Jumpstart on Machine Learning with Apache Spark

Jules S. Damji
Merrimack College, Feb 21, 2018

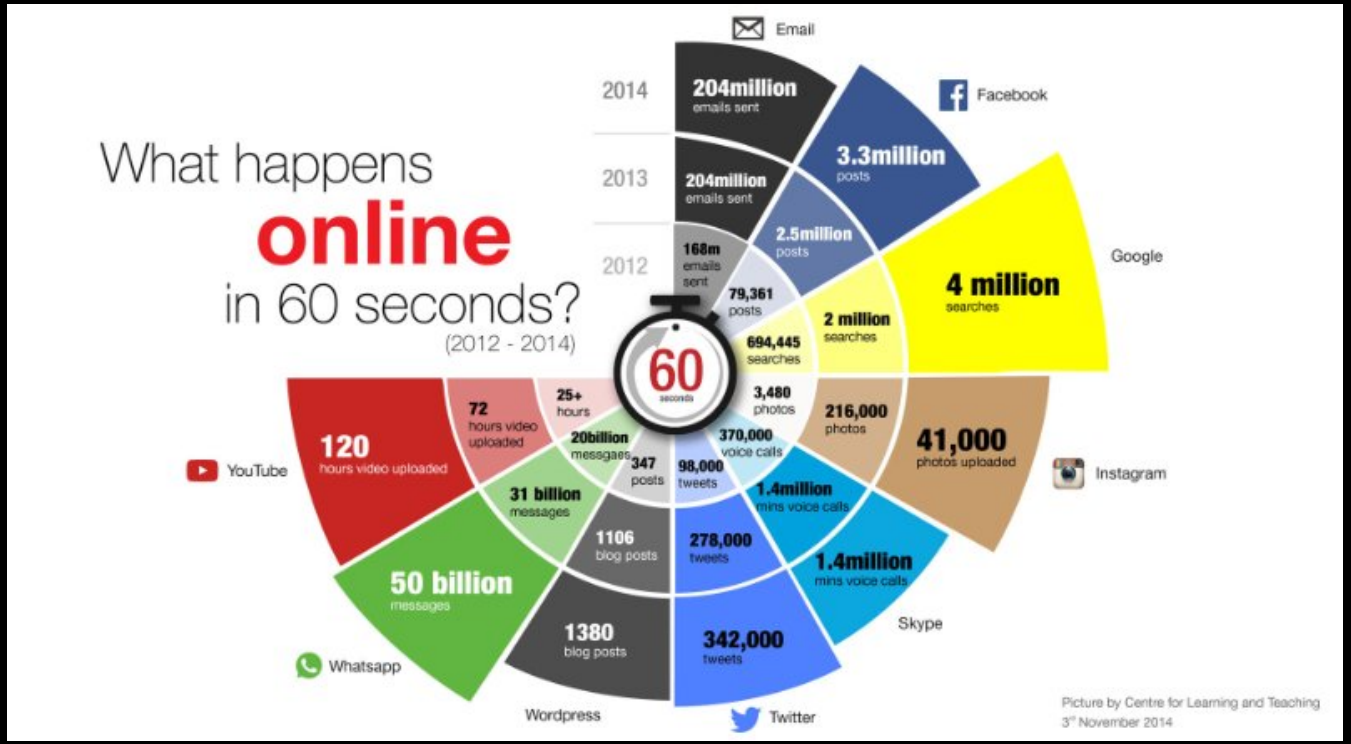
databricks

What's the Problem?

Agenda for Today's & Tomorrow's Problem

- What's Big Data Done To Us
- What and Why Apache Spark
- Survey of Popular Open Source ML Libraries
  - TensorFlow
  - Scikit-learn & Spark-sklearn
  - Spark MLlib
- Supervised Learning:
  - Logistic Regression using Spark MLlib in Databricks Notebook
- Q & A

