That Doesn't Make Sense!

A Case Study in Actively Annotating Model Explanations

Sameer Singh

University of California, Irvine

NIPS 2017 Workshop on

Learning with Limited Labeled Data

Relation Extraction

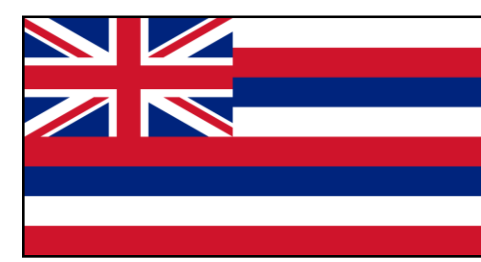
Given two entities, and all the sentences that mention them,
Identify the relations expressed between them.

birthplace
visited

Relation Extractor

Barack Obama

Hawaii



President Obama was born in Hawaii …
Obama is a native of the state of Hawaii..
The president visited Hawaii as part of his …
Barack often wears grass skirts, a common costume in Hawaii..

Understanding and Fixing Errors

Barack is married to Ann Dunham

That's wrong! Why did you predict that?

[-0.18, 0.35, 3.56, -2.45, 8.96, ...]?

Uhh, okay...
Here's more labeled data?
Barack, spouseOf, Ann Dunham ✗

Relation Extractor

User

ML Model

Can we explain predictions to help users understand and debug?

Injecting Knowledge

Most people are married to one person.
"is native to" is same as birthplace relation.

I don't understand. Give me labeled data.

Sigh… okay.
Barack, spouseOf, Michelle ✓
Barack, spouseOf, Ann Dunham ✗

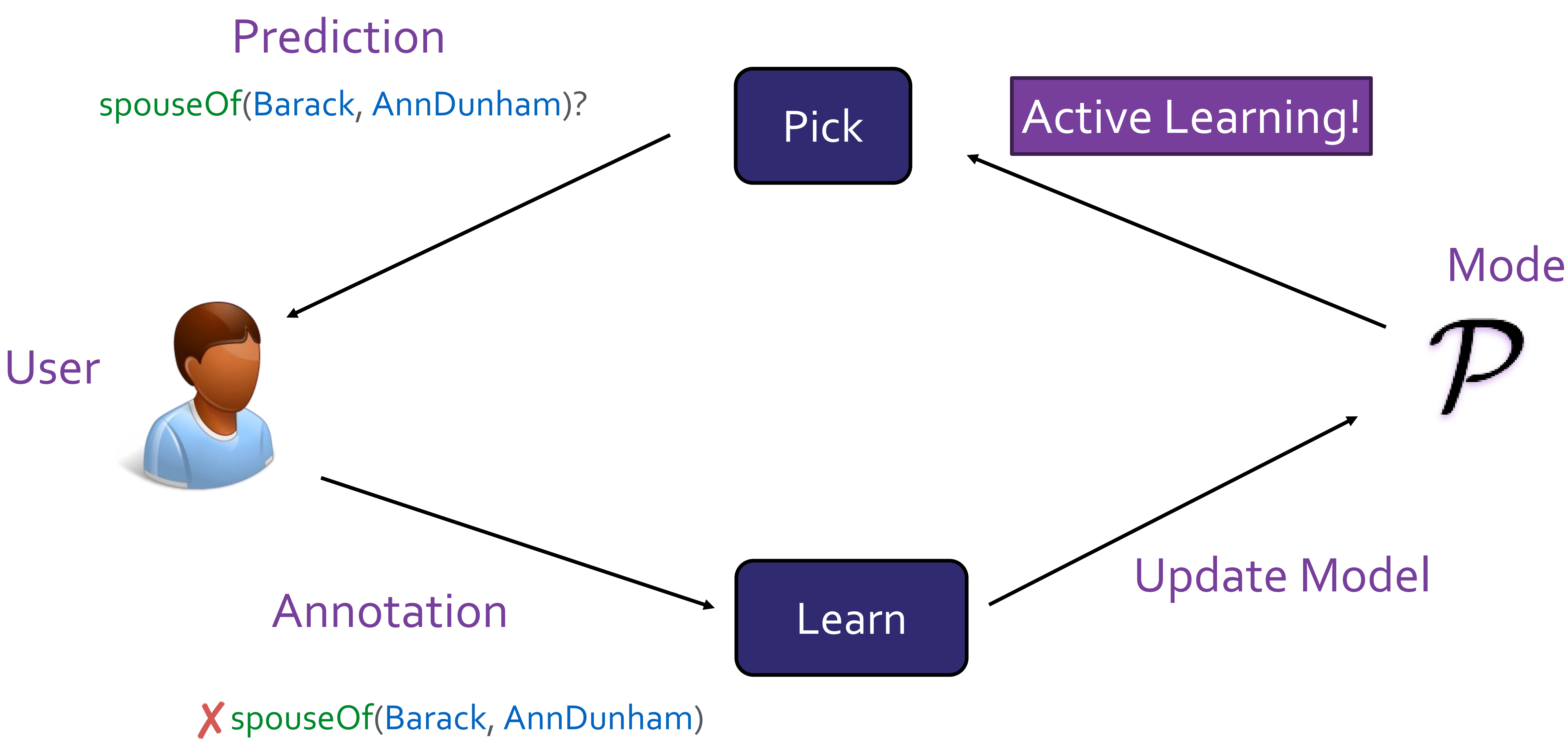
User

Relation Extractor

ML Model

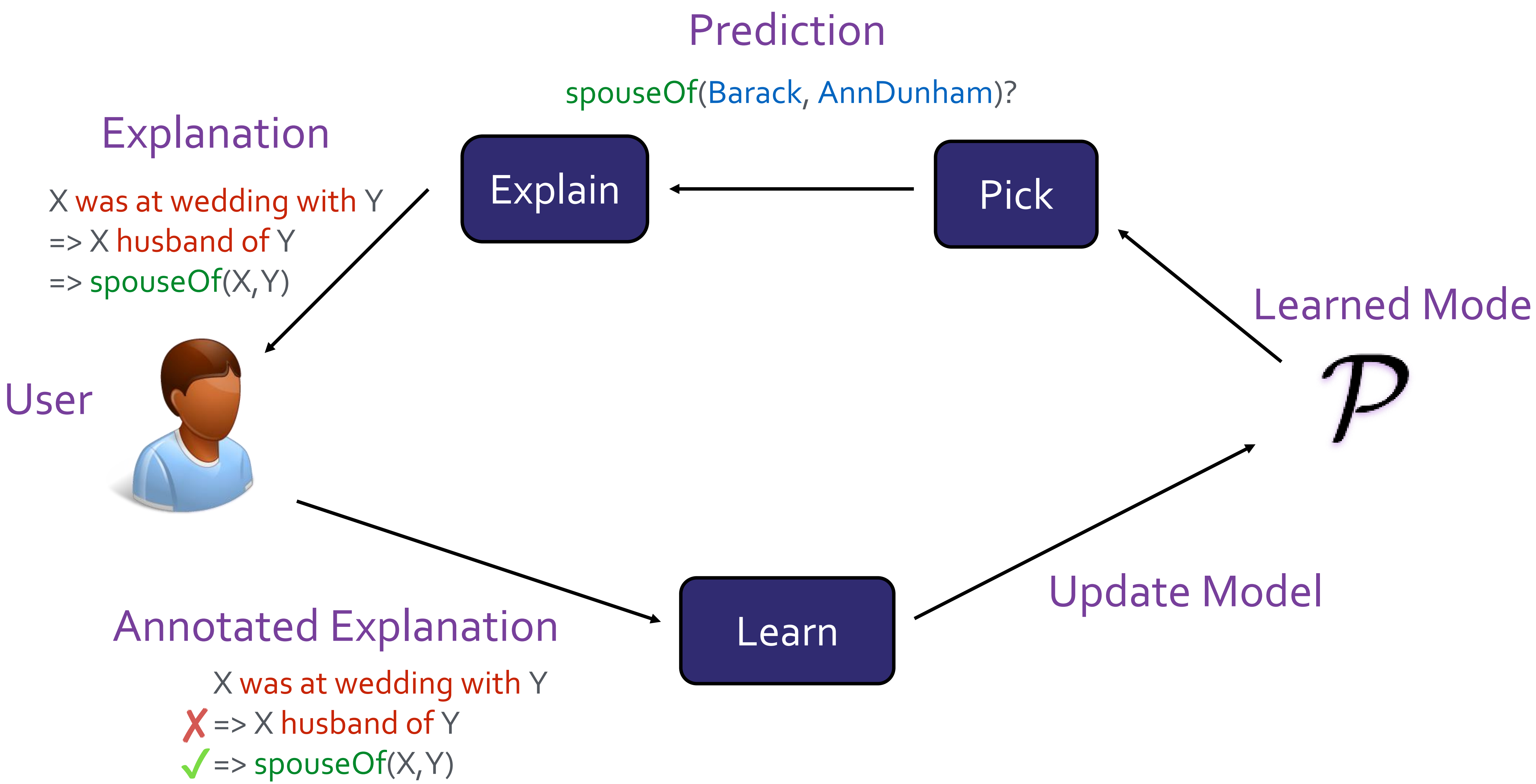
How can we make it easy for users to inject prior knowledge?

