UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

Name Biases in Automated Essay Assessment

Ricardo Muñoz Sánchez

Outline

- Onomastics – the study of names
- Names and (human) essay grading
- Names and (automated) essay grading
- Pseudonymization

PART 1
What's in a name?

How humans perceive names

What are Onomastics?

- Onomastics is the study of proper names

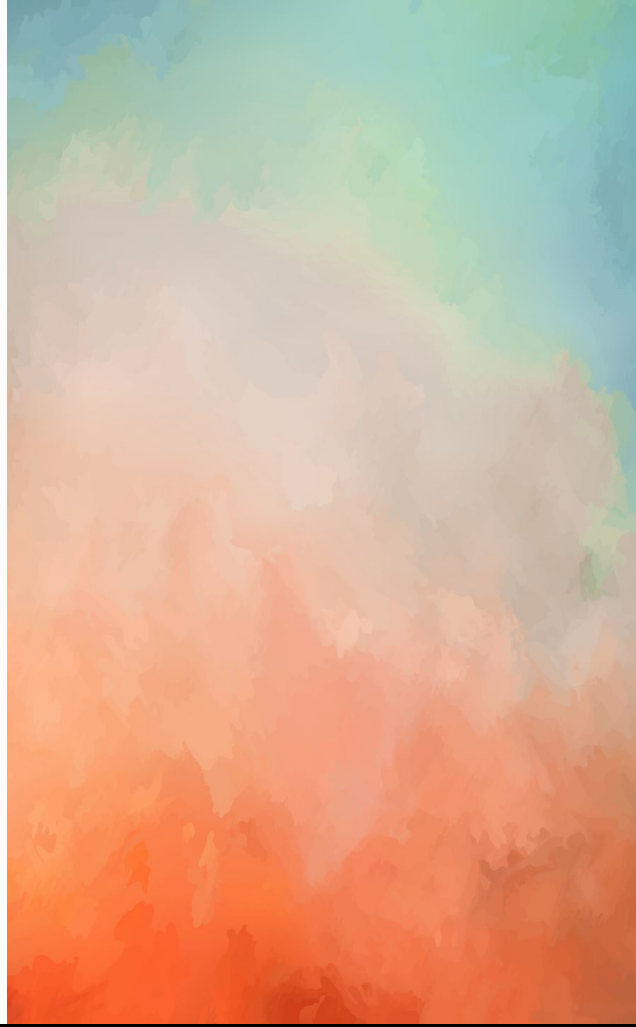
- This can be in a wide variety of contexts
  - Etymological
  - Historical
  - Social

- Names carry social and cultural context

Names Have Power

- We know that proper names affect how people are perceived
  - In job applications (Åslund and Skans 2012)
  - During grading (Anderson-Clark et al. 2008)
  - When looking to rent (Carpusor and Loges 2006)
  - Among many others…

