

Overpass Overpass Overpass [ BoldFont = \*-SemiBold ] Overpass [ UprightFont = \*-ExtraLight, BoldFont = \*-Light ] Overpass [ UprightFont = \*-Light, BoldFont = \*-Regular ]

# WEBINAIRE

# REPRODUCTIBILITÉ EN APPRENTISSAGE AUTOMATIQUE

30 OCTOBRE 2020

# Reproductibilité en apprentissage automatique

## └ Introduction

### └ C'est quoi la reproductibilité ?

#### C'EST QUOI LA REPRODUCTIBILITÉ ?

La reproductibilité est le principe qui dit qu'on ne peut tirer de conclusions que d'un algorithme bien décrit, qui est appliqué plusieurs fois, par des personnes différentes.

Par contre, en apprentissage automatique, la reproductibilité correspond (surtout) soit à être capable de reproduire des résultats, soit d'obtenir des résultats similaires en réécrivant un code source [Fieau et al., 2020].

Être capable de **répliquer** les résultats d'un article d'un projet, à partir du même **jeu de données** ou un jeu de données différent (mais proche), en utilisant la **procédure d'entraînement** de l'article ou en utilisant notre procédure d'entraînement et en utilisant le **code** du projet.

les points clés c'est l'idée que si une idée a de bons résultats, on devrait être capable de la reprendre et de retrouver les mêmes résultats.

En ML, les idées c'est des algos et la reproductibilité devient juste d'être capable de s'assurer que les performances rapportées sont les mêmes.

# Reproductibilité en apprentissage automatique

└ Introduction

└ Pourquoi s'y intéresser ?

POURQUOI S'Y INTÉRESSER ?

70 %<sup>1</sup>

---

1. [Baker, 2016]

70 % des chercheurs en science ont échoué dans leur tentative de reproduire un article d'un autre chercheur

# Reproductibilité en apprentissage automatique

└ Introduction

└ Pourquoi s'y intéresser ?

50 %<sup>1</sup>

50 % n'ont pas réussi à reproduire leurs **propres** expérimentations

# Reproductibilité en apprentissage automatique

## └ Introduction

## └ Pourquoi s'y intéresser ?

40 %<sup>2</sup>

L'informatique ne fait pas exception à cela malgré la simplicité (théorique) de réplcation des résultats. Selon une étude, sur **255** articles, près de 40 % n'étaient pas réplcable [Raff, 2019].



La reproductibilité facilite la **réutilisation** pour d'autres projets de recherche, améliorer votre productivité et permet le transfert vers l'industrie (plus facilement). En bonus, cela aide à vous faire connaître si vous faites du contenu réutilisable par la communauté.



Réutilisation



Productivité

La reproductibilité facilite la **réutilisation** pour d'autres projets de recherche, améliorer votre productivité **et** permet le transfert vers l'industrie (plus facilement). En bonus, cela aide à vous faire connaître si vous faites du contenu réutilisable par la communauté.

# Reproductibilité en apprentissage automatique

└ Introduction

└ Motivation

## MOTIVATION



Réutilisation



Productivité



Transfert

La reproductibilité facilite la **réutilisation** pour d'autres projets de recherche, améliorer votre productivité et permet le transfert vers l'industrie (plus facilement). En bonus, cela aide à vous faire connaître si vous faites du contenu réutilisable par la communauté.





Réutilisation



Productivité



Transfert



Se faire connaître

La reproductibilité facilite la **réutilisation** pour d'autres projets de recherche, améliorer votre productivité et permet le transfert vers l'industrie (plus facilement). En bonus, cela aide à vous faire connaître si vous faites du contenu réutilisable par la communauté.

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Version des données & Étapes de prétraitement



Avec quelle version des données doit-on travailler ?

Comment gérer **facilement** plusieurs versions des données ?

Comment définir **facilement** les étapes de prétraitement des données ?

Il nous faut des data pipelines, des tuyaux que nous pouvons **facilement** raccorder à nos modèles pour l'entraînement et la mise en production.

Il faut somewhat voir nos données comme du code et qu'à chaque fois qu'on fait du traitement sur notre data c'est comme modifier notre codebase et c'est impératif d'avoir une trace, par exemple ...

## Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Version des données & Étapes de prétraitement



Version



Gestion des versions

Avec quelle version des données doit-on travailler ?

Comment gérer **facilement** plusieurs versions des données ?

Comment définir **facilement** les étapes de prétraitement des données ?

Il nous faut des data pipelines, des tuyaux que nous pouvons **facilement** raccorder à nos modèles pour l'entraînement et la mise en production.

Il faut somewhat voir nos données comme du code et qu'à chaque fois qu'on fait du traitement sur notre data c'est comme modifier notre codebase et c'est impératif d'avoir une trace, par exemple ...

# Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Version des données & Étapes de prétraitement



Version



Gestion des versions



Étapes prétraitement

Avec quelle version des données doit-on travailler ?

Comment gérer **facilement** plusieurs versions des données ?

Comment définir **facilement** les étapes de prétraitement des données ?

Il nous faut des data pipelines, des tuyaux que nous pouvons **facilement** raccorder à nos modèles pour l'entraînement et la mise en production.

Il faut somewhat voir nos données comme du code et qu'à chaque fois qu'on fait du traitement sur notre data c'est comme modifier notre codebase et c'est impératif d'avoir une trace, par exemple ...

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Code

CODE



Version

Avec quelle version du code doit-on travailler ?

Comment savoir **rapidement** qu'elle est la différence d'implémentation entre deux versions du modèle ?

Comment gérer **facilement** les divergences d'expérimentations ?

Il nous faut un outil nous permettant de **visualiser** la différence entre des fichiers de code et nous permettant d'avoir **plusieurs** versions du code en même temps, par exemple ...

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Code

CODE



Version



Différence

Avec quelle version du code doit-on travailler ?

Comment savoir **rapidement** qu'elle est la différence d'implémentation entre deux versions du modèle ?

Comment gérer **facilement** les divergences d'expérimentations ?

Il nous faut un outil nous permettant de **visualiser** la différence entre des fichiers de code et nous permettant d'avoir **plusieurs** versions du code en même temps, par exemple ...

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Code

CODE



Version



Différence



Divergences

Avec quelle version du code doit-on travailler ?

Comment savoir **rapidement** qu'elle est la différence d'implémentation entre deux versions du modèle ?

Comment gérer **facilement** les divergences d'expérimentations ?

Il nous faut un outil nous permettant de **visualiser** la différence entre des fichiers de code et nous permettant d'avoir **plusieurs** versions du code en même temps, par exemple ...

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Dépendances



Avec quelle version de cette librairie doit-on travailler ?

Comment gérer les conflits entre des versions ?

Comment installer facilement et rapidement des dépendances ?

Il nous faut un outil nous permettant de **définir** les dépendances, capable de gérer les conflits entre des versions et nous permettant d'installer facilement des dépendances.



# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Dépendances

## DÉPENDANCES



Version



Conflicts

Avec quelle version de cette librairie doit-on travailler ?

Comment gérer les conflits entre des versions ?

Comment installer facilement et rapidement des dépendances ?

Il nous faut un outil nous permettant de **définir** les dépendances, capable de gérer les conflits entre des versions et nous permettant d'installer facilement des dépendances.

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Dépendances

## DÉPENDANCES



Version



Conflits



Installation

Avec quelle version de cette librairie doit-on travailler ?

Comment gérer les conflits entre des versions ?

Comment installer facilement et rapidement des dépendances ?

Il nous faut un outil nous permettant de **définir** les dépendances, capable de gérer les conflits entre des versions et nous permettant d'installer facilement des dépendances.

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Environnement



Comment s'assurer que nos environnements de développement ne rentre pas en conflit entre eux ?

Comment naviguer entre deux environnements ?

Comment gérer plusieurs version des dépendances ?

# Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Environnement

## ENVIRONNEMENT



Comment s'assurer que nos environnements de développement ne rentre pas en conflit entre eux ?

Comment naviguer entre deux environnements ?

Comment gérer plusieurs version des dépendances ?

# Reproductibilité en apprentissage automatique

└─ OK, mais comment ?

└─ Environnement

## ENVIRONNEMENT



Différents environnements



Changer



Versions des dépendances

Comment s'assurer que nos environnements de développement ne rentrent pas en conflit entre eux ?

Comment naviguer entre deux environnements ?

Comment gérer plusieurs versions des dépendances ?

## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Développement des modèles



Réinventer

Ne pas réinventer la roue.

Simplifier l'écriture de code pour développer des modèles.

Qui facilite l'entraînement (GPU, multi-GPU/CPU).

Il nous faut des outils nous permettant de **simplifier** le **développement** de nos modèles, par exemple ...

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Développement

└ Développement des modèles



Réinventer



Simplification

Ne pas réinventer la roue.

Simplifier l'écriture de code pour développer des modèles.

Qui facilite l'entraînement (GPU, multi-GPU/CPU).

Il nous faut des outils nous permettant de **simplifier** le **développement** de nos modèles, par exemple ...

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Développement

└ Développement des modèles



Réinventer



Simplification



Facile

Ne pas réinventer la roue.

Simplifier l'écriture de code pour développer des modèles.

Qui facilite l'entraînement (GPU, multi-GPU/CPU).

Il nous faut des outils nous permettant de **simplifier** le **développement** de nos modèles, par exemple ...



## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Entraînement, configuration et résultats



Avec quelle version du code, du modèle et des données avons-nous fait cet entraînement ?

Quels sont les résultats ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de logger les **paramètres** d'entraînement et les **résultats**, par exemple, ...

## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?
  - └ Développement
    - └ Entraînement, configuration et résultats

Version de  
l'entraînement

Résultats

Avec quelle version du code, du modèle et des données avons-nous fait cet entraînement ?

Quels sont les résultats ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de logger les **paramètres** d'entraînement et les **résultats**, par exemple, ...

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Développement

└ Entraînement, configuration et résultats



Version de  
l'entraînement



Résultats



Visualisation

Avec quelle version du code, du modèle et des données avons-nous fait cet entraînement ?

Quels sont les résultats ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de logger les **paramètres** d'entraînement et les **résultats**, par exemple, ...

# Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Développement

└ Entraînement, configuration et résultats



Version de  
l'entraînement



Résultats



Visualisation



Erreurs d'entraînement

Avec quelle version du code, du modèle et des données avons-nous fait cet entraînement ?

Quels sont les résultats ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de logger les **paramètres** d'entraînement et les **résultats**, par exemple, ...

## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Rapport et analyse des résultats



Tableau des résultats

Comment créer des tableaux de résultats **facilement** (pas à la *mitaine*) ?

Comment s'assurer **facilement** que les résultats sont à jours ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de **créer** des tableaux de résultats **à même les résultats**, soit de diminuer le plus possible le travail manuel, par exemple ...

## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Rapport et analyse des résultats



Tableau des résultats



Mise à jour

Comment créer des tableaux de résultats **facilement** (pas à la *mitaine*) ?

Comment s'assurer **facilement** que les résultats sont à jours ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de **créer** des tableaux de résultats **à même les résultats**, soit de diminuer le plus possible le travail manuel, par exemple ...

## Reproductibilité en apprentissage automatique

└ OK, mais comment ?

└ Développement

└ Rapport et analyse des résultats



Tableau des résultats



Mise à jour



Visualisation configuration

Comment créer des tableaux de résultats **facilement** (pas à la *mitaine*) ?

Comment s'assurer **facilement** que les résultats sont à jours ?

Comment visualiser **rapidement** les résultats et les paramètres de configuration ?

Il nous faut des outils nous permettant de **créer** des tableaux de résultats **à même les résultats**, soit de diminuer le plus possible le travail manuel, par exemple ...

## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Gestion des environnements



Comment s'assurer que nos modèles fonctionnent sur d'autres environnements ?

Comment faciliter la réutilisation de notre code ?



## Reproductibilité en apprentissage automatique

- └ OK, mais comment ?

- └ Développement

- └ Gestion des environnements



Différents environnements



Réutilisation

Comment s'assurer que nos modèles fonctionnent sur d'autres environnements ?

Comment faciliter la réutilisation de notre code ?



Développer des processus rigoureux (par essais, erreurs et journaux) et ne pas prendre tout ce qui a été discuté ici comme l'unique solution.

WEBINAIRE

**MERCI DE VOTRE  
ÉCOUTE !**