FONDAZIONE BRUNO KESSLER
ICT
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

Achieving Explainable AI Through Semantic Technologies: Challenges and Future Directions in Digital Health

Mauro Dragoni

Fondazione Bruno Kessler
Process and Data Intelligence Research Unit
Health and Wellbeing High Impact Initiative

Fondazione Bruno Kessler, Trento, Italy
September 18th, 2020

today's agenda

1. Explainable AI (XAI)
- what does "Explainable" mean ?
- history

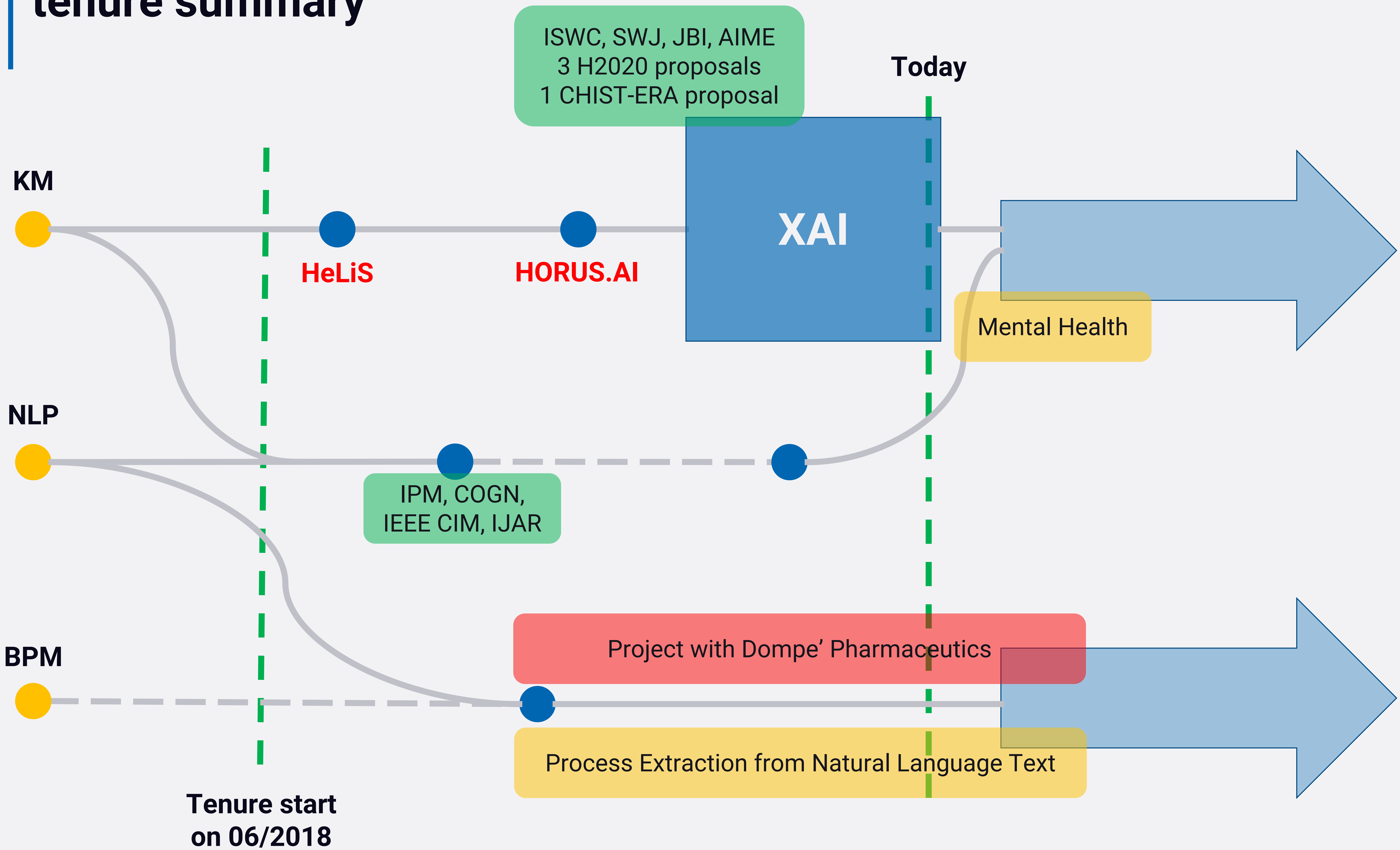
2. XAI in Digital Health
- motivations
- the problem of trust

3. XAI in my tenure
- how semantic technologies can support XAI
- food recognition
- recommendations

4. Final Remarks
- take-home messages
- open challenges
- wrap-up

FONDAZIONE BRUNO KESSLER | ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

tenure summary



FONDAZIONE BRUNO KESSLER
ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

the main question

Is Explainable AI the enabler for adopting artificial intelligence within several domains for supporting our daily lives?

XAI

*What is an **explainable system**, which are its **requirements** and how the research community is working on them?*

explainable AI
a very brief history

Newell et al., 1958

Inference tracing

First explanations

Shortliffe and Buchanan, 1975
Swartout et al., 1991
Johnson, 1994
Lacave and Diez, 2002

Core et al., 2006

“new wave” on explainability

Domain integration

Pearl, 2009
Gershman et al., 2015

Lipton, 2016
Kieseberg et al., 2016
Doran et al., 2017
Montavon et al., 2017
Miller et al., 2017

State of the art techniques

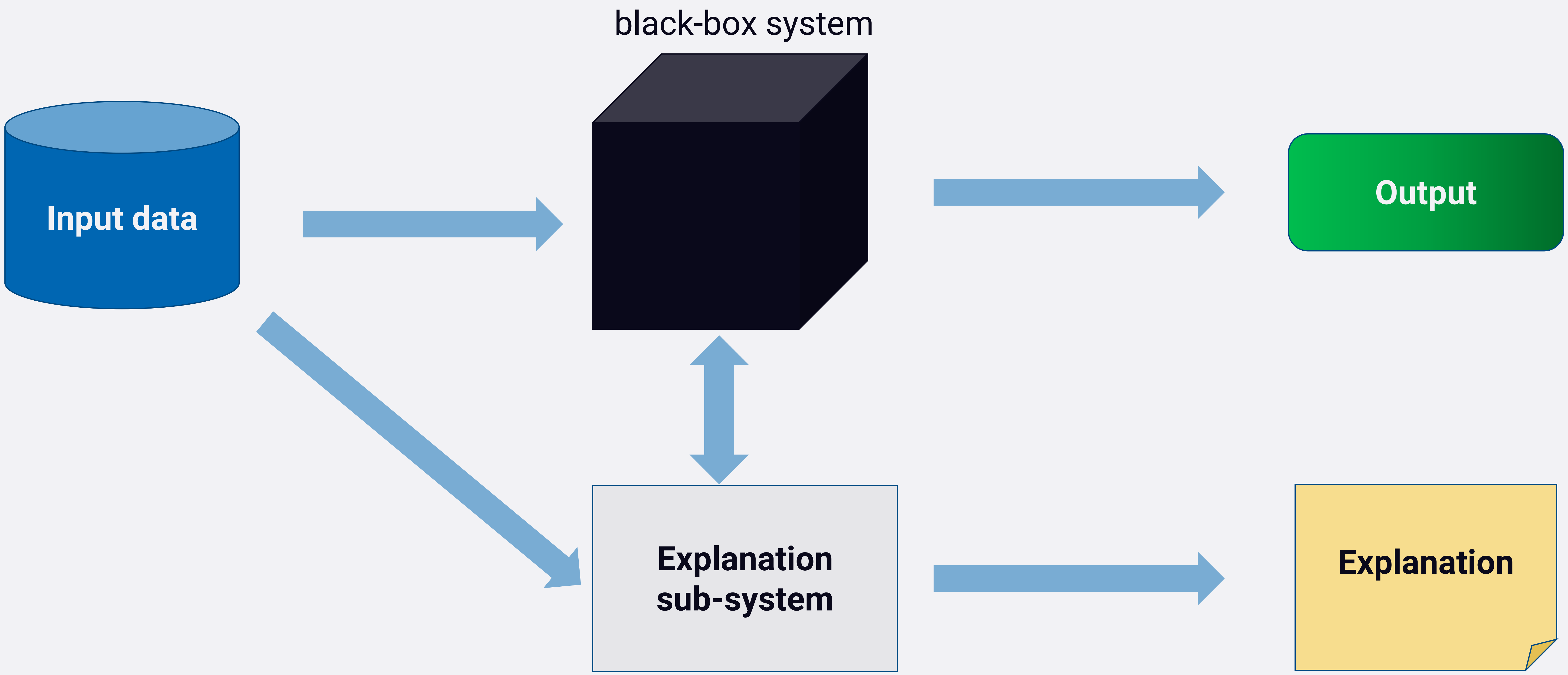
explainable AI
a general view

- No formal, technical, agreed upon definition.
- The comprehensive philosophical overview out of scope of this seminar (Miller, 2017)
- Not limited to machine learning! (Lipton, 2016; Tomsett et al., 2018; Rudin, 2018)

- Two main perspectives (Mittelstadt et al. 2018):

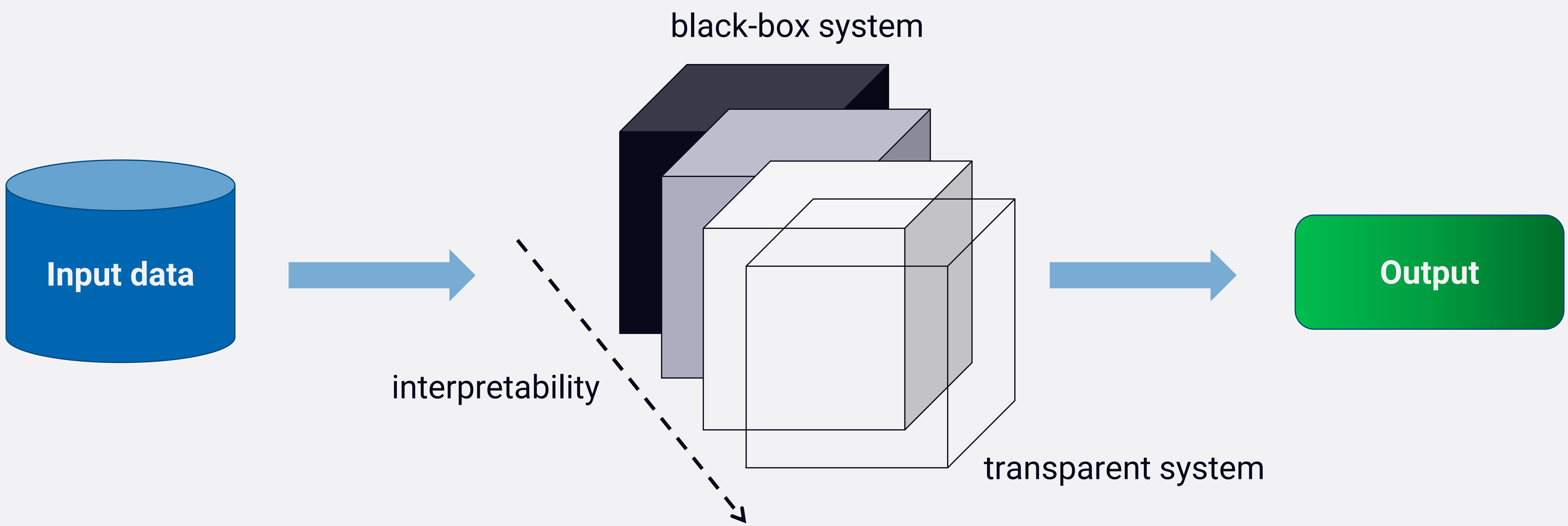
  1. **Post-hoc explanation**: it explains why a black-box model behaved in that way.
  2. **Transparent design**: it reveals how a model works (also know as ante-hoc explanation).

explainable AI
post-hoc explanation



- Post-hoc explanations can be meaningless to many (Rudin, 2018; Mittelstadt et al., 2018).
- Low **Understandability** and Low **Transparency**.

explainable AI
transparent design

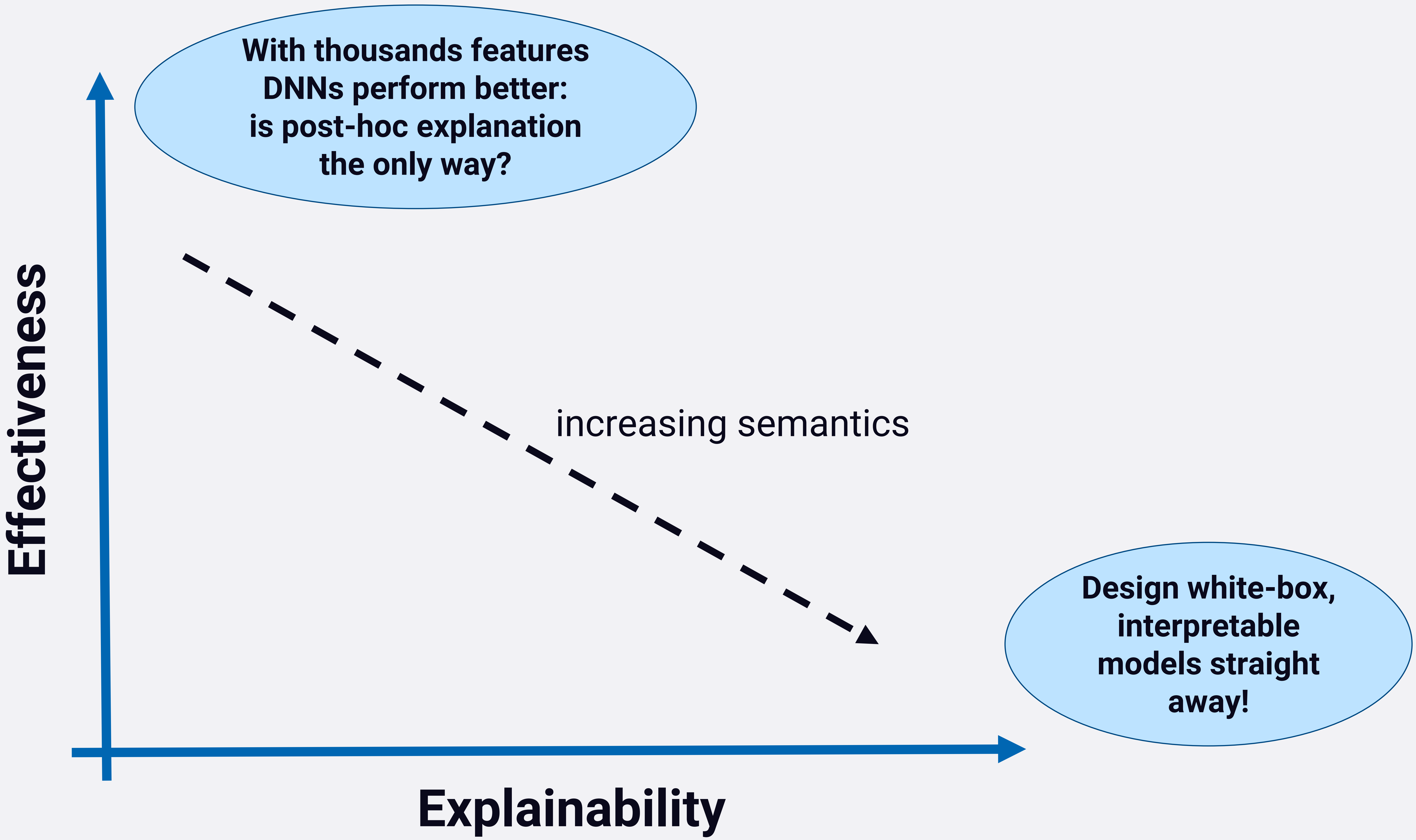


- Three levels of transparency (Lipton, 2016; Lepri et al., 2017; Mittelstadt et al., 2018; Weld and Bansal, 2018):
  1. Simultability
  2. Decomposability
  3. Algorithmic Transparency

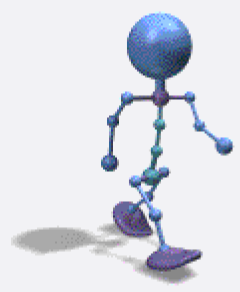
- High **Understandability** and High **Interpretability**.

