Responsible AI: Towards a Hybrid Method for Evaluating Data-Driven Decision-making

Cor Steging

XAI Seminar (Imperial College London)

Just a bit about me…

- Bsc & Msc Artificial Intelligence @ University of Groningen
- Machine Learning Engineer @ Slimmer AI
- PhD candidate @ Hybrid Intelligence & RuG

Explainability and Responsibility

- Explainability:
  - Can the AI **explain** its behavior?

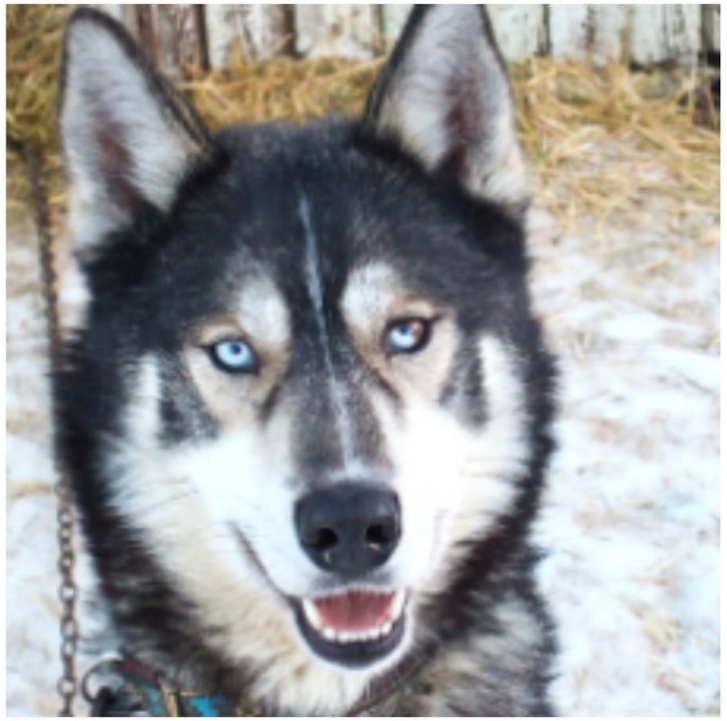
- Responsibility:
  - Does the AI **behave** responsibly?

The questions of responsibility

1. How should the AI behave?
    - Ethical question

2. How do we make the AI behave according to 1?
    - Engineering question

Explainable AI



(a) Husky classified as wolf

(b) Explanation

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pages 1135-1144, 2016.

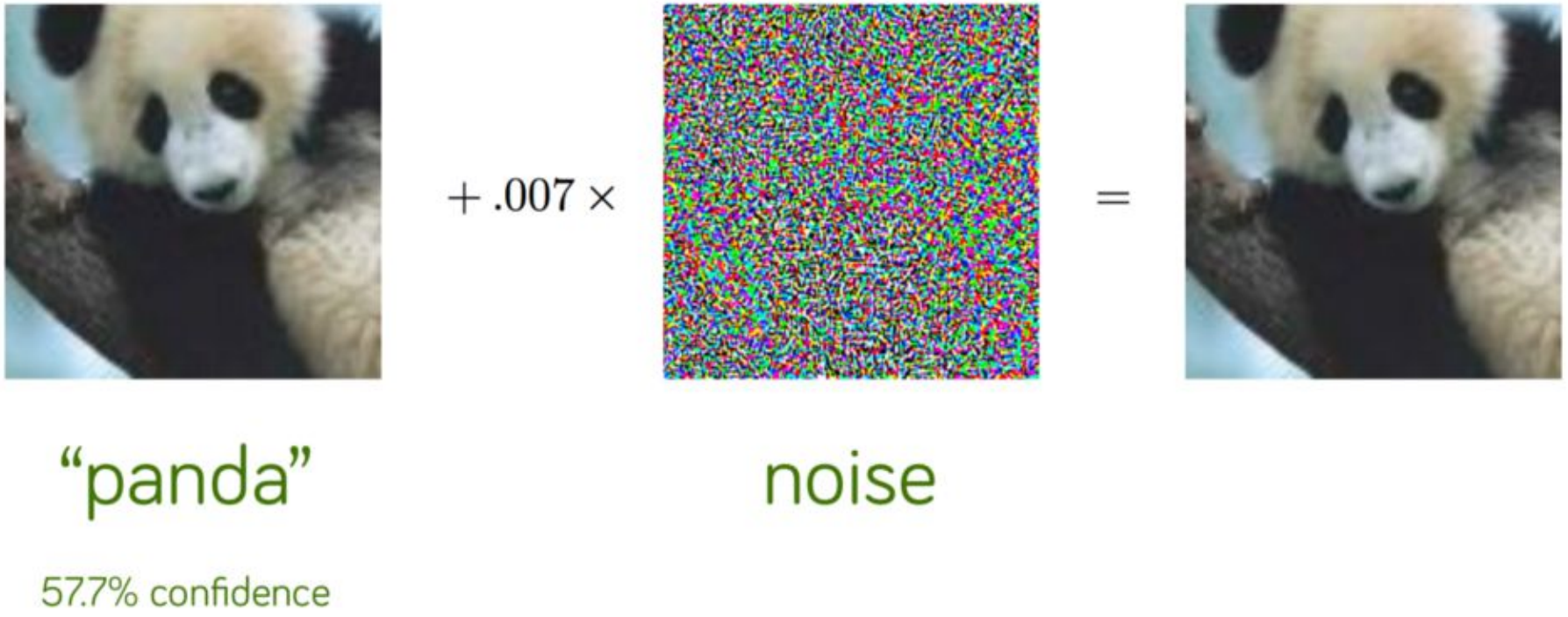
(Ir-)Responsible AI

(a) Wolf classified as wolf

(b) Explanation

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pages 1135-1144, 2016.

Responsible AI



Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015

Responsible AI

- Dutch Childcare Benefit Scandal (Toeslagenaffaire)
- 26K wrongly accused parents
- AI risk assessment system
  - Wrong predictions
  - Wrong reasons

https://nos.nl/artikel/2428355-nog-meer-kinderen-toeslagenaffaire-uit-huis-geplaatst

Responsible AI: knowledge versus data

- Rule-based AI is transparent and uses sound reasoning
  - Based on expert knowledge

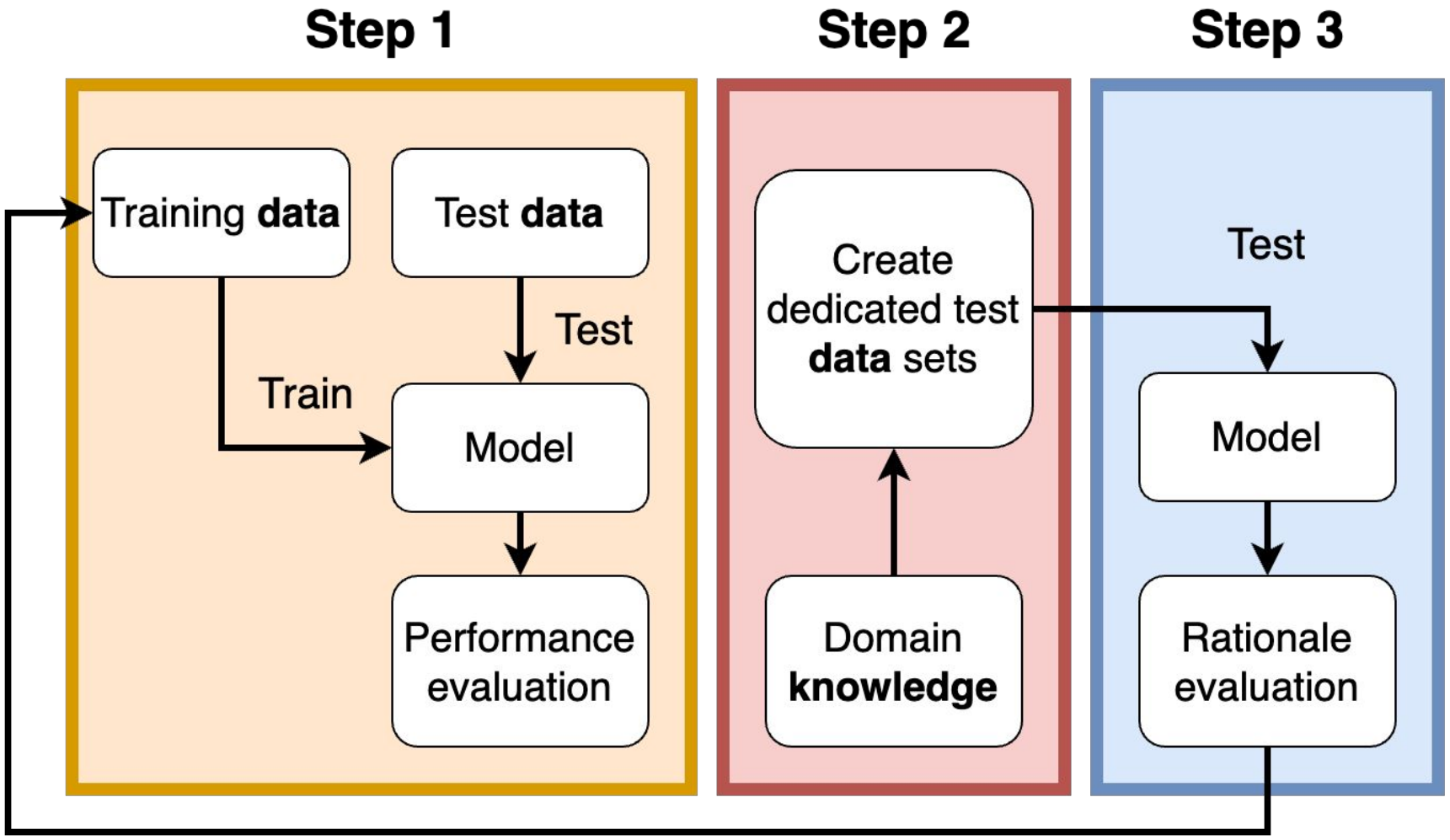
- Do away with data-drive (connectionist) AI?
  - By no means!

Good AI

- AI systems should make the right decisions;
- Make these decisions for the right reasons;
- And explain why they made these decisions.
- Hybrid: Use knowledge to enhance data-driven AI

Current work

- Hybrid method for evaluating the decision-making (rationale) of data-driven AI systems
- We test the behavior of the AI and see if it matches expectations
- Create specialized test cases



Adjust training **data** based on **knowledge** gained from rationale evaluation

Welfare Benefit Domain

Eligible for benefit if and only if:

1. **Pensionable age** (60 for women, 65 for men);
2. At least 4 out of 5 **contributions** were paid;
3. **Spouse** of the patient;
4. Not **absent** from the UK;
5. **Resources** are less than 3000 pounds;
6. Live within **50 miles** of the hospital if the patient is an **'in' patient** or further than **50 miles** if the patient is an **'out' patient**

T. Bench-Capon. 1993. Neural networks and open texture. In Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93). ACM, New York, 292–297.

Welfare benefit domain

$$
\begin{aligned}
&Eligible(x) \iff C_1(x) \wedge C_2(x) \wedge C_3(x) \wedge C_4(x) \wedge C_5(x) \wedge C_6(x)\\
&C_1(x) \iff (Gender(x) = female \wedge Age(x) \geq 60) \vee\\
&\qquad\quad (Gender(x) = male \wedge Age(x) \geq 65)\\
&C_2(x) \iff |Con_1(x), Con_2(x), Con_3(x), Con_4(x), Con_5(x)| \geq 4\\
&C_3(x) \iff Spouse(x)\\
&C_4(x) \iff \neg Absent(x)\\
&C_5(x) \iff \neg Resources(x) \geq 3000\\
&C_6(x) \iff (Type(x) = in \wedge Distance(x) < 50) \vee\\
&\qquad\quad (Type(x) = out \wedge Distance(x) \geq 50)
\end{aligned}
$$

T. Bench-Capon. 1993. Neural networks and open texture. In Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93). ACM, New York, 292–297.

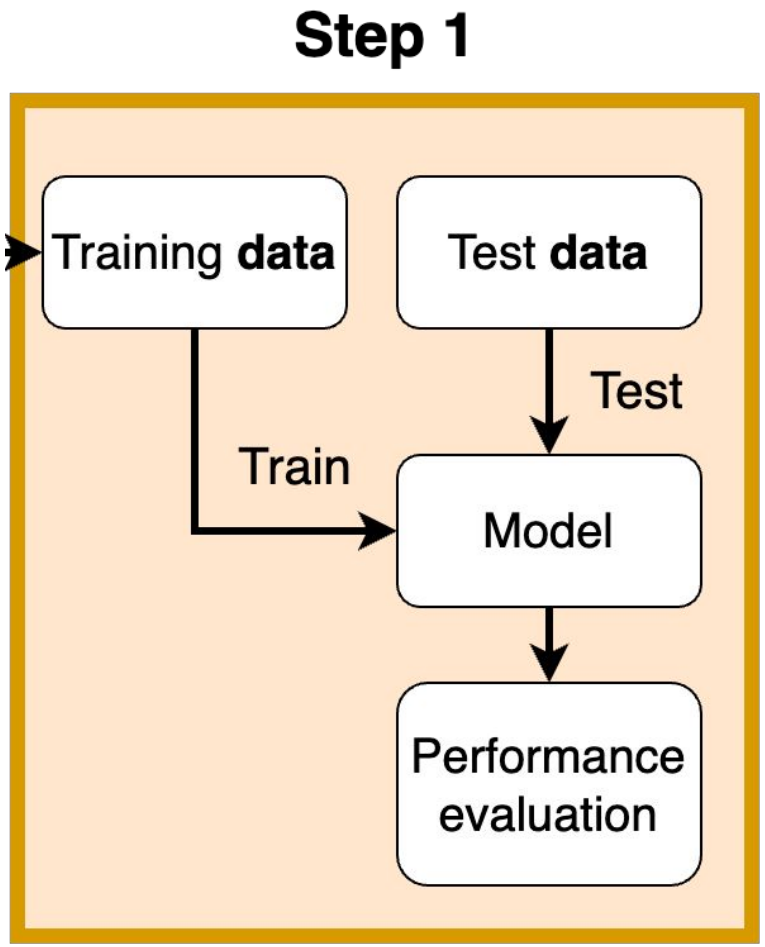
Welfare Benefit Datasets

- Create data:
  - 50% eligibility
  - Ineligible instances due to **multiple** condition
  - A training dataset and a test dataset

Age	Gender	Con1	Con2	Con3	Con4	Con5	Spouse	Absent	Resources	Patient type	Distance	Eligible
84	Female	0	1	1	1	1	1	0	1569	Out	74	True

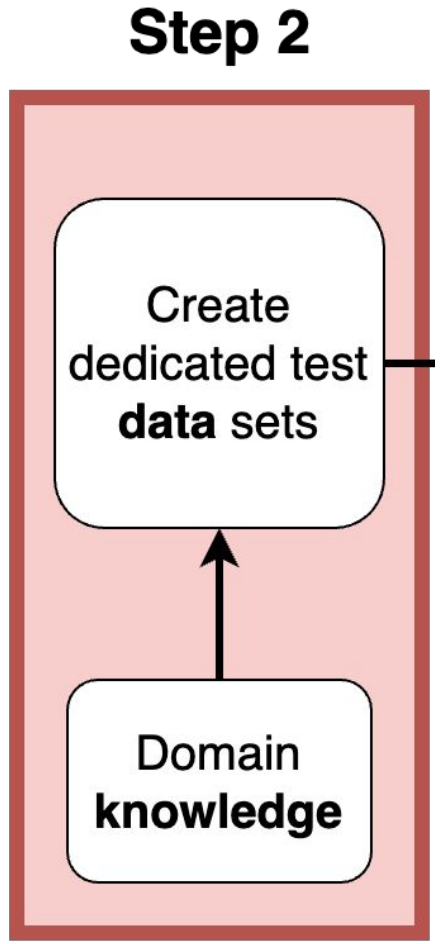
Step 1:

- Train a neural network on training data
- Test the network using test data
- Accuracy of 99.79%



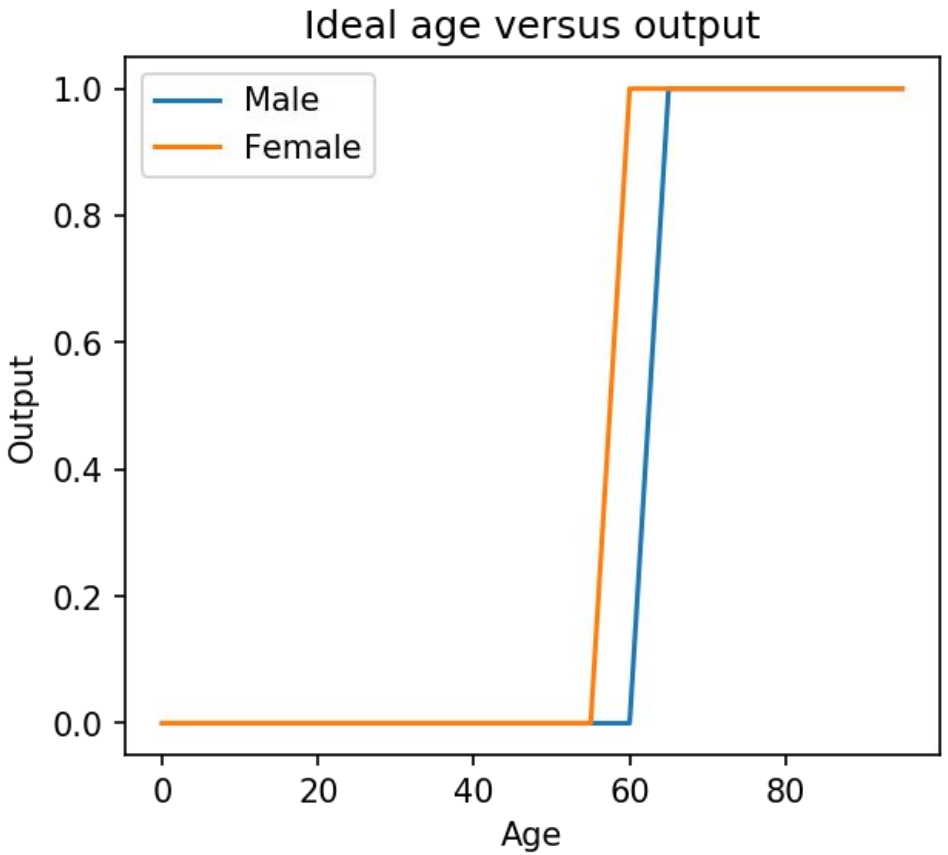
Step 2:

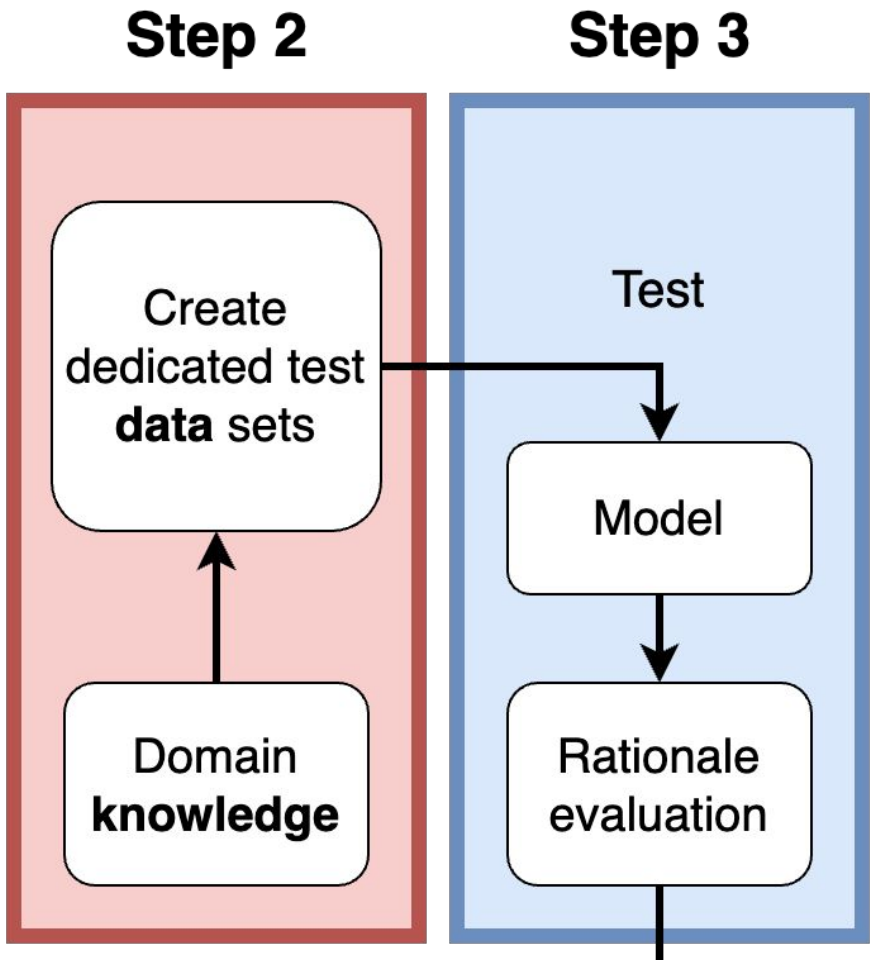
- Domain is defined by 6 conditions
  - Neural network should have learned those
- Focus on condition 1:
  Pensionable age (60 for women, 65 for men)
- Create a dedicated test dataset where
  all conditions are satisfied except Condition 1
  - Value of condition 1 are varied randomly
  - Eligibility is therefore determined solely by C1
  - Neural network can only perform well if it learned C1



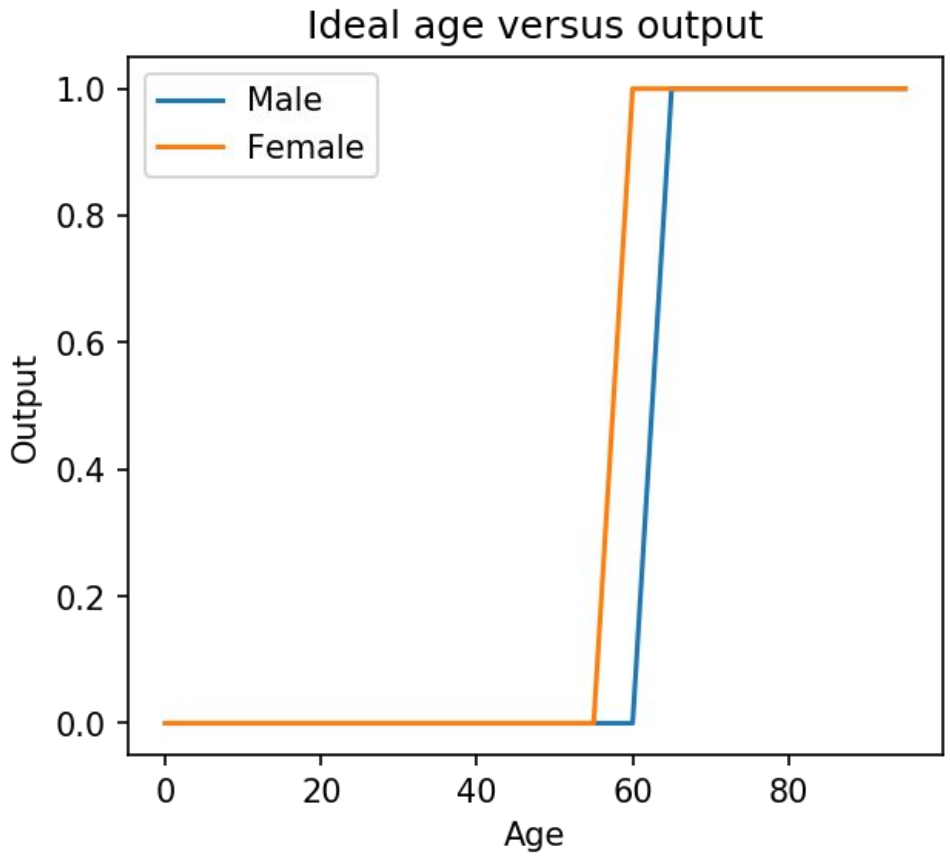
Step 3:

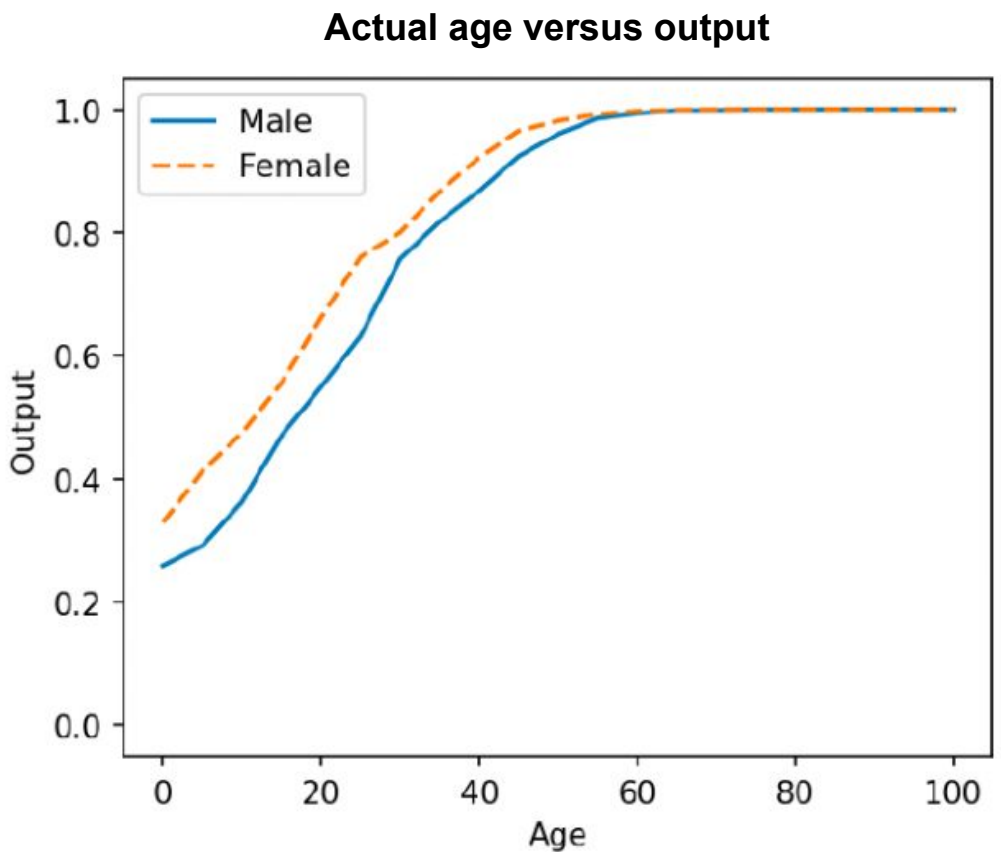
› Accuracy on dedicated test set is 63.24%





Step 3:


Ideal age versus output
Male
Female
Output
Age
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80


Actual age versus output
Male
Female
Output
Age
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

Welfare benefit domain

$$
\begin{aligned}
&Eligible(x) \iff C_1(x) \wedge C_2(x) \wedge C_3(x) \wedge C_4(x) \wedge C_5(x) \wedge C_6(x)\\
&C_1(x) \iff (Gender(x) = female \wedge Age(x) \geq 60) \vee\\
&\qquad\quad (Gender(x) = male \wedge Age(x) \geq 65)\\
&C_2(x) \iff |Con_1(x), Con_2(x), Con_3(x), Con_4(x), Con_5(x)| \geq 4\\
&C_3(x) \iff Spouse(x)\\
&C_4(x) \iff \neg Absent(x)\\
&C_5(x) \iff \neg Resources(x) \geq 3000\\
&C_6(x) \iff (Type(x) = in \wedge Distance(x) < 50) \vee\\
&\qquad\quad (Type(x) = out \wedge Distance(x) \geq 50)
\end{aligned}
$$

T. Bench-Capon. 1993. Neural networks and open texture. In Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93). ACM, New York, 292–297.

Simplified Welfare Benefit domain

$$Eligible(x) \iff C_1(x) \wedge C_6(x)$$

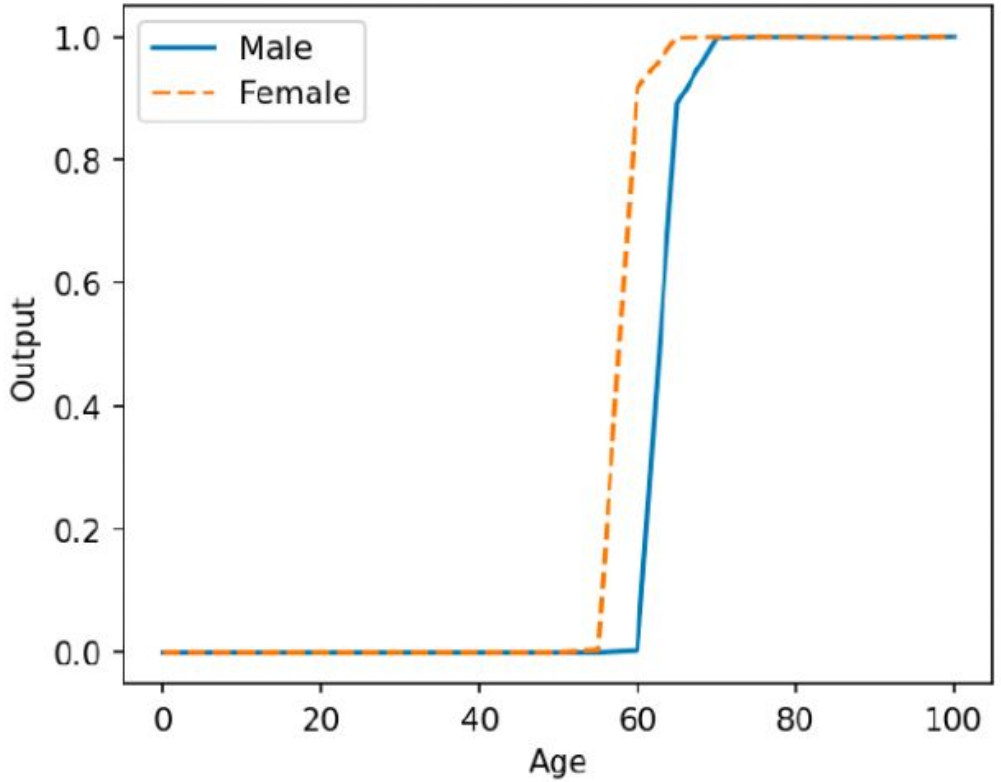
$$C_1(x) \iff (Gender(x) = female \wedge Age(x) \geq 60) \vee (Gender(x) = male \wedge Age(x) \geq 65)$$

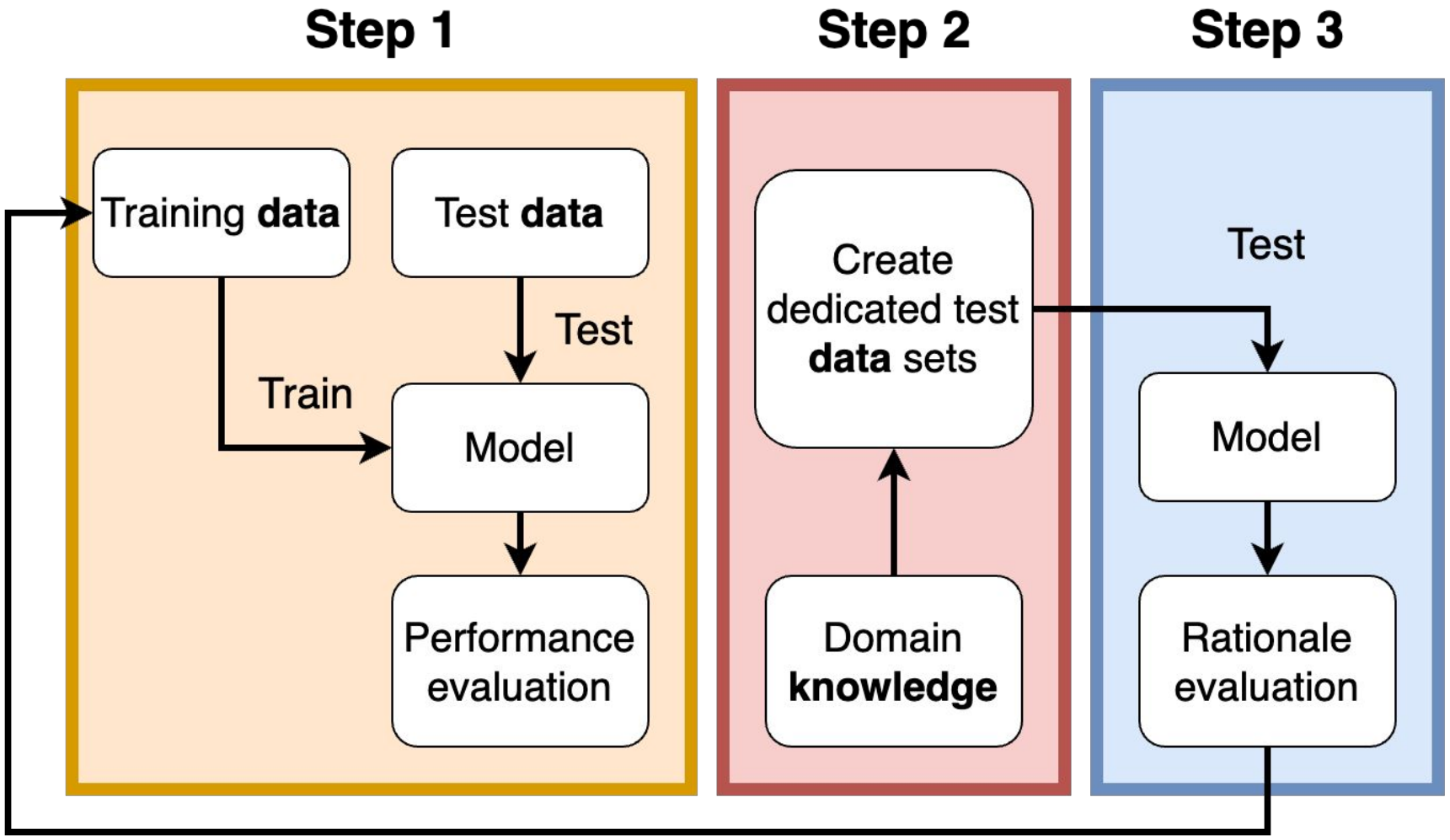
$$C_6(x) \iff (Type(x) = in \wedge Distance(x) < 50) \vee (Type(x) = out \wedge Distance(x) \geq 50)$$

T. Bench-Capon. 1993. Neural networks and open texture. In Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93). ACM, New York, 292–297.

Simplified Welfare Benefit Domain

- Accuracy on regular test set:
  - 99.48%
- Accuracy on dedicated test set:
  - 99.70%





Adjust training **data** based on **knowledge** gained from rationale evaluation

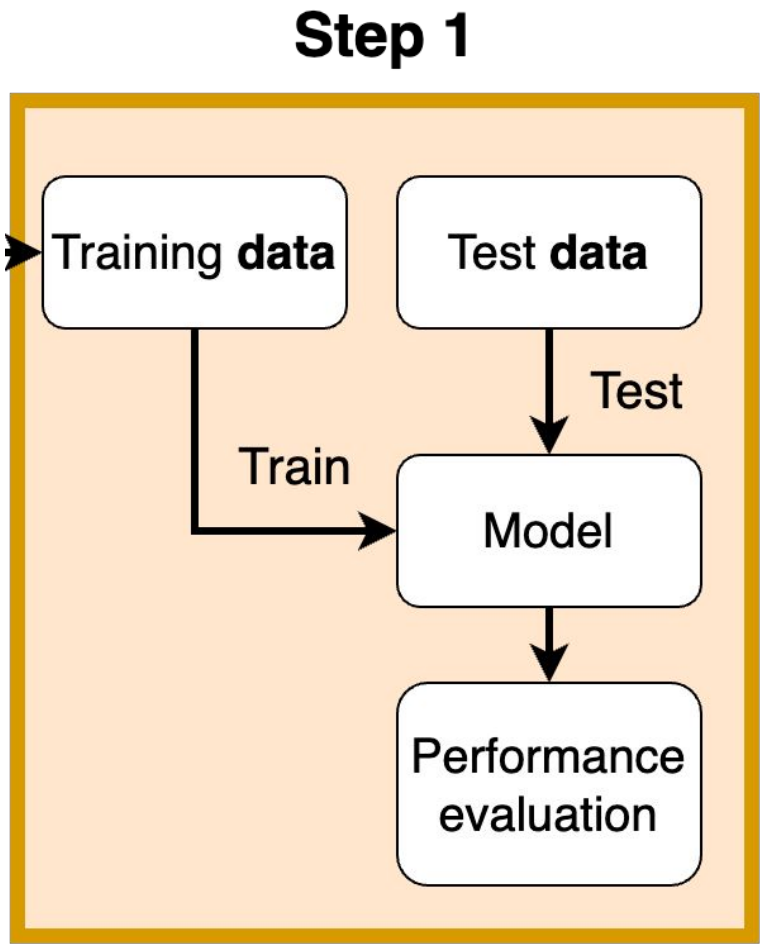
Welfare Benefit Datasets

- Create data:
  - 50% eligibility
  - Ineligible instances due to **a single** condition
  - A training dataset and a test dataset
- Tailored training data

Age	Gender	Con1	Con2	Con3	Con4	Con5	Spouse	Absent	Resources	Patient type	Distance	Eligible
84	Female	0	1	1	1	1	1	0	1569	Out	74	True

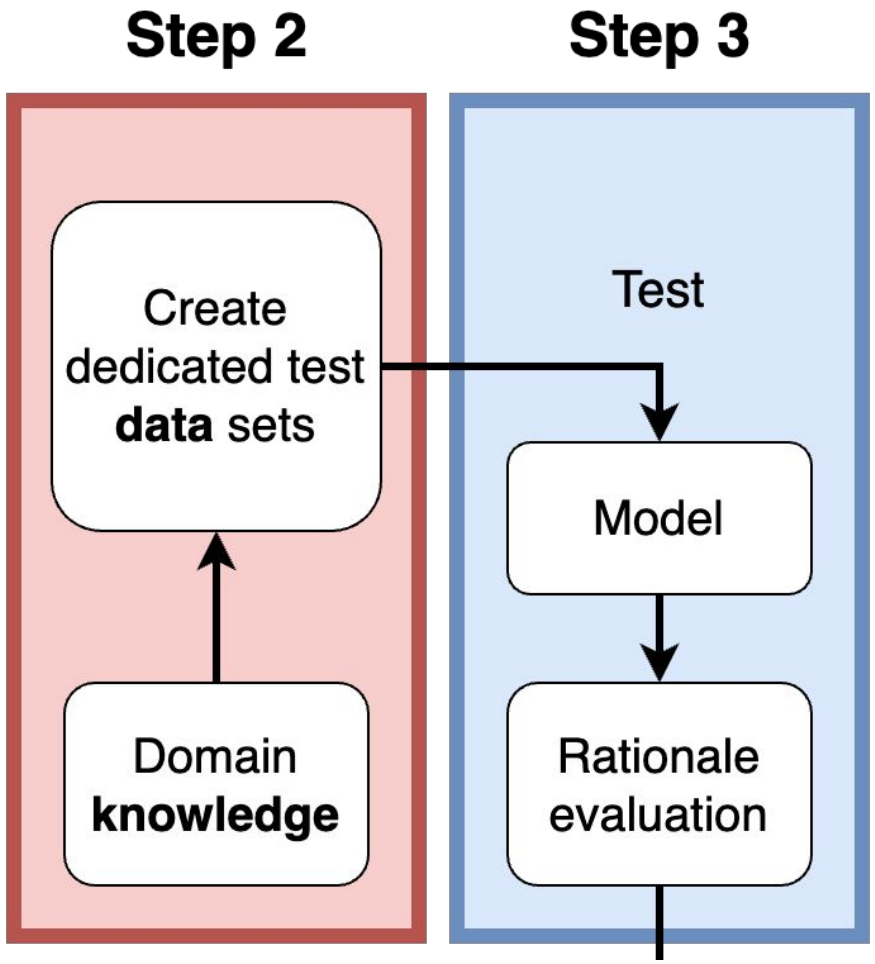
Step 1: (tailored)

- Train a neural network on training data
- Test the network using test data
- Accuracy of 98.03%
  - Previous: 99.79%

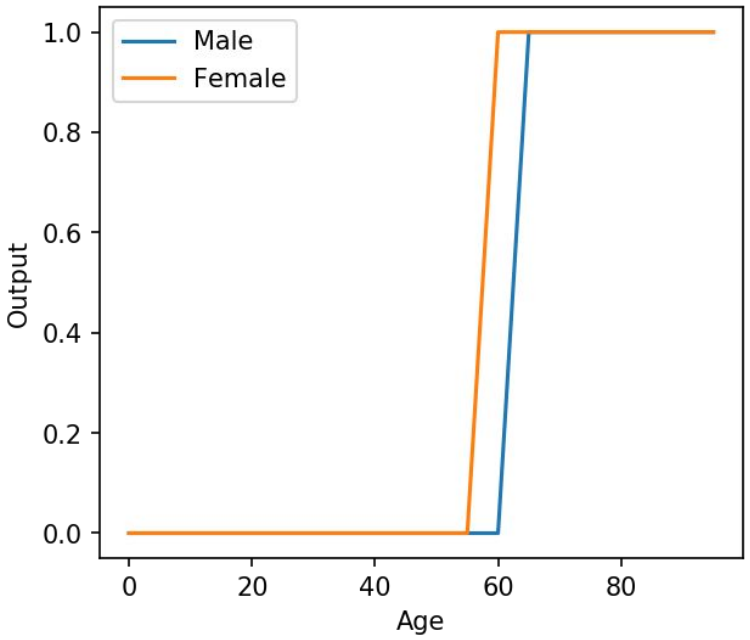


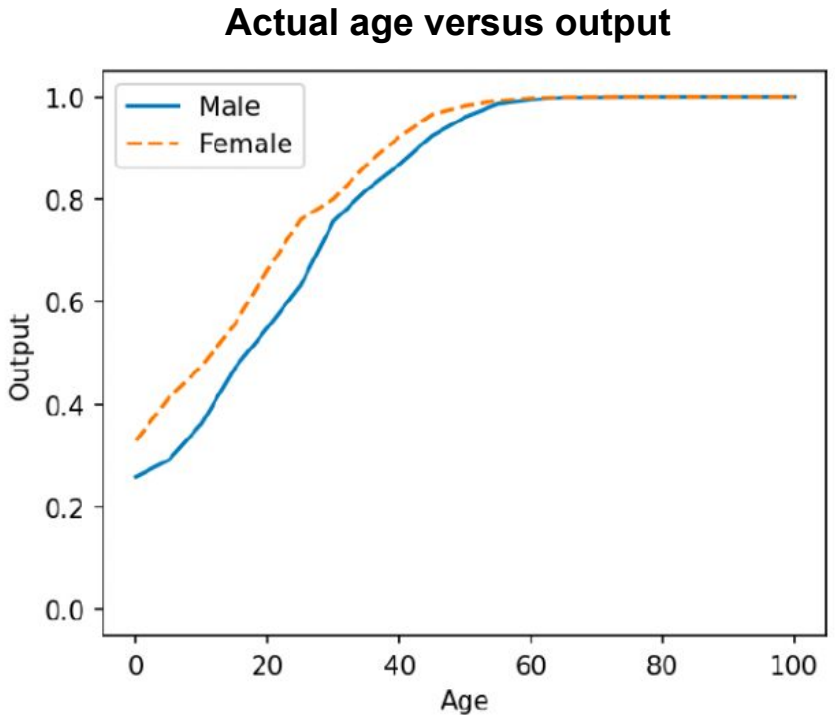
Step 3: (tailored)

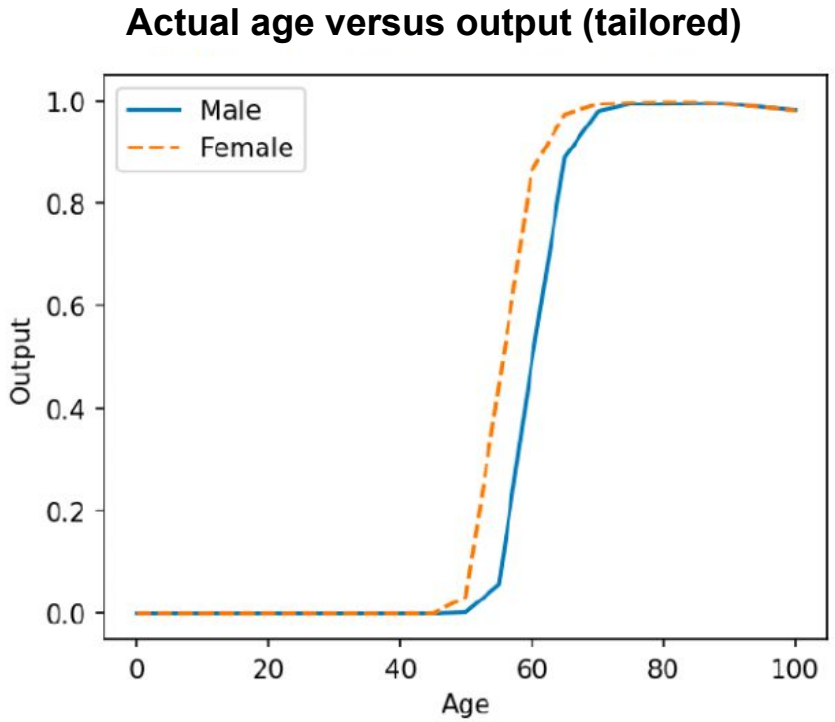
- Accuracy on dedicated test set is 97.66%
  - Previous: 63.24%



Step 3: Tailored







Method for Rationale Evaluation

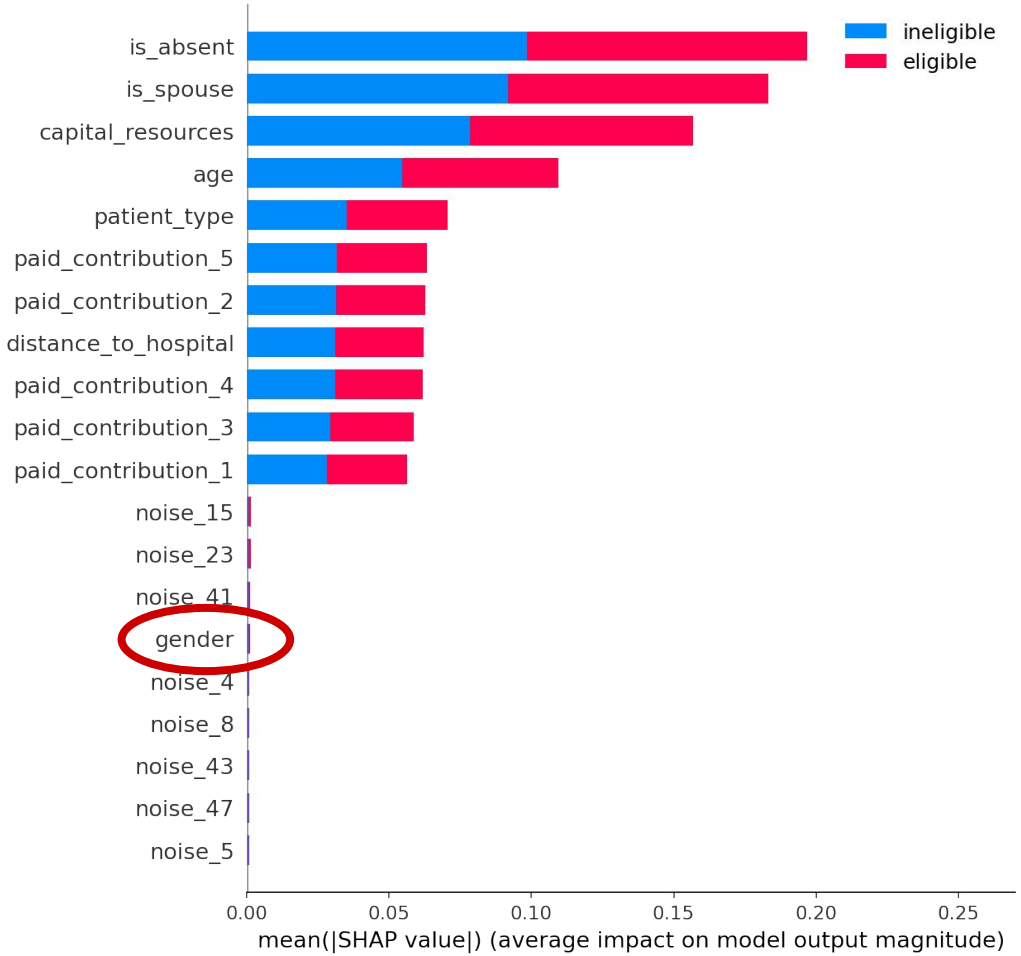
- Evaluate the decision making of machine learning systems
  - Potentially improve
- Model agnostic
- Hybrid: both data and knowledge
  - + human and machine

Explainable AI

- Can XAI be used to expose unsound decision-making?
- Apply SHAP and LIME to our networks:
  - **Original** training data:
    - **Unsound** decision-making
  - **Tailored** training data:
    - **Sound** decision-making

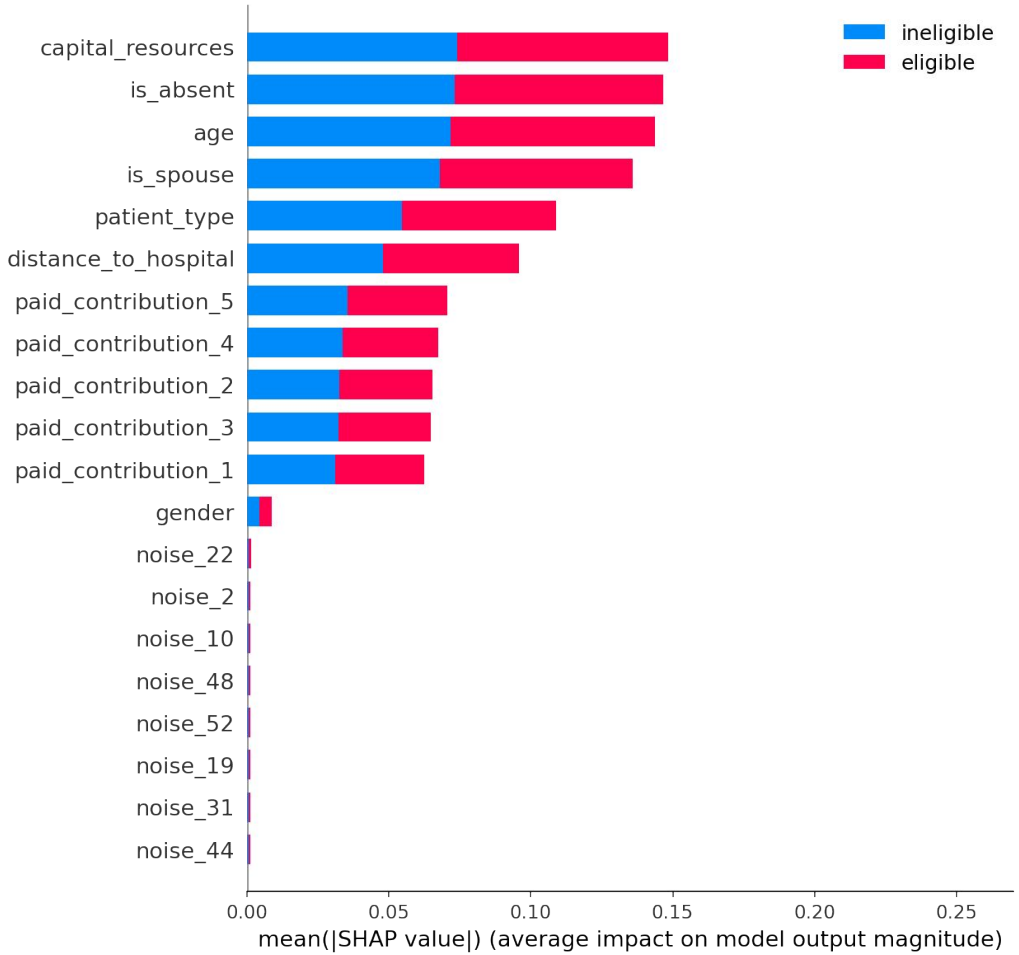
Explainable AI: original training data

- **12** features are relevant
- **11** have high impact values
- Gender can be accounted for
- Conclusion:
  - High accuracy
  - Correct features



Explainable AI: tailored training data

- **12** features are relevant
- **12** have high impact values



Explainable AI

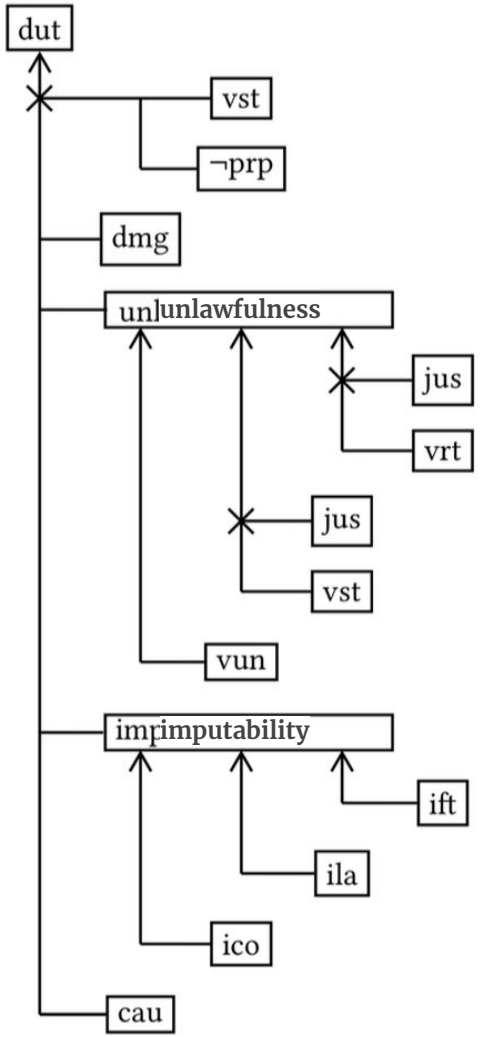
- Can XAI be used to expose unsound decision-making?
  - Yes, but it cannot guarantee sound decision-making

- XAI can incorrectly suggest a sound rationale

Tort law domain

- Real life domain
  - Statutory law
  - Is there a duty to repair damages?

$$
\begin{aligned}
dut(x) &\iff c_1(x) \wedge c_2(x) \wedge c_3(x) \wedge c_4(x) \wedge c_5(x) \\
c_1(x) &\iff cau(x) \\
c_2(x) &\iff ico(x) \vee ila(x) \vee ift(x) \\
c_3(x) &\iff vun(x) \vee (vst(x) \wedge \neg jus(x)) \vee (vrt(x) \wedge \neg jus(x)) \\
c_4(x) &\iff dmg(x) \\
c_5(x) &\iff \neg(vst(x) \wedge \neg prp(x))
\end{aligned}
$$

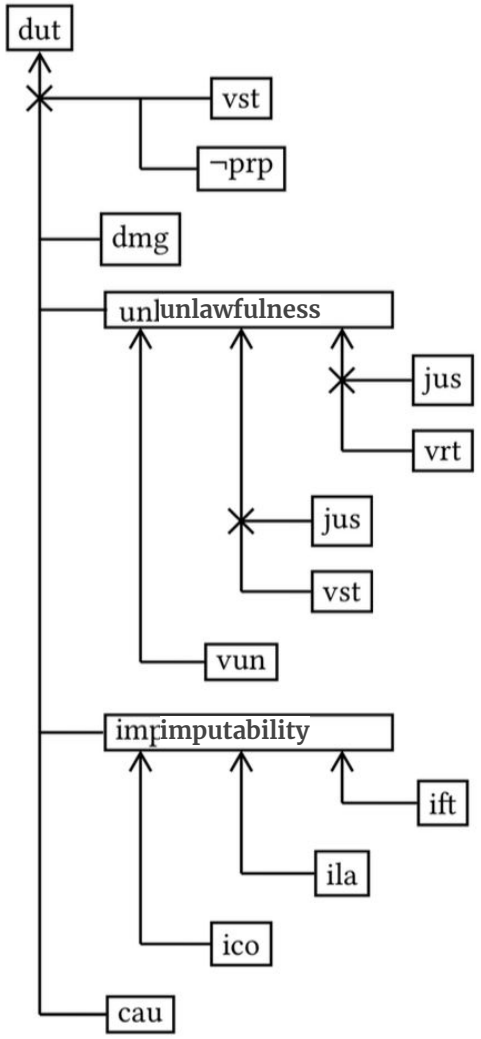


(A) Arguments and their attacks in the domain of Dutch tort law.

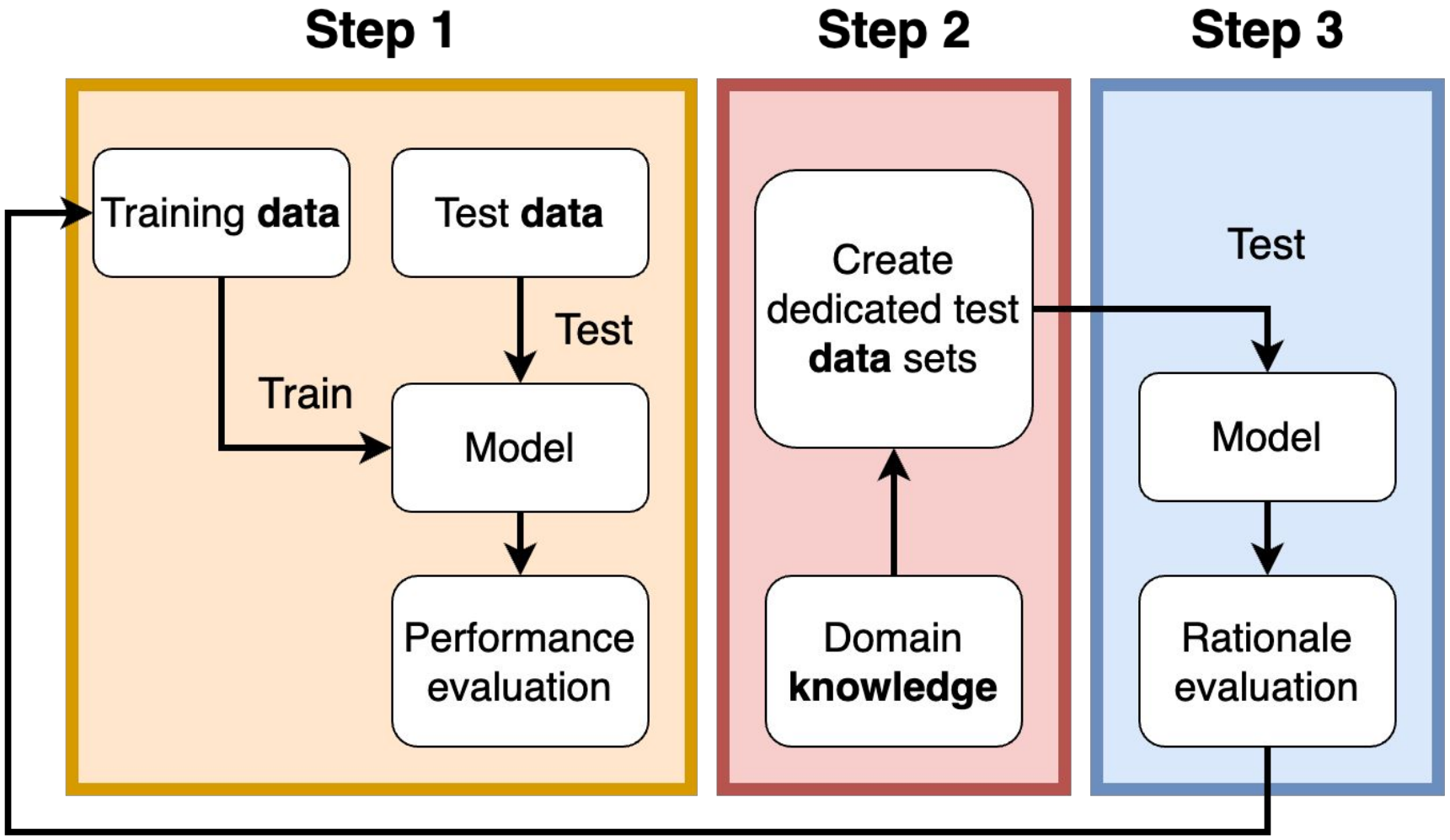
Tort law datasets

- Create data:
  - 50% duty to repair damages
  - Create a small subset (500 instance)
  - A training dataset and a test dataset

vst	prp	dmg	jus	vrt	vun	ift	ila	ico	cau	dut
0	1	0	1	1	1	1	1	0	0	1



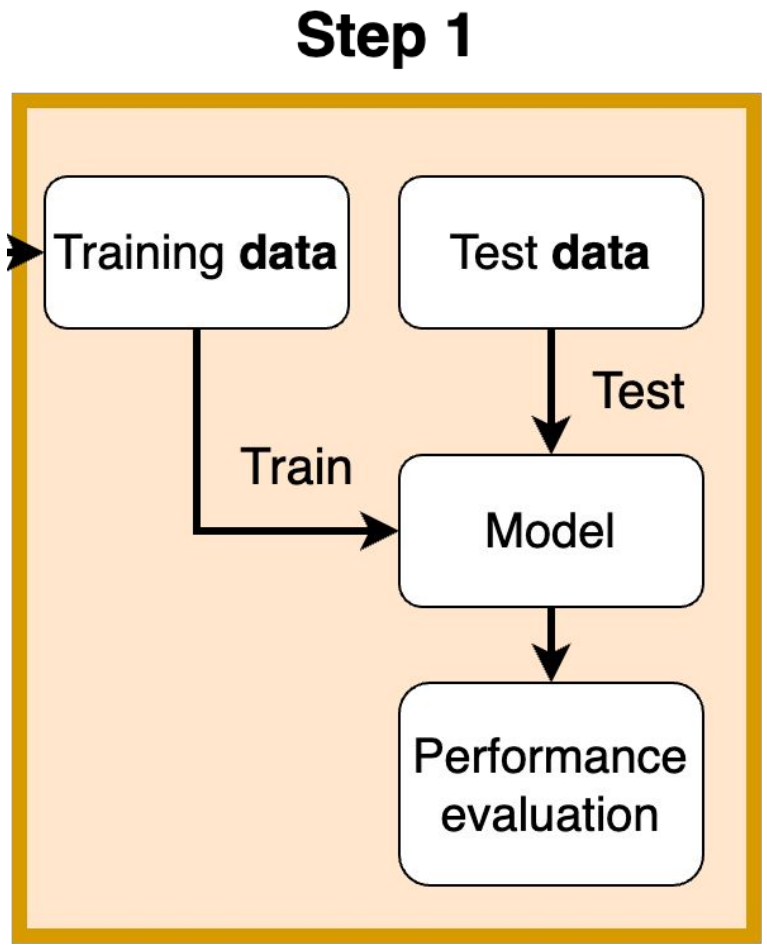
(A) Arguments and their attacks in the domain of Dutch tort law.



Adjust training **data** based on **knowledge** gained from rationale evaluation

Step 1: Tort Law

- Train a neural network on training data
- Test the network using test data
- Accuracy of 98.23%

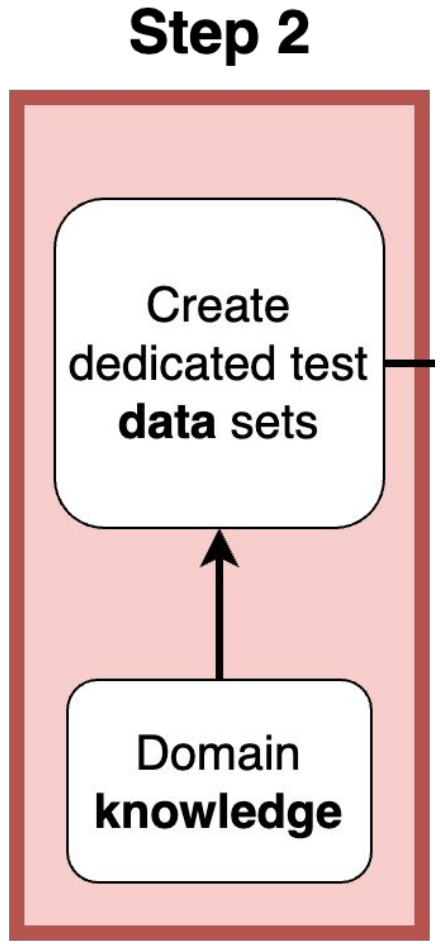


Step 2: Tort Law

- Domain is defined by 5 conditions
  - Neural network should have learned those

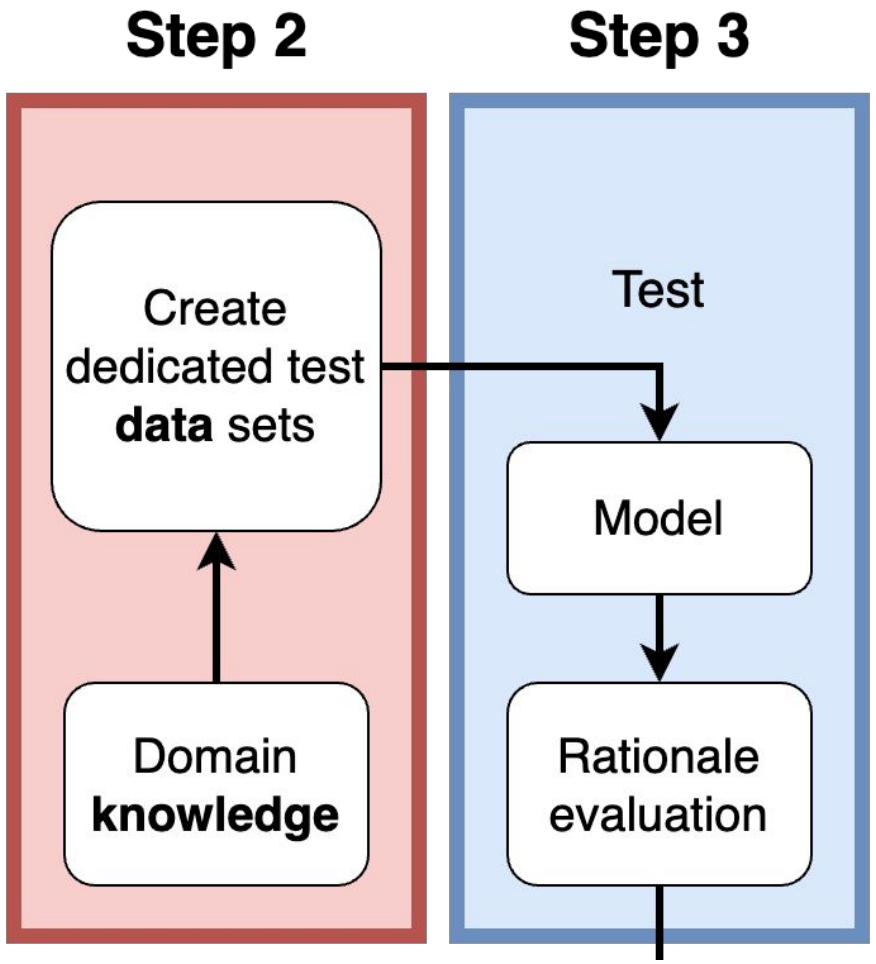
- Focus on condition C2 (imputability):

$$
\begin{aligned}
dut(x) &\iff c_1(x) \wedge c_2(x) \wedge c_3(x) \wedge c_4(x) \wedge c_5(x) \\
c_1(x) &\iff cau(x) \\
c_2(x) &\iff ico(x) \vee ila(x) \vee ift(x) \\
c_3(x) &\iff vun(x) \vee (vst(x) \wedge \neg jus(x)) \vee (vrt(x) \wedge \neg jus(x)) \\
c_4(x) &\iff dmg(x) \\
c_5(x) &\iff \neg(vst(x) \wedge \neg prp(x))
\end{aligned}
$$



Step 3: Tort Law

- Accuracy on dedicated test set is 91.45%
  - Label balance is 87.5% : 12.5%
  - Matthews Correlation Coefficient: 0.2582



Future work

- How do we apply the method to domains without inherent knowledge structure?
  - Test what you do know.

- Making the method work everywhere:
  - Fictional domain, Artificial data (Welfare Benefit)
  - ↓
  - Non-fictional domain, Artificial data (Tort Law)
  - ↓
  - Non-fictional domain, Real data (Court Case predictions)

Conclusion

- Knowledge-driven, model-agnostic method for evaluating decision-making
- Evaluate and improve AI behavior
- Systems can perform well for the wrong reasons
- XAI cannot guarantee a sound rationale

Responsible AI: Towards a Hybrid Method for Evaluating Data-Driven Decision-making

Cor Steging

XAI Seminar (Imperial College London)