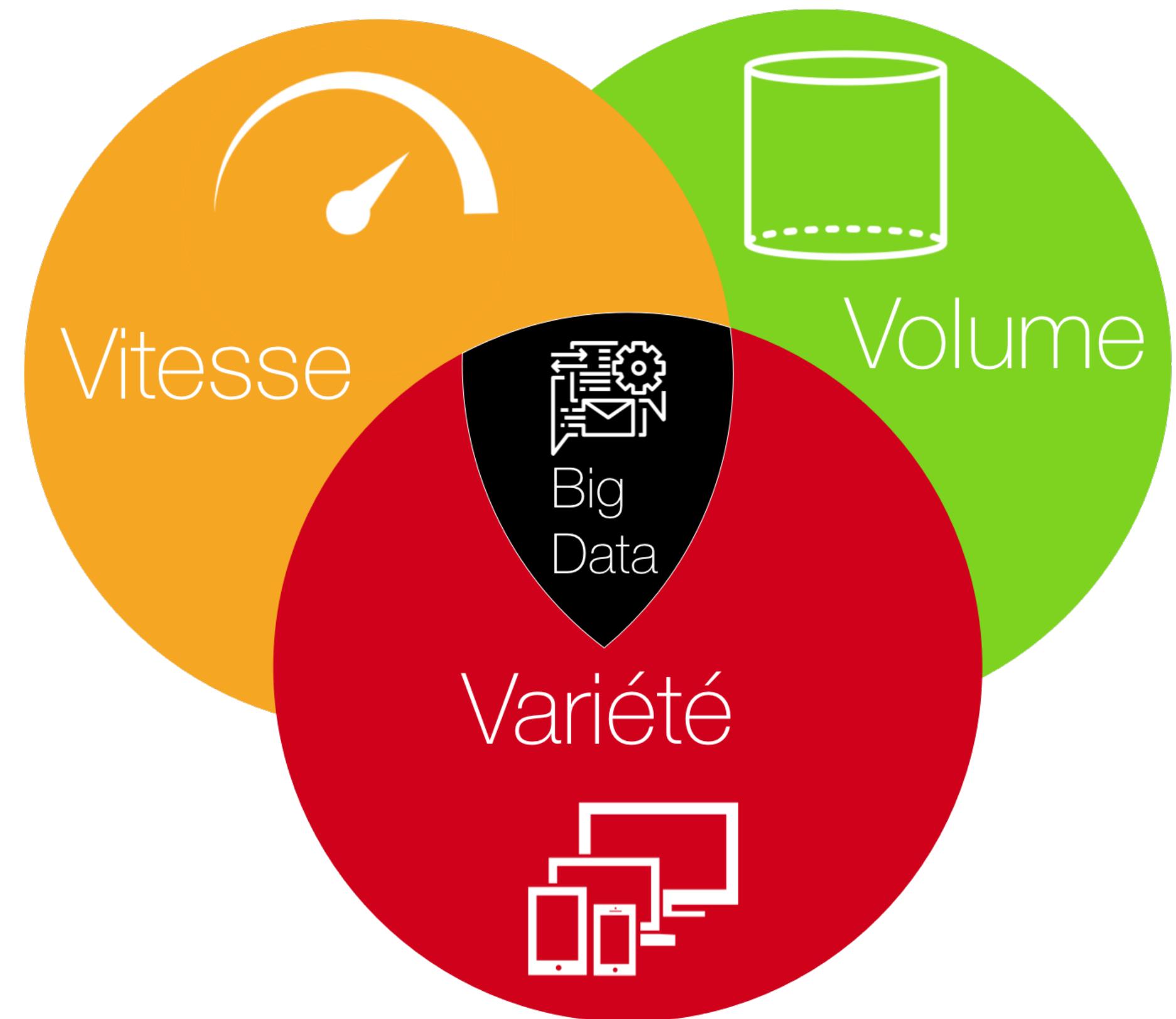


—

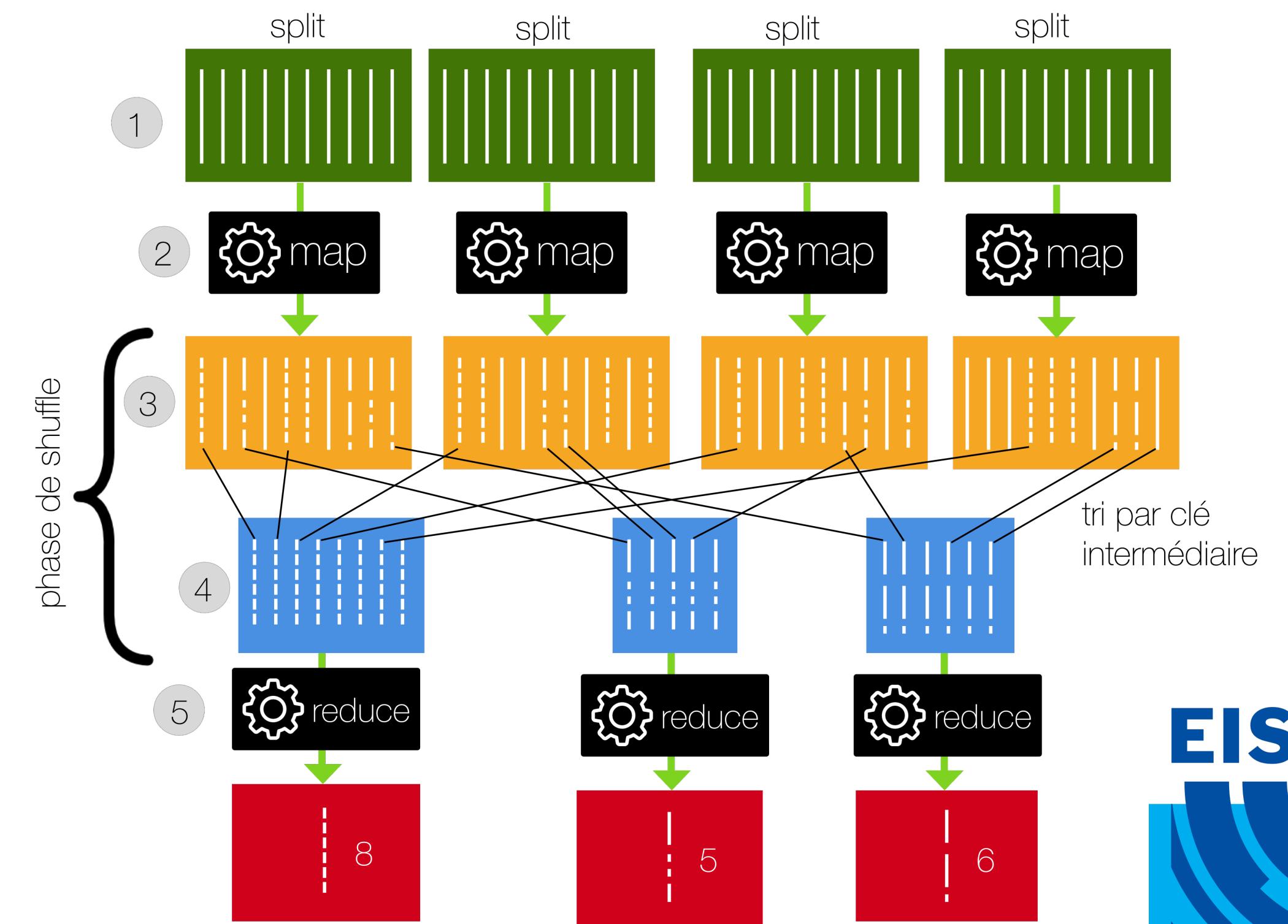
Sujet TIPE CPI1- généralité sur le Big Data



Technologie du Big Data étudiée en CPI1



Map Reduce



Algorithmes PySpark pour traitement de données



TIPE 2016-2017
CPI2

DALAA Mehdi, LUNDQUIST Maxime, MBUNDU Mandry, RIGAUX David



Objectif du TIPE en CPI2

- Etude d'algorithmes de traitement de données
- Appliquer ces algorithmes à des jeux de données et en tirer des résultats exploitables



Sommaire

Éléments techniques

1. Python
2. Spark

Applications et Démos

1. Wordcount
2. K-Means
3. TF.IDF
4. Notes d'informatique

Éléments techniques

1. Python
2. Spark

Python



- Langage Orienté Objet
- Très utilisé par les Data Scientists
- À disposition de nombreuses librairies
- Version Python 3

Framework de traitement Big Data

- Effectuer des analyses sophistiquées
- Conçu pour la rapidité
- Facilité d'utilisation

Spark est basé sur le clustering

- Réparti les tâches sur plusieurs machines
- Traitements beaucoup plus rapide



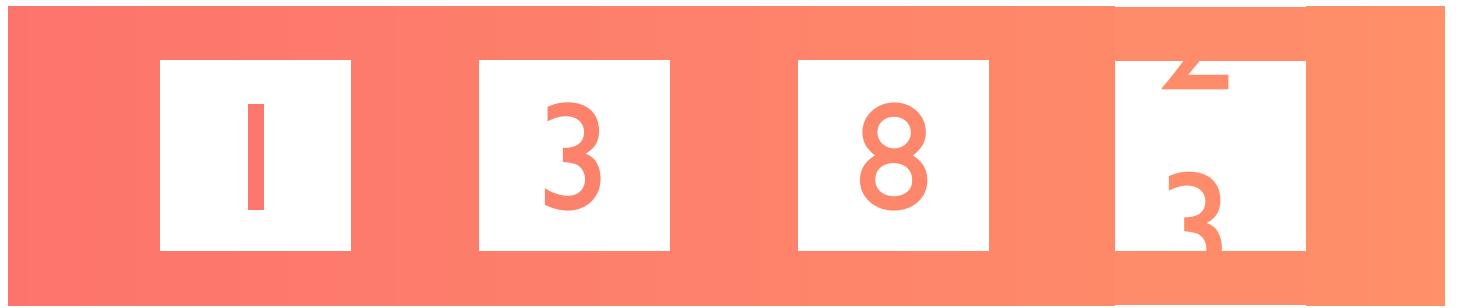
L'API PySpark

- PythonSpark
- Librairie de Machine Learning

Applications et Démonstration

1. Wordcount
2. K-Means
3. TF.IDF
4. Notes d'informatique

Wordcount

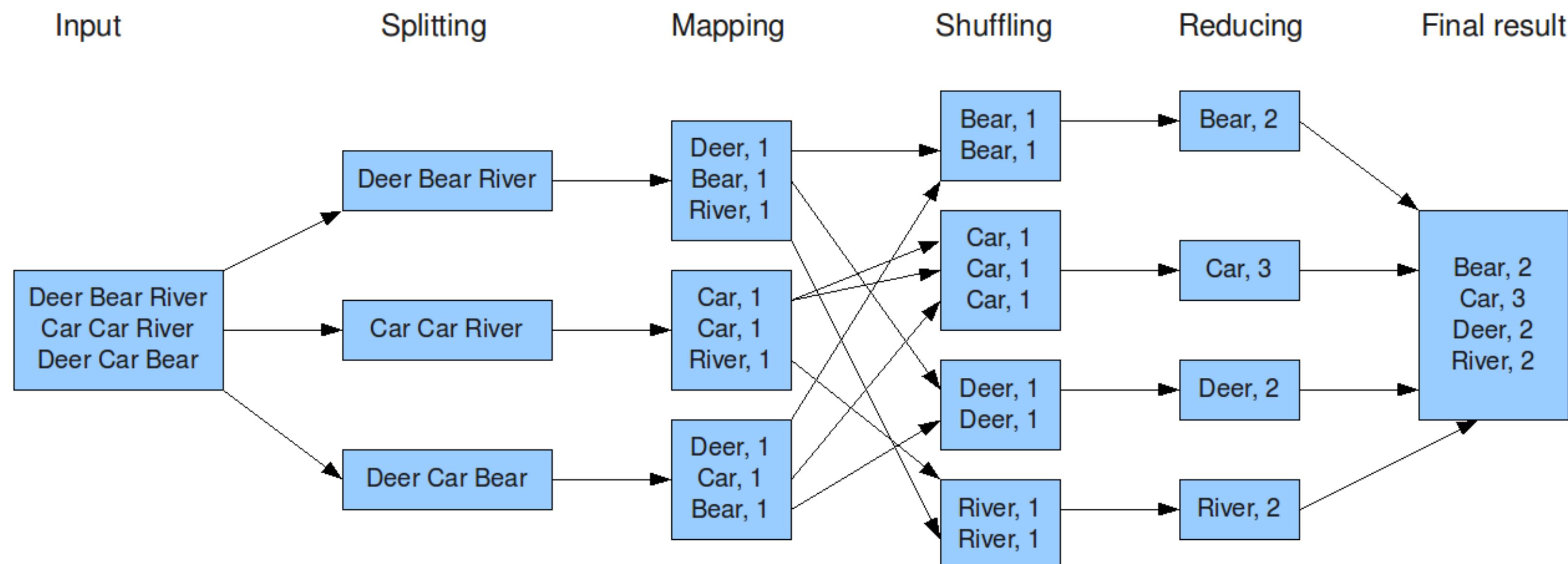


1 3 8 5

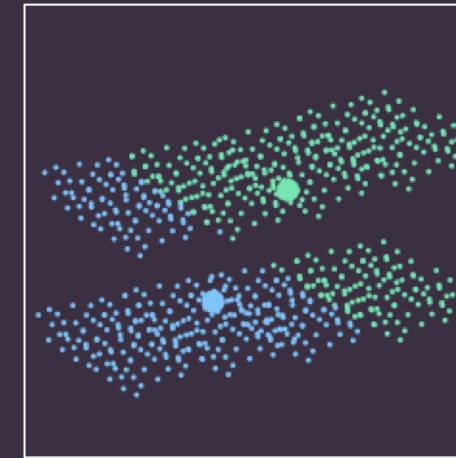
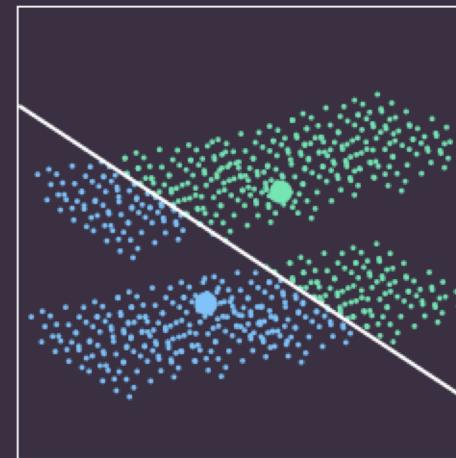
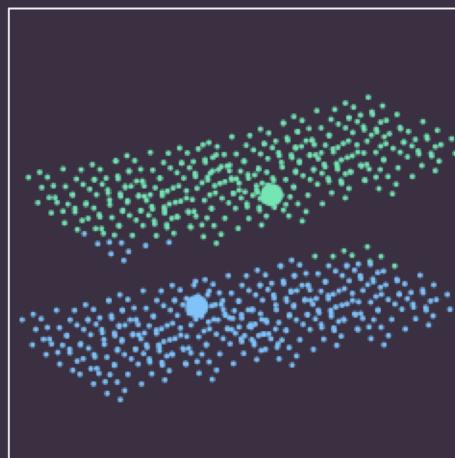
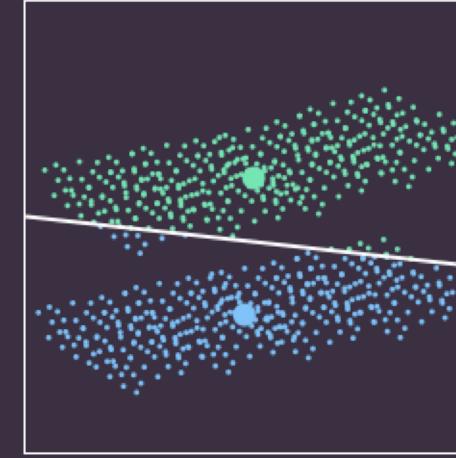
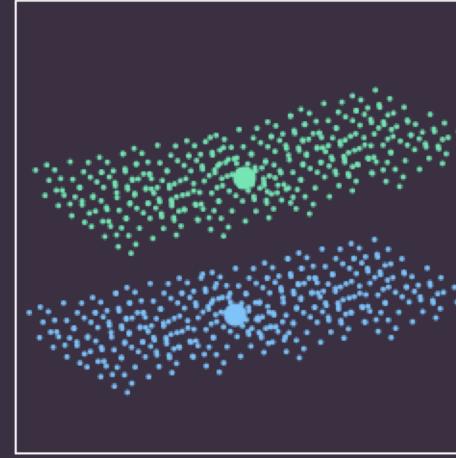
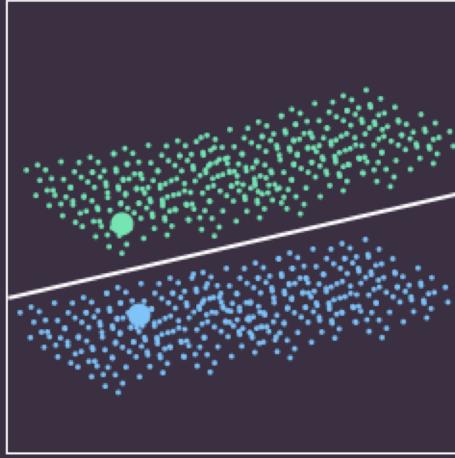
- Programme utilisé pour compter le nombre de mots présents dans un texte
- Utile lorsque l'on traite un texte très long
- Permet aussi de connaître la rapidité de lecture

Wordcount

The overall MapReduce word count process



K-Means



But : Regrouper n-points
en k-groupes

Contraintes: Minimiser la
distance entre chaque point
dans chaque cluster

Formule de K-means

$$\arg_S \min \sum_{i=1}^k \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$$

DISTANCE, PAR EXEMPLE ICI, EUCLIDIENNE

Text-mining :

- Spécialisation du data mining dont le traitement fonctionne en deux étapes dépendantes l'une de l'autre :
 1. Analyse des mots des phrases
 2. Interprétation de l'analyse

- TF-IDF (Term Frequency -Inverse Document Frequency) : mesure statistique qui permet d'évaluer l'importance d'un contenu par rapport à un document ou à un corpus.

- Le TF-IDF se décompose en deux étapes :
 1. TF donnée par le nombre de fois que le mot apparaît divisé par le nombre de mot
 2. L>IDF qui est une mesure de l'importance du terme dans l'ensemble du corpus.

TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

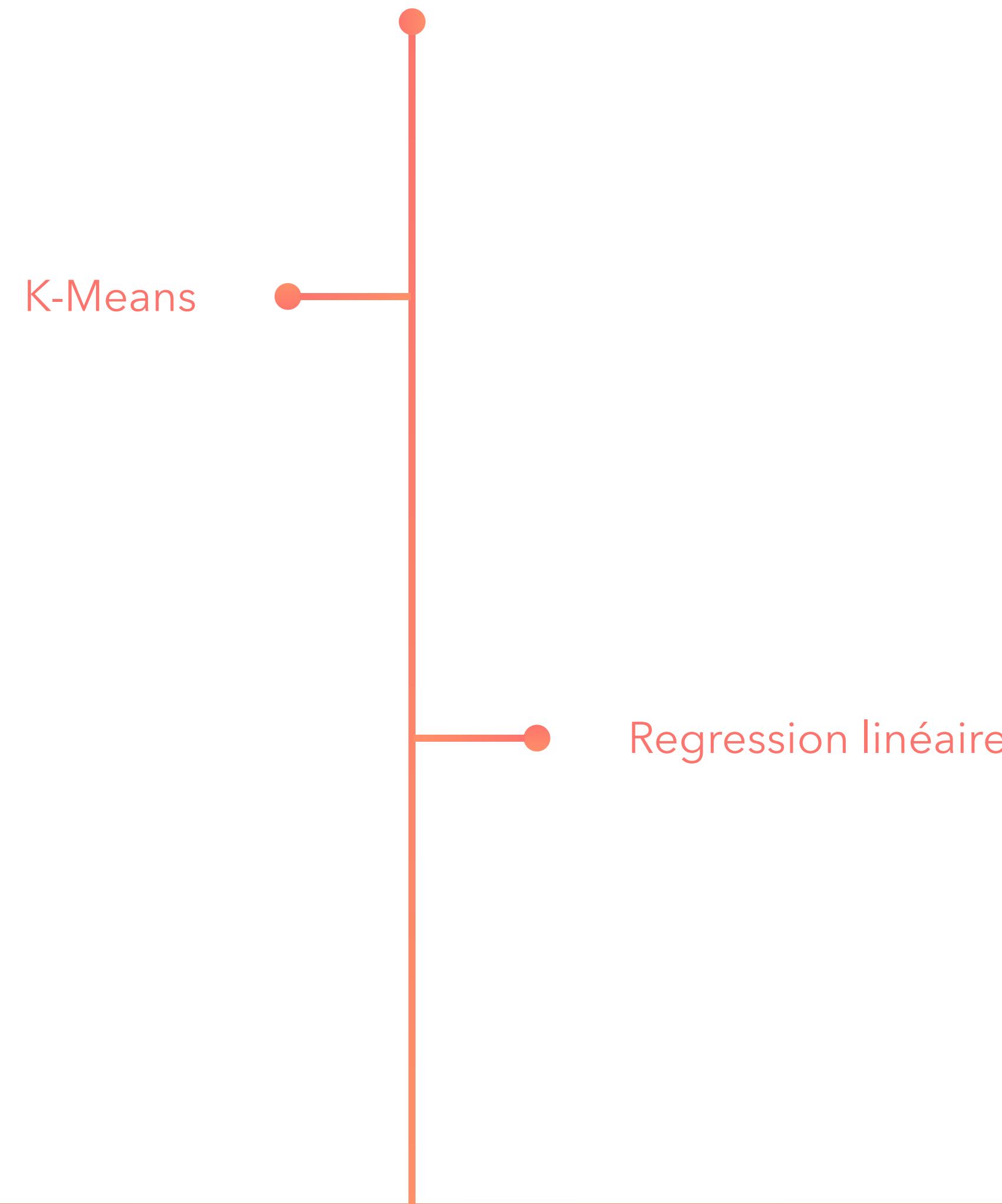
Notes d'informatiques

Objectif :

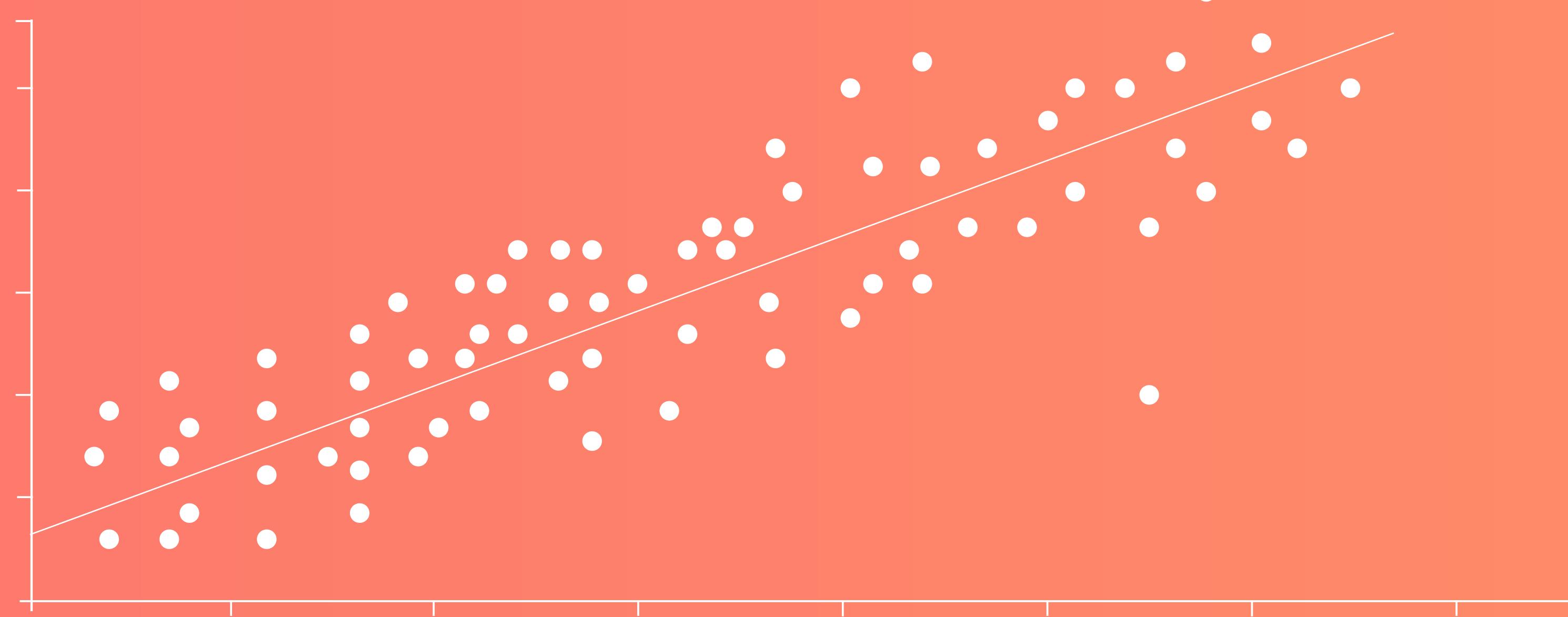
Appliquer des algorithmes de traitement de données pour en tirer un résultat exploitable (prédiction, classification, ...)

id	Info théorique	Programmation
1	12.5	10
2	13	7.5
3	6	5
4	10	12
5	7.5	1
6	13	11
7	11	14
8	10.5	6
9	10	10
10	5	7
11	14.5	10.5
12	9	5
13	10	6
14	14	9
15	17	20
16	7	7
17	9	9.5
18	14	9.5
19	14.5	12
20	8	8.5
21	10	9
22	12.5	13
23	10.5	11

Algorithmes utilisés



Regression Linéaire



- Observer simultanément des individus d'une population sur deux caractères
- Mesurer un lien éventuel entre ces caractères
- Pour deux variables il existe trois croisements possibles :
 1. Quantitatif x Qualitatif
 2. Qualitatif x Quantitatif
 3. Quantitatif x Quantitatif

-
- Enjeu capital dans notre société
 - Le flux d'informations émis, partagé et collecté concerne tout les domaines
 - Difficultés avec Spark et Python
 - Projet enrichissant