Réunion inaugurale du groupe de contact FNRS
« Les humanités des données »

Données, humanités et méthodes quantitatives : premières pistes de réflexion

Nicolas Ruffini-Ronzani (UNamur et AÉN)
Sébastien de Valeriola (ULB)

Bruxelles, lundi 7 novembre 2022

Introduction

« Faut-il en conclure qu'entre les sciences exactes et naturelles, d'une part, les sciences humaines et sociales, de l'autre, la différence est si profonde, si irréductible, qu'on doit perdre tout espoir d'étendre jamais aux secondes les méthodes rigoureuses qui ont assuré le triomphe des premières ? Une telle attitude [...] nous paraît entachée d'un véritable obscurantisme, en prenant ce terme dans son sens étymologique : obscurcir le problème au lieu de l'éclairer. »

LÉVI-STRAUSS, "Les mathématiques de l'homme", 1956, p. 531.

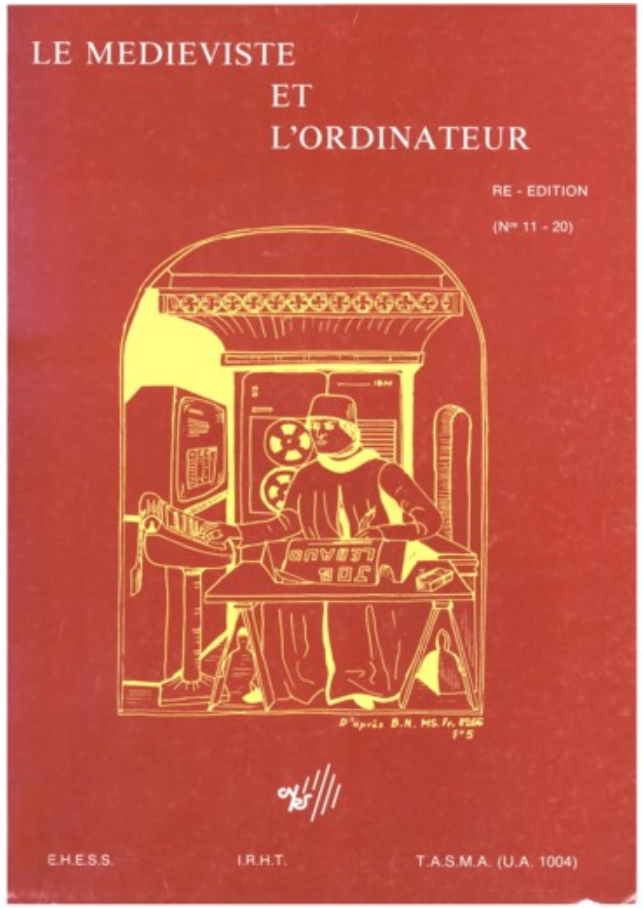
Les premiers projets quantitatifs

- L’application de méthodes quantitatives aux sciences humaines n’est pas une idée nouvelle.
- En 1946, Roberto Busa se lance dans la construction d’un index des œuvres de Thomas d’Aquin avec l’aide d’IBM.
- Dans les années suivantes, on a vu l’apparition de projets visant à appliquer des méthodes quantitatives à de gros corpus.
- De même, des centres dédiés à des entreprises de ce type sont apparus (le LASLA à Liège en 1961, le CETEDOC à Louvain en 1968).

- Les opérations appliquées alors étaient extrêmement simples.
- Depuis cette période, trois grandes évolutions ont eu lieu.

Première évolution (seconde moitié du 20e siècle)

- Les ordinateurs se démocratisent et se perfectionnent, avec une puissance de calcul accrue.
- De nouvelles approches quantitatives sont désormais accessibles aux chercheurs en sciences humaines…
- ...du moins en théorie, car peu d'applications concrètes en sciences humaines (exceptions : Jean-Philippe Genêt et Alain Guerreau en histoire médiévale et François Djindjian en archéologie, par exemple).



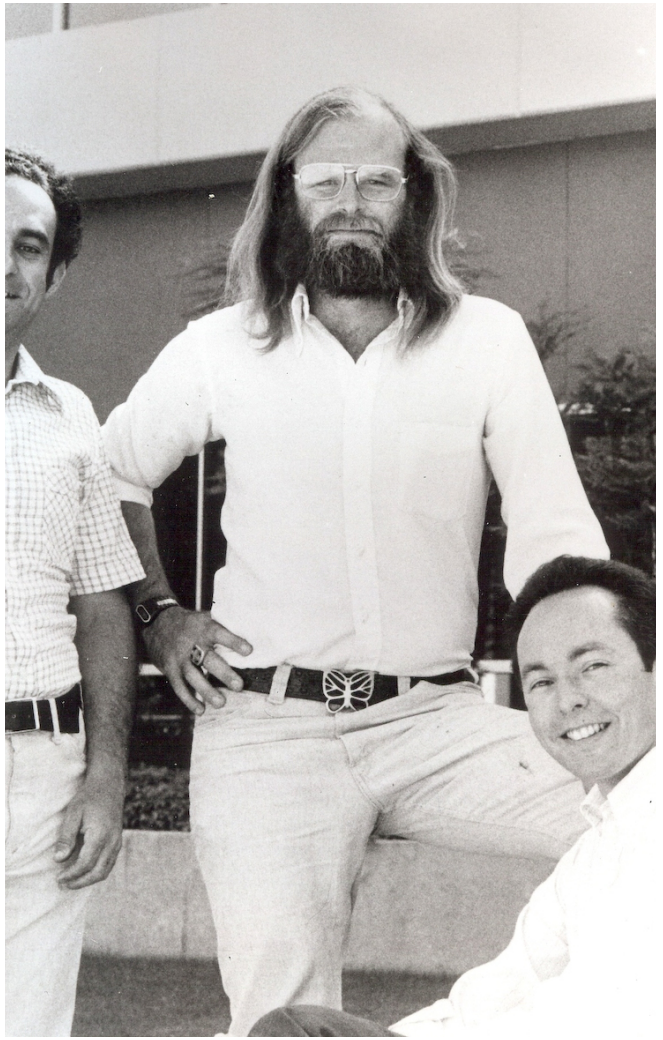
Revue *Le médiéviste et l'ordinateur* (1979-2003)

Deuxième évolution (années 1990 et 2000)

- Avec l'apparition puis la popularisation du Web, au cours des années 1990 et surtout des années 2000, notre société est devenue de plus en plus connectée.
- On a vu se multiplier les corpus disponibles sur le Web : corpus de textes, d'images, d'objets géoréférencés, de sons, de vidéos, etc.
- Les exemples de corpus ainsi disponibles sont aujourd'hui innombrables.
- Cette explosion ne s'est néanmoins pas accompagnée d'une explosion similaire en termes de méthodes quantitatives appliquées de façon globale à ces corpus.
- Bien souvent (pas toujours), l'utilisation de telles techniques n'était tout simplement pas considérée comme une possibilité par les concepteurs de ces sites Web.
- Ils ont cherché à faciliter le travail des chercheurs en sciences humaines, sans penser à une potentielle évolution structurelle (ou en tout cas plus profonde) des méthodes de travail.

Troisième évolution (depuis les années 2000)

- Depuis les années 2000, nous vivons la « *data revolution* ».
- On a pris conscience de la valeur ajoutée que pouvaient représenter la récolte systématique, le traitement et la mise en œuvre de quantités importantes de données.
- Les méthodes *data-driven* sont (ré-)apparues, c'est-à-dire celles qu'on regroupe dans la discipline appelée *data science*, « science des données », à cheval sur la statistique, l'informatique et les sciences de l'information.
- Nous prenons l'expression dans un sens très large, en y incluant toutes les méthodes quantitatives développées dans un cadre statistique ou algorithmique.
- L'impact de leur utilisation sur la recherche scientifique en général est monumental : James Gray (prix Turing 1998), parle d'un nouveau paradigme scientifique : le paradigme *data-intensive.*
- Les sciences humaines sont aussi concernées par ce « *data turn* » !



James Gray

Une typologie des méthodes quantitatives en sciences humaines

1. Acquisition des données: l'ordinateur facilite le dépouillement des sources, documents, images, etc.

2. Analyse exploratoire : l'ordinateur met au jour un phénomène que le chercheur aurait pu découvrir lui-même, mais qui lui est resté invisible pour des raisons d'échelle

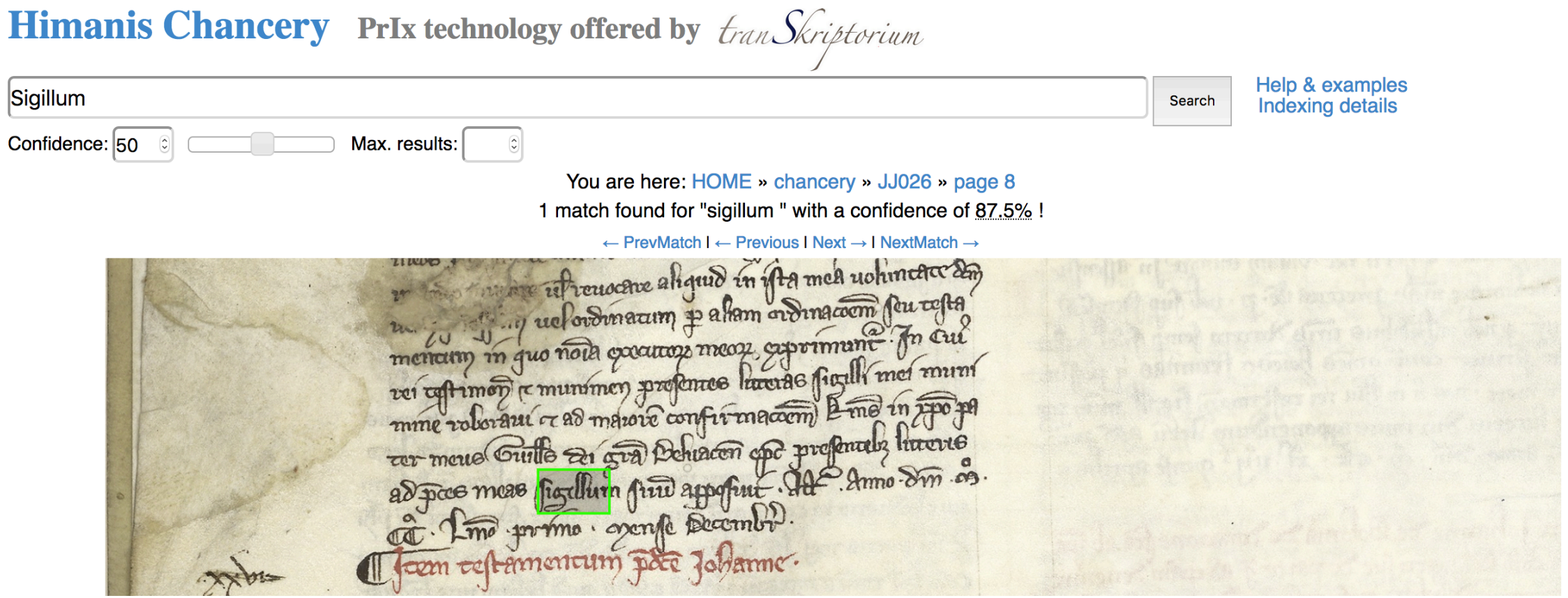
3. Analyse interprétative : l'ordinateur propose des éléments d'interprétation, d'explication, que le chercheur analyse pour conclure

4. Exercice de la critique : l'ordinateur permet de mettre en œuvre une forme de critique complémentaire, en interaction avec le chercheur

➔ Comme toute typologie, elle peut être soumise à la critique. Nous sommes à l'écoute de vos réactions !

1. Acquisition des données : exemples

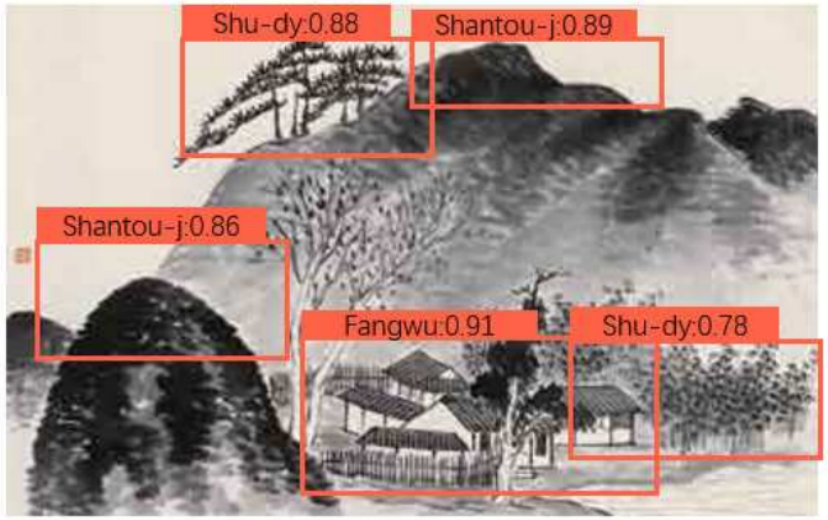
Extraction du texte d'un écrit manuscrit à l'aide d'outils de *Handwritten Text Recognition*

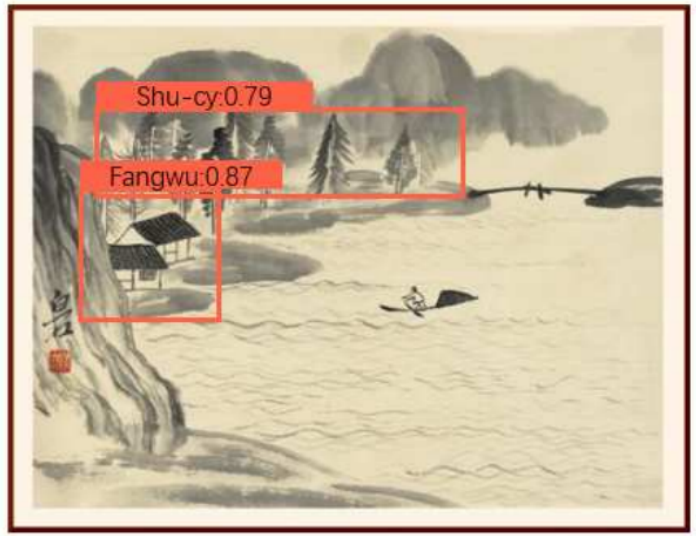


Source : STUTZMANN et al, "HIMANIS project", depuis 2015, url : http://himanis.huma-num.fr/app/

1. Acquisition des données : exemples

Détection de la présence de détails (par exemple des objets) dans les peintures d'un corpus, afin de constituer un sous-corpus

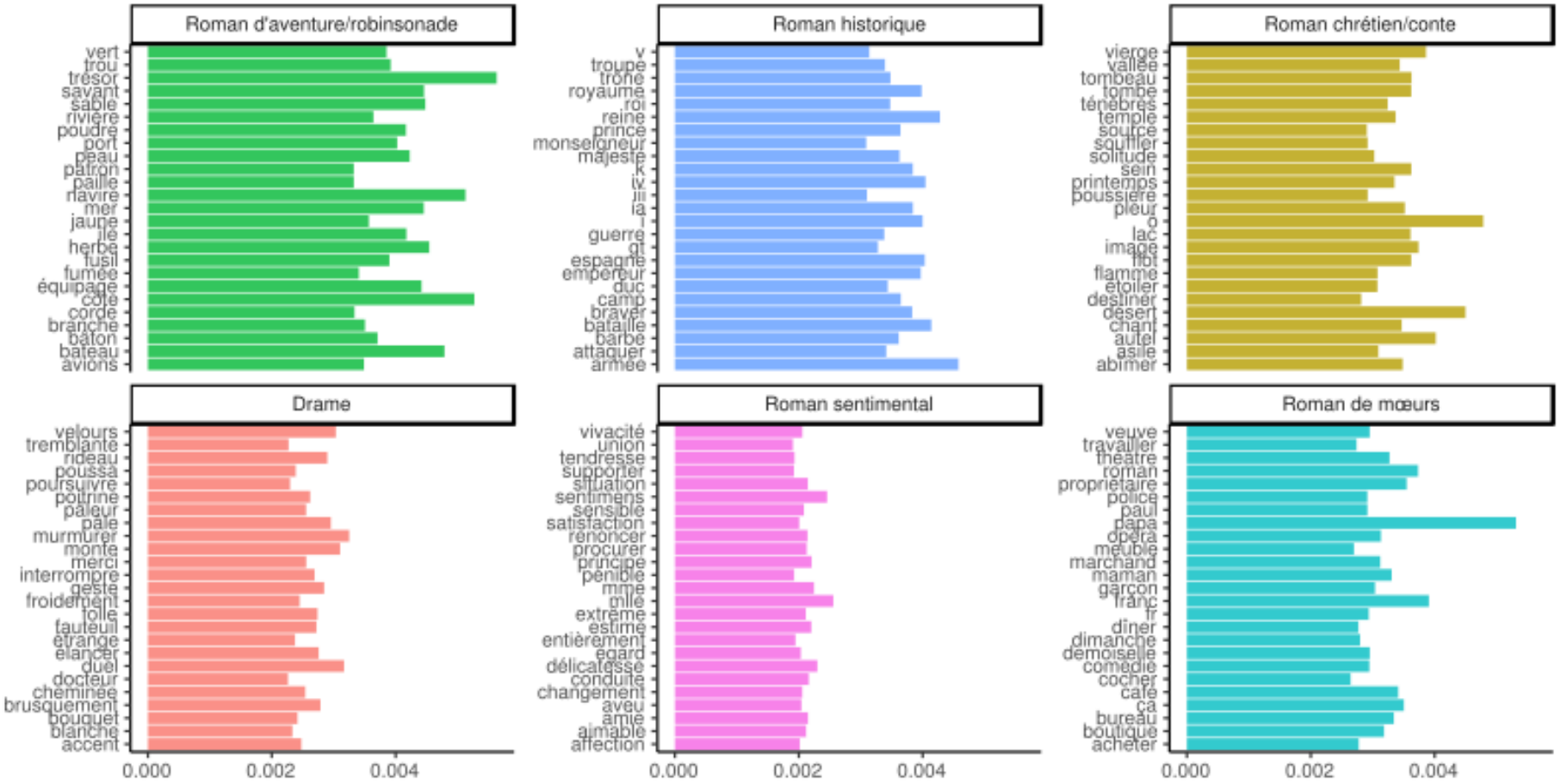




Source : CAI, YANG, ZHOU et LIN, "Automatic Detection of Landscape Painting Elements Based on Machine Learning", 2019

1. Acquisition des données : exemples

Classification automatique des textes d'un corpus en genres, afin de constituer un sous-corpus



Source : LANGLAIS, "Reconstituer les genres romanesque sur Gallica : essai de classification automatisée de 1500 romans", 2019.

2. Analyse exploratoire : exemples

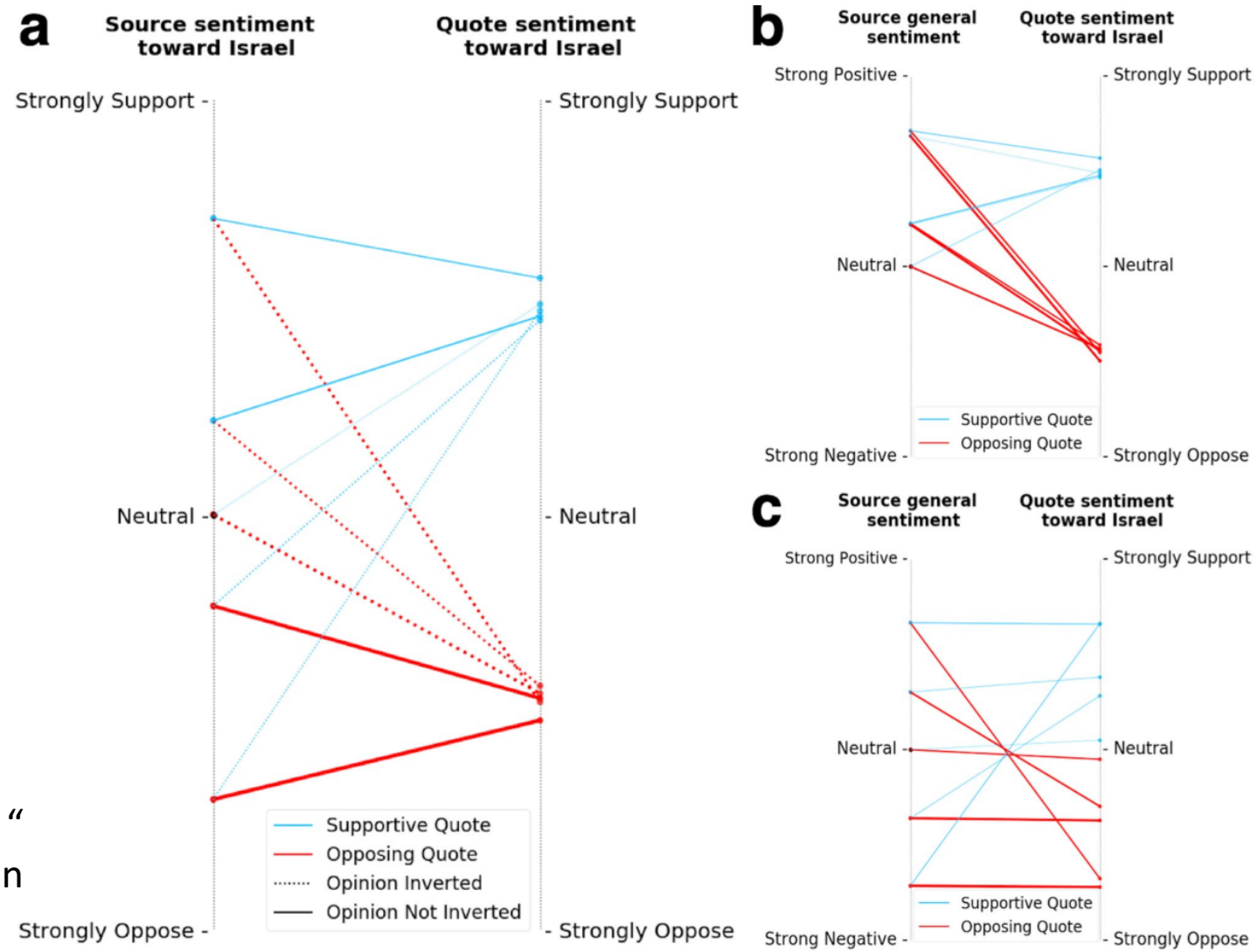
Comparaison d'une grande quantité de textes les uns avec les autres, à la recherche par exemple de similitudes, pour étudier les réutilisations citations

	Source	Target	Match Features	Score
1	**vergil aeneid 1.397**: ut reduces illi ludunt stridentibus alis et coetu cinxere polum cantusque dedere haud aliter puppesque alis et coetu cinxere polum cantusque dedere haud aliter puppesque tuae pubesque tuorum aut portum tenet aut pleno subit ostia velo	**lucan bellum civile 9.283**: et omnes Haud aliter medio revocavit ab aequore puppes Quam simul effetas linquunt examina ceras Atque oblita favi non miscent nexibus alas aliter medio revocavit ab aequore puppes Quam simul effetas linquunt examina ceras Atque oblita favi non miscent nexibus alas Sed sibi quaeque volat nec iam degustat amarum Desidiosa thymum	haud, puppis, alo, aliter, ala	**10**
2	**vergil aeneid 4.256**: Haud aliter terras inter caelumque volabat litus harenosum Libyae ventosque secabat materno veniens ab volabat litus harenosum Libyae ventosque secabat materno veniens ab avo Cyllenia proles	**lucan bellum civile 9.283**: et omnes Haud aliter medio revocavit ab aequore puppes Quam simul effetas linquunt examina ceras Atque oblita favi non miscent nexibus alas Sed sibi quaeque volat aliter medio revocavit ab aequore puppes Quam simul effetas linquunt examina ceras Atque oblita favi non miscent nexibus alas Sed sibi quaeque volat nec iam degustat amarum Desidiosa thymum	aliter, haud, uolo, ab	**10**

Source : COFFEE et al, "The Tesserae Project: Intertextual Analysis of Latin Poetry," 2013.

2. Analyse exploratoire : exemples

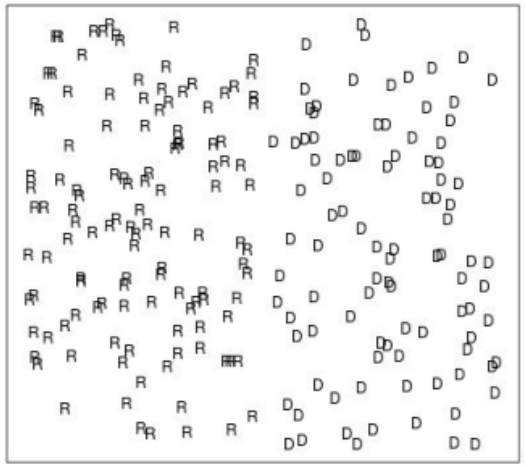
Identification des textes au contenu le plus fortement polarisé au sein d'un corpus (de tweets, par exemple) à l'aide d'outils d'analyse des sentiments



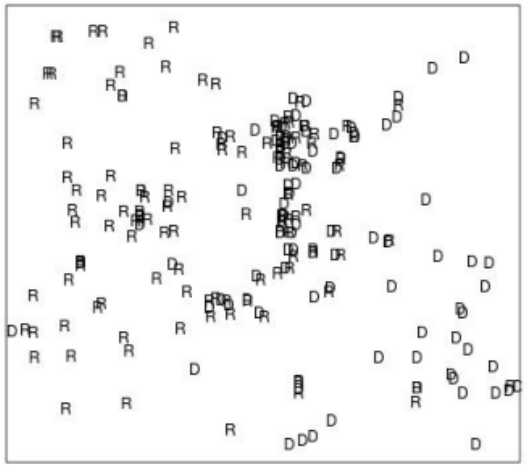
Source : MATALON, MAGDACI, ALMOZLINO et YAMIN, " Using sentiment analysis to predict opinion inversion in Tweets of political communication", 2021

2. Analyse exploratoire : exemples

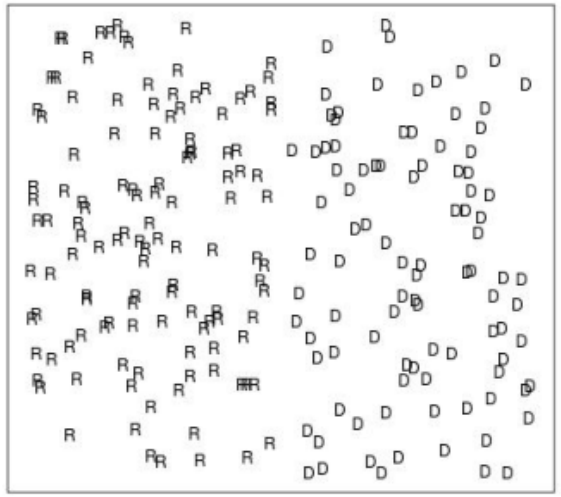
Exploration de la dynamique des votes au sein d'une assemblée représentative à l'aide de l'analyse des vidéos des séances (et du mouvement des représentants dans la salle pendant les périodes de vote)



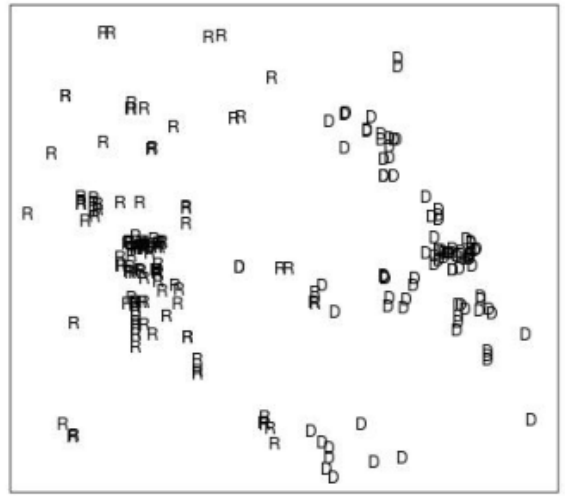
(a) Bipartisan (T = 0)



(b) Bipartisan (T = 100)



(c) Polarization (T = 0)

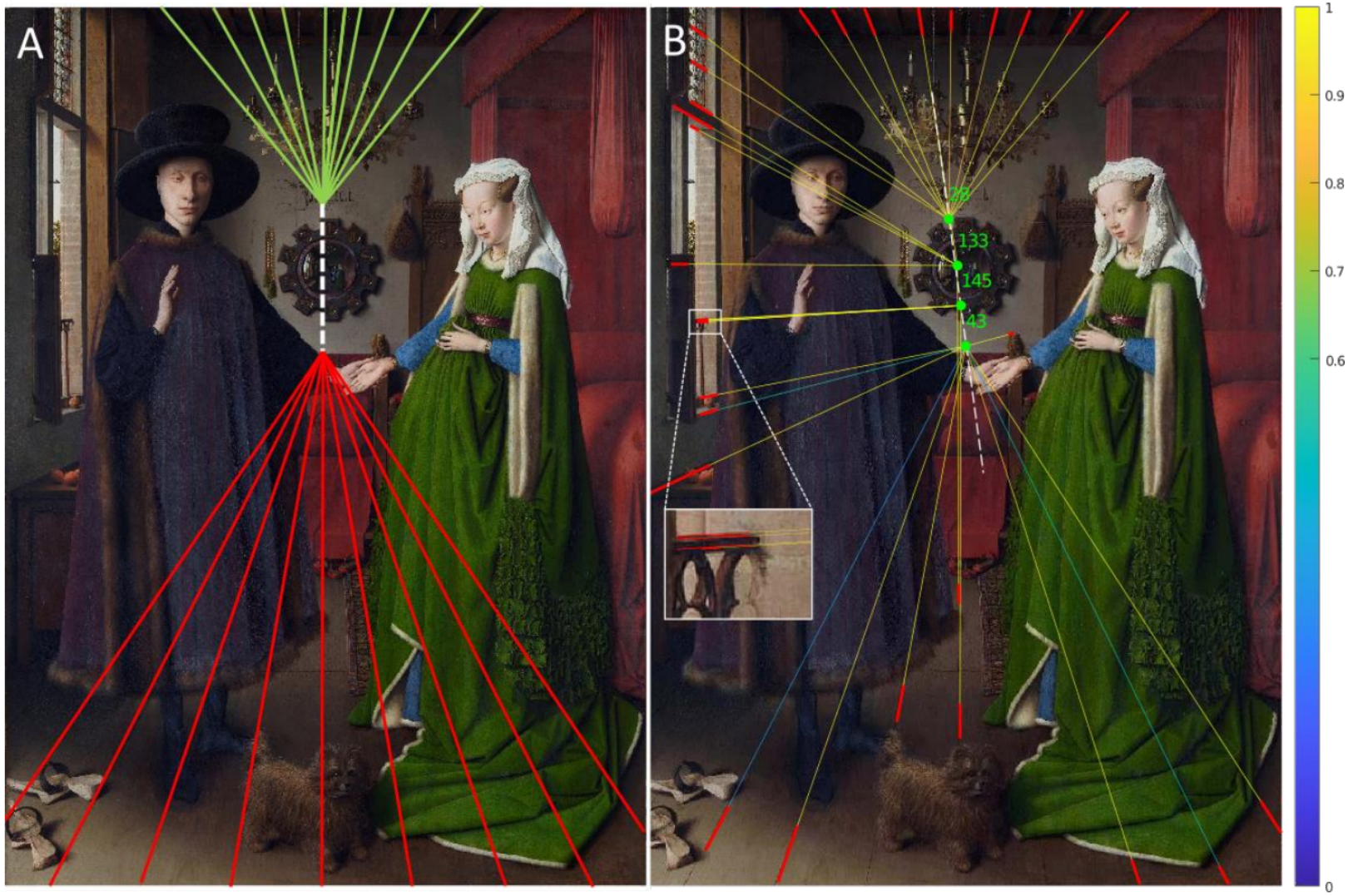


(d) Polarization (T = 100)

Source : DIETRICH, "Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives", 2020

2. Analyse exploratoire : exemples

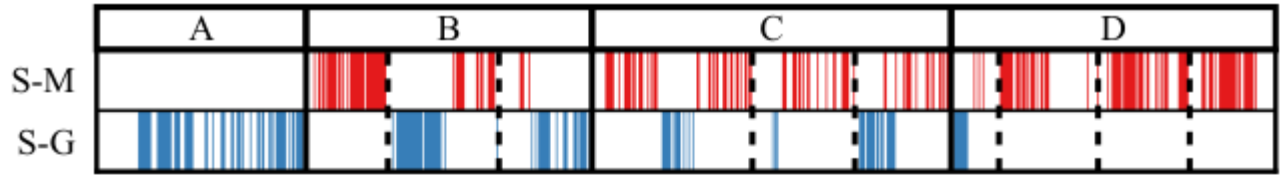
Détection automatique des points de fuite dans des peintures anciennes

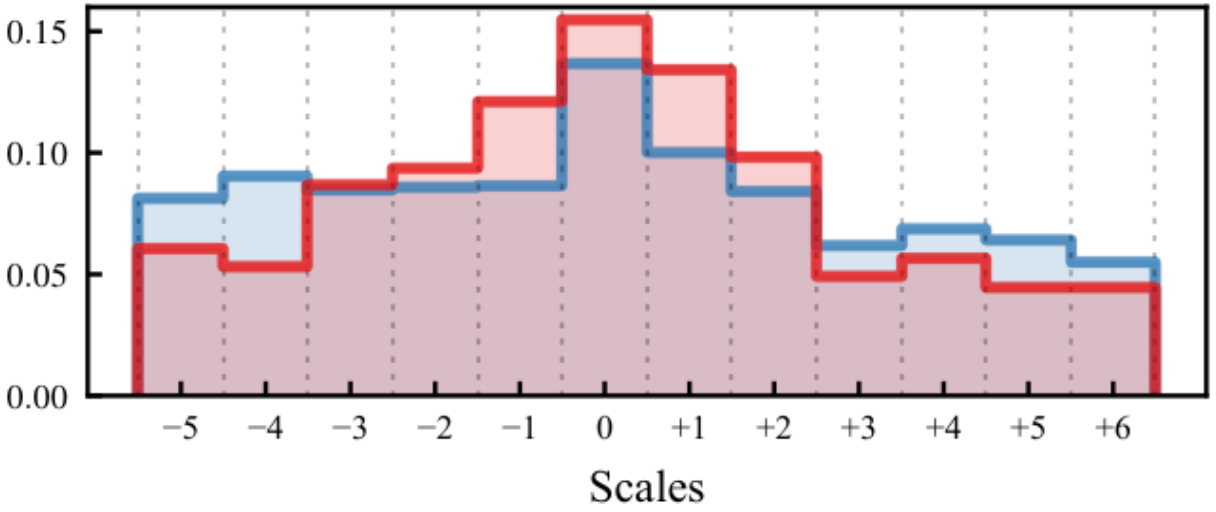


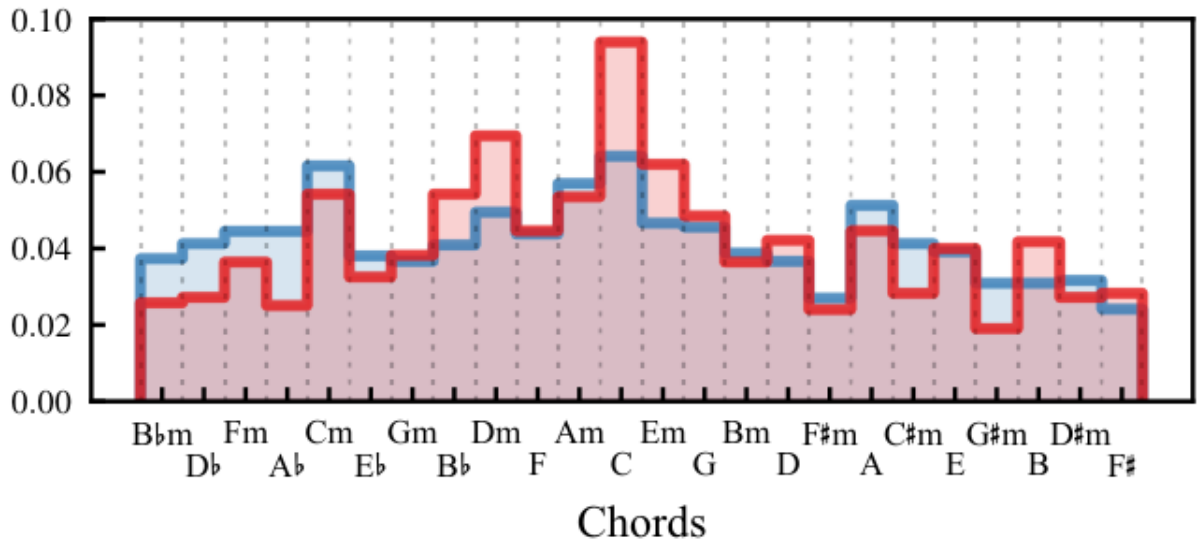
Source : SIMON, “Jan van Eyck's Perspectival System Elucidated Through Computer Vision”, 2021

2. Analyse exploratoire : exemples

Extraction des leitmotivs utilisés dans une œuvre musicale et identification des contextes (harmoniques, scénaristiques, etc.) dans lesquels ils apparaissent



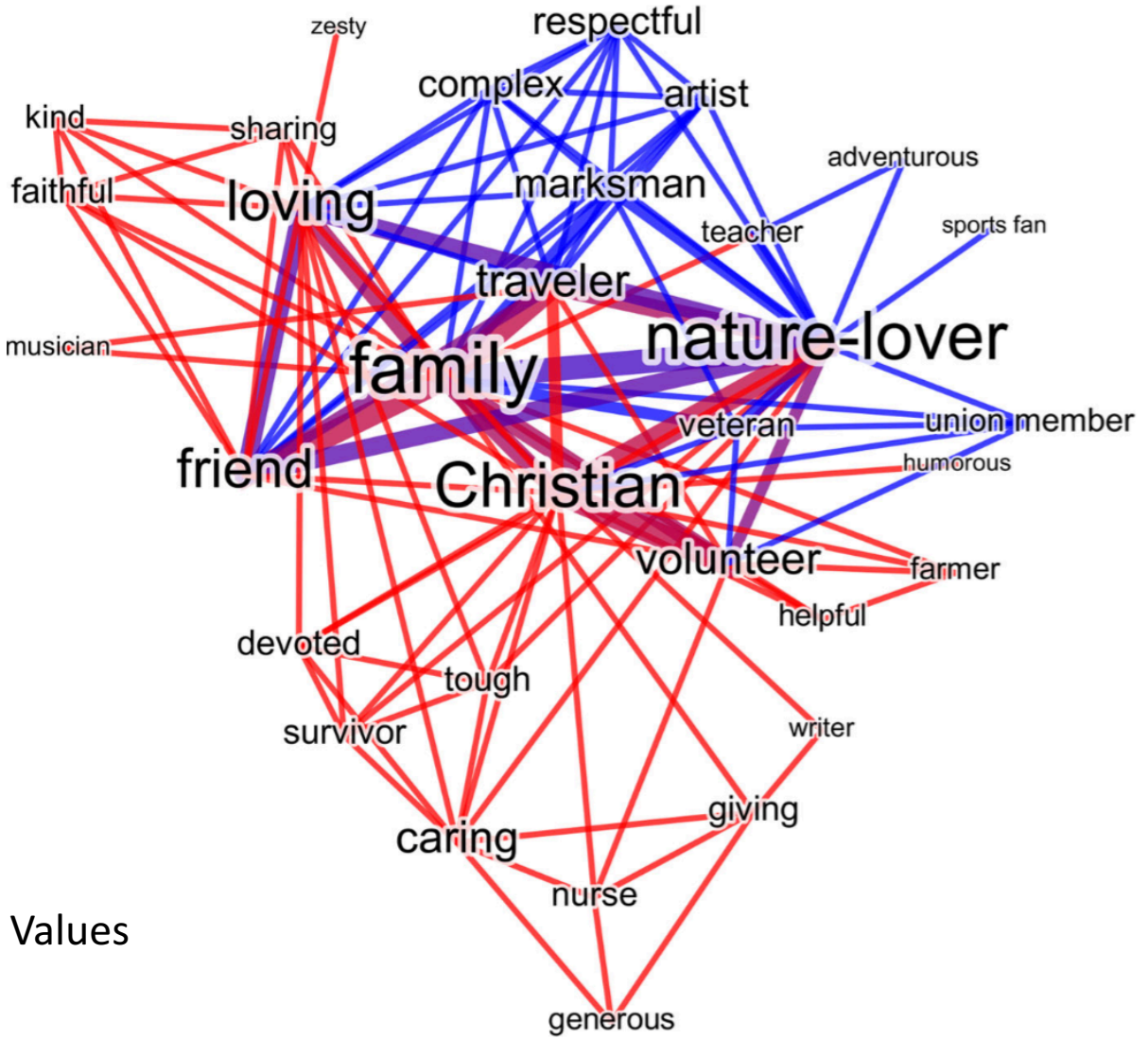




Source : ZALKOW, WEISS et MÜLLER, “Exploring tonal-dramatic relationships in Richard Wagner's ring cycle”, 2017

2. Analyse exploratoire : exemples

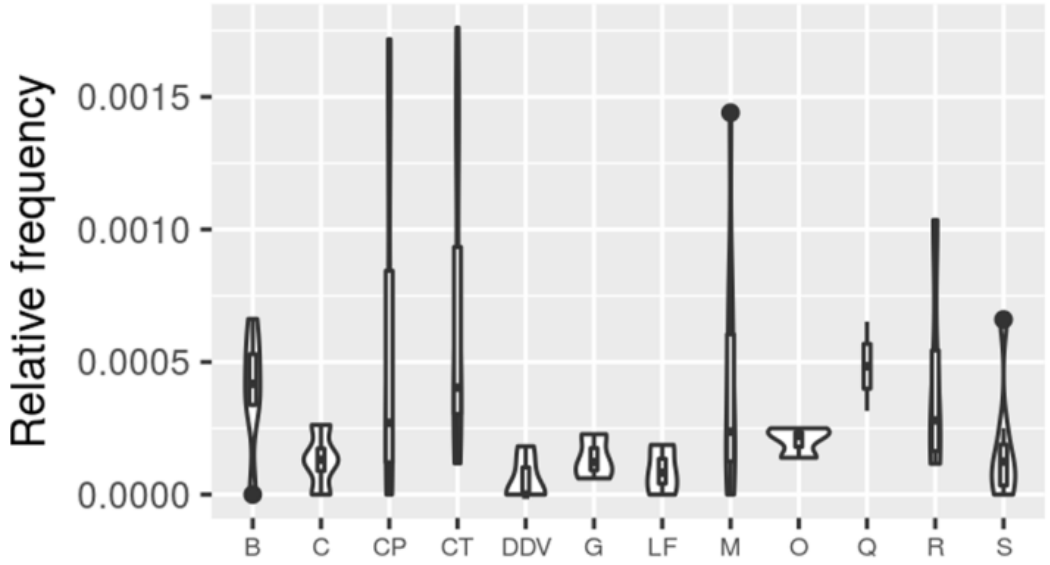
Utilisation de l'analyse de réseaux pour analyser les liens entre des « valeurs morales » mentionnées dans des notices nécrologiques



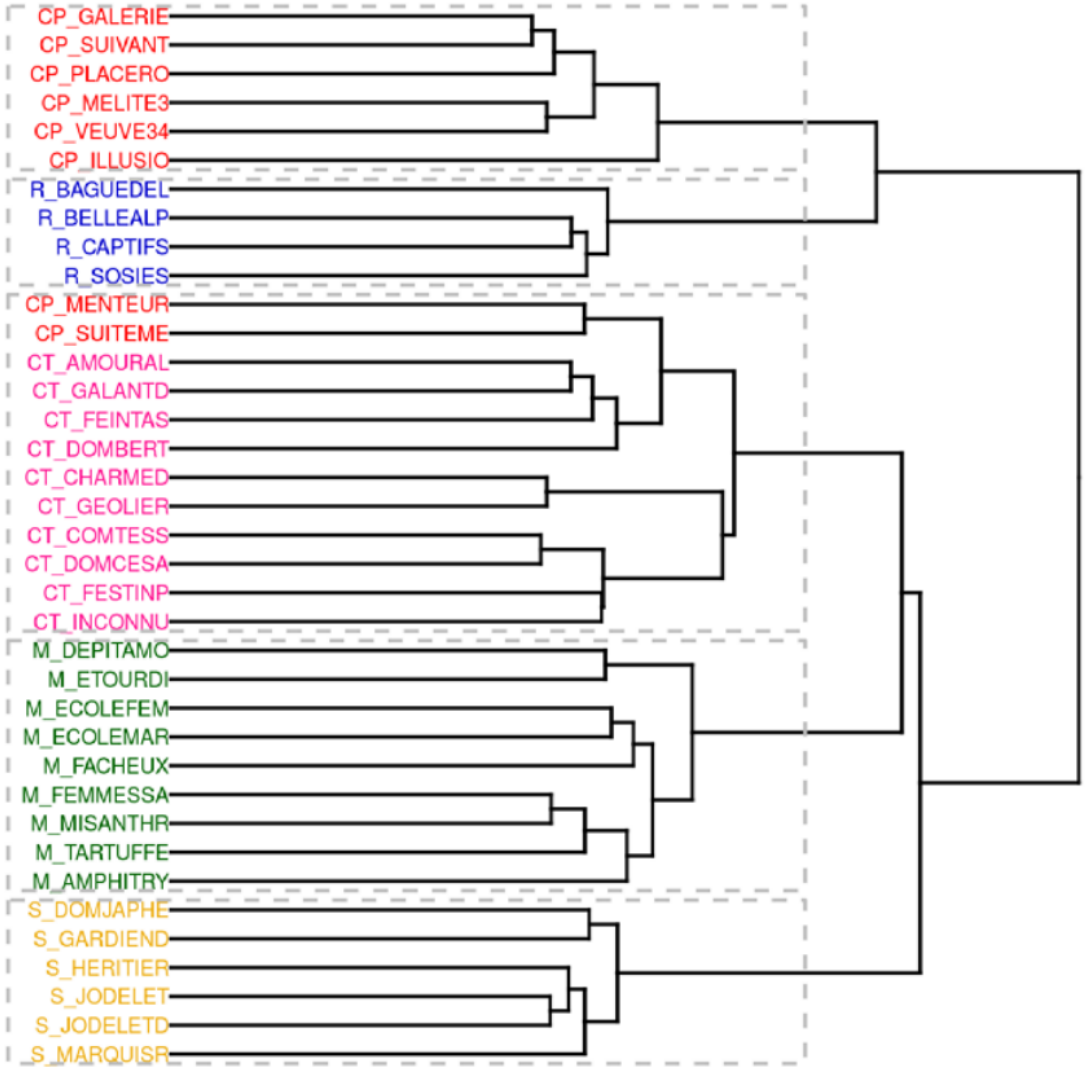
Source : ALFANO, HIGGINS et LEVERNIER, "Identifying Virtues and Values Through Obituary Data-Mining",2018.

3. Analyse interprétative : exemples

Attribution de textes orphelins à des auteurs / confirmation de doutes à propos de la paternité de textes grâce à des méthodes stylométriques

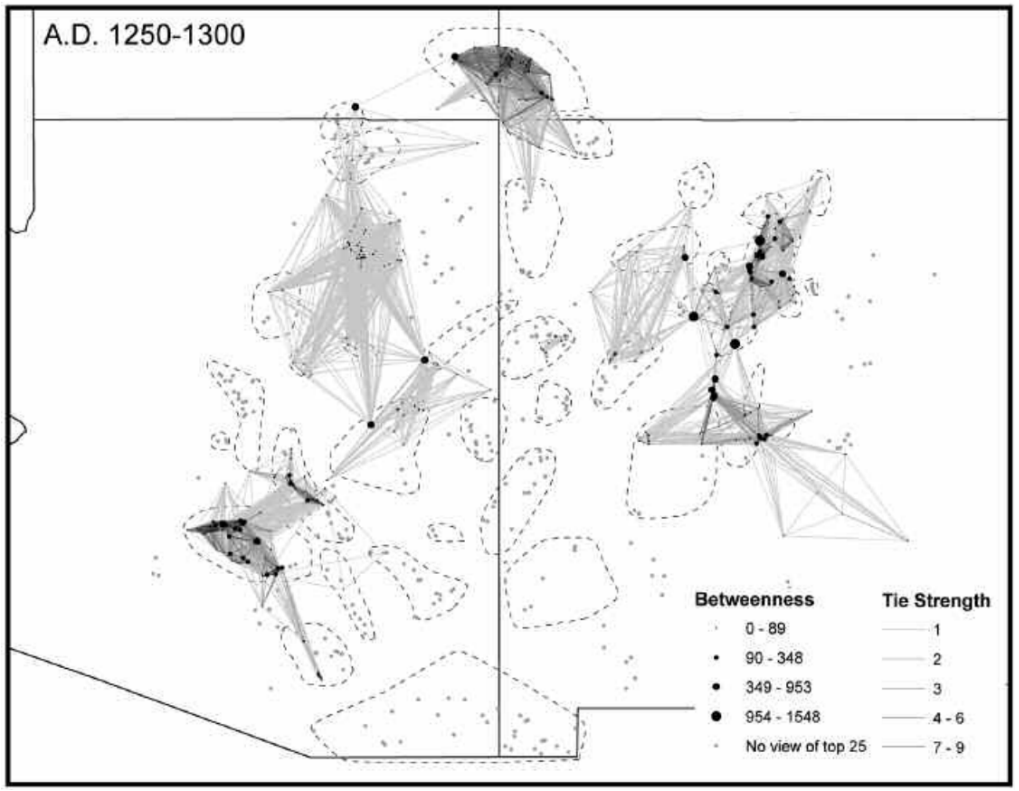


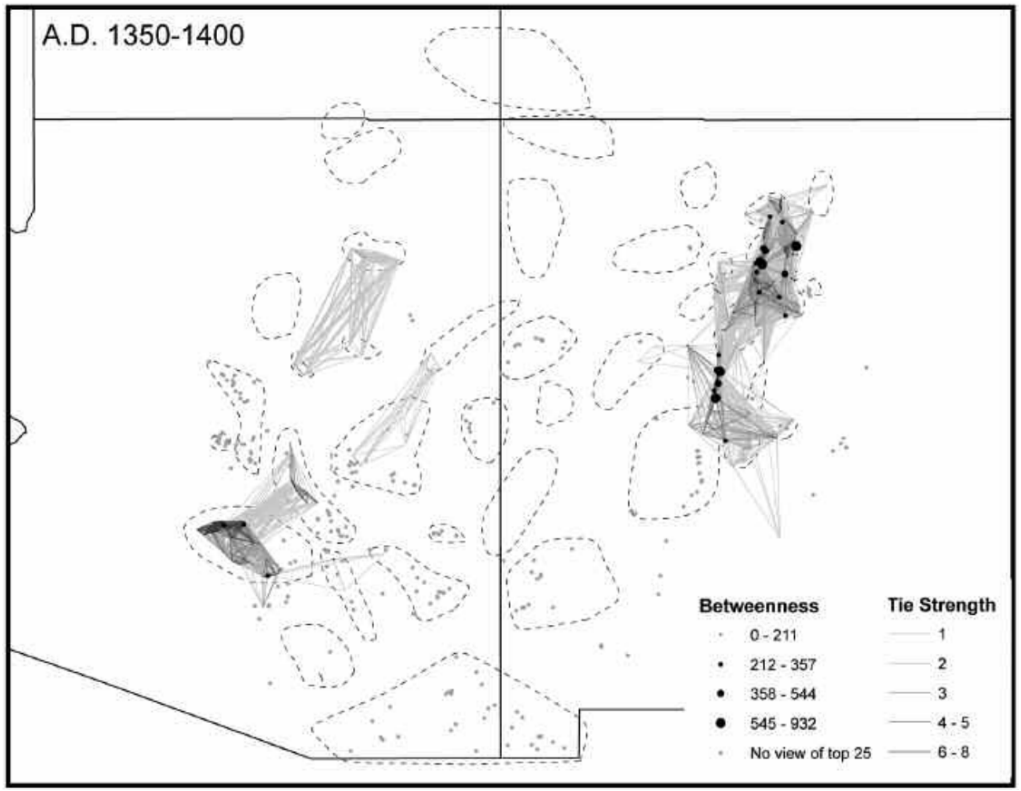
Source : CAFIERO et CAMPS, “Why Molière Most Likely Did Write His Plays” 2019.



3. Analyse interprétative : exemples

Constitution de groupes de villages partageant un paysage visuel similaire et potentiellement une base culturelle commune à l'aide de l'analyse des réseaux

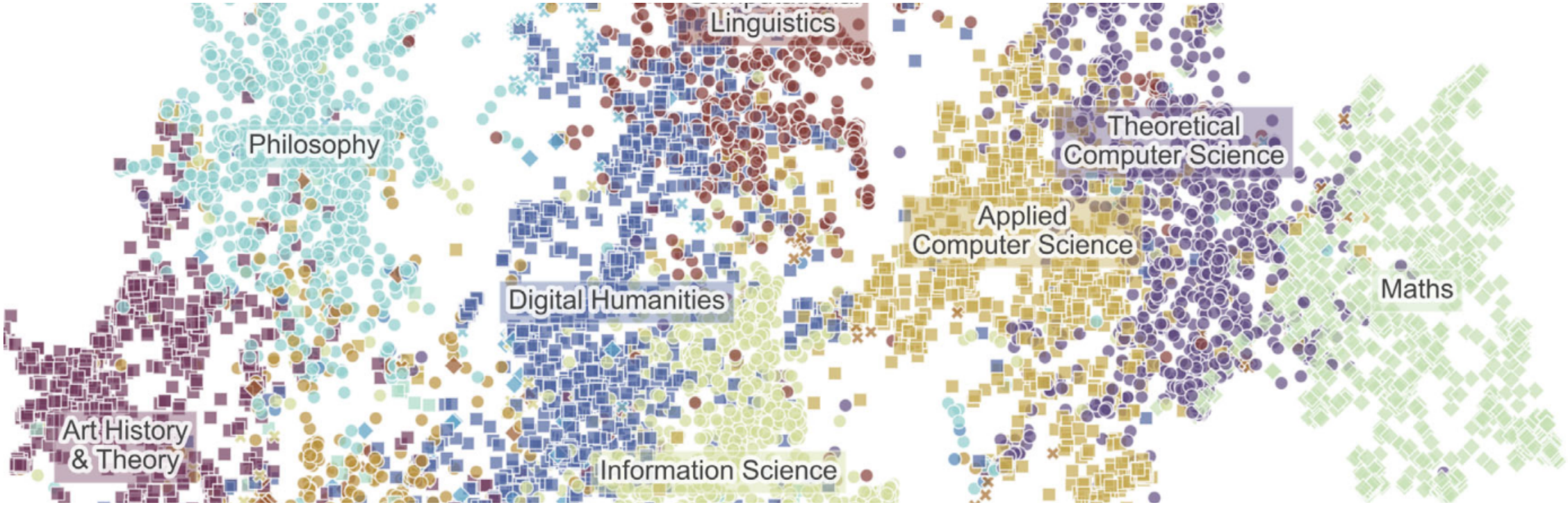




Source : BERNARDINI et PEEPLES, "Sight Communities: The Social Significance of Shared Visual Landmarks", 2015.

3. Analyse interprétative : exemples

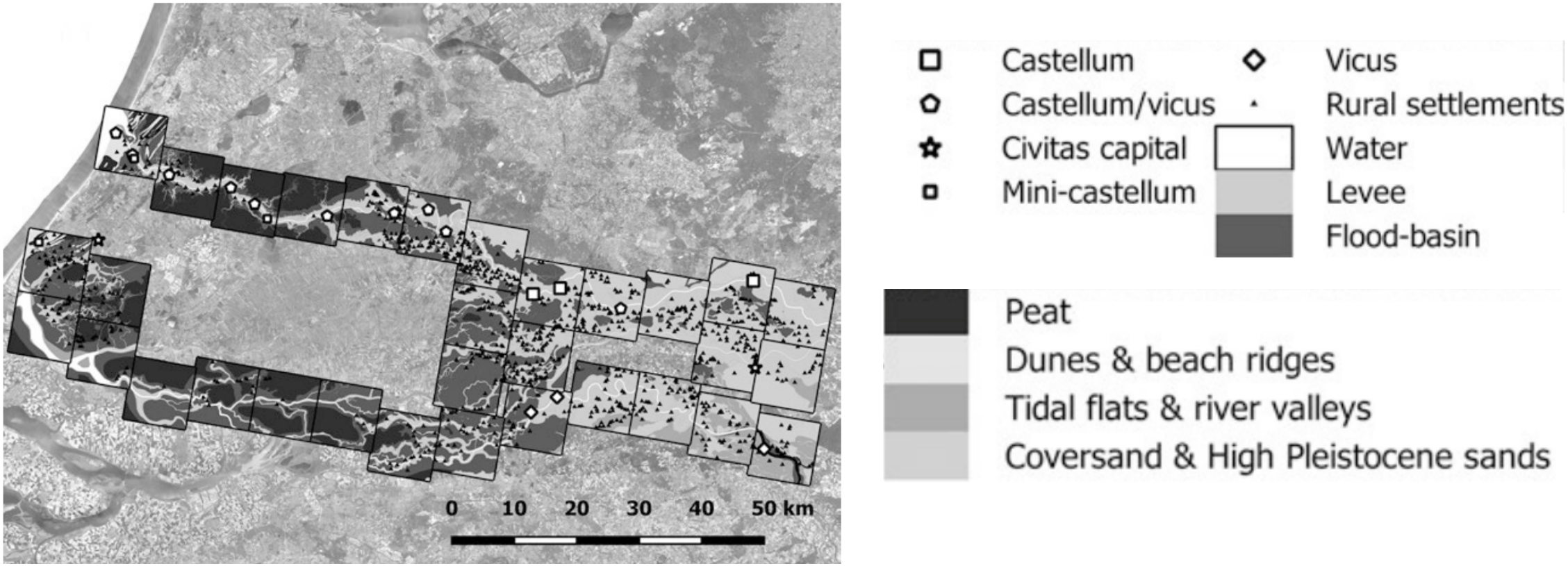
Détermination des thèmes traités le plus souvent dans les articles appartenant à une discipline et situation de cette discipline dans le panorama des disciplines scientifiques



Source : BURGHARDT et LUHMANN, “Digital humanities - A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape”, 2021.

3. Analyse interprétative : exemples

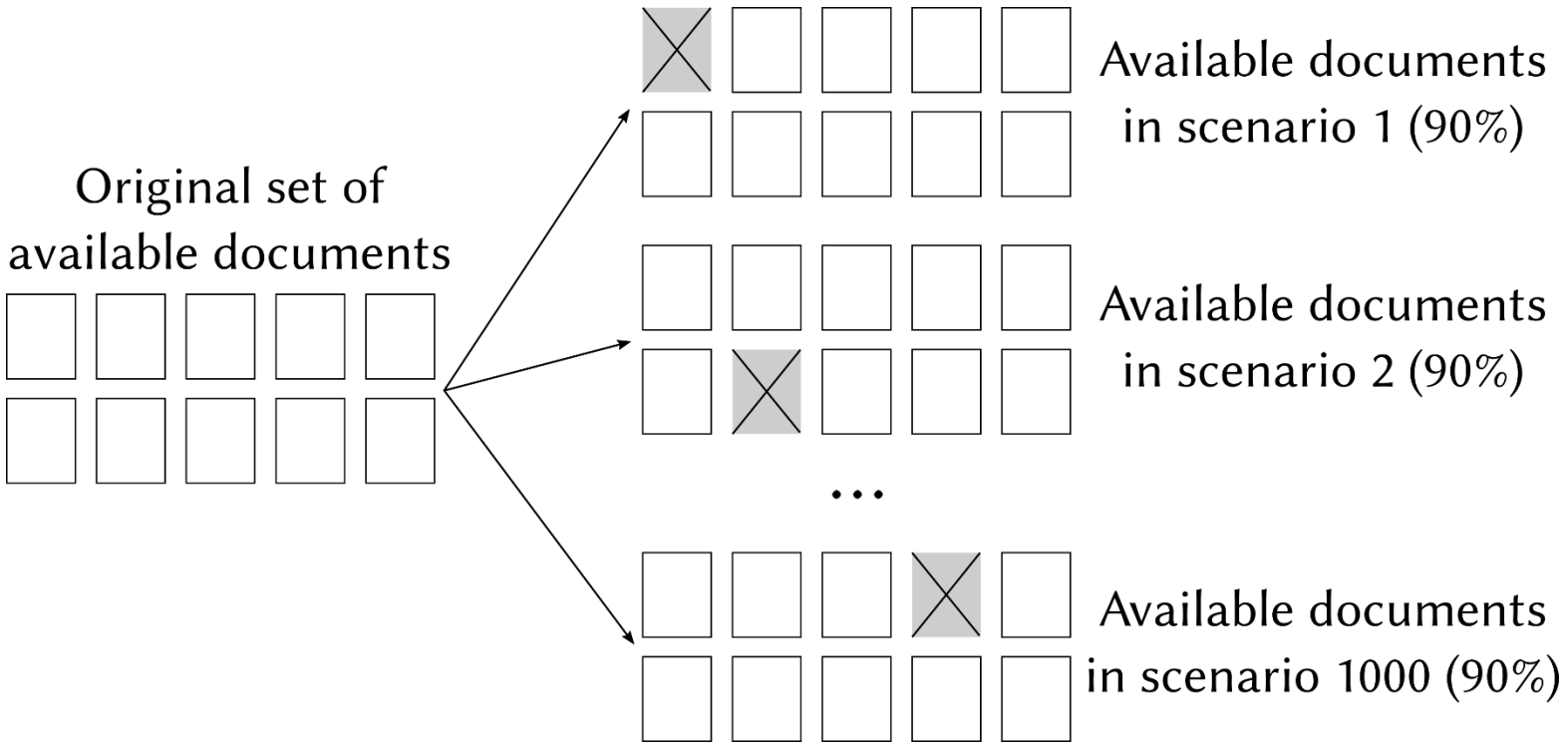
Simulation de processus de peuplement et d'occupation d'un territoire à l'aide de modèles à base d'agents (Agent-Based Modeling).

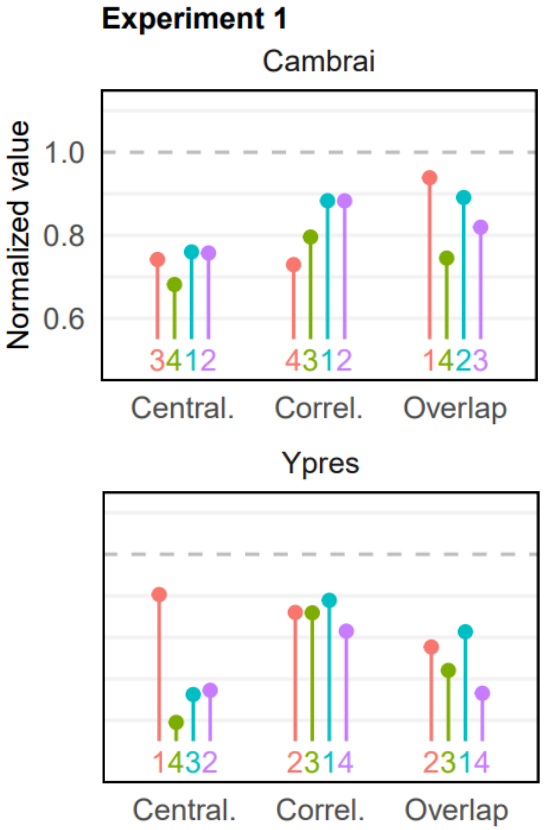


Source : JOYCE, "Modelling Agricultural Strategies in the Dutch Roman Limes via Agent-Based Modelling", 2019.

4. Exercice de la critique : exemples

Estimation de l'impact de la perte de documents et des hypothèses de travail du chercheur sur des résultats de l'analyse des réseaux

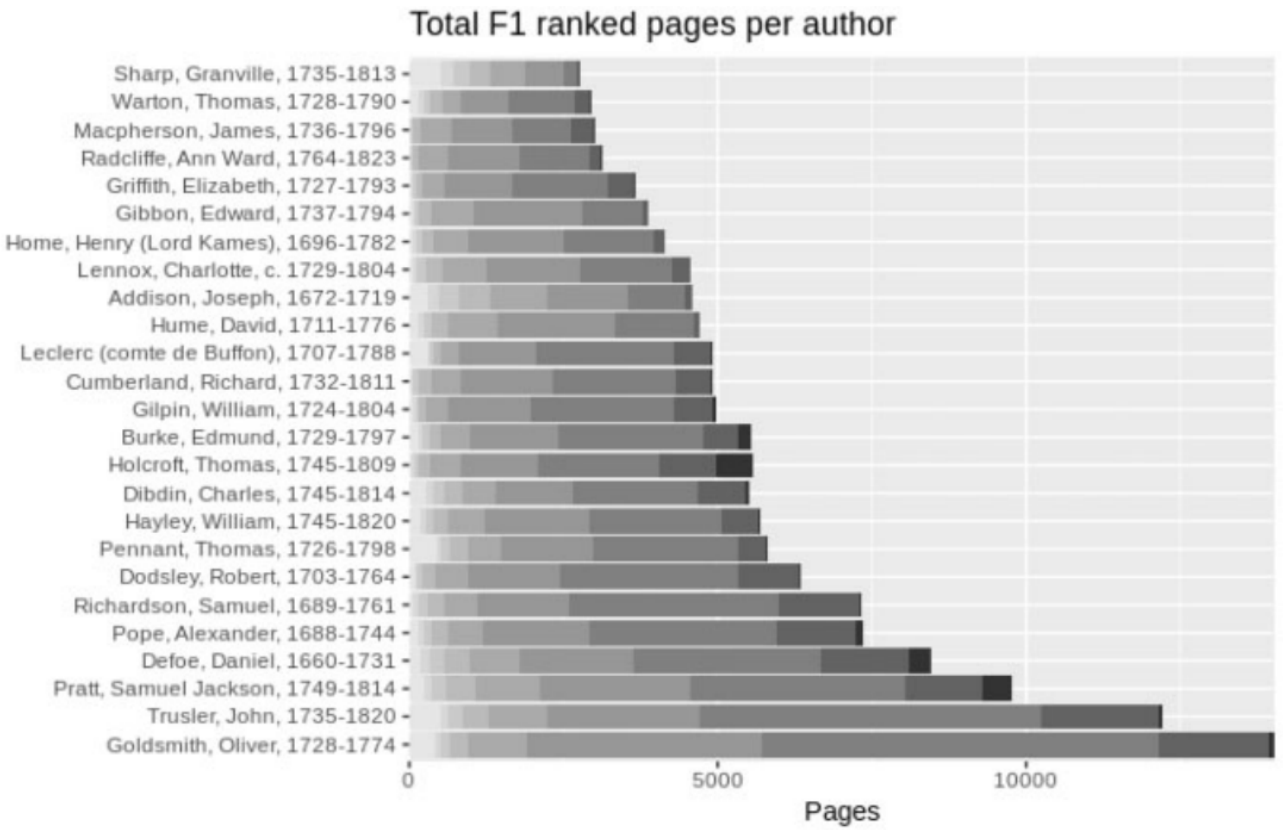


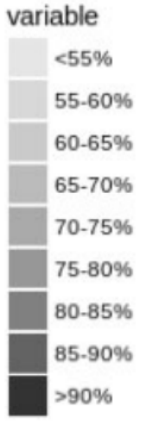


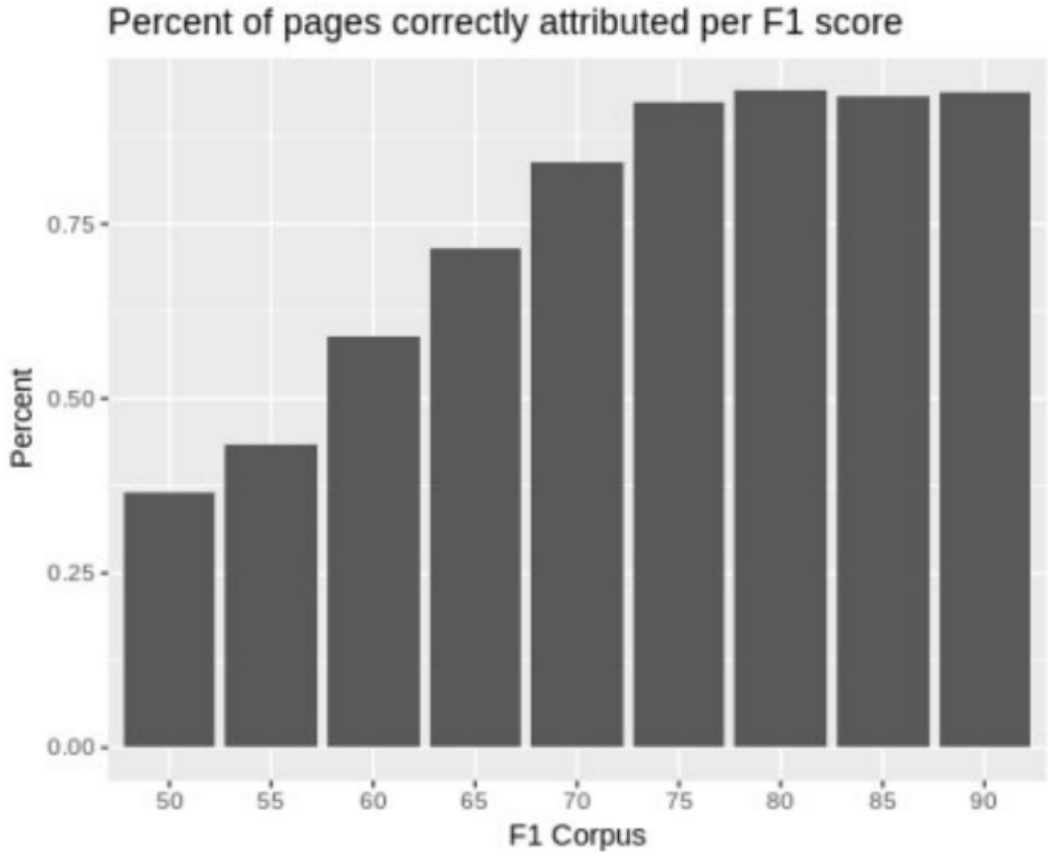
Source : DE VALERIOLA, "Can historians trust centrality? Historical network analysis and centrality metrics robustness ", 2021.

4. Exercice de la critique : exemples

Estimation de l'impact des erreurs d'OCR sur des résultats d'analyses de texte







Source : HILL et HENGCHEN, "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study", 2019.

Problèmes épistémologiques

« Some of the key techniques that are deployed in the digital humanities […] importantly raise questions for epistemology in relation to knowledge production and analysis »

BERRY et FAGERJORD, *Digital Humanities*, p. 105

« Few things will cripple the humanities more than the uncritical “adoption of tools” or the continued encroachment of positivistic research methods borrowed from cognitive science, neuroscience, computer science, or elsewhere »

GALLOWAY, “The Cybernetic Hypothesis”, p. 128

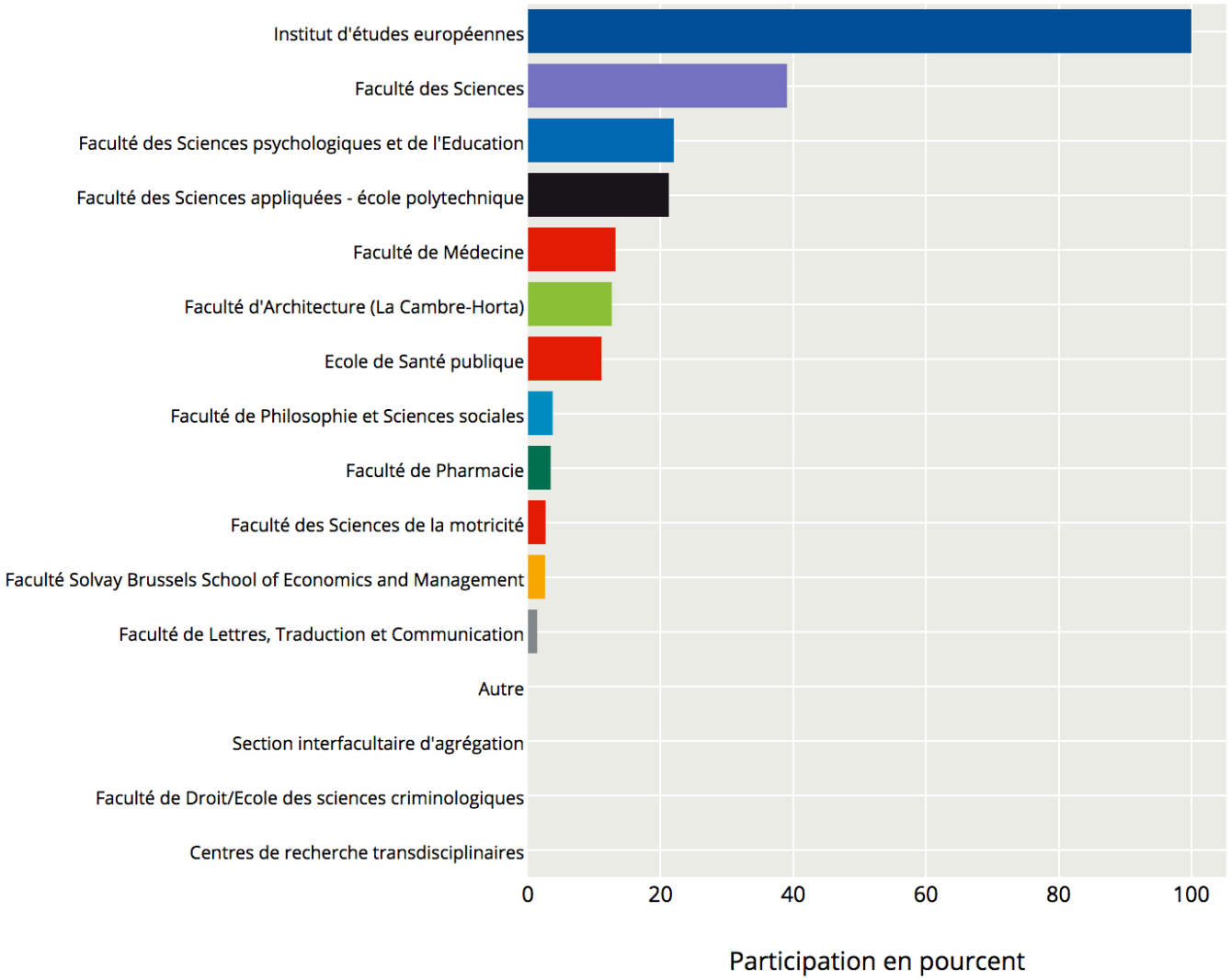
« Digital humanities can no longer afford to take its tools and methods from disciplines whose fundamental epistemological assumptions are at odds with humanistic method »

DRUCKER, “Humanities Approaches to Graphical Display”, § 6

« In fact, there is very little research on the epistemological foundations of digital humanities and on how they differ from the “traditional” humanities; in other words, research that tries to answer the question: what is the impact of computational methods on the production of knowledge in the humanities? »

PIOTROWSKI, “Epistemological Issues in Digital Humanities”, p.2

La question des données en sciences humaines



La gestion des données de recherche à l'ULB

Taux de participation par Faculté : nette différence entre les chercheurs de la Faculté des Sciences (39%) et ceux des Facultés de Philosophie et Sciences sociales (3,5%) et de Lettres, Traduction et Communication (1,5%).

Pourquoi cette différence ? Probablement, car certains chercheurs estiment qu'ils ne manipulent pas de données de recherche.

Source : https://gdr.ulb.ac.be/stat_survey.html

La question des données en sciences humaines

Pourquoi l'utilisation du terme « données » pose-t-elle problème ?

- En dehors des sciences humaines, la donnée est parfois considérée comme de nature factuelle et objective → mal adapté à nos disciplines.

- En sciences humaines, la donnée s'insère dans un contexte, dont elle est extrait. Faudrait-il parler de captées (*capta*), avec Howard Jensen (1950) et Johanna Drucker (2011) ?

- En sciences humaines, les données présentent souvent un caractère incomplet et/ou imprécis.

Johanna Drucker

La question des données en sciences humaines

Les définitions du terme « données » en sciences humaines

- Jean-Philippe Genêt (1986) : la donnée est le résultat d'un processus d'extraction hors du « réel historique ». Les données forment alors une « méta-source » soumise à l'ordinateur.

- Trevor Owens (2011) : la donnée est « un objet multiforme qui peut être mobilisé comme preuve à l'appui d'un argument ». Elle est à la fois un artefact créé par le chercheur, un texte soumis à interprétation, et une information analysable de façon quantitative.

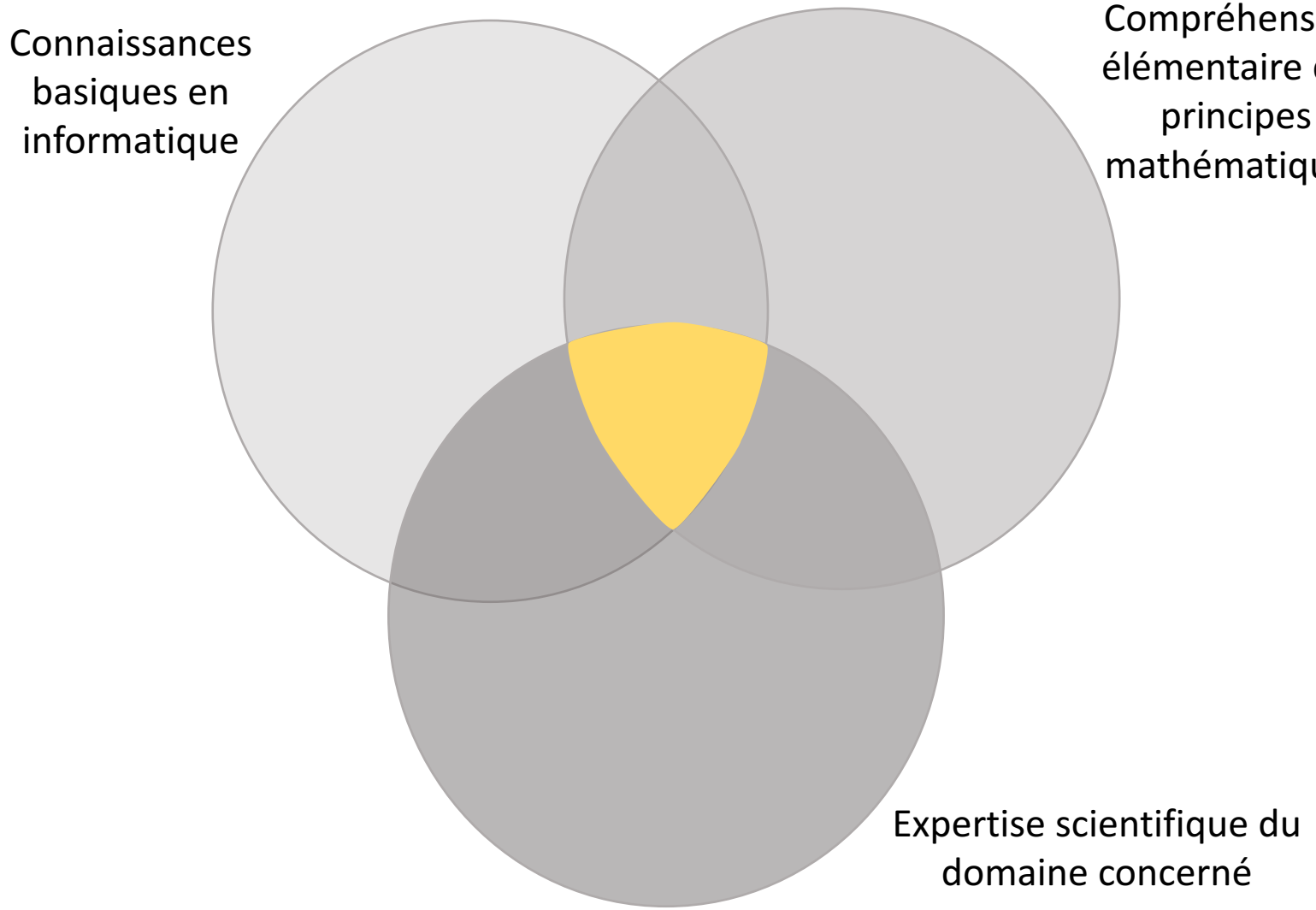
➔Insistance commune sur le processus de création/extraction des données.

➔ Importance des étapes de modélisation faisant intervenir un chercheur spécialiste du domaine.

Jean-Philippe Genet

Trevor Owens

La question des données en sciences humaines



La question des données en sciences humaines

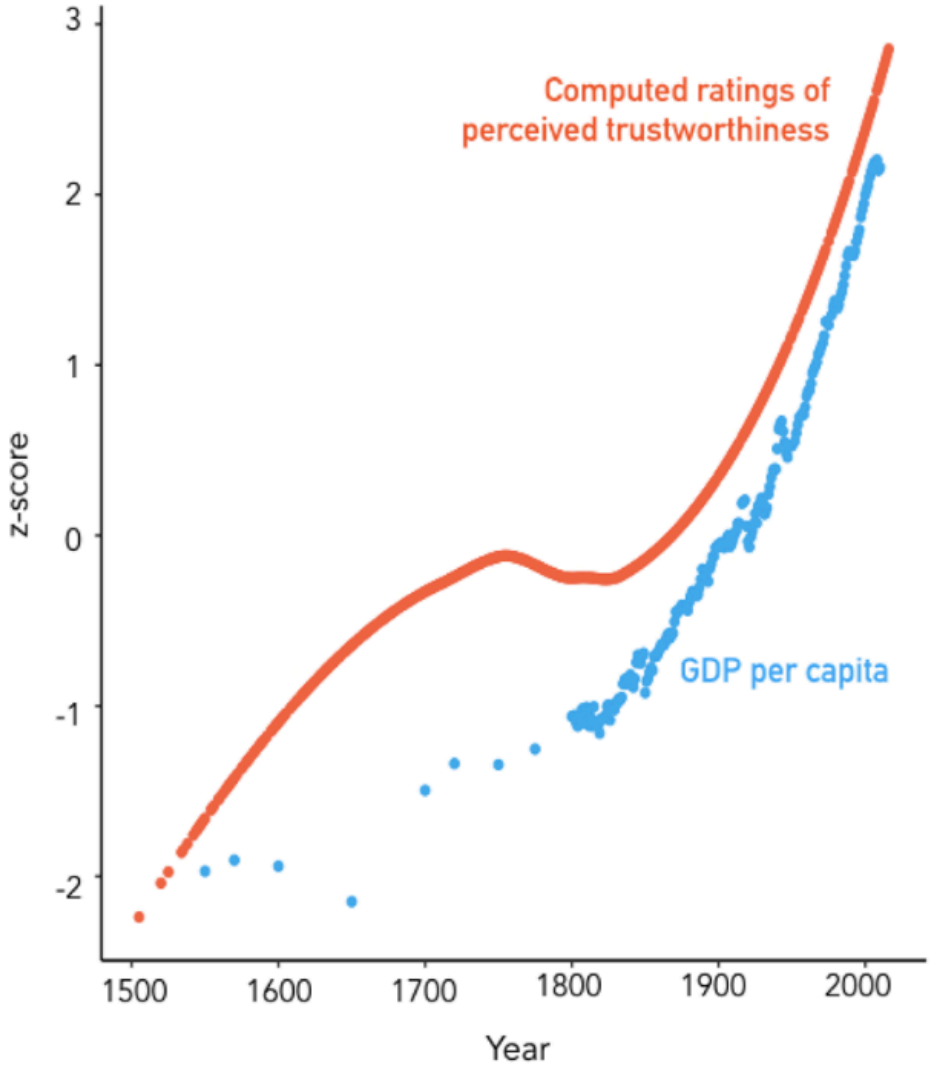
Un lien entre les caractéristiques faciales représentées sur les peintures et le niveau de vie à des époques données ?

Une équipe qui n'intègre pas d'historien de l'art…

Source : SAFRA, CHEVALLIER, GRÈZES et BAUMARD, "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings", 2020.

Portrait with low computed ratings of perceived trustworthiness

Portrait with high computed ratings of perceived trustworthiness



Autres questions qui pourraient être évoquées

- Question de la « vérité absolue »
  - Classification et étiquetage des données
  - Régression et « vérité absolue » (*ground truth*)
  - Applicabilité en sciences humaines ?

- Question de la preuve
  - Prouver un fait ne correspond pas à la même démarche dans toutes les disciplines
  - Confrontation du discours scientifique des sciences humaines avec des techniques conçues pour les sciences exactes et avec des exigences propres du point de vue de l'administration de la preuve.
  - Question de la « réplicabilité »

Autres questions qui pourraient être évoquées

- Question de l'universalité et de la nécessité de particulariser
  - Les méthodes quantitatives sont-elles vraiment agnostiques vis-à-vis des problèmes traités ?
  - L'approche *one size fits all* convient-elle toujours ou les techniques doivent-elles être adaptées à chaque discipline ?

- Question du formalisme
  - Pour les approches quantitatives, le formalisme est indispensable…
  - … mais cela s'applique-t-il bien aux sciences humaines où, pour des raisons pratiques, les concepts se doivent parfois d'être plus "flous" ou "mouvants" ?

Autres questions qui pourraient être évoquées

- Question de la quantité des données
  - Beaucoup de ces méthodes ont été conçues pour manipuler de grandes quantités de données (*big data*).
  - Dans certaines disciplines de sciences humaines, on n'atteindra sans doute jamais ce seuil critique.
  - Certaines méthodes sont-elles dès lors à proscrire, parce que la masse critique de données nécessaires à leur calibration ne sera jamais atteinte ?

- Question de la gestion des données
  - Les pratiques de gestion des données des chercheurs en sciences humaines doivent être identiques à celles des chercheurs en sciences exactes ?
  - Quid de la réutilisation des données, qui est une pratique encore peu diffusée en sciences humaines ?
  - Comment garantir une vraie reconnaissance du travail de préparation des données ?

- …

Un grand merci pour votre attention, place au débat !