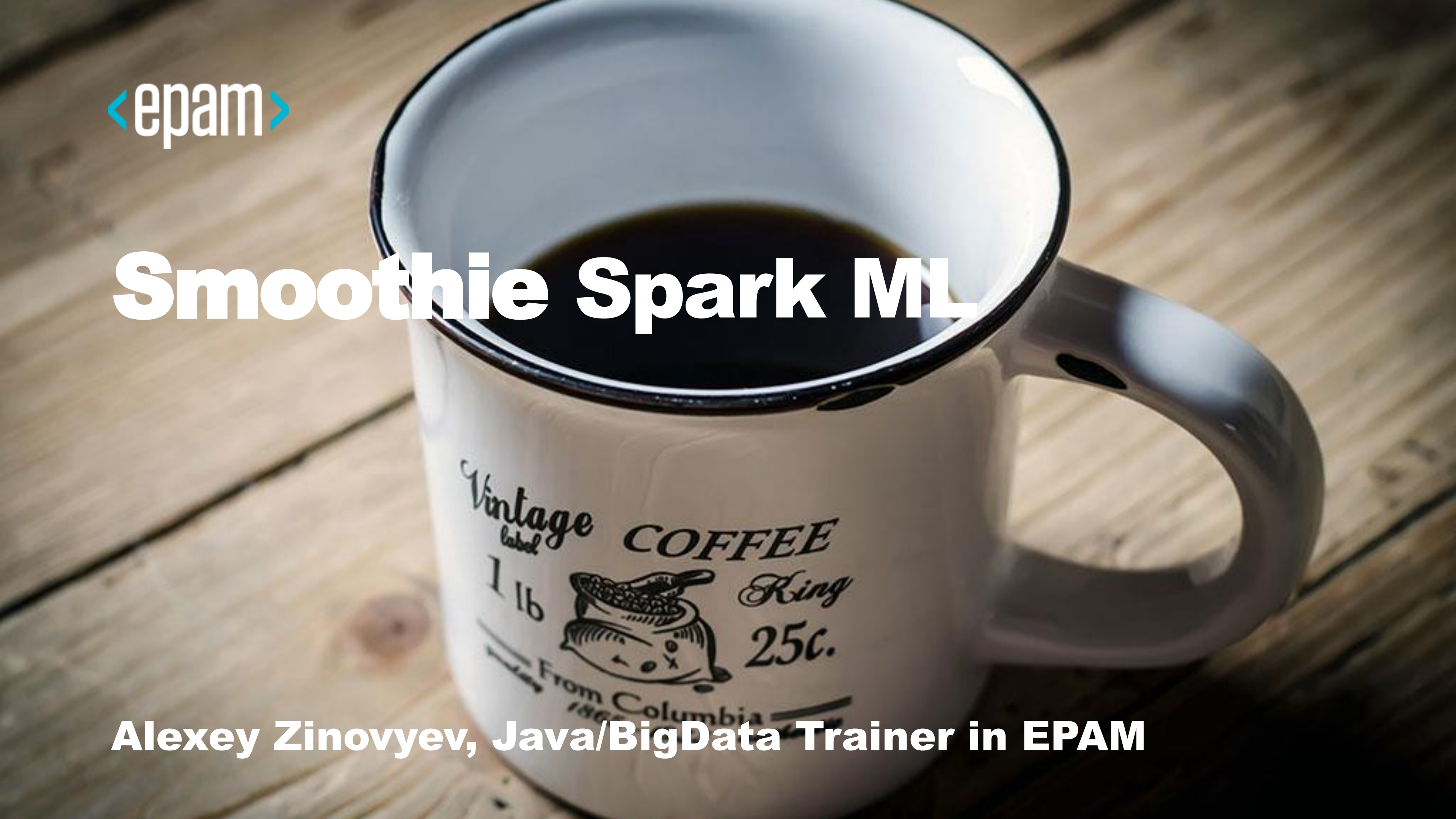


Smoothie Spark ML

Alexey Zinovyev, Java/BigData Trainer in EPAM

About

With IT since 2007

With Java since 2009

With Hadoop since 2012

With Spark since 2014

With EPAM since 2015

Contacts

E-mail : Alexey_Zinovyev@epam.com

Twitter : @zaleslaw @BigDataRussia

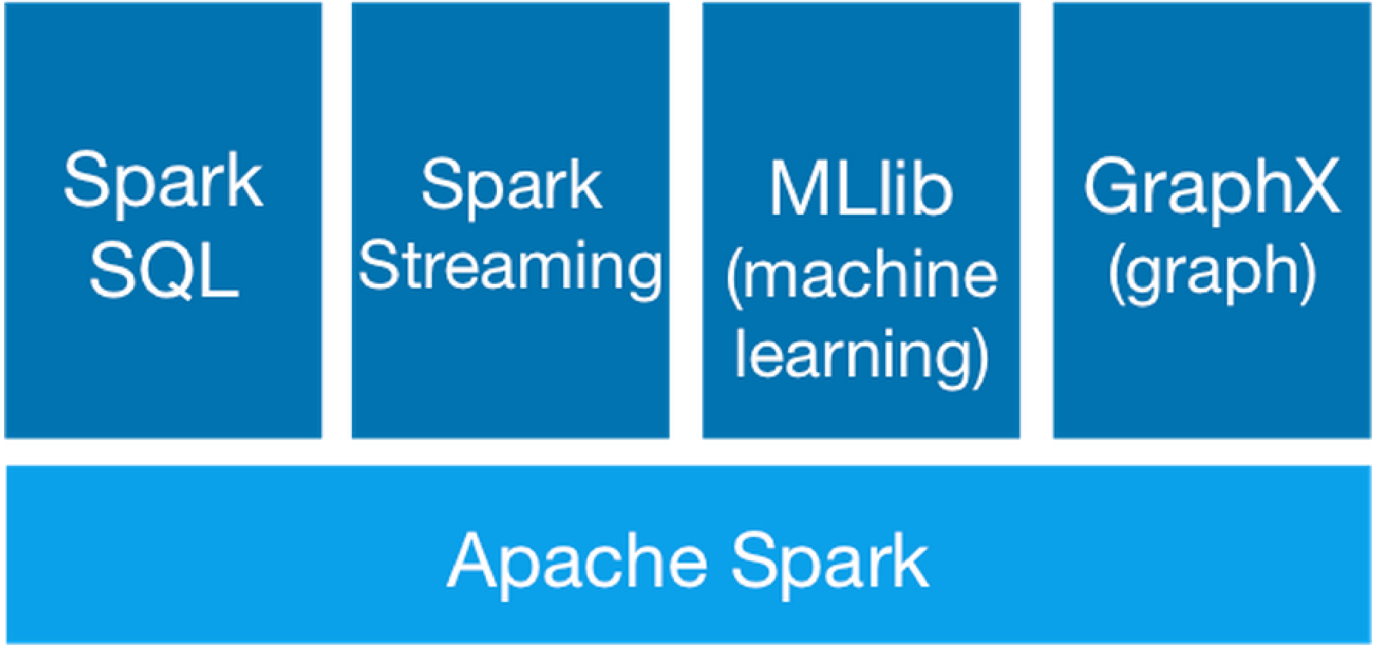
vk.com/big_data_russia **Big Data Russia**

+ Telegram @bigdatarussia

vk.com/java_jvm **Java & JVM langs**

+ Telegram @javajvmlangs

Spark Family



Spark SQL

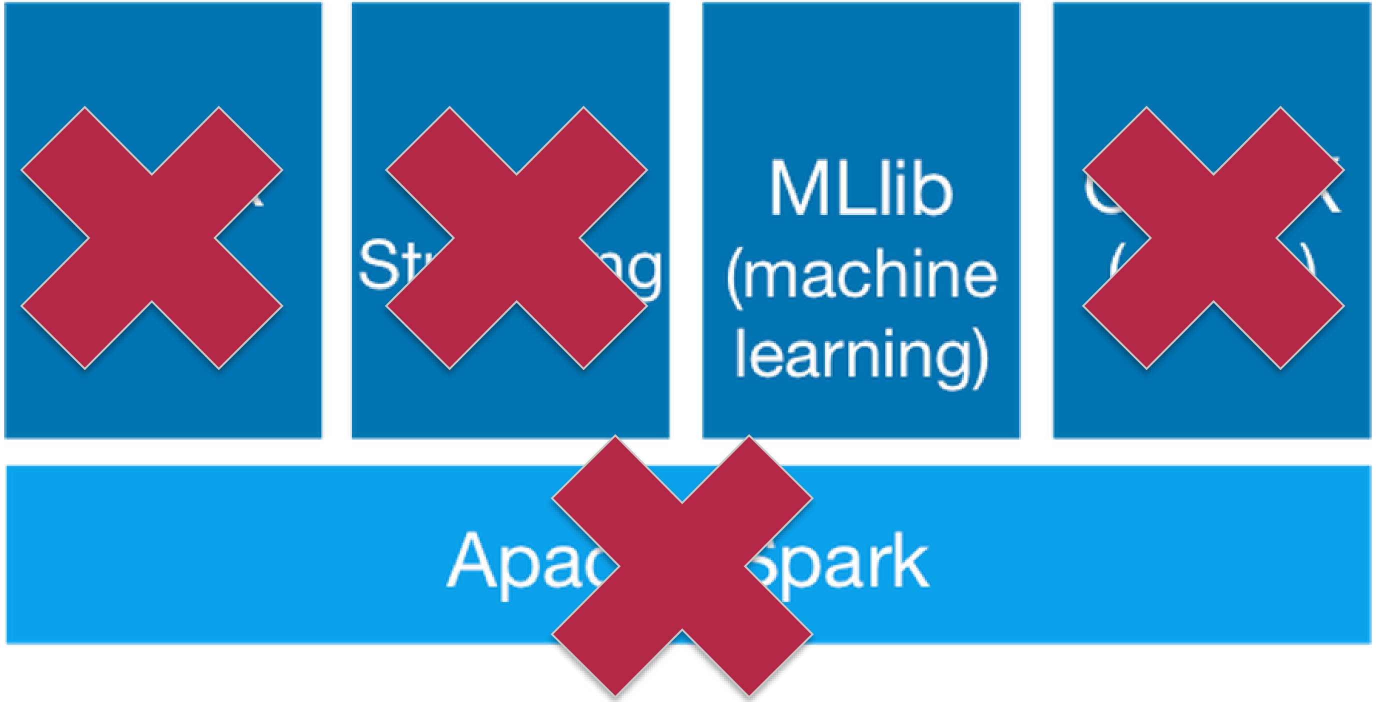
Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Spark Family



Pre-summary

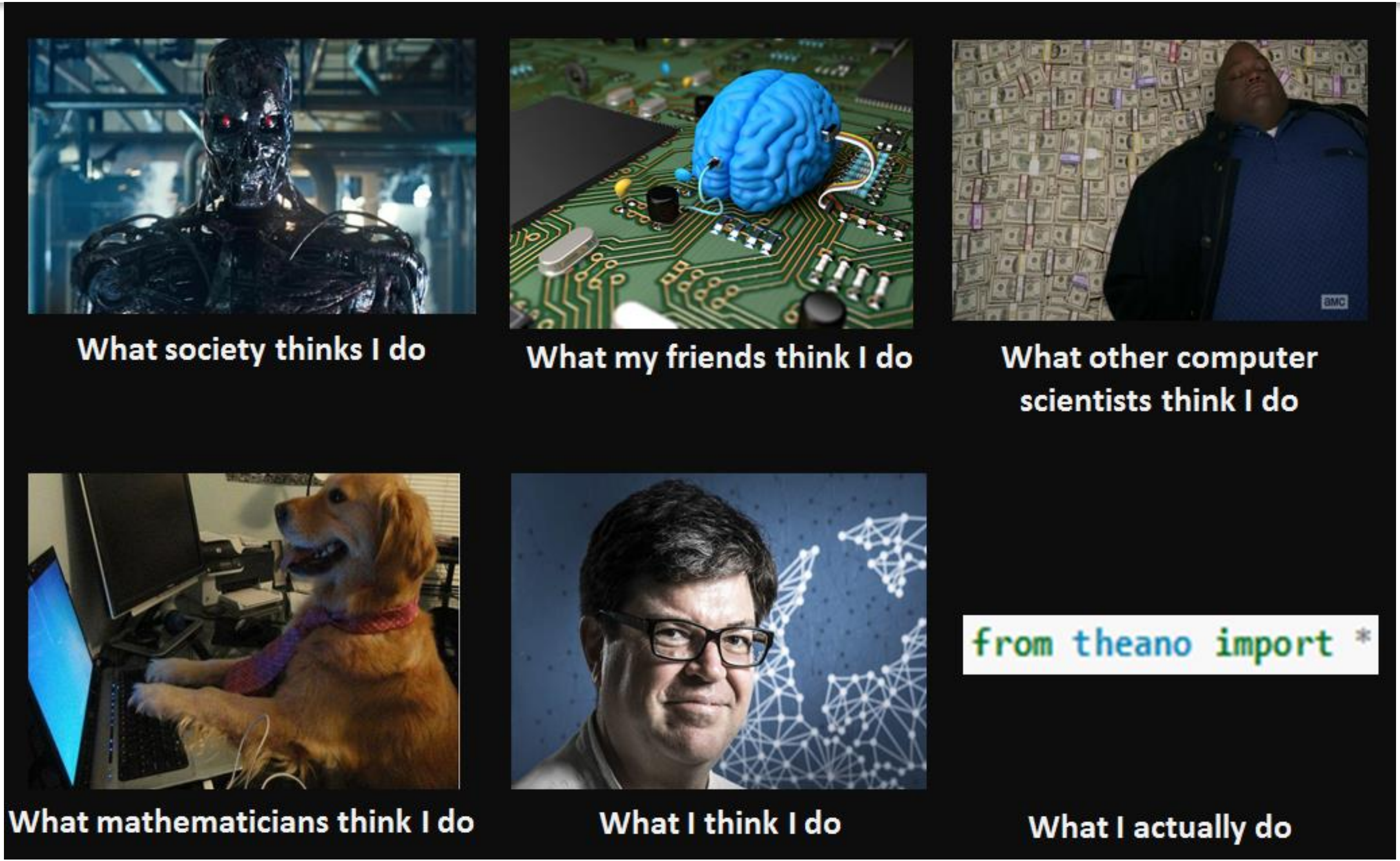
- ML Intro
- Spark Intro
- Data preprocessing steps
- How to build the best Model?
- Demo
- Not only Spark ML for you Data

Let’s go, lover of smoothies

$$\begin{aligned}\bar{x}_1 &= [x_{1,1} \quad x_{1,2} \ldots\ldots\ldots x_{1,p}]\\ \bar{x}_2 &= [x_{2,1} \quad x_{2,2} \ldots\ldots\ldots x_{2,p}]\\ \bar{x}_n &= [x_{n,1} \quad x_{n,2} \ldots\ldots\ldots x_{n,p}]\end{aligned}$$

MACHINE LEARNING

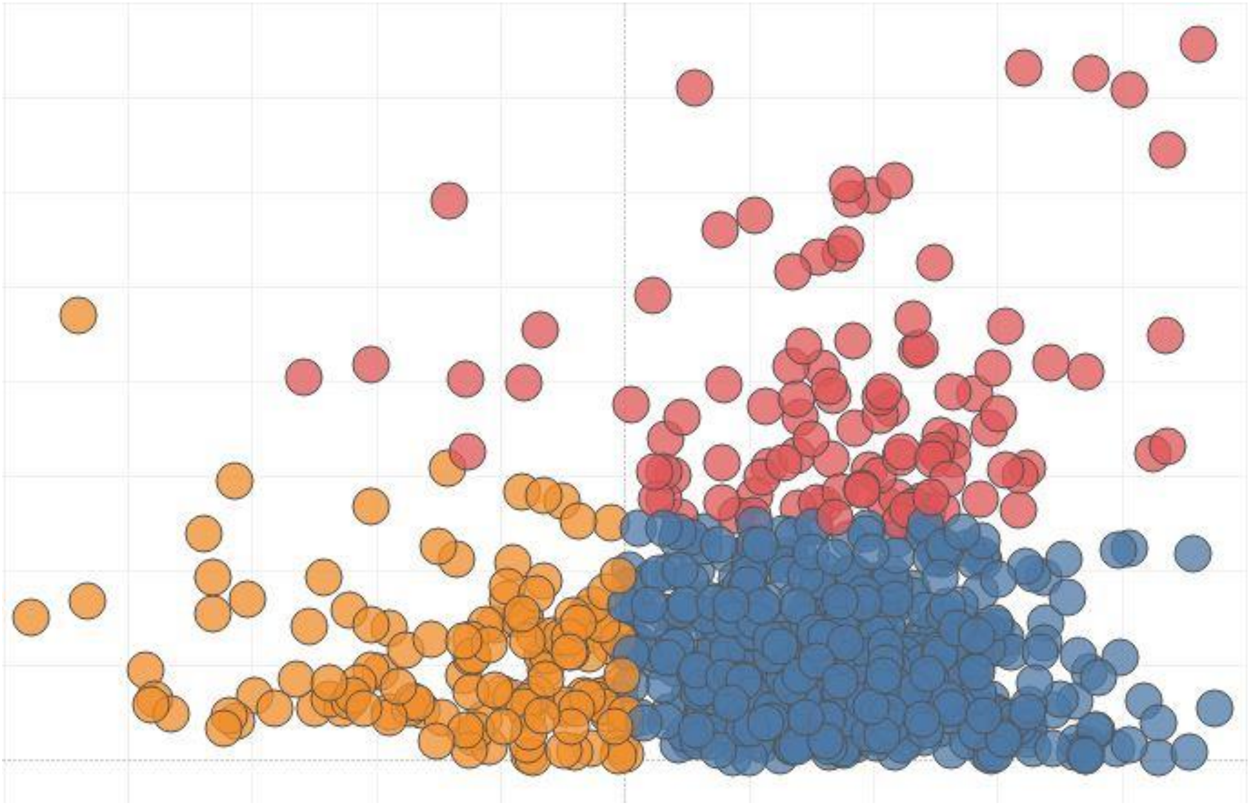
What is Machine Learning?