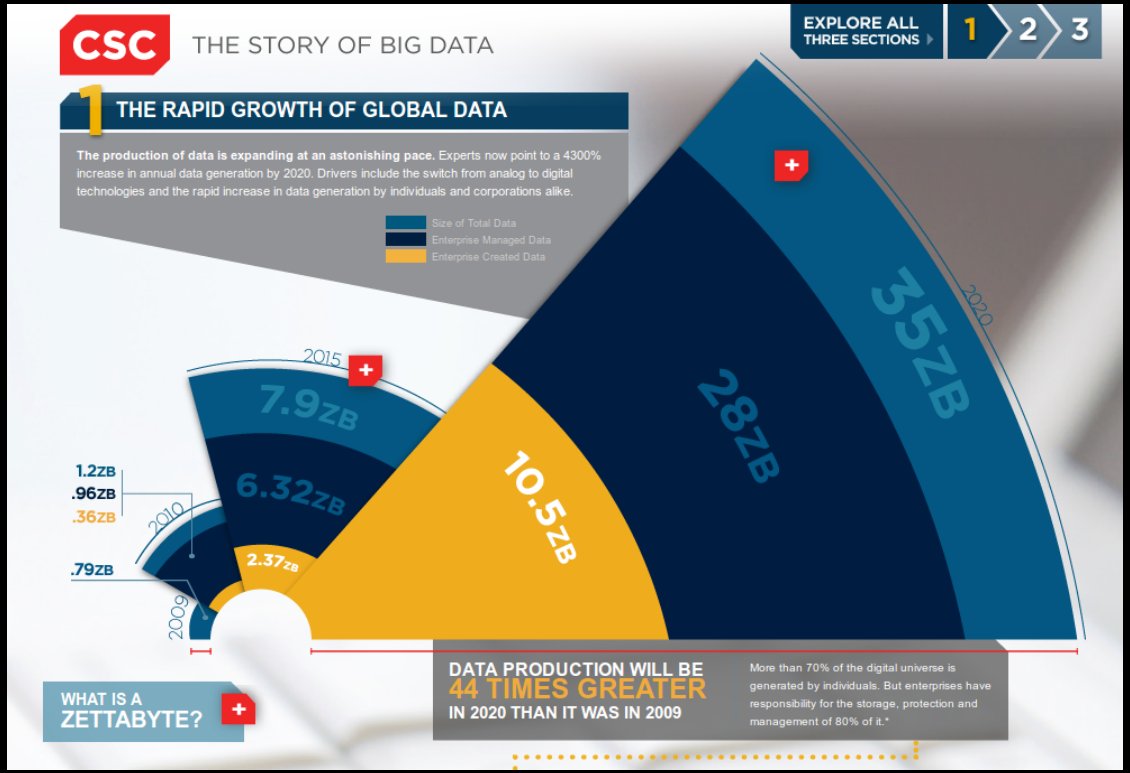
What's Big Data Done to Us



Today's Madness in a Minute

databricks

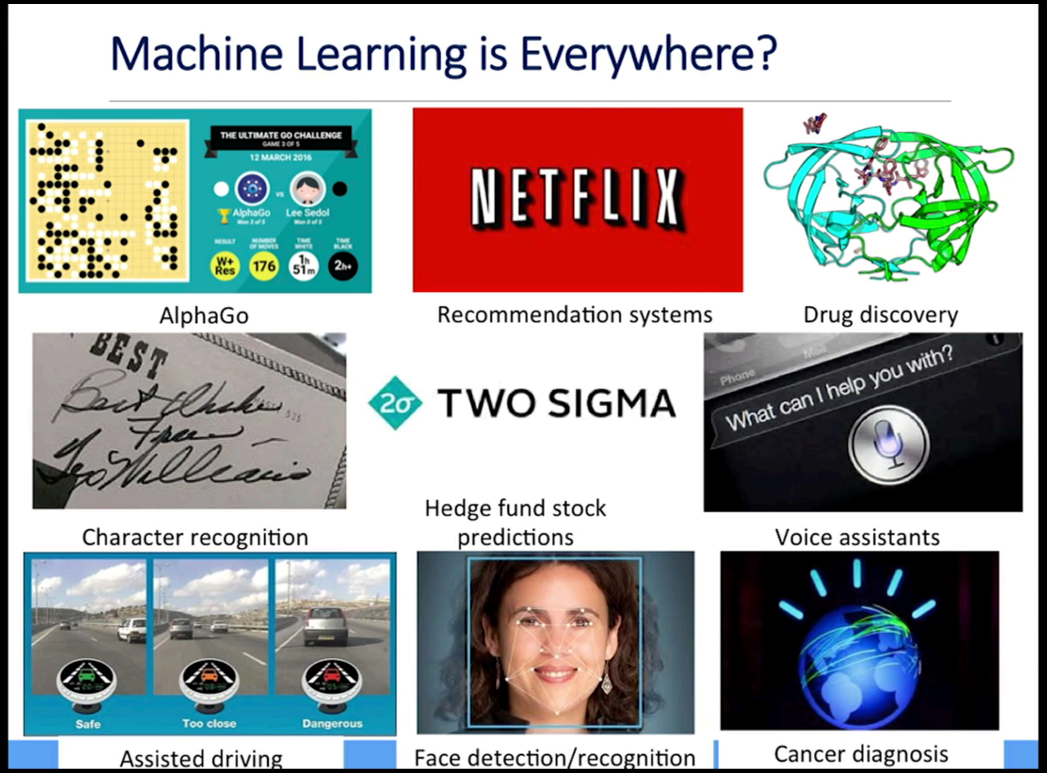
What's Big Data Done to Us



Tomorrow's Madness Magnified

databricks

What's Big Data Done to Us



Permeated our lives

Source : MIT

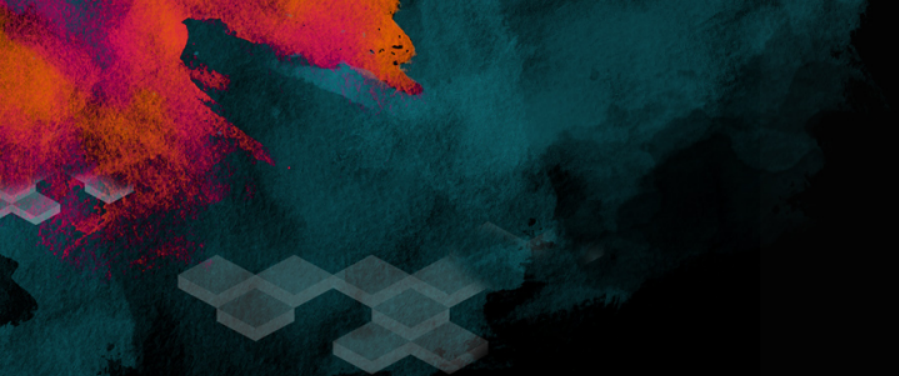
databricks

What's Big Data Done to Us

- High School Dropouts
- Weather
- Cyber Attacks
- Health and disease
- Police Misconduct

Predicting Aspects of our Lives

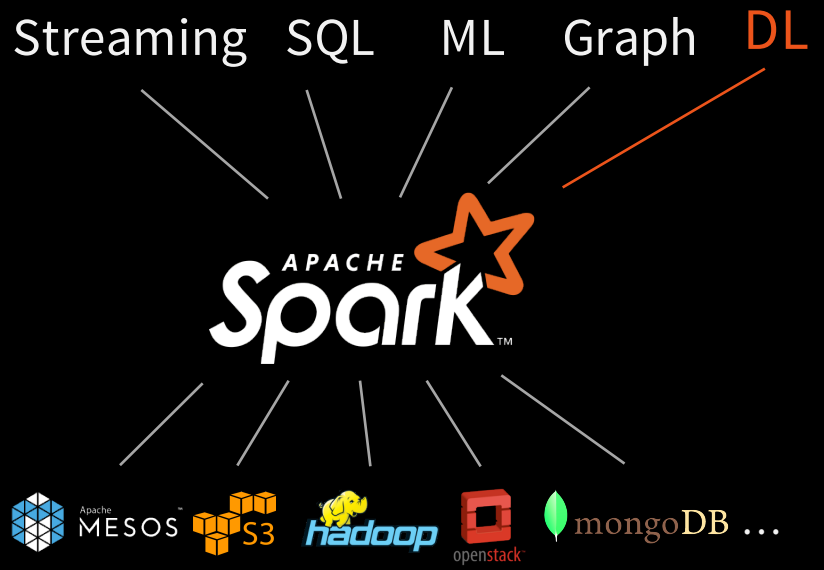
What's Apache Spark & Why



databricks

What is Apache Spark?

- General cluster computing engine that extends MapReduce
- Rich set of APIs and libraries
- Unified Engine
- Large community: 1000+ orgs, clusters up to 8000 nodes



Unique Thing about Spark

- Unification: same engine and same API for diverse use cases
  - Streaming, batch, or interactive
  - ETL, SQL, machine learning, or graph

Why Unification?

