O'REILLY® OSCON

Data Science in Production

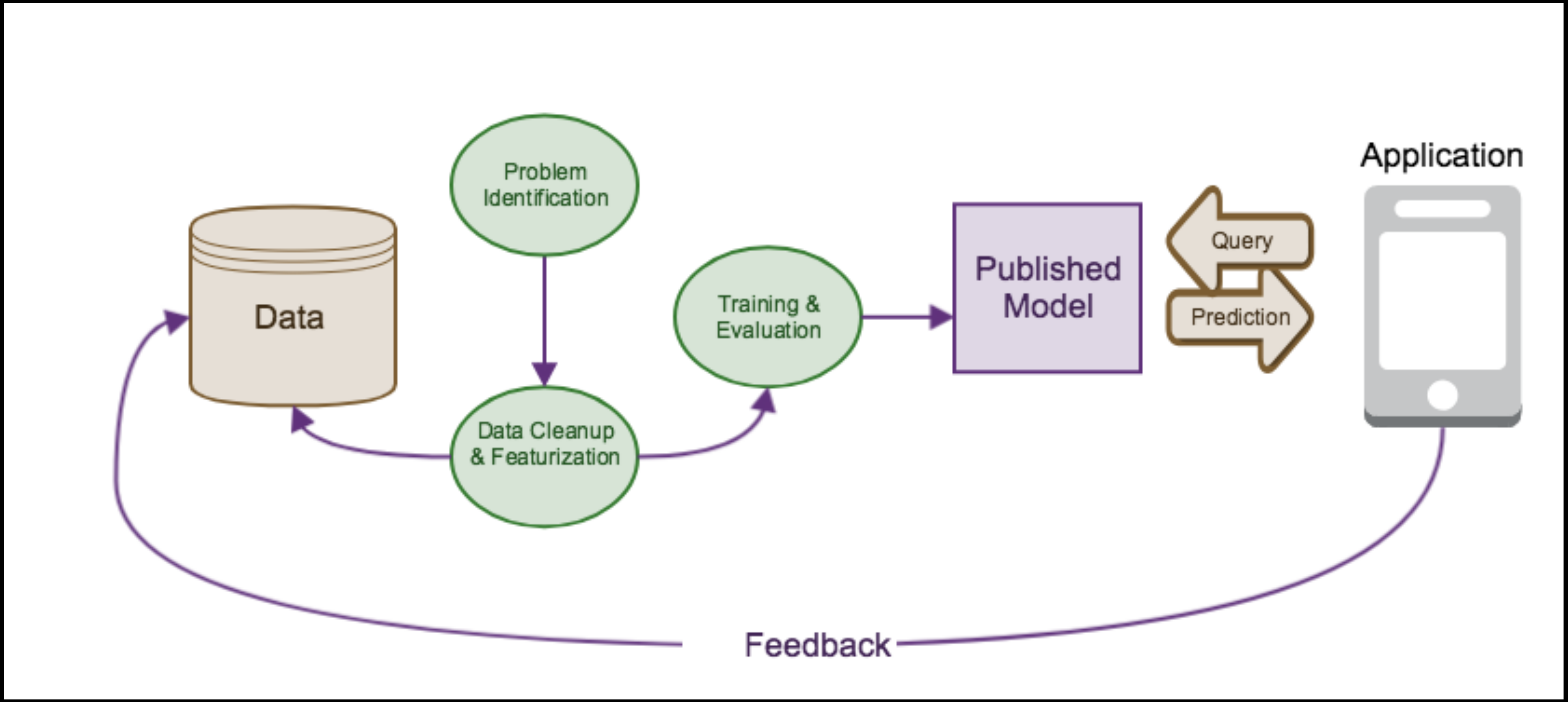
Richa Khandelwal

Sr. Software Engineering Manager @ Nike

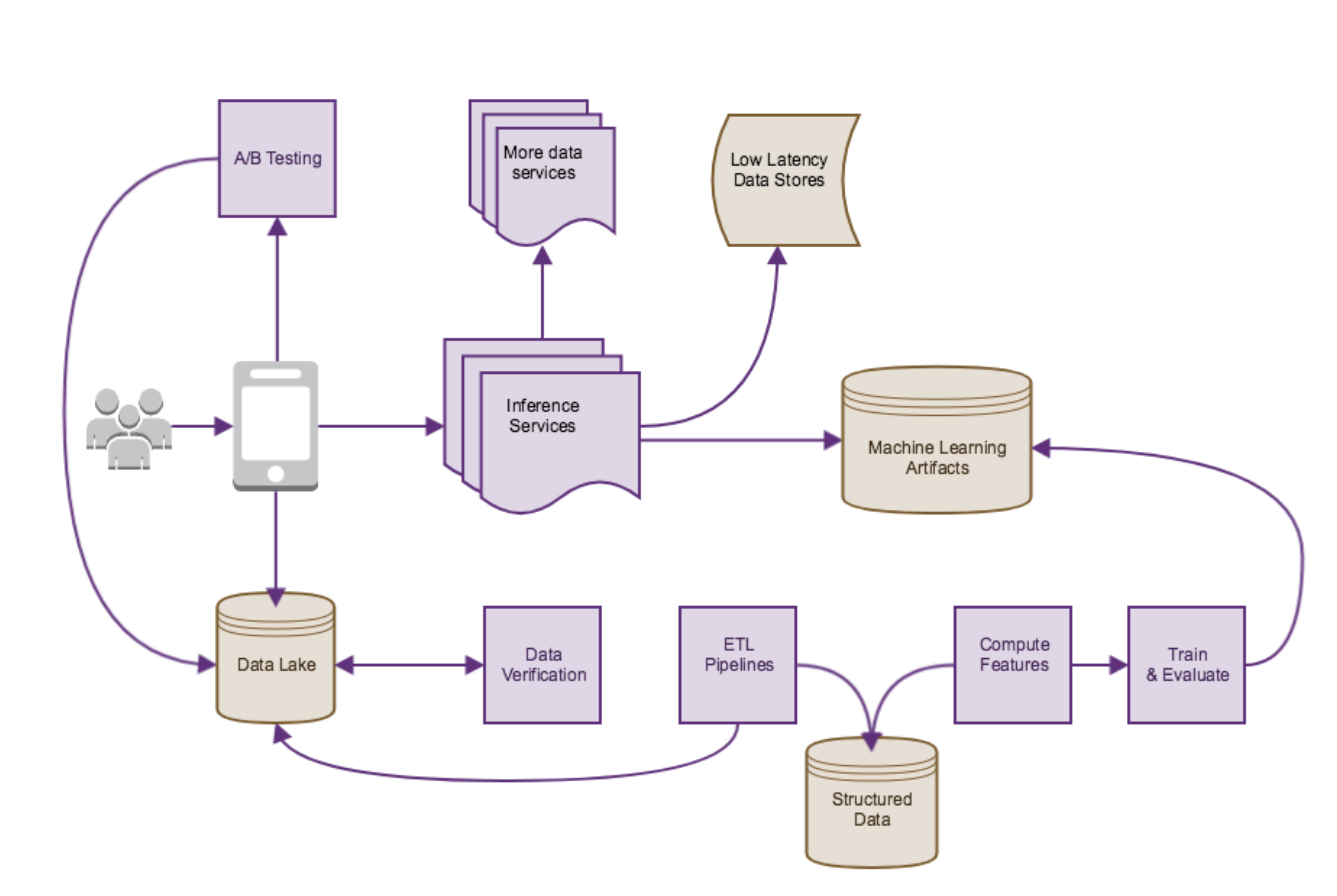
https://www.richakhandelwal.com/

@ri_cha_k

Data Science Solution Lifecycle

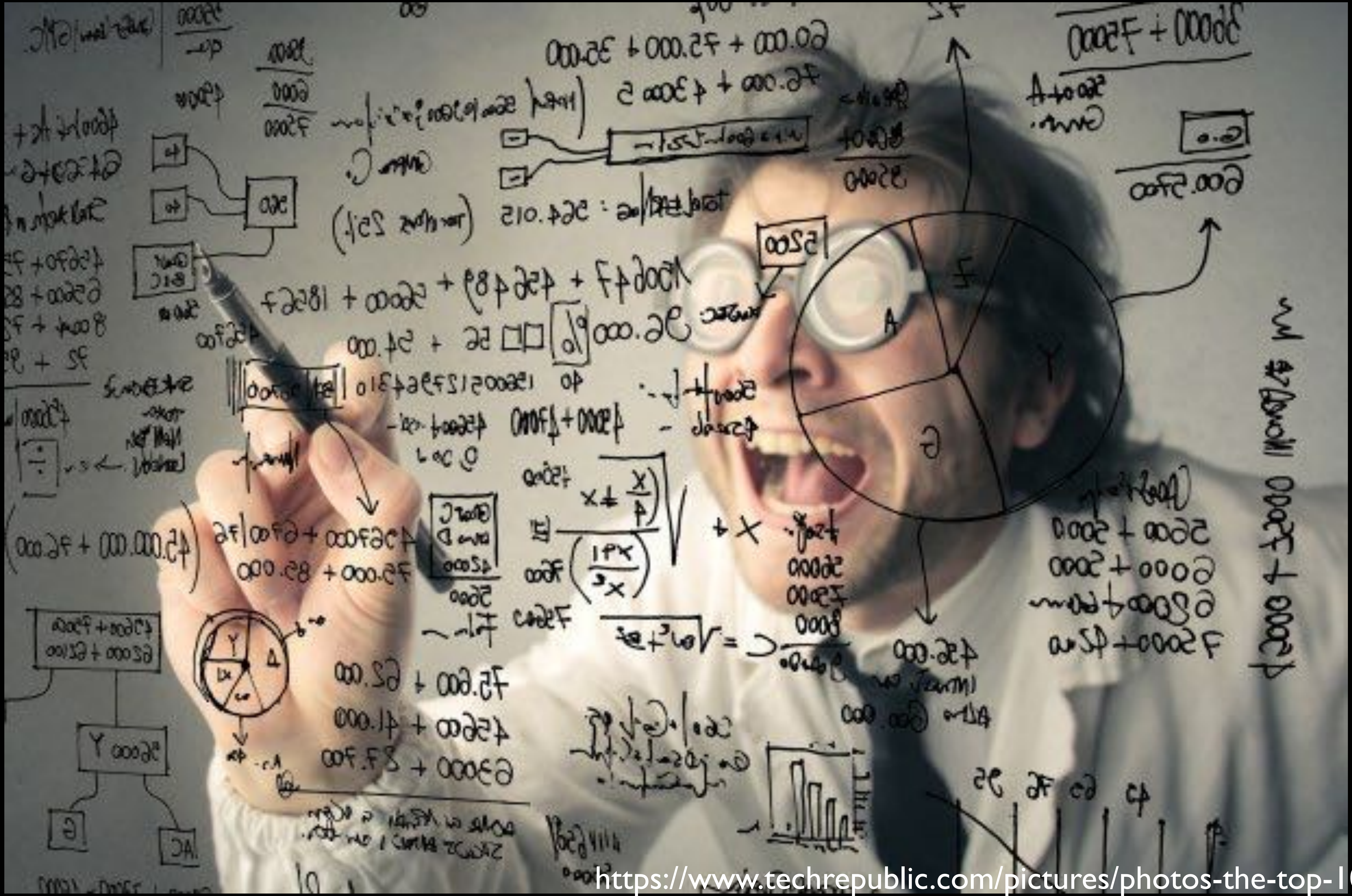


Architecture



What is the Problem Then? Why Are We Here?

