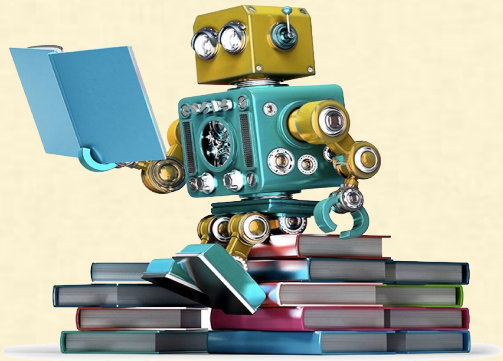
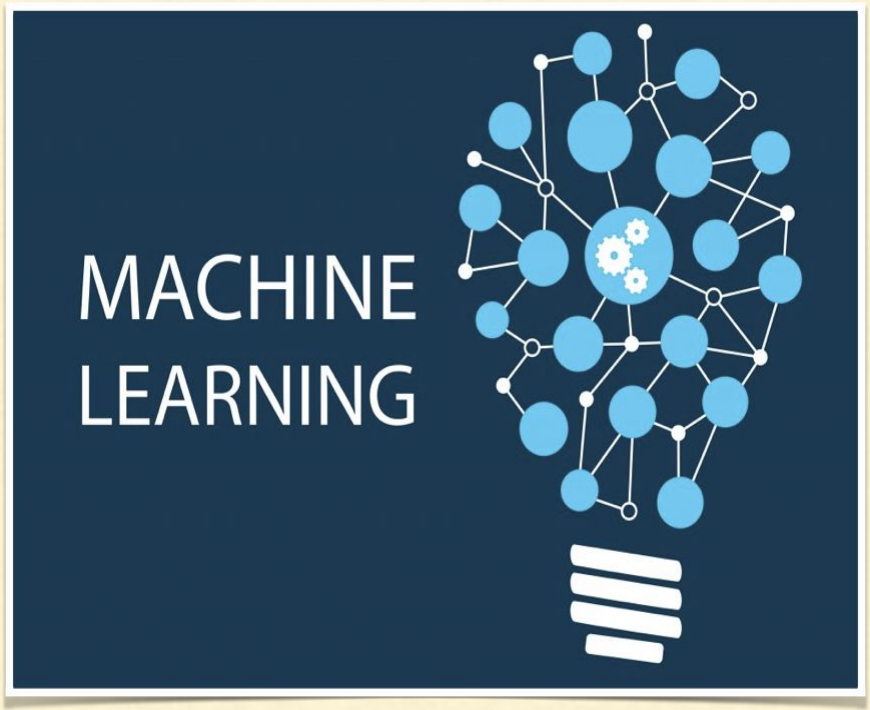
Machine Learning with MLlib

MACHINE LEARNING



“Programming Computers to optimize performance using Example Data or Past Experience”

MACHINE LEARNING?



Field of study that gives "computers the ability to learn without being explicitly programmed."

-- Arthur Samuel, 1959

HAVE YOU PLAYED MARIO?

How much time did it take you to learn & win the princess?

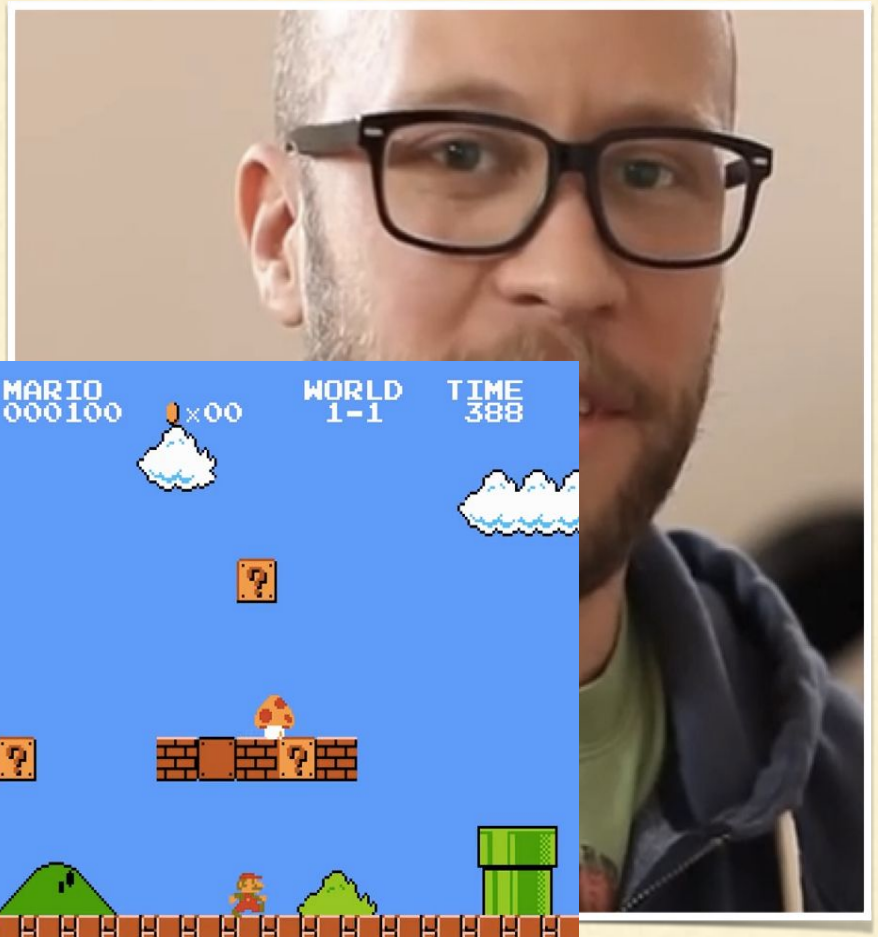
How about automating it?

How about automating it?

- Program Learns to Play Mario

Observes the game & presses keys

Maximises Score

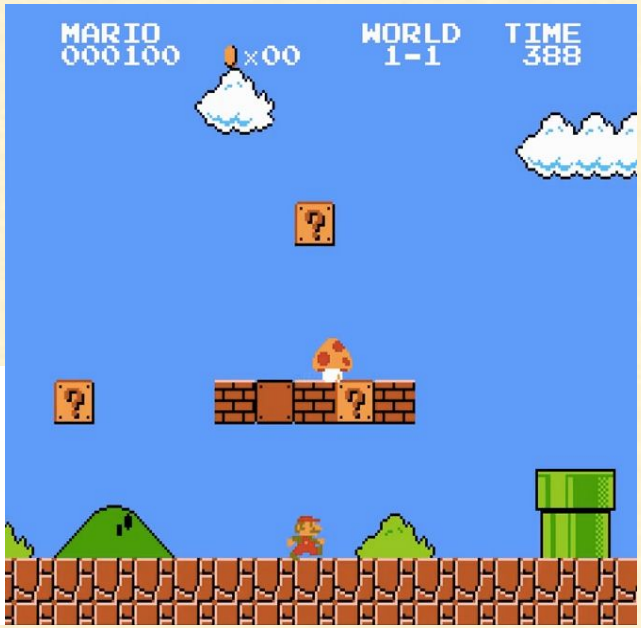


So?

- Program Learnt to play Mario

  and other games

Without any need of programming



Question:

To make this program learn any other games such as PacMan we will have to …

1. Write new rules as per the game
2. Just hook it to new game and let it play for a while

Question:

To make this program learn any other games such as PacMan we will have to …

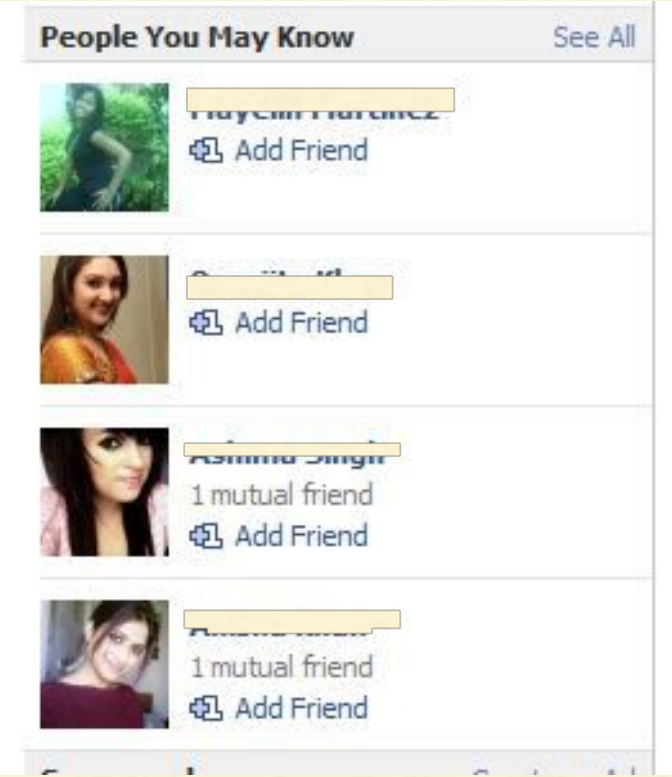
1. Write new rules as per the game
2. Just hook it to new game and let it play for a while

