Towards Lifelong Machine Learning

Christoph Lampert

CHIST-ERA Workshop
June 8, 2016

Ultimate goal

Automatic systems that learn and act autonomously

Ultimate goal

Automatic systems that learn and act autonomously

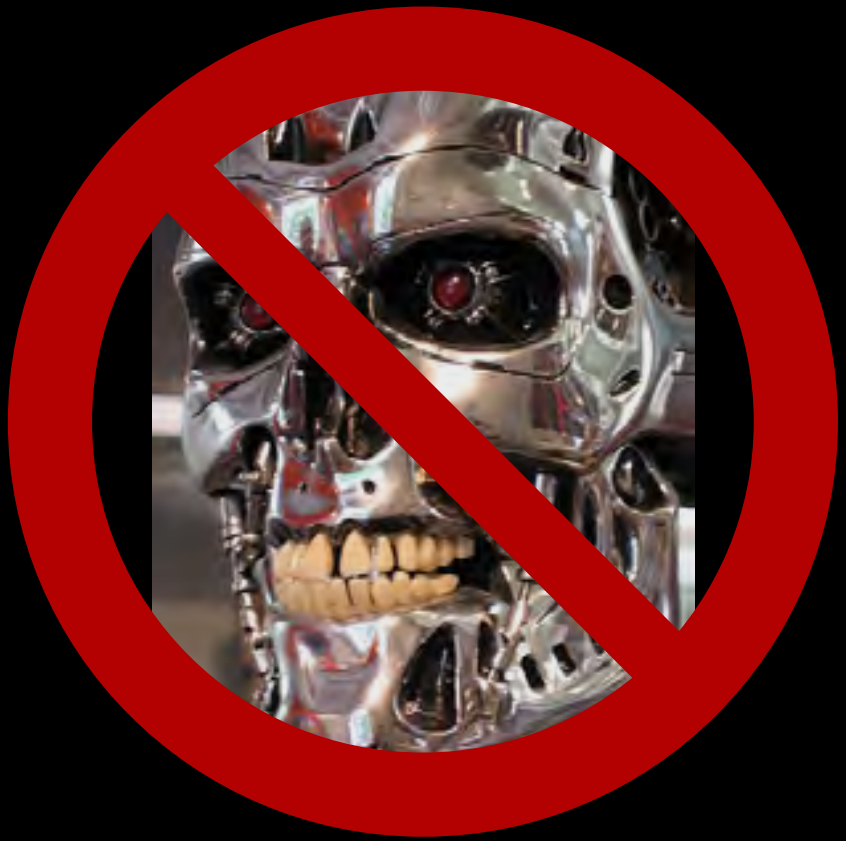


Image: "Terminator's face" by tenaciousme from Tokyo, Japan. Licensed under CC BY 2.0 via Wikimedia Commons
Image: Photo by Basilio Noris

Medium term goal

Automatic systems that can analyze and interpret data

Image Understanding

"Three men sit at a table in a pub, drinking beer. One of them talks while the other two listen."

Image: British Broadcasting Corporation (BBC)

