

Building, Debugging, & Tuning Spark ML Pipelines

Joseph K. Bradley

June 2015

Spark Summit

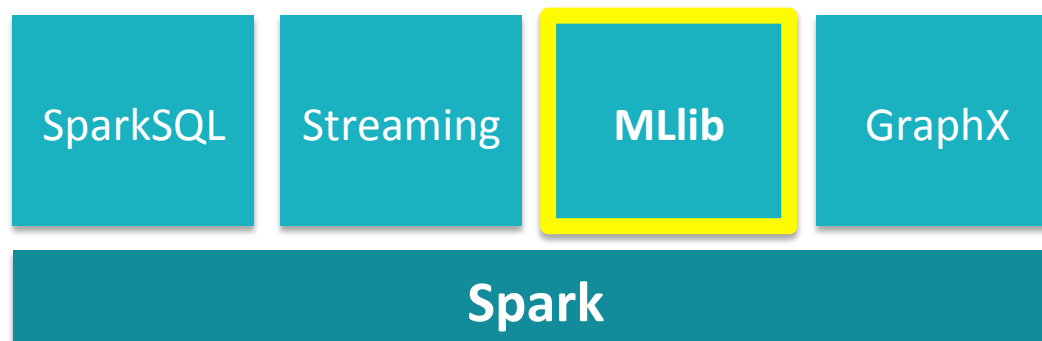


Who am I?

- Joseph K. Bradley
- Ph.D. in Machine Learning from CMU, postdoc at Berkeley
- Apache Spark committer
- Software Engineer @ Databricks Inc.

About Spark MLlib

- Started at Berkeley (Spark 0.8)



- Now (Spark 1.4)
 - Contributions from 50+ orgs, 100+ individuals
 - Growing coverage of *distributed* algorithms

Algorithm Coverage

- **Classification**

- Logistic regression
- Naive Bayes
- Streaming logistic regression
- Linear SVMs
- Decision trees
- Random forests
- Gradient-boosted trees

- **Regression**

- Ordinary least squares
- Ridge regression
- Lasso
- Isotonic regression
- Decision trees
- Random forests
- Gradient-boosted trees
- Streaming linear methods

- **Frequent itemsets**

- FP-growth

Recommendation

- Alternating Least Squares

Feature extraction & selection

- Word2Vec
- Chi-Squared selection
- Hashing term frequency
- Inverse document frequency
- Normalizer
- Standard scaler
- Tokenizer
- One-Hot Encoder
- StringIndexer
- VectorIndexer
- VectorAssembler
- Binarizer
- Bucketizer
- ElementwiseProduct
- PolynomialExpansion

