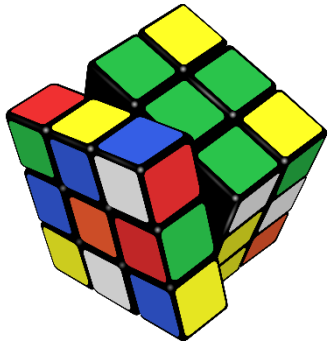


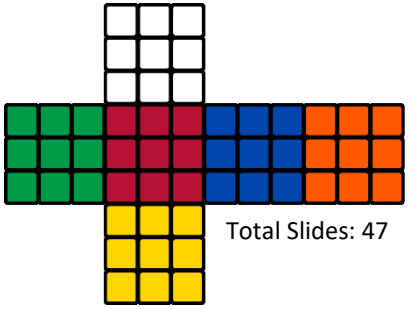
Big Data Engineering For Analytics

Analytics using Spark Machine Learning

Dr LIU Fan

(isslf@nus.edu.sg)

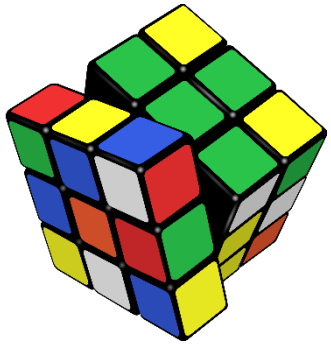
NUS-ISS

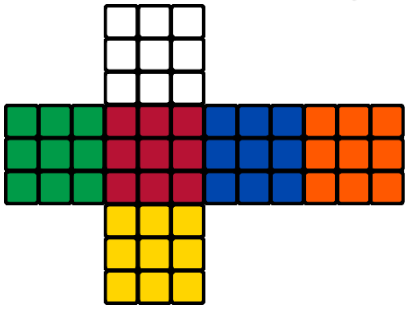


Total Slides: 47

Learning Objectives

- Understand the machine learning Spark special libraries.
- What is Spark ML and Mllib?
  - Understanding Machine Learning
  - Understanding ML Algorithms
  - Understanding Spark Libraries
  - Introducing Mllib and ML





Introducing Machine Learning

A breakthrough in machine learning will worth ten Microsoft.
~Bill Gates

Machine Learning

- Most programs tell computers exactly what to do
  - Database transactions and queries
  - Controllers
  - Phone systems, manufacturing processes, transport, weaponry, etc.
  - Media delivery
  - Simple search
  - Social systems
    - Chat, blogs, email, etc.
- An alternative technique is to have computers *learn* what to do
- Machine Learning refers to programs that leverage collected data to drive future program behavior
- This represents another major opportunity to gain value from data

What is Machine Learning?

- Machine learning is a method of data analysis that automates analytical model building.
- Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Machine Learning Definition

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

Classifying emails as spam or not spam.

Watching you label emails as spam or not spam.

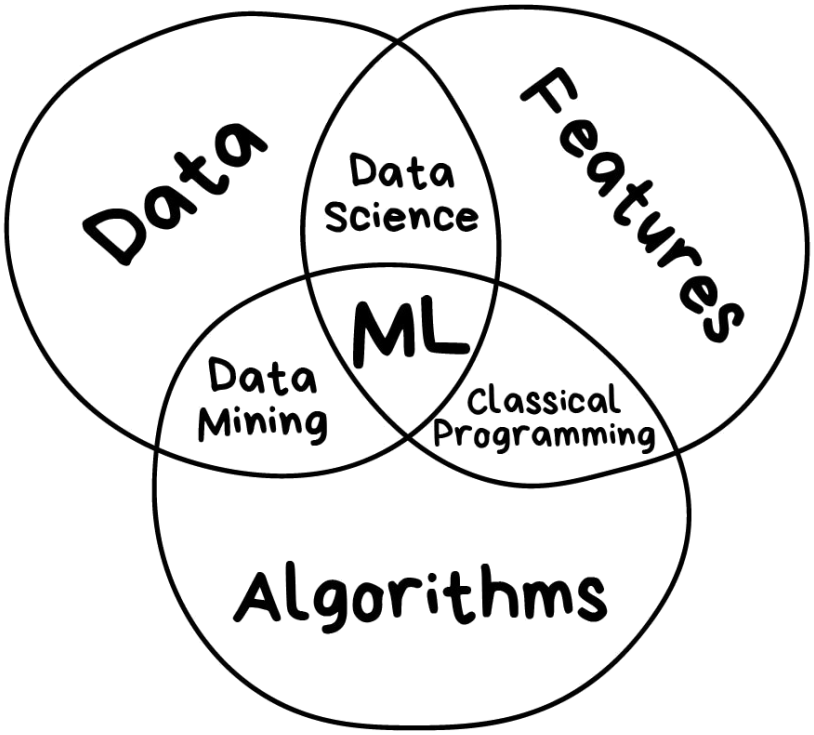
The number (or fraction) of emails correctly classified as spam/not spam.

Machine Learning Applications

- Fraud detection.
- Web search results.
- Real-time ads on web pages
- Credit scoring and next-best offers.
- Prediction of equipment failures.
- New pricing models.
- Network intrusion detection.
- Recommendation Engines
- Customer Segmentation
- Text Sentiment Analysis
- Predicting Customer Churn
- Pattern and image recognition.
- Email spam filtering.
- Financial Modeling

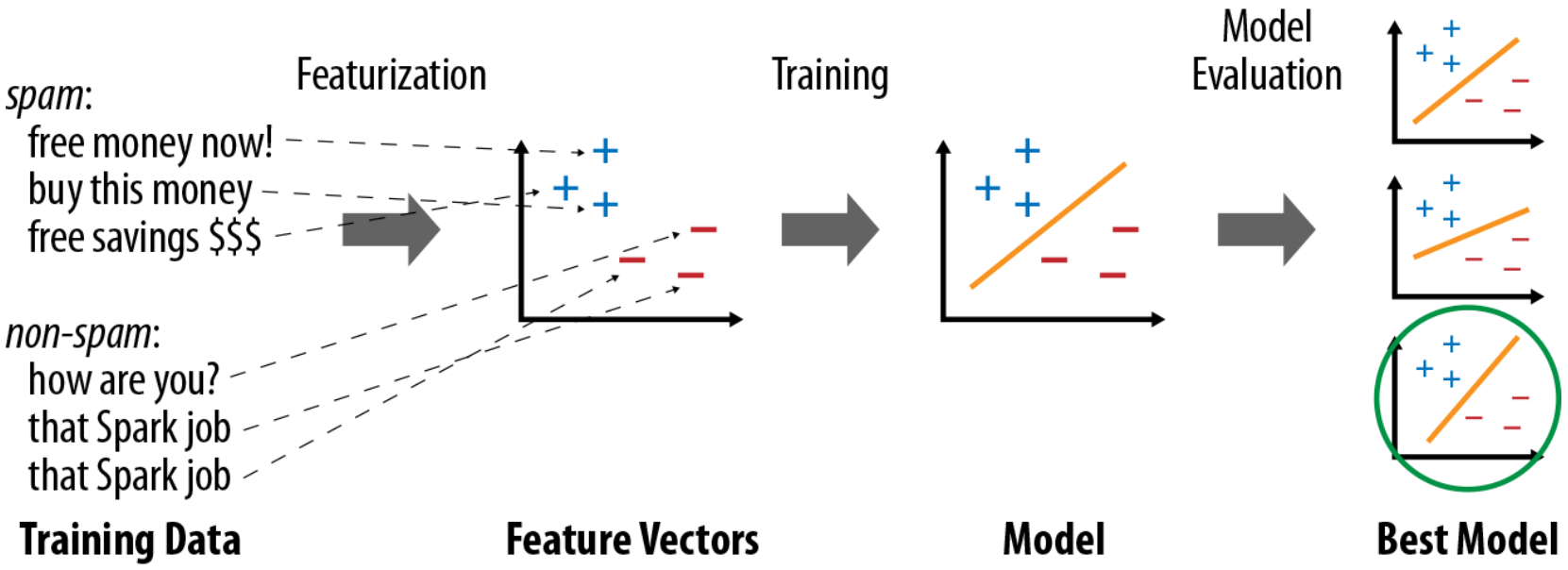
Three components of Machine Learning

- The only goal of machine learning is to predict results based on incoming data.
- The greater the variety of sample in hand the easier it gets to find relevant patterns and hence predict results accurately.
  - **Data**: The more and diverse the data, the better the result. There are two main ways of collecting data — manual and automatic.
  - **Features**: Also known as parameters or variables. These are the factors for a machine to look at. Picking the right set of features is an iterative time consuming process.
  - **Algorithms**: The method chosen to predict results and it affects the precision, performance, and size of the final model.