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with

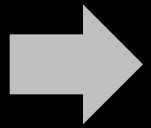
Why Unification?

- MapReduce: a general engine for batch processing

> We wrote the first version of the MapReduce library in February of 2003, and made significant enhancements to it in August of 2003, including the locality optimization, dynamic load balancing of task execution across worker machines, etc. Since that time, we have been pleasantly surprised at how broadly applicable the MapReduce library has been for the kinds of problems we work on. It has been used across a wide range of domains within Google, including:

Big Data Systems Today

MapReduce



Pregel Giraph

Dremel Millwheel

Storm Impala

Drill S4 . . .

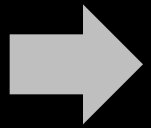
General batch processing

Specialized systems for new workloads

Hard to *combine* in pipelines

Big Data Systems Today

MapReduce

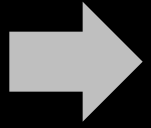


Pregel Giraph

Dremel Millwheel

Storm Impala

Drill S4 . . .



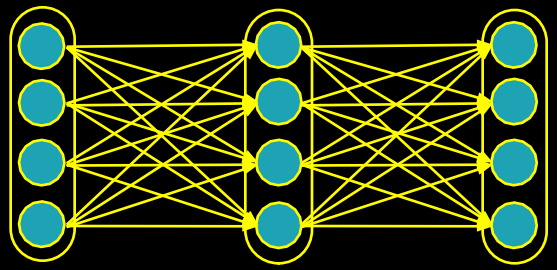
General batch processing

Specialized systems for new workloads

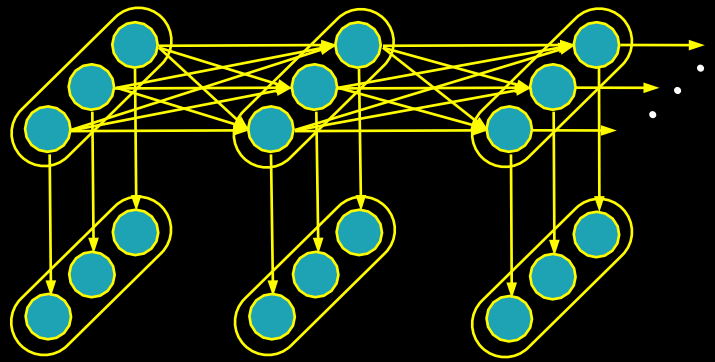
Unified engine

Key Idea

- MapReduce + data sharing (RDDs) captures most distributed apps



Iterative



Streaming

Benefits of Unification

1. Simpler to use and operate

2. Code reuse: e.g. only write monitoring, FT, etc once

3. New apps that span processing types: e.g. interactive queries on a stream, online machine learning

An Analogy

Specialized devices

New applications

Unified device

Survey of Open Source Machine Learning Libraries

“Field of study that gives computers the ability to learn without being explicitly programmed”

Machine Learning: What and Why?

What: ML uses data to identify patterns and make decisions.

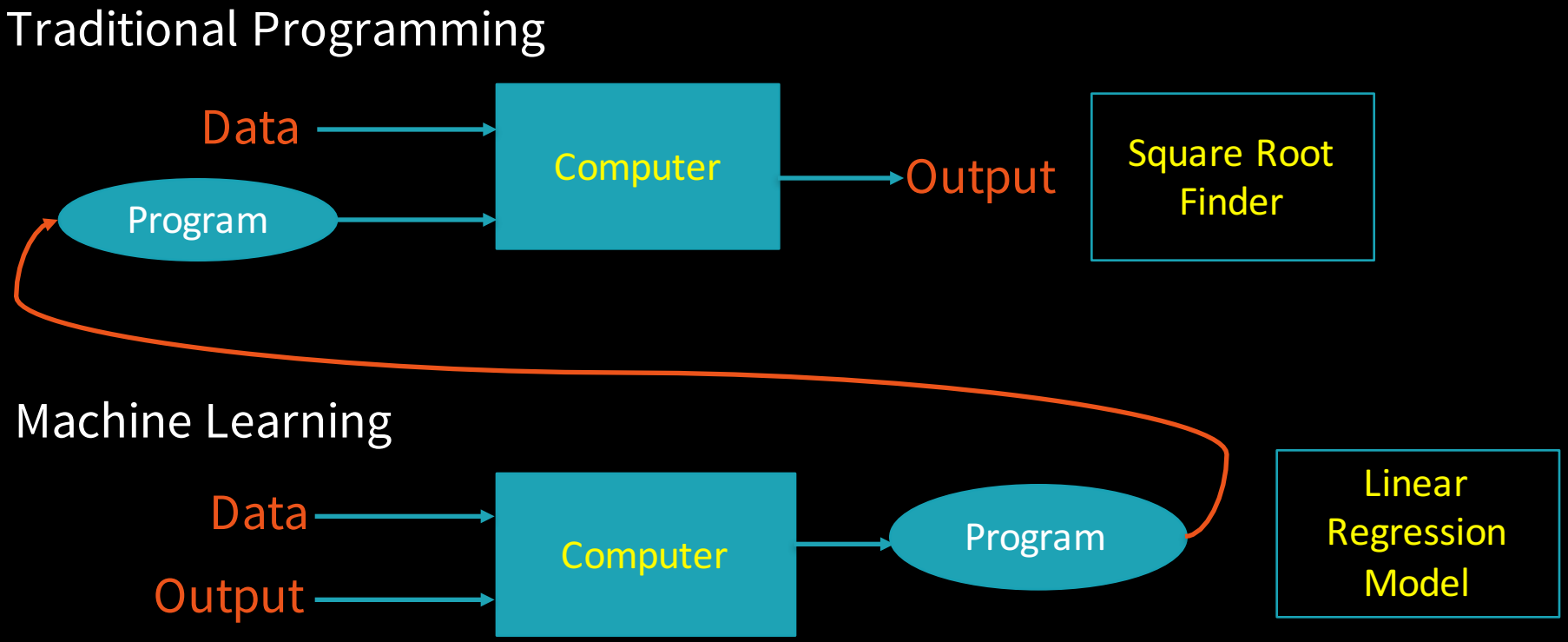
Why: The core value of ML is automated decision making.

- Especially important when dealing with TB or PB of data

Many use cases, including:

- Marketing and advertising optimization
- Security monitoring / fraud detection
- Operational optimizations

Traditional Programming vs Machine Learning?