Model import/export

Pipelines

Clustering

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

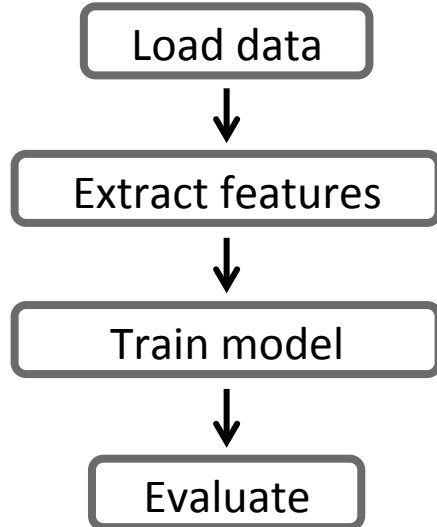
Statistics

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

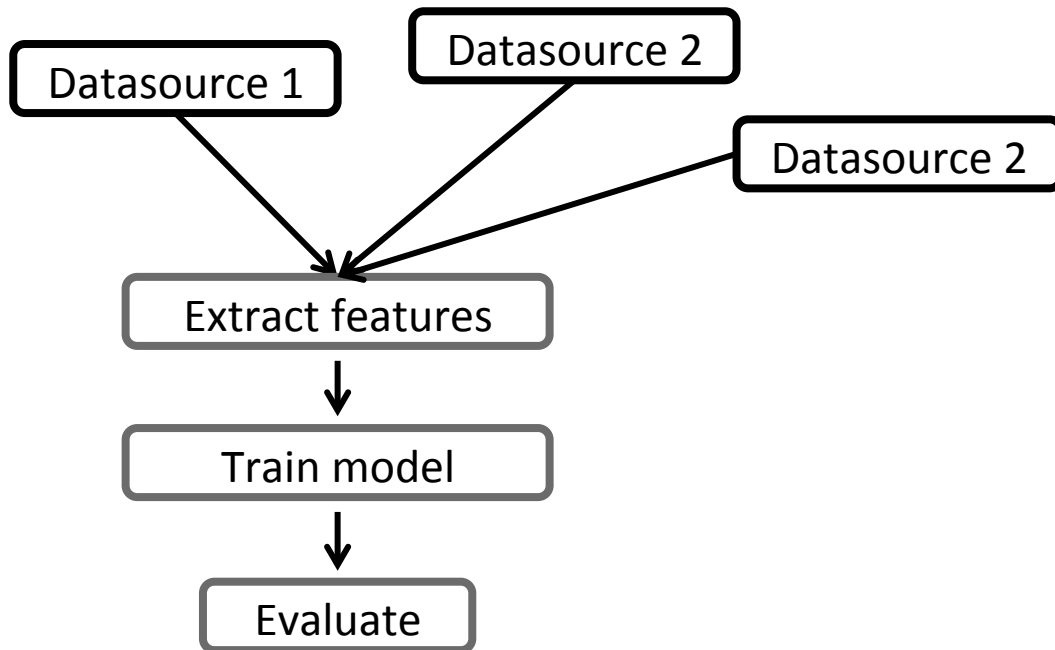
Linear algebra

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

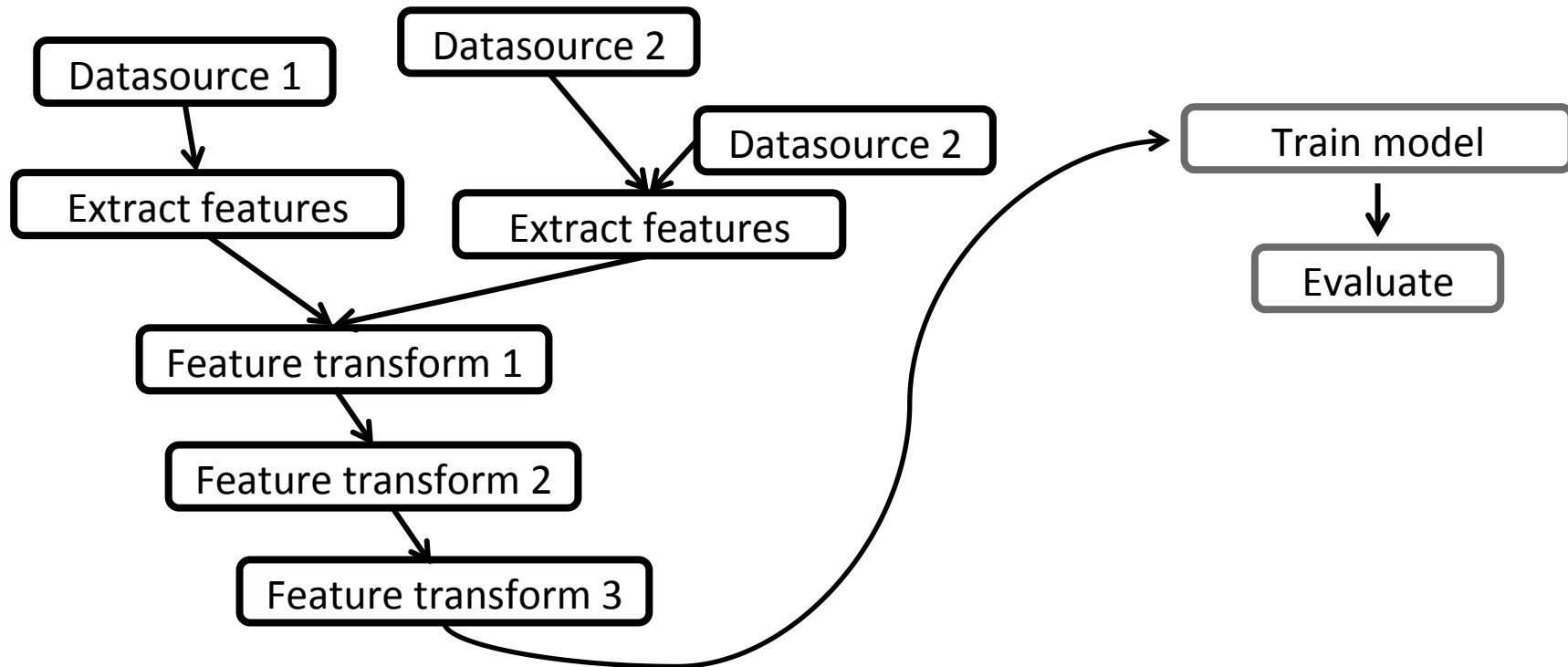
ML Workflows are complex



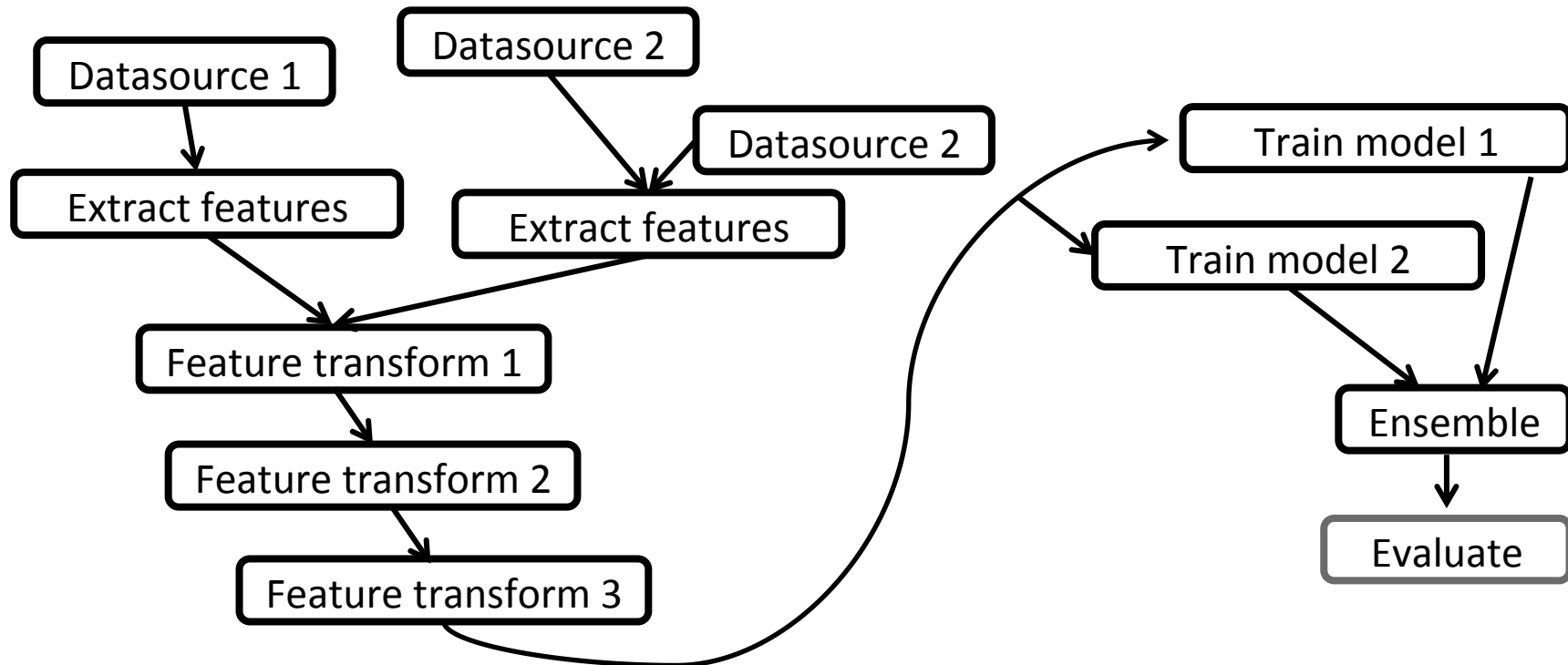
ML Workflows are complex



ML Workflows are complex



ML Workflows are complex



ML Workflows are complex

- Specify pipeline
- Inspect & debug
- Re-run on new data
- Tune parameters

Example: Text Classification