What society thinks I do
What my friends think I do
What other computer scientists think I do
What mathematicians think I do
What I think I do
from theano import *
What I actually do

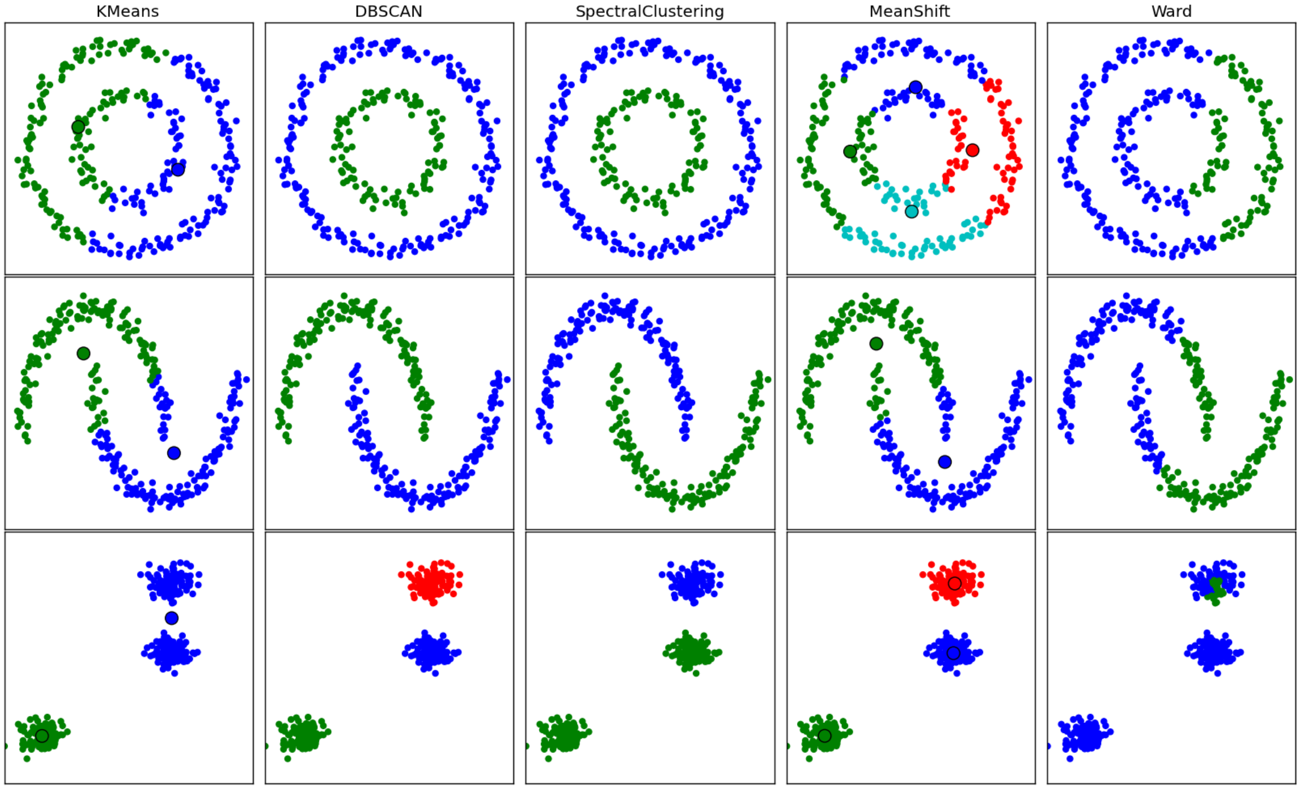
Man or sofa?

Association rule learning

What is Cluster Analysis?



Different algorithms – different results



Example of classification tasks

Input

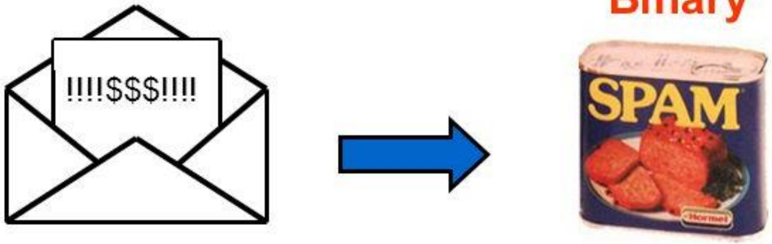
$\mathbf{x} \in \mathcal{X}$

Output

$y \in \mathcal{Y}$

Spam filtering

Binary



Example of classification tasks

Input

$\mathbf{x} \in \mathcal{X}$

Output

$y \in \mathcal{Y}$

Spam filtering

Binary

Character recognition

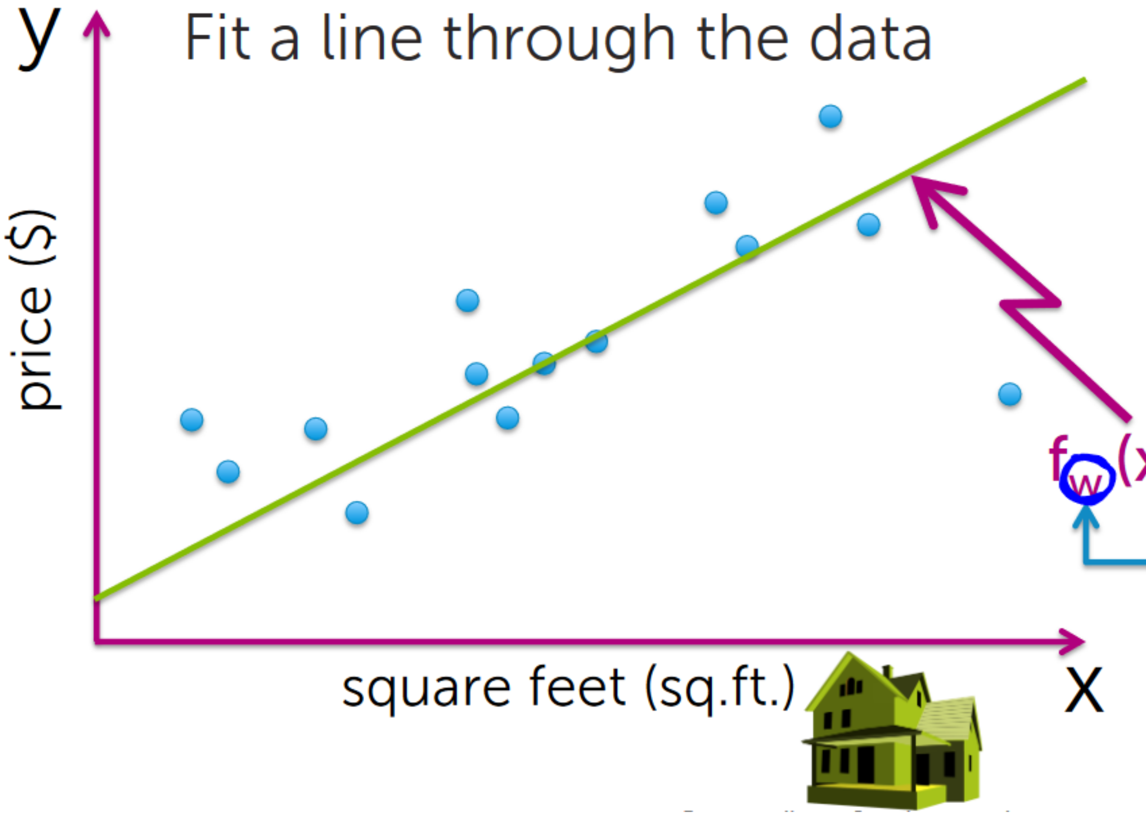
Multi-Class

C

[thanks to Ben Taskar for slide!]

