Announcements

1. Contest due next Saturday at noon. Submit via Gradescope

2. Final exam next Thursday at 12:15p. Review will go out on Wednesday.

3. PSet co-learn party Tuesday at 7p

4. Design future CS109! Friday during class (first, try a co-learn session)

5. Chris Office Hours all 1:1 today and next Monday.
How AI is impacting our lives?

- Smartphones
- Social Media Platforms
- E-Commerce
- Autonomous Vehicles
- Security & Surveillance
- Navigation
- Banking & Finance Sector
- Smart Home
REMEMBER, WITH GREAT POWER COMES GREAT RESPONSIBILITY
We live in a time with real work to be done…

Can we use the affordances of ML to help?

- Access to high quality education
- Smart grids
- Story telling
- Better healthcare
Facebook slammed by UN for its role in Myanmar genocide

**Bitcoin Devours More Electricity Than Many Countries**

Annual electricity consumption in comparison (in TWh)

- **China**: 6,453
- **USA**: 3,990
- **Germany**: 524
- **All the world’s data centers**: 205
- **Bitcoin**: 143
- **Norway**: 124
- **Bangladesh**: 71
- **Switzerland**: 56
- **Google**: 12
- **Facebook**: 5

* Bitcoin figure as of May 05, 2021. Country values are from 2019. Sources: Cambridge Centre for Alternative Finance, Visual Capitalist.
Learning Goals

1. Understand limits in fairness through unawareness

2. Know two ways to measure fairness

3. Know some techniques to mitigate fairness issues
Other learning goal: how to be a responsible scientist (and not show up in the news in a negative way)
New Concepts from philosophy / ethics

- What is a Protected demographic?
- Distributive Harm vs Quality of Service Harm
- What is fairness?
  - Philosophy of procedural vs distributive
  - Different definitions of fairness
Part 0: Review
Machine Learning

$\mathbf{X}$  
(inputs)

$\mathcal{M}$  
(model)

$\hat{y}$  
(prediction)
## Classification Algorithms

<table>
<thead>
<tr>
<th>Heart 1</th>
<th>ROI 1</th>
<th>ROI 2</th>
<th>ROI m</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Heart 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Heart n</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

```
...
```

![Heart Image](image-url)
The Training / Testing Paradigm

Dataset

Training

Testing

Learn your parameters

Make sure that they work

Deployment

If your model passes testing...
Classification Algorithms

\[
\text{argmax } P(y|x) \quad y = \{0,1\}
\]

\[
\hat{y} = 0
\]

Making a prediction…
Logistic Regression

\[
P(Y = 1|x) = \sigma\left(\sum_i \theta_i x_i\right)
\]
Logistic regression is trying to fit a **line** that separates data instances where $y = 1$ from those where $y = 0$.

We call such data (or the functions generating the data) "**linearly separable**".

Naïve Bayes is linear too as there is no interaction between different features.

Turns out Logistic Regression is a Linear Classifier

\[ \theta^T x = 0 \]
\[ \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0 \]
Deep Learning: Logistic Regression Can Be Stacked

A neuron

Your brain

Actually, it’s probably someone else’s brain
Logistic Regression for Image Classification

http://scs.ryerson.ca/~aharley/vis/conv/
Works for any number of layers
1 Trillion Artificial Neurons
GoogLeNet Brain

22 layers deep

Multiple, Multi class output
The Cat Neuron

Le, et al., *Building high-level features using large-scale unsupervised learning*, ICML 2012.

Top stimuli from the test set

Optimal stimulus by numerical optimization
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012
22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012
smoothhound, smoothhound shark, *Mustelus mustelus*
American smooth dogfish, *Mustelus canis*
Florida smoothhound, *Mustelus norrisi*
whitetip shark, reef whitetip shark, *Triaenodon obesus*
Atlantic spiny dogfish, *Squalus acanthias*
Pacific spiny dogfish, *Squalus suckleyi*
hammerhead, hammerhead shark
smooth hammerhead, *Sphyrna zygaena*
smalleye hammerhead, *Sphyrna tudes*
shovelhead, bonnethead, bonnet shark, *Sphyrna tiburo*
angel shark, angelfish, *Squatina squatina*, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, *Pristis pectinatus*
guitarfish
*rougtail stingray, Dasyatis centroura*
butterfly ray
eagle ray
spotted eagle ray, spotted ray, *Aetobatus narinari*
cownose ray, cow-nosed ray, *Rhinoptera bonasus*
manta, manta ray, devilfish
*Atlantic manta, Manta birostris*
devil ray, *Mobula hypostoma*
gray skate, gray skate, *Raja batis*
little skate, *Raja erinacea*

22,000 is a lot
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

0.005%  1.5%  ?

Random guess  Pre Neural Networks  GoogLeNet
0.005%  Random guess  1.5%  Pre Neural Networks  43.9%  GoogLeNet

Szegedy et al, Going Deeper With Convolutions, CVPR 2015
Part 1: Framework of Harm
Quality of Service Harms

Quality-of-service harms
Occur when a system does not work as well for one person as it does for another

Examples:
- Generative Art
- Face Recognition
- Document Search
- Product Recommendation
Distributive Harms

Quality-of-service harms
Occur when a system does not work as well for one person as it does for another

Examples:
- Generative Art
- Face Recognition
- Document Search
- Product Recommendation

Distributive harms
Occur when AI systems extend or withhold opportunities, resources, or information

Examples:
- Hiring
- Lending
- School admissions
### Existential Harms?

<table>
<thead>
<tr>
<th>Quality-of-service harms</th>
<th>Distributive harms</th>
<th>Existential harms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occur when a system does not work as well for one person as it does for another</td>
<td>Occur when AI systems extend or withhold opportunities, resources, or information</td>
<td>Maybe you will just break the whole damn thing</td>
</tr>
</tbody>
</table>

#### Examples:
- Generative Art
- Face Recognition
- Document Search
- Product Recommendation
- Hiring
- Lending
- School admissions
- Genocide?
- Democracy?
- Climate?
- AI Supremacy?
Sticky Challenges

- Existential harms
- Procedural harms

Magnitude of Human Impact

How hard is it to reason about?
Warning: Many of the examples in this lecture come from the US. Much of ML fairness research has been done in the US (AFAIK).

Is that a meta bias?

Every community has axis of discrimination – framing goes beyond the US and beyond racism, sexism.
Part 2: Detecting Hidden Bias
There may be a problem...
How did this happen?
ImageNet classification

22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Trienodon obesus
Atlantic spiny dogfish, Squalus acantrias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sargassum shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish

Moonfish, Torpedo, Manta, Devilfish

Roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish

Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea

...
ImageNet classification challenge

- 22,000 categories
- 14,000,000 images
- 1000 categories
- 1,200,000 images in train set
- 200,000 images in test set

Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

Smooth hammerhead, Sphyra zygaena
Smalleye hammerhead, Sphyra tudes
Shovelnose shark, Sphyra tiburo
Angel shark, Squatina squatina
Monkfish, Lophius piscatorius
Electric ray, Torpedo marmorata
Sawfish, Pristis pectinatus
Guitarfish, Rhinobatos barbatus
Roughtail stingray, Dasyatis centroura
Butterfly ray, Eupomatus annulatus
Eagle ray, Aetobatus narinari
Cow-nosed ray, Rhinoptera bonasus
Manta ray, Manta birostris
Atlantic manta, Manta birostris
Devil ray, Mobula hypostoma
Gray skate, Raja erinacea
Little skate, Raja erinacea

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012
Biases in ImageNet

Imagenet is biased (in a neutral sense) towards texture ...
Biases in ImageNet

Imagenet is biased (in a neutral sense) towards texture ...

Hendrycks et. al. 2020
Machine Learning

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Training Data

New Data

Prediction Function $\theta^*$

Predictions
Machine Learning

Real World Problem

Model the problem

Formal Model $\theta$

Training Data

Learning Algorithm

New Data

Prediction Function $\theta^*$

Predictions
Logistic regression is trying to fit a line that separates data instances where $y = 1$ from those where $y = 0$.

We call such data (or the functions generating the data) "linearly separable".

Naïve bayes is linear too as there is no interaction between different features.
Classification of the minority group may be worse.
Classification of the minority group may be worse.
Classification of the minority group may be worse ... even with “awareness” or “stereotyping.”
Problem 1: Undersampling & Lack of Data

For both gender and race, the majority groups are often undersampled in image databases.

Majority of images in some databases of faces are of white faces.

Faces In The Wild database was 83.5% white and 77.5% male.
Huge Improvement in Face Datasets in 2018

Research and activism by Joy Buolamwini, Timnit Gebru, and many others has led to more representative datasets already.
Another Case Study
Algorithmic Discrimination: The Case of St. George’s Hospital

2,500 applicants to the medical school

Interview approx. 625 (so $\frac{3}{4}$ are rejected)

Offer spots to approx. 425 (so 70% of interviewees accepted)
Algorithmic Discrimination: The Case of St. George’s Hospital

In 1979, Vice Dean Dr. Geoffrey Franglen finishes a classification algorithm to do the job.

2,500 applicants to the medical school

Interview approx. 625 (so \(\frac{3}{4}\) are rejected)

Offer spots to approx. 425 (so 70% of interviewees accepted)
Timeline of a Biased Algorithm

1982: Dr. Franglen argues that 90-95% of classifications agree with the verdict of human assessors on the selection panel.

1982: Algorithm trained on historical data from St. George’s screens all applications.


Internal review questions why applicants are being weighted by factors like name and place of birth.

Commission finds that name and place of birth are used to dock points from female and “Non-Caucasian” applicants.
A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, a computing professional should carefully assess relevant aspects of the situation.
This biased result was predictable

Costs: At least 60 people wrongly rejected each year.

1. Garbage In, Garbage Out.
   Previous admissions process was biased against female applicants and applicants of color. Simply learning from the data will replicate and perpetuate the past bias.

2. Improper use of “Sensitive Features.”
   Algorithm relied on data like name and place of birth that provide no information about the merit of the applicant and are highly correlated with sensitive categories like race and gender.

3. Can be biased without intention to be evil
   Even if you didn’t mean to make a biased algorithm, that doesn’t mean it isn’t biased.
Can we get formal about fairness?
Three Formal Definitions of Fairness

Fairness through Unawareness
Fairness through Awareness: Independence
Fairness through Awareness: Separation
Motivating idea: “The way to stop discrimination on the basis of race is to stop discriminating on the basis of race” – Chief Justice Roberts

Note: Fairness through unawareness of some federally “protected categories” (subset of sensitive features) is legally required in domains like lending.

How to do it:
1. Exclude the sensitive feature (race, gender, age, etc) from your dataset
2. (Recommended) Also exclude proxies for the sensitive feature (name, zip code)
Protected Demographics

Protected Groups
Protected groups under EEO are race, color, national origin, religion, age, sex (gender), sexual orientation, physical or mental disability, and reprisal.

Equal Employment Opportunity, USA

Similarly defined for housing, loans, etc
Case Study: Facebook Ads & Job/Housing Recommendations

Facebook creates “Lookalike” feature for advertisers: upload a “source list” and find users with “common qualities” to target ads, including for housing and jobs

March 2018: National Fair Housing Alliance (NFHA) & other civil rights groups sue Facebook over violations of the Fair Housing Act

March 2019: As part of settlement, Facebook agrees not to use “age, gender, relationship status, religious views, school, political views, interested in, or zip code” in creating lookalike audience
Facebook Input Lookalikes

Create a Lookalike Audience

1. Select Your Lookalike Source
   - Select an existing audience or data source
   - Create New Source

2. Select Audience Location
   - Countries > North America
     - United States
   - Search for regions or countries

3. Select Audience Size
   - Number of lookalike audiences: 1

Create a Special Ad Audience

1. Select Your Source
   - Select an existing audience or data source

2. Select Audience Location
   - Countries > North America
     - United States
   - Search for regions or countries

3. Select Audience Size
   - Number of Special Ad Audiences: 1

Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% lookalike consists of the people in lookalike source. Increasing the percentage creates a bigger, broader audience.

Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% Special Ad Audience consists of the most similar online behavior to your source. Increasing the percentage creates a bigger, broader audience.
New “Special Ad” Audiences Still Biased

Gender: Equally Biased
Age: Almost as Biased
Race: more difficult to measure given the tools provided but still somewhat biased
Political Views: Less Biased

Sapiezynski et. al 2019,

Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite the lack of gender being provided as an input to Special Ad Audiences.

Yo, Piotr, you got your axis backwards 😊
Many Features = Accurate Group Prediction

Sensitive attributes are often “redundantly encoded” in the dataset
Many of the features or datapoints are correlated with the sensitive attribute
Two Philosophic Values of Fairness

**Procedural Fairness:**
Focuses on the decision-making or classification process, ensures that the algorithm does not rely on unfair features.

**Distributive Fairness:**
Focuses on the decision-making or classification outcome, ensures that the distribution of good and bad outcomes is equitable.
Two Philosophic Values of Fairness

**Procedural Fairness:**
Focuses on the decision-making or classification process, ensures that the algorithm does not rely on unfair features.

**Distributive Fairness:**
Focuses on the decision-making or classification outcome, ensures that the distribution of good and bad outcomes is equitable.

Fairness through unawareness (Facebook example shows this is hard)
Let’s Try Fairness Through Awareness!

Awareness of what?
## Fairness Through Awareness Terms

- **$D$:** protected demographic
- **$G$:** guess of your model (aka $y$ hat)
- **$T$:** the true value (aka $y$)

<table>
<thead>
<tr>
<th></th>
<th>$G = 0$</th>
<th>$G = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 0$</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>$T = 1$</td>
<td>0.07</td>
<td>0.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$G = 0$</th>
<th>$G = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 0$</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$T = 1$</td>
<td>0.02</td>
<td>0.08</td>
</tr>
</tbody>
</table>
**Parity**

**Fairness definition #1: Parity**
An algorithm satisfies “parity” if the probability that the algorithm makes a positive prediction ($G = 1$) is the same regardless of being conditioned on demographic variable.

$D$: protected demographic  
$G$: guess of your model (aka y hat)  
$T$: the true value (aka y)

$$P(G=1|D=1) = P(G = 1 | D = 0)$$
Calibration

**Fairness definition #2: Calibration**

An algorithm satisfies “calibration” if the probability that the algorithm is correct \((G = T)\) is the same regardless of demographics.

\[ D: \text{protected demographic} \]
\[ G: \text{guess of your model (aka y hat)} \]
\[ T: \text{the true value (aka y)} \]

\[ P(G = T | D = 0) = P(G = T | D = 1) \]
Calibration (Relaxed)

**Fairness definition #2: Calibration**
An algorithm satisfies “calibration” if the probability that the algorithm is correct \((G = T)\) is the same regardless of demographics.

\[
D: \text{protected demographic} \\
G: \text{guess of your model (aka y hat)} \\
T: \text{the true value (aka y)}
\]

\[
\frac{P(G = T|D = 1)}{P(G = T|D = 0)} \geq 1 - \epsilon
\]

Where \(\epsilon = 0.2\)

US legal standard: “disparate impact,” also known as the 80% rule.
COMPAS: Biased Against Black Inmates

Before: Compas is Biased
What does fairness through awareness fail to capture?

- If the classifier is significantly less good at identifying candidates e.g. for a surgery in a minority group (relative to the data), the candidates accepted might have worse outcomes, leading to future bias.

- Quality of Service Disparity might then lead to an Allocation Disparity.

- Dwork et. al. (including Omer Reingold!) call this a “self-fulfilling prophecy.”
Part 3: What are you going to do about it?
Balanced Training Data
Transparent Reporting
Model Card

- **Model Details.** Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model

- **Intended Use.** Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases

- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors

- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches

- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing

- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.

- **Quantitative Analyses**
  - Unitary results
  - Intersectional results

- **Ethical Considerations**

- **Caveats and Recommendations**
Train bias out
Advanced Idea: Adversarial Learning

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

Christina Wadsworth
Stanford University
Stanford, CA
cwads@cs.stanford.edu

Francesca Vera
Stanford University
Stanford, CA
fvera@cs.stanford.edu

Chris Piech
Stanford University
Stanford, CA
piech@cs.stanford.edu

Seniors at the time they wrote it
COMPAS: Predicting “Recidivism”

\[ \text{argmax } P(y|x) \quad y = \{0,1\} \]

Data about an inmate:
Their zip code, past crimes, etc

\[ \hat{y} = 0 \]

Will they commit a crime again

Was in use in California and Florida
COMPAS: Biased Against Black Inmates

Before: Compas is Biased

Accuracy Calibration Gap Parity Gap

Percentage Points
Can We Train Out Bias?

Model 1: Prediction

\[ \begin{align*}
X & \rightarrow \theta_1 & \hat{y} & \rightarrow \theta_2 & \text{Demographic}
\end{align*} \]

Model 1 should be accurate

Model 2 should be inaccurate

\[ \theta_1, \theta_2 = \arg\max_{\theta_1, \theta_2} L_1(\theta_1) - L_2(\theta_2) \]

*note in the paper these were neural nets*
Can We Train Out Bias?

**Before: Compas is Biased**

- **Accuracy**
- **Parity Gap**

**After: Gaps are reduced**

- **Accuracy**
- **Calibration Gap**
- **Parity Gap**
DON’T USE BLACK BOX ALGORITHMS TO MAKE RECIDIVISM PREDICTIONS
Use A Bayes Net?
Bayes Nets > Black Box?

![Diagram of a Bayesian network with nodes for Demographics, Conditions, and Symptoms, showing Age, Uni, ..., Gender, Cold, H1N1, Influenza, ..., Mono, Fever, Tired, Phlegm, ..., Runny Nose.]

Stanford University
Justice Beyond Distribution
Justice beyond Distribution

Zero-sum:
Resources and outcomes are fixed: the only task of justice is to fairly distribute them between individuals and groups. Improving the outcomes of the least-well-off group means worse outcomes for the best-off group (although in many cases only slightly worse).

Leveling Up & Expanding the Pie:
Outcomes and Resources are not fixed: justice means distributing outcomes fairly and increasing the number of good outcomes. Improving outcomes of the least-well-off group need not come at the expense of any other group.
Activism by Computer Scientists
Before
#TechWontBuildIt

Retail Polaroid cameras had only one flash button, but the ID-2, sold to the South African government, had a second “boost” flash which increased the illumination by 42% to better capture Black skin tones.

This was used to create passbook photographs for the Apartheid government.

http://physical-electrical-digital.nyufasedtech.com/items/show/46
Caroline Hunter: “I worked at Polaroid as a research chemist and my late husband Ken Williams was in the photo department producing advertisements for Polaroid, and one day I went to pick him up for lunch and we discovered an ID badge with a mockup of a black guy that we knew from Polaroid saying ‘Union of South Africa Department of the Mines’”

“We discovered that Polaroid was in South Africa and that they’d been there for quite some time, since 1938, and that they were actually the producers of the notorious passbook photographs which South Africans, black South Africans called their ‘handcuffs.’”
(Pedagogic Pause)
Learning Goals

1. Understand limits in fairness through unawareness
2. Know two ways to measure fairness
3. Know some techniques to mitigate fairness issues
Part 4: The Blind Spots
What are our current blind spots?

(Chris Opinion)
Well intentioned people can break things at scale (especially while moving fast)
Facebook Introduces Free Basics (2015)
Junta Starts a Misinformation Campaign Against Rohingya
Facebook: Two Moderators Who Speak Burmese (2015)
Genocide Against Rohingya Starts (2016)

Almost 1M Displaced
UN Concludes that Facebook Was Critical Component

Human Rights Council
Thirty-ninth session
10–28 September 2018
Agenda item 4
Human rights situations that require the Council’s attention

Report of the independent international fact-finding mission on Myanmar*

The role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet. Although improved in recent months, the response of Facebook has been slow and ineffective.

Silicon Valley’s impact beyond the US was a major blind spot
MOVE FAST AND BREAK THINGS
Aside: Facebook Says the Answer is Better ML

Integrity at Facebook

How we prioritise (NOW)

Reactive (User Reports)

Proactive (ML Classifiers)

Filtering, Ranking, De-duping

Human Review

Non-violating

Content Taken Down

Prioritized queue, with a mixture of reactive and proactive

High pri reports

Low pri reports
One Blind Spot I Want to Highlight
### Bitcoin Devours More Electricity Than Many Countries

Annual electricity consumption in comparison (in TWh)

<table>
<thead>
<tr>
<th>Country</th>
<th>Electricity Consumption (TWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>6,453</td>
</tr>
<tr>
<td>USA</td>
<td>3,990</td>
</tr>
<tr>
<td>Germany</td>
<td>524</td>
</tr>
<tr>
<td>All the world’s data centers</td>
<td>205</td>
</tr>
<tr>
<td>Bitcoin*</td>
<td>143</td>
</tr>
<tr>
<td>Norway</td>
<td>124</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>71</td>
</tr>
<tr>
<td>Switzerland</td>
<td>56</td>
</tr>
<tr>
<td>Google</td>
<td>12</td>
</tr>
<tr>
<td>Facebook</td>
<td>5</td>
</tr>
</tbody>
</table>

* Bitcoin figure as of May 05, 2021. Country values are from 2019.

Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

---

160,000,000,000,000 Hashes per second

But climate change and bitcoin isn’t even part of ethics at Stanford CS (I will update this slide once that changes)
It isn’t too hard to see the trend

We will most almost certainly hit 2x CO2 before 2060, and then blow past it.
The Whole Story is Filled with Uncertainties

Many things are uncertain
- Future Amount of CO2
- Climate Sensitivity
- Impact

But we can reason under uncertainty
We know the physics

https://youtu.be/3v-w8Cyfq8?t=39
Easy to Know Impacts Will Be Harsh

Without prompt, aggressive limits on CO₂ emissions, the Earth will likely warm by an average of 4°-5°C by the century’s end.

How Big a Change Is That?

In the coldest part of the last ice age, Earth’s average temperature was 4.5°C below the 20th century norm. Let’s call a 4.5°C difference one “Ice Age Unit.”

-2 IAU
Snowball Earth (-4 IAU)

-1 IAU
20,000 years ago

0
Average during modern times

+1 IAU
Where we’ll be in 86 years

+2 IAU
Cretaceous Hothouse
+200m sea level rise
No glaciers
Palm trees at the poles

My Neighborhood:
Half a mile of ice

My Neighborhood:
Hurricane

My Neighborhood:
?
Paleoclimate Gives us a Clue

Temperature of Planet Earth

PETM Video: https://www.youtube.com/watch?v=ldLBoErAhz4
Impacts are Here

Cyclone Idai
Impacted over 3M people
But Most Impacts are Far in Time and Space

Cyclone Idai, impacted over 3M people
It is hard to feel like you can do anything...

“I am just going to wait and see what happens”
Not really an ethical stance

“I am just going to wait and see what happens”

Hannah Arendt what is the problem with bureaucrats of Hitler’s empire?

Is this an ethical policy?
What can we do?
Push for some change

Individual  Community  Nation State

Is this our sweet spot?
Reduce CS “Pump” of Proof of Work

Bitcoin Devours More Electricity Than Many Countries

Annual electricity consumption in comparison (in TWh)

- China: 6,453
- USA: 3,990
- Germany: 524
- All the world's data centers: 205
- Bitcoin*: 143
- Norway: 124
- Bangladesh: 71
- Switzerland: 56
- Google: 12
- Facebook: 5

* Bitcoin figure as of May 05, 2021. Country values are from 2019. Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

160,000,000,000,000
Hashes per second

But climate change and bitcoin isn’t even part of ethics at Stanford CS (I will update this slide once that changes)
Advocate for a Clean Grid in CA

Figure 4. Total Renewable Generation Serving California Load by Resource Type

Source: California Energy Commission, staff analysis November 2018
Build?
Your Homework
Give yourself space to reflect on your own sense of what is right. And what you want for your own life’s work
Mencius Philosophy on Ethics

Mencius holds that all humans have innate but incipient tendencies toward benevolence, righteousness, wisdom, and propriety. Employing an agricultural metaphor, he refers to these tendencies as “sprouts” (2A6). The sprouts are manifested in cognitive and emotional reactions characteristic of the virtues.
Thank you!

Feel free to chat about this with Chris or with our Embedded Ethics instructor, Katie Creel
kcreel@stanford.edu