AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 581: Introduction to Trusted AI

Lectures 15 and 16: Mitigations, LLMs, Project Update #2

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

4TH AND 6TH MAR, 2025

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

Credits: **Copyrights of all material reused acknowledged**

Organization of Lectures 15, 16

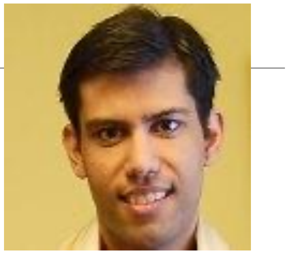
- Introduction Section
  - Recap from Week 7 (Lectures 13 and 14)
  - Announcements and News
- Main Section
  - L15: LLMs, Explanation
  - L16: Project Update #2
- Concluding Section
  - About next week/ non-holiday – Lectures 17, 18
  - Ask me anything

Introduction Section

Recap from Week 7 (Lectures 13, 14)

- We looked at
  - Explanations – LIME method
  - Transparency through documentation – Rating – ARC tool

Announcement: Guest Speaker



Human-Centered Trustworthy Agentic AI, by Dr. Kush Varshney

March 18, 2025, 4:30-5:30pm, Zoom Meeting - https://sc-edu.zoom.us/j/83709414377

Dr. Varshney is an IBM Fellow based at the Thomas J. Watson Research Center, Yorktown Heights, NY, where he directs Human-Centered Trustworthy AI research. He applies data science and predictive analytics to human capital management, healthcare, olfaction, computational creativity, public affairs, international development, and algorithmic fairness, which has led to the Extraordinary IBM Research Technical Accomplishment for contributions to workforce innovation and enterprise transformation, and IBM Corporate Technical Awards for Trustworthy AI and for AI-Powered Employee Journey. He and his team created several well-known open-source toolkits, including AI Fairness 360, AI Explainability 360, Uncertainty Quantification 360, and AI FactSheets 360. AI Fairness 360 has been recognized by the Harvard Kennedy School's Belfer Center as a tech spotlight runner-up and by the Falling Walls Science Symposium as a winning science and innovation management breakthrough.

He independently-published a book entitled 'Trustworthy Machine Learning' in 2022, available at http://www.trustworthymachinelearning.com. He is a fellow of the IEEE.
Personal website: https://krvarshney.github.io

AI News

- AAAI conference
  - Report: AAAI 2025 Presidential Panel on the Future of AI Research: 17 topics, each with sketching its history, current trends and open challenges; contains insights from both expert and survey respondents, https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf
  - Teaching award
    (Prof. Subbaro Kambhampati, 2025;
    Profs. Michael Littman and Charles Isbell, 2024)

Key Insights

Valuable Reading

- AAAI 2025 Presidential Panel on the Future of AI: 17 topics related to AI research, each chapter sketching its history, current trends and open challenges. Has insights both from experts and survey respondents. https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf

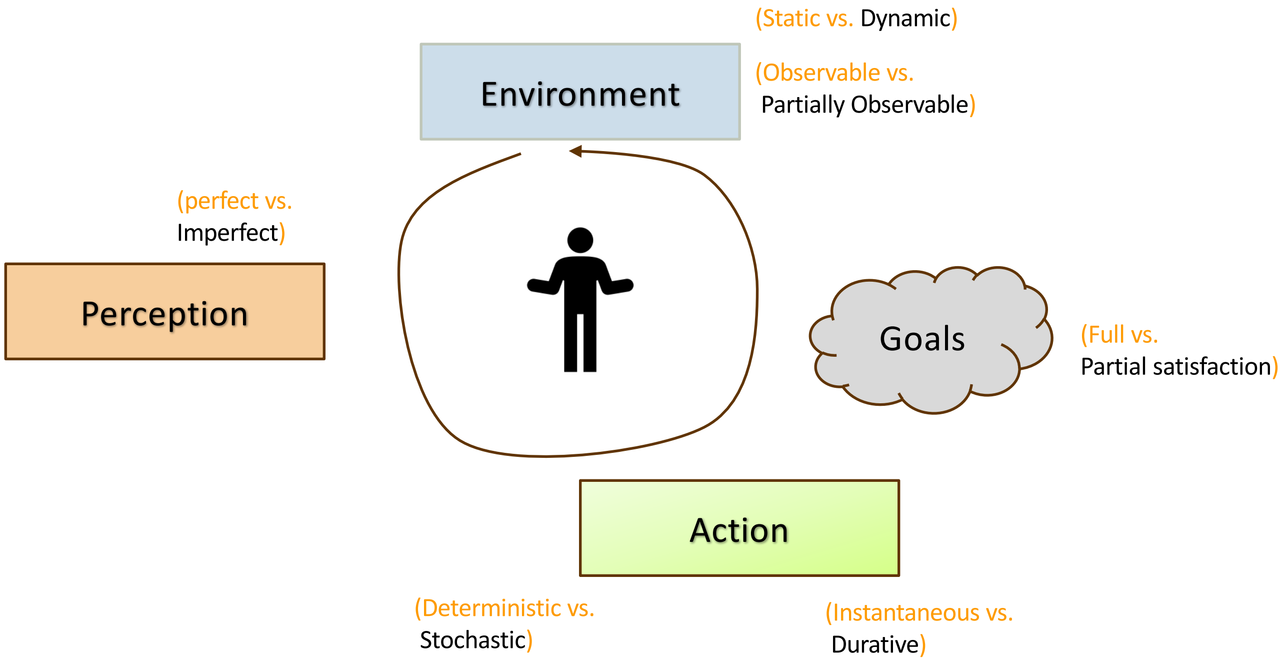
Announcement: Change to Student Assessment

A	=	[920-1000]
B+	=	[870-919]
B	=	[820-869]
C+	=	[770-819]
C	=	[720-769]
D+	=	[670-719]
D	=	[600-669]
F	=	[0-599]

Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



(Static vs. Dynamic)

Environment

(Observable vs. Partially Observable)

(perfect vs. Imperfect)

Perception

Goals

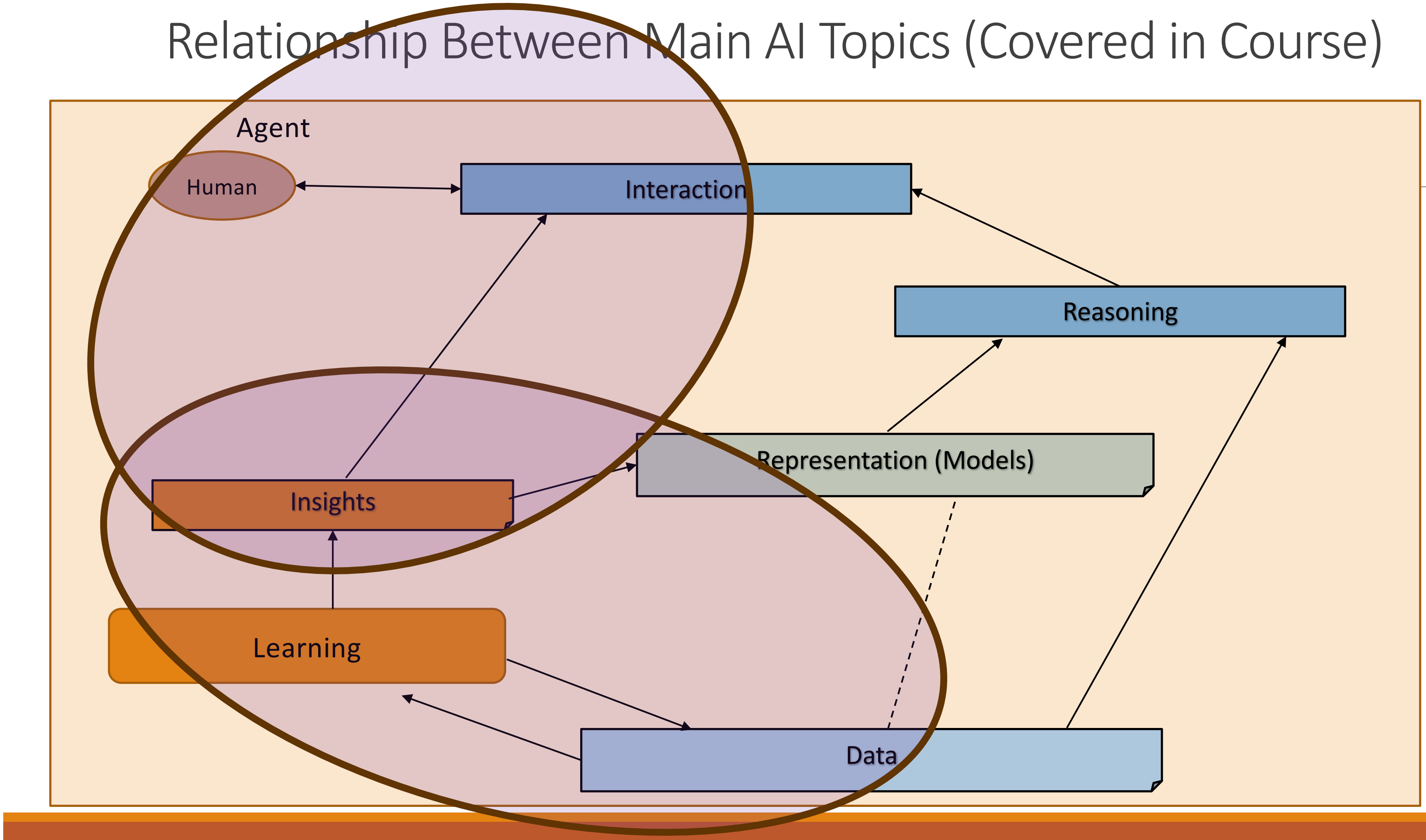
(Full vs. Partial satisfaction)

