# GUEST LECTURE: DATA

School of Informatics Professional Issues Course

27 November 2026

Bríd-Áine Parnell




THE UNIVERSITY *of* EDINBURGH
informatics

Designing Responsible Natural Language Processing




UK Research and Innovation

# A BIT ABOUT ME…

- PhD student in the CDT for Designing Responsible Natural Language Processing
- Interested in multilinguality of digital life, sociotechnical studies, linguistics, data and algorithmic justice.

**SUPERVISION TEAM**

**Alex Taylor**
**Reader in Design Informatics**

**Adam Lopez**
**Reader in Informatics**

# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

# WHAT IS DATA?



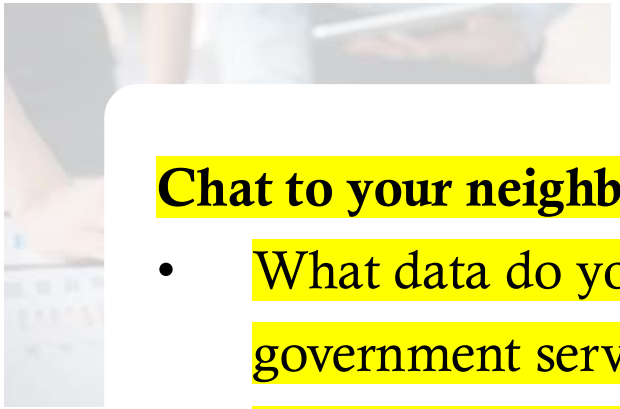






# WHAT IS DATA?

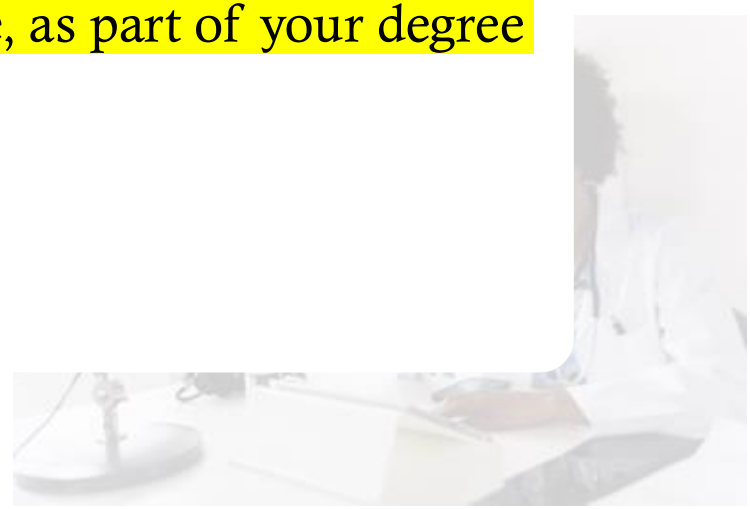




**Chat to your neighbour (2 mins):**

- What data do you use online, for social media, banking, health or government services, etc?
- What data or datasets have you used, or might you use, as part of your degree or work?
- How often do these two overlap?





# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

# RESPONSIBLE DATA USE: KEY ISSUES

## #1 WHAT'S IN THE DATA?

- Privacy[1] – personally identifiable information (PII)
- Representation[2] – large dataset ≠ lots of demographics
- Toxicity[3] – illegal or unwanted content

# RESPONSIBLE DATA USE: KEY ISSUES

## #1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

## #2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4] – was it obtained and *what* was it obtained for?
- Copyright[5] – what is fair use?
- Sovereignty[6] – who 'owns' the data?
- Synthetic data[7]

# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

# CASE STUDY: CRISIS TEXT LINE*

- Crisis Text Line is a not-for-profit mental health texting service.

*https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617

# CASE STUDY: CRISIS TEXT LINE*

- Crisis Text Line is a not-for-profit mental health texting service.
- In 2022, *Politico* broke the story that CTL was giving its "largest mental health data set in the world" to its for-profit spinoff Loris.ai to train a customer service AI.

*https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617

# CASE STUDY: CRISIS TEXT LINE*

- Crisis Text Line is a not-for-profit mental health texting service.
- In 2022, *Politico* broke the story that CTL was giving its "largest mental health data set in the world" to its for-profit spinoff Loris.ai to train a customer service AI.

**Chat to your neighbour (2 mins):**

- What are the key data issues for CTL?
- What are the ethical implications of using its data in this way?

*https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617

# CASE STUDY: CRISIS TEXT LINE*

- Crisis Text Line is a not-for-profit mental health texting service.
- In 2022, *Politico* broke the story that CTL was giving its "largest mental health data set in the world" to its for-profit spinoff Loris.ai to train a customer service AI.
- Company lawyers argued that agreement was legal and "ethically sound" because users gave consent and data sharing was in the Ts&Cs…

*https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617

# CASE STUDY: CRISIS TEXT LINE*

- Crisis Text Line is a not-for-profit mental health texting service.
- In 2022, *Politico* broke the story that CTL was giving its "largest mental health data set in the world" to its for-profit spinoff Loris.ai to train a customer service AI.
- Company lawyers argued that agreement was legal and "ethically sound" because users gave consent and data sharing was in the Ts&Cs…
- …But three days later, it announced it was terminating the agreement.

*https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617

Data harms are intimately connected to the context of data creation and use

# QUESTIONS?

# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

# OVERVIEW OF DATA HARMS

## #1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

## #2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4]
- Copyright? [5]
- Sovereignty[6]
- Synthetic data[7]

## #3 WHO LABELLED IT?

- Subjectivity[8] – what is the position of the participants, annotators and researchers?
- Annotators[9] – where and how are they sourced?

# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

# CASE STUDY: MULTILINGUAL NAME DISCRIMINATION IN LLM HIRING

- LLMs increasingly deployed in HR for decision-making.
- Previous studies have found bias on gender and race.
- Are there biases based on language, specifically names with diacritics:

  Irish – Seán, Máire, Croí
  French – Léna, Timéo, Céleste
  Portugese – João, Inês, Lourenço
  Latvian – Amēlija, Ričards, Luīze

# CASE STUDY: MULTILINGUAL NAME DISCRIMINATION IN LLM HIRING

- LLMs increasingly deployed in HR for decision-making.
- Previous studies have found bias on gender and race.
- Are there biases based on language, specifically names with diacritics:

Irish – Seán, Máire, Croí
French – Léna, Timéo, Céleste
Portugese – João, Inês, Lourenço
Latvian – Amēlija, Ričards, Luīze

**Chat to your neighbour (2 mins):**

- How would you go about gathering your name dataset? Sources? Methods?
- What are the key issues that you might face?

# CASE STUDY: MULTILINGUAL NAME DISCRIMINATION IN LLM HIRING

- Popular names
- Baby name sites
- Archaic
- What are you measuring?

# CASE STUDY: MULTILINGUAL NAME DISCRIMINATION IN LLM HIRING

- What is an Irish name? Or an English name? Or a French one?
- Language contact and –icisations
  - Darragh for Dáire, Gaelicisation: Seosamhín for Josephine
- Names are not good proxies for anything else.
- Again, what are you measuring?

Researchers must understand and account for subjectivity and interpretation of data

# QUESTIONS?

# RESPONSIBLE DATA USE: KEY ISSUES

## #1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

## #2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4]
- Copyright? [5]
- Sovereignty[6]
- Synthetic data[7]

## #3 WHO LABELLED IT?

- Annotators[8]
- Subjectivity[9]

## #4 WHERE IS IT KEPT?

- Secure storage[10]
- Environmental costs[11]
- Open access?

Transparency is key to mitigating data harms.

# OVERVIEW

1. What is data?
2. Responsible data use: key issues
3. **Case Study**: Data in context
4. Responsible data use: key issues
5. **Case Study**: Categories and subjectivity
6. Sum up & recommendations

1 Data harms are intimately connected to the context of data creation and use

2 Researchers must understand and account for subjectivity and interpretation of data

3 Transparency is key to mitigating data harms

- Follow legal and regulatory guidance
- Follow your institution's ethics guidelines
- Think about who the data and the annotations include and exclude
- Think about whether the data can achieve the purposes you want it to
- Document and acknowledge any limitations

# QUESTIONS?

# REFERENCES


1. Lukas, Nils, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. "Analyzing leakage of personally identifiable information in language models." In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346-363. IEEE, 2023.

2. Adams, Julia, Hannah Brückner, and Cambria Naslund. "Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test"." *Socius* 5 (2019): 2378023118823946; Yulin Yu, Xianglong Li, Tianyi Li, Paramveer S. Dhillon and Daniel M. Romero. 2025. Demographic disparity in Wikipedia coverage: a global perspective. *EPJ Data Science*, *14*(1), 15; 10. Kevin Guyan (2022) Constructing a queer population? Asking about sexual orientation in Scotland's 2022 census, Journal of Gender Studies, 31:6, 782-792, DOI: 10.1080/09589236.2020.1866513

3. Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

4. Harry Davies and Yuval Abraham. 2025. " 'A million calls an hour': Israel relying on Microsoft cloud for expansive surveillance of Palestinians". *The Guardian;* Iqbal, Umar, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel J. Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. "Tracking, profiling, and ad targeting in the Alexa echo smart speaker ecosystem." In *Proceedings of the 2023 ACM on Internet Measurement Conference*, pp. 569-583. 2023; Robert Hart. 2024. Clearview AI—Controversial Facial Recognition Firm—Fined $33 Million For 'Illegal Database'. *Forbes;* Halavais, Alexander. "Overcoming terms of service: A proposal for ethical distributed research." *Information, Communication & Society* 22, no. 11 (2019): 1567-1581.

5. Sag, Matthew, and Peter K. Yu. "The globalization of copyright exceptions for AI training." *Emory LJ* 74 (2024): 1163; Dan Milmo. 2025. Anthropic did not breach copyright when training AI on books without permission, court rules. *The Guardian.*

# REFERENCES


6. Kukutai, Tahu, and John Taylor. *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016; Mager, Manuel, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. "Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers." *arXiv preprint arXiv:2305.19474* (2023); Karen Hao, 2022, A new vision of artificial intelligence for the people, *MIT Technology Review.*

7. Whitney, Cedric Deslandes, and Justin Norman. 2024. "Real risks of fake data: Synthetic data, diversity-washing and consent circumvention." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1733-1744.

8. Haliburton, Luke, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. "Investigating labeler bias in face annotation for machine learning." *arXiv preprint arXiv:2301.09902* (2023); Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection." *arXiv preprint arXiv:2111.07997* (2021).

9. Chandhiramowuli, Srravya, Alex S. Taylor, Sara Heitlinger, and Ding Wang, 2024. "Making data work count." *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW1: 1-26; Robert Booth, 2024, "More than 140 Kenya Facebook moderators diagnosed with severe PTSD", *The Guardian.*

10. Joe Tidy, 2025, "Nursery hackers threaten to publish more children's profiles online", *BBC News.*

11. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?🦜." In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610-623. 2021.