

# Documentation sur le sujet

Sonny Klotz - Jean-Didier Pailleux - Malek Zemni

UVSQ

28/02/2017

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

## Plan

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

DCbrain développe des outils qui permettent de visualiser ce qui se passe sur les **réseaux physiques** afin de :

- trouver et prédire les problèmes sur ces réseaux
- optimiser ces réseaux

DCbrain développe des outils qui permettent de visualiser ce qui se passe sur les **réseaux physiques** afin de :

- trouver et prédire les problèmes sur ces réseaux
- optimiser ces réseaux

Les données sont collectées à partir des réseaux physiques puis analysées grâce aux technologies du **Big Data**.

DCbrain développe des outils qui permettent de visualiser ce qui se passe sur les **réseaux physiques** afin de :

- trouver et prédire les problèmes sur ces réseaux
- optimiser ces réseaux

Les données sont collectées à partir des réseaux physiques puis analysées grâce aux technologies du **Big Data**.

## Réseaux physiques

Réseaux industriels, de fluide, de distribution. Exemple : distribution pétrolière, gazière, électrique...

## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

Exemple de sources :



## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

Exemple de sources :

- Capteurs pour collecter les informations climatiques, de trafic, consommation (Smart cities, Internet des Objets)

## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

Exemple de sources :

- Capteurs pour collecter les informations climatiques, de trafic, consommation (Smart cities, Internet des Objets)
- Messages sur les réseaux sociaux

## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

Exemple de sources :

- Capteurs pour collecter les informations climatiques, de trafic, consommation (Smart cities, Internet des Objets)
- Messages sur les réseaux sociaux
- Enregistrements transactionnels d'achat en ligne

## Big Data

Ensembles de très gros volumes de données, à la fois structurées, semi-structurées ou non structurées, qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

Exemple de sources :

- Capteurs pour collecter les informations climatiques, de trafic, consommation (Smart cities, Internet des Objets)
- Messages sur les réseaux sociaux
- Enregistrements transactionnels d'achat en ligne
- Signaux GPS de téléphones mobile

Comment utiliser et donner du sens à ces masses de données enregistrées sur ces réseaux ?

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

## Définition

Ensemble de techniques de statistique descriptive.

## Définition

Ensemble de techniques de statistique descriptive.

- **Objectifs** : une description succincte, regrouper les données.



## Définition

Ensemble de techniques de statistique descriptive.

- **Objectifs** : une description succincte, regrouper les données.
- **Les données** : tableaux de données quantitatives et qualitatives.

## Définition

Ensemble de techniques de statistique descriptive.

- **Objectifs** : une description succincte, regrouper les données.
- **Les données** : tableaux de données quantitatives et qualitatives.
- **Avantages** : traitement en masse, représentations graphiques.

Travail à réaliser :

- détection du type de colonnes
- statistique descriptive sur une colonne

Travail à réaliser :

- détection du type de colonnes
- statistique descriptive sur une colonne

données quantitatives -> effectifs cumulés et fréquence d'apparition

Travail à réaliser :

- détection du type de colonnes
- statistique descriptive sur une colonne

données quantitatives -> effectifs cumulés et fréquence d'apparition

données quantitatives

- tendance centrale : moyennes et mode
- dispersion : quantiles, variance et écart-type
- forme : symétrie et aplatissement

Travail à réaliser :

- détection du type de colonnes
- statistique descriptive sur une colonne

données quantitatives -> effectifs cumulés et fréquence d'apparition

données quantitatives

- tendance centrale : moyennes et mode
- dispersion : quantiles, variance et écart-type
- forme : symétrie et aplatissement

**Open source** : R - Python - ROOT (C++ et Python) - Java

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.



## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.

Les traitements de cette quantité importante de données est massivement "parallélisé" avec MapReduce/Hadoop.

## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.

Les traitements de cette quantité importante de données est massivement "parallélisé" avec MapReduce/Hadoop.

Le Big Data est caractérisé par les 3V :

## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.

Les traitements de cette quantité importante de données est massivement "parallélisé" avec MapReduce/Hadoop.

Le Big Data est caractérisé par les 3V :

- le Volume de données considérable à traiter.

## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.

Les traitements de cette quantité importante de données est massivement "parallélisé" avec MapReduce/Hadoop.

Le Big Data est caractérisé par les 3V :

- le Volume de données considérable à traiter.
- la Variété de ces données qui peuvent être brutes, non structurées ou semi-structurées

## Big Data

Le Big Data fait référence à la masse de données collectée. On considère du Big Data quand le traitement devient trop long pour une seule machine.

Les traitements de cette quantité importante de données est massivement "parallélisé" avec MapReduce/Hadoop.

Le Big Data est caractérisé par les 3V :

- le Volume de données considérable à traiter.
- la Variété de ces données qui peuvent être brutes, non structurées ou semi-structurées
- la Vitesse qui désigne le fait que ces données sont produites, récoltées et analysées en temps réel.

