QA: Different strokes for different systems.

Michael Hilton

Learning goals

- Understand challenges for QA of ML systems
- Be able to test assumptions about the data
- Understand and choose different fairness approaches
- Detect changes in the model over time

Administrivia

- Homework 4D due today
- Homework 5 released soon
- Review retrospective answers
- Visitor on Thursday

What should we start doing?

- Issues with slack (QA via google docs?
- Unclear Homework questions
- Recitation Recordings
- Bigger breakout rooms
- Direct instruction on decision making process
- Docker and Kubernetes via lecture
- Piazza
- WSL ☹
- More TA's
- More technical work
- More bingo-like activities
- More Technical office hours

What should we stop doing?

- Talking about candy we would have had without the pandemic ☹
- Google Slides DDOS, creating a document is stressful
- Less break out rooms
- More break out rooms
- Some people feel the grading is too harsh
- Homework 3 was too short (not enough time)

What should we keep doing?

- Having in class questions. Makes me have a reason to pay extra attention in the lecture.
- Show more of the cat. :-)
- Guest speakers
- Slack
- Informal polls
- Breakout rooms
- Support of non/US East students
- Banter

QA for ML

What does it mean to do QA for a ML System?

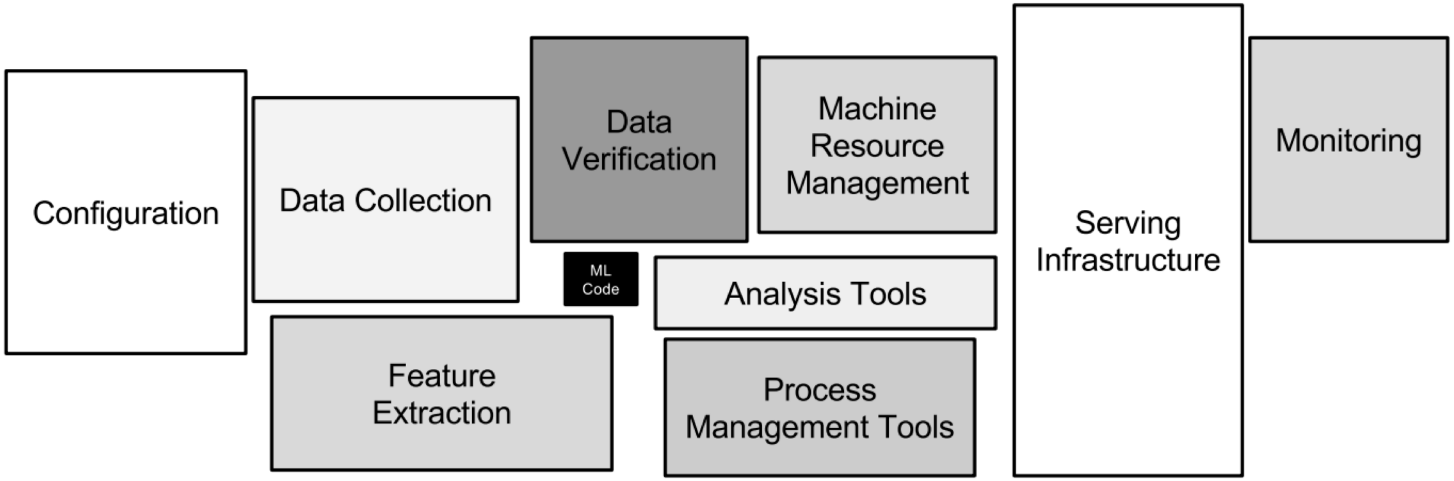


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Broad considerations when testing ML

- Data debugging, validation, and testing
- Model debugging, validation, and testing
- Service debugging, validation, and testing
  - Traditionally testing, Design docs, already covered

Data Debugging

- Validate Input Data Using a Data Schema
  - For your feature data, understand the range and distribution. For categorical features, understand the set of possible values.
  - Encode your understanding into rules defined in the schema.
  - Test your data against the data schema.
- Test Engineered Data: For example:
  - All numeric features are scaled, for example, between 0 and 1.
  - One-hot encoded vectors only contain a single 1 and N-1 zeroes.
  - Missing data is replaced by mean or default values.
  - Data distributions after transformation conform to expectations.
  - Outliers are handled, such as by scaling or clipping.

