# Understanding Machine Learning models for healthcare

## Why, and how?

**Madhumita Sushil**

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

**"Machine Learning has become alchemy."**

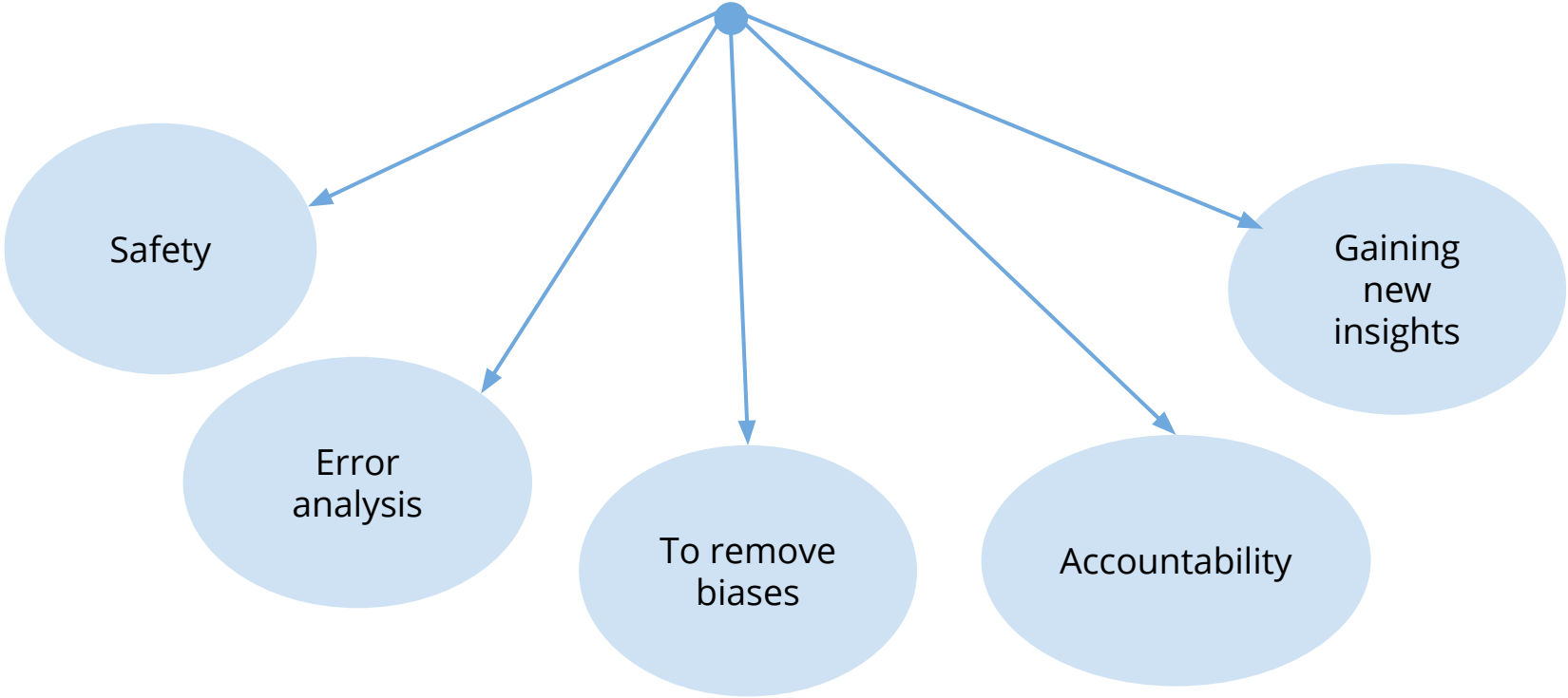
- Ali Rahimi, NeurIPS 2017

**"We're building systems that govern healthcare. I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge and not on alchemy."**

