Traitement et analyse de données biologiques

Objectifs

- **Redonner le sens pratique** à des étudiants gavés de théorie (« je connais mais ne sais pas l'appliquer »); c'est pourquoi cette UE est très axèe sur la pratique.
- Savoir **se débrouiller seul** (passer à la pratique, repérer la situation, travailler vite et bien). **Etre capable d'analyser ses propres données** expérimentales, de production ou de marketing (*contexte stage, labo, bureau d'étude, entreprise, …*)
- Parler un **langage de base en statistique** (et éviter les excès : *« l'inférence de l'inféron »*) permettant de faire appel aux spécialistes lorsque cela est nécessaire
- Ne pas confondre bioinformatique et biostatistique!
- Vous initier aux **data sciences**
- Essayer de **mieux décrypter notre monde**
- Prendre des **responsabilités** (et des risques mesurés)
- En faire profiter le CV

Plan de travail

- Traitement et analyse de données, le contexte spécifique de la biologie

  (Tout ce que vous avez toujours voulu savoir sur les Biostats, le *Big Data*, les Data Sciences, les *Data Scientist* et autres mots sympathiques)

- L'essentiel sur les statistiques descriptives
- Les graphiques de qualité professionnelle
- Analyse univariée
- Analyse multivariée : analyse factorielle, régression linéaire et ANOVA

Et nous disposons de seulement 25 heures pour tout cela !

Traitement et analyse de données biologiques

Méthode de travail

Nous allons (si possible) travailler en salle de la façon suivante :

Travail en groupe

- Analyse et résolution d'un problème en utilisant les fiches de cours (qui peuvent être données en live au tableau)
- Travailler par groupe de 3 étudiants avec 2 ordinateurs, un pour la recherche (utiliser internet et forum) l'autre pour le calcul
- Résoudre une problématique et rendre 1 fiche d'analyse en fin de certaines séance
- Quizz intervenant n'importe quand (y répondre le plus vite possible)

Salle informatique

- Logiciels à disposition
- Site web de soutien (et vidéos de résumé ?)

Synthèses

- Débriefing
- Corrections faites par l'enseignant et par les étudiants eux-mêmes

- Rapide historique

Statistique provient du latin ***status*** signifiant **état**.

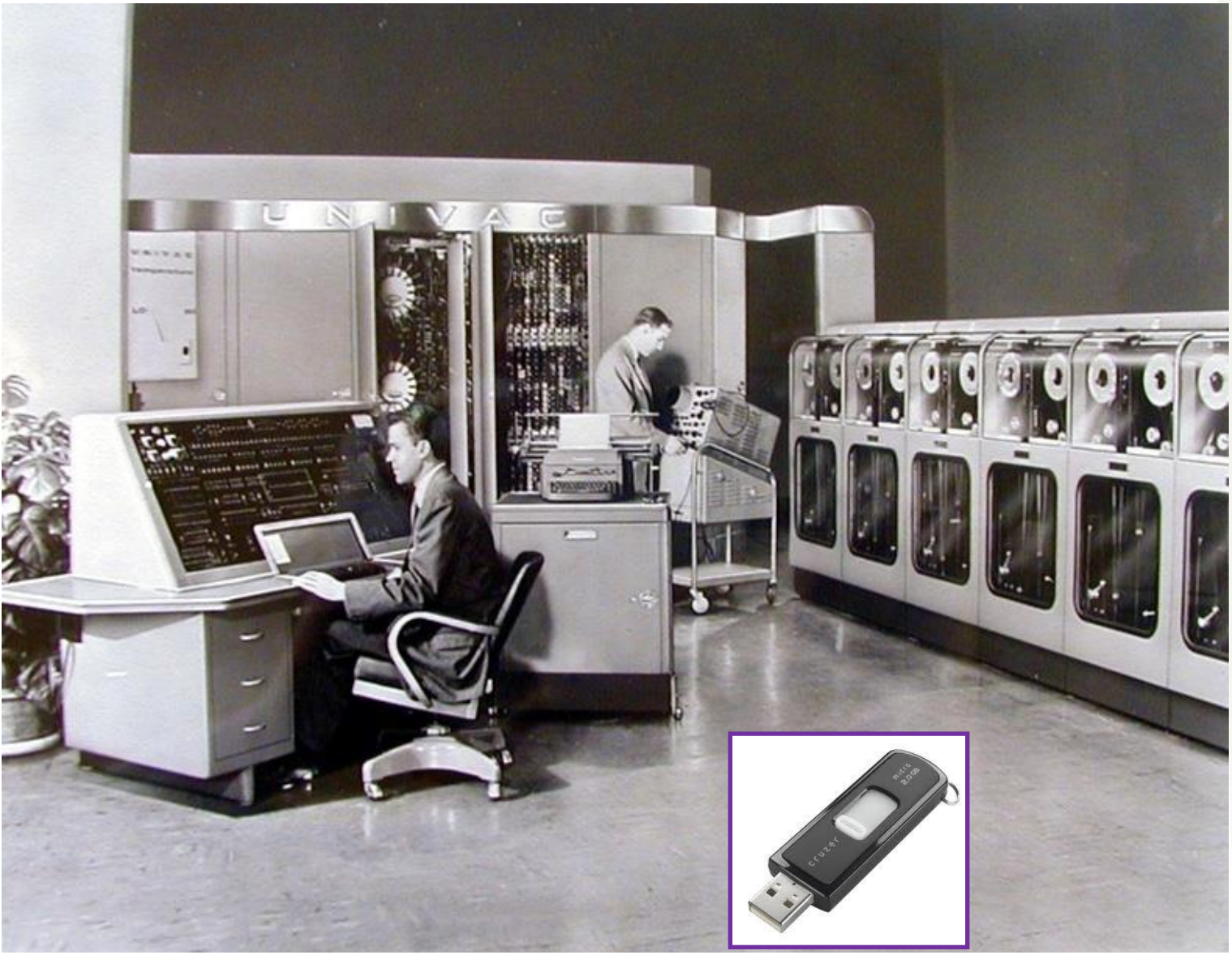
Depuis les temps les plus reculés (Babylone, Égypte, Chine, Grèce, Rome,…), les États voulaient disposer d'informations sur leurs sujets (recensements de population) et sur les ressources qu'ils produisaient ou les biens possédaient *(cf rouleaux de papyrus, colonne ou tablettes d'argile au musée du Louvre).* Les statistiques étaient alors purement descriptives.

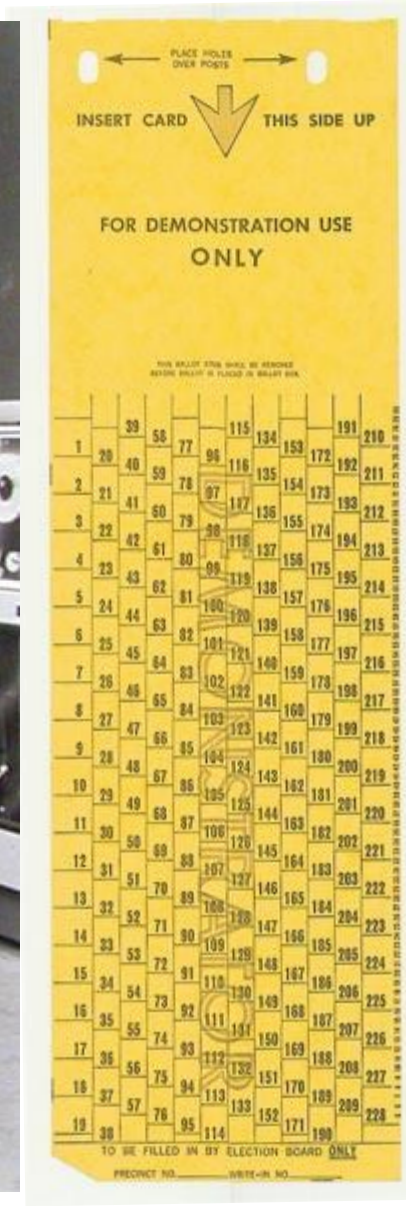
A partir du 17ème siècle s'est développé le calcul des probabilités et des méthodes statistiques sont apparues en Allemagne, en Angleterre et en France. De nombreux scientifiques y ont apporté leur contribution, dont vous avez surement entendu parler : Pascal, Bernoulli, Moivre, Laplace, Gauss, Mendel, Pearson, Fischer, Student, Wilcoxon, …

L'une des missions assignées aux premiers ordinateurs (années 1950) étaient de faciliter le recensement de la population américaine…

Il n'est actuellement pas un domaine ou une discipline, scientifique ou non, qui puisse se passer de l'outil statistique. Où nous mènera le Big Data?

Univac - 1951


UNIVAC
cruzer
micro
2.0GB


PLACE HOLES OVER POSTS
INSERT CARD
THIS SIDE UP
FOR DEMONSTRATION USE
ONLY
TO BE FILLED IN BY ELECTION BOARD ONLY
PRECINCT NO.
WRITE-IN NO.

Questions

Q1 – Qu’est-ce qu’une base de données ?

Q2 – Cela permet-il de faire des stats ?

Q3 - Pouvez vous citer au moins un logiciel de stat ?

Q4 – Comment relier base de données et analyse statistique?

Q5 – Citez une des premières bases de données en Biologie

Système de Gestion de Base de Données (SGBD)
Quelques définitions

- **Une base de données c'est essentiellement une collection structurée d'informations non nécessairement du même type (format)**
- **Une base de données est usuellement localisée en un seul lieu et un seul support (dupliqué en fait) qui est généralement informatique (numérique)**
- **Pièce centrale des dispositifs informatiques qui servent à la collecte, le stockage et l'utilisation des informations**
- **SGBDR, acronyme de Système de Gestion de Base de Données Relationnelles : logiciel moteur qui pilote la base et en permet la manipulation et l'exploitation (interrogation).**
- **Toutes les secondes le volume d'information ne cesse d'augmenter contribuant au Big Data. Mais sans analyse et sans base de données (contribuant à préparer les données à leur analyse), le Big Data n'est rien.**

BIG DATA

Définition

Ensembles de données tellement volumineux qu'ils deviennent difficiles à manipuler et à analyser avec des outils classiques de gestion de base de données ou de gestion de l'information. L'outil statistique doit être repensé pour eux (*data mining* = fouille de données; techniques d'apprentissage,…). On ne sait pas vraiment ce que l'on va y chercher. On essaye de trouver un sens à cette masse considérable d'information (en Zétaoctets).

Synonymes

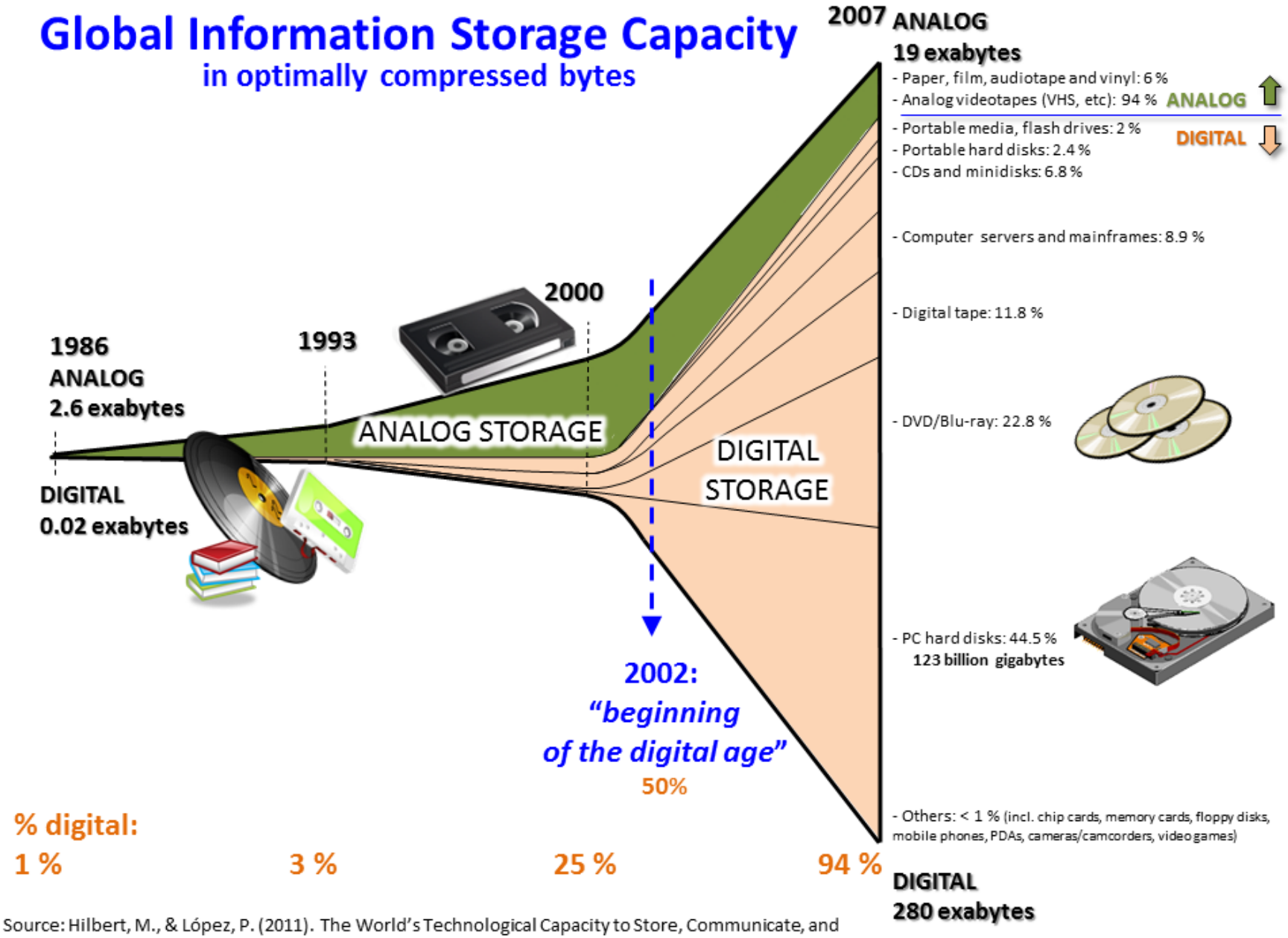
Mégadonnées, données massives, datamasse

Analyse du Big Data

Stockage : l'accès se fait via le réseau (cloud computing). Le Big Data s'accompagne du développement d'applications à visée analytique, qui traitent les données pour en tirer du sens. Ces analyses sont appelées Big Analytics («broyage de données»).

Historique

BIG DATA



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

BIG DATA

Les 5 V

Croissance des données « tri-dimensionnelle » :

les analyses complexes répondent à la règle dite «des 5V»

Volume, Vélocité (Vitesse), Variété, Véracité, Valorisation

Il faudrait 3 ans pour analyser 3 minutes de flux internet (maximum stockable)

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}

Volume

Le volume des données stockées aujourd'hui est en pleine expansion

Environ 20 zettaoctets en 2020

Vélocité

Fréquence à laquelle les données sont générées, capturées et partagées! Elles doivent être analysées en temps réel (*Data Stream Mining*). Risque pour l'Homme de perdre une grande partie de la maîtrise du système >>> Trading haute fréquence : des "robots" sont capables de lancer des ordres d'achat ou de vente de l'ordre de la nanoseconde.

Variété

Les *data centers* sont devant un véritable défi : la variété des données. Il ne s'agit pas de données relationnelles traditionnelles, elles sont brutes, semi-structurées, non structurées >>> analyses d'autant plus complexes qu'elles portent sur les liens entre des données de natures différentes.

Big Data

Le Big Data et la recherche scientifique

Parmi les plus gros pourvoyeurs de données. Les approches massives basées sur une logique d'exploration des données et de recherche d'induction sont **complémentaires** des approches classiques basées sur l'hypothèse initiale formulée (Ouf! On respire!)

Les approches massives en biologie

- Génomique et toutes les « omiques » >>> Décoder le génome humain a originellement pris 10 ans, cela peut désormais être fait en moins d'une semaine : les séquenceurs d'ADN ont progressé d'un facteur 10 000 les dix dernières années, soit 100 fois la loi de Moore (caractérisée par un facteur 100 sur 10 ans)
- Environnement
- Epidémiologie
- Big Brain
- Drug Design …

Autres

- Physique : Large Hadron Collider; Astronomie; Climat et Météo
- Gestion du temps réel
- Banques
- Réseaux sociaux (Facebook : , Twitter,…) et autres monstres (Google, Yahoo, …)
- Assurances …

DATA SCIENCE

Définition

Discipline visant à **l'extraction de connaissances** à partir d'ensembles de données et s'appuyant sur des outils **statistiques**, mathématiques et informatiques (et notamment de visualisation). Elle peut être considérée comme une **science des données numériques.**

Elle permet faire parler et de valoriser les données

Métiers (au moins pour l'instant)

« Le Geek, l'Analyste et le Communicateur ». Une fable ?

Valorisation des données

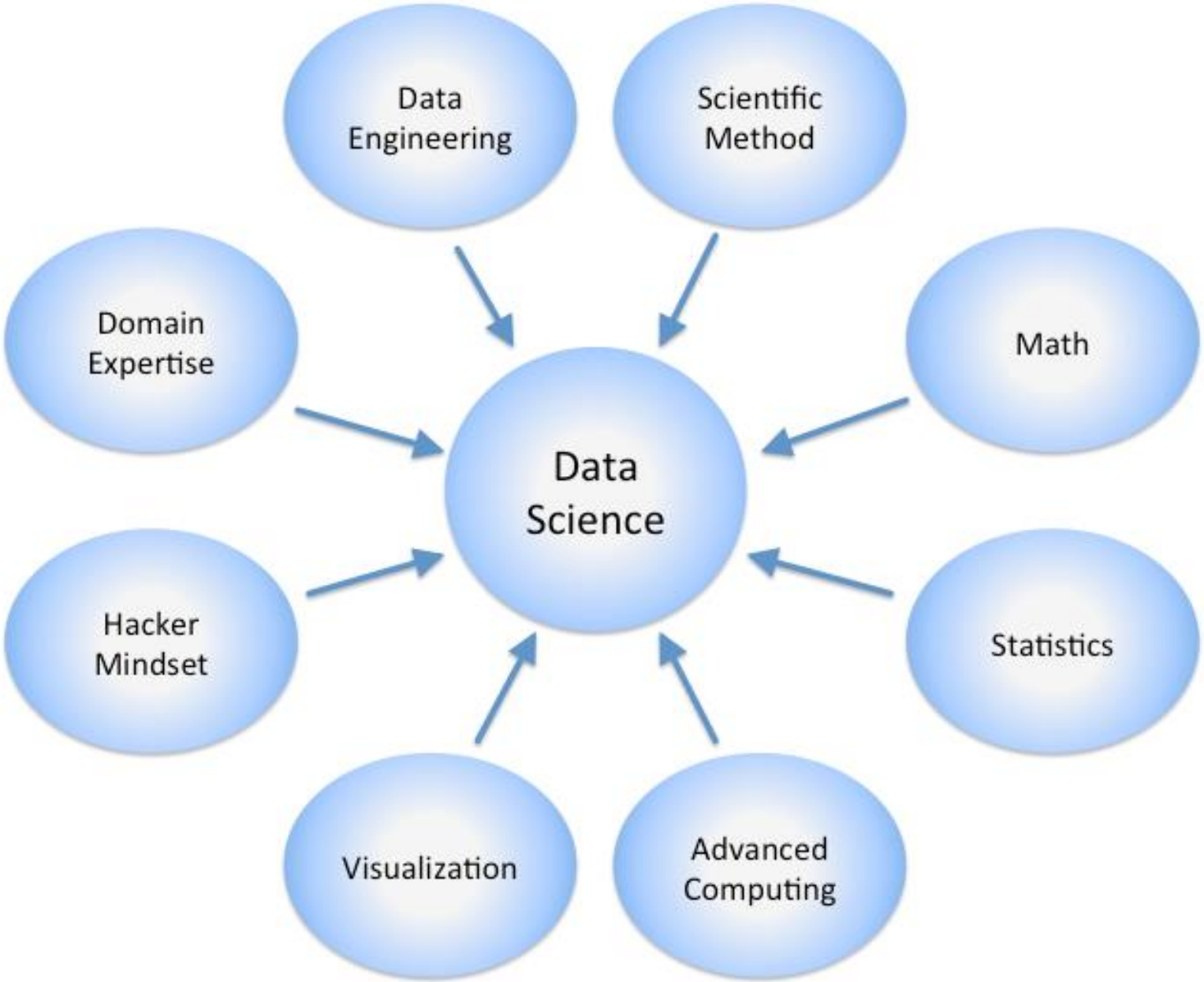
L'élément central

Il va donc falloir apprendre à faire parler les données (préparation, gestions des valeurs manquantes, statistique descriptive, modélisation, statistique inférentielle, …)

Les données : de l'or blanc !

A vous de jouer! La balle est dans votre camp. Saurez vous relever ce défi?

DATA SCIENCE



Présentation du site web de l'UE

www.biostatistique.u-psud.fr