State of the art

Analyze individual aspects: learning tasks

Scene Classification

- indoors
- in a pub

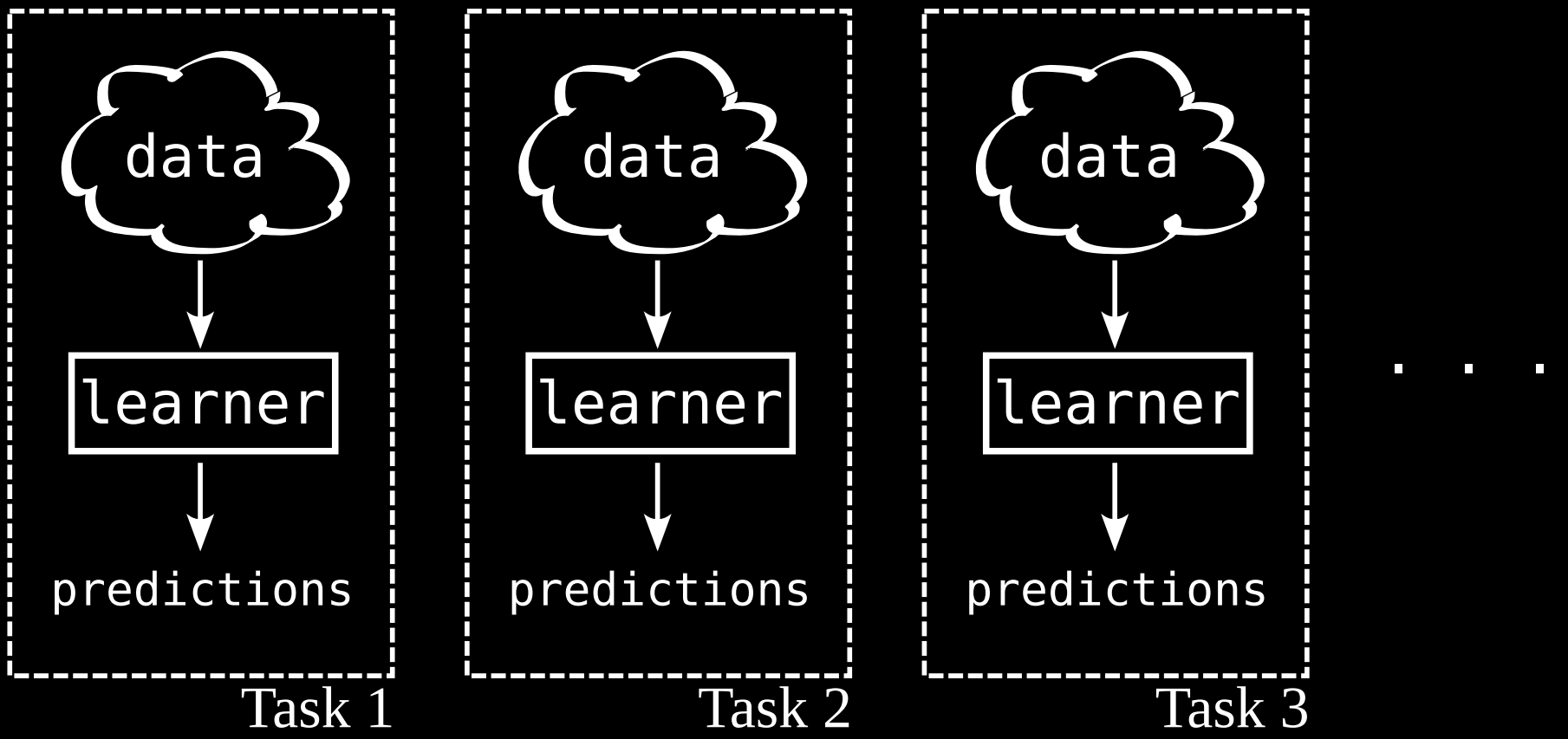
Action Classification

- drinking
- talking

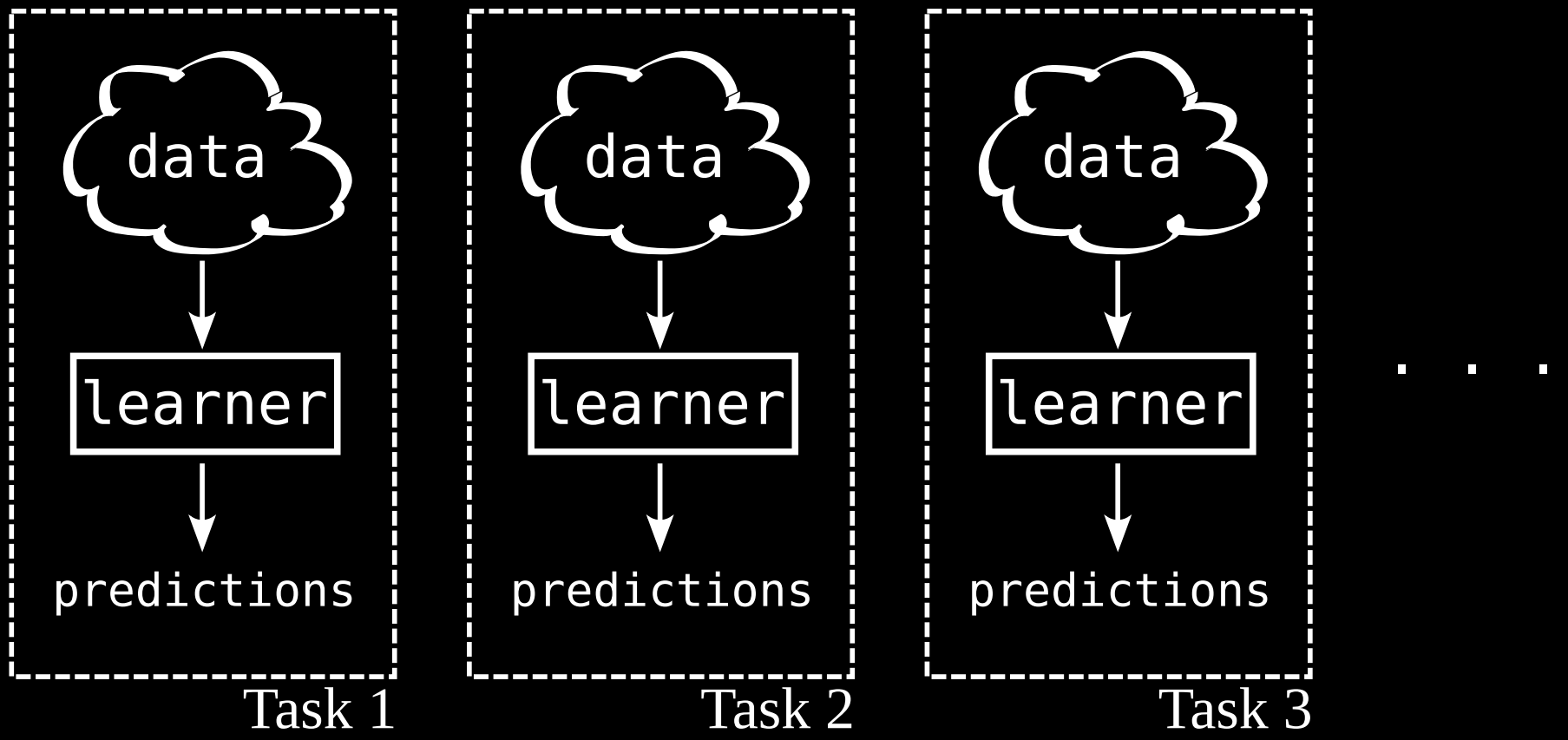
Object Recognition

- three persons
- one table
- three glasses

Learning multiple tasks

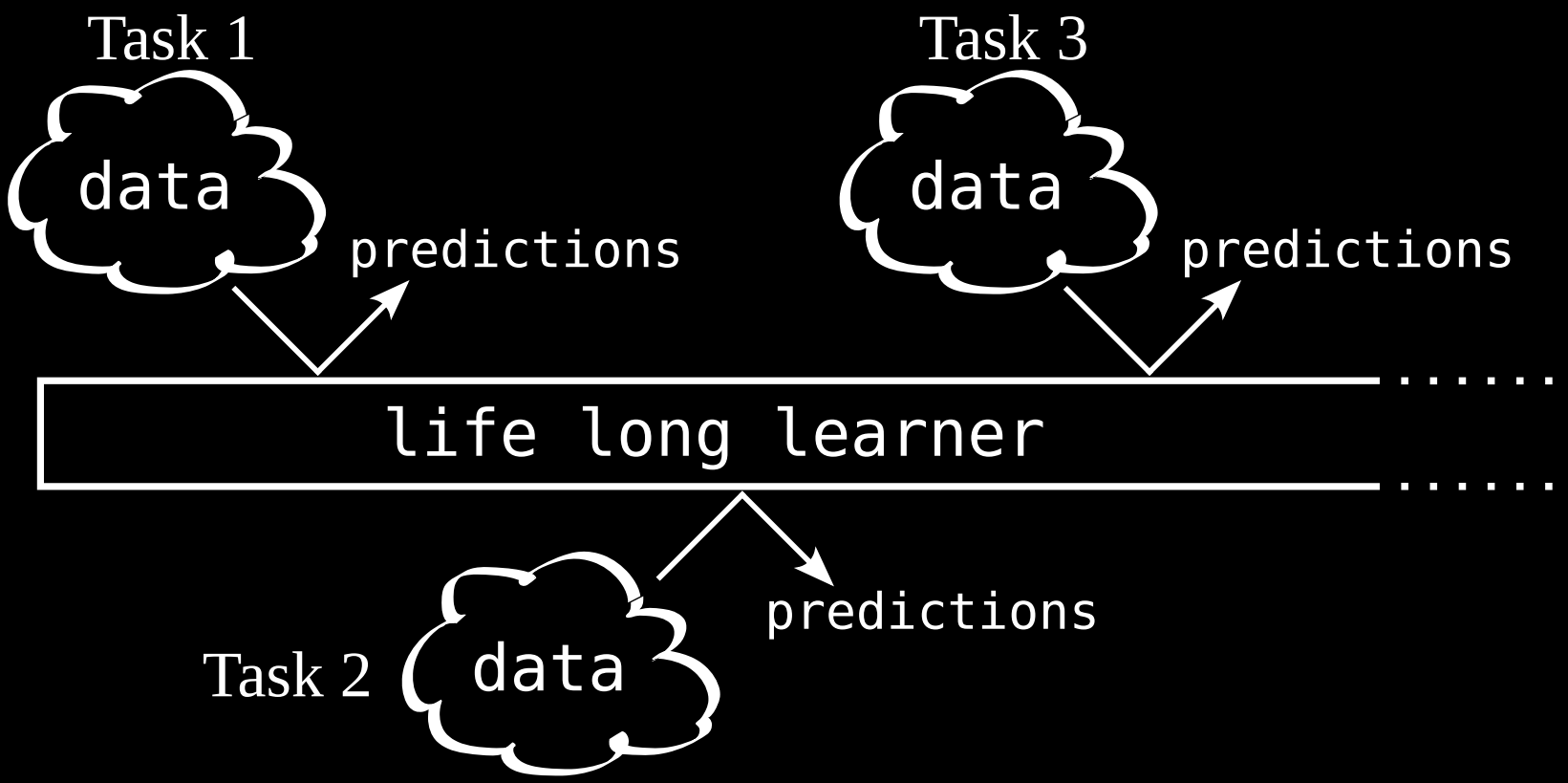


Learning multiple tasks



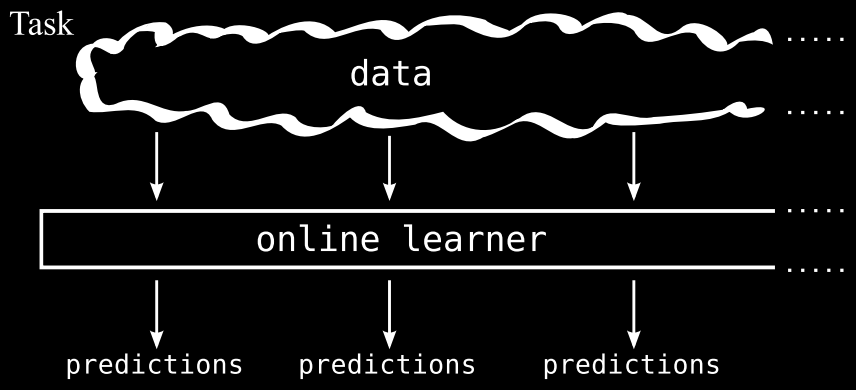
Tabula Rasa Learning

Future challenge: towards continuously improving systems



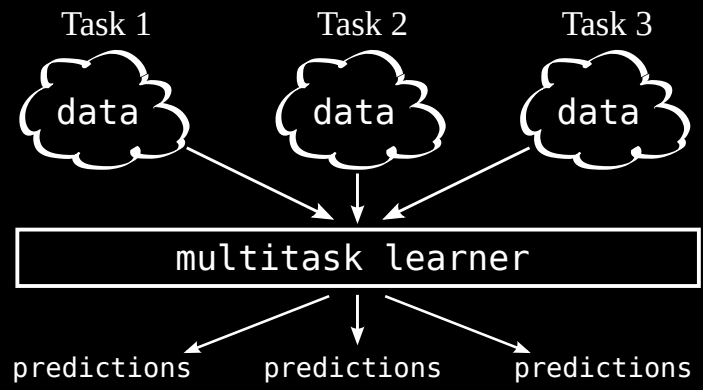
Lifelong Learning

The transfer learning universe...

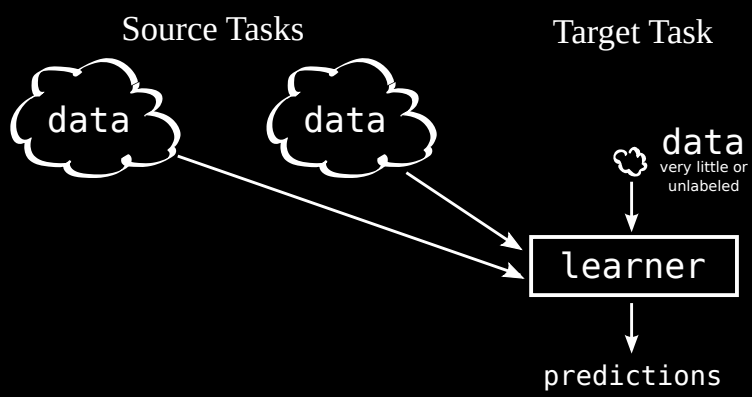


Online Learning

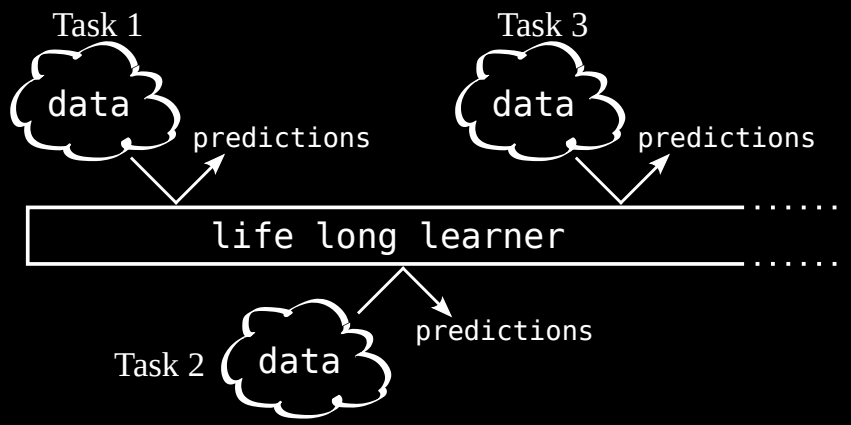
(also: Reinforcement Learning)



Multi-Task Learning



Domain Adaptation



Lifelong Learning

L3ViSU: Lifelong Learning of Visual Scene Understanding

Theory (Statistical Machine Learning)

- Multi-task learning
- Domain adaptation
- Lifelong learning (Learning to learn)

Models/Algorithms

- Zero-shot learning
- Classifier adaptation
- Weakly-supervised learning

Applications (in Computer Vision)

- Object recognition
- Object localization
- Semantic image segmentation

Towards A Theoretical Understanding of Lifelong Learning

Asya Pentina

[A. Pentina, CHL, *"A PAC-Bayesian bound for Lifelong Learning"*, ICML 2014]

[A. Pentina, CHL, *"Lifelong Learning with Non-i.i.d. Tasks"*, NIPS 2015]

[A. Pentina, R. Urner, *"Lifelong Learning with Weighted Majority Votes"*, under review]

What is lifelong learning?

Informal Explanation

- "A system should learn many tasks sequentially over time."
- "From the experience it makes over time it should get better at solving future tasks."
- "E.g., it could build a knowledge representation on-the-fly that helps it to avoid mistakes by providing common sense/context."

What is lifelong learning?

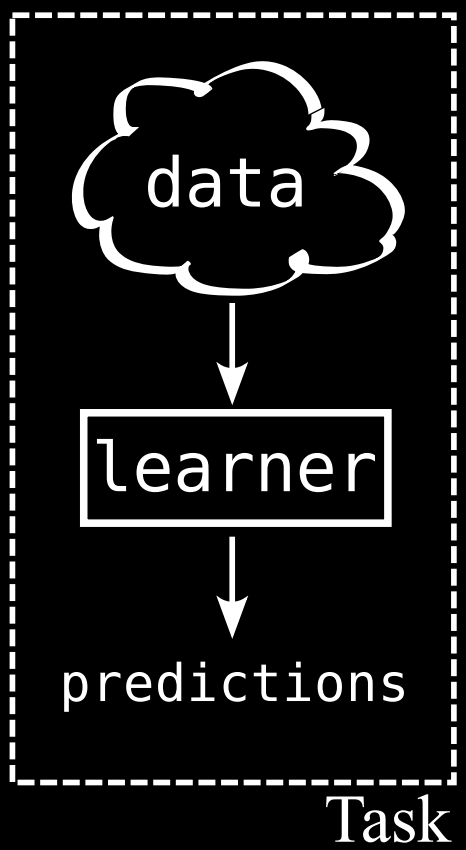
Informal Explanation

- "A system should learn many tasks sequentially over time."
- "From the experience it makes over time it should get better at solving future tasks."
- "E.g., it could build a knowledge representation on-the-fly that helps it to avoid mistakes by providing common sense/context."

Formal Definition?

- What is "learning a task"?
- What are "future tasks"?
- What is a "knowledge representation"?
- How can it help to avoid mistakes?

Learning a task



Setting: inputs $x \in \mathcal{X}$, outputs $y \in \mathcal{Y}$

E.g.: spam classification

$$\mathcal{X} = \{\textit{emails}\}, \mathcal{Y} = \{\texttt{Spam}, \texttt{Not Spam}\}$$

E.g.: automatic translation

$$\mathcal{X} = \{\textit{english texts}\}, \mathcal{Y} = \{\textit{french texts}\}$$

Data: training set with manual annotation

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$$

Learning: find a function/model/classifier, $f : \mathcal{X} \to \mathcal{Y}$, that makes good **predictions**:

$$\text{Prob}\{\, f(x) \neq y \,\} \text{ is small}$$

for future data with the same statistical distribution as the training set.

One classical lifelong learning scenario [J. Baxter. 1997]

"Future tasks"?

- there is a (large) set of all possible tasks
- observed tasks are randomly samples from this set
- future tasks will also be random samples from this set

"Knowledge representation"?

- a low-dimensional data representation that works for all tasks

How does it help?

- low dimension prevents overfitting
  $\rightarrow$ fewer training examples are needed to learn good models

Result [A. Pentina, CHL. "A PAC-Bayesian Bound for Lifelong Learning", ICML 2014]

Theorem. For any $\delta > 0$ the following inequality holds with probability at least $1 - \delta$ (over the training samples $\{S_1, \dots, S_n\}$) for all hyperposterior distributions $\mathcal{Q}$

$$\text{er}(M) \leq \widehat{\text{er}}(M) + \frac{1}{2\sigma n \sqrt{\bar{m}}} \sum_{i=1}^{n} \mathop{\mathbf{E}}_{B \sim D(I_k, M)} \|w_i(B)\|^2 + \text{const}$$

where $w_i(B) = \frac{C}{m_i} \left(I_k + \frac{C}{m_i} B^\top X_i X_i^\top B \right)^{-1} B^\top X_i Y_i$.

In words:
"Look for a low-dimensional representation that leads to small training errors and short weight vectors on the observed tasks."

$\rightarrow$ principled learning algorithm with generalization guarantees

Attribute-Based Classification

Stefan Harmeling
U Düsseldorf

Hannes Nickisch
Philips Research

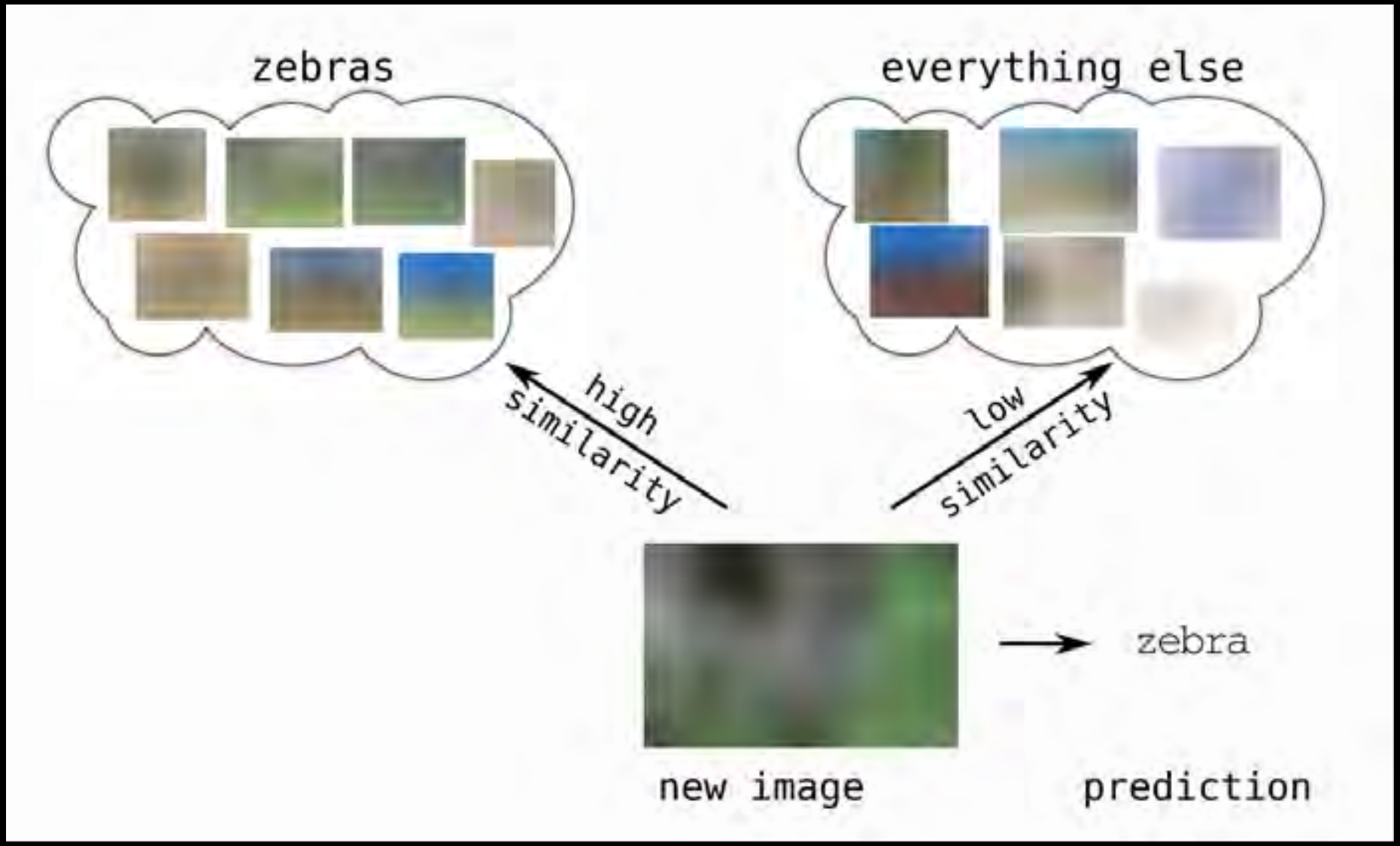
Viktoriia Sharmanska
U Sussex

[CHL, H. Nickisch, S. Harmeling. *"Learning to detect unseen object classes by between-class attribute transfer"*, CVPR 2009]

[CHL, H. Nickisch, S. Harmeling. *"Attribute-Based Classification for Zero-Shot Visual Object Categorization"*, T-PAMI 2014]

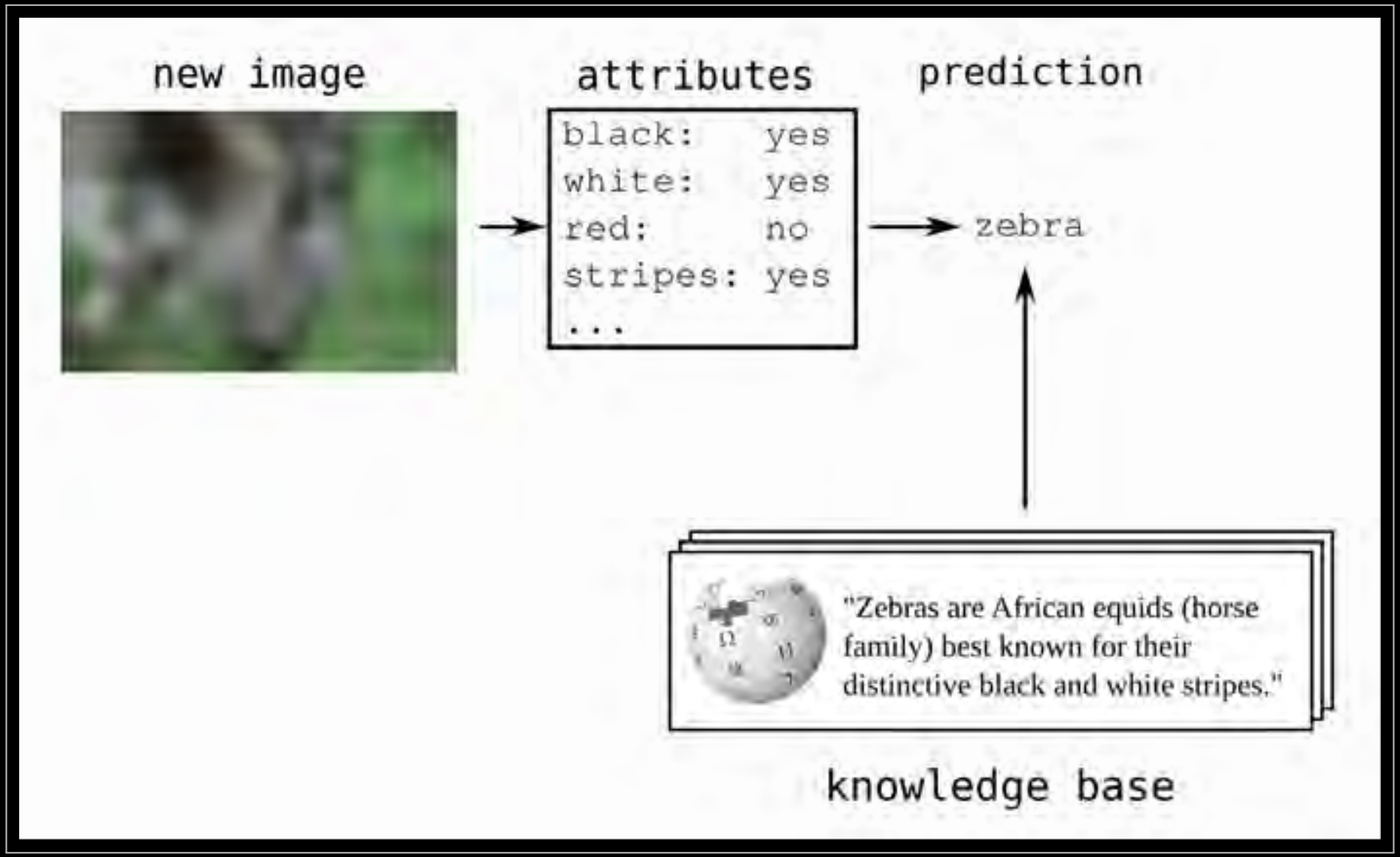
[V. Sharmanska, CHL. *"Augmented attribute representations"*, ECCV 2014]

Object recognition



(blurred for copyright reasons, sorry)

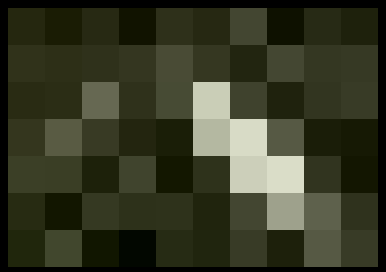
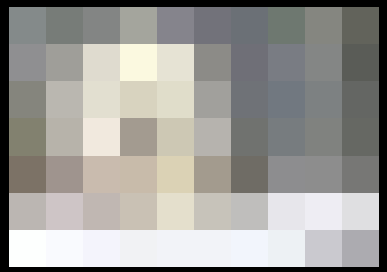
Attribute-based classification



(blurred for copyright reasons, sorry)

Component 1) attribute predictors

Can be learned from earlier classes/tasks:

dove	persian cat	polar bear	horse
			
white **not domestic** **not fluffy**	**white** **domestic** **fluffy**	**white** **not domestic** **fluffy**	**not white** **domestic** **not fluffy**

(blurred for copyright reasons, sorry)

attribute	**positive examples**	**negative examples**
white	{dove} ∪ {cat} ∪ {polar bear}	{horse}
domestic	{cat} ∪ {horse}	{dove} ∪ {polar bear}
fluffy	{cat} ∪ {polar bear}	{dove} ∪ {horse}

Component 2) knowledge base

[Osherson, Stern, Wilkie, Stob, Smith. *"Default probability"*. Cognitive Science, 15(2), 1991]

Results: Animals with Attributes dataset

Recognition of object classes without any training examples:

(blurred for copyright reasons, sorry)

Challenges and Opportunities

Challenges

Challenges

`We need proper foundations.`

- If we don't know what we want, we cannot achieve it.
- In order to communicate, one needs a common language. Concepts must have well-defined names and properties.

Challenges

`We need proper foundations.`

- If we don't know what we want, we cannot achieve it.
- In order to communicate, one needs a common language. Concepts must have well-defined names and properties.

`We need standards and benchmarks.`

- For ideas/algorithms to compete with each other, one must be able to compare them quantitatively.

Challenges

`We need proper foundations.`

- If we don't know what we want, we cannot achieve it.
- In order to communicate, one needs a common language. Concepts must have well-defined names and properties.

`We need standards and benchmarks.`

- For ideas/algorithms to compete with each other, one must be able to compare them quantitatively.

`We need real world applications/data.`

- Teaching a computer to recognize zebras is fun, but purely academic. Which concrete problems of the world would lifelong learning solve better than other approaches?

Opportunities

Opportunities

Go for the messes - that's where the action is.

Stephen Weinberg – Four golden lessons (2003)