h e g

Haute école de gestion
Genève

L'exploration du Big Data par sa visualisation – Application au projet GEoTweet

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Philippe JEANNERET

Conseiller au travail de Bachelor :

Rolf HAURI, Chargé d'enseignement HES

Genève, 22 juin 2015

Haute École de Gestion de Genève (HEG-GE)

Filière Informatique de Gestion

Hes·so GENÈVE
Haute Ecole Spécialisée
de Suisse occidentale

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de Bachelor of Science en Informatique de gestion.

L'étudiant a envoyé ce document par email à l'adresse remise par son conseiller au travail de Bachelor pour analyse par le logiciel de détection de plagiat URKUND, selon la procédure détaillée à l'URL suivante : http://www.urkund.fr/student_gorsahar.asp.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 21 juin 2015

Philippe Jeanneret

Remerciements

Je tiens à remercier Mr. Arnaud Gaudinat, Mme Fanny Béguelin, Mme Romaine et Kaufmann pour leur chaleureux accueil au sein de l'équipe GEoTweet. Ils ont su me faire sentir comme un membre à part entière de l'équipe dès mon arrivée.

Mes remerciements vont également à Mr Rolf Hauri, qui a accepté de suivre ce travail de Bachelor ainsi que me prodiguer des conseils ou pistes par rapport aux difficultés rencontrées.

Je suis reconnaissant du temps que m'a consacré Mr. Jean-Philippe Trabichet, responsable de la filière Informatique de Gestion à la Haute Ecole de Gestion de Genève, ainsi que ses assistants du LTI pour leurs précieux avis sur les interfaces à appliquer au projet GEoTweet.

Je remercie également Aloys Luber, étudiant en marketing digital à l'université CREA de Genève, et Michael Chrusciel, étudiant en sociologie, géographie et management à l'université de Neuchâtel. Leurs précieux avis ont permis d'élargir les pistes disponibles lorsque les choses semblaient confuses.

Finalement un grand merci à Michel Jeanneret pour ses nombreuses relectures et corrections sémantiques.

Résumé

Une multitude de données sont créées grâce à différents outils. Cela va du message écrit sur un réseau social au dernier achat réalisé à l'épicerie de quartier. Toute action laisse une trace digitale. Des données peuvent également être créées passivement comme la récolte d'informations provenant d'un capteur GPS dans véhicule. Tout ceci génère une énorme quantité de données. On estime qu'en 2015, 90% des données mondiales ont été créées au cours des deux dernières années. Cette manne d'informations s'appelle le Big Data.

Ces données permettent d'isoler des tendances comme les produits qui se vendent le mieux par période dans un supermarché ou les destinations qui attirent le plus de touristes en fonction de l'année. Pour répondre à ces questions, il est impensable devoir parcourir chaque enregistrement un à un. Il faut pouvoir trouver les visualiser pour, qu'en un coup d'œil, on puisse donner à sens à nos informations. Ce travail a pour but de proposer des visualisations en fonction du type de données auquel nous sommes confrontés ou bien de ce que nous souhaitons afficher.

Dans un second temps, nous nous pencherons plus en détail sur le projet « GEoTweet » qui a pour but d'utiliser l'énorme quantité données mises à disposition par Twitter. Il s'agit ici d'appliquer des méthodes de visualisations pour afficher des comportements d'utilisateur ou l'évolution de termes, comme les hashtags, grâce au Big Data généré par Twitter.

Table des matières

Déclaration i
Remerciements ii
Résumé iii
Liste des figures vi
1. Qu'est-ce que le Big Data 1
1.1 Qui utilise le Big Data 1
1.1.1 Histoire 1970 1
1.1.2 Histoire 1980 1
1.1.3 Histoire 1990 2
1.1.4 Histoire 2015 2
1.1.5 Impact sociologique 3
1.1.6 Les 4 V 4
1.1.7 Définition 5
2. Quelles sont les catégories existantes 6
2.1 Catégorisation générale 6
2.1.1 Structurée et semi-structurée 6
2.1.2 Non-structurée 8
2.2 Catégorisation IBM 9
3. Quelles sont les familles d'interfaces existantes pour visualiser les Big Data 11
3.1 Affichage classique 11
3.2 Affichage moderne 14
3.2.1 Carte 15
3.2.2 Texte 16
3.2.3 Données 17
3.2.3.1 Comparaison 18
3.2.3.1.1 Entre les instances 18
3.2.3.1.2 En fonction du temps 19
3.2.3.2 Distribution 19
3.2.3.2.1 Une variable 20
3.2.3.2.2 Deux variables 21
3.2.3.3 Relation 21
3.2.3.3.1 Deux variables 21
3.2.3.3.2 Trois variables 22
3.2.3.4 Composition 22
3.2.3.4.1 Statique dans le temps 22
3.2.3.4.2 Evoluant dans le temps 23
3.2.3.5 Connexion 24
3.2.3.6 Cartographique de fond 26
3.2.3.7 Animation 27
3.2.3.8 Infographie 28
3.2.4 Choisir le bon graphique 29

3.2.5 Bonnes pratiques 30
3.2.6 Erreurs à ne pas commettre 30
3.2.6.1 Comment fausser un graphique 31
4. Quelles sont les technologies 33
5. Cas d'étude GEoTweet 35
5.1 Besoins 35
5.2 GEoTweet et le Big Data 36
5.2.1 4 V 36
5.2.2 Catégorisation IBM 37
5.3 Interfaces pertinentes 38
5.3.1 Répartition des langues à Genève 38
5.3.1.1 Thermique 38
5.3.1.1.1 Carte 38
5.3.1.1.2 Classement 39
5.3.1.1.3 Chronologie 39
5.3.1.1.4 Rejouer 39
5.3.1.2 Quartiers 40
5.3.1.2.1 Carte 40
5.3.1.2.2 Classement 40
5.3.1.2.3 Chronologie 40
5.3.1.2.4 Derniers tweets 40
5.3.1.2.5 Ecrire un tweet 40
5.3.2 Rayonnement de Genève dans le monde 40
5.4 Proposition de vues – prototypes 41
5.4.1 Répartition des langues à Genève 41
5.4.2 Rayonnement de Genève dans le monde 43
5.5 Choix technologiques 44
5.6 Validation par l'équipe GEoTweet et analyse des résultats obtenus 44
6. Conclusion 46
Bibliographie 48

Liste des figures

Figure 1 : Google Trend sur le terme « Big Data » 3
Figure 2 : Classification du Big Data selon IBM 10
Figure 3 : Feuille de calcul représentant des statistiques de ventes 12
Figure 4 : Diagramme circulaire sur les proportions d'achats de chaque client.....13
Figure 5 : Graphe à barres des ventes 14
Figure 6 : Carte avec de simples points géographiques 15
Figure 7 : NYC Crime Map 15
Figure 8 : Nuage de mot du discours prononcé par Obama lors de sa victoire aux élections de 2008 17
Figure 9 : Graphique en colonne 18
Figure 10 : Graphique à lignes 19
Figure 11 : Histogramme 20
Figure 12 : Nuage de points 21
Figure 13 : Nuage de points 22
Figure 14 : Diagramme de zones empilées 23
Figure 15 : Regroupement en cercles 24
Figure 16 : Diagramme d'accords.....25
Figure 17 : Cartographie de fond.....26
Figure 18 : Animation des vents 27
Figure 19 : Infographie de la politique américaine 28
Figure 20 : Quel graphique choisir.....29
Figure 21 : Pertes d'emplois selon Palosi.....31
Figure 22 : Pertes d'emplois selon Cage 31
Figure 23 : Graphique manipulé présent sur le site du parti Républicain 32
Figure 24 : Diagramme de Sankey 34
Figure 25 : Carte thermique de la répartition des langues 41
Figure 26 : Carte de la répartition des langues par quartiers 42
Figure 27 : Carte du rayonnement de Genève dans le monde 43

1. Qu’est-ce que le Big Data

1.1 Qui utilise le Big Data

1.1.1 Histoire 1970

Dans les années 1970, les principaux fabricants de produits de grande consommation, comme P&G, Unilever et Kraft ainsi que les grandes surfaces, construisaient leurs stratégies marketing en fonction d’audits bimensuels fournis par la compagnie Nielsen. Cette dernière expédiait des employés dans plusieurs boutiques réparties dans douze villes des Etats-Unis uniquement. Ils avaient pour tâche de réaliser un audit en relevant la quantité de produits sur les étalages, leurs prix, la taille de l’espace qui leur était alloué et les rabais qui y étaient liés. Ces données étaient ensuite transmises aux fabricants et aux détaillants. Un fournisseur pouvait ainsi voir là où se situaient ses produits par rapport à ses concurrents et décider des mesures marketing à mettre en œuvre telles que l’ajustement du prix, des dépenses promotionnelles ou la création d’un nouveau produit.

1.1.2 Histoire 1980

A la fin des années 1980, la société IRI répandit les scanners de codes-barres afin de bouleverser le processus de la grande distribution. Les chaînes de magasins ont alors connu un tournant grâce à l’accès à une quantité de nouvelles données récoltées par les scanners dans chacune de leurs enseignes. Cette manne d’information, croisée avec celle traitant de la fidélité des clients, a offert aux distributeurs l’opportunité d’atteindre une nouvelle profondeur de l’information. Les agissements des consommateurs ont pu être visualisés et les tendances des marchés interprétées, ce qui se traduisit par un net avantage concurrentiel.

Le magazine Fortune classa ainsi Sam Walton, fondateur de la chaine de magasins Wal-Mart, dans les 12 plus grands entrepreneurs de notre ère pour avoir su à l’époque transformer son modèle économique en se basant sur le Big Data.

« La pierre angulaire du succès de l’entreprise était en fin de compte de vendre ses marchandises au prix le plus bas possible, chose qu’il a été à même de faire en écartant les intermédiaires et en négociant directement avec les fabricants pour diminuer les coûts. L’idée consistant à « acheter pas cher, empiler haut et vendre bon marché » est devenue un modèle économique durable en grande partie parce que Walton, avec David Glass qui lui succédera, a investi massivement dans un logiciel qui suivrait le comportement des clients en temps réel à partir des codes-barres lus aux caisses de Wal-Mart. »

Sam Walton a partagé ces données en temps réel avec les fournisseurs pour créer des partenariats qui ont permis à Wal-Mart d'exercer une pression significative sur les fabricants afin qu'ils améliorent leur productivité et deviennent encore plus efficaces. A mesure que l'influence de Wal-Mart grandissait, son pouvoir de pratiquement dicter le prix, le volume, les délais de livraison, le conditionnement et la qualité de nombreux produits augmentait d'autant.

Résultat : Walton a inversé la relation fournisseur-distributeur »[1]

Ainsi, la chaîne de distribution Wal-Mart fut en mesure de connaître les produits qu'achetaient ses clients, les promotions qui fonctionnaient le mieux et les articles qui étaient généralement associés dans un même chariot. L'apparition des cartes de fidélité n'a fait que renforcer la qualité de ces informations en permettant à présent de connaître les produits et les prix qui attiraient chaque client de façon individuelle. L'enseigne a ainsi pu inverser le levier de pression qu'elle subissait de la part des fournisseurs et exiger la quantité d'articles qu'elle voulait vendre (prévision d'offre), à quel prix (amélioration du rendement) et les promotions à mettre en place.

Conscient que ces informations ont de la valeur, Wal-Mart vend aujourd'hui ces données aux fabricants via son service « Retail Link ». Cet outil permet aux fournisseurs de consulter les statistiques de leurs produits et donc d'être plus concurrentiels à leur tour en adaptant leur stratégie.

1.1.3 Histoire 1990

A la fin du XXème siècle, une vanne d'information similaire s'est ouverte aux commerces en ligne où chaque action de l'utilisateur a commencé à laisser une trace digitale qui a été stockée avant d'être déchiffrée : cliques, visites. Ces précieuses données ont permis aux e-commerces et aux agences publicitaires de comprendre les comportements d'achat de leurs utilisateurs afin d'obtenir une avance significative vis-à-vis des boutiques physiques.

1.1.4 Histoire 2015

Aujourd'hui, les données proviennent de nouvelles sources telles que les réseaux sociaux, les téléphones portables ou autres capteurs. Elles révèlent les préférences, affiliations et intérêts des consommateurs. Les téléphones mobiles permettent d'interagir encore plus rapidement avec le client en se basant sur la géolocalisation et ainsi lui proposer une offre toujours plus adaptée. Ainsi si Wal-Mart remarque que je suis un amateur de barbecue et croise des données météo avec ma géolocalisation pour

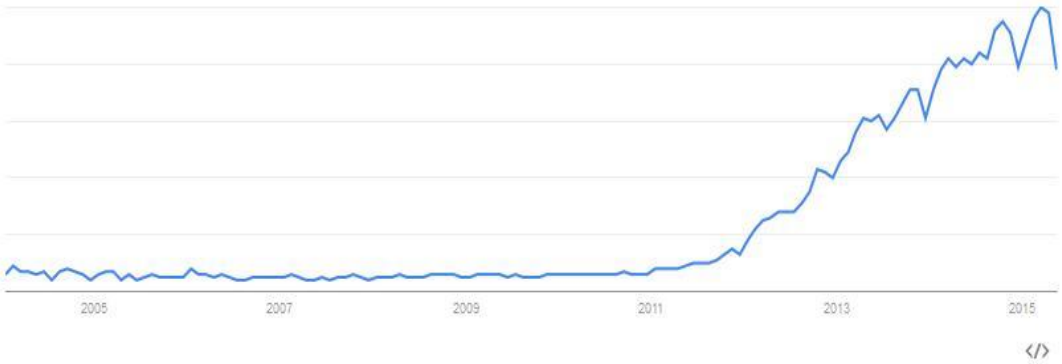
[1] The 12 greatest entrepreneurs of our time, John A. BYRNE

constater qu'il fait beau, je pourrais recevoir un bon de promotion sur du matériel pour grills si ces derniers sont en stock dans les magasins aux alentours. Toutes ces informations, d'apparences bénignes, ont une énorme valeur aux yeux d'Amazon, Google, Facebook et Wal-Mart, qui s'occupent de les revendre à prix d'or, comme dans l'exemple cité précédemment de « Retail Link ».

1.1.5 Impact sociologique

Depuis 2011, l'intérêt lié au Big Data a augmenté de façon exponentielle et il s'agit d'une des technologies informatique qui a reçu le plus d'éclairage de la part des médias. En effet, le Big Data est souvent sujets à controverse et des titres polémiques tel que « Le Big Data : le plus grand bien ou l'invasion de la vie privée ? » sont aujourd'hui monnaie courante dans la presse. L'explosion du Big Data est aussi constatable au travers Google Trends - site affichant la fréquence de recherche d'un terme dans le moteur de recherche en fonction des années – pour voir à quel point ce concept est récent.

Figure 1 : Google Trend sur le terme « Big Data »



Source : GOOGLE TRENDS, 3 mai 2015

Il est évident que le Big Data est lié à des questions socio-techniques, ne serait-ce que pour son côté intrusif dans la vie privée, mais sa définition reste floue. Les premières définitions existantes proviennent de différents domaines (marketing, informatique, …) et cette origine partagée conduit à de multiples définitions, ambigües et parfois contradictoires.

Le Big Data combine deux idées : le stockage de données et leur analyse. Malgré l'intérêt récent porté au Big Data, ces deux idées n'ont rien de nouveau et ont déjà été utilisées par le passé. Dès lors, en quoi le Big Data diffère-t-il ? La réflexion est à porter sur le terme « Big » qui met en valeur l'importance de la taille et son côté vaste, comparé aux techniques classiques de traitement des données qui travaillent sur des échantillons plus

petit. Pour avoir une réponse simplifiée à l'excès, il ne faut pas chercher plus loin que le terme « Big » qui définit la complexité, l'importance et le défi. Malheureusement, le terme « Big » renvoie également à la quantification et c'est là que la difficulté de trouver une définition apparaît.

1.1.6 Les 4 V

Les 4 V

L'une des définitions les plus citées est celle de Meta (aujourd'hui Gartner, entreprise de conseil et de recherche) qui, dans un rapport de 2001, ne faisait aucune mention explicite de « Big Data » - à l'époque ce terme n'était pas aussi courant qu'il ne l'est aujourd'hui. Gartner a proposé une définition appelée les « trois V » : Volume, Vélocité et Variété. Ces trois termes font référence à la quantité croissante de données, à la vitesse à laquelle elles sont produites et au large spectre de formats sous lesquelles elles peuvent être représentées. Bien qu'ils ne définissent pas explicitement ce qu'est le Big Data, ces trois propriétés permettent néanmoins de caractériser le Big Data. Cette définition a été reprise par le NIST et Gartner en 2012 et élargie par IBM afin d'y inclure un 4ème « V » pour « Véracité ». Cet indice aborde la question concernant la qualité des données traitées.

Nous avons donc quatre critères qui définissent les propriétés du Big Data :

- **Volume** : représente la quantité de données à laquelle nous avons accès. On parle ici d'unité de taille de fichiers, de nombre d'enregistrements, transactions, de fichiers ou de tables. Il s'agit du critère clé pour définir subjectivement s'il s'agit de Big Data.

- **Vélocité** : définit la fréquence à laquelle les données sont récupérées. Par exemple, la détection de fraudes, concernant les cartes de crédits, traque les patterns inhabituels de millions de transactions en temps quasi-réel.
- **Variété** : indique le type de données que nous possédons, comme des données financières, des photos, des données de capteurs, de la vidéo ou de l'audio. Il est possible de catégoriser ces données. Ce point sera détaillé au chapitre suivant
- **Véracité** : décrit la pureté des données générées. Un tweet comportant des abréviations, du langage familier aura par exemple une véracité faible.

