

# Science des données I



First class meeting

Philippe Grosjean & Guyliann Engels

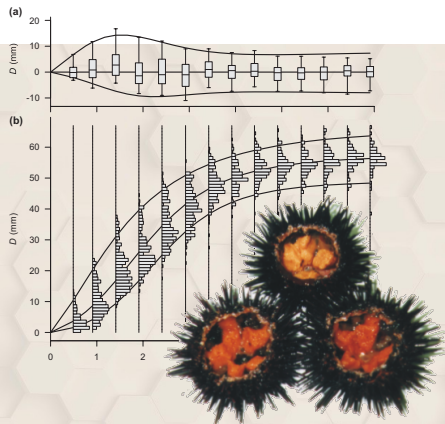
Université de Mons, Belgique  
Laboratoire d'Écologie numérique



<https://wp.sciviews.org>  
[sdd@sciviews.org](mailto:sdd@sciviews.org)

Qui sommes nous ?

# 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

## Guyliann Engels

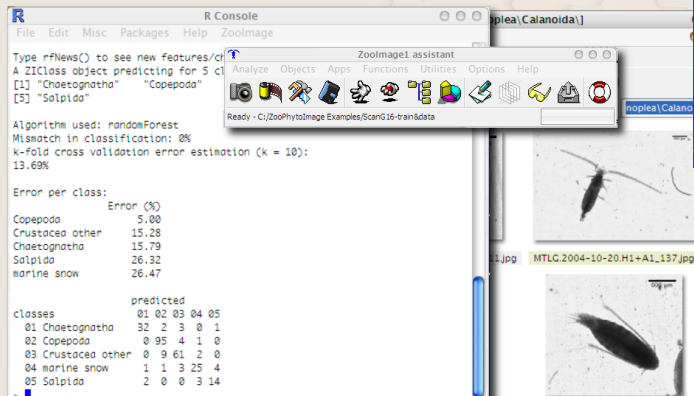
- **Master** en Biologie des Organismes et Écologie à l'UMONS.
- **Mémoire** effectué dans le laboratoire d'Écologie numérique des Milieux aquatiques sur l'écophysiologie et l'écotoxicologie de la posidonie (*Posidonia oceanica*, une plante marine) en Méditerranée.
- **Thèse de doctorat** en cours sur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



# 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).*



The screenshot displays the R Console window with the following output:

```

Type rfNews() to see new features/changes
A ZIClass object predicting for 5 classes
[1] "Chaetognatha"    "Copepoda"
[5] "Salpida"

Algorithm used: randomForest
Mismatch in classification: 0%
k-fold cross validation error estimation (k = 10):
13.69%

Error per class:
              Error (%)
Copepoda      5.00
Crustacea other 15.28
Chaetognatha  15.79
Salpida       26.32
marine snow   26.47

predicted
classes 01 02 03 04 05
01 Chaetognatha 32  2  3  0  1
02 Copepoda     0 95  4  1  0
03 Crustacea other 0  9 61  2  0
04 marine snow  1  1  3 25  4
05 Salpida      2  0  0  3 14
  
```

Overlaid on the R Console is the ZoomImage assistant window, which includes a toolbar with icons for Analyze, Objects, Apps, Functions, Utilities, Options, and Help. Below the toolbar, two plankton images are shown for analysis. The top image is labeled '1.jpg' and the bottom image is labeled 'MTLC.2004-10-20.H1+A1\_137.jpg'. A scale bar in the bottom right of the second image indicates '50 µm'.