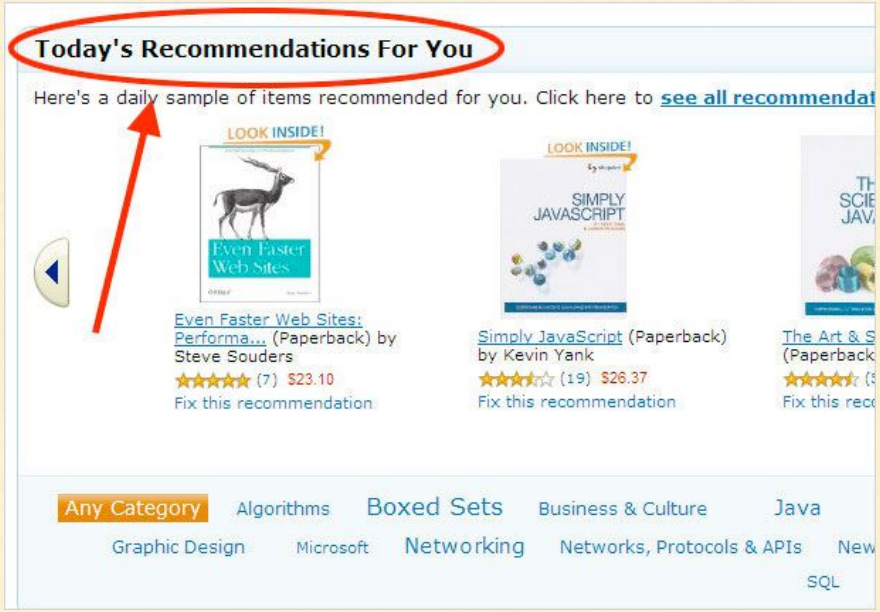
MACHINE LEARNING

- Branch of Artificial Intelligence
- Design and Development of Algorithms
- Computers Evolve Behaviour based on Empirical Data

MACHINE LEARNING - APPLICATIONS

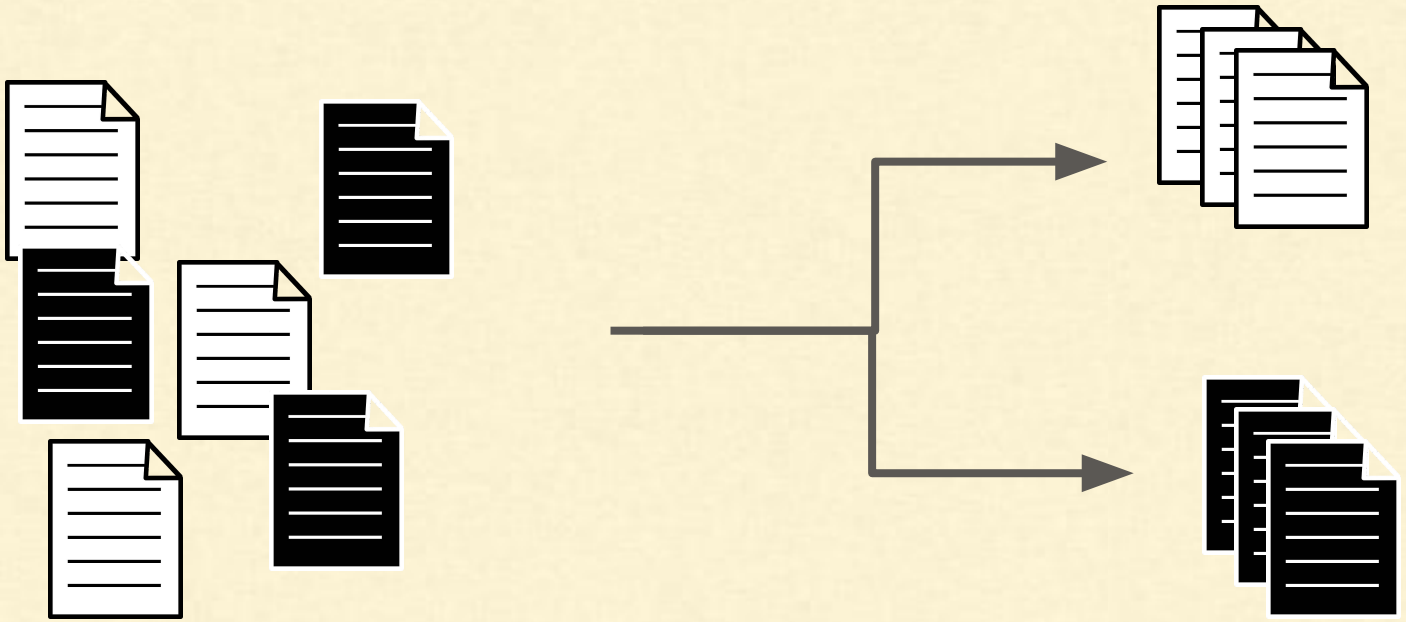
Recommend Friends, Dates, Products to end-user.





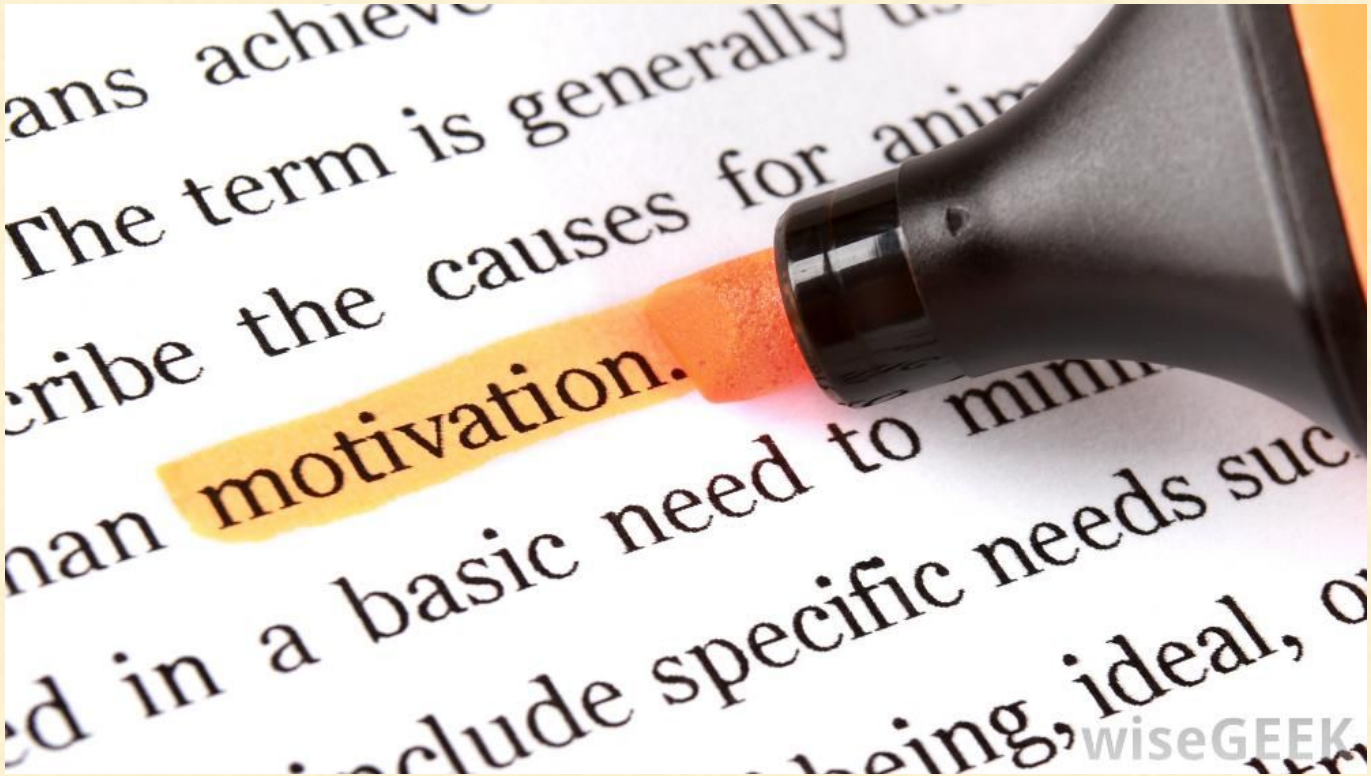
MACHINE LEARNING - APPLICATIONS

Classify content into predefined groups.



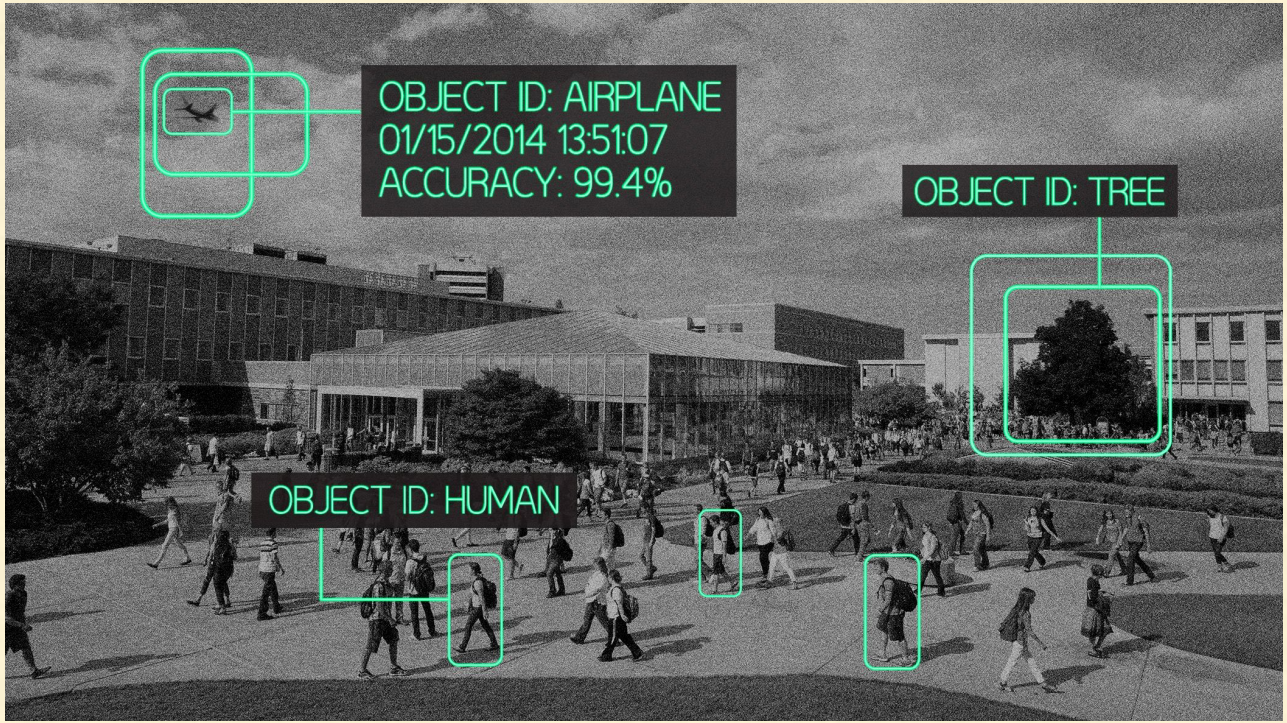
MACHINE LEARNING - APPLICATIONS

Identify key topics in large Collections of Text.



