

Cloud Computing and Big Data

Machine Learning

Oxford University
Software Engineering
Programme
Nov 2015



© Paul Fremantle 2015. Licensed under the Creative Commons 4.0 BY-SA (Attribution-Sharealike) license.
See <http://creativecommons.org/licenses/by-sa/4.0/>

Contents

- Definitions and terminology
- The overall process
- Main techniques
- Algorithms and examples
- Big Data Machine Learning
- R and PMML
- Spark MLlib
- Introduction to the lab



Definition of Machine Learning

- Algorithms that can learn from data



Definition of Machine Learning

- Algorithms that can learn from data

Ok that was a circular definition 😊



Definition take 2

- Algorithms that can analyse a set of data and then make predictions when new data comes in
 - The more data that they analyse, the better the predictions



Learning phase



Usage phase



Terminology

- **Sample**
 - Some incoming data to be analysed
 - E.g. a JPG picture
- **Feature**
 - Some quantifiable data from the sample
 - E.g. colour, height, width, pixel data, etc
- **Label**
 - Some useful information about the sample that we wish to categorise:
 - E.g. looking at a picture this is a person
- **Model**
 - The output of some learning algorithm
 - The parameterization of an algorithm that can be run against new data



Types of learning

- Supervised
 - The required labels are known
 - Aiming to find an algorithm that correctly identifies these
 - Iterative exploration and refinement
- Unsupervised
 - The labels are not known
 - The system identifies new classifications



Types of machine learning

- Classification
- Regression / Prediction
- Clustering
- Recommendation and Collaborative Filtering
- Frequent Pattern mining



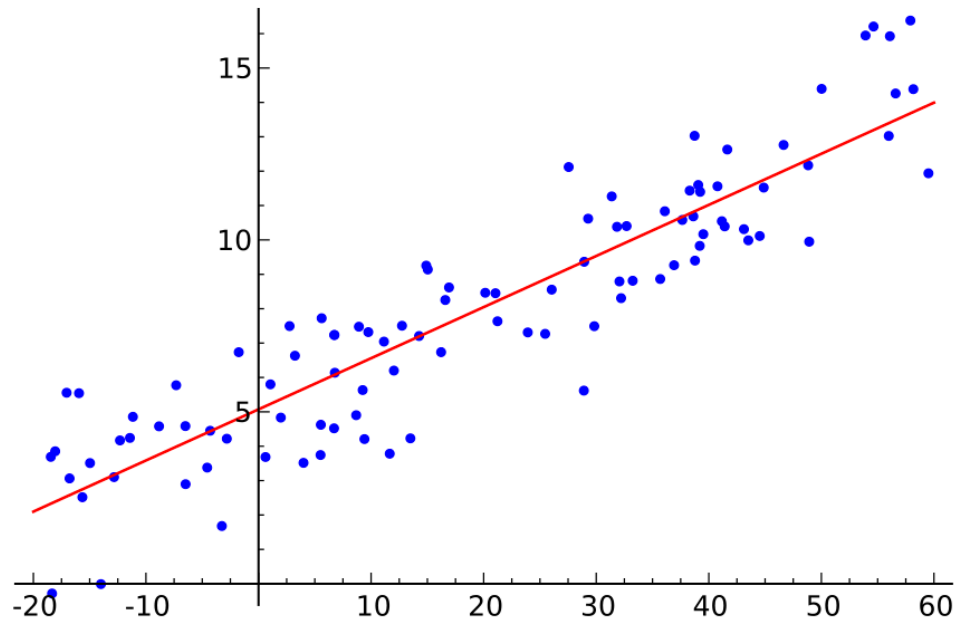
Classification

- Identifying a class into which this sample fits
- E.g. look at a picture and decide this is a piece of fruit
 - Google Photos demo



Regression

- Applying a model based on previous data
 - Allows prediction of future state
- Many statistical techniques

