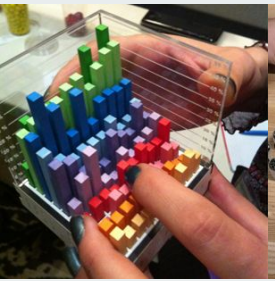
Aviz Visual Analytics Project

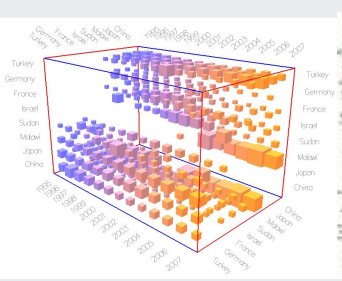
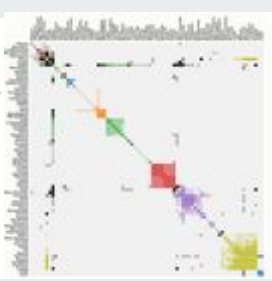
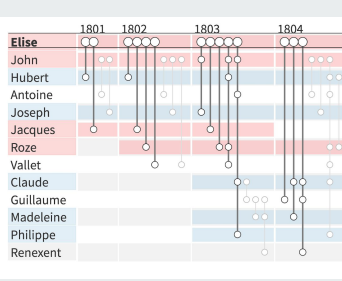
université PARIS-SACLAY

Inria

Visualisation et exploration de grands corpus à partir de projections multidimensionnelles

Jean-Daniel Fekete, Inria & Université Paris-Saclay
https://www.aviz.fr



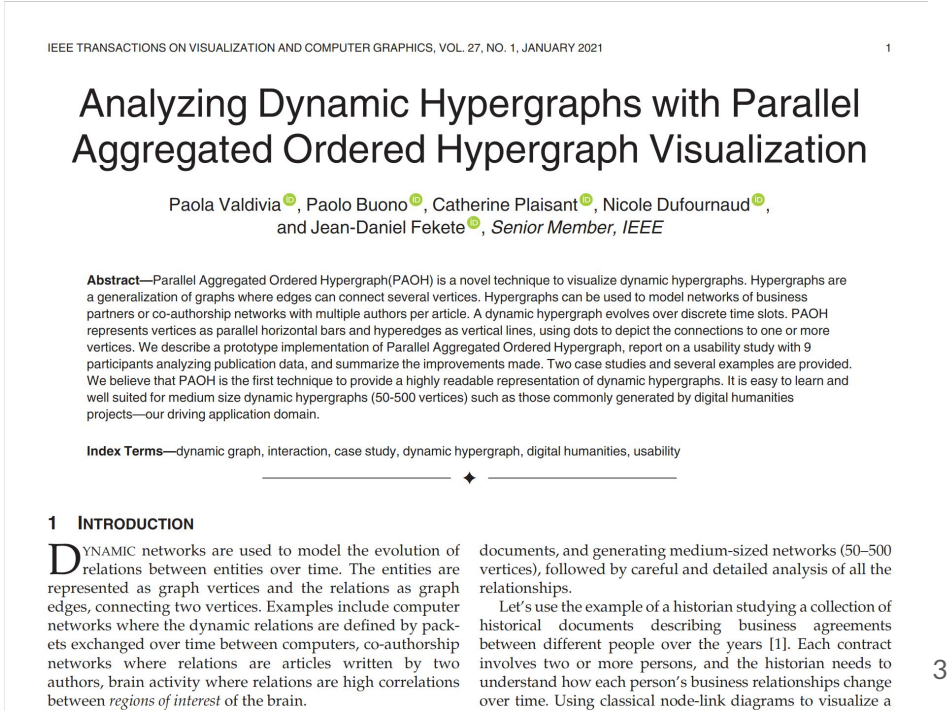
Message

- Les visualisations sont très efficaces pour découvrir et comprendre les données
- Mais elles nécessitent un apprentissage
  - Pas très long
- De nouvelles visualisations arrivent pour voir plus de données et pour les voir mieux
- La crise du COVID-19 a popularisé la visualisation
  - Et montré la nécessité de garder un esprit critique sur les données
- Voici un exemple de nouvelles visualisations qui vont devenir populaire
  - **Les cartes de données**

Au début, il y a des articles

- La principale production de la recherche c'est l'**article**
- Publiés dans une **revue** ou une **conférence**
- Écrites par des **auteurs** qui ont des **affiliations**
    - Université, école d'ingénieur, centre de recherche, société, organisation, etc.

Comment visualiser cette recherche ?



Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization

Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete, *Senior Member, IEEE*

Abstract—Parallel Aggregated Ordered Hypergraph(PAOH) is a novel technique to visualize dynamic hypergraphs. Hypergraphs are a generalization of graphs where edges can connect several vertices. Hypergraphs can be used to model networks of business partners or co-authorship networks with multiple authors per article. A dynamic hypergraph evolves over discrete time slots. PAOH represents vertices as parallel horizontal bars and hyperedges as vertical lines, using dots to depict the connections to one or more vertices. We describe a prototype implementation of Parallel Aggregated Ordered Hypergraph, report on a usability study with 9 participants analyzing publication data, and summarize the improvements made. Two case studies and several examples are provided. We believe that PAOH is the first technique to provide a highly readable representation of dynamic hypergraphs. It is easy to learn and well suited for medium size dynamic hypergraphs (50-500 vertices) such as those commonly generated by digital humanities projects—our driving application domain.

