livelessons
video instruction from technology experts

Lesson 5: Advanced Applications

5.1 Machine Learning on Spark: MLlib and `spark.ml`

At Scale

PySpark

HDFS

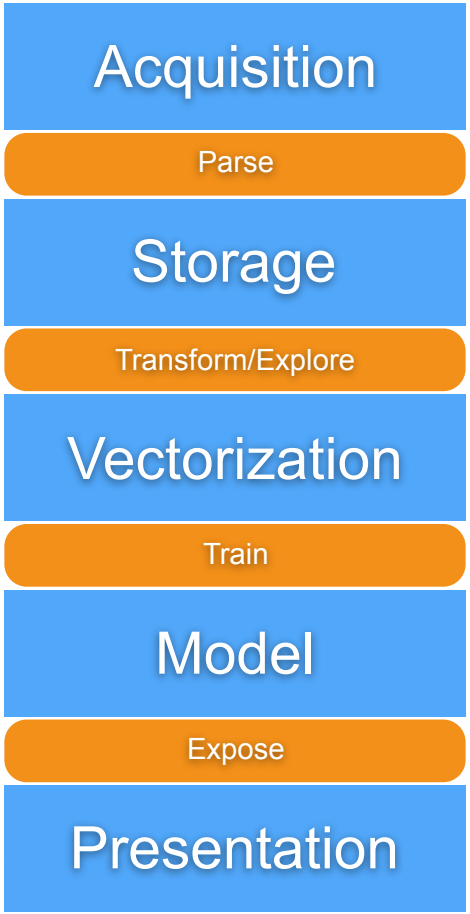
Dataframes/Spark SQL

MLlib/spark.ml

model.save()

