



Week 8

MORPH Algorithmic Fairness

A series of four concentric circles in a light gray color, centered on the left side of the slide.

Agenda

- This week in fairness
- Project progress update
- Course recap
 - Recap what we've learned
 - Related topics
 - Future resources
 - Takeaways





This Week in Fairness

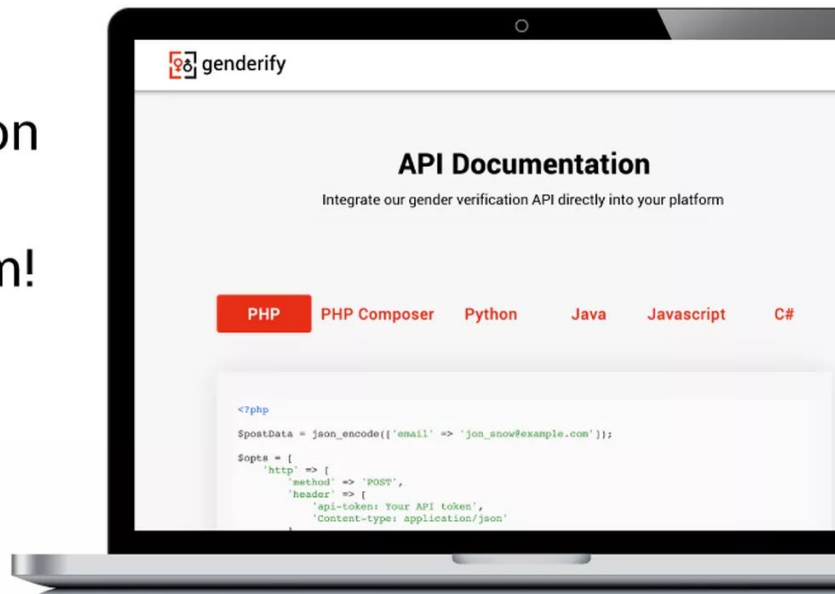
A series of four concentric circles in a light gray color, centered on the left side of the slide.

Genderify

- A new service that launched 2 weeks ago
- Promised to identify someone's gender by analyzing their name, email address, or username with the help of AI



Integrate our
gender verification
API directly
into your platform!



Genderify



Identify the gender of your customers

Enhance your customer data with our real-time AI-based gender verification from name, username or email. Get the best of our unique solution that's the only one of its kind available in the market. Go ahead, **try below!**

Check

Male: 27.70% Female: 72.30%

[Close](#)

Don't agree with the results? Suggest yours.

Submit

Identify the gender of your customers

Enhance your customer data with our real-time AI-based gender verification from name, username or email. Get the best of our unique solution that's the only one of its kind available in the market. Go ahead, **try below!**

Check

Male: 63.60% Female: 36.40%

[Close](#)

Don't agree with the results? Suggest yours.

Submit



Welcome, Product Hunters!

Identify the gender of your customers

We'd love to offer you a special 30-day free plan with 1000 credits per day!
Enjoy your free trial. **Sign up today!**



Check


Male: 95.70% Female: 4.30% Close

[Login](#) or [Register](#) to check up to 100 names, usernames or emails at once!

3. How do you collect the data?



We use integrated multi-source data to deliver the most accurate results. We combine the data from publicly available governmental sources with the information obtained from the social networks to ensure the best possible matches. Each name is added to our database by verifying the data obtained from different sources.

A graphic consisting of four concentric circles of increasing size, centered on the left side of the page.

Genderify team tweeted “Since AI trained on the existing data, this is an excellent example to show how bias is the data available around us.” Hmm...



The word **GIGO** is displayed in a bold, black, sans-serif font. It is positioned to the right of a series of four concentric circles that are centered on the left side of the slide. The circles are thin and light gray.

GIGO

- In CS: “garbage in, garbage out.”
- Models fed by biased data will tend to produce biased predictions
- Many such flawed models may be deployed to market without proper review
 - Genderify was shut down within a day



A series of four concentric circles in a light gray color, centered on the left side of the slide.

Takeaways

- Nothing here is surprising
- Research may be advanced, but much work needs to be done to **bridge research in fairness with industry deployment/applications**
- Your thoughts?



To read more:

Service that uses AI to identify gender based on names looks incredibly biased

<https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify>

AI-Powered 'Genderify' Platform Shut Down After Bias-Based Backlash

<https://syncedreview.com/2020/07/30/ai-powered-genderify-platform-shut-down-after-bias-based-backlash/>

The background is a solid black field. A large, bright white circle is centered on the page. To the upper-left of this white circle, there is a cluster of three overlapping circles in various shades of gray, with the outermost being the lightest and the innermost being the darkest. To the lower-right of the white circle, there is a series of five concentric white circles of decreasing size, creating a ripple effect.

Project Update

The background is black. A large white circle is centered on the slide. To the left of the white circle, there are two overlapping circles: a larger dark gray one and a smaller light gray one. To the right of the white circle, there are several concentric white circles of varying sizes.

