



Big Data: Méthodes et analyses

Visualisation

Visseho Adjewanou, PhD.

Démographe, Département de Sociologie, UQAM

17 June 2024



Remarques

- Présentation adaptée du livre de Foster et al. 2020.



Plan de présentation

- ➊ Introduction à la visualisation des données
- ➋ Outils de visualisation de données
- ➌ Développer des visualisations efficaces
- ➍ Une taxonomie des graphiques
- ➎ Défis

Objectifs du cours

Ce chapitre vous montrera comment utiliser la visualisation pour explorer les données ainsi que pour communiquer les résultats afin que les données puissent être transformées en informations interprétables et exploitables. Il existe de nombreuses façons de présenter des informations statistiques qui transmettent le contenu de manière rigoureuse. L'objectif de ce chapitre est de présenter un aperçu introductif des techniques de visualisation efficaces pour une gamme de types de données et de tâches, et d'explorer les fondations et les défis de la visualisation de l'information à différentes étapes d'un projet.



Objectifs du cours

- Comprendre les principes fondamentaux de la visualisation des données



Objectifs du cours

- Comprendre les principes fondamentaux de la visualisation des données
- Apprendre à utiliser des outils de visualisation courants



Objectifs du cours

- Comprendre les principes fondamentaux de la visualisation des données
- Apprendre à utiliser des outils de visualisation courants
- Appliquer des techniques de visualisation pour explorer et présenter des jeux de données de grande taille



Objectifs du cours

- Comprendre les principes fondamentaux de la visualisation des données
- Apprendre à utiliser des outils de visualisation courants
- Appliquer des techniques de visualisation pour explorer et présenter des jeux de données de grande taille
- Évaluer l'efficacité des visualisations

Section 1

1. Introduction

- L'une des découvertes les plus célèbres en science—que les maladies étaient transmises par des germes plutôt que par la pollution—résulte des insights tirés d'une visualisation de l'emplacement des décès dus au choléra à Londres près d'une pompe à eau [[@snow1855mode](#)].
 - La visualisation de l'information au vingt-et-unième siècle peut être utilisée pour générer des insights similaires :
 - détecter la fraude financière,

- L'une des découvertes les plus célèbres en science—que les maladies étaient transmises par des germes plutôt que par la pollution—résulte des insights tirés d'une visualisation de l'emplacement des décès dus au choléra à Londres près d'une pompe à eau [[@snow1855mode](#)].
 - La visualisation de l'information au vingt-et-unième siècle peut être utilisée pour générer des insights similaires :
 - détecter la fraude financière,
 - comprendre la propagation d'une maladie contagieuse,

- L'une des découvertes les plus célèbres en science—que les maladies étaient transmises par des germes plutôt que par la pollution—résulte des insights tirés d'une visualisation de l'emplacement des décès dus au choléra à Londres près d'une pompe à eau [[@snow1855mode](#)].
 - La visualisation de l'information au vingt-et-unième siècle peut être utilisée pour générer des insights similaires :
 - détecter la fraude financière,
 - comprendre la propagation d'une maladie contagieuse,
 - repérer les activités terroristes, ou

- L'une des découvertes les plus célèbres en science—que les maladies étaient transmises par des germes plutôt que par la pollution—résulte des insights tirés d'une visualisation de l'emplacement des décès dus au choléra à Londres près d'une pompe à eau [[@snow1855mode](#)].
 - La visualisation de l'information au vingt-et-unième siècle peut être utilisée pour générer des insights similaires :
 - détecter la fraude financière,
 - comprendre la propagation d'une maladie contagieuse,
 - repérer les activités terroristes, ou
 - évaluer la santé économique d'un pays.

- Mais le défi est plus grand : de nombreux ($10^2 - 10^7$) éléments peuvent être manipulés et visualisés, souvent extraits ou agrégés à partir de jeux de données encore plus grands, ou générés par des algorithmes pour l'analyse.
 - Les outils de visualisation peuvent organiser les données de manière significative qui réduit **l'effort cognitif** et **analytique** nécessaire pour **comprendre** les données et **prendre des décisions** basées sur ces données.
 - Les utilisateurs peuvent scanner, reconnaître, comprendre et se souvenir des représentations visuellement structurées plus rapidement qu'ils ne peuvent traiter des représentations non structurées.
 - La science de la visualisation s'appuie sur de multiples domaines tels que la psychologie perceptuelle, les statistiques

Introduction

- Exemple: “le quartet d’Anscombe” [[@anscombe1973graphs](#)]

Data Set A	Data Set B	Data Set C	Data Set D
x	y	x	y
10	8.04	10	9.14
8	6.95	8	8.14
13	7.58	13	8.74
9	8.81	9	8.77
11	8.33	11	9.26
14	9.96	14	8.1
6	7.24	6	6.13
4	4.26	4	3.1
12	10.84	12	9.13
7	4.82	7	7.26
5	5.68	5	4.74
x	y	x	y

Statistical Analysis	Average	Variance	Correlation
	X	X	
	9	7.50	11
	X	Y	4.12

All four data sets have the same average, variance and correlation statistics.

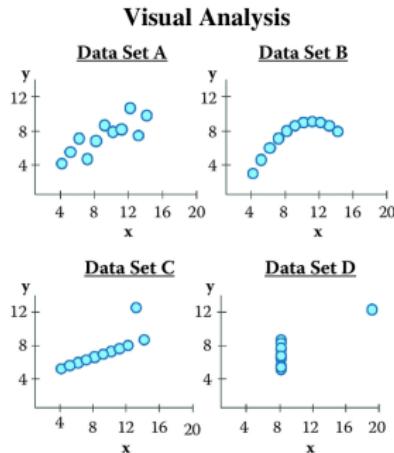


Figure 1: Adapted from Anscombe's quartet [[@anscombe1973graphs](#)]

- Description: La Figure @ref(fig), "le quartet d'Anscombe" [@anscombe1973graphs], fournit un exemple classique de la valeur de la visualisation par rapport à une analyse statistique descriptive de base. Le panneau de gauche inclut des données brutes de quatre petits ensembles de paires de nombres (A, B, C, D), qui ont la même moyenne, médiane et écart-type, et ont une corrélation entre les paires de nombres. Le panneau de droite montre ces ensembles de données visualisés avec chaque point tracé sur des axes perpendiculaires (diagrammes de dispersion), révélant des différences dramatiques entre les ensembles de données, des tendances et des valeurs aberrantes visuellement.

Introduction

- De manière générale, les visualisations sont utilisées :
 - soit pour présenter des résultats,
 - soit pour l'analyse et l'exploration ouverte.
- Ce cours donne un aperçu de la façon dont la visualisation moderne de l'information, ou l'exploration visuelle de données, peut être utilisée dans le contexte des big data.

Section 2

2. Outils



Outils dans R

Voici quelques packages couramment utilisés qui permettent de gérer et visualiser des données à grande échelle, ainsi que des exemples d'application :



ggplot2 :

- ggplot2 est une bibliothèque de visualisation de données qui permet de créer des graphiques élégants et informatifs.
- Bien que ggplot2 soit principalement utilisé pour des ensembles de données de taille modérée à grande, il peut être utilisé efficacement avec des données pré-agrégées ou filtrées pour produire des graphiques interactifs.



plotly :

- plotly est une bibliothèque R qui permet de créer des visualisations interactives et dynamiques.
- plotly est particulièrement efficace pour visualiser des données volumineuses grâce à ses capacités de zoom, de filtrage et d'interactivité.
- Il peut être utilisé pour des graphiques comme des scatterplots, des histogrammes interactifs et des heatmaps.

Shiny :

- Shiny est un framework R pour construire des applications web interactives.
- Il est souvent utilisé pour créer des tableaux de bord interactifs permettant d'explorer et de visualiser des données à grande échelle.
- Avec Shiny, vous pouvez construire des applications web qui chargent et traitent de grandes quantités de données en arrière-plan, puis permettent à l'utilisateur d'explorer ces données via des visualisations interactives.



data.table :

- data.table est une extension de data.frame en R qui offre des performances accrues pour le traitement de grands ensembles de données.
- data.table est utilisé pour l'extraction, le filtrage et l'agrégation rapides des données avant la visualisation.
- Il est particulièrement efficace pour manipuler des millions de lignes de données de manière efficace et rapide.

dygraphs :

- dygraphs est une bibliothèque R pour créer des graphiques interactifs de séries chronologiques.
- dygraphs peut être utilisé pour visualiser des séries chronologiques de données volumineuses, telles que des données économiques ou des données de marché, tout en permettant à l'utilisateur d'explorer les détails à différentes échelles de temps.



SparkR :

- SparkR est une interface R pour Apache Spark, permettant de traiter et d'analyser des données distribuées à grande échelle.
- SparkR peut être utilisé pour l'analyse et la visualisation de grands ensembles de données directement dans R, en exploitant la capacité de Spark à gérer et traiter des données massives de manière distribuée.



Autres applications

- D3.js (Data-Driven Documents) est une bibliothèque JavaScript populaire pour la manipulation de documents basés sur des données.
- Elle est largement utilisée pour créer des visualisations interactives et dynamiques directement dans le navigateur web.
- D3.js permet de manipuler efficacement des millions d'éléments graphiques sur de grands écrans, et offre une flexibilité pour gérer la perception et l'interaction.



Autres applications

- Tableau est une plateforme d'analyse visuelle qui prend en charge l'exploration interactive de grands ensembles de données.
- Il offre des fonctionnalités avancées pour la visualisation des données, y compris des techniques d'agrégation et de filtrage compactes des résultats de requêtes.



Section 3

3. Développer des visualisations efficaces



Visualisations efficaces

- L'efficacité d'une visualisation dépend à la fois :
 - des besoins en analyse et
 - des objectifs de conception.
- Parfois, les questions sur les données sont connues à l'avance ; dans d'autres cas, le but peut être d'explorer de nouveaux ensembles de données, de générer des insights et de répondre à des questions inconnues avant de commencer l'analyse.



Visualisations efficaces

- Le développement d'une visualisation efficace est un processus continu qui inclut généralement les activités suivantes :
- ① Spécifier les besoins des utilisateurs, les tâches, les exigences en matière d'accessibilité et les critères de réussite.



Visualisations efficaces

- Le développement d'une visualisation efficace est un processus continu qui inclut généralement les activités suivantes :
- ① Spécifier les besoins des utilisateurs, les tâches, les exigences en matière d'accessibilité et les critères de réussite.
- ② Préparer les données (nettoyer, transformer).



Visualisations efficaces

- Le développement d'une visualisation efficace est un processus continu qui inclut généralement les activités suivantes :
- ① Spécifier les besoins des utilisateurs, les tâches, les exigences en matière d'accessibilité et les critères de réussite.
- ② Préparer les données (nettoyer, transformer).
- ③ Concevoir des représentations visuelles.



Visualisations efficaces

④ Concevoir l'interaction.



Visualisations efficaces

- ④ Concevoir l'interaction.
- ⑤ Planifier le partage des insights et la traçabilité.



Visualisations efficaces

- ④ Concevoir l'interaction.
- ⑤ Planifier le partage des insights et la traçabilité.
- ⑥ Prototyper/évaluer, y compris les tests d'utilisabilité.



Visualisations efficaces

- ④ Concevoir l'interaction.
- ⑤ Planifier le partage des insights et la traçabilité.
- ⑥ Prototyper/évaluer, y compris les tests d'utilisabilité.
- ⑦ Déployer (surveiller l'utilisation, fournir un support aux utilisateurs, gérer le processus de révision).



Visualisations efficaces

- La conception, le développement et l'évaluation d'une visualisation sont guidés par la compréhension des antécédents et des objectifs du public cible.
- Si l'objectif est de **présenter des résultats**, il existe un large éventail d'utilisateurs et de nombreuses options.
- Si le public est large, des **infographies** peuvent être développées par des graphistes, comme décrit dans des textes classiques (voir @few2009now; @edward2001visual; @edward2006beauty ou les exemples compilés par @harrison2015infographic; @harrisonweb).

Visualisations efficaces

- Si, en revanche, le public est composé d'experts du domaine intéressés par la surveillance continue de l'état général des processus dynamiques, des **tableaux de bord de surveillance** avec peu ou pas d'interactivité peuvent être utilisés.
 - Surveillance des ventes,
 - Nombre de tweets à propos de personnes,
 - Symptômes de la grippe et leur comparaison avec une base de référence [@few2013information].
- De tels tableaux de bord, composés de plusieurs graphiques de différentes données opérationnelles, peuvent augmenter la conscience de la situation afin que les problèmes puissent être remarqués et résolus tôt et que de meilleures décisions puissent être prises avec des informations à jour.

Visualisations efficaces

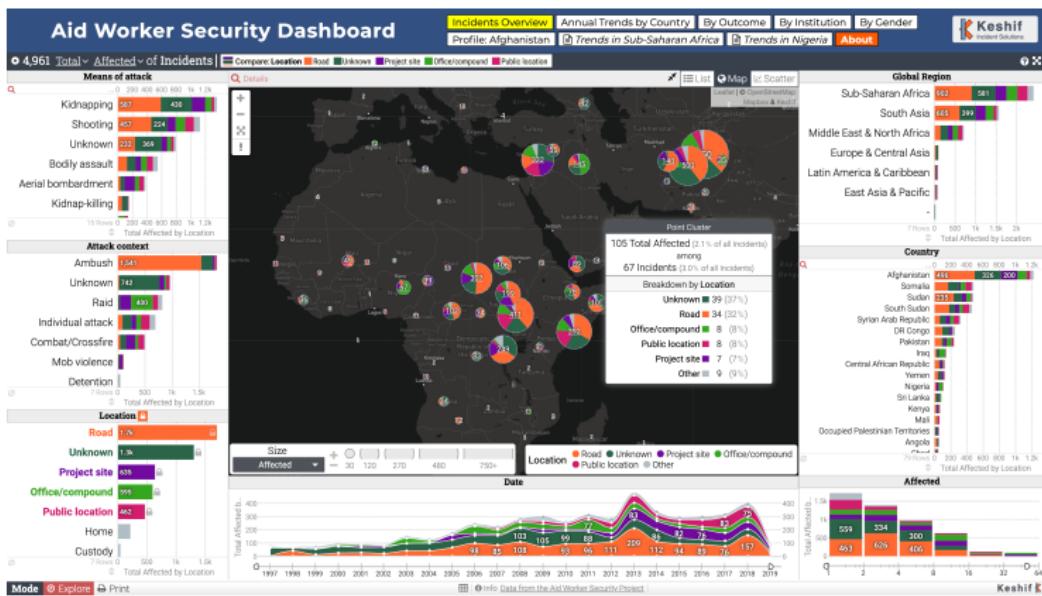


Figure 2: Tableau de bord d'analyse des incidents de sécurité des travailleurs humanitaires (<https://gallery.keshif.me/AidWorkerSecurity>)

Visualisations efficaces

- Un autre objectif de la visualisation est de permettre **l'analyse exploratoire interactive**.
- Cette approche va au-delà d'un instantané visuel des données pour la présentation et fournit de nombreuses fenêtres sur différentes parties et relations au sein des données à la demande.
- Des solutions sur mesure peuvent se concentrer sur des tâches spécifiques de requête et de navigation des données spécifiques.
- Par exemple, le BabyNameVoyager (<http://www.babynamewizard.com/voyager/>) permet aux utilisateurs de taper un nom et de voir un graphique de sa popularité au cours du siècle dernier. À chaque lettre tapée, la page filtre les prénoms commençant par l'entrée (comme

Visualisations efficaces

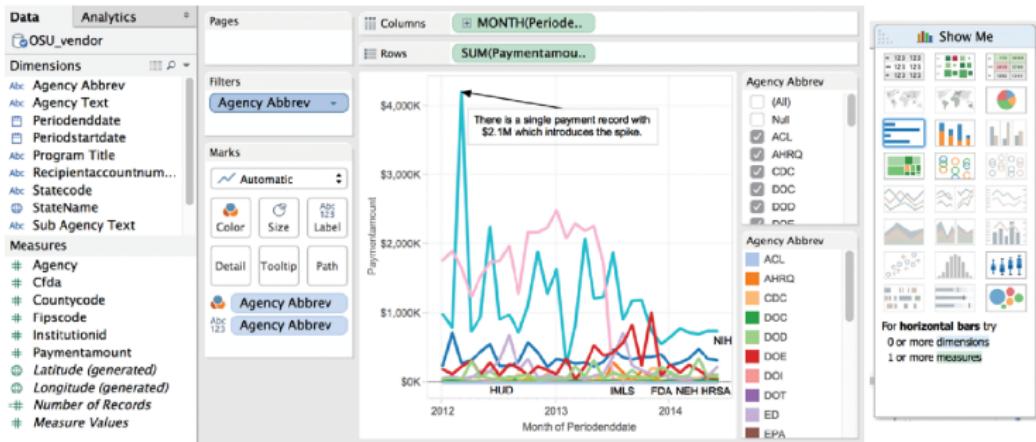


Figure 3: Charting interface of Tableau

Visualisations efficaces

- Pour créer des graphiques et des tableaux de bord interactifs à partir de nouveaux ensembles de données pour l'analyse, des produits et outils tels que Tableau, PowerBI, Keshif, et d'autres (voir la section Ressources), offrent une gamme de types de graphiques avec divers paramètres, ainsi que des environnements de conception visuelle qui permettent de combiner et de partager ces graphiques dans des tableaux de bord puissants.
- Par exemple, la Figure @ref(fig) montre l'interface de création de graphiques de Tableau sur un ensemble de données de transactions. Le panneau de gauche montre la liste des attributs associés aux transactions des fournisseurs pour une université donnée.
- La visualisation (au centre) est construite en plaçant le mois

Visualisations efficaces

- Ce graphique peut être combiné avec d'autres graphiques axés sur d'autres aspects dans des tableaux de bord interactifs.
- La Figure @ref(fig) montre une carte arborescente [@johnson1991tree] pour la répartition des dépenses par agence et sous-agence, combinée avec une carte montrant la dépense moyenne par état.
- L'état de l'Oklahoma se distingue par des dépenses peu nombreuses mais importantes.
- Le survol de la souris sur l'Oklahoma révèle les détails de ces dépenses.
- Un histogramme supplémentaire fournit un aperçu de l'évolution des dépenses sur trois ans.

Visualisations efficaces

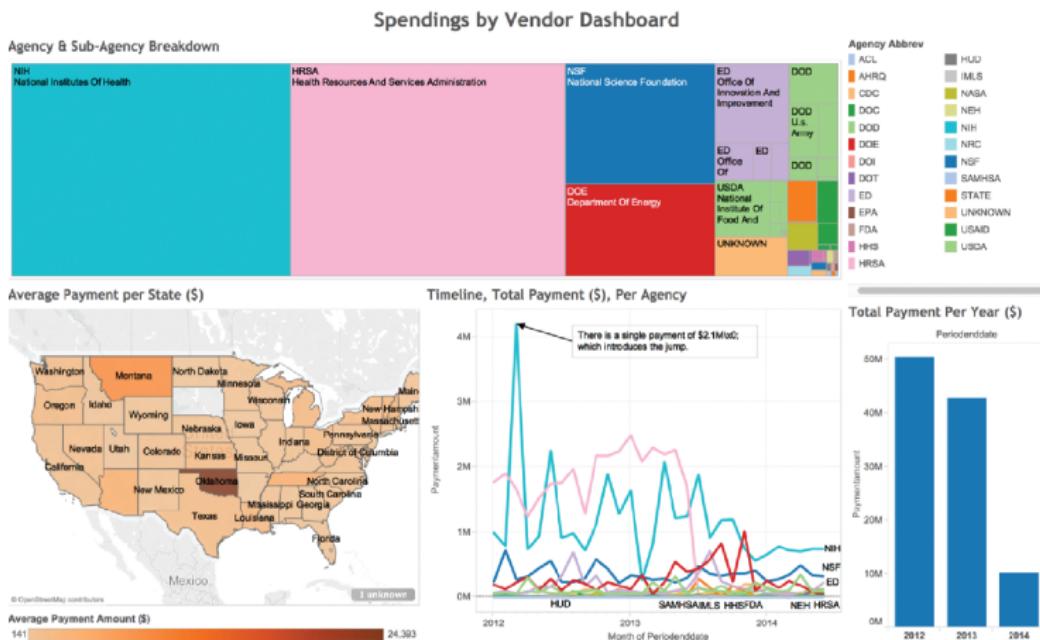


Figure 4: Une visualisation en carte arborescente (treemap) de la

Visseho Adjewanou, PhD.

Big Data: Méthodes et analyses



Visualisations efficaces

- Créer des visualisations efficaces nécessite une considération attentive de nombreux composants.

>- la position,
>- la longueur,
>- la couleur,
>- l'angle,
>- la surface et,
>- la texture (Figure @ref(fig); voir aussi @cleveland1984g)



Visualisations efficaces

- Créer des visualisations efficaces nécessite une considération attentive de nombreux composants.
- Les valeurs des données peuvent être encodées à l'aide d'un ou plusieurs éléments visuels, tels que:

>- la position,
>- la longueur,
>- la couleur,
>- l'angle,
>- la surface et,
>- la texture (Figure @ref(fig); voir aussi @cleveland1984g)



Visualisations efficaces

- Chacun de ces éléments peut être organisé de multiples façons, discutées plus en détail par Munzner [-@munzner2014visualization].



Visualisations efficaces

- Chacun de ces éléments peut être organisé de multiples façons, discutées plus en détail par Munzner [-@munzner2014visualization].
- En plus de l'encodage visuel des données, **des unités pour les axes**, **des étiquettes** et **des légendes** doivent être fournies ainsi que des explications des correspondances lorsque la conception est non conventionnelle.



Visualisations efficaces

- Chacun de ces éléments peut être organisé de multiples façons, discutées plus en détail par Munzner [-@munzner2014visualization].
- En plus de l'encodage visuel des données, **des unités pour les axes**, **des étiquettes** et **des légendes** doivent être fournies ainsi que des explications des correspondances lorsque la conception est non conventionnelle.
- Un exemple visuellement convaincant est la section « comment lire ces données » du projet « A world of Terror » par Periscopic (<https://terror.periscopic.com/>).



Visualisations efficaces

- Chacun de ces éléments peut être organisé de multiples façons, discutées plus en détail par Munzner [-@munzner2014visualization].
- En plus de l'encodage visuel des données, **des unités pour les axes**, **des étiquettes** et **des légendes** doivent être fournies ainsi que des explications des correspondances lorsque la conception est non conventionnelle.
- Un exemple visuellement convaincant est la section « comment lire ces données » du projet « A world of Terror » par Periscopic (<https://terror.periscopic.com/>).
- Des **annotations** ou des **commentaires** peuvent être utilisés pour guider l'attention du lecteur et décrire les insights/aperçus associés.



Visualisations efficaces

- Chacun de ces éléments peut être organisé de multiples façons, discutées plus en détail par Munzner [-@munzner2014visualization].
- En plus de l'encodage visuel des données, **des unités pour les axes**, **des étiquettes** et **des légendes** doivent être fournies ainsi que des explications des correspondances lorsque la conception est non conventionnelle.
- Un exemple visuellement convaincant est la section « comment lire ces données » du projet « A world of Terror » par Periscopic (<https://terror.periscopic.com/>).
- Des **annotations** ou des **commentaires** peuvent être utilisés pour guider l'attention du lecteur et décrire les insights/aperçus associés.

Visualisations efficaces

More accurate



Position

Length

Angle

Slope

Area

Volume

Color

Density

Visualisations efficaces

- Voici une liste succincte de directives :
 - fournir un feedback immédiat lors de l'interaction avec la visualisation ;
 - générer des vues étroitement couplées (c'est-à-dire que la sélection dans une vue met à jour les autres) ; et
 - utiliser un ratio élevé “de données à l'encre”[@edward2001visual].
 - Utiliser la couleur avec précaution et s'assurer que la visualisation est fidèle (par exemple, en évitant les biais perceptuels ou les distorsions).
 - Éviter l'utilisation de représentations en trois dimensions ou d'ornements, car comparer des volumes en 3D est perceptuellement difficile et l'occlusion pose problème.
 - Les étiquettes et légendes doivent être significatives, les mises en page nouvelles doivent être soigneusement expliquées, et les visualisations en ligne doivent s'adapter à différentes tailles



Section 4

4. Une taxonomie des données par tâches



Une taxonomie des données

- Nous présentons un aperçu des approches de visualisation pour six types de données courants :
 - multivariées,
 - spatiales,
 - temporelles,
 - hiérarchiques,
 - réseau, et
 - textuelles.
- Discussion des propriétés
- Questions analytiques courantes, et
- Présentation des exemples



Une catégorisation des tâches pour l'analyse visuelle des données

1. Sélectionner/Interroger

- Filtrer pour se concentrer sur un sous-ensemble des données
- Récupérer les détails d'un élément
- Utiliser la sélection liée pour traverser plusieurs graphiques
- Comparer à travers plusieurs sélections



Une catégorisation des tâches pour l'analyse visuelle des données

2. Naviguer

- Faire défiler le long d'une dimension (1D)
- Faire pivoter le long de deux dimensions (2D)
- Zoomer le long de la troisième dimension (3D)

Une catégorisation des tâches pour l'analyse visuelle des données

3. Dériver

- Agréger des groupes d'éléments et générer des caractéristiques
- Regrouper des groupes d'éléments par techniques algorithmiques
- Classer les éléments pour définir un ordre



Une catégorisation des tâches pour l'analyse visuelle des données

4. Organiser

- Sélectionner le type de graphique et les codages des données pour organiser les données
- Disposer plusieurs composants ou panneaux dans l'interface



Une catégorisation des tâches pour l'analyse visuelle des données

5. Comprendre

- Observer les distributions
- Comparer les éléments et les distributions
- Relier les éléments et les motifs



Une catégorisation des tâches pour l'analyse visuelle des données

6. Communiquer

Annoter les découvertes Partager les résultats Retracer l'historique des actions



1. Données multivariées (Multivariate data)

- Dans les données tabulaires courantes, chaque enregistrement (ligne) possède une liste d'attributs (colonnes), dont la valeur est principalement catégorielle ou numérique.
- L'analyse des données multivariées avec des types de données catégoriels et d'intervalle de base vise à comprendre les motifs au sein des attributs de données et entre eux.
- **Étant donné un plus grand nombre d'attributs, l'un des défis de l'exploration et de l'analyse des données est de sélectionner les attributs et les relations sur lesquels se concentrer.**
- L'expertise dans le domaine des données peut être utile pour cibler les attributs pertinents.

1. Données multivariées (Multivariate data)

- Les données multivariées peuvent être présentées sous différentes formes de graphiques en fonction des données et des relations explorées.
- Les graphiques unidimensionnels (1D) présentent les données sur un seul axe.
 - Un exemple est une boîte à moustaches (box-plot), qui montre les plages de quartiles pour les données numériques.
- Les graphiques dits **1,5D** listent la gamme des valeurs possibles sur un axe et décrivent une mesure des données sur l'autre.
 - Les graphiques à barres sont un type de graphique omniprésent qui peut visualiser efficacement les données numériques, par exemple, une note numérique par étudiant, ou la moyenne des notes pour des groupes d'étudiants agrégés par genre.
 - Les enregistrements peuvent également être regroupés sur des

1. Données multivariées (Multivariate data)

- Les graphiques bidimensionnels (2D), tels que les diagrammes de dispersion, tracent les données selon deux attributs, comme les diagrammes de dispersion.
- Les graphiques matriciels (grille) peuvent également être utilisés pour montrer les relations entre deux attributs.
- Les cartes de chaleur (heatmaps) visualisent chaque cellule de la matrice en utilisant des couleurs pour représenter sa valeur.
- Les matrices de corrélation montrent la relation entre les paires d'attributs.

1. Données multivariées (Multivariate data)

- Pour montrer les relations de plus de deux attributs (3D+), une option consiste à utiliser des codages visuels supplémentaires dans un seul graphique, par exemple, en ajoutant la taille/la forme des points comme variable de données dans les diagrammes de dispersion.
- Une autre option est d'utiliser des conceptions visuelles alternatives qui peuvent coder plusieurs relations dans un seul graphique.
 - Par exemple, un diagramme de coordonnées parallèles [@inselberg2009] a plusieurs axes parallèles, chacun représentant un attribut ;
 - Chaque enregistrement est montré sous forme de lignes connectées passant par les valeurs de l'enregistrement sur chaque attribut.
- Les graphiques peuvent également montrer des relations de

1. Données multivariées (Multivariate data)

- Enfin, une autre approche pour analyser les données multidimensionnelles consiste à utiliser des algorithmes de regroupement pour identifier des éléments similaires.
- Les clusters sont généralement représentés sous forme de structure arborescente (voir la section Données hiérarchiques).
 - Par exemple, le regroupement en k -moyennes commence par l'utilisateur spécifiant combien de clusters créer ;
 - L'algorithme place ensuite chaque élément dans le cluster le plus approprié.
- Des relations surprenantes et des valeurs aberrantes intéressantes peuvent être identifiées par ces techniques sur des algorithmes d'analyse mécanique.
- Cependant, de tels résultats peuvent nécessiter plus d'efforts pour être interprétés.



2. Données spatiale (Spatial data)

- Les données spatiales transmettent un contexte physique, généralement dans un espace 2D, comme les cartes géographiques ou les plans d'étage.
- Plusieurs des exemples les plus courants de visualisation de l'information incluent les cartes, de la représentation de la campagne russe malheureuse de Napoléon par Minard en 1861 (<https://gallica.bnf.fr/ark:/12148/btv1b52504201x>) à l'application interactive HomeFinder qui a introduit le concept de requêtes dynamiques [[@ahlberg1992dynamic](#)].
- Les tâches incluent la recherche d'éléments adjacents, de régions contenant certains éléments ou ayant des caractéristiques spécifiques, et de chemins entre les éléments — et l'exécution des tâches de base (voir Visualisations efficaces).

2. Données spatiale (Spatial data)

- La principale forme de visualisation des données spatiales est la carte.
- Dans les cartes **choroplèthes** (représente des données par des plages de valeurs discrétisées), le codage par couleur est utilisé pour représenter un attribut de données.
- Les cartogrammes visent à coder la valeur de l'attribut avec la taille des régions en déformant l'espace physique sous-jacent.
 - Un cartogramme est une carte pour laquelle une variable thématique, comme la population ou le PIB, remplace la surface des territoires représentés.
 - Exemple: <https://rgeomatic.hypotheses.org/1361>

2. Données spatiale (Spatial data)

- Les cartes en grille de tuiles réduisent chaque zone spatiale à une taille et une forme uniformes (par exemple, un carré) afin que les données codées par couleur soient plus faciles à observer et à comparer ;
 - Elles convertissent chaque région spatiale en une forme fixe, telle qu'une tuile carrée, et arrangent ces tuiles pour approximer et maintenir les positions physiques relatives des régions [@DeBelius2015; @ProtoVis2015].
 - Les cartes en grille facilitent également la sélection de zones plus petites (comme de petites villes ou des États).



2. Données spatiale (Spatial data)

- Les cartes de contours (isoplèthes) connectent des zones avec des mesures similaires et colorient chacune séparément.
- Exemple: <https://www.aquaportal.com/dictionnaire/definition/12888/isoplethe>
- Les **cartes de réseau** visent à montrer la connectivité entre les emplacements, comme les vols à destination/en provenance de nombreuses régions du monde.
- Les données spatiales peuvent également être présentées avec une emphase non spatiale (par exemple, comme une hiérarchie de continents, de pays et de villes, en utilisant un graphique en arbre).

2. Données spatiale (Spatial data)

CDC Home

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

United States Cancer Statistics: An Interactive Cancer Atlas (InCA)

1999-2012 Cancer Incidence and Mortality Data

Section 508 compliant version of these data are available on [United States Cancer Statistics](#) website
[Help](#) | [Glossary](#)

Cancer Event

Incidence Rate

Site All Cancer Sites Combined

Gender Male

Race/Ethnicity All Races

Period 2012

Classification Quantile

Classes 4

- [Make comparison](#)
- [Download data](#)
- [Print page](#)

Age-Adjusted Incidence Rate – All Cancer Sites Combined***

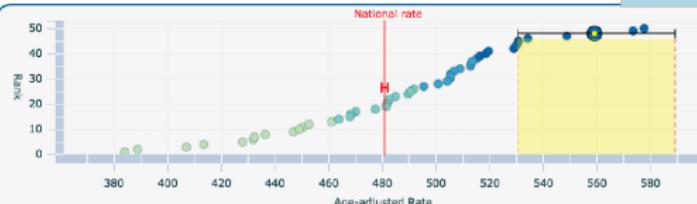
- 2012
- All Races
- Male



Play trend data

State	Rate	Lower CI	Upper CI	Count	Pop
US	483.0	481.9	484.2	767,366	155

AZ	388.8	382.2	395.4	13,612	▲
AR	512.9	501.6	524.3	8,185	▼
CA	446.9	443.8	450.2	78,991	▼
CO	431.9	423.4	440.5	10,685	▼
CT	515.1	504.7	525.6	9,919	▼
DE	548.8	528.0	570.2	2,749	▼
DC	559.1	530.4	589.0	1,513	▼
State: District of Columbia					
Rate: 559.1 (95% CI: 530.4, 589)					
Rank: 48 out of 50					



Footnotes

2. Données spatiale (Spatial data)

- Description: Les cartes sont souvent combinées avec d'autres visualisations. Par exemple, dans la Figure @ref(fig), l'Atlas du cancer des États-Unis combine une carte montrant les motifs à travers les États pour un attribut, avec un tableau triable fournissant des informations statistiques supplémentaires et un diagramme de dispersion permettant aux utilisateurs d'explorer les corrélations entre les attributs. Les Figures @ref(fig) et @ref(fig) démontrent également l'utilisation de différentes conceptions de cartes dans le cadre de solutions analytiques plus larges.

3. Données temporelles (temporal data)

- Le temps est la dimension unique dans notre monde physique qui s'écoule constamment vers l'avant. - Bien que nous ne puissions pas contrôler le temps, nous l'enregistrons fréquemment comme un point ou un intervalle.
- Les Figures @ref(fig) et @ref(fig) illustrent des graphiques en ligne montrant des tendances sur plusieurs années, chaque ligne représentant un sous-ensemble de données pour une comparaison croisée des tendances temporelles.
- Les données temporelles ont également plusieurs niveaux de représentation (année, mois, jour, heure, minute, etc.) avec des irrégularités (année bissextile, différents jours par mois, etc.).



3. Données temporelles (temporal data)

- Comme nous mesurons le temps en fonction des événements cycliques dans la nature (jour/nuit), nos représentations sont également souvent cycliques.
- Par exemple, janvier suit décembre (le premier mois suit le dernier).
- Cette nature cyclique peut être capturée par des codages visuels circulaires, tels que l'horloge conventionnelle avec les aiguilles des heures, des minutes et des secondes.

3. Données temporelles (temporal data)

- Les données de séries temporelles (Figures @ref(fig) et @ref(fig)) décrivent des valeurs mesurées à intervalles réguliers, telles que les données du marché boursier ou les données météorologiques.
- L'analyse vise à comprendre les **tendances** et les **anomalies** temporelles, à rechercher des **motifs spécifiques** ou à effectuer des **prédictions**.
- Pour montrer plusieurs tendances de séries temporelles dans différentes catégories de données dans une zone de graphique très compacte, chaque tendance peut être représentée avec une faible hauteur en utilisant une approche de couleur multicouche, créant des graphiques de type horizon.
- Bien que perceptuellement efficaces après avoir appris à lire leur codage, ces conceptions de graphiques peuvent ne pas

3. Données temporelles (temporal data)

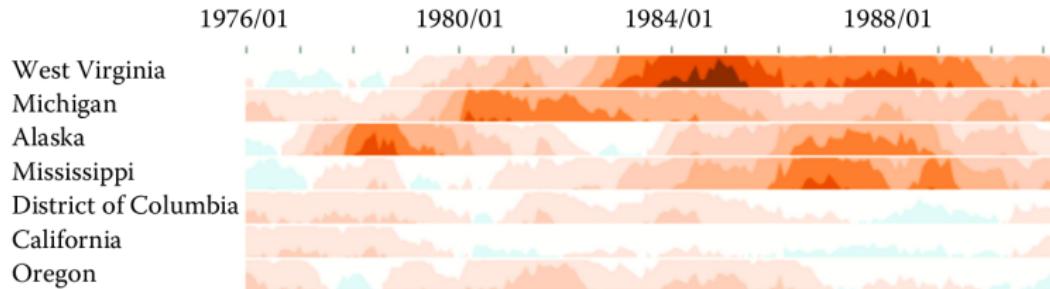


Figure 7: Horizon graphs used to display time series

3. Données temporelles (temporal data)

- Une autre forme d'analyse temporelle consiste à comprendre les **séquences d'événements**.
- L'étude de l'activité humaine inclut souvent l'analyse des séquences d'événements.
 - Par exemple, les dossiers des étudiants incluent des événements tels que la participation à l'orientation, l'obtention d'une note dans une classe, le départ en stage et la diplomation.
- Dans l'analyse des séquences d'événements, il est important de trouver les motifs les plus courants, de repérer les rares, de rechercher des séquences spécifiques ou de comprendre ce qui conduit à des types d'événements particuliers
 - Par exemple, quels événements mènent à l'abandon d'un étudiant, précèdent une erreur médicale ou la faillite d'une entreprise).

3. Données temporelles (temporal data)

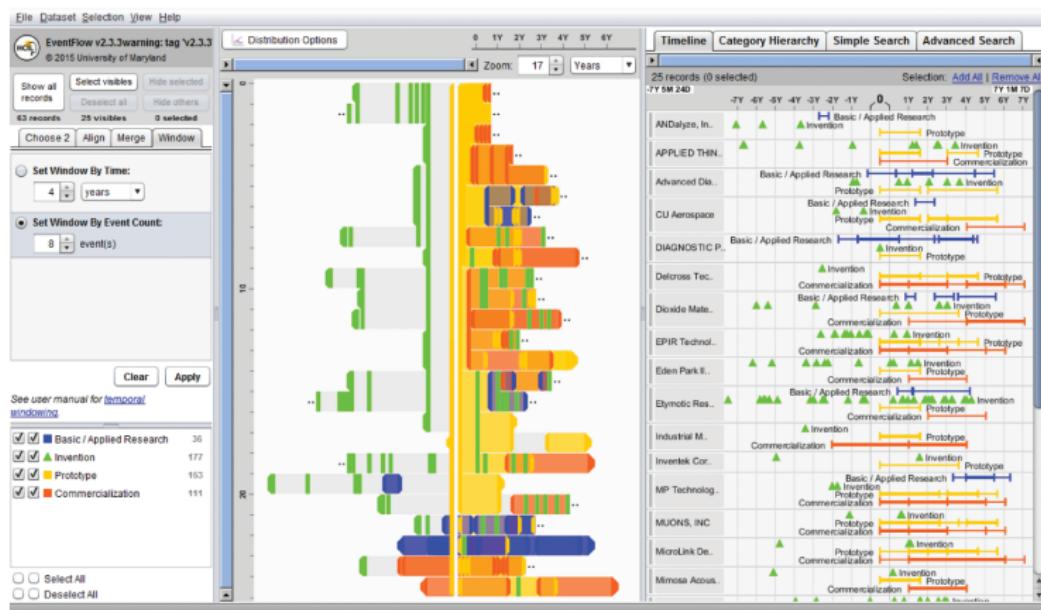


Figure 8: EventFlow (<https://hcil.umd.edu/eventflow/>) is used to visualize sequences of innovation activities by Illinois companies. Created Visseho Adjiwanou, PhD.

3. Données temporelles (temporal data)

- Description: La Figure @ref(fig) montre EventFlow utilisé pour visualiser les séquences d'activités d'innovation des entreprises de l'Illinois. Les types d'activités incluent la recherche, l'invention, le prototypage et la commercialisation. La chronologie (panneau de droite) montre la séquence d'activités pour chaque entreprise. Le panneau de vue d'ensemble (centre) résume tous les dossiers alignés par la première activité de prototypage de l'entreprise. Dans la plupart des séquences montrées ici, le premier prototype de l'entreprise est précédé par deux ou plusieurs brevets avec un décalage d'environ un an.

4. Données hiérarchiques (Hierarchical data)

- Les données sont souvent organisées de manière hiérarchique.
- Chaque élément apparaît dans un regroupement (par exemple, comme un fichier dans un dossier), et les groupes peuvent être regroupés pour former des groupes plus grands (par exemple, un dossier dans un dossier), jusqu'à la racine (par exemple, un disque dur).
- Les éléments et les relations entre les éléments et leur regroupement peuvent avoir leurs propres attributs.
 - Par exemple, la National Science Foundation est organisée en directions et divisions, chacune avec un budget et un nombre de bénéficiaires de subventions.

4. Données hiérarchiques (Hierarchical data)

- L'analyse peut se concentrer sur **la structure des relations**, avec des questions telles que:
 - Quelle est la profondeur de l'arbre ?,
 - Combien d'éléments cette branche a-t-elle ?, ou
 - Quelles sont les caractéristiques d'une branche par rapport à une autre ?.
- Dans de tels cas, la représentation la plus appropriée est généralement un **diagramme nœud-lien** [@plaisant2002spacetree; @card2002degree].

4. Données hiérarchiques (Hierarchical data)

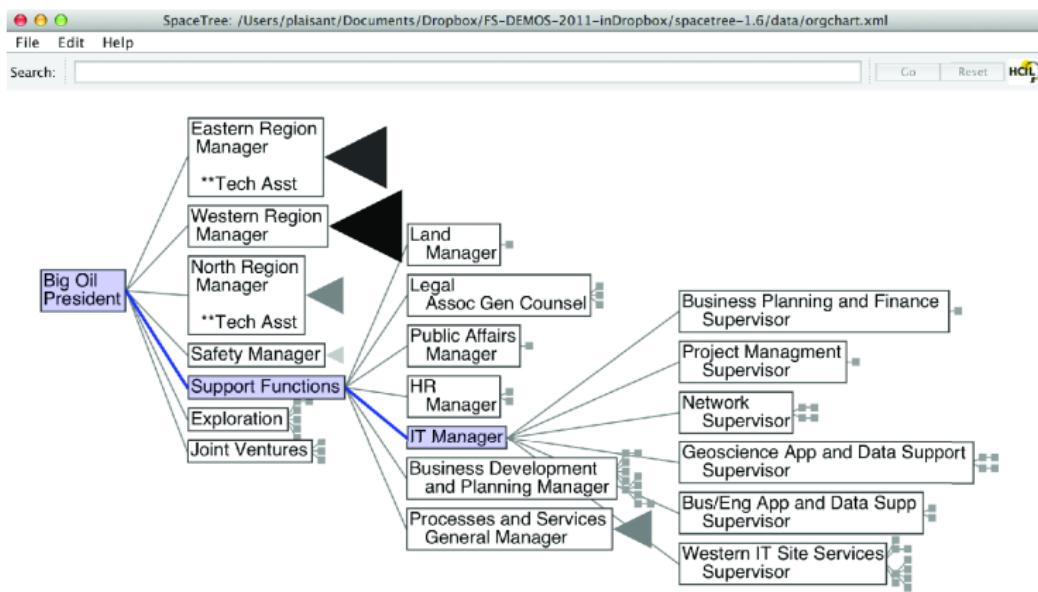


Figure 9: SpaceTree (<http://www.cs.umd.edu/hcil/spacetree/>)

4. Données hiérarchiques (Hierarchical data)

- Description: Dans la Figure @ref(fig), Spacetree est utilisé pour naviguer dans l'organigramme d'une entreprise. Comme tous les nœuds de l'arbre ne tiennent pas à l'écran, nous voyons une représentation iconique des branches qui ne peuvent pas être affichées, indiquant la taille de chaque branche. Lorsque les branches de l'arbre sont ouvertes ou fermées, la disposition est mise à jour avec des animations en plusieurs étapes pour aider les utilisateurs à rester orientés.—>

4. Données hiérarchiques (Hierarchical data)

- Lorsque la structure est moins importante mais que les valeurs des attributs des nœuds feuilles sont d'intérêt principal, les **treemaps**, une approche de remplissage de l'espace, sont préférables car elles peuvent montrer des arbres de taille arbitraire dans un espace rectangulaire fixe et mapper un attribut à la taille de chaque rectangle et un autre à la couleur.

4. Données hiérarchiques (Hierarchical data)



Figure 10: The Finviz treemap helps users monitor the stock market (<https://www.finviz.com/>)

4. Données hiérarchiques (Hierarchical data)

- Description: Par exemple, la Figure @ref(fig) montre la treemap Finviz qui aide les utilisateurs à surveiller le marché boursier. Chaque action est représentée par un rectangle. La taille du rectangle représente la capitalisation boursière, et la couleur indique si l'action monte ou descend. Les treemaps sont efficaces pour la prise de conscience situationnelle : on peut voir qu'aujourd'hui est une assez mauvaise journée car la plupart des actions sont en rouge (c'est-à-dire en baisse). Les actions sont organisées dans une hiérarchie d'industries, permettant aux utilisateurs de voir que la "technologie de la santé" ne se porte pas aussi mal que la plupart des autres industries. Les utilisateurs peuvent également zoomer sur le secteur de la santé pour se concentrer sur cette industrie.

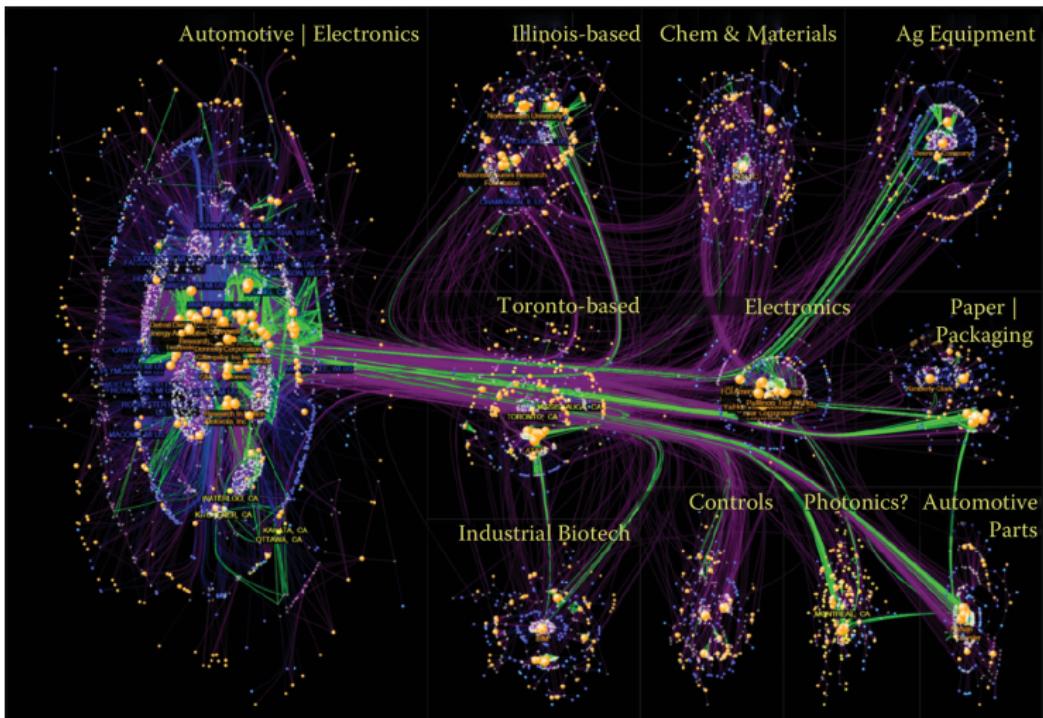
5. Données de réseaux (Network data)

- Les données de réseau encodent les relations entre les éléments par exemple, les modèles de connexion sociale (amitiés, abonnements et reposts, etc.), les modèles de voyage (comme les trajets entre stations de métro) et les modèles de communication (comme les courriels).
- Les aperçus de réseau tentent de:
 - révéler la structure du réseau,
 - montrer les clusters d'éléments liés (par exemple, des groupes de personnes étroitement connectées) et
 - permettre de tracer le chemin entre les éléments.
- L'analyse peut également se concentrer sur les attributs des éléments et des liens entre eux, tels que l'âge des personnes en communication ou la durée moyenne des communications.

5. Données de réseaux (Network data)

- Les diagrammes nœud-lien sont la représentation la plus courante des structures et aperçus de réseaux (Figures @ref(fig) et @ref(fig)). Ils peuvent utiliser des dispositions linéaires (arc), circulaires ou dirigées par la force pour positionner les nœuds (éléments).
- Les matrices ou dispositions en grille sont également une manière précieuse de représenter les réseaux [@henry2006matrixexplorer].
- Des solutions hybrides ont été proposées, avec des algorithmes d'ordonnancement puissants pour révéler des clusters [@hansen2010analyzing].
- Un défi majeur dans l'exploration des données de réseau est de gérer les réseaux plus grands où les nœuds et les arêtes se chevauchent inévitablement en raison de la structure

5. Données de réseaux (Network data)

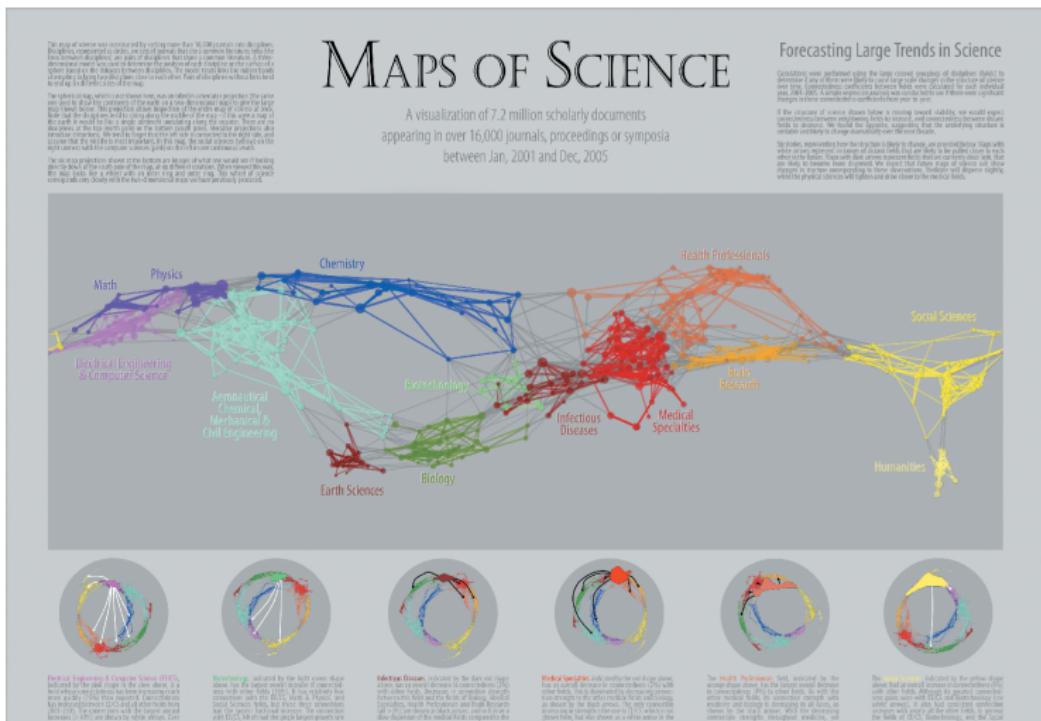


Visseho Adjewanou, PhD.

5. Données de réseaux (Network data)

- Description: La Figure @ref(fig) montre les réseaux d'inventeurs (blancs) et d'entreprises (oranges) et leurs connexions de brevetage (lignes violettes) dans la visualisation de réseau NodeXL. Chaque entreprise et inventeur est également connecté à un nœud de localisation (bleu = USA; jaune = Canada). Les lignes vertes sont des liens faibles basés sur le brevetage dans la même classe et sous-classe, et elles représentent des pistes potentielles de développement économique. Les plus grands clusters technologiques sont montrés en utilisant l'option de disposition group-in-a-box, ce qui rend les clusters plus visibles. Notez le niveau croissant de structure allant du cluster en bas à droite au cluster principal en haut à gauche. NodeXL est conçu pour l'exploration interactive des réseaux ; de nombreux contrôles (non montrés

5. Données de réseaux (Network data)



5. Données de réseaux (Network data)

- Description: La Figure @ref(fig) montre un exemple de visualisation de réseau sur la science en tant que sujet utilisé pour la présentation de données dans un livre et une exposition itinérante. Conçue pour les médias imprimés, elle inclut un titre clair et des annotations et montre une série de clusters de sujets en bas avec un résumé des insights recueillis par les analystes.->

6. Données de texyes (Text data)

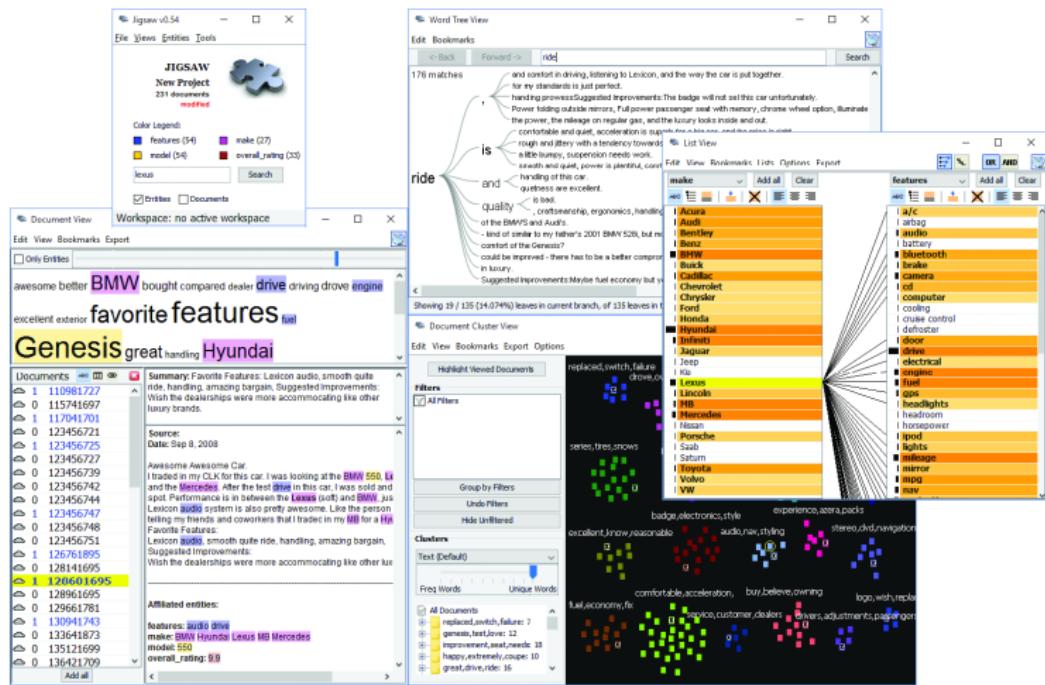
- Le texte est généralement prétraité (pour compter les mots/paragraphes, l'analyse des sentiments, la catégorisation, etc.) afin de générer des métadonnées sur les segments de texte, qui sont ensuite visualisées (voir cours sur text mining).
- Des visualisations simples comme les nuages de tags affichent des statistiques sur l'utilisation des mots dans une collection de textes, ou peuvent être utilisées pour comparer deux collections ou segments de texte.
- Bien qu'attrayantes visuellement, elles peuvent facilement être mal interprétées et sont souvent remplacées par des index de mots triés par un certain comptage d'intérêt.



6. Données de texyes (Text data)

- Les outils spécialisés d'analyse visuelle de texte combinent plusieurs visualisations des données extraites des collections de textes, telles que
 - des matrices pour voir les relations,
 - des diagrammes de réseau, ou
 - des coordonnées parallèles pour voir les relations entre entités (par exemple, entre quoi, qui, où et quand).
- Les chronologies peuvent être cartographiées sur la dimension linéaire du texte.

6. Données de texyes (Text data)



6. Données de texyes (Text data)

- Description de la figure: la Figure @ref(fig) montre un exemple utilisant Jigsaw [@stasko2008jigsaw] pour l'exploration des critiques de voitures. Les entités ont été extraites automatiquement (dans ce cas, marque, modèle, caractéristiques, etc.), et une analyse de cluster a été réalisée, visualisée en bas à droite. Une vue séparée (à droite) permet aux analystes d'examiner les liens entre les entités. Une autre vue permet de traverser les séquences de mots comme un arbre. Lire les documents originaux est essentiel, donc tous les éléments de visualisation sont liés au texte correspondant.



Section 5

Défis

Défis

Bien que la visualisation de l'information soit un outil puissant, il existe de nombreux obstacles à son utilisation efficace. Nous notons ici quatre domaines de préoccupation particulière : l'évolutivité, l'évaluation, les déficiences visuelles et la littératie visuelle.

1. Flexibilité/adaptabilité (Scalability)

- La plupart des visualisations traitent des ensembles de données relativement petits (entre mille et cent mille, parfois jusqu'à des millions, selon la technique), mais faire évoluer les visualisations de millions à des milliards d'enregistrements nécessite une coordination minutieuse des algorithmes d'analyse pour filtrer les données ou effectuer une agrégation rapide, des conceptions efficaces de résumé visuel, et un rafraîchissement rapide des affichages [[@shneiderman2008extreme](#)].
- Le mantra de recherche d'information visuelle, "Vue d'ensemble d'abord, puis zoom et filtrage, ensuite les détails à la demande," reste utile avec les données à grande échelle.

1. Flexibilité/adaptabilité (Scalability)

- Pour accommoder un milliard d'enregistrements, des marqueurs agrégés (qui peuvent représenter des milliers d'enregistrements) et des graphiques de densité sont utiles [@dunne2013motif].
- Dans certains cas, le volume important de données peut être agrégé de manière significative en un petit nombre de pixels.
- Un exemple en est Google Maps et sa visualisation des conditions routières. Un coup d'œil rapide à la carte permet aux conducteurs d'utiliser un résumé très agrégé de la vitesse d'un grand nombre de véhicules et seulement quelques pixels rouges suffisent pour décider quand prendre la route.

1. Flexibilité/adaptabilité (Scalability)

- Alors que des millions d'éléments graphiques peuvent être représentés sur de grands écrans [@fekete2002interactive], les problèmes de perception doivent être pris en considération [@yost2007beyond].
- L'extraction et le filtrage peuvent être nécessaires avant même de tenter de visualiser des enregistrements individuels [@wongsuphasawat2014using].
- Préserver des taux d'interaction élevés lors de l'interrogation de sources de données volumineuses est un défi, avec diverses méthodes proposées, telles que des approximations [@fisher2012trust] et le caching compact des résultats agrégés des requêtes [@lins2013nanocubes].
- Le chargement et le traitement progressifs aideront les utilisateurs à examiner les résultats au fur et à mesure de leur

2. Evaluation

- L'évaluation centrée sur l'humain des techniques de visualisation peut générer des évaluations qualitatives et quantitatives de leur qualité potentielle, les premières études se concentrant sur l'efficacité des variables visuelles de base [@mackinlay1986automating].
- À ce jour, les études utilisateurs restent au cœur de l'évaluation.
- Dans des environnements de laboratoire, des expériences peuvent démontrer une complétion de tâches plus rapide, des taux d'erreur réduits ou une satisfaction accrue des utilisateurs.
- Ces études sont utiles pour comparer les conceptions visuelles et d'interaction.
 - Par exemple, des études rapportent les effets de la latence sur

2. Evaluation

- Les évaluations peuvent également viser à mesurer et étudier la quantité et la valeur des insights révélés par l'utilisation d'outils de visualisation exploratoire [@saraiya2005insight].
- L'évaluation diagnostique de l'utilisabilité reste un pilier du design centré sur l'utilisateur.
- Des études d'utilisabilité peuvent être menées à divers stades du processus de développement pour vérifier que les utilisateurs peuvent accomplir des tâches de référence avec une vitesse et une précision adéquates.

2. Evaluation

- Des comparaisons avec la technologie précédemment utilisée par les utilisateurs cibles peuvent également être réalisées pour vérifier les améliorations.
- Les métriques doivent aborder l'apprentissage et l'utilité du système, en plus de la performance et de la satisfaction des utilisateurs [[@lam2012empirical](#)].
- L'enregistrement des données d'utilisation, les entretiens avec les utilisateurs et les enquêtes peuvent également aider à identifier des améliorations potentielles dans la conception de la visualisation et de l'interaction.

3. Déficience visuelle

- Les troubles de la perception des couleurs sont une condition courante qui doit être prise en considération [@olson1997evaluation].
- Par exemple, le rouge et le vert sont attrayants pour leur association intuitive à des résultats positifs ou négatifs (selon les associations culturelles); cependant, les utilisateurs atteints de daltonisme rouge-vert, l'une des formes les plus courantes, ne pourraient pas différencier clairement de telles échelles.
- Pour évaluer et aider à la conception visuelle sous différentes déficiences de couleur, des outils de simulation de couleur peuvent être utilisés.

3. Déficience visuelle

- L'impact des déficiences de couleur peut être atténué en :
 - sélectionnant soigneusement des schémas de couleurs limités,
 - utilisant le double encodage lorsque approprié (c'est-à-dire en utilisant des symboles qui varient à la fois par la forme et la couleur), et
 - permettant aux utilisateurs de modifier ou personnaliser les palettes de couleurs.
- Pour accommoder les utilisateurs ayant une vision réduite, des paramètres ajustables de taille et de zoom peuvent être utiles.
- Les utilisateurs avec des déficiences visuelles sévères peuvent nécessiter des conceptions d'interface et d'interaction axées sur l'accessibilité en priorité.

4. Littératie visuelle

- Alors que le nombre de personnes utilisant la visualisation continue de croître, tout le monde n'est pas capable d'interpréter avec précision les graphiques et les diagrammes.
- Lors de la conception d'une visualisation pour une population d'utilisateurs qui sont censés comprendre les données sans formation, il est important d'estimer correctement le niveau de littératie visuelle de ces utilisateurs.
- Même les simples graphiques de dispersion peuvent être accablants pour certains utilisateurs.

4. Littératie visuelle

- Des travaux récents ont proposé de nouvelles méthodes pour évaluer la littératie visuelle [@boy2014principled], mais les tests utilisateurs avec des utilisateurs représentatifs aux premiers stades de la conception et du développement resteront nécessaires pour vérifier l'utilisation de conceptions adéquates.
- Il est probable que des formations soient nécessaires pour aider les analystes à se familiariser avec les outils d'analyse visuelle.
- Des démonstrations vidéo enregistrées et un support en ligne pour répondre aux questions sont utiles pour amener les utilisateurs du niveau novice au niveau expert.

Conclusion

- L'utilisation de la visualisation de l'information se répand largement, avec un nombre croissant de produits commerciaux et d'ajouts aux packages statistiques désormais disponibles.
- Des tests utilisateurs approfondis doivent être menés pour vérifier que les présentations visuelles des données vont au-delà du simple désir d'attrait visuel en visualisation, et pour mettre en œuvre des conceptions qui ont démontré des avantages pour des tâches réalistes.
- La visualisation est de plus en plus utilisée par le grand public, et une attention particulière doit être portée à l'objectif de l'utilisabilité universelle afin que la plus large gamme d'utilisateurs puisse accéder et bénéficier des nouvelles approches de présentation des données et d'analyse interactive.

Resources

- Nous avons fait référence à divers manuels tout au long de ce chapitre.
- Les livres de Tufte restent des classiques, aussi inspirants à lire qu'instructifs [[@edward2001visual](#); [@edward2006beauty](#)].
- Nous recommandons également les livres de Few sur la visualisation de l'information [[@few2009now](#)] et la conception de tableaux de bord [[@few2013information](#)].
- “The Visual Display of Quantitative Information” par Edward R. Tufte
- “Storytelling with Data” par Cole Nussbaumer Knaflic
- <https://flowingdata.com/>
- <https://www.storytellingwithdata.com/podcast>
- Gephi pour la visualisation de données de réseaux
<https://gephi.org/>