Source : MIT https://ocw.mit.edu

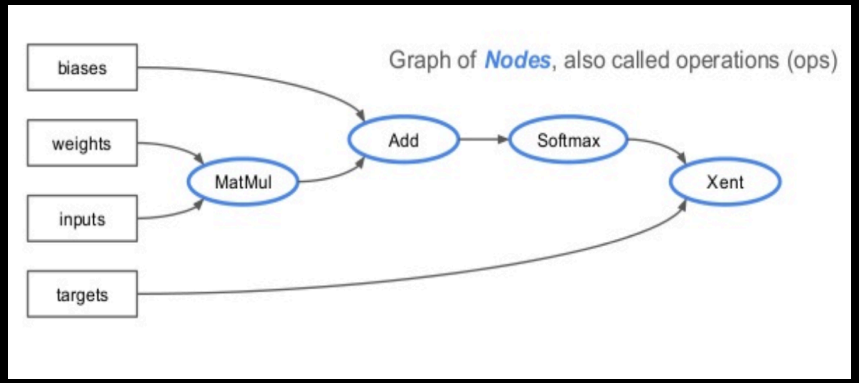
Survey of Open Source ML Libraries

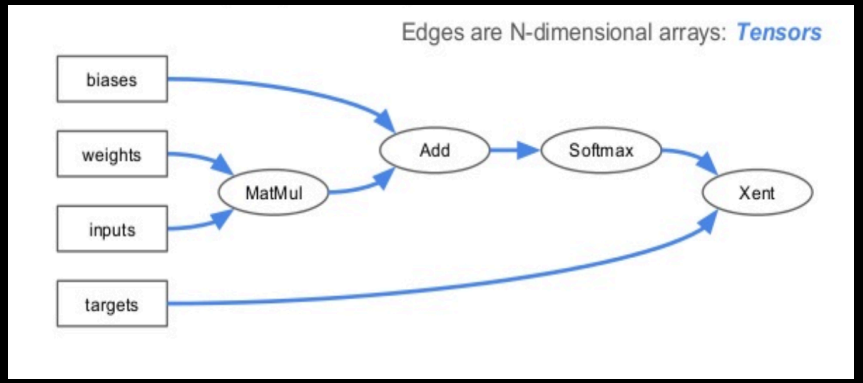
Tensorflow.org scikit-learn.org spark.apache.org

What's TensorFlow?

- Open source from Google, 2015
  - Current v1.5 APIs
- Created for numerical computations
- Mainly suitable: ML & DL Neural Network Applications
- Fast: Backend C/C++
- Data flow graphs
  - Nodes are functions
  - Edges are input or data

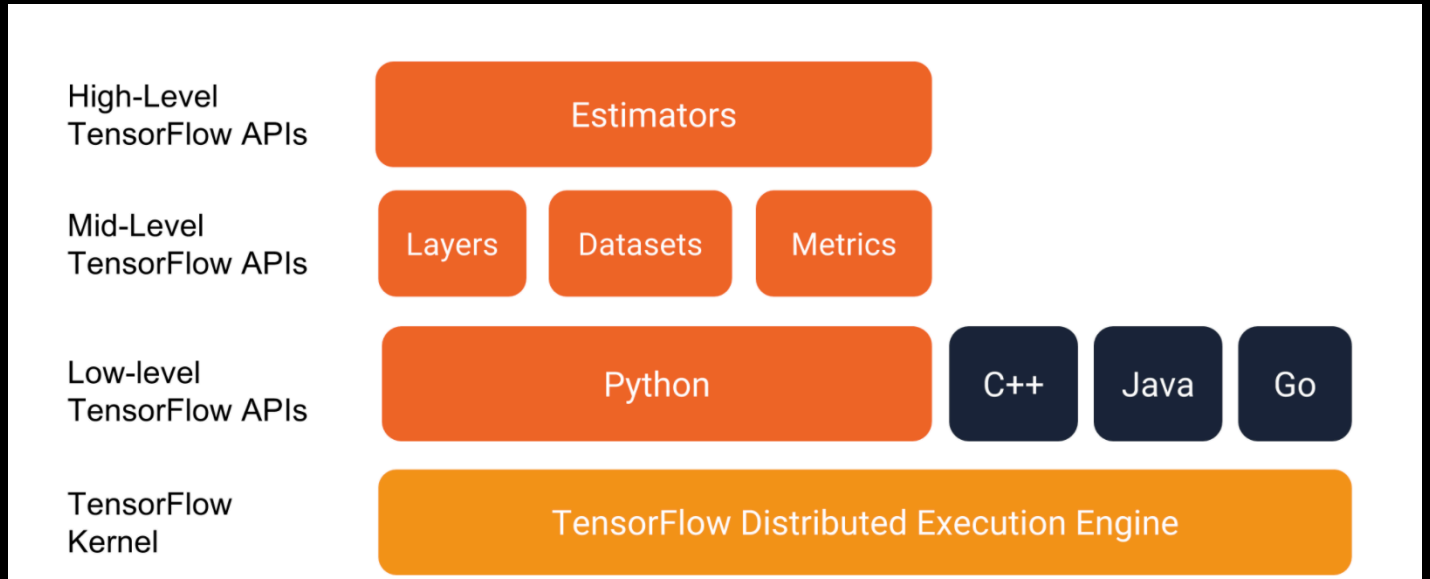
What's TensorFlow?





$W = \{ w, \ldots, wn \}$

TensorFlow Programming Stack



The TensorFlow Programming Environment

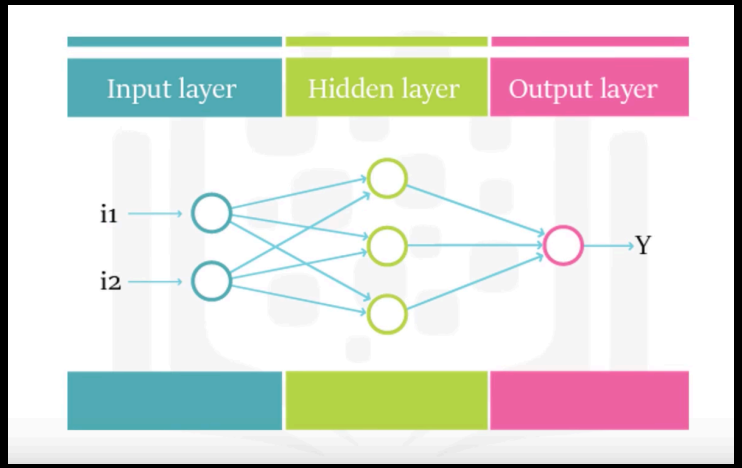
Why TensorFlow?

