Understanding and managing bias in the data science lifecycle

Learning Objectives

- Understand and recognize the steps of the data science life cycle
- Be mindful of bias and ethical considerations throughout the cycle
- Know some of the most common options for putting an ML model into production

Ethics in Data Science

- From analysis of 115 university technology ethics courses
- Wide variation in topic coverage, but consensus on:
  - Ability to critique;
  - Spot issues; and
  - Improve communication

What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis

Casey Fiesler
casey.fiesler@colorado.edu
University of Colorado Boulder
Boulder, CO

Natalie Garrett
natalie.garrett@colorado.edu
University of Colorado Boulder
Boulder, CO

Nathan Beard
nbeard@umd.edu
University of Maryland
College Park, MD

ABSTRACT

As issues of technology ethics become more pervasive in the media and public discussions, there is increasing interest in what role ethics should play in computing education. Not only are there more standalone ethics classes being offered at universities, but calls for greater integration of ethics across computer science curriculum mean that a growing number of CS instructors may be including ethics as part of their courses. To both describe current trends in computing ethics coursework and to provide guidance for further ethics inclusion in computing, we present an in-depth qualitative analysis of 115 syllabi from university technology ethics courses. Our analysis contributes a snapshot of the content and goals of tech ethics classes, and recommendations for how these might be integrated across a computing curriculum.

CCS CONCEPTS

Table 2: The number of courses that had content for each listed topic, out of 115 total courses, organized from most popular to least popular topic.

Topic	Courses
Law & policy	66
Privacy & surveillance	61
Philosophy	61
Inequality, justice & human rights	59
AI & algorithms	55
Social & environmental impact	50
Civic responsibility & misinformation	32
AI & robots	27
Business & economics	27
Professional ethics	25
Work & labor	23
Design	20
Cybersecurity	19
Research ethics	16
Medical/health	12

Ref: Casey Fiesler, Natalie Garrett, and Nathan Beard. 2020. What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. https://doi.org/10.1145/3328778.3366825

Ethics in Data Science

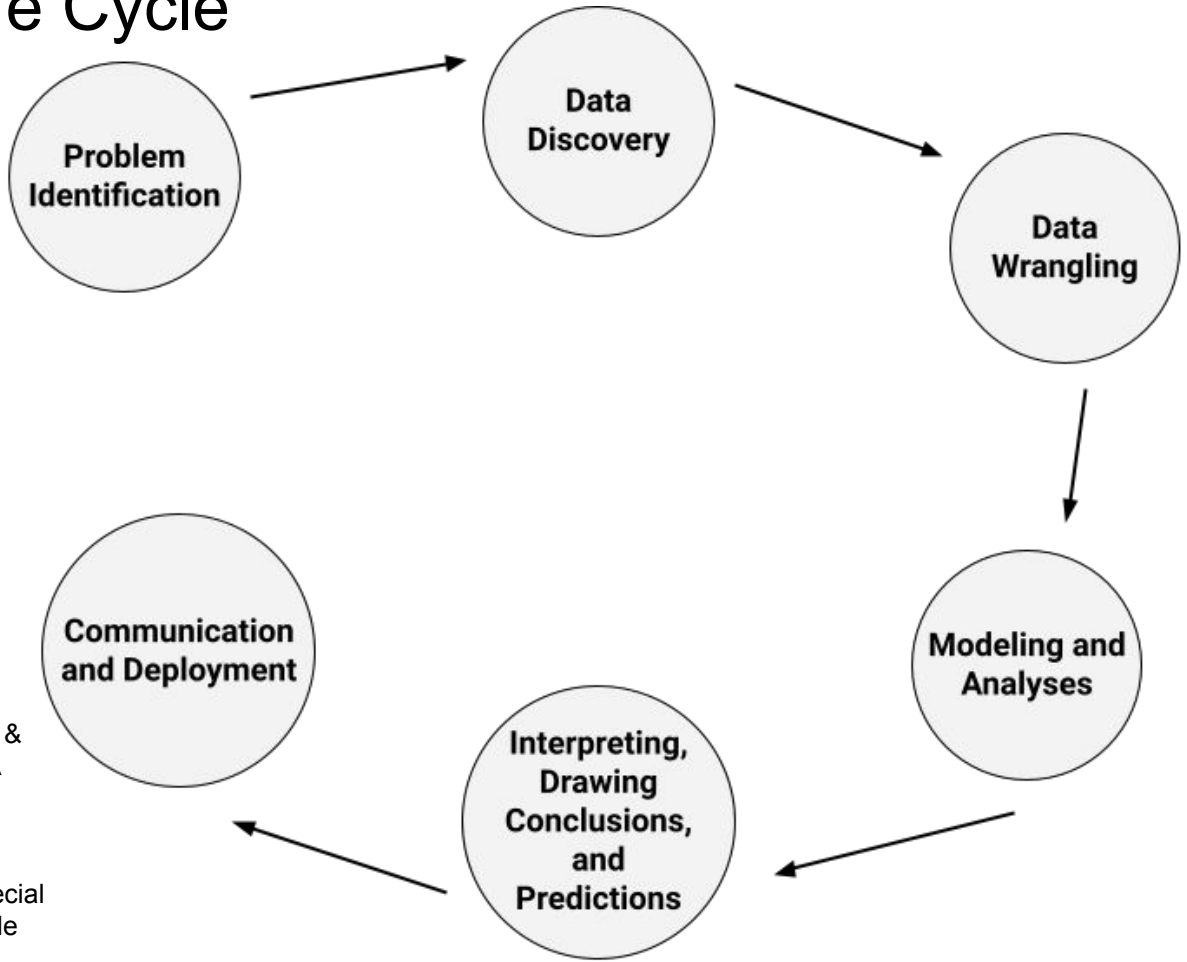
Huge list of purposefully unethical application of AI:

