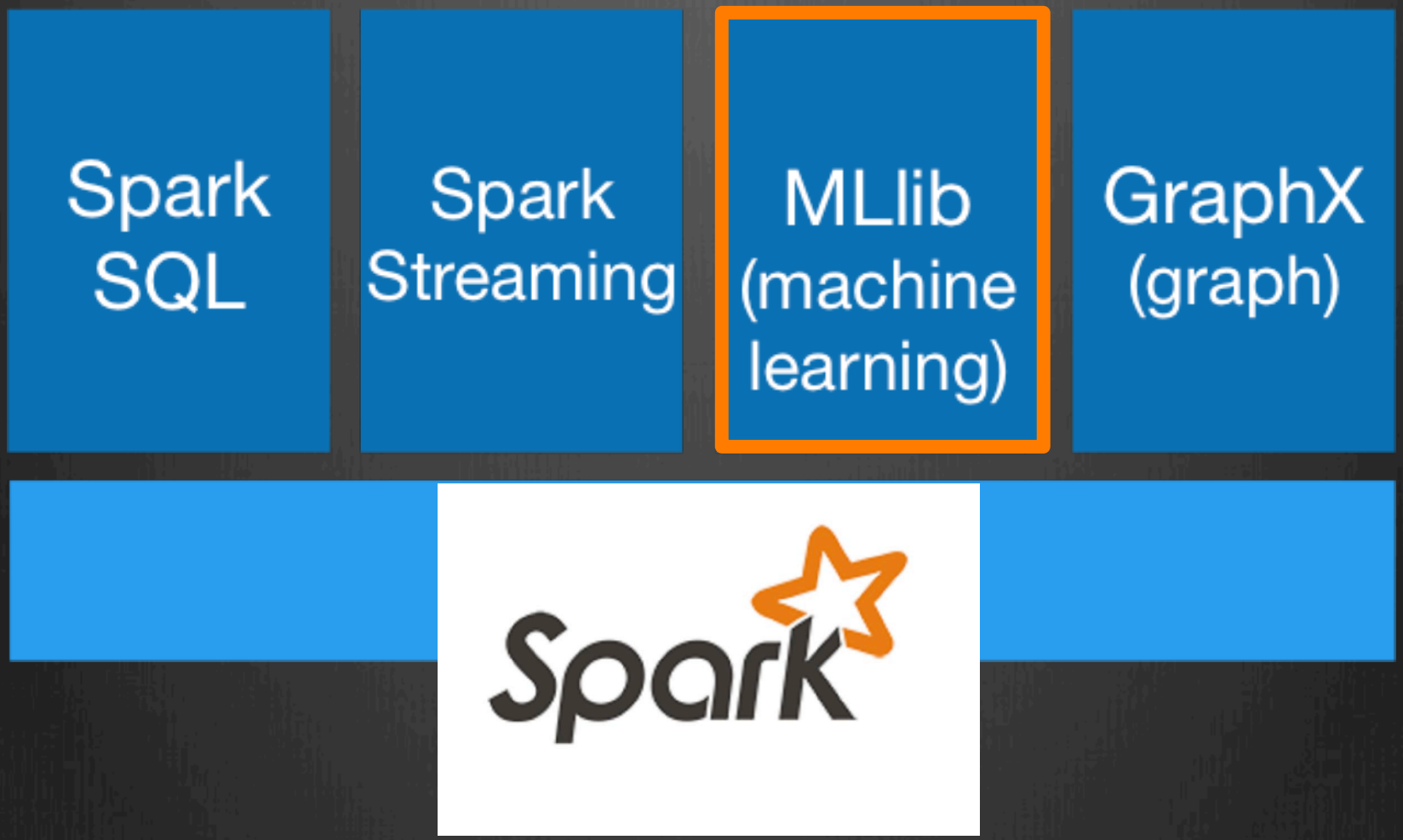
Machine Learning with MLlib and scikit-learn

