



CentraleSupélec

Big Data : Informatique pour les données et calculs massifs


## 11 – Exploration et préparation des données

**Stéphane Vialle**




universitÉ PARIS-SACLAY

ÉCOLE DOCTORALE  
Sciences et technologies  
de l'information  
et de la communication (STIC)




RISEGrid  
Research Institute for Sustainable Electric Grids



Grand Est  
ALSACE CHAMPAGNE-ARDENNE LORRAINE

Stephane.Vialle@centralesupelec.fr  
http://www.metz.supelec.fr/~vialle



CentraleSupélec

## Quelques Pb & Sol avant l'exploitation des données

**Pb 1 : Diversité des données**  
*Sources différentes - Types différents - Formats différents*

**Pb 2 : Qualité variable des données**  
*Exhaustivité insuffisante - Granularité inadaptée - Exactitude ??*

↓

**Sol : Démarche d'exploration préalable des données**  
*Révélation de la structure - Enquête sur la collecte - Stats. descriptives*

**Sol : Démarche de préparation des données**  
*Nettoyage - Transformation - Enrichissement*

*En Big Data : un POC standard est-il démonstratif ?*

2

## Exploration et préparation des données

1. **Pbs engendrés par la diversité des données**
2. Pbs engendrés par la qualité variable des données
3. Explorer les données par précaution
4. Préparer les données pour de meilleures analyses
5. Un POC standard est-il démonstratif ?

3

## Diversité des sources

### Données internes : 1<sup>ères</sup> sources de données

- **Les entreprises débutent souvent leur démarche BigData en analysant leurs données internes** (historiques, archives...):
  - Données de production
  - Données sur les clients (CRM : Customer Relation Management)
  - Transactions d'achats et ventes
  - Processus/Plan de gestion (ERP : Entreprise Resource Planning)

#### Objectifs :

- Améliorer la stratégie de l'entreprise
  - Economiser sur les ressources consommées
  - Anticiper l'avenir avec des modèles prédictifs
- **Ensuite** l'entreprise se tournera vers des sources de **données externes** (achat de BdD, accès à des OpenData, Web crawling)

4

CentraleSupélec

## Diversité des sources

### Données externes : BdD commerciales

- **Ventes de jeux de données par des instituts de sondage, de mesure d'audience...** par des entreprises spécialisées dans la production et la vente de données
- **Vente des données liées à l'activité propre de l'entreprise** (mais dont la diffusion est sans risque pour elle)

La vente de ces données peut devenir une source de revenus significative...

Elle est de plus en plus encouragée...

- Vente « par BdD » ou « à l'unité »

5


CentraleSupélec

## Diversité des sources

### Données externes : OpenData

Des collectivités, des administrations et parfois des entreprises publient des **données libres d'accès** (*OpenData*)

Ex : **data.gouv.fr** publie des *OpenData* sous l'autorité du gouvernement français (résultats du recensement, résultats économiques du pays, données sur l'éducation, le logement, la santé, l'agriculture...)



**data.gouv.fr**

- Agriculture et Alimentation
- Culture, Communications
- Comptes, Économie et Emploi
- Éducation, Recherche, Formation
- International, Europe
- Environnement, Énergie, Logement
- Santé et Social
- Société, Droit, Institutions
- Territoires, Transports, Tourisme

→ données du TD MongoDB

CentraleSupélec

## Diversité des sources

**Données externes : collectées sur le web**

Collecte automatique de contenus de sites web par des robots de *crawling* (des logiciels), puis extraction des données pertinentes

Une source d'information intéressante, mais qui pose 2 problèmes :

- Pb 1 : collecte de **données très déstructurées**  
(pas forcément conçues pour faciliter leur collecte)
- Pb 2 : quels **droits d'exploitation ???**  
(constitue une *zone grise* de la législation)

7

CentraleSupélec

## Diversité des types

**Données comportementales d'utilisateurs / de clients**

- Les **logs des sites web** : contiennent les historiques de navigation des visiteurs
  - Pour reconstituer les parcours et comportements des visiteurs
  - Pour faire des mesures fines d'attractivité des sites web (web analytics)
- Les **données des réseaux sociaux - si elles sont accessibles (!!)**
  - Pour enrichir les profils des utilisateurs/des clients

Elles sont mises à jour par les utilisateurs eux-mêmes (!!)