Because Things Data Scientists Say

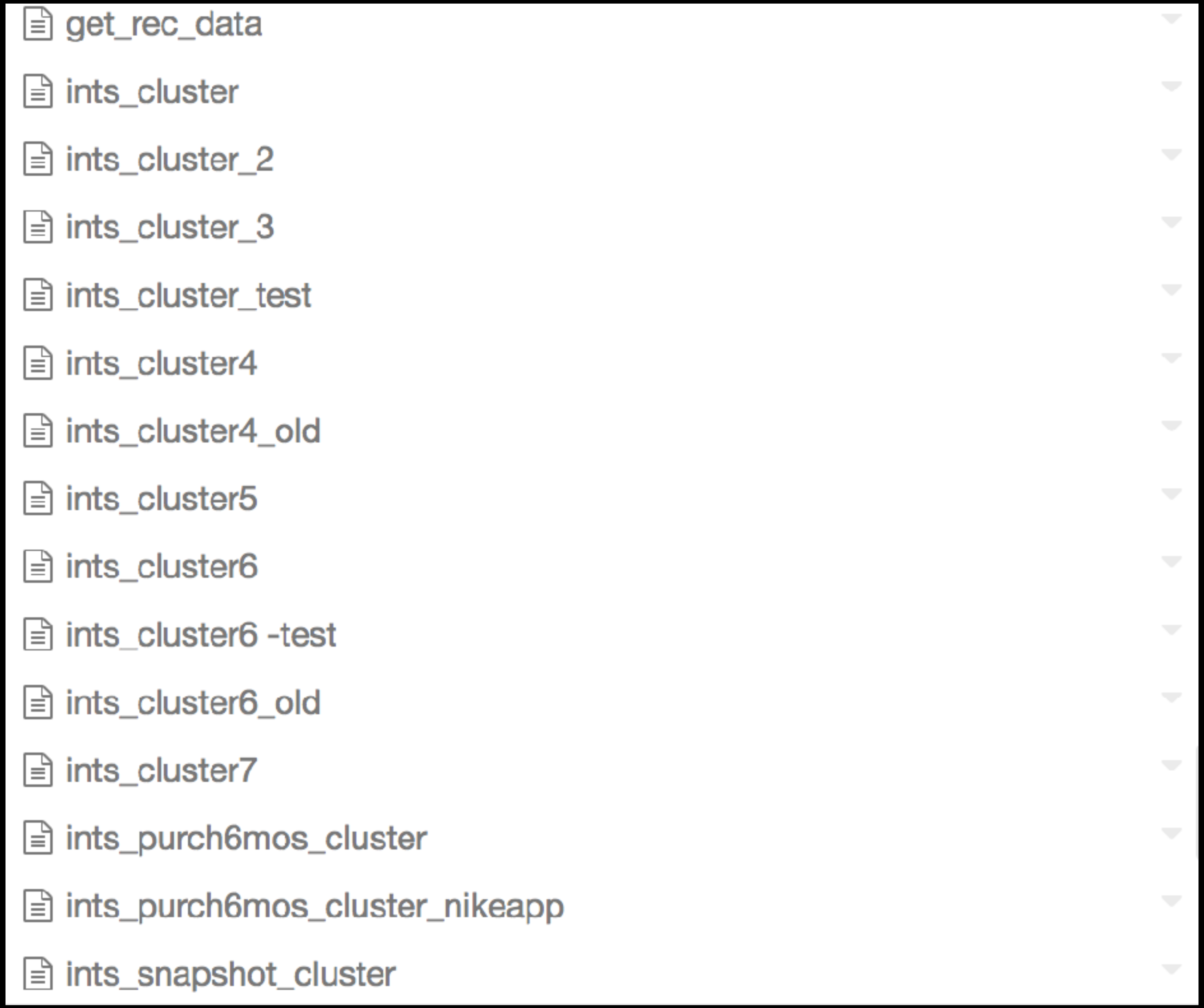


https://www.techrepublic.com/pictures/photos-the-top-10-universities-for-data-science/

Say What?

- My model is ready for production. It is writing results on awesomescientist1/exp-1021/adssgd/result_1096
- Git? What's that?
- What's JIRA?
- What tests?
- I recorded it in a spreadsheet that is saved on my machine
- But this works on my machine/cluster
- Why can't I have all the data? What's GDPR?