Index Terms—dynamic graph, interaction, case study, dynamic hypergraph, digital humanities, usability

1 Introduction

Dynamic networks are used to model the evolution of relations between entities over time. The entities are represented as graph vertices and the relations as graph edges, connecting two vertices. Examples include computer networks where the dynamic relations are defined by packets exchanged over time between computers, co-authorship networks where relations are articles written by two authors, brain activity where relations are high correlations between *regions of interest* of the brain.

documents, and generating medium-sized networks (50–500 vertices), followed by careful and detailed analysis of all the relationships.

Let's use the example of a historian studying a collection of historical documents describing business agreements between different people over the years [1]. Each contract involves two or more persons, and the historian needs to understand how each person's business relationships change over time. Using classical node-link diagrams to visualize a

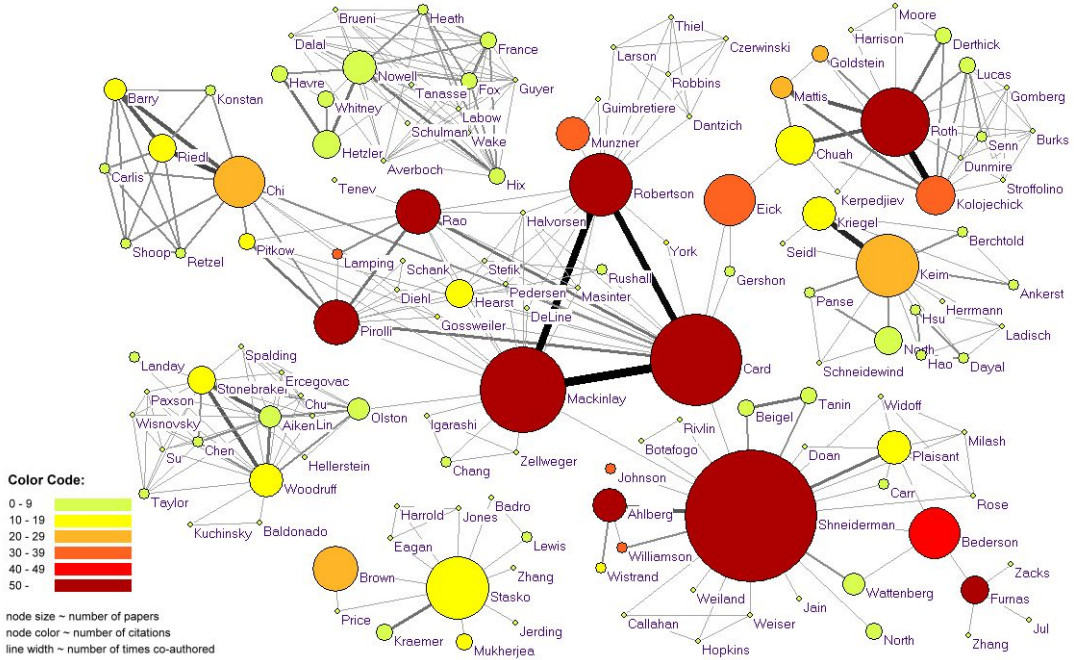
La méthode standard : le réseau

K. Börner et al. 2004

Deux auteurs sont liés s'ils ont écrit un article ensemble

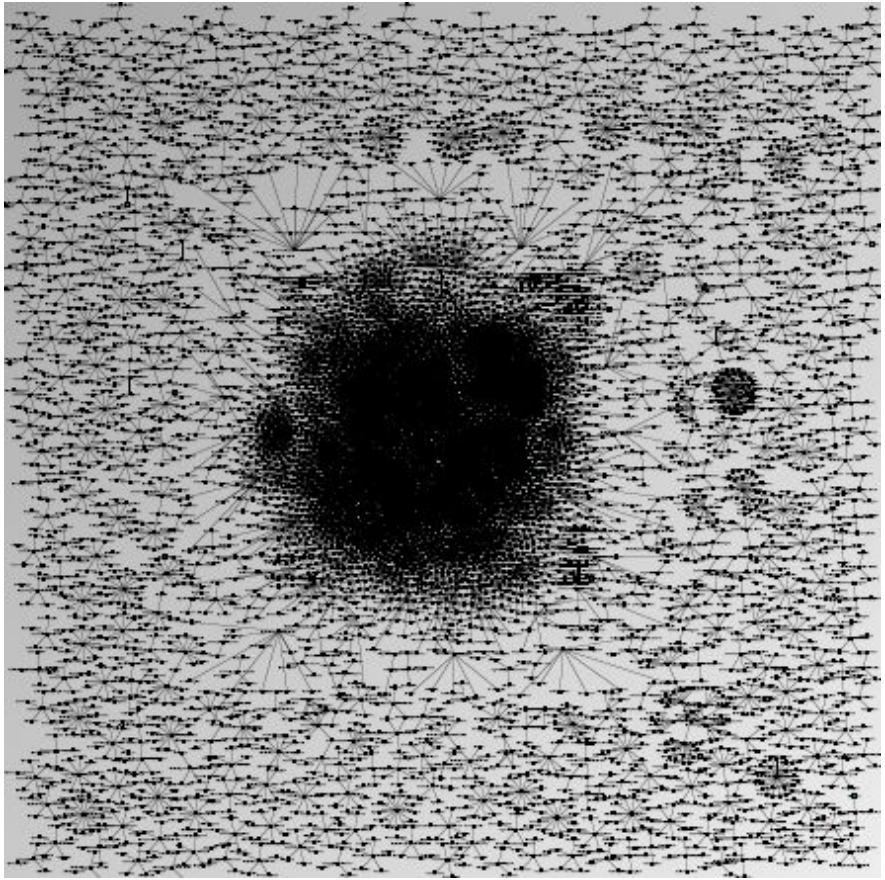
Réseau des co-auteurs à la conférence InfoVis sur 10 ans

Très belle représentation, mais inexacte : De nombreux auteurs ont été retirés



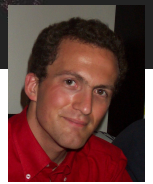
La réalité des réseaux

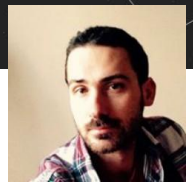
- Ils grandissent mal
- Il faut les tailler pour les rendre lisible
- Mais alors, on devient inexact,
  on déforme la réalité



Cartolabe :
Cartographie de grand corpus

Philippe Caillou Jean-Daniel Fekete Michèle Sebag Jonas Renault Anne-Catherine Letournel
CNRS – INRIA – LRI, Université Paris-Sud

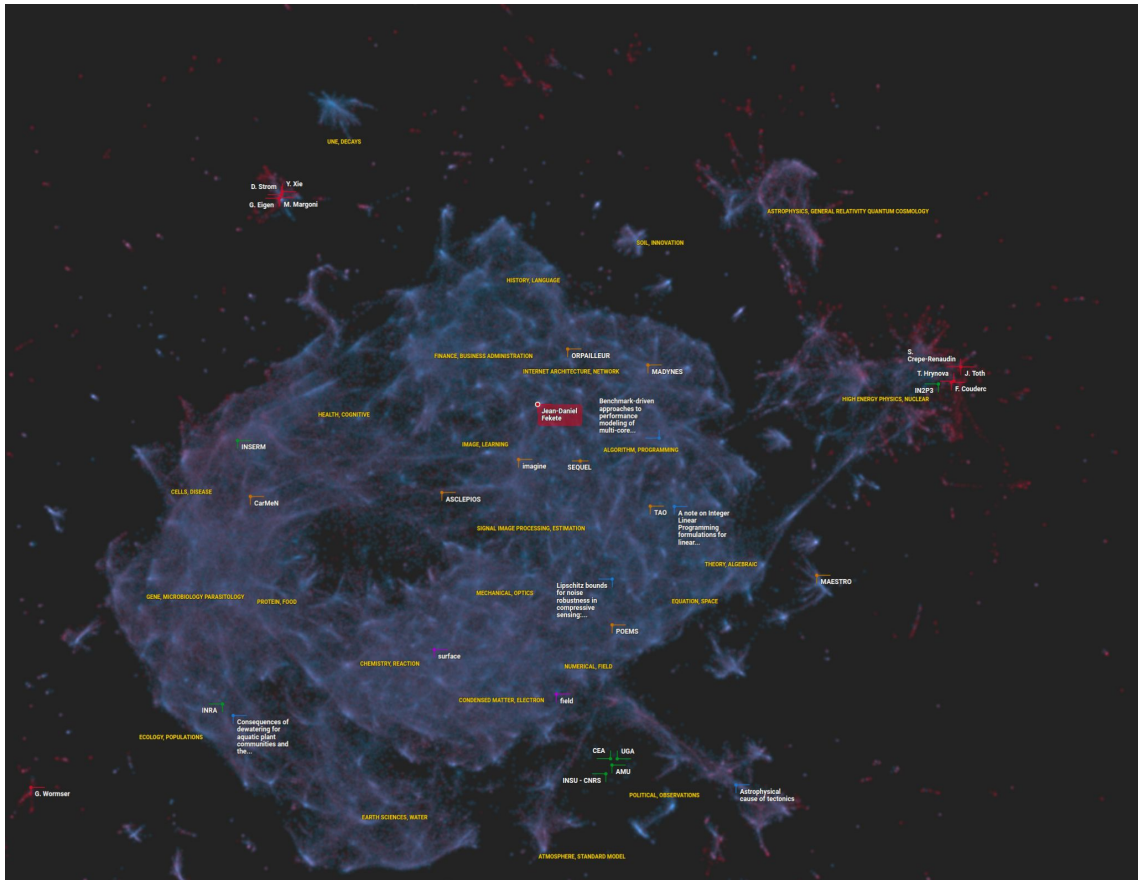


Cartographier un corpus ?

