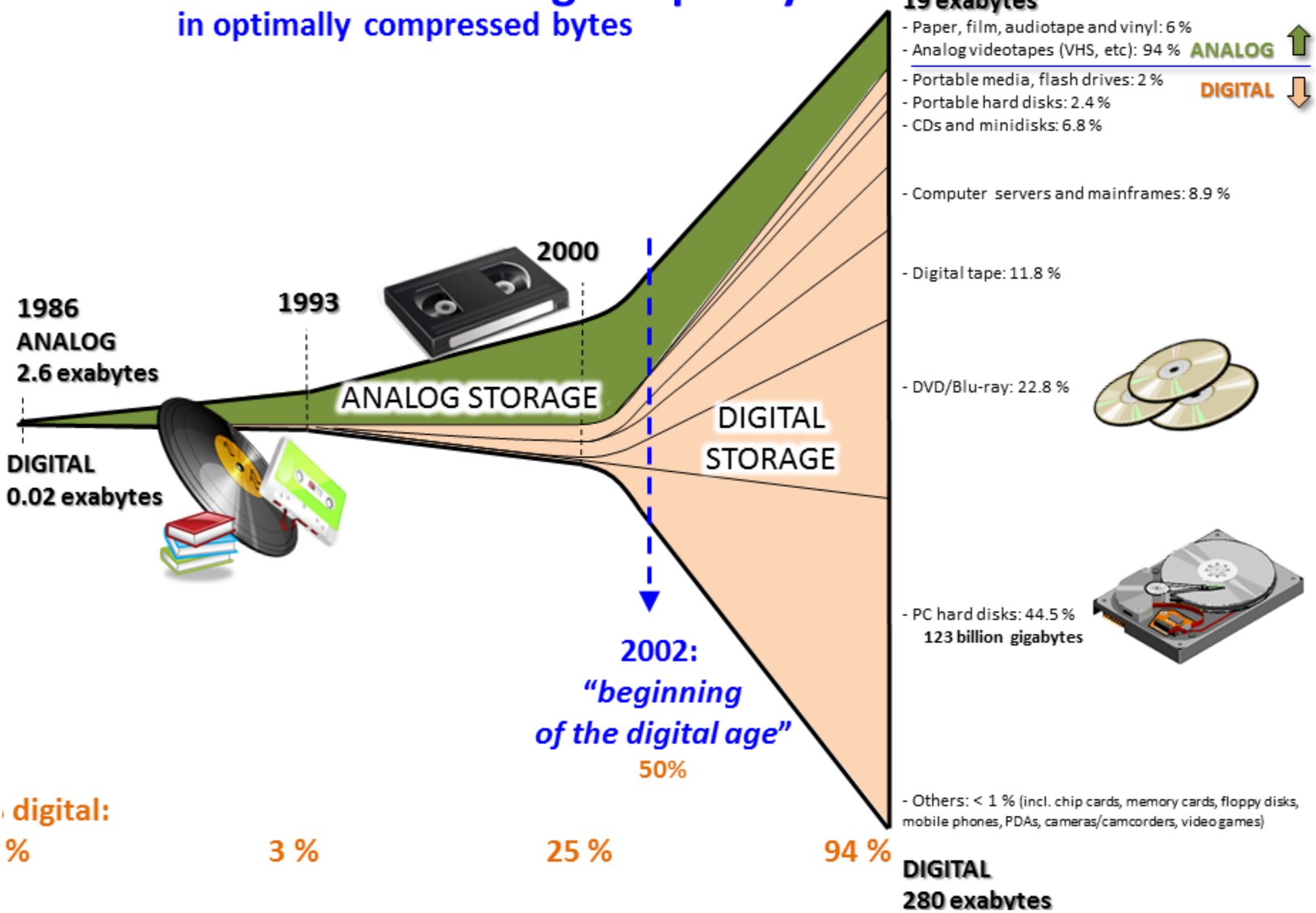
Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

Global Information Storage Capacity

in optimally compressed bytes



19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %

- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

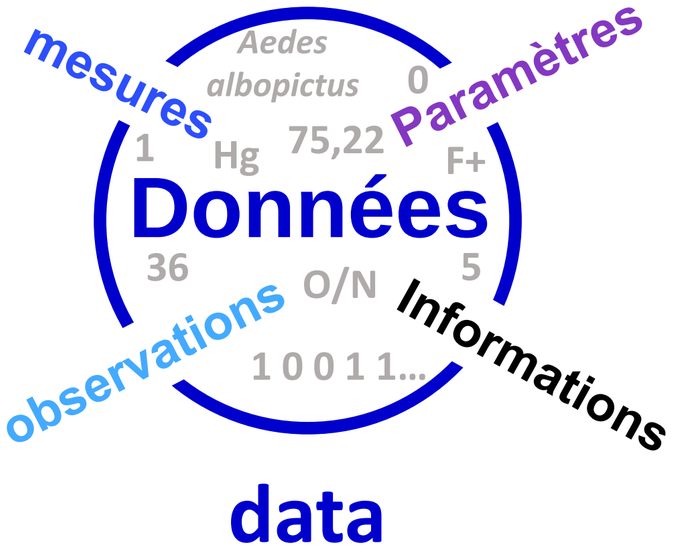
Le volume des données stockées est en pleine expansion. Environ 50 zettaoctets en 2021

Nom	Symbole	Valeur
kilooctet	ko	10^3
mégaoctet	Mo	10^6
gigaoctet	Go	10^9
téraoctet	To	10^{12}
pétaoctet	Po	10^{15}
exaoctet	Eo	10^{18}
zettaoctet	Zo	10^{21}
yottaoctet	Yo	10^{24}

BIG DATA

Chaque année, l'Humanité produit plus de 2000 **Exabit** d'informations nouvelles (10^{21} bits, 1000 milliards de milliards de bits), soit 2 Zettaoctets et a déjà stocké **en 2021 plusieurs dizaines de Zettaoctets** d'information (10^{22} bits,10000 milliards de milliards de bits, 1000 milliards de Gigaoctets, l'équivalent de 1000 milliards de disques durs), environ **50 milliards de téraoctets**. Cela a un coût ⇨ **Matière + Energie + Eau**

Produire des connaissances en faisant parler les données



data

[ordinateurs / serveurs]

Data Sciences

[outils & disciplines]

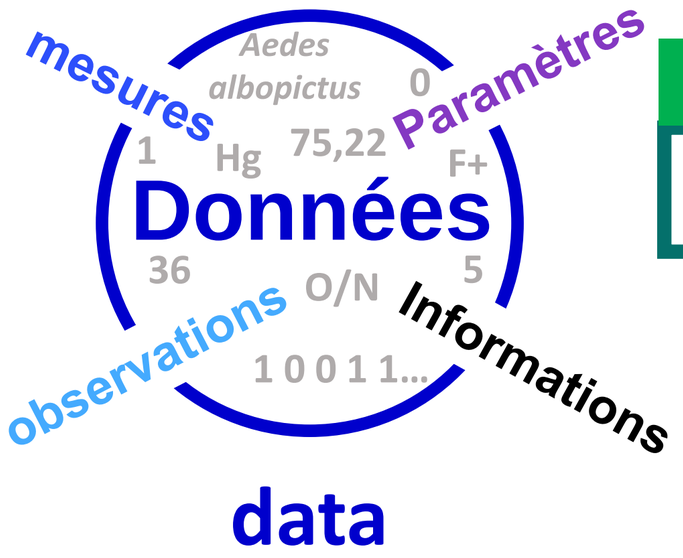
Connaissances

- ⇥ Décisions
- ⇥ Actions
- ⇥ Contrôles
- ⇥ Prévisions

(les data dans « data sciences »)

▷ Approche multidisciplinaire dont les applications dépendent de la puissance des ordinateurs utilisés

Produire des connaissances en faisant parler les données



data

- Les data dans « data sciences »
- Terrain, labos, entreprises,…
- Informations diverses
- Satellites, réseaux, Web

ordinateurs / serveurs

Data Sciences

- Automatisme
- Bases de données
- Informatique
- Mathématiques
- Statistique
- Algorithmique
- Machine learning (*apprentissage*)
- Visualisation et cartographie
- Communication

Connaissances

⇥ **Décisions**

⇥ **Actions**

⇥ **Contrôles**

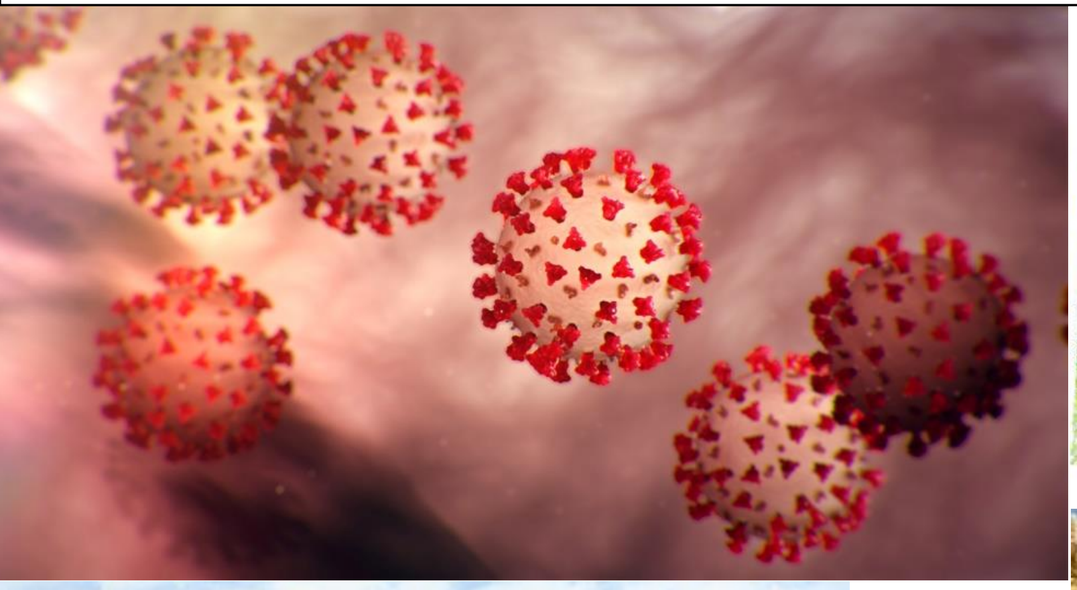
⇥ **Prévisions**

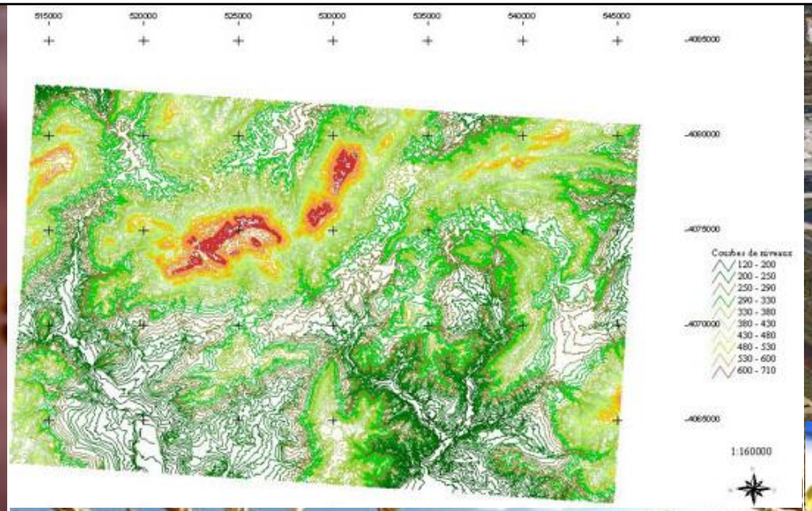
Pilotage voiture autonome,
Développement médicament,
Fidélisation clientelle,…

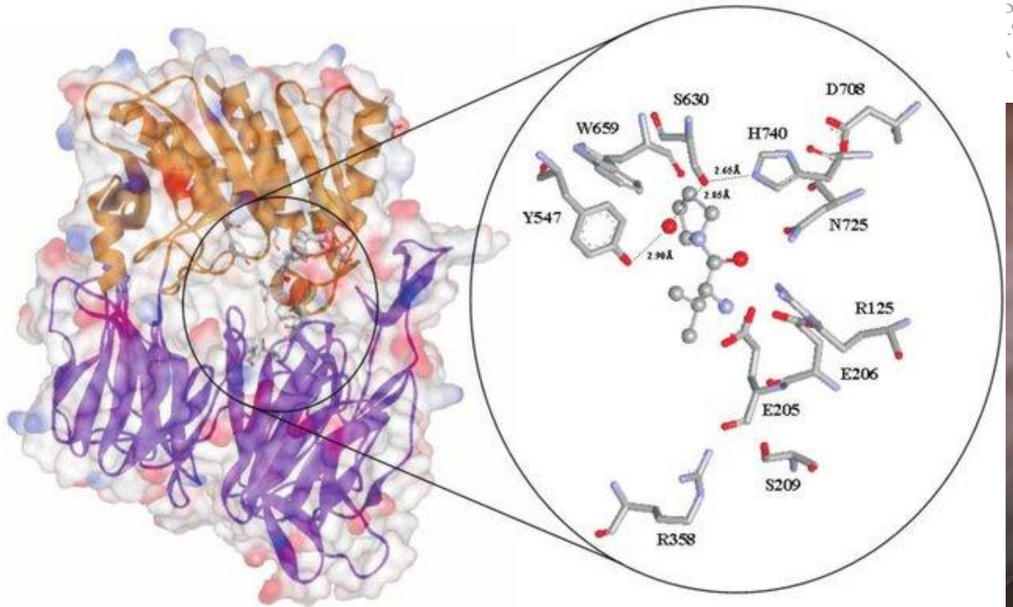
▷ **Gros volume de données / Importante puissance de calcul / Espace de stockage conséquent**

Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)







Data Sciences en Biologie (EN11124 / PR1183, 4 ECTS)

Objectif

Former les étudiants à la **manipulation, au stockage et à l'analyse des données biologiques**

acquises au laboratoire, sur le terrain ou en entreprise et à leur **représentation graphique** dans le contexte du ***Big Data***

Communication des résultats sous forme de présentation synthétique (réalisation d'un rapport de qualité professionnelle)

Une véritable initiation aux **data sciences en biologie**

axée sur la réalisation de **bases de données**, le traitement et l'**analyse des données** et le **machine learning**

Data Sciences en Biologie (EN11124 / PR1183, 4 ECTS)

Objectif

Former les étudiants à la **manipulation, au stockage et à l'analyse des données biologiques**

acquises au laboratoire, sur le terrain ou en entreprise et à leur **représentation graphique** dans le contexte du ***Big Data***

Communication des résultats sous forme de présentation synthétique (réalisation d'un rapport de qualité professionnelle)

Une véritable initiation aux **data sciences en biologie**

axée sur la réalisation de **bases de données**, le traitement et l'**analyse des données** et le **machine learning**

Compétences acquises

- armés pour **répondre aux nouveaux challenges de l'analyse des données en biologie** ;
- capables de **prendre en charge un travail de data sciences** ;
- on pourra vous confier **la responsabilité d'un projet de data sciences** dans le cadre de votre métier dans le secteur privé comme dans le secteur publique (notamment Recherche Scientifique).