Current Supervision Approaches

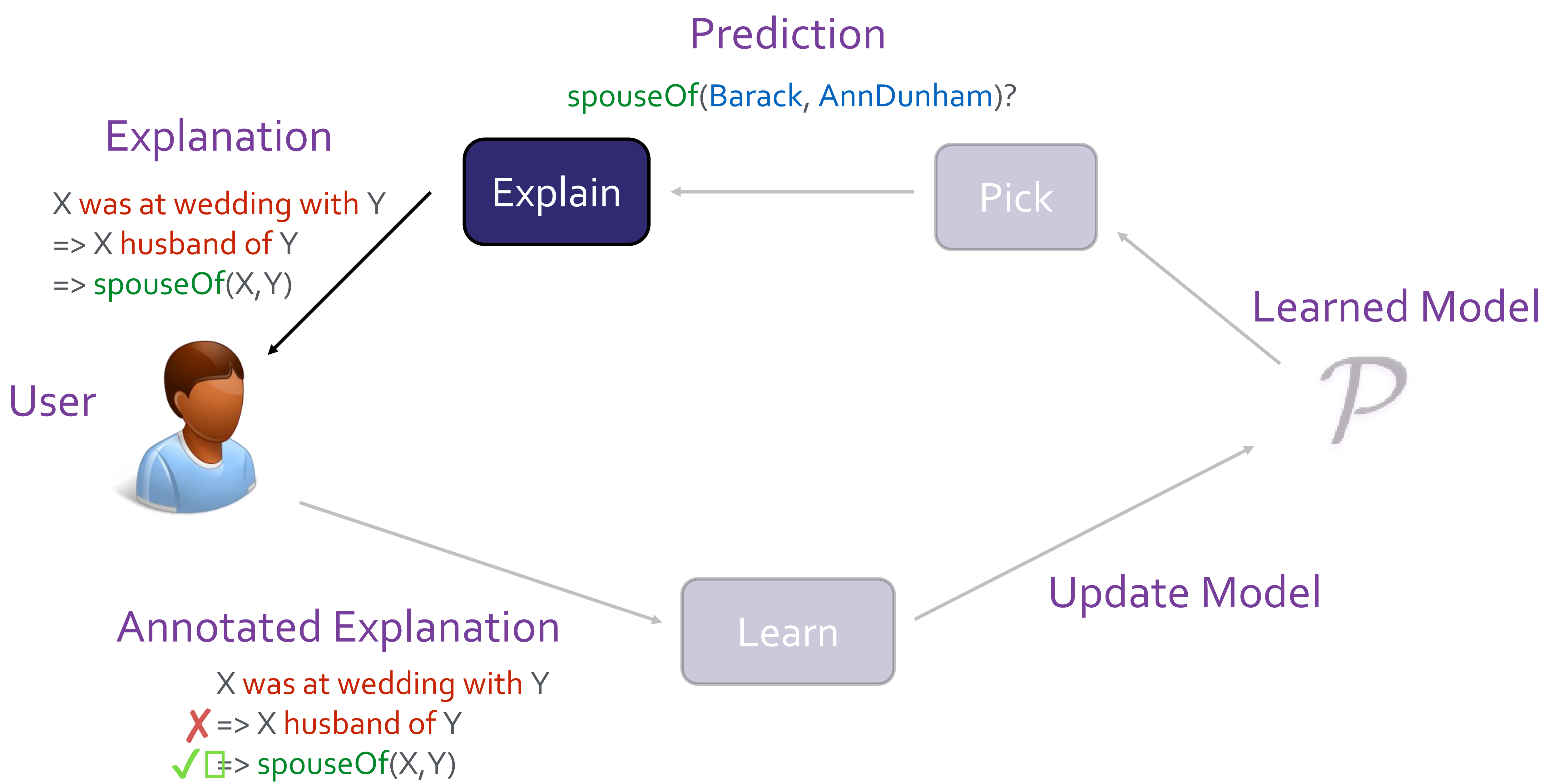


Problem 1: Each annotation takes time
Problem 2: Each annotation is a drop in the ocean

A More Intuitive Paradigm



Explaining Relation Extraction



Implication Chains as Explanations

spouseOf(Barack, AnnDunham)? ⇒

X was at wedding with Y
=> X husband of Y
=> spouseOf(X,Y)

employee(Marvin Minsky, MIT) ? ⇒

X cognitive scientist at Y
=> X professor at Y
=> employee(X,Y)

$$e(x, \theta) = \operatorname*{argmax}_{e \in \mathcal{E}(x)} Q(e, x, \theta) + \mathcal{I}(e)$$

Relation prediction

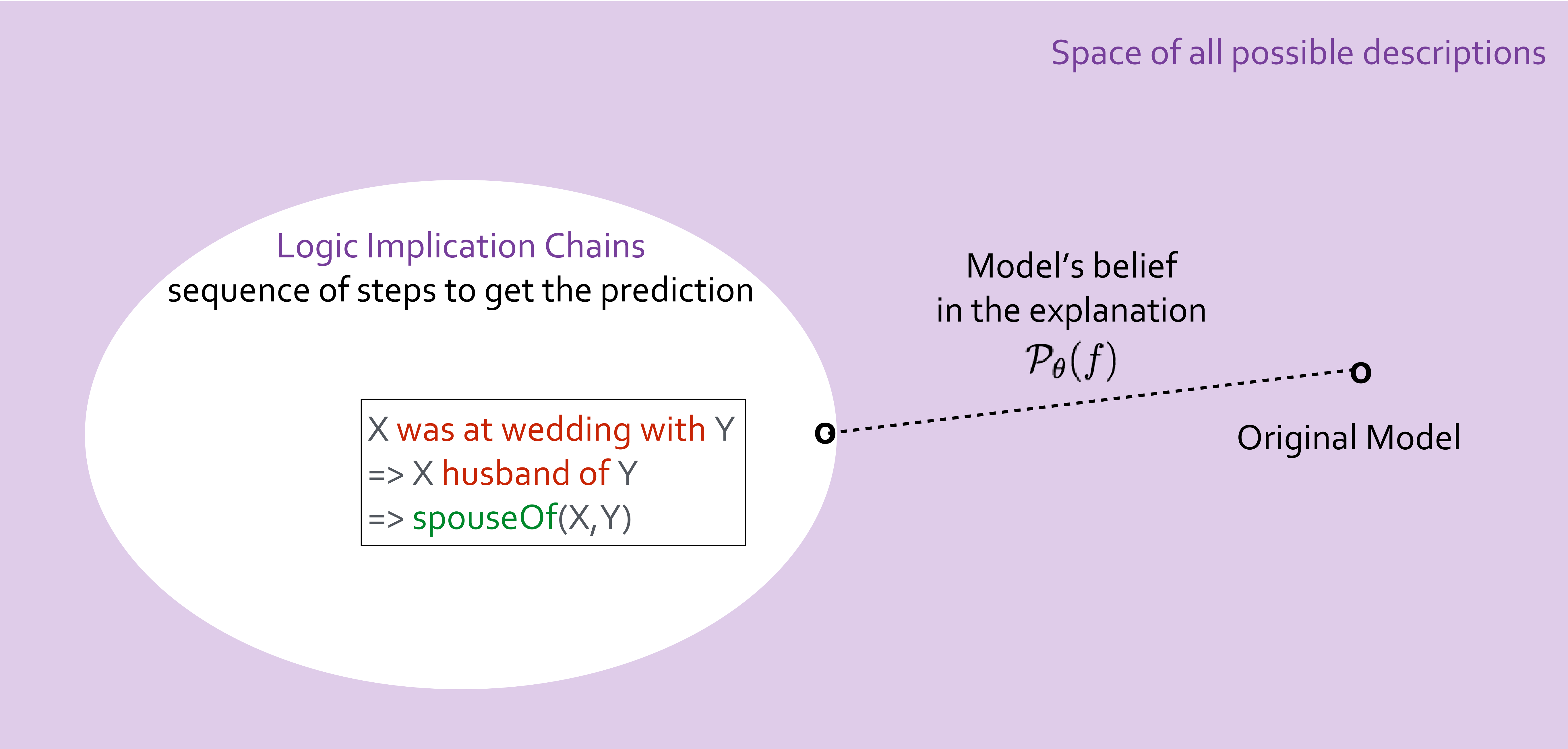
All implication chains
Possible chains that are valid for the prediction *x*

Faithfulness of the chain
How much does the model **believe** in the chain?

Interpretability
Keep chains short

Explaining Relation Extraction

Prediction to explain: spouseOf(Barack, AnnDunham)



Logic Representation of Relations

- Relations are binary predicates

$$\text{bornIn}(a, b) = \top \text{ or } \bot$$
$$\text{was-born-in}(a, b) = \top \text{ or } \bot$$
$$\text{where } a, b \in \{\text{“Bernie Sanders”}, \text{“Brooklyn”}, \text{“Michelle Obama”}, \ldots\}$$

- Facts are ground atoms:

$$\mathcal{F} = \begin{cases} \text{bornIn}(\text{Bernie Sanders}, \text{Brooklyn}) \\ \text{was-born-in}(\text{Bernie Sanders}, \text{Brooklyn}) \\ \text{spouse}(\text{Barack Obama}, \text{Michelle Obama}) \\ \vdots \end{cases}$$

- Relation Extraction models maximize the probability of ground atoms

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{f \in \mathcal{F}} \log \mathcal{P}_\theta(f)$$

Model's belief in a formula f

- For facts, we know this belief: $\mathcal{P}_\theta(f)$
- Otherwise, recurse…

Can be any model!

$$\mathcal{P}_\theta(f) = \begin{cases} R(a,b) & \text{then} \quad \boxed{\text{compute directly}} \\ \neg f' & \text{then} \quad 1 - \mathcal{P}_\theta(f') \\ f_1 \wedge f_2 & \text{then} \quad \mathcal{P}_\theta(f_1)\mathcal{P}_\theta(f_2) \\ \forall_e f(e) & \text{then} \quad \prod_e \mathcal{P}_\theta(f(e)) \end{cases}$$

$$\mathcal{P}_\theta\left(\forall_{a,b}\ \text{was-born-in}(a,b) \Rightarrow \text{bornIn}(a,b)\right) =$$

$$\prod_{a,b} 1 - \mathcal{P}_\theta(\text{was-born-in}(a,b))\,(1 - \mathcal{P}_\theta(\text{bornIn}(a,b)))$$

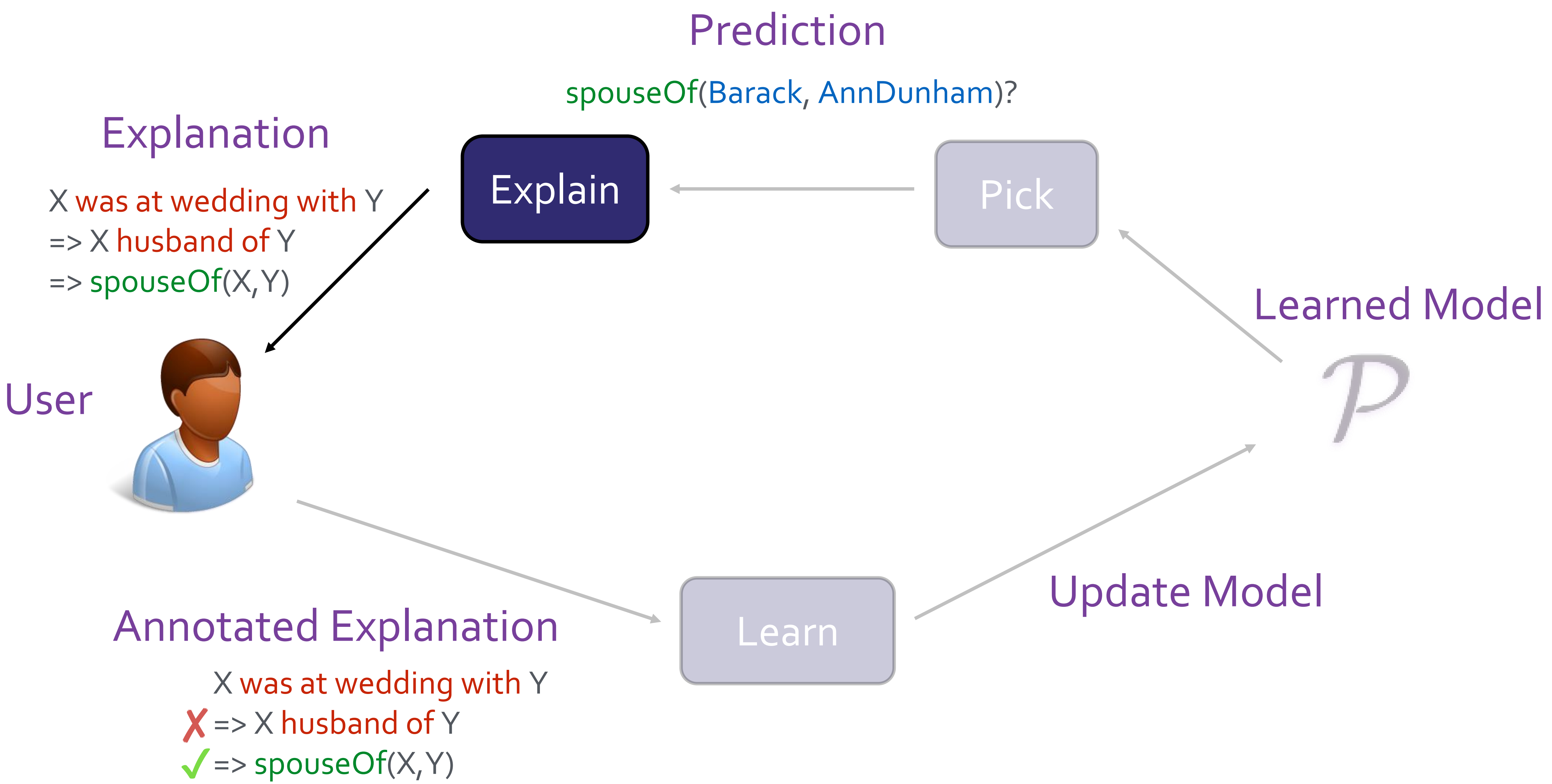
$\text{was-born-in}(a,b)$ $\quad$ $\neg\text{bornIn}(a,b)$

$\text{was-born-in}(a,b) \Rightarrow \text{bornIn}(a,b)$

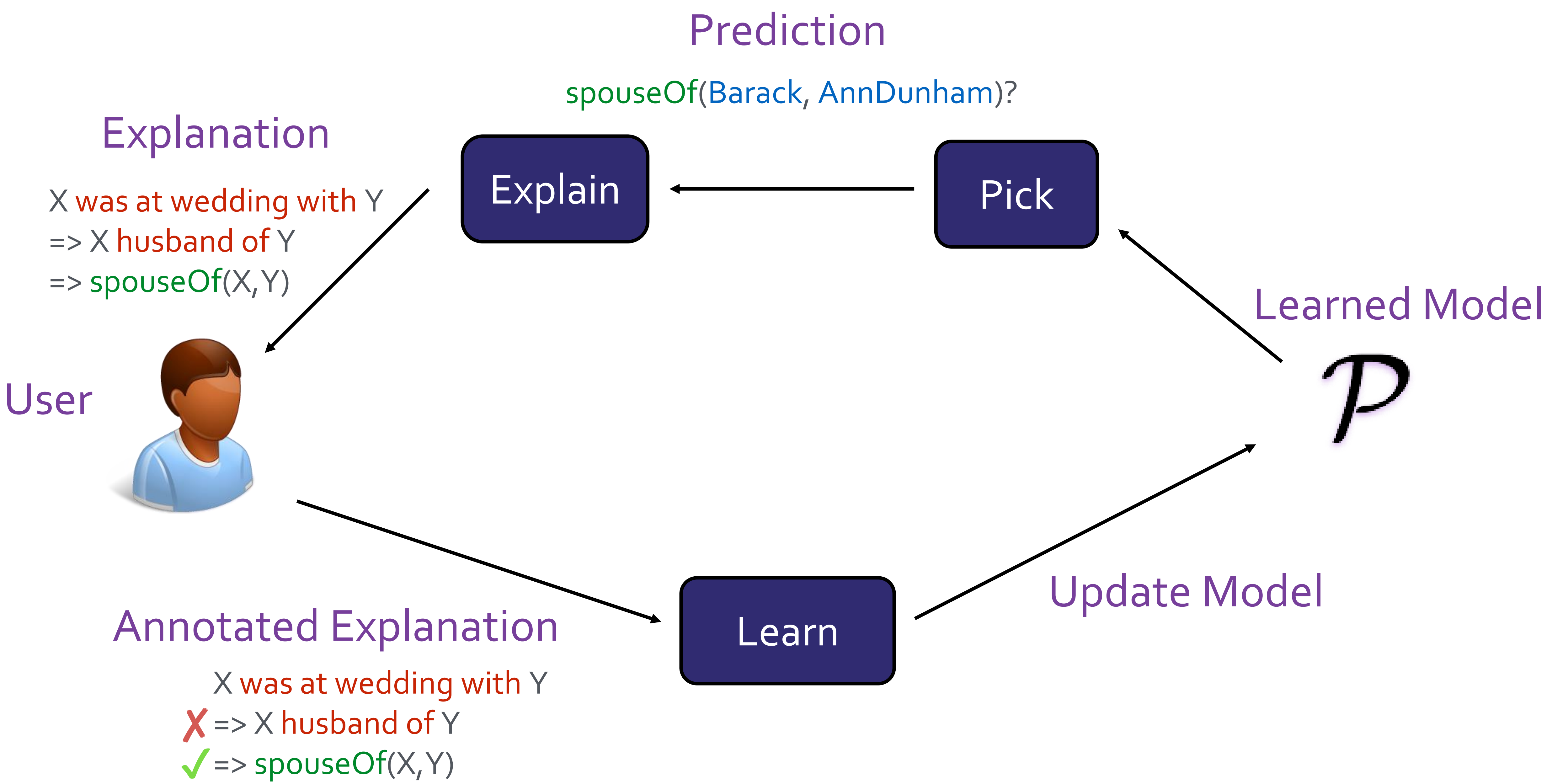
$\forall_{a,b}\ \text{was-born-in}(a,b) \Rightarrow \text{bornIn}(a,b)$

[Rocktaschel et al, NAACL 2015]

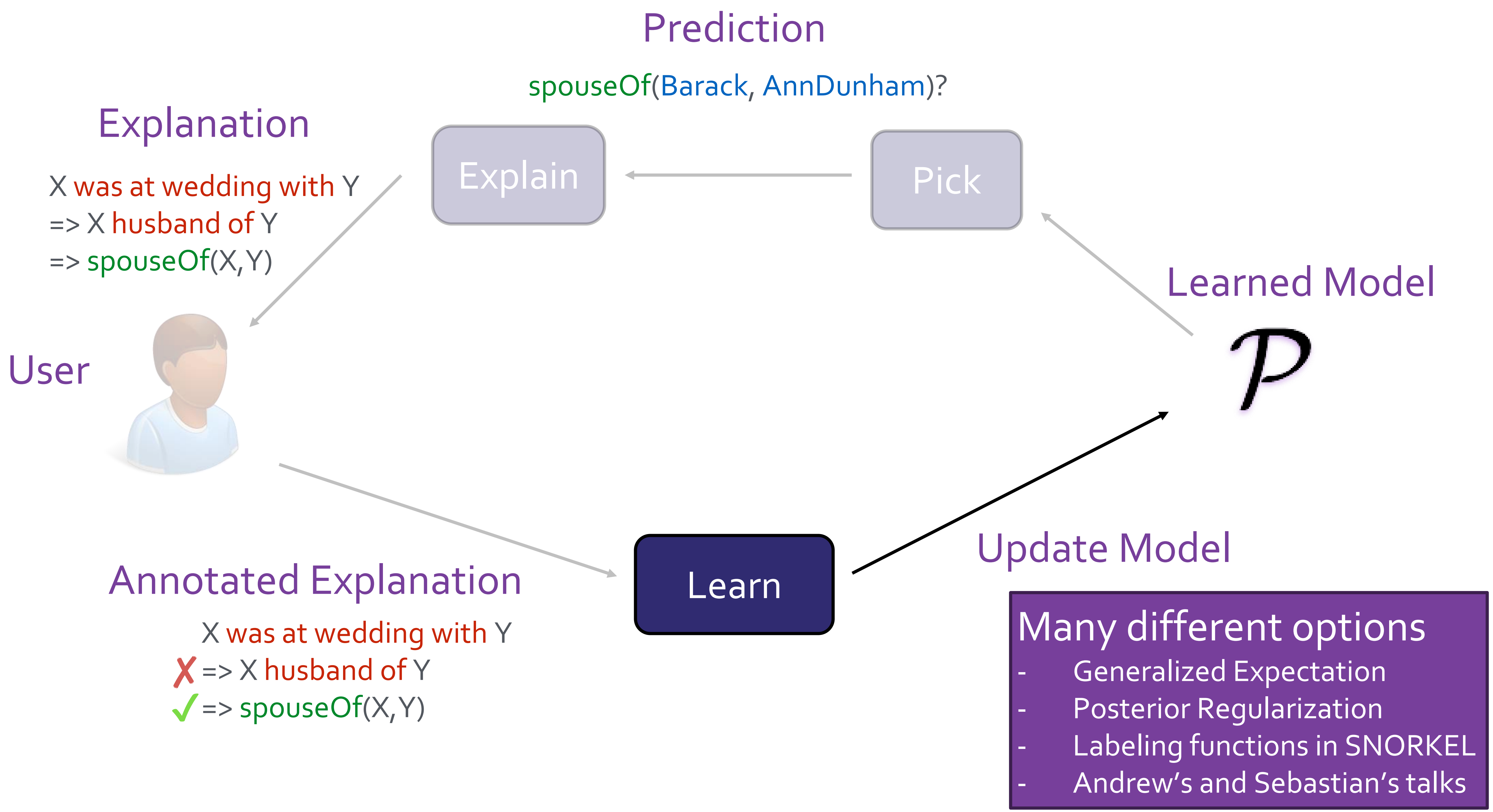
Explaining Relation Extraction



Explaining Relation Extraction



Learning from Logical Knowledge



[Rocktaschel et al, NAACL 2015]

Logical Statements as Supervision

- If you see “was a native of”, it means birthplace

X was native of Y => birthplace(X,Y)

- If a founder of the company is employed by the company, he’s the CEO

X is the founder of Y $\wedge$ employee(X,Y) => ceoOf(X,Y)

- Everyone is married to at most one person

spouse(X, Y) => $\forall$Y’ $\neg$spouse(X,Y’)

Improving the model

- Our model is maximizing probability of ground atoms

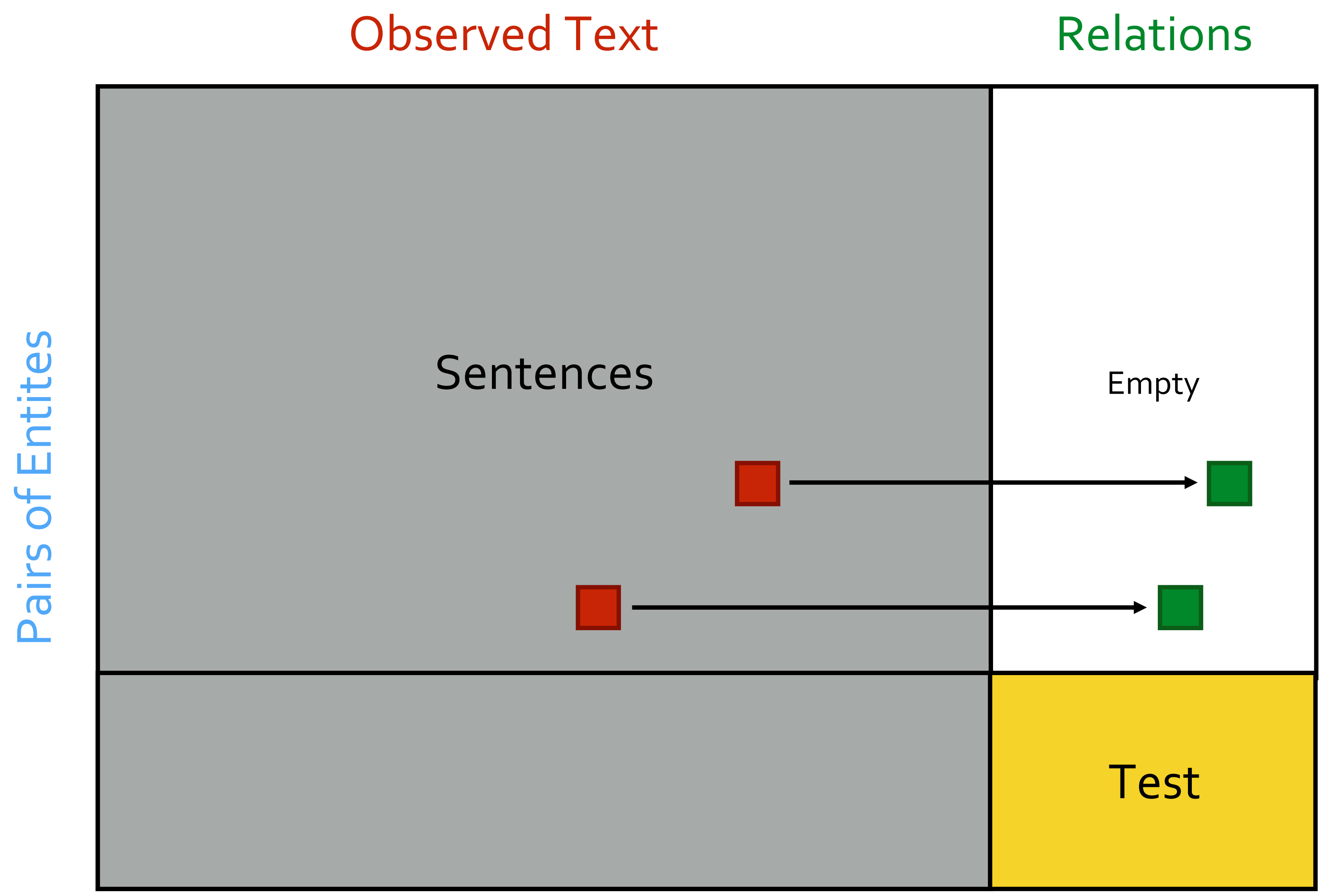
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{f \in \mathcal{F}} \log \mathcal{P}_\theta(f)$$

- But now we have a set of formulae, ground or otherwise

$$\mathcal{F} = \begin{cases} \text{bornIn}(\text{Bernie Sanders}, \text{Brooklyn}) \\ \text{was-born-in}(\text{Bernie Sanders}, \text{Brooklyn}) \\ \text{spouse}(\text{Barack Obama}, \text{Michelle Obama}) \\ \forall_{a,b} \ \text{was-born-in}(a, b) \Rightarrow \text{bornIn}(a, b) \\ \vdots \end{cases}$$

- Still maximizing the probability: $\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{f \in \mathcal{F}} \log \mathcal{P}_\theta(f)$
- Optimized using gradient descent
  - works for most models!

Zero-Shot Learning



~2 million docs

~400,000 Entity pairs

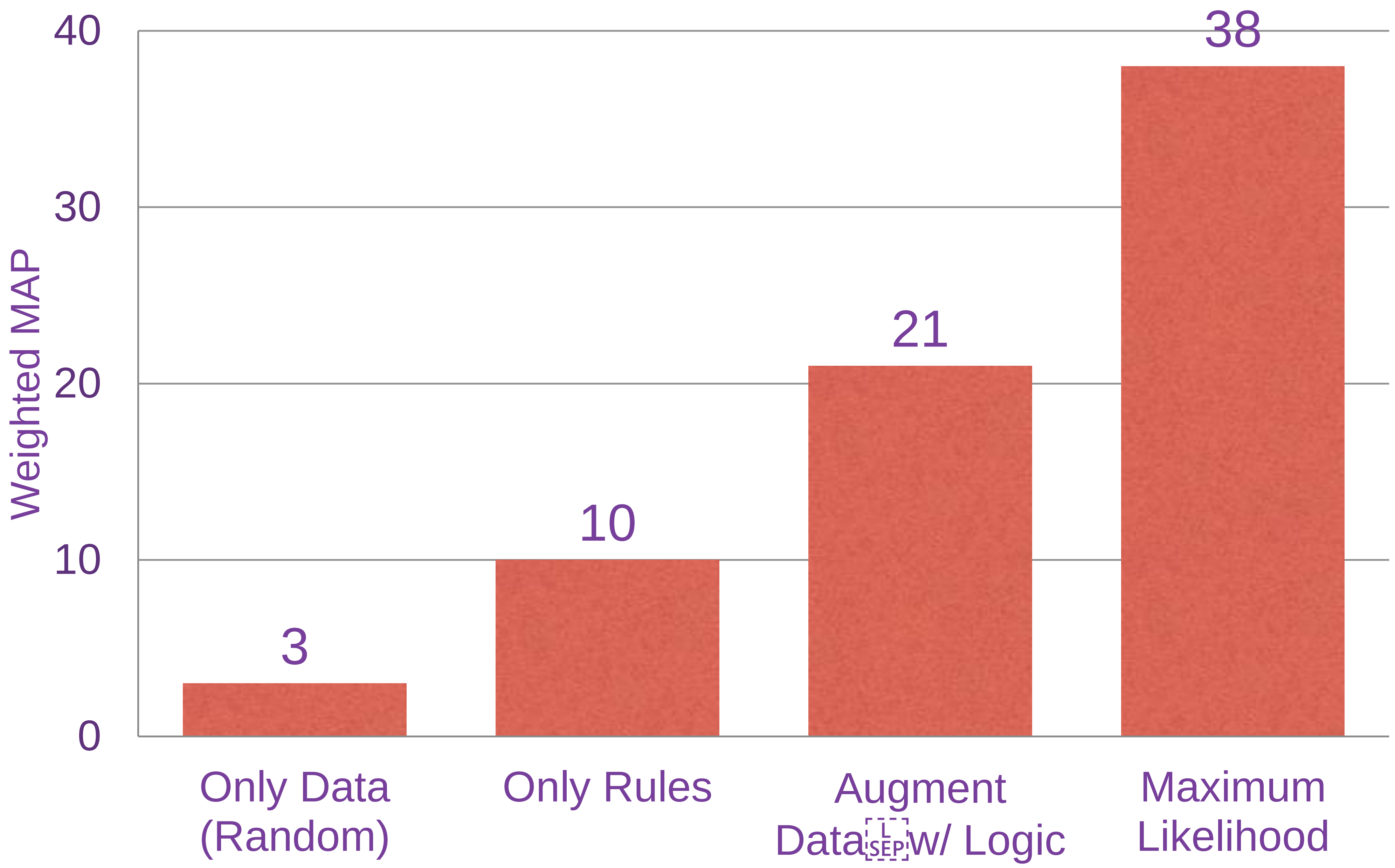
~4000 columns

~50 Relations of interest

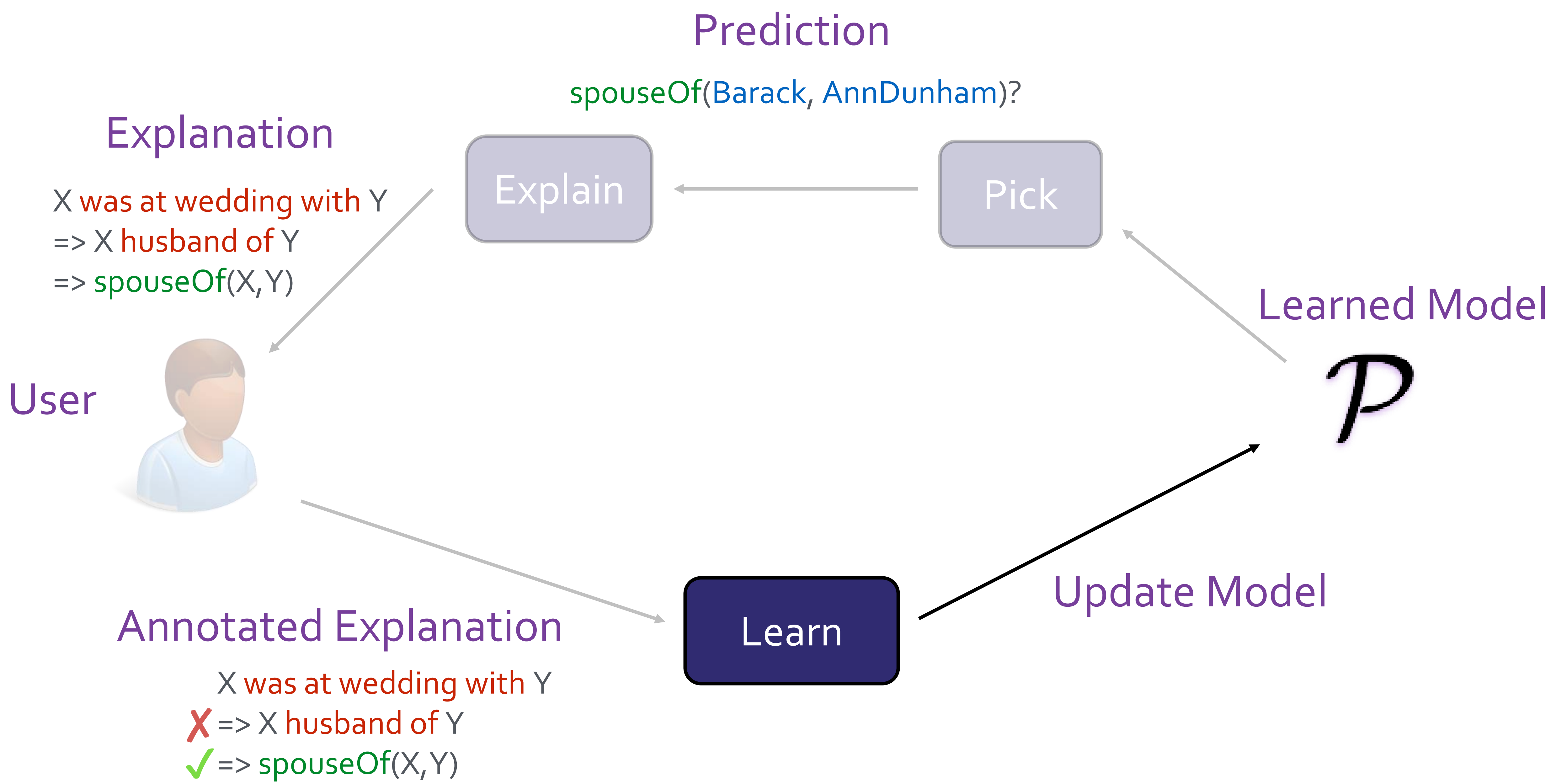
30 Logical Implications

We're evaluating whether formulae can be used instead of labeled data.

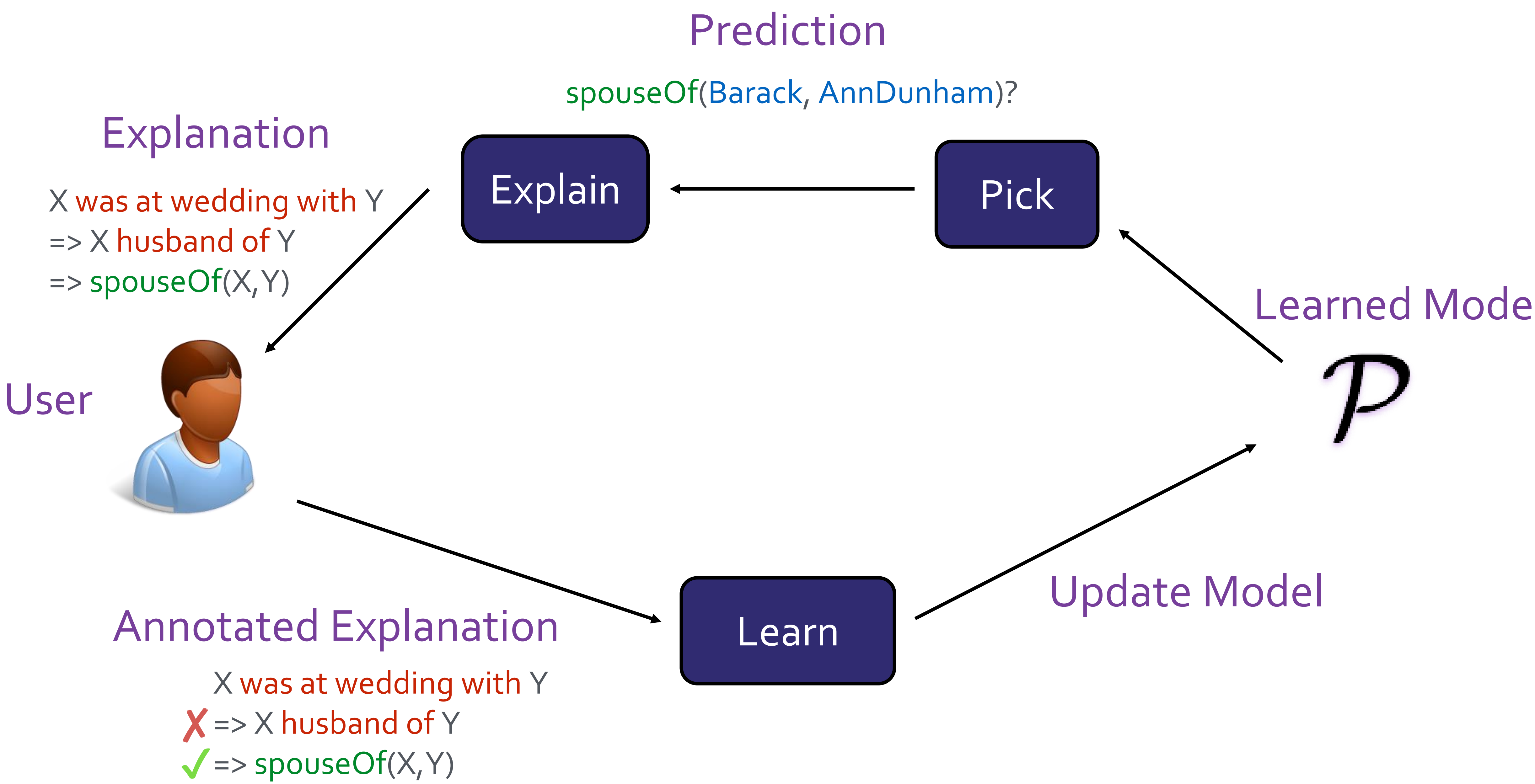
Zero-Shot Learning



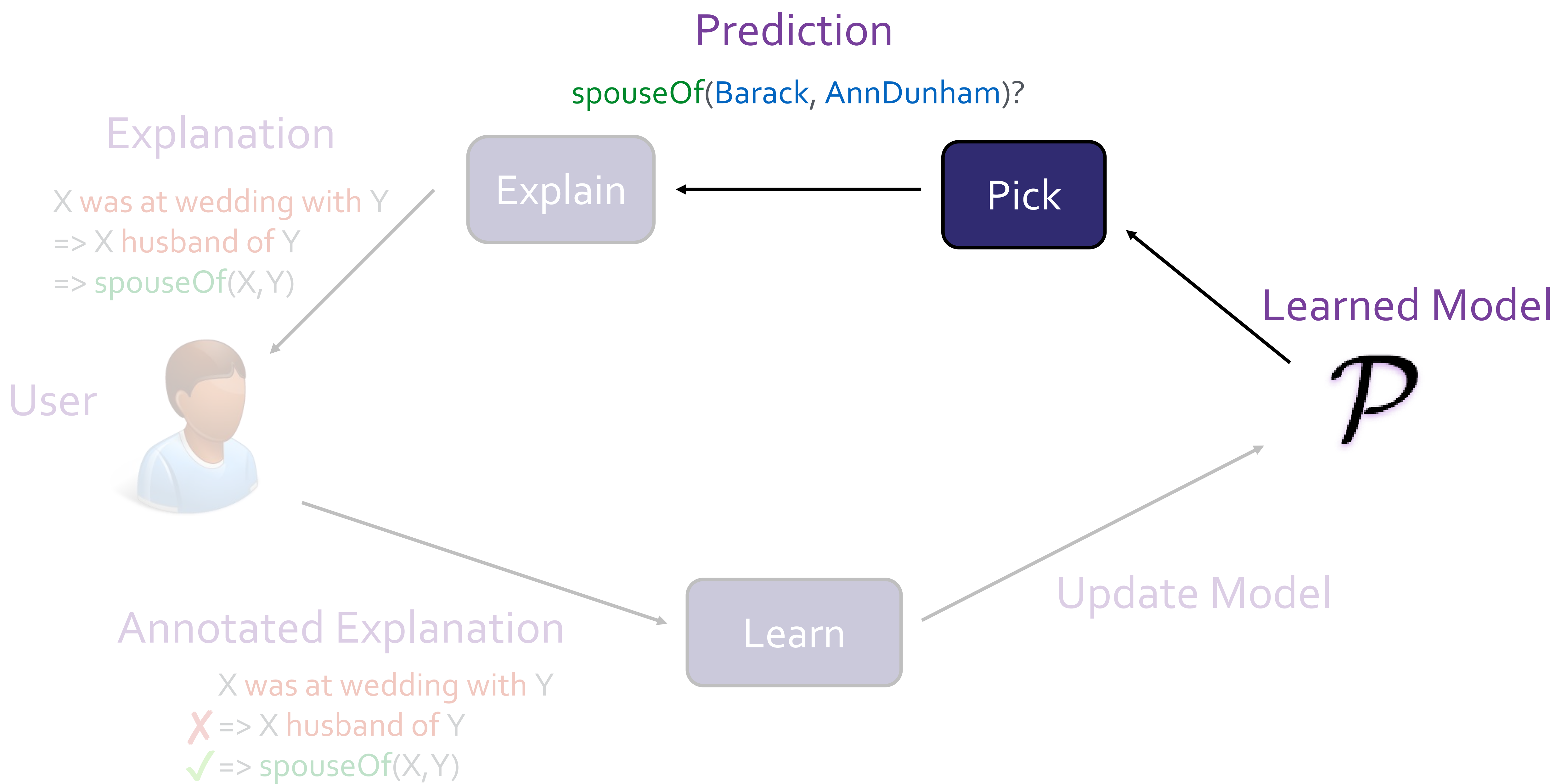
Learning from Logical Knowledge



Learning from Logical Knowledge



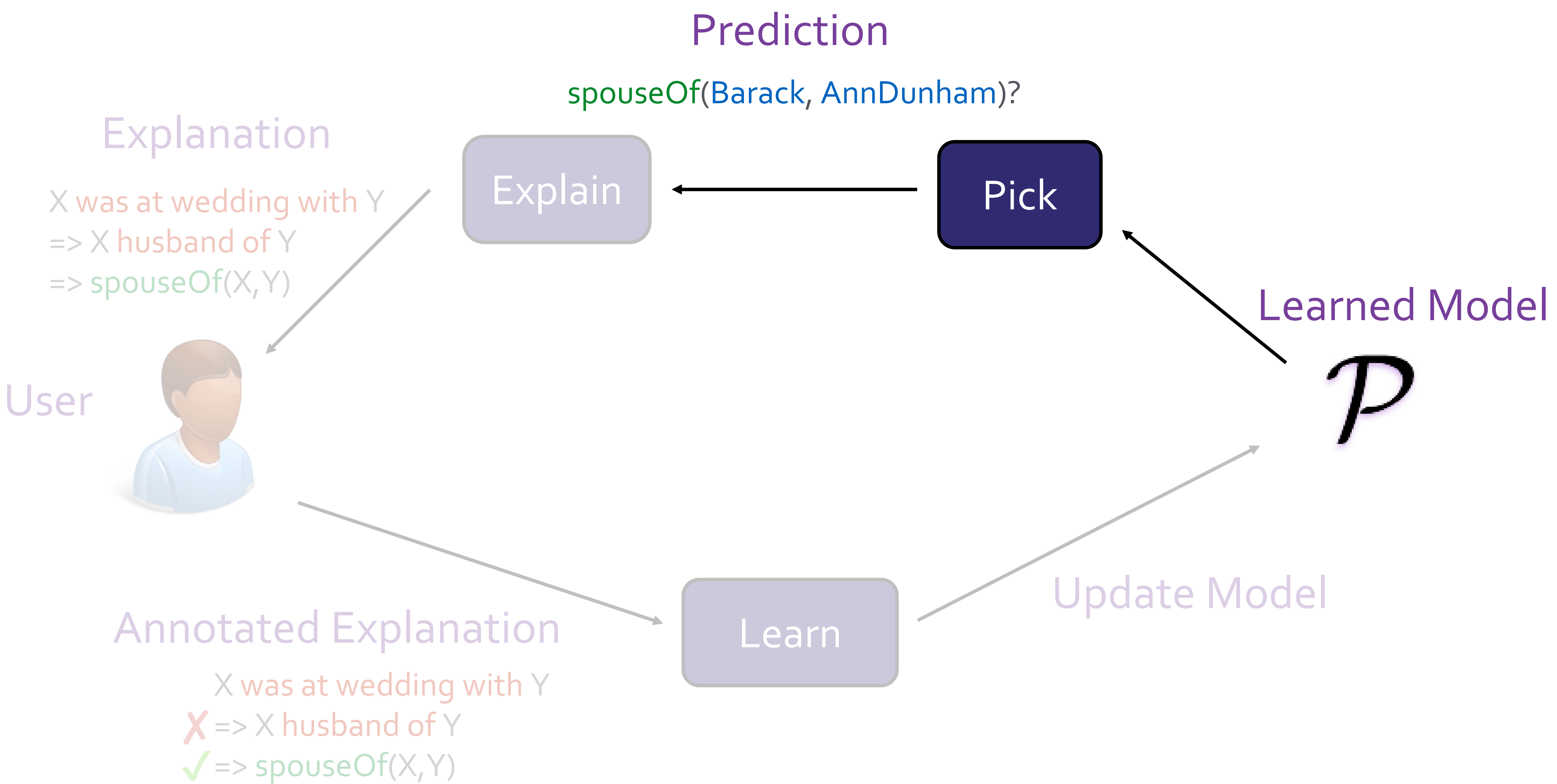
Picking What to Annotate



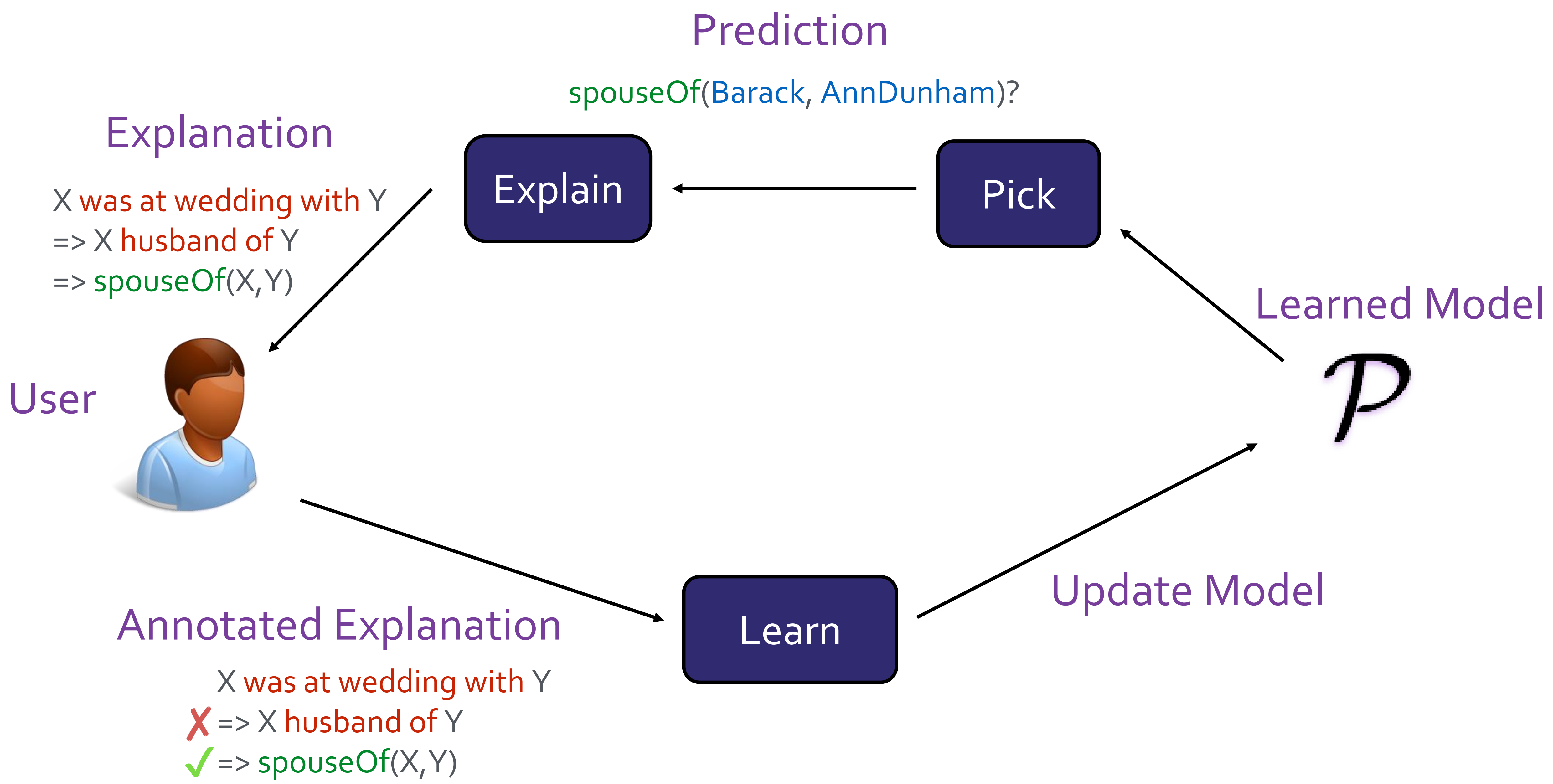
Picking the Constraint

- Active Learning: Annotation that effects the model the most
  - Most uncertain example, since both true and false lead to change
- Should we pick the most uncertain constraint?
  - X was born in Y => X died in Y ✗
  - If model didn't believe it anyway, nothing changes
- Should we pick the most certain constraint?
  - Likely to be correct!
  - X was born in Y => X birthPlace Y ✓
- Pick most confident constraint that is likely to be wrong
  - What we do: Most confident explanation of most uncertain example

Picking What to Annotate



Closing the Loop



Crowd Sourcing Annotations

- Generate textual phrases from dependency paths
- Annotate individual implications, 5 labels each ($0.05 per label)

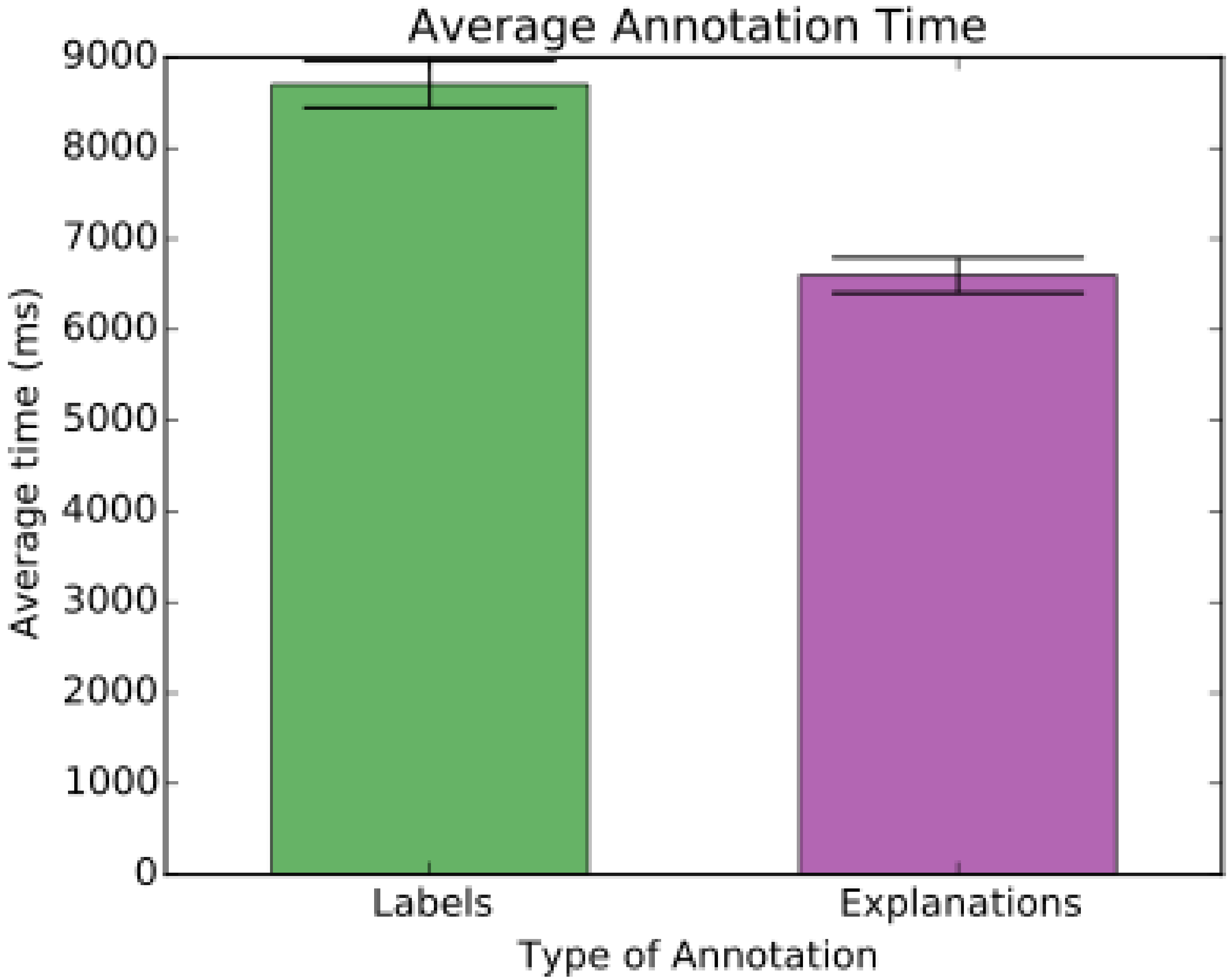
Think about the facts that the following phrases suggest:

1. "X, son of Y", and
2. "X is a child of Y".

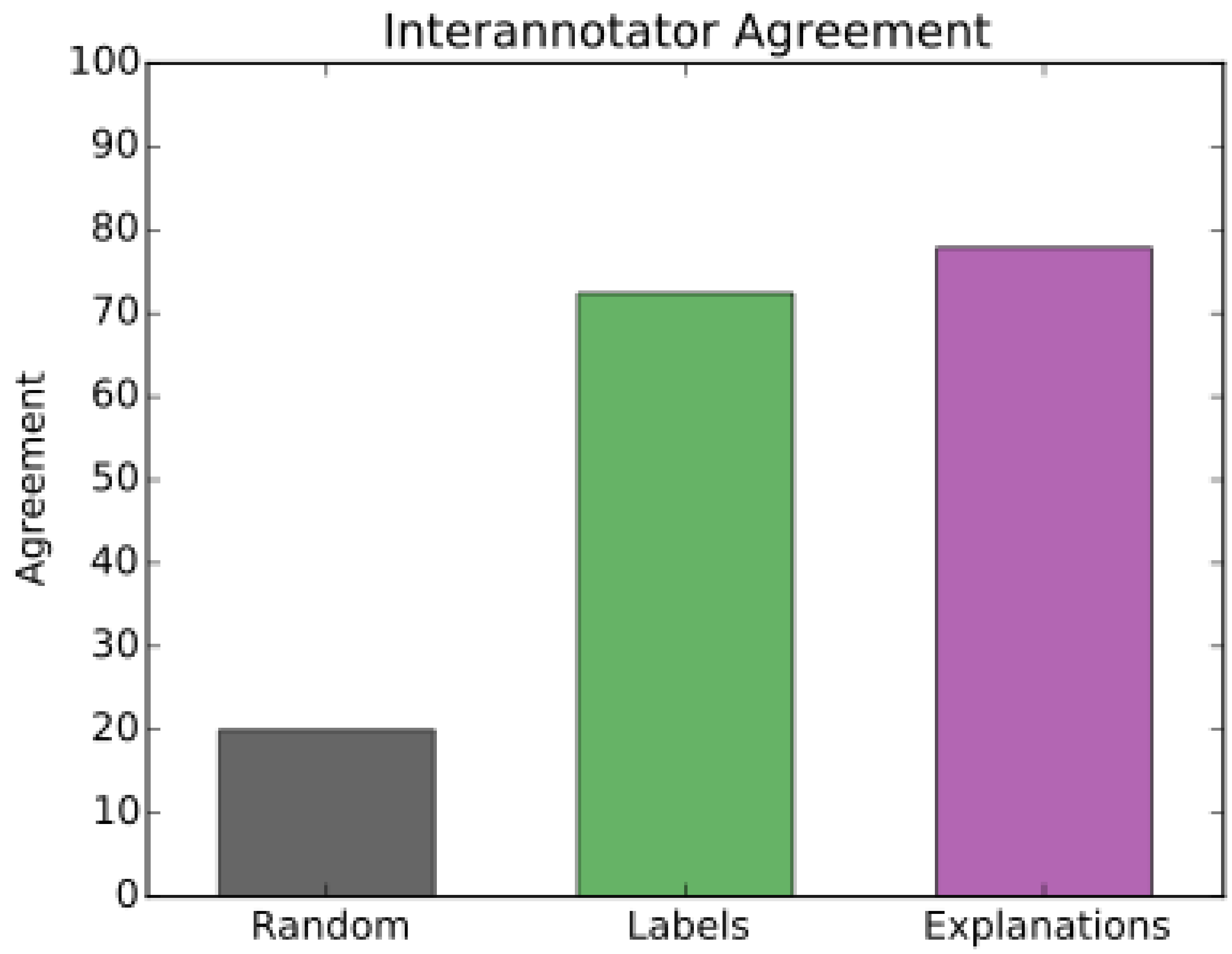
Do you think something in the first statement might imply the second?

- ○ YES, the first phrase strongly conveys the second.
- ○ Yes it does, but only weakly.
- ○ I can't tell, not sure.
- ○ No, the implication is quite weak.
- ○ Not at all, there is very little connection between the two.

Effort of Annotations



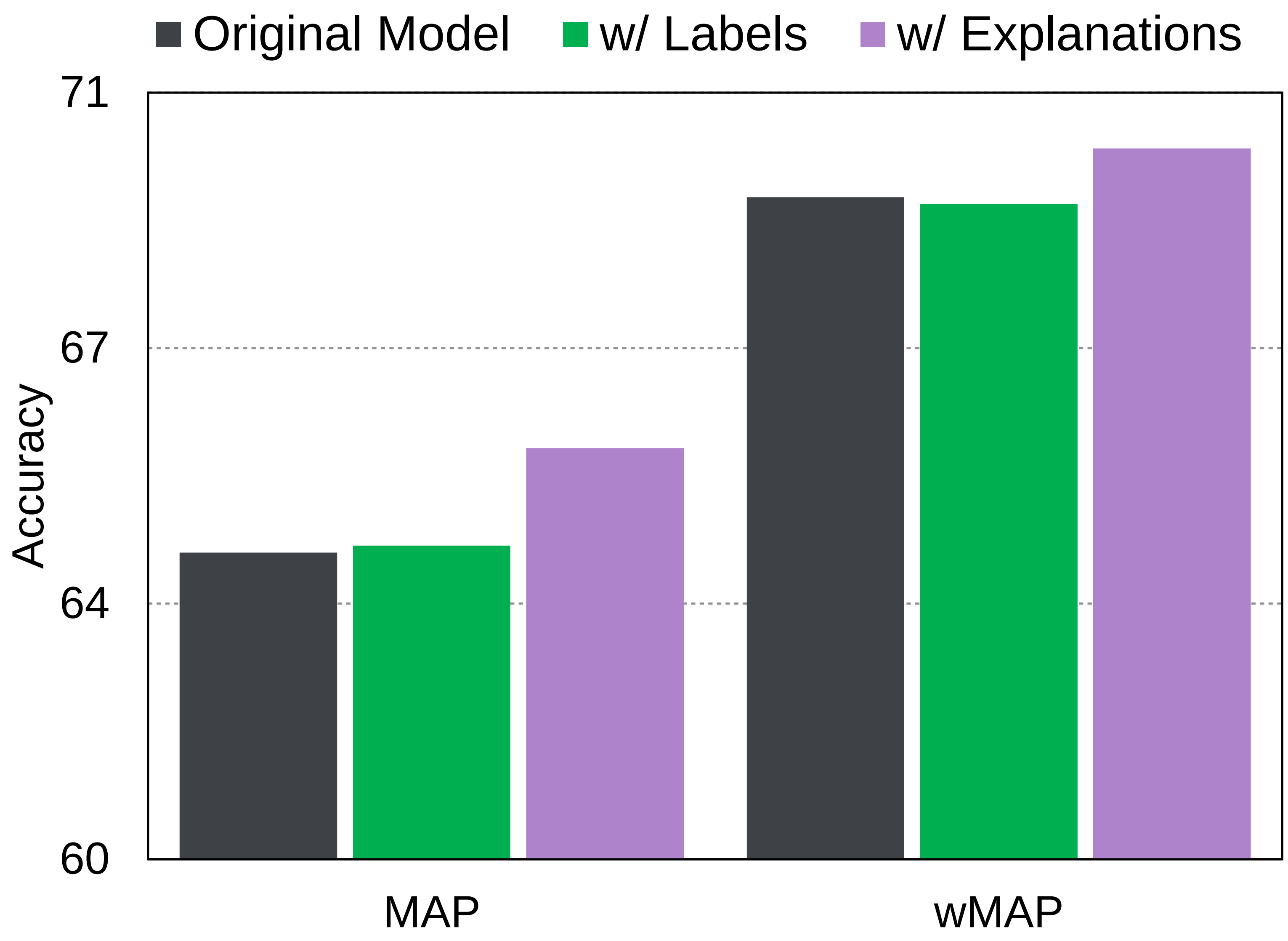
Quality of Annotations



Closing the Loop on the Trained Model

Single round of annotating explanations, and incorporating them

- 150 total implications, 5 annotators each



Interactive Relation Extraction

Real-world, large-scale application of ML and NLP

But suffers from the need for a large amount of labeled data

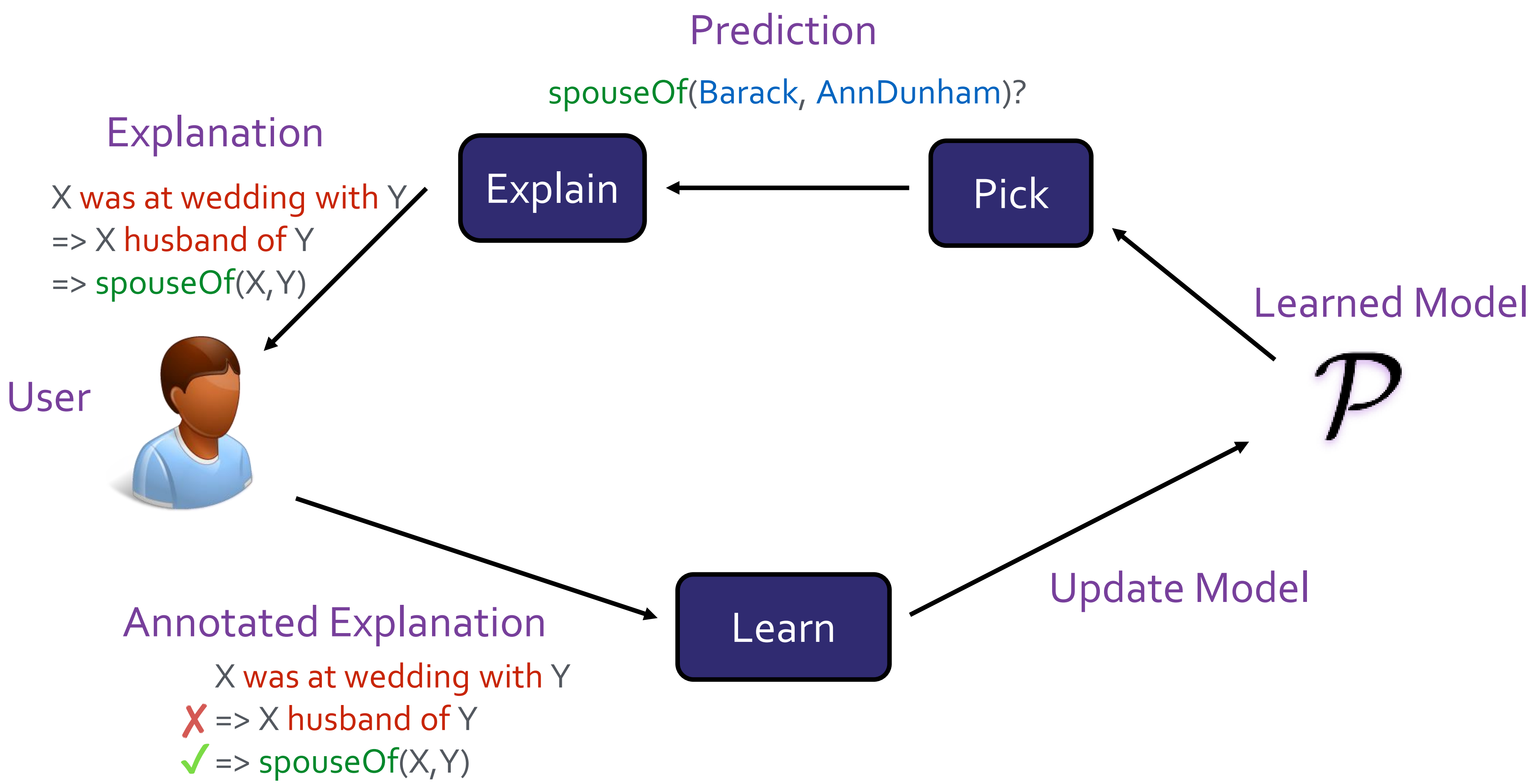
Actively Annotating Model Explanations

Labeled Data: *is expensive, noisy, and time-consuming to obtain*
Explanations: *are simple chains of logical implications*
Feedback on Explanations: *much easier for users to annotate*

Open Questions

- **Evaluation:**
  - How much does labeling explanations help over instances?
  - How much does it help to be “active”?
- **Pick:**
  - What is the optimal explanation to show user?
  - What is a good approximation of that?
- **Explain:**
  - Can the explanations always be black-box?
  - Can we surface latent spaces for annotation directly?
- **Learn:**
  - How do we balance higher-level supervision with observed data?

www.sameersingh.org
@sameer_
sameer@uci.edu



Thank you!

In collaboration with Carlos Guestrin, Sebastian Riedel, Luke Zettlemoyer, Tim Rocktaschel