علم داده ۲

جلسه صفر - مسیر تبدیل شدن به DATA SCEINTIST

DATA SCIENCE 2

HOW TI BECOME A DATA SCIENTIST

Reza Shokrzad

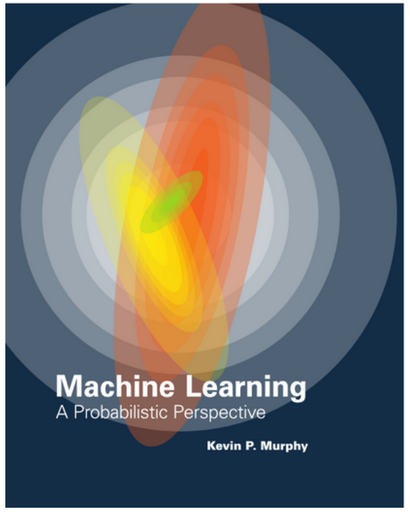
Today

- References
- Recap
- Course Organization
- Data Science in Business
- Data Roles
- Git
- Spark
- Cloud
- A/B testing
- Dashboards
- ML Backgrounds

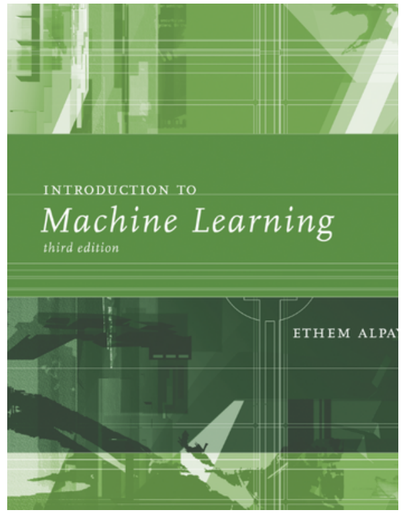
References

References (Traditional ML)

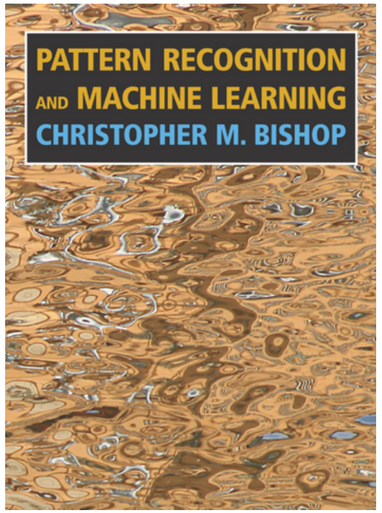
- Machine Learning: A Probabilistic Perspective
- Introduction to Machine Learning
- Pattern Recognition and Machine Learning
- Andrew Ng Coursera Course



Kevin P. Murphy



Ethem Alpaydın



Chris Bishop

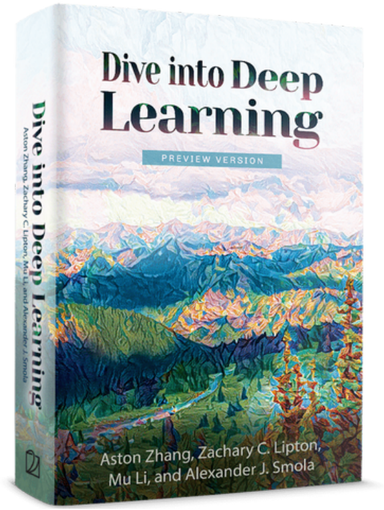
Machine Learning

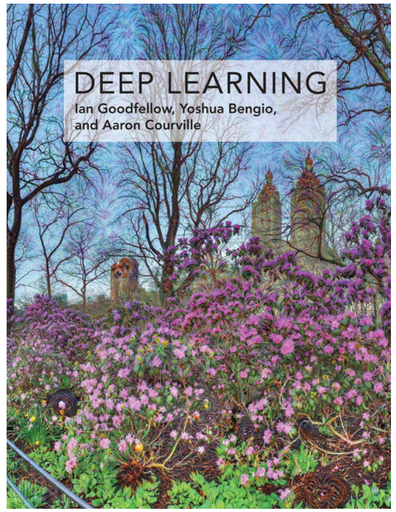
by Andrew Ng

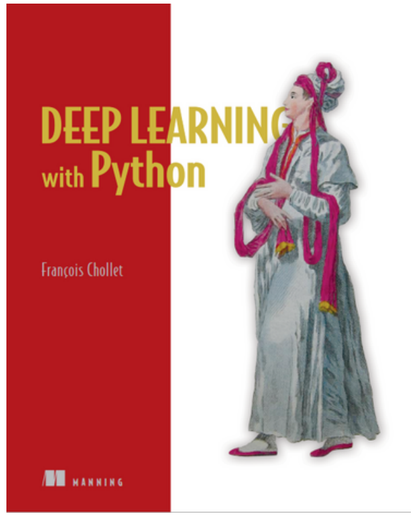
coursera

References (Neural Networks)

- Dive into Deep Learning
- Deep Learning
- Deep Learning with Python
- Andrej Karpathy Course







Aston Zhang

Ian Goodfellow

FRANÇOIS CHOLLET

http://d2l.ai/

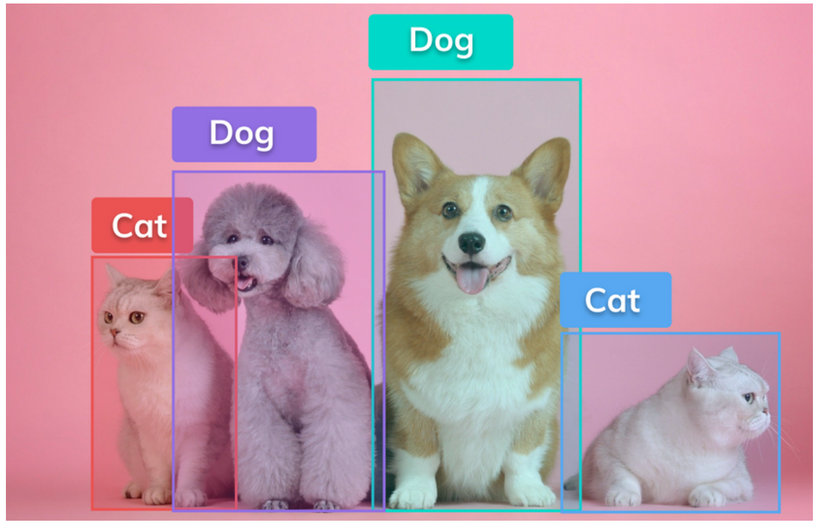
Tasks & Tools

What we already covered in DS 1

- **Learning Types:**
  - Supervised (Regression, Classification)
  - Unsupervised (Clustering)
  - Reinforcement Learning
- **Concepts:**
  - Cost Function
  - Optimization Algorithms (e.g. Gradient Descent)
  - Overfitting (L1/L2 Regularization)
  - Evaluation Metrics (R2, Confusion Matrix)
  - Neurual Networks (ANN, FC, CNN)
- **Packages:**
  - Pandas
  - Numpy
  - Matplotlib & Seaborn
  - Sklearn

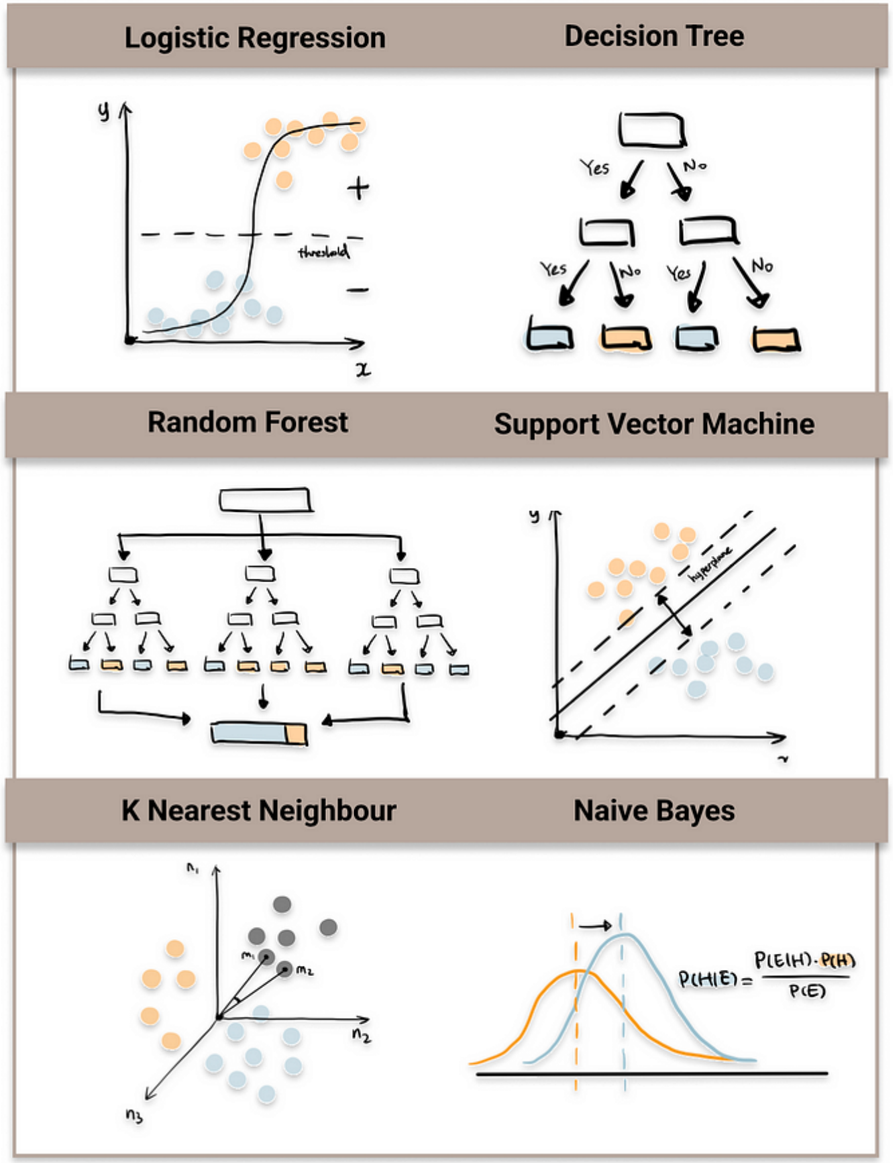
What we will cover in DS 2

- Supervised Learning:
  - Classification
  - Regression
  - Object Detection
  - Image Segmentation
- Unsupervised Learning:
  - Clustering
  - Dimensionality Reduction
  - Anomaly Detection
  - Generation
- Reinforcement Learning



Traditional Supervised Mehotds

- Linear Regression (Regression)
- Logistic Regression (Classification)
- K-Nearest Neighbors (KNN) (Classification, Regression)
- Naive Bayes (Classification, Regression)
- Support Vector Machines (SVM) (Classification, Regression)
- Decision Trees (Classification, Regression)
- Random Forests (Classification, Regression)
- Gradient Boosted Trees (e.g., XGBoost) (Classification, Regression)



Traditional UnSupervised Mehotds

- Principal Component Analysis (PCA) (Dimensionality Reduction)
- UMAP (Dimensionality Reduction)
- tSNE (Dimensionality Reduction)
- K-Means (Clustering)
- DBSCAN (Clustering)
- Agglomerative (Clustering)

Traditional UnSupervised Mehotds

- Z-Score (Anomaly Detection)
- Isolation Forest (Anomaly Detection)
- One-Class SVM (Anomaly Detection)
- Local Outlier Factor (LOF) (Anomaly Detection)
- Collaborative Filtering (Recommender Systems)
- Content-Based Filtering (Recommender Systems)

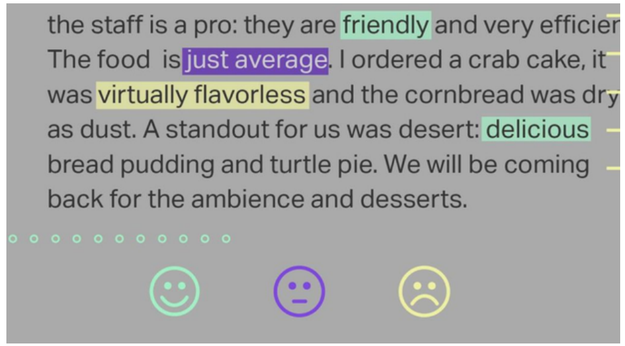
Neural Networks

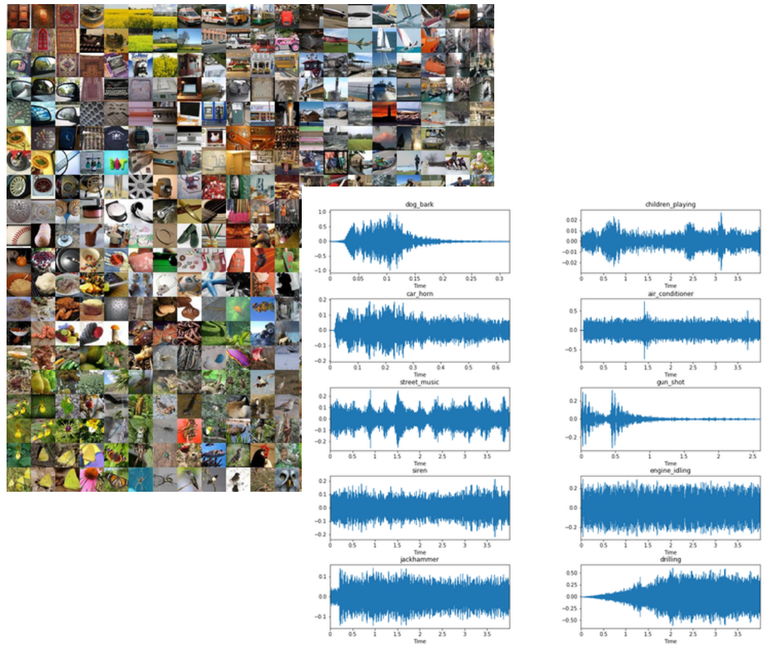
Convolutional Neural Networks (CNN)	Classification, Segmentation, Detection
Recurrent Neural Networks (RNN)	Sequence Prediction, Time Series Analysis, Text Generation
Long Short-Term Memory (LSTM)	Sequence Prediction, Time Series Analysis, Text Generation
Generative Adversarial Networks (GAN)	Generation
Transformer Networks (e.g., BERT, GPT)	Classification, Generation, Sequence Prediction
Autoencoders	Dimensionality Reduction, Generation, Anomaly Detection
Unet	Segmentation
Region-based CNN (R-CNN)	Detection, Localization
Fast R-CNN, Faster R-CNN	Detection, Localization
Single Shot MultiBox Detector (SSD)	Detection
YOLO (You Only Look Once)	Detection, Localization
Mask R-CNN	Detection, Segmentation

Data Types

- Numerical
- Categorical
- Textual (or String)
- Time Series
- Image
- Audio, Signal, Speech
- Video
- Spatial (or Geospatial)
- Graph/Network

Data set	# Features	# Classes	Number of training instances	Number of testing instances	Data type
Iris	4	3	75	75	Continuous
Liver	6	2	173	172	Continuous
Breast	9	2	350	349	Continuous
Echo	11	2	66	66	Continuous
Va-Heart	13	2	100	100	Continuous
Wine	13	3	89	89	Continuous
All-hyper	29	3	486	486	Continuous
Lung	56	3	18	15	Continuous
Soy	208	17	150	139	Binary
Promoter	228	2	53	53	Binary





Language Programming?

- Python (80%)
- SQL (MySQL) (10%)
- R (10%)

Python Frameworks

Course content

- 16 Prerequisite
- 40 Online sessions
- 3 minor quiz
- 1 final quiz

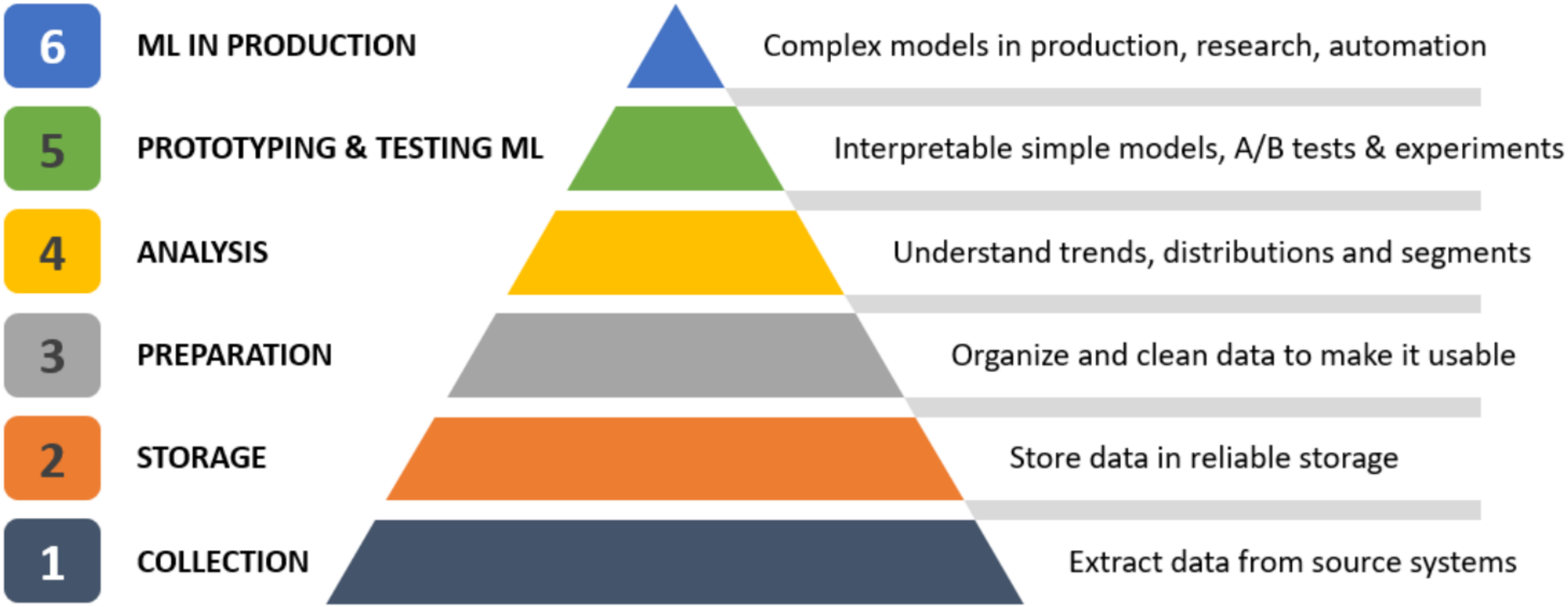
Prerequisite

1	پیش مطالعه شماره 1: پردازش متن (مقدمات)
2	پیش مطالعه شماره 2: پردازش تصویر (مقدمات)
3	پیش مطالعه شماره 3: مقدمات پایتورچ
4	پیش مطالعه شماره 4: مقدمه‌ای بر کتابخانه تنسورفلو
5	پیش مطالعه شماره 5: مرور مقدمات یادگیری ماشین و عمیق
6	پیش مطالعه شماره 6: پیش پردازش‌های لازم برای پکیج کراس
7	پیش مطالعه شماره 7: مقدمه پردازش صوت (تشخیص گفتار)
8	پیش مطالعه شماره 8: مقدمه تصاویر پزشکی - medical imaging

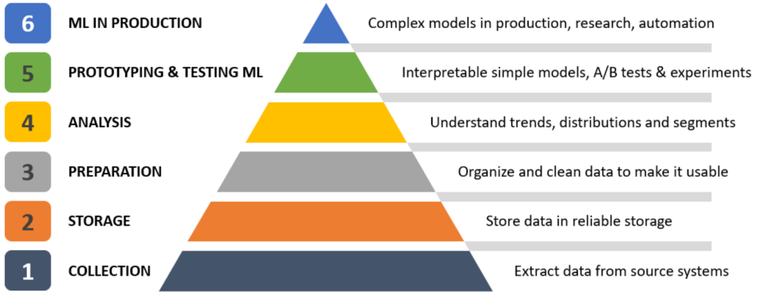
9	پیش مطالعه شماره 9: شبیه سازی مونتی کارلو - onte Carlo Simulation
10	پیش‌مطالعه شماره 10: مقدار ویژه، بردار ویژه و تجزیه منفرد ماتریس
11	پیش‌مطالعه شماره 11: برنامه‌نویسی شئی گرا
12	پیش‌مطالعه شماره 12 بخش اول
13	پیش‌مطالعه شماره 12 بخش دوم
14	پیش مطالعه شماره 13 : پایگاه داده 1
15	پیش مطالعه شماره 14 : پایگاه داده 2
16	پیش مطالعه شماره 15 : پایگاه داده 3

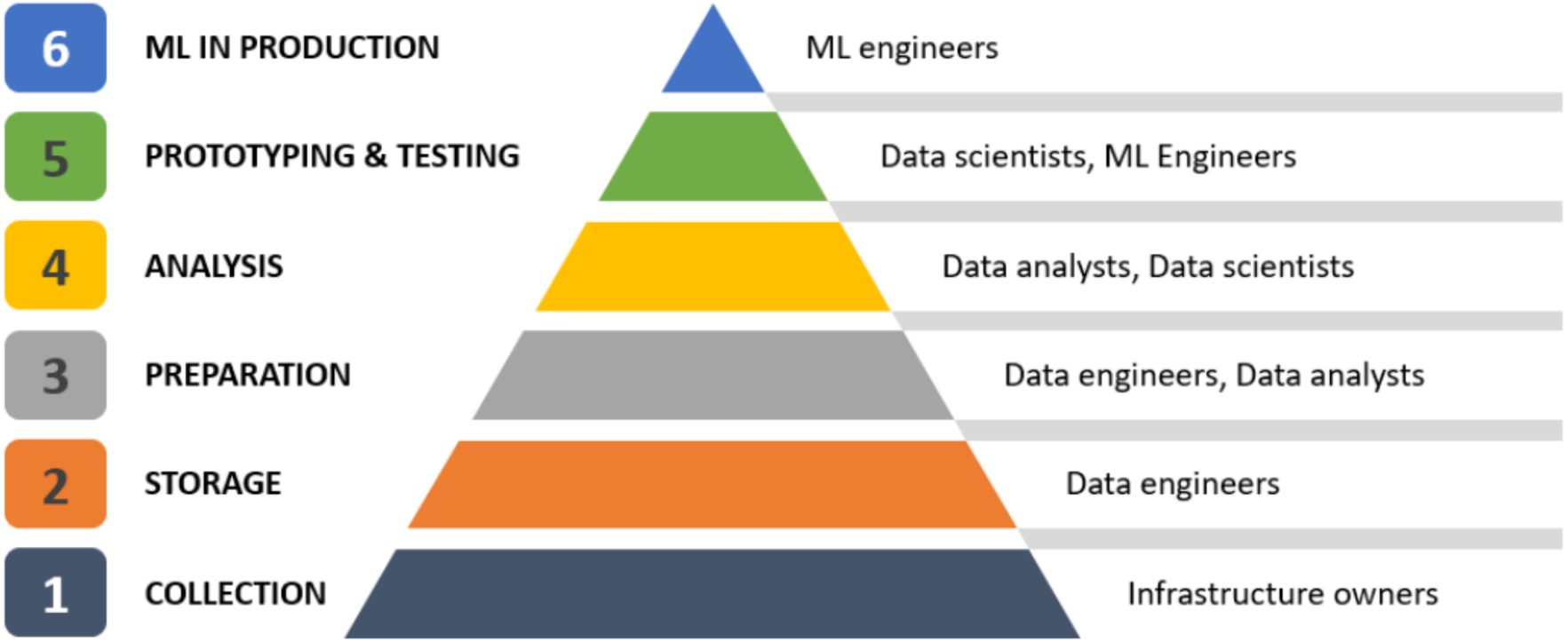
Topics in DS Context

Data Governance



Job Role


6
ML IN PRODUCTION
Complex models in production, research, automation
5
PROTOTYPING & TESTING ML
Interpretable simple models, A/B tests & experiments
4
ANALYSIS
Understand trends, distributions and segments
3
PREPARATION
Organize and clean data to make it usable
2
STORAGE
Store data in reliable storage
1
COLLECTION
Extract data from source systems


6
ML IN PRODUCTION
ML engineers
5
PROTOTYPING & TESTING
Data scientists, ML Engineers
4
ANALYSIS
Data analysts, Data scientists
3
PREPARATION
Data engineers, Data analysts
2
STORAGE
Data engineers
1
COLLECTION
Infrastructure owners

Team Members

Information architects
Build storage solutions
Maintain data access

Tools:

- SQL
  - Storing large quantities of data
- Java, Scala, or Python
  - Programming languages for processing data and automating tasks

Creating dashboards
Hypothesis testing
Data visualization

Tools:

- Spreadsheets (Excel or Google Sheets)
  - Simple storage and analysis
- SQL
  - Large-scale analysis
- BI Tools (Tableau, Power BI, Looker)
  - Dashboarding and sharing information

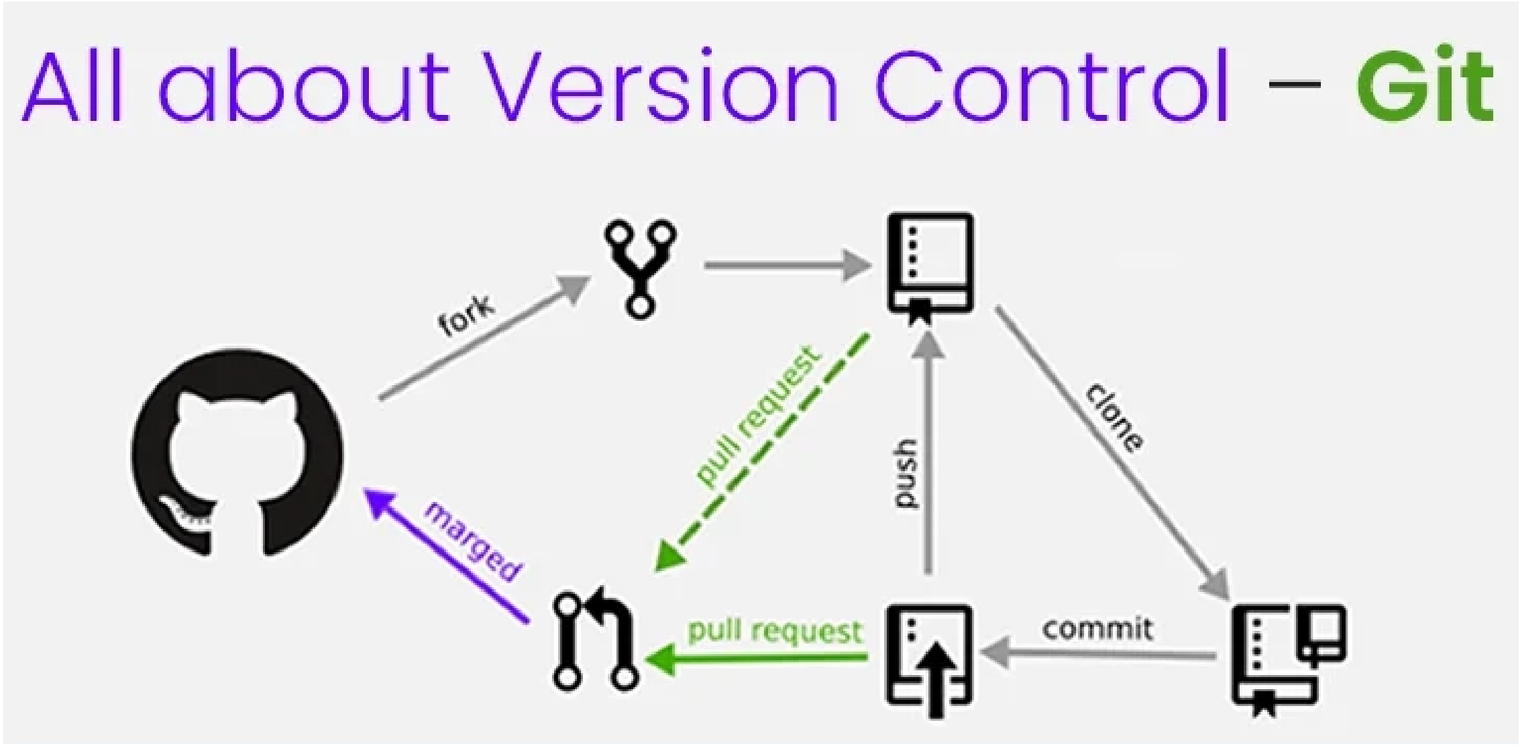
Predictions and extrapolations
Classication
Stock price prediction
Image processing
Automated text analysis

Tools:

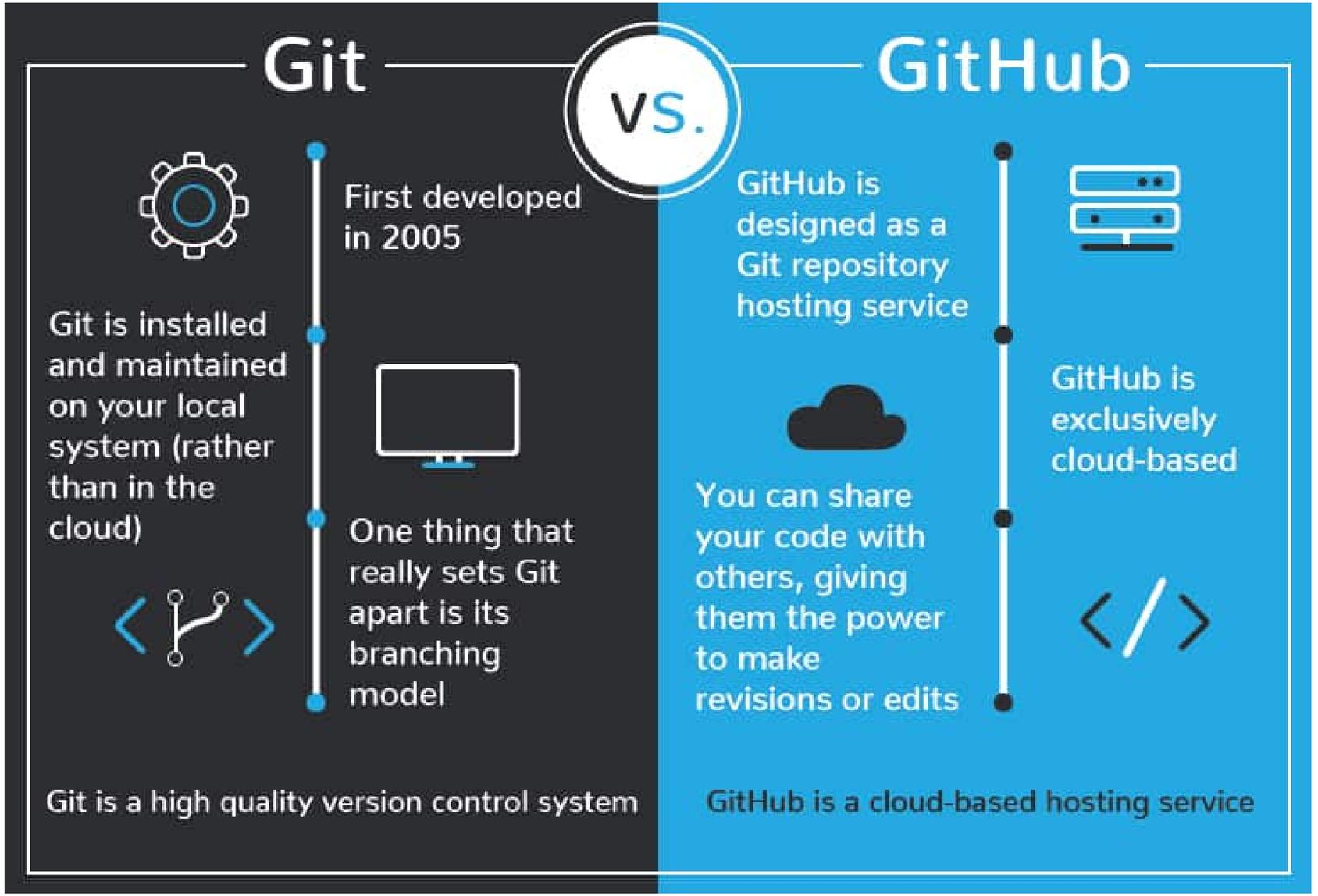
- Python and R
  - Programming languages for creating predictive models

Git / GitHub

Git is software for tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows.



Git / GitHub


Git
vs.
GitHub
First developed in 2005
Git is installed and maintained on your local system (rather than in the cloud)
One thing that really sets Git apart is its branching model
GitHub is designed as a Git repository hosting service
GitHub is exclusively cloud-based
You can share your code with others, giving them the power to make revisions or edits
Git is a high quality version control system
GitHub is a cloud-based hosting service

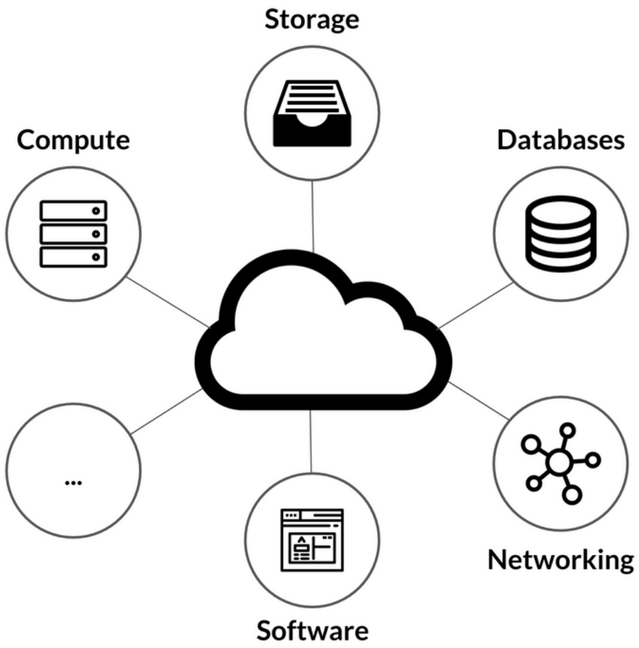
Spark

Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance.

PySpark is an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively analyzing your data in a distributed environment.

Cloud Computing

Cloud computing is the delivery of technology services - including compute, storage, databases, networking, software, and many more - over the internet with pay-as-you-go pricing.



