Delivering Machine Learning Value: A Guide For Humans

Discord Distinguished Speaker Series
3/31/22
Mihail Eric
🐦 mihail_eric

Who am I

- Researcher -> Engineer -> Research Engineer
- Helped start a special projects group at Alexa AI focused on forward-thinking efforts for the platform
    - Built some of first large-scale NLP models
        - Transformers trained on ~1TB of conversational data
    - Built and integrated state-of-the-art models across language understanding, information retrieval, and generation
- Run a consultancy helping organizations build ML systems often in 0-1 phase
- Run Confetti AI a platform for educating data practitioners

How to Quickly Lose Friends

MLOps Is a Mess But That's to be Expected

March 2022

A Brief Recap

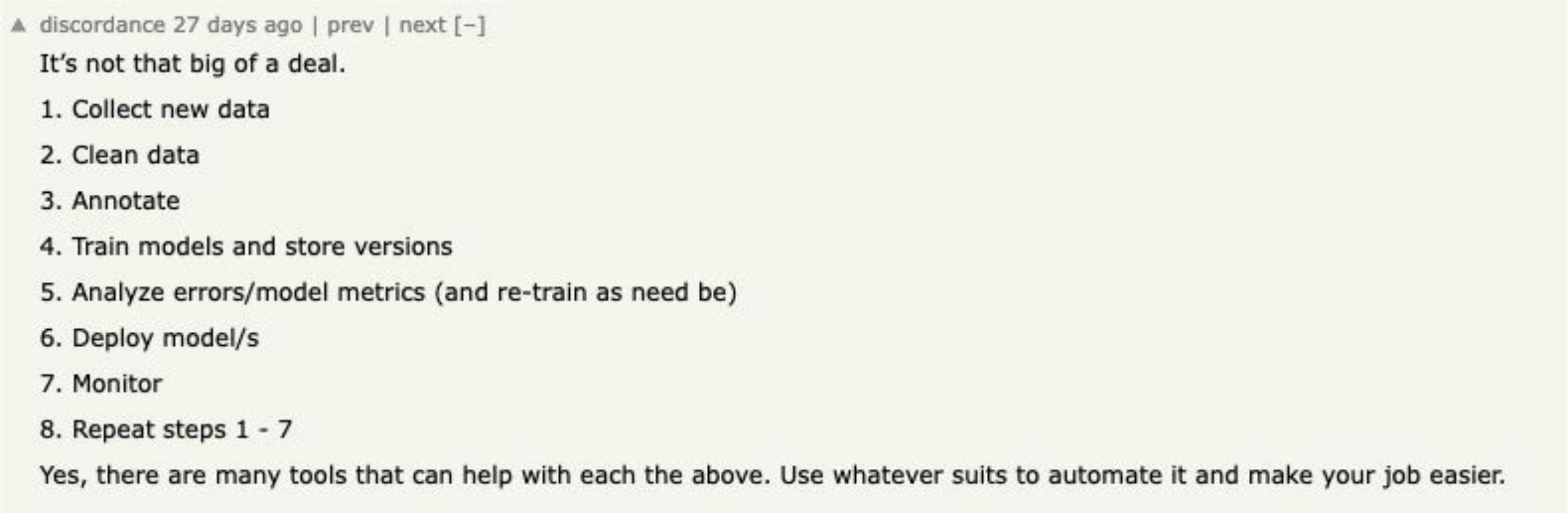
- Machine learning has been on a tear
  - Record funding
  - New tools
  - New companies
- Everything is perfect, right…right?

MLOPs
THIS IS FINE.

A Brief Recap

- Spoke to 20+ ML practitioners
  - What do you think about the state of things?
  - What are your thoughts on different parts of the stack?
  - Where are we going?
- The tooling landscape is fragmented and confused
  - New terms for similar things
  - No real canonical stack
  - Conventions and best practices are largely in flux
- Most companies aren’t ready for the latest and greatest ML
- Tooling will get better but we also need organizational maturation

Hacker News Reception



▲ discordance 27 days ago | prev | next [–]

It’s not that big of a deal.

1. Collect new data
2. Clean data
3. Annotate
4. Train models and store versions
5. Analyze errors/model metrics (and re-train as need be)
6. Deploy model/s
7. Monitor
8. Repeat steps 1 - 7

Yes, there are many tools that can help with each the above. Use whatever suits to automate it and make your job easier.

