Science des données I : module 1

Introduction

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

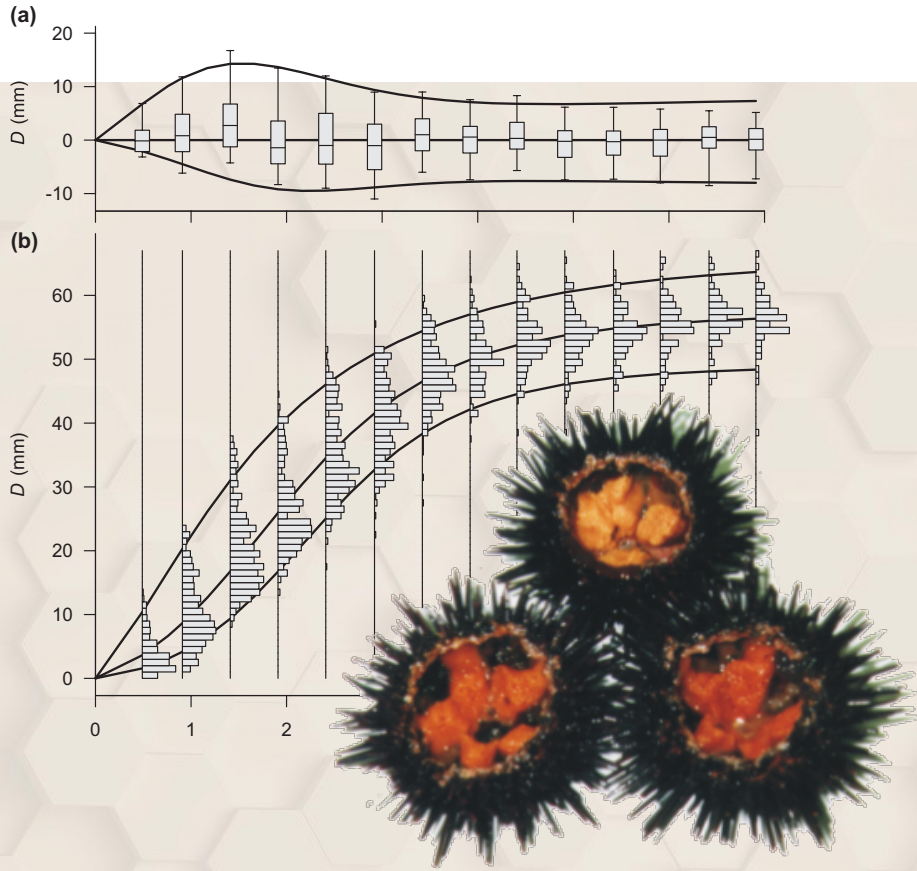
http://biodatascience-course.sciviews.org
sdd@sciviews.org

UMONS

Qui sommes-nous ?

UMONS

Prof. Philippe Grosjean



- **Bioingénieur** + thèse de doctorat en biologie marine (croissance d'oursins)
- Capacités supplémentaires développées en **science des données** durant des post-docs et via de la consultance pendant 4 ans partout en Europe
- **Laboratoire EcoNum** créé en 2004 à l'Université de Mons
- Intéressé par des travaux **interdisciplinaires** : biologie, chimie, modélisation, statistiques, informatique
- **Écrit des logiciels** pour l'écologie en R, Python, ...

UMONS

Guyliann Engels

- **Master** en Biologie des Organismes et Écologie à l'UMONS.
- **Mémoire** effectué dans le laboratoire d'Écologie numérique des Milieux aquatiques sur l'écophysiologie et l'écotoxicologie de la posidonie (*Posidonia oceanica*, une plante marine) en Méditerranée.
- **Thèse de doctorat** en cours sur l'écophysiologie des coraux tropicaux dans le même laboratoire.
- **Assistant** en biologie à l'UMONS depuis septembre 2017.