- Données souvent à jour
- Mise à jour gratuitement


8

CentraleSupélec

## Diversité des types

### Données comportementales de systèmes / d'installations

- Les **données issues de capteurs et objets connectés** : stockées également dans des logs, ou échantillonnées et historisées.
  - Pour faire de la détection et prédiction de pannes
  - Pour réaliser de la maintenance prédictive
 « Comportement d'installations industrielles »
- Données de capteurs et d'objets connectés pour capter des comportements d'utilisateurs... **délicat !!**



9

CentraleSupélec

## Diversité des types

### Données géographiques / spatiales

- Les **données de localisation** permettent **d'enrichir d'autres données**

Ex :

- Carte mentionnant des point d'intérêts (restaurants, hôtels...)
- Données socio-économiques recoupées avec des informations géographiques (revenu par habitant et par pays, par ville...)
- Conséquence sur le trafic routier d'événements de proximité

10

CentraleSupélec

## Diversité des formats

### Données stockées en fichiers

- Fichiers .txt, .csv... : **des lignes de données ASCII**  
Pb : les séparateurs de données et de lignes ne sont pas standardisés
- Fichiers JSON, XML... : **des données structurées et complexes** (hiérarchiques)
  - Souvent sous forme de listes de clé-valeur
  - Contiennent les informations pour les interpréter (auto-descriptif)
  - Peuvent être interrogés à travers des API/bibliothèques (ex : des classes de *parsing* XML en Java)
- Fichiers « *shapefile* » : **des données de nature géométrique** (points, lignes, polygones...)
  - Proviennent de systèmes d'informations géographiques

11

CentraleSupélec

## Diversité des formats

### Données stockées en fichiers

- Fichiers **multimédias au format binaire** : images, vidéos, audio
  - Difficile à exploiter sans une application spécialisée
  - Souvent **étiquetés à partir de leur métadonnées externes**
- Nouveaux **formats de stockage Apache** :
  - Apache parquet : facilite une lecture par colonne pour les calculs de *data-analytics*
  - Apache Avro : mécanisme de *sérialization* incluant le schéma des données (auto-descriptif)

**OpenSource**  
The Apache Software Foundation

**Fichiers binaires plats :**

- Très compacts, économiques
- Sans ~~métadonnées~~ incluses (besoin de connaître leur schéma pour les relire)

Traditionnels en HPC, mais **PAS en BigData** 12

CentraleSupélec

## Diversité des formats

### Données stockées en BdD

- **BdD SQL :**  
Stockage sous forme de **tables relationnelles**, liées par des clés, toutes cohérentes, et respectant des contraintes d'intégrité (ex : il y a bien un « int » et pas un « double » pour un nombre d'étudiants, cet entier est positif ou nul, ...)  
Interrogation par requêtes SQL
- **BdD NoSQL :**  
Stockage sous **formes très variées**
  - tables avec un nombre variable de colonnes par ligne,
  - format JSON structuré-hiérarchique,
  - format XML...
sans cohérence ni intégrité garanties  
Interrogation par API native ou Java ou Python...  
...parfois interrogation dans un langage proche de SQL (ex : Hive)

CentraleSupélec

## Diversité des données

### Bilan de la diversité des données

- **Plusieurs aspects** dans la diversité :  
sources, types, formats ... et d'autres aspects
- **Faut-il exploiter des données plus variées ?**
  - Permettrait une analyse plus riche / pertinente
  - Va demander **plus d'efforts d'ingénierie**
  - Va **augmenter les coûts** de développement

14

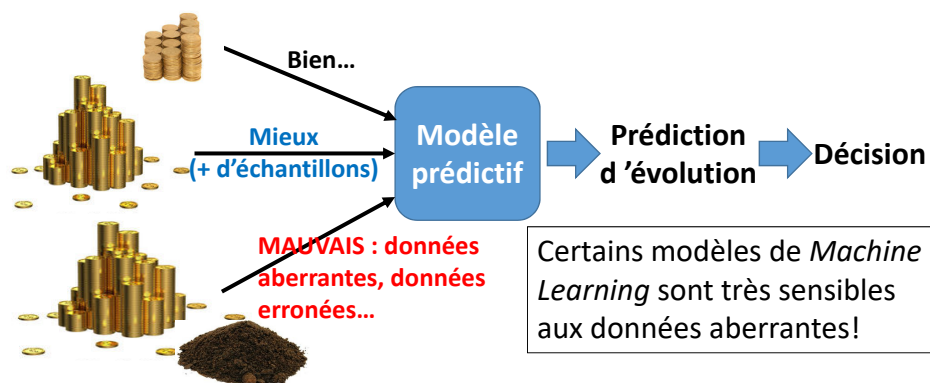
## Exploration et préparation des données

