UNIVERSITAT DE BARCELONA

Master in Fundamental Principles of Data Science

Dr Rohit Kumar

UNIVERSITAT DE BARCELONA

Spark MLLib

UNIVERSITAT DE BARCELONA

High Level Tools

ML Algorithms:	• common learning algorithms such as classification, regression, clustering, and collaborative filtering
Featurization:	• feature extraction, transformation, dimensionality reduction, and selection
Pipelines:	• tools for constructing, evaluating, and tuning ML Pipelines
Persistence:	• saving and load algorithms, models, and Pipelines
Utilities:	• linear algebra, statistics, data handling, etc.

RDD based APIs are retired and in maintenance mode only

UNIVERSITAT DE BARCELONA

Comparing with scitkit-learn

- The ML library is a bit limited compared to scitkit-learn but they are improving with every version
- Spark MLlib is efficient only if the data is large i.e you have too much training data and it is difficult to do it in a single machine with python
- In scitkit-learn, you can take an entire pandas DataFrame and send that to the machine learning algorithm for training. In Spark you have to pack all of your features, **from every column you want to train on, into a single column**, by extracting each row of values and packing them into a Vector.

UNIVERSITAT DE BARCELONA

ML Pipelines

ML Pipelines provide a uniform set of high-level APIs built on top of DataFrames that help users create and tune practical machine learning pipelines.

It is mostly similar concept as you have in scikit-learn!

UNIVERSITAT DE BARCELONA

ML Features

DataFrame	• This ML API uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types. E.g., a DataFrame could have different columns storing text, feature vectors, true labels, and predictions.
Transformer	• A Transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model is a Transformer which transforms a DataFrame with features into a DataFrame with predictions.
Estimator	• An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model.
Pipeline	• A Pipeline chains multiple Transformers and Estimators together to specify an ML workflow.
Parameter	• All Transformers and Estimators now share a common API for specifying parameters.

UNIVERSITAT DE BARCELONA

Transformers

Col 1	Col 2	Col 3
-	-	-
--	--	--
---	---	--

Col 1	Col 2	Col 3	Feature/ Prediction
-	-	-	
--	--	--	
---	---	--	

A **Transformer** implements a method ***transform(),*** which converts one DataFrame into another, generally by appending one or more columns.

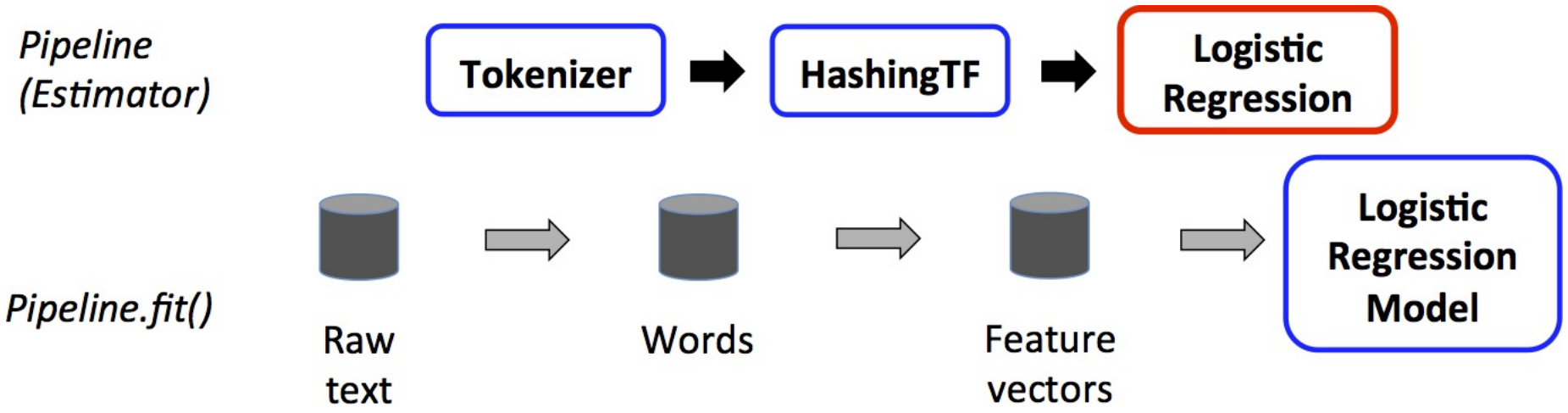
Example of some Transformers:

