# MLeap: Release Spark ML Pipelines

Mikhail Semeniuk and Hollin Wilkins


SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO

# Opening Demo

How much should I rent my house for on AirBnb?

`http://spark-summit.combust.ml`

Yes, open your cell phone and go here :)

Problem Statement: Deploying machine learning algorithms to a production environment is a lot more difficult than it has to be and is a common source of friction at data-driven organizations

| Action | | Reaction |
|---|---|---|
| - Data scientists write data pipelines to construct research datasets | ⟷ | - Engineers re-write the data pipelines for a production-ready system |
| - Engineers write scalable libraries for computing features and algorithms | ⟷ | - Data scientists largely don't use those libraries and maintain/re-write their own copy of the code |
| - Data scientists largely focus on linear/logistic regressions due to engineering constraints | ⟷ | - Talented engineers get largely tired of coding up linear regressions and updating coefficients |

***Everyone wants to do better! The winning technology will be the one that enables Engineers and Data Scientists to collaborate and work across a single platform.***

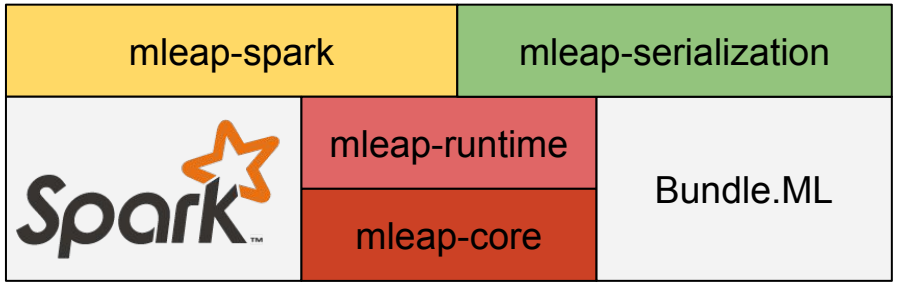
Existing Solutions: You won't believe how many companies are still deploying algorithms in a SQL environment! And these are billion dollar operations.

| | Hard-Coded Models (SQL, Java, Ruby) | PMML | Emerging Solutions (yHat, DataRobot) | Enterprise Solutions (Microsoft, IBM, SAS) | MLeap |
|---|---|---|---|---|---|
| Quick to Implement | ✗ | ✓ | ✗ | ✗ | ✓ |
| Open Sourced | ✗ | ✓ | ✗ | ✗ | ✓ |
| Committed to Spark/Hadoop | ✗ | − | − | ✓ | ✓ |
| API Server Infrastructure | ✗ | − | ✓ | ✗ | ✓ |

