# Language & AI

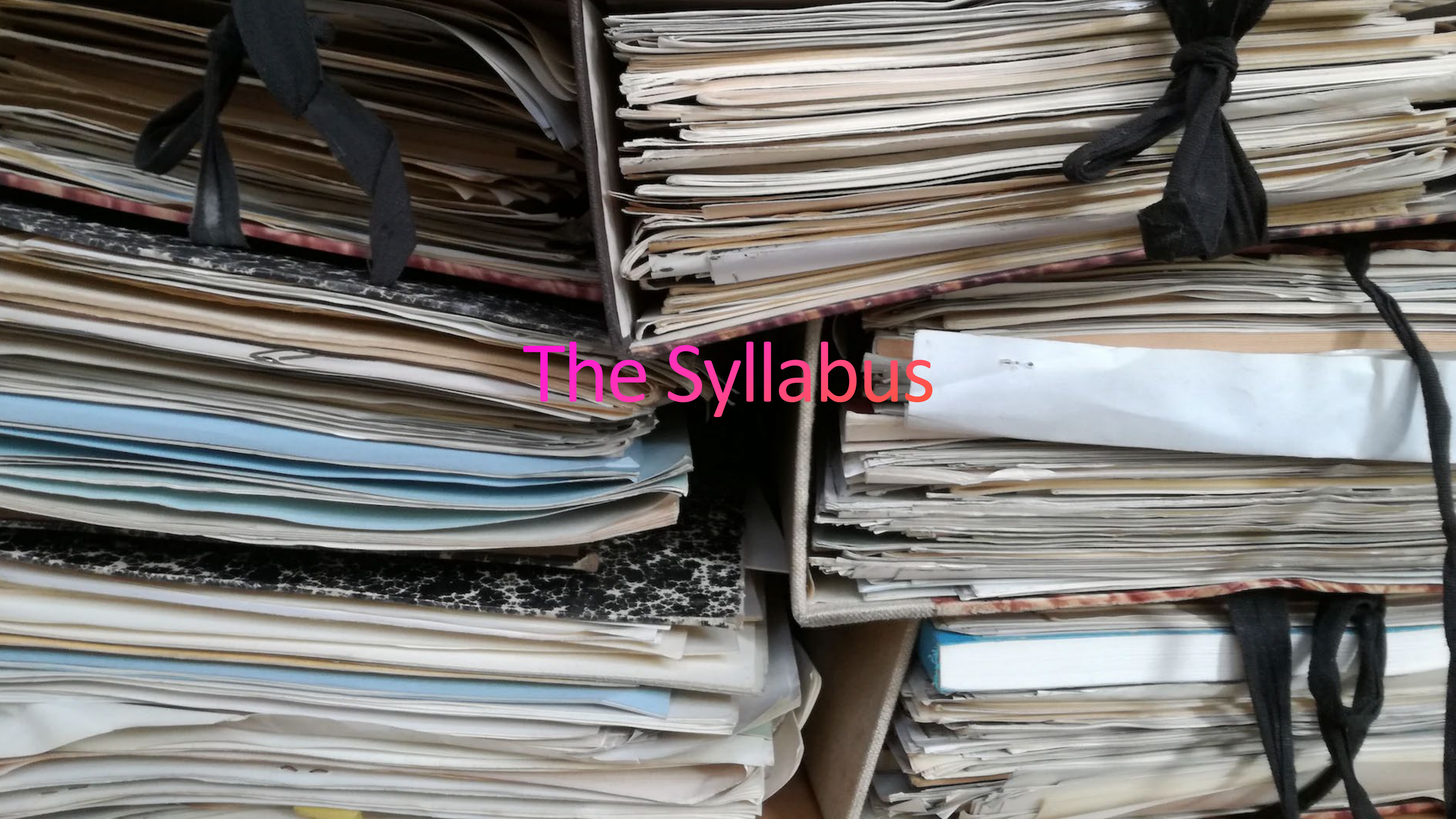
## Lecture 1

Dr. ir. Marijn ten Thij, Nov 13th 2025

# Today's Lecture

## NLP Beginnings

- The Syllabus
- Interim Assignment Details
- Task Information
- Incorporating Lecture Knowledge


The Syllabus

# Coursework

- 📺 Video Lectures
- 🧮 Exercises
- 📚 Reading
- 👨‍🏫 In-Person Lectures <— you are here
  - Will discuss applying video lecture material to research paper assignment and provide contextual information.
- 💻 Lab Sessions
  - Can be used to ask questions, both regarding code and project.
  - Attendance **strongly recommended**.

Dr. ir. Marijn ten Thij
Course Coordinator

Chiara Manna
Tutor

# Assessment

- 🖥️ Interim Assignment (40%) — research paper
  - Groups up to 4 (recommended). See matching board on Canvas.
  - Proposal deadline: Dec 1 (13:00).
  - Final deadline: Fri 9 Jan (17:00). See course page for late policy.
- 📝 Exam (60%)
  - Mon 26 Jan (Mon 13 Apr for Resit)
  - Mix of MC, calculation, and essay-style questions.
  - Example questions, style descriptions, and more info in syllabus.


