Learning Goals

• CT Building Block: Students will be able to explain examples of how computers do what they are programmed to do, rather than what their designers want them to do.

• CT Impact: Students will be able to list reasons that an algorithm might be biased and what its impact will be.

• CT Impact: students will be able to list arguments why a company should or should not change its algorithms based on “fairness”
Algorithms can be compared based on many things

So far we’ve considered:

• Whether they work right
• Time and space they take

But what about if they’re fair?
For some “unambiguous” tasks, like sorting, fairness is a non-issue

Example: Sorting cards:
• *Input*: pile of unsorted cards
• *Output*: pile of cards in sorted order from clubs, diamonds, hearts and spades, with ace's being highest

Example: Sorting flights:
• *Input*: list of flight options from A to B
• *Output*: list sorted by cost/departure time/arrival time/duration etc.
For other tasks, it’s not so clear what the right output is; there’s potential for bias.

Example: Classification tasks

- *Input*: individual's loan application (address, age, gender, credit rating...)
- *Output*: approve/deny a loan

- *Input*: digital image
- *Output*: cat/not a cat

- *Input*: genome sequence from cancerous biopsy tissue and success of treatment
- *Output*: proposed cancer treatments
How do classifiers work?

- Classifiers are derived from patterns or correlations from data.
- The data that classifiers learn the patterns has the “answer” – this data is called **training data**.
  - Some of the training data is held back to check and see if the classifier works. This is called **test data**.
- Classifiers then apply these patterns to new data with no “answer”.
- Example:
  - *Input*: digital image
  - *Output*: cat/not a cat
  - *Training data*: labeled images of cats and images that are not cats.
Example: Classification tasks

- **Input**: individual's loan application (address, age, gender, credit rating...)
- **Output**: approve/deny a loan
- **Training data**: list of loan applications, decisions made, and for those who were approved, whether they repaid the loan or not
Classification task training data example: cancer genes

- Input: genome sequence from cancerous biopsy tissue
- Output: Which cancer treatment is likely to work best
- Training data: labeled genome sequences and which treatments were successful from both cancerous tissue
That was pretty straightforward. But what if I stack the deck?

Setup:

- I have a hand of cards (not necessarily chosen randomly from the deck – it may be biased in some way, e.g., fewer 8’s than average).
- I remove a small number of cards from the hand at random to form the test data. Note that the test data is biased in the same way as the training data.
- Your task: use the remaining cards (on the projector) as training data to build a classifier.
What can this tell us about classifier “fairness”? 

- Suppose that cards classified as high-valued are “rewarded” (loan approved), while those classified as low-valued are “penalized” (loan denied)
- Is it fair if red cards are never rewarded, even though some are high-valued?
- This is a silly question, but it’s not hard to extrapolate to situations where the stakes are higher…
Let’s look at a more complex example: loan applications (from Hardt et al. at Google)

- The bank makes $300 on a successful loan, but loses $700 on a default
- Training data of historical applicants describes the applicant’s credit rating and are labeled as either successful or defaulters
- Light blue are the defaulters, dark blue are successful
Loan application example

Classification task: approve or deny a loan application, based on credit threshold

Group exercise: choose a threshold (credit rating) at which to approve/deny loans and define why you chose that threshold

Light blue are the defaulters, dark blue are successful

Computational Thinking
http://www.ugrad.cs.ubc.ca/~cs100

https://research.google.com/bigpicture/attacking-discrimination-in-ml/
Loan application threshold #1: 50

Threshold Decision

credit rating

0 10 20 30 40 50 60 70 80 90 100

loan threshold: 50

Outcome

Correct 84% loans granted to paying applicants and denied to defaulters

Incorrect 16% loans denied to paying applicants and granted to defaulters

True Positive Rate 83% percentage of paying applications getting loans

Positive Rate 48% percentage of all applications getting loans

Profit: 14800
Loan application threshold #2: 54

Threshold Decision

<table>
<thead>
<tr>
<th>credit rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

- loan threshold: 54

Outcome

Correct 83% loans granted to paying applicants and denied to defaulters

Incorrect 17% loans denied to paying applicants and granted to defaulters

True Positive Rate 75% percentage of paying applications getting loans

Positive Rate 41% percentage of all applications getting loans

Profit: 16600
Changing the problem: there are two groups of people – blue and orange

- Each group has the same # of dots
- Each group has half defaulters/half successful
- Only the distributions are different

Credit rating

[Graph showing credit rating distributions for blue and orange populations]
Loan application example: Consider both populations together

Classification task: approve or deny a loan application, based on credit threshold and/or colour
Let's talk about bias. There are two main ones involved.

- **Conscious bias** is when you're biased and you know it (and you're generally not sorry)
- **Unconscious bias** is when you're biased... and you may not know it (and if you do, you're sorry)... and you may even be biased against what you believe!

An example of unconscious bias

- [http://wwest.mech.ubc.ca/diversity/unconscious-bias/](http://wwest.mech.ubc.ca/diversity/unconscious-bias/)

Test this on yourself

http://www.understandingprejudice.org/iat/

Seriously, test yourself at some point.
### Unconscious bias on gender and work

<table>
<thead>
<tr>
<th>Test Result</th>
<th>% of Test Takers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong association between male and career</td>
<td>40%</td>
</tr>
<tr>
<td>Moderate association between male and career</td>
<td>15%</td>
</tr>
<tr>
<td>Slight association between male and career</td>
<td>12%</td>
</tr>
<tr>
<td>Little or no gender association with career or family</td>
<td>17%</td>
</tr>
<tr>
<td>Slight association between female and career</td>
<td>6%</td>
</tr>
<tr>
<td>Moderate association between female and career</td>
<td>5%</td>
</tr>
<tr>
<td>Strong association between female and career</td>
<td>5%</td>
</tr>
</tbody>
</table>
### Unconscious bias on race

<table>
<thead>
<tr>
<th>Test Result</th>
<th>% of Test Takers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong automatic preference for White people</td>
<td>48%</td>
</tr>
<tr>
<td>Moderate automatic preference for White people</td>
<td>13%</td>
</tr>
<tr>
<td>Slight automatic preference for White people</td>
<td>12%</td>
</tr>
<tr>
<td>Little or no automatic preference</td>
<td>12%</td>
</tr>
<tr>
<td>Slight automatic preference for Black people</td>
<td>6%</td>
</tr>
<tr>
<td>Moderate automatic preference for Black people</td>
<td>4%</td>
</tr>
<tr>
<td>Strong automatic preference for Black people</td>
<td>6%</td>
</tr>
</tbody>
</table>

If your test results showed a preference for a certain group, you may have a hidden, or unconscious, bias in favor of that group. The results of more than one million tests suggest that most people have unconscious biases. For example, nearly two out of three white Americans show a moderate or strong bias toward, or preference for, whites, as do nearly half of all black Americans.
Google search and fake news

Business Insider

Obama signs a nationwide order

About 1,640,000 results (0.60 seconds)

FAKE NEWS

Obama Signs Executive Order Banning The National Anthem At All ...
cnn.com.de › News
11 Nov 2016 - President Obama has signed an Executive Order banning the National Anthem with fines and jail time beginning December 1st of this year.

FAKE NEWS

Obama Signs Executive Order Banning The Pledge Of ... - ABC News
abcnews.com.co › News
President Obama signs an Executive Order banning the Pledge of Allegiance in schools nationwide
President Obama, seen here signing an Executive Order ...

REAL NEWS

Obama Signs Executive Order Banning the Pledge of ... - Snopes
www.snopes.com/pledge-of-allegiance-ban/
16 Aug 2016 - Unimaginative fake news publishers have recycled an old hoax about President Obama's banning the Pledge of Allegiance. ... Claim: President Obama has issued an executive order banning the Pledge of Allegiance in U.S. schools. ... Early this morning, President Obama made what could very ...

REAL NEWS

Obama Did Not Ban the Pledge - FactCheck.org
www.factcheck.org/2016/09/obama-did-not-ban-the-pledge/
2 Sep 2016 - Q: Did President Obama sign an executive order banning the Pledge of ... Order Banning The Pledge Of Allegiance In Schools Nationwide.

Computational Thinking
http://www.ugrad.cs.ubc.ca/~cs100

http://uk.businessinsider.com/google-algorithm-change-fake-news-rankbrain-2016-12
Learning Goals

• CT Building Block: Students will be able to explain examples of how computers do what they are programmed to do, rather than what their designers want them to do.

• CT Impact: Students will be able to list reasons that an algorithm might be biased and what its impact will be.

• CT Impact: students will be able to list arguments why a company should or should not change its algorithms based on “fairness”