

CSCE 590-1: Trusted AI

Lecture 19: AI Unstructured Text - Trust Issues

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

26TH OCT, 2021

Carolinian Creed: “I will practice personal and academic integrity.”

Organization of Lecture 19

- Introduction Segment
 - Recap from invited talks
- Main Segment
 - Review of explanation methods
 - Usage of AIX 360 tool
 - Quiz 3
- Concluding Segment
 - About next lecture – Lecture 20
 - Ask me anything

Introductory Segment

Recap of Lectures 17-18 (Explanation) Lectures

- We looked at explanations for AI in general and IBM point of view
- AIX 360 tool

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

Class 10: Sep 21, 2021: AI Fairness, Diptikalyan Saha

https://lnkd.in/eyJv_XEd

Class 11: Sep 23, 2021: AI Fairness, Diptikalyan Saha

<https://lnkd.in/eJFWdwci>


Class 17: Oct 19, 2021: AI Explainability, Diptikalyan Saha and Vijay Arya

https://lnkd.in/evtSv_5x

Class 18: Oct 21, 2021: AI Explainability, Vijay Arya

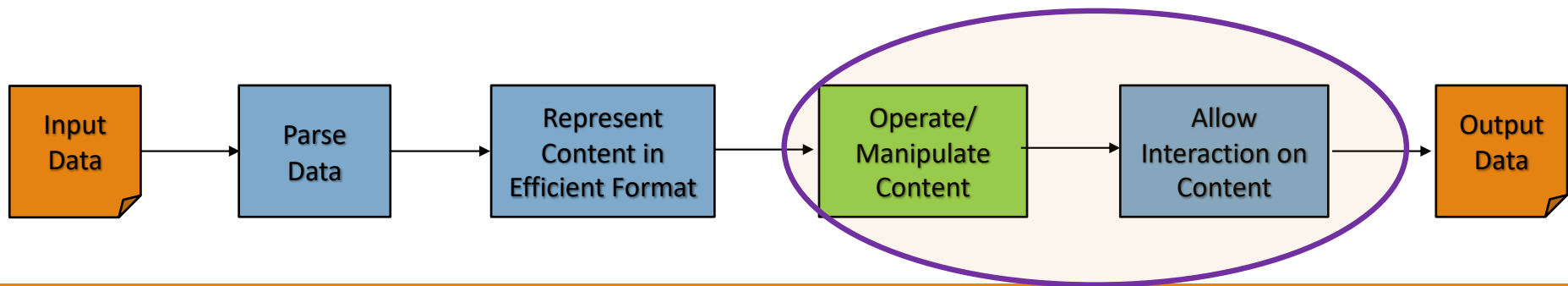
<https://lnkd.in/ee28YVE7>

Schedule Snapshot



Oct 26 (Tu)	Review: Explanation Methods, AIX 360, Discussion	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: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	Quiz 4
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	

Main Segment



What is the Purpose of Explanations (Human or AI)

- Explanation and understanding
 - Frank C Keil, <https://pubmed.ncbi.nlm.nih.gov/16318595/>
- Purposes for explanations in **psychology**
 - To predict similar events in the future: *slippery roads can cause a fall*. Use information later.
 - For diagnosis: *why a system failed and then repair a part to bring it back to its normal function*
 - To affix blame: *for a crime*
 - To justify or rationalize an action: *sweet to an enemy because of the strategic value of being nice on that occasion*
 - In the service of aesthetic pleasure

Source: Lecture 9, Trusted AI, Fall 2021

Terminology Review

- **Decision:** an outcome from an AI/ machine learning model
- **Explanation:** **any information provided** in addition to the decision
- **Interpretable explanation:** an explanation that makes sense to the user **to understand the rationale behind the decision**
- **Intelligible (useful) explanation:** an explanation that helps a user understand the rationale behind the decision **to accept the decision**

Terminology Review - Illustration

- **Decision:** an outcome from an AI/ machine learning model
- **Explanation:** **any information provided** in addition to the decision
- **Interpretable explanation:** an explanation that makes sense to the user **to understand the rationale behind the decision**
- **Intelligible (useful) explanation:** an explanation that helps a user understand the rationale behind the decision **to accept the decision**

*Not a useful explanation
for animal classification*

Decision Tree

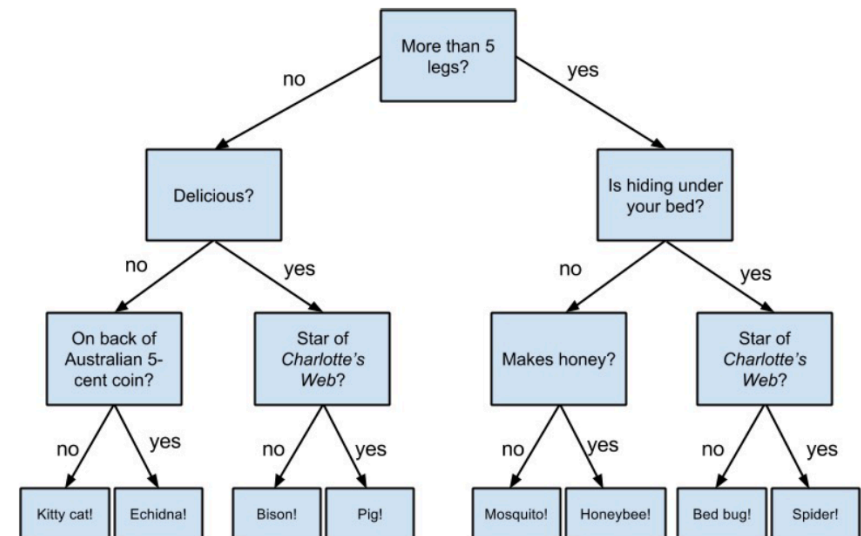
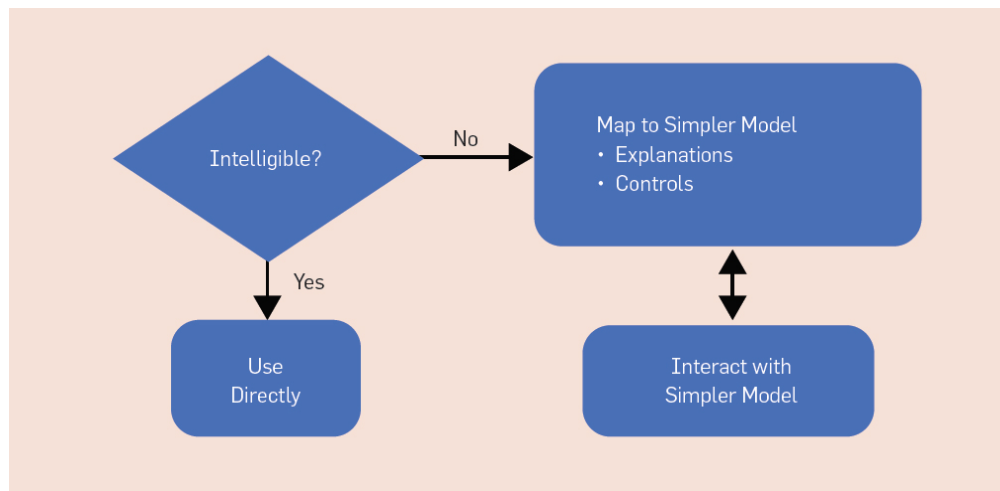


Figure Credit: Diptikalyan Saha and Vijay Arya, Oct 2021

Setting and Terminology: Intelligible Models and Explanations



- Transparency: providing stakeholders with relevant information about how a model works
- Explainability: Providing insights into model's behavior for specific datapoints

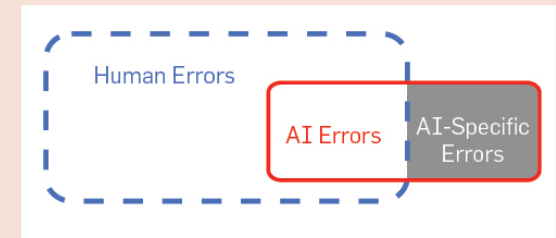
Sources:

1. The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
2. Explainable Machine Learning in Deployment, FAT* 2020.

Source: Lecture 9, Trusted AI, Fall 2021

Need for Intelligibility

The red shape denotes the AI's mistakes; its smaller size indicates a net reduction in the number of errors. The gray region denotes AI-specific mistakes a human would never make. Despite reducing the total number of errors, a deployed model may create new areas of liability (gray), necessitating explanations.



- **AI may have the wrong objective:** is AI right for the right reasons?
- **AI may be using inadequate features:** understand modeling issues
- **Distributional drift:** detect when and why models are failing to generalize
- **Facilitating user control:** guiding what preferences to learn
- **User acceptance:** especially for costly actions
- **Improving human insight:** improve algorithm design
- **Legal imperatives**

Source: Lecture 9, Trusted AI, Fall 2021

Source: The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486

In AI, Stakeholders for Explanations

- Executives
 - Explainability as a market differentiator. Do we need explanations?
- ML engineers
 - How to improve model's performance?
- End-users
 - Understand business decisions emanating from usage of AI
 - Why was my load denied?
 - Why a particular treatment was recommended or de-prioritized ?
- Regulators
 - Prove that you did not discriminate based on existing laws

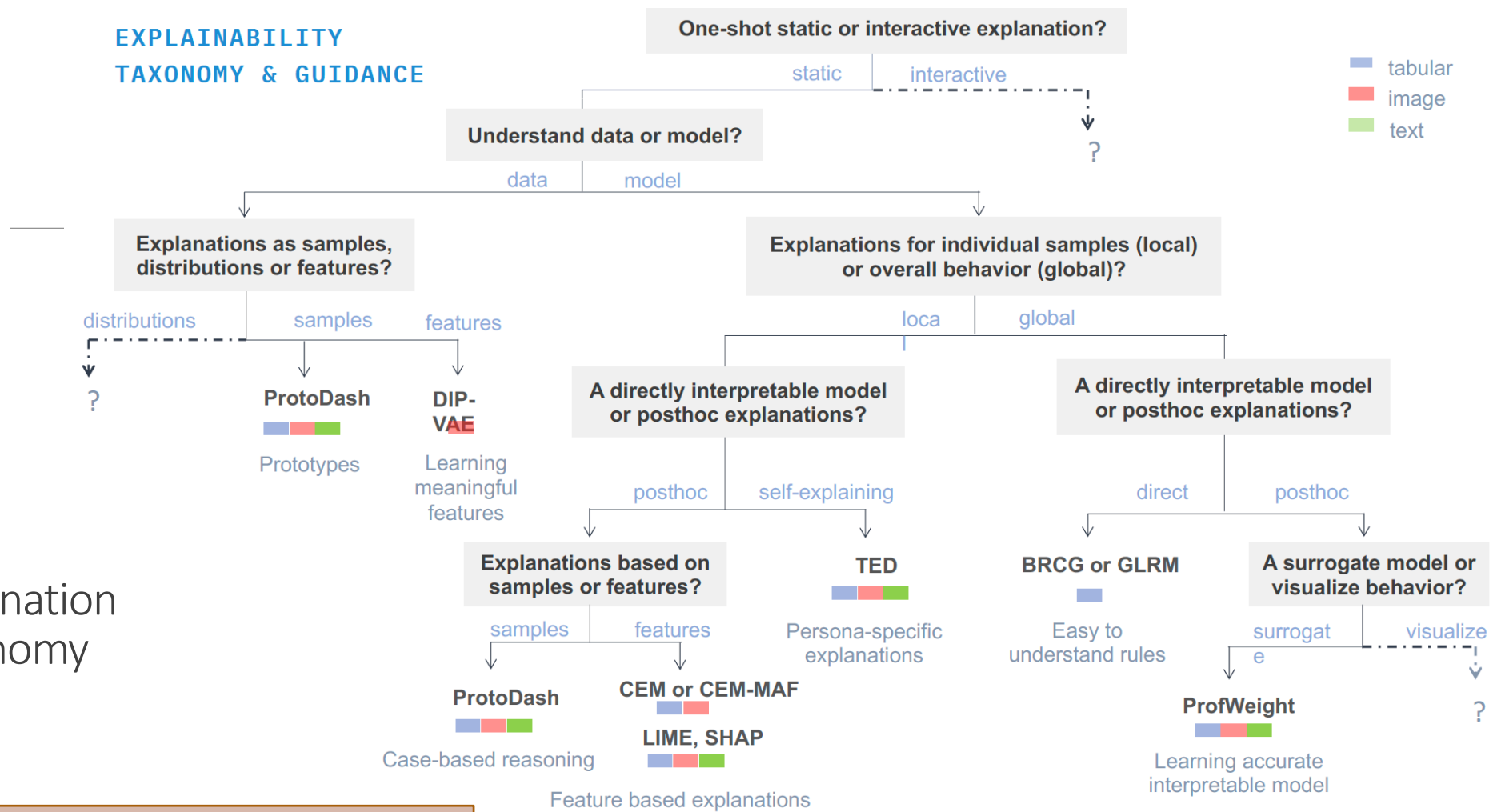
Source: Lecture 9, Trusted AI, Fall 2021