Spark Streaming

Data Pipeline



At Scale

PySpark

HDFS

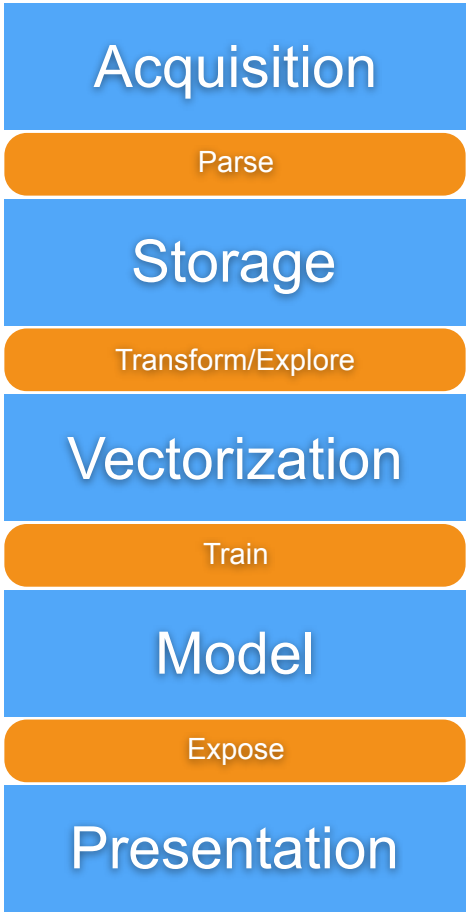
Dataframes/Spark SQL

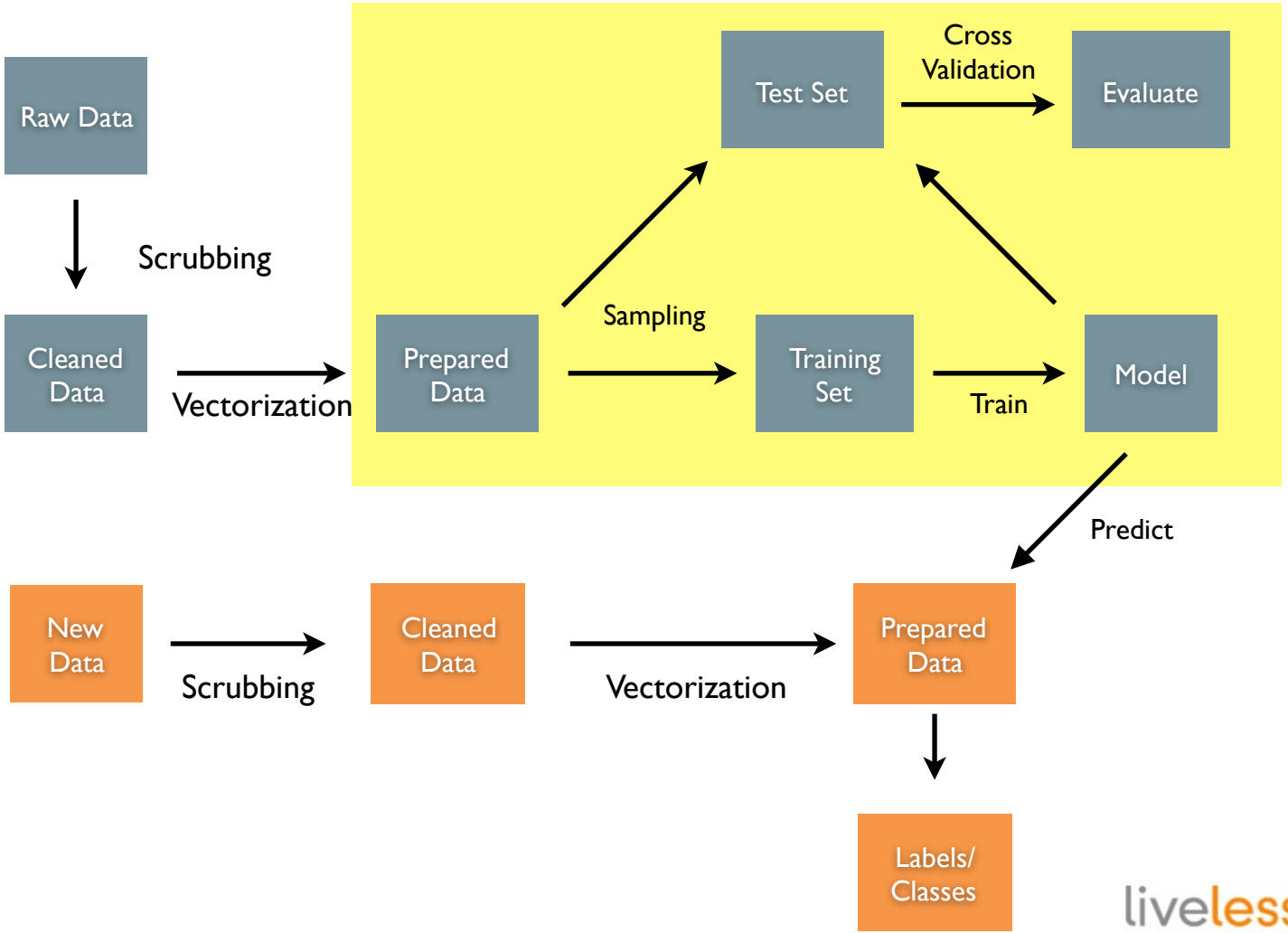
MLlib/spark.ml

model.save()