MACHINE LEARNING - APPLICATIONS

Computer Vision - Identifying Objects



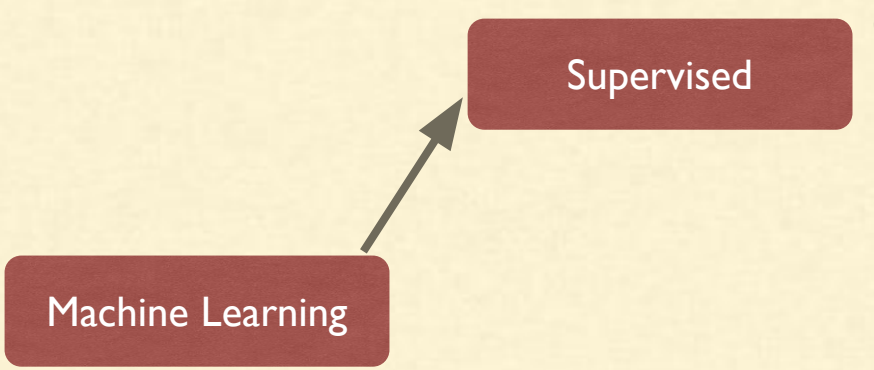
MACHINE LEARNING - APPLICATIONS

Natural Language Processing

MACHINE LEARNING - APPLICATIONS

- Find Similar content based on Object Properties.
- Detect Anomalies within given data.
- Ranking Search Results with User Feedback Learning.
- Classifying DNA sequences.
- Sentiment Analysis/ Opinion Mining
- BioInformatics.
- Speech and HandWriting Recognition.

MACHINE LEARNING - TYPES?



