

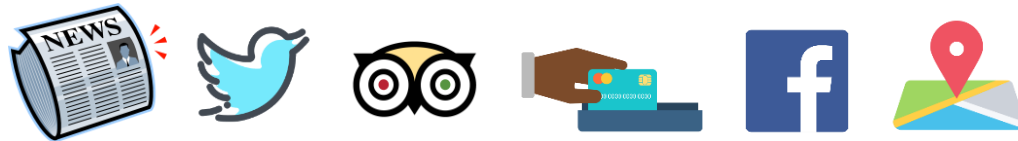
# **Comparative study of instance-level explanation methods for textual and behavioral data**

Yanou Ramon, David Martens  
Applied Data Mining research group

June 25, 2019, 1.30-2PM  
EURO Conference, Dublin

# 1. INTRODUCTION

- Applications using **high-dimensional, sparse** data are ample



## **Behavioral data**

payment data, visited websites or physical locations, FB likes etc.


## **Textual data**

emails, news articles, Twitter posts etc.

# 1. INTRODUCTION

High-dimensional + sparse → Gender prediction using movie viewing data

ACTIVE FEATURE = "EVIDENCE"



	Star wars	Pearl Harbor	Django	...	Home Alone	Target $\hat{y}$ Gender
User 1	1	0	0		1	<i>M</i>
User 2	1	1	0		1	<i>F</i>
...						
User n	1	1	1		0	<i>M</i>

6,040 users

# 1. INTRODUCTION

- High predictive **performance**  $\Leftrightarrow$  **complex** models
- **Interpretability issues**: how are predictions made?

# 1. INTRODUCTION

- High predictive **performance**  $\Leftrightarrow$  **complex** models
- **Interpretability issues**: how are predictions made?



**Relevance?**

- Ethical objectives eg, privacy, fairness, safety
- Model improvements eg, debugging
- Trust/acceptance
- ...

# 1. INTRODUCTION

- High predictive **performance**  $\Leftrightarrow$  **complex** models
- **Interpretability issues**: how are predictions made?



- Ethical objectives eg, privacy, fairness, safety
- Model improvements eg, debugging
- Trust/acceptance
- ...



## INSTANCE-LEVEL EXPLANATIONS

# 1. INTRODUCTION

“Which **model-agnostic, instance-level explanation algorithm** is **most suitable** for explaining model predictions of classifiers built from **textual/behavioral** data?”



- **Overview** and **selection** of instance-level explanation methods (literature review)
- Selection of **quantitative criteria**
- **Comparison** using **behavioral/textual** data

## 2. EXPLANATION METHODS

### Selection criteria

- **Model-agnostic** methods treat the model as a black-box
- **Computational ability** to cope with **high-dimensional** data



## 2. EXPLANATION METHODS

### Selection criteria

- **Model-agnostic** methods treat the model as a black-box
- **Computational ability** to cope with **high-dimensional** data

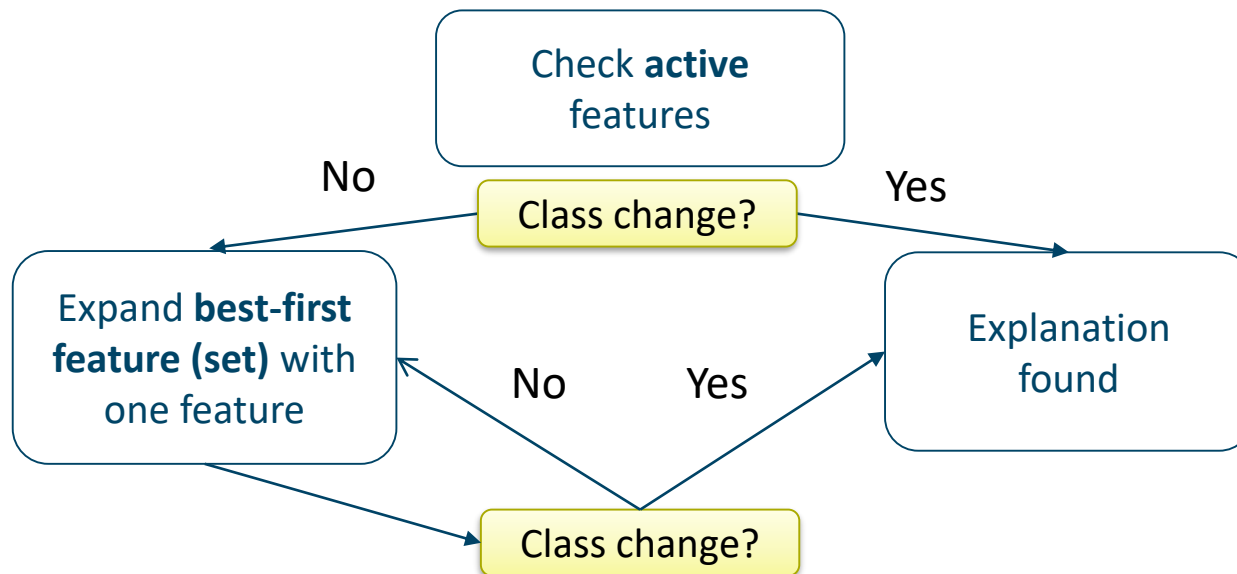


- Evidence Counterfactual (EDC) (Martens & Provost, 2013)
- Linear Interpretable Model-Agnostic Explainer (LIME)  
(Ribeiro et al., 2016)
- Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017)

## 2. EXPLANATION METHODS

### Evidence Counterfactual

- **Minimal set of features** so that “**removing**” them results in a predicted class change
- **Removing** → set feature value to zero
- **Model-agnostic** algorithm (*SEDC*) based on **heuristic best-first**



## 2. EXPLANATION METHODS

### Evidence Counterfactual – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

Sam watched 120 movies

Sam is predicted as male

## 2. EXPLANATION METHODS

### Evidence Counterfactual – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

**WHY?**

Sam watched 120 movies

Sam is predicted as male

## 2. EXPLANATION METHODS

### Evidence Counterfactual – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

**IF** Sam would not have watched *{Taxi driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}*, **THEN** his predicted class would change from male to female

## 2. EXPLANATION METHODS

### Evidence Counterfactual – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

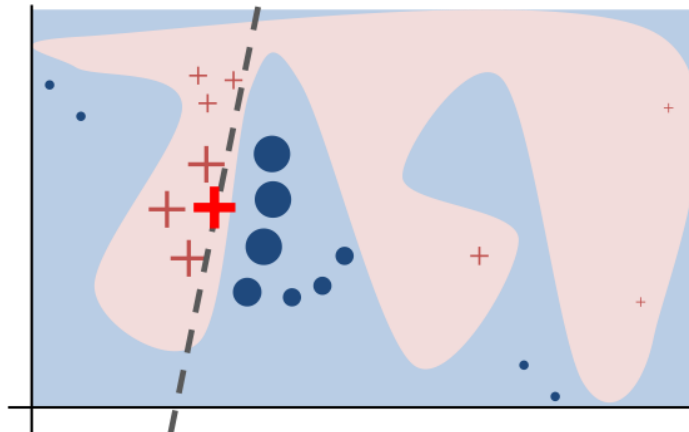
**IF** Sam would not have watched *{Taxi driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}*, **THEN** his predicted class would change from male to female

POSITIVE EVIDENCE = EVIDENCE FOR A PREDICTED CLASS

## 2. EXPLANATION METHODS

### LIME / SHAP

- Explanation model: **sparse, linear model**
- Explanation model **approximates** original model in the **neighborhood of the instance**
- **Perturbed instances**

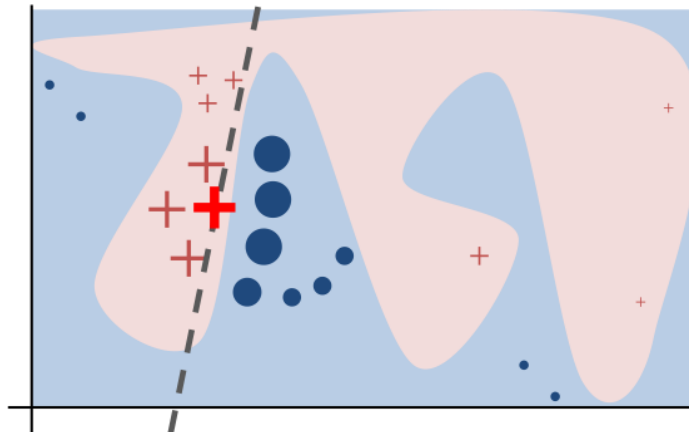


Source: Ribeiro et al., 2016

## 2. EXPLANATION METHODS

### LIME / SHAP – differences

- How they generate perturbed samples
- Distance function
- Complexity control



Source: Ribeiro et al., 2016



## 2. EXPLANATION METHODS

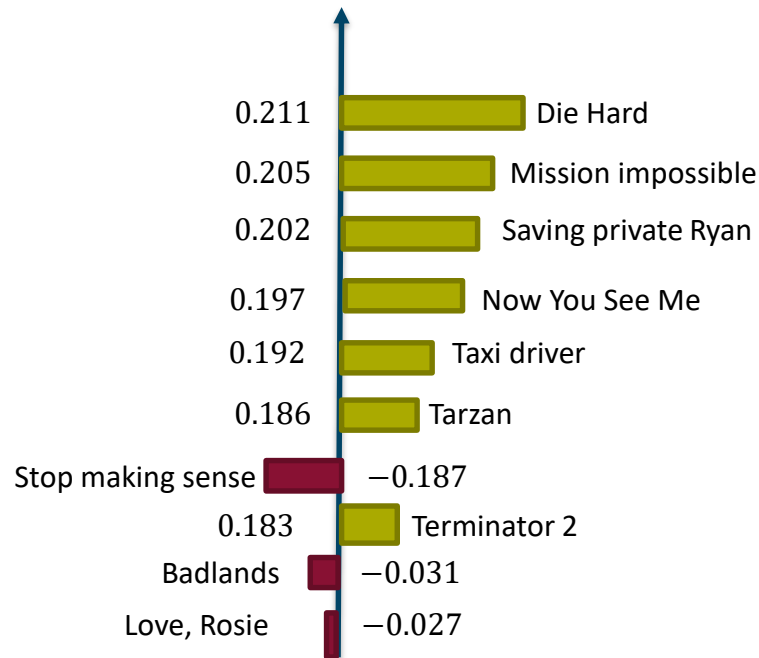
### LIME – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

**$k = 10$  features**  
(feature selection)



## 2. EXPLANATION METHODS

### LIME – example

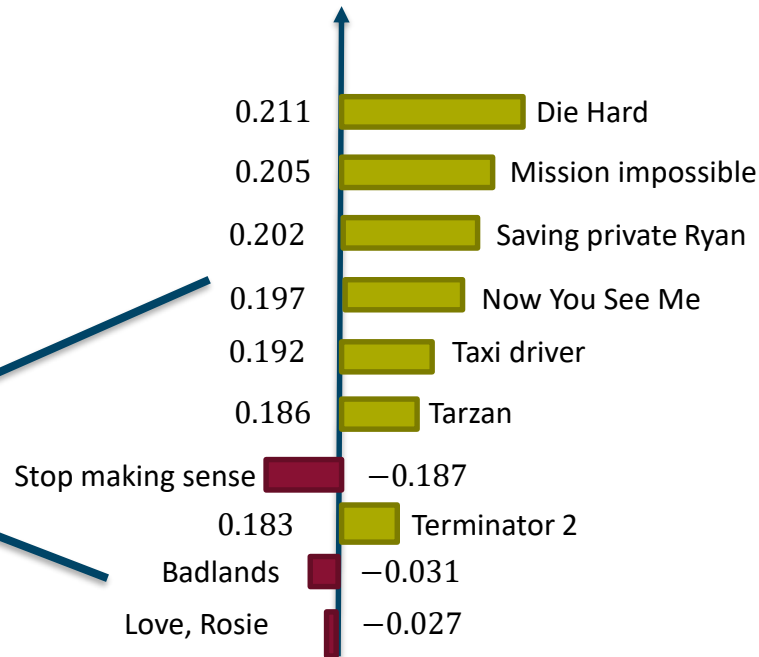
**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam

**$k = 10$  features**  
(feature selection)

**BOTH POSITIVE & NEGATIVE EVIDENCE**



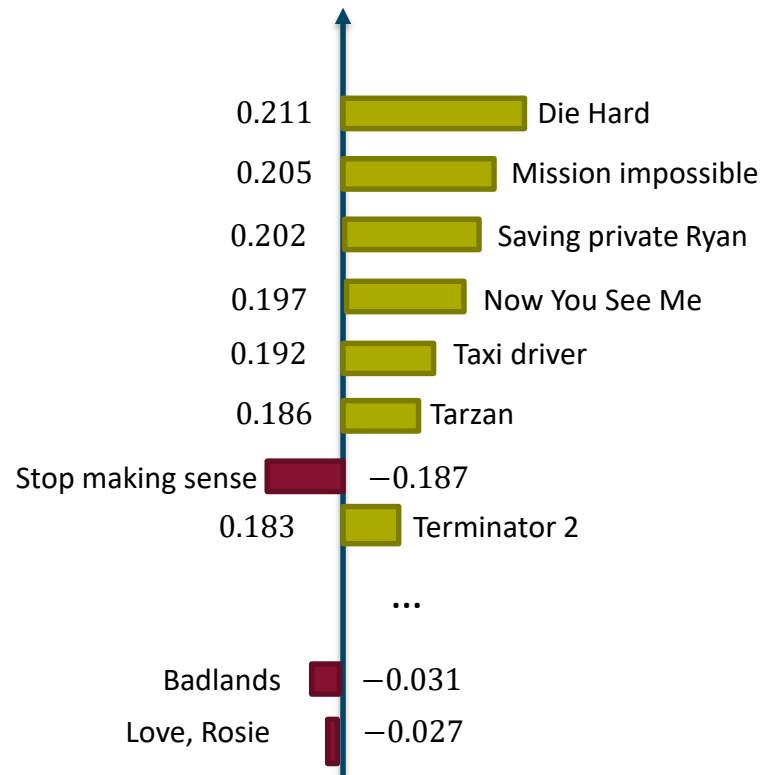
## 2. EXPLANATION METHODS

### SHAP – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam  
Lasso regularization



## 2. EXPLANATION METHODS

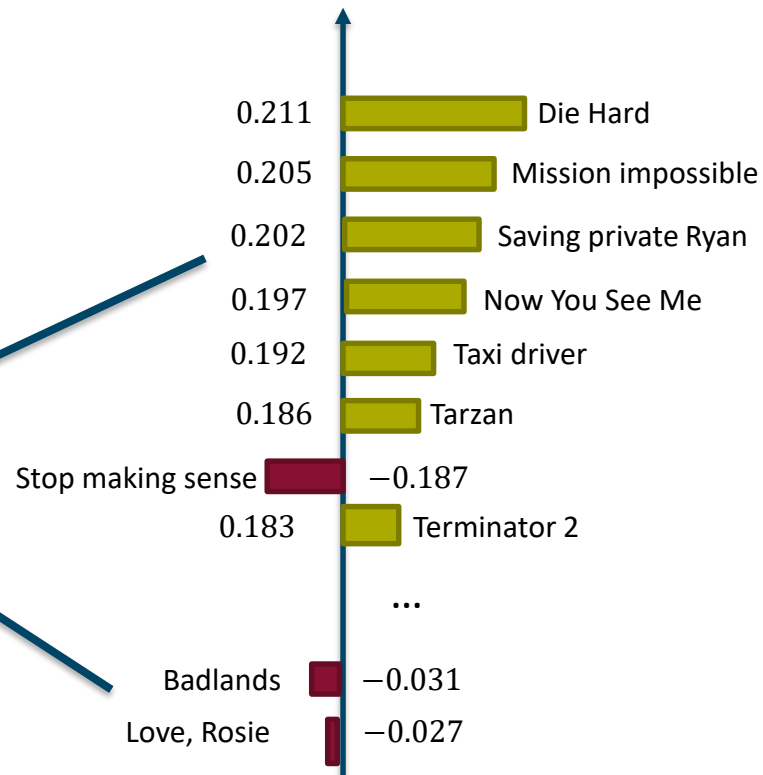
### SHAP – example

**Example:** gender prediction using movie viewing data



User  $x_i$ : Sam  
Lasso regularization

**BOTH POSITIVE & NEGATIVE EVIDENCE**



### 3. EVALUATION CRITERIA

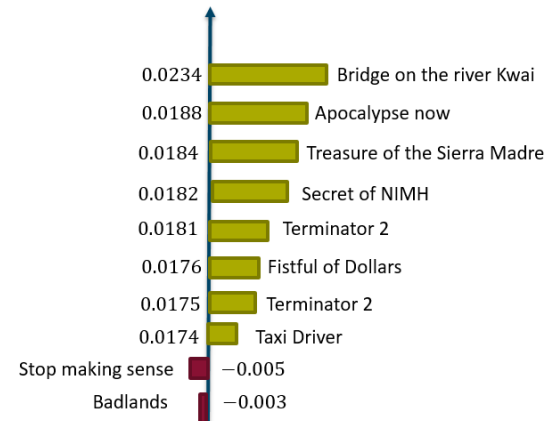
- ⇒ **NOT** a **qualitative** evaluation
- ⇒ No evaluation of *counterfactual* versus *linear model*, *negative evidence*, *output size* etc.

#### Counterfactual

**IF** Sam would not have rated {*Taxi driver*, *North by Northwest*, *Bridge on the river Kwai*, *Terminator 2*, *Hunt for red October*, *Glengarry Glen Ross*}, **THEN** his predicted class would change from male to female

**VS**

#### Additive feature attribution



# 3. EVALUATION CRITERIA

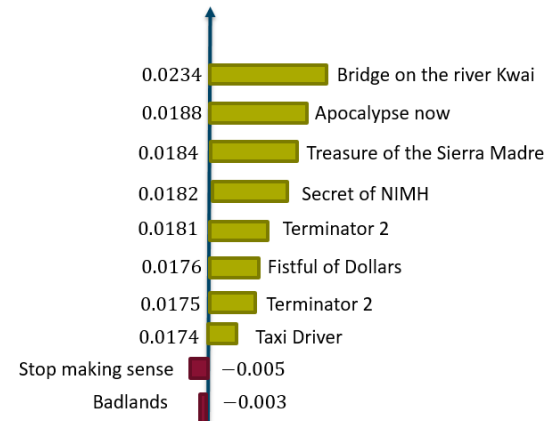
- ⇒ **NOT** a **qualitative** evaluation
- ⇒ No evaluation of *counterfactual* versus *linear model*, *negative evidence*, *output size* etc.

## Counterfactual

**IF** Sam would not have rated {*Taxi driver*, *North by Northwest*, *Bridge on the river Kwai*, *Terminator 2*, *Hunt for red October*, *Glengarry Glen Ross*}, **THEN** his predicted class would change from male to female

**VS**

## Additive feature attribution



⇒ **Quantitative evaluation**

# 3. EVALUATION CRITERIA

## 1. Effectiveness

- **Switching point:** number of features (with positive weight) that need to be removed before the classification changes
- **% of switching points found**
- **% generated output**
- **Output size**

## 2. Efficiency

- **Computation time:** number of seconds to generate explanation

# 4. EXPERIMENTAL SETUP

Collect data sets  
and build models

Textual data:  
linear/rbf SVM

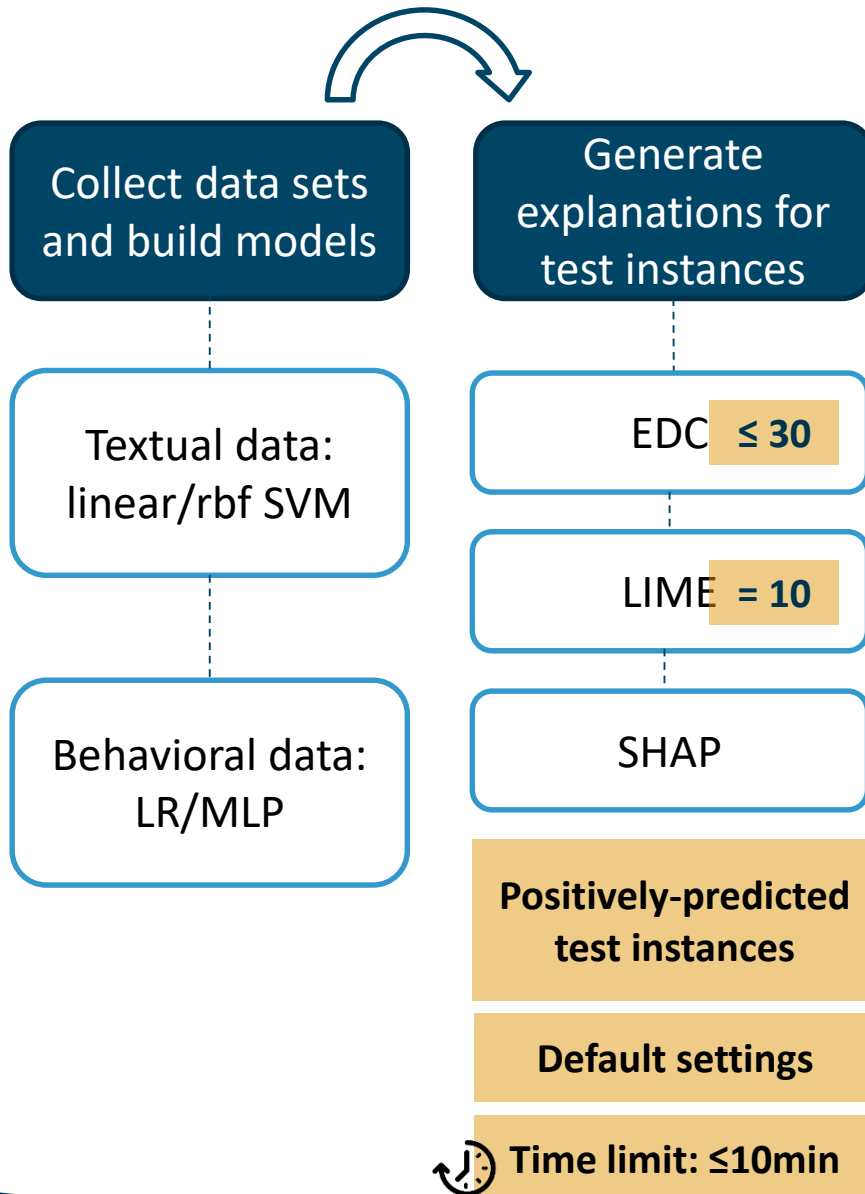
Behavioral data:  
LR/MLP

**Table 1: Data sets and characteristics**

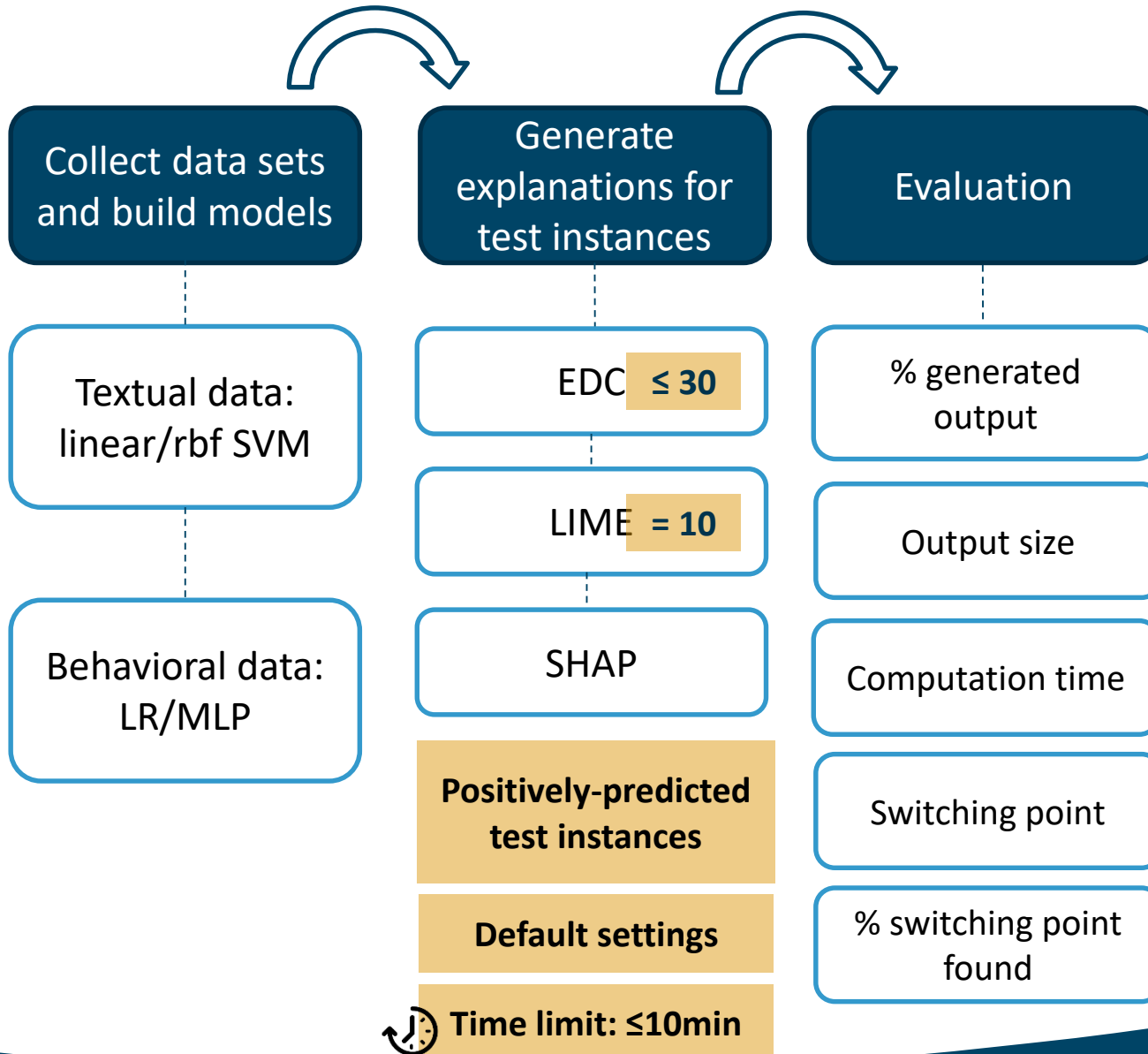
Dataset	Type	Target	Instances	Features	$b$	$p$
Movielens_1m*	B	gender	6,040	3,706	28.29%	95.5316%
Movielens_100k	B	gender	943	1,682	28.95%	93.6953%
YahooMovies*	B	gender	7,642	11,915	28.87%	99.7596%
Ecommerce*	B	gender	15,000	21,880	21.98%	99.9898%
Facebook*	B	gender	386,321	122,924	44.57%	99.9416%
KDD2015*	B	dropout	120,542	4,835	20.71%	99.6707%
Fraud*	B	fraudulent	858,131	107,345	0.000064%	99.9979%
TaFeng*	B	age	31,640	23,719	45.23%	99.9036%
Flickr*	B	comments	100,000	190,991	36.91%	99.9877%
LibimSeTi*	B	gender	137,806	166,353	44.53%	99.9317%
20news	T	atheism	18,846	41,356	4.24%	99.8435%
Airline*	T	sentiment	14,640	5,183	16.14%	99.8191%
Twitter	T	topic	6,090	4,569	9.15%	99.7428%



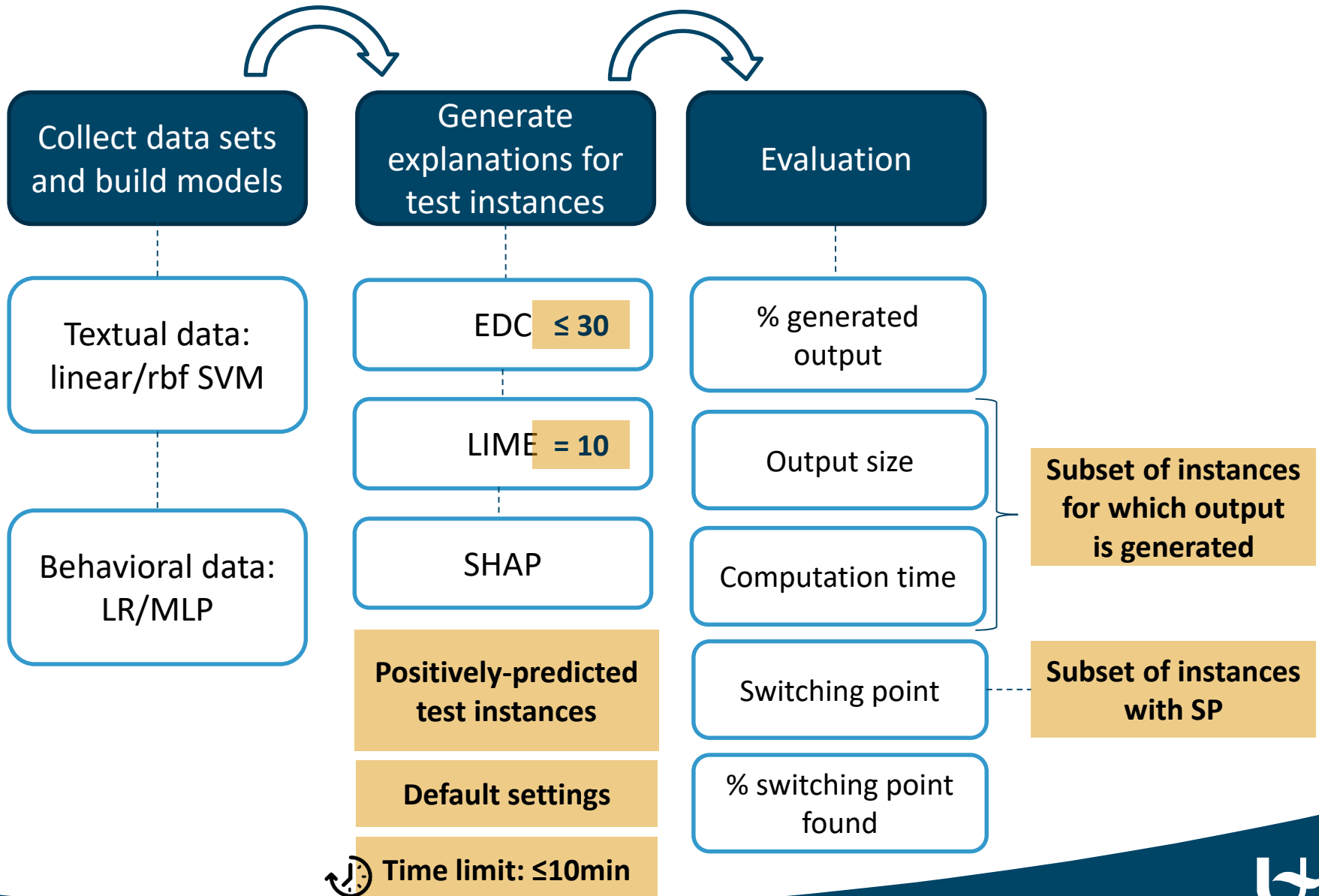
## 4. EXPERIMENTAL SETUP



## 4. EXPERIMENTAL SETUP



# 4. EXPERIMENTAL SETUP



# 5. RESULTS

**Table 4: Percentage of generated output.** For stochastic LIME/SHAP, this are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<u>98.67</u>	<b>100</b>	<b>100</b>	<u>89.67</u>	<b>100</b>	<b>100</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	98.67	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	<b>100</b>	<b>100</b>	<u>95</u>	<b>100</b>	<b>100</b>
Facebook	<u>96.67</u>	<b>100</b>	<b>100</b>	<u>70.33</u>	<b>100</b>	<b>100</b>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	99.67	<b>100</b>	<b>100</b>
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
LibimSeTi	<u>95.67</u>	<b>100</b>	<b>100</b>	<u>77.33</u>	<b>100</b>	<b>100</b>
20news	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	99.31	<b>100</b>	98.59	88.67	<b>100</b>	98.08
# wins	10	<b>13</b>	12	6	<b>13</b>	12

# 5. RESULTS

**Table 4: Percentage of generated output.** For stochastic LIME/SHAP, this are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<u>98.67</u>	<b>100</b>	<b>100</b>	<u>89.67</u>	<b>100</b>	<b>100</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	<u>98.67</u>	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	<b>100</b>	<b>100</b>	<u>95</u>	<b>100</b>	<b>100</b>
Facebook	<u>96.67</u>	<b>100</b>	<b>100</b>	<u>70.33</u>	<b>100</b>	<b>100</b>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	<u>99.67</u>	<b>100</b>	<b>100</b>
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
LibimSeTi	<u>95.67</u>	<b>100</b>	<b>100</b>	<u>77.33</u>	<b>100</b>	<b>100</b>
20news	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	99.31	<b>100</b>	98.59	88.67	<b>100</b>	98.08
# wins	10	<b>13</b>	12	6	<b>13</b>	12

# 5. RESULTS

**Table 4: Percentage of generated output.** For stochastic LIME/SHAP, this are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

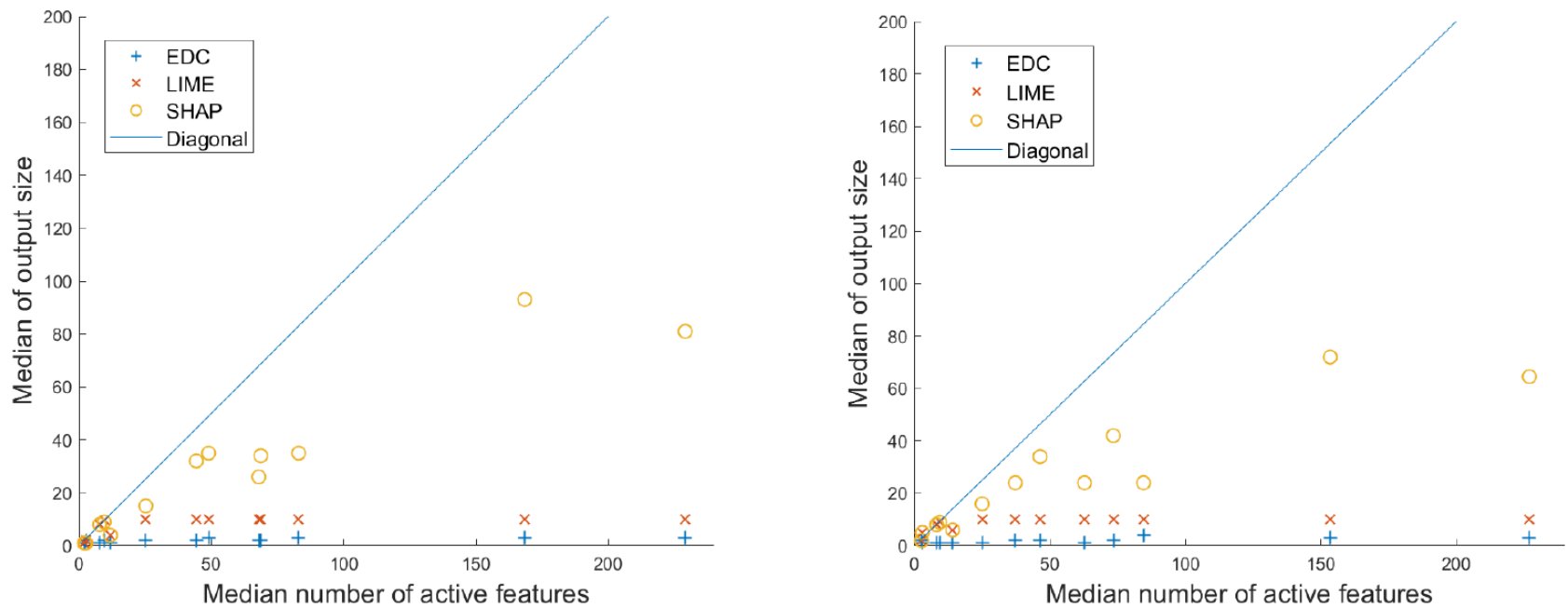
Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<u>98.67</u>	100	100	<u>89.67</u>	100	100
Movielens_100k	100	100	100	100	100	100
YahooMovies	100	100	100	98.67	100	100
Ecommerce	100	100	100	95	100	100
Facebook	<u>96.67</u>	100	100	<u>70.33</u>	100	100
KDD2015	100	100	100	99.67	100	100
Fraud	100	100	<u>81.67</u>	100	100	<u>75</u>
TaFeng	100	100	100	<u>93.33</u>	100	100
Flickr	100	100	100	100	100	100
LibimSeTi	<u>95.67</u>	100	100	<u>77.33</u>	100	100
20news	100	100	100	100	100	100
Airline	100	100	100	100	100	100
Twitter	100	100	100	100	100	100
Average	99.31	100	98.59	88.67	100	98.08
# wins	10	13	12	6	13	12

# 5. RESULTS

**Table 4: Percentage of generated output.** For stochastic LIME/SHAP, this are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<u>98.67</u>	<b>100</b>	<b>100</b>	<u>89.67</u>	<b>100</b>	<b>100</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	<u>98.67</u>	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	<b>100</b>	<b>100</b>	<u>95</u>	<b>100</b>	<b>100</b>
Facebook	<u>96.67</u>	<b>100</b>	<b>100</b>	<u>70.33</u>	<b>100</b>	<b>100</b>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	<u>99.67</u>	<b>100</b>	<b>100</b>
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
LibimSeTi	<u>95.67</u>	<b>100</b>	<b>100</b>	<u>77.33</u>	<b>100</b>	<b>100</b>
20news	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	99.31	<b>100</b>	98.59	88.67	<b>100</b>	98.08
# wins	10	<b>13</b>	12	6	<b>13</b>	12

## 5. RESULTS



**Fig. 2:** Median of output size for linear (right) and nonlinear (left) models as a function of median number of active features.



## 5. RESULTS

**Table 2: Percentage of switching points found (smaller than 30).** For stochastic LIME/SHAP, these are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<b>98.67</b>	<b>98.67</b>	<b>98.67</b>	<u>89.67</u>	<b>95.67</b>	<b>95.67</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	98.67	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	97.33	<b>100</b>	<u>95.00</u>	<u>96.67</u>	<b>99.67</b>
Facebook	<b>96.67</b>	95.33	95.00	<u>70.33</u>	<b>93.67</b>	<u>90.00</u>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	99.67	<b>100</b>	99.67
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	99.33	<b>100</b>	<b>28.67</b>	<b>28.67</b>	<b>28.67</b>
LibimSeTi	<b>95.67</b>	<u>91.00</u>	<u>89.33</u>	<u>77.33</u>	<b>91.33</b>	89.67
20news	<b>100</b>	99.47	<b>100</b>	<b>100</b>	98.94	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	<b>99.31</b>	98.55	97.28	88.67	<b>92.69</b>	90.64
# wins	<b>13</b>	8	10	6	<b>11</b>	9

## 5. RESULTS

**Table 2: Percentage of switching points found (smaller than 30).** For stochastic LIME/SHAP, these are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<b>98.67</b>	<b>98.67</b>	<b>98.67</b>	<u>89.67</u>	<b>95.67</b>	<b>95.67</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	98.67	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	97.33	<b>100</b>	<u>95.00</u>	<u>96.67</u>	<b>99.67</b>
Facebook	<b>96.67</b>	95.33	95.00	<u>70.33</u>	<b>93.67</b>	<u>90.00</u>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	99.67	<b>100</b>	99.67
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	99.33	<b>100</b>	<b>28.67</b>	<b>28.67</b>	<b>28.67</b>
LibimSeTi	<b>95.67</b>	<u>91.00</u>	<u>89.33</u>	<u>77.33</u>	<b>91.33</b>	89.67
20news	<b>100</b>	99.47	<b>100</b>	<b>100</b>	98.94	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	<b>99.31</b>	98.55	97.28	88.67	<b>92.69</b>	90.64
# wins	<b>13</b>	8	10	6	<b>11</b>	9

## 5. RESULTS

**Table 2: Percentage of switching points found (smaller than 30).** For stochastic LIME/SHAP, these are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<b>98.67</b>	<b>98.67</b>	<b>98.67</b>	<u>89.67</u>	<b>95.67</b>	<b>95.67</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	98.67	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	97.33	<b>100</b>	<u>95.00</u>	<u>96.67</u>	<b>99.67</b>
Facebook	<b>96.67</b>	95.33	95.00	<u>70.33</u>	<b>93.67</b>	<u>90.00</u>
KDD2015	<b>100</b>	<b>100</b>	<u>100</u>	99.67	<b>100</b>	<u>99.67</u>
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<u>100</u>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	99.33	<b>100</b>	<b>28.67</b>	<b>28.67</b>	<b>28.67</b>
LibimSeTi	<b>95.67</b>	<u>91.00</u>	<u>89.33</u>	<u>77.33</u>	<b>91.33</b>	89.67
20news	<b>100</b>	99.47	<b>100</b>	<b>100</b>	98.94	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	<b>99.31</b>	98.55	97.28	88.67	<b>92.69</b>	90.64
# wins	<b>13</b>	8	10	6	<b>11</b>	9

## 5. RESULTS

**Table 2: Percentage of switching points found (smaller than 30).** For stochastic LIME/SHAP, these are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC (%)	LIME (%)	SHAP (%)	EDC (%)	LIME (%)	SHAP (%)
Movielens_1m	<b>98.67</b>	<b>98.67</b>	<b>98.67</b>	<u>89.67</u>	<b>95.67</b>	<b>95.67</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<u>100</u>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	<u>98.67</u>	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	97.33	<b>100</b>	<u>95.00</u>	<u>96.67</u>	<b>99.67</b>
Facebook	<b>96.67</b>	95.33	95.00	<u>70.33</u>	<b>93.67</b>	<u>90.00</u>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	99.67	<b>100</b>	99.67
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<u>100</u>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	99.33	<b>100</b>	<u>28.67</u>	<b>28.67</b>	<b>28.67</b>
LibimSeTi	<b>95.67</b>	<u>91.00</u>	<u>89.33</u>	<u>77.33</u>	<b>91.33</b>	89.67
20news	<b>100</b>	99.47	<b>100</b>	<u>100</u>	98.94	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	<b>99.31</b>	98.55	97.28	88.67	<b>92.69</b>	90.64
# wins	<b>13</b>	8	10	6	<b>11</b>	9

## 5. RESULTS

**Table 2: Percentage of switching points found (smaller than 30).** For stochastic LIME/SHAP, these are average percentages over 5 runs. The best percentages are indicated in bold. The percentages are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	EDC (%)	Linear		EDC (%)	Nonlinear	
		LIME (%)	SHAP (%)		LIME (%)	SHAP (%)
Movielens_1m	<b>98.67</b>	<b>98.67</b>	<b>98.67</b>	<u>89.67</u>	<b>95.67</b>	<b>95.67</b>
Movielens_100k	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
YahooMovies	<b>100</b>	<b>100</b>	<b>100</b>	98.67	<b>100</b>	<b>100</b>
Ecommerce	<b>100</b>	97.33	<b>100</b>	<u>95.00</u>	<u>96.67</u>	<b>99.67</b>
Facebook	<b>96.67</b>	95.33	95.00	<u>70.33</u>	<b>93.67</b>	<u>90.00</u>
KDD2015	<b>100</b>	<b>100</b>	<b>100</b>	99.67	<b>100</b>	99.67
Fraud	<b>100</b>	<b>100</b>	<u>81.67</u>	<b>100</b>	<b>100</b>	<u>75</u>
TaFeng	<b>100</b>	<b>100</b>	<b>100</b>	<u>93.33</u>	<b>100</b>	<b>100</b>
Flickr	<b>100</b>	99.33	<b>100</b>	<b>28.67</b>	<b>28.67</b>	<b>28.67</b>
LibimSeTi	<b>95.67</b>	<u>91.00</u>	<u>89.33</u>	<u>77.33</u>	<b>91.33</b>	89.67
20news	<b>100</b>	99.47	<b>100</b>	<b>100</b>	98.94	<b>100</b>
Airline	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Twitter	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	<b>99.31</b>	98.55	97.28	88.67	<b>92.69</b>	90.64
# wins	<b>13</b>	8	10	6	<b>11</b>	9

# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

	Linear				Nonlinear			
Dataset	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	<u>2(1 – 3)</u>	2(1 – 3)	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<u>3(1-7)</u>	<u>3(1-7)</u>	<u>8(3 – 17)</u>	<u>2(1-3)</u>	<b>2(1-3.8)</b>	<u>2(1-4)</u>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	30(13.75 – 55)	<b>2.5(1-5)</b>	<u>3(2 – 8.2)</u>	<u>3(2 – 7.2)</u>	22(9.75 – 43.25)
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<u>1(1-3)</u>	<u>1(1-3)</u>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3

# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

	Linear				Nonlinear			
Dataset	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	2(1 – 3)	2(1 – 3)	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u><b>1(1-2)</b></u>	<b>1(1-1)</b>	<u><b>1(1-1)</b></u>	<b>1(1-1)</b>	<u><b>1(1-1)</b></u>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<u><b>3(1-7)</b></u>	<u><b>3(1-7)</b></u>	<u>8(3 – 17)</u>	<u><b>2(1-3)</b></u>	<b>2(1-3.8)</b>	<u><b>2(1-4)</b></u>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u><b>1(1-1)</b></u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u><b>1(1-2)</b></u>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u><b>1(1-1)</b></u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u><b>1(1-2)</b></u>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	30(13.75 – 55)	<b>2.5(1-5)</b>	<u>3(2 – 8.2)</u>	<u>3(2 – 7.2)</u>	<u>22(9.75 – 43.25)</u>
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<u><b>1(1-3)</b></u>	<u><b>1(1-3)</b></u>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3



# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear				Nonlinear			
	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	<u>2(1 – 3)</u>	2(1 – 3)	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<u>3(1-7)</u>	<u>3(1-7)</u>	<u>8(3 – 17)</u>	<u>2(1-3)</u>	<b>2(1-3.8)</b>	<u>2(1-4)</u>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	<u>30(13.75 – 55)</u>	<b>2.5(1-5)</b>	<u>3(2 – 8.2)</u>	<u>3(2 – 7.2)</u>	<u>22(9.75 – 43.25)</u>
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<u>1(1-3)</u>	<u>1(1-3)</u>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3



# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

	Linear				Nonlinear			
Dataset	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	2(1 – 3)	2(1 – 3)	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<u>3(1-7)</u>	<u>3(1-7)</u>	<u>8(3 – 17)</u>	<del>2(1-3)</del>	<b>2(1-3.8)</b>	<u>2(1-4)</u>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	<u>30(13.75 – 55)</u>	<b>2.5(1-5)</b>	<u>3(2 – 8.2)</u>	<u>3(2 – 7.2)</u>	<u>22(9.75 – 43.25)</u>
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<u>1(1-3)</u>	<u>1(1-3)</u>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3

# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

	Linear				Nonlinear			
Dataset	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<u>3(1-7)</u>	<u>3(1-7)</u>	<u>8(3 – 17)</u>	<u>2(1-3)</u>	<b>2(1-3.8)</b>	<u>2(1-4)</u>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-1)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>1(1-2)</u>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	<u>30(13.75 – 55)</u>	<b>2.5(1-5)</b>	<b>3(2 – 8.2)</b>	<b>3(2 – 7.2)</b>	<u>22(9.75 – 43.25)</u>
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<b>1(1-3)</b>	<b>1(1-3)</b>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3

# 5. RESULTS

**Table 3:** Median and interquantile range of **absolute switching point** with corresponding relative predicted score change. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The switching point is measured over the subset of instances where *all* methods have found a switching point. The best (median) absolute switching points are indicated in bold. The values are underlined if a method is significantly worse than the best method on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear				Nonlinear			
	EDC	LIME	SHAP	Random	EDC	LIME	SHAP	Random
Movielens_1m	<b>3(2-7)</b>	<b>3(2-7)</b>	<b>3(2-7)</b>	9(4 – 19)	<b>3(1-6)</b>	<b>3(2-8)</b>	<b>3(2-8)</b>	7(3 – 14.5)
Movielens_100k	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5.5(3 – 10)</u>	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(2 – 9.75)</u>
YahooMovies	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>4(2 – 7)</u>	<b>1(1-3)</b>	<u>2(1 – 3)</u>	<u>2(1 – 3)</u>	<u>4(2 – 12)</u>
Ecommerce	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-2)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>
Facebook	<b>3(2-8)</b>	<u>4(2 – 8.6)</u>	<u>4(2 – 8)</u>	<u>8(4 – 20)</u>	<u>4.5(1 – 13.25)</u>	<b>4(2-9.2)</b>	<u>4.4(2 – 10.4)</u>	<u>9.5(4 – 20)</u>
KDD2015	<b>3(1-7)</b>	<b>3(1-7)</b>	<b>3(1-7)</b>	<u>8(3 – 17)</u>	<b>2(1-3)</b>	<b>2(1-3.8)</b>	<b>2(1-4)</b>	<u>4.5(2 – 9)</u>
Fraud	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-2)</b>
TaFeng	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>5(3 – 11)</u>	<b>2(1-8)</b>	<b>2(1-3)</b>	<b>2(1-4)</b>	<u>6(3 – 17)</u>
Flickr	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-2)</b>
Libimseti	<b>3(2-7)</b>	<b>3(2-8.2)</b>	<b>3(2-7.4)</b>	<u>30(13.75 – 55)</u>	<b>2.5(1-5)</b>	<u>3(2 – 8.2)</u>	<u>3(2 – 7.2)</u>	<u>22(9.75 – 43.25)</u>
20news	<b>2(1-4)</b>	<b>2(1-4)</b>	<b>2(1-4)</b>	<u>11(4 – 24)</u>	<b>1(1-3)</b>	<b>1(1-3)</b>	<b>1(1-3)</b>	<u>8(3 – 19)</u>
Airline	<b>1(1-2)</b>	<b>1(1-2)</b>	<b>1(1-2)</b>	<u>2(1 – 3)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>2(1 – 3)</u>
Twitter	<b>2(1-3)</b>	<b>2(1-3)</b>	<b>2(1-3)</b>	<u>3(2 – 5)</u>	<b>1(1-1)</b>	<b>1(1-1)</b>	<b>1(1-1)</b>	<u>3(2 – 5.5)</u>
# wins	<b>13</b>	12	12	3	<b>12</b>	11	10	3

# 5. RESULTS

**Table 5:** Median and interquantile range of **computation time in seconds**. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The best (median) computation times are indicated in bold. The values are underlined if a method is **significantly worse than the best method** on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC	LIME	SHAP	EDC	LIME	SHAP
Ecommerce	<b>0.00(0.00-0.01)</b>	0.22(0.19 – 0.24)	0.00(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.27(0.26 – 0.28)	0.01(0.01 – 0.01)
Facebook	<b>0.11(0.02-1.19)</b>	0.35(0.28 – 0.51)	0.90(0.84 – 1.03)	<b>0.17(0.02-2.03)</b>	0.39(0.32 – 0.55)	0.95(0.88 – 1.07)
Flickr	<b>0.00(0.00-0.00)</b>	0.19(0.19 – 0.23)	0.01(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.24(0.23 – 0.25)	0.01(0.01 – 0.08)
Fraud	<b>0.00(0.00-0.02)</b>	0.24(0.22 – 0.28)	0.02(0.01 – 0.17)	<b>0.02(0.02-0.02)</b>	0.65(0.60 – 0.72)	0.05(0.02 – 0.82)
KDD2015	<b>0.09(0.02-0.74)</b>	0.36(0.32 – 0.43)	0.87(0.82 – 0.92)	<b>0.14(0.03-0.53)</b>	0.57(0.52 – 0.64)	1.07(1.02 – 1.13)
Libimseti	<b>0.37(0.13-3.12)</b>	0.70(0.59 – 0.97)	1.17(1.09 – 1.38)	<b>0.84(0.19-3.48)</b>	0.71(0.58 – 0.97)	1.18(1.09 – 1.39)
Movielens_1m	<b>0.34(0.06-2.92)</b>	0.56(0.35 – 0.99)	1.06(0.30 – 1.39)	<b>0.35(0.06-1.59)</b>	0.72(0.51 – 1.24)	2.49(2.21 – 3.33)
TaFeng	<b>0.05(0.02-0.19)</b>	0.53(0.43 – 0.63)	1.99(1.68 – 2.27)	<b>0.03(0.02-0.39)</b>	0.55(0.47 – 0.68)	1.45(1.26 – 1.63)
Movielens_100k	<b>0.07(0.02-0.32)</b>	0.35(0.31 – 0.57)	0.87(0.83 – 1.04)	<b>0.14(0.07-0.70)</b>	0.42(0.34 – 0.66)	0.93(0.88 – 1.13)
YahooMovies	<b>0.07(0.02-0.19)</b>	0.27(0.26 – 0.31)	0.80(0.77 – 0.83)	<b>0.09(0.04-0.30)</b>	0.63(0.62 – 0.67)	1.11(1.08 – 1.16)
20news	<b>0.19(0.05-1.43)</b>	2.95(1.88 – 3.96)	3.36(2.49 – 4.16)	<b>0.10(0.03-0.76)</b>	1.94(1.34 – 2.69)	2.39(1.88 – 2.95)
Airline	<b>0.02(0.01-0.04)</b>	0.79(0.62 – 0.91)	0.08(0.02 – 0.59)	<b>0.02(0.02-0.03)</b>	1.18(0.96 – 1.33)	0.10(0.02 – 0.79)
Twitter	<b>0.03(0.01-0.05)</b>	1.21(1.09 – 1.32)	0.37(0.09 – 1.09)	<b>0.01(0.01-0.01)</b>	0.89(0.82 – 0.95)	0.13(0.03 – 0.43)
# wins	13	0	0	13	0	0

# 5. RESULTS

**Table 5:** Median and interquantile range of **computation time in seconds**. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The best (median) computation times are indicated in bold. The values are underlined if a method is **significantly worse than the best method** on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC	LIME	SHAP	EDC	LIME	SHAP
Ecommerce	<b>0.00(0.00-0.01)</b>	0.22(0.19 – 0.24)	0.00(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.27(0.26 – 0.28)	0.01(0.01 – 0.01)
Facebook	<b>0.11(0.02-1.19)</b>	0.35(0.28 – 0.51)	0.90(0.84 – 1.03)	<b>0.17(0.02-2.03)</b>	0.39(0.32 – 0.55)	0.95(0.88 – 1.07)
Flickr	<b>0.00(0.00-0.00)</b>	0.19(0.19 – 0.23)	0.01(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.24(0.23 – 0.25)	0.01(0.01 – 0.08)
Fraud	<b>0.00(0.00-0.02)</b>	0.24(0.22 – 0.28)	0.02(0.01 – 0.17)	<b>0.02(0.02-0.02)</b>	0.65(0.60 – 0.72)	0.05(0.02 – 0.82)
KDD2015	<b>0.09(0.02-0.74)</b>	0.36(0.32 – 0.43)	0.87(0.82 – 0.92)	<b>0.14(0.03-0.53)</b>	0.57(0.52 – 0.64)	1.07(1.02 – 1.13)
Libimseti	<b>0.37(0.13-3.12)</b>	0.70(0.59 – 0.97)	1.17(1.09 – 1.38)	<b>0.84(0.19-3.48)</b>	0.71(0.58 – 0.97)	1.18(1.09 – 1.39)
Movielens_1m	<b>0.34(0.06-2.92)</b>	0.56(0.35 – 0.99)	1.06(0.30 – 1.39)	<b>0.35(0.06-1.59)</b>	0.72(0.51 – 1.24)	2.49(2.21 – 3.33)
TaFeng	<b>0.05(0.02-0.19)</b>	0.53(0.43 – 0.63)	1.99(1.68 – 2.27)	<b>0.03(0.02-0.39)</b>	0.55(0.47 – 0.68)	1.45(1.26 – 1.63)
Movielens_100k	<b>0.07(0.02-0.32)</b>	0.35(0.31 – 0.57)	0.87(0.83 – 1.04)	<b>0.14(0.07-0.70)</b>	0.42(0.34 – 0.66)	0.93(0.88 – 1.13)
YahooMovies	<b>0.07(0.02-0.19)</b>	0.27(0.26 – 0.31)	0.80(0.77 – 0.83)	<b>0.09(0.04-0.30)</b>	0.63(0.62 – 0.67)	1.11(1.08 – 1.16)
20news	<b>0.19(0.05-1.43)</b>	2.95(1.88 – 3.96)	3.36(2.49 – 4.16)	<b>0.10(0.03-0.76)</b>	1.94(1.34 – 2.69)	2.39(1.88 – 2.95)
Airline	<b>0.02(0.01-0.04)</b>	0.79(0.62 – 0.91)	0.08(0.02 – 0.59)	<b>0.02(0.02-0.03)</b>	1.18(0.96 – 1.33)	0.10(0.02 – 0.79)
Twitter	<b>0.03(0.01-0.05)</b>	1.21(1.09 – 1.32)	0.37(0.09 – 1.09)	<b>0.01(0.01-0.01)</b>	0.89(0.82 – 0.95)	0.13(0.03 – 0.43)
# wins	13	0	0	13	0	0



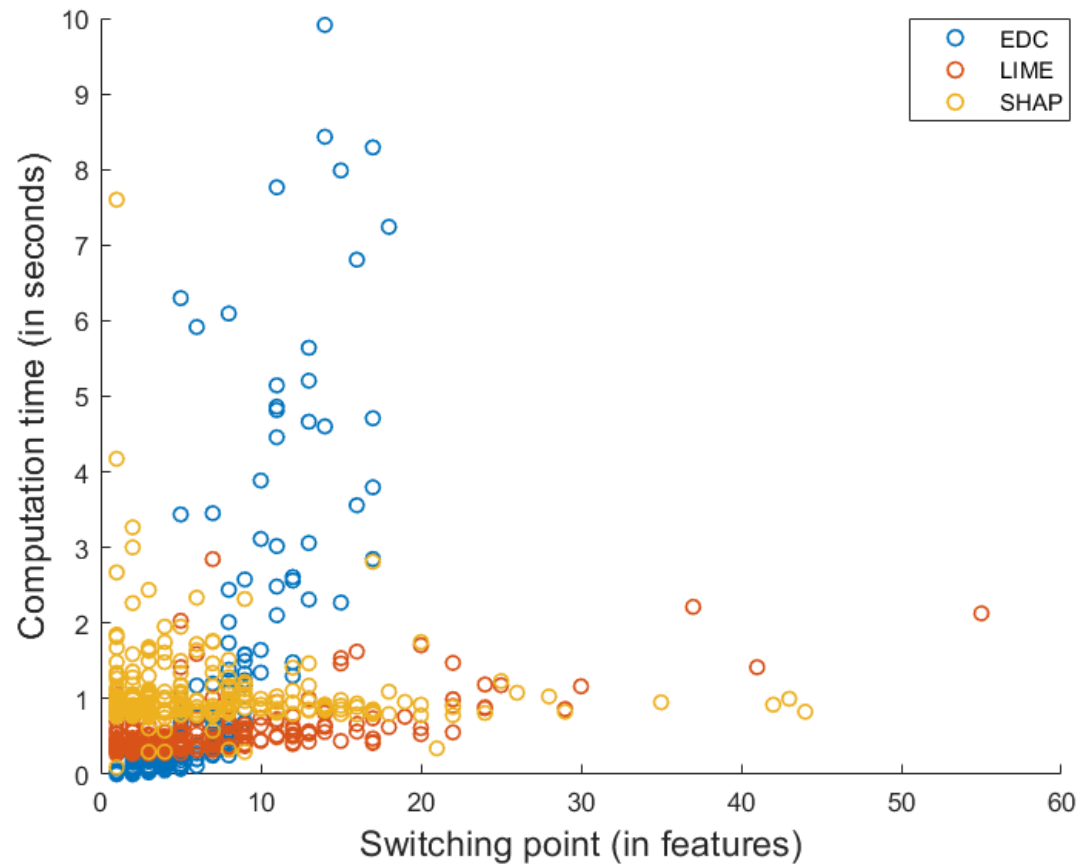
# 5. RESULTS

**Table 5:** Median and interquantile range of **computation time in seconds**. For stochastic LIME/SHAP, this is the average median/range over 5 runs. The best (median) computation times are indicated in bold. The values are underlined if a method is **significantly worse than the best method** on a 1% significance level using a McNemar mid-p test [17].

Dataset	Linear			Nonlinear		
	EDC	LIME	SHAP	EDC	LIME	SHAP
Ecommerce	<b>0.00(0.00-0.01)</b>	0.22(0.19 – 0.24)	0.00(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.27(0.26 – 0.28)	0.01(0.01 – 0.01)
Facebook	<b>0.11(0.02-1.19)</b>	0.35(0.28 – 0.51)	0.90(0.84 – 1.03)	<b>0.17(0.02-2.03)</b>	0.39(0.32 – 0.55)	0.95(0.88 – 1.07)
Flickr	<b>0.00(0.00-0.00)</b>	0.19(0.19 – 0.23)	0.01(0.00 – 0.01)	<b>0.00(0.00-0.02)</b>	0.24(0.23 – 0.25)	0.01(0.01 – 0.08)
Fraud	<b>0.00(0.00-0.02)</b>	0.24(0.22 – 0.28)	0.02(0.01 – 0.17)	<b>0.02(0.02-0.02)</b>	0.65(0.60 – 0.72)	0.05(0.02 – 0.82)
KDD2015	<b>0.09(0.02-0.74)</b>	0.36(0.32 – 0.43)	0.87(0.82 – 0.92)	<b>0.14(0.03-0.53)</b>	0.57(0.52 – 0.64)	1.07(1.02 – 1.13)
Libimseti	<b>0.37(0.13-3.12)</b>	0.70(0.59 – 0.97)	1.17(1.09 – 1.38)	<b>0.84(0.19-3.48)</b>	0.71(0.58 – 0.97)	1.18(1.09 – 1.39)
Movielens_1m	<b>0.34(0.06-2.92)</b>	0.56(0.35 – 0.99)	1.06(0.30 – 1.39)	<b>0.35(0.06-1.59)</b>	0.72(0.51 – 1.24)	2.49(2.21 – 3.33)
TaFeng	<b>0.05(0.02-0.19)</b>	0.53(0.43 – 0.63)	1.99(1.68 – 2.27)	<b>0.03(0.02-0.39)</b>	0.55(0.47 – 0.68)	1.45(1.26 – 1.63)
Movielens_100k	<b>0.07(0.02-0.32)</b>	0.35(0.31 – 0.57)	0.87(0.83 – 1.04)	<b>0.14(0.07-0.70)</b>	0.42(0.34 – 0.66)	0.93(0.88 – 1.13)
YahooMovies	<b>0.07(0.02-0.19)</b>	0.27(0.26 – 0.31)	0.80(0.77 – 0.83)	<b>0.09(0.04-0.30)</b>	0.63(0.62 – 0.67)	1.11(1.08 – 1.16)
20news	<b>0.19(0.05-1.43)</b>	2.95(1.88 – 3.96)	3.36(2.49 – 4.16)	<b>0.10(0.03-0.76)</b>	1.94(1.34 – 2.69)	2.39(1.88 – 2.95)
Airline	<b>0.02(0.01-0.04)</b>	0.79(0.62 – 0.91)	0.08(0.02 – 0.59)	<b>0.02(0.02-0.03)</b>	1.18(0.96 – 1.33)	0.10(0.02 – 0.79)
Twitter	<b>0.03(0.01-0.05)</b>	1.21(1.09 – 1.32)	0.37(0.09 – 1.09)	<b>0.01(0.01-0.01)</b>	0.89(0.82 – 0.95)	0.13(0.03 – 0.43)
# wins	13	0	0	13	0	0

## 5. RESULTS

Fig. 3: Computation time vs switching point for *Facebook/linear*



## 6. Discussion

### Ability to rank positive evidence → Switching point

- EDC provides optimal (*smallest*) switching points for *linear* models
- Heuristic best-first algorithm EDC: worse than LIME/SHAP for *some* non-linear models

### Percentage output generated

- When restricting the output size ( $\leq 30$ ), EDC *not always* generates output
- SHAP difficulties with *Fraud* data

### Explanation output size

- EDC provides *smallest* output sizes
- LIME can be *further reduced* if wanted
- SHAP cannot be *explicitly* restricted ( $\geq 50\%$  of active features included)

### Computational efficiency

- Instances that are small and/or “easy” to explain with counterfactuals  
→ EDC is most efficient
- LIME and SHAP *relatively fast* for all scenarios



# 7. Conclusion

## Comparative study of instance-level explanations EDC, LIME and SHAP for textual and behavioral data

⇒ A **nuanced** conclusion:

- **EDC** seems best for *small* instances and *linear* models
- **SHAP**
  - Consistently relatively fast
  - Low switching points
  - Seems to have difficulties with highly-imbalanced data
  - Very large outputs
- **LIME**: *best* trade-off
  - Consistently relatively fast
  - Low switching points
  - Ability to provide *k*

## **8. FURTHER RESEARCH**

### **1. Extension of quantitative evaluation**

- More data and models

### **2. Application (eg marketing, fraud detection)**

- Specific business needs / domain experts
- *Visualization* of explanations
- *Meta-features*  $\Leftrightarrow$  fine-grained features

### **3. Qualitative evaluation of explanations**

- Relevance of negative evidence?
- Counterfactual versus sparse, linear model?

# Thanks for your attention. Questions?



<https://www.linkedin.com/in/yanou-ramon>



<http://applieddatamining.com/cms/>



[yanou.ramon@uantwerp.be](mailto:yanou.ramon@uantwerp.be)