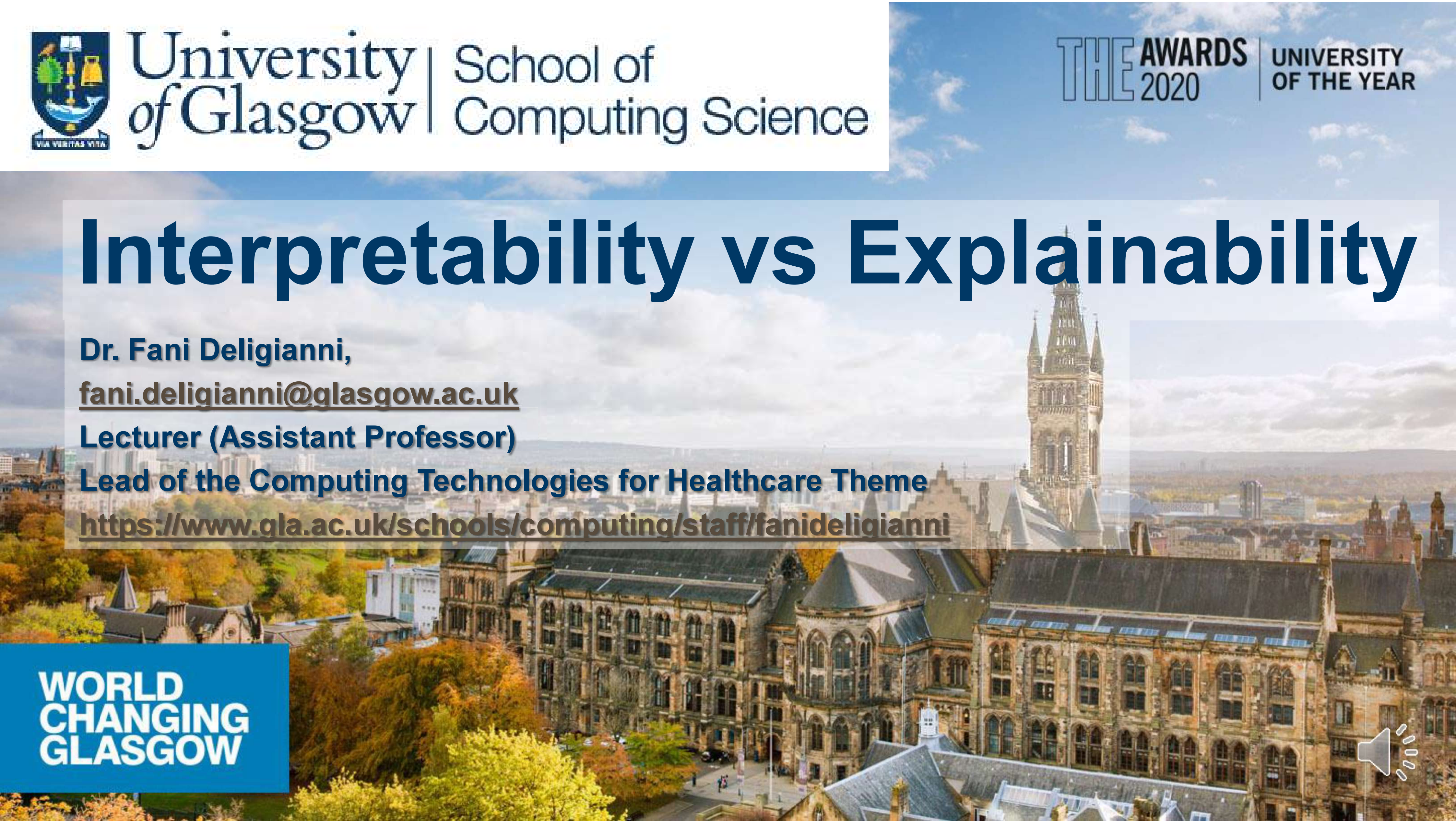
University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

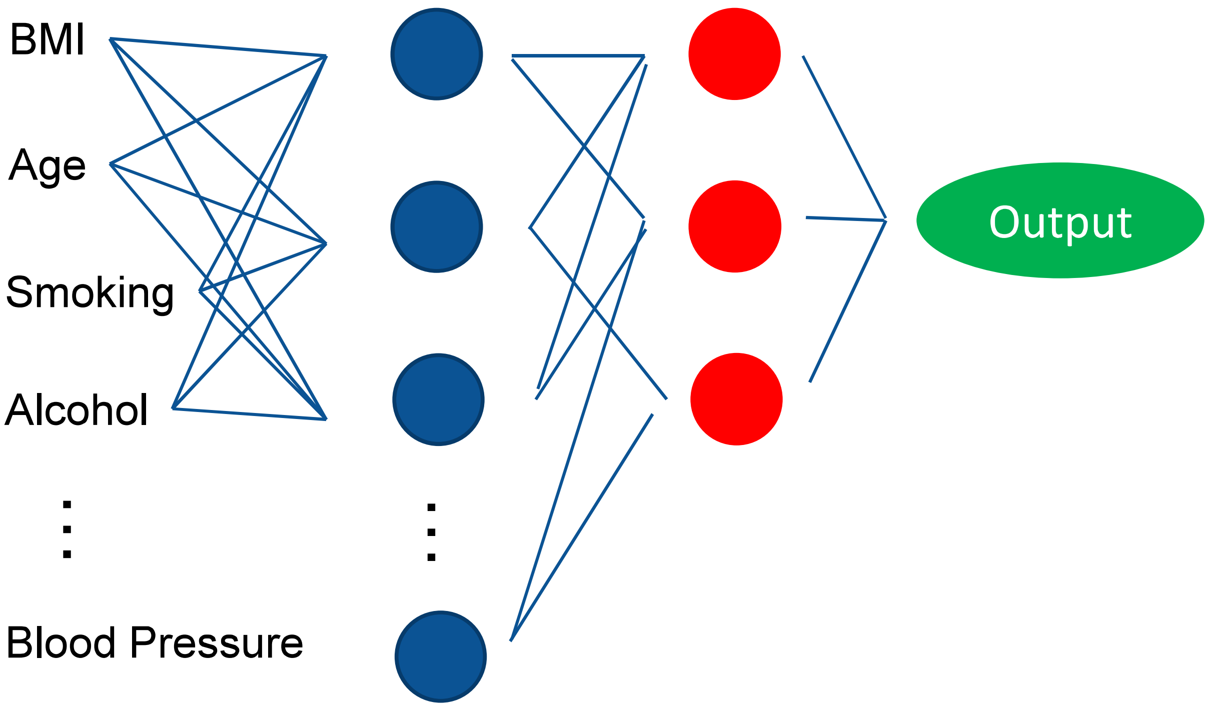


Interpretability vs Explainability

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

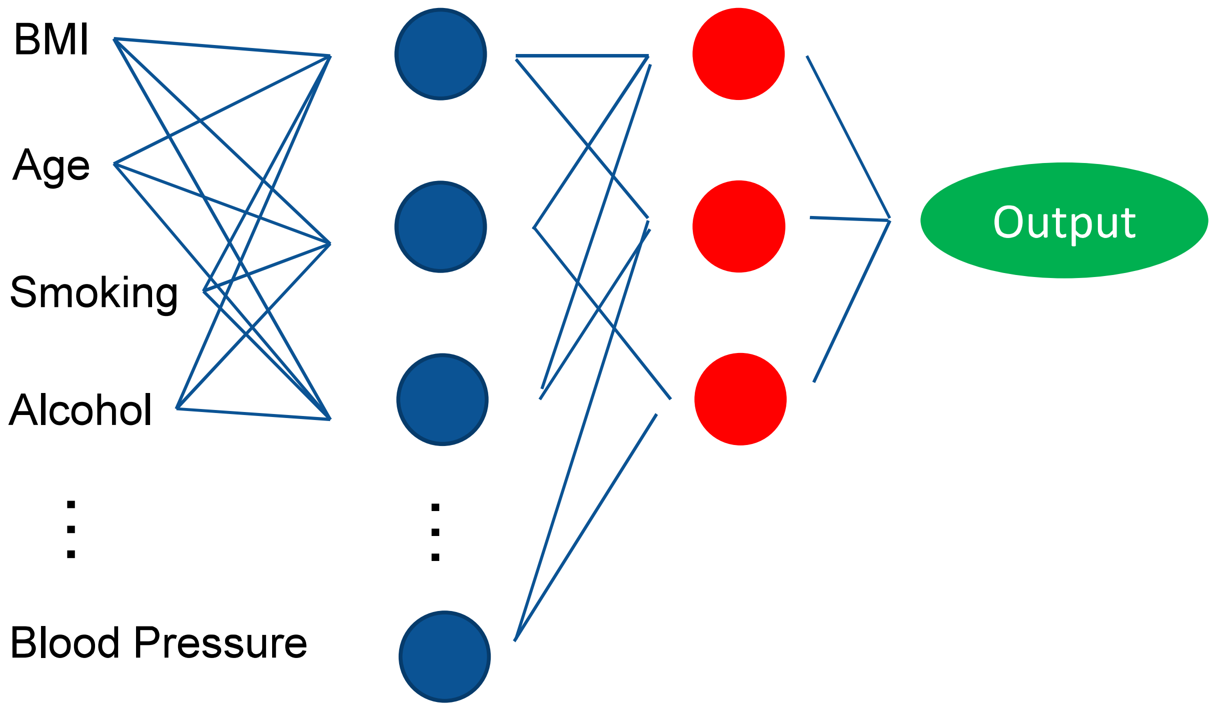
WORLD CHANGING GLASGOW

Explainable Model



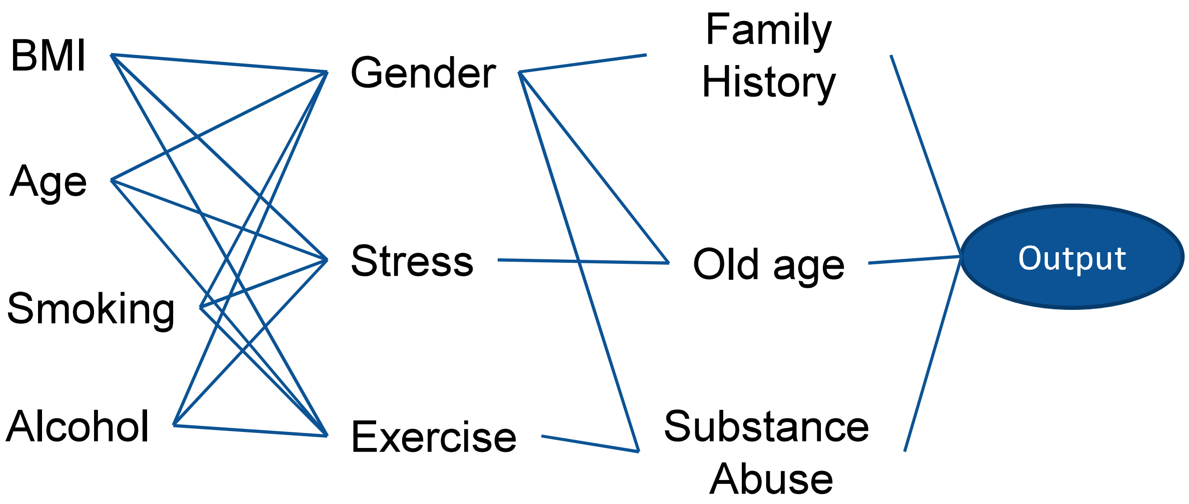
- Do we understand why the model came to this output?
- Do we know the conditions/cases that the model is successful and when it is not?
- Do we know the factors behind this output?

Explainable Model - Factors



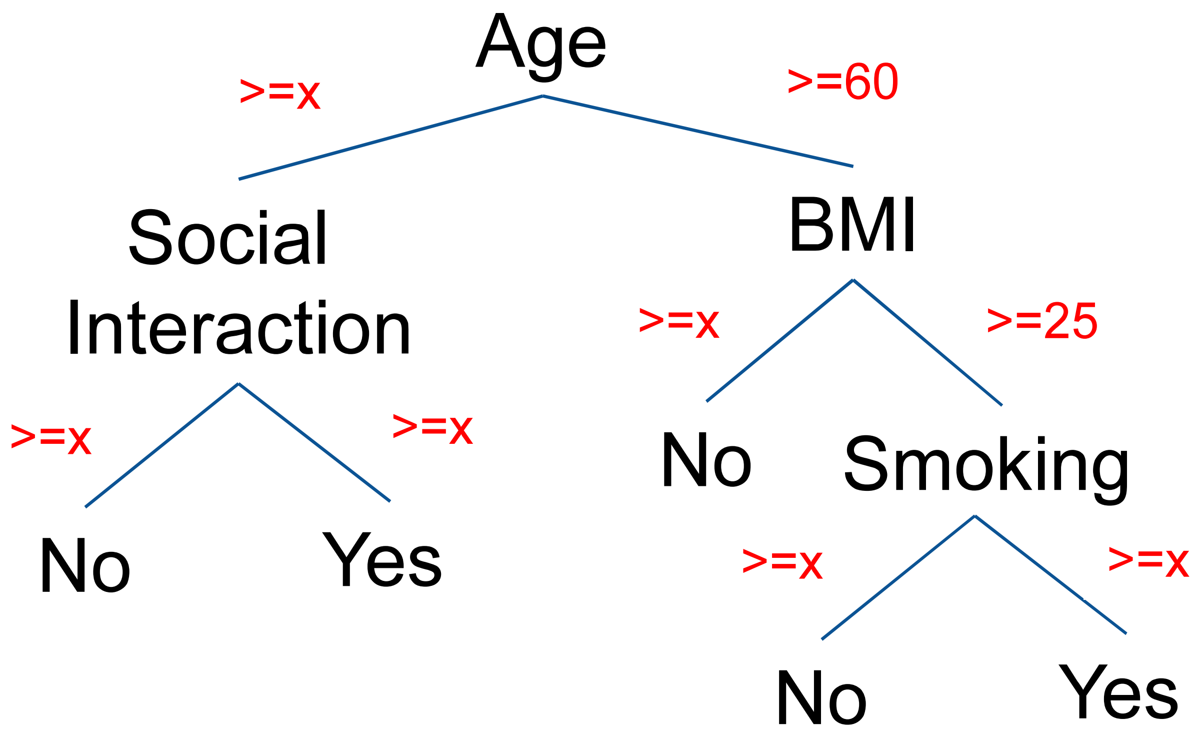
- Age is the most important factor in predicting heart failure.
- Large BMI also increases the probability of a heart attach episode
- History of smoking also increase the probability
- High blood pressure is also associated with heart failure

Explainable Model – Representation Learning



- Knowledge of the what each node represents
- Latent factors that affect the decision process
- How important each node is to the model’s performance

Interpretable Models – Decision Trees



Age

>=x

>=60

Social Interaction

BMI

>=x

>=x

>=x

>=25

No

Yes

No

Smoking

>=x

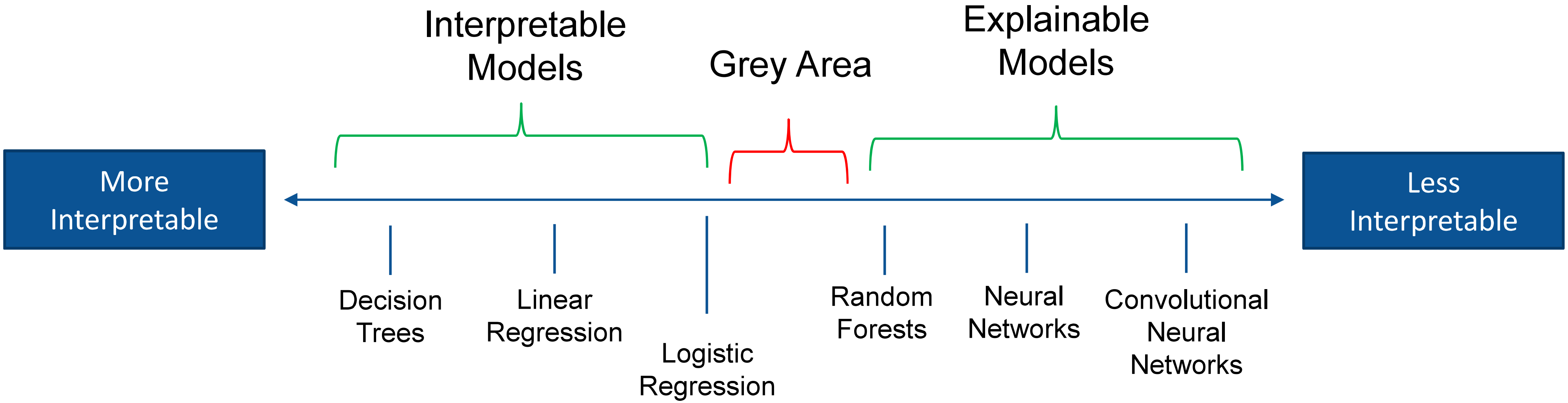
>=x

No

Yes

- It is clearly what each node represents
- Easy to visualize and overview the whole decision operation
- Easy to explain to non-specialists
- Results can be tracked and associated with the output of each node

Interpretable vs Explainable Models



Interpretable Models

Grey Area

Explainable Models

More Interpretable

Less Interpretable

Decision Trees

Linear Regression

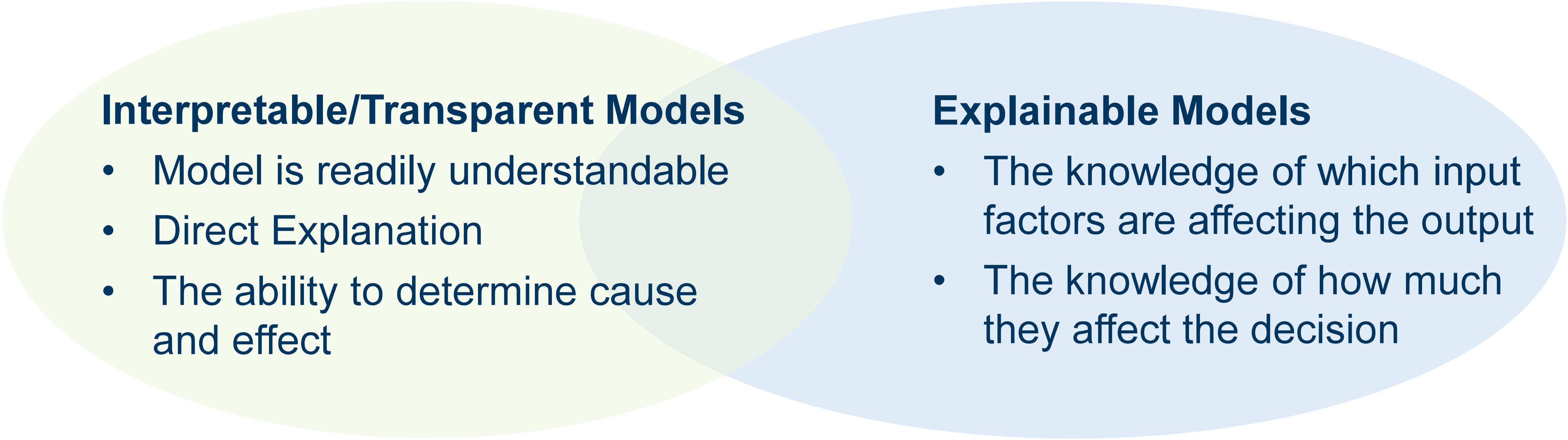
Logistic Regression

Random Forests

Neural Networks

Convolutional Neural Networks

Interpretable vs Explainable Models



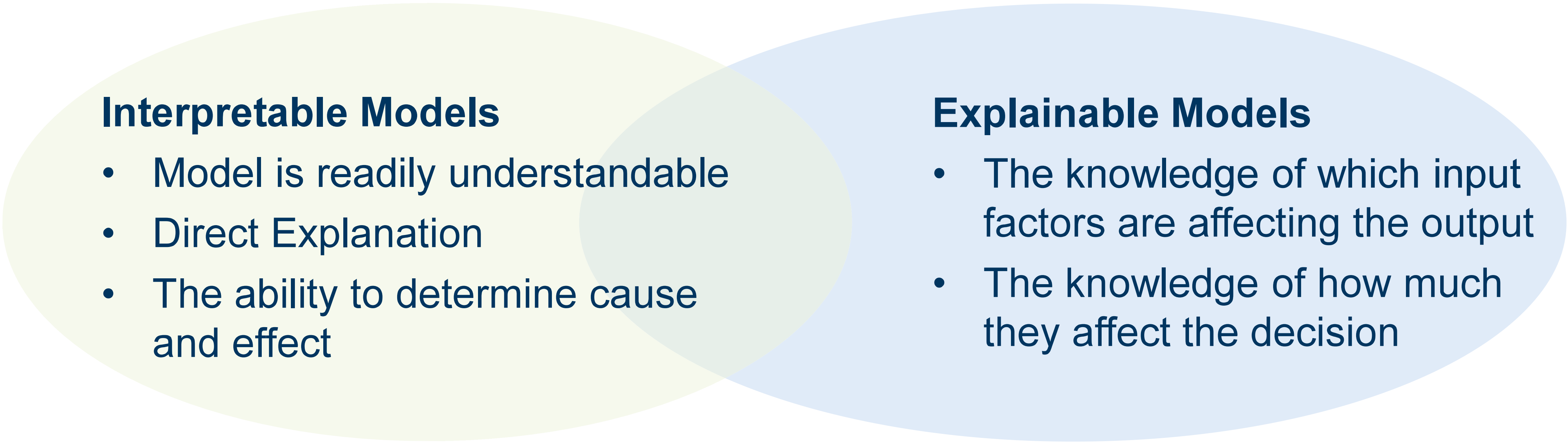
Interpretable/Transparent Models

- Model is readily understandable
- Direct Explanation
- The ability to determine cause and effect

Explainable Models

- The knowledge of which input factors are affecting the output
- The knowledge of how much they affect the decision

Interpretable vs Explainable Models



Interpretable Models

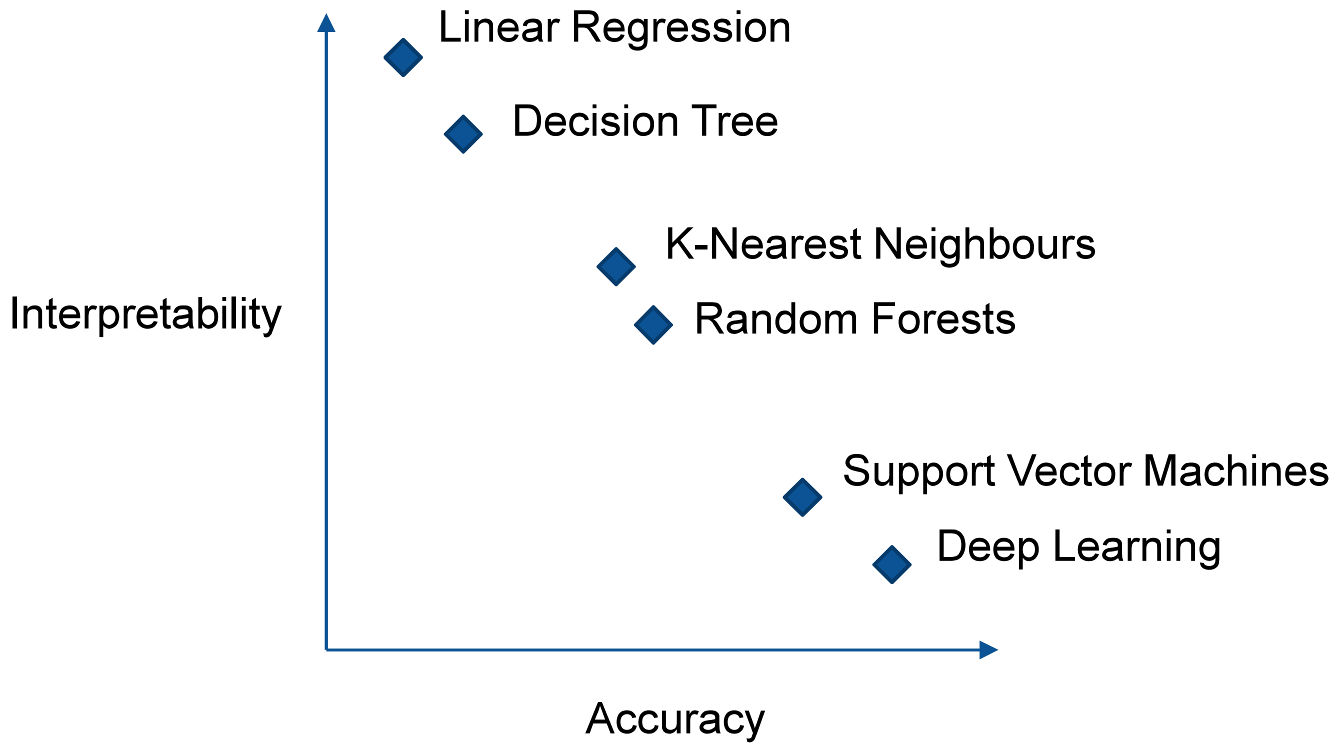
- Model is readily understandable
- Direct Explanation
- The ability to determine cause and effect

Explainable Models

- The knowledge of which input factors are affecting the output
- The knowledge of how much they affect the decision

- The ability to know what each node represents
- The ability to determine cause and effect

Interpretability vs Accuracy



Summary

- Linear models and decision trees are inherently interpretable,
- Complex models can offer better accuracy but they are inherently less interpretable
- Black boxes can be ‘explained’ in a number of different levels:
    - Based on post-hoc models that approximate their function
    - Based on local and global interpretability processes that identify which input factors are most significant and to what degree
    - Based on representation learning that identifies interpretable latent factors
- The ability to determine cause and effect

References

- Arrieta et al. ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 2020.
- Molnar ‘Interpretable Machine Learning - A Guide for Making Black Box Models Explainable’
https://christophm.github.io/interpretable-ml-book/