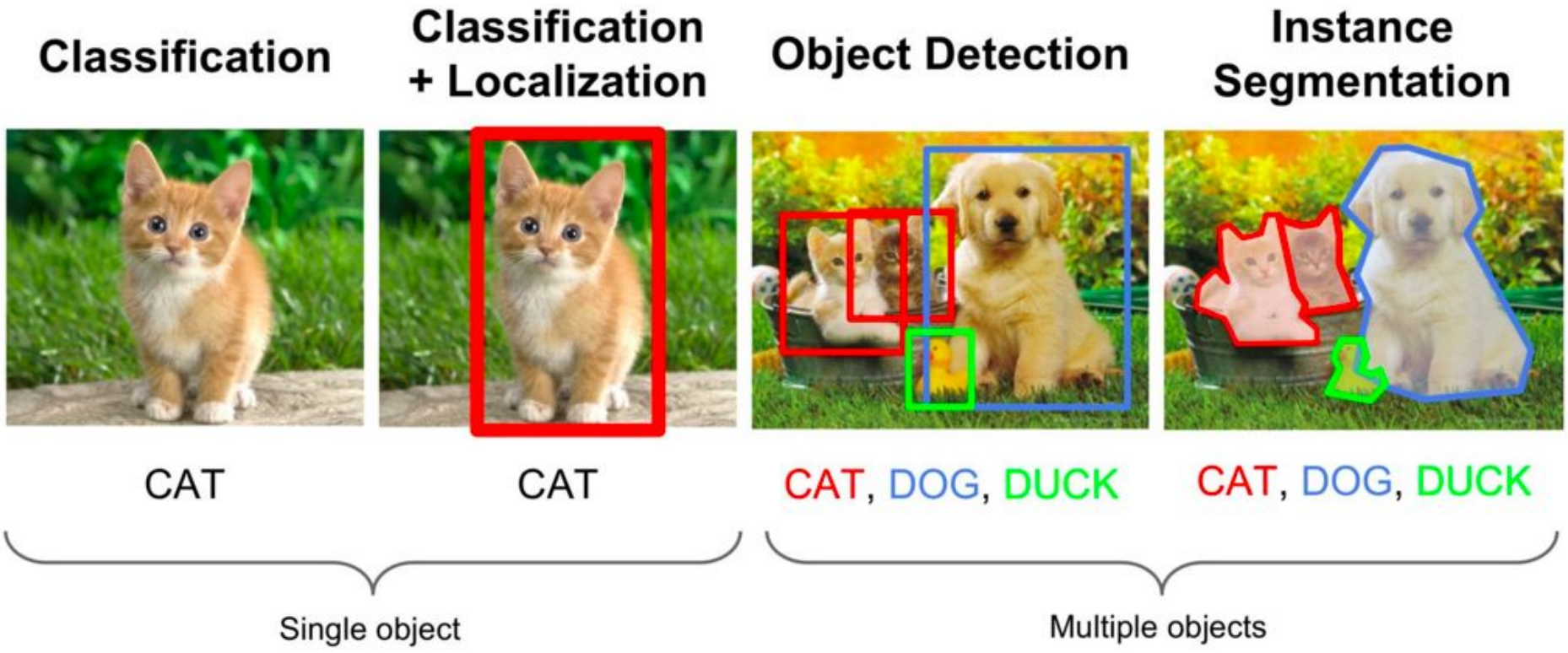
Everything is personalized

Source: InfoQ

Over 80% of what people watch comes from our recommendations

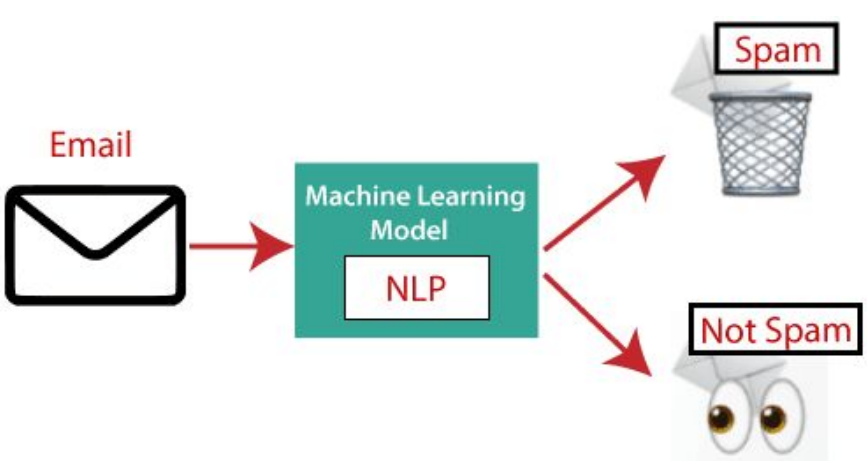
Recommendations are driven by **Machine Learning**