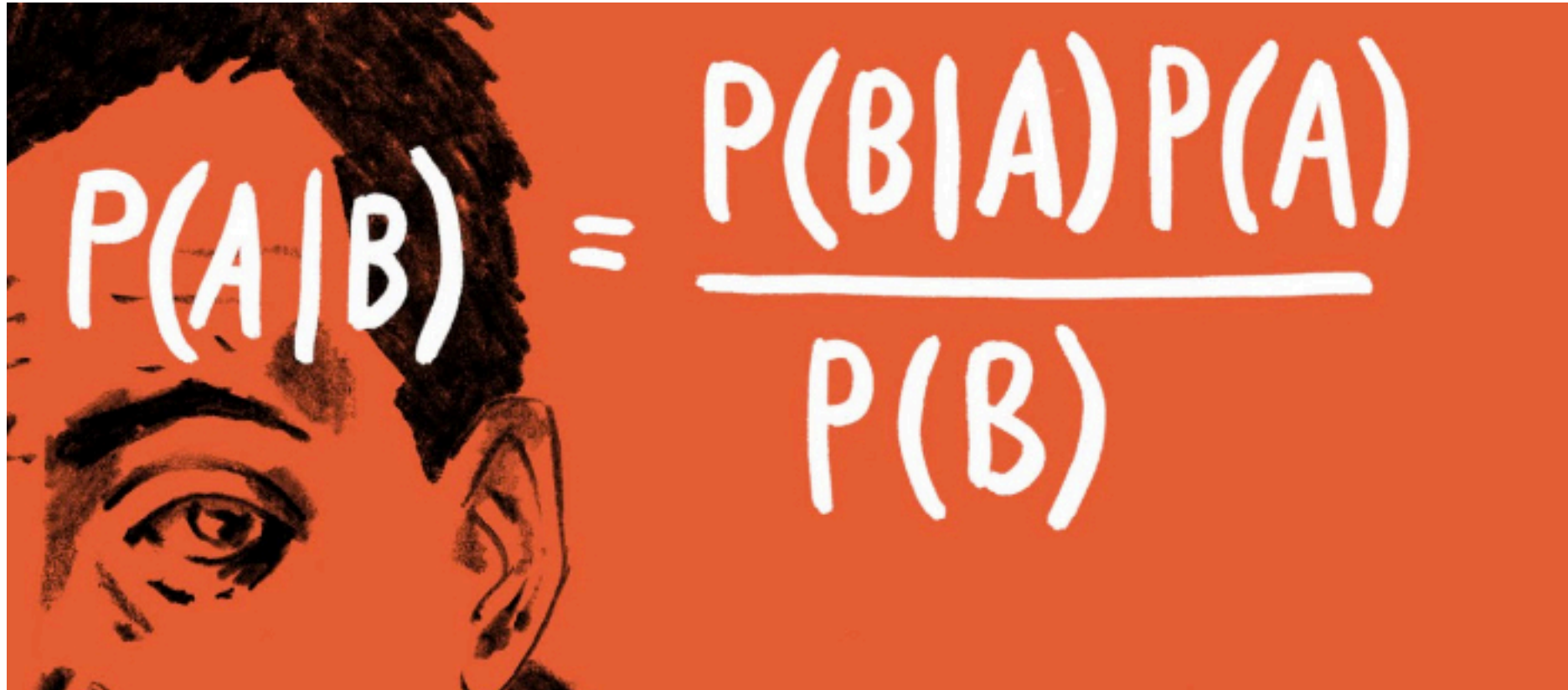


Regression vs Classification

- Regression produces a real number or numbers
 - i.e. a continuously varying answer or answers
- Classification identifies a set or element of a set
 - E.g. False, Blue, Person, High-Value Customer



Bayes Theorem


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

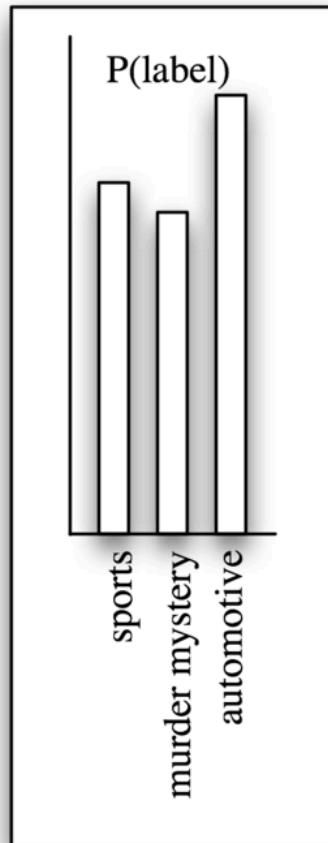


© Paul Fremantle 2015. Licensed under the Creative Commons 4.0 BY-SA (Attribution-Sharealike) license.
See <http://creativecommons.org/licenses/by-sa/4.0/>

