AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 7: AI: Supervised ML and Trust Issues

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

9TH SEP, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 7

- Introduction Segment
  - Recap from Lecture 6
  - Project discussion
- Main Segment
  - Trust Issues Issues – Reading Material – Bias in AI
  - Trust Issues – Reading Material - Turing Institute report on AI for Covid in UK
- Concluding Segment
  - About next lecture – Lecture 8
  - Ask me anything

Image courtesy: Dilbert, Sep 8, 2021

Introductory Segment

Recap of Lecture 6

- Looked at supervised ML methods
  - Random forest
  - Neural methods - MLP
- How to compare classification methods
  - Data preparation is key
  - Choose simplest methods first
  - Use metrics to decide, explainability may be critical
- Quiz 1

Quiz 1 Recap and Discussion

Look at water data from Florida for WaterAtlas project and do four tasks

- WaterAtlas information
  - Website: https://orange.wateratlas.usf.edu/
  - About data collection and group: https://orange.wateratlas.usf.edu/about/
  - APIs to get data - Data download: https://dev.chnep.wateratlas.usf.edu/data-download/beta/
- Local cache with data
  - https://github.com/biplav-s/course-tai/tree/main/sample-code/common-data/water
  - Data for one lake: https://github.com/biplav-s/course-tai/blob/main/sample-code/common-data/water/WaterAtlas-OneLake.csv

Project Discussion – Review Projects

- Information to be shared by students
  - Go to Google sheet: https://docs.google.com/spreadsheets/d/1VAX8ntb5zBQ-vOdsMHMhvEdwoaCZtuBaO4kJdkSA4eQ/edit?usp=sharing
  - Create a Google drive called "CSCE 590-1 Trusted AI (<YourName>)" and share with instructor: firstname.lastname@gmail.com
    - Put shared url in Column E
    - Put project title in column G
    - Create a folder in shared directory call project. Under it, have a Google doc called "Project Description". In it, have the following as bullets with associated details: Problem, User, AI Method, Data, Reliability: Testing, Holding Human Values, Human-AI interaction. See next slide for framework and guidance on what to put.
  - Put Github location for your code in F
    - Create one repository
    - For each quiz, project, etc, create a sub-folder

Course Project – What to Focus On

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
  4. (Data) Explore the data for a solution to work
  5. (Reliability:Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

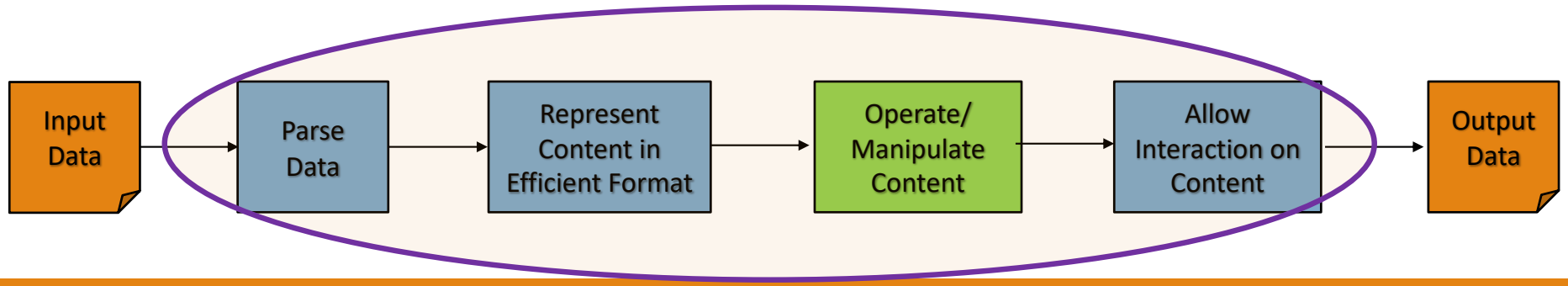
Rubric for Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Main Segment



Context: German Credit Data's Analysis

- Review detailed data exploration at:
https://www.kaggle.com/sanyalush/predicting-credit-risk

- Notice issues in lending ?
  - No single female
  - Discrimination by gender, age, … ?

Discussion on Reading Material - 1

"Biases in AI Systems", Ramya Srinivasan, Ajay Chander
Communications of the ACM, August 2021, Vol. 64 No. 8, Pages 44-49
10.1145/3464903

https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/fulltext

Taxonomy of Biases

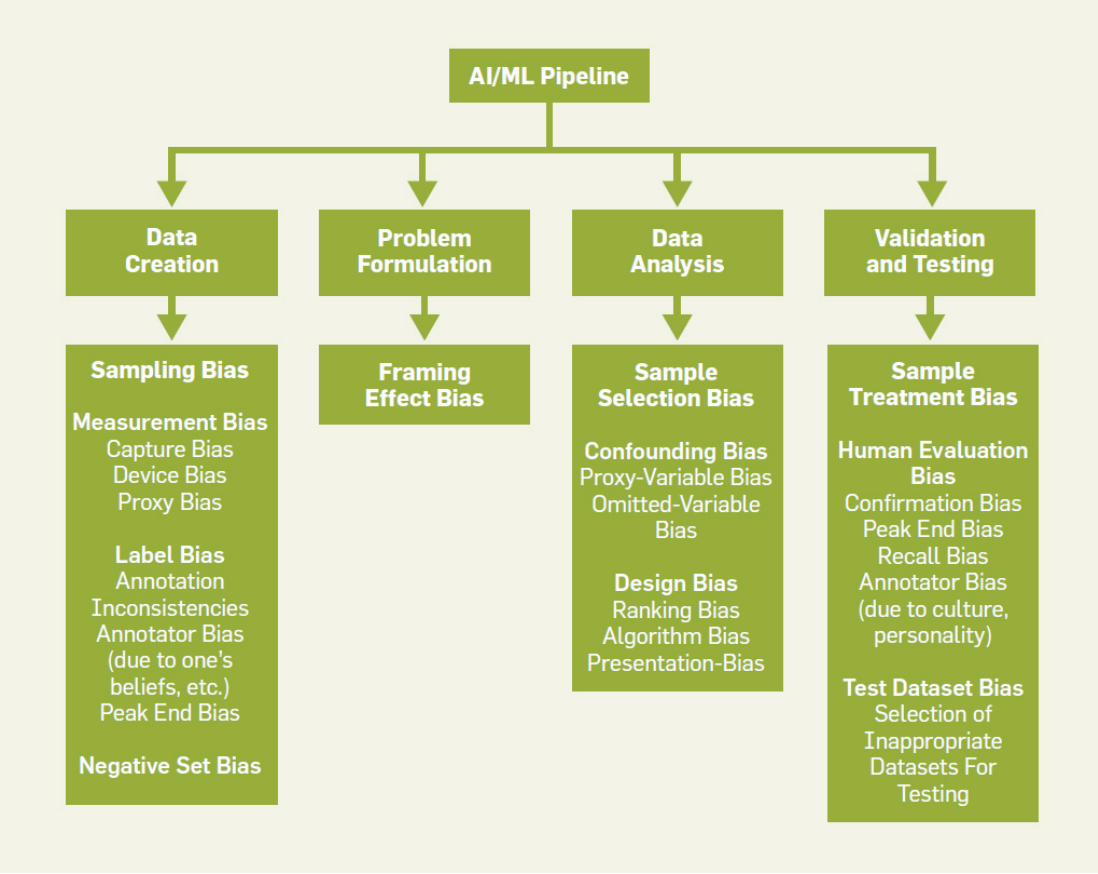
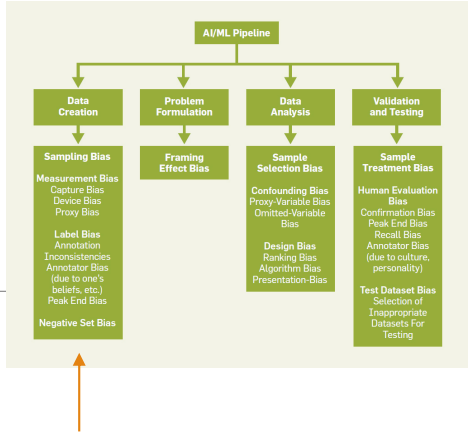


Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

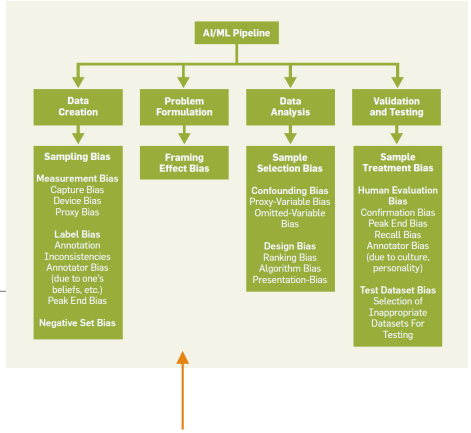
Data Creation Biases



- Sampling bias
  - Selecting specific types of instances more than others
  - Examples:
    - Twitter for finding traffic on a road (assumptions: one social media platform, driver v/s riders)
    - FaceTime for recording accidents (assumptions: a company product, having smart phone)
  - Measurement bias
    - Proxies uses instead of true values (arrests for crime, hospital visits for health)
  - Label bias
    - Inconsistent usage of terms, Subjectivity, peak effect – memory based bias from psychology
    - Serious bottleneck for supervised methods
  - Negative set bias
    - Not having enough data for negative classification

Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Problem Formulation Bias

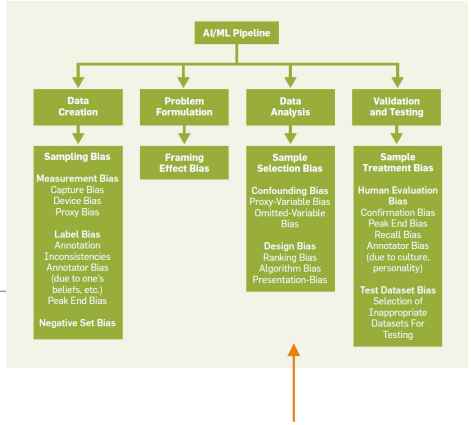


- "What cannot be measured cannot be controlled"

- Framing Effect Bias
  - Compas: predicting recidivism – criminals re-offending
    - Pro-republica claim: biased against black defendants as the group was associated with a higher false-positive rate (**equalized odds** and **equality of opportunity fairness** criteria)
      - **Equalized odds**: if protected and unprotected groups have equal true positive rate and false positive rate
    - Northpointe: scores satisfied fairness from the viewpoint of **predictive rate parity**
      - **Predictive rate parity**: precision, the number of the true positives divided by the total number of examples predicted positive within a group, same within groups

Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Data Analysis Bias

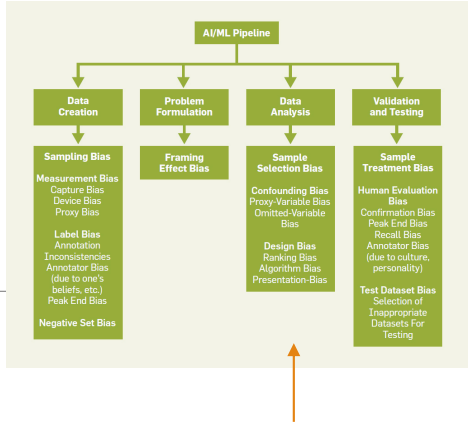


Due to algorithmic properties or computational limitations

- Sample Selection Bias
  - "selection of individuals, groups, or data for analysis in such a way that the samples are not representative of the population intended to be analyzed"
  - Example – *"In analyzing the effect of motherhood on wages, if the study is restricted to women who are already employed, then the measured effect will be biased as a result of conditioning on employed women"*
- Confounding Bias
  - Common causes that affect both inputs and outputs, but are not accounted for. Example – "*admissions to a graduate school are based on the person's previous grade point average. There might be other factors, however, such as ability to get coaching, which in turn may be dependent on sensitive attributes such as race; and these factors may determine the grade point average and admission rates*"
  - Proxy variables: there are proxies for protected variables being omitted; indirect bias
  - Omitted variables: model does not consider variables

Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Data Analysis Bias



Due to algorithmic properties or computational limitations

- Design Bias
  - Ranking Bias: Which top results to be shown? What should be the cut-off?
- Presentation Bias: where and how the result is presented?

Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Validation and Testing Bias

- Human Evaluation Bias
  - Human's biases: confirmation bias, peak end effect, and prior beliefs
- Test Dataset Bias
  - Faces issues similar to (training) dataset selection
  - Further, is it appropriate for application in hand and up-to-date over time?

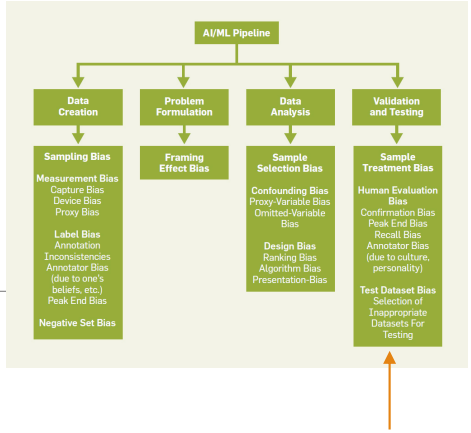


Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Guidelines for Developers

- Draw a structural diagram illustrating various features of interest and their interdependencies
  - To avoid bias, all features that associated with the target feature of interest is important
- Understand which features of the data are sensitive based on the application
  - Understand the laws: protected by law
  - Best practices that are good to have
- Datasets used for analysis should be representative of the true population under consideration
- Standardize rules for annotating data
- Validate model with representative population, not a subset

Discussion on Reading Material - 2

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

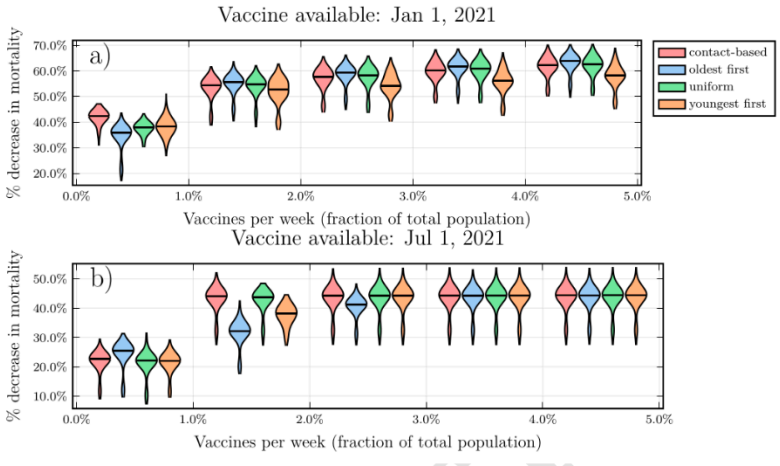
- Approach of report
  - A series of workshops in academic circles in Nov and Dec 2020
  - Discussions and findings distilled into the report

Example AI Usage: Distribution of Vaccines

- **Problem**: Limited supply, larger demand, many technologies, recipient hesitancy; How do distribute equitably, fairly and efficiently
- Possible (automated) solutions
  - Random: pick receiver based on random choice
    - **Benefit**: Easy to implement
    - **Problems**: Equitable but not fair, receiver may not be at risk or not want it, others wanting it may not get it
    - **Question**: assumes we can give vaccine quickly to the selected person
  - Prioritized random: make a prioritized list of groups, assign randomly in each group
    - **Benefit**: identifies affected groups
    - **Problems**: receiver may not want the vaccine
    - **Question**: who comes up with groups?, is it rewarding groups who have not been taking precautions ? Assumes we can give vaccine quickly to the selected person
  - ...
  - Benefit-cost: based on contribution to economy
    - **Benefit**: efficient

Impact of Decisions in Vaccine Distribution

- Article: 'The Pandemic Is a Prisoner's Dilemma Game'
- Prioritising COVID-19 vaccination in changing social and epidemiological landscapes, Sep 2020.
- Choices
  - impact of vaccinating 60+ year-olds first;
  - <20 year-olds first;
  - uniformly by age; and
  - a novel contact-based strategy
- Insights
  - Vaccination reduces median deaths by 32%-77% (22%-63%) for January (July) availability, depending on the scenario.
  - Vaccinating 60+ year-olds first prevents more deaths (up to 8% more) than transmission-interrupting strategies



Discussion on Reading Material

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

- Findings
  1. Researchers responded to COVID need with enthusiasm leading to a large number of projects
     1. **Word-wide* (for context)**: protein to aid disease detection and treatment (**molecular scale**), the analysis of patient data like images and conditions to improve patient care (**clinical scale**) and analysis of cases and social media to predict disease severity, understand mis-information and communicate effectively (**societal scale**).
     2. **UK specific examples**: model disease spread, navigate lockdown
  2. Major hurdle was lack of "robust and timely data", especially access and standardization
     1. Develop protocols for collecting and managing protected data
     2. Develop protocols for generating anonymized and synthetic data

* Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. In Journal of Artificial Intelligence Research 69, 807-845, 2020.

Discussion on Reading Material

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

- Findings
  3. Concern over inequality and exclusion slowed progress. Inadequate representation and engagement from some groups
  4. Challenge in communicating research findings to policy makers and public. Specifically, timeliness, accuracy and clarity.
     - Communication among experts
     - Communication among researchers and policy makers
     - Communication among researchers and public

* Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. In Journal of Artificial Intelligence Research 69, 807-845, 2020.

Concluding Segment

Lecture 7: Concluding Comments

- We looked at trust issues from two perspectives
  - Data analysis with German-credit as example
  - Issues during COVID in UK. Used vaccine distribution to motivate discussion

- Trust issues with AI is real
  - Impacts adoption and business benefit

About Next Lecture – Lecture 8

Lecture 8: Supervised ML – Trust Issues – Mitigation Methods

- Bias definitions
  - Paper: Fairness Definitions Explained
- Mitigation methods
  - Book: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence
  - E.g., Data sampling