Interim Assignment

# Project Proposal

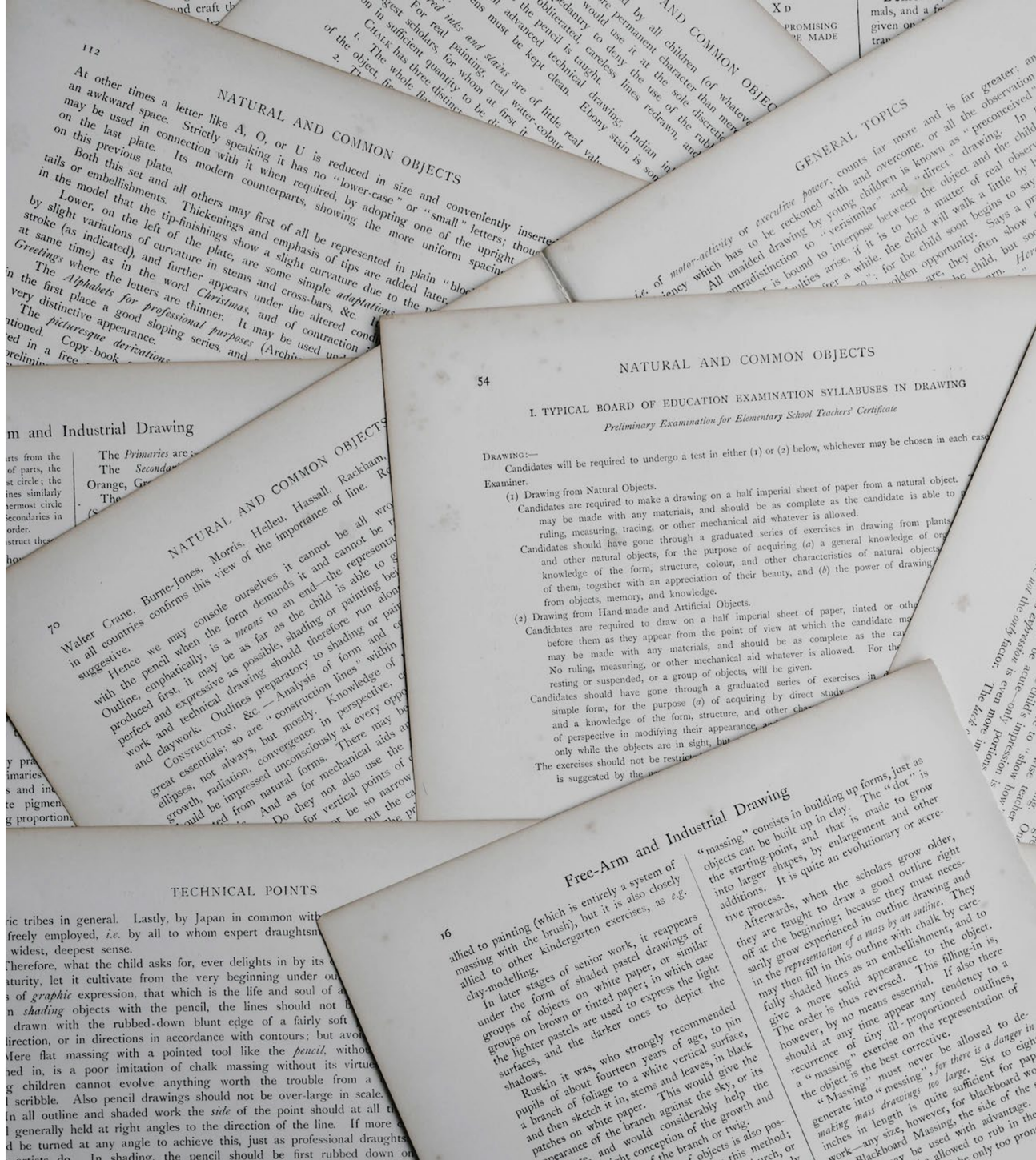
## Get Formative Feedback

- With group: 1-2 pages of content
- Research goal(s) / Hypothes(i/e)s
- Literature Review
- Evaluation (Metrics, Qualitative)
- Models
- General Reasoning
- Progress Summary

# Research Paper

## Final Version (Graded)

- Short paper: 4 pages formatted in ACL template.
- Abstract
- Introduction
- Related Work section
- Experiments (Method)
- Discussion and Conclusion
- Authorship Statement



# Assessment

## We Don’t Care About Performance!

- Implementation: Understanding and strength of the methods and evaluation.
- Interpretation: Extent paper is open and clear-sighted about its limitations.
- Presentation: e.g., writing quality, clarity, and use of visual information.
- Reproducibility: code and documentation quality, matching results in paper.
- Creativity: providing insights that go beyond the material discussed in the lecture.

# Rubric

## Implementation

- The experimental setup is appropriate including the right description of the data, appropriate pre-processing of the data, evaluation, and parameters.

Excellent papers will have robust analyses, both quantitatively as well as qualitatively. They clearly describe the data being used and the method to produce the results.

# Rubric

## Interpretation

- Understanding of the task is clear, research questions / hypotheses have been appropriately formulated, and the discussion is clear-sighted about the implication of the results and the research its limitations.

An excellent paper has a well-structured introduction, a clear results section, and a discussion section that provides context and answers to the research questions / hypotheses, considers the research and tasks limitations, and provides avenues for future work.

# Rubric

## Presentation

- E.g., writing quality, clarity, and use of visual information.

Excellent writing quality is evidenced by clear logic and careful attention to reasoning of points. There are no grammar or other mistakes. The paper is of an appropriate length (4 pages excluding reference section and authorship statement). Appropriate sources are included. Figures and tables are high quality with self-contained captions. Figures and tables are used to allow the reader to understand the data and the results.

# Rubric

## Reproducibility

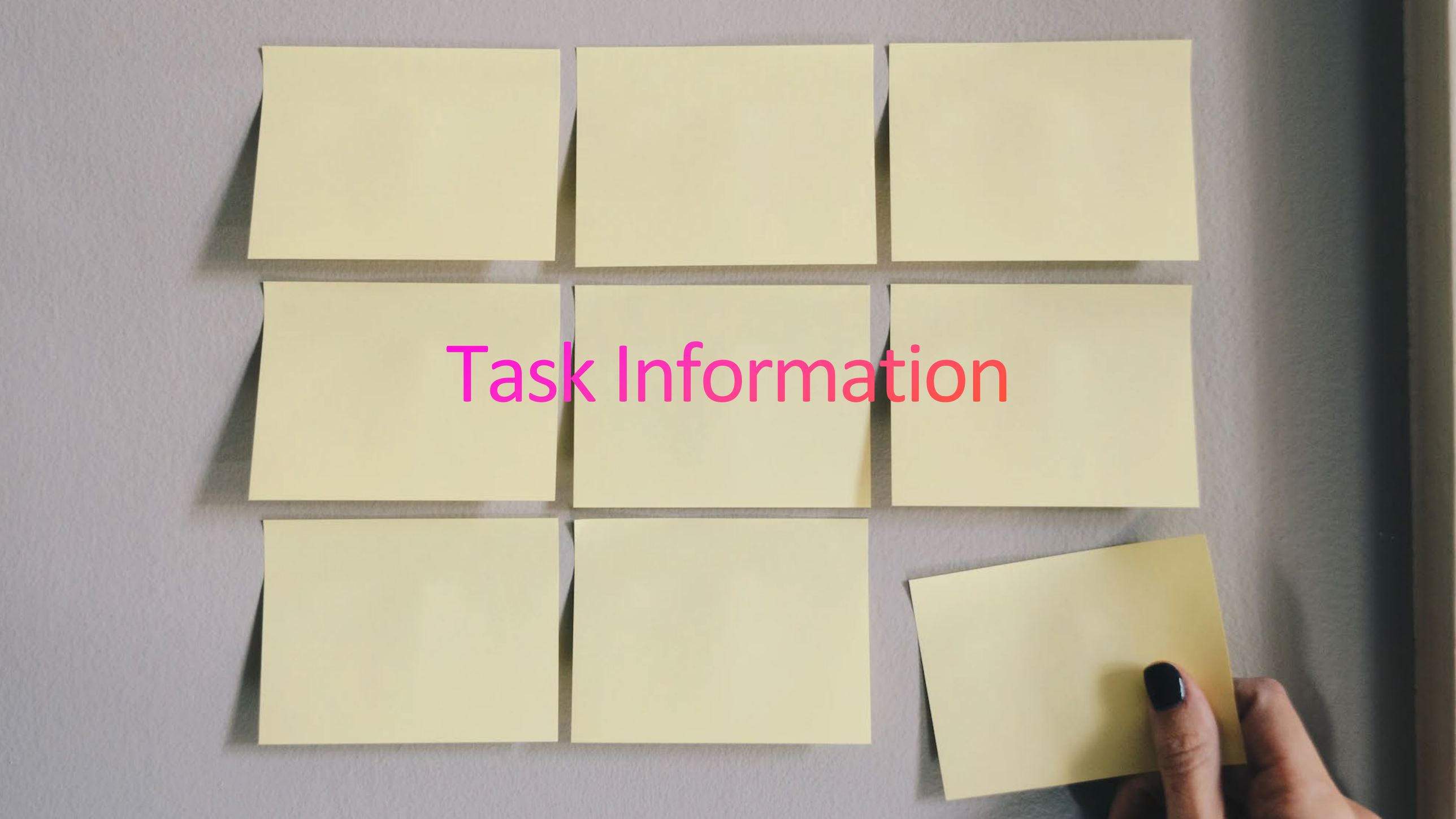
- Code and documentation (of the classes / functions, and the code base itself) is appropriate for the presented research and connects with the results in the paper.

Excellent reproducibility entails a user running the code on their machine will see output matching the paper. Suggestions for extending the code repository and running it on different data are provided as well.

# Rubric
## Creativity

- The method and analyses in the paper expand adequately upon the material discussed in the lecture (both theory and task-related).

Excellent creativity is evidenced by thought-provoking qualitative analyses, visualizations, the use of more advanced or interpretable models, detailed error analyses or bias assessment.


Task Information

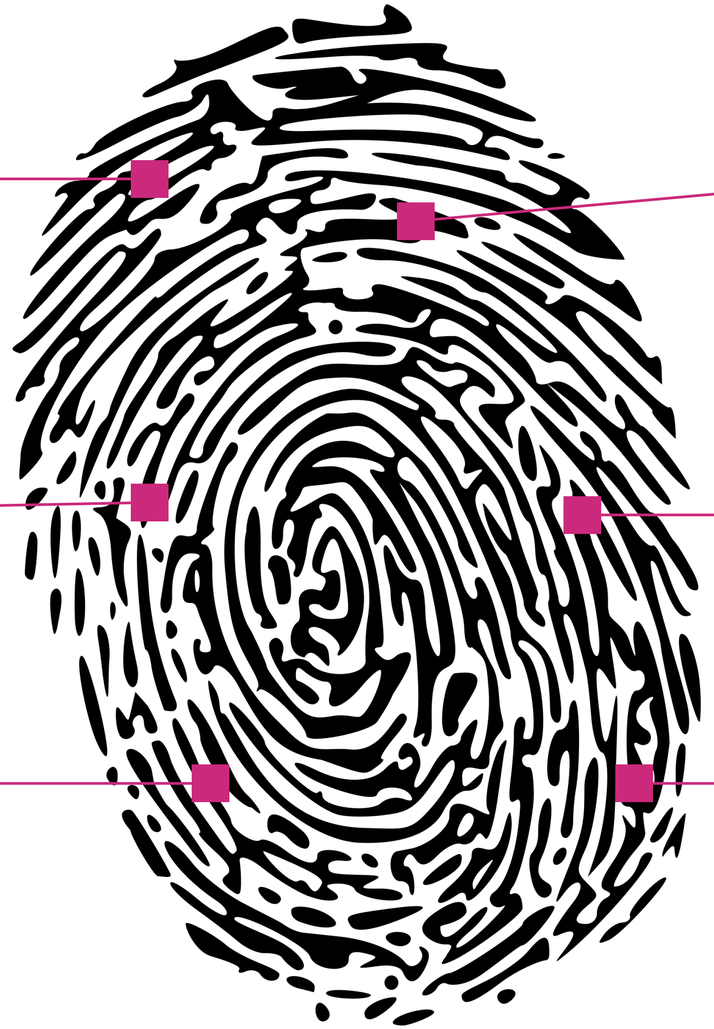
# Stylometry

## Latent Information in Writing

- Language creativity in writing; `problems` with unique word combinations.
- Your writing style may reveal many pieces of personal information.
- More humble beginnings: do factors in our lives determine our language-use?

# What's Writing Style?

## Your Linguistic Patterns



Low frequency word

Slang

Sports metaphor

Passive Voice

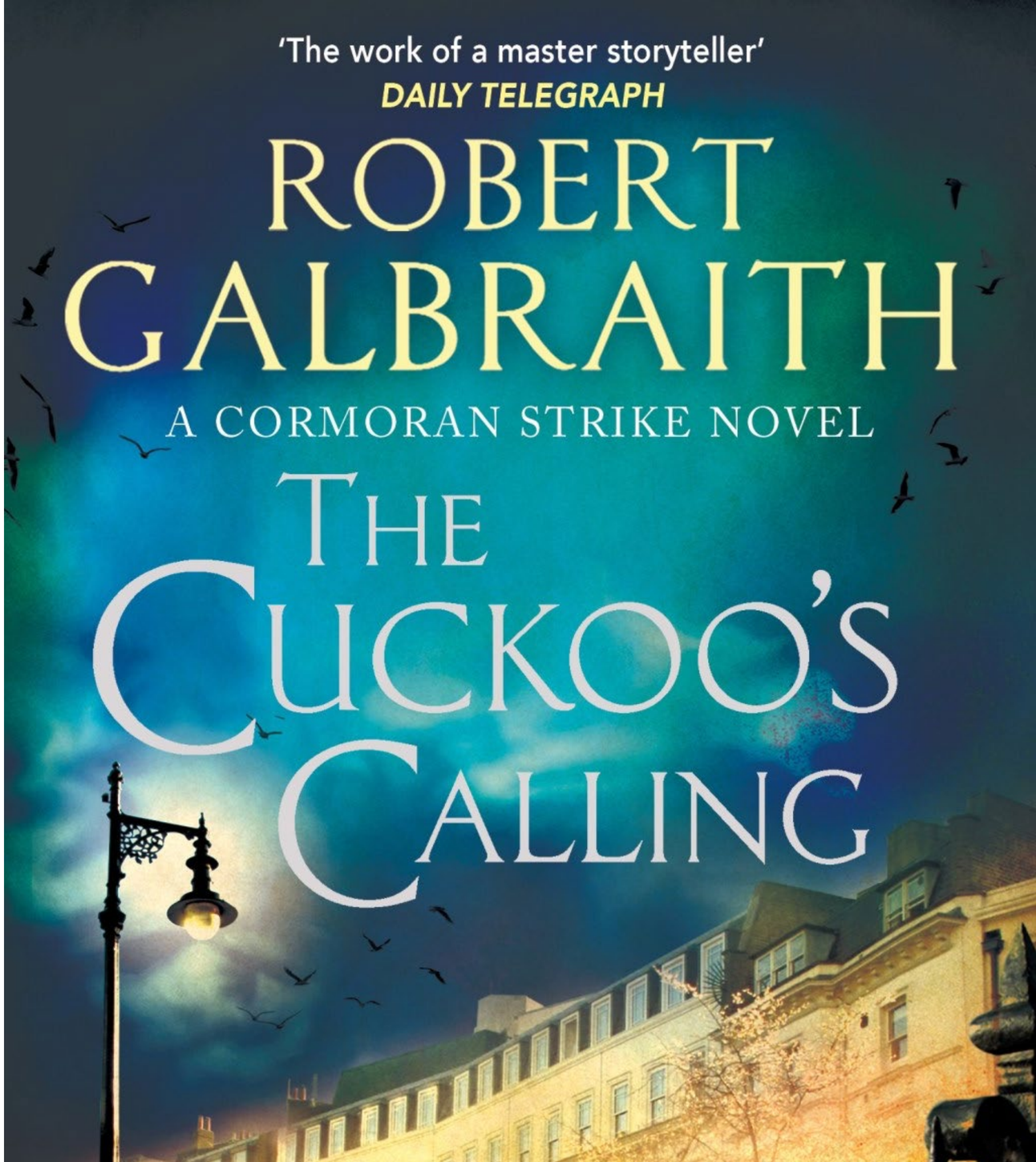
Oxford comma

Perfect Grammar

# Applications

## Stylometry in the Wild

- Forensic investigation
- Marketing (ads)
- Political targeting
- Uncovering anonymous writers through “the use of Latin phrases and a drug-taking scene”.



# Computational Stylometry

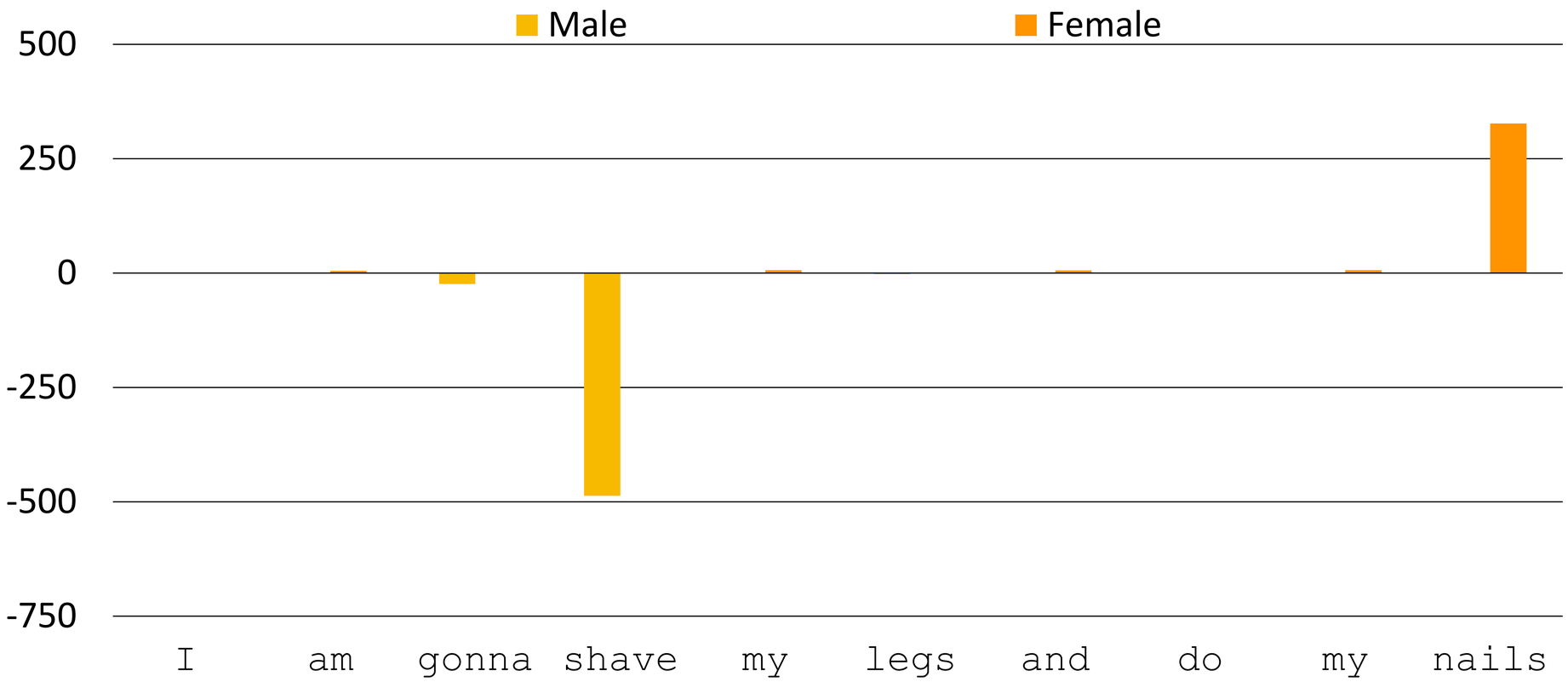
## A range of (sub)tasks

- Authorship Identification — given 📚, by 👥 👥 👥, who is ✍️?
- Authorship verification — given 📖, what's probability 👤 is ✍️?
- Author profiling — given 📚, what's ✍️ ASL (etc.)

- Input:
  - Handcrafted style features (usually raw frequencies).
  - Content words (tf*idf weights for example).

# Does That Work?

## Yes, but stereotypes run rampant!



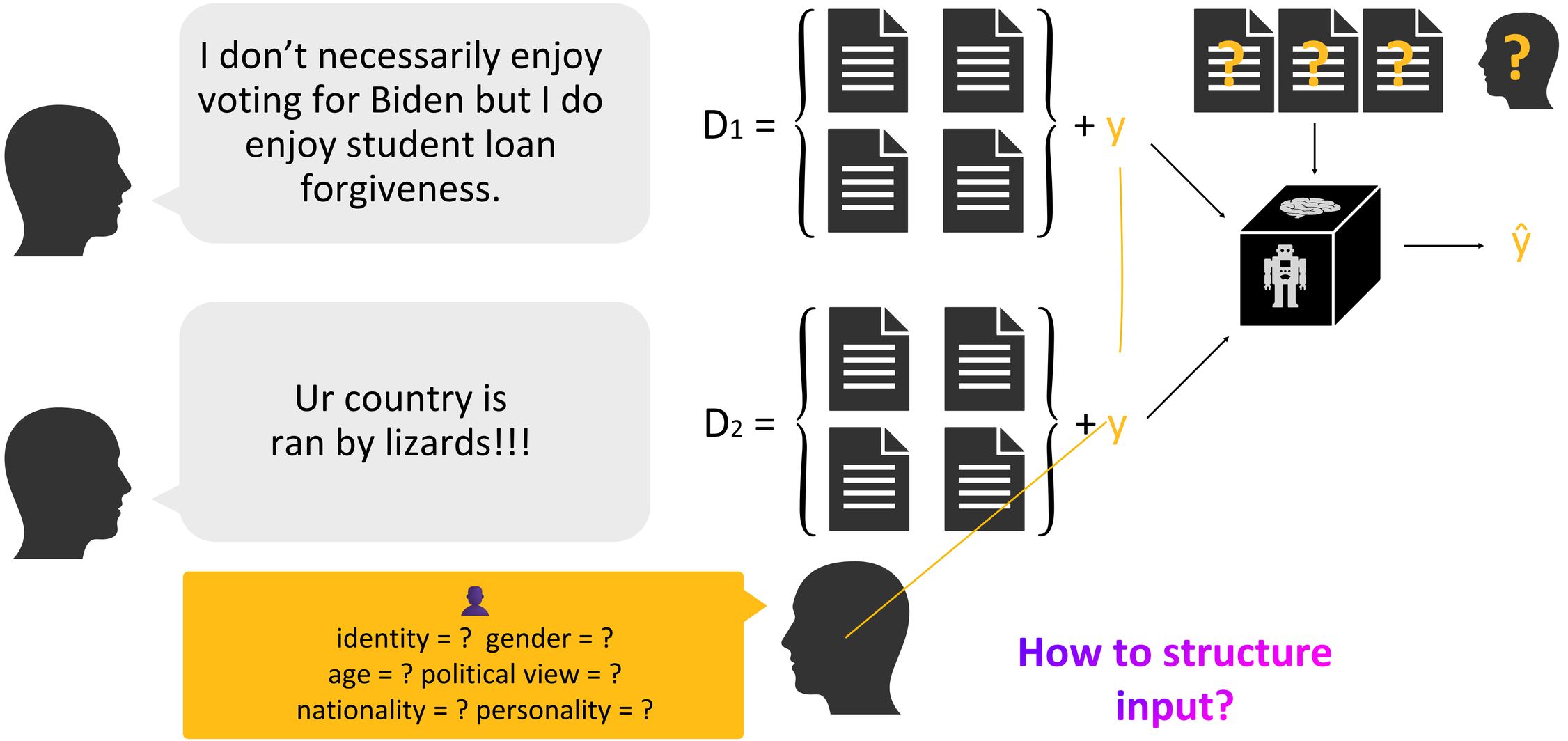

Simple logistic regression weights for words in sentence. Model from Sap et al. (2014).

# Author Profiling

## Beyond Individual Style

- Age
- Gender
- Nationality
- Personality
- Political Leaning
- ???
- Mental Health Issues