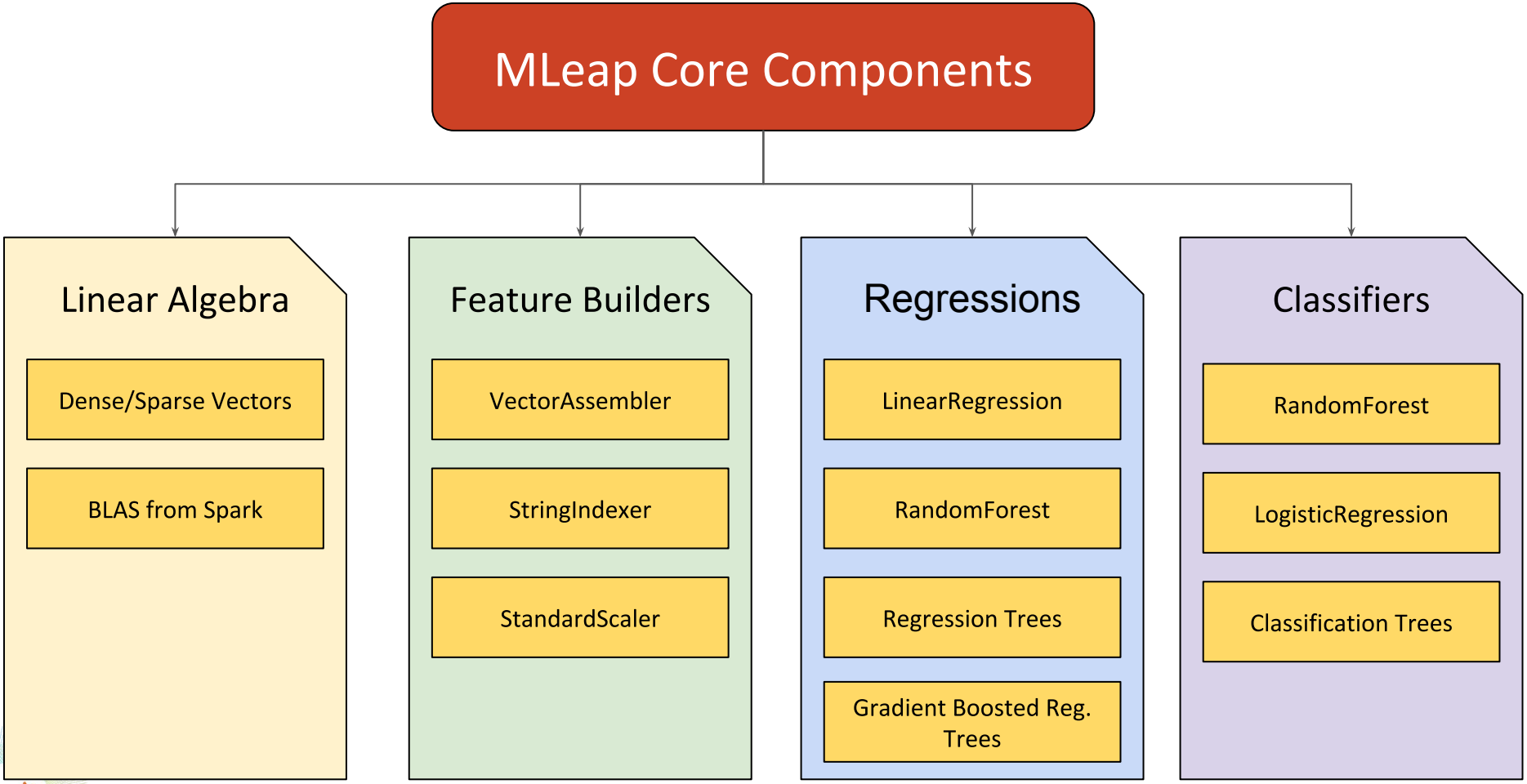
Lesson Learned: Push code down to where the data is, not the other way around!

# MLeap Components

- core - provides linear algebra system, regression models, and feature builders
- runtime - provides DataFrame-like “LeapFrame” and transformers for it
- spark - provides easy conversion from Spark transformers to MLeap transformers
- serialization - common serialization format for Spark and MLeap (Bundle.ML)




New features: expanded serialization formats to include both json and protobuf for large models (i.e. random forests with thousands of features)

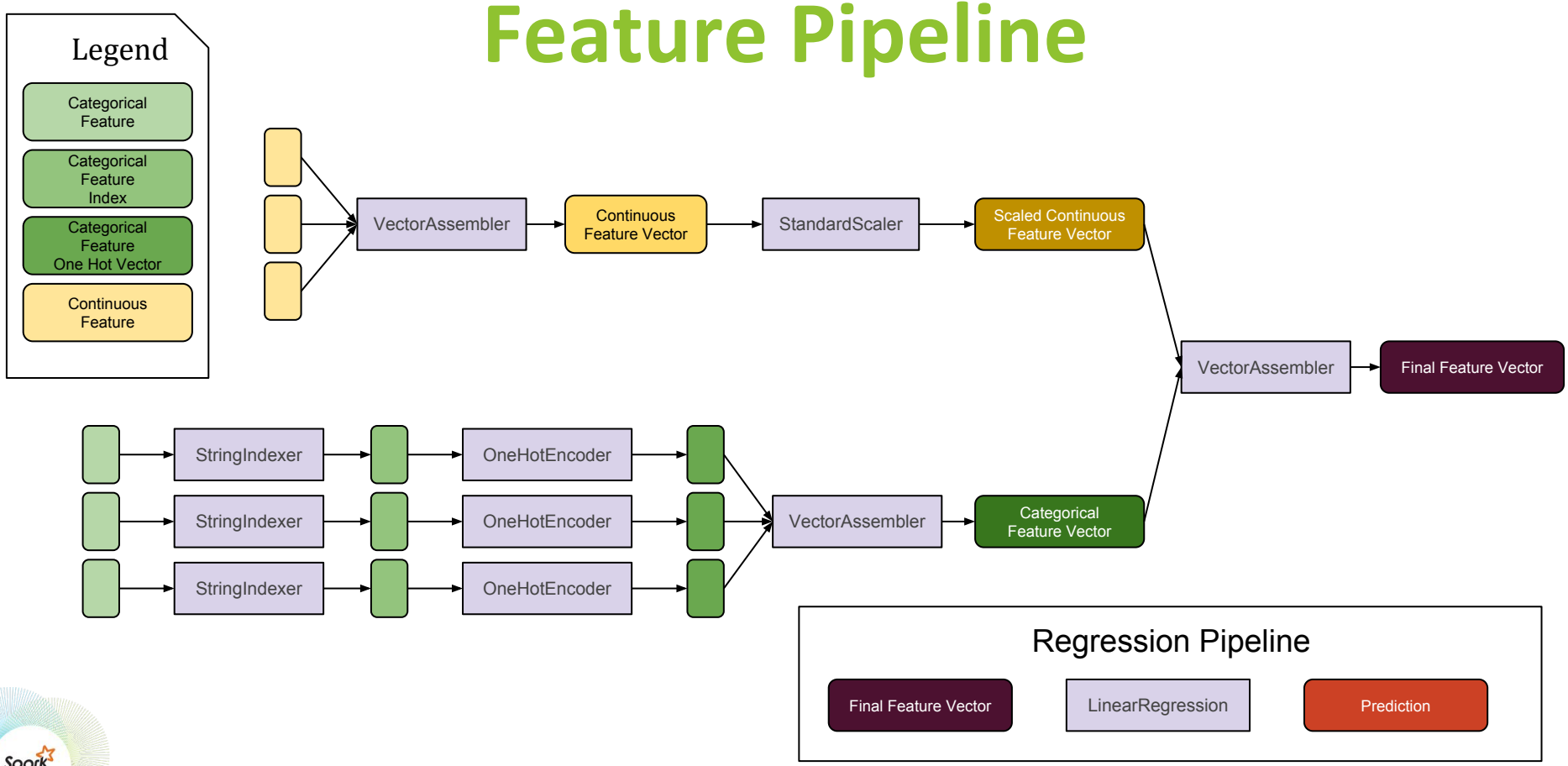

MLeap Core Components
Linear Algebra
Dense/Sparse Vectors
BLAS from Spark
Feature Builders
VectorAssembler
StringIndexer
StandardScaler
Regressions
LinearRegression
RandomForest
Regression Trees
Gradient Boosted Reg. Trees
Classifiers
RandomForest
LogisticRegression
Classification Trees

# MLeap Runtime

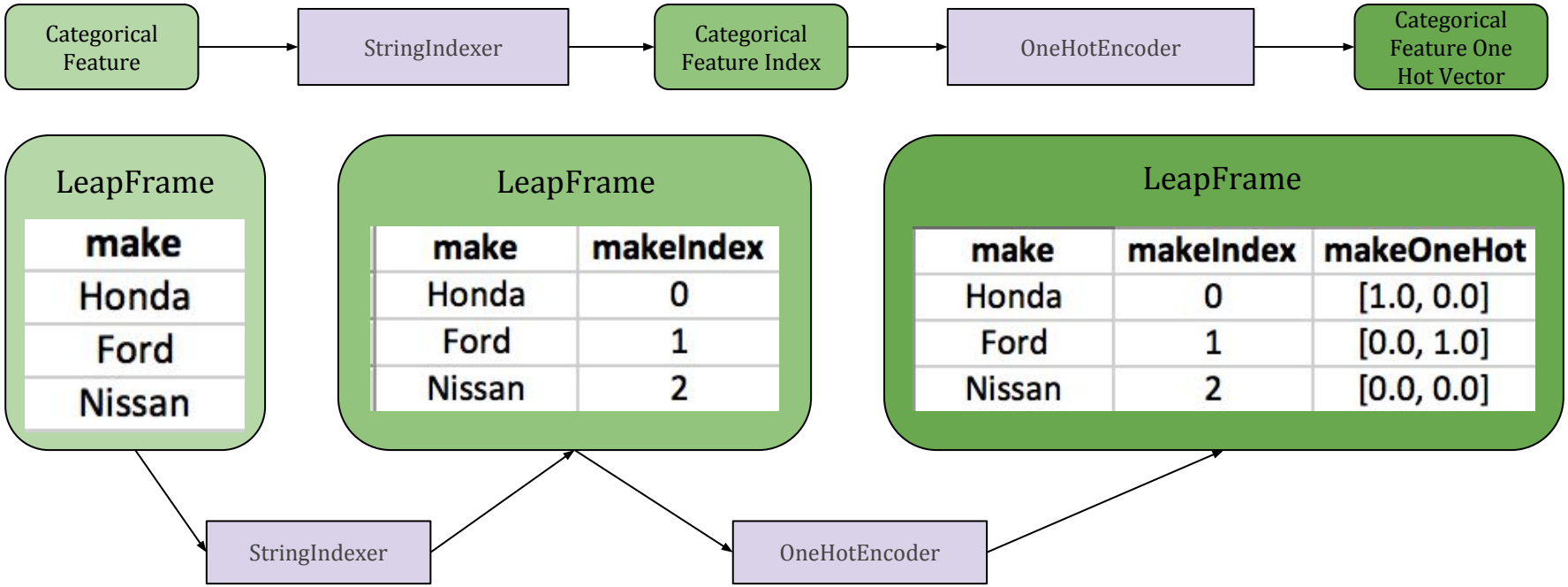
- Provides LeapFrame, which stores data for transformations by MLeap transformers
- MLeap transformers use mleap-core building blocks to transform LeapFrame
- MLeap transformers correspond one-to-one with Spark transformers
- No dependencies on Spark



# Feature Pipeline



# Categorical Pipeline



# MLeap Serialization (Bundle.ML)

- Provides common serialization for both Spark and MLeap
- 100% protobuf/JSON based for easy reading, compact data, and portability
- No dependencies on Parquet *
- Can be written to zip files, file system, HDFS, anywhere with an FS-like structure

String Indexer Model

```
{
  "type": "com.truecar.mleap.runtime.transformer.StringIndexerModel",
  "inputCol": "make",
  "outputCol": "makeIndex",
  "indexer": {
    "strings": ["Ford", "Nissan", "Honda"]
  }
}
```

Linear Regression Model

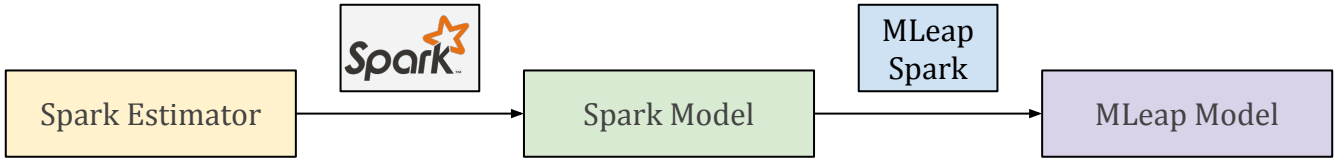
```
{
  "type": "com.truecar.mleap.runtime.transformer.LinearRegressionModel",
  "featuresCol": "features",
  "predictionCol": "listOverMsrpPrediction",
  "model": {
    "weights": [-0.060649238549871816, -0.008726825316488376, ...],
    "intercept": 0.360124036840183
  }
}
```

Linear Regression Model (Code)

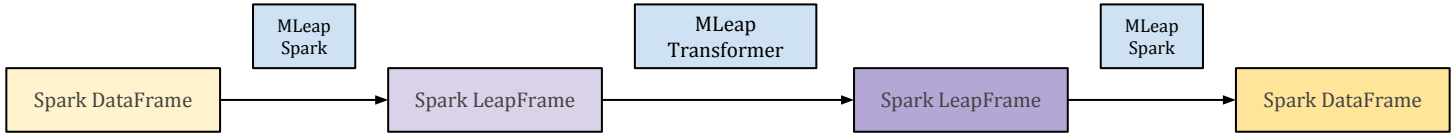
```
case class LinearRegression(coefficients: Vector,
                            intercept: Double) extends Serializable {
  def apply(features: Vector): Double = {
    features.toBreeze.dot(coefficients.toBreeze) + intercept
  }
}
```