Le Machine Learning est :

Le Machine Learning est :

- Une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes automatisables, offrant la possibilité à une machine d'évoluer grâce à un processus d'apprentissage à partir des données et à effectuer des tâches de façon performante.

## Le Machine Learning est :

- Une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes automatisables, offrant la possibilité à une machine d'évoluer grâce à un processus d'apprentissage à partir des données et à effectuer des tâches de façon performante.
- Un traitement statistique de masses de données réunissant à la fois mathématiques appliquées et informatique.



## Le Machine Learning est :

- Une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes automatisables, offrant la possibilité à une machine d'évoluer grâce à un processus d'apprentissage à partir des données et à effectuer des tâches de façon performante.
- Un traitement statistique de masses de données réunissant à la fois mathématiques appliquées et informatique.
- Utilisé lorsque le Big Data rend inopérant les méthodes statistiques traditionnelles.

## Le Machine Learning est :

- Une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes automatisables, offrant la possibilité à une machine d'évoluer grâce à un processus d'apprentissage à partir des données et à effectuer des tâches de façon performante.
- Un traitement statistique de masses de données réunissant à la fois mathématiques appliquées et informatique.
- Utilisé lorsque le Big Data rend inopérant les méthodes statistiques traditionnelles.

Le Machine Learning est composé de plusieurs types d'algorithmes d'apprentissage (supervisé, non supervisé, semi-supervisé, par renforcement).

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

DCbrain emploie une approche basée sur une représentation des réseaux physiques en **graphes de flux**.

DCbrain emploie une approche basée sur une représentation des réseaux physiques en **graphes de flux**.

Les graphes de flux permettent de représenter les données liées au flux du réseau et de les analyser.

DCbrain emploie une approche basée sur une représentation des réseaux physiques en **graphes de flux**.

Les graphes de flux permettent de représenter les données liées au flux du réseau et de les analyser.

Avantages :

DCbrain emploie une approche basée sur une représentation des réseaux physiques en **graphes de flux**.

Les graphes de flux permettent de représenter les données liées au flux du réseau et de les analyser.

Avantages :

- repérer beaucoup plus facilement des anomalies dans le réseau

DCbrain emploie une approche basée sur une représentation des réseaux physiques en **graphes de flux**.

Les graphes de flux permettent de représenter les données liées au flux du réseau et de les analyser.

Avantages :

- repérer beaucoup plus facilement des anomalies dans le réseau
- simuler des évolutions du réseau



Les graphes de flux peuvent être utilisés pour tout réseau physique de fluide, par exemple les réseaux électriques :

Les graphes de flux peuvent être utilisés pour tout réseau physique de fluide, par exemple les réseaux électriques :

### Graphe de flux : d'un réseau électrique :

- Nœuds : des connections
- Arcs : canaux pour acheminer l'électricité (câbles)

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les données représentées par des graphes sont dites **semi-structurées**.

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les données représentées par des graphes sont dites **semi-structurées**.

Il est nécessaire d'employer une méthode d'analyse qui convient à ce genre de données : le **graph mining**.

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les données représentées par des graphes sont dites **semi-structurées**.

Il est nécessaire d'employer une méthode d'analyse qui convient à ce genre de données : le **graph mining**.

Principe du graph mining :

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les données représentées par des graphes sont dites **semi-structurées**.

Il est nécessaire d'employer une méthode d'analyse qui convient à ce genre de données : le **graph mining**.

Principe du graph mining :

- extraire des informations utiles à partir d'une masse de données semi-structurées

Les méthodes d'analyse de données classiques sont limitées au **données structurées**.

Les données représentées par des graphes sont dites **semi-structurées**.

Il est nécessaire d'employer une méthode d'analyse qui convient à ce genre de données : le **graph mining**.

Principe du graph mining :

- extraire des informations utiles à partir d'une masse de données semi-structurées
- càd miner des sous-graphes fréquents qui décrivent l'information recherchée



Utilisation des informations extraites :

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux
- Chimie et biologie

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux
- Chimie et biologie
- Réseaux physiques

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux
- Chimie et biologie
- Réseaux physiques

Quelques algorithmes :



Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux
- Chimie et biologie
- Réseaux physiques

Quelques algorithmes :

- Apriori-based Approach

Utilisation des informations extraites :

- Trouver des relations entre différents éléments du graphe
- Prédire le comportement des éléments d'un graphe

Domaines d'application :

- Réseaux sociaux
- Chimie et biologie
- Réseaux physiques

Quelques algorithmes :

- Apriori-based Approach
- Pattern-Growth Approach

## Plan

- 1 Introduction
- 2 Analyse descriptive de données
  - Généralités
  - Statistique unidimensionnelle
- 3 Big Data et Machine Learning
  - Big Data
  - Machine Learning
- 4 Graphe de flux et Graph Mining
  - Graphe de flux
  - Graph Mining
- 5 Conclusion

Et notre application...