Because Code Looks Like This



get_rec_data

ints_cluster

ints_cluster_2

ints_cluster_3

ints_cluster_test

ints_cluster4

ints_cluster4_old

ints_cluster5

ints_cluster6

ints_cluster6 -test

ints_cluster6_old

ints_cluster7

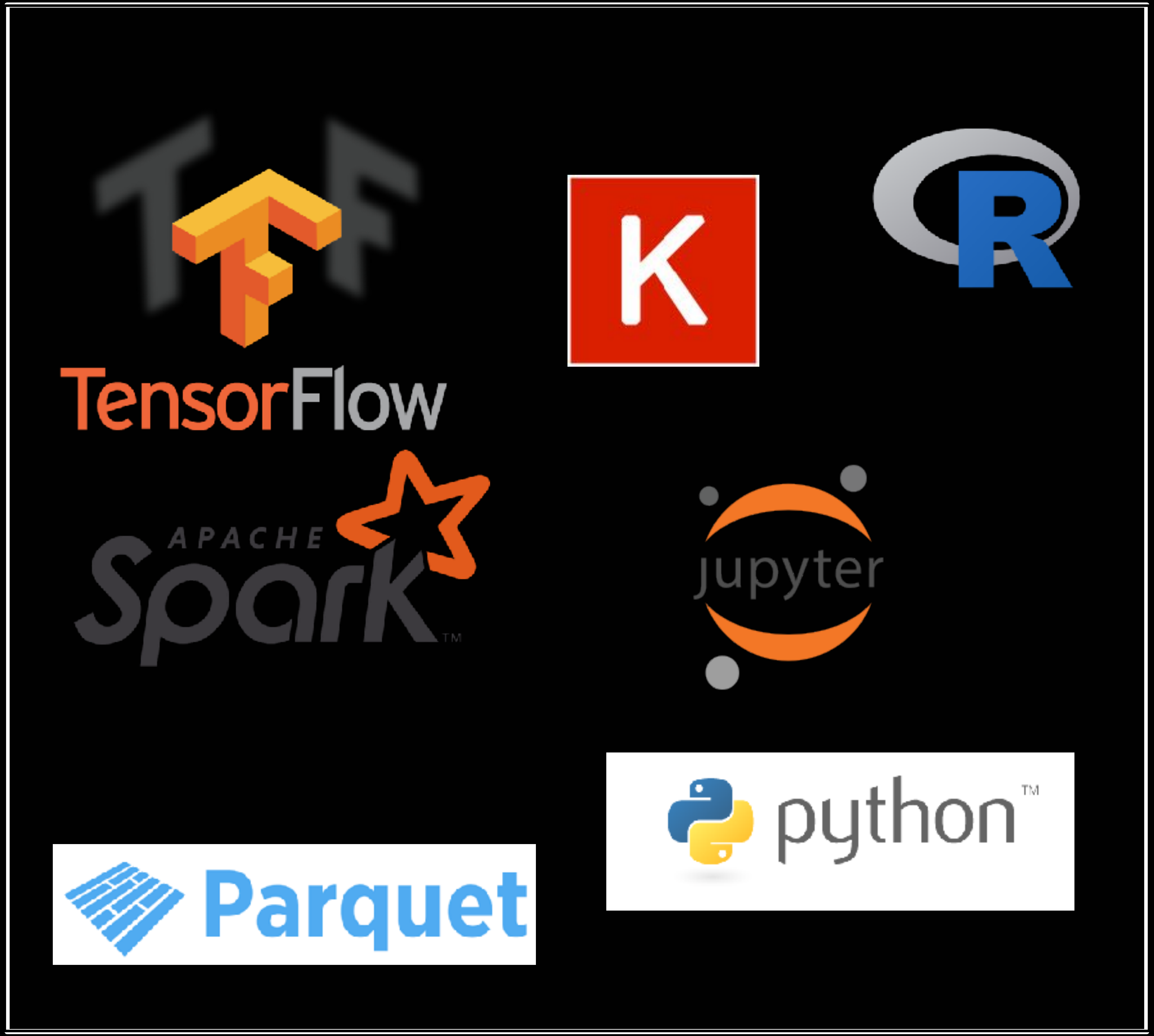
ints_purch6mos_cluster

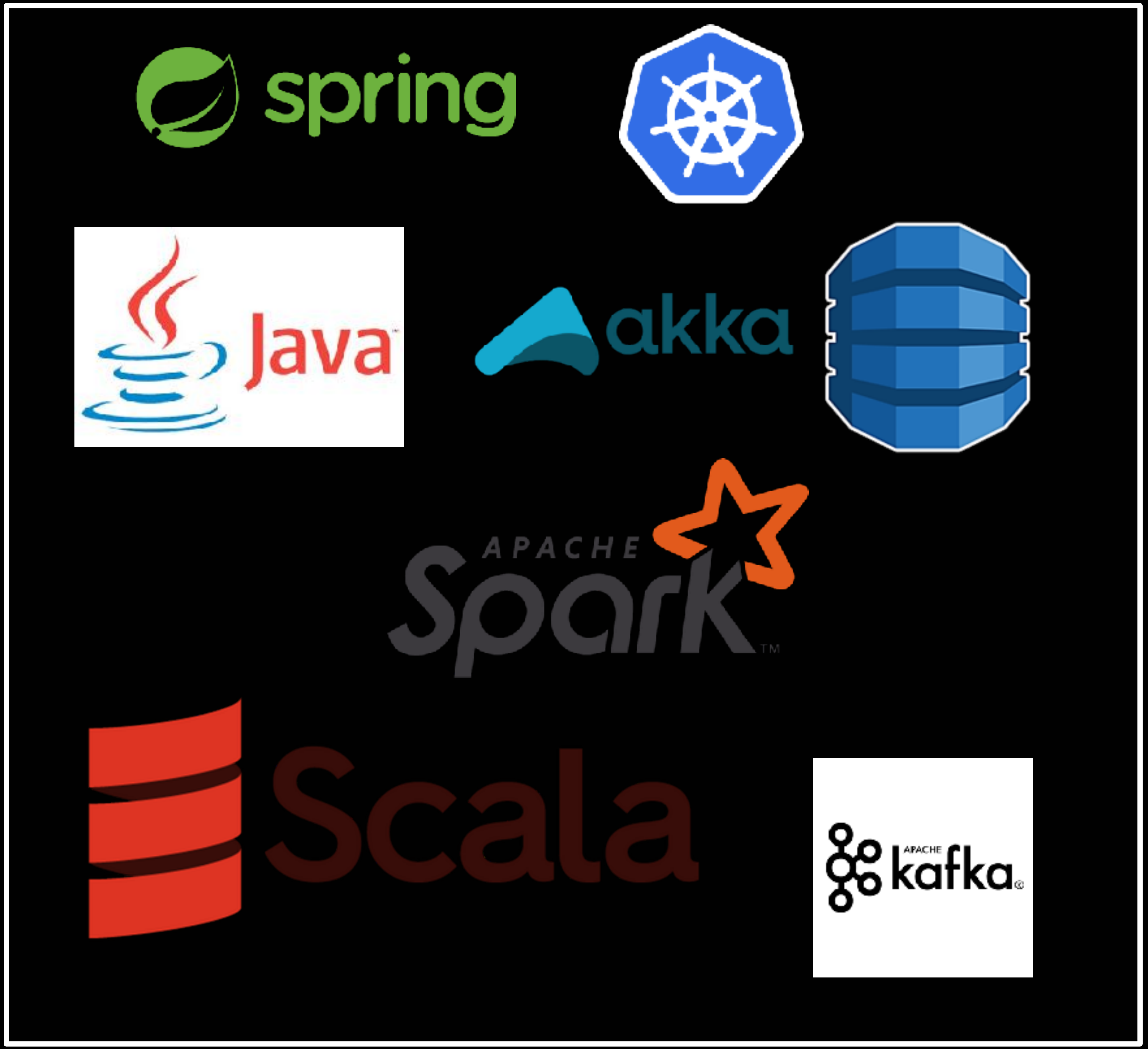
ints_purch6mos_cluster_nikeapp

ints_snapshot_cluster

https://www.finecooking.com/recipe/spaghetti-alla-carbonara

Because Tools Are Different





