



JOHNS HOPKINS

WHITING SCHOOL  
of ENGINEERING

# Fairness, Bias, and Stereotypes: Generation


9/19/2023

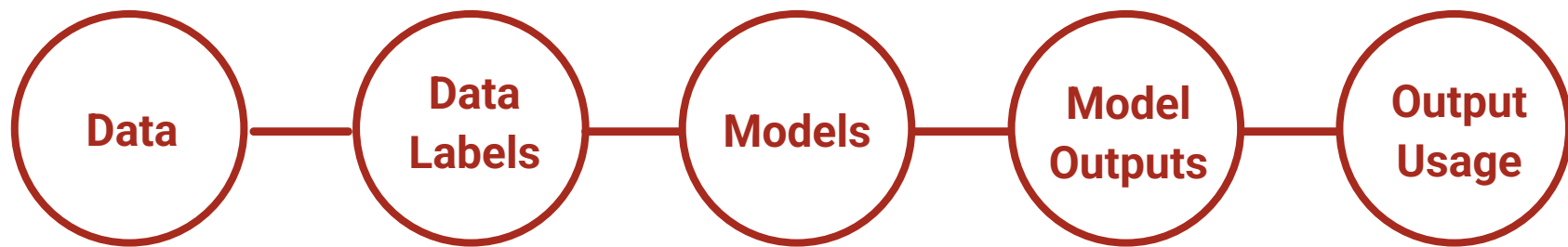
# Discussion

---

# Announcements

---

- Late piazza posts
  - We start grading Piazza posts at 9pm
  - We've been lenient about late posts for the first few weeks, but we won't do that anymore
- Please don't use AI to write your Piazza posts 
  - Using AI to correct grammar or to translate is ok.
  - If we continue to be suspicious of AI-generated posts, we'll need to change the course format, like having in-class quizzes on the readings (we don't want to do this!!)
- Positive disincentive: we will drop your two lowest-scoring Piazza posts

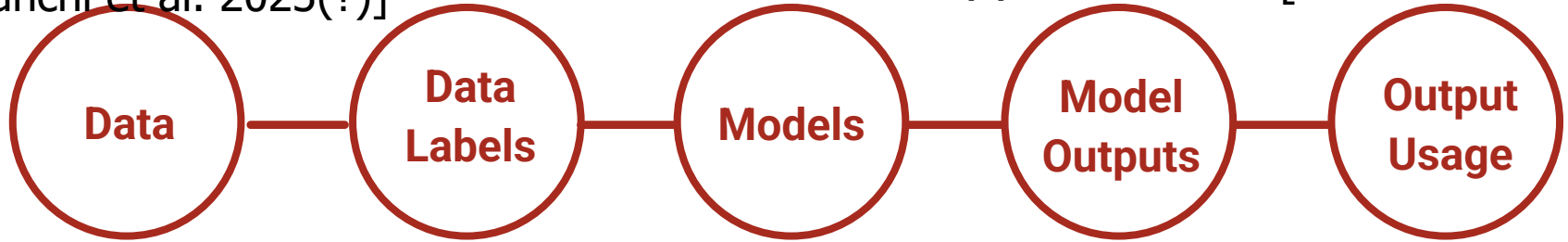


# Where does “bias” originate?

[Feng et al. 2023;  
Bianchi et al. 2023(?)]

Models amplify bias  
[Zhao et al. 2017;  
Bianchi et al. 2023(?)]

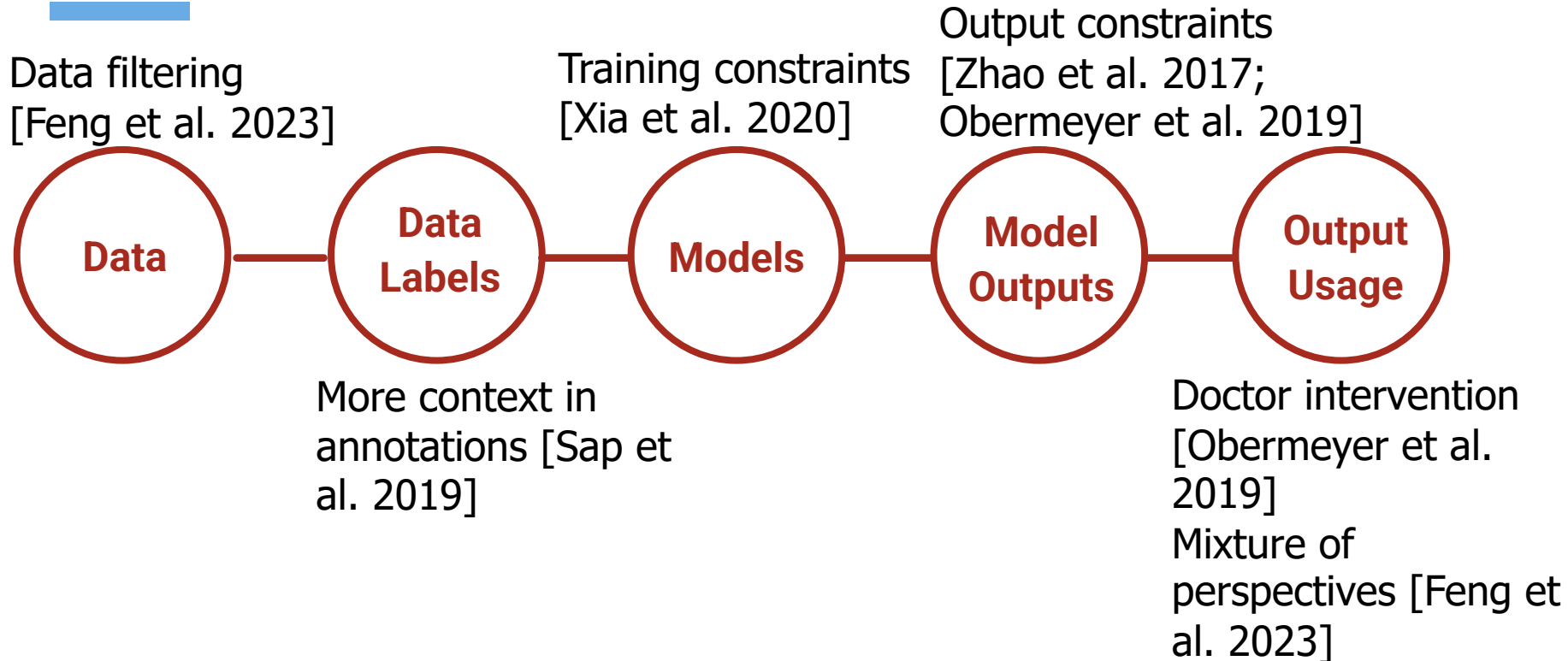
Deployment in  
specific applications,  
“disparate impact”  
[Chouldechova 2017]



Data labels can  
introduce new biases  
[Sap et al. 2019]

Choice of output  
[Obermeyer et al.  
2019]

# Where do developers attempt to mitigate it?



# Next Unit: Values and Design

---

- We've seen some examples of possible mitigation, but how do we decide when to intervene? How to intervene? What is the desired behavior of a system?
  - How can we integrate more thoughtful design from the beginning, rather than post-hoc fixes?
- Some starting points from the perspective of researchers and engineers (e.g. bottom-up) rather than policy and regulation (top-down)
  - Sept 24: Values and Design: Value sensitive design
  - Sept 26: Values and Design: Participatory design
  - Oct 1: Values and Design: Surveyed values