

WEBINAIRE

REPRODUCTIBILITÉ EN APPRENTISSAGE AUTOMATIQUE

30 OCTOBRE 2020

Reproductibilité en apprentissage automatique

2020-10-29

OBJECTIFS DE LA PRÉSENTATION

- Inciter l'intégration des solutions permettant une meilleure reproductibilité dans vos solutions d'affaires et académiques.
- Améliorer la reproductibilité de vos projets.
- Améliorer votre productivité.

2020-10-29

Objectifs de la présentation

- Inciter l'intégration des solutions permettant une meilleure reproductibilité dans vos solutions d'affaires et académiques.
- Améliorer la reproductibilité de vos projets.
- Améliorer votre productivité.

VOTRE CONFÉRENCIER



- Introduit à la recherche reproductible en 2016
(R Markdown et [git](#))
- Participation à REPROLANG de la conférence LREC
[Garneau et al., 2020]
- Membre actif dans le développement d'une librairie
facilitant la reproductibilité ([Poutyne](#))

Reproductibilité en apprentissage automatique

2020-10-29

└ Votre conférencier

VOTRE CONFÉRENCIER



- Introduit à la recherche reproductible en 2016
(R Markdown et [git](#))
- Participation à REPROLANG de la conférence LREC
[Garneau et al., 2020]
- Membre actif dans le développement d'une librairie
facilitant la reproductibilité ([Poutyne](#))

DAVID BEAUCHEMIN
Candidat au doctorat
Département d'informa-
tique et de génie logiciel

AU MENU

...

Reproductibilité en apprentissage automatique

AU MENU



Gestion version



Productivité



Présenter



Réutiliser

└ Au menu

2020-10-29

Introduction

C'EST QUOI LA REPRODUCTIBILITÉ?

Pineau et al., 2020

Reproductibilité en apprentissage automatique

└ Introduction

└ C'est quoi la reproductibilité?

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.

2020-10-29

C'EST QUOI LA REPRODUCTIBILITÉ?

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

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

POURQUOI S'Y INTÉRESSER?

1. [Baker, 2016]

1

2020-10-29

Reproductibilité en apprentissage automatique

└ Introduction

└ Pourquoi s'y intéresser?

70 %¹

1. [Baker, 2016]

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

POURQUOI S'Y INTÉRESSER?

1. [Baker, 2016]

1

2020-10-29

Reproductibilité en apprentissage automatique

└ Introduction

└ Pourquoi s'y intéresser?

50 %¹

1. [Baker, 2016]

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

POURQUOI S'Y INTÉRESSER?

2. [Raff, 2019]

2

POURQUOI S'Y INTÉRESSER?

Reproductibilité en apprentissage automatique

- └ Introduction
- └ Pourquoi s'y intéresser?

40 %²

2. [Raff, 2019]

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

MOTIVATION

Reproductibilité en apprentissage automatique

└ Introduction

└ Motivation

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.

2020-10-29

MOTIVATION

Réutilisation



MOTIVATION

Reproductibilité en apprentissage automatique

└ Introduction

└ Motivation

2020-10-29

MOTIVATION



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.

MOTIVATION

Reproductibilité en apprentissage automatique

└ Introduction

└ Motivation

2020-10-29

MOTIVATION



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.

MOTIVATION

Reproductibilité en apprentissage automatique

└ Introduction

└ Motivation

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.

2020-10-29

MOTIVATION



Réutilisation

Productivité

Transfert

Se faire connaître

Les barrières à la reproductibilité

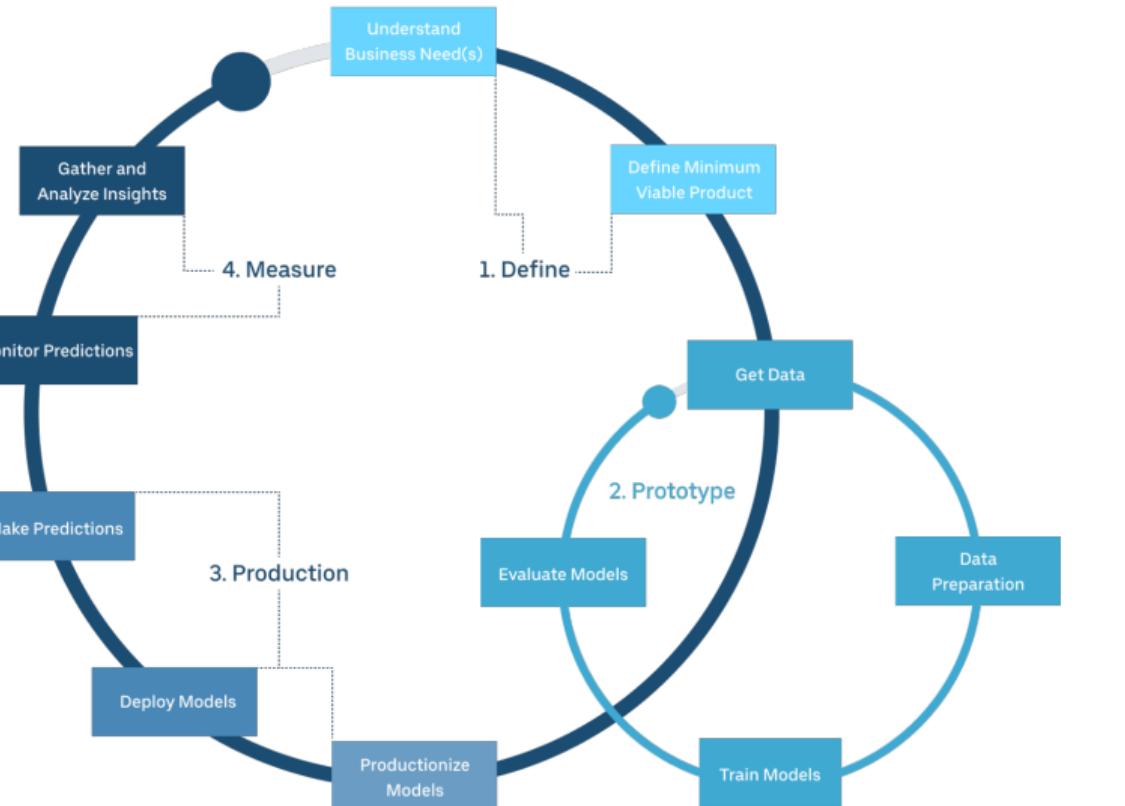
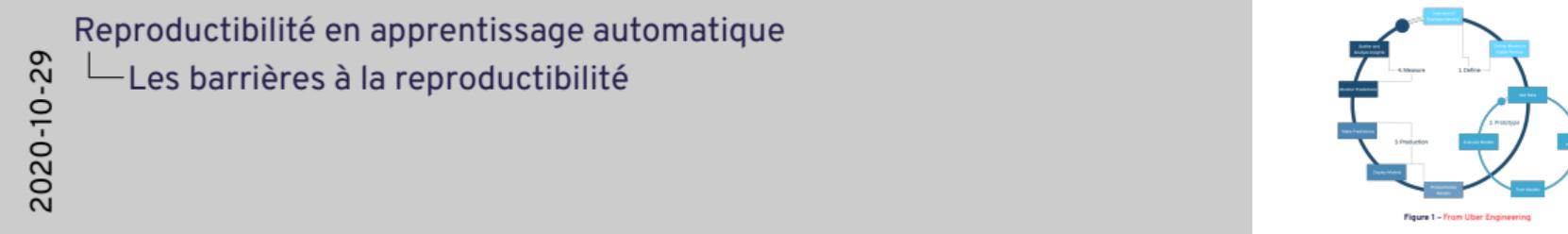


Figure 1 – From Uber Engineering



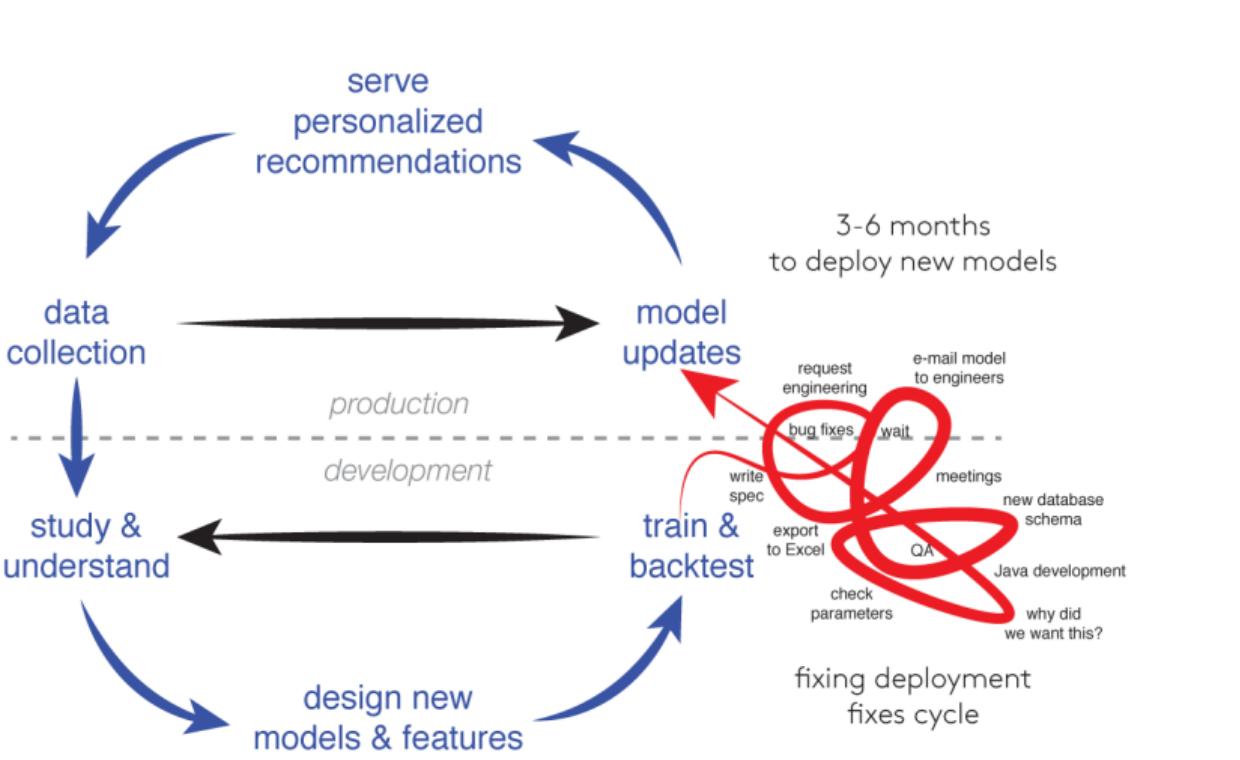
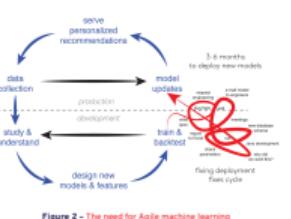


Figure 2 – The need for Agile machine learning

Reproductibilité en apprentissage automatique

- └ Les barrières à la reproductibilité

2020-10-29



OK, mais comment?

Reproductibilité en apprentissage automatique
└ OK, mais comment?

OK, mais comment?

2020-10-29

AU MENU



Productivité



Présenter



Réutiliser

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?

└ Au menu

AU MENU



VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

1

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 ...

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

Version

2020-10-29

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

1

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 ...

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

2020-10-29



Version

Gestion des versions

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

1

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 ...

2020-10-29

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT



Version

Gestion des versions

Étapes prétraitement

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT



Data Version Control



Dask

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?

└ Version des données & Étapes de prétraitement

VERSION DES DONNÉES & ÉTAPES DE PRÉTRAITEMENT

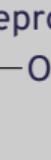


Data Version Control



Dask

CODE



CODE

...

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Code

2020-10-29

CODE



Version



Difference

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 ...

CODE

Version Différence Divergences

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Code

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 ...

2020-10-29

CODE



CODE

Git

GitHub

GitLab

Bitbucket

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?

└ Code

CODE



Git



Github



GitLab



Bitbucket

AU MENU



Gestion version



Présenter



Réutiliser

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?

└ Au menu

AU MENU



Gestion version



Productivité



Présenter



Réutiliser

DÉVELOPPEMENT DES MODÈLES



Reproductibilité en apprentissage automatique

└ OK, mais comment?

 └ Développement

 └ Développement des modèles

2020-10-29

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

...

DÉVELOPPEMENT DES MODÈLES



Réinventer

DÉVELOPPEMENT DES MODÈLES



Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Développement des modèles

2020-10-29

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
...

DÉVELOPPEMENT DES MODÈLES



DÉVELOPPEMENT DES MODÈLES



Reproductibilité en apprentissage automatique

└ OK, mais comment?

 └ Développement

 └ Développement des modèles

2020-10-29

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

...

DÉVELOPPEMENT DES MODÈLES



DÉVELOPPEMENT DES MODÈLES



Poutyne



PyTorch
Lightning



Scikit-learn



Gensim

AllenNLP

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Développement des modèles

DÉVELOPPEMENT DES MODÈLES



ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Entraînement, configuration et résultats



2020-10-29

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, ...

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

1

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Entraînement, configuration et résultats

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, ...

2020-10-29

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

1

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Entraînement, configuration et résultats

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS



Version de l'entraînement



Résultats



Visualisation

2020-10-29

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, ...

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

1

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Entraînement, configuration et résultats

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS



Version de l'entraînement



Résultats



Visualisation



Erreurs d'entraînement

2020-10-29

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, ...

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

mlflow

HYDRA

MLflow

Hydra

Sacred

Notif

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Entraînement, configuration et résultats

ENTRAÎNEMENT, CONFIGURATION ET RÉSULTATS

mlflow

HYDRA

Sacred

Notif

AU MENU



Gestion version



Productivité



Réutiliser

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Au menu

AU MENU



Gestion version Productivité Présenter Réutiliser

RAPPORT ET ANALYSE DES RÉSULTATS

1

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Rapport et analyse des résultats

RAPPORT ET ANALYSE DES RÉSULTATS



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

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

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 ...

2020-10-29



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

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

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 ...

RAPPORT ET ANALYSE DES RÉSULTATS

1

Reproductibilité en apprentissage automatique

└ OK, mais comment?

└ Développement

└ Rapport et analyse des résultats

RAPPORT ET ANALYSE DES RÉSULTATS



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

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

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 ...

2020-10-29

RAPPORT ET ANALYSE DES RÉSULTATS



Python2LaTeX

TensorBoard



Jupyter
notebook



Markdown



Dash

-
1. *I don't like notebooks - Joel Grus*
 2. *New York Oil and Gas*

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Rapport et analyse des résultats

RAPPORT ET ANALYSE DES RÉSULTATS



AU MENU



Gestion version



Productivité



Présenter

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Au menu

AU MENU



Gestion version Productivité Présenter Réutiliser

- └ OK, mais comment?
 - └ Développement
 - └ Environnement

2020-10-29

Comment s'assurer que nos modèles fonctionnent sur d'autres environnements?
Comment faciliter la réutilisation de notre code?

- └ OK, mais comment?
 - └ Développement
 - └ Environnement

2020-10-29

Comment s'assurer que nos modèles fonctionnent sur d'autres environnements?
Comment faciliter la réutilisation de notre code?

ENVIRONNEMENT



Docker



Kubernetes

2020-10-29

Reproductibilité en apprentissage automatique
└ OK, mais comment?
 └ Développement
 └ Environnement

ENVIRONNEMENT



Docker



Kubernetes

La suite



2020-10-29

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

POUR ALLER PLUS LOIN

- Clean code
- Continuous Machine Learning
- Faire des tests!
- Writing Code for NLP Research [Gardner et al., 2018]
- SOLID
- *Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program [Pineau et al., 2020])*

2020-10-29

Reproductibilité en apprentissage automatique

La suite

Pour aller plus loin

POUR ALLER PLUS LOIN

- Clean code
- Continuous Machine Learning
- Faire des tests!
- Writing Code for NLP Research [Gardner et al., 2018]
- SOLID
- *Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program [Pineau et al., 2020])*



PÉRIODE DE QUESTIONS

Reproductibilité en apprentissage automatique

└ La suite

└ Période de questions

2020-10-29

PÉRIODE DE QUESTIONS

WEBINAIRE

MERCI DE VOTRE ÉCOUTE !

REFERENCES i

-  Baker, M. (2016).
1,500 Scientists Lift the Lid on Reproducibility.
Nature News, 533(7604) :452.
-  Gardner, M., Neumann, M., Grus, J., and Lourie, N. (2018).
Writing Code for NLP Research.
In *Conference on Empirical Methods in Natural Language Processing : Tutorial Abstracts*.
-  Garneau, N., Godbout, M., Beauchemin, D., Durand, A., and Lamontagne, L. (2020).
A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings : Making the Method Robustly Reproducible as Well.

Reproductibilité en apprentissage automatique

└ La suite

└ References

2020-10-29

REFERENCES i

-  Baker, M. (2016).
1,500 Scientists Lift the Lid on Reproducibility.
Nature News, 533(7604) :452.
-  Gardner, M., Neumann, M., Grus, J., and Lourie, N. (2018).
Writing Code for NLP Research.
In *Conference on Empirical Methods in Natural Language Processing : Tutorial Abstracts*.
-  Garneau, N., Godbout, M., Beauchemin, D., Durand, A., and Lamontagne, L. (2020).
A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings : Making the Method Robustly Reproducible as Well.

REFERENCES ii

-  Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2020).
Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program).
-  Raff, E. (2019).
A Step Toward Quantifying Independently Reproducible Machine Learning Research.

Reproductibilité en apprentissage automatique

La suite

References

2020-10-29

REFERENCES ii

-  Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2020).
Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program).
-  Raff, E. (2019).
A Step Toward Quantifying Independently Reproducible Machine Learning Research.