livelessons
video instruction from technology experts

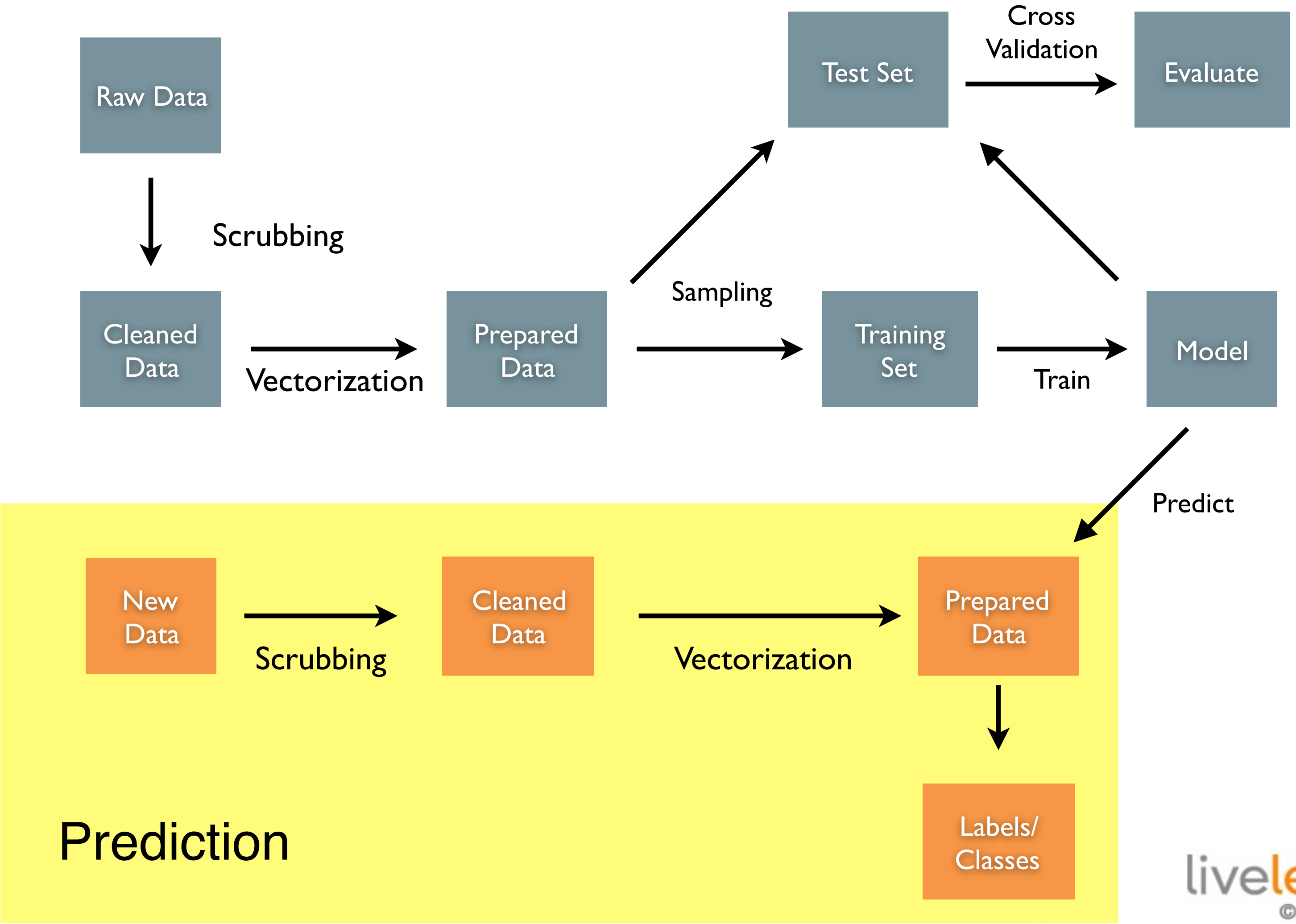
Lesson 5: Advanced Applications

5.7 Training Models with `spark.ml`

livelessons

`spark.ml` API

- **ML Dataset:** Uses Spark SQL `DataFrames` to hold variety of data types (including `Vectors` and predictions)
- **Transformer:** Algorithm which transforms one `DataFrame` into another (models, feature engineering, etc.)
- **Estimator:** Algorithm which can be fit on a `DataFrame` to produce a `Transformer` (learning algorithm which produces a model)
- **Pipeline:** Chains multiple `Transformers` and `Estimators` together into a workflow
- **Param:** All `Transformers` and `Estimators` share a common API for specifying parameters
- **Evaluator:** Operates on a `DataFrame` of predictions and computes an evaluation metric



Raw Data

Scrubbing

Cleaned Data

Vectorization

Prepared Data

Sampling

Test Set

Cross Validation

Evaluate

Training Set

Train

Model

Predict

New Data

Scrubbing

Cleaned Data

Vectorization

Prepared Data

Labels/ Classes

Prediction