1. Pbs engendrés par la diversité des données
2. **Pbs engendrés par la qualité variable des données**
3. Explorer les données par précaution
4. Préparer les données pour de meilleures analyses
5. Un POC standard est-il démonstratif ?

15

## Qualité variable des données

### Données alimentant le modèle



Une approche BigData ne permet pas automatiquement d'accepter des données de mauvaise qualité sous prétexte qu'on en prend énormément





CentraleSupélec

## Qualité variable des données

**Qualité : Exhaustivité**

- Rmq : un jeu de données complet n'est pas toujours indispensable.
  - Recherche d'une trace d'intrusion : **log complets importants**
  - Analyse statistique d'accès à un site : **log complets inutiles**
- **Complétude de la liste des enregistrements**
  - Parfois impossible ou inutile
  - Risque de ramener des données erronées, mal collectées
- **Complétude de chaque enregistrement**
  - que faire d'un enregistrement incomplet ??
    - Le rejeter
    - Le conserver
    - Le compléter avec une valeur moyenne, médiane, précédente ...

**Voir la *préparation des données***

17

CentraleSupélec

## Qualité variable des données

**Qualité : Granularité des données**

La *granularité* est aussi le *degré de finesse* des données


→ finesse spatiale, temporelle, sociale...

Ex. de **besoins en granularité d'un ensemble de mesures** :

- assez précises ?
- assez fréquentes dans le temps ?
- couvrant l'ensemble de la zone étudiée ?
- maillant assez finement la zone étudiée ?
- couvrant toutes les catégories sociales de personnes ?

Ex. d'**études avec des besoins différents en granularité** :

- Des données élémentaires sont nécessaires pour caractériser des individus
- Des données agrégées suffisent pour caractériser des villes

 **Qualité variable des données**


**Qualité : Exactitude des données**

L'**exactitude** des données enregistrées représente aussi leur **fiabilité**

- Les **données saisies manuellement** par l'utilisateur, ou dictées par téléphone, seront **entachées d'erreurs**
  - Besoin de vérification dans des annuaires par exemple.
- Les **sorties de capteurs peuvent être fausses** si un capteur est abimé ou a vieilli
  - Une redondance de la mesure peut permettre de détecter les fausses mesures... mais 2x plus de data !

**Vérifier et assurer l'exactitude des mesures est couteux**

19

 **Qualité variable des données**


**Qualité : Fraicheur des données**

La **fraicheur** des données enregistrées indique si elle sont **assez récentes pour l'étude**

Ex. sensibles sur la fraicheur :

- données **boursières** pour décision d'achat ou de vente
- données **géographiques** pour localiser un produit en mouvement

20



## Qualité variable des données


**Bilan de la qualité des données :**

- Différents critères de qualité des données
- Pas de mesure absolue de la qualité des données
- Sensibilité et besoin très variables selon les données et les traitements
- Il y a un coût à l'augmentation de la qualité des données

**Faut-il augmenter la qualité des données ?**

→ Décider et agir en fonction des coûts de collecte et d'exploitation

21



## Exploration et préparation des données

1. Pbs engendrés par la diversité des données
2. Pbs engendrés par la qualité variable des données
- 3. Explorer les données par précaution**
4. Préparer les données pour de meilleures analyses
5. Un POC standard est-il démonstratif ?

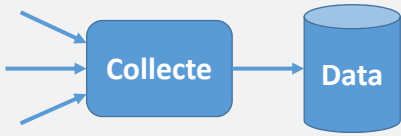
22

CentraleSupélec

## Objectifs de l'exploration

- La phase d'analyse **démarre** avec l'exploration des données

➤ Inspecter les données et leur collecte :



Bonne qualité(s) ?  
 Allure attendue ?  
 Existence de biais ?  
 Existence de valeurs aberrantes ?

