Jean-Louis Martin

Unité Mixte De Recherche
Épidémiologique Transport
Travail Environnement

Umrestte / TS2

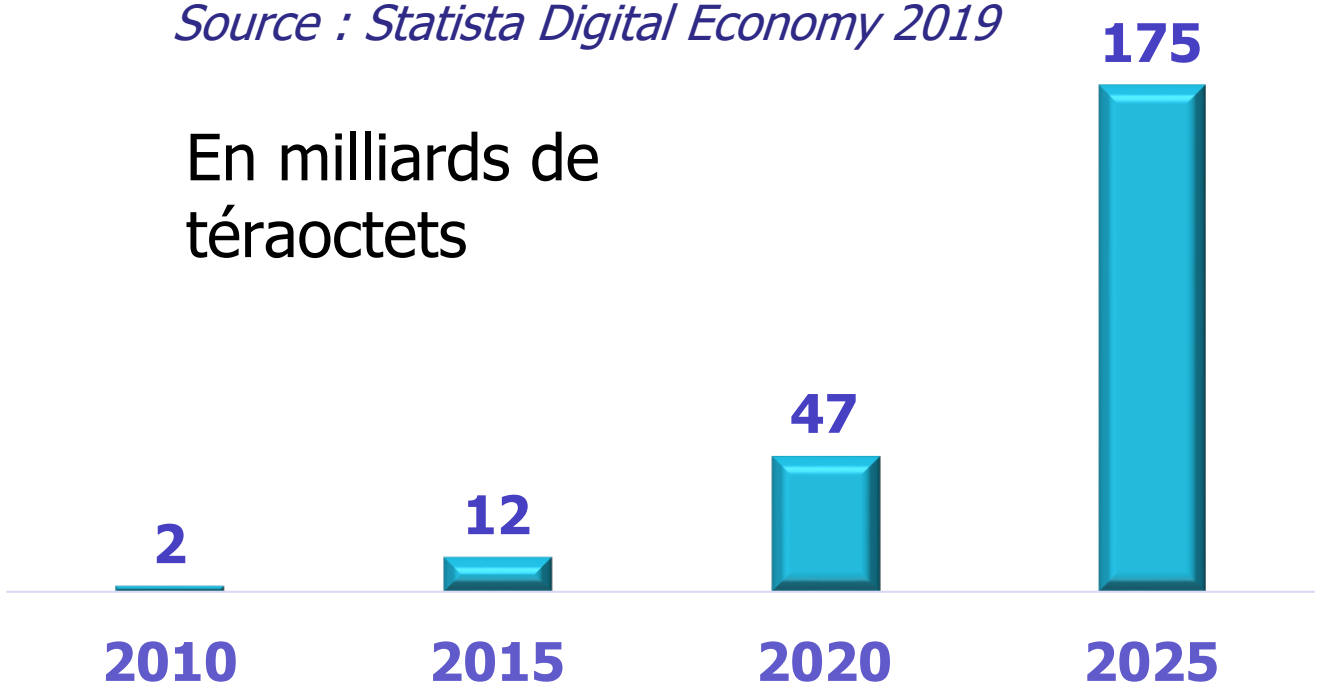
Introduction au Séminaire « grosses données »

Université Gustave Eiffel

15 décembre 2020

Evolution des données numériques mondiales

Source : Statista Digital Economy 2019



- 1,7 Mo : consommation moyenne chaque seconde / personne
- 30 Milliards d'appareils connectés

Deux exemples

Google en 2020

- 100 000 To stockés sur les serveurs Google
- 130 000 Milliards de pages indexées
- Chaque jour :
  - 20 milliards de sites visités
  - 6,9 milliards de requêtes
  - 1 Milliard d'heures de vidéos visionnées sur YouTube

Deezer en 2020

- 70 Millions de fichiers audio en catalogue (musique, podcasts, audiobooks etc), soit plus de 5 Millions de Go
- 7 Millions d'artistes et leurs métadonnées associées (images, biographies, playlists dédiées), nettoyées et mises à jour en continu
- 1 Million de morceaux ajoutés ou mis à jour chaque semaine
- Plus de 2500 lectures audio initiées chaque seconde, alimentant continuellement les algorithmes de recommandation
- Des dizaines de millions de playlists personnalisées générées chaque jour

Les principales questions que cela pose

Les ressources informatiques pour

- Stocker des volumes considérables de données
- Des informations de toutes sortes, de diverses sources
- Avec un très haut niveau de fréquence de création, collecte et partage (En temps réel ➔ vitesse de traitement)

Les méthodes analytiques

- Applicables aux grands volumes de données
- Données non ou peu structurées

Les évolutions technologiques pour le Big Data

Les technologies de stockage, particulièrement le déploiement du Cloud Computing

Plusieurs solutions pour optimiser les temps de traitement sur des bases de données géantes à savoir

- Les bases de données NoSQL (comme MongoDB, Cassandra ou Redis), plus performant que le traditionnel SQL,
- Les infrastructures des serveurs pour du traitement massivement parallèle

Par exemple, Framework Hadoop, combine

- Le système de fichiers distribué HDFS,
- La base NoSQL HBase
- Et l'algorithme MapReduce (concurrencé par Spark)

Le sujet du jour :

Les évolutions des méthodes d'analyse existantes et les méthodes nouvelles

Adaptations de méthodes classiques

- Tests statistiques non discriminants quand n=nombre d'observations très grand
- Sélection des informations pertinentes quand p >> n

Emergence de nouvelles méthodes

- Pour découvrir des « structures » dans de vastes ensembles
- Pour extraire de la connaissance à partir de données non planifiées

Bref historique (1)

Avant 1970, développement de la statistique inférentielle (Fisher, the design of experiments)

- Expérience planifiée
- Petit nombre de facteurs p, sur un petit nombre d'observations
- Test, contrôle du risque de rejeter l'hypothèse nulle
- ➔ ANOVA, modèle linéaire gaussien

Années 70-80, Analyse de données (Benzecri), Exploratory Data Analysis (Tuckey)

- Méthodes descriptives multifactorielles
- Analyse par réduction de dimensions (ACP, AFC, MDS, …)
- Analyse par classification (hiérarchique ou non hiérarchique)

- GLM (Cox, Nelder)

Bref historique (2)

Années 90, data mining, fouille dans des entrepôts de données non planifiées, et recueillies pour des objectifs « comptables »

- Regroupement dans un environnement commun d'outils de gestion de BD, de techniques exploratoires et de modélisation stat
- Emergence de l'apprentissage machine (Vapnik, 1998)

Années 2000, 2010

- Dans plusieurs domaines, c'est le nombre p de variables >> n (biotechnologies omiques, bioinformatique)
  - FDR (False Discovery rate) de Benjamini et Hochberg plutôt que p-valeur
  - Apprentissage stat (Hastie & Tibshirani 2009) pour sélection des modèles par un compromis biais vs Variance

- Avec internet généralisé, datification quasi systématique avec enregistrement des traces numériques du quotidien (y compris géolocalisation) ➔ adaptation des méthodes

Quelques méthodes « récentes »

Approches pénalisées

Machines à vecteurs supports

- Résolution de problèmes de discrimination, recherche de surfaces séparatrices hyperplans ou non linéaires, généralisation au cas de données non séparables

Méthodes à noyaux

- Permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant fondamentalement sur des méthodes linéaires.
- Une fonction noyau correspond à un produit scalaire dans un espace de redescription des données, souvent de grande dimension.
- Dans cet espace, qu'il n'est pas nécessaire de manipuler explicitement, les méthodes linéaires peuvent être mises en œuvre pour y trouver des régularités

Méthodes "Tree-based": Forêts aléatoires, boosting (agrégation de modèles)

Réseaux de neurones profonds (deep learning)

Les principales approches d'apprentissage

Apprentissage supervisé (régression, classification):

- les données ont une étiquette (variable réponse) qu'on cherche à prédire

Apprentissage non-supervisé (clustering):

- les données n'ont pas d'étiquette, le nombre et la définition des classes sont inconnus

Apprentissage semi-supervisé:

- on connait certaines étiquettes, parfois accès à un nombre limité d'étiquettes pour certaines données; lesquelles donneront le plus d'info)

Apprentissage multi-agent ou collaboratif ou fédéré:

- on apprend à partir de plusieurs jeux de données sans les combiner
- Exemple, données de téléphones portables :
  Algo lancé sur chaque téléphone, combinaison des résultats sans stockage des données individuelles (confidentialité des données, etc.)

En bref

Big data → nouveaux ordres de grandeur concernant la capture, le stockage, la recherche, le partage, l'analyse et la présentation des données

Au plan scientifique, les questions classiques demeurent :

- Comment synthétiser l'information ?
- Quelle fiabilité peut-on accorder aux résultats ?
- A quelle population, à quel domaine s'appliquent-ils ?
- Exhaustivité vs échantillonnage ?
- Significativité statistique vs causalité ?

Les quatre interventions

- **Méthodes pénalisées pour l'apprentissage supervisé en grande dimension, applications en cancérologie et en traumatologie,**
  Vivian VIALLON (International Agency for Research on Cancer)

- **Apprentissage non supervisé, application à des données de mobilité,**
  Etienne COME (Univ. Gustave Eiffel, GRETIA)

- **Apprentissage profond en pratique,**
  Jean SAVINIEN (EM Lyon Business School)

- **Fouille de données textuelles, application à l'analyse d'opinion sur le Web,**
  Julien VELCIN (Univ. Lyon 2, ERIC UR 3083)

Jean-louis.martin@univ-eiffel.fr

Université Gustave Eiffel