Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
AGH University of Science
and Technology

# In Search for Model-Driven eXplainable Artificial Intelligence

Antoni Ligęza Dominik Sepioło

AGH University of Science and Technology + KRaKEn Research Group
https://kraken.edu.pl/

22.11.2023

# Kilka uwag na wstępie

AGH

www.agh.edu.pl

- To jest work in progress;
- Bliżej koncepcji niż finalnych rozwiązań,
- Prezentacja bardzo skrótowa i uproszczona,
- Także - tendencyjna; zorientowana na przyjęte tezy,
- Proszę o wyrozumiałość.

# Presentation Outline

AGH

# On AI. Towards XAI

# What is Artificial Intelligence?

## Artificial Intelligence (AI)

The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience (B.J. Copeland).

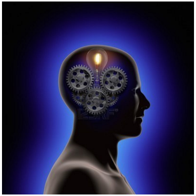
# Definitions of Artificial Intelligence

**Intelligence** = ability to solve new problems

- Artificial Intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals. [1]
- Artificial Intelligence, or AI, is the field that studies the synthesis and analysis of computational agents that act intelligently. [2]
- Artificial Intelligence: `http://aima.cs.berkeley.edu/`:
  - Systems that think like humans;
  - Systems that act like humans;
  - Systems that think rationally;
  - Systems that act rationally;
- Artificial Intelligence = Technology of Machine Thinking
- Artificial Intelligence $\neq$ Computational Intelligence (CI)
- Artificial Intelligence $\simeq$ Algorithmic Intelligence
- Can pass the Turing Test for Artificial Intelligence

# Artificial Intelligence for Problem Solving. What is the Essence of it?





The key issue:
Black-Box Models (hidden knowledge) versus
White-Box Models (explicit knowledge): Knowledge Representation and Reasoning – open, declarative/procedural, undergo analysis, design, verification; reliable and trustable solutions.

# Taxonomy of AI methods: Diversified and Specilized

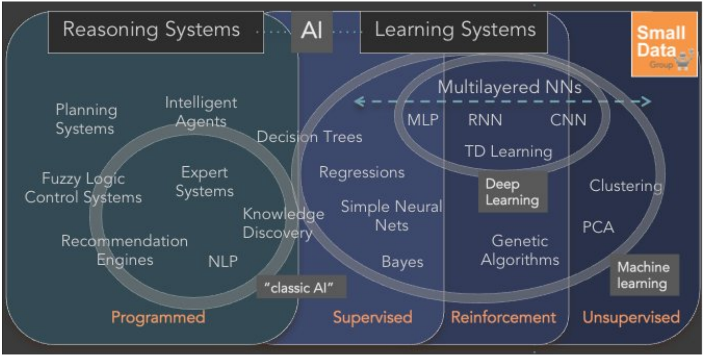



Figure: Source: Small Data Group

# AI-KRR: Basic Ideas behind this Course

Key observation: There is no single, complete, consistent and uniform AI.

- to teach various, but selected methods of AI Knowledge Representation (KR),
- to teach various, but selected methods of AI Automated Reasoning (AR),
- with the focus on Symbolic Knowledge (Logical, Algebraic, Graph-Based)
- with the ultimate goal: Automated Problem Solving (APS);

$$\mathrm{KR} + \mathrm{AR} + \mathrm{Control} \longrightarrow \mathrm{APS}$$

- to keep the course practical rather than *just theory*:
  - some background knowledge — but in an informal way,
  - modern tools — if available (Prolog, MiniZinc, Problog, PDDL, Picat, Logica,...Python, Julia),

# Human Intelligence vs. Artificial Intelligence

- Curiosity - human tend to observe, explore and explain the environment;
- Flexibility - single brain explores a wide spectrum of problems;
- Model Discovery - human tend to understand the problem nature and build a model;
- Combining Knowledge - commonsense and domain knowledge is combined;
- eXplanation - ability to discuss, explain, reconsider and modify the analysis.

- Must be initiated - started by man, event, clock, another program, ...
- Specialization - AI is task oriented, tools are efficient but of narrow domains;
- Pre-programmed Knowledge Processing - applied to solve a particular problem,
- Machine Learning - purely syntax-based, mechanical decision engines;
- Shallow eXplanation - if present, it is based on the same data as used for learning.

# Black Box

In science, computing, and engineering, a **black box** is a device, system, or object which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings. Its implementation is opaque or “black”.

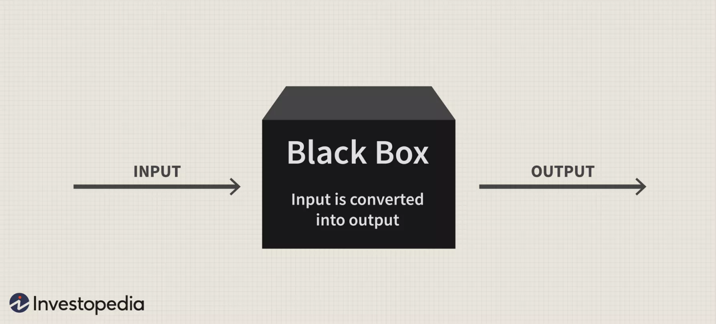
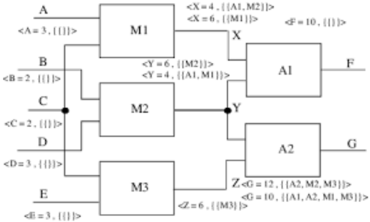



Figure: Source: Black Box Model in Investopedia

# Model-Based Reasoning: Logical Model




## Declarative Modeling Components & Causal Structure

$\mathrm{ADD}(x) \wedge \neg\mathrm{AB}(x) \Rightarrow \mathit{Output}(x) = \mathit{Input}1(x) + \mathit{Input}2(x)$,
$\mathrm{MULT}(x) \wedge \neg\mathrm{AB}(x) \Rightarrow \mathit{Output}(x) = \mathit{Input}1(x) * \mathit{Input}2(x)$,
$\mathrm{ADD}(a1)$, $\mathrm{ADD}(a2)$, $\mathrm{MULT}(m1)$, $\mathrm{MULT}(m2)$,
$\mathrm{MULT}(m3)$,
$\mathrm{Output}(m1) = \mathrm{Input1}(a1)$, $\mathrm{Output}(m2) = \mathrm{Input2}(a1)$,
$\mathrm{Output}(m2) = \mathrm{Input1}(a2)$, $\mathrm{Output}(m3) = \mathrm{Input2}(a2)$,
$\mathrm{Input2}(m1) = \mathrm{Input1}(m3)$,
$\mathit{Input}1(m1) = A \ldots \mathit{Output}(a2) = G$

# Taxonomy of AI methods in ML

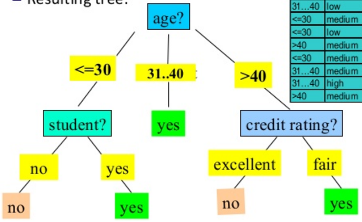


- **White-box models:**
  - Logical Models; declarative KRR,
  - Rule Based Models/Learners, Rule Induction
  - Model-Based Reasoning, Causal Modelling, Structure Discovery, Qualitative Physics, ...
- **Semi-White-box models:**
  - Linear/Logistic Regression
  - Decision Trees
  - Decision Graphs (including BPMN)
- **Black-box models:**
  - Tree Ensembles
  - Computational Intelligence and Bio-Inspired Models
  - Multi–layer Neural Network
  - Convolutional Neural Network
  - Deep Learning

# IS AI just ML: Induction of Trees or Rules

## Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:



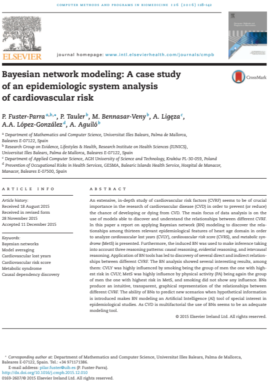

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

10

Problem: shallow knowledge $\Longrightarrow$ Does work — but why?

# Better Model: Bayes Nets *Causal* Model ?

## A further step on...



COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 126 (2016) 128–142

ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

**Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk**

CrossMark

*P. Fuster-Parra*[a,b,*], *P. Tauler*[b], *M. Bennasar-Veny*[b], *A. Ligęza*[c], *A.A. López-González*[d], *A. Aguiló*[b]

[a] *Department of Mathematics and Computer Science, Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain*
[b] *Research Group on Evidence, Lifestyles & Health, Research Institute on Health Sciences (IUNICS), Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain*
[c] *Department of Applied Computer Science, AGH University of Science and Technology, Kraków PL-30-059, Poland*
[d] *Prevention of Occupational Risks in Health Services, GESMA, Balearic Islands Health Service, Hospital de Manacor, Manacor, Baleares E-07500, Spain*

ARTICLE INFO

*Article history:*
Received 18 August 2015
Received in revised form
28 November 2015
Accepted 11 December 2015

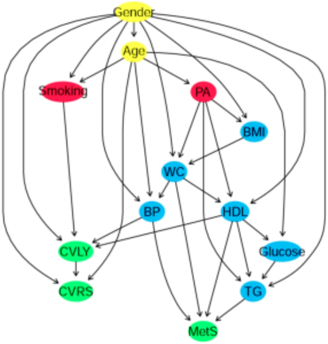
*Keywords:*
Bayesian networks
Model averaging
Cardiovascular lost years
Cardiovascular risk score
Metabolic syndrome
Causal dependency discovery

ABSTRACT

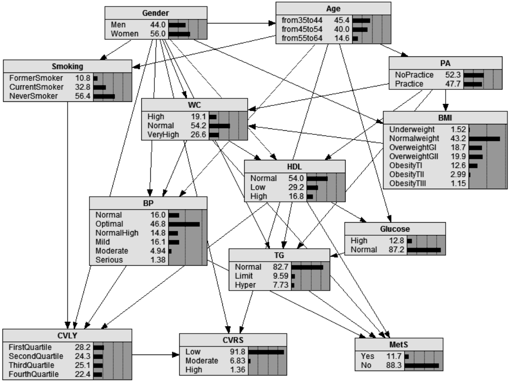
An extensive, in-depth study of cardiovascular risk factors (CVRF) seems to be of crucial importance in the research of cardiovascular disease (CVD) in order to prevent (or reduce) the chance of developing or dying from CVD. The main focus of data analysis is on the use of models able to discover and understand the relationships between different CVRF. In this paper a report on applying Bayesian network (BN) modeling to discover the relationships among thirteen relevant epidemiological features of heart age domain in order to analyze *cardiovascular lost years* (CVLY), *cardiovascular risk score* (CVRS), and *metabolic syndrome* (MetS) is presented. Furthermore, the induced BN was used to make inference taking into account three reasoning patterns: *causal* reasoning, *evidential* reasoning, and *intercausal* reasoning. Application of BN tools has led to discovery of several direct and indirect relationships between different CVRF. The BN analysis showed several interesting results, among them: CVLY was highly influenced by smoking being the group of men the one with highest risk in CVLY; MetS was highly influence by physical activity (PA) being again the group of men the one with highest risk in MetS, and smoking did not show any influence. BNs produce an intuitive, transparent, graphical representation of the relationships between different CVRF. The ability of BNs to predict new scenarios when hypothetical information is introduced makes BN modeling an Artificial Intelligence (AI) tool of special interest in epidemiological studies. As CVD is multifactorial the use of BNs seems to be an adequate modeling tool.



* *Corresponding author at*: Department of Mathematics and Computer Science, Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain. Tel.: +34 971171386.
E-mail address: pilar.fuster@uib.es (P. Fuster-Parra).
http://dx.doi.org/10.1016/j.cmpb.2015.12.010

# Bayes Nets: Even More Precise Model



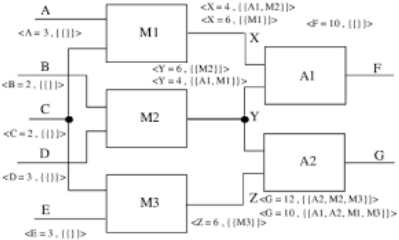
# Towards Model-Driven Explainability

## Model Discovery. Motivation:

- majority of ML models cover *shallow* knowledge only,
- most of them are of decision/classification type; no *functional* output,
- no investigation of the *guts* — what is inside?
- Eternal question: How does it work? No answer...

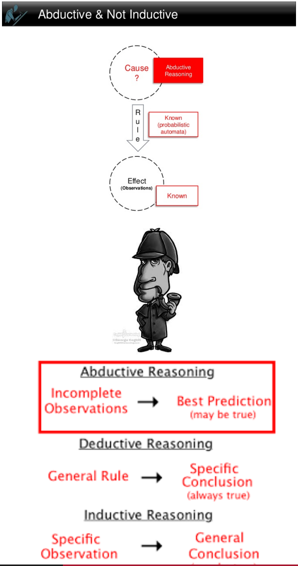
---

- variables, values,
- Components,
- Causal links,
- Connections - struCture,
- input — internal state — output,
- functionality.



# The Role of Abduction and Creativity

- abduction: what, why and where — what for?
- abduction: investigation of causality,
- abduction: a method of logical inference (but invalid),
- abduction vs. deduction,
- abduction: primary method used by Sherlock Holmes!
- abduction: inevitable ambiguity (potential/admissible solutions; many of them),
- abduction: more constraints — better abduction,



AGH

www.agh.edu.pl

# On XAI. Towards Model-Driven XAI

# XAI: Explainable Artificial Intelligence – WHY?





Why do we care about **Explainability in Artificial Intelligence?**

- readable and interpretable models,
- reduction of models (knowledge is more concise than data),
- well-defind domain of application,
- predictable behavior,
- Reliable Artificial Intelligence,
- safe AI solutions,
- easy adaptation and modification,
- Trustable Artificial Intelligence,
- ...

# XAI: Explainable Artificial Intelligence – HOW?



Direct approaches to **Explainable in Artificial Intelligence** (in contrast to a posteriori explanation mechanisms).

- declarative programming; Prolog,
- rule-based systems (Why, How, What-is questions),
- automated deduction; the Resolution Method,
- automated planning systems; STRIPS, PDDL,
- Bayes Networks, causal graphs,
- Model-Based Reasoning (MBR),
- Model Checking,
- Abductive vs. Inductive reasoning,
- Explainability by Design,
- Constraint Programming, Functional Constraint Programming,
- ...

# What is XAI NOW?

AGH

www.agh.edu.pl

## Explainable Artificial Intelligence

XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

**Crucial problems:**

- ML is not AI! AI is not ML!
- no background/commonsense/domain knowledge allowed!
- explanations a created (i) a posteriori (ii) on the base of the same data!

# Hot topic

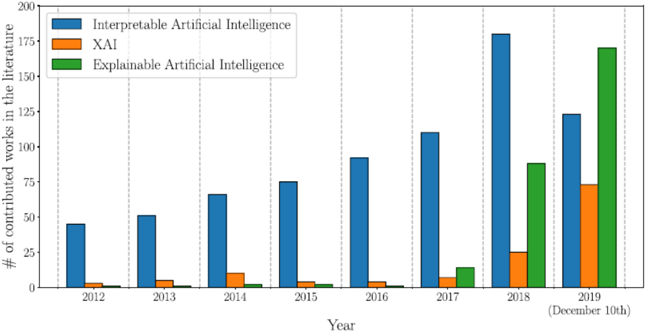



Figure: Source: XAI: Concepts, Taxonomies...

# Who needs explanations?

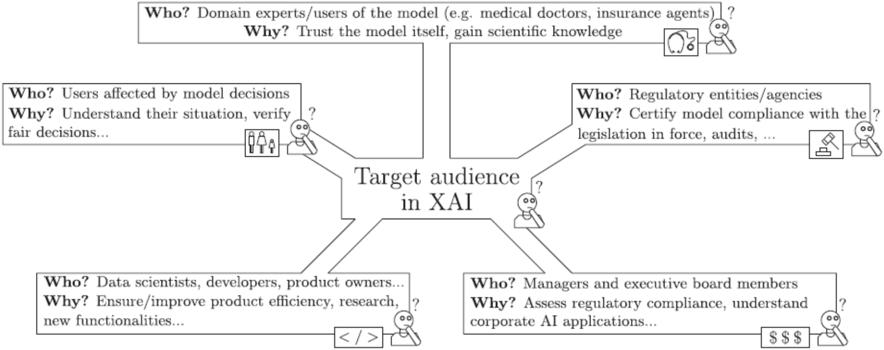



Figure: Source: XAI: Concepts, Taxonomies...

# Explainability workflow

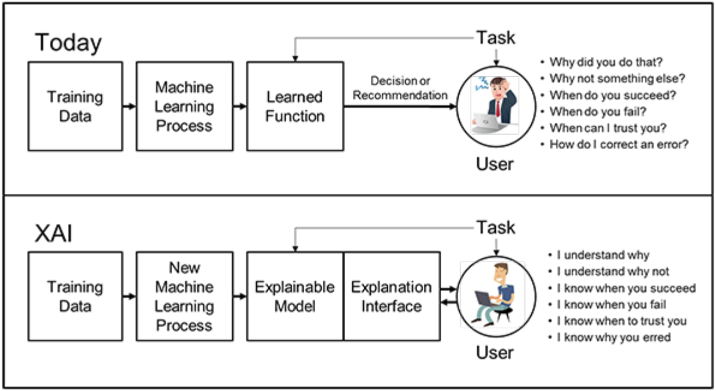



Figure: Source: A brief explanation of XAI, DARPA

# Accuracy vs Explainability

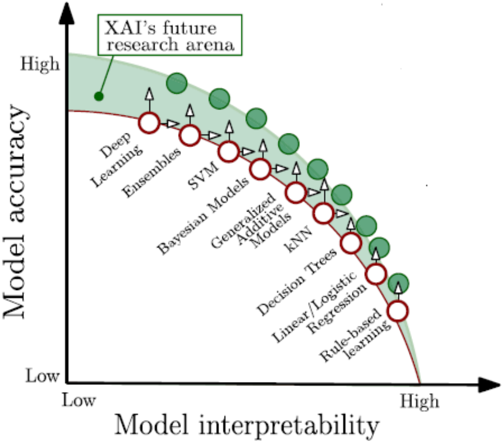



Figure: Source: XAI: Concepts, Taxonomies...

# How to explain black-boxes?

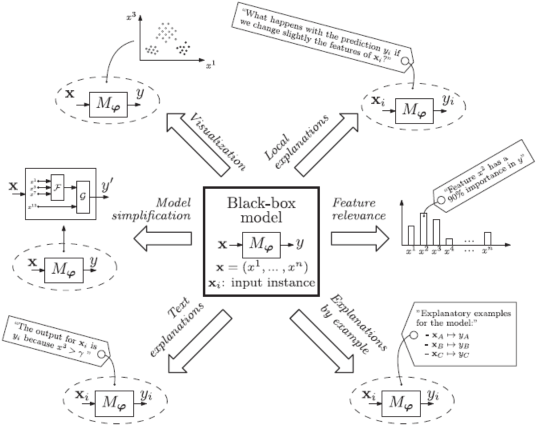



Figure: Source: XAI: Concepts, Taxonomies...

# XAI Approaches

- XAI in AI/ML
  - Transparent Models
    - XAI by Design
    - XAI by Model-Discovery
  - Post-Hoc Explainability
    - Model-Agnostic
    - Model-Specif

# Goal of XAI?

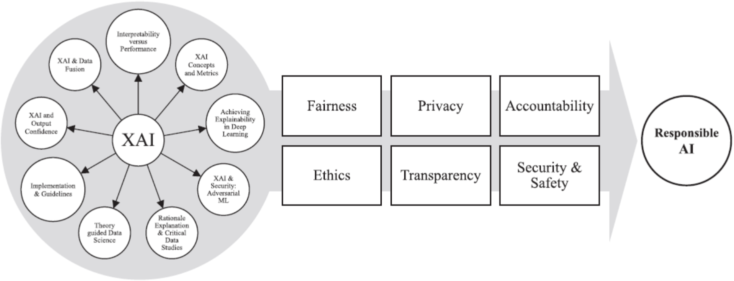



Figure: Source: XAI: Concepts, Taxonomies...

www.agh.edu.pl

# The SiCA Concept. 10 Examples

# XAI: Model-Driven Approach – WHY?



**Explainability in Artificial Intelligence** – observations:

- often only limited, and incomplete data is available,
- on the other hand, some domain knowledge is present:
  - on Components,
  - on Connections,
  - on Causal Dependencies;
- partial inspection of the system is possible,
- abductive reasoning is used instead of induction,
- creative reasoning seems to be the key,
- domain knowledge is available,
- commonsense knowledge plays a role.

In such a case, ML is out of play. Such a case will be referred to as SiCA: Singular Case Analysis. Some 10 examples follow...

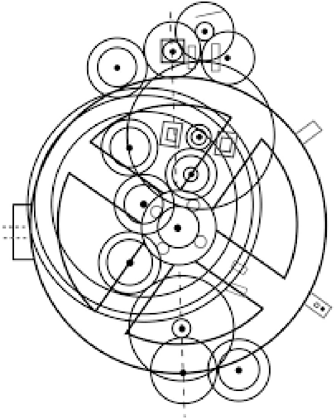
The eternal question: How does it work? is answered with Model-Based Reasoning.

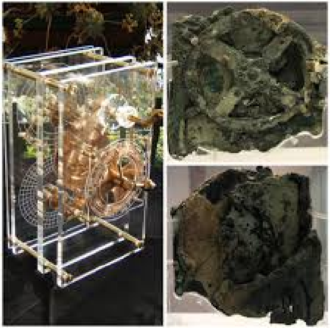
# An Eternal Question: How Does it Work?



Figure: The Antikythera mechanism; recovered on May 17, 1901. The instrument has been variously dated to about 87 BC, or between 150 and 100 BC, or in 205 BC
https://en.wikipedia.org/wiki/Antikythera_mechanism

# How Does it Work? Model-Based Reasoning





Components + Connections + Causality = Operation

# An Eternal Question: How Does it Work?

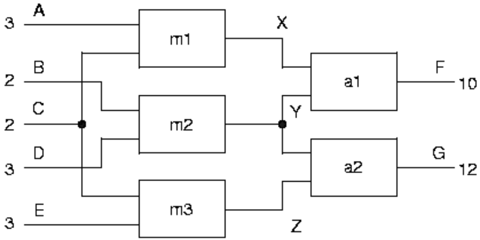



Figure: XAI: Model-Driven eXplainable AI System

# An Eternal Question: How Does it Work?

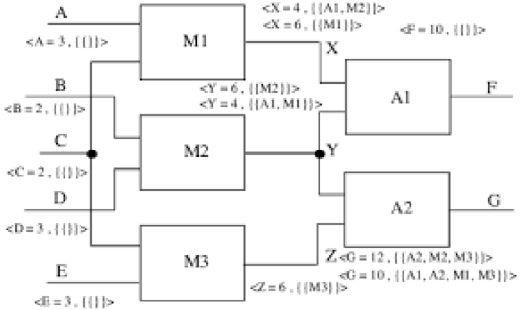
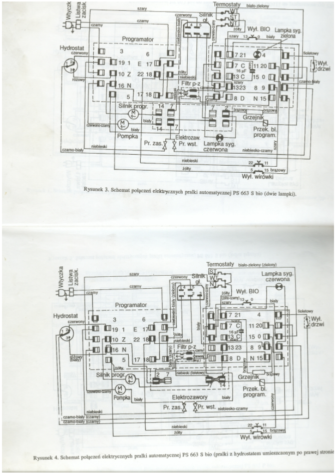



Figure: XAI: Explanatory Reasoning in Case of Fault

# An Eternal Question: How Does it Work?




Components + Connections + Causality = Operation

System: mechanical, hydraulic, pneumatic, electric, electromechanical.

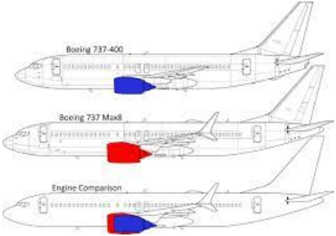
# An Eternal Question: How Does it Work?



Components + Connections + Causality = Operation

System: mechanical, hydraulic, pneumatic, electric, electromechanical.

# An Eternal Question: How Does it Work?




Components + Connections + Causality = Operation

Boeing 737 Max 8: Introduced into exploitation 2017

1. Crash: October 2018, Lion Air (Malaysia)
2. Crash: March 2019, Ethiopian Airlines

Explanation: MCAS System (Maneuvering Characteristics Augmentation System).

# An Eternal Question: How Does it Work?

# An Eternal Question: How Does it Work?


Components + Connections + Causality = Operation

The viaduct was built between 1963 and 1967 costing 3.8 billion Italian lire and opened on 4 September 1967. It had a length of 1,182 meters (3,878 ft), a height above the valley of 45 meters (148 ft) at road level, and three reinforced concrete pylons reaching 90 meters (300 ft) in height; the maximum span was 210 meters (690 ft). On 14 August 2018, a 210-metre (690 ft) section of the viaduct collapsed during a rainstorm, killing forty-three people.

# An Eternal Question: How Does it Work?

| **Input** | **Output** |
|---|---|
| 1 | 5 |
| 2 | 10 |
| 3 | 15 |
| 10 | ??? |

- Find the missing value;
- Find the rule.

# An Eternal Question: How Does it Work?

AGH



| **Input** | **Output** |
|---|---|
| 1 | 5 |
| 2 | 10 |
| 3 | 55 |
| 10 | ??? |

- Find the missing value;
- Find the rule.

From: MindYourDecisions by Presh Talwalkar (`https://www.youtube.com/watch?v=Pb7N6wqhjhg`)

# An Eternal Question: How Does it Work?

| Input | Output |
|---|---|
| 1 | 5 |
| 2 | 10 |
| 3 | 55 |
| 10 | 1490 |

- The missing value is given;
- Find the rule.

From: MindYourDecisions by Presh Talwalkar (https://www.youtube.com/watch?v=Pb7N6wqhjhg)

# An Eternal Question: How Does it Work?

| Input | Output |
|---|---|
| 1 | 5 |
| 2 | 10 |
| 3 | 55 |
| 10 | 1490 |

The output is given by:

$$f(x) = ax^2 + bx + c$$

where: $a = 20$, $b = -55$, $c = 40$.

This explanation is rational, correct, complete, minimal.

From: MindYourDecisions by Presh Talwalkar (https://www.youtube.com/watch?v=Pb7N6wqhjhg)

# An Eternal Question: How Does it Work?

Find all integer functions $f$ such that:

$$f(2a) + 2f(b) = f(f(a+b))$$

- This is a Functional Equation;
- In fact, one looks for Components, Connections and Causality (well, functional dependencies) = the 3C principle;
- From AI point of view this is a Model Discovery task.

---

From: MindYourDecisions by Presh Talwalkar (`https://www.youtube.com/watch?v=uJqbHaFqjmI`)

# AI & XAI Failures

# Chihuahua or Muffin?

Figure: Source:
https://skywell.software/blog/top-artificial-intelligence-fails/

# Husky or wolf?

A husky (on the left) is confused with a wolf, because the pixels (on the right) characterizing wolves are those of the snowy background. This artifact is due to a learning base that was insufficiently representative

Figure: Source: Can Everyday AI be Ethical? Machine Learning Algorithm Fairness

# Shallow Rule Induction – A Naive Example

| Car color | Car turns |
| --- | --- |
| red | left |
| red | left |
| ⋮ | ⋮ |
| black | right |
| black | right |
| ⋮ | ⋮ |

$$Car\_color = red \longrightarrow Car\_turns = left$$

$$Car\_color = black \longrightarrow Car\_turns = right$$

But why ???

Problem: no semantics; no background knowledge.

# Shallow Rule Induction – A Naive Example

| Car color | Car turns |
| --- | --- |
| red | left |
| red | left |
| ⋮ | ⋮ |
| black | right |
| black | right |
| ⋮ | ⋮ |

$$Car\_color = red \longrightarrow Car\_turns = left$$

$$Car\_color = black \longrightarrow Car\_turns = right$$

```
car_turns(X,left) :- drives(X,university).
car_turns(X,right) :- drives(X,court).
drives(X,university) :- young(X).
drives(X,court) :- old(X).
young(X) :- write(X),
            write(' is young and preferes red cars.').
old(X) :- write(X),
          write(' is old and preferes black cars.').
```

www.agh.edu.pl

# In Search for Model-Driven XAI. Towards Model-Discovery

# Model-Driven XAI - The Concept

Model-Driven eXplainable Artificial Intelligence – to be:

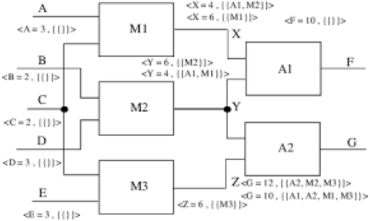
- transparent, readable and interpretable models,
- visible knowledge component,
- maybe simplified and therefore inaccurate,
- but still useful,
- well-defined domain of application,
- predictable behavior,
- Reliable Artificial Intelligence,
- safe AI solutions,
- flexible - an so easy adaptation and modification,
- The 3C Principle:
  - Components,
  - Connections,
  - Causality
- Functional Characteristics – if available.

# XAI: Explainable Artificial Intelligence – HOW?

Direct approaches to **Model-Driven XAI**;
(in contrast to a posteriori explanation mechanisms).

- declarative programming; Prolog,
- rule-based systems (Why, How, What-is questions),
- automated deduction; the Resolution Method,
- automated planning systems; STRIPS, PDDL,
- Bayes Networks, Causal Graphs,
- Model-Based Reasoning (MBR),
- Model Checking,
- Abductive vs. Inductive reasoning,
- Explainability by Design,
- Declarative Programming, Constraint Programming, Logic Programming,
- ...

# Model-Based Reasoning: Logical Model




## Declarative Modeling Components & Causal Structure

$\mathrm{ADD}(x) \wedge \neg\mathrm{AB}(x) \Rightarrow \mathit{Output}(x) = \mathit{Input}1(x) + \mathit{Input}2(x)$,
$\mathrm{MULT}(x) \wedge \neg\mathrm{AB}(x) \Rightarrow \mathit{Output}(x) = \mathit{Input}1(x) * \mathit{Input}2(x)$,
$\mathrm{ADD}(a1)$, $\mathrm{ADD}(a2)$, $\mathrm{MULT}(m1)$, $\mathrm{MULT}(m2)$,
$\mathrm{MULT}(m3)$,
$\mathrm{Output}(m1) = \mathrm{Input1}(a1)$, $\mathrm{Output}(m2) = \mathrm{Input2}(a1)$,
$\mathrm{Output}(m2) = \mathrm{Input1}(a2)$, $\mathrm{Output}(m3) = \mathrm{Input2}(a2)$,
$\mathrm{Input2}(m1) = \mathrm{Input1}(m3)$,
$\mathit{Input}1(m1) = A \ldots \mathit{Output}(a2) = G$

# Model Discovery: If Not Hand-Coded - Then Guided Search

- *Ancient Math:*
  - Differential Equations,
  - Variational Calculus,
  - Functional Equations,
- Inductive Logic Programming,
- Symbolic Regression,
- Grammatical Evolution,
- Functional Constraint Programming,
- Bayes Nets Discovery,
- ...

# Selected Tools

Symbolic Regression[3] – 11 tools, e.g.:

- uDSR (Deep Symbolic Optimization)
- QLattice (a quantum-inspired simulation and machine learning technology)
- geneticengine (Genetic Engine)
- PySR – Python Symbolic Regression,
- ...

Grammatical Evolution[4] – 12 tools, e.g:

- GELab (Matlab),
- PonyGE2 (Python),
- gramEvol (R),
- ...

[3] https://en.wikipedia.org/wiki/Symbolic_regression
[4] https://en.wikipedia.org/wiki/Grammatical_evolution

# XAI by Now: LIME

$$\text{explanation}(x) = \arg\min_{g \in G} L\left(f, g, \pi_x\right) + \Omega(g)$$

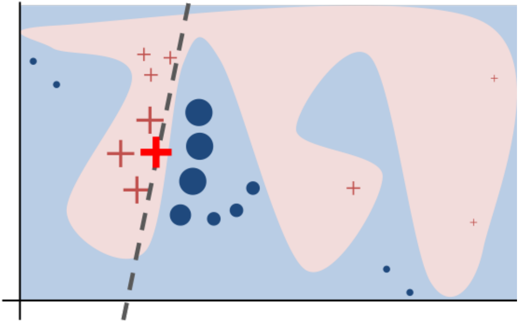


Figure: LIME explanation
source: github.com/marcotcr/lime

# XAI by Now: SHAP

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

SHAP

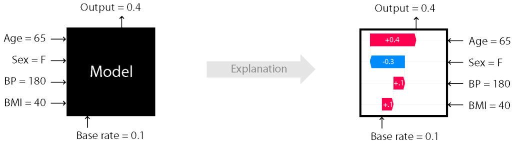



Figure: SHAP explanation
source: github.com/slundberg/shap

# Motivation

AGH

- **Limitations of current XAI task formulation:**
  - No holistic, general framework; only local peep-hole view,
  - No Causal Models; lack of structural approach,
  - The explanation is based on the same data as learning!,
  - No deep, background knowledge is taken into account.
- Towards **Model-Driven XAI**: Introduction of Knowledge-Based Component
- Trustworthy Decision-Making
- Building Model-Based (Model-Driven) XAI techniques

# Grammatical Evolution


Figure: Key research areas in GE
source: Handbook of Grammatical Evolution

# Simple Experiment

- A series of experiments in order to create models that calculate the BMI function
- Decision Tree, Random Forest, Grammatical Evolution
- 10, 50, and 100 observations

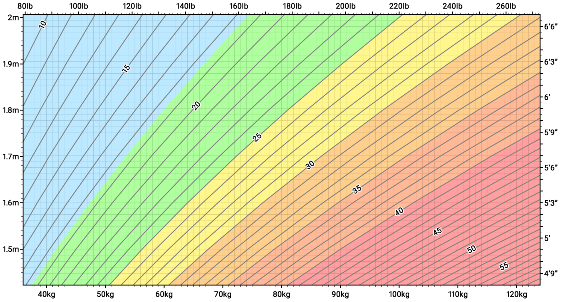
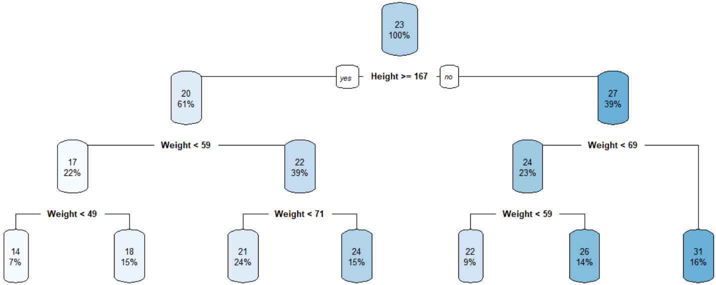



Figure: BMI chart
source: Wikipedia

# Decision Tree



# Random Forest

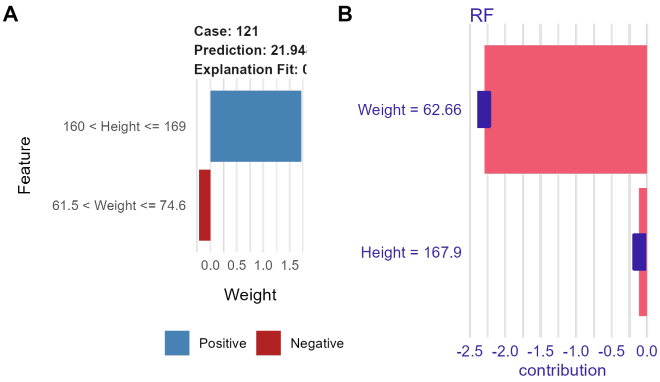



Figure: LIME (A) and SHAP (B) explanation.

# Grammatical Evolution

The Proposed Grammar:

```
<expr>  ::= <op>(<expr>, <expr>) | <func>(<expr>) | <var>
<func>  ::= 'log' | 'sqrt'
<op>    ::= "+" | "-" | "*" | "/" | "^"
<var>   ::= Weight | Height | <n>
<n>     ::= -3 | -2 | -1 | 0 | 1 | 2 | 3
```

The Best Expression: $\mathtt{Weight} * \mathtt{Height}^{-2}$

# A Second Test Example

Table: A schematic presentation of the input data. The table contains a sample of the data used for experiments.

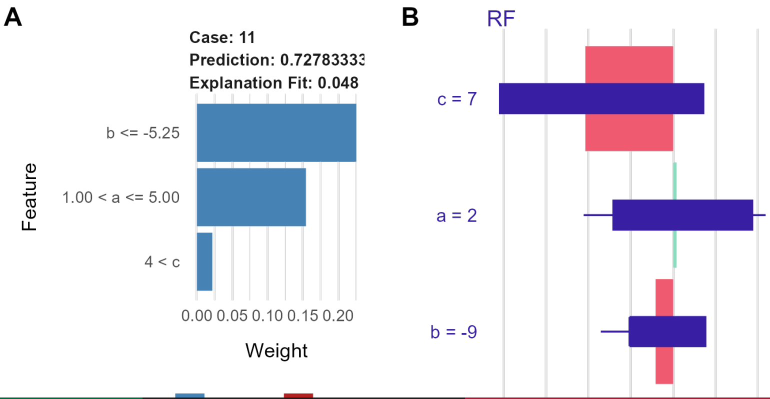
| a | b | c | decision |
|---|---|---|---|
| 3 | 9 | 7 | 0 |
| 1 | 4 | 4 | 1 |
| 5 | 9 | 4 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ |

Table: RMSE for selected ML techniques.

| | RMSE train | RMSE test |
|---|---|---|
| Linear Regression | 0.8267 | 0.8239 |
| Decision Tree | 0.5431 | 0.8142 |
| Random Forest | 0.2284 | 0.5883 |

# Frame Title

| | RMSE train | RMSE test |
|---|---|---|
| Linear Regression | 0.8267 | 0.8239 |
| Decision Tree | 0.5431 | 0.8142 |
| Random Forest | 0.2284 | 0.5883 |



# A Knowledge Component Supplied

Table: A Meaningful Intermediate Variable (MIV) added.

| a | b | c | d | decision |
|---|---|---|---|---|
| 3 | 9 | 7 | -3 | 0 |
| 1 | 4 | 4 | 0 | 1 |
| 5 | 9 | 4 | 16 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Table: RMSE for selected ML models with KC.

| | RMSE train | RMSE test |
|---|---|---|
| Linear Regression | 0.4778 | 0.5779 |
| Decision Tree | 0.0000 | 0.0000 |
| Random Forest | 0.0386 | 0.0266 |

# A GE Model

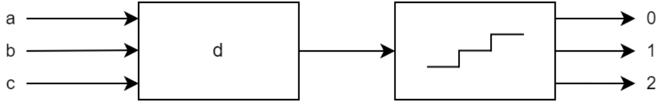
```
<expr> ::= <var> | <op>(<n>, <var>) | <op>(<var>,
              | <op>(<expr>, <expr>)
  <op>   ::= "-" | "*" | "^"
  <var>  ::= a | ‘c‘ | b
  <n>    ::= 1 | 2 | 3 | 4
```

```
Best Expression for d:  b^2 - c * 4 * a
```

```
The rules:

    ifelse(d > 0, "two", "other no. of roots")
    ifelse(d == 0, "one", "other no. of roots")
    ifelse(d < 0, "zero", "other no. of roots")
```

# Final Model-Driven eXplanation

```
Best Expression for d:  b^2 - c * 4 * a

ifelse(d > 0, "two", "other no.  of roots")
ifelse(d == 0, "one", "other no.  of roots")
ifelse(d < 0, "zero", "other no.  of roots")

ifelse(d > 0, "two", ifelse ( d == 0 , "one",
"zero"))
```



AGH

# Summary and What Next?

www.agh.edu.pl

# Observations

AGH

- **Limitations of current XAI task formulation:**
  - No holistic, general framework; only local, peep-hole view,
  - No Causal Models; lack of structural approach,
  - The explanation is based on the same data as learning!; an intrinsic, born-in limitation,
  - No deep, background knowledge is taken into account.
- Towards **Model-Driven XAI**: Introduction of Knowledge-Based Components, Connections, Causality,
- Reliable, Trustworthy Decision-Making,
- Incorporation of Model-Based (Model-Driven) XAI techniques.

# Summary

- Using only shallow explainability techniques can lead to inconsistent and misleading explanations, having nothing to do with the real, Model-Driven understanding of the decision procedure,
- Simple decision models but deep explainability techniques should be preferred,
- Grammatical Evolution can be applied for identification of functional dependencies in data,
- Perhaps the 3C approach seems reasonable for trustworthy XAI:
  - Components identification (also functional ones),
  - Causality – causal dependencies discovery,
  - Connections – struCture identification.

# Further Work

AGH

- Larger data sets... The A case,
- Structure Discovery: combined KB + automated,
- Incorporation of Automated Planning Techniques,
- Incorporation of Functional Constraint Programming,
- Combination of Logical, Causal, Functional and Probabilistic Models.
- Combination with LLM; perhaps some synergy effect?

# Challenges of XAI

AGH

## Some current research trends

- New research field – promising – trustability,
- Explanation quality measures; accuracy vs. simplicity,
- Explanation discrepancies – exceptions,
- Complexity of Model-Generation,
- Implications for Law.

## Alternative Approaches: Deep Models vs. Shallow Ones

- eXplainability by Design,
- eXplainability by Structure Discovery,
- using Model-Based Reasoning,
- Knowledge Graph Models, Functional Constraint Programming, Constructive Abduction,...

# Thank you for your kind attention! Questions? Remarks? Constructive critics?