Toutefois cette définition des 4 V est sujette à la libre interprétation, puisque aucune quantification numérique n'a été définie pour indiquer notre entrée dans du Big Data.

Ex : à partir de 10'000 enregistrements, c'est du Big Data.

1.1.7 Définition

Afin de comprendre à partir de quel moment il s'agit de Big Data, il faut mettre cette notion en perspective. Si on prend la quantité de données générées entre l'aube de la civilisation et 2010, une quantité équivalente est créée chaque minute aujourd'hui.

Dans le domaine de l'astronomie par exemple, le Sloan Digital Sky Survey a comme ambition de créer les cartes 3D les plus détaillées qui soient de l'univers. Ce projet a débuté en 2000 avec un télescope situé au Nouveau-Mexique. Ce dernier a récolté durant les premières semaines plus de données que ce qui a été accumulé jusqu'à ce moment dans toute l'histoire de l'astronomie. En une décennie, ce sont 140 téraoctets qui ont été sauvegardés, ce qui équivaut à 700'000 films. En 2016, lorsque le successeur du télescope du Nouveau-Mexique sera opérationnel, ce seront 140 téraoctets qui seront générés tous les cinq jours.

Ceci démontre qu'aujourd'hui nous avons les moyens de générer, garder et analyser ces données. La combinaison de ces trois facteurs est à la base de la création du Big Data. L'idée primaire se trouvant derrière le concept de Big Data est que tout ce que nous faisons ou entreprenons dans nos vies laisse ou laissera une trace digitale qui pourra être analysée et utilisée.

Le concept de Big Data présente des similitudes avec le Data Mining. Là où le Big Data se contente d'une vue d'ensemble de données collectées, le Data Mining va aller « creuser » ces données afin d'aider les preneurs de décisions à découvrir des pièces d'informations pour développer leur métier.

2. Quelles sont les catégories existantes

Nous avons vu précédemment que le Big Data pouvait être constitué de références d'articles, de vidéos, d'audio, de données traitant de l'astronomie, etc. Il est évident que la gamme d'informations est vaste et de nouveaux types de données apparaissent régulièrement. Il faut donc trouver un moyen pertinent pour regrouper les données de manière durable.

2.1 Catégorisation générale

La plupart des publications universitaires s'accordent pour catégoriser les données de Big Data en fonction de la structure des données et si elles sont internes à l'entreprise ou externes. David Meer, partenaire à Strategy's consumer et auteur de plusieurs articles sur le Big Data, spécifie cette première règle en sous-catégorisant les structures en fonction de l'origine des données.

2.1.1 Structurée et semi-structurée

Il s'agit de données, souvent gérées par SQL – Structured Query Language – ou tout autre langage similaire, possédant un modèle de données prédéfini ou organisé à l'avance.

Un modèle de données est un type de données métier qui sera sauvegardé selon une convention, puis traité. Prenons le cas d'une base de données de clients. Ces derniers seront caractérisés par leur nom, prénom, adresse, numéro de téléphone, etc. Parmi toutes ces conventions, certaines peuvent être restrictives comme le numéro de téléphone qui n'accepterait que des valeurs numériques ou encore un champ titre qui ne laisserait le choix qu'à Mr. Mme, Mlle, Dr.

Les données structurées donnent un nom à chaque champ qui sera saisi dans une base de données, ce qui la rend facile à sauvegarder et à analyser, le tout de manière « structurée ».

Les données semi-structurées sont à la croisée du structuré et du non-structuré qui est décrit plus loin. C'est une sorte de donnée qui peut avoir une structure qui peut être utilisé pour l'analyse, mais dont le modèle de données strict entier n'est pas structuré. Concernant le semi-structuré, les marqueurs de type tags ou autres permettent d'identifier des éléments de données, mais ces éléments n'ont pas de structure.

Prenons l'exemple d'un message Twitter ou diverses informations de premier niveau sont structurées : auteur, date, position géographique, longueur du message. En revanche, le contenu de ces informations est généralement non-structuré : s'agit-il d'un message, d'une photo, d'une vidéo ?

Qu'elles soient structurées ou semi-structurées, ces données peuvent provenir de cinq origines différentes :

- **Données crées** : elles sont créées, car elles n'existeraient pas si nous n'avions pas fait l'effort de poser des questions à des personnes ou tout simplement de sauvegarder des informations qu'ils auraient exprimé, comme lors d'un feedback form ou une étude de marché. Les données créées sont généralement structurées ou semi-structurées et peuvent être internes ou externes.
- **Données provoquées** : elles sont le fruit d'une demande explicite du point de vue d'une personne. L'exemple le plus commun est l'attribution d'une note à une application sur un smartphone. Les données provoquées sont généralement structurées ou semi-structurées et peuvent être internes ou externes.
- **Données transactionnelles** : elles sont générées à chaque fois qu'un utilisateur achète quelque chose ou clic sur un élément d'un site web. Elles résultent d'actions d'utilisateurs. Ces données sont cruciales lorsqu'elles sont croisées avec d'autres, comme la localisation ou la météo. Wal-Mart s'est par exemple rendu compte il y a quelques années que la vente de lampe de poche augmentait significativement avec l'émission d'alertes ouragans. Les données transactionnelles sont généralement structurées et internes.
- **Données compilées** : elles proviennent de sociétés, comme Acxiom, dont le but est de rassembler un maximum d'information sur les individus et compilent tout ce qu'elles trouvent : voitures enregistrées à leur nom, adresses, historique d'achats, etc. pour ensuite vendre ces éléments à des entreprises marketing qui se chargeront de les analyser. Les données compilées sont généralement structurées et externes
- **Données expérimentales** : elles sont un mélange de données créées et transactionnelles. Un exemple serait de créer des données marketings sur des clients fictifs et observer leurs résultats d'analyse dans le monde réel (transactionnelles). Ce cas de figure peut se produire dans des boutiques où une fois qu'on connaît le chemin emprunté par les clients dans le magasin, on peut tester différentes combinaisons d'étalages afin de voir laquelle se vendra le mieux. Les données expérimentales sont généralement structurées ou semi-structurées et peuvent être internes ou externes.

2.1.2 Non-structurée

Les données de type non-structurées sont les plus recherchées actuellement. Selon Bernard Marr, 80% des données non-structurées et semi-structurées ont une pertinence de type business. Ceci prouve que l'analyse et le croisement d'information ne connait aucune limite et que quasiment toutes les données récoltées sont utilisables.

Les données non-structurées comprennent tout ce qui est inconsistant, qui ne peut pas facilement être compartimenté en lignes, colonnes, champs et donc difficilement analysable.

Exemple de données non-structurées :

- Fichiers textes, emails
- Photos
- Vidéos
- Présentations PowerPoint
- Sites web
- Publications sur les réseaux sociaux
- Articles de blogs

Les données de type non-structurées sont tout le temps générées par l'utilisateur. Une nuance existe toutefois quant au fait qu'il est conscient ou non de leur création :

- **Données capturées** : elles sont récoltées passivement auprès des utilisateurs. Ceci inclut aussi bien des termes saisis dans un moteur de recherche que des données GPS d'un téléphone portable. Dans ces deux cas, l'utilisateur n'est pas forcément conscient qu'il génère des données. La quantité d'informations de type capturées a drastiquement augmentée ces dernières années. Les données capturées sont généralement non-structurées et peuvent être internes ou externes.
- **Données générées par l'utilisateur** : à l'inverse des données capturées, l'utilisateur est ici conscient qu'il créé du contenu. Ceci comprend des statuts Facebook, Tweets, vidéos mises en ligne sur YouTube et commentaires d'articles. Les données générées par l'utilisateur sont généralement non-structurées et peuvent être internes ou externes.

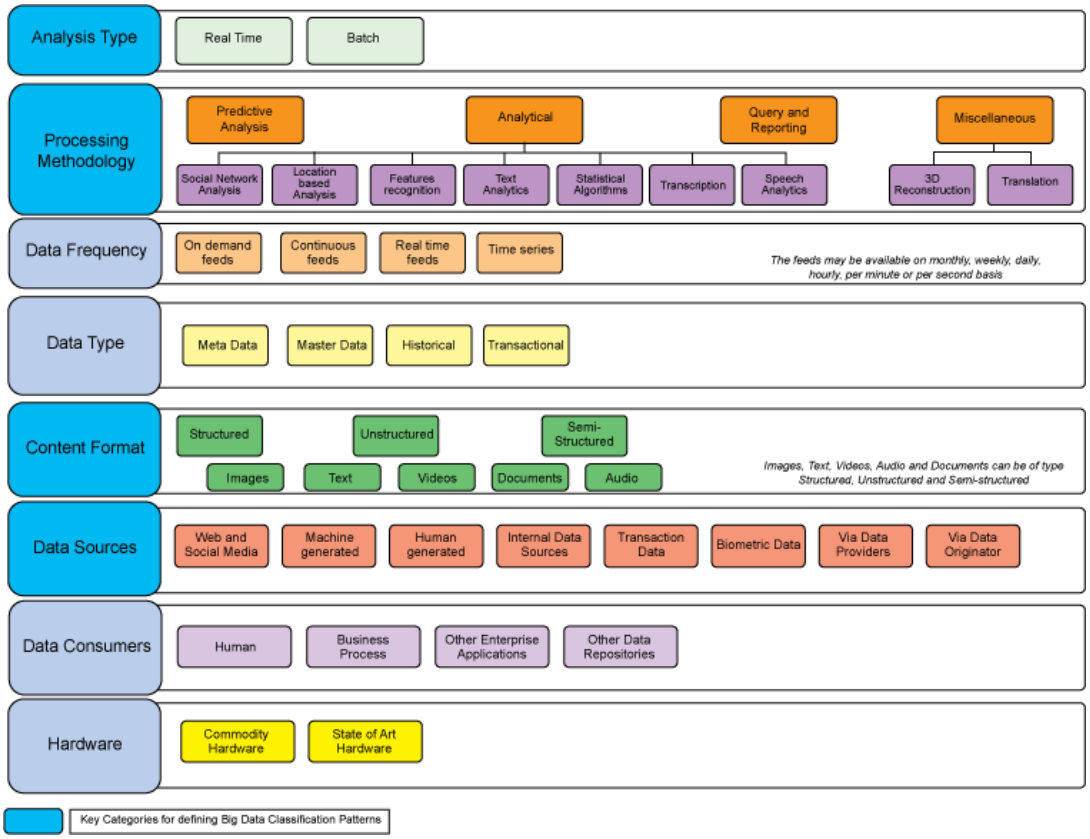
2.2 Catégorisation IBM

IBM, qui a déjà démontré son intérêt pour le Big Data au travers de sa définition des 4 V et la création de SQL dans les années 1970, offre toutefois une caractérisation du Big Data plus complète. Là où les autres acteurs offrent comme seuls critères la structure des données et leur origine, IBM suggère quant à lui 8 critères différents permettant de catégoriser le Big Data avec plus de précision.

- **Type d'analyse** : l'analyse est-elle effectuée en temps réel ou en différé ? Ceci aura un impact sur la vélocité des données et donc le matériel ou logiciel à choisir. L'exemple de la fraude à la carte de crédit évoqué précédemment (cf : Les 4 V) requerra une analyse en temps réel, alors que des données de statistiques pourront être traitées à posteriori.
- **Méthodologie de traitement** : quel type d'analyse est effectué : analyse prédictive, analytique, requêtes ad hoc ou rapports
- **Fréquence des données et taille** : quel est le débit de données ? La fréquence et la taille dépendent de l'origine des données :
    - A la demande. Ex : données d'une chaîne d'information qu'on obtiendra lorsqu'on en exprime la demande.
    - Flux continu. Ex : données financières, balise GPS.
    - Basé sur le temps. Ex : un CRON qui irait remplir une base de données toutes les demi-heures.
- **Type de données** : quel est le type de données à traiter – transaction, historique, métadonnées, données de base ? Connaître le type de données aide à les séparer lors du stockage.
- **Format du contenu** : les données sont-elles structurées, non-structurées ou semi-structurées ?
- **Source des données** : d'où les données sont-elles issues ? IBM différencie 8 différentes sources de données :
    - Web. Ex : réseau social.
    - Machine. Ex : capteur de pollution.
    - Humaine. Ex : formulaire d'inscription.
    - Internes. Ex : gestion des stocks.
    - Transactions. Ex : clic de souris sur un site.
    - Biométriques. Ex : banques d'empreintes digitales.
    - Fournisseurs de données. Ex : achat d'une liste d'adresses email.
    - Créateur des données. Ex : je fournis moi-même ma taille, poids et couleur de cheveux.

- **Client des données** : qui est le client des données traitées ?
    - Processus métier. Ex : le logiciel de la gestion des stocks doit fournir le nombre d'articles qui sont présents dans les entrepôts.
    - Utilisateur métier. Ex : les RH doivent connaître l'âge moyen des employés
    - Application d'entreprise. Ex : le logiciel de gestion des stocks a besoin des chiffres de statistiques de vente
    - Personne individuelle dans différents rôles métier. Ex : l'auditeur de l'entreprise a besoin des chiffres comptables.
    - Partie du flux de processus métier. Ex : l'algorithme de Google se chargeant d'afficher l'ordre des sites en fonction de différents critères (ex : visites).
    - Autres dépôts de données ou applications d'entreprise
- **Matériel** : le type de matériel employé pour mettre en œuvre le Big Data. Le matériel peut-être une limite à la sauvegarde ou l'analyse. Ex : on souhaite pouvoir analyser 1 million d'enregistrements par minute, mais le matériel à disposition ne permet que d'en traiter la moitié. Un choix s'impose.

Figure 2 : Classification du Big Data selon IBM



Source : DIVAKAR Mysore, SHRIKANT Khupat, SHWETA Jain, 17 septembre .2013

3. Quelles sont les familles d'interfaces existantes pour visualiser les Big Data

L'utilité du Big Data réside dans le fait qu'on peut comprendre une grande quantité de données d'un simple coup d'œil. Son analyse ne peut se faire que si le public cible arrive à comprendre les informations véhiculées et les idées qui en découlent. Selon l'Advanced Performance Institute, organisation spécialisée dans le consulting en matière de Big Data, le format le plus courant pour afficher des résultats comporte des tables et des feuilles de calculs complétées par des diagrammes. Alors que les présentations de données sous forme purement numérique, sans aucune forme de graphique sont beaucoup moins usuelles, car trop lourdes et donc difficiles à analyser.

3.1 Affichage classique

Traditionnellement, les rapports analysant les Big Data utilisent différents types de graphiques pour visualiser des résultats, dont voici les plus courants identifiés par le pôle de discussion au sujet du Big Data mis en place par IBM :

- **Diagramme circulaire** : chaque segment du cercle représente un pourcentage du total des données.
- **Graphe à barres** : cette visualisation permet de facilement faire une comparaison entre deux valeurs adjacentes ou plus.
- **Graphique en courbe** : utile pour afficher l'évolution de données dans le temps.

La visualisation de données au travers de graphiques ne leurs octroient pas uniquement plus de sens, mais illustre le lien entre les données. Aujourd'hui, de nombreuses données restent également présentées dans des feuilles de calcul.

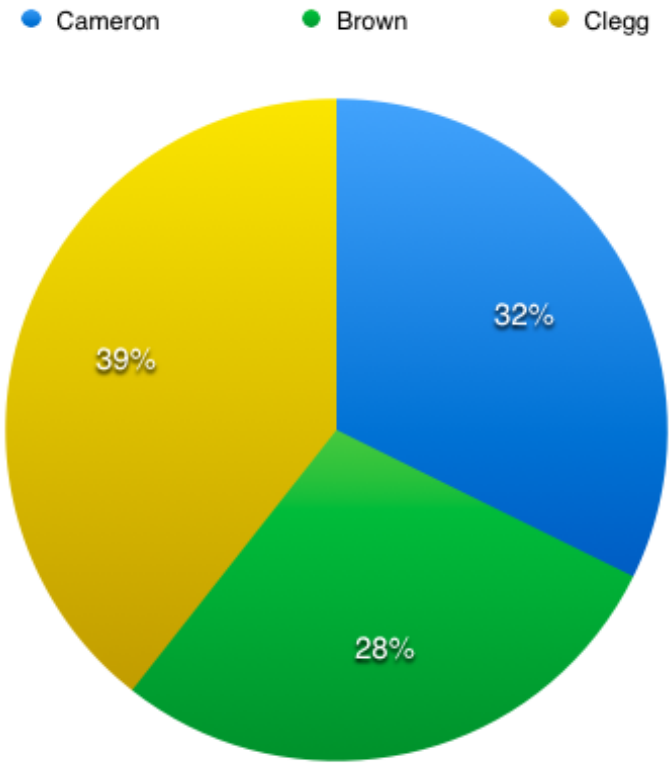
Figure 3 : Feuille de calcul représentant des statistiques de ventes

Date	Client	Produit	Type	Quantité	Prix unitaire	Total
01/01/2009	Brown	Chaussure	Bien	1	CHF 120.00	CHF 120.00
12/01/2009	Cameron	Pantalon	Bien	2	CHF 80.00	CHF 160.00
14/01/2009	Clegg	Réparation	Service	9	CHF 45.00	CHF 405.00
26/01/2009	Clegg	Veste	Bien	1	CHF 280.00	CHF 280.00
04/02/2009	Brown	Ceinture	Bien	2	CHF 60.00	CHF 120.00
10/02/2009	Brown	Livraison	Service	7	CHF 35.00	CHF 245.00
14/02/2009	Cameron	Cravate	Bien	1	CHF 45.00	CHF 45.00
18/02/2009	Clegg	Chemise	Bien	2	CHF 79.00	CHF 158.00
27/02/2009	Cameron	Pantalon	Bien	1	CHF 220.00	CHF 220.00
03/03/2009	Cameron	Ajustement	Service	5	CHF 55.00	CHF 275.00
10/03/2009	Brown	Ceinture	Bien	2	CHF 60.00	CHF 120.00
14/03/2009	Clegg	Boutons de manchette	Bien	4	CHF 15.00	CHF 60.00
26/03/2009	Cameron	Chausettes	Bien	2	CHF 20.00	CHF 40.00
27/03/2009	Brown	Cravate	Bien	1	CHF 45.00	CHF 45.00

Source : JEANNERET Philippe, 17 mai 2014

Actuellement, à cause de la quantité énorme de données, une feuille de calcul ne permet plus d'analyser ou ne serait-ce que de visualiser les informations correctement. Un responsable devrait passer des heures à lire des colonnes et tenter de croiser des données, sans forcément arriver à trouver une réponse à la question originale.

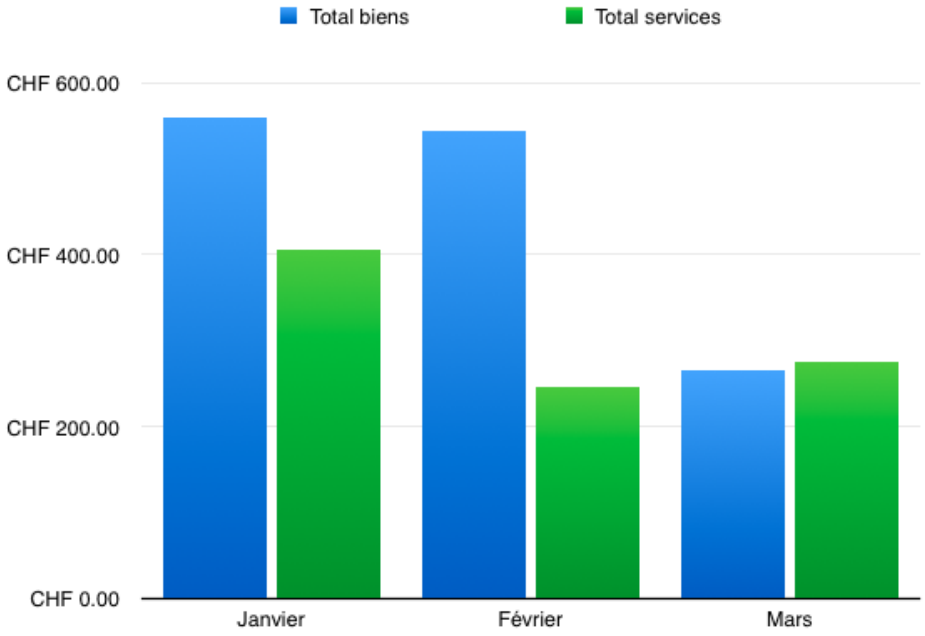
Figure 4 : Diagramme circulaire sur les proportions d'achats de chaque client



Source : JEANNERET Philippe, 17 mai 2015

Les données affichées sous forme de feuille de calcul sont facilement lisibles, mais il est plus ardu d'en tirer des conclusions. Au contraire, un diagramme circulaire permet de se rendre compte rapidement que les trois clients achètent des produits dans des proportions équivalentes. Celui-ci a néanmoins ses limites lorsqu'il arrive au point où trop de segments le rendent illisible. Les données sont bien plus rapidement compréhensibles si elles sont affichées de manière linéaire comme avec un graphe à barres ou en colonnes. De plus, ce dernier a l'avantage de pouvoir afficher des valeurs négatives si nécessaire. L'œil humain peut également prendre du temps à estimer les surfaces d'un diagramme circulaire pour comprendre quelle est la valeur dominante. Il est plus rapide de se rendre compte d'un classement grâce à un graphe à barres horizontales si celles-ci sont déjà triées de façon décroissante.

Figure 5 : Graphe à barres des ventes



Source : JEANNERET Philippe, 17 mai 2015

Placées côte à côte, il devient facile de savoir quel type de bien génère le plus de revenus. Avec ces simples exemples, nous voyons que l'interprétation des données prend un nouveau sens et se traduit par un gain de temps pour l'utilisateur.

3.2 Affichage moderne

Toutefois, l'avènement du Big Data et les types de données à traiter ne nous permettent plus de nous contenter de ces quatre graphiques cités précédemment. Les données semi-structurées et non-structurées ont besoin de techniques d'affichage plus moderne. Bernard Marr de l'Advanced Performance Institute considère qu'il faut regrouper les nouvelles méthodes de visualisation en trois groupes.

- Carte
- Texte
- Données

3.2.1 Carte

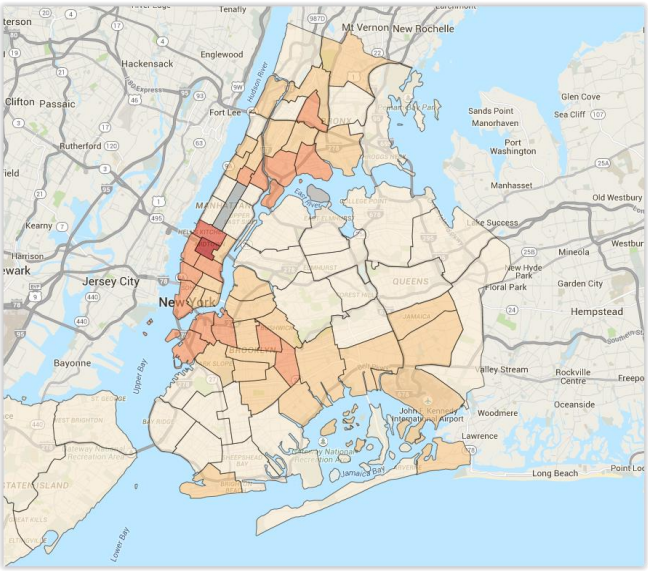
Une carte fournit déjà une forte représentation visuelle. Le gain en termes d'interprétation est énorme, puisqu'on passe d'une coordonnée géographique à plusieurs décimales à quelque chose de déchiffrable en un coup d'œil.

Figure 6 : Carte avec de simples points géographiques

Source : JEANNERET Philippe, 13 juin 2014

L'affichage de cartes se décompose en deux catégories. L'affichage de coordonnées géographiques pures et les données plus complexes liées à des coordonnées. Cette seconde catégorie superpose généralement à la carte les données en question sous forme de couches.

Figure 7 : NYC Crime Map



Source : NYPD, avril 2015

3.2.2 Texte

La visualisation de mots la plus courante passe au travers de nuages de mots. Un amas de mots est créé où chaque mot possède sa propre taille basée sur des données. Par exemple, plus un terme est fréquent, plus il apparaîtra en grand dans le nuage. Cette méthode empêche l'utilisateur de se perdre dans de nombreux sous-ensembles pour savoir ce qu'un individu en particulier a dit. Cette méthode peut être particulièrement utile pour illustrer un sentiment général dans une enquête auprès de clients ou d'employés. La taille nous permet de savoir ce que la plupart des personnes pensent de notre produit ou compagnie et offre un moyen rapide de connaître le sentiment général sans devoir aller lire chaque réponse individuellement. Le nuage de mots est utilisé pour représenter la partie non-structurée des données.

Cette technique possède aussi ses détracteurs, qui estiment qu'un nuage de mots ne fait pas assez. En effet, à part informer sur la proportionnalité d'un terme, nous n'obtenons guère plus de précisions. Jacob Harris, architecte logiciel au New York Times, prend l'exemple de rapports de combats représentés sous forme de nuages. Les termes « voiture » et « explosion » étaient les plus grands.

Cela indique-t-il un grand nombre de rapports sur des explosions de voitures ou d'explosions et de voitures ? Cette technique lui semble appropriée lorsqu'il s'agit d'analyser l'usage d'un terme, mais elle perd son sens quand il faut analyser un sujet complexe comme le serait la guerre en Iraq, car il est donc difficile de mettre en lien des termes entre eux. De plus, la position de chaque mot par rapport aux autres ne signifie rien et il s'agit là d'une autre faiblesse du nuage de mots. Enfin, il y a une liste de termes à bannir d'avance, tels que les déterminants ou éléments de liaisons qui, sinon, domineraient le nuage.

Figure 8 : Nuage de mot du discours prononcé par Obama lors de sa victoire aux élections de 2008

Source : DMARTYR, 5 novembre 2008

3.2.3 Données

Un des défis du Big Data est d'afficher les résultats de l'exploration dans un style qui fasse sens sans être laborieux à comprendre. Néanmoins, il existe peu de références expliquant comment choisir le bon graphique en fonction des données à y représenter. De plus, le peu qui existe est parfois contradictoire et ne concerne pas directement le Big Data. Une des idées consiste à chercher une méthodologie dans un domaine similaire au Big Data ayant également des graphiques comme le Data Mining ou le Marketing.

Les graphiques listés ci-dessous reprennent donc des concepts propres au Marketing, au Data Mining, mais également de nouveaux types de graphiques qui deviennent courants, spécifiques au Big Data et qui répondent à des besoins actuels.

Pour cela, il convient d'établir le type de lien qui existe entre les données :

- Comparaison
- Distribution
- Relation
- Composition
- Connexion

3.2.3.1 Comparaison

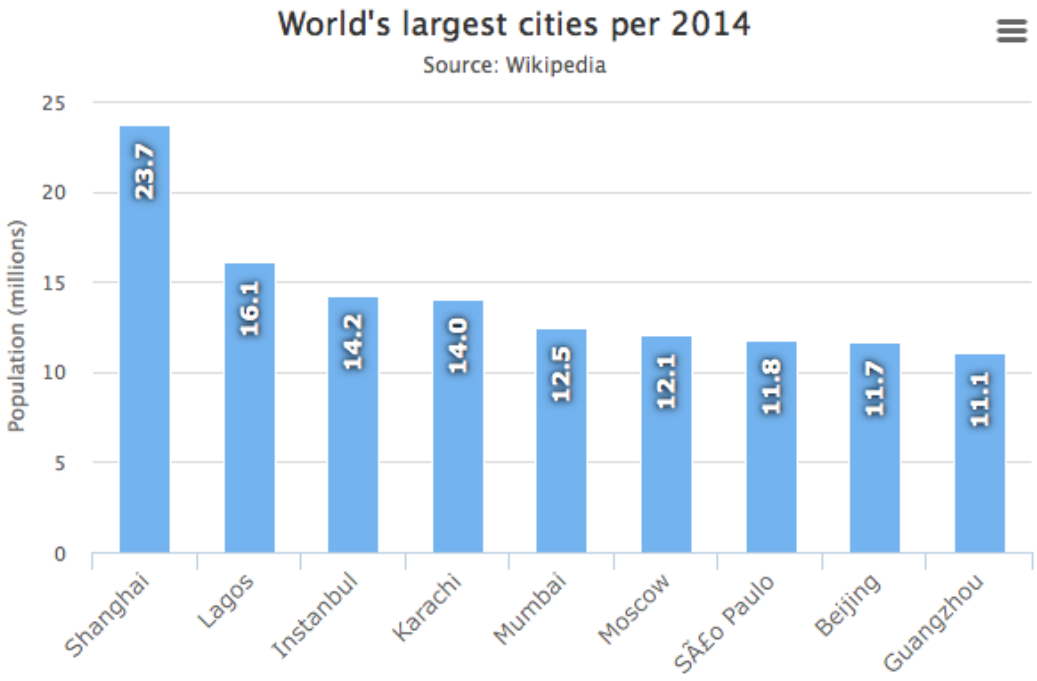
Le principal critère pour choisir un graphique qui doit effectuer une comparaison dépend si celle-ci est à effectuer entre les instances ou en fonction du temps. Il s'agit ici de représenter des données structurées.

3.2.3.1.1 Entre les instances

<u>Graphique en colonne</u>

Un graphique en colonnes permet de comparer des valeurs discrètes ou continues. Il faut le transformer en graphique à barres si le but est d'établir un classement entre les données.

Figure 9 : Graphique en colonne



Source HIGHCHARTS

Il est possible d'avoir deux variables par instance si on souhaite comparer plusieurs variables en même temps pour autant que ces variables partagent le même axe des X et Y.

Rappel : Dans le cadre d'une comparaison de villes en fonction de leur population, Shanghai serait une instance dans le graphique précédant, alors que la valeur de sa population serait une variable.

Ex : On pourrait avoir deux colonnes par villes. L'une représenterait la population en millions comme actuellement et l'autre la somme d'étrangers.

Graphique à barres horizontales

Un graphique à barres horizontales possède les mêmes critères qu'un graphique en colonnes, à la différence que l'alignement horizontale permet de comparer un plus grand nombre d'instances. Il n'est toutefois pas possible d'y afficher des valeurs négatives.

Ex : on utiliserait un graphique à barre si nous devions visualiser le nombre d'habitants de cinquante villes.

3.2.3.1.2 En fonction du temps

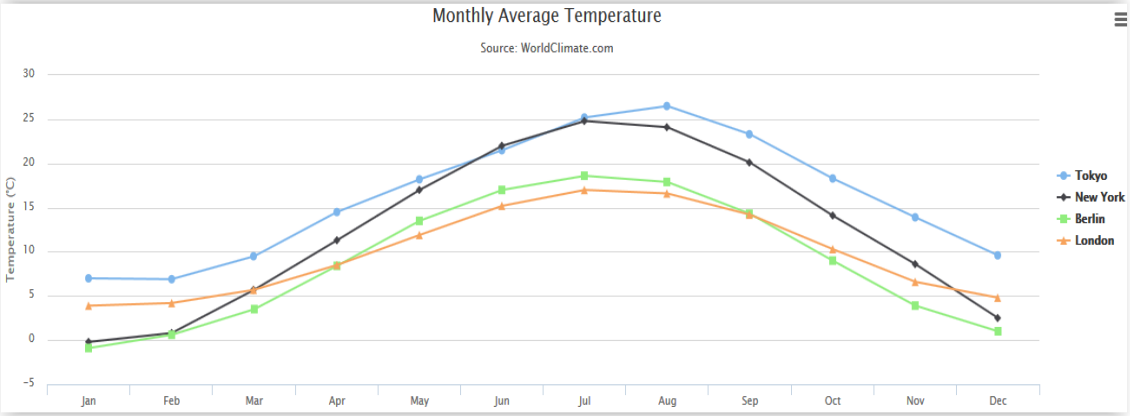
Graphique en colonnes

Il est possible de réutiliser un graphique en colonnes pour effectuer une comparaison en fonction du temps. Il ne faut pas pour autant qu'il y ait un trop grand nombre de variables. De plus, la chronologie sur laquelle les données sont réparties ne doit pas être trop longue afin de garder un graphique compréhensible.

Graphique à lignes

Ce graphique est utilisé principalement pour représenter des données continues en fonction d'un large intervalle temporel sur l'axe des X. Les intervalles doivent être de tailles égales et si on souhaite y superposer un graphique en colonnes, alors les deux diagrammes doivent partager un axe des X et Y commun. Les lignes doivent connecter deux valeurs adjacentes, mais s'il en manque une, alors il convient de l'indiquer par une rupture dans la ligne. Il est également possible d'y comparer plusieurs instances ou variables, pour autant que les axes des X et Y soient identiques.

Figure 10 : Graphique à lignes



Source : HIGHCHARTS

3.2.3.2 Distribution

Une distribution est un jeu de données possédant une seule instance et une ou plusieurs variables. Les graphiques affichant une distribution pour une variable possèdent

généralement une chronologie sur l'axe des Y. La différence entre des graphiques représentant une distribution s'effectuera sur le nombre de variables à représenter. Une distribution affiche des données structurées.

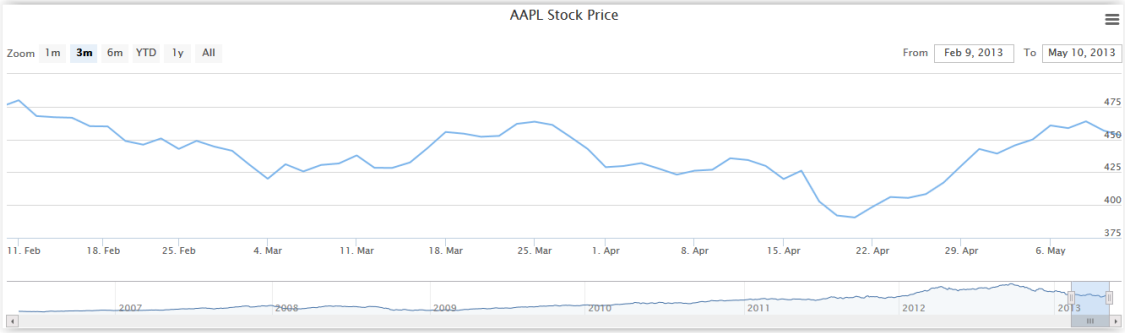
Ex : millimètres de pluies par mois à Londres.

3.2.3.2.1 Une variable

<u>Histogramme à ligne</u>

L'ensemble des règles s'appliquant à un graphique à ligne s'applique également à un histogramme à ligne, à la différence que ce dernier n'affiche qu'une seule variable. Si seuls quelques données sont disponibles, alors il est possible de remplacer l'histogramme à ligne par un histogramme à colonnes. Celui-ci reprend les concepts d'un graphique à colonnes, mais ne peut afficher par contre qu'une seule variable.

Figure 11 : Histogramme



Source : HIGHCHARTS

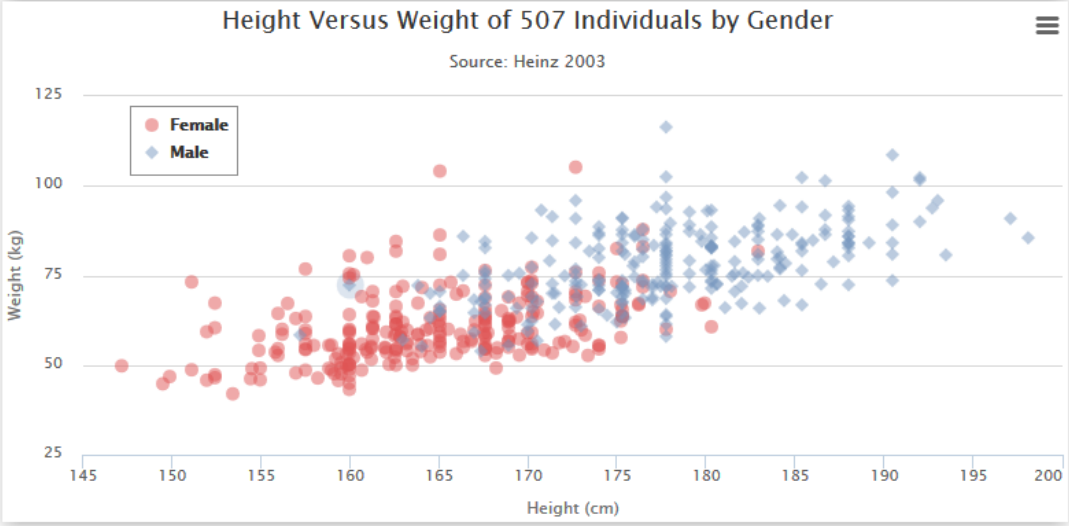
3.2.3.2.2 Deux variables

Nuage de points

Un nuage de point sert à établir une relation entre deux dimensions quantitatives. En jouant sur les formes ou les couleurs, il est possible de représenter une distribution avec deux variables. Il est possible d'appliquer un diamètre élevé aux points afin de se rendre compte plus rapidement d'une tendance. Ceci aura toutefois un impact négatif sur la précision du graphique.

Un nuage de points est utile pour examiner le lien entre les variables sur l'axe des X et celles se trouvant sur l'axe de Y. On peut donc remplacer la chronologie par un autre critère.

Figure 12 : Nuage de points



Source : HIGHCHARTS

3.2.3.3 Relation

Il existe deux graphiques pour visualiser la relation entre des variables. Une relation représente des données structurées sur un même graphique.

3.2.3.3.1 Deux variables

Nuage de points

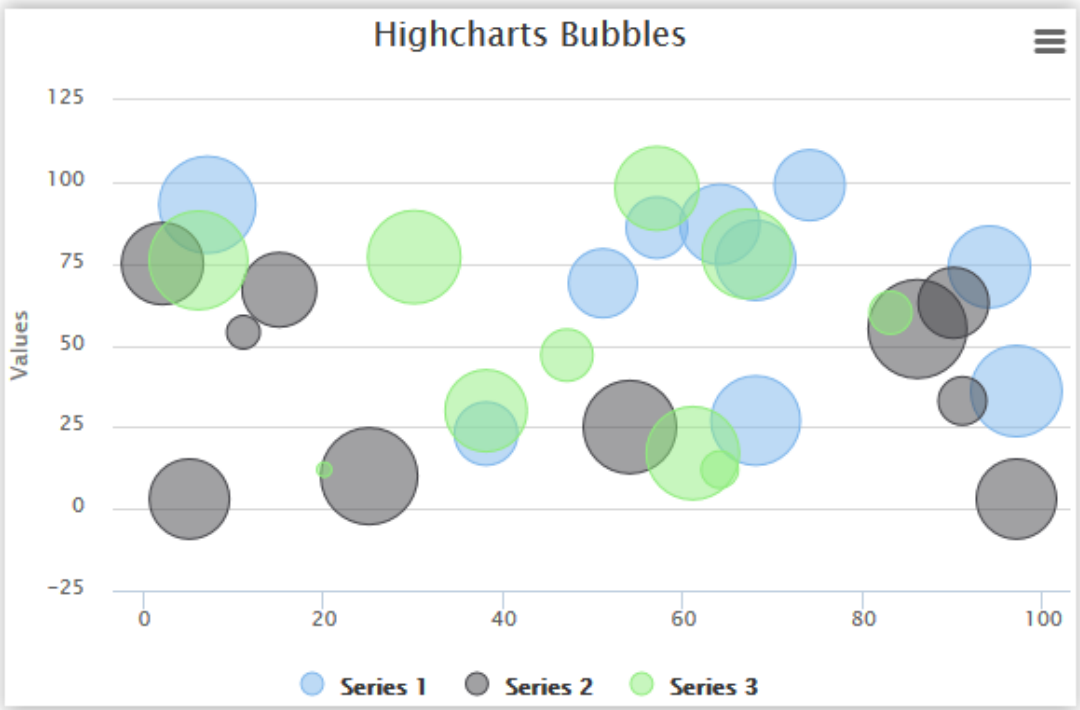
Comme vu précédemment, ce graphique peut être utilisé dans le cadre de l'analyse d'une relation entre deux variables.

3.2.3.3.2 Trois variables

Graphique à bulles

Ce graphique permet de comparer la relation qui existe entre trois variables. Il reprend les propriétés d'un nuage de points, mais transforme les points en bulles de tailles variables. Il est donc possible de présenter ainsi une troisième information en plus de celles définies par l'axe des X et Y.

Figure 13 : Nuage de points



Source : HIGHCHARTS

3.2.3.4 Composition

Une composition représente un regroupement de données formant un tout. Les compositions sont réparties en deux sous-ensembles : les données évoluant au cours du temps et les données n'évoluant pas.

3.2.3.4.1 Statique dans le temps

Diagramme circulaire

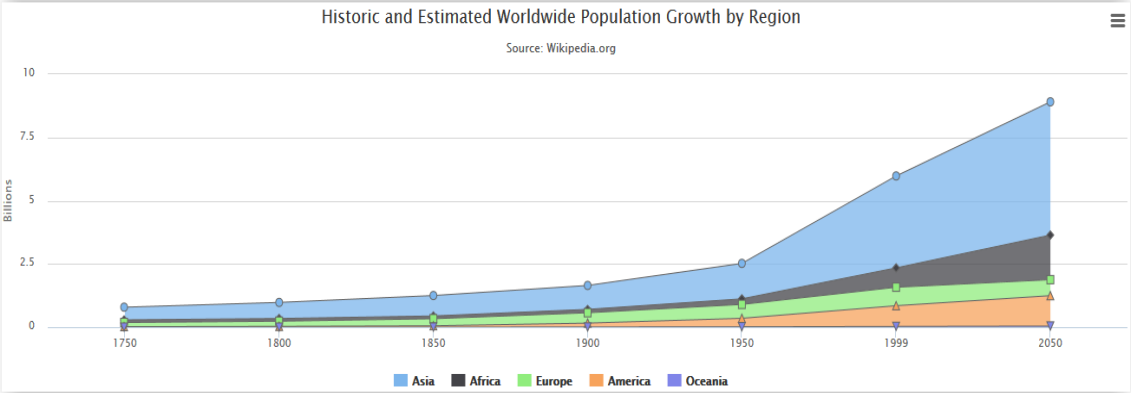
L'exemple le plus commun est le diagramme circulaire où la concaténation de chaque partie forme un cercle complet. Nous avons déjà vu précédemment que les diagrammes circulaires perdent en lisibilité lorsqu'ils sont utilisés pour représenter plusieurs instances. Il est toutefois admis d'en utiliser un à condition qu'il n'utilise pas plus de deux variables.

3.2.3.4.2 Evoluant dans le temps

<u>Diagramme de zones empilées</u>

Lorsque les données évoluent dans le temps, un diagramme à zone d'empilement est à privilégier. Il permet d'indiquer de quelle manière les proportions évoluent au fil du temps.

Figure 14 : Diagramme de zones empilées



Source : HIGHCHARTS

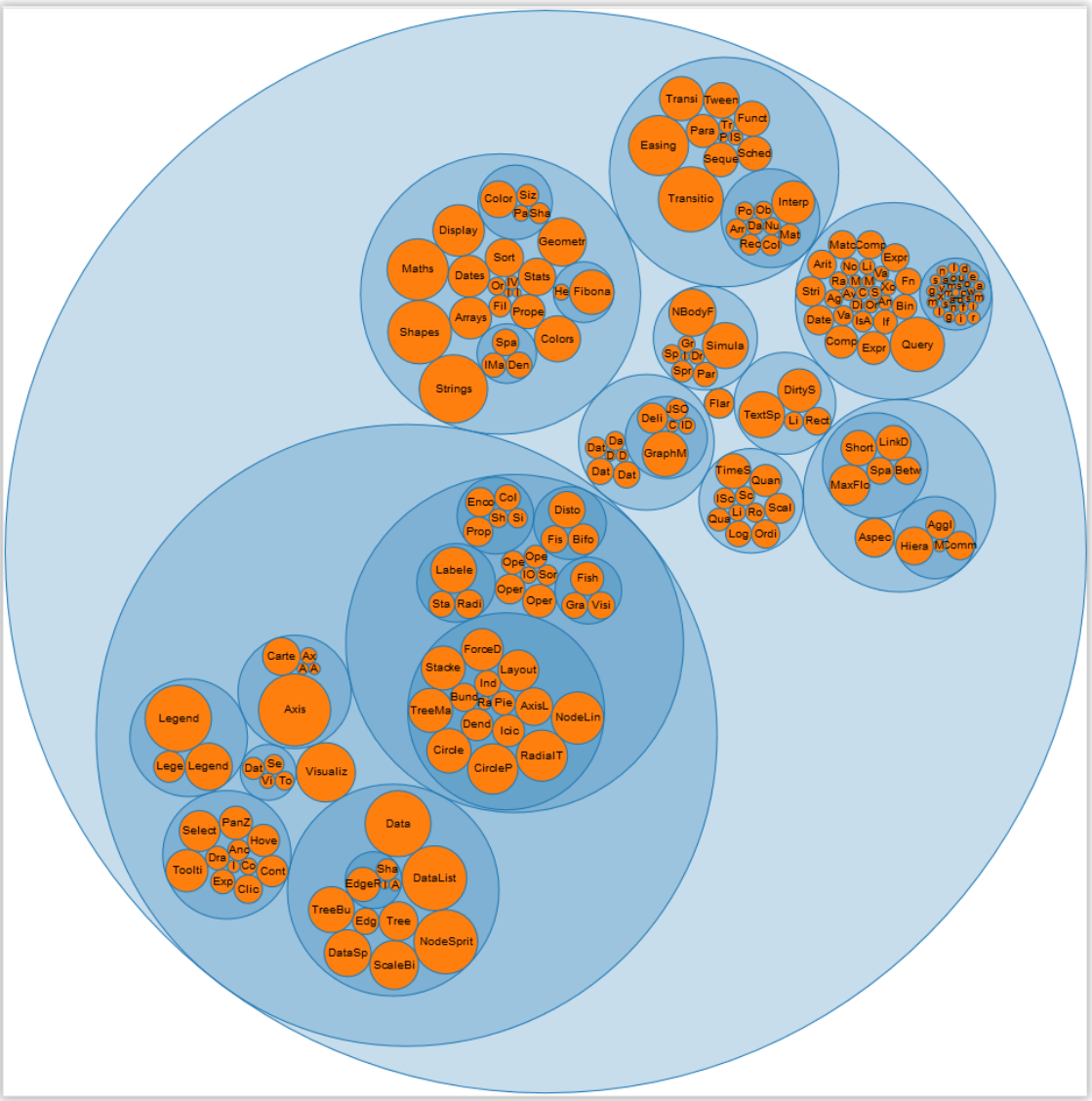
3.2.3.5 Connexion

Le Big Data a ouvert un nouveau type de lien entre les données : la connexion. En effet, il peut être crucial de visualiser ce qui relie des données qui sont de type non-structuré et ce qu'elles ont en commun.

<u>Regroupement en cercles</u>

Cette méthode consiste à regrouper des données, pour lesquelles un lien existe, au sein d'un même cercle. Ce cercle est lui-même un sous-ensemble d'un cercle plus grand et ainsi de suite. Cette méthode a l'avantage d'être très visuelle et donc rapidement compréhensible pour voir où un rapport existe.

Figure 15 : Regroupement en cercles

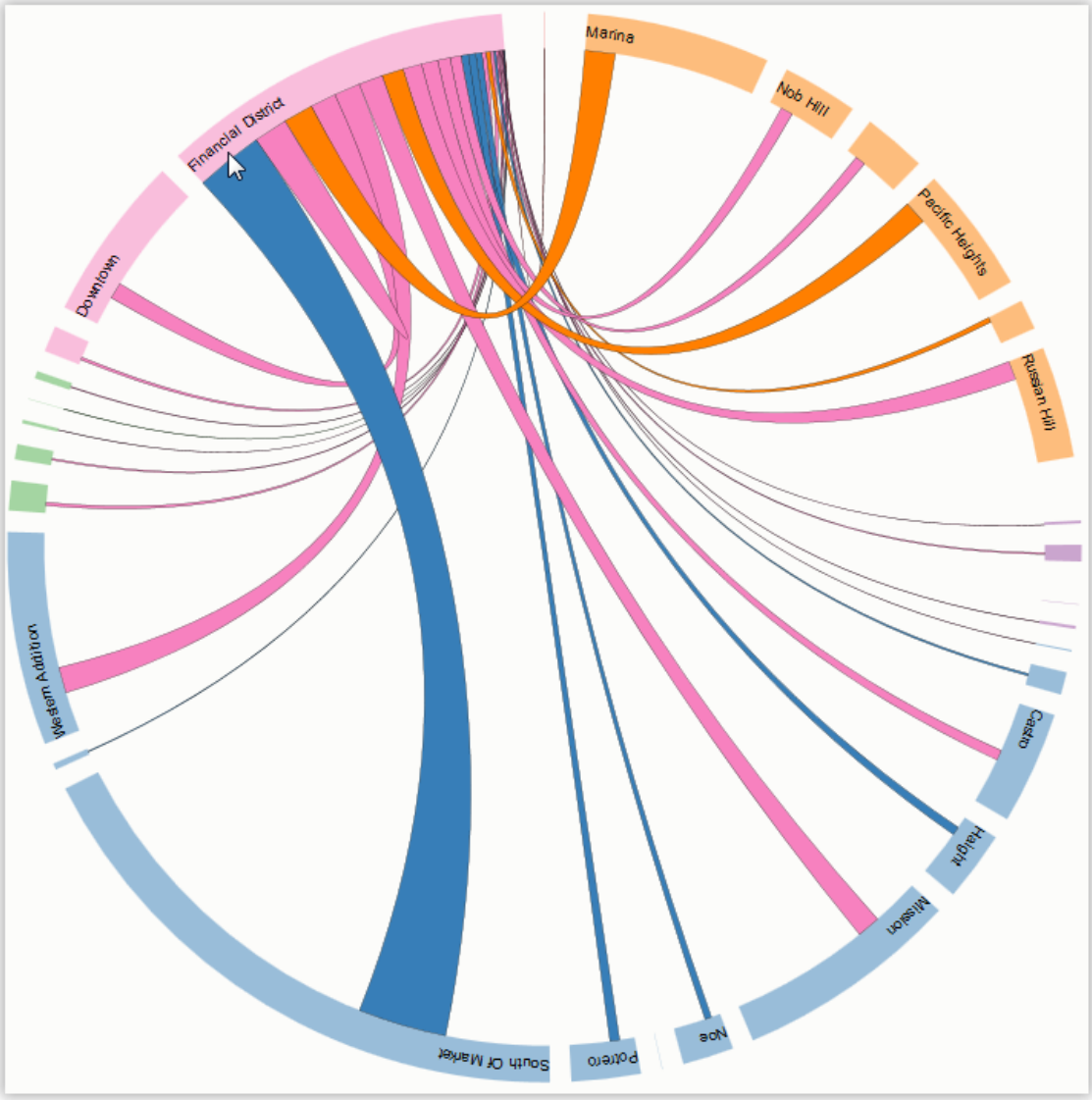


Source : BOSTOCK Mike, 3 novembre 2012

Diagramme d'accords

Un diagramme d'accords sert à afficher le lien qu'il y a entre des instances. Elles sont réparties sur la bordure extérieure d'un cercle. Des arcs de différentes épaisseurs relient les instances entre-elles. L'épaisseur indique l'intensité du lien les liant.

Figure 16 : Diagramme d'accords



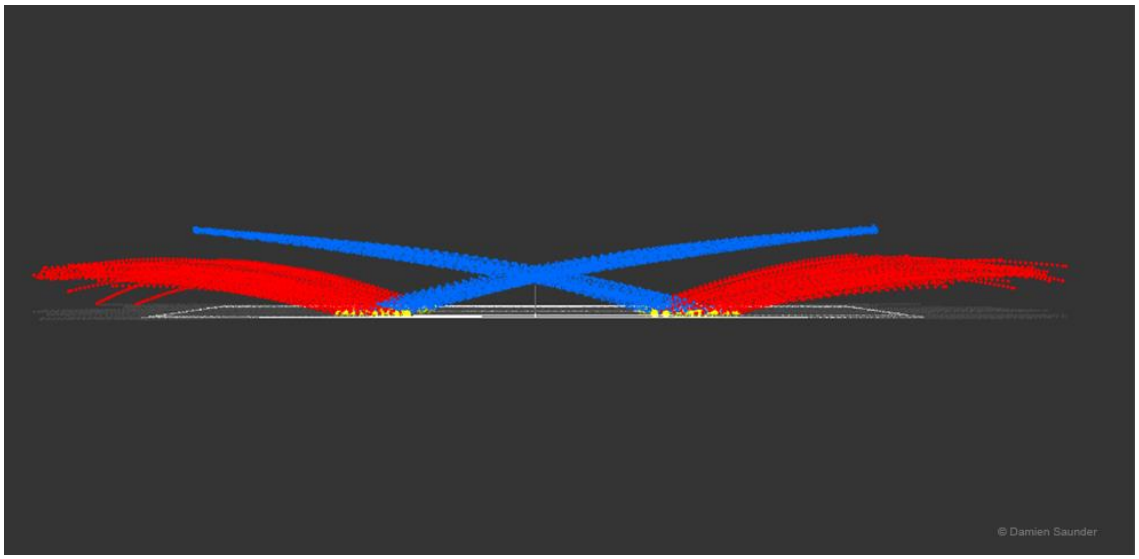
Source : BOSTOCK Mike, 9 juin 2012

Diagramme d'accords représentant la destination de véhicules entre différents quartiers. Cette image représente là où se rendent les véhicules partant du « Financial District » et dans quelles proportions.

3.2.3.6 Cartographique de fond

Ce type de graphique permet de se rendre compte d'une réalité qui est augmentée à l'aide d'un fond imagé. Il est pratique de l'utiliser lorsque la clarté est de rigueur, mais que la précision n'est pas l'effet recherché.

Figure 17 : Cartographie de fond

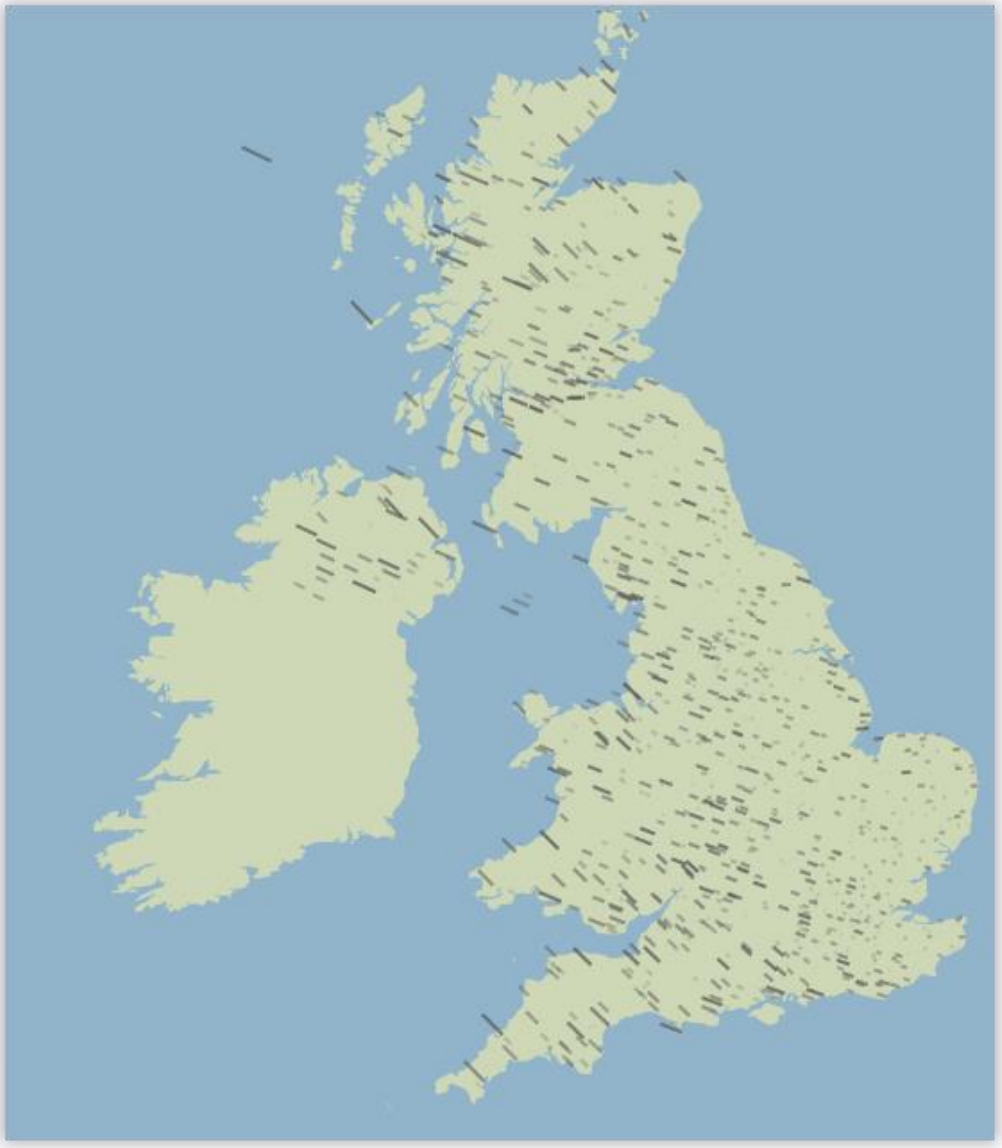


Source : DEMAJ Damien, 26 février 2015

3.2.3.7 Animation

Les animations ne sont pas un type de graphique à part entière, mais une propriété de ce dernier. Elles servent à visualiser une tendance dans le temps. En théorie, ceci concerne tous les graphiques cités précédemment. L'animation a comme but d'apporter de nouvelles données, comme le suivi en direct d'un avion sur une carte. Il ne faut pas confondre animation et interaction. Cette seconde terminologie n'est là que pour permettre à l'utilisateur de jouer avec les données. Toutefois, une interaction peut déclencher une animation.

Figure 18 : Animation des vents

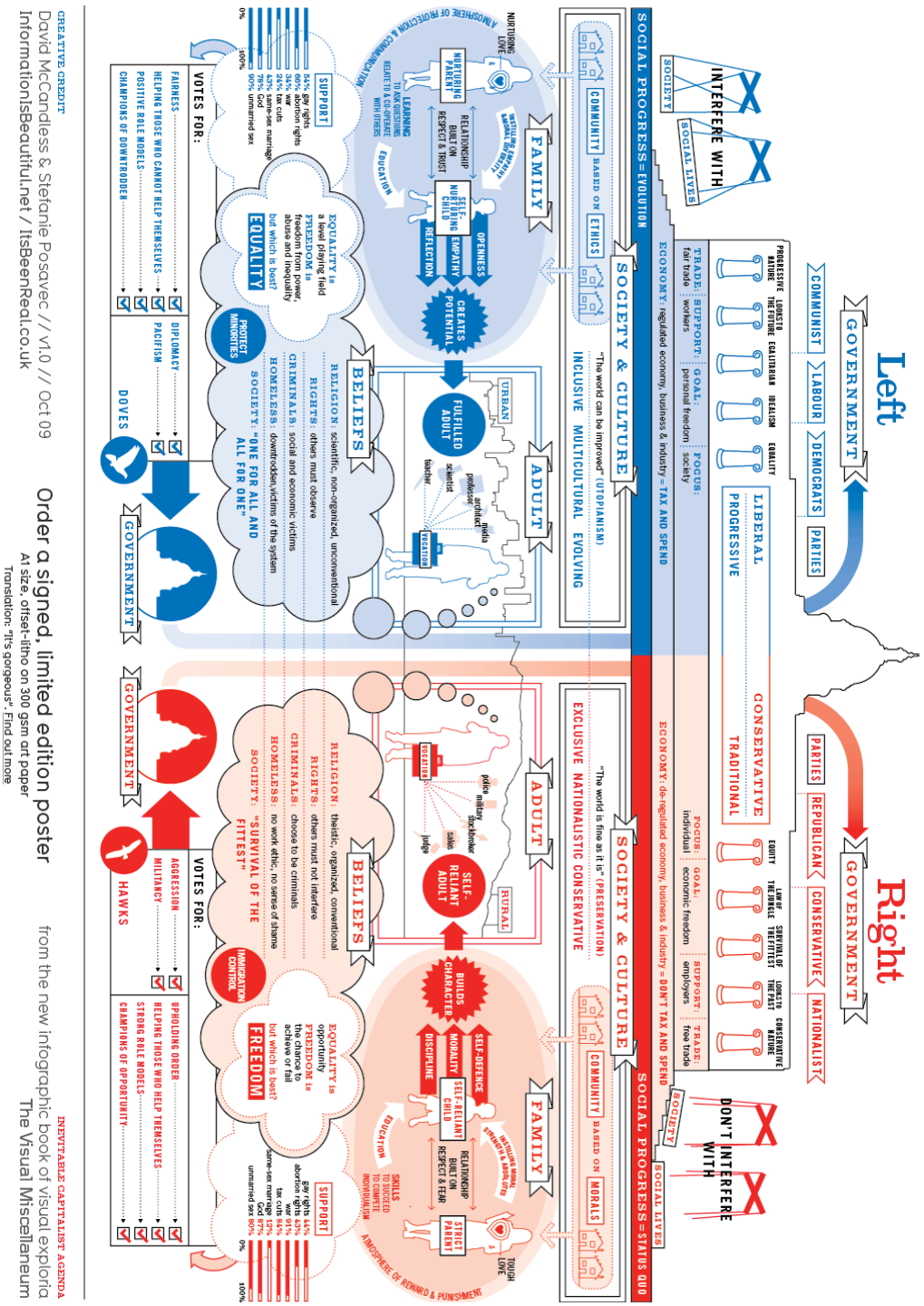


Source : COOK Peter

3.2.3.8 Infographie

Une infographie permet, à la manière d'un livre, de lire une succession de représentations afin de pouvoir en tirer des conclusions. L'utilisateur se créer ici sa propre interprétation sans l'aide d'une personne extérieur qui lui expliquerait ce qu'il voit.

Figure 19 : Infographie de la politique américaine

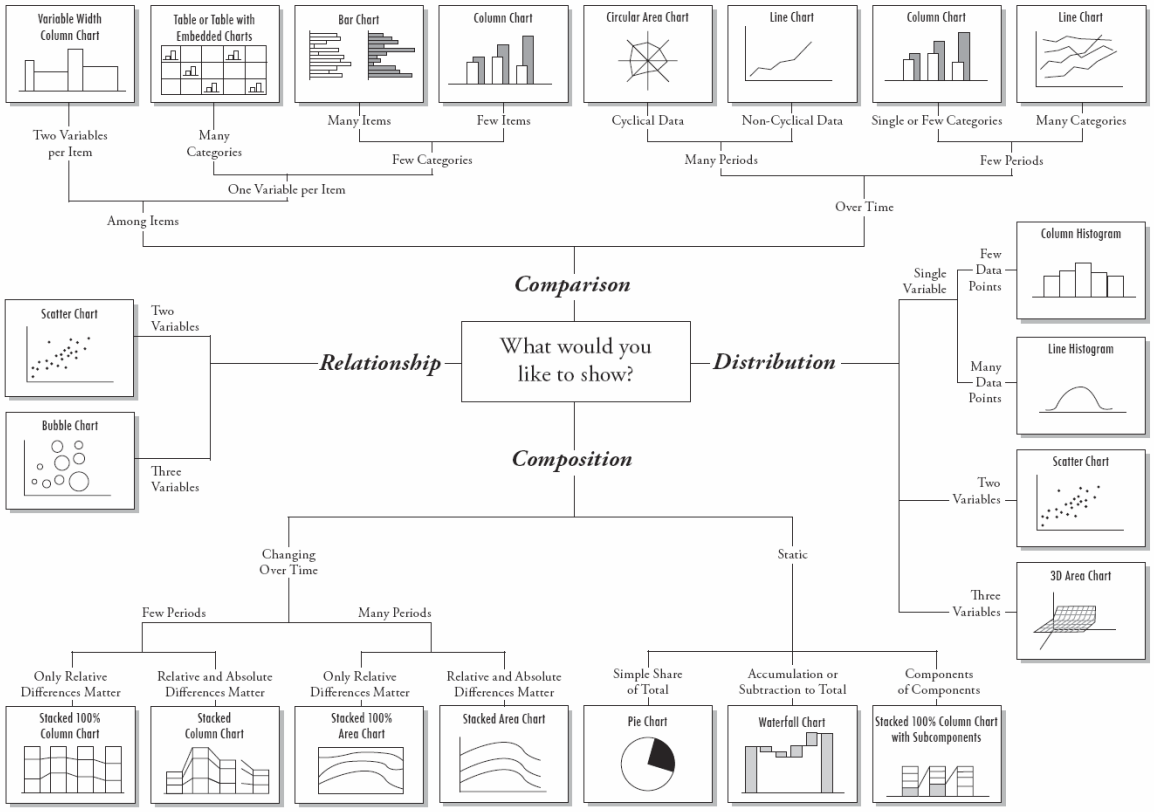


Source : MCCANDLESS David, décembre 2010

3.2.4 Choisir le bon graphique

Les pistes ci-dessus proviennent essentiellement d'un guide écrit par le Dr. Andrew Abela, professeur de marketing à Catholic University Of America, qui est souvent repris en exemple. Bien qu'étant prédestinés à être utilisés en marketing, les principes de ce guide sont souvent respectés lorsqu'il s'agit d'afficher du Big Data. Tous les graphiques ne sont pas forcément bons à prendre pour afficher clairement des données et c'est pour cela que tous n'ont pas été repris ci-dessus. De plus, il manque la partie démontrant la connexion qu'il existe entre des données.

Figure 20 : Quel graphique choisir



Source : ABELA Andrew, 22 avril 2013

3.2.5 Bonnes pratiques

Ces trois groupes cités précédemment (carte, texte, données) ont quelque chose en commun. Premièrement, la visualisation des données est à décomposer en deux sous-parties :

- **Exploration** : pouvoir trouver une histoire que les données nous racontent
- **Explication** : pouvoir raconter cette histoire à une audience

Ensuite suivent les attentes de l'audience auxquelles il faut pouvoir répondre :

- De quelles informations l'audience a-t-elle besoin ?
- De quel niveau de détails l'audience a-t-elle besoin ?
- Quelles hypothèses culturelles peuvent affecter les choix de design ?
- Identité, motivation, langue, contexte social
    - Que signifient les couleurs ?
    - Quelles icônes sont familières ?
- Les couleurs sont-elles adaptées aux daltoniens ?

D'un point de vue esthétique, il faut absolument éviter d'alourdir un graphique. Par exemple, l'espace entre les barres d'un graphique à barres doit être de 50% à 150% la largeur d'une barre. La couleur est également importante puisqu'elle doit être unie et non-intense. Les bordures sont à bannir, sauf pour mettre en évidence une colonne spécifique. Une grille de fond peut faciliter la comparaison de deux valeurs éloignées, à condition que celle-ci ne soit pas trop prononcée.

3.2.6 Erreurs à ne pas commettre

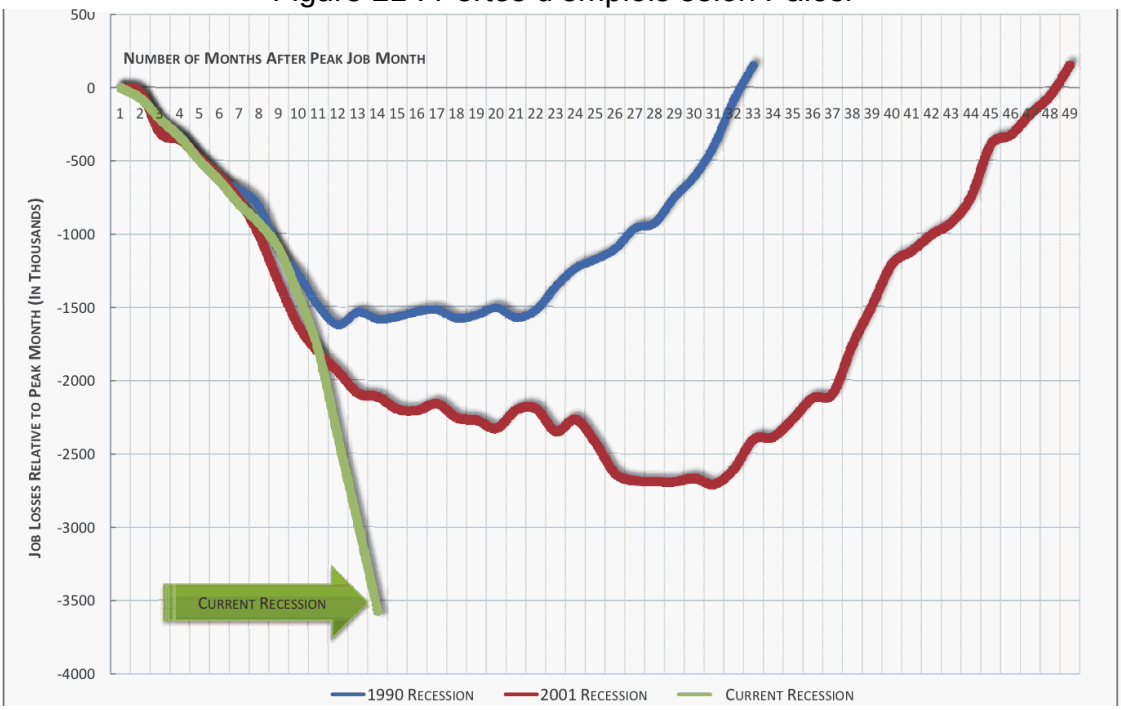
Il y a également des choses communes à tous les graphiques qu'il convient d'éviter. Jen Underwood, analyste de marchés chez Microsoft, a créé une liste des erreurs les plus communes lors de la visualisation de données :

- **Choix d'affichages inapproprié** - ex : les diagrammes circulaires ou les diagrammes en 3D, …
- **Varier les tableaux simplement pour les varier**
- **Trop d'informations**
- **Design pauvre** - ex : diagramme trop dispersé, style des lignes, couleurs saturées/trop claires
- **Insérer des données erronées**
- **Placement et ordre inconsistant** - ex : classer les ventes en fonction des années dans un graphique puis en fonction du chiffre d'affaire dans un autre
- **Echelles inconsistantes ou inversées** - ex : avoir un axe en pourcents à un endroit puis en valeurs brutes ailleurs
- **Echelles des axes proportionnelles** – ex : avoir des axes disproportionnés permet d'accentuer ou de freiner une courbe dans un graphique

3.2.6.1 Comment fausser un graphique

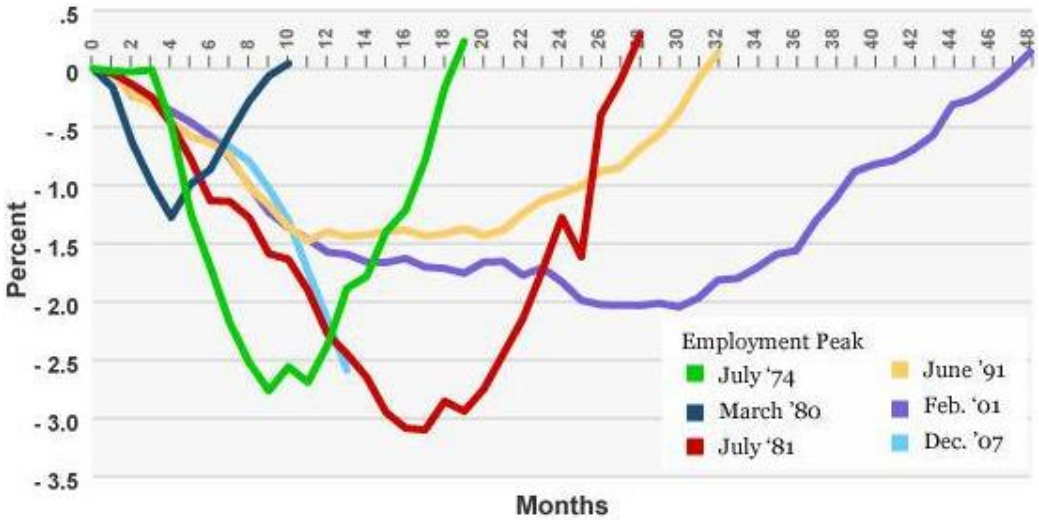
Par exemple, les deux graphiques présentent l'état du chômage suivant une récession où chaque courbe représente une récession passée. Le premier a été présenté par la politicienne Nancy Pelosi et le second par Fielding Cage au TIME Magazine. Tous deux sont basés sur les mêmes données.

Figure 21 : Pertes d'emplois selon Palosi



Source : SPIRA Dan, 8 juillet 2009/

Figure 22 : Pertes d'emplois selon Cage



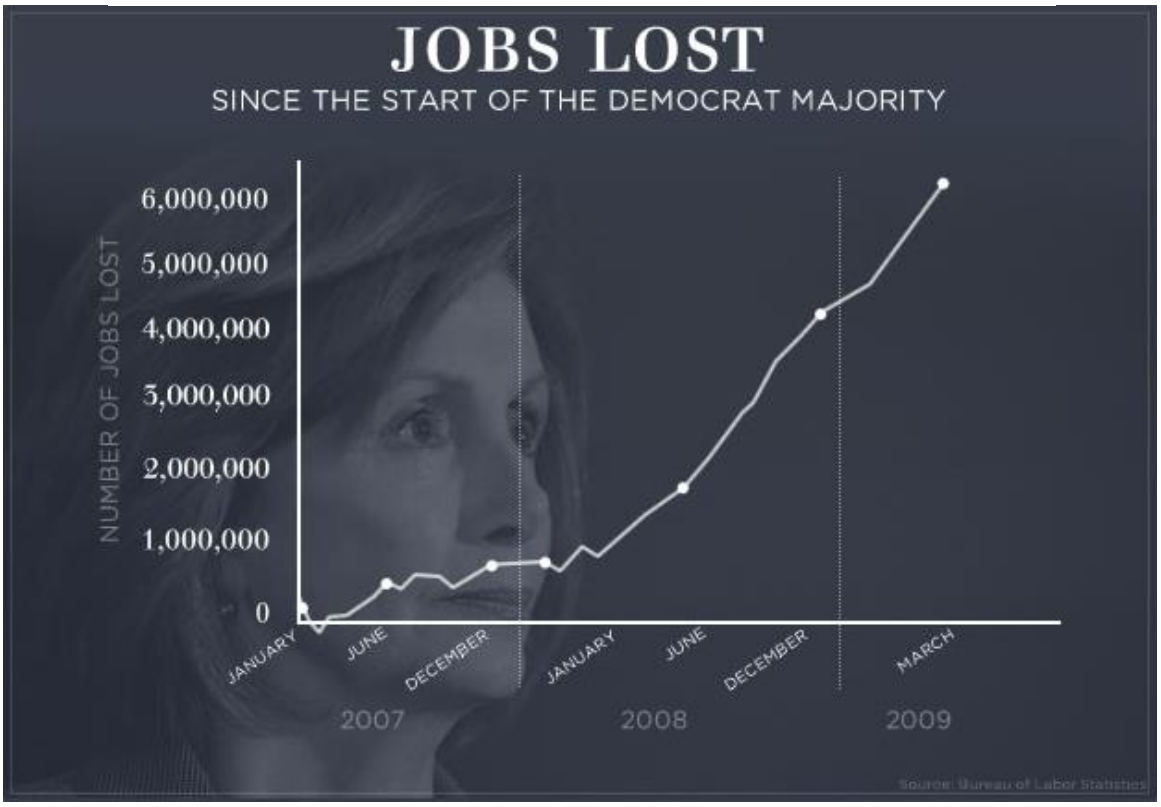
Source : SPIRA Dan, 8 juillet 2009

Relevons que le premier se veut plus alarmiste que le second. Ceci a été réalisé au-travers de trois manipulations :

- **Axe vertical** : les deux graphiques possèdent des axes X et Y qui ne sont pas proportionnels, mais Pelosi augmente cette proportionnalité, ce qui a pour effet d'exagérer la chute de la courbe.
- **Valeurs absolues** : Pelosi travaille avec la somme d'emplois perdus au-lieu du pourcentage d'emplois perdus. Le nombre d'employés ayant augmenté au fil des ans, ceci renforce encore l'effet de chute de la courbe.
- **Contexte réduit** : le graphique de la politicienne réduit notre capacité de comparaison mentale en ne gardant que les récessions les bénéfique à l'emploi. En effet, elle compare la récession actuelle aux deux les violentes (1990 et 2001). Par contre, elle se garde d'afficher le fait que ces deux récessions ont été celles qui ont duré le plus longtemps.

Un graphique, accompagné d'une description, est un outil puissant permettant de raconter une histoire à une audience. Nous venons de voir qu'il est possible de manipuler l'histoire, sans mentir sur les chiffres, afin qu'elle colle à notre.

Figure 23 : Graphique manipulé présent sur le site du parti Républicain



Source : SPIRA Dan, 8 juillet 2009

4. Quelles sont les technologies

Il existe des technologies pour visualiser le Big Data qui sont orientées métier, comme Microsoft SSRS. Ce genre de client lourd est très pratique pour afficher des données structurées, mais son manque de flexibilité lui fait défaut lorsqu'il s'agit de visualiser des données semi-structurées ou non-structurées.

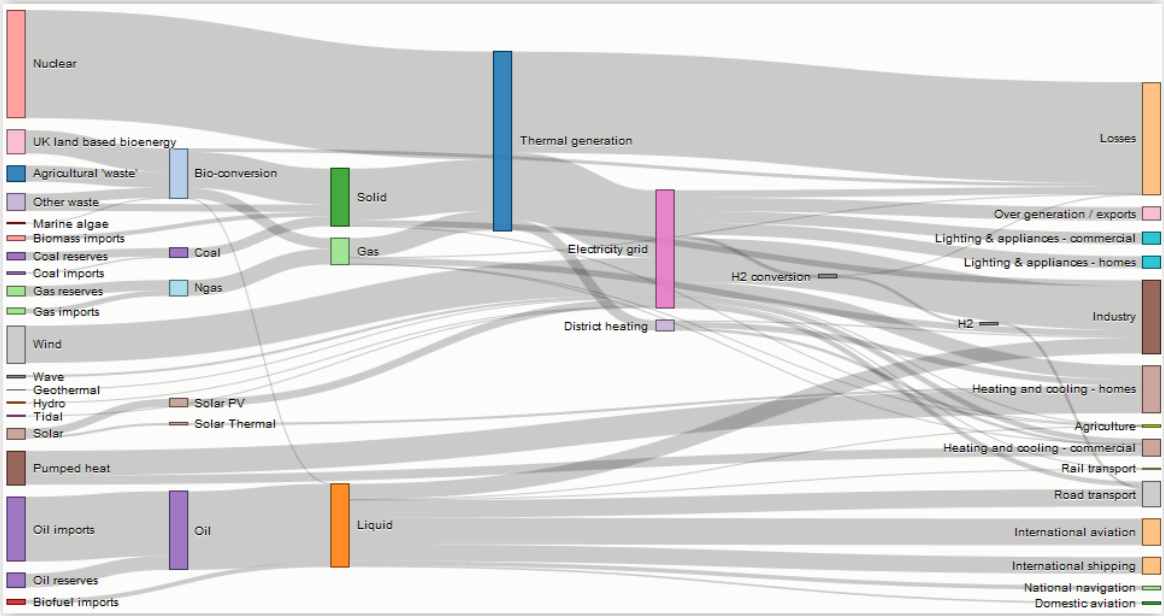
La grande partie des affichages sont traités grâce aux technologies web. Le JavaScript permet, au travers de librairies graphiques comme D3 ou Highcharts, une grande flexibilité en termes de création, en fonction de ce que nous souhaitons représenter. Cette souplesse est à la fois une force est une faiblesse.

L'avantage est qu'il est facilement possible de créer une nouvelle visualisation à partir de zéro ou en adaptant un graphique déjà existant. La grande quantité des personnes ayant des connaissances en technologies web, combiné au métier grandissant de graphiste, fait que les données sont aujourd'hui un véritable terreau de création. Ceci explique la diversité des graphiques existants.

L'inconvénient de cette diversité est qu'il est difficile d'établir un guide expliquant quel graphique choisir. Il ne serait presque pas exagéré de dire qu'il existe autant de graphiques différents que de données.

Prenons l'exemple du diagramme de Sankey. Ce diagramme est utilisé pour visualiser les flux d'énergies entre une source et une consommation. L'origine de l'énergie (nucléaire, pétrole, …) est affichée sur la gauche et son utilisation l'est sur la droite. Des flux partent de l'origine à l'utilisation afin de représenter quelle consommation consomme quelle énergie. Des nœuds intermédiaires servent à indiquer la façon dont l'énergie converge en un point avant d'être transmise pour être consommée. Le diagramme de Sankey ne sert qu'à représenter des flux d'énergies.

Figure 24 : Diagramme de Sankey



Source : BOSTOCK Mike, 22 mai 2012

Cet exemple démontre clairement à quel point les technologies web ont diversifié les graphiques et donc compliqué leur référencement.

5. Cas d'étude GEoTweet

5.1 Besoins

Twitter, avec ses 500 millions de tweets quotidiens, est un outil de choix pour observer les comportements sociaux. En effet, mis à part le contenu, la date et l'auteur du message, Twitter donne accès au total à 59 informations liées à ce message, comme les coordonnées géographiques du tweet ou un analyseur de texte qui retourne la langue du message.

A partir de ces possibilités, la Haute Ecole de Gestion de Genève a décidé de participer au « Mapping, l'événement HES » dont le thème est « frontières et urbanité ». Cette manifestation a pour but d'explorer les frontières existantes à Genève et prendra la forme d'une exposition temporaire pour le grand public. Suivant les directives indiquées par les organisateurs de la manifestation, l'idée de l'étude, intitulée « GEoTweet », est de traiter ce thème au travers des possibilités offertes par Twitter. Un outil mis en place par l'équipe GEoTweet récolte d'ores et déjà les messages émis sur le territoire du Grand Genève.

Il s'agit ensuite de choisir ce qui sera affiché à partir des tweets obtenus. Pour ce faire, deux directions ont été définies. Premièrement, pouvoir observer la répartition des langues à Genève et secondement explorer le rayonnement de Genève dans le monde.

Ce projet est mené par Mr. Patrick Ruche, son assistant Mr. Arnaud Gaudinat ainsi que trois étudiantes de la filière d'information documentaire : Mme Fany Béguelin, Mme Elisa Banfi et Mme Romaine Kaufmann.

5.2 GEoTweet et le Big Data

Cette section a comme but d'étudier le type de Big Data auquel correspond Twitter et le projet GEoTweet par rapport aux aspects théoriques étudiés au début de ce document.

5.2.1 4 V

Les données de Twitter entrent dans le cadre du Big Data tel qu'il est expliqué au début de ce document au travers des 4 V énoncés par IBM :

- **Volume** : rien que sur le territoire du Grand Genève, on dénombre à peu près 200'000 tweets quotidiens. Bien qu'il s'agisse d'un critère subjectif, ce volume peut être considéré comme suffisant pour parler de Big Data.
- **Vélocité** : l'outil récoltant les tweets fonctionne selon un modèle de streaming. Il s'agit d'un filet mis en place qui capture les tweets selon des critères. Un tweet manqué, qui se trouve être en-dehors de la fenêtre géographique analysée, est perdu et ne peut pas être récupéré via le streaming. On parle donc ici de temps réel.
- **Variété** : les données utilisées ici seront des textes et des valeurs numériques. Ce point sera détaillé plus loin.
- **Véracité** : les données de Twitter sont relativement pures du point de vue du contenu. Toutefois il existe certains messages pollués comme des bulletins météo diffusés automatiquement à intervalles réguliers. Ce type de message vient fausser le reste des données générées par des êtres humaines.

Ex : il y a cinquante tweets par jour ayant comme hashtag #météo à Versoix. Il serait légitime de croire que les habitants de Versoix sont concernés par la météo, mais il s'agit en réalité d'un tweet émis automatiquement par une station météo toutes les demi-heures.

De plus, les interfaces choisies travailleront essentiellement avec la géolocalisation contenue dans les propriétés d'un message. Or, pour avoir cet attribut, l'auteur du tweet doit avoir activé la géolocalisation dans les paramètres de son application Twitter. La proportion des tweets dont l'utilisateur a activé cette fonctionnalité représente 2%, soit 28'000 tweets hebdomadaires sur Genève. Malgré le fait que les données soient pures, la véracité s'en retrouve affaiblie car l'échantillon sur lequel nous travaillons représente au mieux 2% de la masse disponible.

5.2.2 Catégorisation IBM

La catégorisation d'IBM est venue spécifier celle des 4 V. On peut donc se rendre compte avec plus de précisions des données avec lesquelles GEoTweet va travailler.

- **Type d'analyse** : l'analyse sera effectuée en temps réel puisqu'un tweet est capturé en streaming, mais il pourra être ajouté à un graphique en différé en fonction de l'interface sur laquelle l'utilisateur se trouve. Cet aspect est spécifié dans la partie « Interfaces pertinentes » qui suit.
- **Méthodologie de traitement** : la méthodologie sera analytique, car on décompose un ensemble en éléments afin de créer un schéma général.

  Ex : on prend toutes les coordonnées géographiques pour les disposer sur une carte
- **Fréquence des données et taille** : la fréquence des données est un flux continu capturant les tweets au fur et à mesure qu'ils sont créés
- **Type de données** :
  - Métadonnées :
    - Auteur
    - Date
    - Coordonnées géographiques
    - Langue du tweet
    - Avatar
  - Données de bases :
    - Message
    - Hashtag
- **Format du contenu** : un tweet est l'exemple même de la « donnée semi-structurée ». En effet, une partie des informations contenues sont structurées : auteur, date, position géographique, longueur du message. Une autre partie est quant à elle non-structurée. : est-ce un message, une photo ou une vidéo ?
- **Source des données** : web
- **Client des données** : il est intéressant de relever que la catégorisation mise en place par IBM porte sur un aspect orienté business, d'où l'absence d'un type de client de données adapté à l'exposition « Mapping ». Il est toutefois imaginable d'indiquer travailler avec des utilisateurs métiers puisque les visiteurs utilisent potentiellement Twitter ou du moins connaissent la plateforme.

5.3 Interfaces pertinentes

Les graphiques et interfaces proposés ne doivent pas avoir comme but une analyse précise du comportement genevois. Il faut garder à l'esprit que le public doit pouvoir comprendre ce qu'il voit durant les 30 secondes qu'il passera devant l'écran lors de sa visite. Il faut donc que les graphiques soient clairs, rapidement compréhensibles, voire même animés tout en étant pertinents et respectueux des concepts vus précédemment. Toutes les interfaces décrites ci-dessous sont représentées sous forme de maquettes au chapitre suivant.

5.3.1 Répartition des langues à Genève

L'interface choisie pour la répartition des langues à Genève serait décomposée en deux sous-parties : thermique et quartiers.

5.3.1.1 Thermique

Cette interface a comme but d'indiquer la répartition des langues à Genève sans poser de frontières comme le ferait un quartier. Il s'agirait ici d'avoir un état global des langues.

5.3.1.1.1 Carte

Etant donné qu'une répartition des langues est souhaitée, il faut profiter du potentiel visuel de des coordonnées géographiques disponibles pour pouvoir les représenter sur une carte. Afin de rester dans la thématique de la répartition des langues, afficher tous les tweets sans distinction serait inutile. L'utilisateur pourrait donc choisir une langue parmi toutes celles qui ont été récoltées et une carte de type thermique afficherait leur répartition géographique.

Exemple : l'affichage des tweets en français a été sélectionné. Les concentrations de tweets en français seraient affichées en rouge alors que les tweets sporadiques le seraient en vert clair. Il n'est pas nécessaire d'indiquer une échelle des couleurs pour savoir à combien de tweets correspond le rouge ou le vert puisque nous nous adressons au grand public. Il est donc suffisant de savoir où et dans quelle mesure une langue apparaît selon les endroits.

Un clic sur une agglomération de points géographiques afficherait une fenêtre comportant un nuage de mots juste au-dessus de la coordonnée choisie. Ce nuage indiquerait les hashtags les plus fréquents de la zone. Bien que décrié par les défenseurs des bonnes pratiques de visualisation, le nuage de mots reste un outil très pratique lorsqu'il s'agit de pouvoir afficher et comprendre une tendance en un coup d'œil. C'est exactement ce que nous recherchons comme effet dans le cadre de notre exposition.

Exemple : un amas de tweets se forme à un endroit de la carte, un clic dessus nous affiche un nuage de mots comme hashtags principaux les termes #TPG #accident. Il n'en faut pas plus pour comprendre la raison de ce regroupement.

Il serait possible d'afficher les termes les plus fréquents au lieu des hashtags uniquement, en ignorant au passage les déterminants, etc.

5.3.1.1.2 Classement

Il est pertinent de connaître les langues principales utilisées globalement au niveau du canton. Pour cela, un classement des langues les plus utilisées serait proposé à l'utilisateur. Afficher un classement de toutes les langues serait trop lourd. Les cinq premières suffiraient amplement avec une catégorisation « autres » pour le reste. Cette hiérarchie serait effectuée sous forme de diagramme à barres horizontales possédant la proportion de la langue dans sa barre. Nous avons vu qu'il était plus facile de comparer des valeurs avec ce graphique qu'un diagramme circulaire.

Une fois la langue choisie, l'utilisateur aurait également un classement des hashtags les plus fréquents. Par exemple, si le français est choisi, il verrait les hashtags « lac, Genève, soleil, vacances » dans sa liste.

5.3.1.1.3 Chronologie

Une timeline serait présente pour permettre à l'utilisateur d'afficher l'état de la carte ou du classement des langues et des hashtags à un instant précis. Une idée serait d'indiquer les événements marquants comme « salon du livre » ou « salon de l'auto » sur la timeline afin de pouvoir comparer ces instants spécifiques au reste.

5.3.1.1.4 Rejouer

Un bouton permettrait de rejouer, par exemple, la dernière semaine à l'utilisateur.

Exemple : si nous sommes le 23 juin à 12h00 et que l'utilisateur appuie le bouton replay, la carte afficherait une animation de son état le 16 juin à 12h00 puis à 12h30 et ainsi de suite jusqu'à la date actuelle du 23 juin 12h00. Une demi-heure pourrait passer en une seconde.

Cette fonctionnalité permettrait de visualiser des mouvements de foules comme autour d'un bord de lac durant un weekend.

5.3.1.2 Quartiers

Cette interface servirait à afficher les répartitions linguistiques en fonction des quartiers.

5.3.1.2.1 Carte

Une carte où les quartiers seraient délimités par des bordures serait présentée à l'utilisateur. Un clic sur un quartier afficherait un diagramme en barres avec les cinq langues les plus présentes ainsi que le hashtag le plus populaire de ce quartier. Il s'agit ici de voir quelle langue prédomine dans quelle région.

5.3.1.2.2 Classement

Comme pour la carte thermique, un classement sous forme de diagramme en barres des cinq langues les plus fréquentes serait proposé.

5.3.1.2.3 Chronologie

La timeline resterait présente avec les mêmes fonctionnalités que la carte thermique.

5.3.1.2.4 Derniers tweets

Une fenêtre afficherait les derniers tweets du quartier sur lequel on a cliqué.

5.3.1.2.5 Ecrire un tweet

Il serait possible de twitter directement depuis la plateforme. N'ayant pas de lien direct avec la visualisation du Big Data, cette fonction ajouterait une forme d'interaction pour l'utilisateur. Il pourrait directement voir son tweet affiché dans le quartier correspondant.

5.3.2 Rayonnement de Genève dans le monde

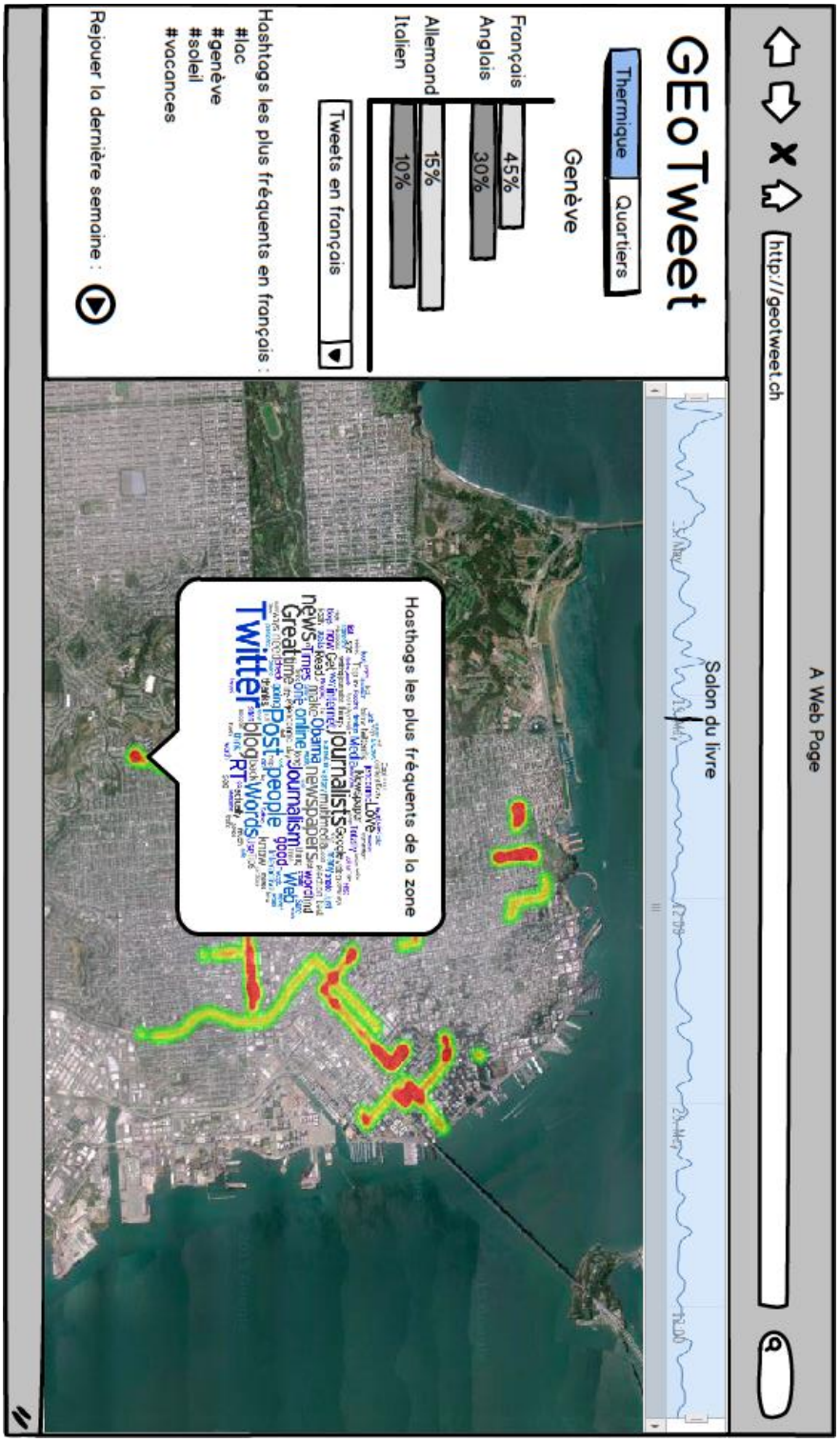
Cette interface devrait être quelque chose de visuellement simple, mais éloquent. Il s'agit dans un premier temps d'identifier un tweet évoquant « Genève ». Cela peut être mis en place par un filtre qui capturerait tous les tweets comportant le terme Genève dans n'importe quelle langue. Une fois un tweet capturé, il serait affiché sous forme d'étincelle sur une carte du monde. Au bout de quelques secondes, cette étincelle disparaîtrait avec l'impossibilité de retrouver ce tweet éphémère. Un clic sur une étincelle ouvrirait une fenêtre affichant le message du tweet en question. De plus, si un tweet est retweeté ailleurs, une ligne partirait du message d'origine pour rejoindre le lieu du retweet où aurait lieu une seconde étincelle à son arrivée. Ainsi on pourrait très facilement voir où Genève apparaît dans le monde.

Cette carte, grandement inspirée par une création d'Artem Zubrov pour la visualisation de transactions bancaires, peut être visualisée en action : http://d3.artzub.com/wbca/

5.4 Proposition de vues – prototypes

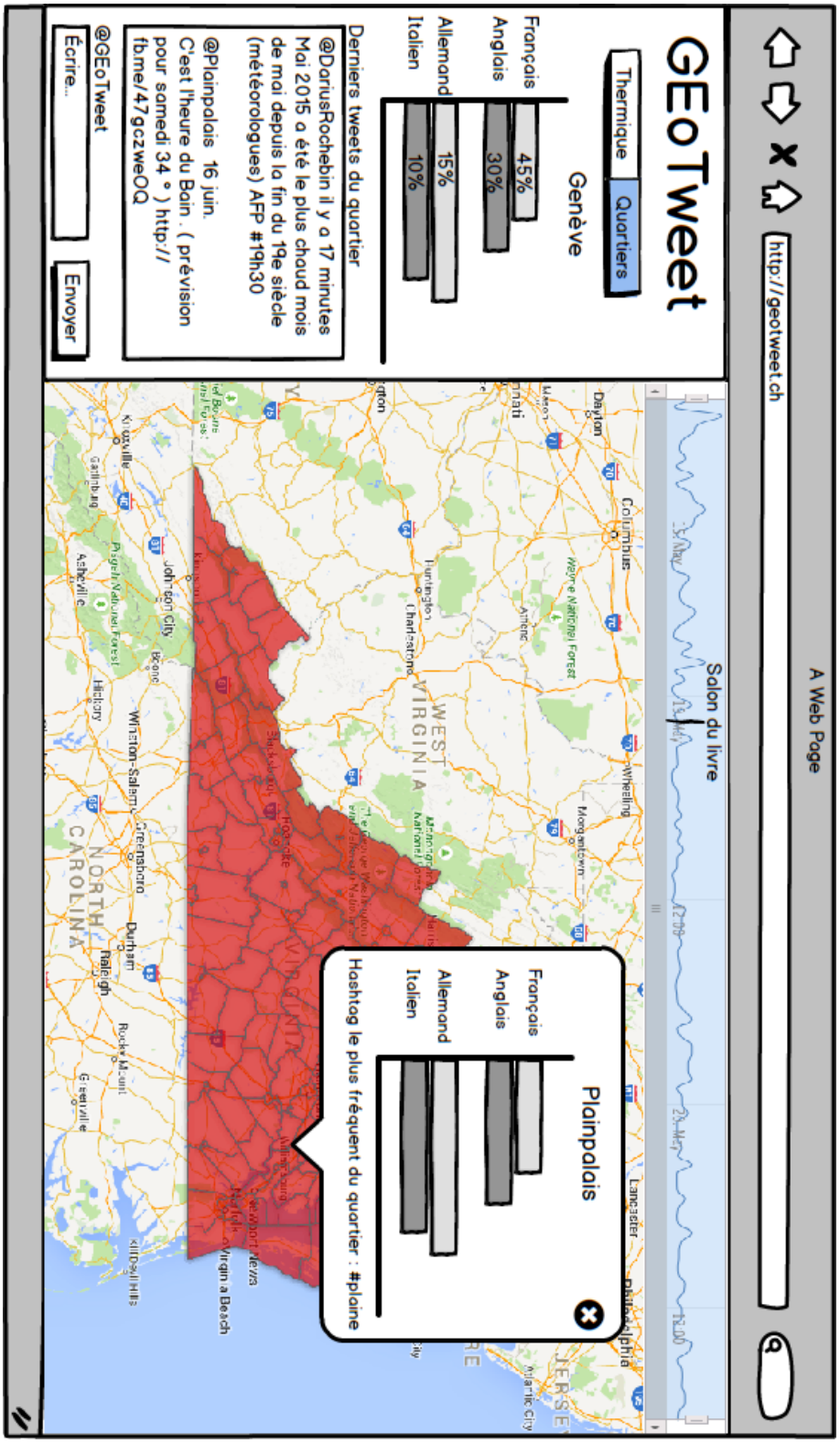
5.4.1 Répartition des langues à Genève

Figure 25 : Carte thermique de la répartition des langues



Source : JEANNERET Philippe, 15 juin 2015

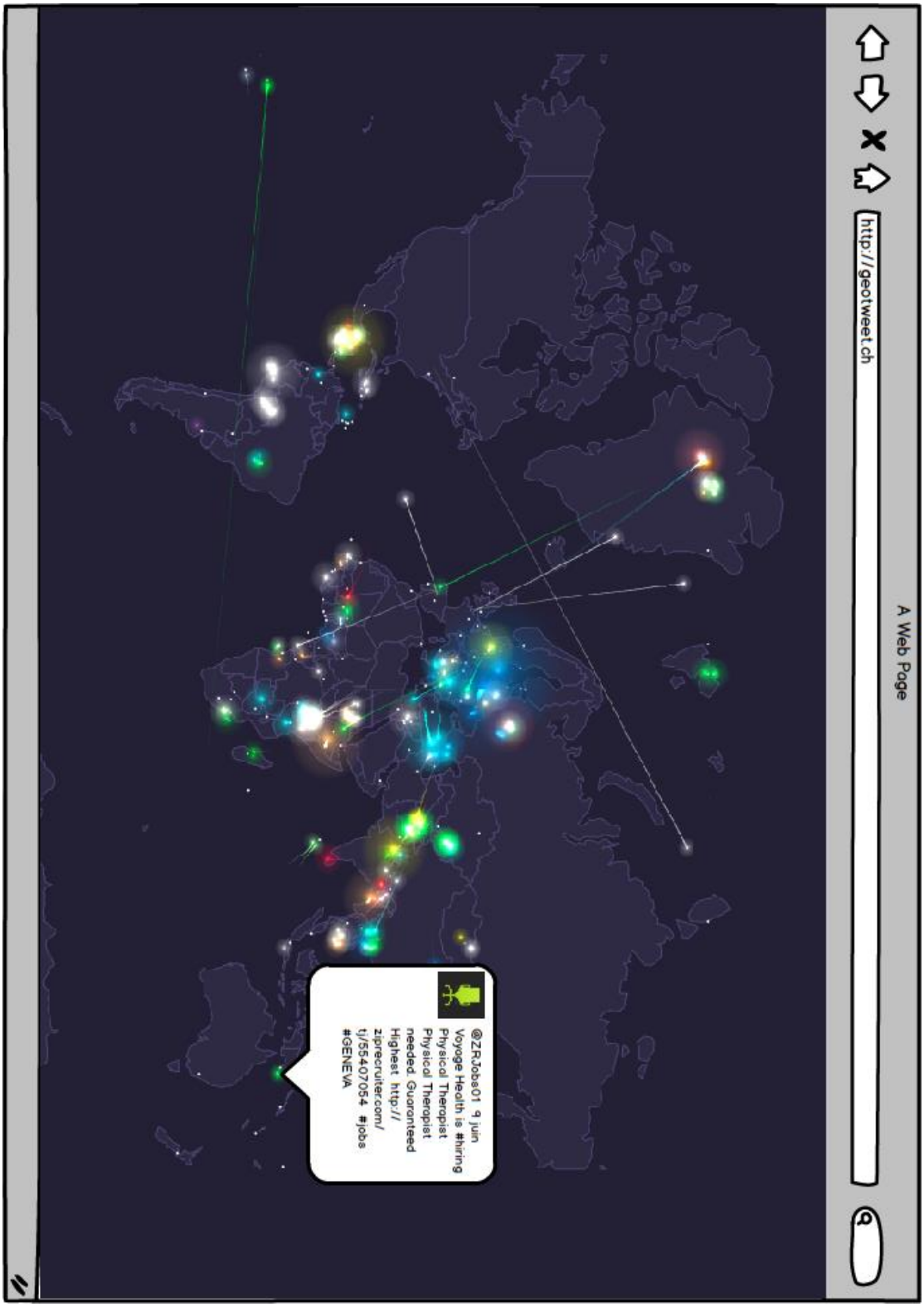
Figure 26 : Carte de la répartition des langues par quartiers



Source : JEANNERET Philippe : 15 juin 2015

5.4.2 Rayonnement de Genève dans le monde

Figure 27 : Carte du rayonnement de Genève dans le monde



Source : adapté de Artem Zubkov, JEANNERET Philippe 16 juin 2015

5.5 Choix technologiques

L'équipe GEoTweet a expressément demandé que la visualisation soit faite grâce à des technologies web.

Il existe trois possibilités pour visualiser des données dans le monde du web :

- **Importer vers un outil en ligne** – ex : le site CartoDB fonctionne comme un SaaS et propose à l'utilisateur d'envoyer sa base de données dans un format spécifique reconnu par le site. L'utilisateur aura ensuite accès à une interface lui permettant de personnaliser la carte selon ses envies. La personnalisation est limitée, puisqu'il n'est pas possible de rajouter des éléments non-prévus comme un bouton spécifique effectuant telle action. En contrepartie les connaissances requises pour l'utilisation de cette catégorie ne sont pas poussées.
- **Télécharger une librairie** – ex : le site CartoDB met à disposition un fichier JavaScript pouvant être téléchargé et importé dans un site web. Il est ainsi possible d'effectuer toutes les actions que ferait l'interface en important ses données. Il s'agit donc d'une base sur laquelle on peut développer sa visualisation. Les connaissances demandées sont ainsi plus importantes, mais le sujet peut être mieux modelé.
- **Partir de zéro** – ex : un graphique pourrait être créé grâce à la librairie D3, un autre grâce à Highcharts et une carte par Google Maps. Il s'agit ici de mélanger les outils. Les connaissances requises sont élevées ainsi que le temps de développement, mais il est possible de créer un graphique tel qu'on le souhaite.

L'équipe GEoTweet a bien identifié ces trois catégories et a choisi de s'orienter sur la dernière d'entre-elles. Toutefois, les membres d'information documentaire n'ayant pas des connaissances en JavaScript, les premiers prototypes seront vraisemblablement effectués avec la première option.

5.6 Validation par l'équipe GEoTweet et analyse des résultats obtenus

Les interfaces présentées au chapitre 5.4 de ce document ont toutes été validées.

Certaines pistes ont néanmoins été avancées par l'équipe : la carte thermique de la répartition des langues pourrait afficher plusieurs langues simultanément. Personnellement, je doute de la lisibilité de ce type d'affichage. En effet, une langue est déjà représentée par quatre couleurs sur une carte thermique (vert, jaune, orange rouge). Deux langues signifient donc huit couleurs à afficher. Il doit être possible de le faire avec des couleurs qui seraient plus parlantes, en allant du bleu clair au bleu foncé ou du vert clair au vert foncé. A ce moment, la question de la superposition des données se pose. Comment représenter un amas bleu foncé et vert foncé qui seraient au même endroit ? Ce sont des questions auxquelles je n'ai pas trouvé de réponses satisfaisantes et c'est pour cela que j'ai préféré abandonner cette piste.

Une autre idée serait de jouer avec un autre sens que la vue en lisant des tweets grâce à un outil de synthèse vocale. J'ai expérimenté ce concept lors d'une exposition des gagnants d'un concours sur la visualisation du Big Data en mai 2015 à Genève. Le gagnant jouait avec le son pour visualiser des données. Le rythme s'accélérait ou ralentissais en fonction de l'importance d'une valeur.

Ex : San Francisco et ses 30°C avait son propre rythme. Celui-ci ralentissait si on passait à Genève où il faisait 23°C.

J'ai trouvé très intéressant de pouvoir voir, mais aussi entendre des données. Ce qui m'a particulièrement frappé fut que cette rythmique n'était absolument pas dure à supporter. Je crains que la lecture de tweets n'apporte pas de nouveauté à la visualisation des données. Dans le cadre de GEoTweet, il s'agirait de lire une donnée telle que l'utilisateur pourrait la lire. Elle ne serait jamais transformée (valeur en rythme). Dès lors, la plus-value est quasi-inexistante à mes yeux, sans parler du fait que cela pourrait devenir pénible pour le visiteur.

6. Conclusion

Le Big Data possède sa littérature comme le prouvent les 4 V, la structuration des données ou encore la catégorisation d'IBM. Ces théories semblent fondées vu la fréquence à laquelle elles sont citées dans le milieu.

Par contre, l'absence de théories sur la visualisation du Big Data est flagrante. On passe d'un monde où tout porte un nom et est expliqué, voir même trop détaillé, à un univers où aucune logique ne règne. Des visualisations sont présentées et on voit ce qu'il est possible de faire, mais quand on cherche à savoir « pourquoi ce graphique et pas un autre ? », la question demeure sans réponse. Il suffit de faire un tour sur la page des exemples de la librairie D3 pour se rendre compte de la multitude de modèles existants. En cliquant sur un auteur, on accède à son site qui contient tout autant d'exemples. Bernard Marr dit qu'il existe autant de graphiques différents que de données. Je dirais plutôt qu'il existe autant de graphiques que de designers. En effet, plus je m'avançais dans la recherche concernant la visualisation, plus je devenais critique en me demandant si quelqu'un avait fait ce graphique ainsi simplement parce qu'il le trouvait esthétique. J'ai donc choisi de lister les graphiques les plus communs plutôt que d'en afficher une multitude de trop spécifiques en ayant la certitude que j'allais passer à côté de certains. L'inconvénient est donc que je passe un peu à côté de ces nouvelles et trop nombreuses méthodes d'affichages.

Concernant GEoTweet, cette absence de règles de visualisation m'a laissé par moment devant un choix, non voulu, pour sélectionner des interfaces. Je me suis retrouvé dans la situation décrite précédemment où j'ai choisi un affichage car il était esthétique. Aucune règle ne vient indiquer qu'il faut mettre une carte thermique. Je me suis simplement dit que cela ferait sens ici. Cette fragilité du côté de GEoTweet s'explique par la faiblesse du lien qui relie la théorie des données du Big Data à la visualisation. Sûrement que cette relation existe chez les professionnels du design, mais je doute qu'ils aient tous une théorie commune.

J'ai apprécié le fait que le Big Data suscite tant de débats. Certains y voient là un mal, comme le pape qui déclarait cette semaine (18 juin 2015) que «La vraie sagesse, fruit de la réflexion, du dialogue et de la rencontre généreuse entre les personnes ne s'acquiert pas avec une simple accumulation des données qui finit par saturer et troubler dans une espèce d'empoisonnement mental.». D'autres au contraire ne jurent que par ça, généralement dans une optique de profit. Alors que certains, que je qualifierais d'explorateurs, visualisent des données pour trouver des questions puis tenter d'y répondre.

Ex : en analysant les sujets des journaux américains, David McCandless s'est rendu compte que les jeux-vidéos violents constituent un sujet récurrent en avril et novembre. Pourquoi ces mois-ci précisément ? Novembre est le moment où les nouveaux jeux sortent et coïncide avec l'approche des fêtes durant lesquelles on les offrira. Il est normal que le sujet soit soulevé par la presse durant cette période. Pourquoi avril ? En cherchant, David McCandless s'est rendu compte qu'une tuerie à Columbine, motivée par les jeux-vidéos violents, a eu lieu en avril 1999. Dès lors, ce souvenir est remis en avant chaque année à la même période.

En conclusion, je dirais qu'il est difficile et restera difficile d'appliquer une théorie pour la visualisation du Big Data. De bonnes pratiques peuvent se mettre en place, mais la quantité de designers, la variété des données, d'émotions que peut susciter le Big Data et le but recherché de l'affichage réduisent la probabilité que deux graphiques se ressemblent. L'avenir me donnera peut-être tort compte tenu de la jeunesse des techniques de visualisation, en témoignent les dates récentes des articles utilisés dans la bibliographie.

Bibliographie

GOOGLE TRENDS, Google Trends, 2015. Google Trends Big Data. *google.ch* [en ligne]. 2015. 3. mai 2015. [Consulté le 03.05.2015]. Disponible à l'adresse : http://www.google.ch/trends/explore#q=Big%20Data

DIVAKAR, Mysore, KHUPAT, Shrikant, JAIN, Shweta, 2013. Big data architecture and patterns, Part 1: Introduction to big data classification and architecture. *ibm.com* [en ligne]. 17 septembre 2013. 17 septembre 2013. [Consulté le 15 mai 2015]. Disponible à l'adresse : http://www.ibm.com/developerworks/library/bd-archpatterns1/

NYPD, Nypd, 2015. NYC Crime Map. *nyc.gov* [en ligne]. avril 2015. 30 avril 2015 [Consulté le 23 mai 2015]. Disponible à l'adresse : http://maps.nyc.gov/crime/

DMARTYR, DMARTYR, 2008. Obama's Victory Speech Word Cloud. *snappedshot.com* [en ligne]. 5 novembre 2008. 5 novembre 2008 [Consulté le 25 mai 2015]. Disponible à l'adresse : http://snappedshot.com/ghost/ky3

HIGHCHARTS, Highcharts, 2015. Column with rotated labels. *highcharts.com* [en ligne]. 2015. 2015 [Consulté le 29 mai 2015]. Disponible à l'adresse : http://www.highcharts.com/demo/column-rotated-labels

HIGHCHARTS, Highcharts, 2015. Basic line. *highcharts.com* [en ligne]. 2015. 2015 [Consulté le 29 mai 2015]. Disponible à l'adresse : http://www.highcharts.com/demo/line-basic

HIGHCHARTS, Highcharts, 2015. Single line series. *highcharts.com* [en ligne]. 2015. 2015 [Consulté le 30 mai 2015]. Disponible à l'adresse : http://www.highcharts.com/stock/demo/basic-line

HIGHCHARTS, Highcharts, 2015. Stacked area. *highcharts.com* [en ligne]. 2015. 2015 [Consulté le 30 mai 2015]. Disponible à l'adresse : http://www.highcharts.com/demo/area-stacked

BOSTOCK, Mike, 2012. Circle Packing. *bl.ocks.org* [en ligne]. 13 novembre 2012. 13 novembre 2012 [Consulté le 31 mai 2015]. Disponible à l'adresse : http://bl.ocks.org/mbostock/4063530

BOSTOCK, Mike, 2012. Uber Rides by Neighborhood. *bost.ocks.org* [en ligne]. 9 janvier 2012. 9 janvier 2012 [Consulté le 1 juin 2015]. Disponible à l'adresse : http://bost.ocks.org/mike/uberdata/

DEMAJ, Damien, 2015. The Symmetry of The Tennis Serve. *gamesetmap.com* [en ligne]. 15 février 2015. 15 février 2015 [Consulté le 2 juin 2015]. Disponible à l'adresse : http://gamesetmap.com/?p=1098

COOK, Peter, . Animated UK Wind Chart. *animateddata.co.uk* [en ligne]. .[Consulté le 4 juin 2015]. Disponible à l'adresse : charts.animateddata.co.uk/ukwind/

MCCANDLESS, David, POSAVEC, Stefanie, 2010. Left vs Right US. *informationisbeautiful.net* [en ligne]. décembre 2010. décembre 2010 [Consulté le 20 juin 2015]. Disponible à l'adresse : http://www.informationisbeautiful.net/visualizations/left-vs-right-us/

ABELA, Andrew, 2013. *Advanced Presentations by Design: Creating Communication that Drives Action*. Pfeiffer. Hoboken : Wiley, 22 avril 2013. Paperback. .,. ISBN 978-1118347911

SPIRA, Dan, 2009. Same Data, Different Graphs. *danspira.com* [en ligne]. 8 juillet 2009. 8 juillet 2009 [Consulté le 8 juin 2015]. Disponible à l'adresse : http://danspira.com/2009/07/08/same-data-different-graphs/

ZUBKOV, Artem, 2003. WorldBank Contract Awards. *artzub.com* [en ligne]. 2013. 2013 [Consulté le 13 juin 2015]. Disponible à l'adresse : http://d3.artzub.com/wbca/

BOSTOCK, Mike, 2012. Sankey Diagrams. *bost.oks.org* [en ligne]. 22 mai 2012. 22 mai 2012 [Consulté le 19 juin 2015]. Disponible à l'adresse : http://bost.ocks.org/mike/sankey/

BYRNE, John A, 2012. The 12 greatest entrepreneurs of our time. *fortune.com* [en ligne]. 9 avril 2012. 9 avril 2012 [Consulté le 7 mai 2015]. Disponible à l'adresse : http://archive.fortune.com/galleries/2012/news/companies/1203/gallery.greatest-entrepreneurs.fortune/12.html

CHATTERJEE, Pratap, 2013. Big data: the greater good or invasion of privacy. *fortune.com* [en ligne]. 12 mars 2013. 12 mars 2013 [Consulté le 8 mai 2015]. Disponible à l'adresse : http://www.theguardian.com/commentisfree/2013/mar/12/big-data-greater-good-privacy-invasion

LANEY, Doug, 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *gartner.com* [en ligne]. 6 février 2001. 6 février 2001 [Consulté le 5 mai 2015]. Disponible à l'adresse : http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

MAYER-SCHONBERGER, Viktor, CUKIER, Kenneth, 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think.* John Murray. Londres : John Murray, 10 octobre 2013. New York Times. .,. ISBN 978-1848547926

MARR, Bernard, 2015. *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Wiley. Hoboken : Wiley, 30 janvier 2015. Wiley. .,. ISBN 978-1118965832

MEER, David, 2013. What is „Big Data“ Anyway. *strategy-business.com* [en ligne]. 25 juin 2013. 25 juin 2013 [Consulté le 8 mai 2015]. Disponible à l'adresse : http://www.strategy-business.com/blog/What-Is-Big-Data-Anyway?gko=28596

MARR, Bernard, 2013. Big Data Analytics and Visualization – with Bernard Marr [podcast]. 2 décembre 2013. [Consulté le 18 mai 2015]. Disponible à l'adresse : http://www.ibmbigdatahub.com/podcast/big-data-analytics-and-visualization-bernard-marr

SHIN, Jennifer, 2015. Data Visualization Playbook: Revisiting the Basics. *ibmbigdatahub.com* [en ligne]. 13 mars 2015. 13 mars 2015 [Consulté le 12 mai 2015]. Disponible à l'adresse : http://www.ibmbigdatahub.com/blog/data-visualization-playbook-revisiting-basics

HARRIS, Jackob, 2011. Word clouds considered harmful. *niemanlab.org* [en ligne]. 13 octobre 2011. 13 octobre 2011 [Consulté le 12 mai 2015]. Disponible à l'adresse : http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

MCCANDLESS, David, 2010. David McCandless: The beauty of data visualization [enregistrement vidéo]. *ted.com* [en ligne]. juillet 2010. [Consulté le 19 juin 2015]. Disponible à l'adresse : http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=fr

BROWN, Brad, CHUI, Michael, MANYIKA, James, 2011. Are you ready for the era of « big data » ?. *t-systems.com* [en ligne]. octobre 2011. octobre 2011 [Consulté le 6 mai 2015]. Disponible à l'adresse : http://www.t-systems.com/solutions/download-mckinsey-quarterly-/1148544_1/blobBinary/Study-McKinsey-Big-data.pdf

STUART WARD, Jonathan, BARKER, Adam, 2013. Undefined By Data: A Survey of Big Data Definitions. *arxiv.org* [en ligne]. 20 septembre 2013. 20 septembre 2013 [Consulté le 6 mai 2015]. Disponible à l'adresse : http://arxiv.org/pdf/1309.5821v1.pdf

LOHR, Steve, 2012. The Age of Big Data. *nytimes.com* [en ligne]. 11 février 2012. 11 février 2012 [Consulté le 7 mai 2015]. Disponible à l'adresse : http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=0

DATACANVAS, Datacanvas, 2015. Sonic Particles 2.0. *datacanvas.org* [en ligne]. 25 mars 2015. 25 mars 2015 [Consulté le 21 juin 2015]. Disponible à l'adresse : http://datacanvas.org/project/sonic-particles-2-0/

SCHMARZO, Bill, 2013. *Tirer parti des données massives pour développer l'entreprise*. Wiley. Hoboken : Wiley, 7 octobre 2013. Paperback. .,. ISBN 978-1118739570

KALOUSIS, Alexandros, 2014. *Introduction au Big Data, découverte de connaissance à partir de données* [document PDF].

Support de cours : Cours « Data Mining », Haute Ecole de Gestion de Genève, filière Informatique de Gestion, année académique 2014-2015