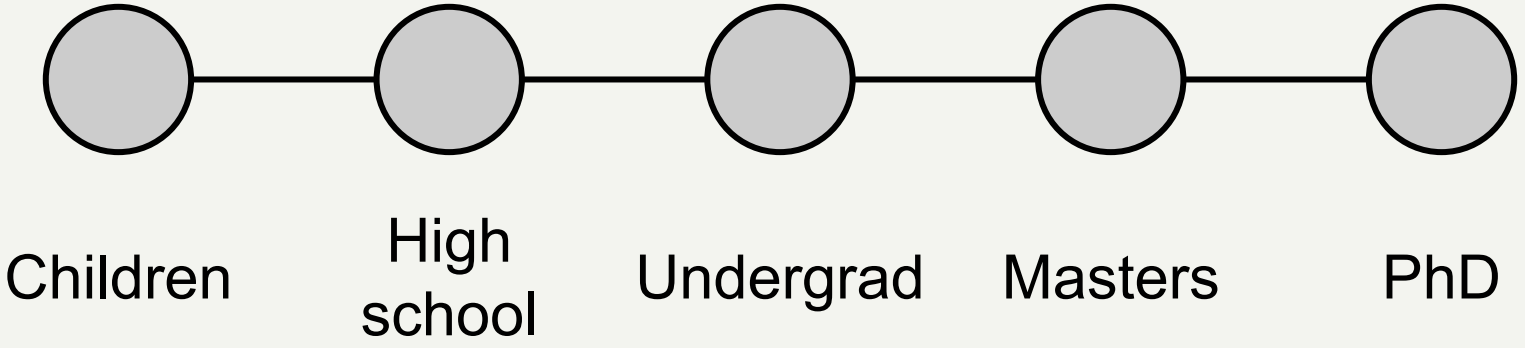
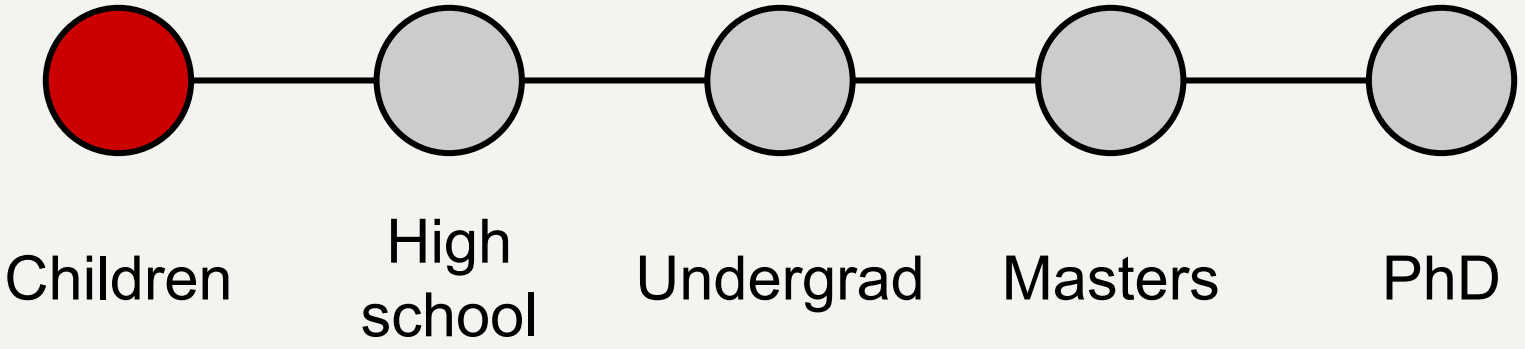
Pedagogy Discussion