Christopher Homa

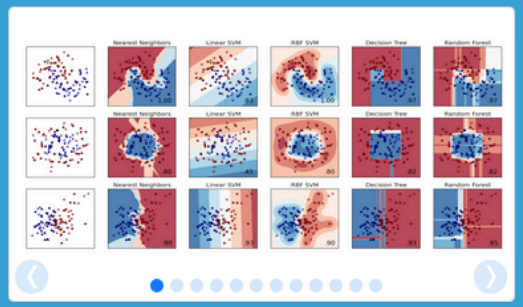

Spark SQL
Spark Streaming
MLlib (machine learning)
GraphX (graph)
Spark

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search

Fork me on GitHub



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *Isomap*, *non-negative matrix factorization*. — *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*. — *Examples*

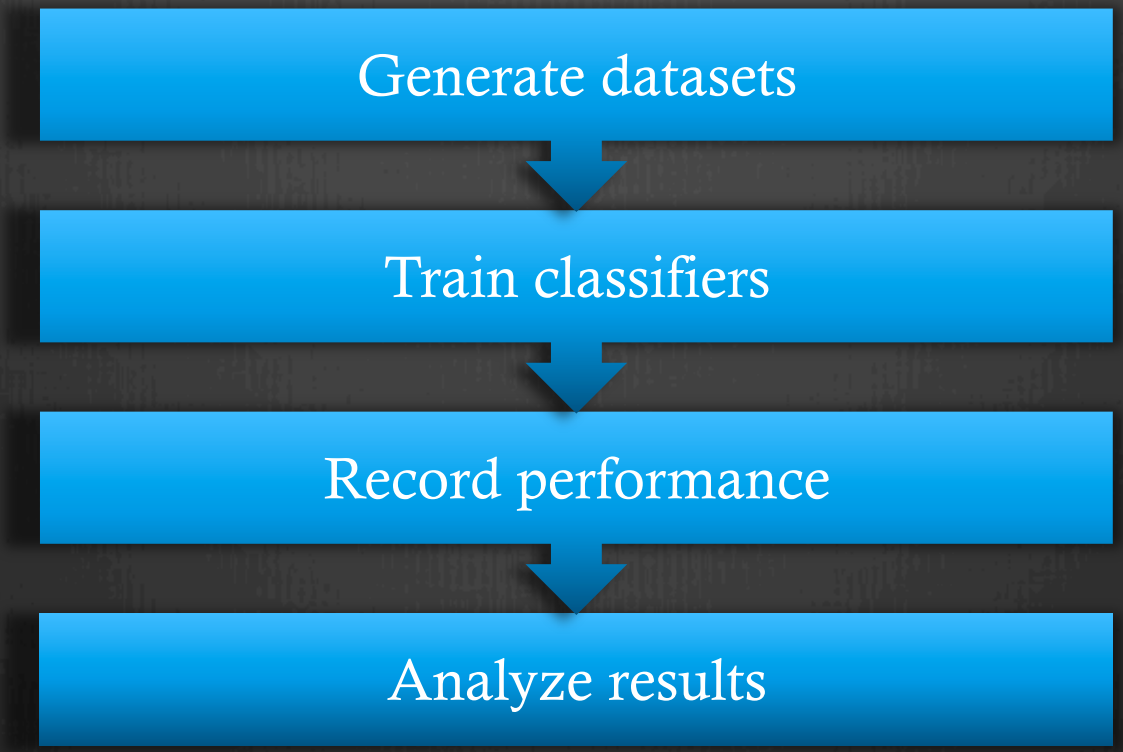
Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*. — *Examples*