- Goal: Given a text document, predict its topic.

Features

Subject: Re: Lexan Polish?
Suggest McQuires #1 plastic
polish. It will help somewhat
but nothing will remove deep
scratches without making it
worse than it already is.
McQuires will do something...



Label

1: about science
0: not about science

Dataset: "20 Newsgroups"
From UCI KDD Archive

DataFrames

dept	age	name
Bio	48	H Smith
CS	54	A Turing
Bio	43	B Jones
Chem	61	M Kennedy

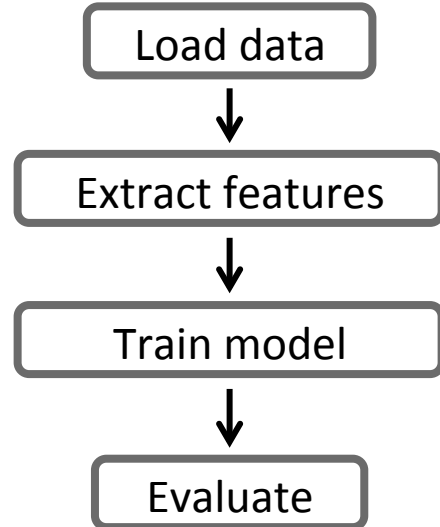
Data grouped into
named columns

DSL for common tasks

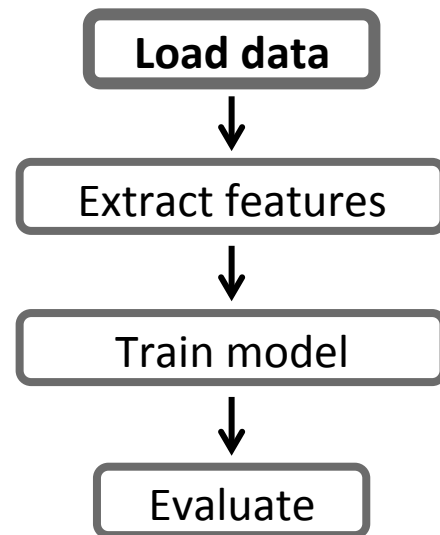
- Project, filter, aggregate, join, ...
- Metadata
- UDFs