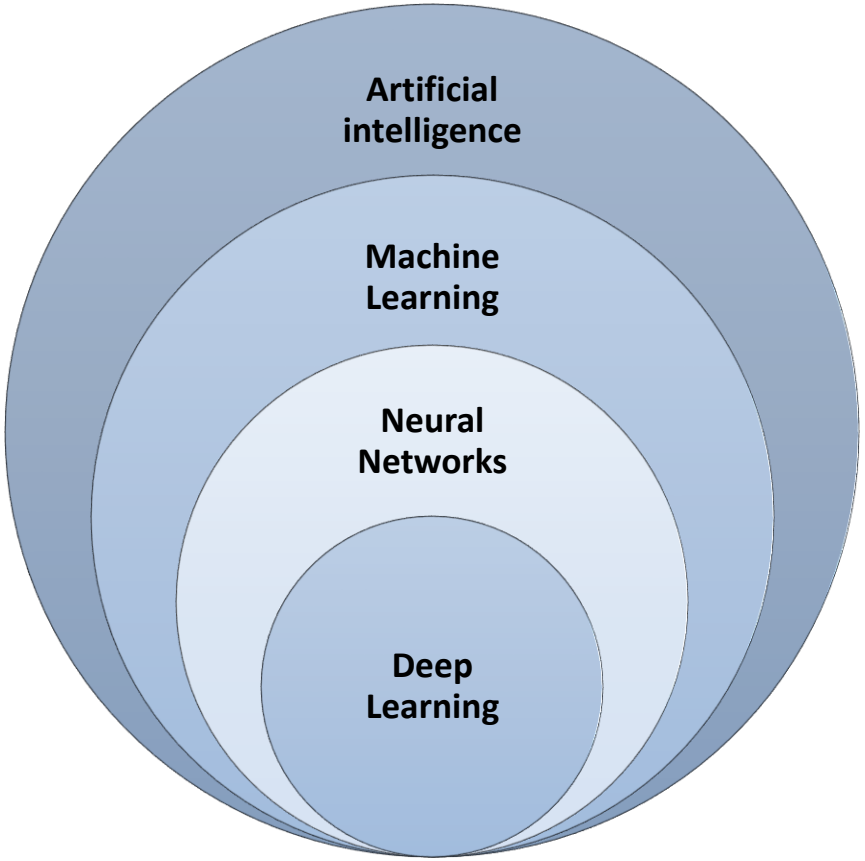
Typical steps in a machine learning pipeline

- Machine learning algorithms attempt to make predictions or decisions based on ***training*** *data*
  - All learning algorithms require defining set of ***features*** for each item, which will be fed into the **learning** function
  - Real-world ML pipelines will ***train*** multiple versions of a model and ***evaluate*** each one.

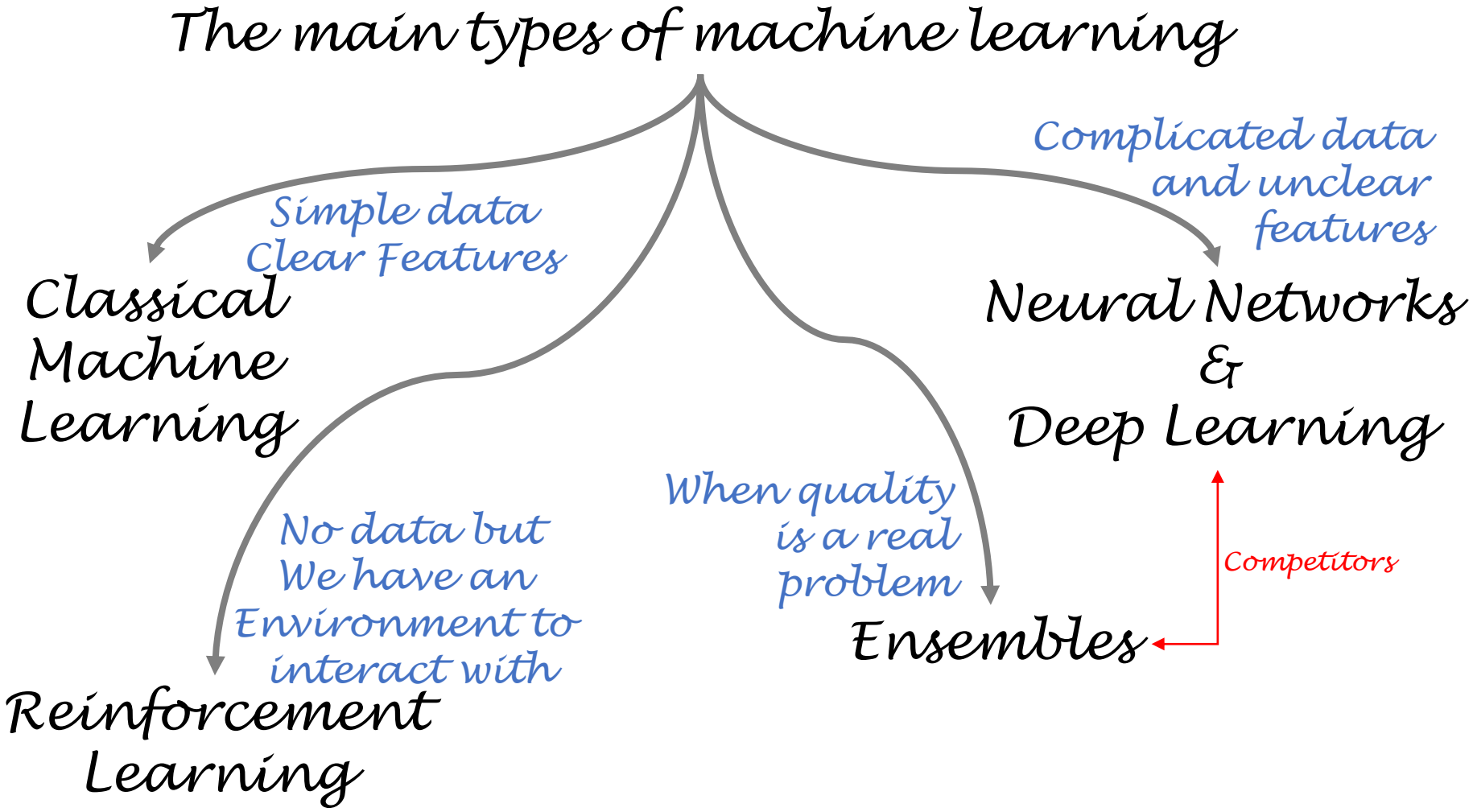


Learning, Intelligence and other jargons

- **Artificial intelligence** is the name of a whole knowledge field.
- **Machine Learning** is a part of artificial intelligence. Important, but not the only one.
- **Neural Networks** is one of machine learning types.
- **Deep Learning** is a modern method of building, training, and using neural networks.