Use a linear model for house price prediction

Fit a line through the data

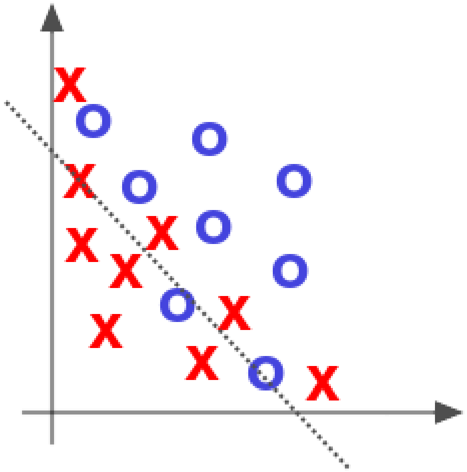


$f_{w}(x) = w_0+w_1 x$

function parameterized by

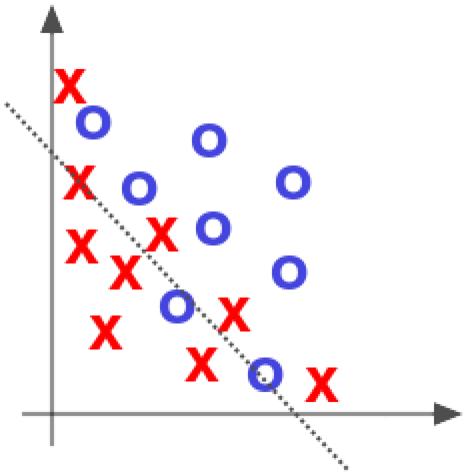
$\boxed{w = (w_0 ,w_1)}$

Underfit vs Overfit

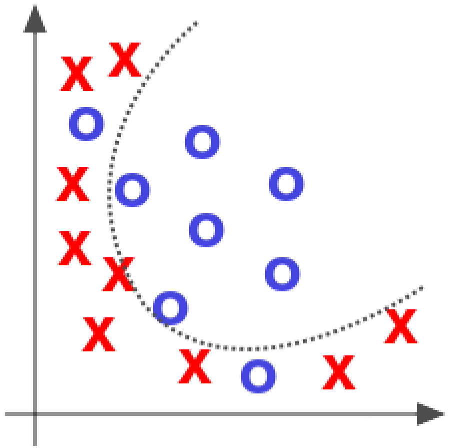


Under Fit

Underfit vs Overfit

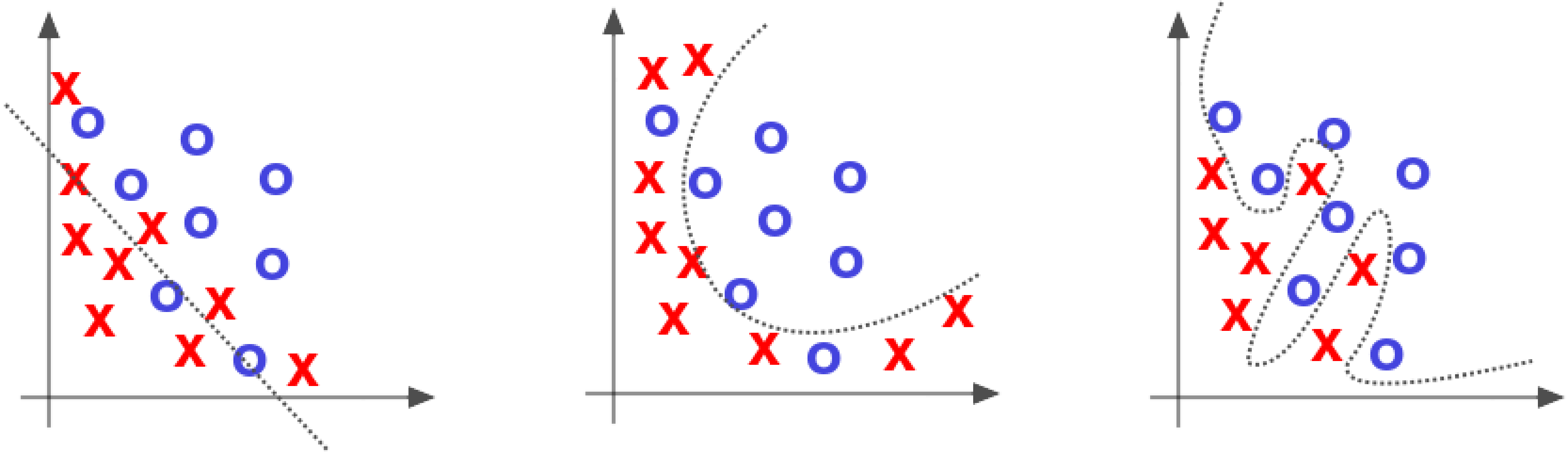


Under Fit



Appropriate

Underfit vs Overfit



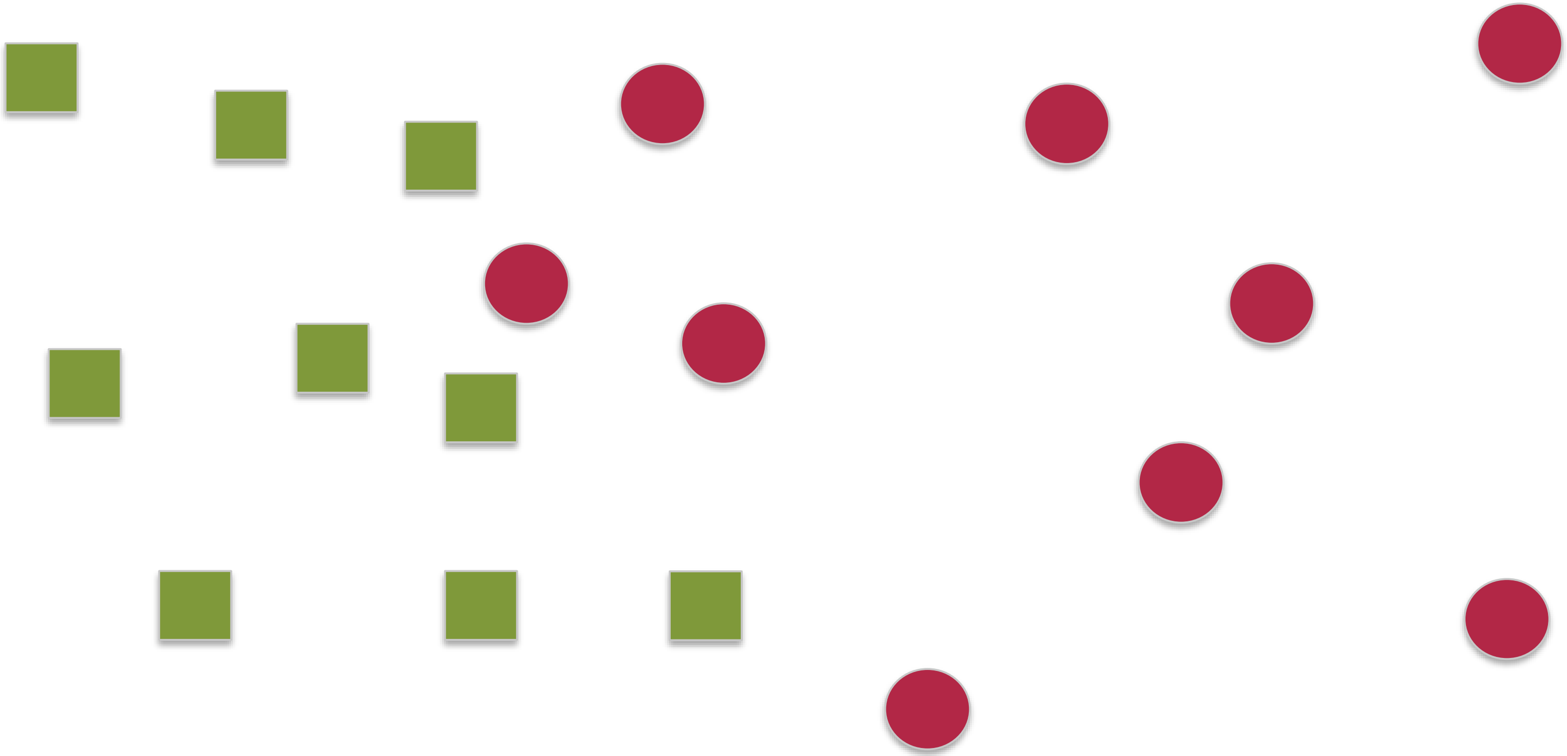
Under Fit

Appropriate

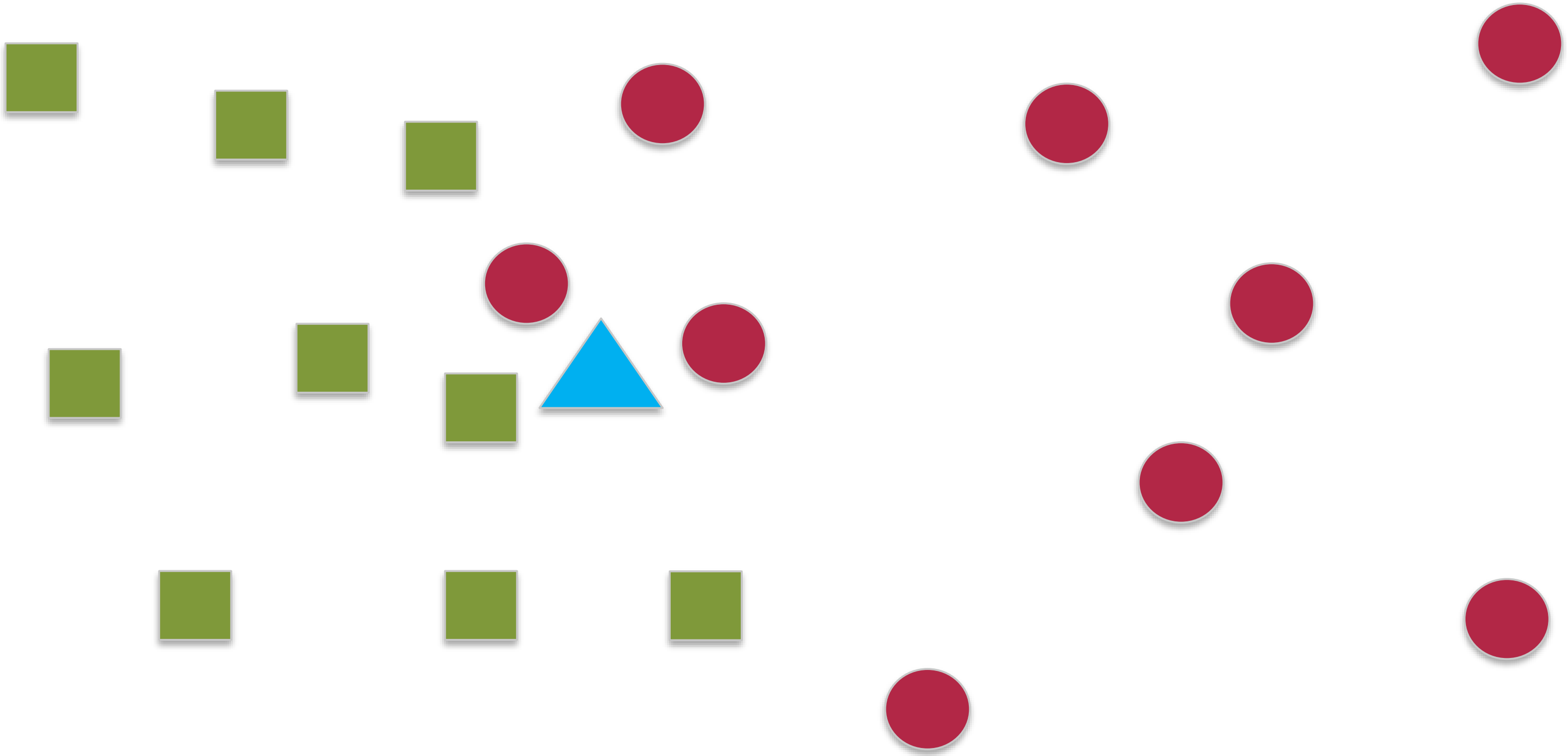
Over Fit

POPULAR ML ALGORITHMS

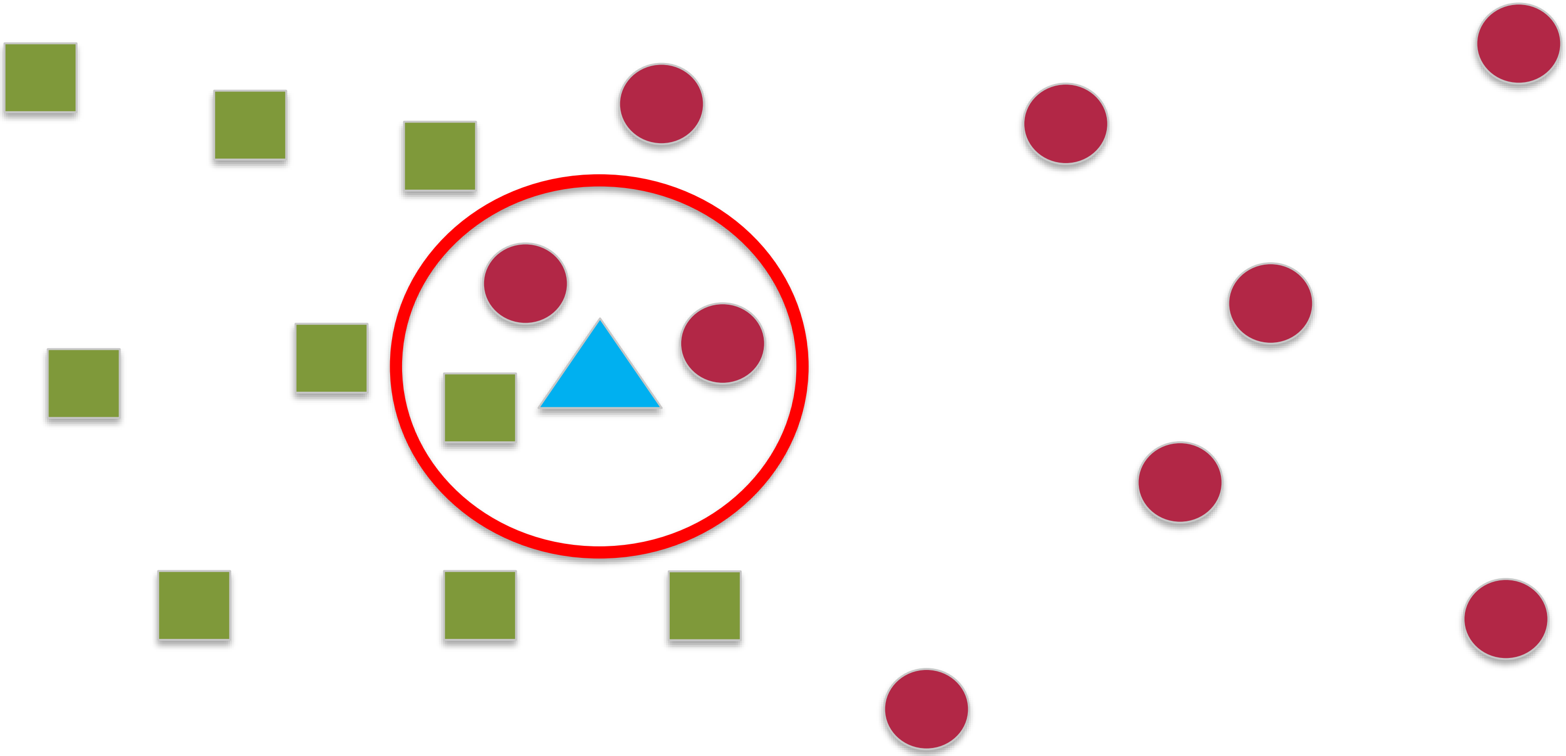
kNN (k-nearest neighbor)



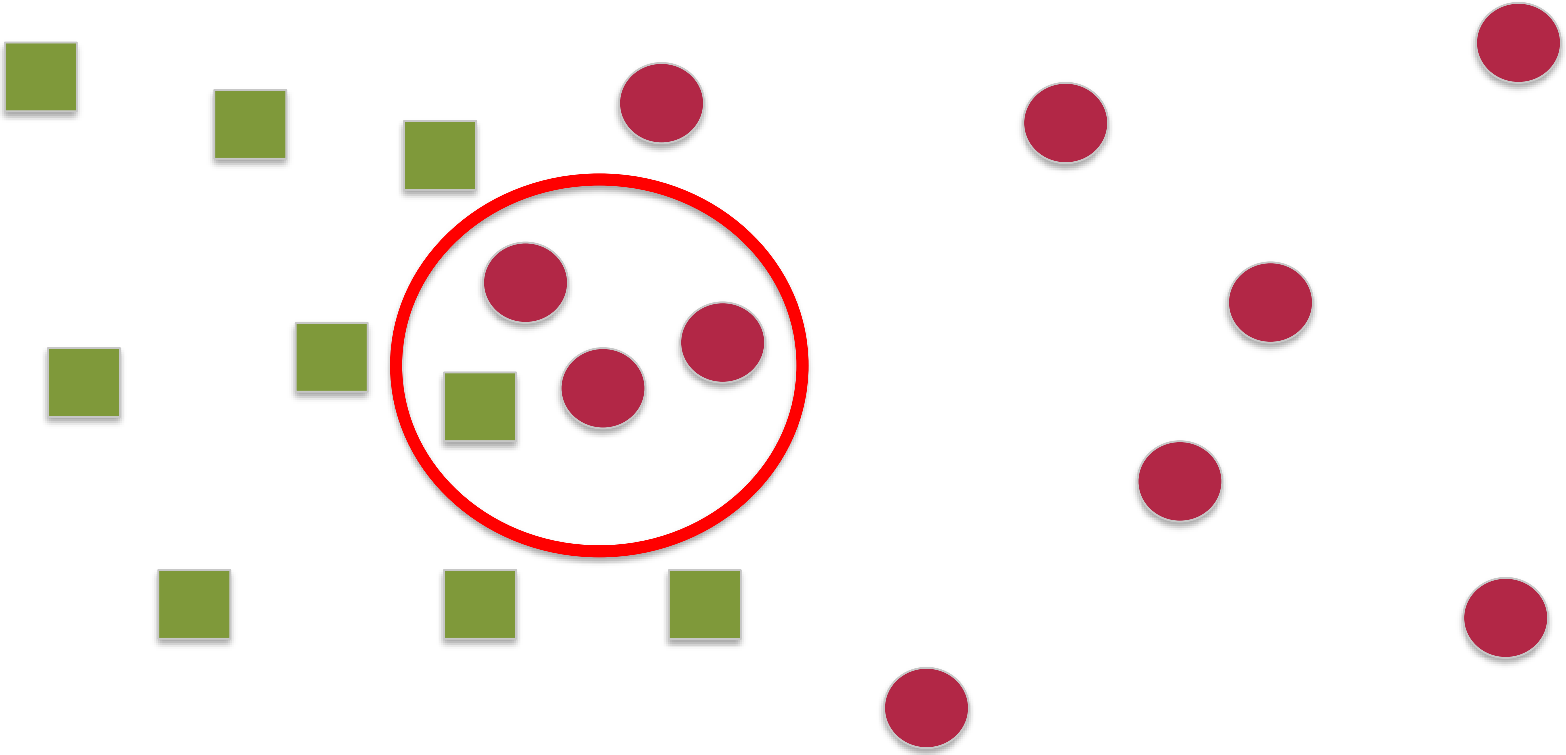
kNN (k-nearest neighbor)



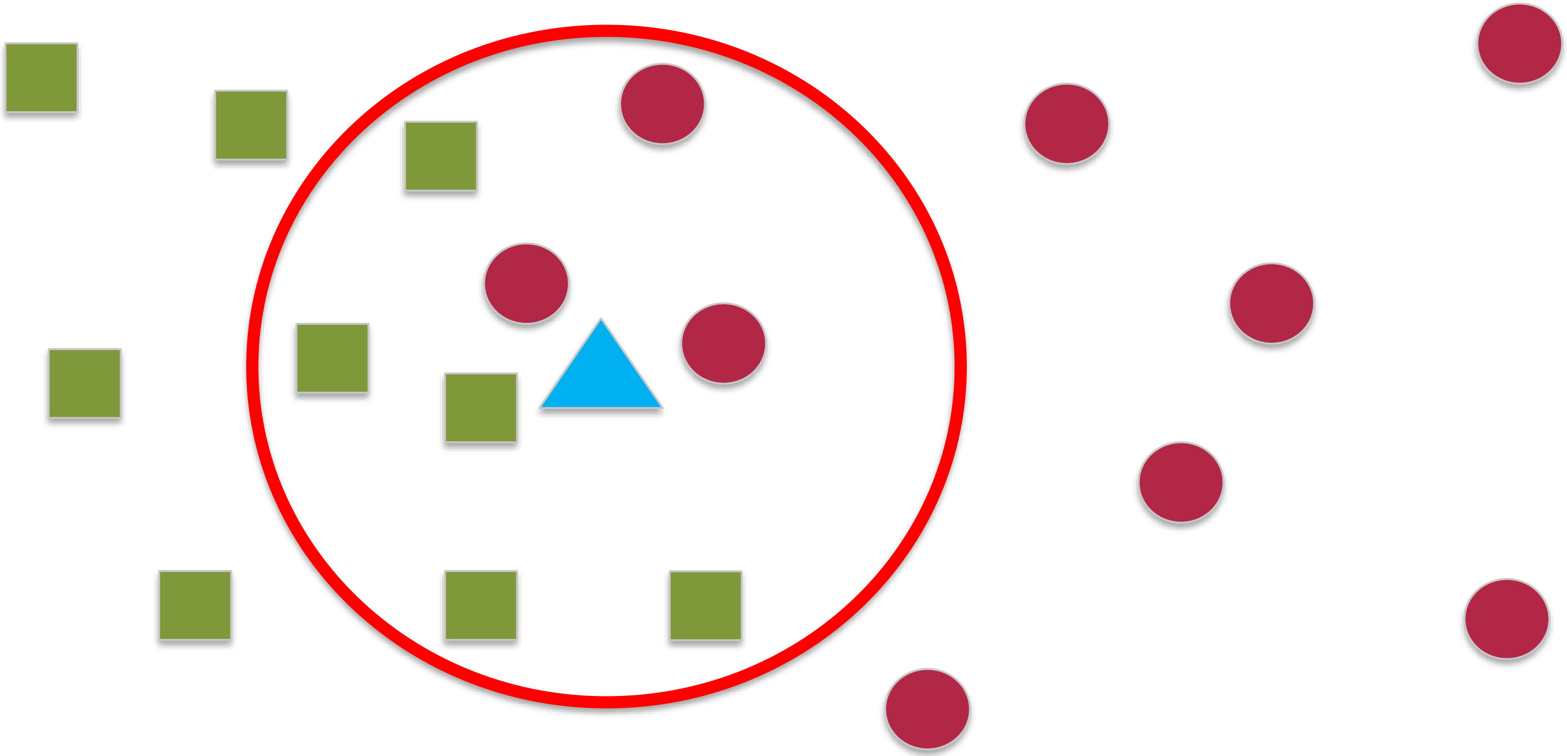
kNN (k-nearest neighbor)



