

Votre mission :
déployer un modèle
de Machine Learning
en Production

Allez vous
l'accepter ?



CLAUDE FALGUIERE

- ▶ Consultante Data Engineering à Sentelis, Paris
@cfalguiere - <https://cfalguiere.github.io>
- ▶ Co-orga du Paris Machine Learning
@ParisMLGroup
- ▶ Mentor au Coder Dojo Paris
@CoderDojoParis
- ▶ Duchess France
@DuchessFr



L'objet de l'étude



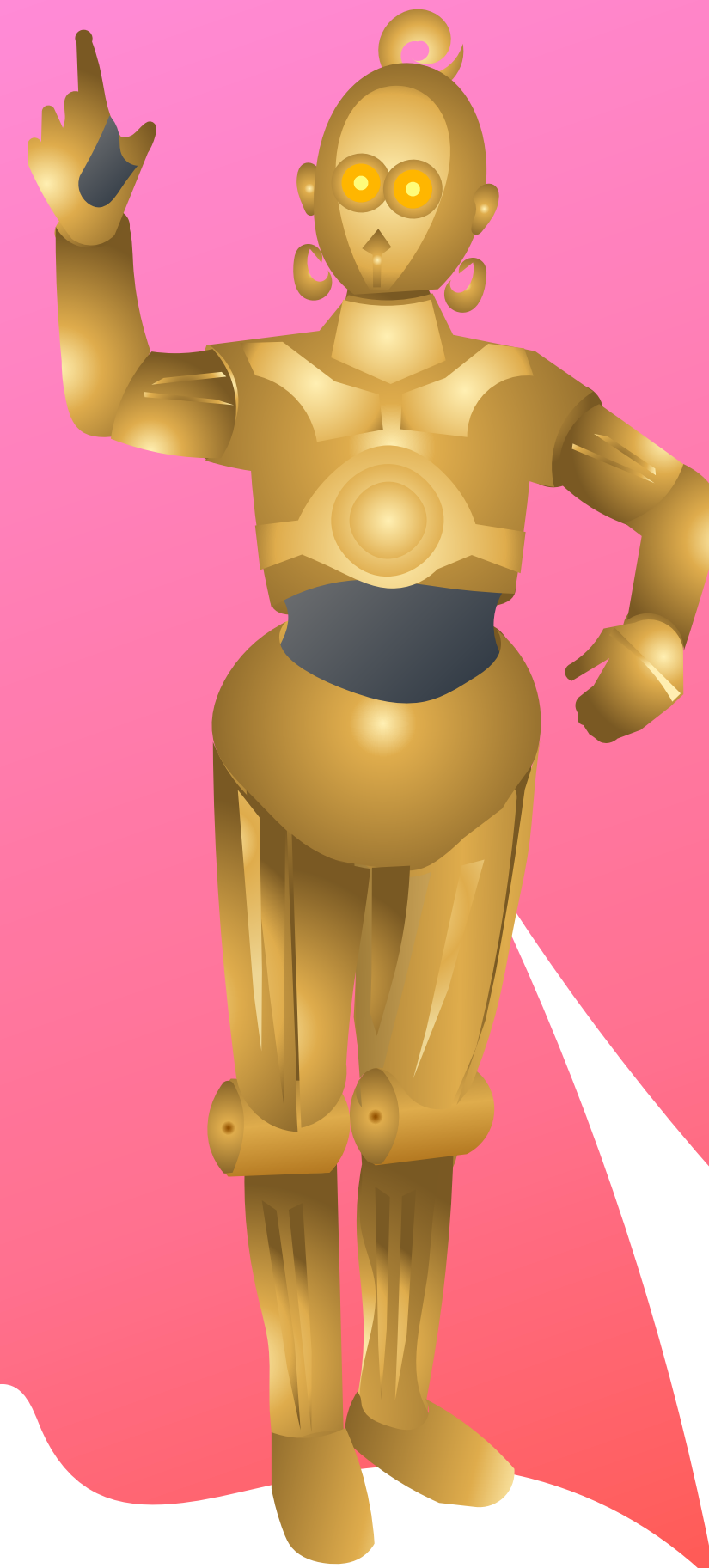
L'objet de l'étude



L'idéation



En route ...



Feuille de route

- ☐ Faire le modèle
- ☐ Encaisser la prime

L'épreuve du POC

Je vous épargne les détails :

- Remix notebook Python et bouts de R
- La moitié du code ne sert plus à rien
- Les variables s'appellent toute data
- Ça ne démarre pas sur votre poste

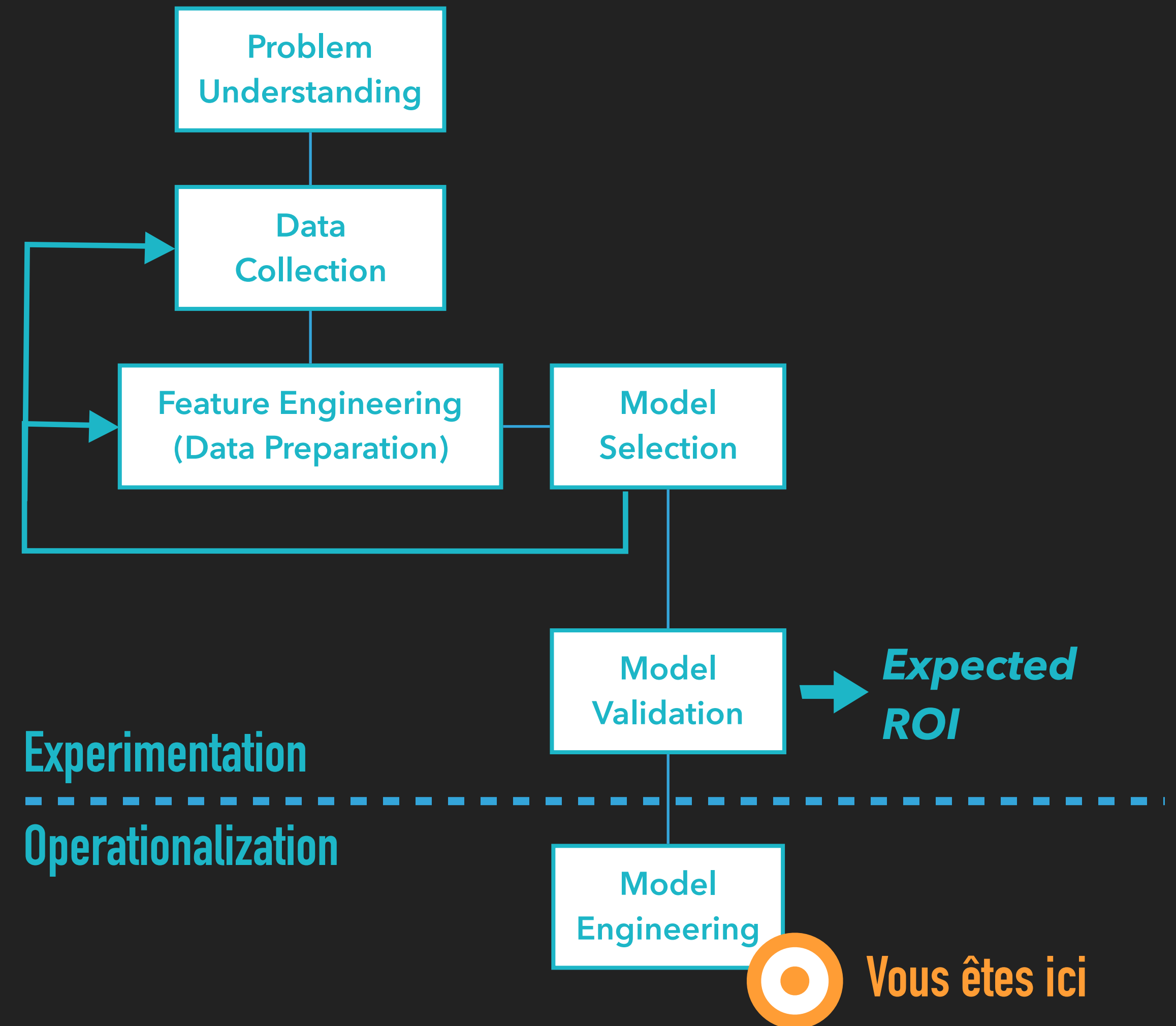
Mais ça donne un résultat, des fois

Bref ça évoque cette chose



L'industrialisation du Modèle

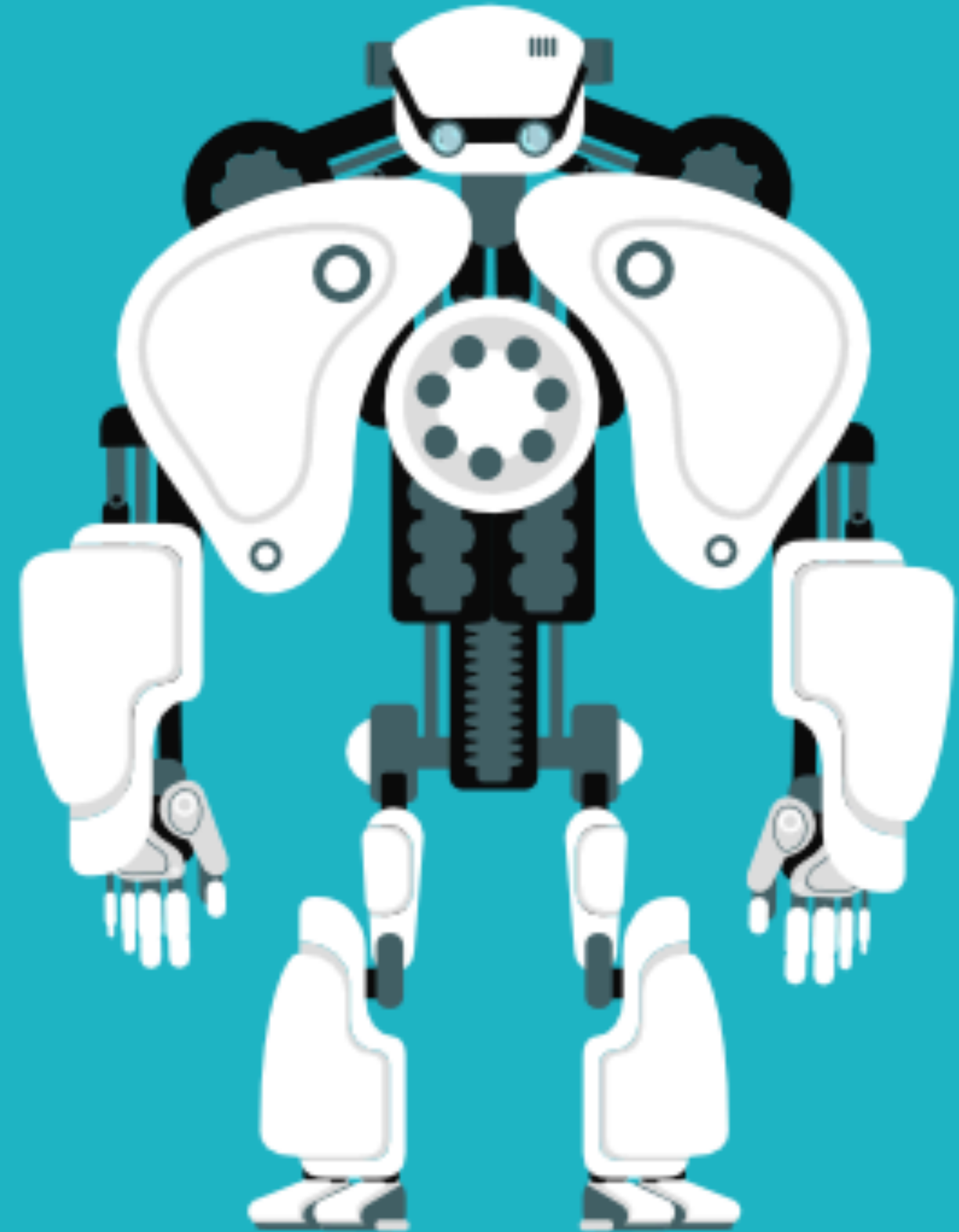
- ▶ Prototype
- ▶ Arbitrages
 - ▶ Adaptation
 - ▶ Refonte complète
 - ▶ Simplification
- ▶ Convergence
 - ▶ Montée en compétence des 2 côtés
 - ▶ Software Craftmanship en Data Lab
 - ▶ Rôle ML Engineer
 - ▶ Studios et outils intégrés



L'épreuve du POC (après toilettage)

C'est toujours moche, mais au moins

- Ça démarre
- On sait quand ça plante
- Ça dit ce que ça fait

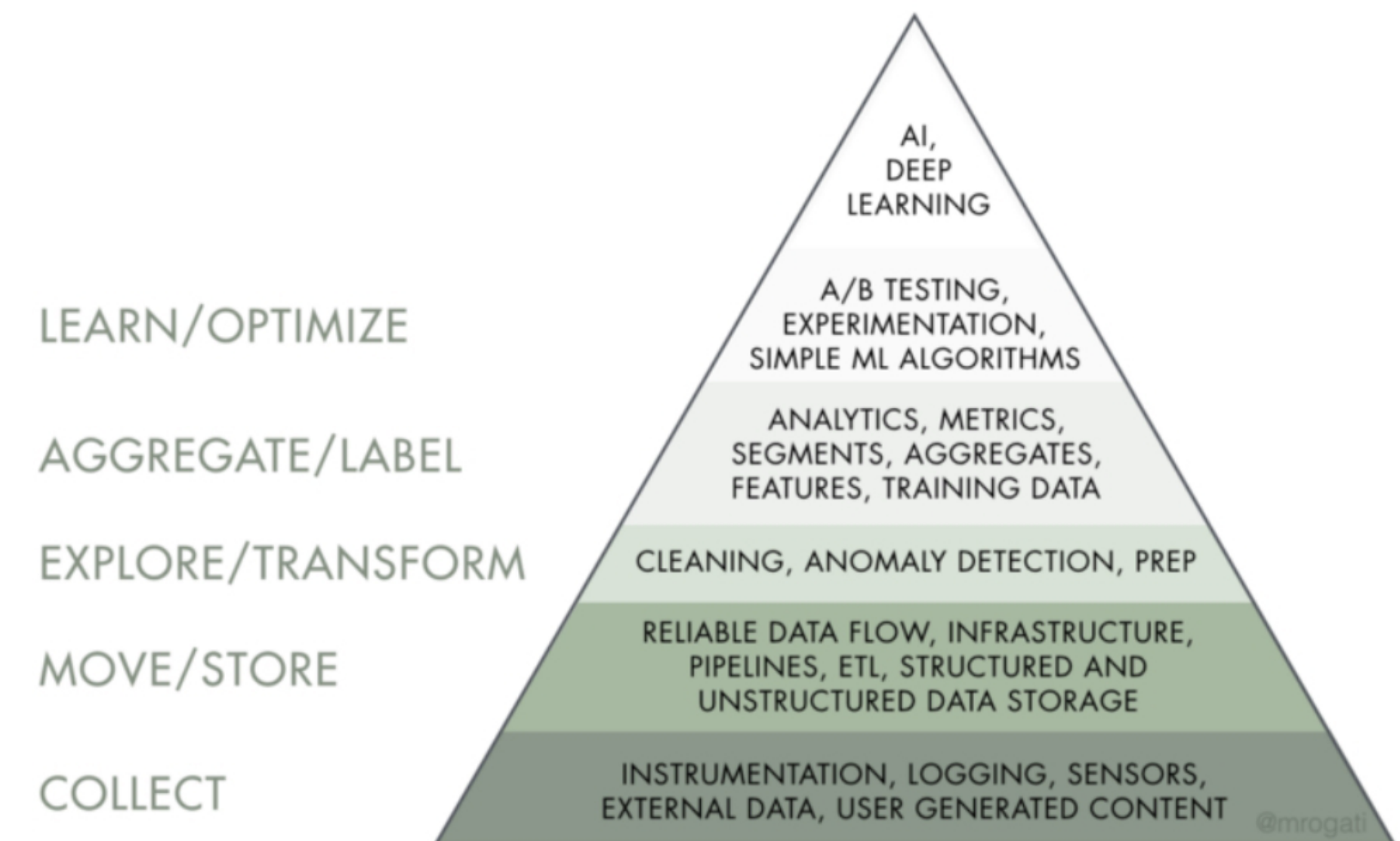


Le Mur des Data

Data, Data, Data

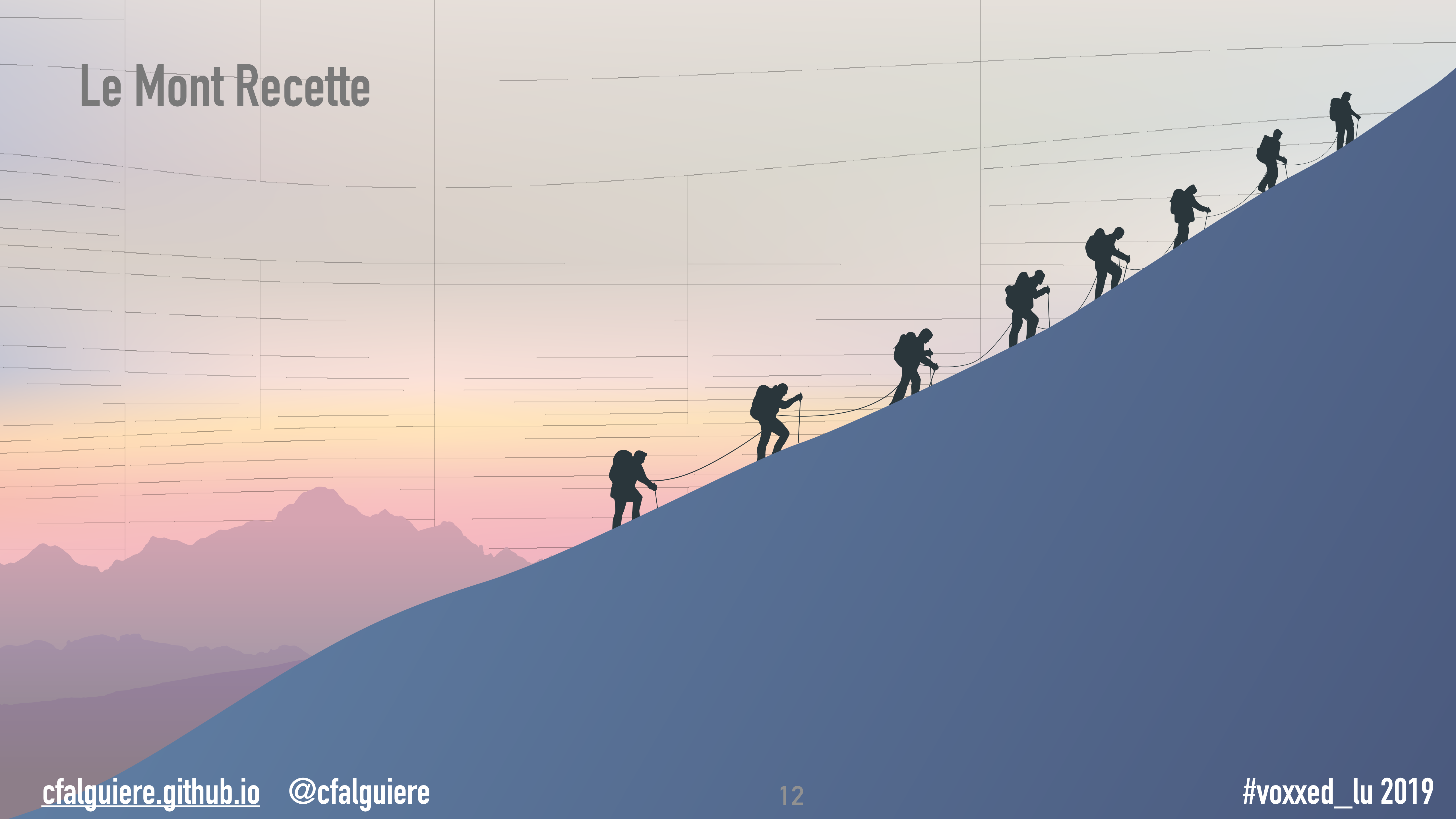
- ▶ Volume, Variété, Vitesse
- ▶ Gouvernance de données
- ▶ Sécurité, GPPR, PCI-DSS ...
 - ▶ Masquage, anonymisation ...
- ▶ Mise en qualité
 - ▶ Données manquantes
 - ▶ Labelisation pour les modèles supervisés
- ▶ Rôle Data Engineer

THE DATA SCIENCE HIERARCHY OF NEEDS



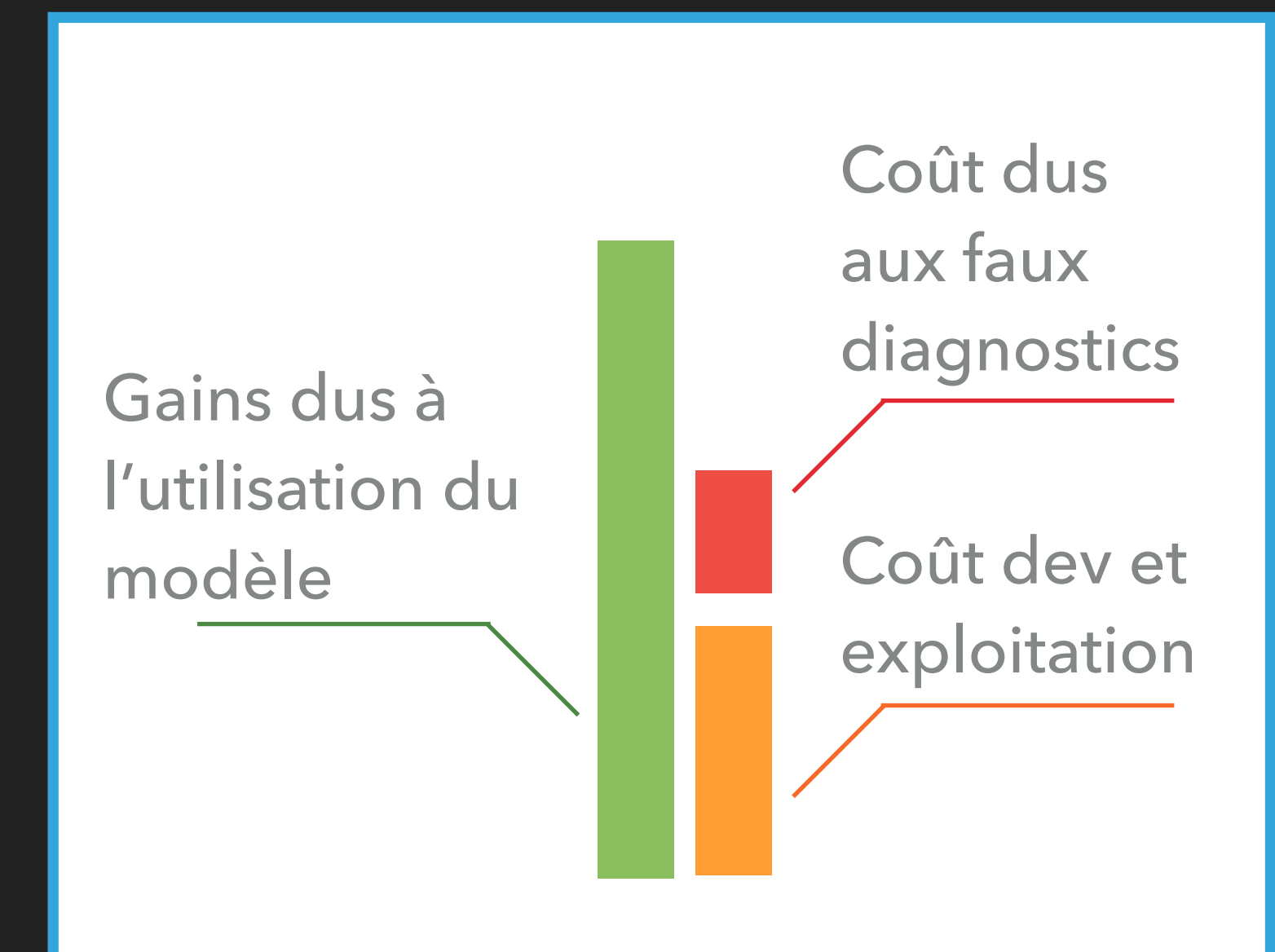
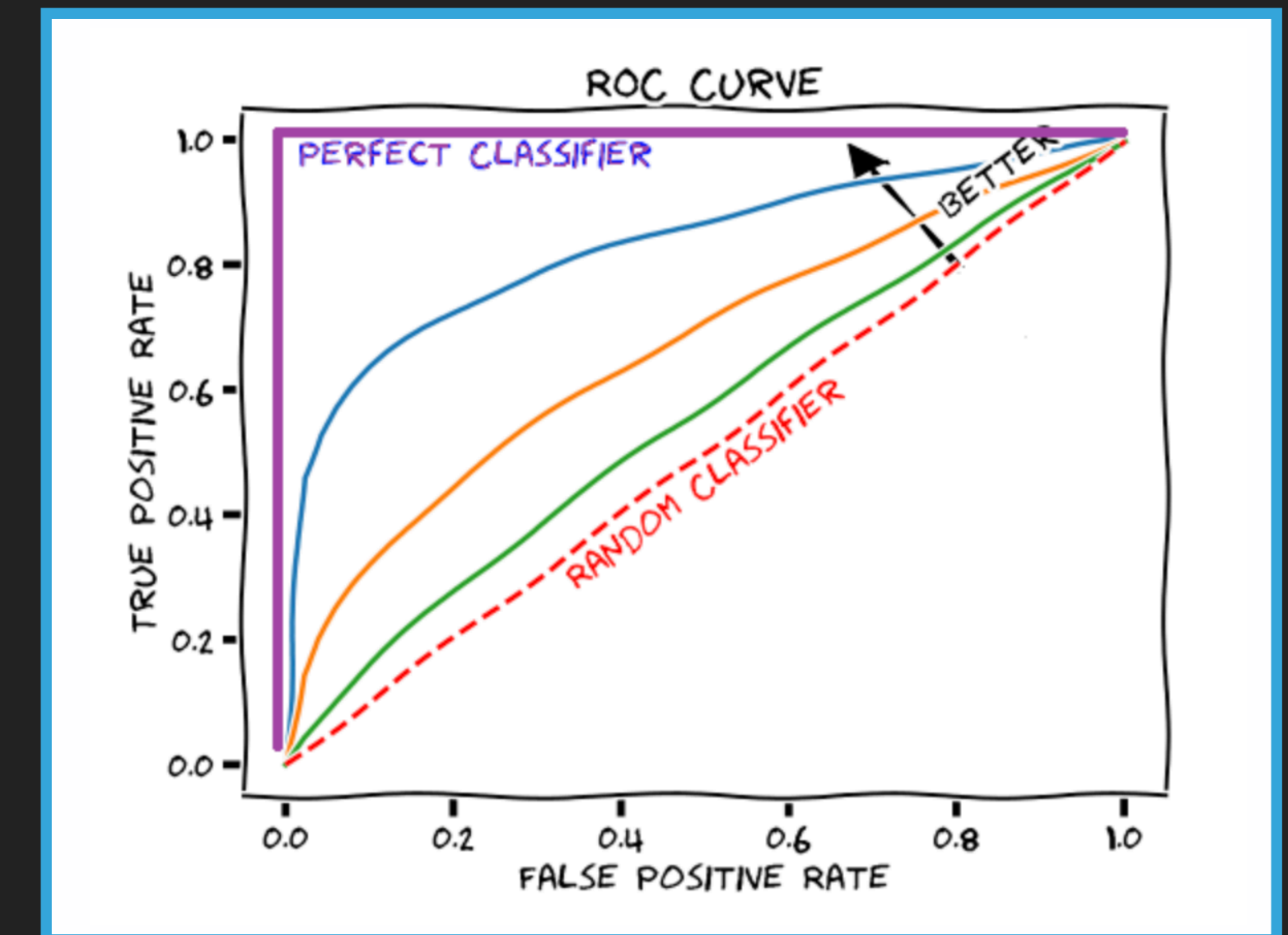
The popular AI Hierarchy of Needs from Monica Rogati

Le Mont Recette



Comment homologuer une modèle ?

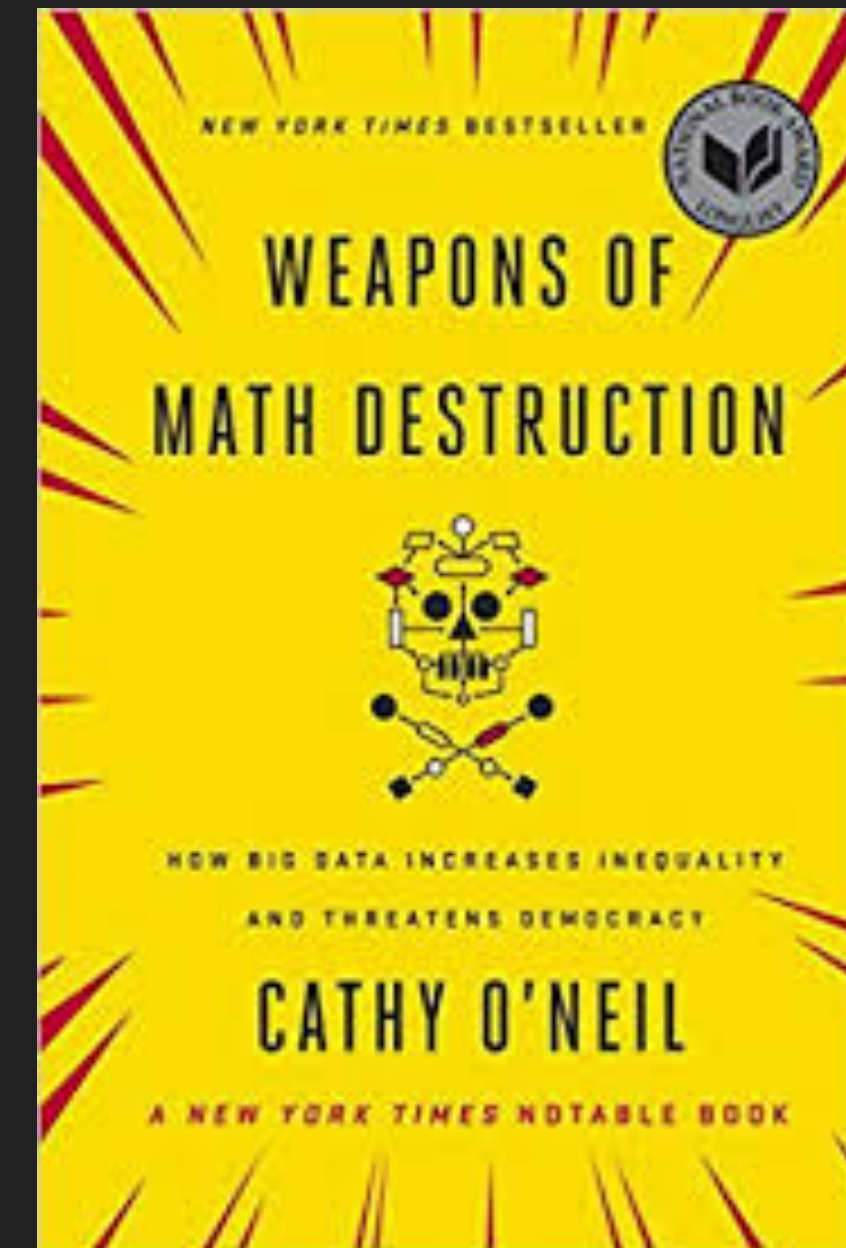
- ▶ Pas de règles explicites
- ▶ Fortement dépendant des données
- ▶ Difficile à reproduire
- ▶ Résultat probabiliste
- ▶ Métriques difficiles à interpréter
- ▶ Estimer un ROI probable



D'autres enjeux

- ▶ Robustesse aux variations de données
- ▶ Explicabilité
 - ▶ Contraintes sur les modèles
 - ▶ Nouveaux algorithmes et outils
- ▶ Équité (Fairness)
 - ▶ Testing pour recherche de biais
- ▶ Adversarial Testing
 - ▶ Recherche de cas limites

Source intéressante sur les enjeux éthiques



En Français :
Algorithmes, La bombe à retardement

Site Web: mathbabe.org

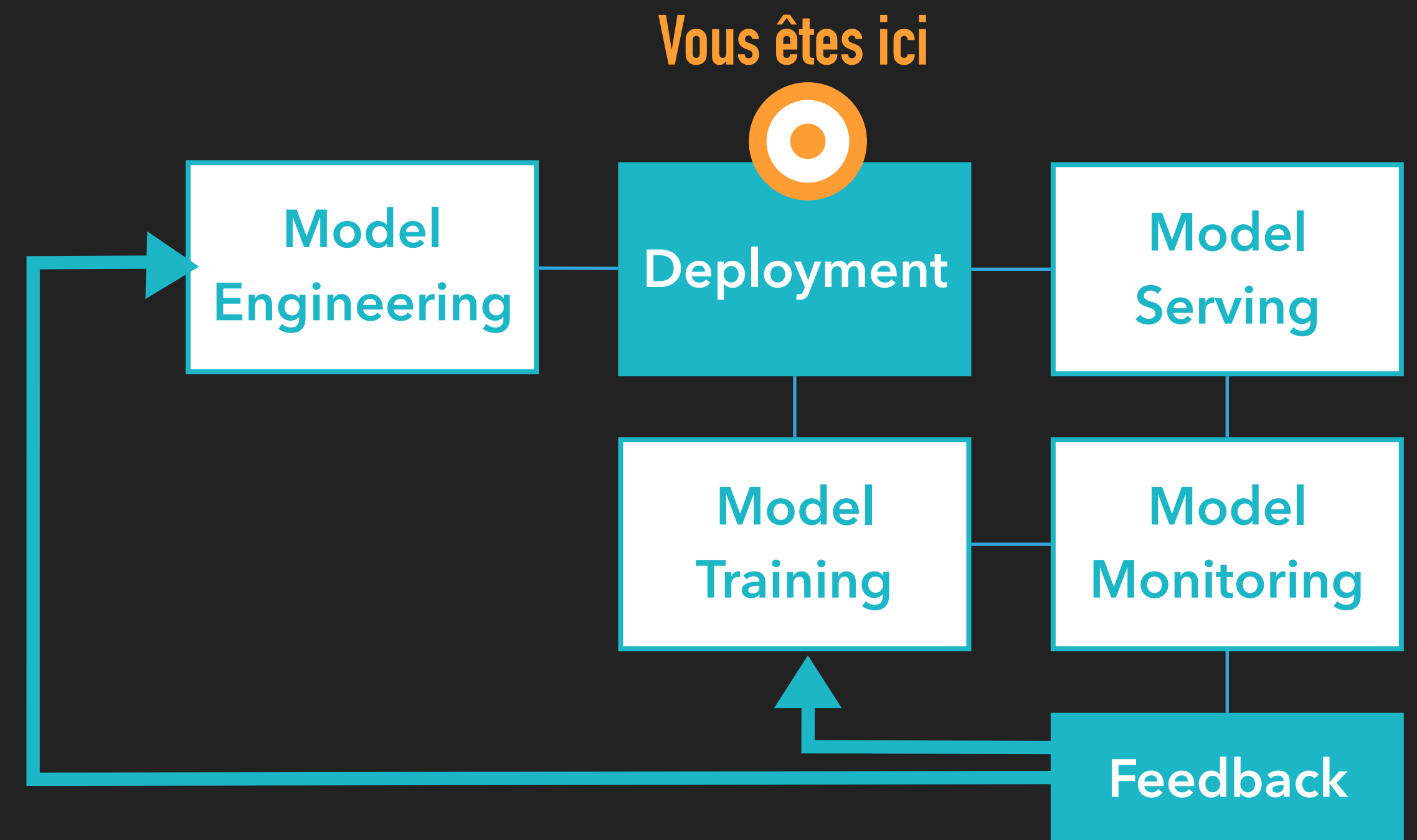
mathbabe.org

Le Marais de la Production

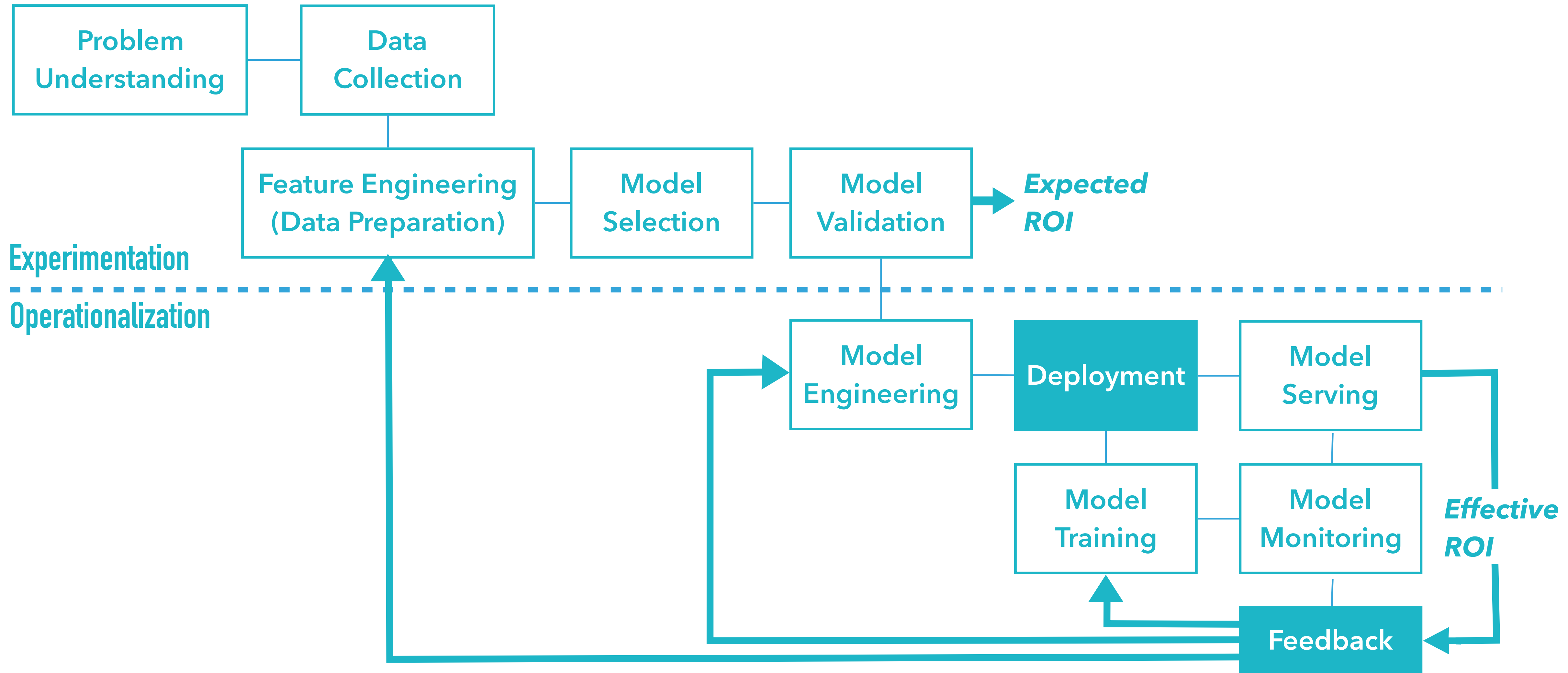
An aerial photograph of a vast wetland area, likely a salt marsh or tidal flat, characterized by a dense network of narrow, winding canals and rectangular ponds. The water is a dark, murky green, while the surrounding land is covered in lush, vibrant green vegetation. The layout is highly organized, suggesting a system designed for water management or agriculture. In the upper right, a larger, more irregular body of water is visible, possibly a reservoir or a natural inlet. The overall scene is a testament to human engineering in a natural environment.

Le déploiement

- ▶ Une myriade d'outils et de technologies
 - ▶ Conduite du changement
- ▶ Monitoring du modèle
 - ▶ Détecter les dérives du modèle
 - ▶ Evaluer sa pertinence
 - ▶ Détecter les changements de contexte
- ▶ Besoin de ré-ajuster
 - ▶ À la volée - sous contrôle
 - ▶ Revenir en développement ou étude



Activités classiques des projets IA



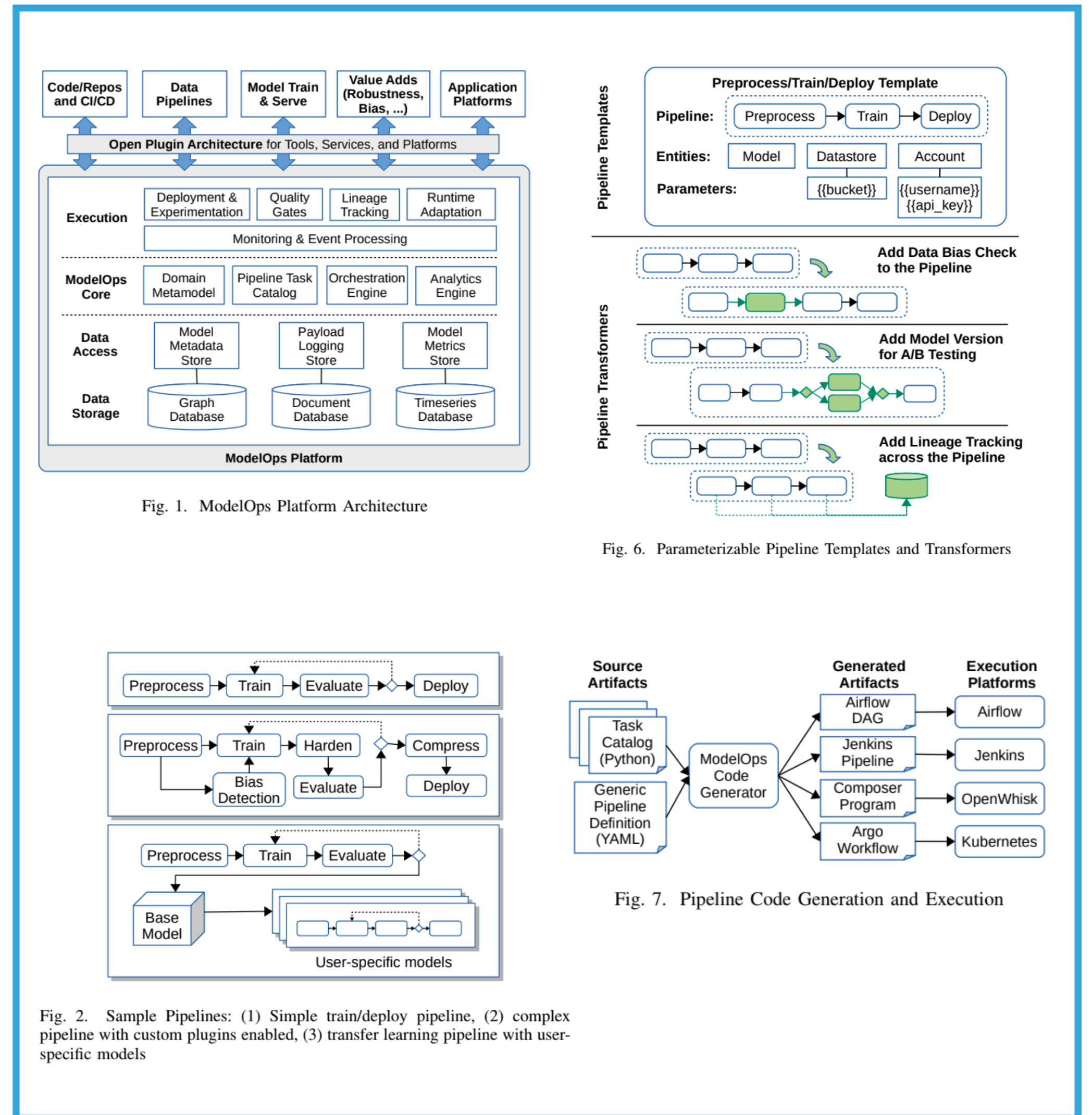
ModelOps

- Isoler la complexité technique
- Workflows et composants standards
- ModelOps: Cloud-based Lifecycle Management for Reliable and Trusted AI

Waldemar Hummer, Vinod Muthusamy, Thomas Rausch, Parijat Dube, Kaoutar El Maghraoui - IBM Research AI

- ModelOps: A programming model for reusable plat-form-independent, and composable AI workflows

Waldemar Hummer, Vinod Muthusamy,- IBM Research AI

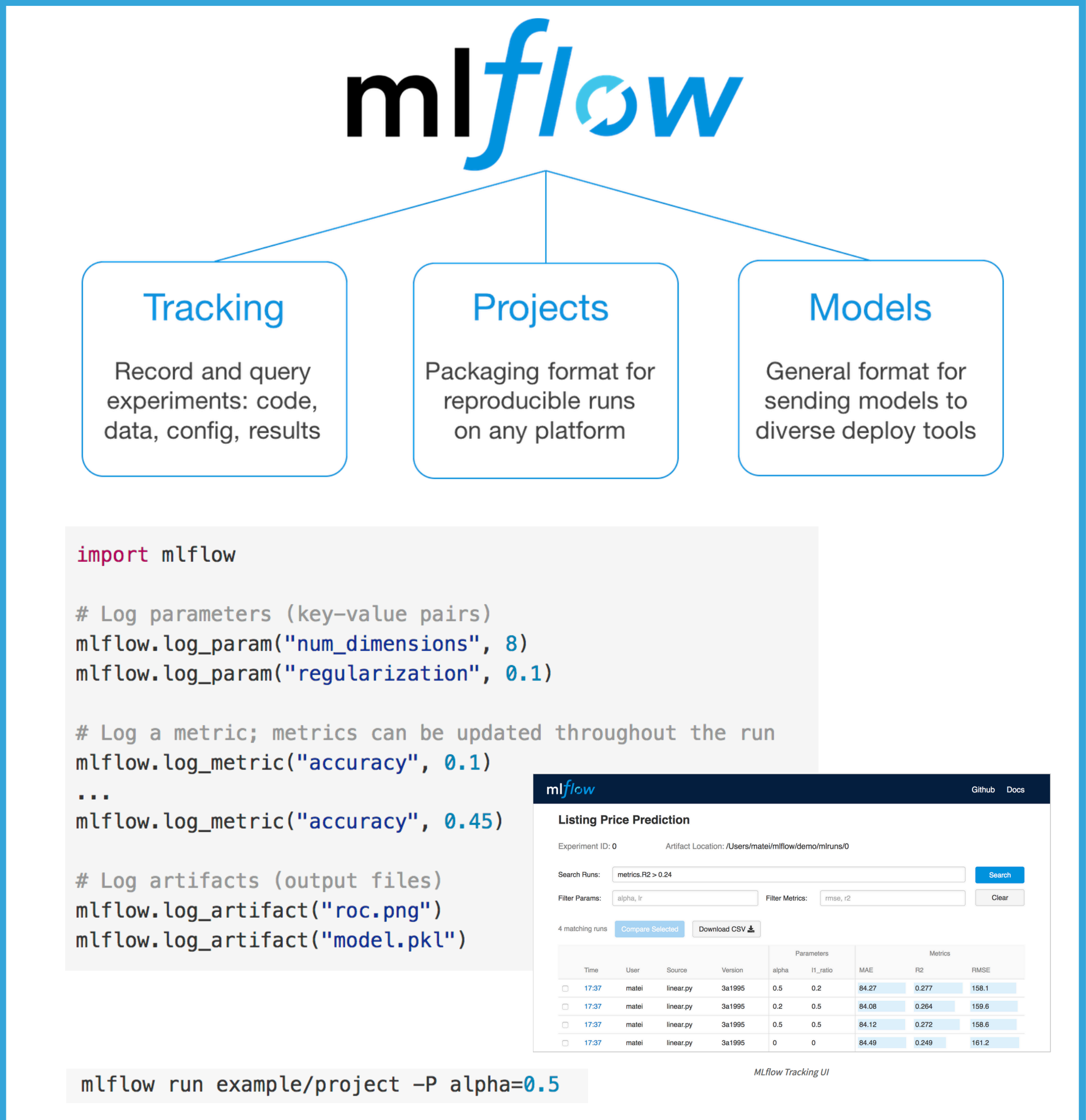


Portabilité

- ▶ Abstraire l'infrastructure distribuée
- ▶ Transposer les modèles
- ▶ ONNX
- ▶ H2O

ML Platforms

- ▶ MLFlow (Open Source Apache - DataBricks)
mlflow.org
- ▶ Introducing MLflow: an Open Source Machine Learning Platform - Blog databricks.com
- ▶ Socle d'exécution des modèles
 - ▶ Suivi paramètres et métriques par essai
 - ▶ Support d'un grand nombre de modèles
 - ▶ API Python, Java, REST
- ▶ Déploiement sur Databricks, SageMaker, AzureML, cluster Docker, Spark



The diagram illustrates the MLflow architecture, showing three main components: Tracking, Projects, and Models. Tracking is used to record and query experiments (code, data, config, results). Projects provide a packaging format for reproducible runs on any platform. Models offer a general format for sending models to diverse deployment tools.

```
import mlflow

# Log parameters (key-value pairs)
mlflow.log_param("num_dimensions", 8)
mlflow.log_param("regularization", 0.1)

# Log a metric; metrics can be updated throughout the run
mlflow.log_metric("accuracy", 0.1)
...
mlflow.log_metric("accuracy", 0.45)

# Log artifacts (output files)
mlflow.log_artifact("roc.png")
mlflow.log_artifact("model.pkl")
```

Below the code, a screenshot of the MLflow Tracking UI is shown. It displays a table of runs for a 'Listing Price Prediction' experiment. The table includes columns for Time, User, Source, Version, Parameters (alpha, l1_ratio), and Metrics (MAE, R2, RMSE). The runs are filtered by 'alpha, lr' and 'metrics.R2 > 0.24'.

Time	User	Source	Version	alpha	l1_ratio	MAE	R2	RMSE
17:37	matei	linear.py	3a1995	0.5	0.2	84.27	0.277	158.1
17:37	matei	linear.py	3a1995	0.2	0.5	84.08	0.264	159.6
17:37	matei	linear.py	3a1995	0.5	0.5	84.12	0.272	158.6
17:37	matei	linear.py	3a1995	0	0	84.49	0.249	161.2

mlflow run example/project -P alpha=0.5

DATA, DATA, DATA (80%)

GESTION DE PROJET

LA GESTION DU FEEDBACK

STANDARDISATION : MODEL OPS, ML PLATFORM, ONNX



Les meetups

- ▶ Paris Machine Learning Applications
Twitter : @ParisMLGroup
YouTube : Machine Learning Paris
- ▶ Data Engineering Paris
- ▶ WiMLDS



Merci
et bonne route