livelessons

video instruction from technology experts

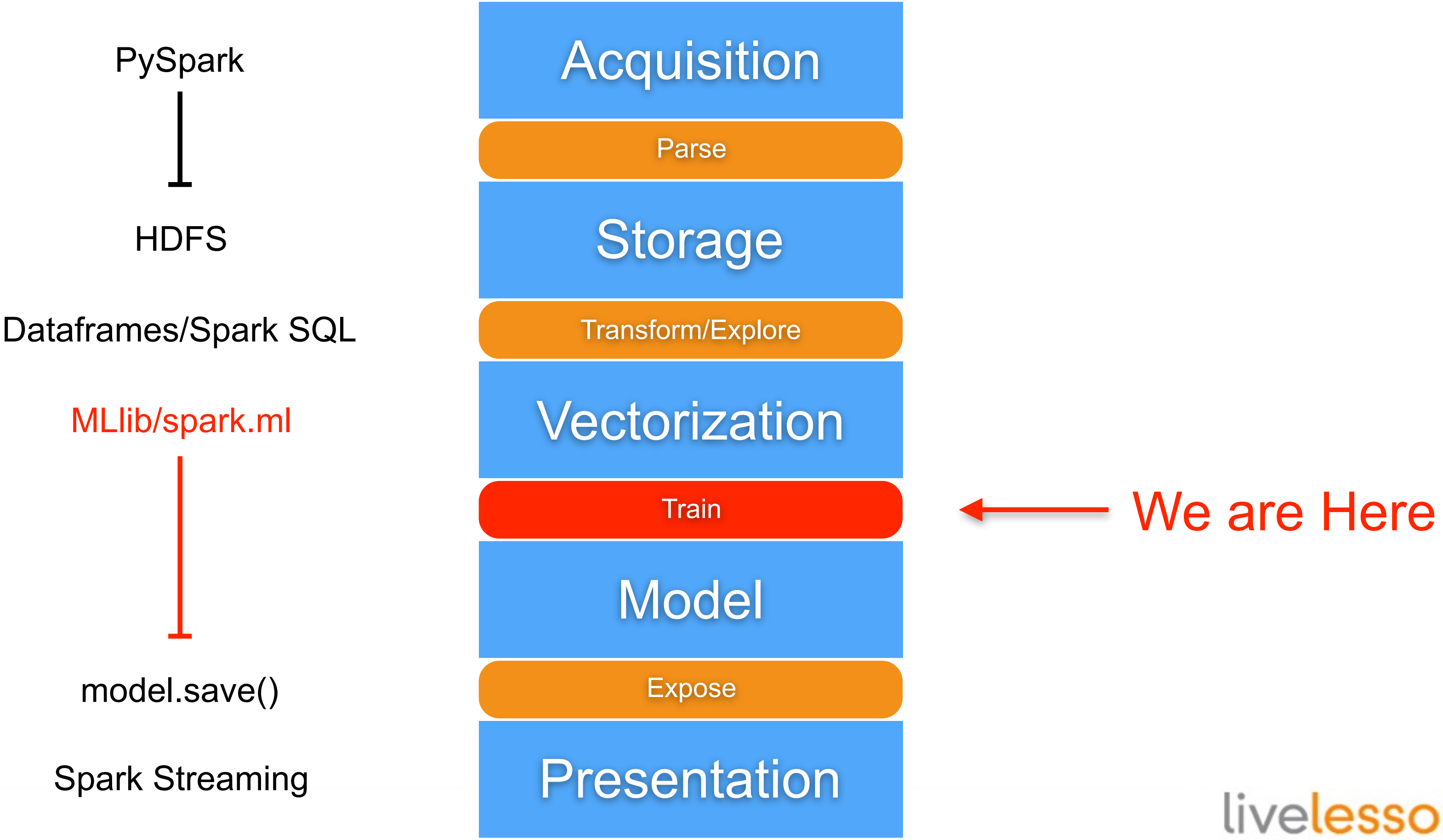
Lesson 3: Your First Spark Application

3.9 Introduction to Machine Learning

livelessons

At Scale

Data Pipeline



PySpark

HDFS

Dataframes/Spark SQL

MLlib/spark.ml

model.save()

Spark Streaming

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

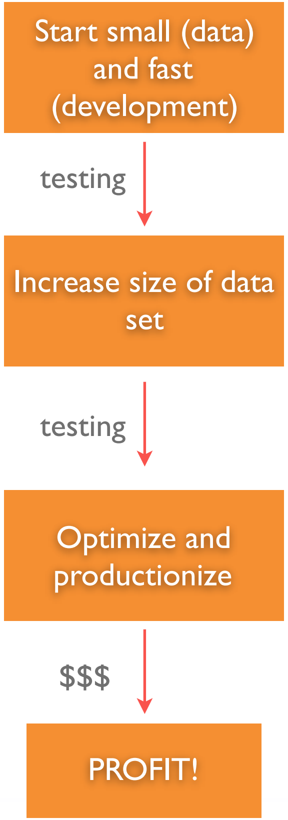
Model

Expose

Presentation

We are Here

Remember: How to Scale



Start small (data) and fast (development)

testing

Increase size of data set

testing

Optimize and productionize

$$$

PROFIT!

> Field of study that gives computers the ability to learn without being explicitly programmed.

-Arthur Samuel circa 1959

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

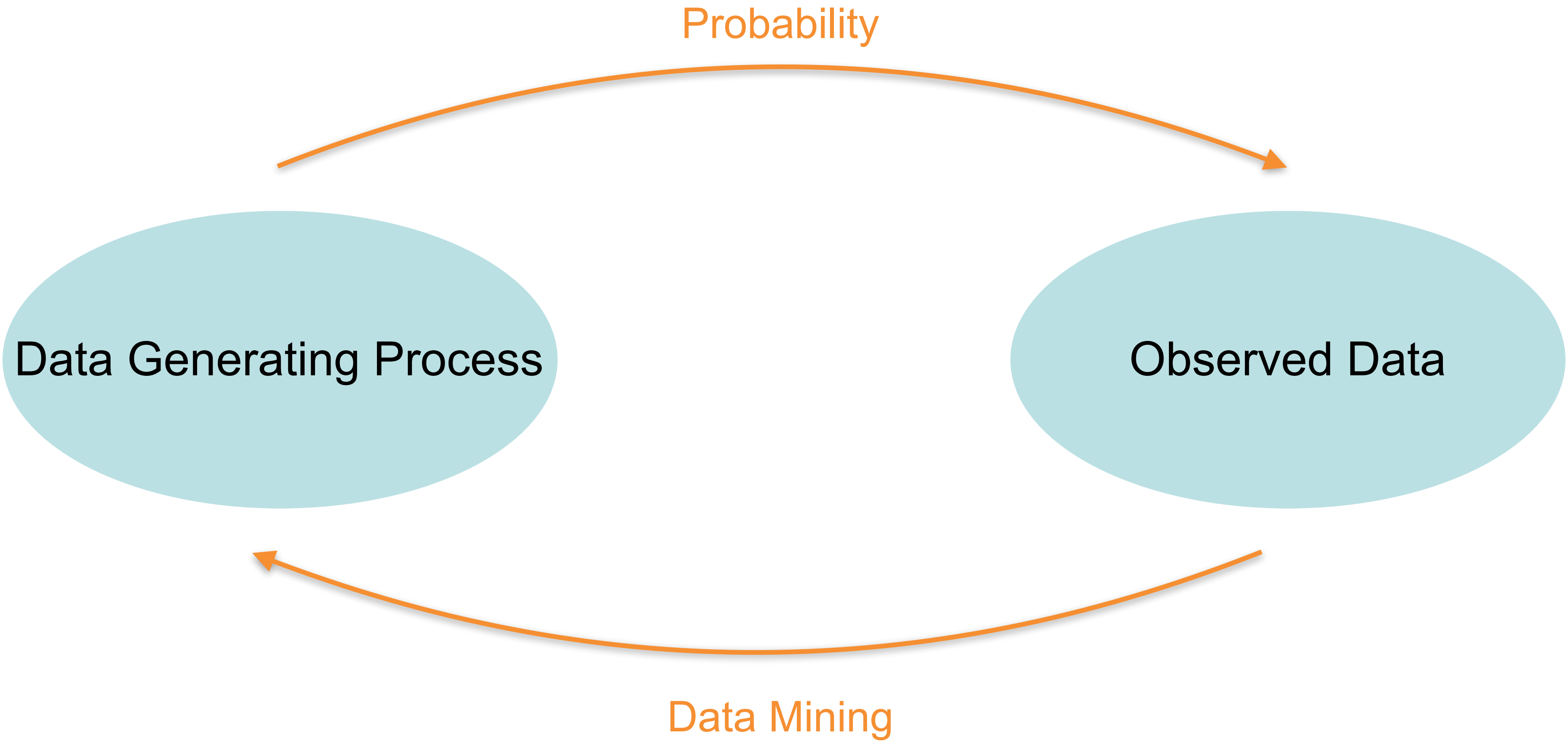
-Tom M. Mitchell

Machine learning is *NOT*:

- Hard coded logic by programmer: **if**s and **else**s...
- Predefined results: completely deterministic
- Burden is placed on programmer at design time
- Must anticipate all inputs to program, and react

Machine learning *is*:

- Automated knowledge acquisition through input
- Iterative improvement as more data is seen
- Adaptive Algorithms



Probability

Data Generating Process

Observed Data

Data Mining

Supervised Learning

- Training Data includes desired output

Unsupervised Learning

- Training Data does not include desired output

Semi-supervised Learning

- Training Data includes a few desired outputs

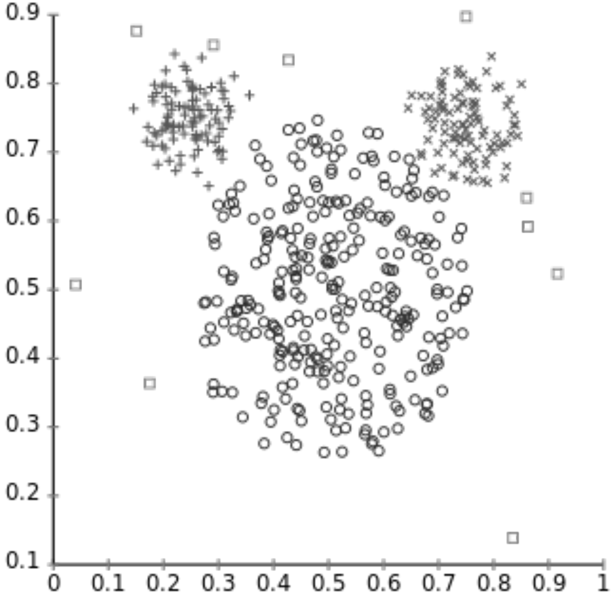
Reinforcement Learning

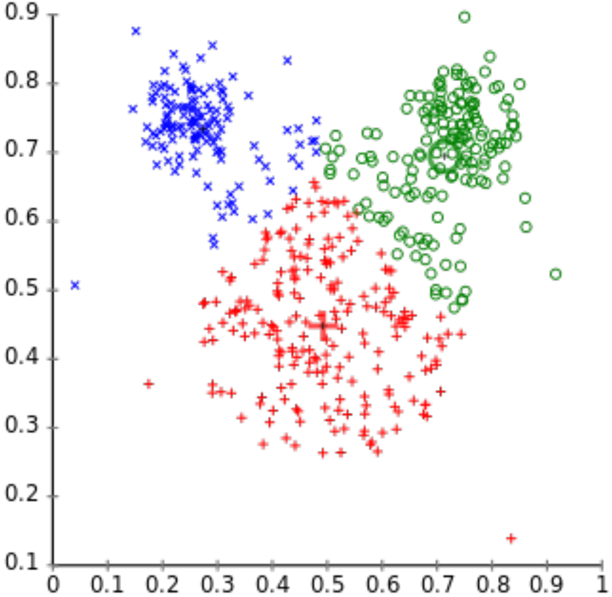
- Rewards from sequence of actions

Unsupervised Learning

- No need for labels
- Discovers latent features (hidden patterns in data)
- Often exploratory in nature
- Since there is no “gold standard” often difficult to validate model (especially with stochastic algorithms)

What Is Clustering?


Original Data


k-Means Clustering

How many subgroups?

What's considered similar?

How can we even do this?

livelessons

Clustering:

- Product Marketing: Cohort Analysis
- Oncology: Malignant cell identification
- Computer Vision: entity recognition
- Census: demographics analysis

See also: https://www.kaggle.com/wiki/DataScienceUseCases

k-means

- Choose initial centroids (randomly)
- Repeat until `num_iter` or convergence:
    - Assign each data point to closest centroid
    - Update centroids to be arithmetic mean of assigned points

k-means

http://stanford.edu/class/ee103/kmeans/kmeans.html