Classification Algorithms

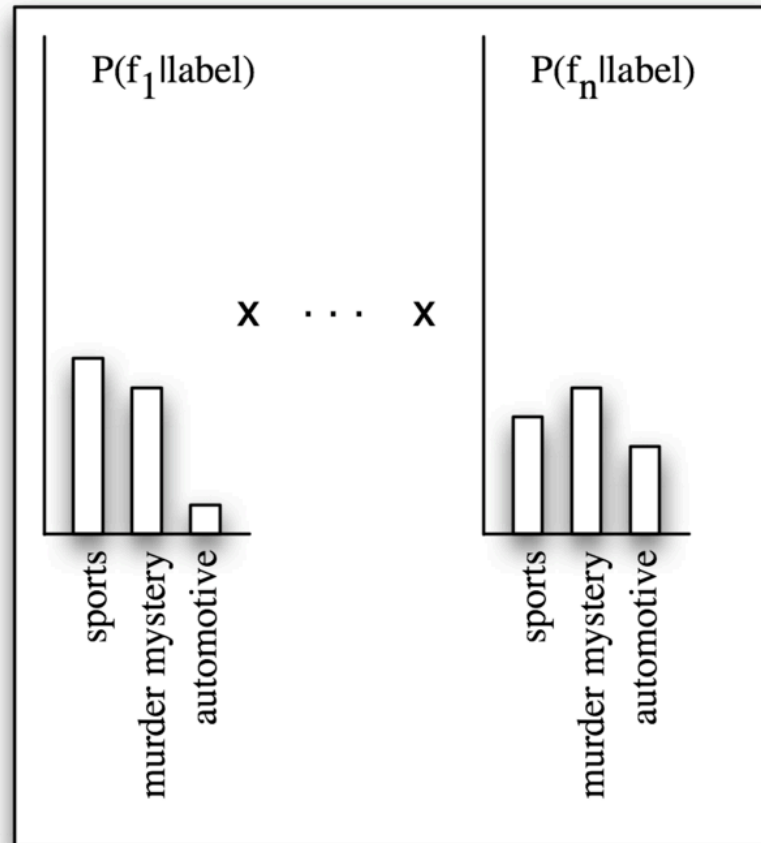
Naïve Bayes

Prior Probabilities



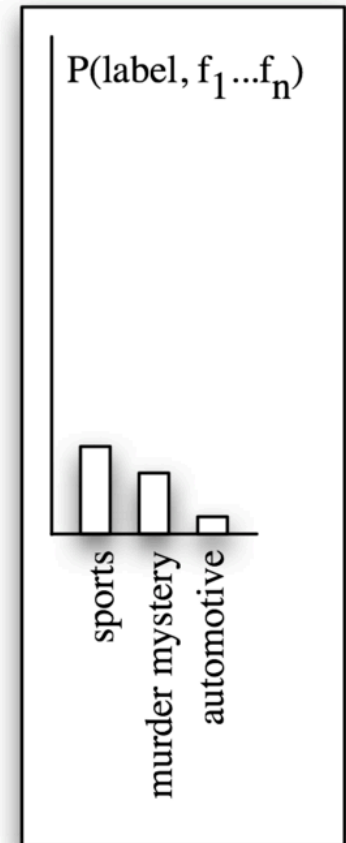
x

Feature Contributions



=

Label Likelihoods



Spark MLlib's algorithms

Problem Type	Supported Methods
Binary Classification	linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes
Multiclass Classification	decision trees, random forests, naive Bayes
Regression	linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression



© Paul Fremantle 2015. Licensed under the Creative Commons 4.0 BY-SA (Attribution-Sharealike) license.
See <http://creativecommons.org/licenses/by-sa/4.0/>