Action

(Deterministic vs. Stochastic)

(Instantaneous vs. Durative)

Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

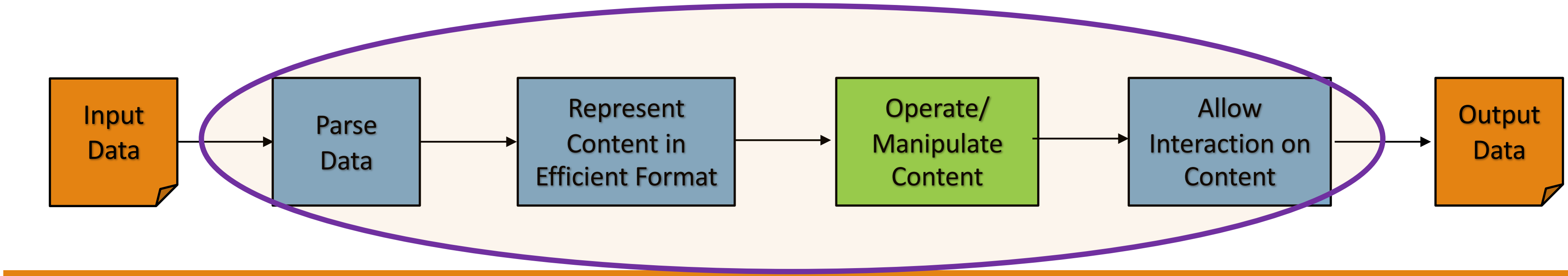
CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification, **Large Language Models**
- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

Increased focus on LLMs and projects now

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

Main Segment


Input Data
Parse Data
Represent Content in Efficient Format
Operate/ Manipulate Content
Allow Interaction on Content
Output Data