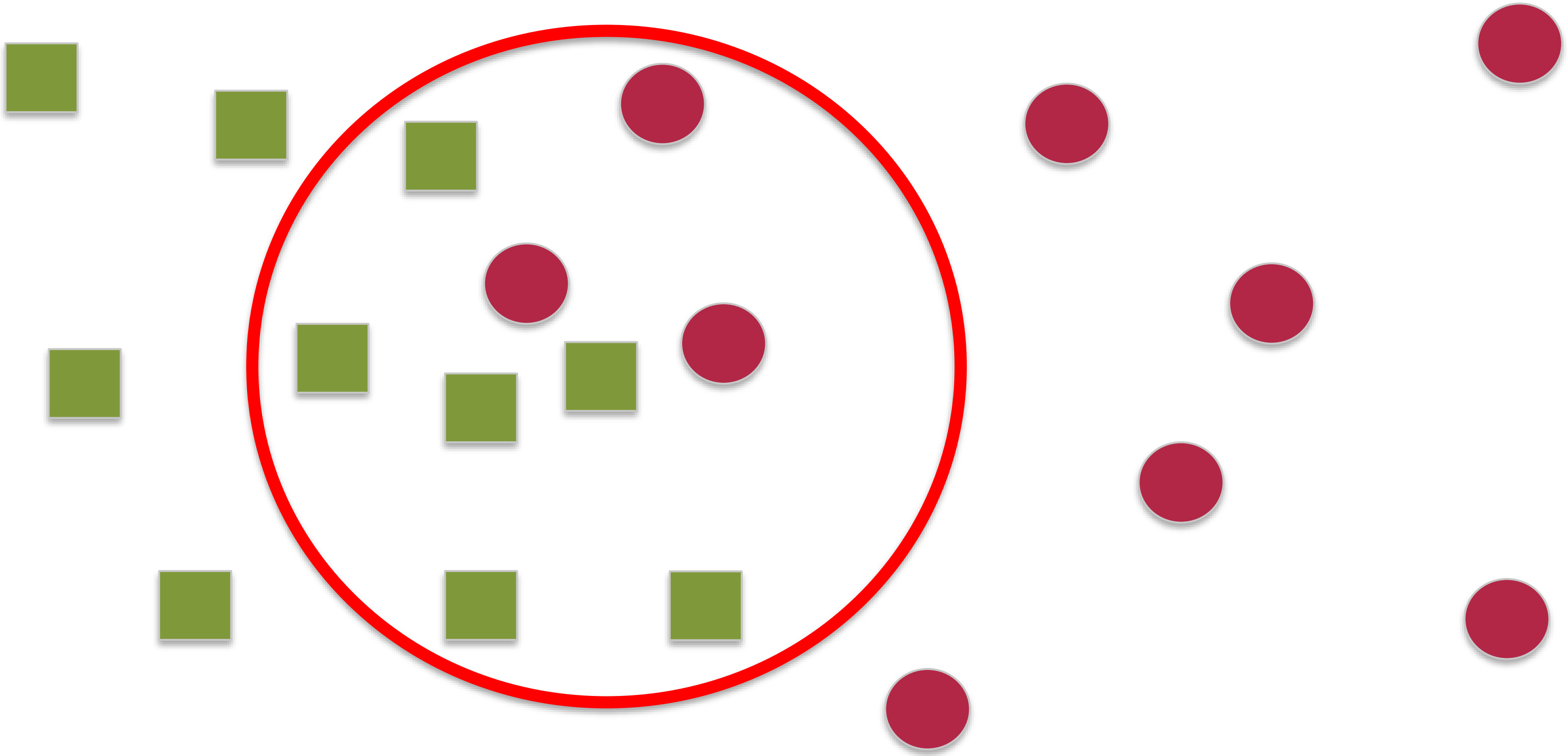
kNN (k-nearest neighbor)



kNN (k-nearest neighbor)



kNN (k-nearest neighbor)



Naive Bayes Classifier

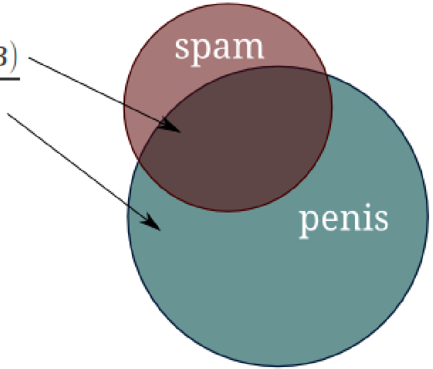
$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$P(C_j \mid A_1, A_2, \ldots, A_n) = \frac{\left(\prod_{i=1}^{n} P(A_i \mid C_j)\right) P(C_j)}{P(A_1, A_2, \ldots, A_n)}$$

Naive Bayes Classifier

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

$$=\frac{P(B|A)P(A)}{P(B)}$$

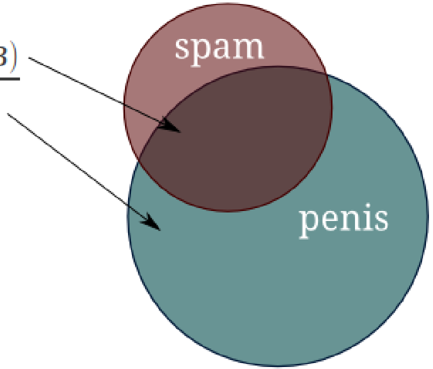


- 30 emails out of a total of 74 are spam messages
- 51 emails out of those 74 contain the word "penis"
- 20 emails containing the word "penis" have been marked as spam

Naive Bayes Classifier

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

$$=\frac{P(B|A)P(A)}{P(B)}$$



$$P(spam|penis)=\frac{P(penis|spam)*P(spam)}{P(penis)}$$

$$=\frac{\frac{20}{30}*\frac{30}{74}}{\frac{51}{74}}=\frac{20}{51}=0.39$$

- 30 emails out of a total of 74 are spam messages
- 51 emails out of those 74 contain the word "penis"
- 20 emails containing the word "penis" have been marked as spam

Naive Bayes Classifier

$$\frac{P(penis|spam \cap viagra) * P(viagra|spam) * P(spam)}{P(penis|viagra) * P(viagra)}$$

$$P(spam|penis, viagra)$$

$$= \frac{P(penis|spam) * P(viagra|spam) * P(spam)}{P(penis) * P(viagra)}$$

$$= \frac{\frac{24}{30} * \frac{20}{30} * \frac{30}{74}}{\frac{25}{74} * \frac{51}{74}} = 0.928$$

- 25 emails out of the total contain the word "viagra"
- 24 emails out of those have been marked as spam
- so what's the probability that an email is spam, given that it contains both "viagra" and "penis"?

Naive Bayes Classifier

$$\frac{P(penis|spam \cap viagra) * P(viagra|spam) * P(spam)}{P(penis|viagra) * P(viagra)}$$

$P(spam|penis, viagra)$

$= \frac{P(penis|spam) * P(vi}{P(p}$

$= \frac{\frac{24}{30}}{\frac{25}{74}}$

- 25 emails out of th

  "viagra"

- been marked as

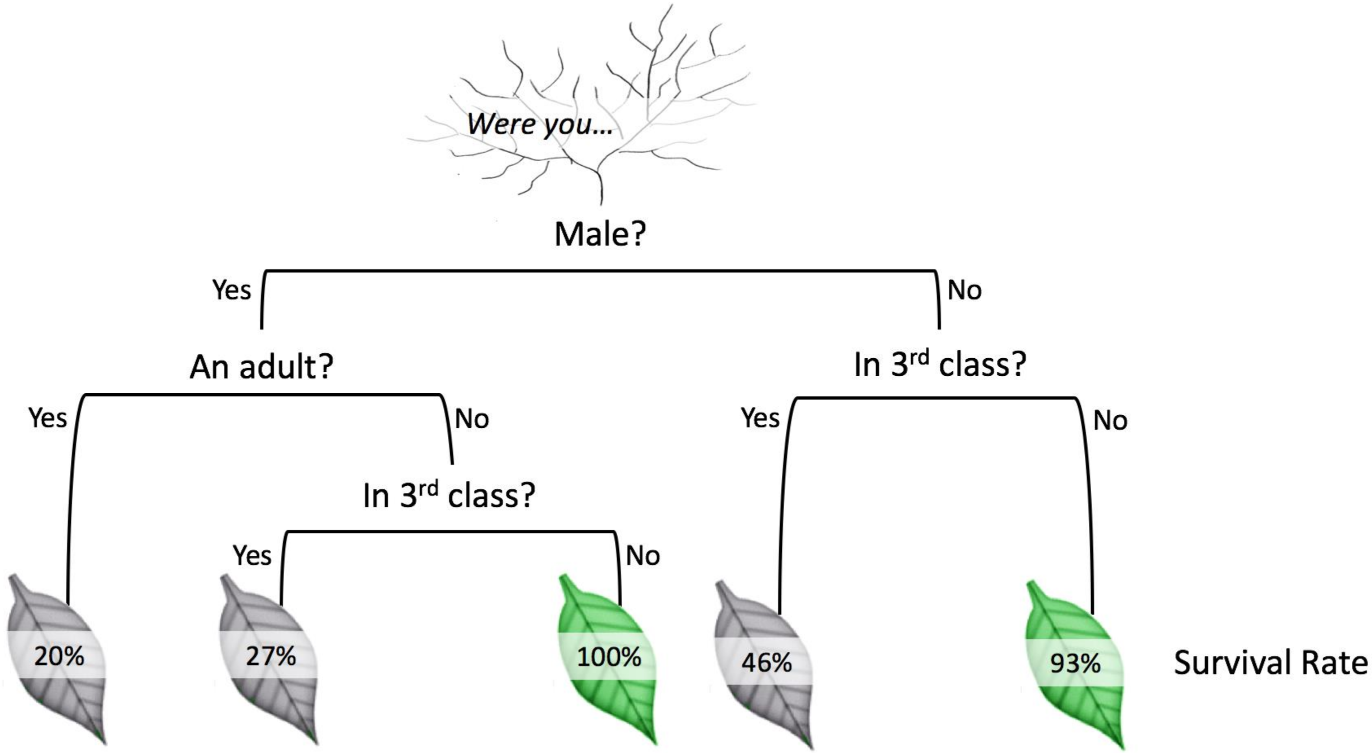
- the probability that an email is

  am, given that it contains both "viagra" and "penis"?

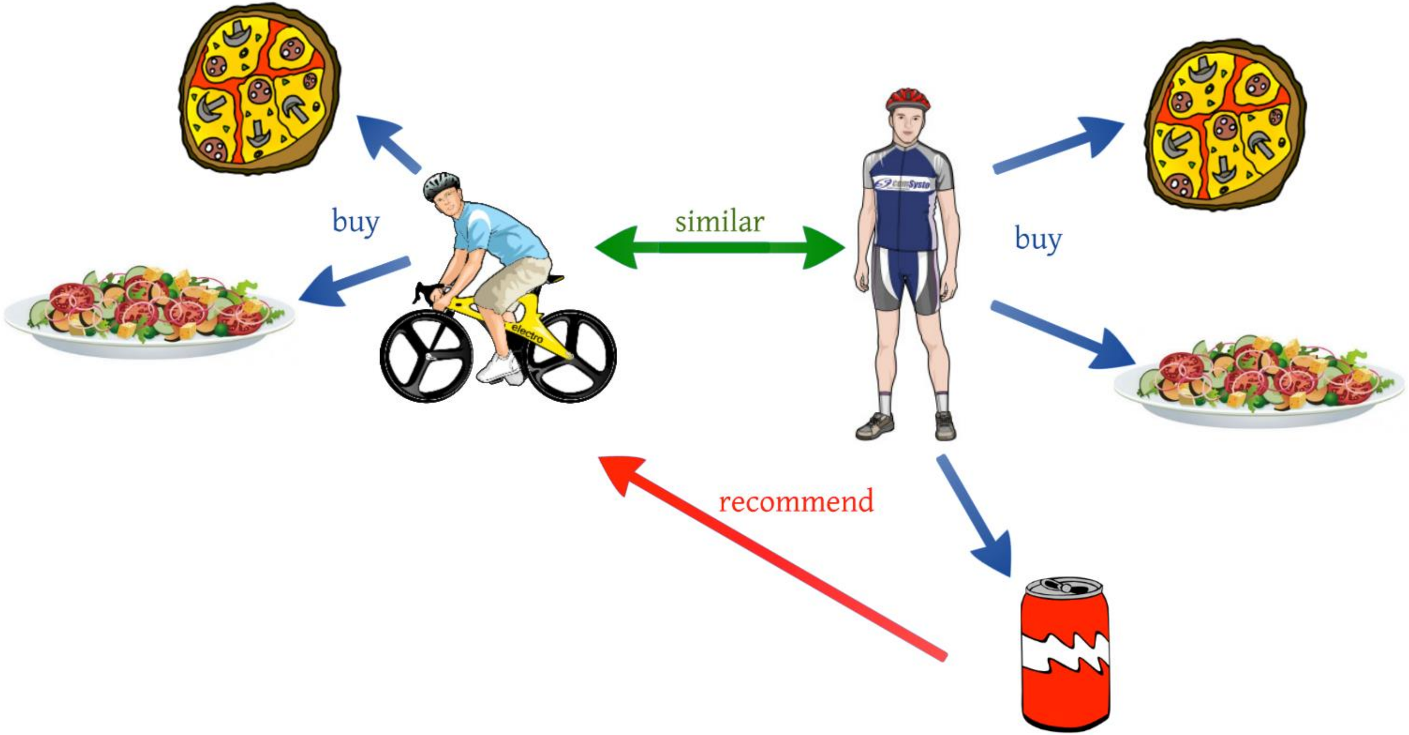
If “penis” and “viagra” are independent

Decision trees & Titanic passengers dataset

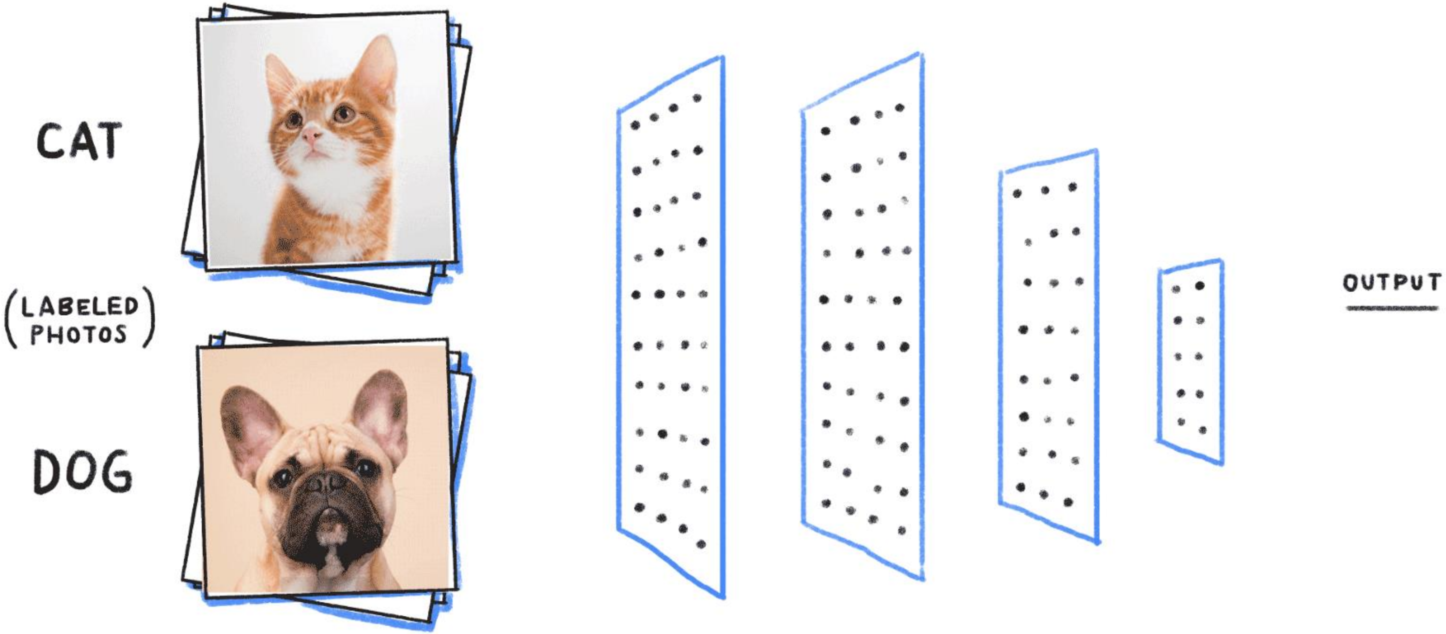
Cruel Tree


Were you…
Male?
Yes
No
An adult?
Yes
No
In 3rd class?
Yes
No
In 3rd class?
Yes
No
20%
27%
100%
46%
93%
Survival Rate

