Les dimensions épistémologiques de l'exploration

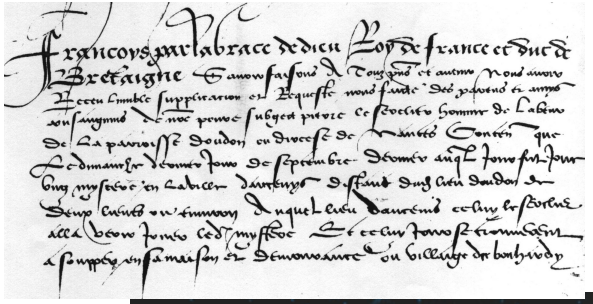
Jean-Daniel Fekete, Inria, Université Paris-Saclay
http://www.aviz.fr/~fekete

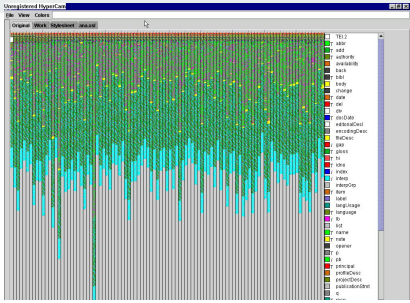
Plan

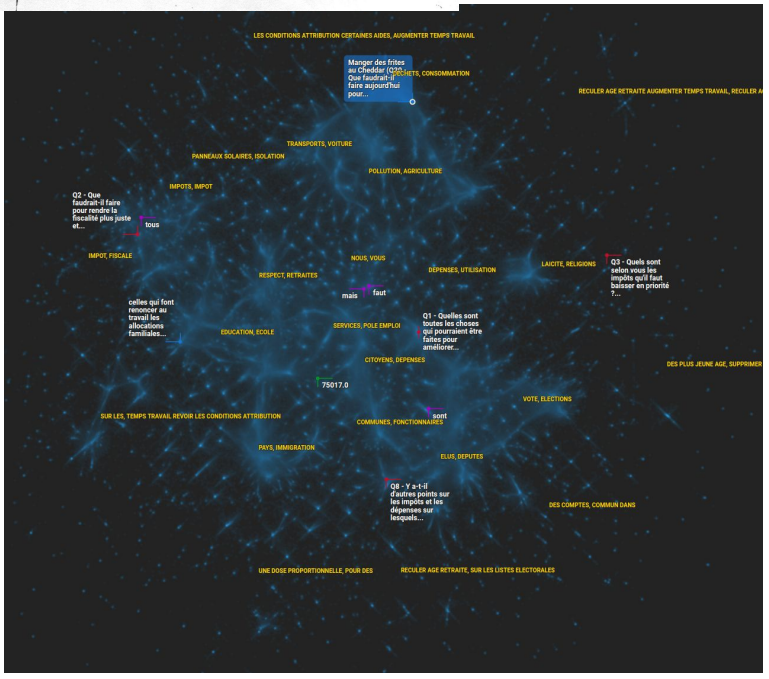
- Deux scénarios
- Les épistémologies
- Les méthodes
  - Transparence !
- Exploration de données
  - Visualisation ?
- Quelques exemples
- Précautions et faux problèmes
- Explorez explorez,
  il en restera toujours quelque chose

Deux scénarios

- Henri IV, pour faciliter l'union de la Bretagne à la France, a-t-il favorisé les nobles bretons ?
  - Vu par les lettres de rémission
- Le grand débat 2019
  - Les frites au Cheddar sont-elles bien prises en compte ?



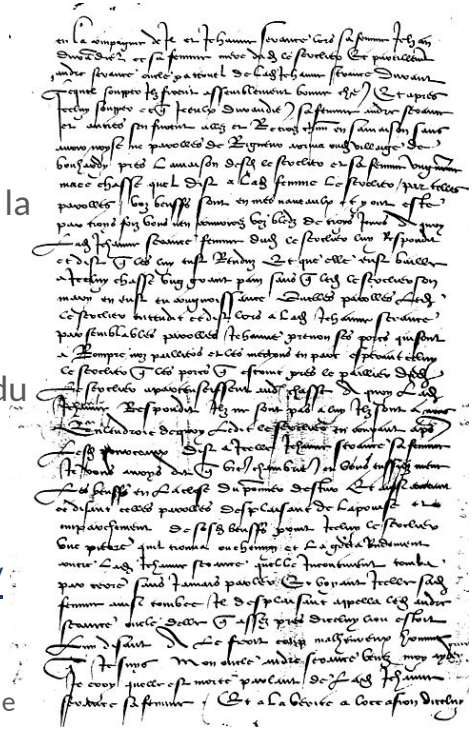


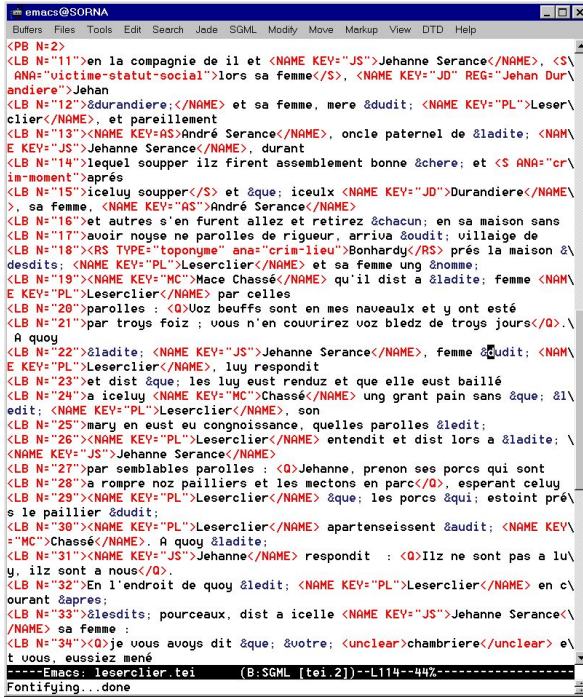


Lettres de rémission

Henri IV, pour faciliter l'union de la Bretagne à la France, réalisée en 1532 avec l'édit de Nantes, a-t-il favorisé les nobles bretons ?

- Corpus de 100 « Lettres de rémission » du duché de Bretagne XVIe siècle »
- Transcrites par Nicole Dufournaud
- Encodées en XML TEI
- http://nicole.dufournaud.org/remission/





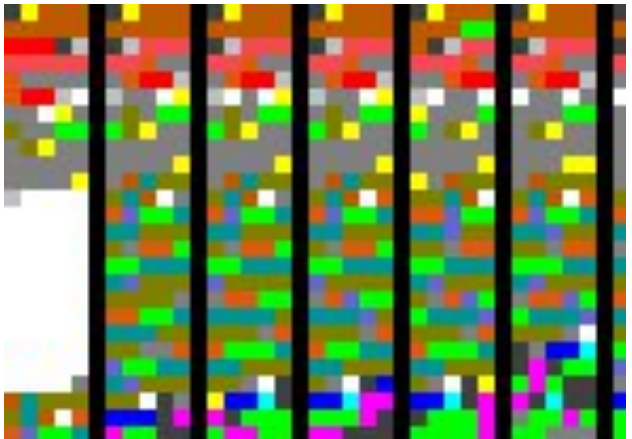
Nicole Dufournaud, Les femmes en Bretagne au XVIe siècle : étude des pratiques sociales et économiques; Perspectives de recherche et méthodologie, (mémoire de DEA), Université de Nantes,faculté des lettres. 2000.

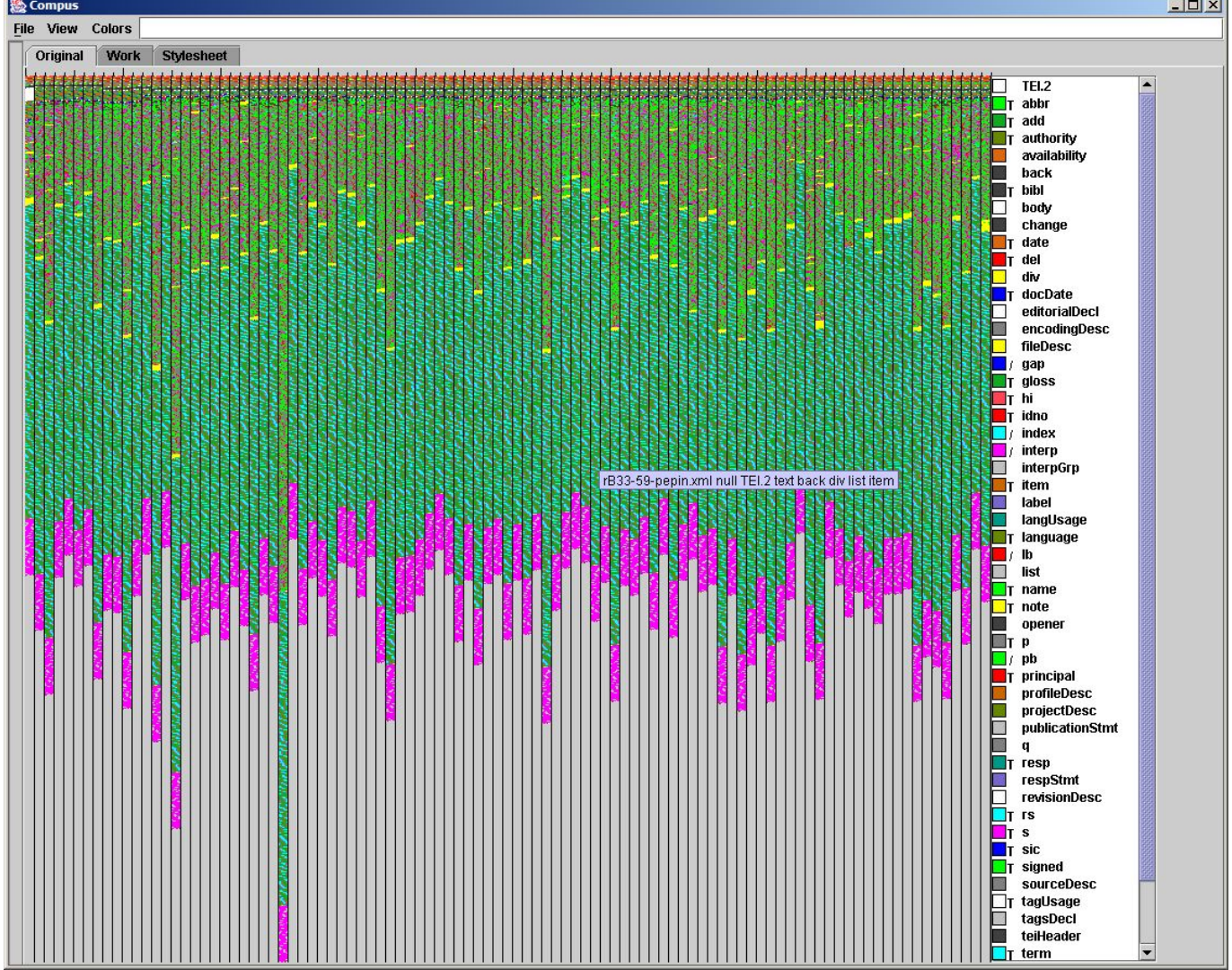
Visualisation du corpus XML

- On transforme le document XML suivant :

```
0         1         2         3         4
01234567890123456789012345678901234567890123456789012345678901234567
<A>abcd<B>efgh</B><C>ijkl<D>mnop</D></C>qrst</A>
```

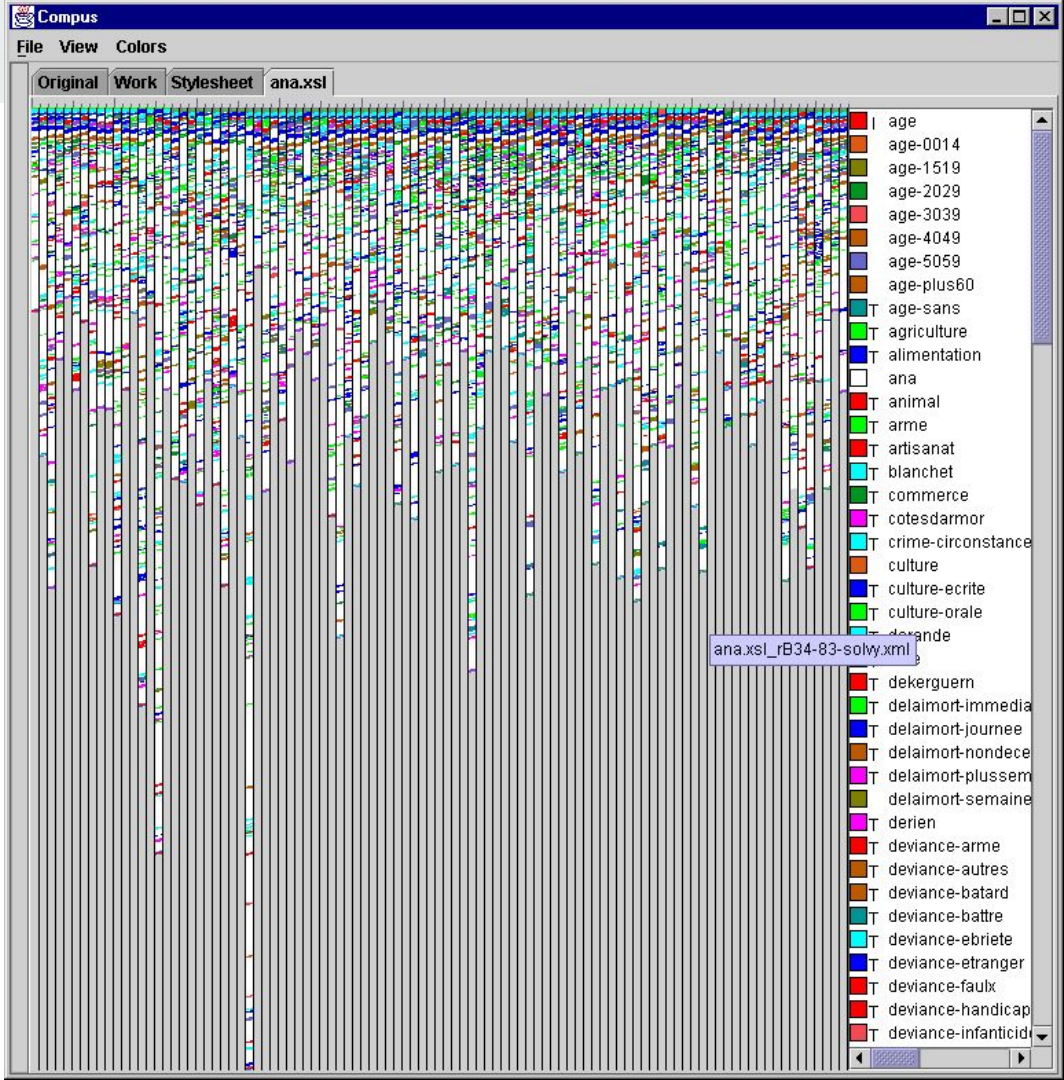
- En une suite d'intervalles:
  A=[0,48[, B=[7,18[, C=[18,40[, D=[25,36[
- On donne une couleur à chaque élément
- On affiche chaque document dans une colonne
- On passe à la ligne quand c'est nécessaire
- Jean-Daniel Fekete and Nicole Dufournaud,
  Compus: Visualization and Analysis of Structured Documents For Understanding Social Life in the 16th Century (Conférence), Proceedings of Digital Libraries (DL00). ACM. June 2000.