- Flexible Programming Stack: Python, C++, Go & Java
- Fast compile times
- Supports standard ML Estimators
  - Logistic Regression, DNNClassifier, etc.
- Supports CPUs, GPUs & Distributed processing on clusters
- Execution supports data flow as graphs
  - Building layers of nodes or neural networks of nodes.
  - Build a graph or layers of nodes and execute in TF session.
- Trained Models Deployed on servers & mobile devices

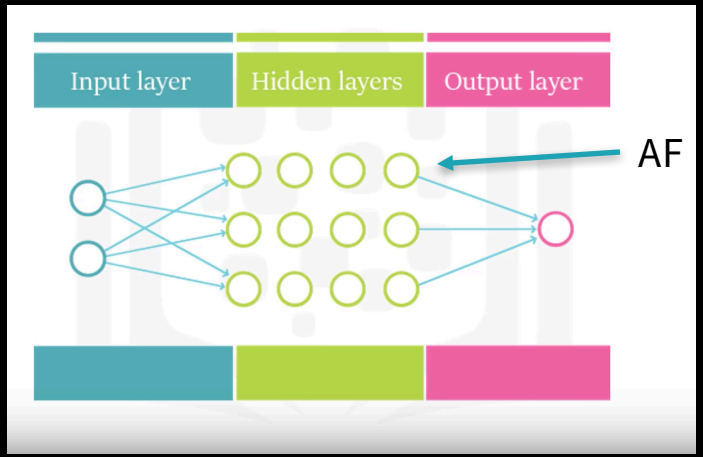
Why TensorFlow?

- Support for deep learning
- Assemble neural networks
- Mathematical functions library for neural networks & ML

Why TensorFlow



Shallow Layer Perceptron



Deep Layered Perceptron

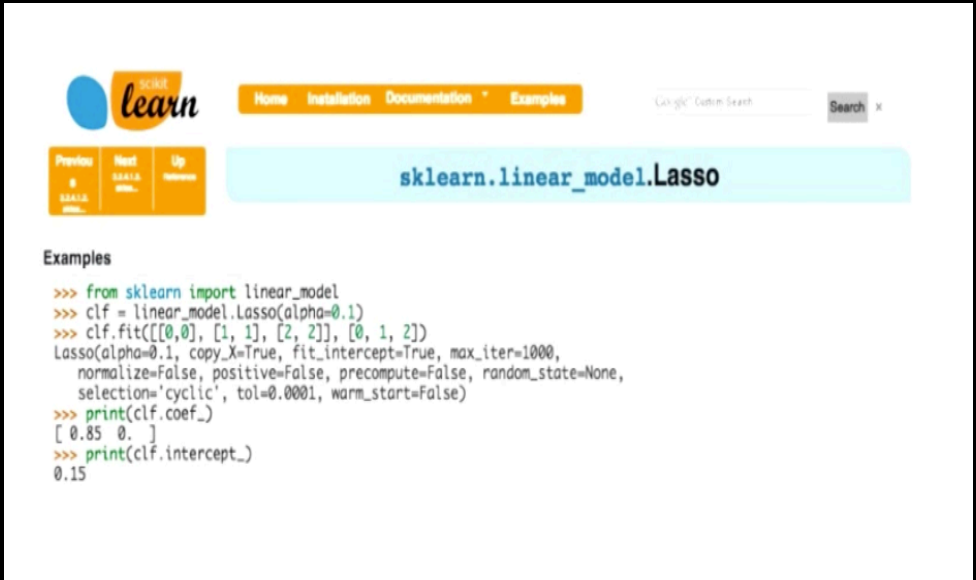
Simple Example: TF Iris Classifier

```
2  # Code example from tensorflow tutoiral in how to build image classifier
3  #
4  from sklearn import metrics
5  import tensorflow as tensorflow
6  import sys
7  from tensorflow.contrib import learn
8
9  def main(argv):
10     #load the iris dataset
11     iris = learn.dataset.load_dataset("iris")
12     (x_train, x_test, y_train, y_test) = cross_validation.train_split(
13         iris.data, iris.target, test_size=0.2, random_state=42)
14
15     #
16     # Build 3 layer DNN with 10, 20, 10 units respectively
17     classifier = learn.DNNClassifier(hidden_units=[10, 20, 10], n_classes = 3
18         )
19     #
20     # fit and predict the model
21     #
22     classifier.fix(x_train, y_train, steps=200)
23     score = metrics.accuracy_score(y_test, classifier.predict(x_test))
24     print ('Accuracy: {0:f}'.format(score))
25
26 if __name__ == '__main__':
27     main(sys.argv)
```

TensorFlow Resources

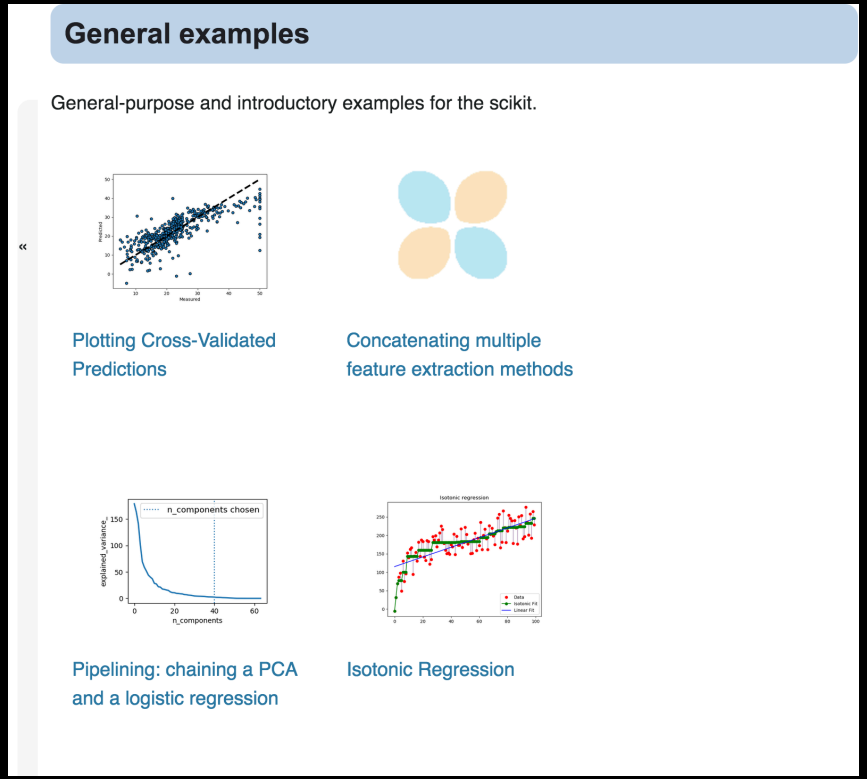
- Tensorflow.org
- TensorFlow for Poets
- Introduction to TensorFlow
- TensorFlow Tutorials

What's Scikit-learn?