Collaborative Filtering



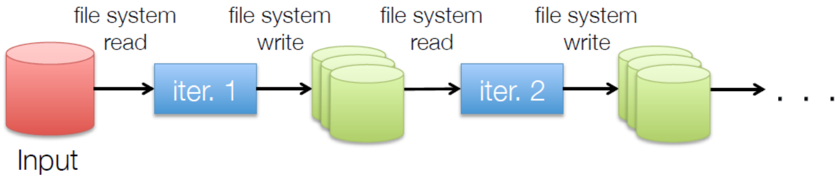
Neural Networks

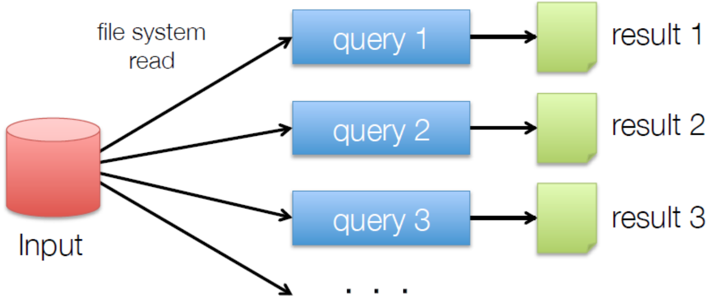


Machine Learning EVERYWHERE

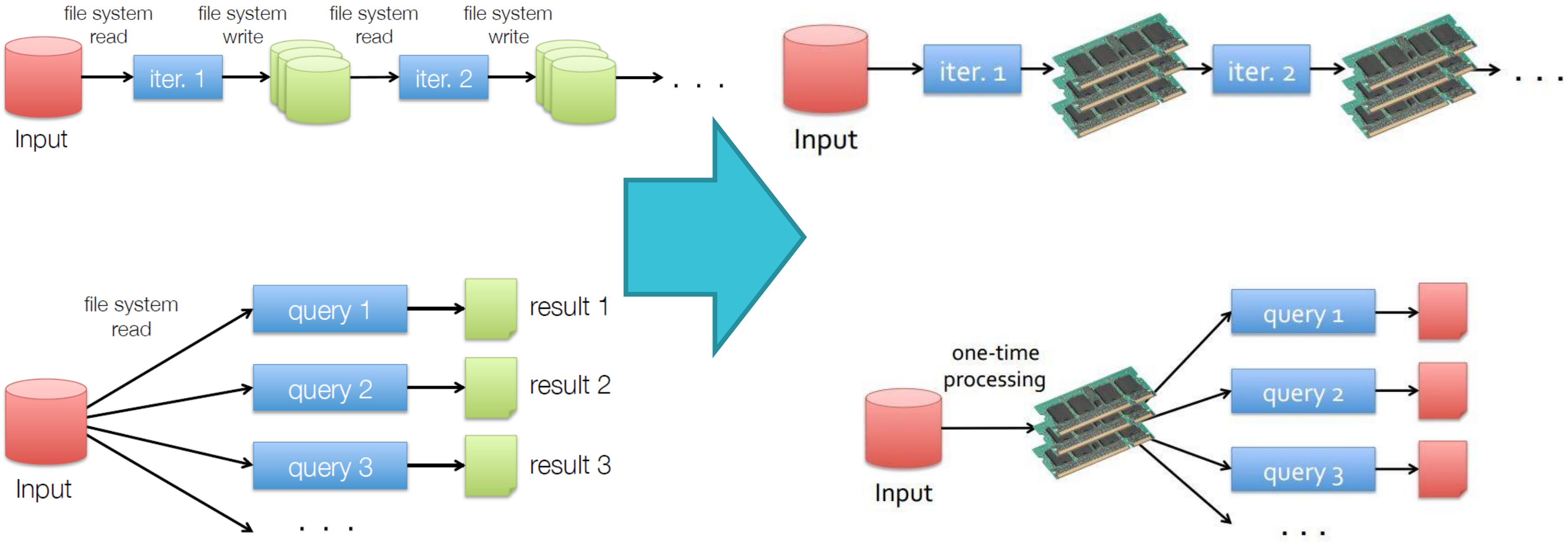
SPARK INTRO

MapReduce vs Spark

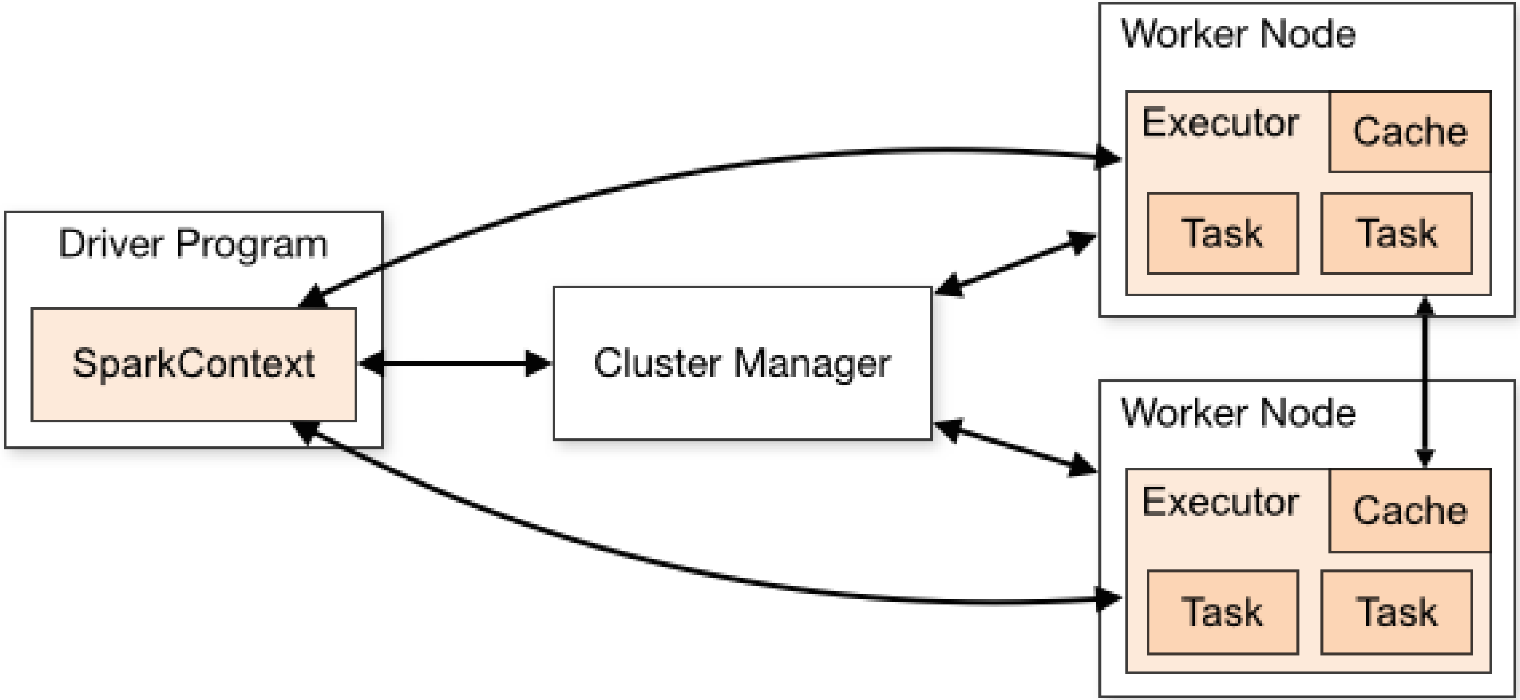




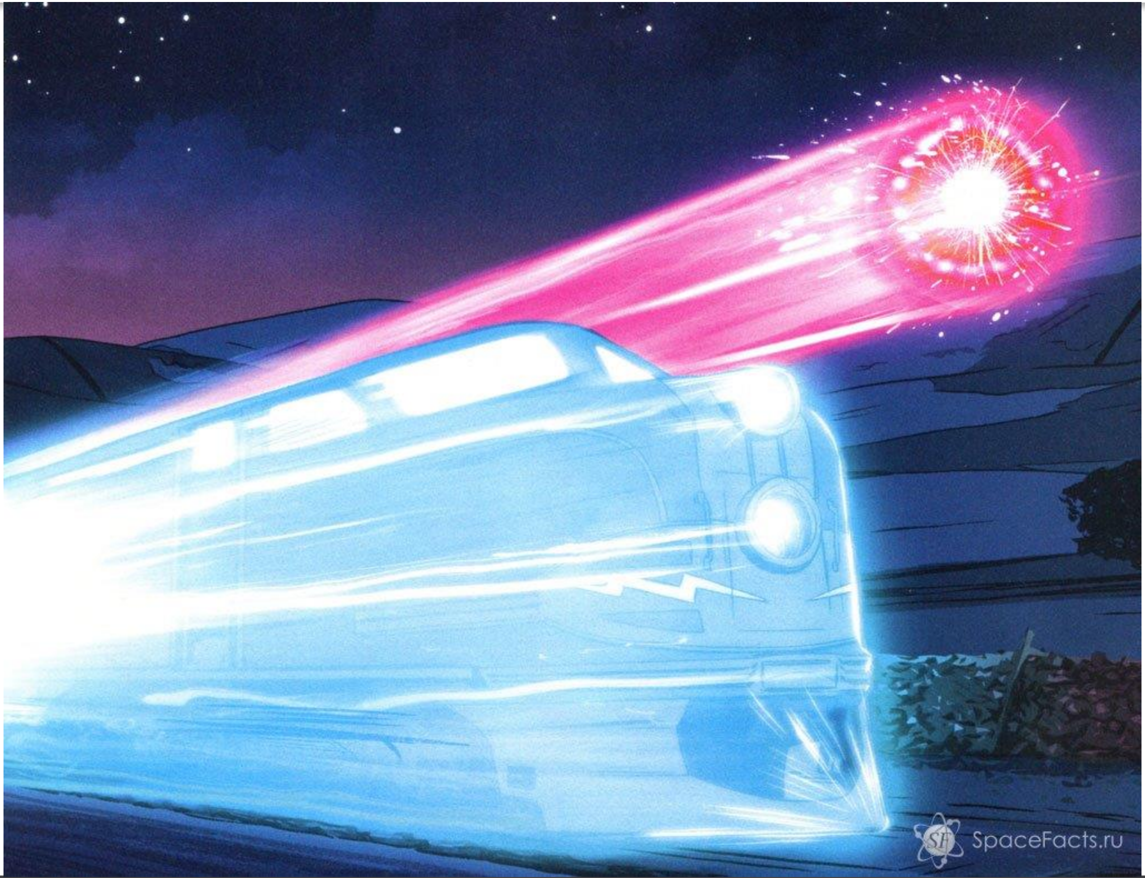
MapReduce vs Spark



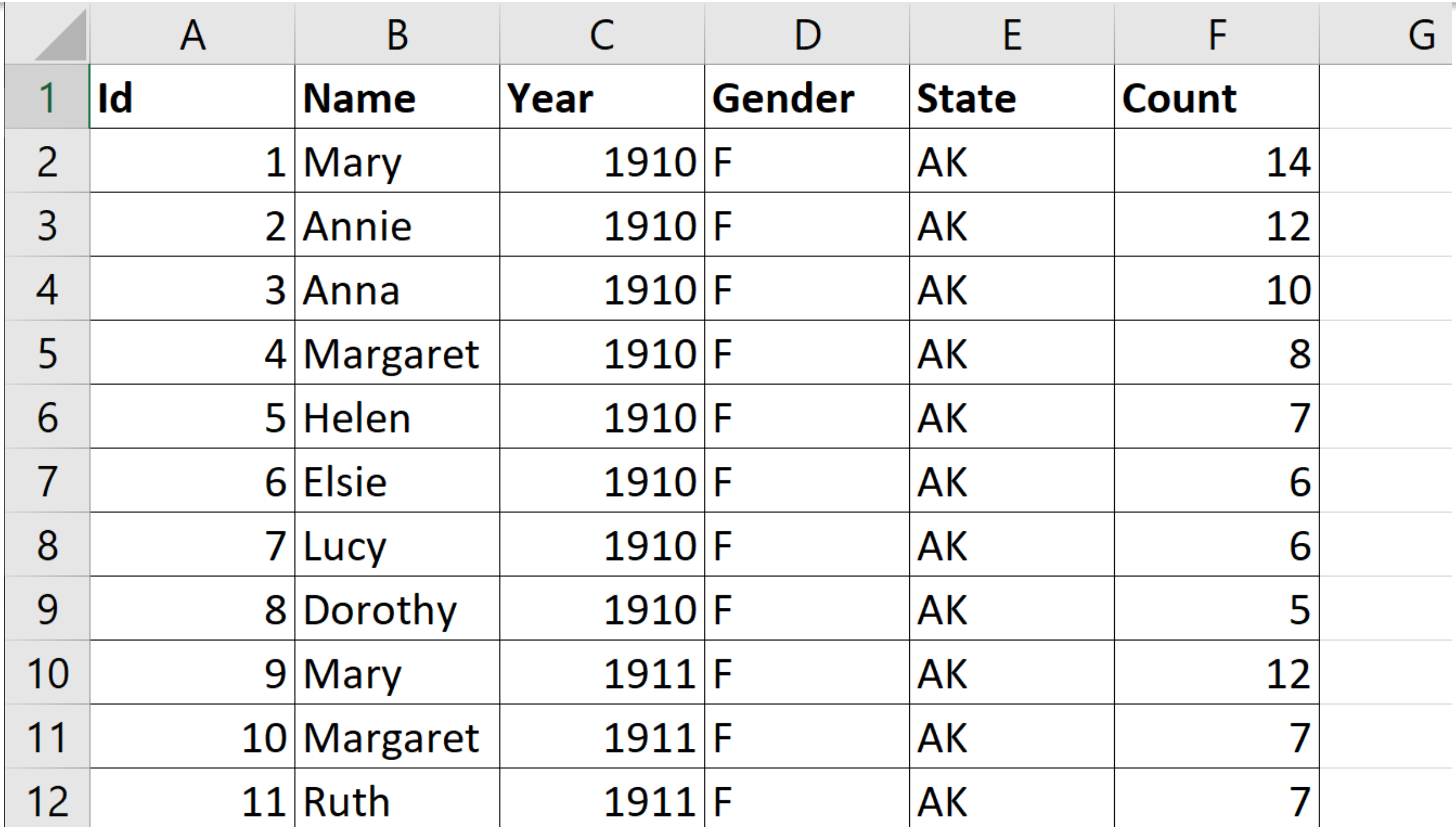
Worker Nodes and Executors



Let’s use Spark. It’s fast!