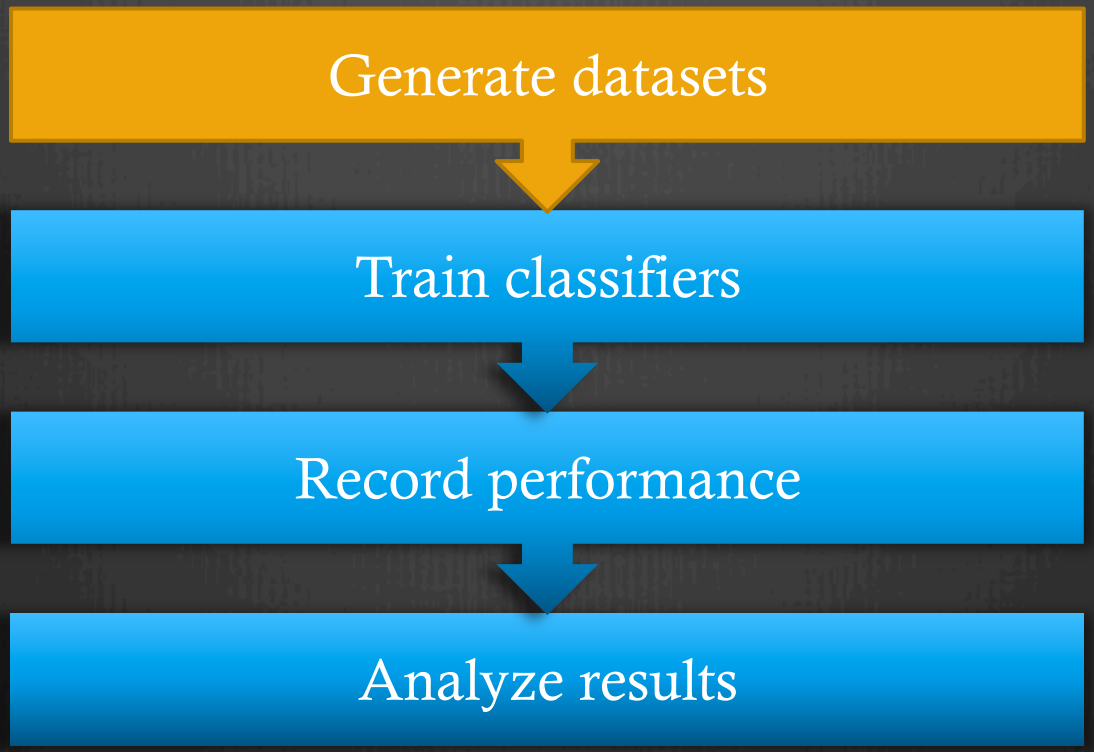
Goal

Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size



Goal

Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size



Generate datasets

Type

- Binary Classification
- Multiclass Regression
- Regression

Size

- Instances
- Features

Generate datasets

Type

- Binary Classification
- Multiclass Regression
- Regression

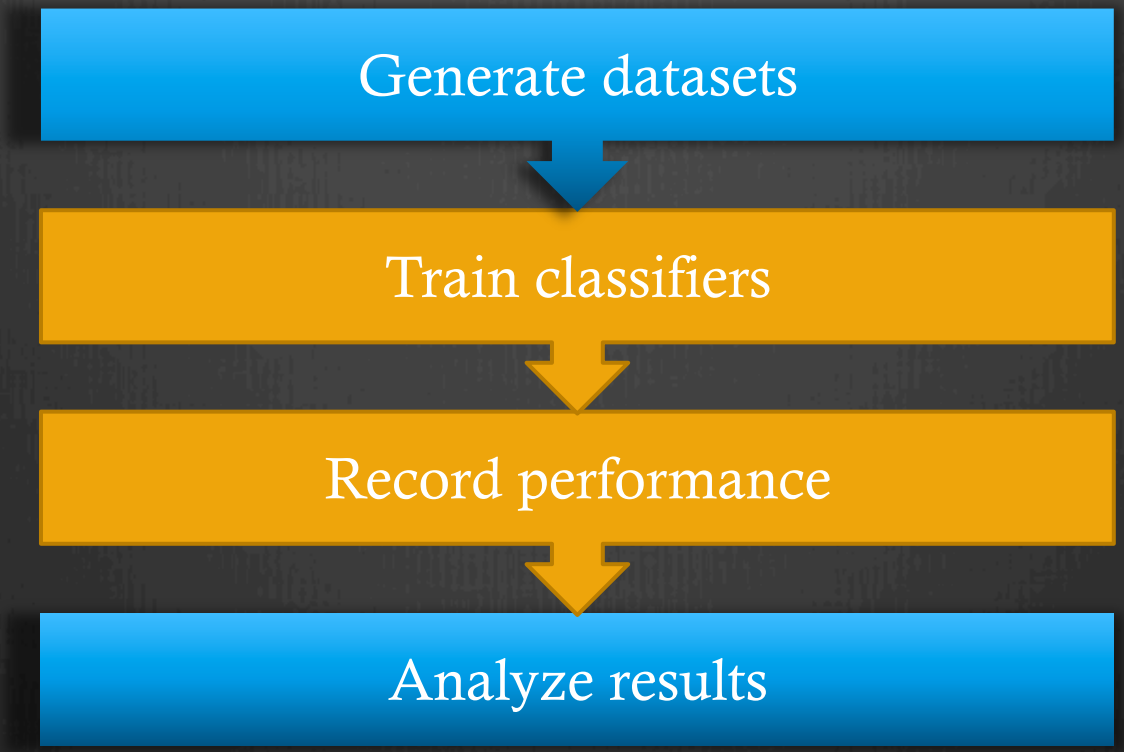
Size

- Instances
- Features

[1000,10] → [2000000,100]

Goal

Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size