# Science des données

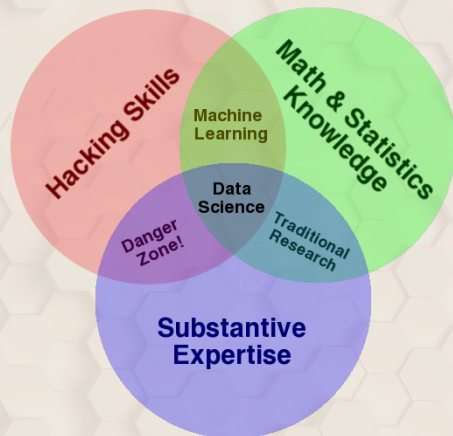
## Science des données : une approche pragmatique

A data scientist is a  
statistician who is useful.  
— *Hadley Wickham*

JS

# 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.

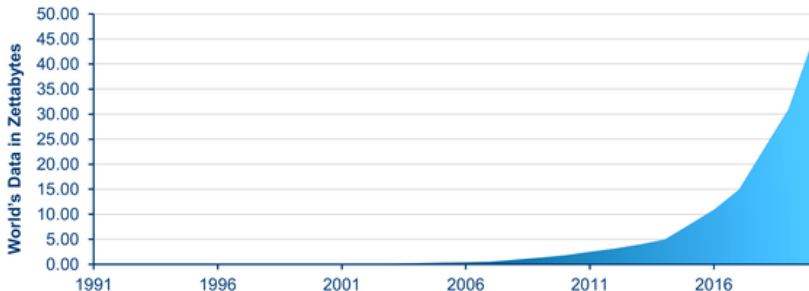




# Pourquoi la science des données ?

- 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 octets).

## Data growth



# La science de données biologiques

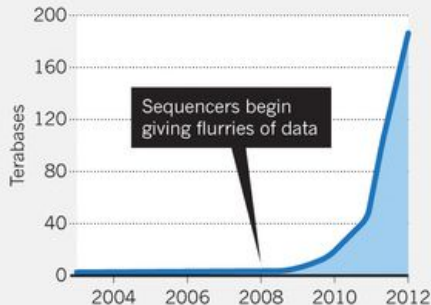
La biologie n'échappe pas au besoin d'analyser des (gros) jeux de données :

- **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 !

## DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.



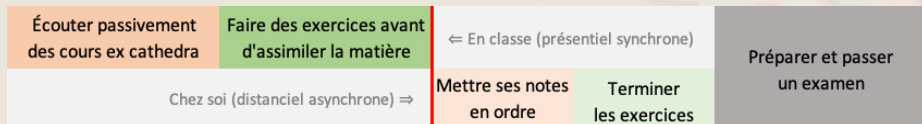
# Objectifs

Nos objectifs principaux durant votre formation en **science des données biologiques** est de vous former afin d'être capable de :

- réaliser des analyses biologiques usuelles,
- présenter clairement ses résultats de manière reproductible avec des outils informatiques et statistiques professionnels
- développer votre esprit critique

**Vous pouvez retrouver via la fiche du cours le détails de tous les objectifs**  
**<http://applications.umons.ac.be/web/fr/pde/2022-2023/ue/US-B2-SCBIOL-006-M.htm>**

# Comment se déroule un cours ?

Cours *ex cathedra* + Travaux pratiques

- Le réel apprentissage se déroule **après** les séances de cours et d'exercices
- Un examen est nécessaire pour vérifier vos acquis

D'après la science lors d'un cours *ex cathedra*, **vous n'apprenez rien**

-> Quelle perte de temps :(

# Analyse de l'apprentissage

*Learning Analytics (LA) : can be defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Lang, Siemens, Wise, & Gasevic, 2017)*

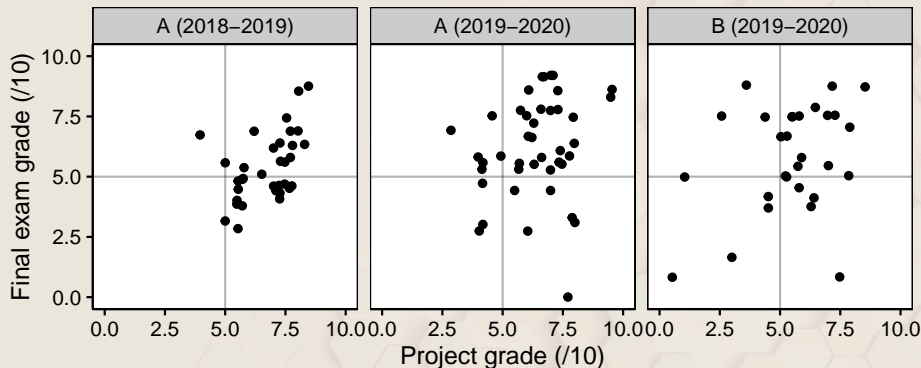


Collecter

Analyser

Agir

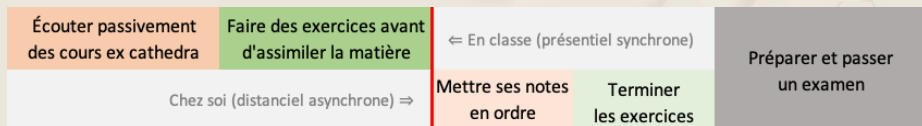
# Examens



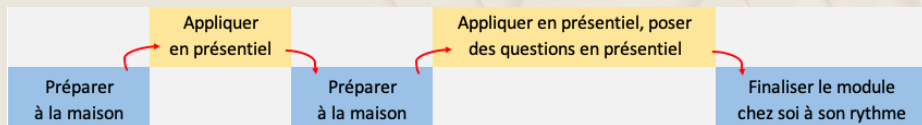
-> les examens ne permettent pas d'évaluer correctement un étudiant

# Pédagogie active et classe inversée

## Cours classique *ex cathedra* + séances d'exercices



## Approche en classe inversée



- *Aucune* séance en présentiel sans préparation
- Chaque heure de travail pleinement consacrée à l'apprentissage
- Vous êtes actifs **tout le temps** et vous gérez à **votre rythme**
- **Pas besoin d'un examen à la fin** : travail évalué dans sa globalité



C'est quoi la classe inversée ?

Vous avez une minute ?

Pour comprendre  
La classe inversée

0:03 / 1:18



(lien vers la vidéo)

## Classe inversée et pédagogie active

Notre approche : **pédagogie active en classe inversée** (vous apprenez *d'abord* à la maison, nous appliquons *ensuite* en présentiel -quand on n'est pas confinés-).

*I hear and I forget.*

*I see and I remember.*

*I do and I understand.*

— Confucius

# C'est quoi la pédagogie active ?

Les  
pédagogies  
actives  
pourquoi ne  
pas essayer?



0:00 / 3:05



(lien vers la vidéo)

UMONS

## Et moi, je fais quoi dans tout cela ?



- Vous êtes **acteur de votre apprentissage**, les enseignants sont des **facilitateurs** (plus en retrait par rapport à l'approche classique).
- Plus de séparation entre **cours théorique** et **exercices** ; vos échanges avec le professeur et le ou les assistants sont similaires.
- Les **élèves-assistants** sont coachés tout autant que vous pour vous faciliter l'apprentissage de manière active.
- Vous posez les **questions**, et vos enseignants vous répondent **individuellement**.

# Organisation du cours

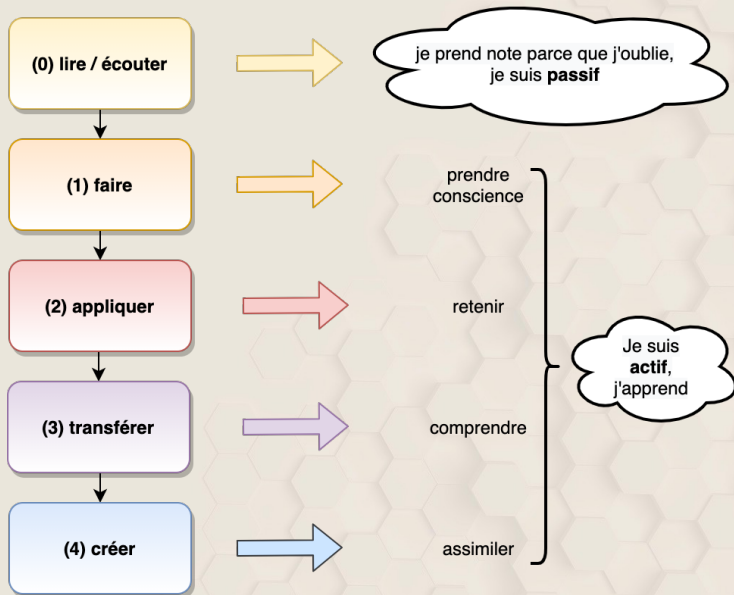
# ECTS

**European Credits Transfer System**, créé en 1988 de manière standardisée par la Commission Européenne comme correspondant à une **charge de travail totale** pour l'étudiant. Ce cours comprend 12 modules pour 6 ECTS (0.5 ECTS/module)



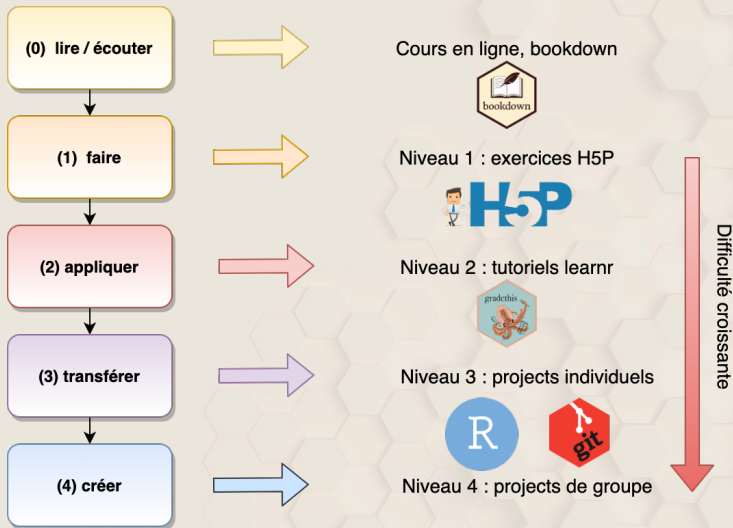
# Apprentissage en 4 niveaux

## Niveaux d'exercice = Apprendre, niveaux 1 à 4

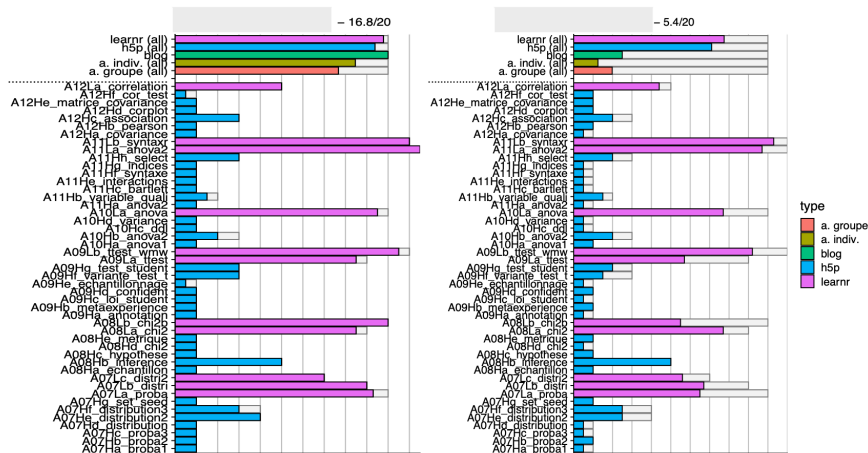




# Quatre niveaux d'exercices



# Construction de la note



■ Vous pouvez suivre votre progression sur moodle ou en fin du module de cours

# Amélioration continue du cours

# Être acteur de l'amélioration

Tout au long de ce cours, nous vous demanderons votre avis ou votre ressenti sur une exercice, sur un chapitre ou encore sur une présentation.

**Soyez honnête et constructif**

Avez-vous des questions ?



### Ressources utiles :

- Site web du cours : <https://wp.sciviews.org/>
- Cette présentation : [https://github.com/BioDataScience-Course/sdd\\_lessons/tree/2022-2023/A01/presentations](https://github.com/BioDataScience-Course/sdd_lessons/tree/2022-2023/A01/presentations)