InterpretML

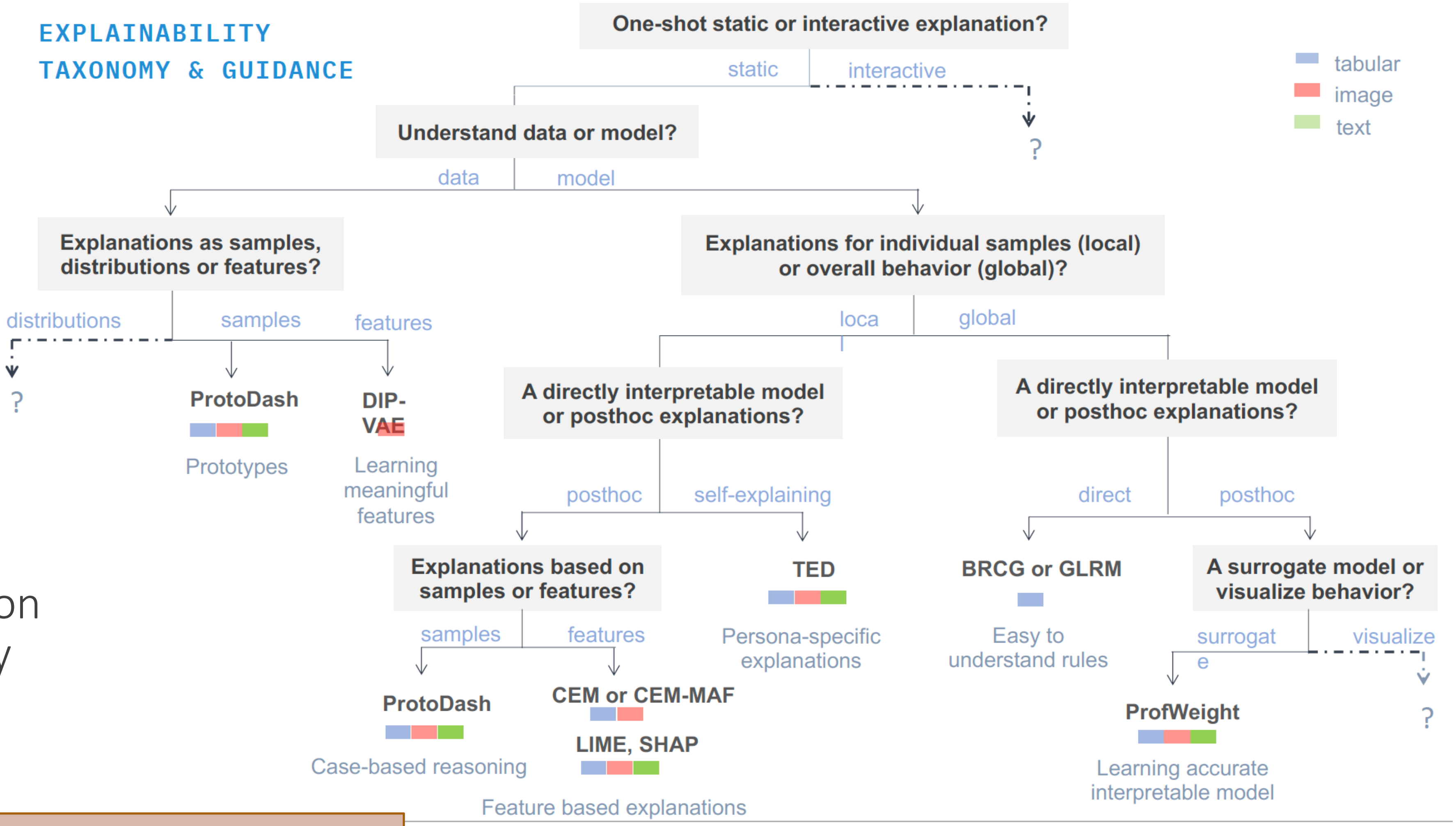
- **Details**: https://github.com/interpretml/interpret
  - Whitebox (Glassbox) models: change learning code to introduce explainability support
  - Blackbox models: don’t change learning code

Interpretability Technique	Type
Explainable Boosting	**glassbox model**
APLR	glassbox model
Decision Tree	glassbox model
Decision Rule List	glassbox model
Linear/Logistic Regression	glassbox model
SHAP Kernel Explainer	blackbox explainer
LIME	**blackbox explainer**
Morris Sensitivity Analysis	blackbox explainer
Partial Dependence	blackbox explainer

InterpretML – Sample Code

Github: https://github.com/biplav-s/course-ai-f24/blob/main/sample-code/l21-explainability/ExploreInterpreatbilityPackageMS.ipynb

Explanation Taxonomy



EXPLAINABILITY
TAXONOMY & GUIDANCE

One-shot static or interactive explanation?

static

interactive

?

🟦 tabular

🟥 image

🟩 text

Understand data or model?

data

model

Explanations as samples, distributions or features?

distributions

samples

features

?

ProtoDash

🟦🟥🟩

Prototypes

DIP-VAE

🟥

Learning meaningful features

Explanations for individual samples (local) or overall behavior (global)?

loca
l

global

A directly interpretable model or posthoc explanations?

posthoc

self-explaining

Explanations based on samples or features?

samples

features

ProtoDash

🟦🟥🟩

Case-based reasoning

CEM or CEM-MAF

🟦🟥

LIME, SHAP

🟦🟥🟩

Feature based explanations

TED

🟦🟥🟩

Persona-specific explanations

A directly interpretable model or posthoc explanations?

direct

posthoc

BRCG or GLRM

🟦

Easy to understand rules

A surrogate model or visualize behavior?

surrogat
e

visualize

ProfWeight

🟦🟥🟩

Learning accurate interpretable model

?

Figure Credit: Diptikalyan Saha and Vijay Arya, Oct 2021

Class 15: Explanation

Methods

- LIME:
  - Tools: LIME, InterpretML
- SHAP:
  - Tools: SHAP, ExplainerBoard

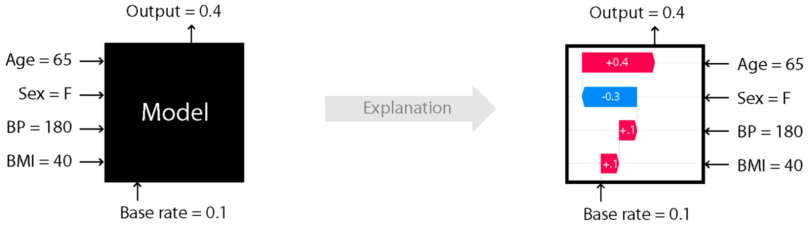
SHAP

- Intuition

SHAP

SHAP (SHapley Additive exPlanations)

Image Credit: https://shap.readthedocs.io/en/latest/index.html



- Features with positive SHAP values positively impact the prediction,
- Negative values have a negative impact.
- Magnitude is a measure strength of effect
- +: Numbers add up to one
- -: Numbers (coefficients) depend on the unit of quantity being measured
- +: Model agnostic
- +: Additive: contribution of each feature to the final prediction can be computed independently and then summed up

- Details and example:
  - https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html
  - https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

GitHub Code

- LIME, SHAP: https://github.com/biplav-s/course-tai-s25/blob/main/sample-code/Demo%20LIME%20and%20SHAP.ipynb
- SHAP/ ExplainerBoard: https://github.com/biplav-s/course-tai-s25/blob/main/sample-code/ExplainerBoard%20Demo.ipynb

Class 15: LLMs

Language Model

Problem:
Given a sentence fragment, predict what word(s) come next

Applications:

- Spelling correction
- speech recognition
- machine translation,
- …

Language Model:
estimate probability of substrings of a sentence

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) = \frac{P(w_1, w_2, \ldots, w_{i-1}, w_i)}{P(w_1, w_2, \ldots, w_{i-1})}$$

Bigram approximation

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \frac{P(w_{i-1}, w_i)}{P(w_{i-1})}$$

From Jurafsky & Martin

Language Model

Markovify library
https://github.com/jsvine/markovify

Language Model:
estimate probability of substrings of a sentence

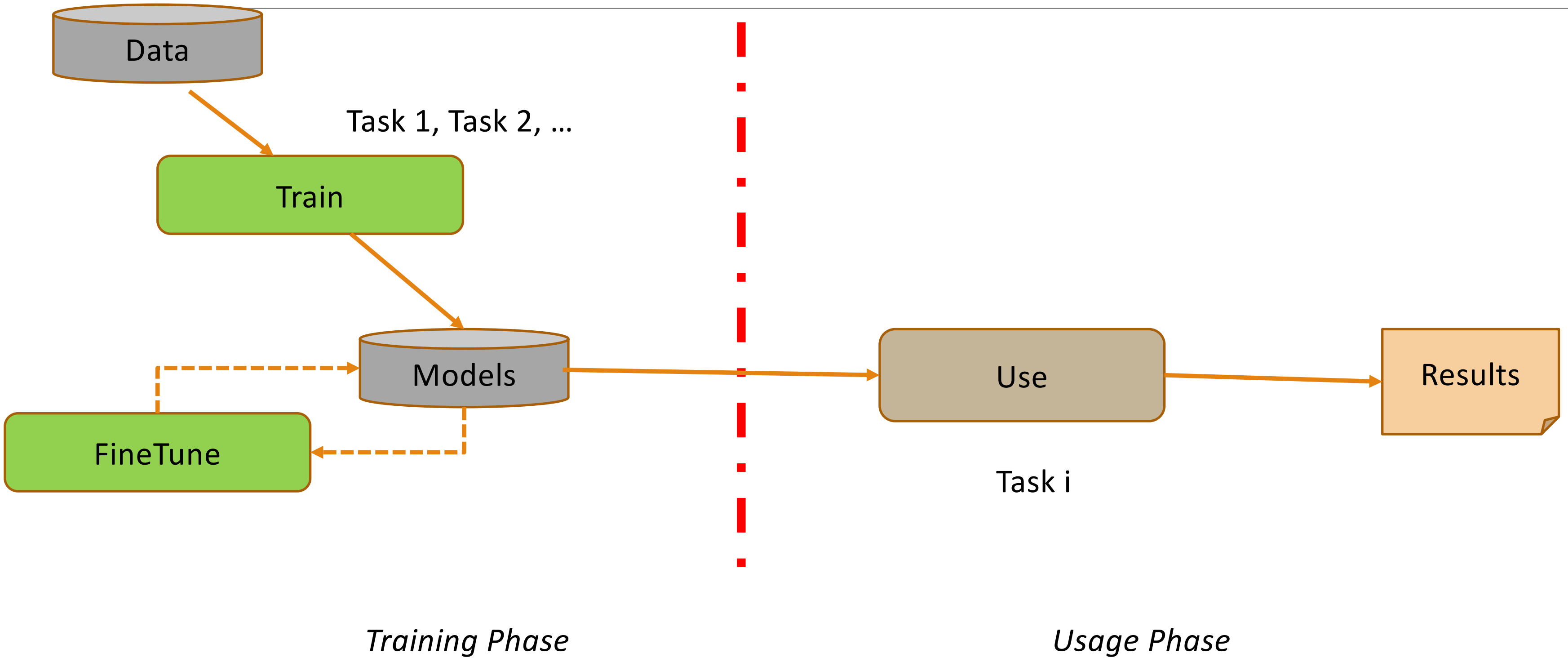
$$P(w_i|w_1, w_2, \ldots, w_{i-1}) = \frac{P(w_1, w_2, \ldots, w_{i-1}, w_i)}{P(w_1, w_2, \ldots, w_{i-1})}$$

See code samples with Markovify library on Github

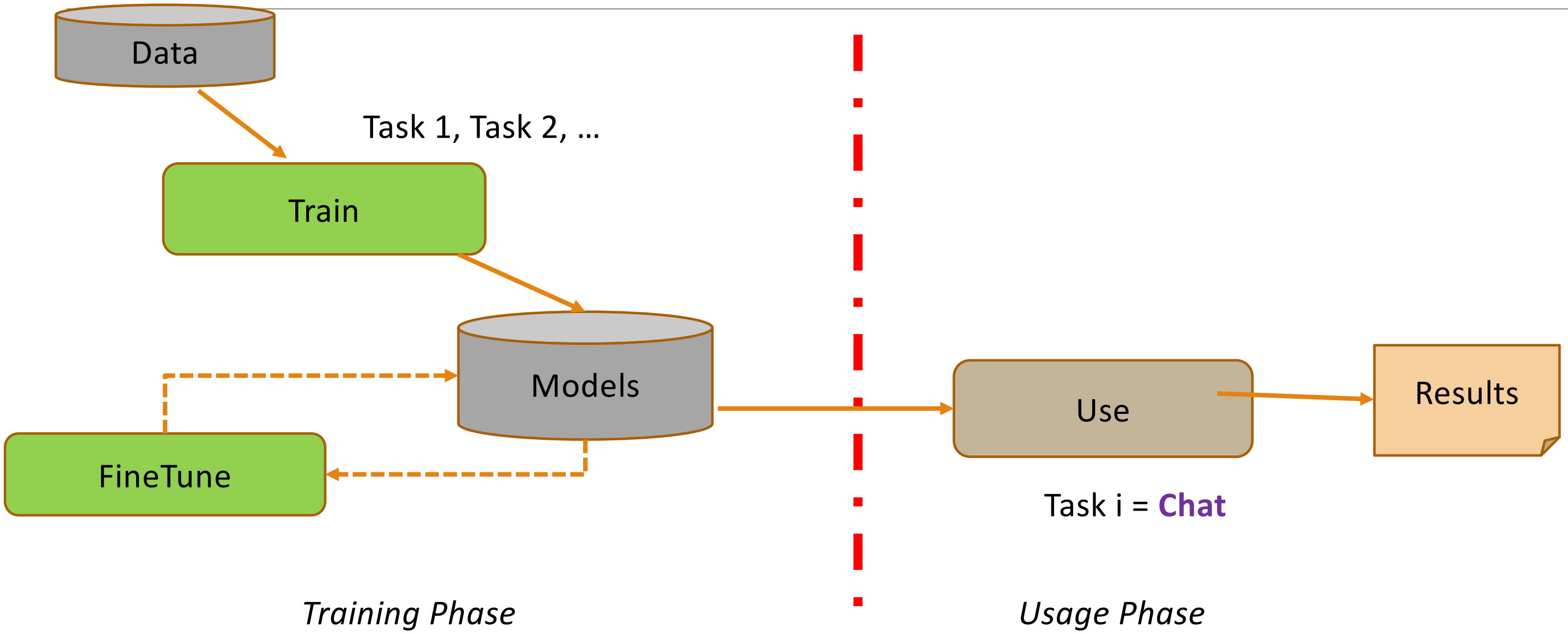
- *Prepare data – two datasets shown*
- *Try generator:*
  - *https://github.com/biplav-s/course-nl/blob/master/l7-language/code/TryMarkovifyLangModel.ipynb*

Large LMs (LLMs)