`data.groupBy("dept").avg("age")`

ML Workflow



Load Data



Data sources for DataFrames

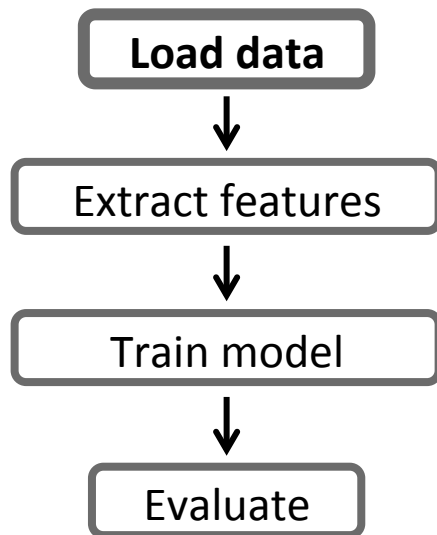
built-in



external



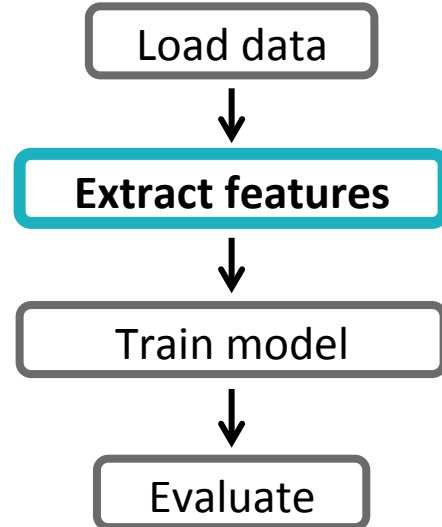
Load Data



Current data schema

```
label: Int  
text: String
```

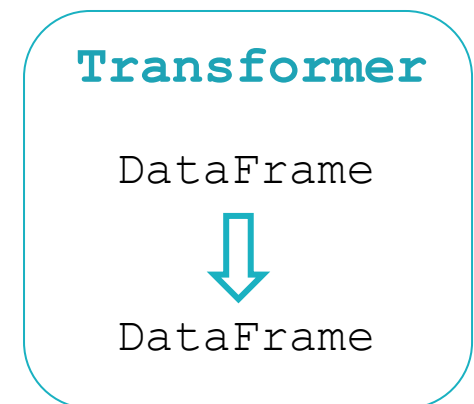
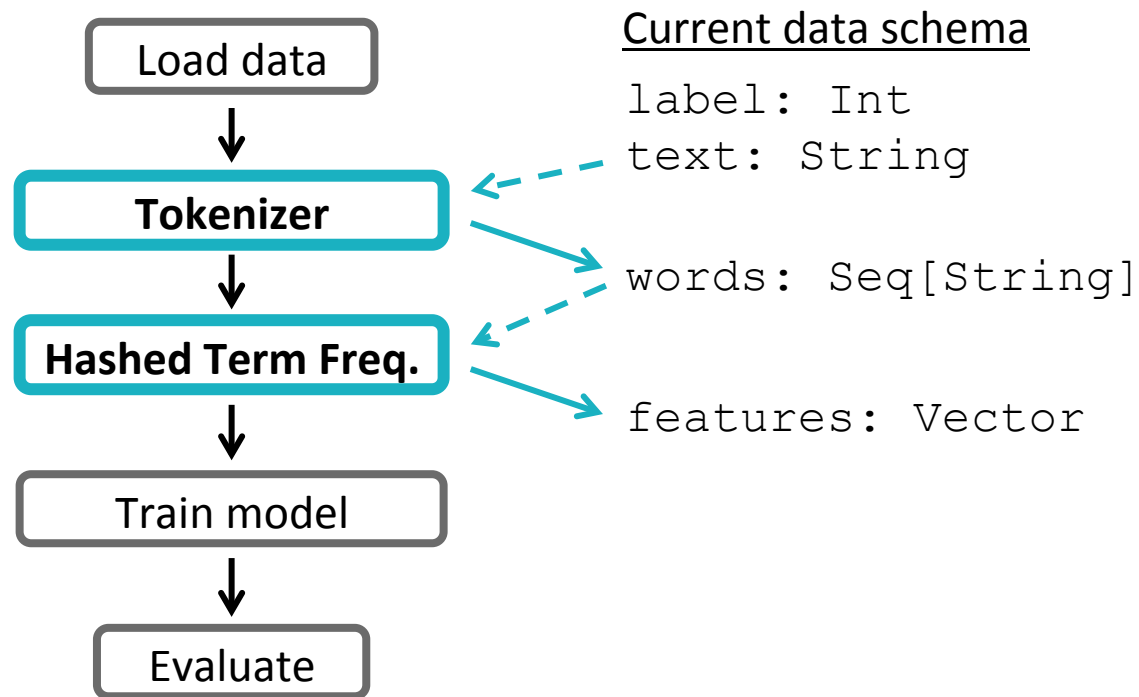
Extract Features