Types of machine learning



CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

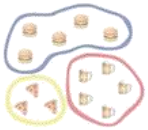
Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»



ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Supervised Learning

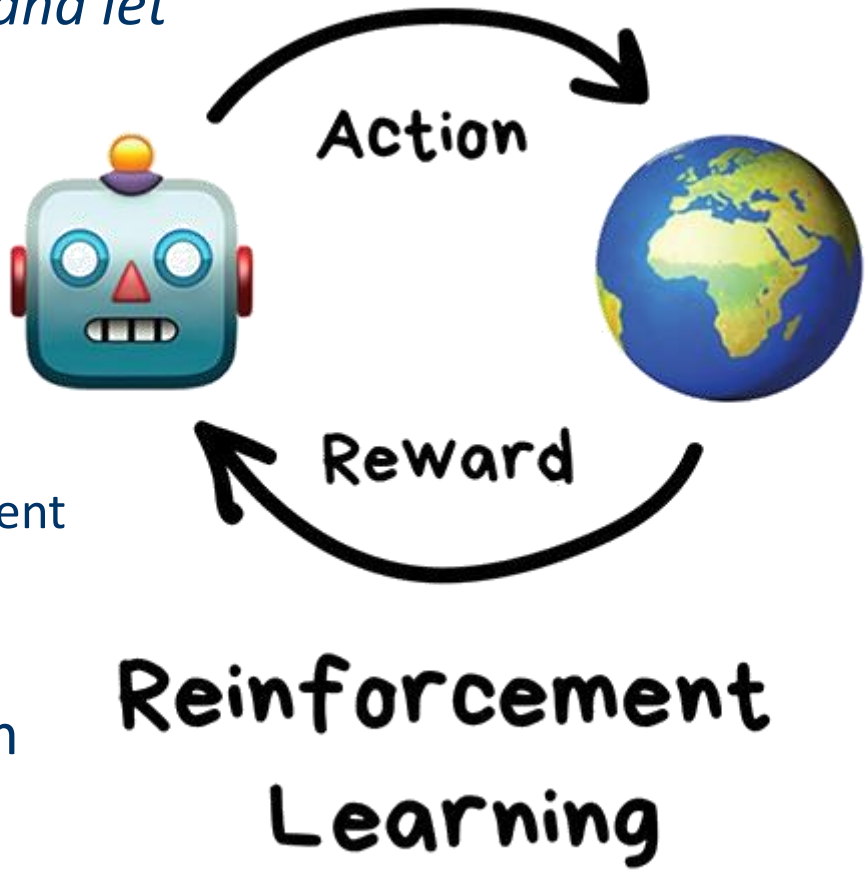
- **Supervised learning** algorithms are trained using labeled examples, such as an input where the desired output is known.
- The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors.
- For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.
- Or it can attempt to predict the price of a house based on different features for houses for which we have historical price data.

Unsupervised Learning

- **Unsupervised learning** is used against data that has no historical labels.
- The system is not told the "right answer." The algorithm must figure out what is being shown.
- The goal is to explore the data and find some structure within.
- For example, it can find the main attributes that separate customer segments from each other.

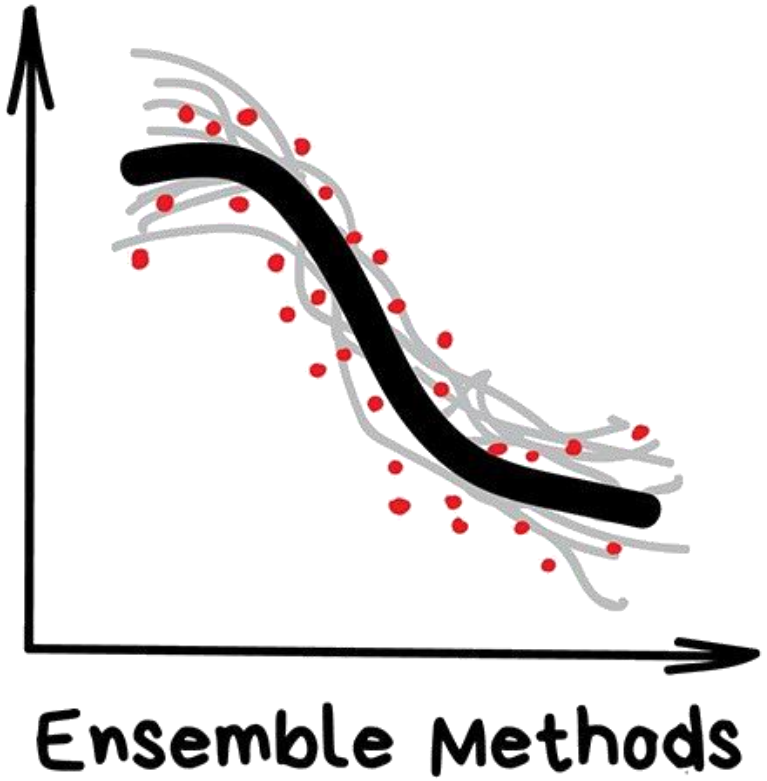
Reinforcement Learning

- *"Throw a robot into a maze and let it find an exit"*
- Nowadays used for:
  - Self-driving cars
  - Robot vacuums
  - Games
  - Automating trading
  - Enterprise resource management
- Popular algorithms: Q-Learning, SARSA, DQN, A3C, Genetic algorithm



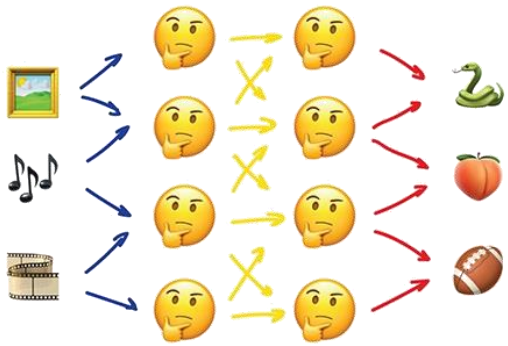
Ensemble Methods

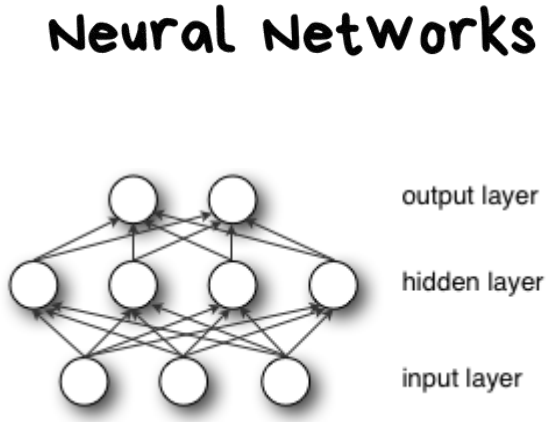
- Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.
- Nowadays is used for:
  - Everything that fits classical algorithms approaches (but works better)
  - Search systems (★)
  - Computer vision
  - Object detection
- Popular algorithms: Random Forest, Gradient Boosting

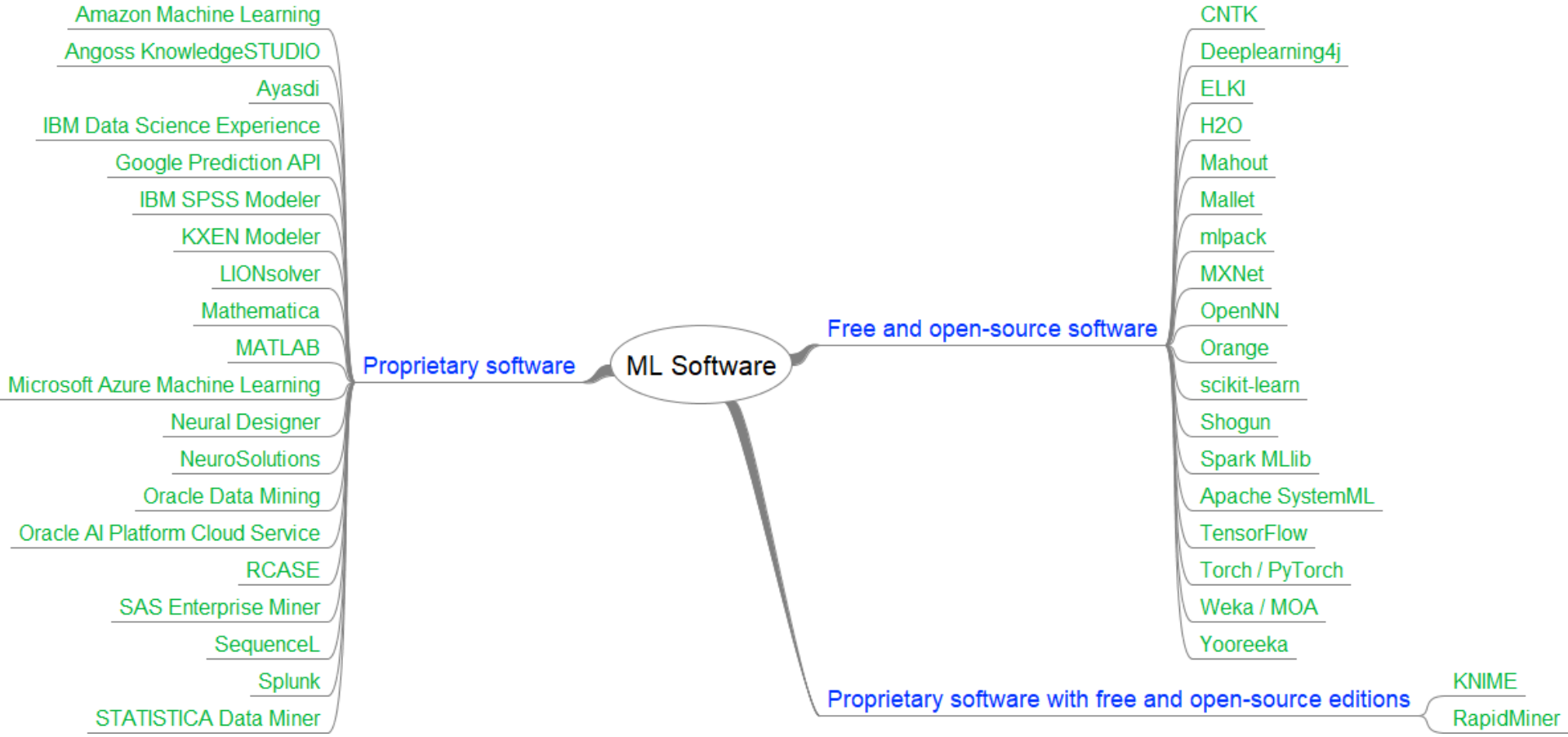


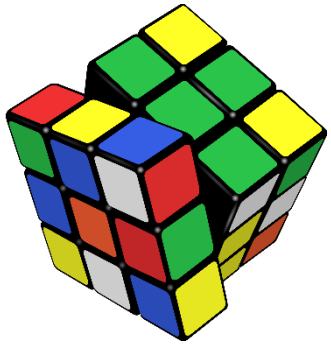
Neural Networks and Deep Learning

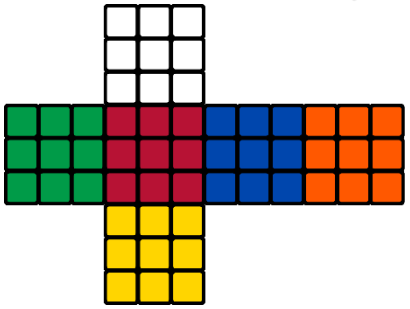
- Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns.
  - They interpret sensory data through a kind of machine perception, labeling or clustering raw input.
  - The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.
- Deep-learning networks are distinguished from single-hidden-layer neural networks by their depth; that is, the number of node layers through which data must pass in a multistep process of pattern recognition.
- Used today for:
  - Replacement of all algorithms above
  - Object identification on photos and videos
  - Speech recognition and synthesis
  - Image processing, style transfer
  - Machine translation
- Popular architectures: Perceptron, Convolutional Network (CNN), Recurrent Networks (RNN), Autoencoders











Spark Machine Learning

"War is 90% information."

*~**Napoleon Bonaparte**, French military and political leader*

Machine Learning On Spark

- Highly computation intensive and iterative
- Many traditional numerical processing systems do not scale to very large datasets
  - e.g., MatLab
- Mllib, ML and H2O part of Apache Spark
- Includes many common ML functions
  - ALS (alternating least squares)
  - k-means
  - Logistic Regression
  - Linear Regression
  - Gradient Descent

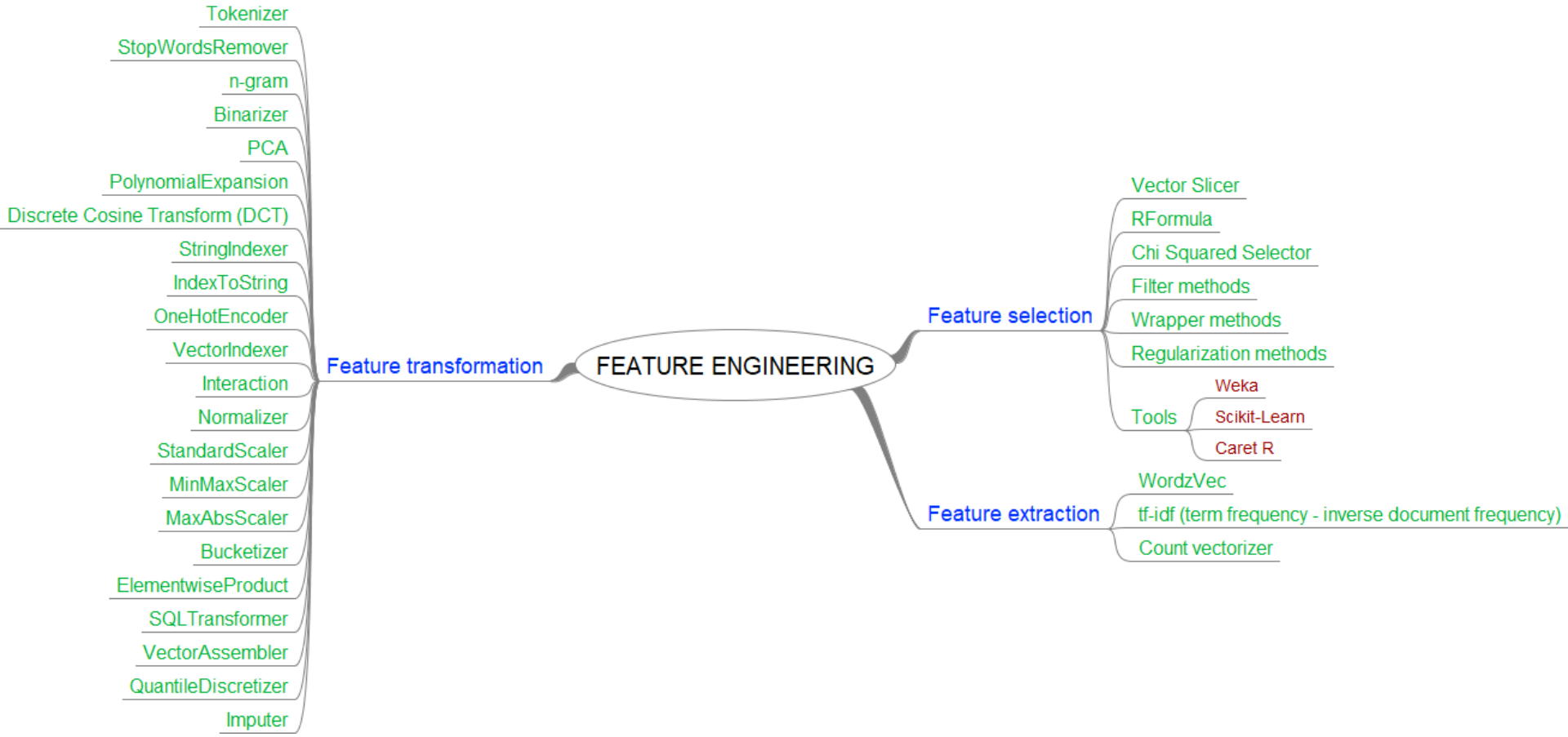
Spark Machine Learning Libraries

- Spark provides two machine learning libraries, **MLlib** and **Spark ML** (also known as the Pipelines API).
  - MLlib extends Spark for machine learning and statistical analysis. It provides a higher-level API than the Spark core API for machine learning and statistical analysis.
  - Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow.
- These libraries provide the following:
  - A set of common machine learning algorithms, including regression, classification, clustering, and collaborative filtering.
  - Features such as extraction, dimensionality reduction, transformation, and pipelines.