# MLeap Spark

- Train an ML pipeline with Spark then export it to MLeap

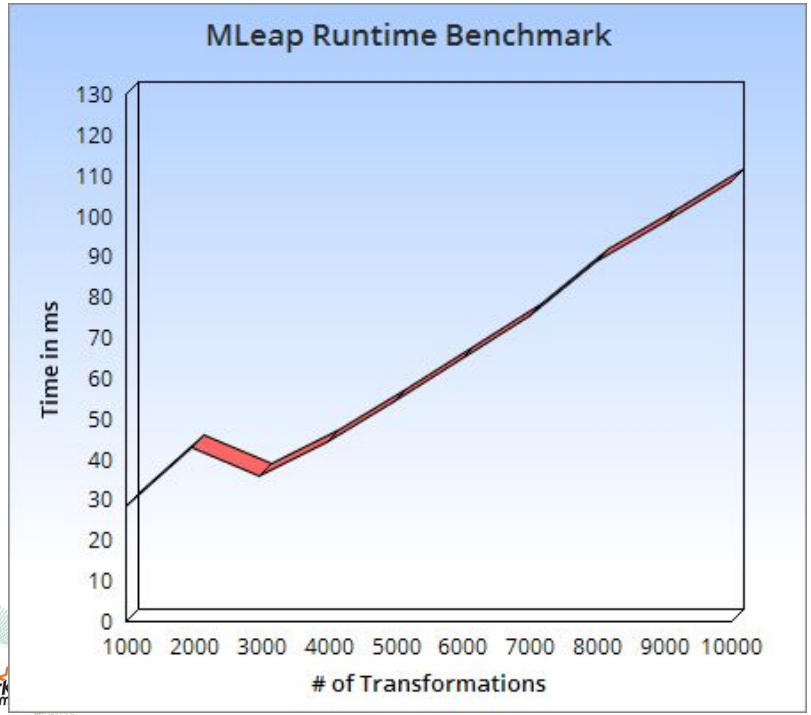



- Execute an MLeap pipeline against a Spark DataFrame

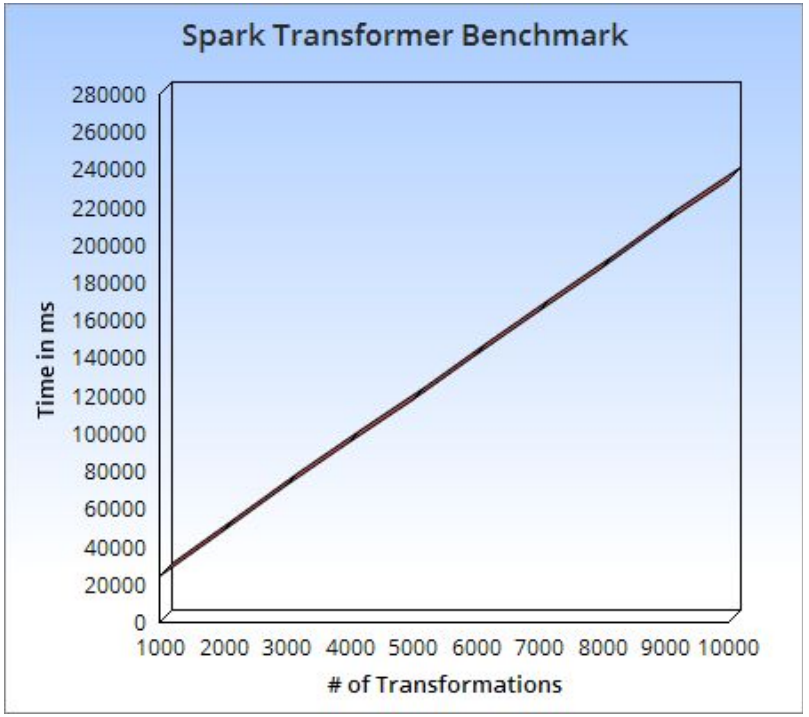


# Benchmarks

MLeap: **0.011ms**/transform




Spark: **23.4ms**/transform



# MLeap Demo

- Train a sample Airbnb listing price model using linear regression and random forest against some AirBnb training data
- Deploy both models to a local API server
- Get real-time results
- IN UNDER 5 MINUTES!

# Future of MLeap

- Unify linear algebra and core libraries with Spark
- Python/R interface (6 months)
- Deploy easily to embedded systems and outside of JVM (1 year)
- Full support for all Spark transformers

# Combust.ML Overview

- Provides a scalable scala-based API server, tuned specifically for MLeap models
- Public interface to drop-in data and deploy restful services
- Feedback loops for verifying model accuracy
- Feature vector definitions for researchers and engineers

# Thank Yous

Spark Saturday - Capital One, D.C.


Roaring Elephant Podcast, Netherlands


Hadoop Summit - Dublin, IR


Spark Summit West - New York, NY


Mishkin Faustini

Prianna Ahsan

Ram Sriharsha

# THANK YOU.

**Mikhail Semeniuk**
email: seme0021@gmail.com
github: https://github.com/seme0021
twitter: https://twitter.com/MikhailSemeniuk
linkedin: https://www.linkedin.com/in/semeniuk

**Hollin Wilkins**
email: hollinrwilkins@gmail.com
github: https://github.com/hollinwilkins
twitter: https://twitter.com/HollinWilkins
linkedin: https://www.linkedin.com/in/hollinwilkins


SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO