Algorithm FATE (Fairness, Accountability, Transparency, & Ethics)

Danna Gurari
University of Texas at Austin
Spring 2021

https://www.ischool.utexas.edu/~dannag/Courses/IntroToMachineLearning/CourseContent.html

Review

- Last week:
  - Machine Learning for Sequential Data
  - Recurrent Neural Networks (RNNs)
  - Training Deep Neural Networks: Hardware & Software
- Assignments (Canvas):
  - Project outline due tonight
  - Prototype of final project ML system due at tomorrow's meeting
  - Final project submission with video due in two weeks
- Questions?

Final Project Video Suggestions

- Video creation/editing resources:
  - https://docs.google.com/document/d/1jBZ1fU1CKDLw1y2ZVM5LvYHjv3iFZoAYPz947_0f2Bs/edit?usp=sharing
- Attributions:
  - Creative commons license generator: https://creativecommons.org/choose/

Plagiarism: Definition

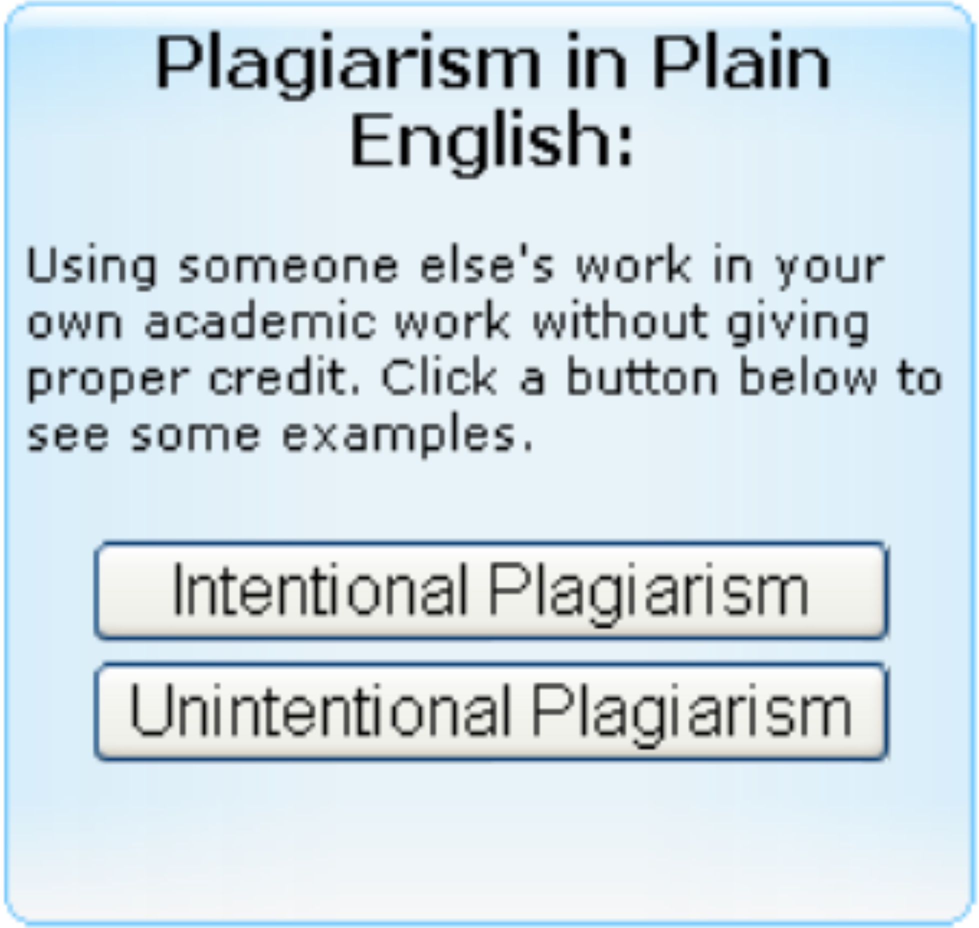
- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html

University of Texas Definition of Plagiarism:

"the appropriation of, buying, receiving as a gift, or obtaining by any means material that is attributable in whole or in part to another source, including words, ideas, illustrations, structure, computer code, and other expression or media, and presenting that material as one's own academic work being offered for credit."

Plagiarism: Definition

- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html



Plagiarism: Play It Safe, Give Credit Generously

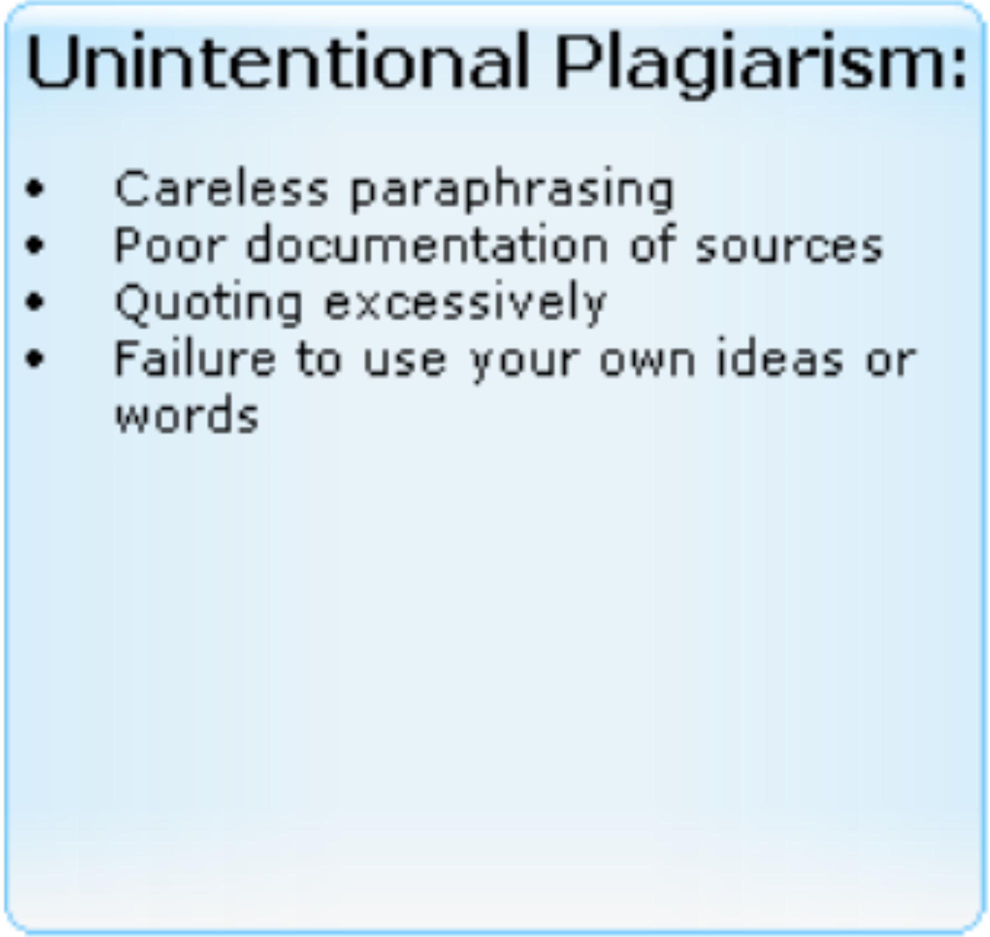
- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html

Intentional Plagiarism:

- Copying a friend's or classmate's work
- Buying or borrowing papers
- Cutting and pasting blocks of text without providing documentation of the original source
- Borrowing images and other media without documentation of the original source
- Publishing work on the Web without the permission of the creator

Plagiarism: Play It Safe, Give Credit Generously

- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html



Plagiarism: Play It Safe, Give Credit Generously

- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html

During the course of your research, you come across an idea that you use in your paper. You don't use the author's exact words or even paraphrase -- just the idea. Cite it?

