(Introduction au) TAL/NLP avec Python

De la lexicométrie aux modèles pré-entrainés

Émilien Schultz

La situation actuelle

Une période troublée : révolution des LLM & explosion des usages

- mais des approches très variées
- venant d'époques différentes
- avec des outils différents

Les LLM sont-ils la solution pour tout ?

- Principales tâches réalisables avec LLM
- Mais des limites :
  - Coûts
  - Fiabilité
  - Biais

Section 1

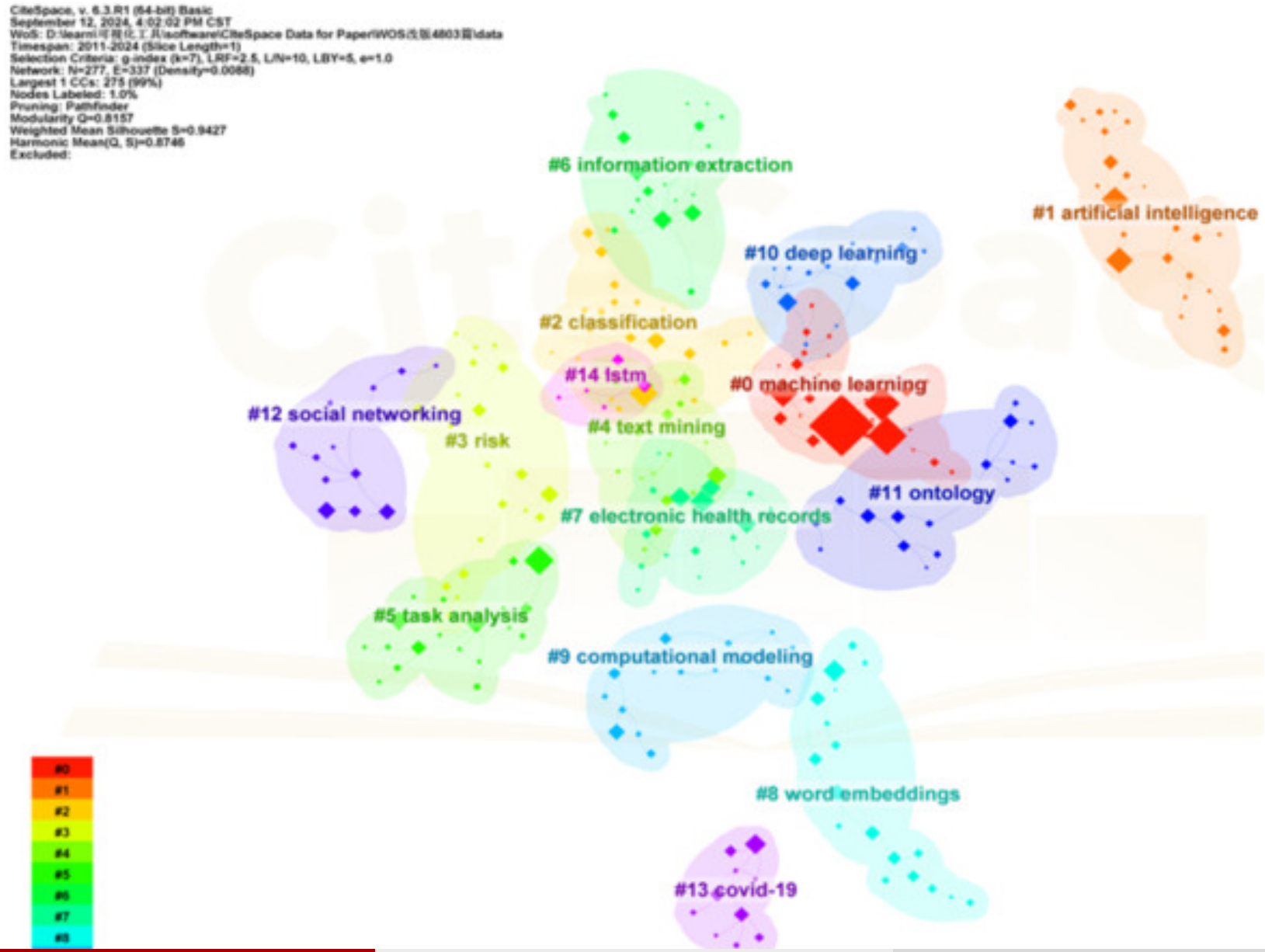
Vous avez dit NLP ?

Le texte : données non structurées

- *Au début* : une suite de caractères
  - Avec un encodage : binary symbole
- *À la fin* : du sens intelligible
  - Proches de l'humain

Entre : comment structurer l'information ?

Un domaine interdisciplinaire



Des techniques différentes

- **Symboliques** ancrage linguistique
  - se baser sur les règles / théorie du language
- **Statistiques** (compter des mots)
  - un peu *old school*
- **Machine Learning** sur données textuelles vectorisées
- Évolution vers les **représentations**
  - *modèles de langages*
  - Approches potentiellement plus proche du langage naturel