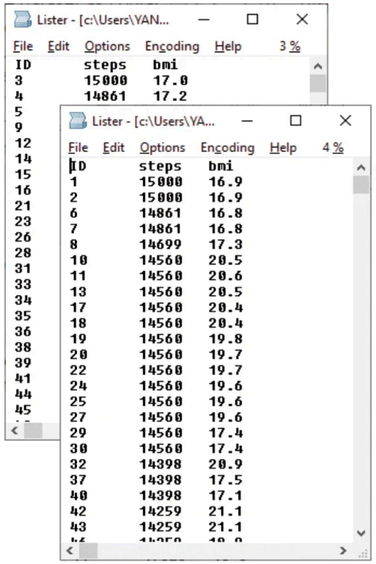
Data Debugging Cont…

- Is your data sampled in a way that represents your users (e.g., will be used for all ages, but you only have training data from senior citizens) and the real-world setting (e.g., will be used year-round, but you only have training data from the summer
- Training-serving skew—the difference between performance during training and performance during serving. During training, try to identify potential skews and work to address them. During evaluation, continue to try to get evaluation data that is as representative as possible of the deployed setting.
- Are any features in your model redundant or unnecessary? Use the simplest model possible.
- Data bias is another important consideration

https://ai.google/responsibilities/responsible-ai-practices/

Examine your data...

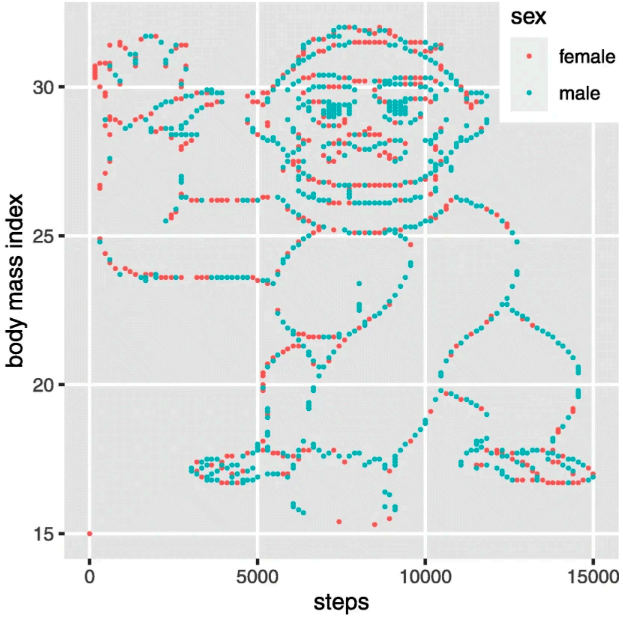
a


Lister - [c:\Users\YAN...
File Edit Options Encoding Help 3 %
Lister - [c:\Users\YA...
File Edit Options Encoding Help 4 %

```
ID        steps    bmi
3         15000    17.0
4         14861    17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

```
ID        steps    bmi
1         15000    16.9
2         15000    16.9
6         14861    16.8
7         14861    16.8
8         14699    17.3
10        14560    20.5
11        14560    20.6
13        14560    20.5
17        14560    20.4
18        14560    20.4
19        14560    19.8
20        14560    19.7
22        14560    19.7
24        14560    19.6
25        14560    19.6
27        14560    19.6
29        14560    17.4
30        14560    17.4
32        14398    20.9
37        14398    17.5
40        14398    17.1
42        14259    21.1
43        14259    21.1
```

b

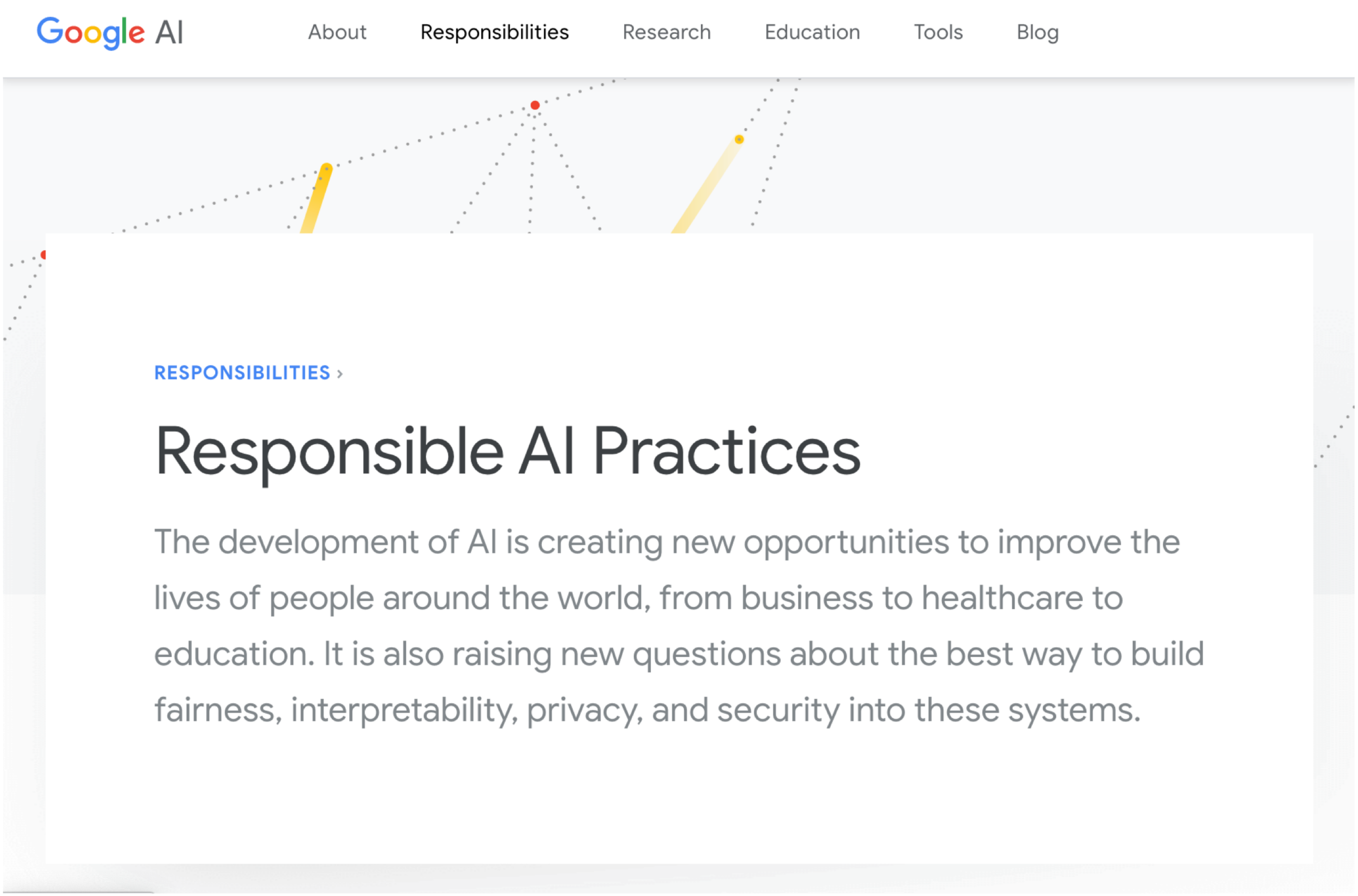


c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Model Debugging

- Check that the data can predict the labels.
  - Use 10 examples from your dataset that the model can easily learn from. Alternatively, use synthetic data.
- Establish a baseline
  - Use a linear model trained solely on most predictive feature
  - In classification, always predict the most common label
  - In regression, always predict the mean value

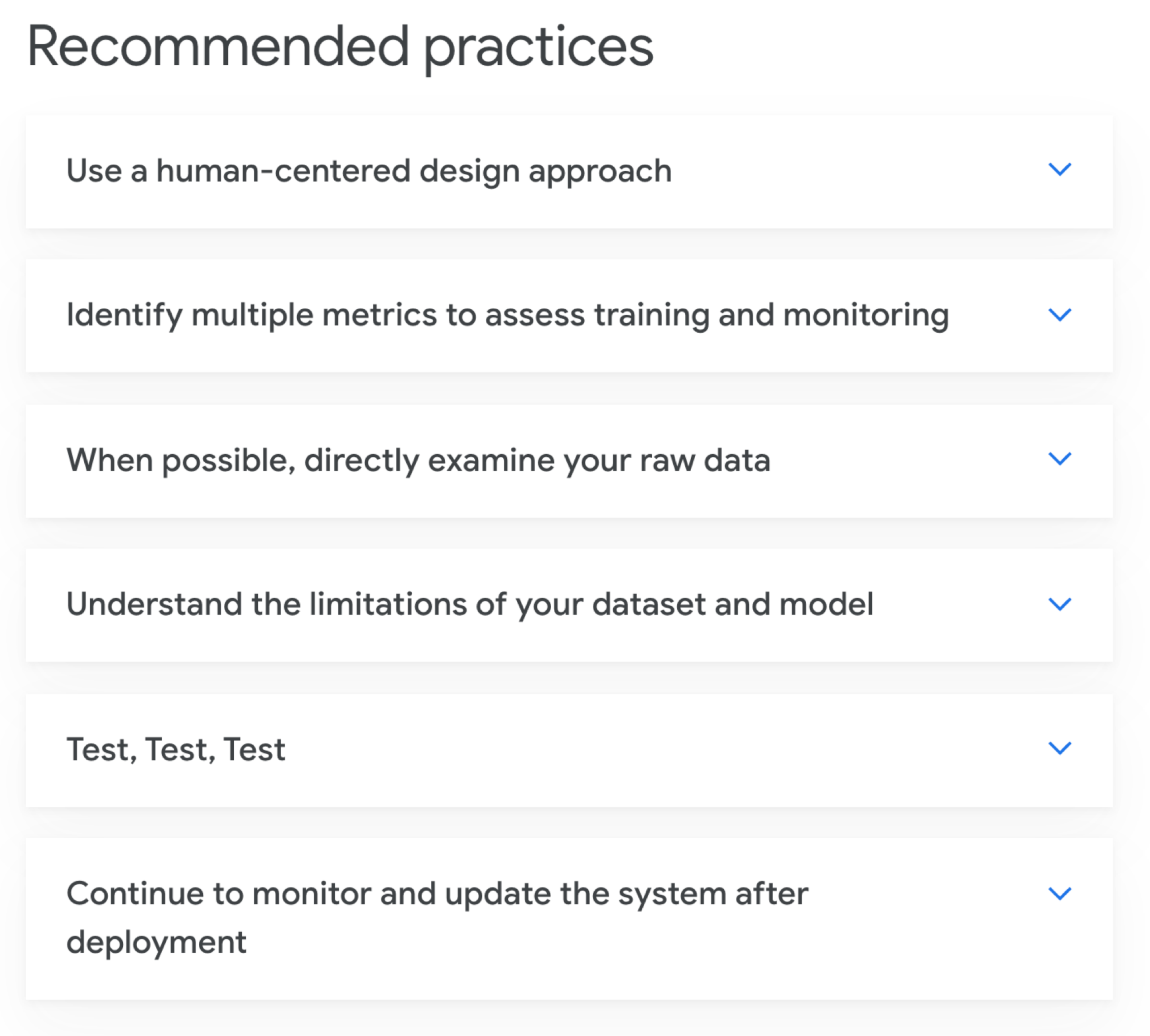


RESPONSIBILITIES ›

Responsible AI Practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build fairness, interpretability, privacy, and security into these systems.

https://ai.google/responsibilities/responsible-ai-practices/



Recommended practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

Test, Test, Test

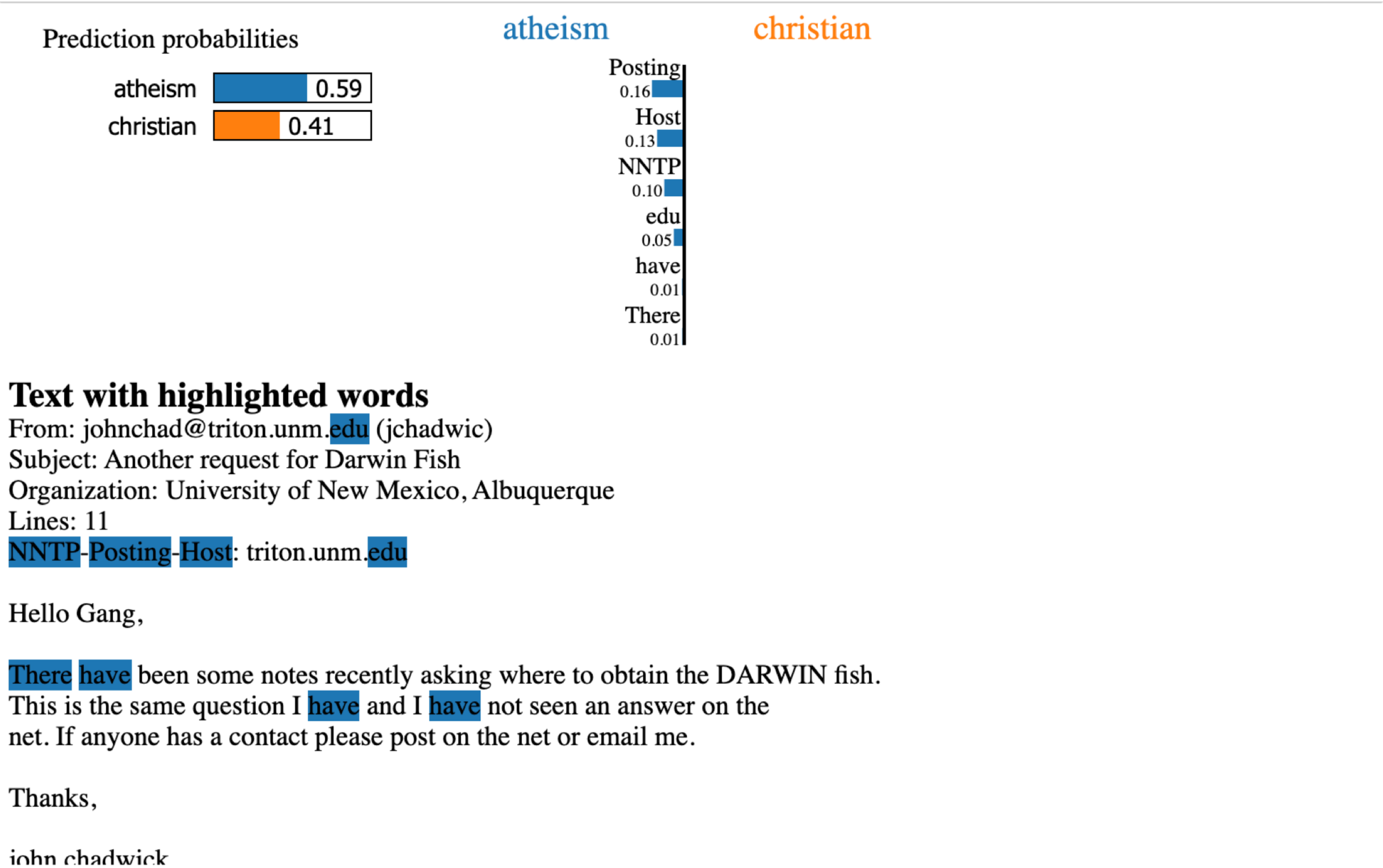
Learn from software engineering best test practices and quality engineering to make sure the AI system is working as intended and can be trusted.

- Conduct rigorous **unit tests** to test each component of the system in isolation.
- Conduct **integration tests** to understand how individual ML components interact with other parts of the overall system.
- Proactively detect **input drift** by testing the statistics of the inputs to the AI system to make sure they are not changing in unexpected ways.

Test, Test, Test - Continued

- Use a gold standard dataset to test the system and ensure that it **continues to behave as expected**. Update this test set regularly in line with changing users and use cases, and to reduce the likelihood of training on the test set.
- Conduct iterative user testing to incorporate a diverse set of users' needs in the development cycles.
- Apply the quality engineering principle of poka-yoke: build quality checks into a system, so that unintended failures either cannot happen or **trigger an immediate response** (e.g., if an important feature is unexpectedly missing, the AI system won't output a prediction).

Beyond Correctness....



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

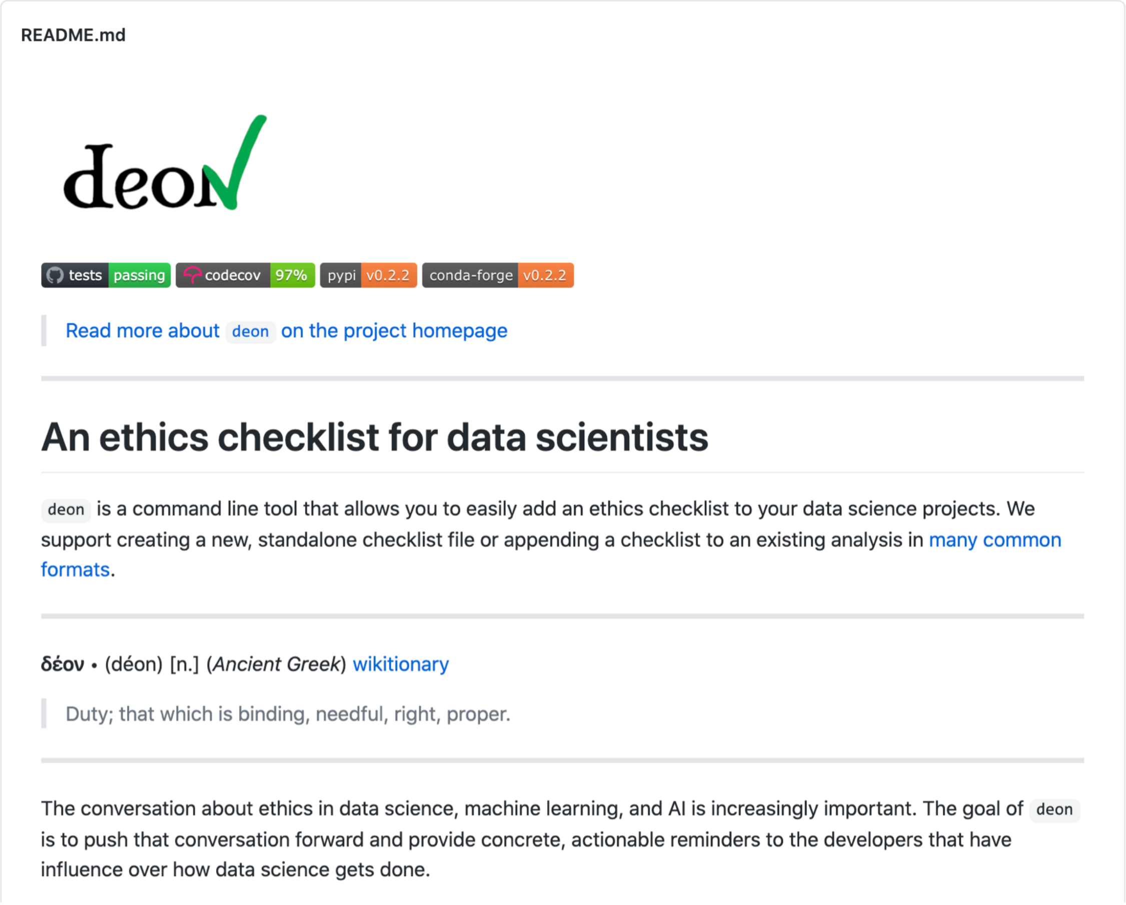
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick

https://github.com/drivendataorg/deon



> Read more about `deon` on the project homepage

An ethics checklist for data scientists

`deon` is a command line tool that allows you to easily add an ethics checklist to your data science projects. We support creating a new, standalone checklist file or appending a checklist to an existing analysis in many common formats.

