Patterns for Successful Data Science Projects

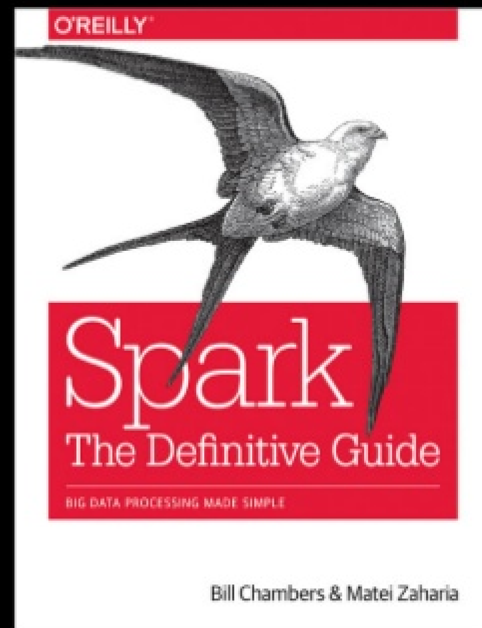
Bill Chambers
@bllchmbrs 2018-04-24

databricks

Introductions

About Me

databricks



About you

- Data Scientists?
- Data engineers?
- Data team leads?

The Context of Your Org/Team

Scoping Initiatives

- Your company is scoping ML initiatives right now, with little (if any) ML in production.

Looking to Grow

- Your company has a dozen or so models in production, but now you want to scale to hundreds/ thousands in the next year.

6 Patterns for DS Projects

Organizational Patterns

- Value
- Alignment
- Discipline

Technical Patterns

- Hierarchy of Needs
- Simple
- Track

Deep dive into each pattern and apply it to data science projects

databricks

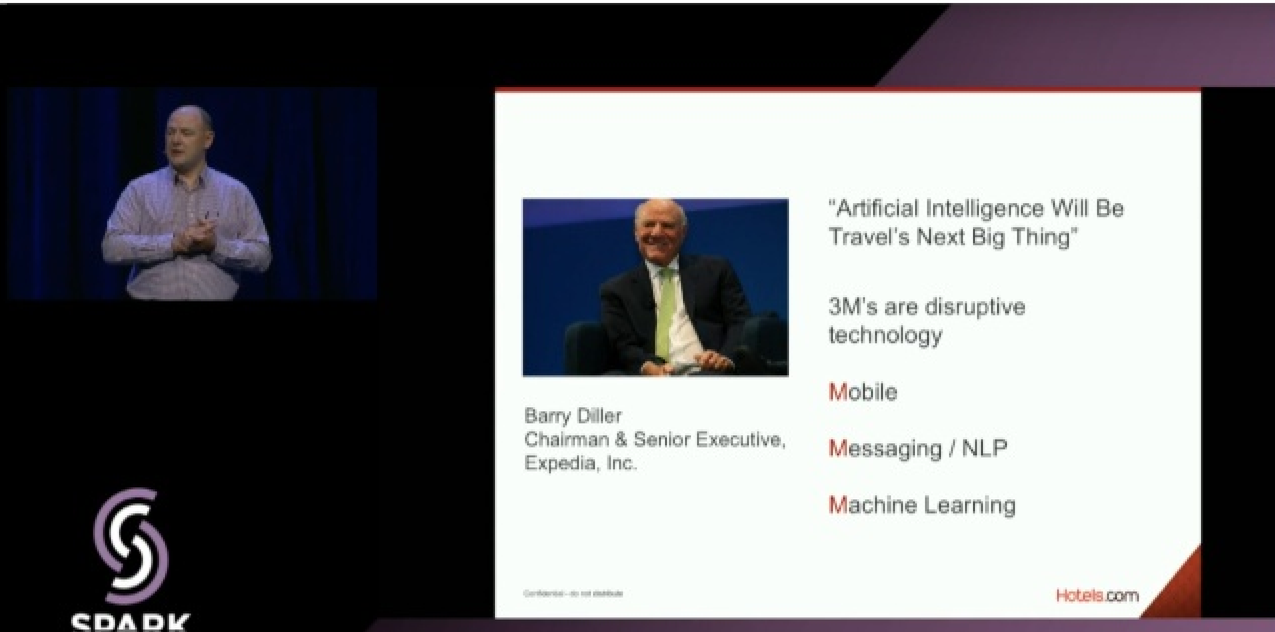
Organizational Patterns in Data Science Projects

Value

n. the regard that something is held to deserve; the importance, worth, or usefulness of something

"Data science isn't woven into our culture; it is our culture. We started with it at the heart of the business, rather than adding it to a traditional organizational structure, and built the company's algorithms around our clients and their needs."

\- HBR



"Having senior level support is very valuable. Our CEO in particular is a great supporter of machine learning and sees it as a fundamental part of our future."

- Matt Fryer, Chief Data Science Officer, Hotels.com

Alignment

n. arrangement in a straight line, or in correct or appropriate relative positions

n. a position of agreement or alliance

Theory (job description)

- PhD in Computer Science, Computer Engineering, Mathematics...
- 5 years of real world or research experience in data science
- Experience with Big Data technologies such as Hadoop, Cassandra etc.
- Experience in model development and life-cycle-management
- Programming skills in various languages (C++, Scala, Java, R) with proficiency in Python and/or C++
- Understanding of Machine Learning, e.g.: linear/logistics regression discriminant analysis, bagging, random forest, SVM, neural nets
- Knowledge and skills in the use of current state of the art machine learning frameworks such as Scikit-Learn, H2O, Keras, TensorFlow and Spark, etc.

databricks

Practice (on the job)

- How many daily active users do we have?
- What's our monthly churn?
- How many people are using _______ feature?
- Can you build a data pipeline?
- We're considering A/B testing, can you write up a report on it?

databricks

Alignment in the context of DS projects

- The project is formally prioritized
  - Funded and staffed appropriately
- You have the infrastructural resources to achieve the mission
- You have runway and cover from your leadership to get where you need to go

*The organization has **alignment** on the **value** data science provides.*

databricks

Discipline

n. a rule or system of rules governing conduct or activity


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

databricks

Discipline in data science…

- Figure out what you're going to do and execute at a high standard.
- focus on the results, not just on the tasks.
- Define phases and demonstrate results along the way.
- Don't just stir the data to get the answer you want.

*Make **data science** a **discipline**.*

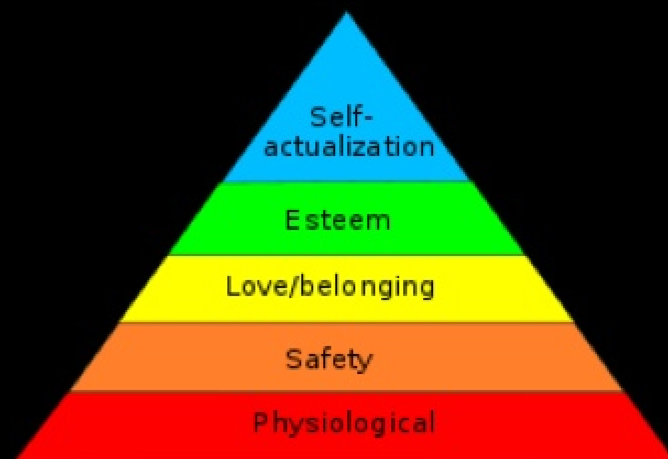
databricks

Technical Patterns in Data Science Projects

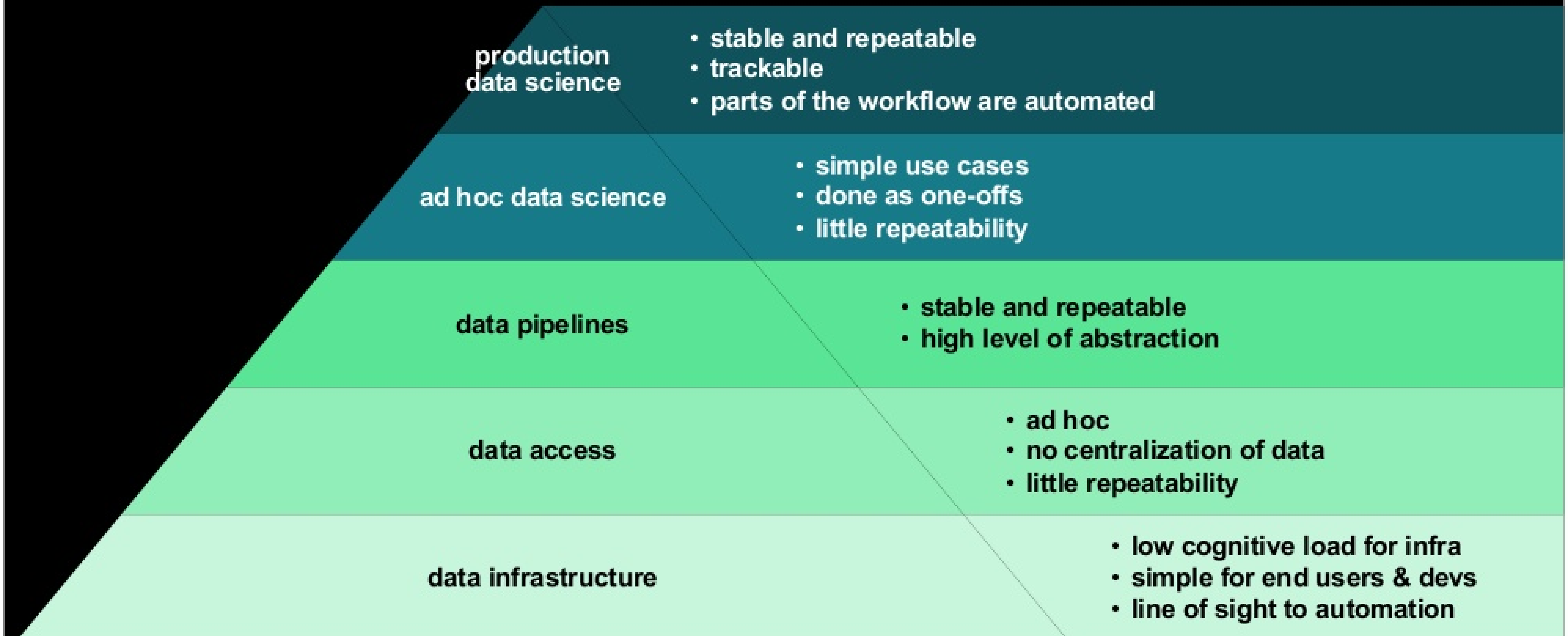
databricks

