THE UNIVERSITY OF MELBOURNE

Social Media Sentiment Matters

data science, social data & their ethics

Dr Marc Cheong
Senior Fellow (MLS) & Senior Research Fellow in Digital Ethics (CAIDE) & Honorary Burnet Institute Senior Fellow

14th October 2021

About me

CompSci - Did Twitter research ca. 2009-2013

- Was a wonderful world back then, generous support by Twitter, less toxicity online
- Lucky to jump in then, ~3 dozen papers or so worldwide!

Pivoted to philosophical research

- Introducing data science to philosophers (digital humanities/experimental philosophy)
- Digital Ethics (philosophical ethics + data science + practical aspects)
- Contemporary social media usage and social networking trends.

All cover image credits: Unsplash (@firmbee /@prateekkatyal)

Ethics in social media data analysis

Not just the realm of ethicists/philosophers, lawyers, etc.

The ML/Data Science community is fast realizing its importance – e.g., FACCT (left) and NeurIPS (right).

Algorithmic systems are being adopted in a growing number of contexts, fueled by bi
filter, sort, score, recommend, personalize, and otherwise shape human experience, i
informing decisions with major impact on access to, e.g., credit, insurance, healthcare
and immigration. Although these systems may bring myriad benefits, they also contai
codifying and entrenching biases; reducing accountability, and hindering due process
information asymmetry between individuals whose data feed into these systems and
inferring potentially relevant information.

ACM FAccT is an interdisciplinary conference dedicated to bringing together a diverse
from computer science, law, social sciences, and humanities to investigate and tackle
area. Research challenges are not limited to technological solutions regarding potenti
question of whether decisions should be outsourced to data- and code-driven compu
particularly seek to evaluate technical solutions with respect to existing problems, ref
benefits and risks; to address pivotal questions about economic incentive structures,
distribution of power, and redistribution of welfare; and to ground research on fairne
transparency in existing legal requirements.

Announcing the NeurIPS 2021 Datasets and Benchmarks Track

COMMS CHAIRS | Uncategorized | 0

Joaquin Vanschoren and Serena Yeung

There are no good models without good data (Sambasivan et al. 2021). The vast majority of the NeurIPS community focuses on algorithm design, but often can't easily find good datasets to evaluate their algorithms in a way that is maximally useful for the community and/or practitioners. Hence, many researchers resort to data that are conveniently available, but not representative of real applications. For instance, many algorithms are only evaluated on toy problems, or data that is plagued with bias, which could lead to biased models or misleading results, and subsequent public criticism of the field (Paullada et al. 2020).

Researchers are often incentivized to benchmark their methods on a handful of popular datasets that have been well established in the field, with state-of-the-art results on these key benchmark datasets helping to secure a paper acceptance. Conversely, evaluations on lesser known real-world datasets, and other benchmarking efforts to connect models to real world impacts, are often harder to publish and are consequently devalued within the field.

Overview

Case study: Social media metadata – interactive lecture 💬

- Studying the public reaction to the Black Lives Matter – movement for equality & social good
- **The ethics behind this!**

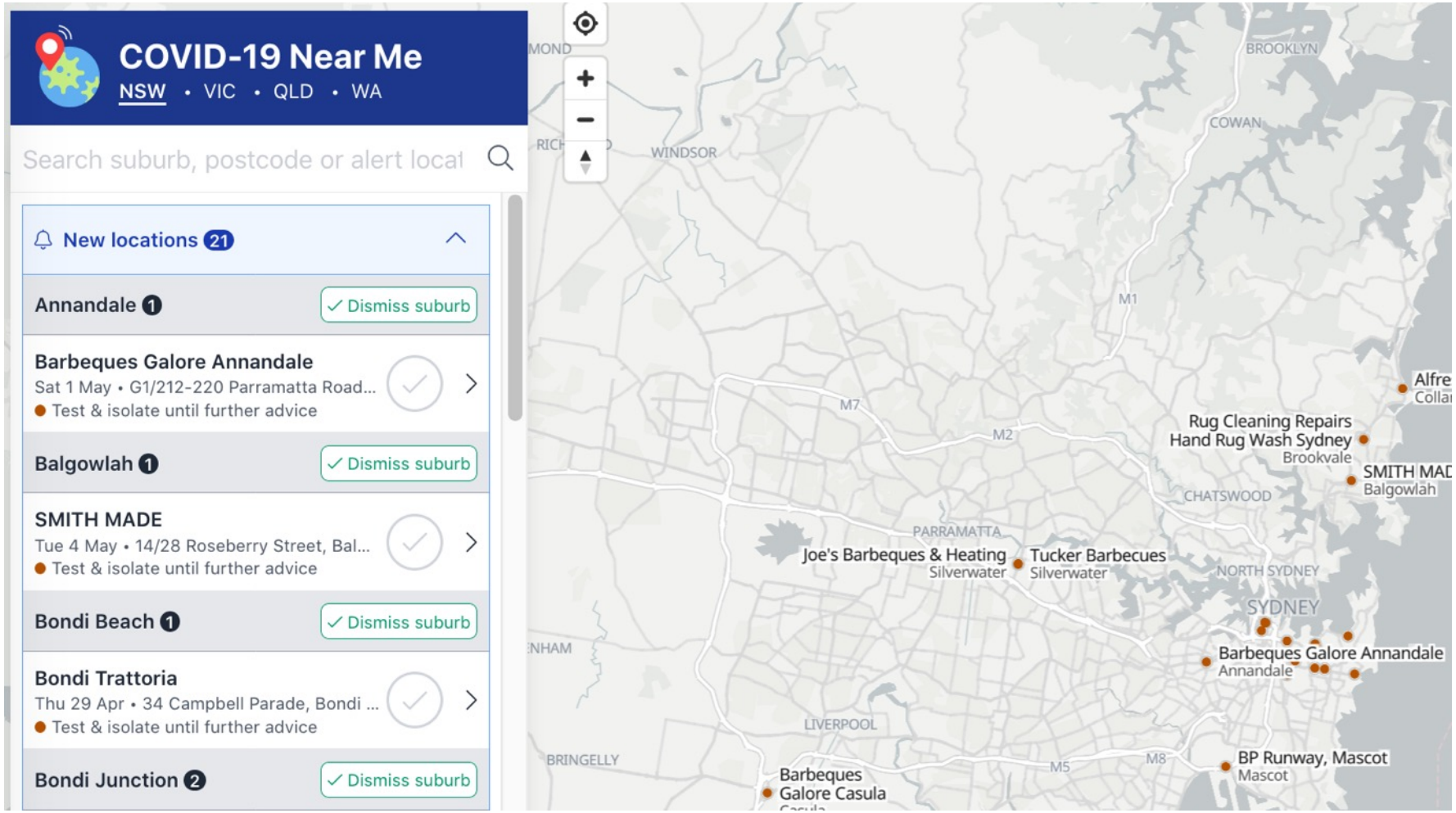
Enter: Digital ethics & social media sentiment mining – interactive lecture

- Ethics and the ML pipeline.
  - 💬 **Ethical discussions include scenarios, mitigating risks, techniques?**
- Coghlan et al – the ethical principles
- Further Discussion.

THE UNIVERSITY OF MELBOURNE

Case study: Geo data and social media

Geo data: helping us in COVID-19…



Aggregating data from official public health websites and visualizes them (including time metadata etc)
Tsang (2021). https://covid19nearme.com.au/

…beware danger!

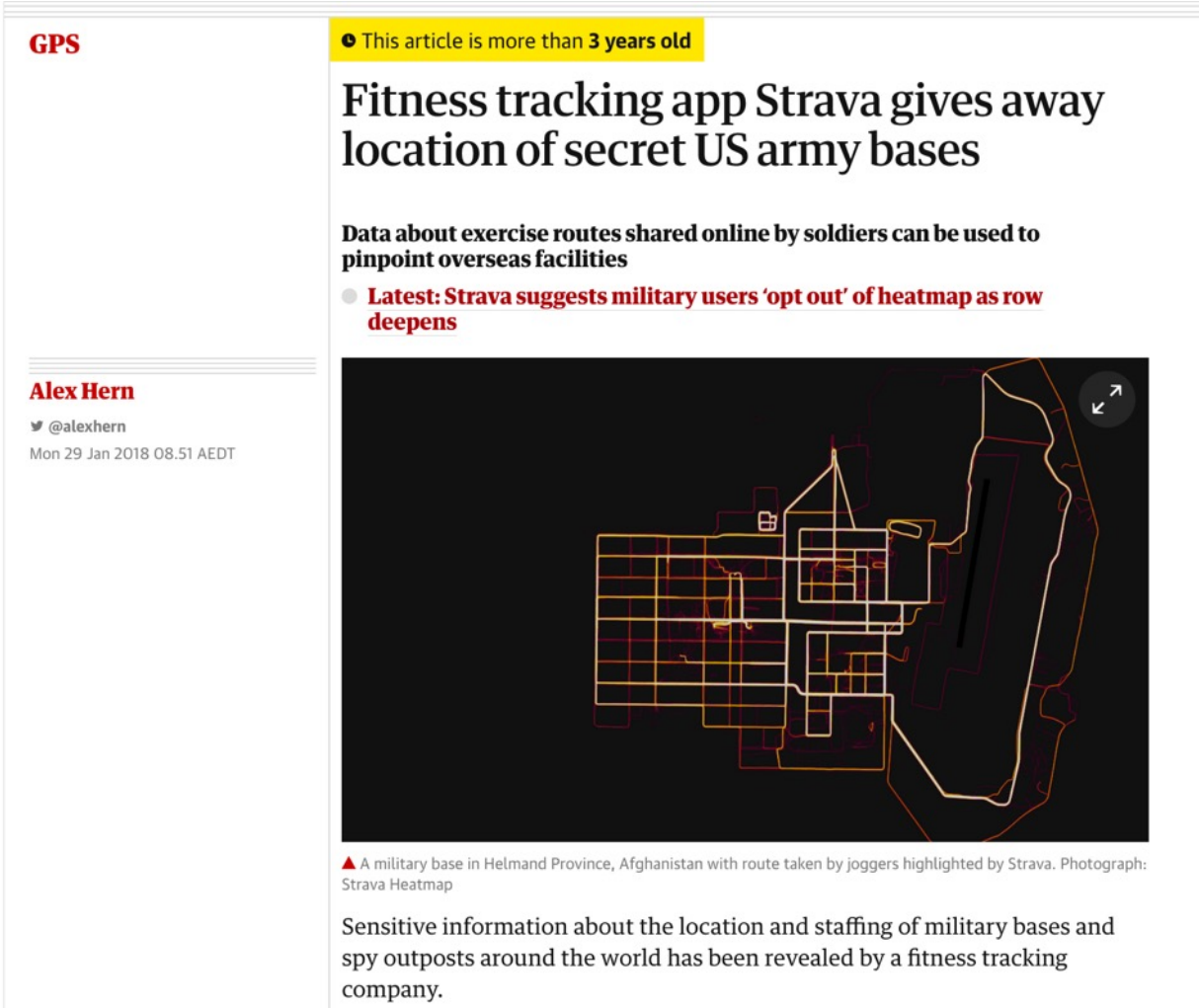
"Sensitive information about the location and staffing of military bases and spy outposts around the world has been revealed by a fitness tracking company…

…data visualisation map that shows all the activity tracked by users of its app, which allows people to record their exercise and share it with others.

The map, released in November 2017, shows every single activity ever uploaded to Strava – more than 3 trillion individual GPS data points, according to the company.."

Hern (2018) for The Guardian

https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases



GPS

This article is more than **3 years old**

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

Latest: Strava suggests military users 'opt out' of heatmap as row deepens

Alex Hern

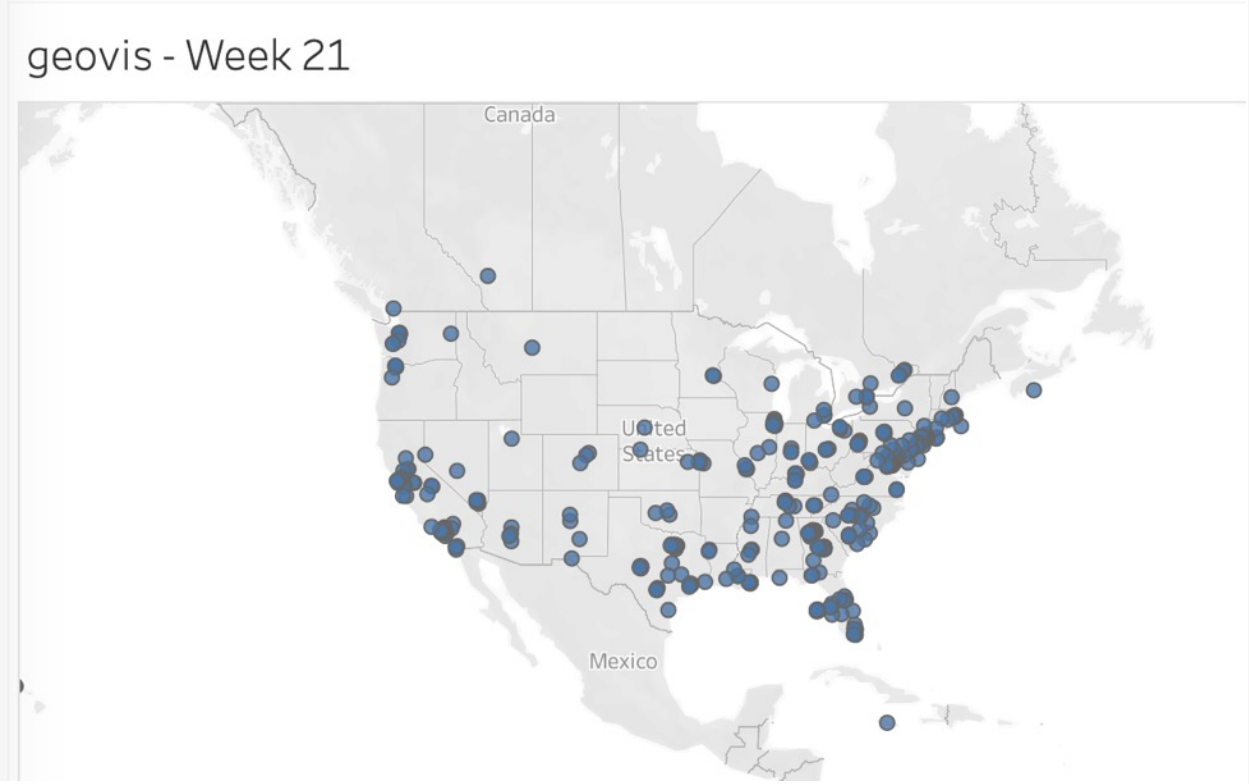
@alexhern

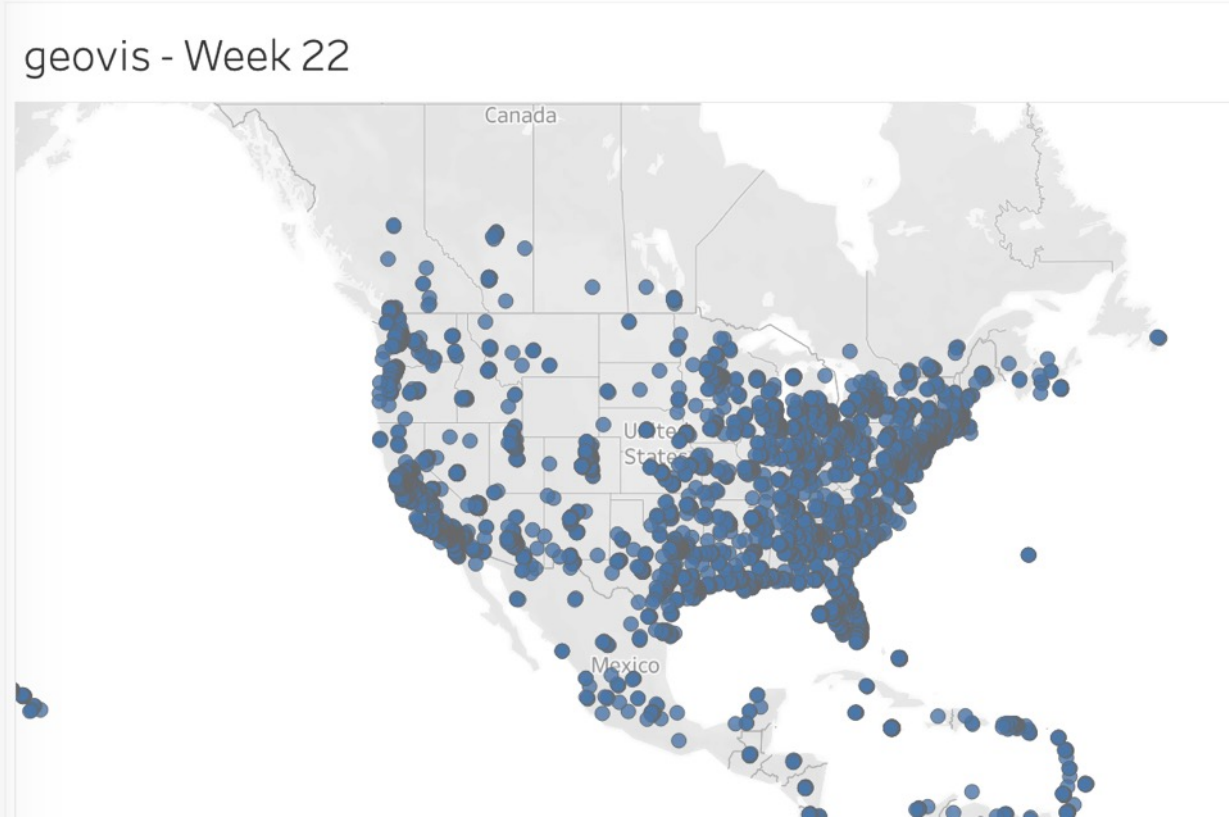
Mon 29 Jan 2018 08.51 AEDT

A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

Sensitive information about the location and staffing of military bases and spy outposts around the world has been revealed by a fitness tracking company.

Case Study: Twitter geo-metadata
BLM peaceful protest movement (2020)





Work in progress, Please do not cite, circulate or quote without permission –
credits: Ignacio Ojea Quintana; Ritsaart Reimann; Mark Alfano; Marc Cheong; Colin Klein; Emily Sullivan & Amir Fard.

Ethical discussion…

For peaceful protests (BLM protests, 2020) –
is it problematic if geotags are used by police?

- Why?
- Consider Amnesty International (2020)'s report → https://www.amnesty.org/en/latest/news/2020/08/usa-law-enforcement-violated-black-lives-matter-protesters-human-rights/

USA: Law enforcement violated Black Lives Matter protesters' human rights, documents acts of police violence and excessive Force

4 August 2020, 10:00 UTC

Amnesty International USA Recorded 125 Separate Incidents of Police Violence Against Protesters, Medics, Journalists and Legal Observers in 40 States and D.C. During May and June Protests

The Report Chronicles the Stories of More Than 50 People Affected by Police Violence as Black Lives Matter Grows Into the Largest Social Movement in U.S. History

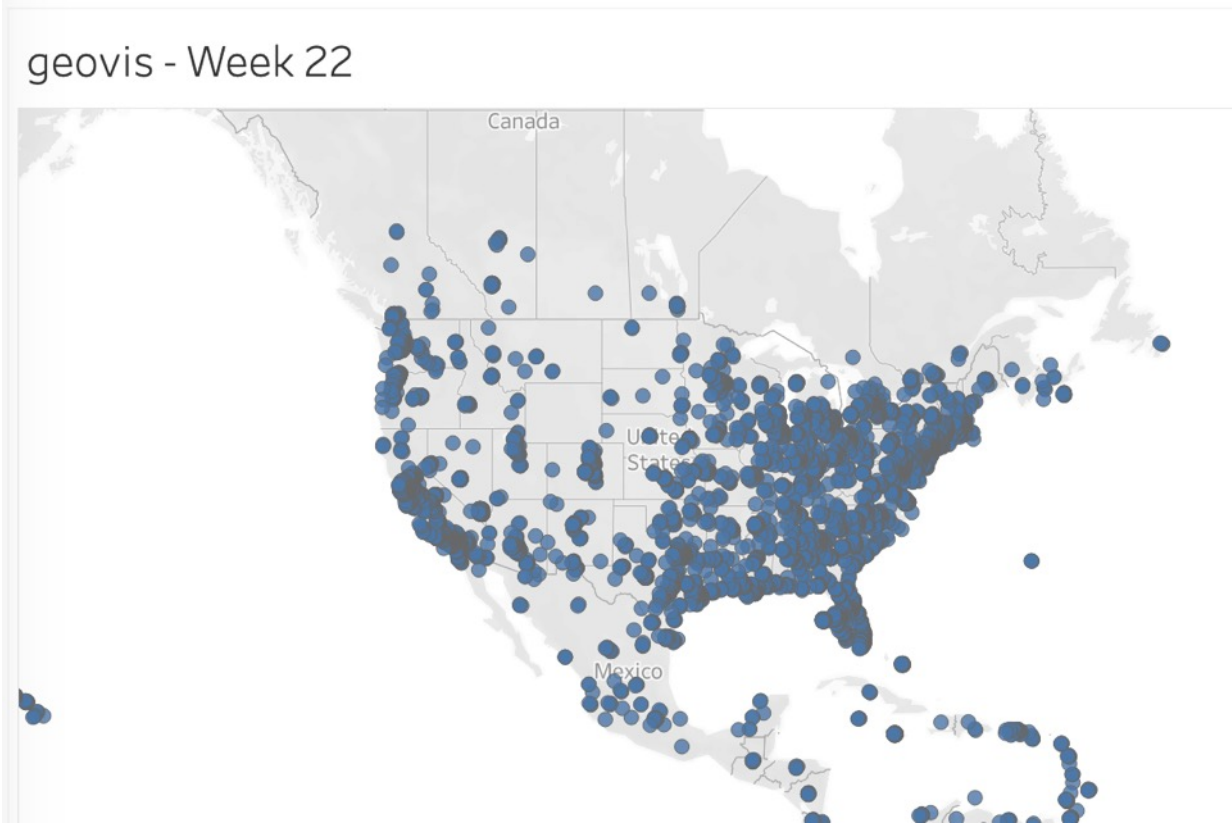
Today, Amnesty International USA released a report documenting widespread and egregious human rights violations by police officers against protesters, medics, journalists and legal observers who gathered to protest the unlawful killings of Black people by the police and to call for systemic reform in May and June of 2020. The report, *The World is Watching: Mass Violations by US Police of Black Lives Matter Protesters' Rights,* builds on Amnesty's interactive mapping of violence against protesters and new findings on the use of lethal force by the police. It is the most comprehensive human rights analysis of police violence against protesters to date.

💬 **Class discussion – as data scientists, discuss:**

What determines the ethics of use of social media geo data? (5 minutes)

Risk mitigation: data precision and risk of re-ID

Take, say, the BLM case study –
this is sensitive data!



Why?

Modern GPS's in phones → accurate geotags!

(What did we do? Added lots of random noise –
hence some data points in the ocean.)

What else can we do?

THE UNIVERSITY OF MELBOURNE

Digital ethics in NLP!

Interactive lecture.

Case Study: Twitter text for sentiment analysis

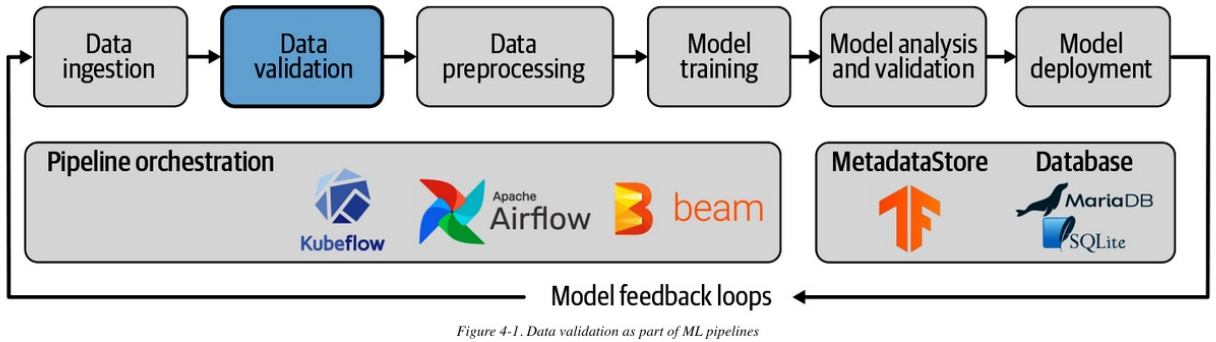


Figure 4-1. Data validation as part of ML pipelines

Typical ML pipeline, from Hapke & Nelson (2021)
https://www.oreilly.com/library/view/building-machine-learning/9781492053187/ch04.html

Case Study: Twitter text for sentiment analysis

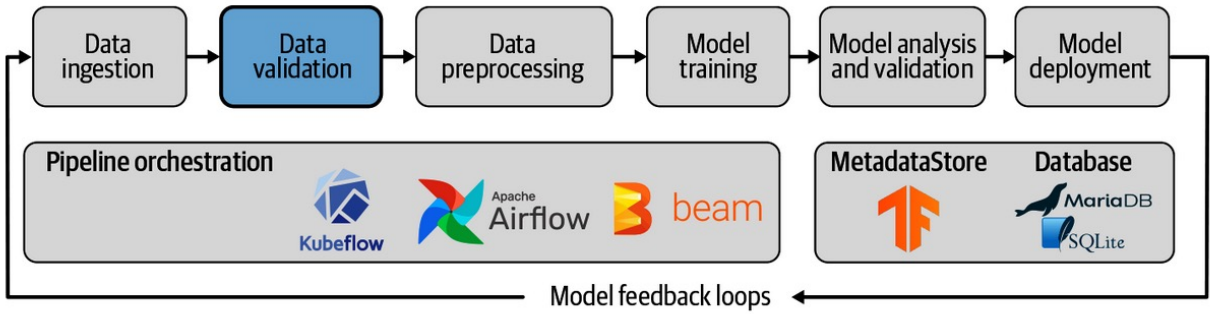


Figure 4-1. Data validation as part of ML pipelines

Data:

- Where does it come from: issues of consent, ethics, representation, harms?

Preprocessing and feature engineering:

- What assumptions do we make: issues of bias, discrimination?

Model construction:

- Language models – contemporary issues? *
- Long term effects – entrenchment?

Typical ML pipeline, from Hapke & Nelson (2021)
https://www.oreilly.com/library/view/building-machine-learning/9781492053187/ch04.html

Data.

Where does it come from: issues of consent, ethics, representation, harms?

- Issues to be aware of when collecting the data (terms of use, reproduction of data)
  - E.g. Twitter hydration policy.
  - Anecdote: In university research – Ethics Approval process.

- Harms to individuals?
  - No collection of private tweets, even as an approved follower.
  - 💬 Ethical dilemma next slide.

- Representation
  - Clearly outline assumptions and parameters of data collected.
  - 💬 Ethical dilemma next slide.

Data: Ethical discussion…

💬 **Ethical question 1:**

- You have completed your prototype classification model. (Good work!)
  As part of your report, you mention the following:
  The classifier performs lower than expected if tested on short tweets with text emoticons and common abbreviations/non-dictionary words, such as “@bob5678 hey whatsup? :)” and “@realdonaldtrump we miss u ^_^”
- **What ethical issue can we run into?**

💬 **Ethical question 2:**

- You have built a classifier which works very well in classifying movie-related sentiment on Twitter.
  For another project for an NGO/charity, you repurpose the model, but retrain it to classify e.g. public health sentiment. Assume your model is technically sound and has been peer-reviewed at the NGO.
- **What ethical issue can we run into?**

Preprocessing and feature engineering.

What assumptions do we make: issues of bias, discrimination?

- Document the assumptions adequately!
  - Good not only from a DS/ML perspective…
  - … but also, an ethical perspective!

💬 **Ethical question:**

- Assume your preprocessing pipeline consists of the removal of non-alphanumeric characters, e.g.
  `text = re.sub(r'[^a-zA-Z0-9]', '', input)`
- **What ethical issues of representation can you run into?**
- **What accuracy issues can you also run into? (Hint: 😏)**

Models.

Current issues in NLP.

- Language models – contemporary issues?
- Long term effects – entrenchment?
- (General assumptions on accuracy – the ethical discussion **on stakes:** **FN/FP** not so important for sentiment analysis; but very important for e.g. automated sanctioning/enforcement – COMPAS).

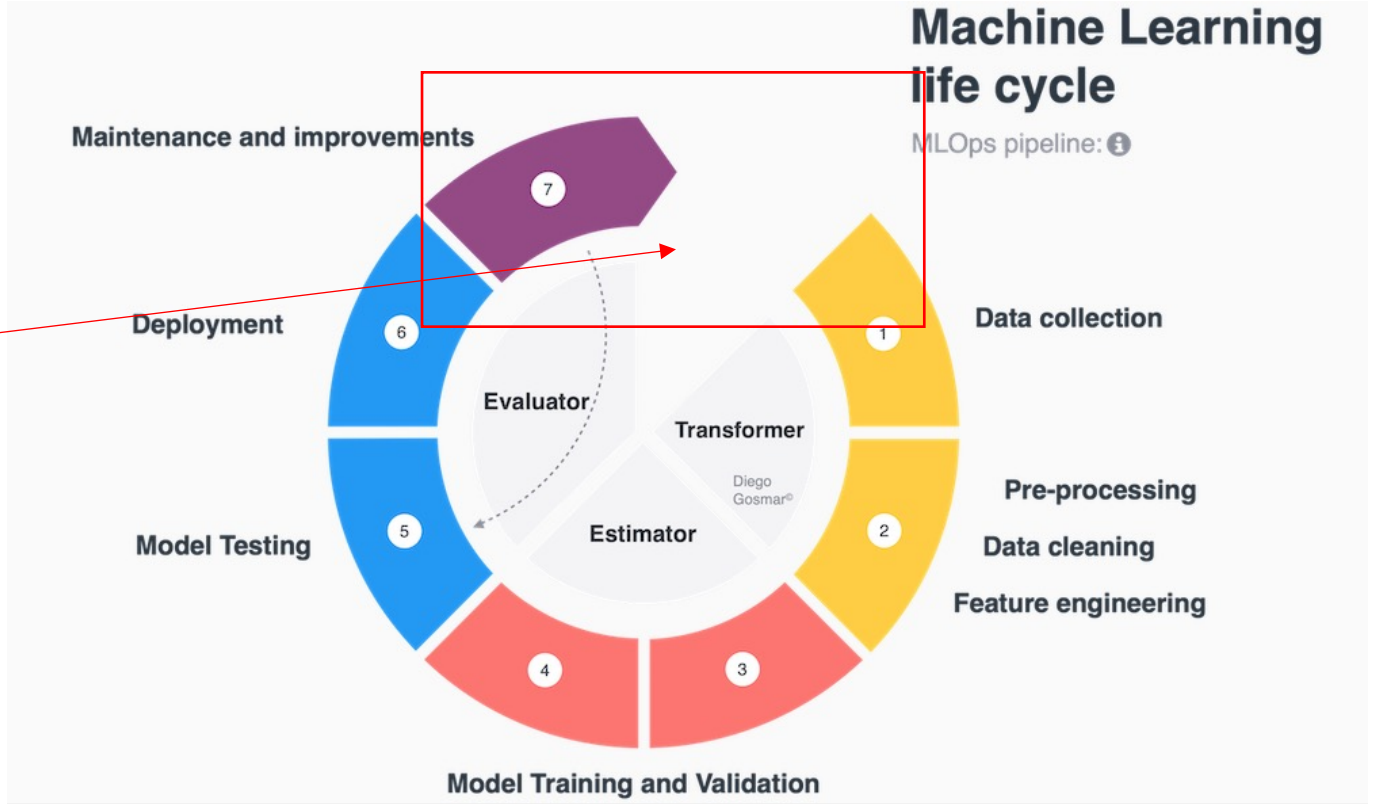


Image source: Diego Gosmar (2020)
https://www.gosmar.eu/machinelearning/2021/01/02/mlops-scalability/

Model Bias in NLP: Hutchinson et al (2020)

https://arxiv.org/abs/2005.00813

Social Biases in NLP Models as Barriers for Persons with Disabilities

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton,
Kellie Webster, Yu Zhong, Stephen Denuyl
Google
{benhutch,vinodkpg,dentone,websterk,yuzhong,sdenuyl}@google.com

Abstract

Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions

Sentence	Toxicity
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

Table 1: Example toxicity scores from Perspective API.

of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals.

Model Bias in NLP

Jane Austen

"as the daughter of an attorney Mrs. Bennet married up when she captivated the landed Mr. Bennet"

- Pride and Prejudice, as cited in http://www.diva-portal.org/smash/get/diva2:207053/FULLTEXT01.pdf

(extrapolated to 'big data'…)

Gender bias in word embeddings
(Duman, Kalai, Leiserson, Mackey, Sursesh, 2017)
http://wordbias.umiacs.umd.edu/

he (267)

guy (0.29)
heir_apparent (0.24)
maestro (0.24)
successor (0.23)
mercurial (0.22)
statesman (0.22)
genius (0.21)

she (33)

muse (0.13)
compassion (0.09)
intuition (0.09)
transformative (0.08)
philanthropy (0.08)
problem_solving (0.07)
originality (0.06)

Slide adapted from "Gender Bias: From language models to disparate impact" – credits to [CAIDE] G. Bush, S. Coghlan, K. Leins, A. Lodders, T. Miller, J. Paterson; and [CIS/Policy Lab] L. Frermann; S. Njoto; L. Ruppanner (in alphabetical order).

Models and *Entrenching* Bias

Once an ML model is deployed, past decisions (en masse) will play a role in determining future decisions.

Take, say, current biases in NLP models (say **M**).

- Reuse of models: Any ML model (say **M'**) based on a current NLP model will inherit these biases. Any further project building upon **M'** will continue inheriting these biases.
- Also beware feedback loops: if M misclassifies data point X_0 (e.g. misclassifies a particular tweet as 'not harmful' even though it is clearly harmful/discriminatory), it is likely that other data points similar to X_0 will be similarly misclassified… … and if that misclassification is left unchecked when fine-tuning the model…

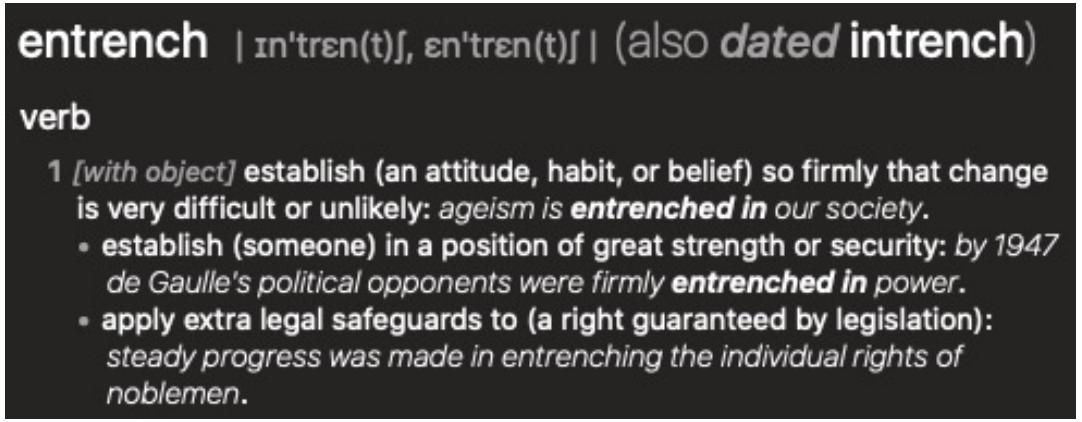


Image source: Oxford Dictionary of English, per Dictionary.app (MacOS Catalina).

Models: Ethical discussion…

💬 **Ethical question:**

- You've completed your model for the assignment. (Good work!)
  Open ended question: What should we capture in the documentation/report about the model? (We can't audit an entire model for bias, but what CAN we do?)

- **Recall Wednesday's topics** ☺️
  - **Human Bias[1], Data Bias[2], Model Bias[3], Deployment Bias[4], Evaluation Bias[5]…**
- **What can data scientists – i.e. you – do, for [1-5] and beyond?**
  - **Auditing the entire pipeline (not just the code!!!)**
  - **Including domain experts (+ legal/ethical experts) in the construction/deployment of the system.**
  - **Tools that may help: Amazon SageMaker Clarify, Azure FairLearn, LIME.**
  - **XAI to the rescue?**
  - 💬 **Open discussion**

Ethical discussion – just a start

Adapted from my talk given at the Victorian Centre for Data Insights (VIC Govt), May 2021.

How do we put academic ideas of digital ethics into practice

- *As a start: learning basic key tenets (cf Coghlan, Miller, Paterson, 2020) https://arxiv.org/pdf/2011.07647.pdf*
- ***"fairness, non-maleficence, transparency, privacy, respect for autonomy, liberty, and trust."***

THE UNIVERSITY OF MELBOURNE

Thank you

Any questions?