CSCE 590-1: Trusted AI

Lecture 9: AI: Supervised ML / Trust / Explanations

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

16TH SEP, 2021

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

Organization of Lecture 9

- Introduction Segment
  - Recap from Lecture 8
  - Project discussion
- Main Segment
  - Explanation
  - Explanation Method: LIME
  - Book: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence
- Concluding Segment
  - About next lecture – Lecture 10
  - Ask me anything

Introductory Segment

Recap of 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

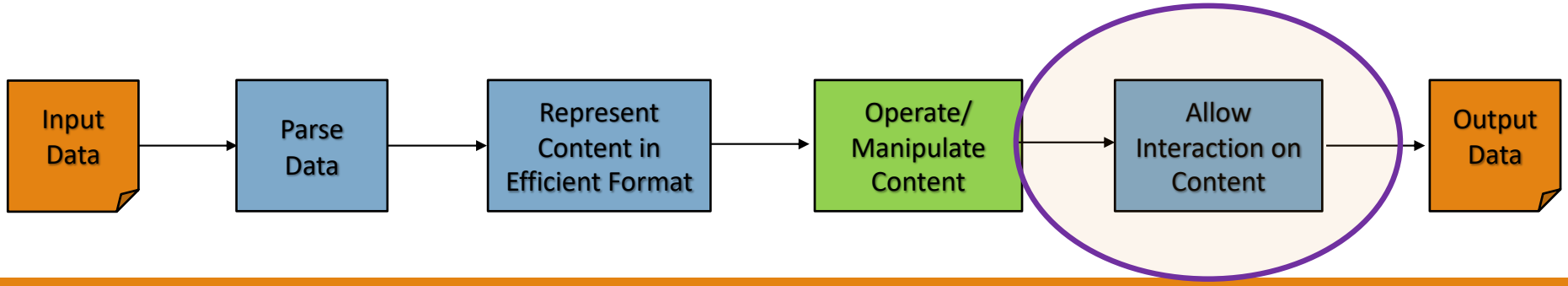
- Metrics should not only be technically sound but practically useful
  - Did role-playing to discuss: for loans, loan applicant, banker and regulator roles
  - We can consider for the class: coding language (Python, Java, …) domain. Roles: Student, instructor, university ?

- Most definitions are theoretical exercises while law catches up; little technical guidance to developers

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

Main Segment



Generating Explanations

What is the Purpose of Explanations

- 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

AI Explainability

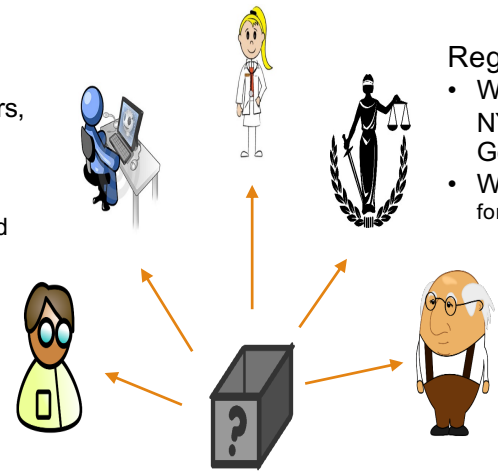
Meaningful explanations depend on the explanation consumer

he General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing an profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

End Users
- Who: Physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights(

AI System builders, stakeholders
- Who: data scientists, developers, prod mgrs
- Why: ensure/improve performance



Regulatory Bodies
- Who: EU (GDPR), NYC Council, US Gov't, etc
- Why: ensure fairness for constituents

Affected Users
- Who: Patients, accused, loan applicants, teachers
- Why: understanding of factors

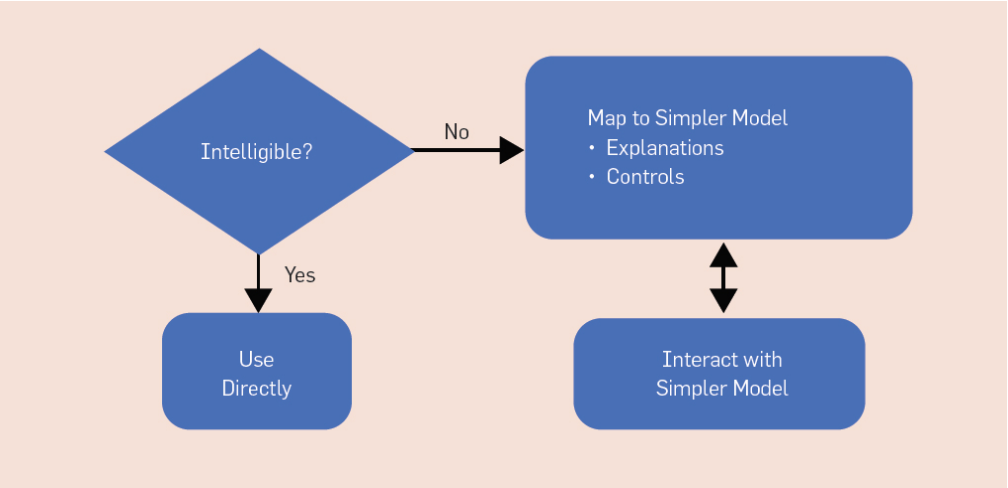
Must match the complexity capability of the consumer
Must match the domain knowledge of the consumer

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: Explainable Machine Learning in Deployment, FAT* 2020

Setting and Terminology: Intelligible Models and Explanations

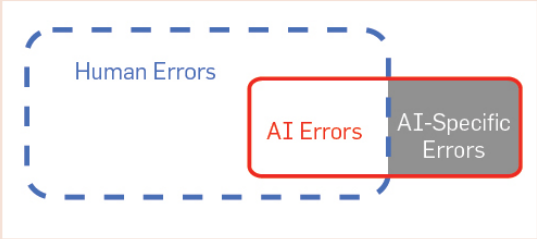


- Transparency: providing stakeholders with relevant information about how a model works
- Explainability: Providing insights into model's behavior for specific datapoints

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.

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: 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

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

Source: Explainable Machine Learning in Deployment, FAT* 2020

References for AI Explainability

Papers

- 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
- "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, in ACM's Conference on Knowledge Discovery and Data Mining, KDD2016; https://homes.cs.washington.edu/~marcotcr/blog/lime/, https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/
- Explainable Machine Learning in Deployment, FAT* 2020, https://arxiv.org/pdf/1909.06342.pdf; Video: https://www.youtube.com/watch?v=Hofl4uwxtPA

Tutorial: XAI tutorial at AAAI 2020, https://xaitutorial2020.github.io/

Tool: AIX 360

Tool: https://aix360.mybluemix.net/

Video: https://www.youtube.com/watch?v=Yn4yduyoQh4

Paper: https://arxiv.org/abs/1909.03012

LIME – Local Interpretable Model-Agnostic Explanations

Paper: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM's Conference on Knowledge Discovery and Data Mining, KDD2016

Blogs:

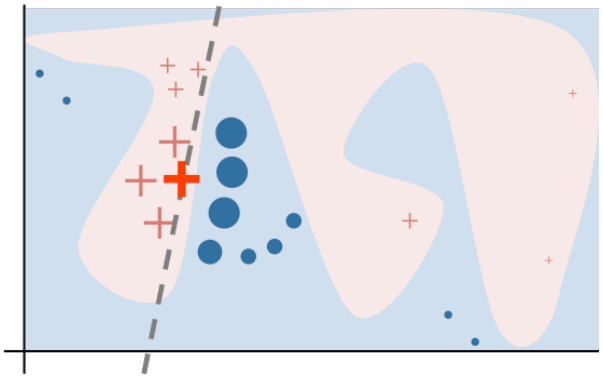
- https://homes.cs.washington.edu/~marcotcr/blog/lime/
- https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Code: https://github.com/marcotcr/lime

Figures credit: Marco Tulio Ribeiro

LIME Key Idea

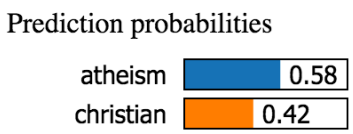
- Generate a local, linear explanation of any model
- How
  - Perturb near the neighborhood of a point of interest, X (**Local**)
  - Fit a linear function to the model's output (**Linear**)
  - Interpret coefficients of the linear function (**Explain**)
  - **Visualize**

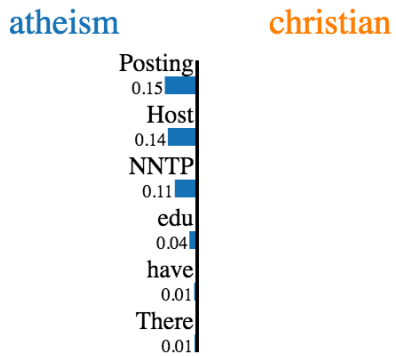


LIME on Text

Question: Why is a classifier with >90% accuracy predicting based on ?

"If we remove the words **Host** and **NNTP** from the document, we expect the classifier to predict **atheism** with probability 0.58 - 0.14 - 0.11 = 0.31"





Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Source: https://github.com/marcotcr/lime

Code Examples

- LIME
  - Iris dataset and supervised classifiers – random forest and logistic regression, rabular data: https://github.com/biplav-s/course-tai/blob/main/sample-code/l9-explanations/LIME%20explanations%20on%20tabular%20data.ipynb

- Many other examples
  - https://github.com/biplav-s/course-d2d-ai/tree/main/sample-code/l12-explanability-autoai

LIME on Image

Question: Why is this a frog?

Divide image into interpretable components - contiguous superpixels

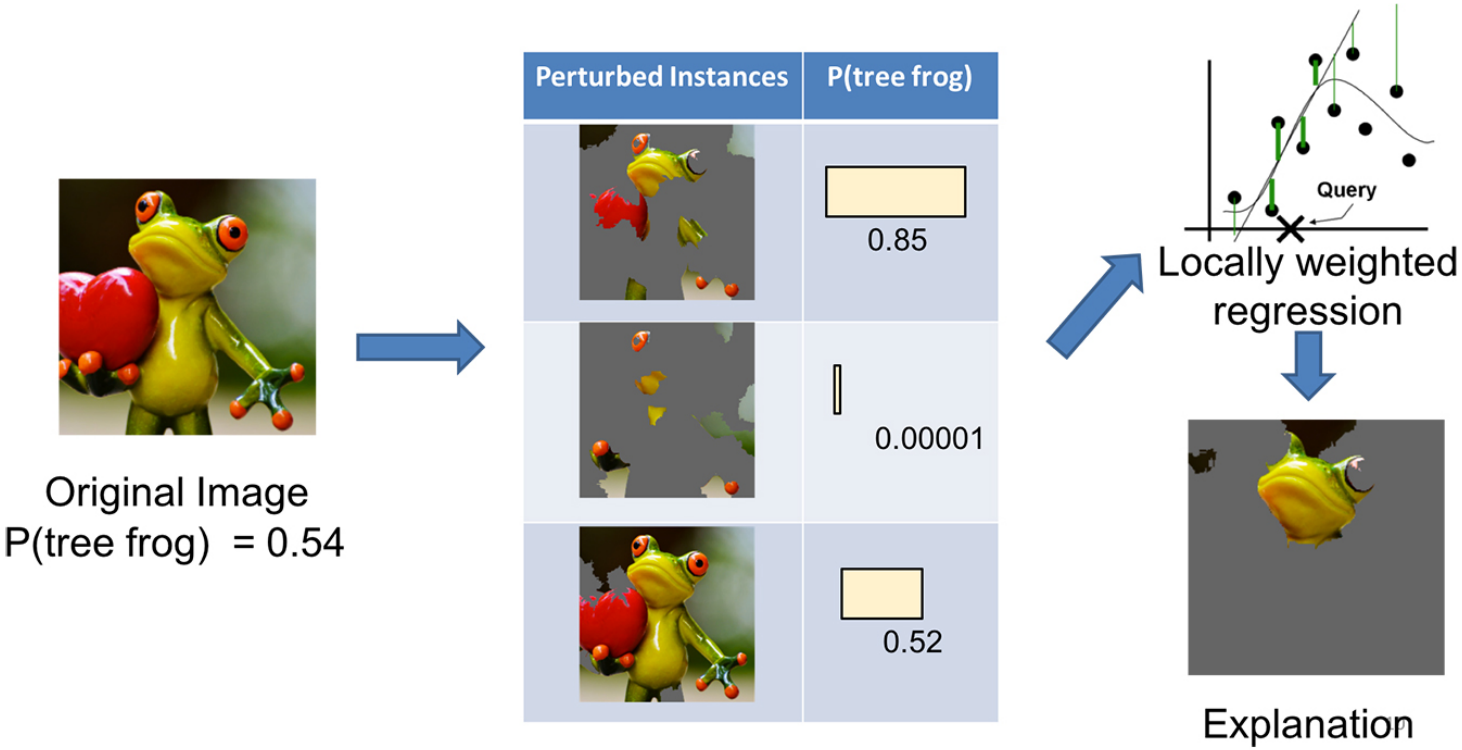
Original Image

Interpretable Components

Source: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

LIME

1. Generate a data set of perturbed instances by turning some of the interpretable components “off” (gray).
2. For each perturbed instance, calculate probability that a tree frog is in the image according to the model.
3. Learn a simple (linear) model on this data set, which is locally weighted
4. Output regions with highest positive weights as an explanation, graying out everything else.



Source: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Explanation and Practical Implications

- Context
  - Problem: detect common cardiovascular conditions
  - Data: ECG data
  - Explanation: LIME
- References
  - Blog: https://www.ucsf.edu/news/2021/08/421301/ai-algorithm-matches-cardiologists-expertise-while-explaining-its-decisions
  - Paper: https://jamanetwork.com/journals/jamacardiology/article-abstract/2782549

Many Explanation Methods

- Review paper on many methods and data types (image, text, audio, and sensory domains):
  - **How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods,** *Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani Srivastava,* Advances in Neural Information Processing Systems 33 (NeurIPS 2020), https://proceedings.neurips.cc/paper/2020/hash/2c29d89cc56cdb191c60db2f0bae796b-Abstract.html

Handbook on Data Protection and Privacy for Developers of Artificial Intelligence

- Review .pdf

Concluding Segment

Lecture 9: Concluding Comments

- We looked explanation method
  - LIME

- Book: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence

About Next Lecture – Lecture 10

Lecture 10: Guest Speaker – Techniques from IBM

- Modified Logistics
  - Time: 10:00 – 11:15 am
  - Location: Seminar room, AI Institute – 5th floor at 1112 Greene St., Columbia; Science and Technology Building
  - Speaker to join over Blackboard

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

Sep 21 (Tu)	**Invited Guest – AI - Supervised ML: External Talk/ IBM's Approach**
Sep 23 (Th)	**Invited Guest – AI - Supervised ML: External Talk/ Working Session**