CY Cergy Paris Université
DUDA 2023

Visualisation des données : UE 2

Auteur :
Francisco Martin-Gomez

Référent :
Pr. Mathieu CISEL

23 août 2023

Résumé

Cette étude est un cas pratique synthétique des principaux enjeux de la data visualisation pour un data analyst. À ce titre, nous y présenterons de manière détaillée deux figures archétypiques : un *area chart* présentant l'évolution des thèses soutenues par discipline entre 1985 et 2018, et un *histogramme* présentant la distribution des langues utilisées pour la rédaction de ces thèses. L'ensemble des données, codes et figures est disponible en ligne.

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte et objectifs

Cette étude a porté sur les soutenances de thèses entre 1984 et 2020. Axée tout particulièrement sur les enjeux et les contraintes de visualisation des données en analyse de données (ou data science), cette étude s'est attachée à mettre en lumière l'évolution du nombre de thèses soutenues en fonction de leurs disciplines universitaires, en interrogeant parallèlement l'évolution des langues utilisées pour leurs rédactions.

Organisation du rapport

Ce rapport est organisé en deux sections :

La première section, « **Présentation des données** », s'attachera à décrire rapidement le jeu de données à notre disposition (PhD_v3.csv).

La deuxième section, « **Résultats préliminaires** », montrera dans une première partie l'évolution des thèses soutenues par disciplines universitaires. Dans une seconde partie, nous montrerons l'évolution des langues choisies pour la rédaction de ces thèses.

1 Présentation des données

— *Le jeu de données PhD_ v3 (.csv) est disponible en ligne à l'adresse suivante : lien vers les données (drive.google)*
— *Le code et les graphiques sont disponibles en ligne à l'adresse suivante : lien vers le dépôt github.com*

Le Tableau 1 présente l'ensemble des variables et des observations du jeu de données PhD_v3, issu d'un travail de scrapping à partir du site www.theses.fr (Cisel *et al.*, 2015). Ce jeu de données est composé de 23 variables (21 de type *charactère*, et 2 de type *numérique*) avec 448047 observations (ici chaque ligne/observation correspond à une thèse répertoriée dans la base de données du site). Le jeu de données reflète le référencement officiel des thèses soutenues ou en cours dans les universités françaises, entre 1984 et 2018.

Ce jeu de données est une version actualisée et enrichie du jeu de données PhD_v2 ayant fait l'objet d'une analyse détaillée dans un précédent rapport (cf. Manipulation et pré-traitement des données : UE1). A ce titre, considérant l'objet principal de cette étude, nous faisons le choix de reproduire la méthodologie de nettoyage et de pré-traitement précédemment mise en œuvre et exposée dans le rapport de l'UE1. Pour des raisons de clarté et de lisibilité, nous n'exposerons pas dans ce rapport les étapes et les résultats de ce travail de préparation. Le code et les résultats du wrangling sur le jeu de donnée PhD_v3 sont disponibles sur le github présenté plus haut. Le jeu de données nettoyé est également disponible sur le drive (PhD_v3_clean).

Nous attirons cependant l'attention du lecteur sur la présence de nouvelles variables calculées, i.e. non directement issues du scrapping du site www.these.fr mais construites à partir du travail d'imputation de Cisel et ses collaborateurs (Cisel *et al.*, 2020, 2021, 2022). Ainsi, apparaissent dans ce dataset les variables "Genre" (retraçant le genre des auteurs) et "Discipline_predi". Cette dernière est une variable construite par machine learning (avec imputation complémentaire manuelle lorsque cela fut possible) à partir de la nomenclature officielle des disciplines et sections du CNU (Conseil National des Universités) afin de réduire la dimensionnalité anormale de la variable d'origine "Discipline". Cette dernière variable présente en effet un nombre anormalement élevé de catégories et de sous-catégories (24262, Tableau 1), provenant d'une indexation humaine libre lors des dépôts des manuscrits. Il en résulte une difficile comparaison dans le temps des productions universitaires selon les disciplines, puisqu'il est possible avec cette modalité de gestion de créer une discipline/sous-disciplines universitaire unique par thèse. Le recours à une ontologie contrôlée basée sur les sections CNU permet une nette réduction des catégories d'indexation (15, Tableau 1).

