Python for Data Science

Vue d'ensemble

marie.szafranski@ensiie.fr

Début : 09h20

Datascience in Real Life

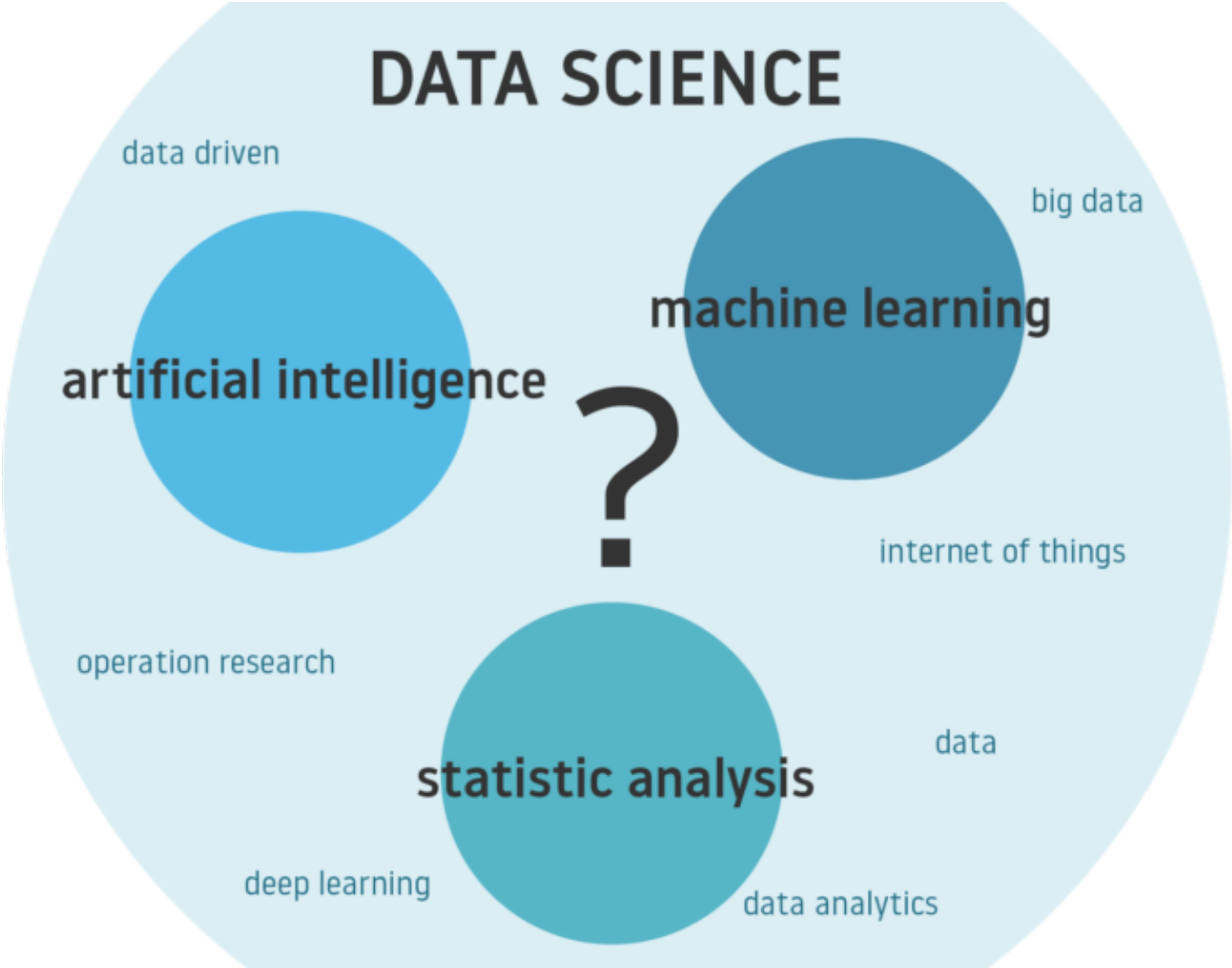
Vue d'ensemble

marie.szafranski@ensiie.fr

Début : 09h20

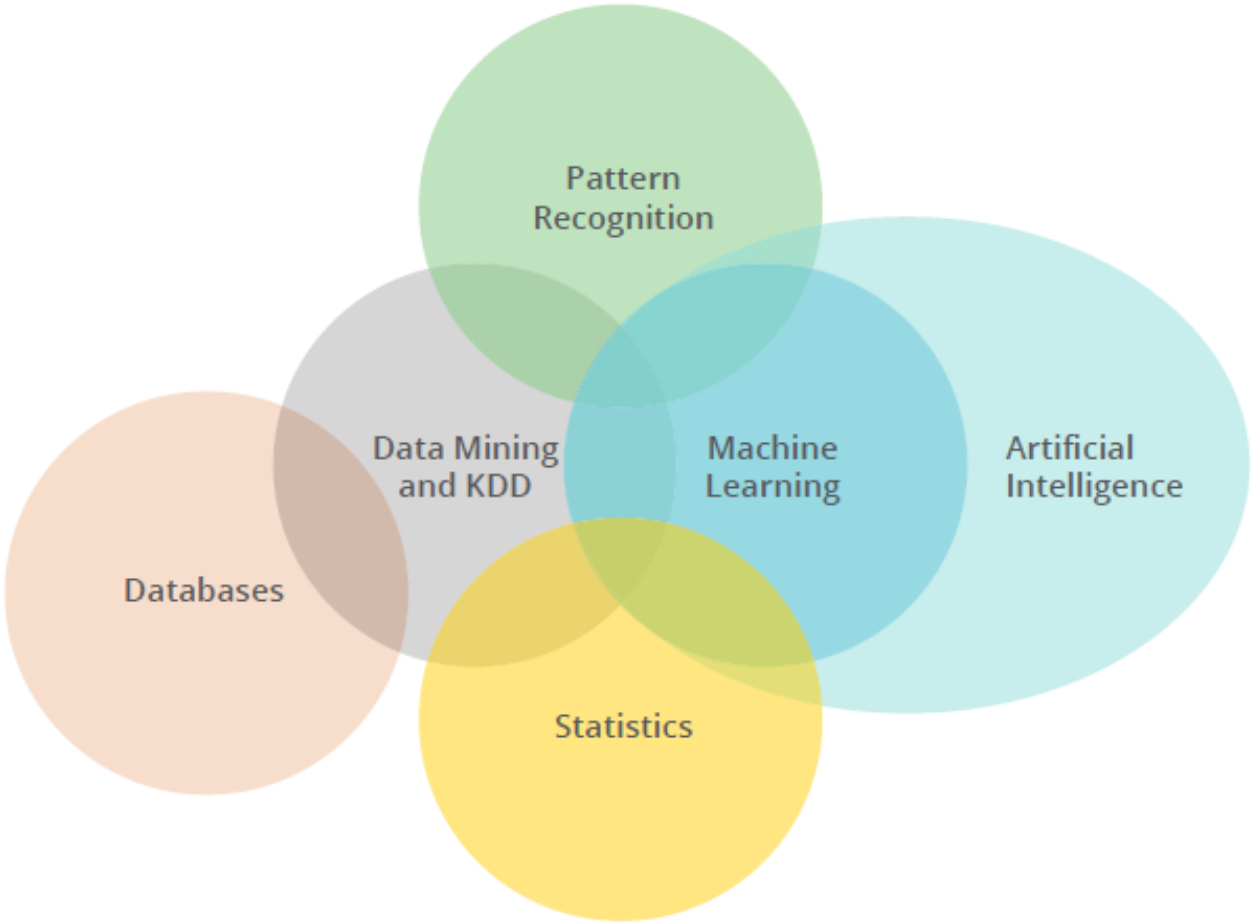
De quoi parle-t-on ?

Buzzwords en vrac



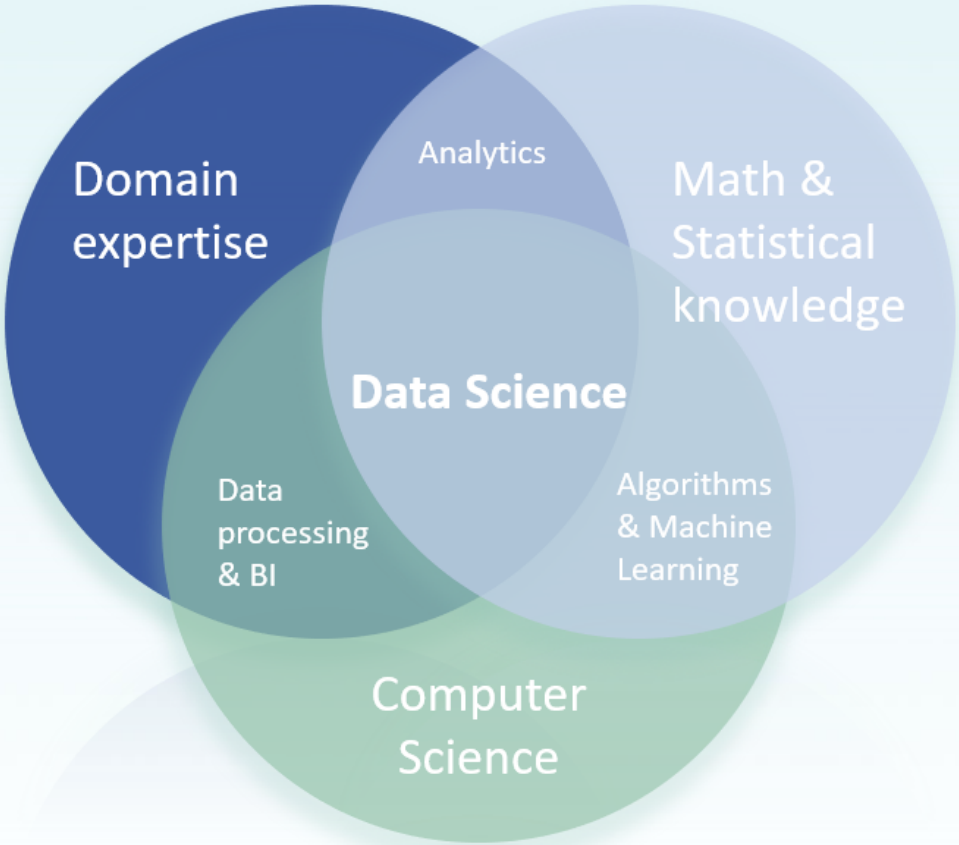
De quoi parle-t-on ?

Une structure « académique »



De quoi parle-t-on ?

Une structure « opérationelle »



Quelques ordres de grandeur [Wikistat, 2016]

Quelles données ?

N individus décrits par D variables

- **kO** 1970s
  Analyse de données
- **MO** 1980s
  Apprentissage automatique $\rightsquigarrow$ réseaux de neurones
- **GO** 1990s
  Exploration et fouille de données [Fayyad et al., 1996]
  Apprentissage statistique $\rightsquigarrow$ SVM [Vapnik, 1995]
- **TO** 2000s
  Bioinformatique $D \gg N$
- **PO** 2010s
  Réseaux sociaux, e-commerce, etc. . . $N \gg D$

Des données à grande échelle

N ou/et D très grands

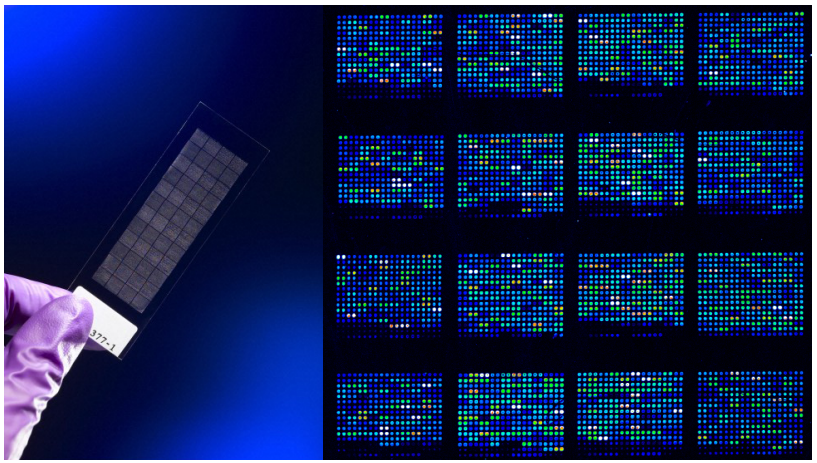
Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Génomique

- N : quelques centaines de patients
- D : mesure de l'expression de plusieurs millions de variants génétiques



Des données à grande échelle

N ou/et D très grands

Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Astronomie

- N : plusieurs millions de corps célestes
- D : quelques centaines de mesures (positions, vitesses, etc.)

Des données à grande échelle

N ou/et D très grands

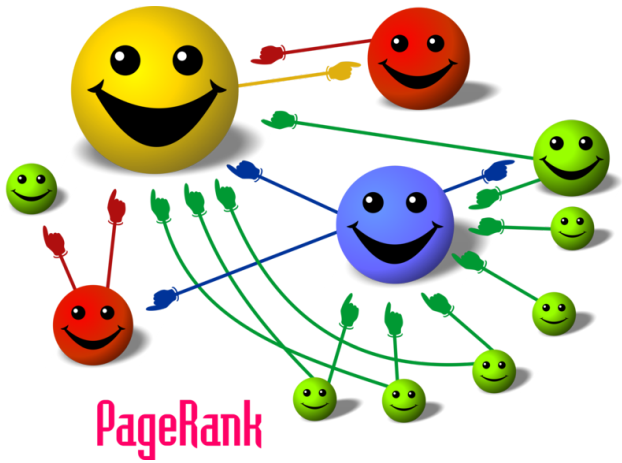
Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Web et text mining

- N : plusieurs millions de sites
- D : liens entrants et sortants, ancres, nom de domaine, hébergeur, etc.



Quelle finalité pour ces données ?

Définir un modèle et réaliser l'algorithme associé dans une perspective prédictive et / ou explicative

Exemples

- **Bioinformatique :** identifier les gènes qui permettent de distinguer des patients sains de patients atteints d'une maladie
- **Astronomie :** étudier des relations liant des paramètres de positions et de vitesses de corps célestes à leur composition chimique
- **Web mining :** analyser le comportement d'internautes pour définir un algorithme de recherche sur les sites web (PageRank de Google)

Schéma global

À la croisée de l'informatique et des statistique

1. Phase de collecte et d'intégration

- Architecture des données
- Web et réseaux distribués

Schéma global

À la croisée de l'informatique et des statistique

2. Phase exploratoire

- Analyse de données
- Recherche d'information et de motifs

Schéma global

À la croisée de l'informatique et des statistique

3. Phase explicative ou décisionnelle

- Machine Learning
- IA interprétable (≠ XAI)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

3. Phase explicative ou décisionnelle

- Machine Learning
- IA interprétable (≠ XAI)

Environnement : vraie R&D

Concurrence : +++ (> doctorat)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

2. Phase exploratoire

- Analyse de données
- Recherche d'information et de motifs

Environnement : partout

Concurrence : ++ (ingé. XP / doct.)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

1. Phase de collecte et d'intégration

- Architectures des données
- Web et réseaux distribués

Environnement : partout

Concurrence : $-/=$ (ingé. / doct.)

DataOps $\rightsquigarrow$ MLOps

Contenu de l'UE

Aspects exploratoires, explicatifs et décisionnels

⇝ ensiie selon les parcours + M2 Datascale

J1. Introduction
J2. Apprentissage non supervisé
J3. Apprentissage supervisé

Mise en situation

⇝ aperçu de l'ensemble des phases, avec un accent sur l'une d'elles

- Cas d'étude — PAYPS
- Projet — PAYPS ou Criteo

Informations pratiques

Planning prévisionnel

Séances	Matin ∼ 3h30	Après-midi ∼ 3h30
12/09	Présentation des projets Criteo	Cas d'étude PAYPS + Travail sur le cas d'étude
19/02	Travail sur le cas d'étude	Travail sur le cas d'étude
26/09	Solution élèves sur le cas d'étude	Solution PAYPS sur le cas d'étude + Présentation projets PAYPS
03/10	J1. Méthodologie // projets	J1. Méthodologie // projets
10/10	J2. Non supervisé // projets	J2. Non supervisé // projets
17/10	Semaine entreprise // projets	Semaine entreprise // projets
24/10	J3. Supervisé // projets	Soutenances projets PAYPS
sem 31/10	Projets Criteo (immersion) et soutenances	

Informations pratiques

Supports

https://pydio.pedago.ensiie.fr/public/pub/FISE_PYDS35

Language pour les TP (Jx.) libre...

`Scikit-learn`, `R`, ... [Pedregosa et al., 2011][CRAN]

Ceci n'est pas un cours de `Python` (ni de `R`)

Modalités d'évaluation

Cas pratique : compte-rendu ⇝ 3 pages
participation (pitch, discussions, etc)

Projets : compte-rendu ∼ 3 pages
soutenance avec les enreprises

Références I

CRAN. The Comprehensive R Archive Network. URL `https://cran.r-project.org`.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011. URL `https://scikit-learn.org/stable/index.html`.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Team Wikistat. Wikistat, 2016. URL `http://wikistat.fr/`.