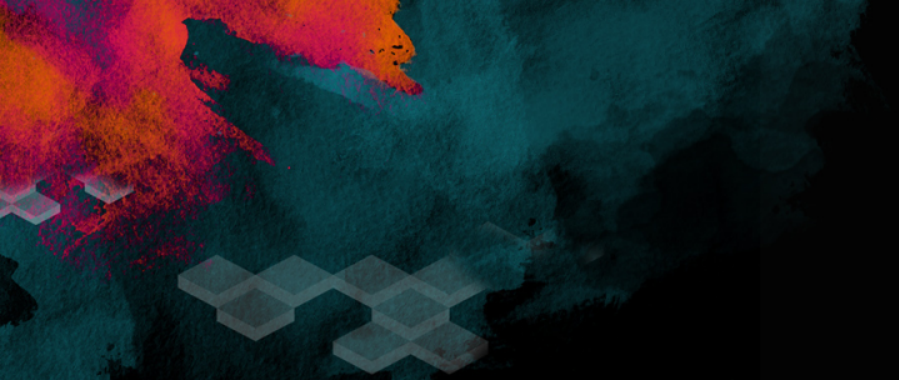


- Open source Python ML Library
- Algorithms: Supervised/Unsupervised
- Only runs on a single machine, not clusters

Why Use Scikit-learn



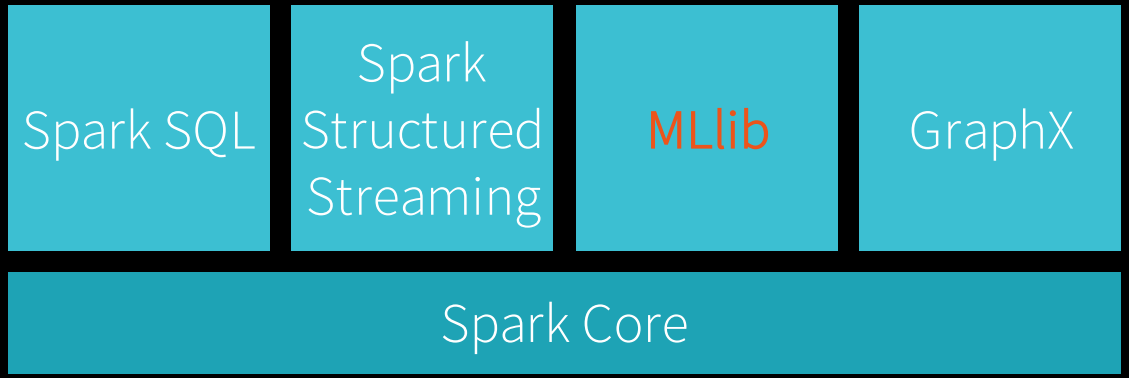
- Comprehensive Documentation
- Easy and Simple APIs
- Easy to learn ML concepts
- Large Community Base
- Built on NumPy, SciPy, and matplotlib
- Apache Spark MLlib Integration
  - spark_sklearn



Apache Spark MLlib

databricks

Apache Spark Components



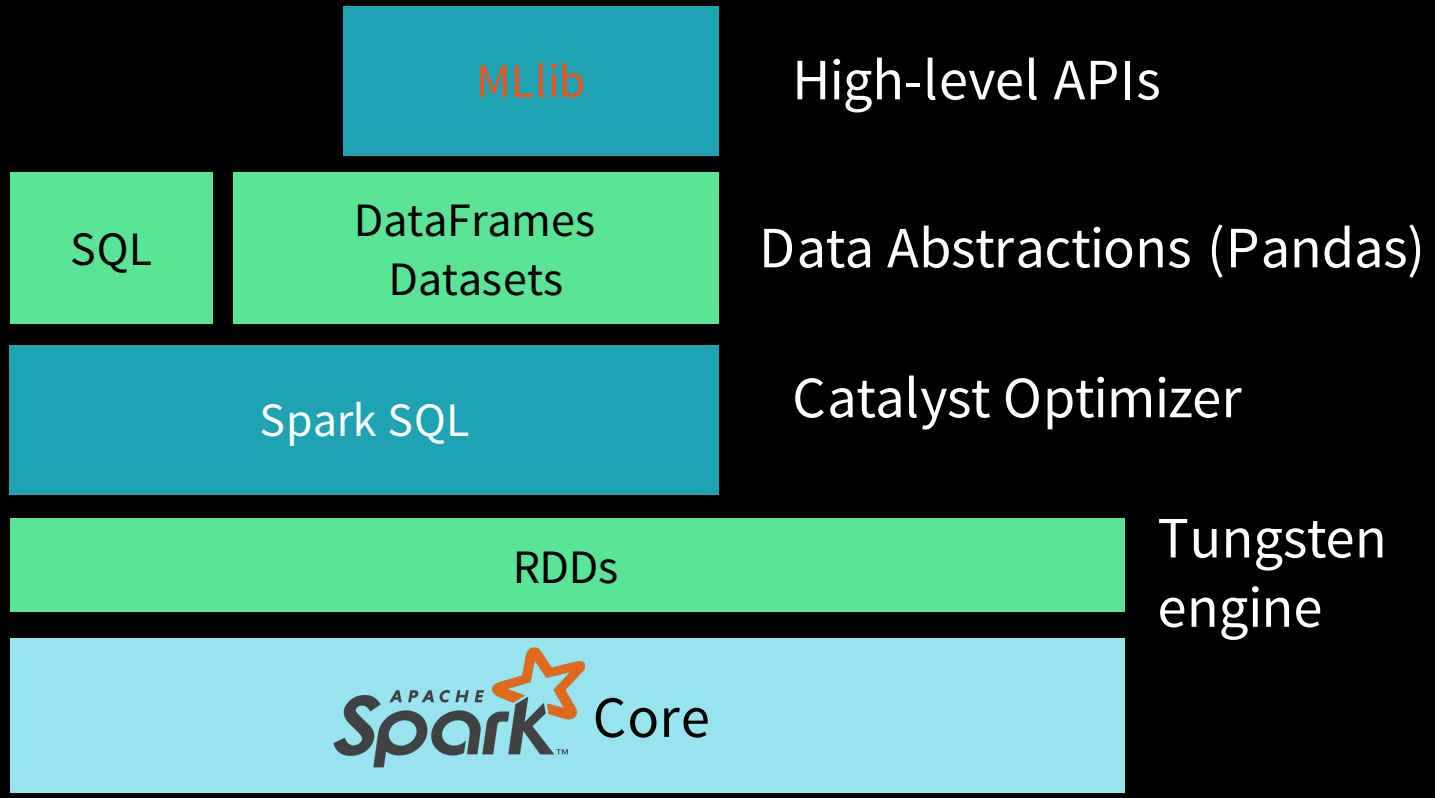
Scale out, fault tolerant

Python, Java, Scala, and R APIs

Standard libraries

Unified engine across diverse workloads & environments

Apache Spark MLlib Programming Stack



DataFrame-based API for MLlib

In 2.0, the DataFrame-based API became the primary MLlib API.

- Voted by community
- `org.apache.spark.ml, pyspark.ml`

The RDD-based API is in *maintenance mode*.

- Still maintained with bug fixes, but no new features
- `org.apache.spark.mllib, pyspark.mllib`

Why Spark MLlib

Provide general purpose ML algorithms on top of Spark

- Hide complexity of distributing data & queries, and scaling
- Familiar API based on scikit-learn
- Leverage Spark improvements (DataFrames, Tungsten, Datasets)

Advantages of MLlib's design:

- Simplicity
- Scalability
- Streamlined end-to-end
- Compatibility (other Frameworks)

Why Spark MLlib: High-level functionality

Learning Tasks

- Regression
- Classification
- Clustering

Data Utilities

- Feature extraction & selections
- Statistics
- Linear Algebra

Workflow

- Pipelines
- Model Import/Export
- Cross validation

Why Spark MLlib: Algorithm Coverage

Classification

- Logistic regression w/ elastic net
- Naive Bayes
- Streaming logistic regression
- Linear SVMs
- Decision trees
- Random forests
- Gradient-boosted trees
- Multilayer perceptron
- One-vs-rest

Regression

- Least squares w/ elastic net
- Isotonic regression
- Decision trees
- Random forests
- Gradient-boosted trees
- Streaming linear methods

Recommendation

- Alternating Least Squares (ALS)

Why Spark MLlib: Algorithm Coverage

Feature Extraction

- Binarizer
- Bucketizer
- Chi-Squared selection
- CountVectorizer
- Discrete cosine transform
- ElementwiseProduct
- Hashing term frequency
- Inverse document frequency
- MinMaxScaler
- Ngram
- Normalizer
- One-Hot Encoder
- PCA
- PolynomialExpansion
- RFormula
- SQLTransformer
- Standard scaler
- StopWordsRemover
- StringIndexer
- Tokenizer
- StringIndexer
- VectorAssembler
- VectorIndexer
- VectorSlicer
- Word2Vec

Clustering

- Gaussian mixture models
- K-Means
- Streaming K-Means
- Latent Dirichlet Allocation
- Power Iteration Clustering

Statistics

- Pearson correlation
- Spearman correlation
- Online summarization
- Chi-squared test
- Kernel density estimation

Why Spark MLlib: Algorithm Coverage

Linear Algebra

- Local dense & sparse vectors & matrices
- Distributed matrices
  - Block-partitioned matrix
  - Row matrix
  - Indexed row matrix
  - Coordinate matrix
- Matrix decompositions

Model Export

- APIs to save/load ML models

Pipelines

- API to string together Estimators & Transformers
- ML Workflow

Our Demo task

Sentiment analysis

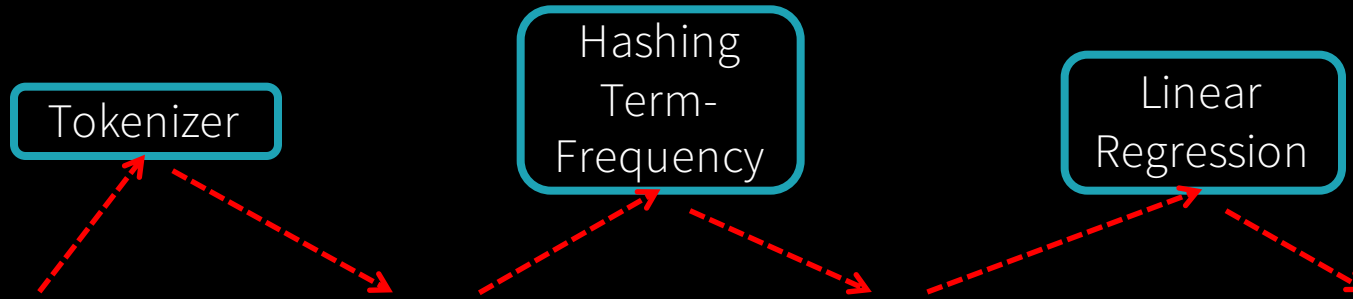
Given a review (text)

Predict the user's rat

rating	review
4	GOOD FOR THE PRICE, SO IT CANNOT BE MORE RUGGE...
5	It's made in China (like everything), but it s...
5	Considering the price, I wasn't expecting much...
5	This tool has the right bend to remove radiato...
2	I shoot weddings professionally and there are ...
4	Really like the swivel feature because I use i...
5	Love, Love, Love this brand. The pancake mix i...
2	They are not a easy to put on as the metal cli...
3	I wasn't sure this product would work on my 17...
5	Did some research on this bluetooth and decide...

https://snap.stanford.edu/data/web-Amazon.html

Our Machine Learning Workflow



Text
This scarf I bought is very strange. When I ...

