

Introduction to .NET for Apache Spark

Luis Quintanilla

[@ljquintanilla](#)

Welcome



ML.NET Content Developer, Microsoft



lqdev@outlook.com



<http://luisquintanilla.me>



[@ljquintanilla](https://twitter.com/ljquintanilla)



<https://github.com/lqdev>

Code & Slides

<http://bit.ly/NYCCodeCamp2019>

Agenda

- 01** What is Big Data?
- 02** Apache Spark
- 03** .NET for Apache Spark
- 04** Working with Data
- 05** Build Apps with .NET for Apache Spark

What is **Big Data**?

3 V's of Big Data

Volume

- Terabytes
- Petabytes

Variety

- Structured
- Unstructured

Velocity

- Batch
- Streaming

Big Data Software



What is **Apache Spark**?

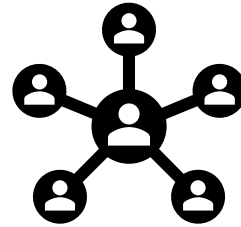
Apache Spark



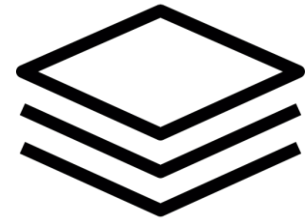
Open
Source



Cross
Platform

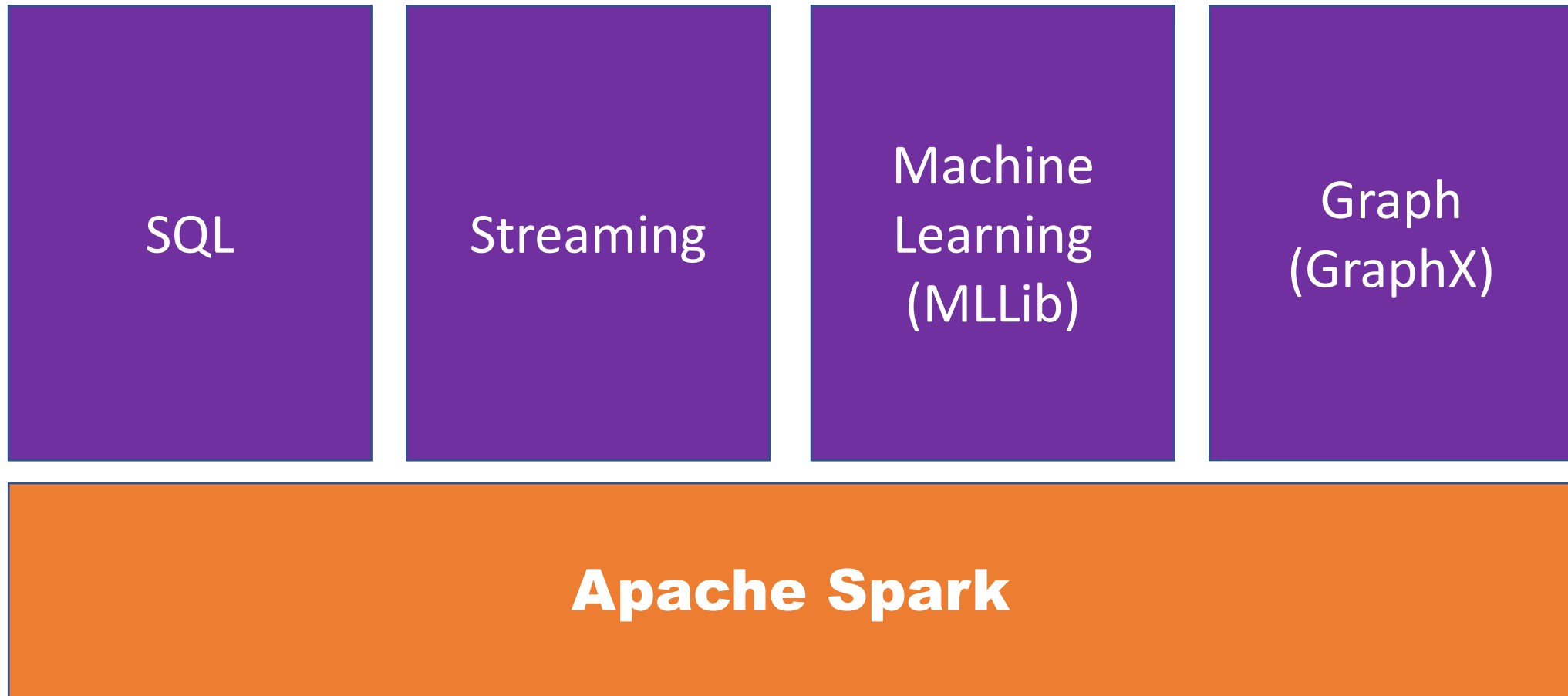


Distributed*

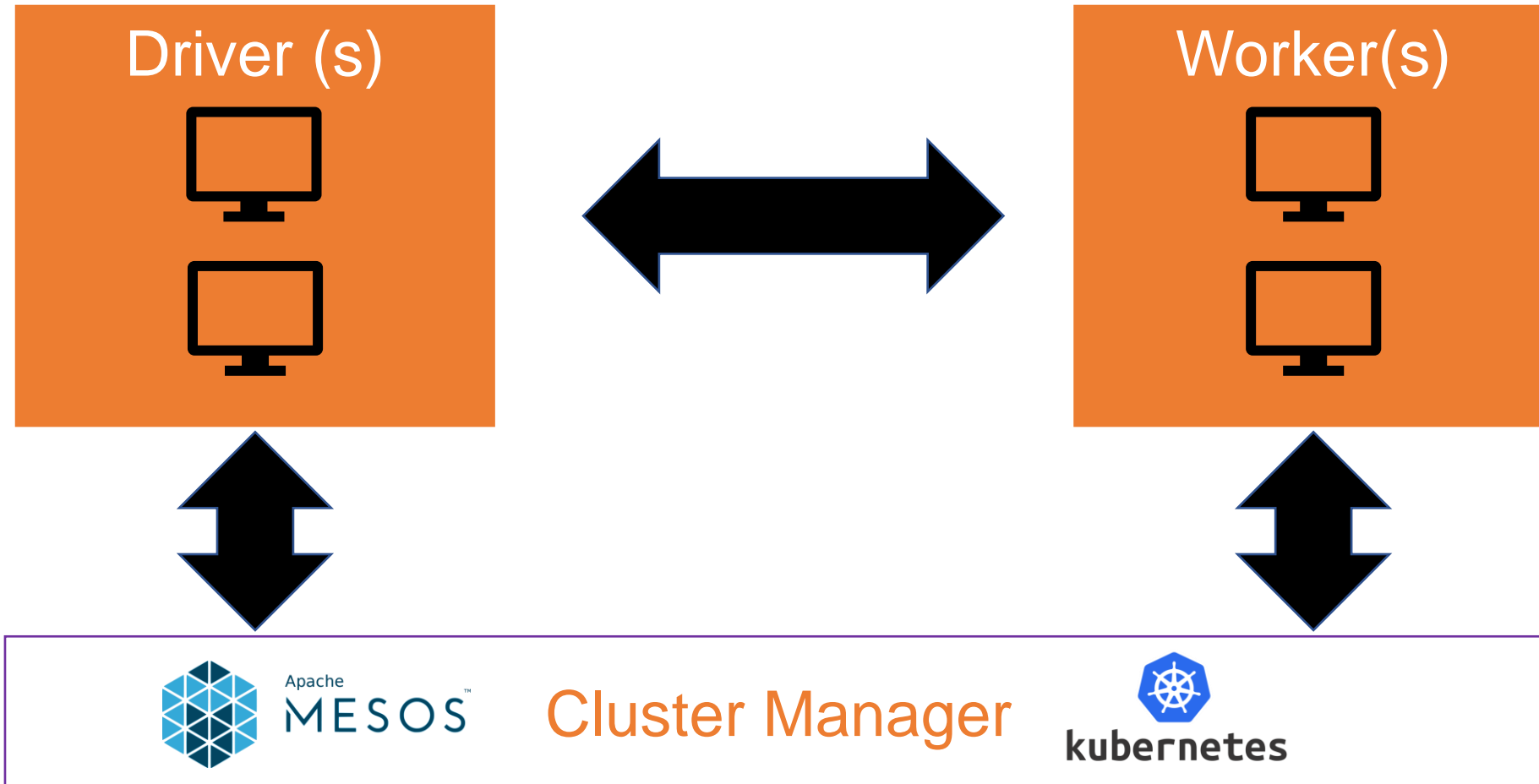


Extensible

Apache Spark Architecture



Apache Spark Execution



Interacting with Apache Spark

```
b.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.properties
To adjust logging level use sc.setLogLevel("INFO")
Welcome to

  ____  _
 / ___|| | | |
| |___| |_| |