Vagrant Gautam (they/xe)
Saarland University

My teaching experience

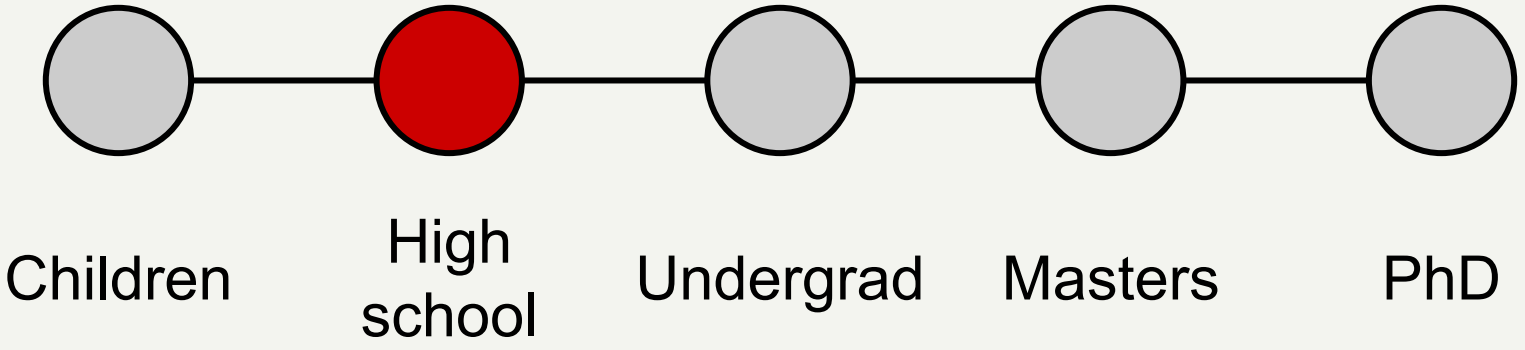


My teaching experience: Coding with Scratch, Python



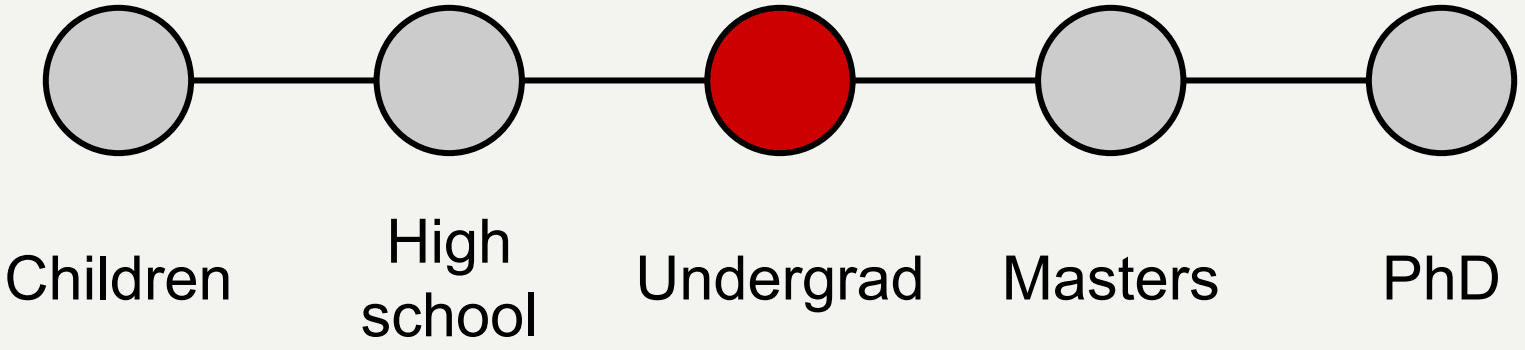
- Stride Avenue Community School, Burnaby 🇨🇦
- Girls Learning Code, Vancouver 🇨🇦
- Kids Code Jeunesse, Vancouver 🇨🇦

My teaching experience: Coding with Python



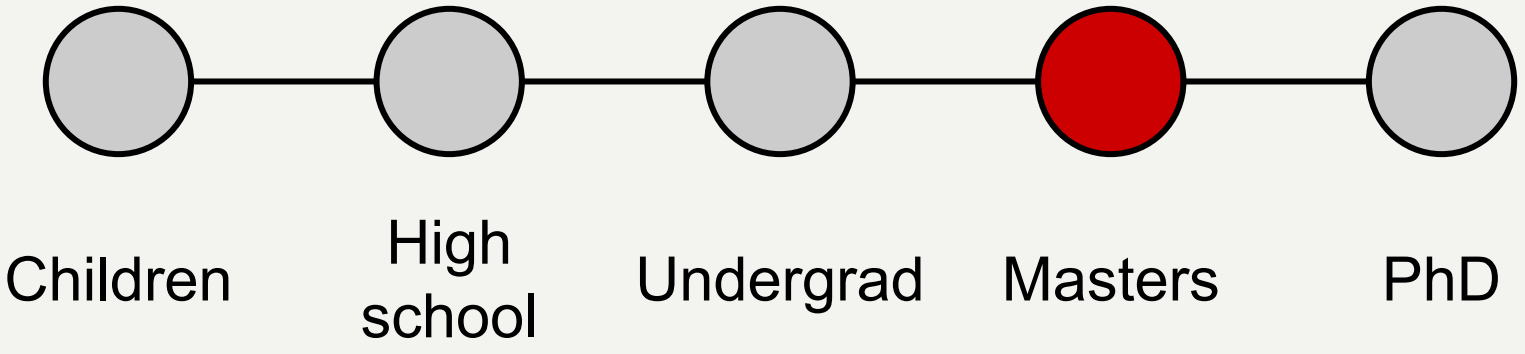
- Girls Learning Code, Vancouver 🇨🇦
- Try/CATCH Computing Conference @ SFU 🇨🇦

