Transforming ML models into robust services in production using MLOps

Carlos Maestre

March 2021

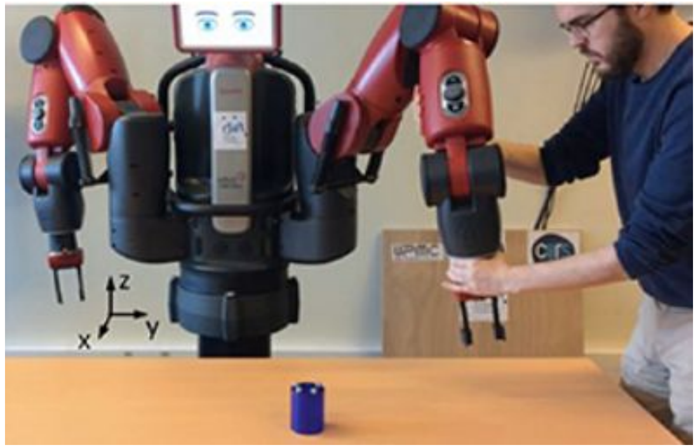
- Background
  - Researcher (PhD. in Robotics and Machine learning)
  - Software engineer
- Actual position
  - Machine learning engineer

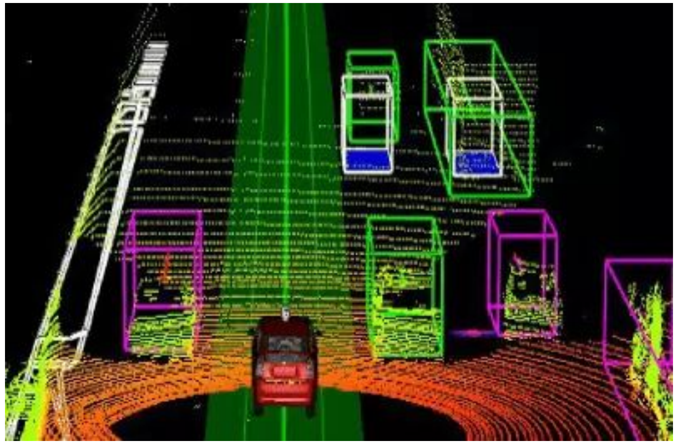
frontiers in Neurorobotics

Action Generation Adapted to Low-Level and High-Level Robot-Object Interaction States

Carlos Maestre[1*], Ghanim Mukhtar[1], Christophe Gonzales[2] and Stephane Doncieux[1]

[1]UMR 7222, ISIR, Sorbonne Université and CNRS, Paris, France





What is the actual status of Machine learning in the industry?

The majority of business analytics and AI projects are still failing

By **Yulia Kosarenko** April 30, 2020 4482 1

Failure rates for analytics, AI, and big data projects = 85% – yikes!

July 23, 2019 by Brian T. O'Neill

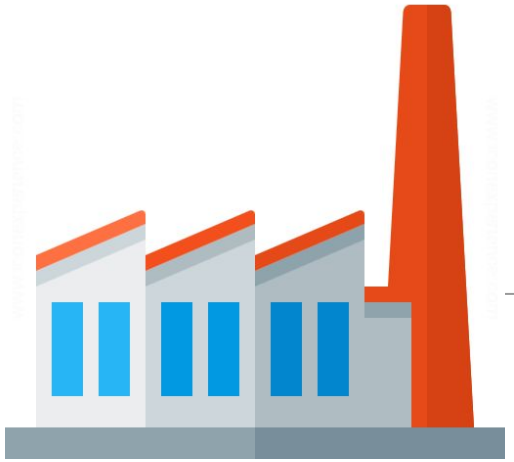
If Data Scientists are so smart, why do 70% of their projects fail?

Published on June 20, 2019

https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/
https://www.datadriveninvestor.com/2020/04/30/the-majority-of-business-analytics-and-ai-projects-are-still-failing/
https://www.linkedin.com/pulse/data-scientist-so-smart-why-do-70-projects-fail-mike-fish/

Why?

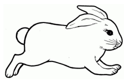
Tomatoes classifier : Actual approach



PRODUCTION

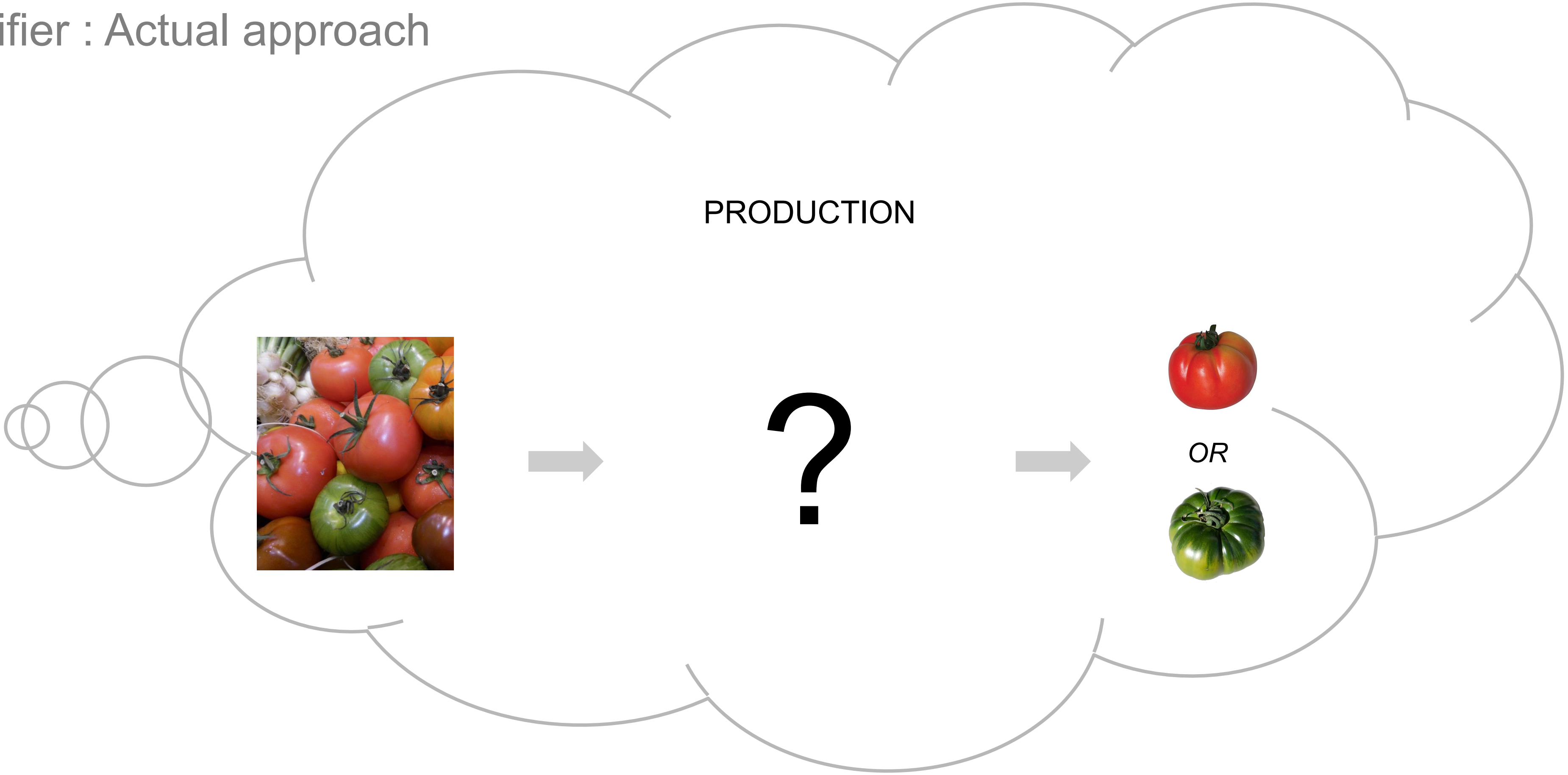
OR

COST : 

SPEED : 

SUCCESS RATE : 

Tomatoes classifier : Actual approach



COST : Lower

SPEED : Faster

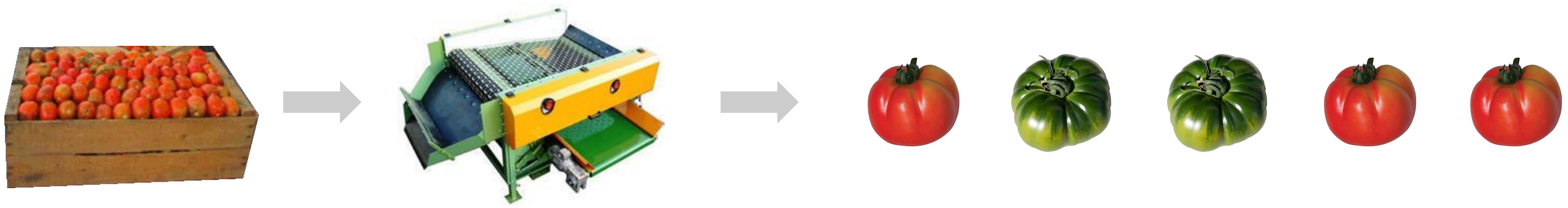
SUCCESS RATE : Similar or better

Tomatoes classifier : Fast classifier machine

Tomatoes classifier : Fast classifier machine

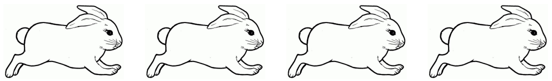
Tomatoes classifier : Machine : Working in a controlled environment

DEVELOPMENT



(HALF) COST :

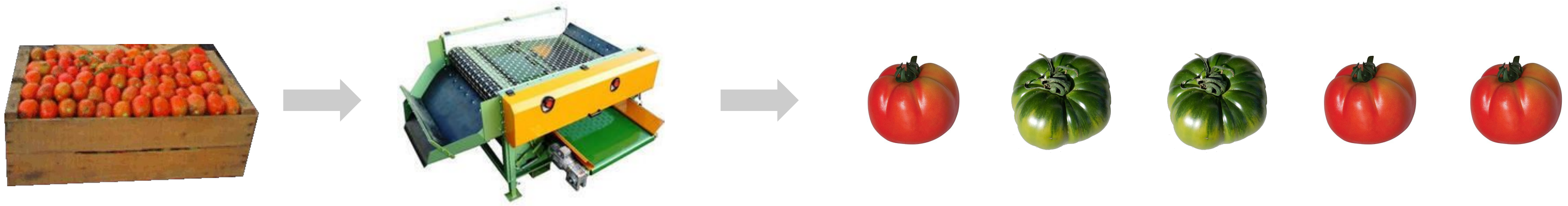
(DOUBLE) SPEED :



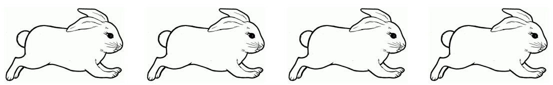
(SIMILAR) RATE :

Tomatoes classifier : Machine : Working under similar conditions

PRODUCTION TEST



COST : 

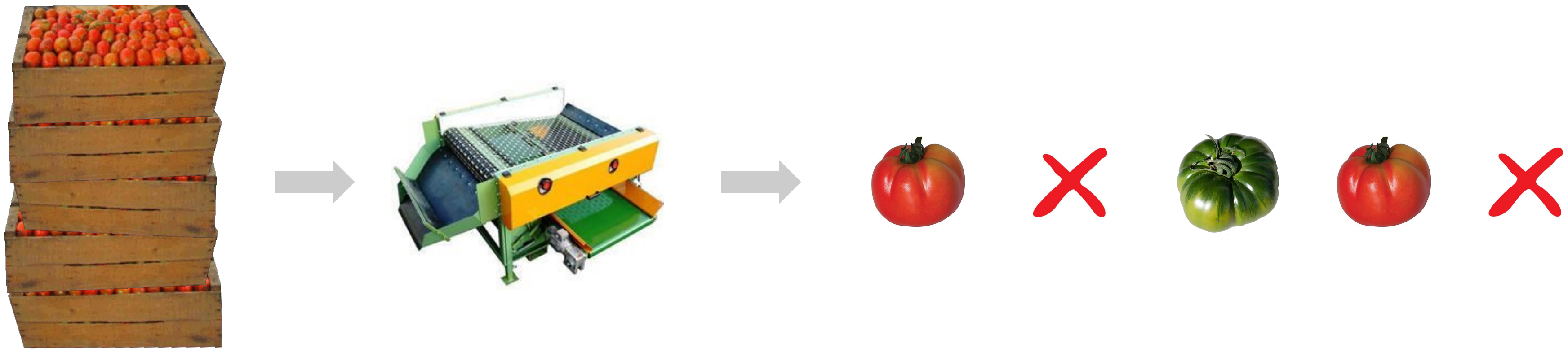
SPEED : 

SUCCESS RATE : 

UNEXPECTED PRODUCTION COST

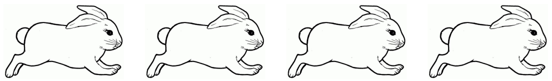
Tomatoes classifier : Machine : Stops working with high workload

PRODUCTION



COST :

SPEED :

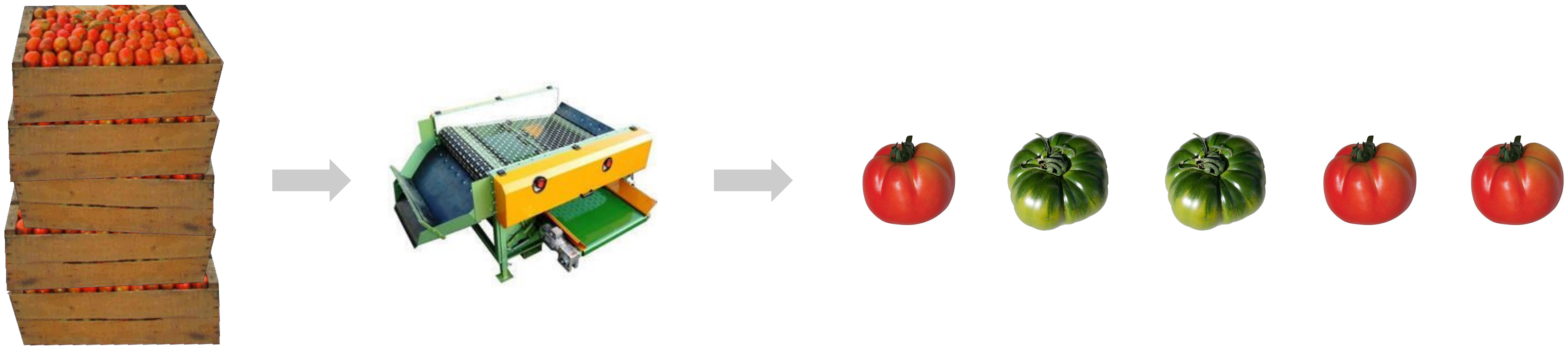


SUCCESS RATE :

UNEXPECTED
WORKLOAD FAILURE

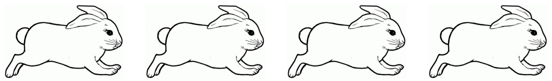
Tomatoes classifier : Machine : Stops working with high workload

PRODUCTION



COST :

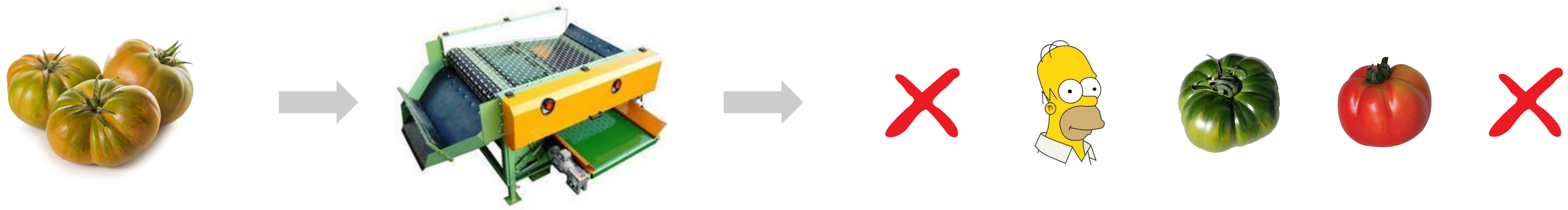
SPEED :



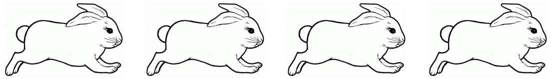
SUCCESS RATE :

Tomatoes classifier : Machine : Misclassified unseen types of tomato

PRODUCTION



COST : 

SPEED : 

SUCCESS RATE : 

UNEXPECTED
INPUT TYPE

Tomatoes classifier : Machine : Where to invest more money ?

VS

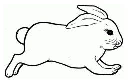
UNREALISTIC EXPECTATIONS

WELL KNOWN APPROACH

Tomatoes classifier : Final approach

PRODUCTION

COST : 

SPEED : 

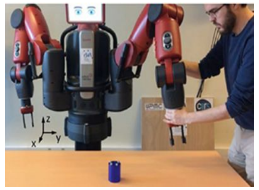
SUCCESS RATE : 

Tomatoes classifier : Machine : Failure analysis

Problem	Description	Solution
UNEXPECTED PRODUCTION COST	Lost of traceability between environments	Development trace
UNEXPECTED WORKLOAD FAILURE	Small number of tomatoes	Workload test
UNEXPECTED INPUT TYPE	Same type of tomatoes	Add tomato diversity during development
UNREALISTIC EXPECTATIONS	ML hype	Define realistic metrics

Tomatoes classifier : Machine : Failure analysis

Problem	Description	Solution
UNEXPECTED PRODUCTION COST	Lost of traceability between environments	Development trace
UNEXPECTED WORKLOAD FAILURE	Small number of ~~tomatoes~~ images	Workload test
UNEXPECTED INPUT TYPE	Same type of ~~tomatoes~~ image	Add tomato diversity during development
UNREALISTIC EXPECTATIONS	ML hype	Define realistic metrics

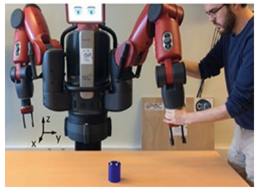


→ Human? → ✓ OR ✗

Tomatoes classifier : Machine : Failure analysis

Problem	Description	Solution	Best practices
UNEXPECTED PRODUCTION COST	Lost of traceability between environments	Development trace	Governance
UNEXPECTED WORKLOAD FAILURE	Small number of ~~tomatoes~~ images	Workload test	Quality assurance (QA)
UNEXPECTED INPUT TYPE	Same type of ~~tomatoes~~ image	Add tomato diversity during development	Robust services
UNREALISTIC EXPECTATIONS	ML hype	Define realistic metrics	Business-oriented metrics

MLOps



Human?

✓ OR ✗

MLOps

From Wikipedia, the free encyclopedia

MLOps (a compound of “machine learning” and “operations”) is a practice for collaboration and communication between data scientists and operations professionals to help manage production ML (or deep learning) lifecycle.

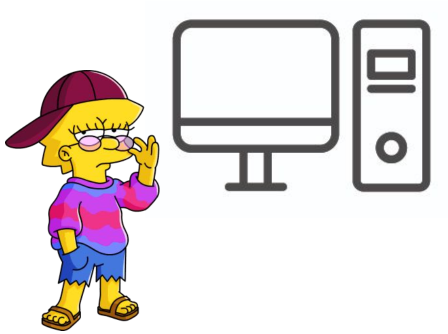
to build robust ML services and

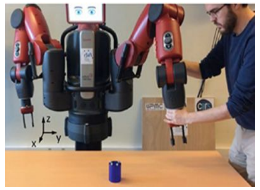
- ➔ New field: few standards
- ➔ New mindset
- ➔ New frameworks
- ➔ Fast changes

https://www.wikipedia.org

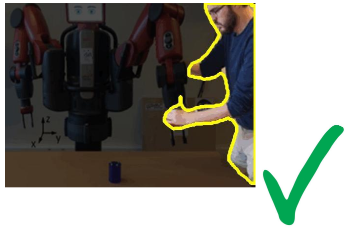
ML service

Model (development environment)





Human?
(model)



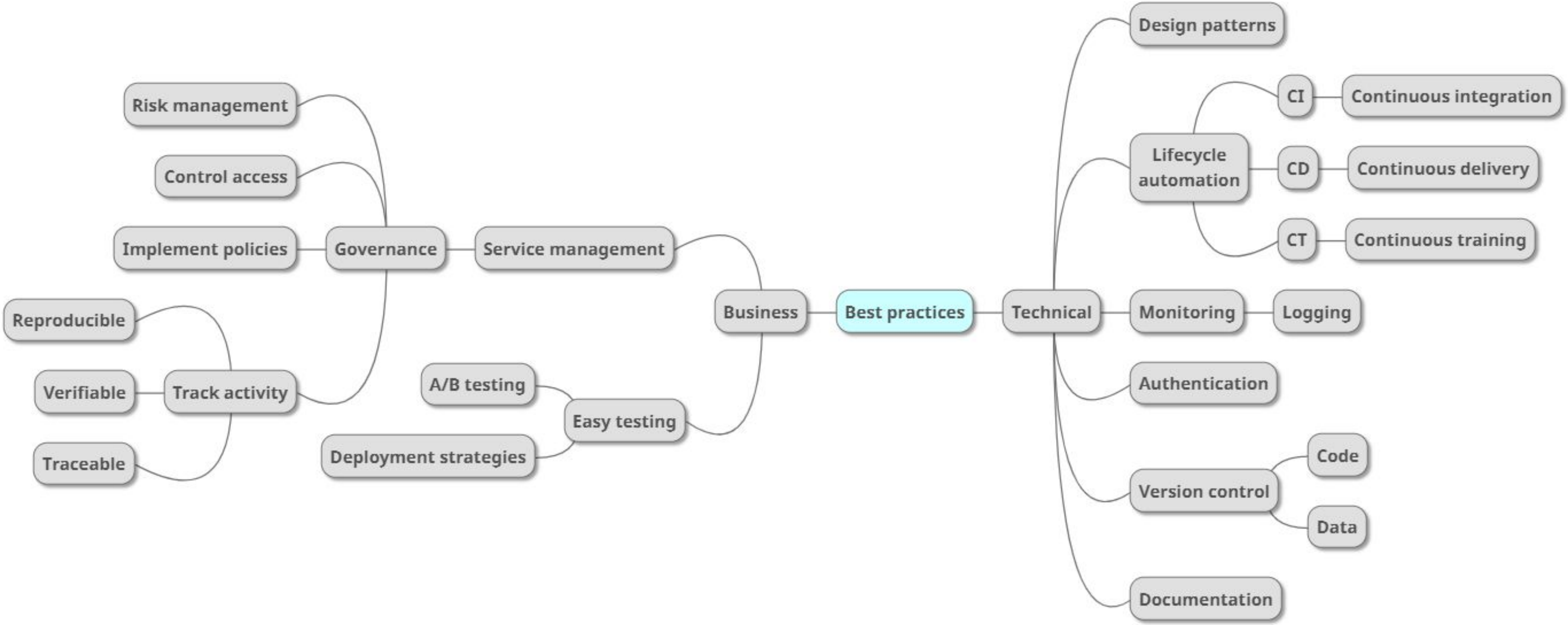
Service (production environment)

Human?
(model)

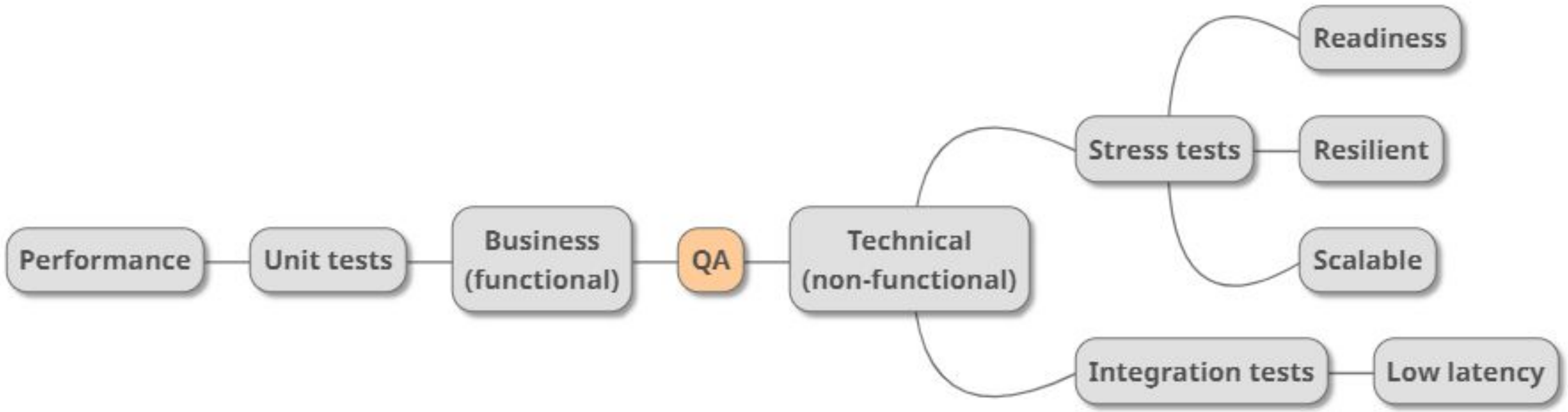
(service)

What are the features of a robust ML service?

MLOps : Best practices

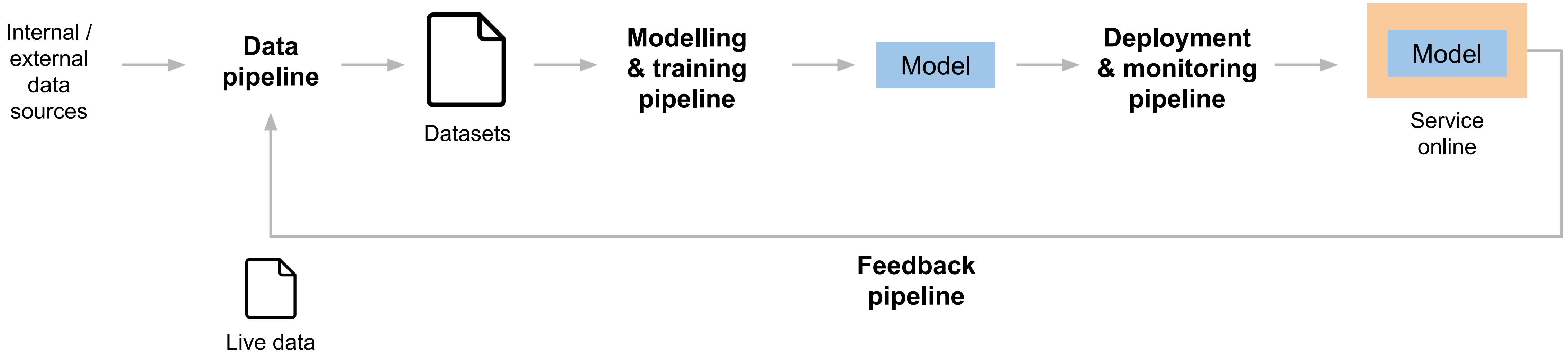


MLOps : QA based on indicators

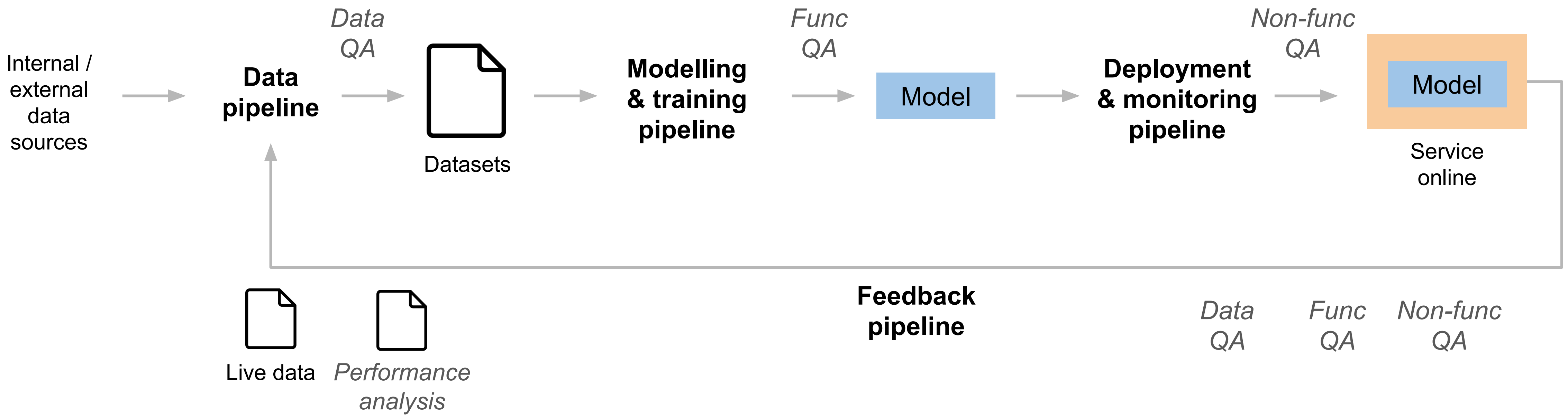


How to build these features?

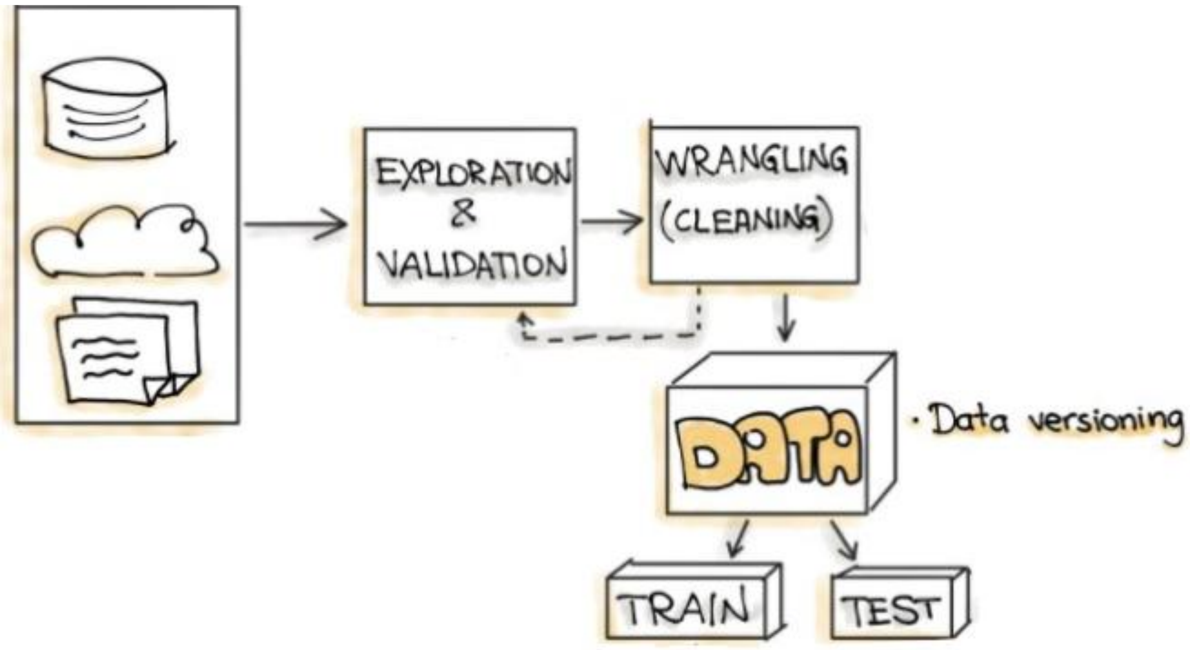
MLOps : ML service lifecycle



MLOps : ML service lifecycle



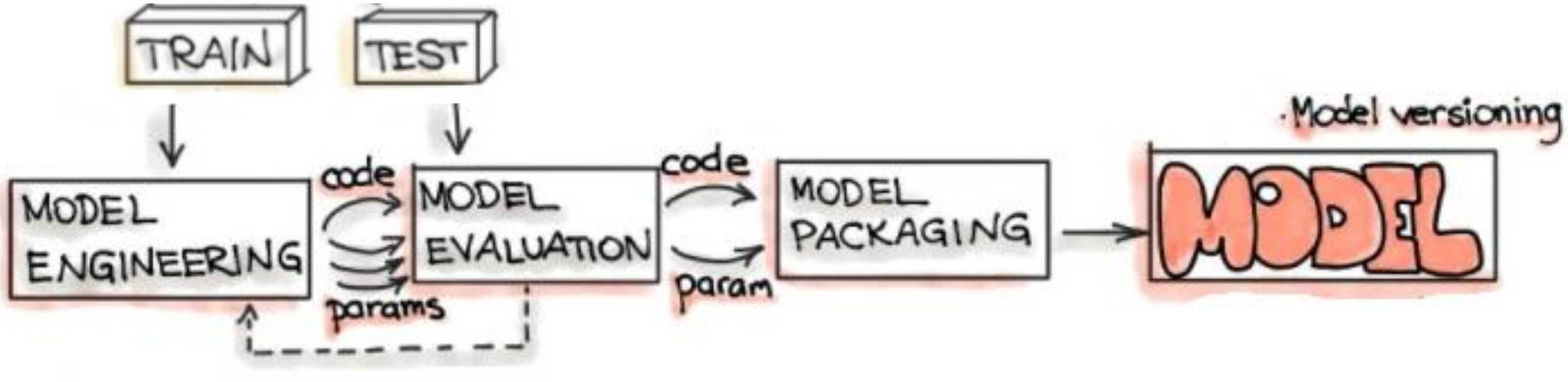
MLOps : Data pipeline



- Data is the new oil !
  - Poor data + good model = poor results
  - Good data + poor model = fine results
  - Good data + good model = good results

https://ml-ops.org

MLOps : Modelling / Training pipeline



Creating a model comprehends 3 steps:

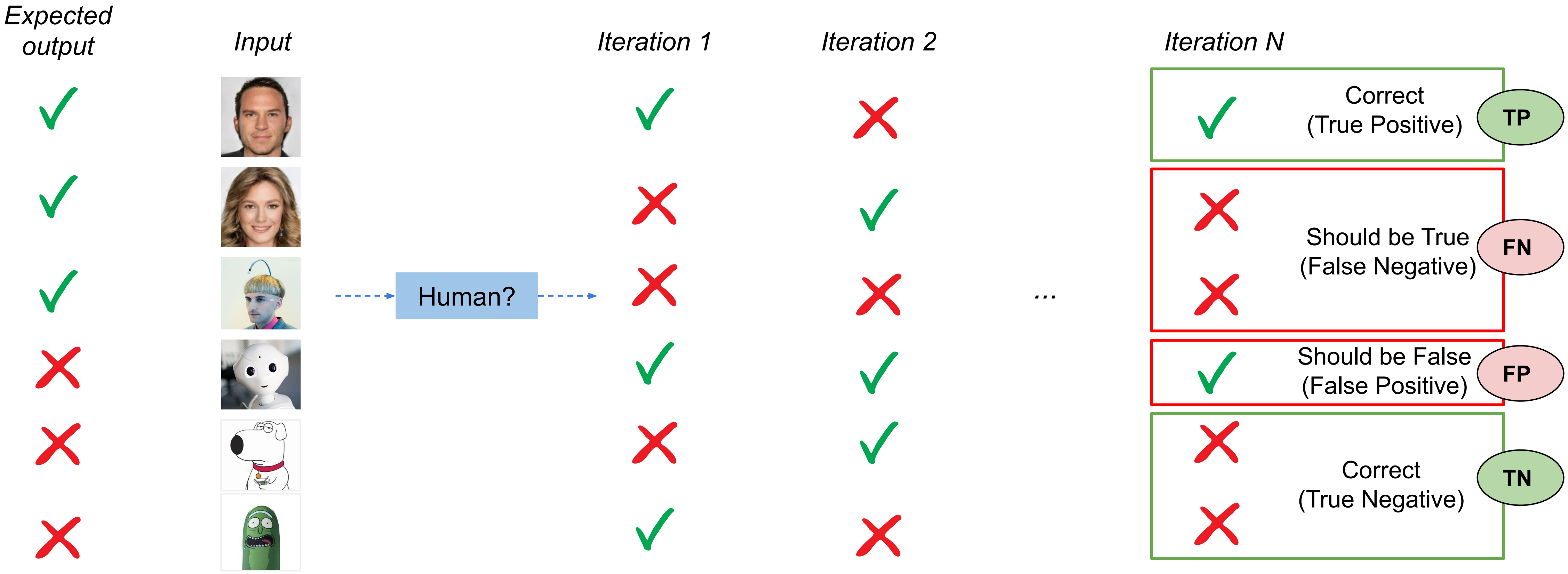
1. Design
2. Training
3. Evaluation

MLOps : Modelling / Training pipeline

Model training consist of given

- some examples as input
- and some corresponding outputs

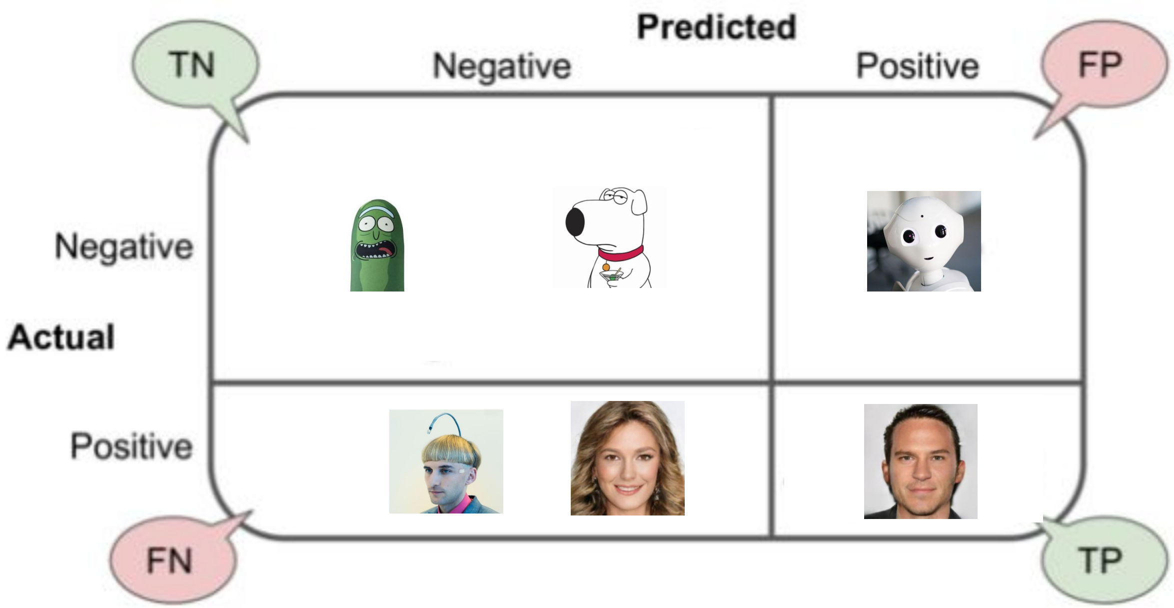
a model generates an internal configuration that reproduces the outputs (and generalises to unseen inputs)



MLOps : Modelling / Training pipeline

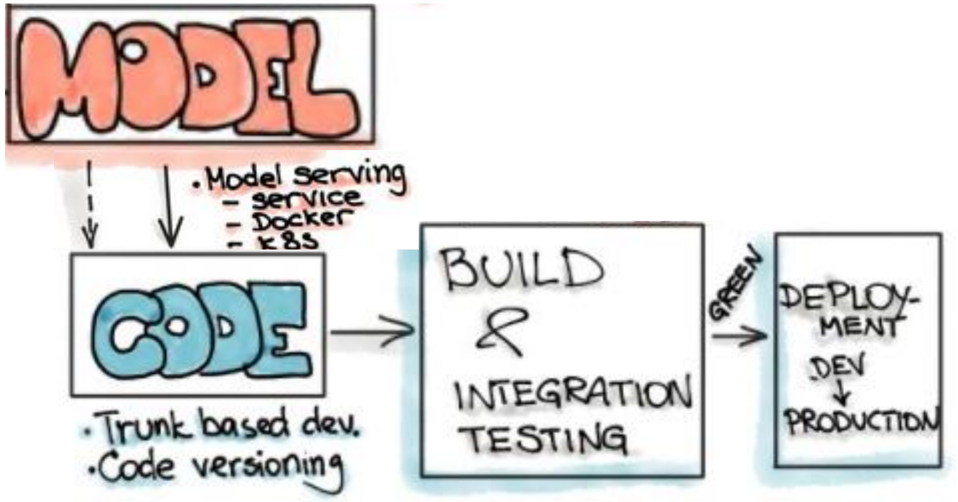
<u>Model evaluation</u> (functional indicators)

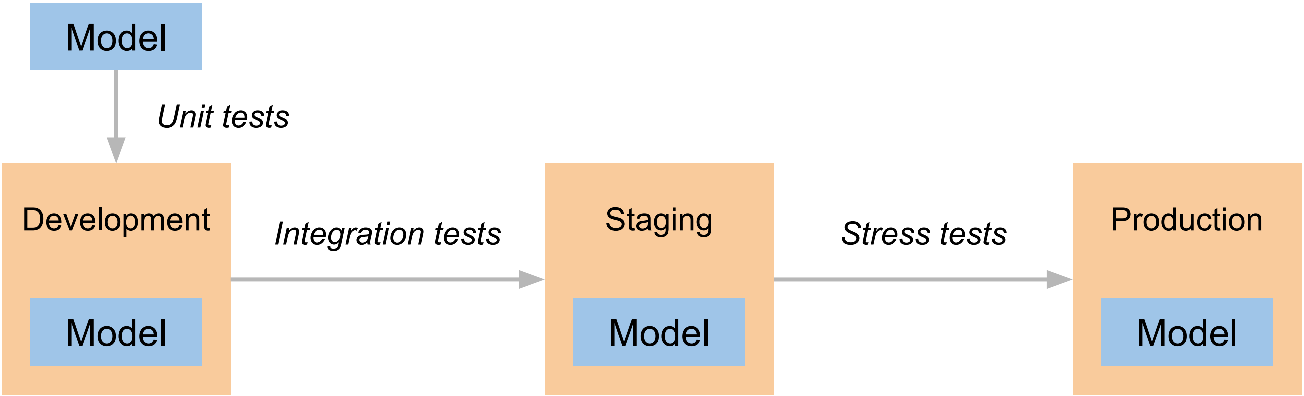
Precision : fraction of cases predicted positive (human) which are actually positive (human)
Recall : fraction of positives (all humans) that have been correctly predicted positive (human)
F1-score : value in range [0, 1] computed using precision and recall



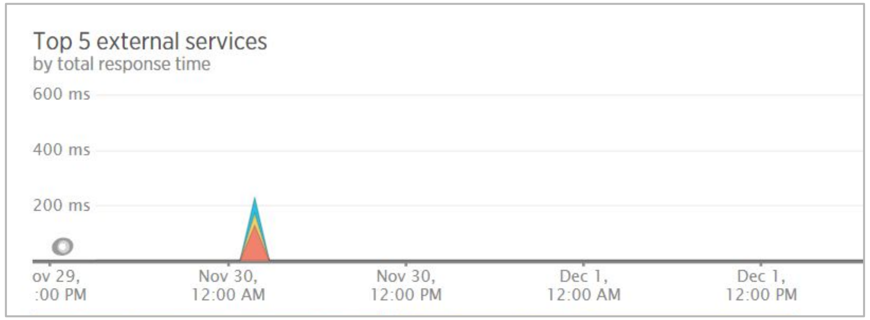
Precision : 50% (1/2)
Recall : 33% (1/3)
F1 : ~ 0.4

MLOps : Deployment / Monitoring pipeline

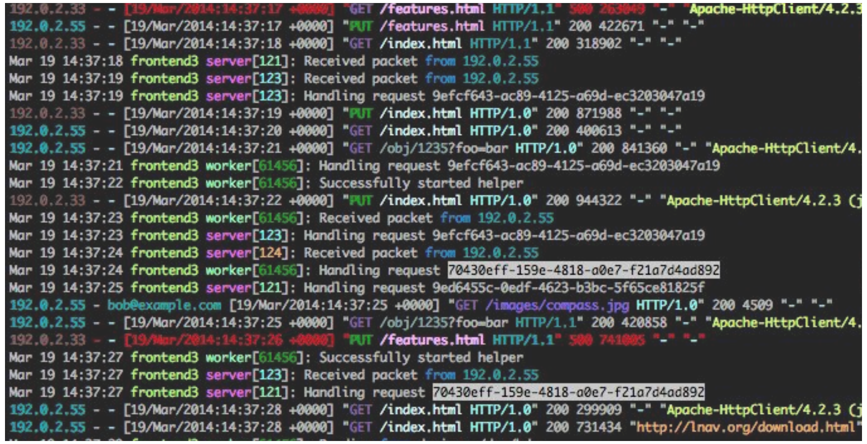




Monitoring

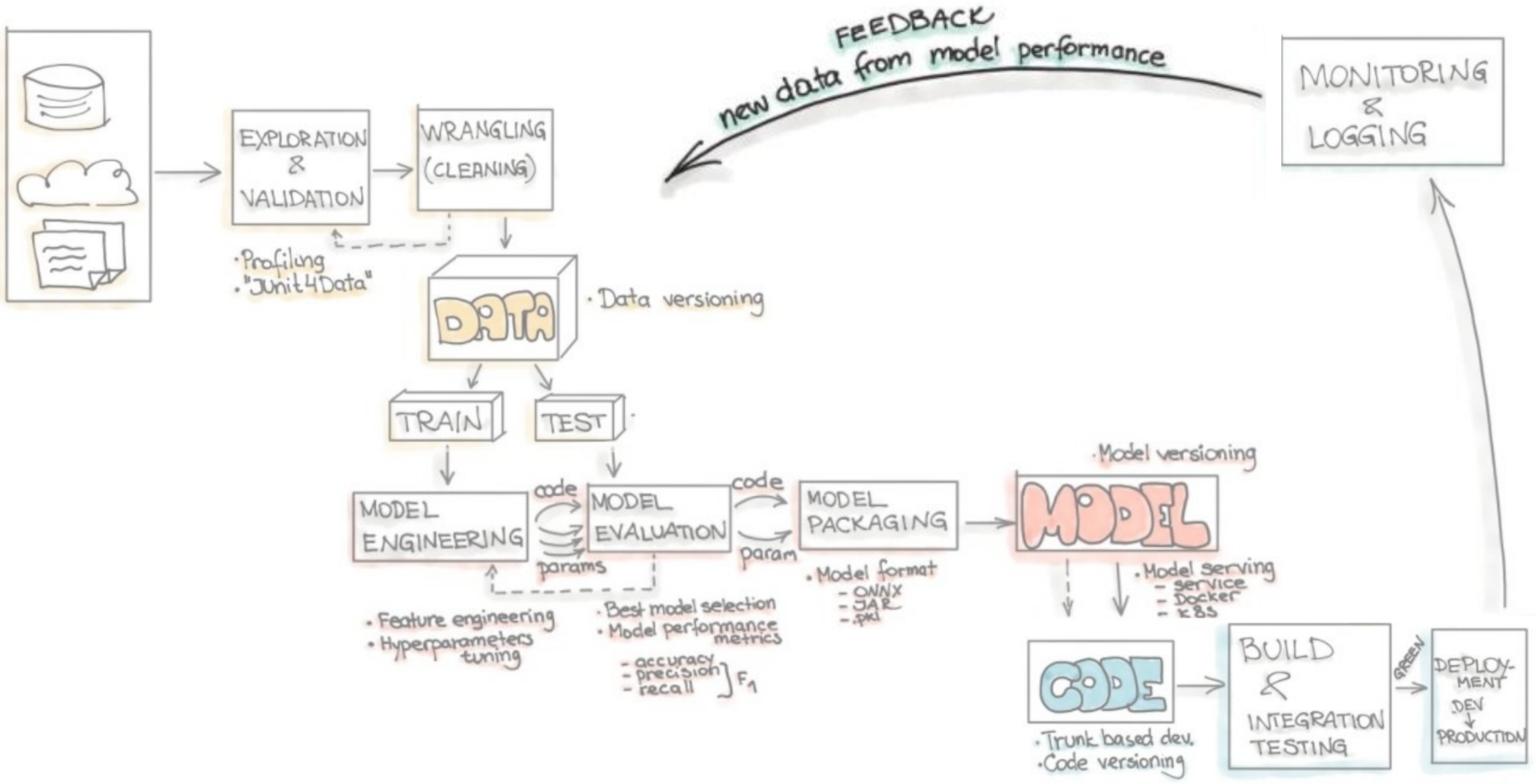


Logging



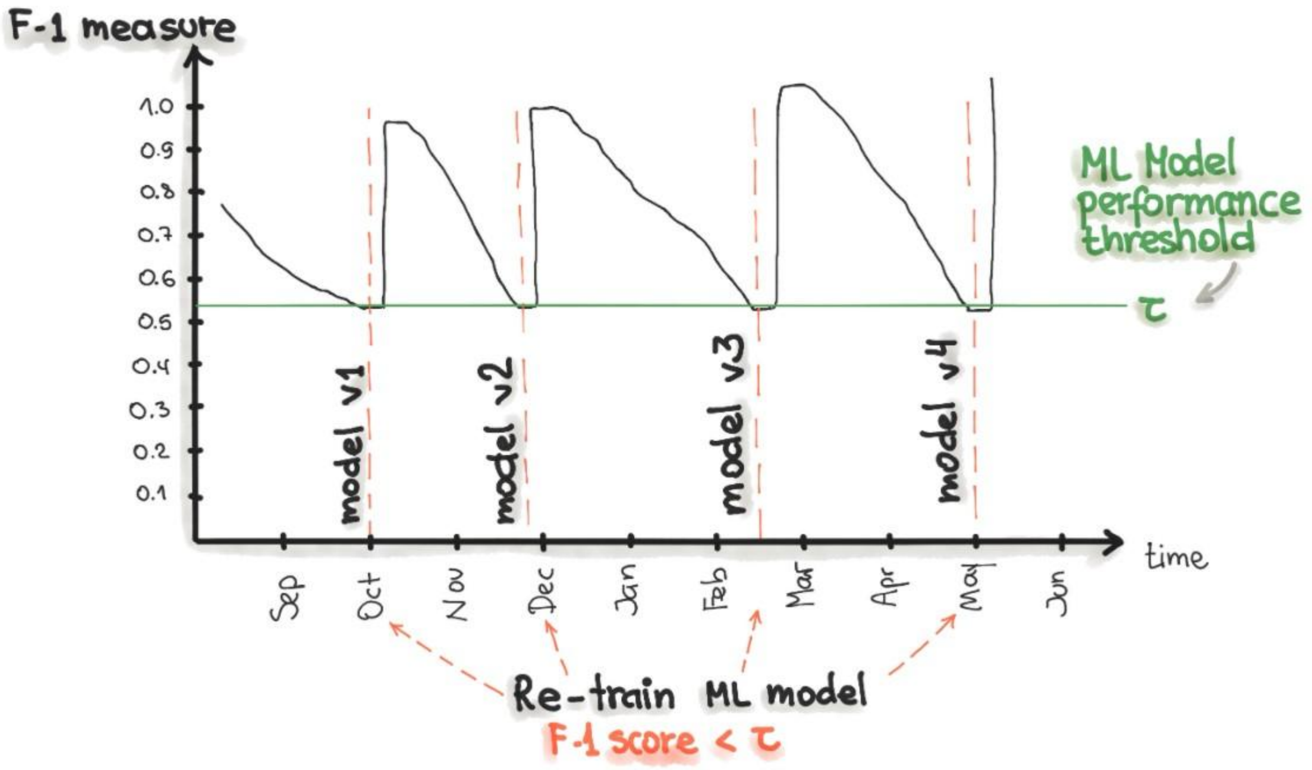
https://ml-ops.org

MLOps : Feedback pipeline



https://ml-ops.org

MLOps : Feedback pipeline



- Feedback data :
  - Model performance (precision, recall, etc)
  - Service performance (timeouts, latency, etc)
- Relevant for :
  - model training and evaluation
  - triggering alerts when model performance decays
- Generate performance analysis report

MLOps : ML service lifecycle

Internal / external data sources → **Data pipeline** → *Data QA* → Datasets → **Modelling & training pipeline** → *Func QA* → Model → **Deployment & monitoring pipeline** → *Non-func QA* → Model

Service online

Feedback pipeline

Live data

Performance analysis

Data QA *Func QA* *Non-func QA*

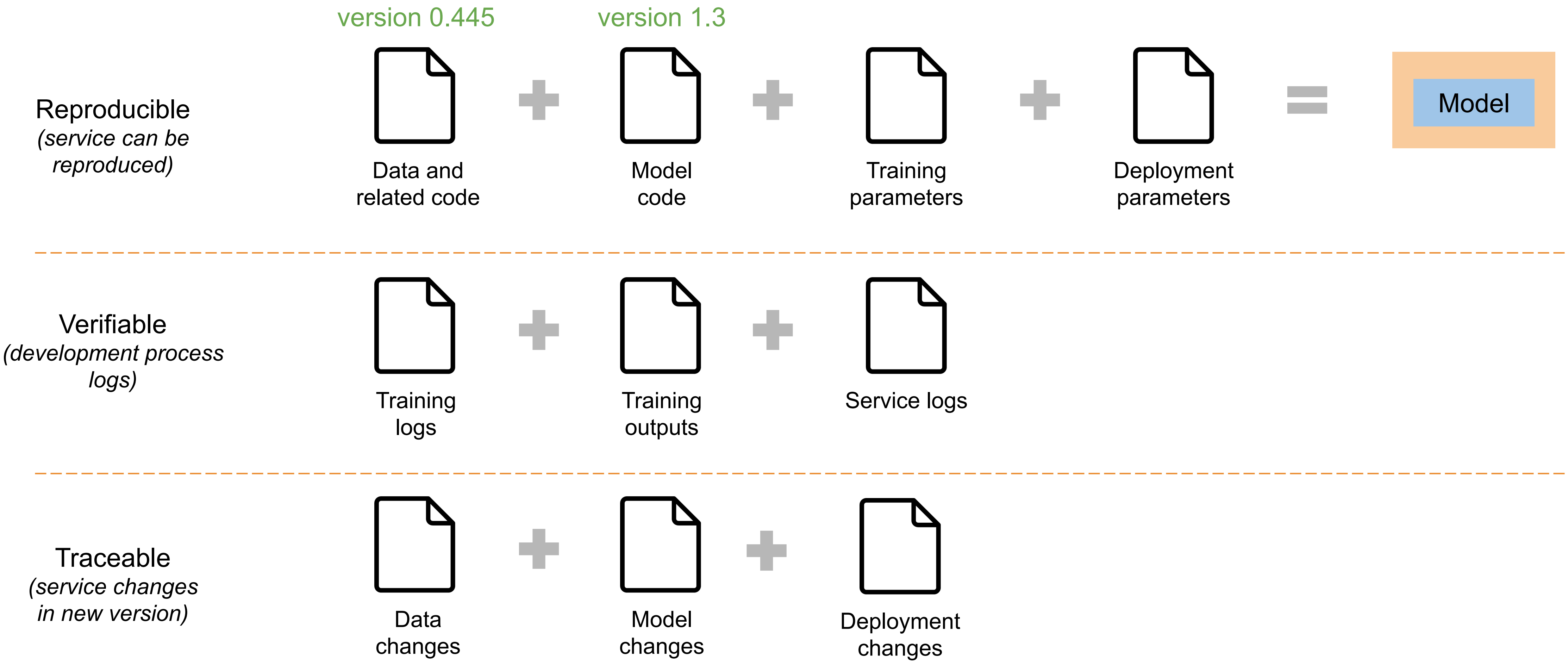
We have deployed a robust service. Is that all?

Governance

Risk management
Control access
Implement policies
Governance
Reproducible
Verifiable
Track activitiy
Traceable

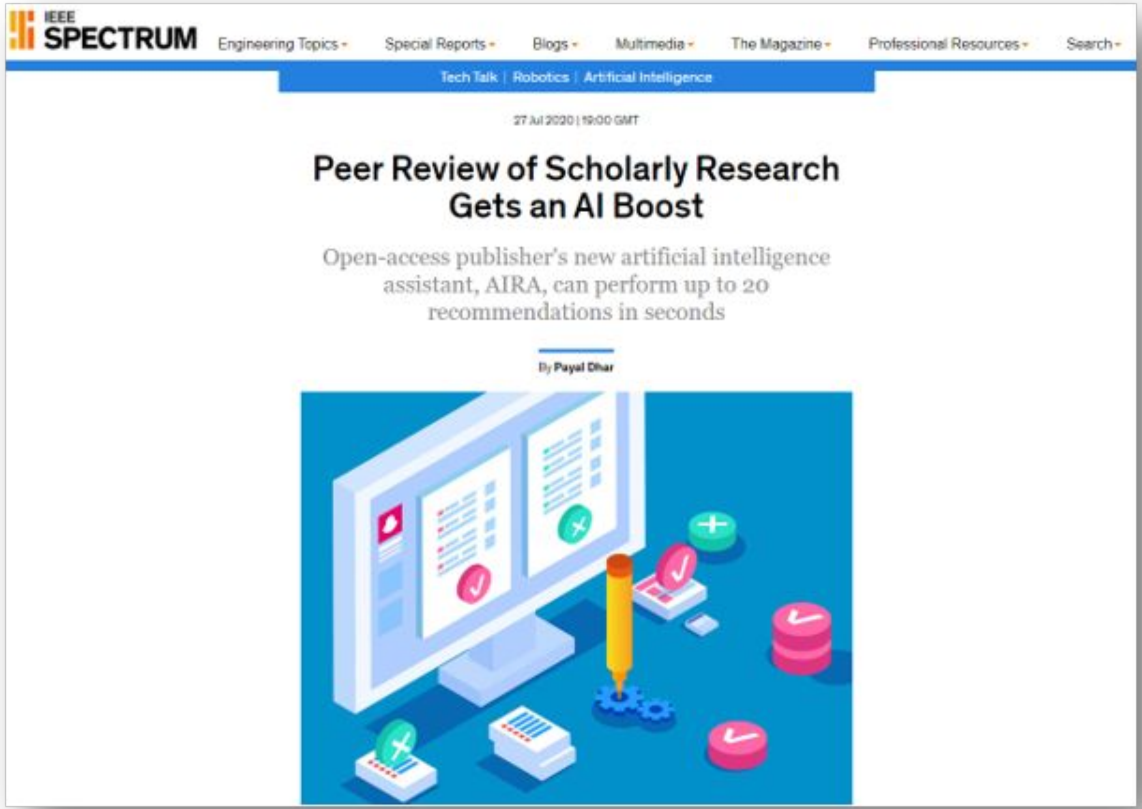
MLOps : Governance

The objective is to ensure that evidence is provided that demonstrates that a model and its delivery process can generate results that are :



ML in Frontiers

A I R A



IEEE SPECTRUM

Engineering Topics · Special Reports · Blogs · Multimedia · The Magazine · Professional Resources · Search

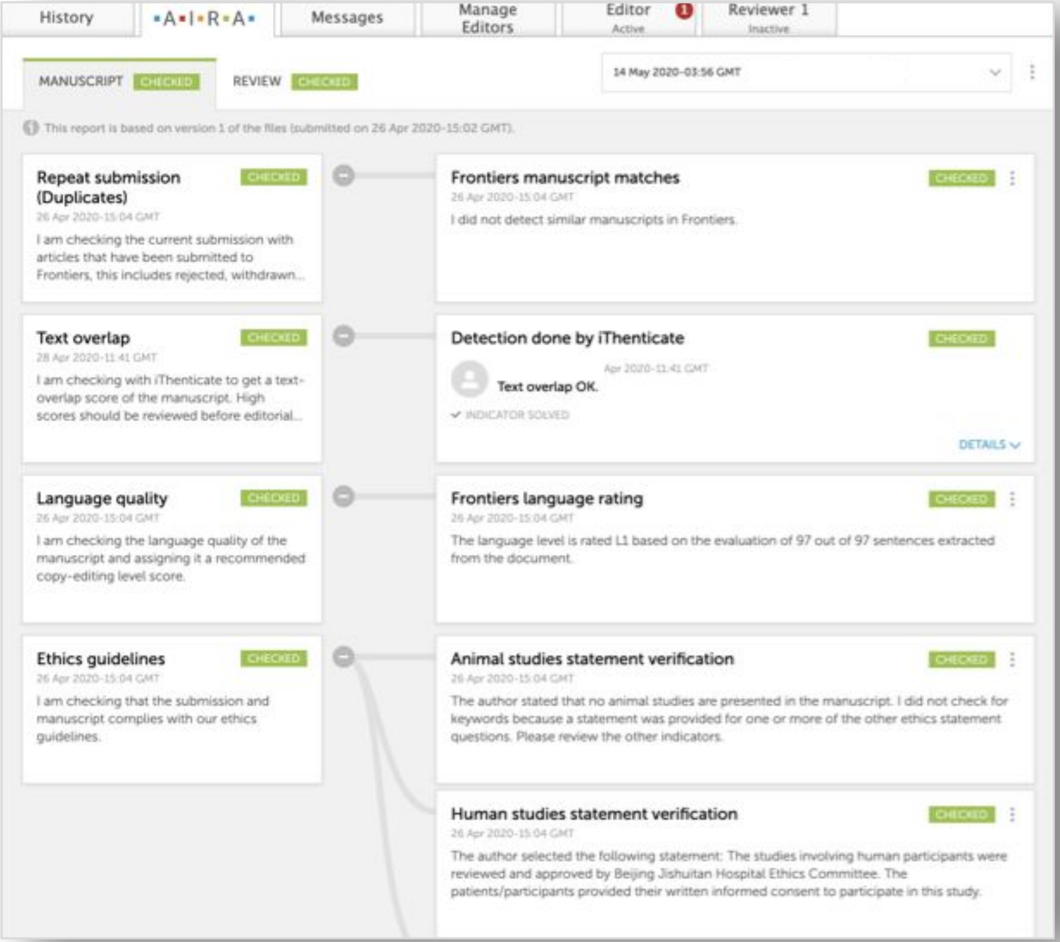
Tech Talk | Robotics | Artificial Intelligence

27 Jul 2020 | 19:00 GMT

Peer Review of Scholarly Research Gets an AI Boost

Open-access publisher's new artificial intelligence assistant, AIRA, can perform up to 20 recommendations in seconds

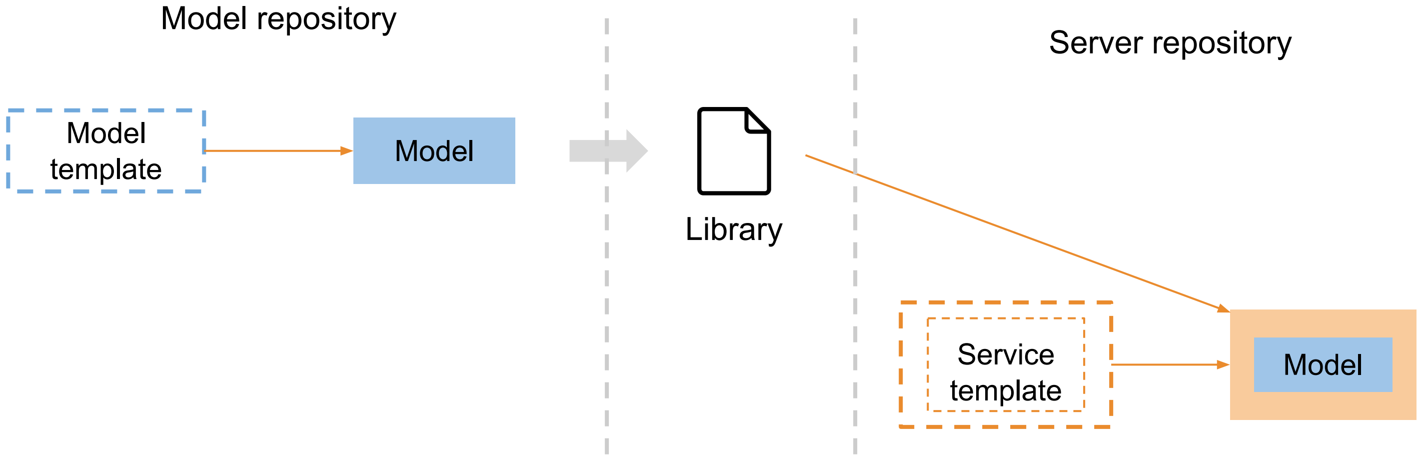
By Payal Dhar



The churros machine

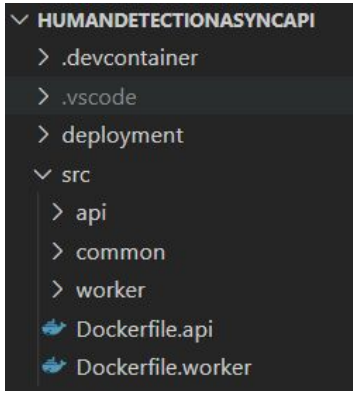
- Template that includes tons of best practices
  - Design patterns
  - Code version control
  - Documentation
  - Authentication
  - Monitoring
  - Logging
  - QA

From POC to deploy a robust service (in dev) in 1 or 2 days !



Cookie-cutter templates

```json
{
    "project_name": "Natural Name Of The Service",
    "repo_name":  "{{ cookiecutter.project_name.replace(' ', '') + 'SyncAPI' }}",
    "package_name": "{{ cookiecutter.project_name.lower().replace(' ', '').replace('-', '')}}",
    "package_version": "1.0.0",
    "model_name": "{{ cookiecutter.project_name.replace(' ', '').replace('-', '') }}",
    "project_description": "This is a sample Sync API for an ML Python package.",
    "author": "Frontiers DS Team",
    "api_port": "8000",
    "ui_port": "8001",
    "model_type": ["image", "text"],
    "cerberus_role": ""
}
```



Take-home messages

- Define a set of realistic indicators aligned to business objectives
  - Make decisions based on the indicator values
- Build a service lifecycle
  - Using best practices (logging, monitoring, etc)
  - Define a strong QA
  - Focus on having high-quality data: https://www.youtube.com/watch?v=06-AZXmwHjo&ab_channel=DeepLearningAI
- Trace all developments and changes using a governance framework

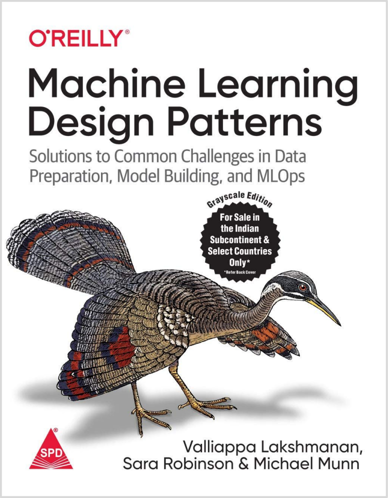
https://towardsdatascience.com/ai-and-machine-learning-governance-692b245bb6d7

References

Best practices :
https://www.idealista.com/labs/blog/data/machine-learning-en-produccion-lecciones-aprendidas/

Chip Huyen course :
https://huyenchip.com/machine-learning-systems-design/toc.html

Overview :
https://ml-ops.org



Design patterns :
https://www.oreilly.com/library/view/machine-learning-design/9781098115777/

Transforming ML models into robust services in production using MLOps

https://www.linkedin.com/in/carlosmaestreterol/