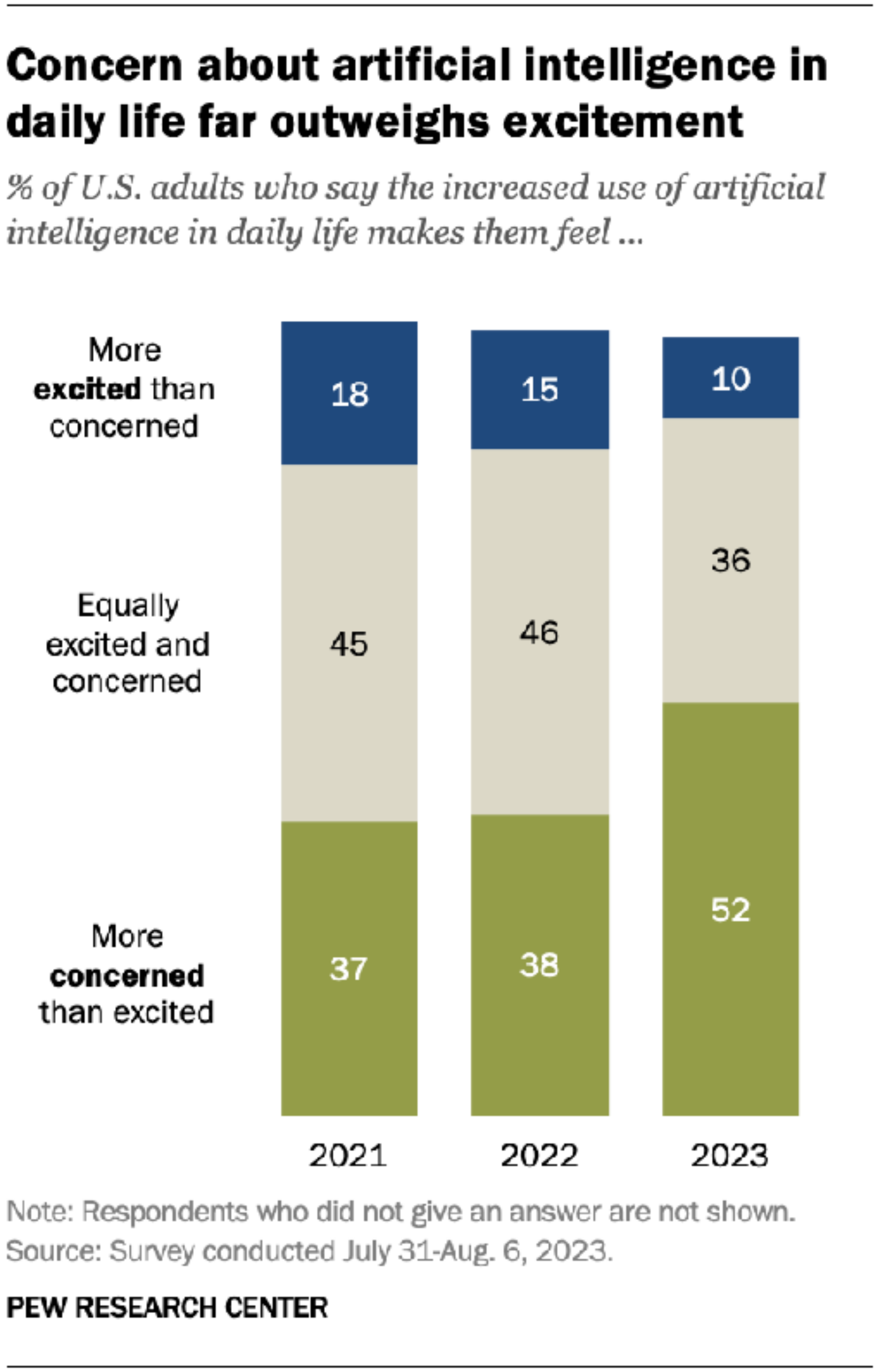
Neurosymbolic Learning and Reasoning for Trustworthy AI

Zhe Zeng, Rice Hall 105

September 18th, 2025 - **CS6190**

Trustworthy AI is A Pressing Need

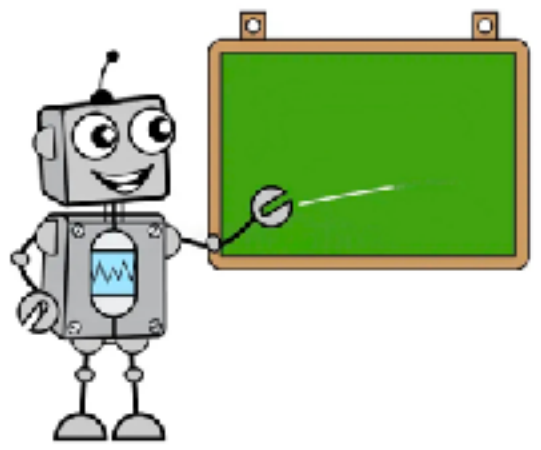


Goal

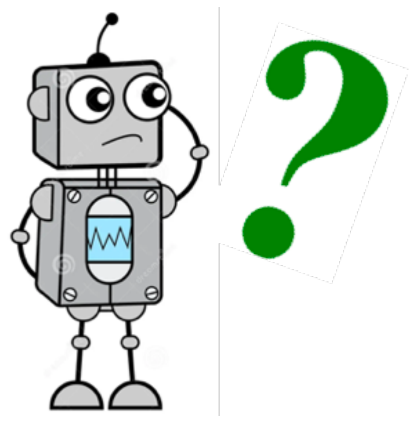
How to build AI systems that *behave as intended*?

Challenges in Trustworthy AI

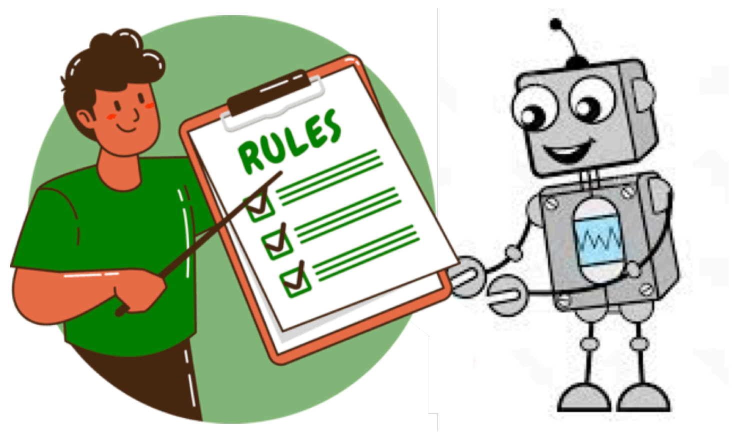
When an AI model makes decisions,



- Explain *why*



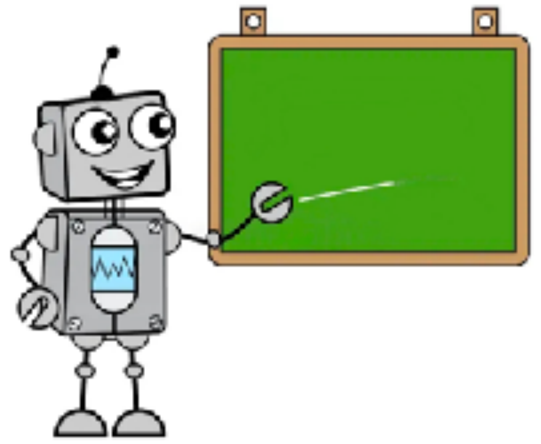
- Know what they *don't know*



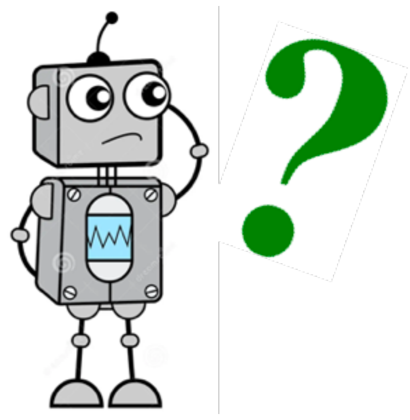
- Follow *rules/domain knowledge*

Challenges in Trustworthy AI

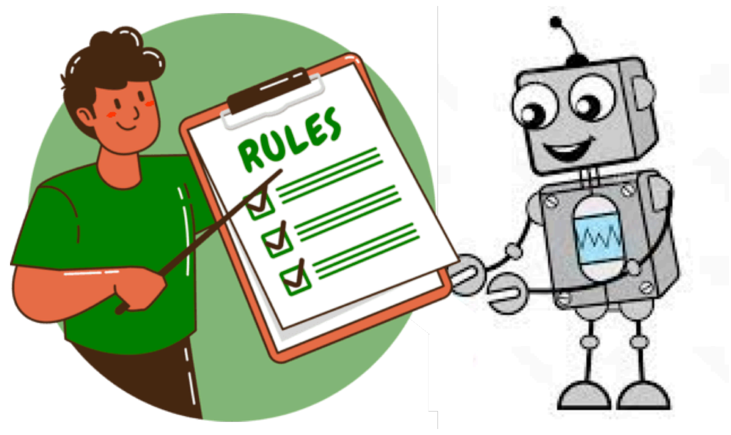
When an AI model makes decisions,



- Explain *why*



- Know what they *don't know*



- Follow *rules/domain knowledge*

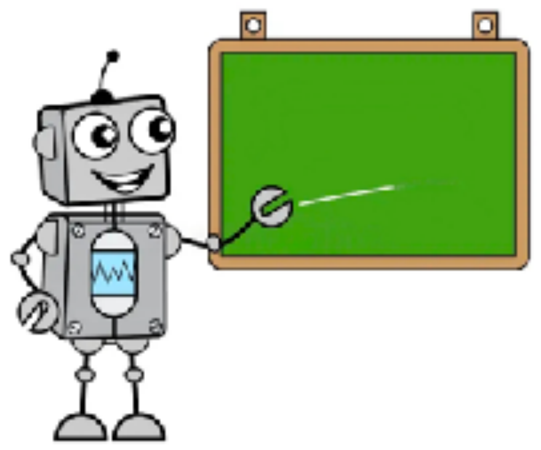
Solution:

Deep Learning Models?

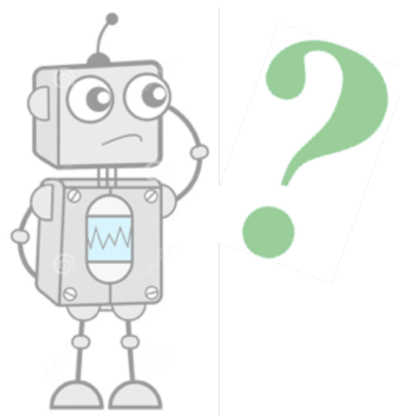
Fundamental Limitations

Challenges in Trustworthy AI

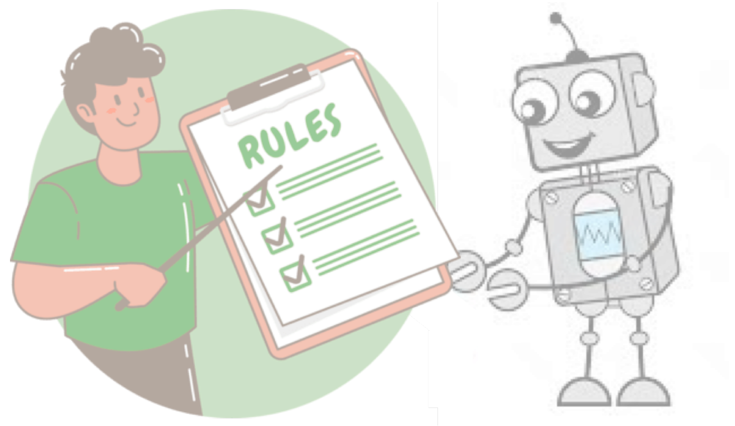
When an AI model makes decisions,



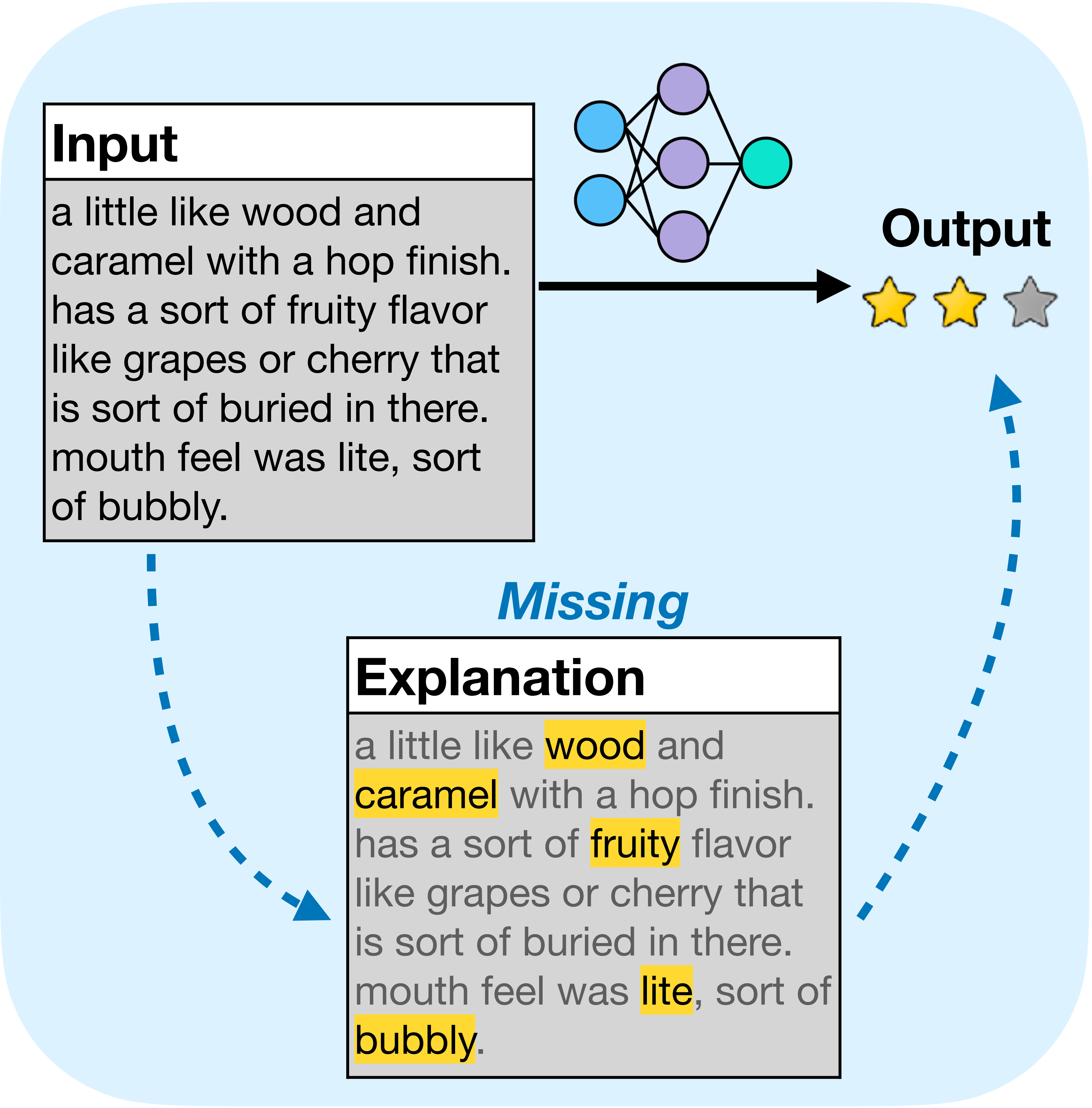
- Explain *why*



- Know what they *don't know*

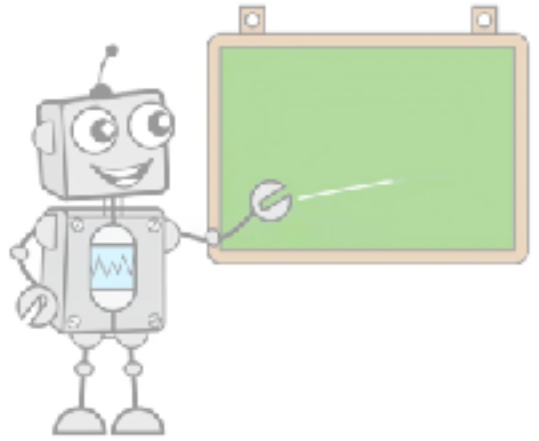


- Follow *rules/domain knowledge*

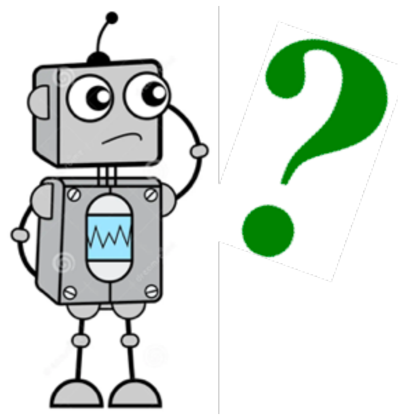


Challenges in Trustworthy AI

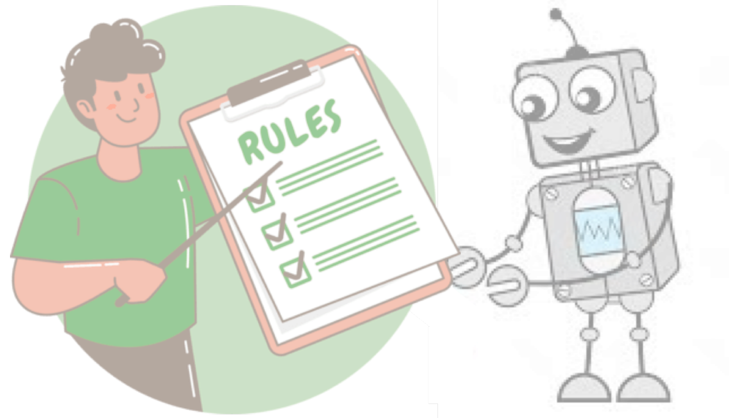
When an AI model makes decisions,



- Explain *why*



- Know what they *don't know*



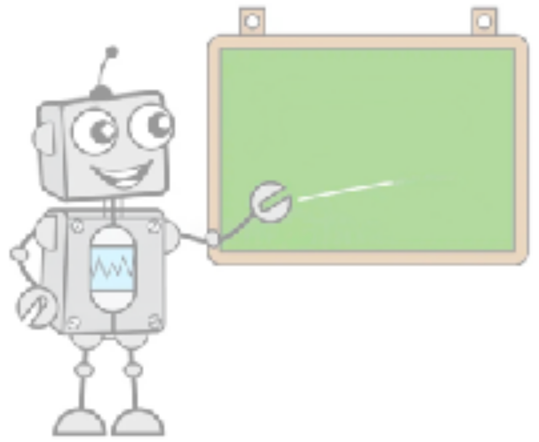
- Follow *rules/domain knowledge*

Image Classification

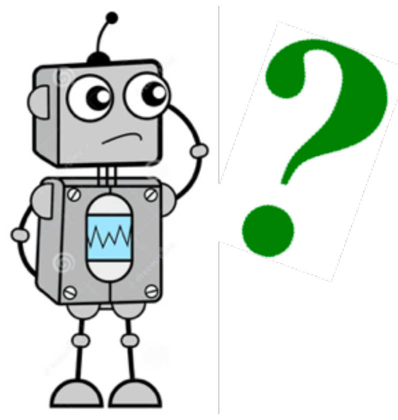
Confidence: 80%

Challenges in **Trustworthy AI**

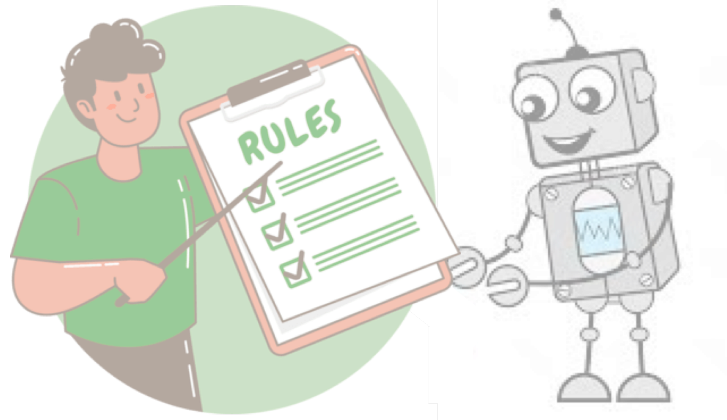
When an AI model makes decisions,



- Explain *why*



- Know what they *don't know*



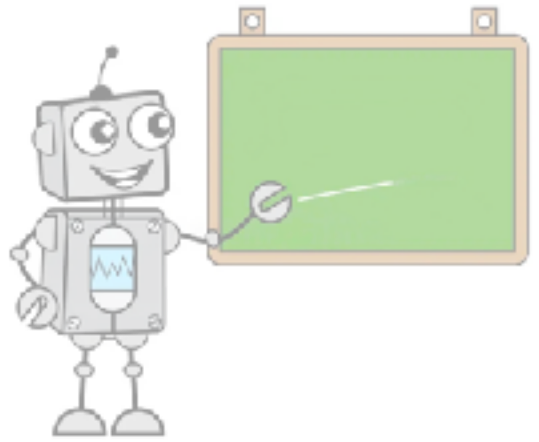
- Follow *rules/domain knowledge*

Image Classification

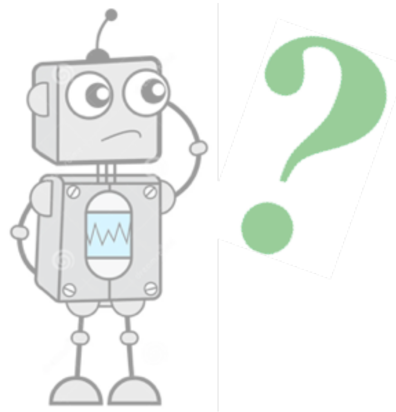
Confidence: 80% ***Accuracy: 50%***

Challenges in Trustworthy AI

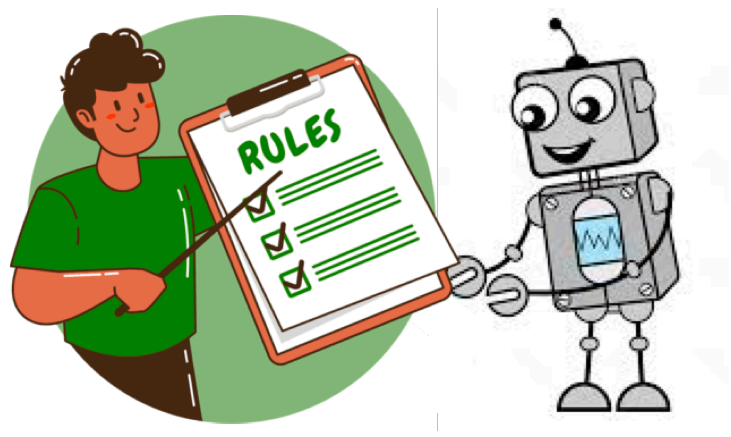
When an AI model makes decisions,



- Explain *why*

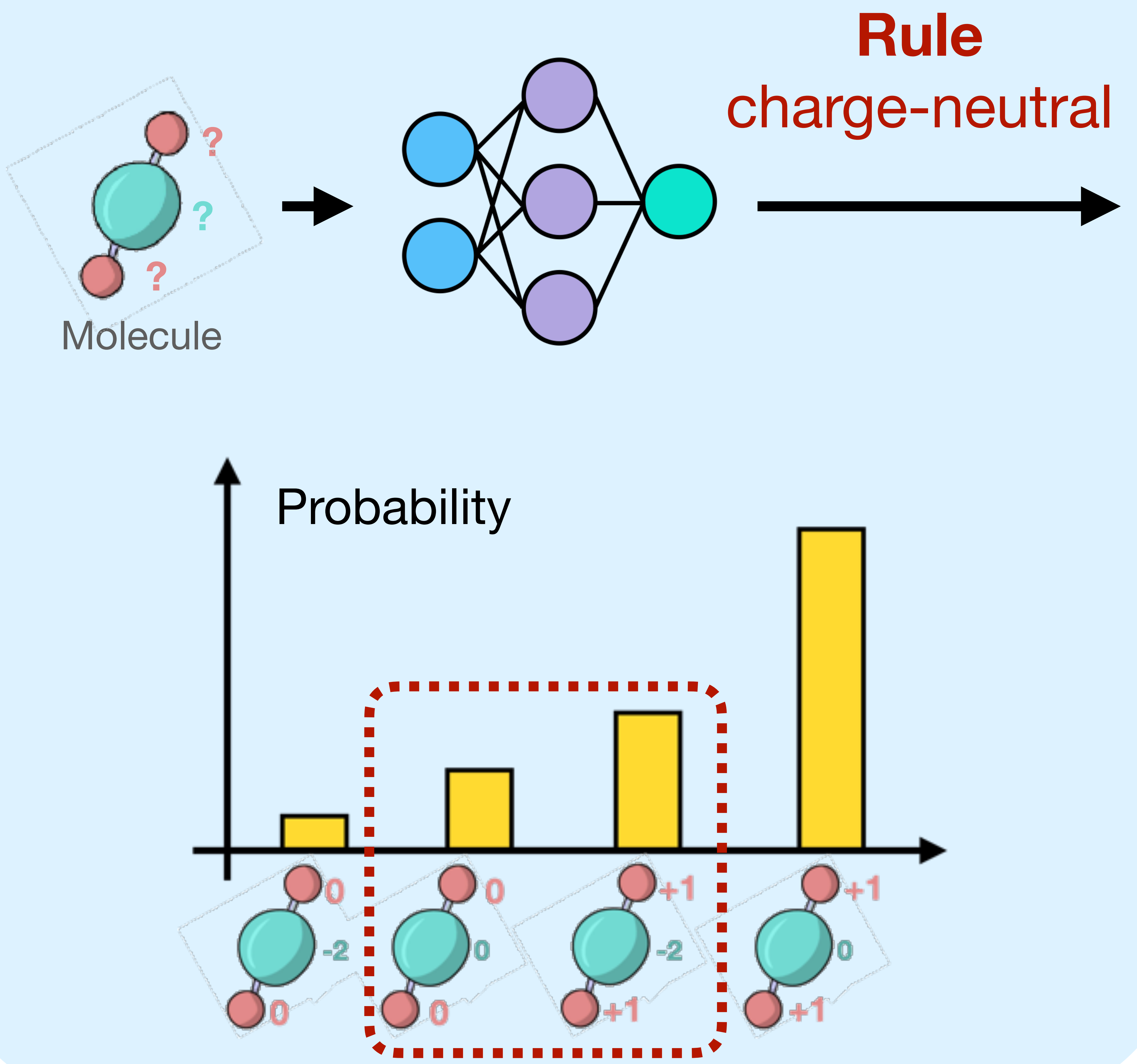


- Know what they *don't know*



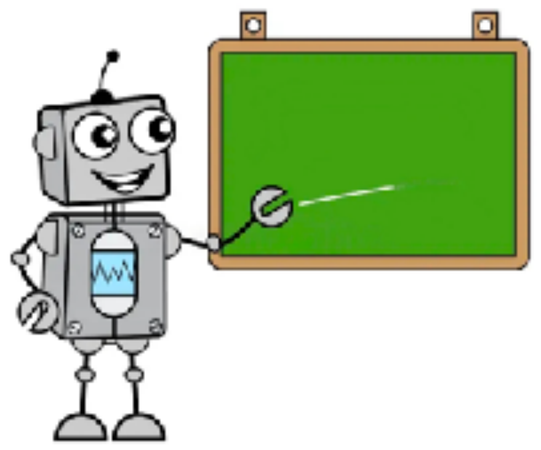
- Follow *rules/domain knowledge*

Molecule Modeling

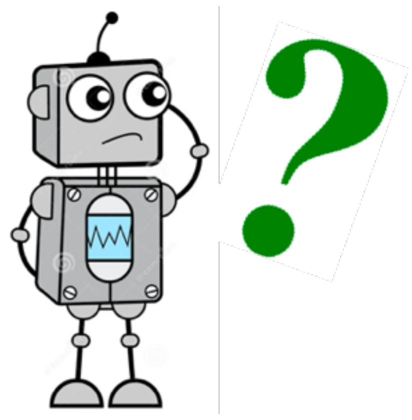


Challenges in Trustworthy AI

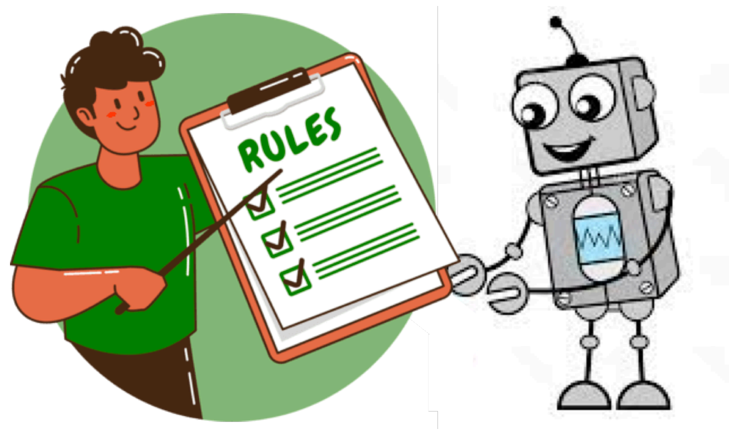
When an AI model makes decisions,



- Explain *why*



- Know what they *don't know*



- Follow *rules/domain knowledge*

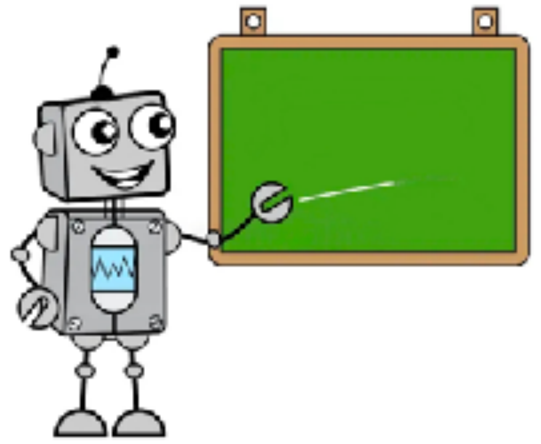
Generate a sentence with "**Virginia**" as the fifth word.

The journey through the mountains in **Virginia** revealed breathtaking views.

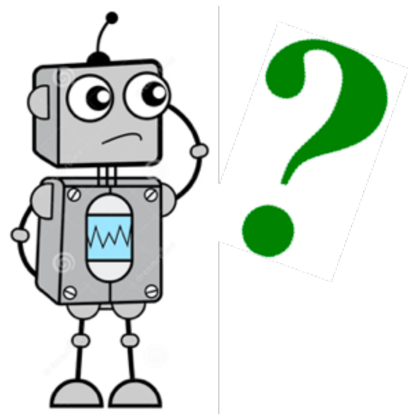
7th

Challenges in Trustworthy AI

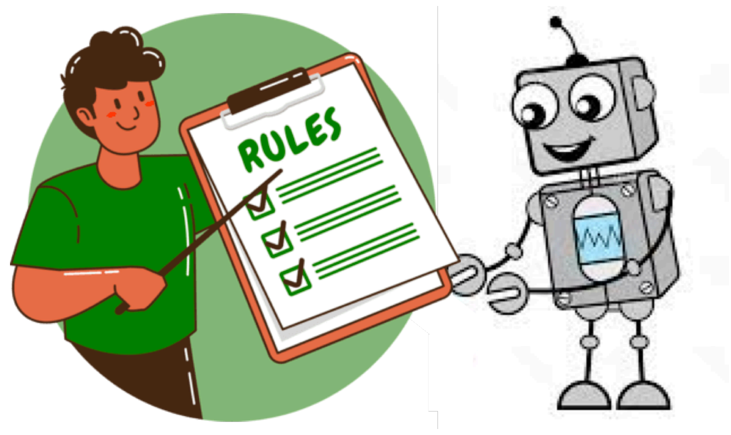
When an AI model makes decisions,



- Explain *why*



- Know what they *don't know*



- Follow *rules/domain knowledge*

Trustworthy AI
can not be achieved
by *pure deep learning* alone.

Symbolic AI

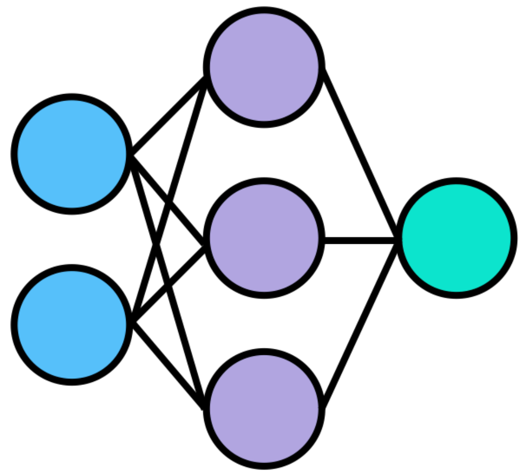
Knowledge	*Symbolic Constraints*	
“The fifth word being Virginia”	$X_5 = \text{“}Virginia\text{”}$	(Symbol)
“A short explanation”	$\bigvee_{s:\lvert s\rvert=k} \bigwedge_{i\in s} X_i \bigwedge_{i\notin s} \overline{X}_i$	(Logical)
“A charge-neutral molecule”	$\sum_i X_i = 0$	(Arithmetic)
“ReLU Neural Network”	$(WX < 0 \Rightarrow Y = 0) \bigwedge (WX \geq 0 \Rightarrow Y = WX)$	(Logical & Arithmetic)

Neurosymbolic AI

Earlier Example

Input
a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.

→



+

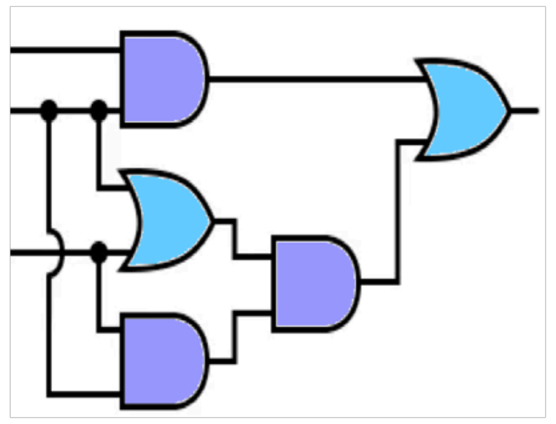
$$\bigvee_{s:|s|=k} \bigwedge_{i\in s} X_i \bigwedge_{i\notin s} \overline{X}_i$$

→

Explanation
a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.

Neurosymbolic AI

Symbolic

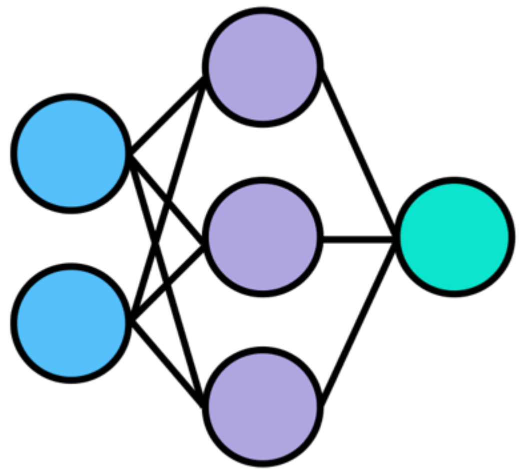


Explainability

Knowledge

Data Efficient

Neural

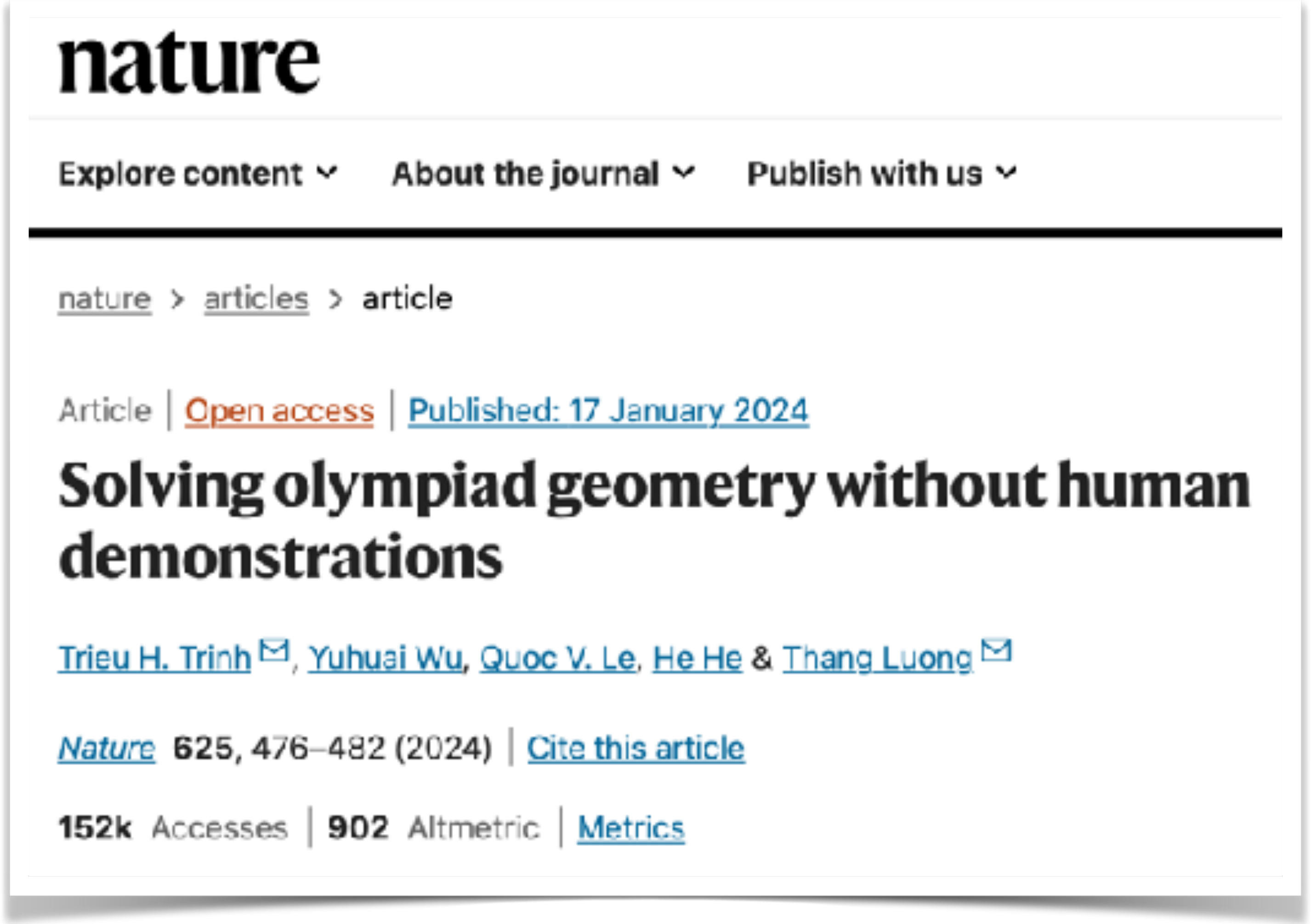


Recognition

Flexibility

Scalability

Another Example



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

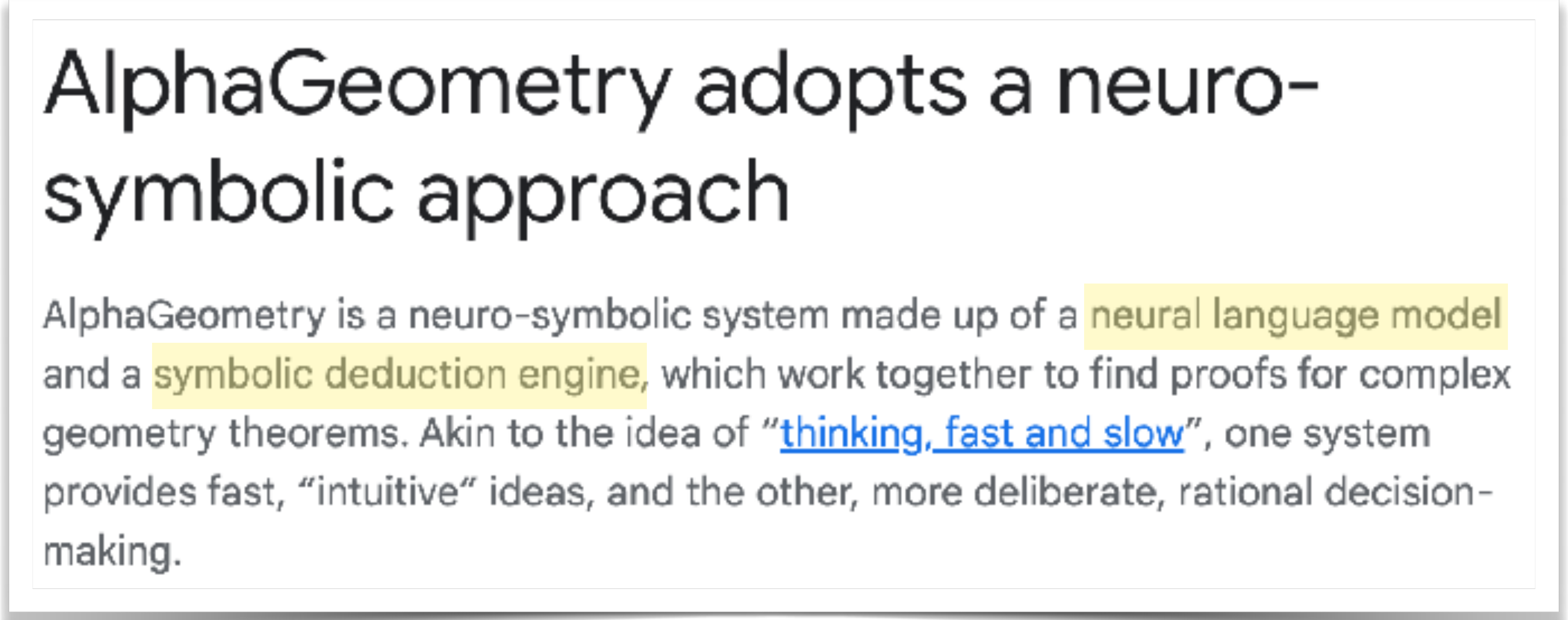
Article | Open access | Published: 17 January 2024

Solving olympiad geometry without human demonstrations

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He & Thang Luong

Nature 625, 476–482 (2024) | Cite this article

152k Accesses | 902 Altmetric | Metrics



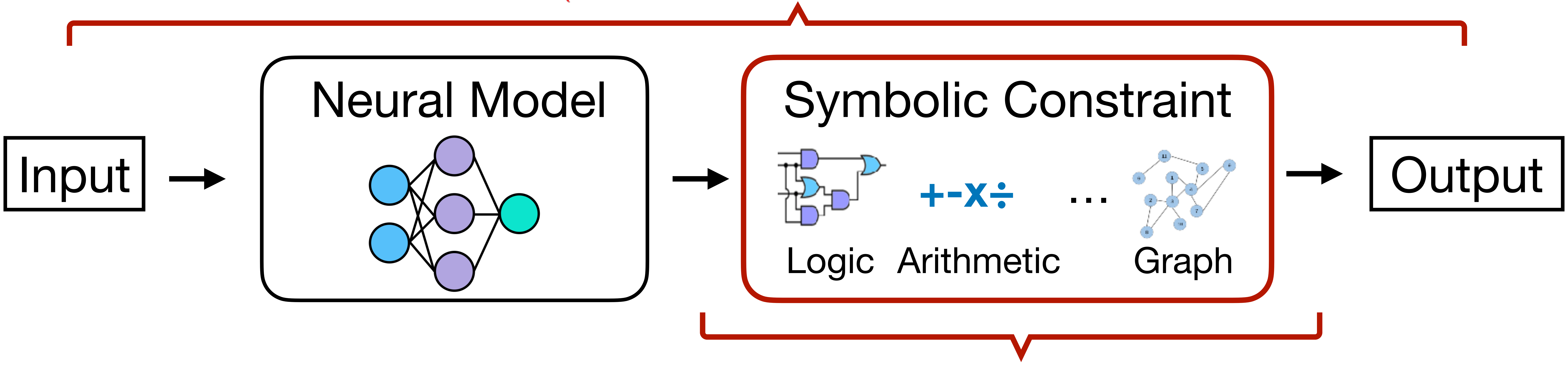
AlphaGeometry adopts a neuro-symbolic approach

AlphaGeometry is a neuro-symbolic system made up of a neural language model and a symbolic deduction engine, which work together to find proofs for complex geometry theorems. Akin to the idea of "thinking, fast and slow", one system provides fast, "intuitive" ideas, and the other, more deliberate, rational decision-making.

Unique Challenges in Neurosymbolic AI

Challenge 1 Most symbolic modules are *non-differentiable*

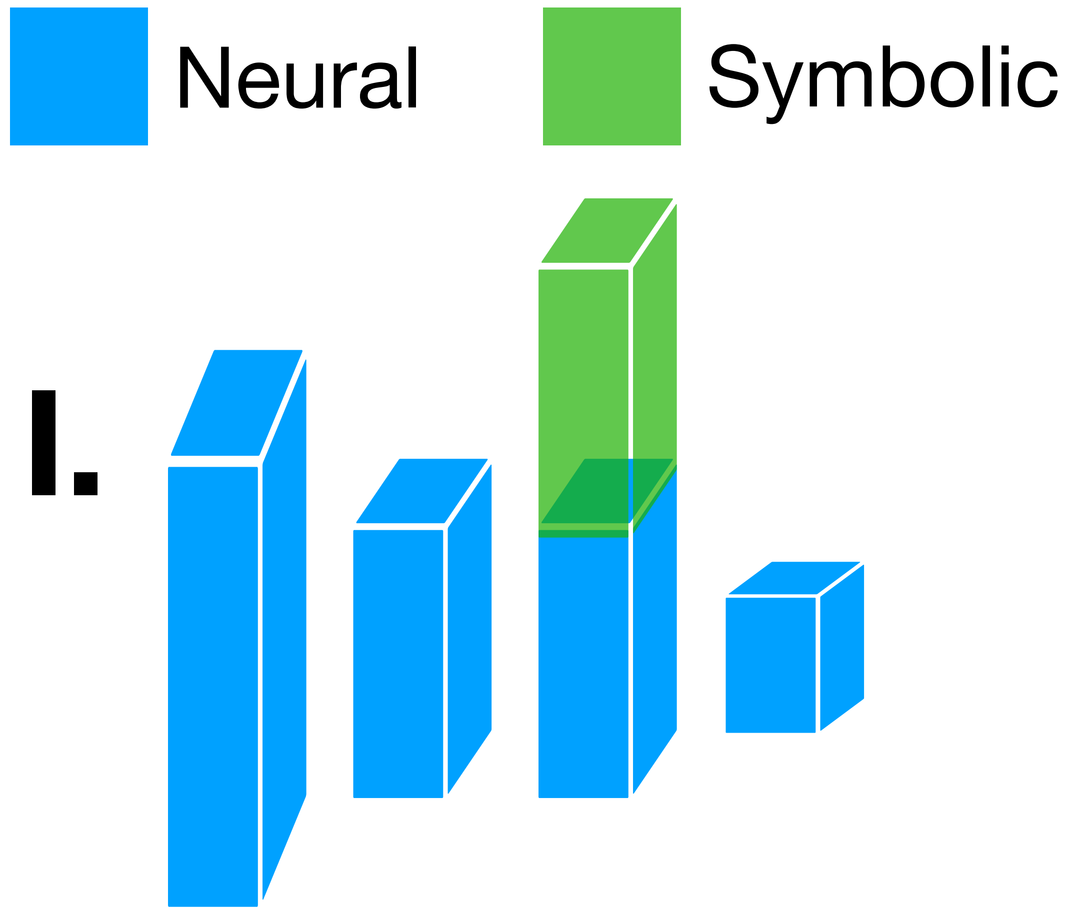
❌ **End-to-end training**



❌ **Reliable & efficient reasoning**

Challenge 2 Reasoning over symbolic constraints is *#P-hard* in general

My *Contributions* to Neurosymbolic AI



I.

- *Differentiable learning* through the symbolic constraint
  - Flexible integration of the symbolic constraint

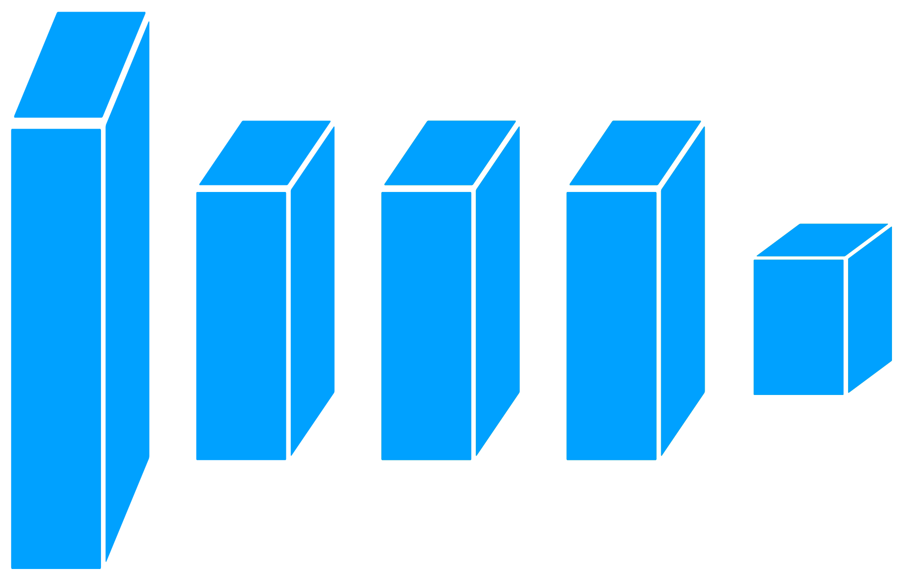
My *Contributions* to Neurosymbolic AI

Neural   Symbolic

I.

- *Differentiable learning* through the symbolic constraint
    - Flexible integration of the symbolic constraint

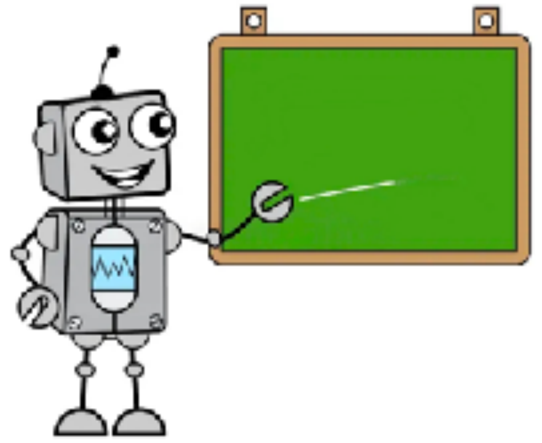
II.



- *Reliable Reasoning* over neural models via symbolic tools
    - Deep connection between neural and symbolic reasoning

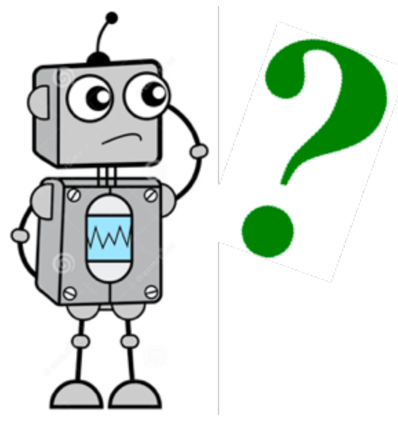
My *Contributions* to Neurosymbolic AI

Part 1
Differentiable Learning



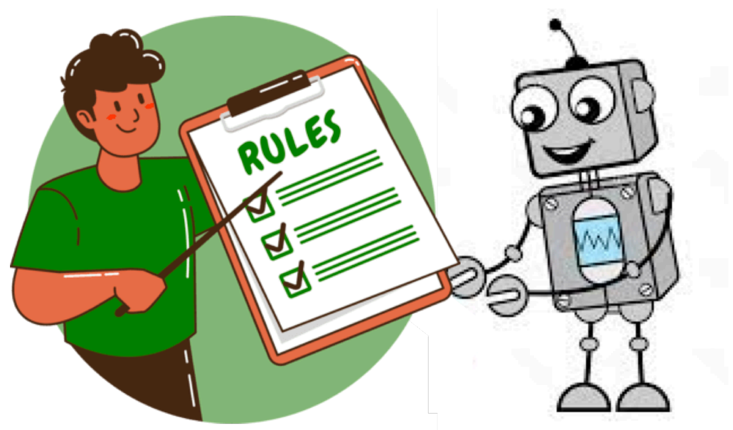
- Explain *why*

Part 2
Reliable Reasoning



- Know what they *don't know*

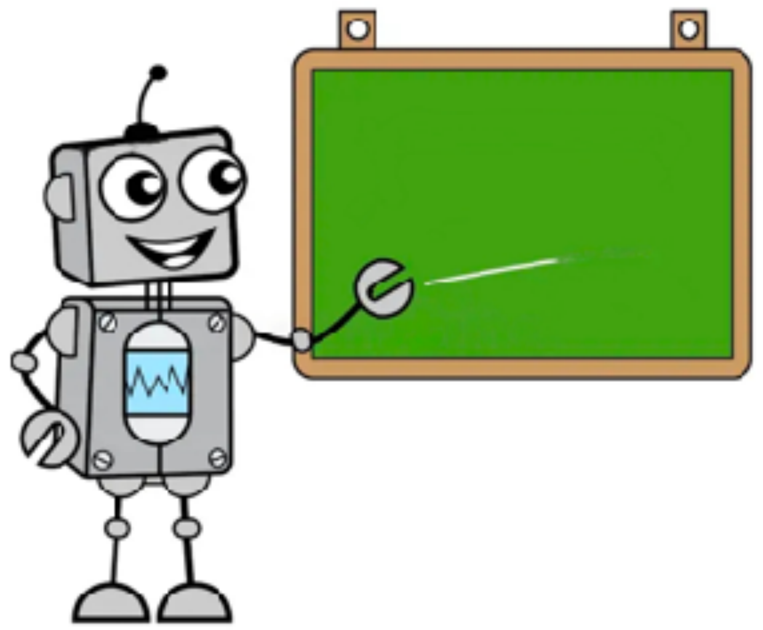
Part 3
Applications



- Follow *rules/domain knowledge*

Part 1

How to do *differentiable learning* under symbolic constraints?



- Explain *why*

Learn to Explain (L2X) — *Definition*

Data Beer review *(review, score)*

Goal Learn a model that outputs

- Predicted scores
- ***k***-word explanation

Input (*review*)

a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.

Explanation (*k*=5)

a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.

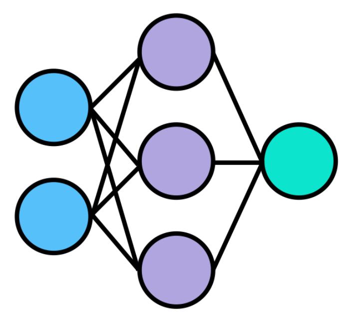
Output

Chen, Jianbo, et al. "Learning to explain: An information-theoretic perspective on model interpretation." ICML, 2018.

Learn to Explain (L2X) — *Existing Method*

Input (*review*)

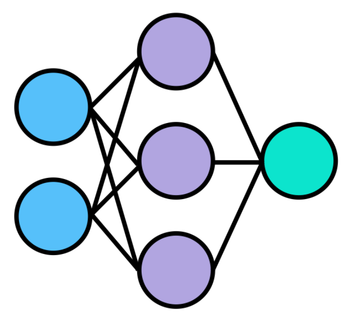
a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.



word$_i$	**prob$_i$**
little	θ_1
wood	θ_2
caramel	θ_3
…	…

Explanation

word$_i$	**select? z_i**
little	TRUE
wood	TRUE
caramel	FALSE
…	…



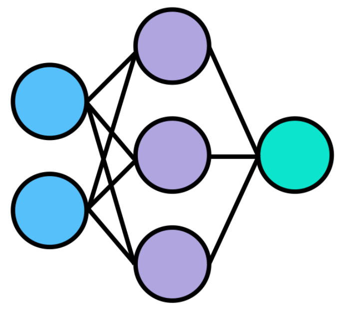
Output

Pure neural method? sparsity regularization doesn't really work

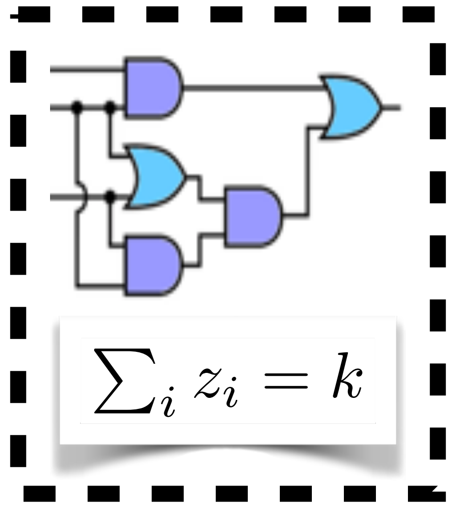
Why? Optimizing for $\mathbb{E}[\sum z_i] = \sum \theta_i = k$ does not guarantee $\sum z_i = k$

Learn to Explain (L2X) — *Symbolic Method*

Input (*review*)
a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.



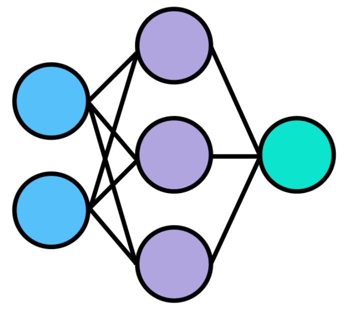
word$_i$	**prob$_i$**
little	θ_1
wood	θ_2
caramel	θ_3
…	…



Symbolic

Explanation

word$_i$	**select? z_i**
little	TRUE
wood	TRUE
caramel	FALSE
…	…



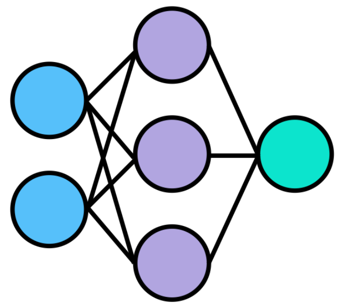
Output

subset	**prob**
ø	$(1-\theta_1)(1-\theta_2)(1-\theta_3)$
{little}	$\theta_1(1-\theta_2)(1-\theta_3)$
{wood}	$(1-\theta_1)\theta_2(1-\theta_3)$
{caramel}	$(1-\theta_1)(1-\theta_2)\theta_3$
{little, wood}	$\theta_1\theta_2(1-\theta_3)$
{little, caramel}	$\theta_1(1-\theta_2)\theta_3$
{wood, caramel}	$(1-\theta_1)\theta_2\theta_3$
{little, wood, caramel}	$\theta_1\theta_2\theta_3$

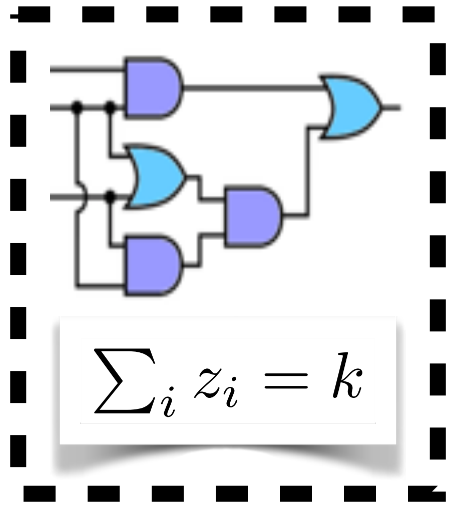
Learn to Explain (L2X) — *Symbolic Method*

Input (*review*)

a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.



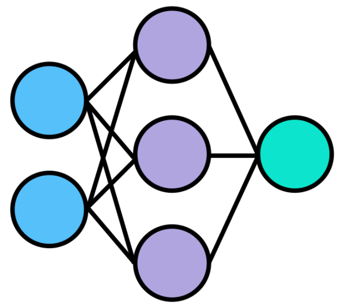
word$_i$	**prob$_i$**
little	θ_1
wood	θ_2
caramel	θ_3
…	…



Symbolic

Explanation

word$_i$	**select? z_i**
little	TRUE
wood	TRUE
caramel	FALSE
…	…



Output

subset	**conditional prob***
{little, wood}	$\theta_1\theta_2(1-\theta_3)\ /\ A$
{little, caramel}	$\theta_1(1-\theta_2)\theta_3\ /\ A$
{wood, caramel}	$(1-\theta_1)\theta_2\theta_3\ /\ A$

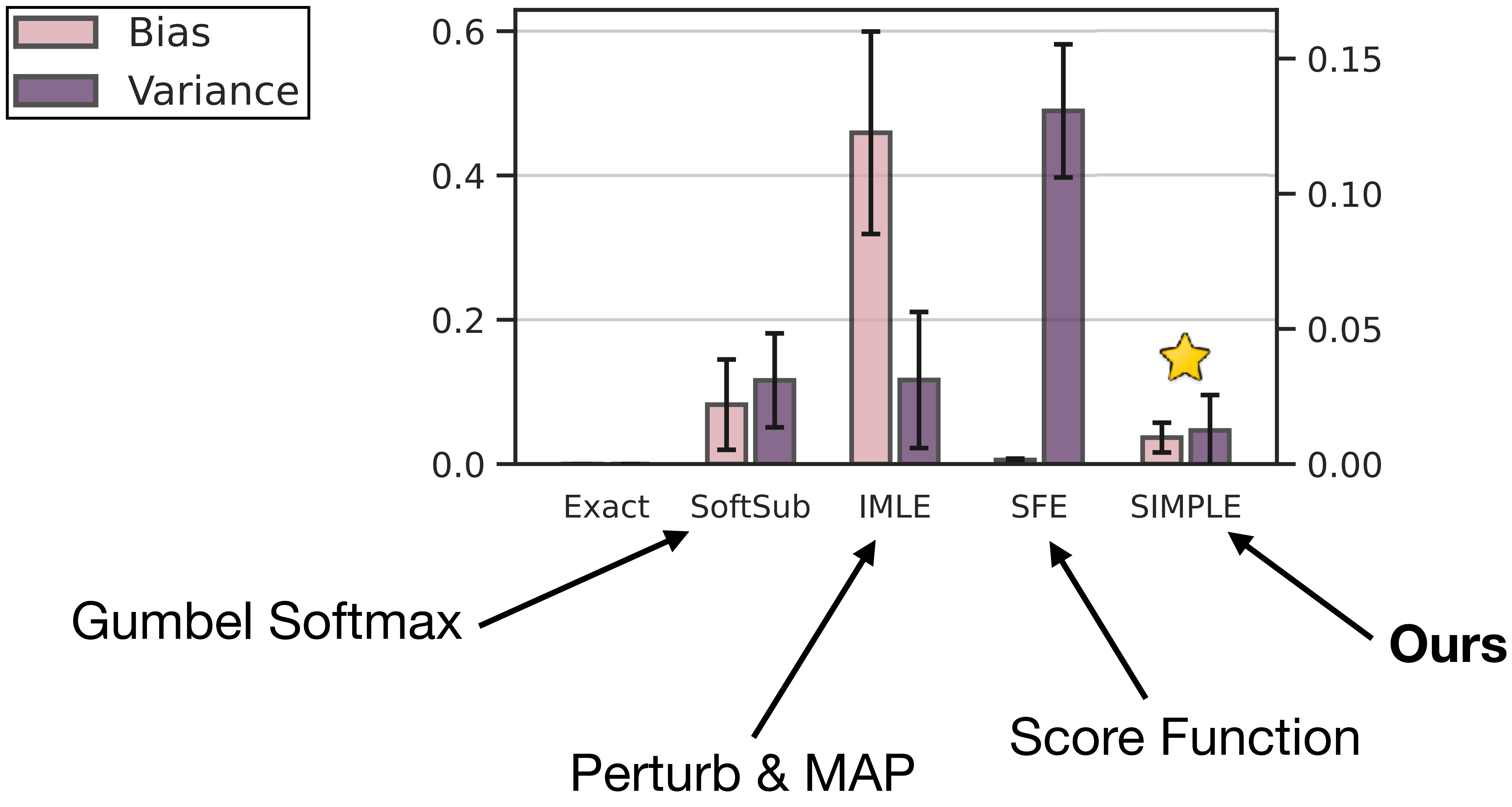
*normalizing constant:

$$A = \theta_1\theta_2(1-\theta_3) + \theta_1(1-\theta_2)\theta_3 + (1-\theta_1)\theta_2\theta_3$$

- How to ***sample*** from the constrained distribution?
- How to ***differentiate*** the sampling?

SIMPLE: Gradient Estimator for k-Subset Constraint

We achieve ***lower bias and variance*** by ***exact*** sampling and ***exact*** constraint probability



[AZNB, ICLR'23]

SIMPLE: Gradient Estimator for k-Subset Constraint

Symbolic constraint $\sum_i z_i = k$

$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \sum_i z_i = k)$

Log probability $\boldsymbol{\theta}$

Distribution

[AZNB, ICLR'23]

SIMPLE: Gradient Estimator for k-Subset Constraint

Symbolic constraint $\sum_i z_i = k$

Log probability $\boldsymbol{\theta}$

$$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \textstyle\sum_i z_i = k)$$

Distribution

Thm. Exact sampling can be achieved in a *dynamic-programming* way

sample $= [\,], j = k$
for $i = n$ **to** 1 **do**
  // $a[i, j] := p(\sum_{m=1}^{i-1} z_m = j - 1)$ ← **Constraint Probability**
  $p = a[i-1, j-1]$
  $z_i \sim \text{Bernoulli}(p \cdot p_{\theta_i}(z_i = 1)/a[i, j])$
  // Pick next state based on value of sample
  if $z_i = 1$ **then** $j = j - 1$
  sample.append(z_i)

[AZNB, ICLR'23]

SIMPLE: Gradient Estimator for k-Subset Constraint

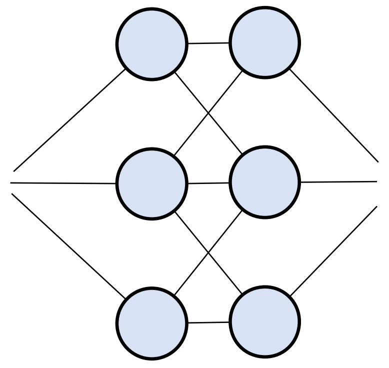
Symbolic constraint $\sum_i z_i = k$

Log probability $\boldsymbol{\theta}$

$$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \textstyle\sum_i z_i = k)$$

$\mathbf{z}$

$f_{\mathbf{u}}$



ℓ

Distribution

[A**Z**NB, ICLR'23]

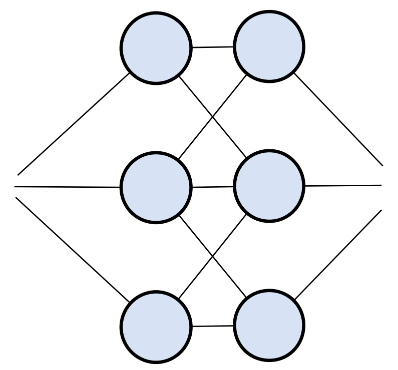
SIMPLE: Gradient Estimator for k-Subset Constraint

Symbolic constraint $\sum_i z_i = k$

$$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \textstyle\sum_i z_i = k)$$

$\mathbf{z}$

$f_{\mathbf{u}}$



ℓ

Log probability $\boldsymbol{\theta}$

$$\boxed{\partial_{\boldsymbol{\theta}} \mathbf{z}} \nabla_{\mathbf{z}} \ell$$

discrete sample

Distribution

[AZNB, ICLR'23]

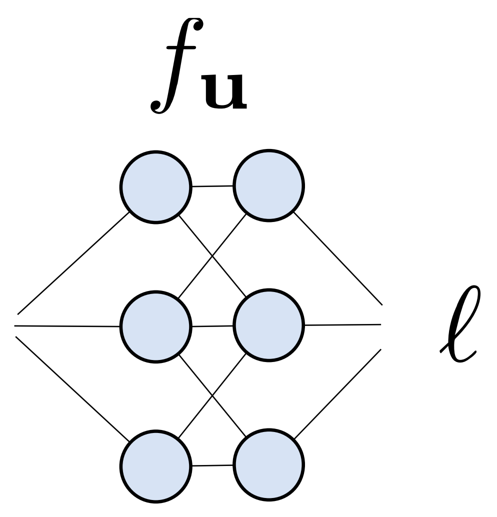
SIMPLE: Gradient Estimator for k-Subset Constraint

Symbolic constraint $\sum_i z_i = k$

$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \sum_i z_i = k)$

Log probability $\boldsymbol{\theta}$

$\mathbf{z}$



$\boxed{\partial_{\boldsymbol{\theta}} \mathbf{z}} \nabla_\mathbf{z} \ell$

discrete sample

Distribution

Intuition: *"Don't select 1st word"*

z_1	1	➡	0
$p_\theta(z_1 = 1 \mid \sum_i z_i = k)$	0.8	➡	0.3

[A**Z**NB, ICLR'23]

SIMPLE: Gradient Estimator for k-Subset Constraint

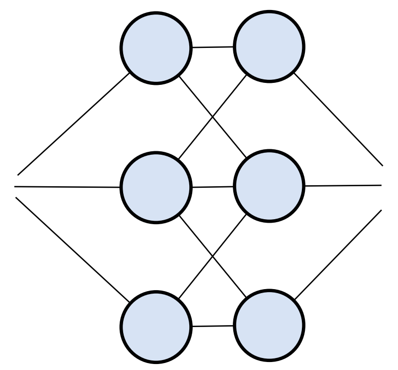
Symbolic constraint $\sum_i z_i = k$

Log probability $\boldsymbol{\theta}$

$$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \textstyle\sum_i z_i = k)$$

$\mathbf{z}$

$f_{\mathbf{u}}$



ℓ

$$\boxed{\partial_{\boldsymbol{\theta}} \boldsymbol{\mu}(\boldsymbol{\theta})} \nabla_{\mathbf{z}} \ell$$

marginal

$$\mu(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}\left(z_i \mid \textstyle\sum_j z_j = k\right)$$

Distribution

[A**Z**NB, ICLR'23]

SIMPLE: Gradient Estimator for k-Subset Constraint

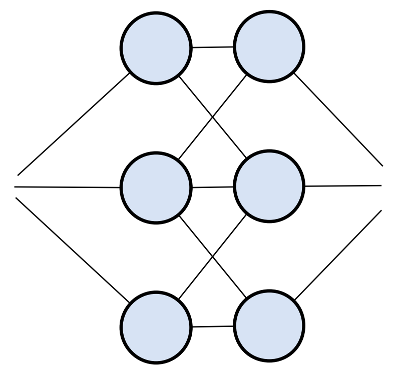
Symbolic constraint $\sum_i z_i = k$

$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \sum_i z_i = k)$

Log probability $\boldsymbol{\theta}$

$\mathbf{z}$

$f_{\mathbf{u}}$



ℓ

$\boxed{\partial_{\boldsymbol{\theta}} \boldsymbol{\mu}(\boldsymbol{\theta})} \nabla_{\mathbf{z}} \ell$

marginal

Prop. conditional marginals can be obtained by $\mu(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}\left(z_i \mid \sum_j z_j = k\right) = \frac{\partial}{\partial \theta_i} \log \underline{p_{\boldsymbol{\theta}}(\sum_j z_j = k)}$

Constraint Probability

[A**Z**NB, ICLR'23]

Evaluation — *Learn to Explain*

Dataset

- Beer Review

Input (*review*)	Output
a little like wood and caramel with a hop finish. has a sort of fruity flavor like grapes or cherry that is sort of buried in there. mouth feel was lite, sort of bubbly.	0.75

- Human annotations of the words that best describe the review

Predictive performance

	Mean Squared Error ↓	Precision ↑
SIMPLE (Ours)	**2.11**	**42.31**
L2X	6.92	32.23
SoftSub	2.18	41.98
IMLE	2.38	41.38

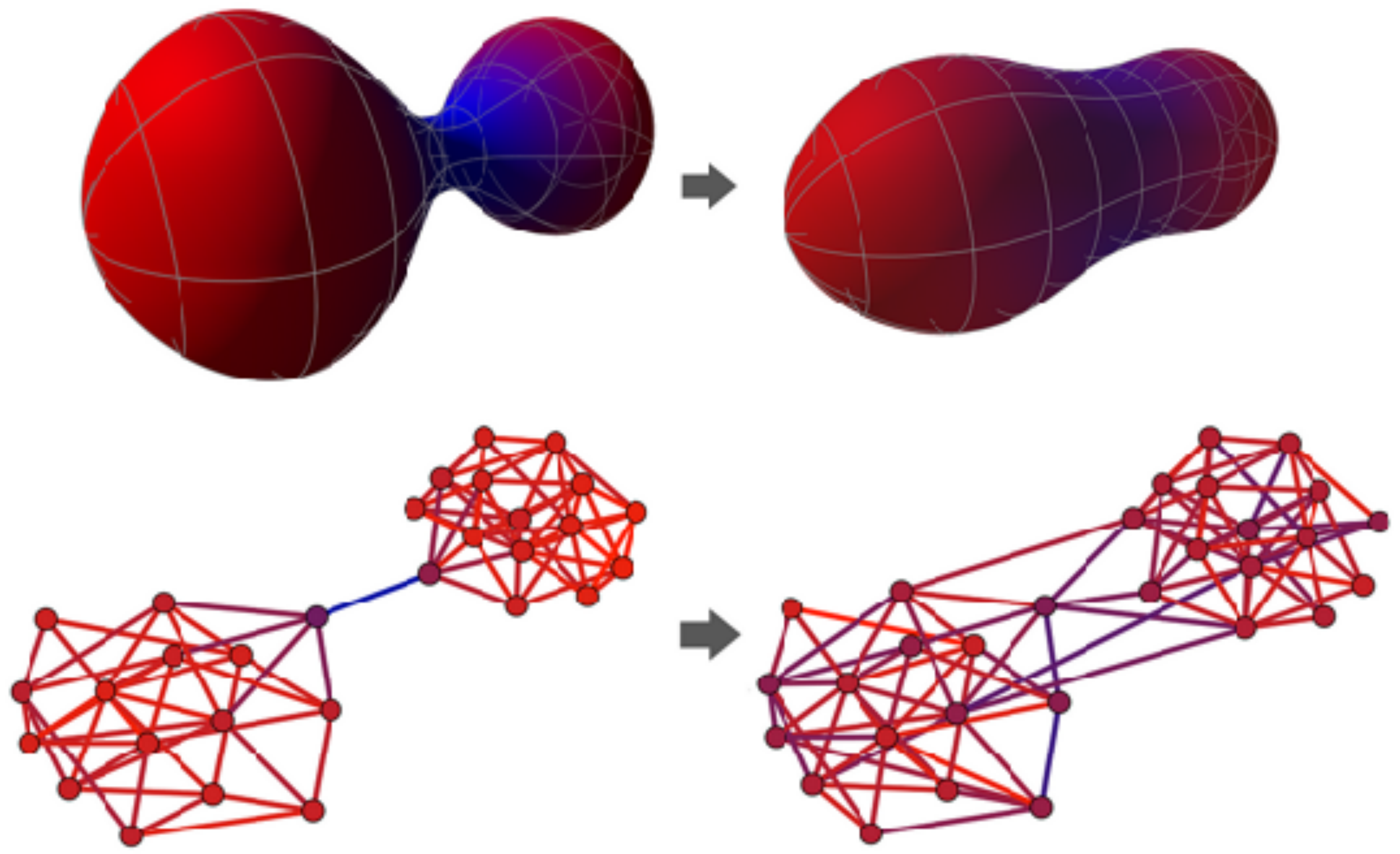
How much the learned explanations match the ones from human

[A**Z**NB, ICLR'23]

Evaluation — *Graph Neural Networks*

Graph rewiring for learning graph structures

Best 2nd-best 3rd-best



Model	MUTAG	PTC_MR	PROTEINS	NCI1	NCI109
GK ($k=3$) (Shervashidze et al., 2009)	81.4±1.7	55.7±0.5	71.4±0.3	62.5±0.3	62.4±0.3
PK (Neumann et al., 2016)	76.0±2.7	59.5±2.4	73.7±0.7	82.5±0.5	N/A
WL kernel (Shervashidze et al., 2011)	90.4±5.7	59.9±4.3	75.0±3.1	**86.0**±1.8	N/A
DGCNN (Zhang et al., 2018)	85.8±1.8	58.6±2.5	75.5±0.9	74.4±0.5	N/A
IGN (Maron et al., 2019b)	83.9±13.0	58.5±6.9	76.6±5.5	74.3±2.7	72.8±1.5
GIN (Xu et al., 2019)	89.4±5.6	64.6±7.0	76.2±2.8	82.7±1.7	N/A
PPGNs (Maron et al., 2019a)	90.6±8.7	66.2±6.6	77.2±4.7	83.2±1.1	82.2±1.4
Natural GN (de Haan et al., 2020)	89.4±1.6	66.8±1.7	71.7±1.0	82.4±1.3	83.0±1.9
GSN (Bouritsas et al., 2022)	92.2±7.5	68.2±7.2	76.6±5.0	83.5±2.0	83.5±2.3
CIN (Bodnar et al., 2021)	92.7±6.1	68.2±5.6	77.0±4.3	83.6±1.4	**84.0**±1.6
CAN (Giusti et al., 2023a)	**94.1**±4.8	**72.8**±8.3	**78.2**±2.0	84.5±1.6	83.6±1.2
CIN++ (Giusti et al., 2023b)	**94.4**±3.7	**73.2**±6.4	**80.5**±3.9	**85.3**±1.2	**84.5**±2.4
FoSR (Karhadkar et al., 2022)	86.2±1.5	58.5±1.7	75.1±0.8	72.9±0.6	71.1±0.6
DropGNN (Papp et al., 2021)	90.4±7.0	66.3±8.6	76.3±6.1	81.6±1.8	80.8±2.6
PR-MPNN (10-fold CV) (Ours)	**98.4**±2.4	**74.3**±3.9	**80.7**±3.9	**85.6**±0.8	**84.6**±1.2

[QMA**Z**BNM, ICLR'24]

SIMPLE: Gradient Estimator for k-Subset Constraint

Symbolic constraint $\sum_i z_i = k$

```
sample = [ ], j = k
for i = n to 1 do
    // a[i, j] := p(∑_{m=1}^{i-1} z_m = j − 1)
    p = a[i − 1, j − 1]
    z_i ∼ Bernoulli(p · p_{θ_i}(z_i = 1)/a[i, j])
    // Pick next state based on value of sample
    if z_i = 1 then j = j − 1
    sample.append(z_i)
```

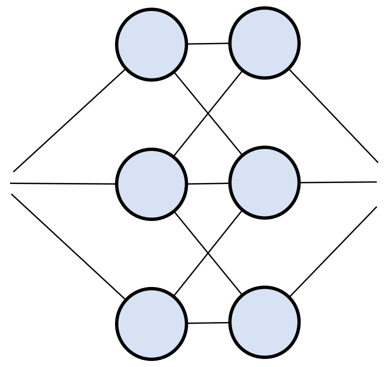
Constraint Probability

Log probability $\boldsymbol{\theta}$

$$\mathbf{z} \sim p_\theta(\mathbf{z} \mid \textstyle\sum_i z_i = k)$$

$\mathbf{z}$

$f_{\mathbf{u}}$



ℓ

$$\boxed{\partial_{\boldsymbol{\theta}} \boldsymbol{\mu}(\boldsymbol{\theta})} \nabla_{\mathbf{z}} \ell$$

marginal

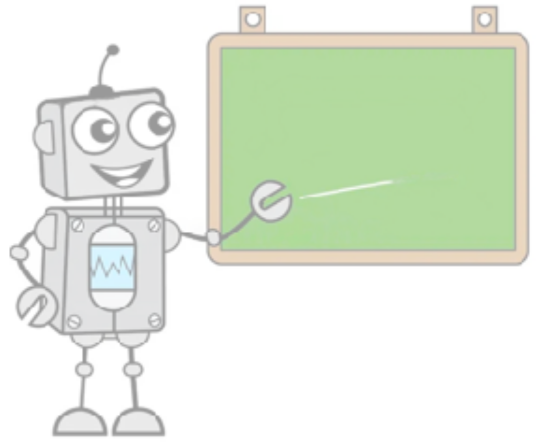
$$\mu(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}\Big(z_i \mid \textstyle\sum_j z_j = k\Big) = \dfrac{\partial}{\partial \theta_i} \log \underbrace{p_{\boldsymbol{\theta}}(\textstyle\sum_j z_j = k)}_{\text{Constraint Probability}}$$

Key: *constraint probability*!

[AZNB, ICLR'23]

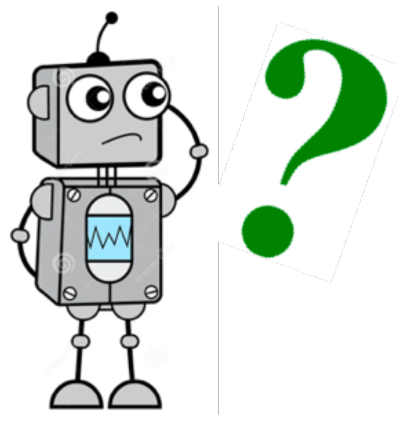
Outline

Part 1
Differentiable Learning



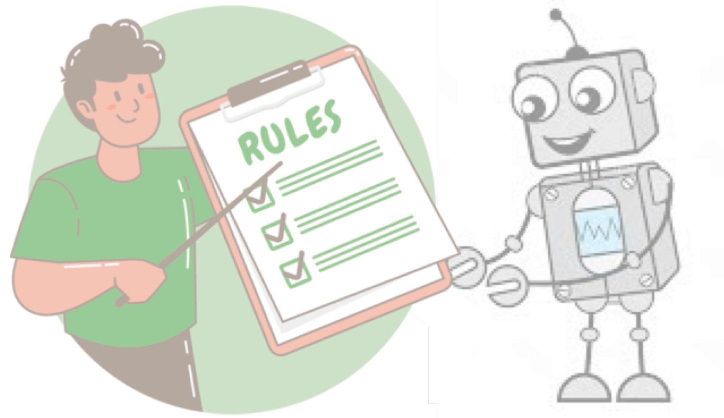
- Explain *why*

Part 2
Reliable Reasoning



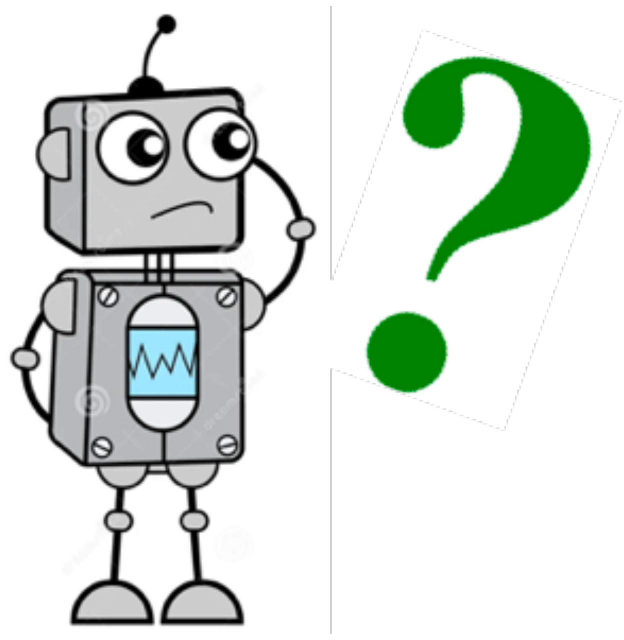
- Know what they *don't know*

Part 3
Applications



- Follow *rules/domain knowledge*

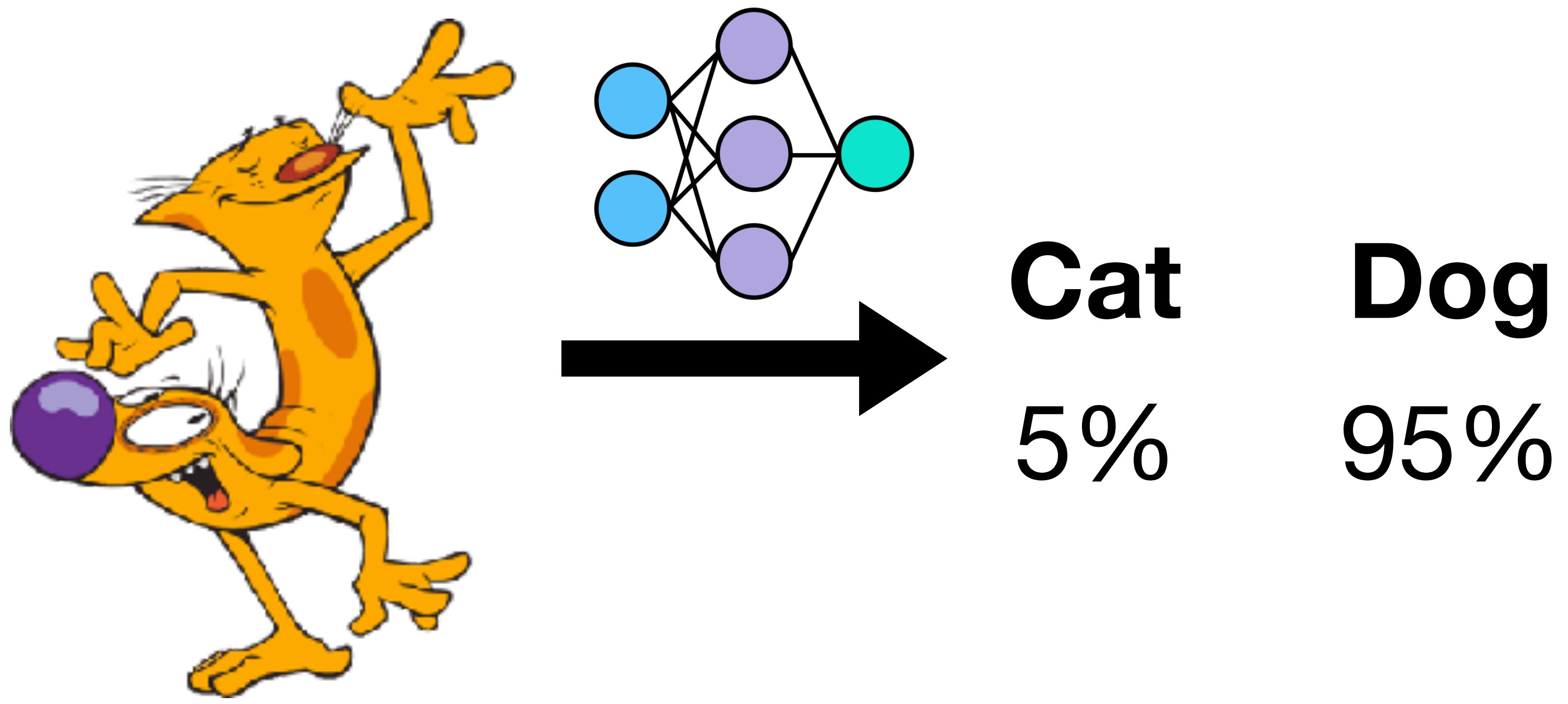
How to *quantify uncertainty* using symbolic reasoning?

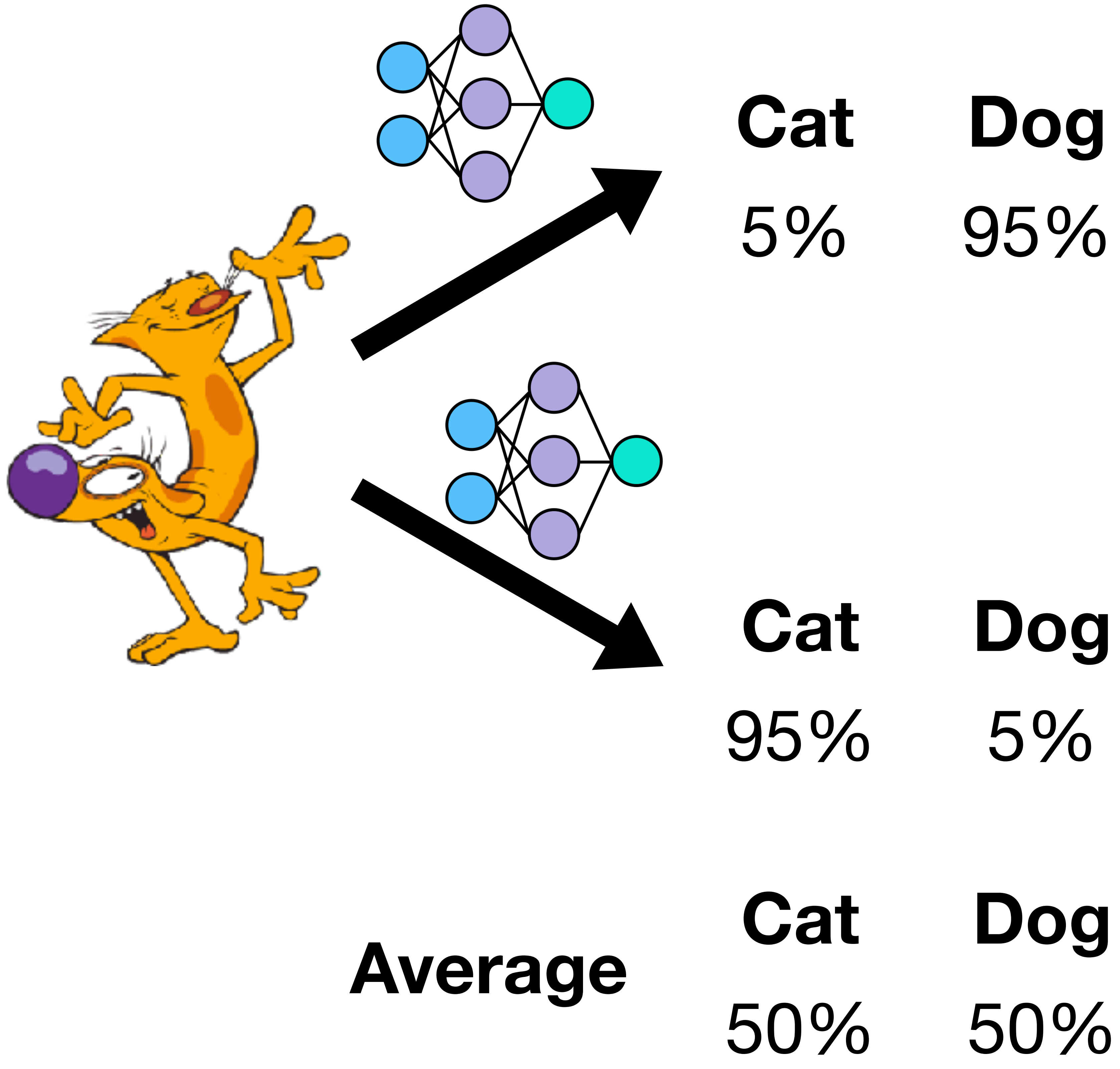


- Know what they *don't know*

Uncertainty Quantification

- Neural networks are overconfident
- Point estimate is risky

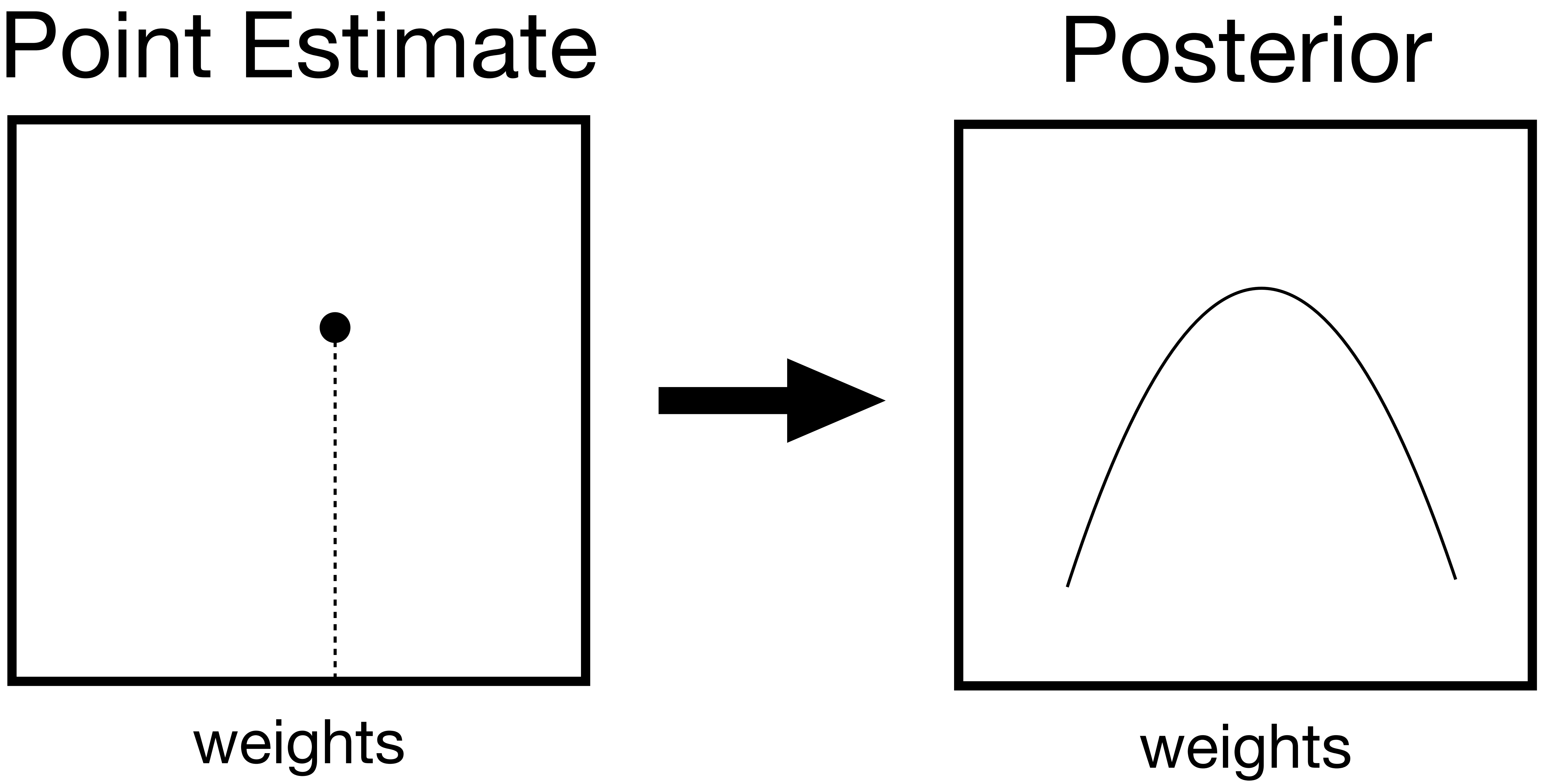




➡**Bayesian Deep Learning** for *robust* and *reliable* predictions

Bayesian Model Average (BMA)

Key idea



$p(y \mid x, w)$

$$p(y \mid x) = \int p(y \mid x, w) p(w) \, dw$$

Motivation

- **Goal:** Bayesian model average

  Predictive posterior $p(y \mid x) = \int p(y \mid x, w) p(w)\, dw$

  Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x)\, dy$

- **Challenge:** DNNs are too big!

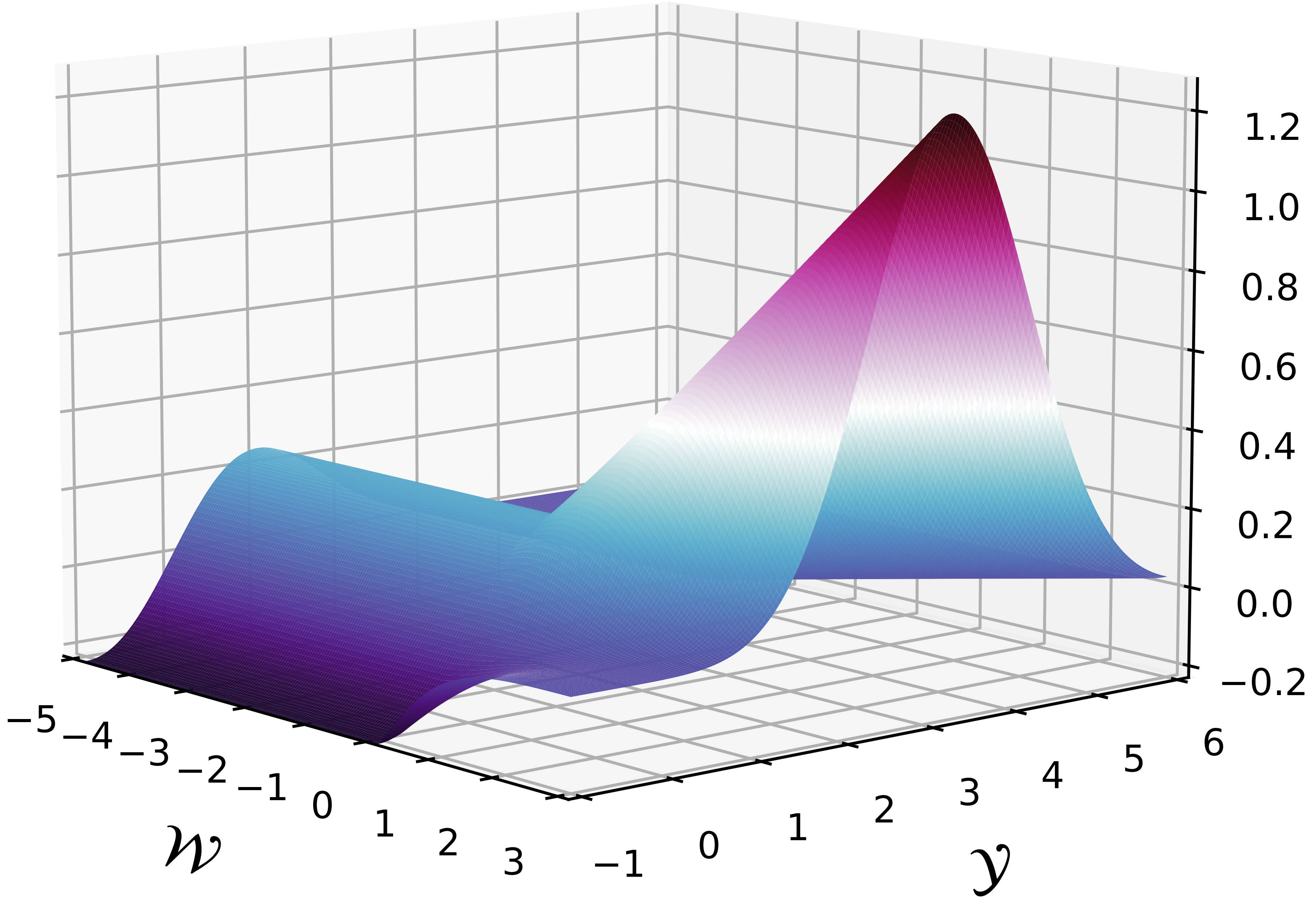
  ➡ *Costly* to maintain too many samples

  ➡ *Low sample efficiency* given the complex integrand

How complex? 🤔

How *complex* is the integrand? Simplest case!

Expected prediction $\mathbb{E}[y] = \int y \, \underbrace{p(w \mid D)}_{\text{Uniform}} \, \underbrace{p(y \mid \underbrace{f(x)}_{\text{Single ReLU}}, w)}_{\text{Gaussian}} \, dw \, dy$



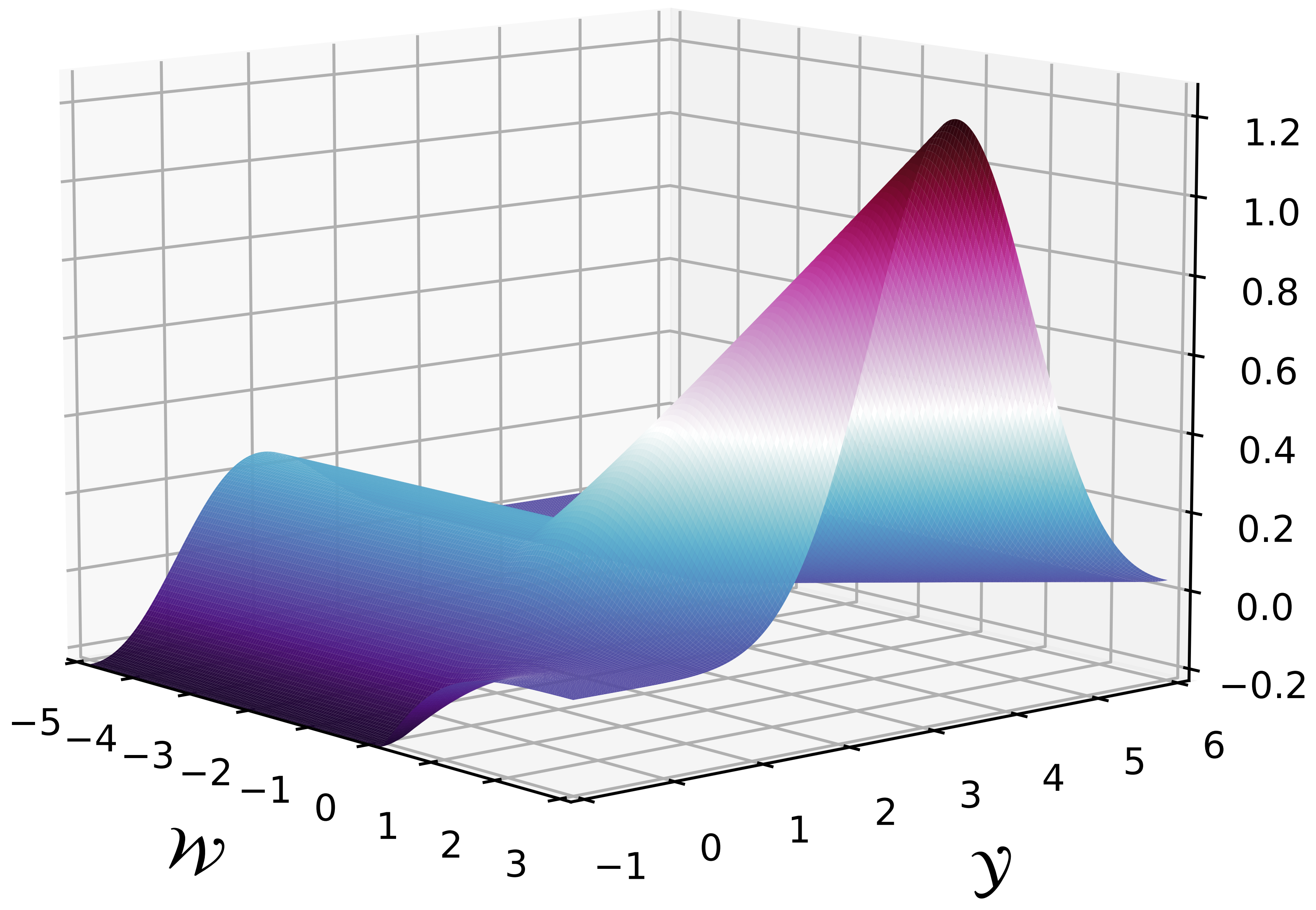
Non-convex, multi-modal, no closed form 🤯

Motivation

- **Goal:** Bayesian model average

Predictive posterior $p(y \mid x) = \int p(y \mid x, w) p(w) \, dw$

Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x) \, dy$

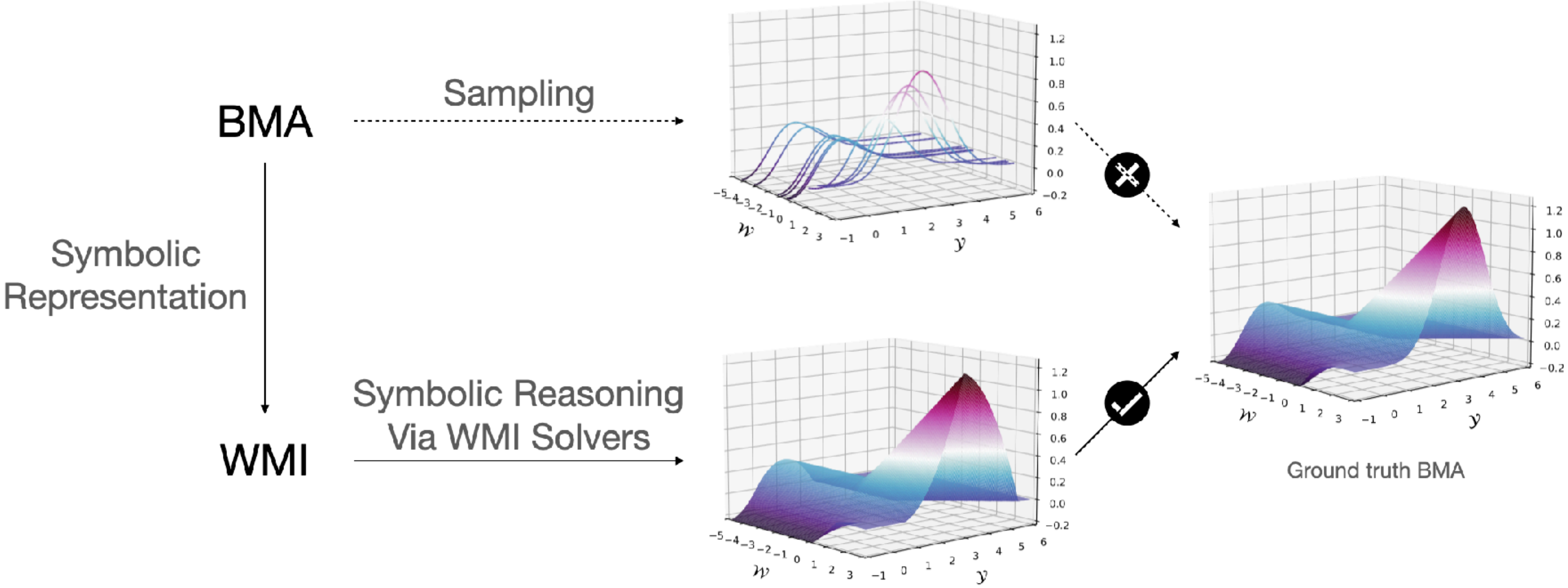


Is there a *better* way to estimate the integral than *sampling*?

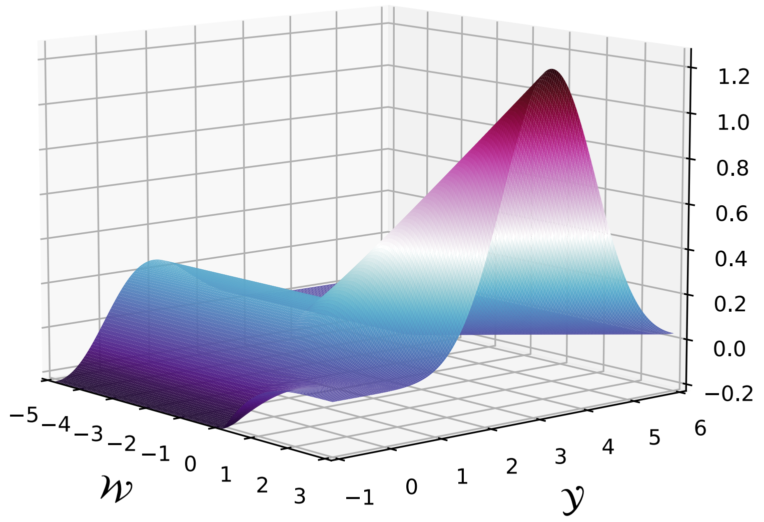
Symbolic! 🤩

Idea

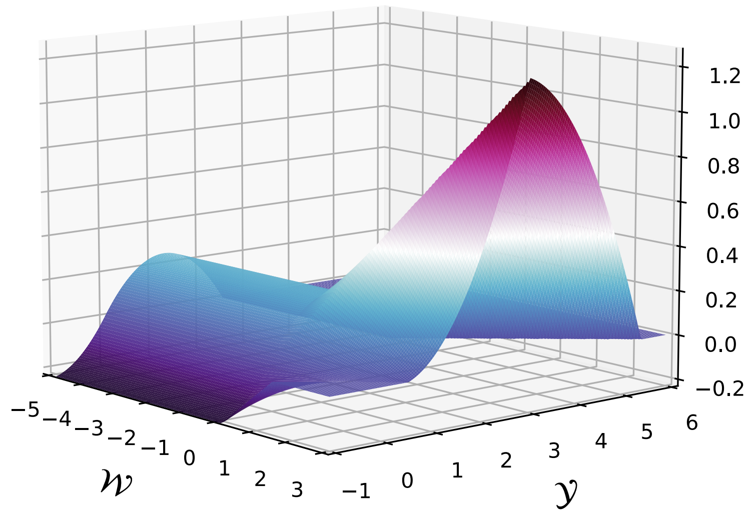
A reduction from BMA to Exact Integration



Ground truth BMA



$\approx$



Accurate approximation!

… but scalability?

Limitations

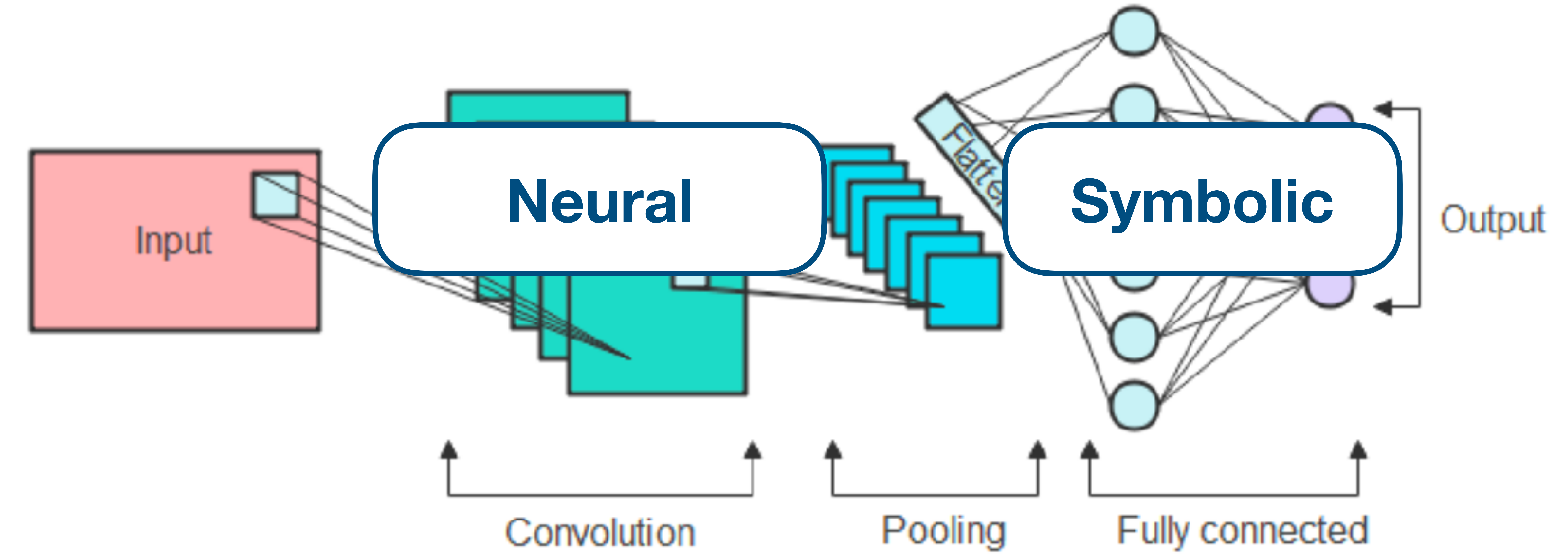
	Sampling	Symbolic
Accuracy	❌	✅
Flexibility	✅	❌*
Scalability	✅	❌**

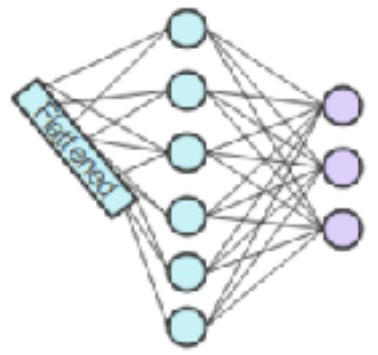
* Limited to fully connected layers

** Integration over polytopes in arbitrarily high dimensions is #P-hard

How to combine good from both worlds? 🤔

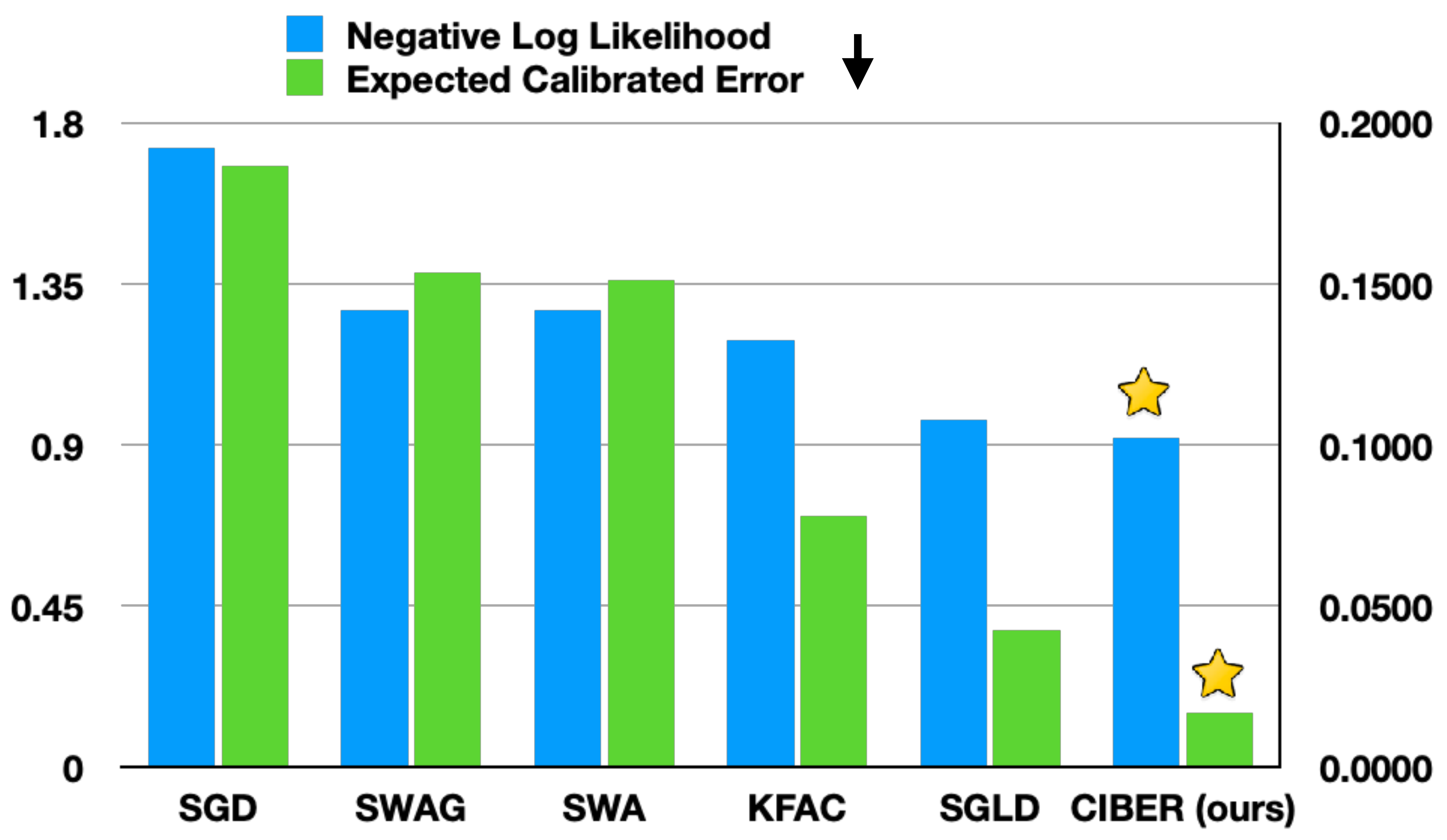
Collapsed Inference

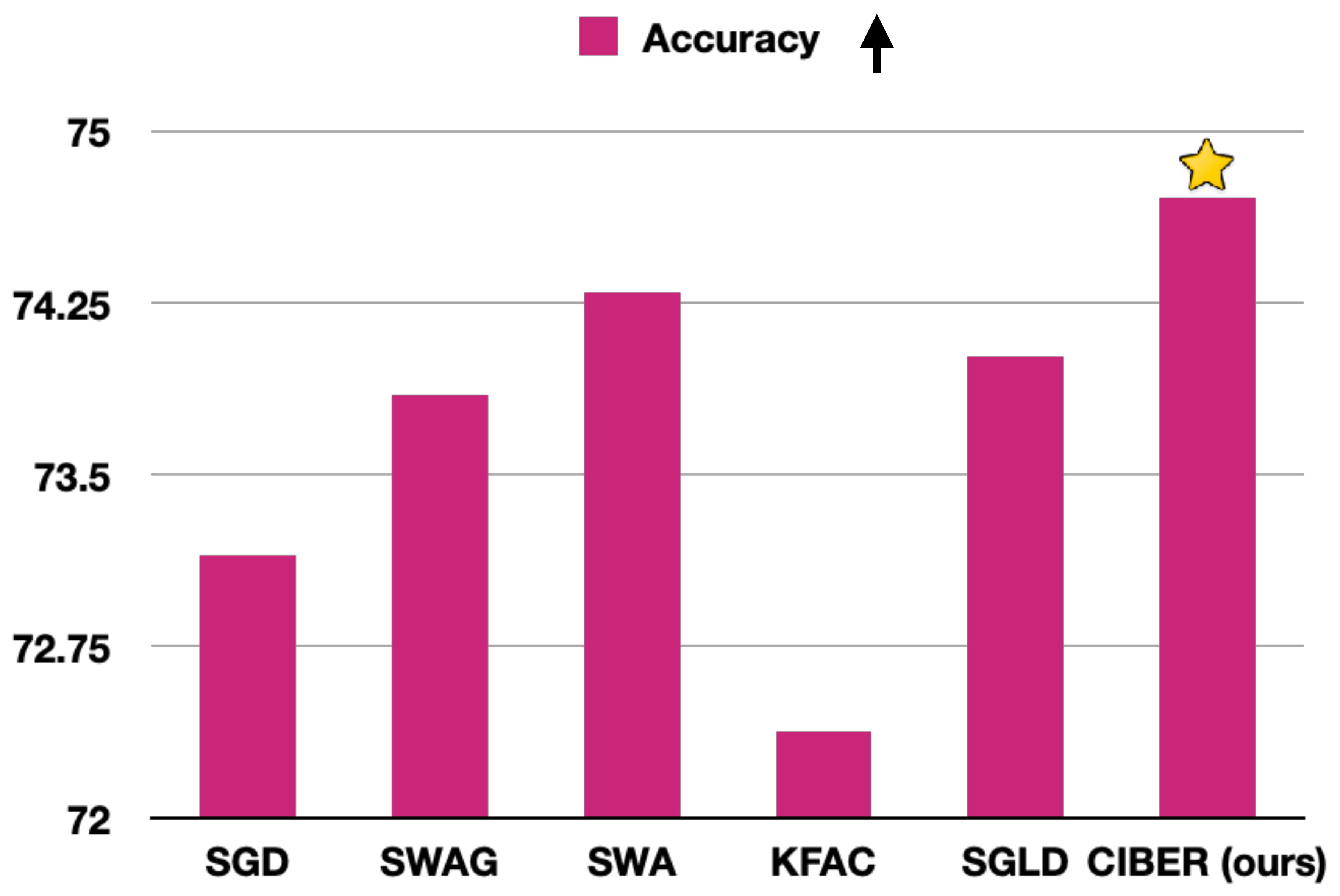


Expected prediction in BMA $\mathbb{E}[y] = \frac{1}{n}\sum_{w_s} \texttt{WMI}($  $)$

Accuracy + Flexibility, Scalability! 🥳

Evaluation *Image Classification on CIFAR-100*

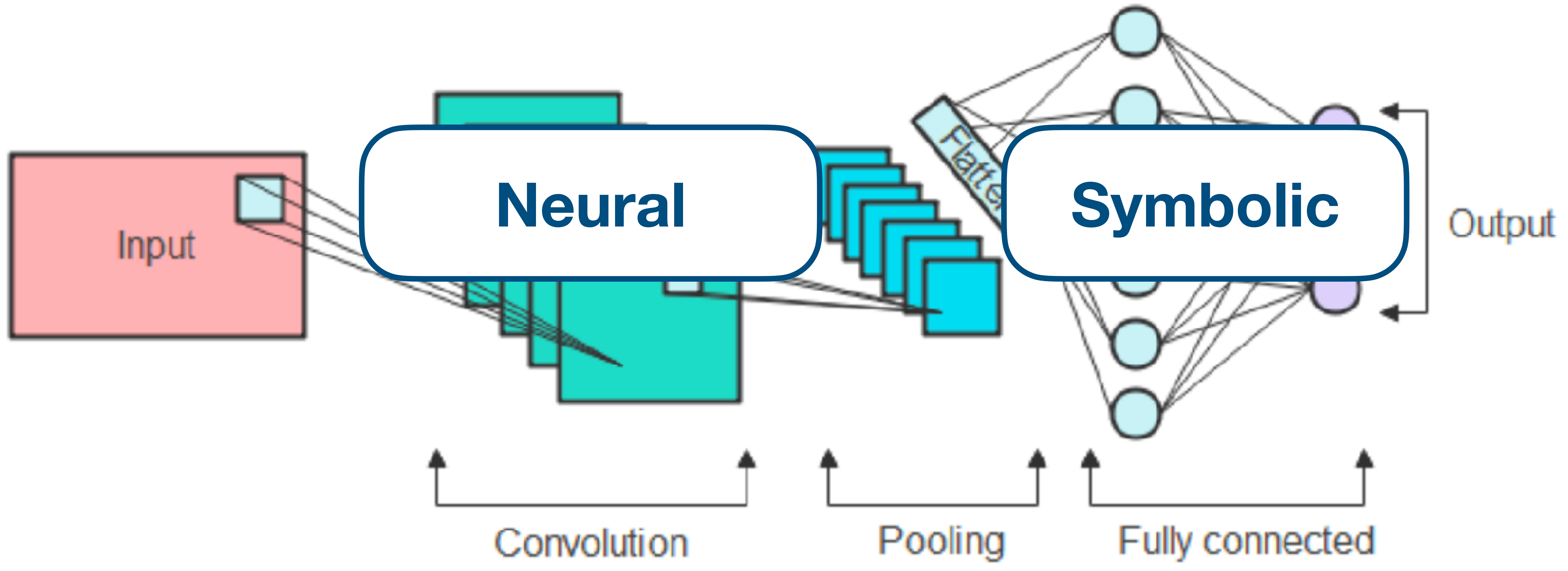




➡ *Accurate estimates of uncertainty*

➡ *Boost predictive performance*

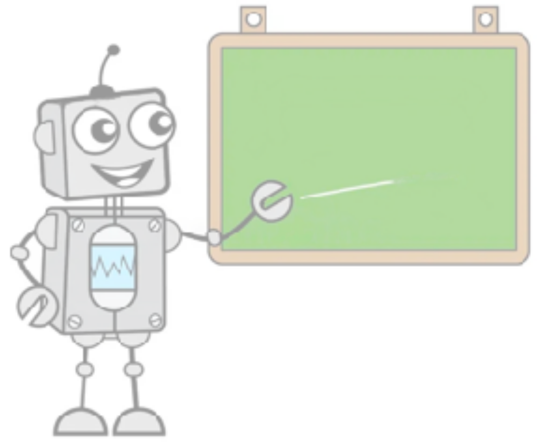
Collapsed Inference



Key: *A bit symbolic + neural models = immense power*

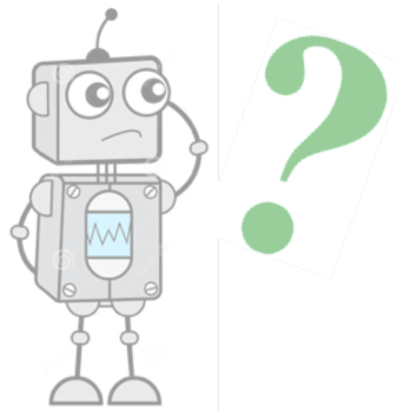
Outline

Part 1
Differentiable Learning



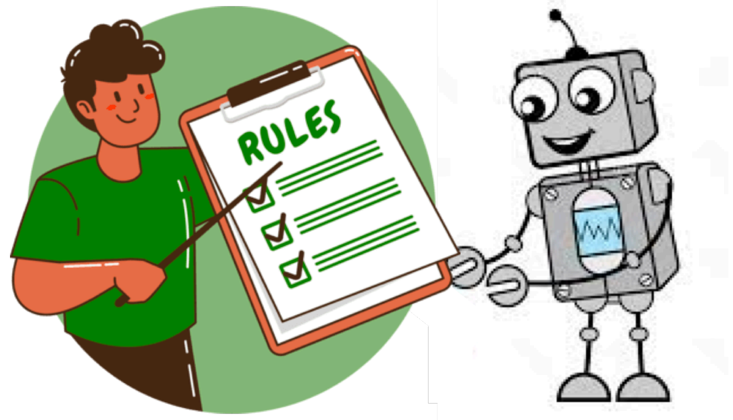
- Explain *why*

Part 2
Reliable Reasoning



- Know what they *don't know*

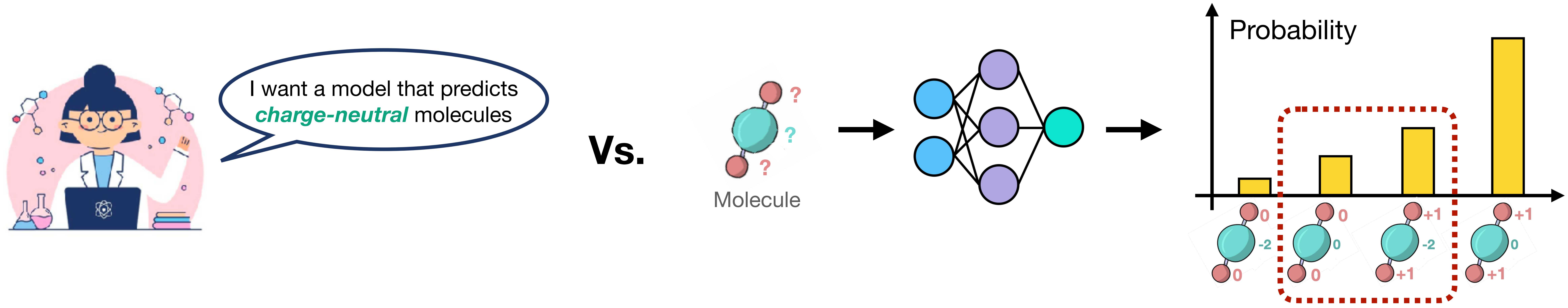
Part 3
Applications

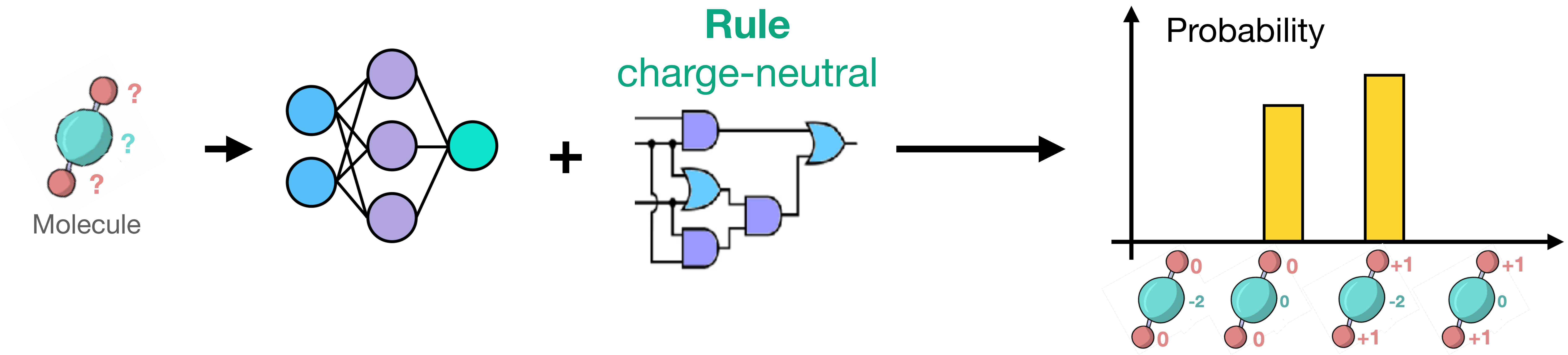


- Follow *rules/domain knowledge*

Neurosymbolic AI *for Science*

Molecule Modeling for Charge Prediction



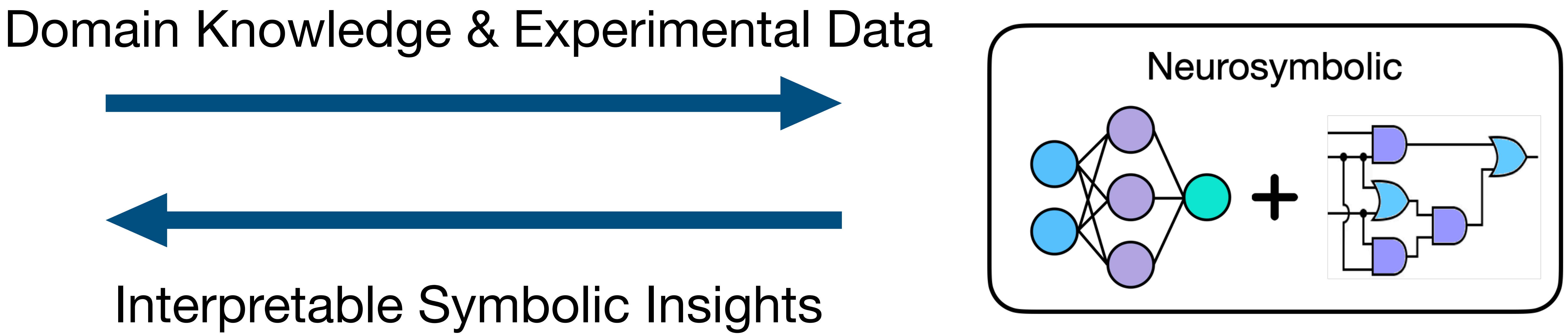


[LSB**Z**, NeurIPS Workshop'23]

Neurosymbolic AI *for Science*

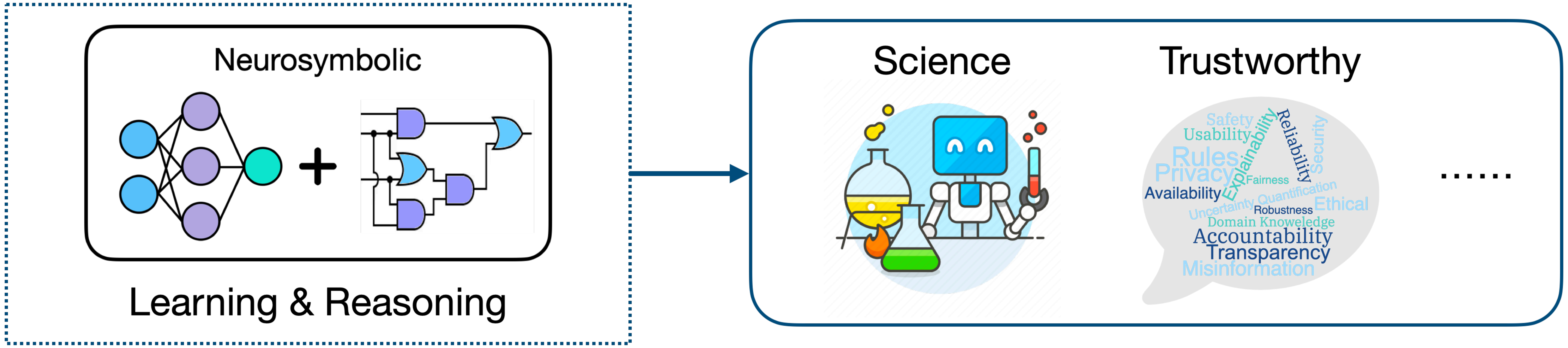
Future Research

Goal To establish a deep integration of AI/ML with science and accelerate scientific discovery



Future Research

Release the huge power of neurosymbolic and build the better AI ☺



Thank you!