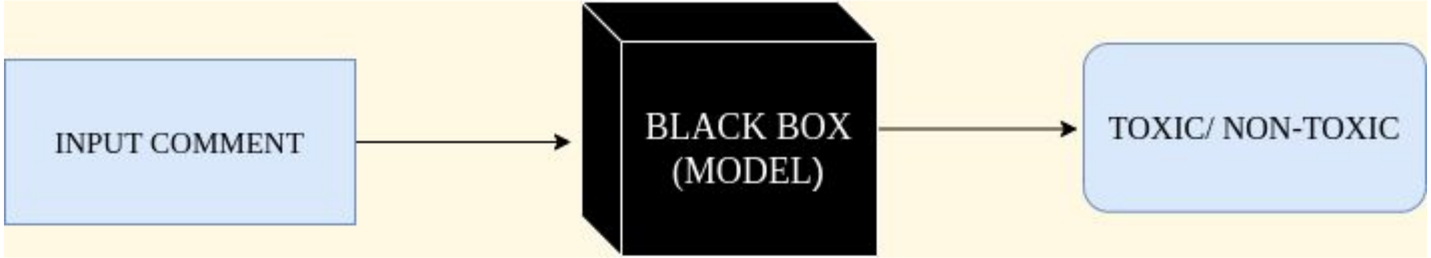
Explainable Artificial Intelligence

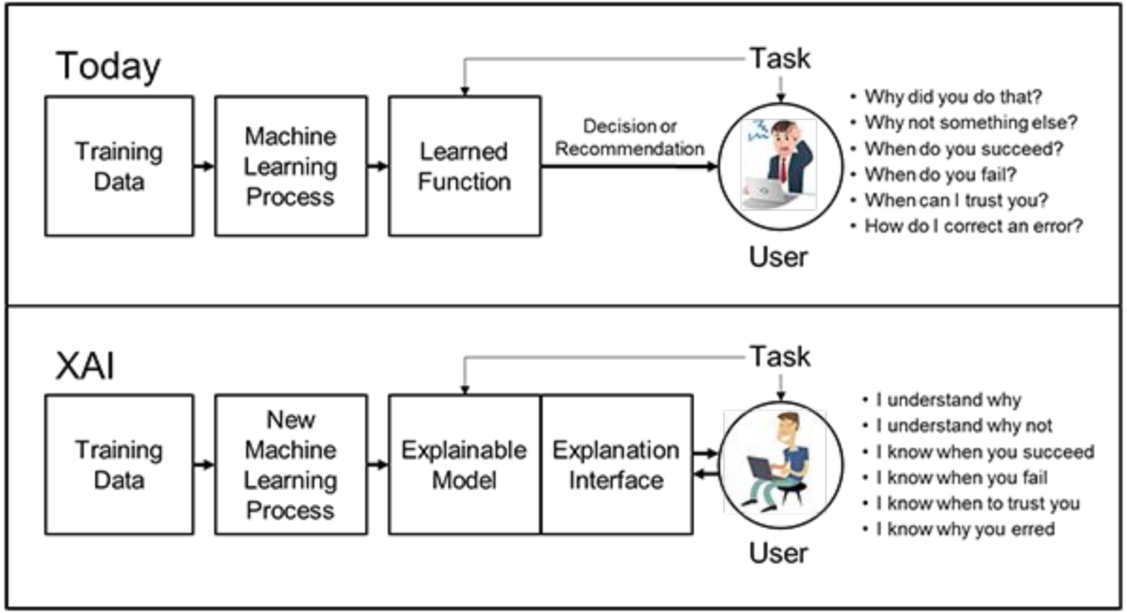
‘XAI’

Need



- Why did the model make a specific prediction or description
- When did the AI system fail

Concept



Today
Training Data
Machine Learning Process
Learned Function
Decision or Recommendation
Task
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
Training Data
New Machine Learning Process
Explainable Model
Explanation Interface
Task
User
• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

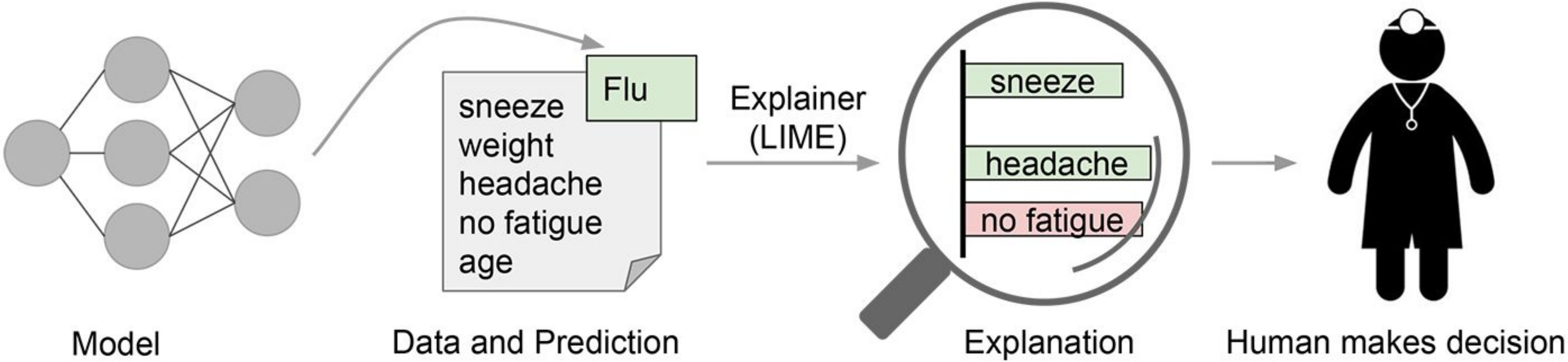
Ideal Model Classifier

1. Interpretable
2. Local Fidelity
3. Model Agnostic
4. Global perspective

1. Provide qualitative understanding between input variables and responses
2. Might not be possible for an explanation to be completely faithful unless it itself describes the model completely
3. Able to explain any model
4. Explain a representative set to user to give global intuition of the model

Locally Interpretable Model-Agnostic Explanation

- A python library which tries to solve for model interpretability by producing locally faithful explanations
- LIME is a novel explanation technique that explains prediction of any classifier in an interpretable and faithful manner by learning an interpretable model locally around the prediction



Flu

sneeze
weight
headache
no fatigue
age

Explainer
(LIME)

sneeze
headache
no fatigue

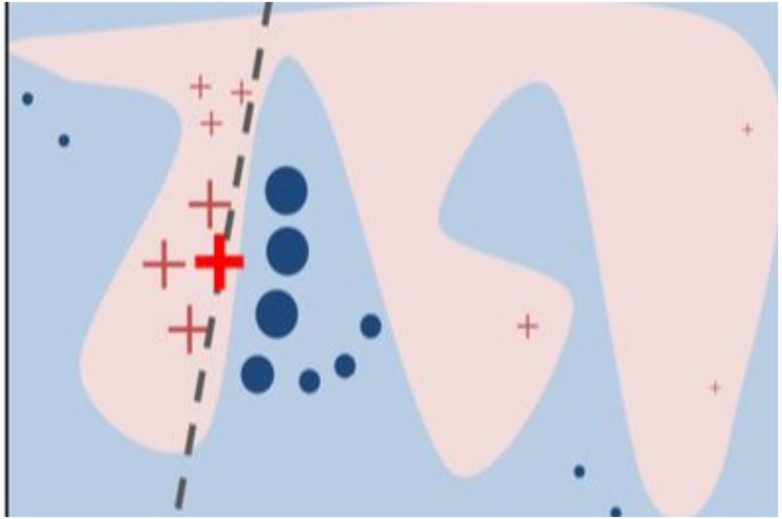
Model

Data and Prediction

Explanation

Human makes decision

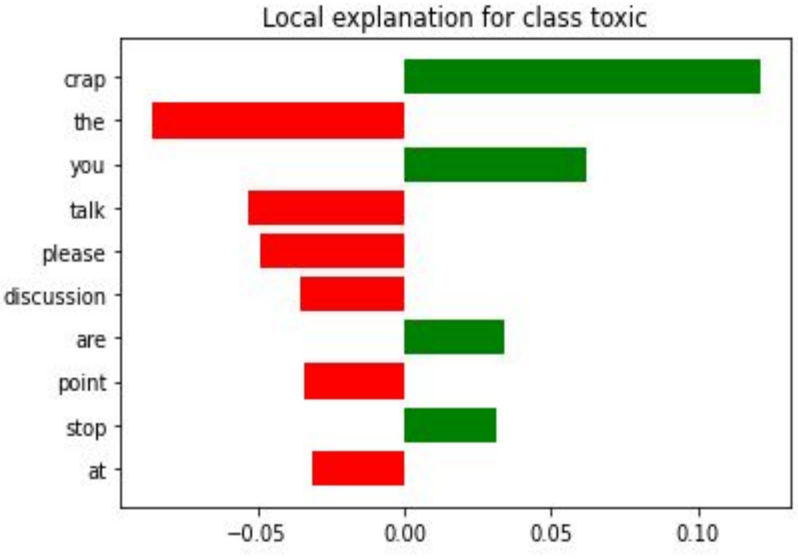
Working



- The black box model's complex decision function f is represented by the blue/pink background
- The red cross is the instance being explained
- The dashed line is the learned explanation that is locally faithful

LIME on Imbalanced data

Logistic Regression

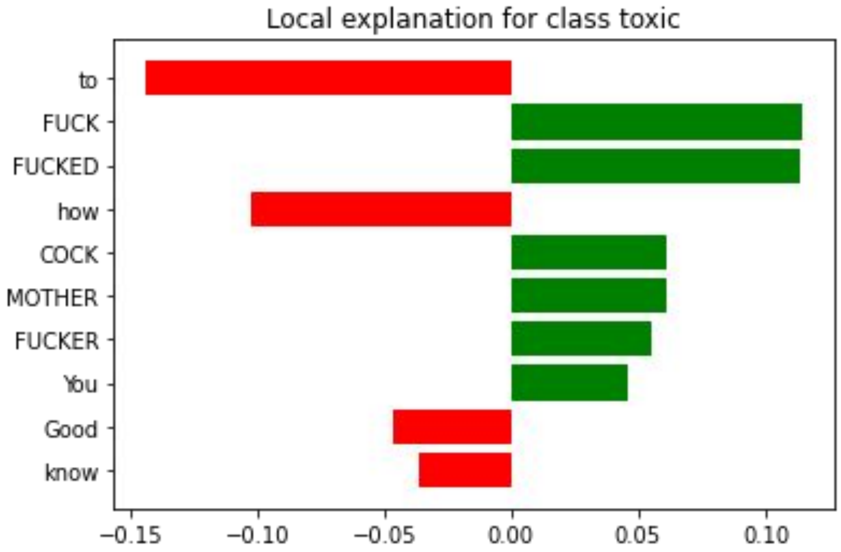


"Drappel, please stop reverting or read the god damned talk page, some medical practitioners..."" is weasel word crap, and the discussion of phentype is DIRECTLY contradicted in the following paragraph. I am not a vandal, but at this point you are."

```
[('crap', 0.12135146549087317), ('the',
-0.08618127602306407), ('you', 0.062367789540932056),
('talk', -0.053072472438244186), ('please',
-0.04913042520893451), ('discussion',
-0.0356355724391451), ('are', 0.03423936508446596),
('point', -0.0340468183082469), ('stop',
0.031431635519624315), ('at', -0.0313282075101204)]
```

LIME on balanced data

Random Forest Classifier



FUCK YOU Jdelanoy. You are German COCK SUCKER | FUCKER MOTHER FUCKER. Good to know how you FUCKED face looks.

[('crap', 0.12135146549087317), ('the', -0.08618127602306407), ('you', 0.062367789540932056), ('talk', -0.053072472438244186), ('please', -0.04913042520893451), ('discussion', -0.0356355724391451), ('are', 0.03423936508446596), ('point', -0.0340468183082469), ('stop', 0.031431635519624315), ('at', -0.0313282075101204)]