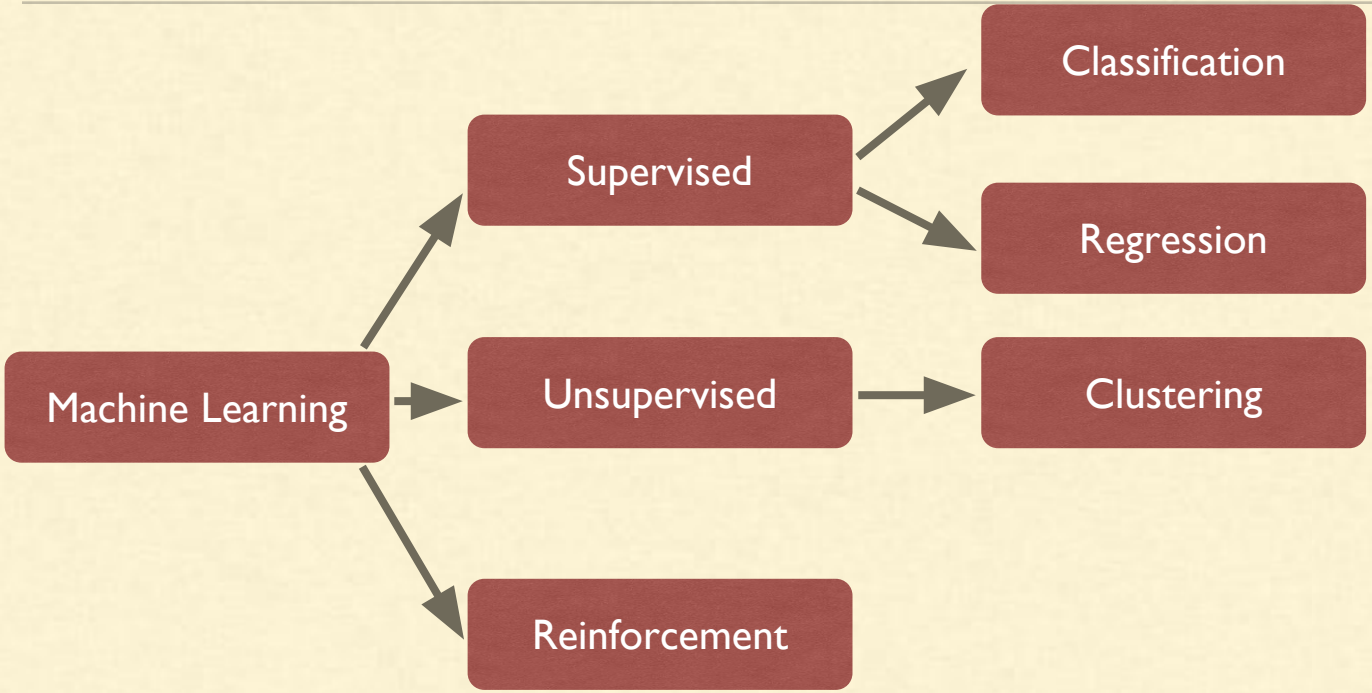
Given example inputs & outputs,

learn to map inputs to outputs

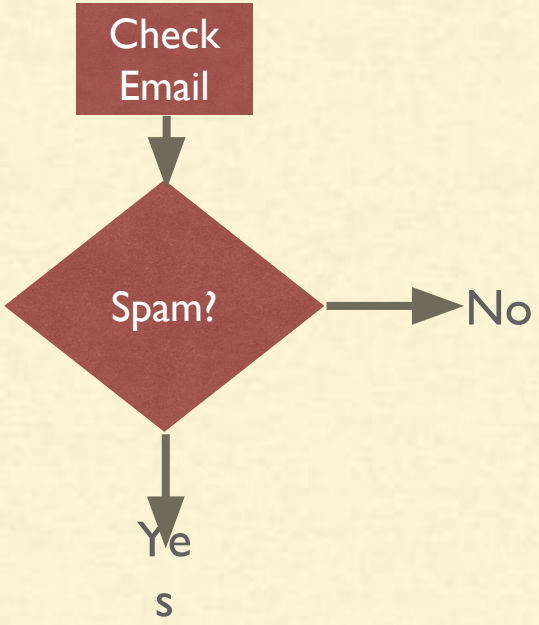
MACHINE LEARNING - TYPES?

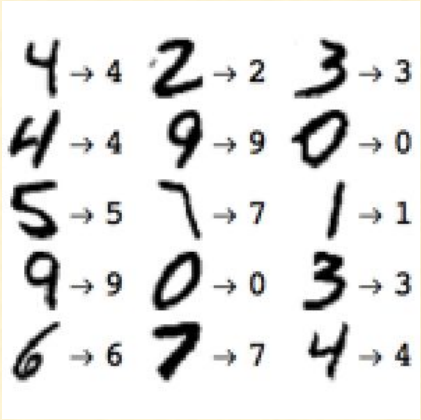
MACHINE LEARNING - TYPES?

MACHINE LEARNING - TYPES?



MACHINE LEARNING - CLASSIFICATION?





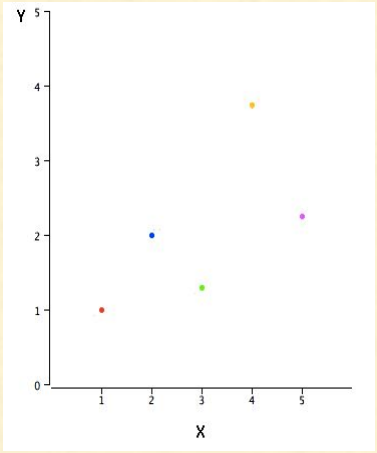
We Use Logistic Regression

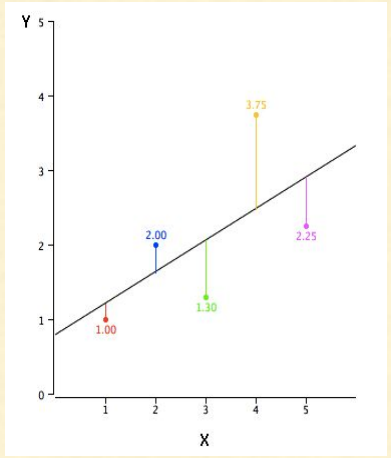
MACHINE LEARNING - REGRESSION?

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

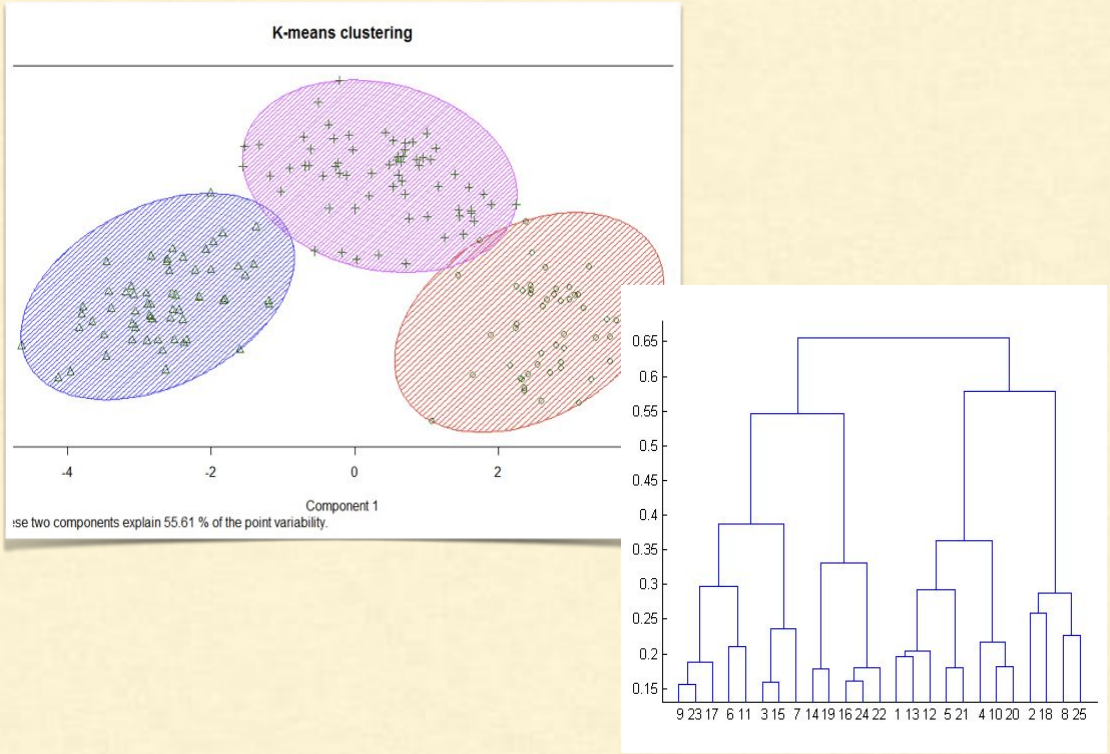
Predicting a continuous-valued attribute associated with an object.





In linear regression, we draw all possible lines going through the points such that it is closest to all.

MACHINE LEARNING - CLUSTERING?



- To form a cluster
- based on some definition of nearness

MACHINE LEARNING - TOOLS

DATA SIZE	CLASSFICATION	TOOLS
Lines Sample Data	Analysis and Visualization	Whiteboard,…
KBs - low MBs Prototype Data	Analysis and Visualization	Matlab, Octave, R, Processing,
MBs - low GBs Online Data	Analysis	NumPy, SciPy, Weka,
	Visualization	Flare, AmCharts, Raphael, Protovis
GBs - TBs - PBs Big Data	Analysis	MLlib, SparkR, GraphX, Mahout, Giraph

Machine Learning Library (MLlib)

Goal is to make practical machine learning scalable and easy

Consists of common learning algorithms and utilities, including:

- Classification
- Regression
- Clustering
- Collaborative filtering
- Dimensionality reduction
- Lower-level optimization primitives
- Higher-level pipeline APIs

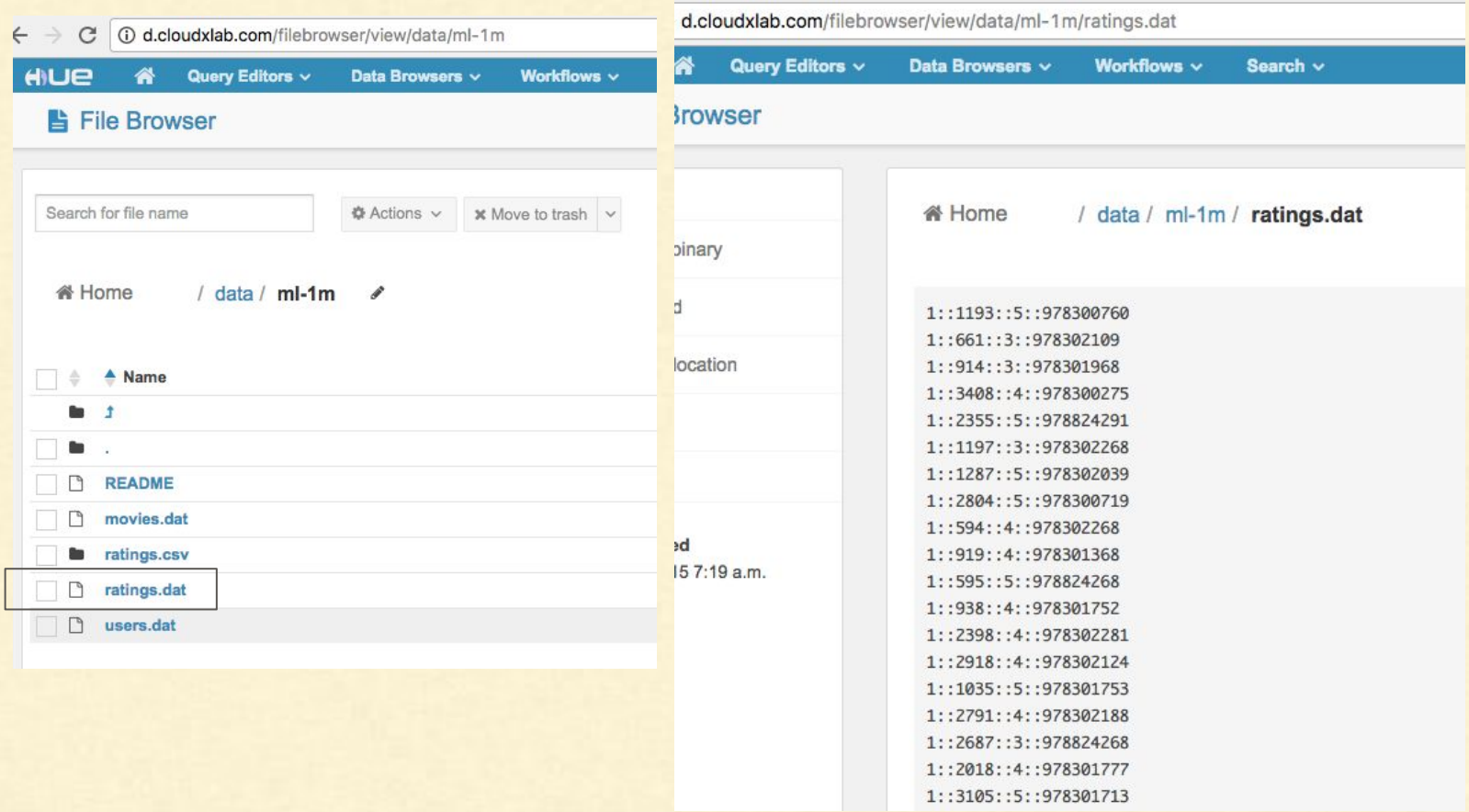
MLlib Structure

ML Algorithms Common learning algorithms e.g. classification, regression, clustering, and collaborative filtering 	**Featurization**  Feature extraction, Transformation, Dimensionality reduction, and Selection
Pipelines  Tools for constructing, evaluating, and tuning ML Pipelines	**Persistence**  Saving and load algorithms, models, and Pipelines
Utilities  Linear algebra, statistics, data handling, etc.	