- Les moteurs de recherche permettent de «chercher » !
- Les cartes permettent de représenter l'ensemble d'un corpus
- Règle du jeu :
  - Un document doit être placé près des documents similaires,
    et moins près de documents moins similaires
  - Seule la distance compte, la position absolue est arbitraire

Visualisation de millions de documents



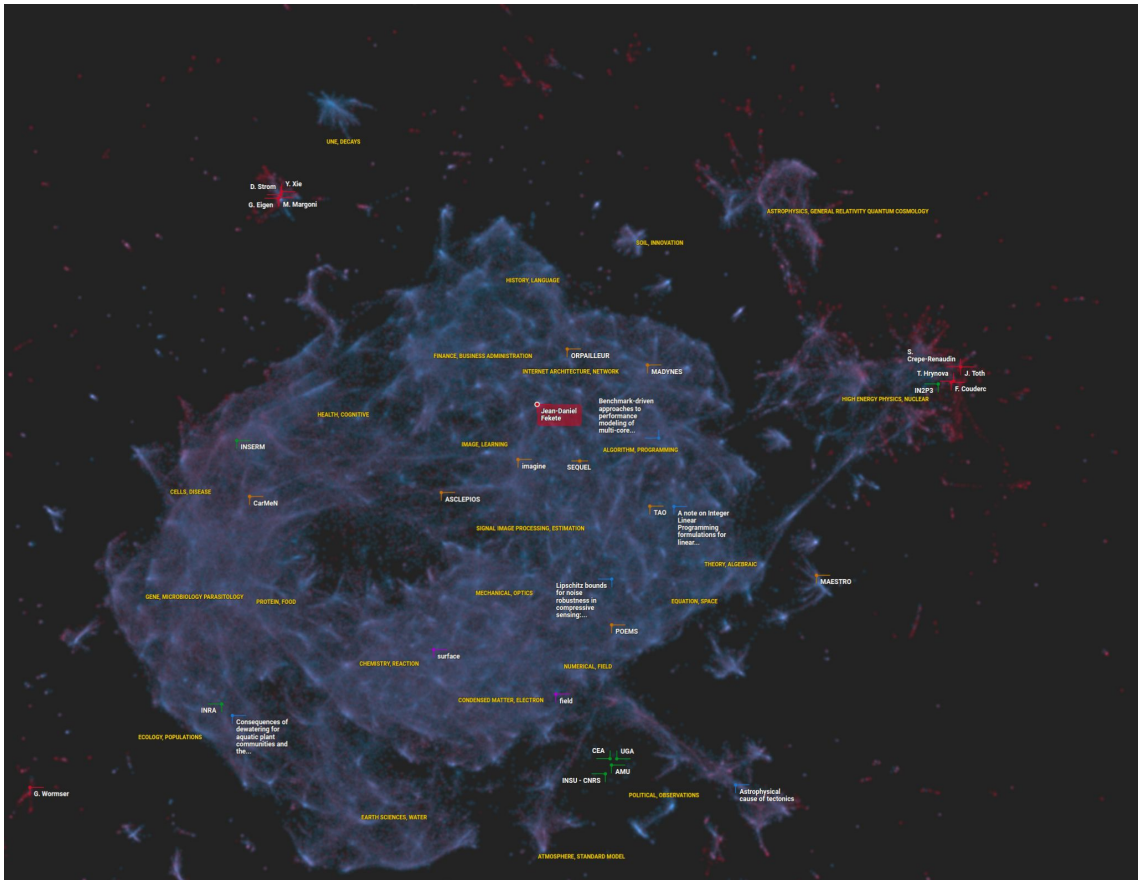
Tâches réalisables avec une carte

- Avoir une vue d'ensemble
- Découvrir un voisinage
- Balayer avec un pointeur
- Naviguer
- Découvrir un chemin
- Découvrir un groupe
- Découvrir une exception

- Trouver une exception de classe
- Évaluer la pureté d'une classe
- Évaluer la compacité d'une classe
- Mettre en correspondance classe et cluster

L. G. Nonato and M. Aupetit, "**Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment**," in *IEEE Transactions on Visualization and Computer Graphics*. doi: 10.1109/TVCG.2018.2846735

Démo à https://cartolabe.fr



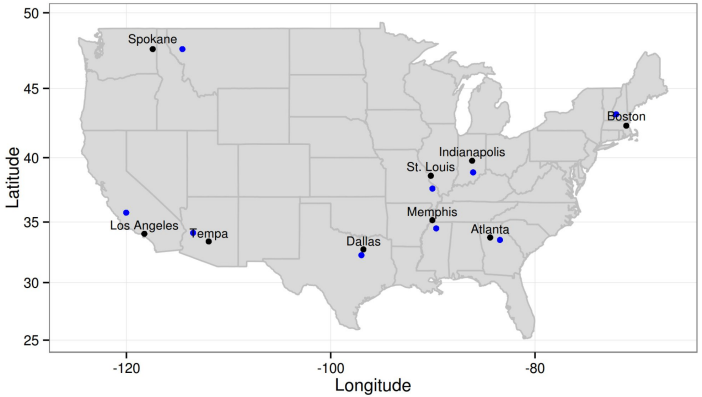
Carte ET moteur de recherche

- Trouver des items connus par nom ou contenu
- Visiter à proximité pour trouver des items similaires, connus ou inconnus
- Comprendre la diversité des items et leur structure
- Explorer par contenu ou groupe identifié (laboratoires, institutions, mot clé)
  - Couverture d'Inria par rapport au reste du monde?
  - Couverture de Saclay par rapport à Inria ?

Positionner des document ?

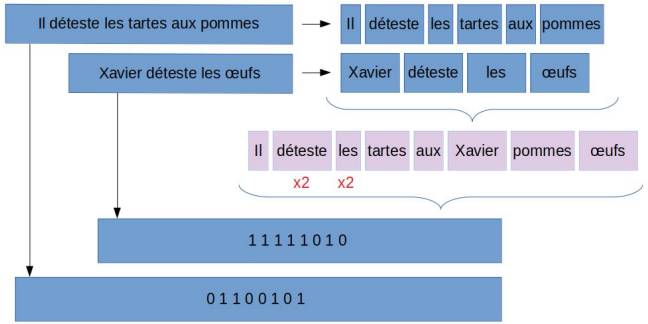
https://renecutura.eu/compadre/

- Sur une carte les villes ont une **position**
- Les documents n'ont pas de **position**
  - Un document doit être placé près de documents similaires,
  - et moins près de documents moins similaires
- On doit calculer la distance entre les documents
- Puis on essaye de faire une carte qui a les mêmes distances
  - Ce n'est pas si facile, mais on a beaucoup progressé depuis 10 ans



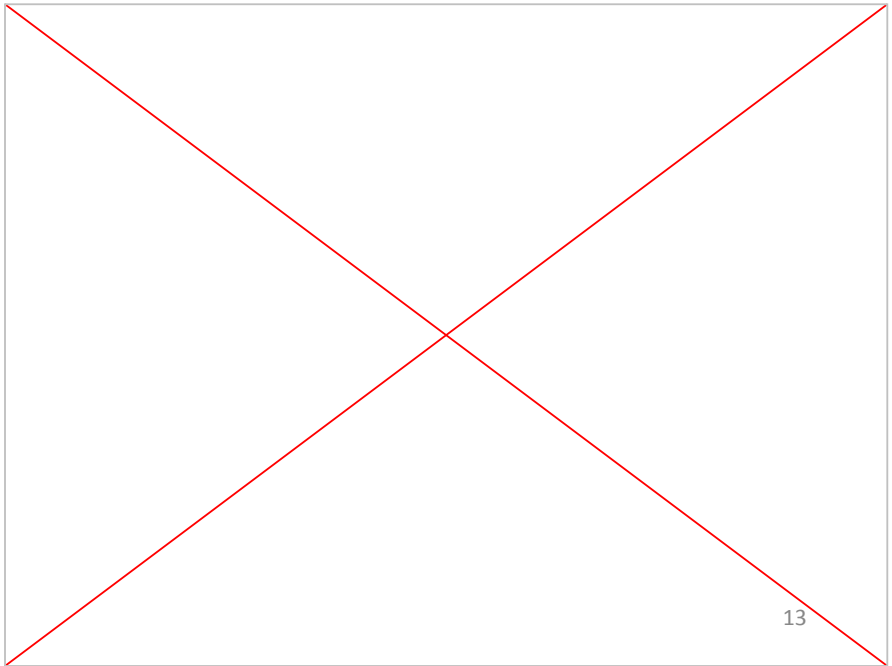
https://www.datacorner.fr/bag-of-words/

Similiarité de documents?



Comment calculer la distance entre documents ?

- Problème de Traitement Automatique du Langage (TAL)
  - Un document devient une longue colonne de nombres
  - Sac de mots
  - Concepts
  - Word Embeddings (IA)
- Après, on mesure une distance entre les colonnes de nombres



Pourquoi est-ce de la recherche ?

Visualisation

- Comment visualiser de manière lisible un large corpus ?
  - Agrégation à plusieurs niveaux
  - Mélange de couleurs devant rester lisibles et séparables
- Comment visualiser rapidement ?
- Comment rendre les interactions fluides ?
- Comment afficher des labels sur la carte ?

Apprentissage automatique

- Quels traitements linguistiques appliquer?
- Comment calculer la distance entre documents?
- Comment projeter les points pour rendre la structure compréhensible?

Comment fonctionne Cartolabe ?

Deux parties très différentes :

1. Module de visualisation générique applicable à n'importe quel corpus (Cartolabe-vis)
   - Peut être spécialisé ensuite
   - Open source (BSD2)
2. Module de traitement de la langue spécifique à chaque corpus (Cartolabe-pipeline)
   - Plusieurs pipelines de traitement en préparation
   - Certains BSD2, pour les vôtres, c'est comme vous voulez

Cartolabe Pipeline HAL

1. Collecte des résumés sur HAL → 500k docs
2. Traitement linguistique → 500k textes
3. LSA ou Bert → 700k vecteurs en dimension ~700
4. Projection UMAP → 500k positions x,y
5. Création de tuiles d'images multi-résolution
6. Clustering hiérarchique k-Means -> 100 clusters
7. Extraction de mots discriminants → 100 labels hiérarchiques pour les labels de régions
8. Transformation au format Cartolabe-Vis

Utilisation de Cartolabe

- Le système est distribué en open-source
  - https://gitlab.inria.fr/cartolabe/cartolabe-data
  - https://gitlab.inria.fr/cartolabe/cartolabe-visu
- Plusieurs corpus disponibles
  - Wikipedia (5 millions de pages)
  - Le Grand Débat (5 millions de contributions)
  - arXiv (2 millions d'articles)

Diffusion de Cartolabe

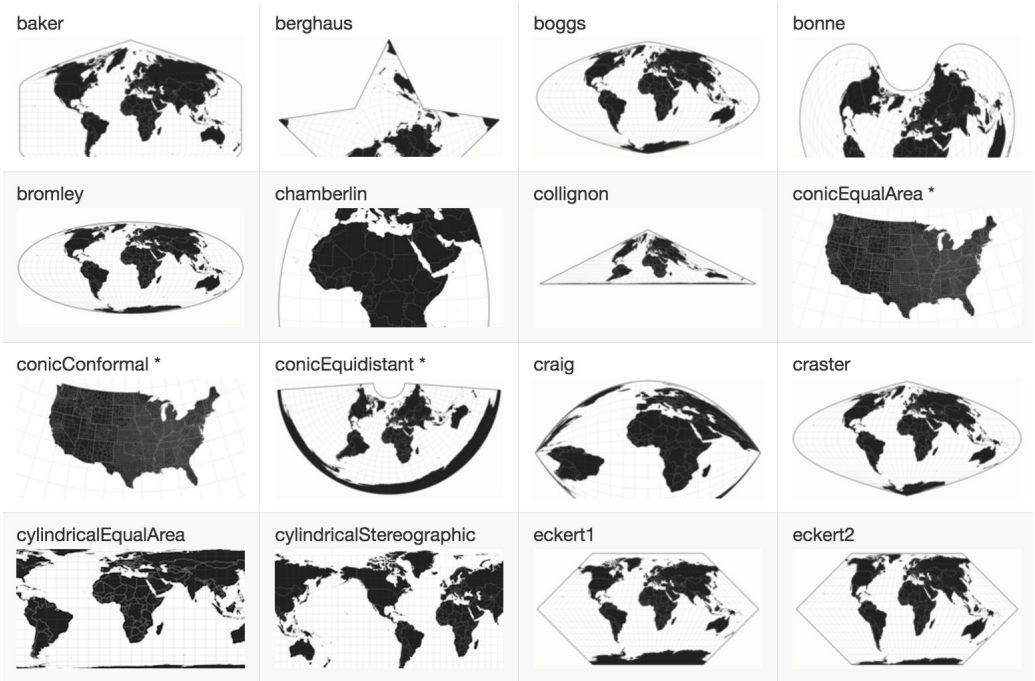
Source d'inspiration pour les usages de demain

- Prenez-le, adaptez-le, améliorez-le

Demande une prise en main

- Lire une carte topographique n'est pas immédiat non plus
  - Nos enfants l'apprennent à l'école
  - Beaucoup d'adultes ont toujours du mal ...

Les cartographes s'en mêlent



https://observablehq.com/@d3/projection-transitions

Comment passer à 2 dimensions ?

Il faut **déplier** et **mettre à plat**

Provoque des :

- Coupures
- Superpositions
- Distortions des distances
- Beaucoup d'artéfacts

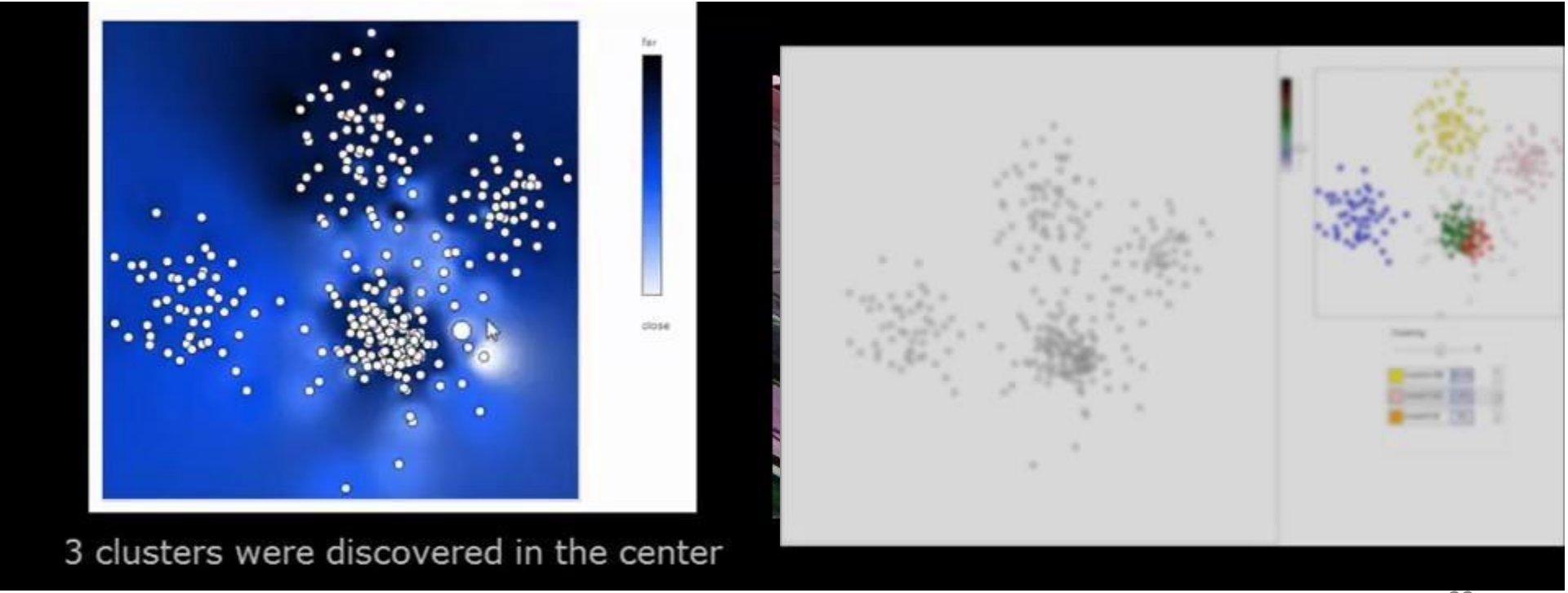
Plus de 3 dimensions ?

Des déchirures et superpositions sont inévitables

- Il faut permettre de
  - les détecter et de
  - les corriger

Détecter et corriger en HD

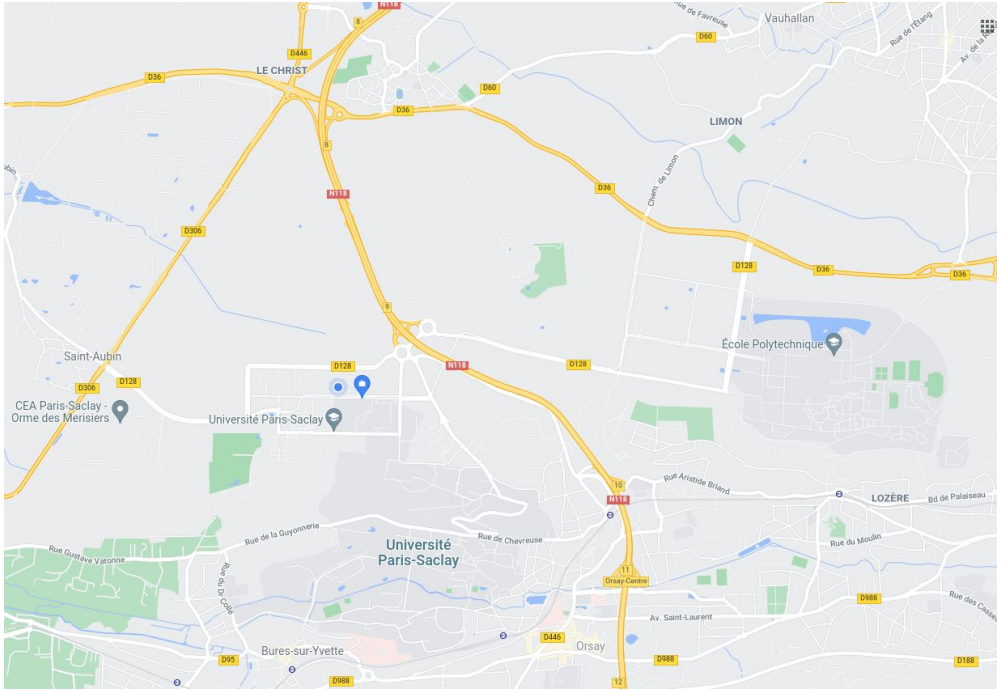
Michaël Aupetit, Nicolas Heulot, Jean-Daniel Fekete. A multidimensional brush for scatterplot data analytics. IEEE. *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, Oct 2014, Paris, France. IEEE, pp.221 - 222, 2014



Détecter et corriger les erreur

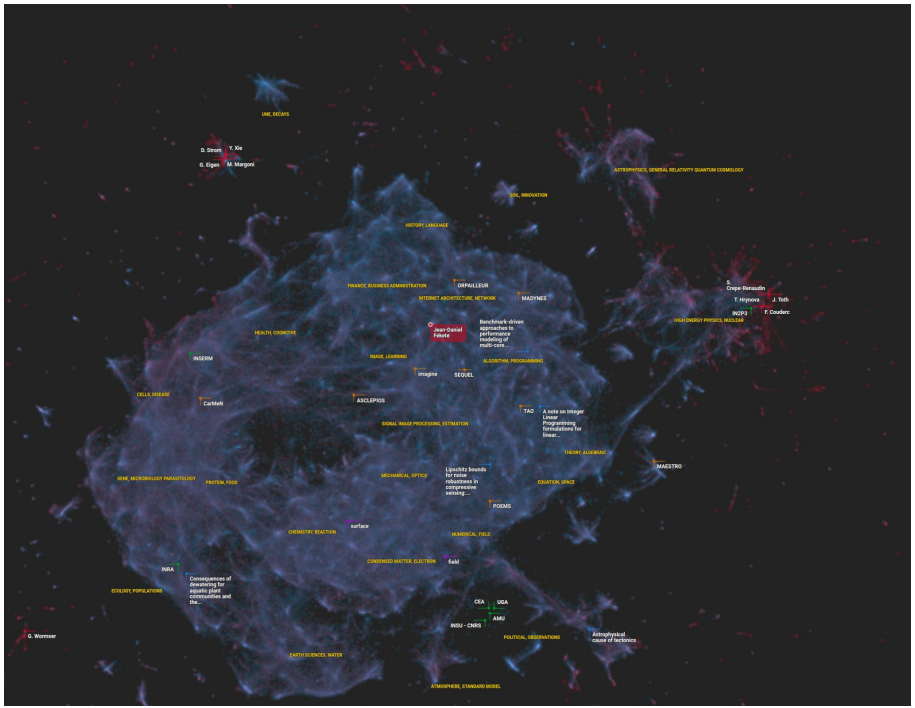
On le fait déjà

- Les voies rapides produisent des distortions
- Les reliefs produisent des déchirures

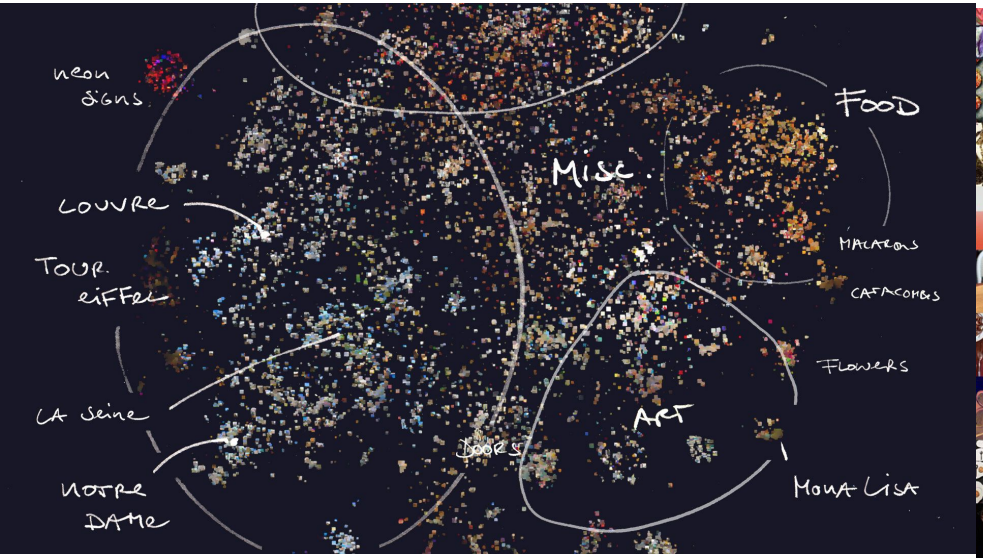


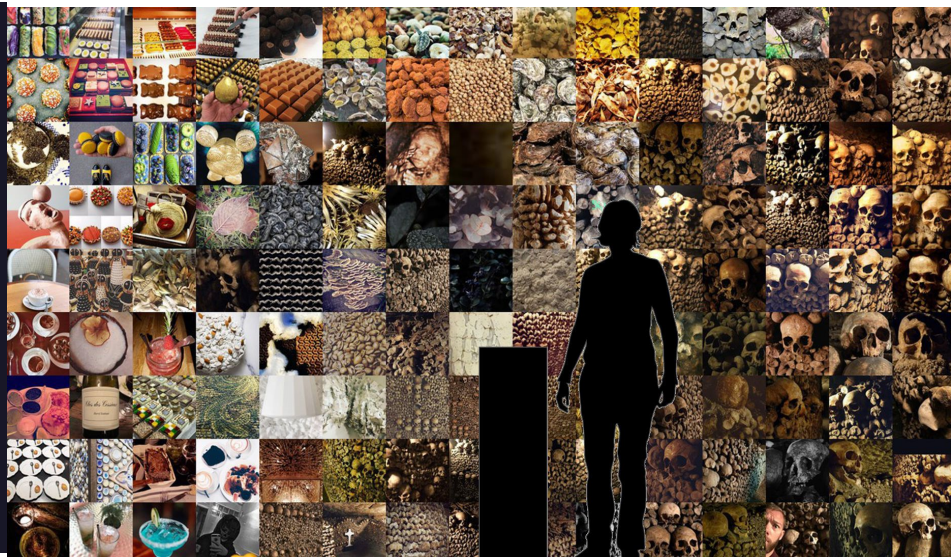
Conclusion

- On peut maintenant découvrir et explorer de grands corpus
- Il va falloir s'y habituer et apprendre à gérer les artéfacts
- Pour mieux explorer les données



Multiplicity - Moritz Stefaner





https://truth-and-beauty.net/projects/multiplicity