aws

Microsoft Azure

Google Cloud

Cloud vs On-premise

On-premise

Less scalable

Takes time to set up

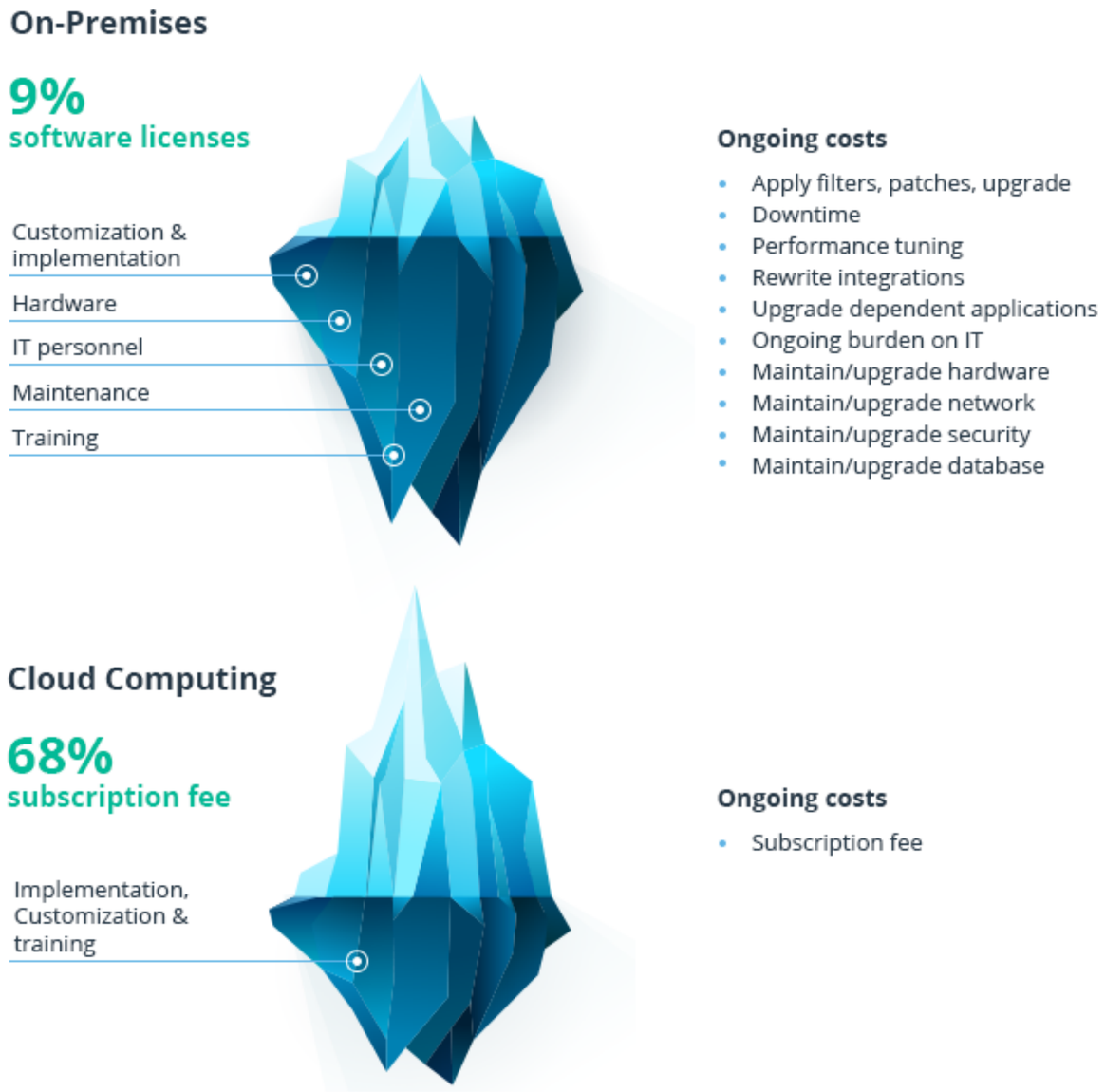
Ongoing costs

Cloud

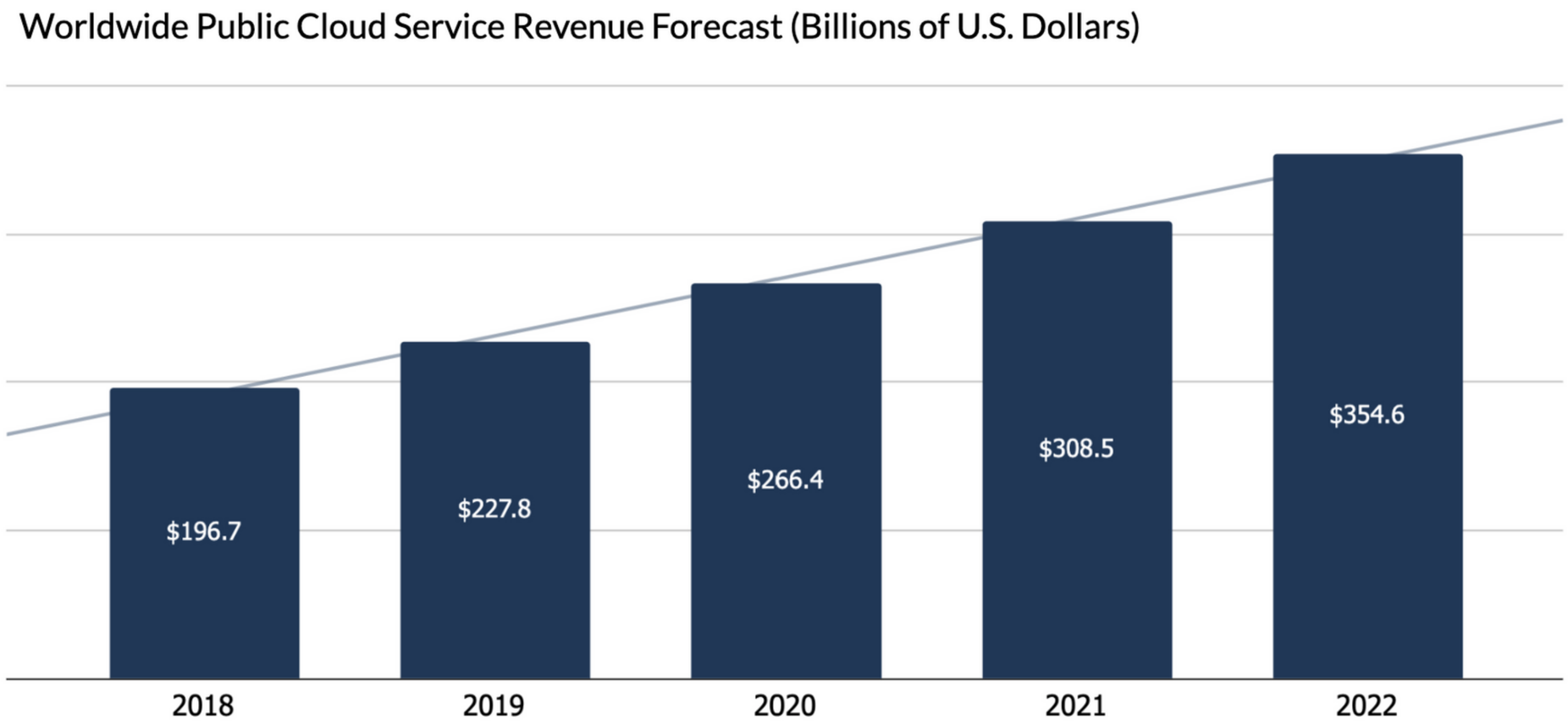
Fast set-up speed

Scalable

Pay-as-you-go



Cloud Computing

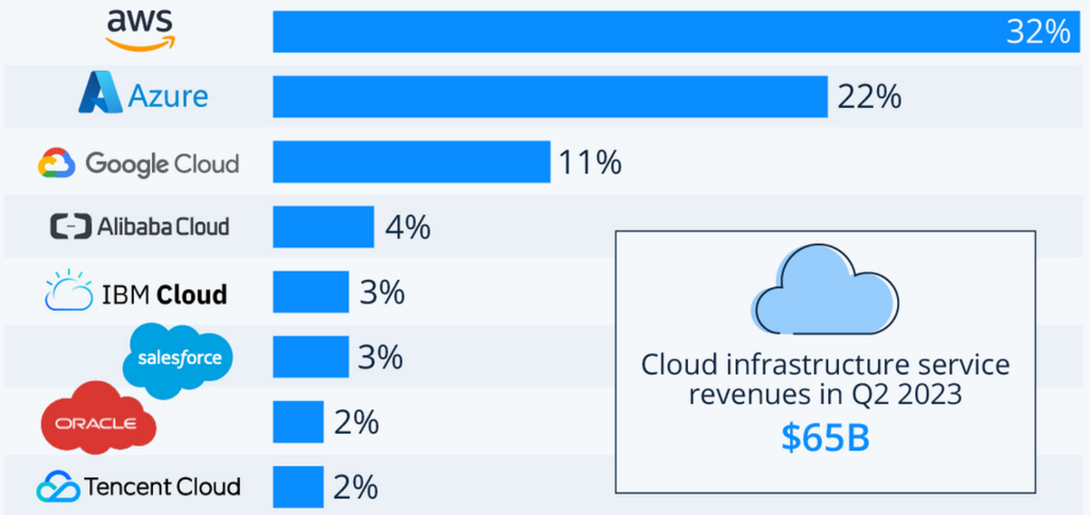


Source: https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020

Cloud Computing

Amazon Maintains Lead in the Cloud Market

Worldwide market share of leading cloud infrastructure service providers in Q2 2023*



* Includes platform as a service (PaaS) and infrastructure as a service (IaaS) as well as hosted private cloud services

Source: Synergy Research Group

Team Structure

- **Centralized**
  - all data functions in one central team. Works well for small companies, startups, new organizations. Gets slow once business matures and requires focus

- **Decentralized**
  - each business unit, geography or department have their own data functions. Works well for larger companies. Introduces issues with data governance, dierences in denitions, redundancies, and added complexity

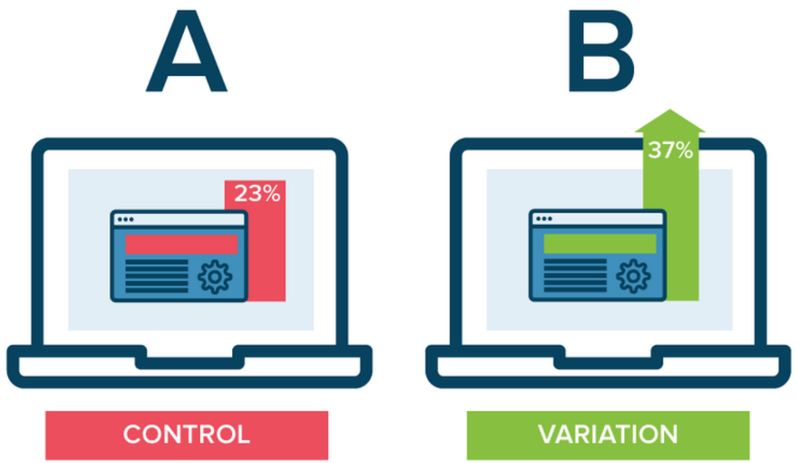
- **Hybrid**
  - infrastructure, definitions, methods and tooling are centralized, while application and prototyping decentralized

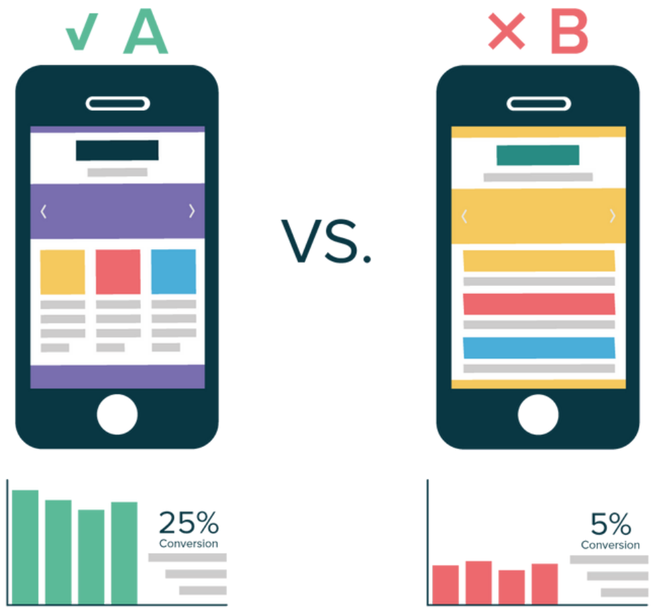
Ethics and Law in Cyberspace

Cyber ethics is the study of ethics pertaining to computers, covering user behavior and what computers are programmed to do, and how this affects individuals and society.

Cyberspace law generally encompasses various legal issues involving the communication, distribution, and transactions that occur over the internet and other types of networked devices and technologies.

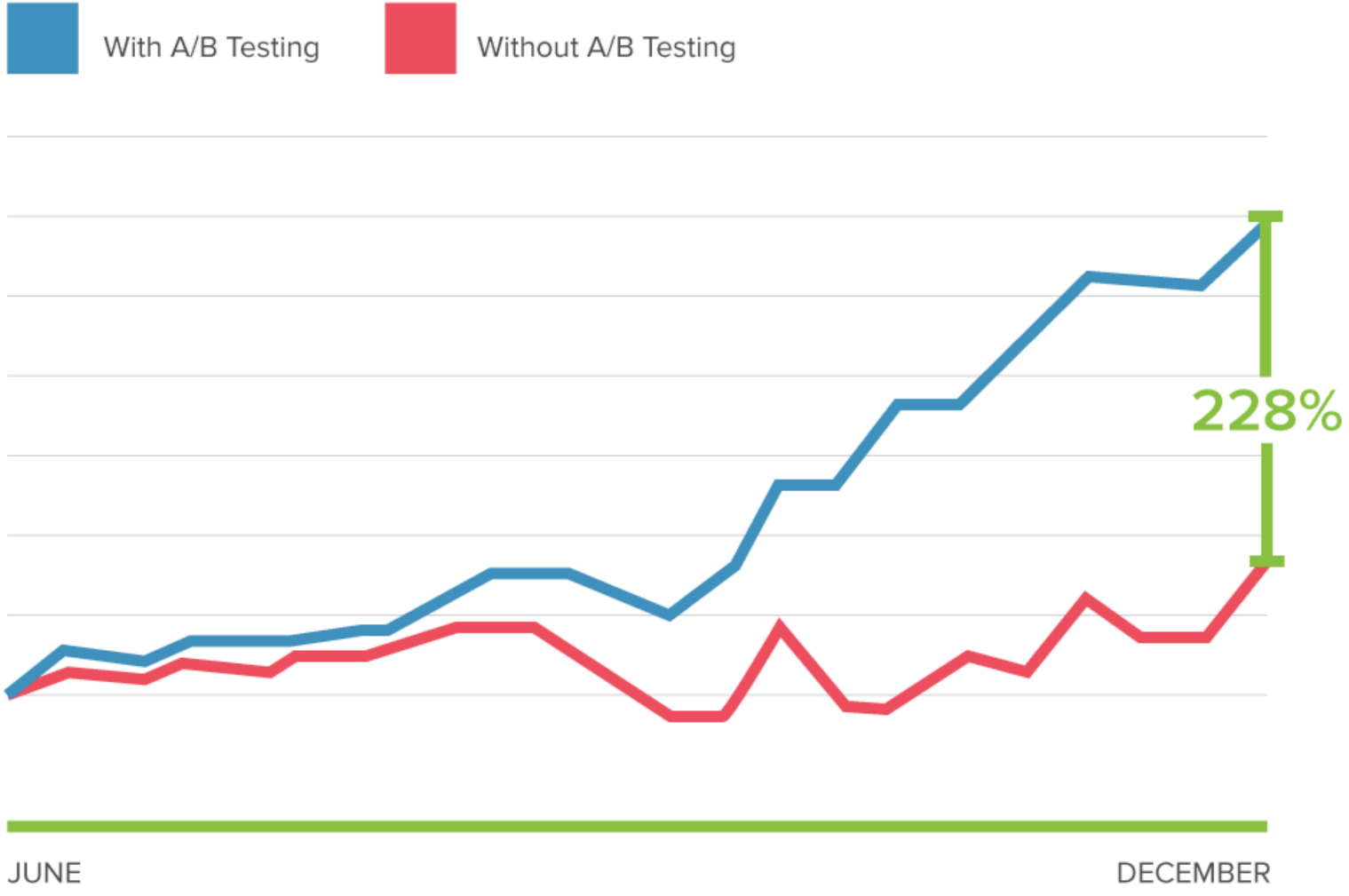
A/B Testing

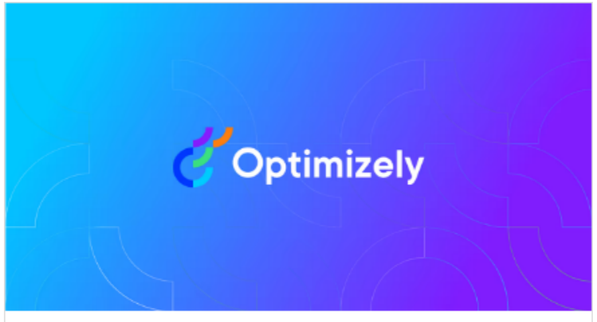




A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Why A/B Testing



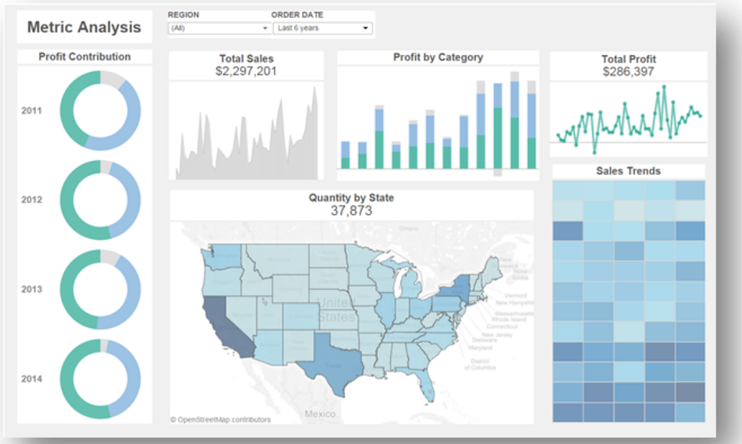


What is A/B testing? With examples

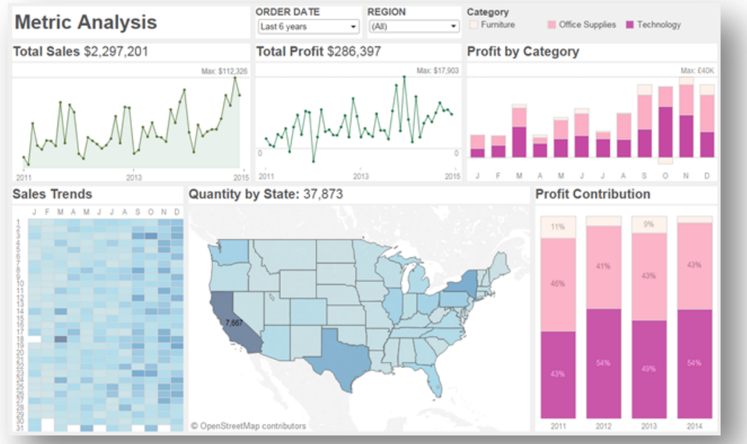
A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

Dashboards

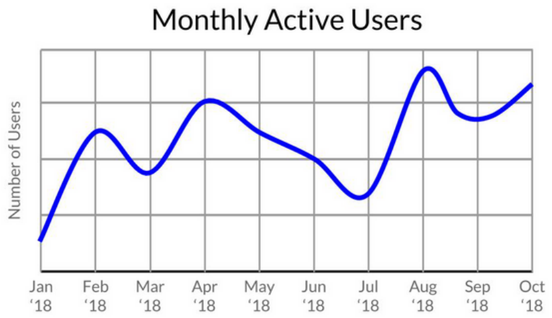
Before



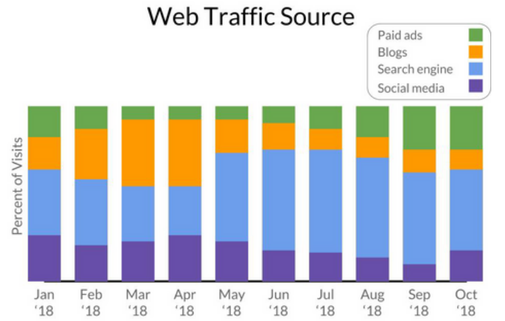
After



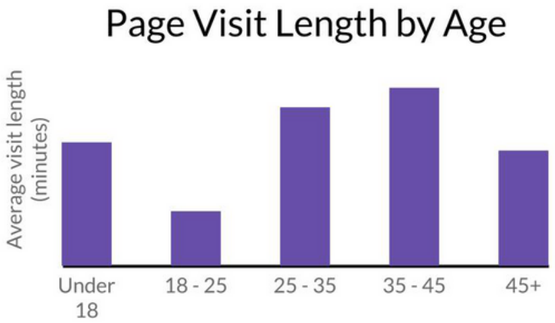
Tracking a value over time



Tracking composition over time



Categorical comparison



Highlighting a single number

3,495,108

visits today

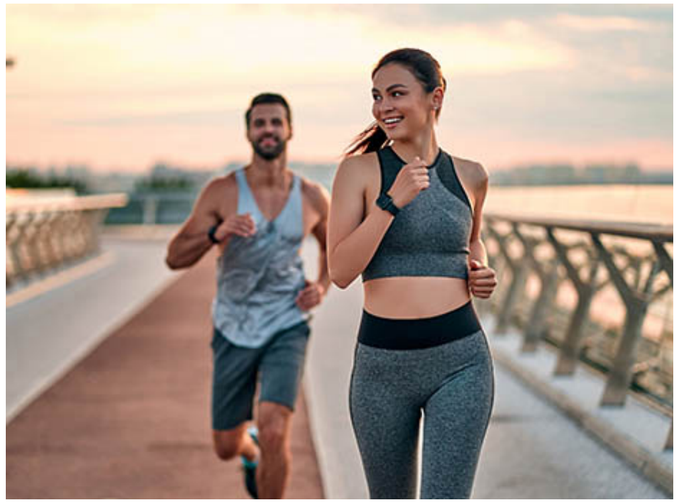
8% from yesterday

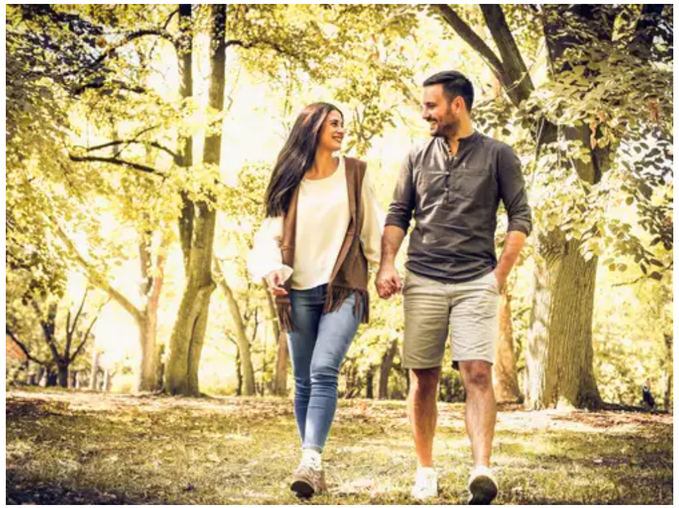
Machine Learning

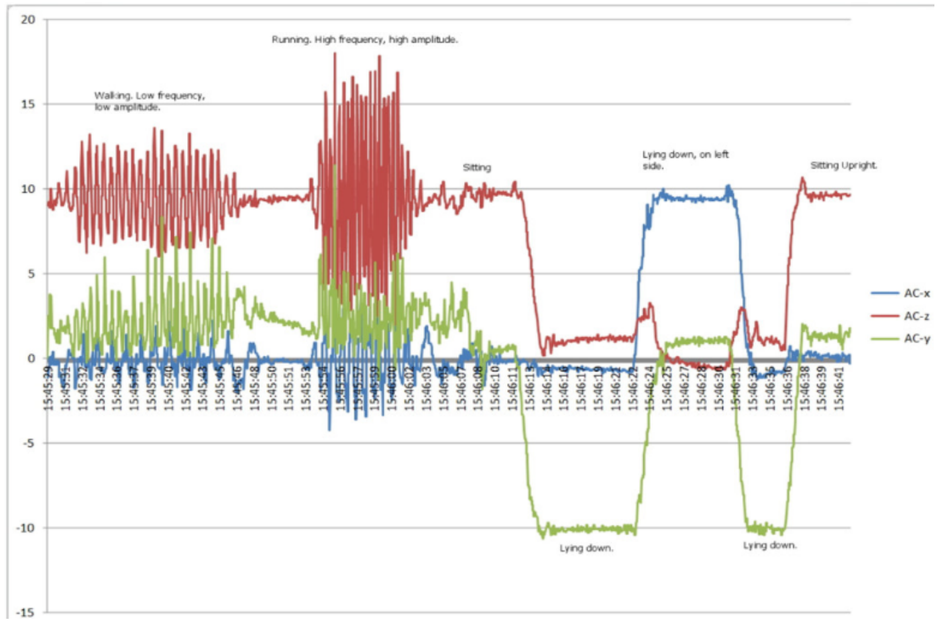
What do we need for machine learning?

- A well-defined question
  - "What is the probability that this transaction is fraudulent?"
- A set of example data
  - Old transactions labeled as "fraudulent" or "valid"
- A new set of data to use our algorithm on
  - New credit card transactions

Case Study - Internet of Things (IOT)







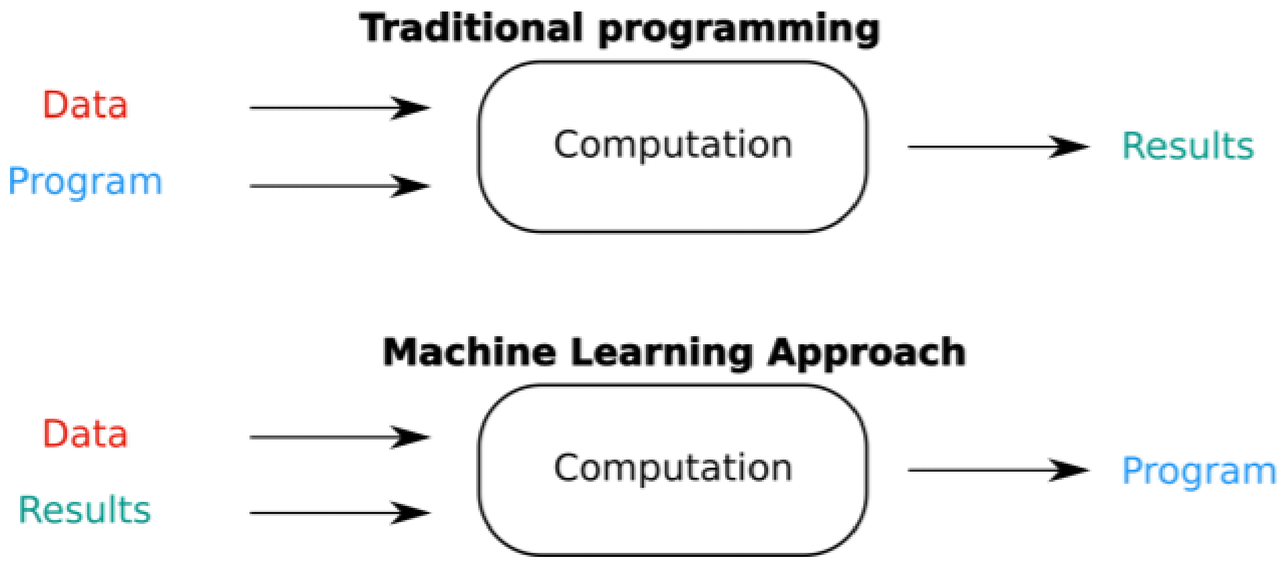
Case Study - Internet of Things (IOT)

- Smart watches
- Internet-connected home security systems
- Electronic toll collection systems
- Building energy management systems
- Much, much more!

Case study - Image Recognition

1	1	1	1	1	1	1	1	2	2	3	3	3	1	3	1	1	2
1	1	1	1	1	1	1	1	2	2	3	3	3	1	3	1	1	2
1	1	1	1	1	1	1	1	2	1	2	3	1	3	3	1	1	2
8	8	8	8	8	8	8	8	8	8	8	8	48	20	20	20	8	8
6	6	7	6	6	6	6	6	6	6	6	6	6	20	20	20	8	5
4	4	4	4	4	4	4	4	4	4	4	4	4	20	20	20	5	5
4	4	4	4	4	4	4	5	5	5	5	5	5	5	4	4	4	4
4	4	4	4	5	5	5	5	5	5	4	4	4	4	4	4	4	4
5	5	5	5	5	5	4	4	4	4	4	4	4	4	4	4	4	4

What is Machine Learning

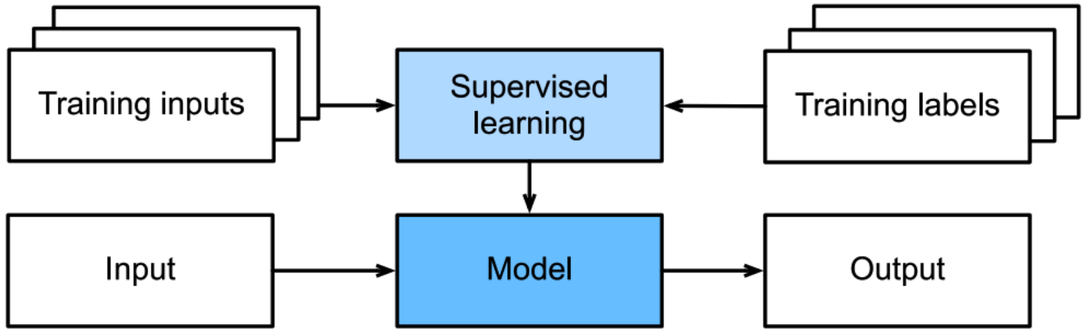


Working with data

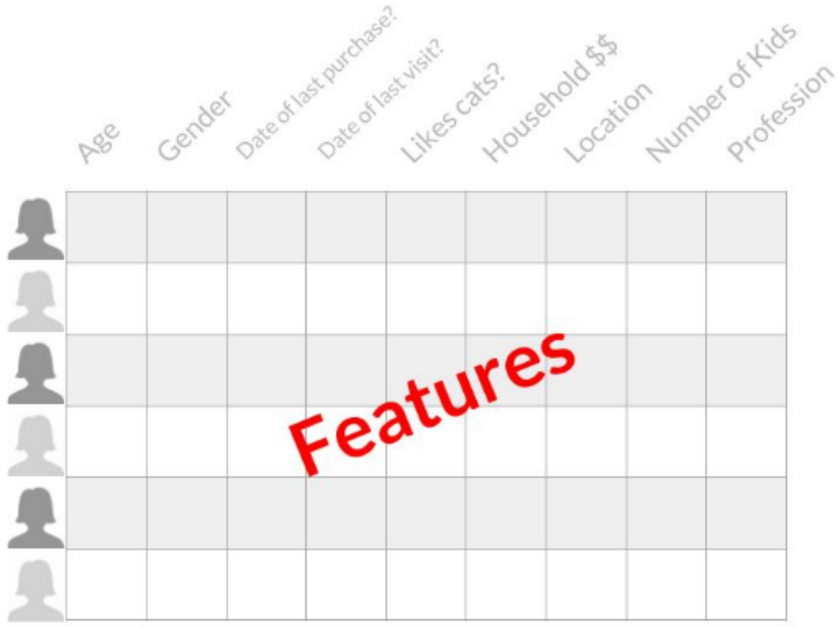
- Training set
  - To train model parameters
- Validation set
  - To choose a model
- Test set
  - To evaluate your model

Machine Learning

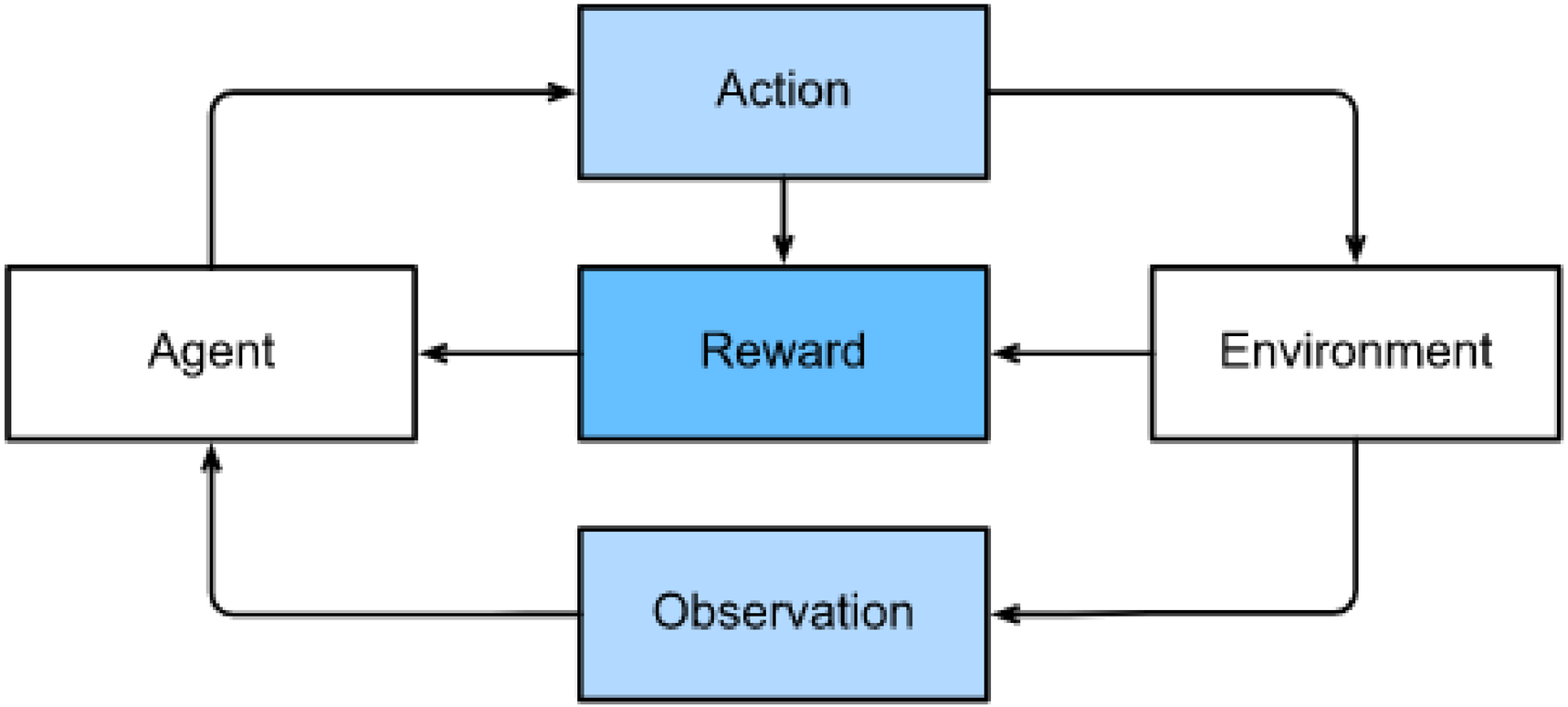
Supervised Machine Learning



Unsupervised Machine Learning



Reinforcement Learning



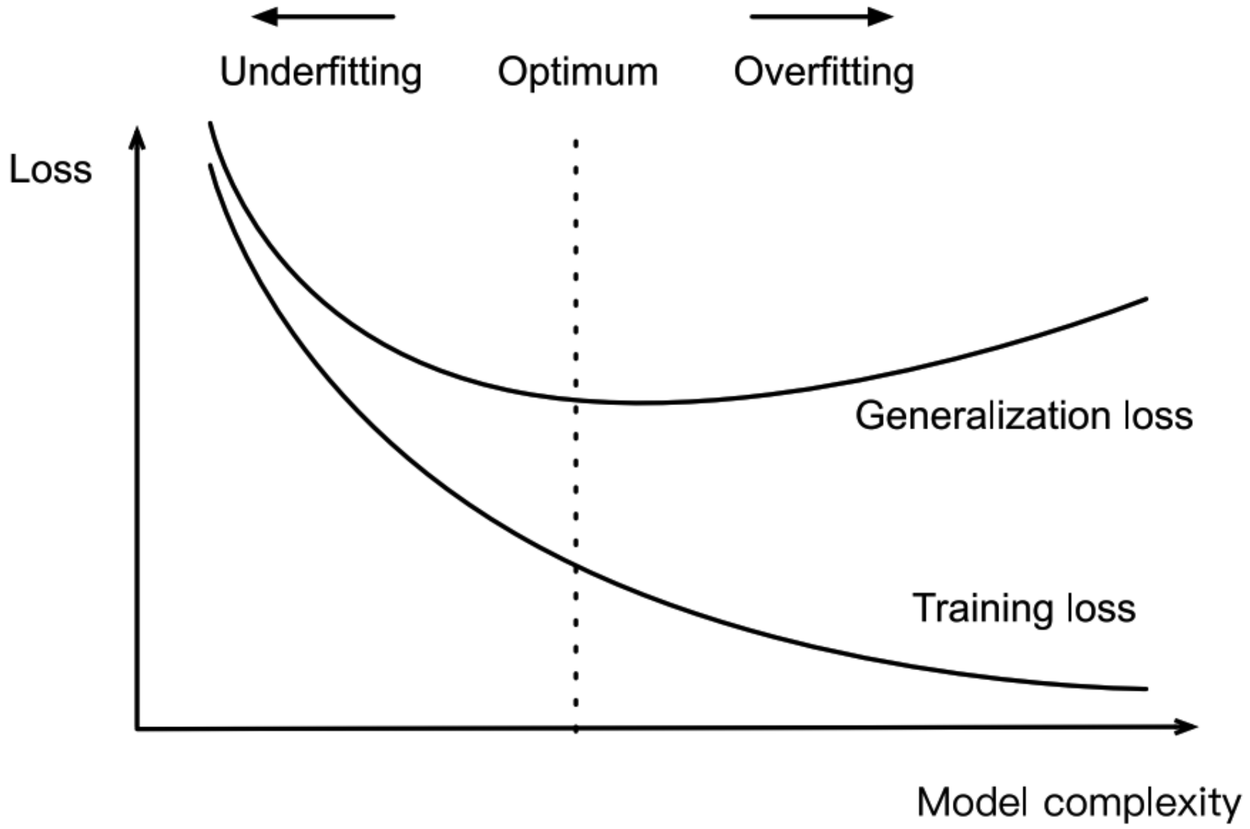
Empirical Risk Minimization

Loss $L(y, \hat{y})$.

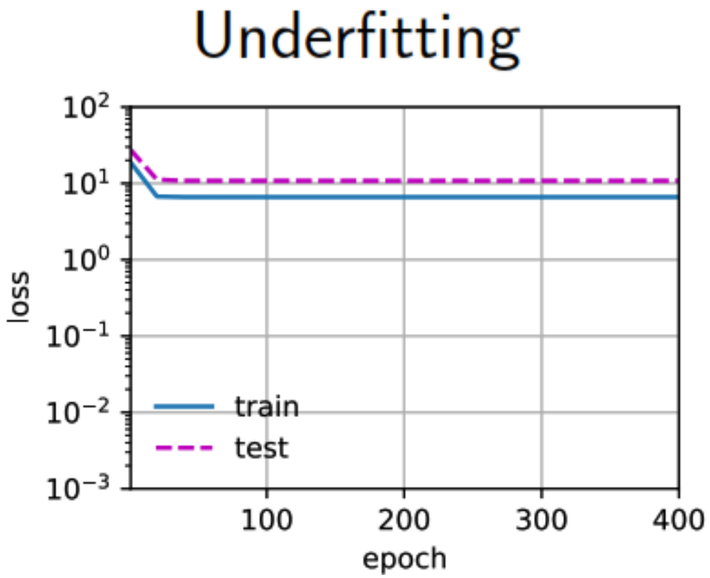
How bad is it to predict $\hat{y}$ when the true answer is y?

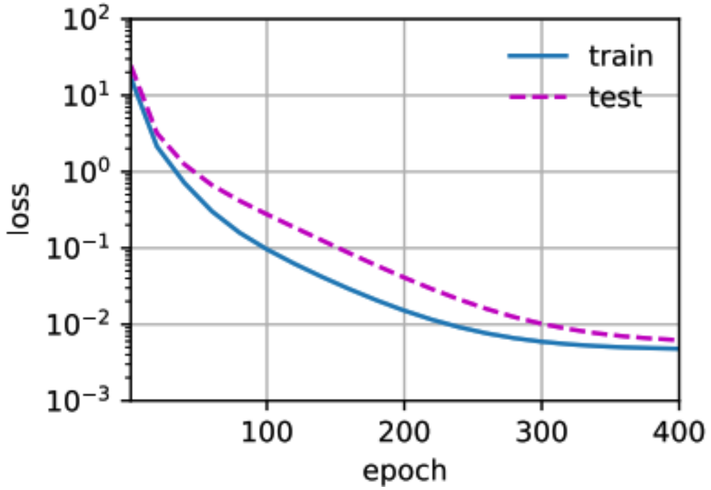
Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

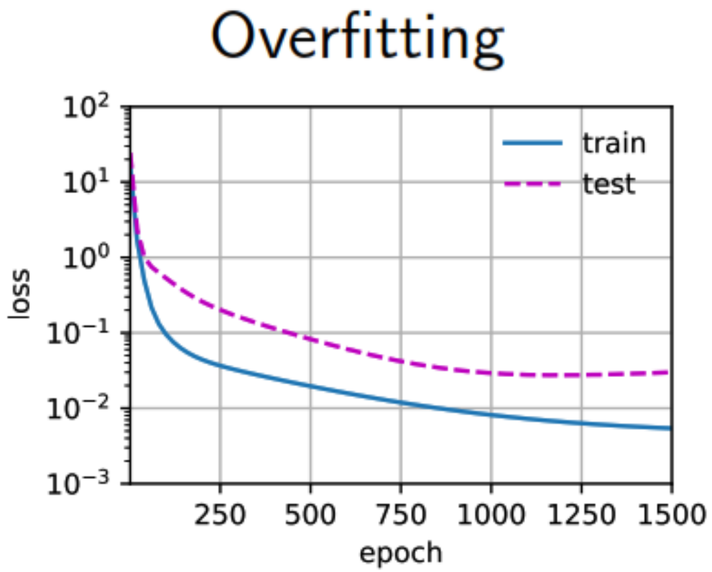
What data occurs in the real world



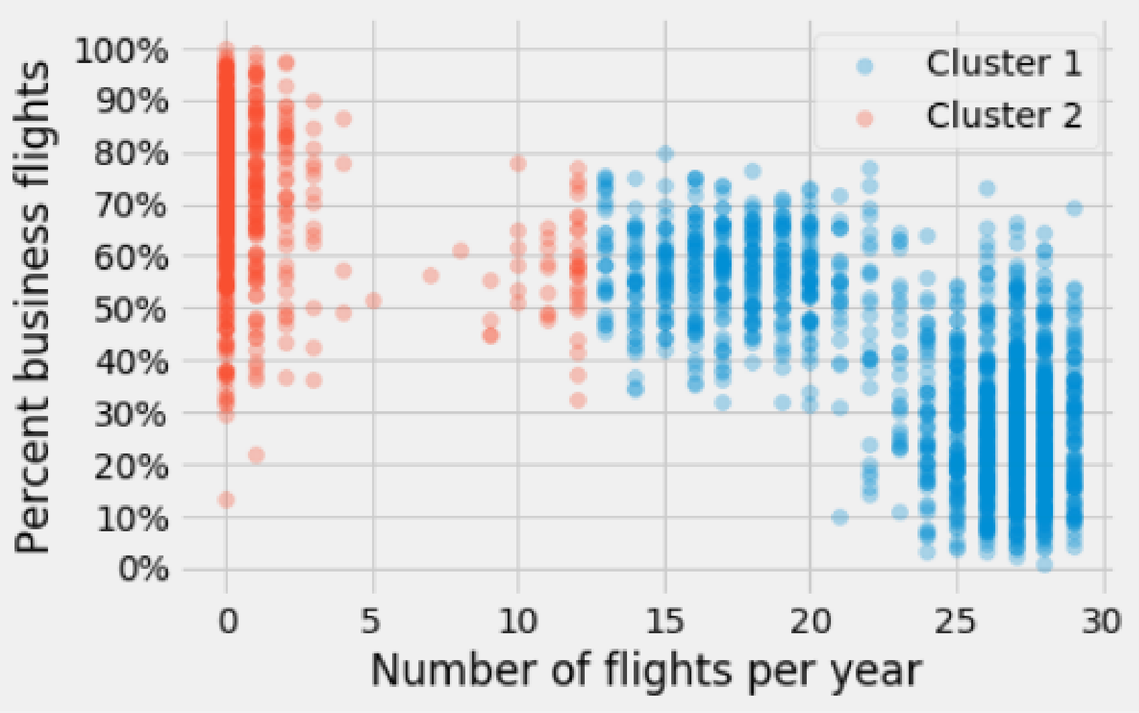
Which fitting?

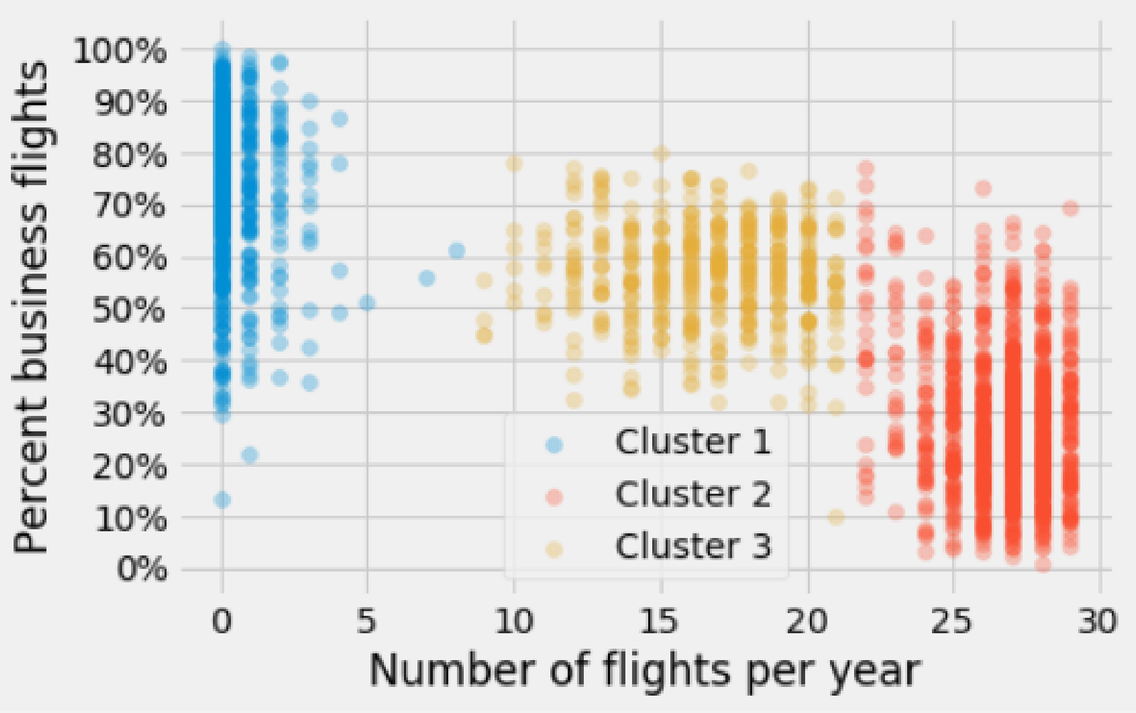






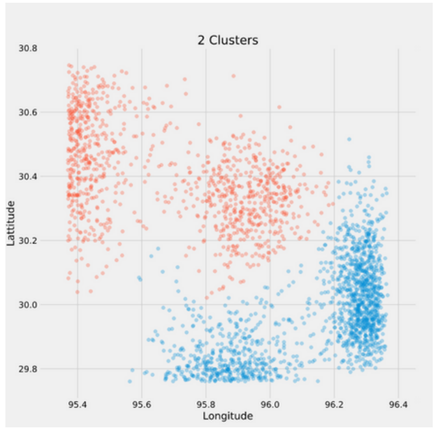
Case study: customer segmentation

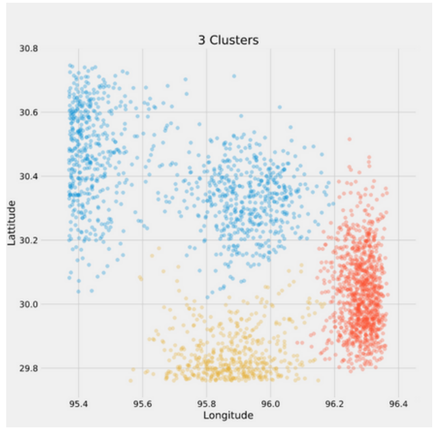


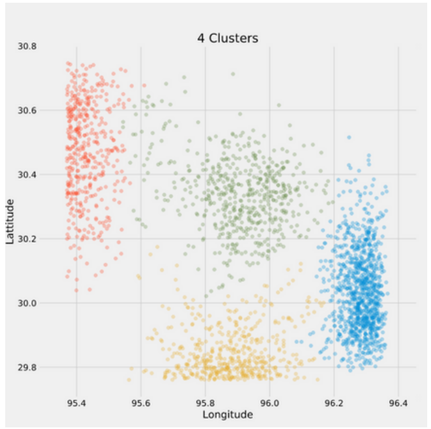


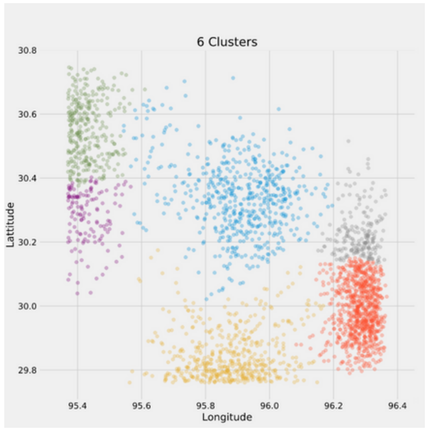
How many Cluster?

Crystal manages operations for a meal delivery service in Houston, Texas. She has latitude and longitude data for meals requests in the past month. She'd like to create several new "delivery hubs" where demand has been greatest. In order to do this, she's asked a Machine Learning Scientist to use a clustering algorithm to divide the orders into groups based on their location.









What is the most suitable number of clusters?

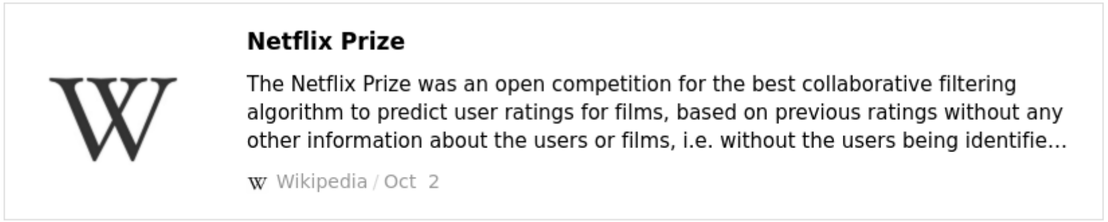
ML examples - Marketing

SUPERVISED Machine Learning:

- Predict which customers are likely to purchase next month
- Predict each customer's expected lifetime value

UNSUPERVISED Machine Learning:

- Group customers into segments based on their past purchases



ML examples - Finance

SUPERVISED Machine Learning:

- Identify key transaction attributes that indicate a potential fraud
- Predict which customers will default on their mortgage payments

UNSUPERVISED Machine Learning:

- Group transactions into segments based on their attributes to understand which segments are the most profitable

ML examples - Manufacturing

SUPERVISED Machine Learning:

- Predict which items in production are likely faulty and should be manually inspected
- Predict which machines are likely to break and need maintenance

UNSUPERVISED Machine Learning:

- Group readings from machine sensors and identify anomalies for potential manufacturing malfunctions

ML examples - Transportation

SUPERVISED Machine Learning:

- Predict the expected delivery of the parcel
- Identify the fastest route for driving
- Predict product demand to prepare enough stock, rent/buy vehicles and hire workers

5-min Quiz

Sup or Unsup?

1. Create groupings of clothing items that have similar features.
2. Divide customers into different categories based on their shopping habits.
3. Predict if a customer will purchase an item based on what previous customer purchased.
4. Predict whether a new clothing style will be successful based on previous session's trends.

5-min Quiz

A/B Test - ML - Dashboard

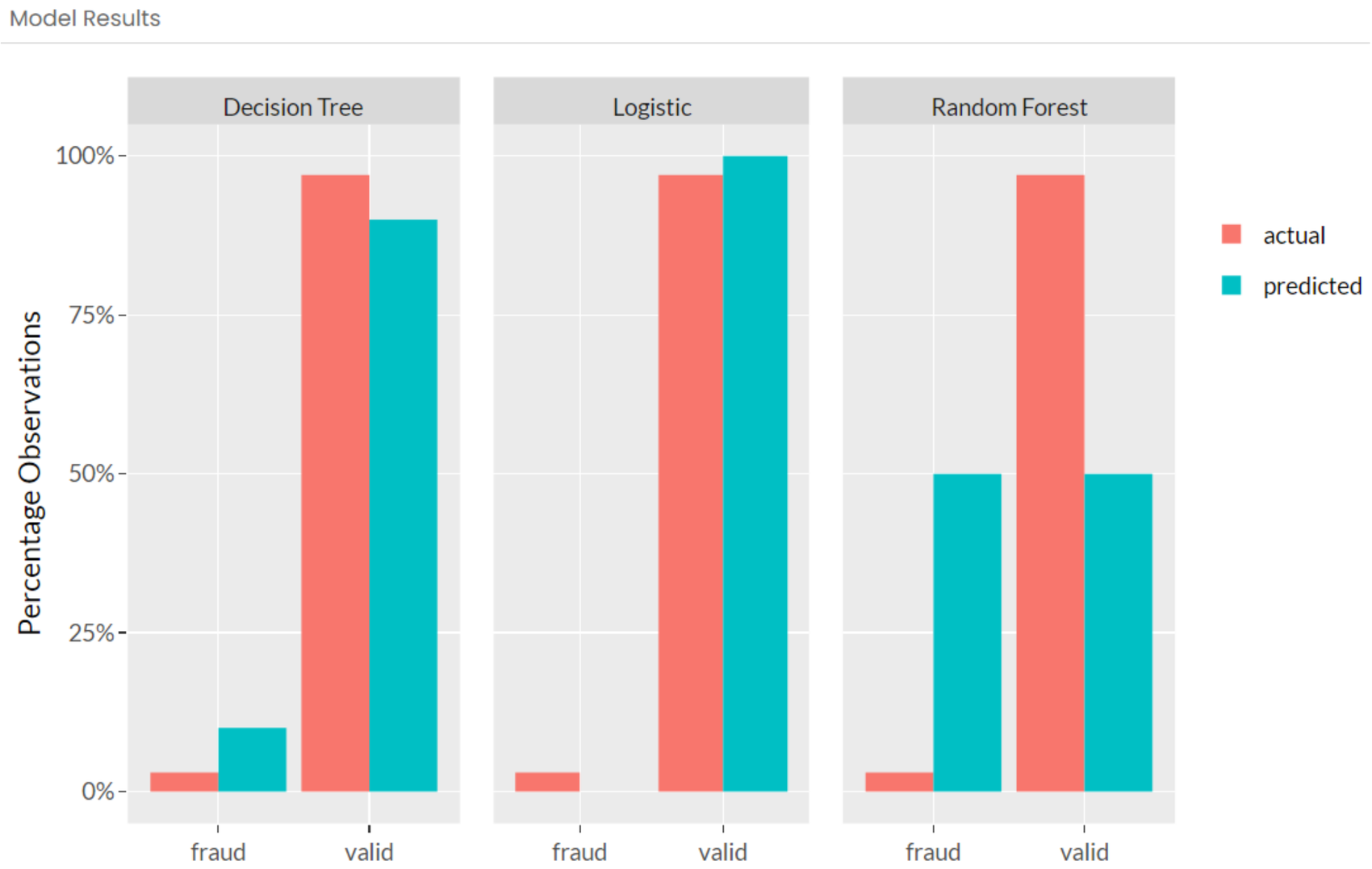
1. Growth team wants to know which banner ad gets the most click.
2. CEO wants a chart correlating customer age with life time value.
3. Finance team wants to know which customer accounts are likely to miss payment based on previous payments and demographics.
4. Customer success team wants to know which customer are likely to churn based on previous purchases and social media interactions.
5. Sales team wants to know which script maximized purchases from calls with new clients.

Recap

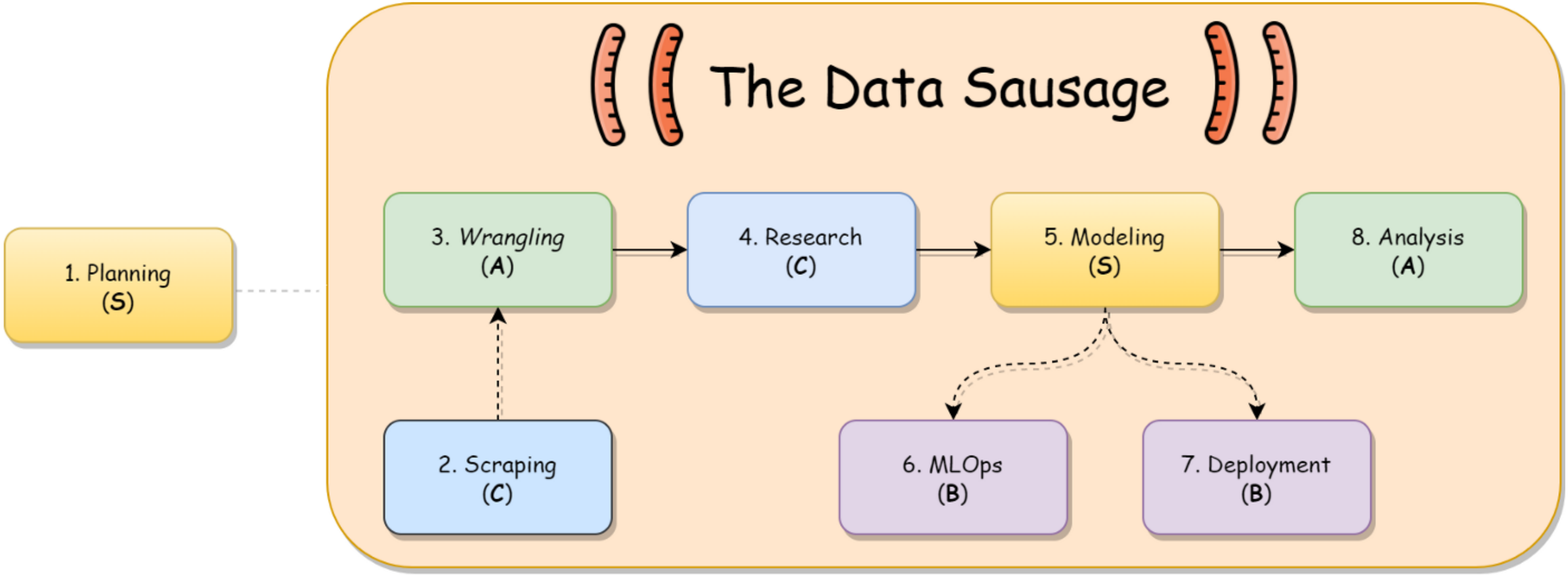
1. A/B tests help us pick between two choices.
2. Machine Learning helps us make predictions using features and labels.
3. Dashboards help us display information.

Question. Which model performs better?

Detecting Fraudulent Bank Transactions.



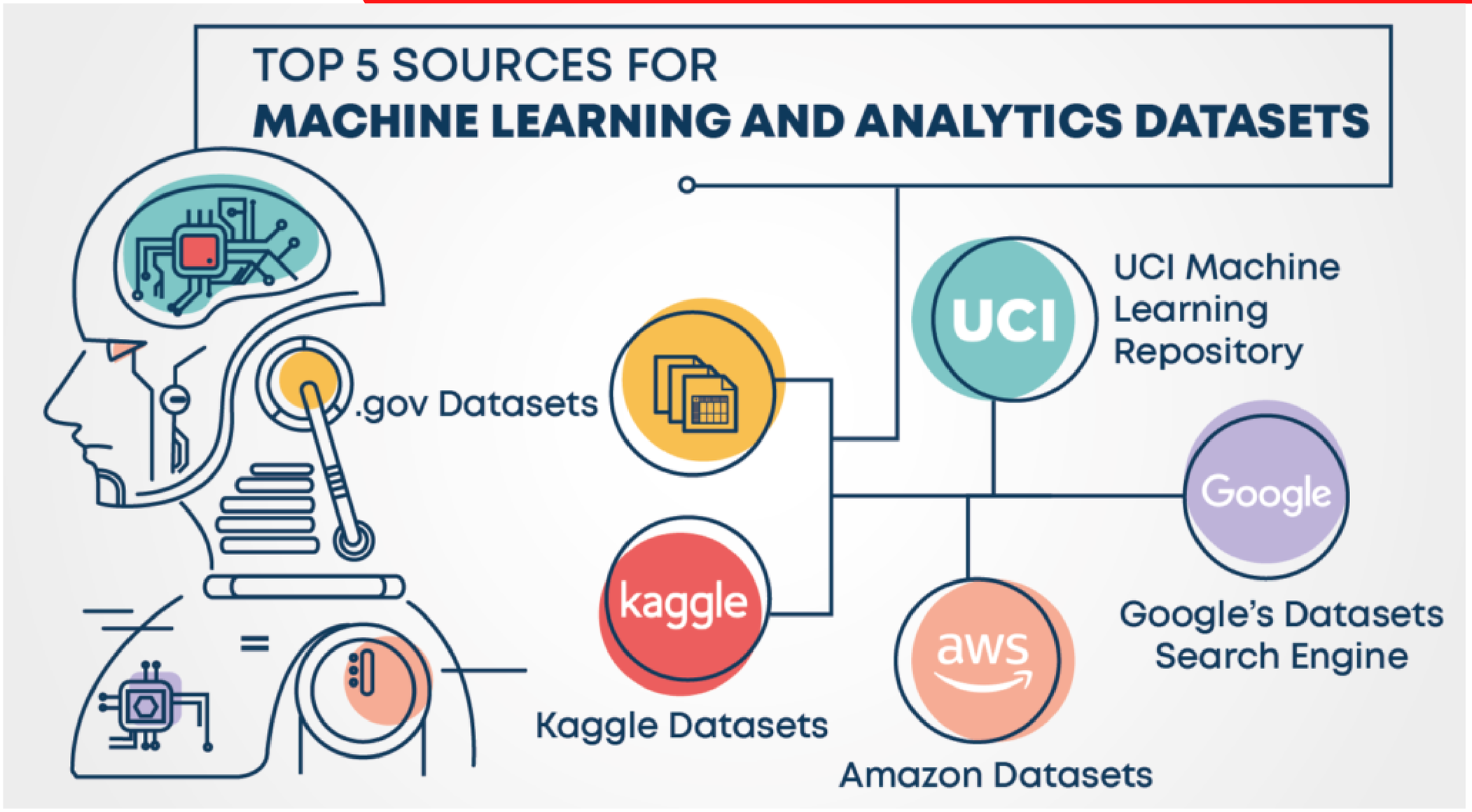
The Skeleton of a Data Science Project



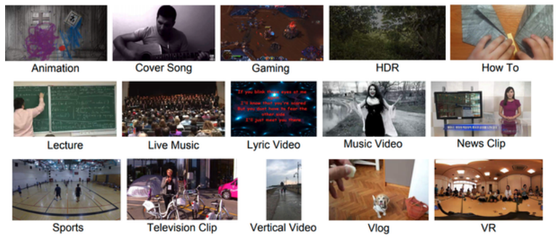
The Skeleton of a Data Science Project

1. It all starts with the planning phase, where you describe the objective of your project, the inputs and outputs of each processing step, and the roles of your team in them.
2. If we don't have the raw data at hand we must scrape (Obtain) it from somewhere, be it from an API or a webpage.
3. Formatting that raw data is done during the wrangling (Scrub) phase.
4. Sometimes we need that state-of-the-art model, and that's made during the research phase.
5. With everything in place, we can start modeling (Explore/Model).
6. Off to the engineering part of town! We need to take care of the lifecycle of our model using MLOps...
7. And deploy our solution in a production-level environment.
8. Finally, we gather our enriched data and analyze (iNterpret) it using dashboards and reports.

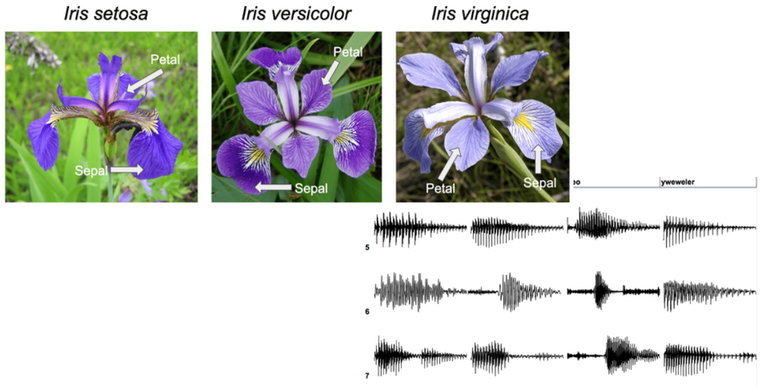
DataSet Providers



DataSet Examples



- ImageNet
- Coco dataset
- Iris Flower dataset
- Breast cancer Wisconsin (Diagnostic) Dataset
- Twitter sentiment Analysis Dataset
- MNIST dataset (handwritten data)
- Fashion MNIST dataset
- Amazon review dataset
- Spam SMS classifier dataset
- Spam-Mails Dataset
- Youtube Dataset



- CIFAR -10
- IMDB reviews
- Sentiment 140
- Facial image Dataset
- Wine Quality Dataset
- The Wikipedia corpus
- Free Spoken digit dataset
- Boston House price dataset
- Pima Indian Diabetes dataset
- Iris Dataset

- Diamond Dataset
- mtcars Dataset
- Boston Dataset
- Titanic Dataset
- Pima Indian Diabetes Dataset
- Beavers Dataset
- Cars93 Dataset
- Car-seats Dataset
- msleep Dataset
- Cushings Dataset
- ToothGrowth Dataset

Farsi DataSet

۱. مجموعه داده دیجی کالا

۲. شناسایی موجودیت‌های نامدار برای تشخیص مشاغل در متن

۳. ۱۰۰ موجودیت اسمی برای تشحیص رویداد

۴. شناسایی موجودیت‌های نامدار برای مکان‌های جغرافیایی

۵. اسامی اشخاص حقیقی برای تشخیص موجودیت اسمی

۶. اسم تجهیزات (appliances) برای تشخیص موجودیت اسمی

۷. ارقام دست نوشت فارسی (هدی)

۸. اشعار مثنوی معنوی مولوی و دیوان شمس / اشعار حافظ

۹. متن دعای ابوحمزه ثمالی با ویرگول و قطعه‌بندی شده

۱۰. اخبار فارسی

مجموعه داده ocr

Memory

Decade	Dataset	Memory	Floating point calculations per second
1970	100 (Iris)	1 KB	100 KF (Intel 8080)
1980	1 K (House prices in Boston)	100 KB	1 MF (Intel 80186)
1990	10 K (optical character recognition)	10 MB	10 MF (Intel 80486)
2000	10 M (web pages)	100 MB	1 GF (Intel Core)
2010	10 G (advertising)	1 GB	1 TF (Nvidia C2050)
2020	1 T (social network)	100 GB	1 PF (Nvidia DGX-2)

Dataset vs. computer memory and computational power

What we have done today!

- Course Organization
- Data Science in Business
- Data Roles
- Git
- Spark
- Cloud
- A/B testing
- Dashboards
- ML Backgrounds