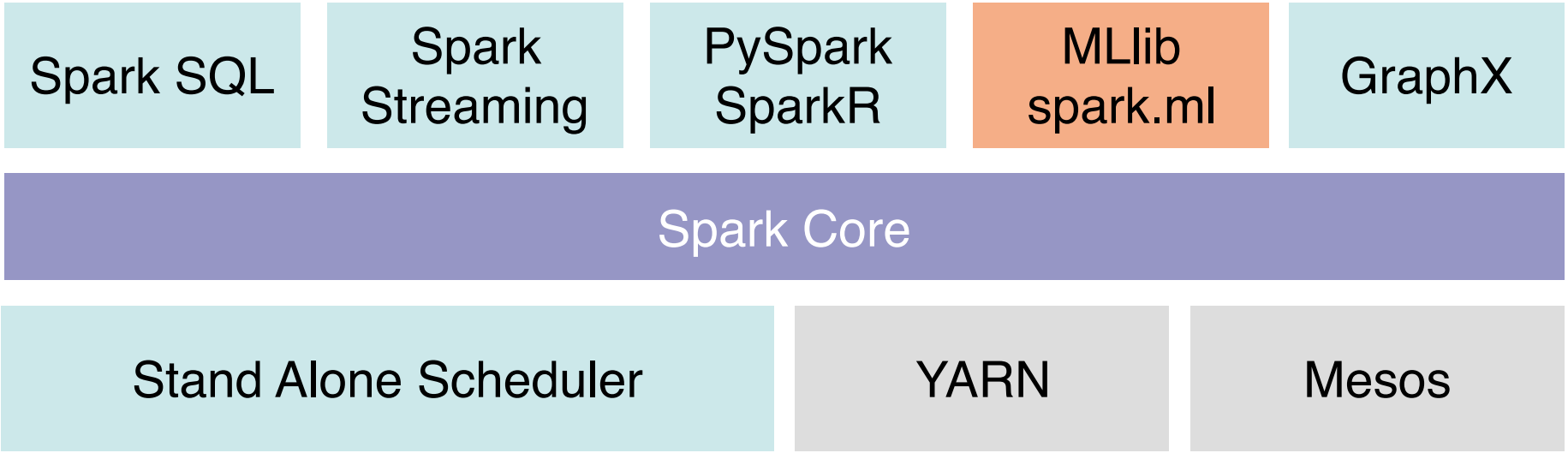
Spark Streaming

Data Pipeline





Unified Platform



Unified Platform

MLlib	`spark.ml`
original primary ML API	*Higher level API to construct ML workflows/ pipelines*
Library	*Framework*
Statistics, Classification/Regression, Clustering, Feature Extraction, etc.	*Dataset, Transformer, Estimator, Pipeline, Parameter*

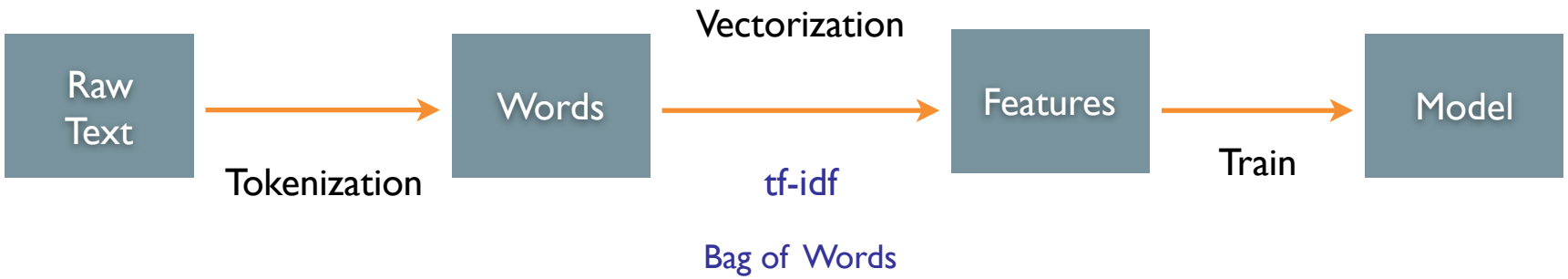
MLlib

- Data types
- Basic Statistics
- Classification and Regression
- Collaborative Filtering
- Clustering
- Dimensionality Reduction
- Feature Extraction
- Optimization
- Model Export

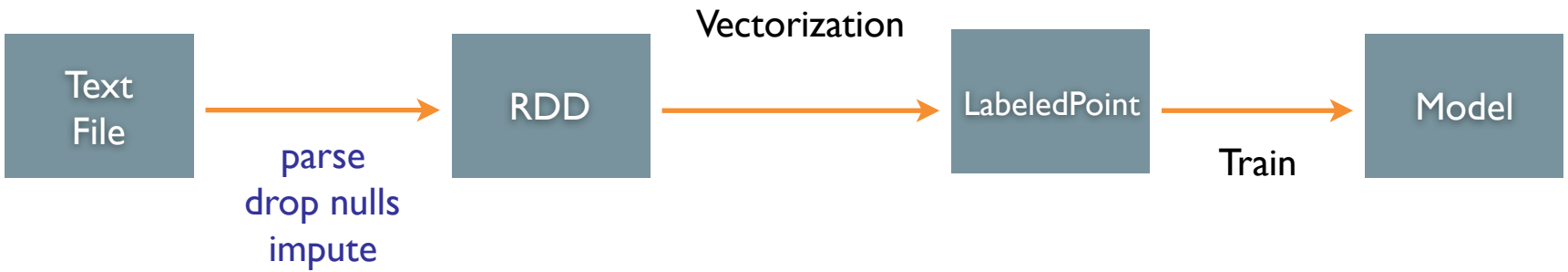
MLlib

- Data types
- Basic Statistics
- Classification and Regression
- Collaborative Filtering
- Clustering
- Dimensionality Reduction
- Feature Extraction
- Optimization
- Model Export

Machine Learning Pipeline



Machine Learning Pipeline



2009 KDD Cup: Customer Relation Prediction

- French telecom company Orange
- Optimize customer relationship:
    - Churn: switching providers
    - Appetency: buying new products and services
    - Up-selling: buying upgrades/add-ons for current sale

Challenges

- “Very large dataset”
- Heterogeneous (continuous and discrete) noisy data
- Imbalanced classes
- (no context on data — anonymized)