Table 1 – Présentation des variables de l'ensemble de données des thèse (PhD_v3)

Variables	Nbr_NA	Type	Nbr_Levels
Auteur	3	character	
Identifiant auteur	129983	character	
Titre	11	character	
Directeur de these	15	character	
Directeur de these (nom_prenom)	15	character	
Identifiant directeur	49168	character	
Etablissement de soutenance	1	character	
Identifiant etablissement	17082	character	
Discipline	0	factor	24262
Statut	0	character	
Date de premiere inscription en doctorat	383991	Date	
Date de soutenance	221165	Date	
Year	56738	numeric	
Langue de la these	63760	character	
Identifiant de la these	0	character	
Accessible en ligne	0	character	
Publication dans theses.fr	0	character	
Mise a jour dans theses.fr	177	character	
Discipline_predi	0	factor	15
Genre	0	factor	6
etablissement_rec	3723	character	
Langue_rec	63760	factor	4

2 Résultats préliminaires

2.1 Évolution des thèses par disciplines

Le graphique 1 (p.4) illustre l'évolution des thèses par disciplines au fil des années (1985 - 2018). Le fait le plus remarquable est la très forte diminution du nombre de thèses en *Médecine* durant cette période. Alors que ces dernières représentaient jusqu'à 35,40 % des thèses référencées en 1989, année du pic de dépôt pour cette matière, les thèses en *Médecine* ne représentent en 2018 que 1,20 % des thèses soutenues (voir Tableau 4, en Annexes, pour plus de détails). Cette forte diminution du nombre de thèses s'explique selon nous par une modification des comportements de référencement au sein des universités, les thèses en médecine ne correspondant pas aux standards des autres thèses scientifique (durée d'un an, portant sur un cas pratique ou exercice hospitalier, entre 50 et 100 pages en moyenne, considérée comme plus proche d'un mémoire de Master et ne relevant pas de la recherche du point de vue méthodologique).Notons que pour l'année 2018, il y avait encore 158 thèses de *Médecine* référencées sur thèse.fr.

Comme le soulignent la Figure 1, la *Biologie* est la matière universitaire connaissant le plus grand nombre de thèses soutenues entre 1985 et 2018. Le nombre de thèse en *Biologie* est en augmentation sur la période, passant de 695 thèses en 1985 (1ère position), à 2394 en 2002 (1ère position), et 3245 en 2018 (1ère position) (Tableau 2). Sur la période 1985-2018, si on excepte la *Médecine* dont le cas é été exposé plus haut, les trois principales disciplines en termes de thèses soutenues sont donc la *Biologie*, les *SHS* et les *Matériaux, Milieux et Chimie*. A ce titre, la plus forte progression de thèses soutenues est celle de thèses soutenues en *Materiaux, Milieux et Chimie*, passant de 331 thèses en 1985 (4ème position), à 1779 en 2002 (2ème position) et 2469 en 2018 (2ème position) (Tableau 2)

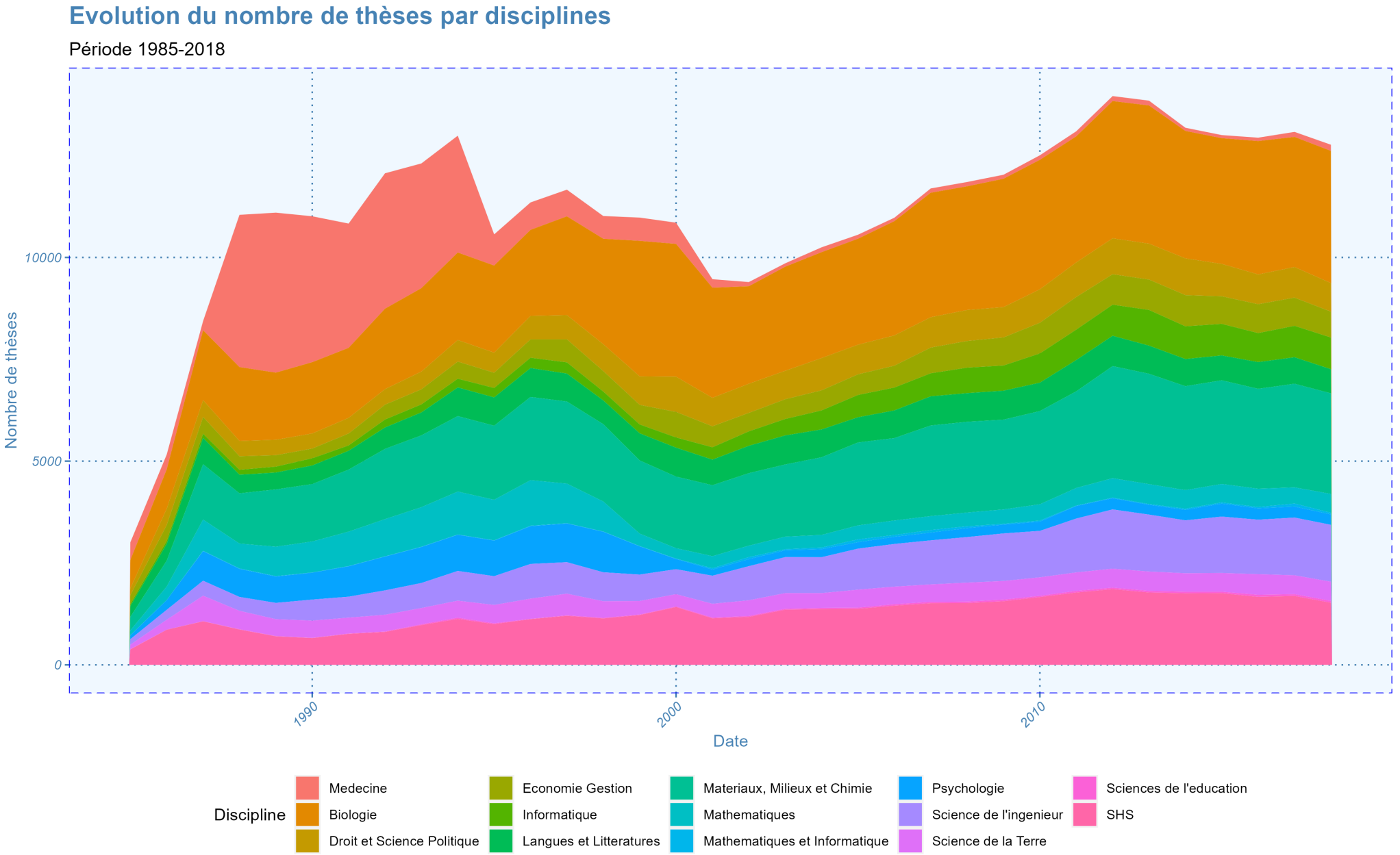

FIGURE 1 – Evolution du nombtre de thèses soutenues par disciplines entre 1985 et 2018

Table 2 – Nombre de thèses soutenues en 1985, 2002 et 2018 : les 5 disciplines les plus référencées

Années	Discipline_predi	Thèses par discipline
1985.00	Biologie	695
1985.00	Medecine	434
1985.00	SHS	380
1985.00	Materiaux, Milieux et Chimie	331
1985.00	Langues et Litteratures	268
2002.00	Biologie	2394
2002.00	Materiaux, Milieux et Chimie	1779
2002.00	SHS	1177
2002.00	Science de l'ingenieur	837
2002.00	Droit et Science Politique	715
2018.00	Biologie	3245
2018.00	Materiaux, Milieux et Chimie	2469
2018.00	SHS	1525
2018.00	Science de l'ingenieur	1396
2018.00	Informatique	777

2.2 Evolutions des langues de rédactions des thèses

La croissance des thèses rédigée en anglais (11 en 1985, 177 en 2002 et 3429 en 2018 ; Tableau 3) reste cependant très forte et s'explique selon nous par l'ouverture internationale de la recherche française. Les financements des projets de recherche sont par exemple fortement dépendant de la dimension collaborative à l'échelle européenne et mondiale, faisant de l'anglais une seconde langue de référence pour la rédaction des travaux de recherche en cotutelle internationale. Il est également possible de voir un effet conjoint de deux phénomènes à l'œuvre dans les milieux académiques : une montée en compétences linguistique des étudiants français (qui pratiquent davantage et mieux l'anglais que leurs prédécesseurs) et la nécessité de plus en plus forte de publier qui pèsent sur les doctorants désireux de devenir chercheurs. En effet, ces derniers sont souvent incités à rédiger de nombreux manuscrits en anglais (des articles basés sur leurs travaux de thèses), et il est possible de penser que rédiger une thèse en anglais représente un gain de temps et d'efforts.

Table 3 – Nombre de thèses soutenue par langue de rédaction, en 1985, 2002 et 2018

Année	Langue_redac	Nb. de thèses
1985.00	Français	2917
1985.00	Bilingue	73
1985.00	Anglais	11
1985.00	Autre	6
2002.00	Français	8671
2002.00	Bilingue	404
2002.00	Anglais	177
2002.00	Autre	100
2002.00		44
2018.00	Français	7807
2018.00	Anglais	3429
2018.00	Bilingue	741
2018.00		673
2018.00	Autre	155

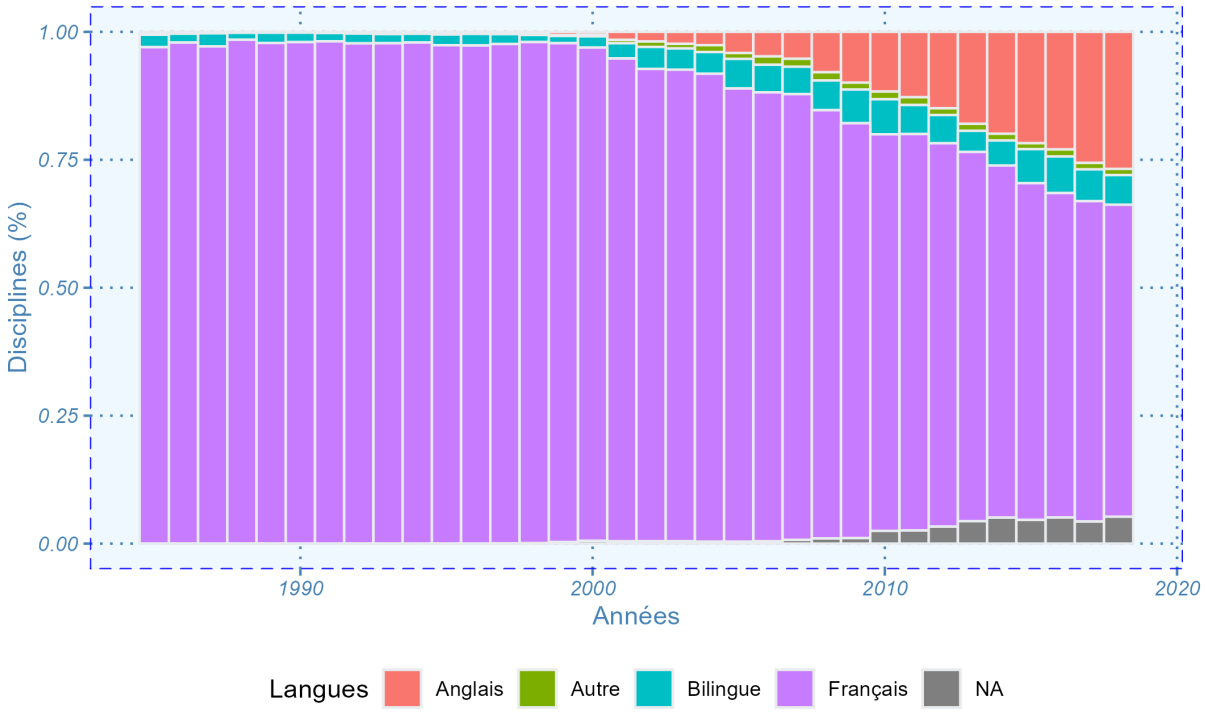

FIGURE 2 – Evolution des langues de rédactions des thèses soutenues entre 1985 et 2018

3 Annexes

TABLE 4 – Evolution de la discipline Medecine entre 1985 et 2018

Année	Discipline	Nb. these par discipline/an	Total annuel de theses	Proportion
1985.00	Medecine	434	3007	14.43
1986.00	Medecine	364	5162	7.05
1987.00	Medecine	225	8439	2.67
1988.00	Medecine	3736	11045	33.83
1989.00	Medecine	3926	11102	35.36
1990.00	Medecine	3581	11011	32.52
1991.00	Medecine	3050	10831	28.16
1992.00	Medecine	3323	12065	27.54
1993.00	Medecine	3061	12309	24.87
1994.00	Medecine	2867	12991	22.07
1995.00	Medecine	765	10569	7.24
1996.00	Medecine	673	11354	5.93
1997.00	Medecine	654	11669	5.60
1998.00	Medecine	557	11023	5.05
1999.00	Medecine	571	10982	5.20
2000.00	Medecine	520	10855	4.79
2001.00	Medecine	209	9468	2.21
2002.00	Medecine	102	9396	1.09
2003.00	Medecine	83	9857	0.84
2004.00	Medecine	119	10250	1.16
2005.00	Medecine	98	10562	0.93
2006.00	Medecine	84	10975	0.77
2007.00	Medecine	105	11697	0.90
2008.00	Medecine	105	11854	0.89
2009.00	Medecine	98	12033	0.81
2010.00	Medecine	111	12516	0.89
2011.00	Medecine	124	13110	0.95
2012.00	Medecine	122	13985	0.87
2013.00	Medecine	118	13868	0.85
2014.00	Medecine	78	13202	0.59
2015.00	Medecine	76	13023	0.58
2016.00	Medecine	83	12965	0.64
2017.00	Medecine	123	13123	0.94
2018.00	Medecine	151	12805	1.18

4 Références

Jones, L. (2018). *Deep Learning for Natural Language Processing* (thèse de doct.). University of Techland.

Miller, A. (2020). *Introduction to Data Science*. Tech Publishers.

Smith, J., & Doe, J. (2021). Data visualization techniques in modern research. *Journal of Modern Data Science*, *1*(1), 1-12.

Williams, R., & Thompson, M. (2019). Neural networks and their applications in image recognition. *Proceedings of the 5th International Conference on Machine Learning*, 224-230.