Course Recap & Future Directions

Course Recap: Algorithmic Fairness

- Individual fairness
 - Disparate treatment vs impact
 - Similar individuals should be treated similarly
 - Frame as optimization (linear program)
- Group fairness
 - Statistical parity, equalized odds, predictive rate parity
- Aware vs Unaware



Course Recap: Algorithmic Fairness

- Causality
 - Directed acyclic graphs
 - Causal models
- Counterfactual fairness
 - A model is counterfactually fair if it produces the same prediction for both an individual and its counterfactual
 - Probabilistic construction



Course Recap: Fairness Applications in AI

- NLP
 - Bias in word embeddings (including programming demo)
 - Lipstick on a Pig
- Computer vision
 - Gender Shades
 - Importance of intersectional consideration



Course Recap: Fairness Applications in AI

- Deepfakes
 - History
 - Advantages and disadvantages
 - Recent advances in audio & video
- Federated learning
 - Preserving privacy in de-centralized manner

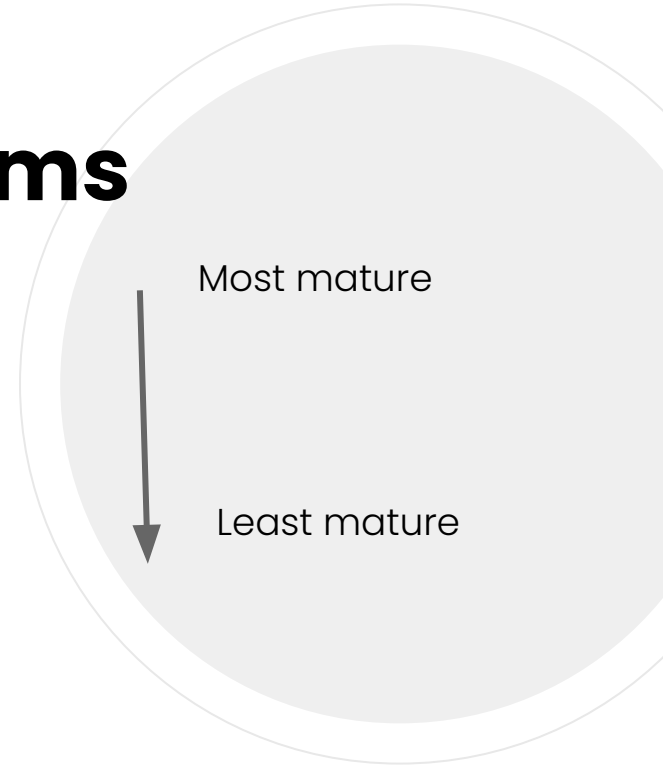
Towards Ethical Algorithms

- It's important to embed social values in algorithms
- In doing so, we need to be precise about their definitions and consequences
 - Don't let the algorithm choose what fairness means



Towards Ethical Algorithms

1. Privacy
2. Fairness
3. Interpretability
4. Morality





Privacy

- How to preserve people's privacy?
- **Differential privacy** (2006)
 - Publicly share information about a dataset by describing patterns of groups within the data without revealing information about individuals
 - Used in, iOS, Android, Chrome; 2020 Census
- Cynthia Dwork, pioneer of differential privacy:
"Anonymized data isn't"
 - Either your data isn't really anonymized, or you've anonymized it so much it's not data anymore



A series of four concentric circles in a light gray color, centered on the left side of the slide.

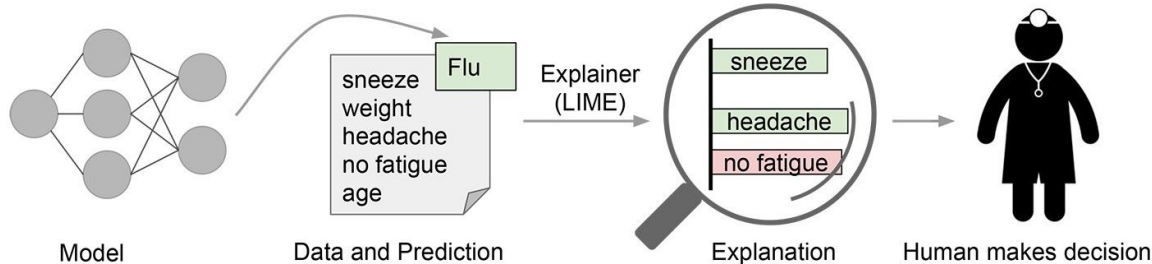
Fairness

- Work in progress
- This class!
 - As we saw — many papers, theorems, proposed properties; differs per domain
 - Not all definitions agreed upon
 - Only beginning to understand tradeoffs between different kinds of fairness, and between fairness and accuracy



Interpretability

- Even more of a work in progress
- “The degree to which a human can understand the cause of a decision”
- Models being created to explain how a ML algorithm is making its predictions
 - LIME, SHAP
- Good textbook to reference: [Interpretable ML](#)



Future Resources

- Math, statistics, computer science courses
- Some books for all audiences:
 - Weapons of Math Destruction
 - Ethical Algorithms
 - <https://www.goodreads.com/shelf/show/ai-ethics>
- If there are topics you are specifically interested in, see what resources, blog articles, books, papers there are, or reach out to me!



Course Takeaways: Always Remember These Considerations

- How do we (mathematically) define what it means for an algorithm to be fair?
- How do we use these definitions to construct algorithms that are fair?
- How do these algorithms impact all populations and subgroups? Who is affected?



Course Takeaways: Always Remember These Considerations

- Who designed and created these algorithms?
- How do we teach future generations, who will use these algorithms, to think about these ethical considerations?
- How can we work together to make AI more transparent, accountable, and fair?





Course Takeaways

- The goal is NOT to spot something and immediately look for/claim bias, racism, sexism
 - A lot of difficult questions that technology poses for us – cannot blindly commit to good or bad
- Instead, the goal is to carry these considerations with you as you are confronted with more algorithms and technology going forward
 - **Intersect both quantitative and ethical thinking**





Thank you.