- Ali Rahimi, NeurIPS 2017

# Why?

# Why understand models?

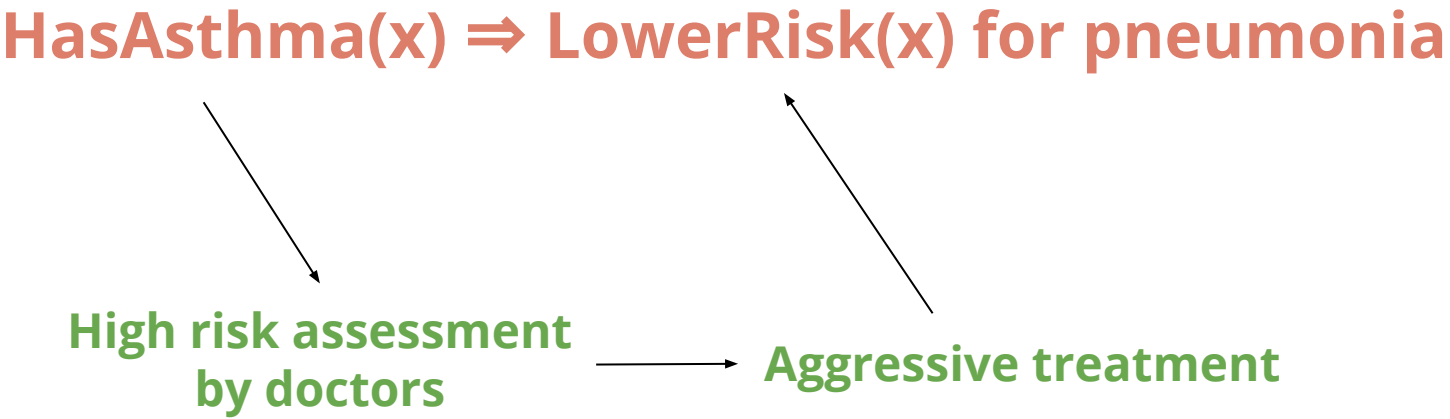


# Safety

**HasAsthma(x) ⇒ LowerRisk(x) for pneumonia**

Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

# Safety




Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

# Error analysis and improving models

IBM Watson recommended a 65 year old patient with severe bleeding a drug **that could lead to severe or fatal haemorrhage**

https://www.siliconrepublic.com/machines/ibm-watson-cancer-treatment

# Error analysis and improving models

IBM Watson recommended a 65 year old patient with severe bleeding a drug **that could lead to severe or fatal haemorrhage**

**Could this be avoided by understanding the model better?**

https://www.siliconrepublic.com/machines/ibm-watson-cancer-treatment

# Removing biases

**"AI systems are only as good as the data we put into them." - IBM**

| | |
|---|---|
| **Implicit data bias** | • Racial bias limits **model transferability**<br>• Socioeconomic treatment bias results in **inaccurate models** |
| **Missing data bias** | • Missing data about healthy patients causes **risk overestimation**<br>• Missing information about other hospital visits causes **inaccurate predictions** |
| **Small sample bias** | • **Overfitting** underrepresented subgroups of patients |

Gianfrancesco, Milena A., et al. "Potential biases in machine learning algorithms using electronic health record data." JAMA internal medicine 178.11 (2018): 1544-1547.

# Accountability

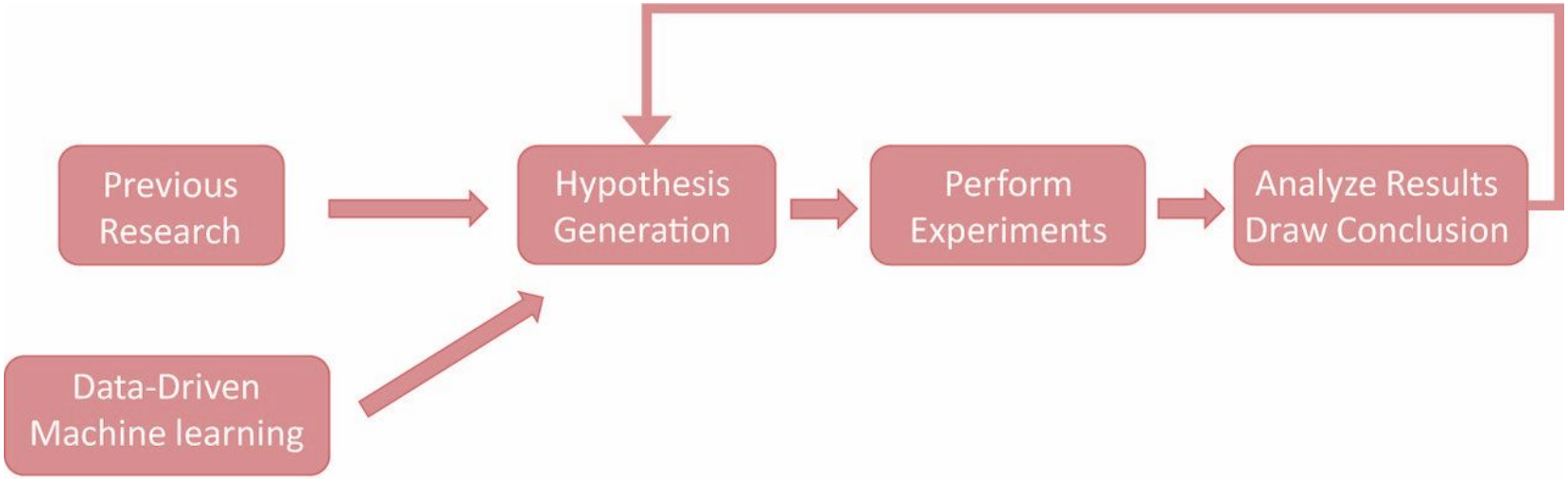
## GDPR

The data subject shall have right to obtain from the controller the confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, the following information:

- The existence of automated decision making, and at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

1. https://gdpr-info.eu/art-15-gdpr/
2. Opening Quotation Mark by Oliver Kittler from the Noun Project

# Gaining new insights




Vu, Mai-Anh T., et al. "A shared vision for machine learning in neuroscience." Journal of Neuroscience 38.7 (2018): 1601-1607.

# Defining interpretability

# Interpretability of AI systems

**Interpretability is the degree to which a human can understand the cause of a decision.**

Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017).

# Interpreting Machine Learning

WHAT

TO WHOM

Providing **explanations to humans** to facilitate them **to understand the cause of a model's decision**

WHY

# What is an explanation?

Anything that the target audience can understand

1. ML people: Maths is all we need!
2. Domain experts: It could use our jargon and domain knowledge
3. Layman: Plain old English?

# What do we explain?

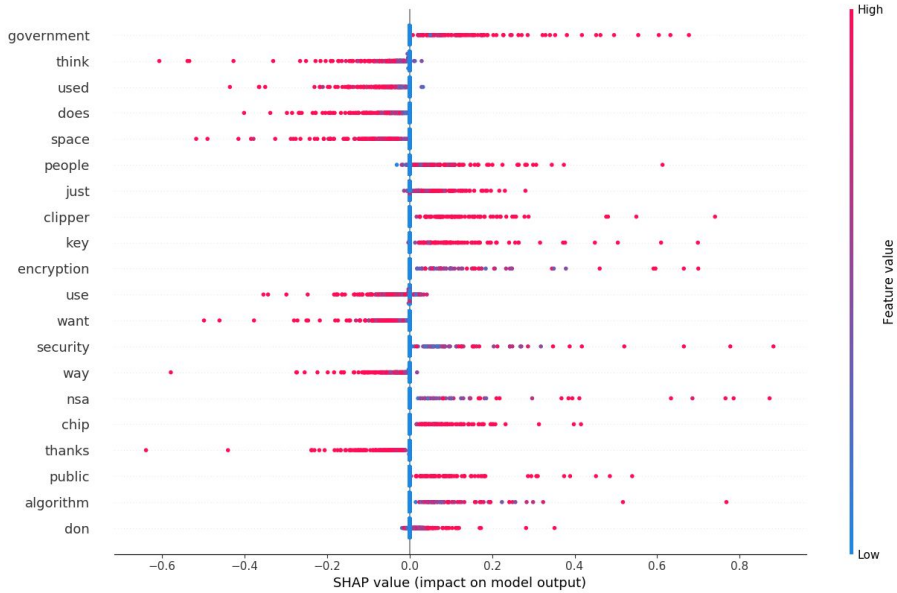
**Single output decisions**

**Complete trained models**

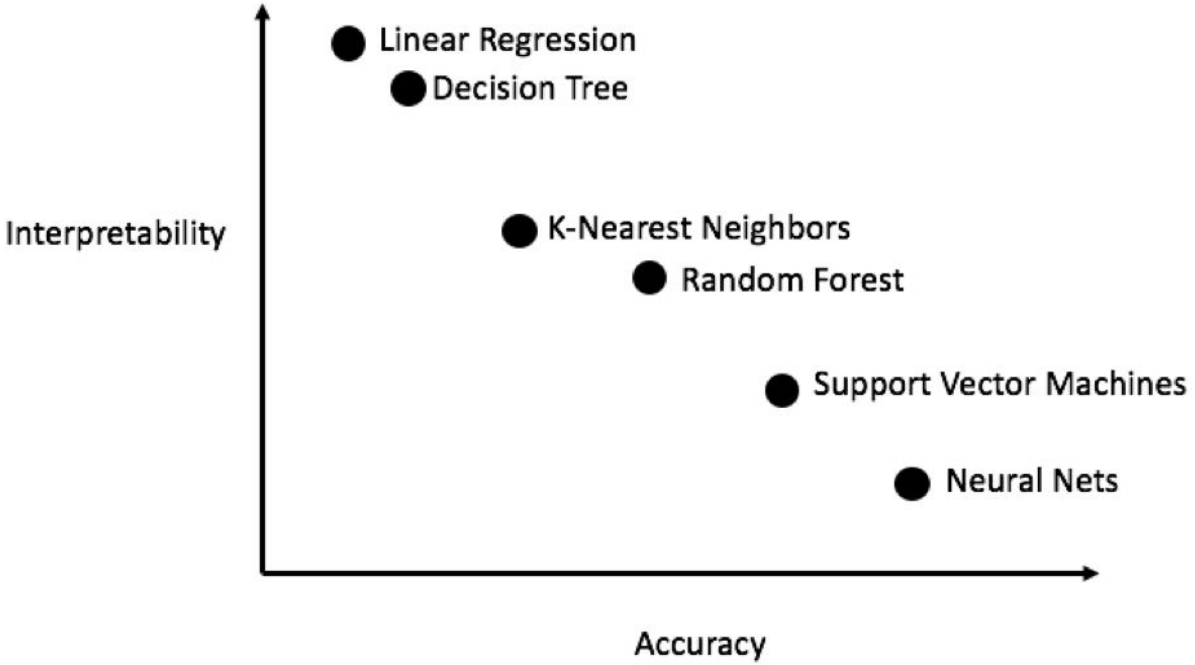
# Established practices

low dose prednisone was started at 10 mg . the patient was given morphine at 2:00 a.m. for tachypnea . lisinopril 40 mg b.i.d. 5 . discharge condition : stable , satting 97 -99 % on ra no murmur , mean blood pressures 52 -69 . endocrine : the patient has been noted to have hyperglycemia on this admission . lp was also performed to evaluate for meningitis . hematocrit was 24.1 . we had shut off the heparin before the operation . he was not bleeding further . # leukocytosis : afebrile , ? position of tubes and lines and the left perihilar region opacities are unchanged . ugib also would lead to altered mental status given hypoperfusion of brain . she also wanted to delay her surgery . the right dorsalis pedis pulse is dopplerable . status post hypothermia . sinus versus atrial tachycardia .

0.01

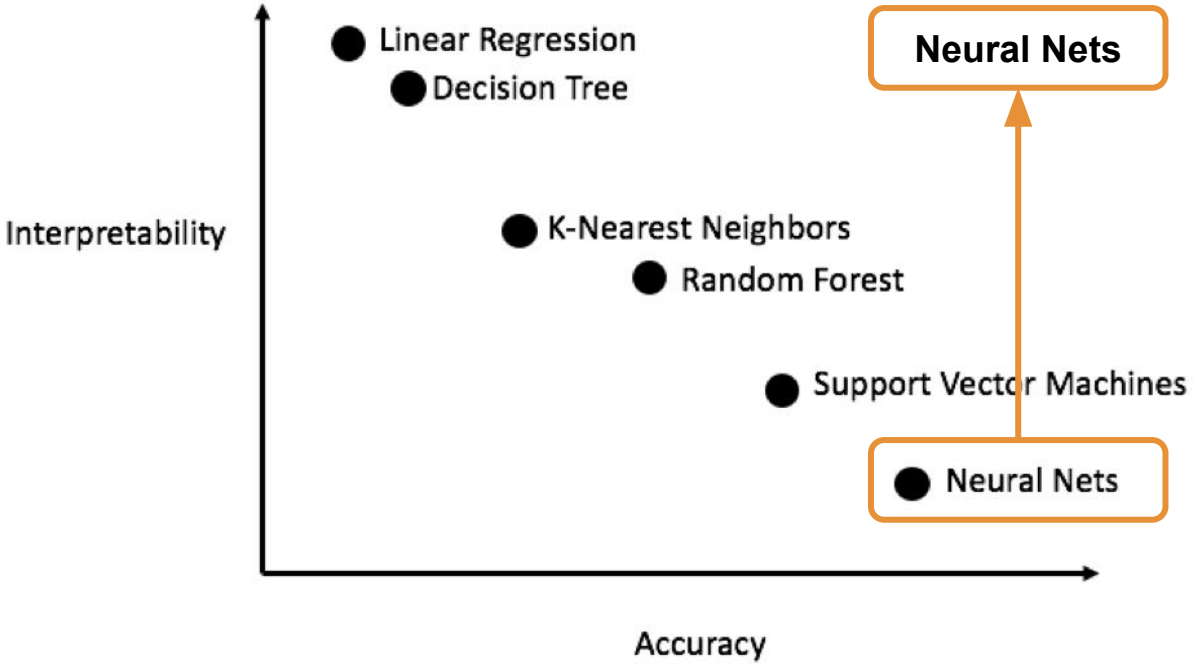


# Interpretability vs. Accuracy




https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9

# Interpretability vs. Accuracy




https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9

# Opening the black box

# If-then-else rules as explanations

Hierarchical list of if-then-else rules:

if ***<condition1>*** and ***<condition2>*** and ... ⇒ **class1**

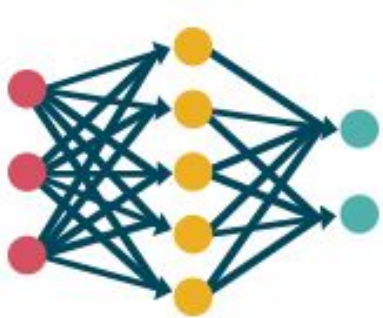
elif ***<condition3>*** ... ⇒ **class1**

else **class2**

Quantifies associations between features and classes

# If-then-else rules to interpret neural nets

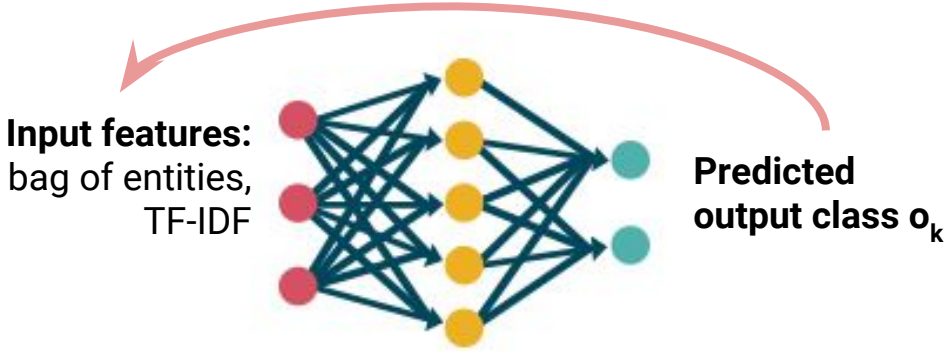
**Input features:**
bag of entities,
TF-IDF



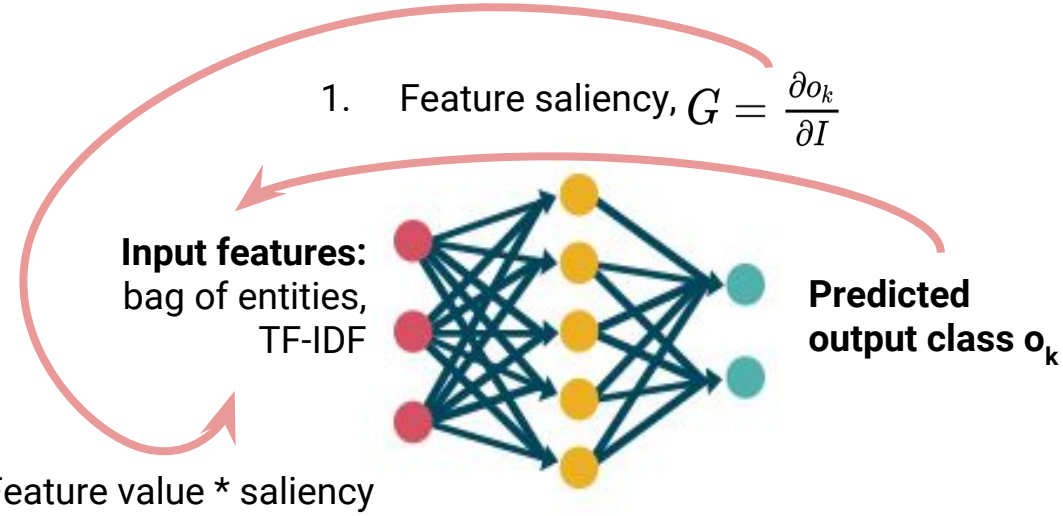
**Predicted output class $o_k$**

# If-then-else rules to interpret neural nets

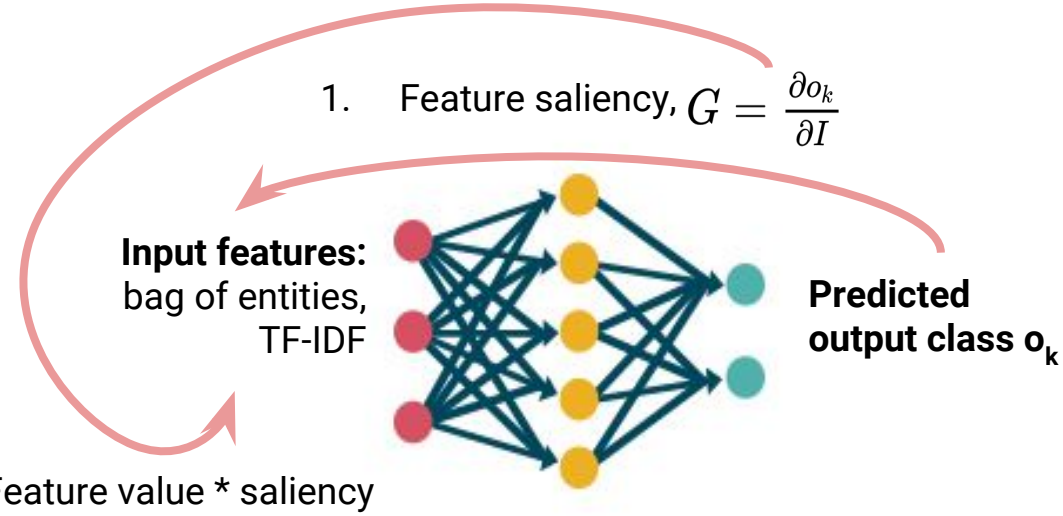
1. Feature saliency, $G = \frac{\partial o_k}{\partial I}$



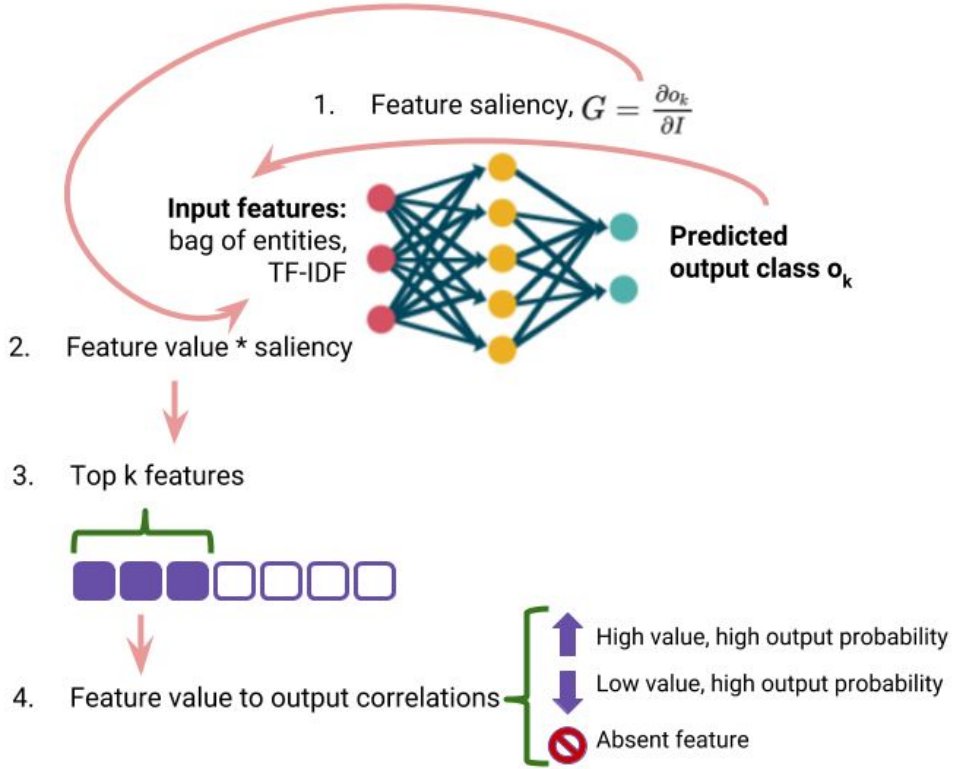
# If-then-else rules to interpret neural nets



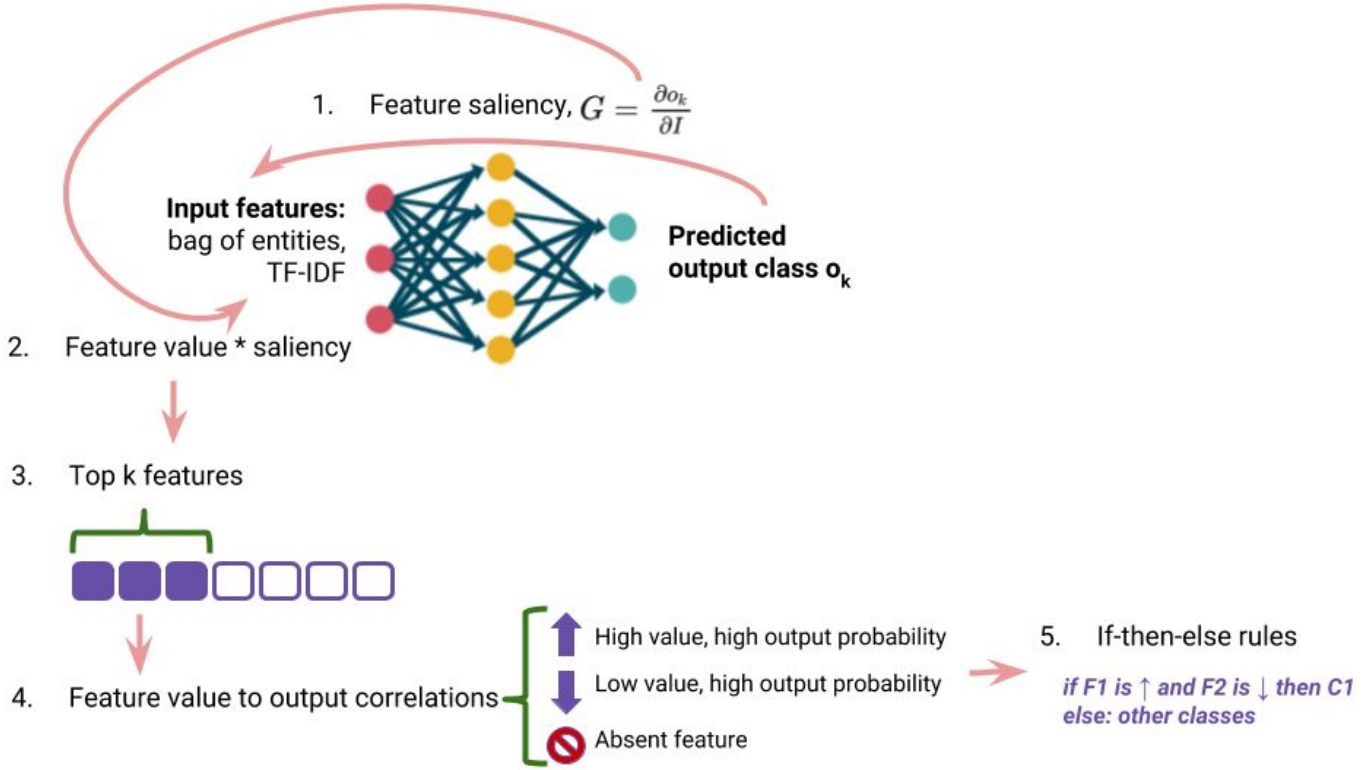
# If-then-else rules to interpret neural nets



# If-then-else rules to interpret neural nets



# If-then-else rules to interpret neural nets




Madhumita Sushil, Simon Šuster, Walter Daelemans. Rule induction for global explanation of trained models. Workshop on Analyzing and interpreting neural networks for NLP (BlackboxNLP), EMNLP 2018

# Resulting explanations

↑ **Take blood pressure (treatment)** and

⊘ **Nothing by mouth** and

⊘ **Flagyl**

→**Diseases of the circulatory system** (✔84/90)

# Resulting explanations

↑ **Pneumonia** and

↑ **Lung opacity** and

↓ **Non-specific ST-T changes by ECG** and

⊘ **CT of pelvis w/o contrast**

→**Diseases of the respiratory system** (✔7/7)

# Resulting explanations

↑ **Physical examination** and

↑ **Pregnancy with medical condition**

→**Dies within hospital** (✔221/222)

# Open questions

When is an explanation better than another?

- If it matches domain knowledge
- If it is plausible according to existing knowledge
- If it provides new counter-intuitive insights

# Open questions

Which explanation form is ideal?

Question by Elves Sousa from the Noun Project