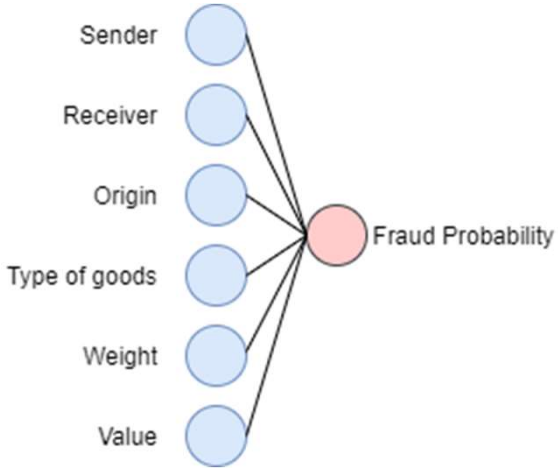
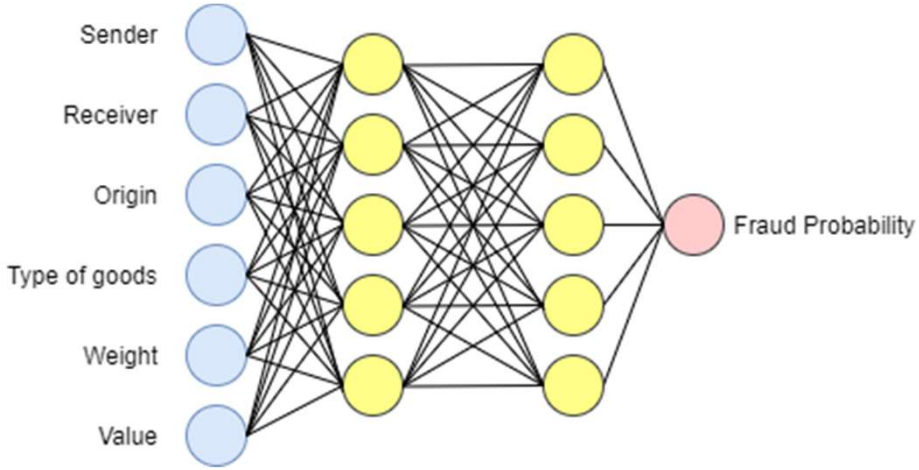
Counterfactual explanations

Transparant



- Straightforward relationship between input and output
- Interpretable

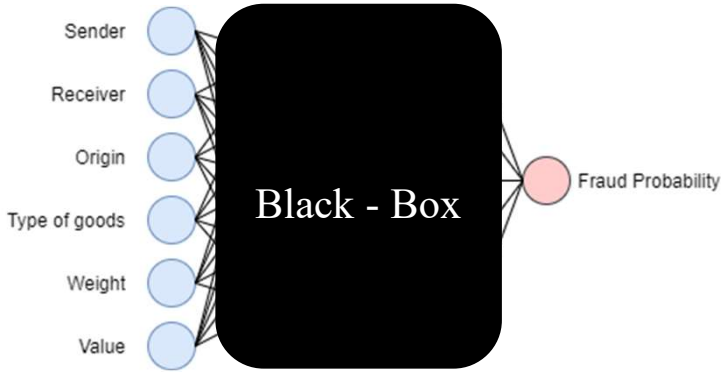
Opaque?



- Non-linear relationships
- Huge amount of parameters
- Not interpretable

Counterfactual explanations

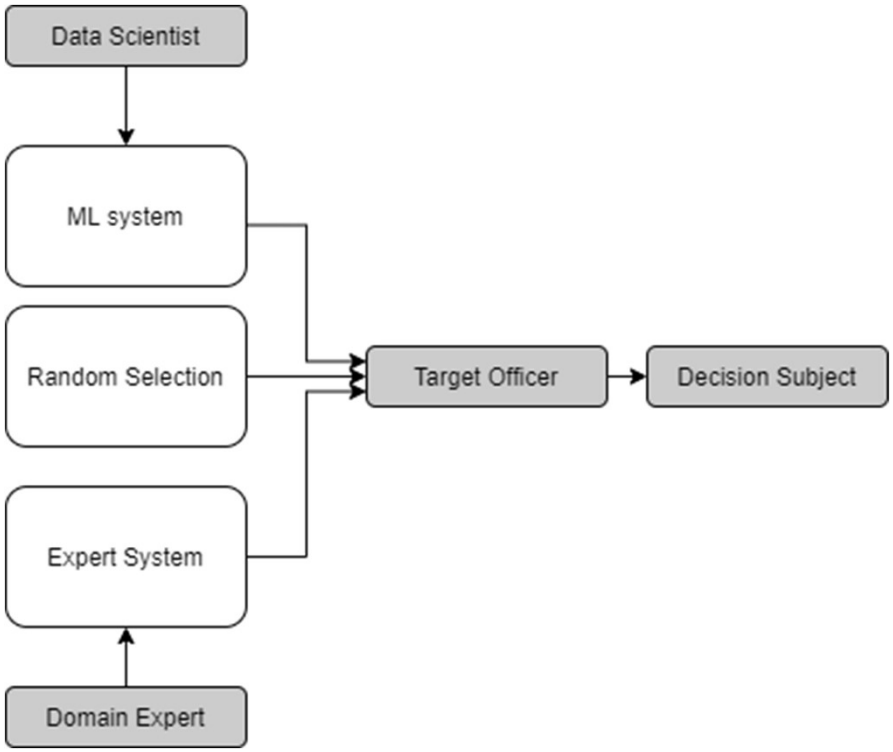
- Local explanation method
- Consider the model a black-box
- “What is the minimal change of input that would change the output?”
- Reduce the feature-space to an interpretable size



Example

Feature	Declaration 1	Declaration 2	Explanation
Sender	Dieter	Dieter	
Receiver	John	John	
Origin	Belgium	Belgium	
Type of goods	Wood	Wood	
Weight	2000 kg	**1900 kg**	-100kg
Value	€ 15 000	**€ 16 000**	€ 1 000

Stakeholders at the Belgian Customs



Data Scientist

- Debugging
- Model Improvement
- Detect Bias

Target officer

- Increase efficiency
- Contribute to model improvement/debugging

Domain Expert

- New insights
- Build complementary systems

Data Subject

- Could be exploited
- Data scientist has to supervise fairness

Antwerp Tax Academy
University of Antwerp