It sort of is a big deal

What Should You Be Doing as an Organization

- Well… it depends
- Different companies need (and should expect) different things from ML

Focus on business context

Case Study 1: ML Newcomers

- The Organization
  - Most companies in the early stages really just need their first wins
  - Small team of ~1-2 data scientists
  - Get executive buy-in for machine learning efforts
  - Emphasize POC projects with clear path to ROI
- The Tech
  - Use some off-the-shelf end-to-end platform (Sagemaker, Datarobot, etc)
  - Some components like a monitoring solution may not be immediate priority
    - If appropriate start with no ML!
- Education
  - Set reasonable expectations for what ML can do
  - Bring business stakeholders to the table (provide transparency in understandable terms)

Case Study 1: ML Newcomers

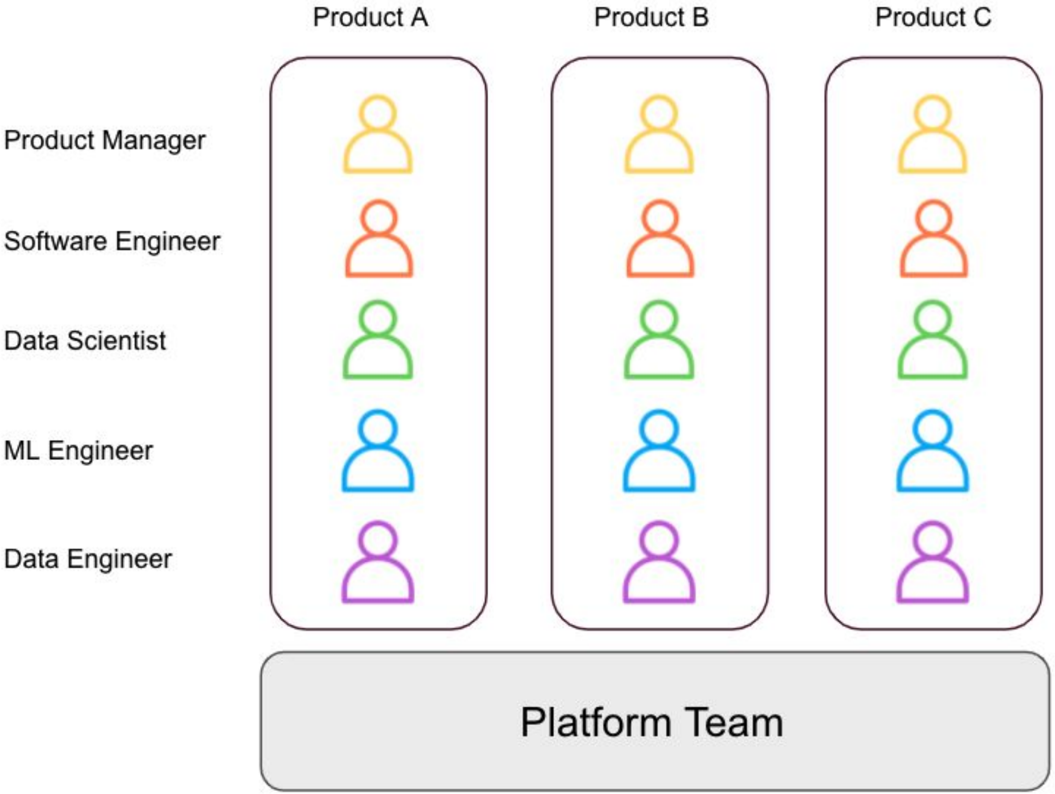
- Pitfalls
    - Oftentimes resource constrained (people, compute, data)
    - Need strong alignment with business function
    - Time to value is crucial
        - Avoid ML disillusionment

Case Study 2: ML Natives

- Companies that recognize ML is core to what they do
- The Organization
  - Already have buy-in for projects
  - More deliberate about defining scalable operational processes
- The Tech
  - Heterogeneous stack with your best-in-breed sample of the tooling buffet

Organizational Blueprint

- Tight collaboration and cross-functional teams
  - No silos please!
  - Minimize communication latency, maximize context
- Enable fast iteration of problem and solution
  - Agile principles at all parts of ML pipeline
  - Emphasize easy developer experience/onboarding for data practitioners
- Recognize that ML systems are living entities that balance experimentation **and** engineering
  - Experimentation is about doing the ML
  - Engineering is about delivering the ML



Product A

Product B

Product C

Product Manager

Software Engineer

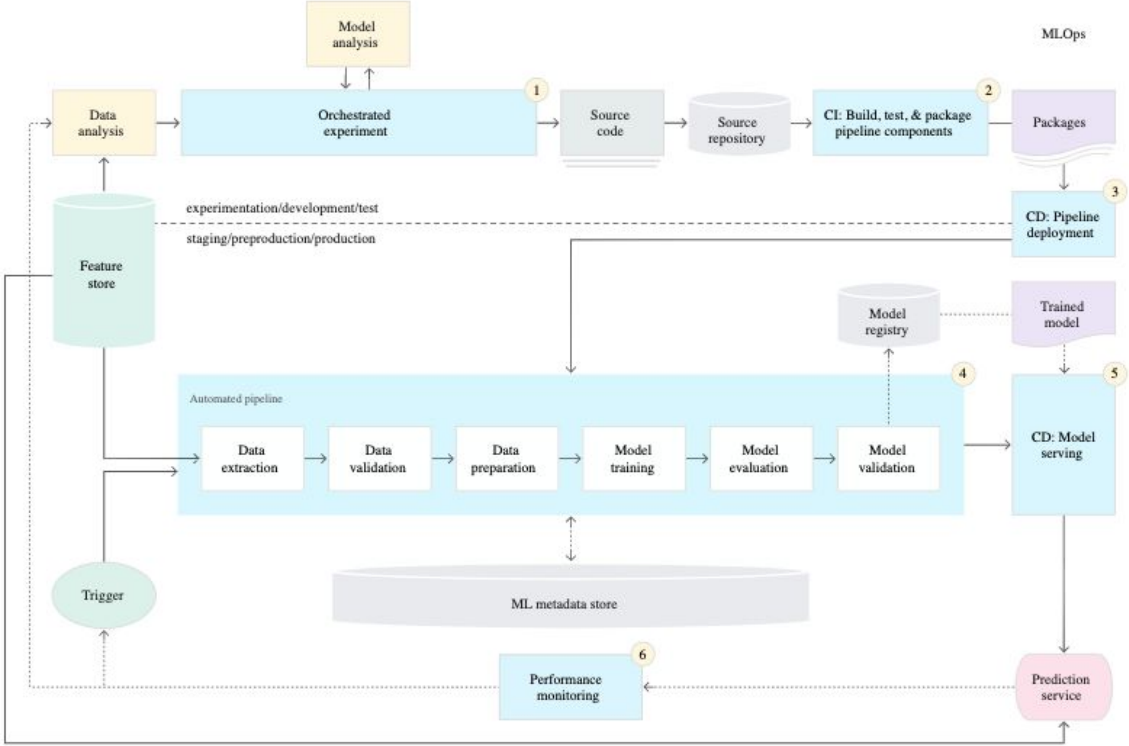
Data Scientist

ML Engineer

Data Engineer

Platform Team

“The Stack”



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Important Considerations

- Automation and configs galore
- Ensure there is robust data management and infrastructure
    - Remember the AI hierarchy of needs
- Effectively abstract away resource concerns for experimental work
- Need well-defined triggers for different pipeline processes
- Programmatic documentation
    - The tests are the documentation!

Some Unsolved Problems

- Monitoring
  - What to measure and how to visualize
  - Rolling windows for metrics are tricky to get right
- Serving
  - How to support resource requirements of new model types (foundation models, etc.)
- Productionizing
  - Often still takes too long for machine learning models to go live
    - Undermines ML ROI
  - Tools such as Metaflow are steps forward

Further Reading

- MLOPs Community
- https://storage.googleapis.com/pub-tools-public-publication-data/pdf/43146.pdf
- https://github.com/eugeneyan/applied-ml
- https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
- https://fullstackdeeplearning.com/spring2021/lecture-13/
- https://www.shreya-shankar.com/rethinking-ml-monitoring-1/
- https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Thanks for Listening!

@mihail_eric