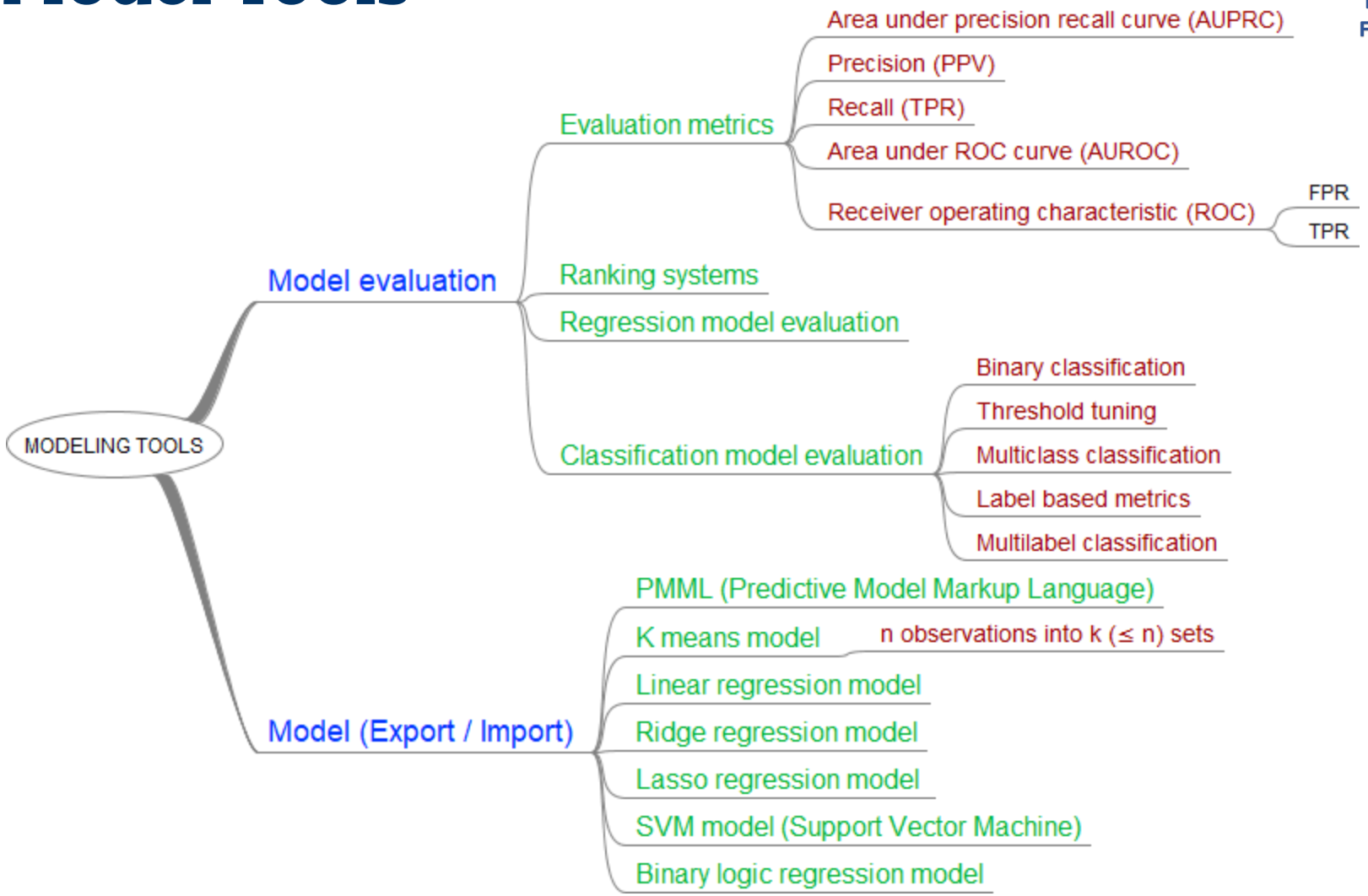
MLlib and ML Libraries

- MLlib provides a set of data types that uses RDD to represent data points.
  - Some of the common datatypes are vectors and labeled points.
- Spark's ML is another library that makes use of the DataFrame API.
  - All the new features have now been added to the ML library, and MLlib is now kept in maintenance mode.
  - Apart from using structured APIs, Spark's ML lets you define the machine- learning *pipeline,* which is similar to the pipeline concept in scikit-learn.
- The pipeline API provides two main core features:
  - **Transformer**: A transformer takes a DataFrame as input and produces a new DataFrame as an output
  - **Estimator**: An estimator is an algorithm that uses a DataFrame to produce a new transformer

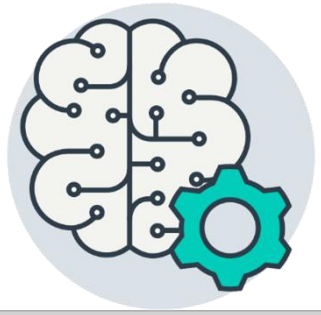
Spark Feature Engineering Libraries



Model Tools



Machine Learning

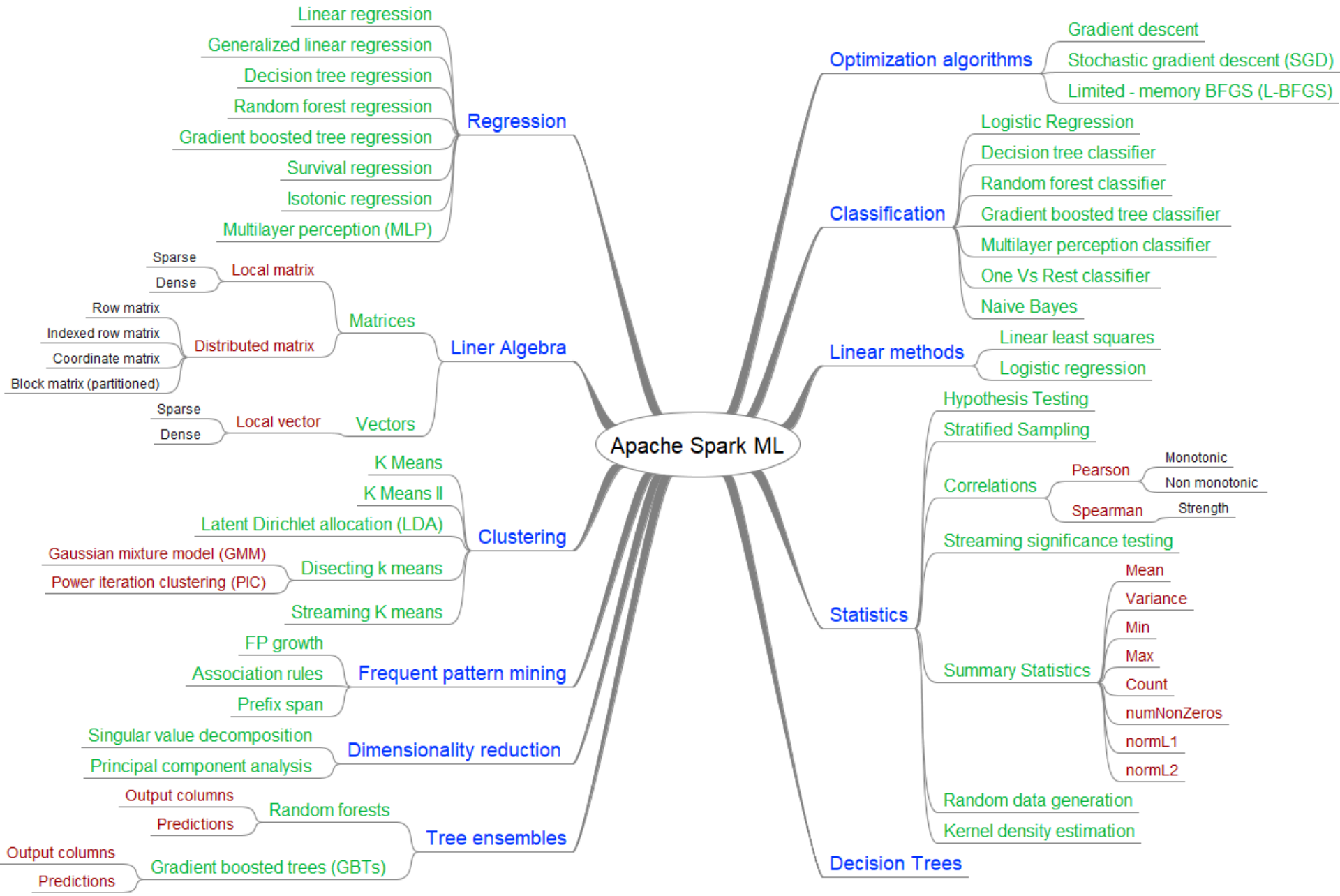


Supervised

- Classification
  - Naïve Bayes
  - SVM
  - Random Decision Forests
  - . . .
- Regression
  - Linear
  - Logistics
  - . . .

Unsupervised

- Clustering
  - K Means
  - . . .
- Dimensionality Reduction
  - Principal Component Analysis
  - SVD
  - . . .



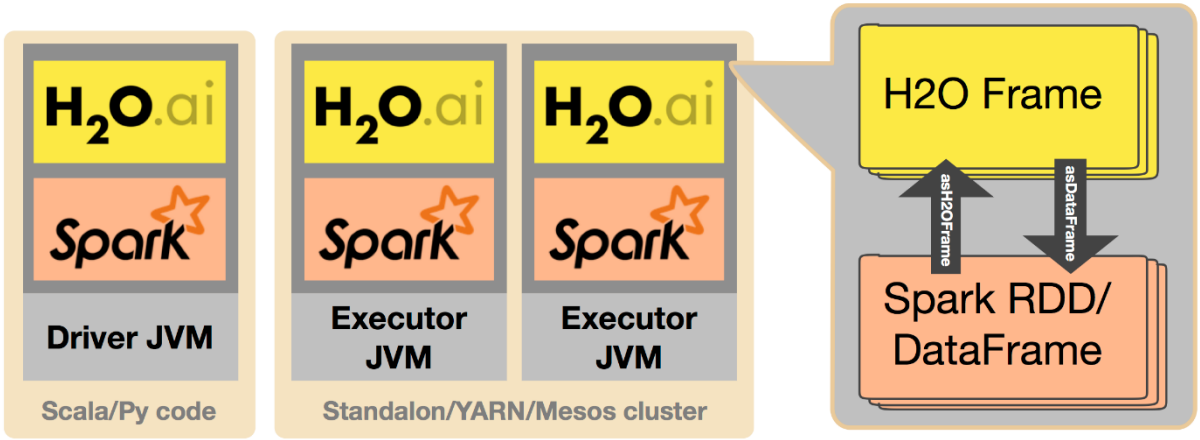
ML Algorithms vs H2O

- Decision Trees
- Statistics
- Model evaluation
- Regression
- Clustering
- Optimization algorithms
- Classification
- Linear methods
- Features
- Model (Export / Import)
- Tree ensembles

H2O and ML share many of the same algorithms but differ in both their implementation and functionality. Also H2O provides grid search for hyper parameters.

Design of Sparkling Water

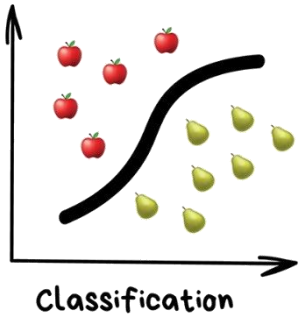
- H2O starts services, including a distributed key-value (K/V) store and memory manager, and orchestrates them into a cloud
- Sparkling Water enables transformation between different types of RDDs/DataFrames and H2O's frame, and vice versa
- Data stored in an H2O frame is heavily compressed and does not need to be preserved as an RDD anymore



The 'Three Cs'

- There are well-established categories of techniques for exploiting data
  - Collaborative filtering (recommendations)
  - Clustering
  - Classification
- Beyond that we have techniques specializing in
  - Regression
  - Anomaly detection
  - Recommendation
  - Dimensionality reduction
  - Neural Processing with Artificial Neural Network

Classification

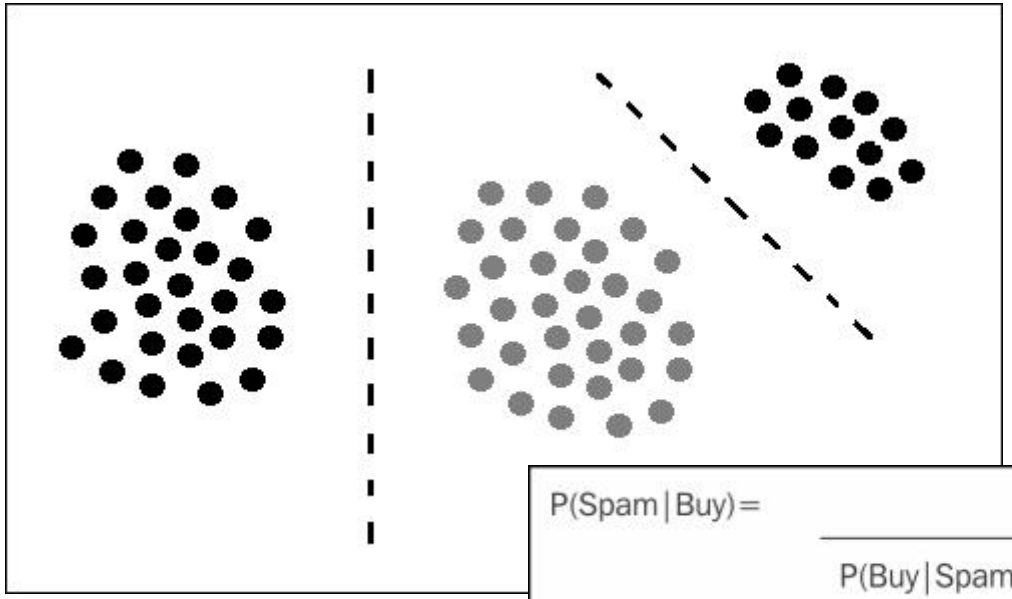


- Classification algorithm discovers groups
  - Classification is a form of 'supervised' learning
  - **Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Trees, Ensembles and Neural Network, K-Nearest Neighbor**
- The goal while solving a classification problem is to predict a class or category for an observation
  - A classification system takes a set of data records with known labels
  - Learns how to label new records based on that information
- Examples
  - Given a set of e-mails identified as spam/not spam, label new e--mails as spam/not spam
  - Given tumors identified as benign or malignant, classify new tumors

Classification with Naïve Bayes

Theory

In order to use the Naïve Bayes algorithm to classify a data set, the data must be linearly divisible, that is, the classes within the data must be linearly divisible by class boundaries.



