Mastering MLOps at a Reasonable Scale at Allegro

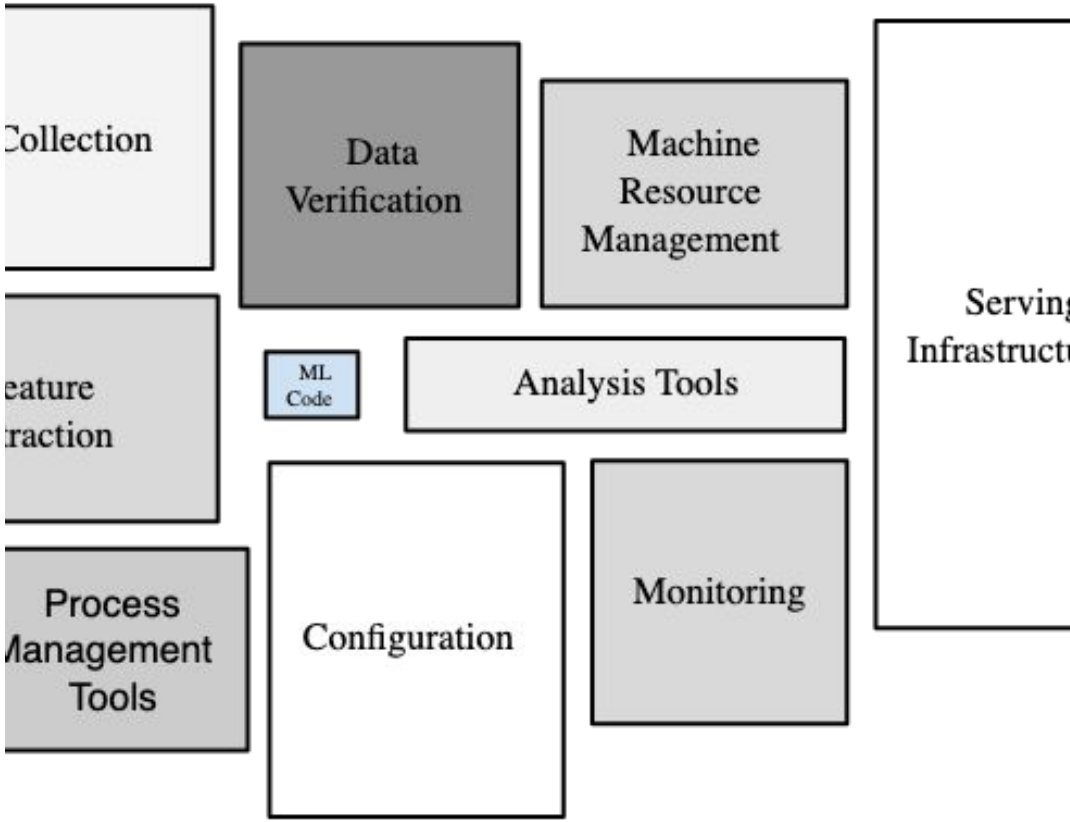
Who we are

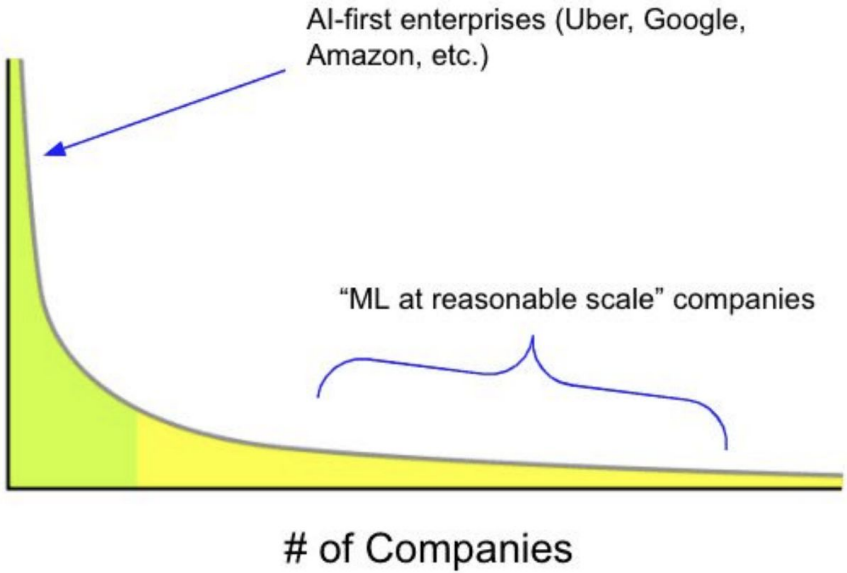
Piotr Januszewski

Marcin Cylke

From working model to production

- reproducible and orchestrated pipelines,
- alerts and monitoring,
- versioned and traceable models,
- auto-scalable model serving endpoints,
- data versioning and data lineage,
- feature stores,
- and so much more.