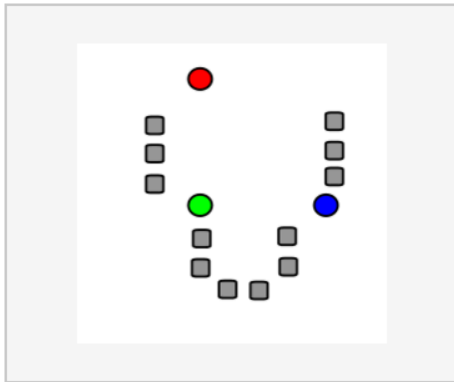
Clustering

- Grouping items into clusters
 - Where items in a cluster are more similar to each other than to items in other clusters
- Basically creating the classifications from the data rather than applying them a priori

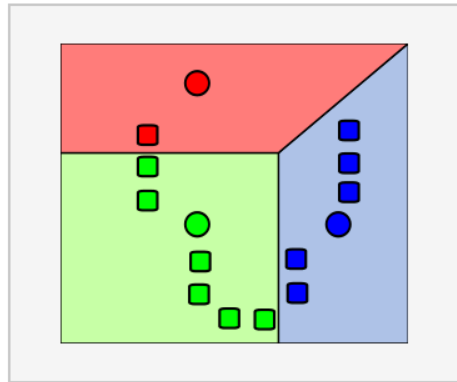


K-Means Clustering

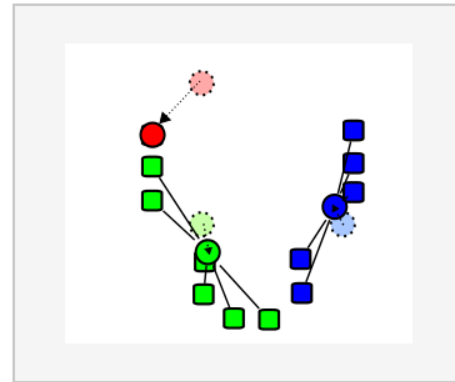
Demonstration of the standard algorithm



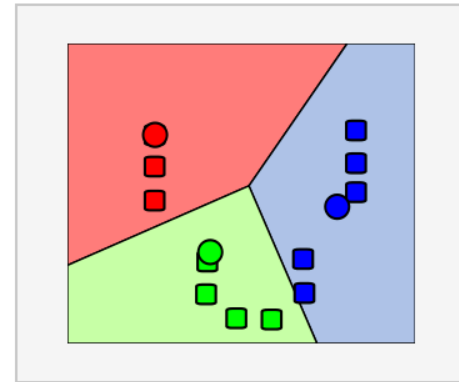
1. k initial "means" (in this case $k=3$) are randomly generated within the data domain (shown in color).



2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the [Voronoi diagram](#) generated by the means.



3. The [centroid](#) of each of the k clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

MLLib's clustering

- K-means
- Gaussian mixture
- Power iteration clustering (PIC)
- Latent Dirichlet allocation (LDA)
- Streaming k-means



Recommendation and Collaborative Filtering

- Given a user's interaction with items, what else are they likely to prefer

Large-scale Parallel Collaborative Filtering for the Netflix Prize

Yunhong Zhou, Dennis Wilkinson, Robert Schreiber and Rong Pan


HP Labs, 1501 Page Mill Rd, Palo Alto, CA, 94304

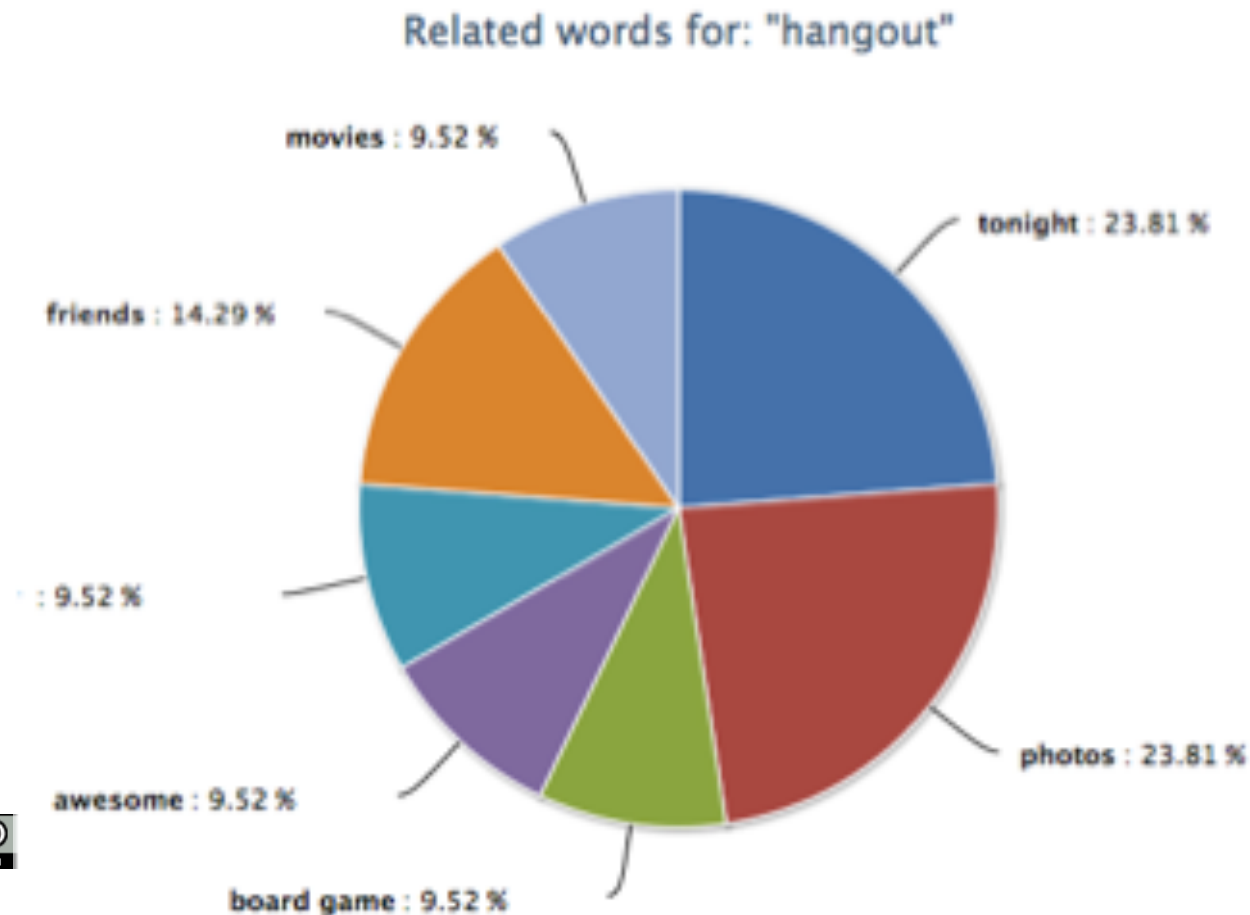
{yunhong.zhou, dennis.wilkinson, rob.schreiber, rong.pan}@hp.com

Abstract. Many recommendation systems suggest items to users by utilizing the techniques of collaborative filtering (CF) based on historical records of items that the users have viewed, purchased, or rated. Two major problems that most CF approaches have to resolve are scalability and sparseness of the user profiles. In this paper, we describe *Alternating-Least-Squares with Weighted- λ -Regularization* (ALS-WR), a parallel algorithm that we designed for the Netflix Prize, a large-scale collaborative filtering challenge. We use parallel Matlab on a Linux cluster



Frequent Pattern Mining

Related topics:  @cassiomelo Your last post was about a hangout. These are the topics you relate to hangout: [tonight](#), [movies](#), [board game](#), [friends](#), [awesome](#), [photos](#), [bar](#) and [NBA](#).



