



UNIVERSITY OF  
GOTHENBURG



UNIVERSITY OF HELSINKI

# Did the Names I Used within My Essay Affect My Score? Diagnosing Name Biases in Automated Essay Scoring

**Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna,  
Therese Lindström Tiedemann, Elena Volodina**



# Outline

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



# PART 1

## What's in a name?

How humans perceive names

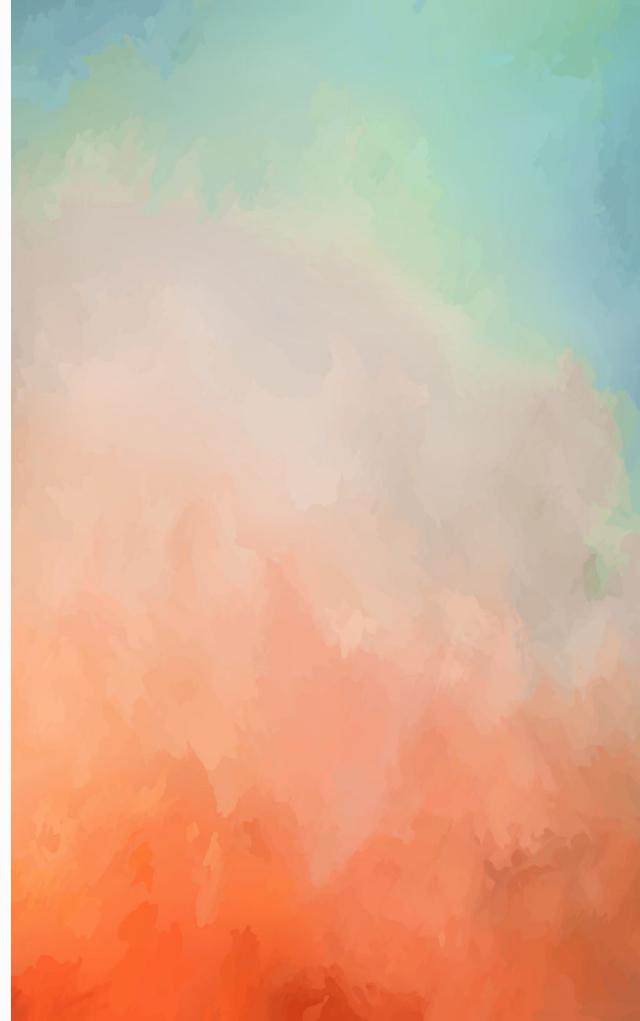
# What are Onomastics?

- 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

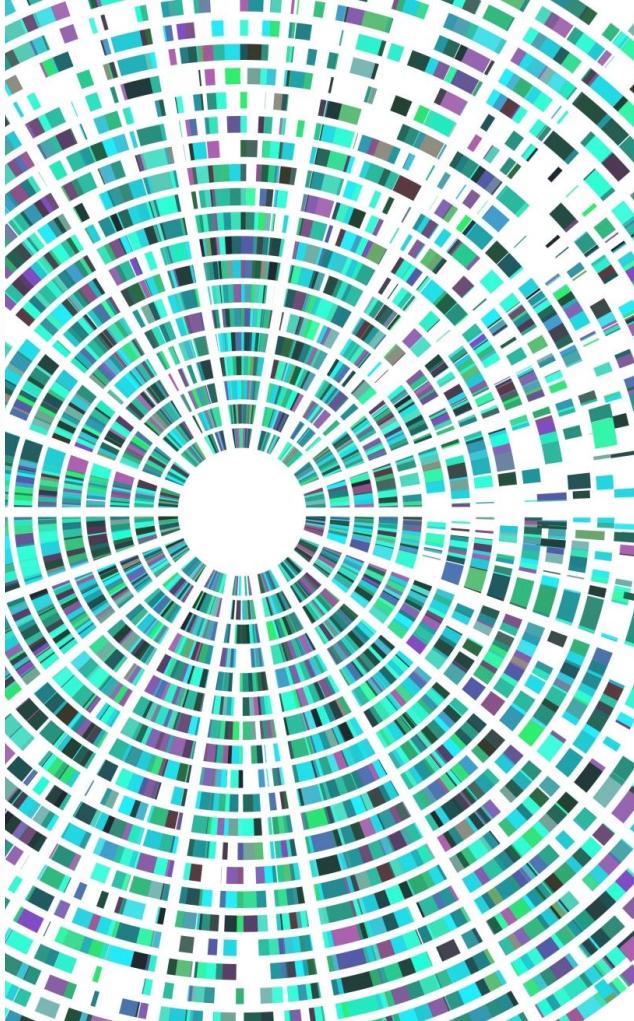
# Checking for Human Biases

- 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

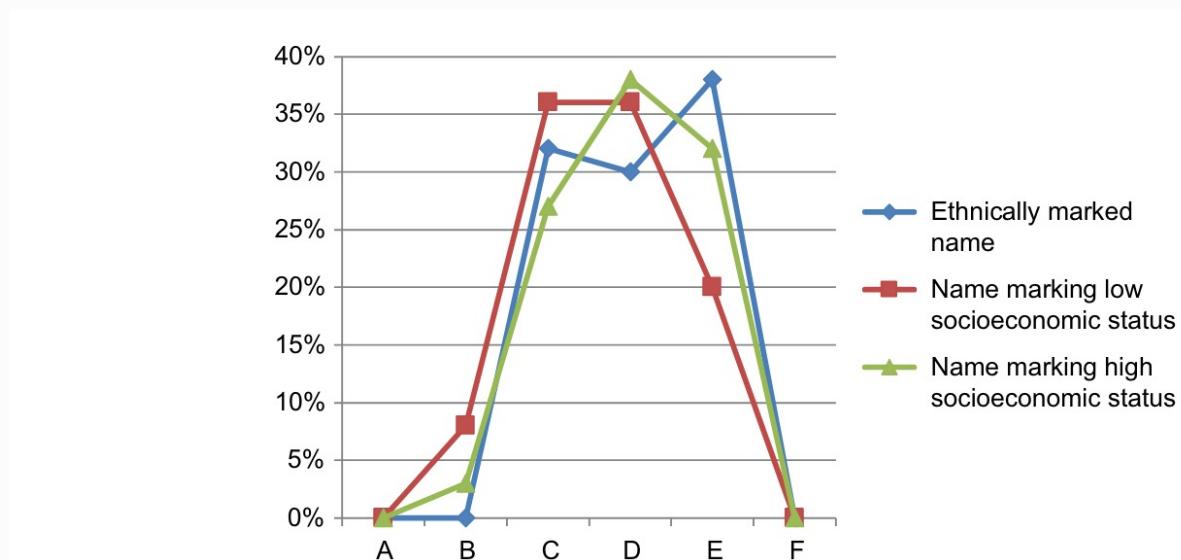
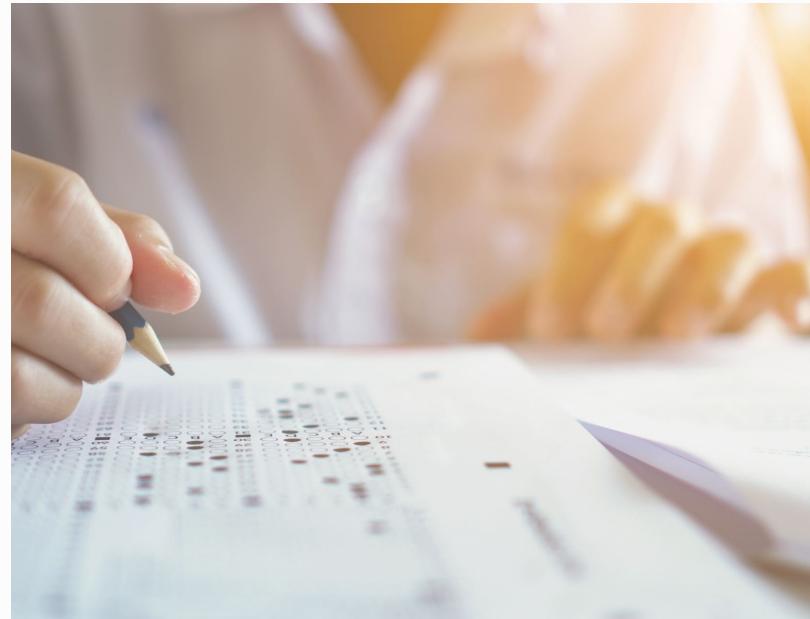


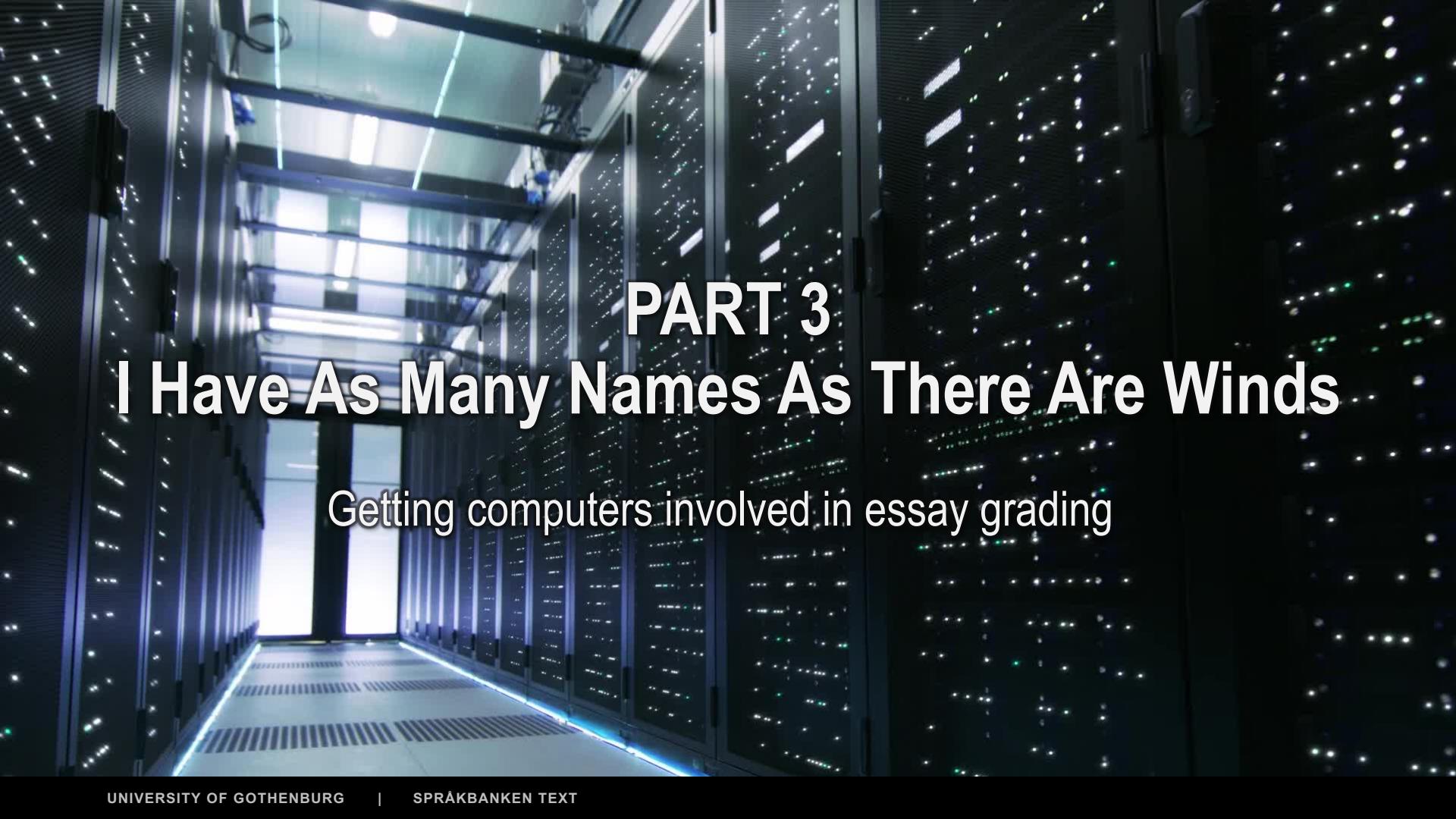
FIGURE 1 Teachers' general assessment of the text correlated to inserted name.

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

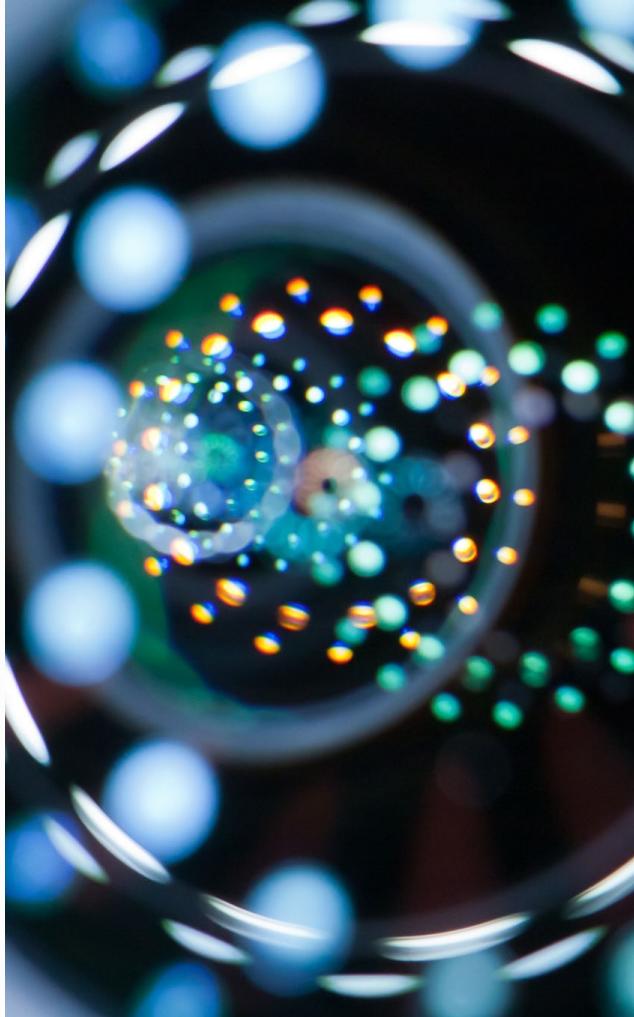
# Bias in Machine Learning

- 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



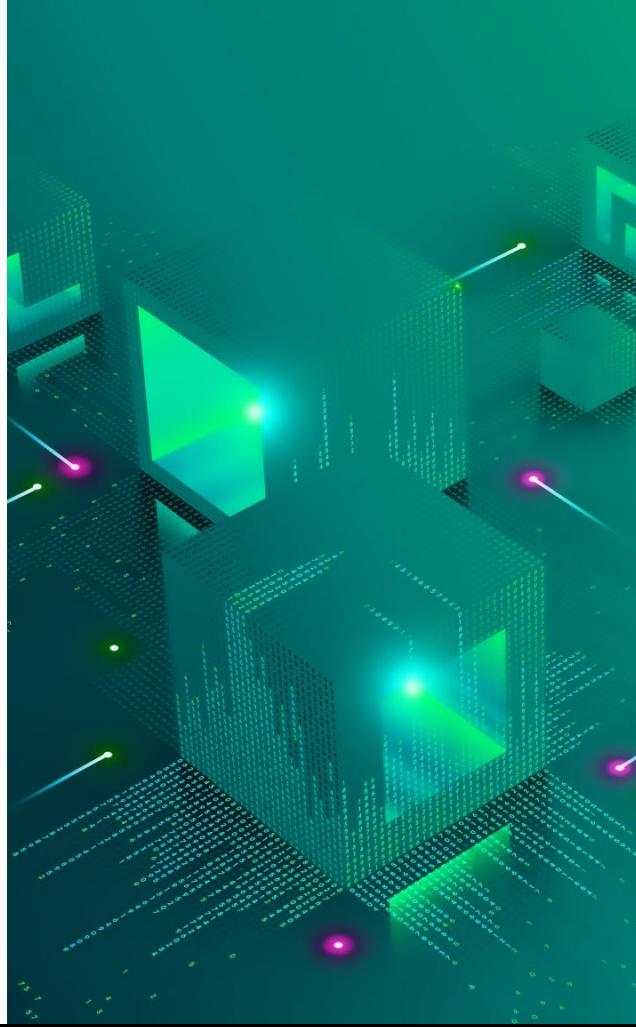
# Measuring Fairness

- 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



# Our Models

- Feature-based model (Pilán et al. 2016; Volodina et al. 2016b)
  - Uses lexical, morphological, syntactic and semantic features
  - We expected to find little to no bias at all
- Swedish BERT (Malmsten et al. 2020)
  - Learns syntax and semantics through context
  - This means it might have picked up biases during any stage of its training process



# Checking for Human-Like Biases

- Take 9 essays
  - Two for each CEFR level except for C1 (one essay) and C2 (no essays)
  - From the Swell-Pilot corpus of L2 Swedish learner essays

The image shows a blackboard with various mathematical notes and diagrams:

- At the top right, there is a diagram of a rectangle divided into four quadrants. The top-left quadrant is shaded, and the formula  $\frac{1}{2}ab = S$  is written next to it.
- Below the rectangle, there is a circle with radius  $c$  and center  $C$ . The formula  $c(x, y) \left\{ \begin{array}{l} xy = c \\ cx - cy = c \\ x^2 + y^2 = c^2 \end{array} \right.$  is written near the circle.
- In the center, there is a large bracketed equation involving  $x$ ,  $y$ , and  $c$ : 
$$\left[ \frac{xy}{c} + \frac{x^2 + y^2}{c} + \frac{xy}{S} \right] = 384. + n^{30} (x^2 + 30)$$
- To the left of the center equation, there is a note:  $n=14!$
- Below the center equation, there is a series of formulas:
  - $\sum N^{50} \cdot x - \frac{1}{2} [964 + x]$
  - $x \geq 2$
  - A graph of a function  $f(x)$  is shown, with a vertical tangent line at a point on the curve.
  - The formula  $\beta = 9 + x^2$  is written on the right.

# Checking for Human-Like Biases

- Take 9 essays
- Generate a list of 20 names, for each of four ethnic groups
  - Swedish
  - Finnish
  - Anglo-American
  - Arabic

A blackboard filled with mathematical calculations and diagrams. At the top right, there is a diagram of a rectangle divided into four quadrants, with arrows indicating a flow from one quadrant to another. Below this is a circle with a shaded sector. To the right of the circle is a system of equations:

$$\begin{cases} xy = c \\ cx - cy = s \\ 2\pi = c \end{cases}$$

Below the equations is a complex fraction:

$$\frac{2y+x}{y} + \frac{c^2 + s^2}{c} + \overrightarrow{x}s$$

Further down, there is a calculation involving factorials:

$$n! = 14!$$

On the left, there is a summation formula:

$$\sum N_{50} \cdot x - \frac{1}{2} [964 + x]$$

At the bottom, there is a graph of a function labeled  $\beta = 9 + x^2$ .

# Checking for Human-Like Biases

- Take 9 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

A blackboard filled with mathematical calculations and diagrams. At the top right, there is a diagram of a rectangle divided into four quadrants, with arrows indicating a flow from left to right and top to bottom. Below this is a circle with a shaded sector. To the right of the circle is a system of equations:

$$\begin{cases} xy = c \\ cx - cy = s \\ 2\pi = c \end{cases}$$

Below the equations is a complex fraction:

$$\frac{2y+x}{y} + \frac{c^2 + 3^2}{c} + \overline{x}^2 s$$

Further down, there is a calculation involving factorials:

$$n! = 14!$$

At the bottom, there is a graph of a function labeled  $\beta = 9 + x^2$ .



# Part 4

# My Name is “Nobody”

Results and conclusions

# Performance on the Test Set

<b>Model</b>	<b>Accuracy</b>	<b>F1 Macro</b>	<b>F1 Weighted</b>
Feature-Based	0.25	0.08	0.1
BERT	0.66	0.65	0.65

# Performance on the Diagnostic Set

<b>Name Groups</b>	<b>Feature-Based</b>		<b>BERT</b>	
	<b>Accuracy</b>	<b>Recall</b>	<b>Accuracy</b>	<b>Recall</b>
Swedish	0.14	0.20	0.86	0.60
Finnish	0.14	0.20	0.86	0.60
Anglo-American	0.14	0.20	0.86	0.60
Arabic	0.14	0.20	0.86	0.60

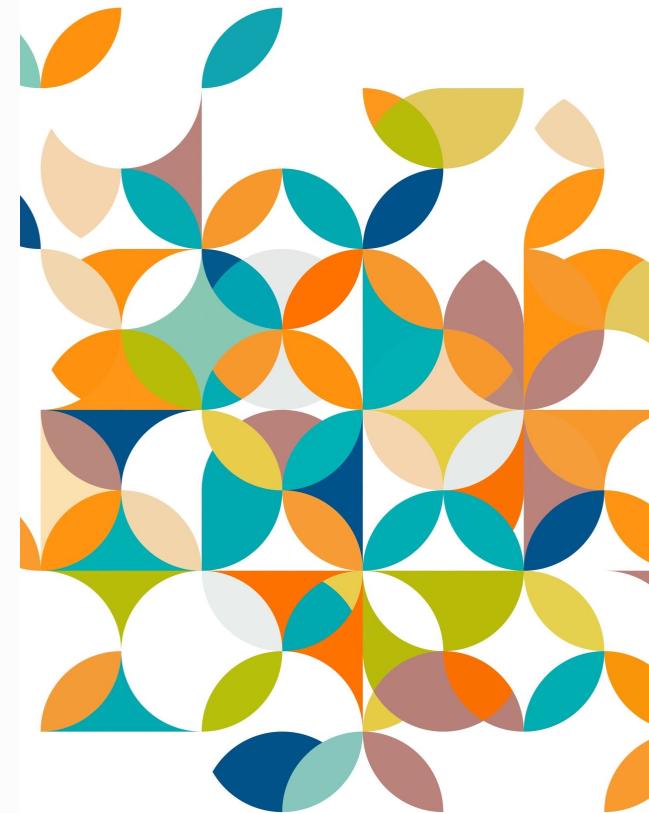
# What Does this Mean?

- Changing a single name within an essay did not change the models' performance
- The performance does not change either when taking (binary) gender into account
- This is what we would expect from a fair automated essay assessment system



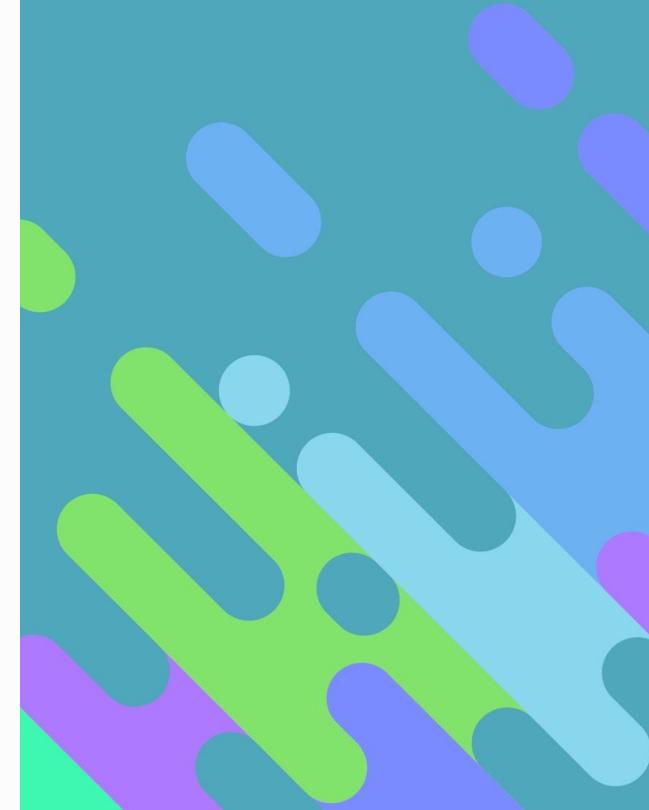
# Things to Keep in Mind

- The sample size from which the diagnostic set was generated is small
- We did not account whether the names were in BERT's dictionary or not
- This does not mean that neither the models nor the data are free of biases



# 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



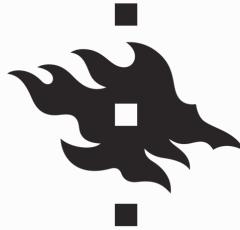


Dogs have human names.  
It's what keeps them from  
being wolves.

- T. Kingfisher, Nettle & Bone



GÖTEBORGS  
UNIVERSITET



UNIVERSITY OF HELSINKI

# SPRÅKBANKENTEXT

**Ricardo Muñoz Sánchez**  
[ricardo.munoz.sanchez@svenska.gu.se](mailto:ricardo.munoz.sanchez@svenska.gu.se)  
[rimusa.github.io](http://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