UMONS

Ecophysiologie des coraux en mésocosmes

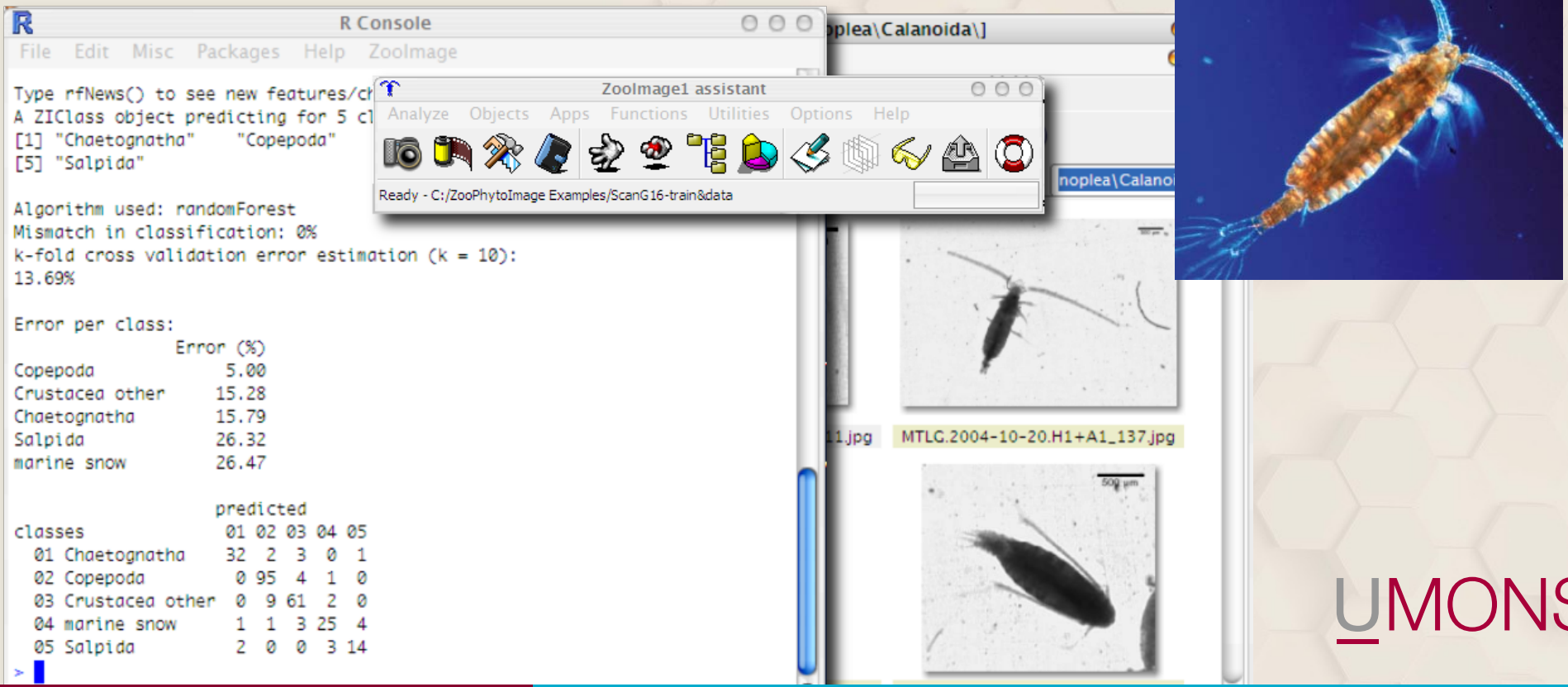
Les récifs de coraux tropicaux forment des écosystèmes riches et diversifiés, mais ils sont en danger face aux changement climatiques globaux, la surpêche et la pollution.

Au laboratoire EcoNum, nous étudions comment l'environnement affecte la croissance, la reproduction et la santé des coraux tropicaux en mésocosmes récifaux artificiels.

Identification automatisée du plancton

Le plancton (constitué des organismes aquatiques qui dérivent en pleine eau) forme des communautés très diversifiées. Un litre d'eau de mer contient typiquement des milliers d'espèces de plancton.

Au laboratoire EcoNum, nous développons des outils pour énumérer automatiquement plancton via l'analyse d'image combinée à la classification supervisée (une technique statistique que nous étudierons en Master 1).