- This can be an issue when dealing with high-stakes situations



PART 2
Call Me by Your Name

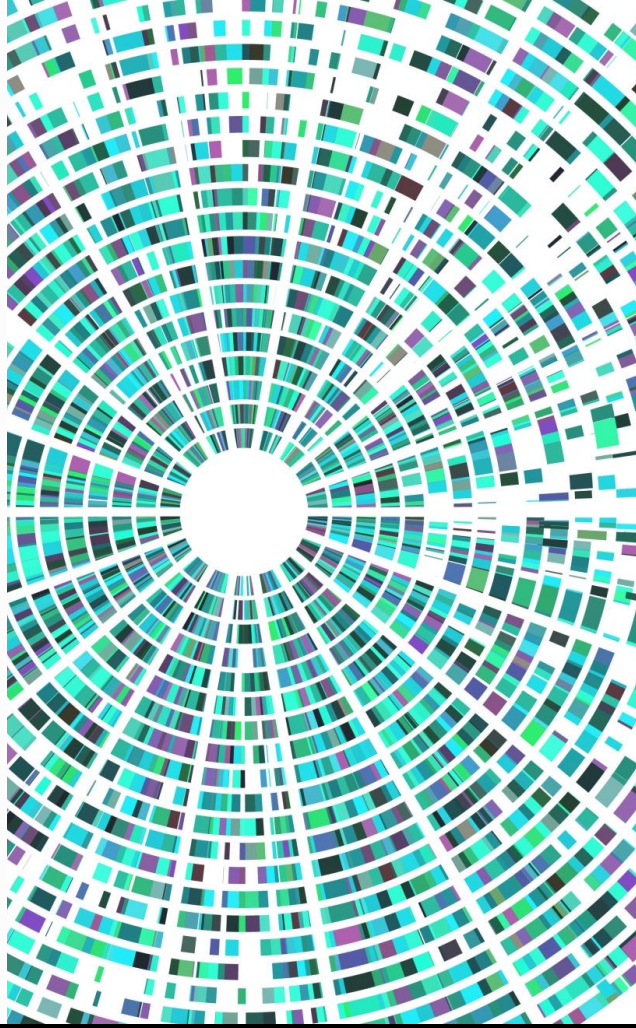
How names affect human grading

Checking for Human Biases

- We know that humans have implicit biases (Greenwald et al. 1998)
- These can reflect on how we perform on our day-to-day tasks
- On high-stakes situations, this can lead to undesirable results

Biases in Essay Grading

- Essay grading can be a high-stakes situation
- Students should be graded based on their knowledge and skills
- Discovering and acknowledging biases can reduce the impact they have



Assessing Names? (Aldrin 2017) – Design

Take an essay where a given name appears once

- The topic was “my childhood”
- The language of the essay was Swedish

Select three names with different sociocultural implications

- Carl, commonly associated with higher economic status
- Kevin, commonly associated with lower economic status
- Mohammed, an ethnically marked Muslim name