My teaching experience: Coding with VBA



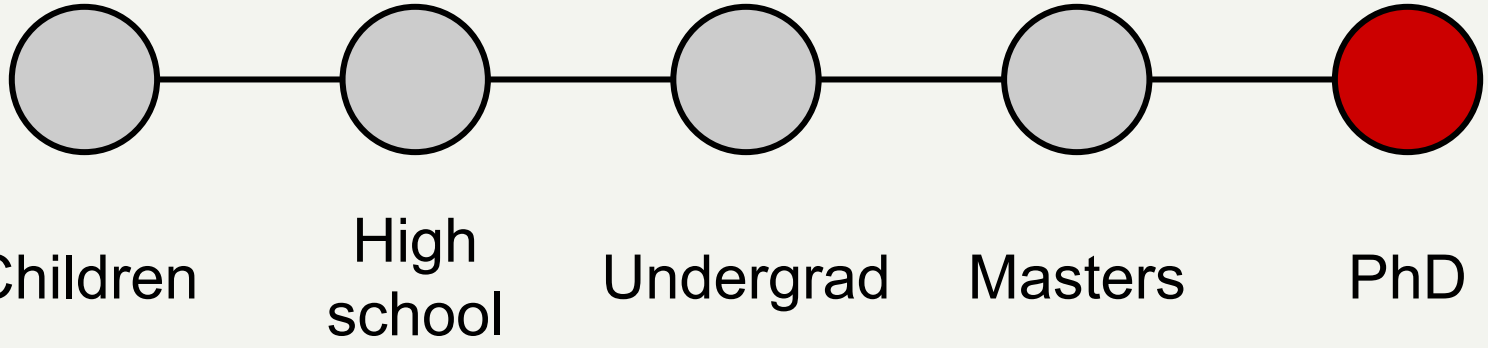
- WISE Workshop Series @ SFU 🇨🇦

My teaching experience: Coding with Git



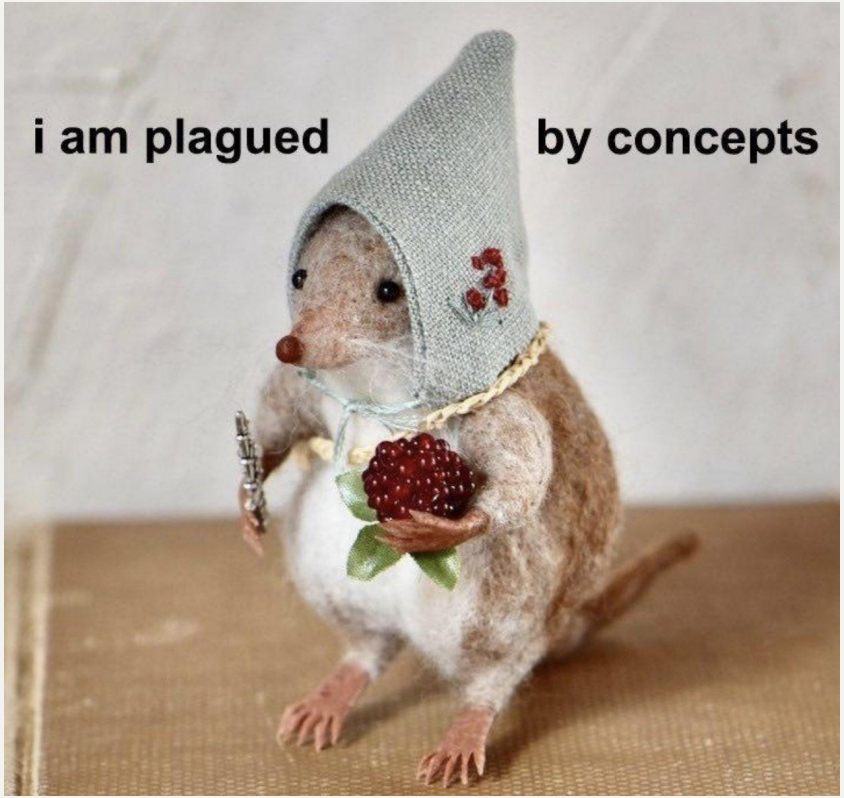
- Dialpad 🇨🇦
- Computational Linguistics @ Saarland U 🇩🇪

My teaching experience: Coding with Git



- Research Process Management @ Saarland U 🇩🇪

Course development: Concepts class 2025



Concepts class: Pitch

- **Abstract concepts** in NLP:

  interpretability, *bias*, *stereotypes, …*

- We have a **shared understanding** within the field
- But what do these terms actually **mean**?
- And what ***should*** they mean?

Teaching Philosophy

Transparency and structure

Diversity and inclusion

Hands-on engagement

Establishing norms

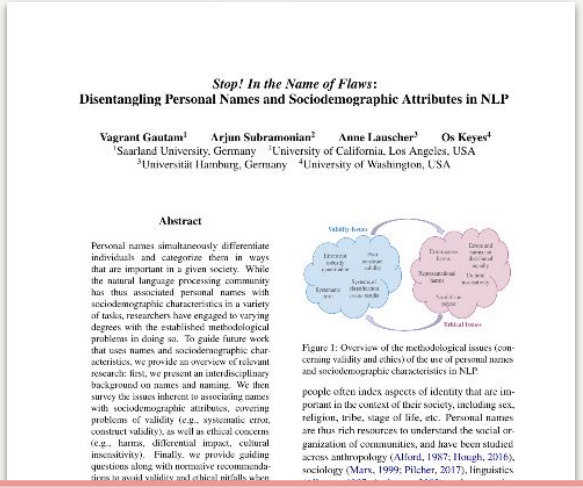
Interdisciplinarity

Concepts class: Learning objectives

Transparency and structure

- **Read** and **critique** papers
- **Critically evaluate** aspects of
  a. *conceptualization* (defining an abstract concept)
  b. *operationalization* (creating empirical measures of it)
- **Design projects** in ways that address critiques and *push the discipline forward*

Concepts class: Logistics



Stop! In the Name of Flaws:
Disentangling Personal Names and Sociodemographic Attributes in NLP

Vagrant Gautam[1] Arjun Subramonian[2] Anne Lauscher[3] Os Keyes[4]
[1]Saarland University, Germany [2]University of California, Los Angeles, USA
[3]Universität Hamburg, Germany [4]University of Washington, USA

Abstract

Personal names simultaneously differentiate individuals and categorize them in ways that are important in a given society. While the natural language processing community has thus associated personal names with sociodemographic characteristics in a variety of tasks, researchers have engaged to varying degrees with the established methodological problems in doing so. To guide future work that uses names and sociodemographic characteristics, we provide an overview of relevant research: first, we present an interdisciplinary background on names and naming. We then survey the issues inherent to associating names with sociodemographic attributes, covering problems of validity (e.g., systematic error, construct validity), as well as ethical concerns (e.g., harms, differential impact, cultural insensitivity). Finally, we provide guiding questions along with normative recommendations to avoid validity and ethical pitfalls when

Figure 1: Overview of the methodological issues (concerning validity and ethics) of the use of personal names and sociodemographic characteristics in NLP.

people often index aspects of identity that are important in the context of their society, including sex, religion, tribe, stage of life, etc. Personal names are thus rich resources to understand the social organization of communities, and have been studied across anthropology (Alford, 1987; Hough, 2016), sociology (Marx, 1999; Pilcher, 2017), linguistics

A Rose by Any Other Name would *not* Smell as Sweet: Social Bias in Name Mistranslations