Chacune a ses particularités/limites/coûts/outils

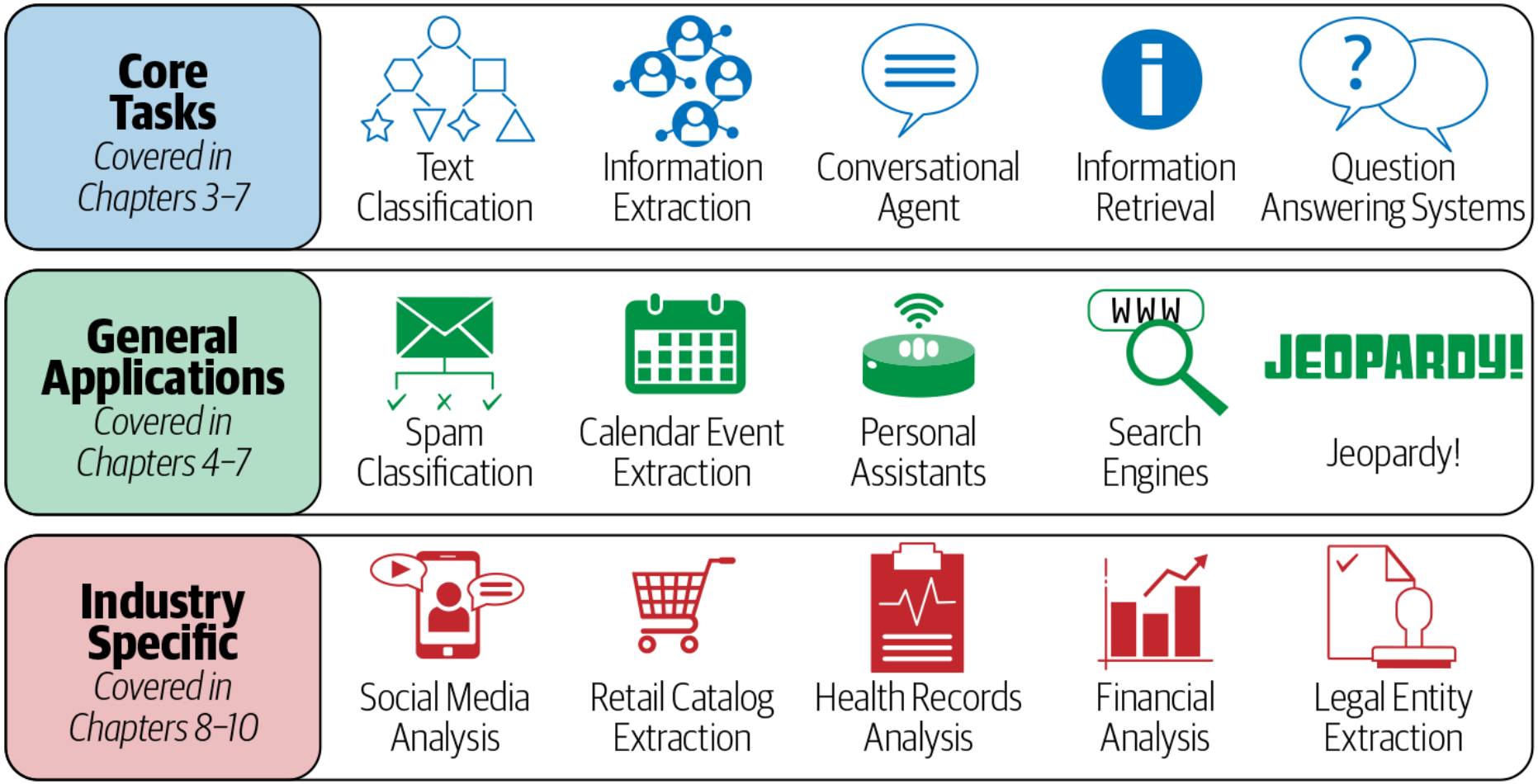
Quoi faire avec du texte ?

Des traitements très variés

- Classifier des textes
  - Faire des groupes
  - Retrouver des éléments
- Identifier des éléments spécifiques
  - Noms propres, etc.
- Générer des textes ...

Un domaine découpé en tâches

Pour de nombreuses applications



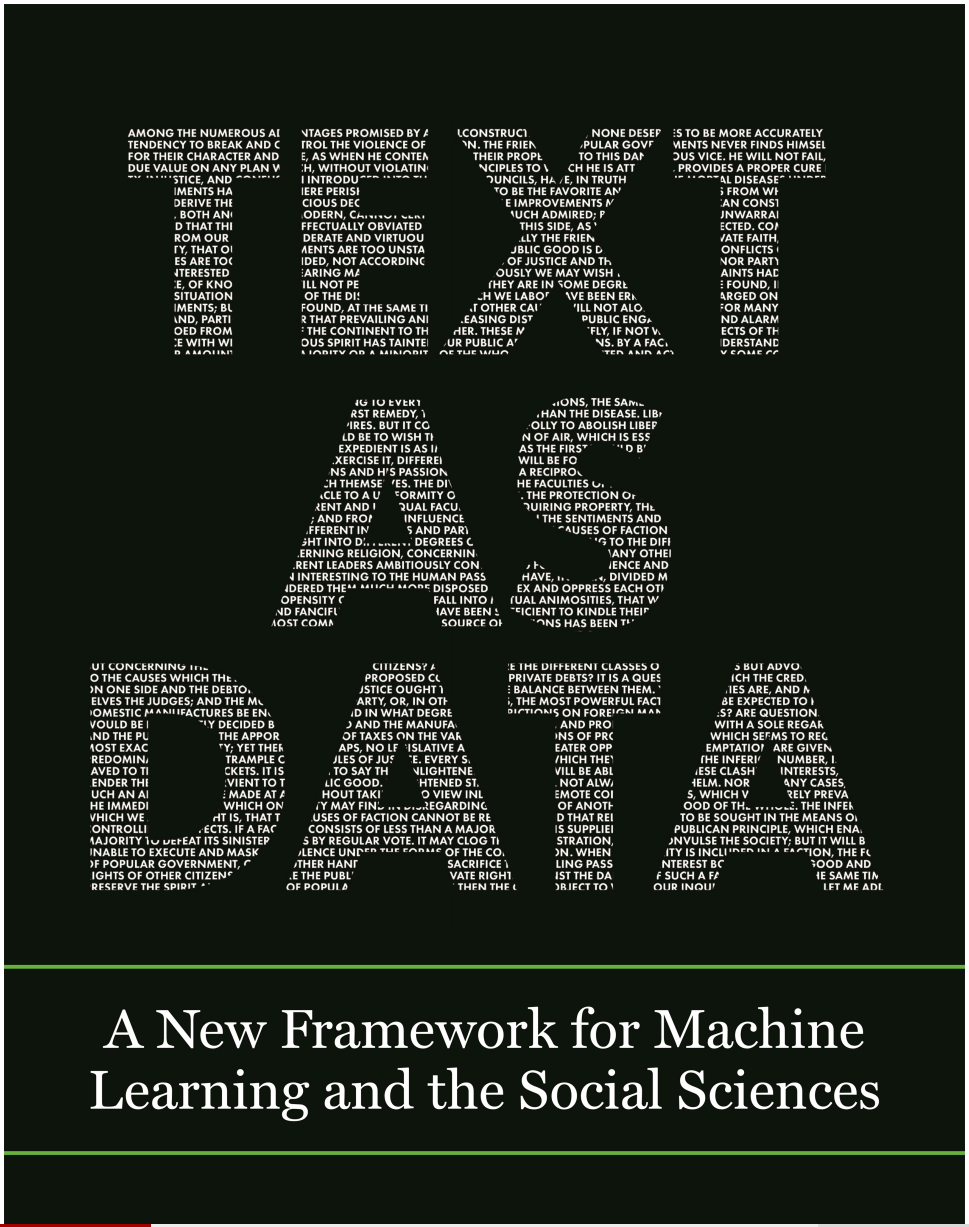
Spécificités de la recherche

Du NLP dans de nombreux domaines scientifiques

Différences fortes :

- humanités numériques : corpus bien structurées, textes anciens
- sciences sociales, ...computationnelles : grands corpus de réseaux sociaux, ...