 \___ \|  _/
      |_|

 version 1.6.1

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_91)
Type in expressions to have them evaluated.
Type :help for more information.

█
```

Shell

Notebooks

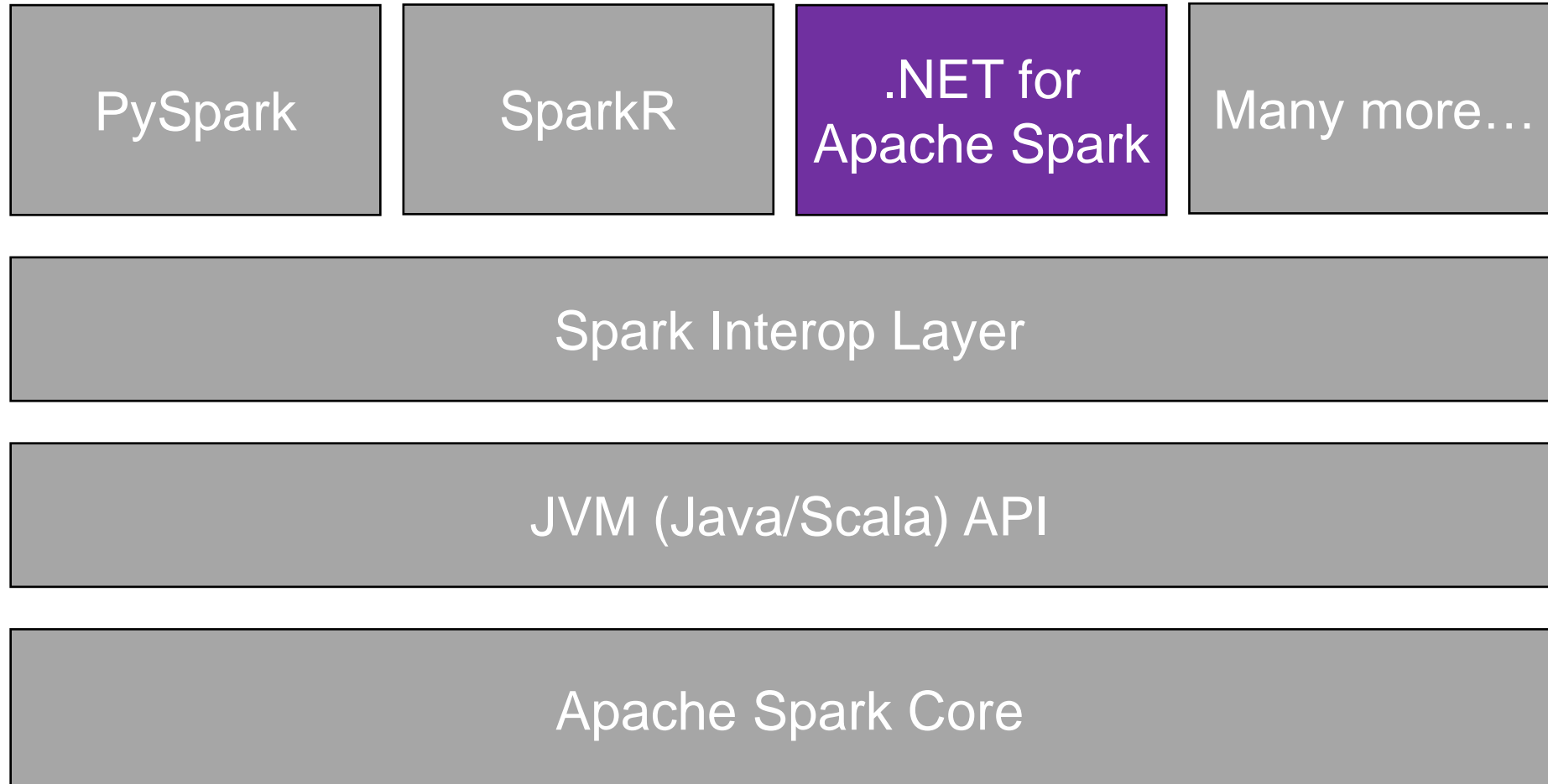
@ljquintanilla

```
// Create SparkSession
var sc =
    SparkSession
        .Builder()
        .AppName("Restaurant_Inspections_ETL")
        .GetOrCreate();

// Load data
DataFrame df =
    sc
        .Read()
        .Option("header", "true")
        .Option("inferSchema", "true")
        .Csv("Data/NYC-Restaurant-Inspections.csv");
```