➤ afin de révéler/vérifier la structure des données

➤ et pouvoir répondre à la question :  
 → Le modèle prévu semble-t-il toujours adapté aux données ?

- Une exploration par **visualisation** est souvent une bonne méthode (l'œil humain perçoit très bien des structures au sein d'ensembles de données)

CentraleSupélec

## Démarche d'exploration

**1 - « Enquêter » sur les données et sur leur collecte**

- Saisies manuelle ou automatique ?
- Dans le même but que l'analyse prévue ?
- Collectées en une fois ou en plusieurs étapes ? (mêmes conditions pour toutes les étapes ?)
- Référentiels et unités de mesures utilisés ?
- ...

} Recherche de metadonnées

On n'a jamais toutes les metadonnées que l'on souhaite !  
 → **besoin d'enquêter ! ... peut être long et délicat**

**Objectifs :**

- Identifier la **provenance** des données
- Appréhender la **structure des données** et leurs **formats**
- Détecter les **incohérences** entre les données
- Détecter les biais des mesures

24

CentraleSupélec

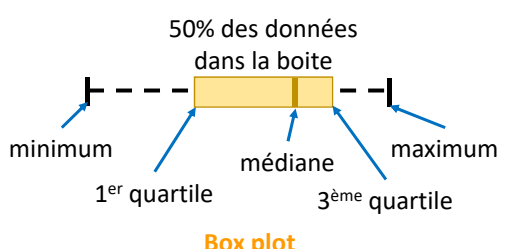
## Démarche d'exploration

### 2 - Calculer et visualiser des **statistiques descriptives**

Objectif des statistiques descriptives :  
*Cerner le centre d'un ensemble de données et sa dispersion autour de ce centre*

→ moyenne, **médiane**, variance, écart –type, **quantiles**  
 → **représentation graphique**

**Ex. de visualisation**  
 pour un ensemble de valeurs **numériques**



50% des données dans la boîte

minimum 1<sup>er</sup> quartile médiane 3<sup>ème</sup> quartile maximum

**Box plot**

25

CentraleSupélec

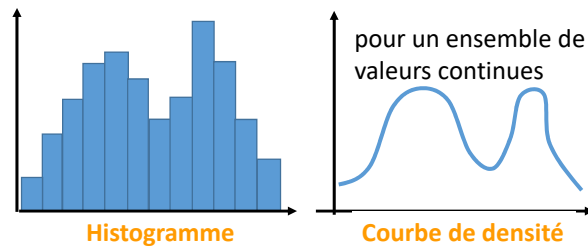
## Démarche d'exploration

### 2 - Calculer et visualiser des **statistiques descriptives**

Objectif des statistiques descriptives :  
*Cerner le centre d'un ensemble de données et sa dispersion autour de ce centre*

→ moyenne, **médiane**, variance, écart –type, **quantiles**  
 → **représentation graphique**

**Ex. de visualisation**  
 pour un ensemble de valeurs **numériques**



**Histogramme**

**Courbe de densité**

pour un ensemble de valeurs continues

26

CentraleSupélec

## Démarche d'exploration

**3 - Multiplier les visualisations en cas de données complexes**


**Rappel des objectifs :**

- Avoir une bonne idée des données qui vont alimenter les modèles d'analyse et d'apprentissage
- Détecter les valeurs aberrantes, les biais...

→ En cas de données complexes plusieurs visualisations peuvent être nécessaires

**Négliger/sauter cette étape peut mener à faire des analyses longues et coûteuses totalement fausses !**

Ex : certains algorithmes sont très sensibles aux valeurs aberrantes...



27

CentraleSupélec

## Exploration et préparation des données

1. Pbs engendrés par la diversité des données
2. Pbs engendrés par la qualité variable des données
3. Explorer les données par précaution
- 4. Préparer les données pour de meilleures analyses**
5. Un POC standard est-il démonstratif ?

