Introduction

Algorithmique du traitement des données
USID07 – Master MEDAS

Auteur : Raphaël Fournier-S'niehotta (fournier@cnam.fr)

Département d'informatique
Conservatoire National des Arts & Métiers, Paris, France

Plan

1 Introduction

2 Données

3 Tableur

Plan

1 Introduction

Infos pratiques

- 9 séances de cours (9h-10h45)
- 8 séances de TP (11h-12h30)
- Évaluation sur table, sans documents, 11h-12h30 le 12 décembre
- Mail : fournier@cnam.fr

Objectifs

MÉgaDonnées et Analyse Sociale

- Faire un panorama de la manipulation des données
- Apprendre les bases de l'algorithmique et du traitement des données (développé dans les autres cours du Master)

Aujourd'hui

- Distribution des certificats de scolarité
- Prise de contact
- Émargement / appel
- Cours introductif : Algorithmique du traitement des données

Plan

2 Données
- Types de données

Données

- une donnée (data en anglais) est **la représentation d'une information**
- chaque jour, on en génère tellement qu'on est amené de ne pas y prêter trop attention :
  - la moindre connexion internet,
  - la poussière sur le plancher
  - la température de l'eau,
  - la fuite d'air à la fenêtre
  - l'heure à laquelle on mange/dort
  - le nombre de pas/mètres/calories que l'on effectue
- ça devient une donnée dès qu'on décrit/mesure/analyse

Popularité actuelle des données

- une des raisons est qu'elles restent plus longtemps
- la poussière sur mon plancher disparaît avec l'aspirateur.
- la connexion à un site internet restent plusieurs mois dans plusieurs fichiers de plusieurs machines différentes.
- comme elles restent plus longtemps, on a plus de temps pour les observer et leur donner du sens.

Popularité actuelle des données

- une des raisons est qu'elles restent plus longtemps
- la poussière sur mon plancher disparaît avec l'aspirateur.
- la connexion à un site internet restent plusieurs mois dans plusieurs fichiers de plusieurs machines différentes.
- comme elles restent plus longtemps, on a plus de temps pour les observer et leur donner du sens.

Pourquoi ?

Comment ?

Pourquoi ?

- la plupart des gens ne manipulent pas des données simplement pour leur plaisir :
- elles sont utilisées pour rendre visible des phénomènes.

- on commence souvent par une question !
- combien de fois le soleil brille dans ma ville natale ?
- comment mon gouvernement dépense-t-il son argent ? Et d'où proviennent les fonds ?
- quelles sont les connaissances de la population française entre 18 et 25 ans sur la contraception féminine ?

Comprendre

- une question est un bon point de départ pour explorer des données,
- cela permet de préciser votre recherche et aide à détecter des tendances intéressantes.
- quelle que soit votre problématique, vous devez toujours rester attentif aux observations inattendues, aux résultats inhabituels, ou à tout ce qui pourra vous surprendre
- souvent, les phénomènes les plus intéressants ne sont pas ceux que vous recherchez.

Donner du sens aux données

- nous sommes tous très amnésiques (comparativement aux machines) mais une des façons qui nous permet de retenir est la répétition.
- une observation, une donnée, commence à prendre du sens dès qu'elle se répète.

Enquête sur l'entendement humain, David Hume (1748)

De causes qui paraissent semblables, nous attendons des effets semblables. Telle est la somme de toutes nos conclusions expérimentales.

Donner du sens aux données

- les marins utilisaient les étoiles pour se repérer.
- ils ont su associer la position d'une étoile dans le ciel (une donnée) de la même étoile à la même position une année plus tard
- la donnée est répétée.
- c'est le début de la connaissance : chaque année, la même étoile est à la même position dans le ciel.
- on peut l'utiliser pour se repérer.

Traces

- la somme des données qui se rapporte à la même personne est quasiment infini.
- seulement, aujourd'hui, elle persiste.
- qu'en faire ? C'est tellement énorme que ce serait comme découvrir toute la voie lactée le même jour.
- il faudrait une vie pour l'étudier...
- sauf que...on a maintenant des ordinateurs qui font beaucoup de calculs très rapidement.

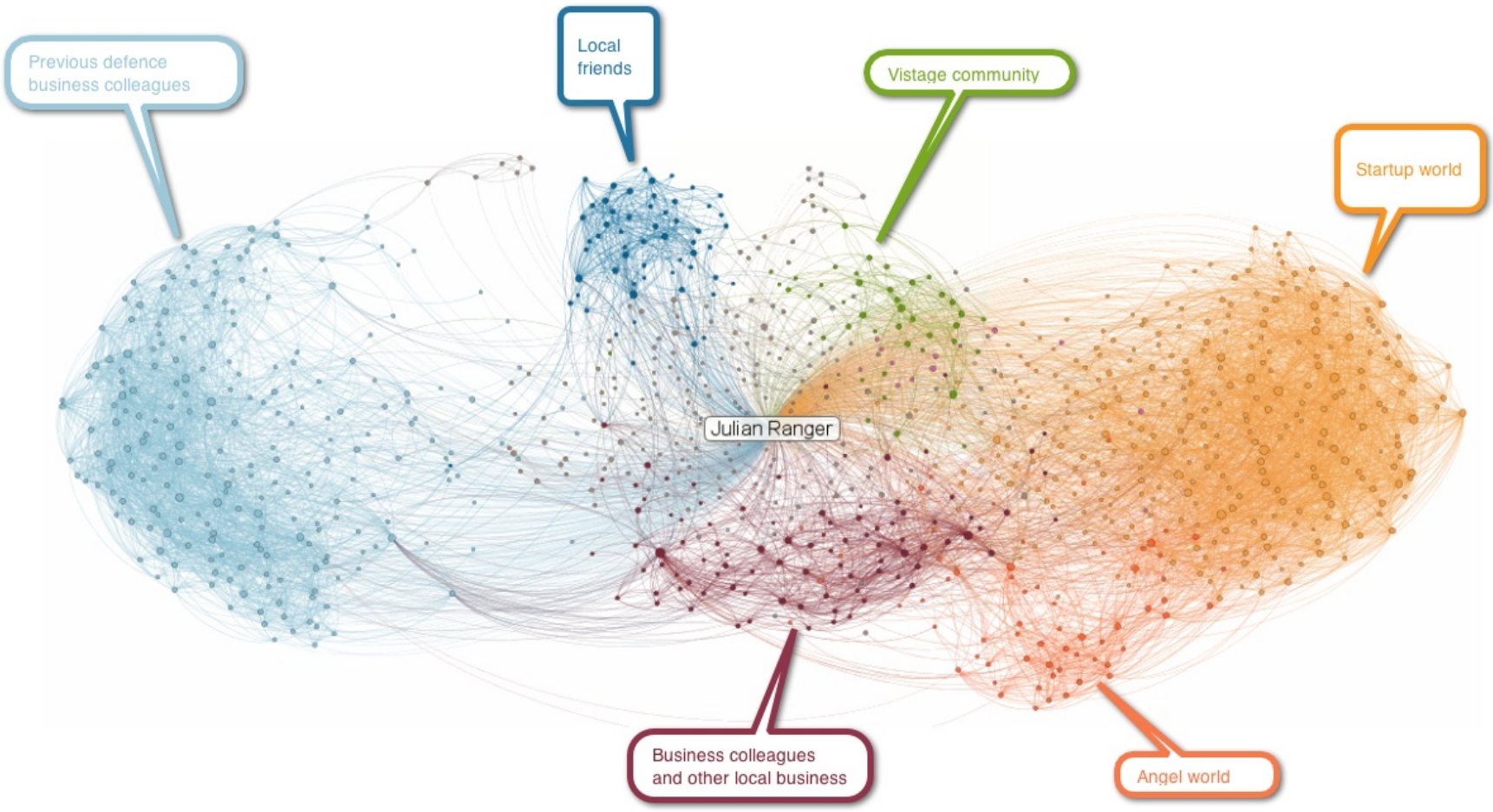
Traces

- la somme des données qui se rapporte à la même personne est quasiment infini.
- seulement, aujourd'hui, elle persiste.
- qu'en faire ? C'est tellement énorme que ce serait comme découvrir toute la voie lactée le même jour.
- il faudrait une vie pour l'étudier...
- sauf que...on a maintenant des ordinateurs qui font beaucoup de calculs très rapidement.