Because Science Workflow is Different

- Data Science work is research oriented
- Majority of code is thrown away
- Small changes may not show full impact
- Unit tests can’t capture problems that appear only with full dataset
- Data plays major role and is equally, sometime more, important than code
- Slower feedback loop

Because This is Not Sustainable

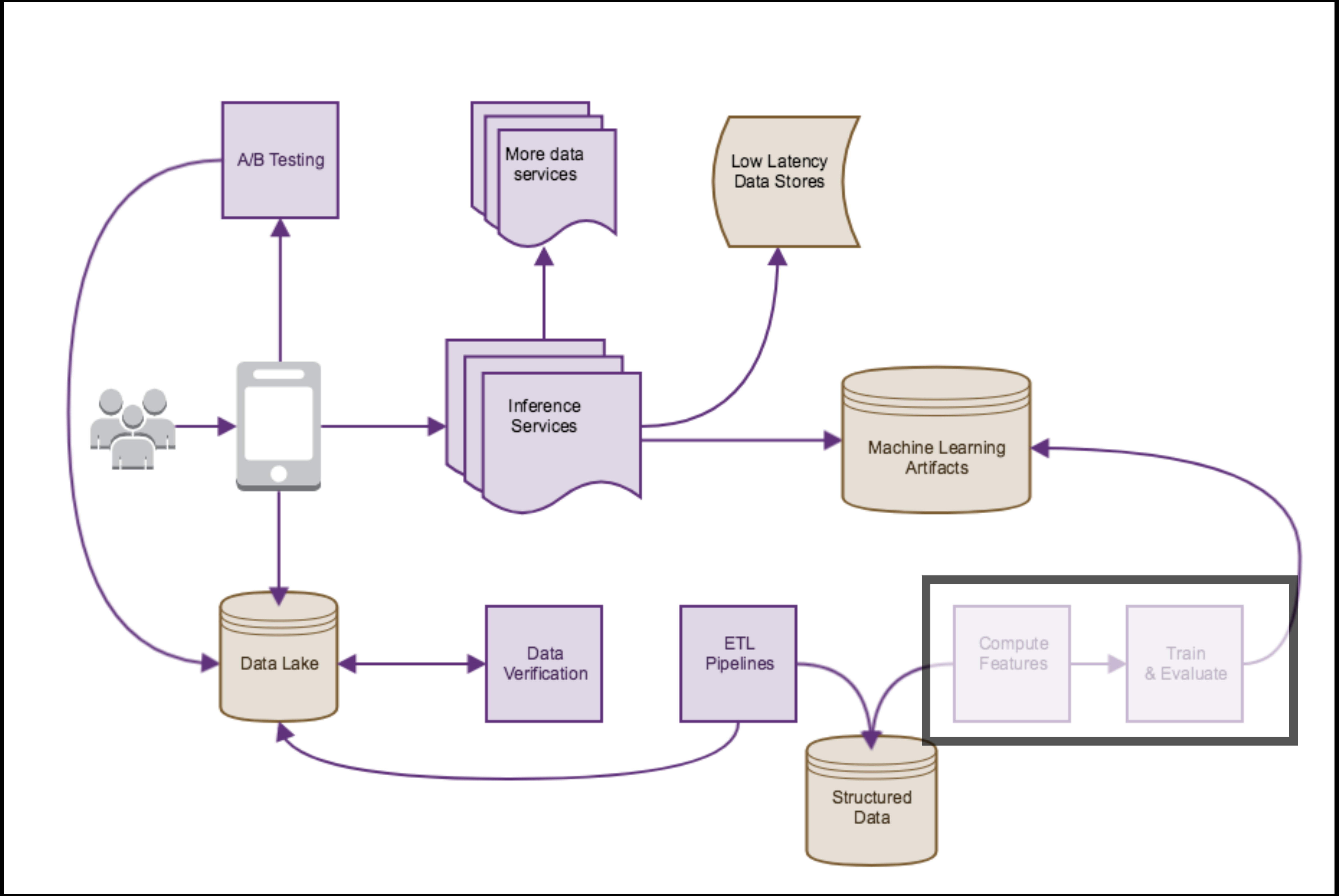
Hidden Debt is Dangerous Because it Compounds Silently.

Hidden Technical Debt in Machine Learning Systems Paper

We Can Come Back From This

http://coolspotters.com/characters/hermione-granger/and/accessories/vine-wood-and-dragon-heartstring-core-wand/media/1222722#medium-1222722

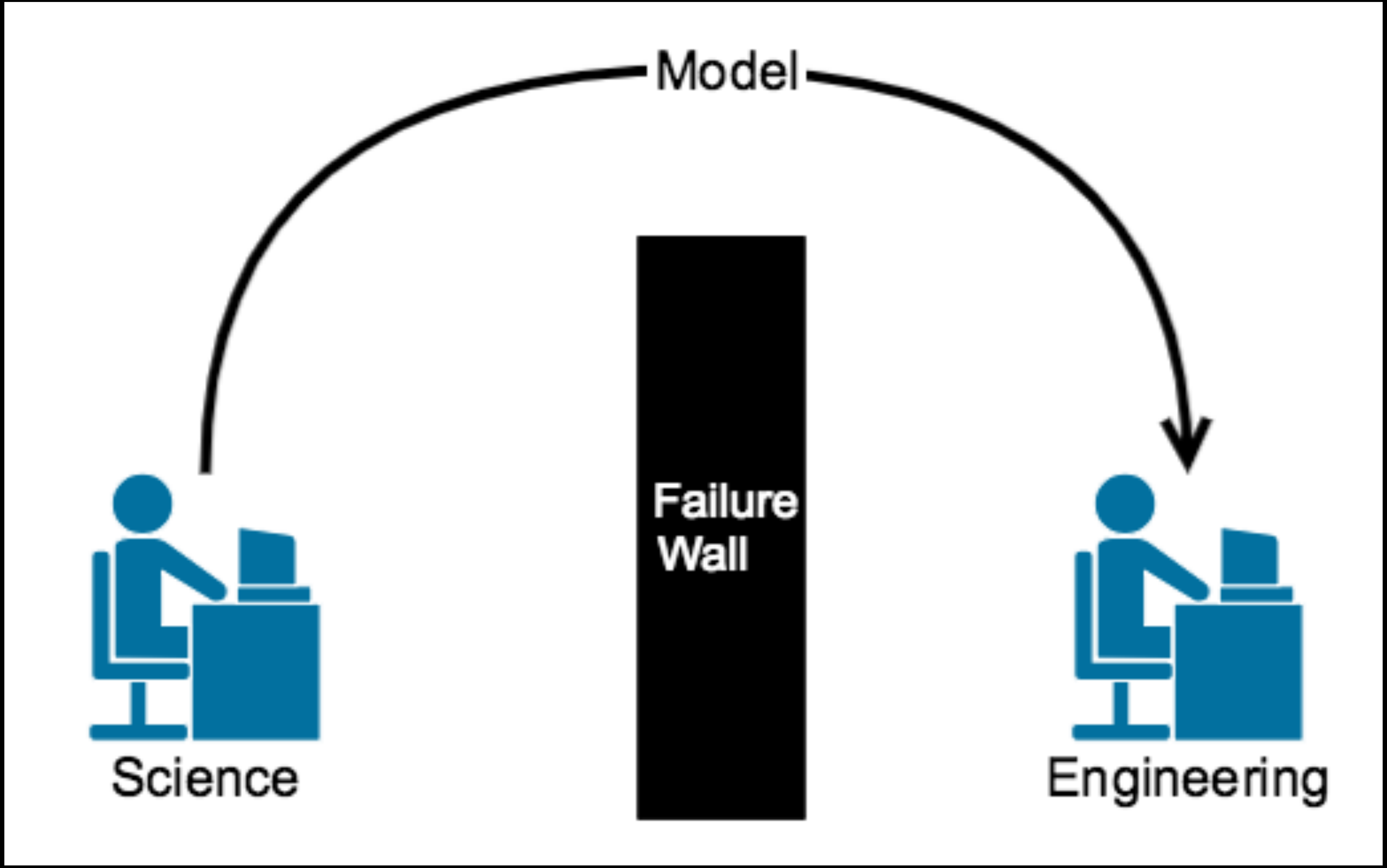
ML is a Small Component - Improve Holistically



The Basics: Adopt Production Grade Code Practices

- Reusability
- Unit Tests
- Logging
- Code Optimizations
- Version Control and PR Process
- Readability

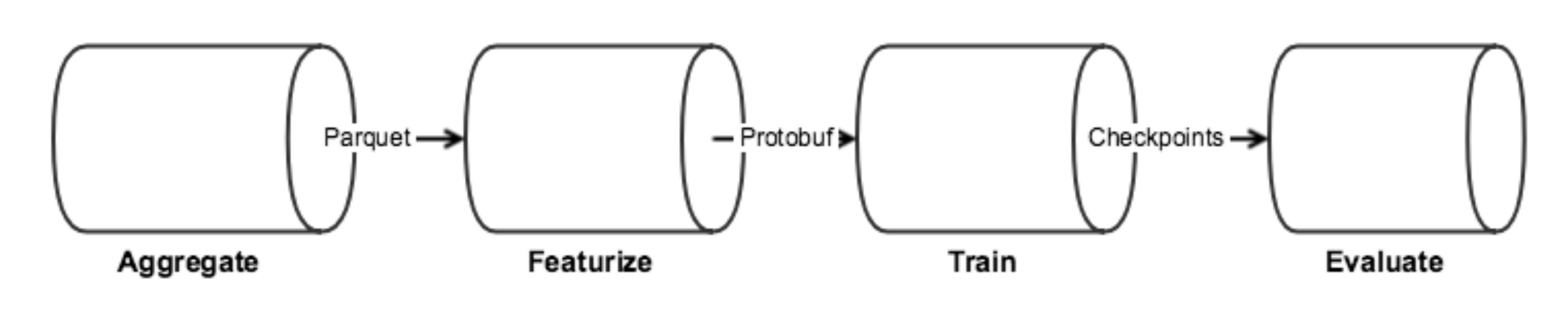
Move Away From This Workflow



Embed Engineers Early On

- Early optimizations
- Avoid rewrites
- Enforcement of approved tools
- Frequent iterations
- Avoid data pipeline maze
- Early contract definition

Contract First Development



Build Definition of Done For a Data Science Solution

- Data pipelines for input
- Acceptable output formats
- Repeatable pipelines for retraining
- Scalability
- Cluster optimizations
- Configurations are trackable through code

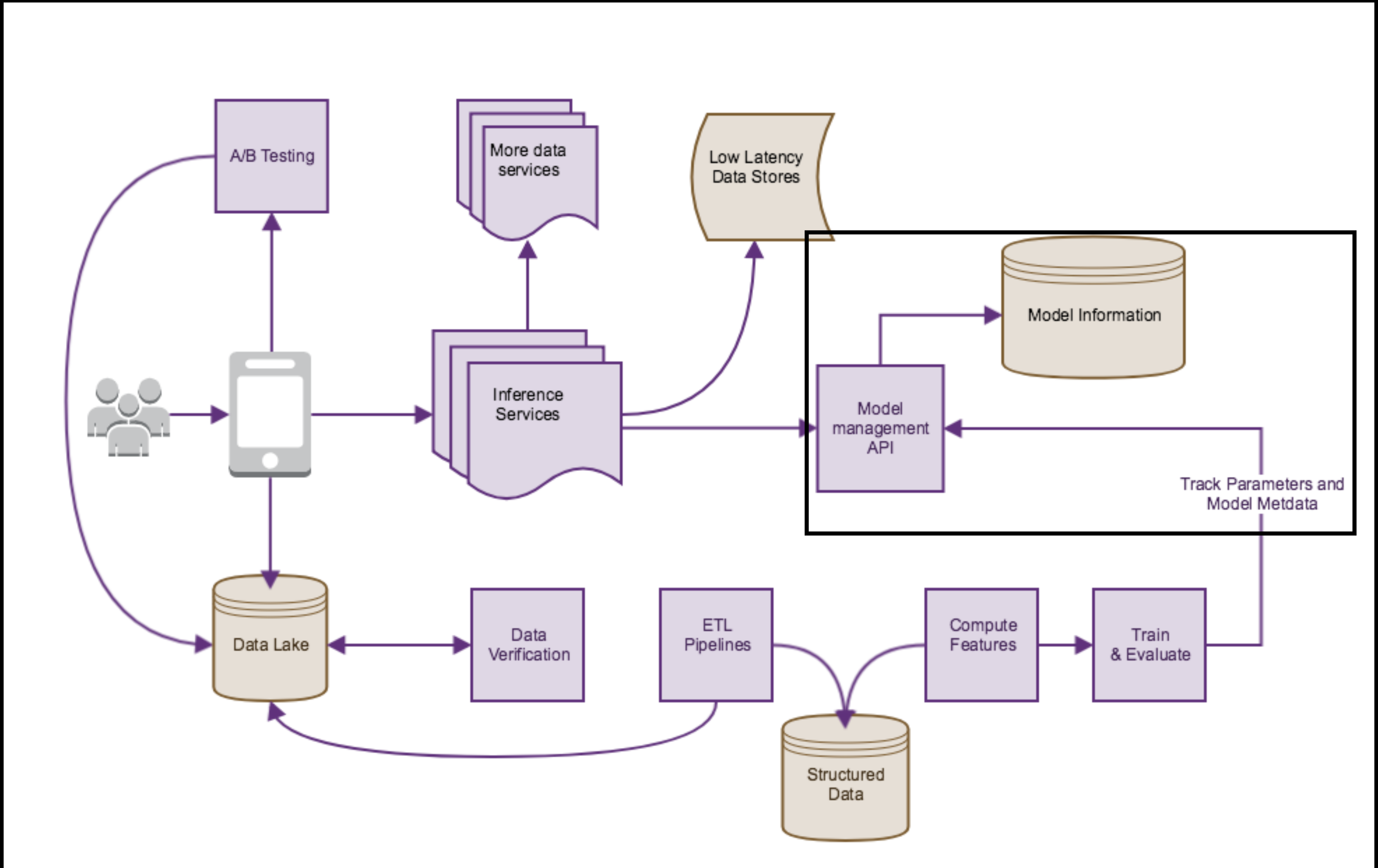
Go Beyond the Basics

https://wallpapersite.com/movies/hermione-granger-emma-watson-harry-potter-hd-4k-8936.html

Build Tools For Model Lifecycle

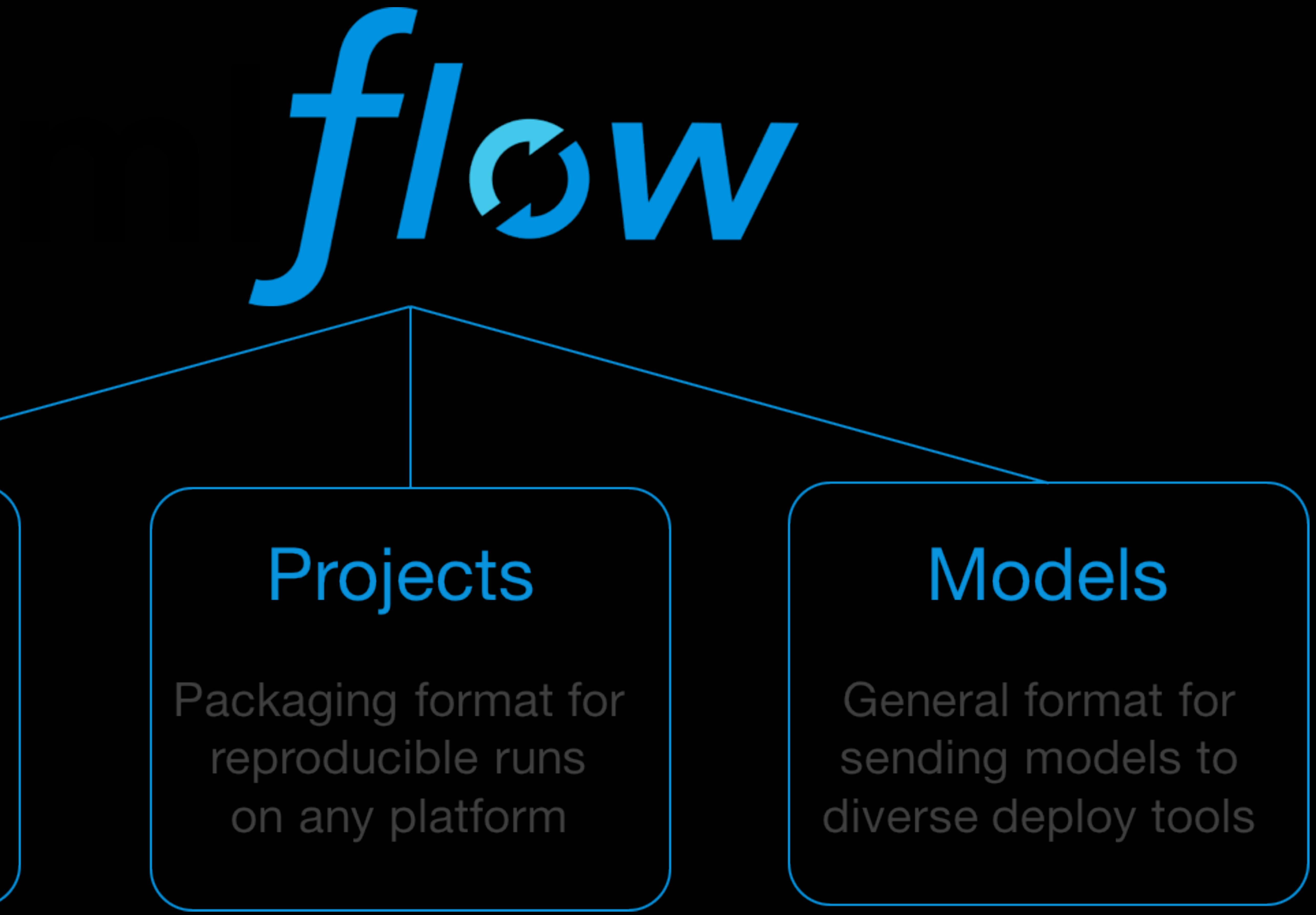
- Track Hyperparameters during multiple iterations
- Track input locations that can help reproduce a model
- Retraining schedule
- Track offline accuracy and model metadata for traceability.
- Single place to discover models developed for different domains
- Packaging to allow running on multiple platforms