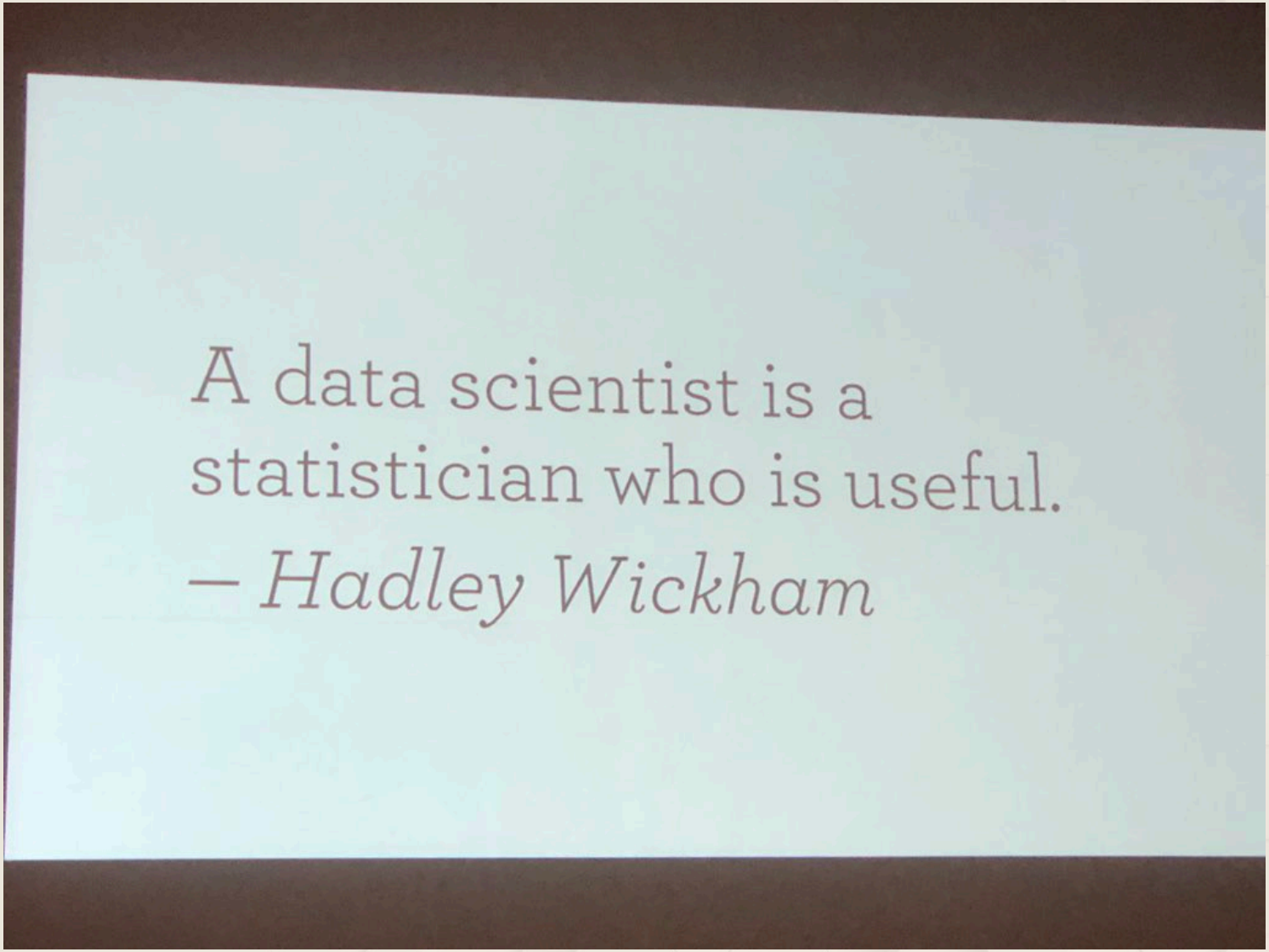


UMONS

Qu'est-ce que la science des données ?