- Disinformation
- Surveillance/privacy
- Autonomous weapons

(***Activity***: *Add another unethical AI case study to EtherPad, if you can think of another example)*

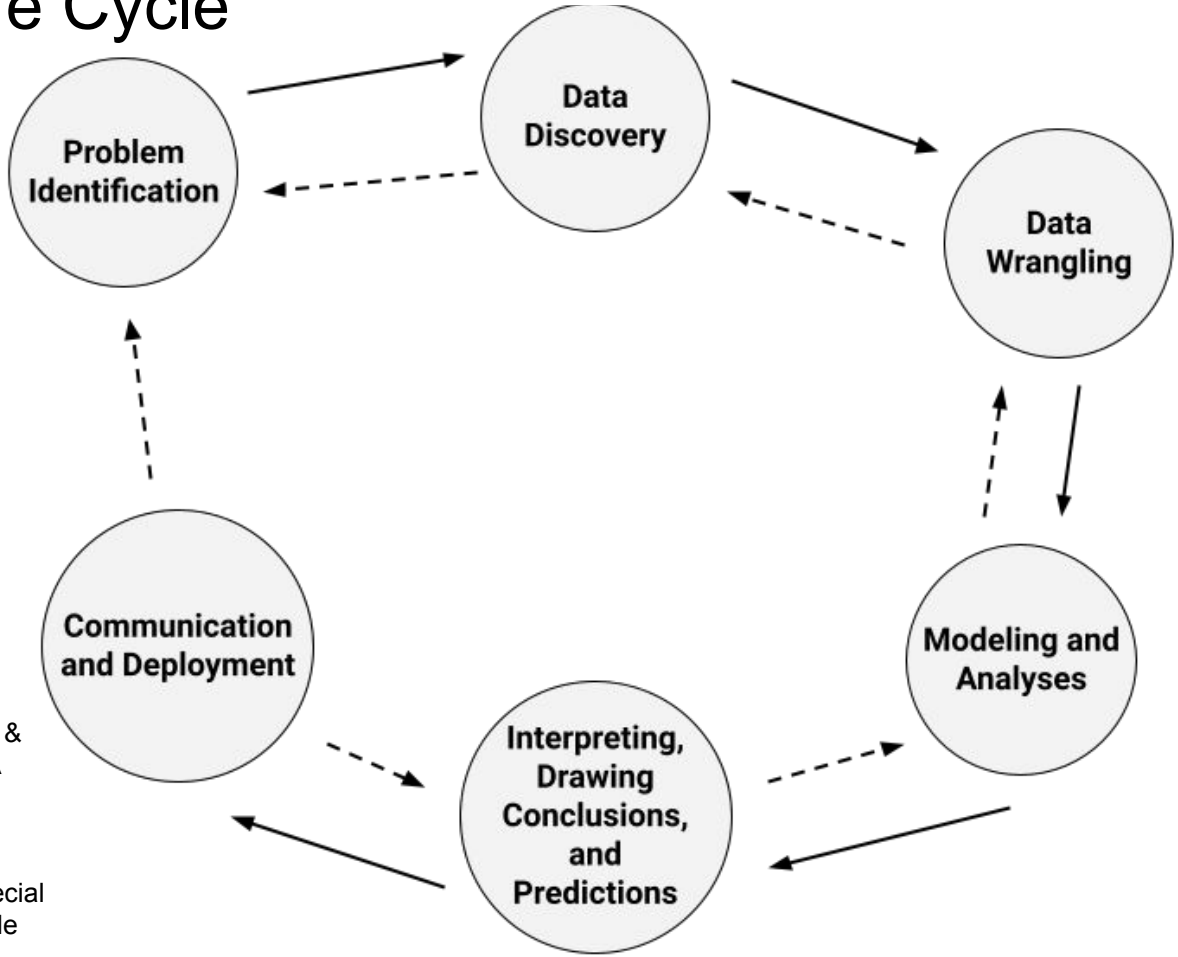
But we will be focusing on spotting unintentional biased or unethical practices that arise in the Machine Learning process.

Data Science Life Cycle



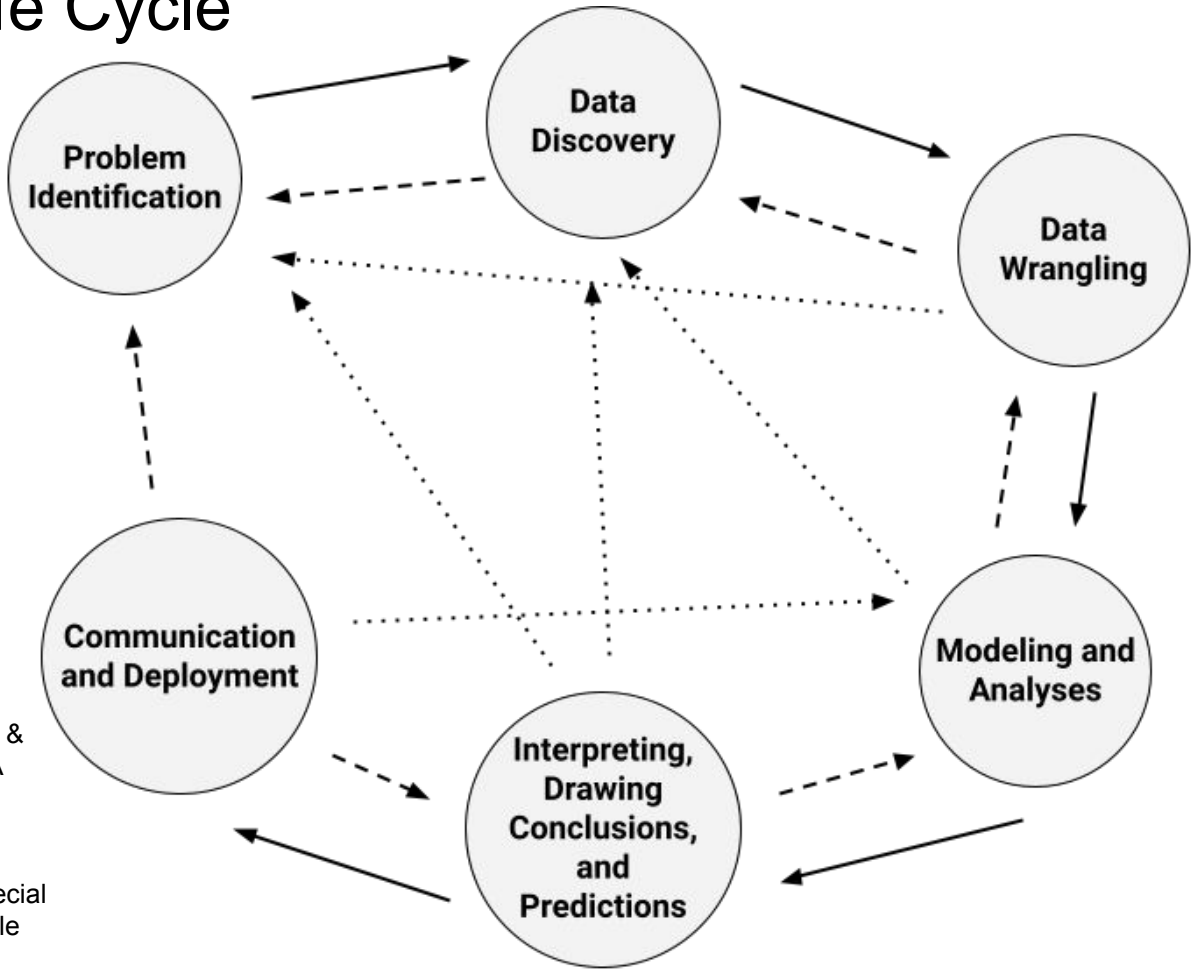
Refs:
Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.2d83f7f5
Academic Data Science Alliance Ethics Special Interest Group. Data Science Ethos Lifecycle

Data Science Life Cycle



Refs:
Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.2d83f7f5
Academic Data Science Alliance Ethics Special Interest Group. Data Science Ethos Lifecycle

Data Science Life Cycle



Refs:
Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.2d83f7f5
Academic Data Science Alliance Ethics Special Interest Group. Data Science Ethos Lifecycle

Bias and ethics should be considered at every step of the cycle

GLAM as field(s) that ML can learn from

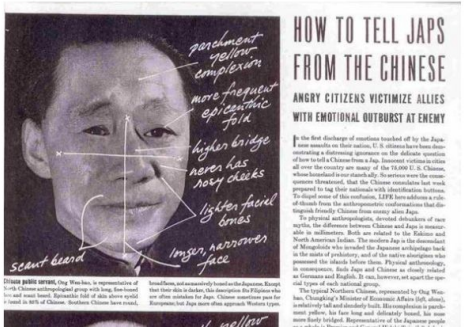
Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning

Eun Seo Jo
Stanford University
eunseo@stanford.edu

Timnit Gebru
Google
tgebru@google.com

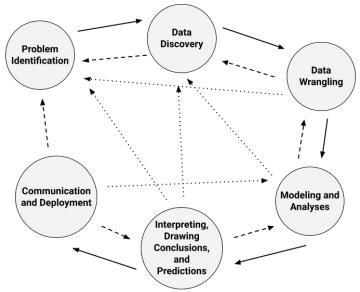
ABSTRACT

A growing body of work shows that many problems in fairness, accountability, transparency, and ethics in machine learning systems are rooted in decisions surrounding the data collection and annotation process. In spite of its fundamental nature however, data collection remains an overlooked part of the machine learning (ML) pipeline. In this paper, we argue that a new specialization should be formed within ML that is focused on methodologies for data collection and annotation: efforts that require institutional frameworks and procedures. Specifically for sociocultural data, parallels can be drawn from archives and libraries. 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



Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 306–316.
DOI:https://doi.org/10.1145/3351095.3372829

Life cycle: Problem Identification



The life cycle of a data science project starts with the definition of a problem or issue. Once the scope of the problem is defined, this is when it is important to set goals and confirm them with the project team.

Questions to ask:

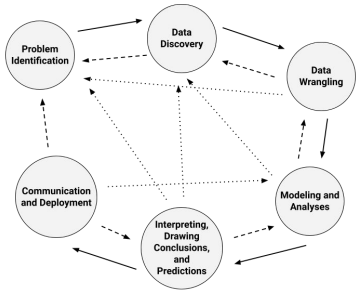
- What assumptions do you bring to the problem?
- Does the project plan incorporate regular checks, discussion, and documentation about the ethical dimensions of the project?
- What will the scope of this project be, and what are potential off-shoots to table for later?

Data Science Team

You will need the right mix of people and skills to successfully apply machine learning in a GLAM setting:

- *Domain expertise*: it is important that people with intimate knowledge of the collections or ‘business’ need to be addressed are involved in the development of machine learning approaches.
- *Project management*: to keep track of progress
- *IT*: if you are going to work with existing infrastructure you might need support from IT for things like the storage of data
- *Communication skills*: to communicate the goals of the project internally and potentially to external audiences

Data Discovery



Search for and identify potential data sources that fit the problem boundaries.

If existing datasets don't exist -- make them!

Questions to ask:

- What restrictions are there on access to the data?
- Who collected the data or did the annotations? For what purposes?
- Do the data include disproportionate coverage for different communities under study?
- Do data have adequate geographic coverage?

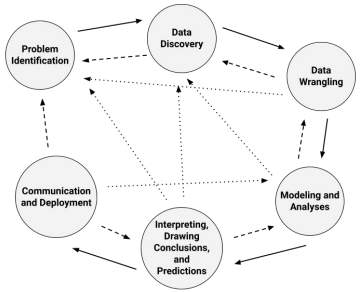
Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Wrangling

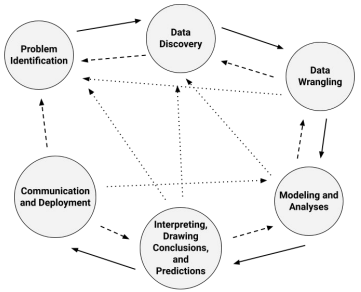


Data that you pull from available data sources are almost never in a format that you can plug directly into a machine learning model or visualization tool. You will likely need to do a lot of data "cleaning" and re-formatting.

Questions to ask:

- What is the quality of the data?
- If data fields are empty or missing, why?
- Are "corrections" of misspellings or standardizations documented?

Data Wrangling



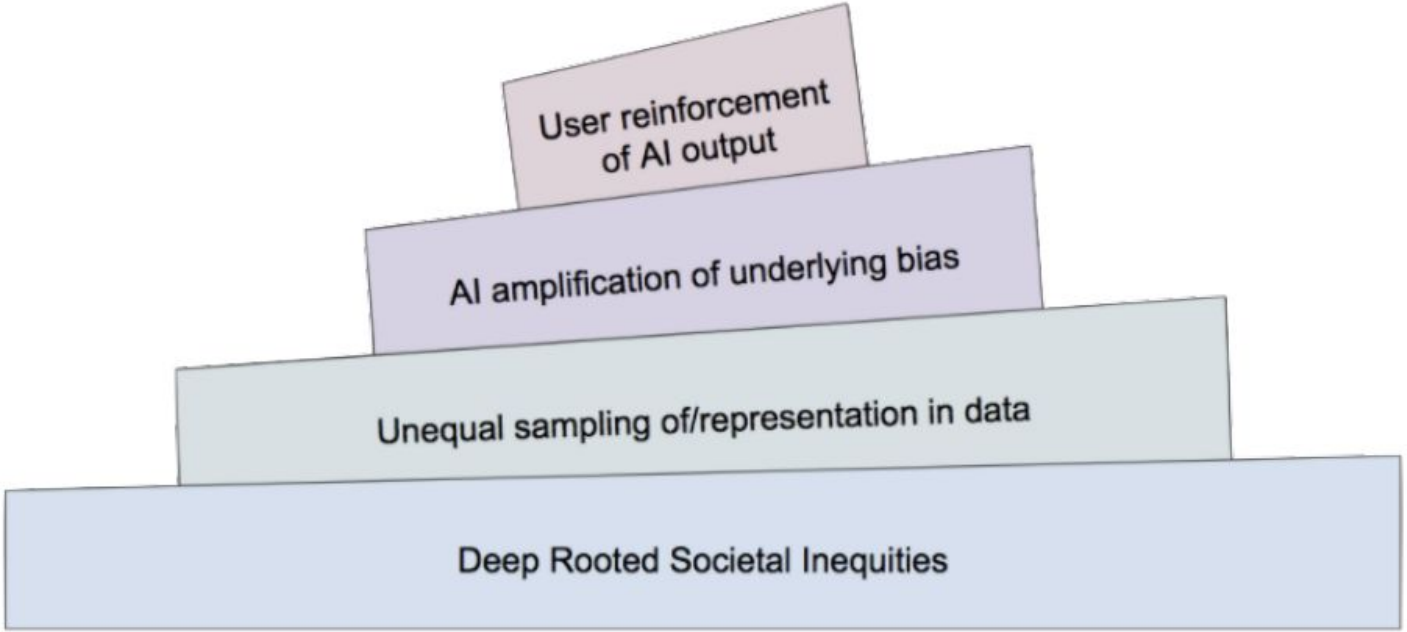
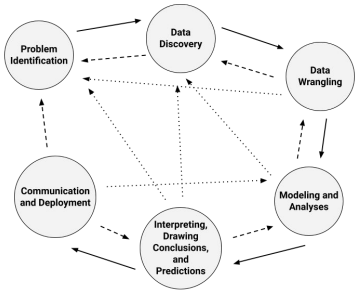


Figure originally created by Eric Wang, turnitin.com

Data Wrangling



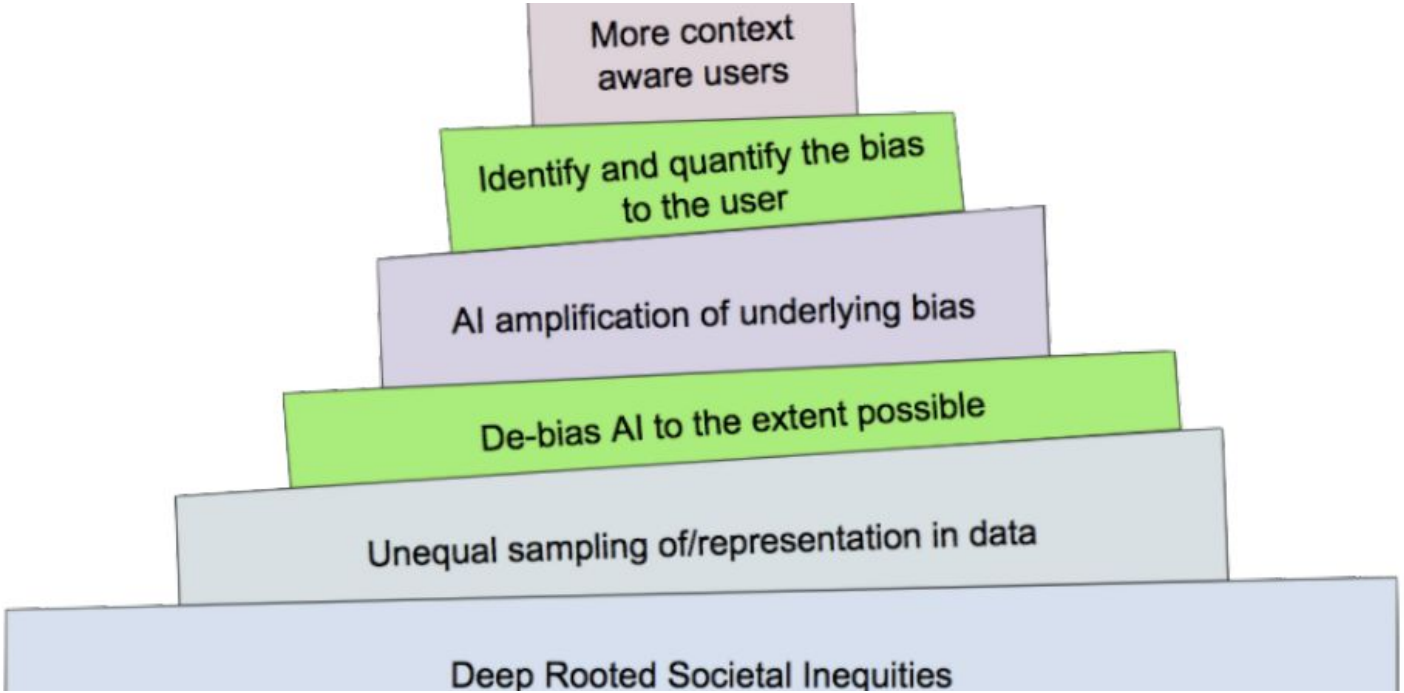
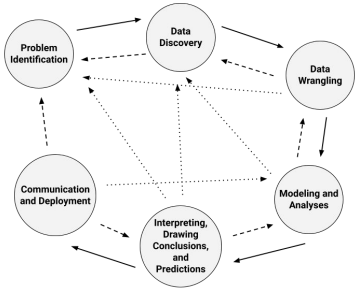


Figure originally created by Eric Wang, turnitin.com

Modeling and Analyses

Here is where the model-building discussed in previous lessons happens.



Questions to ask:

- Is the correct metric (such as classification accuracy) being optimized?
- Is a machine learning model necessary, or could a rules-based approach work?
- Could the results be improved by using additional kinds of data or other methods?
- If your model is a fine-tuned version of another model, do you know how that model was created?

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 learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and

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

1 INTRODUCTION

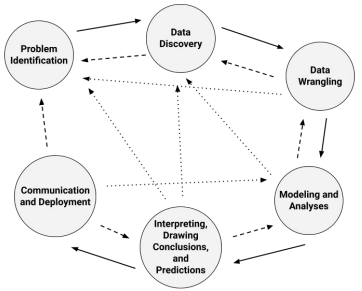
Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. DOI:https://doi.org/10.1145/3287560.3287596

Interpreting, Drawing Conclusions, and Predictions

Using the results of the prior analysis to infer new knowledge.

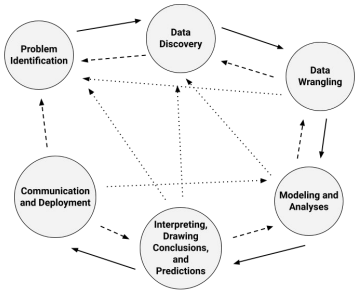
It is very important to examine the incorrect predictions, and involve domain experts in determining why the model failed in those instances.



Questions to ask:

- Are there benchmarks to compare the results?
- What do the false positive and false negative rates mean in context of this research?
- What actions do the results recommend? To whom?
- What are limitations in generalizing the results to other situations?

Communication and Deployment



The last step is to communicate the results with the research team and the wider public via conference presentation, journal articles or social media. Ongoing communication and dissemination are critical to ensure processes and results are transparent and reproducible. This step includes sharing data, source code, and instructions.

Questions to ask:

- What could have been done better?
- Are all data sources credited?
- What kinds of actions are possible now that the results have been disseminated?

Group Activity

Case study: "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings": https://doi.org/10.1038/s41467-020-18566-7

A group of researchers assembled a dataset of all portrait paintings from the UK National Portrait Gallery, and fine-tuned a facial recognition model to quantify the "trustworthiness" of people. They then used those trustworthiness scores to make associations with other variables (like year: https://www.nature.com/articles/s41467-020-18566-7/figures/1 or GDP) to make conclusions about how trustworthiness of individuals changes over time.

As a group, go through the steps of the data science life cycle, and see if you can identify where unethical or biased decisions were made. How could they have been avoided?

Options for Deployment: Batch

The most common method of deploying a model in a research setting is to run a batch of predictions -- load an entire dataset of files into a model, and collect the predictions.

This is the technically easiest mode to set up, but typically only 1 person (or a small team) are able to run predictions.

Options for Deployment: API

It is possible to connect a model to a Web endpoint, so that users can feed it data one example at a time -- on-demand.

Typically the results are returned in a machine-readable format (like JSON or XML), that can then be turned into a customized web portal -- or fed into additional steps of a pipeline.

Options for Deployment: API

Request URL

```
https://vision.googleapis.com/v1/images:annotate
```

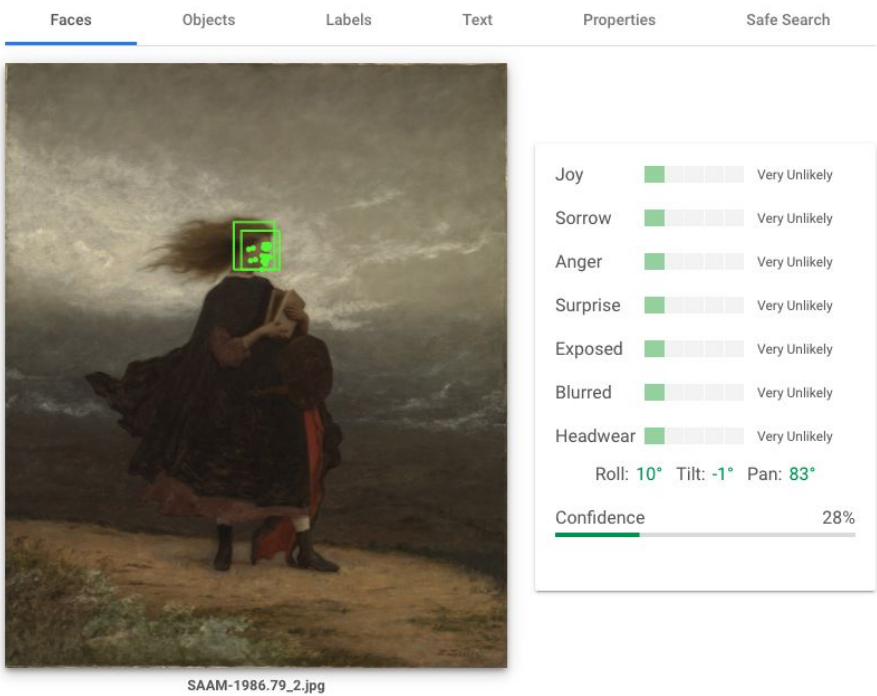
Request

```
{
  "requests": [
    {
      "features": [
        {
          "maxResults": 50,
          "type": "LANDMARK_DETECTION"
        },
        {
          "maxResults": 50,
          "type": "FACE_DETECTION"
        },
        {
          "maxResults": 50,
          "type": "OBJECT_LOCALIZATIO
N"
        },
        {
          "maxResults": 50,
          "type": "LOGO_DETECTION"
        },
        {
```

Response

```
{
  "cropHintsAnnotation": {
    "cropHints": [
      {
        "boundingPoly": {
          "vertices": [
            {},
            {
              "x": 960
            },
            {
              "x": 960,
              "y": 1199
            },
            {
              "y": 1199
            }
          ]
        },
        "confidence": 0.224588,
        "importanceFraction": 1
      }
```

JSON response from an API



The same response converted to visualize the results

Activity: Try out an image classification API

Go to https://www.si.edu/spotlight/open-access, and select an image to classify. Download it to your computer.

Now go to https://cloud.google.com/vision#try-the-api to test out the image on the Google Vision API. (Make sure to tell the site you are not a robot)

Do the results make sense?

What bias or ethics questions do the results prompt?