Learning Symbols for Trustworthy AI

Rajeev Alur
University of Pennsylvania

August 2025

Penn Engineering | ASSET Center
AI-enabled Systems: Safe, Explainable, and Trustworthy

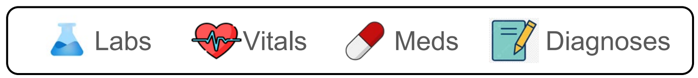
The Promised Land of Generative AI

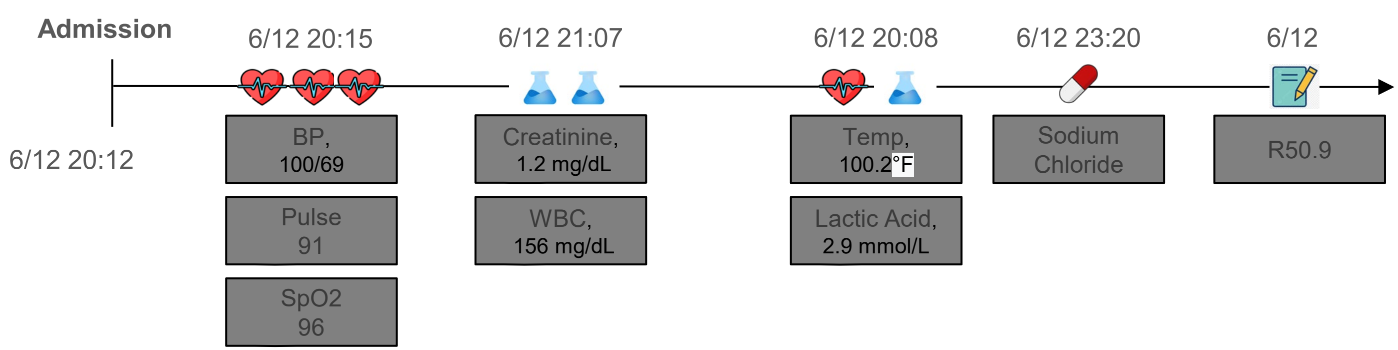
Patient Info

Age: 53.29
Race: Asian
Sex: Female
Zip: 19104

Hospital Visit Info

Admitting Code: R10.11
Principal Code: A40.1





Computational problem: predict whether the patient will develop Sepsis

Solution: Just ask an LLM!

Seek not just the answer, but also why

Prompt

You are a medical AI assistant focused on the early detection of sepsis in Emergency Department (ED) patients. Your primary task is to analyze provided Electronic Health Record (EHR) data to assess the likelihood of sepsis development within the next 90 minutes from the last recorded data point.

The input will be a chronologically ordered sequence of real-time patient data from ED admission. Each entry follows the format: ⟨time_in_mins⟩: ⟨data⟩, where ⟨time_in_mins⟩ denotes minutes since admission. The data will include patient demographics, triage vitals, subsequent real-time vitals, laboratory results, and medications.
Pay close attention to trends and critical changes in the following parameters, as they relate to general sepsis indicators:

- Vital Signs: Heart Rate (HR), Respiratory Rate (RR), Blood Pressure (BP, MAP), Temperature, Oxygen Saturation (SpO2).
- Laboratory Results: White Blood Cell Count (WBC), Lactate levels, Creatinine, Bilirubin, Platelet Count.
- Clinical Assessment: Altered Mental Status (e.g., changes in Glasgow Coma Scale - GCS), evidence of organ hypoperfusion (e.g., skin changes, urine output).
- Medications: Administration of vasopressors, intravenous fluids, and antibiotics.

To arrive at your prediction, follow these reasoning steps:

- Establish Patient Context: ...
- eSOFA Component Assessment: ...
- Risk Synthesis & Prediction: ...

Your final output MUST strictly adhere to the following format:
Reasoning: 2-3 paragraphs based on the above steps
Prediction: Yes/No
Risk: [0, 1)
'Yes' indicates a high risk of sepsis development within the next 90 minutes, 'No' indicates a low risk, and 'Risk' is a numerical value between 0 (no risk) and 1 (high risk). For example, "Prediction: Yes, Risk: 0.95", indicates a high risk of sepsis. Base your analysis solely on the provided EHR text and your embedded medical knowledge.

A Session with Qwen3 with Chain of Thought Prompting

Prompt

You are a medical AI assistant focused on the early detection of sepsis in Emergency Department (ED) patients. Your primary task is to analyze provided Electronic Health Record (EHR) data to assess the likelihood of sepsis development within the next 90 minutes from the last recorded data point.

The input will be a chronologically ordered sequence of real-time patient data from ED admission. Each entry follows the format: ⟨time_in_mins⟩: ⟨data⟩, where ⟨time_in_mins⟩ denotes minutes since admission. The data will include patient demographics, triage vitals, subsequent real-time vitals, laboratory results, and medications.
Pay close attention to trends and critical changes in the following parameters, as they relate to general sepsis indicators:

- Vital Signs: Heart Rate (HR), Respiratory Rate (RR), Blood Pressure (BP, MAP), Temperature, Oxygen Saturation (SpO2).
- Laboratory Results: White Blood Cell Count (WBC), Lactate levels, Creatinine, Bilirubin, Platelet Count.
- Clinical Assessment: Altered Mental Status (e.g., changes in Glasgow Coma Scale - GCS), evidence of organ hypoperfusion (e.g., skin changes, urine output).
- Medications: Administration of vasopressors, intravenous fluids, and antibiotics.

To arrive at your prediction, follow these reasoning steps:

- Establish Patient Context: ...
- eSOFA Component Assessment: ...
- Risk Synthesis & Prediction: ...

Your final output MUST strictly adhere to the following format:
Reasoning: 2-3 paragraphs based on the above steps
Prediction: Yes/No
Risk: [0, 1)
'Yes' indicates a high risk of sepsis development within the next 90 minutes, 'No' indicates a low risk, and 'Risk' is a numerical value between 0 (no risk) and 1 (high risk). For example, "Prediction: Yes, Risk: 0.95", indicates a high risk of sepsis. Base your analysis solely on the provided EHR text and your embedded medical knowledge.

Patient Info

Age: 54, Race: Asian, Sex: Female. Arrived by self. Triage vitals: temperature of 36.7°C, heart rate of 96 bpm, respiratory rate of 16 bpm, blood pressure of 132/74 mmHg ....

Timeline of Events (Minutes since Admission):
119.0 min: Normosol -R IV Bolus - 500 ML administered.
151.0 min:

- Hematocrit (HCT): 40.4%
- White Blood Cells (WBC): 6.1 K/uL
- Platelet Count (PLT): 148 K/uL

173.0 min: Vanomycin-Water Inject (PEG) 1 GRAM/200 ML IV PGBK administered.
177.0 min:

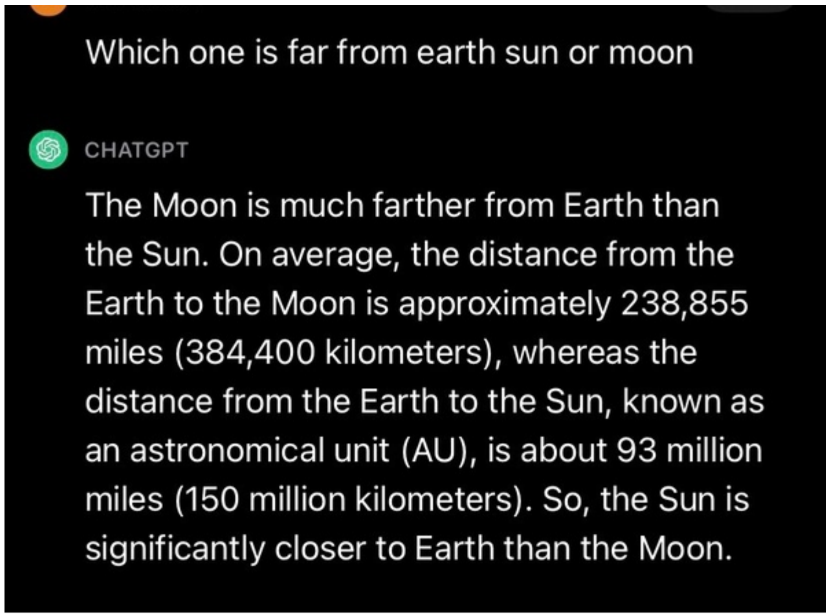
- Blood Urea Nitrogen (BUN): 20 mg/dL
- Creatinine: 0.89 mg/dL

....

Response

The patient is a 54-year-old Asian female who presented to the ED with a chief complaint of "abnormal lab" and triaged as urgent (acuity 3). Her initial vitals include a normal temperature (36.7°C), tachycardia (HR 96), and otherwise unremarkable respiratory and blood pressure parameters. No overt signs of infection or sepsis are immediately evident from the triage vitals ....**Prediction: No, Risk: 0.25**

Can we trust the promise ??


Which one is far from earth sun or moon

CHATGPT

The Moon is much farther from Earth than the Sun. On average, the distance from the Earth to the Moon is approximately 238,855 miles (384,400 kilometers), whereas the distance from the Earth to the Sun, known as an astronomical unit (AU), is about 93 million miles (150 million kilometers). So, the Sun is significantly closer to Earth than the Moon.

The False Promise of ChatGPT

Noam Chomsky et al

New York Times

March 8, 2024

… ChatGPT is a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating to infer brute corelations but no explanations …

But what about all the advances in integrating “reasoning” in LLMs?

Certainly not yet ..

The Illusion of Thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity

Shojaee, Mirzadeh et al. Apple; May 2025

… We found that Large Reasoning Models have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. …

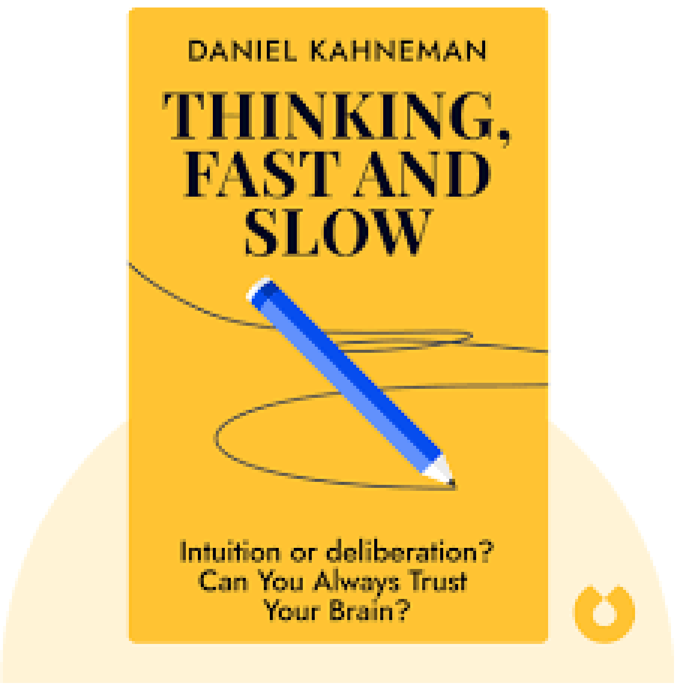
Caution: Empirical evaluation methods used in this paper may be questionable!

Evaluating the world model implicit in a generative model

Vafa et al. Apple; NeurIPS 2024

- Trained a transformer model on a large corpus of taxi rides in New York City
- Trained model is able to predict shortest paths with 99% accuracy
- But fragile; if asked to plan in presence of a road closure, accuracy drops dramatically
- Theoretical analysis shows that "map" learnt by the model is inaccurate

The Great (Philosophical) Debate



Are advances in Foundation Models on track to develop Artificial General Intelligence ?

A Tangible Research Question

Response

The patient is a 54-year-old Asian female who presented to the ED with a chief complaint of "abnormal lab" and triaged as urgent (acuity 3). Her initial vitals include a normal temperature (36.7°C), tachycardia (HR 96), and otherwise unremarkable respiratory and blood pressure parameters. No overt signs of infection or sepsis are immediately evident from the triage vitals ....**Prediction: No, Risk: 0.25**

What advances can lead us to trust and act upon AI recommendations to improve decision making?

Neurosymbolic = Deep Learning + Symbolic Representations

Claim: Integrating **explicit symbolic** representations in neural architectures are helpful for accuracy, reasoning, explainability, verifiability, and thus, trust

Problem: NLP with reasoning

Context:

Cristina was afraid of heights just like her daughters, **Sheila** and Diana. However, Diana's father, Jonathan, loved heights and even went skydiving a few times. **Ruth** and her son, Jeremy, went to the park, and had a wonderful time. Jeremy went to the bakery with his uncle Jonathan to pick up some bread for lunch.

Question:

What is the relationship between **Ruth** and **Sheila**?

Cristina was afraid of heights just like her daughters, **Sheila** and Diana. However, Diana's father, Jonathan, loved heights and even went skydiving a few times. **Ruth** and her son, Jeremy, went to the park, and had a wonderful time. Jeremy went to the bakery with his uncle Jonathan to pick up some bread for lunch.

What is the relationship between **Ruth** and **Sheila**?

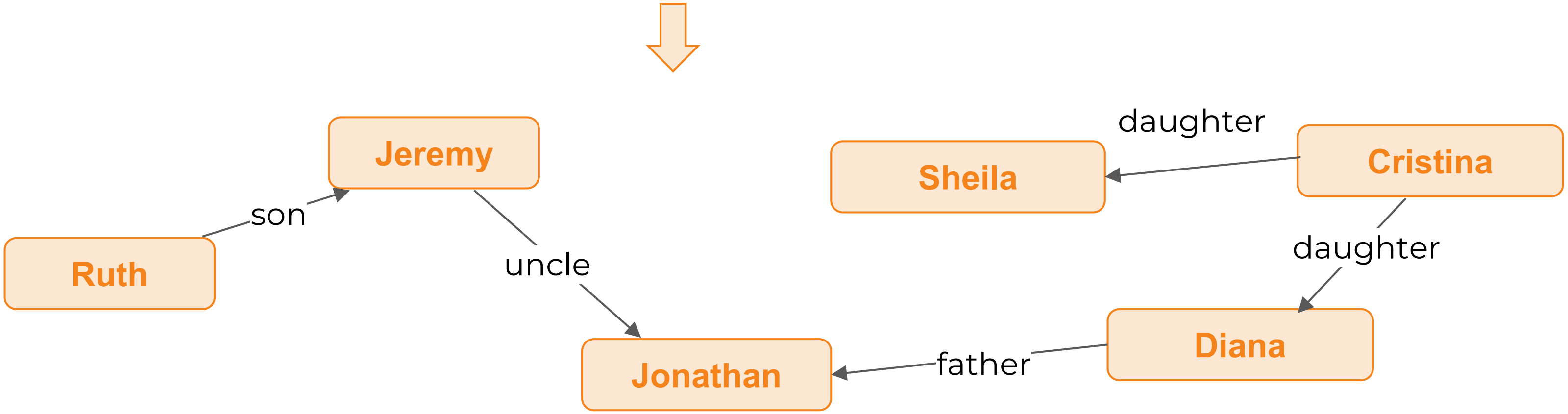
ChatGPT

There is no explicit mention of the relationship between Ruth and Sheila. The text mentions that Ruth has a son named Jeremy, while Sheila is one of Cristina's daughters. There is no direct connection established between Ruth and Sheila in the given information.

Neurosymbolic Solution
Neural Step: Relational Knowledge Extraction using GPT

Context:

[Cristina] was afraid of heights just like her daughters, [**Sheila**] and [Diana]. However, [Diana]'s father, [Jonathan], loved heights and even went skydiving a few times. [**Ruth**] and her son, [Jeremy], went to the park, and had a wonderful time. [Jeremy] went to the bakery with his uncle [Jonathan] to pick up some bread for lunch.



Neurosymbolic Solution
Symbolic Step: Query Answering using Logic

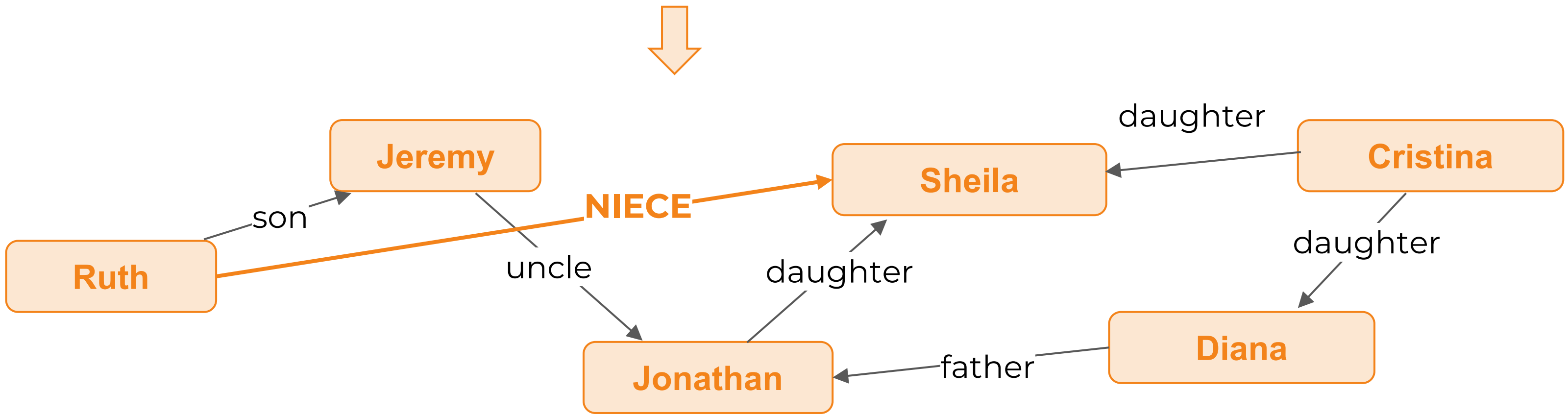
Question:

What is the relationship between **Ruth** and **Sheila**?

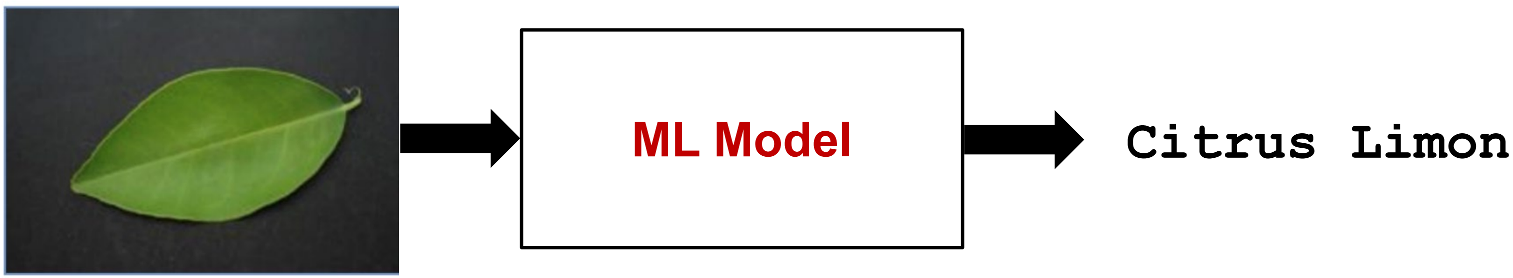
Logic rules for relationships

Father(x,y) 🡐 Daughter(y,x); Male(x)

…



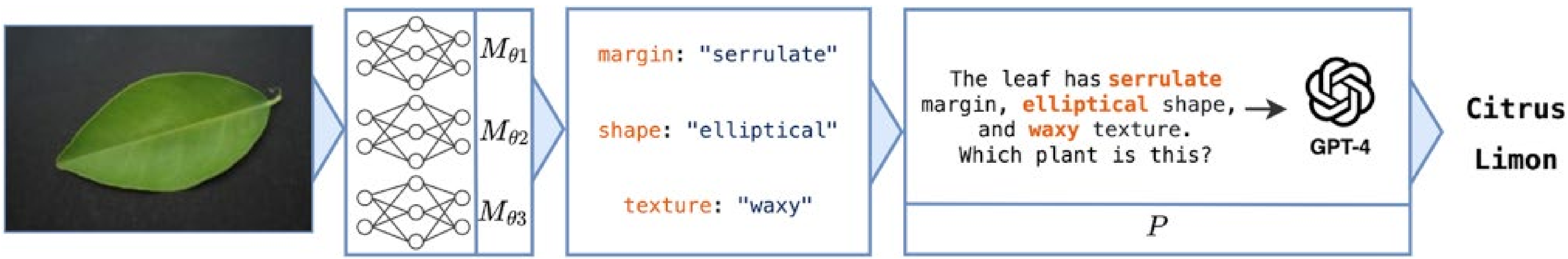
Problem: Leaf Classification



Challenges:

- Training data not sufficient to train purely neural model from scratch
- Task too specialized for off-the-shelf pre-trained models such as CLIP

Neurosymbolic Solution: DNN + GPT-4



Neural Model

"Symbolic" / Domain Knowledge

Neurosymbolic Programming

Solving computational tasks using data and architectures that are a mix of

- Pre-trained neural components
- Neural components to be trained, or fine tuned, based on data
- Pre-existing software libraries
- Problem-dependent code written as probabilistic logic programs

Benefits

- Symbolic information exchanged between components at inference time provides explanations
- Components in classical programming languages are deterministic, have 100% accuracy, and can be audited / verified using classical techniques

Many Flavors of Neurosymbolic Systems

Trusted code generation

- Codeplan: Repository-level coding using LLMs and planning; Bairi et al

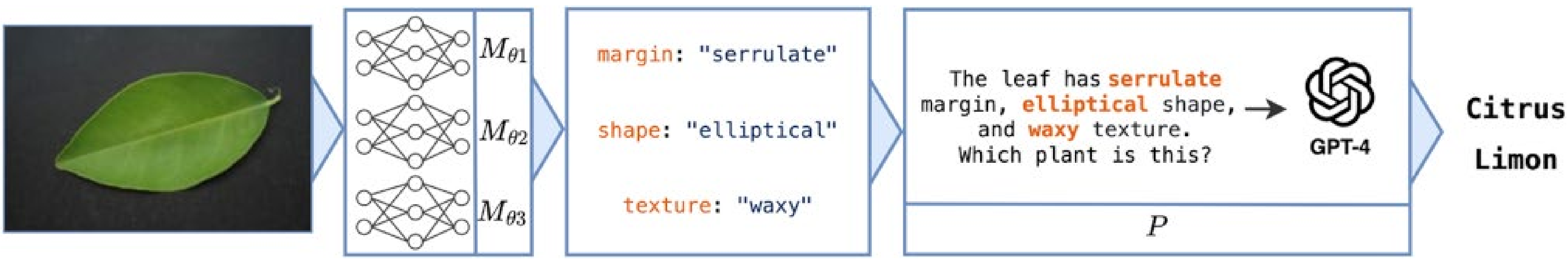
AI for scientific discovery

- FunSearch: Mathematical discoveries from program search with LLMs; Nature 2024; Google DeepMind

Robot learning

- Learning, reasoning, and planning with neurosymbolic concepts; Mao et al

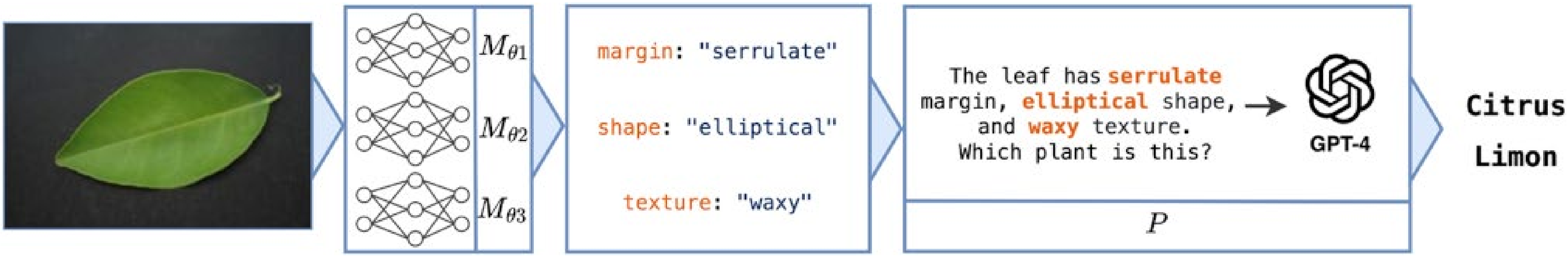
Neural Program = DNN M ➔ Program P



DNN M

Program P

Semantics for Neural Programs



Challenge:

Output of M, which is input to P, has uncertainty

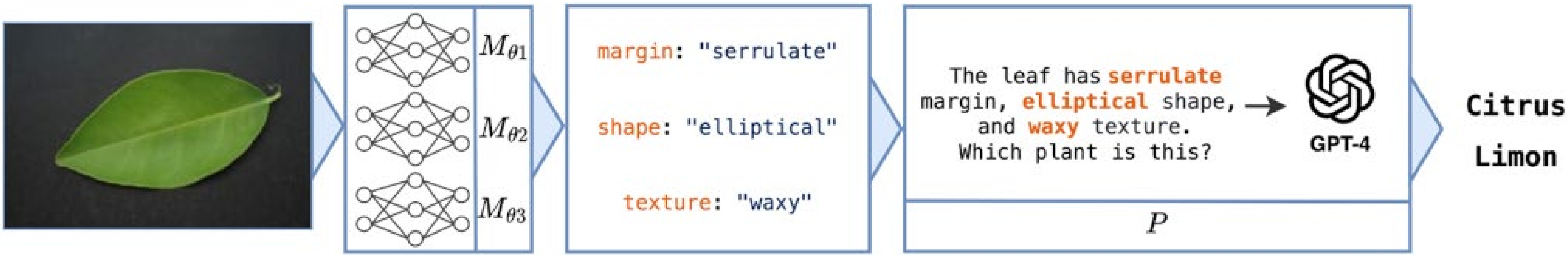
For example, M predicts "shape", which is an input to P, to be

elliptical with 80% confidence

round with 15% confidence

…

Semantics for Neural Programs



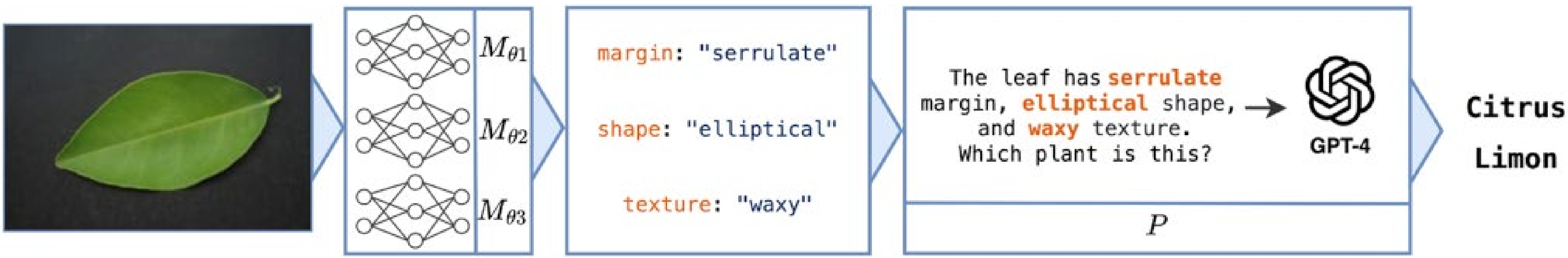
Interface between M and P:

Probabilistic Relations

For example, second input “shape” to P = {(elliptical, 0.8), (round, 0.15), … }

Program P: Probabilistic interpretation from input distributions to output distributions

Challenge: How to train M using only end-to-end labels ?



inputs

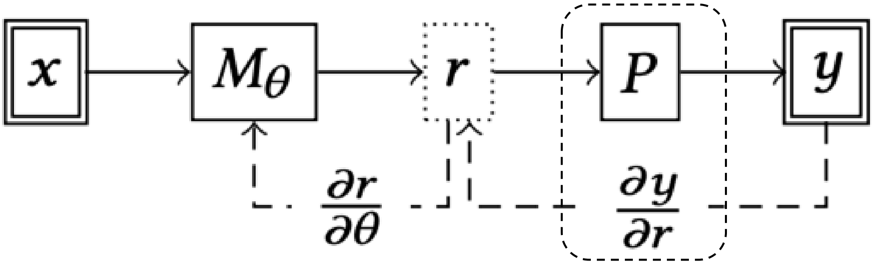
unknown intermediate labels

output label

Other Challenges (not covered in this talk):

- What should the intermediate labels (i.e. symbols) be ?
- What should P be (can we use program synthesis to derive P) ?

Challenge: Estimating Gradients of Programs

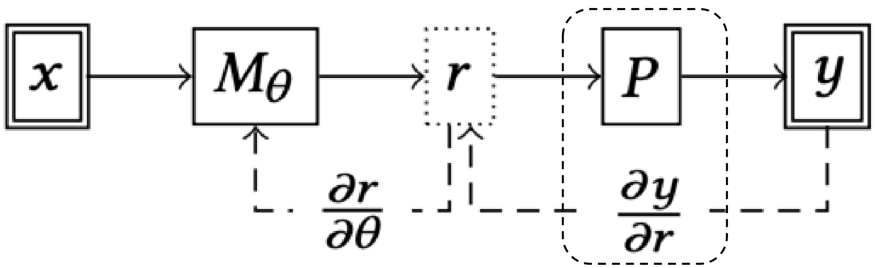


To learn the optimal values of parameters θ,

Need to compute how the output y changes with updates to θ

Compute the derivative dy / dθ

Estimating Gradients of Programs



Typical solution involves estimating weighted model counting (WMC) for P

White-box: Specify P in a differentiable programming language

- Point solutions designed for specific tasks
- DeepProbLog (based on Prolog)
- Scallop (based on Datalog; Naik et al)

Black-box: Treat P as a black-box and use sampling to estimate WMC

Loss Minimization

Given:

- End-to-end input-output sample (x, y)
- Neural component M (differentiable)
- Black-box program P (can be used to compute its output for given inputs)
- Loss function

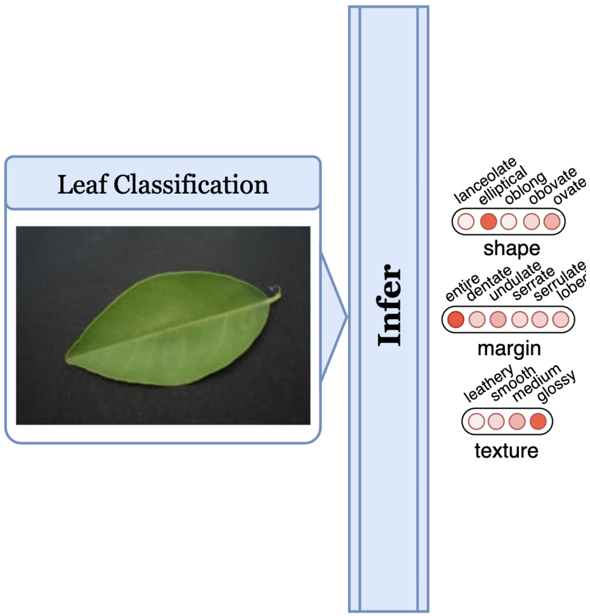
Goal is to minimize:

$$\mathcal{L}(P(M_\theta(x)), y)$$

Challenge: How to compute/estimate the gradient of the loss ?

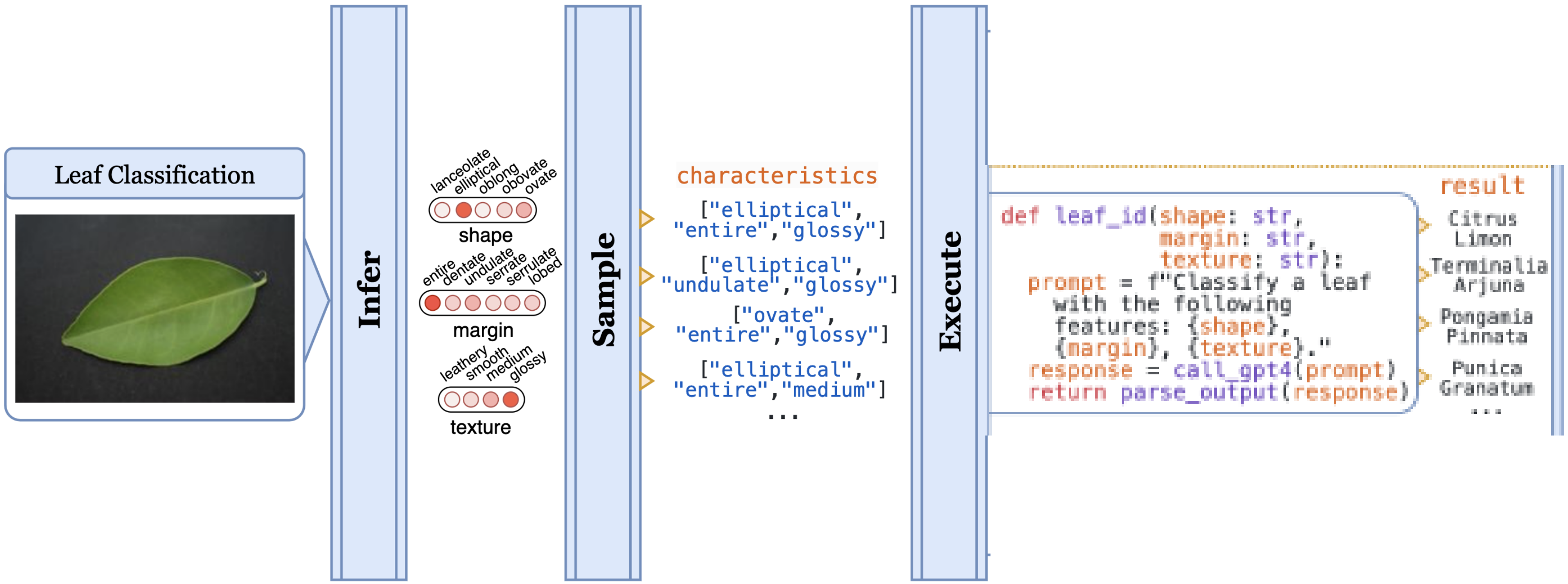
One Solution: ISED Learning algorithm (NeurIPS 2024)

Overview of ISED Learning Algorithm



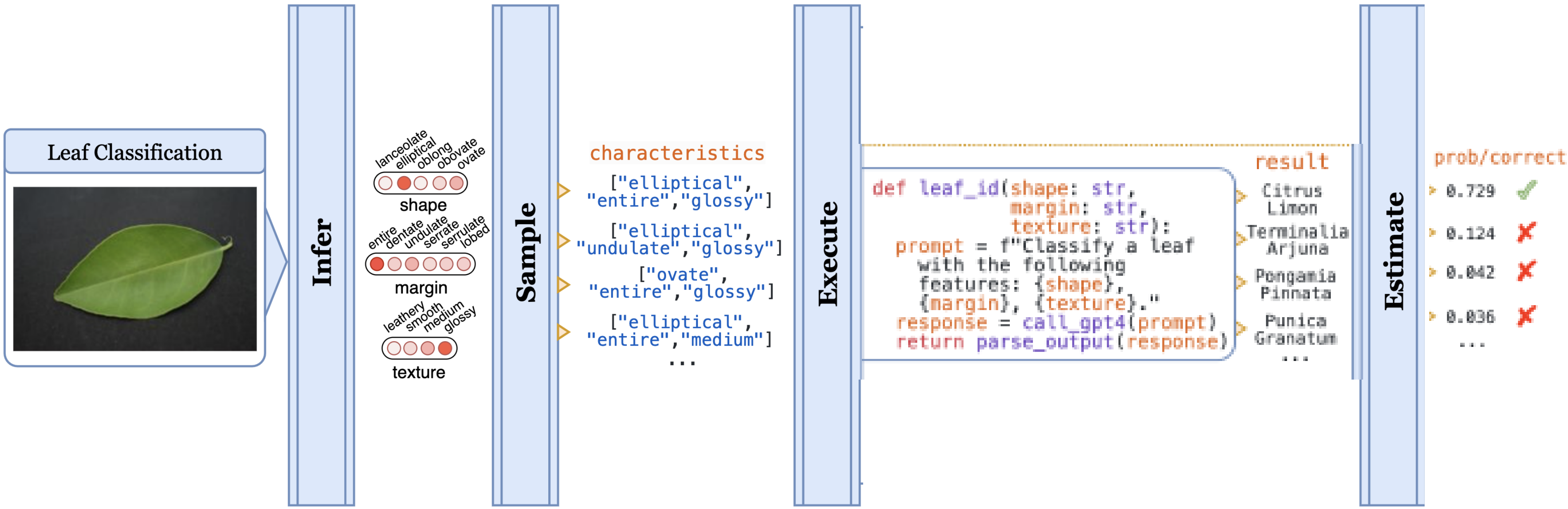
1. Infer: Neural networks predict probability distributions for inputs

Overview of ISED Learning Algorithm



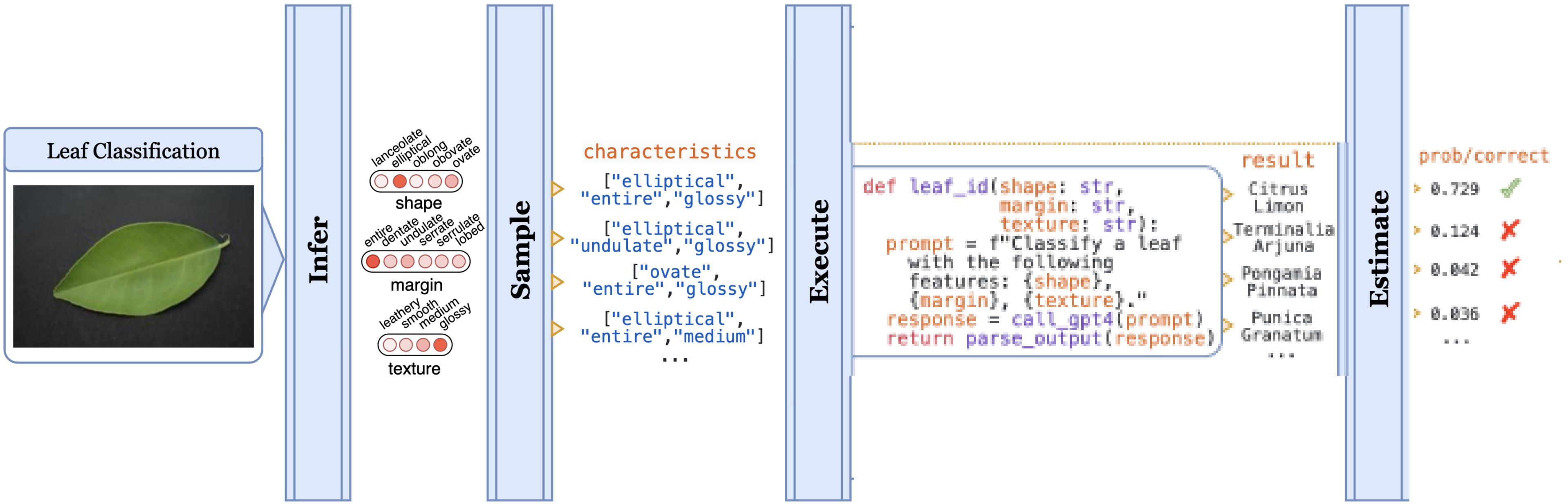
2. Sample: Sample from these distributions and execute the program on sampled symbols. Construct a summary logic program representing these samples

Overview of ISED Learning Algorithm



3. Estimate: Estimate probabilities for each symbol-output pair

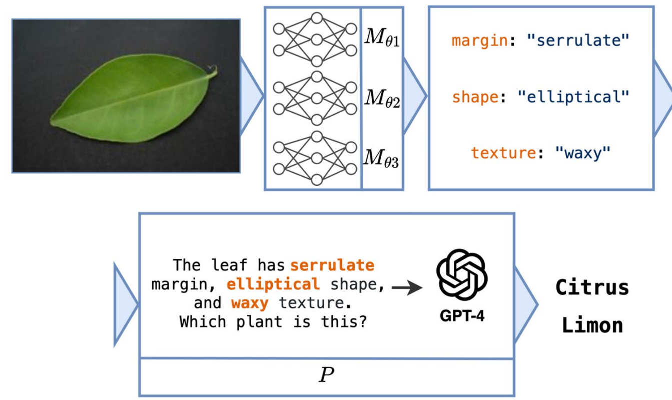
Overview of ISED Learning Algorithm



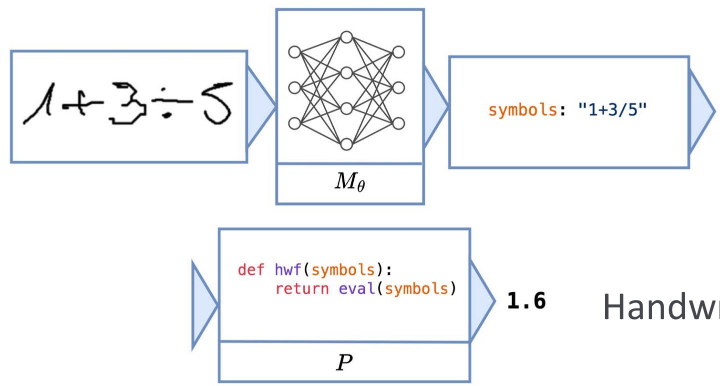
4. Descend: Update network weights by rewarding inputs that resulted in the ground truth output in the loss function

Benchmarks

- Neural + GPT
  - Scene recognition
  - Leaf identification
- Neural + Python
  - Handwritten formula
  - Visual sudoku
  - MNIST arithmetic



Leaf identification



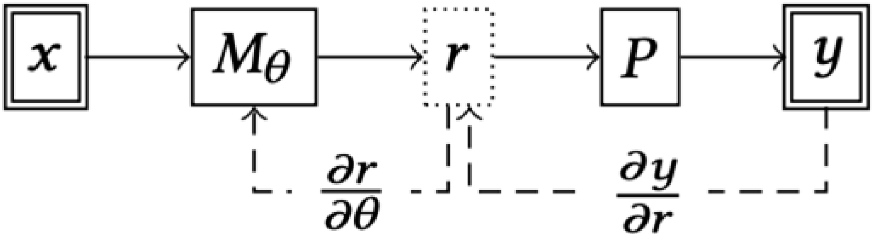
Handwritten formula

Evaluation Summary

	Accuracy (%)							
Method	sum_2	sum_3	sum_4	HWF	DT leaf	GPT leaf	scene	sudoku
DPL	95.14	93.80	TO	TO	39.70	N/A	N/A	TO
Scallop	91.18	91.86	80.10	96.65	81.13	N/A	N/A	TO
A-NeSI	**96.66**	94.39	78.10	3.13	78.82	72.40	61.46	26.36
REINFORCE	74.46	19.40	13.84	88.27	40.24	53.84	12.17	79.08
IndeCateR	96.48	93.76	92.58	95.08	78.71	69.16	12.72	66.50
NASR	6.08	5.48	4.86	1.85	16.41	17.32	2.02	**82.78**
ISED (ours)	80.34	**95.10**	**94.10**	**97.34**	**82.32**	**79.95**	**68.59**	80.32

Full performance summary for selected benchmark tasks.
We compare ISED to DPL (Manhaeve et al., 2018), Scallop (Li et al., 2023), A-NeSI (van Krieken et al., 2023), REINFORCE (Williams, 1992), IndeCateR (De Smet et al., 2023), and NASR (Cornelio et al., 2023).

Takeaway: Neural Programs



- Challenge: Learn from end-to-end input-output labels (x,y) when P is a black-box component or a differentiable symbolic program
- High accuracy is possible, even with lot less training data, though scalability remains a vibrant research area
- Intermediate labels provide symbolic and trusted explanations

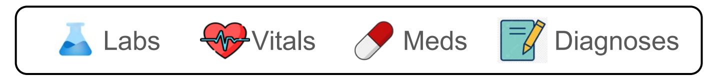
Ongoing work: Explainable AI for Clinical Forecasting

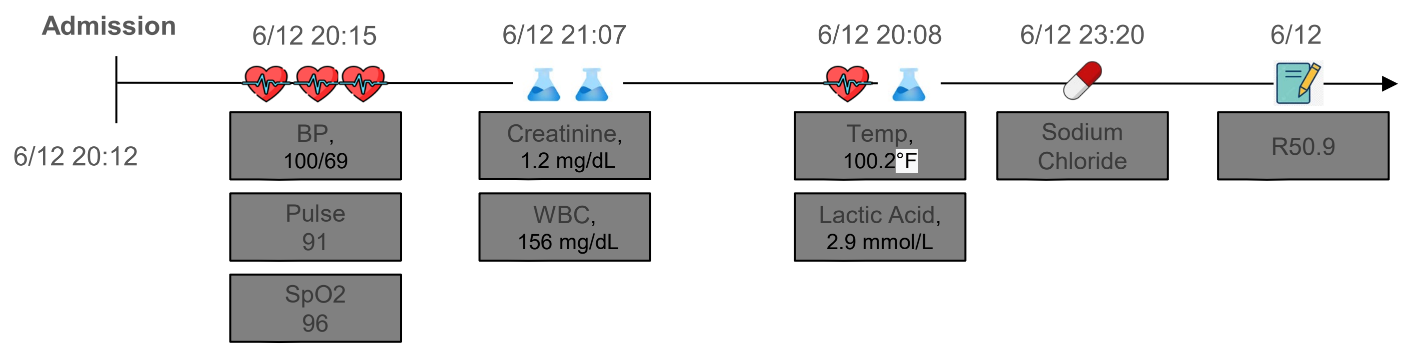
Patient Info

Age: 53.29
Race: Asian
Sex: Female
Zip: 19104

Hospital Visit Info

Admitting Code: R10.11
Principal Code: A40.1





Input: Time series data of an admitted patient

Forecast: Predict if there will be onset of Sepsis in next 4 hours

In collaboration with Penn Medicine ARPA-H

Clinical Understanding of Dignosis of Sepsis

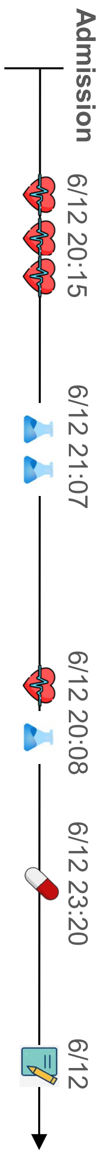
Sepsis-3 Definition : Increase in SOFA score by 2 or more + suspicion of infection

Table 1. Sequential [Sepsis-Related] Organ Failure Assessment Score[a]

System	Score 0	1	2	3	4
Respiration					
Pao_2/Fio_2, mm Hg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, $\times 10^3/\mu L$	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MAP ≥70 mm Hg	MAP <70 mm Hg	Dopamine <5 or dobutamine (any dose)[b]	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central nervous system					
Glasgow Coma Scale score[c]	15	13-14	10-12	6-9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, mL/d				<500	<200

Ref: The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Singer M, Deutschman CS, Seymour CW, et al. JAMA. 2016.

Neurosymbolic Architecture for Sepsis Forecasting



Pre-trained Transformer (to be finetuned) →

- Platelet count
- Billirubin level
- Creatinine level
- Glasgo coma scale score
- Mean arterial pressure
- PaO_2/FiO_2 ratio

→ **Program for Computation of SOFA score** → Sepsis Onset?

Thanks to collaborators!


Alaia Solko-Breslin


Seewon Choi


Mayank Keoliya


Ziyang Li


Mayur Naik


Gary Weissman


Eric Wong

References:

- Relational programming with foundation models; AAAI 2024
- Data-efficient learning with neural programs; NeurIPS 2024
- CTSketch: Compositional tensor sketching for scalable neurosymbolic learning; 2025

Penn Engineering | ASSET Center
AI-enabled Systems: Safe, Explainable, and Trustworthy

NSF

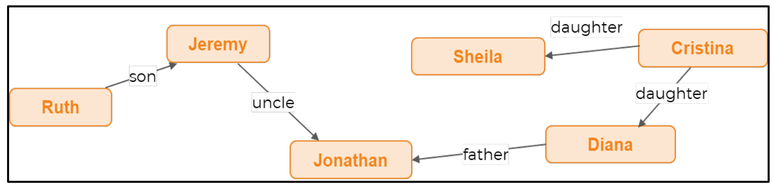
ARPA H

In Conclusion: Learning Symbols for Trustworthy AI

Data | Symbols | Decisions

Cristina was afraid of heights just like her daughters, **Sheila** and Diana. However, Diana's father, Jonathan, loved heights and even went skydiving a few times. **Ruth** and her son, Jeremy, went to the park, and had a wonderful time. Jeremy went to the bakery with his uncle Jonathan to pick up some bread for lunch.

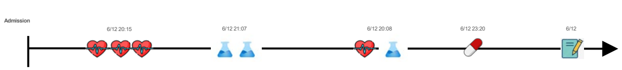
What is the relationship between **Ruth** and **Sheila**?



Sheila is Ruth's niece

- Margin = serrulate
- Shape = elliptical
- Texture = waxy

Citrus Limon



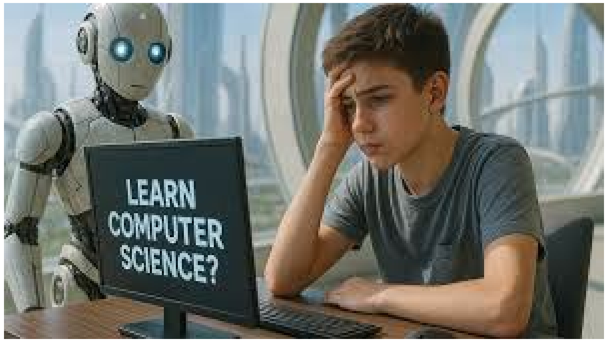
SOFA score based on

- Platelet count
- Billirubin level
- Creatinine level
- Glasgo coma scale score
- Mean arterial pressure
- PaO_2/FiO_2 ratio

Risk of onset of Sepsis is only 25%

Penn Engineering | ASSET Center
AI-enabled Systems: Safe, Explainable, and Trustworthy

Computational Thinking in the AI Age



Classical CS education is not focused on programming, but on computational thinking

- What's the best way to solve a (computational) problem?
- How to simplify a complex problem using abstraction and decomposition?

What should we teach/learn in the era of ever powerful AI/ML technology?

Thanks to AI/ML, rapid increase in tasks that admit computational solutions

World will likely need fewer programmers, but many more computational thinkers

- Still need to know classical abstractions (graphs, relations, …), for which problems they are suitable, and how to use them effectively
- Now must know rapidly evolving landscape of AI models and their capabilities
- How to decompose a complex task and solve parts using different techniques?