28

CentraleSupélec

## Objectifs de la préparation


**Constituer un jeu de données apte à être analysé :**

- de qualité homogène
- avec des structures et des formats bien définis
- exploitable par des algorithmes d'analyse automatique

**Corriger les défauts** (révélés par l'exploration) :

- Traiter les valeurs manquantes et aberrantes
- Redéfinir certaines échelles pour homogénéiser des variabilités
- Utiliser des données externes pour enrichir le jeu initial

**3 étapes :**



```

graph LR
  A[nettoyage] --> B[transformation]
  B --> C[enrichissement]
  
```

29

CentraleSupélec

## Les 3 étapes de la préparation

**nettoyage** → **transformation** → **enrichissement**

### 1 - Nettoyage des données

**Identifier toutes les données indésirables :**

- Erronées / aberrantes
- Trop inexactes (manque de précision)  
*Ex : valeurs trop éloignées de la théorie ou des autres, enregistrement d'utilisateurs sans leurs emails...*
- Inutiles pour l'analyse voulue ... faut-il les supprimer ?

**Actions possibles :**

- **Supprimer une valeur** non conforme et garder le reste de l'enregistrement
- **Supprimer tout l'enregistrement** (toute la ligne)
- **Ajouter un flag** indiquant si l'enregistrement est conforme ou non (exploitable ou non)
- **Remplacer la valeur** non conforme par : la médiane, la moyenne, la valeur précédente (cas des données ordonnées)

30

CentraleSupélec

## Les 3 étapes de la préparation

nettoyage → transformation → enrichissement

### 2 - Transformation des données

Objectifs :

**Manipuler, modifier, créer** de nouvelles informations à partir de celles contenues dans les données initiales

Exemples :

- Identifier et Lister les mots les plus fréquents dans des articles à analyser ensuite
- Identifier et Taguer les logs ou événements dont les dates correspondent à des vacances scolaires ou à des jours fériés

*Rmq : ce genre de transformations sur de grands ensembles de données se prête bien à des traitements Map-Reduce*

31

CentraleSupélec

## Les 3 étapes de la préparation

nettoyage → transformation → enrichissement

### 2 - Transformation des données

**Opérations de Séparation et Extraction**

- Analyse syntaxique d'une chaîne et séparation en une liste d'éléments
- Extraction d'un élément particulier dans une chaîne : nom, email...

**Opérations d'Agrégation de valeurs**

Remplacement d'un ensemble de données par :

- Le nombre des occurrences d'un terme
- Les termes les plus fréquents
- Une valeur moyenne, médiane, maximale, minimale...

32



CentraleSupélec

## Les 3 étapes de la préparation

nettoyage → transformation → enrichissement

### 2 - Transformation des données

#### Opérations de **Modification de valeurs**

- Transformation des dates pour s'adapter aux fuseaux horaires, calcul du nombre de jours entre deux dates...
- Remplissage des valeurs manquantes
- Réorganisation des données par catégories pertinentes

*Ce dernier point empiète sur l'analyse de données !!*

33

CentraleSupélec

## Les 3 étapes de la préparation

nettoyage → transformation → enrichissement

### 3 - Enrichissement des données

Objectif :

**Croiser les données existantes avec de nouvelles données, parfois extérieures à l'entreprise**

**Ex. d'enrichissement géographique :**

- Ajouter des info de géocodage (latitude, longitude) à une adresse
- Ajouter une adresse à un géocodage (géocodage inverse)
- Ajouter un géocodage à une adresse IP (démarche approximative)

**Ex. d'enrichissement temporel :**

- Déterminer si une date d'événement à une adresse donnée, se situe dans les congés scolaires du pays concerné

**Ex. d'enrichissement socio-économique :**

- Associer des informations juridiques et financières au nom d'une entreprise

34

CentraleSupélec

## Les 3 étapes de la préparation

nettoyage → transformation → enrichissement

### 3 - Enrichissement des données

Objectif :

**Croiser les données existantes avec de nouvelles données, parfois extérieures à l'entreprise**

→ **Correspond souvent à une opération de jointure**

- Pour une BdD relationnelle : jointure classique
- Pour une BdD NoSQL :
  - patron Map-Reduce de jointure
  - jointure + restructuration des données

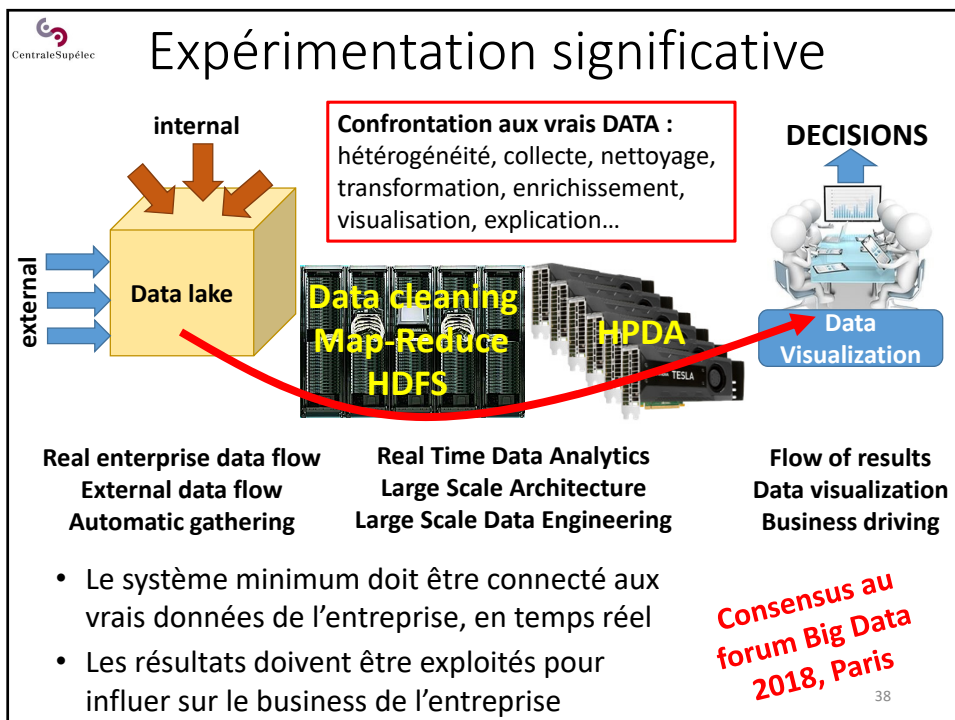
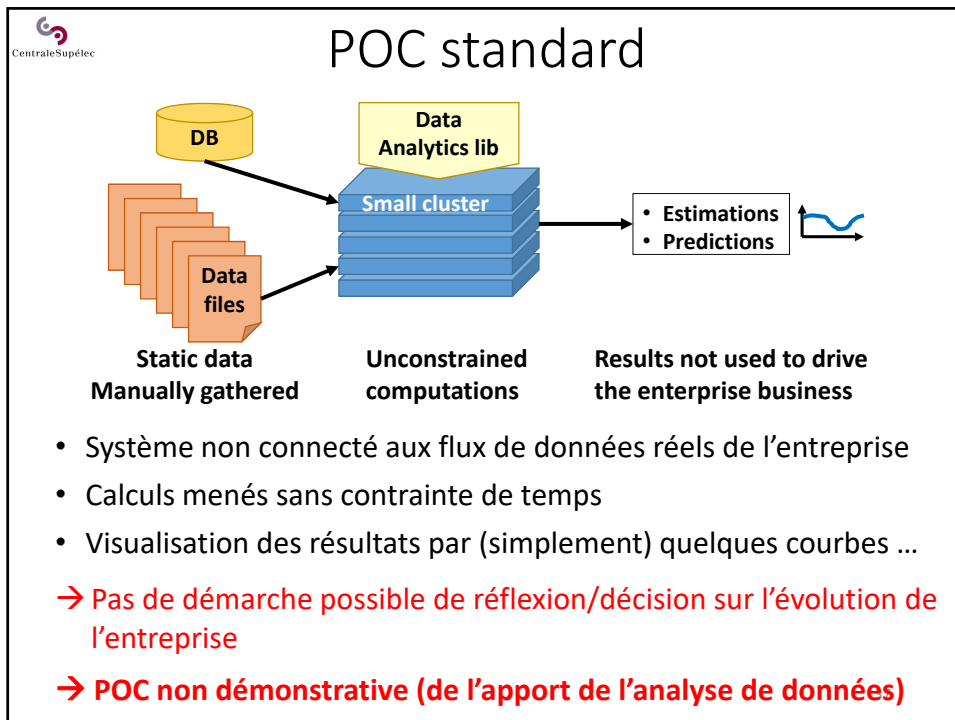
35

CentraleSupélec

## Exploration et préparation des données

1. Pbs engendrés par la diversité des données
2. Pbs engendrés par la qualité variable des données
3. Explorer les données par précaution
4. Préparer les données pour de meilleures analyses
5. **Un POC standard est-il démonstratif ?**

36



# Exploration et préparation des données