AI-first enterprises (Uber, Google, Amazon, etc.)
“ML at reasonable scale” companies
# of Companies

Reasonable scale	Hyper-scale (FAANG)
Models generate 100s thousands to 10s of millions per year	Models generate 100s of millions or billions of USD per year
10s of engineers	100s or 1000s of engineers
“Mere” terabytes	Petabytes (or more) of data
Finite amount of computing budget	Godlike amount of computing budget

We did not coin this term. This guy did:

Jacopo Tagliabue

You Don't Need a Bigger Boat

An end-to-end (Metaflow-based) implementation of an intent prediction (and session recommendation) flow for kids who can't MLOps good and wanna learn to do other stuff good too.

After few mo

- Our MLC
- A new op project,
- A second Merlin-fo

recs-at-resonable-scale

Recommendations at "Reasonable Scale": joining dataOps with deep learning recSys with Merlin and Metaflow (blog)

Overview

February 2023: aside from behavioral testing, the ML pipeline is now completed. A blog post on the NVIDIA Medium was just published!

This project is a collaboration with the Outerbounds, NVIDIA Merlin and Comet teams, in an effort to release as open source code a realistic data and ML pipeline for cutting edge recommender systems "that just works". Anyone can ~~cook~~ do great ML, not just Big Tech, if you know how to pick and choose your tools.

Allegro - The Move to the Cloud

- Cloud offers wealth of tools

Allegro - The Move to the Cloud

- Cloud offers wealth of tools
- We come with our old habits

Reasonable scale - DataOps

The definition of Big Data changed from the time Spark was first introduced: popular datasets from the Big Data era … can now be processed comfortably in one machine.

Most of what we considered internet scale at the time would be considered a “reasonable scale” today.

~ Jacopo Tagliabue et al., Building a Serverless Data Lakehouse from Spare Parts. VLDB Workshops 2023

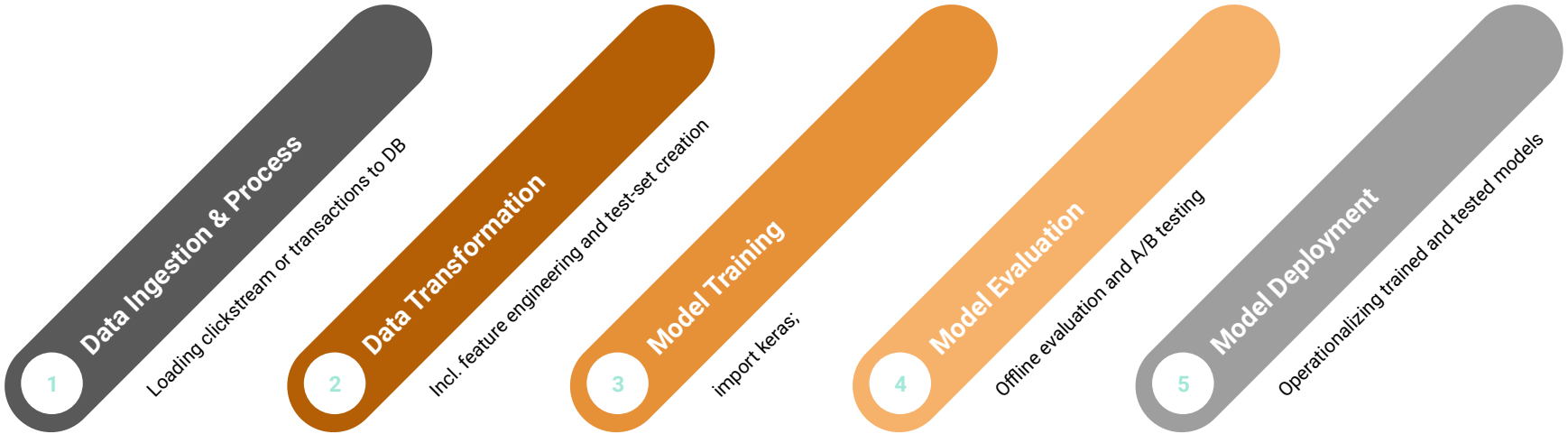
Allegro - The Move to the Cloud

- Cloud offers wealth of tools
- We come with our old habits
- With great force comes great responsibility

The four pillars of ML at Reasonable Scale

1. Data is superior to modeling.
2. Log then transform.
3. PaaS & FaaS is preferable to IaaS.
4. Vertical cuts deeper than distributed.

Reasonable Scale Pipeline



Orchestrators

Don’t tightly couple too many small steps where each needs to wait for the cloud resources allocation to run.

Use containerized environment in which you can quickly run each pipeline step in separation, with one command.

Data Ingestion & Process

Don’t store bunch of loose, unstructured files.

Load your raw data into the modern, serverless data warehouse where you can transform it in reproducible and reliable fashion.

Data Transformation

You don’t need Spark!

Use wealth of in-memory, GPU accelerated, parallel execution DataFrame libraries.

Model Training

Don’t rewrite complex models to then train them on your devel machine, and poke around your log dirs in search for a specific run results.

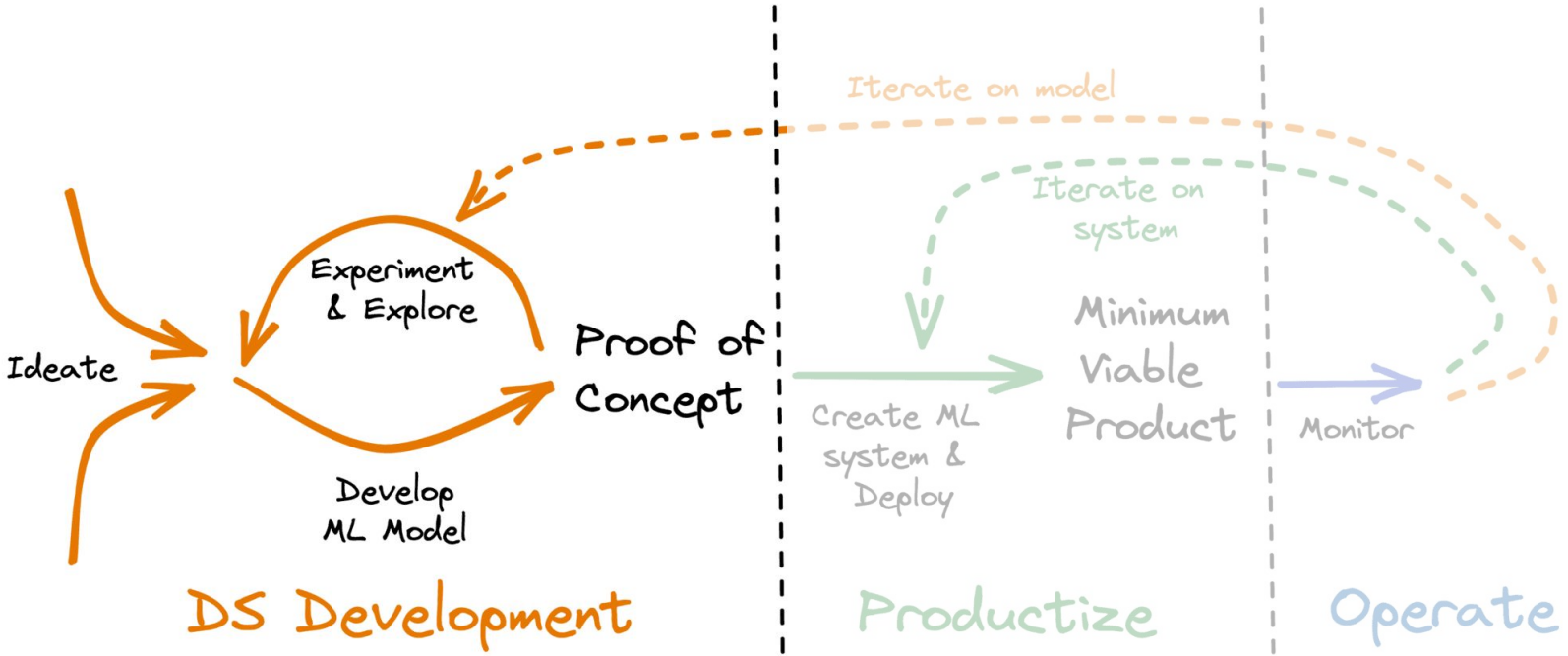
Reuse models from model hubs, pack your training scripts and scale them horizontally using cloud capabilities, and track your experiments with integrated services for easy comparison in one place.

Model Evaluation & Deployment

Don’t send your model to production because it does something, hoping for the best.

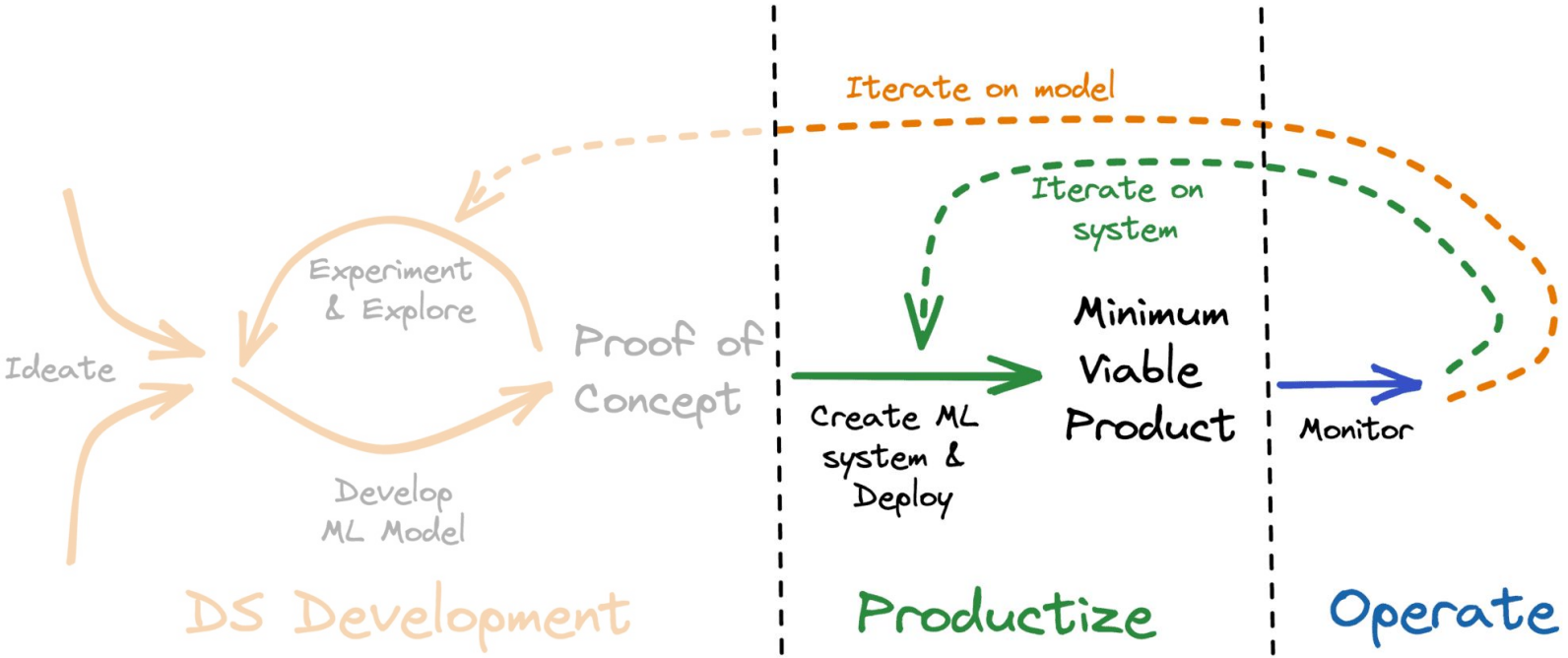
Run fast and repeatable evaluation of online and offline correlated metrics; A/B testing when your solution is mature enough.

Reasonable Scale + Modern Tools = Fast Iteration



MLOps: why and how to build end-to-end product teams by Xebia

Fast Iteration -> Productization



MLOps: why and how to build end-to-end product teams by Xebia

Parting thoughts

If you write Github actions before you drew one plot...

If you collected 1 GB of data but won't touch it until you gather 1 TB...

If you are not delivering insights to your stakeholders on a daily basis because you're busy building planetary-scale infrastructure...

You are probably at risk of overengineering and not adding value.

– Quote from Alexander Kislukhin


That's all Folks!

Awesome MLOps templates!

- https://github.com/jacopotagliabue/you-dont-need-a-bigger-boat
- https://github.com/jacopotagliabue/recs-at-resonable-scale

A whole bunch of links

- https://neptune.ai/blog/mlops-at-reasonable-scale
- Recs at Reasonable Scale NVIDIA RecSys Summit 2022
- ML Ops at Reasonable Scale feat. Jacopo Tagliabue | Stanford MLSys Seminar Episode 35
- Just Build It! Tips for Making ML Engineering and MLOps Real // Andy McMahon // MLOps Meetup #91
- Building End-to-End Recommender Systems with Nvidia Merlin
- MLOps at a Reasonable Scale
  - ML and MLOps at a Reasonable Scale
  - Hagakure for MLOps: The Four Pillars of ML at Reasonable Scale