INTRODUCTION TO SPARK WITH SCALA

Spark MLlib

Agenda

- AI Overview
- Machine Learning Overview
- Spark MLlib overview

AI Overview

Welcome To AI

GTC 2019 Opening Keynote - https://www.youtube.com/watch?v=Z2XlNfCtxwI

AI Overview

AI is the new "***Electricity***"

— Andrew Ng

Joke - Why are there so many shocking results in AI?

AI Overview

[AI] is the holy grail, it's the big dream that anybody who's ever been in computer science has been thinking about

— Bill Gates

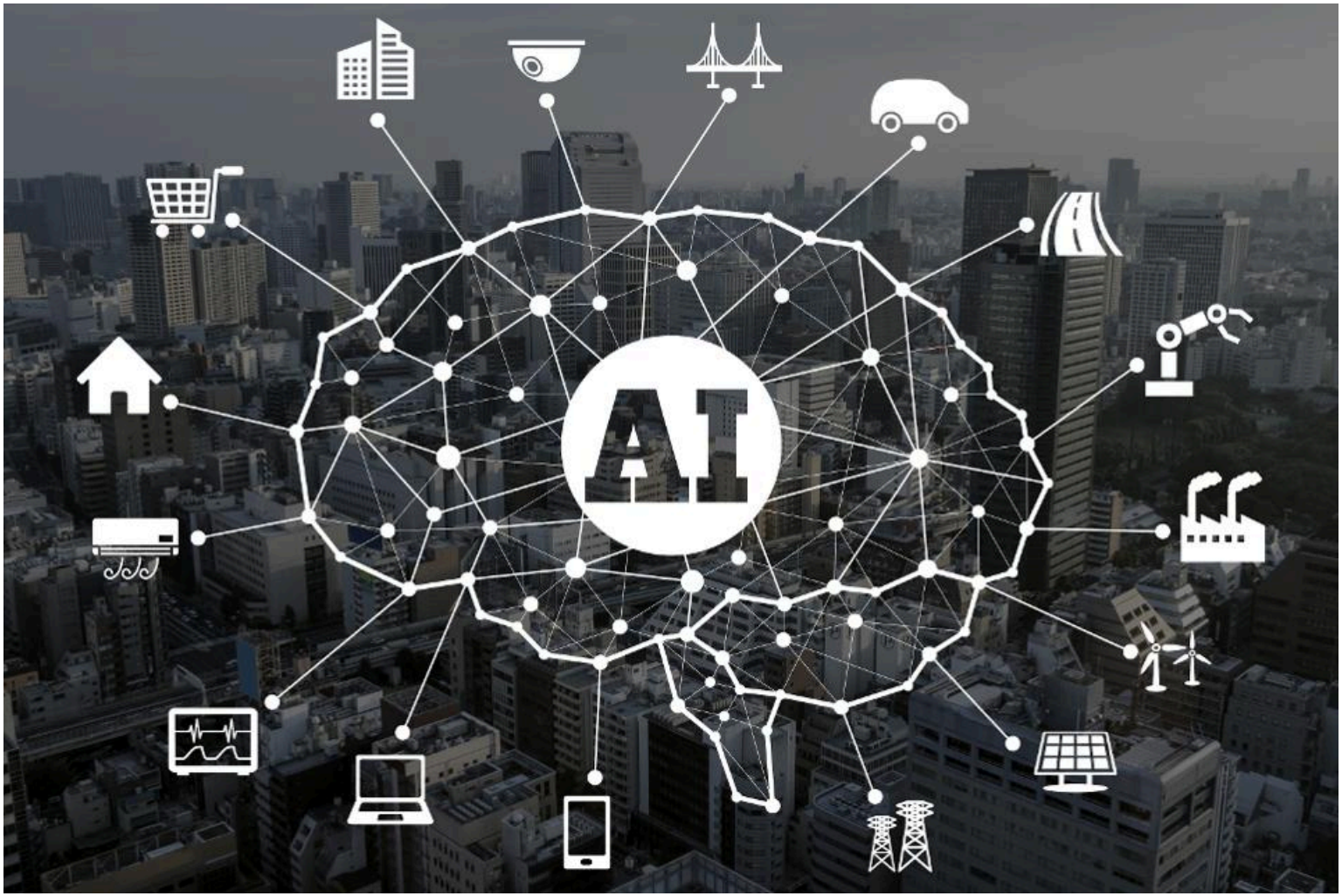
AI Overview

Artificial intelligence is shaping up as the **next industrial revolution**, poised to rapidly reinvent business, the global economy and how people work and interact with each other

- WSJ 03/06/2017

https://www.wsj.com/articles/how-artificial-intelligence-will-change-everything-1488856320

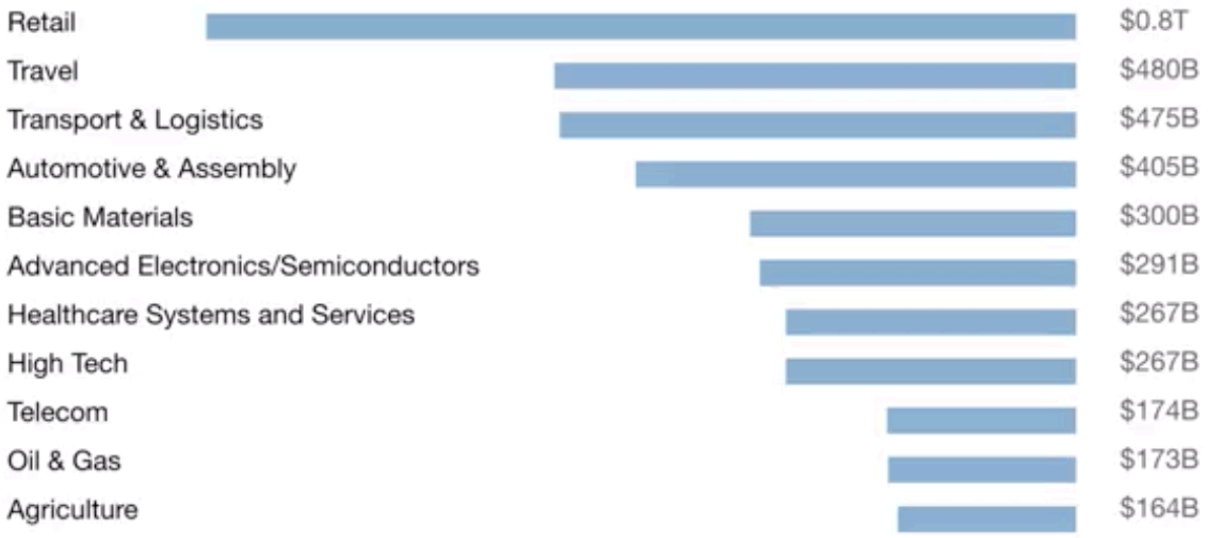
AI Overview



AI Overview

AI value creation by 2030

$13 trillion



[Source: McKinsey Global Institute.]

AI Overview

Impact On Worldwide Jobs

Jobs displaced by 2030

400-800 mil

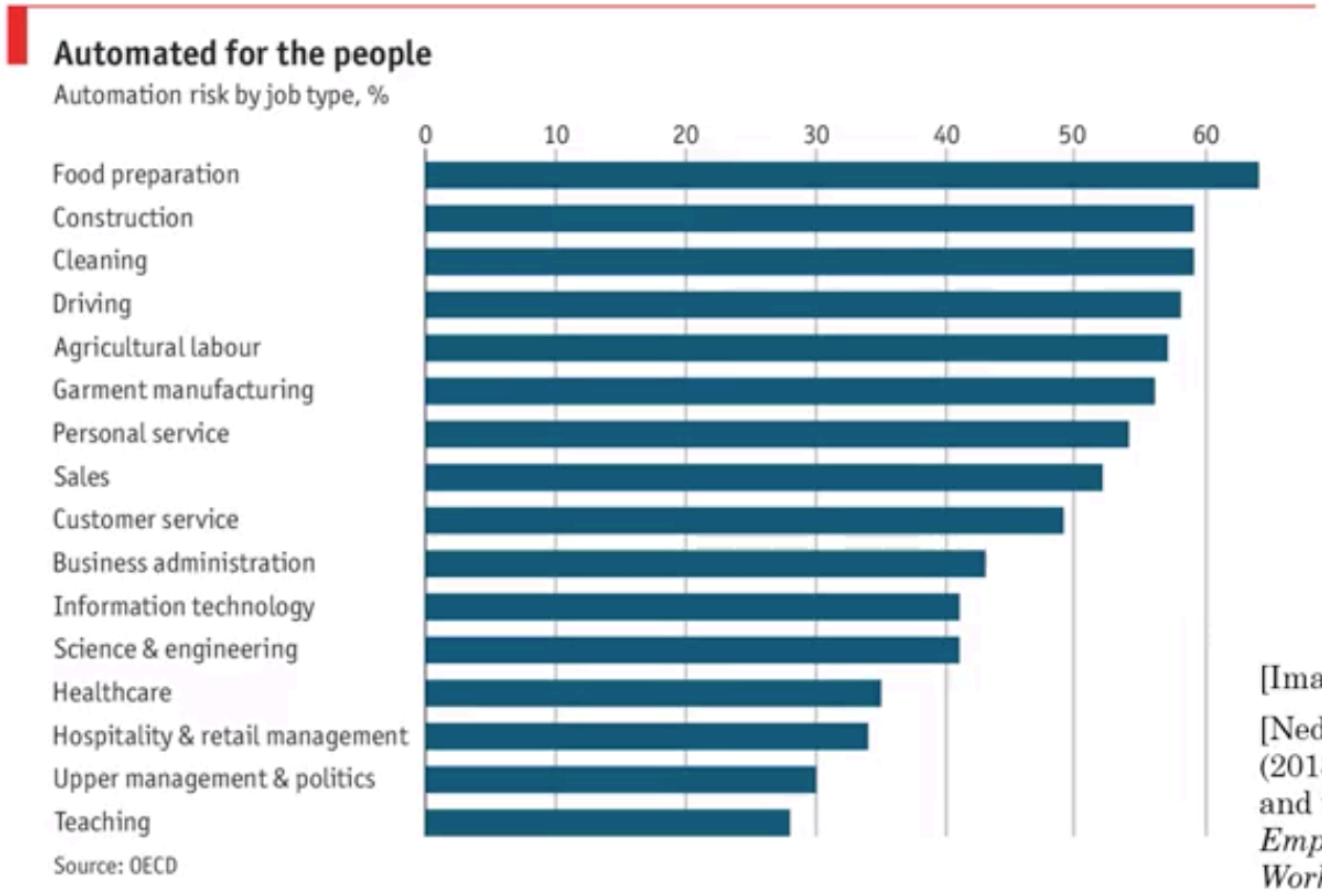
Jobs created by 2030

555-890 mil

[Source: McKinsey Global Institute.]

AI Overview

Impact On Worldwide Jobs



[Image credit: Economist.com]

[Nedelkoska, L. and G. Quintini. (2018). Automation, skills use and training. *OECD Social, Employment and Migration Working Papers*, No. 202.]

AI Overview

Online Dating – Recommendation System

One third of US marriages begin with online dating

AI Overview

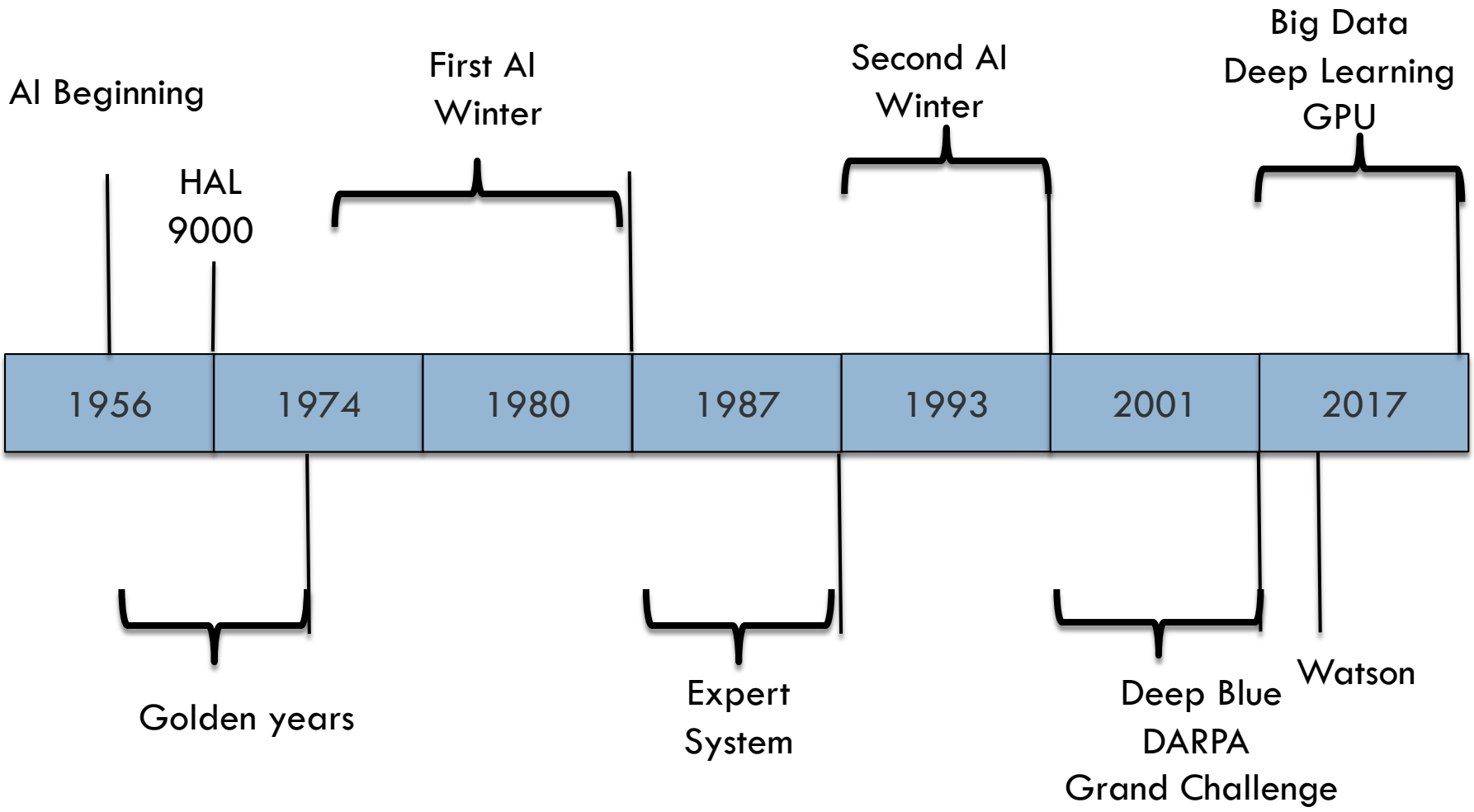
What is Artificial Intelligence?

