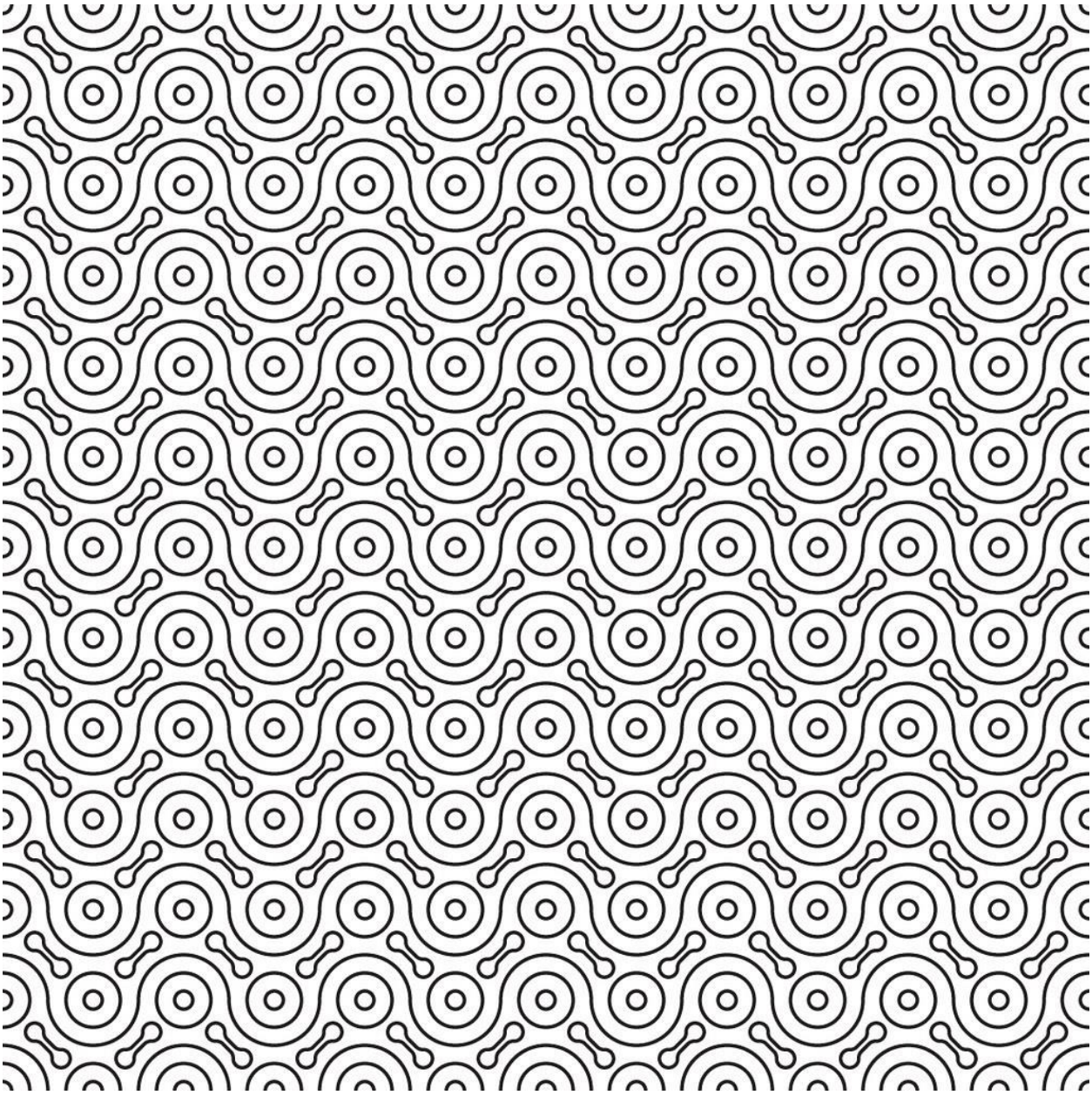


Understanding and Managing Bias in Machine Learning (Beta)

A Carpentries Beta Episode presented by Nora McGregor, Digital Curator as part of AI4LAM applying and deploying AI in Libraries, Galleries, Archives, and Museums online workshops
30 March 2021

Today we'll be considering the following questions:

- What are common types of bias and their effect in machine learning?
- At what points can bias enter the machine learning pipeline?
- Can we manage bias? Some lessons from and for the GLAM sector

At the end of this session, you'll be able to:

- Understand bias in the context of machine learning
- Identify common types of bias and how and at what stages these may impact model predictions
- Consider a range of bias mitigation strategies available to GLAM staff

Bias in Machine Learning

Though AI and machine learning (ML) systems **may appear** to us as **objective**, dealing solely in facts and numbers, **free from troublesome human proclivities** in their decision making, there are abundant opportunities for **human bias** to enter ML systems at all stages of the pipeline.

Human bias can have a **broad and complex range of effects** on the classification and predictions of models, with consequences of varying degrees. Bias in AI can be understood in most general terms as an **error where incorrect assumptions** lead to **systematically prejudiced results**.

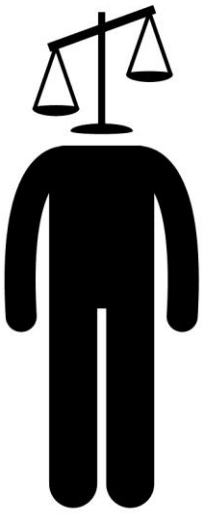
When we typically hear about bias in AI these days, particularly in the news, it is most often presented and understood in the context of **societal prejudice** and discrimination, where human prejudice in the development and application of algorithms results in the perpetuation of unfairness, inequities, and stereotypes of the real world (as has happened with predictive policing algorithms).

But bias may exist in other forms, such as an image dataset that inadvertently contains objects that always happen to appear in the centre of the image, making it hard for a classifier to recognise objects that are not in the centre of images.

Bias may even be **introduced** to an algorithm in order to **correct** an unfair model. Consider an image search algorithm, based on real world data where a majority of men are CEOs. A user searching for "CEO" will find images of primarily men, and a good bias may be introduced in order to insure proper representation of a diversity of CEOs.

In whatever form it takes, it is crucial for model builders to be able to identify bias and manage it.

Let's take a look at that canonical example of algorithmic bias...

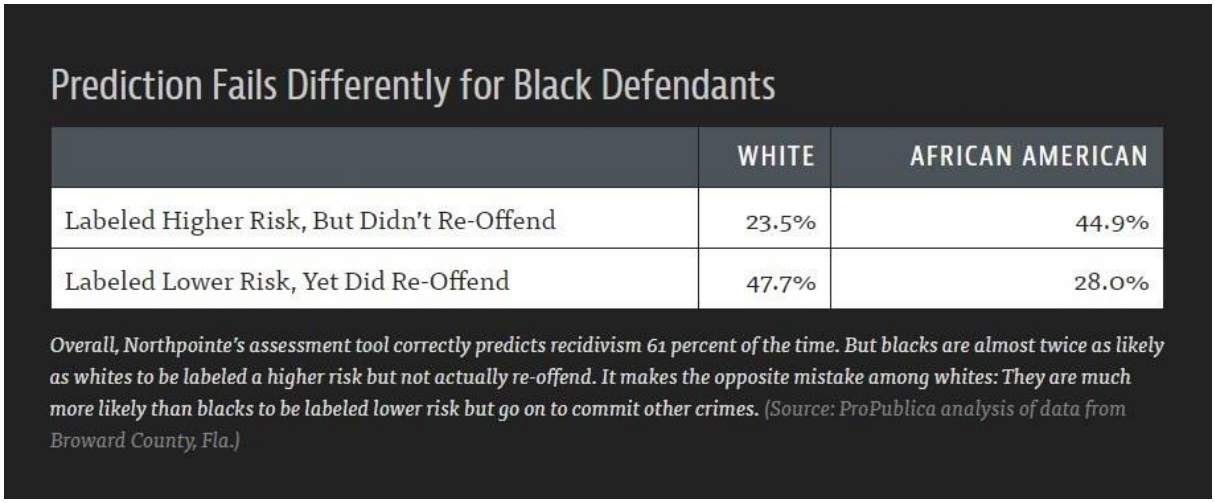


Predictive Policing Algorithms

Machine learning systems are being relied upon as **tools for courts** to use in sentencing, aiming to **predict** the likelihood of defendants committing a future crime.

Analysis by ProPublica in 2016 of one system used widely throughout the United States, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool, uncovered **significant racial disparity** between the system predictions for white and black defendants.

The COMPAS tool assigns scores from 1 to 10 to a defendant based on over 100 or so factors such as age, sex and criminal history although **notably race has been excluded**.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Common Bias Types in Machine Learning

Let's take a closer look at some specific and **common types of bias** that may manifest in the undertaking of machine learning approaches at your institution.

This is of course **only a small handful of potential sources** of bias that may affect our judgment and skew a model's predictions. It's important for model builders to be vigilant about finding and remedying bias in whatever form it may enter the machine learning pipeline.

Type	Definition	Example
Prejudice bias	Arises when data incorporates cultural, race, gender or stereotypes it should be ignorant of	A model is designed to differentiate between men and women in a museum's photograph collection The training data contains more pictures of women in kitchens than men in kitchens, or more pictures of men coding than women, then the algorithm is trained to make incorrect inferences about the gender of people engaged in those activities due to prejudices that occur in the real world, represented in the data.
Selection bias	Introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed	A model is trained to predict future sales of a new product line for the museum gift shop. To build the training set, the first 200 subscribers to the museum's newsletter were offered a small gift voucher to fill in a survey. Instead of randomly targeting consumers, the dataset targets newsletter subscribers who don't necessarily represent the museum's potential paying customers. It's entirely possible the newsletter subscribers population may be more inclined to be signed up to learn about free events and giveaways, while typical consumers may not be enticed by small gift vouchers or even signed up at all.
Confirmation bias	In the process of refining and reinforcing a models learning, unconsciously or consciously processing data in ways that confirm preexisting beliefs and hypotheses.	An engineer is building a model that predicts aggressiveness in dogs based on a variety of features (height, weight, breed, environment). The engineer had an unpleasant encounter with a hyperactive toy poodle as a child, and ever since has associated the breed with aggression. When the trained model predicted most toy poodles to be relatively docile, the engineer retrained the model several more times until it produced a result showing smaller poodles to be more violent.
Correlation bias	Correlation is not causation. Correlation implies the mutual relation, covariation, or association between two or more variables. It only questions whether the variable varies together or not.	The height of the father and his children is correlated, but one can't say that the father's height is caused by determining his children's height on the only assumption of the hereditary factor. There are several other factors present such as environment, genetics, etc.
Exclusion bias	Removing data from a set that we think isn't relevant	For example, imagine you have a dataset of customer sales in America and Canada. 98% of the customers are from America, so you choose to delete the location data thinking it is irrelevant. However, this means your model will not pick up on the fact that your Canadian customers spend two times more.

Consider this riddle: A father and son get in a car crash and are rushed to the hospital. The father dies. The boy is taken to the operating room and the surgeon says, “I can’t operate on this boy, because he’s my son.” **How is this possible?**

The surgeon is a **woman**. In research conducted on 197 BU psychology students (where women outnumbered men two-to-one) and 103 children, ages 7 to 17, only a small minority of subjects—15 percent of the children and 14 percent of the BU students—came up with this answer. Of self-described feminists in the student group, **a majority**, 78 percent, **did not guess the surgeon was the mother.**

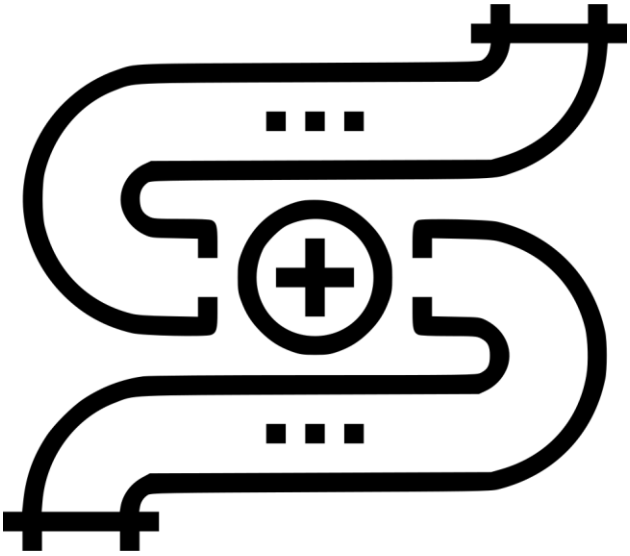
This illustrates how a training dataset may **unintentionally** come to encode **societal gender bias** through human applied labels such as “Surgeon” vs. “Female Surgeon”.

“Although neural networks might be said to write their own programs, they do so towards **goals set by humans**, using data collected for **human purposes**. If the data is skewed, even by accident, the computers will amplify injustice.” — The Guardian

When might human bias enter a machine learning pipeline?

There are abundant opportunities for human bias to enter ML systems at all stages of the pipeline including:

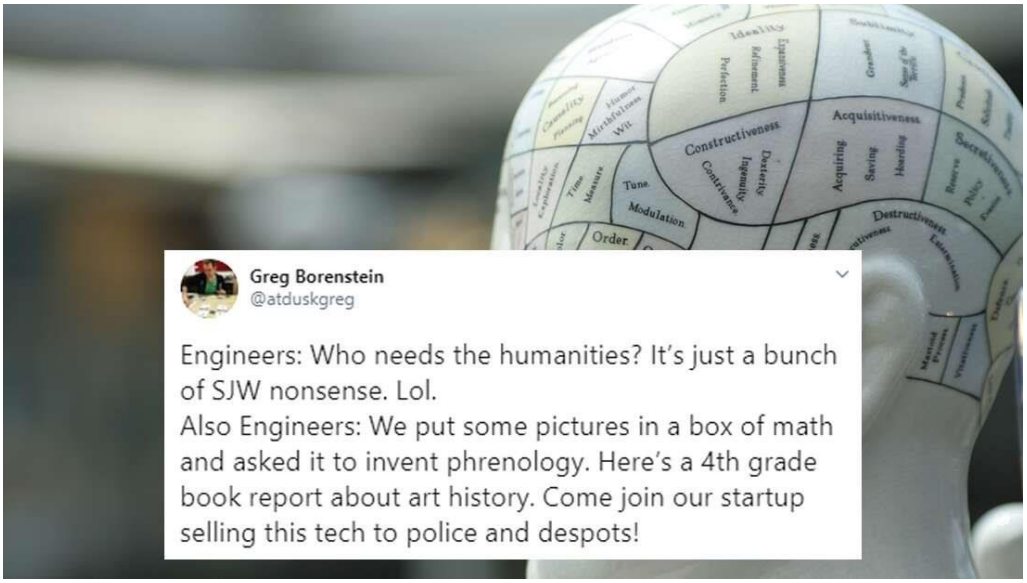
- When the study is being **designed**
- When **datasets** are **constructed**
- When **decisions** are being made to **refine** and **reinforce** a models learning
- When interpreting and **applying** decisions made by the model to **real world scenarios**



Bias arising in the study design

Some machine learning systems are quite simply **built on ethically unsound foundations** from the outset.

A recent controversial study, *Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings*, published in Nature Communications, garnered significant controversy for its proximity to the thoroughly debunked pseudoscience phrenology which aims to assess an individuals personality and (or in the case of this study, trust) based on facial structure.



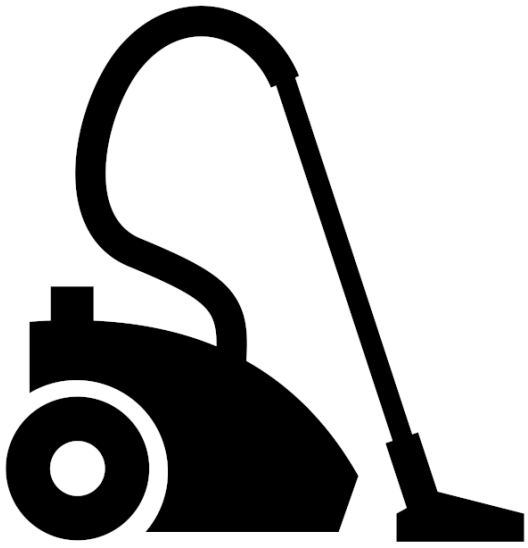
Bias arising in dataset collection and construction

Data is never neutral. It is about people. And the manner in which data is **sourced** and **constructed** for training sets will have important **implications** on your model's output.

Because of the many complexities around copyright and licensing, privacy issues and high costs involved in getting access to large quantities of quality datasets, data scientists have favoured **scraping** what they can find **freely**, **en masse** and **indiscriminately**, across the internet from open sources such as Wikipedia, FlickR or Google News.

Datasets collected in this brute force manner, will naturally **reflect the biases of these sources**, based on their demographic and geographic composition (overrepresentation of young internet users from developed countries for example). When data is **incomplete**, or **skewed** in any way either **intentionally** or **unintentionally** bias will arise.

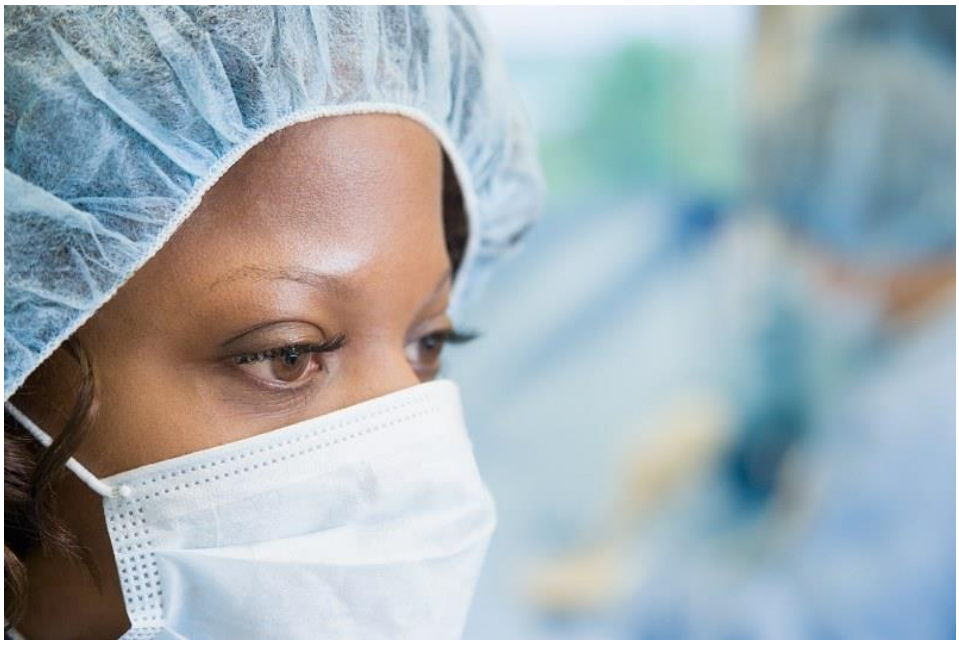
"It's all too easy to forget that **data is about human beings** and their behaviors. Data is not an abstraction…Data encodes the stories of our lives, capturing not only our tastes and interests but also our hopes and fears. Data isn't an abstract idea or a set of numbers or qualitative responses. It can be and is, ultimately, human."



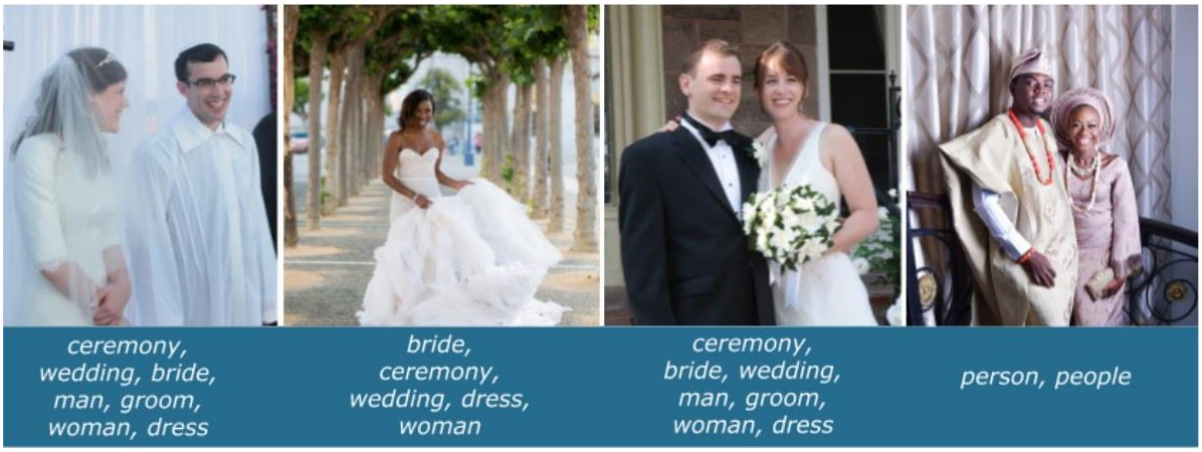
When training data is **labelled (annotated)**, typically through some kind of **crowdsourcing** mechanism such as Amazon's Mechanical Turk, and as we saw in the first activity, bias can crop up, either consciously or unconsciously, in the course of this.

GLAM staff will be more than **familiar** with this phenomenon as we grapple with **historical bias in library and archives descriptions** for instance. As demand for GLAM collections and catalogue data for use in machine learning increases, it is **vital** that model builders are **made aware** of the **biases** that may be **encoded within cultural heritage data**.

Whether personally undertaking or crowdsourcing your **data annotation**, it's important to **be aware** of how different **demographics** and **social constructs** may introduce bias and **implicit associations** into your pipeline.



__Consider__ these four images and __write a list of terms__ you would use to annotate each. __Compare__ your outputs with your nearest neighbour(s). __Discuss__ the __differences__ and how this could potentially __affect__ a model. How might you __mitigate__ these differences in annotations?



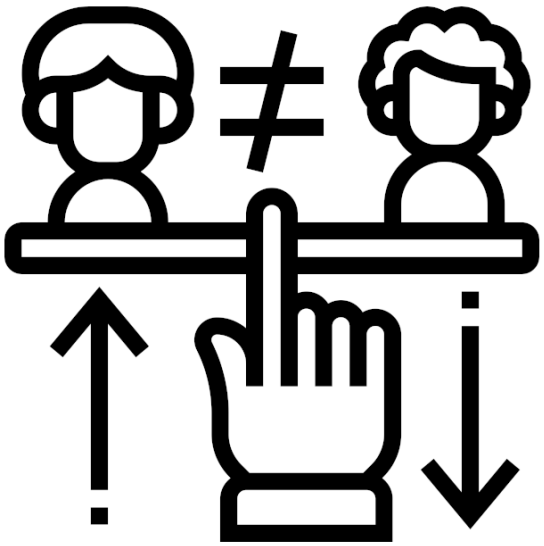
Wedding photographs (donated by Googlers), ***labeled by a classifier*** *trained on the Open Images dataset. The classifier's label predictions are recorded below each image.*
How did you compare?

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

---

Bias arising when refining and reinforcing a models learning

Machine learning models are **refined** and **reinforced** based on **reactions to its results**. In this process, there is a risk of certain outcomes being ignored and others privileged over others, skewing a models learning.



Example:

A model builder is using named entity recognition across multilingual newspapers. They might determine they are satisfied when the model gets to 90% accuracy and will aim to improve to this result. However this overall accuracy can hide the fact that some particular 'slices' of our data might have much worse accuracy. Your overall accuracy might be very good but your model may underperform on one language. This might not be addressed by changing your data but changing how you approach training and evaluating your model.

Bias arising in the application of machine learning decisions to real world scenarios

Algorithmic bias is defined as **unjust**, **unfair**, or **prejudicial treatment** of people related to race, income, sexual orientation, religion, gender, and other **characteristics historically associated with discrimination** and marginalization, when and where they manifest in algorithmic systems and algorithmically aided decision-making.

1 - Amazon scraps secret AI recruiting tool that showed bias against women

How can GLAM staff help manage human bias in machine learning approaches?

The presence of human bias in the classifications and predictions of machine learning is a **key challenge** today, but being **aware of** and **transparent about** the problem allows us to take **proactive** steps to **mitigate** their effects.

For GLAM professionals, this should be **relatively familiar ground** and in *Reference Lessons from archives: strategies for collecting sociocultural data in machine learning* the authors argue that the **document collection practices in archives** present the ideal **ethical** and **practical framework** for mitigating bias in data collection for the field of machine learning.

GLAM professionals have an opportunity to **apply** these **professional tools** to the management of bias.

"Archives are the longest standing communal effort to gather human information and archive scholars have already developed the language and procedures to address and discuss many challenges pertaining to data collection such as consent, power, inclusivity, transparency, and ethics & privacy."

Responsible Operations: Data Science, Machine Learning, and AI in Libraries outlines several great recommendations and is a great place to start!

- GLAM institutions need to ensure a diverse workforce-monoculture cannot effectively manage bias and diversity is not an option, it is an imperative (Collections as Data)
- Hold symposia focused on surfacing historic and contemporary approaches to managing bias with an explicit social and technical focusing on the challenges libraries faced in managing bias while adopting technologies like computation, the internet, and currently with data science, machine learning, and AI.
- Contribute diverse language materials, collections and texts to sources where model builders are finding their data such as Wikipedia
- Collaborate directly with computer scientists to build new diverse, curated data sets for use in machine learning (Arabic HTR)
- Enlist the help of staff with the right domain expertise to review training data construction before and after, they may see biases that you have overlooked
- When collecting and annotating data make sure to recruit diversified crowds for the task and carefully communicate instructions
- Employ toolkits for detecting and removing bias from machine learning models, for instance the IBM open sourced AI fairness 360 tool
- Consider your partnerships and collaborators closely, including ramifications of outsourcing your AI to external companies/partners
- Know your data and be transparent about it by creating a datasheet describing and documenting it in full.

In small groups, consider the following potential machine learning project. Discuss 2-3 points at which bias may enter the pipeline, and questions/strategies GLAM staff might want to consider in order to manage it:

An art museum is keen to make a newly acquired digitised collection of 20,000 Southeast Asian photographs more discoverable within the main museum image search. Aside from a collection level description noting the provenance of the collection from a 19th century English explorer, the individual images have very little in the way of item level description except for some captions ascribed by the collector. A model will be trained to classify these images. In order to build the training data set, the art museum is considering setting up a public crowdsourcing project is set up asking members of the public to tag a set of images with as many descriptive words as they can.

Resources Consulted & Recommended Reading

Barlow, R. (2014). *BU Research: A Riddle Reveals Depth of Gender Bias | BU Today | Boston University*. Boston University. Retrieved 29 March 2021, from https://www.bu.edu/articles/2014/bu-research-riddle-reveals-the-depth-of-gender-bias.

Catanzaro, B. (2019, December 4). "Datasets make algorithms: how creating, curating, and distributing data creates modern AI." [Video file]. Retrieved from https://library.stanford.edu/projects/fantastic-futures.

Coleman, C. (2020). Managing Bias When Library Collections Become Data. *International Journal Of Librarianship*, *5*(1), 8-19. https://doi.org/10.23974/ijol.2020.vol5.1.162.

Ekowo, M. (2016). *Why Numbers can be Neutral but Data Can't*. New America. Retrieved 29 March 2021, from https://www.newamerica.org/education-policy/edcentral/numbers-can-neutral-data-cant/.

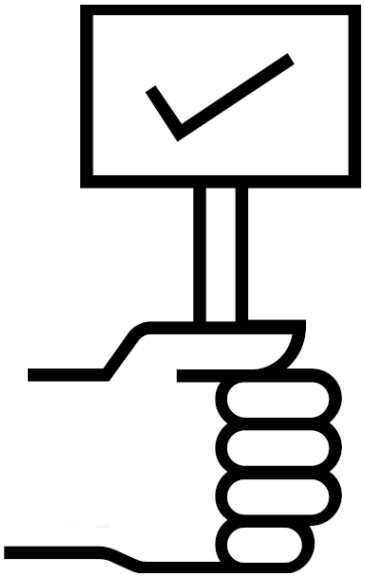
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumeé III, H., & Crawford, K. (2020). *Datasheets for Datasets*. arXiv.org. Retrieved 29 March 2021, from https://arxiv.org/abs/1803.09010v3 .

Jo, E., & Gebru, T. (2020). Lessons from archives. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. https://doi.org/10.1145/3351095.3372829.

Mayson, Sandra Gabriel, Bias In, Bias Out (2019). 128 Yale Law Journal 2218, University of Georgia School of Law Legal Studies Research Paper No. 2018-35, Available at SSRN: https://ssrn.com/abstract=3257004.

Padilla, T. (2019). Responsible Operations: Data Science, Machine Learning, and AI in Libraries. OCLC Research Position Paper. https://doi.org/10.25333/xk7z-9g97.

Barbosa, N., & Chen, M. (2021). *Rehumanized Crowdsourcing | Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Dl.acm.org. Retrieved 29 March 2021, from https://dl.acm.org/doi/10.1145/3290605.3300773.



Thank you!

To revisit this lesson please visit the (beta) episode at:

https://carpentries-incubator.github.io/machine-learning-librarians-archivists/index.html

Questions? Feedback?

digitalresearch@bl.uk

Icons and stock images used in this presentation are CC and sourced from The Noun Project.