State_Names.csv



	A	B	C	D	E	F	G
1	**Id**	**Name**	**Year**	**Gender**	**State**	**Count**	
2	1	Mary	1910	F	AK	14	
3	2	Annie	1910	F	AK	12	
4	3	Anna	1910	F	AK	10	
5	4	Margaret	1910	F	AK	8	
6	5	Helen	1910	F	AK	7	
7	6	Elsie	1910	F	AK	6	
8	7	Lucy	1910	F	AK	6	
9	8	Dorothy	1910	F	AK	5	
10	9	Mary	1911	F	AK	12	
11	10	Margaret	1911	F	AK	7	
12	11	Ruth	1911	F	AK	7	

Create context

```
val spark = SparkSession.builder
      .master("local[2]")
      .appName("DataFrameIntro")
      .getOrCreate()
```

Read from file

```
val stateNames = spark.read
       .option("header", "true")
       .option("inferSchema", "true")
       .csv("/home/data/StateNames.csv")
```

Prepare report

```
// Registered births by year in US since 1880
    nationalNames
      .groupBy("Year")
      .sum("Count").as("Sum")
      .orderBy("Year")
      .show(200)
```

HOW TO DEVELOP?

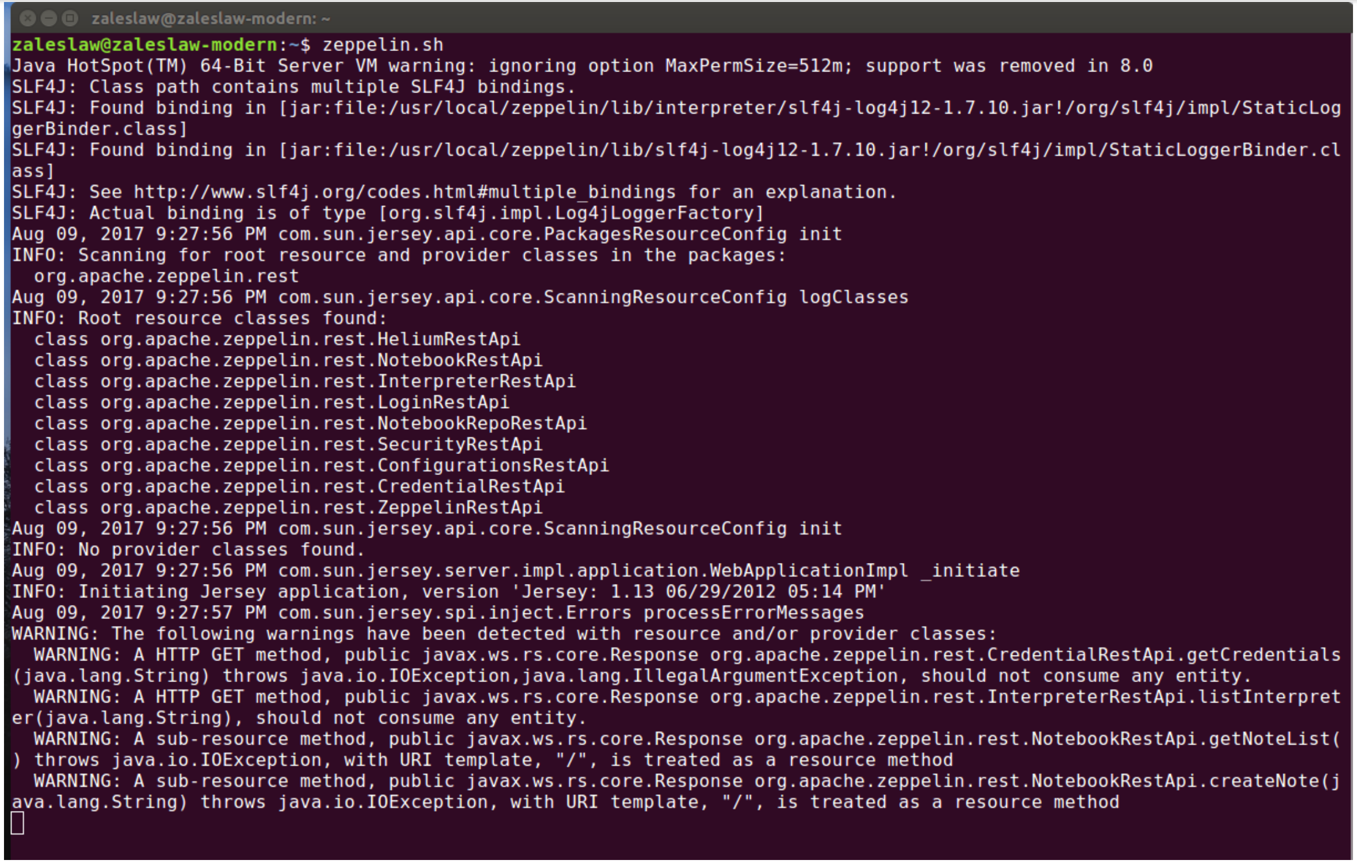
Development tools

- Console REPL ($SPARK_HOME/sbin/spark-shell)

Development tools

- Console REPL ($SPARK_HOME/sbin/spark-shell)
- Apache Zeppelin

Run Zeppelin



```
zaleslaw@zaleslaw-modern:~$ zeppelin.sh
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512m; support was removed in 8.0
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/zeppelin/lib/interpreter/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLog
gerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/zeppelin/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.cl
ass]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Aug 09, 2017 9:27:56 PM com.sun.jersey.api.core.PackagesResourceConfig init
INFO: Scanning for root resource and provider classes in the packages:
  org.apache.zeppelin.rest
Aug 09, 2017 9:27:56 PM com.sun.jersey.api.core.ScanningResourceConfig logClasses
INFO: Root resource classes found:
  class org.apache.zeppelin.rest.HeliumRestApi
  class org.apache.zeppelin.rest.NotebookRestApi
  class org.apache.zeppelin.rest.InterpreterRestApi
  class org.apache.zeppelin.rest.LoginRestApi
  class org.apache.zeppelin.rest.NotebookRepoRestApi
  class org.apache.zeppelin.rest.SecurityRestApi
  class org.apache.zeppelin.rest.ConfigurationsRestApi
  class org.apache.zeppelin.rest.CredentialRestApi
  class org.apache.zeppelin.rest.ZeppelinRestApi
Aug 09, 2017 9:27:56 PM com.sun.jersey.api.core.ScanningResourceConfig init
INFO: No provider classes found.
Aug 09, 2017 9:27:56 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
INFO: Initiating Jersey application, version 'Jersey: 1.13 06/29/2012 05:14 PM'
Aug 09, 2017 9:27:57 PM com.sun.jersey.spi.inject.Errors processErrorMessages
WARNING: The following warnings have been detected with resource and/or provider classes:
  WARNING: A HTTP GET method, public javax.ws.rs.core.Response org.apache.zeppelin.rest.CredentialRestApi.getCredentials
(java.lang.String) throws java.io.IOException,java.lang.IllegalArgumentException, should not consume any entity.
  WARNING: A HTTP GET method, public javax.ws.rs.core.Response org.apache.zeppelin.rest.InterpreterRestApi.listInterpret
er(java.lang.String), should not consume any entity.
  WARNING: A sub-resource method, public javax.ws.rs.core.Response org.apache.zeppelin.rest.NotebookRestApi.getNoteList(
) throws java.io.IOException, with URI template, "/", is treated as a resource method
  WARNING: A sub-resource method, public javax.ws.rs.core.Response org.apache.zeppelin.rest.NotebookRestApi.createNote(j
ava.lang.String) throws java.io.IOException, with URI template, "/", is treated as a resource method
```

Development tools

- Console REPL ($SPARK_HOME/sbin/spark-shell)
- Apache Zeppelin
- IntelliJ IDEA Community + Scala Plugin

Development tools

- Console REPL ($SPARK_HOME/sbin/spark-shell)
- Apache Zeppelin
- IntelliJ IDEA Community + Scala Plugin
- Don’t forget about SBT or adding spark’s jars

SBT build

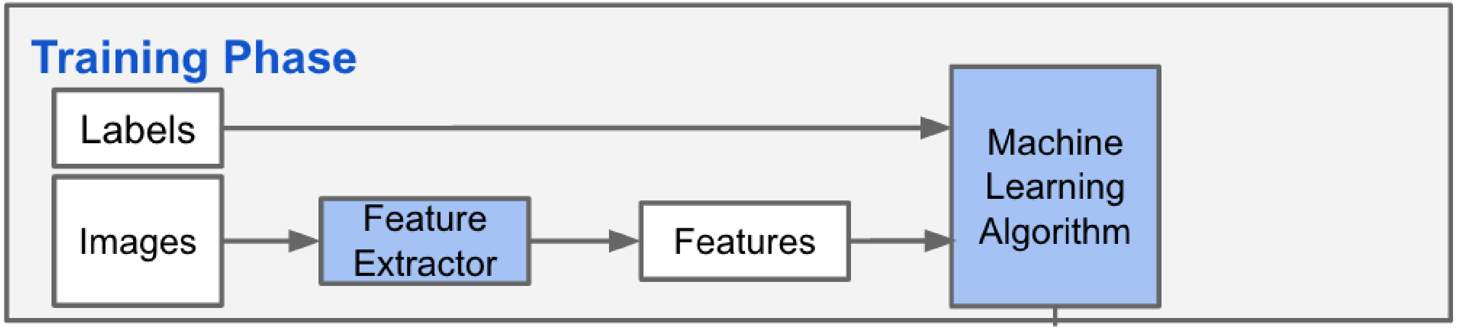
```
name := "Spark-app"
version := "1.0"
scalaVersion := "2.11.11"

libraryDependencies += "org.apache.spark" % "spark-core_2.11" % "2.2.0"
libraryDependencies += "org.apache.spark" % "spark-sql_2.11" % "2.2.0"
libraryDependencies += "org.apache.spark" % "spark-mllib_2.11" % "2.2.0"
```

PREPROCESSING

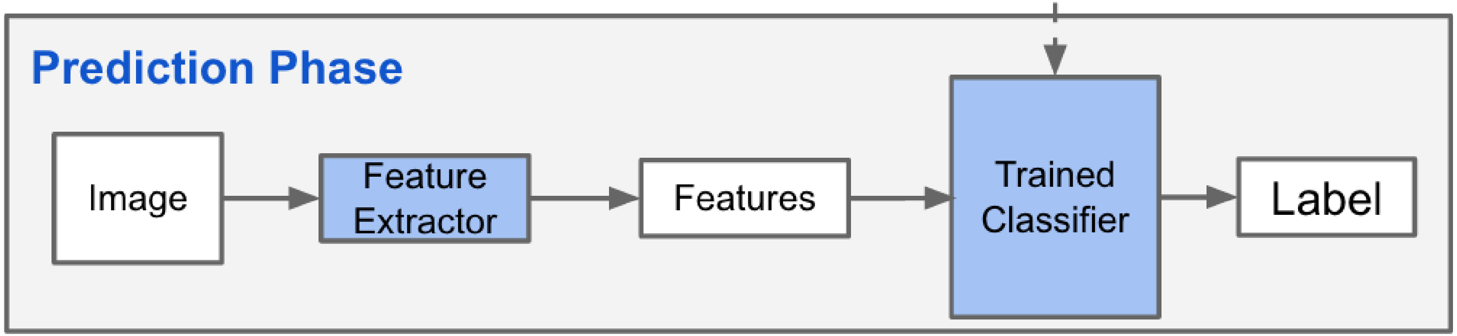
The main concept of tabular dataset

Features = columns, observations = rows

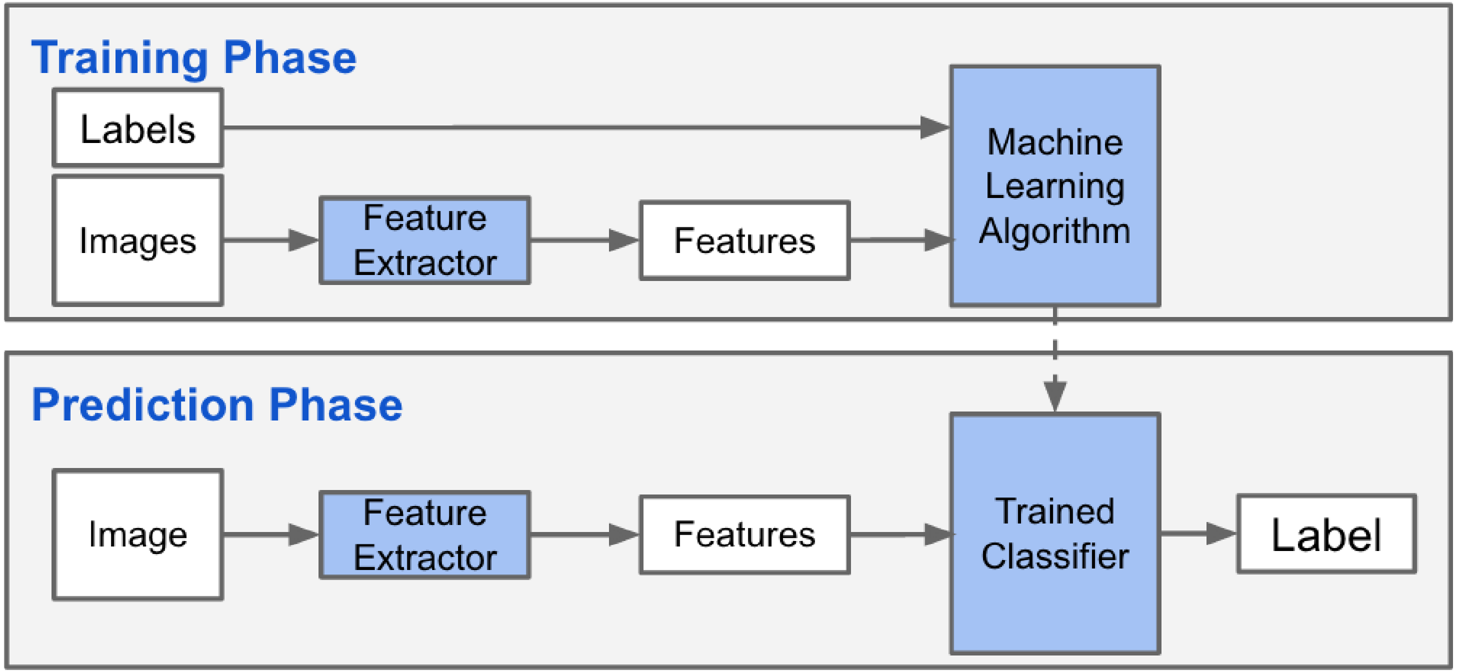
Machine Learning Phases



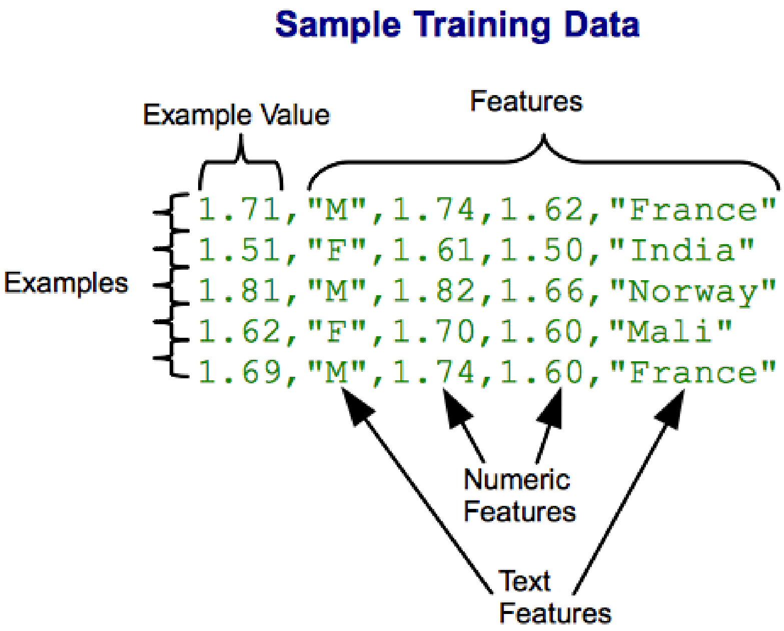
Machine Learning Phases



Machine Learning Phases

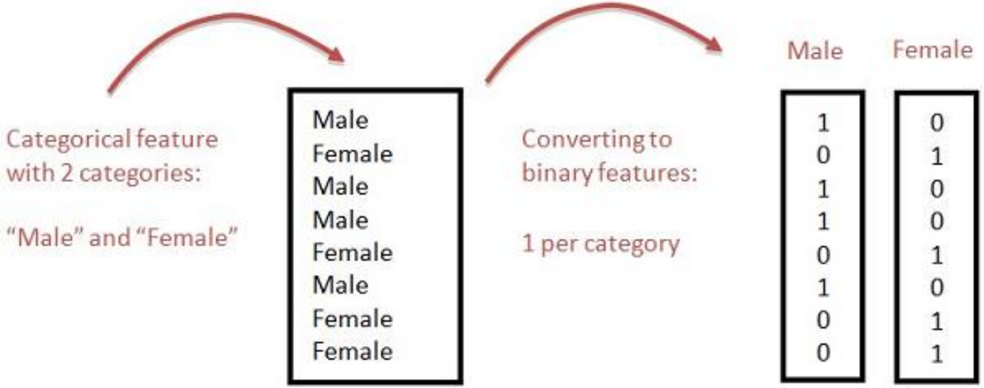


Feature types and transformation



Sample Training Data

```
1.71,"M",1.74,1.62,"France"
1.51,"F",1.61,1.50,"India"
1.81,"M",1.82,1.66,"Norway"
1.62,"F",1.70,1.60,"Mali"
1.69,"M",1.74,1.60,"France"
```