MLlib - Collaborative Filtering

- Commonly used for recommender systems
- Techniques aim to fill in the missing entries of a user-item association matrix
- Supports model-based collaborative filtering,
- Users and products are described by a small set of latent factors
  - that can be used to predict missing entries.
- MLlib uses the alternating least squares (ALS) algorithm to learn these latent factors.

Example - Movie Lens Recommendation (1)



Example - Movie Lens Recommendation

Demo

https://github.com/cloudxlab/bigdata/blob/master/spark/examples/mllib/ml-recommender.scala

Exercise - Movies suggestions for you!

1. Find the maximum user id
2. Create the next user id denoting yourselves
3. Put your ratings of various movies
4. Generate your movies recommendations
5. Write down the steps in your Google Doc and share with support@cloudxlab.com.

spark.mllib - DataTypes

Local vector

integer-typed and 0-based indices and double-typed values

dv2 = [1.0, 0.0, 3.0]

Labeled point

a local vector, either dense or sparse, associated with a label/response

pos = LabeledPoint(1.0, [1.0, 0.0, 3.0])

Matrices:

Local matrix

Distributed matrix

RowMatrix

IndexedRowMatrix

CoordinateMatrix

BlockMatrix

Pipe Lines

DataFrame: This ML API uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types. E.g., a DataFrame could have different columns storing text, feature vectors, true labels, and predictions.

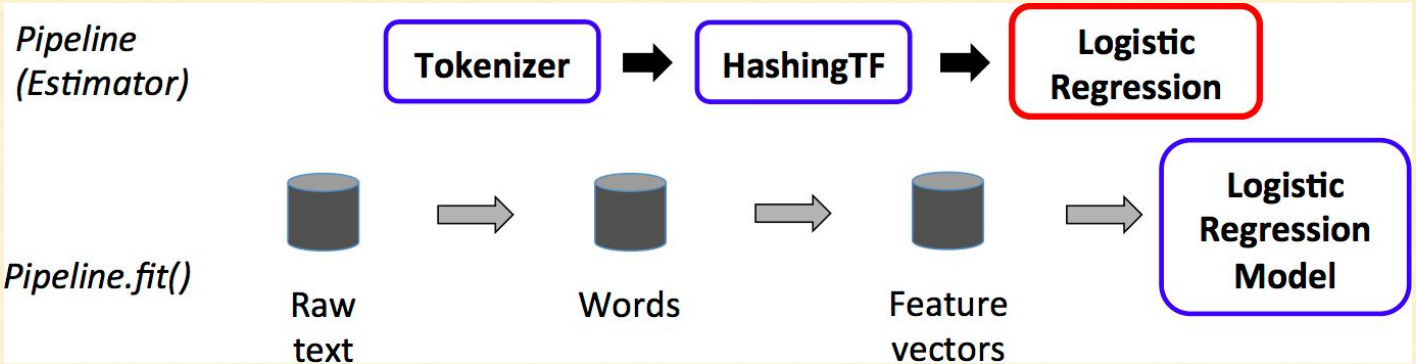
Transformer: A Transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model is a Transformer which transforms a DataFrame with features into a DataFrame with predictions.

Estimator: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model.

Pipeline: A Pipeline chains multiple Transformers and Estimators together to specify an ML workflow.

Parameter: All Transformers and Estimators now share a common API for specifying parameters.

Pipe Lines


Pipeline
(Estimator)
Tokenizer
HashingTF
Logistic Regression
Pipeline.fit()
Raw text
Words
Feature vectors
Logistic Regression Model

spark.mllib - Basic Statistics

Summary statistics
- Correlations
- Stratified sampling
- Hypothesis testing
- Random data generation
- Kernel density estimation
- See https://spark.apache.org/docs/latest/mllib-statistics.html

MLlib - Classification and Regression

MLlib supports various methods:

Binary Classification

linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes

Multiclass Classification

logistic regression, decision trees, random forests, naive Bayes

Regression

linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression

More Details>>

MLlib - Other Classes of Algorithms

Dimensionality reduction:

https://spark.apache.org/docs/latest/mllib-dimensionality-reduction.html

Feature extraction and transformation:

https://spark.apache.org/docs/latest/mllib-feature-extraction.html

Frequent pattern mining:

https://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html

Evaluation metrics:

https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html

PMML model export:

https://spark.apache.org/docs/latest/mllib-pmml-model-export.html

Optimization (developer):

https://spark.apache.org/docs/latest/mllib-optimization.html

MLLib

Thank you!

reachus@cloudxlab.com