Other people's words aren't the only thing you need to cite. You also need to cite ideas. So in this case, you should give the author credit for the idea by citing them.

Plagiarism: Play It Safe, Give Credit Generously

- Material from: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html

You are doing a presentation for your Chemistry class and use an image of the Periodic Table you found on a government web site. Cite it?

You should cite images. Even government websites in the public domain need to be cited.

Plagiarism: Play It Safe, Give Credit Generously

- What can happen if you are accused of plagiarism?
  - Redo assignment
  - Receive a failing grade
  - Be suspended
  - Be expelled

- What resources can help you to avoid plagiarism?
  - Review: https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html
  - Review: https://legacy.lib.utexas.edu/d7/sites/default/files/services/instruction/AvoidingPlagiarism_guide.pdf
  - Visit writing center: http://uwc.utexas.edu/

- Neither you (I believe) nor I have any desire to talk about plagiarism ☺
- Play it safe and give credit generously!!!

Give Credit Generously

- Idea: add credit page to your presentation for resources used
  - e.g., Microsoft Azure
  - e.g., freely-shared code/libraries
  - e.g., links to all images
  - …

Today’s Topics

- Machine Learning Algorithms that Discriminate
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

Today’s Topics

- Machine Learning Algorithms that Discriminate
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

Observation: World Population is Diverse

Image Source: https://www.rocketspace.com/corporate-innovation/why-diversity-and-inclusion-driving-innovation-is-a-matter-of-life-and-death

Algorithms Discriminate: Google Search

Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

Algorithms Discriminate: Google Search

A search for "Jew" returned many anti-Semitic web pages:

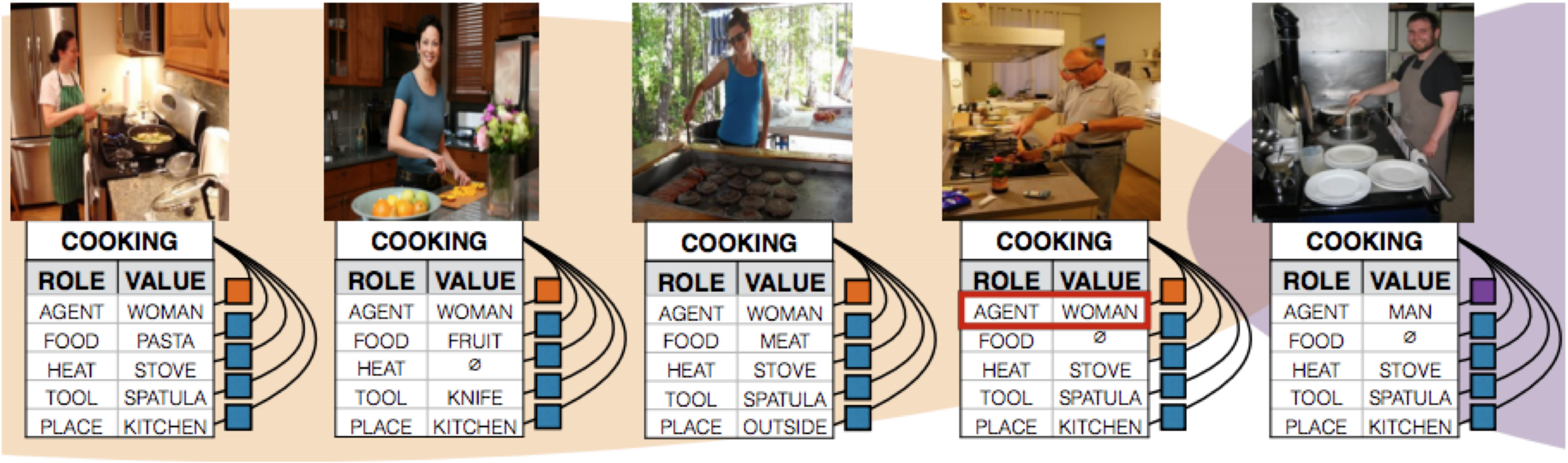
Safiya U. Noble ; [Algorithms of Oppression: How Search Engines Reinforce Racism]

Algorithms Discriminate: Image Tagging

Using Twitter to call out Google's algorithmic bias

https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

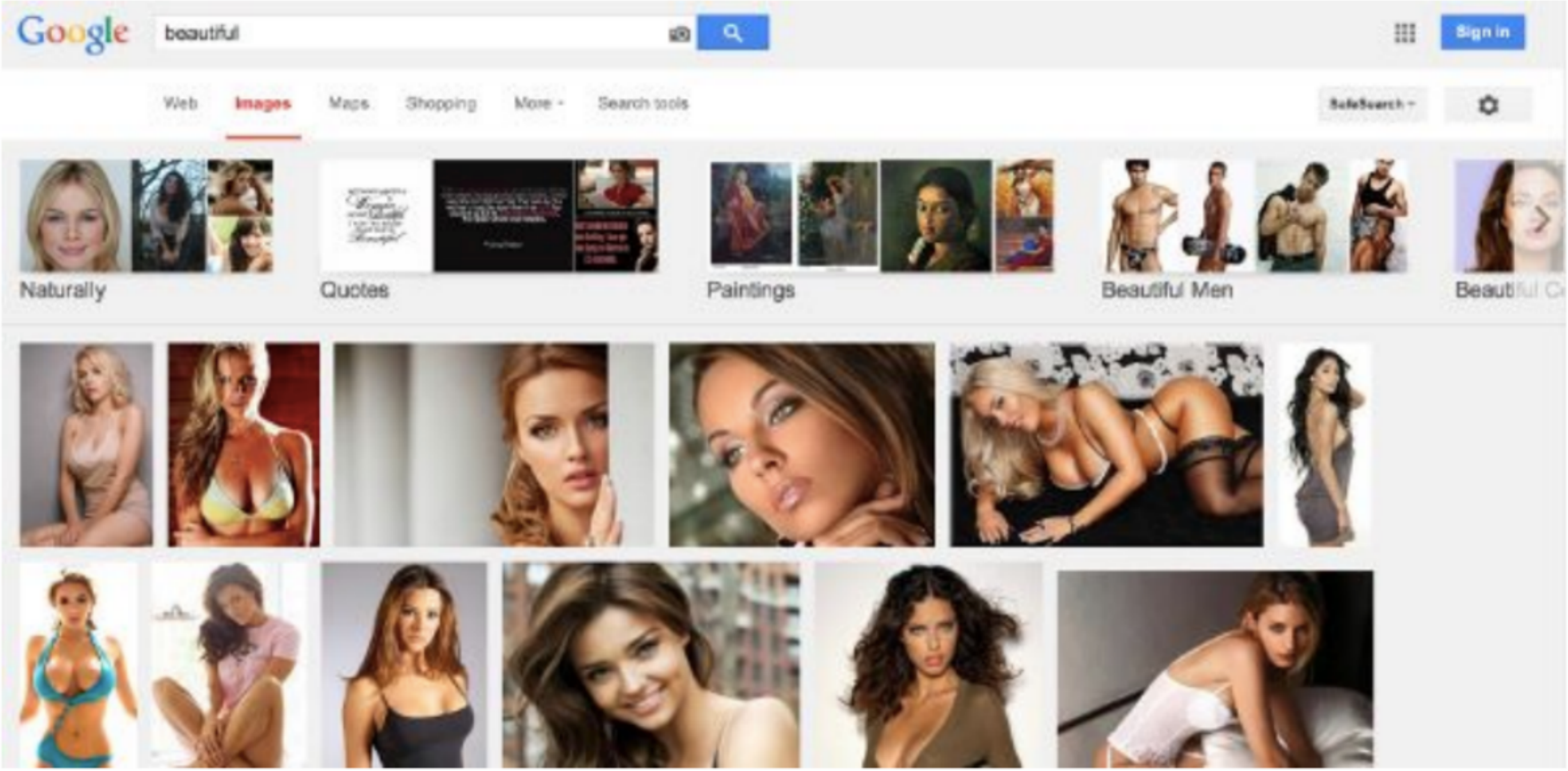
Algorithms Discriminate: Image Tagging



Algorithm identifies men in kitchens as women. Learned this example from given dataset. (Zhao, Wang, Yatskar, Ordonez, Chang, 2017)

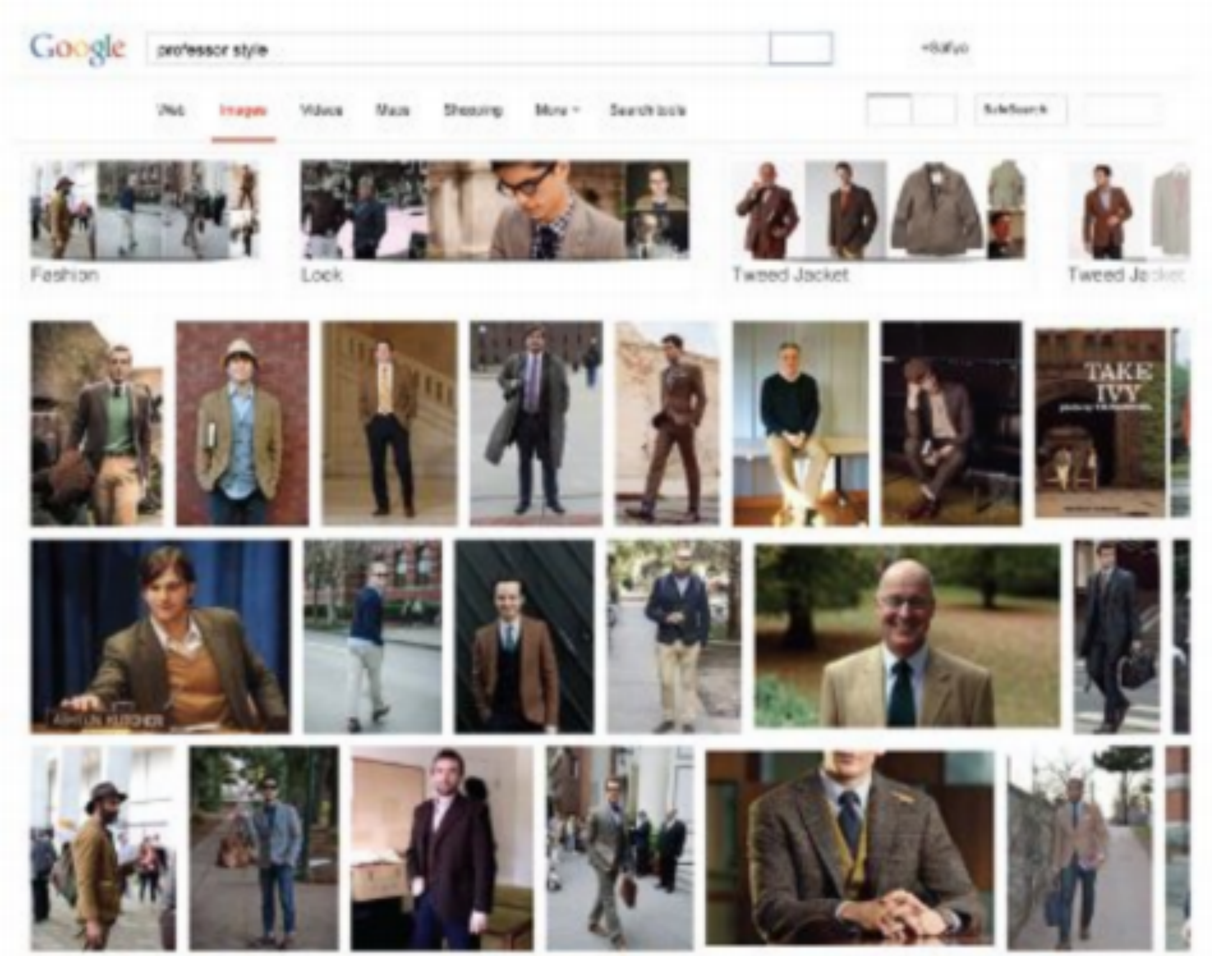
https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/ç

Algorithms Discriminate: Image Tagging ("beautiful"; 2014)



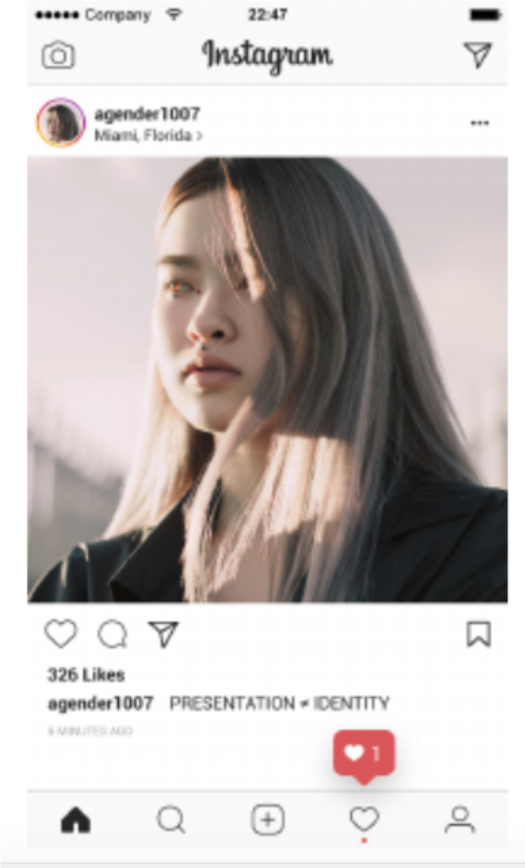
Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

Algorithms Discriminate: Image Tagging ("professor style"; 2014)



Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

Algorithms Discriminate: Image Tagging




```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

```
{
  "class": "woman",
  "score": 0.813,
  "type_hierarchy": "/person
  /female/woman"
},
{
  "class": "person",
  "score": 0.806
},
{
  "class": "young lady (heroine)",
  "score": 0.504,
  "type_hierarchy": "/person/female
  /woman/young lady (heroine)"
}
...
```


Person identifies as agender (gender-less, and so non-binary)

Morgan Klaus Scheurman, Jacob M. Paul, and Jed R. Brubaker, “How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services.” CSCW 2019.

Algorithms Discriminate: “Hotness” Photo-Editing Filter

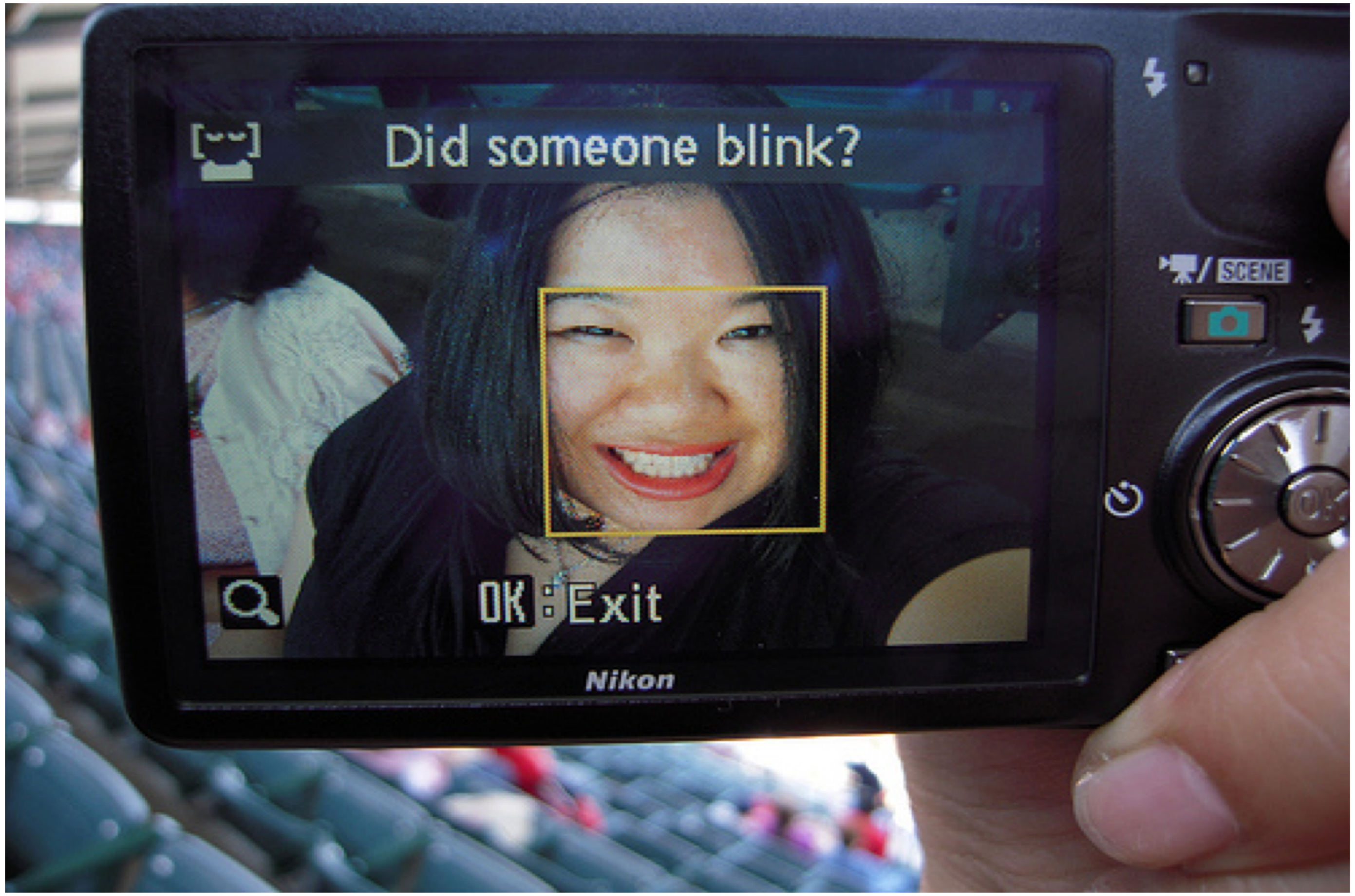
FaceApp apologizes for building a racist AI

Natasha Lomas @riptari / 2 years ago Comment

https://techcrunch.com/2017/04/25/faceapp-apologises-for-building-a-racist-ai/

Algorithms Discriminate: Nikon Blink Detection

Two kids bought their mom a Nikon Coolpix S630 digital camera for Mother's Day… when they took portrait pictures of each other, a message flashed across the screen asking, "Did someone blink?"



http://content.time.com/time/business/article/0,8599,1954643,00.html

Algorithms Discriminate: Face Recognition

Software engineer at company: “It got some of our Asian employees mixed up,” says Gan, who is Asian. “Which was strange because it got everyone else correctly.”

Gfycat's facial recognition software can now recognize individual members of K-pop band Twice, but in early tests couldn't distinguish different Asian faces. GFYCAT

https://www.wired.com/story/how-coders-are-fighting-bias-in-facial-recognition-software/

Algorithms Discriminate: Book Shopping

Anti-Semitic Bias:

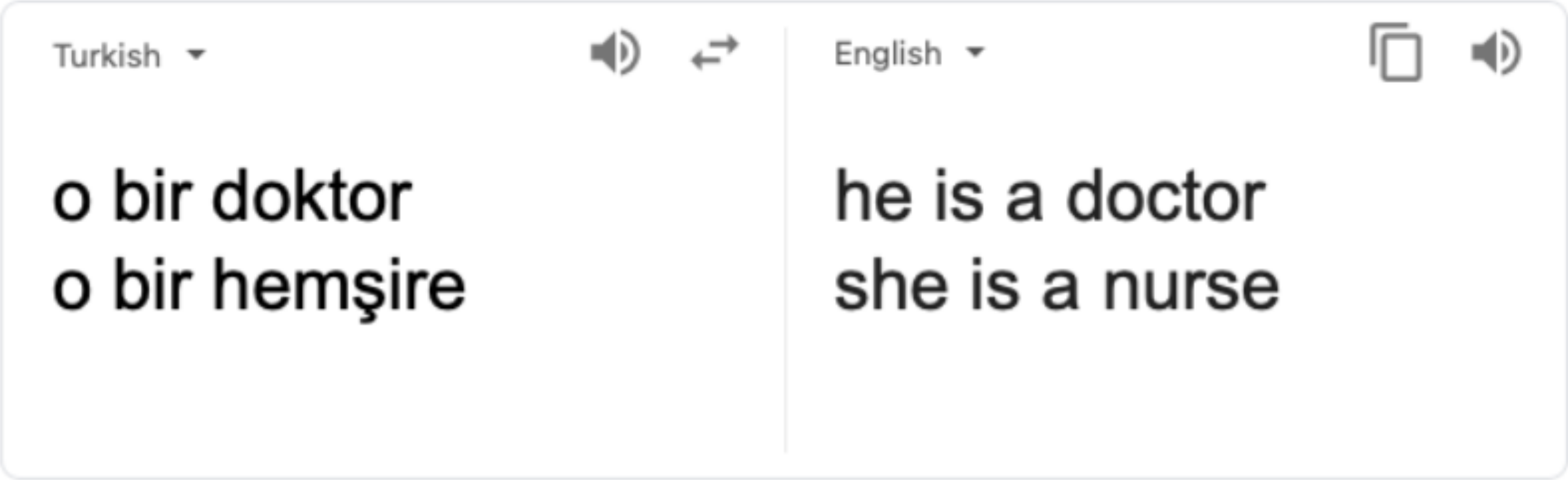
https://www.tabletmag.com/scroll/275042/amazons-algorithm-has-an-anti-semitism-problem

Algorithms Discriminate: Job Recruiting

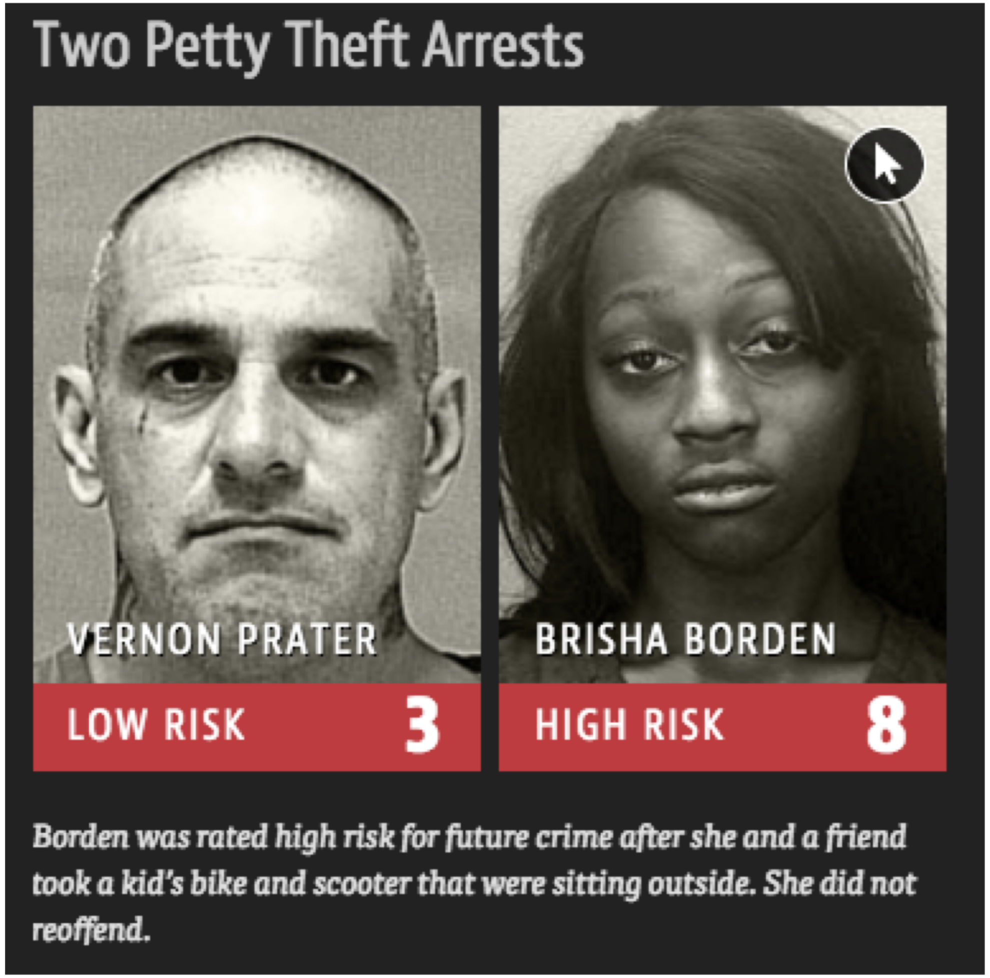
Amazon’s algorithm learned to systematically downgrade women's CVs for technical jobs such as software developer.

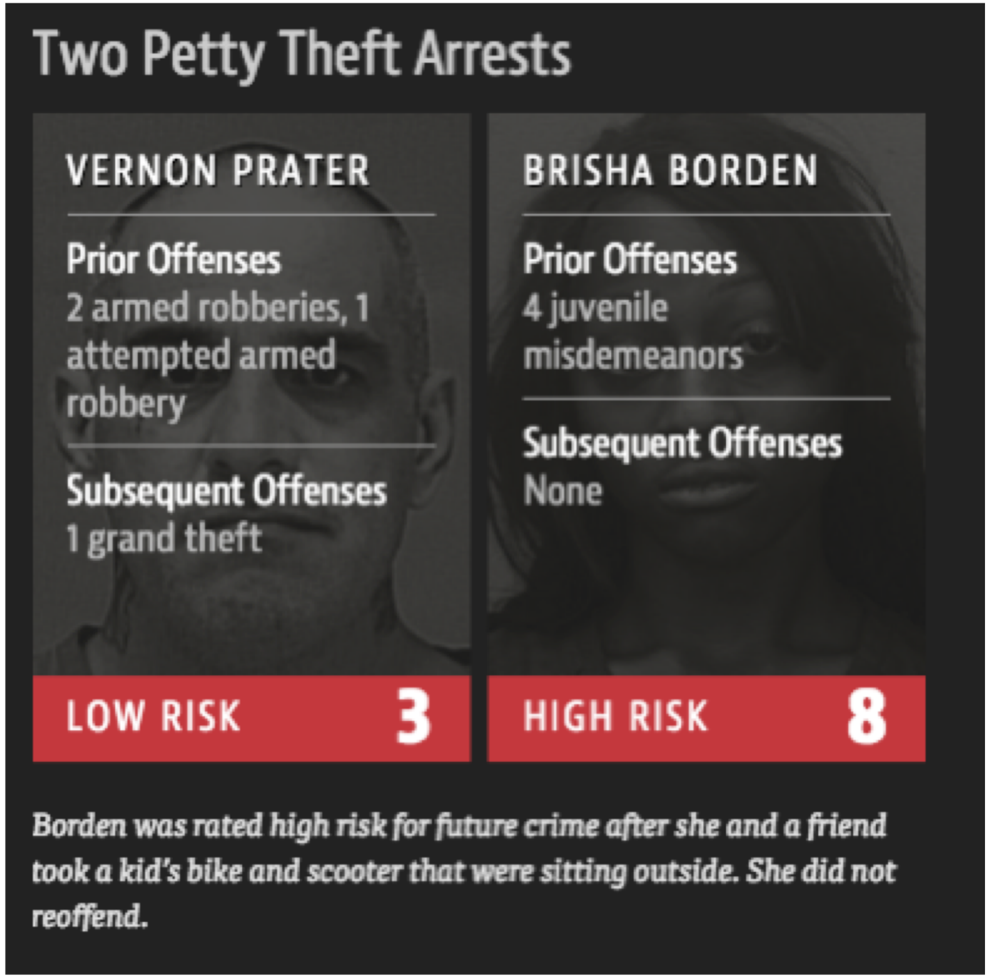
https://phys.org/news/2018-11-amazon-sexist-hiring-algorithm-human.html

Algorithms Discriminate: Language Translation



Algorithms Discriminate: Criminal Sentencing





https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Algorithms Discriminate: And MANY more…

- e.g.,

README.md

Awful AI

Awful AI is a curated list to track *current* sca

Artificial intelligence in its current state is un
Often, AI systems and predictions amplify ex
more and more concerning the uses of AI te
hope that *Awful AI* can be a platform to spur
fight back!).

Discrimination

AI-based Gaydar - Artificial intelligence can
their faces, according to new research that s
[summary]

Infer Genetic Disease From Your Face - Dee
photograph of a patient's face. This could le

https://github.com/daviddao/awful-ai

Gender, Race, and Power in AI

A Playlist

AI Now Institute Apr 17, 2019 · 6 min read

Gender, Race, and Power in AI is the product of a year-long survey of literature at the nexus of gender, race, and power in the field of artificial intelligence. Our study surfaced some astonishing gaps, but it also made clear that scholars of diverse gender and racial backgrounds have been sounding the alarm about inequity and discrimination in artificial intelligence for decades.

We are concerned that in the rush to diagnose and solve 'new' problems, this critical scholarship is deserving of greater attention. So, we're offering up what we like to think of as a playlist — some of the greatest hits and deep cuts from the literature on gender, race and power in AI — by sharing the work that has inspired us, we hope that others might read along with us.

https://medium.com/@AINowInstitute/gender-race-and-power-in-ai-a-playlist-2d3a44e43d3b

Algorithms Discriminate

How would you try to fix issues like these?

Today’s Topics

- Biased Machine Learning Algorithms
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

We know that algorithms are not perfect.

How can we alleviate the issue that ML algorithms that discriminate?

FAT Machine Learning: In Vague, Lay Terms

- **Fairness:** treat people fairly
- **Accountability:** mimic infrastructure to oversee human decision makers (e.g., policymakers, courts) for algorithm decision-makers
- **Transparency:** clearly communicate algorithms’ capabilities and limitations

FAT Machine Learning: Fairness

- How to make more fair methods?
  - Pre-processing:
    - Training data: modify it
  - Optimization at training:
    - Algorithm: e.g., add regularization term to objective function to penalize unfairness
    - Features: remove those that reflect bias; e.g., gender, race, age, education, sexual orientation, etc.
  - Post-process predictions
    - Counterfactual assumption: check impact of modifying single feature

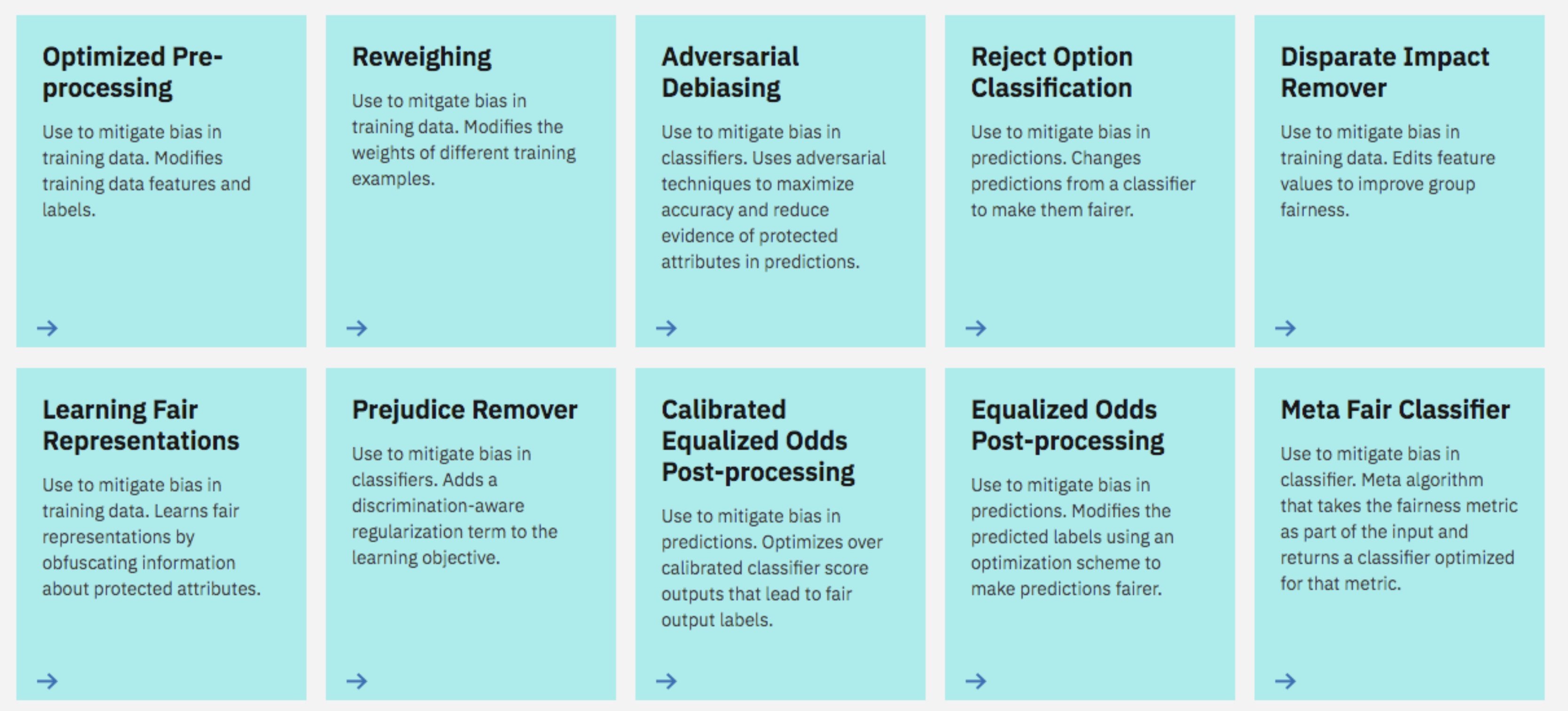
https://fairmlclass.github.io/; https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

FAT Machine Learning: Fairness

- Fairness – how to define this mathematically?
  - e.g., group fairness (proportion of members in protected group receiving positive classification matches proportion in the population as a whole)
  - e.g., individual fairness (similar individuals should be treated similarly)

e.g., IBM's AI Fairness 360 Open Source Toolkit

70+ fairness metrics and 10+ bias mitigation algorithms

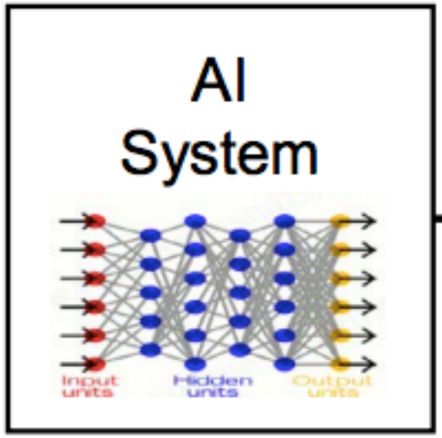


FAT Machine Learning: Accountability

- Accountability: who is accountable for ML algorithm behavior?
  - e.g., developers who must design algorithms so that oversight authorities meet pre-defined rules (“procedural regularity”)?
  - e.g., data providers?
  - e.g., regulators who determine scope of oversight (e.g., require describing and explaining failures in ML systems)?

Joshua Kroll et al. “Accountable Algorithms.” University of Pennsylvania Law Review, 2017.

FAT Machine Learning: Transparency



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

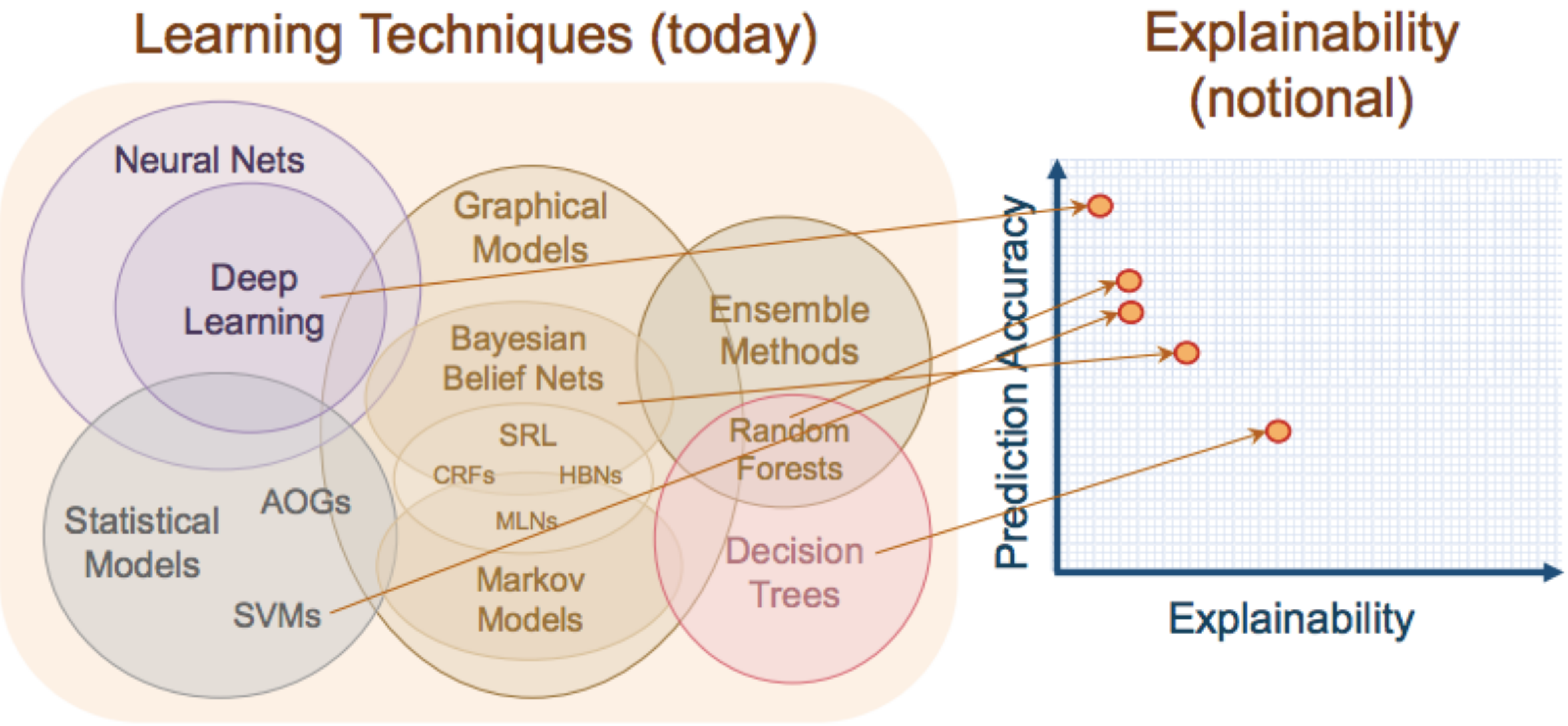
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

FAT Machine Learning: Transparency

New Approach

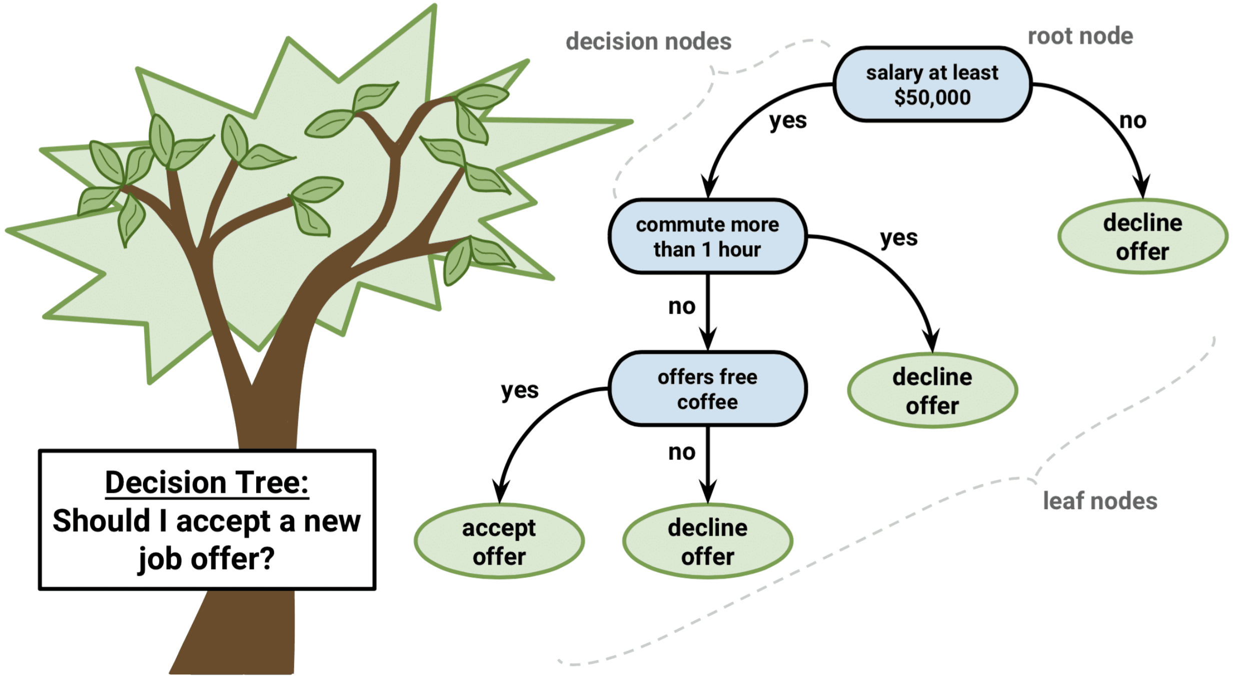
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance



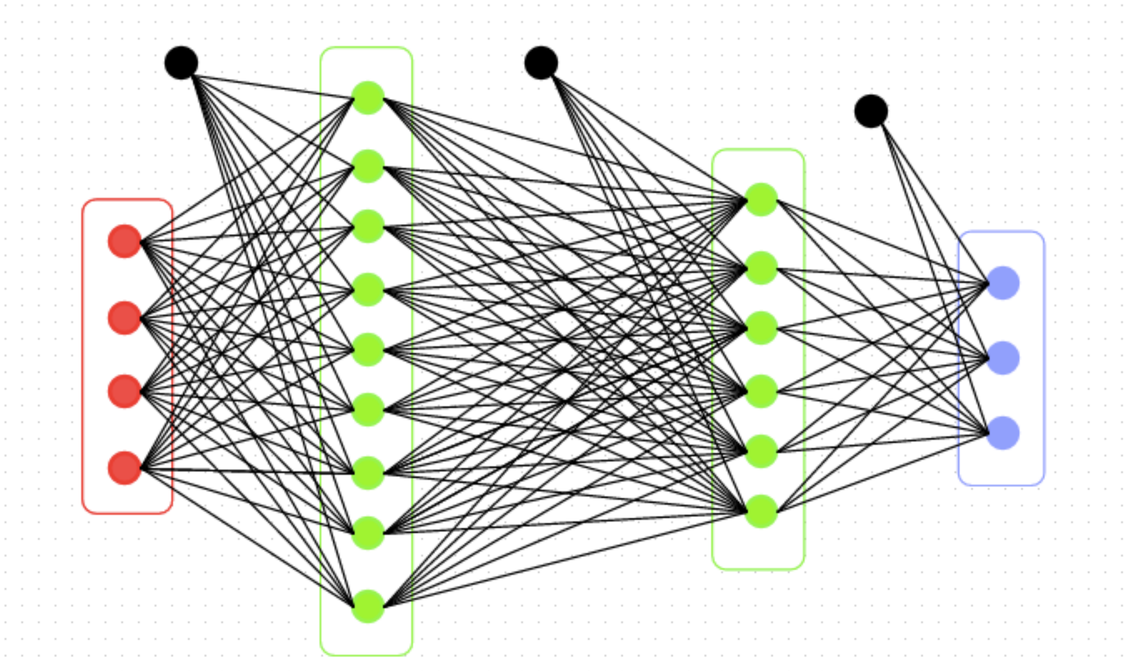
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

FAT Machine Learning: Transparency

- Transparency: how are predictions made by black box ML algorithms?
  - e.g.,



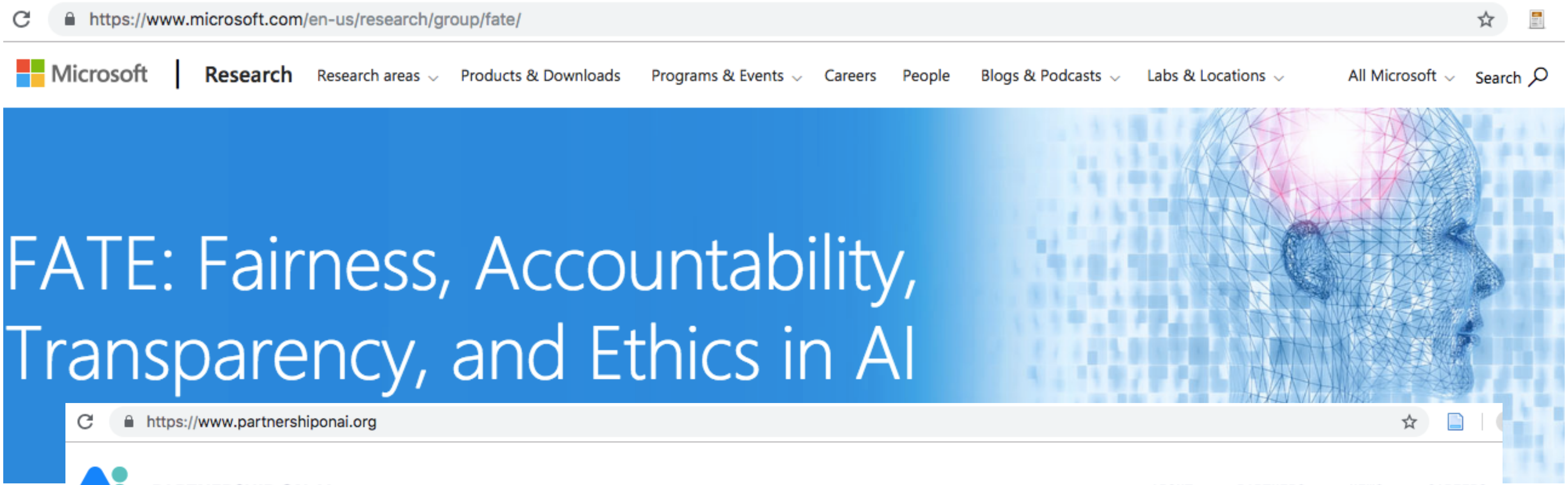
VS

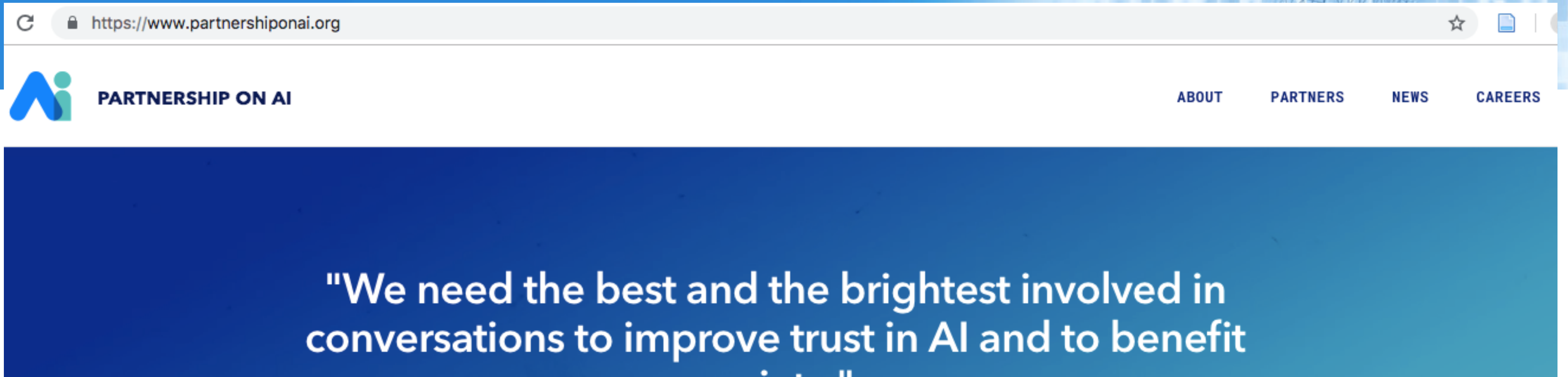


Source: http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

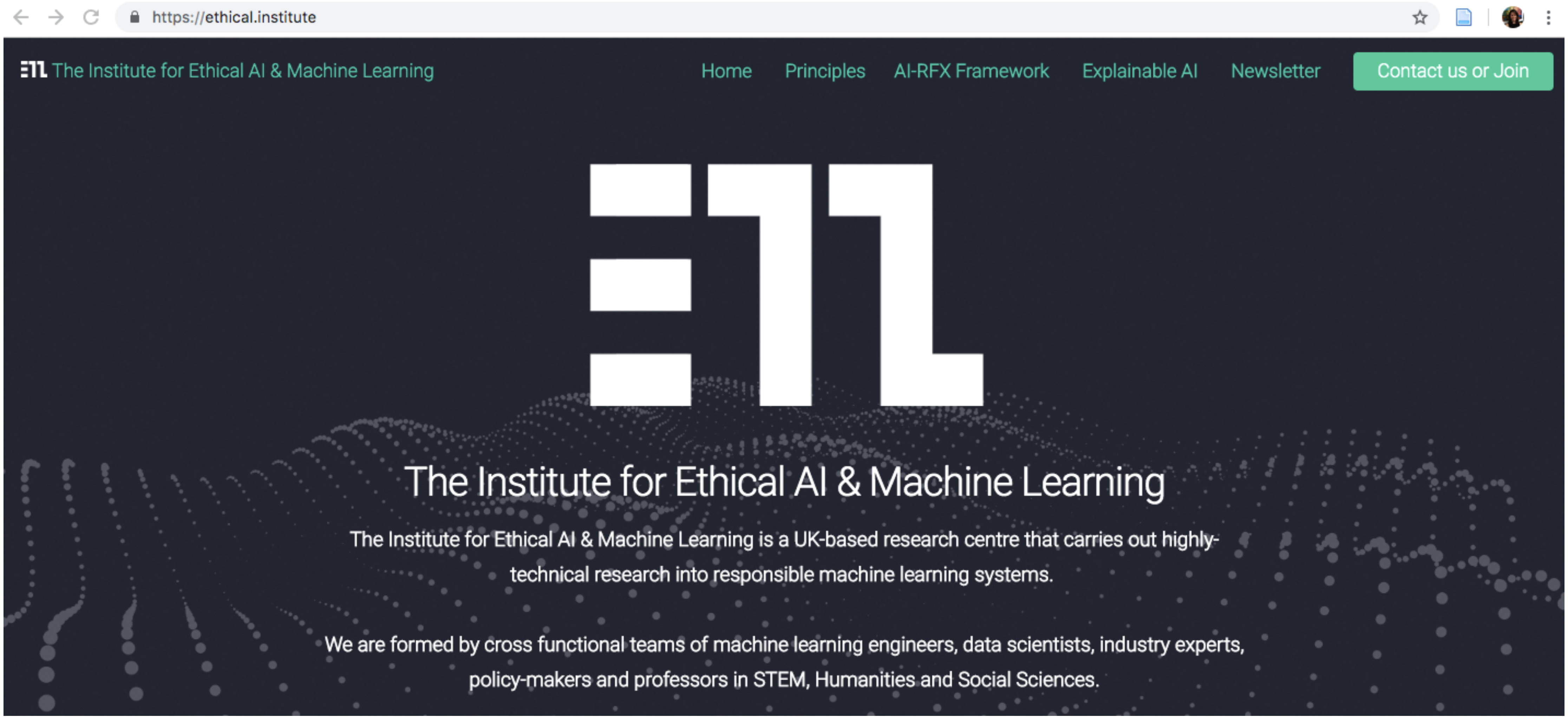
Source: https://towardsdatascience.com/build-your-first-deep-learning-classifier-using-tensorflow-dog-breed-example-964ed0689430

Industry (Facebook, Google, Uber, & more…)





Institutes



https://www.fast.ai/2018/09/24/ai-ethics-resources/

Academia: Workshops

Academia: Workshops

Academia: Workshops

Academia: Annual Workshop Since 2014…

Not Secure | www.fatml.org/schedule/2014/page/scope-2014

FAT/ML 2018 2017 2016 2015 **2014** Organization Resources Mailing list

Scope Attend Schedule Speakers Organizers

Scope

This interdisciplinary workshop will consider issues of fairness, accountability, and transparency in machine learning. It will address growing anxieties about the role that machine learning plays in consequential decision-making in such areas as commerce, employment, healthcare, education, and policing.

Academia: Annual Workshop Scope…

Questions to the machine learning community include:

- How can we achieve high classification accuracy while eliminating discriminatory biases? What are meaningful formal fairness properties?
- How can we design expressive yet easily interpretable classifiers?
- Can we ensure that a classifier remains accurate even if the statistical signal it relies on is exposed to public scrutiny?
- Are there practical methods to test existing classifiers for compliance with a policy?

Academia: And Many More Resources…

https://fatconference.org/resources.html

Today’s Topics

- Biased Machine Learning Algorithms
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

We know that algorithms are not perfect.
Algorithms can be biased.

Are they ethical to use?

Time for a group activity!

Unacceptable to acceptable:
Using ML to sentence people for a crime

Unacceptable to acceptable:
Using ML to diagnose diseases

Unacceptable to acceptable:
Using ML to filter resumes for jobs

Unacceptable to acceptable:
Using ML to determine eligibility for a loan

Today’s Topics

- Biased Machine Learning Algorithms
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

Google Form: Guest Speaker

- Guest: Dr. Mehrnoosh Sameki, Senior Technical Program Manager at Microsoft (https://www.linkedin.com/in/mehrnoosh-sameki-a2a02245/)
  - Share one question for her for tomorrow's visit