



Improving Decision Making with Machine Learning, Provably

Manuel Gomez Rodriguez

Includes joint work with Eleni Straitouri, Luke Wang & Nastaran Okati

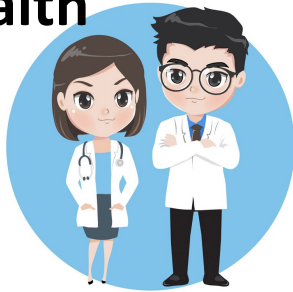


MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

Machine learning to improve decision making

Machine learning promises a new generation of automated decision support systems in many high-stakes domains:

Health



Hiring



Content moderation



Education



Security



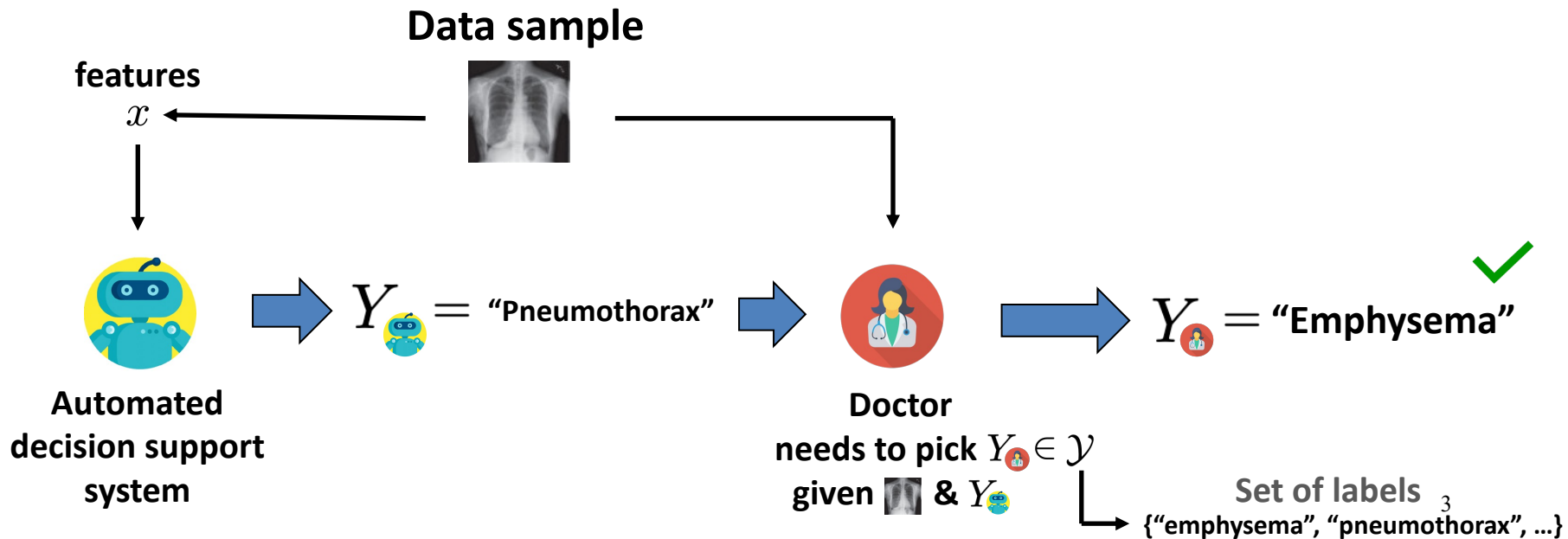
Justice



Finance

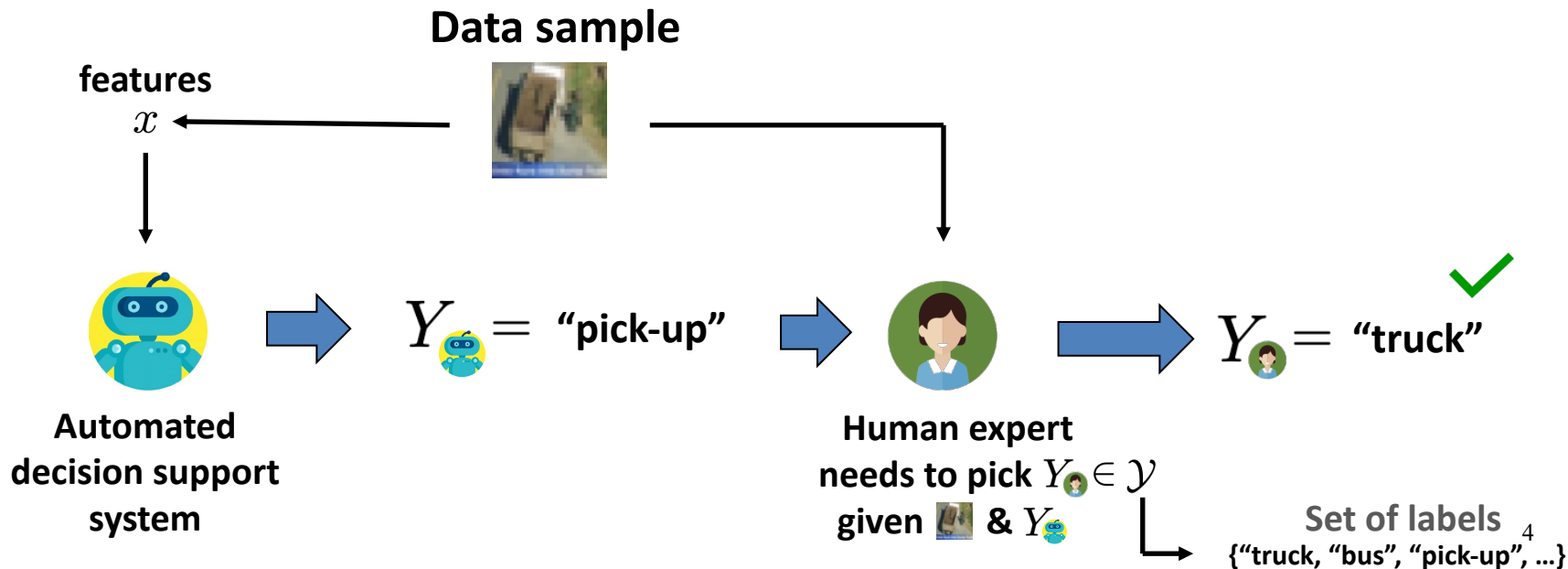
Decision support systems for classification tasks

Machine learning has mainly focused on **decision support systems** for **classification tasks**



Decision support systems for classification tasks

Machine learning has mainly focused on **decision support systems** for **classification tasks**



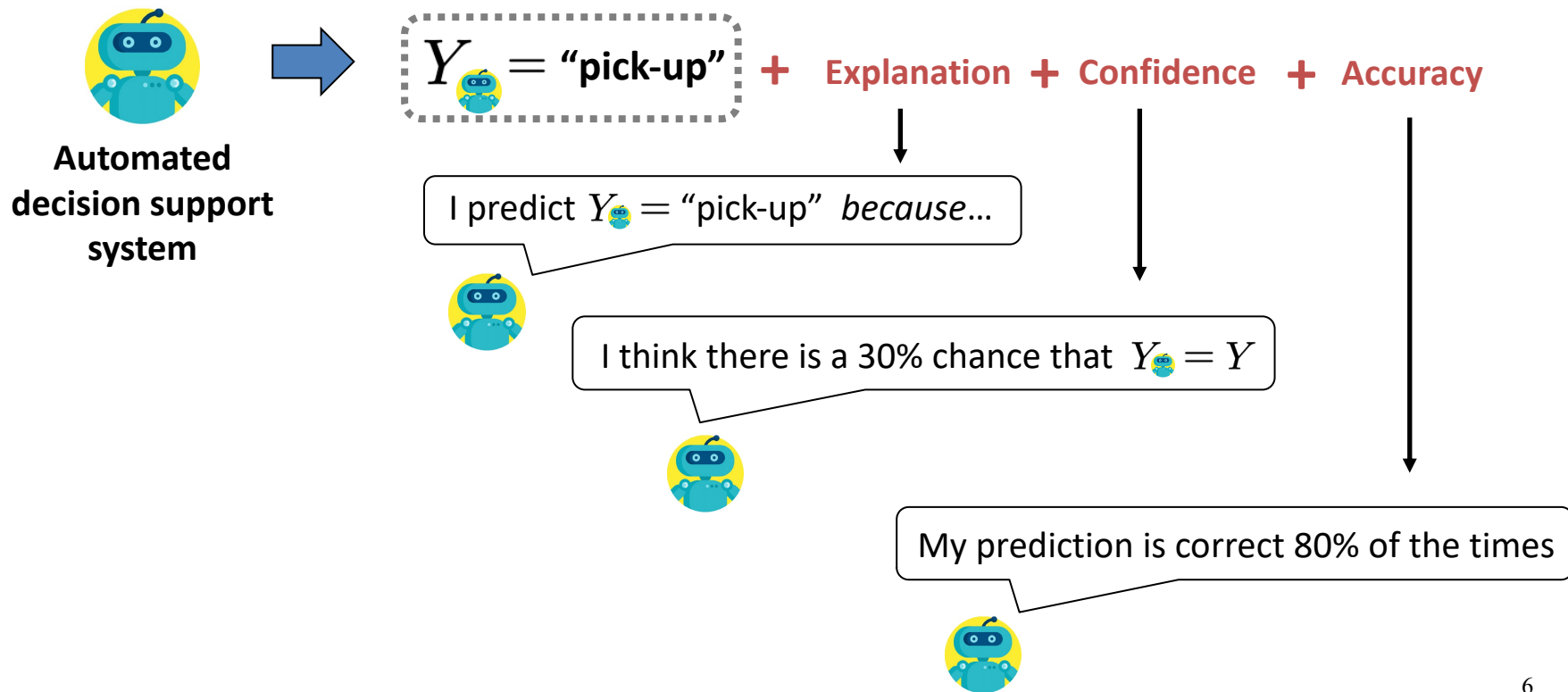
Human experts need to understand when to trust the classifier



Human needs to **understand when to trust** a prediction Y_{robot} made by the decision support system

- ➡ This follows from the fact that, in general, the accuracy of the system differs across data samples
- ➡ **Otherwise, they may be better off on their own**

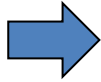
How do decision support systems *modulate* trust?



How do decision support systems *modulate* trust?



Automated
decision support
system



$$Y = \text{"pick-up"}$$

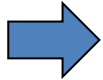
+

Explanation + Confidence + Accuracy

Does this **additional information** help humans understand when to trust a prediction?

Not always. The empirical findings are mixed and seem to depend on the application domain.

How do decision support systems *modulate* trust?



$$Y_{\text{robot}} = \text{"pick-up"}$$

+

Explanation + Confidence + Accuracy

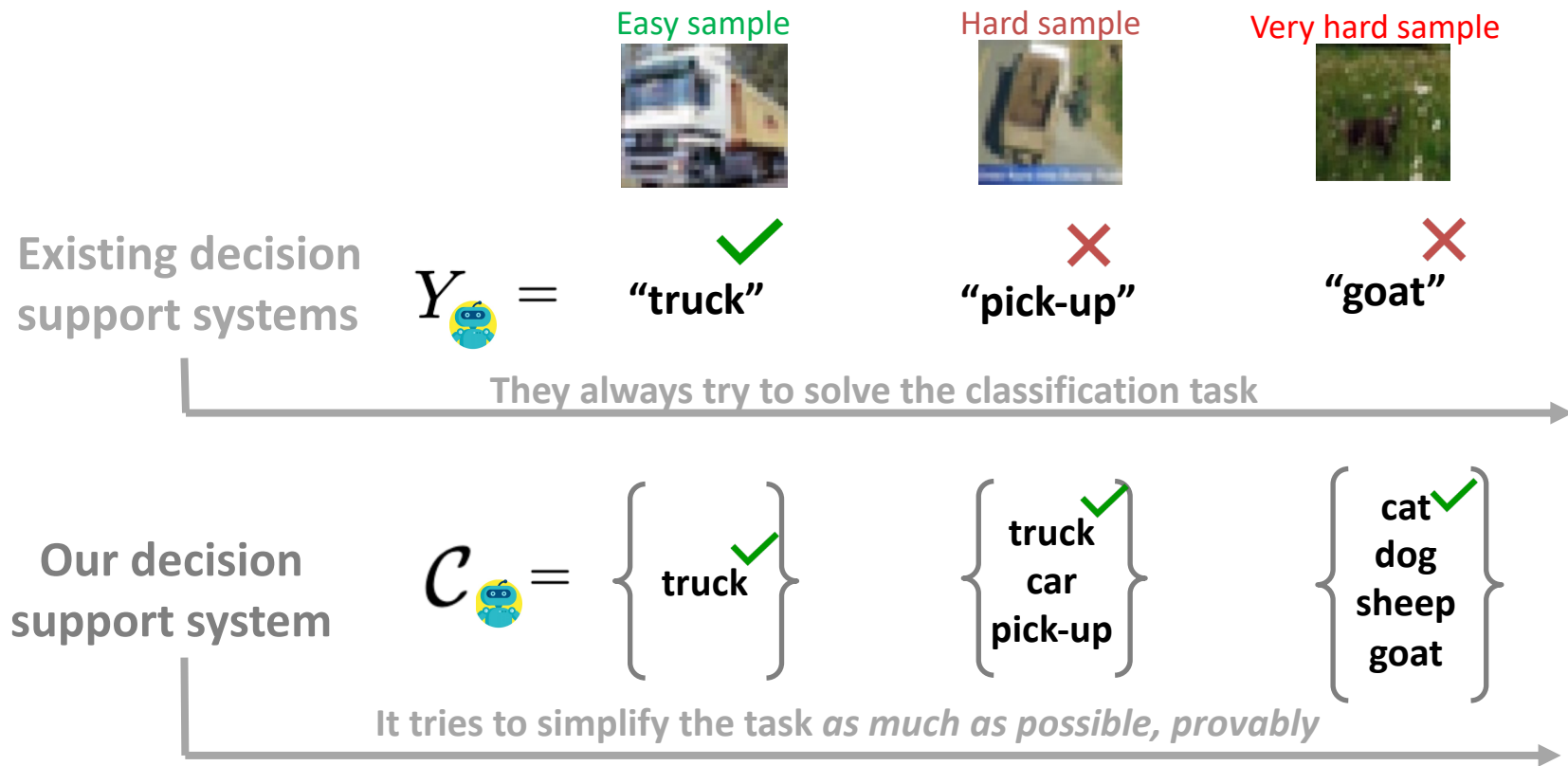
Automated
decision support
system

Does the system
when the

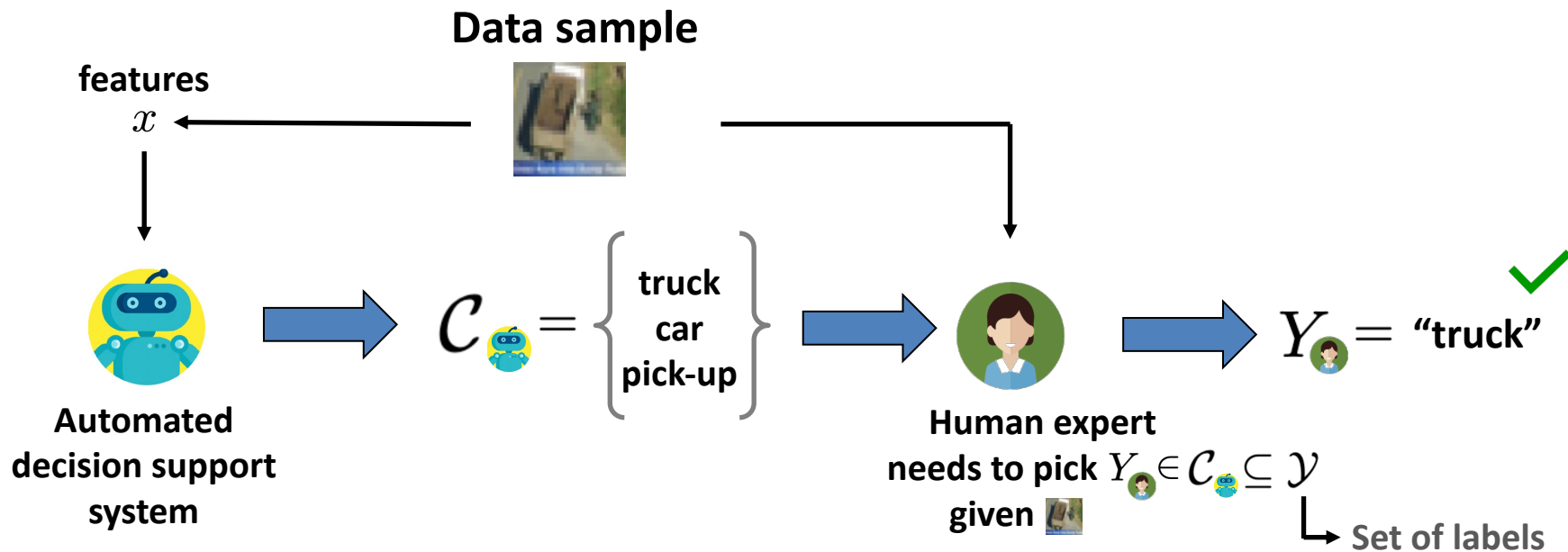
Not yet clear how to make sure experts do
not develop a misplaced trust

Not always. The empirical findings are mixed and
seem to depend on the application domain.

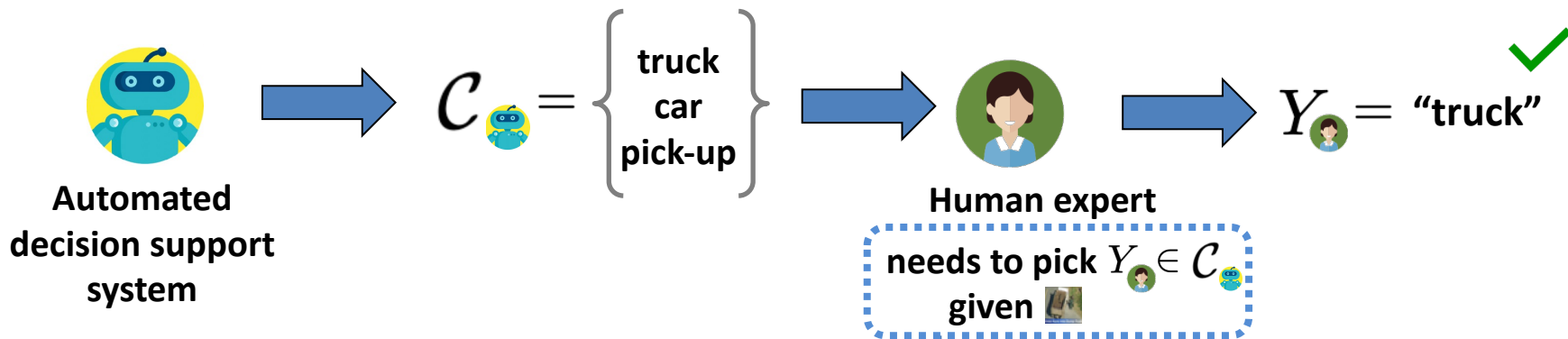
A decision support system that simplifies, rather than solves



A new type of decision support systems for classification



Humans do not need to understand when to trust the system



The human **does not need to understand** when to **trust** the system

→ However, we need to ensure that the subset $\mathcal{C}$ contains the **true label Y with high probability**

↓
"truck"

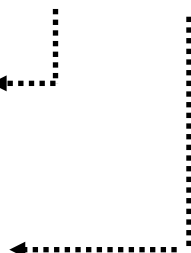
Trustworthy subsets using conformal prediction

To ensure that the subsets $\mathcal{C}_{\text{robot}}$ contain the **true label** Y with **high probability**, we rely on **conformal prediction**.

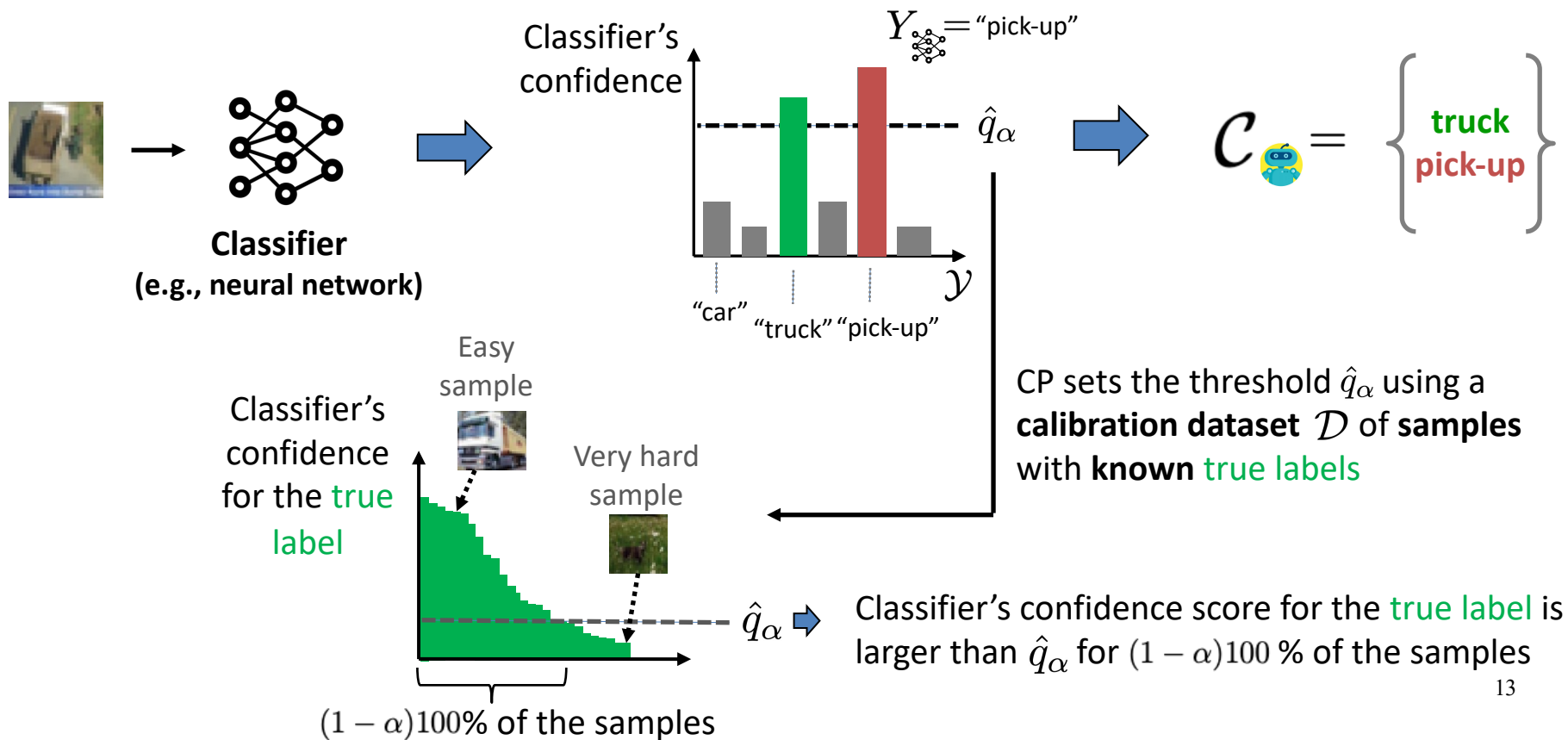
Conformal prediction (CP) is a statistical technique to construct *trustworthy subsets* $\mathcal{C}_{\text{robot}}$

$1 - \alpha$
Desired coverage probability $\Rightarrow$ CP guarantees that $\Pr(Y \in \mathcal{C}_{\text{robot}}) = 1 - \alpha$

$$1 - \alpha = \frac{3}{4}$$

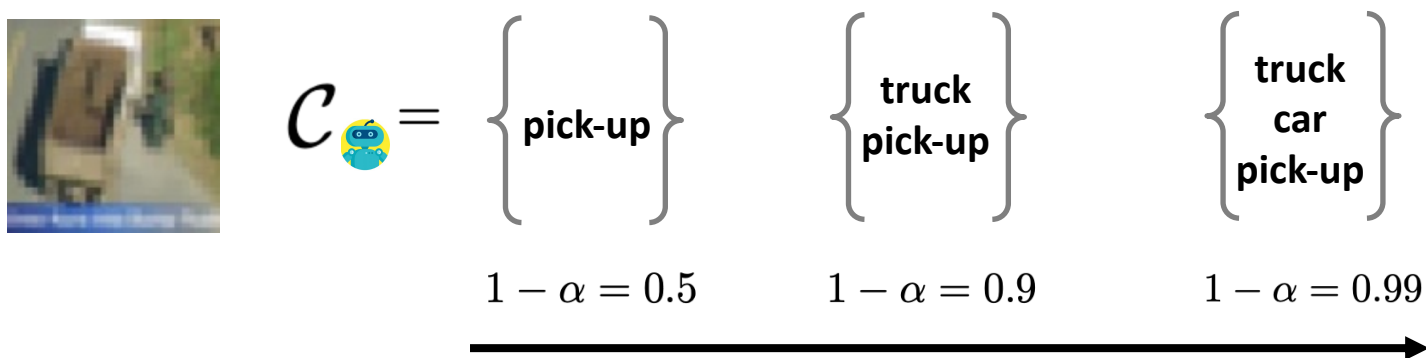


Conformal predictors in a nutshell



Conformal prediction is not magic

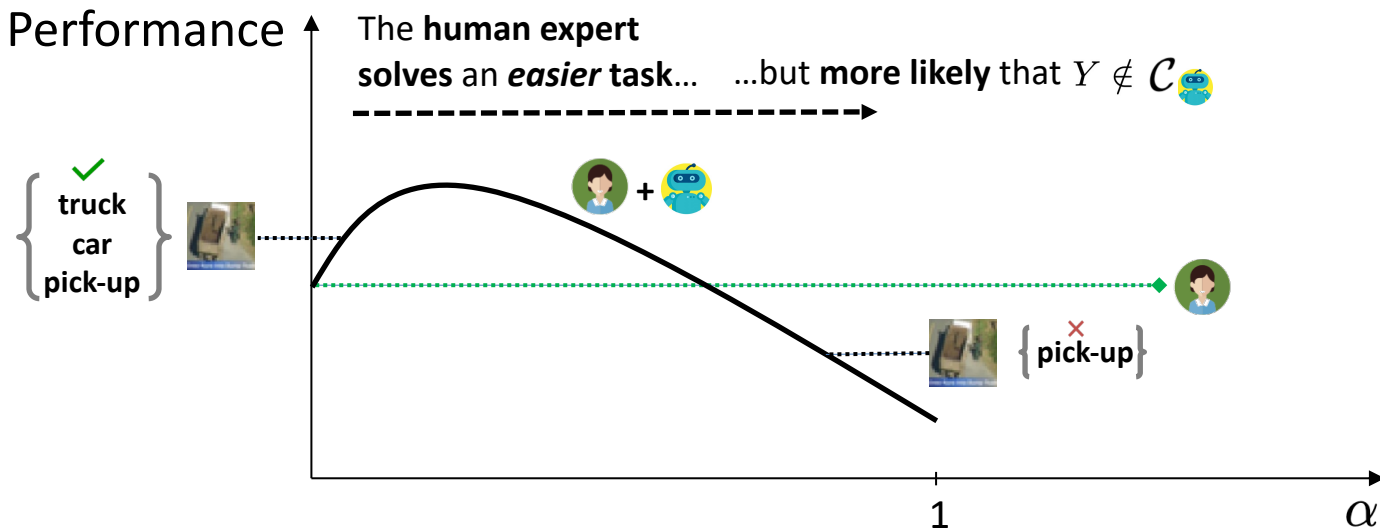
Depending on the desired coverage probability $1 - \alpha$, the **size of the subsets** $\mathcal{C}_{\text{robot}}$ constructed by a **conformal predictor** varies



The *larger* the desired coverage probability $1 - \alpha$, the larger the subsets $\mathcal{C}_{\text{robot}}$

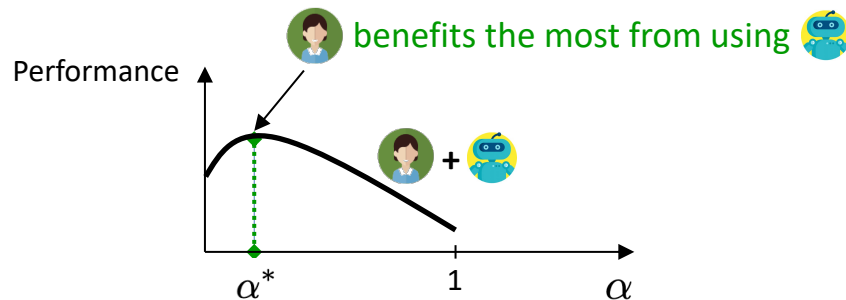
Optimizing across conformal predictors

The parameter α trade-offs **how frequently** the **system** will **mislead** the **human expert** & the **difficulty** of the **task** the **human** needs to solve



Bandit algorithms to find the optimal conformal predictor

To find the optimal parameter α^* ,
we resort to **bandit algorithms**



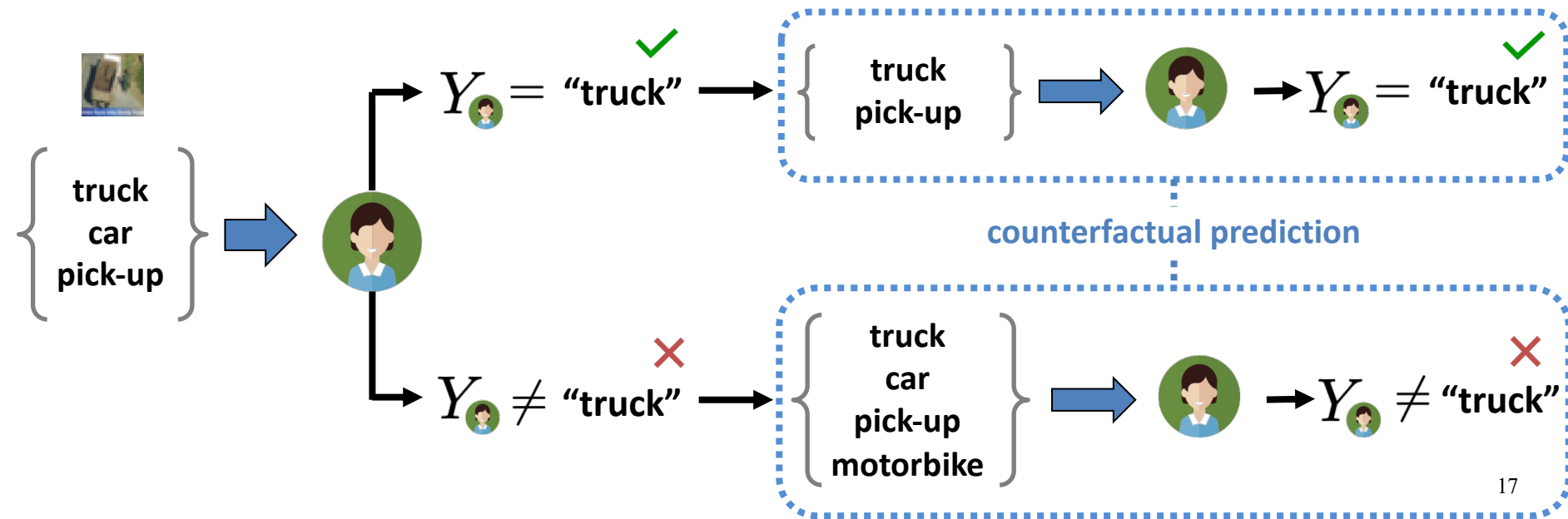
Bandit algorithms sequentially gather predictions by human experts using our system under different α values...

...**prioritizing** α values that seem **more promising over time**.

many bandit algorithms, e.g.,
successive elimination, UCB1

Efficient bandit algorithms using counterfactual monotonicity

We **speed-up** how quickly **bandit algorithms** gather predictions for different α values using a **counterfactual monotonicity assumption**



Exponential improvement in regret in successive elimination

For **successive elimination (SE)**, a well-known bandit algorithm, we show that **counterfactual monotonicity** allows for an **exponential improvement in regret**

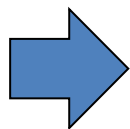
number of time steps
↓

$$\underbrace{R(T)}_{\text{regret}} = T \cdot \underbrace{\mathbb{E}_{\alpha^*} [\mathbb{I}\{Y_{\text{person}} = Y\}]}_{\text{performance under optimal } \alpha^*} - \underbrace{\sum_{t=1}^T \mathbb{E}_{\alpha_t} [\mathbb{I}\{Y_{\text{person}} = Y\}]}_{\text{performance under } \alpha_t \text{ values chosen by SE}}$$

Vanilla SE

$$\mathbb{E}[R(t)] \leq O(\sqrt{\underbrace{m}_{\text{size of the calibration dataset } \mathcal{D} \text{ used by CP}} t \log T})$$

size of the calibration
dataset $\mathcal{D}$ used by CP



SE with counterfactual monotonicity

$$\mathbb{E}[R(t)] \leq O(\sqrt{t \log m \log T})$$

Large-scale human subject study

We gather **194,407** predictions from **2,751** human subjects over **19,200** different pairs of natural images and subsets.

Which one of the following categories fits better the image below?



- ☐ Car
- ☐ Airplane
- ☐ Truck

Strict implementation

Human experts are allowed to

pick $Y_{\text{human}} \in \mathcal{C} \subseteq \mathcal{Y}$

- ☐ Car
- ☐ Airplane
- ☐ Truck
- ☐ Other

Lenient implementation

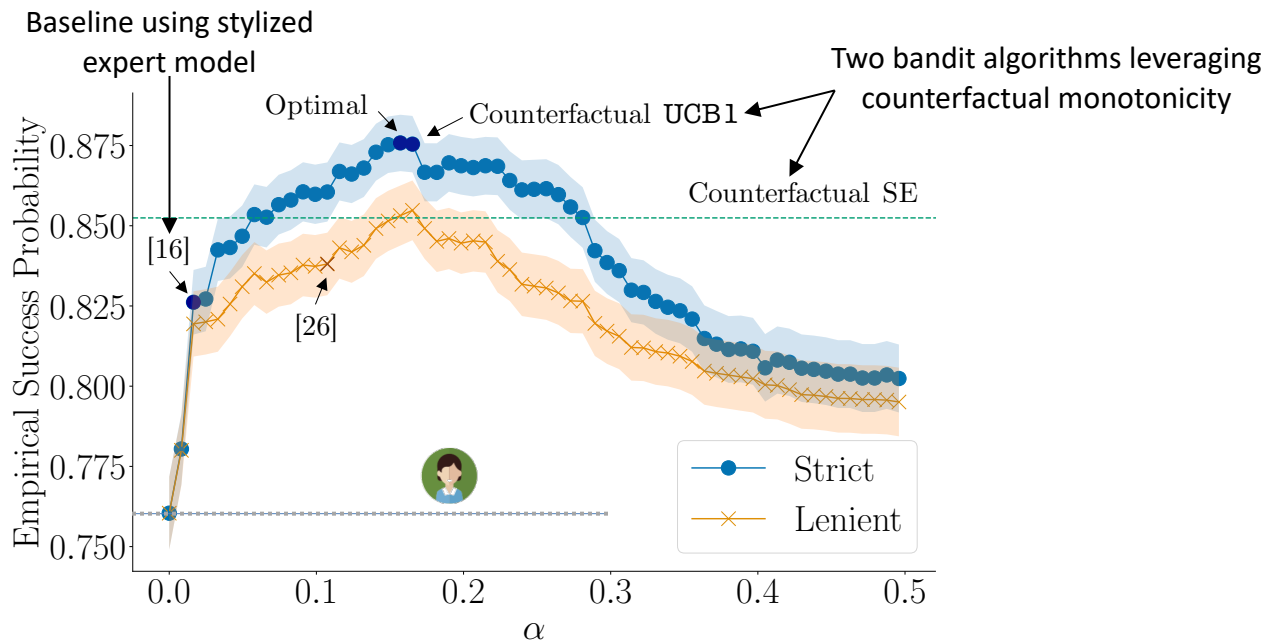
Human experts are allowed to

pick $Y_{\text{human}} \in \mathcal{Y}$

1. If you chose 'Other' above, please choose a category:

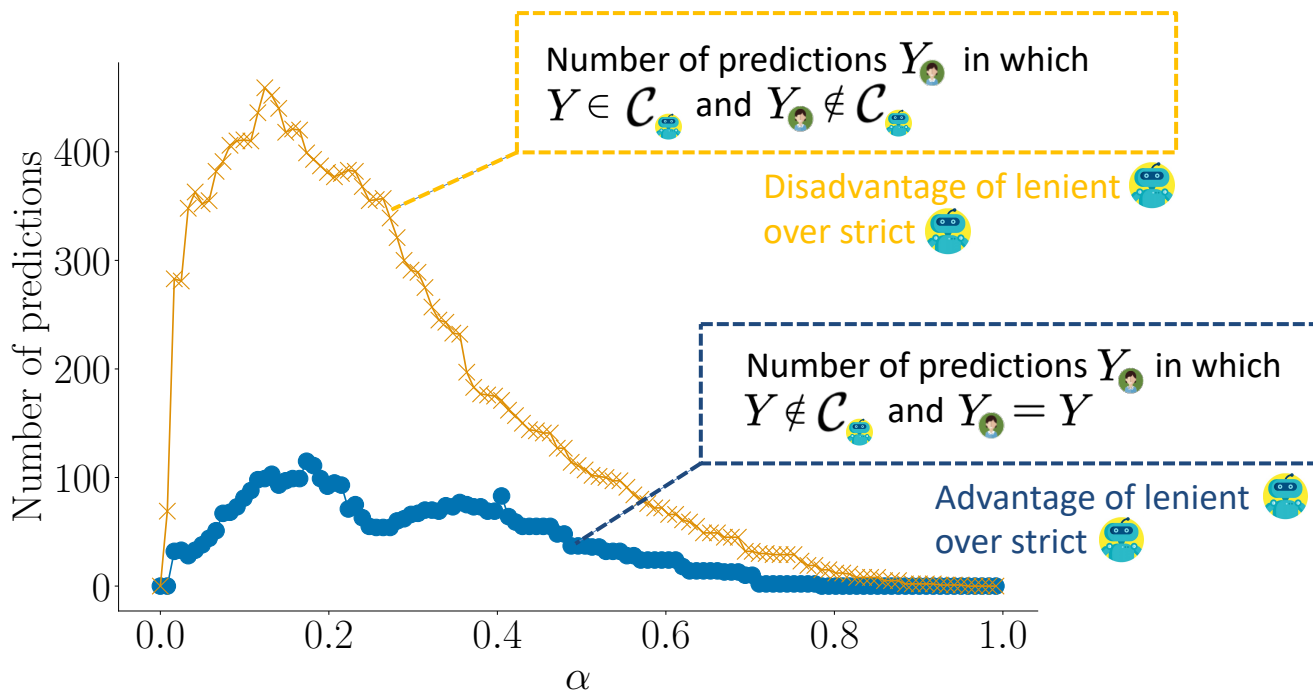
Choose

Limiting expert's level of agency offers greater performance



The **strict implementation**, which **adaptively limits experts' agency**, beats the **lenient implementation**, which **allows experts to always exercise their agency**

Allowing experts to exercise their own agency does not pay off



Moving beyond classification tasks

There are **many decision making processes** where one does not need to solve **classification tasks** but **other types of tasks**.

Huge amount of excitement about the possibility of using **sophisticated LLMs (e.g., ChatGPT)** to improve **decision making**.

→ However, human experts still need to **understand when to trust** the answers provided by LLMs.

Developing **trustworthy decision support systems** using **LLMs** is **highly non trivial**.

Thanks!

Improving Expert Predictions with Conformal Prediction, ICML 2023

<https://arxiv.org/abs/2201.12006>

<https://github.com/Networks-Learning/improve-expert-predictions-conformal-prediction>

Designing Decision Support Systems Using Counterfactual Prediction Sets, Arxiv 2023

<https://arxiv.org/abs/2306.03928>

<https://github.com/Networks-Learning/counterfactual-prediction-sets>



Eleni



Nastaran



Luke

Learn more about our research at
learning.mpi-sws.org