UMONS

Science des données : une approche pragmatique

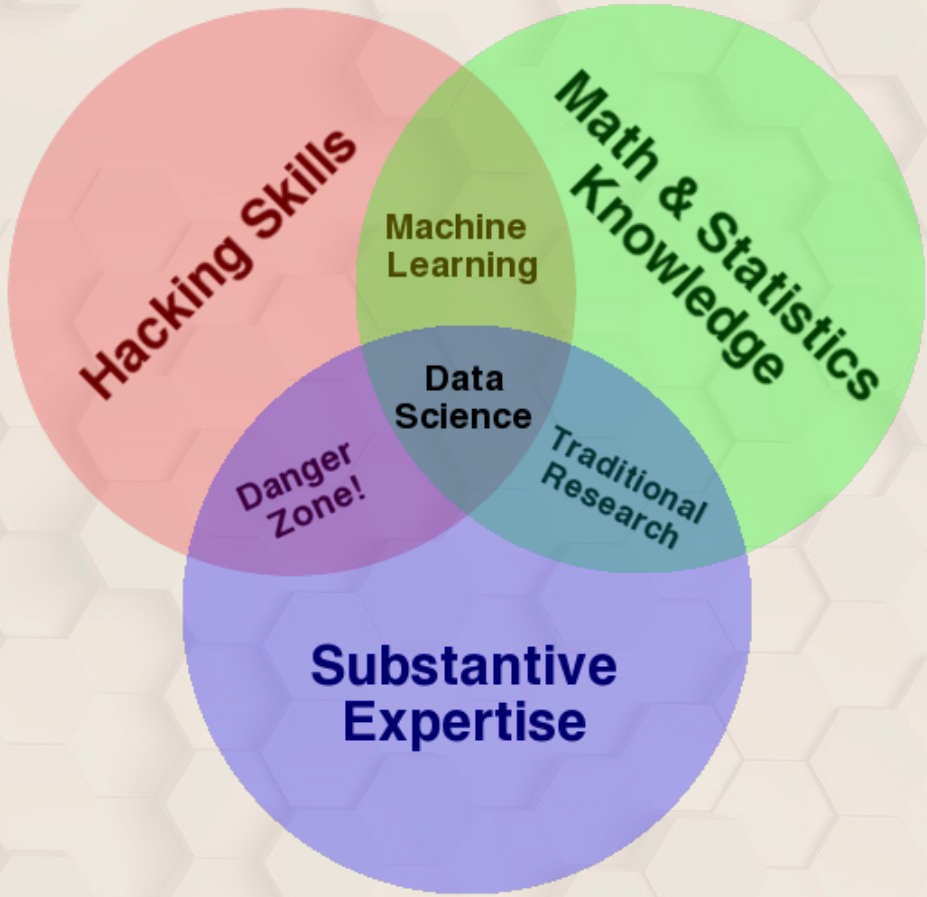


UMONS

Science des données : à l'interface entre plusieurs disciplines

- La Science des Données, c'est la discipline qui s'intéresse à l'analyse de données *sous toutes ses formes*
- Très large et **interdisciplinaire** :
  - **(Bio)statistiques** et visualisation