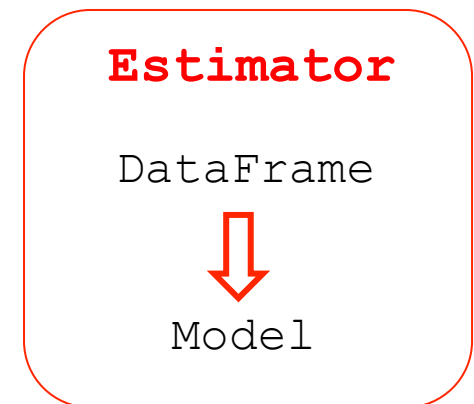
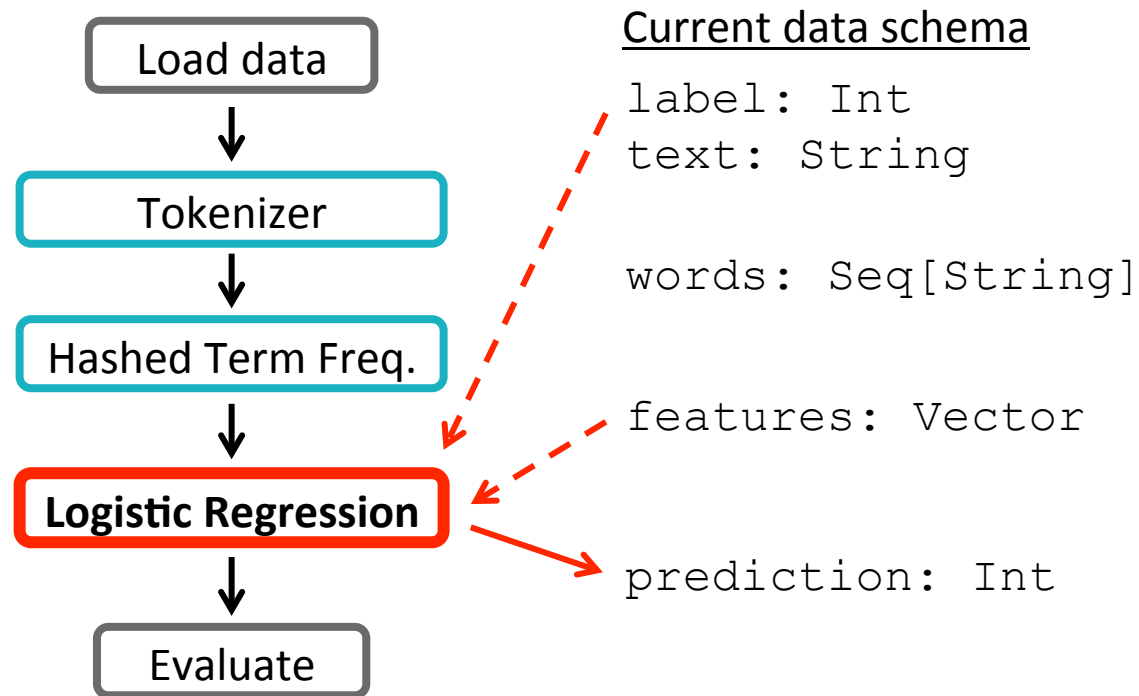
Current data schema

```
label: Int  
text: String
```

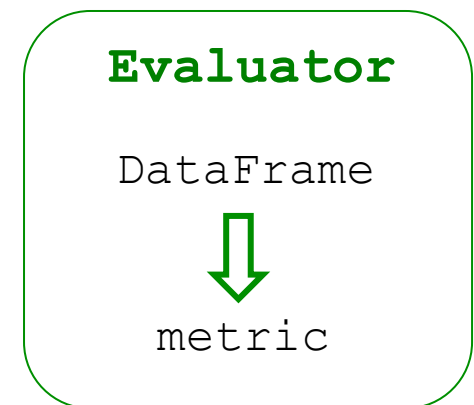
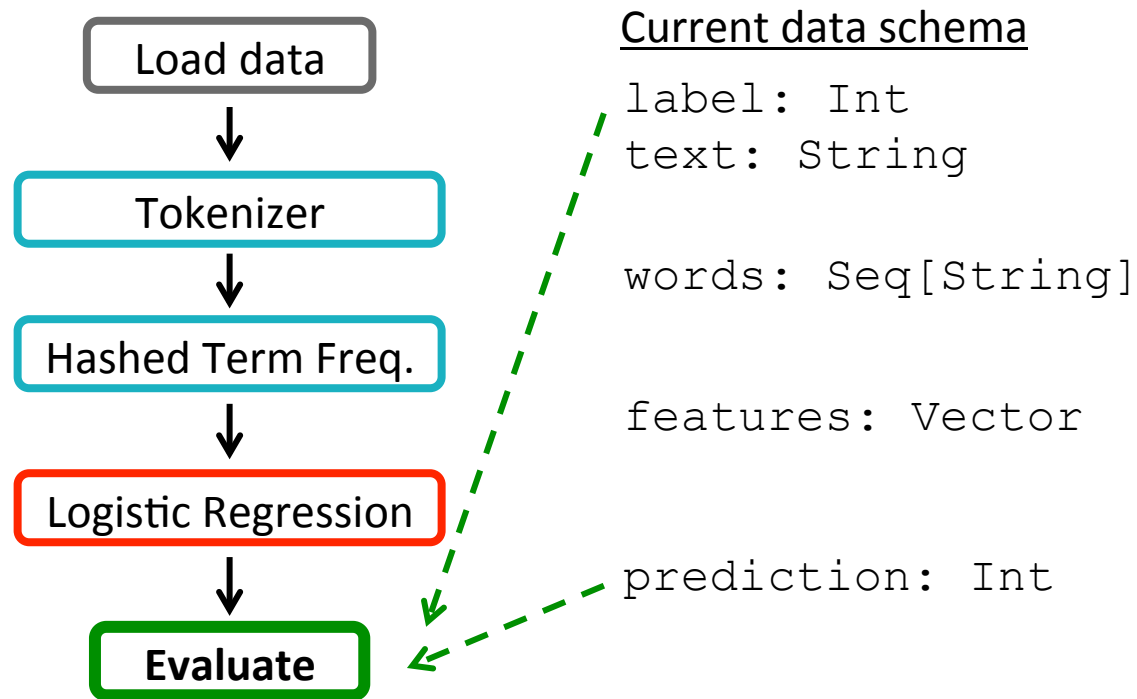
Extract Features



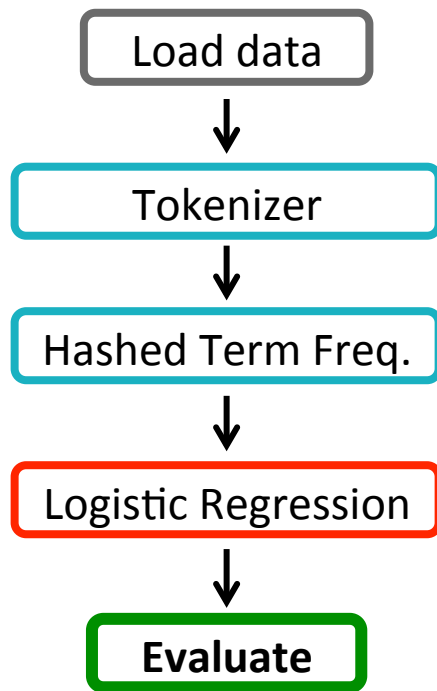
Train a Model



Evaluate the Model



Data Flow



Current data schema

label: Int
text: String

words: Seq[String]

features: Vector

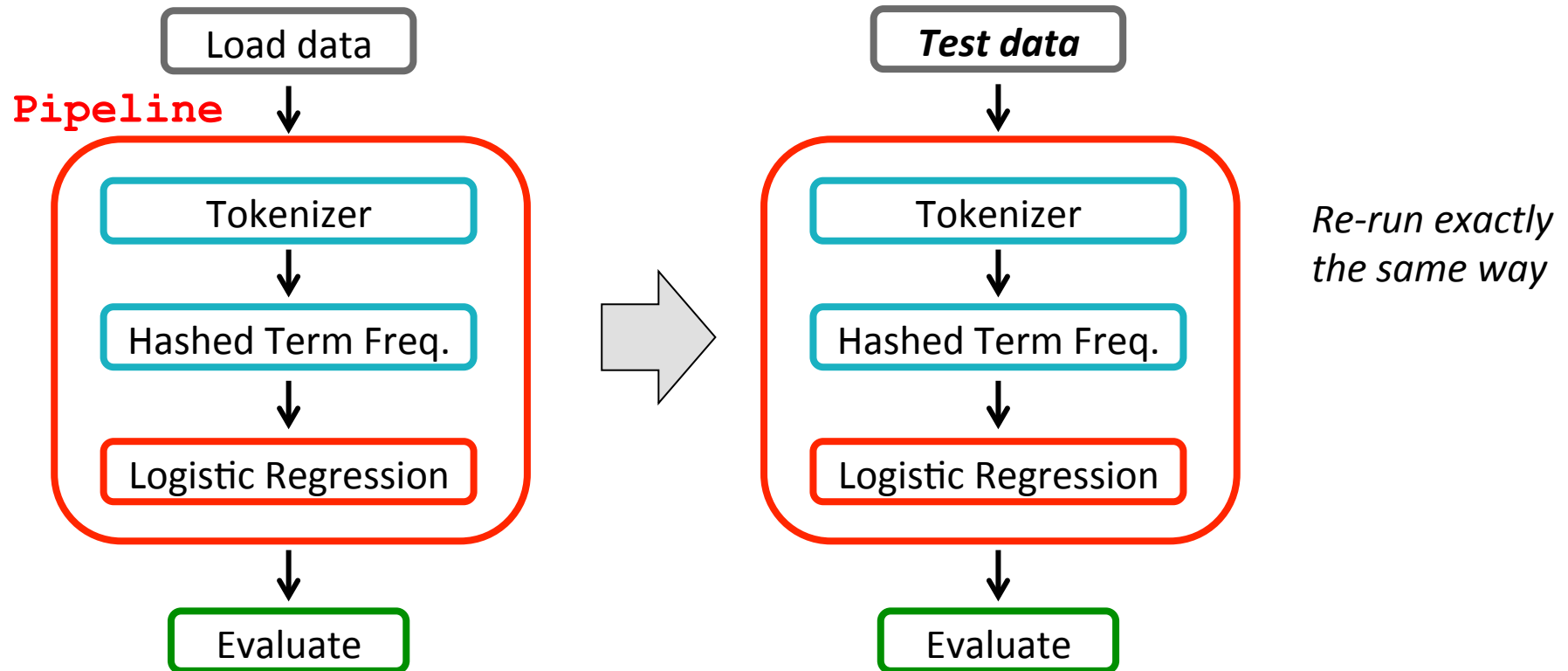
prediction: Int

By default, always
append new columns

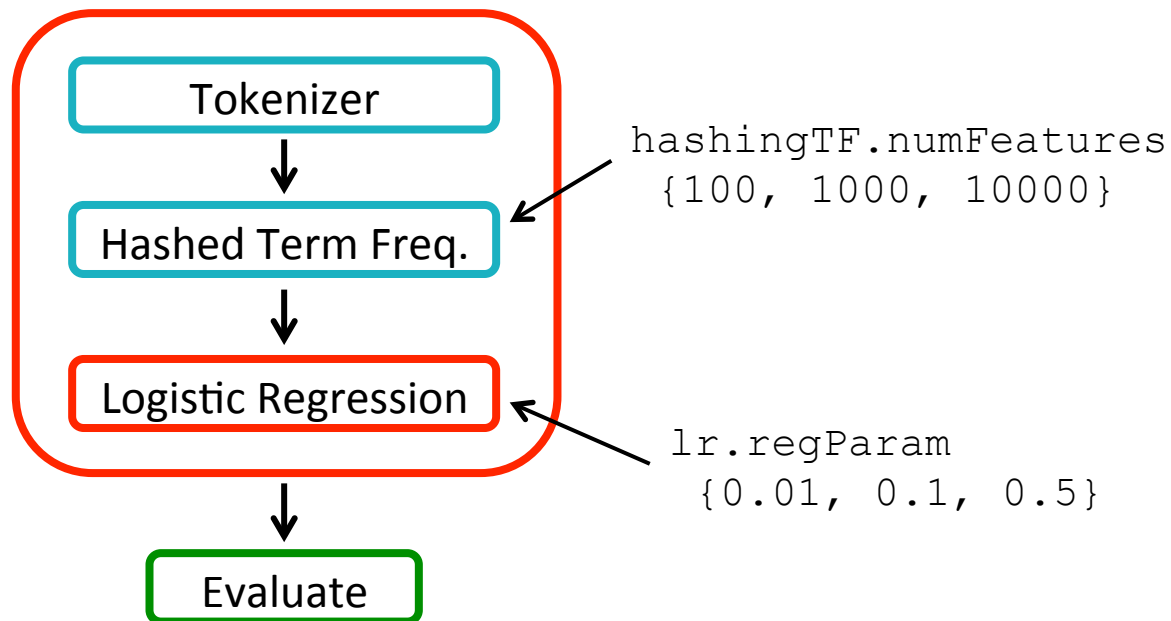
→ Can go back & inspect
intermediate results

→ Made efficient by
DataFrames

ML Pipelines



Parameter Tuning



CrossValidator

Given:

- Estimator
- Parameter grid
- Evaluator

Find best parameters

Demo: ML Pipelines *in Databricks Cloud*

Recap

- ML Pipelines provide:
 - Integration with DataFrames
 - Familiar API based on scikit-learn
 - Easy workflow inspection
 - Simple parameter tuning