Source: Explainable Machine Learning in Deployment, FAT* 2020,
<https://arxiv.org/pdf/1909.06342.pdf>; Video: <https://www.youtube.com/watch?v=Hofl4uwxtPA>

Types of Explanations

- **Feature-based**: from the features of the data, which feature(s) were most important for given decision output
 - Example: For a loan, is it income or the person's age ?
- **Sample-based**: from data in training, which data points were important for given test point; helps understand sampling and its representation in wider population
 - Example: For a loan, what instances similar to the loan application would have gotten the loan ?
- **Counter-factual**: what-ifs – what do you change about the input to change the decision output
 - Example: For a loan, does getting an additional borrower insurance increase chance of getting the loan?
- Natural language

EXPLAINABILITY TAXONOMY & GUIDANCE



Explanation
Taxonomy

Figure Credit: Diptikalyan Saha and Vijay Arya, Oct 2021

Explanation Methods Covered

- Local Explainability – LIME, ANCHORS
- Global Explainability – TREPAN
- Counterfactual Explainability – Wachter
- HELOC / Credit Rating
 - Loan Officer: ProtoDash - Prototypical explanations
 - Data scientist: Boolean Rule and Logistic Rule Regression models
 - Customer: Contrastive explanations - Pertinent positives and negatives

AIX Tool

- HELOC: <https://github.com/Trusted-AI/AIX360/blob/master/examples/tutorials/HELOC.ipynb>

Concluding Segment

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/>

Lecture 19: Concluding Comments

- We looked at

Review of What We Covered


- Perspective
 - Explanation, types
 - Usage, Taxonomy
- Methods
 - Overview of Three types of Explainability
 - Local Explainability – LIME, ANCHORS
 - Global Explainability – TREPAN
 - Counterfactual Explainability – Wachter
 - Others in AIX tutorials
- Tool
 - Notebook walkthrough of Algorithms in AIX360

Adapted from: Diptikalyan Saha and Vijay Arya, Oct 2021

Quiz 3: Explanation

About Next Lecture – Lecture 20

Schedule Snapshot



Oct 26 (Tu)	Review: Explanation Methods, AIX 360, Discussion	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: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	Quiz 4
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	

Lecture 20:

- Review of Projects
- Use 1-slide template, upload to shared Google drive
 - Template and link shared via slack