

Interpretability in AI (?)

Mor Vered

Dept of Data Science and AI

Faculty of IT

What Do We Mean When We Say 'AI'?



Biased Data



VERNON PRATER Prior Offenses 2 armed robberies, 1 attempted armed robbery Subsequent Offenses 1 grand theft LOW RISK 3	BRISHA BORDEN Prior Offenses 4 juvenile misdemeanors Subsequent Offenses None HIGH RISK 8
DYLAN FUGETT LOW RISK 3	BERNARD PARKER HIGH RISK 10

JAMES RIVELLI LOW RISK 3	ROBERT CANNON MEDIUM RISK 6
JAMES RIVELLI Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking Subsequent Offenses 1 grand theft LOW RISK 3	ROBERT CANNON Prior Offense 1 petty theft Subsequent Offenses None MEDIUM RISK 6

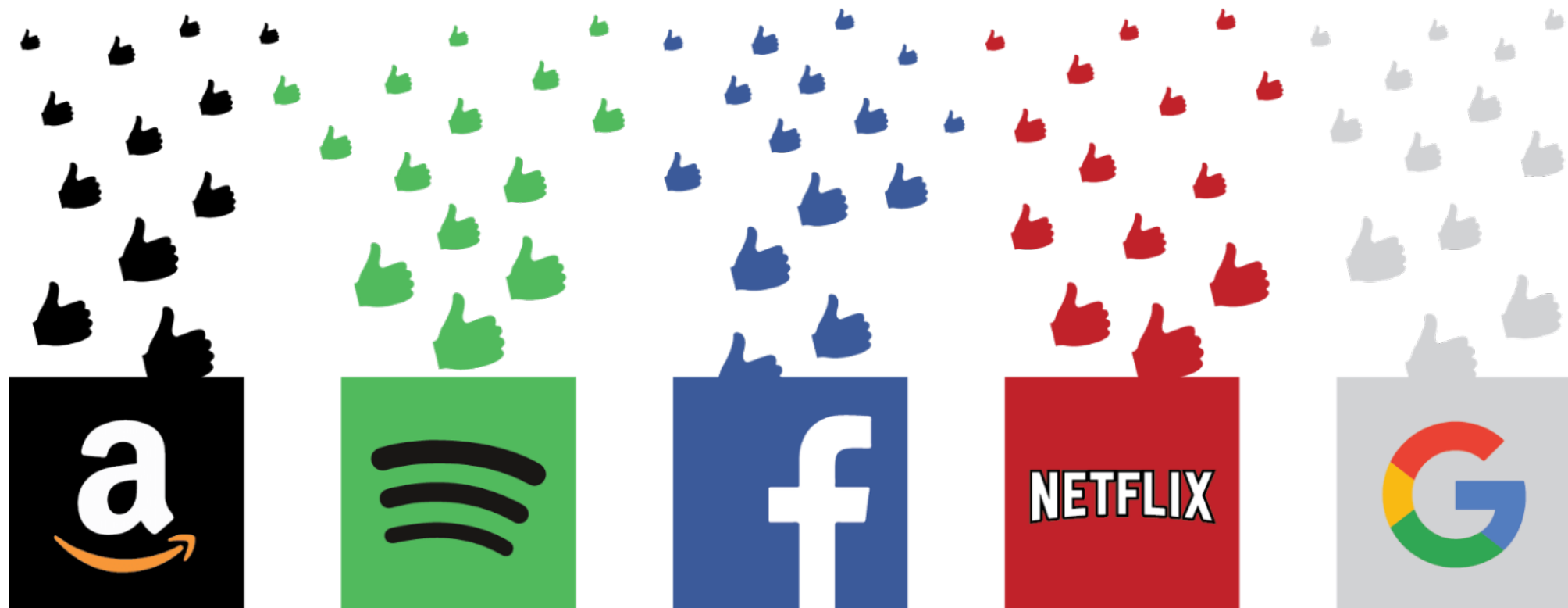
Data Not Fully Representative

Guidelines based on the study of one sex are often generalized and applied to both.

- **Research funding for coronary artery disease in men is far greater than for women. Even though the at risk population of women, which are an older age group, suffer more morbidity and mortality**
- **Prior to 1994 women have been excluded from early studies of most drugs, there is little information about the effects of these drugs in women.**

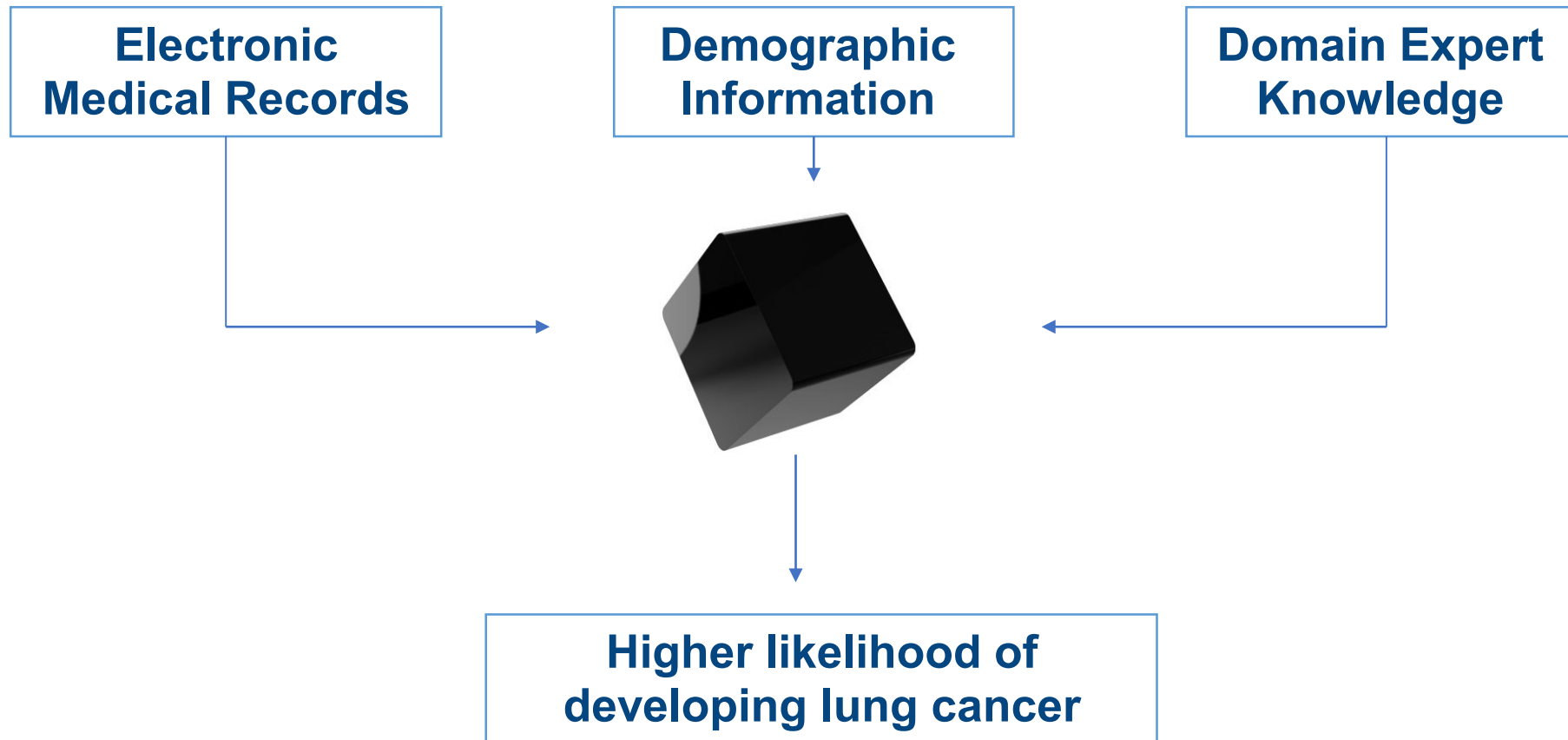


Not Stationary Environment



Sometimes The Output Is Not Enough

Example

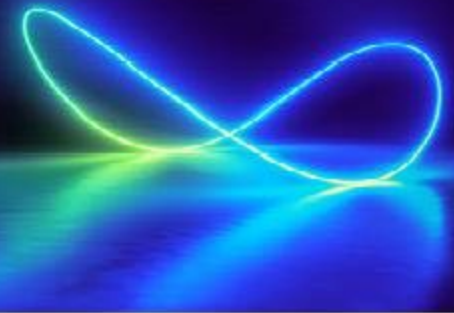


Explainable AI

Interpretability, Transparency

- **Explainability** has been defined as one mode in which an observer may obtain understanding and interpretability of an AI model.
- The purpose of XAI is for the output of an AI system to be understood by humans so as to increase trust and acceptance.
- In contrast with “black box” ML models.
- Originally targeted explanations for ML experts.
- Only recently considering human centred explanations for end-users.

LIME



- **Local Interpretable Model-Agnostic Explanations.**
- **Interpretable - “provide qualitative understanding between the input variables and the response”.**
- **Approximating the output of the classifier locally with a different, interpretable model.**

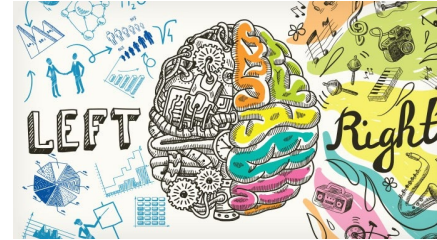
Human – Centered Explainable AI

Multidisciplinary approach

Contrastive explanations



Artificial
Intelligence



Psychology

Levels of
transparency

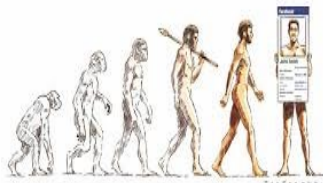
AI?

Human-Centered
Explainable AI



Neuroscience

Causal explanations



Anthropology

Global vs. Local
explanations

Trigger causes vs.
Enabling causes

Questions ?

Mor Vered

Dept of Data Science and AI
Monash University

mor.vered@monash.edu