FONDAZIONE BRUNO KESSLER | ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

explainable AI
considerations



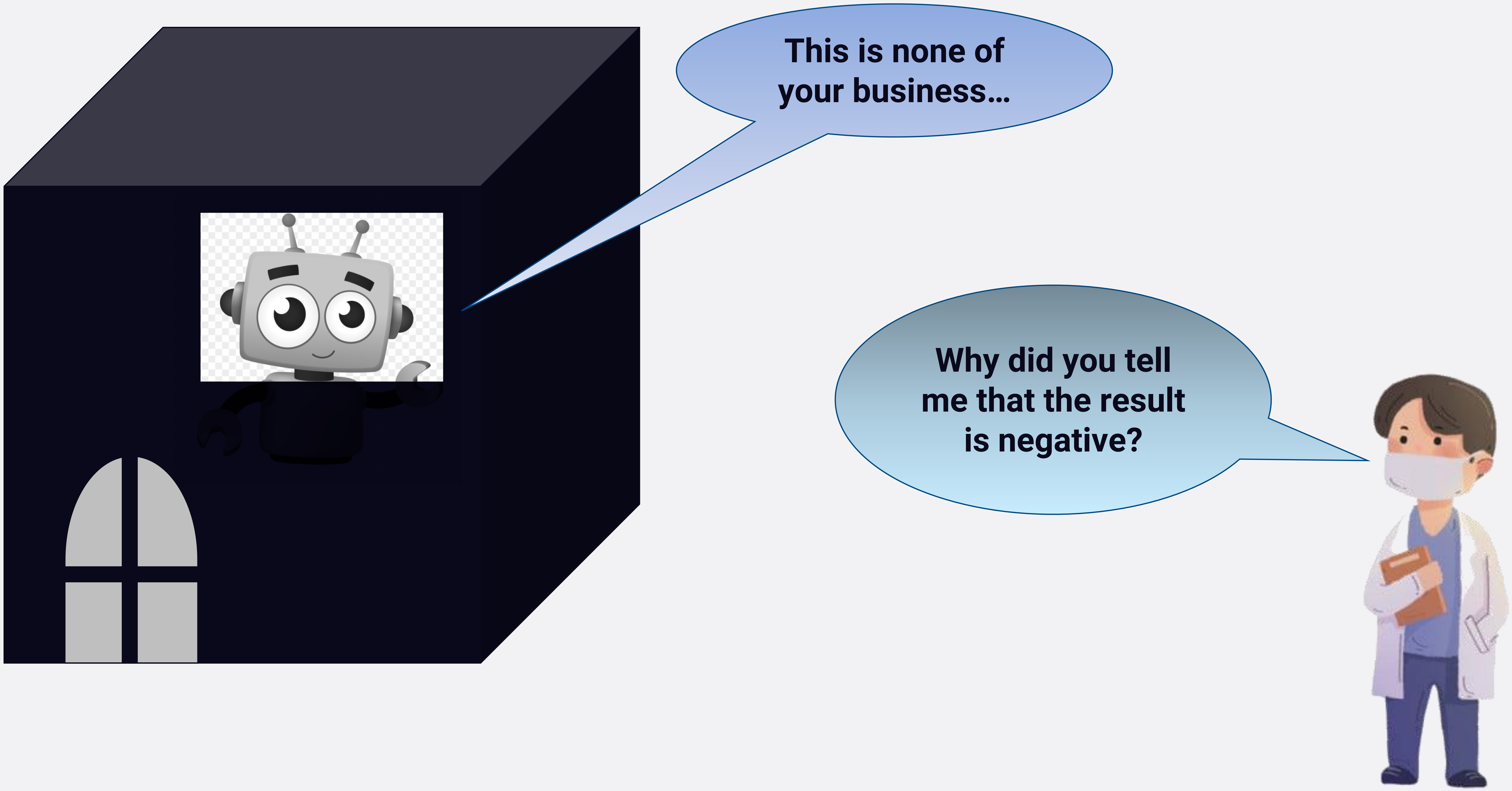
explainable AI
from theory to practice



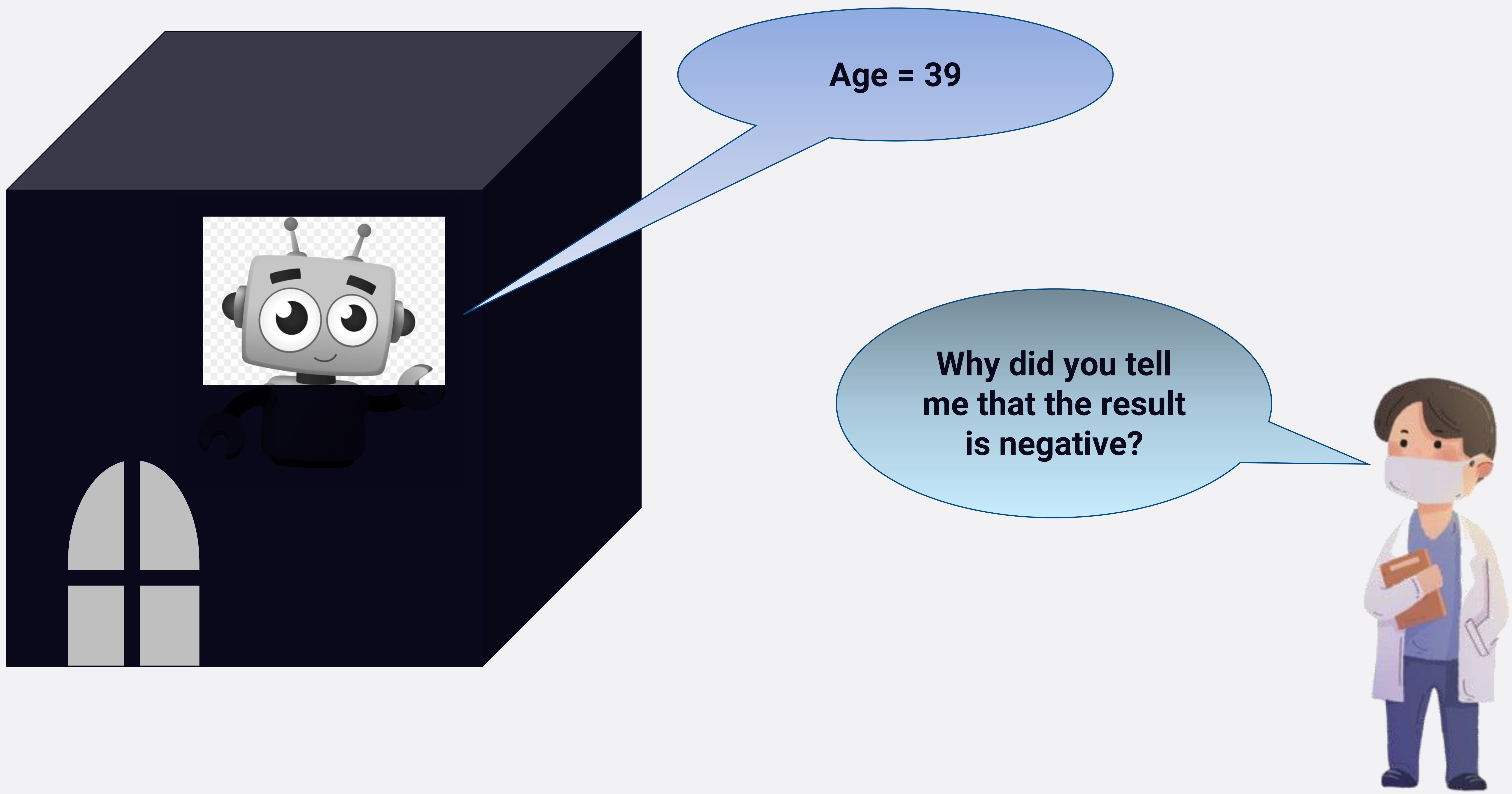
healthcare

*Why are the **challenges of XAI** amplified within **real-world domains** and in particular within the **Digital Health** one?*

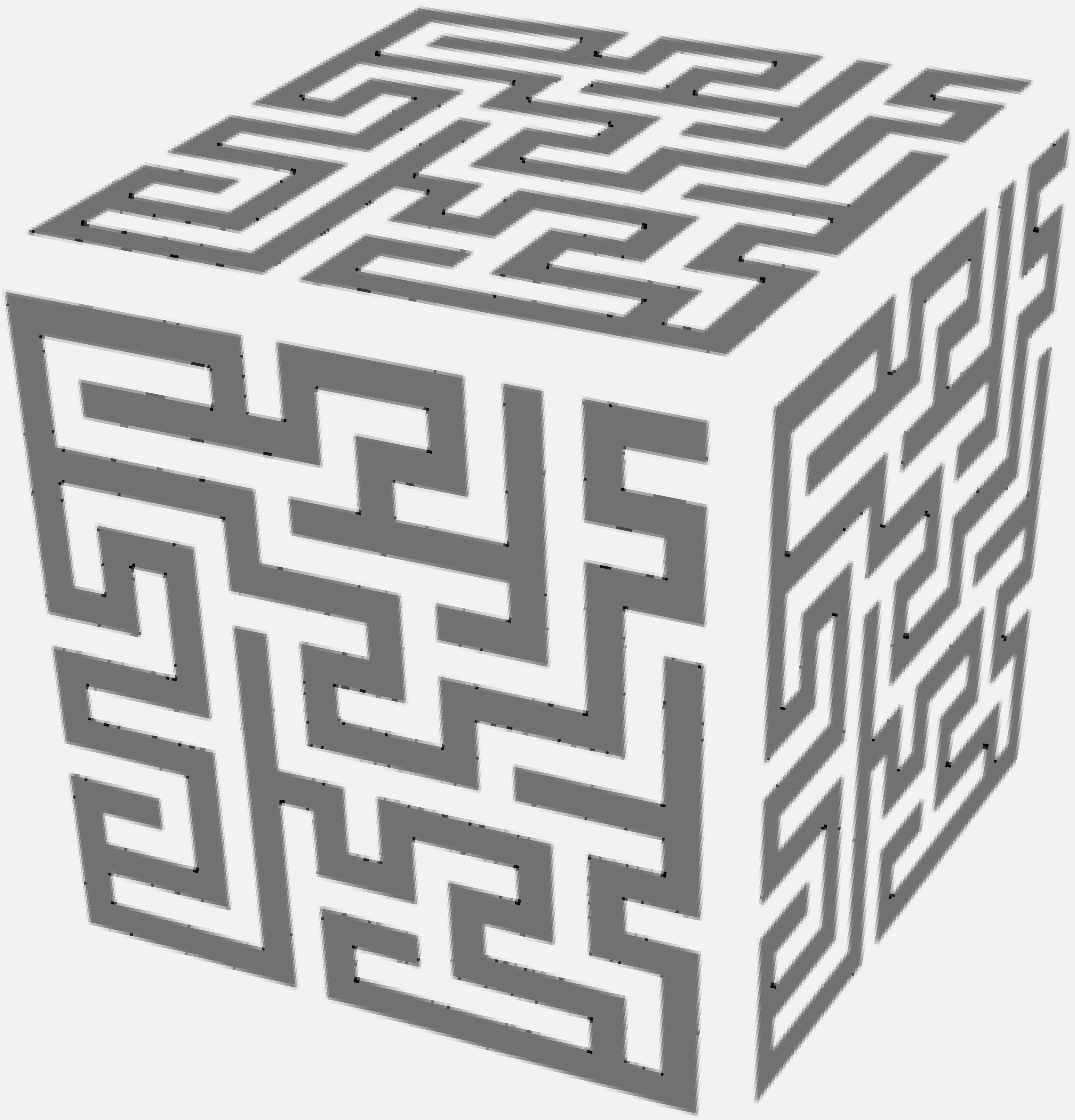
explainable AI and digital health
a problem with trust



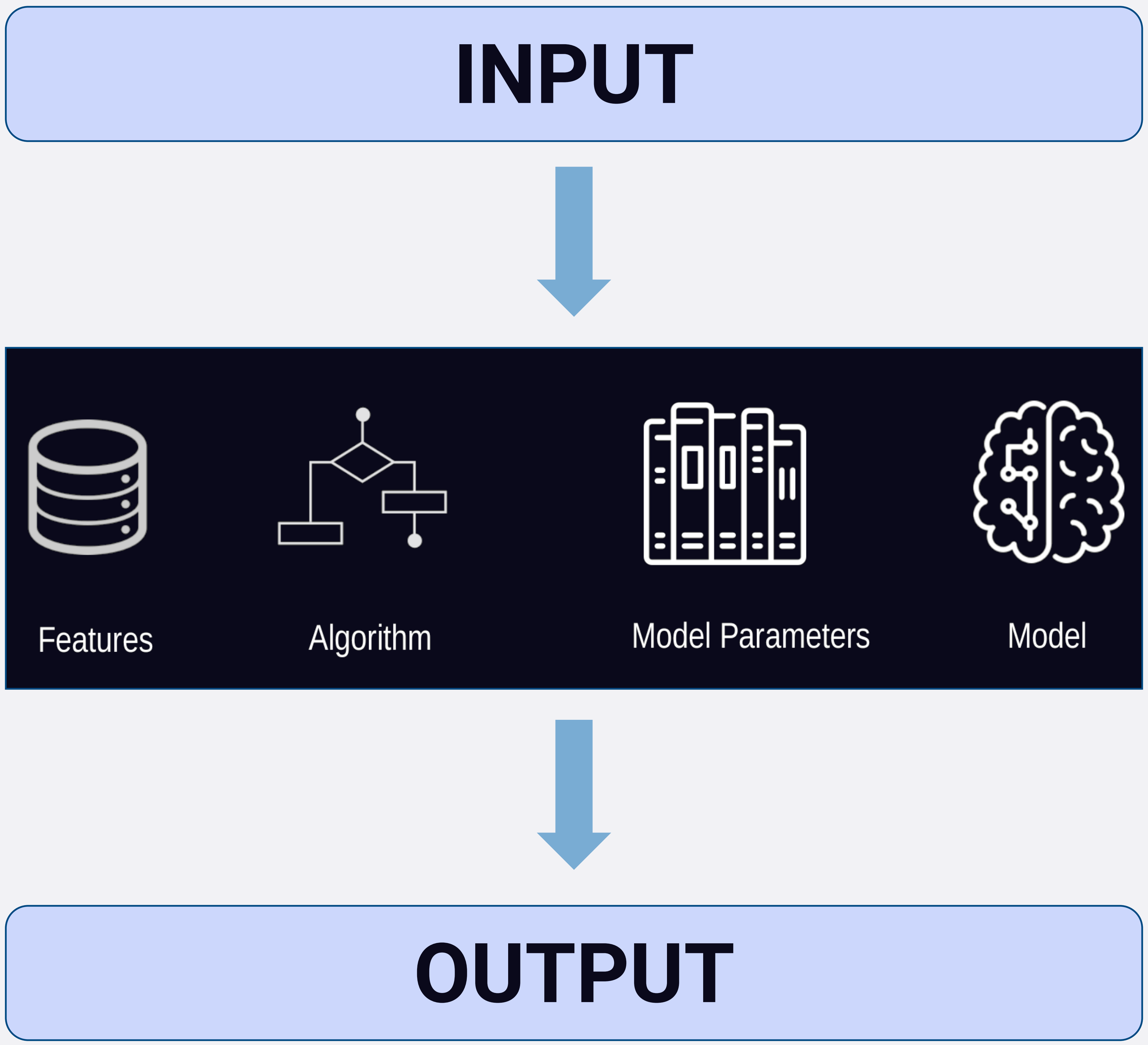
explainable AI and digital health
a problem with trust



explainable AI and digital health
does more transparency mean more trust?



explainable AI and digital health
a problem with trust



Main experts' desiderata:

- To have the certainty that specific input data provide a specific output.
- To have the possibility of changing dynamically the cautiousness of the model.
- To understand how each single feature is treated by the model.

explainable AI and digital health
when do we need explanations?

- **When fairness is critical**: any context where humans are required to provide explanations so that people can not hide behind machine learning models.

- **When consequences are far-reaching**: predictions can have far reaching consequences; e.g., recommend an operation, recommend sending a patient to hospice etc.

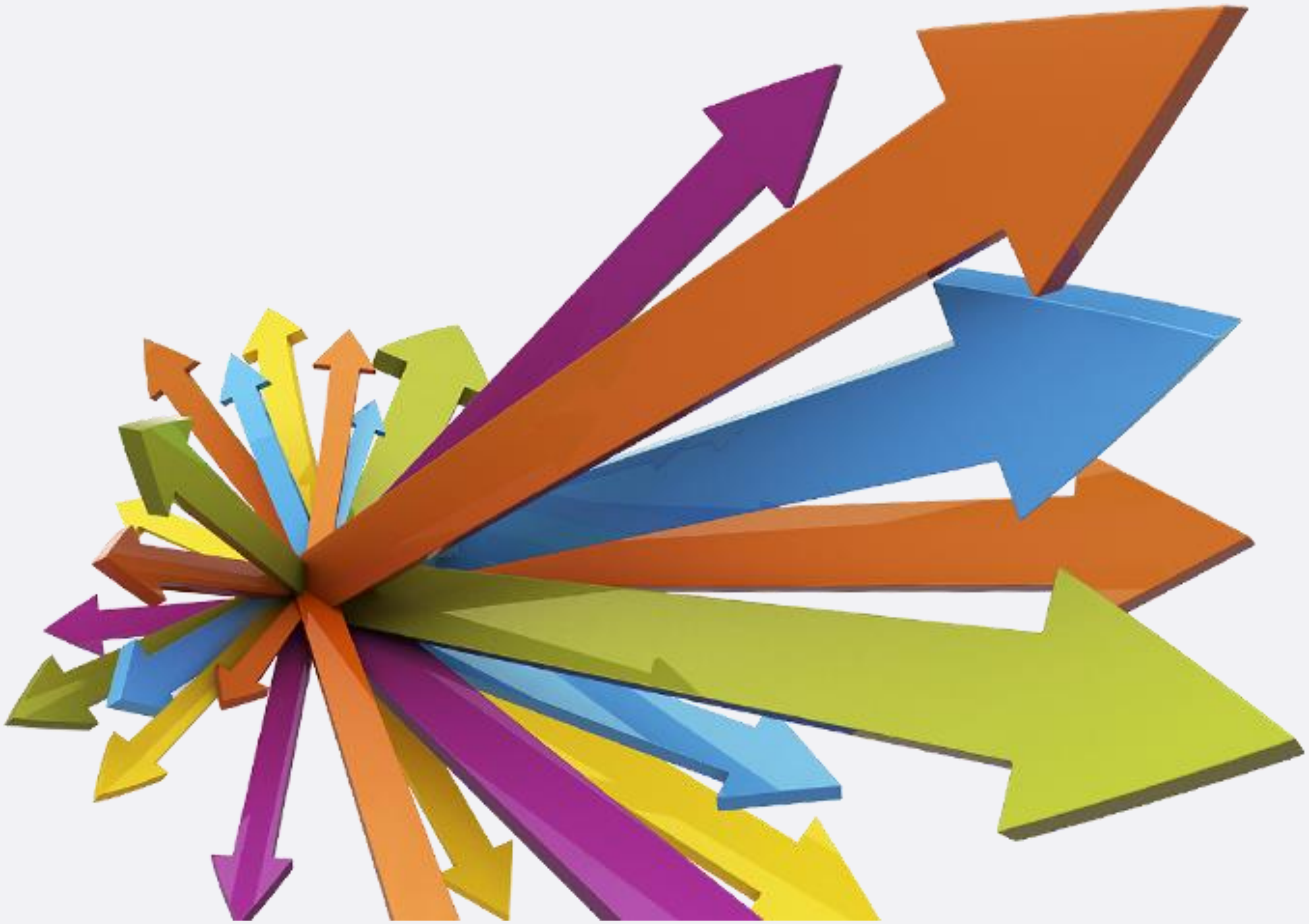
- **When the cost of a mistake is high**: e.g., misclassification of a disease can be costly and dangerous

- **When a new/unknown hypothesis is drawn**: e.g. “Pneumonia patients with asthma had lower risk of dying”

explainable AI and digital health
explanations are role based

- Explanations have to be meaningful.
- A physician requires different explanations as compared to a staff member or to a user.
- Explanations need to be provided with the proper language and also within the proper context.

explainable AI and digital health
how to solve these challenges?



explainable AI in my tenure

XAI in my tenure

*How did **I contribute** to the **explainable AI** research field during **my tenure**?*

explainable AI in my tenure
contributions

01

integration of semantic technologies for enabling the generation of explanations.

02

design of an explainable solution for food image classification.

03

exploitation of knowledge graphs for generating explanations as recommendations for users.

to integrate semantic technologies for enabling the generation of meaningful explanations

(Dragoni and Donadello, 2019)

01

the role of semantic technologies
explanation with background knowledge

- We tend to give explanation in terms of our current knowledge.
- When we see any image of dog our thinking automatically try to capture those objects.
- We always want to conform with our previously acquired knowledge (Background Knowledge).

Will not it be better if we can explain in terms of our knowledge?

the role of semantic technologies
how to use background knowledge?

- Hard to make connection between our knowledge and a model which is trained by reducing loss.

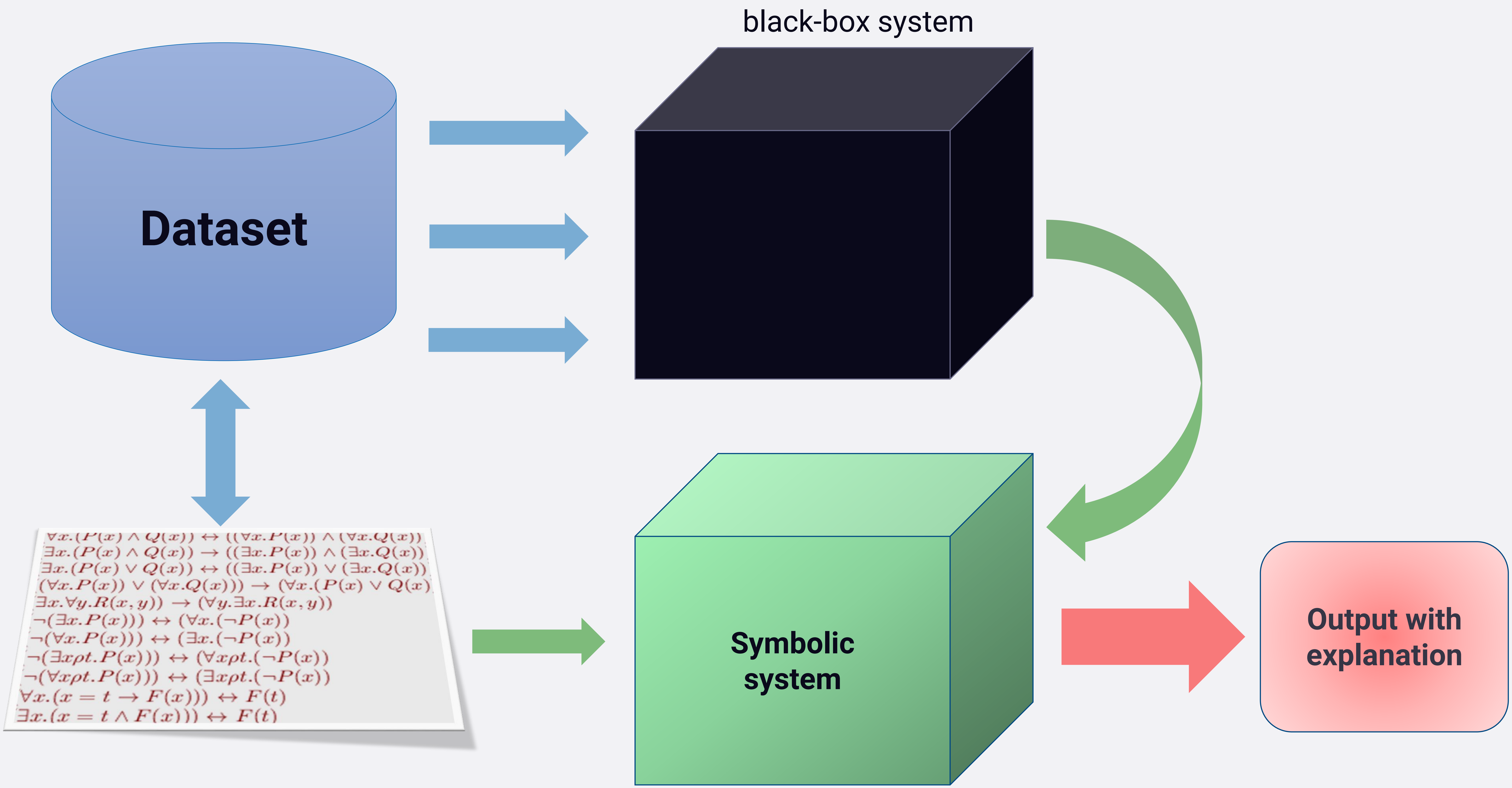
- A three-steps proposal (Dragoni and Donadello, 2019):

  1. Use background knowledge in the form of linked data and ontologies to help explain.

  2. Link inputs and outputs to background knowledge.

  3. Use a symbolic system to generate an explanatory theory.

the role of semantic technologies
how to use background knowledge?



the role of semantic technologies
input needed for these kind of systems

- Background information, ontology, and knowledge graphs
  - Common sense knowledge resources (e.g. Cyc, Wordnet, Suggested Merged Upper Ontology (SUMO), Dbpedia, Freebase)
  - Domain specific resources (e.g. HeLiS (Dragoni et. al., 2018))

- Positive and/or negative examples containing concept-related contextual information (Sarker and Hitzler, 2019).

- Mapping between model dataset and the ontology
  - Mapping each instance as an individual and put it in exact hierarchy.

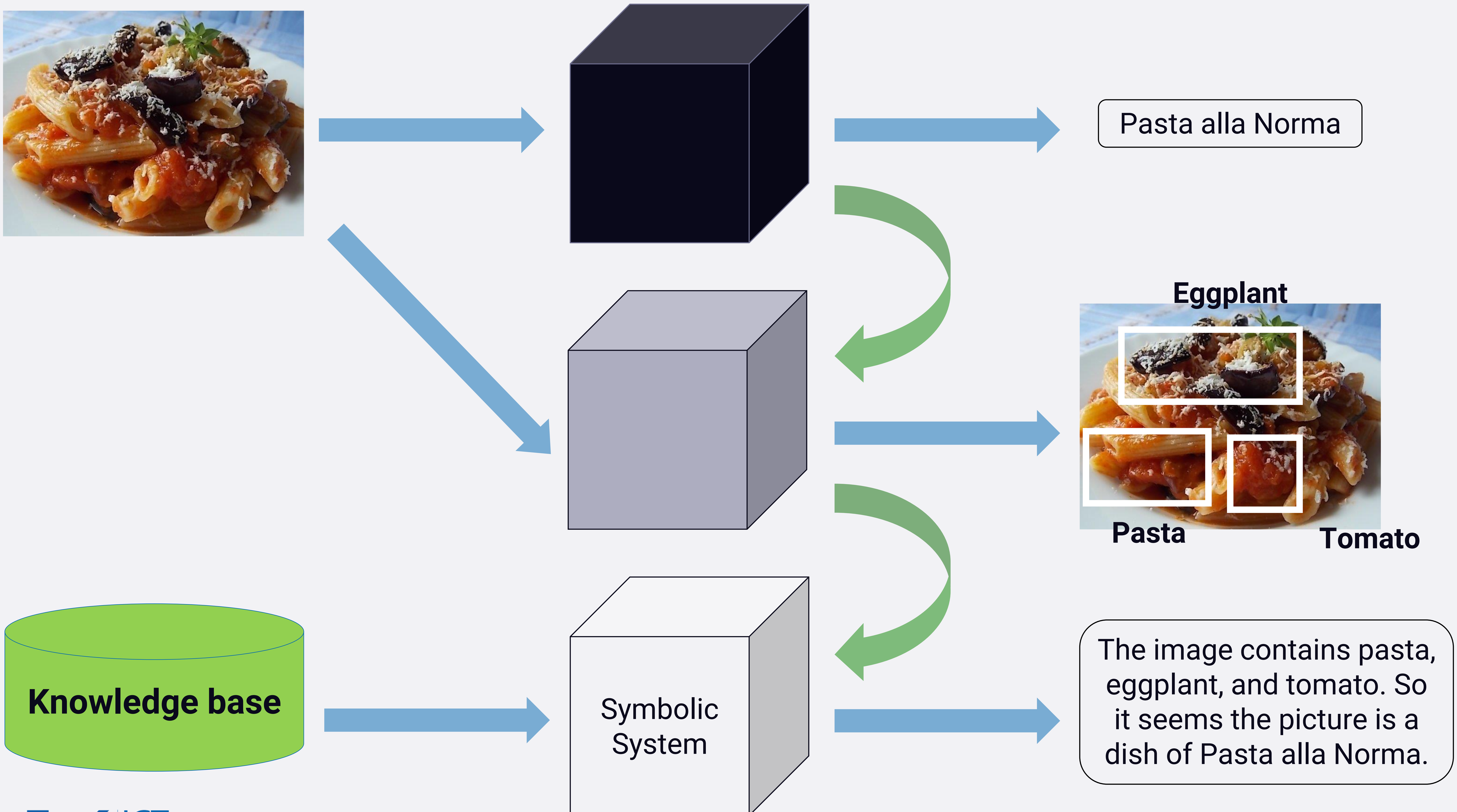
FONDAZIONE BRUNO KESSLER | ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

the role of semantic technologies
pasta image classification example

- Images come with annotations of objects in the picture.
- Objects in image annotations became individuals (constants), which can be typed with the ontology.

```
contains Pasta
contains Eggplant
contains Tomato
contains Ricotta
```

the role of semantic technologies
pasta image classification example



the role of semantic technologies
open questions

- This is just beginning of using background information to enhance explanation.

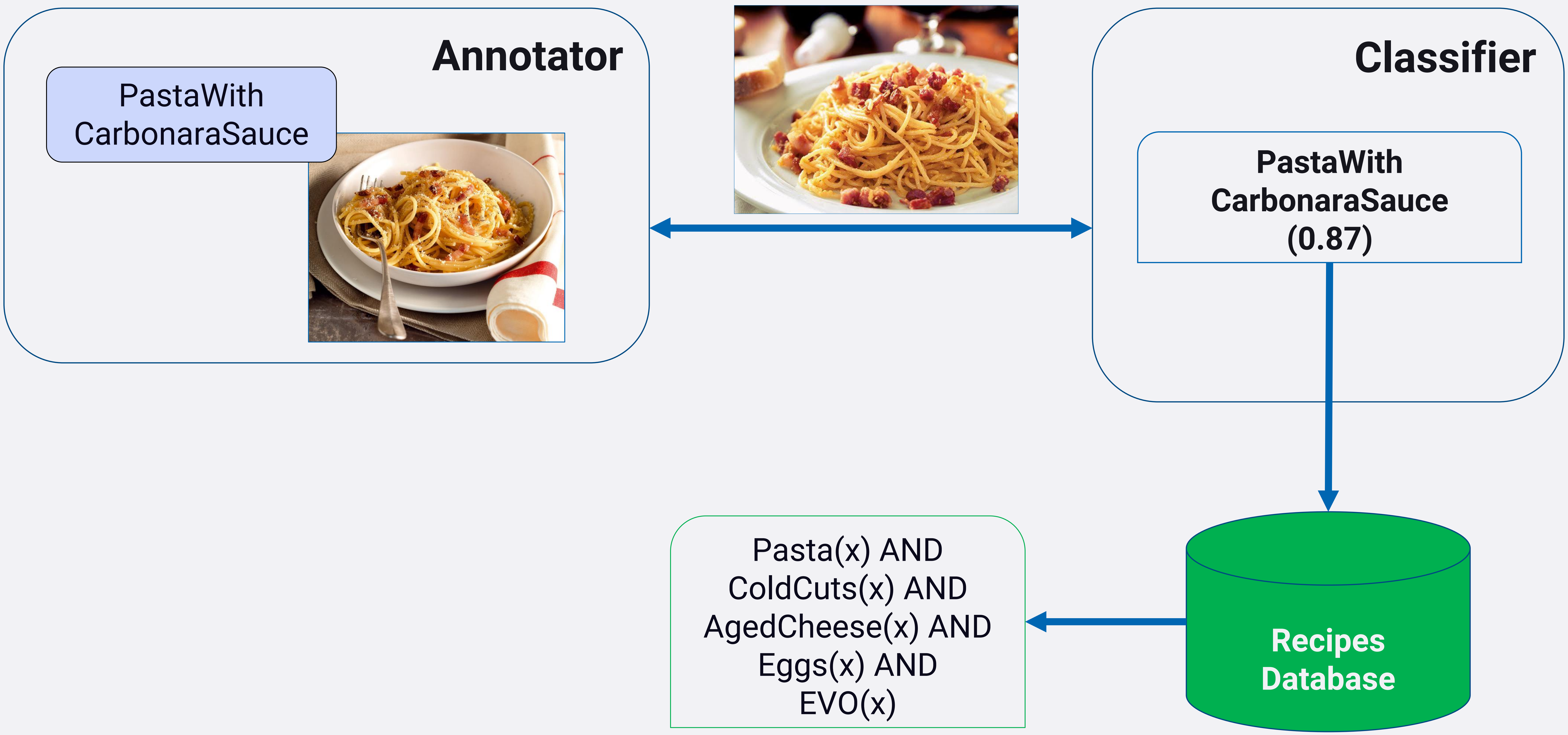
- There are some interesting open questions like:
  - Where can we get effective background information?
  - How to relate already available background information with models?
  - Are those explanations enough to satisfy users' quests?

to classify recipe images through the recognition of ingredients

(Dragoni and Donadello, 2019)

02

food category recognition
state of the art competitor



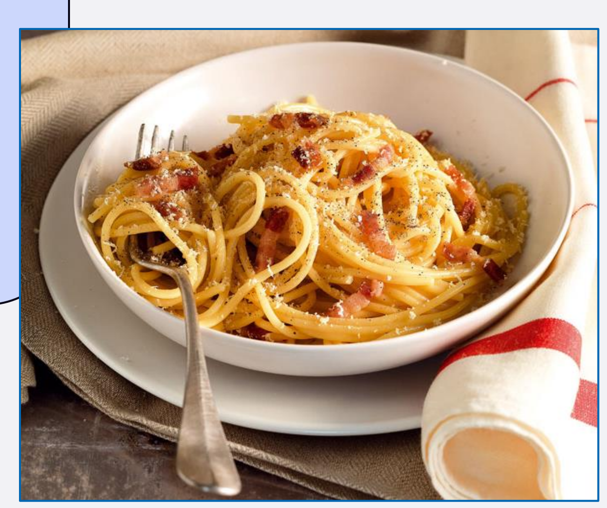
food category recognition
multi-label annotation and classification

Annotator

PastaWith CarbonaraSauce

HeLiS Ontology

Pasta(x) AND
ColdCuts(x) AND
AgedCheese(x) AND
Eggs(x) AND
EVO(x)



Classifier

Pasta - (0.86)
ColdCuts - (0.96)
AgedCheese - (0.88)
Eggs - (0.72)
Tomato - (0.53)

HeLiS Inference Engine

PastaWith CarbonaraSauce

evaluation
effectiveness of classification models

- We enabled explanations and we observed that the proposed strategy can improve the effectiveness of classification models.

Method	Micro-AP (%)	Macro-AP(%)
Multi-label	**76.24**	**50.12**
Single-class (without uncertainty)	50.53	31.79
Single-class (with uncertainty)	60.21	42.51

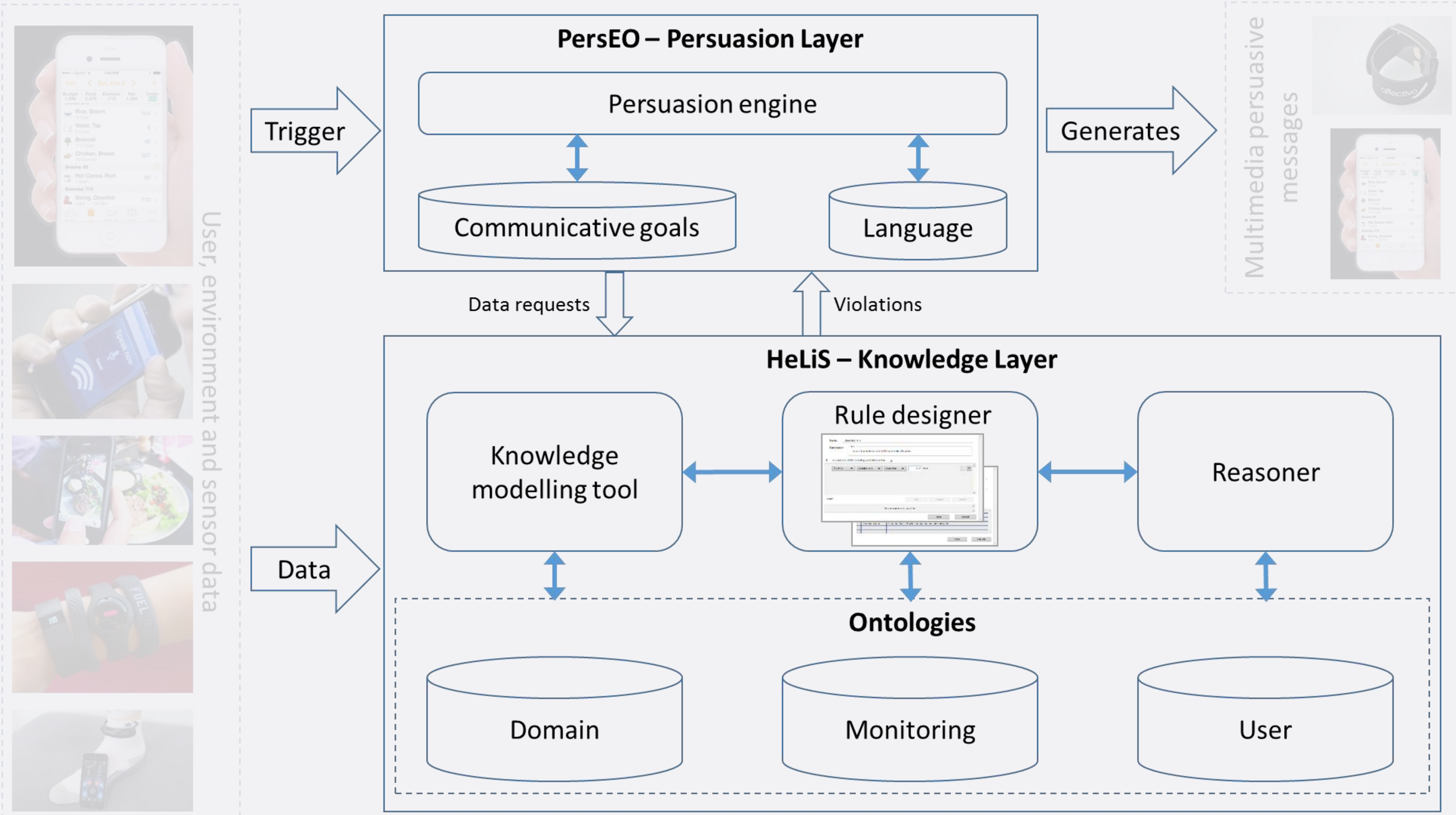
- How can we evaluate the content of the generated explanations?

03

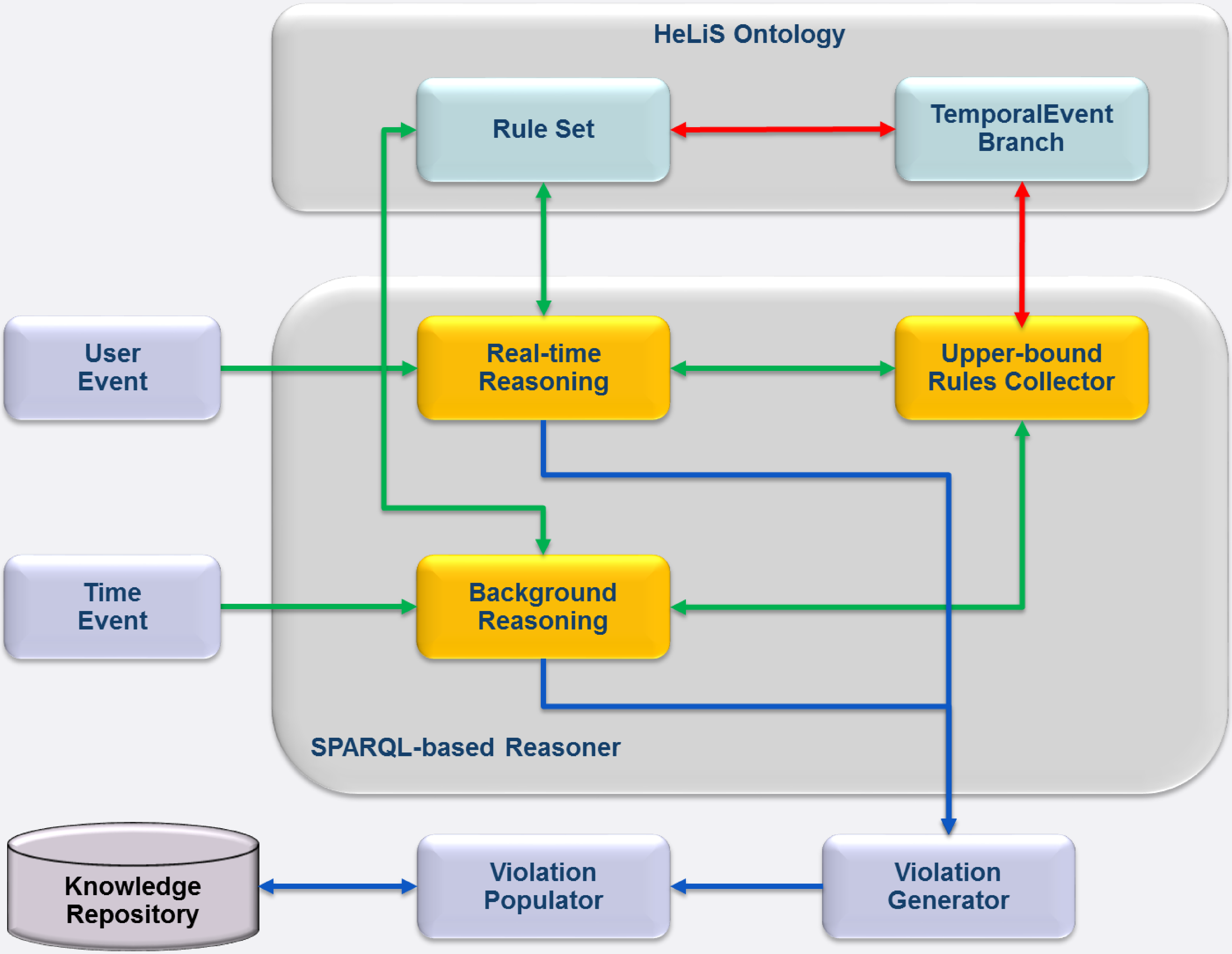
to provide recommendations to users by means of knowledge graphs

(Dragoni et al., 2018; Dragoni et al., 2020)

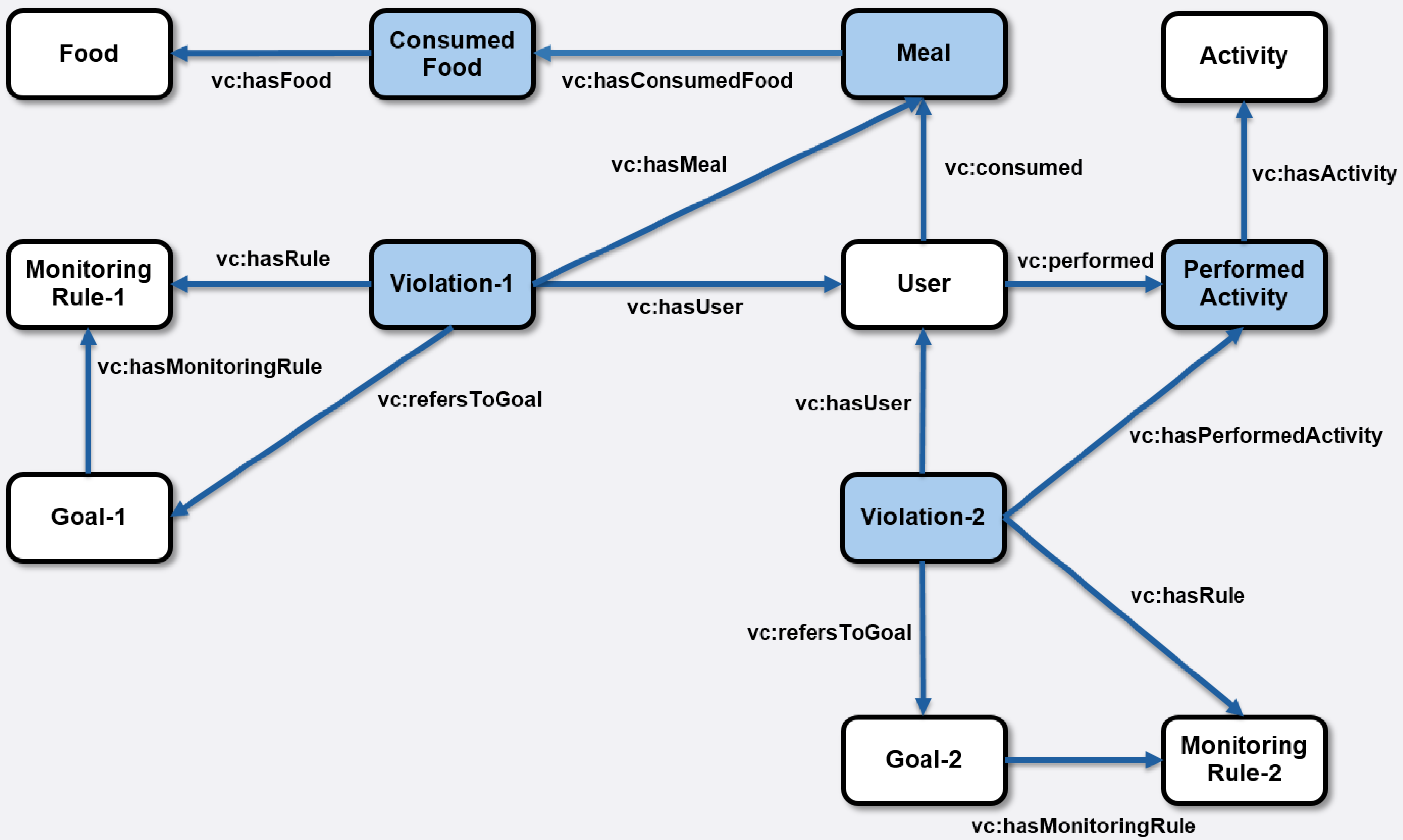
the HORUS.AI platform



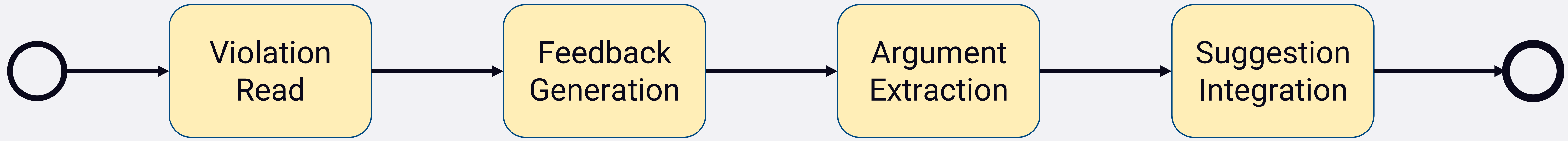
the knowledge layer
the reasoning process

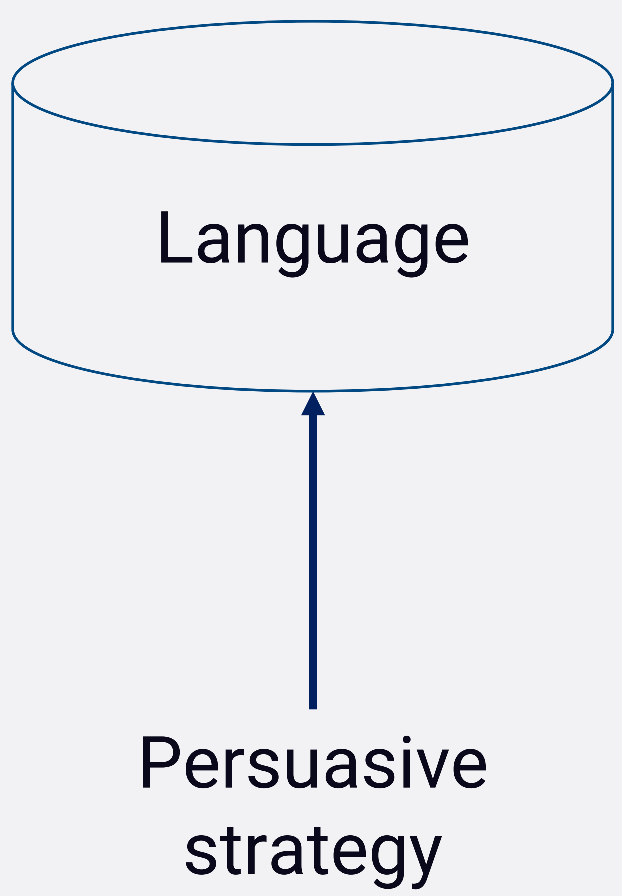


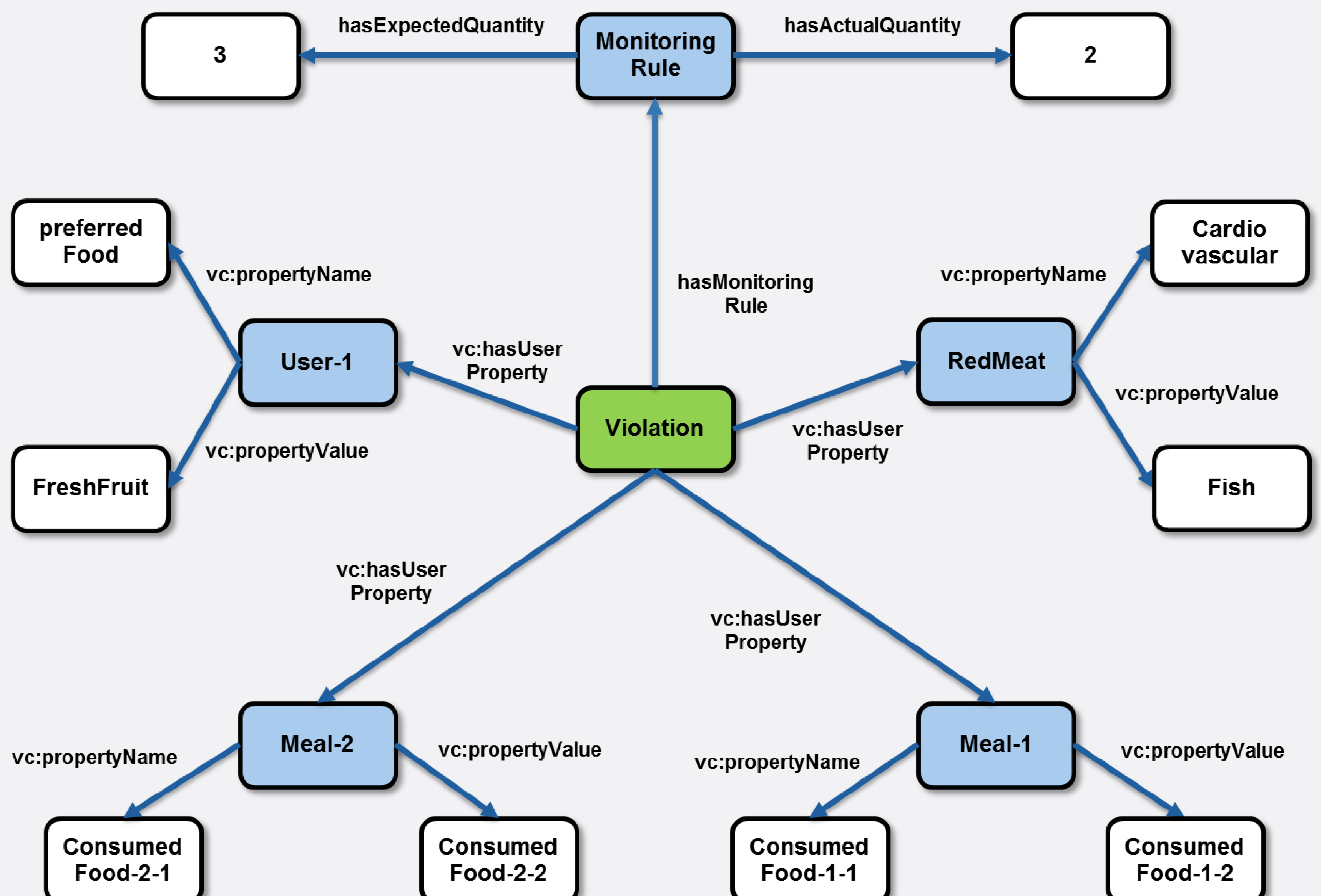
the knowledge layer
population of the knowledge base with detected behaviors



the persuasive layer
message generator process





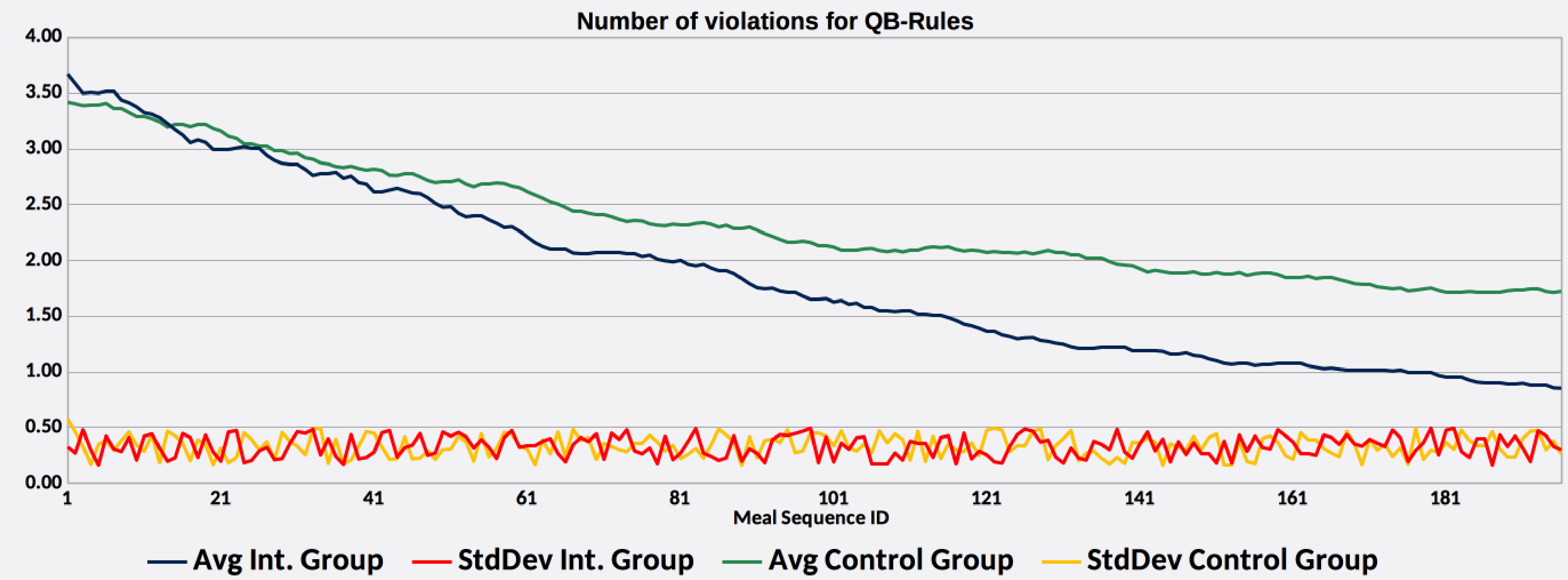


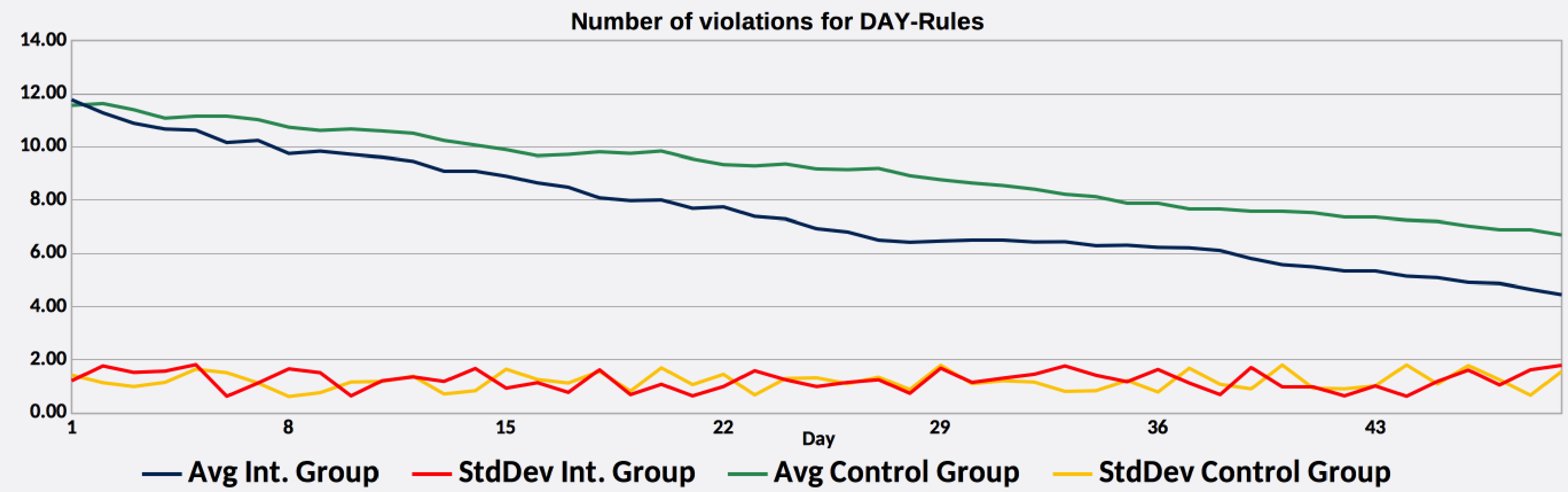
evaluation
living lab

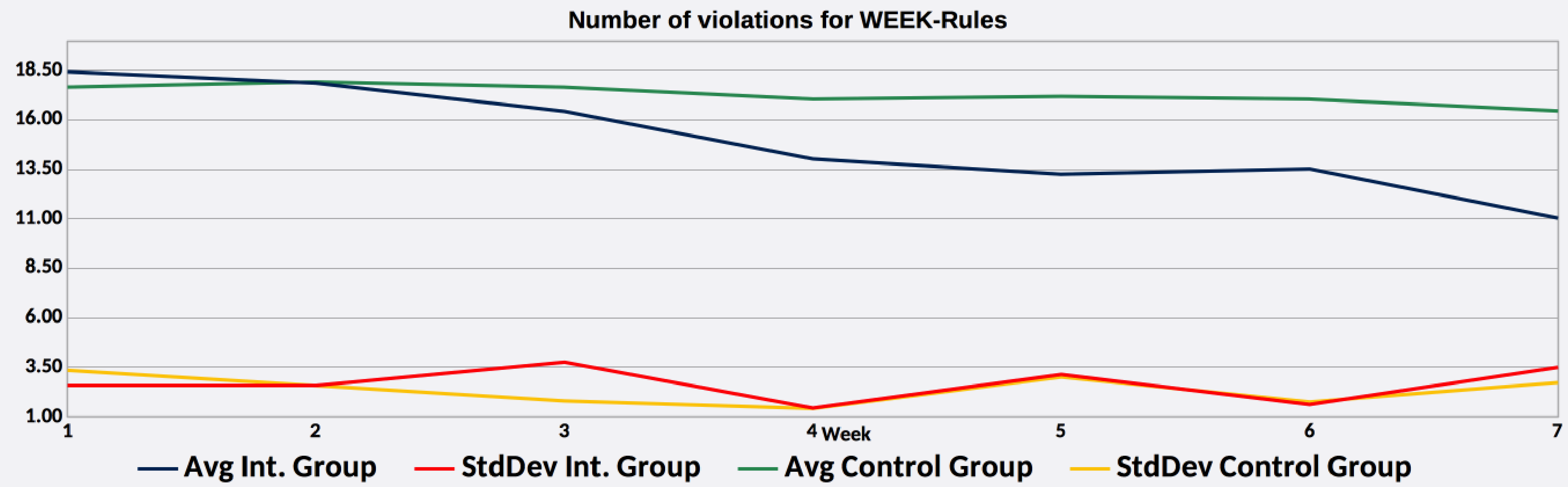
- The evaluation of explanations is an ongoing research activity (Holzinger et al., 2020).

- 120 users have been monitored for 7 weeks.
    - 92 users in the intervention group;
    - 28 users in the control group.

- We observed and reported the effectiveness of generated explanations.







final remarks

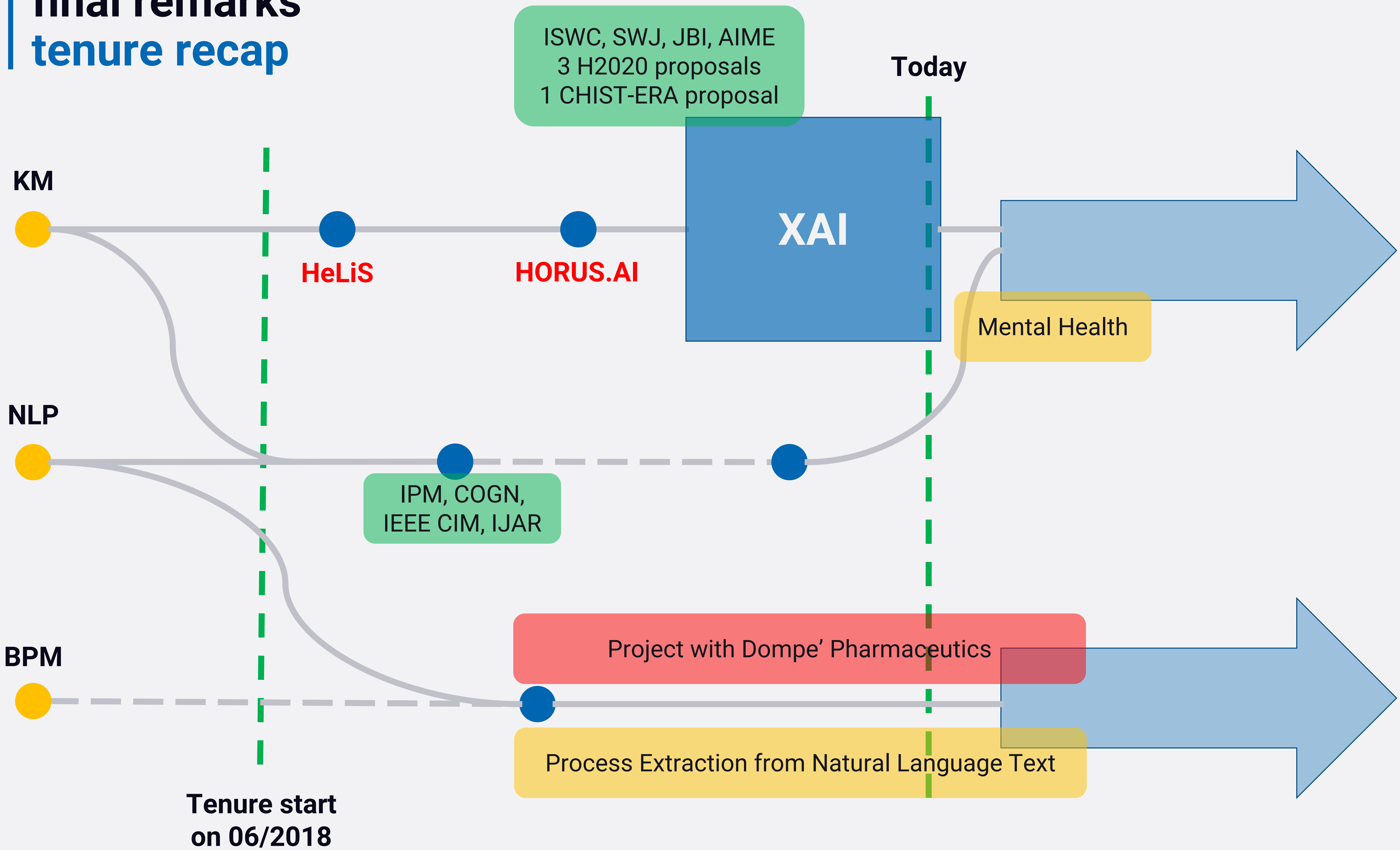
so, in the end?

FONDAZIONE BRUNO KESSLER | ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

final remarks
take-home messages

- Explainable AI is motivated by real-world application of AI.
- Multi-disciplinary: multiple AI fields, HCI, social sciences (multiple definitions).
- Transparent design or post-hoc explanation?
- Background knowledge matters!
- Evaluation:
    - need of benchmark;
    - rigorous, agreed upon, human-based evaluation protocols.

final remarks
tenure recap



FONDAZIONE BRUNO KESSLER
ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

thank you.

references

- M.G. Core, H.C. Lane, M. Van Lent, D. Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In AAAI, pages 1766-1773. MIT Press, 2006.
- E.H. Shortliffe and B.G. Buchanan. A model of inexact reasoning in medicine. Mathematical biosciences, 23(3-4):351-379, 1975.
- W. Swartout, C. Paris, and J. Moore. Explanations in knowledge systems: Design for explainable expert systems. IEEE Expert, 6(3):58-64, 1991.
- W.L. Johnson. Agents that learn to explain themselves. In 12th AAAI, pages 1257-1263, 1994.
- C. Lacave and F.J. Diez. A review of explanation methods for Bayesian networks. The Knowledge Engineering Review, 17(2):107-127, 2002.
- A. Newell, J.C. Shaw, and H.A. Simon. Chess-playing programs and the problem of complexity. IBM Journal of Research and Development, 2(4):320-335, 1958.
- J. Pearl. Causality: Models, Reasoning, and Inference (2nd Edition). Cambridge University Press, Cambridge, 2009.
- S.J. Gershman, E.J. Horvitz, and J.B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245):273-278, 2015.
- Z.C. Lipton. The mythos of model interpretability. arXiv:1606.03490, 2016.
- P. Kieseberg, E. Weippl, and A. Holzinger. Trust for the doctor-in-the-loop. European Research Consortium for Informatics and Mathematics (ERCIM) News: Tackling Big Data in the Life Sciences, 104(1):32-33, 2016.
- D. Doran, S. Schulz, and T.R. Besold. What does explainable ai really mean? A new conceptualization of perspectives. arXiv:1710.00794, 2017.
- G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. arXiv:1706.07979, 2017.
- T. Miller, P. Howe, and L. Sonenberg. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences.arXiv:1712.00547, 2017.
- A. Holzinger, A.M. Carrington, H. Müller: Measuring the Quality of Explanations: The System Causability Scale (SCS). Künstliche Intell. 34(2): 193-198 (2020)
- M.K. Sarker, P. Hitzler. Explaining Input Output Relationship of Training Neural Networks : First Steps, Nesy 2019.
- B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in AI. arXiv preprint arXiv:1811.01439 (2018).
- C. Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. arXiv preprint arXiv:1811.10154 (2018).
- D. Weld and G. Bansal. The challenge of crafting intelligible intelligence. Communications of ACM (2018).

final remarks
publications and metrics

Type	#
Top-ranked journals (Q1, Q1/Q2)	**8**
Other journals	2
Top conferences	**4**
Other conferences	7
Other publications (workshops and demos)	16

Metric	Tenure start	Now
H-index (Google Scholar – Scopus)	11 - 6	**24 - 18**
Citations (Google Scholar – Scopus)	611 - 353	**1457 - 910**