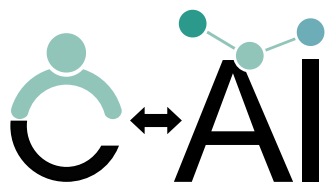
The Third Wave of Artificial Intelligence

From Blackbox Machine Learning to Explanation-Based Cooperation

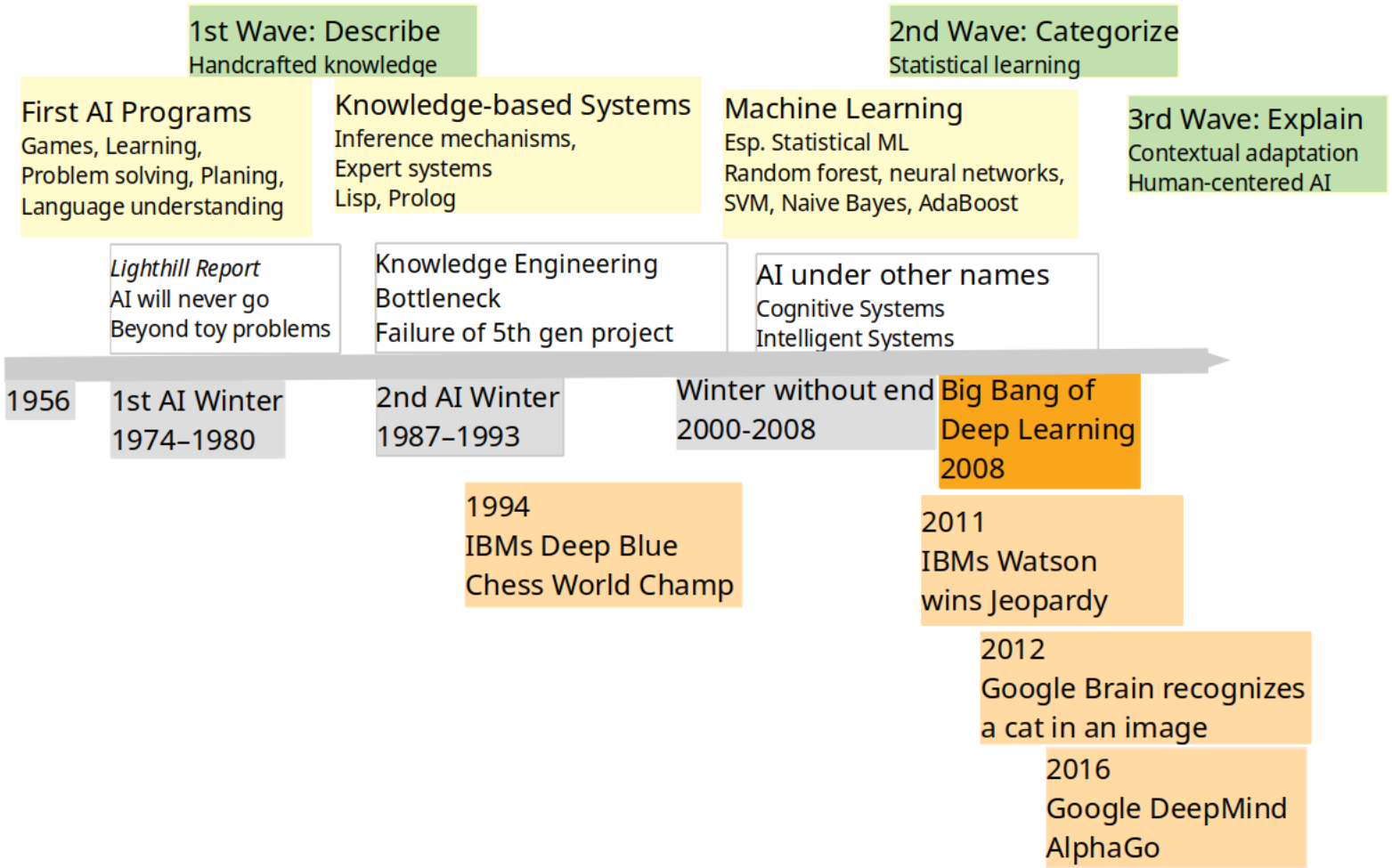
Ute Schmid

Cognitive Systems Group
Otto-Friedrich Universität Bamberg
Fraunhofer IIS Project Group Comprehensible AI
www.uni-bamberg.de/cogsys www.iis.fraunhofer.de/explainable-AI



Women in Data Science, Regensburg, April 13-14 2021

A Very Short History of AI

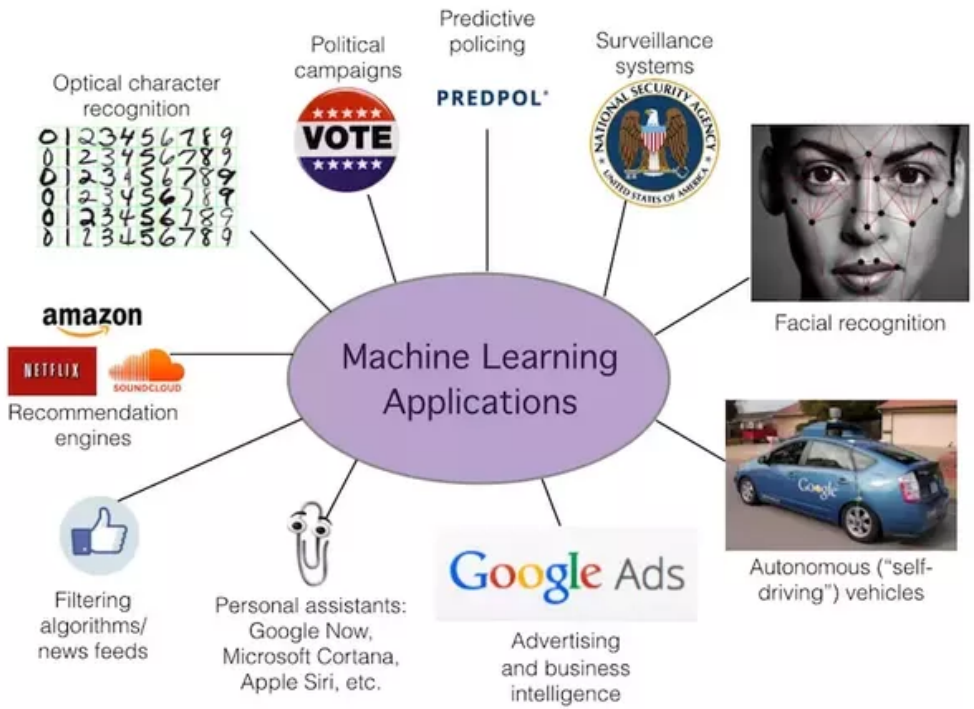


winter is coming?

AI from Lab to Application

We (AI researchers) never asked ourselves "What if it really works?"

Stuart Russell, 2019

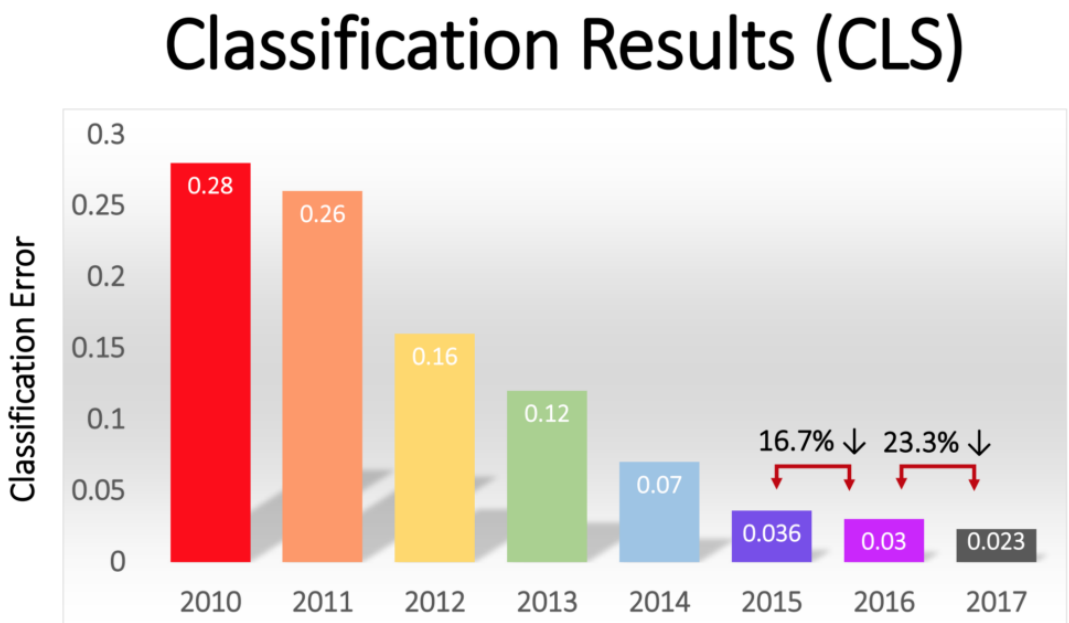


https://goo.gl/images/kEjk48

Who can comprehend and control decisions of an AI system (a learned model), if it learns and acts autonomously?