At Nike


A/B Testing
More data services
Low Latency Data Stores
Model Information
Inference Services
Model management API
Track Parameters and Model Metdata
Data Lake
Data Verification
ETL Pipelines
Compute Features
Train & Evaluate
Structured Data

ML Flow - Open Source

Open source tool for tracking, packaging and deploying a model



mlflow

Tracking

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

Projects

Packaging format for reproducible runs on any platform

Models

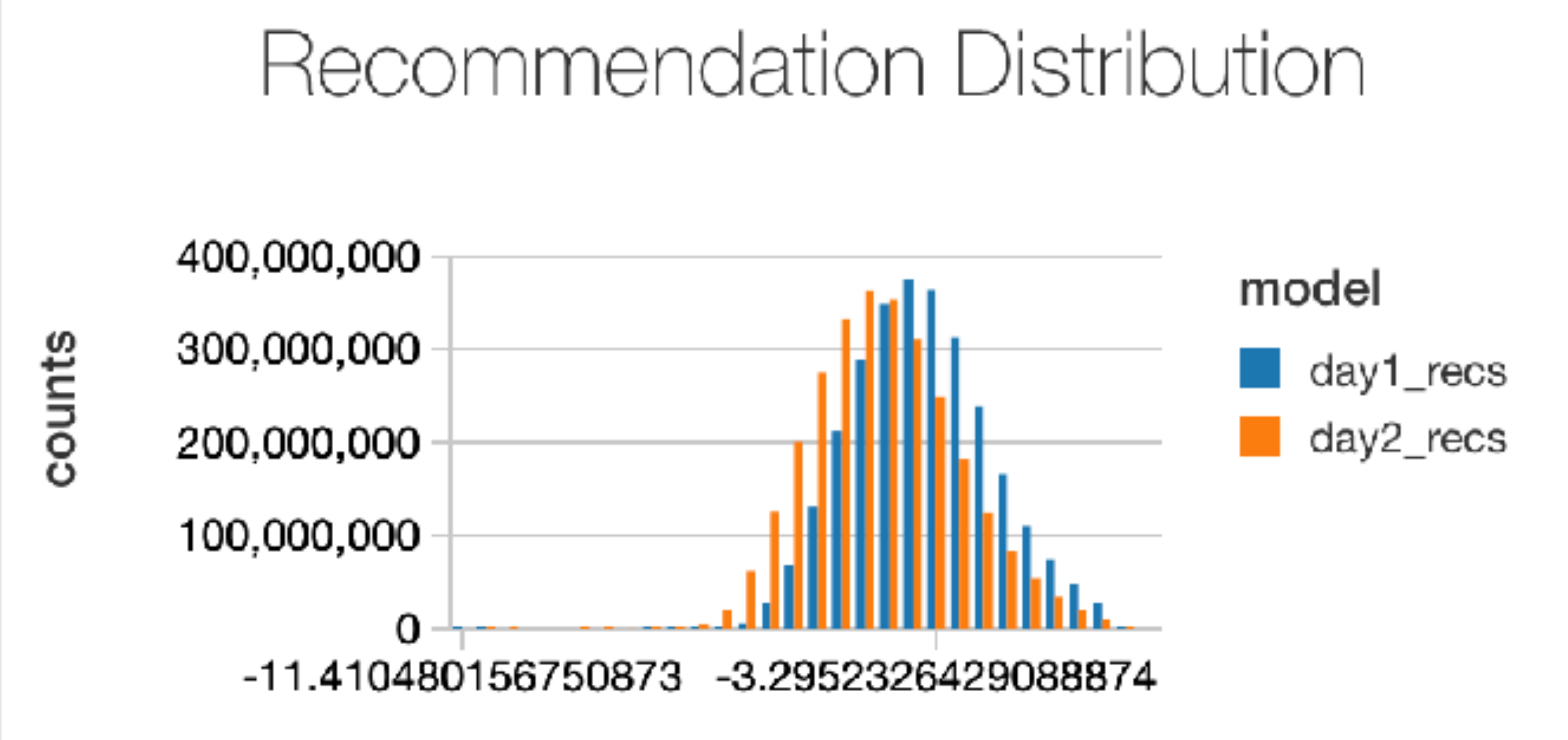
General format for sending models to diverse deploy tools

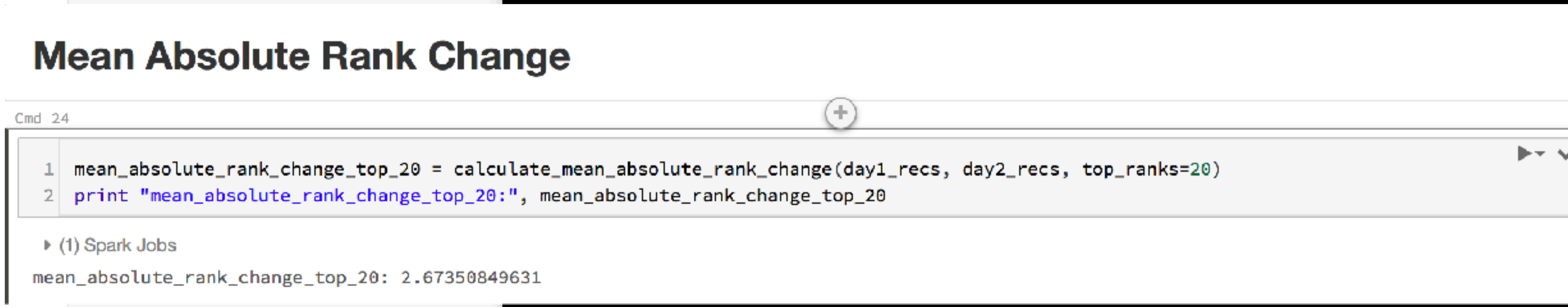
https://databricks.com/blog/2018/06/05/introducing-mlflow-an-open-source-machine-learning-platform.html

Alerting and Monitoring

- Prediction biases
- Upstream data dependencies
- Large prediction variances
- Automated mitigations
- Retraining variances
- Ask scientists for model verification scripts

