19 septembre 2024

Moindzé Soilahoudine
Paul Clabaut
DGDIN - DAR

La DGDIN dans vos problématiques d'IA

AMI Animation Scientifique IA

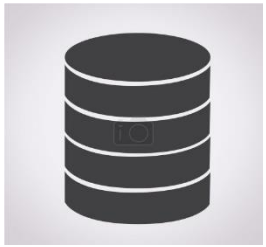
Université Gustave Eiffel

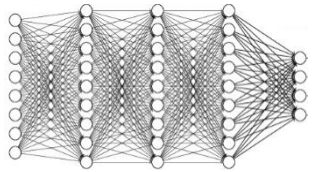
IA

Marvin Lee Minsky, 1956

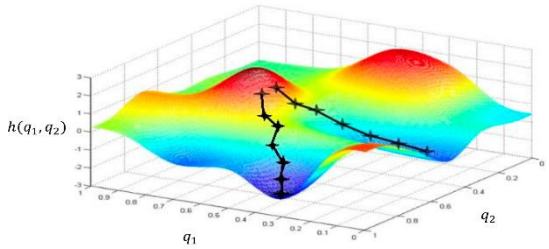
Programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau

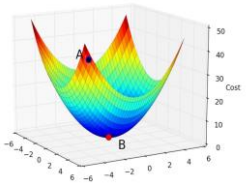
Problème → Données → Entraînement → Produit → Usage

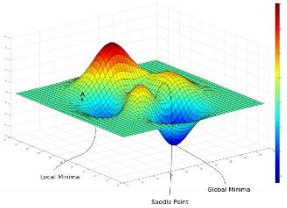


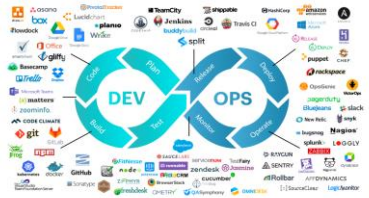


$$x \quad \hat{y} \leftrightarrow F(x)$$



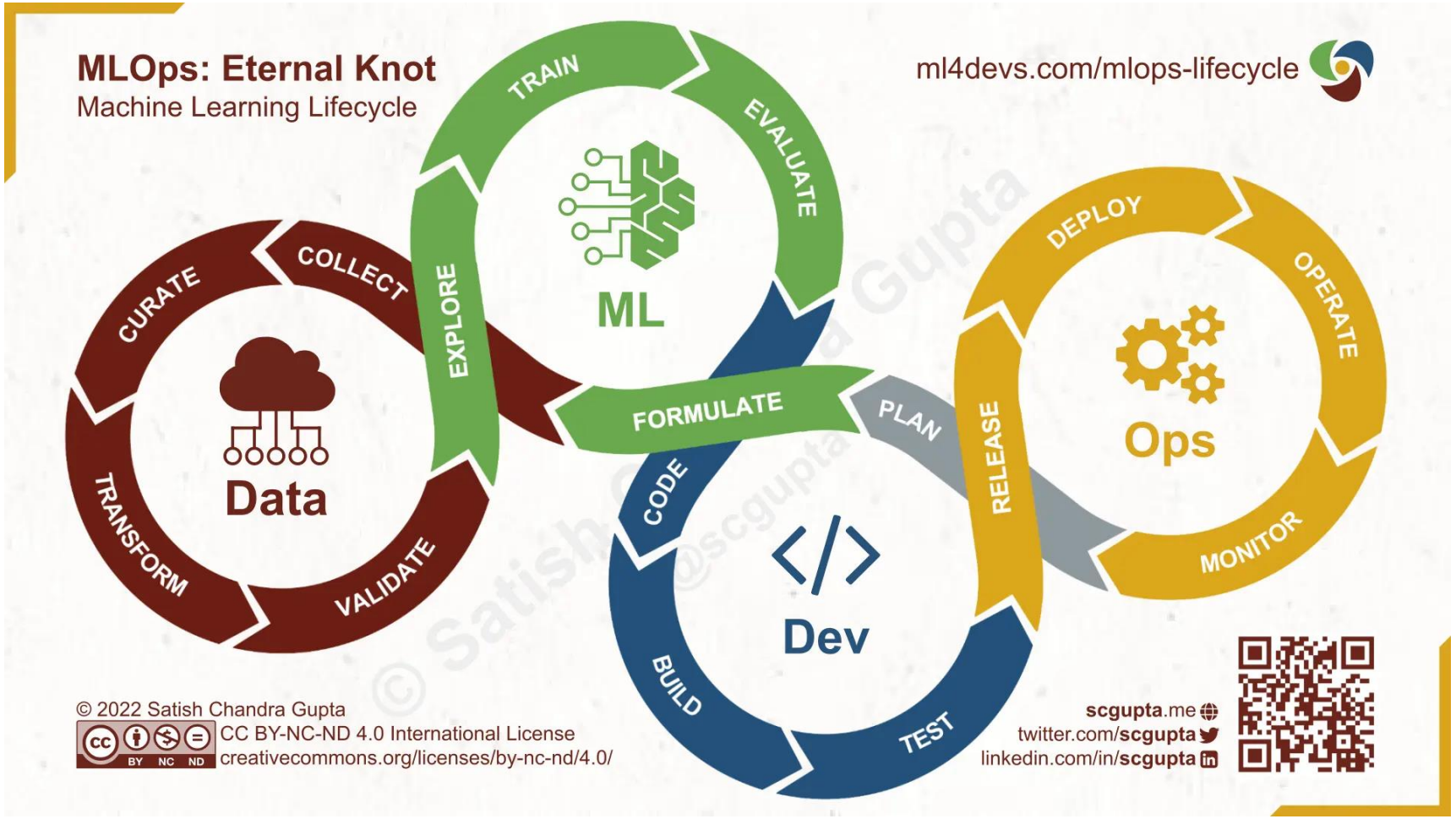




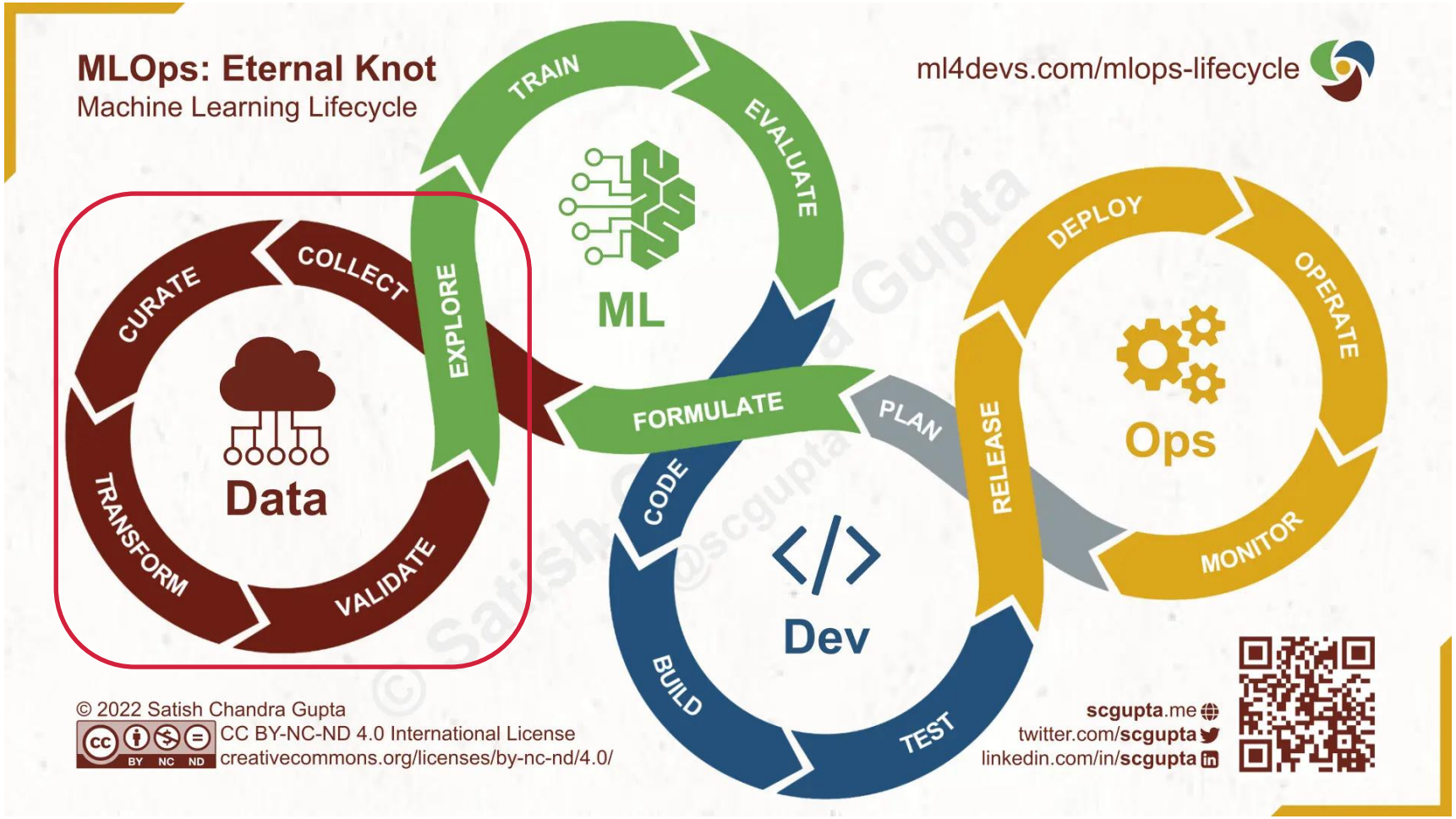


Université Gustave Eiffel

Cycle de vie d'un projet de Machine Learning



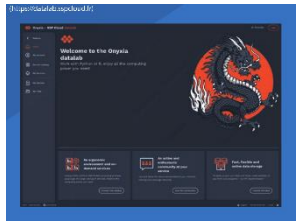
Cycle de vie d'un projet de Machine Learning



DATA : Ingrédients

Stockage

PANDORE



GENCI
Le calcul intensif au service de la connaissance

aws

Google Cloud Platform

Microsoft Azure

Cadre logiciel

python

NumPy

pandas

matplotlib

seaborn

plotly

jupyter

DVC

Université Gustave Eiffel

Data : qualité

Sélection de caractéristiques:

- Suppression
- Extraction
- Transformation

Fail sur les filles

Quand Amazon fabrique par accident une intelligence artificielle qui n'aime pas les femmes

Par **Magazine Marianne**

Publié le 12/10/2018 à 13:00

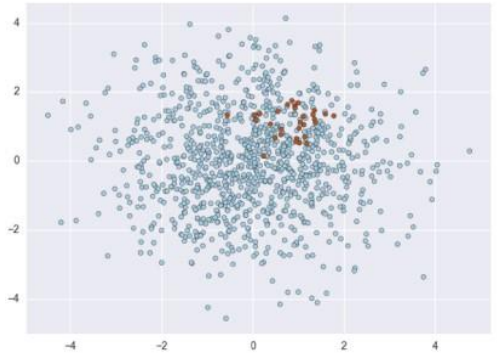
Data : qualité

Sélection de caractéristiques:

- Suppression
- Extraction
- Transformation

Déséquilibre de classes:

- Sous-échantillonnage
- Sur-échantillonnage



Fail sur les filles

Quand Amazon fabrique par accident une intelligence artificielle qui n'aime pas les femmes

Par **Magazine Marianne**

Publié le 12/10/2018 à 13:00

Université Gustave Eiffel

Data : qualité

Sélection de caractéristiques:

- Suppression
- Extraction
- Transformation

Fail sur les filles

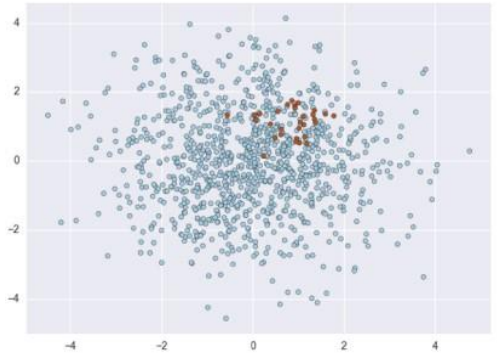
Quand Amazon fabrique par accident une intelligence artificielle qui n'aime pas les femmes

Par **Magazine Marianne**

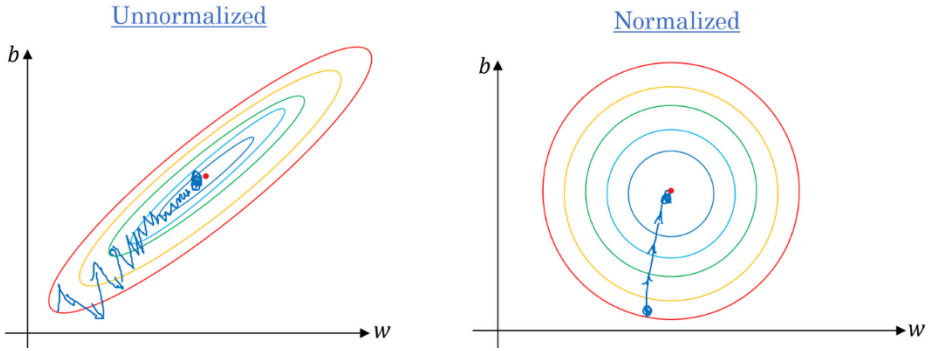
Publié le 12/10/2018 à 13:00

Déséquilibre de classes:

- Sous-échantillonnage
- Sur-échantillonnage



Mise à l'échelle:



Université Gustave Eiffel

Data : qualité

Sélection de caractéristiques:

- Suppression
- Extraction
- Transformation

Fail sur les filles

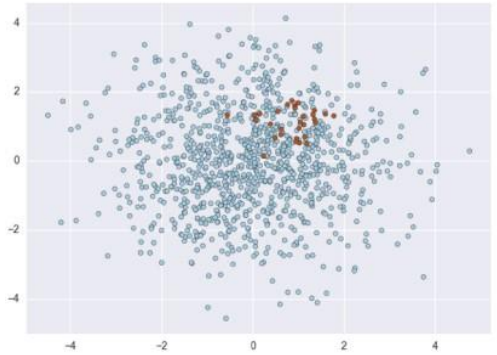
Quand Amazon fabrique par accident une intelligence artificielle qui n'aime pas les femmes

Par **Magazine Marianne**

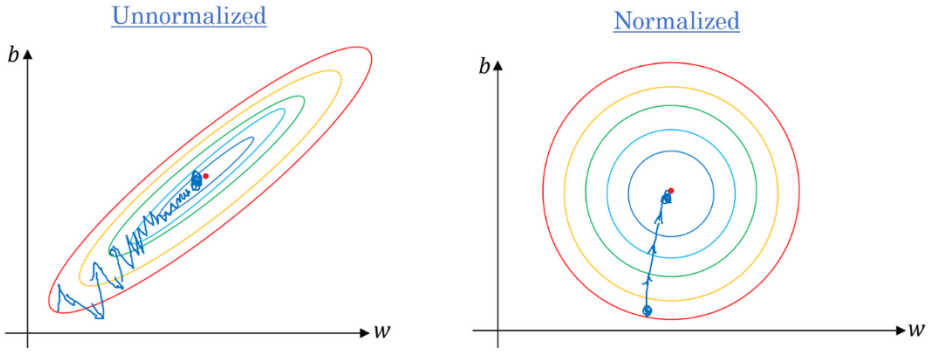
Publié le 12/10/2018 à 13:00

Déséquilibre de classes:

- Sous-échantillonnage
- Sur-échantillonnage



Mise à l'échelle:

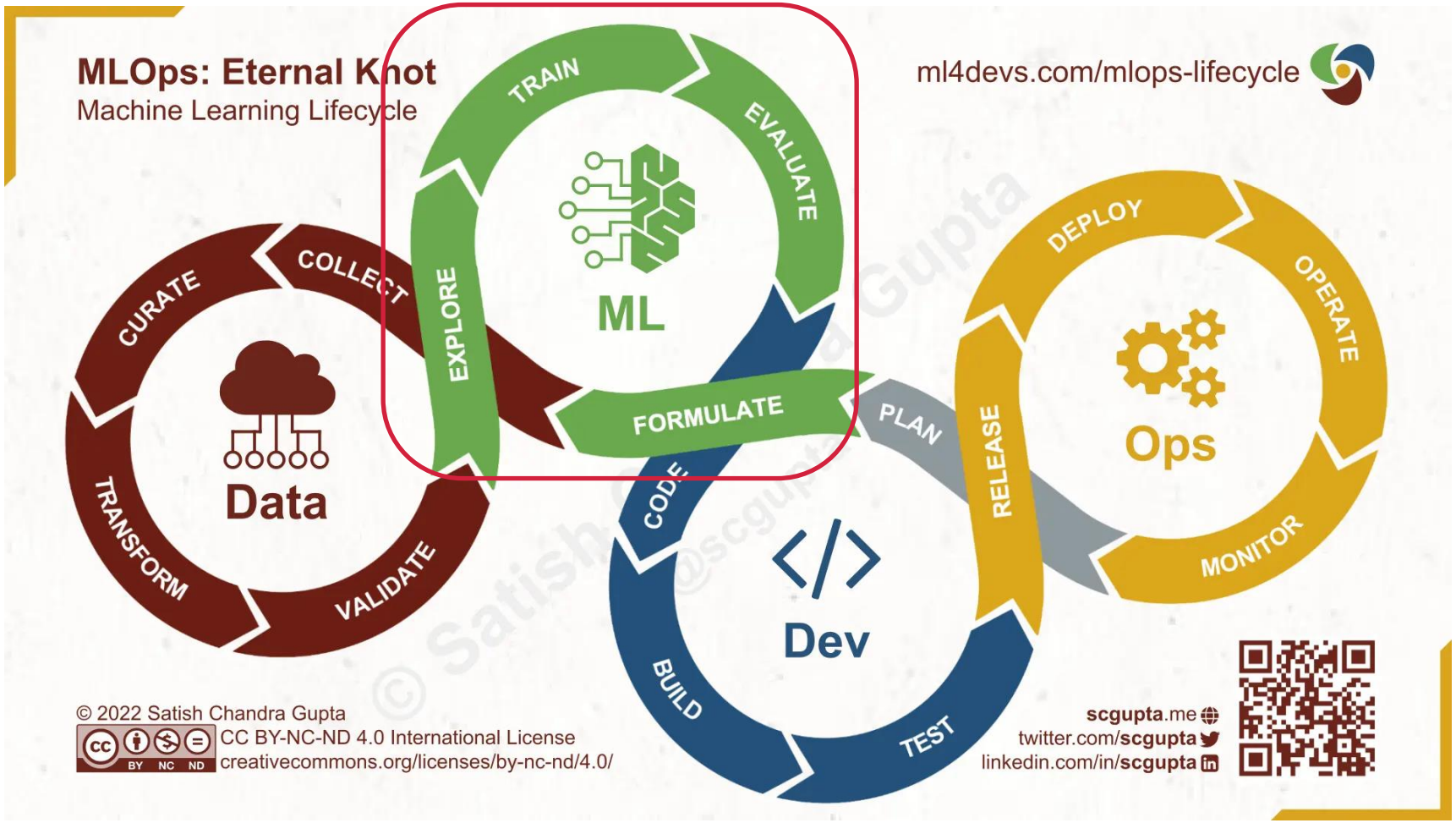


Sans oublier:

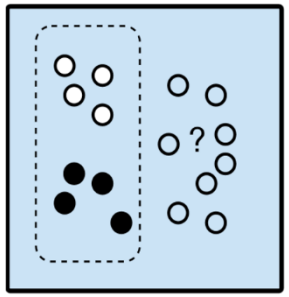
- Analyse exploratoire
- Traitement des valeurs aberrantes
- Traitement des valeurs manquantes

Université Gustave Eiffel

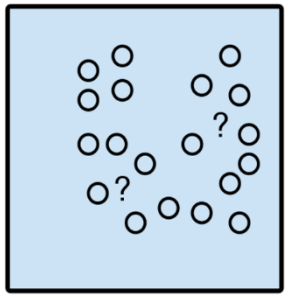
Cycle de vie d'un projet de Machine Learning



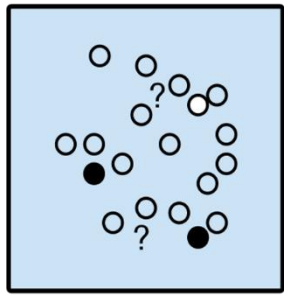
ML : types de modèles



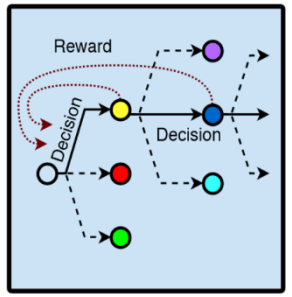
Apprentissage supervisé



Apprentissage non-supervisé

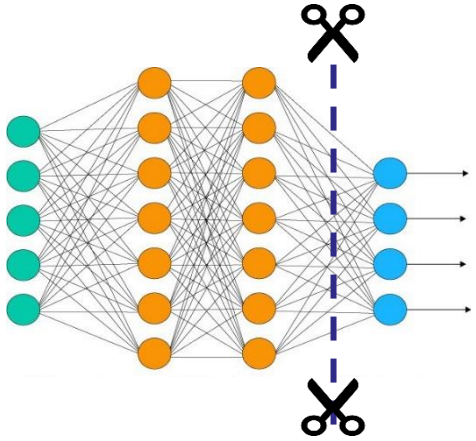


Apprentissage semi-supervisé

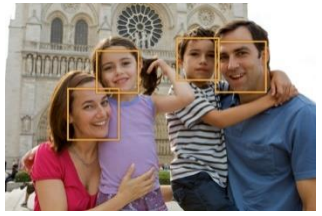


Apprentissage par renforcement

Réseau open-source entraîné pour une application proche

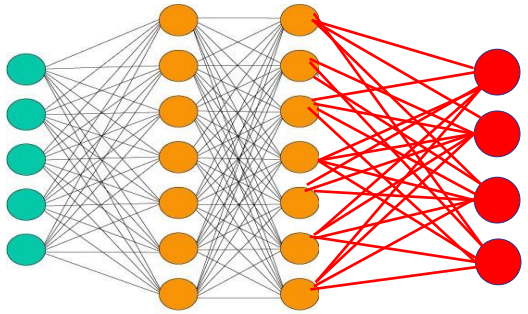


Reconnaissance faciale

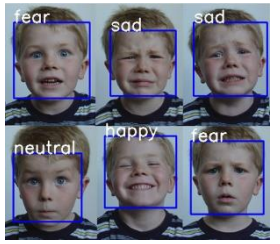


Réseau transféré pour cas d'usage personnalisé

Une seule couche à ré-entraîne



Reconnaissance d'émotions

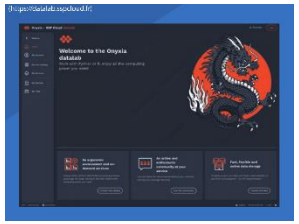


Université Gustave Eiffel

ML : Ingrédients

calcul

ICARE



GENCI
Le calcul intensif au service de la connaissance

aws

Google Cloud Platform

Microsoft Azure

Cadre logiciel

python

NumPy

pandas

jupyter

scikit learn

TensorFlow

PyTorch

Keras

gensim

Transformers

mlflow

Université Gustave Eiffel

ML : Entraînement

Recherche:

- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

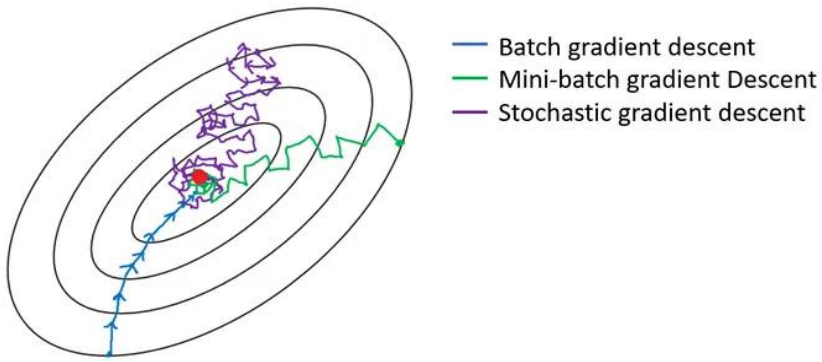
ML : Entraînement

Recherche:

- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

Entraînement:

- Descente de gradient



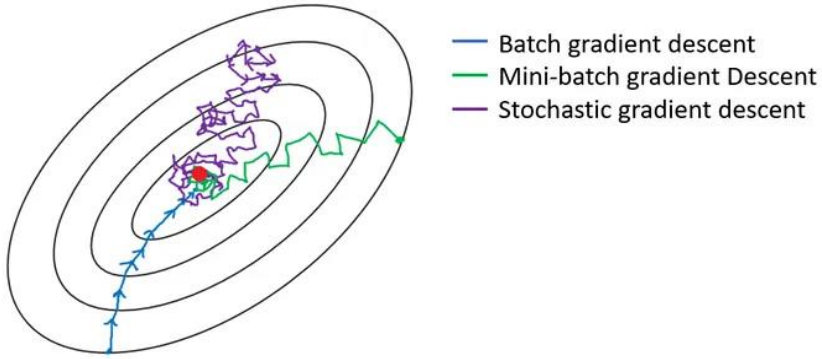
ML : Entraînement

Recherche:

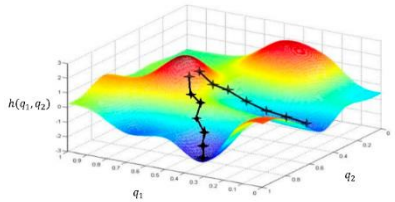
- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

Entraînement:

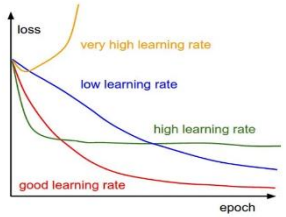
- Descente de gradient



- Initialisation



- Learning rate



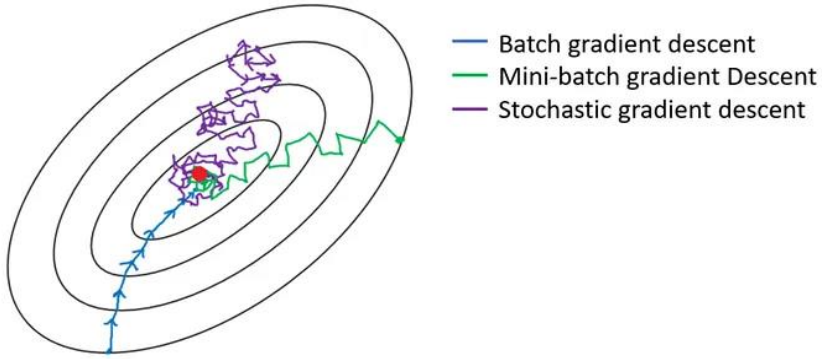
ML : Entraînement

Recherche:

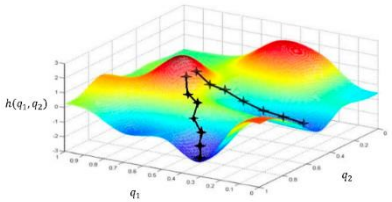
- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

Entraînement:

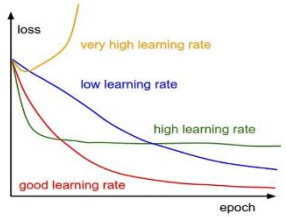
- Descente de gradient



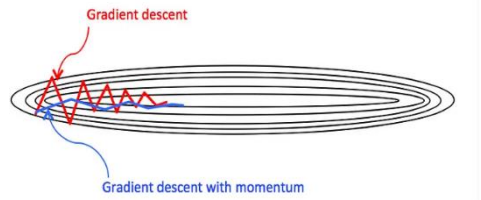
- Initialisation



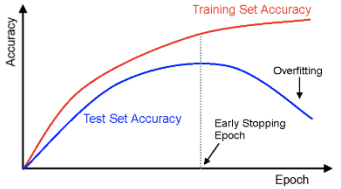
- Learning rate



- Momentum



- Early stopping



Université Gustave Eiffel

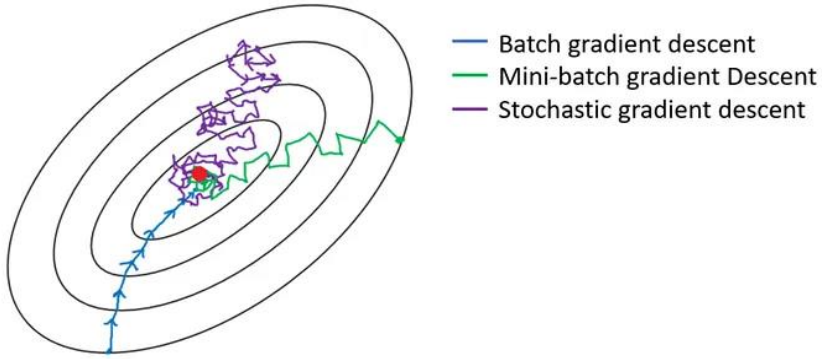
ML : Entraînement

Recherche:

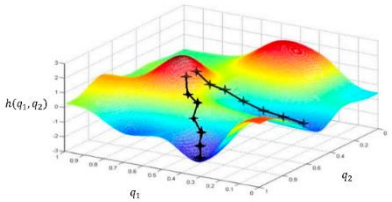
- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

Entraînement:

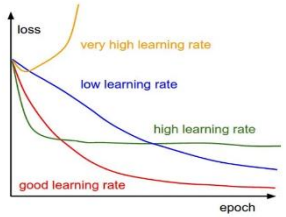
- Descente de gradient



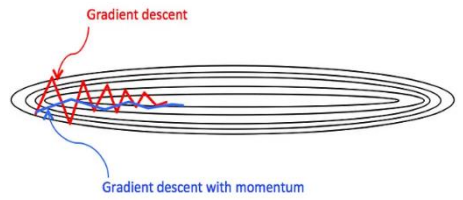
- Initialisation



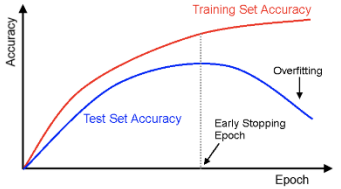
- Learning rate



- Momentum

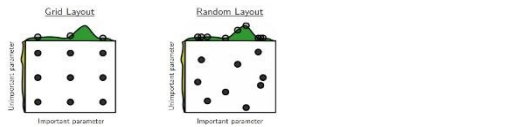


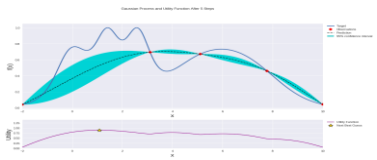
- Early stopping



Optimisation des hyperparamètres:

- Babysitting
- Grid Search
- Random Search
- Bayesian Optimization
  - Optuna
  - Hyperopt





Université Gustave Eiffel

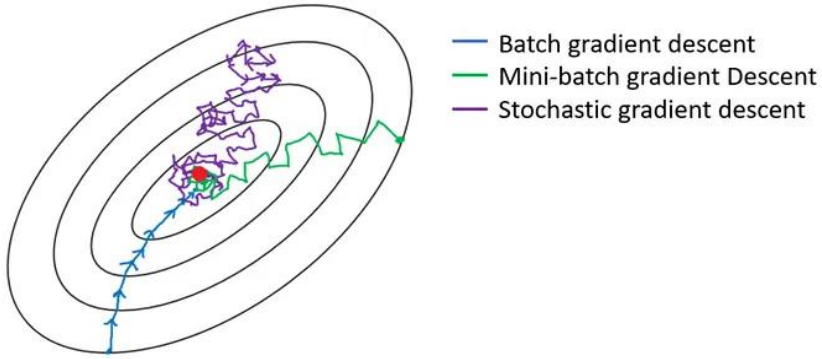
ML : Entraînement

Recherche:

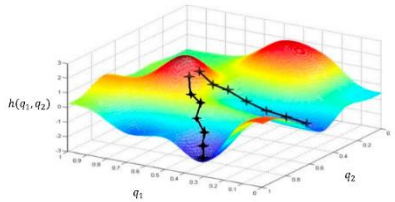
- Tester des algorithmes au hasard peut prendre beaucoup de temps avec des résultats incertain

Entraînement:

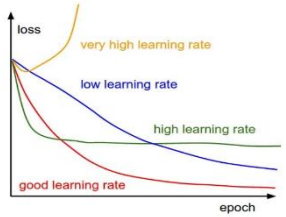
- Descente de gradient



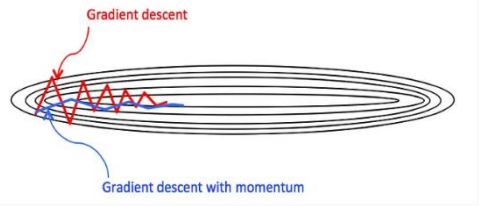
- Initialisation



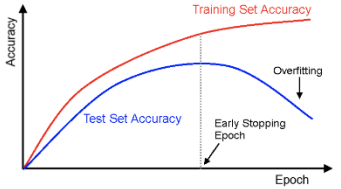
- Learning rate



- Momentum

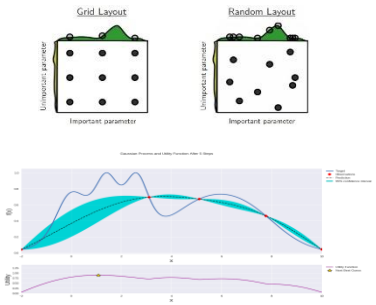


- Early stopping

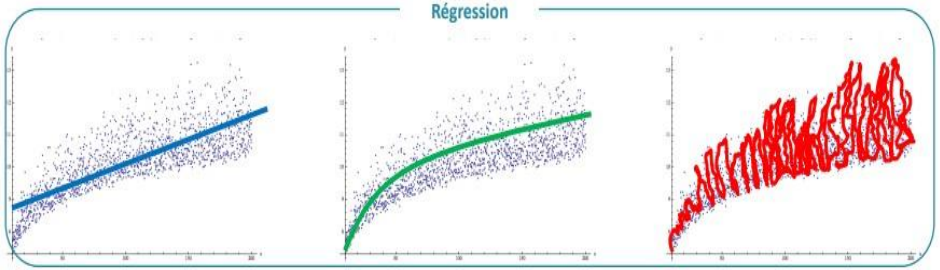


Optimisation des hyperparamètres:

- Babysitting
- Grid Search
- Random Search
- Bayesian Optimization
  - Optuna
  - Hyperopt

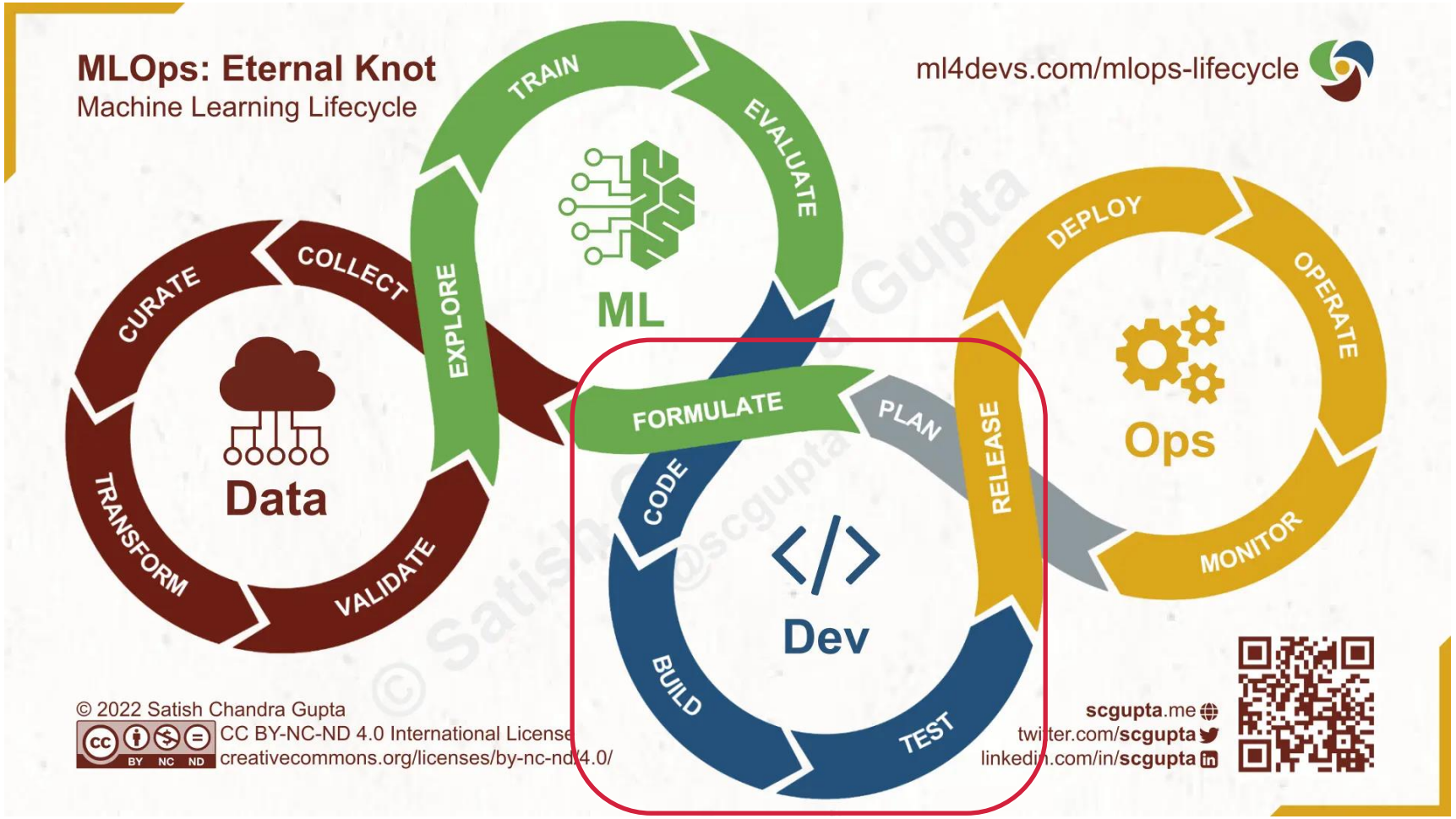


sous-apprentissage et sur-apprentissage:



Université Gustave Eiffel

Cycle de vie d'un projet de Machine Learning



DEV

Gitflow :

DEV

Gitflow :

Environnements virtuels :

	venv		pipenv	Poetry
Native integration with python	✓			
Package & environment management	✓	✓	✓	✓
Sub-dependency management			✓	✓
Active community	✓	✓		✓
Comments		*Use miniconda to avoid huge Anaconda*		*Recommended tool to use!*

DEV

Gitflow :

Environnements virtuels :

	venv		pipenv	Poetry
Native integration with python	✓			
Package & environment management	✓	✓	✓	✓
Sub-dependency management			✓	✓
Active community	✓	✓		✓
Comments		*Use miniconda to avoid huge Anaconda*		*Recommended tool to use!*

Tests :

- **Tests unitaires** : Rapides, déterministes, automatisés
- Code coverage
- TDD
- **Tests des données:** Validation des données par rapport à des contraintes attendues, typage, indicateurs statistiques, etc.
- TDD
- **Tests du modèle :** Tests de performance, tests de la logique du code ML (shapes, …), tests de comportements spécifiques

DEV

Gitflow :

Environnements virtuels :

	venv		pipenv	Poetry
Native integration with python	✓			
Package & environment management	✓	✓	✓	✓
Sub-dependency management			✓	✓
Active community	✓	✓		✓
Comments		*Use miniconda to avoid huge Anaconda*		*Recommended tool to use!*

Tests :

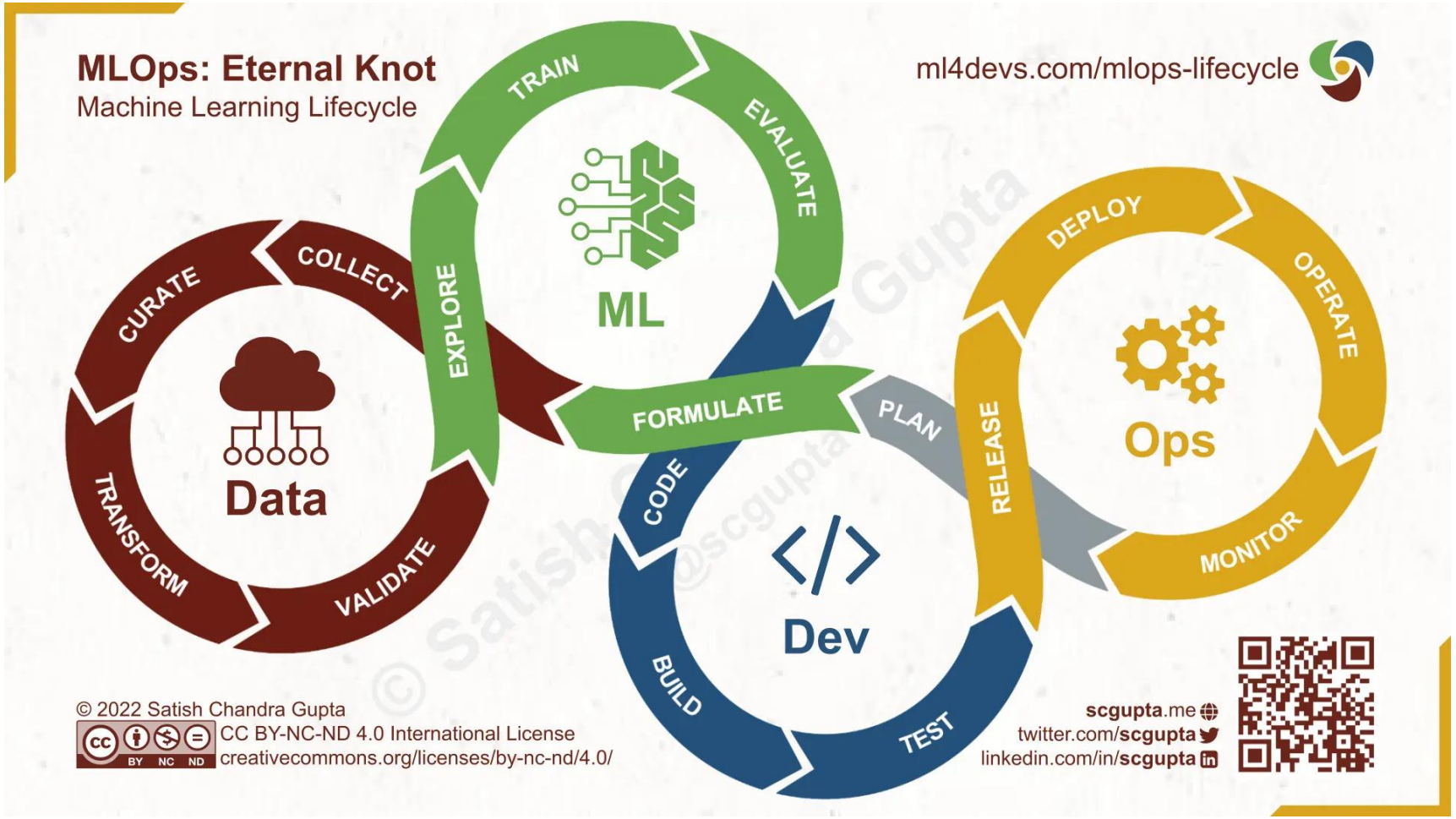
- **Tests unitaires** : Rapides, déterministes, automatisés
- Code coverage
- TDD
- **Tests des données:** Validation des données par rapport à des contraintes attendues, typage, indicateurs statistiques, etc.
- TDD
- **Tests du modèle :** Tests de performance, tests de la logique du code ML (shapes, …), tests de comportements spécifiques

Refactoring :

- Eviter les God Classe : les fonctions ou classe qui ont trop de responsabilités
- Eviter les Feature Envy : les fonctions qui font trop d'appels aux méthodes d'autres
- Utiliser des fichiers, classes et fonctions de petite taille
- Retirer Code mort
- Bien nommer (et convention : snake_case en python, camelCase en scale, …)
- Bien commenter

Soyez attentif aux aspects juridiques (licences) et au RGPD

Cycle de vie d'un projet de Machine Learning



Moyens de calcul

Interne :

- Ordinateur portable ou de bureau
- Workstation
- VM dans ICARE

+ Familier
+ Ressources internes

- Pas de possibilité d'utiliser un GPU
- Administration interne
- Stack à installer

Moyens de calcul

Interne :

- Ordinateur portable ou de bureau
- Workstation
- VM dans ICARE

+ Familier
+ Ressources internes

- Pas de possibilité d'utiliser un GPU
- Administration interne
- Stack à installer

SSP Cloud :

- Plateforme orientée datascience
- Basée sur le projet Onyxia et installée dans les serveurs de l'Insee
- Technologies cloud (conteneurs, Kubernetes, stockage objets compatible S3)
- Ressources de calcul (y compris des GPU)
- Données de type open data ou conforment RGPD

+ Facile d'accès
+ Utiliser un GPU
+ Catalogue de service couvrant l'ensemble du cycle de vie des projets datascience (Jupyter, Rstudio, …)
+ Aucun enfermement propriétaire
+ Interface conviviale pour les utilisateurs

- Environnement de travail volatiles
- Utiliser git
- Utiliser stockage type S3
- Externe à l'UGE

Moyens de calcul

Interne :

- Ordinateur portable ou de bureau
- Workstation
- VM dans ICARE

+ Familier
+ Ressources internes

- Pas de possibilité d'utiliser un GPU
- Administration interne
- Stack à installer

SSP Cloud :

- Plateforme orientée datascience
- Basée sur le projet Onyxia et installée dans les serveurs de l'Insee
- Technologies cloud (conteneurs, Kubernetes, stockage objets compatible S3)
- Ressources de calcul (y compris des GPU)
- Données de type open data ou conforment RGPD

+ Facile d'accès
+ Utiliser un GPU
+ Catalogue de service couvrant l'ensemble du cycle de vie des projets datascience (Jupyter, Rstudio, …)
+ Aucun enfermement propriétaire
+ Interface conviviale pour les utilisateurs

- Environnement de travail volatiles
- Utiliser git
- Utiliser stockage type S3
- Externe à l'UGE

Cluster :

- Ensemble de machines occupant des salles entières (circuit électrique, climatisation, bâtiment dédiés)
- Administrée par une équipe technique dédiée

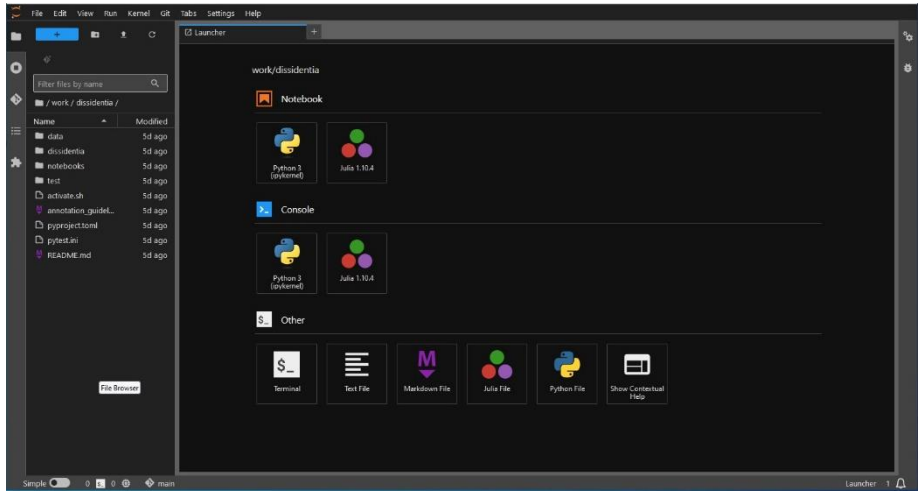
+ Possibilité d'utiliser plusieurs GPU
+ Très haute disponibilité

- Modification nécessaire des codes pour en tirer des performances
- Accès par console
- Queue d'attente (1/2j max)
- Accès pas forcément facile

Moyens de calcul : accès

SSP Cloud

- Créez un compte sur le SSP Cloud en utilisant votre adresse e-mail univ-eiffel.fr
- Connectez-vous à votre compte
- Ajoutez un token d'accès personnel pour Forge git
- Choisir un service via le catalogue de service (Jupyter, Rstudio, Vscode)
- Configurer le service
- Lancer le service
- Ouvrez le service et saisissez le mot de passe du service



Université Gustave Eiffel

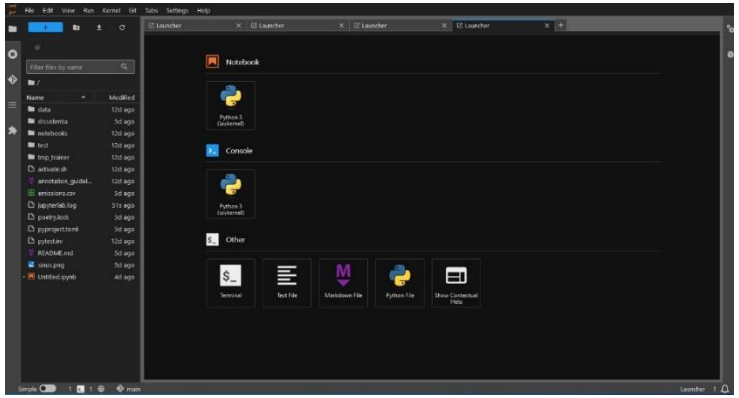
Moyens de calcul : accès

SSP Cloud

- Créez un compte sur le SSP Cloud en utilisant votre adresse e-mail univ-eiffel.fr
- Connectez-vous à votre compte
- Ajoutez un token d'accès personnel pour Forge git
- Choisir un service via le catalogue de service (Jupyter, Rstudio, Vscode)
- Configurer le service
- Lancer le service
- Ouvrez le service et saisissez le mot de passe du service

ICARE

- Demandez l'ouverture d'une VM via l'outil « Ticket »
- Une VM sera mise à votre disposition, en fonction des ressources disponibles et les informations d'accès vous seront communiquées
- Nous pouvons vous accompagner pour la mise en place de votre environnement de travail



https://intranet.univ-eiffel.fr/informatique/informatique-scientifique/tutoriel-ia

Application

Problématique

- Objectif
  - Mettre au point un outil IA de démonstration capable de détecter des phrases typiques d'insatisfaction à l'égard de l'action publique
- Enjeu
  - Améliorer la qualité les services publiques
- Contraintes
  - Il s'agit de la première phase d'un projet en 3 phases
  - Les deux prochaines phases doivent pouvoir être traitées par une autre équipe
  - Données en open data
  - Délais de 2 mois maximum pour le livrable

Collecte des données

- 4 Thèmes
  - Transition écologique
  - Fiscalité
  - Démocratie
  - **Organisation de l' État**

- Question : « Que pensez-vous de l'organisation de l' État et des administrations en France ? De quelle manière cette organisation devrait-elle évoluer ? »
- 80 000 réponses à cette question téléchargeables sur le site de l'État (data.gouv) sous forme de csv

Application : DATA

Exemples de réponses sur le thème de l'organisation de l'État

On doit en finir avec le SENAT qui est un refuge, d'inutiles et qui coûtent pour rien »

Administration trop lourde, il faut alléger le service public, allez vers plus de numérique

C'est un empilement de gabegie financière et de copinage

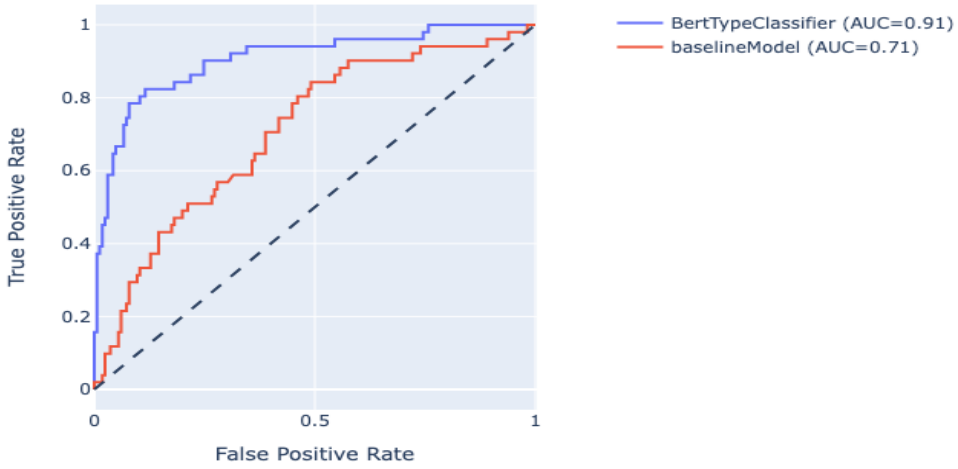
L' État doit se limiter au régalien, sécurité, santé, éducation

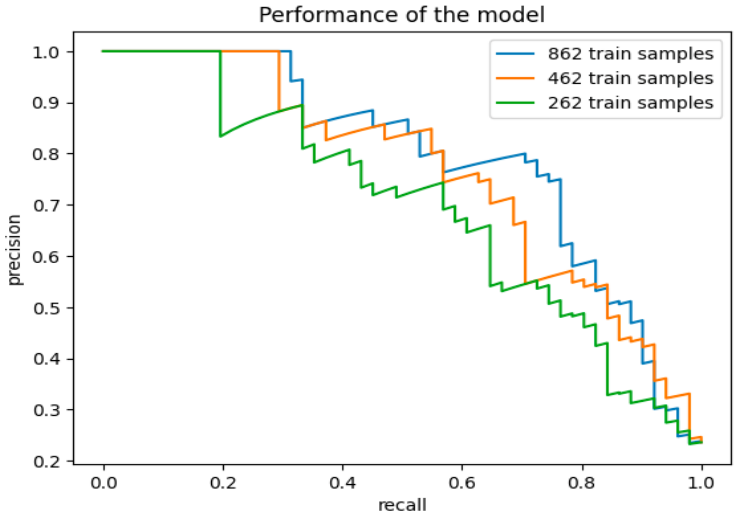
Labellisation

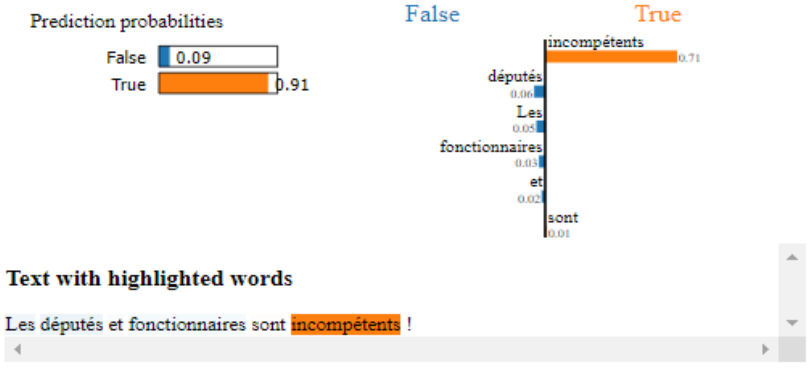
- **Mise au point d'une guideline d'annotation**
- **Annotation d'un échantillon par 3 personnes**
- **Comparaison des résultats et amélioration de la guideline**
- **Recherche de similarité de phrases avec un modèle**

Application : ML

- Solution « classiques » Régression Logistique, RNN
- Solution à l'état de l'art BERT
- Apprentissage par Transfert
- Transformer (API Trainer) : CamemBERT (110M de paramètres, 135GB de text, basé sur RoBERTa)

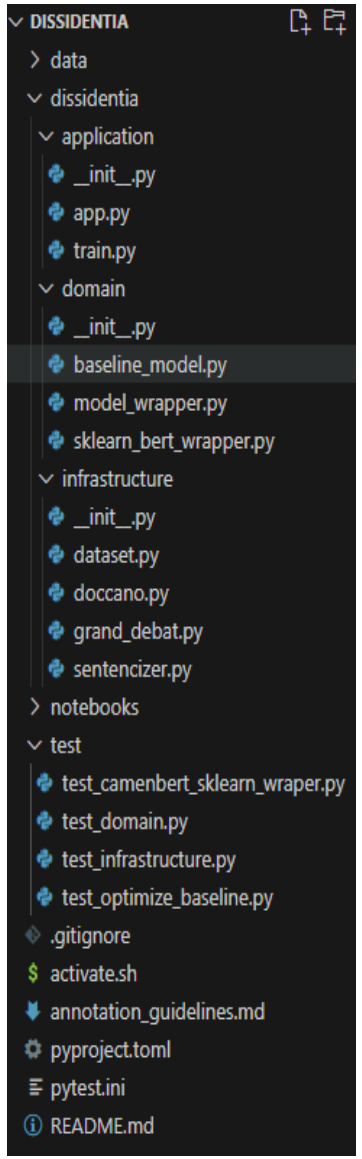




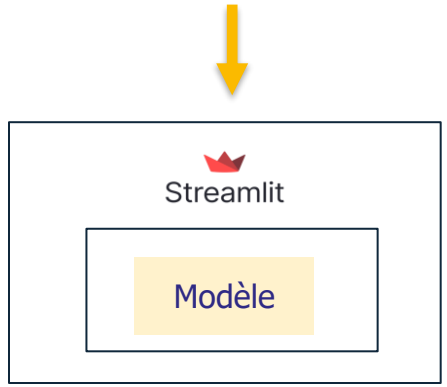


Université Gustave Eiffel

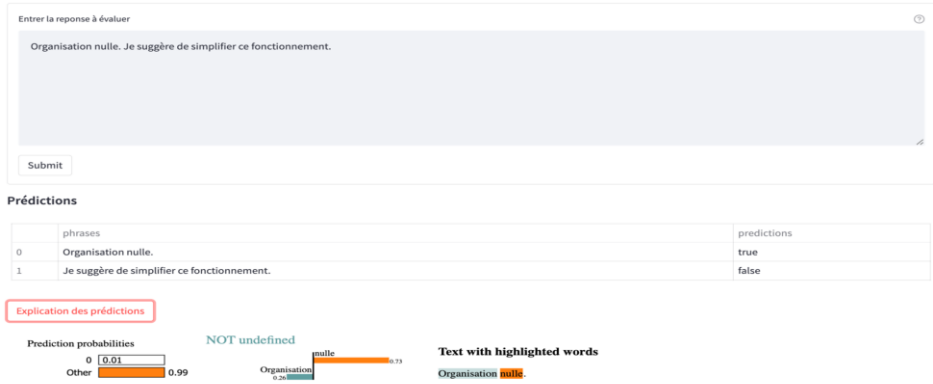
Application : DEV



Réponses du grand débat → Doccano (serveur Azur Quantmetry)



- TDD
- Sprint 0 : Mise en place du socle technique



Model

A scikit-learn wrapper to fine-tuning Bert type model using huggingface Trainer API for dissidentIA detection is proposed and can be used as follow:

```
from dissidentia.domain.sklearn_bert_wrapper import BertTypeClassifier

# define model with default parameters
model = BertTypeClassifier(val_dataset=(x_val, y_val))

# fine-tuning model
model.fit(x_train, y_train)

# make predictions
y_pred = model.predict(x_test)

# make probabilty predictions
y_pred = model.predict_proba(x_test)

# evaluate model on val_dataset for different metrics
model.evaluate()

# save model
model.save(save_path)

# load model
new_model = model.load(save_path)
```

Basic python example

```
from dissidentia.domain.model_wrapper import DissidentModelWrapper

# model BertClassifier is supposed to be already trained and saved (See "Train" section)
model = DissidentModelWrapper.load("BertTypeClassifier").model
model.predict(["Bravo à nos dirigeants pour cet excellent travail !!.",
               "Ce gouvernement est vraiment nul!",
               "Tous des pourris. Démission !!"])
```


Université Gustave Eiffel

Conclusion

Types d'accompagnement

- Vers de bonnes pratiques à des moments clés de vos projets
- Vers des moyens de calcul internes ou externes à l'université
- Vers la mise en place d'environnements de travail

Illustration par un exemple concret

Pour plus d'informations

https://intranet.univ-eiffel.fr/informatique/informatique-scientifique/tutoriel-ia

Moindzé Soilahoudine & Paul Clabaut

moindze.soilahoudine@univ-eiffel.fr

paul.clabault@univ-eiffel.fr

Université Gustave Eiffel