Productionizing ML Pipelines with Airflow, Kedro, & Great Expectations

Introductions

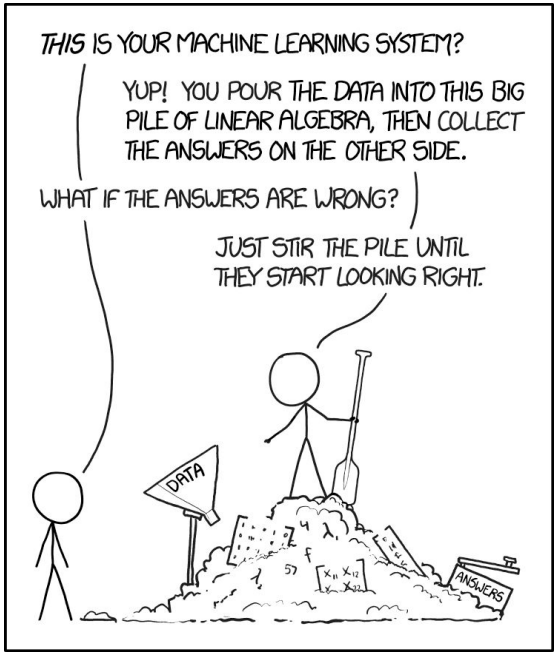
Kenten Danas, Field Engineer @ Astronomer

Based in Seattle, WA. Almost always outside when not working with Airflow

Background in data engineering consulting and helping companies adopt Airflow for many different use cases

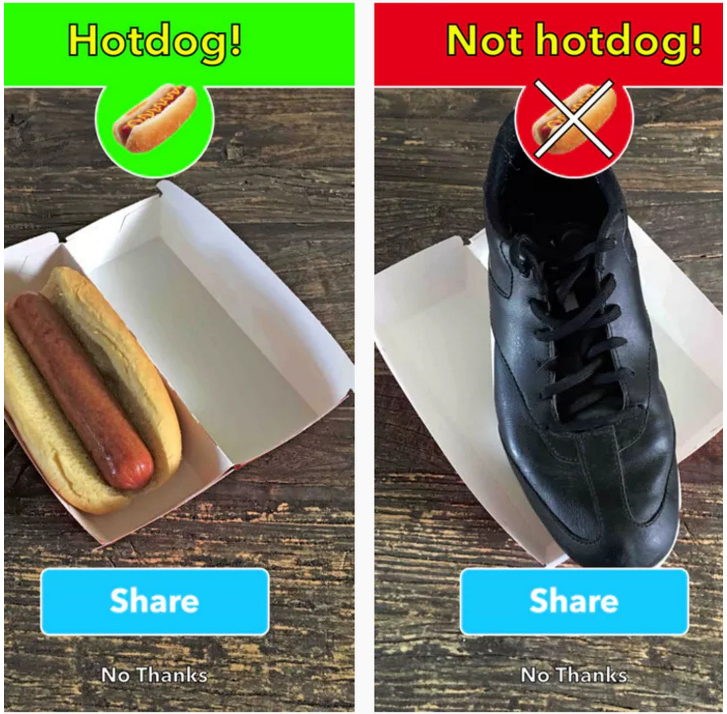
Productionizing ML Models

Making an ML model is fun! But automating it so results can be used in production adds another layer (or multiple layers) of complication.



Orchestrating ML jobs

Automating the orchestration of your ML models allows you to rely on their results in production systems



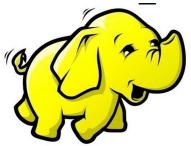
Data Validation

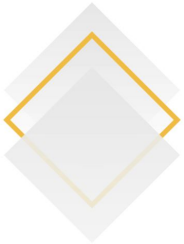
Building data validation checks into your ML pipelines can help prevent aberrant data from causing issues in production.

Why Open Source?

Using open source software brings the benefit of tools developed by broad, active communities at no cost.

Users have a say in the direction of the project, and in most cases can contribute back themselves.

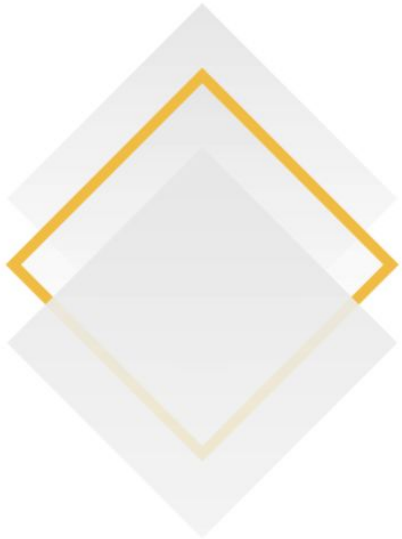




Kedro

“Open source development workflow tool that helps structure reproducible, scaleable, deployable, robust and versioned data pipelines.”

- **https://kedro.readthedocs.io/en/0.15.3/01_introduction/01_introduction.html**
- **https://github.com/quantumblacklabs/kedro**



Kedro

Great Expectations

Open source Python framework for data validations

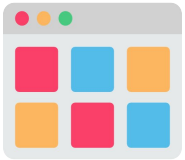
- **Define expectations, automated data validations and profiling**
- **https://greatexpectations.io/**
- **Airflow provider!**
  - **https://registry.astronomer.io/providers/great-expectations**

great_expectations

Airflow - to tie it all together!

The best way to orchestrate your data pipelines!

Pipelines as Code



Extensible

Scalable

Apache Airflow

Demo