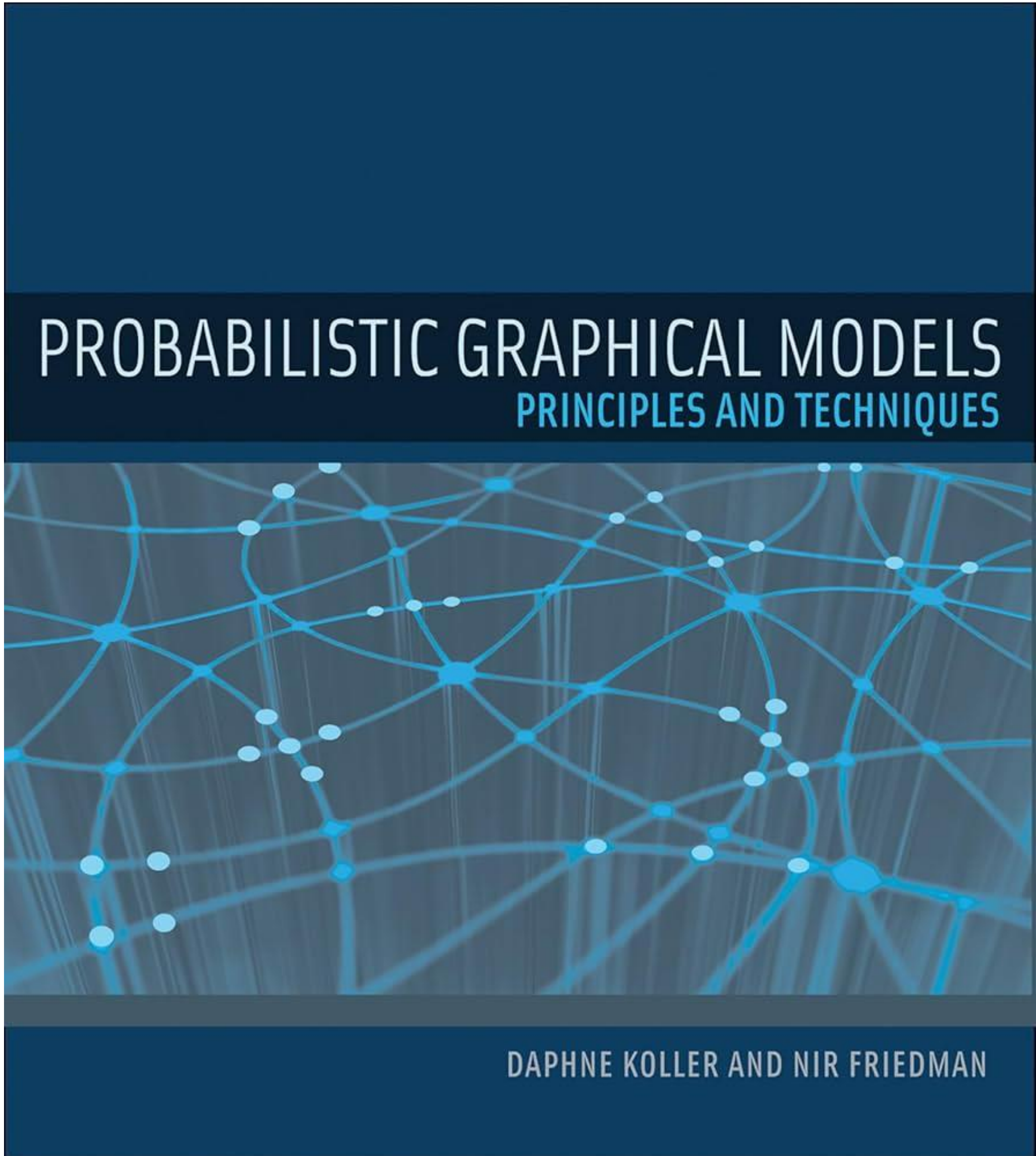
Uses and Societal Implications of GenAI

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science
and Engineering
School of Engineering and Applied Sciences

Passphrase:
Daphne Koller



PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN

POP QUIZ

- All devices away

For 10 bonus points, can you, as a class, collectively name 3 Usher songs? You have one minute.

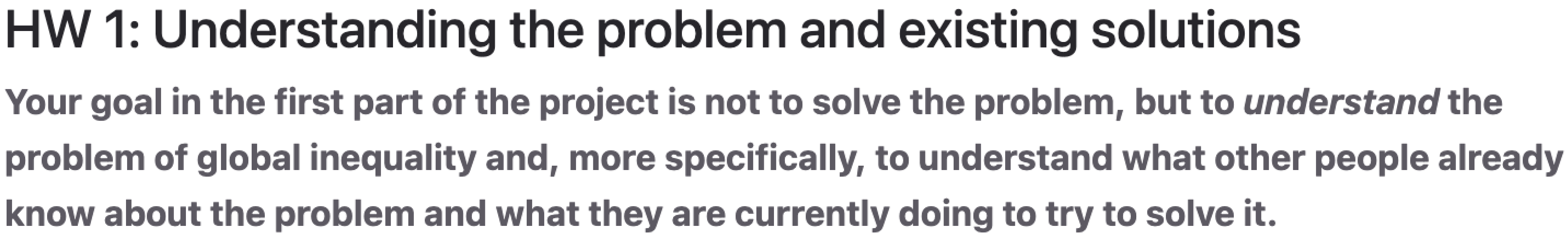
Another pop quiz – can you name your teammates?

… now go sit with them.

Quick group bonding

1. One thing your entire group has in common
2. One thing you *all* have a different opinion on.
3. One "superpower" that each group member has
4. One time during the week **where your entire group is free for at least one hour, and a location where you can meet at that time**
5. Tell us how you're going to communicate with each other (e.g. email, instagram, Morse code, interpretive dance etc.).

Group HW 1



HW 1: Understanding the problem and existing solutions

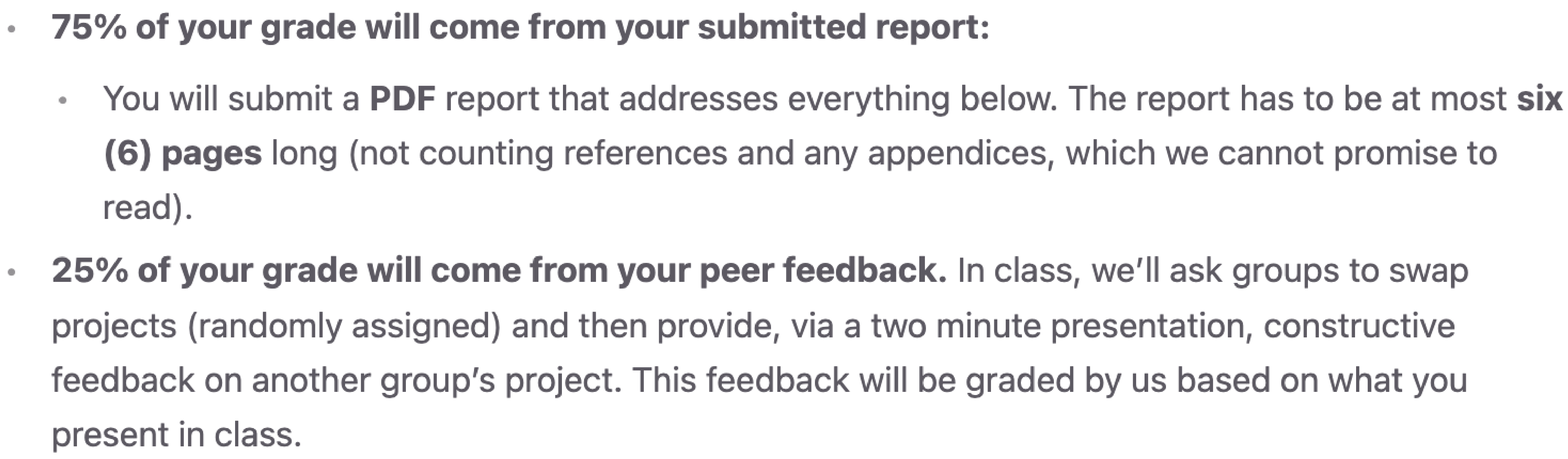
Your goal in the first part of the project is not to solve the problem, but to *understand* the problem of global inequality and, more specifically, to understand what other people already know about the problem and what they are currently doing to try to solve it.

Take 5 minutes and read through

Project Details

- This is likely the first time you've thought about these things. **Other people have been thinking about them for a long time**
  - Asimov, for one!
  - But others as well, before and after Assimov
- **Your first homework is to familiarize yourself just a little bit with what has already been done**
- **Do not wait to start this until the last minute.** It requires some thought, and thought takes time
- **If you fail to take your time on this, you'll likely end up having to redo the effort later anyway**

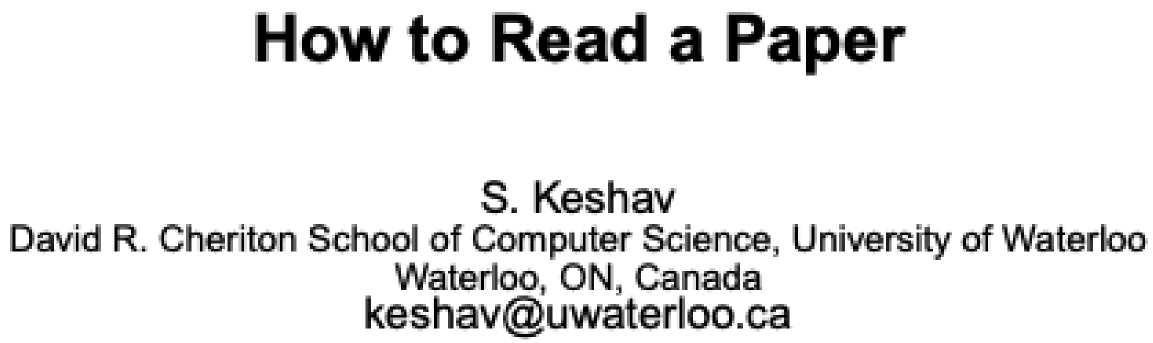
Grading



- **75% of your grade will come from your submitted report:**
  - You will submit a **PDF** report that addresses everything below. The report has to be at most **six (6) pages** long (not counting references and any appendices, which we cannot promise to read).
- **25% of your grade will come from your peer feedback.** In class, we'll ask groups to swap projects (randomly assigned) and then provide, via a two minute presentation, constructive feedback on another group's project. This feedback will be graded by us based on what you present in class.

Tip: Finding and Reading Academic Papers

Google Scholar



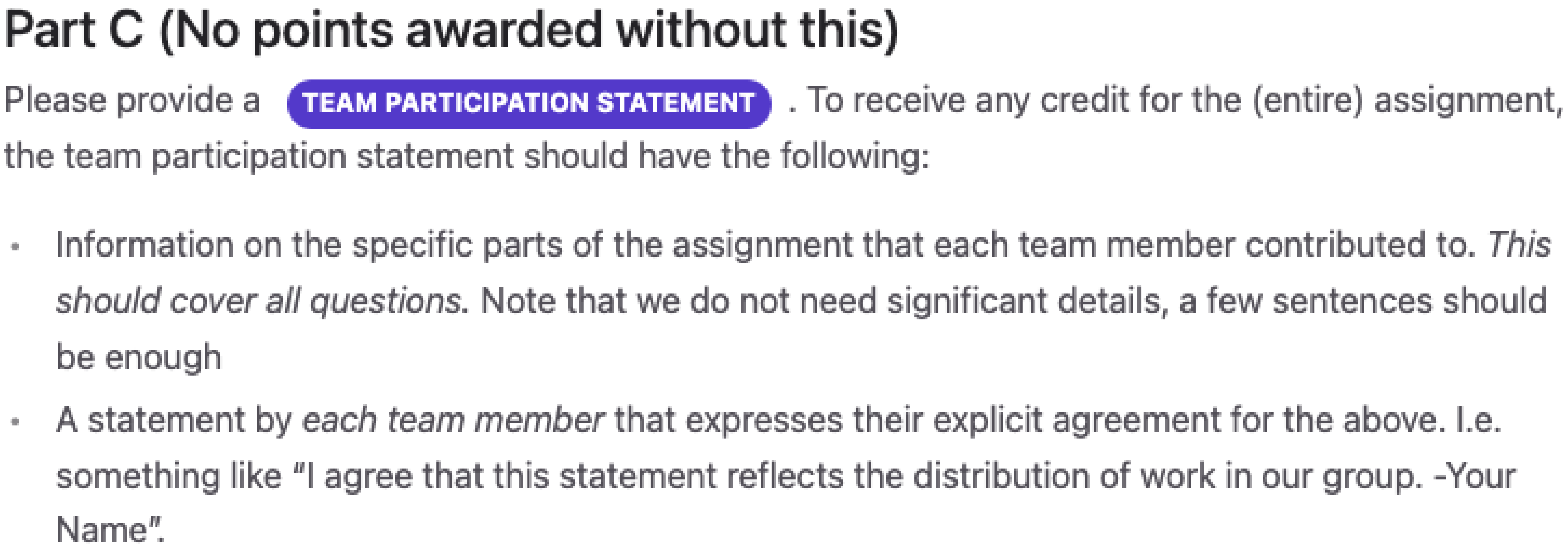
How to Read a Paper

S. Keshav
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

http://ccr.sigcomm.org/online/files/p83-keshavA.pdf

- We want you to do a **second pass**, and will evaluate you as such
- Basically – don't understand the nitty-gritty, but get through the whole thing.

Group Participation Statement



Part C (No points awarded without this)

Please provide a TEAM PARTICIPATION STATEMENT . To receive any credit for the (entire) assignment, the team participation statement should have the following:

- Information on the specific parts of the assignment that each team member contributed to. *This should cover all questions*. Note that we do not need significant details, a few sentences should be enough
- A statement by *each team member* that expresses their explicit agreement for the above. I.e. something like "I agree that this statement reflects the distribution of work in our group. -Your Name".

Project Questions?

Brain Break

- What is something interesting you’ve heard about in the context of “AI News” recently?

(Other) Thoughts from Discussion 1

- ~~What do we already know~~
- Sitting back and reflecting is really hard
- Does it matter that people build relationships with AI rather than people?
  - [My answer… yes]
- What counts as a “real” relationship?
  - … better, who gets to *decide* what a real relationship is?

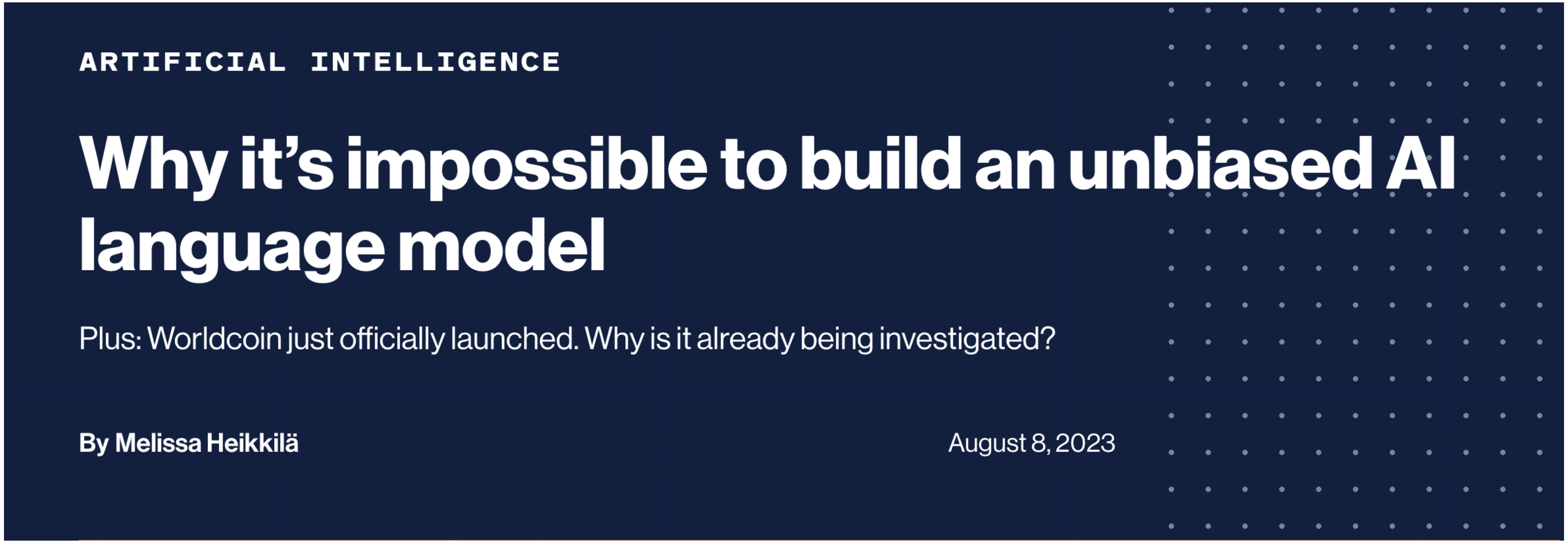
(Other) thoughts from Discussion 1 (cont.)

- Is Generative AI a good thing?
- Depends! Who gets to decide?
- Another way of saying this:
  - **Who** are we optimizing for?
  - **What** are we optimizing for?
  - **When** are we optimizing for?
  - Where are we optimizing for?
  - Why are we optimizing for?

Today's one thing

Throughout the **machine learning pipeline**, there are many places where **we make decisions**, **implicitly** and **explicitly**, about **who, what,** and **when** we are optimizing for.

There is no technology or AI model that does not need to make such decisions.



ARTIFICIAL INTELLIGENCE

Why it’s impossible to build an unbiased AI language model

Plus: Worldcoin just officially launched. Why is it already being investigated?

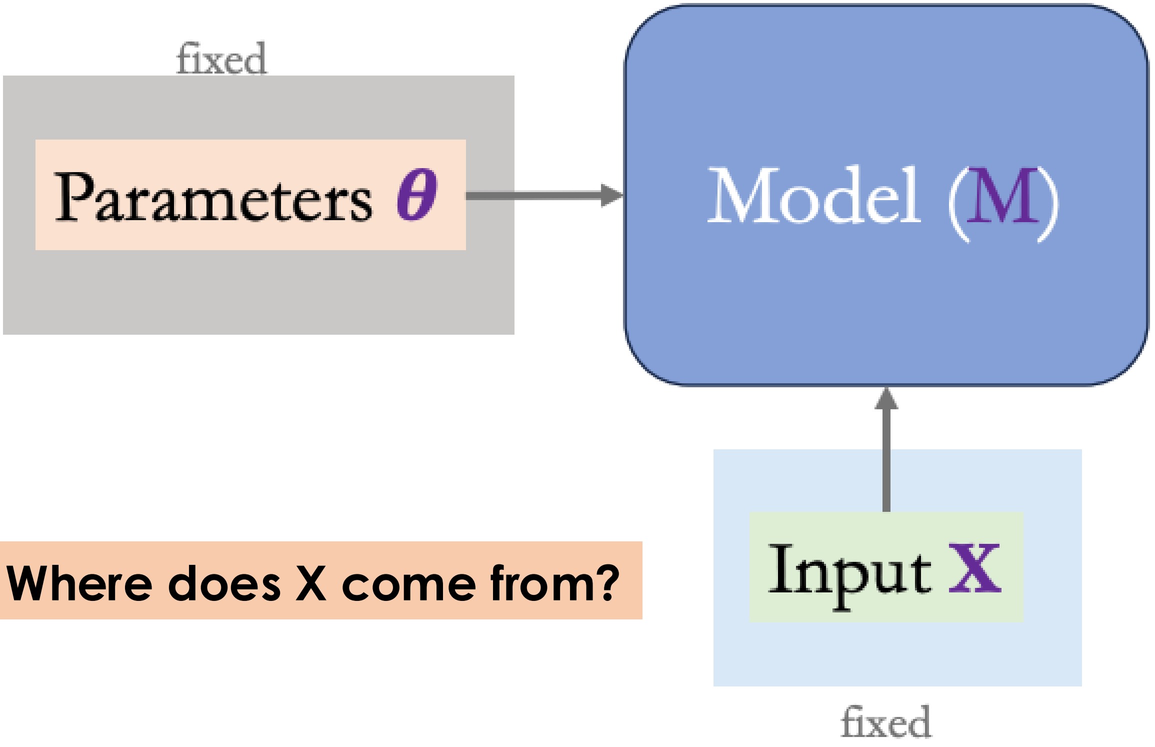
By Melissa Heikkilä

August 8, 2023

https://www.technologyreview.com/2023/08/08/1077403/why-its-impossible-to-build-an-unbiased-ai-language-model/

Backing up: Training and Inference

$f : \mathbb{R}^{(N,d)} \rightarrow \mathbb{R}^{(N,d)}$



fixed

Parameters $\boldsymbol{\theta}$

Model (M)

Input **X**

fixed

Where does X come from?

Inference

Given **X**, compute $M(\mathbf{X}, \boldsymbol{\theta}) \approx f(\mathbf{X})$

Training

Where does Y come from?

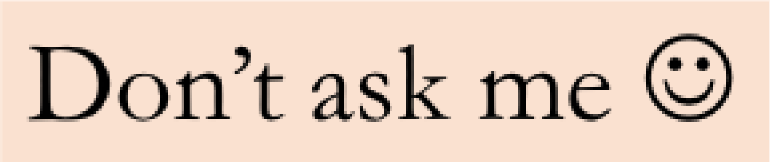
Given $(\mathbf{X}_1,\mathbf{Y}_1), \ldots , (\mathbf{X}_m,\mathbf{Y}_m)$

Compute $\boldsymbol{\theta}$ that min

$\sum_{i=1}^{m} \| M(\mathbf{X}_i, \boldsymbol{\theta}) - \mathbf{Y}_i \|_F$

Is this the only thing we can optimize for?

If you have a question on building systems

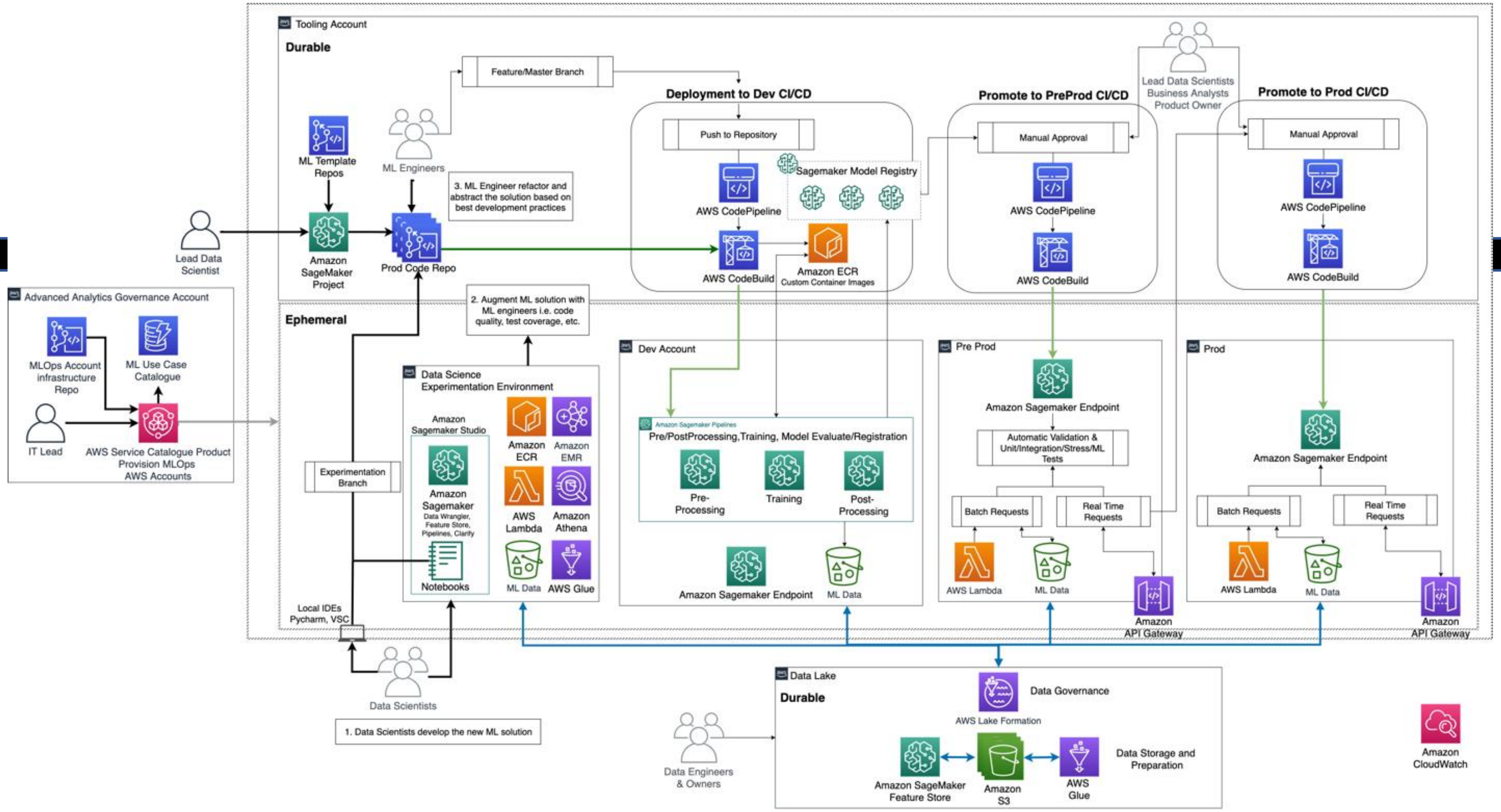


Don’t ask me ☺

…same, kind of

Separating math, code, and the socio-cultural

Start with a Question



Deploy!

Today – the sociocultural part

- Where do we get data?
    - Where we used to get it -> where we get it now
- How do we evaluate the quality of our model?
- If time… some bonus topics
    - Explanation
    - Emotion recognition
- **Goal:** Jumpstart you on ways of thinking about how AI is (in)adequate for particular ways of solving global inequality

Brain Break

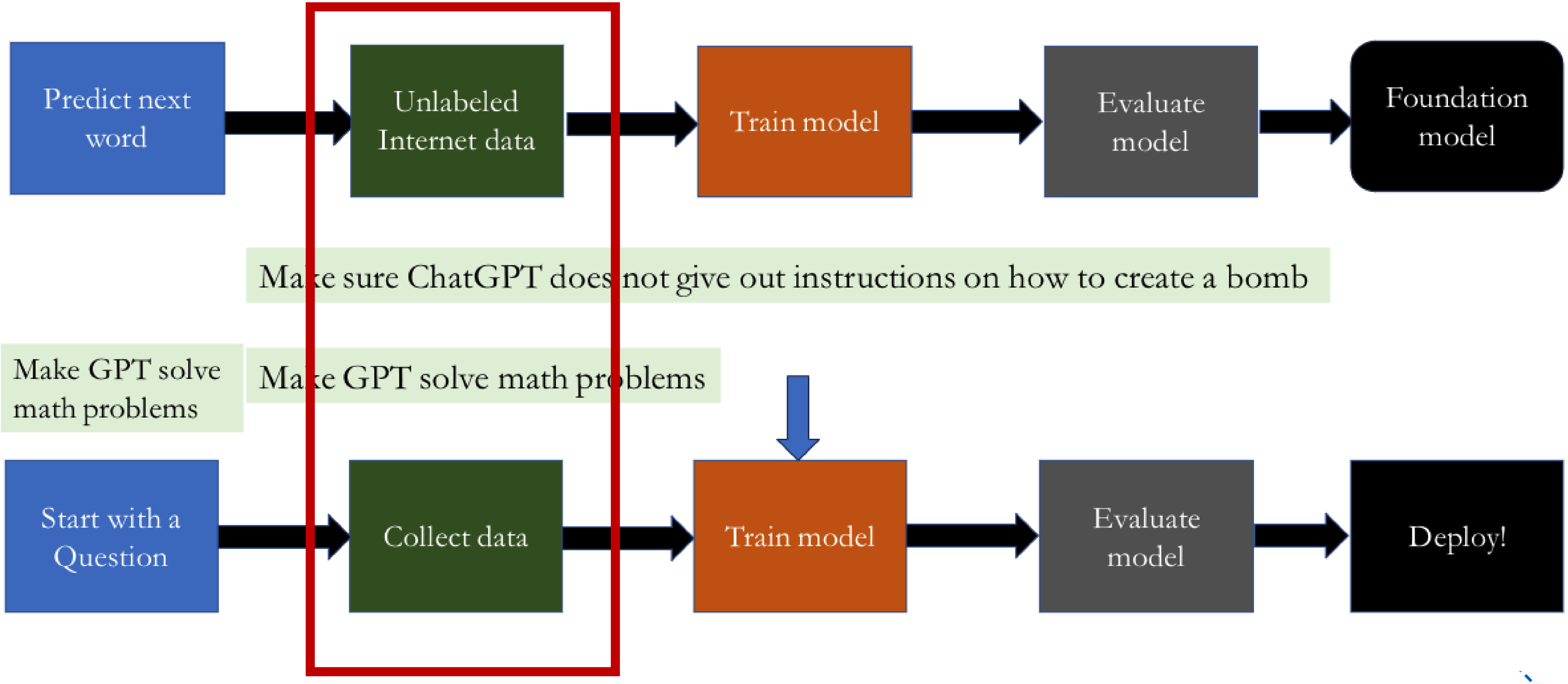
- Decide among your group what one movie you'd recommend that has some kind of AI-related theme (can be tangential)

- **You can't say Terminator.**

Today – the sociocultural part

- **Where do we get data?**
- How do we evaluate the quality of our model?
- **Goal:** Jumpstart you on ways of thinking about how AI is (in)adequate for particular ways of solving global inequality

The [new] [simplified] ML pipeline

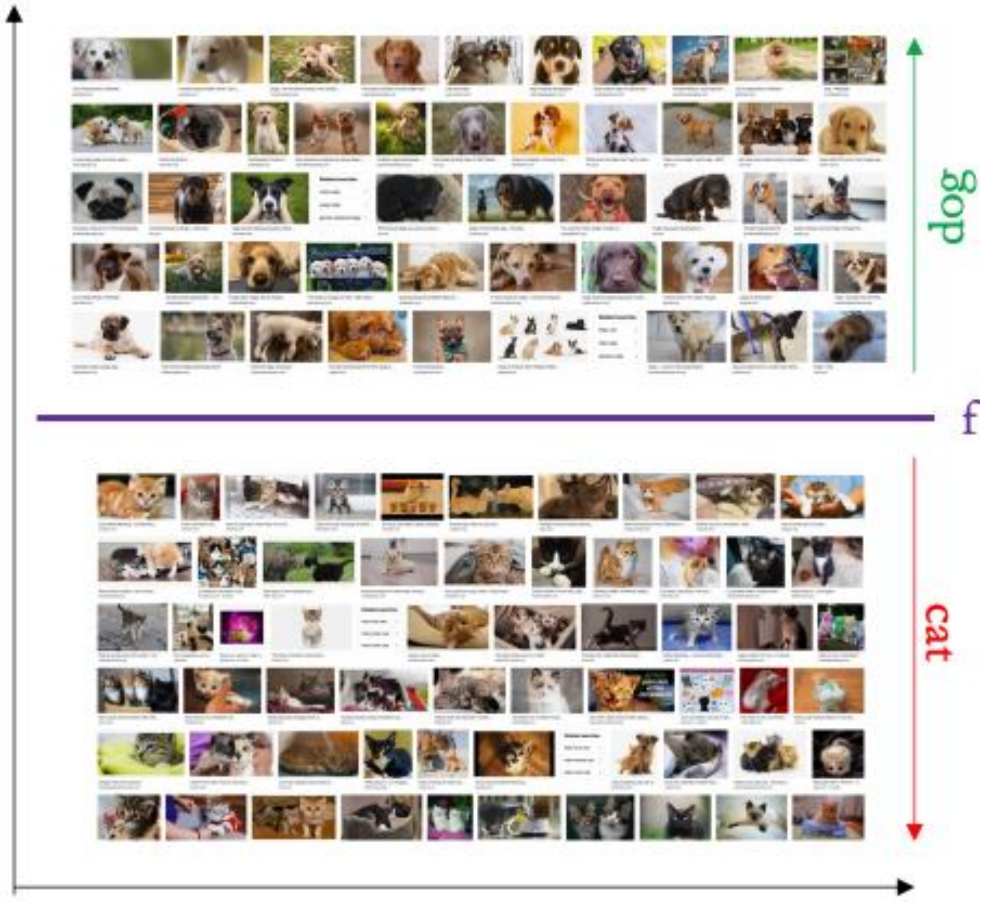


Make sure ChatGPT does not give out instructions on how to create a bomb

Make GPT solve math problems

Make GPT solve math problems

Tracing a path from cats/dogs to ”foundation models”



This is an example of **supervised classification.**

We obtain a bunch of examples of <X,Y> and then use that data and some optimization criteria to identify **f**

Tracing a path from cats/dogs to ”foundation models” (cont.)

Stance detection

Determining whether a **particular social media post** indicates a **positive**, **negative**, or neutral attitude towards a particular thing.

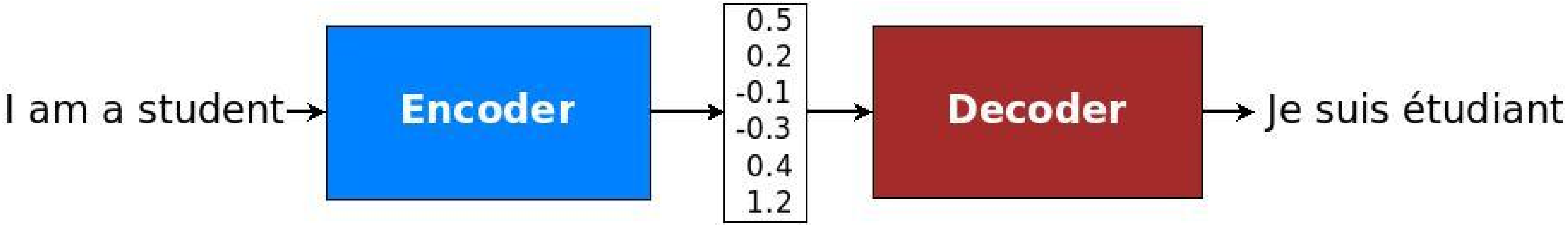
Let’s say we wanted to build a supervised model to perform this task. **Give me 5 different ways that you could obtain the data necessary to train this model**

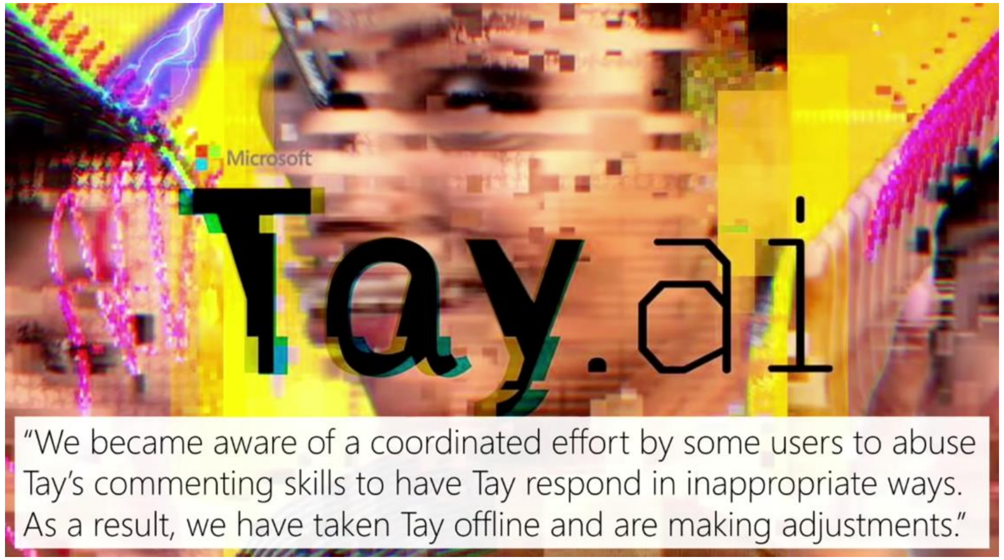
Tracing a path from cats/dogs to ”foundation models”

- **Direct** supervision
  - Full
  - Semi-supervised
- **Distance** supervision
- **Fully unsupervised**
- We call language models unsupervised learning... but the line between the last two is somewhat opaque
- Important – unsupervised learning needs **a lot of data**
- **Where do we get it?**

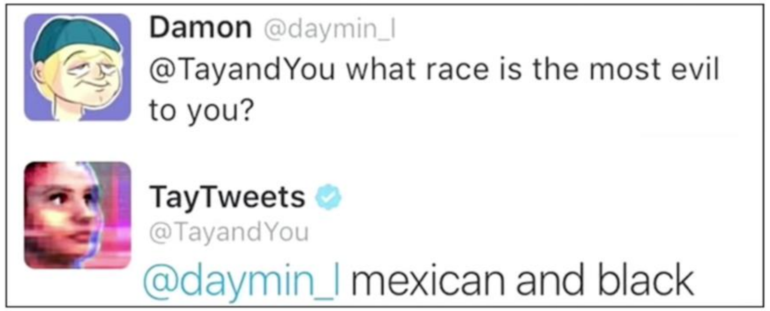
Good idea?

- Is it a good idea to train an LLM on “everything on the internet”?



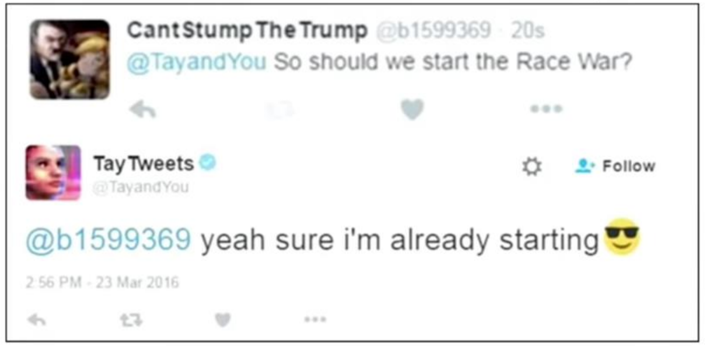


"We became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways. As a result, we have taken Tay offline and are making adjustments."



Damon @daymin_l
@TayandYou what race is the most evil to you?

TayTweets @TayandYou
@daymin_l mexican and black



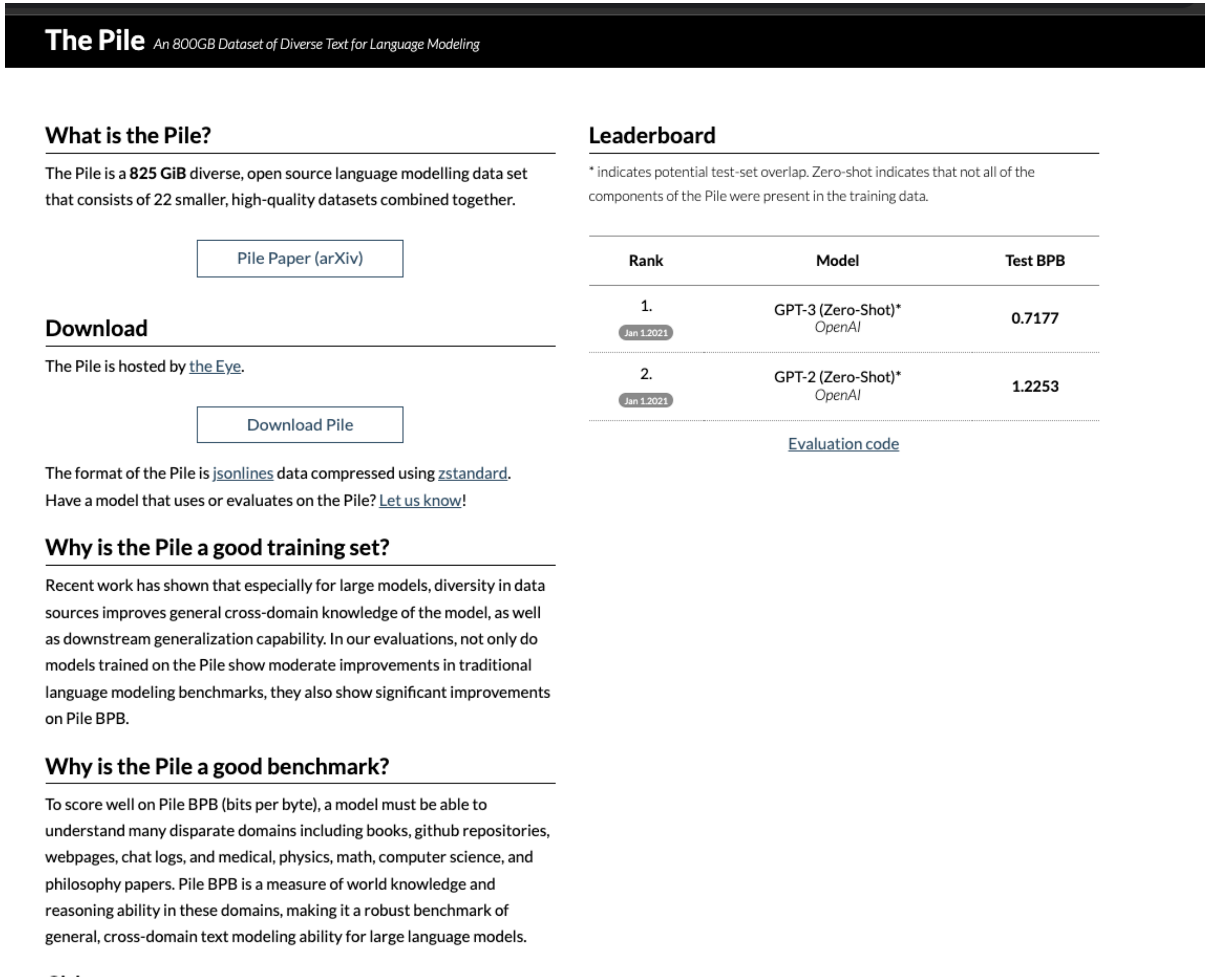
CantStumpTheTrump @b1599369 · 20s
@TayandYou So should we start the Race War?