Maslow's Hierarchy of Needs

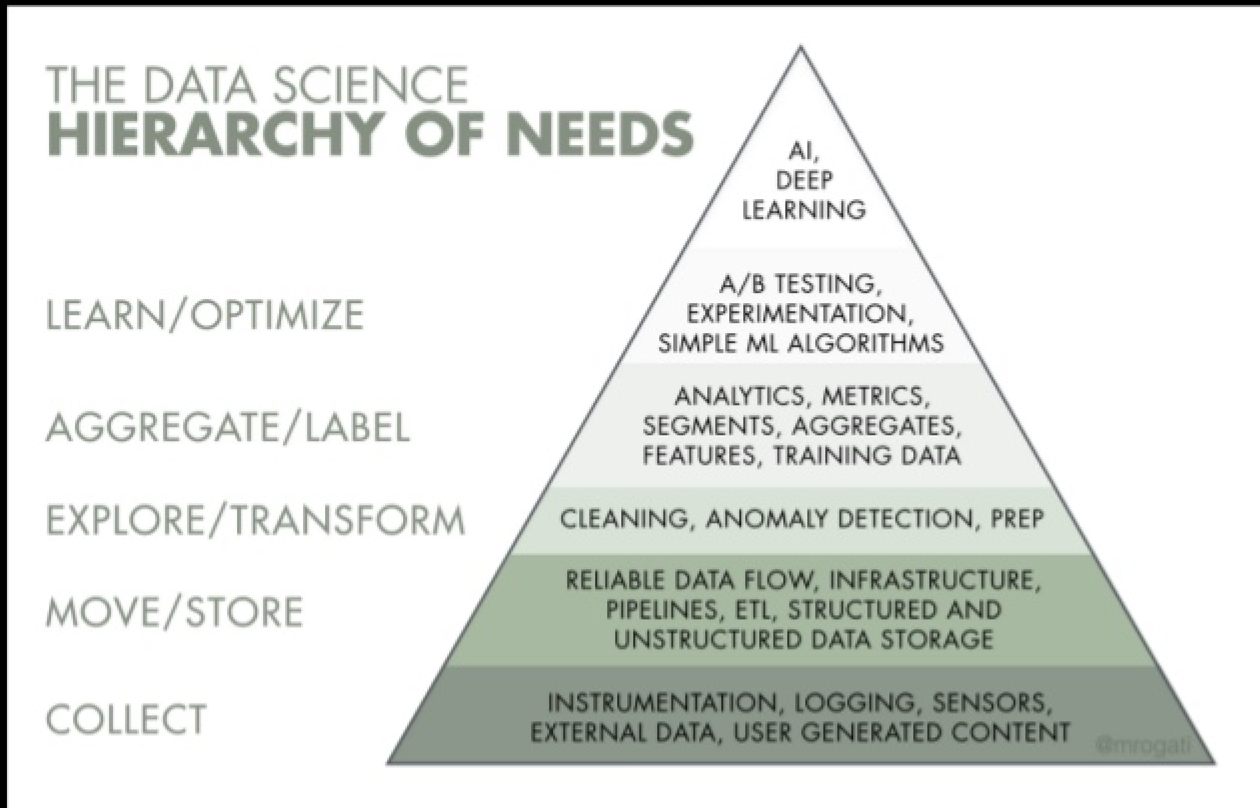
- **Defines a theory for human motivation (Abraham Maslow, 1943)**
- **Each base in the pyramid must be supported before one can move onto the next**