Statistiques

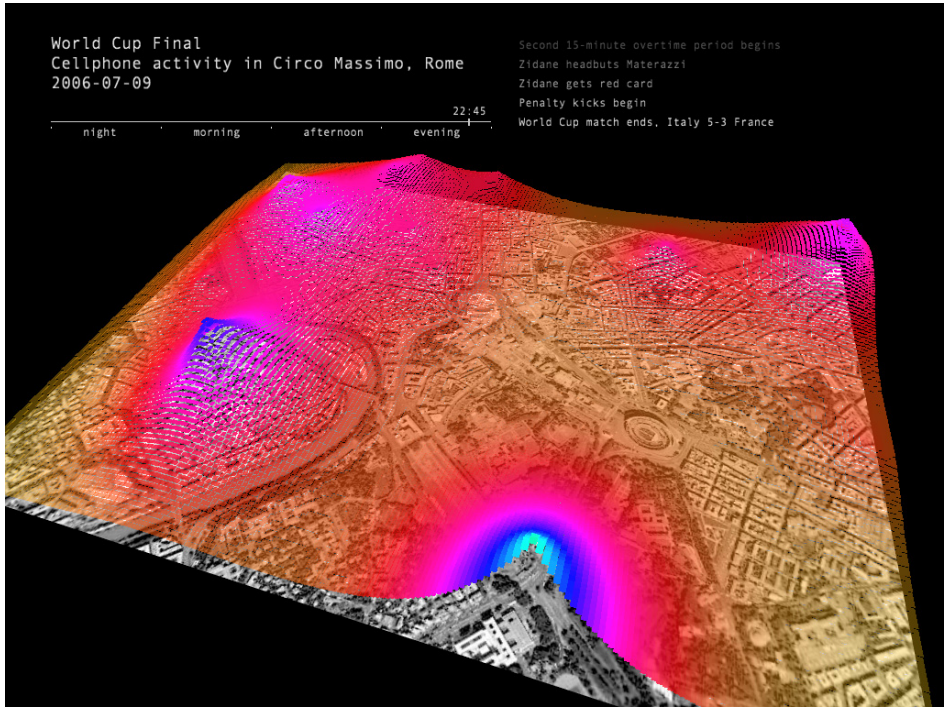
Informatique

Exemple : Réseau personnel



Exemple : déplacement via communications téléphoniques

- suivi de communications :
  - date, heure, durée, type, correspondant
  - type d'appelant, mobilité, ...
  - http://senseable.mit.edu



Autres exemples d'études

- informatique : pages Web, routeurs, P2P, etc.
- biologie : protéines, neurones cérébraux, etc.
- sciences sociales : amitiés, collaboration, etc.
- économie : échanges financiers
- linguistique : synonymie, co-occurrence
- transports : réseau aérien, électrique

Big Data

- volume
- variété
- vélocité (Vitesse)
- youtube : 400 heures de vidéos crées chaque minute
- environ 1000 jours de nouvelles vidéos par heure
- chaque plateforme a des chiffres équivalents

Démarche

Héritée de la démarche en statistique descriptive

- étude d'un problème
- avec des objectifs bien définis
- **collecter des données**
- décrire, explorer
- **modéliser**
- **analyser** : estimer, tester, valider
- synthétiser : communiquer, proposer, revenir au réel

Autre approche :

- centrée sur les données
- on trouve la question / les phénomènes pertinents **pendant** l'exploration, non avant

Exemple

- quelles sont les connaissances de la population française entre 18 et 25 ans sur la contraception féminine ?
- objectif : améliorer la prévention
- sondage, avec des questions
- calculer un score par individu
- décrire les résultats selon les catégories possibles dans le problème (hommes, femmes, 14/15/16)
- “les hommes de 22 ans et plus ne sont pas suffisamment au courant des risques encourus avec la pilule XYZ”
- “les femmes de moins de 19 ans ne connaissent pas les risques avec le préservatif”

Données qualitatives et quantitatives

- les données qualitatives se réfèrent à la qualité :
  - la description d'une couleur, de textures et l'aspect d'un objet, la description d'une expérience sont toutes des données qualitatives.
- les données quantitatives sont des données qui se réfèrent aux chiffres.
  - le nombre de balles de golf, la taille, le prix, le résultat d'un test, etc.

Données qualitatives et quantitatives

- les données qualitatives se réfèrent à la qualité :
  - la description d'une couleur, de textures et l'aspect d'un objet, la description d'une expérience sont toutes des données qualitatives.
- les données quantitatives sont des données qui se réfèrent aux chiffres.
  - le nombre de balles de golf, la taille, le prix, le résultat d'un test, etc.
- les données catégorielles permettent de classer les objets que vous traitez par catégories. Dans notre exemple, l'aspect « usagé » serait une catégorie au sein de la typologie suivante : « nouveau », « usagé », « cassé », etc.
- les données discrètes sont des données dénombrables. Ex: le nombre de balles de golf. Il ne peut y avoir qu'un nombre entier de balles de golf (il ne peut pas y avoir 0,3 balles de golf).
- les données continues sont des données numériques non entières. Ex: le diamètre des balles de golf (ex: 10,53mm, 10,56mm, 10.536mm), ou la taille précise du pied (en opposition à la pointure, discrète). Toutes les valeurs sont admises.

Données

- données numériques
  - âge
  - revenu mensuel
  - échelle de Likert (entre 1 et 5, ou variantes)
- données textuelles
  - noms, prénoms
  - oui/non
  - profession

Structurons un peu

couleur	blanche
Catégories	Sport, Golf
État	usagé
Diamètre	43 mm
Prix	36 centimes

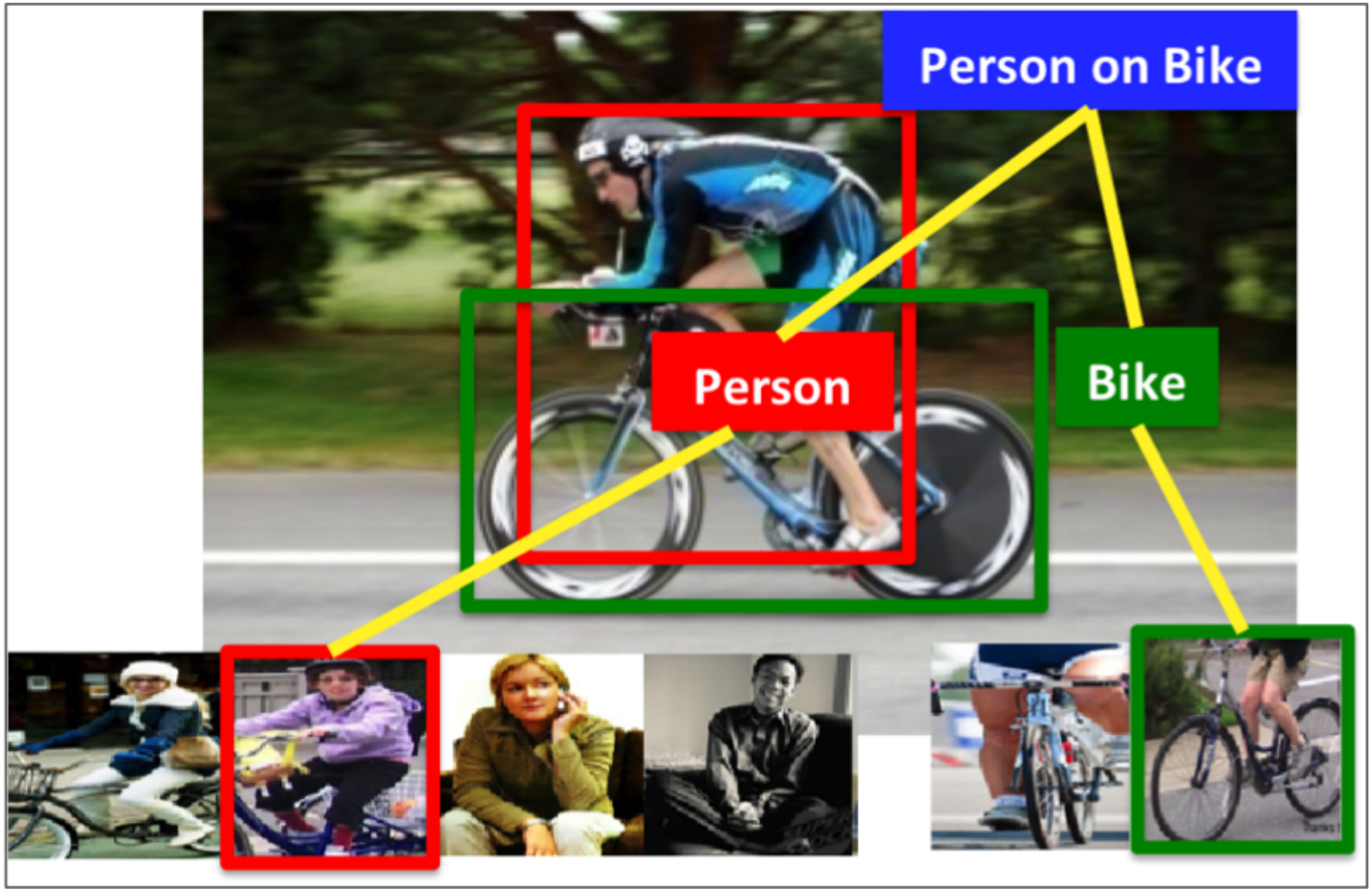
Interprêter

- ces données n'ont pas de sens exploitées individuellement. Pour faire émerger l'information, nous devons les interpréter.
- un diamètre de 43 mm ne signifie rien.
- il devient intéressant quand il est comparé à une autre donnée, un autre diamètre.
- dans certains sports, il y a une réglementation pour les équipements. La taille minimale d'une balle de golf en compétition est de 42,67 mm. Nous pouvons donc utiliser cette balle en compétition. C'est une information.
- en revanche, ce n'est toujours pas de la connaissance. La connaissance est créée lorsque l'information est apprise, appliquée et comprise

Données non structurées, données structurées

- “Il y a 5 balles de golf usagées avec un diamètre de 43 mm à 0,5 € chacune”
- phrase facilement compréhensible pour un humain, mais compliquée à comprendre par un ordinateur
- donnée non structurée
  - beaucoup de tâches facilement réalisables par un individu sont encore difficilement exécutables par les machines
  - exemple : comprendre le langage naturel, décrire une image, etc.

Image décrite par ordinateur



Structurer

- si l'on veut que l'ordinateur analyse la donnée, il faut qu'il soit capable de la lire et de la traiter
- elle doit être structurée dans un format lisible par la machine
- un des formats les plus couramment utilisé pour l'échange de données est le format CSV, pour "comma-separated values" (données/valeurs séparées par des virgules)
- exemple :

  quantité, couleur, condition, objet, catégorie, diamètre(mm), prix unitaire (€)

  5,blanc,usagé,balle,golf,43,0.36

Format CSV

- c'est un format simplifié pour l'ordinateur et lisible par des tableurs.
- vous noterez que les mots sont entourés de guillemets, ce qui les distingues en tant que texte (chaîne de caractères dans le langage informatique), alors que les nombres n'ont pas de guillemets.

Formats de stockage

- fichiers “plats” non structurés : texte brut
- fichiers structurés : XML, JSON, CSV
- formats complexes ouverts ou propriétaires (PDF, Word)
- bases de données
  - mySQL, SQLite, PostgreSQL
  - oracle, SQL Server
  - mongoDB
- il existe beaucoup d’autres formats structurés et lisibles par une machine : XML, JSON, etc. cf cours de mon collègue, G. Kembellec

Encodage

- pour transcrire les langues naturelles, il faut un codage des caractères
- code ASCII (7 bits : 128 caractères)
- pas de caractères accentués, de cédilles, tilde espagnol
- puis norme ISO-8859-1/ISO-8859-15 en France
- aujourd'hui : **Unicode** (UTF-8, 1 à 4 octets, 1 million de caractères)
- attention : toutes les plateformes informatiques n'ont pas les mêmes pratiques par défaut
- il faudra parfois effectuer des transmformations

Sources de données

Il existe principalement trois moyens de se procurer des données :

- collecter vous-mêmes les données et les stocker dans le format de votre choix
  - appareil de mesure, entretiens, etc.
- trouver des données déjà disponibles et les télécharger (gratuitement ou non)
  - open Data, API, autres (IGN, Infogreffe)
- réclamer des données auprès des sources officielles, par exemple en application des lois sur le droit d'accès et de réutilisation des données publiques (loi CADA de 1978 en France). Parfois une donnée est publiée en ligne mais n'est pas directement téléchargeable.
- il faut identifier les sources de données qui peuvent vous aider à répondre aux problèmes que vous traitez

Source de données

- l'État et les collectivités territoriales
  - depuis quelques années, les acteurs publics ont commencé à ouvrir une partie de leur données.
  - ils ont parfois créé des portails dédiés pour mettre à disposition les données publiques ouvertes. Par exemple, le gouvernement français publie des données sur le portail data.gouv.fr.
  - datacatalogs.org
- les organisations internationales : La Banque mondiale, l'Organisation mondiale de la santé (OMS), l'OCDE, ...publient régulièrement des rapports d'études et des jeux de données.
- les sources scientifiques. Les institutions de recherche publient des données à destination de leurs communautés scientifiques et du grand public.
  - exemple : CNRS, NASA, etc.
  - pour la plupart des disciplines scientifiques, il existe des répertoires spécialisés de données, parfois librement réutilisables.

Autres projets Open Data

- de plus en plus de projets ont pour objectif de faciliter l'accès aux données déjà publiées.
- l'annuaire des répertoires de données scientifiques Open Access Directory
- le site datahub.io (Open Knowledge Foundation)
- en France, le site Nosdonnees.fr (maintenu par Regards Citoyens et l'Open Knowledge Foundation France)
- ils recensent les sources de données, ou parfois les jeux de données eux-mêmes.

Chaîne de traitement des données

- il est important de bien documenter la provenance des données (l'origine et l'historique d'un jeu de données).
- chaque utilisateur qui a modifié un jeu de données doit pouvoir être identifié.
- il est responsable des traitements et des nettoyages des données qu'il a effectués.

Outils

- outil dédié (entreprise)
- tableurs : Excel / LibreOffice Calc
- ligne de commande (Linux OSX surtout)
- système de gestion de base de données (SGBD)
- langage de programmation

Plan

3 Tableur

Tableur

- L'outil le plus basique pour la manipulation des données est la feuille de calcul.
- Les données contenue dans une feuille de calcul sont dans un format structuré, lisible par les machines, qui peut être trié et filtré.
- Avec un tableur on utilise de simples feuilles de calcul,
- un outil cependant puissant pour faire des opérations basiques (trouver le total, la moyenne, etc.), appliquer un traitement de masse, ou créer des graphiques et des tableaux.

Tableurs

- Une grande variété de tableurs et d'applications existent.
- Microsoft Office propose Excel,
- LibreOffice avec Calc
- Google Docs a aussi un tableur
- etc.

Comparatif rapide

	Google Spreadsheet	LibreOffice Calc	Excel
Usage	gratuit	gratuit	commercial
Stocakge	Google Drive	Disque dur	Disque dur
Internet requis	oui	non	non
Installation requise	non	oui	oui
Collaboration	oui	non	non

Demo

- Demo