Large Language Models (LLMs) Basics



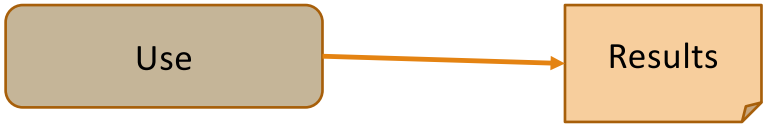
ChatGPT: Large Language Models (LLMs) based Chatbot



Another “Turning Point” Moment In Technology

Raised interest about Chatbots among public

- Excitement about new use-cases
- Concerns about social impact – cheating, jobs, misinformation
- Renewed calls for regulations



Task i = **Generally speaking: content generation – text, image, video, audio, …**

Usage Phase

BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Learns with two tasks

- Predicting missing words in sentences
  - mask out 15% of the words in the input, predict the masked words.
- Given two sentences A and B, is B the actual next sentence that comes after A, or just a random sentence from the corpus?

Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon

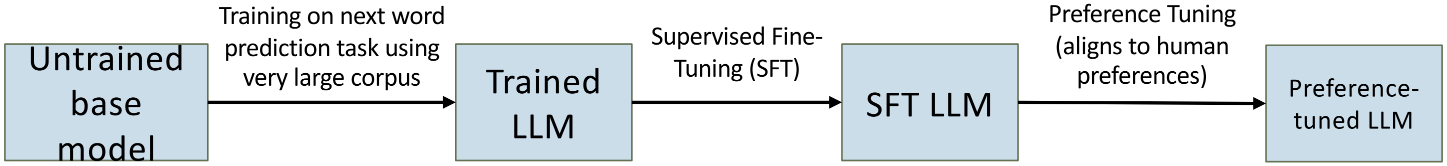
Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence

Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence

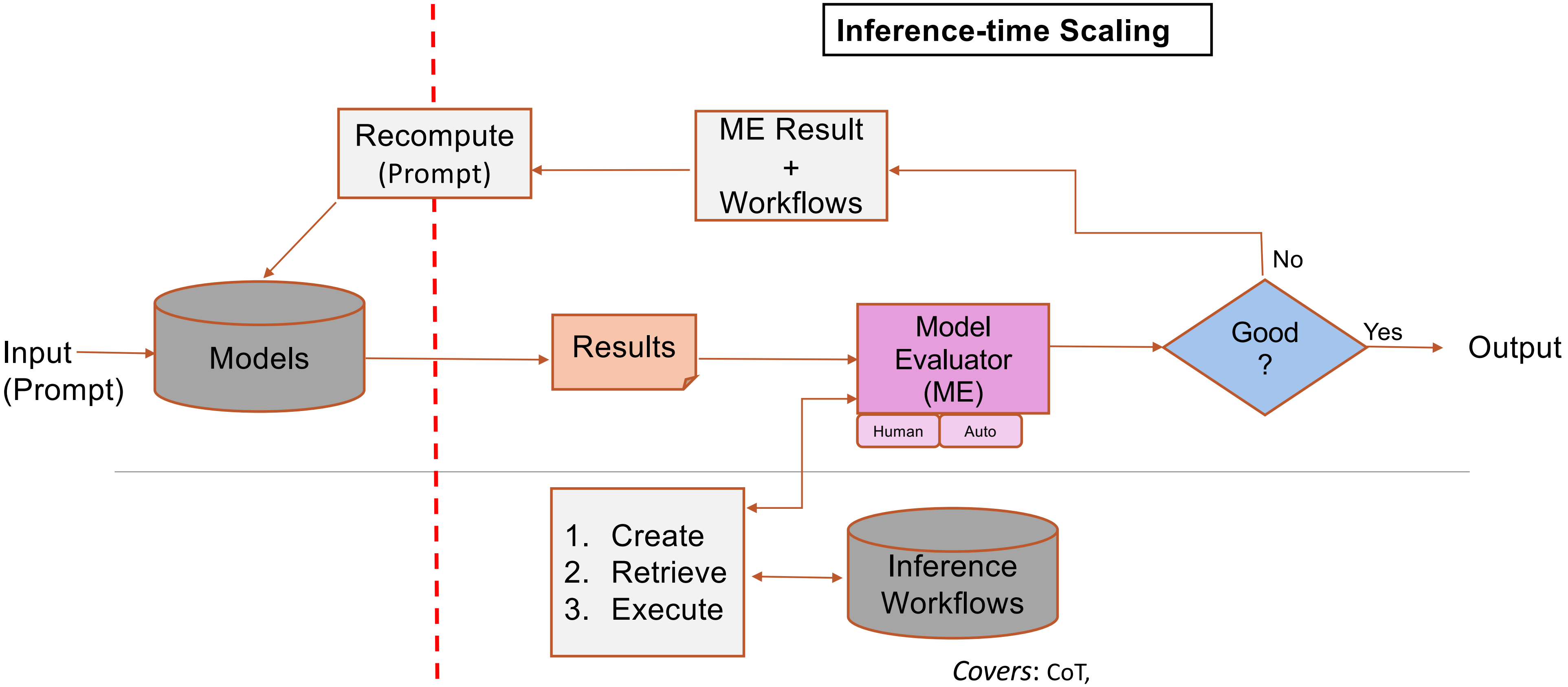
(12-layer to 24-layer Transformer)
on (Wikipedia + BookCorpus)