User-defined Pipeline components:
Estimators, Transformers, & Evaluators

Schema validation

Complex Pipelines:
compositions & DAGs

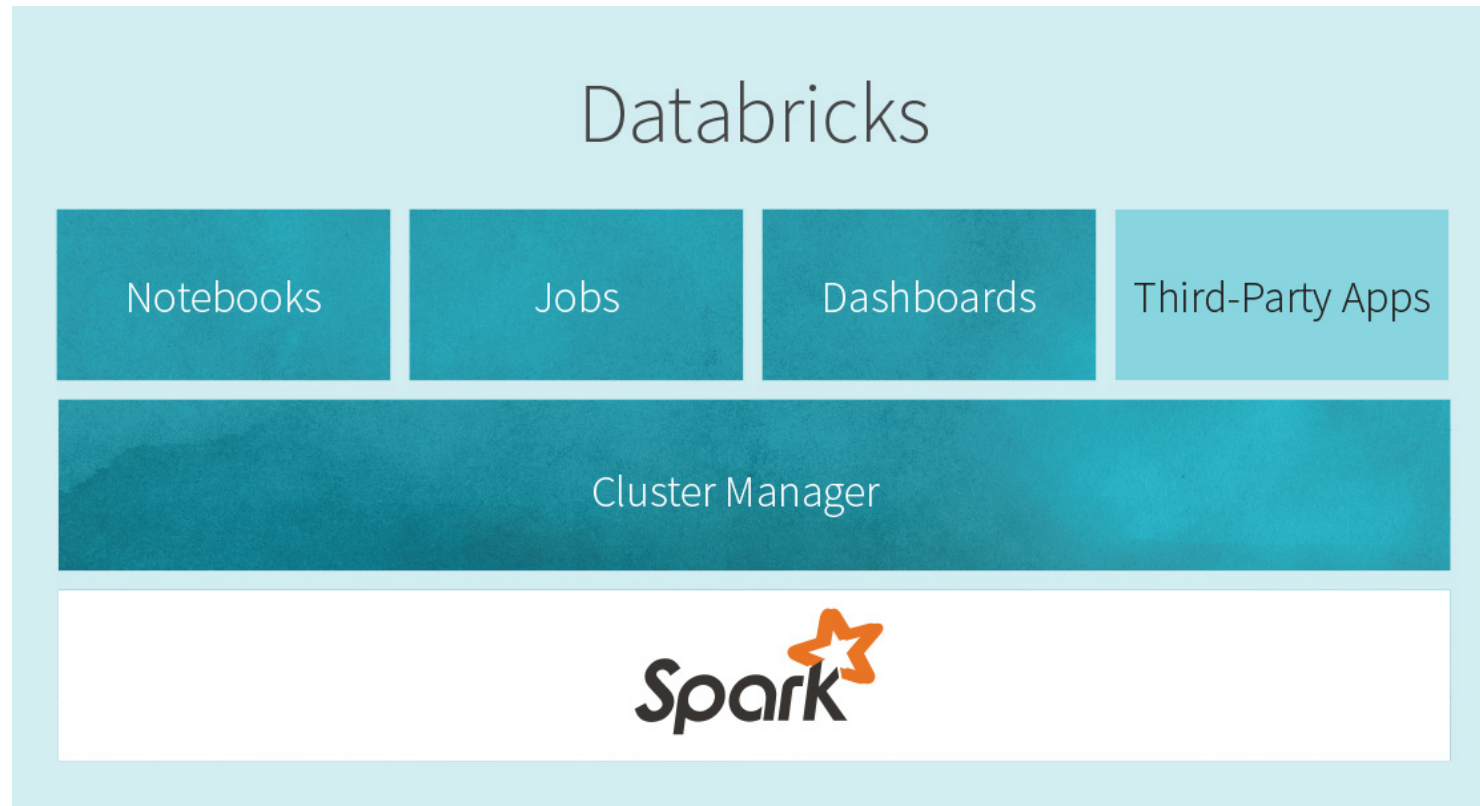
Support for complex types
(via DataFrames & UDTs)

In Spark 1.4

- Pipelines graduating from alpha!
- *Many* feature transformers
- More complete Python API
- Spark R
- PMML model export
- More algorithms (OneVsRest, OnlineLDA, ...)

Looking Ahead

- MLlib API for SparkR
- Improved model inspection
- Pipeline optimizations



Databricks, Inc.: *Founded by the creators of Spark & remains largest contributor*

Spark documentation

spark.apache.org

Pipelines blog post

databricks.com/blog/2015/01/07

DataFrames blog post

databricks.com/blog/2015/02/17

Databricks

databricks.com/product

Spark ML edX MOOC

[ML with Spark](#)

Spark Packages

spark-packages.org

Thank you!

databricks.com/company/careers

