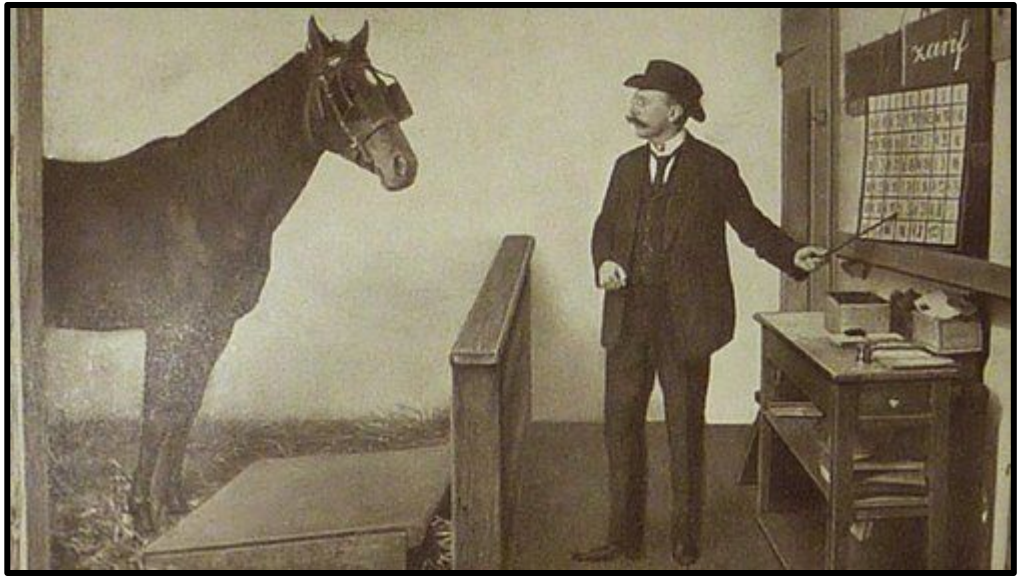
"Why Should I Trust You?": Explaining the Predictions of Any Classifier (*Explainable AI using LIME*)

Author: Marco Tulio Ribeiro

Story of Clever Hans

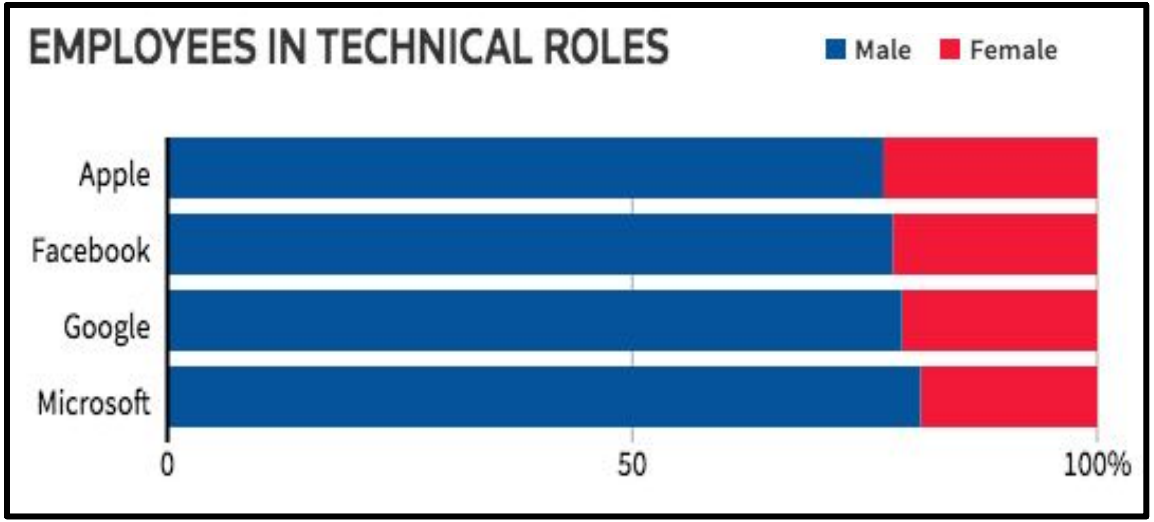
The horse that can do **Arithmetics!**



Amazon's sexist AI resume parser

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" **until** the company discovered the problem.

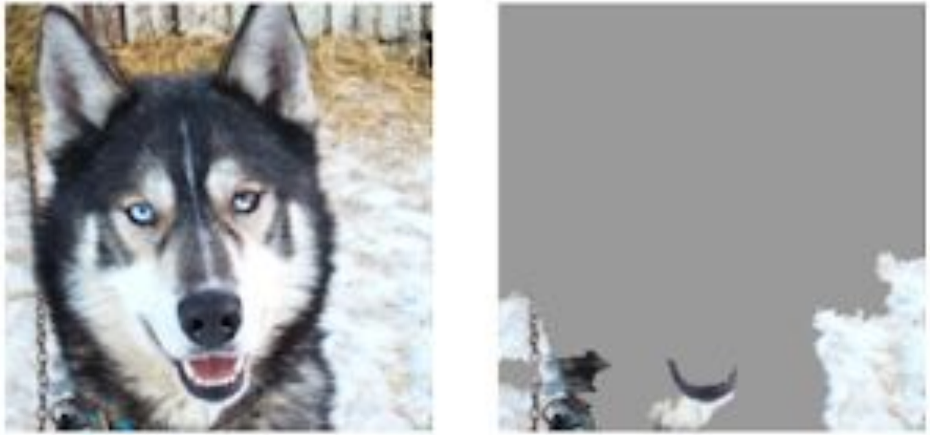
- *Reuters (Oct, 18)*



Husky vs Wolf example

The algorithm was using the background of the picture and totally ignoring animal characteristics.

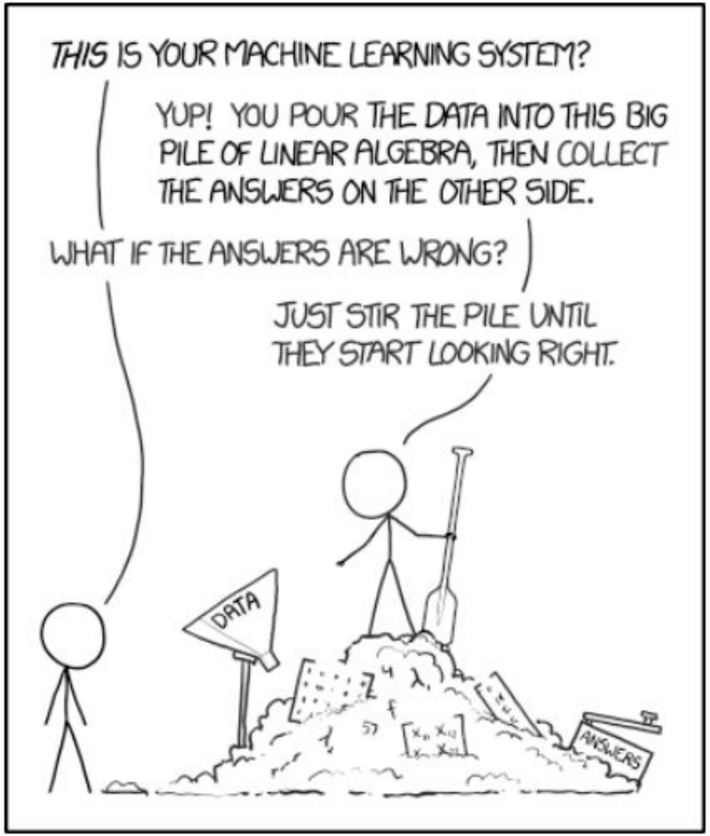
According to model, snow = wolf



(a) Husky classified as wolf

(b) Explanation

Current industry scenario

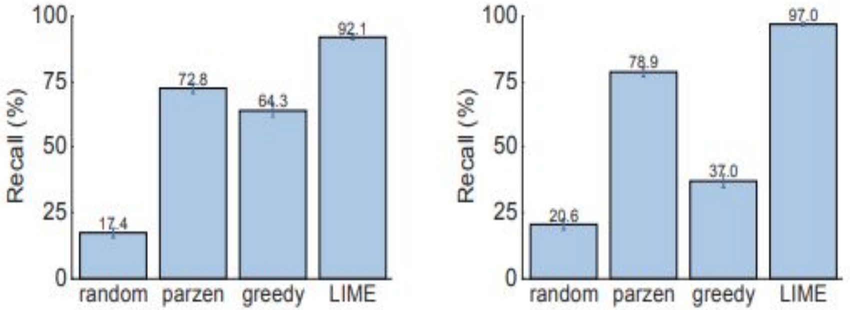


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Author's work: LIME

LIME: Local interpretable model-agnostic explanations

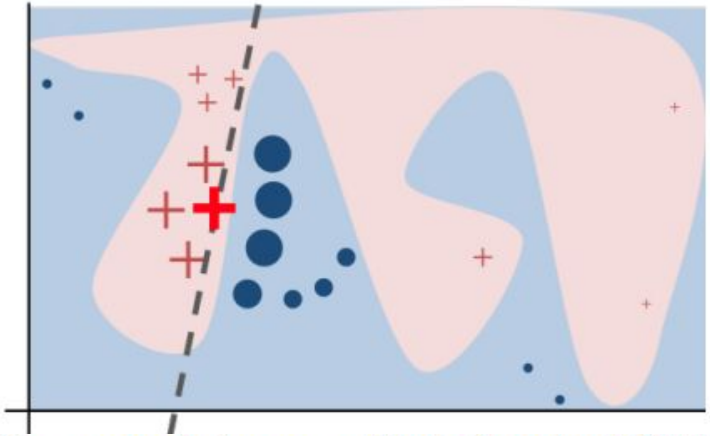
$$\xi(x) = \operatorname*{argmin}_{g\in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



(a) Sparse LR

(b) Decision Tree

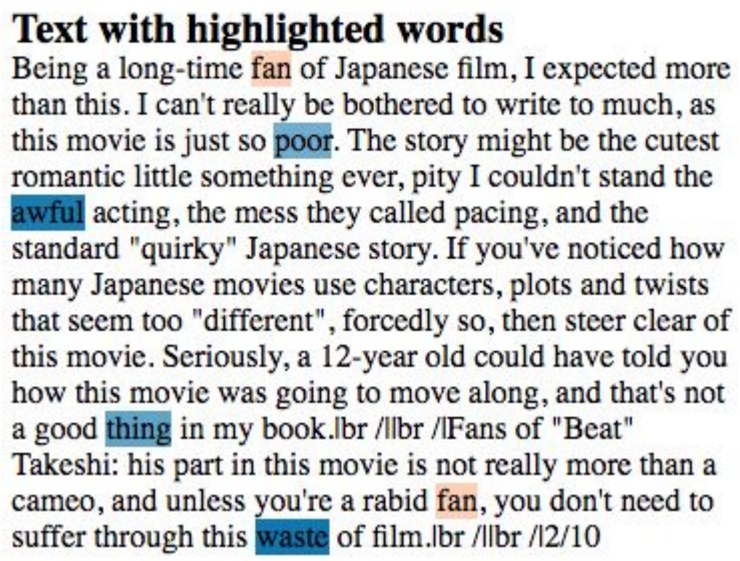
Recall on truly important features



Toy example to present intuition for LIME.

My implementation on Sentiment Analysis

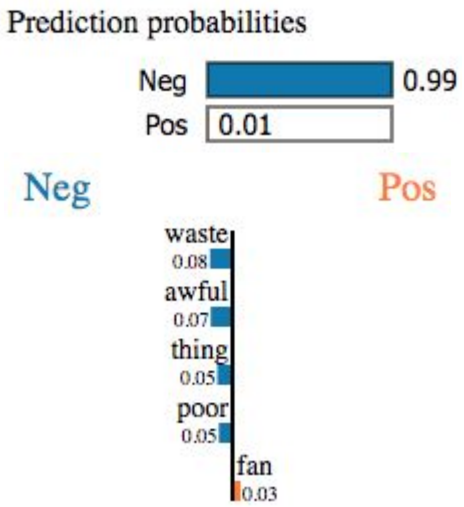
Link-https://nbviewer.jupyter.org/github/rohancode/xai-tutorials/blob/master/Lime/Lime%20Imdb%20SVM%20%28LSA%20features%29.ipynb



Text with highlighted words

Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write to much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcedly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book.lbr /llbr /lFans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.lbr /llbr /l2/10

```
SVM model prediction : Neg sentiment
True value : Neg sentiment
```



Explainable AI (XAI)benefits

- Builds **trust** in the model
- Involvement of subject matter experts in model building (ex: Doctor in disease prediction)
- Gain complex hidden insights in the data

A necessity of future: General Data Protection Regulation (EU - 2016/679)

“Regulation (EU) 2016/679 Regulation on the **protection** of natural persons with regard to the **processing of personal data** and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)”

THANK-YOU, ANY QUESTIONS?...