Examples





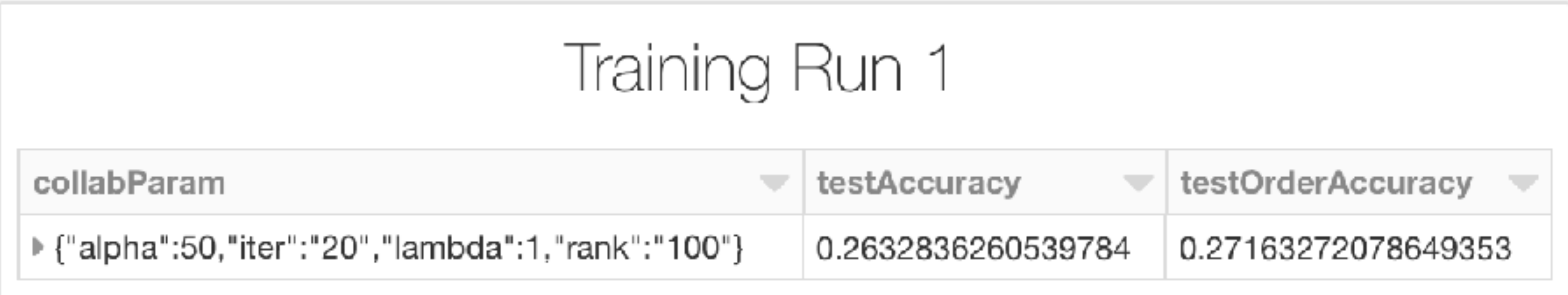
Mean Absolute Rank Change

Cmd 24

```
mean_absolute_rank_change_top_20 = calculate_mean_absolute_rank_change(day1_recs, day2_recs, top_ranks=20)
print "mean_absolute_rank_change_top_20:", mean_absolute_rank_change_top_20
```

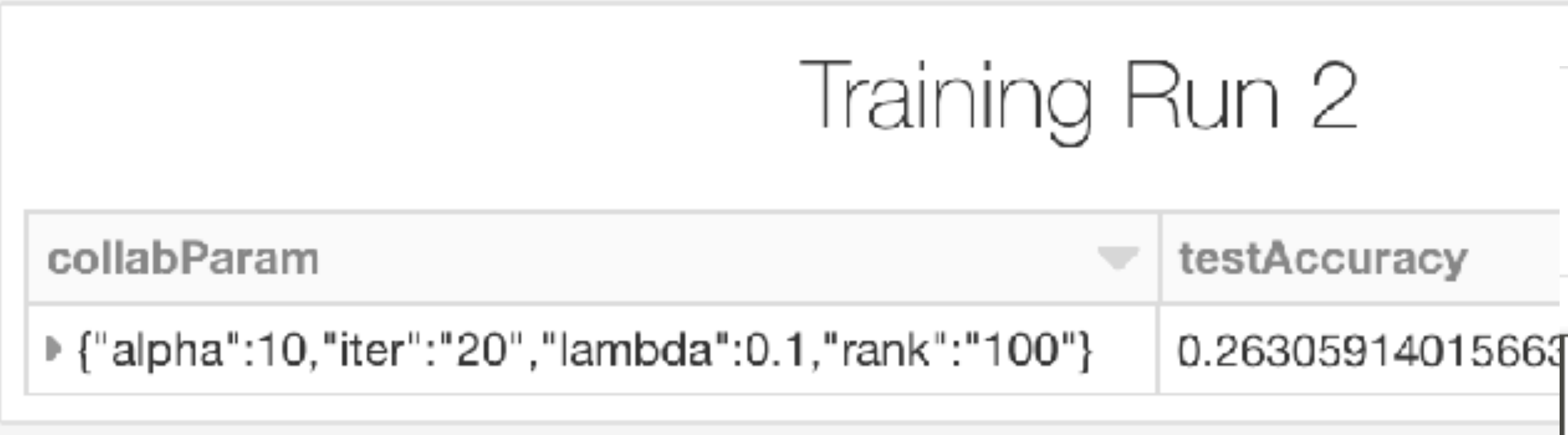
▸ (1) Spark Jobs

```
mean_absolute_rank_change_top_20: 2.67350849631
```



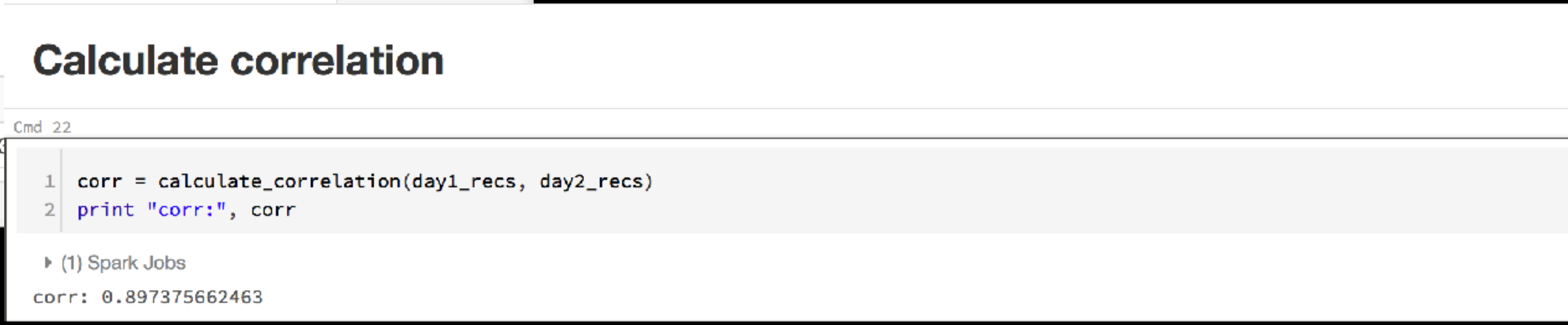
Training Run 1

collabParam	testAccuracy	testOrderAccuracy
▸ {"alpha":50,"iter":"20","lambda":1,"rank":"100"}	0.2632836260539784	0.27163272078649353



Training Run 2

collabParam	testAccuracy
▸ {"alpha":10,"iter":"20","lambda":0.1,"rank":"100"}	0.26305914015663



Calculate correlation

Cmd 22

```
corr = calculate_correlation(day1_recs, day2_recs)
print "corr:", corr
```

▸ (1) Spark Jobs

```
corr: 0.897375662463
```

Summary

Blur the lines between engineering and science to create sustainable data science systems. Reward reduction of overhead just as much as increase in accuracy of a model. Collaborate, Collaborate and then Collaborate more.

Questions?

http://www.mugglenet.com/2017/01/extreme-harry-potter-fan/hand-raise-2/

Thank You!

- Talk to your data scientists and understand their pain points.
- Talk to your engineers and understand their pain points.
- Collaborate closely to build a model together from inception to production
- Read the Hidden Technical Debt in Machine Learning Systems Paper
- Automate one thing that shouldn’t live on somebody’s laptop
- Join Nike Personalization - Hiring Data Scientists and Engineers!