# Common Ethical Challenges (ii)

for Data Practitioners and Users

** based on Introduction to Data Ethics module (Part 2) developed by Shannon Vallor, Ph.D.*

# Summary of Previous Lecture

**Ethical Challenges in Appropriate Data Collection and Use**

- Purpose of data collection, context, dissemination of data, choice in data sharing, compensation, control/rights...

**Data Storage, Security and Responsible Data Stewardship**

- Storage of data, risk estimation, mitigation strategies, privacy-preserving techniques, ethical risks of keeping data longer...

# 3. DATA HYGIENE AND DATA RELEVANCE

How dirty (inaccurate, inconsistent, incomplete, or unreliable) is our data, and how do we know?

- Big data era is dangerous since researchers focus less on clean data.
- Dirty data means dirty models...
- Dirty models can do significant harms (e.g., criminal justice, hiring, finance etc.).

# Only 3% of Companies' Data Meets Basic Quality Standards


https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards *(published in 2017 by Harvard Business Review)*

# What are our established tools and practices for scrubbing dirty data?

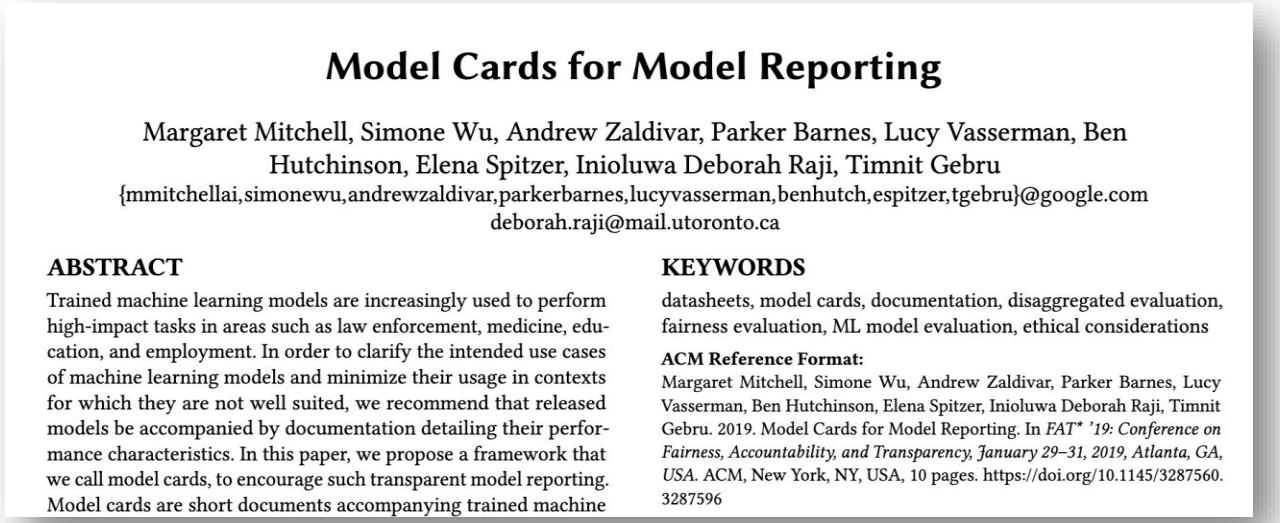
- A set of policies should be followed for data cleaning.
- This ensures data integrity, especially when data is coming from various datasets.

# Datasheets for datasets*

- Bridging the gap between dataset creators and data consumers
- Good for:
  - Reproducibility
  - Increasing accountability and transparency
  - Mitigating unwanted biases
  - Deciding on the use of a dataset
- Similar datasets could be created based on datasheets


*Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (December 2021)

# Democratization of ML --- Model Cards



**Model Cards for Model Reporting**

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

**ABSTRACT**

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine

**KEYWORDS**

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

**ACM Reference Format:**

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

**MMitchell**
@mmitchell_ai

Okay, if I were to make an internship @huggingface almost solely focused on filling out Model Cards, would anyone actually be interested?

8:45 PM · Jan 25, 2022 · Twitter Web App

**28** Retweets **3** Quote Tweets **233** Likes

Tweet your reply

Reply

**MMitchell** @mmitchell_ai · Jan 25
Replying to @mmitchell_ai
Ideally the project would also involve identifying patterns in how they're used/filled out, pain points, & specifying what writing could be used to auto-fill the card based on some specific questions/model types.

3 1 38

**MMitchell** @mmitchell_ai · Jan 25
Probably reasonable for someone just outside of undergrad or @ a Master's level, who's good at writing documentation and "translating" between what a sole-ML-engineer would say to what, e.g., a regulator would benefit from reading.

11 1 37

https://modelcards.withgoogle.com/about

What are our practices and procedures for validation and auditing of data in our context?

- We should have means to check if an organization's constraints are satisfied.
  - Manual processes
  - Automated processes
- We will cover external/internal auditing later in the class.

How do we establish proper parsing and consistency of data field labels?

- Data may be gathered from **different sources**.
- Ensuring the **integrity** of our data is critical.

Have we considered the diversity of the data sources and/or training datasets?

- Does the dataset reflect the population we focus on? Is it representative enough?
- If not, a new dataset may be used, or it needs to be collected.

# Is our data appropriately relevant to the problem?

- This question is critical when we are the data consumers.
- An adoption of datasheets (or similar) is needed for data consumers to make informed decisions.
- As data collectors, we may sometimes forget about the research questions; and instead, we may focus on collecting the maximum data and missing relevant ones.

How long is this data likely to remain accurate, useful or relevant?

- For some research questions, time aspect is critical.
- Keeping historical or non-updated data may lead to unethical decisions.

# Use of historic criminal data

- Goal: predicting criminal behavior to allocate resources accordingly, whether for rehabilitation or for prison sentences in the US.
- AI gives a recidivism score to people. Judges consider such scores in making decisions for defendants.

**AI is sending people to jail—and getting it wrong**

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

By Karen Hao

January 21, 2019

Low-income and minority communities have **historically** been disproportionately targeted by law enforcement

# 4. IDENTIFYING AND ADDRESSING ETHICALLY HARMFUL DATA BIAS

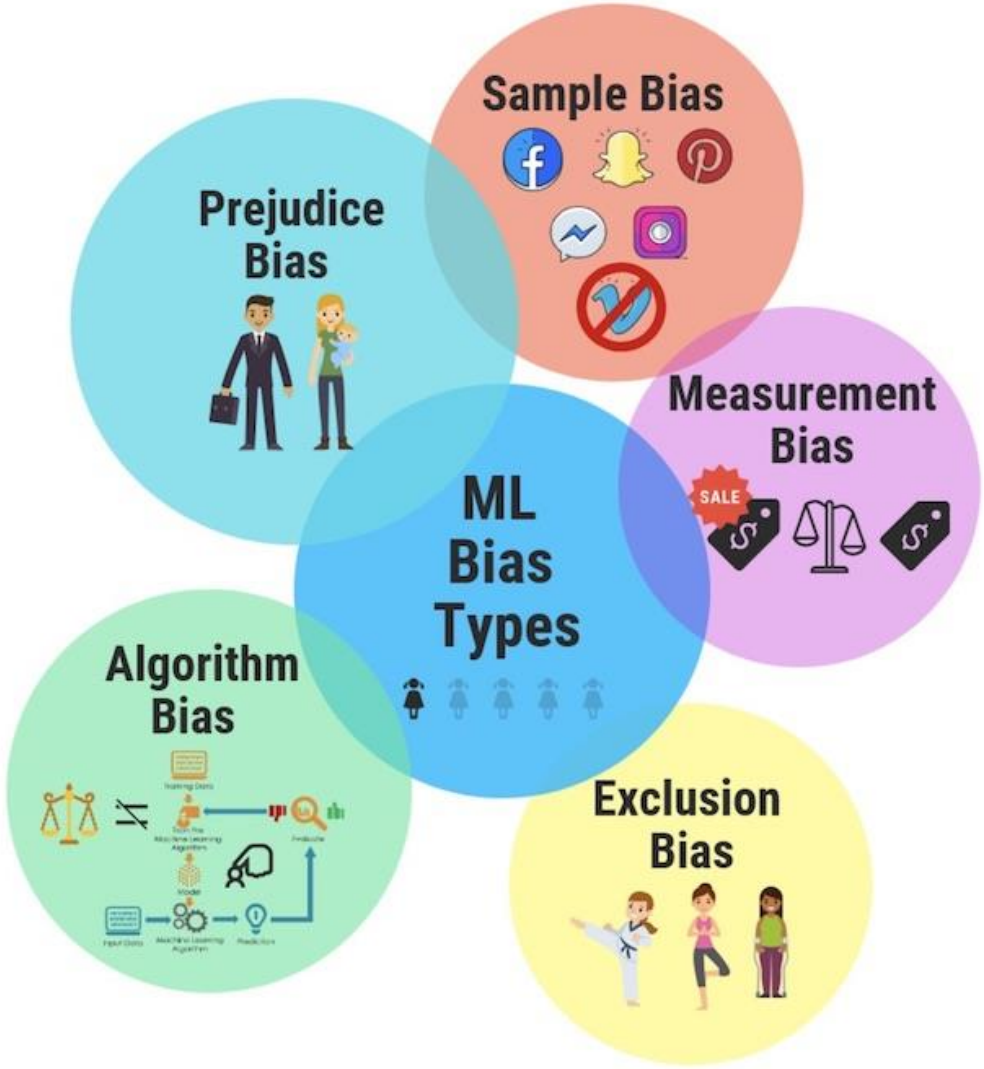
What inaccurate, unjustified, or otherwise harmful human biases are reflected in our data?

How might harmful or unwarranted bias in our data get magnified, transmitted, obscured, or perpetuated by our use of it?

- Biases can be explicit or implicit.
- Identification of harms is important, but the impact on the stakeholders should be thought about carefully.
- The wide use of non-documented datasets could perpetuate harmful/unwarranted bias.

# ML/AI Bias

- Sample Bias: Training data contains partial/incorrect information
- Prejudice Bias: Occurs mostly because of belonging to a social group (unconscious bias)
- Exclusion Bias: Inclusive policies should be adopted by organizations
- Algorithm Bias: Human error again...
- Measurement Bias: Results from poorly measuring the outcome (e.g., surveys)




https://predera.com/reimagining-ai-building-togetherness-with-bias-monitoring/

# 5. VALIDATION AND TESTING OF DATA MODELS & ANALYTICS

How can we ensure that we have adequately tested our analytics/data models to validate their performance?

Have we fully considered the ethical harms that may be caused by inadequate validation and testing?

- The datasets that we train/validate/test our models is critical. They should be representative enough for the focused population.
- Validation/testing may be rushed because of internal/external pressures. This may lead to more ethical harms.

# How can we test our data analytics and models to ensure their reliability across new, unexpected contexts?

What if the same system does harm to various stakeholders in another context?


In what cases might we be ethically obligated to ensure that the results, applications, or other consequences of our analytics are audited for disparate and unjust outcomes?

- For example, in the privacy context, the users are protected under GDPR. They have rights to submit complaints, and organizations must respond in a specific time-frame.
- Frequent audits can identify such problems earlier and prevent damage before it happens.

# 6. HUMAN ACCOUNTABILITY IN DATA PRACTICES AND SYSTEMS

Who will be designated as responsible for each aspect of ethical data practice?

Who should and will be held accountable for various harms that might be caused by our data or data practice?

- A team leader is responsible of the actions of their team members.
- Team members, who are doing or not doing a task, are accountable and they are the ones who should provide answers.
- Accountability is related to answerability.

THE UNIVERSITY of EDINBURGH

# Informatics researchers to help make autonomous systems more responsible

Nadin Kokciyan and Michael Rovatsos will be working on one of the strands of a multi-disciplinary project that seeks to address responsibility gaps in autonomous systems. Informatics researchers will focus on the development of new techniques and tools for making autonomous systems more answerable. The project is led by Professor Shannon Vallor, Director of the Centre for Technomoral Futures at the Edinburgh Futures Institute.

Drawing on research in philosophy, cognitive science, law and AI, the project will develop new ways for autonomous system developers, users and regulators to bridge responsibility gaps by boosting the systems' answerability.

## Making machines better at giving answers

We are surrounded by autonomous systems, be it self-driving cars, virtual assistants or plane autopilots, and increasingly, these solutions are also making their way to high-stakes areas such as health and finance. This creates a vital need to ensure we can trust these systems. A key to maintaining social trust is the ability to hold others responsible for their actions and it is no different for autonomous systems.

https://www.ed.ac.uk/informatics/news-events/stories/2021/answerable-autonomous-systems-nadin-kokciyan

Do we have a clear and effective process for any harmful outcomes of our data practice to be surfaced and investigated?

What processes should we have in place to allow an affected party to appeal the result or challenge the use of a data practice?

PIS Example

**What are my data protection rights?**

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

**Who can I contact?**

If you have any further questions about the study, please contact the lead researcher, [name redacted].

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

https://web.inf.ed.ac.uk/infweb/research/ethics-and-integrity/ethics-resources

To what extent should our data systems and practices be open for public inspection and comment?

- We do not observe this behavior much in current online systems.
- Some examples:
  - How does Facebook decide which posts to display in a user's feed?
  - How does Twitter search works and what does their ranking mean?

# AI Ethics in Practice -- AlgorithmWatch


- AlgorithmWatch is a non-profit research and advocacy organization.
- They analyze automated decision-making systems to measure their impact on society.
- AlgorithmWatch maintains AI Ethics Guidelines Global Inventory that includes 173 guidelines (April 2020).
- They have many projects to investigate how algorithms work in practice.
- They are not liked by big tech companies.

# AI Ethics in Practice -- AJL


- The Algorithmic Justice League is an organization that combines art, research, policy guidance and media advocacy to illuminate the social implications and harms of AI.
- AJL is a cultural movement towards
  - Equitable AI (agency and control, affirmative consent, centering justice)
  - Accountable AI (transparency, continuous oversight, redress harms)
- AJL recognizes the limitations of Ethical AI, which does not create any mandatory requirements or ban certain uses of AI. They focus on creating action.

# 7.EFFECTIVE CUSTOMER/USER TRAINING IN USE OF DATA AND ANALYTICS

Have we placed data tools in appropriately skilled and responsible hands, with appropriate levels of instruction and training?

- Do organizations invest in appropriate tools to monitor how the data is used?
- Training staff is essential, since a user cannot do much without proper training.

Are our data customers/users given an accurate view of the limits and proper use of the data, data practice or system we offer, not just its potential power?

- The data collected should be used for the purposes specified.
- Selling inappropriate technology to third-parties may cause ethical harms to many people (without their consent).
- Do organizations check what power their employees have?

# 8.UNDERSTANDING PERSONAL, SOCIAL, AND BUSINESS IMPACTS OF DATA PRACTICE

Have we fully considered how our data/data practice or system will be used, and how it might impact data subjects or other parties later on?

Has sufficient input been gathered from other stakeholders?

Has the testing of the practice taken into account how its impact might vary?

- Diversity in teams is something desirable. It is difficult to be aware of ethical harms that a system could do to groups unlike ourselves.
- Focus groups are great to consult stakeholders of a system during the design phase.
- The final product should be also tested by their stakeholders.

# AI Ethics in Practice – Diverse Voices


https://techpolicylab.uw.edu/project/diverse-voices/

Does the collection or use of this data violate anyone's legal or moral rights, limit their fundamental human capabilities, or otherwise damage their fundamental life interests?

Would information about this data practice be morally or socially controversial or damaging to professional reputation of those involved?

- Personal and social impacts of a data practice should be considered prior to the production stage.
- Organizations may take decisions not aligned with ethical values of their employees.

# Google workers reject company's account of AI researcher's exit as anger grows

**Timnit Gebru's colleagues challenge claims she resigned while more than 1,800 sign petition of solidarity**

Gabrielle Canon

@GabrielleCanon

Mon 7 Dec 2020 22.14 GMT

Timnit Gebru said she was fired from Google after sending an email to an internal group for women. Photograph: Kimberly White/Getty Images for TechCrunch

Gebru, a Black female scientist who is highly respected in her field, said on Twitter last week that she had been fired after sending an email to an internal company group for women and allies, expressing frustration over discrimination at Google and a dispute over one of her papers that was retracted after initially being approved for publication.

https://www.theguardian.com/technology/2020/dec/07/timnit-gebru-google-firing-resignation-ai-research

# Next week on CSAI

Monday – 1st Case Study Week (bring laptops!)

Monday – Release of CW1 (finalize your groups!)