Data Sciences en Biologie (EN11124 / PR1183, 4 ECTS)

Objectif

Former les étudiants à la **manipulation, au stockage et à l'analyse des données biologiques**

acquises au laboratoire, sur le terrain ou en entreprise et à leur **représentation graphique** dans le contexte du ***Big Data***

Communication des résultats sous forme de présentation synthétique (réalisation d'un rapport de qualité professionnelle)

Une véritable initiation aux **data sciences en biologie**

axée sur la réalisation de **bases de données**, le traitement et l'**analyse des données** et le **machine learning**

Compétences acquises

- armés pour **répondre aux nouveaux challenges de l'analyse des données en biologie** ;
- capables de **prendre en charge un travail de data sciences** ;
- on pourra vous confier **la responsabilité d'un projet de data sciences** dans le cadre de votre métier dans le secteur privé comme dans le secteur publique (notamment Recherche Scientifique).

Un plus pour votre CV

Acquisition d'une **compétence très demandée** (recherche / ingénierie / entreprise)

Data Sciences en Biologie (EN11124 / PR1183, 4 ECTS)

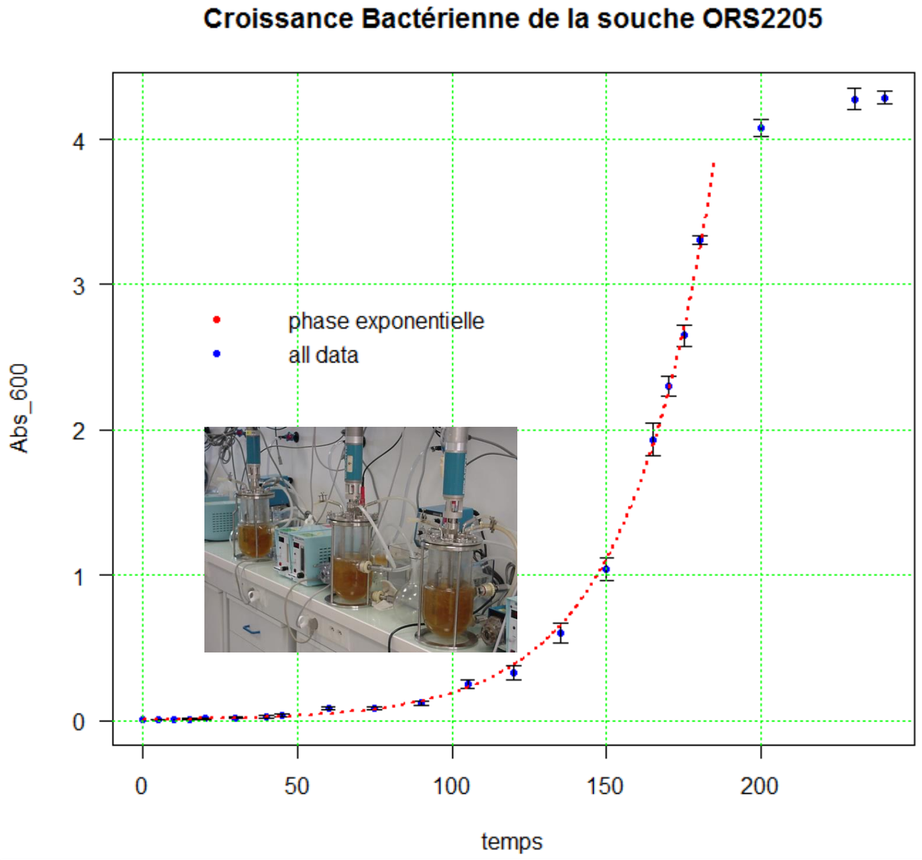
- Volume horaire : **40 h - Cours/Ateliers Pratiques/Projets**
- Travail sur des **jeux de données réels**
- Thématiques : Environnement, Epidémiologie, Biomédical, Pollution, Agronomie, Aide au diagnostique…
- Apprentissage d'une démarche
- **L'accent** est mis sur **la pratique** ⇨ ***« mains à la pâte »***
- Travail organisé en **groupes de 4 en salle informatique**
- **Utilisation de logiciels** dédiés
- **Jeux** individuels et par équipe
- **Préparation et avancement des projets en classe**
- Vous servira dans vos **stages** *ainsi que dans la* ***suite de votre cursus*** *et dans votre* ***métier***
- **Apprendre à raisonner** avant d'appliquer le modèle adapté à la problématique.

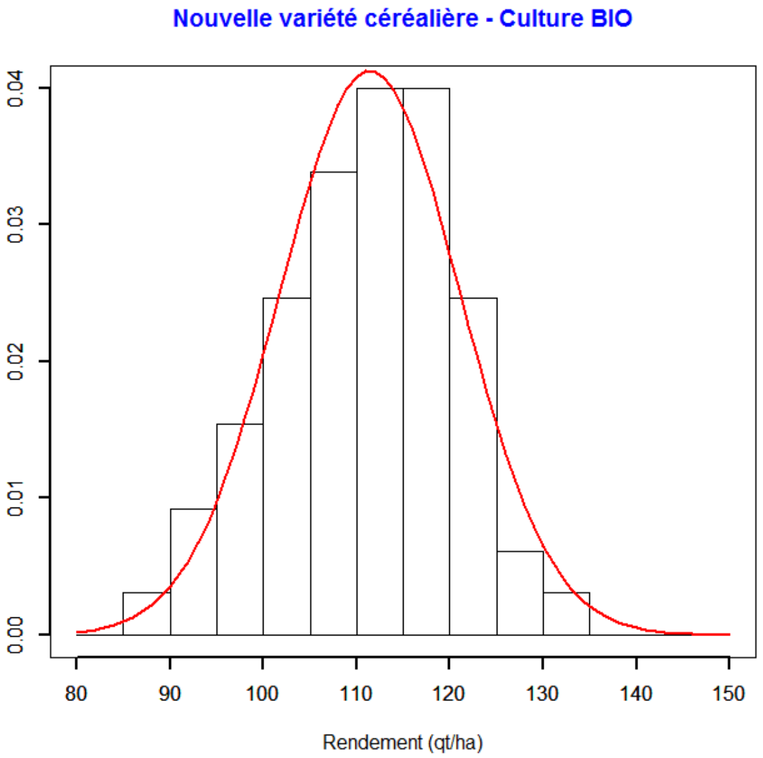
Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

La représentation graphique

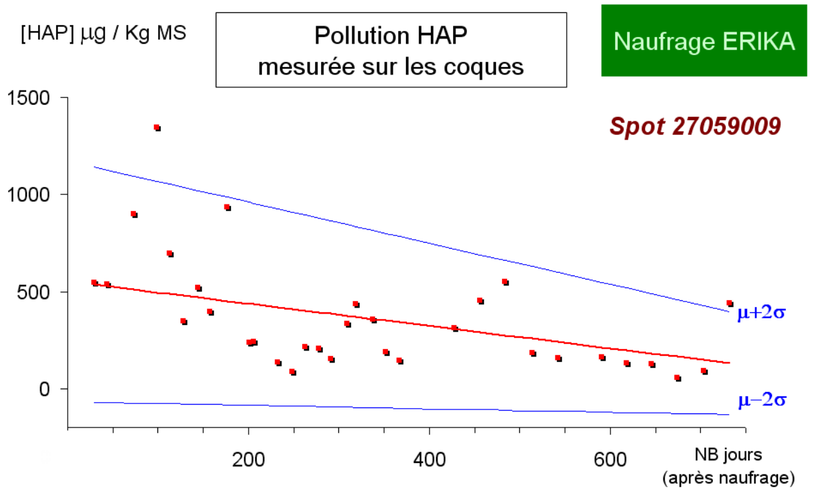
- méthode
- gestion des barres d'erreurs
- apprentissage de logiciels

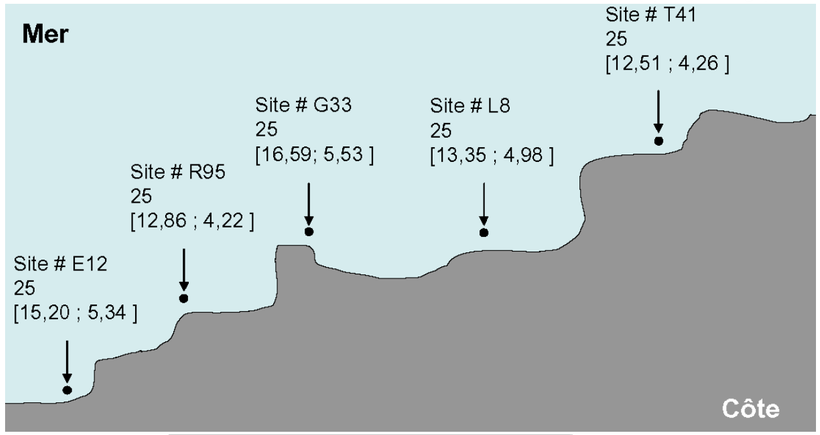




Col	1	2	3	4	5	6	7
Line	temps	DO1	DO2	DO3	DO4	DO5	DO6
1	0	0.005	0.005	0.003	0.005	0.005	0.005
2	5	0.004	0.008	0.006	0.007	0.005	0.006
3	10	0.008	0.007	0.007	0.006	0.006	0.007
4	15	0.007	0.005	0.005	0.005	0.013	0.012
5	20	0.005	0.013	0.010	0.014	0.010	0.016
6	30	0.012	0.022	0.016	0.021	0.018	0.017
7	40	0.024	0.036	0.031	0.018	0.018	0.012
8	45	0.042	0.021	0.028	0.038	0.038	0.033
9	60	0.082	0.080	0.067	0.090	0.091	0.073
10	75	0.086	0.084	0.067	0.096	0.099	0.080
11	90	0.097	0.132	0.139	0.112	0.118	0.121
12	105	0.258	0.217	0.251	0.214	0.254	0.303
13	120	0.371	0.327	0.273	0.276	0.326	0.380
14	135	0.583	0.666	0.580	0.624	0.499	0.657
15	150	1.000	0.920	1.104	1.072	1.135	1.014
16	165	1.917	1.892	1.825	2.137	1.894	1.937
17	170	2.410	2.267	2.250	2.245	2.286	2.361
18	175	2.599	2.748	2.585	2.662	2.593	2.708

Modèle Linéaire Généralisé



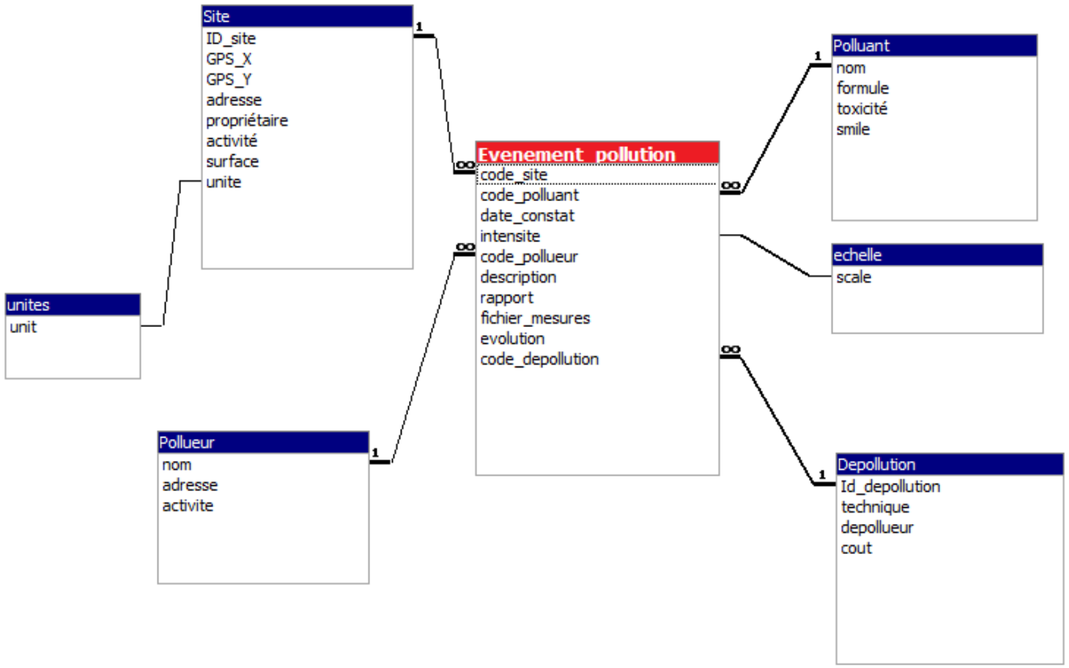


Activité enzymatique	
Individus Témoins	Individus exposés pendant 2 jours à une dose d'insecticide
27.0	26.0
28.8	19.0
26.0	27.1
34.5	19.6
28.8	22.0

▷ Représentation graphique

Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)



Formulaire_accueil | cartographie | region_bretagne | **Evenement_pollution** | Depollution | Pollueur

Nom du champ	Type de données	Description (facultative)
code_site	Numérique	code pointant sur l'index du site
code_polluant	Texte court	code pointant sur l'index du polluant
date_constat	Date/Heure	date du constat de a pollution (non celle à laquelle la pollution s'est déclanchée)
intensite	Numérique	échelle sur 5 niveaux décrivant la gravité de la pollution (de 1: faible à 5 catastrophiqu
code_polluteur	Texte court	pointe vers l'ID du polluteur
description	Texte court	description synthétique de la pollution à la date du constat (temps t0)
rapport	Lien hypertexte	Rapport complet contenant la description, l'évolution, images de la pollution sur le si
fichier_mesures	Lien hypertexte	fichier excel/csv (exploitable par R) compilant l'évolution des concentration de pollu
evolution	Texte court	compilation formatée (separateur ";") des données d'évolution
code_depollution	Numérique	pointe vers la dépollution sélectionnée

code_site	code_pollua	date_constat	intensite	code_polluteur
1	benzène	08/01/2020	4	TAUTALE
1	mercure	08/01/2020	2	CHAILE
2	mercure	08/01/2020	3	TAUTALE
3	Cadmium	12/02/2019	5	Inconnu
3	HAP	09/10/2019	4	
4	benzène	08/11/2018	3	HAP2
4	mercure	05/02/2020	3	CHAILE
4	mercure	07/02/2020	2	
5	Cadmium	05/02/2020	1	
6	Cadmium	05/02/2020	3	TAUTALE
6	mercure	04/02/2020	3	CHAILE
0				

▷ Développement et gestion d'une base de données

Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

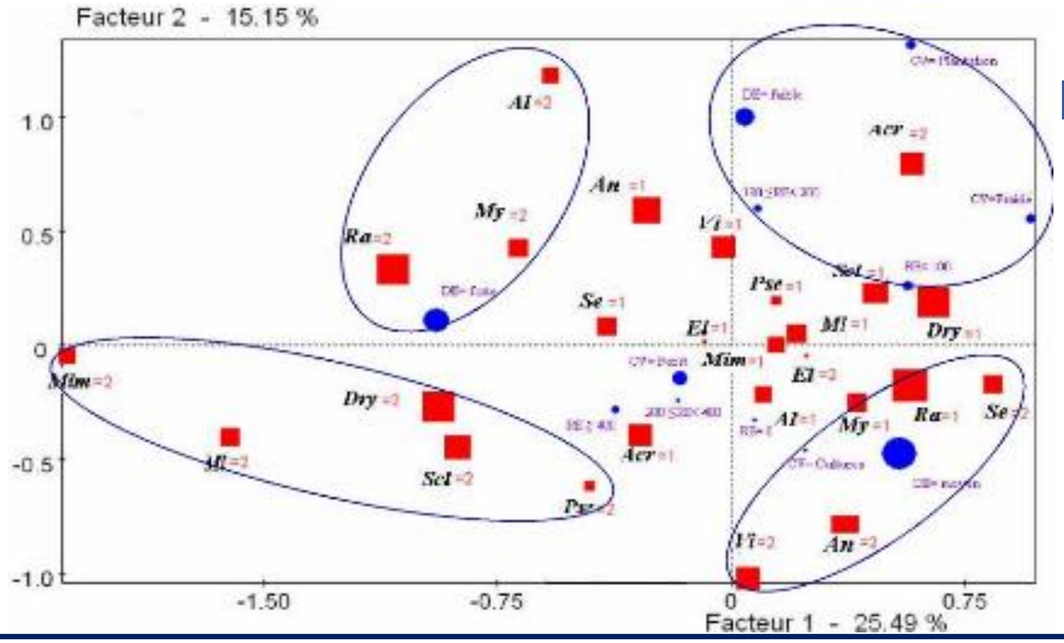
Analyse Factorielle (AFC)

Facteurs influençant la **production piscicole des étangs** (Ouest-Cameroun)

Mise en place d'un étang d'inondation

HISTOGRAMME DES 13 PREMIERES VALEURS PROPRES

NUMERO	VALEUR	POURCENTAGE PROPRE	POURCENTAGE CUMULE	
1	0.2549	25.49	25.49	**********
2	0.1515	15.15	40.64	******
3	0.1087	10.87	51.50	****
4	0.0926	9.26	60.76	****
5	0.0798	7.98	68.74	***
6	0.0661	6.61	75.35	***
7	0.0521	5.21	80.57	**
8	0.0503	5.03	85.60	**
9	0.0430	4.30	89.90	**
10	0.0365	3.65	93.55	**
11	0.0280	2.80	96.35	*
12	0.0208	2.08	98.43	*
13	0.0157	1.57	100.00	*



La **production** piscicole d'un étang **dépend** de **l'environnement végétal** des sites et des **conditions hydrologiques** du cours d'eau qui l'inonde

Décision *Action*

- réhabiliter d'autres étangs afin d'augmenter la surface exploitable ;
- organisation des pisciculteurs autour d'un groupe (partage d'expérience);
- protéger les couverts forestiers car ils assurent les meilleurs productions des étangs

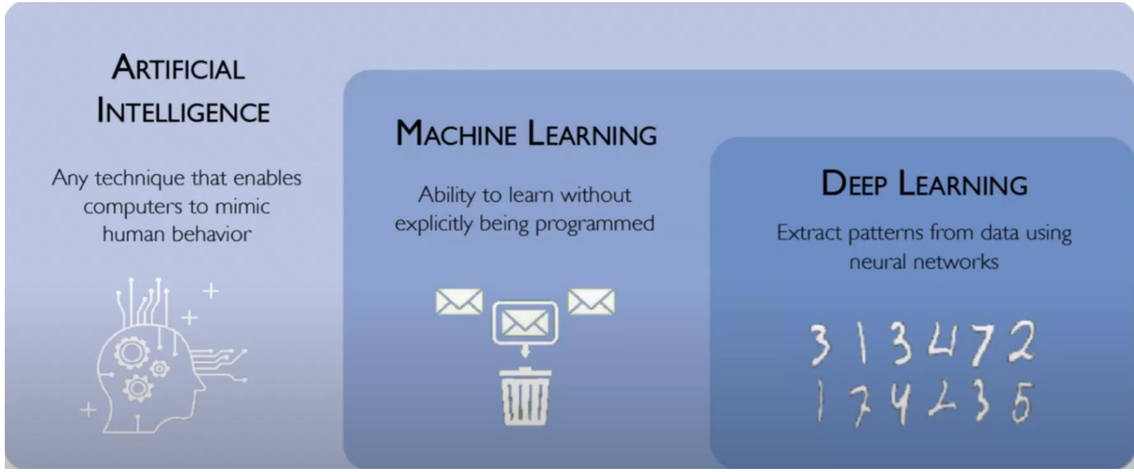
Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

- ⇨ Apprentissage supervisé, non supervisé, semi-supervisé et par renforcement
- ⇨ Quelques grandes familles d'algorithmes :

  Régression logistique, Régression linéaire, Machine à vecteurs de support,

  Arbres de décision, Random Forest, K-means, Model-based, …



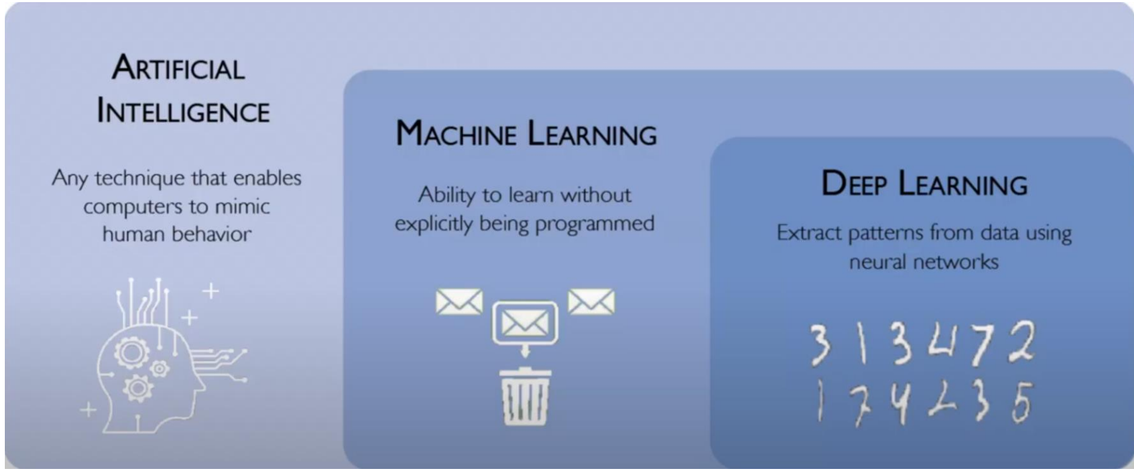
Data Sciences en Biologie

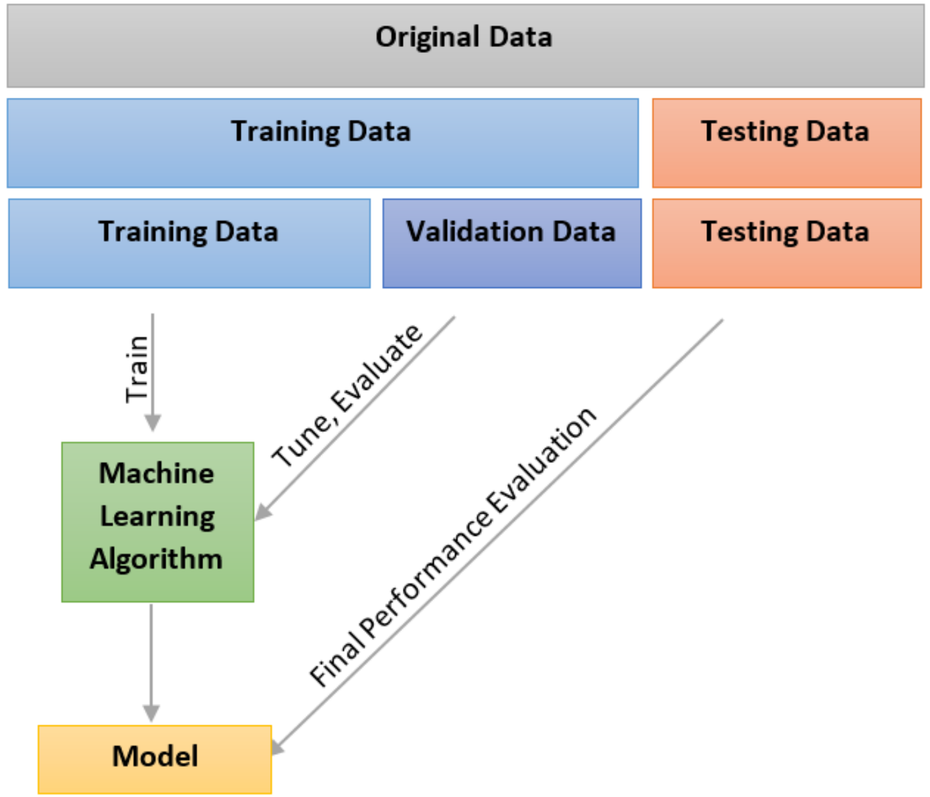
(EN11124 / PR1183, 4 ECTS)

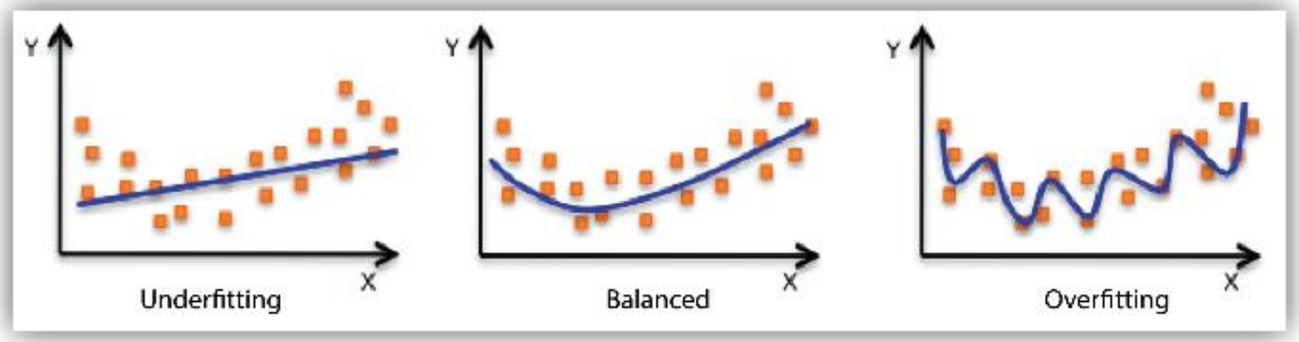
- ⇨ Apprentissage supervisé, non supervisé, semi-supervisé et par renforcement
- ⇨ Quelques grandes familles d'algorithmes :

  Régression logistique, Régression linéaire, Machine à vecteurs de support,

  Arbres de décision, Random Forest, K-means, Model-based, …
- ⇨ Entrainement (algorithmes), data set, training set, validation set et test set







- ⇨ Robustesse, précision et généralisation de la prédiction

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics

# Create a classifier
classifier = LogisticRegression(multi_class='multinomial',
                                penalty='none', solver='newton-cg')

# Train classifier using training dataset
classifier.fit(X_train_scaled, y_train)

# Predict labels for test dataset
y_pred_test = classifier.predict(X_test_scaled)

# Calculate accuracy comparing prediction
# with real labels in test dataset
accuracy = metrics.accuracy_score(y_test, y_pred_test)
```

⇨ Exemple d'applications :

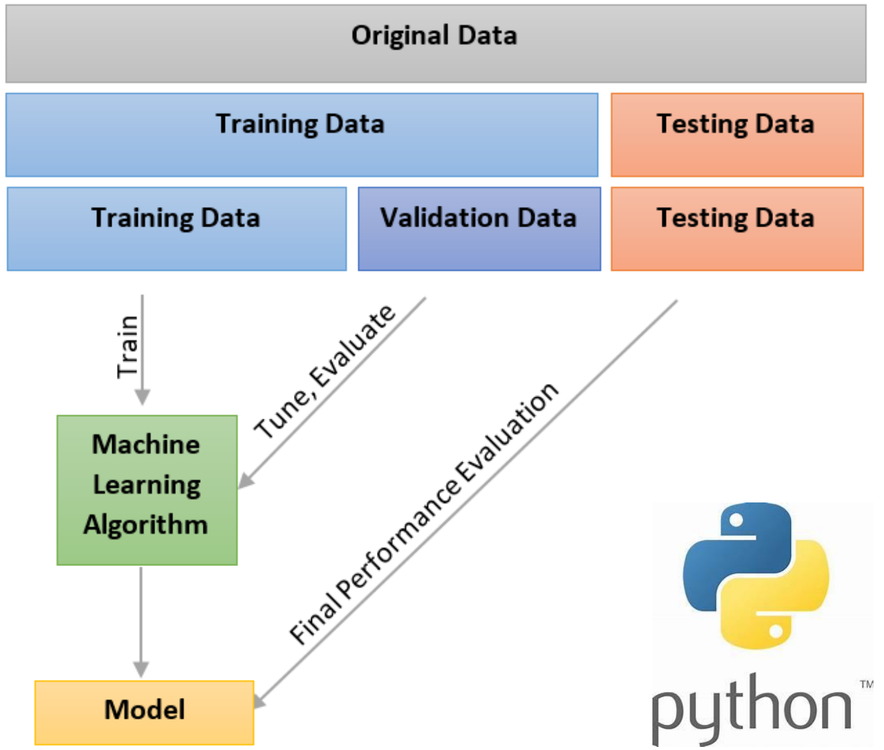
Protection de l'environnement

Mieux comprendre, détecter et protéger

Lutter contre l'extinction des éléphants d'Afrique

L'intelligence artificielle pour suivre la démographie

et le comportement d'animaux en Afrique




```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics

# Create a classifier
classifier = LogisticRegression(multi_class='multinomial',
                                penalty='none', solver='newton-cg')

# Train classifier using training dataset
classifier.fit(X_train_scaled, y_train)

# Predict labels for test dataset
y_pred_test = classifier.predict(X_test_scaled)

# Calculate accuracy comparing prediction
# with real labels in test dataset
accuracy = metrics.accuracy_score(y_test, y_pred_test)
```

Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

⇨ Exemple d'applications :

Protection de l'environnement

Mieux comprendre, détecter et protéger

Lutter contre l'extinction des éléphants d'Afrique

L'intelligence artificielle pour suivre la démographie

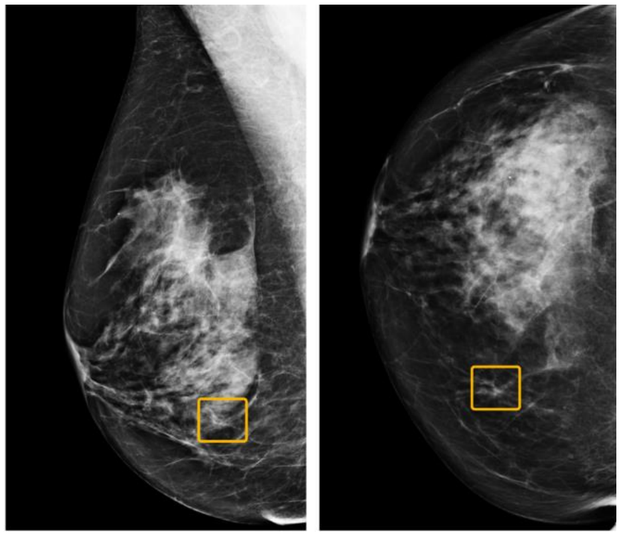
et le comportement d'animaux en Afrique

Diagnostic des cancers précoces

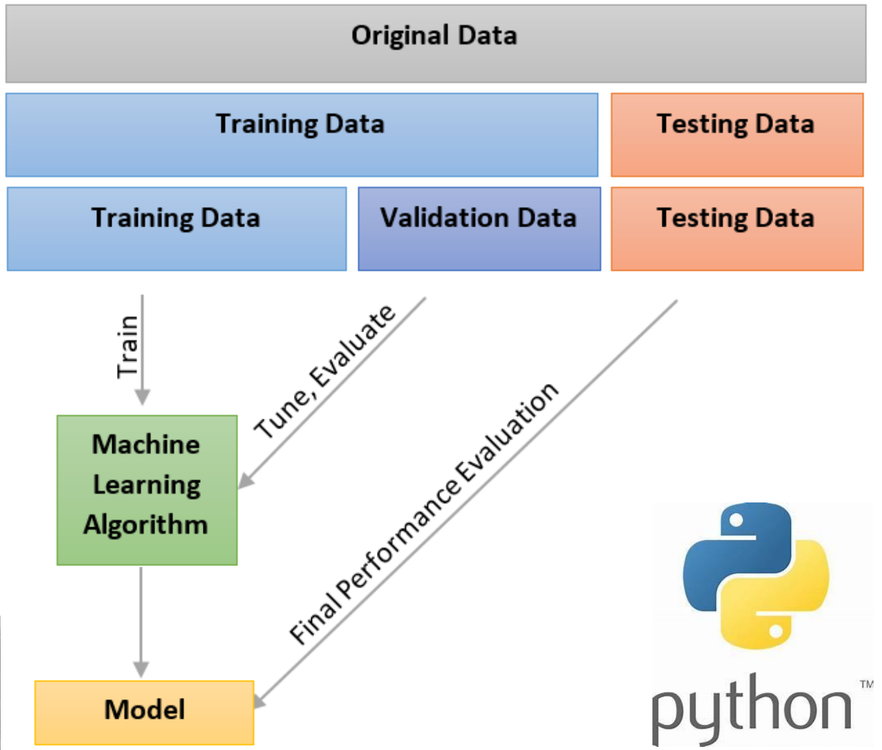
Cas de cancer du sein qui n'a pas été diagnostiqué

par 6 radiologues mais correctement identifié

par un système d'intelligence artificielle



McKinney *et al.* International evaluation of an AI system for breast cancer screening (2020) **Nature**. 577:89-94




```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics

# Create a classifier
classifier = LogisticRegression(multi_class='multinomial',
                                penalty='none', solver='newton-cg')

# Train classifier using training dataset
classifier.fit(X_train_scaled, y_train)

# Predict labels for test dataset
y_pred_test = classifier.predict(X_test_scaled)

# Calculate accuracy comparing prediction
# with real labels in test dataset
accuracy = metrics.accuracy_score(y_test, y_pred_test)
```

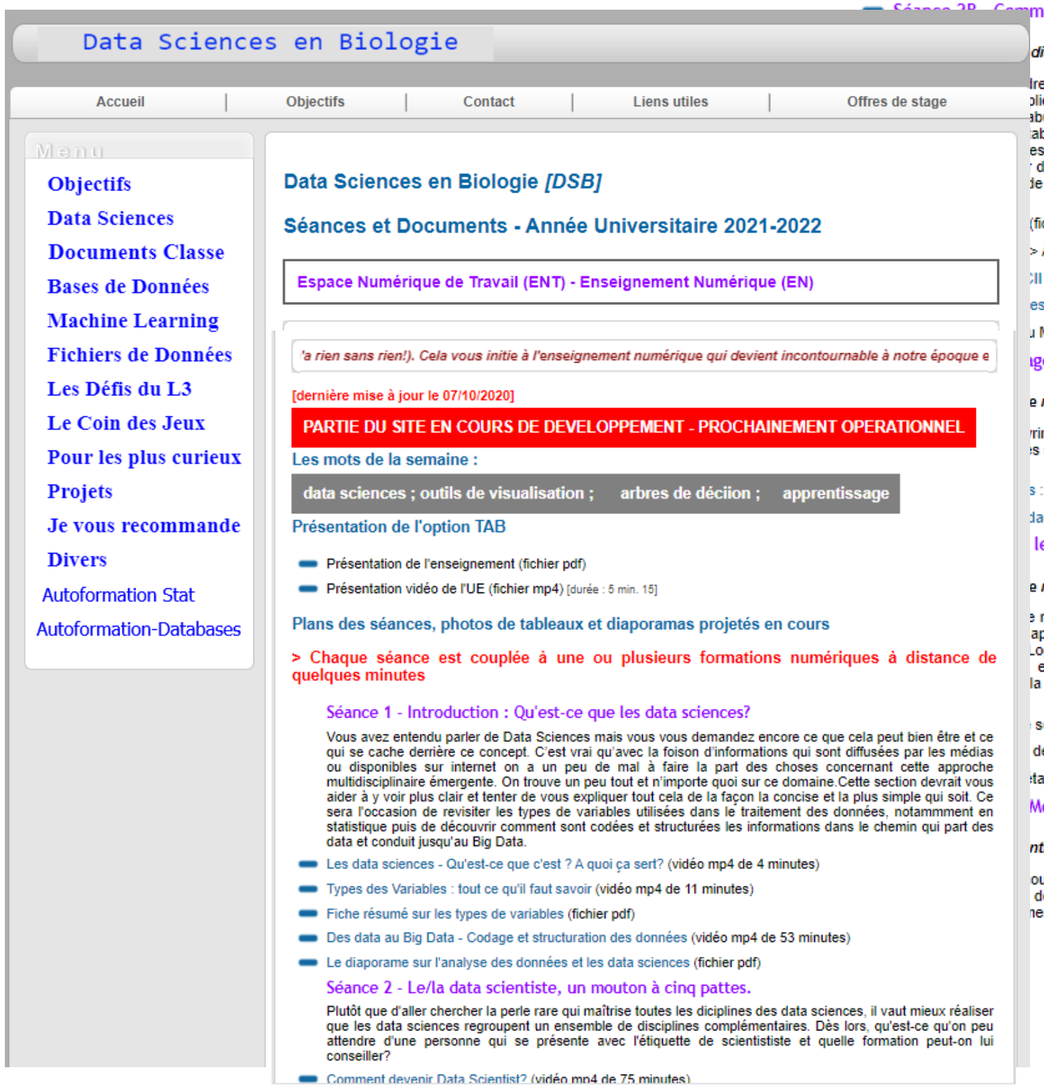
Data Sciences en Biologie

(EN11124 / PR1183, 4 ECTS)

⇨ Un site web où vous retrouverez cours, défis, jeux (équipes), corrigés, auto-évaluations, **avancement des projets**, forum, liaison avec l'enseignant

⇨ vidéos de cours

⇨ Présentation enseignement



Responsable : Pascal RIGOLET
(pascal.rigolet@universite-paris-saclay.fr)

Les Projets, un travail d'équipe

Projets : les 4 compétences

> Base de données,

> Analyse statistique (avec Rcmdr),

> Machine learning (en programmation python ou R),

> Communication

Et

> le/la chef(fe) de projet (une charge et une responsabilité)

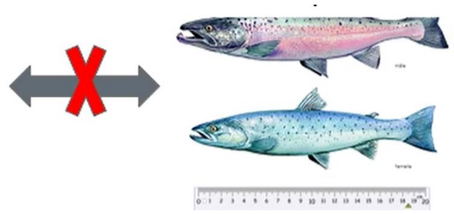
⇨ *rôles très utiles des postit, du tableau blanc (paper board) et du story telling!*

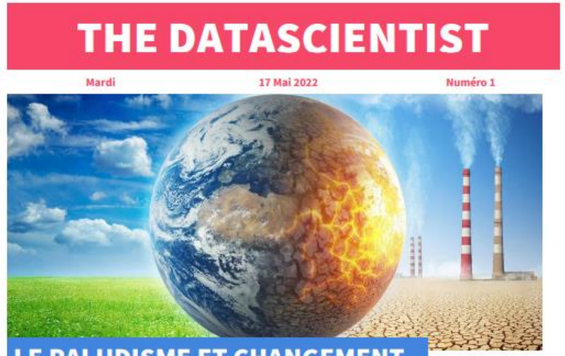
Data Sciences en Biologie - Les Projets - (EN11124 / PR1183, 4 ECTS)

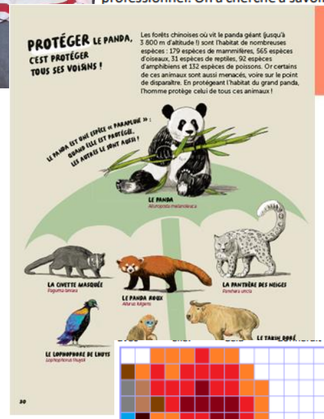
RAPPORT DE RETOUR SUR LE PROJET PODCAST

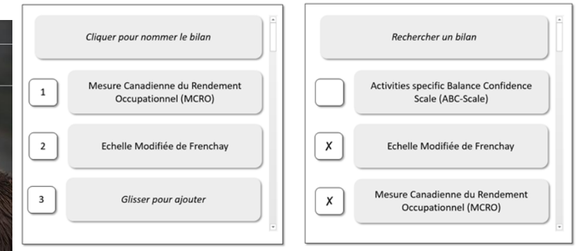
Contexte :

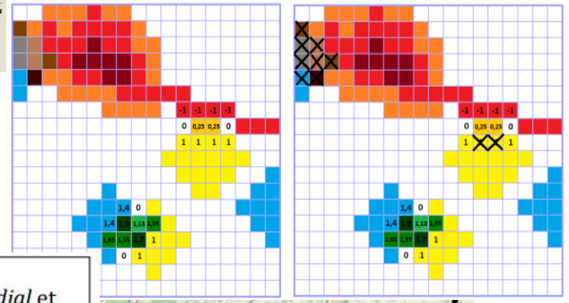
Nous avons tous choisi de participer à ce projet podcast car cela nous permettait d'arbore
autre facette des data sciences. cela permet d'aborder quelques notions mais d'un côté nous a
carte blanche au niveau créativité (nom, style d'écriture, ambiance sonore...) ainsi que de la ma
dont on va aborder la data analyse. Nous avons décidé dans ce second cas de relayer l'expér
d'étudiants ou de professionnels qui sont en lien avec les data au cœur de leurs études ou post
travail, d'où le nom « Danta Science ». Le but de cette émission est de voir comment ces notio
tests statistiques, gestions de données et autres sont utilisées dans le secteur de l'enseigneme
professionnel. On a cherché à savoir comment ces personnes les utilisent, quel est leur avis conce





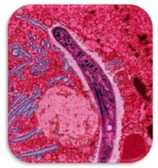






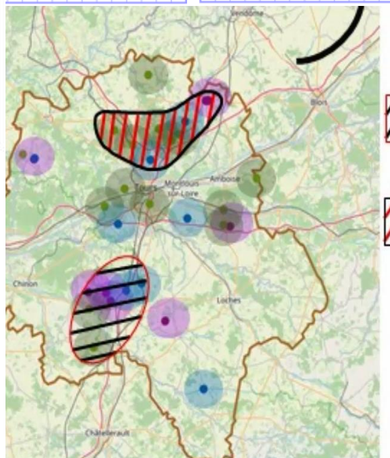
Effets de deux antifongiques chimiques *Polygodial* et *Indoles* sur le développement de Botrytis Cinerea à travers l'exemple de la vigne

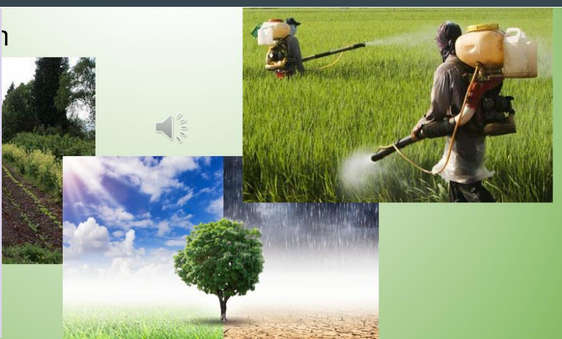
Le paludisr



Plasmodium falciparum

Un moustique anophèle





Data Sciences en Biologie (EN11124 / PR1183, 4 ECTS)

Contrôle des connaissances

- **Réalisation d'un projet de data sciences sur une thématique choisie par le groupe d'étudiants**

  [fichier de données + résultats de l'analyse numérique + rapport écrit (15 pages environ) + présentation orale]
- **Contrôle continu** : défis et jeux savants
- **Première session :** NF = 0,30*CC + 0,70*PROJET
- **Deuxième session :** NF = 0,30*CC[REP] + 0,70*PROJET

Le contrôle des connaissances portera essentiellement sur un **projet** qui mettra l'accent non pas sur des calculs mais sur des stratégies de traitement des données et sur l'analyse de résultats présentés sous forme de graphiques et de tableaux. Ce projet sera démarré en classe (ateliers)

Enseignant intervenant

- Pascal RIGOLET (responsable du module) : **pascal.rigolet@universite-paris-saclay.fr**

Retrouvez la vidéo de présentation de l'enseignement (EN11124 / PR1183, 4 ECTS)

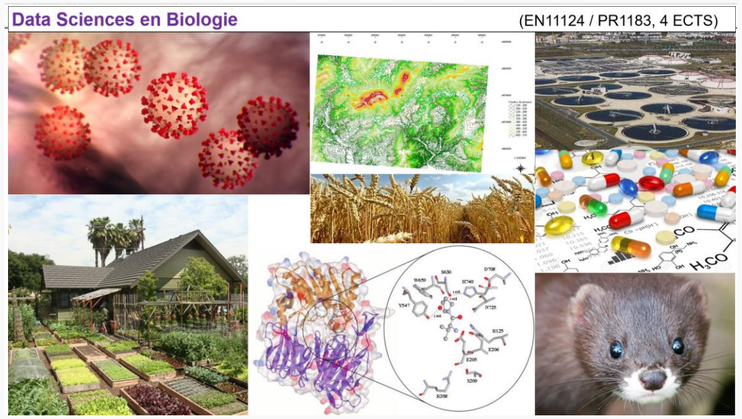
Sur Youtube entrez :

“data sciences biologie”

YouTube FR

UE Data sciences Biologie L3S6

- Accueil
- Explorer
- Abonnements
- Bibliothèque
- Historique
- Vos vidéos
- À regarder plus tard
- Vidéos "J'aime"

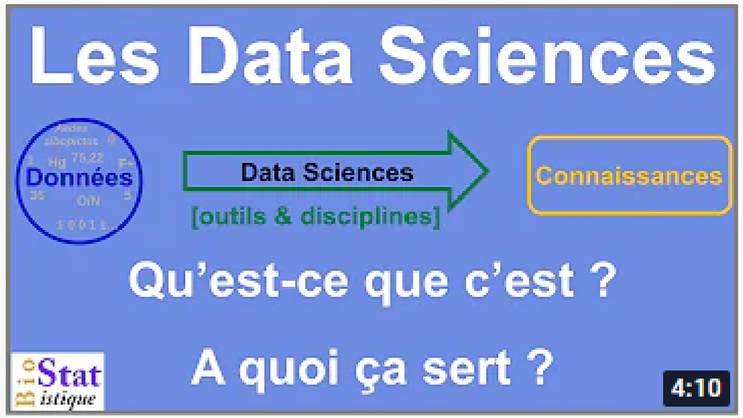


Présentation de l'UE Data Sciences en Biologie du L3S6 (Paris-Saclay)

datascientist

Vidéo présentant en Détail l'UE Data Sciences proposée par Pascal RIGOLET en L3 Biologie (L3S6) à Saclay.

Nouveau



Les data sciences - Qu'est-ce que c'est ? A quoi ça sert?

biostatistique

Les data sciences expliquées en 3 minutes ---------- Vous avez t de ...