Compus
File View Colors
Original Work Stylesheet
rB33-59-pepin.xml null TEI.2 text back div list item
TEI.2
abbr
add
authority
availability
back
bibl
body
change
date
del
div
docDate
editorialDecl
encodingDesc
fileDesc
gap
gloss
hi
idno
index
interp
interpGrp
item
label
langUsage
language
lb
list
name
note
opener
p
pb
principal
profileDesc
projectDesc
publicationStmt
q
resp
respStmt
revisionDesc
rs
s
sic
signed
sourceDesc
tagUsage
tagsDecl
teiHeader
term

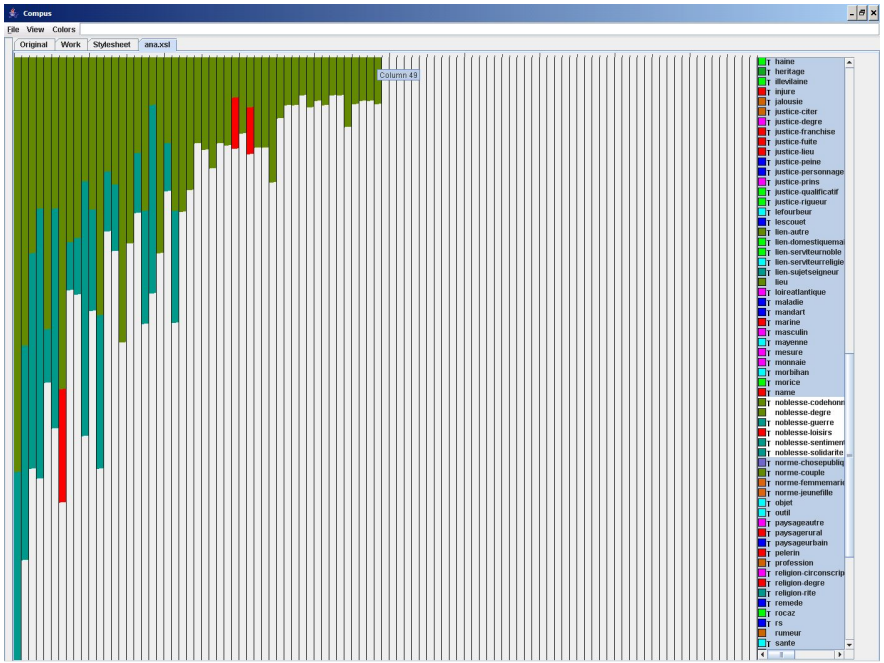
Transformation

Au lieu de visualiser les éléments, on visualise les attributs analytiques associés



Henri IV a-t-il favorisé les nobles bretons ?

- On ne visualise que les segments portant sur des nobles
- On trie selon la taille de ces segments
- Et on regarde le nombre de documents ayant trait à des nobles : 49 / 100
- Conclusion : non, Henry IV n'a pas vraiment pardonné plus aux nobles qu'aux autres
- CQFD



Exploration ?

- On a fait tout ce travail de transcriptions exhaustives et d'encodage TEI XML avec annotations analytiques pour ça ?
  - Non en fait
- Mais le plus coûteux est la transcription
- Alors pourquoi ne pas ajouter 20% d'efforts supplémentaires pour l'encodage ?
- Et permettre d'autres utilisations
- Ainsi que la **transparence** dans l'analyse ?
- http://nicole.dufournaud.org/remission/

Et pouvoir explorer :

- Où sont les enfants de moins de 14 ans ?
- Quels sont les rôles des femmes décrites dans ces lettres de rémission ?

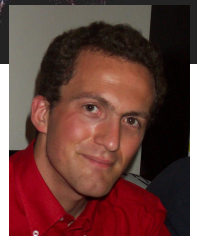
Situation typique :

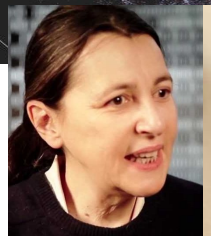
- Dépouiller des sources avec quelques questions
- Exploration pour trouver d'autres questions et d'autres réponses

Cartolabe

Grand Debat

Philippe Caillou Jean-Daniel Fekete Michèle Sebag Jonas Renault Anne-Catherine Letournel

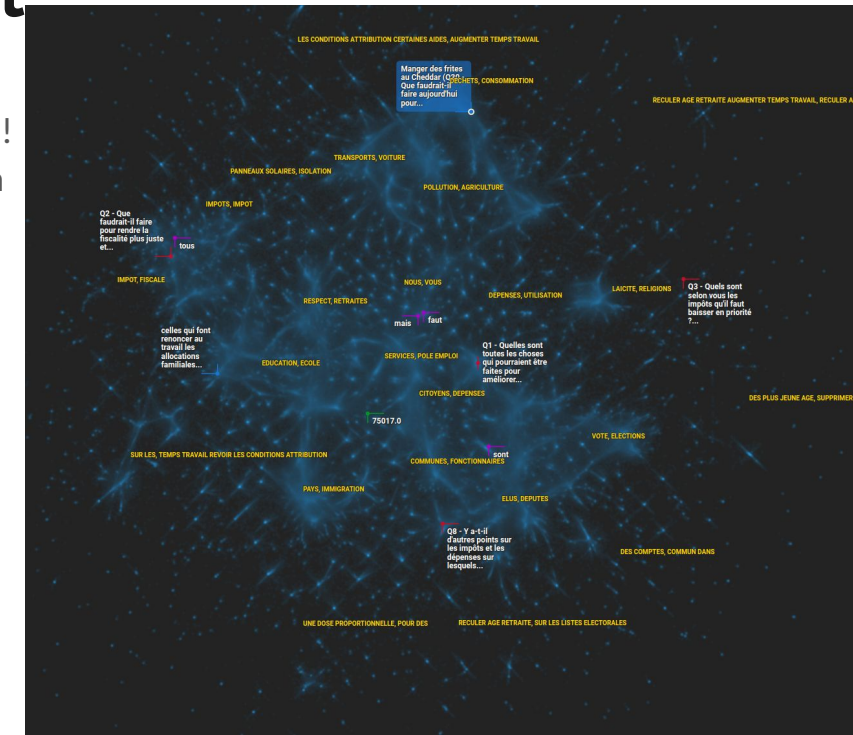





LABORATOIRE DE RECHERCHE EN INFORMATIQUE
université PARIS-SACLAY
cnrs dépasser les frontières
Inria

Cartolabe et le grand débat

- Les moteurs de recherche permettent de «chercher » !
- Les cartes permettent de représenter l'ensemble d'un corpus et de naviguer
    - Chaque point est un document
    - Un document doit être placé près de documents similaires,
    - et moins près de documents moins similaires



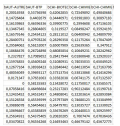

Le grand
débat national
Accueil
Réunions locales
Contributions
Ressources
À propos
Inscription
Connexion
Le grand
débat national
À l'initiative du Président de la République, le Gouvernement engage un Grand Débat National permettant à toutes et tous de débattre de questions essentielles pour les Français.

87 questions réparties en 4 thèmes
Au 2 mars:
200k contributeurs
2M de réponses avec au moins 3 mots

Grand Débat Pipeline

4 CSVs with 200k people answering

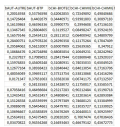
Text Processing



2M propositions
87 questions
1k villes
250k (1-5)grams

X 2M propositions

Projection: LSA



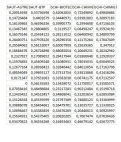
2M propositions
87 questions
1k villes
250k (1-5)grams

X 400 dims

Scoring: Count

Scores (nombre de propositions)

Neighbors: KNN (L2 Dist)



Neighbors
(Direct/
Biased for high scores/
Indirect via the articles)

2D Projection: UMAP

7/28/84/252
Clusters

Cluster names:
Discriminant Words

7/28/84/252
Clusters names

Clustering: K-Means

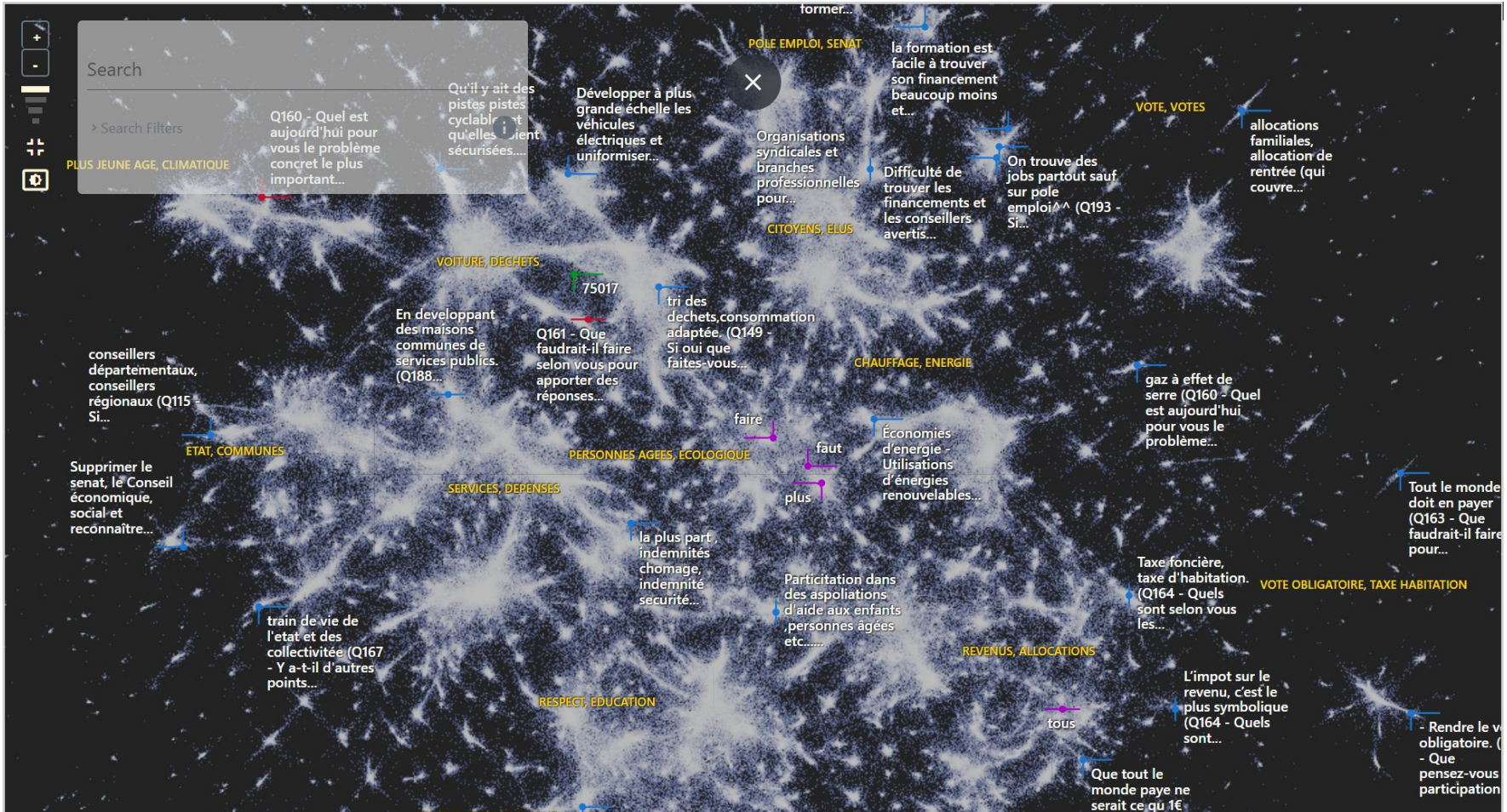
2M propositions
87 questions
1k villes
250k (1-5)grams

