It's all in the data

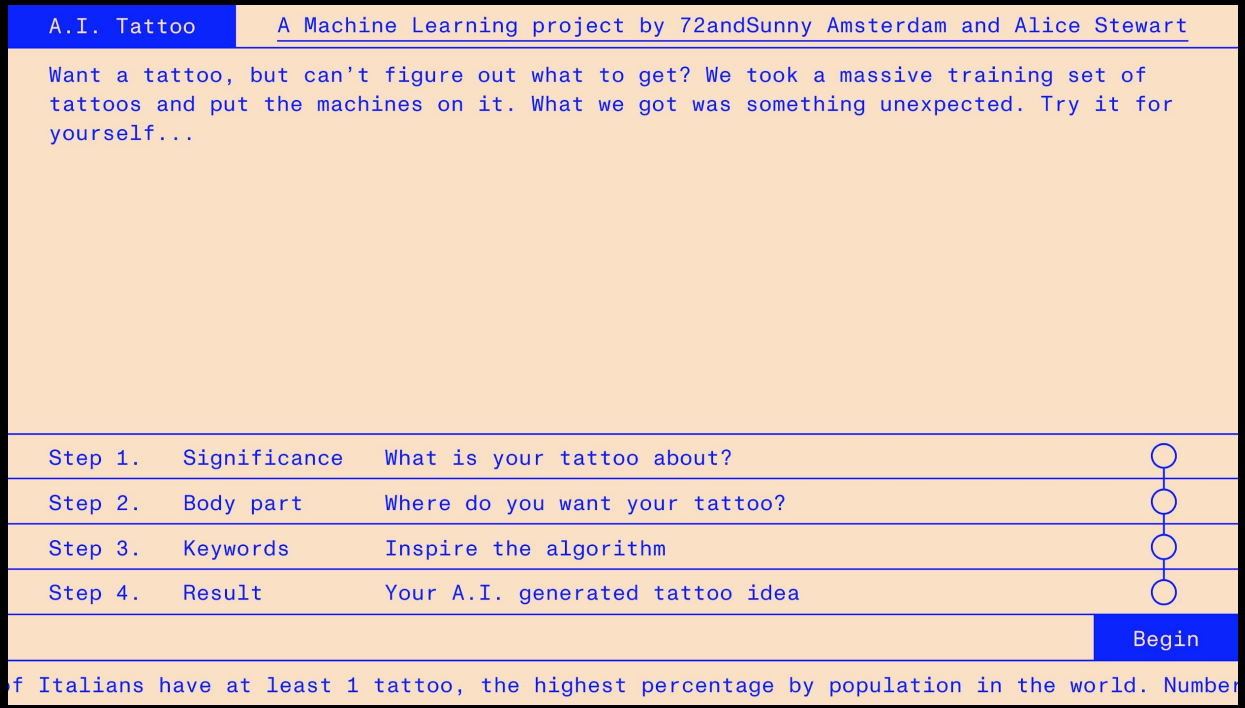
Alex Fefegha
Computational Futures & AI
a.fefeghaetta@arts.ac.uk

ual: creative computing institute

Alice Stewart :)

Creative Technologist

ual: creative computing institute



https://ai.tattoo/

ual: creative computing institute

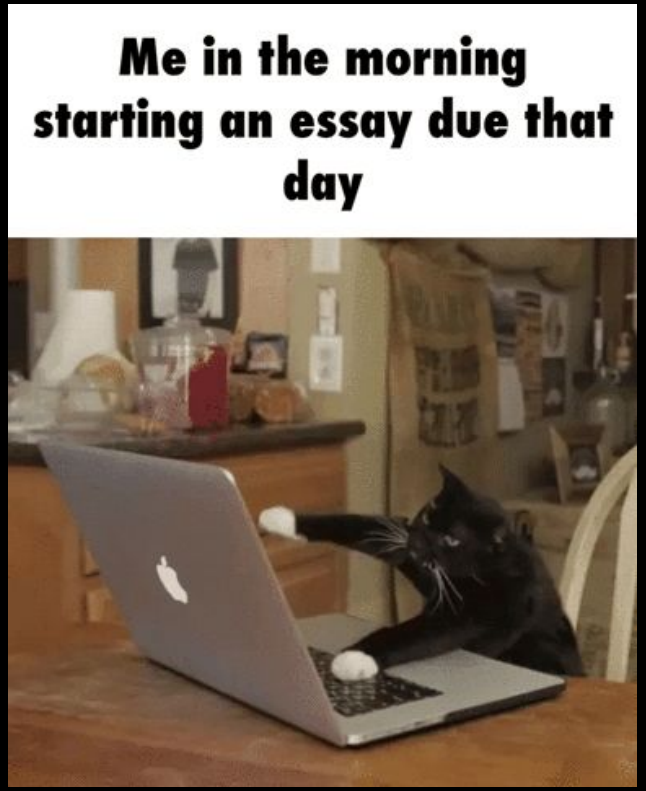
The snowball activity!


TARY

Slide here for a Break

How is the homework going?

Let's recap on the essay.



Using examples, identify and develop an essay containing three case studies that exemplify an argument of how artificial intelligence is culturally constructed or explained..

(2000 words)

Your case studies should triangulate a key argument that explores how narratives, myths and rhetoric develops around AI and how these are challenged or counteracted. misused or exploited, where deviance enters the use case, or where they are used out of their assumed context.

Your case studies may be drawn from commercial, activist, artistic or other fields.

400 word introduction.

What is machine learning?
How do machines learn?
Historical overview?

400 words per case study.

Work by artists?
Work by ML researchers?
Work by authors?
Work by designers?
Work by governments?
Work by companies?

400 words conclusion.

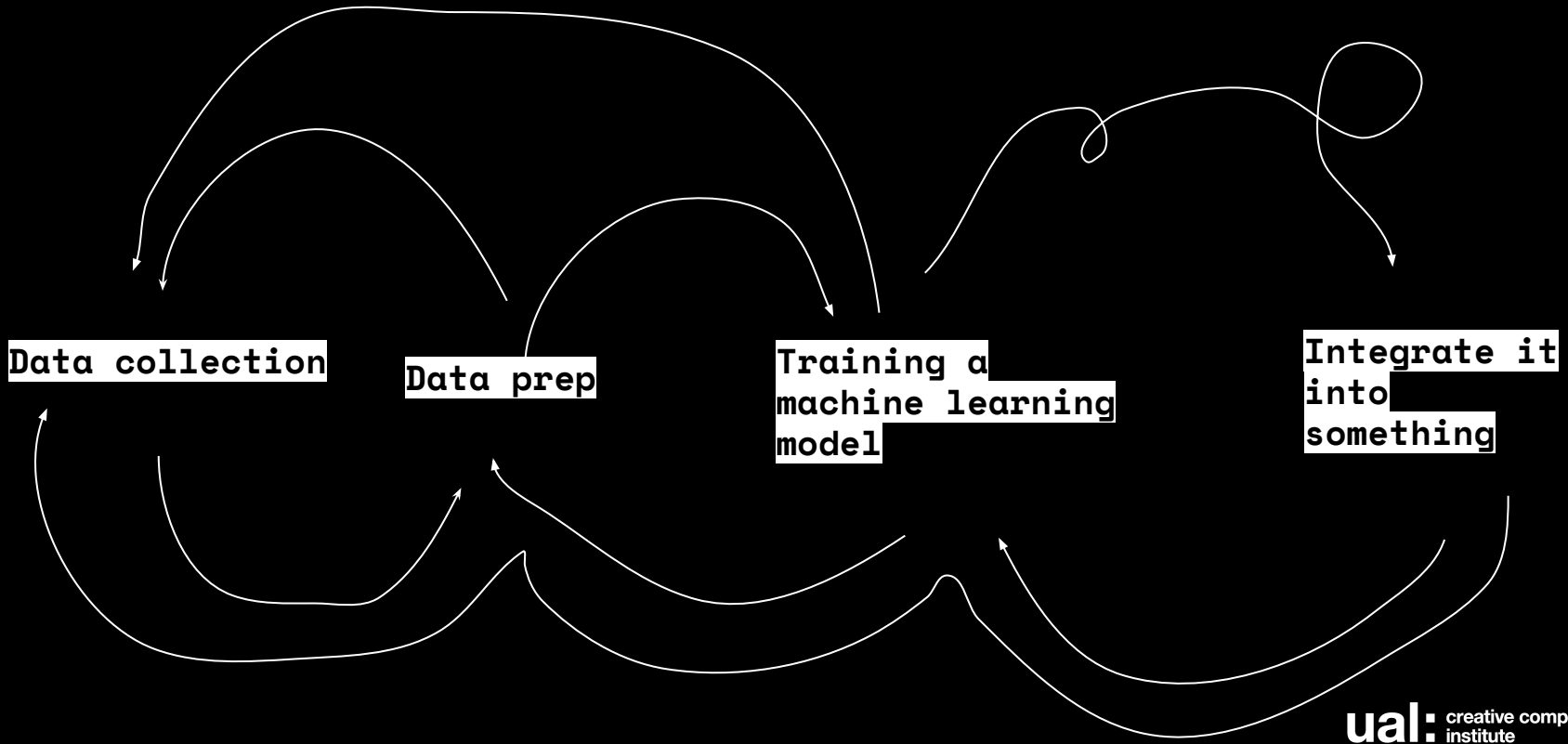
Based on all you absorbed and conducted, what would your artistic speculation provocation on machine learning be? (*something like that*)

Let’s talk about data!

Without data, we can’t train any model.

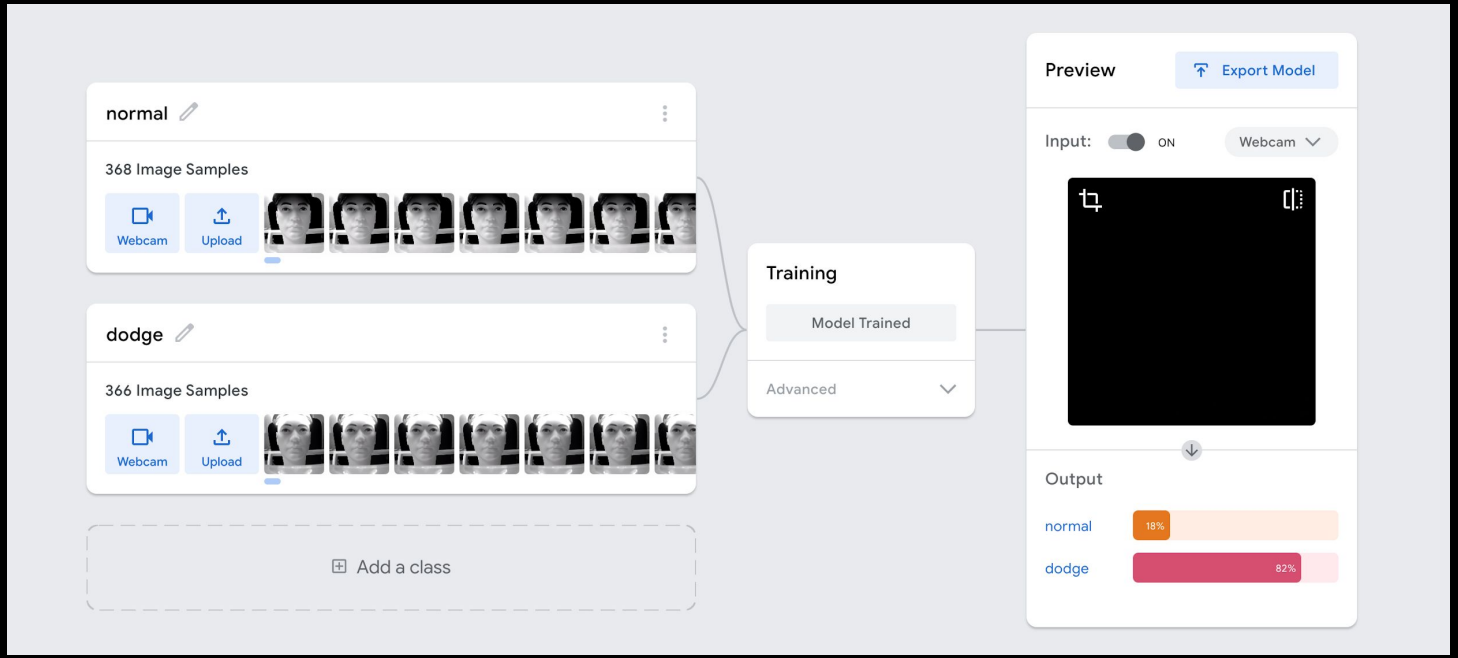
For a AI powered experience to work, the ML model needs to be able to recognize patterns and correlations in data.

How the machine learning process works

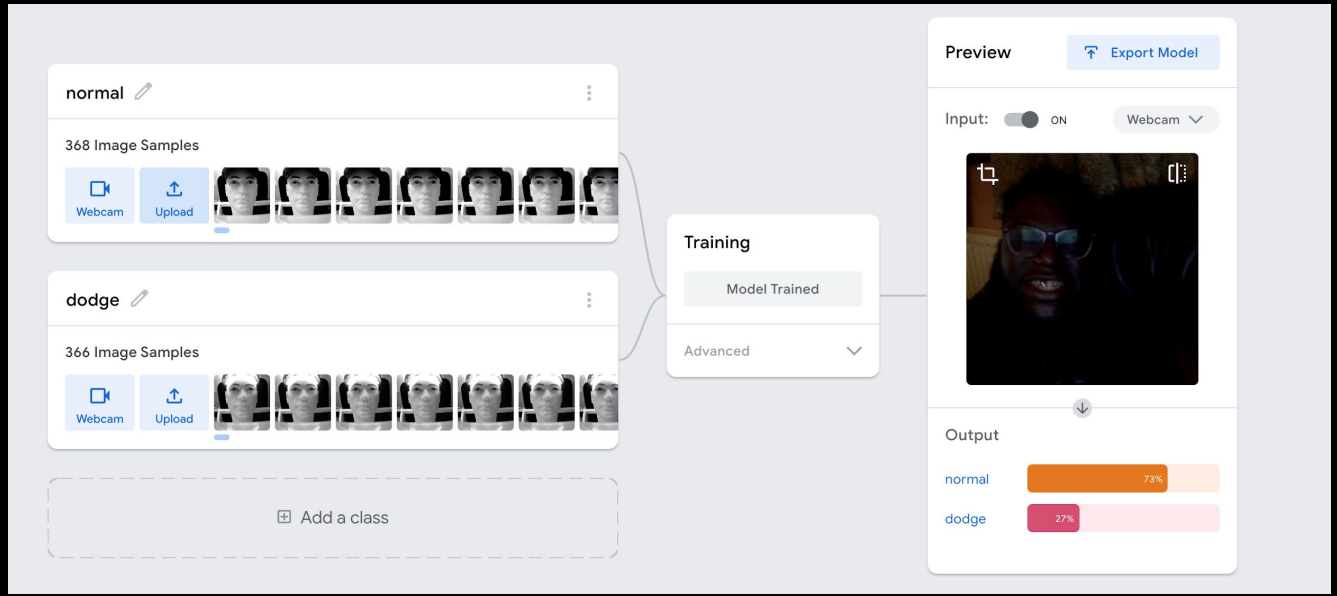


Data sets

A collection of related sets of information that is composed of separate elements.



https://teachablemachine.withgoogle.com/



https://teachablemachine.withgoogle.com/

Data in ML is split into three different sets

1. Training Data **(Our Focus)**
2. Validation Data
3. Test Data

Training Data

Datasets that you use to teach your ML model which outcomes correspond to which inputs.

(could be images, videos, text, audio and more)

Datasets formats we will focus on

1: JSON

JSON stands for JavaScript Object Notation
JSON uses JavaScript syntax, but the JSON format is text only.

2: CSV (Tabular Data)

CSV stands for comma-separated values.
Stores database/spreadsheet data

Supervised Learning

Training data is **labeled**.