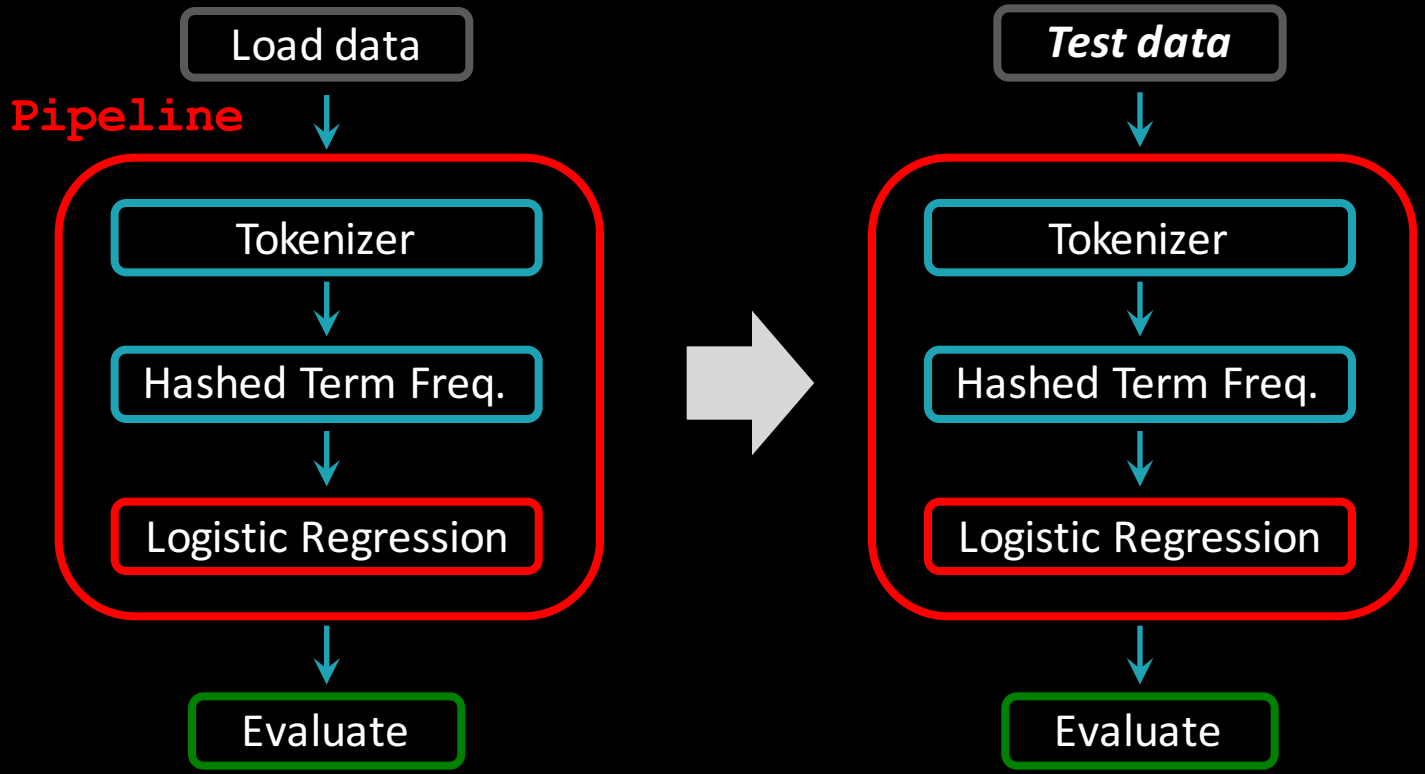
Label
Rating = 3.0

Words
[This, scarf, I, bought, ...]

Features
[2.0, 0.0, 3.0, ...]

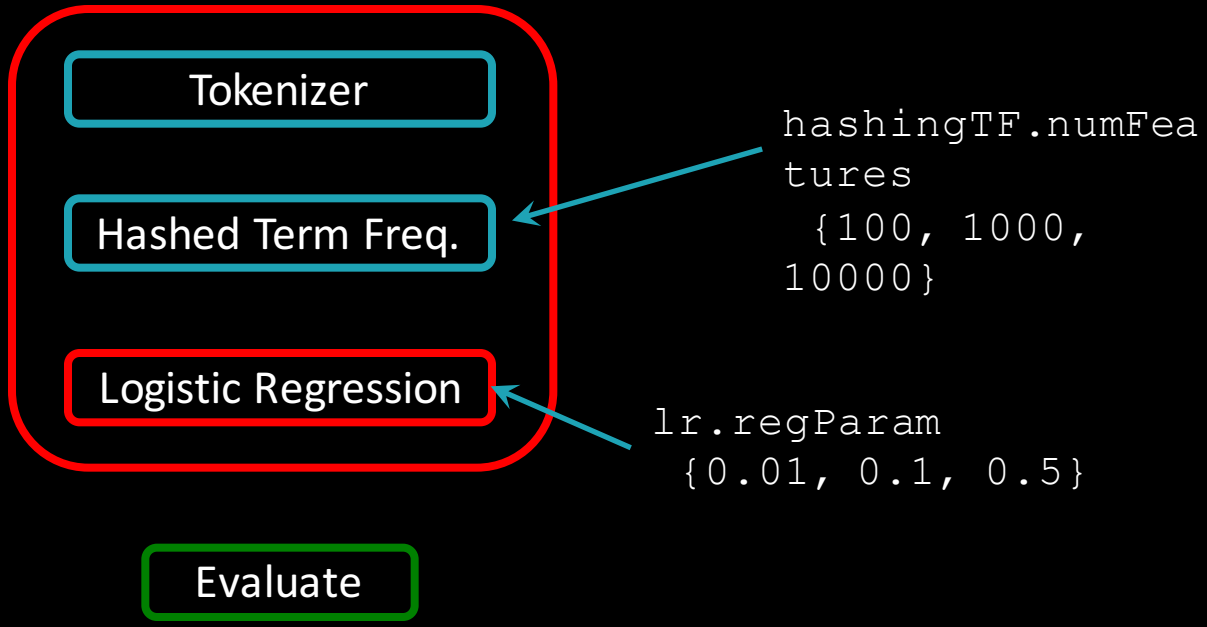
Prediction
Rating = 2.7

Our Machine Learning Workflow Pipelines



Re-run exactly the same way

Parameter Tuning



CrossValidator

Given:

- Estimator
- Parameter grid
- Evaluator

Find best parameters

Spark MLlib Demo

databricks

"All models are wrong, but some are useful"

databricks

Summary

- Data Deluge Eating the World
- Why Apache Spark Makes Sense
- Glimpse into OSS ML libraries
- Spark MLlib Merits

Resources

- Apache Spark MLlib User Guide
- docs.databricks.com
- Data Science Central
- KDnuggets

Thank You!

Apache Spark Makes Big Data Simple