Every aspect of ***learning*** or any other feature of intelligence can in principle be so precisely described that a machine can be made to ***simulate*** it

- John McCarthy

Mission statement of the Dartmouth Conference, 1956

AI Overview

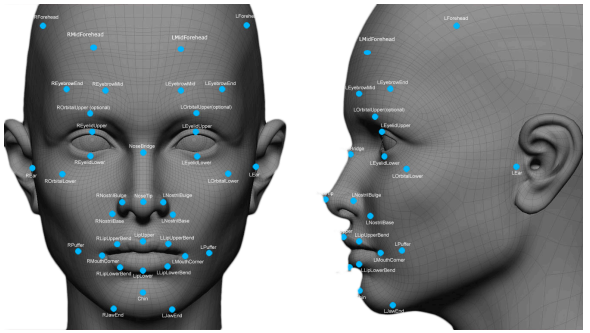

AI Beginning
HAL 9000
First AI Winter
Second AI Winter
Big Data
Deep Learning
GPU
1956
1974
1980
1987
1993
2001
2017
Golden years
Expert System
Deep Blue
DARPA
Grand Challenge
Watson

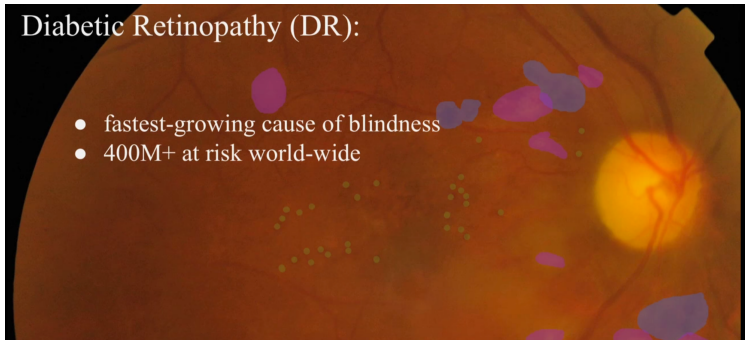
AI Overview

Computer can now see and hear, before they couldn't

99% Accurate

3% Error

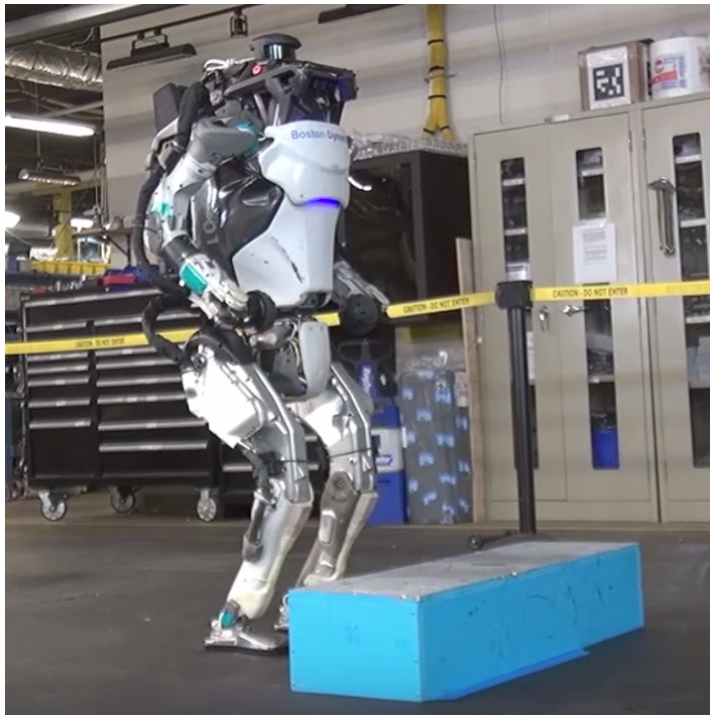




AI Overview

Moley Kitchen Robot

https://www.youtube.com/watch?v=6wgi_XGITio



https://www.youtube.com/watch?v=fRj34o4hN4I

Boston Dynamics - Atlas

AI Overview

Go game is the holy grail for AI

There are more possible positions in Go than there are atoms in the universe

AI Overview

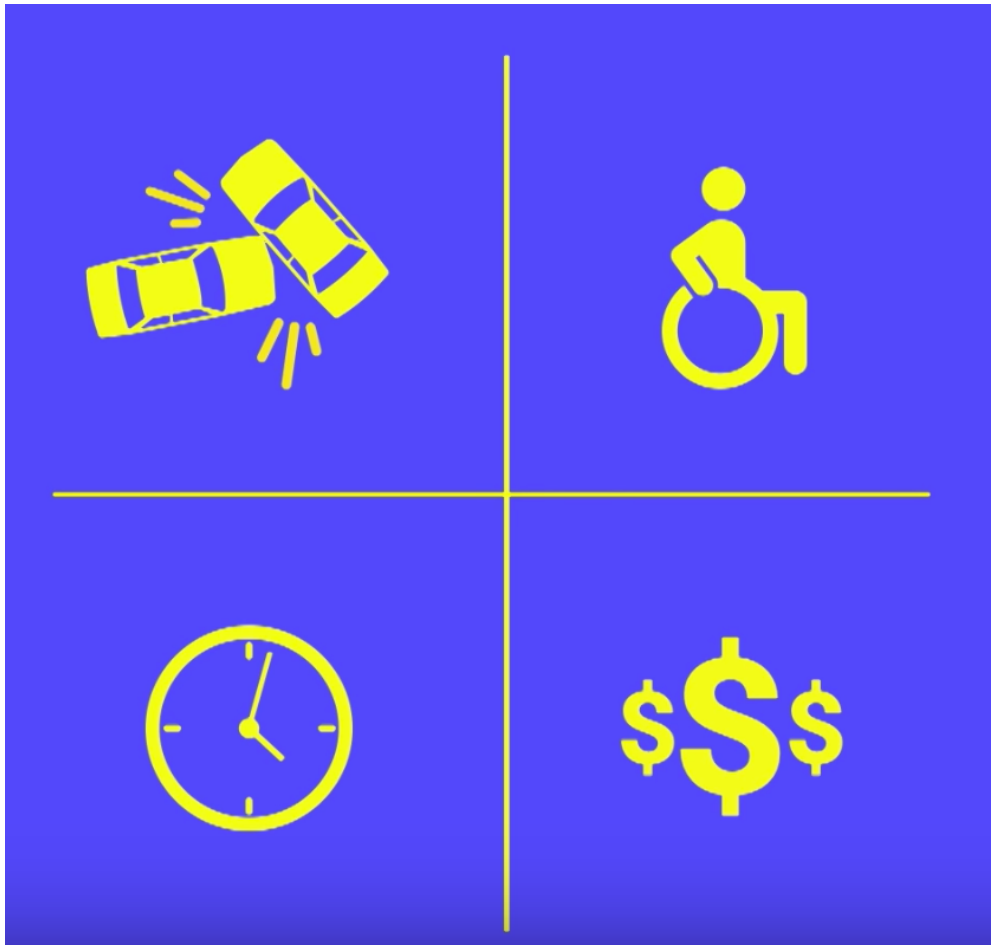
NIO - Eve

https://www.youtube.com/watch?v=x4qYjqC7eVQ - 12:40 (2 mins), 16:30

AI Overview

Why building a better driver?

1.2M

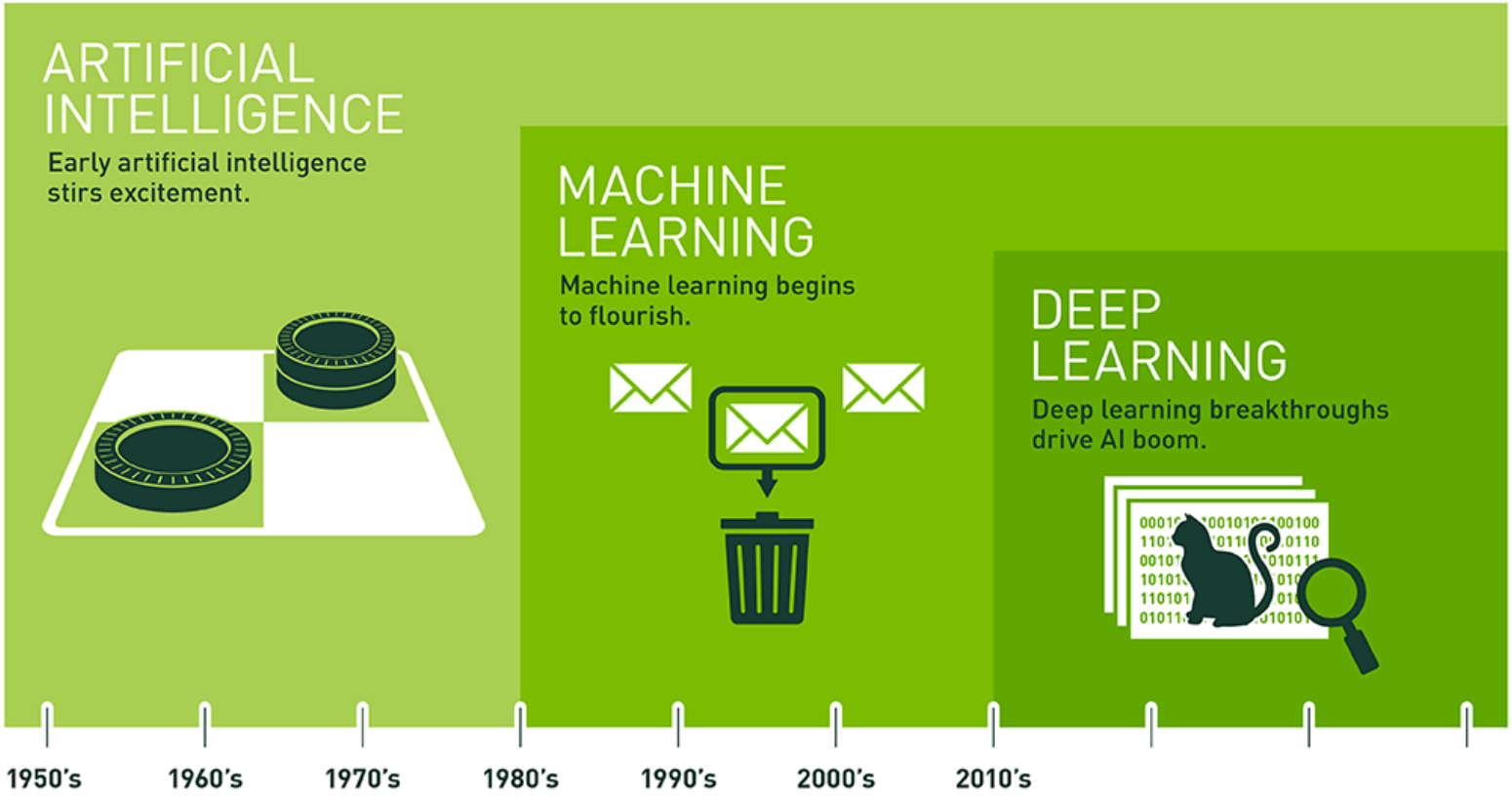


20M

50min

94%

AI Overview



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

AI Overview

Machine Learning vs Data Science

Machine Learning Overview

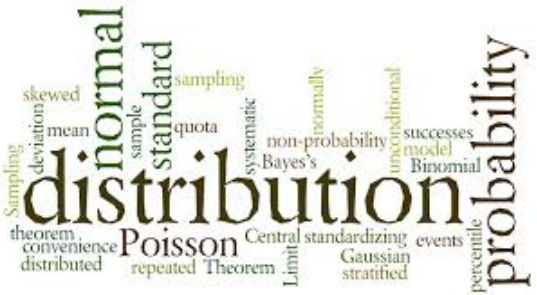
ML Definition

A field of study that gives computers the ability to *learn* without being *explicitly programmed*

- - Arthur Samuel (1959)

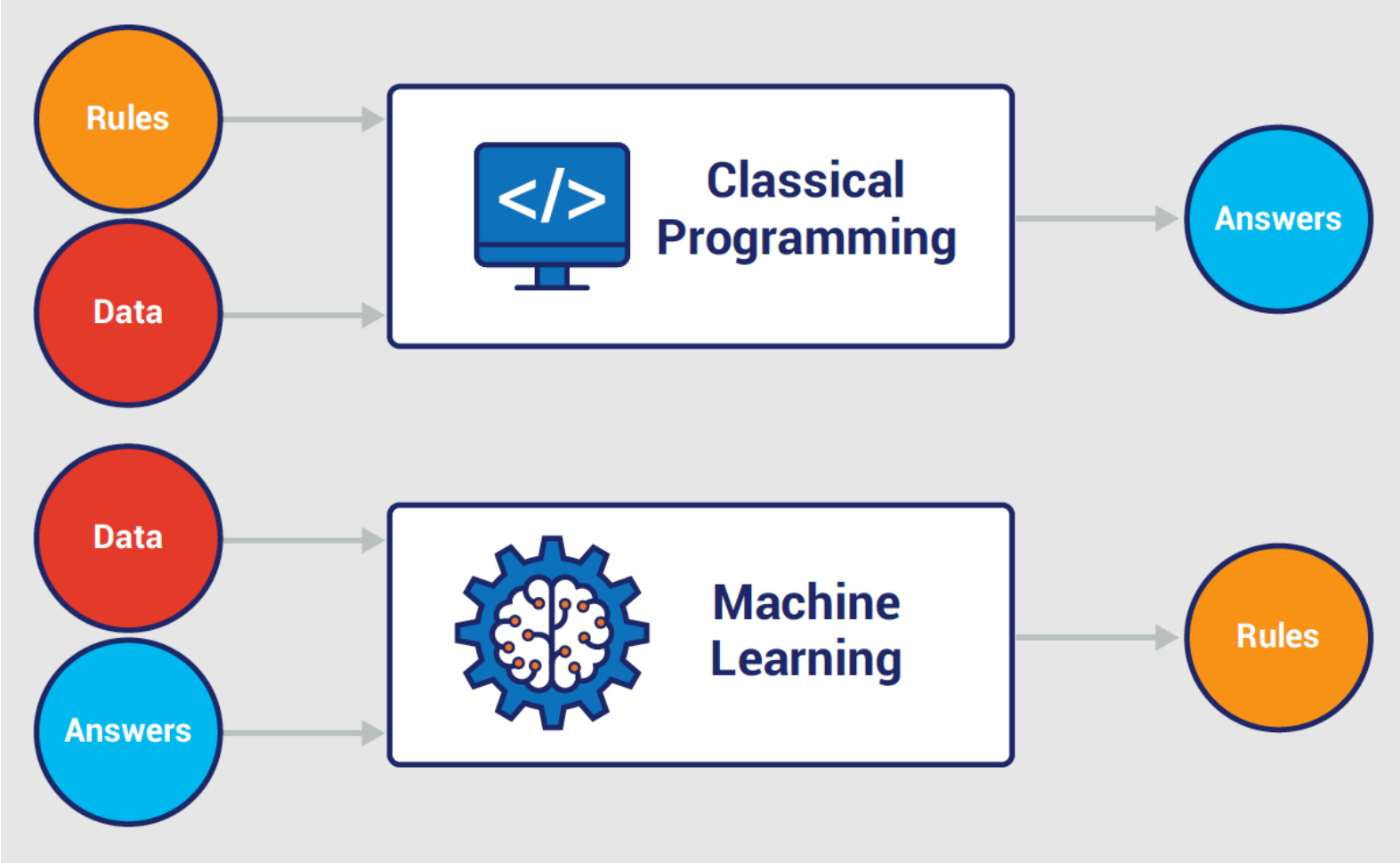
Machine Learning Overview

Involving tools and ideas from various domains



Machine Learning Overview

Machine Learning Programming Paradigm



Machine Learning Overview

ML Glossary

- Classification
- Regression
- Active learning
- Reinforcement learning
- No free lunch
- Clustering
- K-means
- Chinese restaurant process
- Indian buffet process
- Naïve Bayes
- Discrimination rule
- Latent feature
- Linear regression
- Logistic regression
- Additive regression
- Autoregression
- Markov Chain
- Hidden Markov model
- Couple hidden Markov model
- Monte Carlo
- Markov Chain Monte Carlo
- Stochastic gradient descent