What is the probability that an e-mail that contains the word buy is spam? Well, this would be written as **P (Spam|Buy)**. Naïve Bayes says that it is described by the equation in the below:

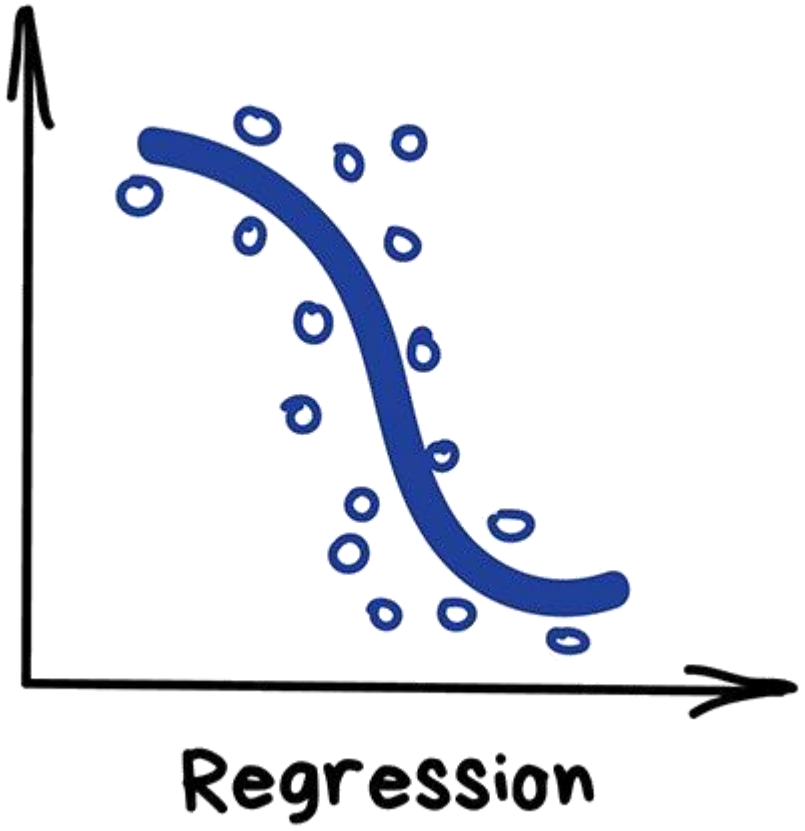
$$P(Spam|Buy) = \frac{P(Buy|Spam) * P(Spam)}{P(Buy|Spam) * P(Spam) + P(Buy|Not\ Spam) * P(Not\ Spam)}$$

Classification Use Cases

- Predicting disease
  - A doctor or hospital might have a historical dataset of behavioral and physiological attributes of a set of patients.
  - They could use this dataset to train a model on this historical data and then leverage it to predict whether or not a patient has heart disease or not.
  - This is an example of binary classification (healthy heart, unhealthy heart) or multiclass classification (healthly heart, or one of several different diseases).
- Classifying images
  - There are a number of applications from companies like Apple, Google, or Facebook that can predict who is in a given photo by running a classification model that has been trained on historical images of people in your past photos.
  - Another common use case is to classify images or label the objects in images.
- Predicting customer churn
  - A more business-oriented use case might be predicting customer churn—that is, which customers are likely to stop using a service. You can do this by training a binary classifier on past customers that have churned (and not churned) and using it to try and predict whether or not current customers will churn.
- Buy or won't buy
  - Companies often want to predict whether visitors of their website will purchase a given product. They might use information about users' browsing pattern or attributes such as location in order to drive this prediction.

Regression

- Regression is about predicting real values from observations.
  - Unlike classification, the predicted value is not discrete, but rather it is continuous.
- Used for:
  - Stock price forecast
  - Demand and sales volume analysis
  - Medical diagnosis
  - Any number-time correlations
- Popular algorithms are Linear and Polynomial regressions.

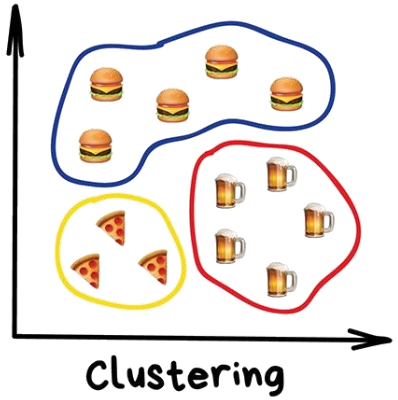


Regression Use Cases

- Predicting sales
  - A store may want to predict total product sales on given data using historical sales data. There are a number of potential input variables, but a simple example might be using last week's sales data to predict the next day's data.
- Predicting height
  - Based on the heights of two individuals, we might want to predict the heights of their potential children.
- Predicting the number of viewers of a show
  - A media company like Netflix might try to predict how many of their subscribers will watch a particular show.

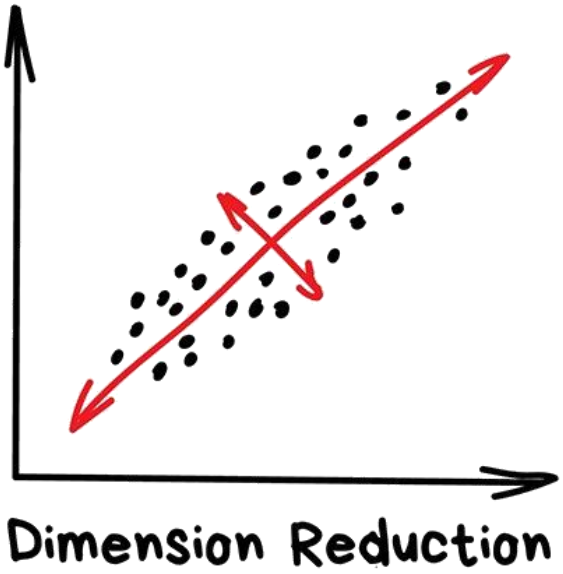
Clustering

- In clustering, a dataset is split into a specified number of clusters or segments.
  - Elements in the same cluster are more similar to each other than to those in other clusters.
- Popular algorithms: **K-means_clustering, Mean-Shift, DBSCAN**
- Clustering algorithms discover structure in collections of data
  - Where no formal structure previously existed
  - They discover what clusters, or groupings, naturally occur in data
- Examples
  - Finding related news articles
  - Computer vision (groups of pixels that cohere into objects)
  - Customer Segmentation
  - Market segmentation (types of customers, loyalty)
  - To merge close points on the map
  - For image compression
  - To analyze and label new data
  - To detect abnormal behavior



Dimensionality Reduction

- Assembles specific features into more high-level ones
- Nowadays is used for:
  - Recommender systems
  - Beautiful visualizations
  - Topic modeling and similar document search
  - Fake image analysis
  - Risk management



- Popular algorithms: Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA, pLSA, GLSA), t-SNE (for visualization)

Recommenders -Collaborative Filtering

- Collaborative Filtering is a technique for recommendations
  - Example application: given people who each like certain books, learn to suggest what someone may like in the future based on what they already like
- Helps users navigate data by expanding to topics that have affinity with their established interests
- Collaborative Filtering algorithms are agnostic to the different types of data items involved
  - Useful in many different domains

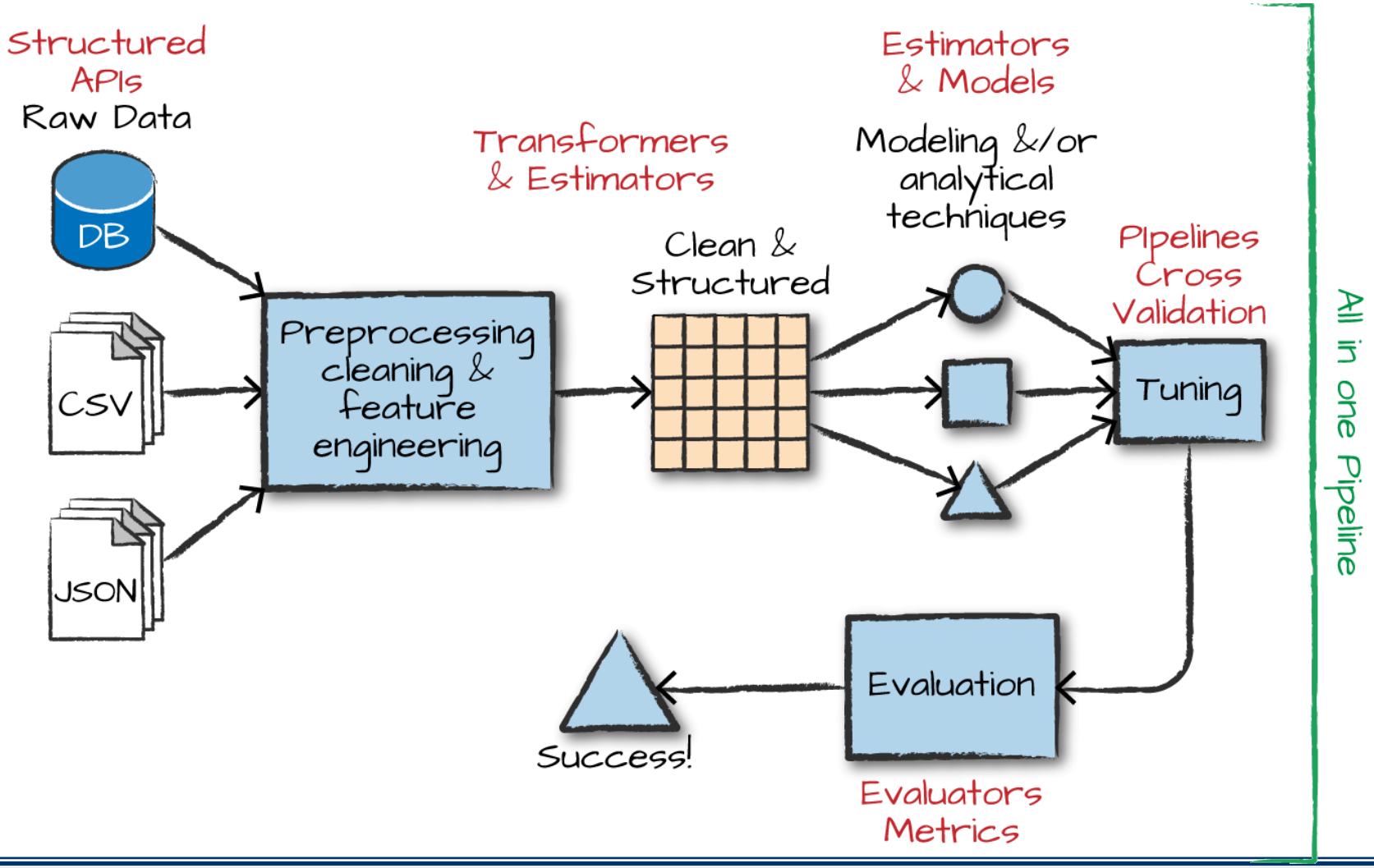
Recommendation – Use Cases

- Movie recommendations
  - Netflix uses Spark, although not necessarily its built-in libraries, to make large-scale movie recommendations to its users.
  - It does this by studying what movies users watch and do not watch in the Netflix application.
  - In addition, Netflix likely takes into consideration how similar a given user's ratings are to other users'.
- Product recommendations
  - Amazon uses product recommendations as one of its main tools to increase sales.
  - Based on the items in our shopping cart, Amazon may recommend other items that were added to similar shopping carts in the past.
  - Likewise, on every product page, Amazon shows similar products purchased by other users.

More Unsupervised Learning Use Cases

- Anomaly detection
  - Given some standard event type often occuring over time, we might want to report when a nonstandard type of event occurs.
  - For example, a security officer might want to receive notifications when a strange object (think vehicle, skater, or bicyclist) is observed on a pathway.
- User segmentation
  - Given a set of user behaviors, we might want to better understand what attributes certain users share with other users.
  - For instance, a gaming company might cluster users based on properties like the number of hours played in a given game.
  - The algorithm might reveal that casual players have very different behavior than hardcore gamers, for example, and allow the company to offer different recommendations or rewards to each player.
- Topic modeling
  - Given a set of documents, we might analyze the different words contained therein to see if there is some underlying relation between them.
  - For example, given a number of web pages on data analytics, a topic modeling algorithm can cluster them into pages about machine learning, SQL, streaming, and so on based on groups of words that are more common in one topic than in others.

The machine learning workflow, in Spark

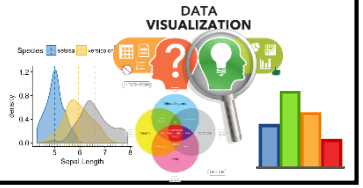


Typical steps in a machine-learning project

1. Data Collection

2. Data clean up and preparation

3. Data analysis and visualization



4. Training the model

The Model

5. Model Evaluation

The Model

New Data

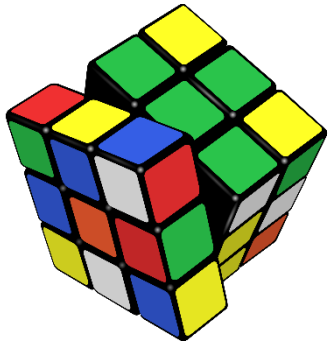
6. Use the model for predictions

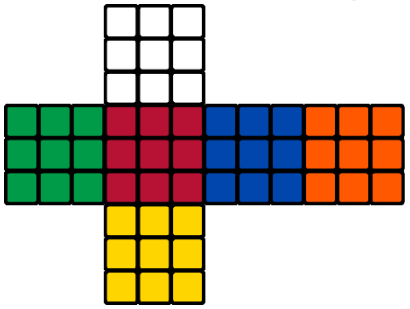
Project consists of multiple steps - 1

1. Collecting data
   - First the data needs to be gathered from various sources.
   - The sources can be log files, database records, signals coming from sensors, and so on.
   - Spark can help load the data from relational databases, CSV files, remote services, and distributed file systems like HDFS, or from real-time sources using Spark Streaming.
2. Cleaning and preparing data
   - Data isn't always available in a structured format appropriate for machine learning (text, images, sounds, binary data, and so forth). At times unstructured data is transformed into numerical features.
   - Additionally, you need to handle missing data and the different forms in which the same values can be entered (for example, VW and Volkswagen are the same carmaker).
   - Often, data also needs to be scaled so that all dimensions are of comparable ranges.

Project consists of multiple steps – 2.

3. Analyzing data and extracting features
   - Analyze the data, examine its correlations, and visualize them.
   - Choose the appropriate machine-learning algorithm (or set of algorithms) and split the data into training and validation subsets
   - Decide on a cross-validation method, where the dataset is split into different training and validation datasets and average the results over the rounds.
4. Training the model
   - Train a model by running an algorithm that learns a set of algorithm-specific parameters from the input data
5. Evaluating the model
   - Validate performance, decide if feature space needs refinement or different algorithms need to be tuned.
6. Using the model
   - Deploy the built model to the production environment





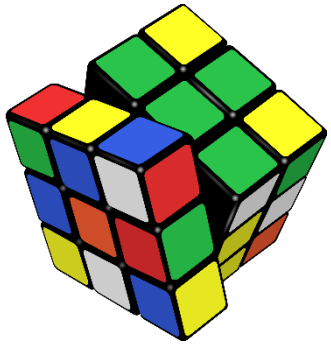
Summary

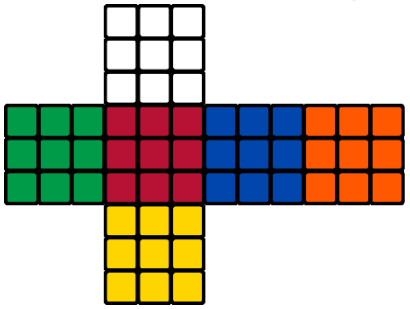
"The real question is, when will we draft an artificial intelligence bill of rights? What will that consist of? And who will get to decide that?"

Gray Scott

In Essence

- Apache Spark project is ideally positioned to help tackle these topics, which range from data ingestion and feature extraction/creation to model building and deployment.
- The three basic steps in feature engineering are feature extraction, feature transformation, and selection. Spark ML provides implementation of several algorithms to make these steps easier.
- Spark ML also provides several **classifications** (logistic regression, decision tree classifier, random forest classifier, and more), **regression** (liner regression, decision tree regression, random forest regression, survival regression, and gradient-boosted tree regression), **decision tree** and **tree ensembles** (random forest and gradient-boosted trees), as well as **clustering** (K-means and LDA) algorithms





References

"The key to artificial intelligence has always been the representation."

—Jeff Hawkins

References

- Machine Learning and Security, by David Freeman , Clarence Chio, O'Reilly Media, Inc, February 2018.
- Machine Learning, by Eihab Mohammed Bashier , Muhammad Badruddin Khan , Mohssen Mohammed, CRC Press, August 2016.
- Machine Learning, 2nd Edition by Stephen Marsland, Chapman and Hall/CRC, September 2015
- http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html
- https://spark.apache.org/docs/2.4.6/sql-programming-guide.html
- https://medium.com/@josemarcialportilla
- https://spark.apache.org/docs/2.4.6/ml-guide.html