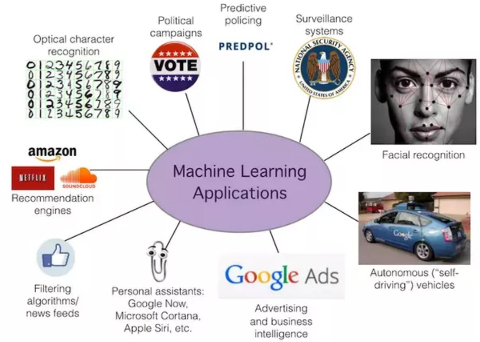
Learning from Mutual Explanations for Cooperative Decision Making in Medicin

Ute Schmid & Bettina Finzel

Cognitive Systems, Applied Computer Science, University of Bamberg
www.uni-bamberg.de/cogsys

AI+Health, University Luxembourg, Nov. 4, 2020

Machine Learning – From the Lab to the Real World



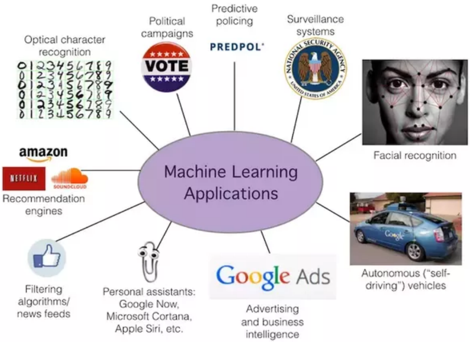
- Impressive success of deep learning
- CNNs, LSTMs, GANs
- AlexNet, AlphaGo, DeepFace, DeepL

`www.researchgate.net/publication/323108787_Introduction_to_Machine_Learning/`

BUT: $\hookrightarrow$ **Data engineering bottleneck**

- Some domains only have **small data sets**
- Data sets often are highly **imbalanced**
- **Sampling biases** may be not avoidable
- **Ground truth labeling** either not existent or expensive

Machine Learning – From the Lab to the Real World



- Impressive success of deep learning
- CNNs, LSTMs, GANs
- AlexNet, AlphaGo, DeepFace, DeepL

www.researchgate.net/publication/323108787_Introduction_to_Machine_Learning/

BUT: $\hookrightarrow$ **Necessity of explainability**

- **Transparency** for ML experts: recognize overfitting
- For critial domains: Provable **safety guaranties** for deep neural nets
- **Legal/ethical** obligation to explain decisions to end-users
- Establish **trust** for joint human-machine decision making

Tim Miller, Explanation in artificial intelligence, AIJ 2019

In the Following

- **Inductive Logic Programming** (ILP) as highly expressive approach to interpretable machine learning
- **Multimodal Explanation Generation** for Interactive Learning
- **Application Examples**
  - Classification of medical images (BMBF ML3 – TraMeExCo)
  - Facial expression analysis (DFG PainfaceReader)

Inductive Logic Programming (ILP)

- Proposed by Stephen Muggleton 1991
- Highly expressive interpretable ML approach
- Learning declarative programs from small data sets
- Models can capture arbitrary relations, can be defined over variables, allow recursion
- Exploiting relational structure in data
- Natural combination of reasoning and learning
- Allows translation into verbal explanations

Machine Learning
July 2018, Volume 107, Issue 7, pp 1119–1140 | Cite as

Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP

Authors Authors and affiliations

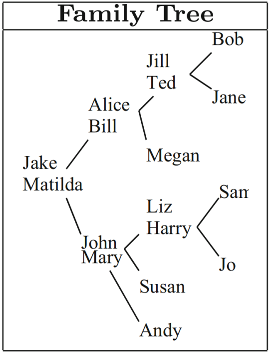
Stephen H. Muggleton, Ute Schmid, Christina Zeller, Alireza Tamaddoni-Nezhad, Tarek Besold

ILP: Learning Prolog Programs

- Hypotheses/models are represented as Prolog programs
- Examples are presented by target predicates (positive and negative) and by background knowledge
- In some approaches: also by background theories
- Uniform representation as Horn clauses

Gulwani, Hernandez-Orallo, Kitzelmann, Muggleton, Schmid, Zorn, Inductive Programming meets the real world, CACM, 2015

Example: Family Domain




```
% Background Knowledge
father(jake,bill).    mother(matilda,bill).
father(jake,john).    mother(matilda,john).
father(bill,ted).     mother(alice,jill).
father(bill,megan).   mother(alice,ted).
father(john,harry).   mother(alice,megan).
father(john,susan).   mother(mary,harry).
father(ted,bob).      mother(mary,susan).
father(ted,jane).     mother(mary,andy).
father(harry,san).    mother(jill,bob).
father(harry,jo).     mother(jill,jane).
mother(liz,san).      mother(liz,jo).
```

```
% Examples
grandparent(matilda,megan).    not grandparent(megan,matilda).
grandparent(matilda,harry).    not grandparent(jake,jake).
grandparent(jake,susan).       not grandparent(matila,alice).

% Learned hypothesis    (parent can be background theory or invented)
grandparent(X,Y) :- parent(X, Z), parent(Z,Y).
parent(X,Y) :- father(X,Y).
parent(X,Y) :- mother(X,Y).
```

ILP Algorithms

- FOIL (Quinlan, 1990): Generate-and-test, sequential covering
  cf. ID3, C4.5, simulteneous covering by the same author
- Golem, Progol, Aleph, Metagol (Muggleton, since 1990ies): inverse deduction principle in different variants
- Igor (Kitzelmann & Schmid, 2006; Schmid & Kitzelmann, 2011): Inductive (functional) programming
- ProbLog (de Raedt, 2007): combining logical and statistical learning

Algorithm

FOIL($Target\_predicate, Predicates, Examples$)

- $Pos \leftarrow$ those *Examples* for which the *Target_predicate* is *True*
- $Neg \leftarrow$ those *Examples* for which the *Target_predicate* is *False*
- $Learned\_rules \leftarrow \{\}$
- while *Pos*, Do
  - $NewRule \leftarrow$ the rule that predicts *Target_predicate* with no precondition
  - $NewRuleNeg \leftarrow Neg$
  - while *NewRuleNeg*, Do
    - $Candidate\_literals \leftarrow$ generate new literals for *NewRule*, based on *Predicates*
    - $Best\_literal \leftarrow argmax_{L \in Candidate\_literals} FoilGain(L, NewRule)$
    - add *Best_literal* to preconditions of *NewRule*
    - $NewRuleNeg \leftarrow$ subset of *NewRuleNeg* that satisfies *NewRule* preconditions
  - $Learned\_rules \leftarrow Learned\_rules + NewRule$
  - $Pos \leftarrow Pos - \{$members of *Pos* covered by *NewRule*$\}$
- Return *Learned_rules*

Probabilistic Inductive Logic Programming

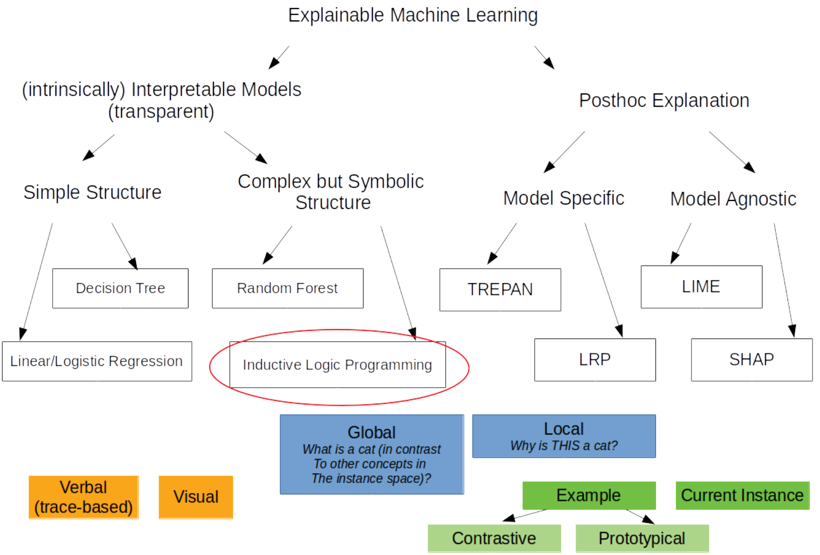
- Statistical Relational Learning (StarAI)
- Motivation: Biological Graphs
  `path(gene_620, disease_altzheimer)`
  edges are typically probabilistic

Example 1 *As an example, consider:*

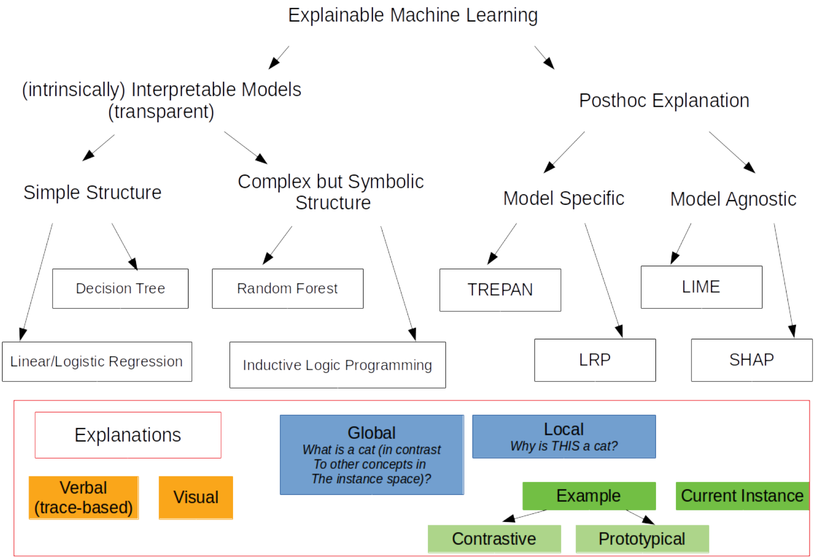
```
1.0: likes(X,Y):- friendof(X,Y).
0.8: likes(X,Y):- friendof(X,Z), likes(Z,Y).
0.5: friendof(john,mary).
0.5: friendof(mary,pedro).
0.5: friendof(mary,tom).
0.5: friendof(pedro,tom).
```

De Raedt, Kimmig, Toivonen, ProbLog: A probabilistic Prolog and its application in link discovery, IJCAI 2007

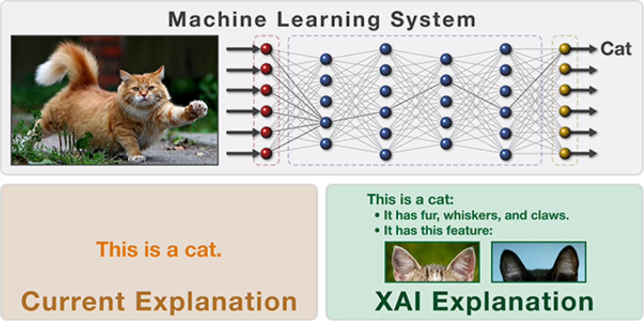
Interpretable ML with ILP



Explanation Generation



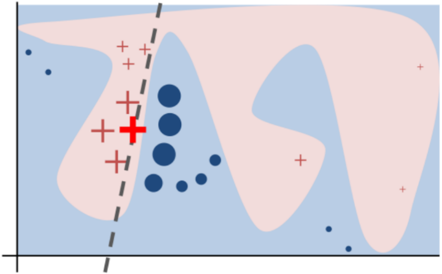
Explanation Interfaces



http://www.darpa.mil/program/explainable-artificial-intelligence

LIME

Local Interpretable Model-Agnostic Explanations

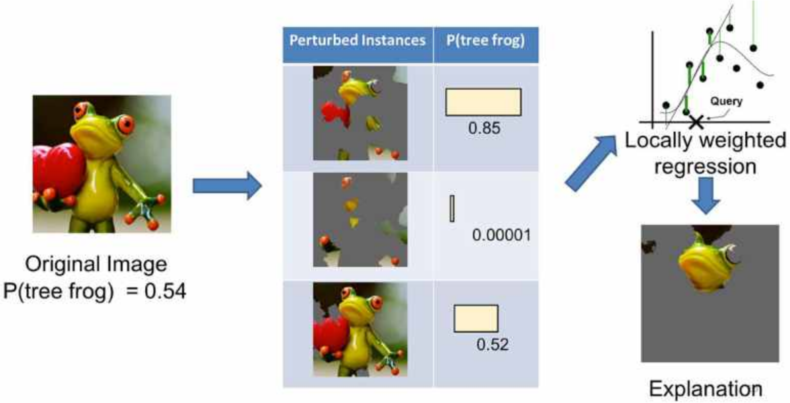


- Blue/pink background: The black-box model's complex decision function f (unknown to LIME)
- Bold red cross: instance being explained
- LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented by size)
- Dashed line: explanation that is locally (but not globally) faithful

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)

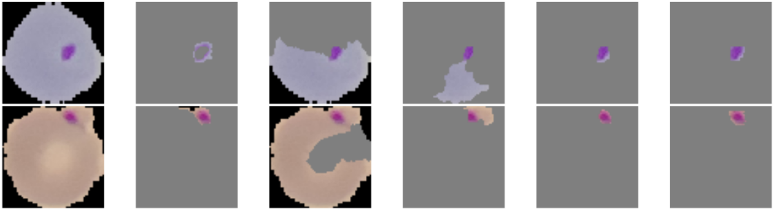


Ribeiro, Singh, Guestin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD 2016

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**

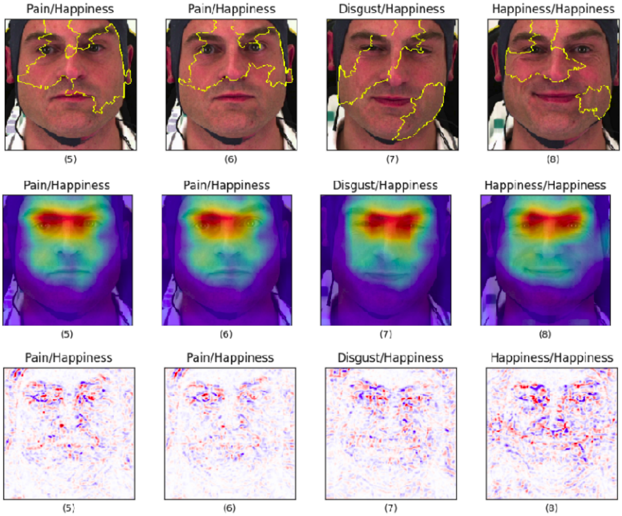


(a) Original (b) Felzenszwalb (c) Quick-Shift (d) Quick-Shift opt. (e) SLIC (f) Compact-Watershed

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

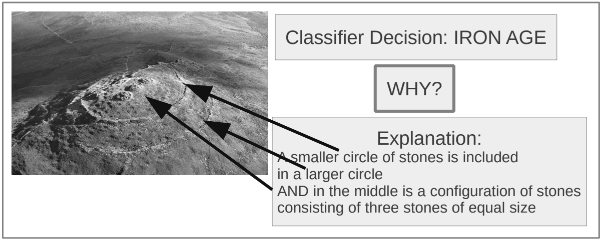
LRP

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 not Enough

- Can be helpful to detect unwanted overfitting
- Allow fast communication of information
- BUT – highlighting relevant areas cannot express
  - Spatial relations (the tumor tissue is **touching** fat tissue)
  - General characteristics (**all** metastases are smaller than 1 mm)
  - Negation (there is **not** a mole but a scar)

$\hookrightarrow$ Combining visual and verbal explanations



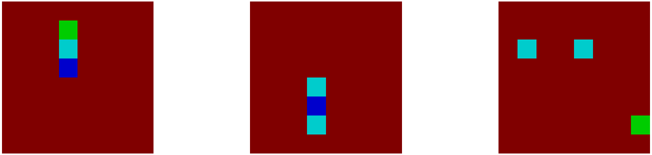
Rabold, Siebers, Schmid, ILP 2018; Rabold, Deininger, Siebers, Schmid, Enriching Visual with Verbal Explanations for Relational Concepts – Combining LIME with Aleph, AIMLA 2019

LIME-Aleph Algorithm

Require: Instance $x \in X$
Require: Classifier f, Selection size k, Threshold θ
Require: Attribute pool $\mathcal{A}$, Relation pool $\mathcal{R}$

```
S ← LIME(f, x, k)                          ▷ Selection of k most important super-pixels.
A ← extract_attribute_values(S, 𝒜)         ▷ Find all attribute values A_i for all i ∈ S.
R ← extract_relations(S, ℛ)                ▷ Find all relations r : S × S between all i ∈ S.
E⁺ ← {⟨A, R⟩}
E⁻ ← {}
for each r(i, j) ∈ R do
    z ← flip_in_image(x, i, j)             ▷ Flip the super-pixels in the image space.
    r′ ← r(j, i)                           ▷ Obtain new predicate for the BK by flipping parameters.
    R′ ← R \ {r} ∪ {r′}                    ▷ All relations in the BK; also the altered one.
    R′ ← calculate_side_effects(R′, r′)
                                           ▷ Re-calculate relations that are affected by the flipped relation.
    c′ ← f(z)                              ▷ Obtain new estimator for the perturbed image.
    if c′ ≥ θ do                           ▷ If estimator reaches threshold, add new positive example.
        E⁺ ← E⁺ ∪ {⟨A, R′⟩}
    else                                   ▷ Else, add negative example.
        E⁻ ← E⁻ ∪ {⟨A, R′⟩}
end for
T ← Aleph(E⁺, E⁻)                          ▷ Obtain theory T with Aleph.
return T
```

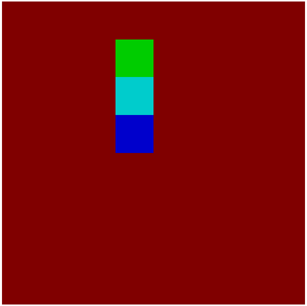
Experiment 2 - Concept "Tower"



(a) (b) (c)

Positive (a) and negative (b, c) examples for the concept "tower".

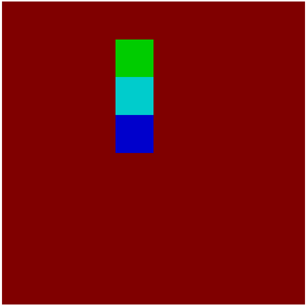
Experiment 2 - Concept “Tower“



$k = 3$, $\theta = 0.8$

```
concept(A) :- contains(B, A), has_color(B, cyan),
              contains(C, A), on(B, C).
```

Experiment 2 - Concept “Tower“

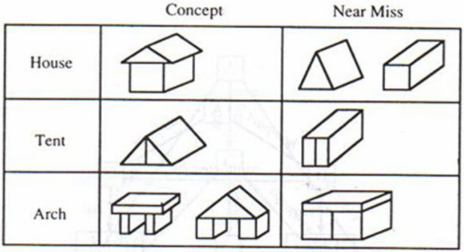


$k = \mathbf{4}$, $\theta = 0.8$

```
concept(A) :- contains(B, A), has_color(B, cyan),
contains(C, A), has_color(C, blue), top_of(B, C).
```

Contrastive Examples

- AI: Winston, 1970, **learning from near-miss examples**



- **Counterfactual reasoning**, Wachter et al., 2018:

  You were denied a loan because your annual income was 30,000 Euro. If your income had been 45,000 Euro, you would have been offered a loan.

Contrastive Examples

- Cognitive science: Gentner & Markman, 1994, **structural alignment**

APPENDIX

Table A1. *High- and low-similarity word pairs used in Experiments 1 and 2*

Similar pairs		Dissimilar pairs	
Light bulb	Candle	VCR	Lounge chair
Kitten	Cat	Hammock	Horse track
Magazine	Newspaper	Bed	Hockey
Bowl	Mug	Football	Boutique
Phone book	Dictionary	Kite	Painting
Microphone	Stereo speaker	Sculpture	Navy
Piano	Organ	Army	Abacus
Air conditioner	Furnace	Calculator	Escalator
Freezer	Refrigerator	Stairs	Stool
Hammer	Mallet	Broom	Sailboat
Bicycle	Tricycle	Yacht	Missile
Dumpster	Garbage can	Chair	Banana split
Lake	Ocean	Ice cream sundae	Clock
Telephone	CB radio	McDonald's	Couch
Diamond	Ruby	Police car	Burger King
Sponge	Towel	Rocket	Motel
Computer	Typewriter	Hotel	Tape deck
Staple	Paper clip	Watch	Ambulance
Shoe	Sandal	Casino	Mop
Chemistry	Biology	Stove	Hang glider
VCR	Tape deck	Light bulb	Cat
Hammock	Lounge chair	Kitten	Newspaper

Natural Language Explanations: From Shallow to Deep

- The rule which is applied to classify a current image can be explained verbally (using a template for text generation from Prolog)
- Verbal explanations can be shallow – referring only to the predicates in the body – or deep by expanding them to additional rules or facts.

Why did you classify this expression as pain?

Sequence "sub13_seq37" is a pain sequence because

shallow event "e1" starts the sequence, finishes the sequence, and is an Action Unit 7 of intensity "C".

```
starts(e1,sub13_seq37), finishes(e1,sub13_seq37), is_au(e1,au7),
has_intensity(e1,c)
```

Natural Language Explanations: From Shallow to Deep

Why did you classify this expression as pain?

Sequence "sub13_seq37" is a pain sequence because

shallow `starts(e1,sub13_seq37), finishes(e1,sub13_seq37), is_au(e1,au7), has_intensity(e1,c)`

deep event "e1" begins in frame 17 in sequence "sub13_seq37", there is no other event in this sequence which starts in an earlier frame, event "e1" ends in frame 82 in sequence "sub13_seq37", there is no other event in this sequence which ends in a later frame, and is an Action Unit 7 of intensity "C".

```
event_starts(e1,sub13_seq37,17), not(event_starts(E,sub13_seq37,F),
F<17),
event_ends(e1,sub13_seq37,82), not(event_ends(E,sub13_seq37,F), F>82),
is_au(e1,au7), has_intensity(e1,c)
```

even deeper An event is the occurance of an action unit with a specific intensity in a sequence of action units shown by a specific person

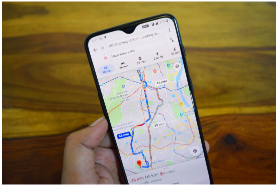
Multi-Level Explanations – Demo

```
Please enter the identifier of the example that you want to classify and explain: sub1_seq2_pain03
Ask me a question and I will tell you more! You can ask me (1) What class is <example>...? - to ask for a class label of an instance,
 (2) Explain why <part of explanation> - to ask for more details, (3) Get back to the last explanation - to get back to the previous
explanation (one-level higher)
Please ask a valid question.
Your question: What class is sub1_seq2_pain03
 sub1_seq2_pain03 is of class pain
Your question: Explain why sub1_seq2_pain03 is of class pain
B overlaps C in sub1_seq2_pain03 and C is au06 and B is au07.
Your question: Explain why B overlaps C
Start of B is earlier than Start of C and End of B is later than Start of C and End of B is earlier than End of C.
Your question: Get back to the last explanation
Getting back...
B overlaps C in sub1_seq2_pain03 and C is au06 and B is au07.
Your question: Explain why au06
Displaying frame that shows au06 and description that says that au06 is the cheek raiser.
Your question: Explain why au04
You want me to tell you something more about something undefined. Please ask a valid question.
Your question: Explain why au07
Displaying frame that shows au06 and description that says that au07 is the lid tightener.
Your question: Quit
Program quits now...
>>> |
```

Human-AI-Partnership

- Combine the strengths of humans and strength of AI-systems
- Joint performance should surpass performance of human or machine alone (cf. Donals Michie's Ultra-stong machine learnin criterion)
- Need for suitable interaction interfaces based on extended/novel machine learning approaches
  - transparent and comprehensible ML
    explanations, justifications, argumentations
  - incremental learning by feedback from human experts

Assistance System

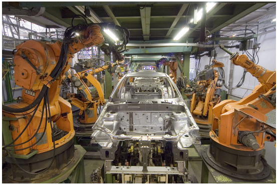


human decides

Interactive ML System

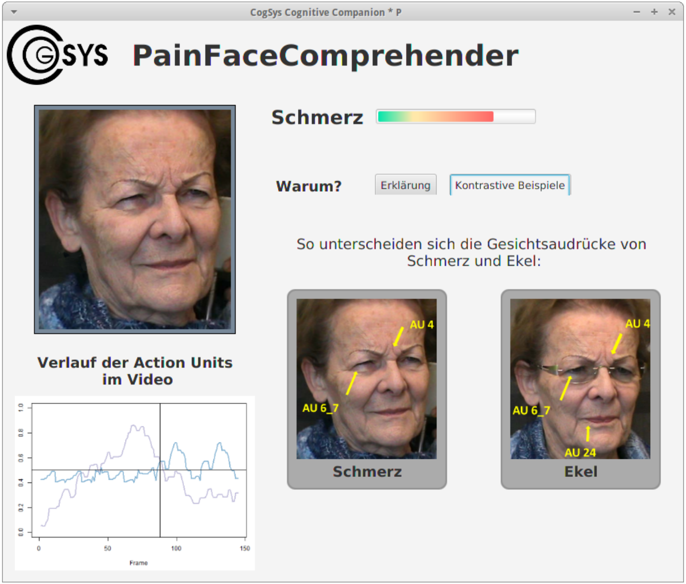
joint decision making

Autonomous System

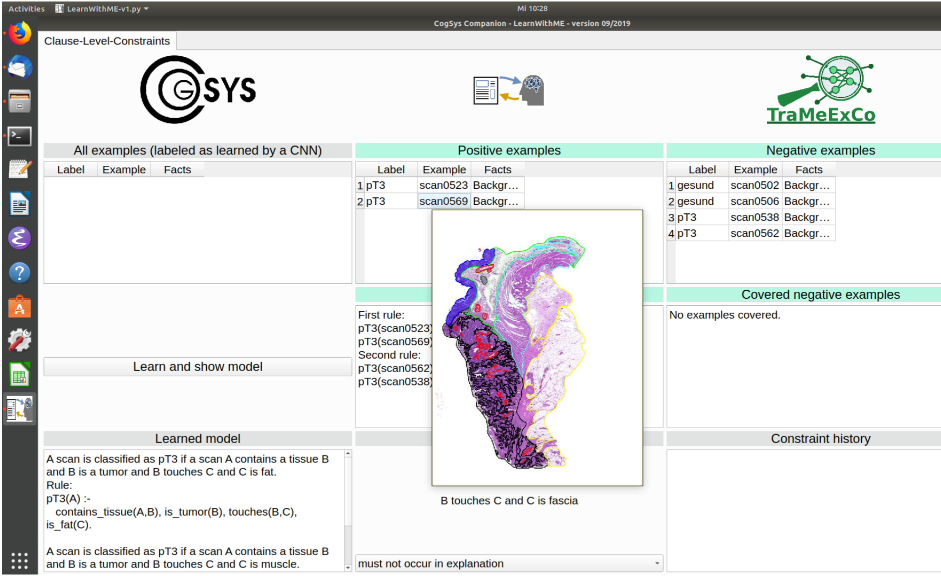


machine decides

Supporting Nurses – Pain Classification



Medical Expert Support for Tissue Classification



```
% 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 ...
```

Take Away

- Machine learning in the real world has many requirements which cannot be met by blackbox machine learning alone
- Inductive Logic Programming is a highly expressive approach to interpretable machine learning
- Combination of deep learning and ILP allows relational explanations
- Verbal explanations can be generated from reasoning traces
- Contrastive examples can highlight relevant aspects
- Mutual explanations allow to introduce expert knowledge (less data needed, correcting wrong labels) as constraints for learning

Bundesministerium für Bildung und Forschung
TraMeExCo (ML-3)

DFG Deutsche Forschungsgemeinschaft
Dare2Del (SPP 1921)
PainFaceReader

Fraunhofer IIS

In cooperation with
fortiss
Bayerisch-Tschechische Hochschulagentur
Česko-bavorská vysokoškolská agentura
BTHA-AP-2019-27 VSE Prag (Tomáš Kliegr)