- Tokenizer
- HashingTF
- Vector Assembler
- Any trained ML model

https://spark.apache.org/docs/latest/ml-features.html

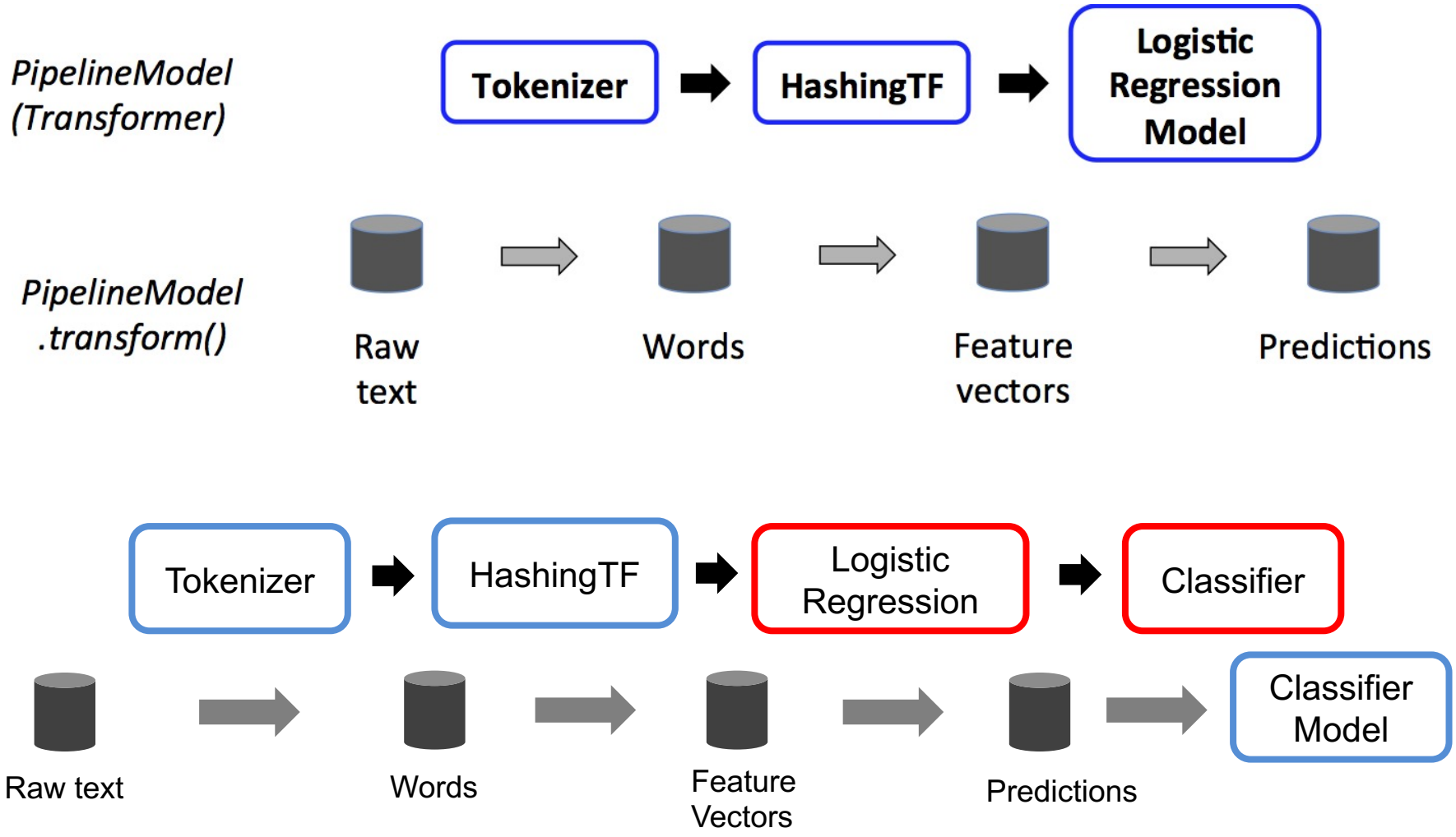
UNIVERSITAT DE BARCELONA

Pipelines



UNIVERSITAT DE BARCELONA

Pipelines



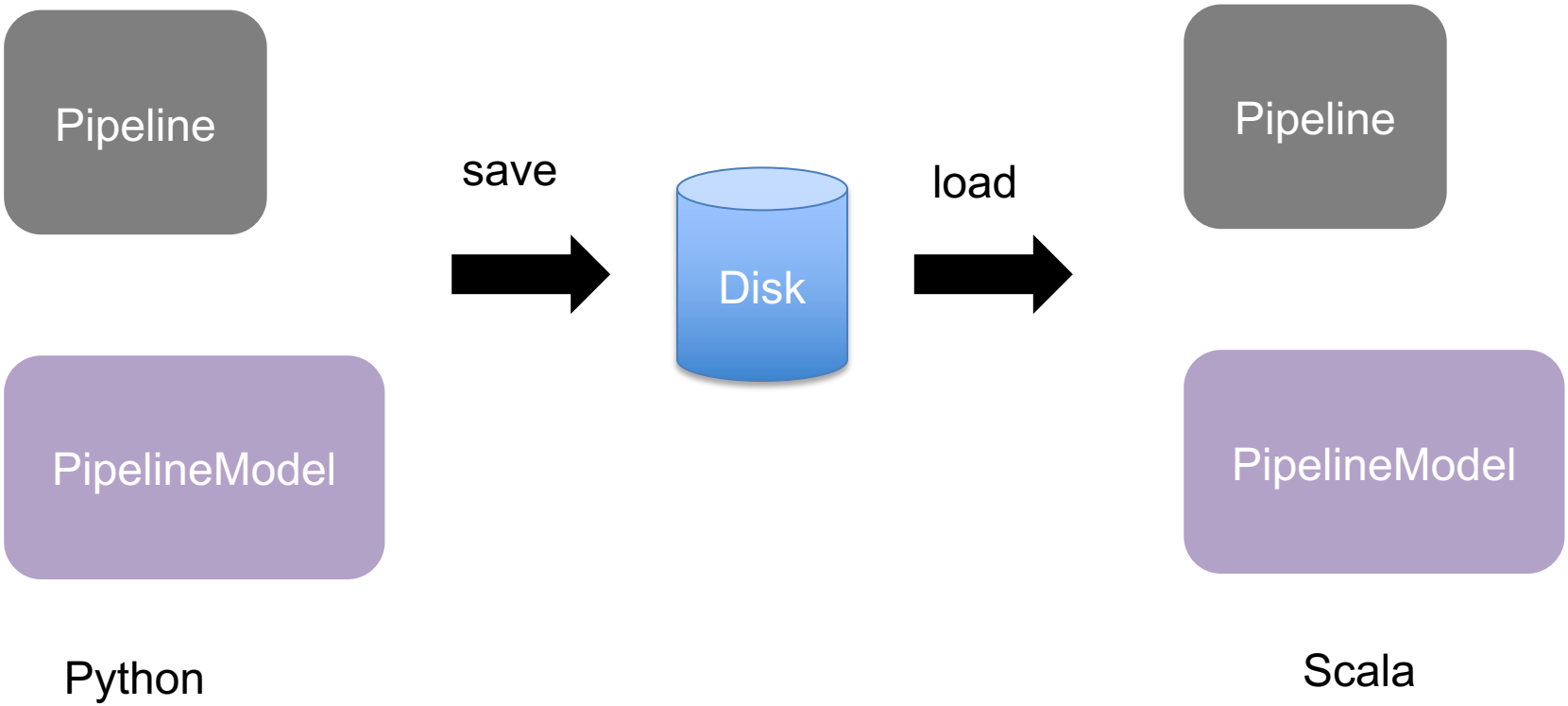
UNIVERSITAT DE BARCELONA

Some Details

- ***DAG Pipelines*:** It is possible to create non-linear Pipelines as long as the data flow graph forms a Directed Acyclic Graph (DAG).
- ***Runtime checking*:** Pipelines and PipelineModels do runtime checking before actually running the Pipeline of the parameters.
- ***Unique Pipeline stages*:** A Pipeline's stages should be unique instances. E.g., the same instance myHashingTF should not be inserted into the Pipeline twice since Pipeline stages must have unique IDs. However, different instances myHashingTF1 and myHashingTF2 (both of type HashingTF) can be put into the same Pipeline since different instances will be created with different IDs.

UNIVERSITAT DE BARCELONA

Persistence



ML persistence works across Scala, Java and Python. However, R currently uses a modified format, so models saved in R can only be loaded back in R

UNIVERSITAT DE BARCELONA

Compatibility

Model persistence: Is a model or Pipeline saved using Apache Spark ML persistence in Spark version X loadable by Spark version Y?

- Major versions: No guarantees, but best-effort.
- Minor and patch versions: Yes; these are backwards compatible.

Model behavior: Does a model or Pipeline in Spark version X behave identically in Spark version Y?

- Major versions: No guarantees, but best-effort.
- Minor and patch versions: Identical behavior, except for bug fixes.

UNIVERSITAT DE BARCELONA

References

- https://databricks.com/session/combining-the-strengths-of-mllib-scikit-learn-and-r
- https://spark.apache.org/docs/latest/ml-guide.html
- https://spark.apache.org/docs/latest/ml-pipeline.html