MLOPS:
AS BOAS PRÁTICAS PARA O DESENVOLVIMENTO DE MODELOS DE ML

Ana Flávia Souza

SUMÁRIO

- Contextualizando MLOps
- Cinco palavras que resumem tudo
- Ciclo de vida dos modelos de ML
  - Desenvolvimento
  - Pré-produção
  - Deployment
  - Monitoramento
  - Extra: Governança
- Momento Pseudo-*Sommelier* de Cloud Services: automatização com GCP

TREVEIL, Mark *et al*, **Introducing MLOps: How to Scale Machine Learning in the Enterprise**, [s.l.]: O'Reilly Media, Incorporated, 2020.

CONTEXTUALIZANDO MLOPS

CONTEXTUALIZANDO MLOPS

- MLOps é um acrônimo para *Machine Learning Operations;*
- Originado do conceito de DevOps, segue alguns princípios bem semelhantes com uma pitada de problemas próprios a serem resolvidos;
- Esses princípios buscam tornar a linha de produção de modelos de ML mais rápida, eficiente e escalável, sem perder a qualidade do produto final

CINCO PALAVRAS QUE RESUMEM TUDO

(Na minha humilde opinião)

CINCO PALAVRAS QUE RESUMEM TUDO

O processo de ML consiste em diversos times que nem sempre conversam na mesma "língua" ou nas mesmas tecnologias.

O ideal é que cada indivíduo envolvido na produção e usuário dos modelos saibam exatamente o que está acontecendo na pipeline.

Automatizar uma tomada de decisão impacta muito a vidas das pessoas e os riscos envolvidos nisso devem ser altamente considerados na produção.

Melhorias e alterações nos modelos ou na estrutura de produção devem ser feitas de forma rápida.

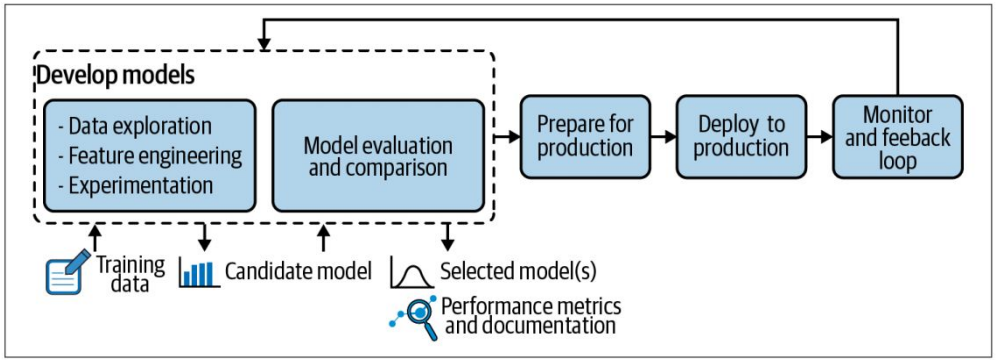
Entregas rápidas não significam entregas malfeitas.

CICLO DE VIDA DOS MODELOS

(Quais são os pontos de atenção em cada fase)

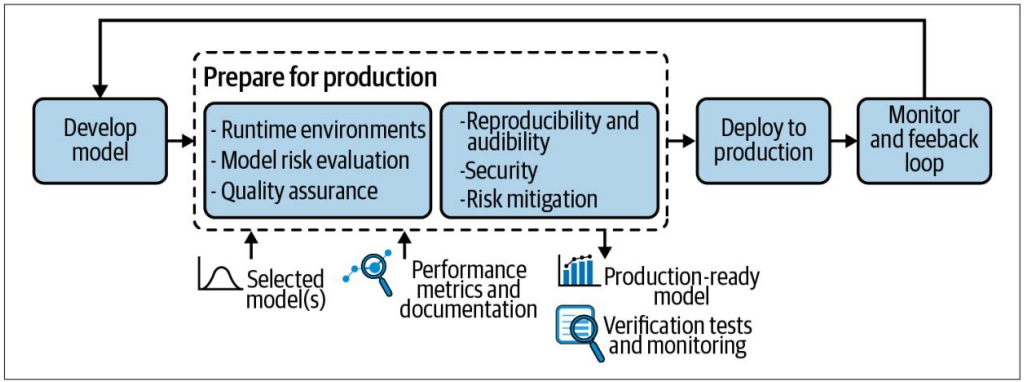
DESENVOLVIMENTO

- Qualidade e bom entendimento dos dados;
- Seleção do modelo segundo risco, custos e benefícios;
- Boas práticas de modelagem (sets de treino e teste, validação cruzada);
- Experimentação;
- Avaliação das métricas apropriadas;
- Explicabilidade (*feature importance plots* e *Shap values*)



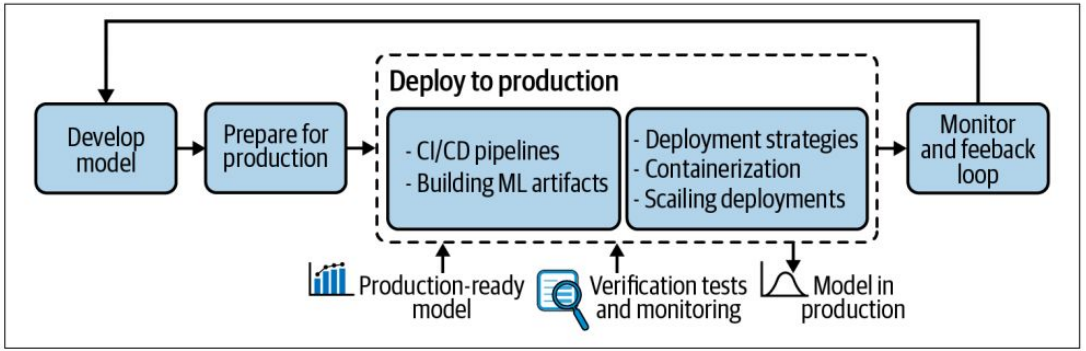
PRÉ-PRODUÇÃO

- O ambiente de desenvolvimento é compatível com o ambiente de produção?
- Testes de validação do modelo (*fairness*, explicabilidade, reprodutibilidade);
- *Quality Assurance* (QA)



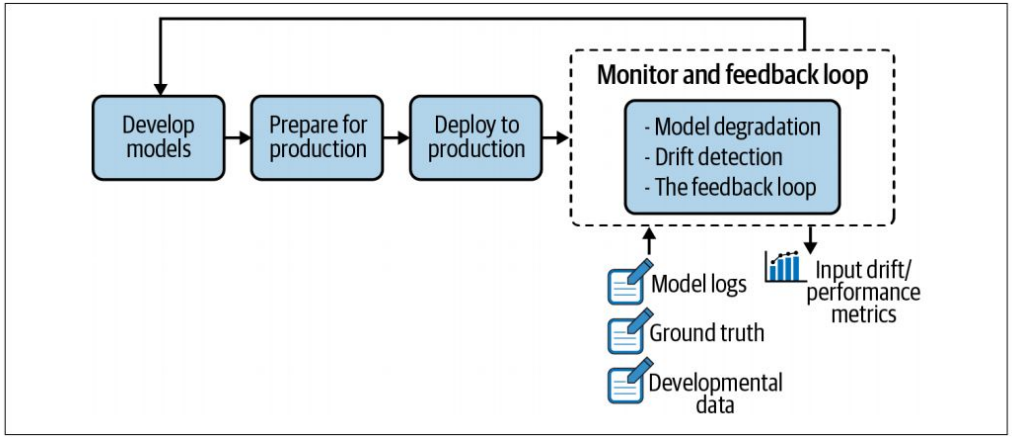
DEPLOYMENT

- CI/CD pipelines associadas ao treino contínuo;
- Artifacts e controle de versão;
- Máxima documentação possível;
- Como lidar com downtime? (estratégias de deployment)
- Monitoramento dos recursos computacionais e das métricas dos modelos;
- Estrutura escalável (containers, Kubernetes)



MONITORAMENTO

- Feedback loops:
  - Logging
  - Análise de performance dos modelos por ground truth ou input drifts;
- Problemas no modelo em execução podem ser resolvidos por retreino ou construção de novos modelos



GOVERNANÇA

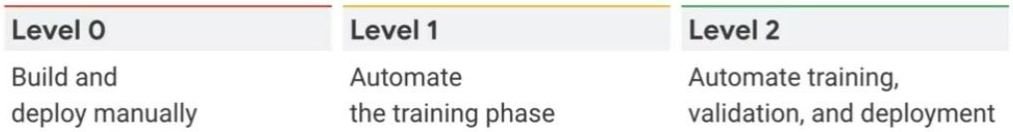
- Esse ponto se refere ao famoso “fazer IA seguindo os protocolos da instituição”;
- Responsible AI e seus impactos na organização

AUTOMATIZAÇÃO COM NUVEM

(Porque a nuvem é uma boa opção para MLOps)

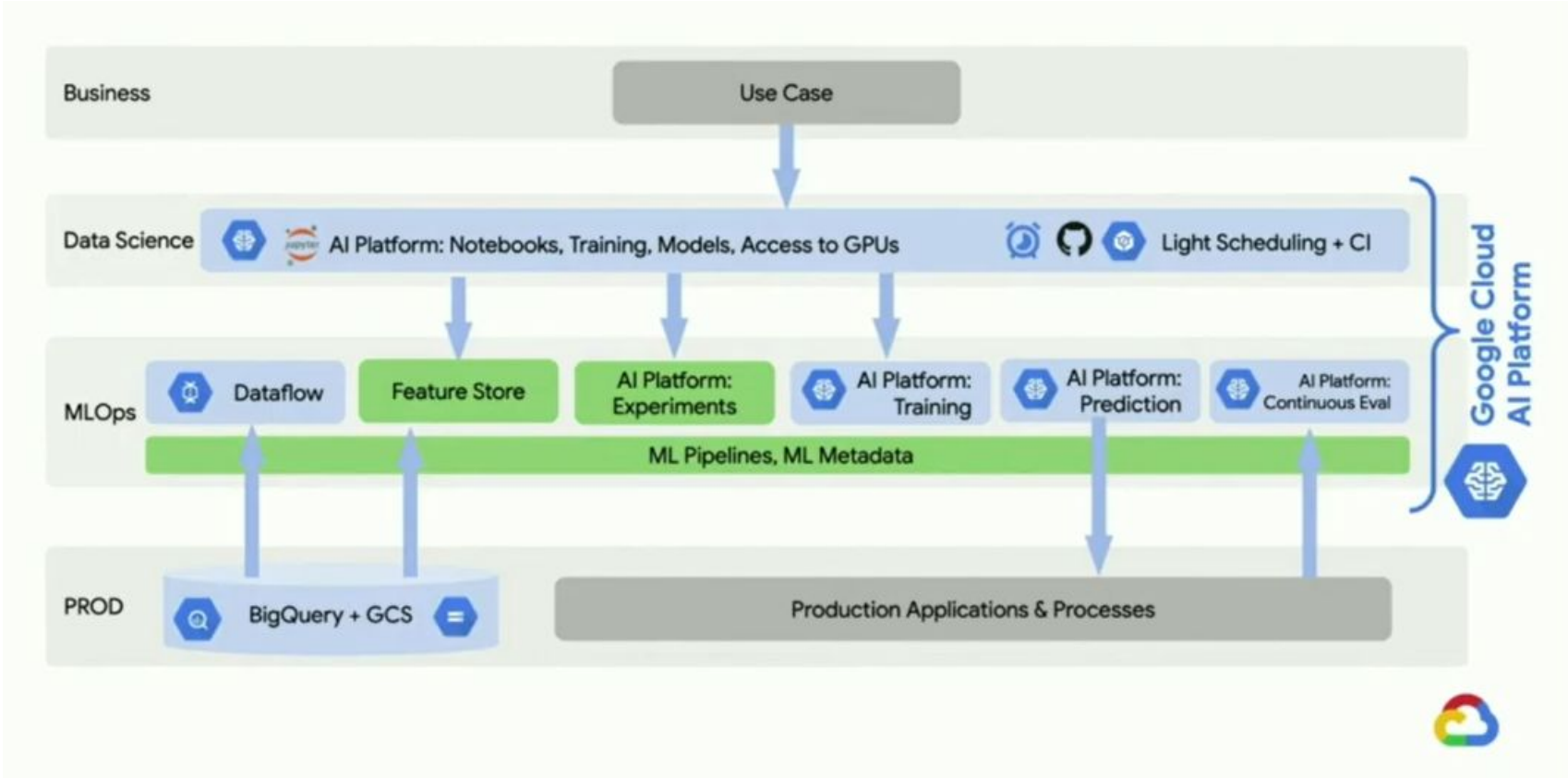
SOBRE A AUTOMATIZAÇÃO

- Fazer todos esses processos na mão pode levar a vários problemas e erros no futuro;
- Integrar tecnologias isoladas entre si também pode apresentar problemas na fase de produção;
- Automatizar é a melhor estratégia para lidar com esses problemas e podemos dizer que quanto mais automatizado o ambiente mais maduro o seu processo;



SOBRE A AUTOMATIZAÇÃO (CONT.)

- A vantagem de usar os serviços de Nuvem (aqui, vamos falar de GCP, mas os outros fazem isso também) é que ela oferece diversos produtos para as diferentes fases da do processo de produção, que podem ser adaptados de acordo com a necessidade do projeto;
- Isso facilita os processos de criação das pipelines e do ambiente de deployment, e do deployment em si.


Business
Use Case
Data Science
AI Platform: Notebooks, Training, Models, Access to GPUs
Light Scheduling + CI
MLOps
Dataflow
Feature Store
AI Platform: Experiments
AI Platform: Training
AI Platform: Prediction
AI Platform: Continuous Eval
ML Pipelines, ML Metadata
PROD
BigQuery + GCS
Production Applications & Processes
Google Cloud AI Platform

OBRIGADA!