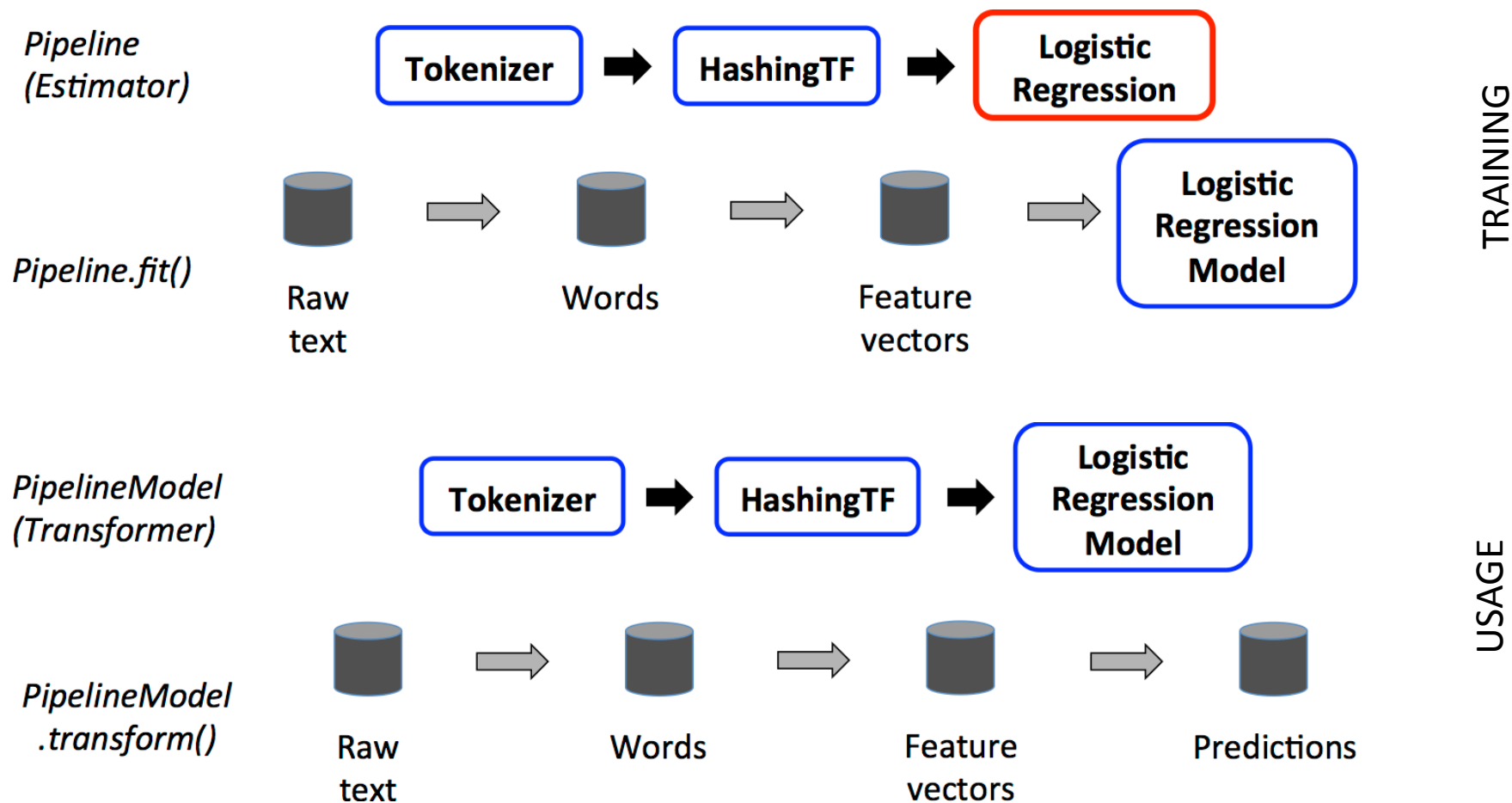
ke) license.

MLLib FPM

- Frequent pattern mining
 - FP-growth
 - association rules
 - PrefixSpan



Spark MLLib Pipelines



Big Data ML

- Obviously we can learn more insights with more data
- Many examples
 - Netflix competition
 - Google, Facebook, Twitter etc are all doing big data ML
- Obviously we want the right algorithms:
 - E.g. Kmeans++ is a parallelizable version of Kmeans
- MLLib and Mahout come pre-built with these



Recap

- Machine Learning is a powerful way of gaining insight and value from big data
 - Recommendation
 - Classification and prediction
 - Clustering and understanding
 - Etc



Questions?



© Paul Fremantle 2015. Licensed under the Creative Commons 4.0 BY-SA (Attribution-Sharealike) license.
See <http://creativecommons.org/licenses/by-sa/4.0/>