δέον • (déon) [n.] (*Ancient Greek*) wikitionary

> Duty; that which is binding, needful, right, proper.

The conversation about ethics in data science, machine learning, and AI is increasingly important. The goal of `deon` is to push that conversation forward and provide concrete, actionable reminders to the developers that have influence over how data science gets done.

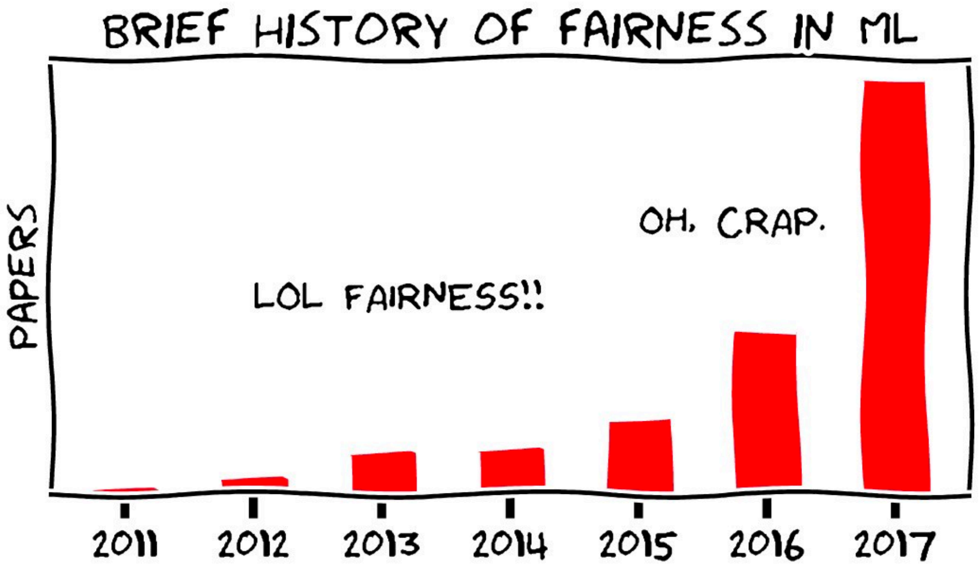
Where ML has gone wrong

- Data Collection
  - Facebook uses phone numbers provided for two-factor authentication to target users with ads.
  - Personal information on taxi drivers can be accessed in poorly anonymized taxi trips dataset released by New York City.
- Data Storage
  - FedEx exposes private information of thousands of customers after a legacy s3 server was left open without a password.
- Analysis
  - A widely used commercial algorithm in the healthcare industry underestimates the care needs of black patients, assigning them lower risk scores compared to equivalently sick white patients.
  - Strava heatmap of exercise routes reveals sensitive information on military bases and spy outposts.
  - Excel error in well-known economics paper undermines justification of austerity measures.
- Modeling
  - Google Photos tags two African-Americans as gorillas.
  - Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names.

https://deon.drivendata.org/examples/

ML Fairness

- Getting answers is the easy part... Asking the right questions is the hard part.

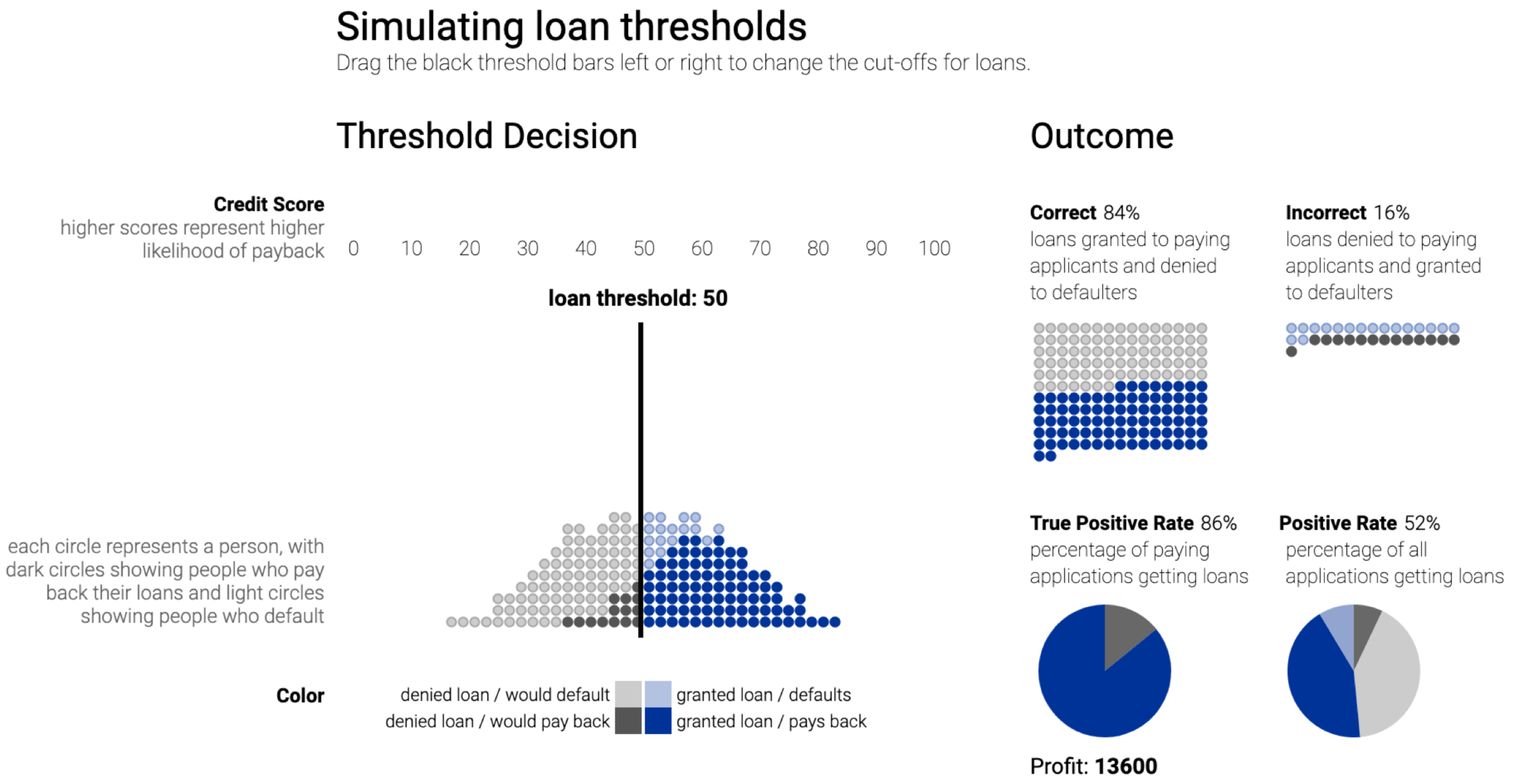


https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Varieties of fairness

- Group unaware
  - Ignore group data (one group could get excluded)
- Group thresholds
  - Different rules per group (rules differ by group)
- Demographic parity
  - Same percentage in pool as outcomes (might result in random selection)
- Equal opportunity
  - Equal chance out positive outcomes regardless of groups (focus on individual, rules differ per group)
- Equal accuracy
  - Equal chance of both outcomes per group (focus on group, rules differ per group)

Explainability



Simulating loan thresholds

Drag the black threshold bars left or right to change the cut-offs for loans.

Threshold Decision

Credit Score
higher scores represent higher likelihood of payback

each circle represents a person, with dark circles showing people who pay back their loans and light circles showing people who default

Color

denied loan / would default | granted loan / defaults
denied loan / would pay back | granted loan / pays back

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 86%
percentage of paying applications getting loans

Positive Rate 52%
percentage of all applications getting loans

Profit: **13600**

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Activity

Consider the different approaches to fairness. Can you come up with different scenarios where each fairness approach might be appropriate?

Remember the fairness approaches are:

- Group unaware
- Group thresholds
- Demographic parity
- Equal opportunity
- Equal accuracy