Des méthodes qui doivent être intégrées dans les enquêtes

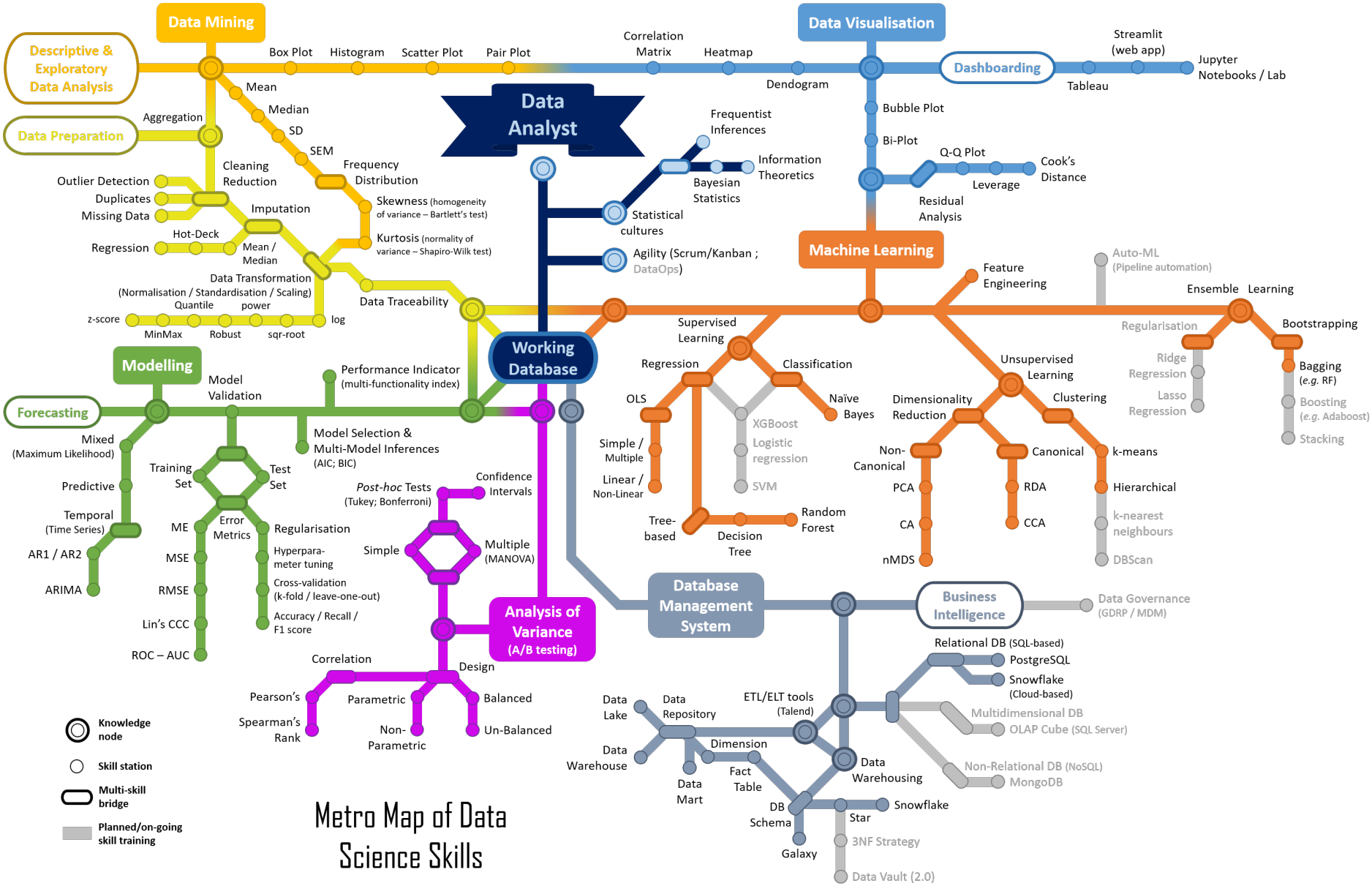


Donc

- Pour une même tâche, des techniques très variées
  - par exemple identifier les thématiques
- Pour une techique, des applications très différentes
  - retrouver des éléments vs. les visualiser
- Des métriques d'évaluation différentes
  - F1, corrélation, etc.
- Des mises en oeuvres multiples
  - Logiciels, programmation, etc.

Difficile de tout faire :)

Rappel de l'extension des méthodes



Quelques exemples

New Media & Society

Impact Factor: **4.5** / 5-Year Impact Factor: **7.0**

Open access | Research article | First published online June 6, 2023

Men who hate women: The misogyny of involuntarily celibate men

Michael Halpin, Norann Richard, [...], and Finlay Maguire +2 View all authors and affiliations

Volume 27, Issue 1 | https://doi.org/10.1177/14614448231176777

Abstract

This article uses computational data and social science theories to analyze the misogynistic discourse of the involuntary celibate ("incel") community. We analyzed every comment ($N = 3{,}686{,}110$) produced over 42 months on a popular incel discussion board and found that nearly all active participants use misogynistic terms. Participants used misogynistic terms nearly one million times and at a rate 2.4 times greater than their use of neutral terms for women. The majority of participants' use of misogynistic terms does not increase or decrease with post frequency, suggesting that members arrive (rather than become)

Une méthode très classique

Data and methods

This article examines the discussion board on incels.is, a popular English language incel website. At the time of our data collection (15 April 2021), the site had 13,700 registered members who produced nearly 6 million comments and spent more than 54,000 days on the website. These numbers do not count people who view the site without commenting; the site receives several million visits per month (Similarweb, 2021).

Procedure

We collected all posts (e.g. text posted by a user) that appeared on the incels.is discussion board ("Inceldom Discussion") from 8 November 2017 until 16 April 2021 ($N = 3{,}686{,}110$). We employed custom scraping scripts[1] to extract all public post data. This script automatically traversed and downloaded the corresponding HTML for every page of posts within each thread of "Inceldom Discussion." Individual comments and user information were then extracted from the HTML and saved as thread-specific text files. The extracted data include post text along with the associated participant user IDs, posting time, thread titles, threads, and the respective order of posts in threads. Our analyses do not include posts that were deleted before we completed our data collection and excludes text quoting other posts in the same thread.

Analytic approach

We use computational methods to collect and describe incel misogyny, while using social science theories to inform our research questions, as well as interpret and discuss our results. In this sense, we aim to leverage

Exemple de topic analysis

American Political Science Review (2019) 113, 4, 883–901
doi:10.1017/S0003055419000352

Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data

PABLO BARBERÁ *University of Southern California*
ANDREU CASAS *New York University*
JONATHAN NAGLER *New York University*
PATRICK J. EGAN *New York University*
RICHARD BONNEAU *New York University*
JOHN T. JOST *New York University*
JOSHUA A. TUCKER *New York University*

Are legislators responsive to the priorities of the public? Research demonstrates a strong correspondence between the issues about which the public cares and the issues addressed by politicians, but conclusive evidence about who leads whom in setting the political agenda has yet to be uncovered. We answer this question with fine-grained temporal analyses of Twitter messages by legislators and the public during the 113th US Congress. After employing an unsupervised method that classifies tweets sent by legislators and citizens into topics, we use vector autoregression models to explore whose priorities more strongly predict the relationship between citizens and politicians. We find that legislators are more likely to follow, than to lead, discussion of public issues, results that hold even after controlling for the agenda-setting

Utiliser de l'apprentissage

La part du genre

Genre et approche intersectionnelle dans les revues de sciences sociales françaises au XXIe siècle

Résumé
Quand mobilise-t-on une perspective de genre dans les publications de sciences sociales françaises, et qui le fait ? Recourt-on davantage à cette approche que par le passé, et cette dernière a-t-elle remplacé d'autres grilles d'analyse comme on l'entend parfois ? Parce qu'il est impossible de consulter l'ensemble de la production universitaire sur une longue période, les spéculations l'emportent souvent sur les analyses empiriques. Afin d'apporter une contribution à ces réflexions, ainsi qu'à l'histoire et à la sociologie des sciences sociales, cet article s'appuie sur des outils issus du traitement automatique des langues – des *grands modèles de langage (LLMs)* – pour repérer les invocations du genre dans un vaste corpus. Appliqué aux articles scientifiques de près de 120 revues de sciences sociales françaises publiées depuis le début du siècle, il offre plusieurs résultats parlants. Le recours à une approche de genre a indéniablement progressé au cours des vingt-cinq dernières années, même si le point de départ était souvent bas, et qu'il reste finalement limité. Des différences marquées existent entre disciplines, tout comme diffèrent les formes d'institutionnalisation de cette approche. Cette perspective est, en outre, encore très majoritairement mobilisée par des femmes. Enfin, l'approche intersectionnelle augmente légèrement, tout en restant très minoritaire, et sans que le genre ne se substitue à la classe.

Mots-clés : Genre, Classe, Sociologie des sciences sociales, Intelligence artificielle, Approche intersectionnelle.

Des besoins différents

digital humanities quarterly

submissions | about dhq | dhq people | news | contact Search

2024
Volume 18 Number 3

2024 18.3 | XML | PDF | Print

Towards a differentiated digital-hermeneutic analysis tool for the detection of short quotations using the example of the Church Father Jerome

Franziska Schropp <franziska_dot_schropp_at_uni-konstanz_dot_de>, University of Konstanz
Thomas E. Konrad <thomas_dot_eugen_dot_konrad_at_uni-konstanz_dot_de>, University of Konstanz
https://orcid.org/0000-0002-0568-9420
Marie Revellio <marie_dot_revellio_at_uni-konstanz_dot_de>, University of Konstanz
Barbara Feichtinger <barbara_dot_feichtinger_at_uni-konstanz_dot_de>, University of Konstanz

Abstract

Late Latin literature is characterized by numerous references to classical texts and authors. For Jerome of Stridon in particular, manual-hermeneutic research has revealed various intertextuality phenomena usually published in encyclopaedic collections of quotations. In this paper, we present a digital-hermeneutic analysis toolkit primarily designed to detect *short* text-text congruencies that have a high chance of being evaluated as an intentional quotation. We favour a mixed-methods approach, which is based on findings from manual-hermeneutic research. Our aim is to focus on Jerome's citation technique: Based on hermeneutic analysis of confirmed quotations, we formulate differentiated criteria that lead to a deeper understanding of the phenomenon of quoting and thus also have the potential to optimize our toolkit.

Et les promesses des LLM

De nombreuses promesses

PNAS OPINION

Check for updates

Large Language Models based on historical text could offer informative tools for behavioral science

Michael E. W. Varnum[a,1], Nicolas Baumard[b], Mohammad Atari[c], and Kurt Gray[d]

Et leurs dangers

Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation

Joachim Baumann[1], Paul Röttger[1], Aleksandra Urman[2], Albert Wendsjö[3], Flor Miriam Plaza-del-Arco[4], Johannes B. Gruber[5], and Dirk Hovy[1]

[1]*Bocconi University*
[2]*University of Zurich*
[3]*University of Gothenburg*
[4]*LIACS, Leiden University*
[5]*GESIS, Leibniz Institute for the Social Sciences*

Abstract

Large language models (LLMs) are rapidly transforming social science research by enabling the automation of labor-intensive tasks like data annotation and text analysis. However, LLM outputs vary significantly depending on the implementation choices made by researchers (e.g., model selection or prompting strategy). Such variation can introduce systematic biases and random errors, which propagate to downstream analyses and cause Type I (false positive), Type II (false negative), Type S (wrong sign for significant effect), or Type M (correct but exaggerated effect) errors. We call this phenomenon where configuration choices lead to

Section 2

Quelques notions

Point de départ : du texte

Texte Représentation numérique

C'est quoi un texte ? Diversité de supports

- Document avec des chaâines de caractères numériques
- Pas encore mis en forme (PDF, images)
  - Enjeux d'OCR, de spatialisation (frame)

Constituer un corpus exploitable

Une étape à part entière (qui peut prendre plus de temps que tout le reste)

- techniques de traitement d'image (segmentation)
- image to text (OCR)
- manipulation de données

Et les questions de droits d'usages

Numériques mais non structuré

- Différentes langues (mélangées)
- Des erreurs (OCR)
- Des éléments supplémentaires non textuels (émoticones)
- Et plus ...

Et tous les problèmes liés aux représentations numériques des textes (formats, encodage)

Structurer implique de faire des choix

Quelle unité de base choisir ?

- Document entier
- Paragraphe
- Phrase
- Mot / couple de mot (bigramme)
- ...

(Dépendance à la langue évidemment)

Représentation d'un texte

Il n'y a pas une seule façon de penser un texte :

- Une suite de lettres
- Une suite de mots
- Des mots liés les uns aux autres
- Des éléments pertinents (noms propres, mots clé, etc.) liés aux autres

Différentes méthodes de représentation

- Par la présence de certains mots
  - Approches par dictionnaires
  - Ou par motifs : expressions régulières
- Par l'ensemble des mots
  - Approches par *sacs de mots* (*bags of words*)
- Par encodage de la structure
  - Approches par plongement (embeddings) contextuels ou non

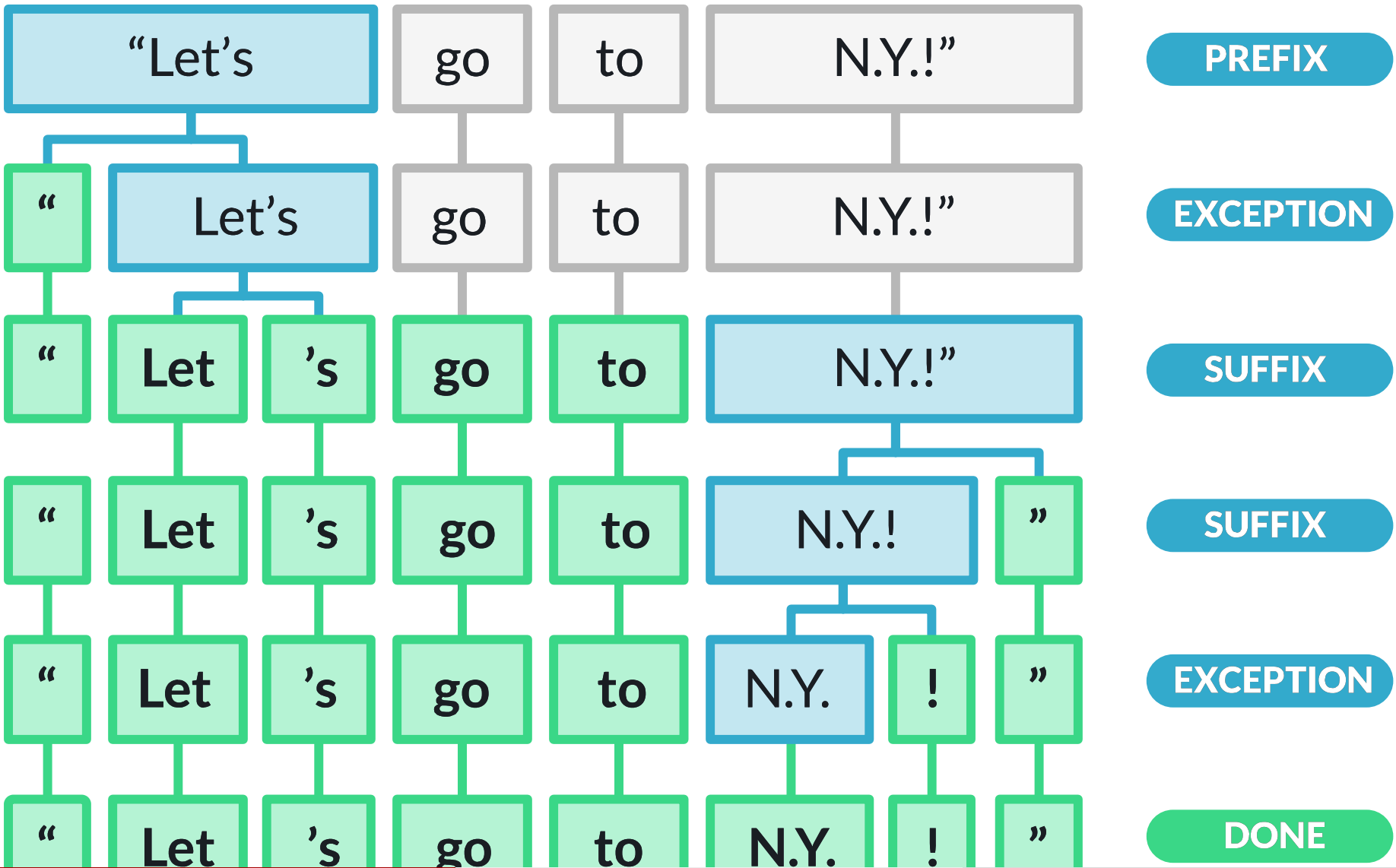
L'importance historique du mot

Importance de l'unité de base du mot

- découpage en mots
- suppression des mots vides
- lemmatisation/stemmisation

Pour de nombreux besoins spécifiques, intéressant de maîtriser les manipulations de bas niveaux

Tokénizer : une opération complexe



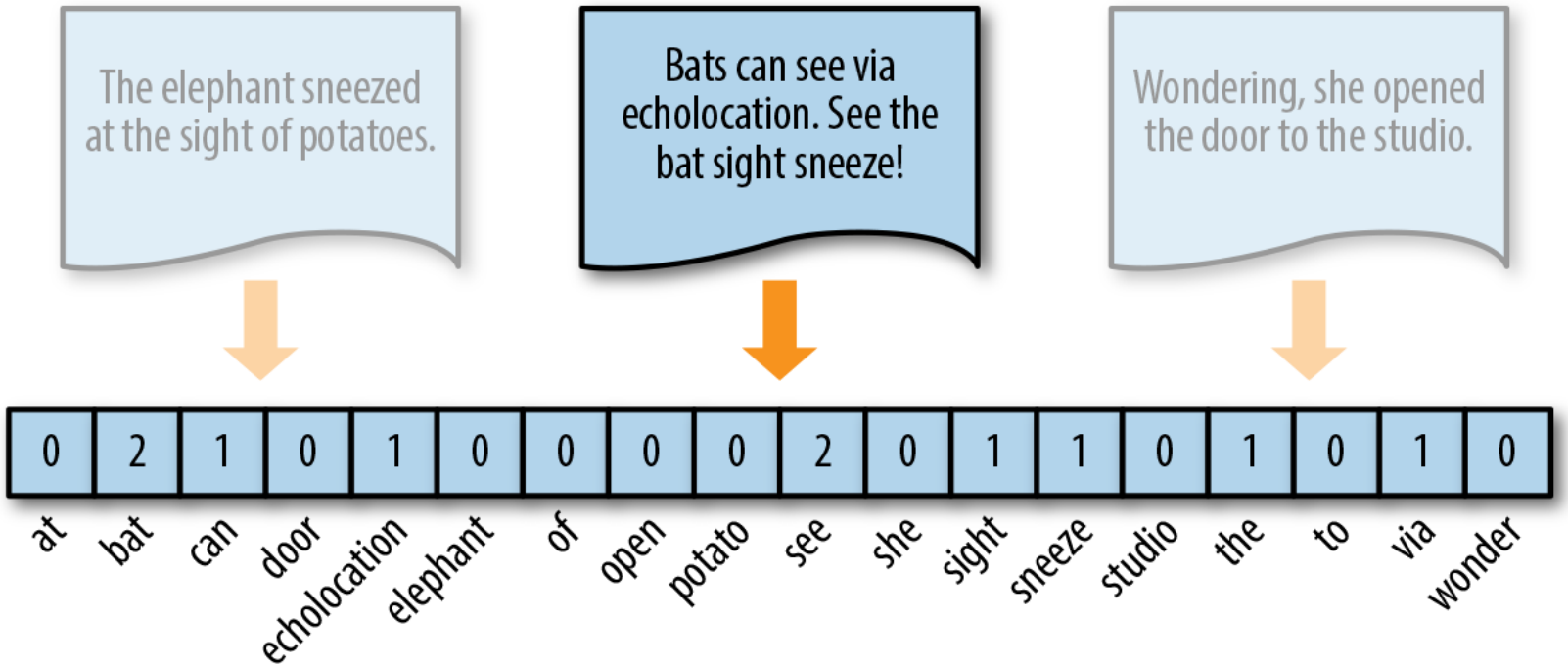
Faire des statistiques - la lexicométrie

Analyse *fréquentiste*

- Comptage
- Spécificité sur certains documents
- Indicateurs spécifiques (mots compliqués, hapax, etc.)
- Evolution / croisement avec d'autres variables
- Modélisation des distributions
  - LDA que je ne présenterai pas :)

Du token à la représentation

Passer d'un texte à un vecteur numérique sur l'ensemble de l'unité textuelle représentée.



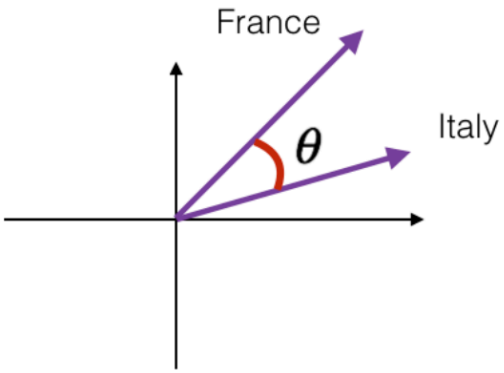
Du token mot au token segment

Avec le ML, d'autres unités sont importantes

- Passage en unités discrètes
  - **Tokenisation par mots** : "Je vais bien" → `["Je", "vais", "bien"]`
  - **Sous-mots (Byte-Pair Encoding, WordPiece)** :
  - "inconnue" → `["in", "##con", "##nue"]`
  - **Caractères** : chaque caractère est un token → `["J", "e", " ", "v", ...]`
- Dépend du pipeline/conséquences importantes

Que faire avec une représentation

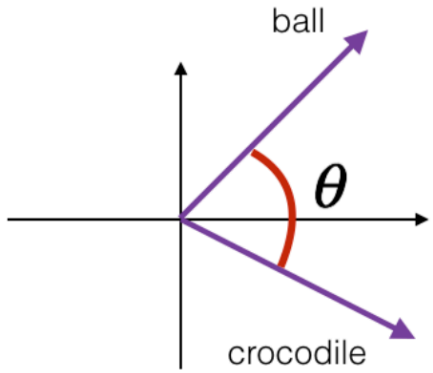
- Facilité de comparer deux vecteurs
- Calculer des distances
- Utiliser des modèles “classiques” de machine learning (ML)
- Faire des représentations (décompositions factorielles, etc)



France and Italy are quite similar

θ is close to 0°

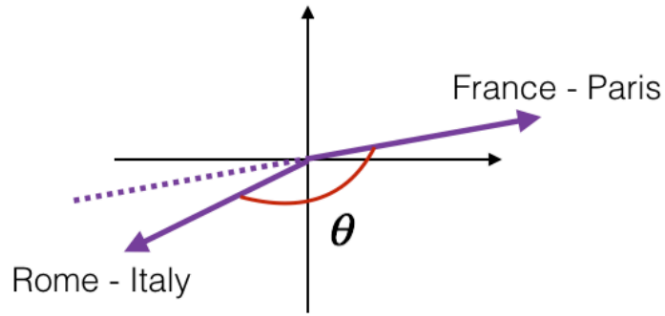
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$

Parenthèse : notions de machine learning

Notions plus générales que le NLP

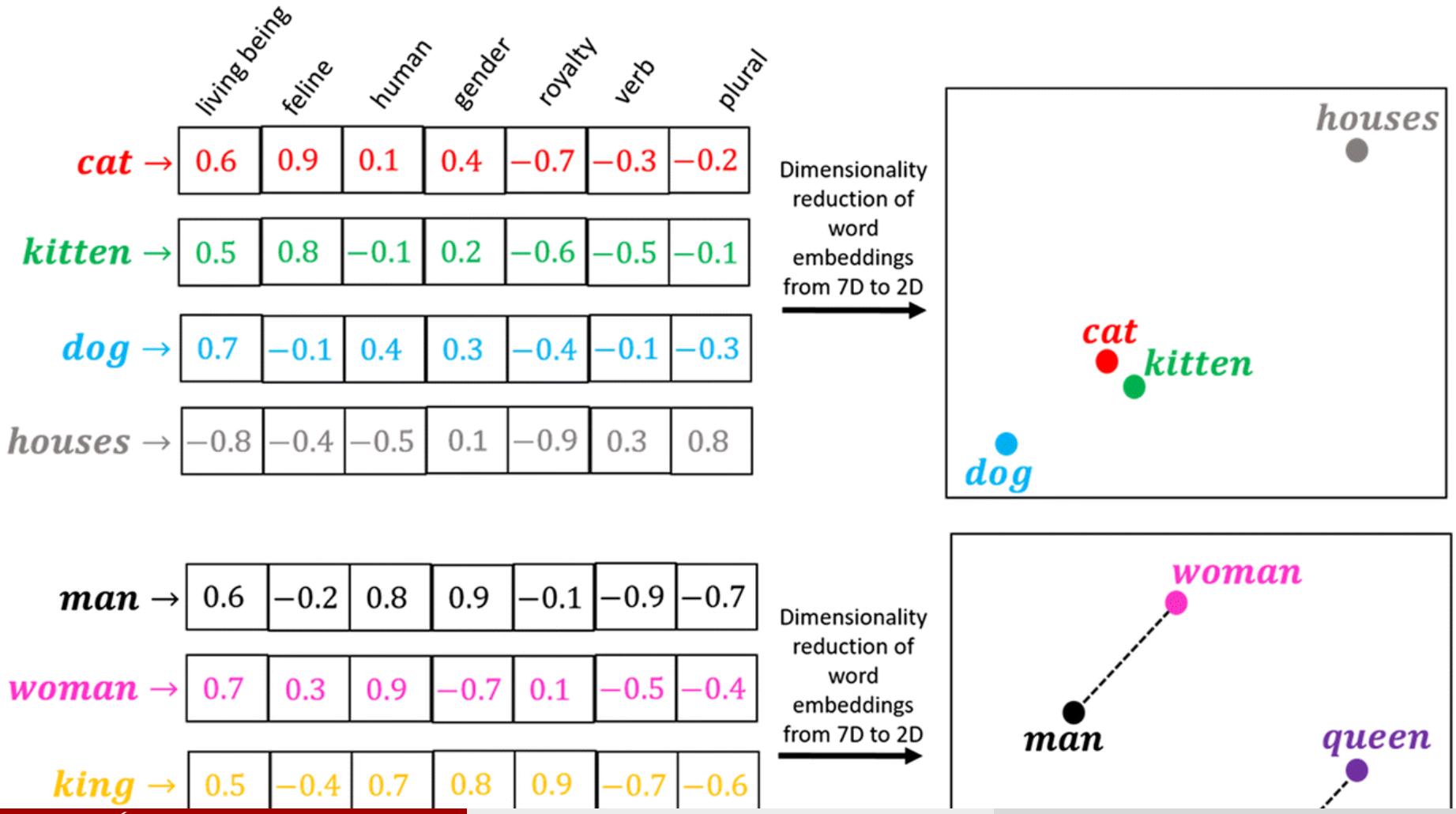
- Apprentissage non-supervisé
  - Utiliser la structure propre d'un jeu de données (ex. cluster, représentations)
- Apprentissage supervisé
  - Utiliser de l'information donnée par l'utilisateur pour guider

Deep learning et modèles

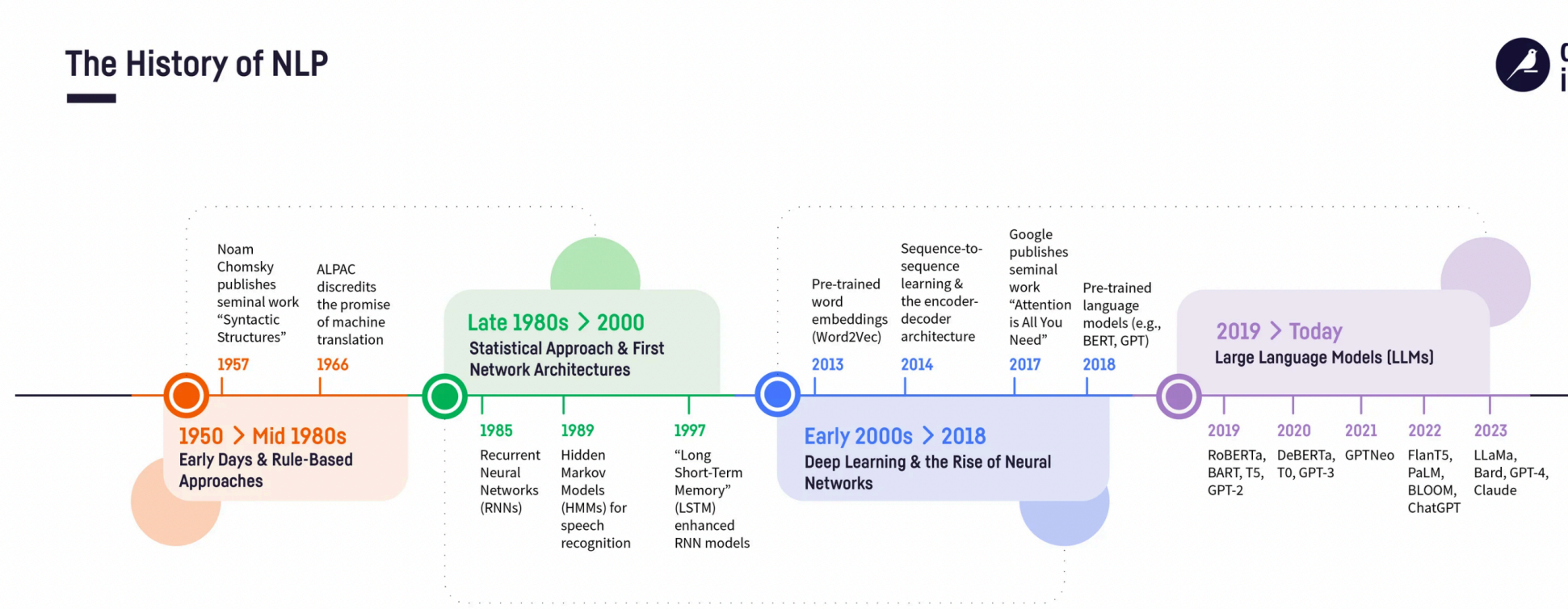
- Augmentation des corpus & des tailles de modèles
- Possibilité d'entraîner des modèles à
  - représenter
  - prédire
- De plus en plus de modèles préentrainés

Arrivée des embeddings

Espaces latents construits par entrainement de modèles sur des grands corpus (prédiction)



Une multitude de modèles



https://blog.dataiku.com/nlp-metamorphosis

Importance des modèles

- Outils pré-entrainés permettant la prédiction
  - Importance du corpus d'entrainement
  - Spécifiques à la langue / type de textes
- Dépendent de plusieurs niveaux
  - Tokenisation
  - Corpus d'entrainement
  - Méthodes (RLHF)...
- De nouvelles notions :
  - fenêtre de contexte
  - BERT, GPT, ...

En ce qui nous concerne

- Transformers
  - BERT (encoder only, 2018+)
    - pour le français CamemBert ou FlauBERT
    - Récemment, ModernBERT
- Depuis 2022, explosion des LLM
  - Dépasse le NLP (par exemple, Whisper)
  - HuggingFace

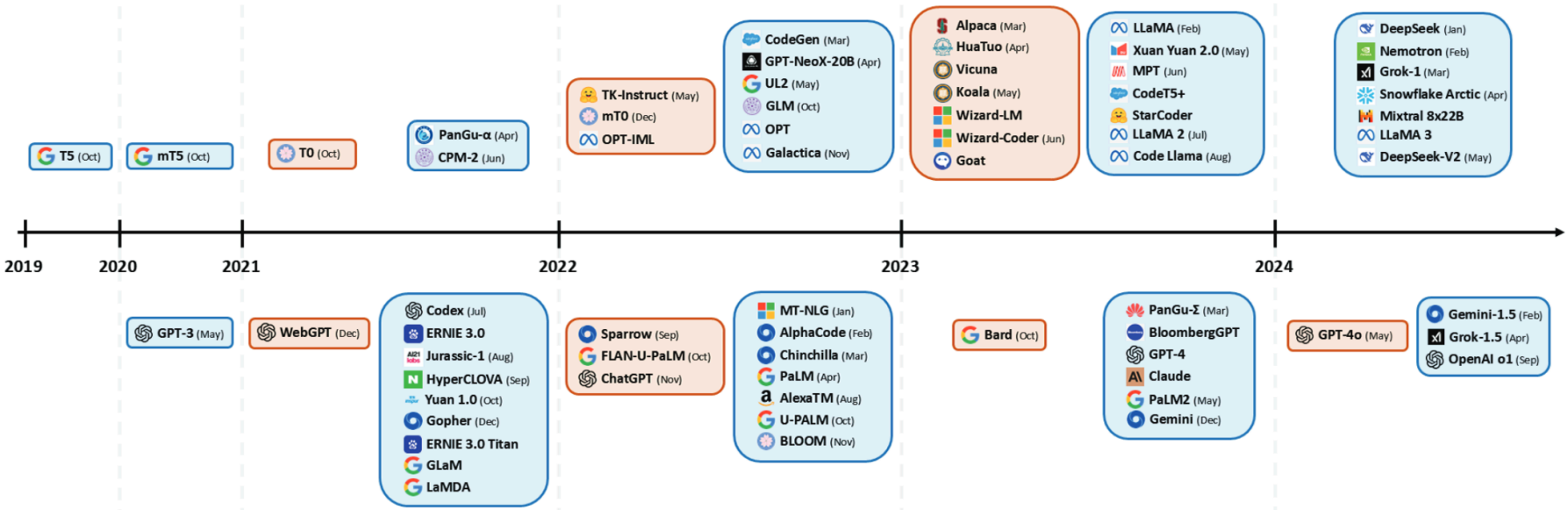


Fig. 2. Chronological display of LLM releases: blue cards represent "pre-trained" models, while orange cards correspond to "instruction-tuned" models. Models on the upper half signify open-source availability, whereas those on the bottom are closed-source. The chart illustrates the increasing trend toward instruction-tuned and open-source models, highlighting the evolving landscape and trends in NLP research.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, et al. 2025. « A Comprehensive Overview of Large Language Models ». ACM Transactions on Intelligent Systems and Technology 16 (5): 1-72. https://doi.org/10.1145/3744746.

Entrainement, tâches et fine-tuning

- Les grands modèles fondationnels sont lourds à entraîner (jours avec beaucoup de ressources)
- Une pratique de fine-tuning (apprentissage supervisé ou semi-supervisé)
- Modèle pré-entrainé $>$ utilisation sur différentes tâches

Des modèles de toutes tailles

Un petit tour sur Hugging Face

- Indicateur : nombre de paramètres
- Des modèles de grande taille
  - Certains nécessitent des GPU
  - Penser à utiliser des services dédiés si nécessaires

Un enjeu : manipuler ces modèles

Enjeu en général

Suivant les besoins, trouver la bonne tâche :

- rapidité
- robustesse
- efficacité
- ...

Et importance d'évaluer la qualité du traitement.

Faisons un petit tour des lieux

Section 3

Faire du NLP avec Python

Python dans tout ça

- Language de programmation permettant la manipulation des données
- Au coeur de la révolution IA actuelle

Les bibliothèques Python

- Avant, un peu périmée `NLTK`
- Pour faire du ML avec `Scikit-learn`
- Le plus pratique : `SpaCy`
- Des modèles dédiés : `GenSim`
- Utiliser directement des modèles de HuggingFace avec `Transformers`
  - Ou des bibliothèques construites dessus ...

Mélange savoir spécialisés / compétences génériques

- Manipuler des données
  - petites/larges
- Notions de ML
- Bibliothèques spécialisées

Outils complémentaires

- Extraction d'information
- Annotation
  - Label Studio, Doccano
- Sémantisation (TEI)

Section 4

Passons à la pratique

Pour aller plus loin

Beaucoup de littérature & de tutoriaux

- le cours de Lino Galliana
- Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin
- Text As Data