  - Utilisation d'**outils informatiques**
  - Expertise dans le domaine (**biologie**)
- Il faut maîtriser simultanément les 3 domaines pour être un scientifique des données.

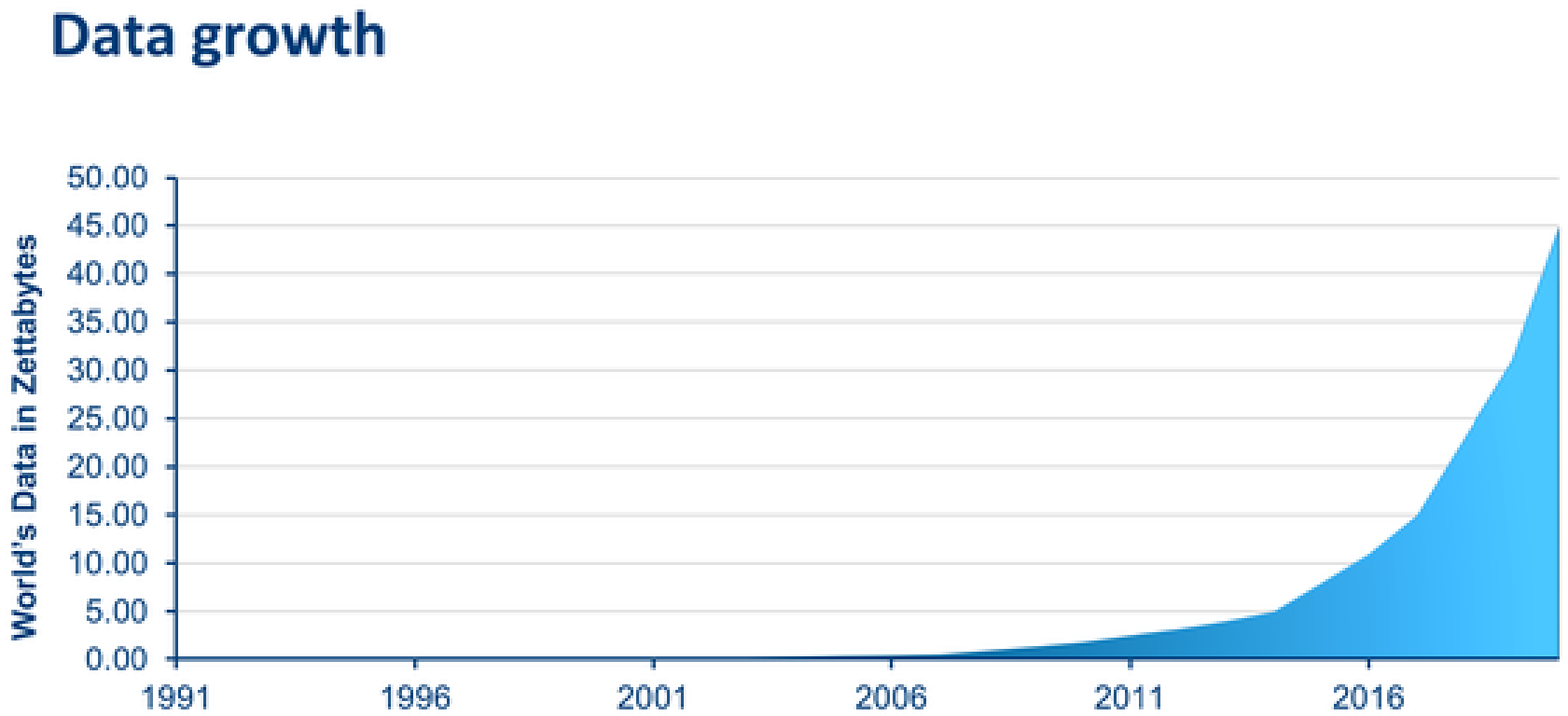
C'est notre objectif durant votre formation "science des données biologiques" qui s'étalera sur 4 année.



UMONS

Pourquoi la science des données ?

- Discipline à la fois ancienne et **récente**
  - Evolution des statistiques, avec ses prémices dans les années 1960 (John Tukey).
  - Emerge comme science à part : 2001 William S. Cleveland, *"Data Science : An Action Plan for Expanding the Technical Area of field of Statistics"*.
  - Le terme **Data Scientist** n'est d'usage courant que depuis 2008.
- Besoin issu de la **quantité de données** disponibles (1 zettabyte = 1 milliard de terabytes = 1 000 000 000 000 000 000 000 octets).

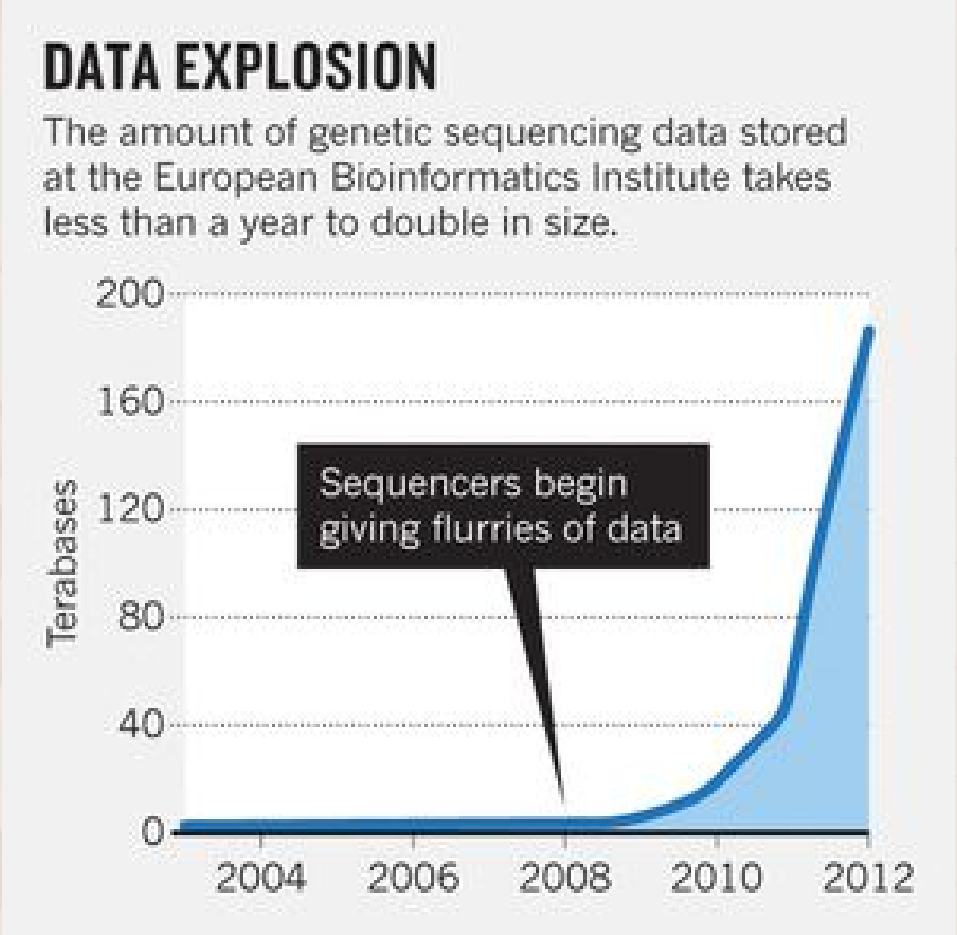


La science de données biologiques

La biologie n'échappe pas au besoin d'analyser des (gros) jeux de données :

- **Génétique**, bases immenses
- **Biodiversité** animale et végétale
- **Etudes écologiques** avec images satellites, capteurs haute vitesse
- **Littérature** scientifique
- etc.

Un biologiste analyse des données pratiquement quotidiennement sous une forme ou l'autre !

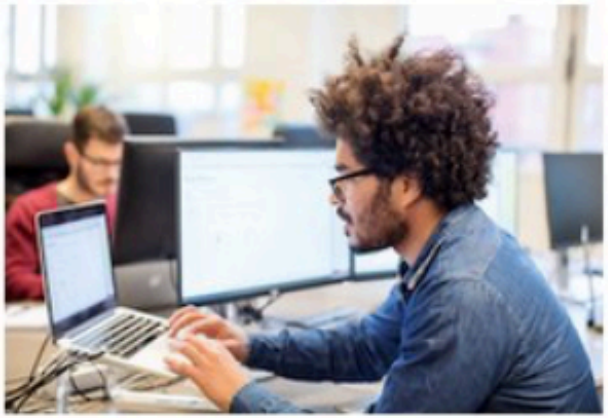


UMONS

Scientifique des données : le meilleur job !

... aux USA, voir https://www.kdnuggets.com/2018/01/glassdoor-data-scientist-best-job-america-3years.html

1 Data Scientist



4.8 / 5
Job Score

4.2 / 5
Job Satisfaction

$110,000
Median Base Salary

4,524
Job Openings

View Jobs

In 2017, Glassdoor also ranked Data Scientist as the best job in America, with the same job score and the same median base salary, but slightly higher job satisfaction. In 2016, Glassdoor ranked Data Scientist as no. 1 job in USA for the first time, with median salary $117K, and about 1,700 jobs listed.

Here are the top 5 jobs in 2018 and also other jobs in top 50 related to Analytics, Big Data, Data Science, according to Glassdoor:

2018 Rank	Job Title	Job Score	Job Satisfaction	Median Base Salary	Job Openings
1	Data Scientist	4.8	4.2	$110,000	4,524

UMONS

Ce que cela signifie pour vous...

- Croissance exponentielle des données = besoin de spécialistes
- Aujourd'hui, un plus, demain une **obligation**
- Tout biologiste a le *même* diplôme. C'est les *spécialisations* qui les différencient
- Un spécialiste en science des données trouve un travail immédiat, intéressant et bien payé !

Prévoyez dès maintenant de rajouter cette compétence dans votre C.V. Nous sommes là pour vous y aider.

UMONS

Approche des cours de science des données biologiques

UMONS

Science des données biologiques à l'UMONS

- Nouveau cours, plus axé sur l'utilisation des outils et moins sur les statistiques de base qu'avant
- Matériel didactique riche et varié :
  - Syllabus en ligne http://biodatascience-course.sciviews.org/sdd-umons/, en évolution permanente,
  - Tutoriels interactifs (dans la machine virtuelle),
  - Capsules vidéos (à venir),
  - Dépôts Github (Classroom) pour les exercices.
  - Matériel complémentaire sur Moodle.
  - Site Web : http://biodatascience-course.sciviews.org
- Travail en **classe inversée** : vous étudiez et préparez chez vous *avant* les séances
- Compléments et travail supplémentaire en séance

Apprendre la science des données, c'est comme apprendre une nouvelle langue : sur la durée et en pratiquant souvent.

UMONS