Fanoos: Multi-Resolution, Multi-Strength, Interactive Explanations for Learned Systems

David Bayani , Stefan Mitsch
dcbayani@alumni.cmu.edu
smitsch@cs.cmu.edu

This material is based upon work supported by the United States Air Force and DARPA under Contract No. FA8750-18-C-0092. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.

IJCAI-XAI
Jan. 7, 2021

Individual ML systems are part of a larger whole in tackling problems

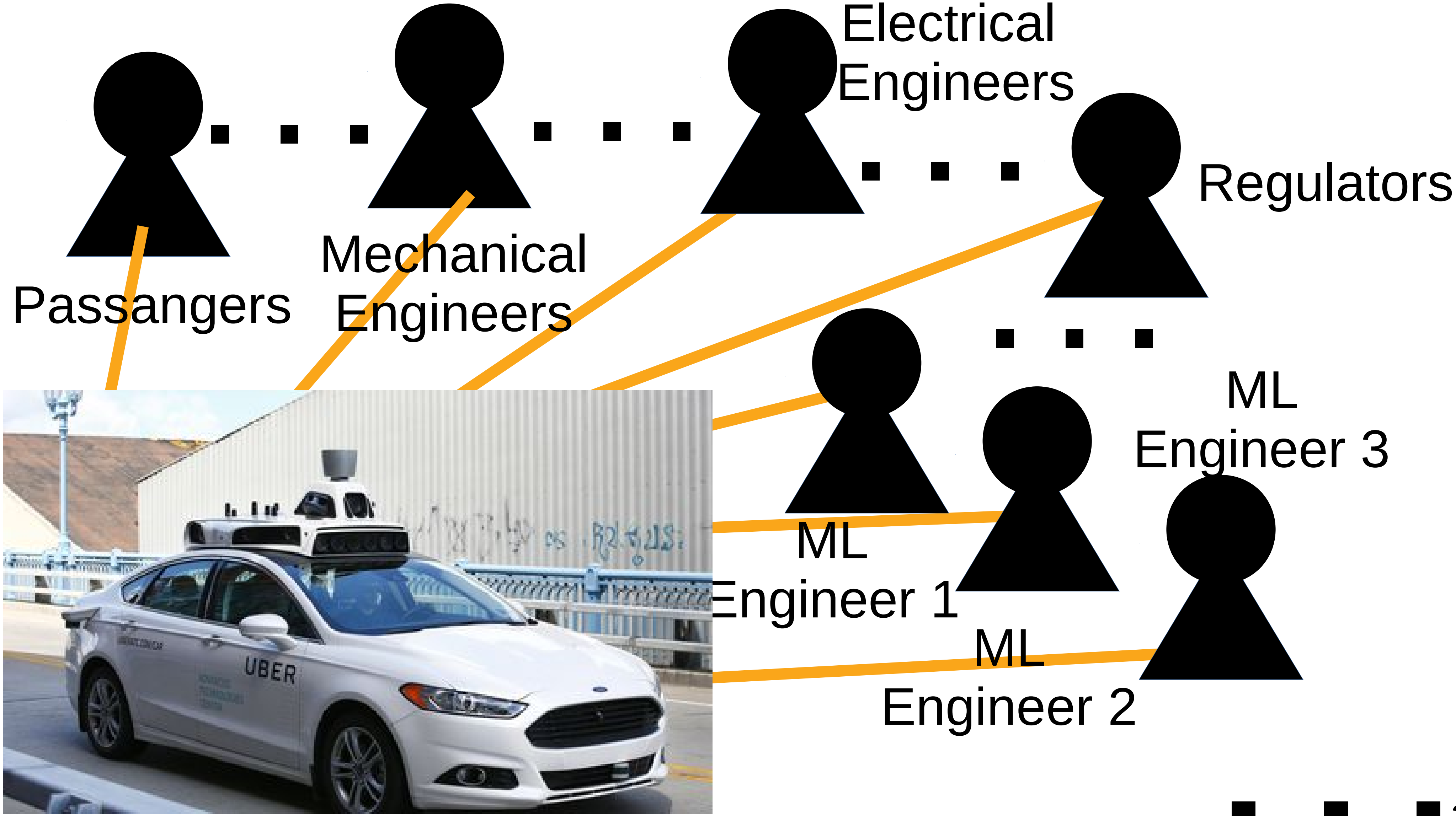


Image credit: https://www.chicagotribune.com/business/blue-sky/ct-uber-self-driving-cars-pittsburgh-20160906-story.html

Some Observations

Actors have many different:

- Interests
- Needs – depends on actor ***and task at hand***
  - Stakes vary. Safety : high stakes, efficiency tweaks: lower stakes
- Backgrounds / Expertise

Unfilled Desiderata for XAI

- *Interactive* (so can explore as needed)

Unfilled Desiderata for XAI

- *Interactive* (so can explore as needed)
- *Can provide multiple abstraction levels of information* (so can suite multiple audiences and needs)

Unfilled Desiderata for XAI

- *Interactive* (so can explore as needed)
- *Can provide multiple abstraction levels of information* (so can suite multiple audiences and needs)
- Can provide strong guarentees about info. provided (so *explanations necessarily reflect system behaviour*)
  - Should be as pedantic about details as user needs (*sometimes want / don’t want corner-case info.*)

Our Solution:
Fanoos

- Fanoos (فــانوس) - “Lantern” in Farsi

“Shining a Light on Black-Box AI”

Masrawy.com

Plan For Next Few Slides

- Overview of What User Sees
- Description of the Mechanics
- Breifly overview experiments

Fanoos Overview:

Initial Setup

Fanoos

Fanoos Overview:

Initial Setup

L, the Learned Sytem to Explain (E.g., Neural Net)

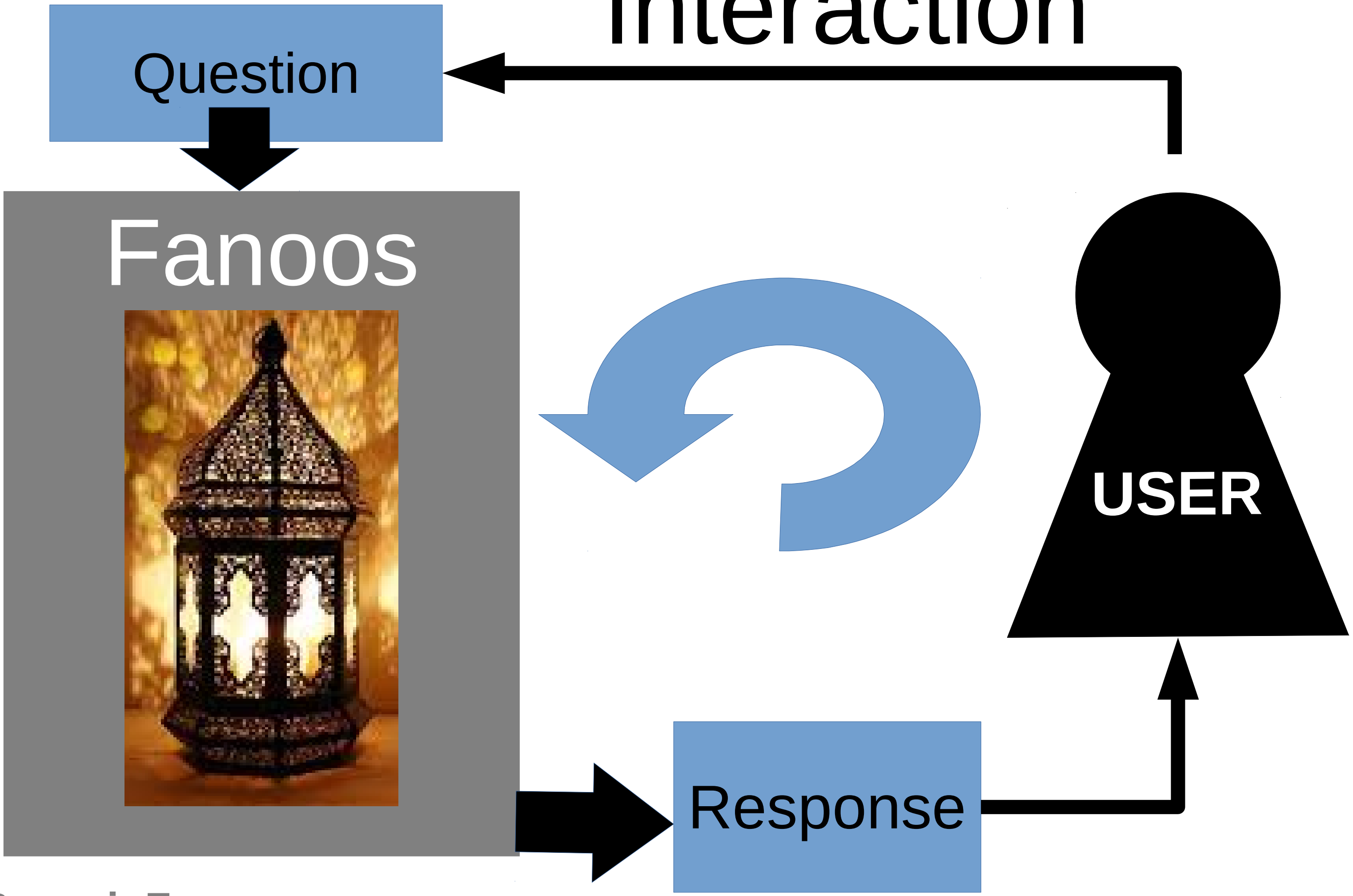
Domain Info. (can have data-assisted construction)

Initial Configuration Inputs

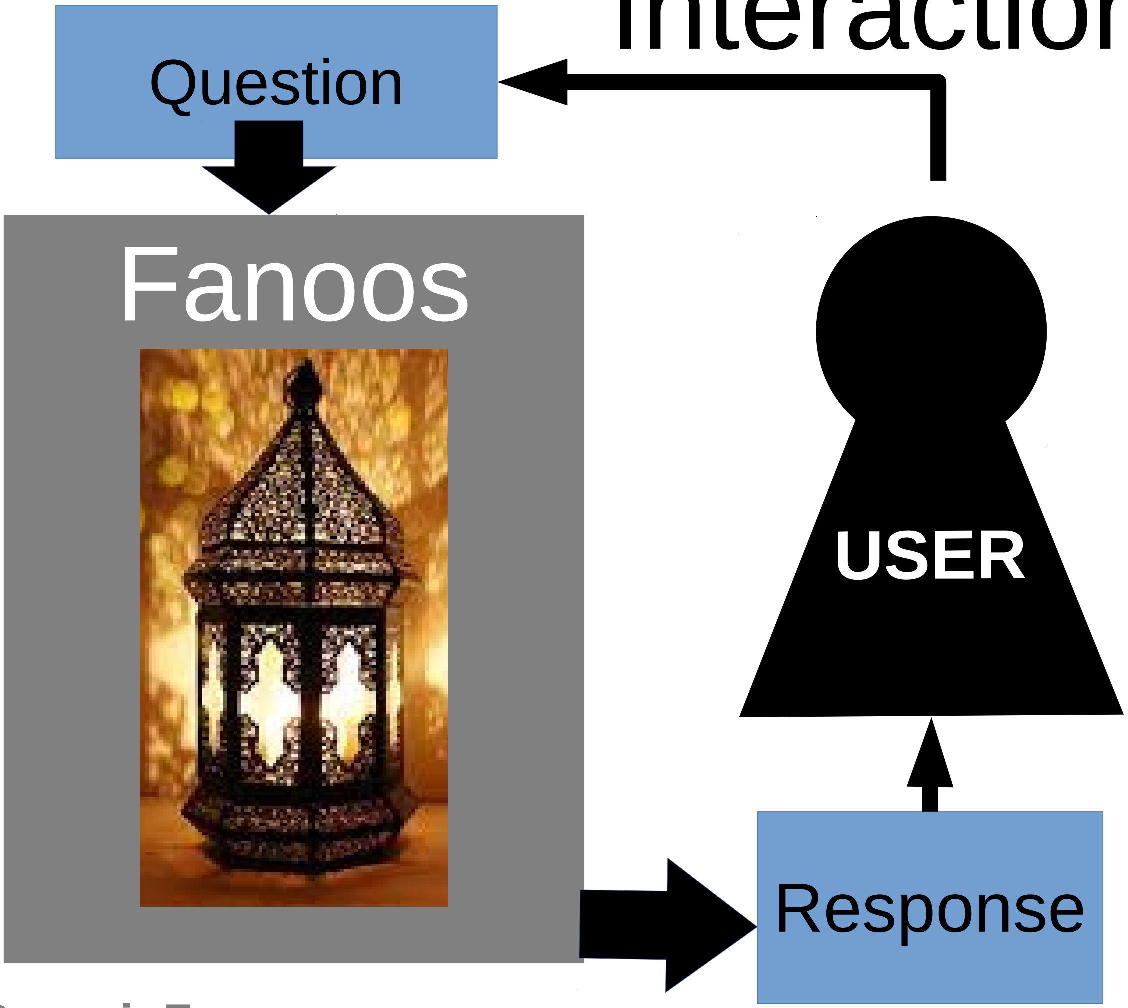
Fanoos

Fanoos Overview:

Interaction



Fanoos Overview:
Interaction



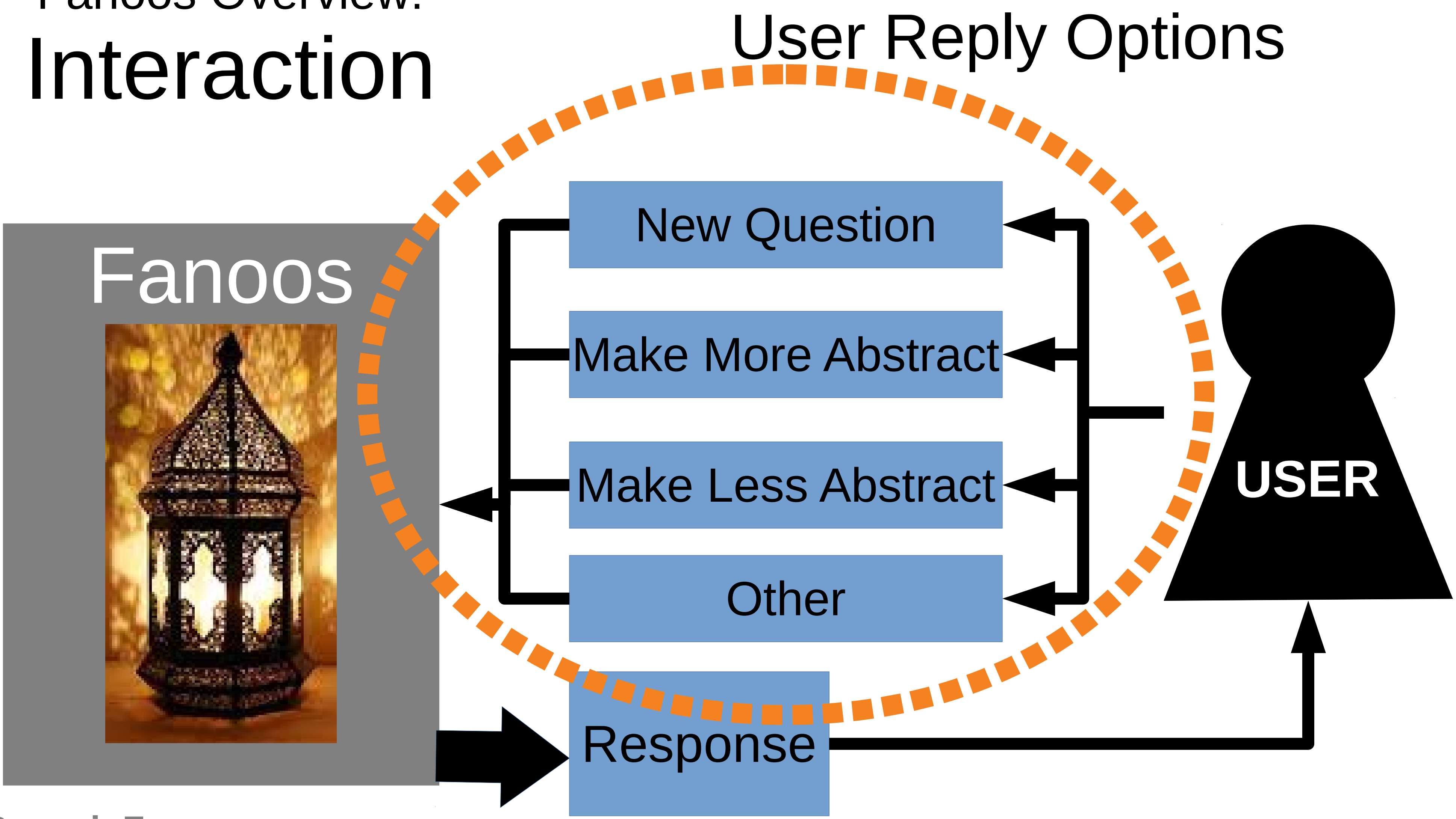
Question Types:

- When does ***L*** do X?
- What does ***L*** do when Y?
- In what circumstances is ***L*** doing X during Y?

Can be formally sound or probabilistically guaranteed

Fanoos Overview:

Interaction



Fanoos Overview:

Interaction Example

Fanoos

Example from robotics

Initial Question

```
(Fanoos) what_are_the_circumstances_in_which and(
    pole1angle_rateofchange_low__magnitude ,
    outputtorque_high__magnitude )?
```

Initial Response

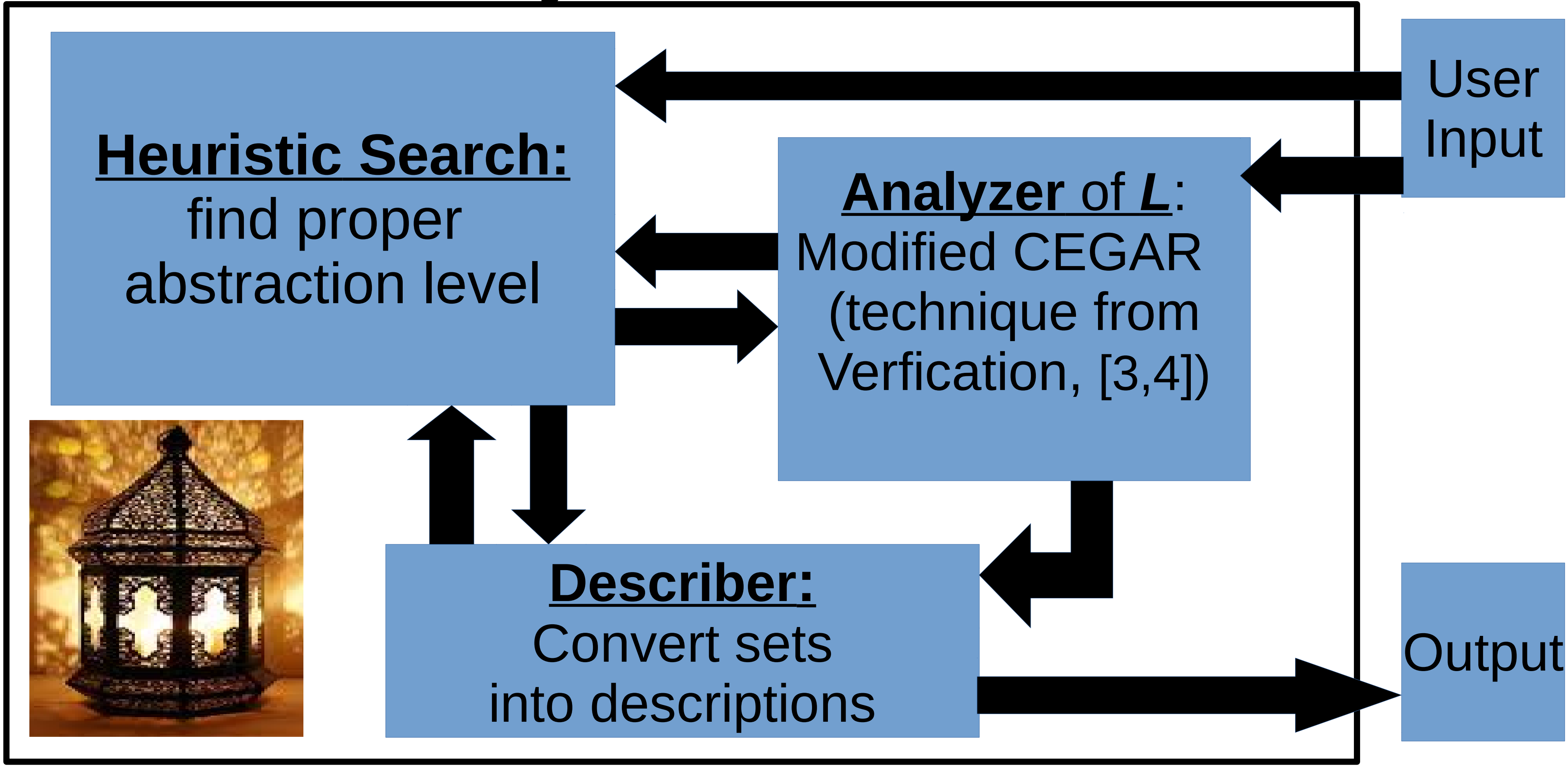
```
(0.10147897, 0.17831770, pole1angle_on_the_left ,
    pole2angle_on_the_left ,
    pole2angle_rateofchange_low__magnitude)
(0.09885232, 0.16335186, pole1angle_on_the_left ,
    pole2angle_on_the_left ,
    pole2angle_turning_counterclockwise)
(0.07900125, 0.14467123, pole1angle_on_the_right ,
    pole2angle_on_the_right ,
    pole2angle_turning_clockwise)
(0.06693577, 0.12822191, pole1angle_down ,
    pole2angle_to_right ,
    statevalueestimate_very_low )q
```

User Request ***More*** abstract

New Response

```
(0.44378316, 0.48588134, pole2 not near target
    position)
(0.33605014, 0.36551887,
    pole2angle_rateofchange_high__magnitude)
(0.22016670, 0.23739381, pole2angle_to_right ,
    statevalueestimate_very_low)
```

Briefly, Inside Fanoos

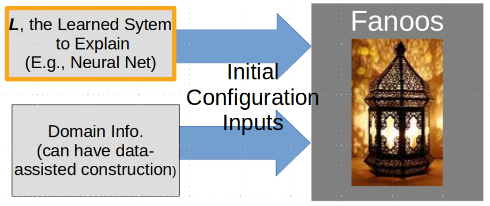


Fanoos:

Mechanics

Fanoos

Domain Knowledge that User Provides



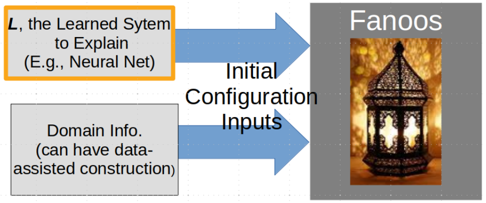
- The learned system, *L*
- Universe Bounding-Box for Input Space
  - Ex: for a constant-v Dubin car:

$$(x, y, \theta) \in [-1, 1] \times [50.3, 100.0] \times [0, 2\pi]$$

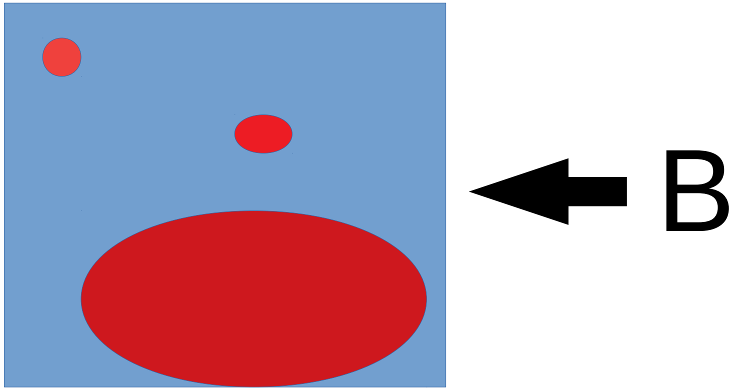
- Predicates: connecting sets to something user grasps. Ex:
  - "left arm higher than right arm" : y_arm1 > y_arm2
  - “attempting spiral roll” :

$$\exists cx, cy \in B. |(x - cx)^2 + (y - cy)^2 - r^2| \leq \epsilon_1 \wedge |2(x - cx)dx - 2(y - cy)dy| \leq \epsilon_2 \wedge \ldots$$

Domain Knowledge that User Provides



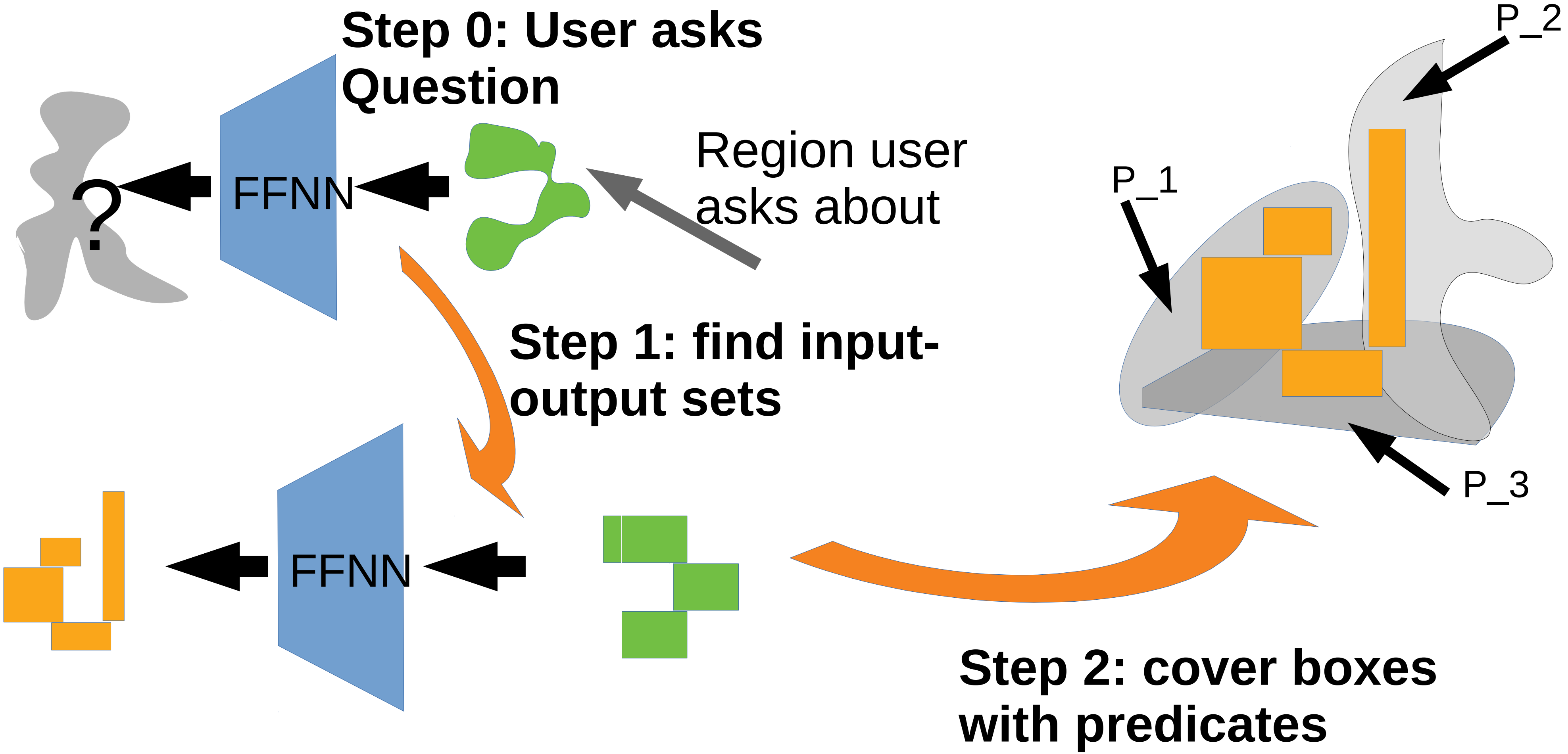
- Note: Using a SAT-solver, we can say whether predicate holds:
  - Everywhere on a set
  - At least one place in a set



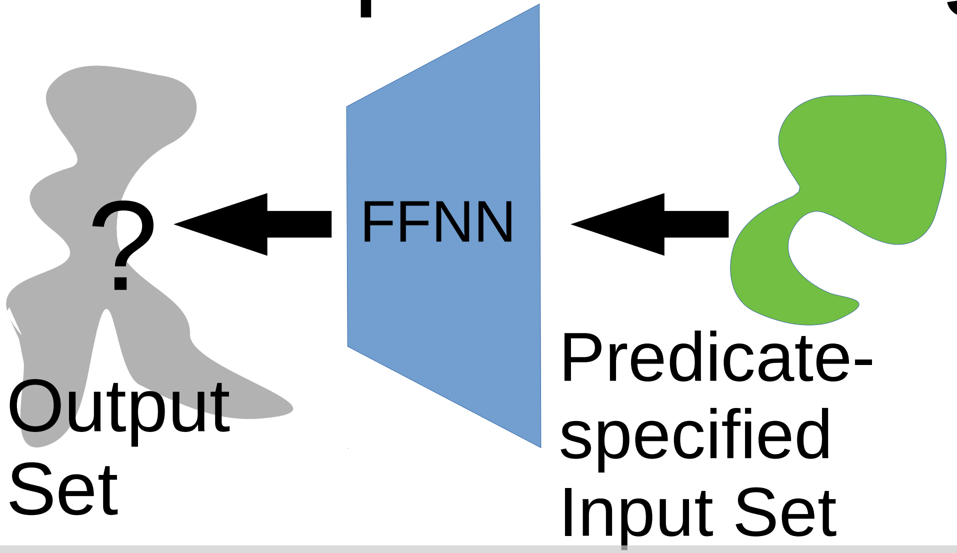
“$\forall v \in B . P(v)$” is false
“$\exists v \in B . P(v)$” is true

Red : P fails to hold
Blue: P holds

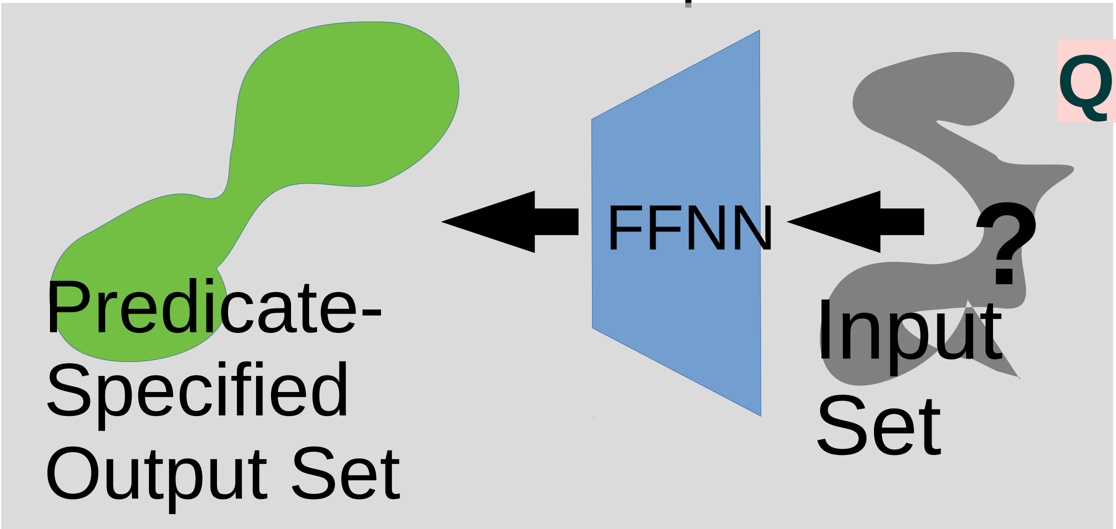
Overview: Responding to Questions



Step 1: "Finding the Other Set"

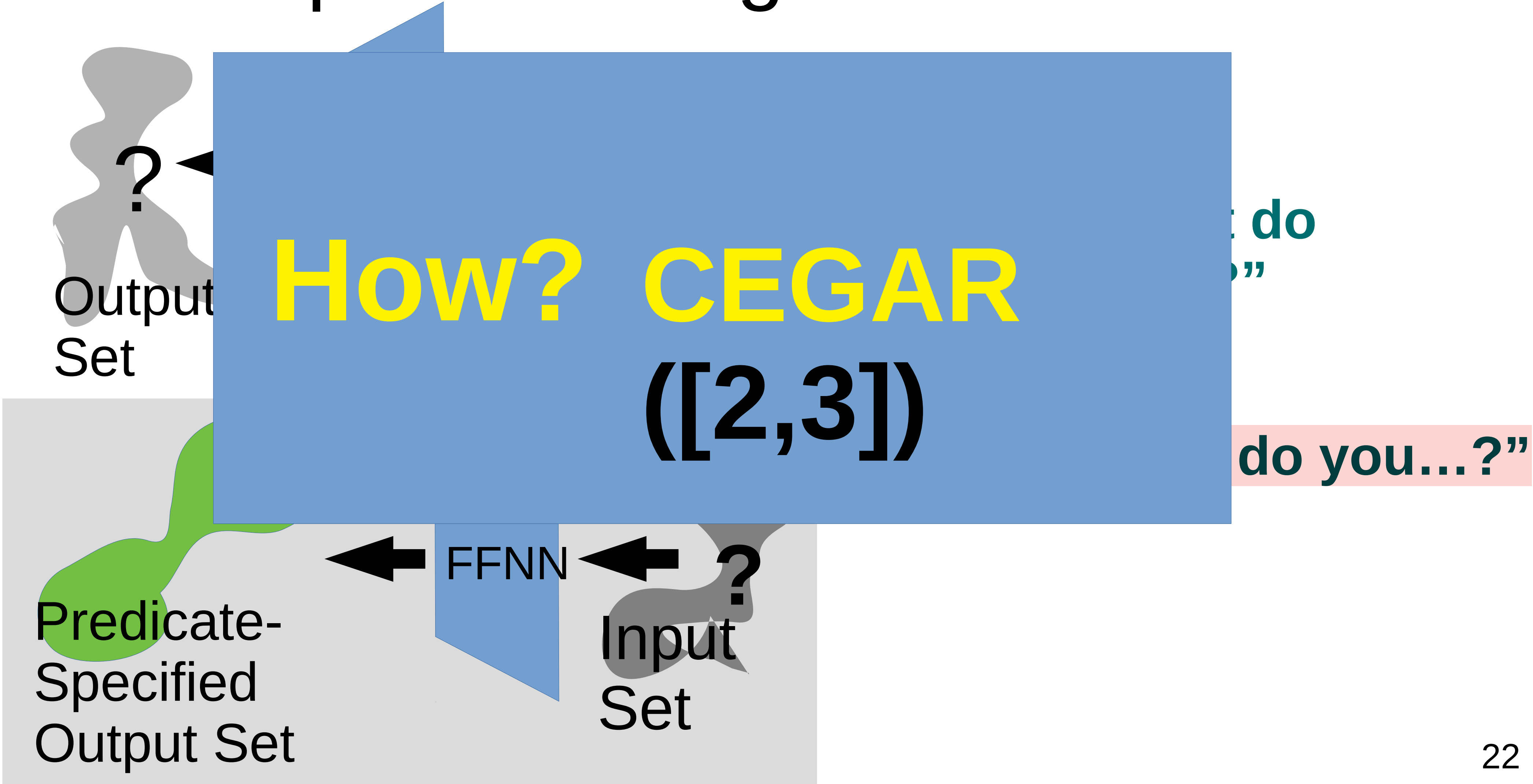


Question: "What do you do when …?"



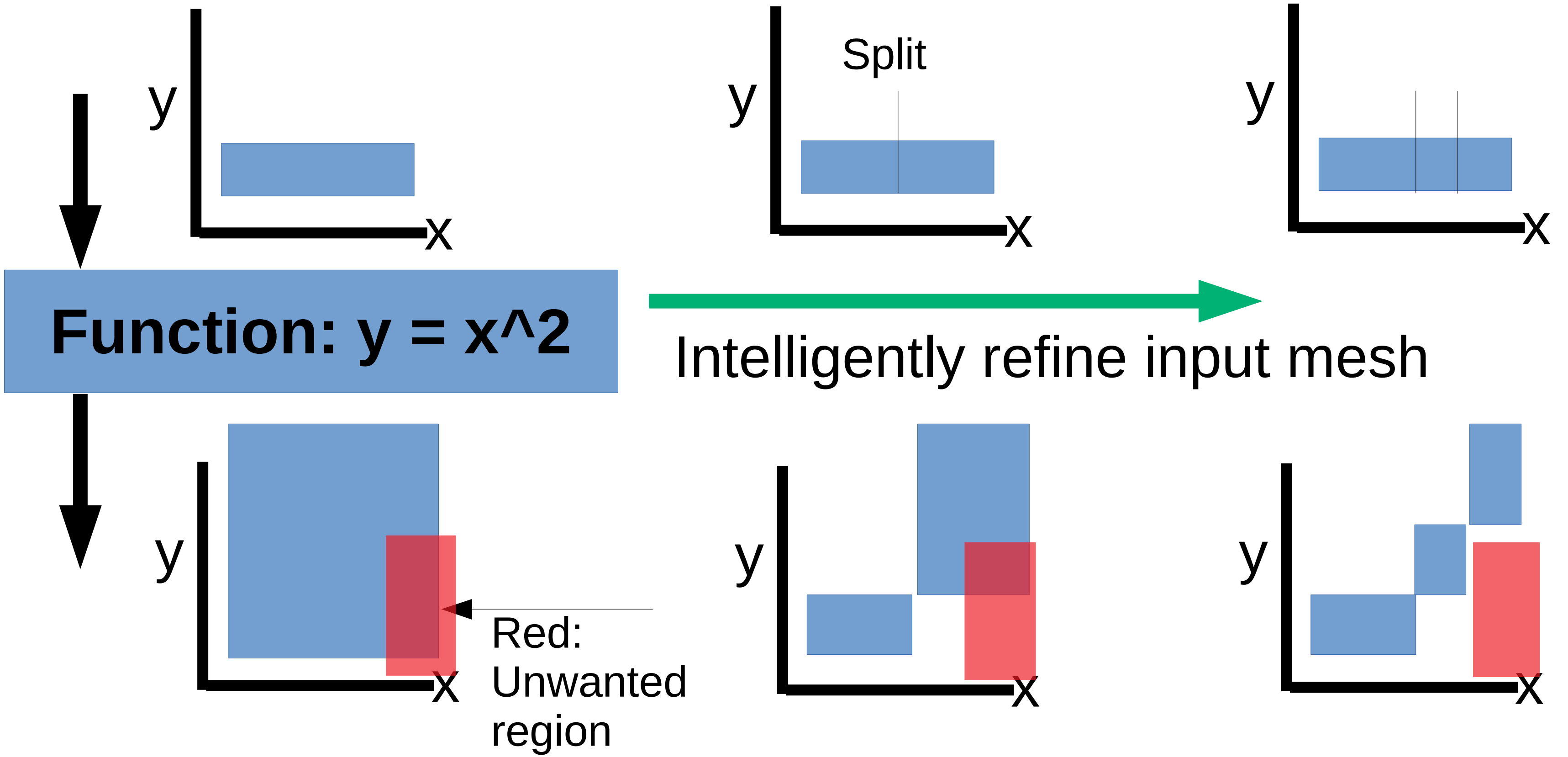
Question: "When do you…?"

Step 1: “Finding the Other Set”



CEGAR, Cnt

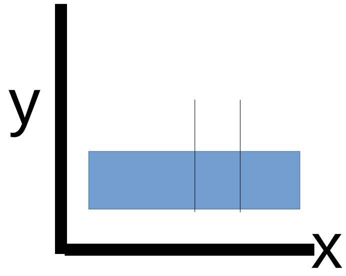
- CEGAR solution: dynamically refine based on property you want to check for
- For us, used hyper-cubes as the abstraction



CEGAR, Cnt

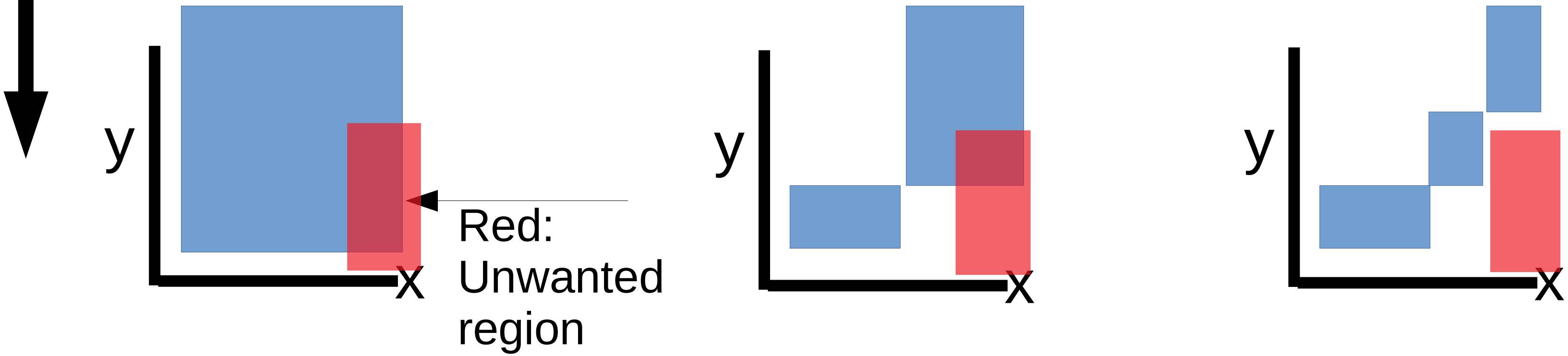
n property you want to check for

Fanoos:
(1) bi-/tri-sect longest (normalized) axis,
(2) continue to refine as desired *while overlapping with* user's predicate-desribed region

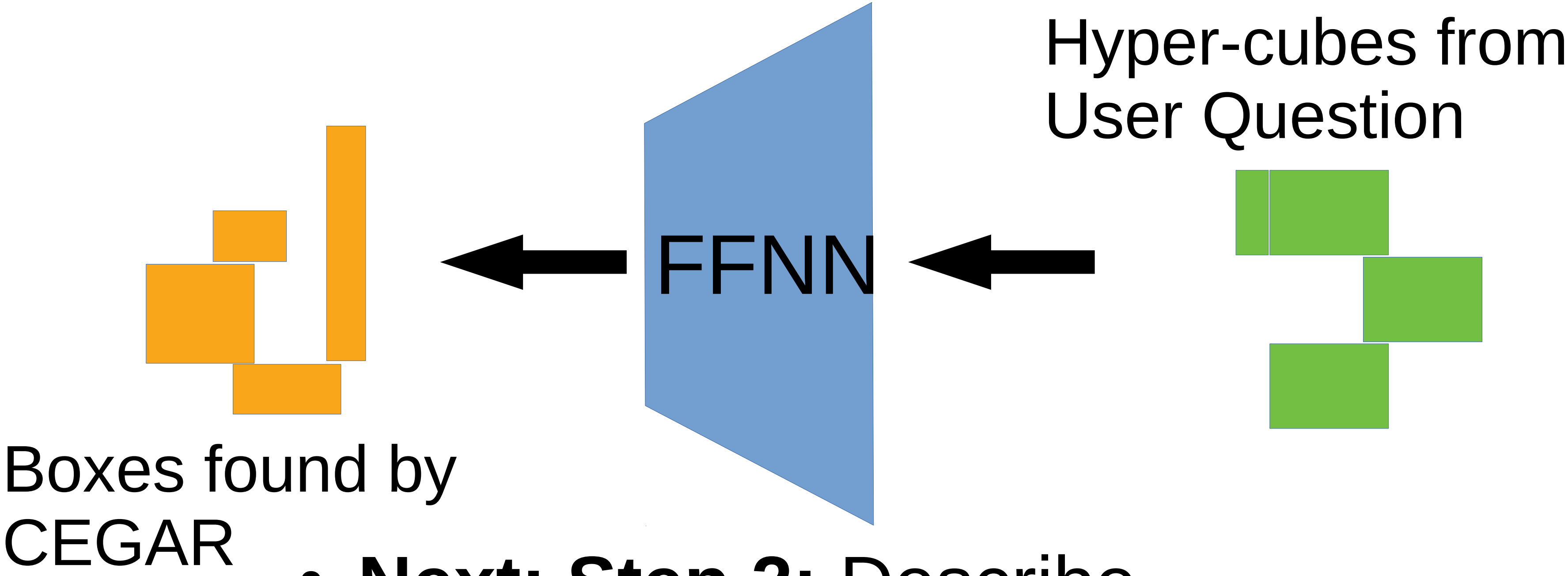


x

efine input mesh



What we have so far:



- **Next: Step 2:** Describe the found boxes
  - How? Form a covering with the predicates.

What we have so far:

Sub-steps:
2.1) Get candidate predicates for each box
2.2) Form global covering from the candidates

Hyper-cubes from User Question

N

Note: Might merge boxes a bit first

- What do we do next?
  - Describe the found boxes

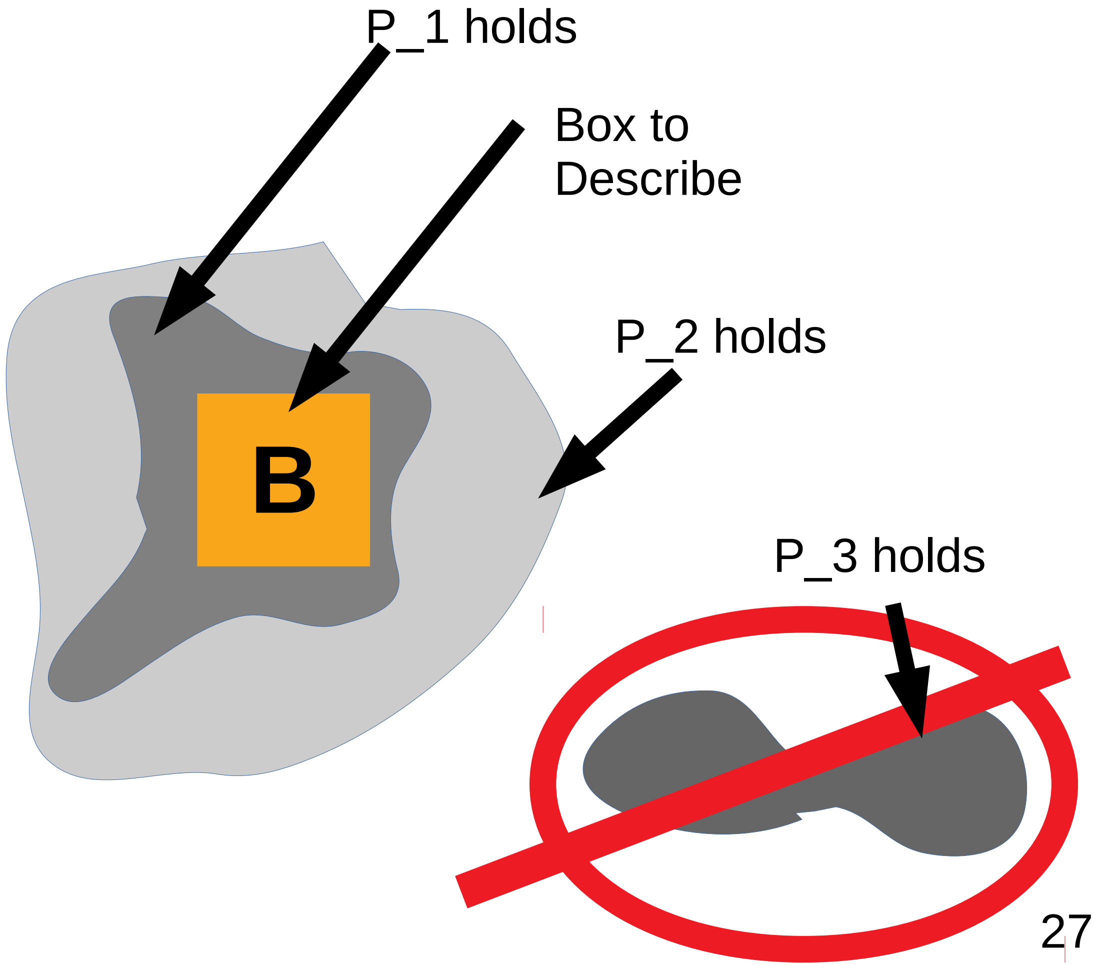
- How? Form a covering with the predicates.

Step 2.1: Getting Candidate Preds. For Each Box

For each box, B:

1) Get preds that hold over B
    1. "feasibility check": try on random sample from B first
    2. Check with SAT-Solver

2) Get most specific preds

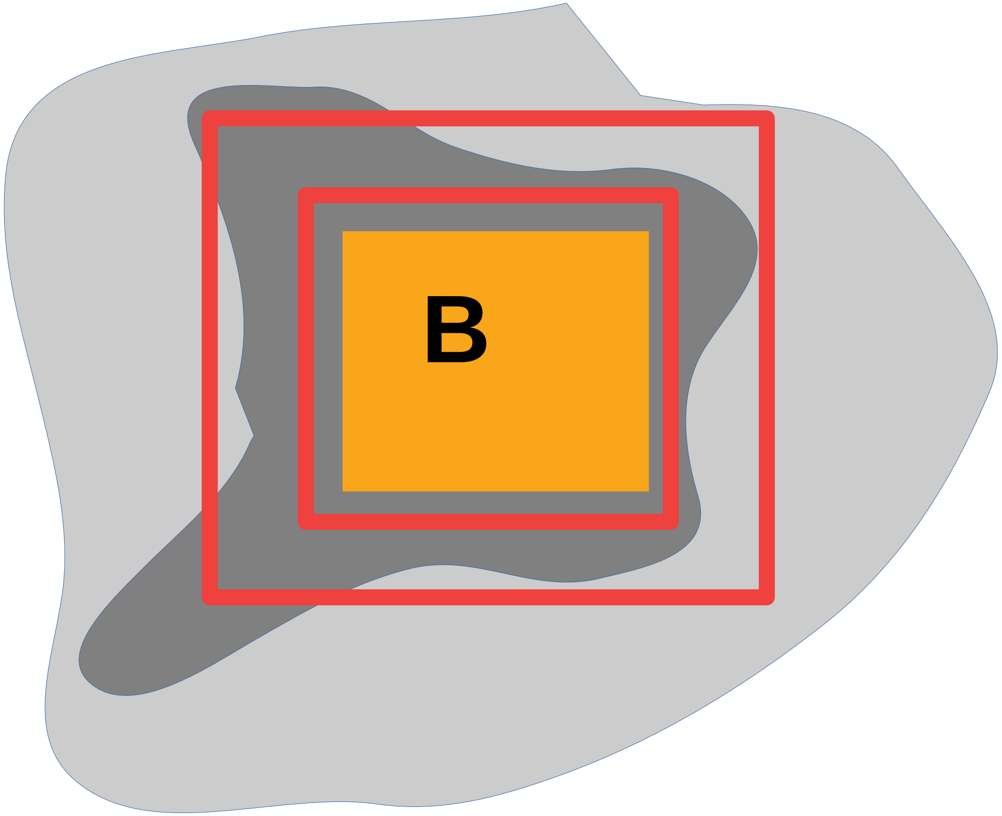


Step 2.1: Getting Candidate Preds. For Each Box

For each box, B:

1) Get preds that hold over B
   1. "feasibility check": try on random sample from B first
   2. Check with SAT-Solver

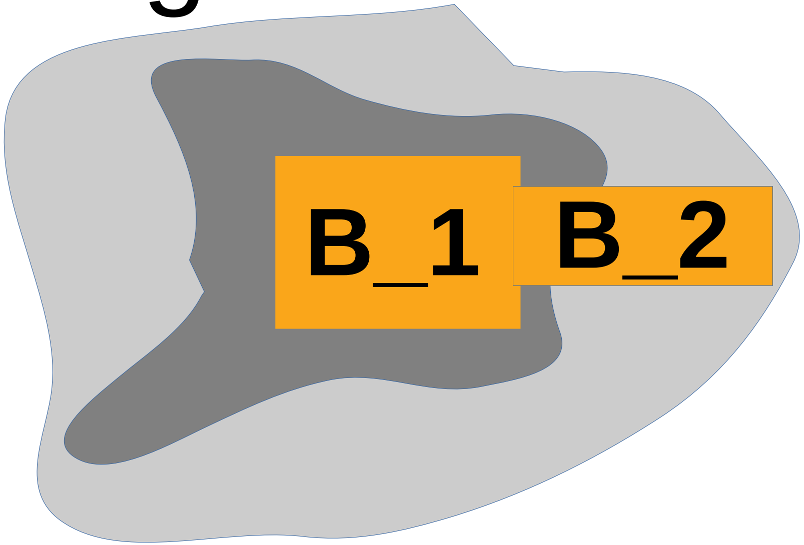
2) Get most specific preds



Note: Can utilize a taxonomy for filtering, If provided

Step 2.2: Forming Global Covering

- Iterative greedy based on score
  - Submodularity
  - Need to throw out “dominated” preds
- Some subtleties for mult-dimensional setting
  - Ex: may need multiple preds to cover box; one pred variable x, another might cover y



Cleaning and Presenting to User

- Some further post-processing
- Gather and show

Normalize “unique” box volumes covered

Normalize total box volumes covered

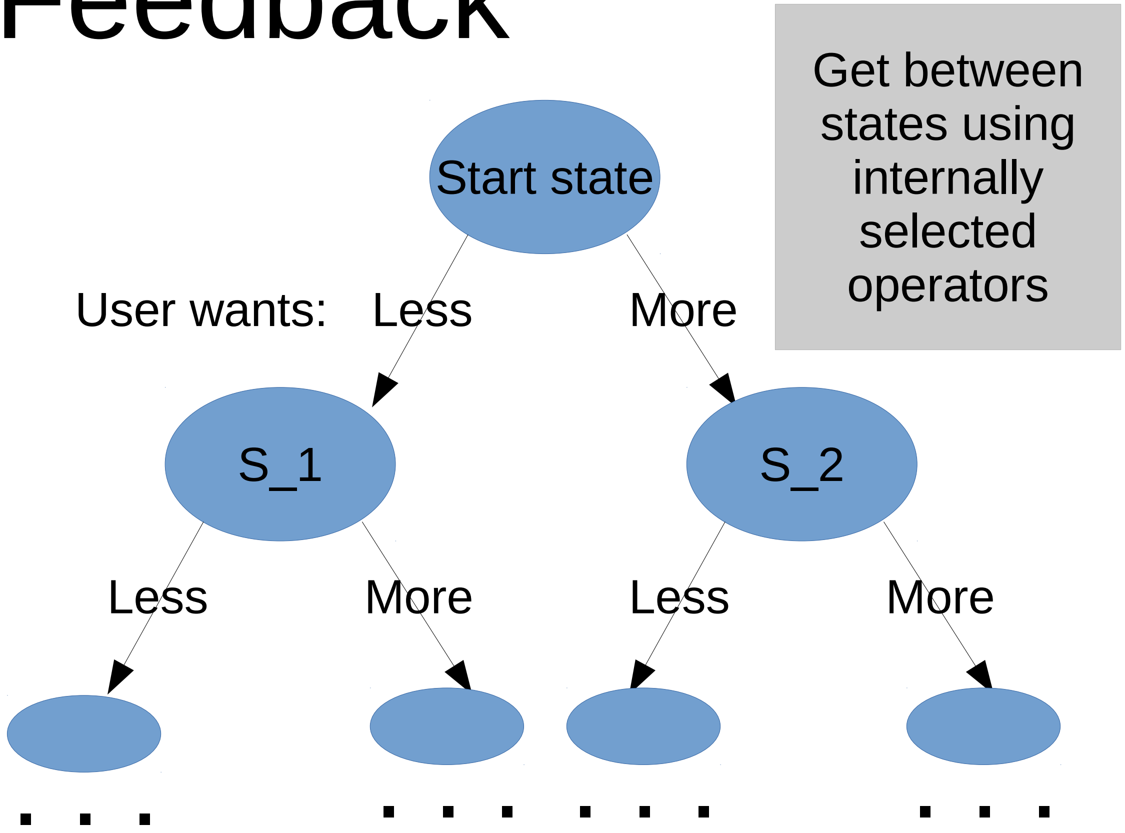
(0.11, 0.34, And(pole1_on_left cart_moving_right))

```
(0.44378316, 0.48588134, pole2 not near target position)
(0.33605014, 0.36551887, pole2angle_rateofchange_high__magnitude)
(0.22016670, 0.23739381, pole2angle_to_right ,
     statevalueestimate_very_low)
```

Output is a Weighted DNF

Using Feedback

- Fanoos has many internal parameters for:
  - CEGAR
  - Box-merging
  - Predicate
  - Etc.
- Use state-operator model
  - Feedback changes state and internal params
  - View as search for proper abstraction level



Using Feedback

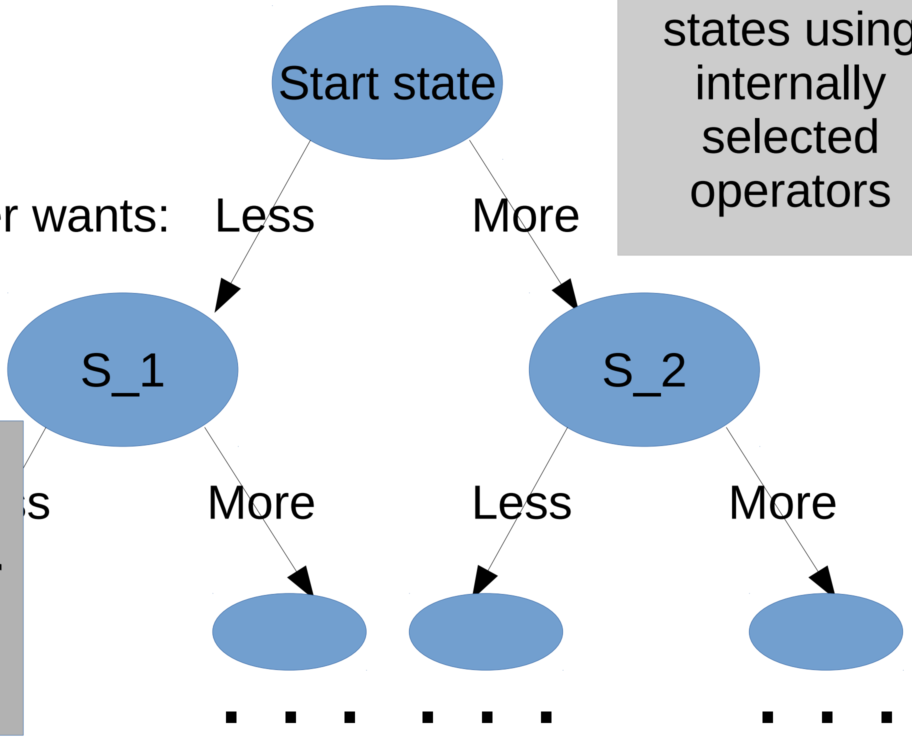
- Fanoos h… paramete…
  - CEGAR
  - Box-me…
  - Predica…
  - Etc.

How select operators?
Hand-written **heuristics**.
Generally try to get smaller boxes, and looser descriptions for greater abstraction

- Use s…
  - Fe… and…
  - Vie… abstraction level

On-going future work: Using ML-back operator selection

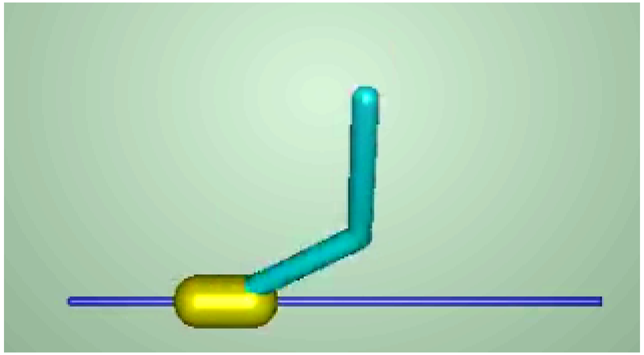
…er wants:



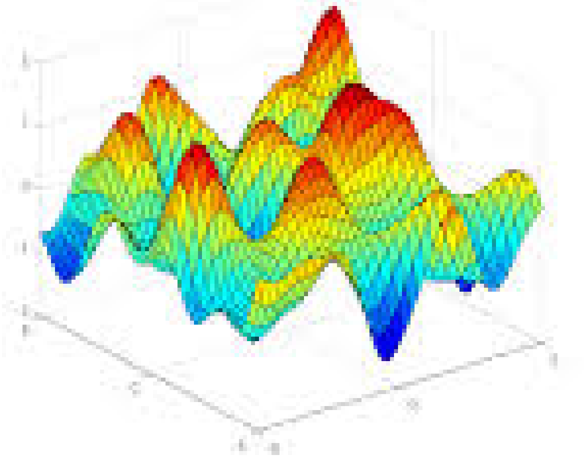
Get between states using internally selected operators

Experiments

- Ran on
  - Invertible double pendulum policy
    - 6D input, 2D output
  - A 3-degree polynomial regression for CPU Usage
    - 5D pre-featurization input, 3D output
- Preds. formed by mix of hand, data statistics, and templates



Openai gym
InvertedDoublePendulum-v2



Experiments

- 130+ Start Questions, Several Hundred Replies Total
  - Questions randomly generated based on some criteria
  - Half asked to make **more abstract (MA)**,
    half asked to make **less abstract (LA)**
- Compared befores-and-afters for:
  - Reachability
  - Result structure
  - Some approximation of an expert agreement

Experiments

- Reachability: MA tend to result in fewer, larger boxes. Opposite for LA
- Structural:
  - MA tend to be shorter, and have fewer conjuncts
  - Based on Jaccard and Overlap score, not just becoming more verbose

Table 2: Median *relative* change in description before and after Fanoos adjusts the abstraction in the requested direction

			CPU	CPU	IDP	IDP
		Request	LA	MA	LA	MA
Reachability	Boxes	Number	8417.5	-8678.0	2.0	-16.0
	Volume	Max	-0.015	0.015	-0.004	0.004
		Median	-0.003	0.003	-0.004	0.004
		Min	-0.001	0.001	-0.003	0.003
		Sum	-0.03	0.03	-0.168	0.166
Structural	Jaccard		0.106	0.211	0.056	0.056
	Overlap coeff.		0.5	0.714	0.25	0.25
	Conjuncts		1.0	-2.0	0.5	-2.5
	Disjuncts		7.0	-7.5	2.0	-2.5
	Named preds.		1.0	-1.0	1.0	-4.5
	Box-Range preds.		2.0	-2.0	1.5	-1.5
Expert	MA term	Multiplicity	3.0	-3.0	24.0	-20.0
		Uniqueness	0.0	0.0	1.0	-1.5
	LA term	Multiplicity	20.0	-21.5	68.5	-86.0
		Uniqueness	2.0	-2.0	12.0	-14.0

Experiments

- Approximate "Expert" Judgement:
  - Labeled each predicate as higher or lower abstractness
  - "grain of salt measure": course labels and did not review whole output
  - As expected: LA requests tended for more lower abstraction terms, opposite for MA requests

Table 2: Median *relative* change in description before and after Fanoos adjusts the abstraction in the requested direction

			CPU	CPU	IDP	IDP
		Request	LA	MA	LA	MA
Reachability	Boxes	Number	8417.5	-8678.0	2.0	-16.0
	Volume	Max	-0.015	0.015	-0.004	0.004
		Median	-0.003	0.003	-0.004	0.004
		Min	-0.001	0.001	-0.003	0.003
		Sum	-0.03	0.03	-0.168	0.166
Structural	Jaccard		0.106	0.211	0.056	0.056
	Overlap coeff.		0.5	0.714	0.25	0.25
	Conjuncts		1.0	-2.0	0.5	-2.5
	Disjuncts		7.0	-7.5	2.0	-2.5
	Named preds.		1.0	-1.0	1.0	-4.5
	Box-Range preds.		2.0	-2.0	1.5	-1.5
Expert	MA term	Multiplicity	3.0	-3.0	24.0	-20.0
		Uniqueness	0.0	0.0	1.0	-1.5
	LA term	Multiplicity	20.0	-21.5	68.5	-86.0
		Uniqueness	2.0	-2.0	12.0	-14.0

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Edmund Clarke, Ansgar Fehnker, Zhi Han, Bruce Krogh, Olaf Stursberg, and Michael Theobald. 2003. Verification of hybrid systems based on counterexample-guided abstraction refinement. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 192–207.

[3] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. 2000. Counterexample-guided abstraction refinement. In *International Conference on Computer Aided Verification*. Springer, 154–169.

[4] Patrick Cousot and Radhia Cousot. 1977. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. In *Conference Record of the Fourth ACM Symposium on Principles of Programming Languages, Los Angeles, California, USA, January 1977*, Robert M. Graham, Michael A. Harrison, and Ravi Sethi (Eds.). ACM, 238–252. https://doi.org/10.1145/512950.512973

[5] Bradley Hayes and Brian Scassellati. 2016. Autonomously constructing hierarchical task networks for planning and human-robot collaboration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5469–5476.

[6] Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 303–312.

[7] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43, 2 (01 Feb 2019), 309–326. https://doi.org/10.1007/s10514-018-9771-0

[8] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. (2018), 577–593 pages. https://doi.org/10.1007/978-3-030-01216-8_35

[9] Anurag Koul, Alan Fern, and Sam Greydanus. 2019. Learning Finite State Representations of Recurrent Policy Networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=S1gOpsCctm

[10] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[11] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (2016), 1135–1144 pages. https://doi.org/10.1145/2939672.2939778