Credit and details: https://github.com/google-research/bert

LLM (Fine-tuning) Training Procedure


Untrained base model
Training on next word prediction task using very large corpus
Trained LLM
Supervised Fine-Tuning (SFT)
SFT LLM
Preference Tuning (aligns to human preferences)
Preference-tuned LLM

Inference Time with LLMs



Covers: CoT, ToT, …RAG

LLM/ FM Tools

Using LLMs/FMs

- Choose a LLM/FM
  - Open-source: Llama, Mistral, DeepSeek, Bloom, …
  - Closed-source: GPT, Gemini, Claude, …
- API interface
  - Huggingface
  - Ollama
- Use chat interface
  - POE
  - ChatGPT
  - DeepSeek
  - …

Class 16: Project Update #2

Extra Task for Update #2

- Give project input to a LLM
- Collect its result
- Discuss
  - Is the result a good baseline?
  - Is your solution beating the baseline?
  - Where do you go from here?

Project Discussion

Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
     1. Solution => ML task (e.g. classification), recommendation, text summarization, …
     2. Use a foundation model (e.g., LLM-based) solution as the baseline
  4. (Data) Explore the data for a solution to work
  5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Project Discussion

1. Create a private Github repository called "CSCE581-Spring2025-<studentname>-Repo". Share with Instructor (biplav-s)

2. Create a folder called "Project". Inside, create a text file called "ProjectPlan.md" (or "ProjectPlan.txt") and have details by the next class (Jan 30, 2025)

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

Concluding Section

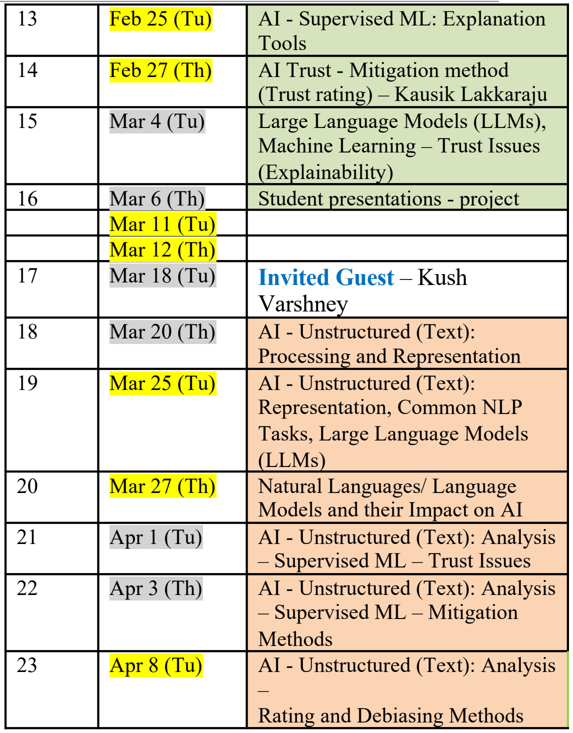
Week 8 (L15 and 16): Concluding Comments

- We looked at
  - Revised explanation methods
  - Did an overview of LLM/ FM basics and tools
  - Reviewed projects, especially in the context of a LLM/FM

About Next Week – Lectures 17, 18

Lectures 17, 18:

- Invited talk
- Text processing



13	Feb 25 (Tu)	AI - Supervised ML: Explanation Tools
14	Feb 27 (Th)	AI Trust - Mitigation method (Trust rating) – Kausik Lakkaraju
15	Mar 4 (Tu)	Large Language Models (LLMs), Machine Learning – Trust Issues (Explainability)
16	Mar 6 (Th)	Student presentations - project
	Mar 11 (Tu)	
	Mar 12 (Th)	
17	Mar 18 (Tu)	**Invited Guest** – Kush Varshney
18	Mar 20 (Th)	AI - Unstructured (Text): Processing and Representation
19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods