

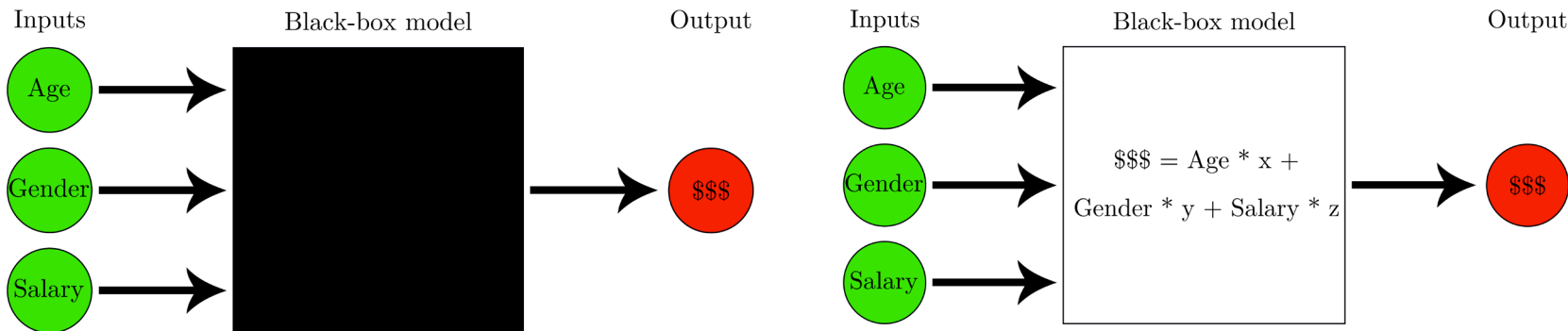
“Honey, Tell Me What's Wrong”, Global Explainability of NLP Models through Cooperative Generation

Antoine Chaffin
Julien Delaunay

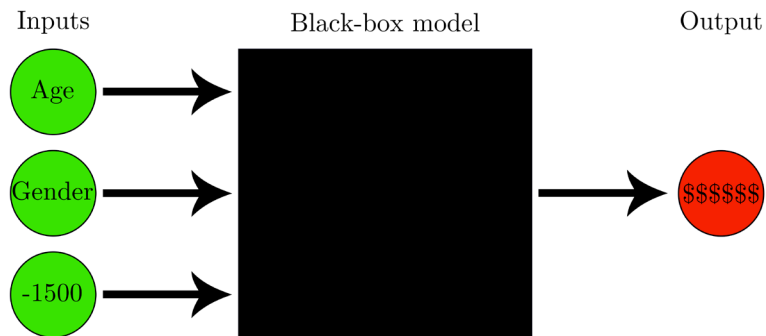
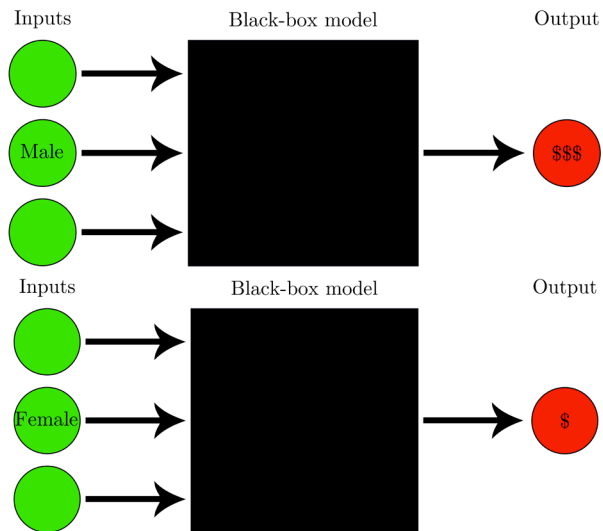


Introduction

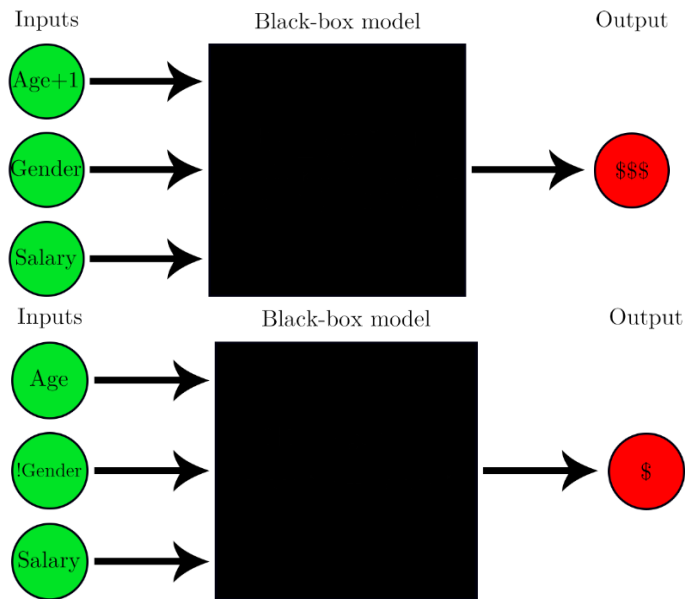
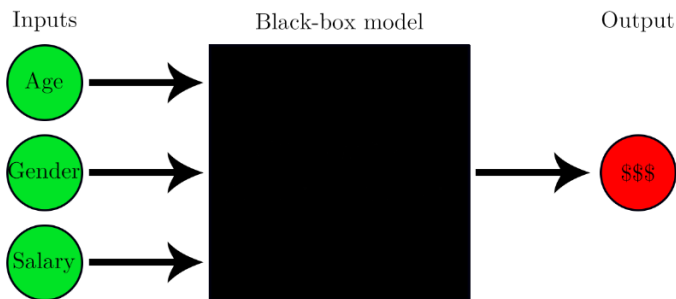
- (Deep) neural networks learn a **complex mapping** between inputs and outputs
 - Larger models have more capacity**, can approximate even more complex functions
- Explainability try to give **insights about the decision**



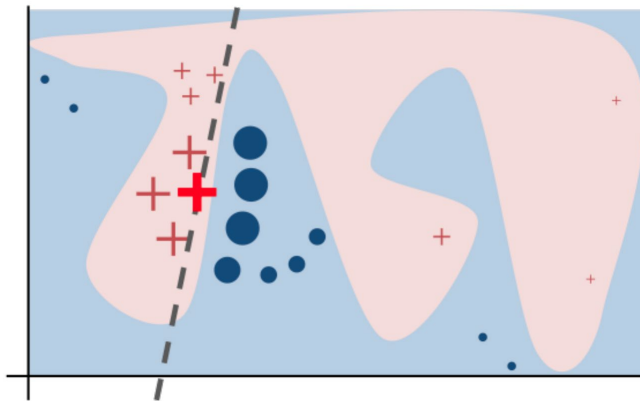
- Being able to justify the decision of a deployed model (**legal**)
- Identify biases (**fairness**)
- Identify edge cases and understand failures of the model (**performance**)



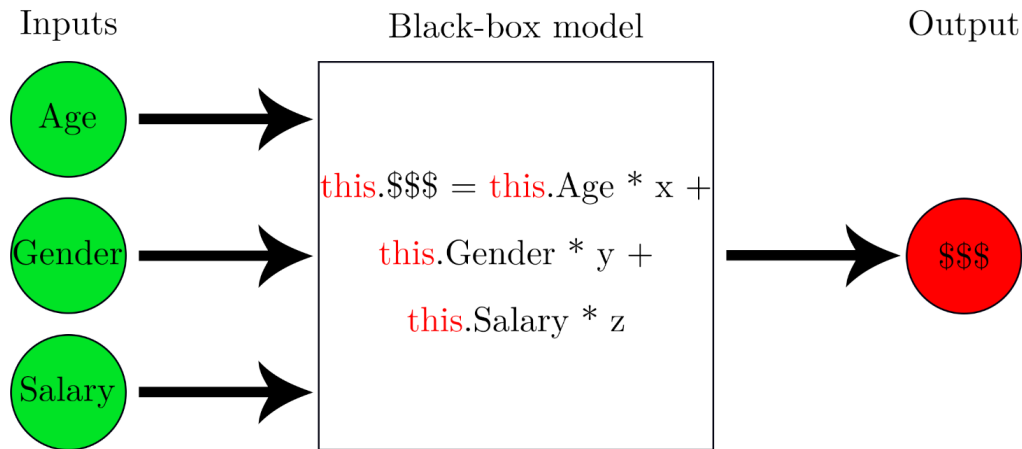
- Do not rely on the inner working of the model
 - Model agnostic to work on every model**
- Explain the **decision for a given input**: local explanation



1. Generate **neighbors** of the instance to explain
2. Use the complex model to get decision for these samples
3. Learn a **linear model** with these decisions
4. The linear model is used as an **local approximation of the complex model**

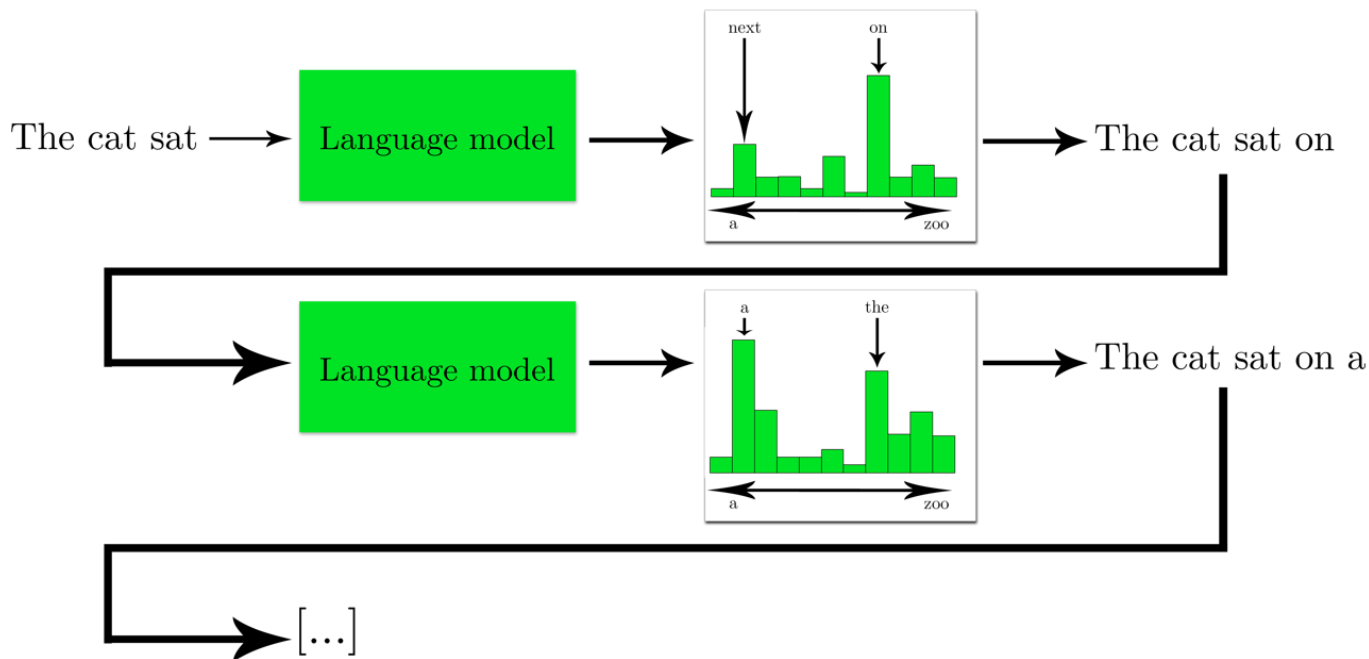


1. Requires data (**confidentiality/privacy**)
2. Selecting **representative data** is hard
3. Explain the decision for **this input and this input only**



Therapy

- Probability of the **next word given past ones**
- **Iteratively add tokens** to produce text



- Few options to control the generation besides the **prompt**
- Add some **constraints** on the generated text (writing style, emotion/polarity, detoxification, etc.)

Text
generation

I feel → Language model → I feel normal

Emotion: fear 🤪

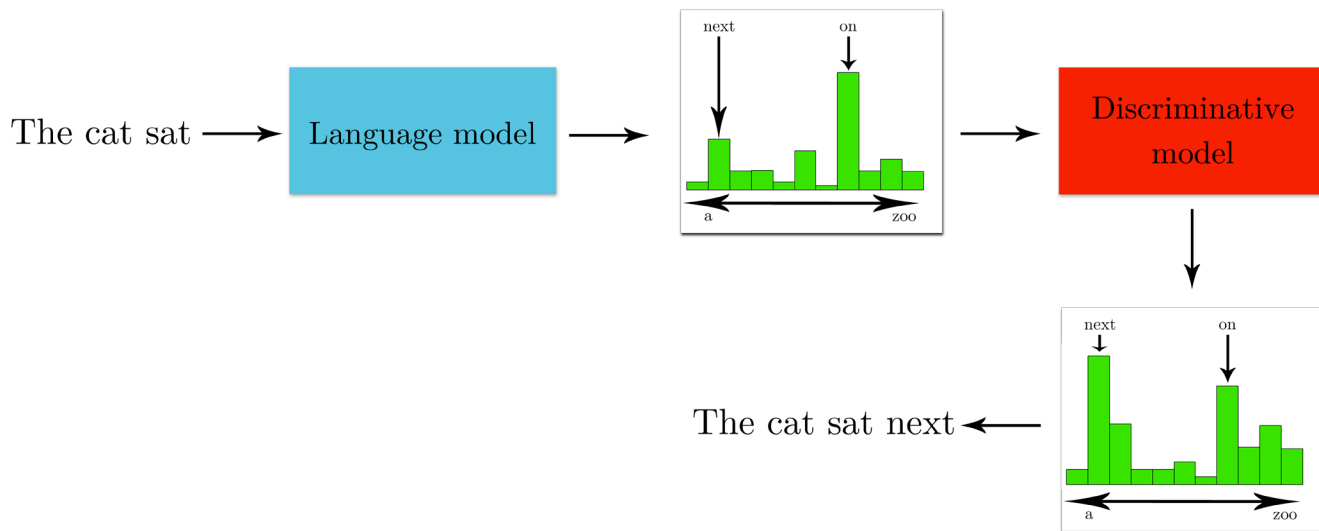


Constrained text
generation

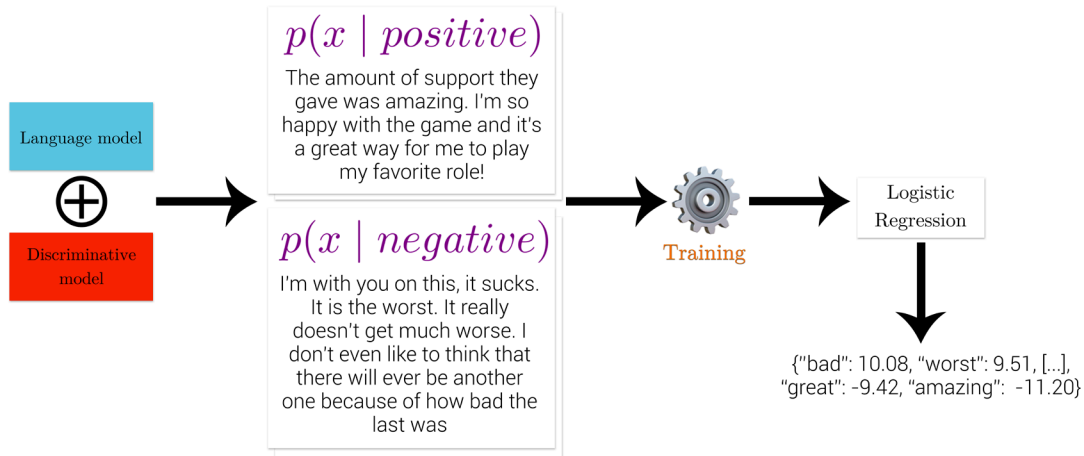
I feel → Language model → I feel terrified

- Guide the generation using the **score of an external model**
- Generate text following the **conditional distribution** (product of the **language model likelihood** and the **score of the discriminative model**)

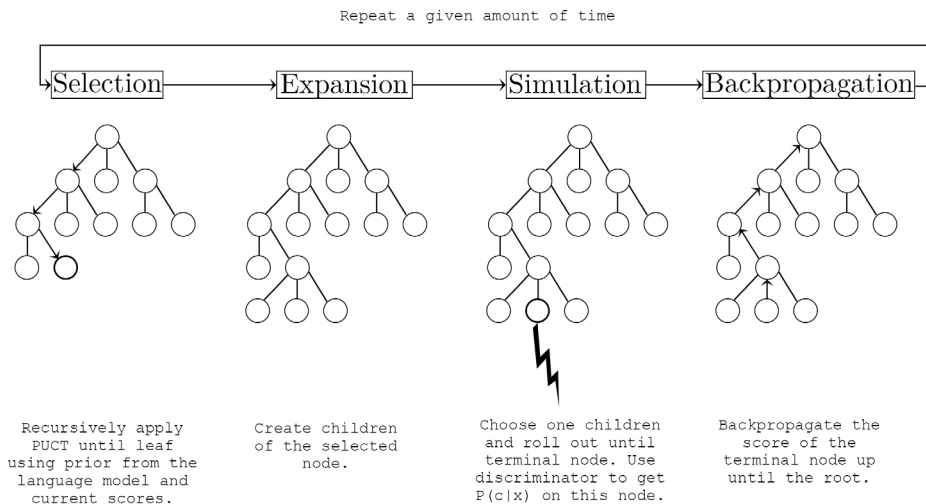
$$p(x \mid c) \propto p(x) * p(c \mid x)$$



- Use the **distribution of cooperatively generated texts** to explain the model: **words with high frequencies are likely to be important**
- Learn a logistic regression to predict class of generated texts using **tf-idf**
 - **Weights associated to each word** can be returned as explanation
 - Words that are frequent because of the language model or across different class will be **filtered out**



- Monte Carlo Tree Search (MCTS) properties:
 - Long-term vision:** scores the next token using finished sequences (rollout)
 - Efficient:** exploration of sub-optimal paths has an upper bound
 - Modular:** outputs a solution according to the computational budget
 - Plug and play:** can be used with any LM and discriminator without any tuning



Experiments

- Two tasks: **polarity** (amazon_polarity) & **topic** (ag_news) classification

amazon_polarity

[POSITIVE] Stuning even for the non-gamer. This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^ _ ^

[NEGATIVE] "A complete waste of time. Typographical errors, poor grammar, and a totally pathetic plot add up to absolutely nothing. I'm embarrassed for this author and very disappointed I actually paid for this book."

ag_news

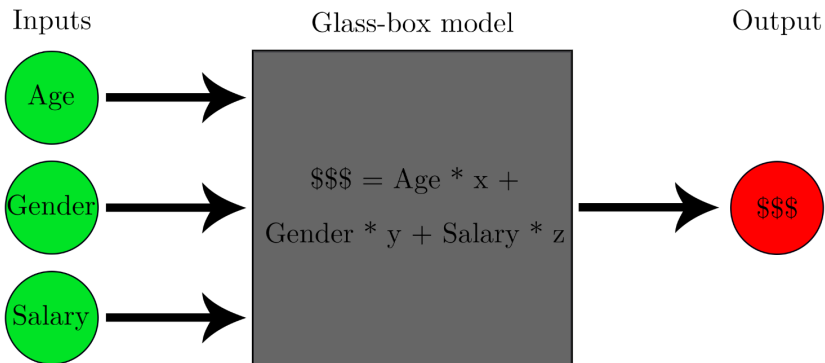
[World] Talks End With No U.S. Climate Deal A U.N. conference ended early Saturday with a vague plan for informal new talks on how to slow global warming but without a U.S. commitment to multilateral negotiations on next steps, including emissions controls.

[Sports] Wenger Ready To Prove Doubters Wrong Arsene Wenger has hit back at critics who claim that Arsenal cannot perform against Europe; big guns after being drawn against Bayern Munich in the Champions League.

[Business] Carlyle Looks Toward Commercial Aerospace (Reuters) Reuters - Private investment firm Carlyle Group, which has a reputation for making well-timed and occasionally\controversial plays in the defense industry, has quietly placed its bets on another part of the market.

[Sci/Tech] Hacker Cracks Apple's Streaming Technology (AP) AP - The Norwegian hacker famed for developing DVD encryption-cracking software has apparently struck again; this time breaking the locks on Apple Computer Inc.'s wireless music streaming technology.

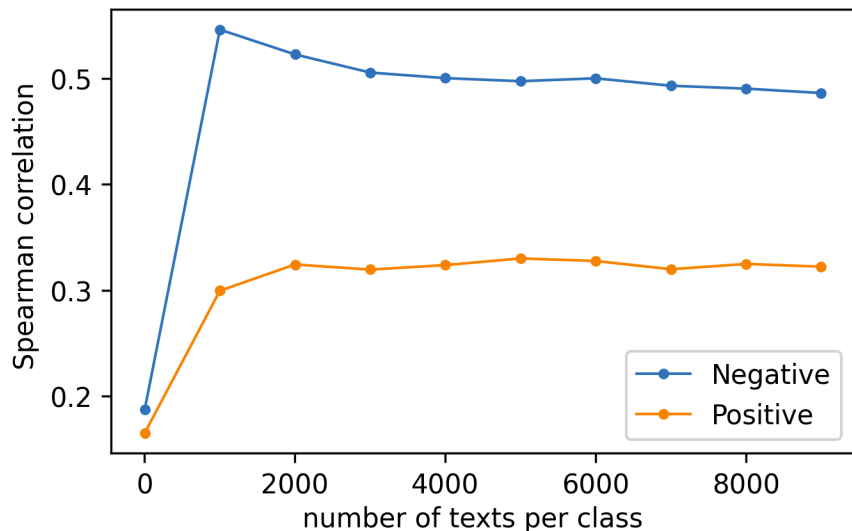
- **No « ground truth explanations »** available
- Use of a **glass-box**: a model explainable by design but used as a black-box
- List of features that contains important features and link them to similar (relative) weights
 - **Spearman correlation** of the explanation and glass-box weights



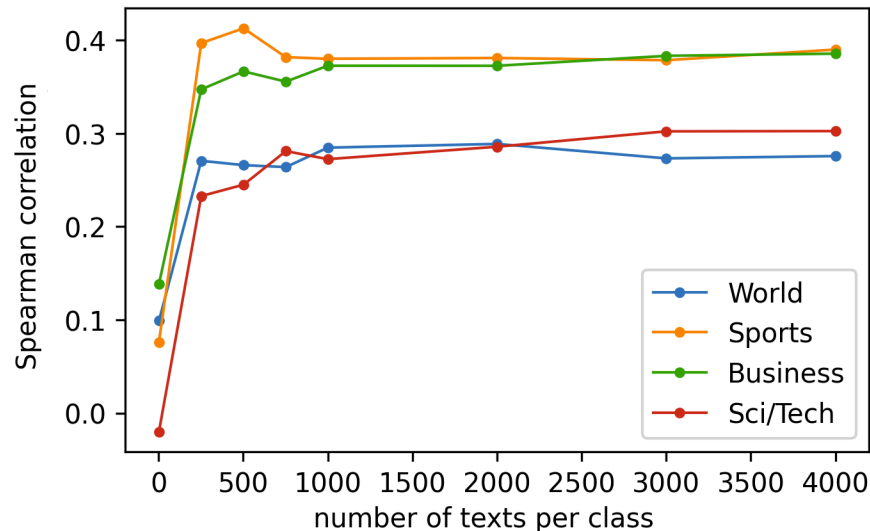
	Glass-box	LIME	SHAP	Therapy
X	5.1	1.7	10	4.9
y	8.2	2.3	15.1	7.6
Z	-1.2	-0.1	-5.4	-1.5

- Correlation **quickly grows** with the number of generated texts **until plateauing**
 - Only a **small number of samples is needed**

Spearman correlation w.r.t number of texts per class on amazon_polarity



Spearman correlation w.r.t number of texts per class on ag_news



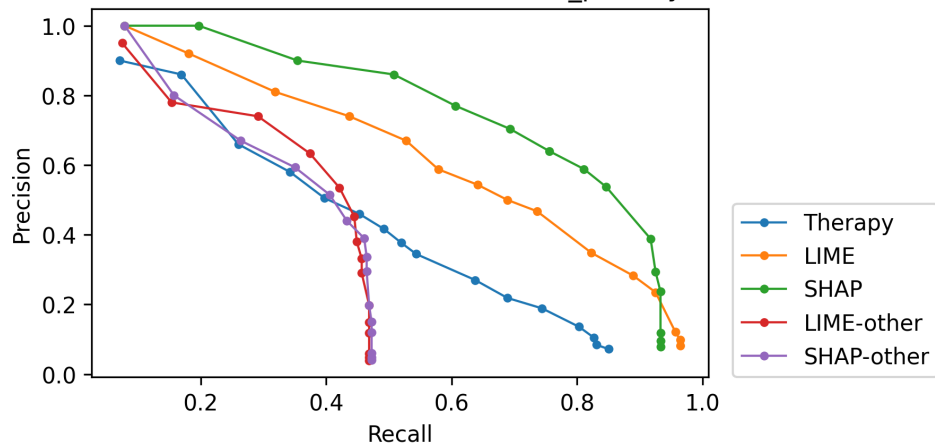
- SHAP is better than LIME and Therapy on both datasets
- Therapy is better than LIME on ag_news but worse on amazon_polarity
- When using data from the other dataset: **SHAP & LIME collapse, way below Therapy**

Dataset	AMAZON_POLARITY		AG_NEWS			
Class	Positive	Negative	World	Sports	Business	Sci/Tech
LIME	0.64 (5.0e-7)	0.44 (1.5e-3)	0.09 (0.53)	0.16 (0.27)	0.20 (0.16)	0.19 (0.19)
LIME-other	0.21 (0.14)	0.18 (0.21)	-0.03 (0.85)	0.23 (0.12)	0.09 (0.52)	0.29 (0.04)
SHAP	0.71 (7.6e-9)	0.76 (1.6e-10)	0.47 (6.2e-4)	0.62 (1.7e-06)	0.53 (8.0e-5)	0.61 (2.4e-6)
SHAP-other	0.02 (0.87)	0.26 (0.06)	-0.05 (0.71)	0.04 (0.77)	0.15 (0.31)	0.12 (0.41)
Therapy	0.49 (3.3e-08)	0.31 (1.0e-4)	0.27 (1.6e-07)	0.37 (4.0e-12)	0.38 (5.6e-13)	0.3 (8.9e-09)

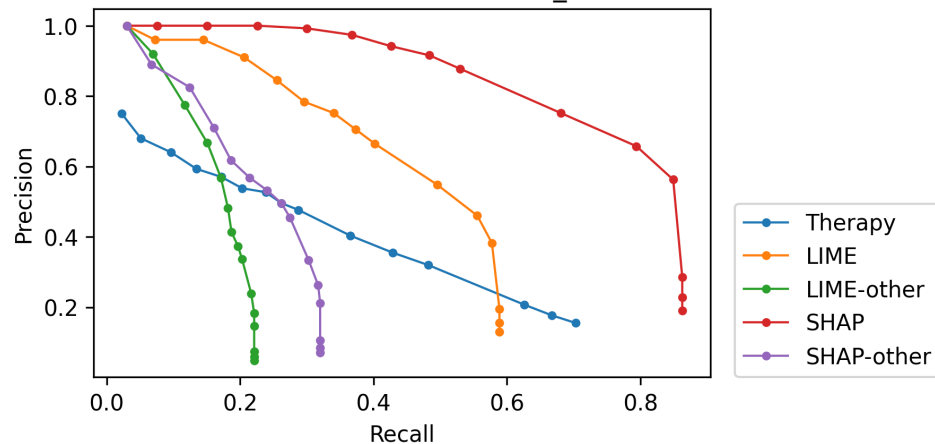
Table 1: Spearman correlation (p-value) between the top words of a logistic regression glass box and the four explanation methods. Results are shown per class and dataset. 'other' indicate that the explanations are generated using the other dataset.

- Besides correct scores: **returned features are important and most important features are found**
- Precision/recall curves** using top-words of the glass-box as targets
- Again, Therapy is below LIME and SHAP (although competitive), but those **collapse because limited to terms present in the data**

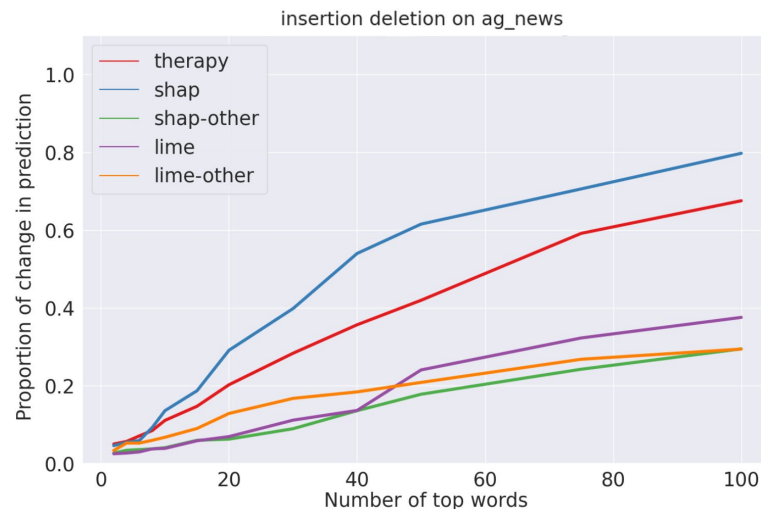
Precision/recall curves on amazon_polarity



Precision/recall curves on AG_news



- Assert **whether the features returned affect the model predictions**
 - Removing the « cause »** should force the black-box to change its decision^[1]
 - Adding words from the other classes** should also lower its confidence
- Percentage of **classification changes** w.r.t the number of swapped words



- Therapy is a **model-agnostic global explanation method that works without input data**
- Rather than using input data, it leverage **cooperative generation** to generate **texts following the distribution learned the studied model**
 - **Search is driven by a pre-trained LM**
- It achieves **competitive results against usual methods** when those have access to very specific data that contains target features
- In the **more realistic case** where **no data or not very specific data** is available, performances drop **substantially below Therapy**
- **Code based on Hugging Face transformer library available on Github**
 - Experiments with other type of model, e.g **CLIP (cross-modal regression)**

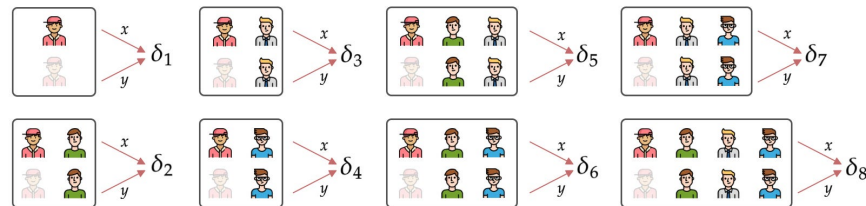
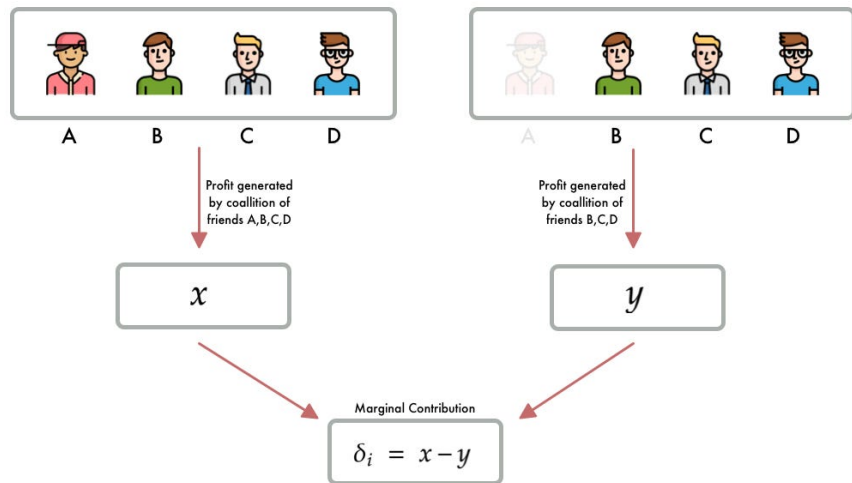
Thank you for your attention !
Any question ?




antoine.chaffin@irisa.fr  @antoine_chaffin

Institut de Recherche en Informatique et Systèmes Aléatoires

1. Compute permutations of the input
2. Compute marginal contribution of each element for every permutation



The Shapley value for member 

is given by:

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$