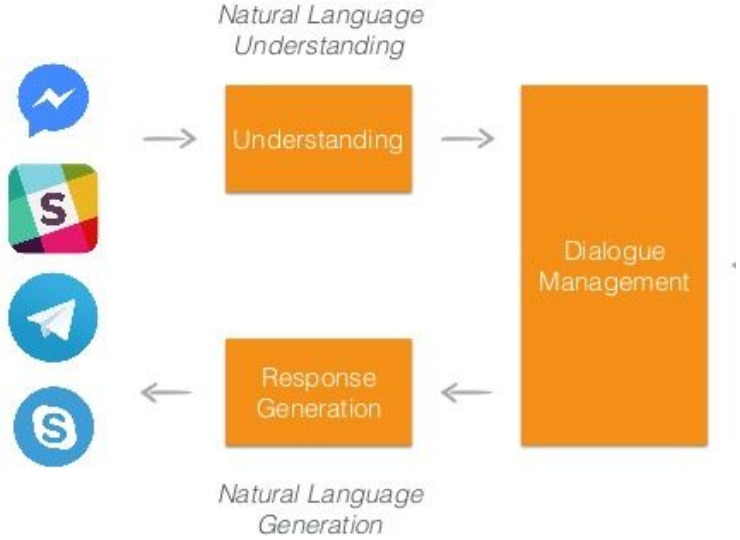
Image processing/ Computer Vision



Facial Recognition

Natural Language Processing (NLP)





Virtual Assistants (Google Duplex)

Why Now?

Data Scientist:

The Sexiest Job of the 21st Century

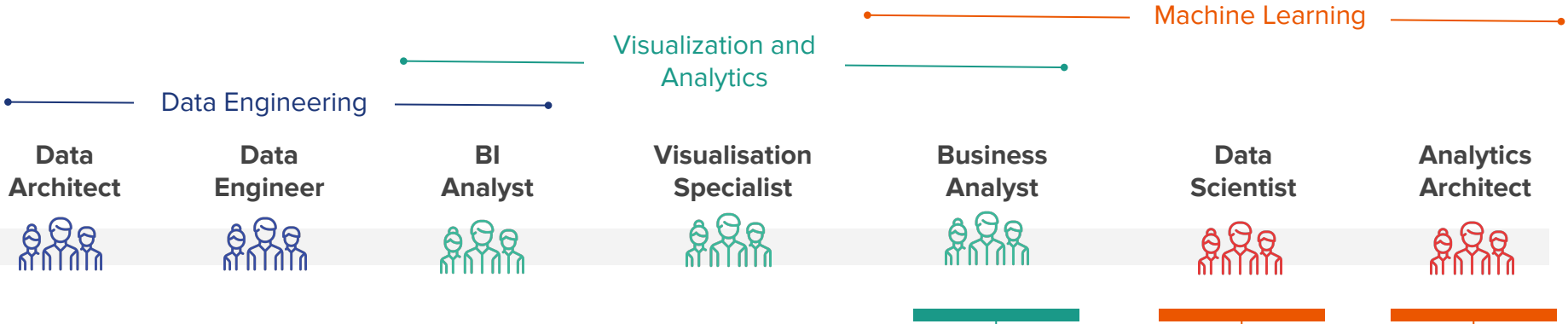
Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

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 experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012

Roles in Data Science



Business Analyst

- Improves business processes
- Conscious listener and storyteller
- Intermediary between business & IT
- Good BI understanding

SQL, Excel, Tableau, Domain Knowledge

Data Scientist

- Uses advanced machine learning
- Well versed in distributed computing
- Cleans, massages and organises data
- Design and run A/B tests
- Create predictive models

R, Python, SQL, Spark, Statistics, Communication

Analytics Architect

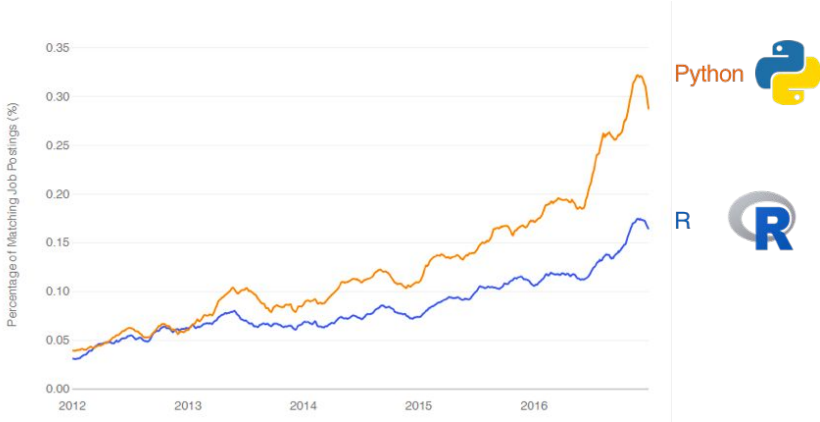
- Uses deep analytics experience
- Shapes requirements with clients - Translates requirements into analytics services
- Good understanding of data engineering

R, Python, SQL/No-SQL, Spark

Python vs R

Anaconda is a python distribution that contains hundreds of packages, many of which are useful for data mining and analysis, such as..

- **NumPy:** objects for multidimensional arrays and matrices
- **SciPy:** algorithms & high-level commands for manipulating, visualizing data
- **Pandas:** Data structures and tools for shaping, merging, and slicing datasets.
- **Scikit-learn (sklearn):** common machine learning and data mining tasks
- **Matplotplib:** standard Python library for creating 2D plots and graphs



Where to get started?

- Python programming
  - Corey Schafer
  - Install Anaconda: https://docs.anaconda.com/anaconda/install/

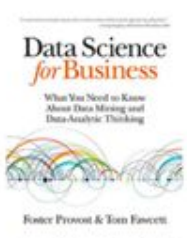
- Basic stats
  - StatQuest with Josh Starmar

Books!!

An introduction to Statistical Learning (ISLR) can be downloaded from - https://statlearning.com/

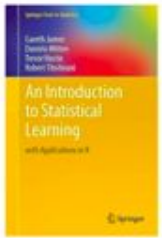
Data Science from Scratch: First Principles with Python
› Joel Grus
54

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and…
› Wes McKinney



Data Science for Business: What You Need to Know about Data Mining and…
› Foster Provost

Reproducible Research with R and R Studio, Second Edition…
Christopher Gandrud

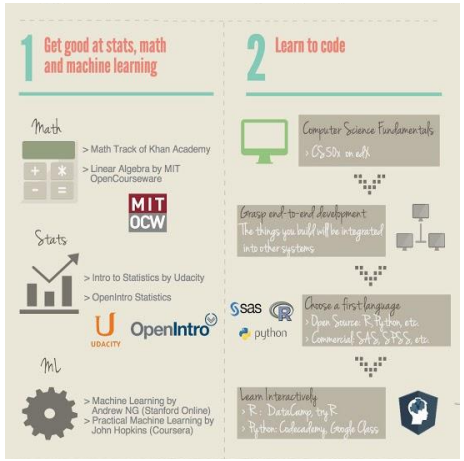


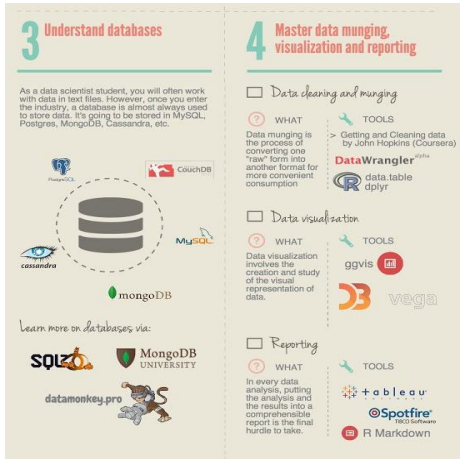
An Introduction to Statistical Learning: with Applications in R…
› Gareth James

Data Smart: Using Data Science to Transform Information into Insight
› John W. Foreman

The Statistical Sleuth: A Course in Methods of Data Analysis
Fred Ramsey

8 easy steps to start data science journey





Thank You!!

Please fill the class survey - https://ga.co/sgintro