AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

# *CSCE 590-1:* Trusted AI

## Lecture 17 and 18: AI Explanations – Invited Lectures

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

19TH AND 21ST OCT, 2021

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

# Organization of Lectures 17 and 18

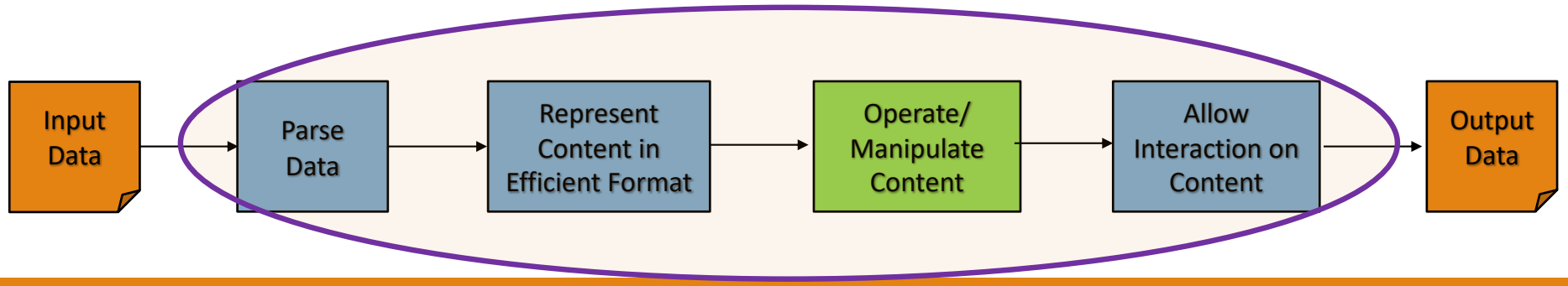
- Introduction Segment
  - Recap from Recent Lectures
  - About invited speaker
- Main Segment
  - Invited Talk – AI Explanation
  - Work Session – AIX 360
- Concluding Segment
  - About next lecture – Lecture 18
  - Ask me anything

# Introductory Segment

# Recap of Recent Lectures

- Lectures 13 -14: Text representation
- Lectures 14-15: Common NLP tasks
- Lecture 16: Text classification
- Trust issues like fairness and transparency of decisions are again relevant.

# Main Segment



| | | |
|---|---|---|
| Oct 19 (Tu) | **Invited Guest** – AI - Supervised ML: External Talk/ AI Explanation Methods (AIX) | 10 am EST |
| Oct 21 (Th) | **Invited Guest** – AI - Supervised ML: External Talk/ Working Session on AIX360 | 10 am EST |

# About Speaker - 1

- **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.

| | | |
|---|---|---|
| Oct 19 (Tu) | **Invited Guest** – AI - Supervised ML: External Talk/ AI Explanation Methods (AIX) | 10 am EST |
| Oct 21 (Th) | **Invited Guest** – AI - Supervised ML: External Talk/ Working Session on AIX360 | 10 am EST |

# About Speaker - 2

- **Vijay Arya**, IBM Research
  - https://researcher.watson.ibm.com/researcher/view.php?person=in-vijay.arya

Vijay Arya is a senior researcher in IBM Research AI at the IBM India Research Lab where he works on problems related to Trusted AI. Vijay has 15 years of combined experience in research and software development. His research work spans Machine learning, Energy & smart grids, network measurements & modeling, wireless networks, algorithms, and optimization. His work has received outstanding technical achievement awards at IBM and has been deployed by power utilities in USA. Before joining IBM, Vijay worked as a researcher at National ICT Australia (NICTA) and received his PhD in Computer Science from INRIA, France, and a Masters from Indian Institute of Technology (IIT) Delhi. He has served on the program committees of IEEE, ACM, and IFIP conferences, he is a senior member of IEEE & ACM, and has more than 60 conference & journal publications and patents.

# 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 17 and 18: Concluding Comments

- We looked at
  - AI Explanation methods
  - Explanations provide transparency and one component of trust
  - Working sessions of IBM tools

# About Next Lecture – Lecture 19

# Schedule Snapshot

| | | |
|---|---|---|
| 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 – NLP Tasks | |
| Oct 14 (Th) | AI – Unstructured (Text): Analysis – Supervised ML | |
| Oct 19 (Tu) | **Invited Guest** – AI - Supervised ML: External Talk/ AI Explanation Methods (AIX) | 10 am EST |
| Oct 21 (Th) | **Invited Guest** – AI - Supervised ML: External Talk/ Working Session on AIX360 | 10 am EST |
| 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<br>Trust: AI Testing | |

# Lecture 19:

- Review of Explanation lectures
- Trust issues with Classification methods
- Quiz 3