

# 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

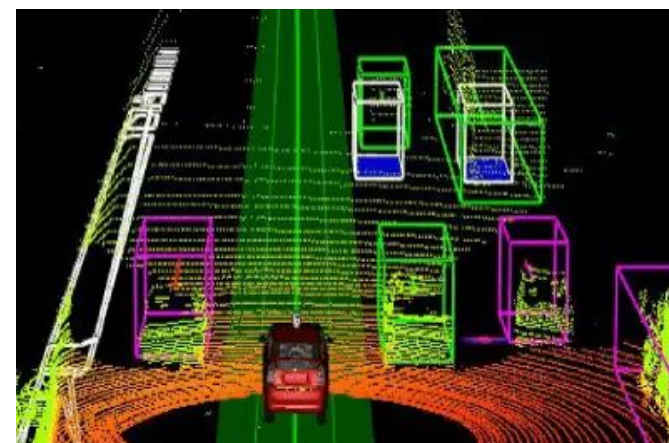
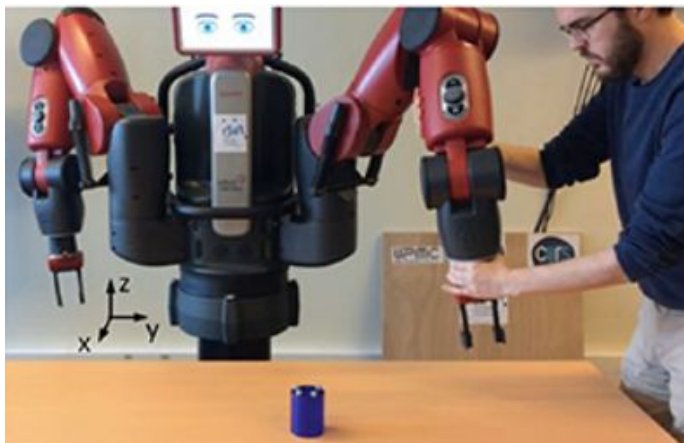


**frontiers**  
in Neurorobotics

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

 Carlos Maestre<sup>1\*</sup>,  Ghanim Mukhtar<sup>1</sup>,  Christophe Gonzales<sup>2</sup> and  Stephane Doncieux<sup>1</sup>

<sup>1</sup>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

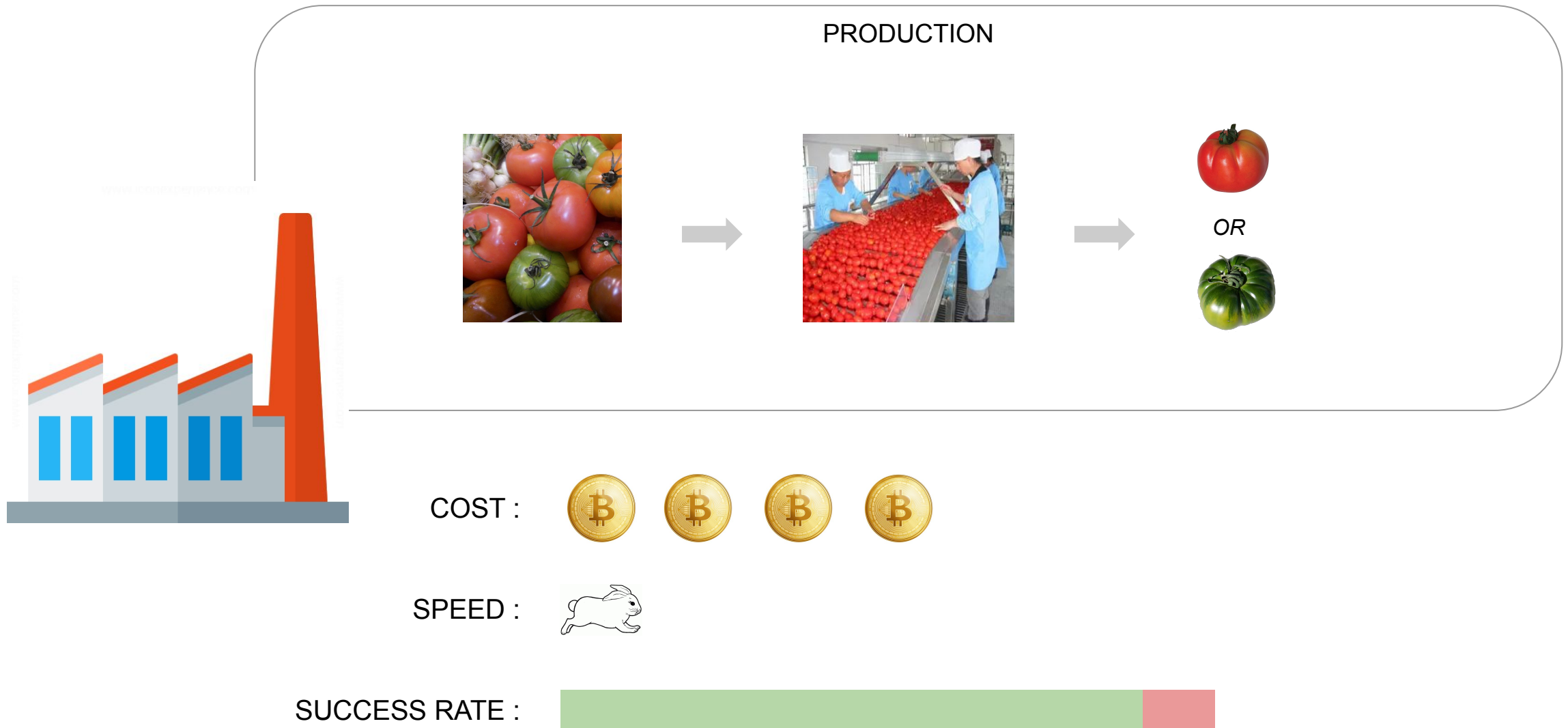
**If Data Scientists are so smart, why do 70% of their projects fail?**

Published on June 20, 2019

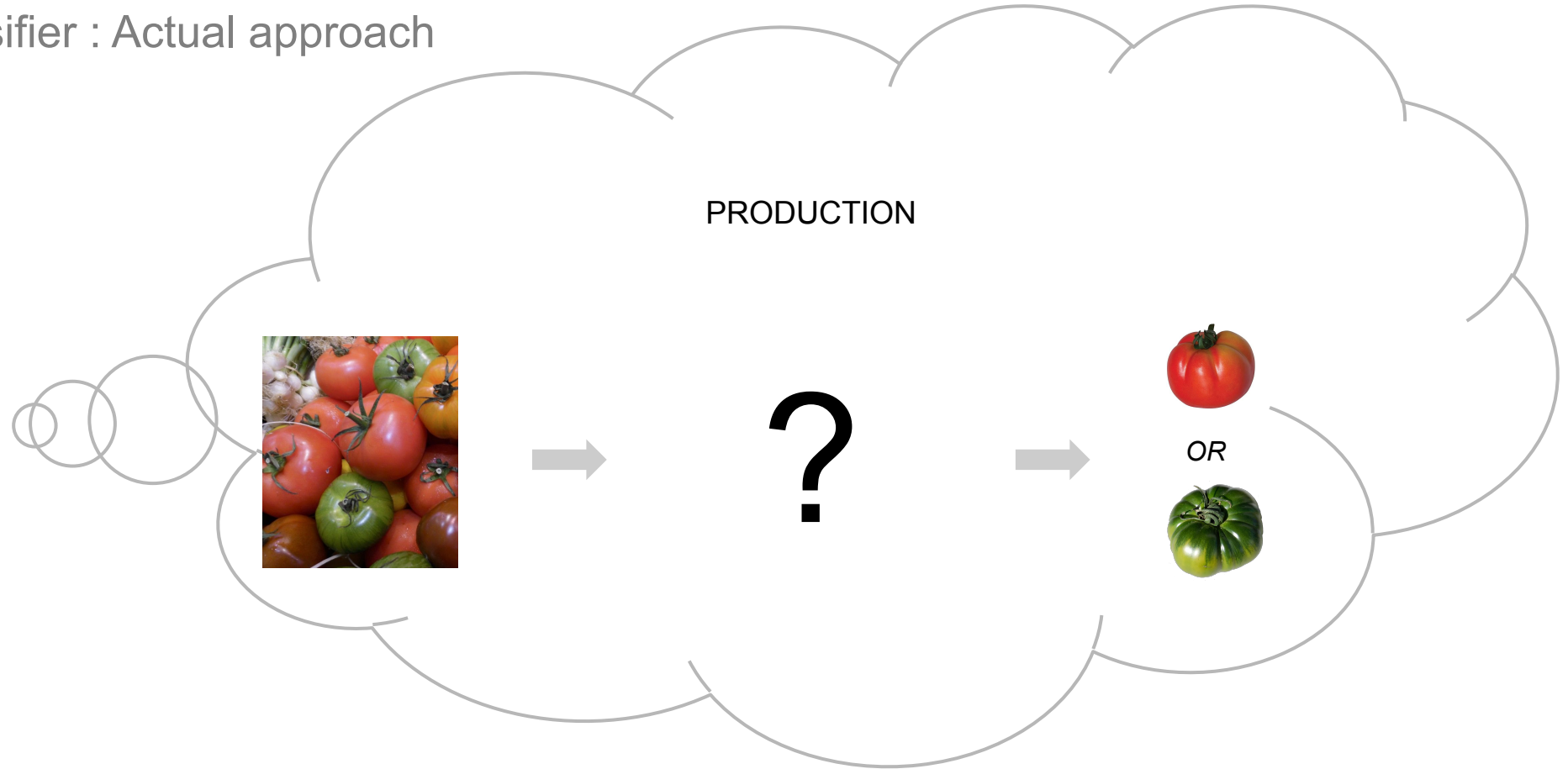
<https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/>  
<https://www.datadriveninvestor.com/2020/04/30/the-majority-of-business-analytics-and-ai-projects-are-still-failing/>  
<https://www.linkedin.com/pulse/data-scientist-so-smart-why-do-70-projects-fail-mike-fish/>

Why?

## Tomatoes classifier : Actual approach



## Tomatoes classifier : Actual approach



COST : Lower

SPEED : Faster

SUCCESS RATE : Similar or better

## Tomatoes classifier : Fast classifier machine

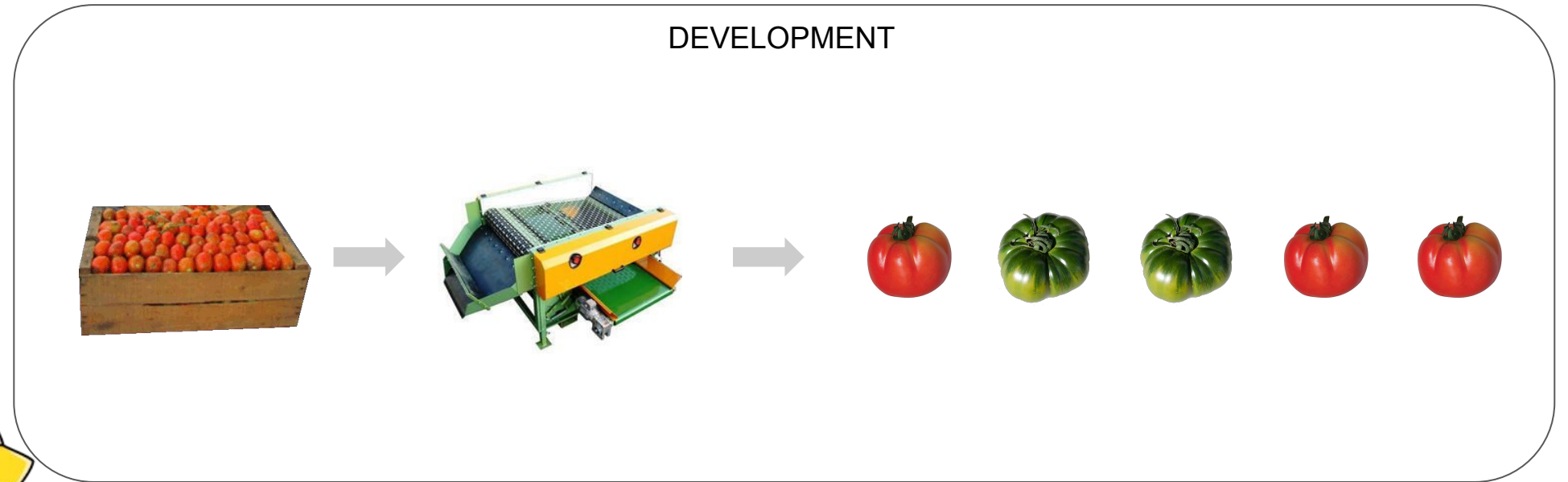





## Tomatoes classifier : Fast classifier machine



# Tomatoes classifier : Machine : Working in a controlled environment

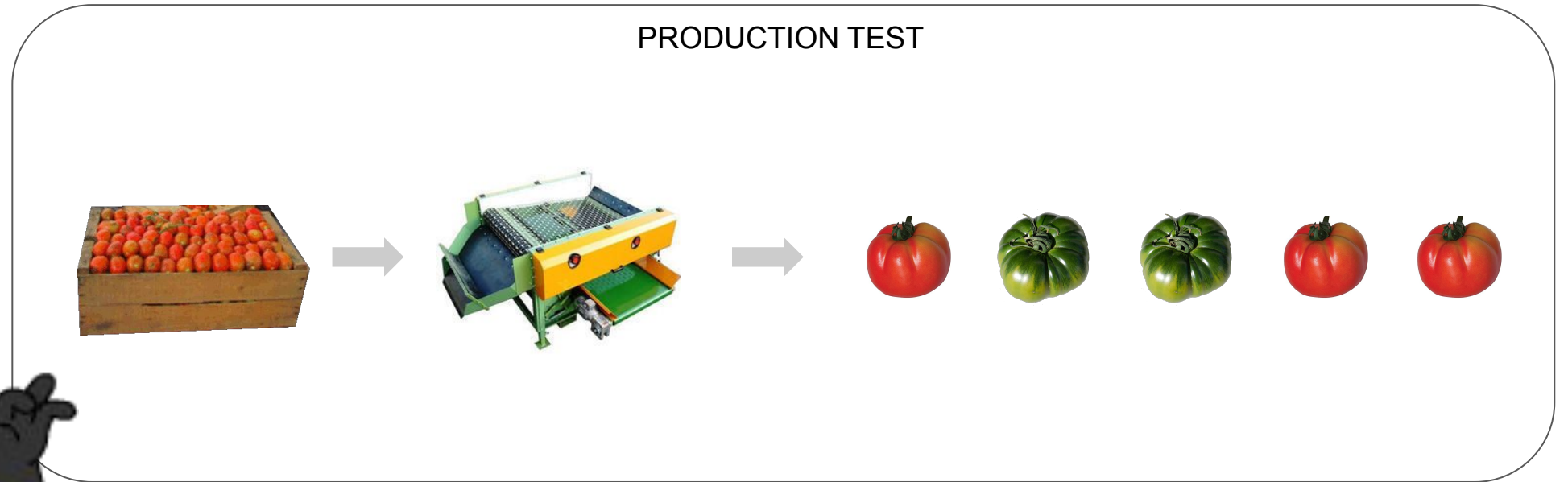


(HALF) COST : 


(DOUBLE) SPEED : 

(SIMILAR) RATE : 

# Tomatoes classifier : Machine : Working under similar conditions



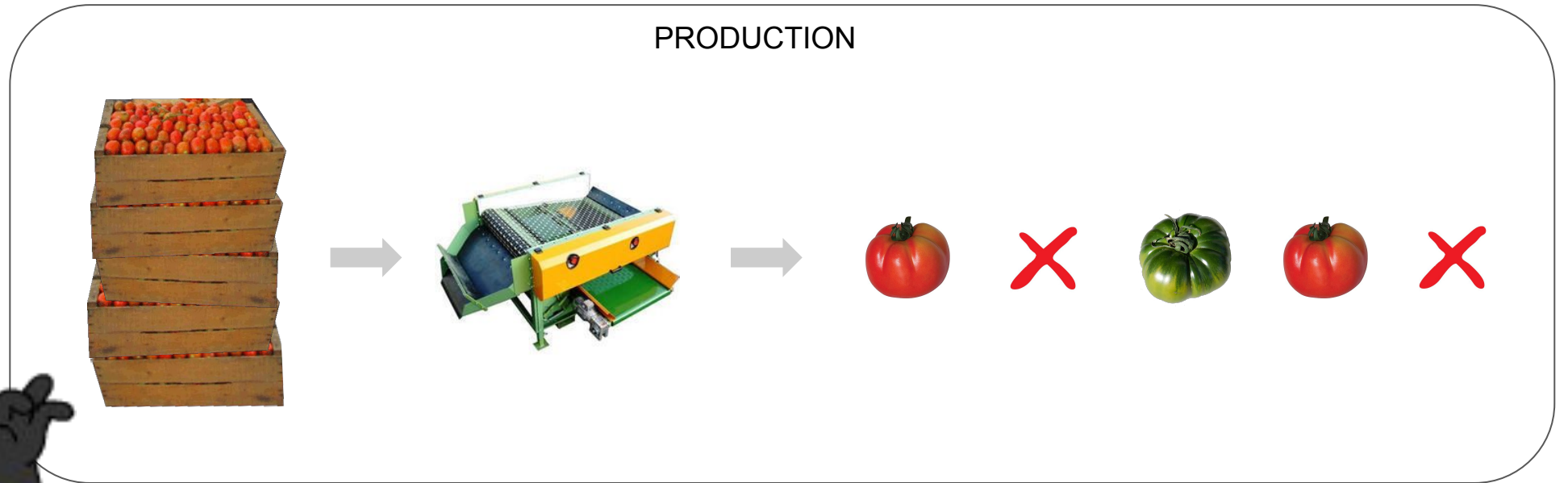
COST : 

SPEED : 


SUCCESS RATE : 

UNEXPECTED  
PRODUCTION COST

# Tomatoes classifier : Machine : Stops working with high workload



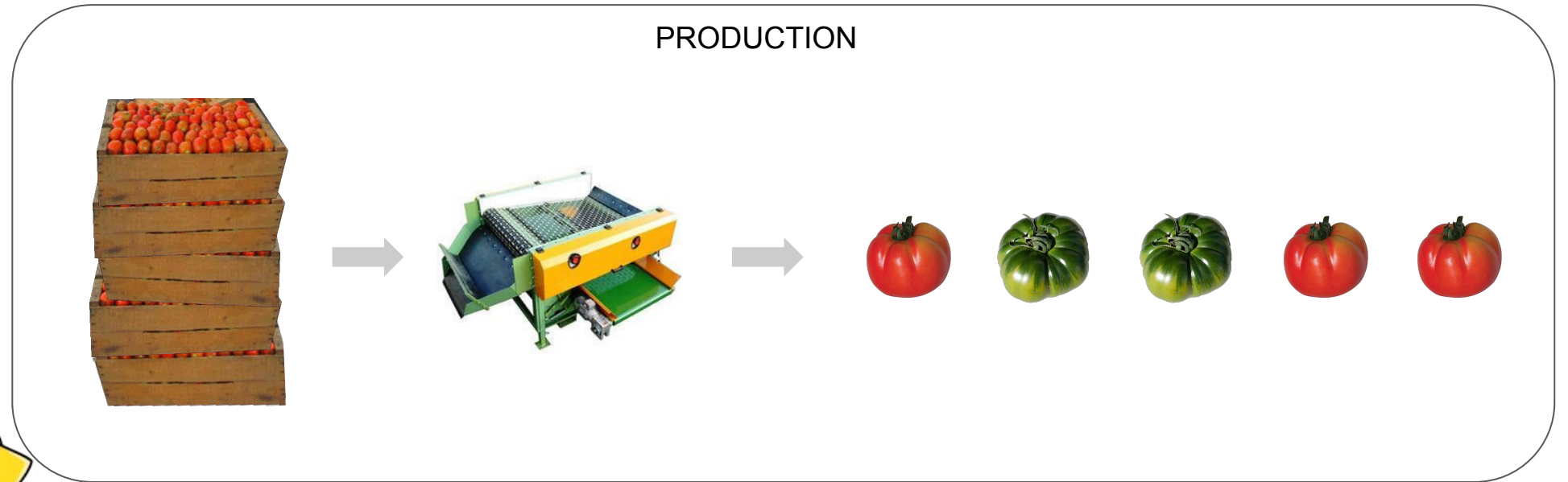
COST : 

SPEED : 





SUCCESS RATE : 

UNEXPECTED  
WORKLOAD FAILURE

# Tomatoes classifier : Machine : Stops working with high workload



COST :   

SPEED :    

SUCCESS RATE : 

# Tomatoes classifier : Machine : Misclassified unseen types of tomato



## PRODUCTION



COST :



SPEED :



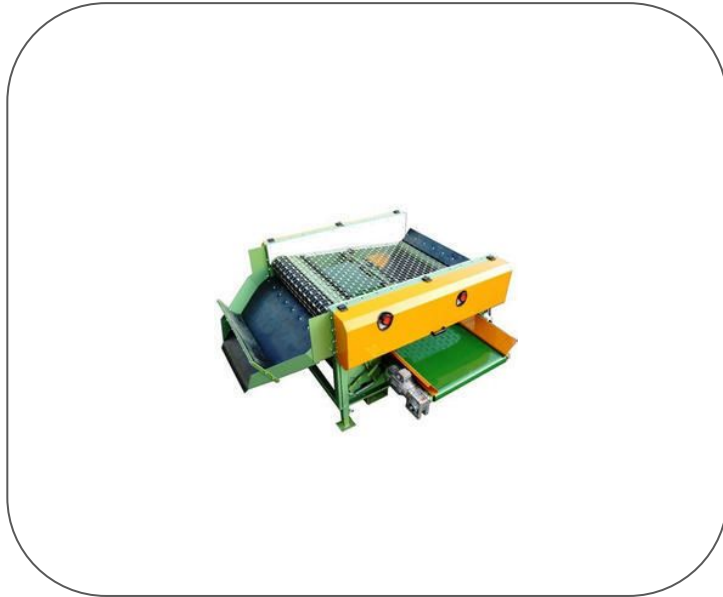
SUCCESS RATE :



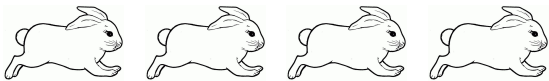
UNEXPECTED  
INPUT TYPE



# Tomatoes classifier : Machine : Where to invest more money ?



VS



UNREALISTIC  
EXPECTATIONS



WELL KNOWN  
APPROACH



## Tomatoes classifier : Final approach



### PRODUCTION



COST :



SPEED :



SUCCESS RATE :



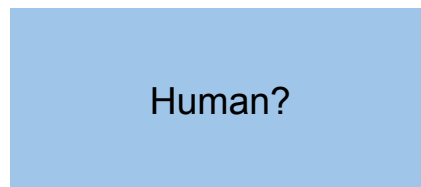
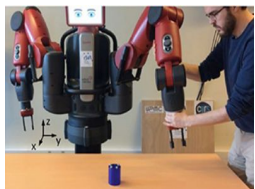


## Tomatoes classifier : Machine : Failure analysis

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

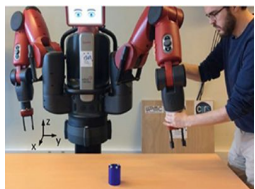
## Tomatoes classifier : Machine : Failure analysis

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



## 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 <del>tomatoes</del> images	Workload test	Quality assurance (QA)
UNEXPECTED INPUT TYPE	Same type of <del>tomatoes</del> image	Add tomato diversity during development	Robust services
UNREALISTIC EXPECTATIONS	ML hype	Define realistic metrics	Business-oriented metrics
MLOps			



Human?



# MLOps

---

From Wikipedia, the free encyclopedia

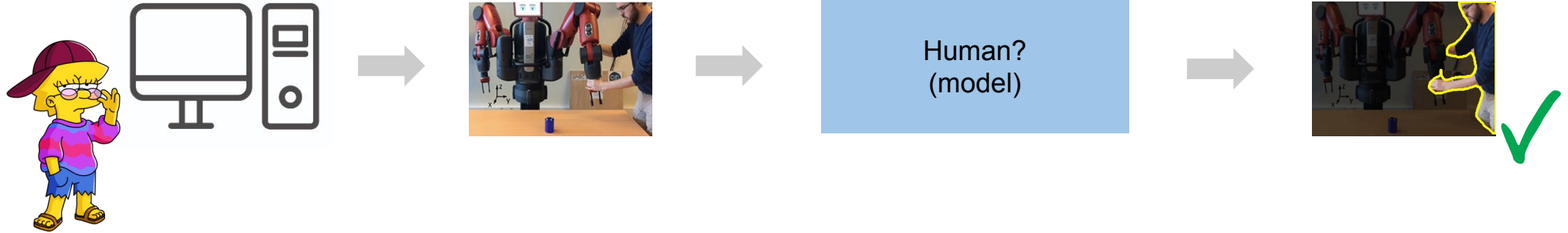
**MLOps** (a compound of “machine learning” and “operations”) is a practice for collaboration and communication between data scientists and operations professionals to help manage production ML (or deep learning) lifecycle.

*to build robust ML services and*

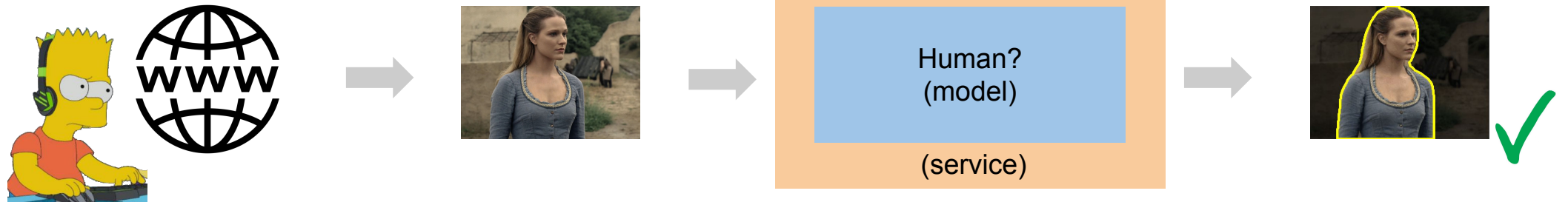
- New field: few standards
- New mindset
- New frameworks
- Fast changes

# ML service

*Model (development environment)*

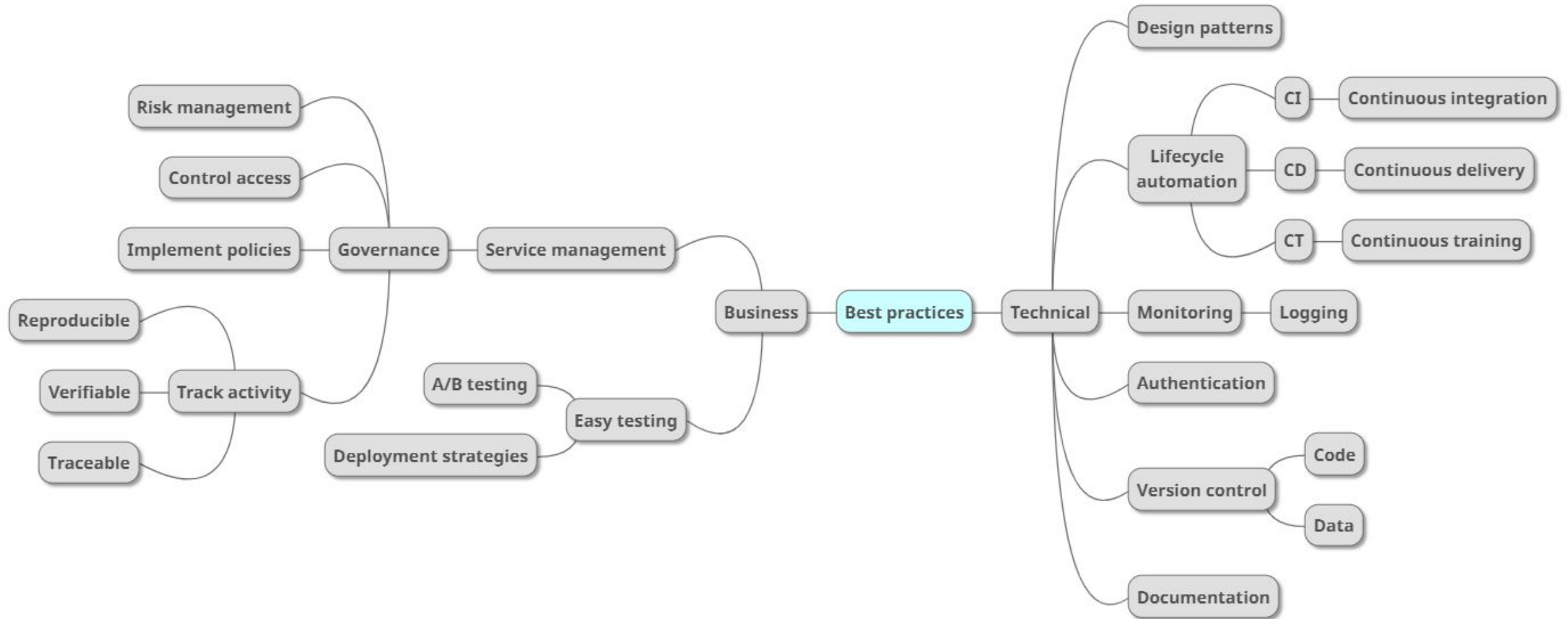


*Service (production environment)*

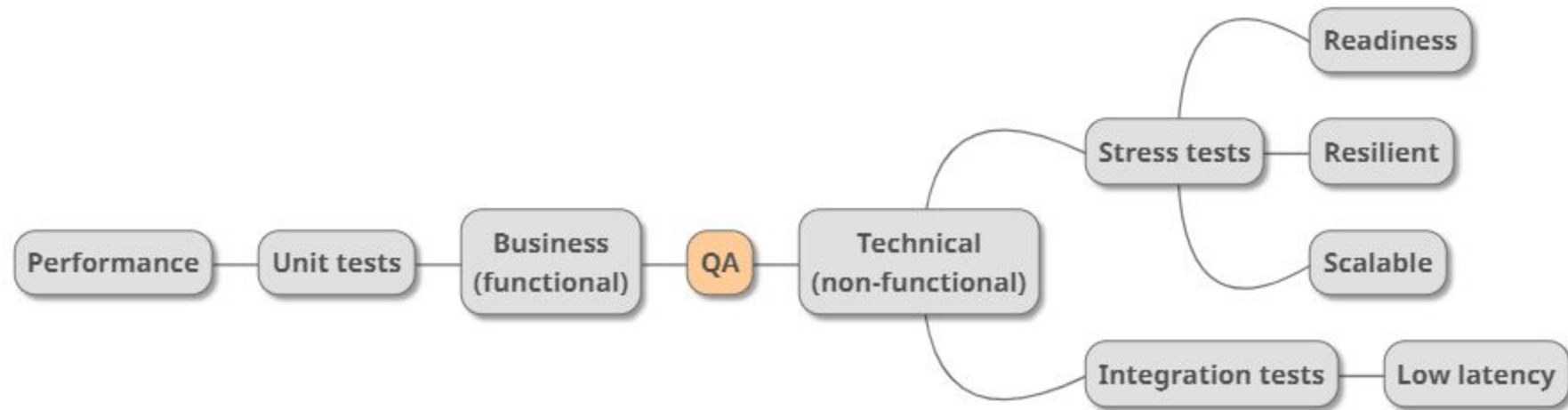


What are the features of  
a robust ML service?

# MLOps : Best practices



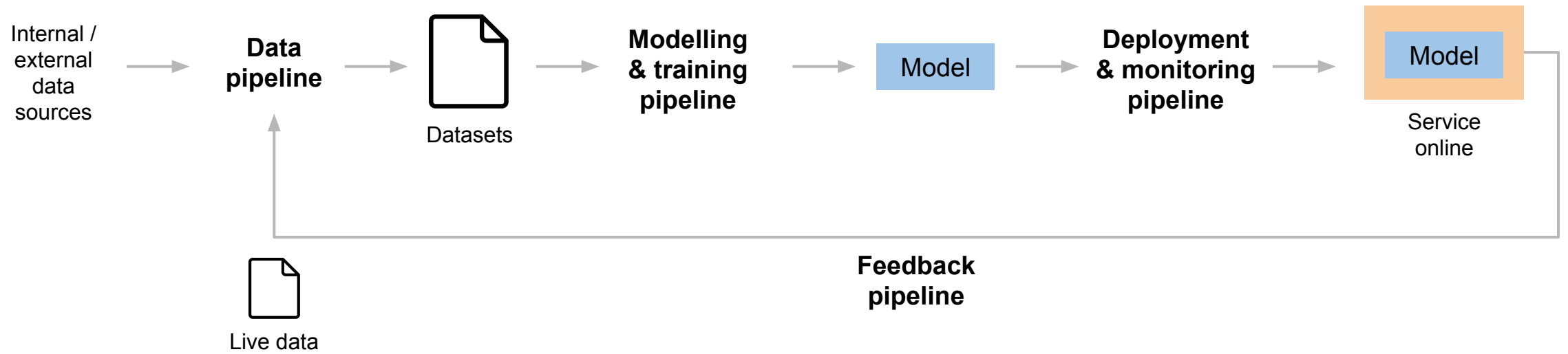
## MLOps : QA based on indicators



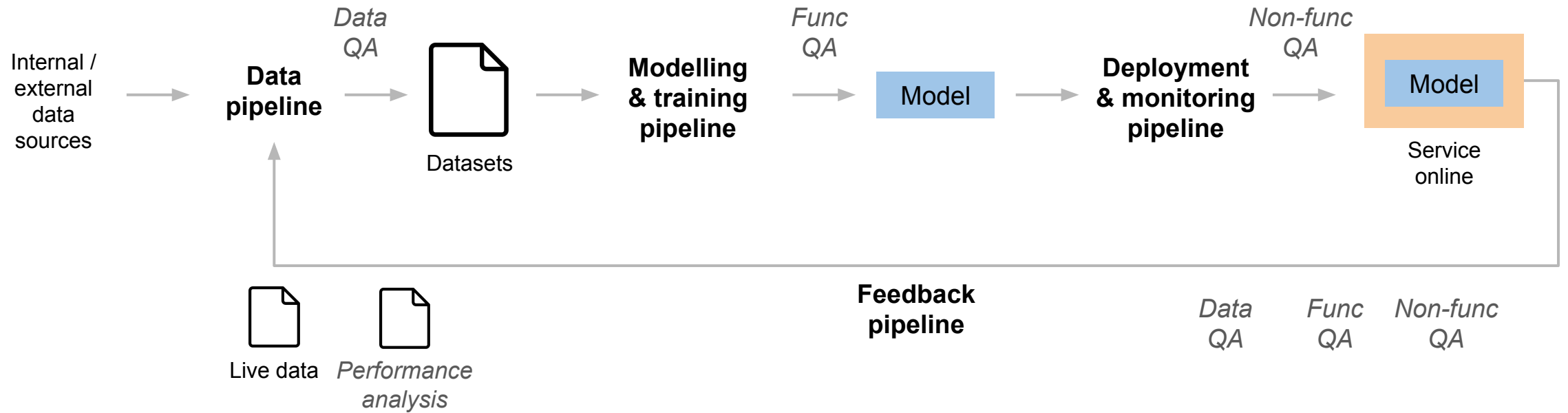


How to build these features?

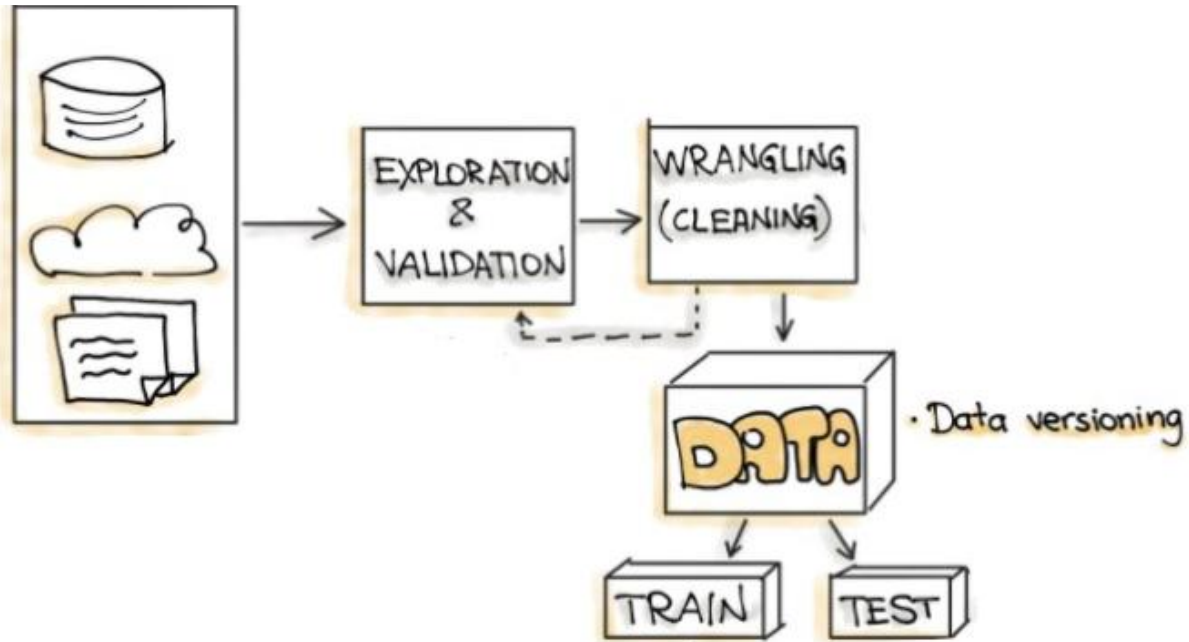
## MLOps : ML service lifecycle



# MLOps : ML service lifecycle

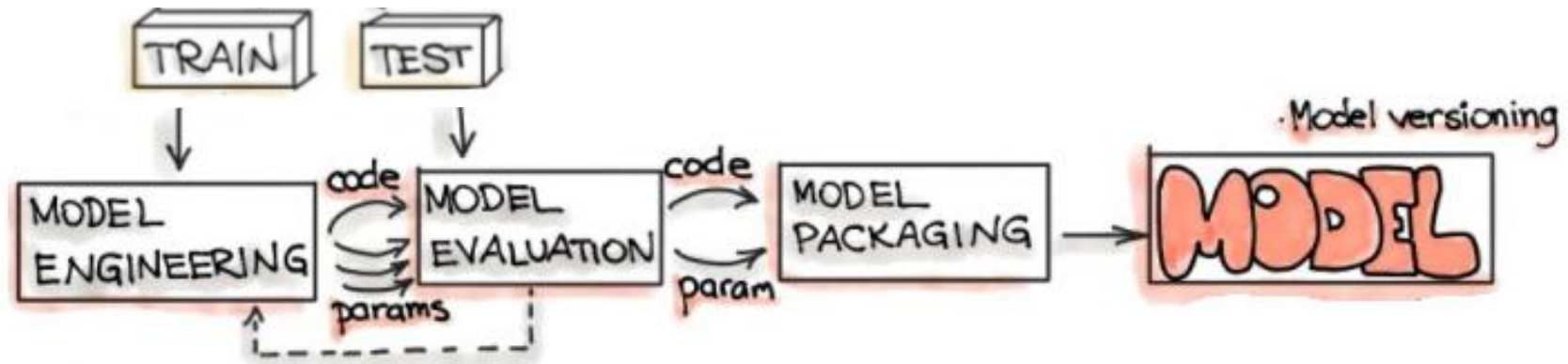


## MLOps : Data pipeline



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

## MLOps : Modelling / Training pipeline



Creating a model comprehends 3 steps:

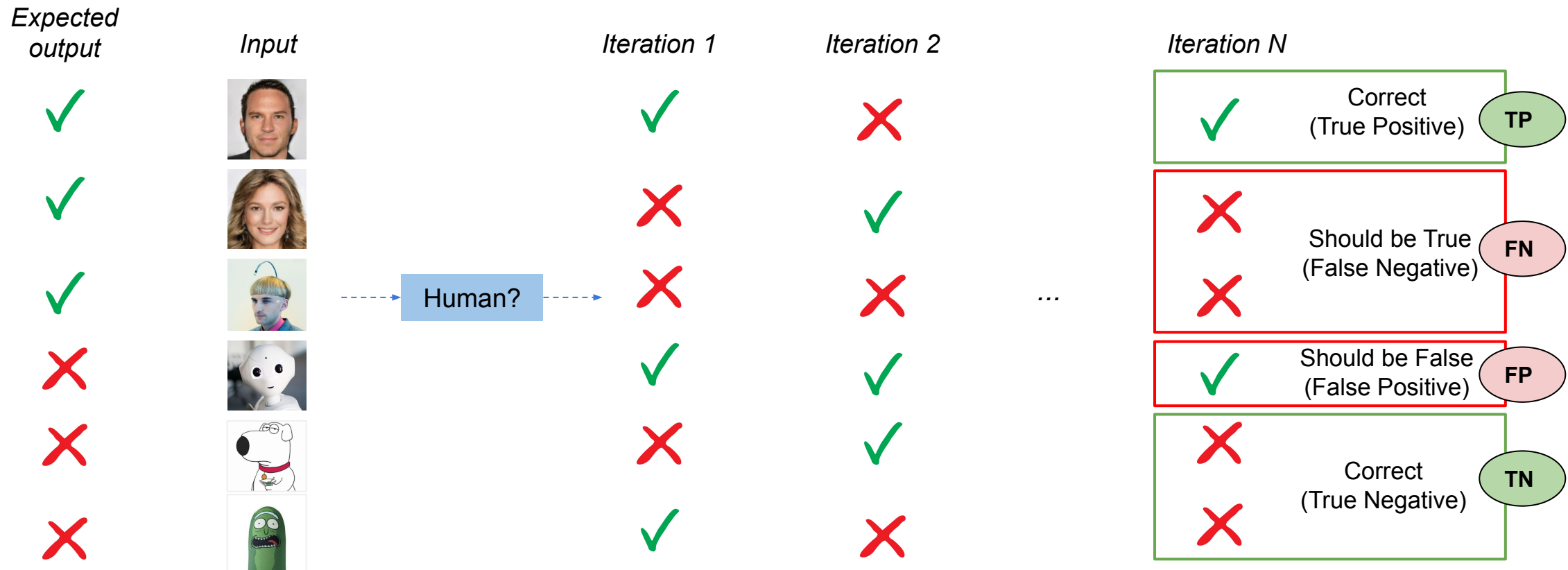
1. Design
2. Training
3. Evaluation

# 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)



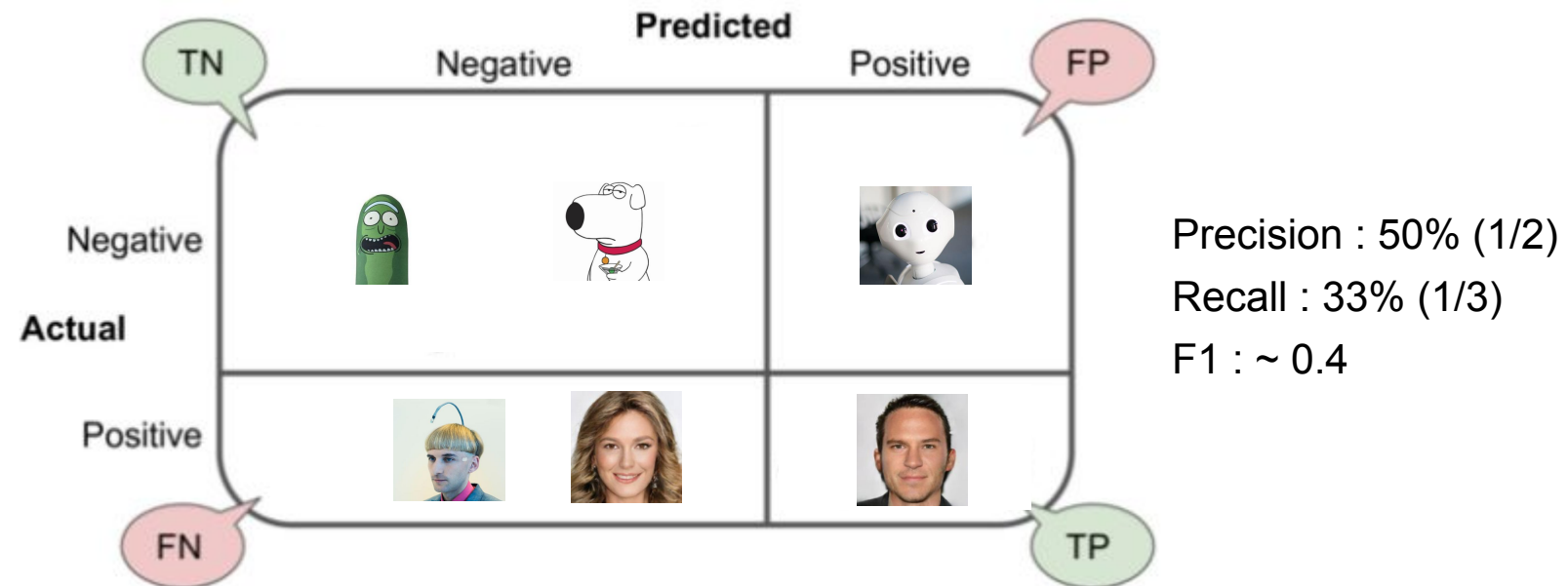
# MLOps : Modelling / Training pipeline

## Model evaluation (functional indicators)

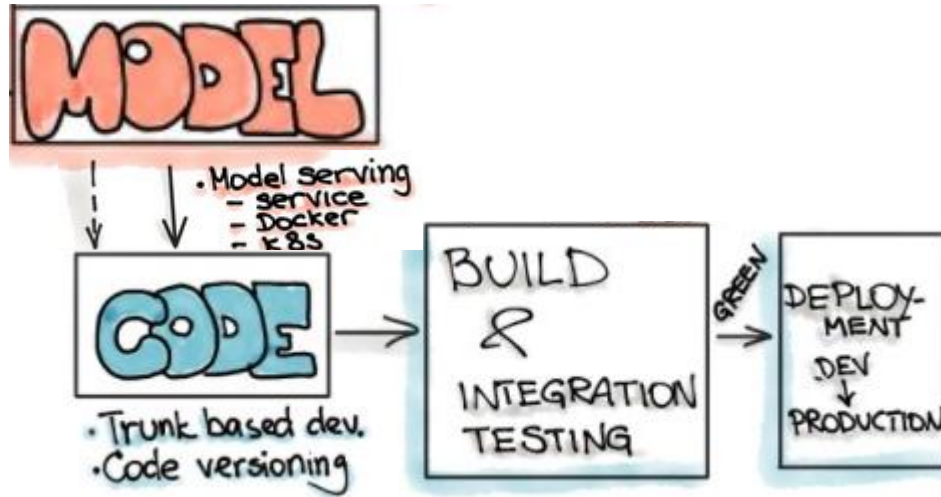
Precision : fraction of cases predicted positive (human) which are actually positive (human)

Recall : fraction of positives (all humans) that have been correctly predicted positive (human)

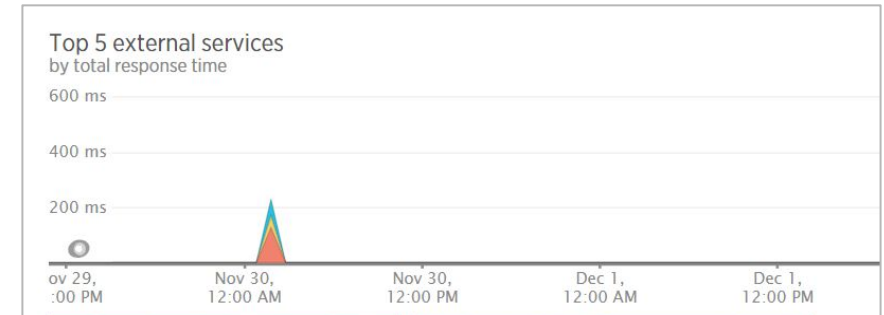
F1-score : value in range [0, 1] computed using precision and recall



# MLOps : Deployment / Monitoring pipeline



## Monitoring

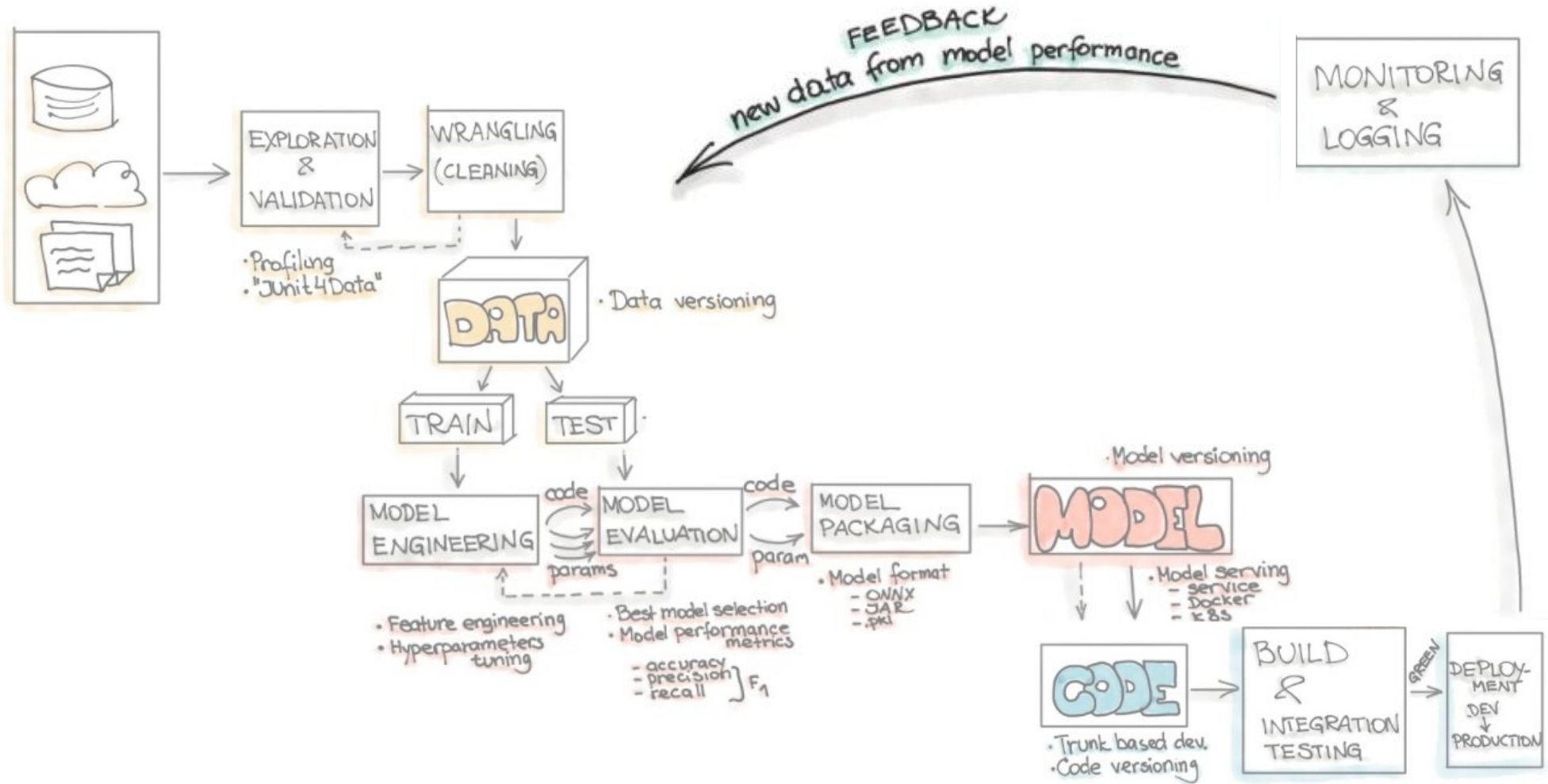


## Logging

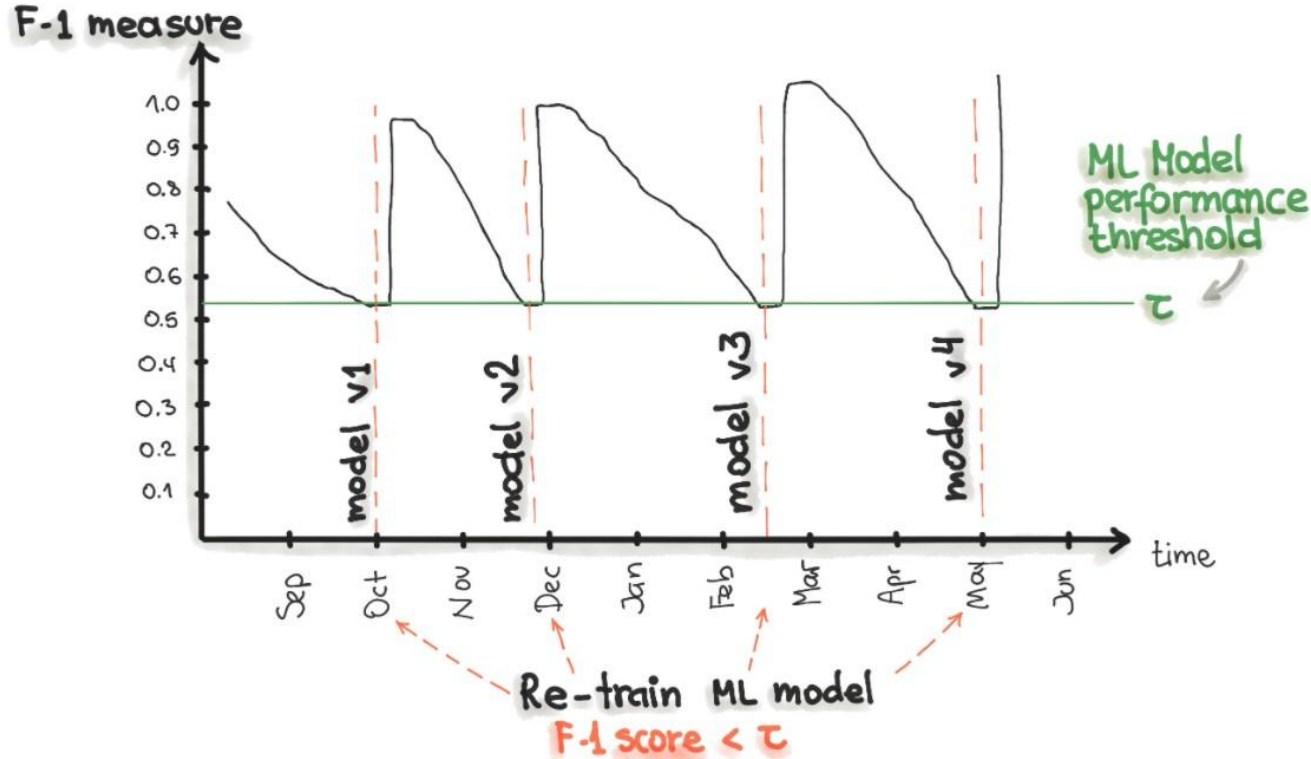
```
192.0.2.33 - - [19/Mar/2014:14:37:17 +0000] "GET /features.html HTTP/1.1" 200 263049 "-" Apache-HttpClient/4.2.3
192.0.2.55 - - [19/Mar/2014:14:37:17 +0000] "PUT /features.html HTTP/1.1" 200 422671 "-" "-"
192.0.2.33 - - [19/Mar/2014:14:37:18 +0000] "GET /index.html HTTP/1.1" 200 318902 "-" "-"
Mar 19 14:37:18 frontend3 server[121]: Received packet from 192.0.2.55
Mar 19 14:37:19 frontend3 server[123]: Received packet from 192.0.2.55
Mar 19 14:37:19 frontend3 server[123]: Handling request 9efcf643-ac89-4125-a69d-ec3203047a19
192.0.2.33 - - [19/Mar/2014:14:37:19 +0000] "PUT /index.html HTTP/1.0" 200 871988 "-" "-"
192.0.2.55 - - [19/Mar/2014:14:37:20 +0000] "GET /index.html HTTP/1.0" 200 400613 "-" "-"
192.0.2.55 - - [19/Mar/2014:14:37:21 +0000] "GET /obj/12357foo-bar HTTP/1.0" 200 841360 "-" "Apache-HttpClient/4.
Mar 19 14:37:21 frontend3 worker[61456]: Handling request 9efcf643-ac89-4125-a69d-ec3203047a19
Mar 19 14:37:22 frontend3 worker[61456]: Successfully started helper
192.0.2.33 - - [19/Mar/2014:14:37:22 +0000] "PUT /index.html HTTP/1.0" 200 944322 "-" "Apache-HttpClient/4.2.3
Mar 19 14:37:23 frontend3 worker[61456]: Received packet from 192.0.2.55
Mar 19 14:37:23 frontend3 worker[123]: Handling request 9efcf643-ac89-4125-a69d-ec3203047a19
Mar 19 14:37:24 frontend3 server[124]: Received packet from 192.0.2.55
Mar 19 14:37:24 frontend3 worker[61456]: Handling request 70430eff-159e-4818-a0e7-f21a7d4ad892
Mar 19 14:37:25 frontend3 server[121]: Handling request 9ad6455c-0edf-4623-b3bc-5f65ce81825f
192.0.2.55 - bob@example.com [19/Mar/2014:14:37:25 +0000] "GET /images/compass.jpg HTTP/1.0" 200 4509 "-" "-"
192.0.2.55 - - [19/Mar/2014:14:37:25 +0000] "GET /obj/12357foo-bar HTTP/1.1" 200 420858 "-" "Apache-HttpClient/4.
192.0.2.33 - - [19/Mar/2014:14:37:26 +0000] "PUT /features.html HTTP/1.1" 200 741805 "-" "-"
Mar 19 14:37:27 frontend3 worker[61456]: Successfully started helper
Mar 19 14:37:27 frontend3 server[123]: Received packet from 192.0.2.55
Mar 19 14:37:27 frontend3 server[121]: Handling request 70430eff-159e-4818-a0e7-f21a7d4ad892
192.0.2.55 - - [19/Mar/2014:14:37:28 +0000] "GET /index.html HTTP/1.0" 200 299909 "-" "Apache-HttpClient/4.2.3
192.0.2.55 - - [19/Mar/2014:14:37:28 +0000] "GET /index.html HTTP/1.0" 200 731434 "http://lnav.org/download.html"
```



# MLOps : Feedback pipeline

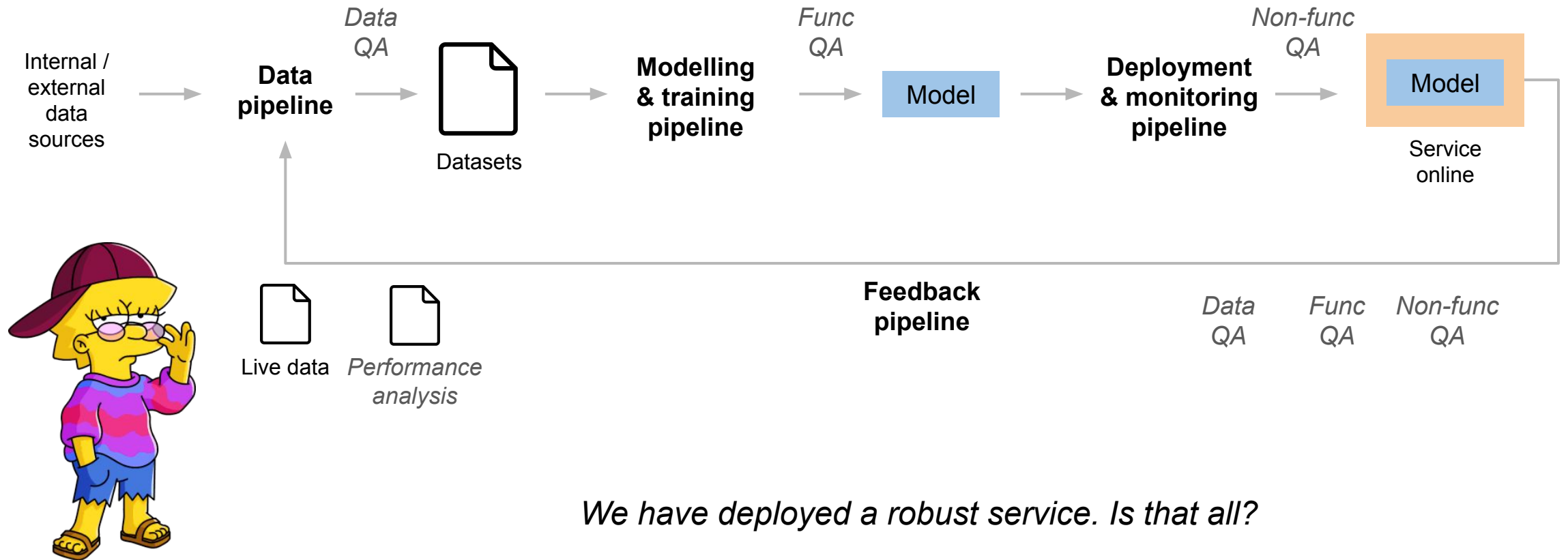


## MLOps : Feedback pipeline

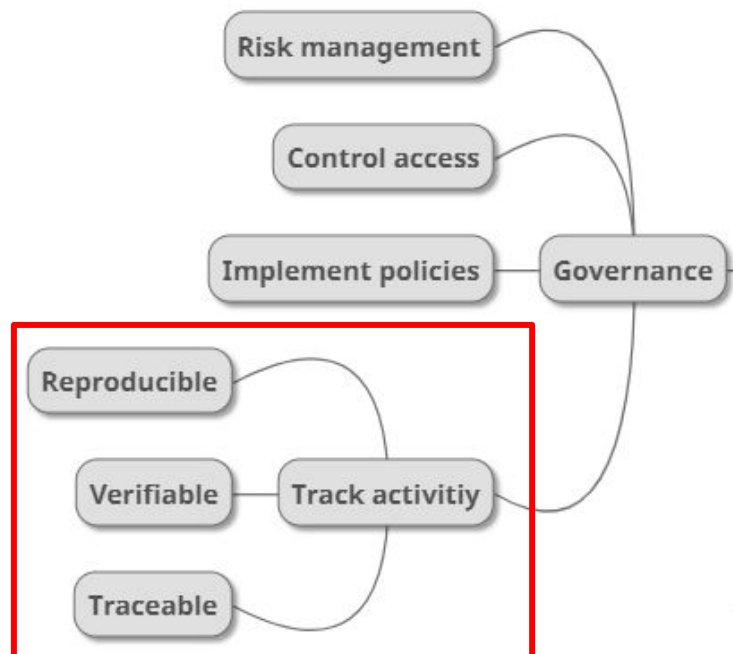


- Feedback data :
  - Model performance (precision, recall, etc)
  - Service performance (timeouts, latency, etc)
- Relevant for :
  - model training and evaluation
  - triggering alerts when model performance decays
- Generate performance analysis report

# MLOps : ML service lifecycle

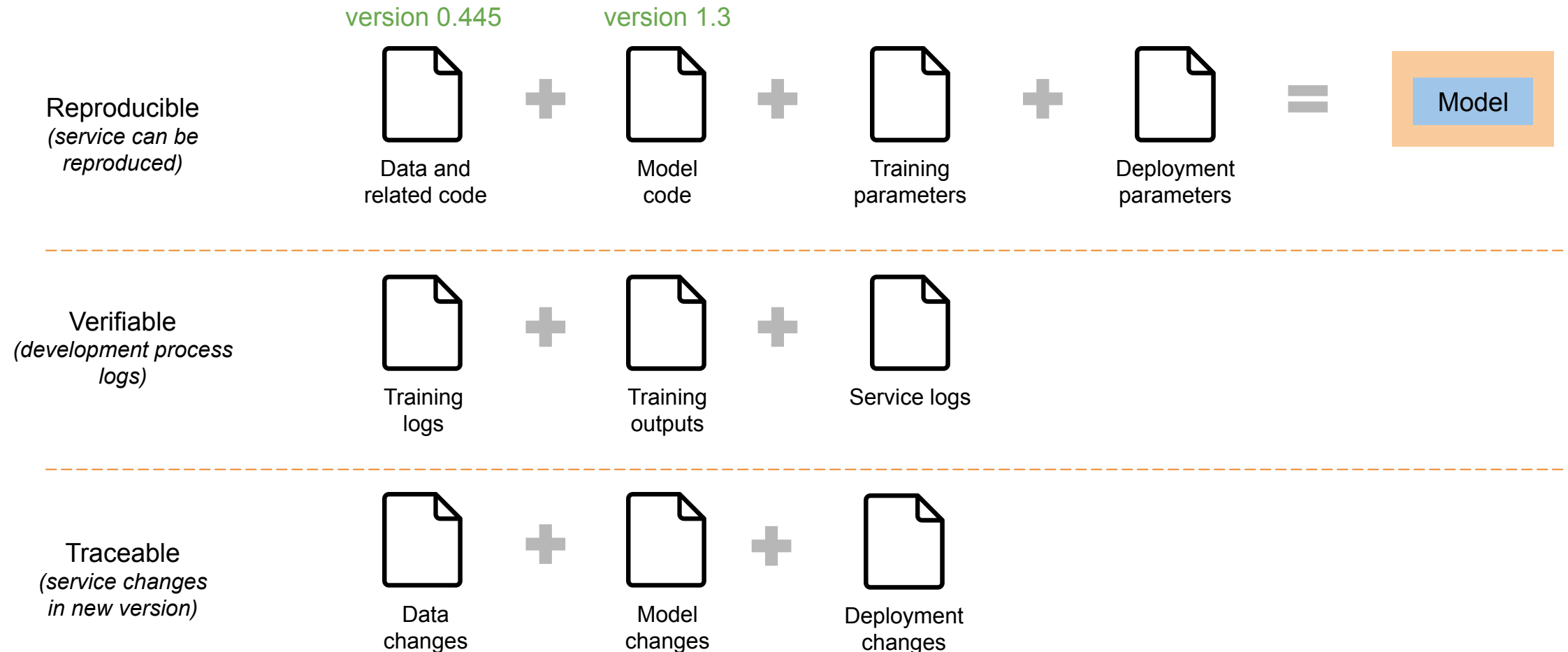


# Governance



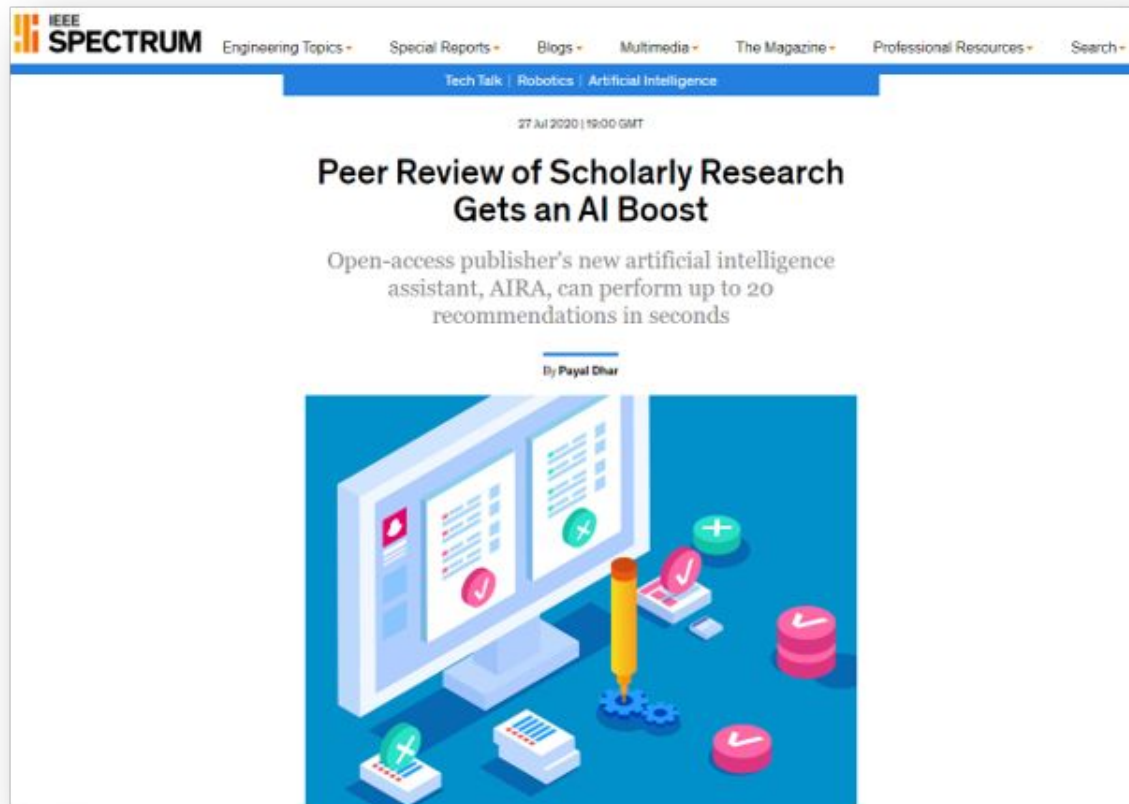
# MLOps : Governance

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

# AIRA



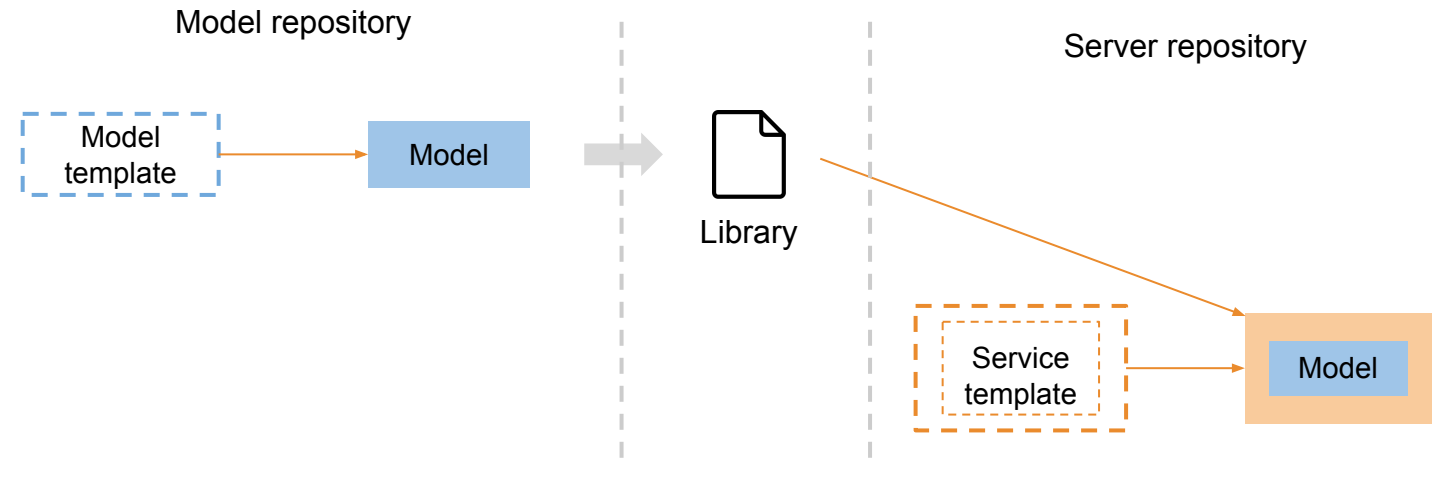
The image shows a screenshot of the AIRA interface. The top navigation bar includes History, Messages, Manage Editors, Editor (Active), and Reviewer 1 (Inactive). The main content area is titled "MANUSCRIPT" and "REVIEW". It displays a report for a manuscript submitted on 26 Apr 2020-15:02 GMT. The report includes several sections, each with a "CHECKED" status:

- Repeat submission (Duplicates)**: 26 Apr 2020-15:04 GMT. I am checking the current submission with articles that have been submitted to Frontiers, this includes rejected, withdrawn...
- Text overlap**: 26 Apr 2020-15:41 GMT. I am checking with iThenticate to get a text-overlap score of the manuscript. High scores should be reviewed before editorial...
- Language quality**: 26 Apr 2020-15:04 GMT. I am checking the language quality of the manuscript and assigning it a recommended copy-editing level score.
- Ethics guidelines**: 26 Apr 2020-15:04 GMT. I am checking that the submission and manuscript complies with our ethics guidelines.
- Frontiers manuscript matches**: 26 Apr 2020-15:04 GMT. I did not detect similar manuscripts in Frontiers.
- Detection done by iThenticate**: Apr 2020-15:41 GMT. Text overlap OK. ✓ INDICATOR SOLVED. DETAILS
- Frontiers language rating**: 26 Apr 2020-15:04 GMT. The language level is rated L1 based on the evaluation of 97 out of 97 sentences extracted from the document.
- Animal studies statement verification**: 26 Apr 2020-15:04 GMT. The author stated that no animal studies are presented in the manuscript. I did not check for keywords because a statement was provided for one or more of the other ethics statement questions. Please review the other indicators.
- Human studies statement verification**: 26 Apr 2020-15:04 GMT. The author selected the following statement: The studies involving human participants were reviewed and approved by Beijing Jishuitan Hospital Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

# 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

```
{
  "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": ""
}
```



```
▼ HUMANDETECTIONASYNCAPI
  > .devcontainer
  > .vscode
  > deployment
  ▼ src
    > api
    > common
    > worker
    🐳 Dockerfile.api
    🐳 Dockerfile.worker
```



## 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](https://www.youtube.com/watch?v=06-AZXmwHjo&ab_channel=DeepLearningAI)
- Trace all developments and changes using a governance framework

# References



Best practices :

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



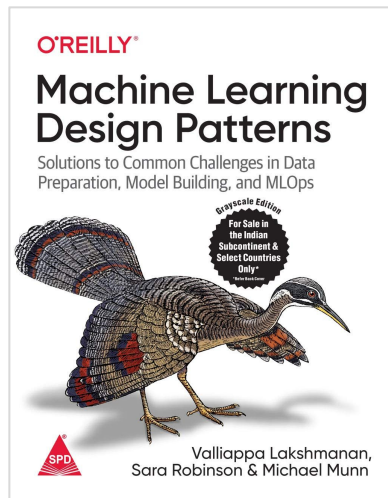
Chip Huyen course :

<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/>