3NG TECH
INTELLIGENT SOLUTIONS

Software Engineering and DevOps for ML

By /in/renzoe-guarise/ ML Engineer at MercadoLibre & CTO at 3ng.tech

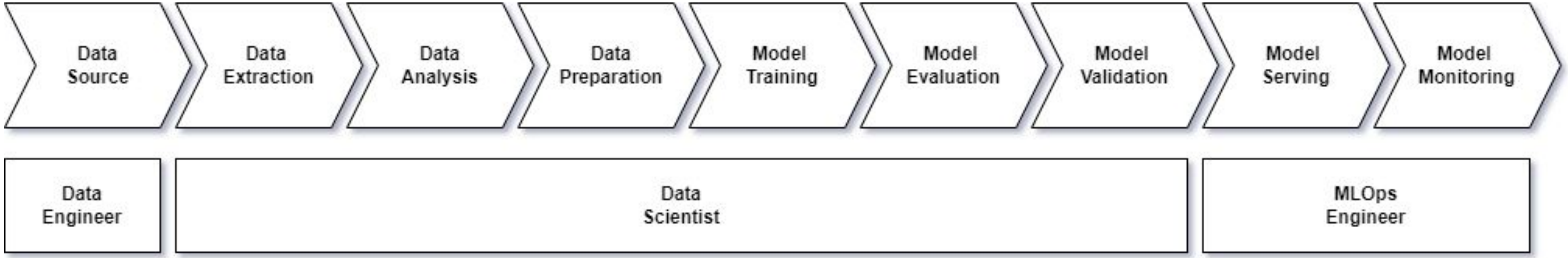
Agenda

1. Objectives
2. ML Basics
3. What is MLOps?
4. Implementing MLOps
5. Planning and Management
6. Q/A

1.1 Objectives

- Key concepts
- Advantages of its implementation
- The life cycle of a Machine Learning project
- Challenges of its implementation
- Planning and Management of a ML project

2.1 ML Life Cycle

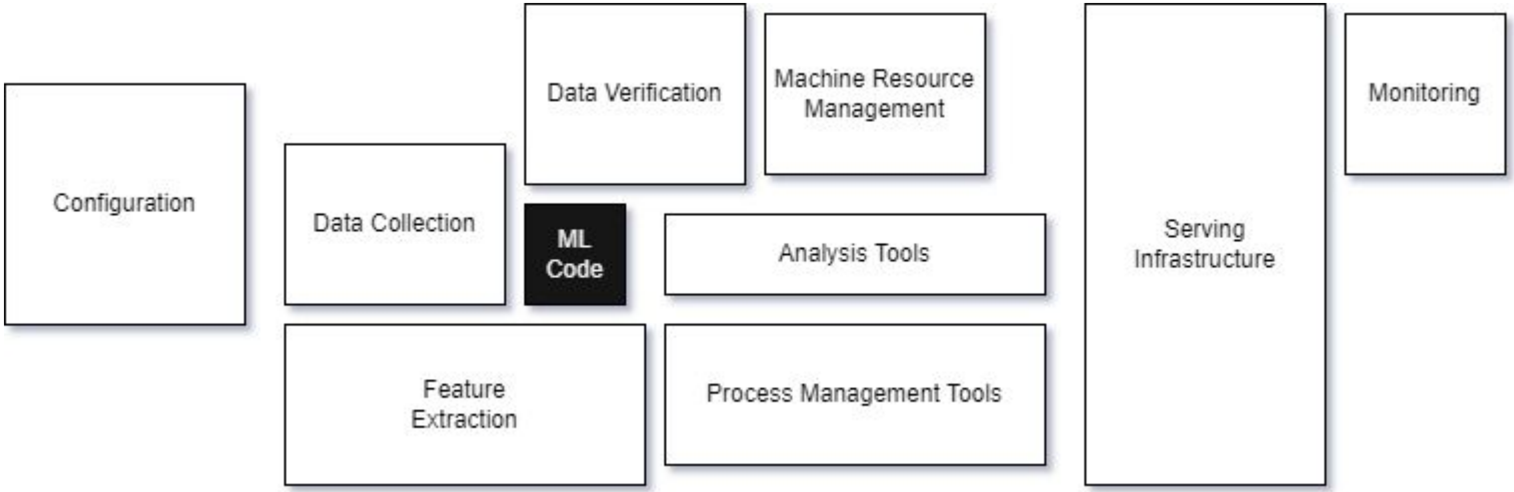


2.1 ML Basics

“Machine Learning is like the raisins in a raisin bread:

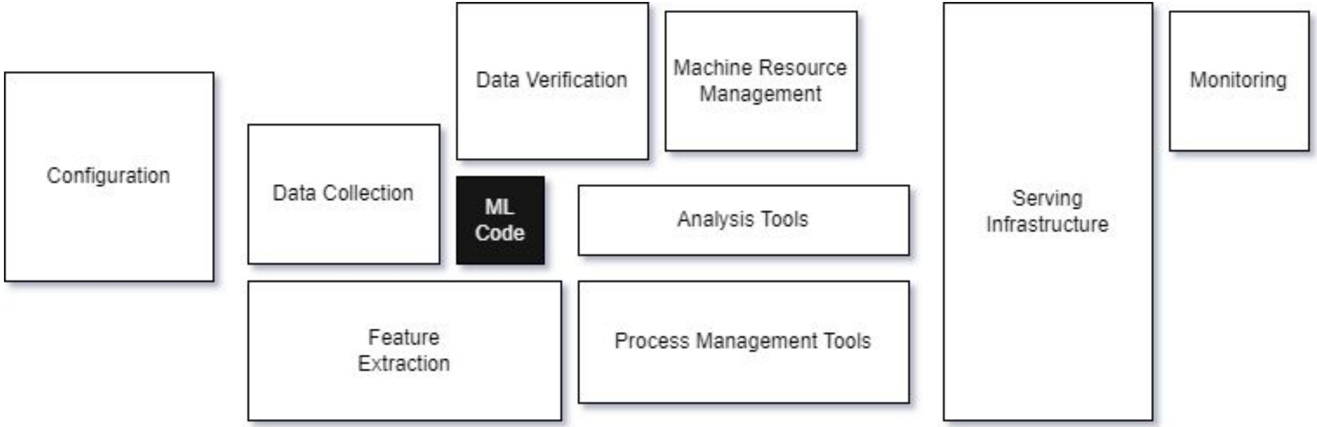
it’s just a few tiny raisins but without it you would just have plain bread”.

Peter Norvig



2.1 ML Basics

To do Machine Learning is to do Software, so good old Software Engineering applies.



But! …Do these raising come with specific Software Engineering practices?

3.1 What is MLOps?

MLOps is a set of good practices that provides benefits such as:

- shortening the development cycles,
- increasing deployment and releases velocity.

In order to achieve this benefits, we introduce:

- CD (Continuous Delivery)
- CI (Continuous Integration)

but with more seasoning

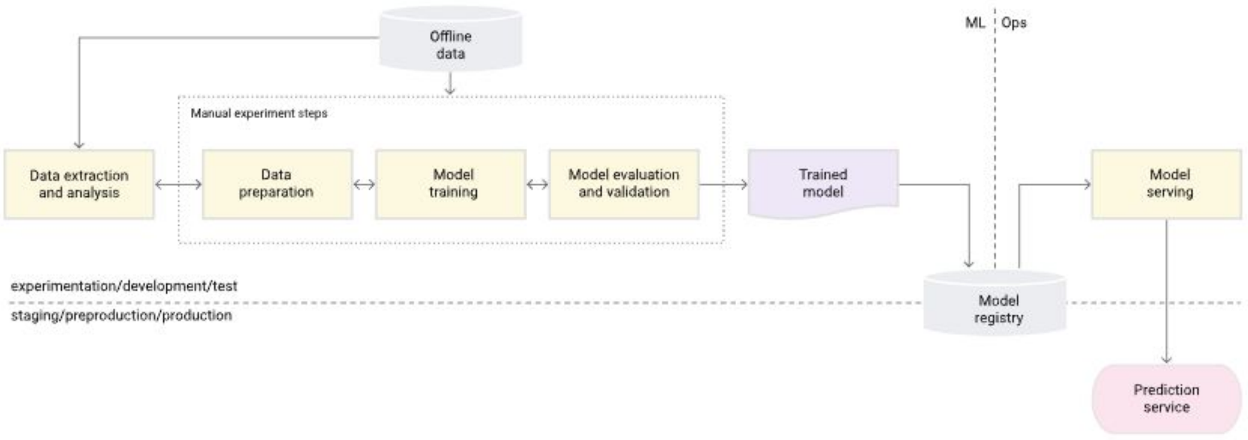
- CT(Continuous Train)

3.2 What is different?

- Team skills
- Development
- Testing
- Deployment
- Production

4.1 Why do we need MLOps?

MLOps level 0: Manual process



Source: [GCP Cloud Architecture Center]()

4.1 Why do we need MLOps?

MLOps level 0: Manual process

Characteristics

- Manual, script-driven, and interactive process
- Disconnection between ML and operations
- Infrequent release iterations
- No CI/CD
- Deployment refers to the prediction service
- Lack of active performance monitoring

4.1 Why do we need MLOps?

MLOps level 0: Manual process

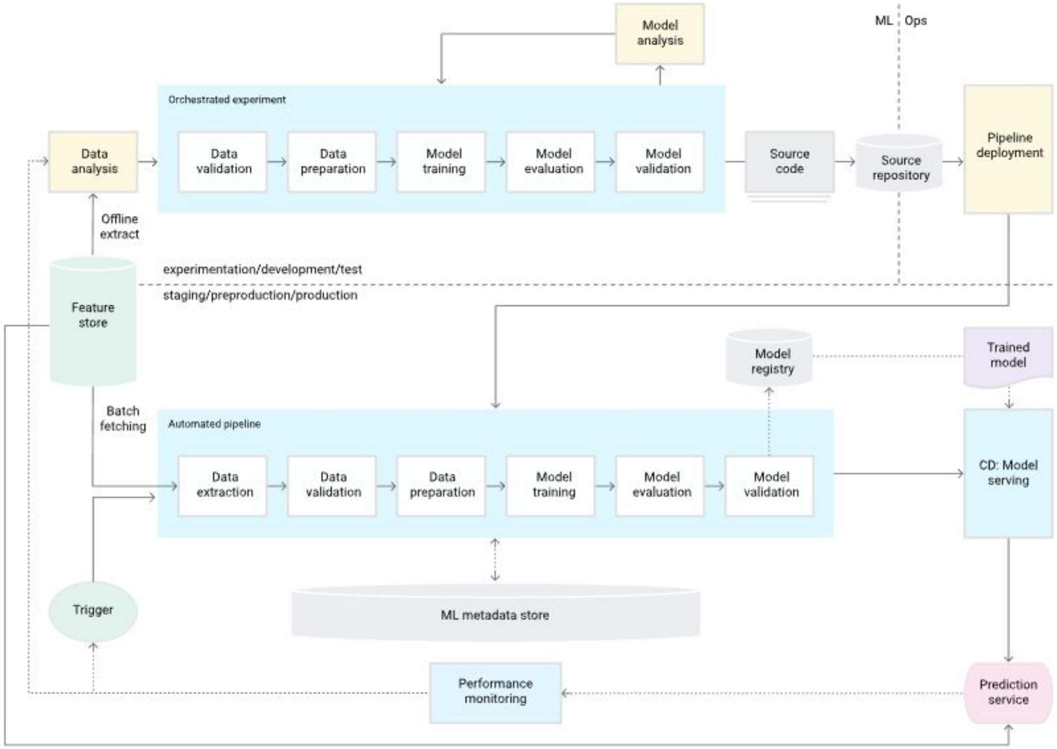
Challenges

- Actively monitor the quality of your model in production
- Frequently retrain your production models
- Continuously experiment with new implementations to produce the model

4.2 How do we apply MLOps?

MLOps level 1:

ML pipeline automation



Source: [GCP Cloud Architecture Center]

4.2 How do we apply MLOps?

MLOps level 1:

ML pipeline automation

Characteristics

- Rapid experiment
- CT of the model in production
- Experimental-operational symmetry
- Modularized code for components and pipelines
- Continuous delivery of models
- Pipeline deployment

4.2 How do we apply MLOps?

MLOps level 1: ML pipeline automation

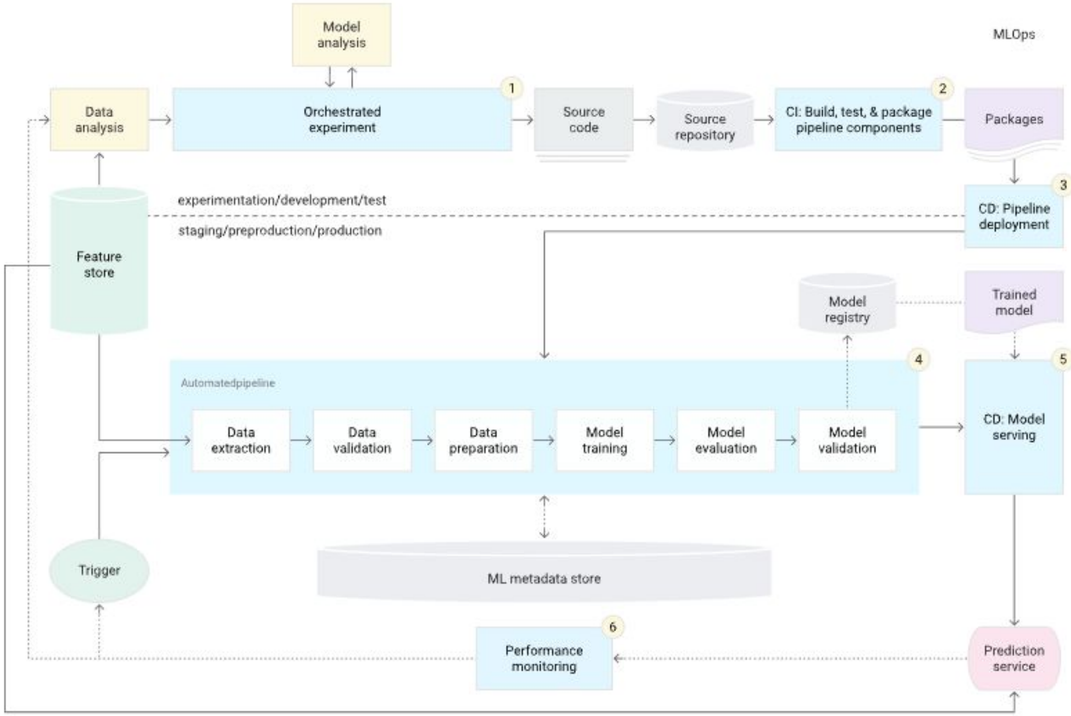
Challenges

- This setup is suitable when you deploy new models based on new data, rather than based on new ML ideas.

4.2 How do we apply MLOps?

MLOps level 3:

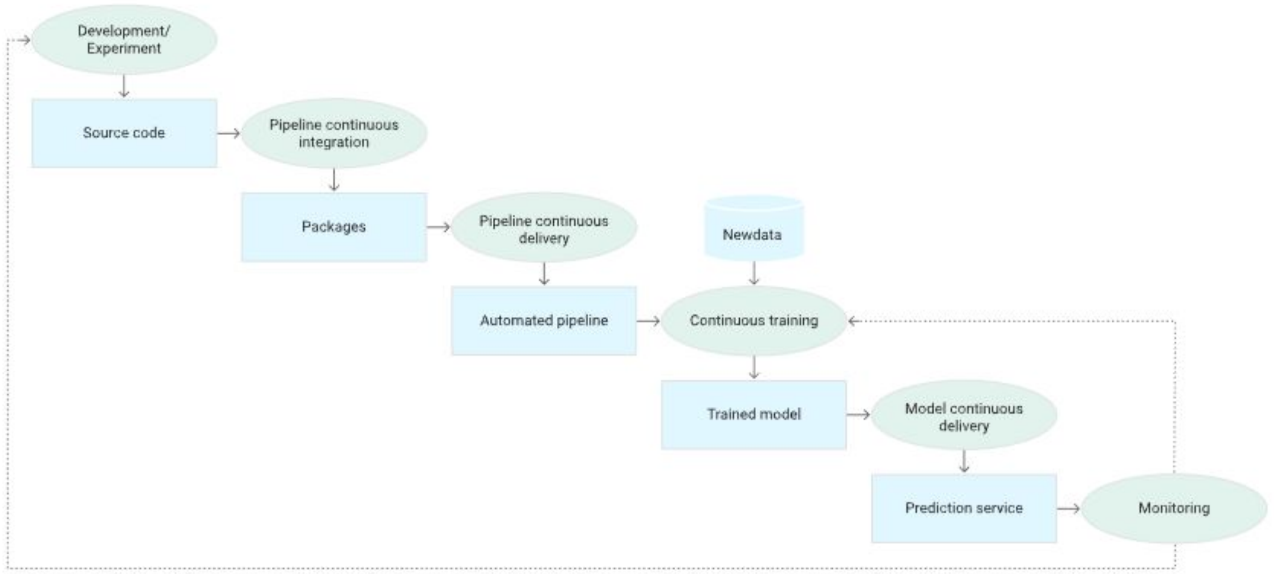
CI/CD pipeline automation



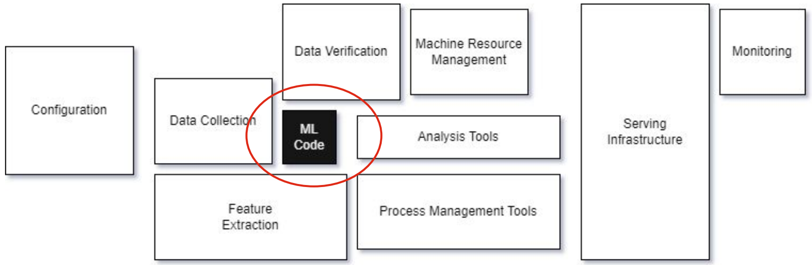
Source: [GCP Cloud Architecture Center]

4.2 How do we apply MLOps?

MLOps End Game



5.1 Uncertainty in planning



- Is it feasible? Can this be solved at all?
- How much predictive quality will we achieve?
- How many months to reach production?
- Is it suitable for Machine Learning?
- Wich attack angle for this problem has the best Roi?
- Should we implement online predictions or pre-computed serving?
- Which framing classification, ranking-retrieval, pairwise comparison, etc. should we put into practice?
- What delay can we achieve?

5.2 Tools #1

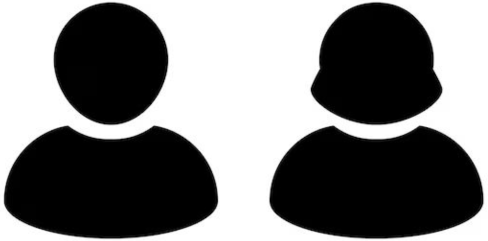
Human Feasibility

5.2 Tools #1

Human Feasibility

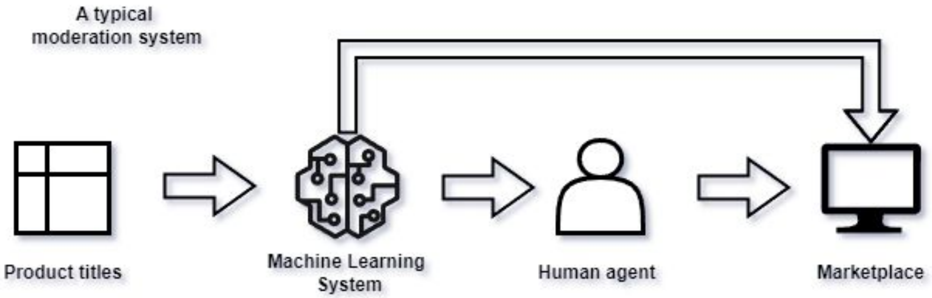
Rule of thumb:

Human agreement also limits
Machine Learning performance



5.2 Tools #2

Time-Quality tradeoff



5.2 Tools #2

Time-Quality tradeoff

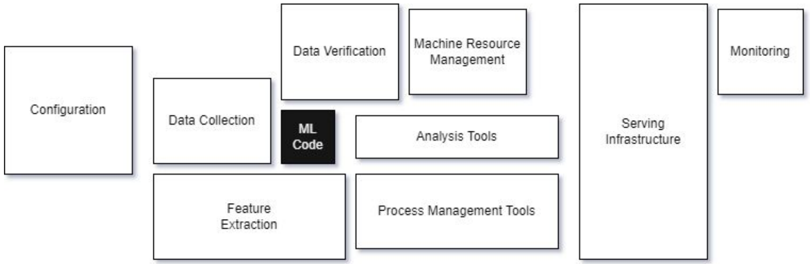
```python
def totally_Acurate_meds_detector(product_title):
    if 'antibiotic' in product_title:
      return 'meds'
    return 'not_meds'
```

1 minute –worst accuracy
1 day – baseline accuracy
1 week – decent accuracy
1 year – awesome accuracy

5.1 Uncertainty in planning

Human Feasibility
Time-Quality tradeoff

- Is it feasible? Can this be solved at all?
- How much predictive quality we will achieve?
- How many months to reach production?
- Is this good fit for Machine Learning?
- Wich attack angle for this problem has the best Roi?
- Online predictions or pre-computed serving?
- Which framing: classification / ranking-retrieval/ pairwise comparison / etc
- What latency can we achieve?



3NG TECH
INTELLIGENT SOLUTIONS

Estrategia y Gobierno de Datos

Arquitectura e Ingeniería de Datos

Visualización y Analytics

Ciencia, ML e IA

Productos de Datos

Cloud Services

3NG TECH
INTELLIGENT SOLUTIONS

¡Muchas Gracias!

Seguinos y contactanos:

/company/3ng-tech

/3ng.tech

3ng.tech

/3ngtech