Machine Learning Overview

ML Concepts &Terminologies

Observations
(Items for entities used for learning, i.e emails)

Label values
(Spam or Not)

Features
(Word count, sender)

Training Data | Test Data

Machine Learning Overview

Machine Learning In Action

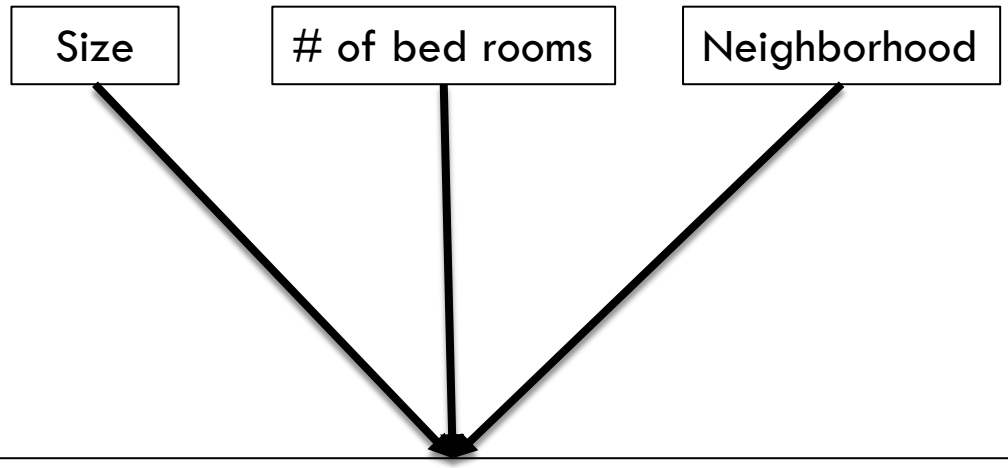
380K

560K

990K

Machine Learning Overview

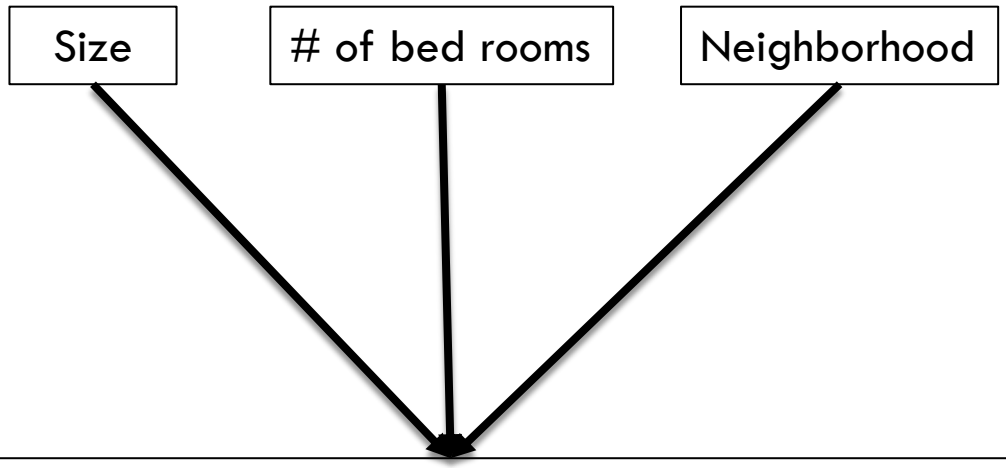
Rule Based Approach



price = function(size, bed rooms, neighborhood)

Machine Learning Overview

Rule Based Approach



price = function(size, bed rooms, neighborhood)

Housing market changes overtime

Machine Learning Overview

House Price Prediction

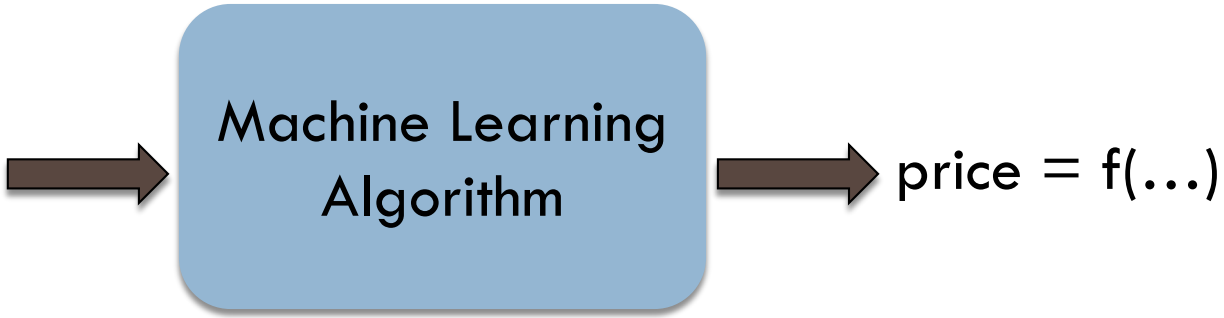
Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

Machine Learning Overview

ML Based Approach

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

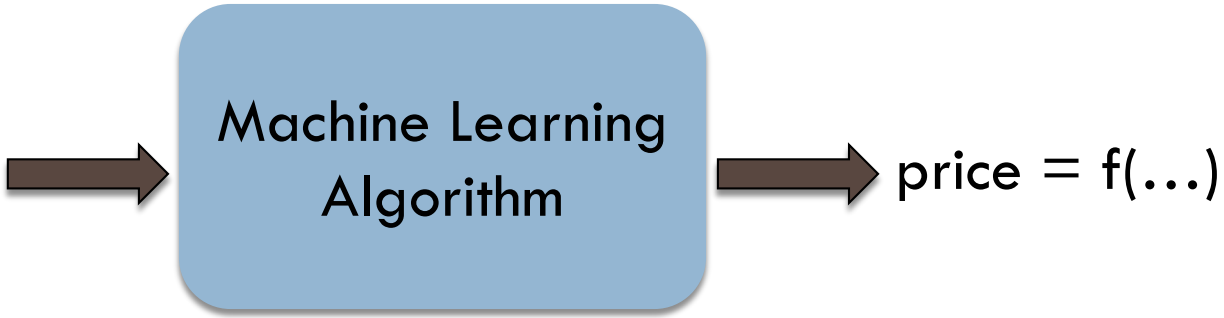


Without explicitly programmed

Machine Learning Overview

ML Based Approach

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

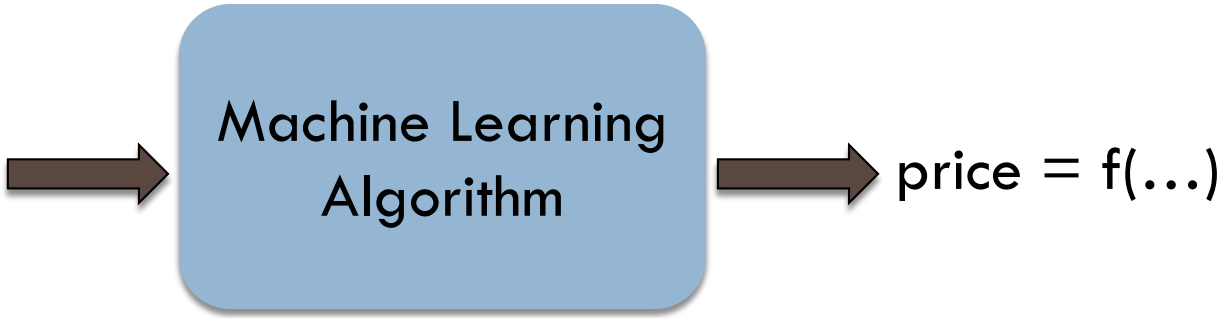


The ML algorithm doesn't know anything about those houses. It only sees numbers.

Machine Learning Overview

ML Based Approach

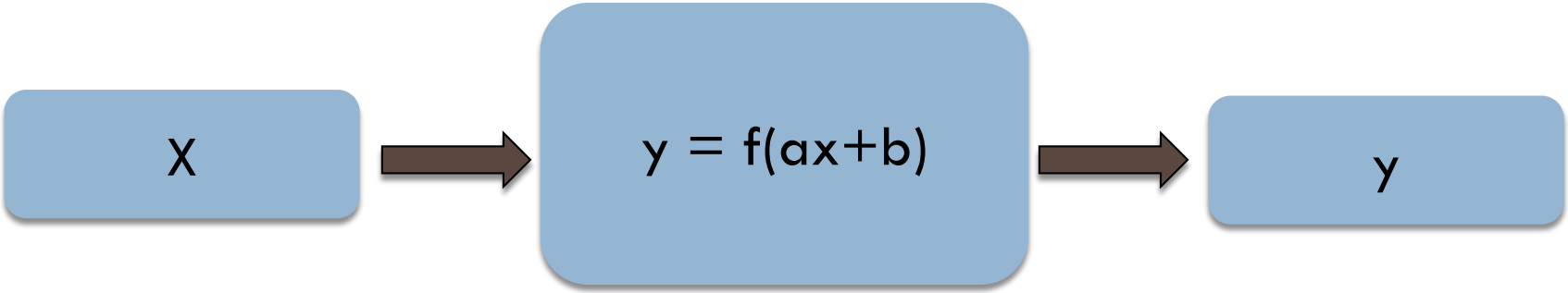
Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000



The more data you throw at it, the more accurate the predictions will be

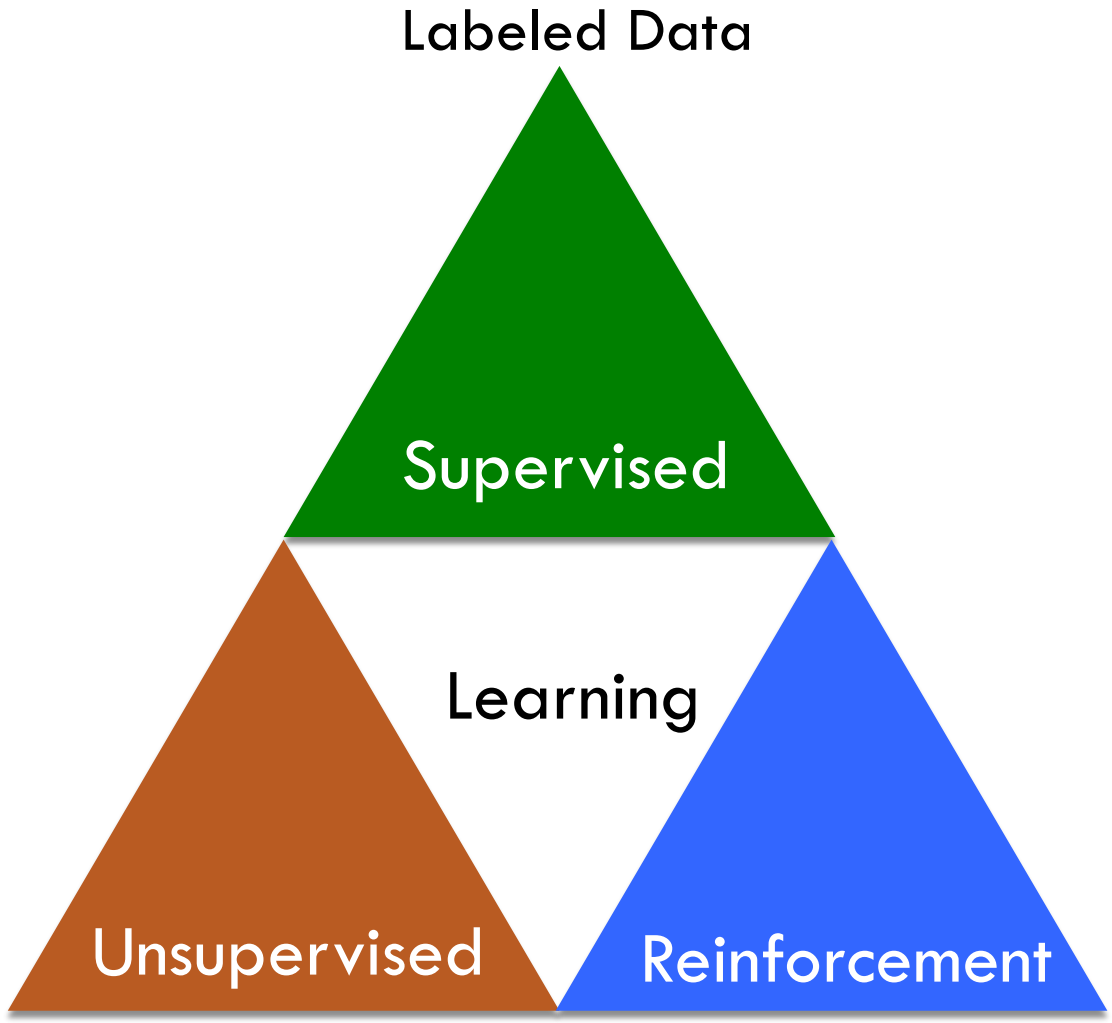
Machine Learning Overview

Machine Learning Model



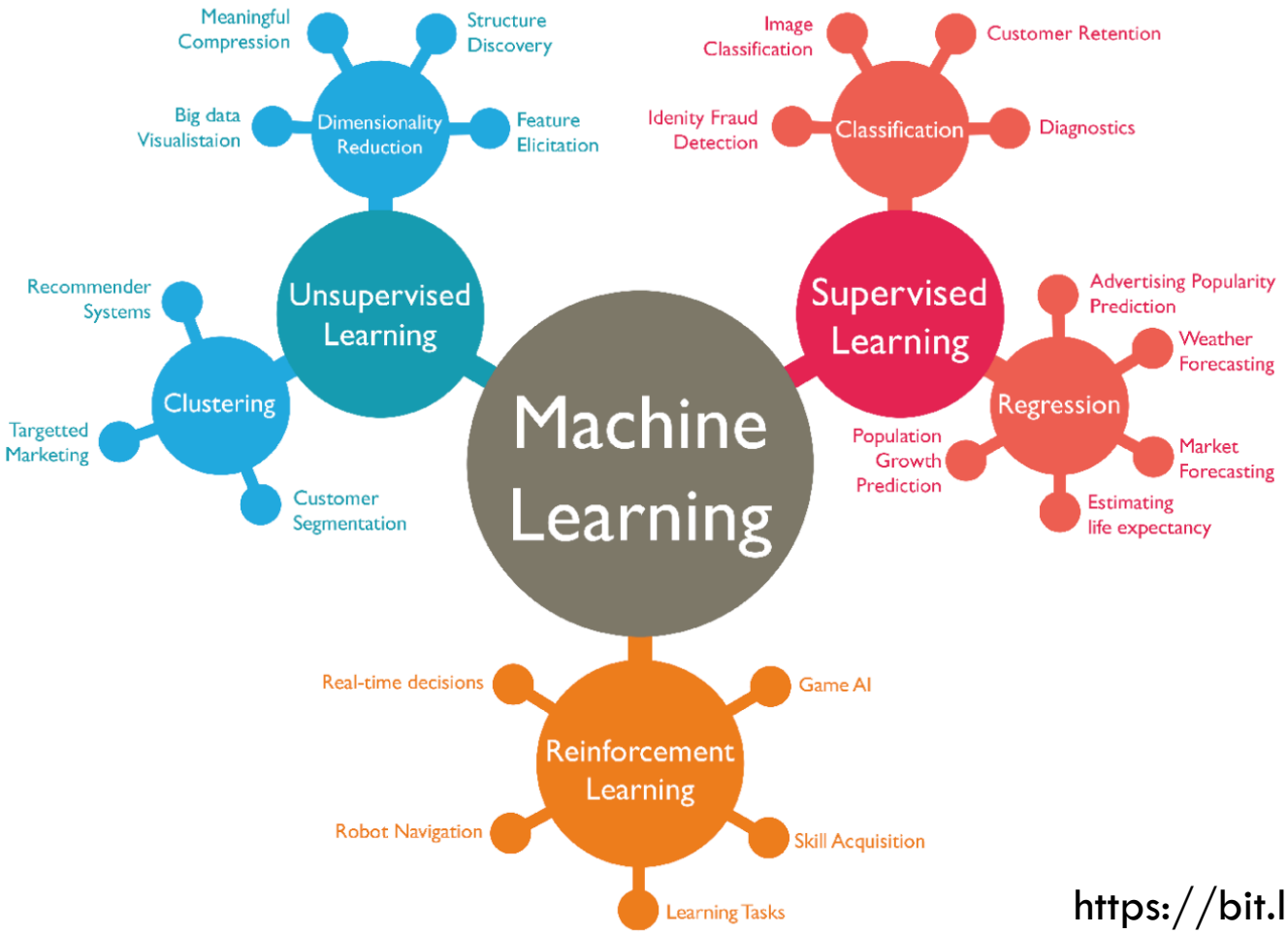
Beautiful Math

Machine Learning Overview



Machine Learning Overview

Machine Learning Types



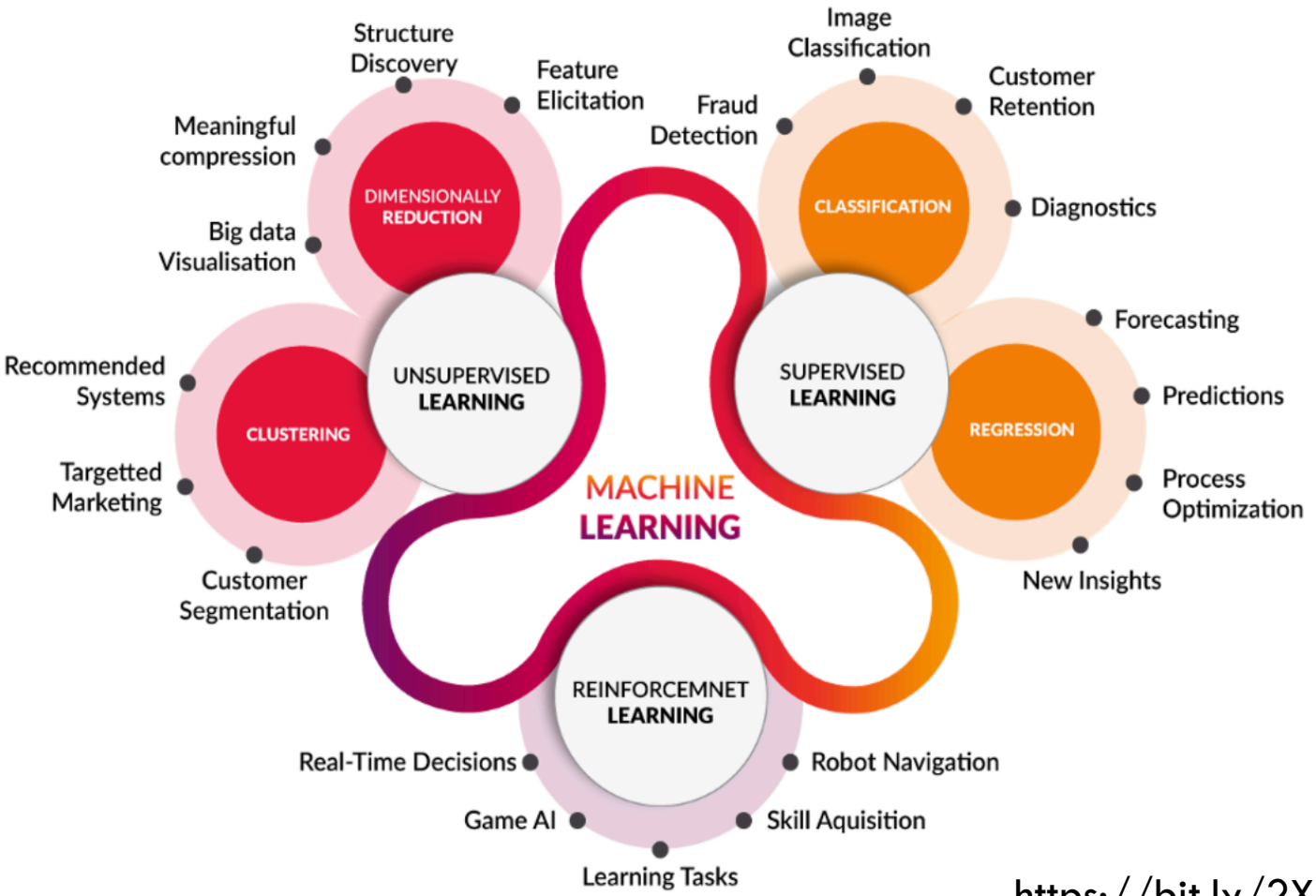
https://bit.ly/2pyxziU

Machine Learning Overview

- Supervised – learning from labeled observations
  - Algorithm to learn mapping from observations to labels
  - Training data includes desired outputs (labeled data)
  - Goal is to predict class or value
- Unsupervised – learning from unlabeled observations
  - No knowledge of output class or value
  - Labeled data are not available
  - Goal is to extract data patterns or groupings or structure
- Reinforcement learning
  - Learn a policy of how to act given an observation

Machine Learning Overview

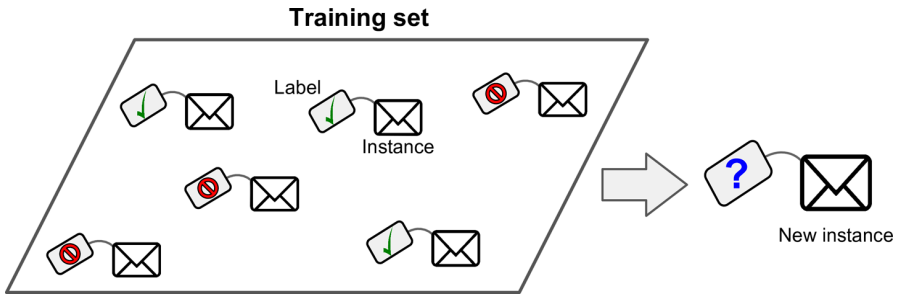
Applications of Machine Learning Types

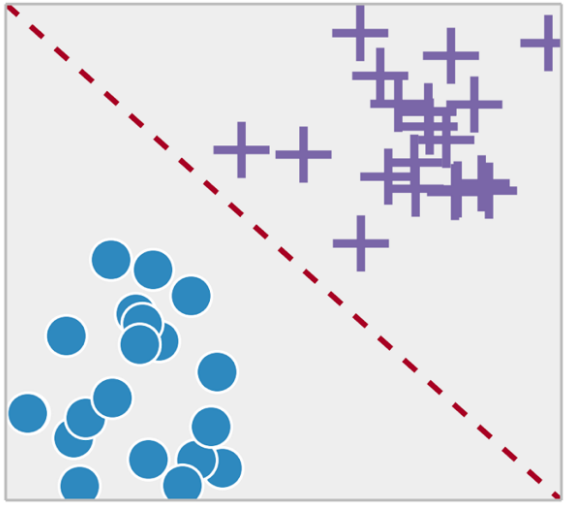


https://bit.ly/2XLQIQC

Machine Learning Overview

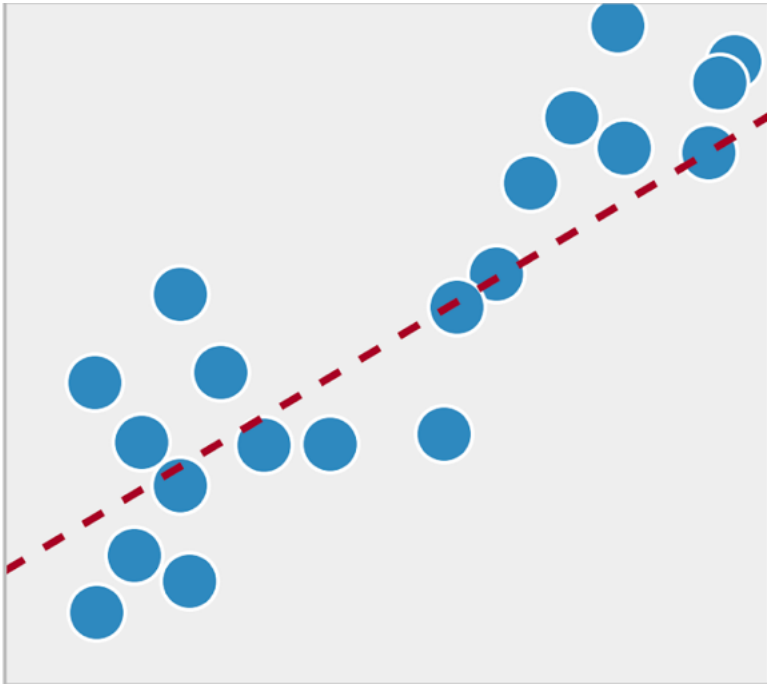
- Supervised Learning
  - The training data includes the *desired outcome*
  - Classification – assign a category to each observation
    - Spam or no spam
    - Cat, dog, dolphin





Machine Learning Overview

- Supervised Learning
  - Regression – label value is contiguous
    - House price, income level, (other examples ???)

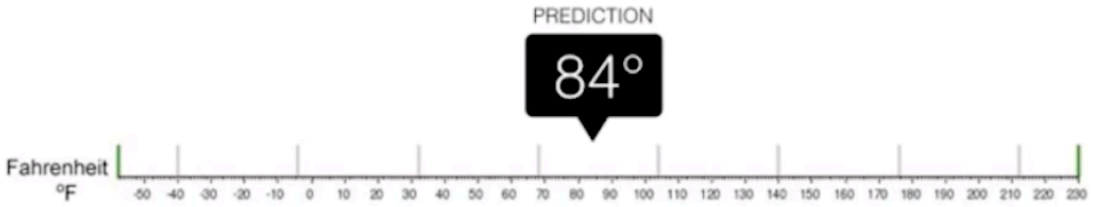


Machine Learning Overview

Regression vs Classification

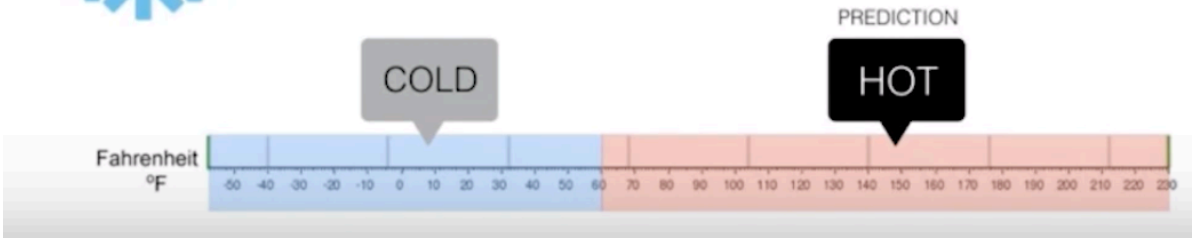
Regression

What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?



Machine Learning Overview

Supervised Learning Applications

Input	Output	Applications
Voice recording	Transcript	Speech Recording
Stock market data	Future market data	Trading bots
Photograph	Caption	Imaging tagging
Store transaction details	Is the transaction fraudulent?	Fraud detection
Recipe ingredients	Customer reviews	Food recommendation
Faces	Names	Face recognition
Purchase history	Future purchase behavior	Customer retention

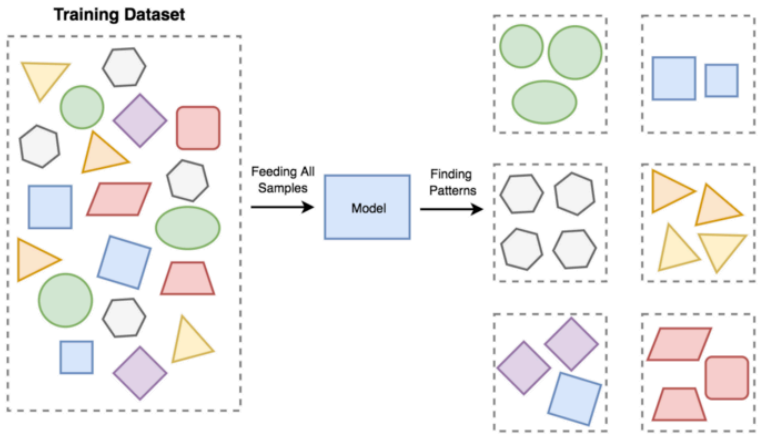
https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Machine Learning Overview

- Supervised Learning Algorithms
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVMs)
  - Decision Trees
  - Random Forests
  - Neural Networks

Machine Learning Overview

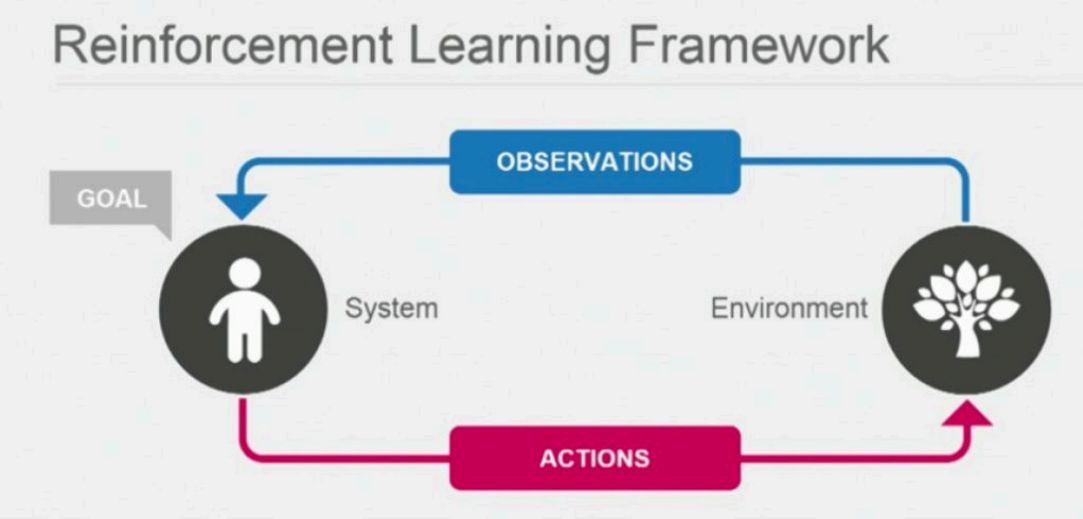
- Unsupervised Learning
  - Identify/infer commonalities or patterns in the data
    - The underlying structure of the data
  - The training data does not include the *desired outcome*
  - Examples
    - Clustering – customer segmentation
    - Anomaly/fraud detection



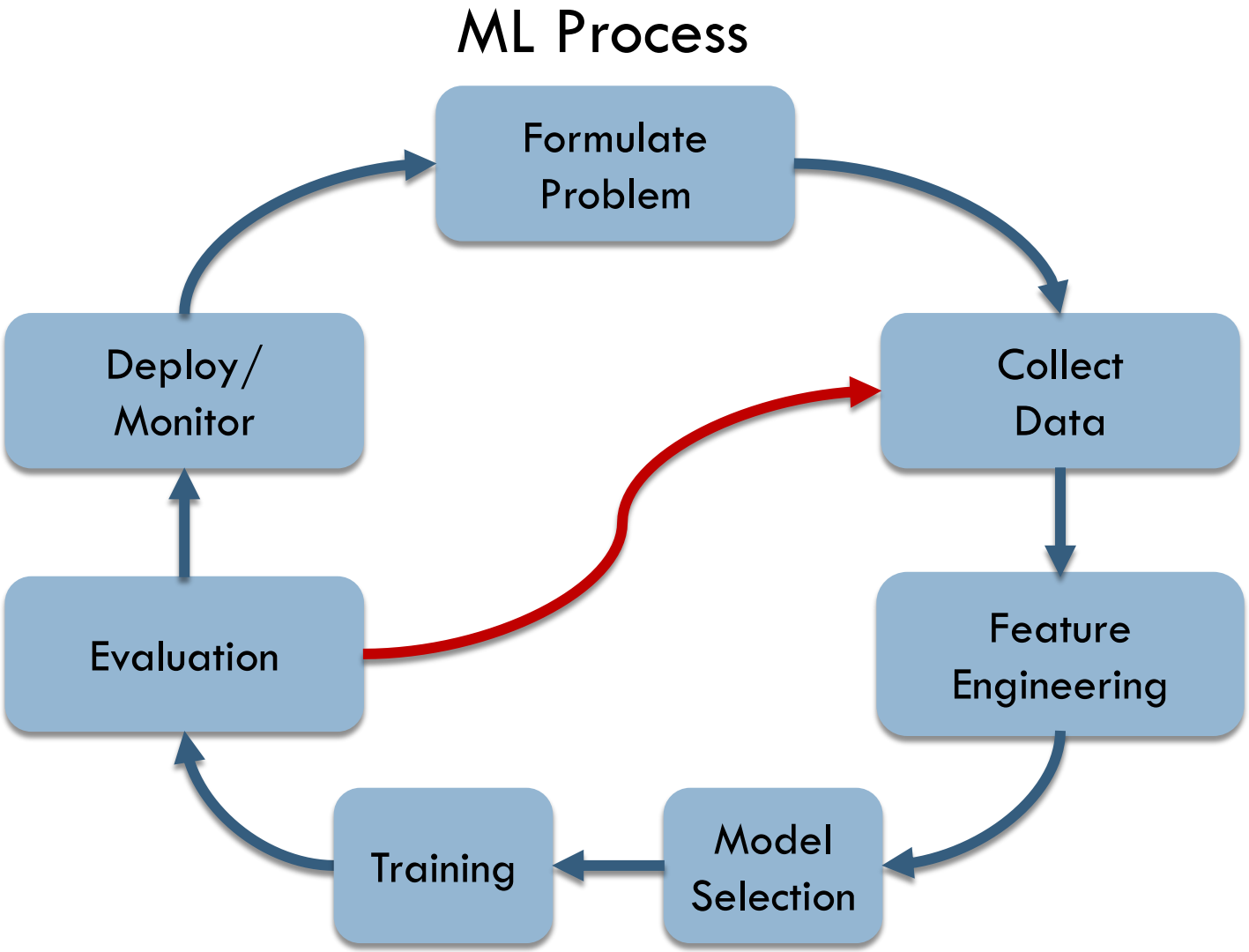
https://bit.ly/2XLQIQC

Machine Learning Overview

- Reinforcement Learning
  - Learning through taking action and evaluate reward
  - Policy
    - Best strategy to get most rewards over time
    - Determine what action agent should choose in given state



Machine Learning Overview



Machine Learning Overview

- Formulating Problem
  - Clearly define business problem
    - Reduce churn
    - Improve certain business metric or earning predictions
  - Clearly define success metrics
    - Improve CTR by 5%
    - Increase engagement by 10%
  - Understand ROI
    - Is there a simpler approach?
  - Define MVP to get quick feedback and iterate

Machine Learning Overview

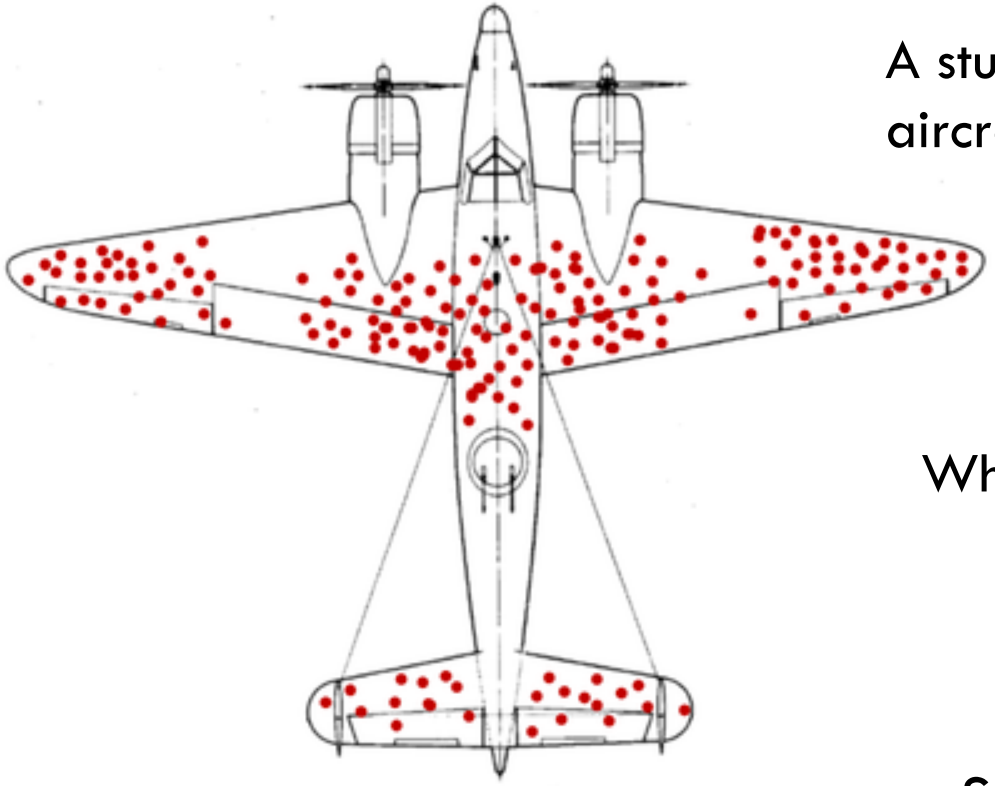
- Collect data
  - Garbage in, garbage out
  - The more (relevant) data the better
  - Make sure data is representative
    - Generalize to new observations
  - Dealing with cold start problem
    - No labeled data
    - Crowd sourcing
  - Avoid survivorship bias

Machine Learning Overview

- Feature Engineering
  - Longer computing time – more features more time
  - Unnecessarily complex models, hard to interpret
  - Process
    - Data cleansing
      - Real world reality
    - Cleaning missing values
    - Reducing noise – outliers, erroneous values
    - Data normalization
      - Range, categorical values into integers
    - Dimensionality reduction
      - Reducing # of features
    - Transform features in numerical values

Machine Learning Overview

WW II Story – US Navy Planes



A study of the damage done to aircraft that had returned from missions

Where to optimally armor the planes?

Survivorship Bias

Machine Learning Overview

Feature Engineering

- "Cheap" or "Prince"
- Spelling mistake
- Missing title
- Weird sender email

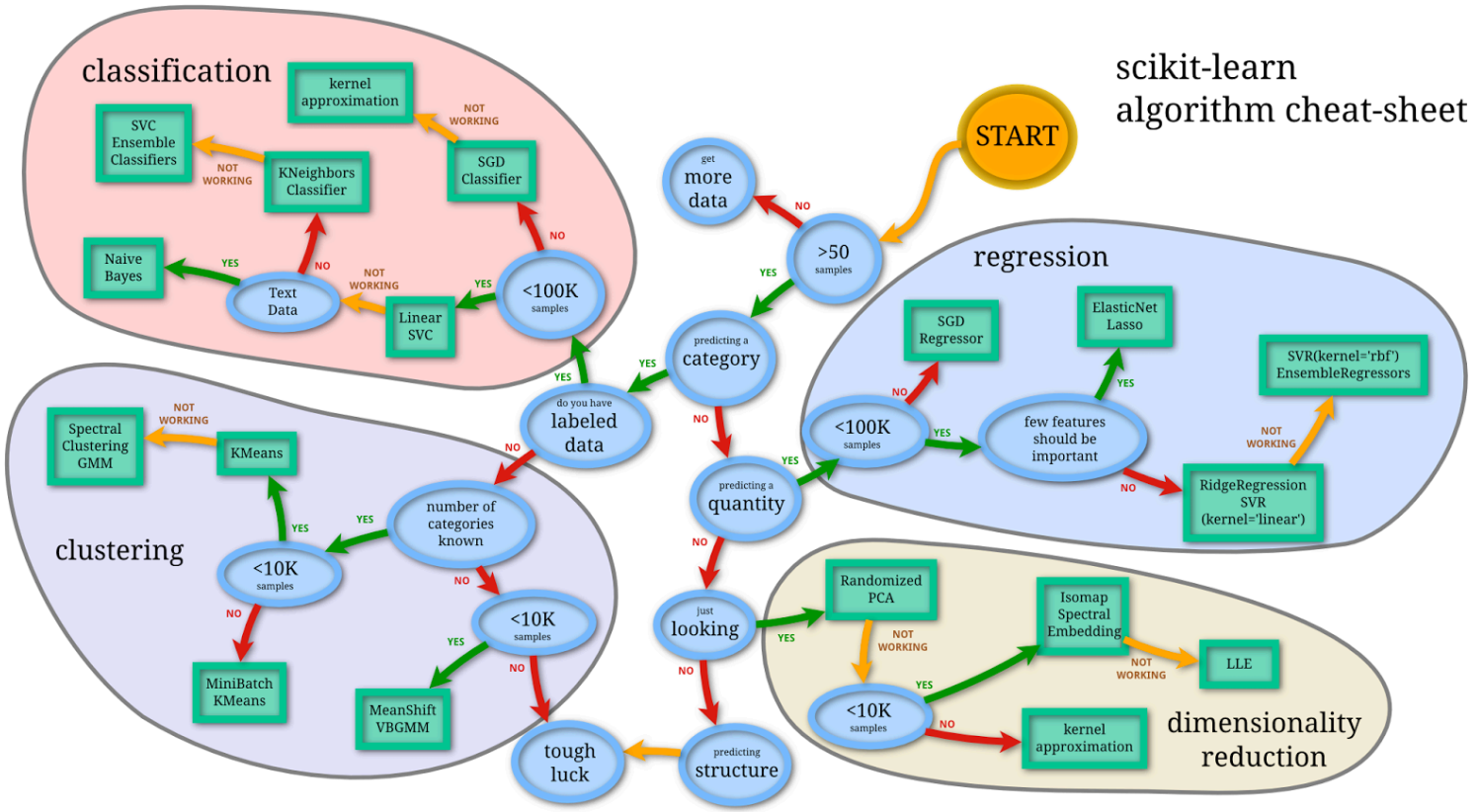
- Location
- Size
- School district
- # of bath rooms

Machine Learning Overview

- Model Selection (half art, half science)
  - Choosing the right ML algorithm
    - Based on the task – prediction or finding pattern
    - Size, quality, and nature of the data
  - Start with simple algorithm to validate hypothesis
  - Try out several algorithms and compare result
  - Consult your in house ML experts

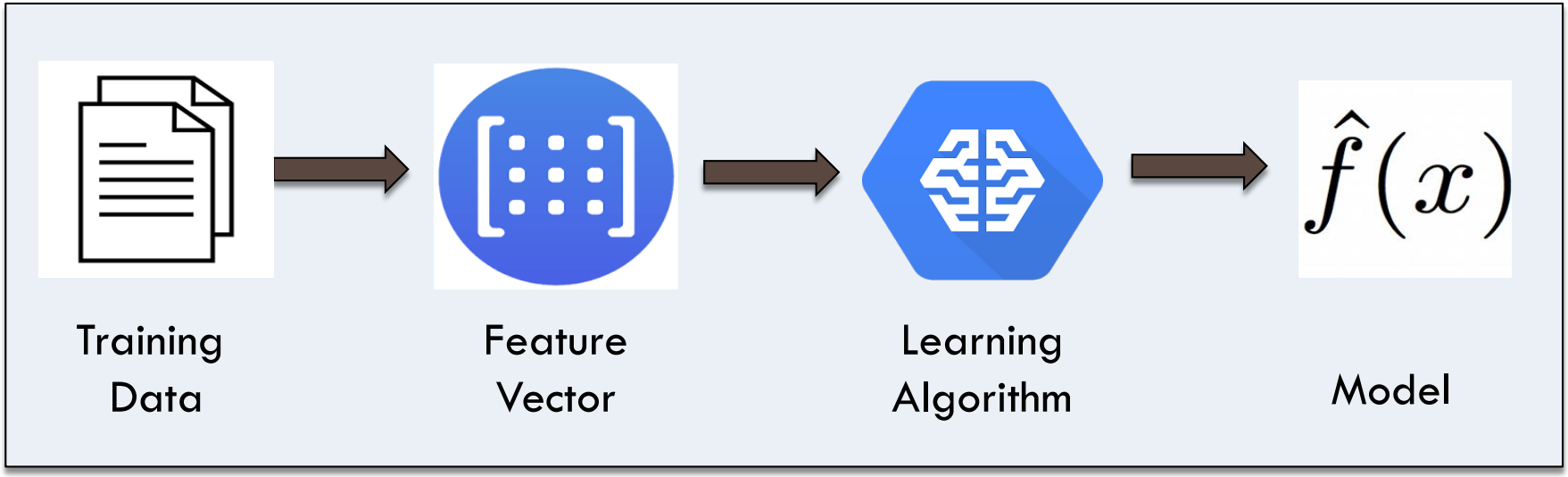
Machine Learning Overview

ML Algorithms



Machine Learning Overview

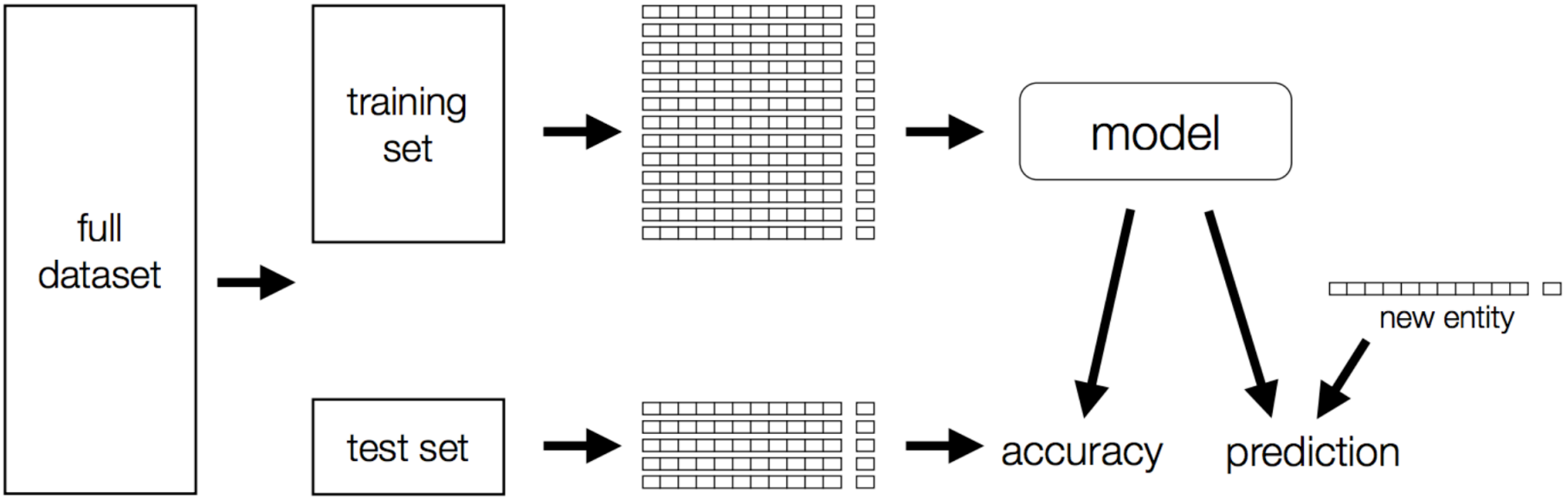
- Training
  - Configure the selected ML algorithm
  - Feed data into the ML algorithm
  - Get back a trained model



Training Step

Machine Learning Overview

Training & Test Data



Machine Learning Overview

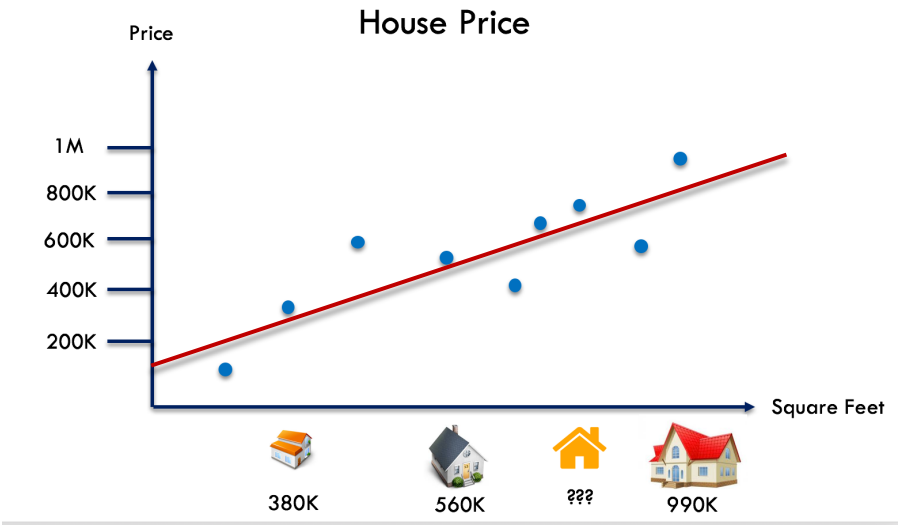
Machine Learning Magic

How does a ML algorithm learn?

What does the machine learn?

Machine Learning Overview

Machine Learning Algorithm – Linear Regression

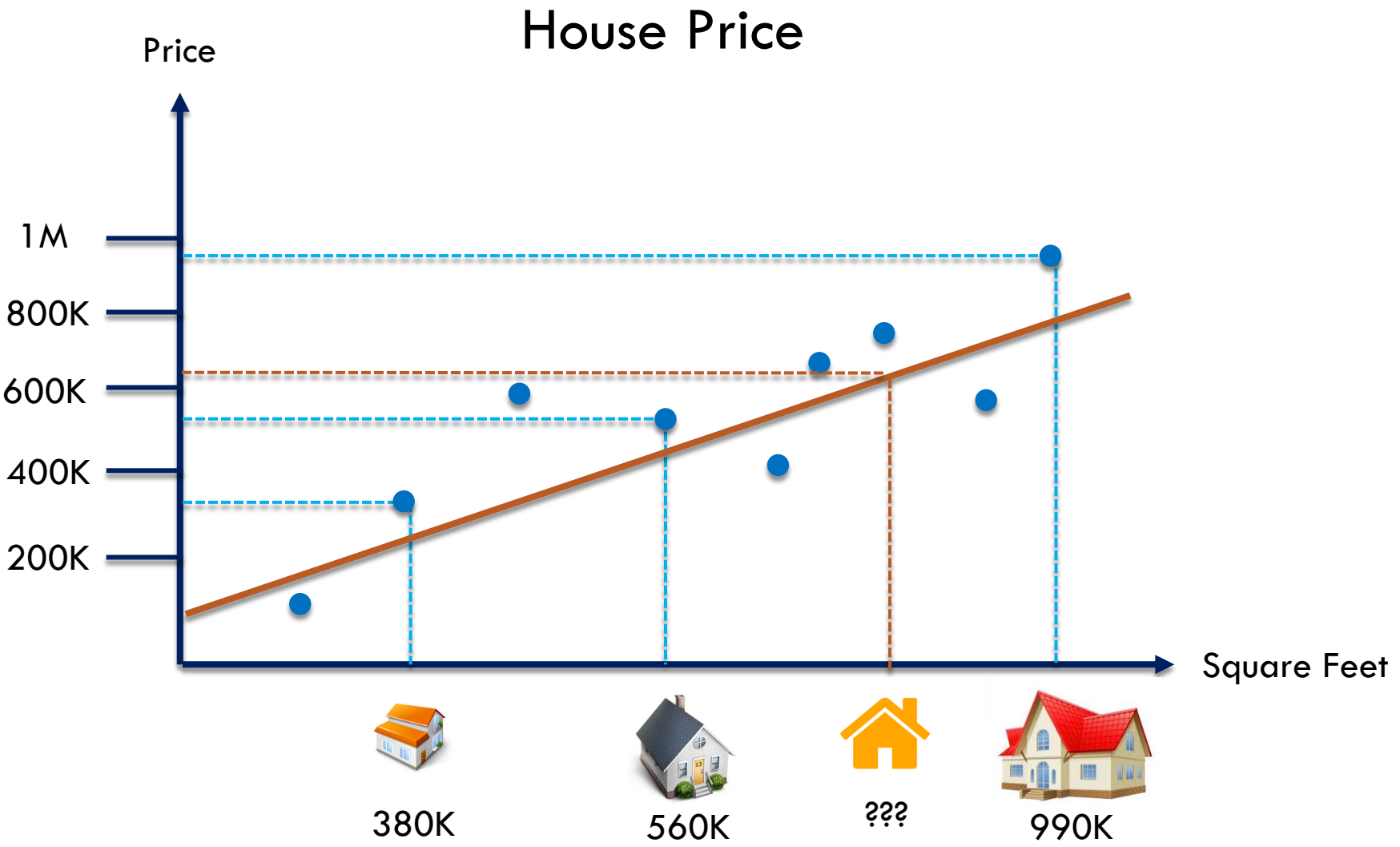


Hypothesis: y ~= h(x) h(x) = b + a*x

Parameters: b, a (to be learned)

Cost function: Pick b and a such that h(x) is close to y

Machine Learning Overview



Machine Learning Overview

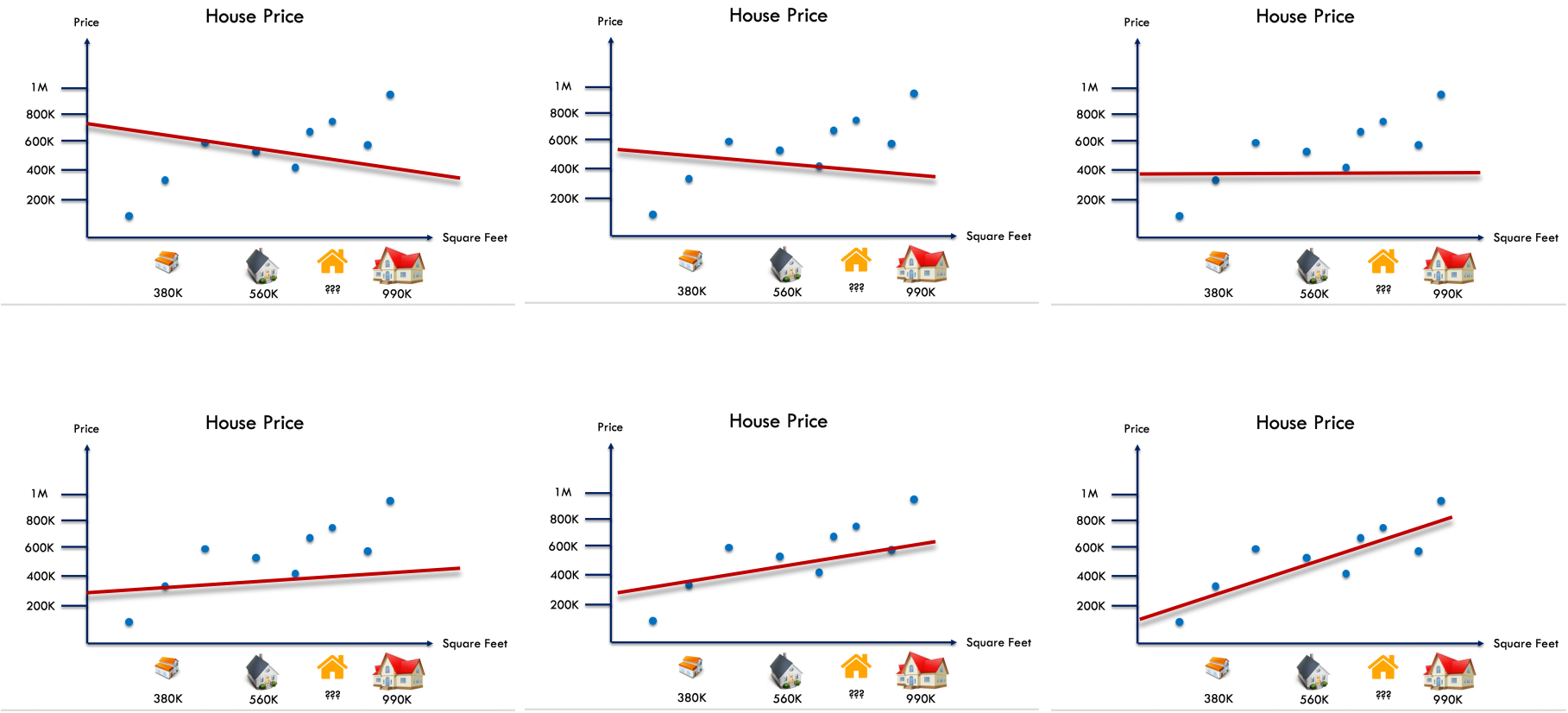
Machine Learning Overview

Machine Learning Overview

Machine Learning Overview

Machine Learning Overview

Minimize Cost Function – Gradient Descent

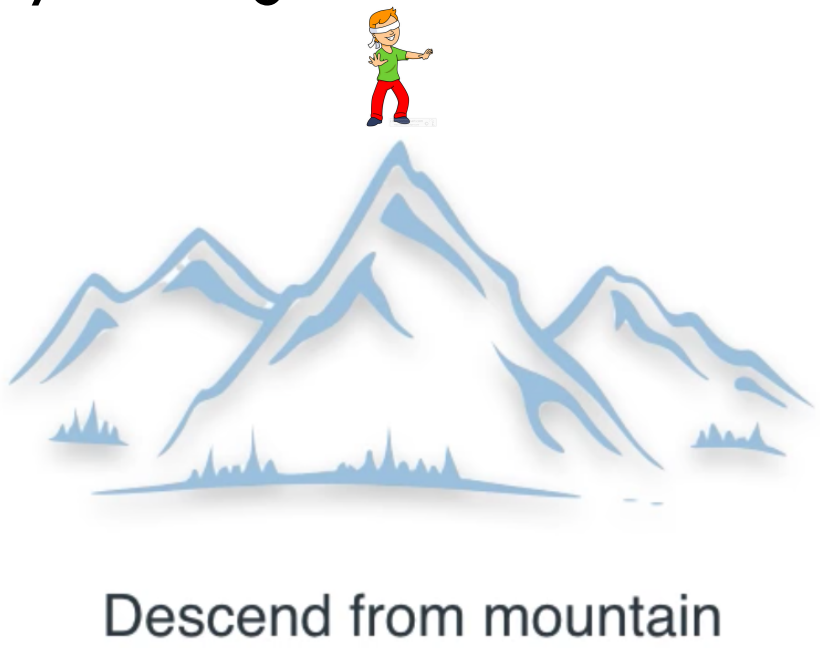


Taking small steps toward direction of minimizing cost function

Machine Learning Overview

Gradient Descent

- A well-known efficient optimization technique
  - Involve calculus
- Widely used in many ML algorithms
- Iterative process



Machine Learning Overview

Essence of ML

- Cost function
  - Find a metric to determine how far from solution
  - Find a way to calculate that metric
- Minimize the error
  - Leverage gradient descent optimization technique

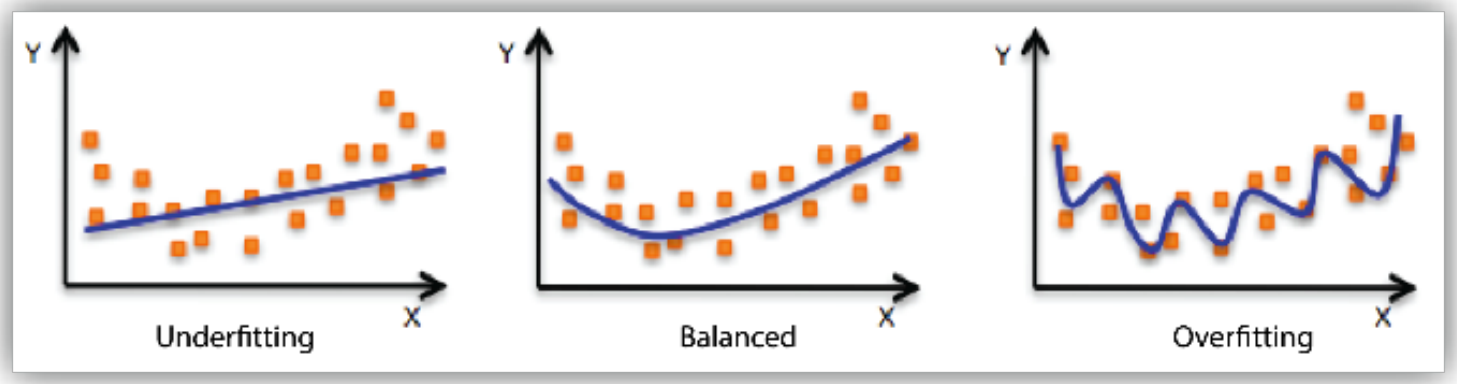
Machine Learning Overview

Model Performance Evaluation Metrics

- Classification
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - ROC & AUC
- Regression
  - Sum of square errors
  - Mean absolute errors
  - RMSE

Machine Learning Overview

- Evaluation – Cardinal Sins of ML
  - Overfitting
    - Perform well on the training data, not on testing data
    - Simplify model and features, collect more data
  - Underfitting
    - Perform not well on training data and testing data
    - Better model and features



Machine Learning Overview

Classification Evaluation Metrics

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Nike Shoe Prediction

- TP: predict Nike shoe as Nike
- FP: predict Nike shoe, but it was Adidas
- TN: predict Adidas shoe, and it was Adidas (not Nike))
- FN: predict Adidas shoe, and it was Nike

Incorporae this - https://www.datasciencecentral.com/profiles/blogs/removing-confusion-from-confusion-matrix-hawaii-false-missile

Machine Learning Overview

- Classification Evaluation metrics
  - Accuracy – how often is the classifier correct?

$$Accuracy = \frac{TP+TN}{Total}$$

  - Precision – when it predicts yes, how often is it correct?

$$Precision = \frac{TP}{TP + FP}$$

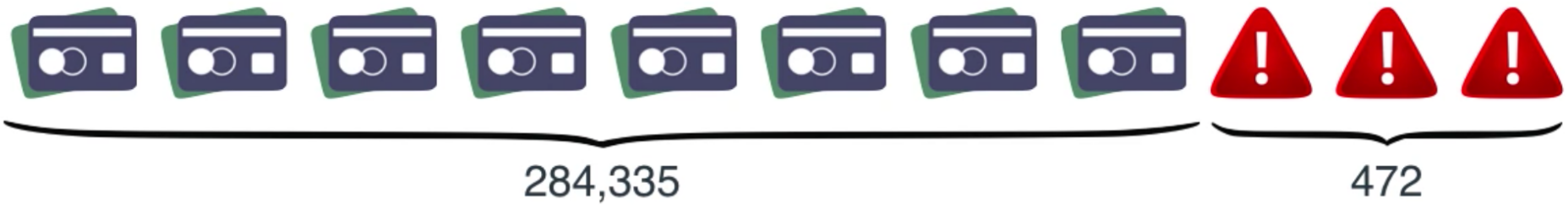
  - Recall – when it is actually yes, how often does it predict yes?

$$Recall = \frac{TP}{TP + FN}$$

Machine Learning Overview

- Good model to predict correctly over 90%?

Credit Card Transaction Data



If a model predicts all transactions are good

$$Accuracy = \frac{TP+TN}{Total}$$

$$Accuracy = \frac{284{,}335}{284{,}807} = 99.83\%$$

Is this a good model?

ML Development & Process

Confusion Matrix Example

Medical Diagnosis

	Actually Sick	Actually Healthy
Diagnosed Sick	TP ✓	FP ✗
Diagnosed Healthy	FN ✗	TN ✓

Spam Detector

	Actually Spam	Actually Not Spam
Predicted Spam	TP ✓	FP ✗
Predicted Not Spam	FN ✗	TN ✓

✓ Prediction matches reality

✗ Prediction does not matches reality

What are the consequences in FN and FP?

Machine Learning Overview

Medical Diagnosis

Accuracy: 90%
Precision: 55.7%
Recall: 83.3%
F1 Score: 66.76%

Spam Detector

Accuracy: 80%
Precision: 76.9%
Recall: 37%
F1 Score: 49.96%

F1 Score = Harmonic Mean = (2*P*R) / (P + R)

Why is this better?

Machine Learning Overview

Strengths & Weaknesses

Strengths

- Learning simple concepts
- Where there is lots of data

Weaknesses

- Learning complex concepts from small amount of data
- Perform on new types of data

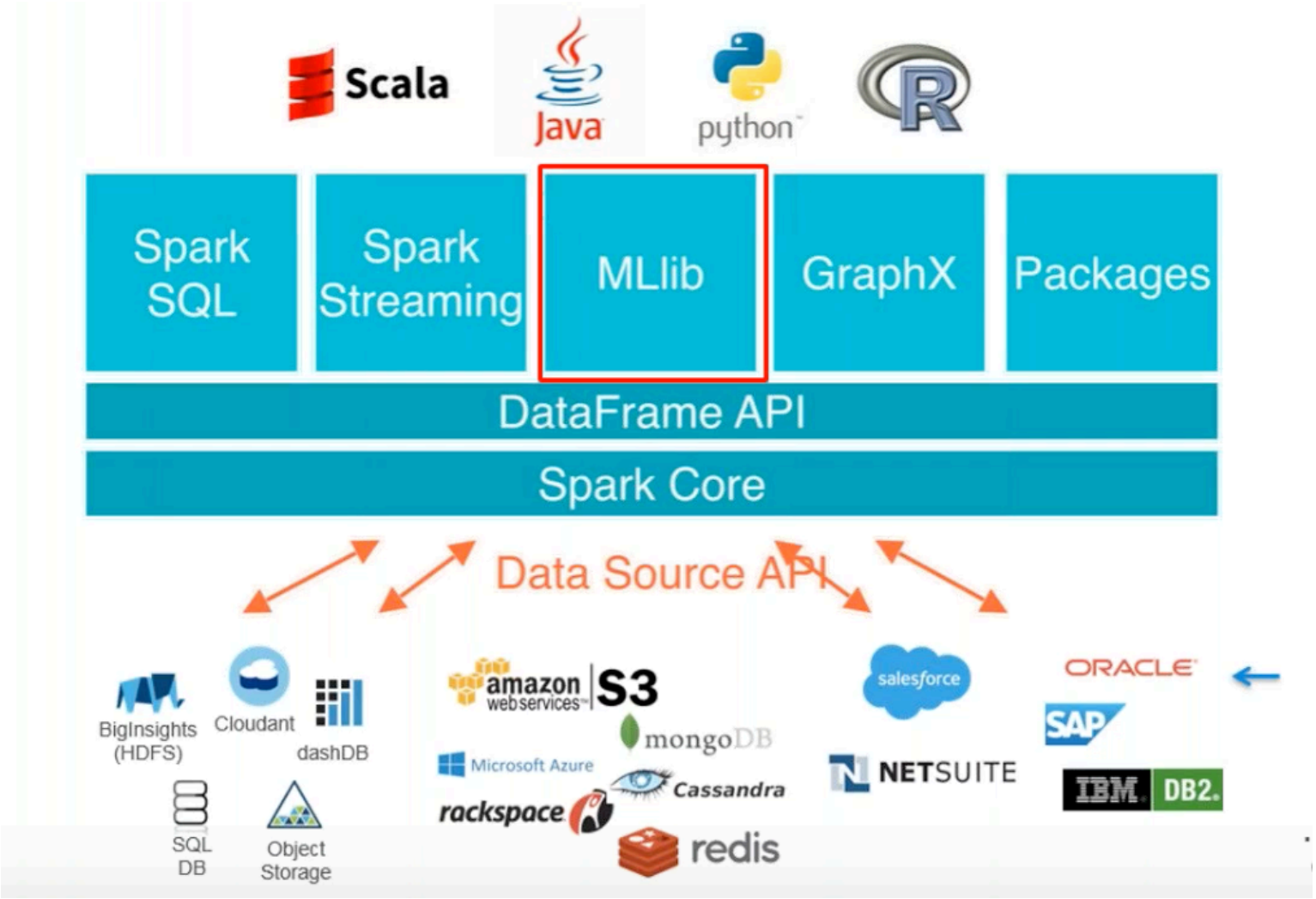
Machine Learning Overview

- Skills for ML Beginners
  - Let the data speak
  - Learn to spot the better model
  - Be suspicious of your conclusion
  - Build many models to throw away
  - Don't be afraid to change the game
  - Start small and go big
  - Domain knowledge still matters
  - Coding skills still matter

Machine Learning Overview

- Tips for ML Beginners
  - Data-driven instead of design-driven
  - Look from the empirical point of view
  - Rely on experimental proofs
    - Let the data speak
  - No tight specifications
    - Only data that represents the past experience
  - Build a system that will work in the future
    - How to tell it is working?

Spark MLlib Overview



Spark MLlib Overview

Practical machine learning scalable and easy

Simplify the development and deployment of scalable machine learning pipelines

Spark MLlib Overview

MLlib Components

Algorithms
- Classification
- Regression
- Clustering
- Collaborative Filtering

Pipeline
- Constructing
- Evaluating
- Tuning
- Persistence

Featurization
- Extraction
- Transformation

Utilities
- Linear algebra
- Statistics

Spark MLlib Overview

ML Algorithms

Classification	Regression	Other
Logistic Regression	Linear Regression	K-means
Decision Tree	Decision Tree	Bisecting k-means
Random Forest	Random Forest	Gaussian Mixture
Gradient-boosted Tree	Gradient-boosted Tree	
Linear SVM	Survival Regression	Alternate Least Square
Naïve Bayes	Isotonic Regression	

Spark Mllib Overview

The Tale of Two Libraries

spark.mllib

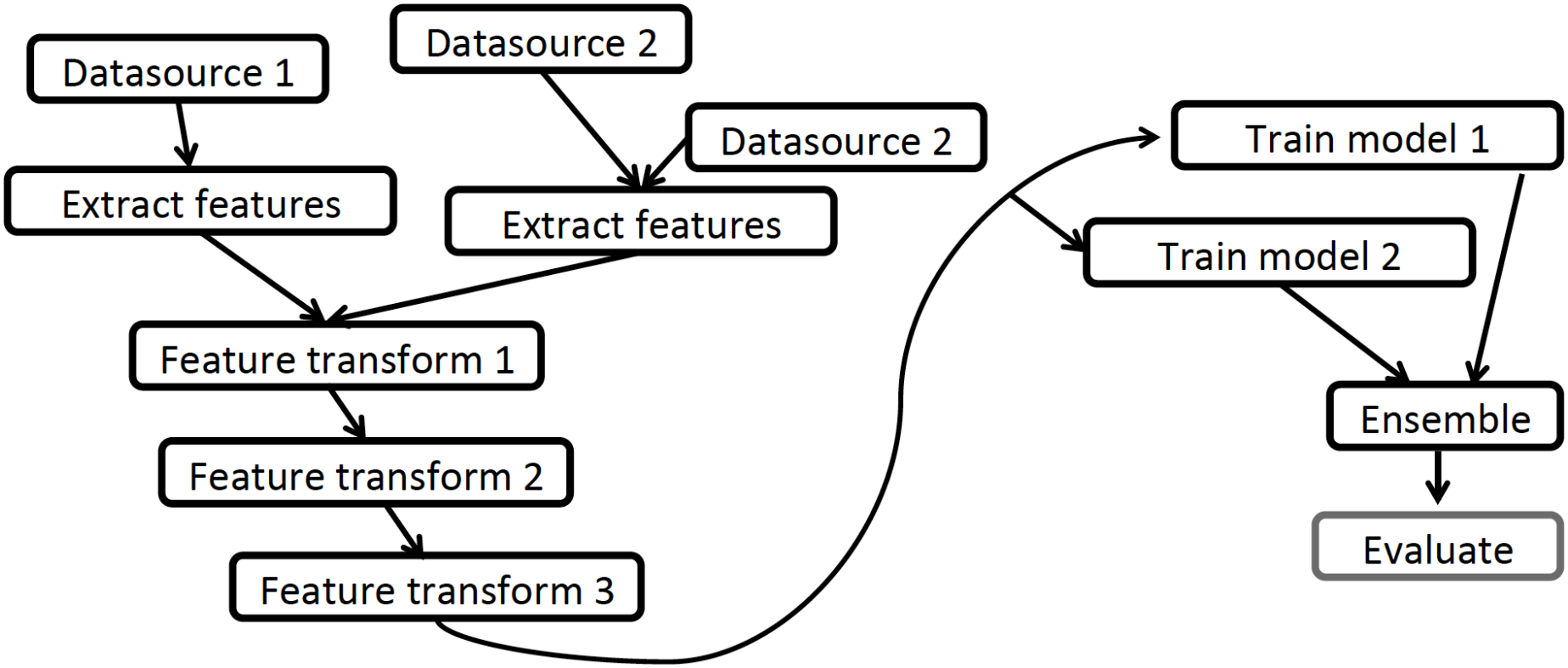
RDD

spark.ml

DataFrame

Spark MLlib Overview

ML Workflows are complex



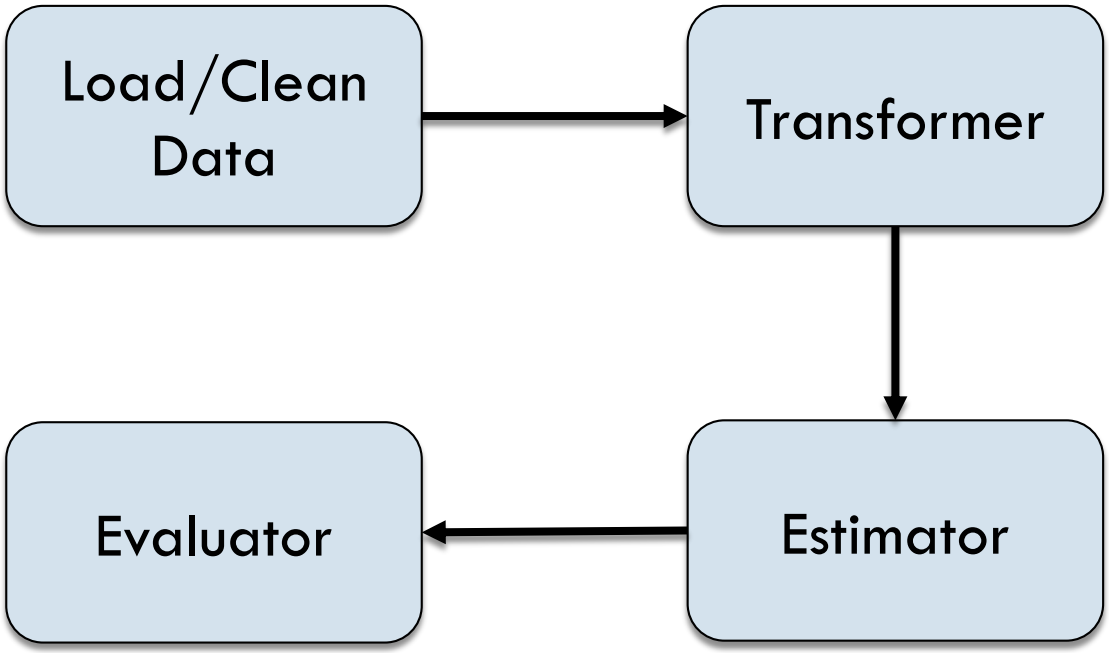
https://spark-summit.org/2015/events/practical-machine-learning-pipelines-with-mllib-2/

Spark MLlib Overview

- ML Pipelines
  - Inspired by scikit-learn
  - A new high level APIs with integration with DataFrames
  - A sequence of stages (data preprocessing, feature extraction, model fitting and validation)
  - Each transformation takes an input and produces an output, which becomes the input to the next stage
  - Data import/export is the start/end of a ML pipeline

Spark MLlib Overview

Mllib Pipeline Concepts



Spark MLlib Overview

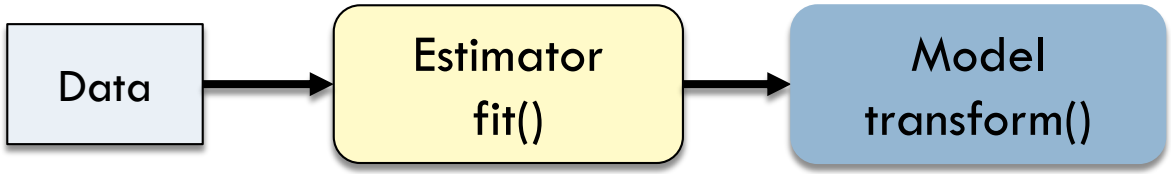
- Transformer
  - Preprocessing step of feature extraction
  - Transforming data into consumable format
  - Take input column, transform it to an output column
  - Examples
    - Normalize the data
    - Tokenization – sentences into words
    - Converting categorical values into numbers

Spark MLlib Overview

- Transformer Examples
  - Working with text data
    - Tokenization, removing stop words, n-grams
    - Term frequency
  - Working with categorical data
    - Indexing (convert string to index)
  - Working with numeric data
    - Bucketizing, scaling or normalizing data

Spark MLlib Overview

- Estimator
  - Another kind of Transformer
    - Transform data by requiring two passes over data
  - Learning algorithm that trains (fits) on data
  - Return a model, which is a type of Transformer
  - Examples
    - LogisticRegression.fit() => LogisticRegressionModel

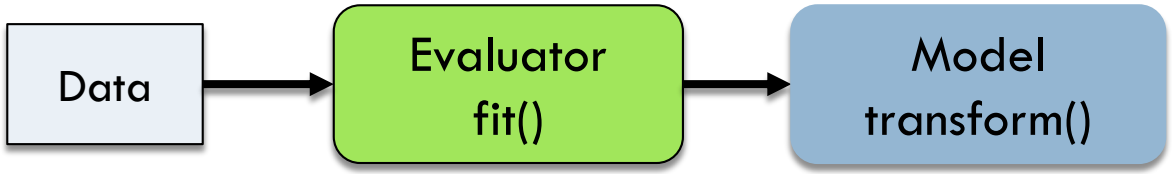


Spark MLlib Overview

- Estimator Examples
  - Algorithm - classification, regression, etc.
  - One hot encoding (covert string to an array of indexes)
    - To remove implied ordering
  - Word to vector
    - Map each word to a unique fixed–size vector
  - Pipeline

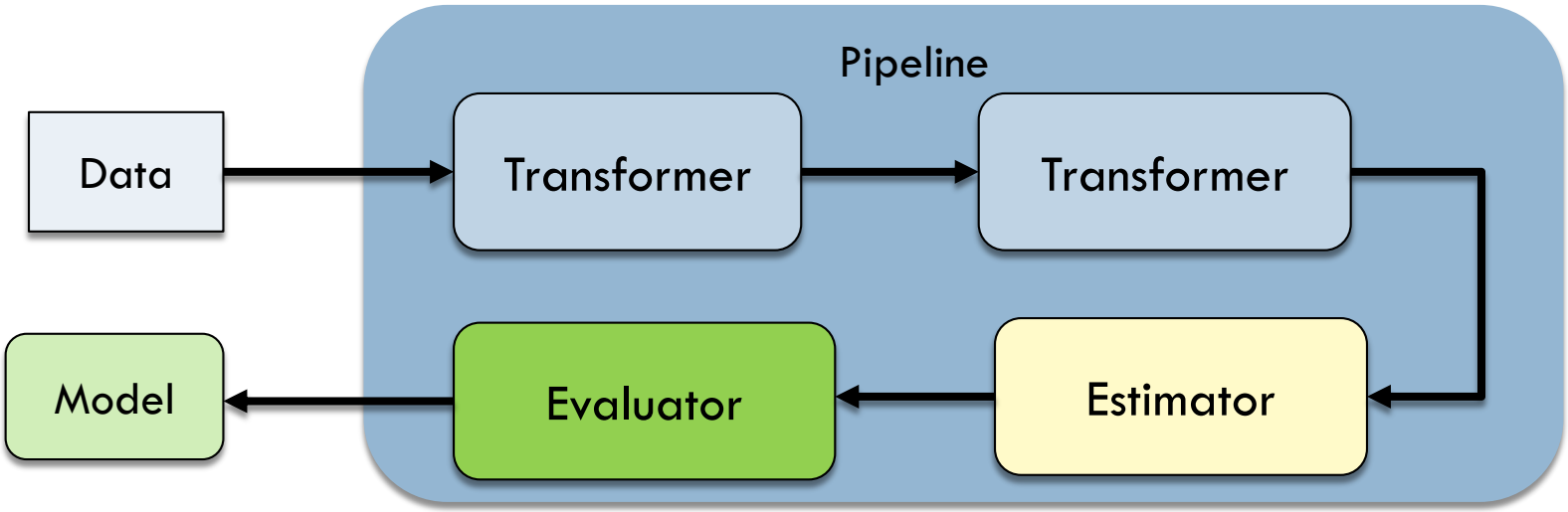
Spark MLlib Overview

- Evaluator
  - Evaluate the model performance based certain metric
    - ROC, RMSE
  - Help with automating the model tuning process
    - Comparing the model performance
    - Select the best model for generating predictions
  - Examples
    - BinaryClassificationEvaluator, CrossValidator



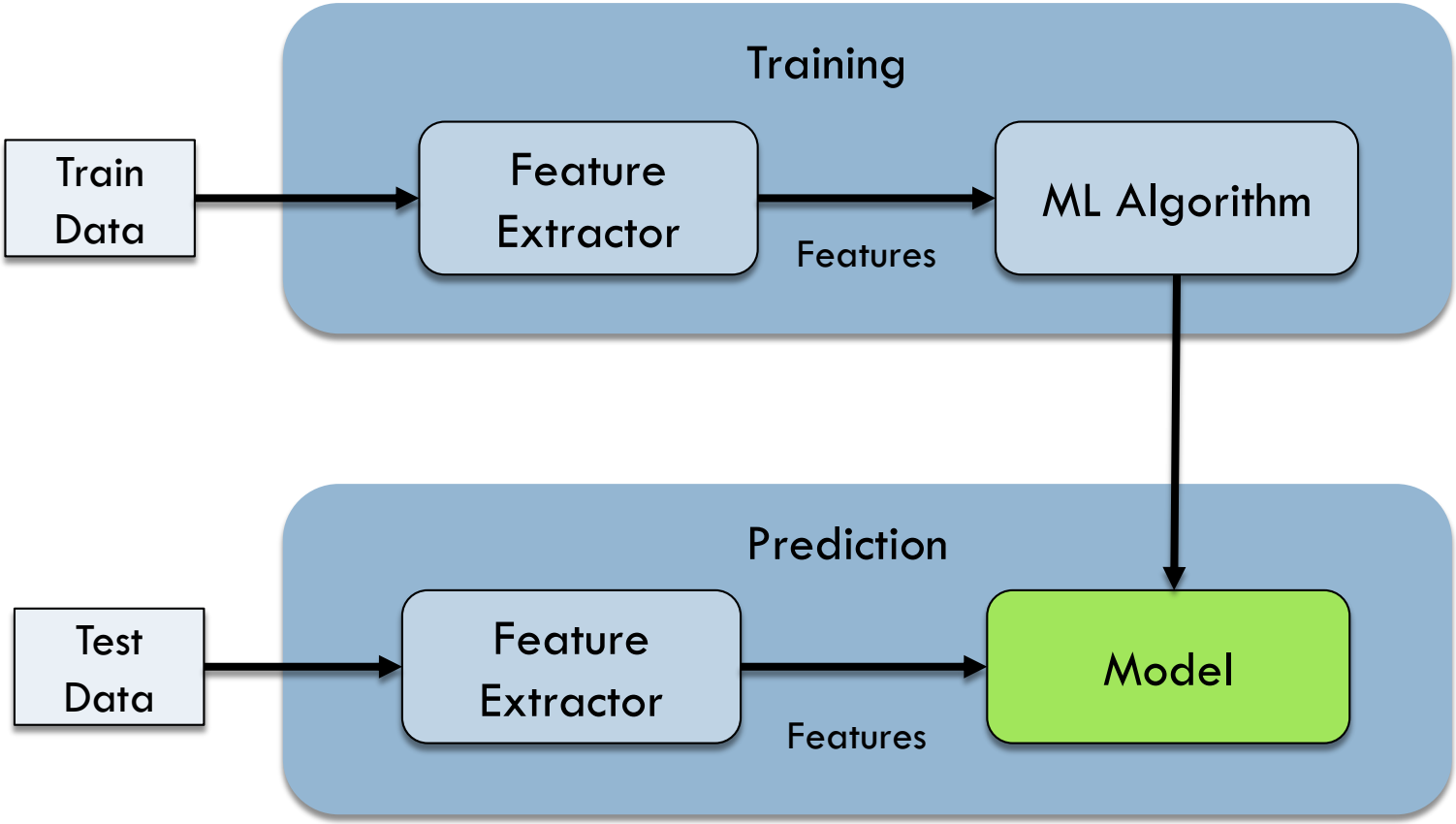
Spark MLlib Overview

- Pipeline
  - To represent a ML workflow
  - Consist of a set of stages
  - Leverage the uniform API of Transformer & Estimator
  - A type of Estimator
  - Can be persisted



Spark MLlib Overview

Training & Prediction Pipeline



Let's have fun with Machine Learning

MLlib Classification with Flight Data

MLlib Regression with HK Properties Value