Train classifiers

Choose classifiers

- Stochastic Gradient Descent
- Gradient Boosted Decision Trees
- Random Forests

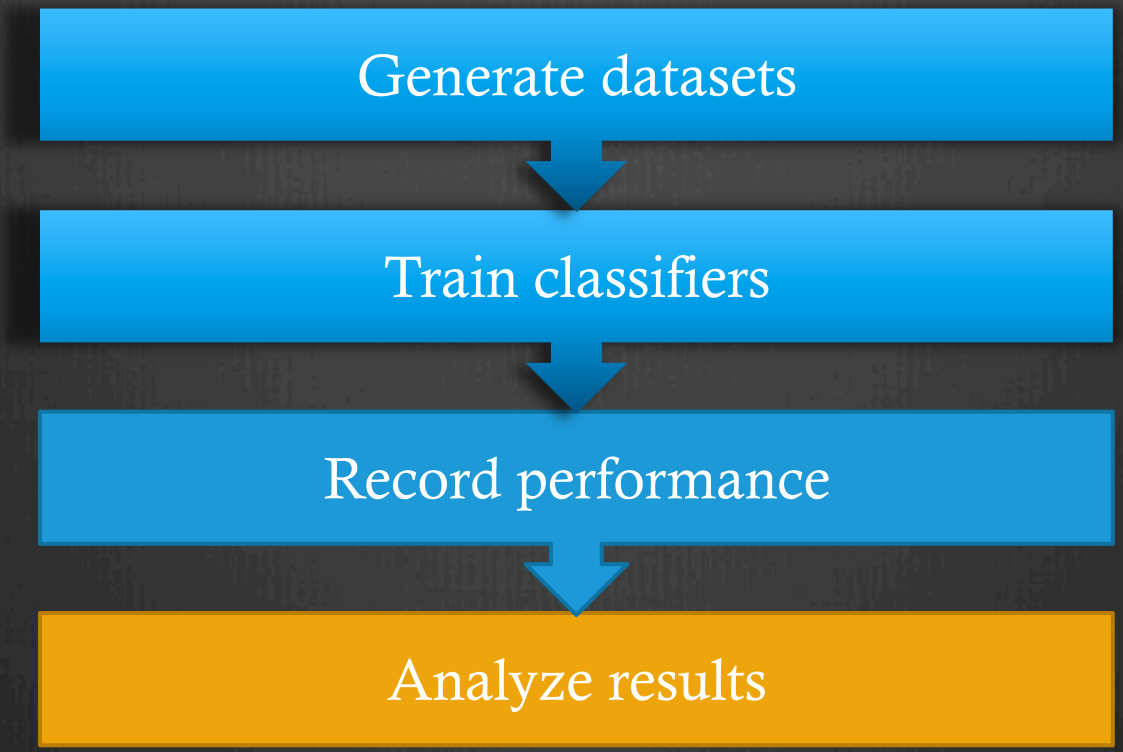
Match parameters

- Iterations
- Depth
- Most defaults match

Iteratively train classifiers on all datasets and record training times

Goal

Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size



Analyze results

Analyze results

Future Considerations

- Fewer, (much) larger datasets
- Utilize EC2 instances to run sk-learn scripts
- Improve data storage