X 2D

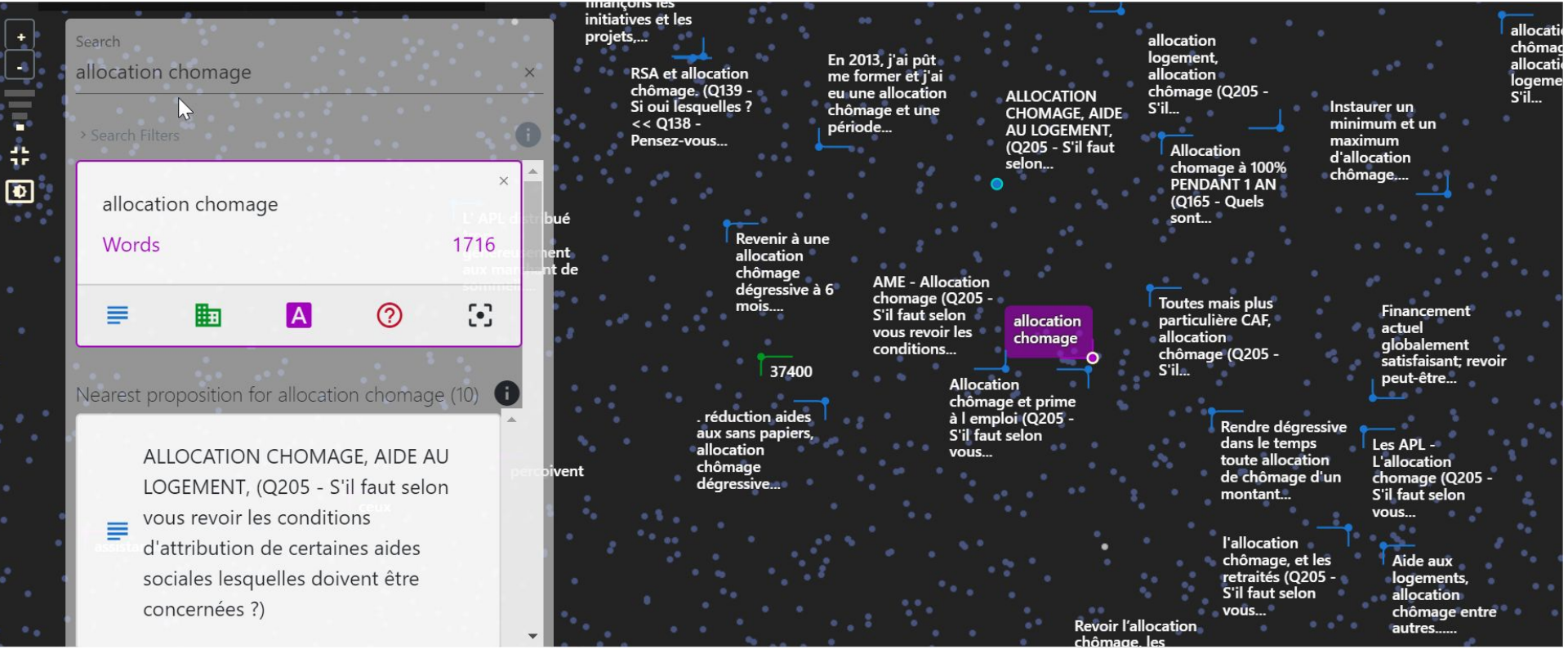
Export

JSON
Cartolabe
File

S'utilise comme une carte



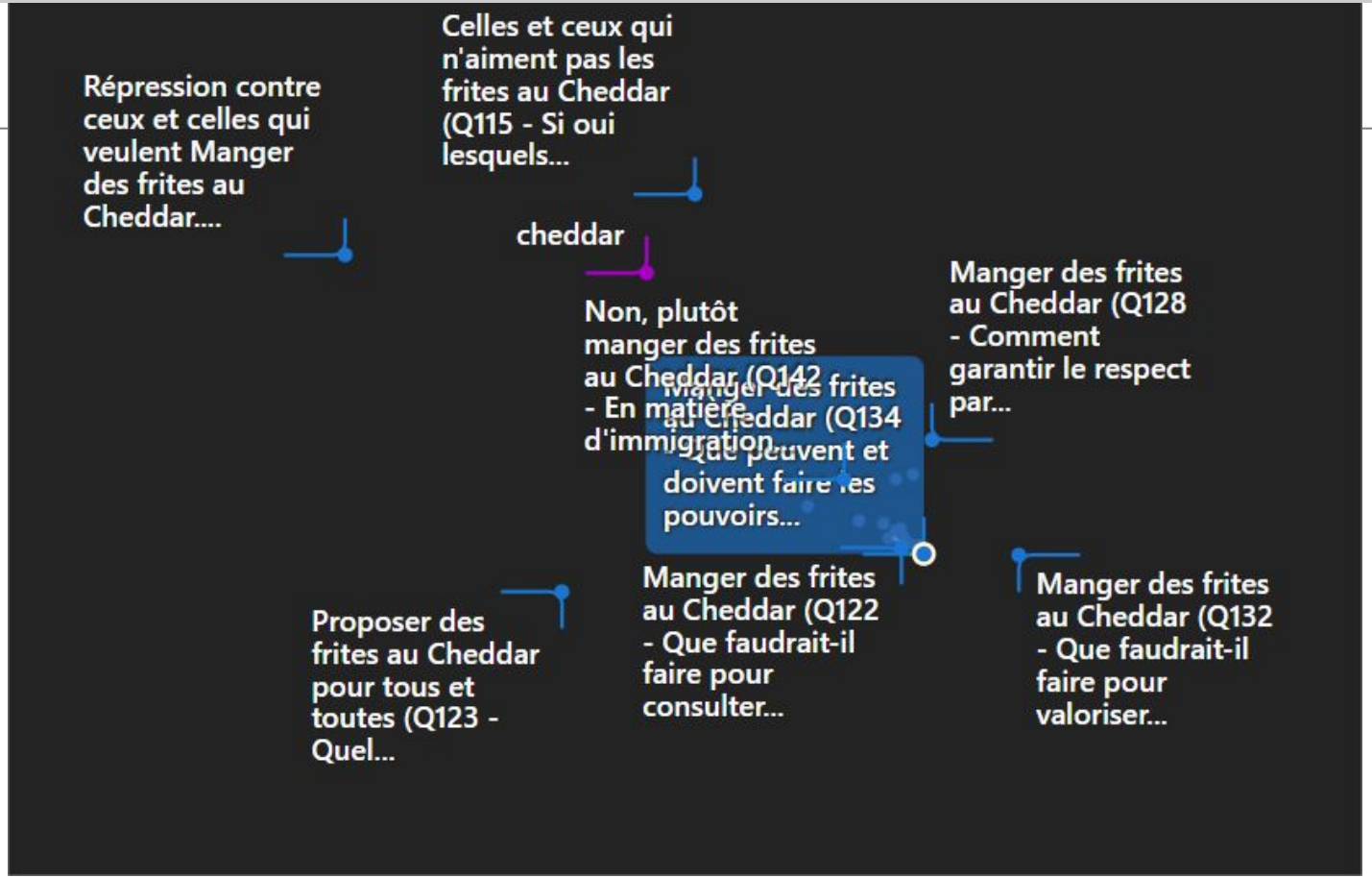
Recherche autour d’un mot-clé/ville…



Les attentes pour le grand débat : exhaustif et non biaisé?

- Toutes les propositions sont présentes sans filtres
- Tous les mots sont utilisés, seuls quelques stopwords sont retirés (est, ai, et, …)
- Les thèmes (clusters) sont générés automatiquement
- *Mais toute représentation est “biaisée” par ses choix de procédure et hyperparamètres*
- Seules des procédures sociales peuvent y remédier
  - Vues multiples, vérifications contre les biais, etc.

Les frites au Cheddar sont-elles bien prises en compte ?



Cartolabe et les frites au Cheddar

Travail avec les garants du gouvernement :

- Analyses **exhaustives** et **non biaisées**
- Possibilité **d'explorer** dans la carte des 4 millions de contributions / propositions
- Utilisé pour valider les analyses produites par les "entreprises" :
  - Les gens veulent moins d'impôts, plus de protection, et de meilleurs services de proximité !
  - Pas très surprenant …
- Comment valider les analyses sans explorer ?
- Comment trouver des propositions intéressantes autour d'un thème ?
  - Pas utilisé par les groupes politiques jusqu'à aujourd'hui hélas

Synthèse

Dans un projet SHS lié à des données :

- Avoir une hypothèse initiale à confirmer n'empêche pas d'explorer les données récoltées
  - Et permet souvent de trouver des questions inattendues et intéressantes
- Parfois, on a des données mais pas d'hypothèse, l'exploration permet d'en trouver
  - C'est une situation qui devient plus fréquente
  - On peut donc partir de donnée pour y chercher des questions ou des surprises
- Dans tous les cas, lorsqu'on a des données, il est utile de commencer par les explorer
  - Mais attention au statut des résultats de l'exploration …

Ne pas explorer n'empêche pas de regarder

Dès que vous avez collecté des données, avant tout calcul statistique, regardez vos données. Examiner les données, ce n'est pas "fourrer son nez dedans". Ce n'est pas une opportunité pour supprimer des données ou changer des valeurs pour favoriser vos hypothèses. Cependant, si vous évaluez vos hypothèses sans examiner vos données, vous risquez de publier des inepties.

L'inspection graphique des données permet de détecter des erreurs d'intégrité sérieuses. La raison en est simple : les graphiques ouvrent le champ de vision, les statistiques le rétrécissent.

Leland Wilkinson, "Statistical Methods in Psychology Journals: Guidelines and Explanations", American Psychologist, Aug. 1999 https://www.apa.org/pubs/journals/releases/amp-54-8-594.pdf

Épistémologies

L'étude critique des sciences et de la connaissance scientifique

Quatre grandes familles

- les mathématiques
- les sciences naturelles
- les sciences humaines et sociales

Quel est le rôle de l'exploration dans ces familles, en particulier en SHS ?

Les mathématiques (logique, statistiques, etc.)

Statut de la vérité et de la preuve :

- Une clause est vraie ou fausse !
  - Le plus souvent
- Dans un cadre axiomatique donné.

C'est la seule épistémologie qui peut démontrer le vrai !

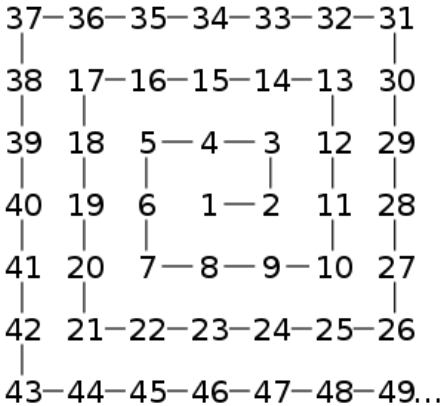
Peut-on néanmoins utiliser l'exploration ?

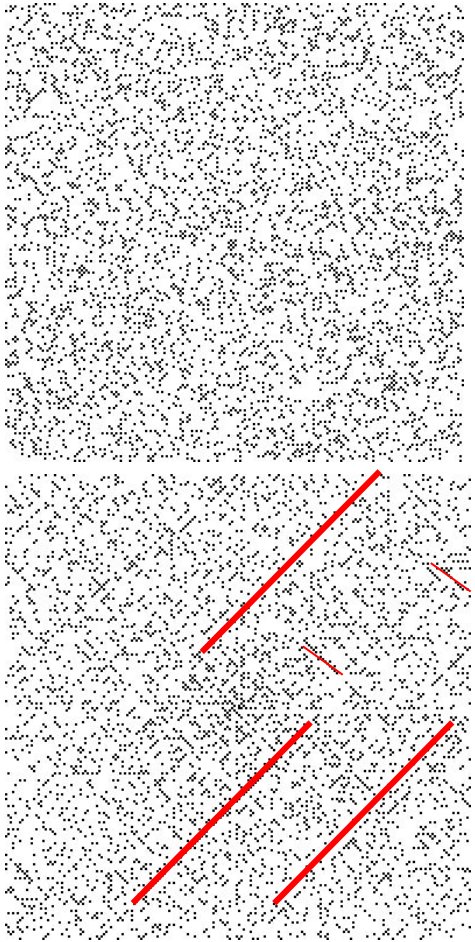
Syllogisme

Tous les humains qui font des SHS
sont intelligents
Tu fais des SHS
Donc, tu n'es pas humain
[sur une porte du MIT, USA]

Mathématiques et exploration?

- Les nombres premiers sont-ils distribués aléatoirement ?
- Spiral d'Ulam
  - 1 au centre
  - Puis on tourne : 2 à droite, 3 au dessus à droite, 4 au dessus, etc.
  - On regarde et on s'attend à du hasard
  - Main on voit des motifs inattendus
  - Pourquoi ?
- Les mathématiciens sont aussi inspirés par des constatations empiriques





Les sciences naturelles

Travail magistral de Karl Popper (1902-1994) sur :

- le problème de l'induction
  - On ne peut pas **induire** de théorie à partir de faits
  - Le soleil s'est levé tous les matins => il se lèvera donc tous les matins
  - C'est une hypothèse non justifiable formellement, mais elle peut être plausible
  - Mais une hypothèse plus forte est encore moins justifiable (par ex. sauf le mardi 13 octobre)
- le problème de la démarcation
  - Une théorie scientifique doit être **réfutable**
  - Une hypothèse peut être **corroborée** par des tests destinés à la mettre en échec (la **falsifier**)
  - Elle ne devient pas **vraie** pour autant, mais reste **corroborée** jusqu'à preuve du contraire (épreuve du temps)
  - Les critères de réfutation sont à la charge de celui qui propose la théorie
  - D'où un besoin de transparence

Les sciences humaines et sociales

Pas de travaux aussi magistraux que Popper (que je connaisse), mais beaucoup de travaux quand même (Bachelard, Durkeim, Kuhn, etc.)

Pas de vérité non plus, mais deux différences par rapport aux sciences naturelles :

- Complexité (au sens de la décomposition en systèmes plus simples) :
  - Impossibilité d'isoler les phénomènes observés pour simplifier les études
  - Réactions non déterministes (dans la même situation, les humains ou les groupes réagissent différemment)
- Boucle fermée : les humains et groupes sociaux réagissent à leur observation
  - Impossibilité d'empêcher un humain qui se sait observer d'en tenir compte
  - Les observations modifient les comportements et les résultats d'expérimentations

Où est l'exploration dans tout ça ?

Ce que l'on sait :

- Pas d'induction
- Pas de vérité
- La charge de la méthode de réfutation vient de l'émetteur de l'hypothèse

Que qu'on voudrait :

- Faciliter la corroboration
- Trouver des hypothèse plausibles plus facilement ou rapidement
- Falsifier des théories plus facilement ou rapidement

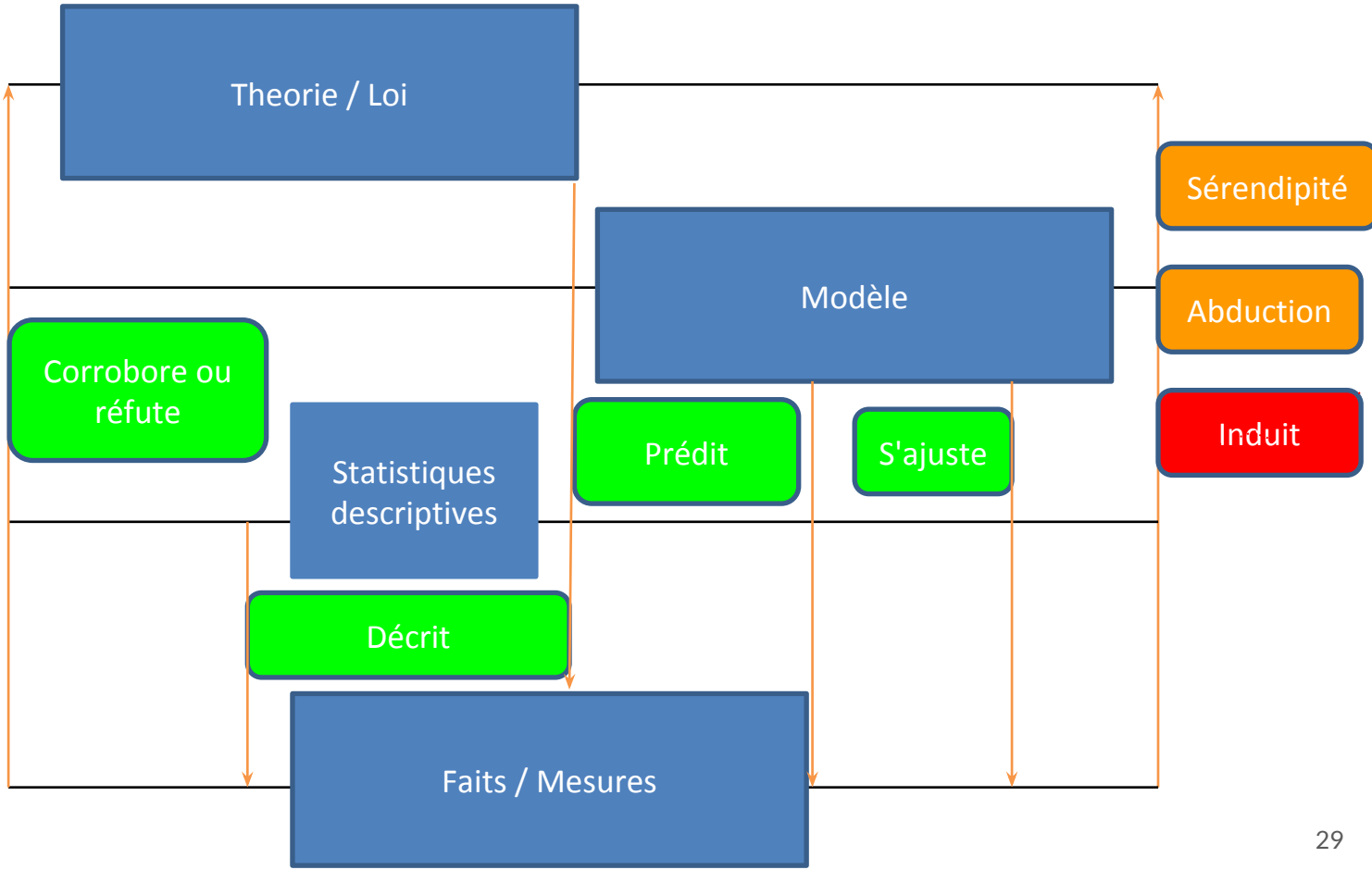
L'exploration permet :

- De trouver des hypothèses
- De les falsifier plus facilement et rapidement
- Par de les prouver
  - que peut-on appeler preuve ?
  - À suivre...

