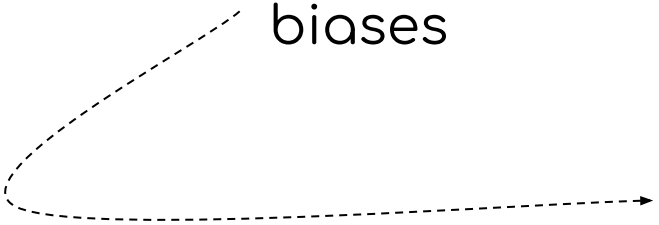
Machine Learning Ethics

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

When models are trained on historical data, predictions will perpetuate historical biases



Predictive Analysis Ethics

Dare Obasanjo
@Carnage4Life

Product leader at Microsoft. My team is responsible for advertiser experience for Bing Ads; mobile apps, web UX, desktop apps & SDKs.

Dare Obasanjo
@Carnage4Life

Follow

Machine learning algorithms are driven more by the training data than math. Give an algorithm biased data then results will be biased. E.g.
• Amazon's resumé referral algo which auto rejected women
• Search ads algo which showed background check ads for "black sounding names"

Ryan Saavedra @RealSaavedra
Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist

8:59 PM - 22 Jan 2019

https://twitter.com/carnage4life/status/1087892442349678592?s=21

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / A MONTH AGO

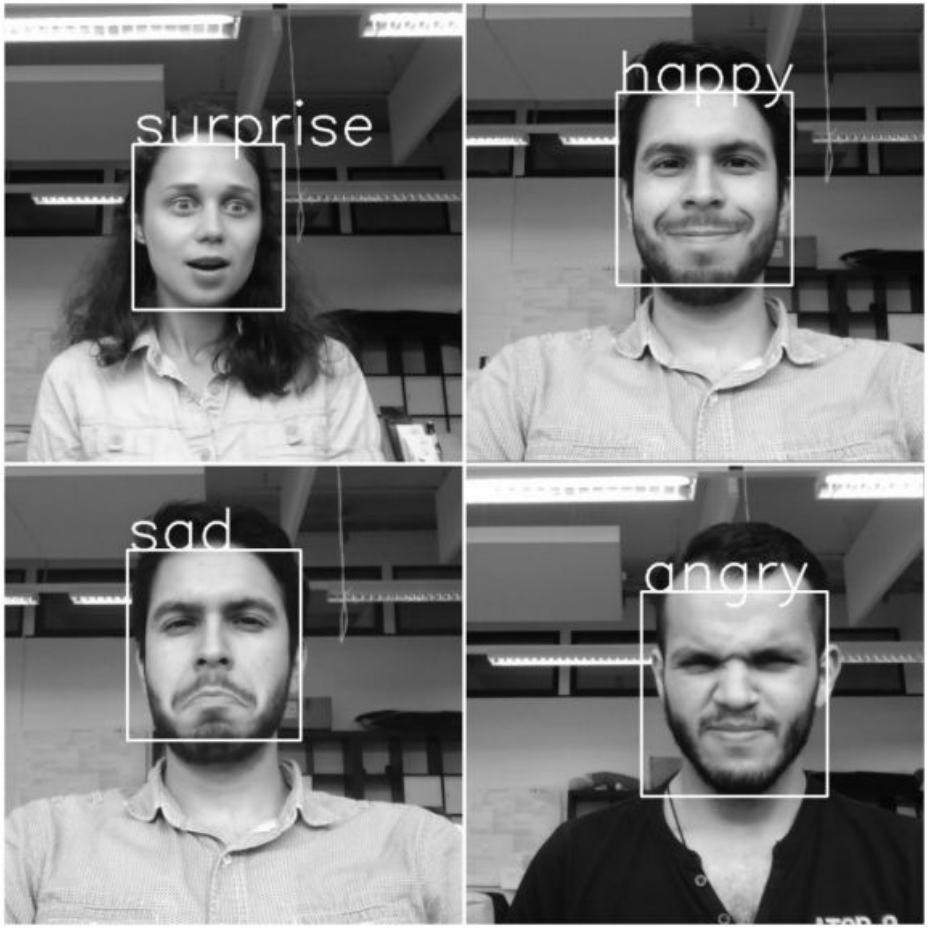
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G



https://github.com/oarriaga/face_classification

https://twitter.com/nke_ise/status/897756900753891328

What to do about bias...

1. Anticipate and plan for potential biases before model generation. Check for bias after.
2. Have diverse teams.
3. Use machine learning to improve lives rather than for punitive purposes.
4. Revisit your models. Update your algorithms.
5. You are responsible for the models you put out into the world, unintended consequences and all.

Discussed so far…

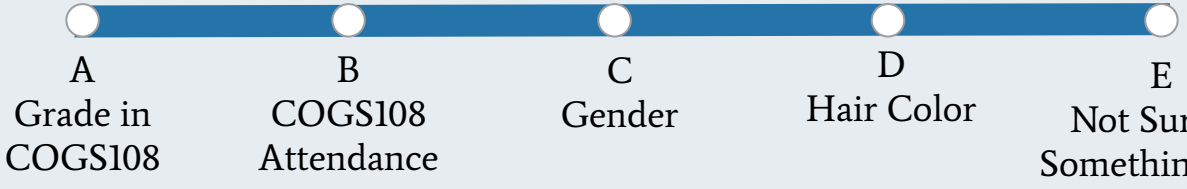
- data partitioning
- feature selection
- supervised & unsupervised machine learning
  - Continuous variables: regression (supervised) and dimensionality reduction (unsuperfied)
  - Categorical variables: classification (supervised; decision trees) or clustering (unsupervised)
- model assessment
  - Continuous: RMSE (& Accuracy)
  - Categorical: Accuracy, Sensitivity, Specificity, AUC
- biased data can & will lead to biased predictions

Data Science Question

Based on data I have about you all, can I predict who in this course will be successful?

Prediction Approach

Which would be the most predictive of your future success?



N = 254

train the model:
N = 178
(70% of the data)

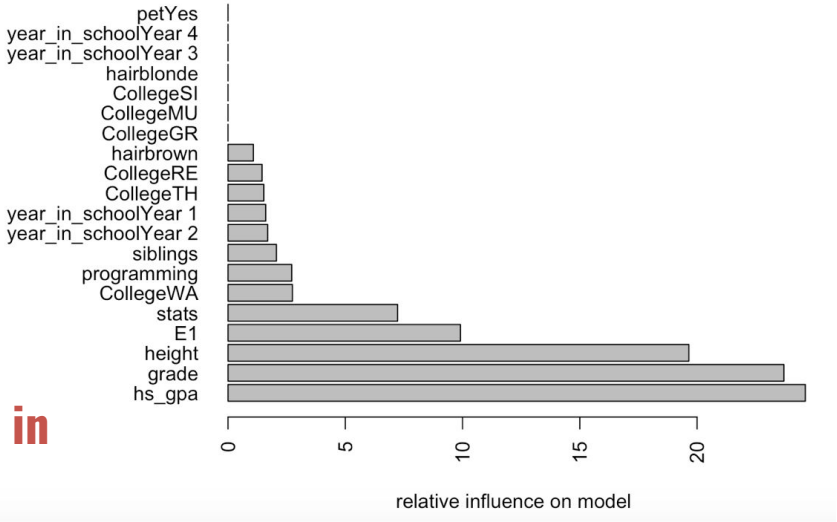
test the model:
N = 76
(30% of the data)

train the model

predicted success in test set

	Accuracy	Sensitivity	Specificity
training set	71.2%	76%	67%
test set	49.1%	40%	60%

Assess Prediction Model



N = 254

train the model:
N = 178
(70% of the data)

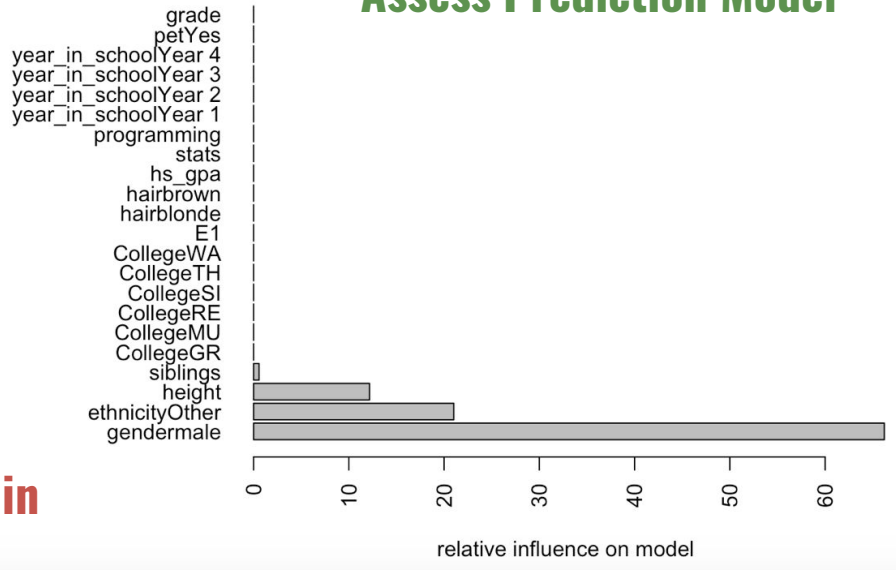
test the model:
N = 76
(30% of the data)

train the model

predicted success in test set

	Accuracy	Sensitivity	Specificity
training set	100%	100%	100%
test set	100%	100%	100%

Assess Prediction Model



What if I were using these data to determine who I should write recommendation letters for?

Or to determine which students I focus my attention on?

Or whose projects I read?

Or who I allow to come to office hours?

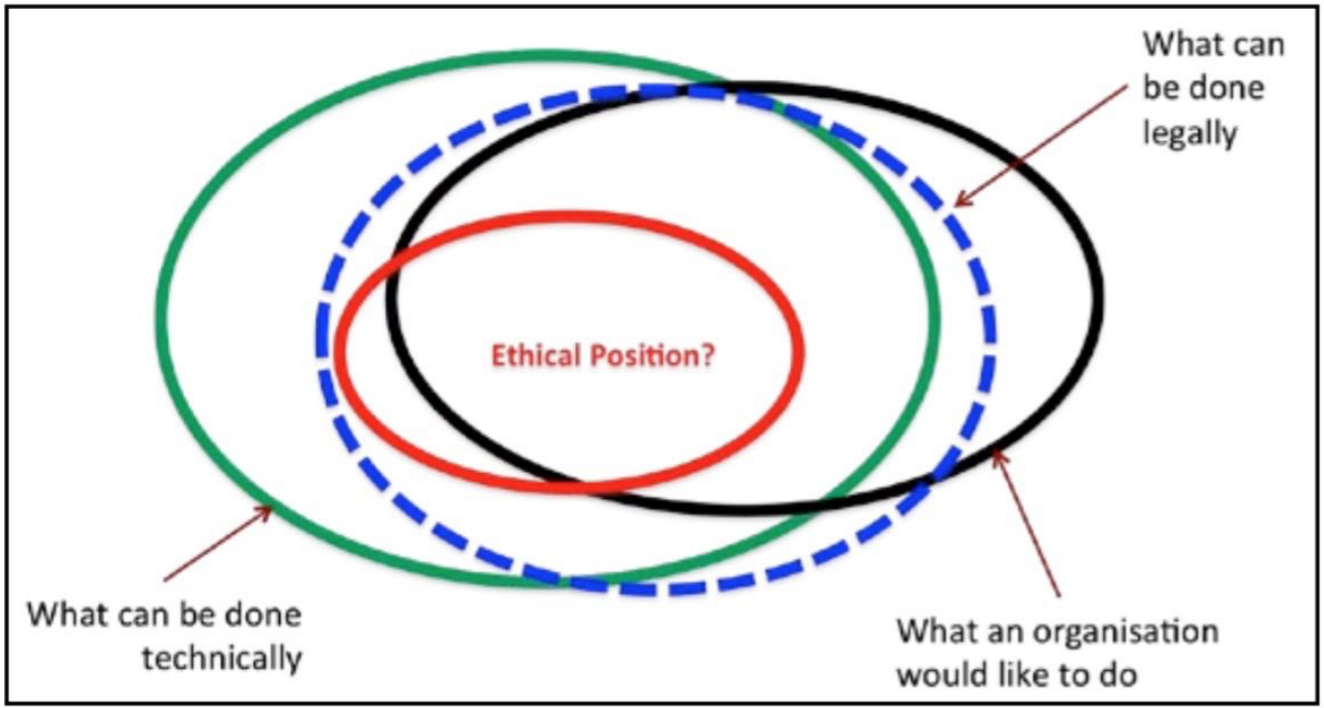
Or who UCSD allows to be data science majors?

What to do about bias...

1. Anticipate and plan for potential biases before model generation. Check for bias after.
2. Have diverse teams.
3. Use machine learning to improve lives rather than for punitive purposes.
4. Revisit your models. Update your algorithms.
5. You are responsible for the models you put out into the world, unintended consequences and all.

Think about whether the models you're building should even be built.

Big Data Ethics



Source: IBM

adapted from Brad Voytek

Predictive algorithms should (*at a minimum*) be FAT

Fair: lacking biases which create unfair and discriminatory outcomes

- For whom does this algorithm fail?
- Steps to take:
  1. Verify data about individual is correct
  2. Carry out “sensitivity test”

Accountable/Accurate: answerable to the people subject to them

- Correct data used? Is there a mechanism for appeal?

Transparent: open about how and why particular decisions were made

- Think *carefully* about what transparency is (Handing over source code likely isn’t the answer)

Sources: https://www.datacamp.com/community/podcast/weapons-math-destruction & https://ironholds.org/resources/papers/mulching.pdf

A Mulching Proposal

Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry

Os Keyes
Department of Human Centered Design & Engineering
University of Washington
Seattle, WA, USA
okeyes@uw.edu

Jevan Hutson
School of Law
University of Washington
Seattle, WA, USA
jevanh@uw.edu

Meredith Durbin
Department of Astronomy
University of Washington
Seattle, WA, USA
mdurbin@uw.edu

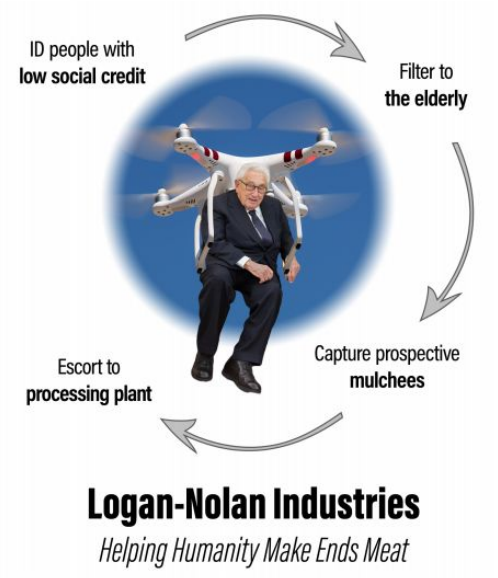
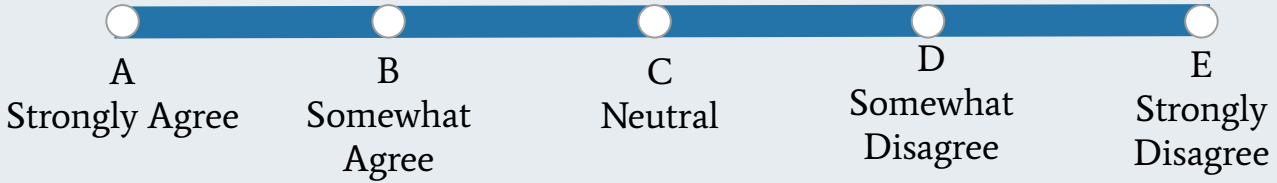


Figure 1: A publicity image for the project, produced by Logan-Nolan Industries

https://dl.acm.org/citation.cfm?id=3310433 (*CHI* 2019)

Prediction Thoughts

We should start using this algorithm to mulch up the elderly



A Mulching Proposal

FAIR - equally considers all elderly individuals

ACCURATE - pre- has mechanism for appeal; post - compensation

TRANSPARENT - website with all features; testable

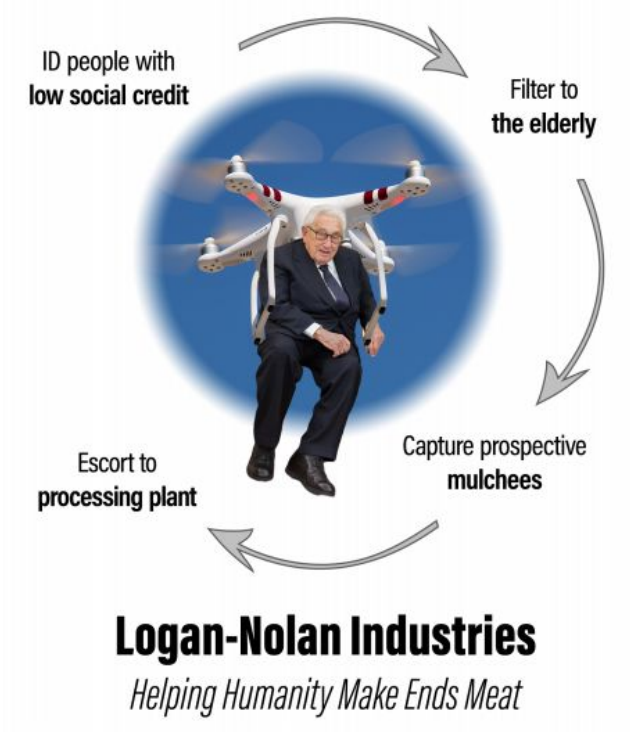


Figure 1: A publicity image for the project, produced by Logan-Nolan Industries

Checklists are helpful, but they’re not and excuse for thoughtlessness.