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
<table>
<thead>
<tr>
<th>Feature</th>
<th>Declaration 1</th>
<th>Declaration 2</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sender</td>
<td>Dieter</td>
<td>Dieter</td>
<td></td>
</tr>
<tr>
<td>Receiver</td>
<td>John</td>
<td>John</td>
<td></td>
</tr>
<tr>
<td>Origin</td>
<td>Belgium</td>
<td>Belgium</td>
<td></td>
</tr>
<tr>
<td>Type of goods</td>
<td>Wood</td>
<td>Wood</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>2000 kg</td>
<td><strong>1900 kg</strong></td>
<td>-100kg</td>
</tr>
<tr>
<td>Value</td>
<td>€ 15 000</td>
<td><strong>€ 16 000</strong></td>
<td>€ 1 000</td>
</tr>
</tbody>
</table>
Stakeholders at the Belgian Customs
Data Scientist
• Debugging
• Model Improvement
• Detect Bias

Domain Expert
• New insights
• Build complementary systems

Target officer
• Increase efficiency
• Contribute to model improvement/debugging

Data Subject
• Could be exploited
• Data scientist has to supervise fairness