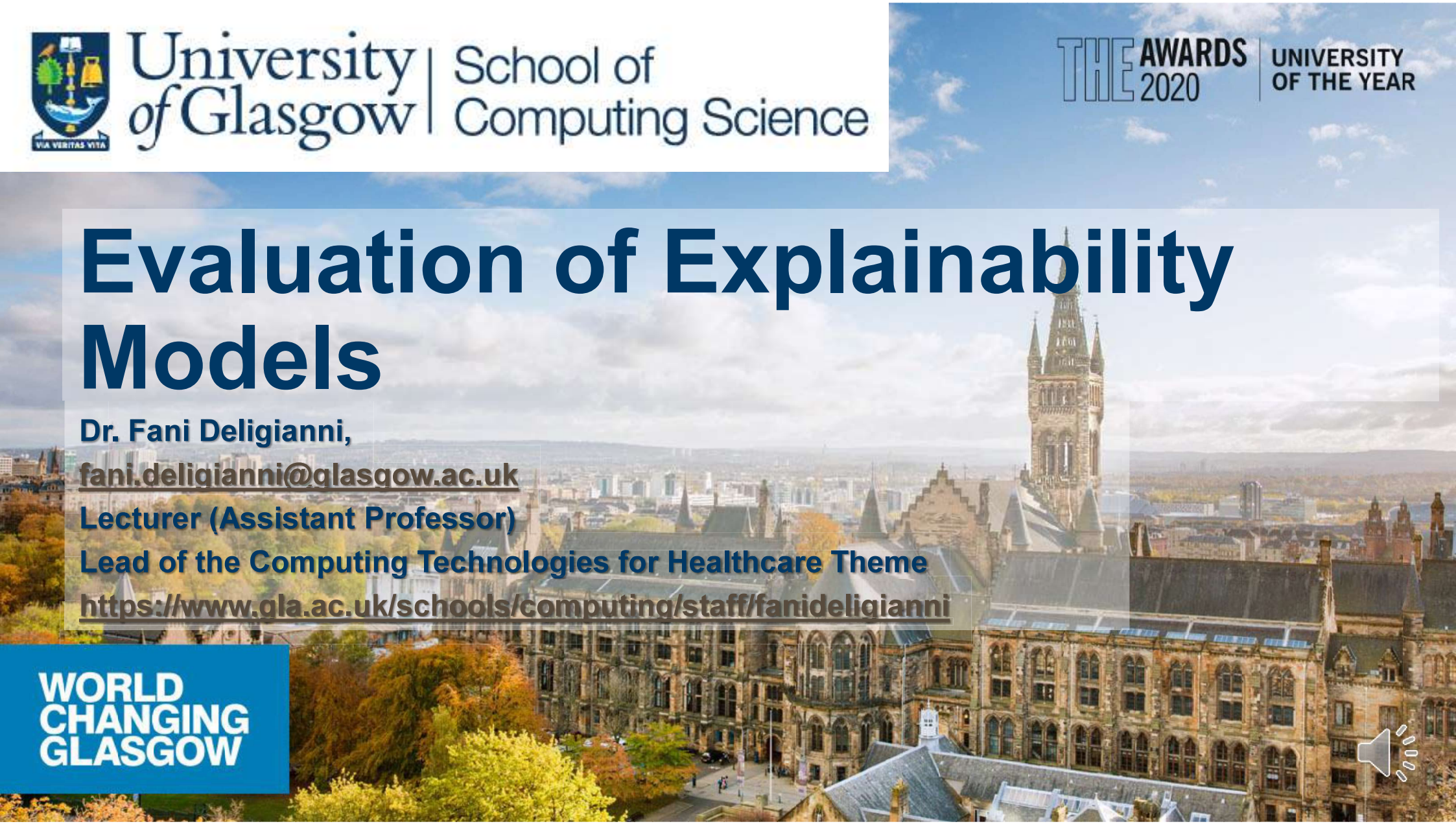
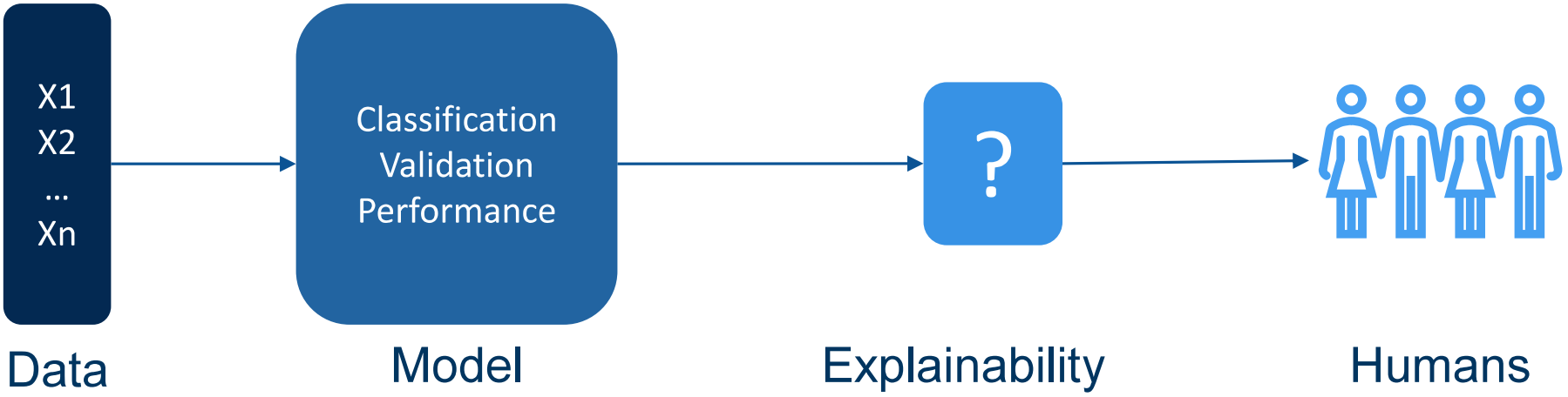


Evaluation of Explainability Models

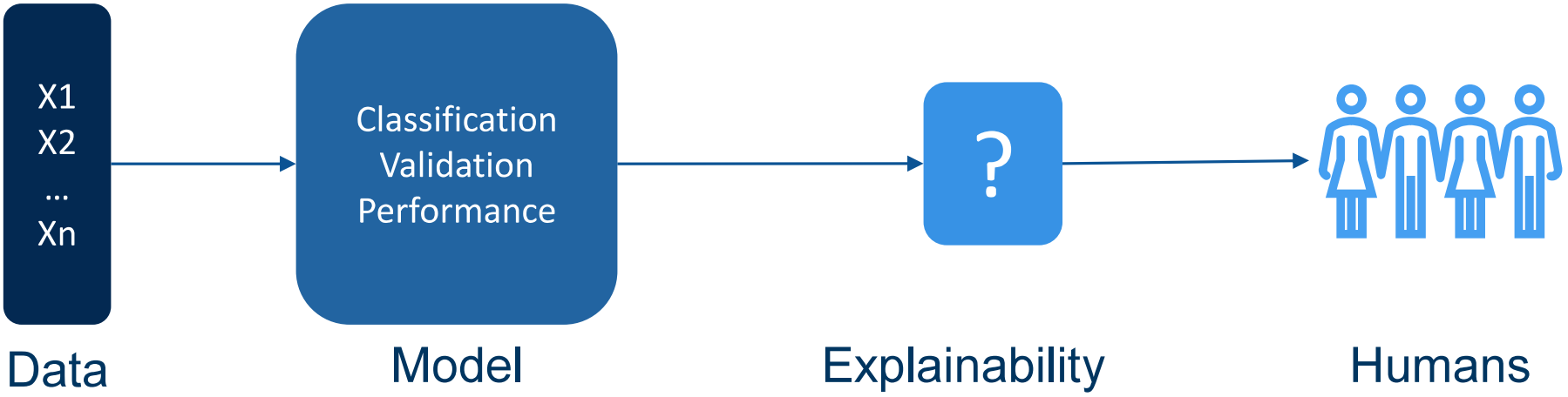
Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Evaluate Explanation



- Allow a formal comparison between explanation methods
- There is no ground truth for post-hoc explanations
- **What are the desirable objectives**

Types of Evaluation

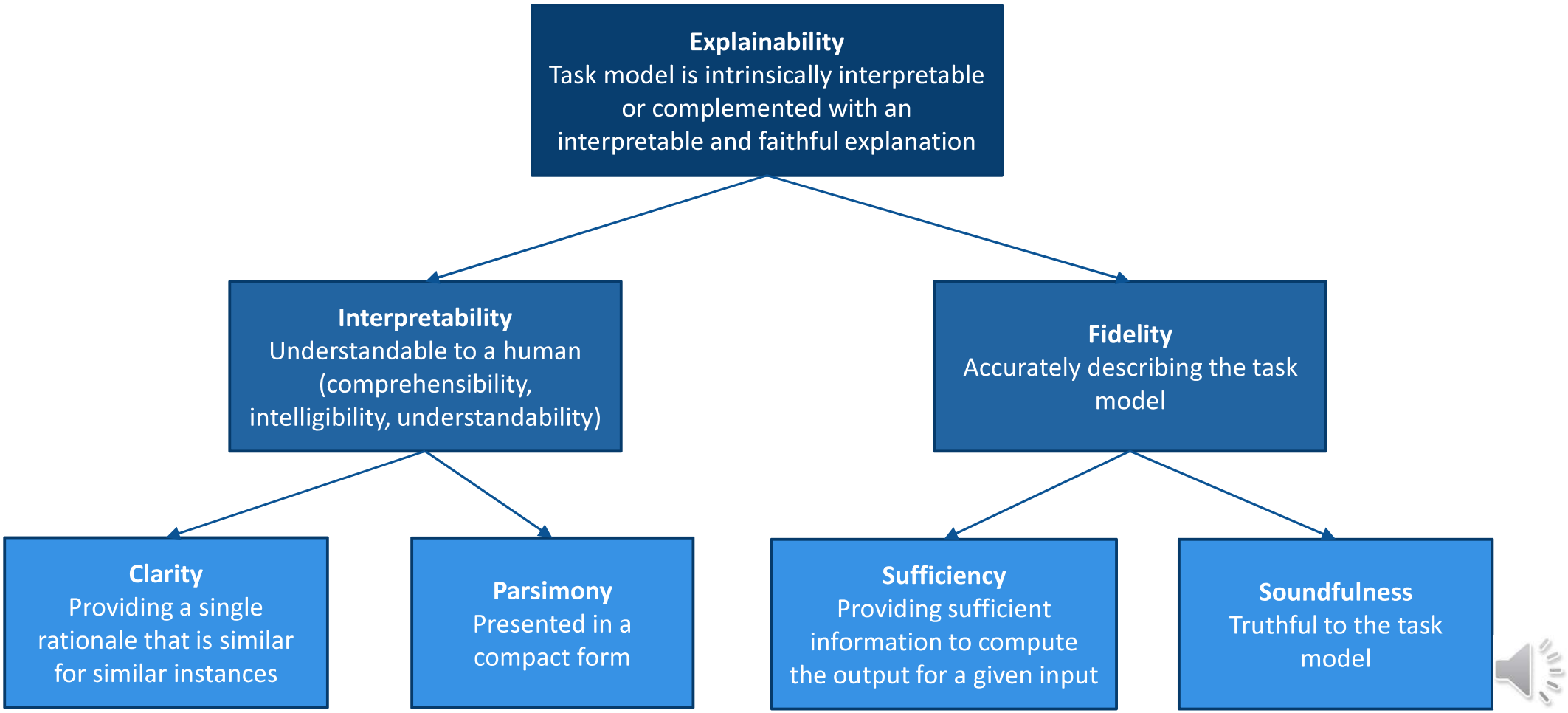


- Application-grounded (experiments with end-users)
- Human-grounded (experiments with lay humans)

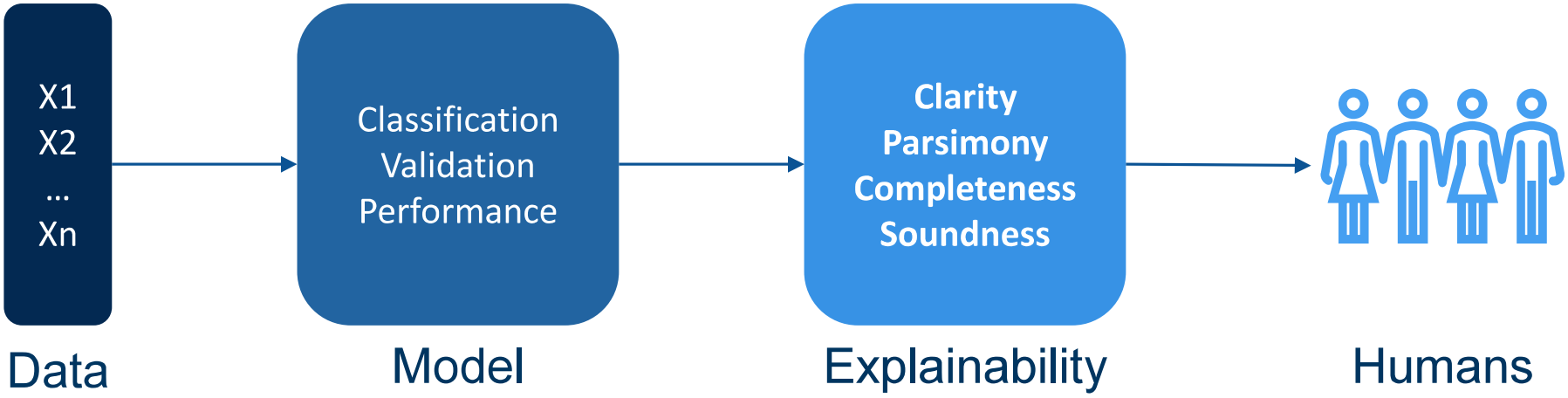
Understand Explanation

- **What features/attributes are important to the model?**
  - You should be able to extract information about what features are important as well as how features interact to create powerful information.
- **Why did the model come to this conclusion?**
  - You should also can extract information about specific predictions in order to validate and justify why the model produced a certain result.
- **Can we approximate the model with a surrogate interpretable model?**
  - Fuzzy models, IF-THEN rule-based system can provide the level of explainability required.

Evaluate Explanations - Characteristics

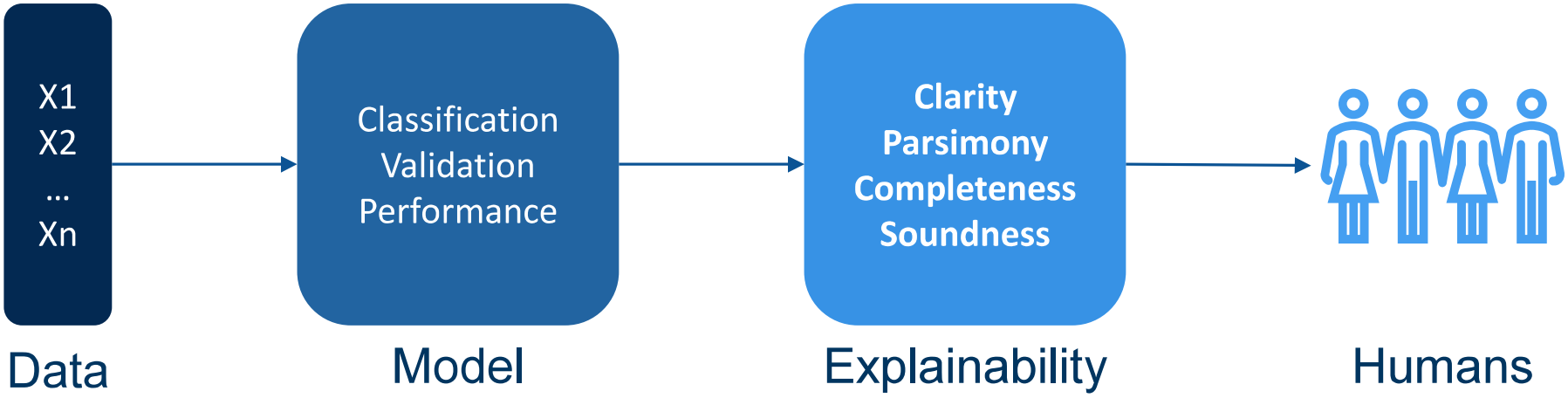


Attribution-based Explanations



- Normally they provide partial explanation and thus do not satisfy **sufficiency**
- **Parsimony** is satisfied as long as the feature itself is understandable

Global vs Local explanations



- Local explanations can be different between similar samples
- Global explanations satisfy **clarity**

Summary

- Another way to evaluate explanations is based on the end users
- User-based evaluation can be quantitative and qualitative:
  - Qualitative: Questionaries
  - Quantitative: Performance based
- User-based evaluations are important to understand how trust in AI models affects overall system performance

References

- Markus et al. ‘The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies’, Journal of Biomedical Informatics, 2021.