



# Comprehension, Correction and Competence: Combining XAI and ITS for Human and Machine Teaching

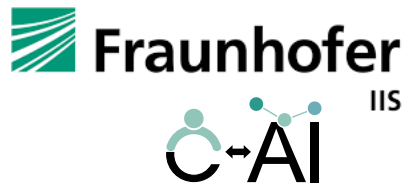
**Ute Schmid**

Cognitive Systems  
University of Bamberg

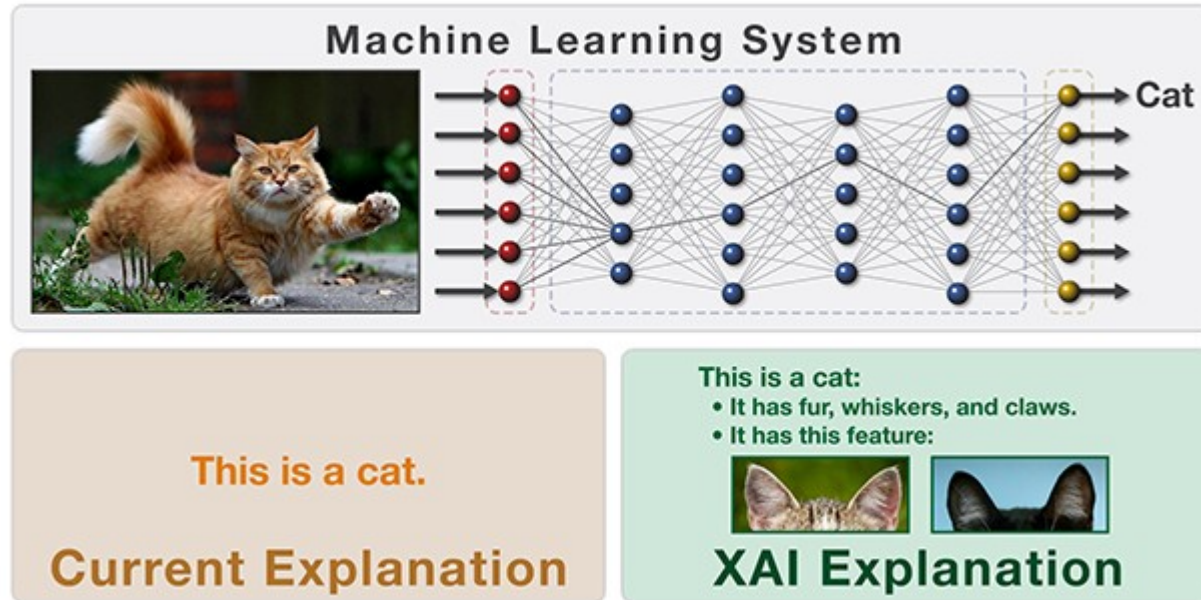
Machine Teaching for Explainable AI

Workshop

Valencia, Spain  
January 2024



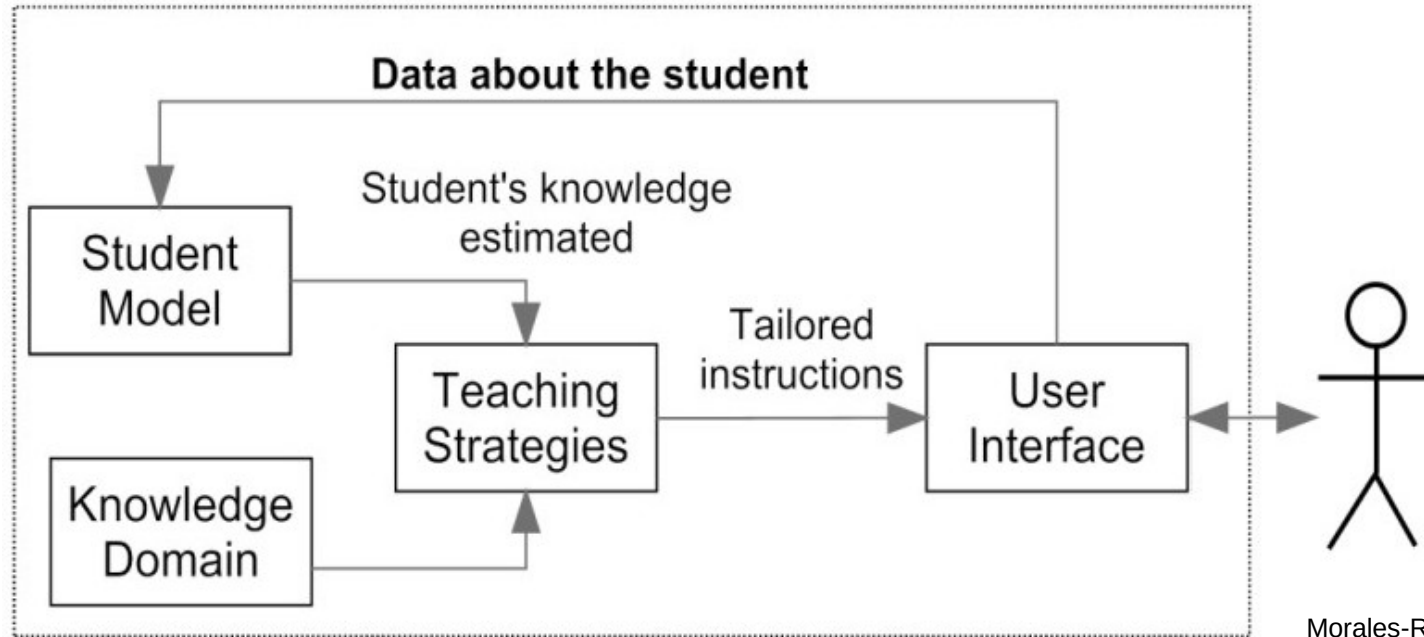
# XAI is for comprehending ML models



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

David Gunning, IJCAI 2016

# ITS is for comprehending a knowledge domain



**Fig. 1.** Basic architecture of an ITS [7].

Morales-Rodríguez, M.L., Ramírez-Saldivar, A., Hernández-Ramírez, A., Sánchez-Solís, J.P., & Flores, J.A. (2012). Architecture for an Intelligent Tutoring System that Considers Learning Styles. *Res. Comput. Sci.*, 47, 37-47.



# How can XAI and ITS be interleaved?

- XAI offers an ever growing set of explanation methods
- ITS provides methods to model the knowledge states (and maybe other psychological aspects) of humans
- XAI can profit for tailoring explanations to fit the specific information need of humans
- ITS can profit by extending the set of didactic interventions

# Explanation Methods in XAI

What a good explanation is, is context-dependent:

- **what is explained**

a current output or the model (local vs. global), a classification, a detected anomaly, a generated text ... (type of model output)

- **to whom**

an ML expert, a domain expert, a end-user

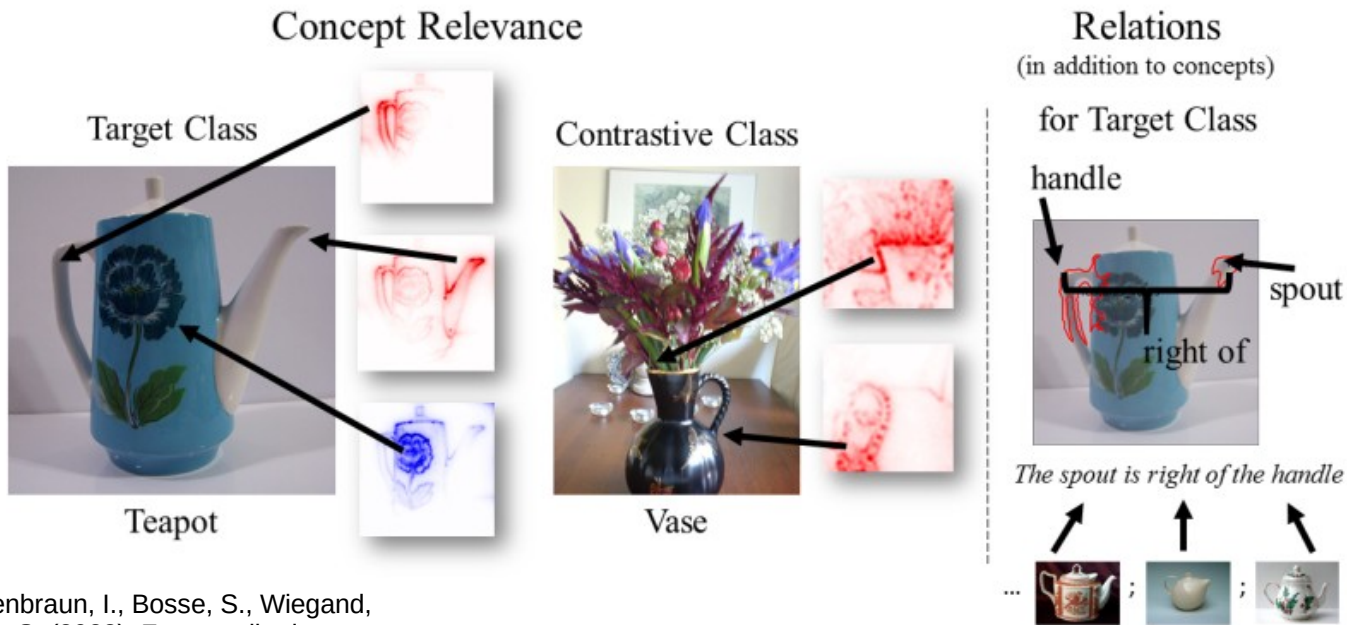
- **in what way (how)**

highlighting, symbolic/verbal, example, prototype, counterfactual, contrastive

- **for what reason (why)**

Understand reason for possible wrong output, decision boundaries, understand to correct (XIML)

# Explaining with concepts and relations



See esp.:

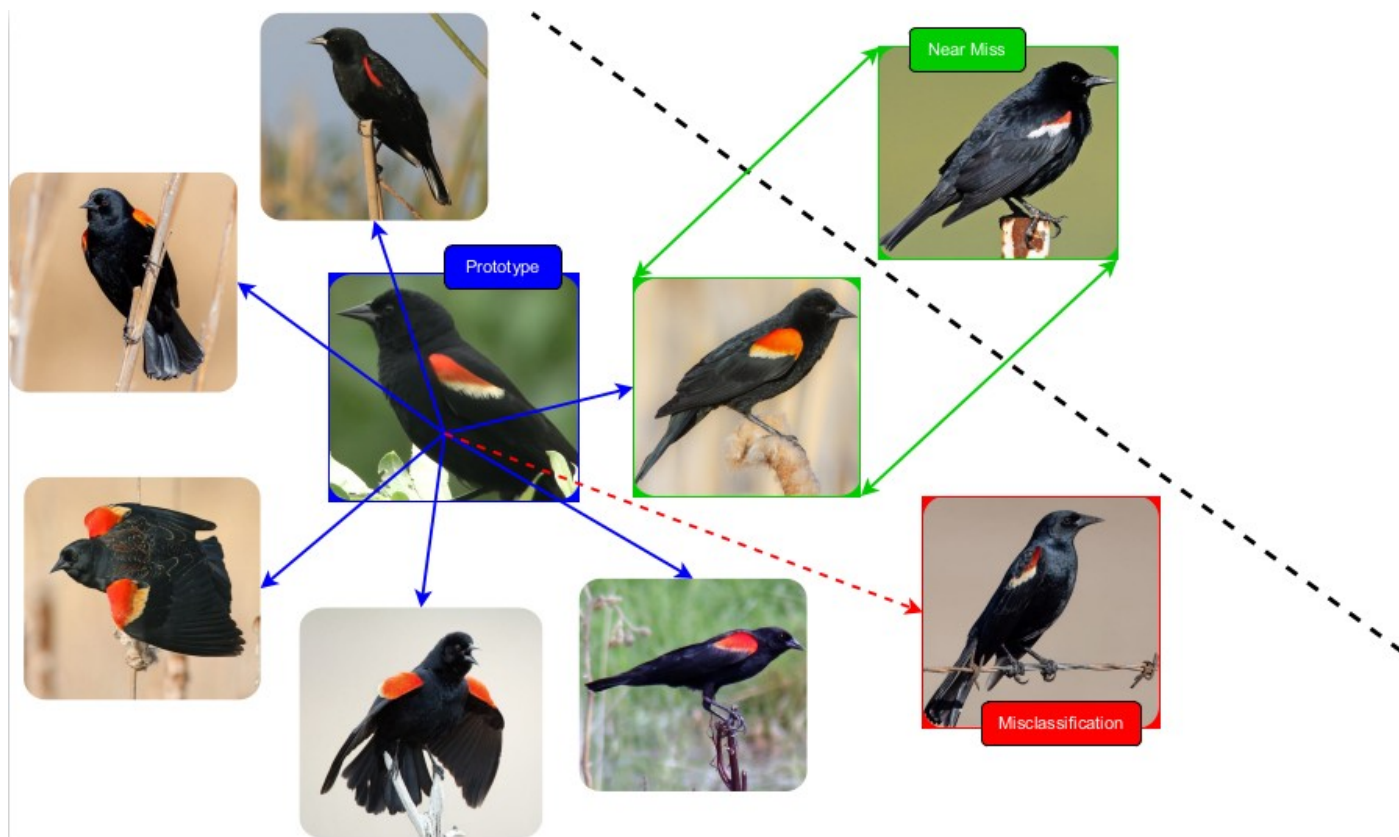
Achibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5(9), 1006-1019.

# Explanation by Near Misses and Prototypes

See also:

Herchenbach, M., Müller, D., Scheele, S., & Schmid, U. (2022, May). Explaining Image Classifications with Near Misses, Near Hits and Prototypes: Supporting Domain Experts in Understanding Decision Boundaries. In International Conference on Pattern Recognition and Artificial Intelligence (pp. 419-430). Cham: Springer International Publishing.

Rabold, J., Siebers, M., & Schmid, U. (2022). Generating contrastive explanations for inductive logic programming based on a near miss approach. Machine Learning, 111(5), 1799-1820.





# Near Miss Explanations for Effective Teaching

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    |

Gentner & Markman. Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5(3):152–158, 1994.



# Contrastive Explanations and Causality

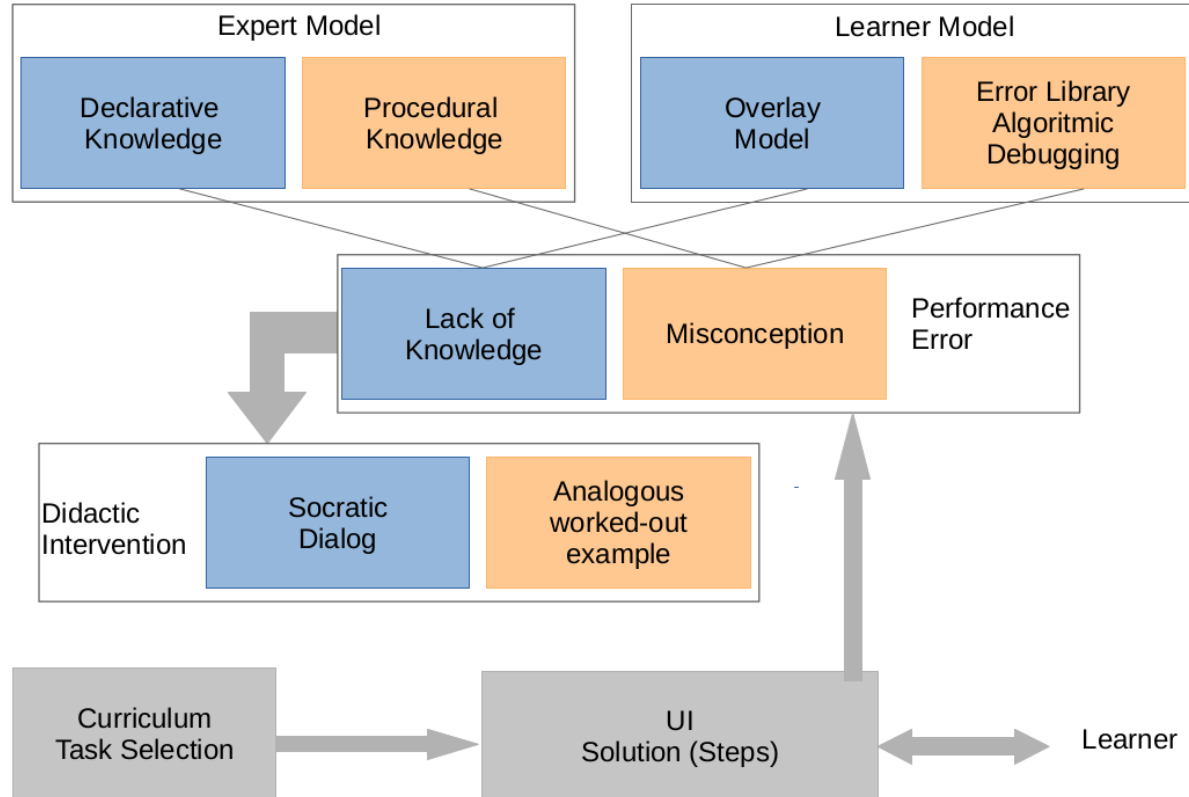
**Causal explanations are contrastive** (Tim Miller, 2019, referring to P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27, 1990)

- To be a beetle, an arthropod must have six legs, but this does not cause an arthropod to be a beetle – other causes are necessary.
- But, to answer the question: “Why is image J labelled as a Beetle instead of a Spider?” it is sufficient to cite the fact that the arthropod in the image has six legs.
- We do not need information about eyes, wings, or stingers to answer this, whereas to explain why image J is a spider in a non-contrastive way, we must cite all causes.

| Type   | No. Legs | Stinger | No. Eyes | Compound Eyes | Wings |
|--------|----------|---------|----------|---------------|-------|
| Spider | 8        | ✗       | 8        | ✗             | 0     |
| Beetle | 6        | ✗       | 2        | ✓             | 2     |
| Bee    | 6        | ✓       | 5        | ✓             | 4     |
| Fly    | 6        | ✗       | 5        | ✓             | 2     |

Tim Miller, Explanation in artificial intelligence: Insights from the social sciences. AIJ 2019

# Modeling knowledge states



# Example: Overlay Model in Scholar (Carbonell, 1970)

APPROX WHAT IS THE AREA OF BRAZIL?

\*2500000 SQUARE MILES\*

WRONG.  
PLEASE INDICATE IF THE FOLLOWING STATEMENT IS CORRECT  
OR INCORRECT:

THE AREA OF PARAGUAY IS APPROX 47432 SQUARE MILES.

\*WHAT IS THE AREA OF BRAZIL?\*

THE AREA OF BRAZIL IS APPROX 3600000. SQUARE MILES.  
NOW ANSWER THE QUESTION YOU DIDN'T ANSWER BEFORE.

\*FALSE\*

REPHRASE YOUR STATEMENT. I CAN'T UNDERSTAND IT.

\*INCORRECT\*

VERY GOOD.

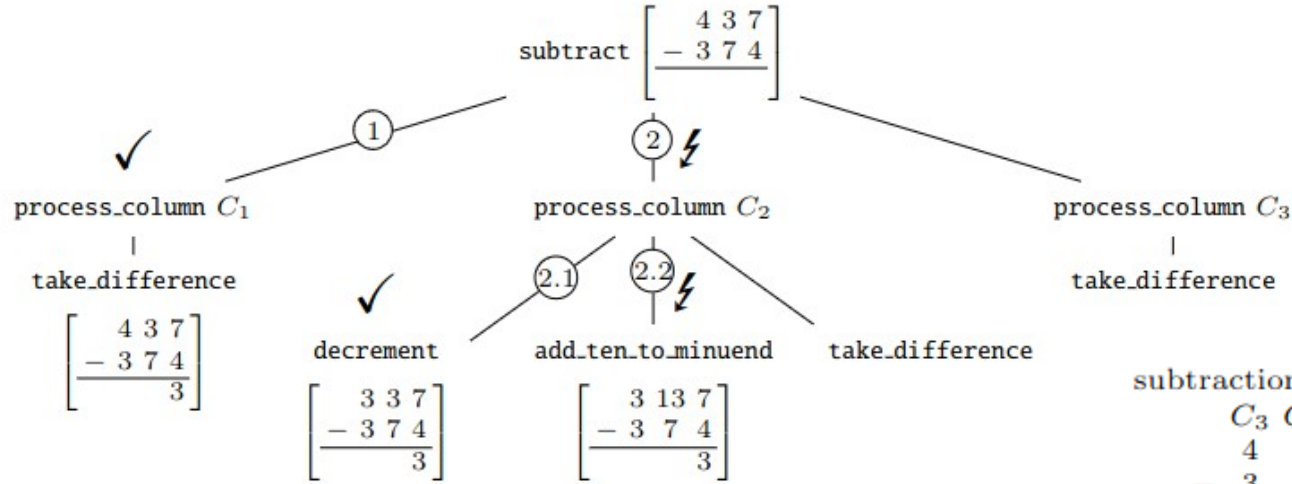
THE AREA OF PARAGUAY IS APPROX 157000 SQUARE MILES.

The diagram illustrates the 'Overlay Model' in the Scholar system, showing a network of interconnected boxes representing different entities and their attributes. The boxes are:

- LATITUDE**: A simple box with dashed lines for input.
- CONTINENT**: A box with dashed lines for input.
- ARGENTINA**: A box containing the following text:  
(SUPERC COUNTRY)  
(LOCATION SOUTH AMERICA)  
(LATITUDE (RANGE -22 -55))  
(BORDERING COUNTRIES (EASTERN BRAZIL URUGUAY))
- SOUTH AMERICA**: A box containing the following text:  
(SUPERC CONTINENT)  
(COUNTRIES ARGENTINA  
URUGUAY VENEZUELA)
- COUNTRY**: A box containing the following text:  
(SUPERC (STATE INDEPENDENT))  
(SUPERC CONTINENT)  
(EXAMPLES ARGENTINA  
BOLIVIA BRAZIL  
URUGUAY U. S. VENEZUELA)
- URUGUAY**: A box containing the following text:  
(SUPERC COUNTRY)

Lines connect these boxes, indicating relationships and data flow between the entities and their attributes. For example, 'ARGENTINA' is connected to 'SOUTH AMERICA' and 'COUNTRY', while 'SOUTH AMERICA' is connected to 'CONTINENT' and 'COUNTRY'. 'URUGUAY' is connected to 'COUNTRY' and 'ARGENTINA'.

# Identifying misconceptions with Algorithmic Debugging



And giving feedback  
with an analog worked-out example

subtraction problem

$$\begin{array}{r} C_3 \ C_2 \ C_1 \\ 4 \ 3 \ 7 \\ - \ 3 \ 7 \ 4 \\ \hline \end{array}$$

typical student solution

$$\begin{array}{r} m_{-1} \ m^{+10} \ m \\ 3 \ 3 \ 7 \\ - \ 3 \ 7 \ 4 \\ \hline 0 \ 4 \ 3 \end{array}$$

analogous problem

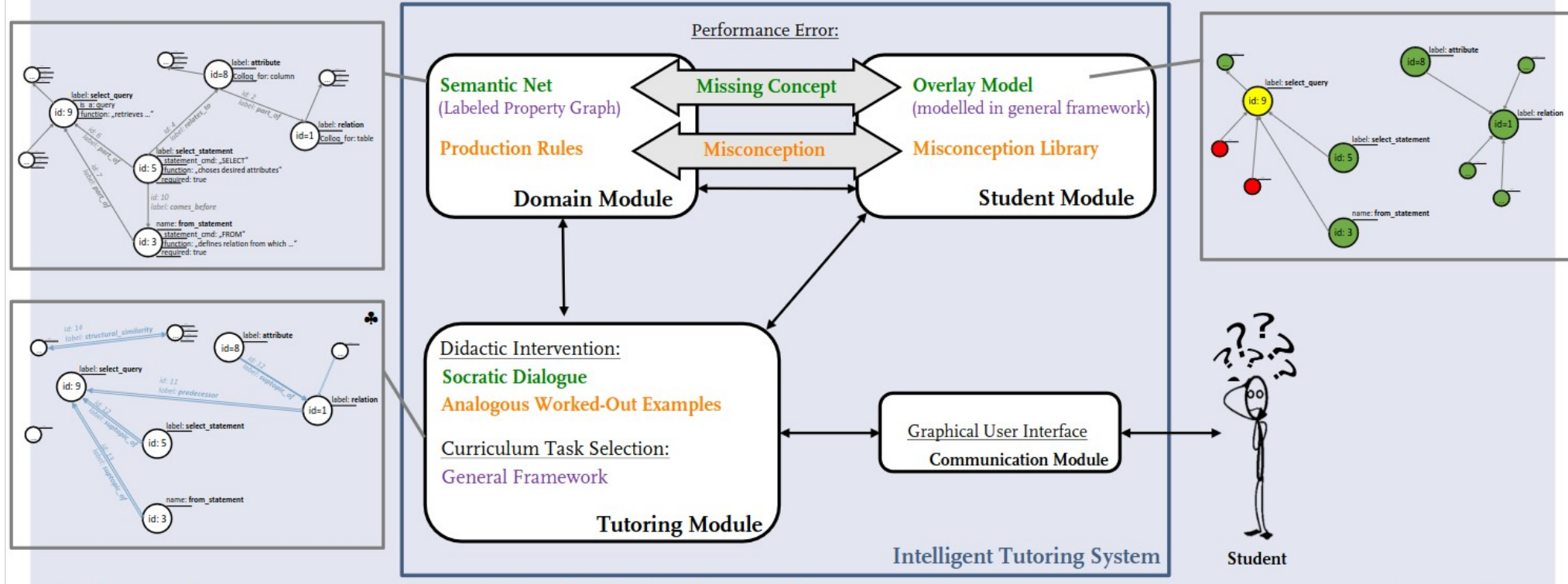
$$\begin{array}{r} C_3 \ C_2 \ C_1 \\ 4 \ 1 \ 0 \\ - \ 1 \ 8 \ 0 \\ \hline \end{array}$$

analogous problem solution

$$\begin{array}{r} m_{-1} \ m^{+10} \ m \\ 3 \ 11 \ 0 \\ - \ 1 \ 8 \ 0 \\ \hline 2 \ 3 \ 0 \end{array}$$

Zeller, C., & Schmid, U. (2016). Automatic generation of analogous problems to help resolving misconceptions in an intelligent tutor system for written subtraction.

# General Structure to ITS with declarative and procedural knowledge



Thaler, A. M., Mitrovic, A., Schmid U. (2022). Worked Examples as Application of Analogical Reasoning in Intelligent Tutoring and their Effects on SQL Competencies. Biannual Conference of the German Cognitive Science Society. (KogWis, 2022, Freiburg, Germany)

# An ITS for Primary School Education

Learn to look closely and be aware of relevant feature values to discriminate local trees

- Shape: tripartite, lobed, elliptic
- Margin: even, wavy toothed, saw toothed

<https://en.wikipedia.org/wiki/Leaf>

Near miss on feature value:  
show a leaf which indeed is saw toothed  
and one which is wavy toothed  
(but from another category)

CogSys Cognitive Companion \* C

**GSYS Im heimischen Wald**

Antwort 1: Das Blatt ist eiförmig.

Frage 2: Ist das Blatt gesägt oder gezähnt?

☐ gesägt

☒ gezähnt

Weiter

**Bist du sicher?**

Sieh dir diese beiden Beispiele an!

 Dies ist ein gesägtes Blatt

 Dies ist ein gezähntes Blatt

Schmid, U. (2020). AI goes to school: learning about and learning with artificial intelligence. In Proceedings of the 15th Workshop on Primary and Secondary Computing Education.

# ITS for Recursive Programming – Identifying misconceptions by testing

```
def sum(n):  
    if n == 0:  
        0  
    else:  
        n + sum(n-1)
```

testing: 1 → None  
          4 → None  
         10 → None

RCnoReturnRequired\*  
No return statement  
required

```
def sum(n):  
    if n == 1:  
        return 0  
    else:  
        return n + sum(n-1)
```

testing: 1 → 0  
          4 → 9  
         10 → 54

BCevaluation\*  
Execution either too often or  
not often enough

\*Sally Hamouda, Stephen H. Edwards, et al., (2017) A basic recursion concept inventory, Computer Science Education, 27:2, 121-148

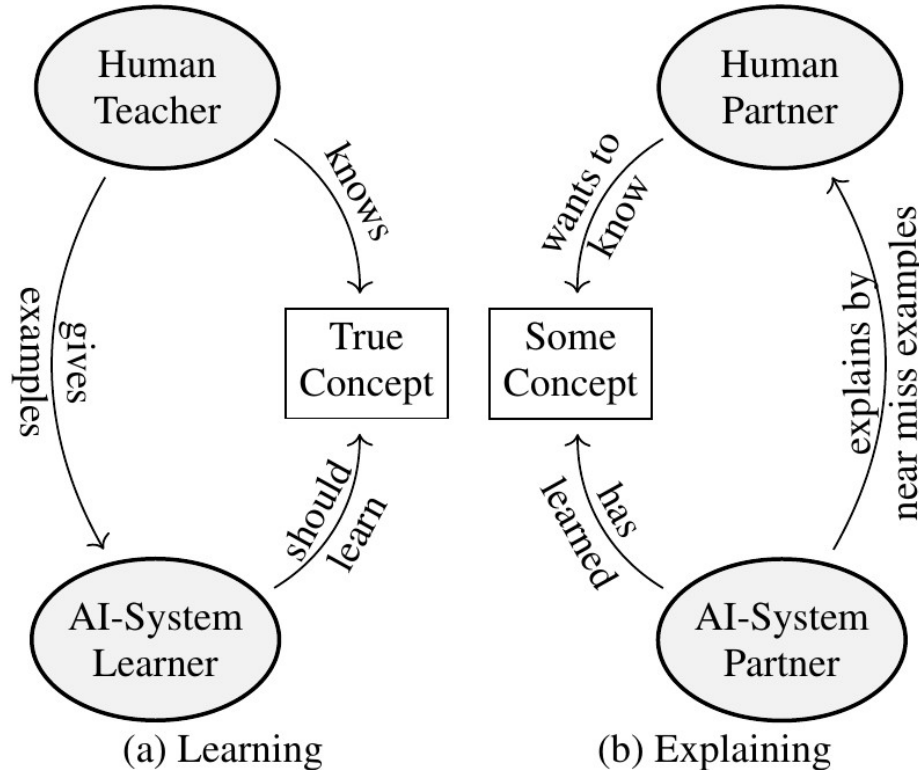
See esp.

XAI & ITS

Nguyen, M. H., Tschitschek, S., & Singla, A. (2023). Large Language Models for In-Context Student Modeling: Synthesizing Student's Behavior in Visual Programming from One-Shot Observation. arXiv preprint arXiv:2310.10690.



# Near Miss Explanations for Effective Learning and Effective Teaching



Patrick Winston, Learning structural descriptions from examples.

MIT/LCS/TR-76, 1970.

## Principles of efficient teaching

Shafto, Goodman, & Griffiths, A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71, 55-89, 2014

Telle, J. A., Hernández-Orallo, J., & Ferri, C. (2019). The teaching size: computable teachers and learners for universal languages. *Machine Learning*, 108(8), 1653-1675.

# Decision Making in Medicine

Activities LearnWithME-v1.py MI 10:28  
CogSys Companion - LearnWithME - version 09/2019

Clause-Level-Constraints

**Gsys**

**TraMeExCo**

| All examples (labeled as learned by a CNN) |          |           | Positive examples |          |           | Negative examples |          |           |
|--|----------|-----------|-------------------|----------|-----------|-------------------|----------|-----------|
| Label                                      | Example  | Facts     | Label             | Example  | Facts     | Label             | Example  | Facts     |
| 1 pT3                                      | scan0523 | Backgr... | 1 pT3             | scan0523 | Backgr... | 1 gesund          | scan0502 | Backgr... |
| 2 pT3                                      | scan0569 | Backgr... | 2 pT3             | scan0569 | Backgr... | 2 gesund          | scan0506 | Backgr... |
|  |          |           |                   |          |           | 3 pT3             | scan0538 | Backgr... |
|  |          |           |                   |          |           | 4 pT3             | scan0562 | Backgr... |

Learn and show model

Learned model

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.  
Rule:  
pT3(A) :-  
contains\_tissue(A,B), is\_tumor(B), touches(B,C),  
is\_fat(C).  
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 muscle.

First rule:  
pT3(scan0523)  
pT3(scan0569)  
Second rule:  
pT3(scan0562)  
pT3(scan0538)

B touches C and C is fascia

Covered negative examples  
No examples covered.

Constraint history

must not occur in explanation

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

Explanation as foundation for  
Interactive ML

HUMAN PARTNERSHIP WITH MEDICAL  
ARTIFICIAL INTELLIGENCE

Association for the Advancement of Artificial Intelligence Fall 2021  
Symposium

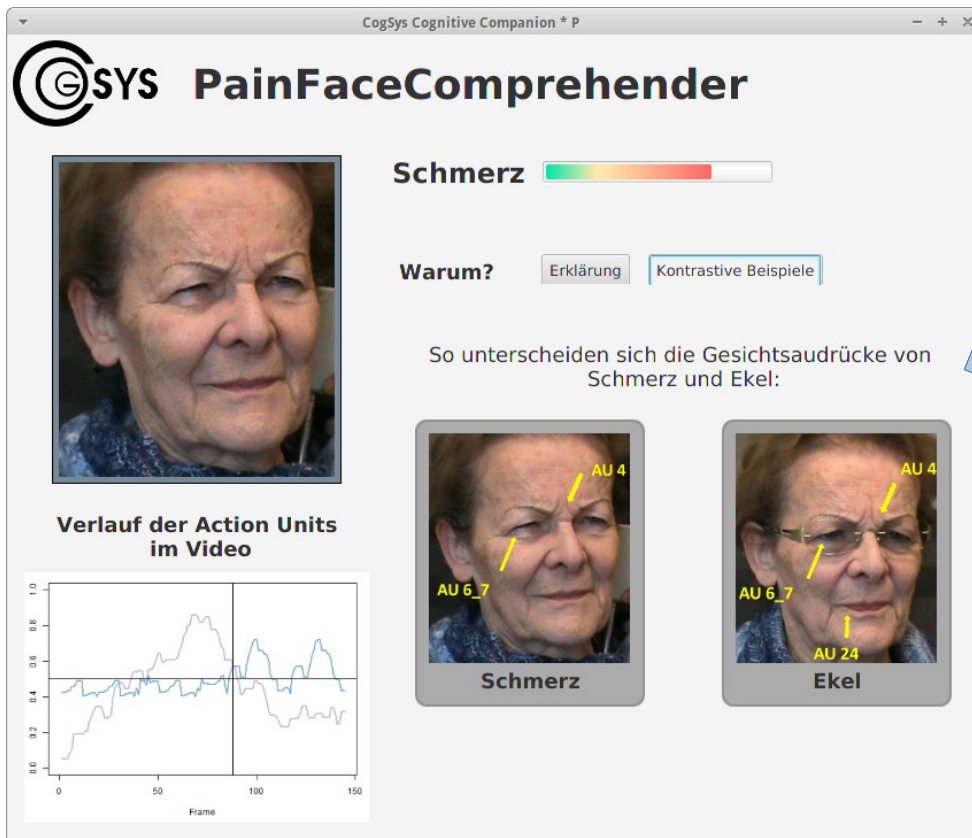
Schmid, U., & Finzel, B. (2020). Mutual explanations  
for cooperative decision making in medicine. KI-  
Künstliche Intelligenz, 34(2), 227-233.

# Near-miss Explanations for interactive ML and ITS

- Near-miss explanations can be used for interactive explanatory ML (e.g., CAIPI by Teso & Kersing)  
taking the human as teacher for an AI system
- Near-miss explanations can also be used for intelligent tutoring systems  
taking the AI system as teacher (ultra-strong ML)

TESO, Stefano; KERSTING, Kristian. Explanatory interactive machine learning. In: Proceedings of the 2019 AAI/ACM Conference on AI, Ethics, and Society. 2019. S. 239-245.

# Educating Nurses



Learned model + explanation  
as input to educate nurses  
= ultra-strong ML

Hassan, T., Seuß, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., ... & Schmid, U. (2019). Automatic detection of pain from facial expressions: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6), 1815-1831.

# Ultra-Strong Machine Learning

Donald Michie (1988):

- **Weak ML:** machine learner produces improved predictive performance with increasing amounts of data
- **Strong ML:** additionally requires the learning system to provide its hypotheses in symbolic form (interpretable machine learning, e.g. Rudin, Nature ML, 2019)
- **Ultra-strong ML:** extends the strong criterion by requiring the learner to teach the hypothesis to a human, whose performance is consequently increased to a level beyond that of the human studying the training data alone

Human to Machine  
Teaching

Machine to Human  
Teaching

# Explanatory Dialog

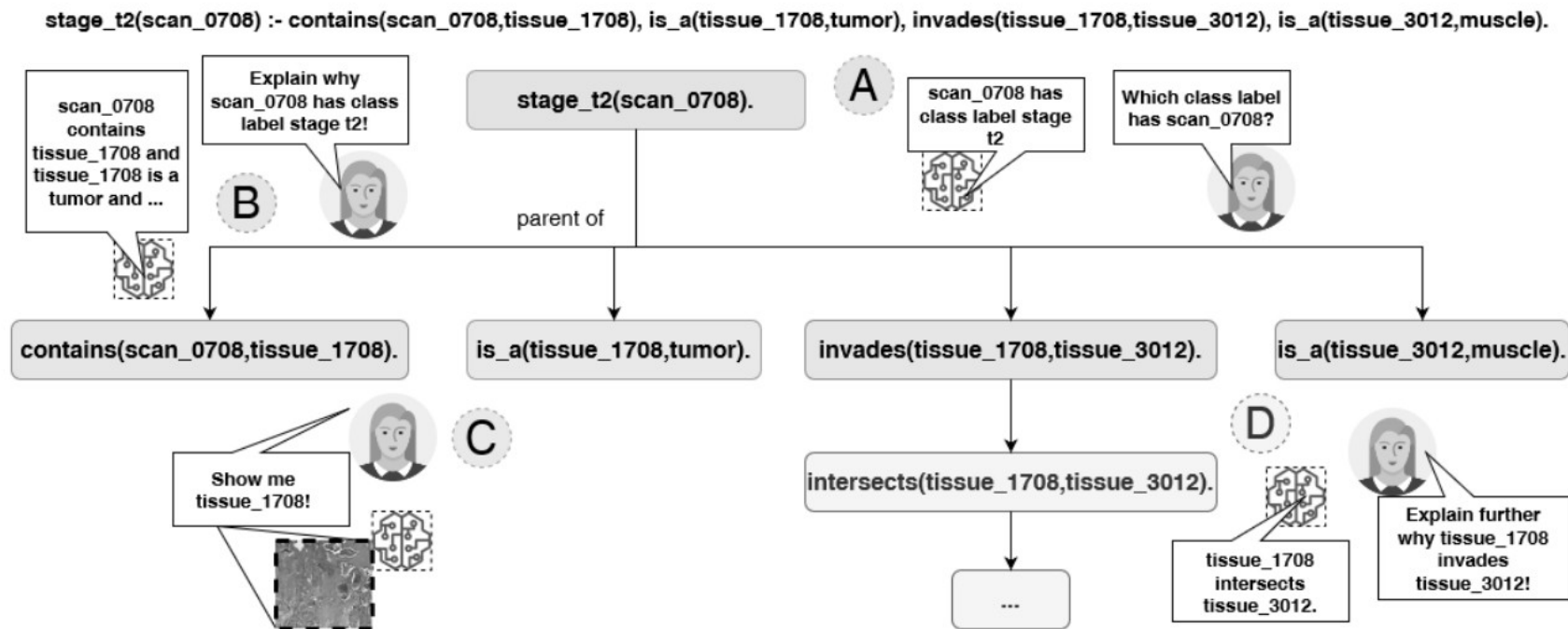


Figure 2: An explanatory tree for *stage\_t2(scan\_0708)*, that can be queried by the user to get a local explanation why scan\_0708 is labeled as T2 (steps A and B). A dialogue is realized by further requests, either to get more visual explanations in terms of prototypes (step C) or to get more verbal explanations in a drill-down manner (step D).

# Wrapping Up

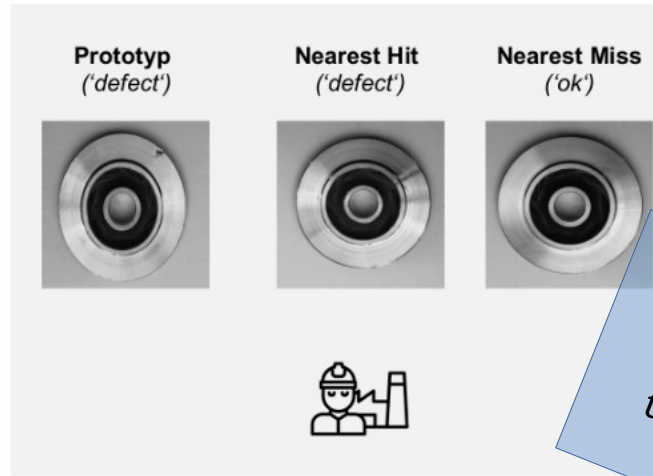
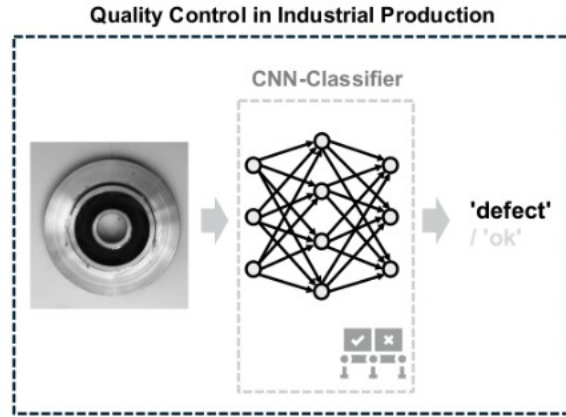
- XAI offers a growing set of approaches to explain ML models but for beneficiary human-AI partnerships, it is important to take into account the specific information need of the human in a given context, here methods from ITS may be helpful
- Human-AI partnerships can be realized in two ways
  - Human as provider of information for model adaptation (XIML)
  - Learned model as provider of new insights (ultrastrong ML)
- Interleaving XAI and ITS might be a way to address both perspectives

Learning without  
thought is labor  
lost Confucius





# Example-based Explainable AI (XAI) Demonstrator



Help the quality engineer to understand classification boundaries of the model to provide helpful examples for model adaptation

Re-implementation of Kim, Khanna, Koyejo: Examples are not Enough – Learn to Criticize!  
Criticism for Interpretability, NeurIPS 2016

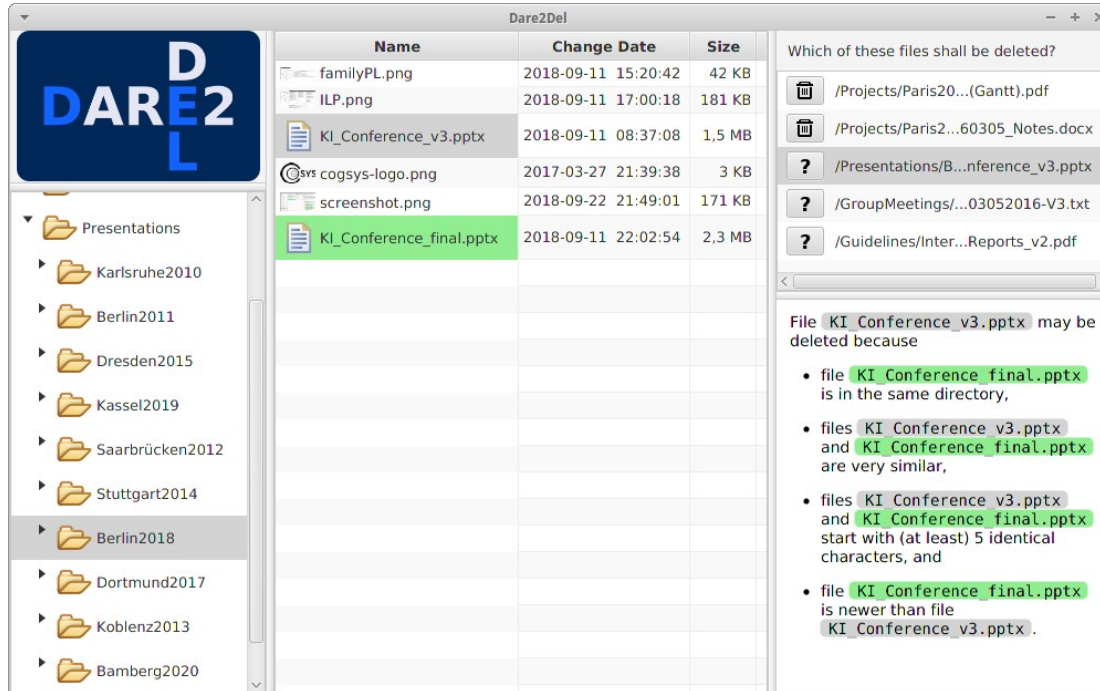
$$\text{MMD}^2(X, Y) := \frac{1}{|X|^2} \sum_{x_1, x_2 \in X} k(x_1, x_2) + \frac{1}{|Y|^2} \sum_{y_1, y_2 \in Y} k(y_1, y_2) - \frac{2}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} k(x, y)$$

Maximum Mean Discrepancy, similarity measure on distributions

Extended to Near Miss Explanations

Herchenbach, Müller, Scheele, & Schmid, Explaining image classifications with near misses, near hits and prototypes. ICPRAI 2022.

# Deleting Irrelevant Files/Data



**DFG** Deutsche  
Forschungsgemeinschaft

What must be minimally changed  
that this file is not classified as  
Irrelevant?

Schmid, U. (2021). Interactive learning with mutual explanations in relational domains. In: S. Muggleton and N. Chater, Human-like Machine Intelligence,(chap.~17). 338-354, OUP.