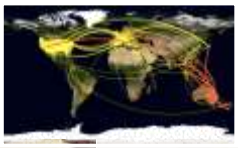
Aviz

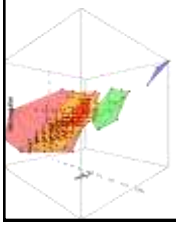
Visual Analytics Project

Visualisation d'information : des principes au passage à l'échelle

Jean-Daniel Fekete

www.aviz.fr



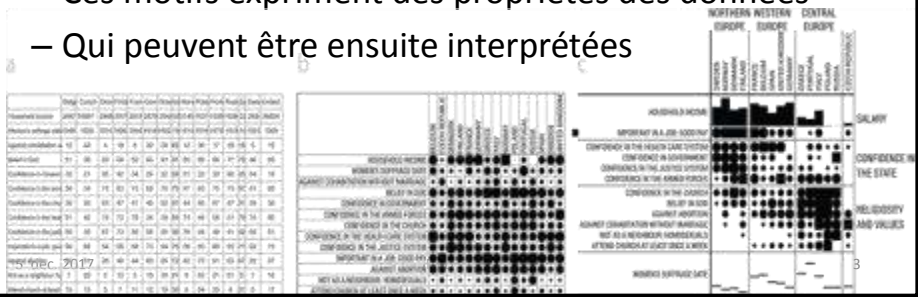


AVIZ fait de la visualisation

www.aviz.fr/Research/Projects

Visualisation de données

- Représenter graphiquement des données et interagir avec ces représentations
  - Utiliser nos capacités perceptives et cognitives
  - Pour percevoir des **motifs visuels** dans les graphiques
  - Ces motifs expriment des propriétés des données
  - Qui peuvent être ensuite interprétées



Visualisation vs. statistiques

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Raw Data from Anscombe's Quartet

[Source: Anscombe's quartet, Wikipedia]

Analyse statistique

Pour les 4 jeux des données, les valeurs sont identiques

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

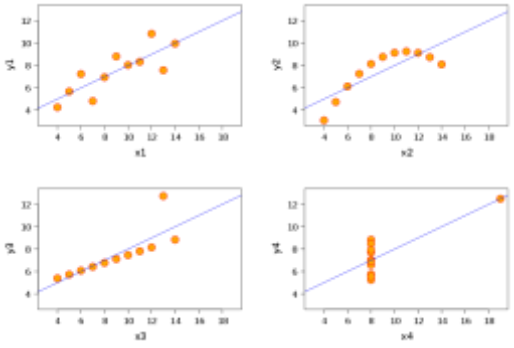
Moyenne des *x*	9.0
Variance des *x*	11.0
Moyenne des *y*	7.5
Variance des *y*	4.12
Correlation entre *x* et *y*	0.816
Régression linéaire	$y = 3 + 0.5x$

[Source: Anscombe's quartet, Wikipedia]

Représentation visuelle

L'histoire est différente

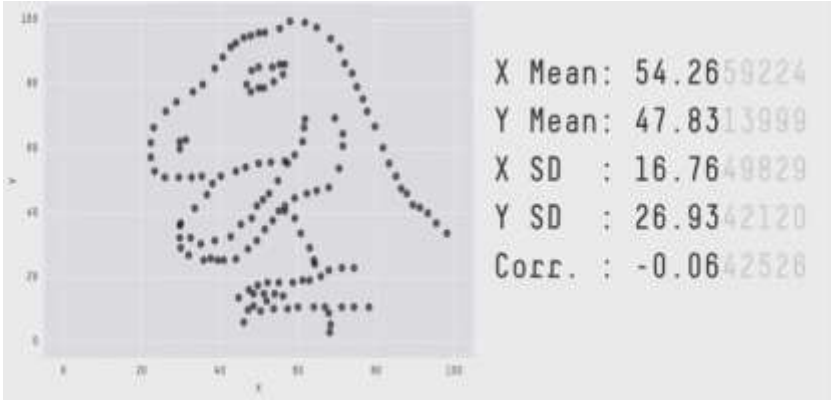
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



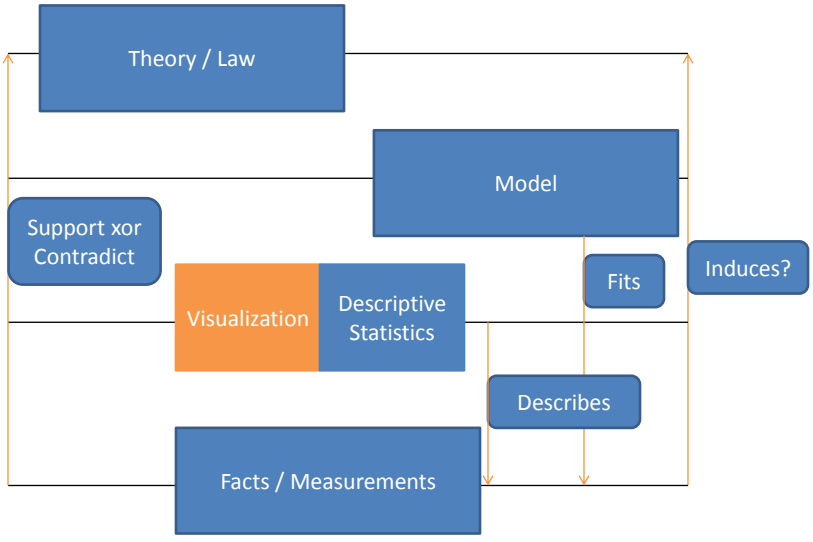
[Source: Anscombe's quartet, Wikipedia]

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing [CHI17]

https://www.autodeskresearch.com/publications/samestats



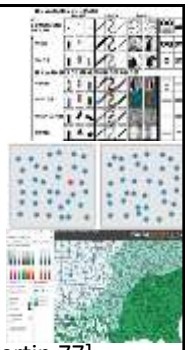
Où se place la visualisation ?



Visualisation d'information

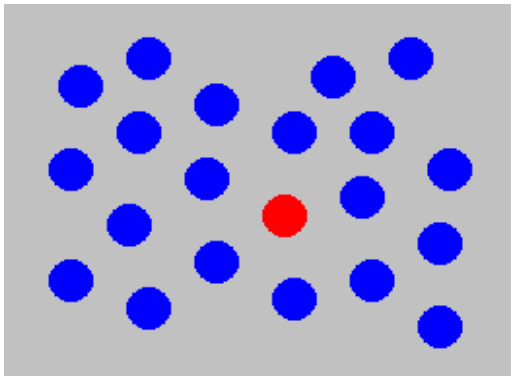
- The use of computer-supported, interactive, visual representations of abstract data to amplify cognition. [Card et al., 1999]
- Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively. [Munzner 2014]

Bases de la visualisation

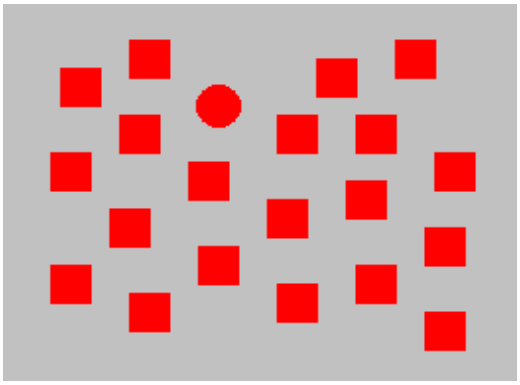


- Représentation visuelle
  - Sémiologie graphique [Bertin 67]
  - Perception préattentive [Triesman 86]
  - Perception des couleurs [Brewer etc.]
- Interaction
  - La graphique et le traitement graphique de l'information [Bertin 77]
  - Dynamic Queries [Shneiderman et al. 93] http://www.cs.umd.edu/hcil/spotfire/
- Systèmes
  - Toolkit D3 [Bostock&Heer 11]
  - Produits commerciaux comme Tableau
  - Grammar of Graphics [Wilkinson 99]
    - Implementé dans GGPlot2, Vega-Lite, G2
  - Livre de cours: Visualization Analysis&Design [Munzner 14]

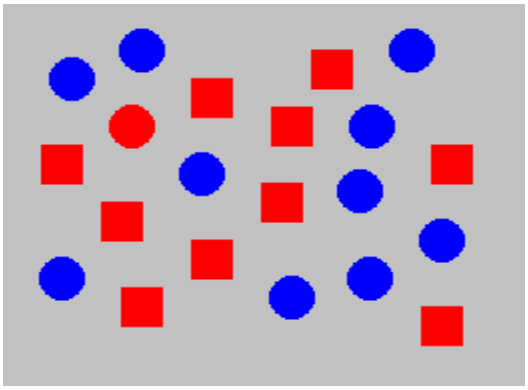
Perception préattentive (1)



Perception préattentive (2)



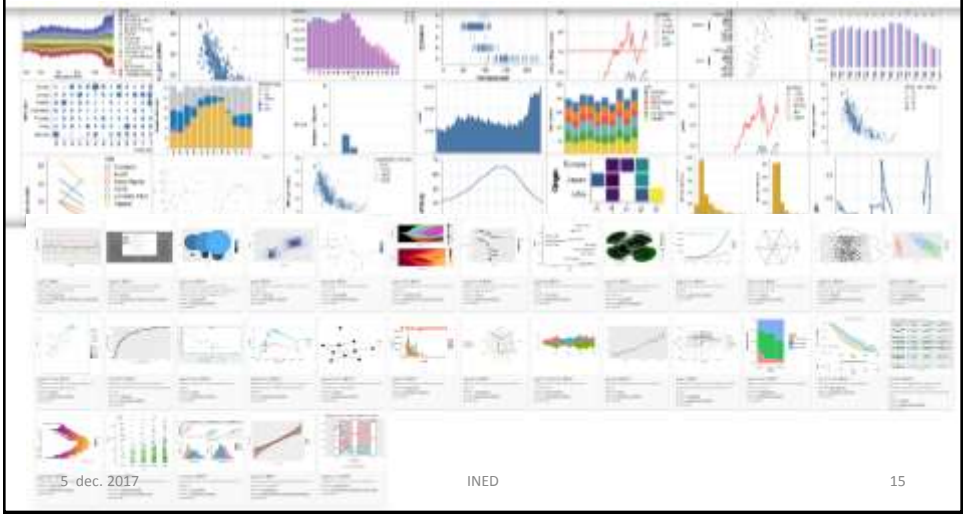
Perception préattentive (3)



Perception préattentive : théorie

- Notre système visuel de bas niveau (25 millions de cellules) fait de la reconnaissance de motif en parallèle en permanence
- Les caractéristiques préattentives sont reconnues à ce niveau
- Les autres nécessitent un parcours séquentiel !
- On a parfois besoin de données visuelles non préattentives
  - Labels/étiquettes sur les données
  - Représentations traditionnelles acceptables par les utilisateurs novices
- Excellents théories psychologiques
  - Information Visualization: Perception for Design de Colin Ware
- Besoin de conception et réalisation de techniques qui fonctionnent
  - Recours au designer / informaticien
- Les traitements informatiques automatiques peuvent-ils faire mieux ?
  - Pas toujours

Beaucoup de solutions “clé en main”



Passage à l’échelle

Problèmes :

- Limitations perceptive et cognitive
- Limitations graphique
- Limitations de calculs

Problème : Limitations perceptive et cognitive

Perception

- Le nombre de cônes et bâtonnets est limité
- La perception préattentive est limitée

Cognition

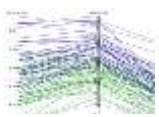
- Certains biais peuvent fausser l'interprétation
- La latence nuit gravement à l'exploration
  - 0.1 s pour la sensation de continuité (animation)
  - 1 s pour le feedback d'une commande
  - 10 s pour le retour d'une requête

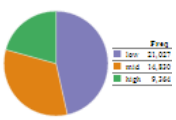
Problème : Limitations graphiques

- Plus de points de données que de pixels
  - Agréger/échantillonner
- Envoyer les données à la carte graphique ou au navigateur
  - Bande passante bus ou internet
- Taux de rafraichissement en dessous de 100ms
  - Latence

Problème : Limitations de calculs

- Seules quelques visualisations passent à l'échelle dans certaines limites
  - Barres, camemberts
- Les autre nécessitent un traitement
  - Échantillonnage
  - Agrégation
  - Projection multidimensionnelle
- Ces traitements sont coûteux
  - Ils introduisent des artéfacts
  - Ils dépassent les seuils de latence




Problème typique : le filtrage croisé

Limitations techniques

- La puissance des ordinateurs augmente
- La densité de stockage augmente
  - le prix des SSD diminue
- Mais des ordinateurs plus petits apparaîssent
  - Téléphones mobiles, montres connectées
  - Leur puissance de calcul est en compétition avec leur autonomie
- Les débits des canaux n'augmentent plus

Passage à l'échelle

Cinq stratégies :

- Pré-calcul
- HPC
- Systèmes distribués
- Calcul approché / échantillonnage
- Calcul progressif

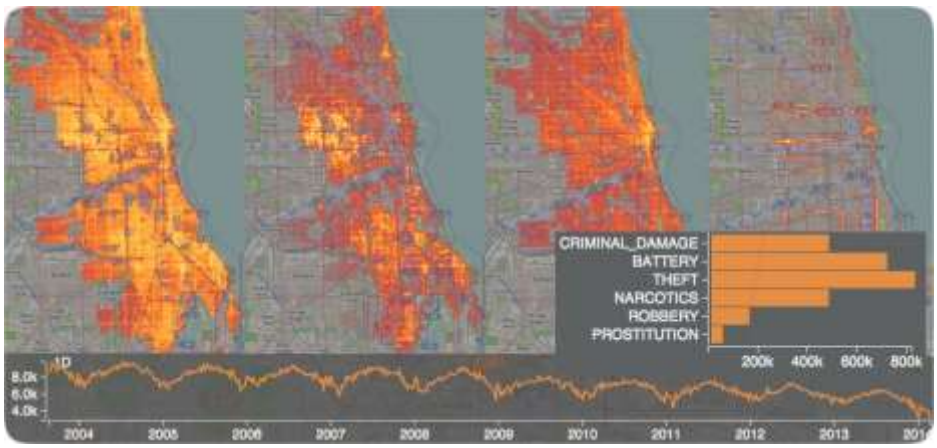
Pré-calcul

Séparation entre le calcul des agrégations utiles et de leurs explorations

- Google Maps (création de tuiles par niveau)
- Nanocubes
- imMens
- HeatMapReduce

Nanocubes (Lins et al. 2013)

http://nanocubes.net/



Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. Visualization and Computer Graphics IEEE Transactions on 19, no. 12 (2013): 2456-2465.

Nanocubes

dataset	objects (N)	memory	time	size	sharing	keys ($\|K\|$)	$\|K^*\|$	schema
brightkite	4.5 M	1.6 GB	3.50 m	149.0 M	3.00x	3.5 M	2^{74}	lat(25), lon(25), time(16), weekday(3), hour(5)
customer tix	7.8 M	2.5 GB	8.47 m	213.0 M	2.93x	7.8 M	2^{69}	lat(25), lon(25), time(16), type(3)
flights	121.0 M	2.3 GB	31.13 m	274.0 M	16.50x	43.3 M	2^{75}	lat(25), lon(25), time(16), carrier(5), delay(4)
twitter-small	210.0 M	10.2 GB	1.23 h	1.2 B	3.72x	116.0 M	2^{57}	lat(17), lon(17), time(16), device(3)
twitter	210.0 M	46.4 GB	5.87 h	5.2 B	4.00x	136.0 M	2^{60}	lat(17), lon(17), time(16), lang(5), device(3), app(2)
splom-10	1.0 B	4.3 MB	4.13 h	51.2 K	5.67x	7.4 K	2^{20}	d1(4), d2(4), d3(4), d4(4), d5(4)
splom-50	1.0 B	166.0 MB	4.72 h	8.8 M	16.00x	1.9 M	2^{30}	d1(6), d2(6), d3(6), d4(6), d5(6)
cdrs	1.0 B	3.6 GB	3.08 h	271.0 M	18.60x	96.3 M	2^{69}	lat(25), lon(25), time(16), duration(3)

Fig. 13. Summary of resource usage for our reported experimental results (K=10^3, M=10^6, B=10^9). The numbers in parentheses on the sche column denote the number of bits necessary to refer to a value of that dimension, and their sum is the exponent of 2 in the $|K^*|$ column.

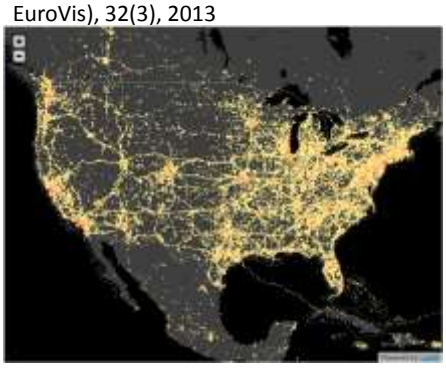
imMens

- Calcul des agrégations multidimensionnelles
- Compresse les résultats en tuiles
- Utilise le GPU, WebGL pour naviguer interactivement
- Pas d'indication de temps de création des donées agrégées

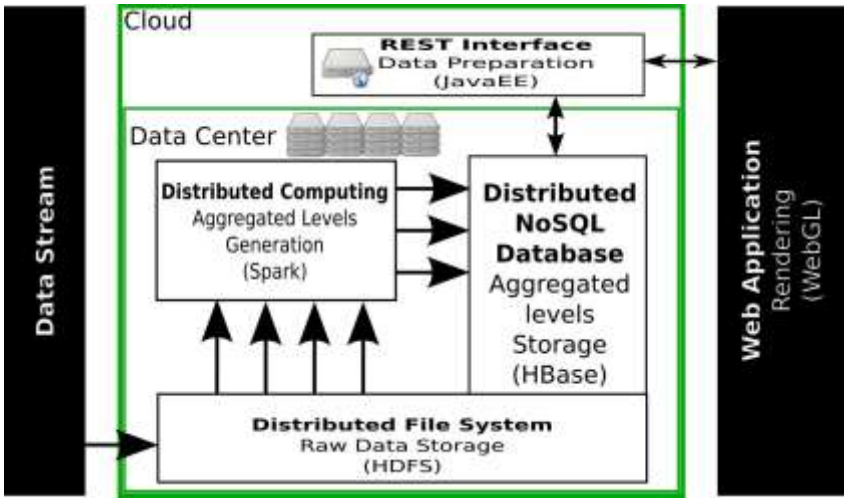
imMens: Real-time Visual Querying of Big Data

Zhicheng Liu, Biye Jiang, Jeffrey Heer

Computer Graphics Forum (Proc. EuroVis), 32(3), 2013



HeatMapReduce [Perrot et al. 15]



Alexandre Perrot, Romain Bourqui, Nicolas Hanusse, Frédéric Lalanne, David Auber. Large Interactive Visualization of Density Functions on Big Data Infrastructure. 5th IEEE Symposium on Large Data Analysis and Visualization, October 25-26, 2015. Chicago, Illinois.

HeatMapReduce [Perrot et al. 15]

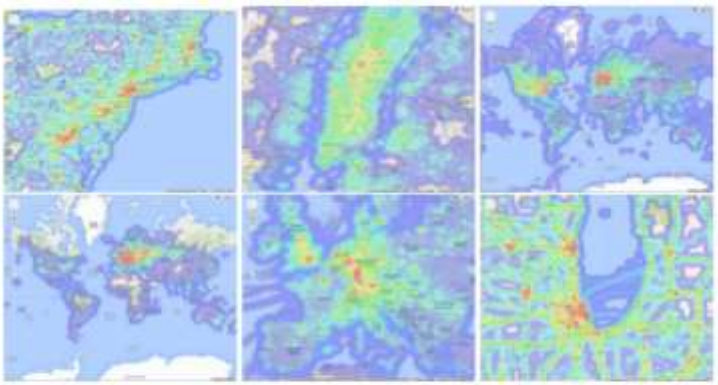
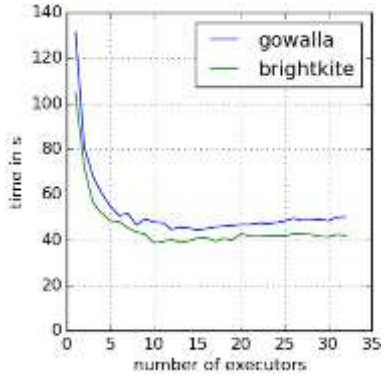
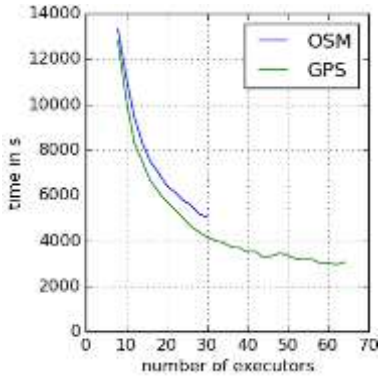


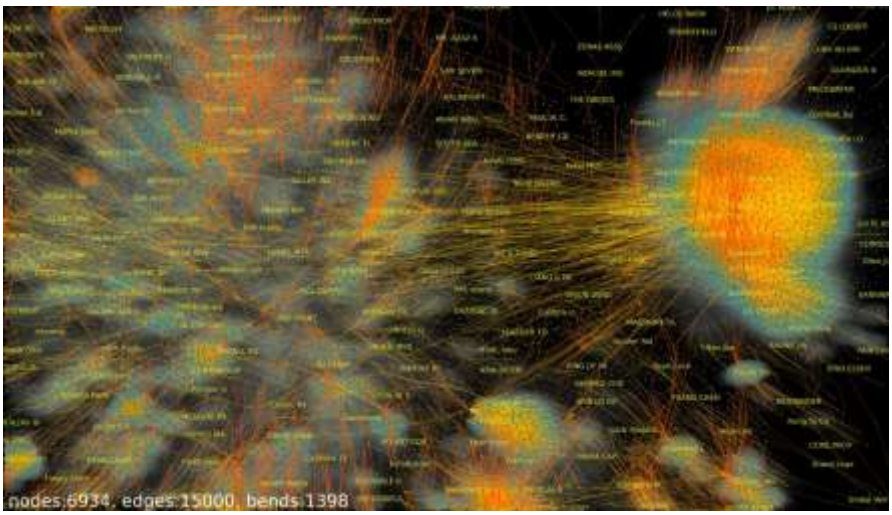
FIGURE 7.8 – Vues des quatre jeux de données utilisés, avec le nombre de points affichés. Sur la rangée supérieure, de gauche à droite : brightkite (1671), gowalla (1334) et OpenStreetMap (2873). Sur la rangée inférieure, le jeu de données GPS, de gauche à droite : vue globale (3185), Europe (3218) et Chicago (1468).

HeatMapReduce [Perrot et al. 15]





HeatMapReduce [Perrot et al. 15]



HPC

- Extension de l'ordinateur multi-cœurs
  - Très rapide mais très cher
  - Un peu compliqué à programmer, mais les choses évoluent rapidement
- Permet de repousser la quantité de données explorable, mais pour un coût prohibitif
- La plupart des algorithmes de projection ne passe pas bien à l'échelle sur ces architectures

Systèmes distribués

- Plusieurs machines sur un réseau
- Très grande puissance potentielle

- Latence difficile à contrôler, parfois très importante (plusieurs minutes)
- Systèmes industriels orientés tolérance aux pannes, pas vitesse ou latence
  - MapReduce, Hadoop, Spark
- Parfait pour le pré-calcul ou la gestion de caches

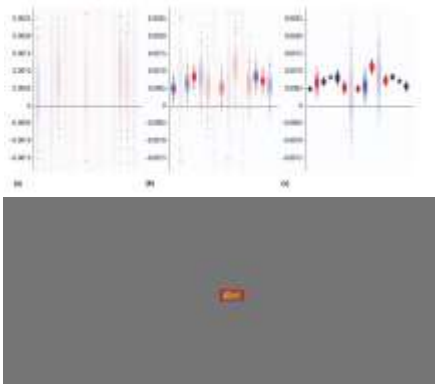
Calcul approché / échantillonnage

- Approximate computing / Sampling
- Extension “naturelle” des calculs
- Des bases de données commencent à l’offrir
  - BlinkDB
  - Mot clé “approximate” dans les requêtes
- Mais difficile à contrôler
  - Quel niveau est acceptable ?

5 dec. 2017

Calcul progressif

- Calcul approché
  - Amélioré dans le temps
  - Les estimations sont mises à jour toutes les 1/5/10 secondes
- Nouveau paradigme
  - Nécessite de reconstruire toutes les couches de programmation
- Sujet très « chaud » en recherche et dans l’industrie



Analyse de données progressive

1. Produit des estimations qui s'améliorent
   – Avec une latence bornée
2. Converge vers le résultat
3. Permet le contrôle
   – Les paramètres peuvent être changés
4. Génère des mesures de
   – Qualité
   – Progression

Avantages de l'analyse progressive

- Passage à l'échelle
  – Exploration de larges quantités de données et d'algorithmes complexes
- Décisions en avance
  – Arrêter le calcul lorsque les résultats sont inutiles ou l'algorithe mal configuré
- Compréhension de l'algorithme
  – Voir les résultats de l'algorithme au fur et à mesure de leur calcul permet de mieux comprendre le fonctionnement de l'algorithme (parfois)

Conclusion

- La visualisation exploratoire de données ne passe pas bien à l'échelle avec les architectures logicielles et matérielles actuelles
- Pour des cas particuliers où les types de requêtes sont connues, le pré-calcul fonctionne
- De nouvelles méthodes sont en cours de mise au point ou de recherche
  - Approximations, Calcul progressif
  - Pas encore sur les étagères, mais bientôt (2 à 5 ans)

Bibliographie

- Livres
  - Tamara Munzner. Visualization Analysis and Design. A K Peters Visualization Series, CRC Press, 2014. www.cs.ubc.ca/~tmm/vadbook
  - Beautiful Data (McCandless)
  - Now You See it (Few)
  - Livres de Tufte: Visual Display of Quantitative Information (et autre)
  - Sémiologie Graphique, J. Bertin, (1967) Réédition Editions EHESS 2013
- "Datavisualisation : des données à la connaissance " - Revue I2D Volume 52, N° 2, 25 juin 2015
  corist-shs.cnrs.fr/Datavisualisation_RevueI2D_2015
- Blogs
  - http://infosthetics.com/
  - http://fellinlovewithdata.com/
  - http://eagereyes.org/
  - http://flowingdata.com/
  - http://www.informationisbeautiful.net/