Using A Hackathon To Crowd-Source The First Release Of A Clinico-Genomics Feature Store

James Black
Senior Director, Insights Engineering | Roche

11th September, 2023
Pharma Data Congress | London

Agenda

1. Why feature stores?
2. Data-centric AI
3. Crowd-sourcing
4. Learnings

Why feature stores?

Data

NGS test

Vitals

Labs

Demographics

Features

labs_sunday
tumour_burden

Model

fast_frugal_tree

Why feature stores?

Data

Jan 2023 

Aug 2023 

Features

labs_sunday
tumour_burden
tumour_burden_2

Models

fast_frugal_tree
fast_frugal_tree_v2
fast_frugal_tree_v3
xgboost_v1

Why feature stores?

Data	Features	Models
Jan 2023 	**labs_sunday** **tumour_burden** tumour_burden_2	**fast_frugal_tree** fast_frugal_tree_v2 fast_frugal_tree_v3 xgboost_v1
Aug 2023 		

Why feature stores?

Data	Features	Models
Jan 2023 	**labs_sunday** tumour_burden **tumour_burden_2**	fast_frugal_tree **fast_frugal_tree_v2** fast_frugal_tree_v3 xgboost_v1
Aug 2023 		

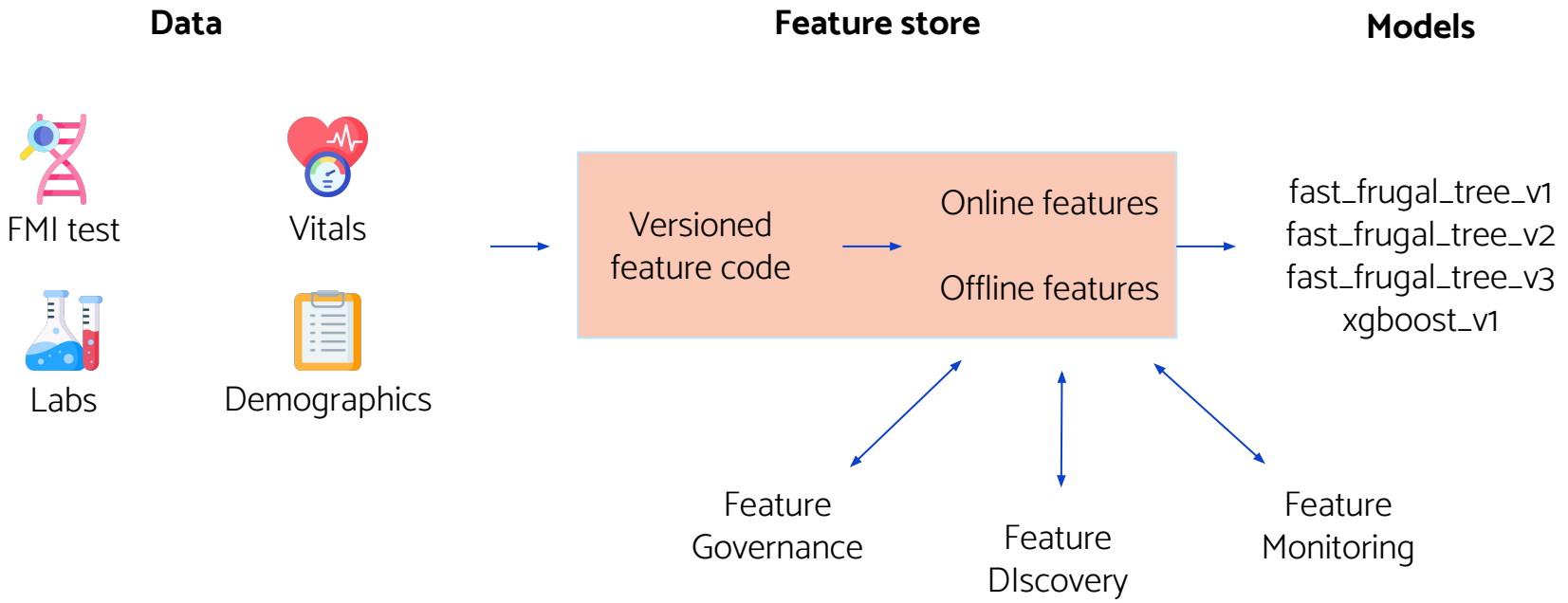
Why feature stores?

Data	Features	Models
Jan 2023 	**labs_sunday** tumour_burden **tumour_burden_2**	fast_frugal_tree fast_frugal_tree_v2 **fast_frugal_tree_v3** xgboost_v1
Aug 2023 		

Why feature stores?

Data	Features	Models
Jan 2023 	**labs_sunday** tumour_burden **tumour_burden_2**	fast_frugal_tree fast_frugal_tree_v2 fast_frugal_tree_v3 **xgboost_v1**
Aug 2023 		

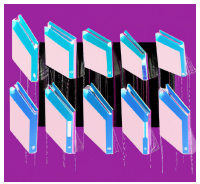
Why feature stores?