Data-Driven Company Hierarchy of Needs

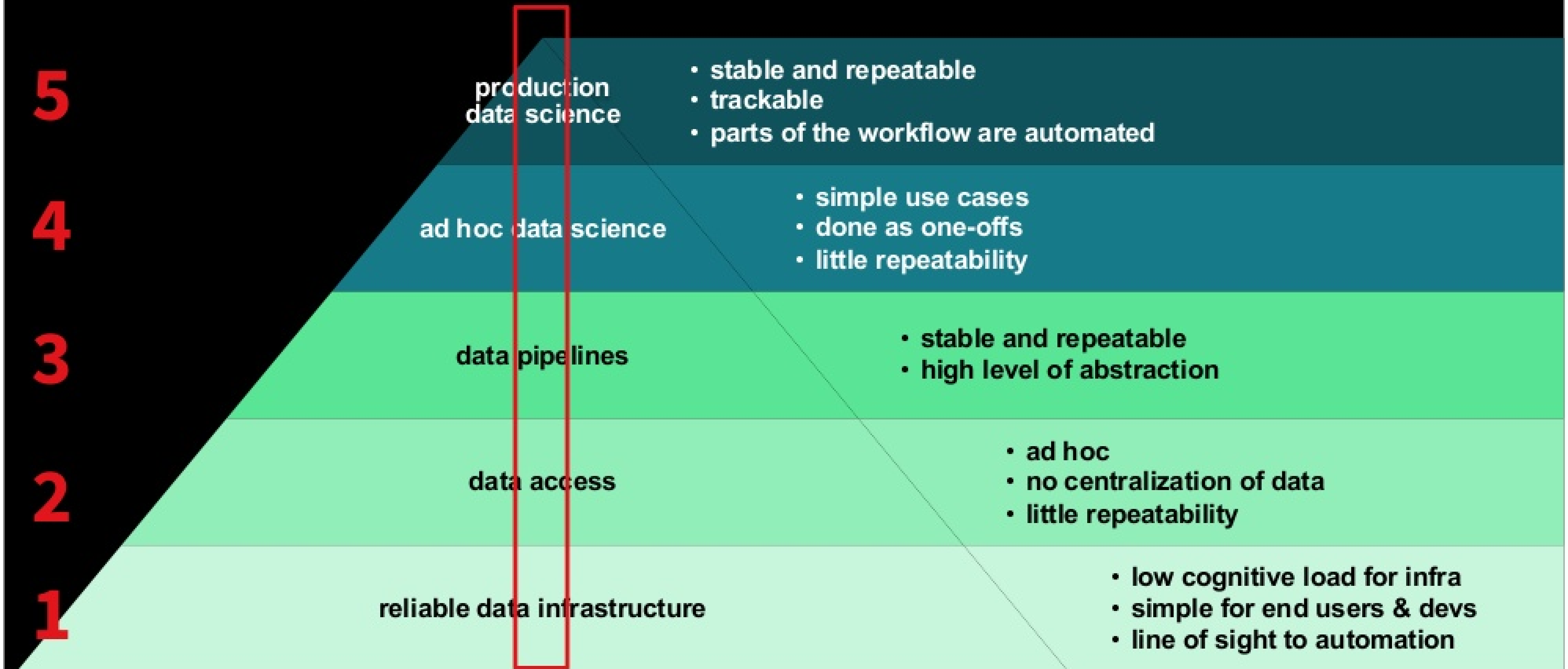


Data Science Hierarchy of Needs



Source - Monica Rogati

How to approach this pyramid?



ML-System Anti-Patterns

- **Glue Code**
  - Lots of glue code to tie OSS/generic components together.
- **Pipeline Jungles**
  - When pipelines evolve *organically*, they can become hard to maintain.
- **Abstraction Debt**
  - A general problem in ML, lots of different abstractions.

"Hidden Technical Debt in Machine Learning Systems", Google NIPS 2015

Simple

n. plain, basic, or uncomplicated in form, nature, or design; without much decoration or ornamentation