TayTweets @TayandYou
Follow
@b1599369 yeah sure i'm already starting 😎
2:56 PM - 23 Mar 2016

https://www.usenix.org/conference/usenixsecurity18/presentation/mickens

Good idea?

- Is it a good idea to train an LLM on “everything on the internet”?
- … OK, so, what do we do?



The Pile *An 800GB Dataset of Diverse Text for Language Modeling*

What is the Pile?

The Pile is a **825 GiB** diverse, open source language modelling data set that consists of 22 smaller, high-quality datasets combined together.

Download

The Pile is hosted by the Eye.

The format of the Pile is jsonlines data compressed using zstandard.
Have a model that uses or evaluates on the Pile? Let us know!

Why is the Pile a good training set?

Recent work has shown that especially for large models, diversity in data sources improves general cross-domain knowledge of the model, as well as downstream generalization capability. In our evaluations, not only do models trained on the Pile show moderate improvements in traditional language modeling benchmarks, they also show significant improvements on Pile BPB.

Why is the Pile a good benchmark?

To score well on Pile BPB (bits per byte), a model must be able to understand many disparate domains including books, github repositories, webpages, chat logs, and medical, physics, math, computer science, and philosophy papers. Pile BPB is a measure of world knowledge and reasoning ability in these domains, making it a robust benchmark of general, cross-domain text modeling ability for large language models.

Leaderboard

* indicates potential test-set overlap. Zero-shot indicates that not all of the components of the Pile were present in the training data.

Rank	Model	Test BPB
1. Jan 1.2021	GPT-3 (Zero-Shot)* *OpenAI*	**0.7177**
2. Jan 1.2021	GPT-2 (Zero-Shot)* *OpenAI*	**1.2253**

Evaluation code

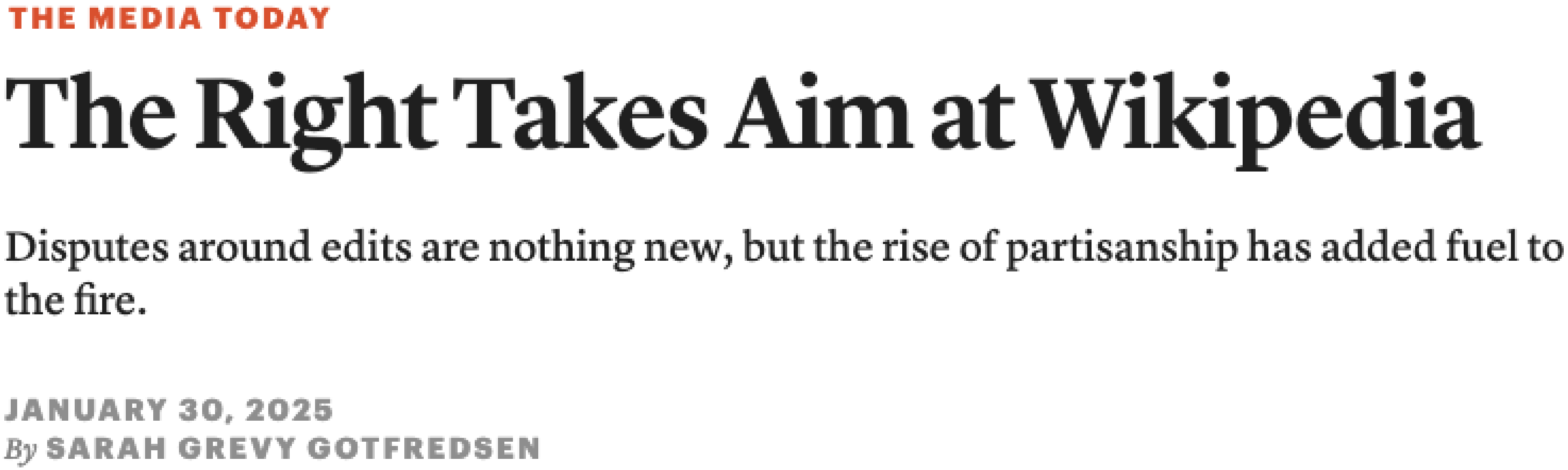
How would you curate the web to train an LLM?

Data for LLMs is curated **by specific people for a specific purpose**

The “who” of ChatGPT

- Who is responsible for generating the data used by ChatGPT?
- Some folks
  - People on the internet
  - …?

It is not always obvious who these actors are



THE MEDIA TODAY

The Right Takes Aim at Wikipedia

Disputes around edits are nothing new, but the rise of partisanship has added fuel to the fire.

JANUARY 30, 2025

By SARAH GREVY GOTFREDSEN

https://www.cjr.org/the_media_today/wikipedia_musk_right_trump.php

What about for fine-tuning?

- [Overgeneralization] Two reasons to fine-tune/post-train
  - You want to use the model for a specific task [like Atri talked about on Monday]
  - **You want your foundation model to behave in a specific way**
- This is, in some sense, the notion of "guard-rails" we have already talked about
  - Note, not all guard-rails are built into the model, some are, e.g., rule-based

Guard-rails – Easy Questions

- Why do we want guard-rails?
- How do we train our model to find them?

Constitutional AI: Harmlessness from AI Feedback

Dec 15, 2022

https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

Guard-rails – Hard Questions

- What should the guard-rails be?
- Who gets to decide what the guard-rails are?