Data-centric AI

“Instead of focusing on the code, companies should focus on developing systematic engineering practices for improving data in ways that are reliable, efficient, and systematic. In other words, companies need to move from a model-centric approach to a data-centric approach.”

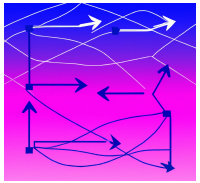
A feature store as a pillar of data centric AI



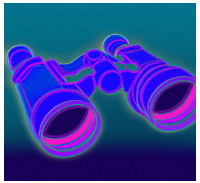
Versioned feature code

Parameterise features

Monitor features



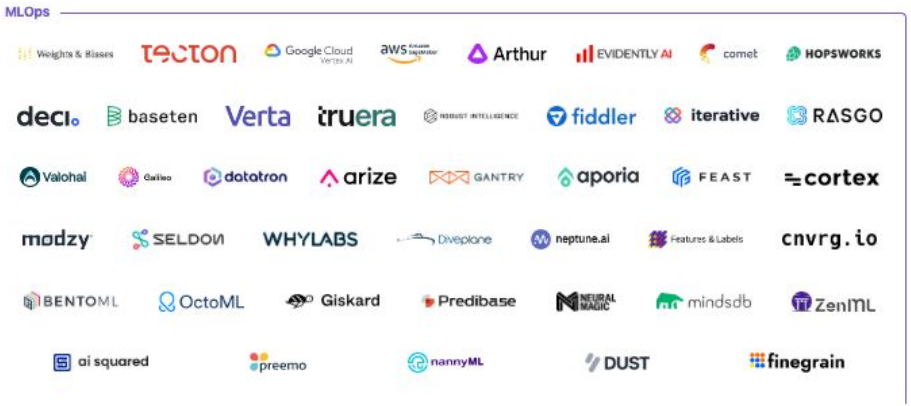
Across versioned data



Discover features

Serve features

Picking a feature store



- MLOps platforms usually have a feature store
- Some stand-alone feature stores exist already
- Discovery tends to be poor, and interfaces technical (e.g. CLI)

https://mad.firstmark.com/ is source of MLOps icons

The realities of ML in Pharma

Statistician

Data Scientist

ML Engineer (bioinformatics)

ML Engineer (imaging)

Common feature store?

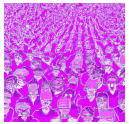
OR

Expose feature metadata in common discovery layer?

How to kick start a shared feature store?

A company wide hackathon

Roche runs a yearly company wide hackathon



Usually ~500 data scientists take part

2022 theme was data-centric AI

Benefits of a hackathon to collect an initial set of features

Have a useful feature store at launch

Ensure relevant across departments

Bring attention to a data-centric AI mindset

Data-Centricity: data for CGDB


FMI test


Vitals


Labs


Demographics

Feature generation

Model


Prediction target is time till death


96,686 patients


118 variables


58,066 events


744 MBs of data


93,673,017 data points


Train - 60%


Leaderboard - 20%


Test - 20%

Hackathon output

Data centricity scores

RAAD Challenge 3.0

	Score	Division	Country
Top 10			
The CubeEneers	0.6804	Diagnostics	DE
Dynamic Databenders	0.6673	Diagnostics	US
CGDBuilders	0.6610	Pharmaceuticals	US
The Left Hamsters	0.6601	Pharmaceuticals	US
GenGenix	0.6592	Pharmaceuticals	CH, GB
Support Victory Machine	0.6590	Pharmaceuticals	CN, US
Hyperion	0.6550	Pharmaceuticals	CH
Three Data Prophets	0.6546	IT/HR, Pharmaceuticals	CH, GB
Gotluck_2	0.6496	Pharmaceuticals	US
Team Python	0.6401	Pharmaceuticals	US
Not in top 10			
DunderMifflinDataCompany	0.6345	Diagnostics	US, CH, DE
Pattern Hunters	0.6048	Diagnostics, Pharmaceuticals	US, CH
ZostawmynakonieC	0.6030	IT/HR	PL
covid-404	0.6030	Pharmaceuticals	CN
Tutorial model	0.5846	NA	NA
cymax	0.5660	Pharmaceuticals	US
gNewbies	0.5586	Pharmaceuticals	US

Where are we today?

1. 170 features submitted
2. Expert panel reviewed top 10 teams features
3. From top teams 100 features, prioritised taking to production features for feature store

Next steps

4. Decide on recommended feature store approach (2023 Q4)
5. Deploy feature store (2024)

Learnings

Learnings

The hackathon approach:

- Brought attention to feature sharing / re-use
- Exposed us to diversity of how features are written
- Where many teams tackled the same concept (e.g. ECOG), there were learnings in reviewing and combining approaches
- Features often needed to be parameterised (e.g. pick window, pick index)

Learnings

For feature stores:

- Does discoverability and ease of use trump MLOps tooling like drift detection?
    - If true - do we need MLOps tooling, or is it really just another form of derived data so can push metadata into existing data data catalogs?