

Data Engineering for Data Science

Sharon Xu
February 2020



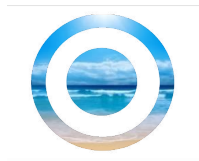
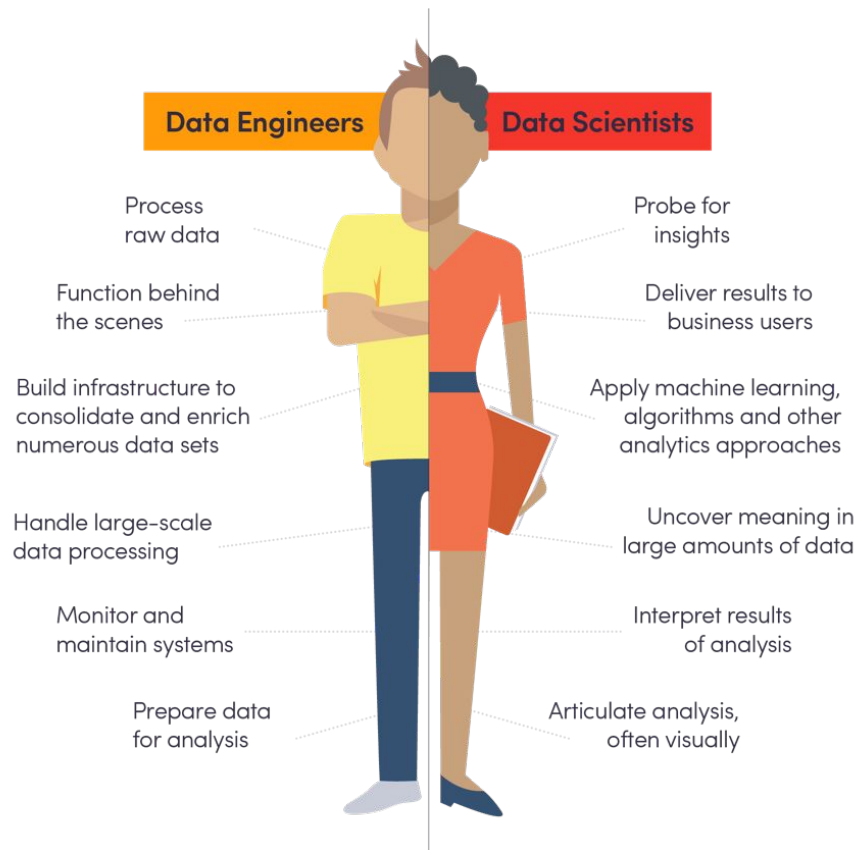


Table of Contents

Workshop Agenda

1. Data Engineering at TripAdvisor
2. MapReduce, Hadoop, and Spark
3. Hands-on Exercise

Scientists “vs.” Engineers



Engineering ↔ Modeling

As data scientists we should...

- Understand ETLs and how features are created. Need to be aware of bias and noise that may be introduced.
- Consider real-time complexity when choosing features and building models.

The background of the slide is a wide-angle photograph of a desert landscape. In the foreground and middle ground, there are several prominent, jagged red rock formations. The most notable one on the right is a tall, craggy spire. To its left, there are more rounded rock formations. The sky is filled with large, white and grey clouds, and the sun is low on the horizon, casting a warm, golden light across the scene. The overall mood is majestic and serene.

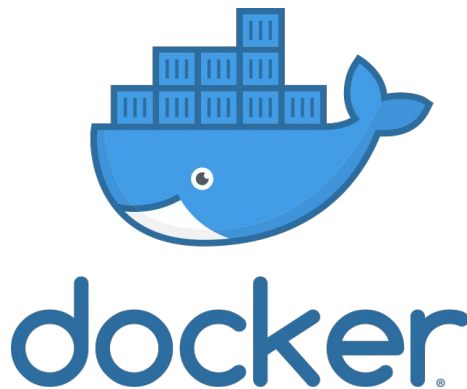
Data Engineering at TripAdvisor



The Tools



Java

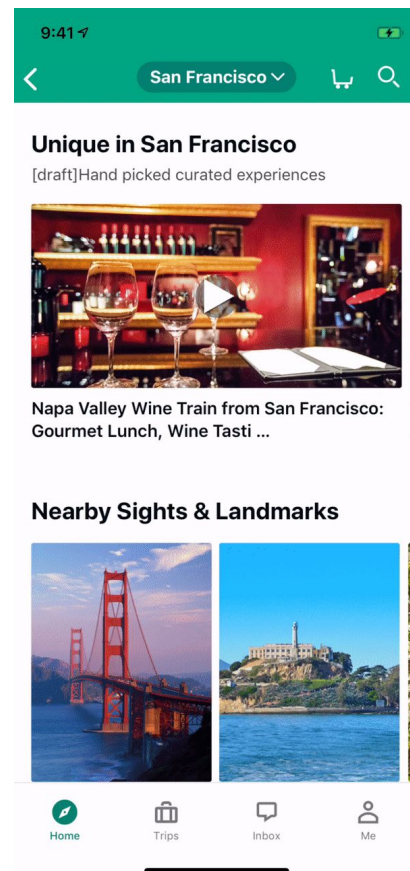


What We Work On

A diverse set of applications

- Learn to rank models
- Recommender systems
- Machine vision
- Search results ranking
- Causal inference
- Review fraud detection
- Bot detection
- Booking / content models
- Automated collection generation
- Keyphrase generation
- Location classification

Blog: <https://www.tripadvisor.com/engineering/tag/machine-learning>

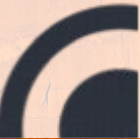


Questions?



The background image is a wide-angle shot of a desert canyon. In the foreground and midground, there are several prominent red rock formations. On the right, a large, craggy rock face rises steeply, with a smaller, more rounded rock formation to its left. Further left, another large rock formation is visible. The sky is filled with soft, white and grey clouds, and the sun is low on the horizon, casting a warm, golden light across the scene. The overall mood is serene and majestic.

Spark: A Hands-On Exercise



Setup

1. Clone the repository

```
git clone https://github.com/sharonxu/data-engineering-tufts.git
```

2. Open the Colab Notebook : [Data Exploration with Spark](#)
3. Upload data.zip to the notebook
4. Run cells up to “Step 1 - What inputs do you have?”

MapReduce, Hadoop, and Spark



MapReduce

Other Data
Processing
Frameworks

YARN

Resource Management

HDFS

Distributed File Storage



Spark SQL +
DataFrames

Streaming

MLlib
Machine Learning

GraphX
*Graph
Computation*

Spark Core API

R

SQL

Python

Scala

Java

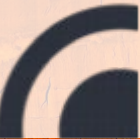


Time to code!



A wide-angle photograph of a desert landscape featuring prominent red rock formations. The sky is filled with large, white and grey clouds, with a warm, golden light from the sun low on the horizon. The rock formations are rugged and layered, with some showing distinct horizontal strata. A semi-transparent white banner is overlaid across the middle of the image, containing the text "What Next?".

What Next?



Resources and Next Steps

Data Engineering

- Learn SQL ([SQLZOO](#))
- [Big Data Analysis with Apache Spark](#)
- Play around with an API (like [Spotify's](#))

Data Science

- [RecSys Challenge](#)
- [Google's Dataset Search](#)
- [Kaggle](#)

Cool Colab Notebooks

- [Intro Notebook with ML Resources](#)
- [Language Models from N-grams to GPT2](#) ([Harvard ComputeFest 2020](#))
- [Solutions for our PySpark Exercise](#)

Thank You!

Questions?

LinkedIn: <https://www.linkedin.com/in/sharon6>

Github: <https://github.com/sharonxu>