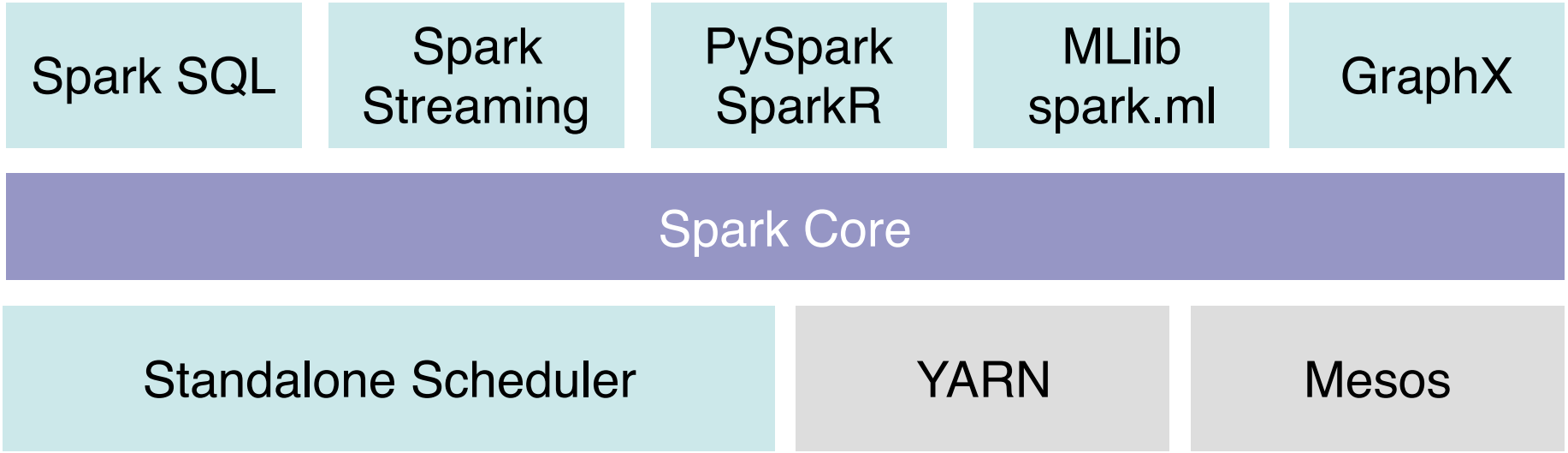
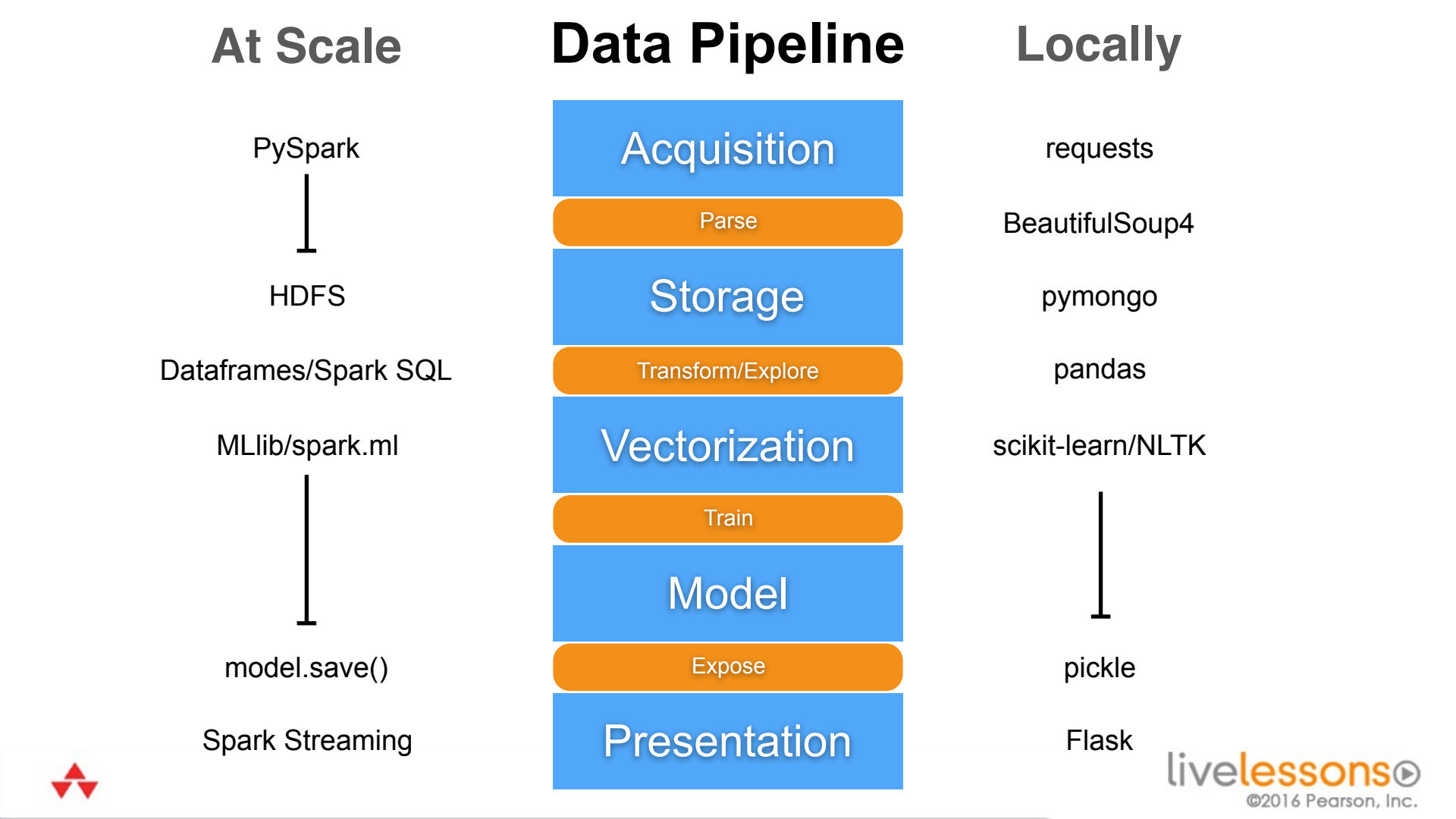
Summary

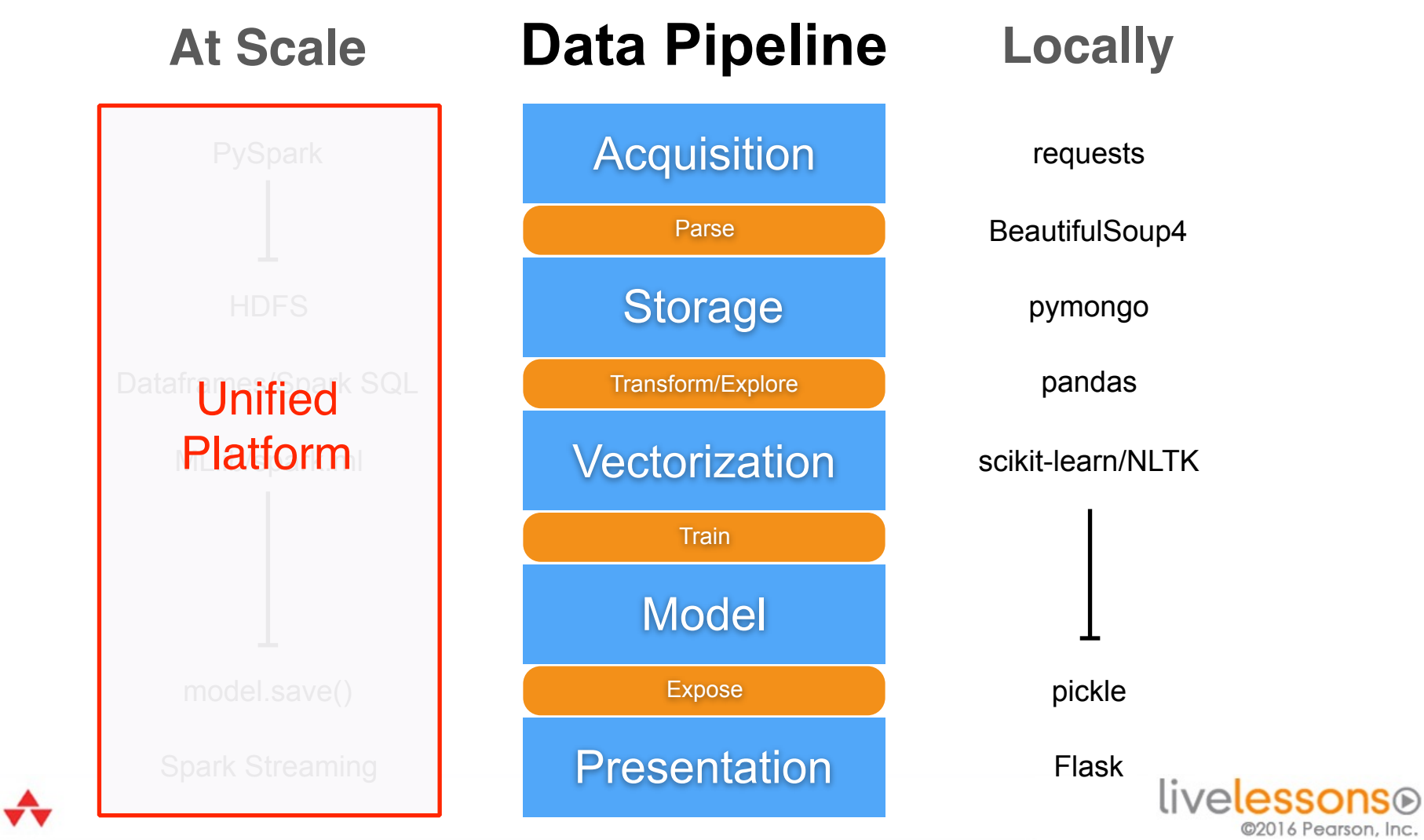
©2016 Pearson, Inc.

Unified Platform



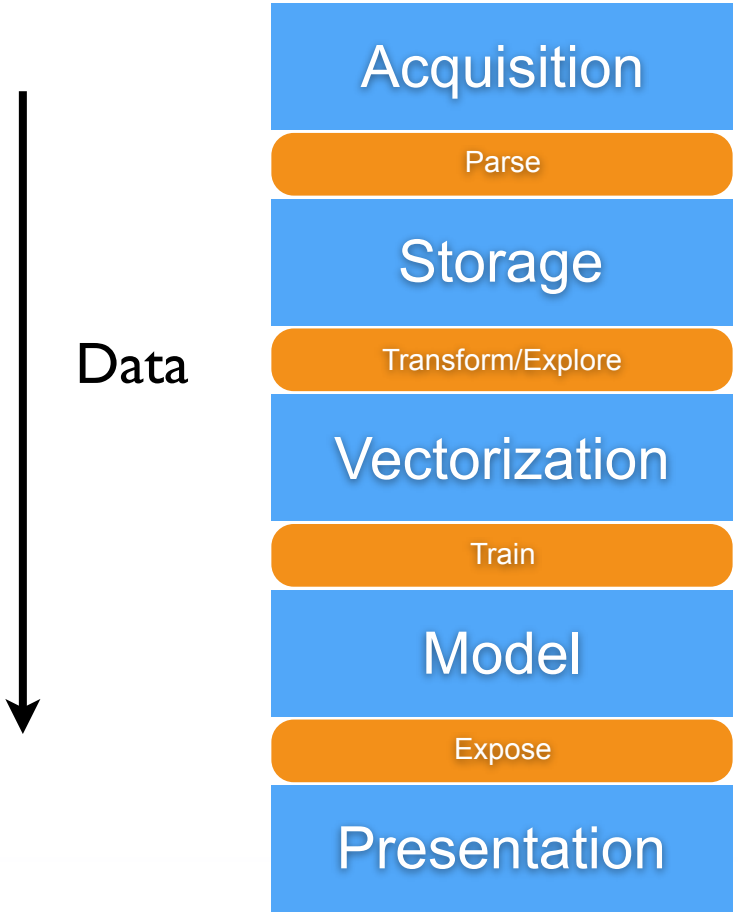
livelessons
©2016 Pearson, Inc.


At Scale
Data Pipeline
Locally
PySpark
HDFS
Dataframes/Spark SQL
MLlib/spark.ml
model.save()
Spark Streaming
Acquisition
Parse
Storage
Transform/Explore
Vectorization
Train
Model
Expose
Presentation
requests
BeautifulSoup4
pymongo
pandas
scikit-learn/NLTK
pickle
Flask
livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.

Data Pipeline



Immutable append only set of Raw Data

Remember to Remember (Lineage)

Computation is a view on data

livelessons
©2016 Pearson, Inc.

Remember: How to Scale

livelessons
©2016 Pearson, Inc.

Thanks for Watching!

- **Code:** http://bitly.com/galvanize-spark-workshop
- **Resources:** http://galvanize.com/resources/spark
- **Forum:** http://bitly.com/galvanize-spark-forum

- **Data:**
  - Donors Choose: http://data.donorschoose.org/open-data/overview/
  - KDD Cup 2009: http://kdd.org/kdd-cup/view/kdd-cup-2009/Data

livelessons
©2016 Pearson, Inc.