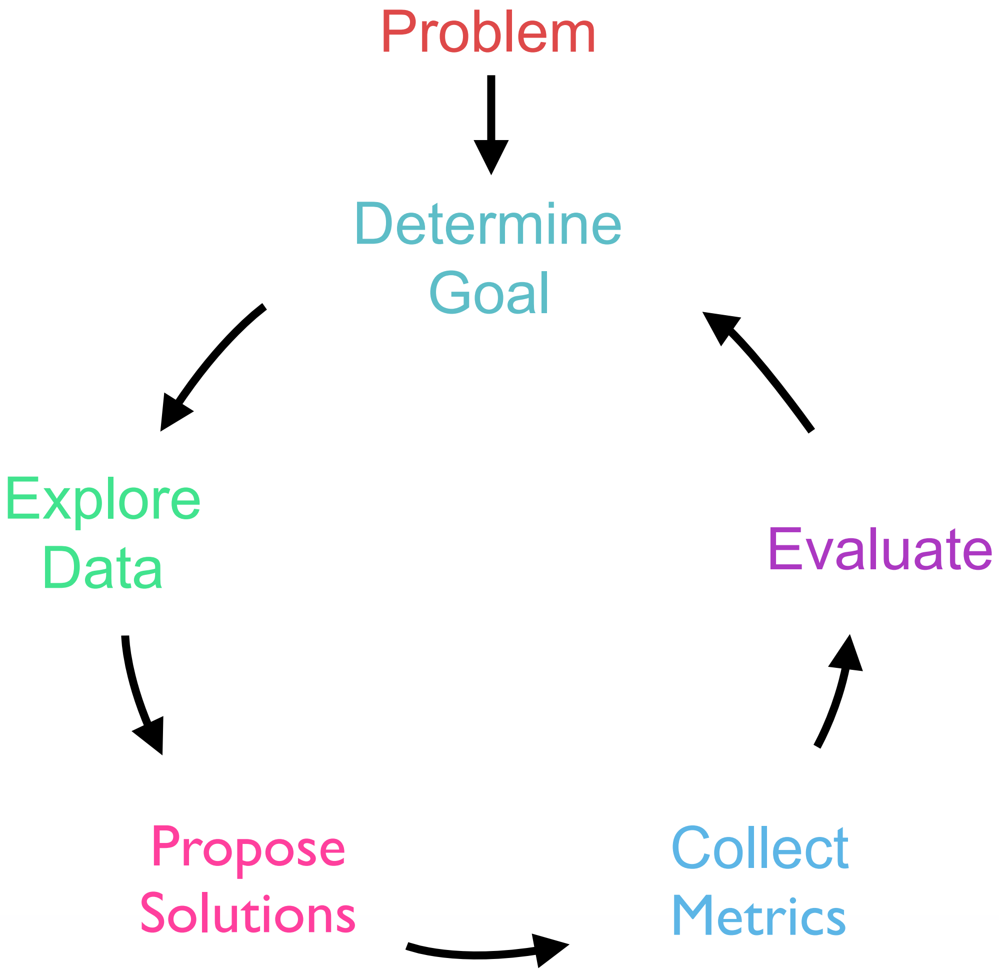
livelessons
video instruction from technology experts