Quatre voies possibles

- Déduction / Confirmation
- Induction / Exploration des théories possibles à partir de causes observées
- Abduction / Exploration des causes possibles à partir d'effets observés
- Sérendipité ou *découverte fortuite* / Exploration par chance et attention
    - « faculté de discerner l'intérêt, la portée d'une découverte inattendue lors d'une recherche » [OQF]
- « Pour résumer,
    - la **déduction**, qui repose sur des causes et des effets certains, aboutit à des énoncés *certains* ;
    - l'**induction**, qui propose des causes certaines à des effets probables, aboutit à des énoncés *probables* ; et
    - l'**abduction**, qui recherche des causes probables à des effets certains, aboutit à des énoncés *plausibles*. » — Nicolas Chevassus-au-Louis, Théories du complot

Science 1.0


Theorie / Loi
Modèle
Faits / Mesures
Corrobore ou réfute
Statistiques descriptives
Décrit
Prédit
S'ajuste
Sérendipité
Abduction
Induit

Méthodes

Myriades de méthodes en sciences mais ...

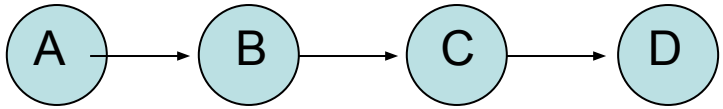
Distinguer clairement les approches :

- Exploratoires
- Confirmatives

La nature de la "preuve" apportée par ces approches est très différente.

Ex: Noeuds-Liens ou Matrice d'Adjacence ?

Graphe orienté



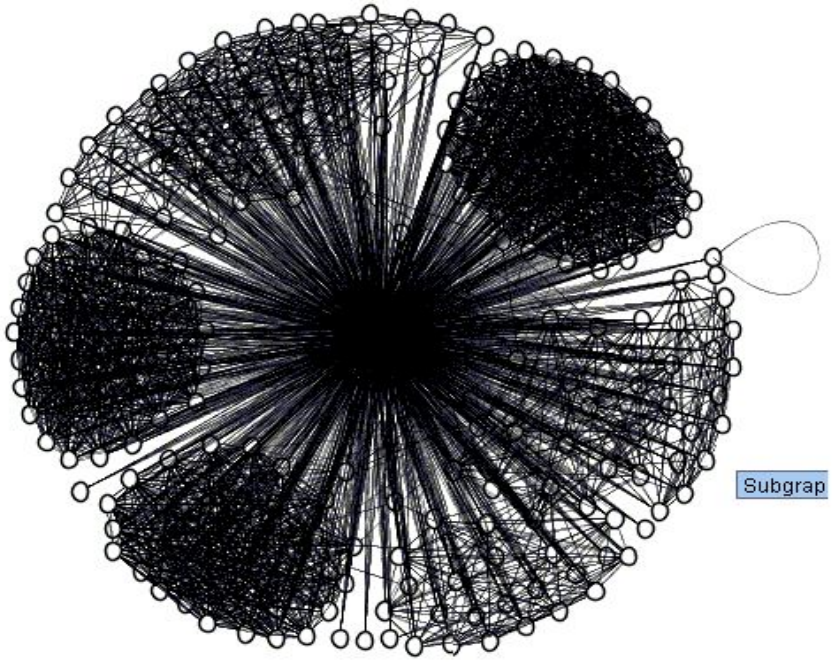
Matrice d'adjacence:

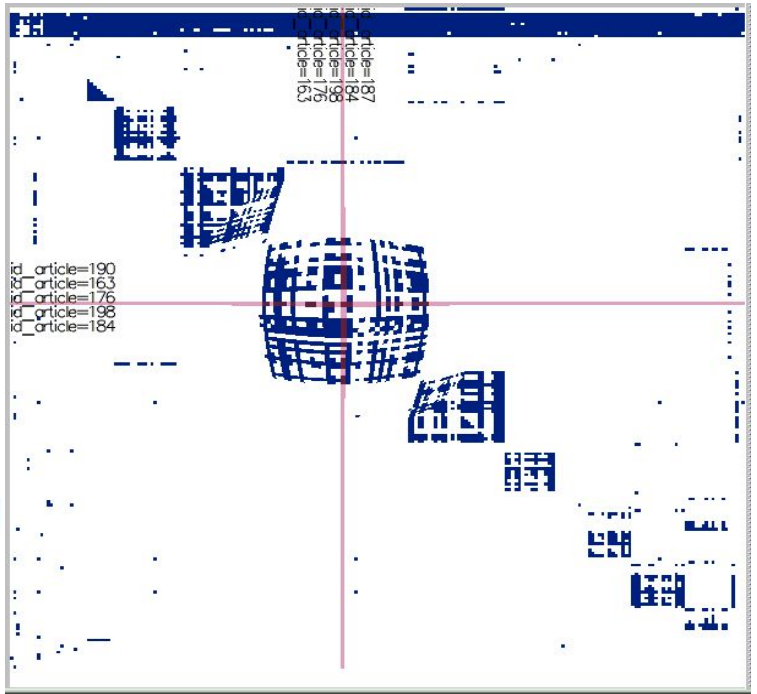
Destination

Source

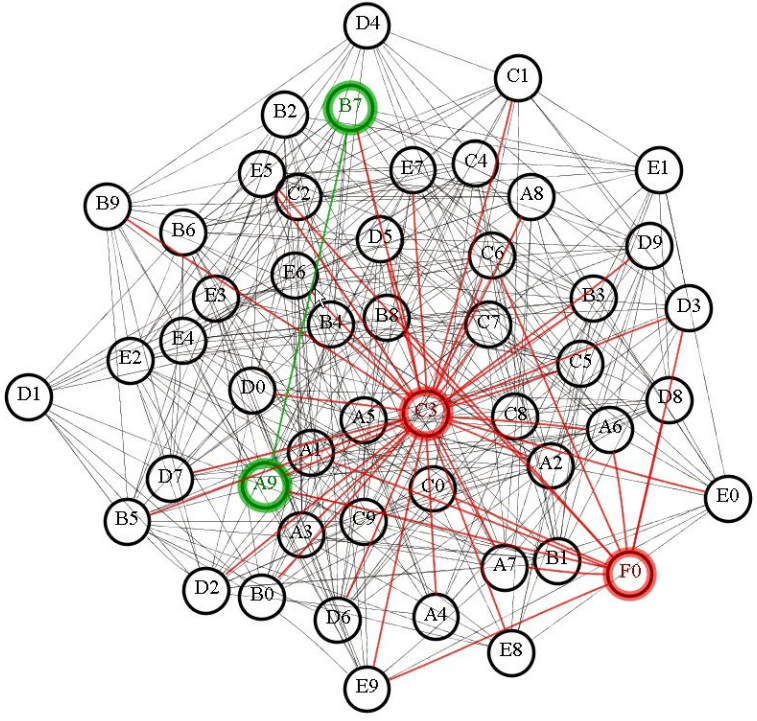
	A	B	C	D
A	0	1	0	0
B	0	0	1	0
C	0	0	0	1
D	0	0	0	0

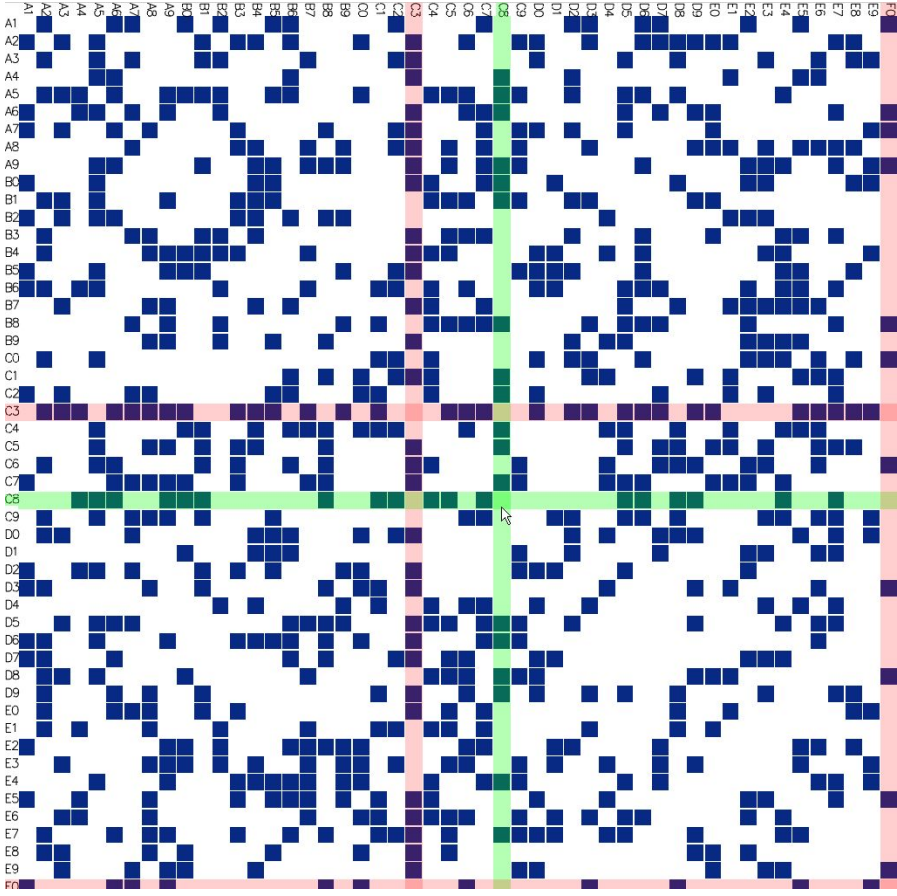
Noeuds-Liens ou Matrice d'Adjacence ?





Expérimentation de lisibilité





Expérimentation : Noeuds-Liens vs Matrice d'Adjacence

Les taches :

- Vue d'ensemble
  - Estimer le nombre de sommets
  - Estimer le nombre d'arêtes
- Details de le graphe
  - Trouver un sommet lié à ...
  - Trouver le sommet le plus connecté à ...
  - Trouver un voisin commun
  - Trouver un chemin
- Graphes aléatoires (3 tailles et 3 densités)
- 2 representations: NL + Matrice

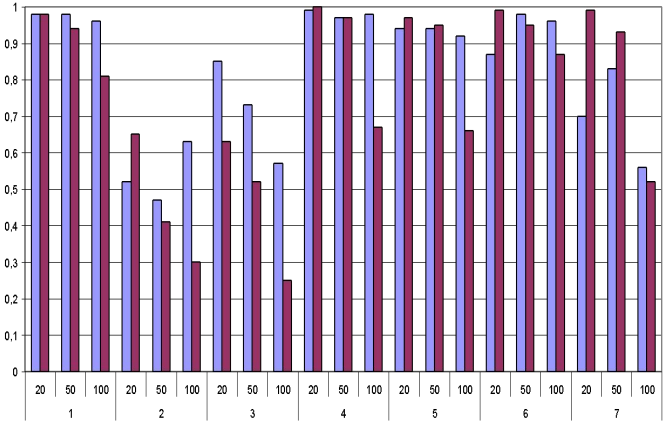
Resultats:

- NL préférable pour des petits graphes peu denses (20 sommets)

Matrices plus lisibles pour les graphs denses et plus grands

References:

Mohammad Ghoniem, Jean-Daniel Fekete and Philippe Castagliola *Readability of Graphs Using Node-Link and Matrix-Based Representations: Controlled Experiment and Statistical Analysis*, Information Visualization Journal, 4(2), Palgrave Macmillan, Summer 2005, pp. 114-135.



Percentage of correct answers for the 7 tasks, 3 densities and 2 representations. NL in purple, Matrix in blue

Résumé de la situation

- On a des données
- On veut suivre une méthode acceptable scientifiquement
- Pour répondre à des questions ou hypothèses
- Et éventuellement trouver des nouvelles questions et hypothèse intéressantes
- Comment doit-on procéder ?

La méthode hypothetico-deductive

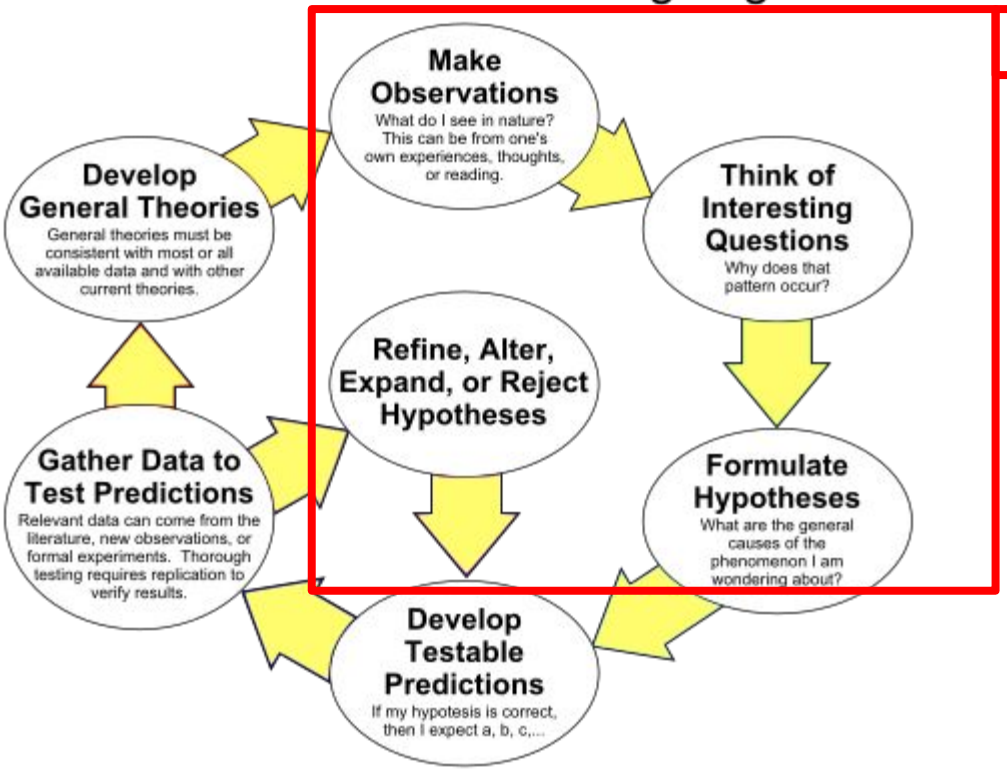
Formuler une hypothèse, en déduire des conséquences observables, permettant d'en déterminer la validité.

Mettre en marche des méthodes confirmatives pour valider. OK

Les méthodes exploratoires, en plus des intuitions, permettent de trouver des hypothèses.

- Les méthodes exploratoires sont-elles suffisantes pour valider ?
  - … ça dépend
- Peut-on toujours trouver des meilleurs méthodes pour valider ?
  - … ça dépend

The Scientific Method as an Ongoing Process



Exploration ?

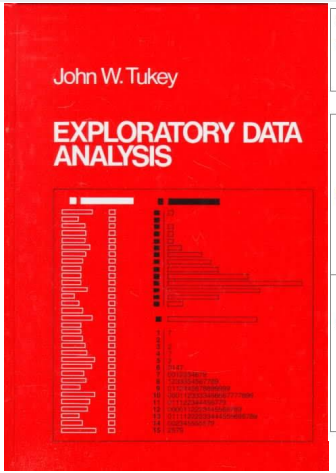
Exploration et visualisation

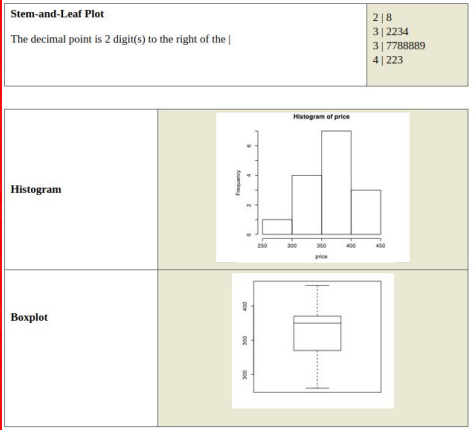
John W. Tukey, Exploratory Data Analysis, 1977

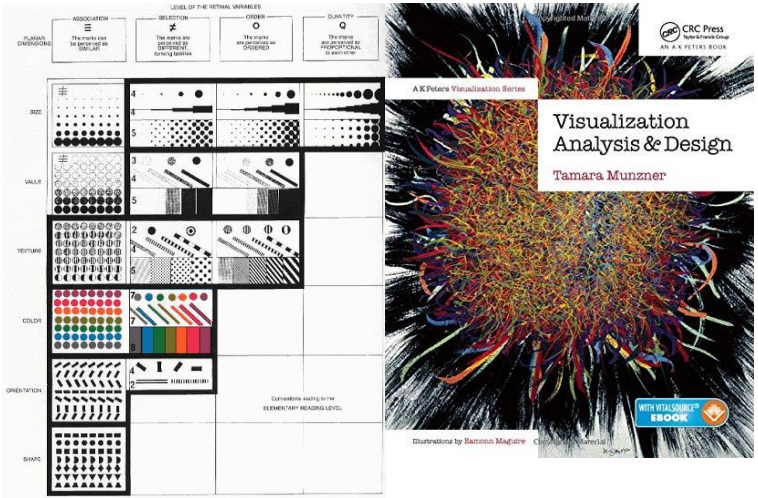
- Présente sa méthode pour explorer les données
- Statistiques descriptives
- Graphiques statistiques (boxplots, histogrammes, etc.)

Trois familles de méthodes deviennent populaires :

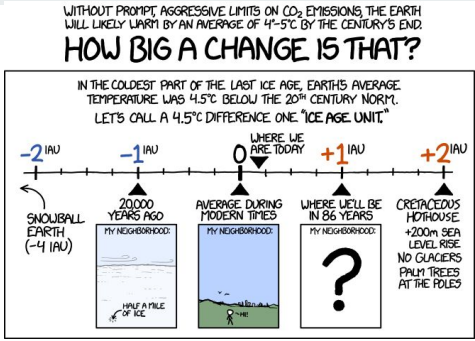
- Statistiques pour aider à l'exploration
- Visualisation pour aider à l'exploration
- Ignorer l'exploration et la critiquer







Deux visualisations - deux rôles



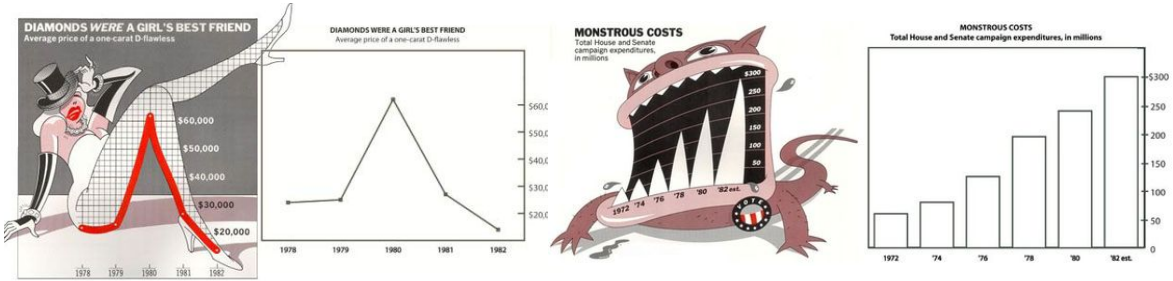
Visualisation pour l'exploration

- Sujet de cette présentation

Visualisation pour la communication

http://aviz.fr/~bbach/datacomics/

- Méthode rhétorique
- Tout ce qui peut aider à communiquer est utile
- Mais doit être compréhensible par le plus grand nombre



Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, Christopher Brooks, Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts. *ACM* CHI, 2010.

Visualisation pour l'exploration

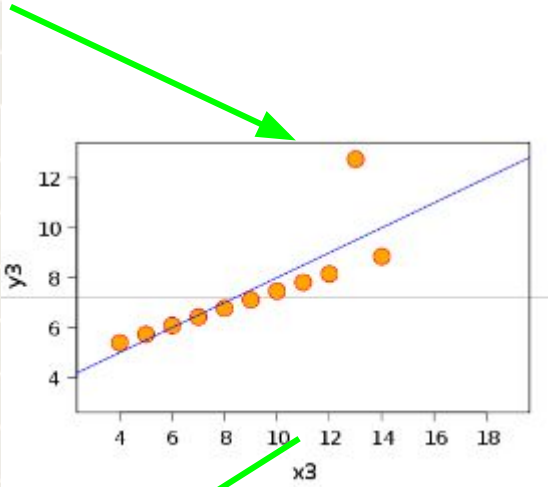
Représentation graphique compacte et interface utilisateur
pour manipuler un grand nombre d'items,
potentiellement extrait d'un jeu de données bien plus grand.
Permet de :
    faire des découvertes, prendre des décisions, ou trouver des explications
Concernant des :
    motifs (tendances, groupes, exceptions, …), groupes d'items ou individus.
[Plaisant, 2001]

Étude de représentations visuelles (interactives) de données [abstraites] pour assister la cognition humaine [Wikipedia, information visualization, 2018]

Visualisation Literacy

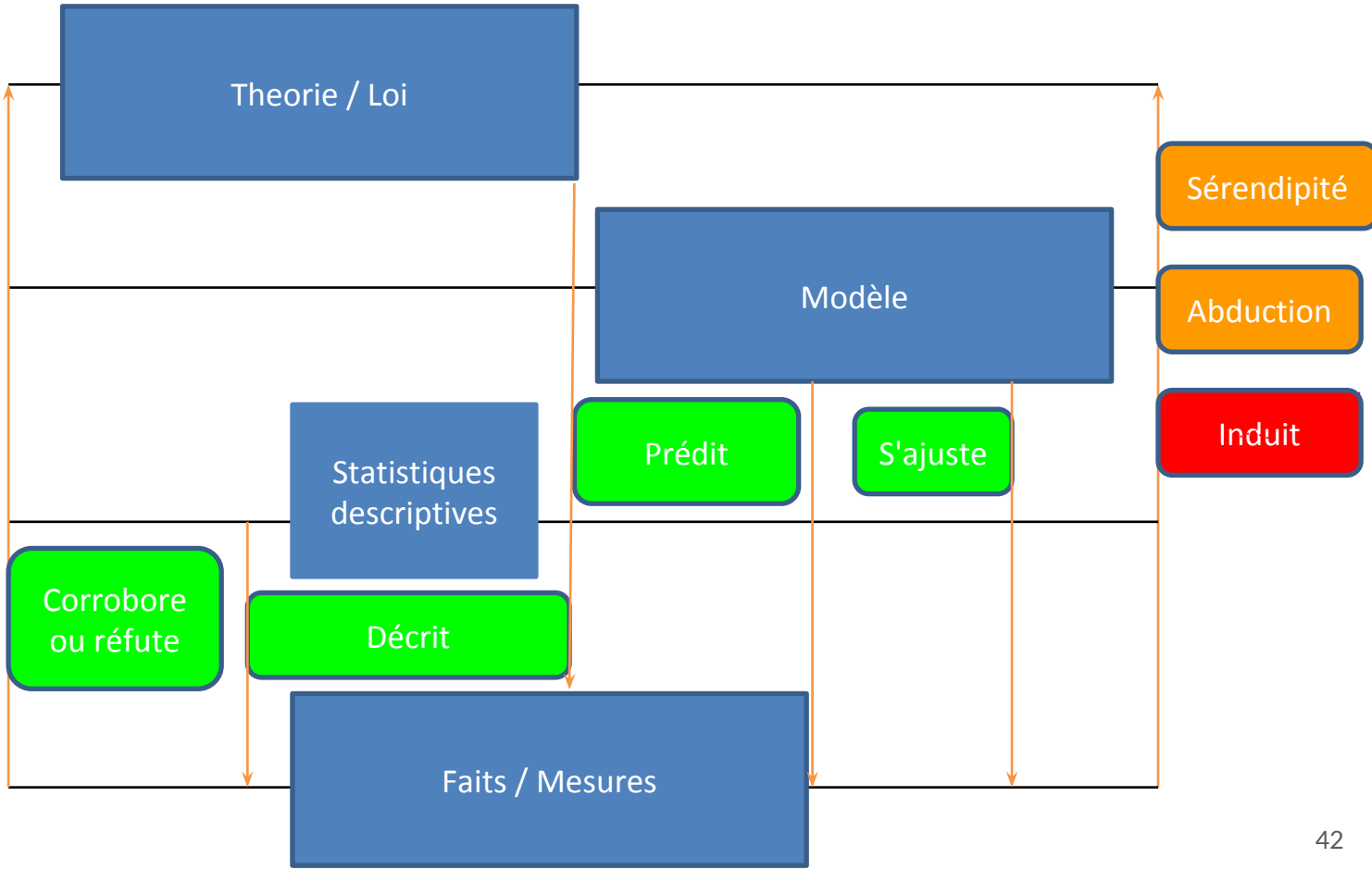
La capacité à utiliser *avec assurance* une visualisation pour traduire des *questions* dans le *domaine des données* en des *requêtes visuelles* dans le *domaine visuel*,
ET
à interpréter des *motifs visuels* dans le *domaine visuel*, comme des *propriétés* dans le *domaine des données*.

III	
x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73



[Boy et al. A Principled Way of Assessing Visualization Literacy, TVCG 2014]

Science 1.0


Theorie / Loi
Modèle
Statistiques descriptives
Faits / Mesures
Sérendipité
Abduction
Induit
Prédit
S'ajuste
Corrobore ou réfute
Décrit

Stats vs. visualisation : le Quartet d'Anscombe

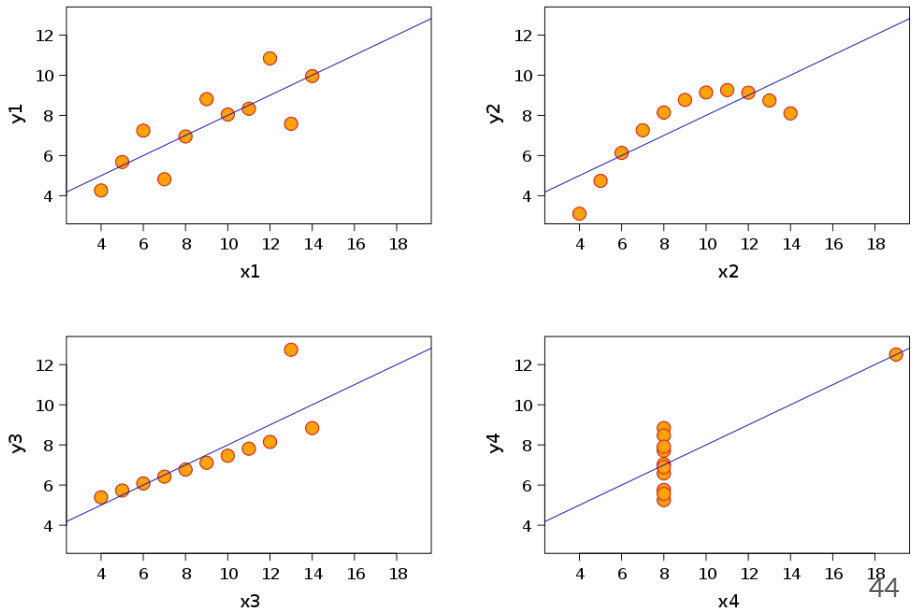
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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 de x	9.0
Variance de x	11.0
Moyenne de y	7.5
Variance de y	4.12
Corrélation entre x et y	0.816
Régression linéaire	$y = 3 + 0.5x$

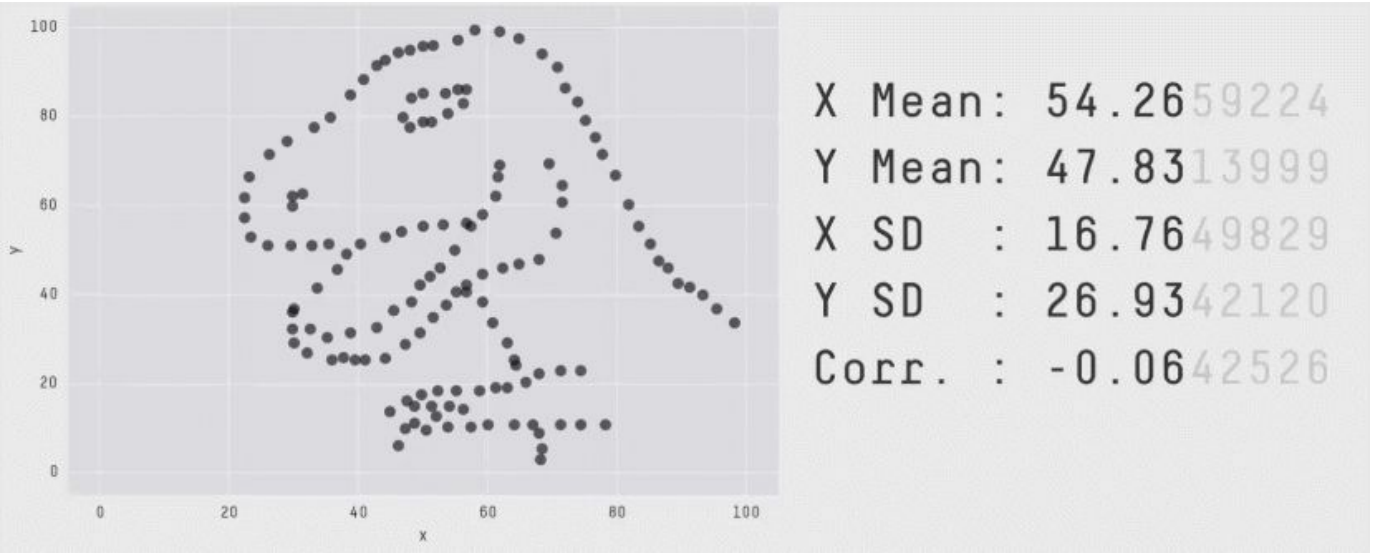
[Source: Anscombe's quartet, Wikipedia]

La visualisation révèle une histoire 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

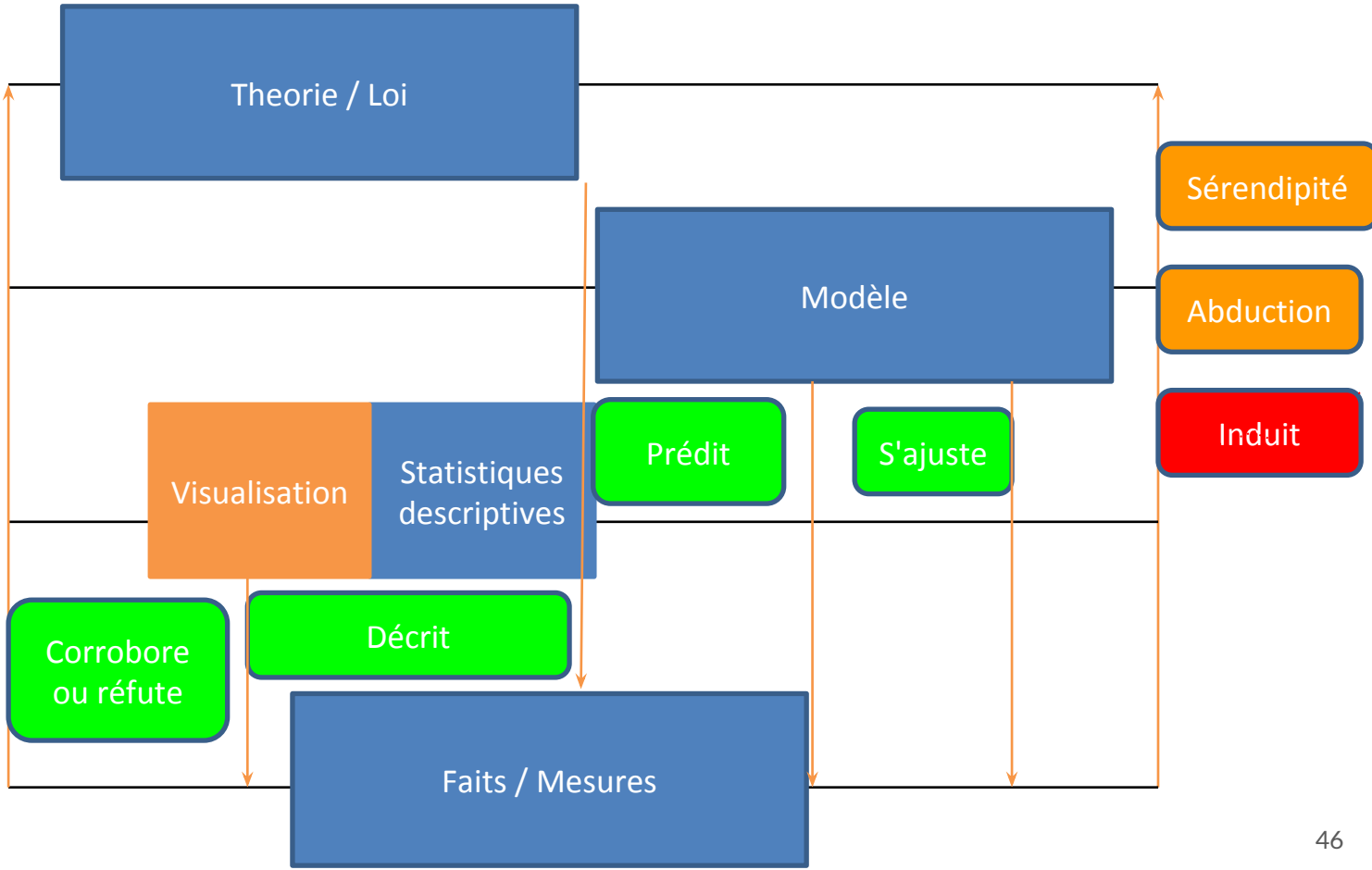


Des stats peuvent avoir n'importe quelle forme



J. Matejka and G. Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM CHI'17
https://www.autodeskresearch.com/publications/samestats

Science 1.0


Theorie / Loi
Modèle
Faits / Mesures
Visualisation
Statistiques descriptives
Prédit
S'ajuste
Décrit
Corrobore ou réfute
Sérendipité
Abduction
Induit

La science 2.0 repose plus sur les données

- Plus de mémoire, plus de données disponibles, plus d'opportunités pour l'exploration
- Plus de données sont collectées à partir de capteurs (pollution, photos, vidéos, audio, etc.)
- Plus de données sont disponibles sur le Web
- Il est possible d'explorer sans hypothèse a priori
  - Comme le fait la métagénomique par exemple
- La science tirée par les données (Data Science) est devenu populaire
- La visualisation permet une exploration très rapide des données
  - Avec une boucle de rétroaction rapide qui permet d'infirmer rapidement des hypothèses
- Cette méthode est-elle valide ?

La visualisation pour l'exploration

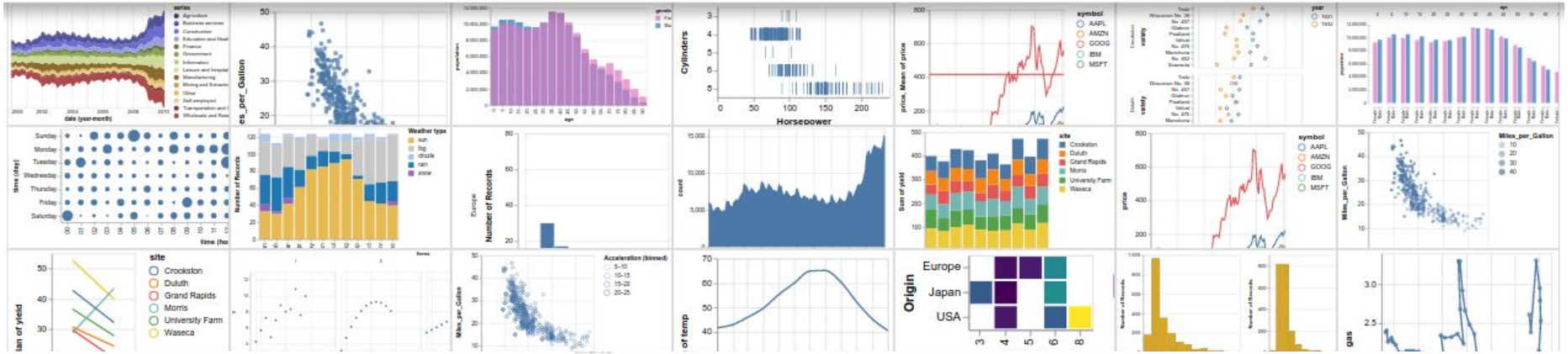
Des progrès très important dans les 15 dernières années

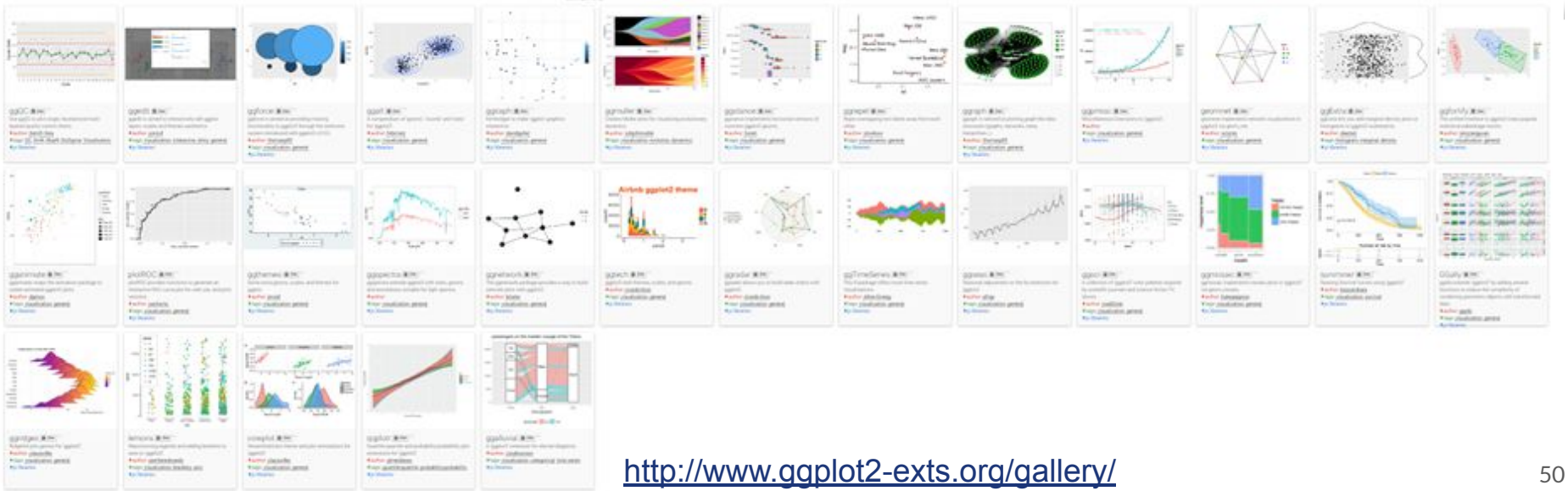
- Un modèle conceptuel avancé pour construire les visualisations
- Une compréhension en voie d'amélioration de l'interaction
- Des progrès dans la compréhension de la perception visuelle
- Une bonne compréhension des couleurs, de la perception à la cognition
- Une compréhension de la perception d'ensemble (corrélation par ex.) et de ses limitations
- Une compréhension des biais de perception (cécité au changement et limitations des animations)
- Une meilleure compréhension des biais cognitifs pour la prise de décision
- Une compréhension plus fine des techniques de visualisation spécifiques (plots, graphes, etc.)

Construire des visualisation rationnellement

- Grammar of Graphics (Wilkinson, 2005) explique comment décrire et construire la plupart des visualisations à partir de quelques constructions primitives et de principes
- La visualisation n'est pas limitée à un ensemble fini de composants prédéfinis
  - On peut décrire et construire de nouveaux composants rationnellement
- gglot2 with R (Wickham, 2010) implémente cette grammaire avec quelques variations
  - Hadley Wickham. A layered grammar of graphics. Journal of Computational and Graphical Statistics, vol. 19, no. 1, pp. 3–28, 2010.
- Vega en JavaScript, l'implémente aussi avec des extensions pour spécifier des interactions
  - A. Satyanarayan, D. Moritz, K. Wongsuphasawat, J. Heer, Vega-Lite: A Grammar of Interactive Graphics, IEEE Trans. Visualization & Comp. Graphics 2017
  - Vega est accessible à partir d'autres langages comme Python (Altair), Juli, R, et tous ceux compilés en JS

https://vega.github.io/vega-lite/examples/





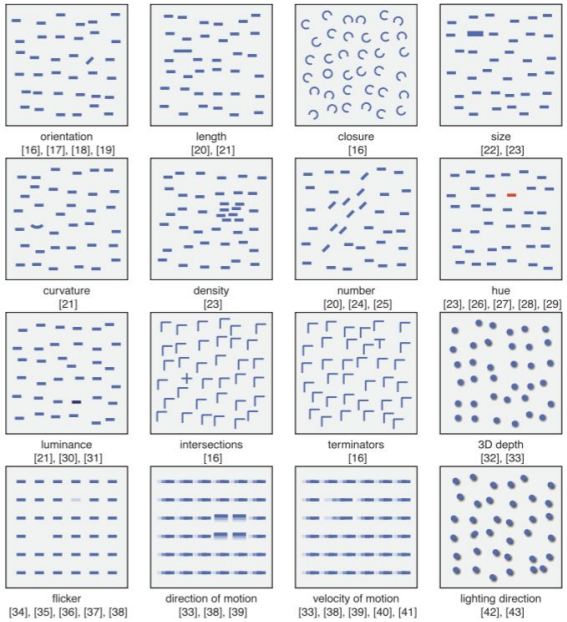
http://www.ggplot2-exts.org/gallery/

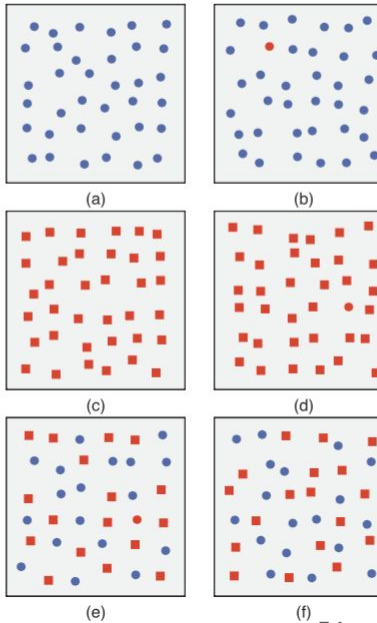
La vision bénéficie de la *perception préattentive*

“tasks that can be performed on large multi-element displays in less than 200-250 milliseconds”: preattentive processing is done quickly, effortlessly and in parallel without any attention being focused on the display.

[Treisman, 1985] A. Treisman, Preattentive Processing in Vision, *Computer Vision, Graphics, and Image Processing*, 31(2):156-177, August 1985.

[Treisman, 1986] A. Treisman, Features and Objects in Visual Processing, *Scientific American*, 255(5):114-125, 1986.

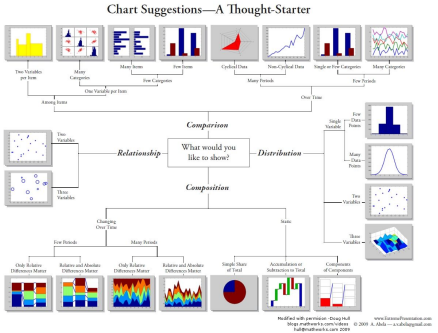


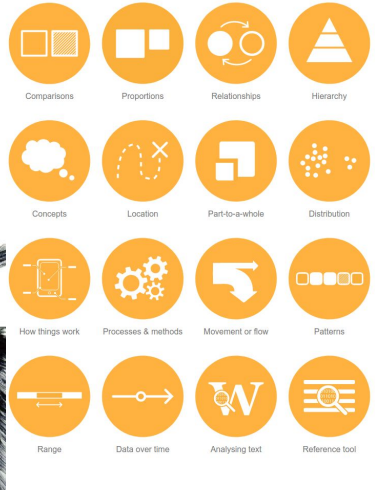


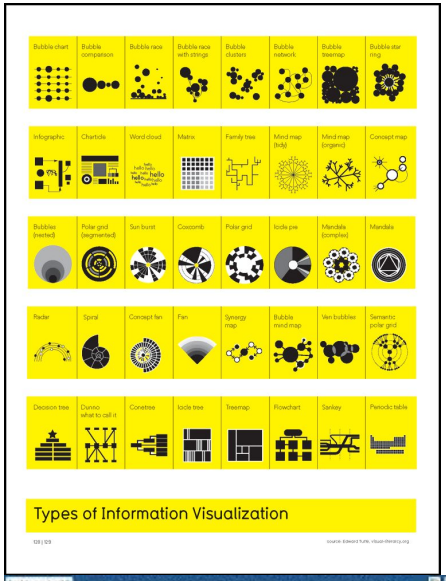
Utilisation raisonnée

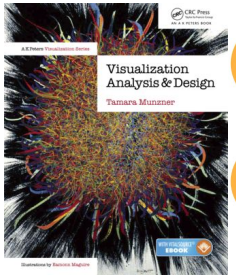
Beaucoup d'informations disponibles : https://github.com/widged/data-for-good/wiki/Visualisation-::-Choosing-a-chart

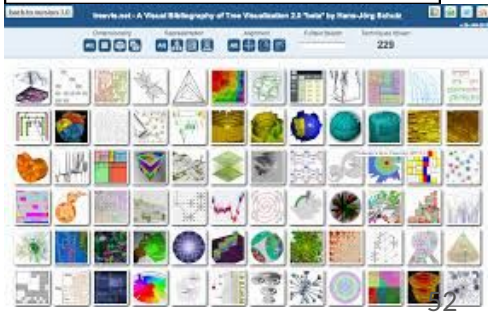
- A Tour through the Visualization Zoo https://queue.acm.org/detail.cfm?id=1805128
- Sites web résumant les techniques
  - treevis.net
  - textvis.lnu.se
  - www.timeviz.net





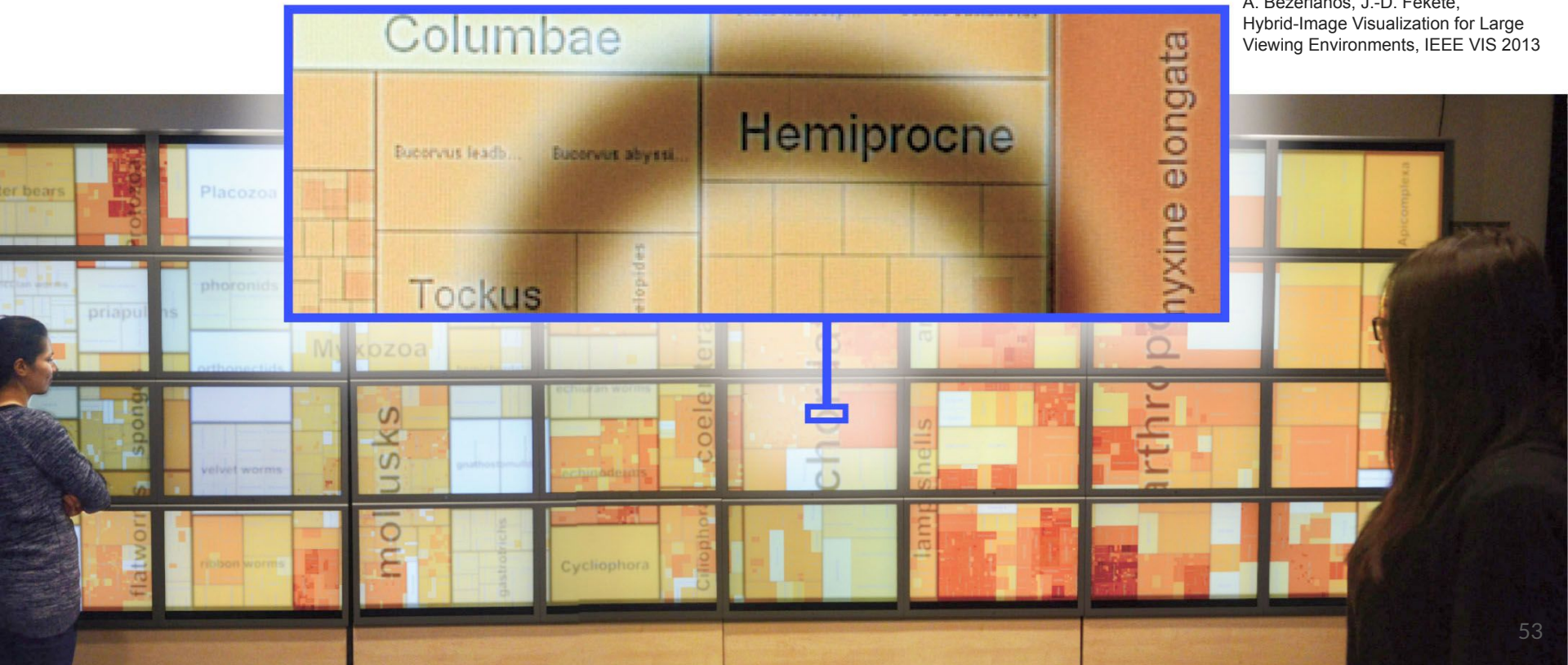






De nouvelles surfaces sont disponibles

P. Isenberg, P. Dragicevic, W. Willett, A. Bezerianos, J.-D. Fekete, Hybrid-Image Visualization for Large Viewing Environments, IEEE VIS 2013



La méthode hypothetico-deductive

Formuler une hypothèse, en déduire des conséquences observables, permettant d'en déterminer la validité.

Mettre en marche des méthodes confirmatives pour valider. OK

Les méthodes exploratoires permettent de trouver des hypothèses.

- La visualisation ou les méthodes exploratoires sont-elles suffisantes pour valider ?
  - … ça dépend
- Peut-on toujours trouver des meilleurs méthodes pour valider ?
  - … ça dépend
- En SHS, on utilise généralement les statistiques pour valider des hypothèses
  - Parfois des maths directes en économie, mais leur statut est "plausible", mais "certain"

Généralisation et statistiques

En tout temps, en tous lieux, …

- Toutes les questions ne sont pas des généralisations
  - ex : Qui est le personnage le plus influent d'un groupe ?
  - Qui est l'espion ?
  - Pas d'inférences sur la population générale
- Pour les généralisations
  - Il faut contrôler les inférences
  - Éviter des erreurs de type 1 et 2

- Décision entre deux hypothèses : H0, H1
- H0 est vraie a priori (hypothèse nulle)
- H1 est alternative
- Erreur Type 1 : accepter H1 si H0 est vraie
- Erreur Type 2 : accepter H0 si H1 est vraie

	Hypothèse H_0 vraie	Hypothèse H_1 vraie
Hypothèse H_0 acceptée	Bonne décision (1-α)	Risque β
Hypothèse H_1 acceptée	Risque α	Bonne décision (1-β)

Test statistique visuel

"Le protocole est simple :

- Générer n−1 leurres (données vérifiant l'hypothèse nulle).
- Faire des visualisation des leurres et positionner au hasard la visualisation des vraies données
- Les montrer à des observateurs impartiaux.

Peuvent-ils détecter les vraies données ?
Pratiquement, on choisit n = 19, pour obtenir une chance sur 20 d'avoir la bonne réponse au hasard, soit p = 0.05, la borne traditionnelle de la représentativité statistique pour un test.
Comparer 20 visualisation est faisable pour un humain."

Peut-être, mais combien de fois ?

Graphical Inference for Infovis

Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja

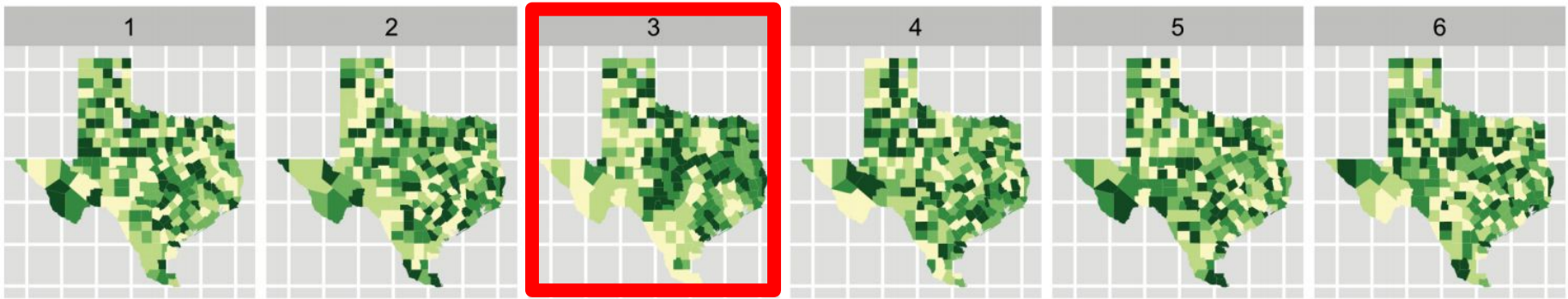


Fig. 1. One of these plots doesn't belong. These six plots show choropleth maps of cancer deaths in Texas, where darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence? If so, you've provided formal statistical evidence that deaths from cancer have spatial dependence. See Section 8 for

Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. Graphical inference for infovis. IEEE Transactions on Visualization and Computer Graphics 16, 6 (November 2010), 973-979.

Le problème des comparaisons multiples

"As more comparisons are made, the probability rapidly increases of encountering interesting-looking (e.g., data trend, unexpected distribution, etc.), but still random events. Treating such inevitable patterns as insights is a false discovery (Type I error) and the analyst `loses' if they act on such false insights." [Zgraggen, Zhao, Zeleznik and Kraska, Investigating the Effect of the Multiple Comparison Problem in Visual Analysis in CHI 2018.]

- On a besoin d'un ensemble de validation pour tester si l'hypothèse trouvée dans un échantillon est valide sur tout le jeu de données
- On doit aussi contrôler les biais ajoutées aux données lorsqu'on a fait des multiples filtrages et agrégations durant le processus d'exploration

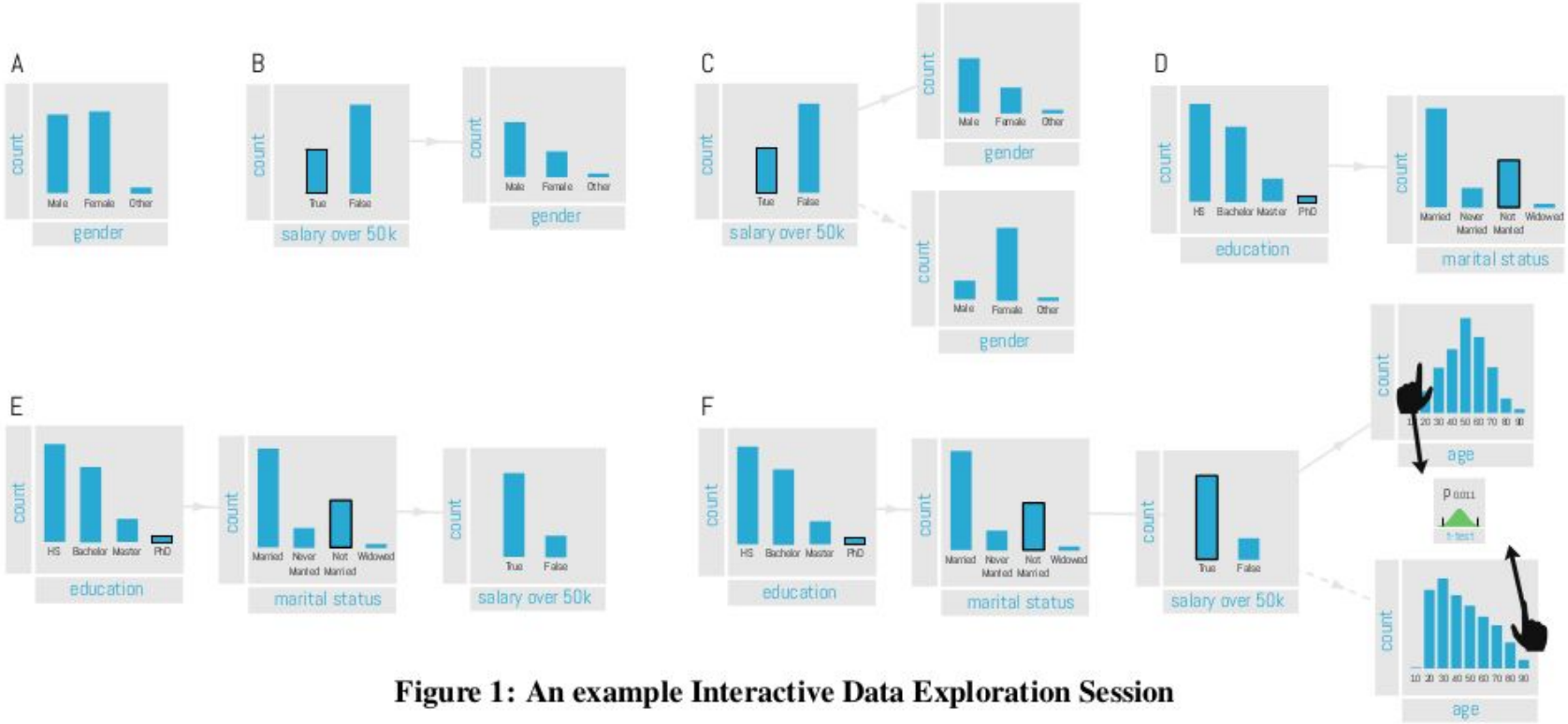
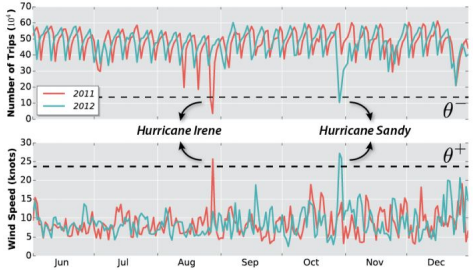


Figure 1: An example Interactive Data Exploration Session

Le problème des comparaisons multiples

- Le problème ne vient pas de la visualisation
- Quelques solutions existent, mais il est nécessaire que le chercheur soit conscient du problème
  - pour faire des tests a posteriori
- La question provoque une controverse :
  - **l'expertise peut-elle éviter les erreurs statistiques durant l'exploration ?**
- Réponse :
  - ça dépend ...



F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. **Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets**.SIGMOD '16

Figure 1: Variation of the number of taxi trips in NYC and its relationship with wind speed.

Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska. **Controlling False Discoveries During Interactive Data Exploration**.SIGMOD '17

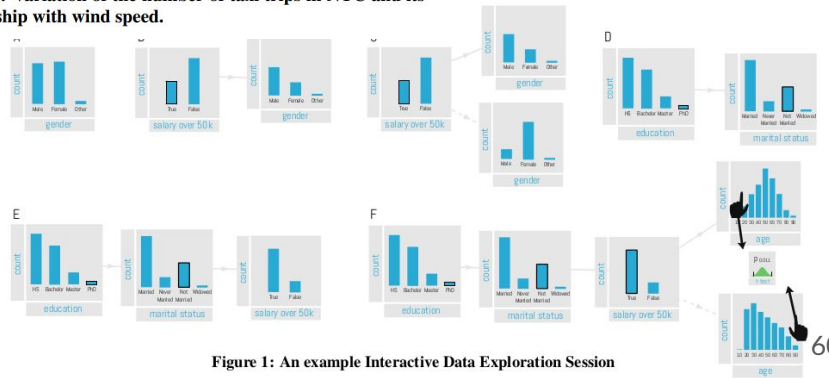


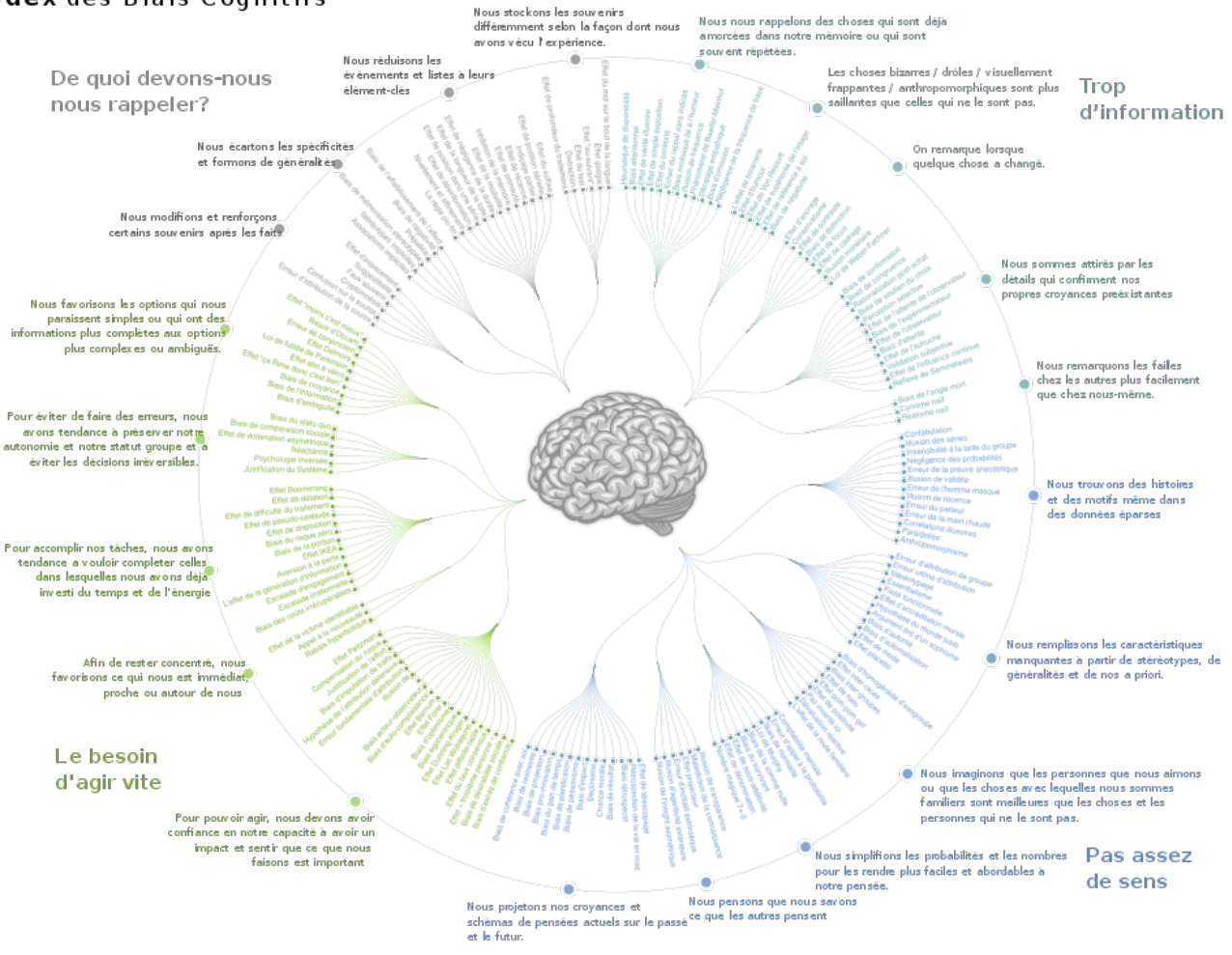
Figure 1: An example Interactive Data Exploration Session

Les biais cognitifs

Déviation systématique de la pensée logique et rationnelle par rapport à la réalité

- Terme introduit au début des années 70 par Daniel Kahneman et Amos Tversky pour expliquer des décisions irrationnelles dans le domaine économique
  - Prix nobel d'économie 2002 (D. Kahneman seul, A. Tversky était décédé)
- **Si on ne les connaît pas, on ne peut pas les combattre**
- Ce sont généralement des problèmes de **prise de décision**
  - Daniel Kahneman, Thinking, Fast and Slow, Allen Lane, 2011
  - Traduction en français : Daniel Kahneman, Système 1 : Système 2 : les deux vitesses de la pensée, Paris, Flammarion, 2012

Codex des Biais Cognitifs



DESIGNHACKS.CO · CATEGORIZATION BY BUSTER BENSON · ALGORITHMIC DESIGN BY JOHN MANOOGIAN III (JM3) · DATA BY WIKIPEDI

Les biais cognitifs dangereux pour l'exploration

- Ancrage mental
  - Influence laissée par la première impression
- Biais de confirmation
  - Tendance à valider ses opinions auprès des instances qui les confirment, et à rejeter d'emblée les instances qui les réfutent.
- Biais de représentativité
  - Considérer un ou certains éléments comme représentatifs d'une population.
- Effet Stroop
  - Incapacité d'ignorer une information non pertinente.
- Illusion de savoir
  - Dans une situation en apparence identique à une situation commune, réagir de manière habituelle, sans éprouver le besoin de rechercher les informations complémentaires qui auraient mis en évidence une différence par rapport à la situation habituelle.

Se protéger des biais cognitifs ?

- S'informer sur leur existence
- Valider ses raisonnements avec des collègues critiques
- Mettre en place des méthodes pour les éviter
    - Par ex. pour mener des enquêtes (contre-espionnage, sécurité, journalisme), Matrice de décision, https://en.wikipedia.org/wiki/Decision_matrix
- Même pour Daniel Kahneman, certains biais sont très difficiles à éviter
- Mais ils posent des problèmes aussi dans la vie de tous les jours et dans la vie scientifique
    - Pas seulement pour l'exploration

Quelques controverses

- L'exploration n'est pas reproductible ou réplicable
  - Est-ce vraiment un problème ? Einstein est-il replicable ?
  - Non, mais ses découvertes sont corroborées avec des expérimentations réplicables !
  - Seules les hypothèses doivent être corroborées de manière reproductibles
- La visualisation est utilisée par des humains qui peuvent voir des motifs là où il n'y en a pas
  - Oui, ce sont des erreurs de Type I en statistiques.
  - La validation devrait les détecter
- La visualisation est utilisée par des humains qui peuvent louper des patterns importants
  - Oui, ce sont des erreurs de Type II en statistiques.
  - On n'est jamais sûr d'avoir trouvé tout ce qui est intéressant. OK

Résumé des messages

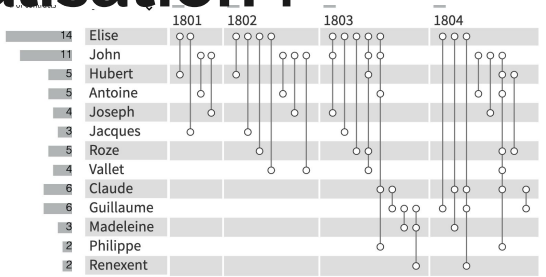
- L'exploration, en particulier avec de la visualisation, est une méthode efficace et valide pour trouver des hypothèses et parfois les tester
- Les méthodes exploratoires ont leurs propres artefacts
  - Il faut développer un certain niveau d'éducation pour les éviter
- La visualisation permet d'explorer très rapidement
  - Il faut prendre en compte le problème de comparaisons multiples
- Il faut aussi tester si les motifs trouvés ne sont pas fallacieux
  - Garder des données de validation pour faire plus de tests
  - Utiliser des tests plus puissant
  - Parfois, l'expertise humaine peut valider les motifs, mais attention...
- Décrivez la procédure que vous avez utilisé pour explorer,
  Pas seulement la validation de l'hypothèse a posteriori
  - Cela aidera vos collègues à trouver plus d'hypothèses

Que font les chercheurs en visualisation ?

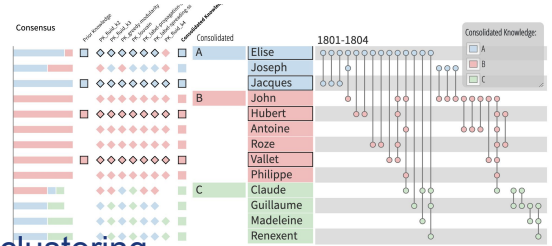
Entre autres, trouver des représentations nouvelles qui facilitent l'exploration :

- Pour des problèmes de plus en plus spécifiques (mais pas toujours)
  - ex : hypergraphes dynamiques, Bitcoin
- Combinent exploration et validation
  - ex : PK-Clustering
- Passage à l'échelle
  - ex : Cartolabe

Contactez-moi si vous êtes dans ces cas.



https://aviz.fr/paohvis



https://aviz.fr/pkclustering

Résumé des messages

- L'exploration, en particulier avec de la visualisation, est une méthode efficace et valide pour trouver des hypothèses et parfois les tester
- Les méthodes exploratoires ont leurs propres artefacts
  - Il faut développer un certain niveau d'éducation pour les éviter
- La visualisation permet d'explorer très rapidement
  - Il faut prendre en compte le problème de comparaisons multiples
- Il faut aussi tester si les motifs trouvés ne sont pas fallacieux
  - Garder des données de validation pour faire plus de tests
  - Utiliser des tests plus puissant
  - Parfois, l'expertise humaine peut valider les motifs, mais attention...
- Décrivez la procédure que vous avez utilisé pour explorer,
  Pas seulement la validation de l'hypothèse a posteriori
  - Cela aidera vos collègues à trouver plus d'hypothèses

Bibliographie

- Dufournaud N., Les femmes en Bretagne au XVIe siècle ..., (mémoire de DEA), Université de Nantes, 2000.
- Caillou P., Renault J., Fekete J.-D., Letournel A.-C., Sebag M., Cartolabe: A Web-Based Scalable Visualization of Large Document Collections, 2020, arXiv preprint, 2003.00975
- Wilkinson L., "Statistical Methods in Psychology Journals: Guidelines and Explanations", American Psychologist, 1999
- Popper, K. (2002) [1959]. The Logic of Scientific Discovery. Abingdon-on-Thames: Routledge
- Boy, J. Rensink, R. A., Bertini, E. and Fekete, J.-D., A Principled Way of Assessing Visualization Literacy, IEEE TVCG 2014
- Matejka, J.Fitzmaurice, G. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM CHI'17
- Satyanarayan, A.,Moritz, D., Wongsuphasawat, K., Heer, J., Vega-Lite: A Grammar of Interactive Graphics, IEEE TVCG 2017
- Bachelard, G. 1984. The new scientific spirit. Boston: Beacon Press.
- Durkheim, E. 1964. What is a Social Fact? In: The rules of sociological method. New York: Free Press of Glencoe
- Kuhn, T.S. 1962. The structure of scientific revolutions. Chicago: University of Chicago Press.
- Mohammad Ghoniem, Jean-Daniel Fekete and Philippe Castagliola Readability of Graphs Using Node-Link and Matrix-Based Representations: Controlled Experiment and Statistical Analysis, Information Visualization Journal, 4(2), Palgrave Macmillan, Summer 2005
- John W. Tukey, Exploratory Data Analysis, 1977
- Tamara Munzner, Visualization Analysis and Design, AK Peters Visualization Series, 2014
- Wickham, H. A layered grammar of graphics. Journal of Computational and Graphical Statistics, vol. 19, no. 1
- Treisman A., Preattentive Processing in Vision, Computer Vision, Graphics, and Image Processing, 1985.

Bibliographie (2)

- P. Isenberg, P. Dragicevic, W. Willett, A. Bezerianos, J.-D. Fekete, Hybrid-Image Visualization for Large Viewing Environments, IEEE TVCG 2013
- Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. Graphical inference for infovis. IEEE TVCG, 2010
- Zgraggen, Zhao, Zeleznik and Kraska, Investigating the Effect of the Multiple Comparison Problem in Visual Analysis in ACM CHI 2018
- Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska. Controlling False Discoveries During Interactive Data Exploration.SIGMOD '17
- F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets.SIGMOD '16
- Daniel Kahneman, Thinking, Fast and Slow, Allen Lane, 2011
- Daniel Kahneman, Système 1 : Système 2 : les deux vitesses de la pensée, Paris, Flammarion, 2012
- Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N. and Fekete, J-D., Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. IEEE TVCG, 2020
- Alexis Pister, Paolo Buono, Jean-Daniel Fekete, Catherine Plaisant, Paola Valdivia. Integrating Prior Knowledge in Mixed Initiative Social Network Clustering. IEEE TVCG, 2021