AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 10 and 11: AI Trust – Invited Lectures

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

21ST AND 23RD SEP, 2021

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

Organization of Lectures 10 and 11

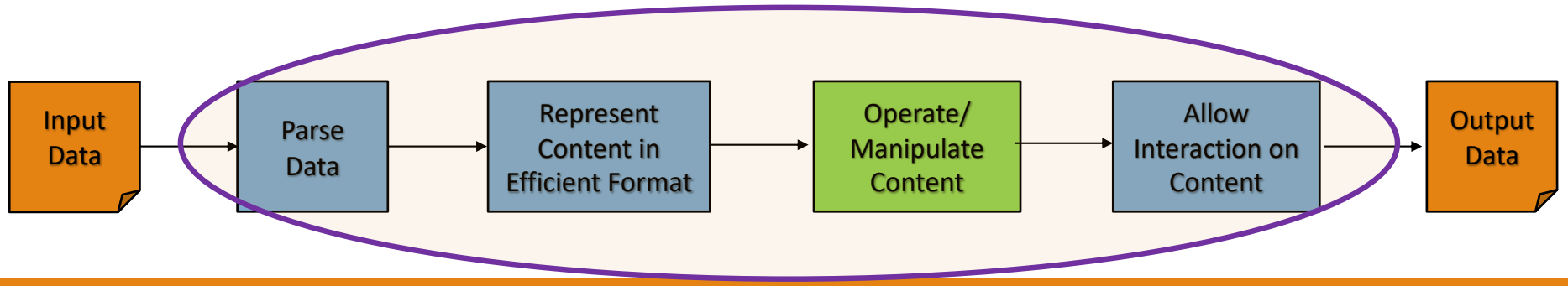
- Introduction Segment
  - Recap from Lecture 9
  - About invited speaker
- Main Segment
  - Invited Talk
- Concluding Segment
  - About next lecture – Lecture 12
  - Ask me anything

Introductory Segment

Recap of Recent Lectures

- Lecture 8
  - We looked at bias definitions: Five categories: C1: predicted outcome, C2: predicted and actual outcome, C3: predicted probabilities and actual outcome , C4: similarity based, C5: causal reasoning
  - Reviewed with respect to German-credit as example; Did role-playing to discuss: for loans, loan applicant, banker and regulator roles
  - Metrics should not only be technically sound but practically useful. Most definitions are theoretical exercises while law catches up; little technical guidance to developers
- Lecture 9
  - We looked at explanation – purpose from psychology literature and an automated explanation - LIME
  - Reviewed guidance on model development from book: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence

Main Segment



About Speaker

- **Diptikalan Saha**, IBM Research
  - https://researcher.watson.ibm.com/researcher/view.php?person=in-diptsaha

Dr. Diptikalan Saha (Dipti) is a Senior Technical Staff Member and manager of Reliable AI team in Data&AI department of IBM Research at Bangalore. His research interest includes Artificial Intelligence, Natural Language Processing, Knowledge representation, Program Analysis, Security, Software Debugging, Testing, Verification, and Programming Languages. received my Ph.D. degree in Computer Science from the State University of New York at Stony Brook. My advisors were Prof. C. R. Ramakrishnan and Prof. Scott A. Smolka. I received my B.E. degree in Computer Science and Engineering from Jadavpur University. His group's work on Bias in AI Systems is available through AI OpenScale in IBM Cloud as well as through open-source AI Fairness 360.

Topics to be Covered

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

Preparatory Reading Material

- Blogs:
  - https://medium.com/@diptikalyan?p=5ce7347f5f75
  - https://www.ibm.com/blogs/watson/2021/06/trustworthy-ai-assessment-mitigation/
  - https://www.ibm.com/blogs/watson/2020/10/how-ibm-makes-ai-based-on-trust-fairness-and-explainability/
- Suveys:
  - Fairness: https://arxiv.org/pdf/1908.09635.pdf
  - Explainability: https://christophm.github.io/interpretable-ml-book/
  - AI Testing: https://www.researchgate.net/publication/334048996_Machine_Learning_Testing_Survey_Landscapes_and_Horizons
  - Counterfactual: https://arxiv.org/abs/2010.10596
- Tools:
  - AIF360: https://aif360.mybluemix.net/
  - AIX360: https://aix360.mybluemix.net/

Concluding Segment

Lectures 10 and 11: Concluding Comments

- We looked at
  - Trustworthy AI: Fairness, Robustness, Explainability, Transparency
  - AI testing: Test, Debug, Repair
  - Bias definitions and their implications
  - Working sessions of IBM tools

About Next Lecture – Lecture 12

Lecture 12:

- Unstructured text – processing
  - Parsing
  - Entity extraction
- Representation