Lesson 3: Your First Spark Application

3.2 Introduction to Exploratory Data Analysis

Data Science Process



Problem

Determine Goal

Explore Data

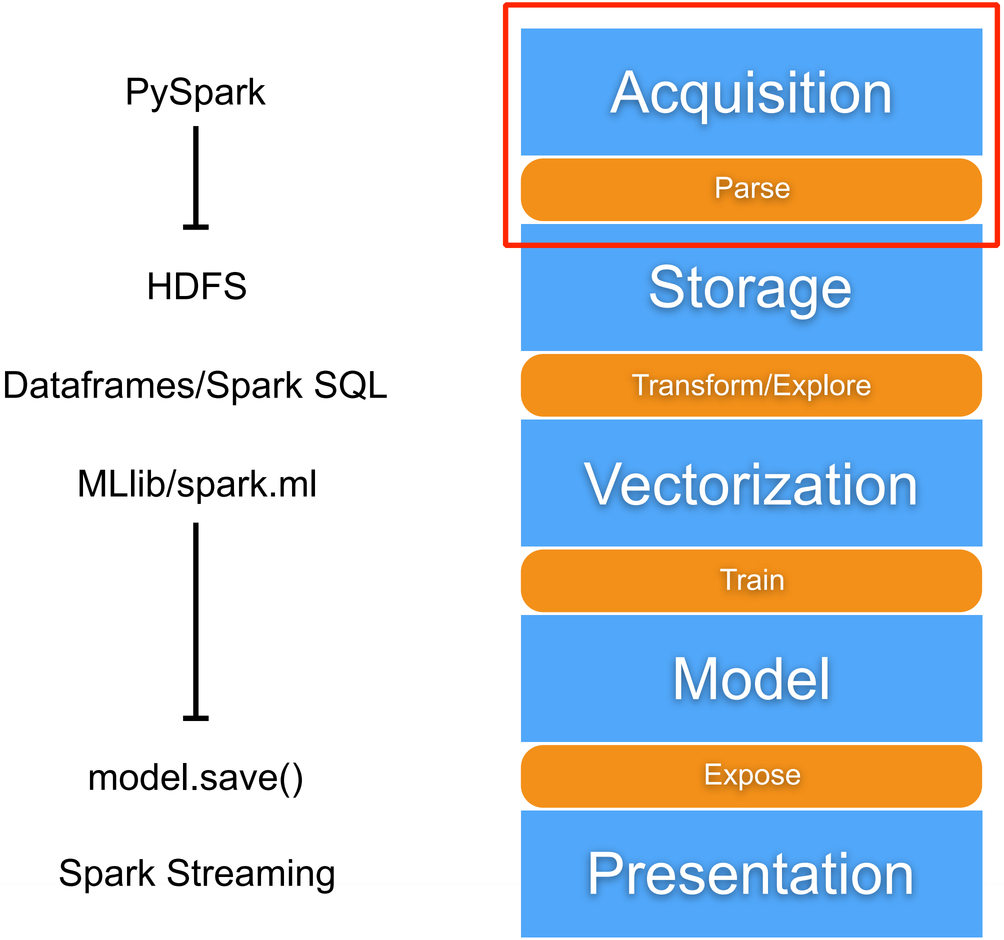
Propose Solutions

Collect Metrics

Evaluate

At Scale

Data Pipeline



PySpark

HDFS

Dataframes/Spark SQL

MLlib/spark.ml

model.save()

Spark Streaming

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

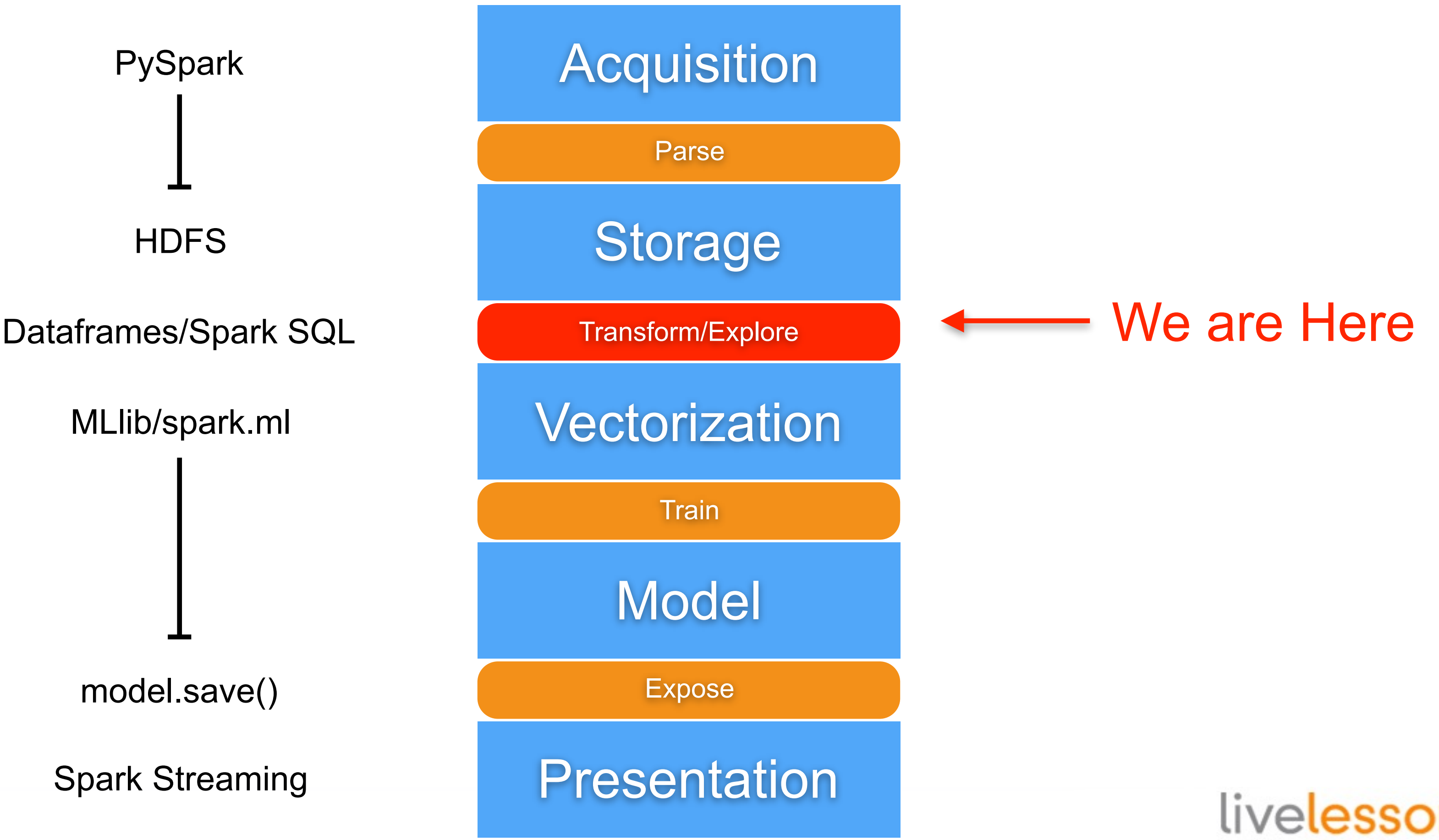
Model

Expose

Presentation

At Scale

Data Pipeline



PySpark

HDFS

Dataframes/Spark SQL

MLlib/spark.ml

model.save()

Spark Streaming

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

Model

Expose

Presentation

We are Here

What Is Exploratory Data Analysis?

> *But as much as EDA is a set of tools, it's also a mindset. And that mindset is about your relationship with the data… EDA happens between you and the data and isn't about proving anything to anyone else yet.*

- Cathy O'Neil (Doing Data Science)

What Is Exploratory Data Analysis?

- Developed at Bell Labs in the 1960's by John Tukey
- Techniques used to visualize and summarize data
    - Five-number summary: `describe()`
    - Distributions: box plots, stem and leaf, histogram, scatterplot

Goals of Exploratory Data Analysis

- Gain greater intuition
- Validate our data (consistency and completeness)
- Make comparisons between distributions
- Find outliers
- Treat missing data
- Summarize data (a statistic -> one number that represents many #'s)

How Can Spark Help?

- Interactive REPL
- Rapid computation (especially aggregates) on large amounts of data
- High level abstractions for querying data
- “Condense” data for easier local exploration and visualization