Sandra Sandoval
University of Maryland
sandracs@umd.edu

Jieyu Zhao
University of Southern California
jieyuz@usc.edu

Marine Carpuat
University of Maryland
marine@umd.edu

Hal Daumé III
University of Maryland
Microsoft Research
hal3@umd.edu

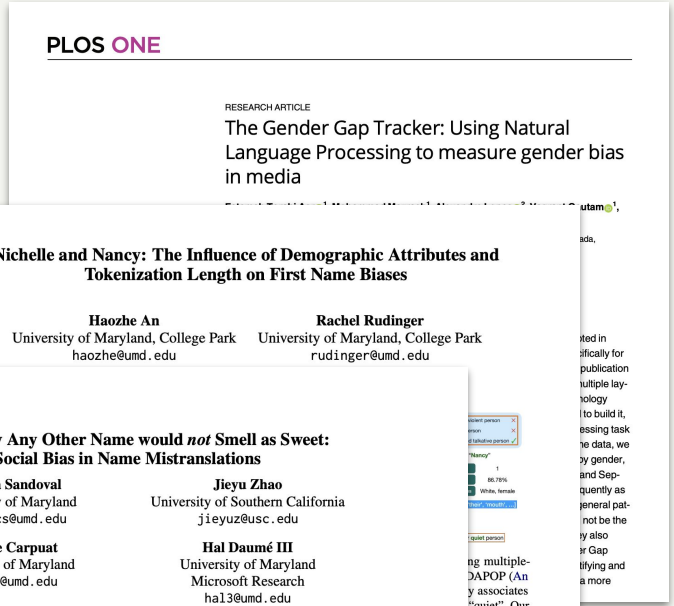
Abstract

We ask the question: *Are there widespread disparities in machine translations of names*

Source: Journee es una poeta británica de fuerza, claridad y oficio honesto.
Translation: Journee is a British poet of force, clarity and

translate female-associated names is significantly lower than male-associated names. This effect is particularly pronounced for female-associated names that are also associated with racial (Black) and ethnic (Hispanic) minorities. This disparity in translation quality between so-

frequent in online settings, whether via LinkedIn, professional email, Twitter, or Slack, or any other place where people address others or are addressed by name. In multilingual personal or workplace settings, displaying names incorrectly due to mis-



PLOS ONE

RESEARCH ARTICLE

The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media

Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases

Haozhe An
University of Maryland, College Park
haozhe@umd.edu

Rachel Rudinger
University of Maryland, College Park
rudinger@umd.edu

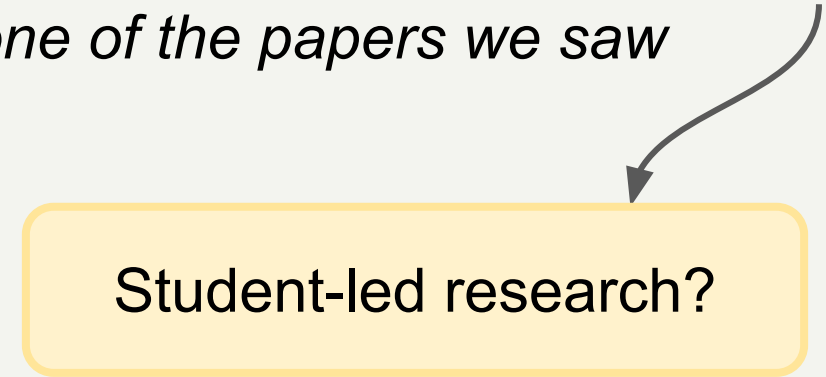
Analysis/critique + Papers to discuss

Concepts class: Logistics

- **Present** a concept + lead a discussion
- **Engage** with other presentations
- **Write** a report *designing a novel research project or reimagining one of the papers we saw*

Concepts class: Logistics

- **Present** a concept + lead a discussion
- **Engage** with other presentations
- **Write** a report *designing a novel research project or reimagining one of the papers we saw*



Concepts class: “Names”

Stop! In the Name of Flaws: Disentangling Names and Sociodemographic Attributes in NLP[1]

- Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases[2]
- The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media[3]

[1] Gautam et al. (2024)
[2] An and Rudinger (2023)
[3] Asr et al. (2021)

Other classes I could teach at REDACTED University

- First and second year courses
- ✨ **Bias**, **Fairness** and **Justice** in *Linguistics* ✨
- ✨ **Bias**, **Fairness** and **Justice** in *NLP* ✨
- ✨ Mixing Research Methods in **Interdisciplinary** Work ✨

Discussion topics?

- **Generative language models** in the classroom
- **Diversity**, **equity**, **inclusion** and **power** in teaching
- **Computational methods** in linguistics
- Connecting linguistics to **society**

Backup slides

Concepts class: “Explainability”

Explanation in artificial intelligence: Insights from the social sciences[1]

- Attention is not Explanation[2]
- Attention is not not Explanation[3]
- Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals[4]

[1] Tim Miller (2019)
[2] Jain and Wallace (2019)
[3] Wiegreffe and Pinter (2019)
[4] Elazar et al. (2021)