Traditional Apps

Extending Apache Spark



.NET for Apache Spark

.NET for Apache Spark



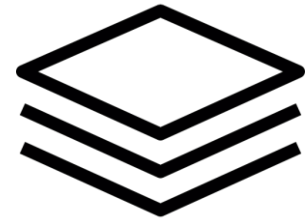
Open
Source



Cross
Platform

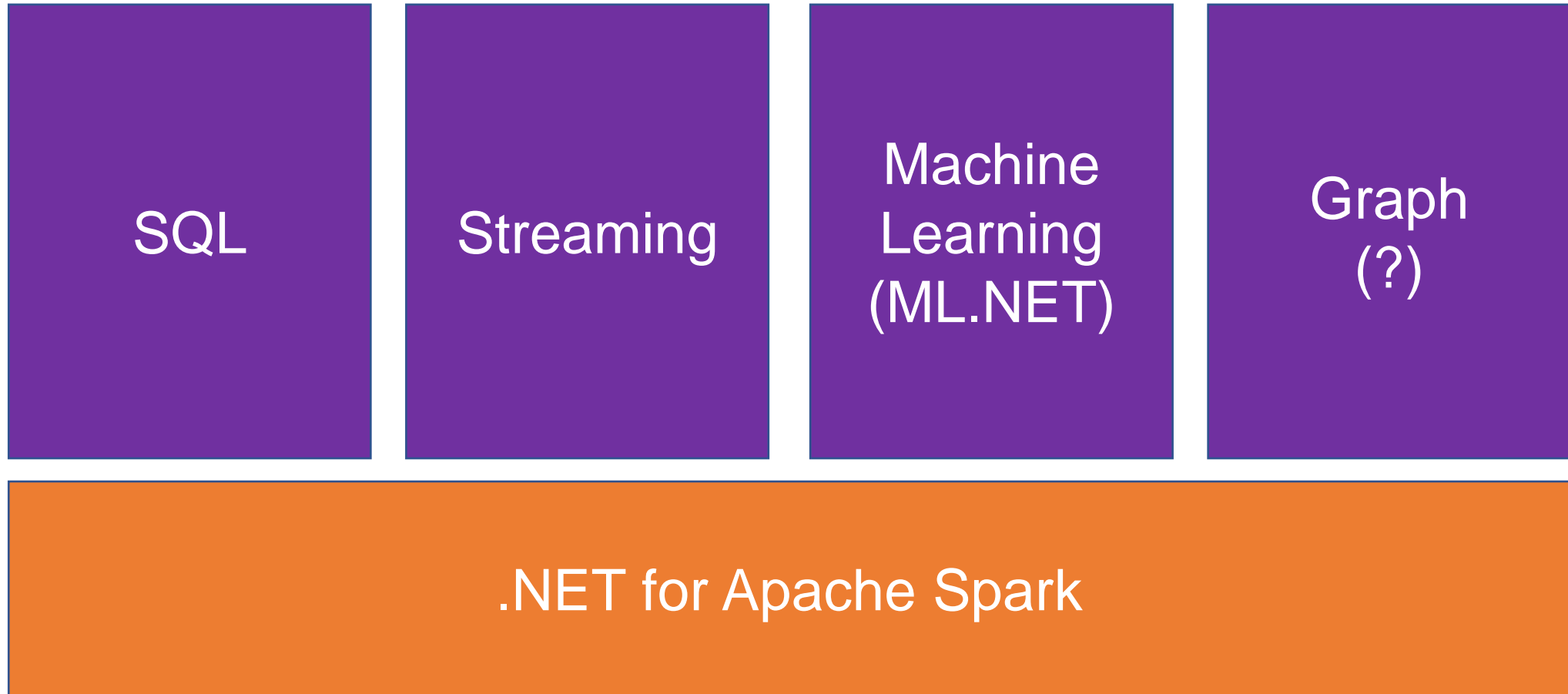


.NET
Standard



Extensible

.NET for Apache Spark Architecture



ML.NET Framework

Transforms

- Missing Values
- Feature Selection
- Normalization

Trainers

- SVM
- K-Means
- Boosted Trees
- Logistic Regression

Misc

- Data Loaders
- Evaluators

Extensions

- TensorFlow
- ONNX

Data in Apache Spark

Resilient
Distributed
Datasets (RDD)

DataFrames

Datasets

Demo

.NET for Apache Spark

DataFrames

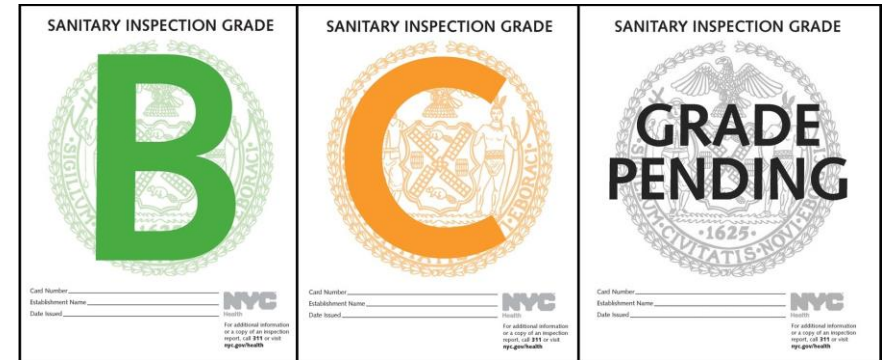
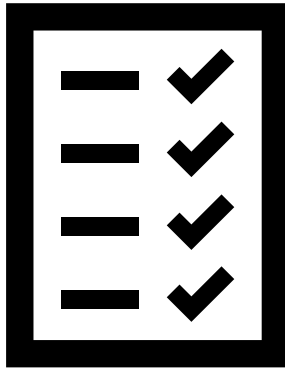
Demo

.NET for Apache Spark

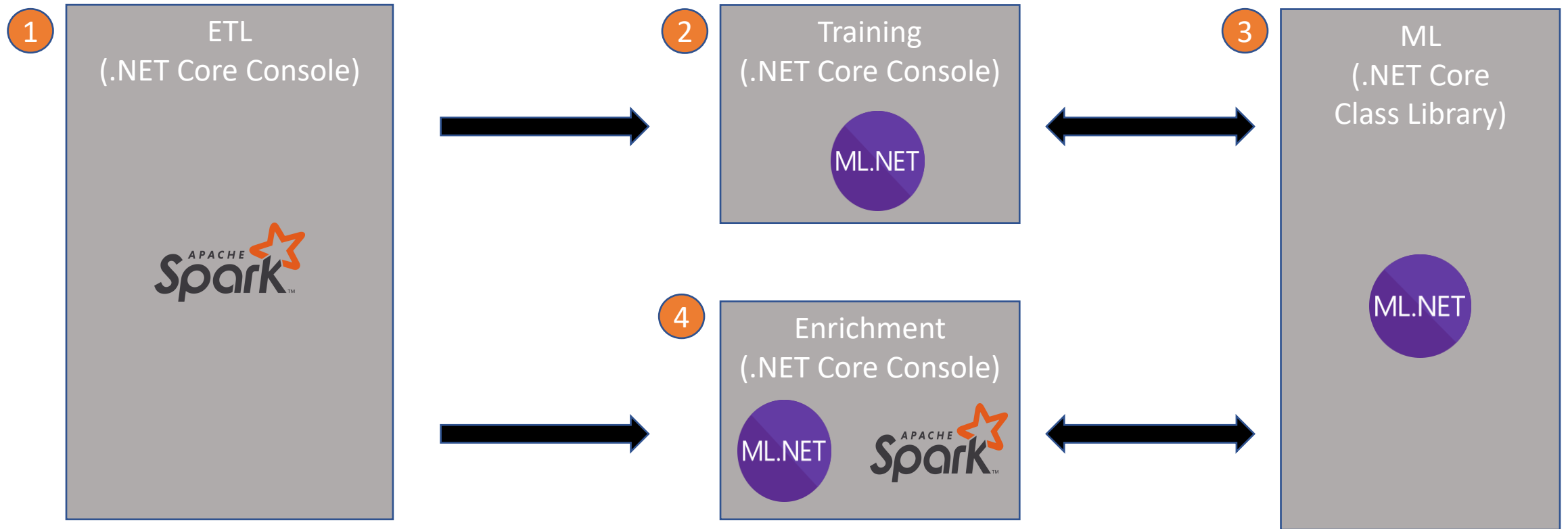
Streaming

Building Applications with .NET for Apache Spark

Restaurant Inspections



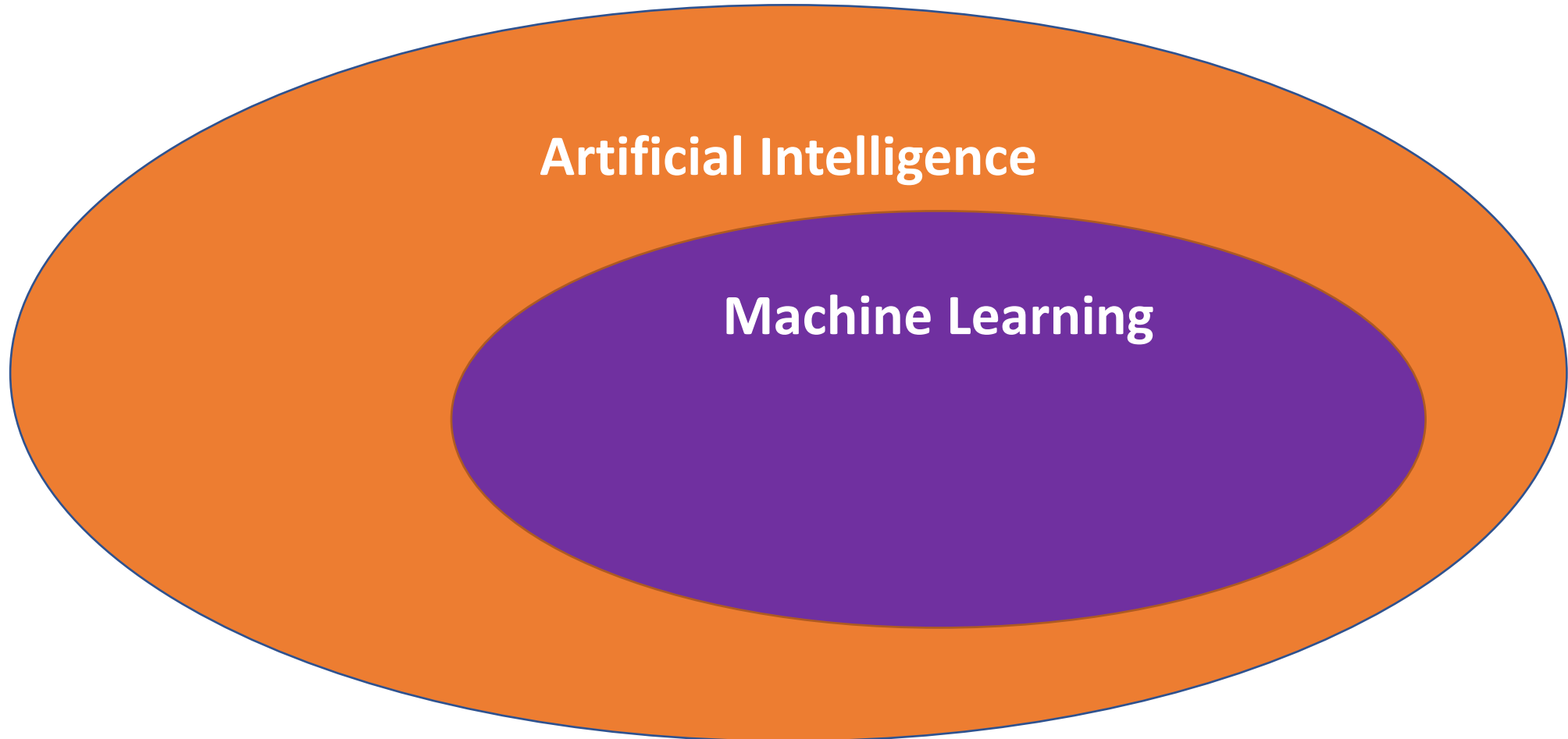
Restaurant Inspections Architecture



Demo

Restaurant Inspections ETL

What is Machine Learning?



Machine Learning Tasks

Supervised Learning

Regression

What is
the price
of a home
in NYC?

Classification

Is this a
dog or
cat?

Unsupervised Learning

Clustering

Customer
segments
in a
database

Classification Example

Training Data

Species	Is Independent	Class
Canine	False	Dog
Feline	True	Cat
Feline	True	Cat
Canine	False	Dog
Canine	True	Dog

Features
(input)

Label
(output)

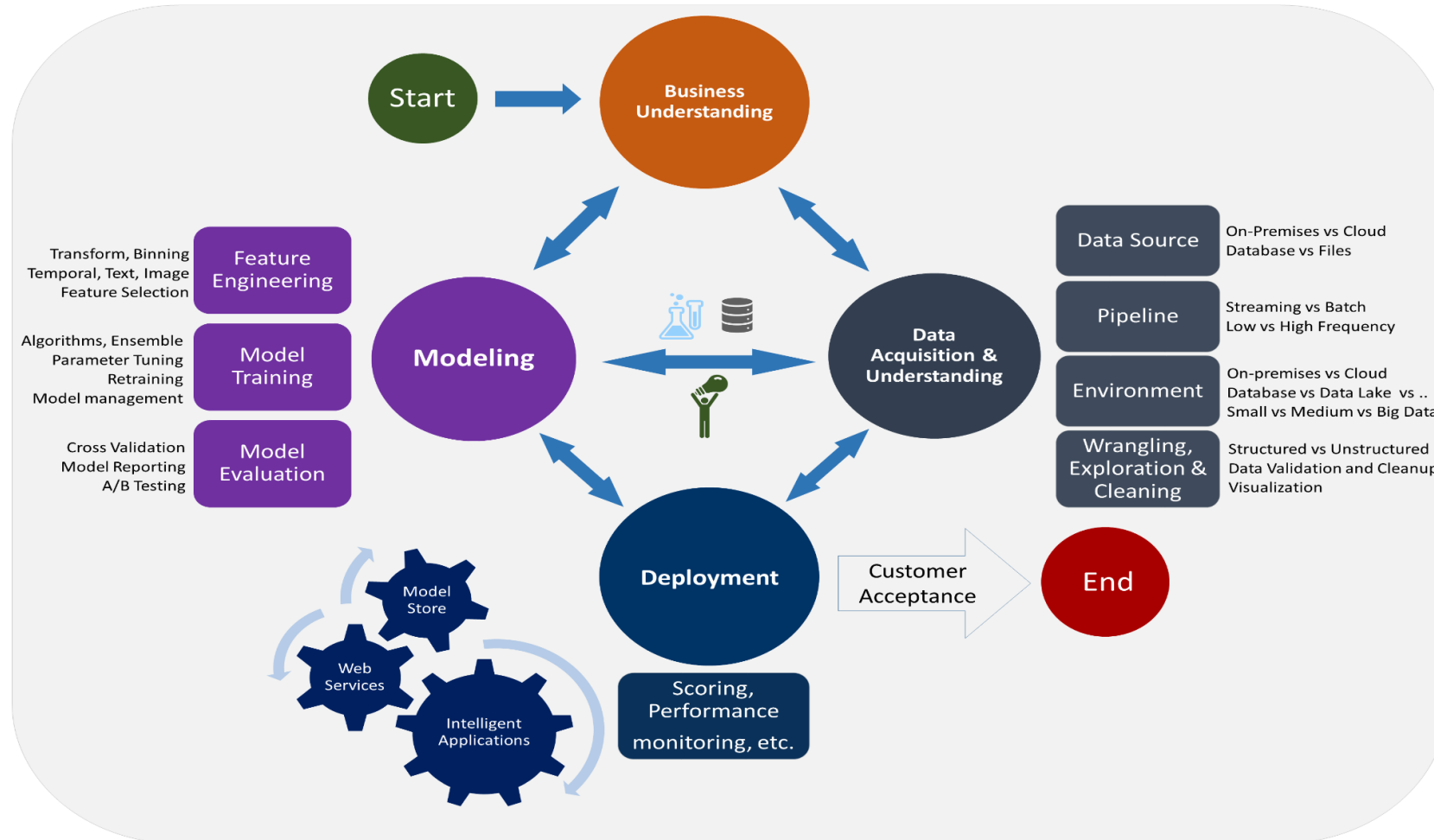
New Data

Species	Is Independent
Canine	False

Prediction

Class
Dog

The Continuous Machine Learning Process



What is a **model**?

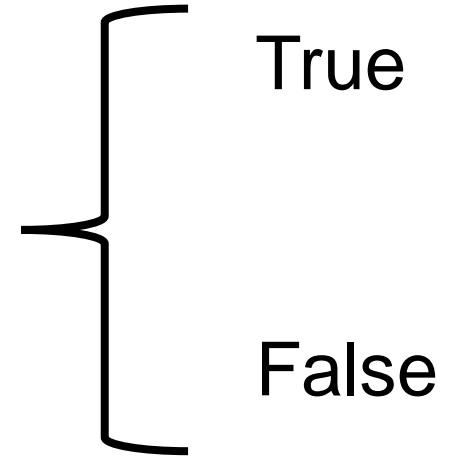


Input



$f(x)$

Model



Output

Demo
Restaurant Inspections
Training
(ML.NET Automated (Auto) ML)

Demo
Restaurant Inspections
Enrichment
(.NET for Apache Spark + SQL)

Demo Restaurant Inspections Enrichment (Azure HDInsight)

Takeaways

- Apache Spark is a fast and extensible computing platform for processing Big Data workloads.
- .NET for Apache Spark brings the power of Apache Spark to the .NET ecosystem.
- .NET for Apache Spark provides easy integration with other .NET libraries / frameworks such as ML.NET.

Resources

- <http://luisquintanilla.me/2019/09/15/restaurant-inspections-etl-data-enrichment-spark-auto-ml-net/>
- <https://docs.microsoft.com/en-us/dotnet/spark/>
- <https://github.com/dotnet/spark>
- <https://docs.microsoft.com/en-us/dotnet/machine-learning/>
- <https://github.com/dotnet/machinelearning>
- <https://spark.apache.org/>

Questions ?