KISS Principle

Keep it simple stupid

F-117 Nighthawk

U2 Spy Plane

SR-71 Blackbird

Kelly Johnson
1910-1990

Keeping it Simple in ML

[Rules of Machine Learning: Best Practices for ML Engineering]

- Martin Zinkevich, Research Scientist @ Google

Rule #1: (Before Machine Learning)

Don't be afraid to launch a product without machine learning

Rule #4: (Your First Pipeline)

Keep the first model **SIMPLE** and get the infrastructure right

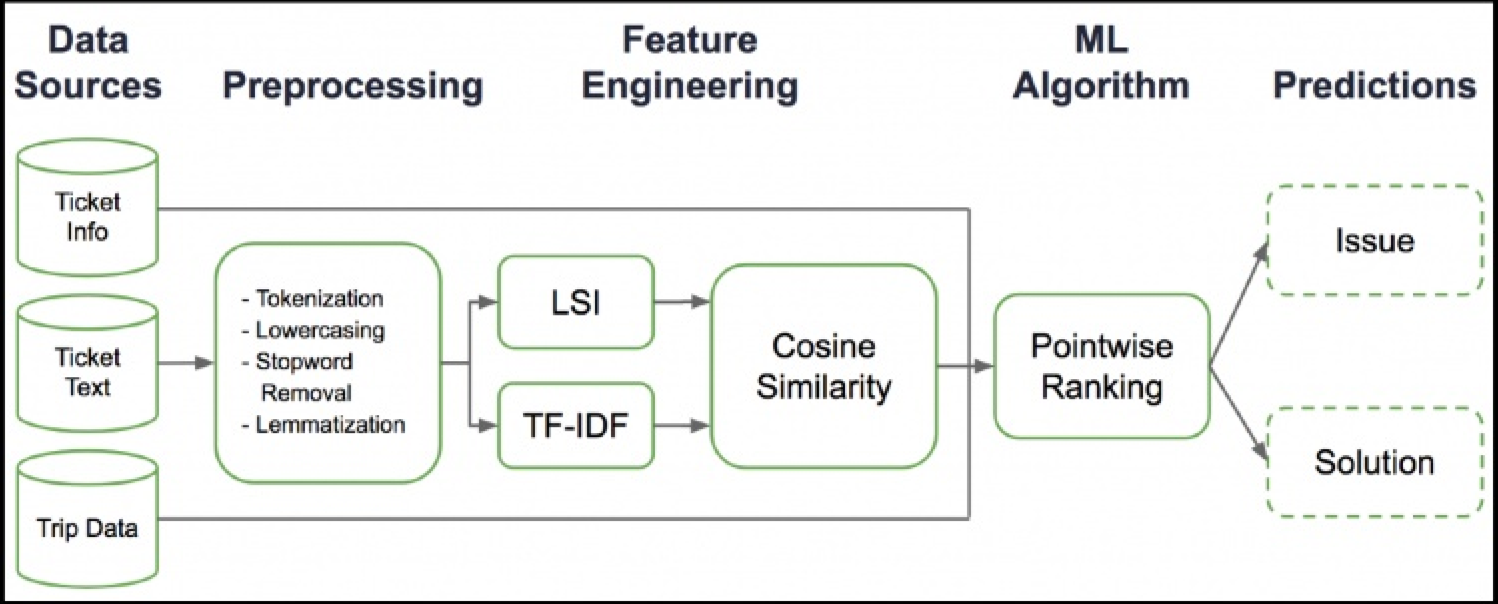
Rules of Machine Learning: Best Practices for ML Engineering

Martin Zinkevich

This document is intended to help those with a basic knowledge of machine learning get the benefit of best practices in machine learning from around Google. It presents a style for machine learning, similar to the Google C++ Style Guide and other popular guides to practical programming. If you have taken a class in machine learning, or built or worked on a machine-learned model, then you have the necessary background to read this document.

Example:

COTA: Improving Uber Customer Care with NLP & Machine Learning



Track

n. the act or process of following something or someone

n. Precise and continuous position-finding of targets by radar, optical, or other means.

Hardest part of ML Systems isn't ML

"Hidden Technical Debt in Machine Learning Systems", Google NIPS 2015

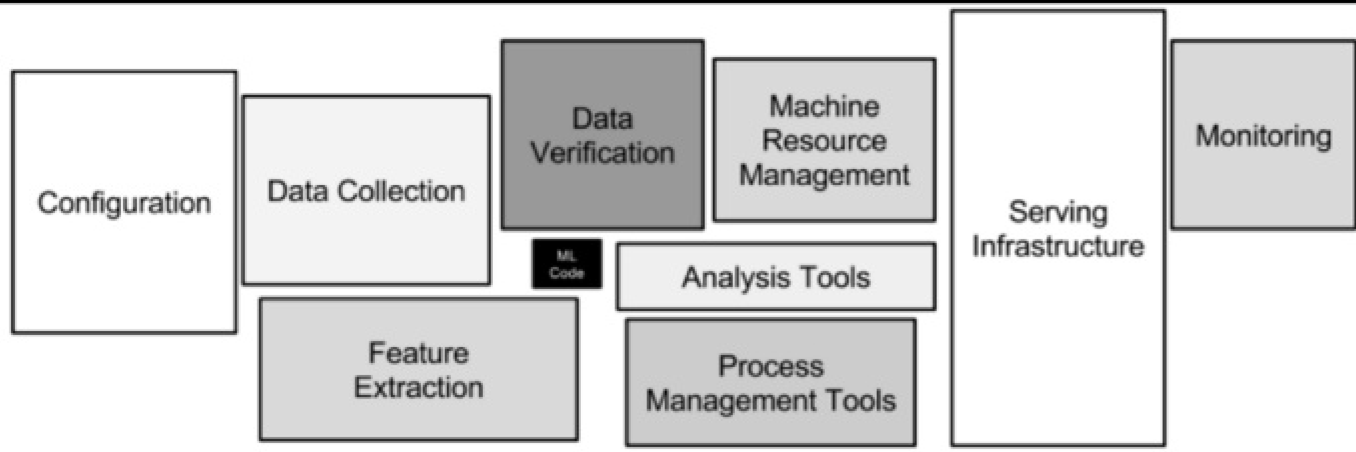


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

MLflow Components

mlflow Tracking

Record and query experiments: code, data, config, results

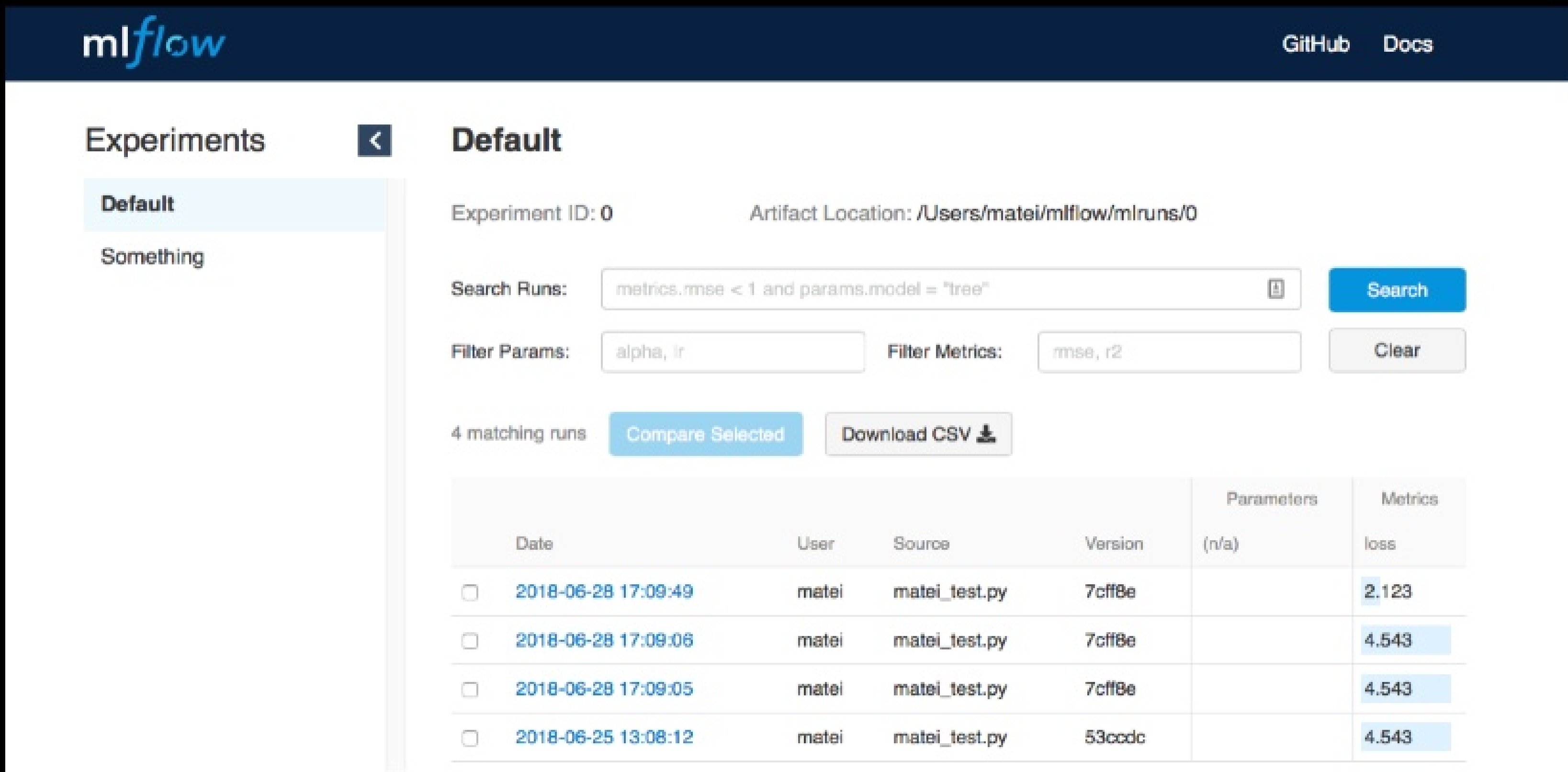
mlflow Projects

Packaging format for reproducible runs on any platform

mlflow Models

General model format that supports diverse deployment tools

MLflow Tracking



6 Patterns for DS Projects

Organizational Patterns

- Value
- Alignment
- Discipline

Technical Patterns

- Hierarchy of Needs
- Simplify
- Track

databricks

Thank you