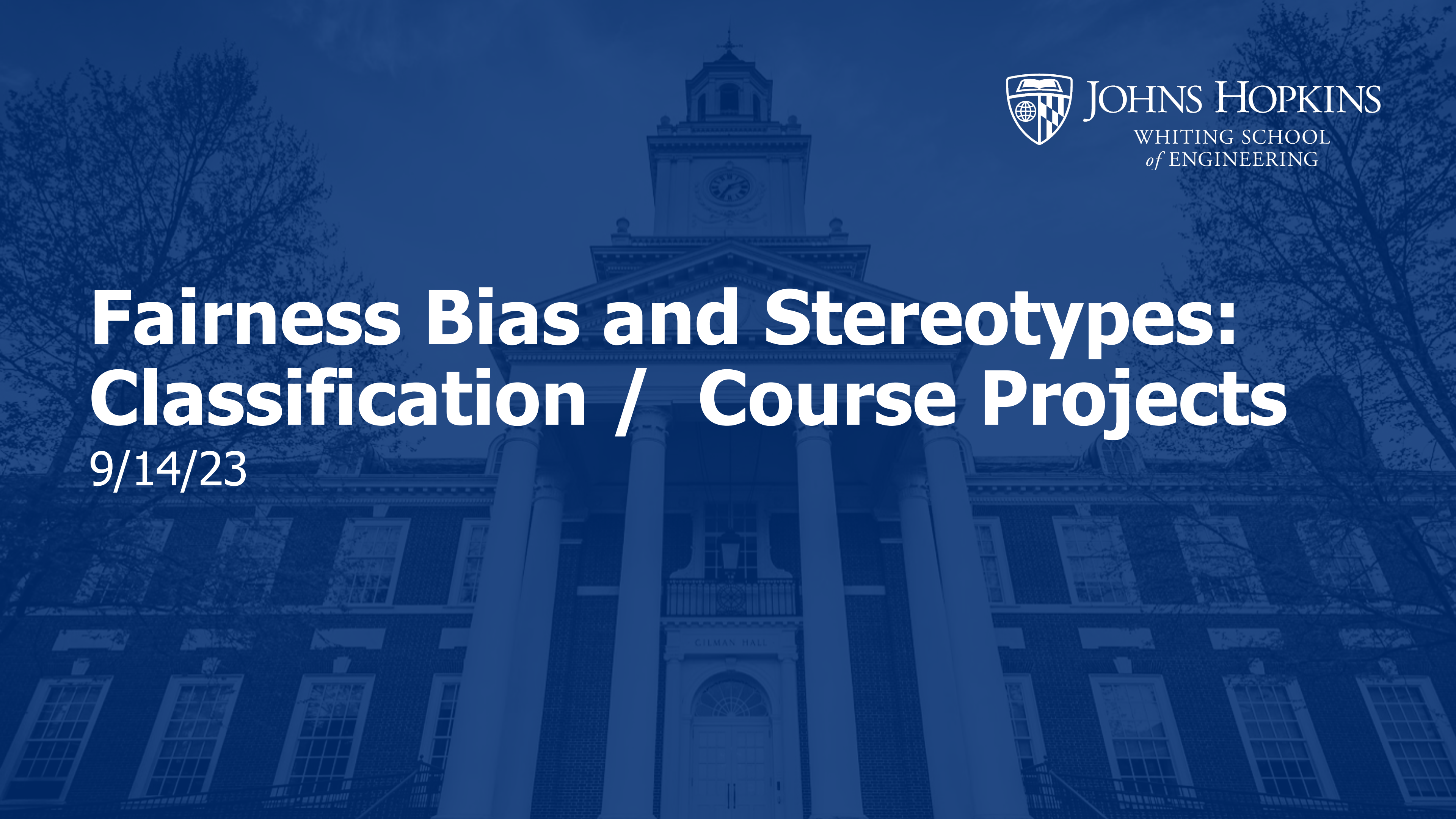

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Fairness Bias and Stereotypes: Classification / Course Projects

9/14/23

Course Project

- http://ai-ethics-601-770.cs.jhu.edu/fa2023/project.html

1. Literature Review [Due September 28]
2. Project Proposal (4-8 pages) [Due about 2 weeks later]

    Written Report

    Short in-class presentation
3. Final Report [Due at the end of the semester]

    In-class poster session

Literature Review

- Short paper (min 4 pages, max 8 pages) summarizing and synthesizing several papers (or works in general)
- Groups of one should review 5 papers, groups of two should review 7 papers, and groups of three should review 9 (you are welcome to review more)
- Most people will choose the same topic for your lit review and final paper, but this is not required, and the lit review will be graded on its own
- More tips on the course website: http://ai-ethics-601-770.cs.jhu.edu/fa2023/project.html

Please use the ACM template, with “nonacm” and “sigconf parameters (2-column format) https://www.acm.org/publications/proceedings-template. No need to include the teaser figure, CSS concepts, or keywords

Project Topics

- Can be anything related to the course
- We encourage you to talk to instructors about ideas (via email, Piazza, office hours, schedule additional appointment)
- Literature review focuses on identifying broad topic area – you have more time for specific project ideas!

Project Examples

- Conduct a technical assessment of a model or deployed system, such as probing for unfairness
  - Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. ACL 2023
  - Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." FAccT 2023.
  - Inna Wanyin Lin*, Lucille Njoo*, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. "Gendered Mental Health Stigma in Masked Language Models" EMNLP (2022)

Project Examples

- Write a critical survey of a topic in AI
  - Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Annual Meeting of the Association for Computational Linguistics (ACL)
  - Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. "A Survey of Race, Racism, and Anti-Racism in NLP" ACL (2021).

Project Examples

- Conduct a technical study, focused on building models to address ethical concerns.
  - Methods for reducing toxicity in model outputs
  - Methods for reducing stereotyping in model outputs
  - Methods for privacy preservation and anonymization

Project Examples

- Write a survey structured around a policy or regulation question (what should policy makers understand about AI in writing regulations?)
  - Some starting points:
    - AI Bill of Rights: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
    - https://www.vox.com/future-perfect/2023/7/3/23779794/artificial-intelligence-regulation-ai-risk-congress-sam-altman-chatgpt-openai
    - https://techpolicy.press/

Project Examples

- Conduct a manual or automated meta-analysis of research processes in papers or other avenues where research discussions take place (social media, news)
  - Birhane, Abeba, et al. "The values encoded in machine learning research." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022
  - Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. *A Systematic Review of Ethics Disclosures in Predictive Mental Health Research*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023
  - Media hype around AI
  - Prevalence of corporate research at different ML venues
  - Policy discussions around AI

Discussion Time

- Zhao et al. 2017 introduces corpus constraints to reduce bias amplification. What are some of the strengths and limitations of this approach? Where might it be useful and where might it be difficult or impossible to apply?
- [Zhao et al. 2017] What further experiments and data analysis might be helpful to better understand the potential impact of the methods?
- [Sap et al. 2019] What are some of the broader challenges involved in hate/offensive speech classification (or even text classification) suggested by this work? What might be alternatives to building hate speech classifiers?
- [Sap et al. 2019] What are some of the assumptions made in this analysis, e.g. about race, AAE, prevalence of hate speech? How might they be relaxed in follow-up studies?
- Given Tuesday's readings, what are your thoughts on the choice of metrics in these works? Are there other metrics that might matter?
- What are some of the potential impacts of model biases discussed in these papers? Where might these models be deployed? Who could be harmed?
- What do these papers say about decisions developers make in building AI pipelines rather than just focusing on data bias?

Ideas for data sources

- Generally, data that has been used in prior research
  - Identify from reading research papers, which may have directly released the data or cited how they collected it (for example, this paper has a pointer to this archive of US Congressional Records). Authors are also often willing to share data if you reach out to them, even if they didn't post it publicly.
  - Shared tasks associated with workshops. For example, past versions of the workshop on NLP for Internet Freedom has had shared tasks on misinformation and censorship
  - Benchmark/shared task data is useful as analysis data, not just as model training data, for example this paper and this paper criticize standard NLP benchmarks
- There are some tools for data collection
  - Semantic Scholar API provides data on research publications (more details in the paper), which could be used for a variety of meta-analyses, like this paper on Big Tech influence in research
  - Many websites are possible to scrape (if you pay attention to terms of service and rate limits)
- There are some pre-collected archives of data
  - Common Crawl https://commoncrawl.org/ - large archive of web data
  - Twitter releases archives of data they've identified as potential information manipulation operations. It looks like they have not taken this down: https://transparency.twitter.com/en/reports/moderation-research.html

Ideas for data sources

- Some random data Anjalie has:
  - Wikipedia biography pages (with inferred race and gender information): https://anjalief.github.io/wikipedia_bias_viz/
  - Tweets about the Russia-Ukraine War from this paper
  - Online media articles about the #Metoo movement from this paper
  - Tweets about #BlackLivesMatter protests from this paper
  - Articles from the New York Times (1990s-2016) originally collected in this paper, also analyzed in this paper
  - Tweets collected in late 2022 about the protests in Iran