The Hype of Deep Learning

- End-to-end learning
  2012 a Convolutional Neural Network (AlexNet) wins the ImageNet Large Scale Visual Recognition Challenge (ILSVRC, $>$ 1 Mio images, 1000 categories) with significant improvement compared to classical computer visions methods
- Learning directly from raw data, without the need of previous feature extraction
- **Hope: Replace the need to think by sampling data**



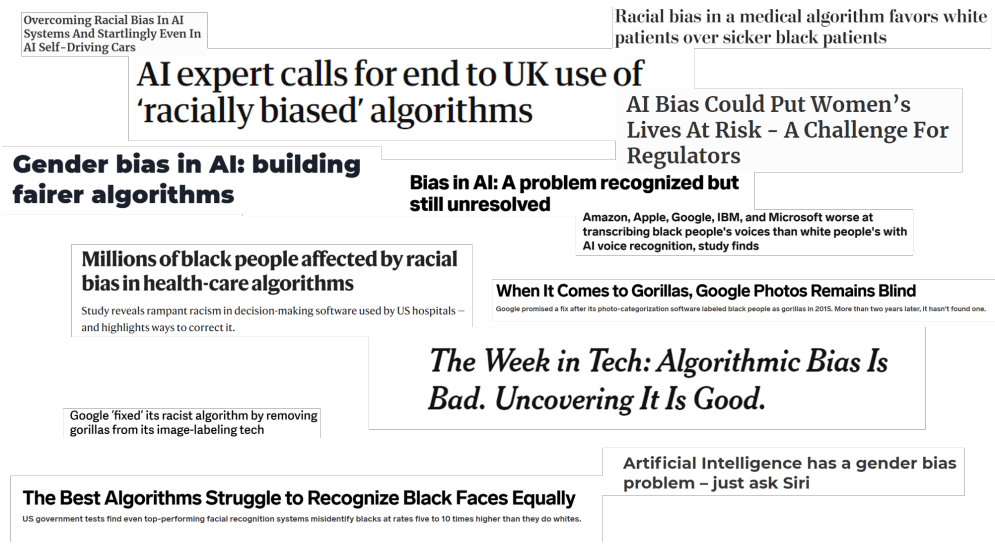
Problems with purely data-driven machine learning

- **Problem 1: Data Quality**
  - In many domains there are **not enough data available** to train deep learning models (in the intended manner)
  - Supervised machine learning presupposes that training data are labelled/annotated with ground truth (for general categories such as animals or traffic signs this might be done with suitable reliability by crowd sourcing – but what about expert domains such as medical diagnostics or quality control in production?) $\hookrightarrow$ **labeling can be as expensive as knowledge engineering**!
  - Distribution of data in the training set should be similar to the true distribution, however: many ML approaches are sensitive to imbalanced data: resulting in **sampling biases and unfair models**

Problems with purely data-driven machine learning

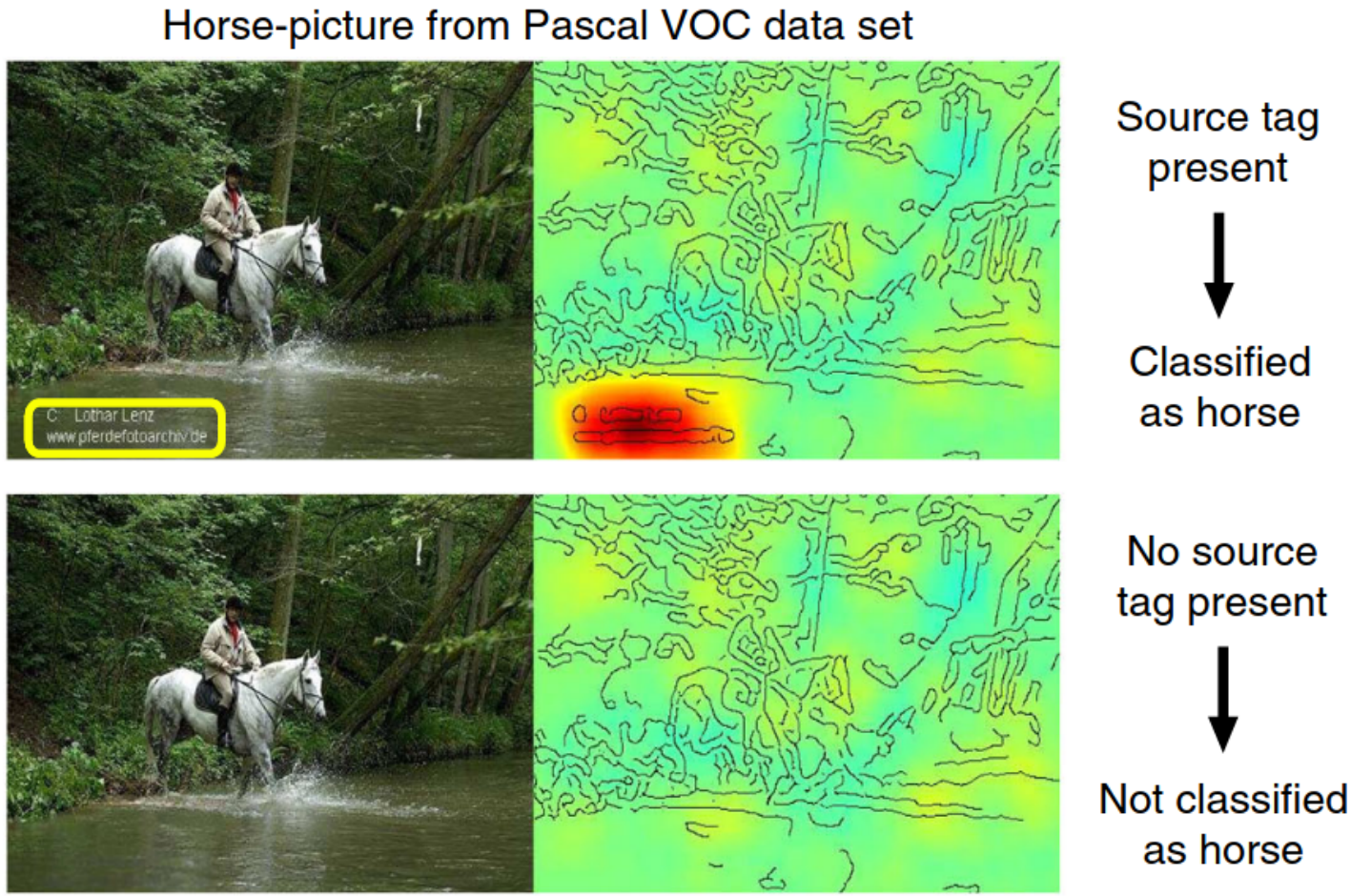
Be aware:

- there is no bias-free learning (without induction bias no learning, neither in machines, nor in humans!)
  stereotypes and predjudice are the downside of human's eager generalization learning
- learned models cannot be error-free (a modell with 99% predictive accuracy committs an error for one in 100 cases – miss oder false alarm)



https://miro.medium.com

Unmasking Clever Hans Predictors



(Lapuschkin et al., 2019, LRP)

The next AI-Winter: Data Engineering Bottleneck

Nuremberg Funnel, 1910; `https://de.wikipedia.org/`

What we can learn from human learning

- Learning has implicit and explicit components
- Learning is lifelong – incremental, correctable (not batch and static)
- Learning from few examples should be possible
- What one already knows must not be learned!
  ↪ Hybrid approaches: Combine Learning an Reasoning, see e.g. International Joint Conference on Learning & Reasoning (IJCLR 2021)

Learning from very few examples

(Josh Tenenbaum)

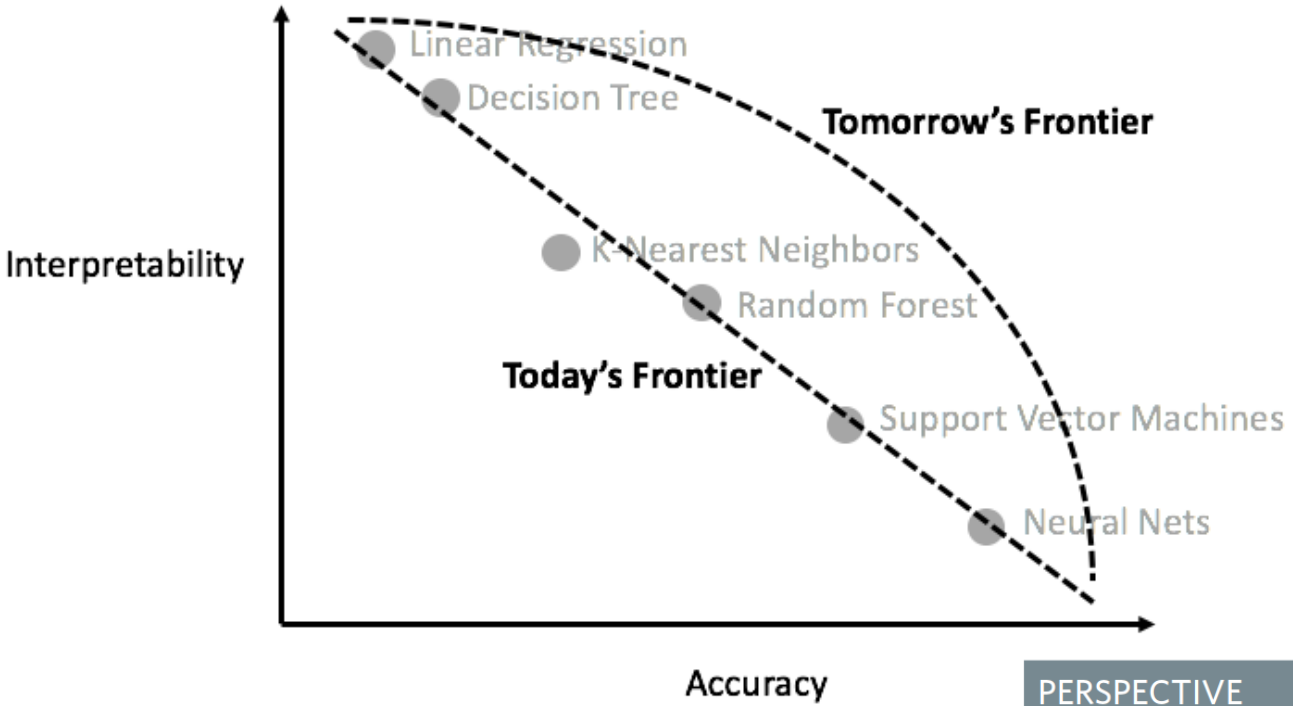
Problems with purely data-driven machine learning

- **Problem 2: Comprehensibility, transparency and control**
  - At least the **developers** should be able to look in the blackbox
    - ⋆ Recognize overfitting (Unmasking Clever Hans predictors, Lapuschkin et al., nature comm 2019)
    - ⋆ Evaluate functional safety of (embedded/autonomous) systems which inlucde AI components (A Survey on Methods for the Safety Assurance of Machine Learning Based Systems, Schwalbe & Schels, ERTS 2020)
  - **Domain experts** (e.g. a medical doctor or a quality engineer) must be able to understand system decisions (from what information what inferences have been drawn) and **be able to override and correct** system decisions
  - Right to transparency should be possible to implement in a broad area of applications (finance, marketing, health, etc.) for **end-users**

Problems with purely data-driven machine learning

- **Problem 2: Comprehensibility, transparency and control**
  - ... for different stakeholders: devops, domain experts, end-users
  - Comprehensibility (understandability/explainability) should be given at least for critical domains $\hookrightarrow$ 3rd wave – XAI
    but: beware of wrong explanations and unjustified trust!
  - Combine XAI and interactive learning to keep humans in the loop
    enhancement instead of decrease of human competence, human feedback to correct system decisions