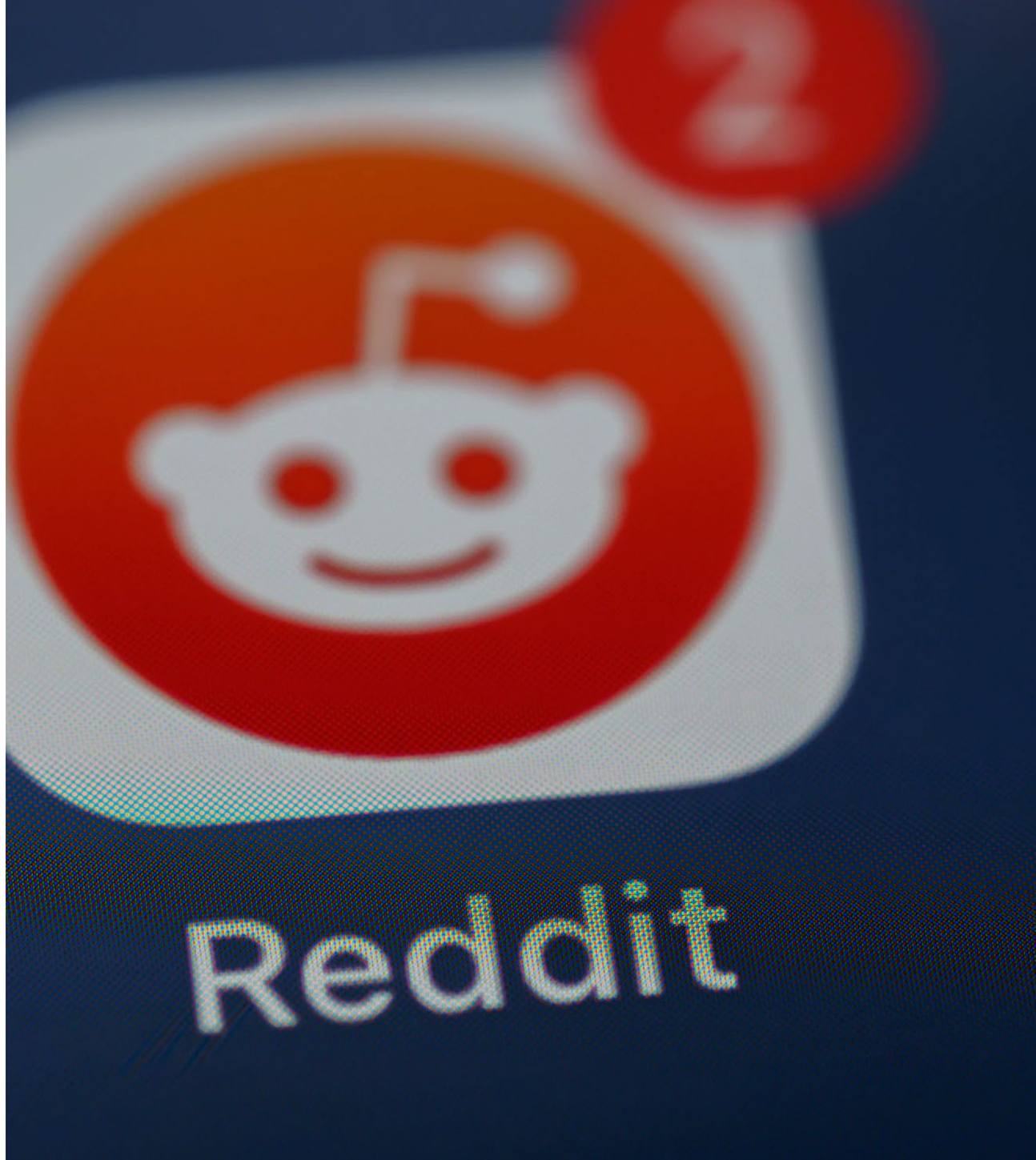
# Typical Prediction Pipeline



# Your Project's Data

## Reddit Author Profiles

- Reddit posts (from Pushshift).
- Distantly annotated.
- Used flairs.
  - Previous work: text reports.
- Presented @ LREC-COLING 2024.
- Data sharing agreement.



# Accompanying Paper

## Will Also Provide Examples From Last Year

**SOBR: A Corpus for Stylometry, Obfuscation, and Bias on Reddit**

**Chris Emmery**[1], **Marilù Miotto**[2], **Sergey Kramp**[1], **Bennett Kleinberg**[2,3]
[1]Department of Cognitive Science & Artifical Intelligence, Tilburg University
[2]Department of Methodology and Statistics, Tilburg University
[3]Department of Security and Crime Science, University College London
cmry@pm.me

**Abstract**

Sharing textual content in the form of public posts on online platforms remains a significant part of the social web. Research on stylometric profiling suggests that despite users' discreetness, and even under the guise of anonymity, the content and style of such posts may still reveal detailed author information. Studying how this might be inferred and obscured is relevant not only to the domain of cybersecurity, but also to those studying bias of classifiers drawing features from web corpora. While the collection of gold standard data is expensive, prior work shows that distant labels (i.e., those gathered via heuristics) offer an effective alternative. Currently, however, pre-existing corpora are limited in scope (e.g., variety of attributes and size). We present the SOBR corpus: 235M Reddit posts for which we used subreddits, flairs, and self-reports as distant labels for author attributes (age, gender, nationality, personality, and political leaning). In addition to detailing the data collection pipeline and sampling strategy, we report corpus statistics and provide a discussion on the various tasks and research avenues to be pursued using this resource. Along with the raw corpus, we provide sampled splits of the data, and suggest baselines for stylometric profiling. We close our work with a detailed set of ethical considerations relevant to the proposed lines of research.

**Keywords:** corpus, author identification, author profiling, author obfuscation, computational stylometry, bias

### 1. Introduction

The increasing computational capabilities of language models do not bode well for safety in online public spaces. A large variety of pre-trained Large Language Models (LLMs) made readily available through platforms such as the HuggingFace Model Hub (Wolf et al., 2020) can be used to generate (Pan et al., 2020; Carlini et al., 2021) and infer (Tesfay et al., 2019; Kleinberg et al., 2022), sensitive information. While these often deal with concrete mentions of personal information, a handful (so far) seeks to uncover latent author attributes through computational stylometry.

Stylometry posits that one's unique writing style might encode features about an author's identity, which eventually extended to sociodemographic et al., 2008; Ott et al., 2011; Banerjee et al., 2014; Fornaciari and Poesio, 2014), it enables malicious actors to infer potentially sensitive information unbeknownst to the user. This is particularly harmful to individuals in a vulnerable position regarding race, political affiliation, mental health, or any other personal details made explicitly unavailable.

Historically, collecting high-quality labels for stylometric classification tasks was a costly process (both in time and resources) requiring trained annotators. Collecting the data itself, and fine-tuning models, would also require expertise and computational infrastructure. Works such as Beller et al. (2014) and Emmery et al. (2017), however, showed that targeting Twitter users that post self-reported attributes ("I'm a ...") generates enough distantly labeled data to train models that match the per-

# Data Agreement

This Agreement is dated and in effect as of the submitted date, between: the submitter of this form and their group (hereinafter the "Recipient''), and the Department of Cognitive Science & AI, Tilburg University, Warandelaan 2 5037 AB Tilburg (hereinafter the ``Discloser''). This Agreement is with respect to the disclosure of confidential research data (hereinafter the "Data'') to the Recipient. Whereas, Recipient is: a group in the Language & AI course who wish to use the Data for: a research paper assignment for the Language & AI course, to be submitted the 9th of January 2025 (hereinafter ``the Purpose''); Whereas, Discloser wishes Recipient to be able to access the Data under strict conditions solely for the Purpose; Now, therefore, in consideration of the foregoing premises and the mutual covenants hereinafter set forth and other valuable considerations, the parties hereto agree as follows:

1. The Recipient undertakes not to use the Data for any purpose except the Purpose, without first obtaining the written agreement of the Discloser.

2. The Recipient undertakes to keep the Data secure, not to disclose it to any third party, and to ensure the confidentiality and sensitivity of the data is not breached in any manner. The Recipient also agrees to abide by the rules of data protection legislation.

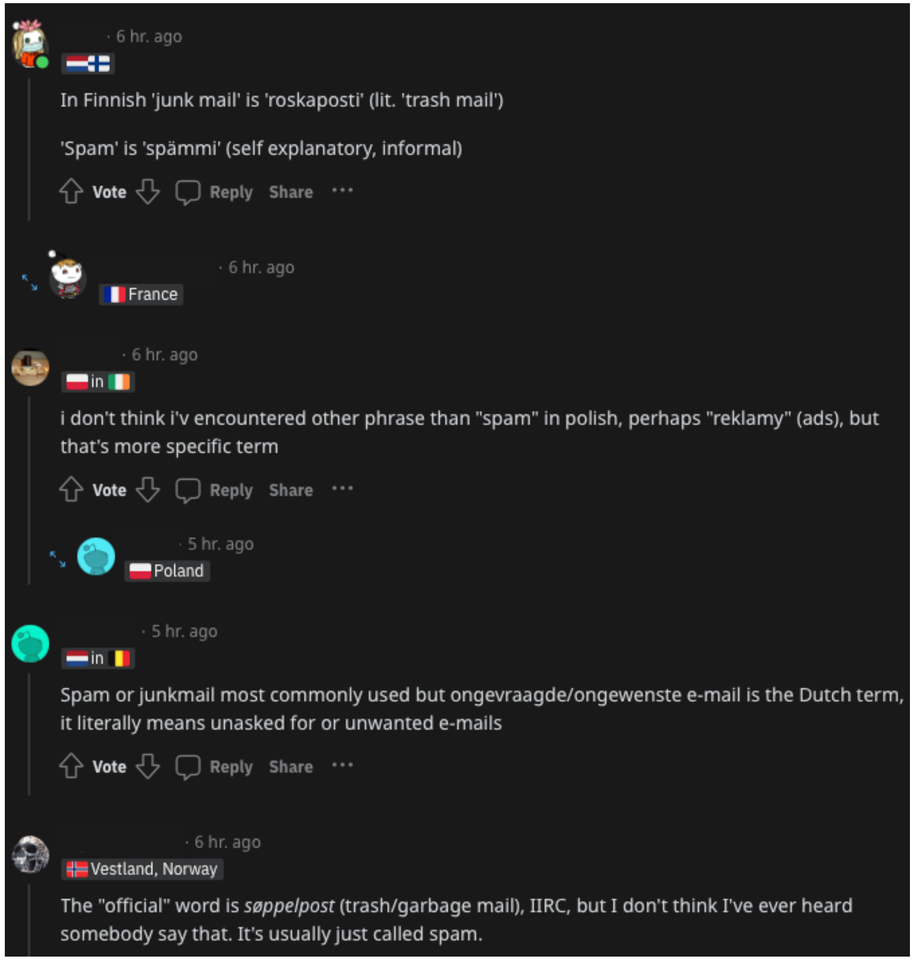
3. The undertakings in clauses 1 and 2 above apply to all of the Data disclosed by the Discloser to the Recipient, regardless of the way or form in which it is disclosed or recorded. However, they do not apply to data which is or in future comes into the public domain (unless as a result of the breach of this Agreement).

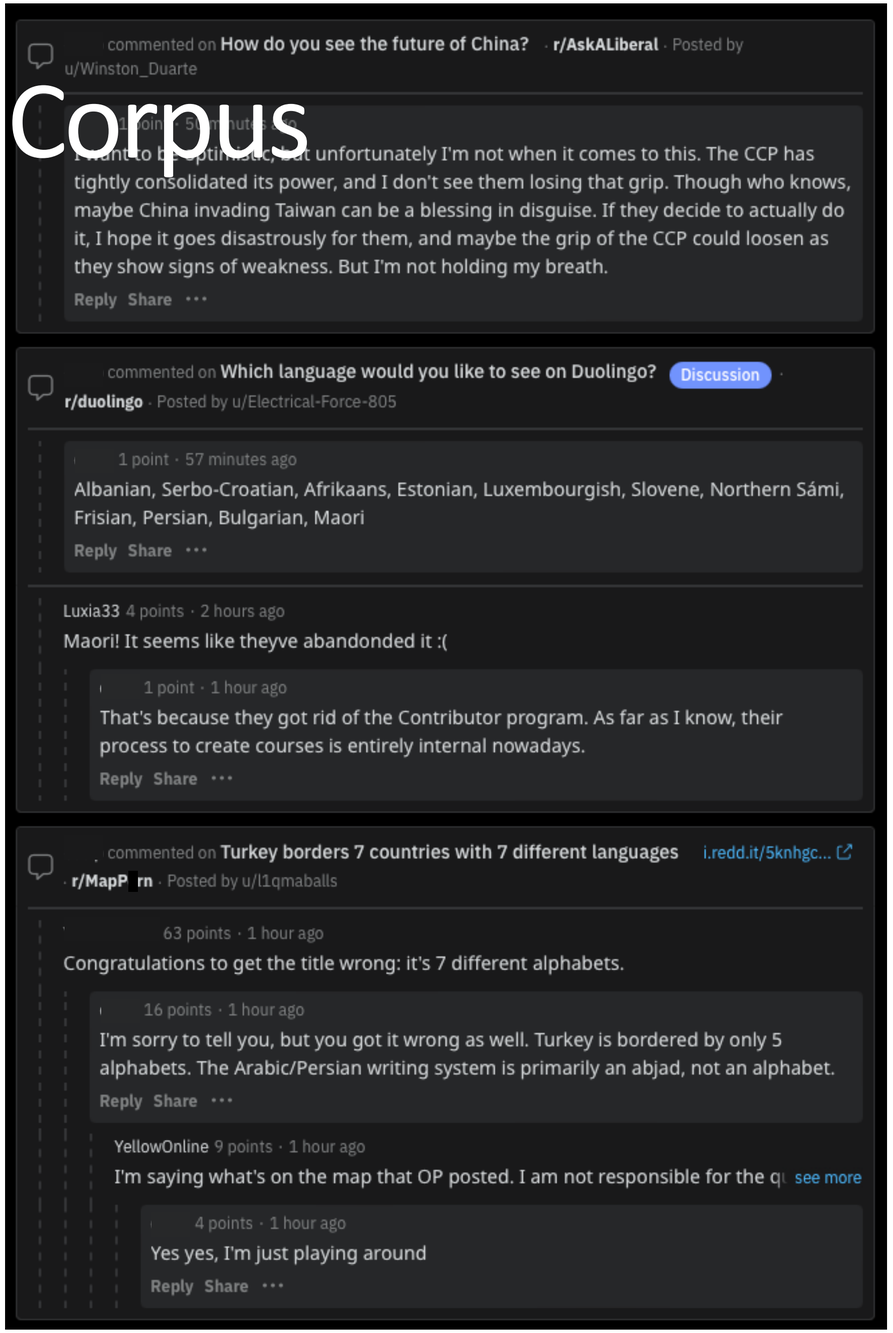
4. Nothing in this Agreement will prevent the Recipient from making any disclosure of the Data required by law or by any competent authority.

5. The Recipient will, on request from the Discloser, or after termination of the Purpose, not retain any copies, records, or derivatives of the Data.

6. Neither this Agreement nor the supply of any information grants the Recipient any license, interest or right in respect of any intellectual property rights of the Discloser except the right to copy the Data solely for the Purpose.

# Reddit as a Corpus




**Potential Biases?**

# Debug the Corpus

This Year’s Spin

# Incorporating Lecture Knowledge

# (Psycho)linguistics

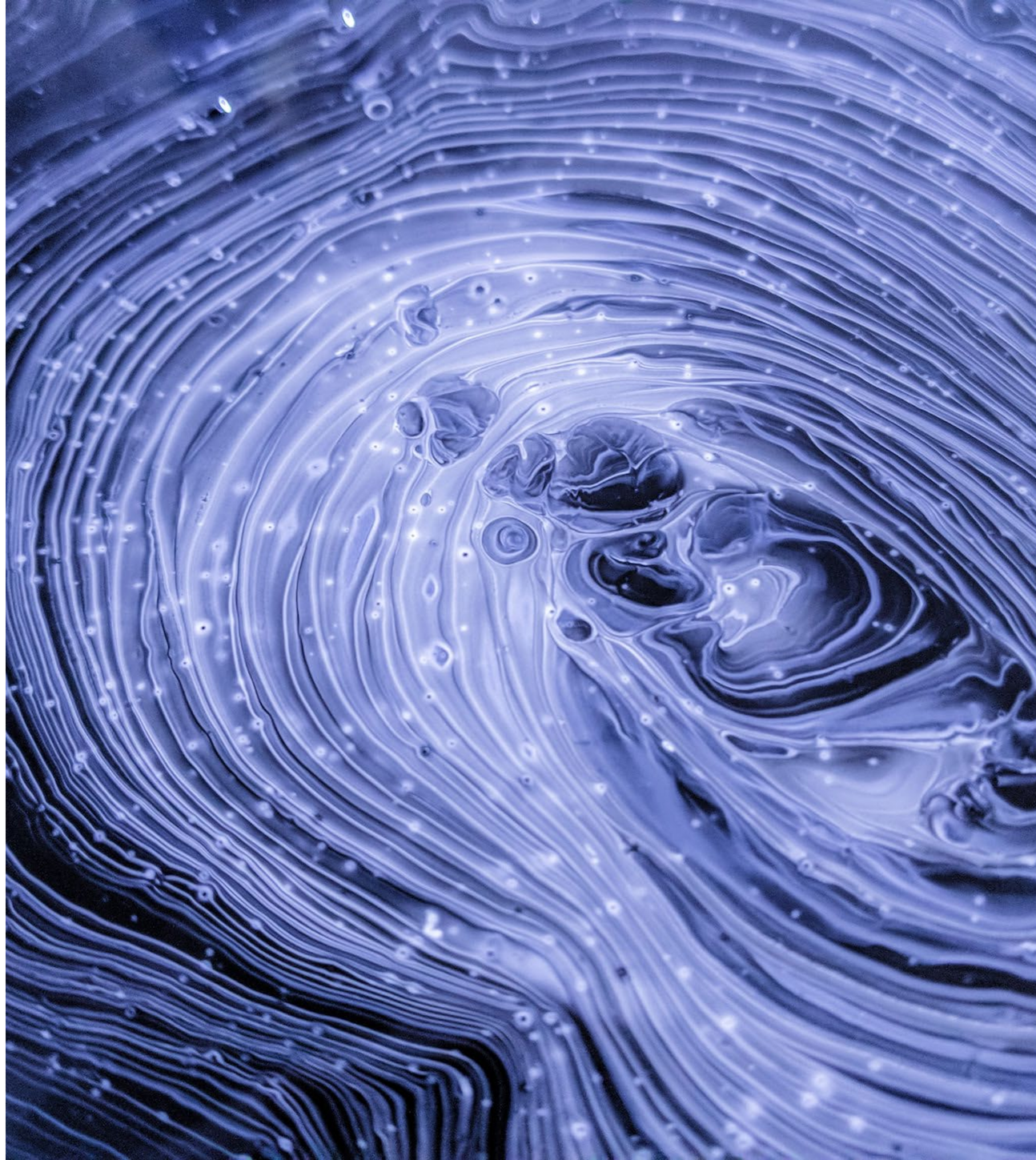
## Debugging the Task

- Investigate bias in data and models.
- Interpret / ‘explaining’ models.
- Create challenging test sets!

# Feature Extraction

## More to Come!

- Token counts can serve as input. Weighting important.
- Next lectures will go into more detail on extracting richer linguistic patterns (style features!).
- Distances may be used per author for qualitative analysis.



# Any Questions?