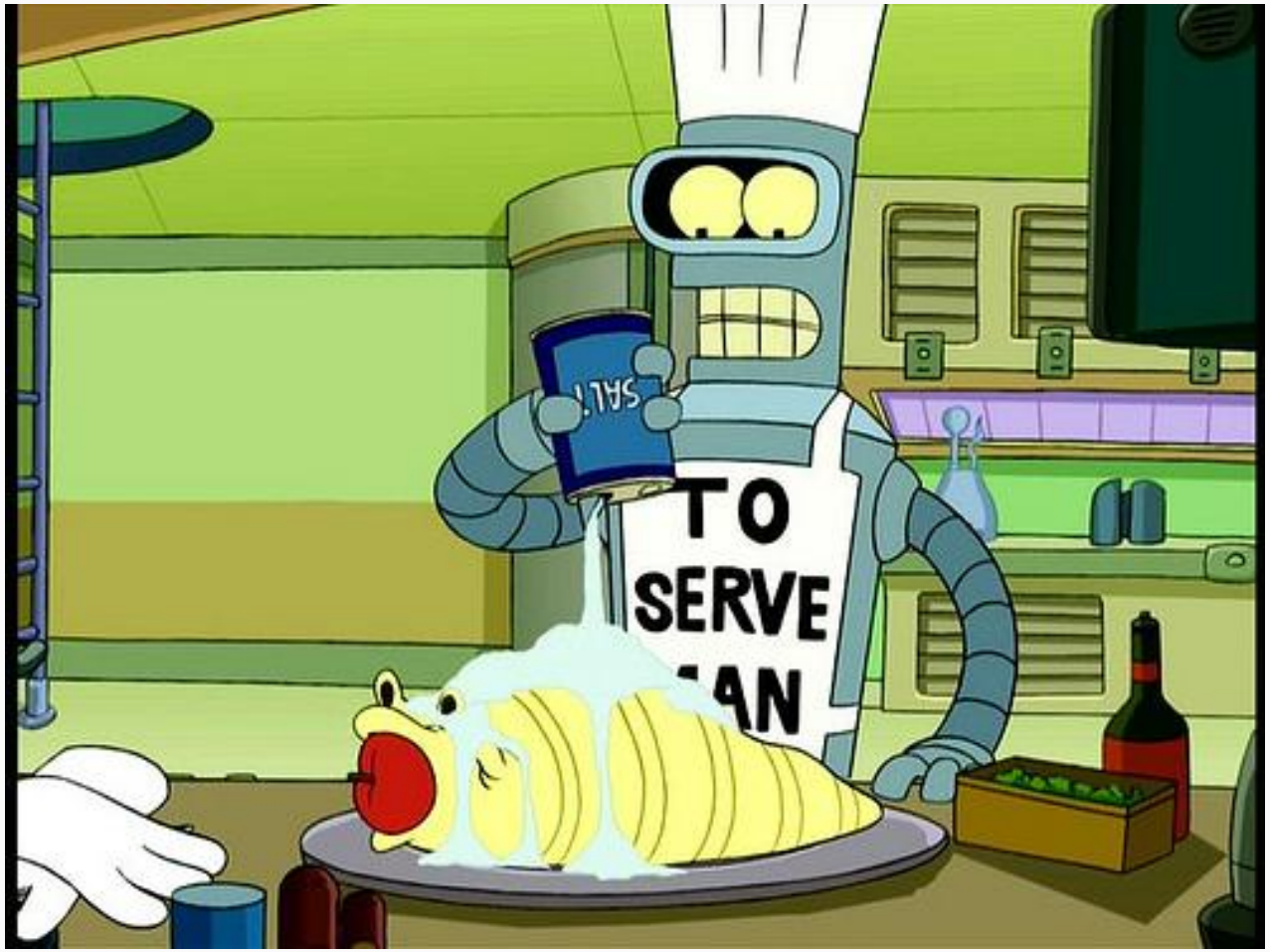
Five or so actionable tips for building trust and being trustworthy (in interactive learning)

Stefano Teso, University of Trento

stefano.teso@unitn.it



ML increasingly being used in sensitive domains like *criminal justice*, *hiring*, ...



Black-box ML models can be whimsical and hard to control

How can we build justifiable trust into black-box ML models?

How do humans establish/reject trust into others?

[1] R Hoffman et al. "Trust in automation". In: *IEEE Intelligent Systems* (2013).

[2] Luke Chang et al. "Seeing is believing: Trustworthiness as a dynamic belief". In: *Cognitive psychology* (2010).

How do humans establish/reject trust into others?

Understanding Trust involves understanding the other's beliefs & intentions; it depends on the perceived **understandability** and **competence**[1]

Interaction Trust is updated dynamically, **interaction** establishes expectations[2]; it relies on **directability**

(We have "hardware support" for all of this: *theory of mind*, *mirror neurons*, . . .)

---

[1] R Hoffman et al. "Trust in automation". In: *IEEE Intelligent Systems* (2013).

[2] Luke Chang et al. "Seeing is believing: Trustworthiness as a dynamic belief". In: *Cognitive psychology* (2010).

How do humans establish/reject trust into others?

Understanding Trust involves understanding the other's beliefs & intentions; it depends on the perceived **understandability** and **competence**[1]

Interaction Trust is updated dynamically, **interaction** establishes expectations[2]; it relies on **directability**

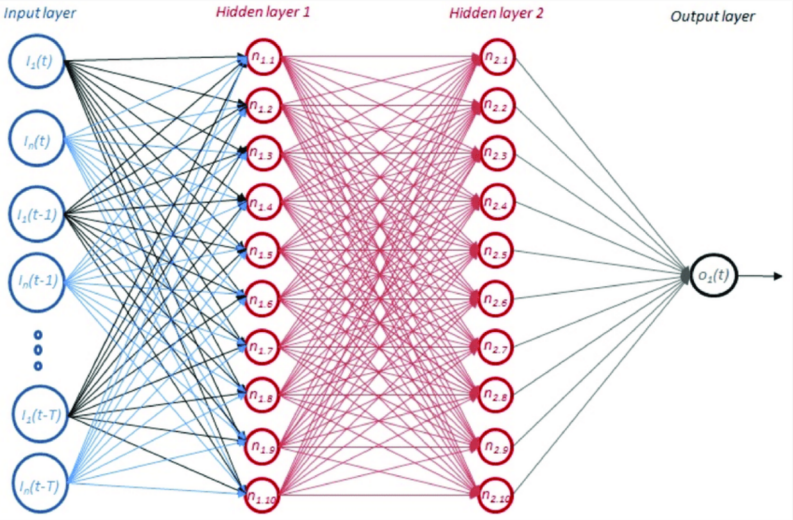
(We have "hardware support" for all of this: *theory of mind*, *mirror neurons*, . . .)

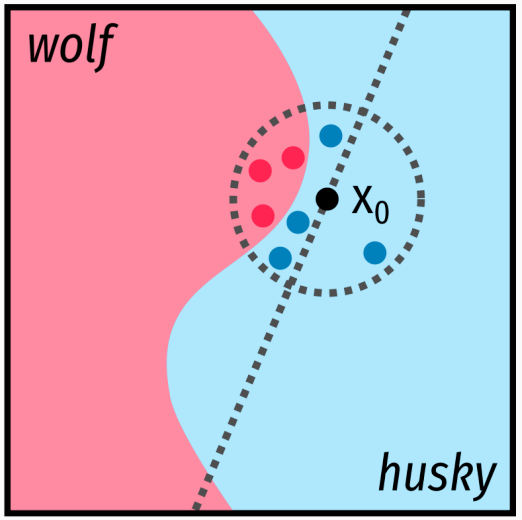
Alas, **explainable AI is passive** and **interactive ML is opaque**

[1] R Hoffman et al. "Trust in automation". In: *IEEE Intelligent Systems* (2013).

[2] Luke Chang et al. "Seeing is believing: Trustworthiness as a dynamic belief". In: *Cognitive psychology* (2010).

Local Explanations with LIME

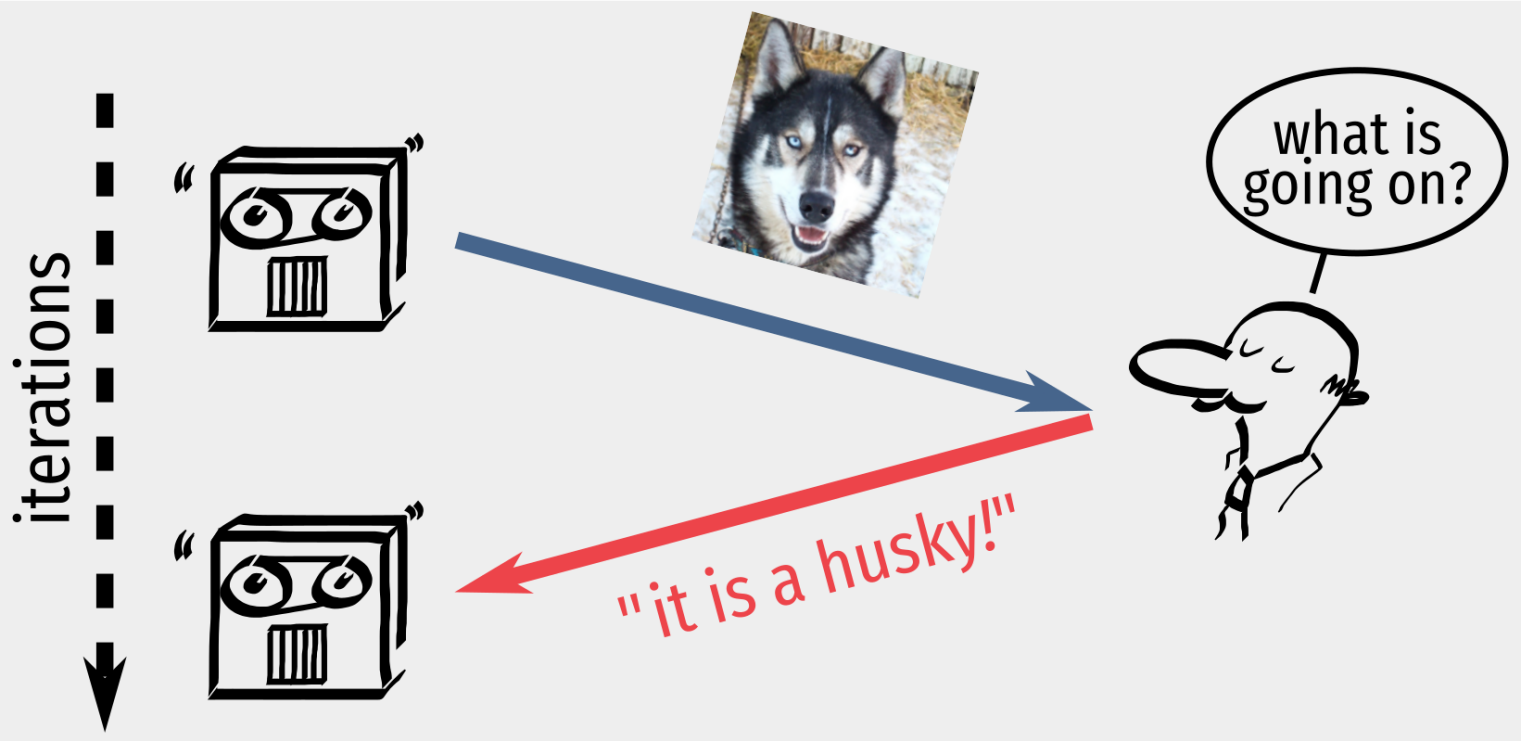




This helps to **identify** potential "Clever Hans" behavior[3], but it does not provide the means to **fix** it

[3] Sebastian Lapuschkin et al. "Unmasking clever hans predictors and assessing what machines really learn". In: (2019).

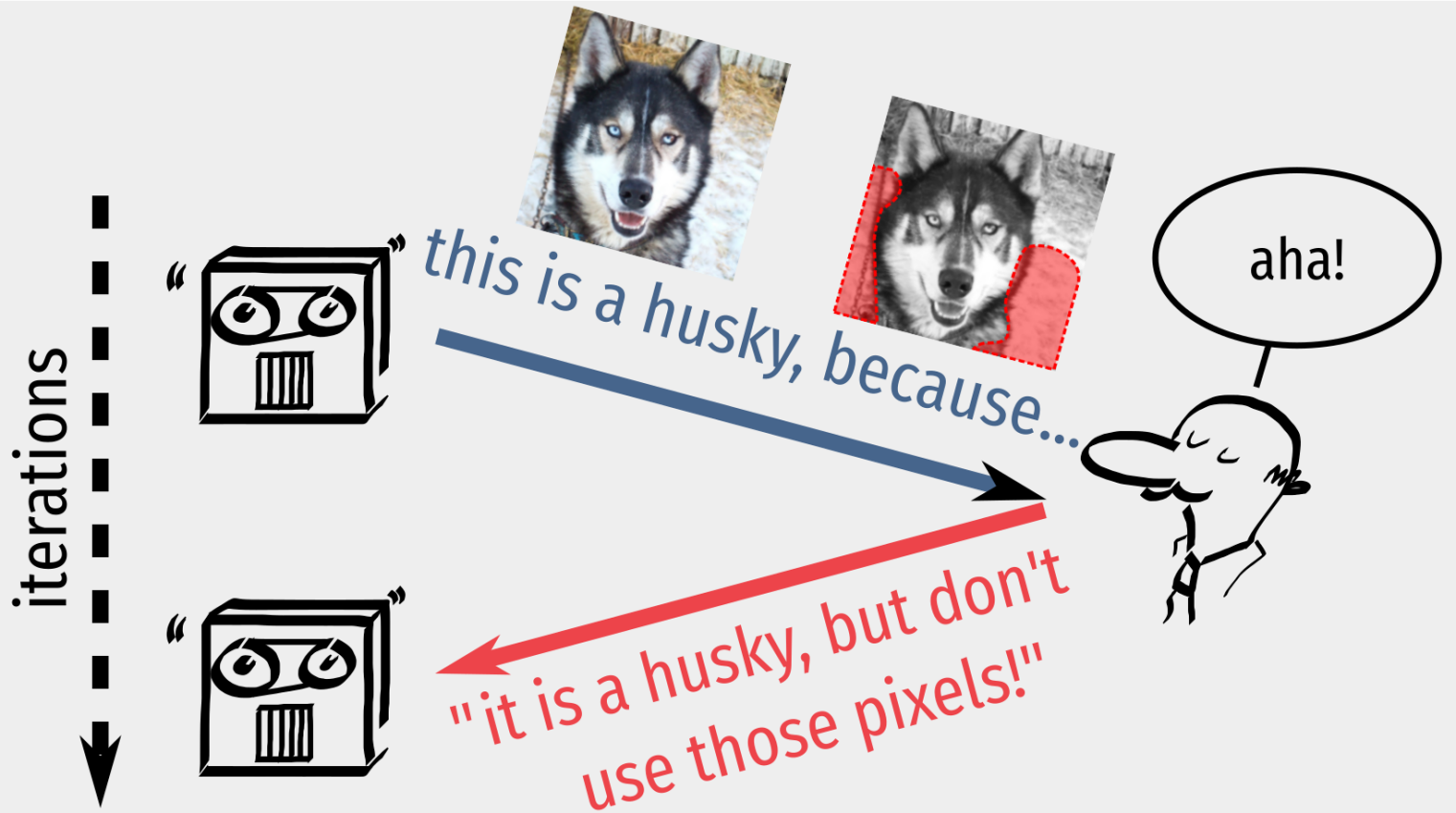
Active Learning



The user **a**) does not know the **model's beliefs**, **b**) cannot **affect** them directly, **c**) has no clue of what his **feedback does**!

Explanatory Active Learning

Explanatory Active Learning[4]



a) Explain predictions (**competence**, **understandability**),

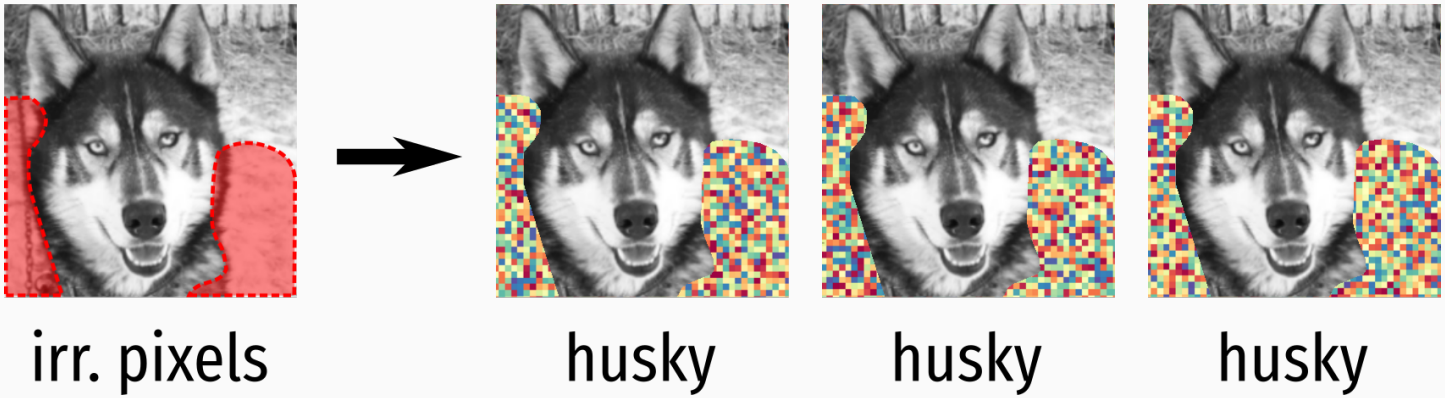
b) Allow to correct explanations (**directability**)

[4] Stefano Teso and Kristian Kersting. "Explanatory interactive machine learning". In: 2019.

Explanation Corrections

1. User's correction indicates the **false positive** segments
2. converts correction to **counterexamples**, i.e., fill in random values while keeping the same label

Example: husky predicted right for the wrong reasons

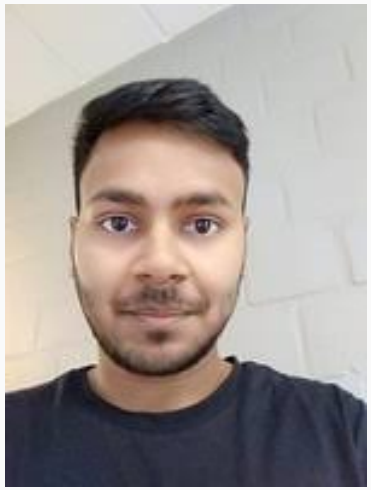


Explanatory Guided Learning and Beyond

On-going work

with **Teodora Popordanoska** and **Mohit Kumar** (KU Leuven)





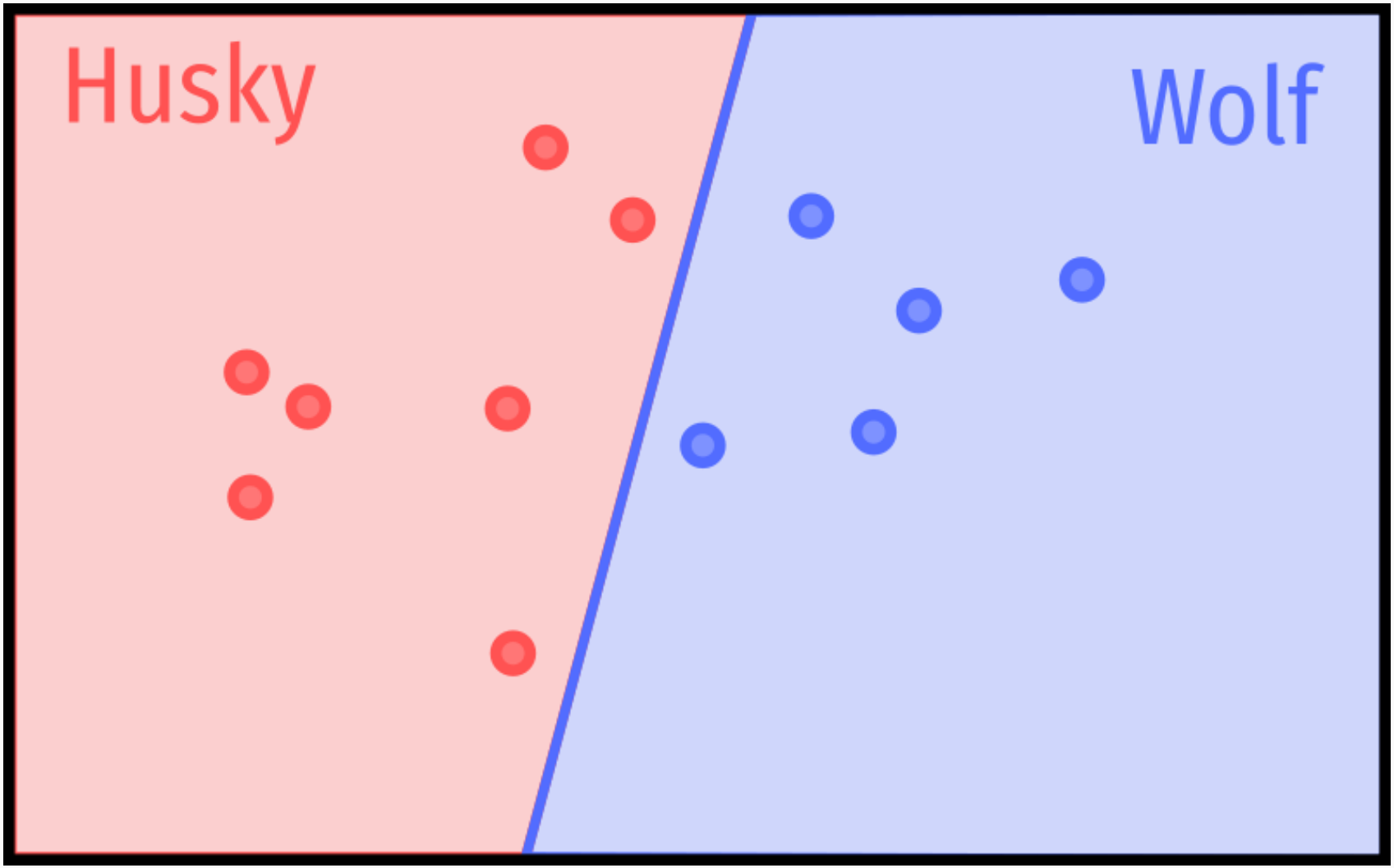
By witnessing that the model's beliefs improve over time, the human "teacher" builds trust into the "student" model

By witnessing that the model's beliefs improve over time, the human "teacher" builds trust into the "student" model

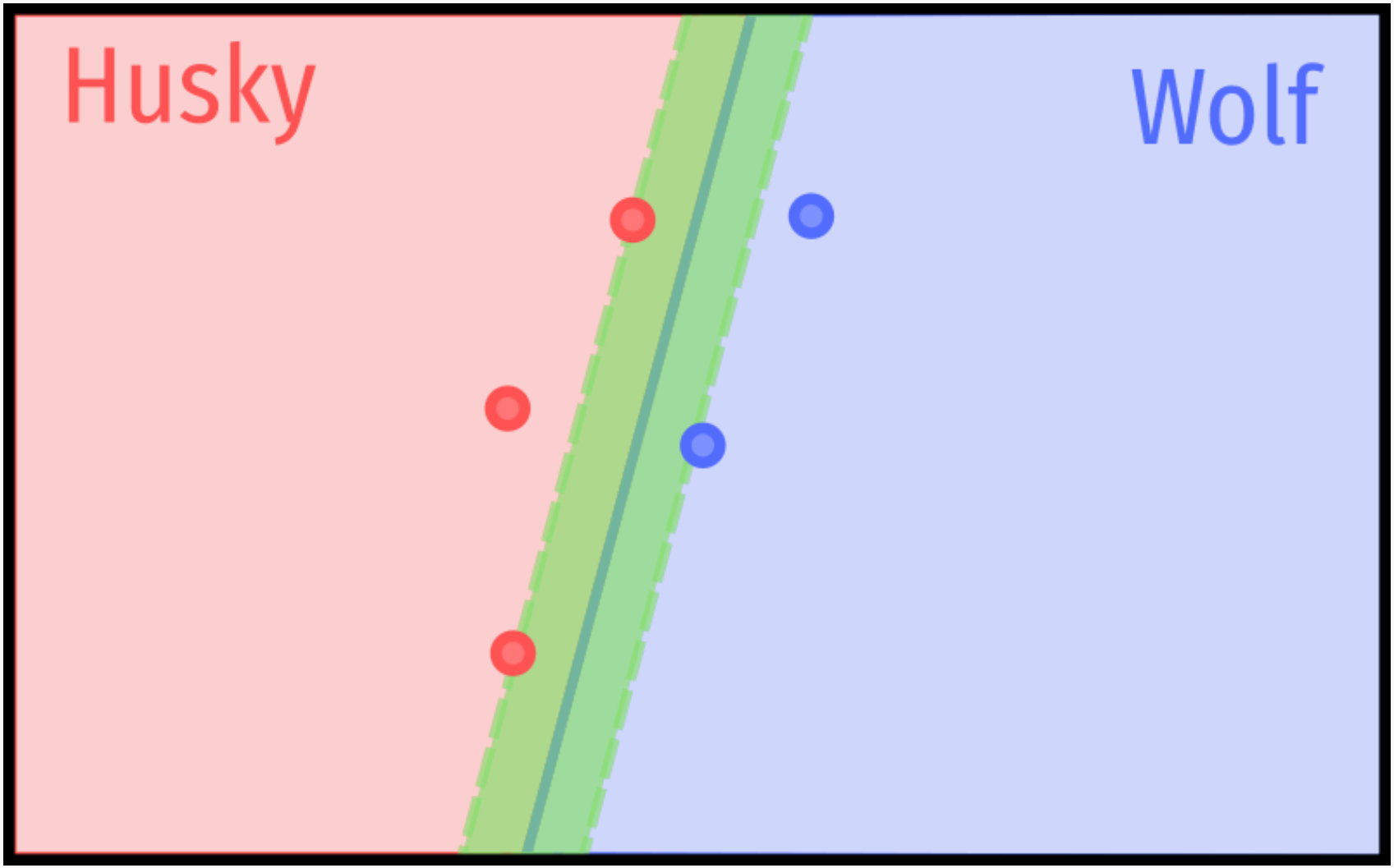
Problem: nothing prevents the machine from repeatedly choosing instances where it does well.

Not so far-fetched: the machine does not know how to choose difficult instances, think of **high-loss unknown unknowns**

Example: dog vs wolf, machine **very certain** everywhere

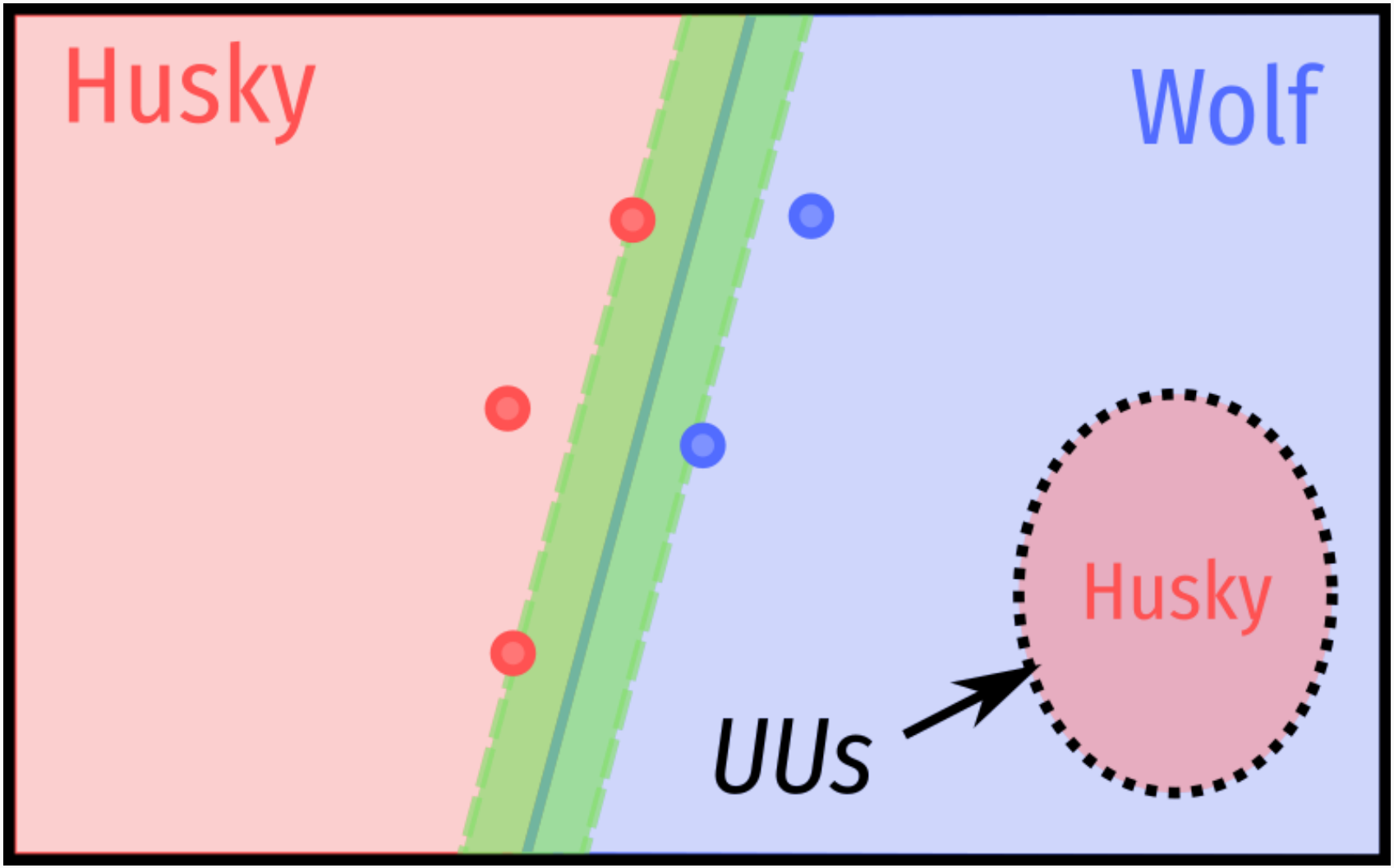


Example: dog vs wolf, machine **very certain** everywhere



What about **unknown unknowns**?

Example: dog vs wolf, machine **very certain** everywhere



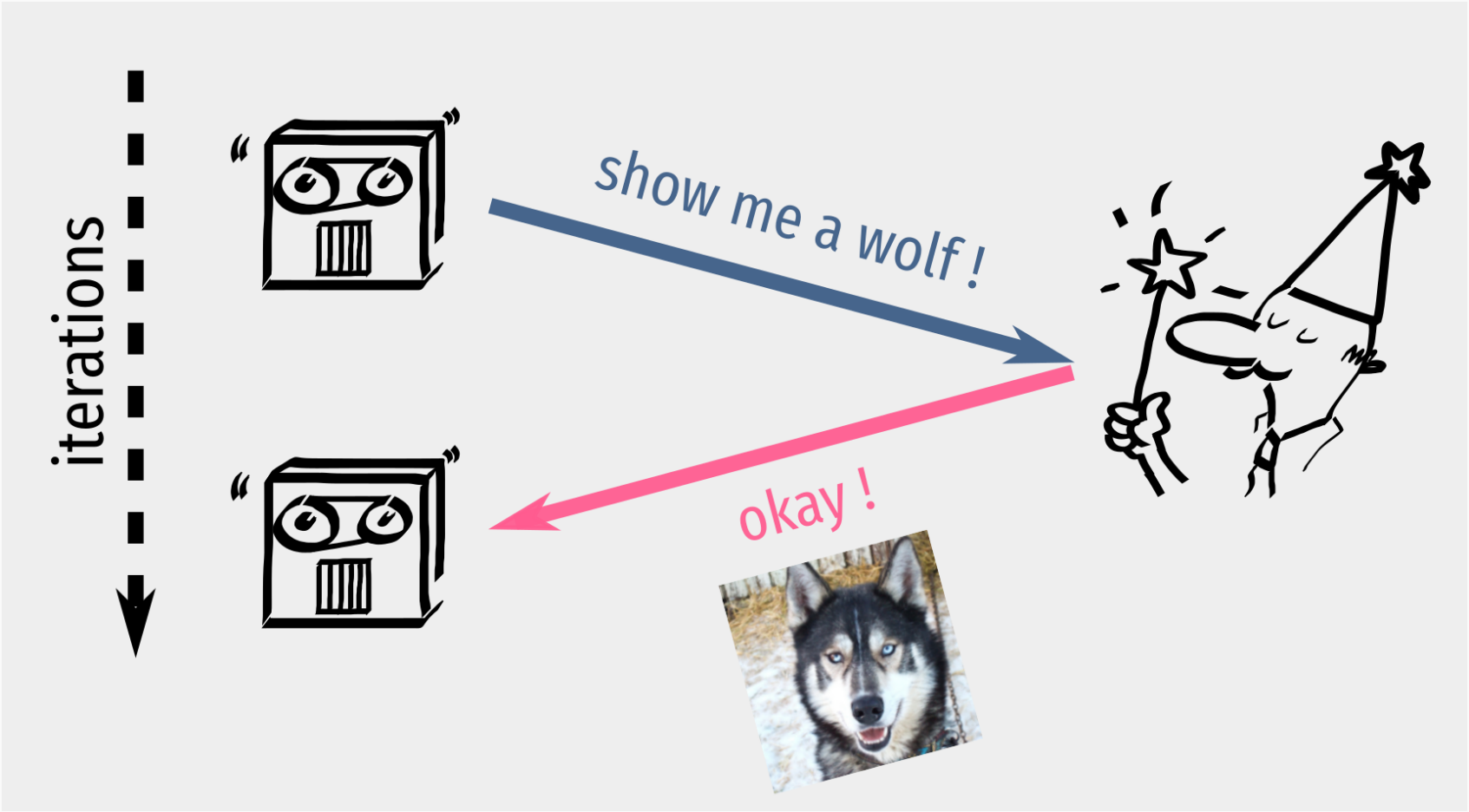
AL **doesn't** help with UUs, uncertainty may be wrong too

Idea: *let the user choose the challenging instances*

This is what professors do when testing students

Piggy-back on Guided Learning

machine chooses rare label, **user** searches for example



Useful for tackling *class unbalance* where AL fails

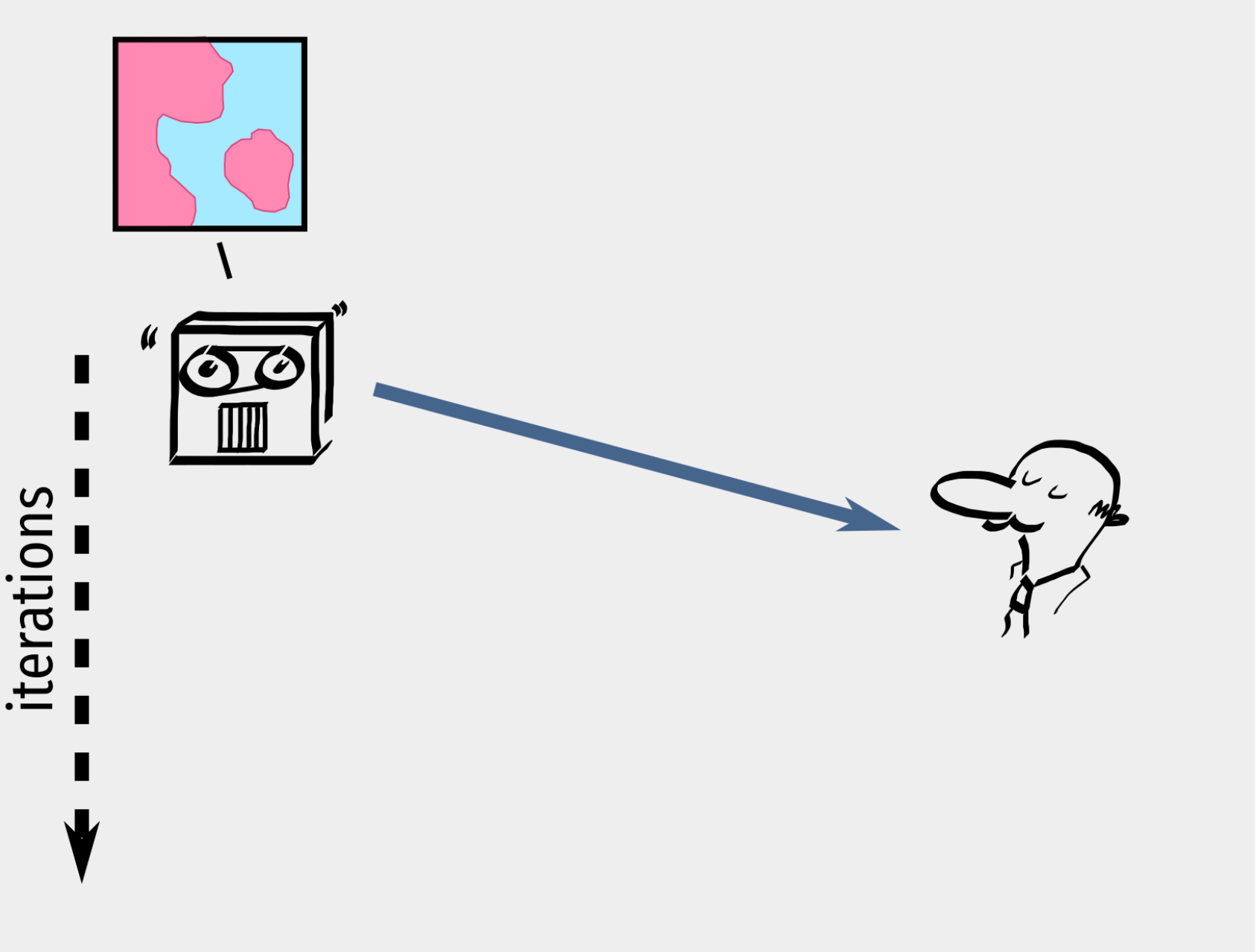
The human teacher is **blind**:

- impossible to establish justifiable **trust**
- she may provide examples that teach **nothing new** to the machine

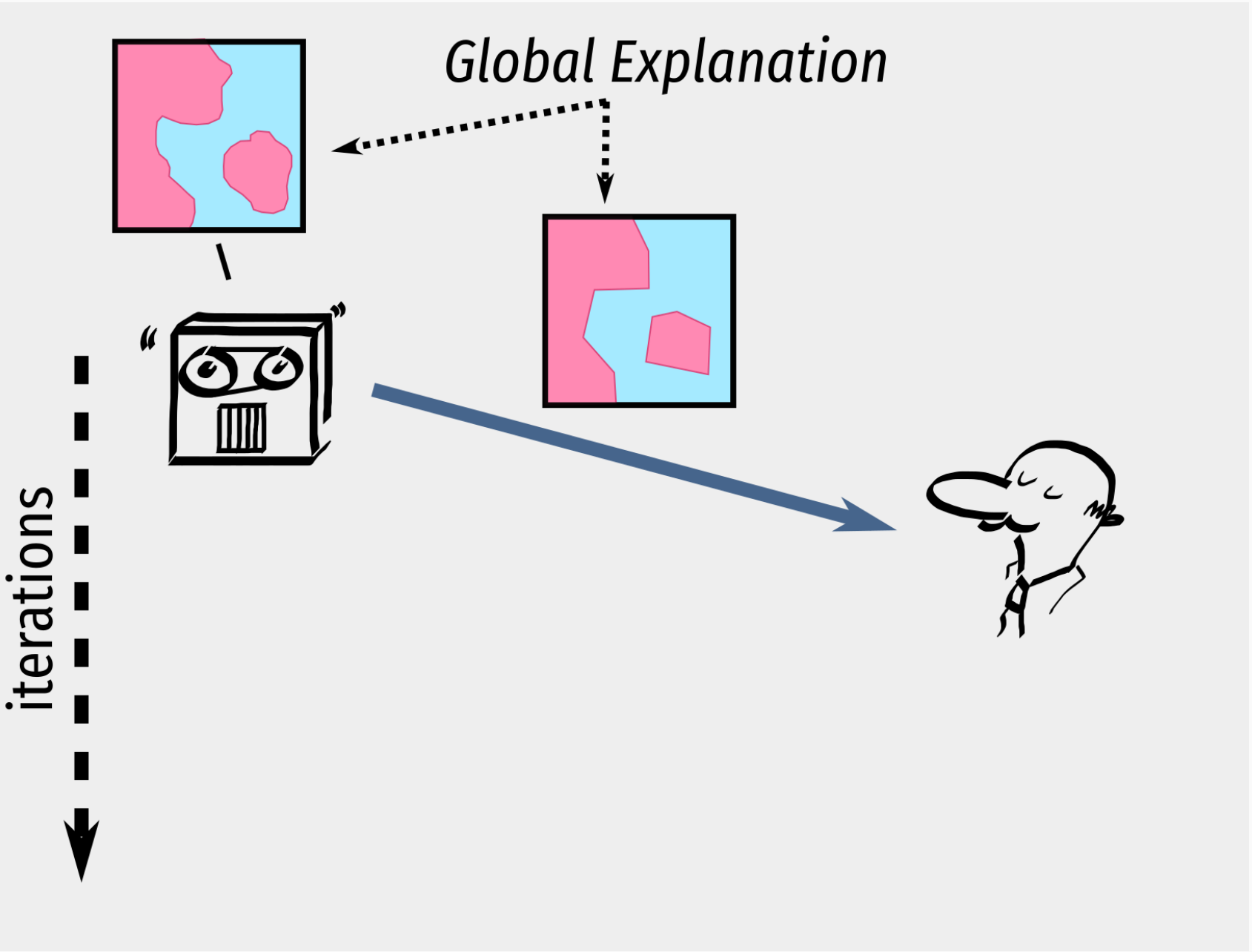
How can we expect the human to provide useful examples?[5]

[5] Interactive machine teaching with black-box models shows that blind teachers cannot do better than random teachers.

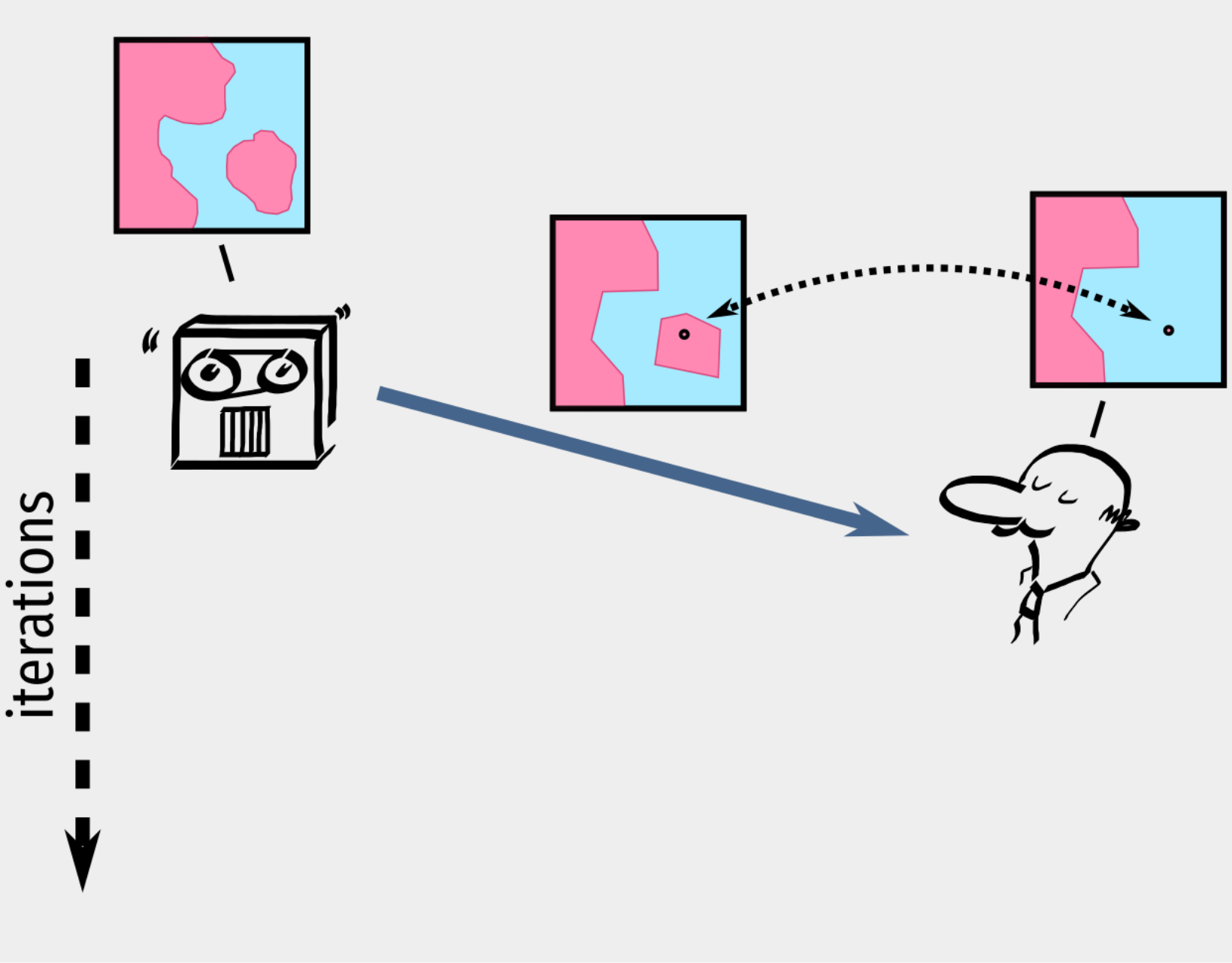
Explanatory Guided Learning



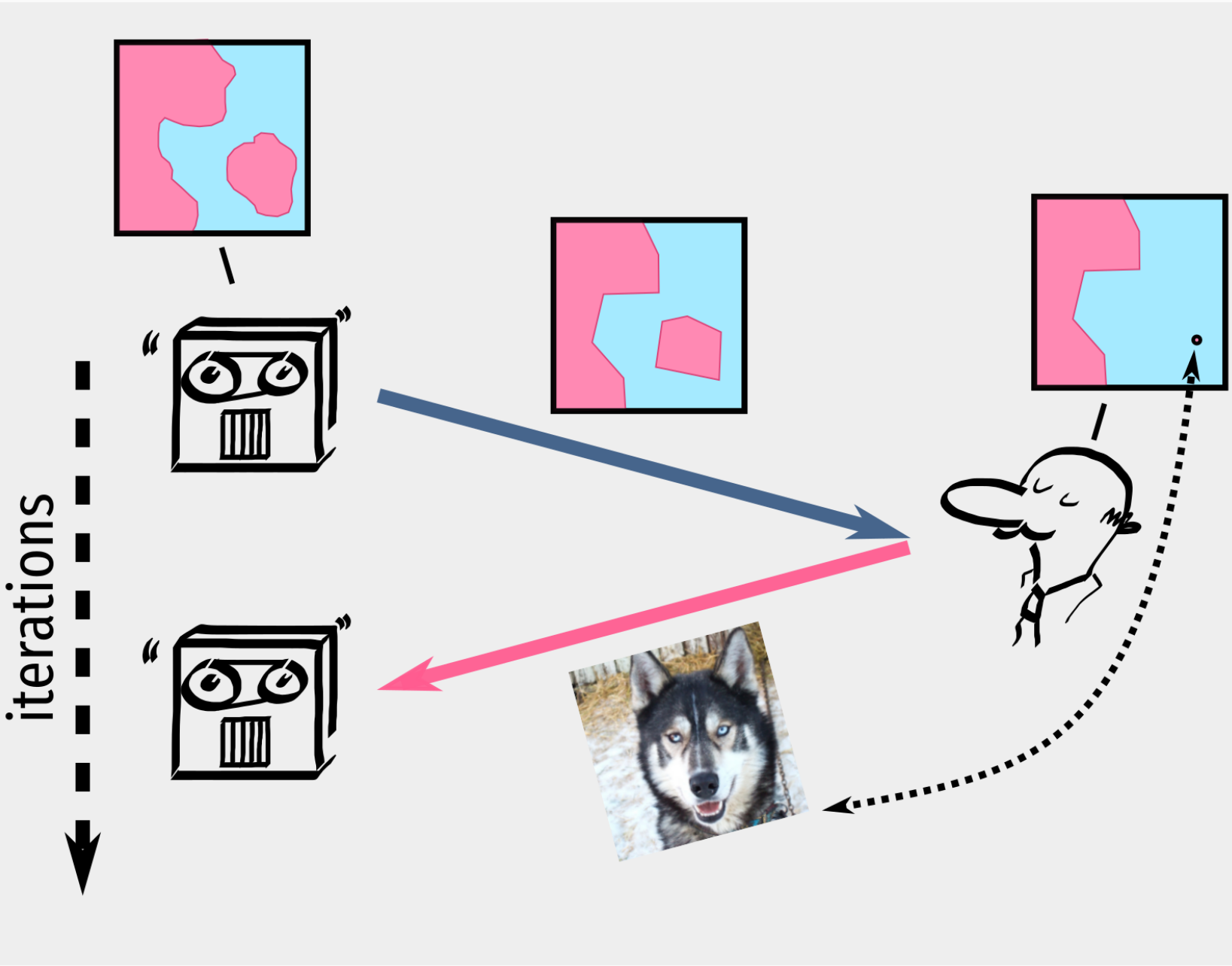
Explanatory Guided Learning



Explanatory Guided Learning



Explanatory Guided Learning

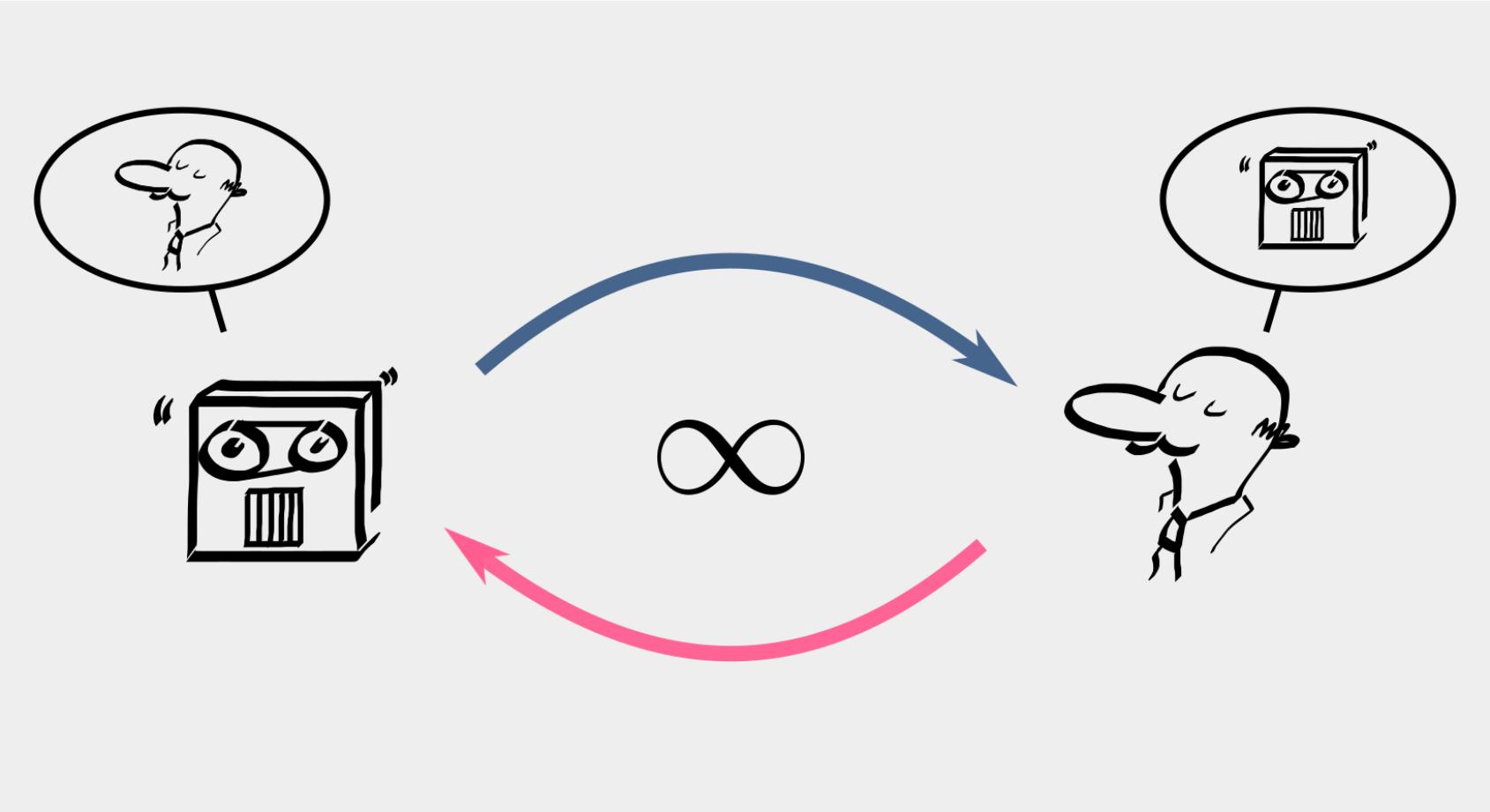


RESULTS

Plan:

1. Polish experiments with "imperfect user"
2. Case study with real users (!)
3. Hook up iterative machine teaching theory

A “Theory of Mind” for Machine-Human Teams



mutual understanding guides trust, learning, and teaching