Predictive Accuracy & Comprehensibility of Models/Decisions

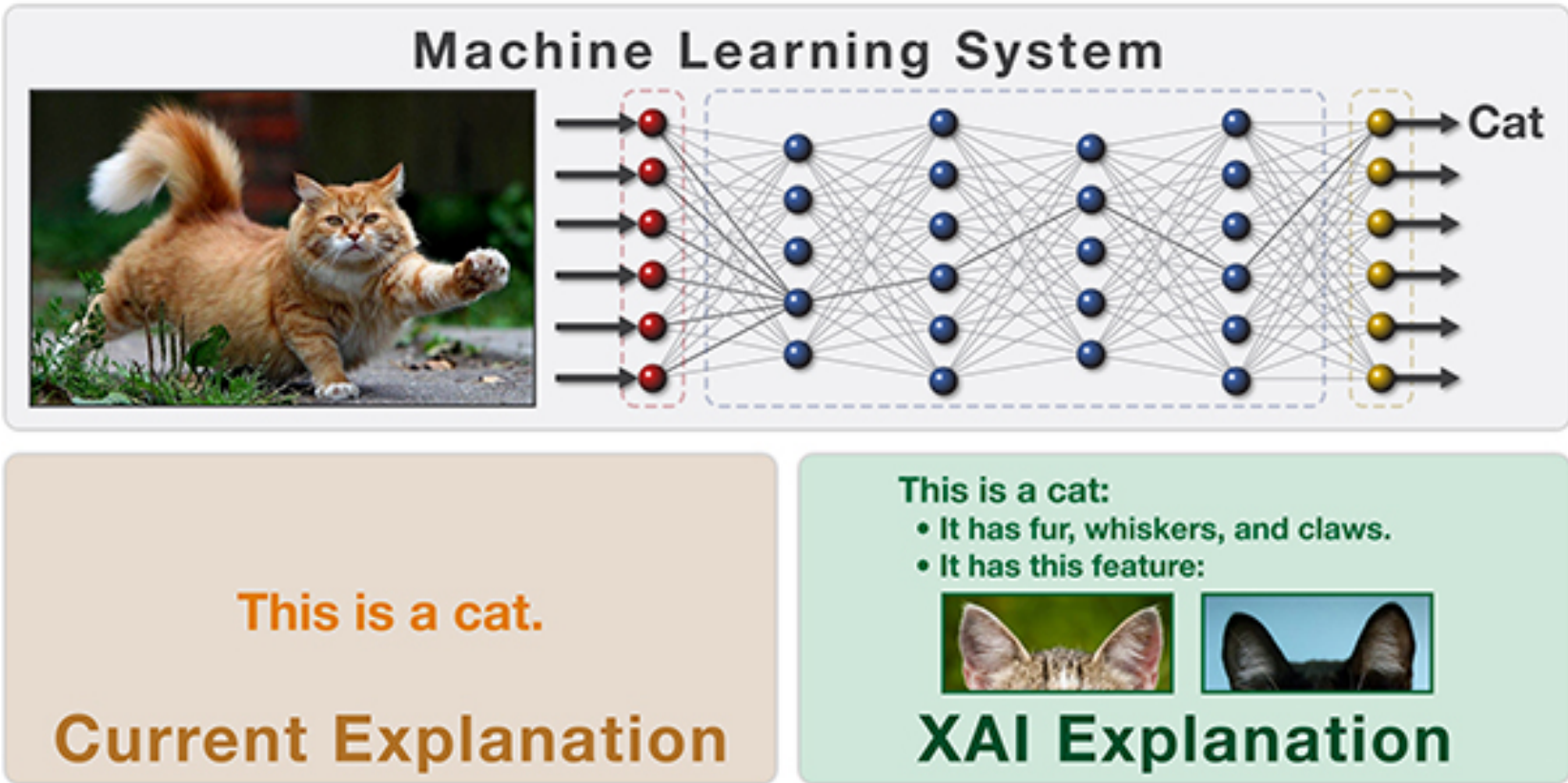


PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

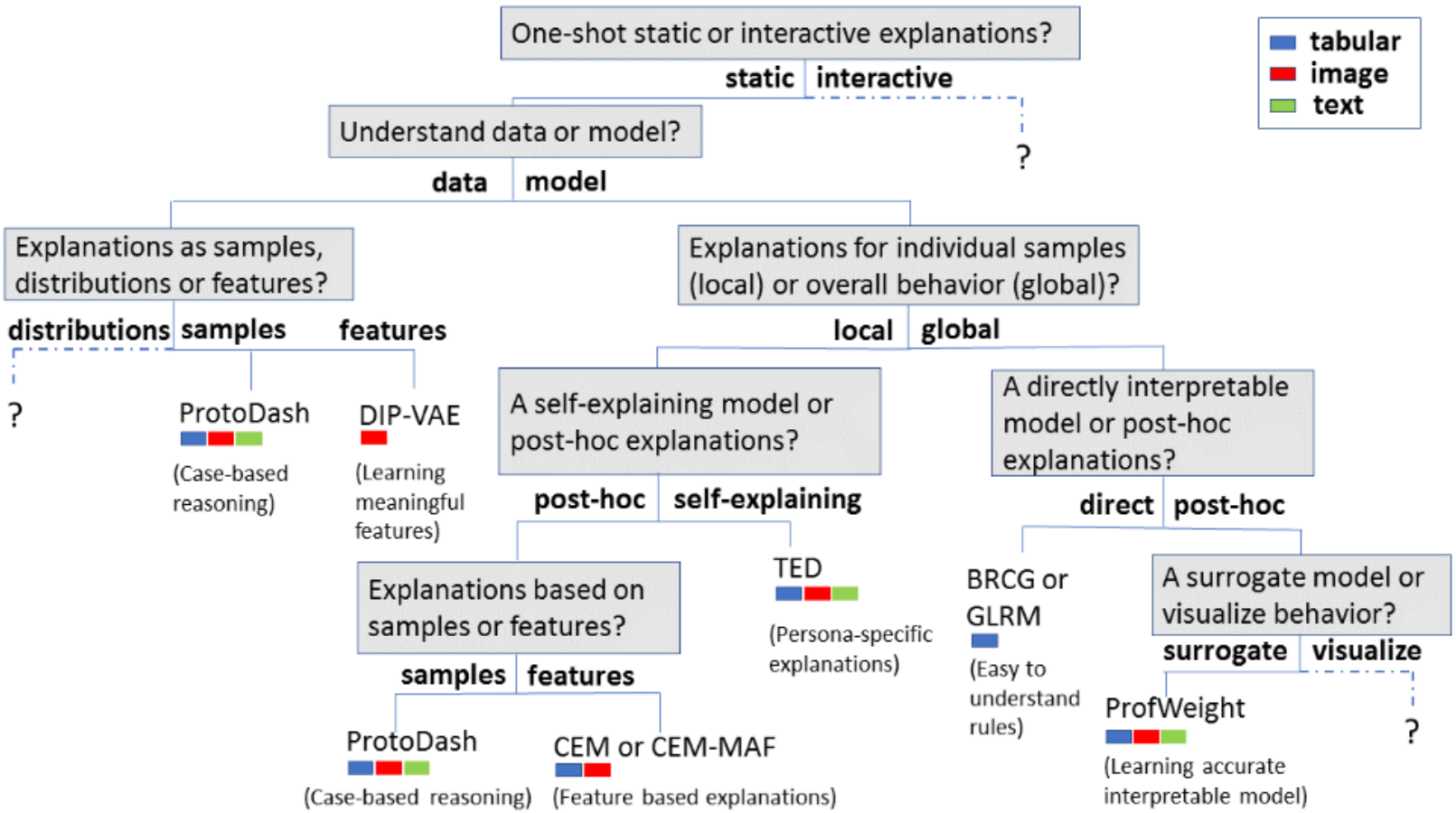
Cynthia Rudin

Explainable Machine Learning

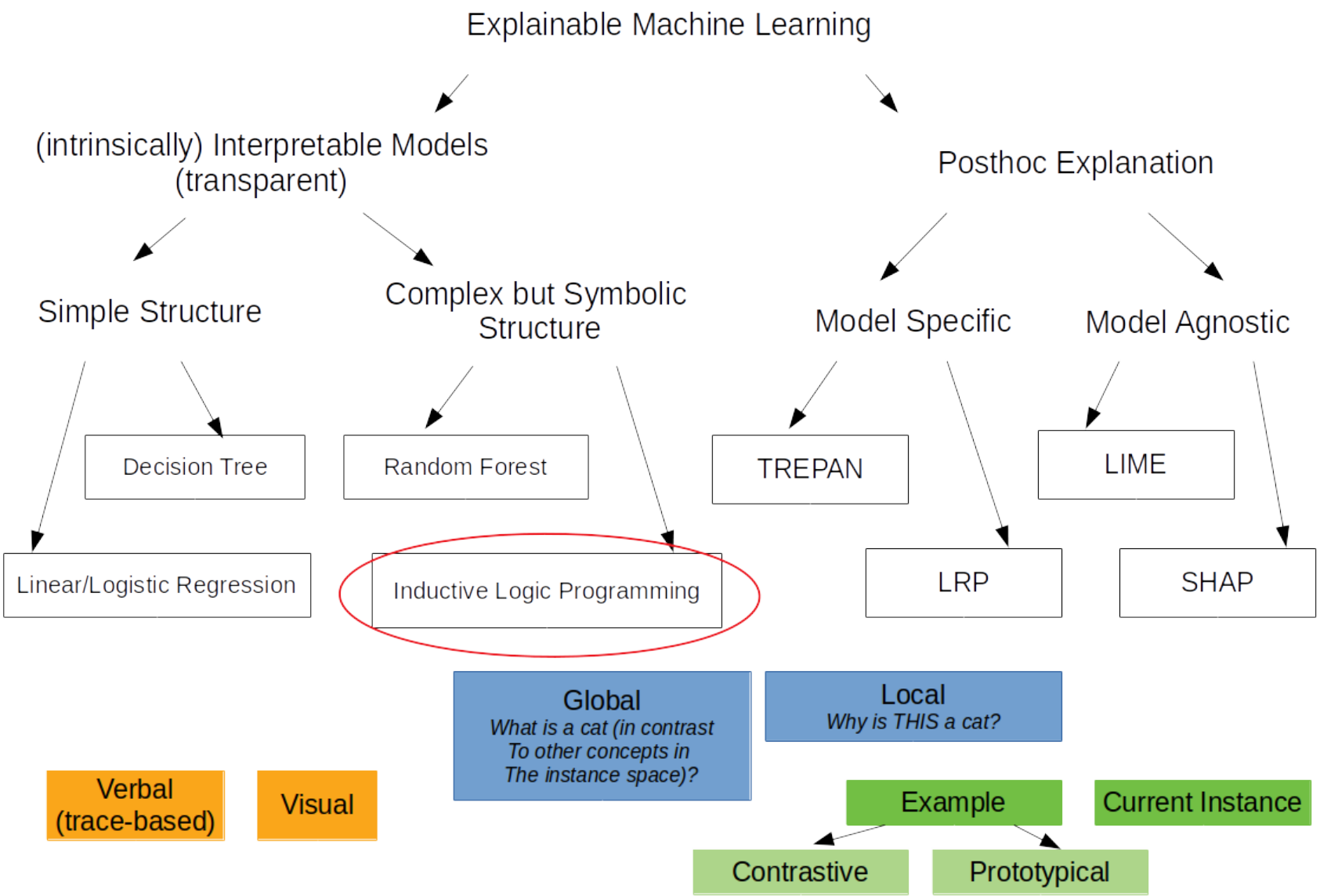


http://www.darpa.mil/program/explainable-artificial-intelligence David Gunning, IJCAI 2016

IBM – AI Explainability 360 Toolkit



Explainable/Explanatory Machine Learning (X-ML)



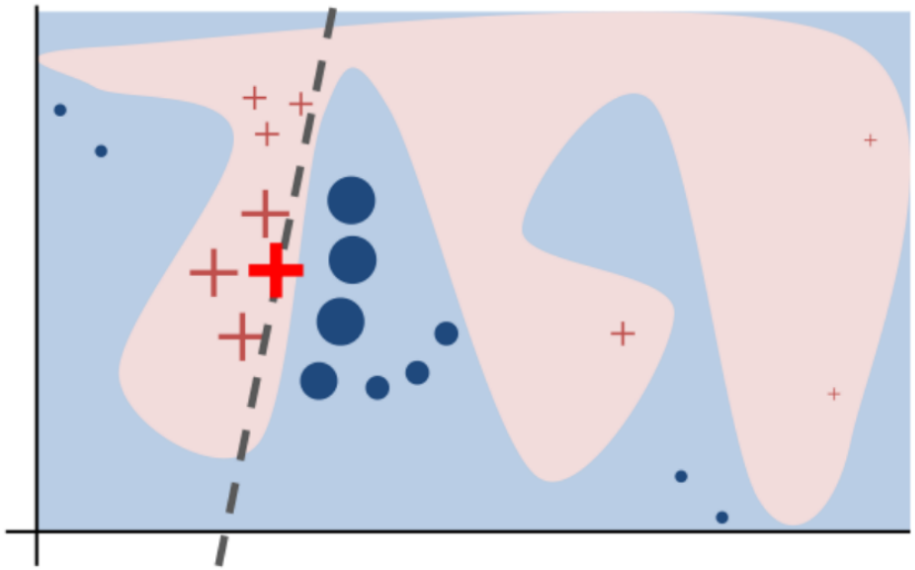
Some Observations on Explanations

- There are different possibilities to explain something to someone
  - verbal (different degrees of detail)
  - visual (maybe with symbolic annotations)
  - prototyical examples
  - contrastive (near miss) example
- There is no one-size fits all (context specificity)
- Explanations can be wrong (*right for the wrong reasons*, Teso & Kersting, AAAI/ACM Conference on AI, Ethics, and Society, 2019)
- Explanations are not always helpful (*Beneficial and Harmful Explanatory Machine Learning*, Ai, Muggleton, ..., Schmid, MLJ to appear)
- Explanations might lead to unjustified trust

Tim Miller, AIJ 2019; Tania Lombrozo, TiCS 2006

LIME

Local Interpretable Model-Agnostic Explanations

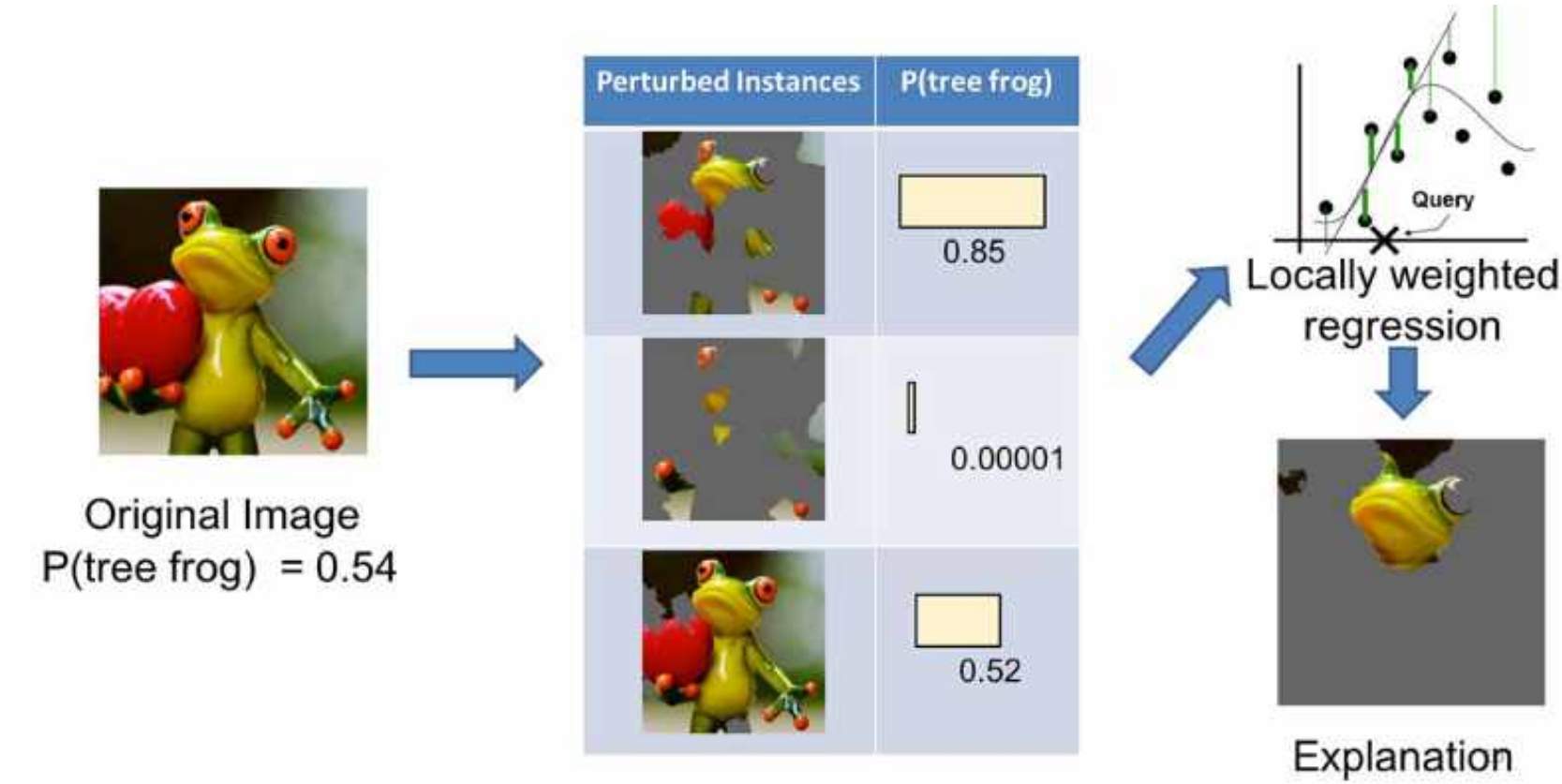


- Blue/pink: Complex decision function f of the learned classifier (unknown to LIME)
- Bold red cross: instance whose classification is to explain
- LIME samples perturbed examples, gets their class pediction from f, weights predictions wrt to their distance (size of circles and crosses)
- Dashed line: local explanation (no global lokale faithfulness)

Ribeiro, Singh, Guestin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD 2016

LIME

"Perturbed" samples (deleting part of information, e.g., superpixels, words)

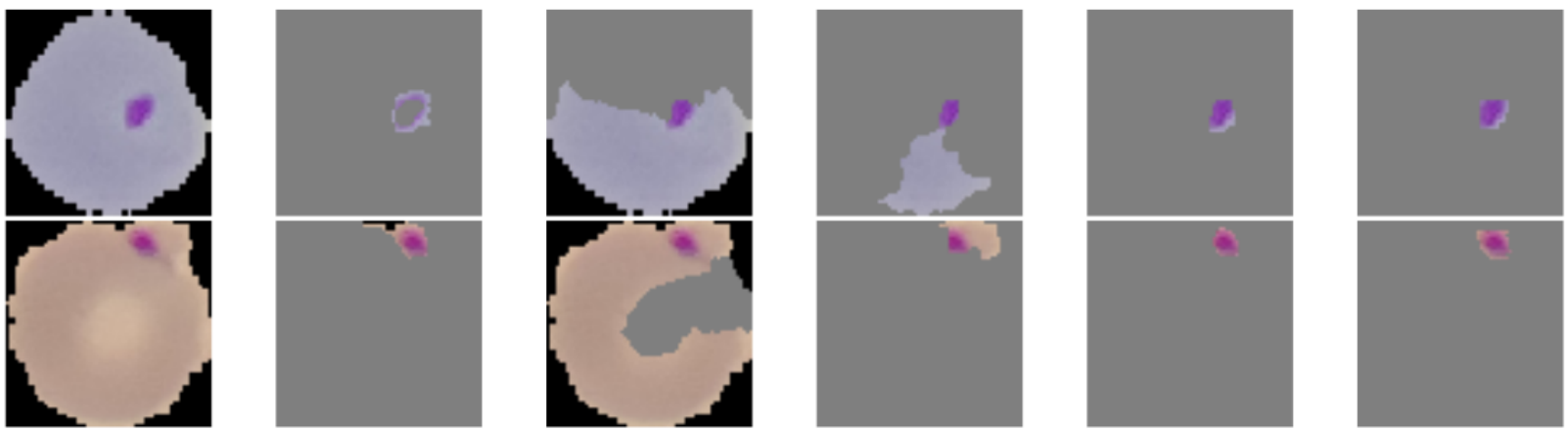


Ribeiro, Singh, Guestin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD 2016

LIME's Superpixel Approach Quick-Shift

Table 2: Jaccard Coeffficient of the different superpixel methods

Superpixel method	Mean Value	Variance	Standard deviation
Felzenszwalb	0.85603243	0.03330687	0.18250170
Quick-Shift	0.52272303	0.04613085	0.21478094
Quick-Shift optimized	0.88820585	0.00307818	0.05548137
SLIC	0.96437629	0.00014387	0.01199452
Compact-Watershed	**0.97850773**	**0.00003847**	**0.00620228**

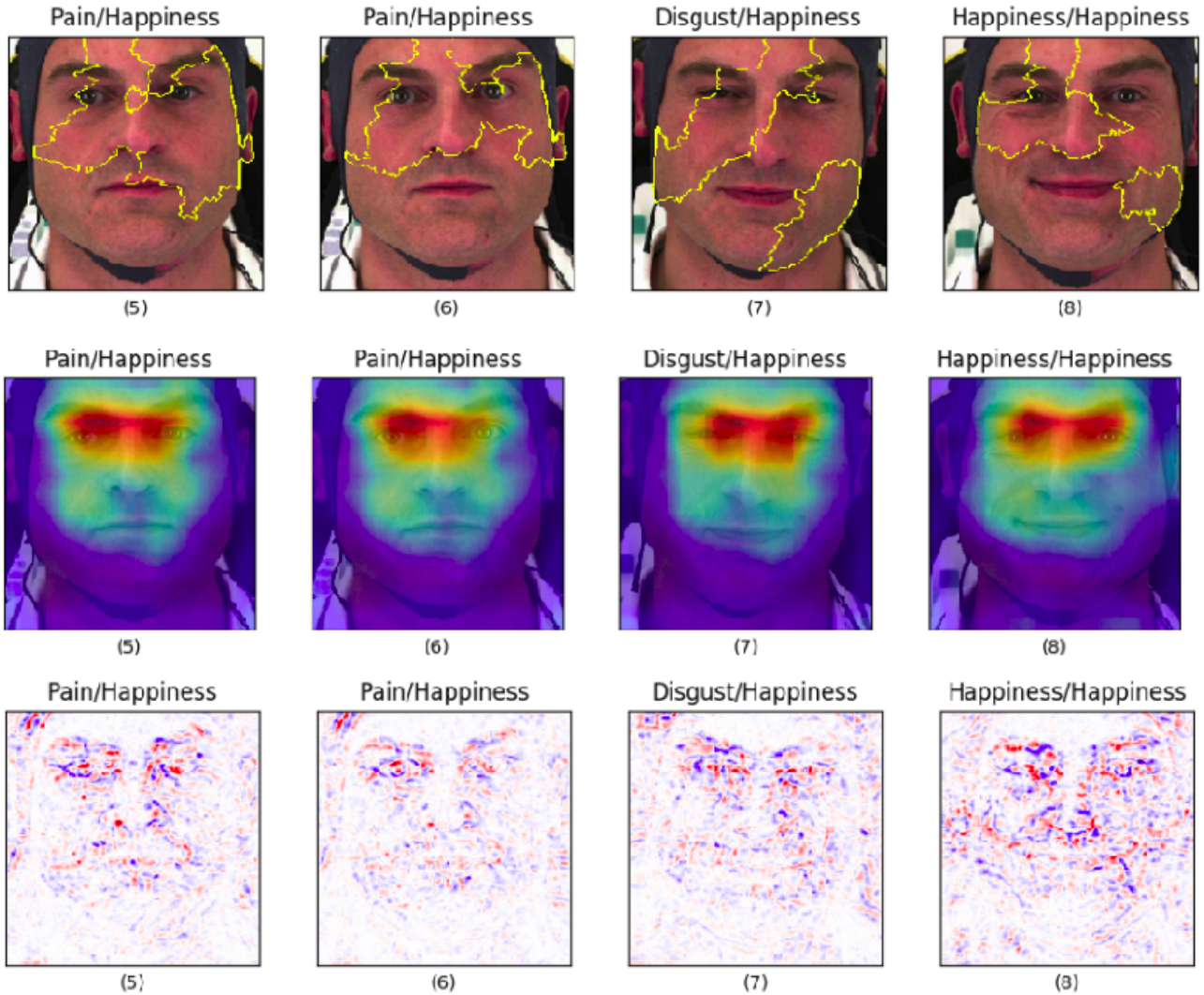


(a) Original (b) Felzenszwalb (c) Quick-Shift (d) Quick-Shift opt. (e) SLIC (f) Compact-Watershed

Fig. 4: LIME results for true positive predicted malaria infected cells

Schallner, Rabold, Scholz, Schmid, Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data, AIMLA 2019

Visual Explanations



LIME

CAM

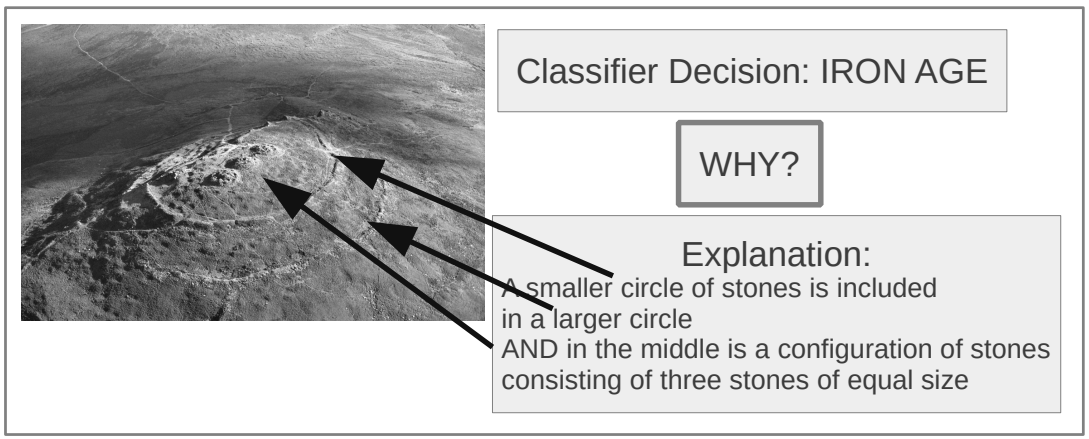
LRP

Weitz, Hassan, Schmid, Garbas, Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods, tm-Technisches Messen, 2019

Visual explanations are often not sufficient

- Helpful to recognize overfitting
- Fast communication of information (attention, relevance)
- BUT – visual highlighting is not expressive enough for
  - spatial relations (the blowhole is **on** a supporting part)
  - quantification (**all** blowholes are smaller than 1 mm)
  - feature values (the eyes are **shut** not open)
  - negation (there is **not** a blowhole but a hairline crack)
  - recursion (an arbitraty number of objects of increasing size)

$\hookrightarrow$ combining visual and verbal explanations

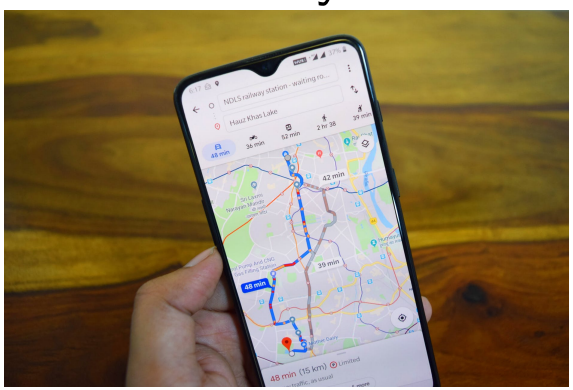


Rabold, Siebers, Schmid, ILP 2018; Rabold, Deininger, Siebers, Schmid, Enriching Visual with Verbal Explanations for Relational Concepts – Combining LIME with Aleph, AIMLA 2019

Human-AI Partnerships

- Interactive, cooperative learning as a means to prevent decrease of human competences and a possibility to correct machine learned models and to guide model adaptation
- Possibility to avoid the *Data Engineering Bottleneck*

Assistance System

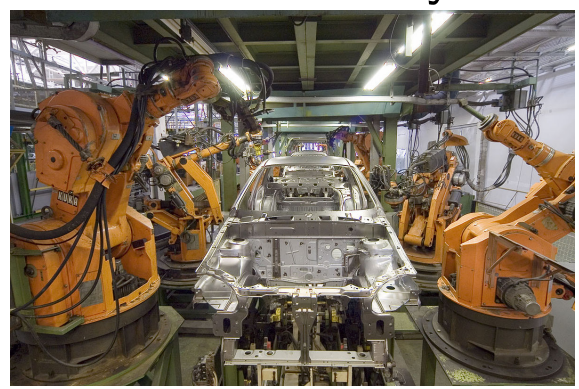


Human decides

Human-AI-Partnership

Joint decision making

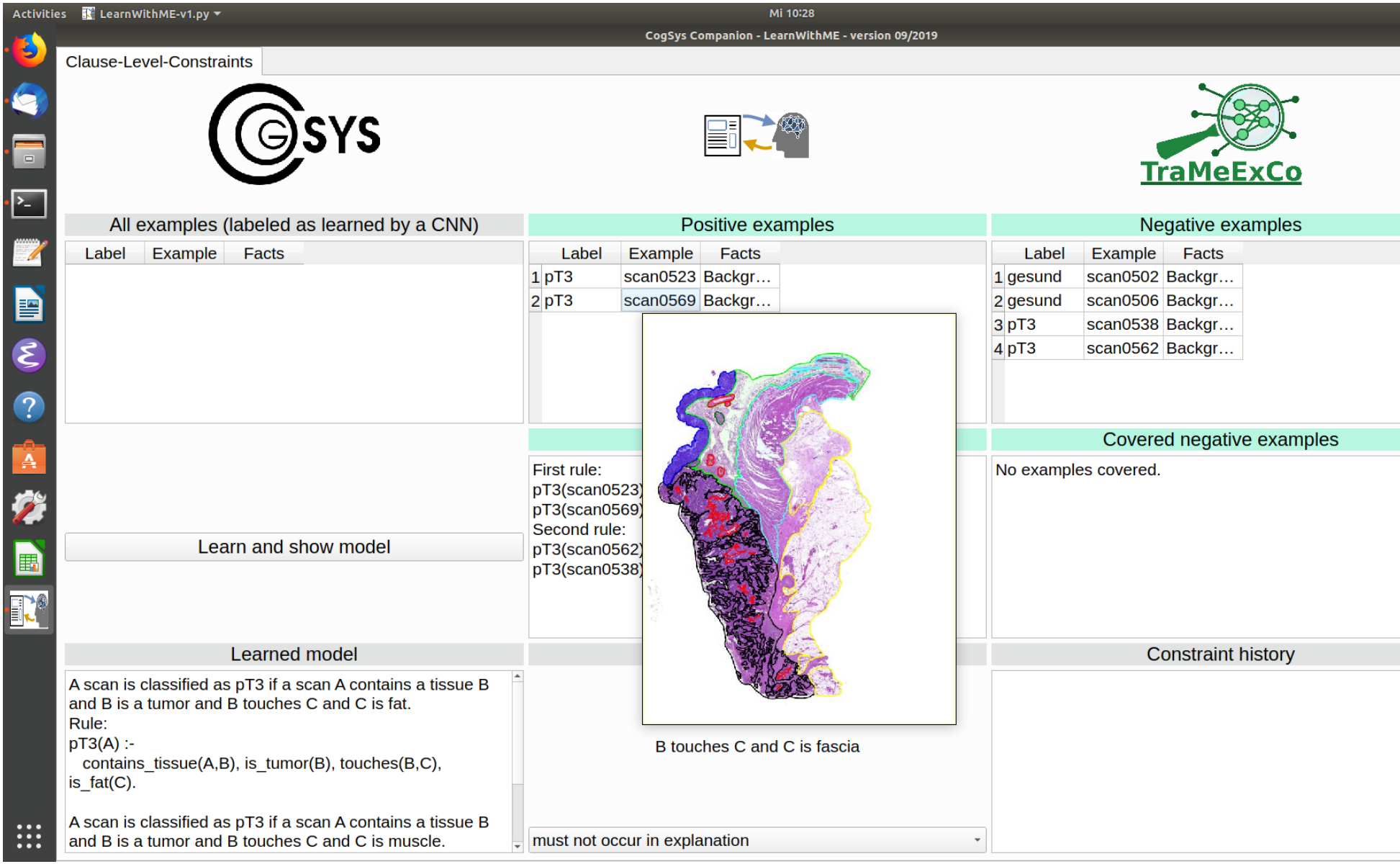
Autonomeous System



AI system decides

Special Topic Interactive ML, KI Zeitschrift 6/2020 (e.g., Mutual Explanations for Cooperative Decision Making in Medicine, Schmid & Finzel)

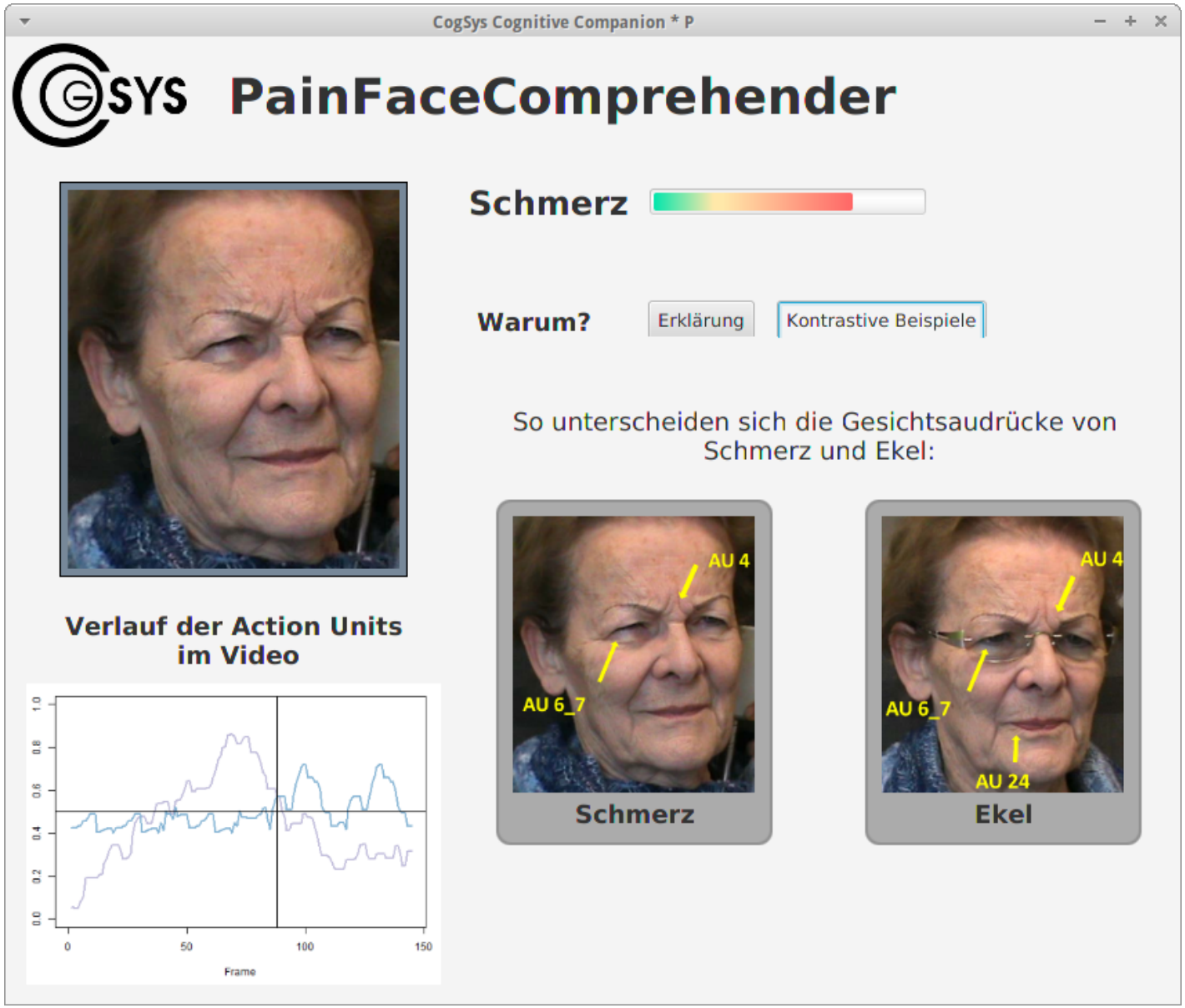
TraMeExCo: XI-ML for Medical Diagnosis



```
% Background Theory for Spatial Relations
% ---------------------------------------
% Area X touches area Y if holds that they have at least one boundary point
% in common, but no interior points.
touches(X,Y) :- I is intersection(X,Y), not(empty(I)),
InteriorX is interior(X), InteriorY is interior(Y),
J is intersection(InteriorX,InteriorY), empty(J).
% disjoint(X,Y) :- ...
% includes (X,Y) :- ...
% ...
% positive examples for diagnostic class pT3
% ----------------------------------------
% scan123 is classified as pT3. The scan is composed of areas of
% different tissues such as fat and tumor which are in specific spatial relations.
pt3(scan123).
contains_tissue(scan123,t1). contains_tissue(scan123,f1).
contains_tissue(scan123,f2).
is_tumor(t1). is_fat(f1). is_fat(f2)
touches(t1,f1). disjoint(f1,t1).
% negative examples for diagnostic class pT3 (e.g. pT2, pT4)
% ----------------------------------------
% ...
% Induced Rules: (learned from data with ILP)
% ----------------------------------------
% A scan is classified as pT3 if a scan A contains a tissue B
% and B is a tumor and B touches C and C is fat.
pT3(A) :-
   contains_tissue(A,B), is_tumor(B), touches(B,C), is_fat(C).
% further rules ...
```

Bruckert, Finzel, Schmid, The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions, Frontiers in AI, 2020

Support in Care – Identifying Pain



Dare2Del: Interactive learning to delete irrelevant digital objects

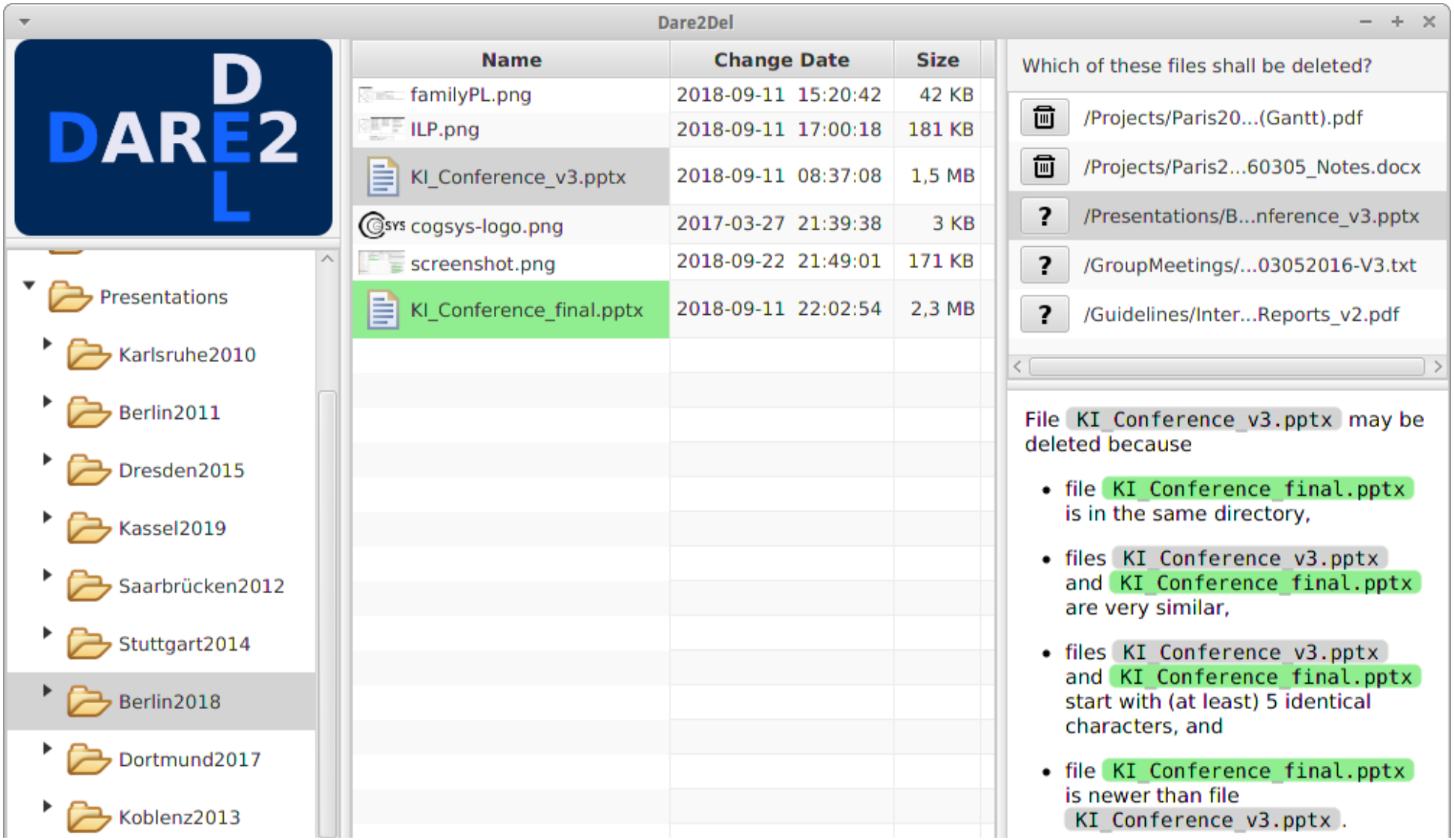


Bild Nürnberg vom 14.01.2020

Seite: 9 bis 9
Rubrik: NÜRNBERG
Mediengattung: Tageszeitung
Jahrgang: 2020
[1] IVW 3/2019
[2] AGMA ma 2019 Tageszeitungen

Nummer: 11
Auflage: 43.045 (gedruckt) [1] 31.324 (verkauft) [1] 31.365 (verbreitet) [1]
Reichweite: 0,214 (in Mio.) [2]

FRANKEN-SOFTWARE HILFT DATEN-MESSIES

Bamberg – **Kennen Sie das? 900 ungelese-**

Sie heißt „Dare-2Del" (Deutsch: „Wage es, zu

on oder es gibt eine Kopie in einem anderen

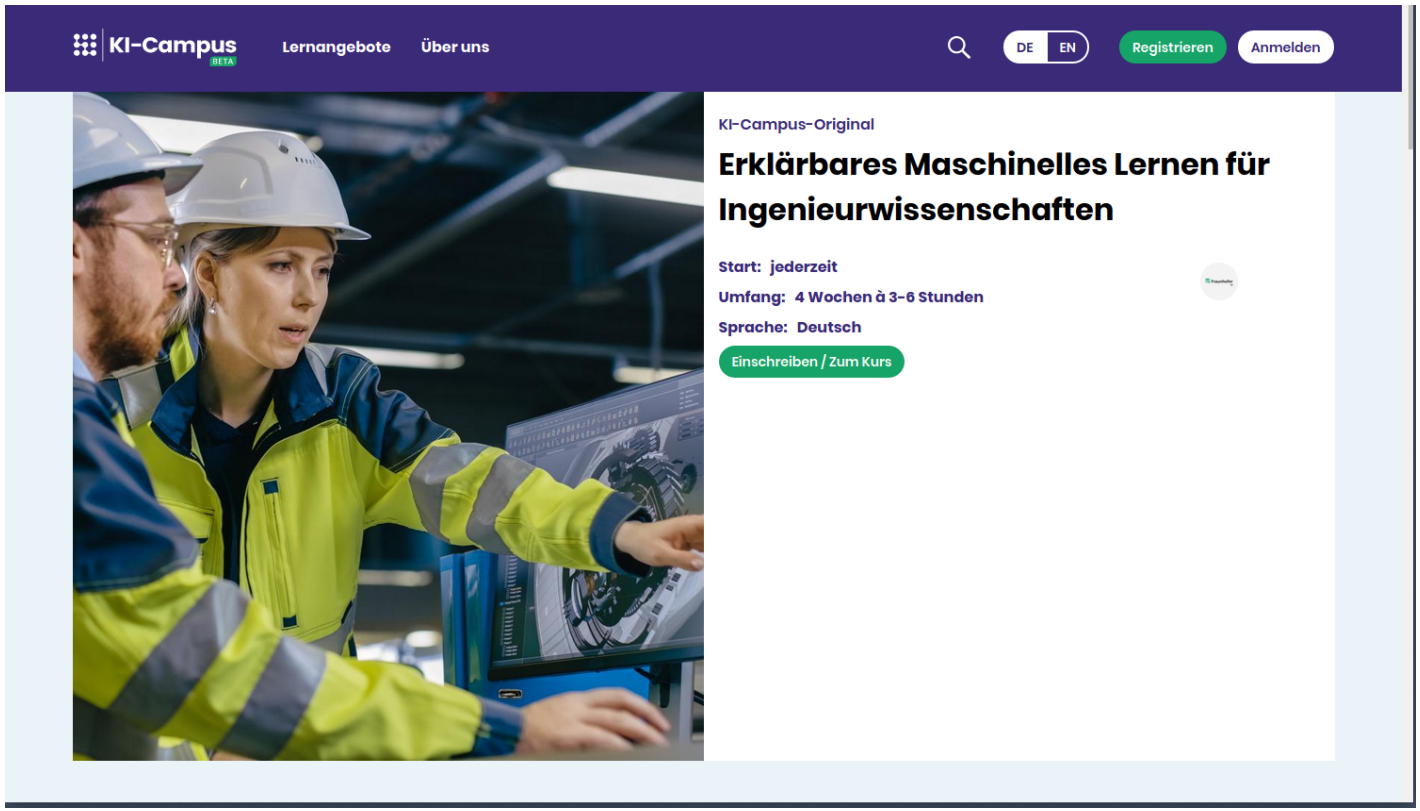
Take Away

- Many application domains have requirements which cannot be met by data intensive blackbox approaches of machine learning alone
- Interpretabel ML approaches such as regression, decision tree learning, Inductive Logic Programming (ILP) are less data intensive and inherently transparent
- Combining deep learning and ILP supports learning of classifiers for image data together with relational explanations
- Mutual explanations and interactive learning allow to integrate expert/common sense knoweldge into the learning process resulting in less need for data and allowing to correct errouneos decisions of the learned model

Bundesministerium für Bildung und Forschung
TraMeExCo (ML-3)

DFG Deutsche Forschungsgemeinschaft
Dare2Del (SPP 1921)
PainFaceReader

Learn more about XAI



STIFTERVERBAND

DFKI

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

mmb Institut

NEOCOSMO

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung