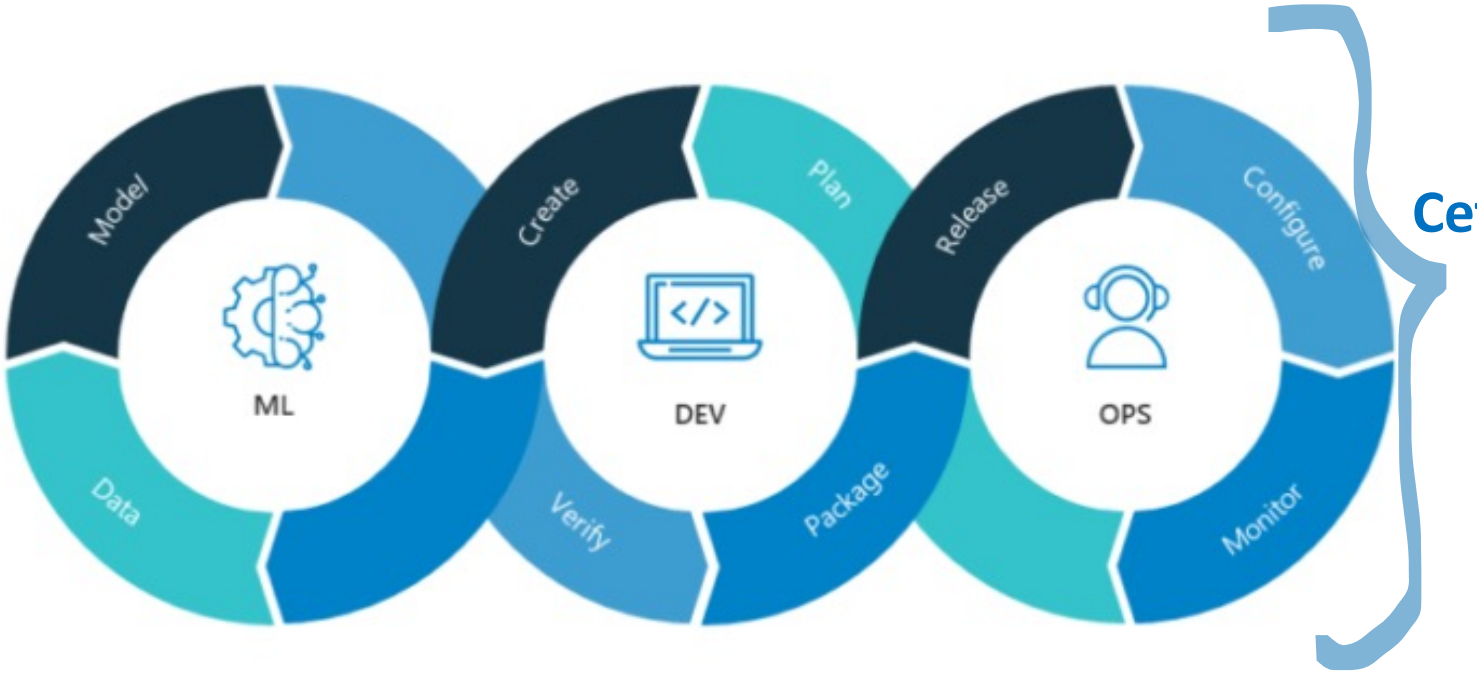
Déploiement des systèmes d'intelligence artificielle

MGL 7320: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD (+ Laurent Magnin, PhD)
Université du Québec à Montréal

Machine Learning Operations (Partie 2)



Problèmes et solutions

1

Learning under Concept Drift: A Review

Jie Lu, *Fellow, IEEE,* Anjin Liu, *Member, IEEE,* Fan Dong, Feng Gu, João Gama, and Guangquan Zhang

Abstract—Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. Concept drift research involves the development of methodologies and techniques for drift detection, understanding and adaptation. Data analysis has revealed that machine learning in a concept drift environment will result in poor learning results if the drift is not addressed. To help researchers identify which research topics are significant and how to apply related techniques in data analysis tasks, it is necessary that a high quality, instructive review of current research developments and trends in the concept drift field is conducted. In addition, due to the rapid development of concept drift in recent years, the methodologies of learning under concept drift have become noticeably systematic, unveiling a framework which has not been mentioned in literature. This paper reviews over 130 high quality publications in concept drift related research areas, analyzes up-to-date developments in methodologies and techniques, and establishes a framework of learning under concept drift including three main components: concept drift detection, concept drift understanding, and concept drift adaptation. This paper lists and discusses 10 popular synthetic datasets and 14 publicly available benchmark datasets used for evaluating the performance of learning algorithms aiming at handling concept drift. Also, concept drift related research directions are covered and discussed. By providing state-of-the-art knowledge, this survey will directly support researchers in their understanding of research developments in the field of learning under concept drift.

Index Terms—concept drift, change detection, adaptive learning, data streams

MLOps: Continuous delivery and automation pipelines in machine learning

This document discusses techniques for implementing and automating continuous integration (CI), continuous delivery (CD), and continuous training (CT) for machine learning (ML) systems.

Continuous Delivery for Machine Learning

Automating the end-to-end lifecycle of Machine Learning applications

Concept Drift (Dérive de concept)

Met l'accent sur la présentation d'un aperçu des méthodes pour :

1. Identifier
2. Comprendre et
3. Réagir

... à la dérive de concept

1

Learning under Concept Drift: A Review

Jie Lu, *Fellow, IEEE,* Anjin Liu, *Member, IEEE,* Fan Dong, Feng Gu, João Gama, and Guangquan Zhang

Abstract—Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. Concept drift research involves the development of methodologies and techniques for drift detection, understanding and adaptation. Data analysis has revealed that machine learning in a concept drift environment will result in poor learning results if the drift is not addressed. To help researchers identify which research topics are significant and how to apply related techniques in data analysis tasks, it is necessary that a high quality, instructive review of current research developments and trends in the concept drift field is conducted. In addition, due to the rapid development of concept drift in recent years, the methodologies of learning under concept drift have become noticeably systematic, unveiling a framework which has not been mentioned in literature. This paper reviews over 130 high quality publications in concept drift related research areas, analyzes up-to-date developments in methodologies and techniques, and establishes a framework of learning under concept drift including three main components: concept drift detection, concept drift understanding, and concept drift adaptation. This paper lists and discusses 10 popular synthetic datasets and 14 publicly available benchmark datasets used for evaluating the performance of learning algorithms aiming at handling concept drift. Also, concept drift related research directions are covered and discussed. By providing state-of-the-art knowledge, this survey will directly support researchers in their understanding of research developments in the field of learning under concept drift.

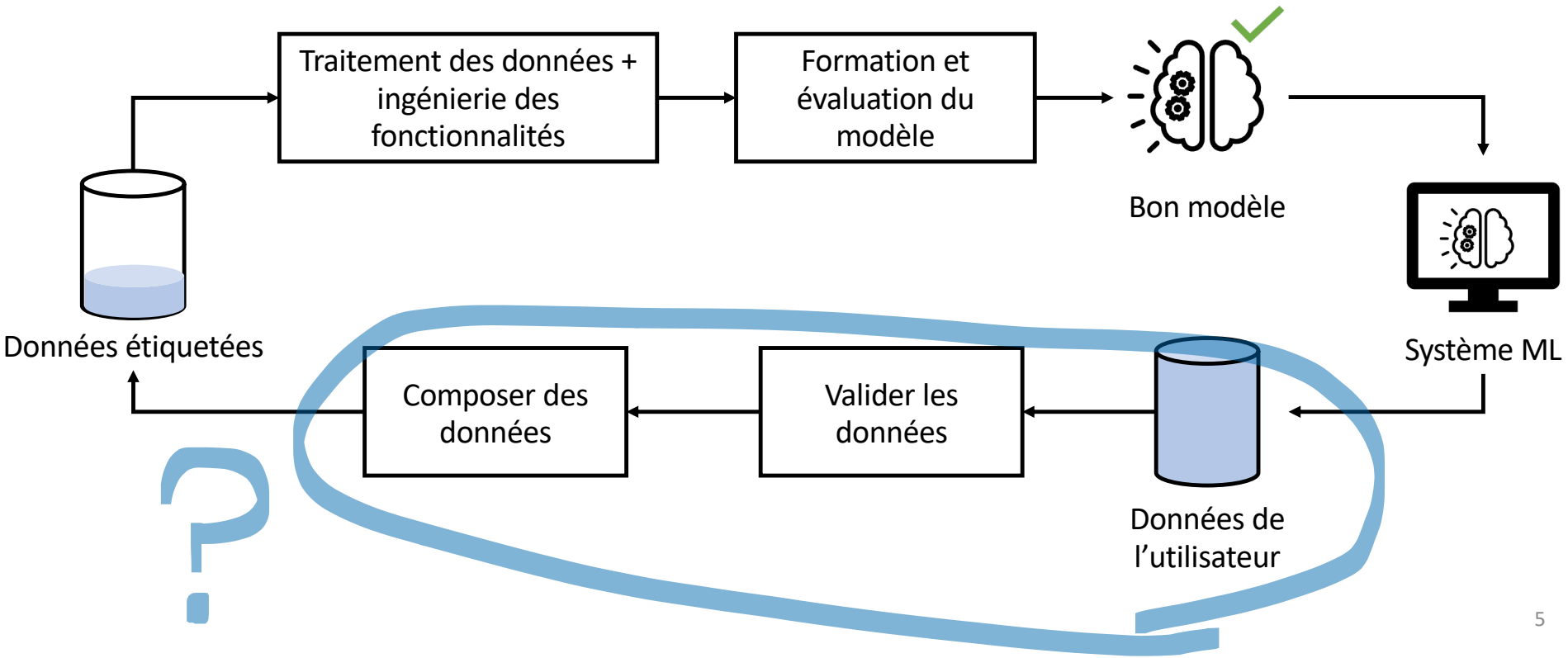
Index Terms—concept drift, change detection, adaptive learning, data streams

Study Design?

- Revue systématique de la littérature (SLR)
- Revue de 130 publications

Publié le IEEE Transactions on Knowledge and Data Engineering

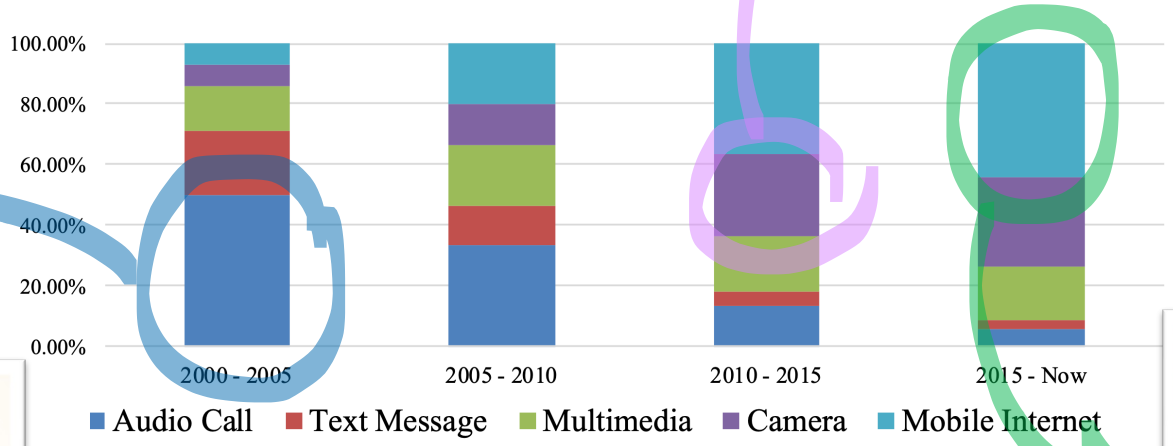
Pipeline de données des systèmes ML



ChangEMents

Exemples de changements

- Changement **progressif** dans l'utilisation des téléphones portables

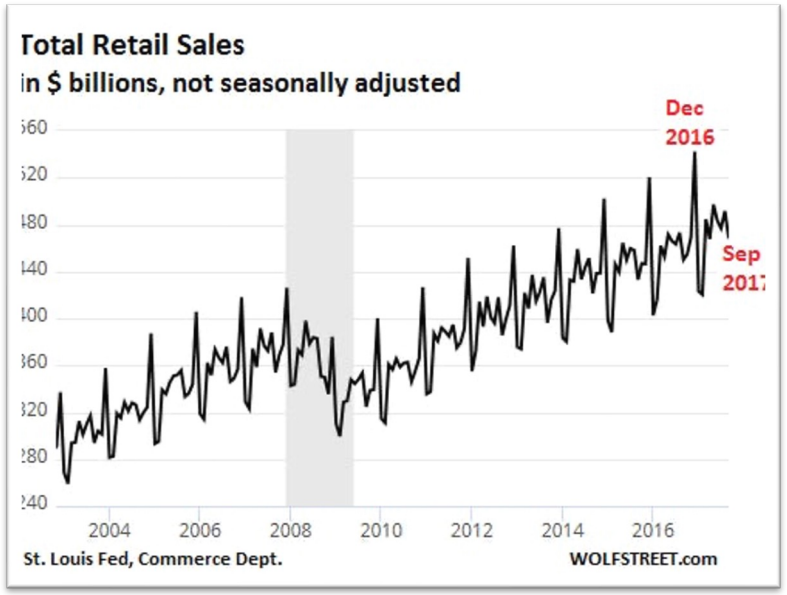


Exemples de changements

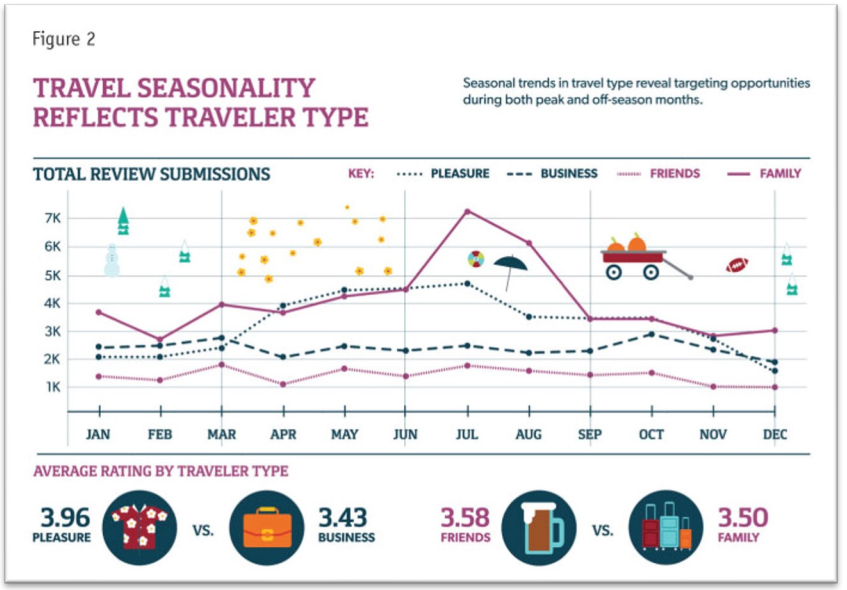
- Changement **soudain** et global sur les marchés boursiers

Exemples de changements

- Changements **saisonniers**



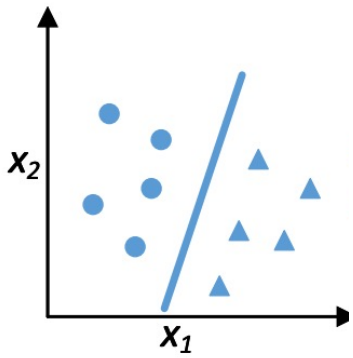
https://www.businessinsider.com/retail-sales-september-what-everyone-got-wrong-2017-10



https://www.quirks.com/articles/how-consumer-feedback-can-change-what-you-know-about-seasonal-buying-patterns

Sources de la dérive de concept

- Changements dans les caractéristiques des données : $P_t(X)$
- Changements dans l'étiquetage (classe) : $P_t(y|X)$



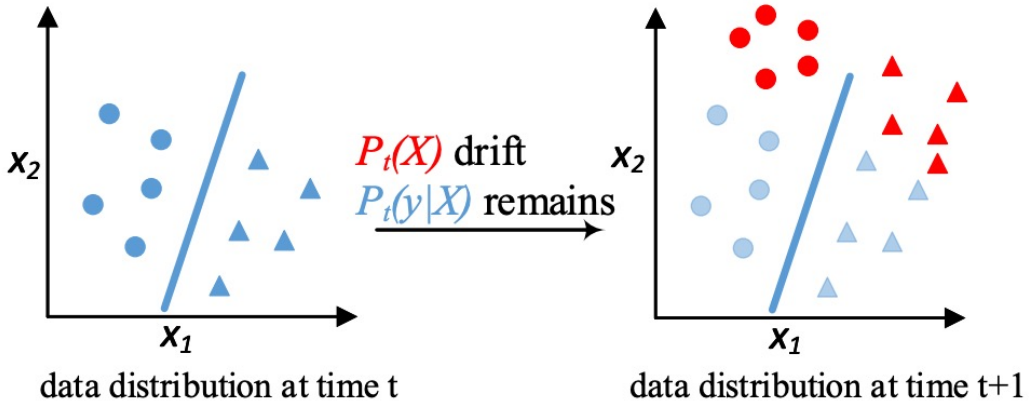
data distribution at time t

x_1

data distribution at time t+1

Sources de la dérive de concept

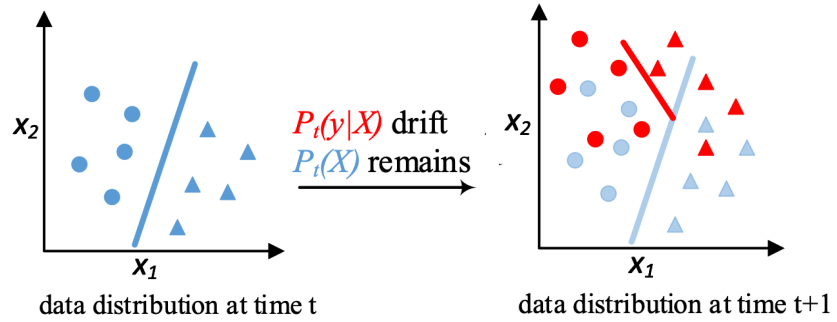
- Changements dans les caractéristiques des données : $P_t(X)$
- Changements dans l'étiquetage (classe) : $P_t(y|X)$

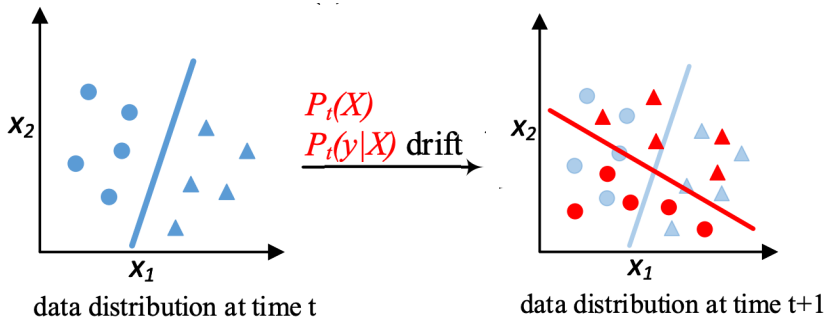


Source I

- Frontières de décision inchangées
- Distribution des données modifiée
- Dérive virtuelle

Sources de la dérive de concept (cont.)





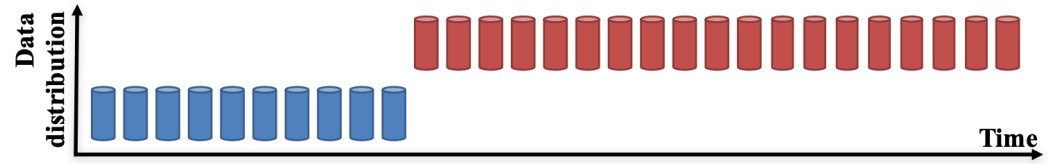
Source II

- Les frontières de décision changent
- Distribution des données inchangée
- Dérive d'étiquetage

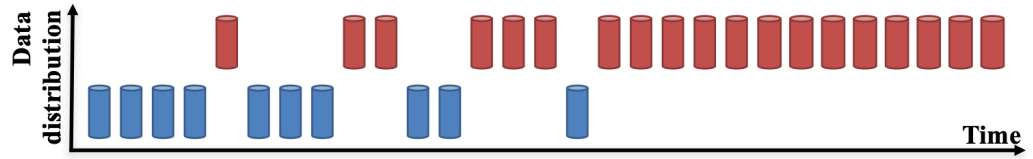
Source III

- Les frontières de décision changent
- Distribution des données modifiée
- Dérive de concept

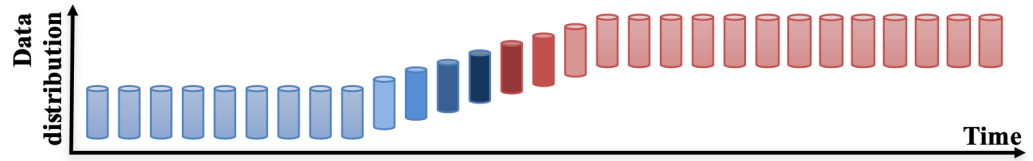
Types de dérive de concept



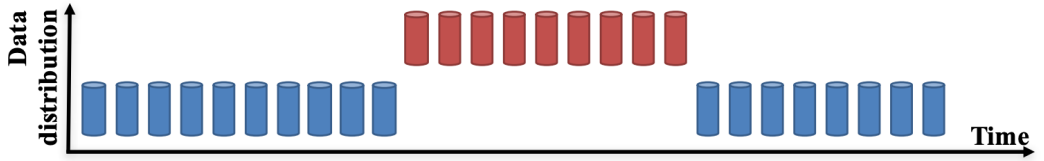
Sudden Drift:
A new concept occurs within a short time.



Gradual Drift:
A new concept gradually replaces an old one over a period of time.

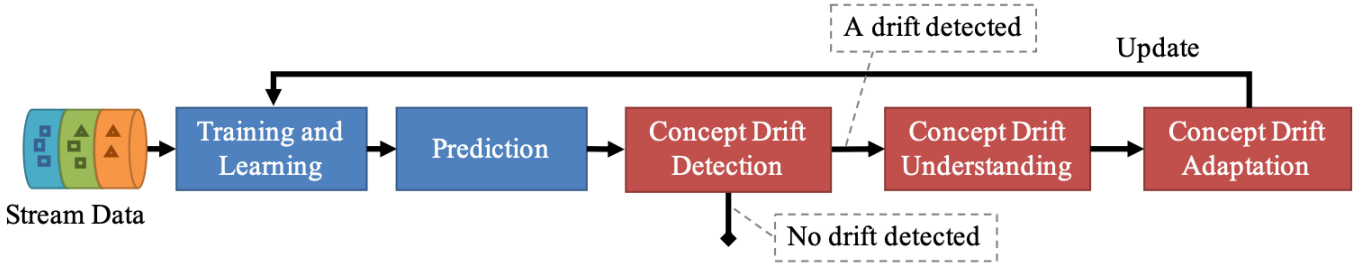


Incremental Drift:
An old concept incrementally changes to a new concept over a period of time.



Reoccurring Concepts:
An old concept may reoccur after some time.

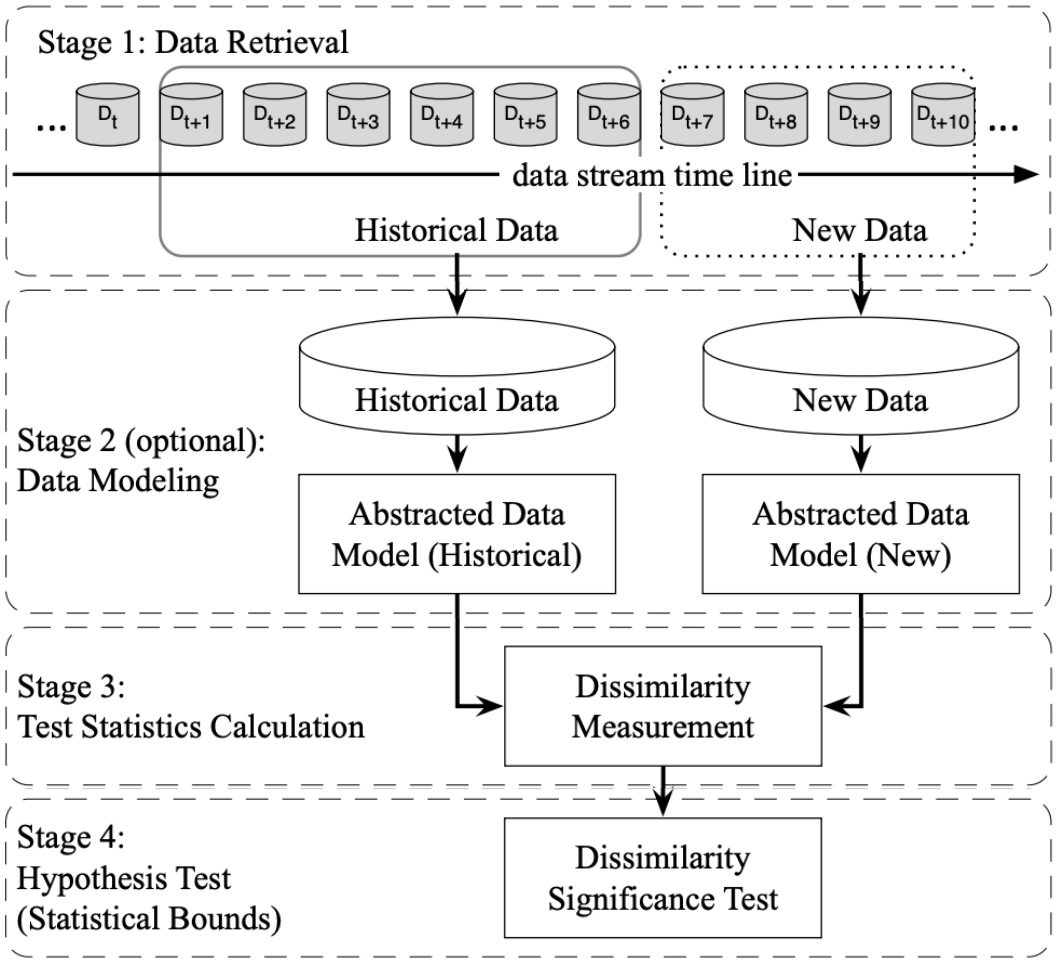
Gestion de la dérive de concept



L'article se concentre sur trois tâches principales liées à la derive

- Identifier la dérive de concept
  - Se produit-elle ?
- Comprendre la dérive de concept
  - Pourquoi se produit-elle ?
- Réagir à la dérive de concept
  - Comment garantir la qualité de notre système d'IA au fil du temps ?

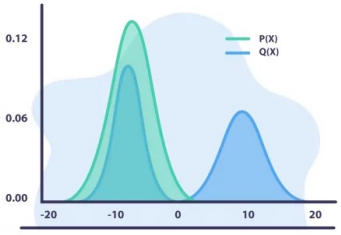
Identifier la dérive de concept



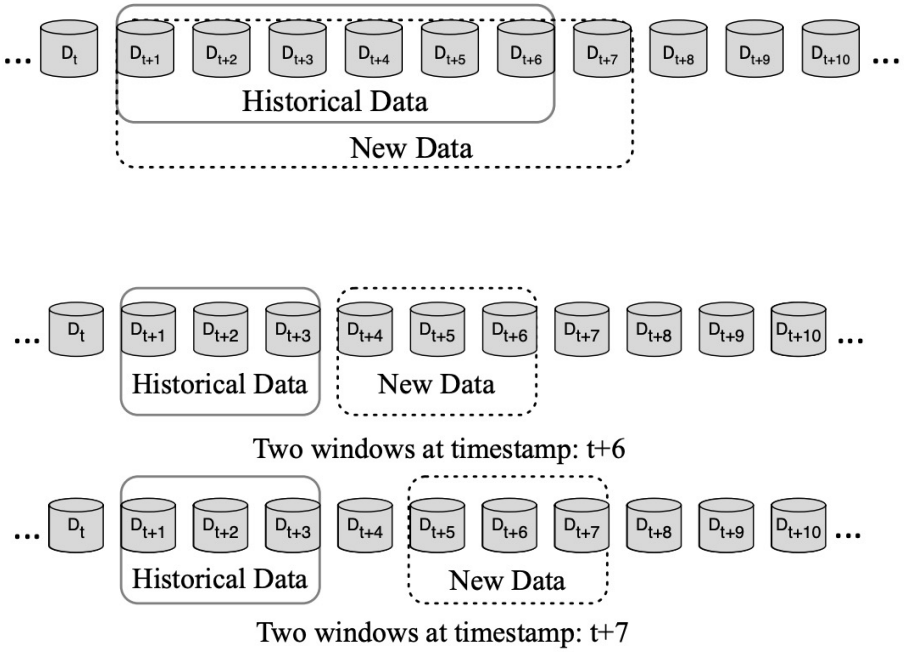
Identifier la dérive de concept (cont.)

Méthodes basées sur l'erreur

- Drift Detection Method (DDM)
  - Calculer le taux d'erreur
  - Un taux d'erreur élevé déclenche un nouveau modèle
- Méthodes basées sur la distribution



Kullback–Leibler Divergence

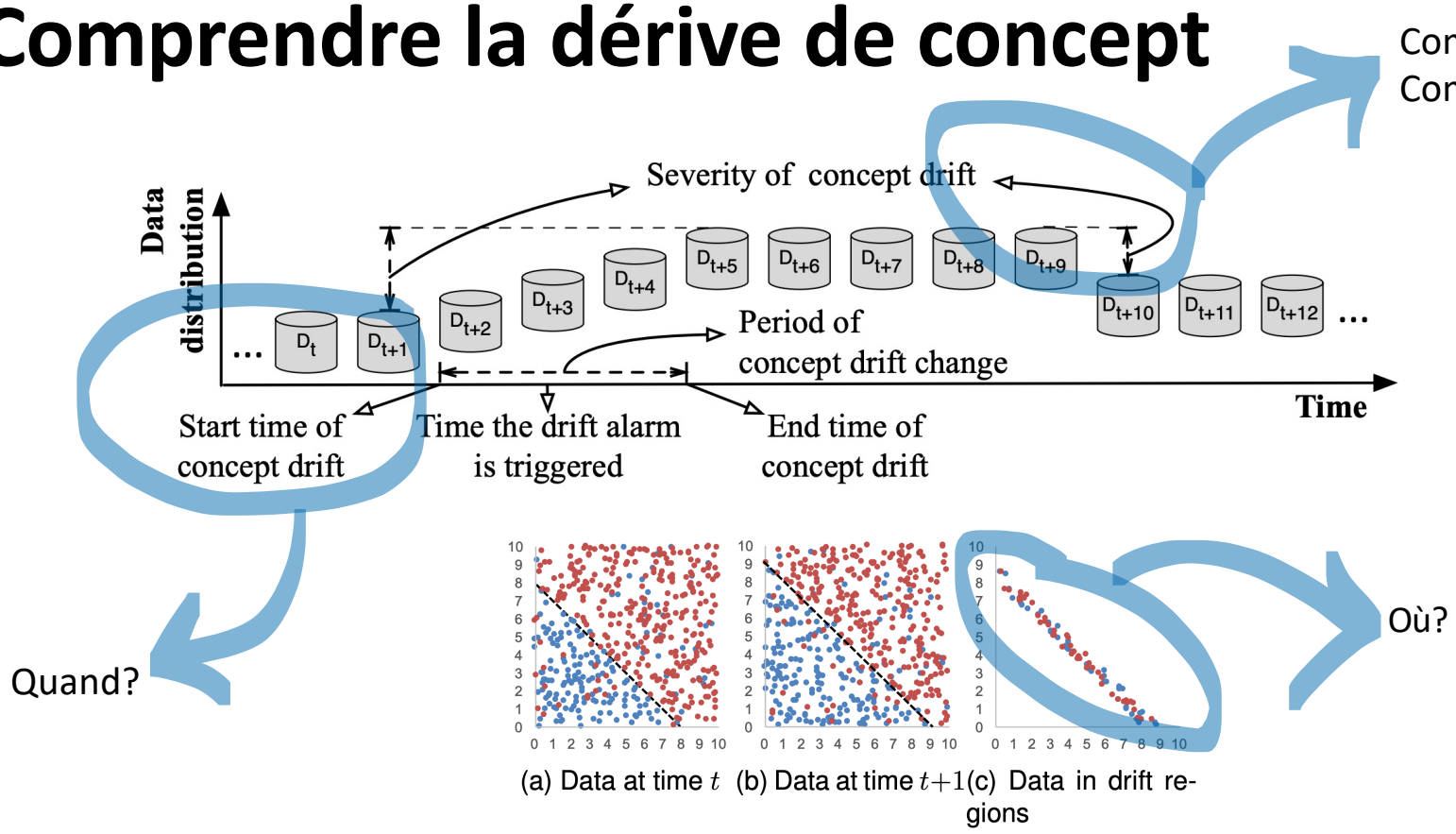


https://www.aporia.com/learn/data-drift/concept-drift-detection-methods/

Identifier la dérive de concept (cont.)

Category	Algorithms	Stage 1	Stage 2	Stage 3	Stage 4
Error rate-based	DDM [20]	Landmark	Learner	Online error rate	Distribution estimation
	EDDM [26]	Landmark	Learner	Online error rate	Distribution estimation
	FW-DDM [5]	Landmark	Learner	Online error rate	Distribution estimation
	DEML [27]	Landmark	Learner	Online error rate	Distribution estimation
	STEPD [30]	Predefined w_{hist}, w_{new}	Learner	Error rate difference	Distribution estimation
	ADWIN [31]	Auto cut w_{hist}, w_{new}	Learner	Error rate difference	Hoeffding's Bound
	ECDD [29]	Landmark	Learner	Online error rate	EWMA Chart
	HDDM [23]	Landmark	Learner	Online error rate	Hoeffding's Bound
	LLDD [25]	Landmark, or sliding w_{hist}, w_{new}	Decision trees	Tree node error rate	Hoeffding's Bound
Data distribution-based	kdqTree [22]	Fixed w_{hist}, Sliding w_{new}	kdqTree	KL divergence	Bootstrapping
	CM [2], [3]	Fixed w_{hist}, Sliding w_{new}	Competence model	Competence distance	Permutation test
	RD [37]	Fixed w_{hist}, Sliding w_{new}	KS structure	Relativized Discrepancy	VC-Dimension
	SCD [38]	Fixed w_{hist}, Sliding w_{new}	kernel density estimator	log-likelihood	Distribution estimation
	EDE [40]	Fixed w_{hist}, Sliding w_{new}	Nearest neighbor	Density scale	Permutation test
	SyncStream [36]	Fixed w_{hist}, Sliding w_{new}	PCA	P-Tree	Wilcoxon test
	PCA-CD [39]	Fixed w_{hist}, Sliding w_{new}	PCA	Change-Score	Page-Hinkley test
	LSDD-CDT [21]	Fixed w_{hist}, Sliding w_{new}	Learner	Relative difference	Distribution estimation
	LSDD-INC [41]	Fixed w_{hist}, Sliding w_{new}	Learner	Relative difference	Distribution estimation
	LDD-DSDA [4]	Fixed w_{hist}, Sliding w_{new}	k-nearest neighbor	Local drift degree	Distribution estimation
Multiple Hypothesis Tests	JIT [19]	Landmark	Selected features	4 configurations	Distribution estimation
	LFR [46]	Landmark	Learner	TP, TN, FP, FN	Distribution estimation
	Three-layer [47]	Sliding both w_{hist}, w_{new}	Learner	$P(y)$, $P(X)$, $P(X\|y)$	Distribution estimation
	e-Detector [48]	depends on base detector	depends	depends	depends
	DDE [49]	depends on base detector	depends	depends	depends
	TSMSD-EWMA [52]	Landmark	Learner	Online error rate	EWMA Chart
	HCDTs [50]	Landmark	Depending on layers	Depending on layers	Depending on layer
	HLFR [51]	Landmark	Learner	TP, TN, FP, FN	Distribution estimation
	HHT-CU [53]	Landmark	Learner	Classification uncertainty	Layer-I Hoeffding's Bound, Layer-II Permutation Test
	HHT-AG [53]	Fixed w_{hist}, Sliding w_{new}	N/A	KS statistic on each attribute	Layer-I KS test, Layer -II 2D KS test

Comprendre la dérive de concept



(a) Data at time t (b) Data at time $t+1$ (c) Data in drift regions

Comprendre la dérive de concept (cont.)

TABLE 2
Summary of drift understanding for drift detection algorithms

Category	Algorithms	When	How	Where
Error rate-based	DDM [20]	√		
	EDDM [26]	√		
	FW-DDM [5]	√		
	DEML [27]	√		
	STEPD [30]	√		
	ADWIN [31]	√		
	ECDD [29]	√		
	HDDM [23]	√		
	LLDD [25]	√		√
Data distribution-based	kdqTree [22]	√	√	√
	CM [2], [3]	√	√	√
	RD [37]	√	√	
	SCD [38]	√	√	
	EDE [40]	√		
	SyncStream [36]	√	√	
	PCA-CD [39]	√	√	
	LSDD-CDT [21]	√		
	LSDD-INC [41]	√		
	LDD-DSDA [4]	√	√	√
Multiple hypothesis tests	JIT [19]	√		
	LFR [46]	√		
	Three-layer drift detection [47]	√		
	e-Detector [48]	√		
	DDE [49]	√		
	EWMA [52]	√		
	HCDTs [50]	√		
	HLFR [51]	√		
	HHT-CU [53]	√		
	HHT-AG [53]	√		

Réagir à la dérive : Apprentissage en binôme

- Paired learners (Apprentissage en binôme)
  - Apprenant stable : entraîné sur de vieilles données
  - Apprenant réactif : entraîné sur de nouvelles données.
  - Changement lorsque l'erreur de l'apprenant stable > apprenant réactif."

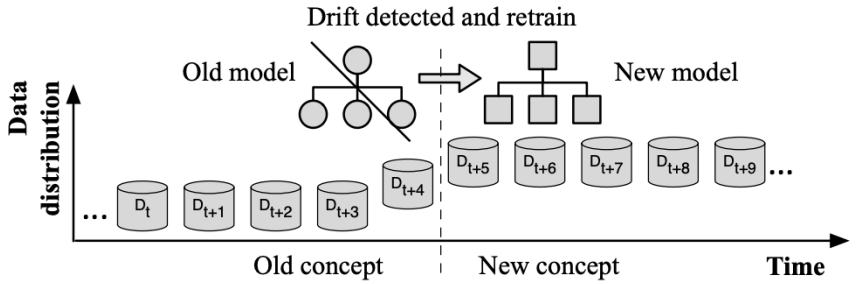


Fig. 13. A new model is trained with latest data to replace the old model when a concept drift is detected.

Réagir à la dérive : Modèles en ensemble

- Modèles en ensemble
  - Ideal pour la dérive récurrente
  - Inclure un nouveau modèle dans l'ensemble une fois qu'une dérive est identifiée
    - Le nouveau modèle classifie mieux les nouveaux concepts.

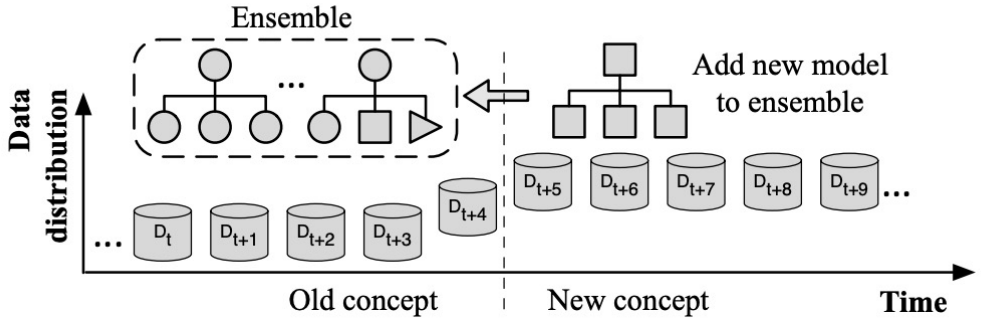


Fig. 14. A new base classifier is added to the ensemble when a concept drift occurs.

Conclusions

Un article d'introduction incontournable sur la dérive de concept.

- Présente un aperçu des méthodes.

Défis futurs :

- La compréhension de la dérive de concept est encore limitée.
- Les méthodes actuelles reposent sur des hypothèses irréalistes.
- Les méthodes supposent la présence de données étiquetées.
  - Les données étiquetées sont coûteuses et ne sont pas toujours possibles à obtenir dans un court laps de temps.

Problèmes et Solutions

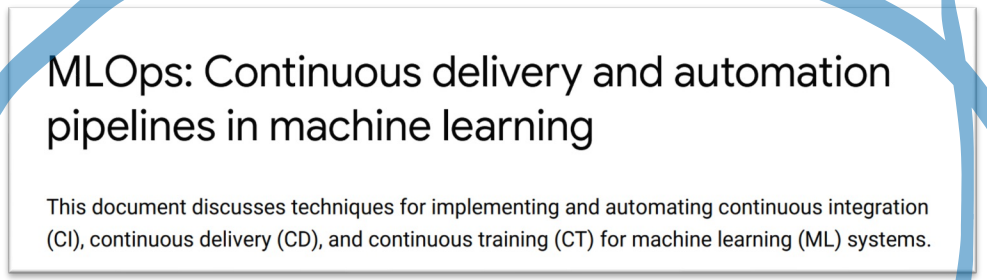
1

Learning under Concept Drift: A Review

Jie Lu, *Fellow, IEEE,* Anjin Liu, *Member, IEEE,* Fan Dong, Feng Gu, João Gama, and Guangquan Zhang

Abstract—Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. Concept drift research involves the development of methodologies and techniques for drift detection, understanding and adaptation. Data analysis has revealed that machine learning in a concept drift environment will result in poor learning results if the drift is not addressed. To help researchers identify which research topics are significant and how to apply related techniques in data analysis tasks, it is necessary that a high quality, instructive review of current research developments and trends in the concept drift field is conducted. In addition, due to the rapid development of concept drift in recent years, the methodologies of learning under concept drift have become noticeably systematic, unveiling a framework which has not been mentioned in literature. This paper reviews over 130 high quality publications in concept drift related research areas, analyzes up-to-date developments in methodologies and techniques, and establishes a framework of learning under concept drift including three main components: concept drift detection, concept drift understanding, and concept drift adaptation. This paper lists and discusses 10 popular synthetic datasets and 14 publicly available benchmark datasets used for evaluating the performance of learning algorithms aiming at handling concept drift. Also, concept drift related research directions are covered and discussed. By providing state-of-the-art knowledge, this survey will directly support researchers in their understanding of research developments in the field of learning under concept drift.

Index Terms—concept drift, change detection, adaptive learning, data streams



MLOps: Continuous delivery and automation pipelines in machine learning

This document discusses techniques for implementing and automating continuous integration (CI), continuous delivery (CD), and continuous training (CT) for machine learning (ML) systems.

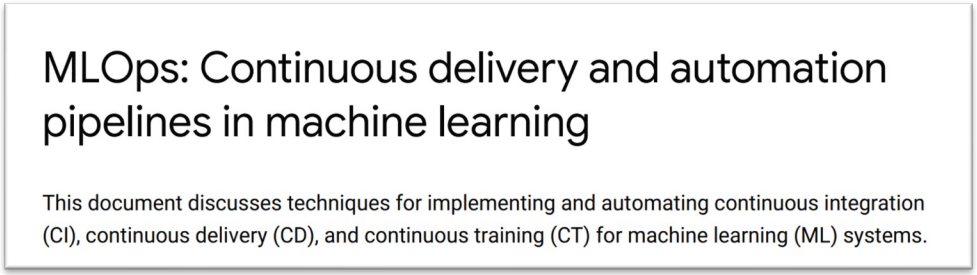
Continuous Delivery for Machine Learning

Automating the end-to-end lifecycle of Machine Learning applications

Study Design

- Rapport de l'industrie (blog).
  - Littérature grise.
  - Axé sur le partage d'expérience.

- Aucune évaluation.
- Aucune publication.
- Plus pratique.



Causé par des facteurs externes.
Imprevisible.
Nécessite une surveillance.

Quelle est la différence entre la dérive de concept et la dérive de la performance du modèle ?

Peut être causé par des facteurs externes (dérive de concept).
Peut également être causé par des facteurs internes.

- Qualité du code (bugs) + qualité des données (valeurs aberrantes) + ..
- Nécessite de bonnes practices (tests + déploiement).

MLOps = DevOps + AI systems

MLOps est une **culture** et **pratique** d'ingénierie ML visant à unifier le **développement de systèmes ML** (Dev) et **l'exploitation de systèmes ML** (Ops).

- Promouvoir l'automatisation et la surveillance à toutes les étapes du système ML

DevOps vs MLOps

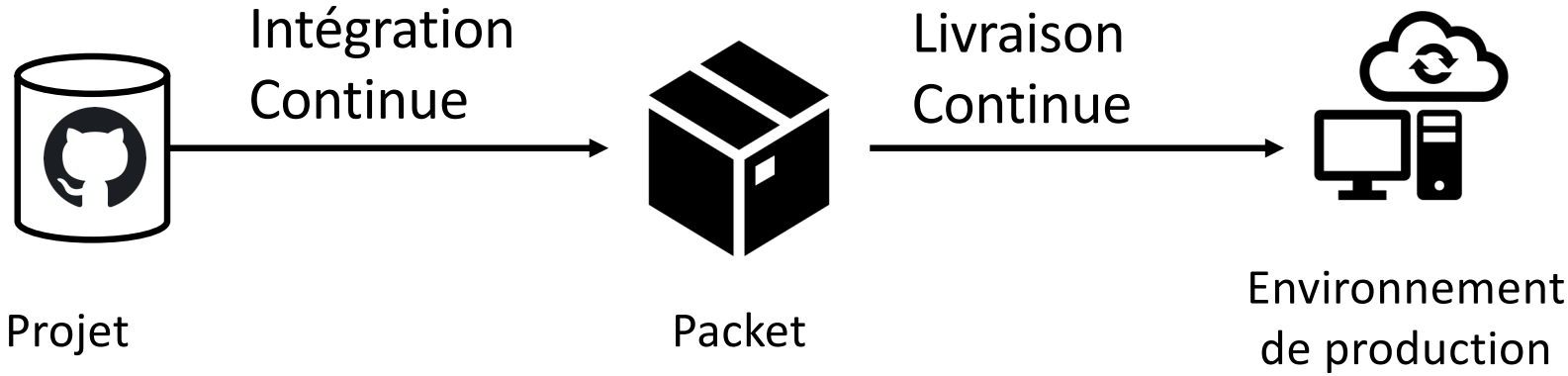
Les pratiques DevOps offrent de nombreux avantages.

- Réduction des cycles de développement.
- Augmentation de la vitesse de développement.
- Augmentation des déploiements fiables

- Deux pratiques principals
  - Intégration Continue (CI)
  - Livraison Continue (CD)

Quelle est la différence ?

Intégration Continue (CI) vs Livraison Continue (CD)



Les systèmes ML sont différents des systèmes traditionnels

Les systems ML ont toute la complexité des systemes traditionnels plus:

- L'équipe de développement est multidisciplinaire.
- Le développement est de nature expérimentale.
  - La reproductibilité est un défi.
- Les tests incluent également la validation des données, les tests de qualité du modèle, etc.
- Le déploiement est plus complexe.
  - Comprend une chaîne pour re-entraîner et déployer automatiquement les modèles.
- *Distinction non précise entre environnements de DEV & de PROD*

Les pratiques de MLOps

- Intégration Continue (CI)
  - Test et validation du code et des composants
  - Test et validation des données + des schémas de données + des modèles
- Déploiement Continu (CD)
  - Déploiement de packages et de services
  - Déploiement du pipeline de maintenance (re-entraînement du modèle)
- Formation continue (CT)
  - Réentraînement automatique des modèles

Propre aux systèmes ML !

Niveaux de maturité des MLOps

- **Niveau 0 : Processus manuel**
  - La chaîne de traitement de ML est entièrement manuelle
  - Aucune pratique de CI/CD - les changements fréquents ne sont pas supposés

- **Niveau 1 : Automatisation du pipeline de traitement de ML**
  - Le réentraînement est automatisé

- **Niveau 2 : Automatisation du pipeline CI/CD**

Niveau 0 : Processus manuel

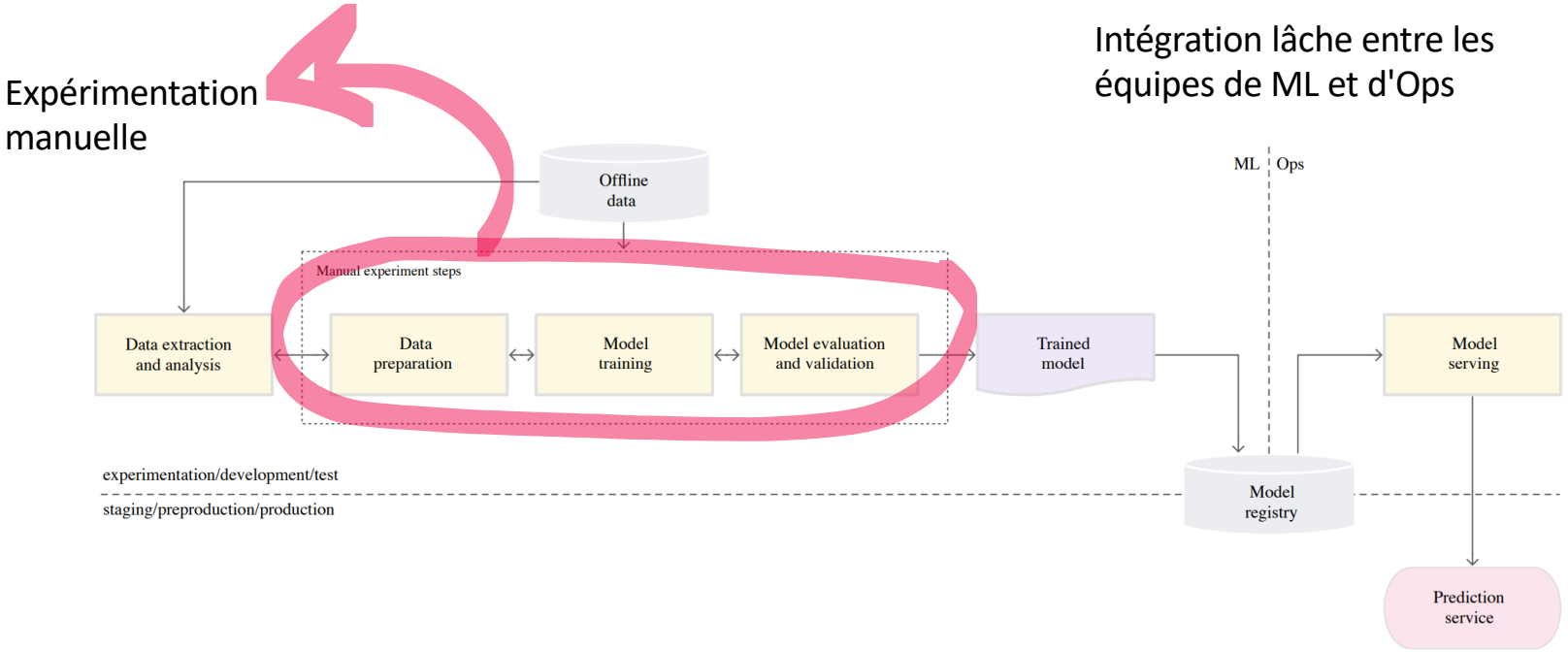
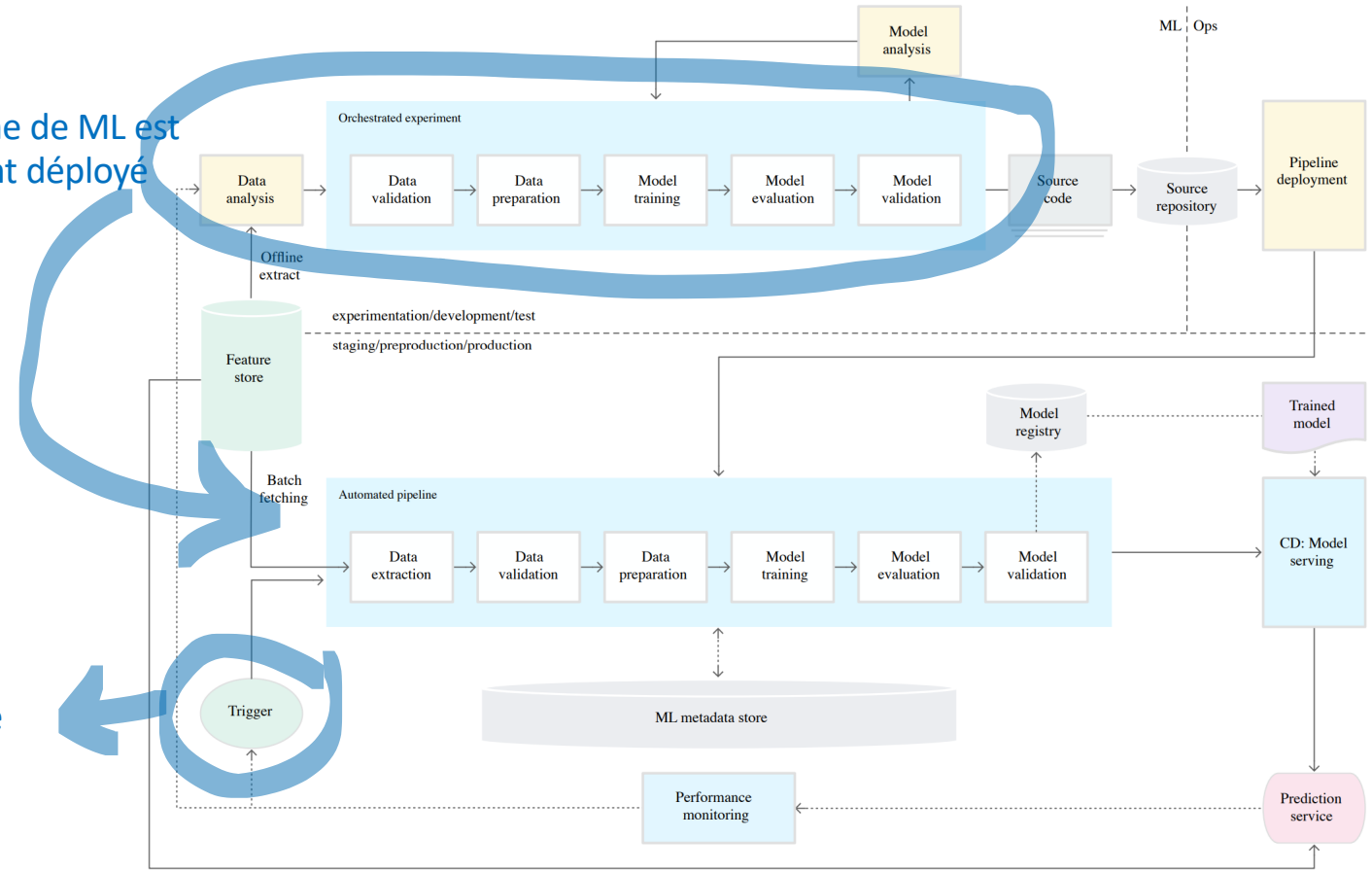


Figure 2. Manual ML steps to serve the model as a prediction service.

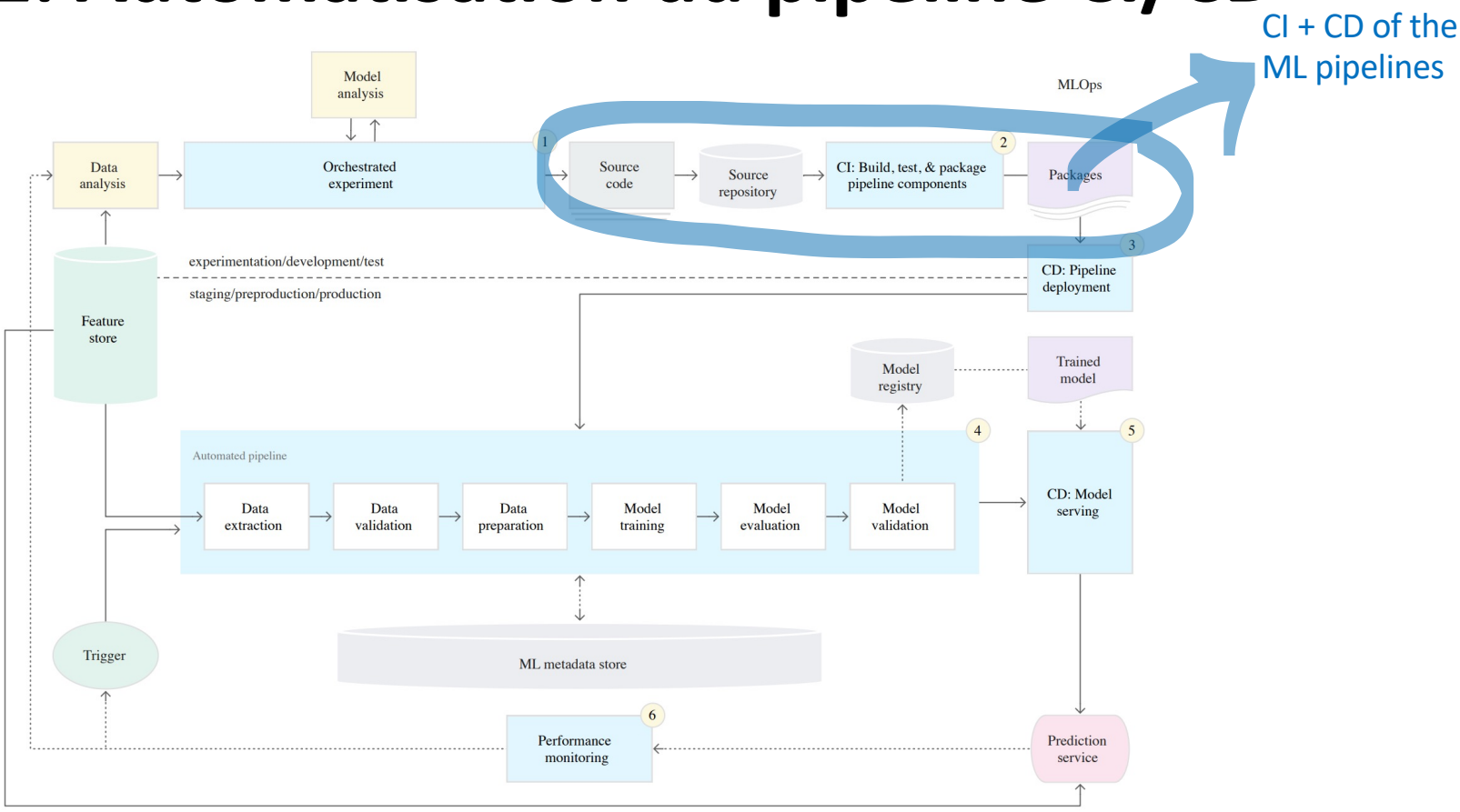
Level 1: Automatisation du pipeline de CT

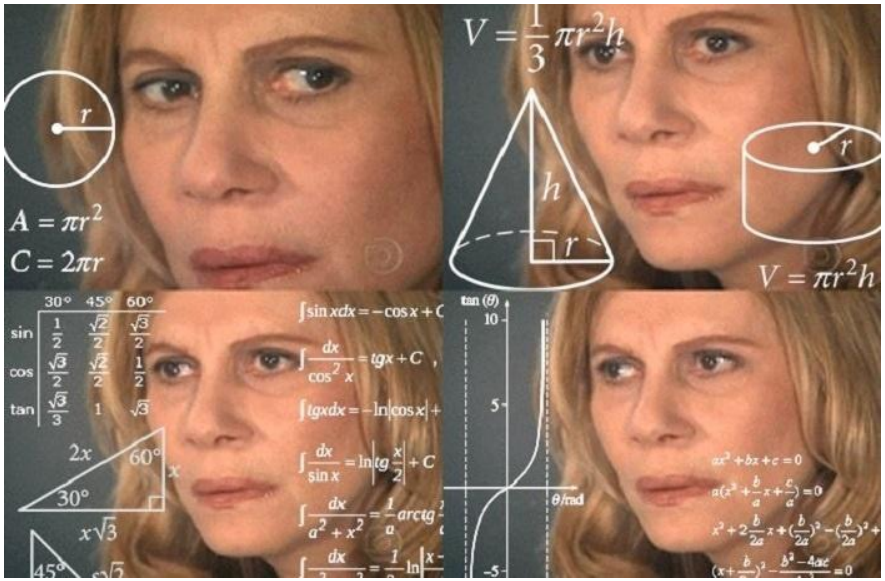
Le pipeline de ML est également déployé

Pipeline est activé



Level 2: Automatisation du pipeline CI/CD





Abstrait ? Plongeons plutôt dans un exemple concret de système ML.

Problèmes et Solutions

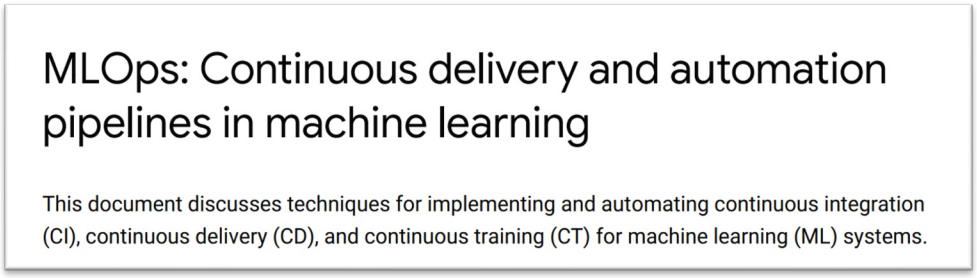
1

Learning under Concept Drift: A Review

Jie Lu, *Fellow, IEEE,* Anjin Liu, *Member, IEEE,* Fan Dong, Feng Gu, João Gama, and Guangquan Zhang

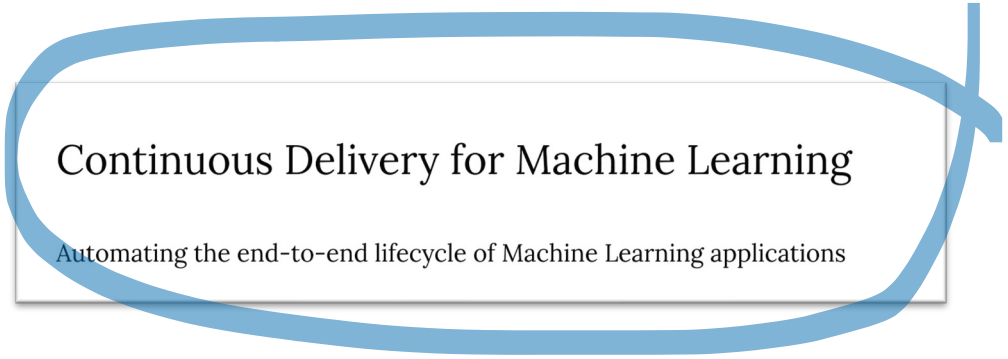
Abstract—Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. Concept drift research involves the development of methodologies and techniques for drift detection, understanding and adaptation. Data analysis has revealed that machine learning in a concept drift environment will result in poor learning results if the drift is not addressed. To help researchers identify which research topics are significant and how to apply related techniques in data analysis tasks, it is necessary that a high quality, instructive review of current research developments and trends in the concept drift field is conducted. In addition, due to the rapid development of concept drift in recent years, the methodologies of learning under concept drift have become noticeably systematic, unveiling a framework which has not been mentioned in literature. This paper reviews over 130 high quality publications in concept drift related research areas, analyzes up-to-date developments in methodologies and techniques, and establishes a framework of learning under concept drift including three main components: concept drift detection, concept drift understanding, and concept drift adaptation. This paper lists and discusses 10 popular synthetic datasets and 14 publicly available benchmark datasets used for evaluating the performance of learning algorithms aiming at handling concept drift. Also, concept drift related research directions are covered and discussed. By providing state-of-the-art knowledge, this survey will directly support researchers in their understanding of research developments in the field of learning under concept drift.

Index Terms—concept drift, change detection, adaptive learning, data streams



MLOps: Continuous delivery and automation pipelines in machine learning

This document discusses techniques for implementing and automating continuous integration (CI), continuous delivery (CD), and continuous training (CT) for machine learning (ML) systems.



Continuous Delivery for Machine Learning

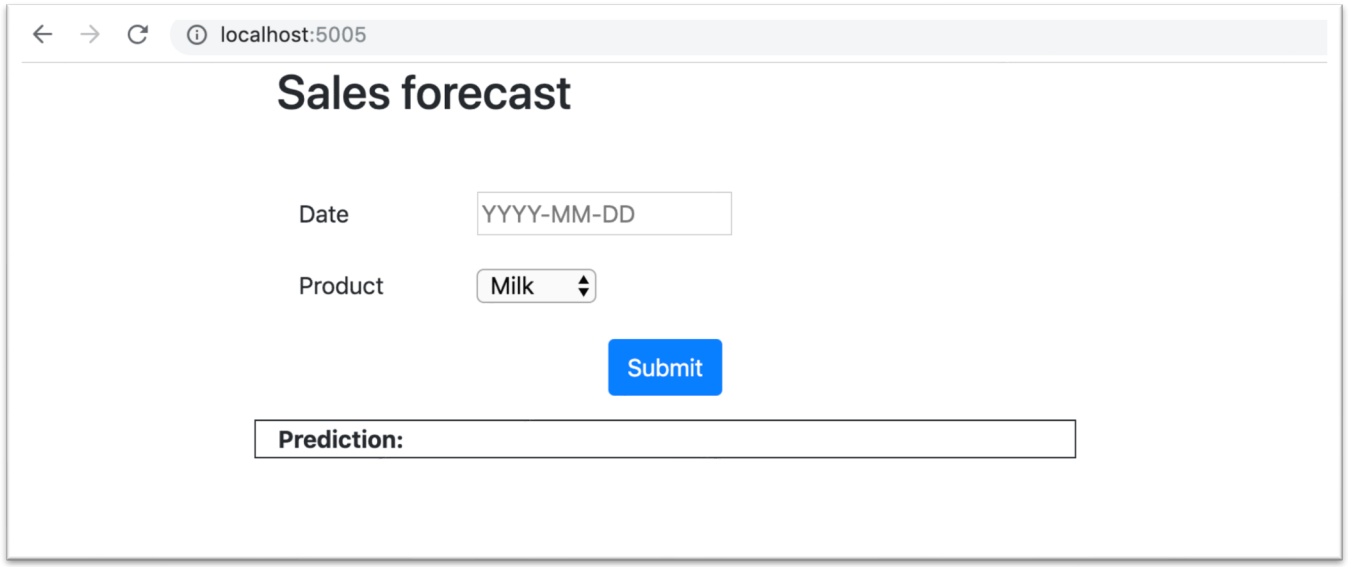
Automating the end-to-end lifecycle of Machine Learning applications

Les changements dans un système ML

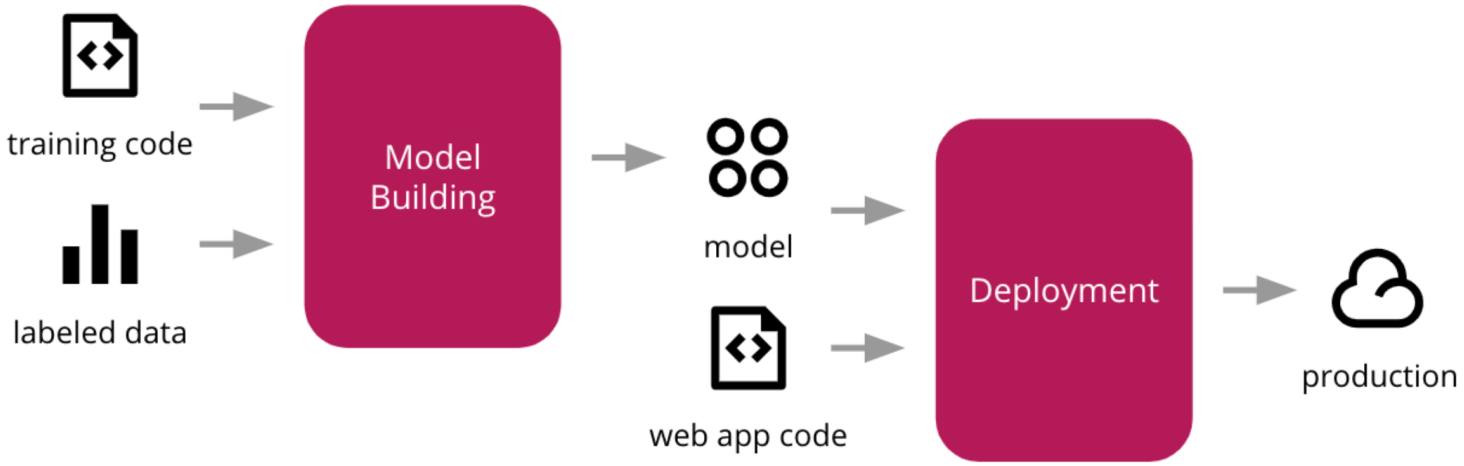
Figure 1: the 3 axis of change in a Machine Learning application — data, model, and code — and a few reasons for them to change

Étude de cas : Sales forecast

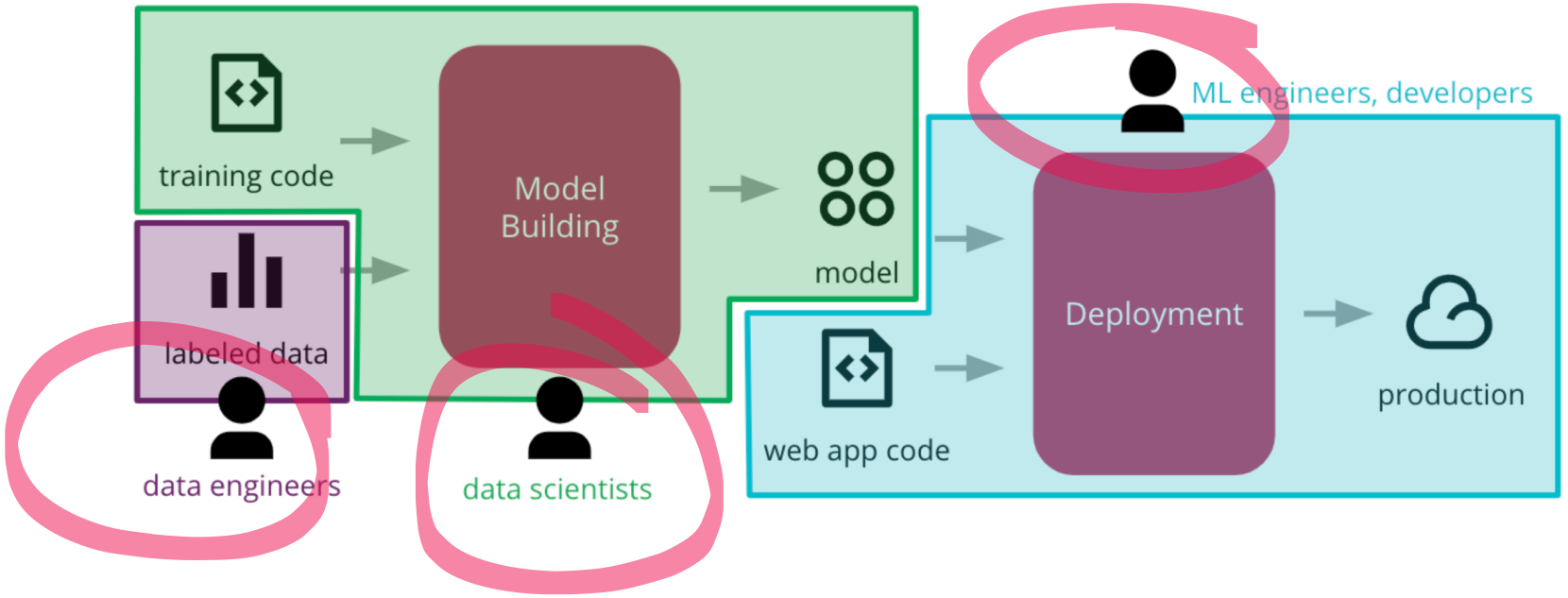
- En fonction d'un produit et d'une date spécifiques
- Le système devrait prédire le nombre d'unités qui seront vendues



Un simple pipeline de ML



Un simple pipeline de ML (cont.)



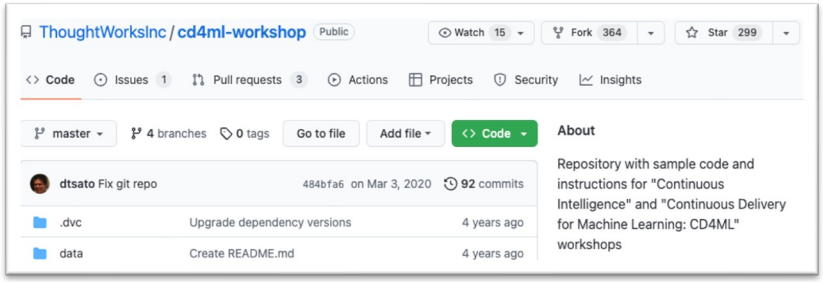
Les défis pour maintenir les prévisions de vente

- **Défi 1** : Comment orchestrer le travail entre différentes équipes ayant des expertises différentes ?
  - Chaque équipe transfère des artefacts aux autres (sans supervision)

- **Défi 2** : Comment rendre le processus vérifiable et reproductible ?

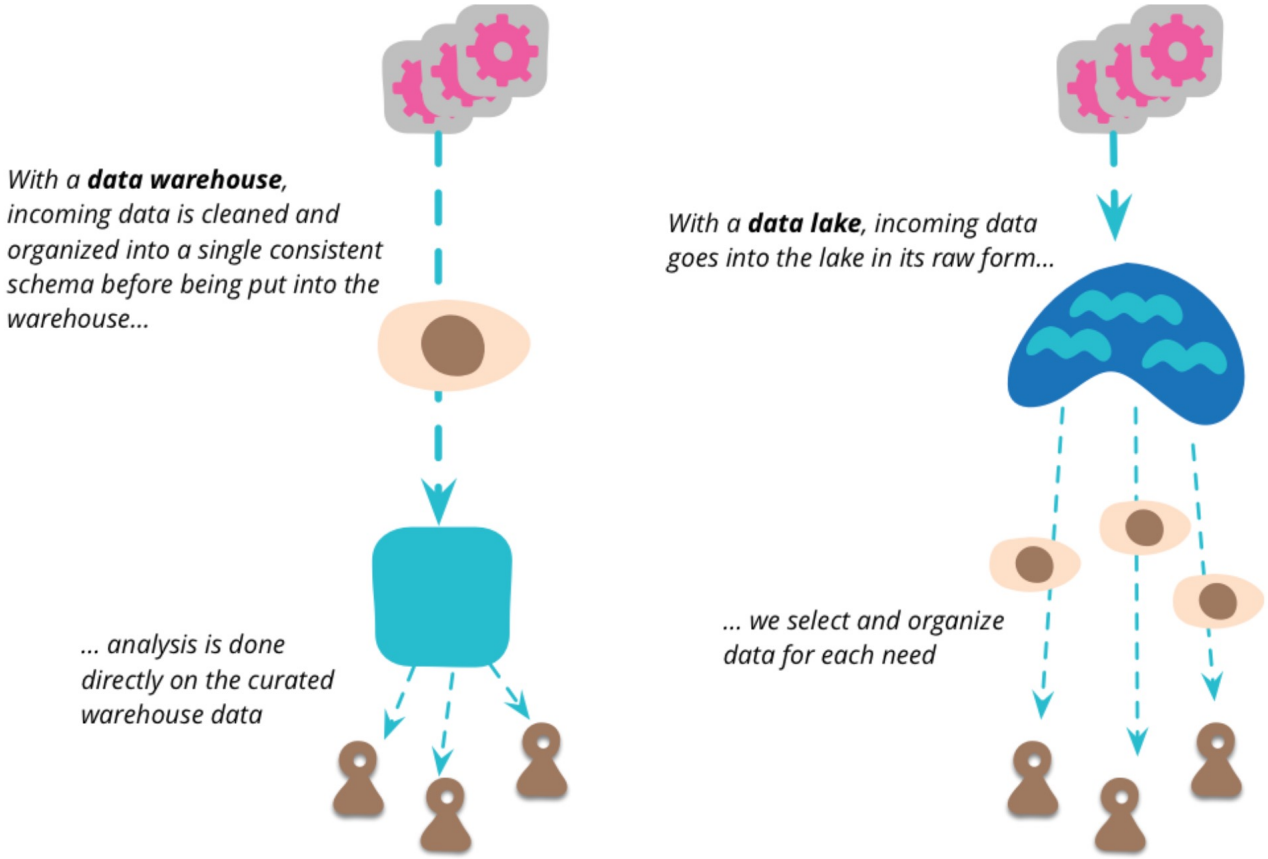
CD4ML: Continuous Deployment for ML Systems

Code accessible CD4ML workshop



- Données découvrables et accessibles
- Entraînement de modèles reproductible
- Déploiement de modèles
- Test et validation
- Suivi des expériences
- Orchestration de la livraison continue
- Surveillance et observabilité des modèles

Données découvrables et accessibles



Entraînement de modèles reproductible

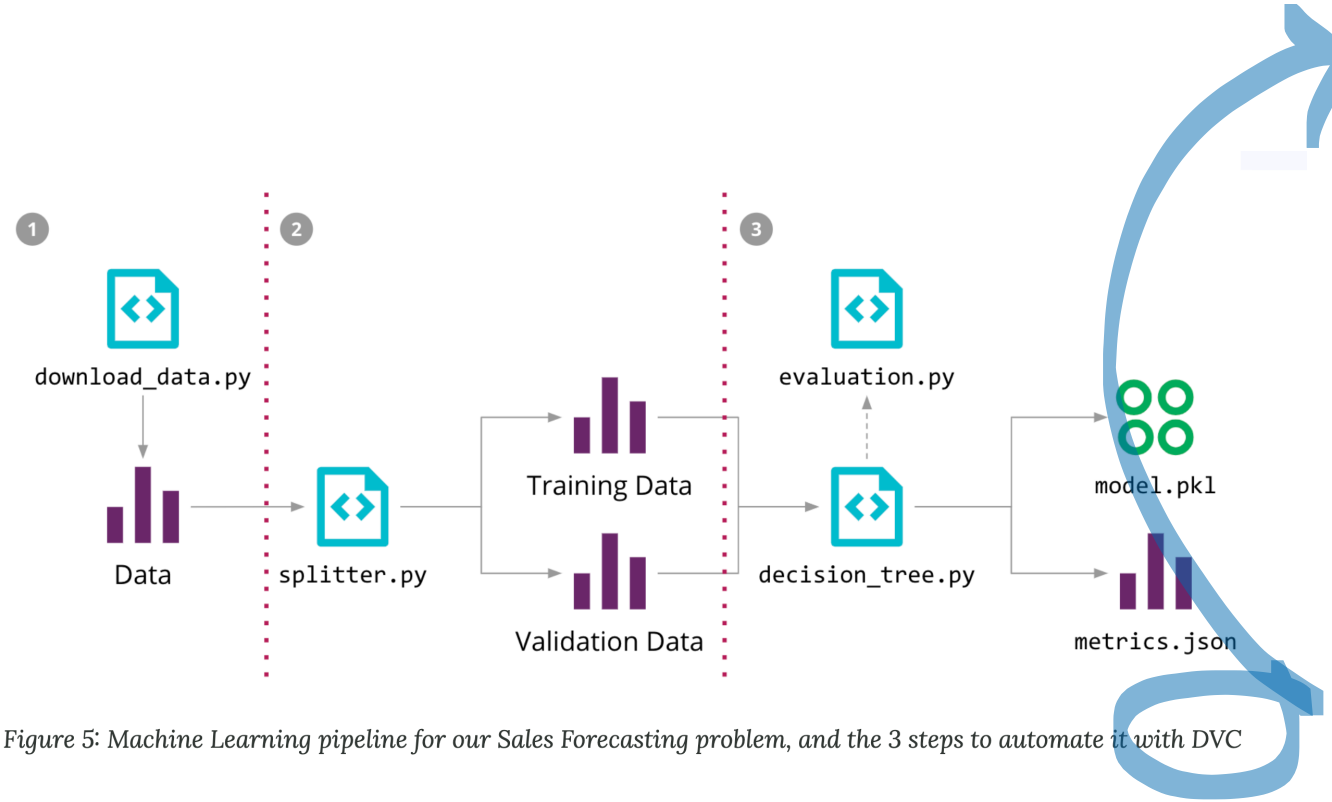


Figure 5: Machine Learning pipeline for our Sales Forecasting problem, and the 3 steps to automate it with DVC

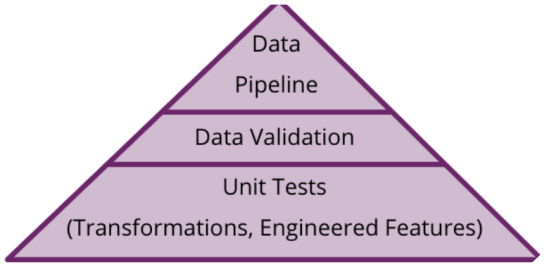
[Data Version Control]

- Outil CLI simple
- Exécuter des parties spécifiques de la chaîne de traitement
- Générer le modèle
- `Dvc repro` exécute l'ensemble de la chaîne de traitement

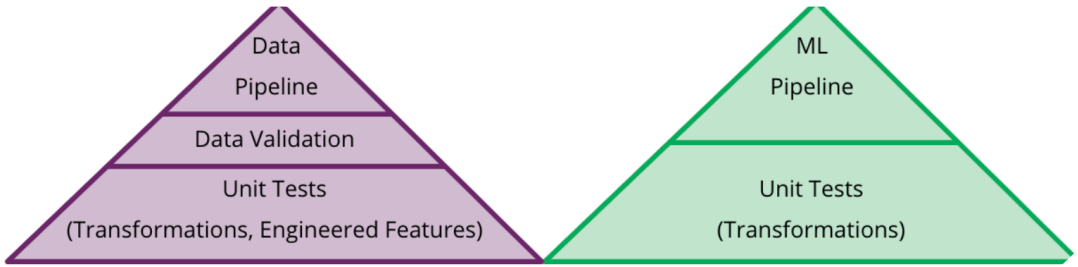
Déploiement de modèles

- Décider comment servir le modèle en production
  - **Modèle intégré**
    - Artéfact du modèle en tant que dépendance de l'application
  - **Modèle déployé en tant que service**
    - Déployé de manière indépendante
  - **Modèle publié en tant que données**
    - Les données sont consommées par l'application cible

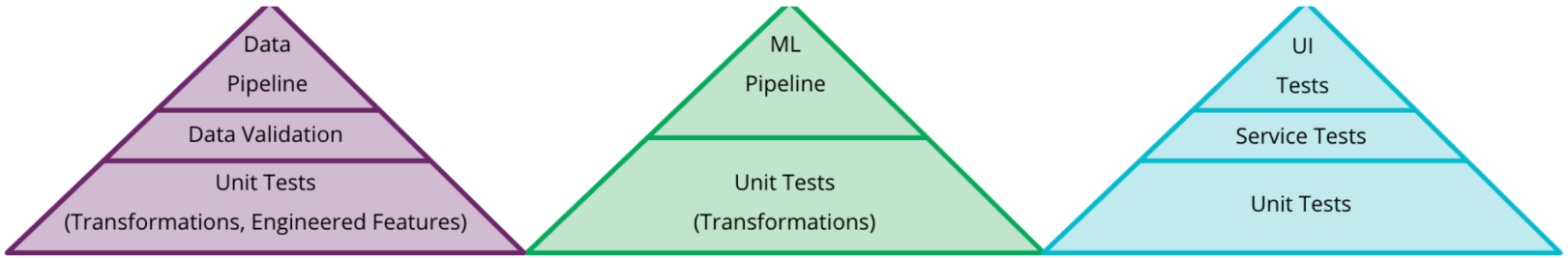
Test et validation



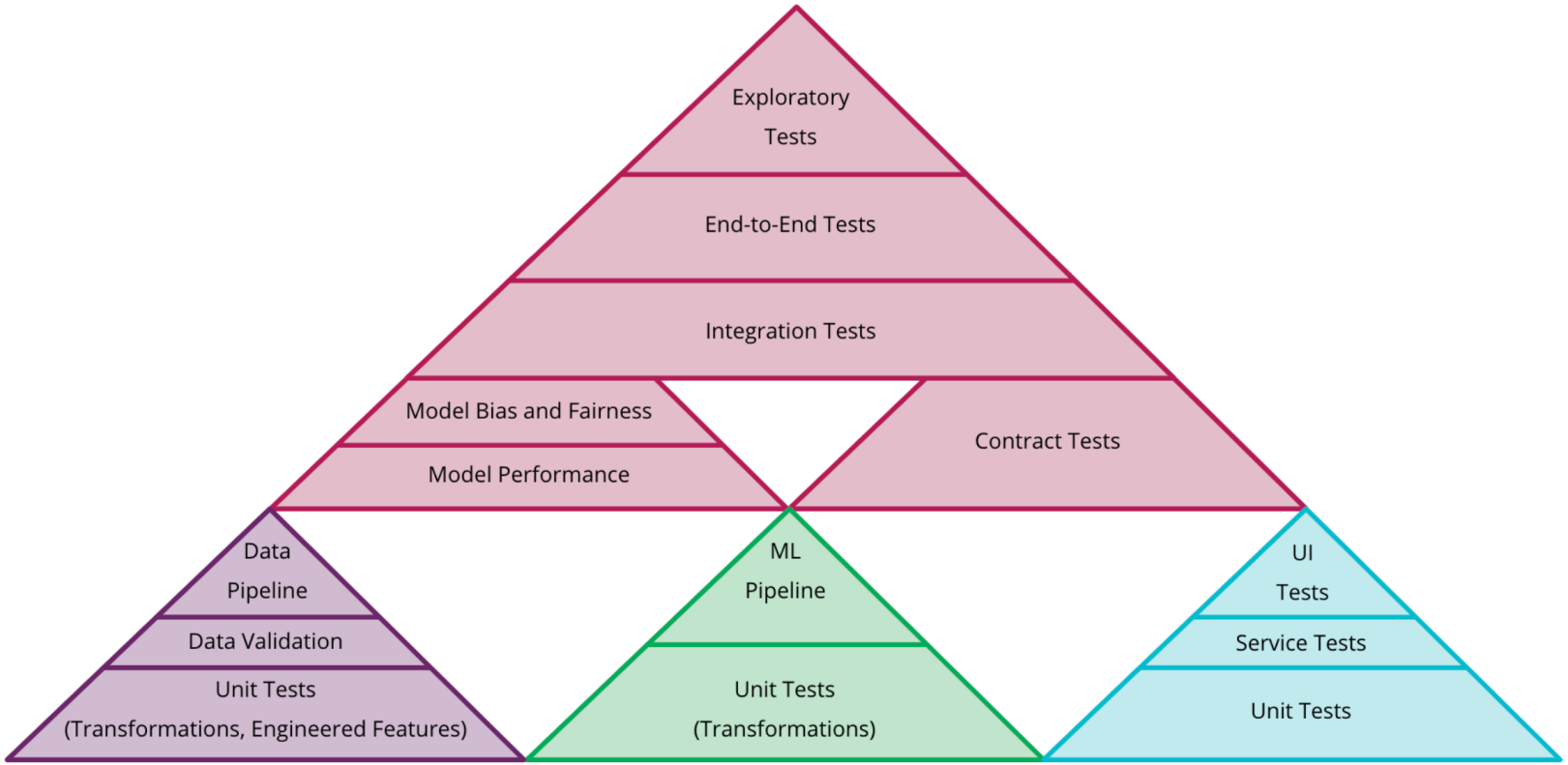
Test et validation



Test et validation

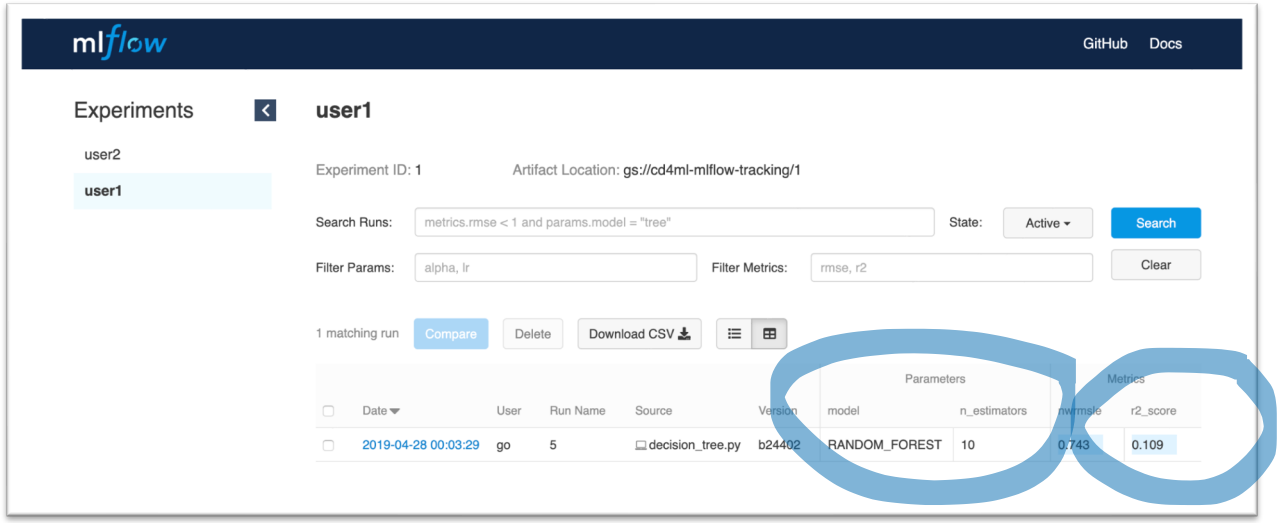


Test et validation



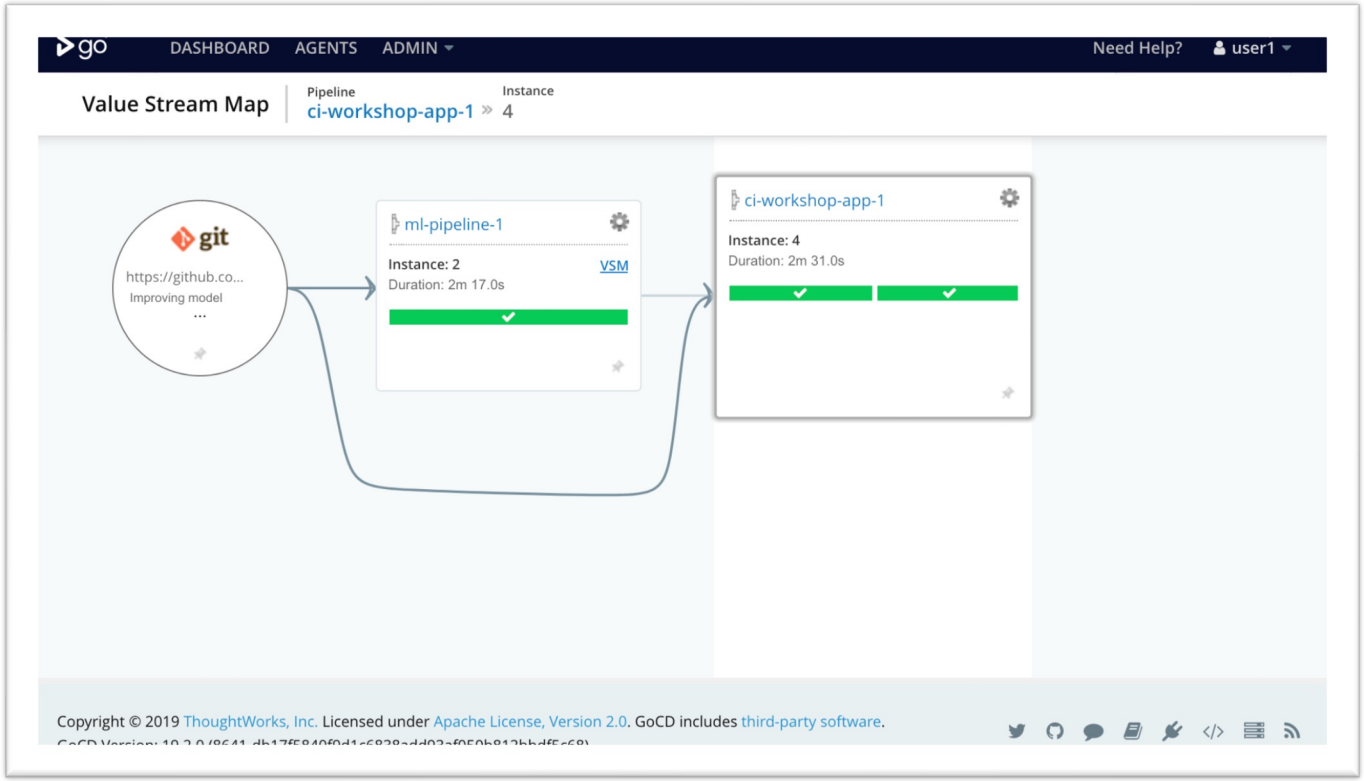
Suivi des expériences

- L'utilisation d'un processus de branche de fonctionnalités génère trop de surcharge
- Mlflow
  - Service pour suivre les performances du modèle à travers les expériences



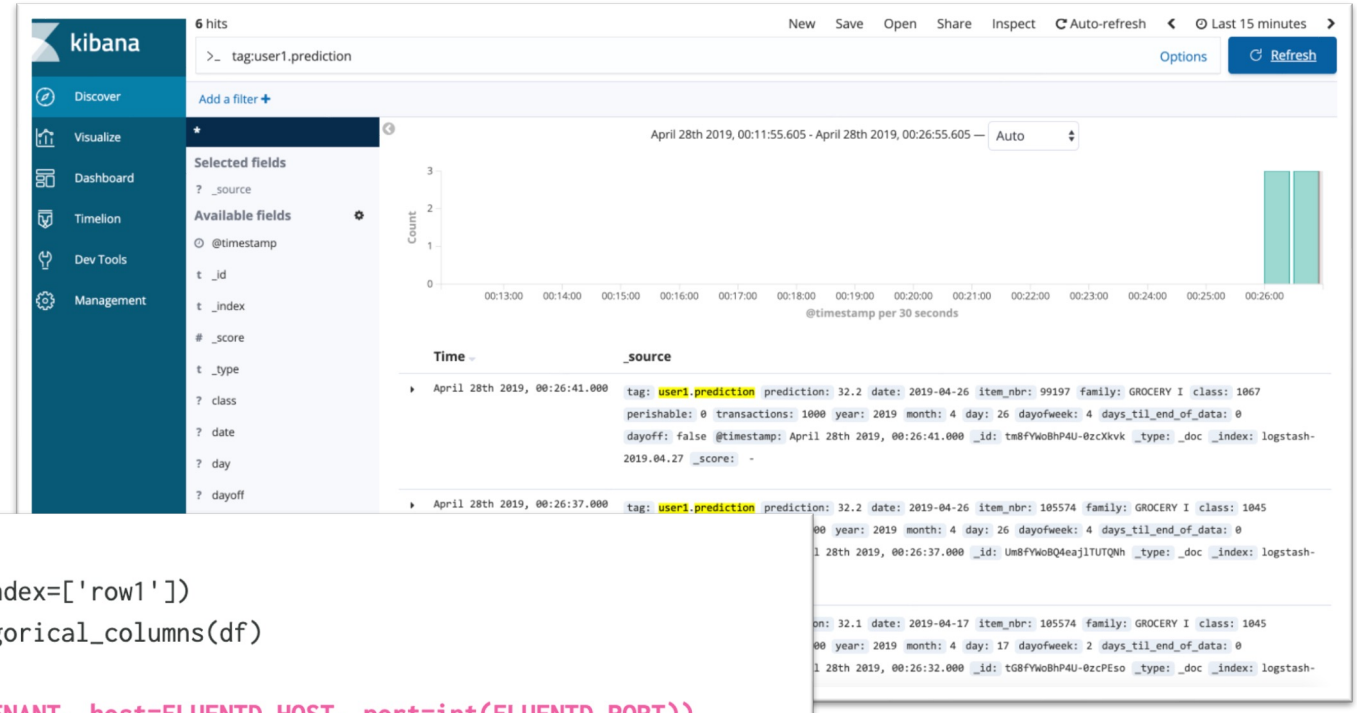
Orchestration de la livraison continue

- Use GoCD tool
  - Open source
  - System workflow
- Utile pour
  - Définir des processus de retour en arrière (rollback)



Surveillance et observabilité des modèles

- EFK stack
  - Elastic search
  - FluentD
  - Kibana



predict_with_logging.py...

```
df = pd.DataFrame(data=data, index=['row1'])
df = decision_tree.encode_categorical_columns(df)
pred = model.predict(df)
logger = sender.FluentSender(TENANT, host=FLUENTD_HOST, port=int(FLUENTD_PORT))
log_payload = {'prediction': pred[0], **data}
logger.emit('prediction', log_payload)
```

Continuous Delivery for Machine Learning