Categorical feature with 2 categories:

“Male” and “Female”

Converting to binary features:

1 per category

	Male	Female
Male	1	0
Female	0	1
Male	1	0
Male	1	0
Female	0	1
Male	1	0
Female	0	1
Female	0	1

Normalizer

```
val normalizer = new Normalizer()
  .setInputCol(“old_features”)
  .setOutputCol(“new_features”)
  .setP(2.0) // L ^ P - norm

val l2NormData = normalizer.transform(dataset)
```

Normalizer

```
val scaler = new MinMaxScaler()
  .setInputCol(“old_features”)
  .setOutputCol(“new_features”)

// calculate stat over the data
val scalerModel = scaler.fit(dataset)

// rescale each feature to range [min, max]
val scaledData = scalerModel.transform(dataset)
```

DATA TYPES

Data Types in MLlib

- Vector (mllib.linalg.Vectors class)
- LabeledPoint (mllib.regression.LabeledPoint)
- Rating (mllib.recommendation.Rating)
- Local matrix loaded from LibSVM
- RowMatrix
- BlockMatrix

Vectors

```
import org.apache.spark.mllib.linalg.{Vector, Vectors}

val v1: Vector = Vectors.dense(1.0, 0.0, 3.0)

val v2: Vector = Vectors.sparse(3, Array(0, 2), Array(1.0, 3.0))

val v3: Vector = Vectors.sparse(3, Seq((0, 1.0), (2, 3.0)))
```

Labeled Point

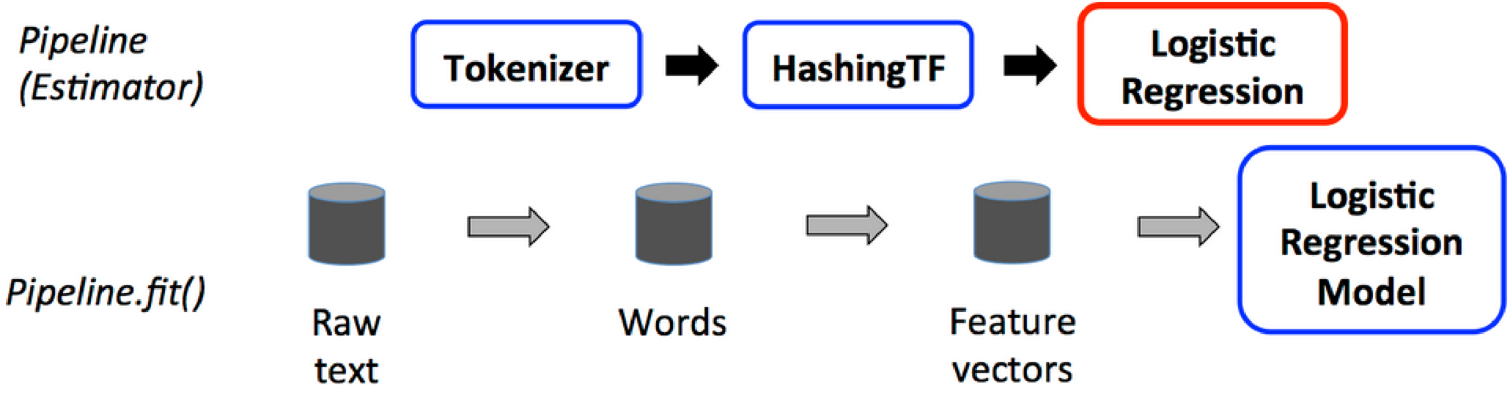
```
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.regression.LabeledPoint

val lp1 = LabeledPoint(1.0, Vectors.dense(1.0, 0.0, 3.0))

val lp2 = LabeledPoint(0.0, Vectors.sparse(3, Array(0, 2), 
Array(1.0, 3.0)))
```

How to use DataFrames?

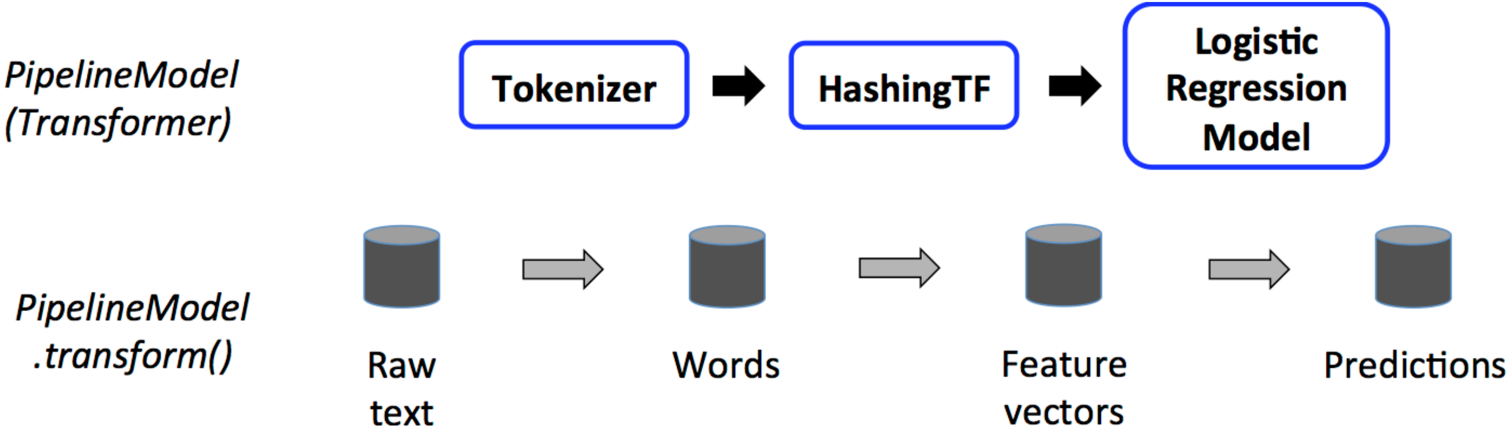
Build ML Pipelines with …

- DataFrame
- Transformer
- Estimator
- Pipeline
- Parameter

Pipeline: Model Generation



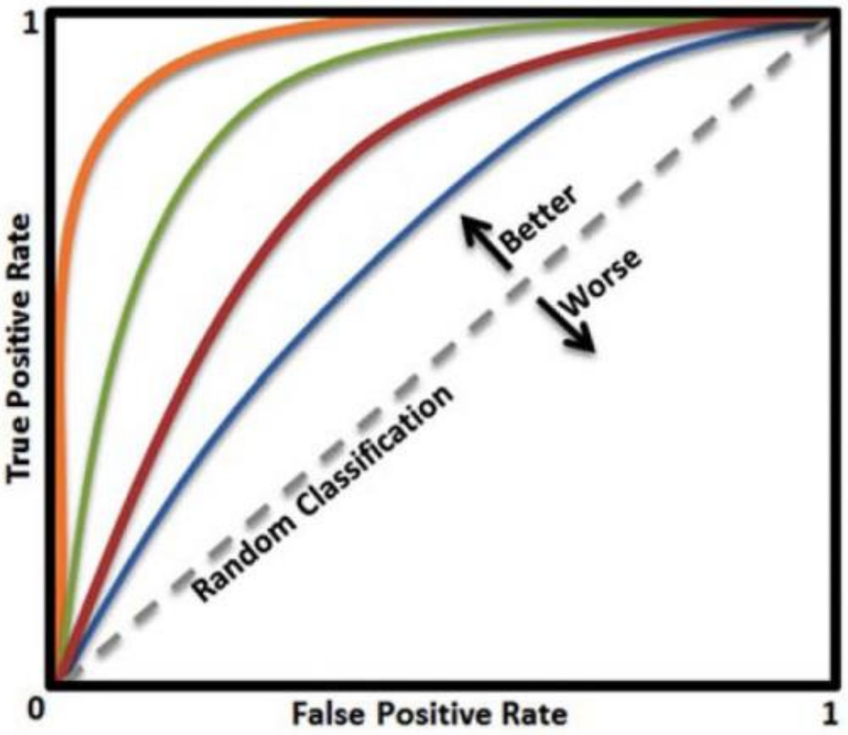
Pipeline: Model Usage


PipelineModel (Transformer)
Tokenizer
HashingTF
Logistic Regression Model
PipelineModel .transform()
Raw text
Words
Feature vectors
Predictions

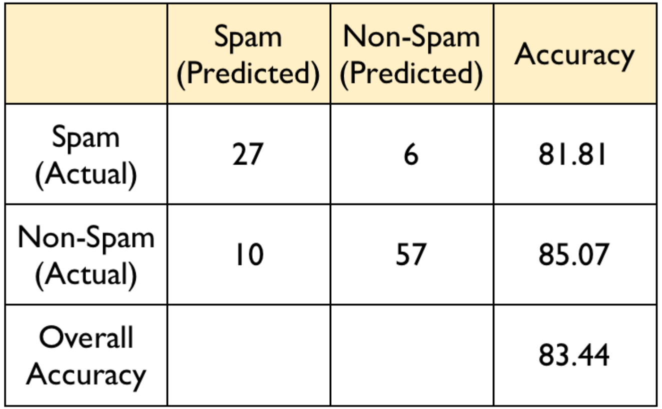
How to choose the best model?

MODEL ACCURACY

ROC AUC for binary classification



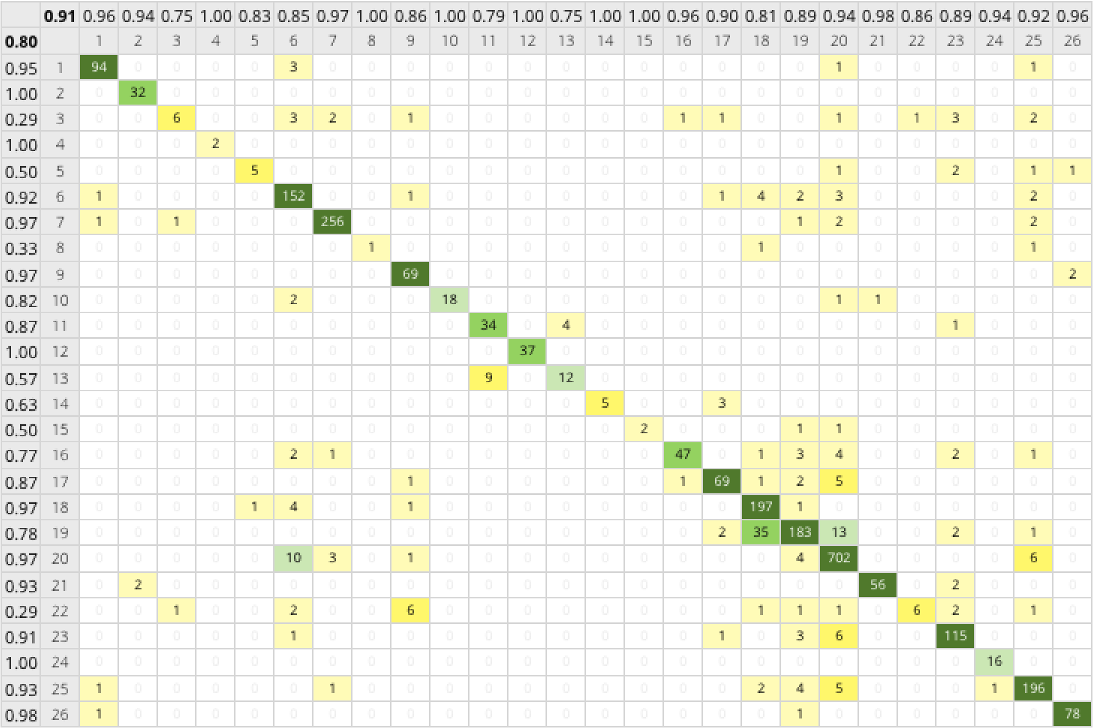
Confusion Matrix for two



	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

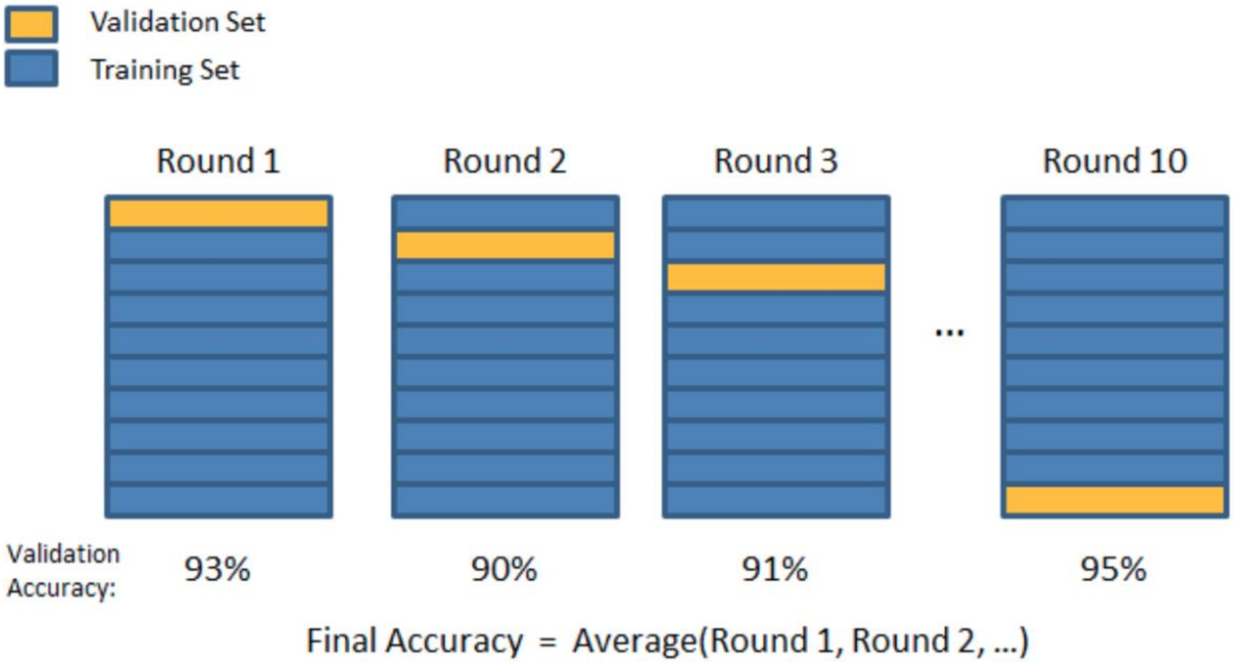
Confusion Matrix for two and more classes

	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44



	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78

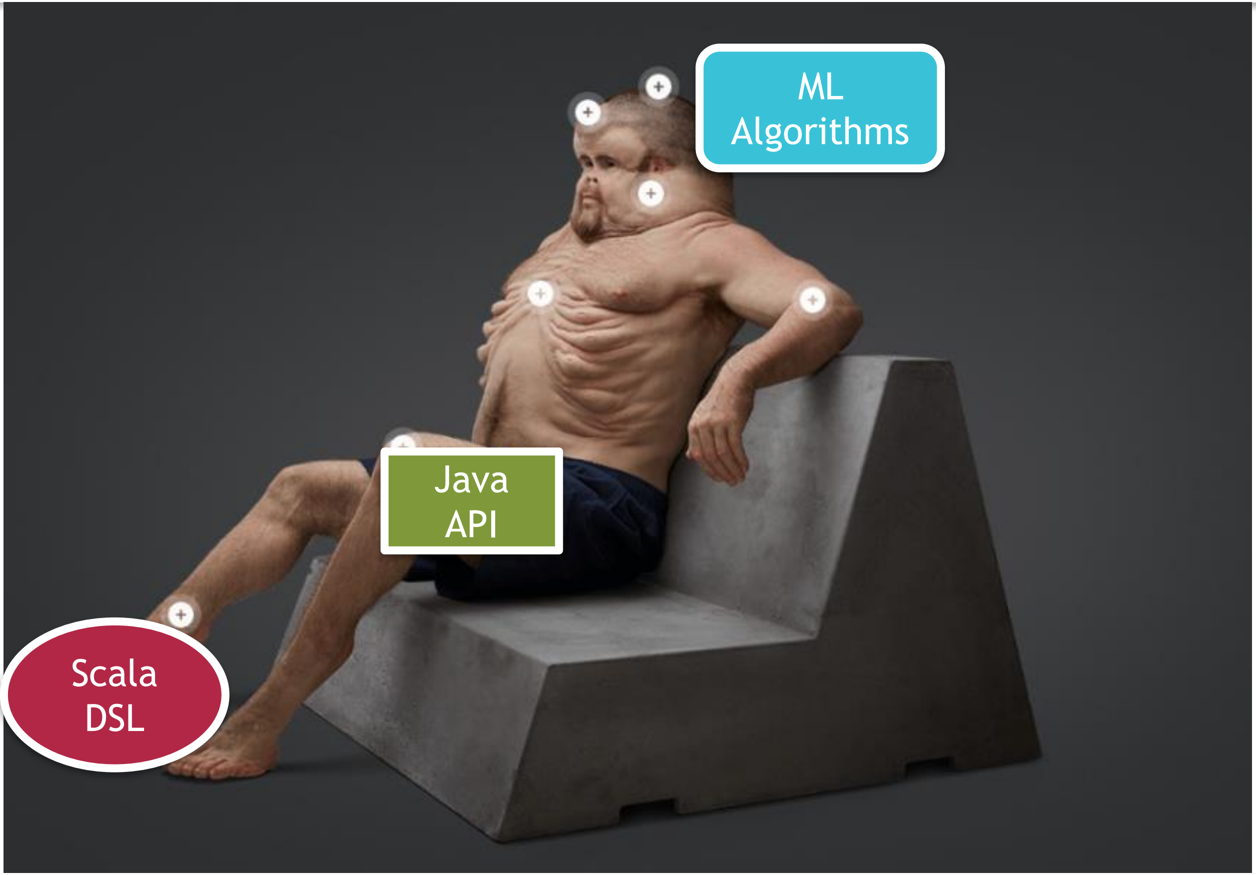
Cross-validation



MLlib Demo

MLLIB VS ALL

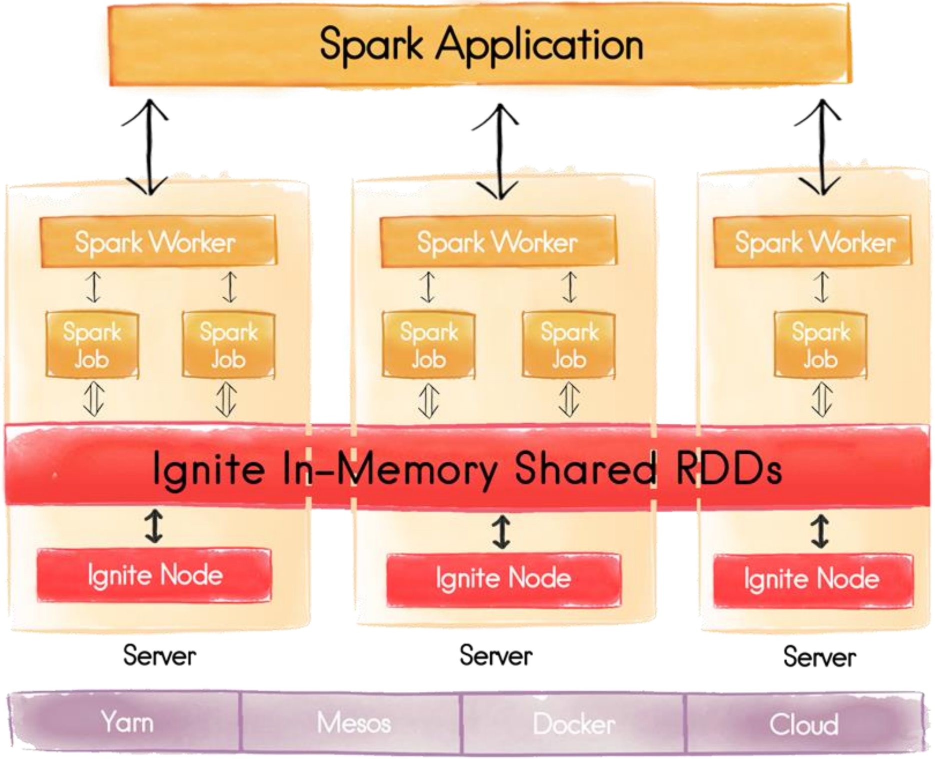
Mahout



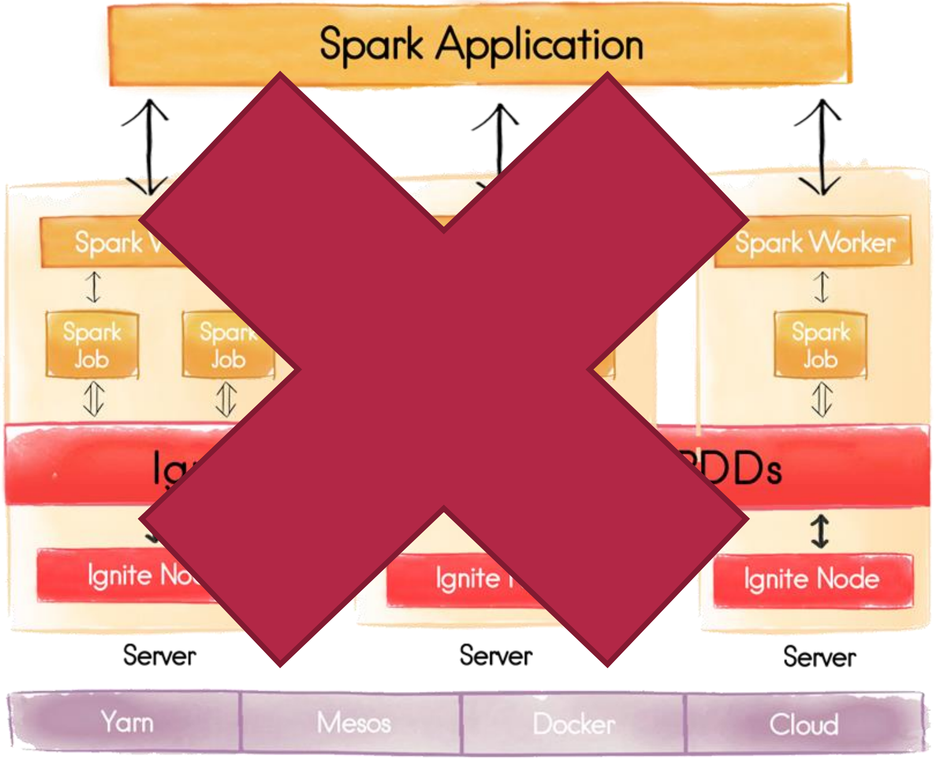
Integration issues

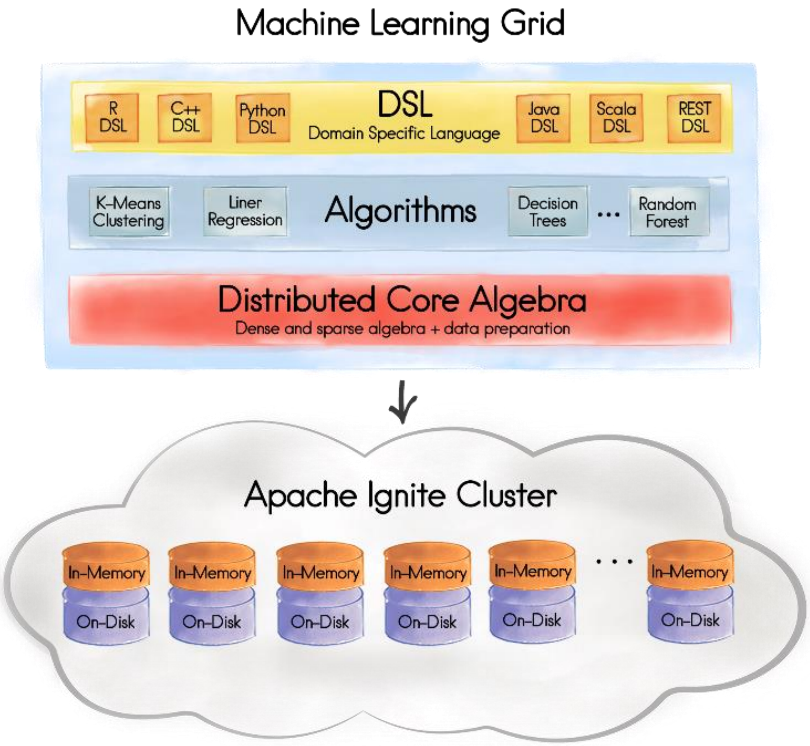
Flink ML

Ignite ML


Spark Application
Spark Worker
Spark Job
Spark Job
Spark Worker
Spark Job
Spark Job
Spark Worker
Spark Job
Ignite In-Memory Shared RDDs
Ignite Node
Ignite Node
Ignite Node
Server
Server
Server
Yarn
Mesos
Docker
Cloud

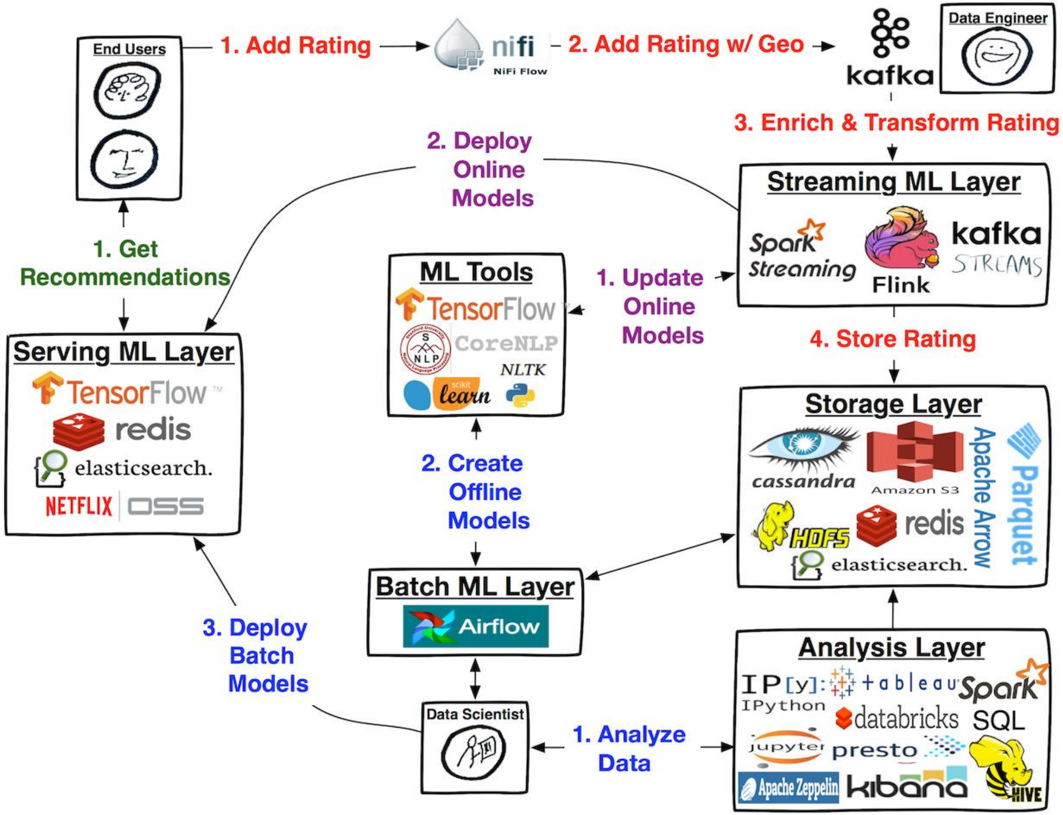
Ignite ML


Spark Application
Spark W
Spark Worker
Spark Job
Spark Job
Spark Job
Ig
DDs
Ignite No
Ignite
Ignite Node
Server
Server
Server
Yarn
Mesos
Docker
Cloud


Machine Learning Grid
R DSL
C++ DSL
Python DSL
DSL
Domain Specific Language
Java DSL
Scala DSL
REST DSL
K-Means Clustering
Liner Regression
Algorithms
Decision Trees
...
Random Forest
Distributed Core Algebra
Dense and sparse algebra + data preparation
Apache Ignite Cluster
In-Memory
On-Disk
In-Memory
On-Disk
In-Memory
On-Disk
In-Memory
On-Disk
In-Memory
On-Disk
...
In-Memory
On-Disk

IN CONCLUSION

Reality



End Users
1. Add Rating
nifi
NiFi Flow
2. Add Rating w/ Geo
kafka
Data Engineer
3. Enrich & Transform Rating
2. Deploy Online Models
Streaming ML Layer
Spark Streaming
Flink
kafka STREAMS
1. Get Recommendations
ML Tools
TensorFlow
CoreNLP
NLTK
scikit learn
1. Update Online Models
4. Store Rating
Serving ML Layer
TensorFlow
redis
elasticsearch.
NETFLIX OSS
Storage Layer
cassandra
Amazon S3
Apache Arrow
Parquet
HDFS
redis
elasticsearch.
2. Create Offline Models
Batch ML Layer
Airflow
3. Deploy Batch Models
Analysis Layer
IP[y]: IPython
tableau
Spark
databricks
SQL
jupyter
presto
Apache Zeppelin
kibana
HIVE
Data Scientist
1. Analyze Data

Contacts

E-mail : Alexey_Zinovyev@epam.com

Twitter : @zaleslaw @BigDataRussia

vk.com/big_data_russia **Big Data Russia**

+ Telegram @bigdatarussia

vk.com/java_jvm **Java & JVM langs**

+ Telegram @javajvmlangs

APACHE Spark™

Any questions?