AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 12: Review of Lectures and Projects

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

28TH SEP, 2021

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

Organization of Lectures 12

- Introduction Segment
  - Recap from Lectures 10-11
  - Discussion of external talk
- Main Segment
  - AI Fairness, AIF 360 installation and running of tutorial
  - Review of students course projects – peer discussion
  - Quiz 2
- Concluding Segment
  - About next lecture – Lecture 13
  - Ask me anything

Introductory Segment

Recap of Dr. Diptikalyan's Talk – 21 and 23 Sep

1st Lecture:

- Pillars of Trustworthy AI: Fairness, Robustness, Explainability, Transparency
- The challenges of AI testing: Test, Debug, Repair
- [Short Recap] The notion of group and individual discriminations, definitions of group and individual discriminations
- Mitigation method classification:
  - One example algorithm of mitigation method of pre, in, and post-processing
  - Individual Discrimination Testing: 2 simple algorithms for testing
  - one augmentation-based repair algorithm for individual discrimination
- [Recap] The notion of Local, Global and Counterfactual Explanation
  - Local Explainability: Recap Lime for Tabular to motivate Anchor (Ribeiro et al.)
  - Global explainability - Trepan (Shavlik et al.) algorithm (since it is based on decision tree what you are already taught)
  - Counterfactual Explainability - Wachter et al.

2nd Lecture: Hands on session

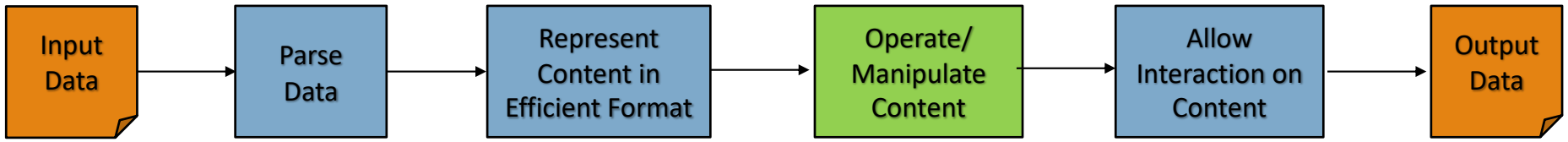
- Recap
- Demonstration of our AI Testing tool.
- Fairness: Hands-on fairness issue detection using AIF360 with the tabular dataset (adult-income)
- Explainability: AIX360

- Could not be covered
- Future classes on Oct 19 and 21 at 10am EST

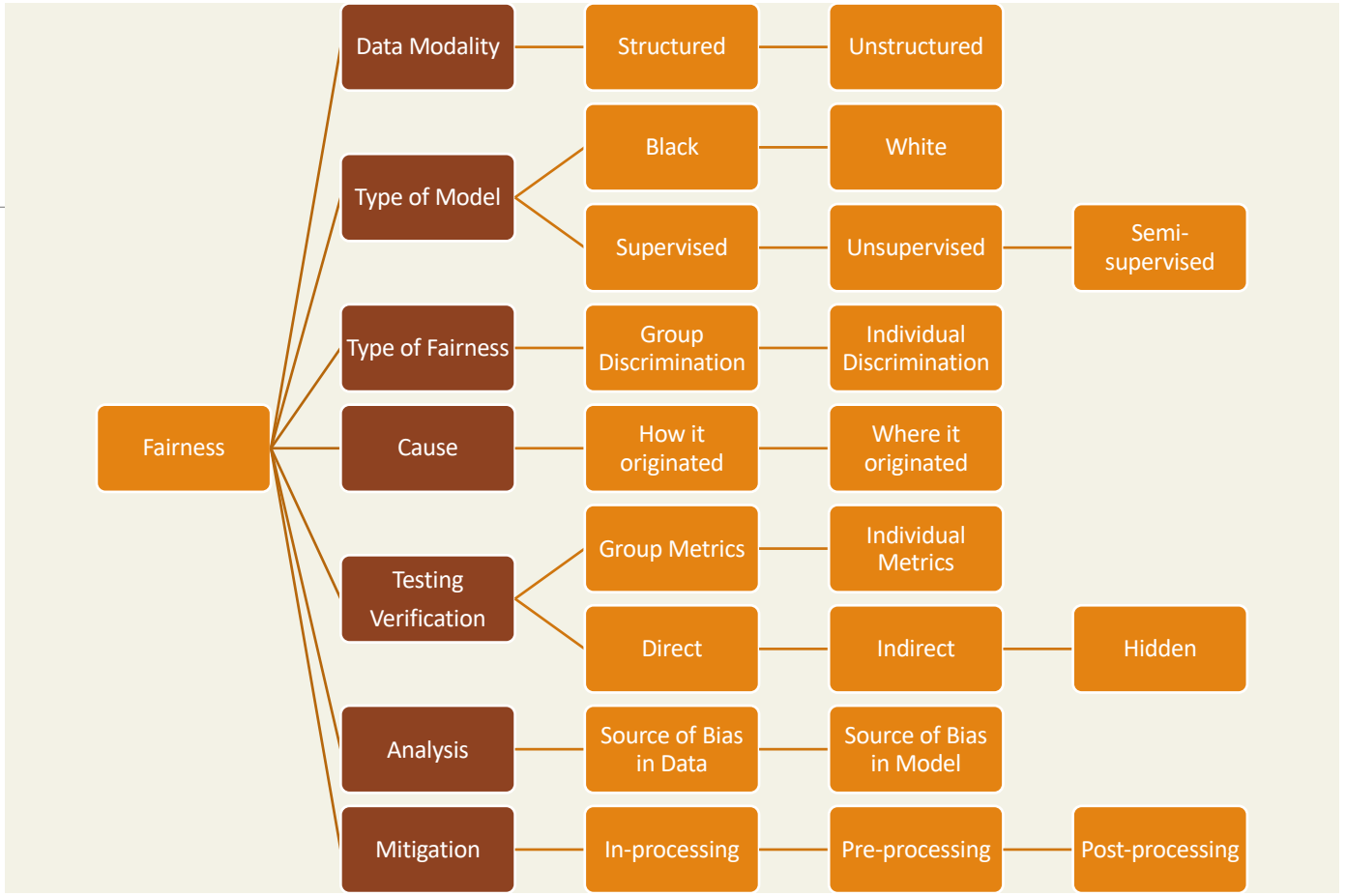
Schedule Re-adjustment

Sep 28 (Tu)	**Review: AI Fairness, Project presentations, Discussion**	Quiz 2
Sep 30 (Th)	AI - Unstructured (Text): Processing and Representation	
Oct 5 (Tu)	AI - Unstructured (Text): Common NLP Tasks	Mid-sem Project Review
Oct 7 (Th)	FALL BREAK	NO Classes, Course Midpoint
Oct 12 (Tu)	AI – Unstructured (Text): Analysis – Supervised ML	
Oct 14 (Th)	AI – Unstructured (Text): Analysis – Supervised ML	
Oct 19 (Tu)	**Invited Guest** – AI - Supervised ML: External Talk/ AI Explanation Methods (AIX)	
Oct 21 (Th)	**Invited Guest** – AI - Supervised ML: External Talk/ Working Session on AIX360	
Oct 26 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	Quiz 3
Oct 28 (Th)	**Review: project presentations, Discussion**	
Nov 2 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	
Nov 4 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods	
Nov 9 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Explanation Methods	
Nov 11 (Th)	Trust: Data Privacy Trust: Fairness & Bias , Trust: Explanation Methods, Trust: AI Testing	

Main Segment



Discussion: Fairness Dimensions



Slide Courtesy: Diptikalyan Saha's Talk on 21 Sep 2021

AI Fairness Tools

IBM AIF360 Tool

https://github.com/Trusted-AI/AIF360

Try Tutorial

- Install tool
- Clone for examples
- Run example

Class Exercise

Other Tools

- Aequitas : University of Chicago
- AI Fairness 360 : IBM Research
- Audit-AI : pymetric.ai
- FairML : Julius Adebayo
- Fairness Comparison : Haverford College, Univ. of Arizona, Univ. of Utah
- Fairness Measures :
- FairTest : Multiple Universities
- Themis™ : University of Massachusetts, Amherst
- Themis-ML : arena.io
- What-If tool : Google
- … Accenture, Facebook, Microsoft …

Project Name:
Student Name:

Problem

User of Results

Data

Approach

Trust Issue

Project Review and Trust Issues

Quiz 2

- Individual component – question 1
- Group component – questions 2 and 3

Concluding Segment

Lecture 12: Concluding Comments

- We reviewed the course
- We reviewed external talk and tool on AI fairness
- Quiz 2

About Next Lecture – Lecture 13

Lecture 13: Unstructured Text – Processing and Representation

- Parsing
- Entity extraction
- Text representation