Substitute the names on the original essay

- This leads to three different versions of the essays

Randomly give a professional grader one of these three versions

- 113 high school teachers across Sweden graded the essays

Assessing Names? (Aldrin 2017) – Results

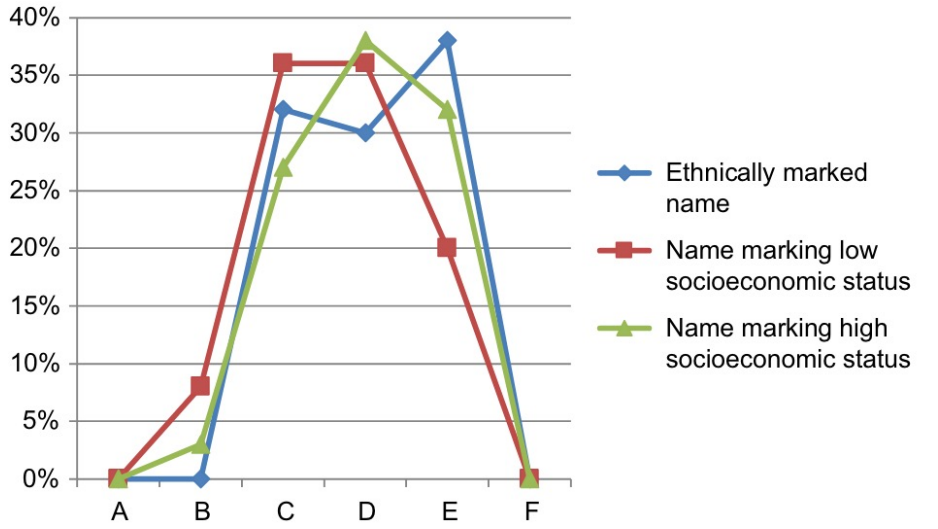
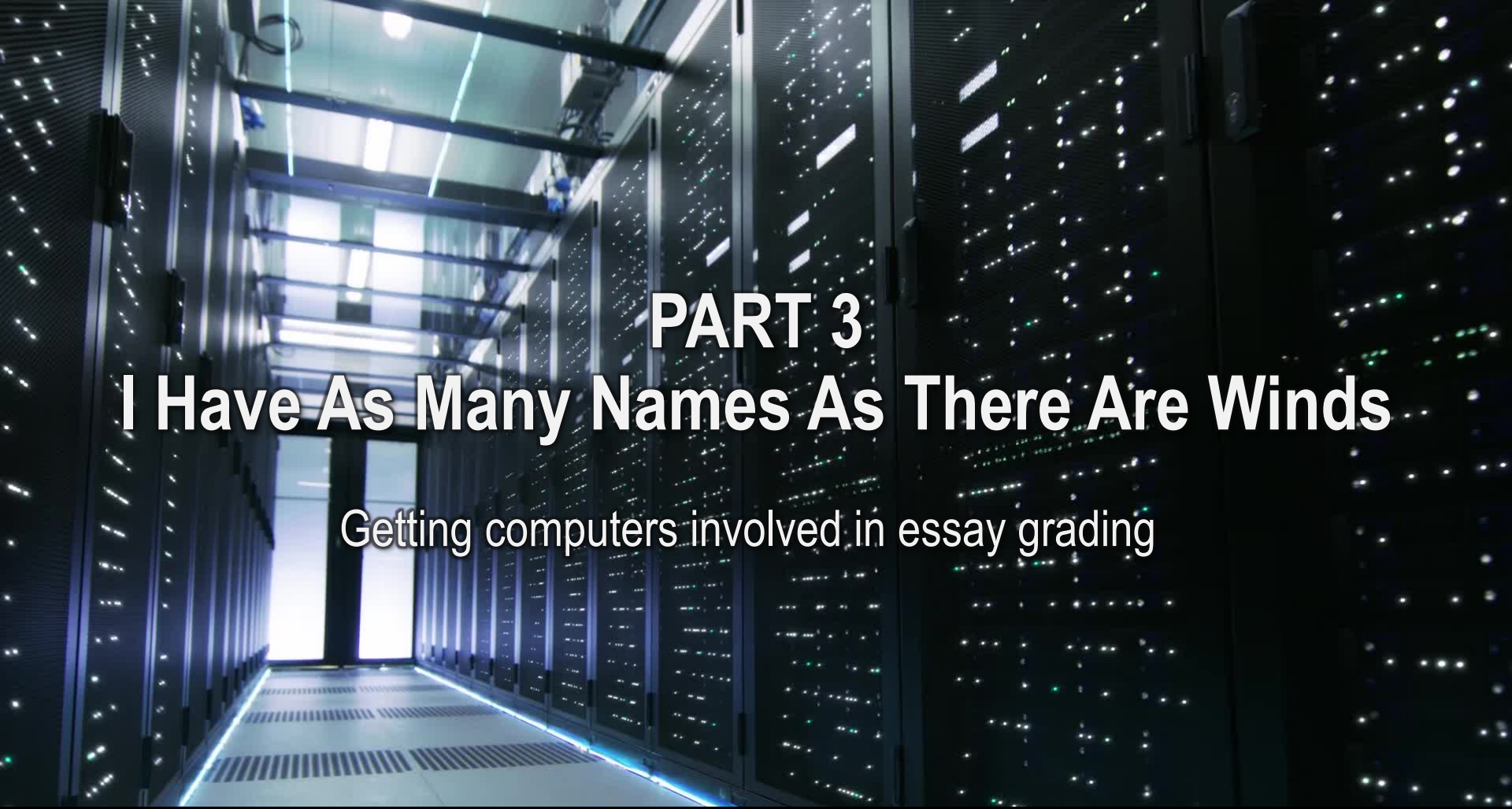


FIGURE 1 Teachers' general assessment of the text correlated to inserted name.

From "Assessing Names? Effects of Name-Based Stereotypes on Teachers' Evaluations of Pupils' Texts" by Aldrin (2017) [Link]

Assessing Names? (Aldrin 2017) – Conclusions

- The quantitative differences were small and not statistically significant
- The essay version with the Muslim-marked name
  - Tended to get lower grading across all rubrics
  - It also got the most comments on its deficiencies across three dimensions



PART 3
I Have As Many Names As There Are Winds

Getting computers involved in essay grading

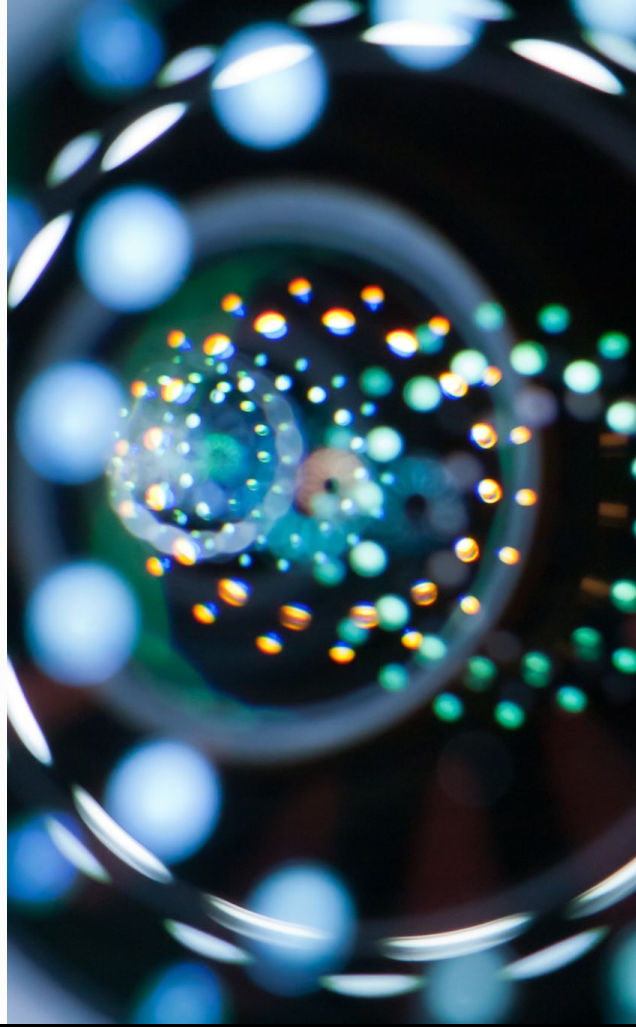
Bias in Machine Learning

- AI looks at insane amounts of data to learn
- It does so by looking at patterns and exploiting them
- However, human biases are reflected as patterns in the data
- This can affect the fairness of AI models

AI for Second Language Evaluation

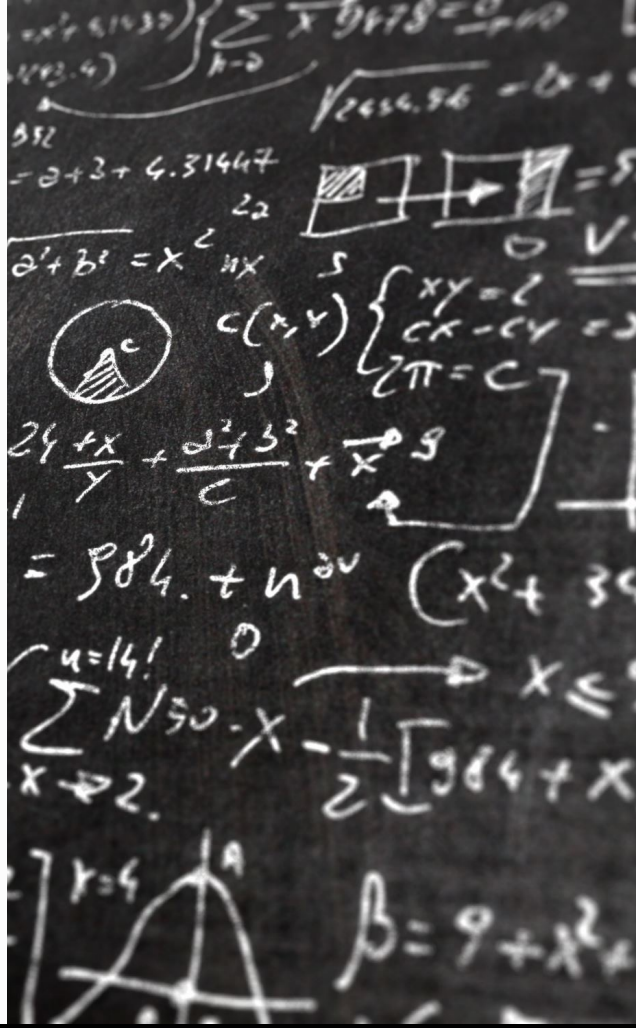
- The task
  - Given a second-language learner's essay, determine the CEFR level it belongs to

- Several ways to do it
  - Extract linguistic features + classical ML
  - Language models (BERT or GPT)
  - Smaller models to test different things



Checking for Human-Like Biases

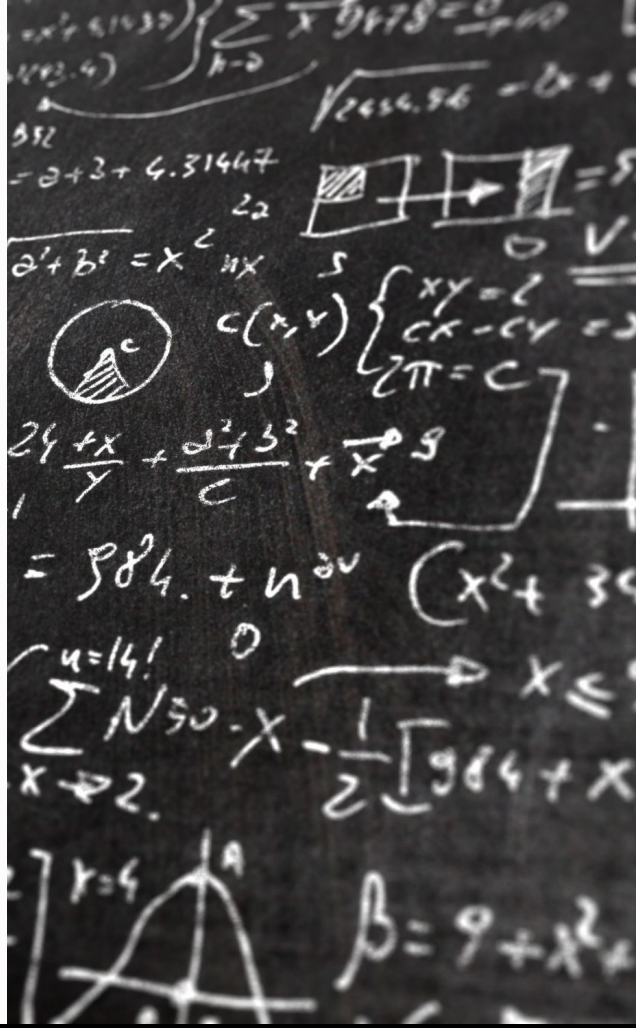
- Take 10 essays
  - Two for each CEFR level
  - From the Swell-Pilot corpus of L2 Swedish learner essays



Checking for Human-Like Biases

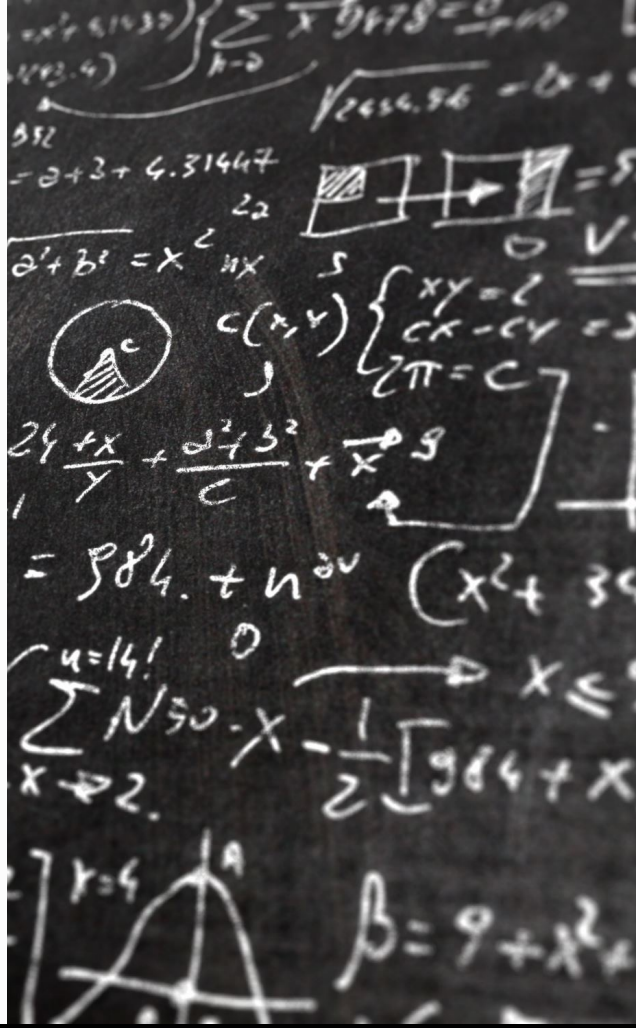
- Take 10 essays

- Generate a list of 20 names, for each of four ethnic groups
  - Swedish
  - Finnish
  - Anglo-American
  - Arabic



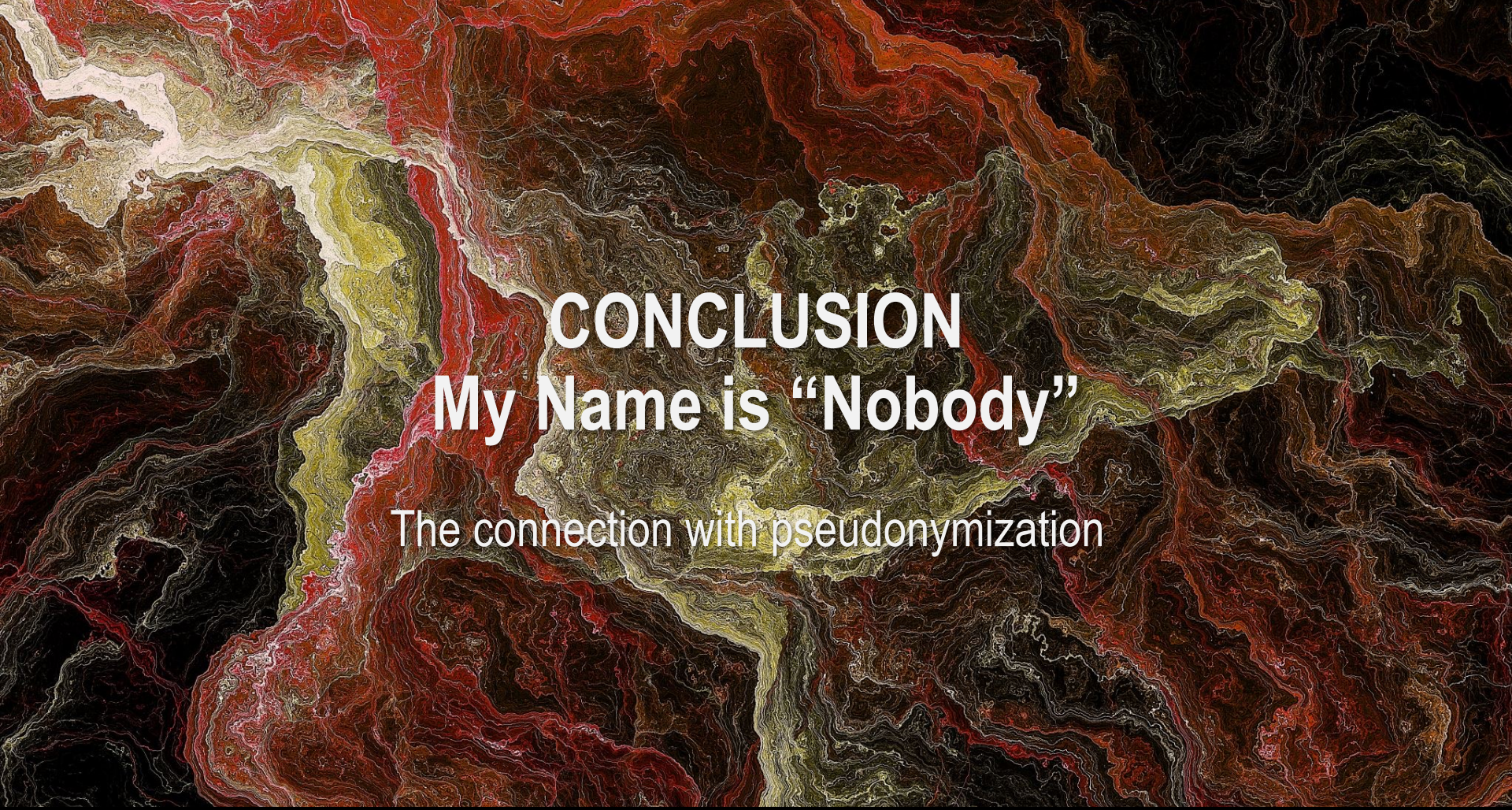
Checking for Human-Like Biases

- Take 10 essays
- Generate a list of 20 names, for each of four ethnic groups
- Substitute a given name in the original essay for one on the list



Measuring Fairness

- A model is fair if it performs equally for different subgroups
- An essay with a Swedish name in its text should be graded the same as the same essay with an Arabic name in its text
- If we find biases in one or more models, we can explore where they come from


CONCLUSION
My Name is “Nobody”
The connection with pseudonymization

Changing Names with Fairness in Mind

- We need to make sure these names don’t affect the outcome of automated systems
- There is a trade-off between privacy, fairness, and performance
- In the end we’re doing this to support people

Dogs have human names. It's what keeps them from being wolves.

- T. Kingfisher, Nettle & Bone

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io

Section Titles

- **Section 1** – most famously from Romeo and Juliet, a play by Shakespeare
- **Section 2** – is title of a book by André Aciman, later adapted into film by Luca Guadagnino
- **Section 3** – is from American Gods, a book from Neil Gaiman
- **Conclusion** – is the Pseudonym taken by Odysseus when talking to the cyclops Polyphemus in the Odyssey, an epic poem by Homer

Bibliography – Onomastics

- Aldrin, Emilia. “Assessing Names? Effects of Name-Based Stereotypes on Teachers’ Evaluations of Pupils’ Texts.” *Names* 65, no. 1 (January 2, 2017): 3–14. https://doi.org/10.1080/00277738.2016.1223116.

- Carpusor, Adrian G., and William E. Loges. “Rental Discrimination and Ethnicity in Names1.” *Journal of Applied Social Psychology* 36, no. 4 (2006): 934–52. https://doi.org/10.1111/j.0021-9029.2006.00050.x.

- Kozlowski, Diego, Dakota S. Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R. Sugimoto. “Avoiding Bias When Inferring Race Using Name-Based Approaches.” *PLOS ONE* 17, no. 3 (March 1, 2022): e0264270. https://doi.org/10.1371/journal.pone.0264270.

- Åslund, Olof, and Oskar Nordströum Skans. “Do Anonymous Job Application Procedures Level the Playing Field?” *ILR Review* 65, no. 1 (January 1, 2012): 82–107. https://doi.org/10.1177/001979391206500105.

Bibliography – Machine Learning

Bibliography – Bias and Fairness in AI