Pas d’IA sans data.

Vers des jeux de données libres et ouverts pour l’éducation.

@jmaupetit // @open_fun

Photo de Susan Q Yin sur Unsplash

> Cette présentation ne parle pas d’« Intelligence Artificielle ».

De l’importance des jeux de données pour apprendre

Problématique

- Dépendance : un jeu de données comportementales
- Comment obtenir un tel jeu de données ?
  - données maison
  - données ouvertes
- Très peu de jeux de données sont disponibles (OULAD, MOOCCubeX, StanfordMoocPosts, et ?)

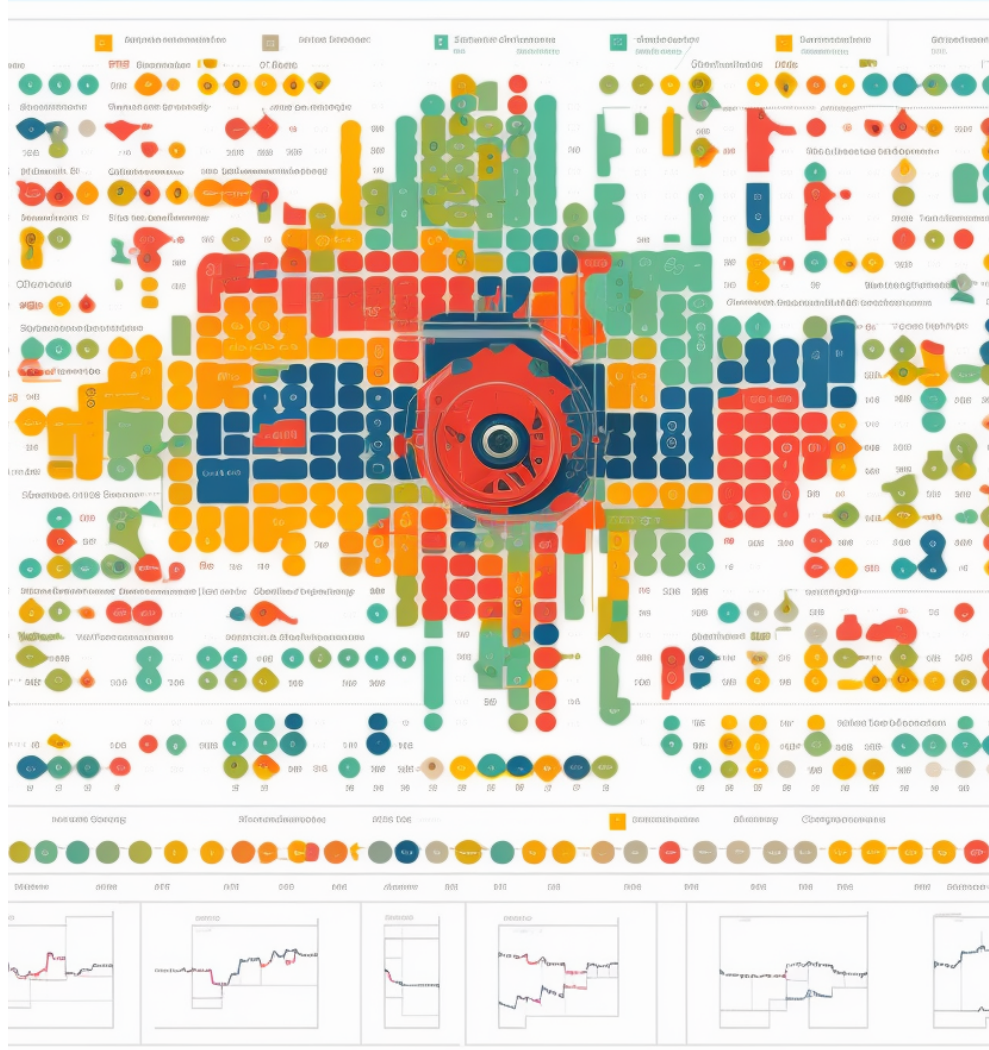

Illustration réalisée par MidJourney

Des données ouvertes pour qui ?

- Chercheurs
- Entreprises
- Citoyens

Illustration réalisée par MidJourney

Freins à l’ouverture

- Pas d’incitation à libérer ces données
- Les traces d’apprentissage sont des données identifiantes

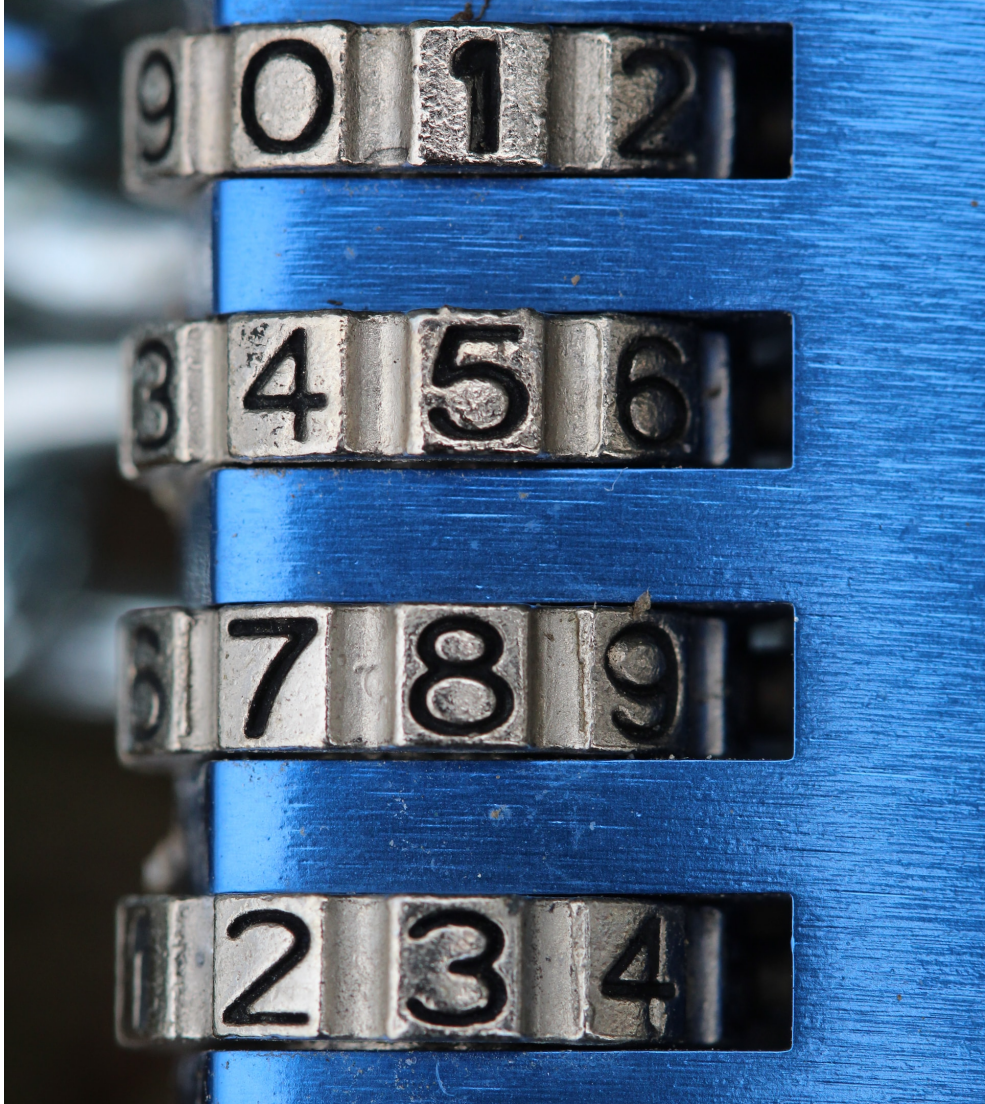

Photo de Anne Nygård sur Unsplash

FUN est un Groupement d'Intérêt Public

Notre crédo :

> Argent public, code public … données publiques.

- Argent public ✅
- Code public * ✅
- Données publiques ⬛

* https://github.com/openfun

Photo de Marco Oriolesi sur Unsplash

Vers une ouverture de nos données d’apprentissage

Défis

1. Quel standard adopter pour ces données ?
2. Comment les anonymiser ?

Photo de Rui Silvestre sur Unsplash

Standard adopté : profils xAPI

- Un vocabulaire contrôlé pour des statements xAPI
- Fusion de datasets de sources différentes
- github.com/Gaia-X-DaSES

Photo de Ferenc Almasi sur Unsplash

Anonymisation

En 2022, FUN est lauréat du Bac à Sable Education de la CNIL.

Premières étapes :

- Minimisation des traces d’apprentissage
- Hachage identifiant utilisateur
- Obfuscation de la clé de cours / ressource pédagogique

Photo de Jesús Rocha sur Unsplash

Anonymisation

Deuxième étape (avant publication) :

- Analyse statistique des traces
- Evaluer la génération de données synthétiques
- Participation à des challenges de ré-identification

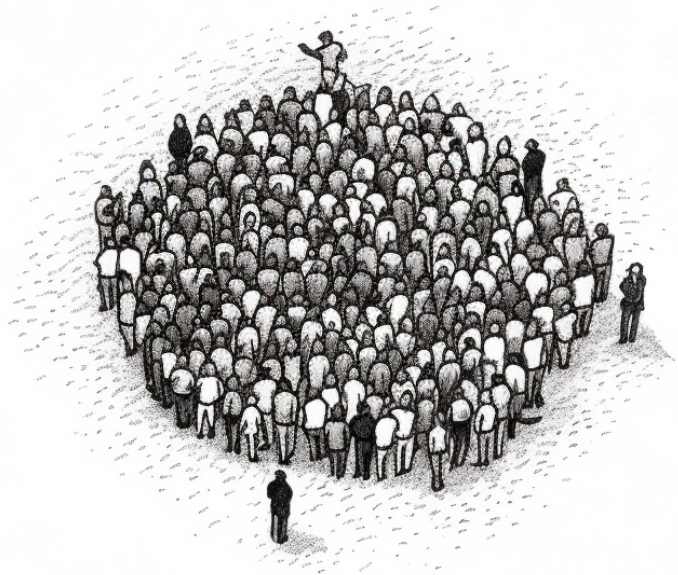

Illustration réalisée par MidJourney

Des questions, des remarques ?

Parlons-en.