EXAMPLES OF LABELED DATA

USER-CREATED HASHTAGS

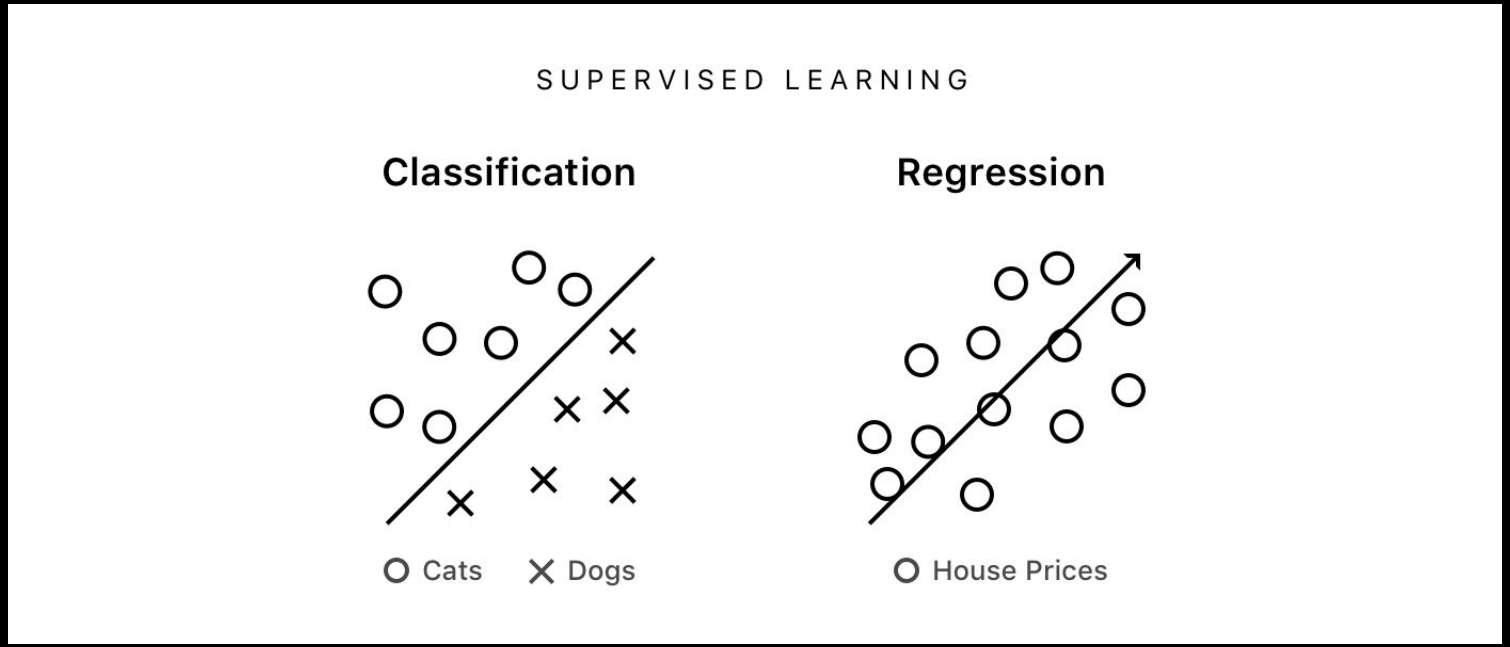
#burger #fries #food #happy

The hastags are the *labeled data*

Year Built	Bedrooms	Bathrooms	Sq. Ft.	Price
1901	3	1	1,800	**$200,000**
1995	4	3	2,500	**$350,000**
1980	2	1	1,300	**$150,000**
1922	5	3	1,900	**$400,000**
1950	3	2	2,200	**$220,000**

House prices are the *labeled data*

Source - https://uxdesign.cc/an-intro-to-machine-learning-for-designers-5c74ba100257



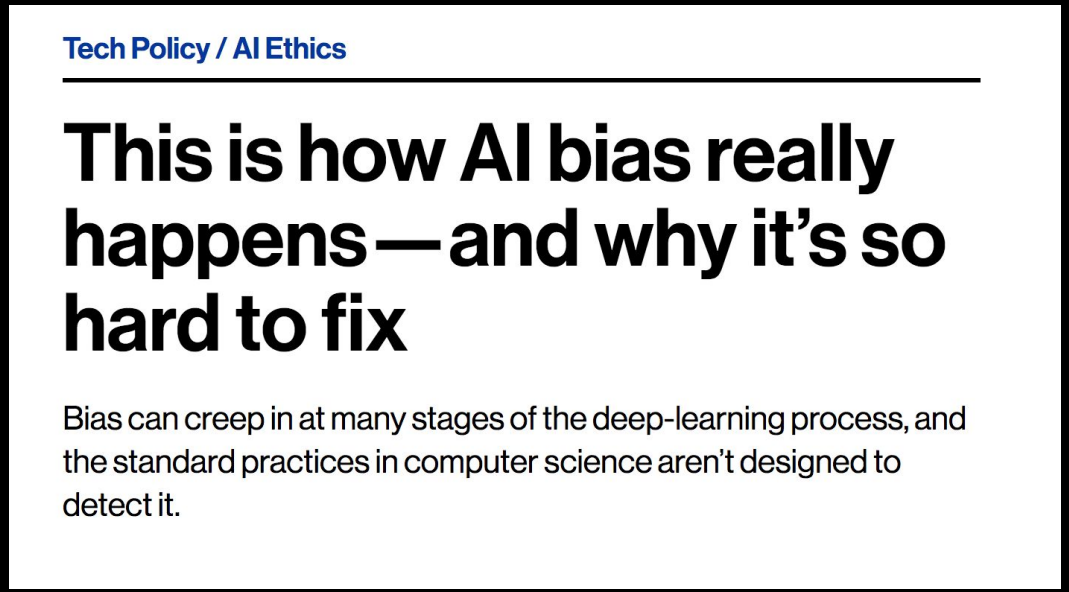
Classification - predict a list of tags or values

Regression - predict a number

The training data you source or collect, and how those data are labeled, directly determines the output of your system — and the quality of the user experience.

That’s where bias can creep in.

Bias occurs in AI systems when they reflect human biases held by the people involved in coding, collecting, selecting, or using data to train the algorithms that power the AI



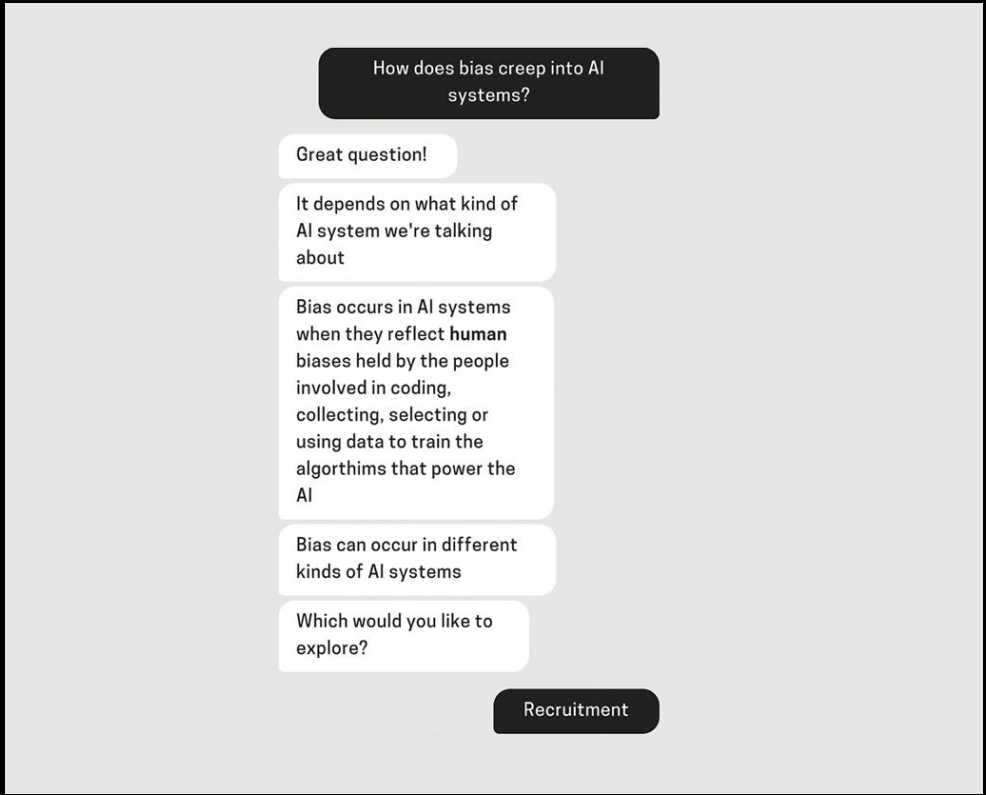
Tech Policy / AI Ethics

This is how AI bias really happens—and why it's so hard to fix

Bias can creep in at many stages of the deep-learning process, and the standard practices in computer science aren't designed to detect it.

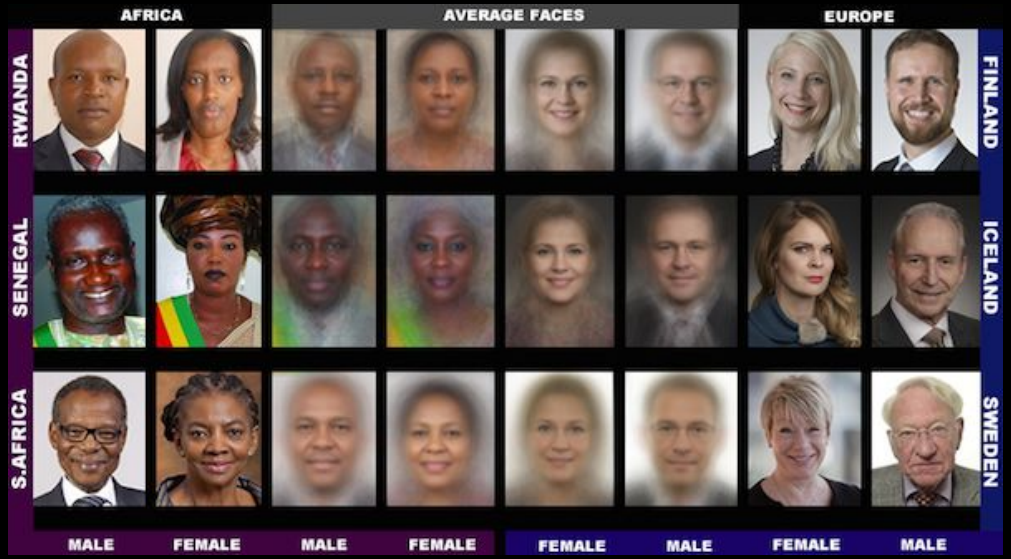
This is how AI bias really happens - and why it's so hard to fix.

Karen Hao



F’xα

Feminist Internet & COMUZI



Gender Shades

Joy Buolamwini & Timnit Gebru

Feminist Dataset

Caroline Sinders

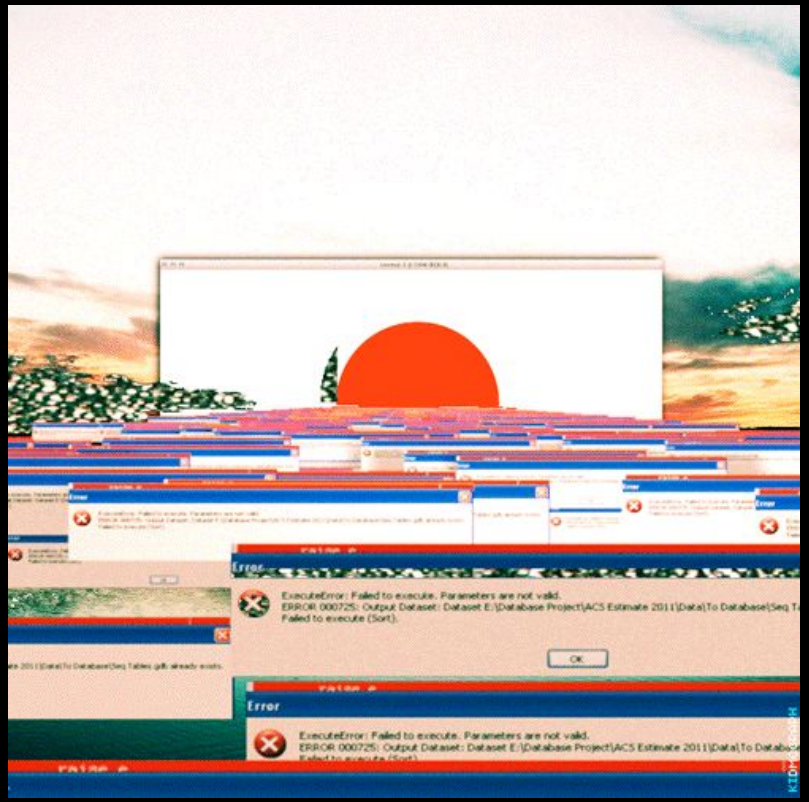
ual: creative computing institute

Suggestions have made that decision-support systems powered by AI can be used to augment human judgement and reduce both conscious and unconscious biases (Anderson & Anderson, 2007). However, machine learning data, algorithms, and other design choices that shape AI systems may reflect and amplify existing cultural prejudices and inequalities (Sweeney, 2013). While, historian of technology Melvin Kranzberg (1986) constructed the viewpoint that technology is regarded as neutral or impartial.

Racial Bias and Gender Bias Examples in AI systems

Alex Fefegha

Let’s go find some data sets.



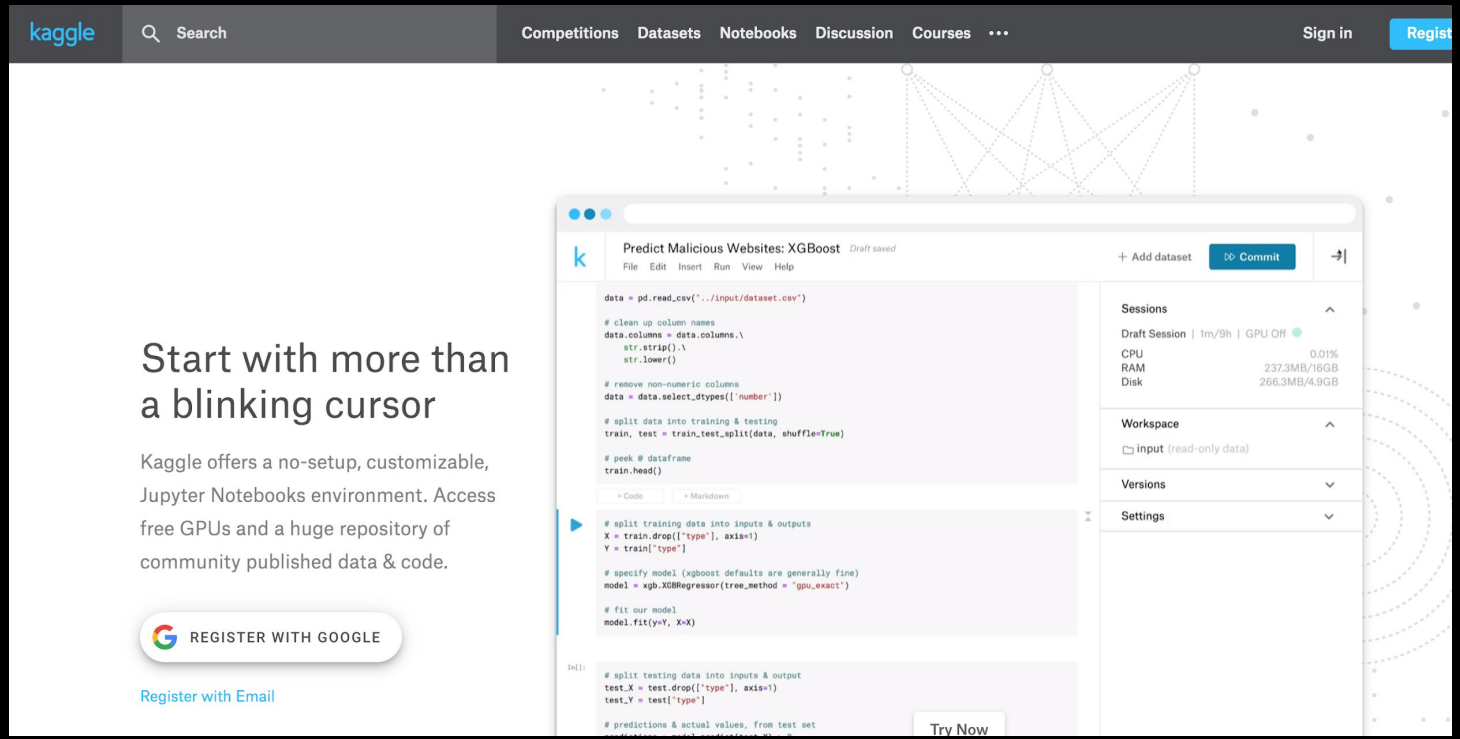
Exercise

Find or make a dataset in your group!

Talk about it

Make a slide deck

What is (are!) the data?
What format is the data in? (CSV, JSON, PDF, or . .)
What are the dimensions of the data (rows and columns)?
What are the "variables" (also known as "data items").
(In a CSV these would be the column headings. Do you recognize the data types (numbers, strings, images, etc.)?)
Is there missing, incorrect, or otherwise problematic data?
How and why was this data collected?
For whom is this data accurate or useful? What is this data unrepresentative of? (Who is missing and left out of the data?)
Knowing what you know now about machine learning, what will a model trained on this data help you do? Are there are alternative (non-machine learning) methods you could use instead?



https://www.kaggle.com/

List of datasets

https://www.kaggle.com/se18m502/bee-hive-metrics

https://www.kaggle.com/qqfiddler/scotus-opinions

https://www.kaggle.com/PromptCloudHQ/world-happiness-report-2019

https://www.kaggle.com/brittabettendorf/berlin-airbnb-data

https://www.kaggle.com/zynicide/wine-reviews#winemag-data-130k-v2.csv

https://www.kaggle.com/metmuseum/the-met

https://www.kaggle.com/usforestservice/usfs-fia

https://www.kaggle.com/datasnaek/youtube-new

http://www.shieldsgenealogy.com/recipe-nyc-dl.html

https://www.kaggle.com/new-york-city/nyc-dog-names

https://www.kaggle.com/destring/metacritic-reviewed-games-since-2000

https://www.kaggle.com/zaeemnalla/premier-league#stats.csv

https://www.kaggle.com/usforestservice/usfs-fia

https://www.kaggle.com/datasnaek/youtube-new

http://www.shieldsgenealogy.com/recipe-nyc-dl.html

https://www.kaggle.com/new-york-city/nyc-dog-names

https://www.kaggle.com/AnalyzeBoston/crimes-in-boston

https://www.kaggle.com/ruiqurm/lianjia

https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and#aac_shelter_outcomes.csv

https://www.kaggle.com/lishuyangkaggle/cocktails-hotaling-co

https://www.kaggle.com/epa/epa-historical-air-quality

https://www.kaggle.com/aashita/nyt-comments#ArticlesFeb2018.csv

https://www.kaggle.com/nycparks/tree-census

Present.

Homework!

Think about how machine learning can assist you in your project for your Creative Practice: Visual Coding and Physical Computing Class.

Think about what are the data sets you need?

Do you need to create them from the blank canvas or you could source data sets from elsewhere?

A 3 MINUTE PRESENTATION NEXT WEEK

Class done.
You are free!