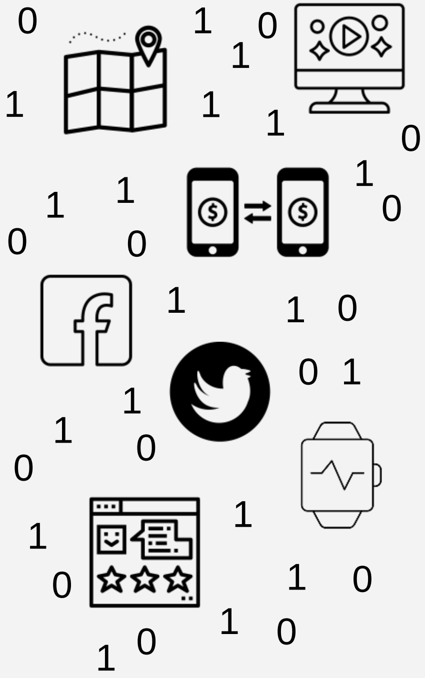
Counterfactual Algorithms for Explaining Prediction Models on Behavioral and Textual Data

Yanou Ramon, David Martens, Foster Provost, Theodoros Evgeniou

AIMLAI workshop (CIKM) – Oct. 20, 2020 – 3:40pm



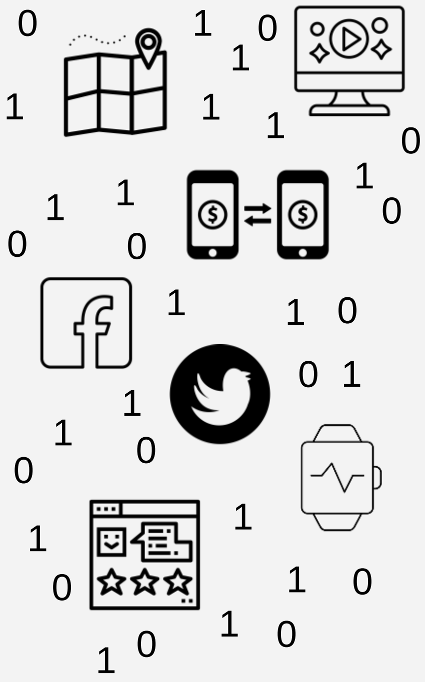
Behavioral and textual data

(High-dimensional & sparse)

Behavioral and textual data
(High-dimensional & sparse)

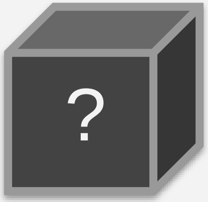
Prediction model

Predicted value of target variable



Behavioral and textual data
(High-dimensional & sparse)

Prediction model



“Black Box”

$\Rightarrow$ ***Thousands of coefficients***
$\Rightarrow$ ***Nonlinear techniques***

Predicted value of target variable

LOCATION DATA NYC: tourist or citizen?

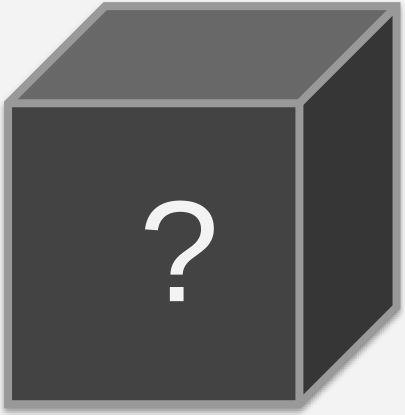
evidence = active feature

	Columbia University	Time Square	DUMBO	…	Chelsea Market	*Target $\hat{y}$ Tourist*
Anna	1	1	1	…	0	***1***
Jack	1	0	0	…	1	***0***
…	…	…	…	…	…	***…***
Bill	0	0	1	…	0	***0***

➔ **data is high-dimensional and sparse**

	Columbia University	Time Square	DUMBO	...	Chelsea Market	*Target $\hat{y}$ Tourist*
Anna	1	1	1	…	0	***1***
Jack	1	0	0	…	1	***0***
…	…	…	…	…	…	***…***
Bill	0	0	1	…	0	***0***

LOCATION DATA NYC



$\hat{y} = 1$ ***if*** *tourist*
else $\hat{y} = 0$

“Black Box” model
$\Rightarrow$ **Thousands of coefficients**
$\Rightarrow$ **Nonlinear techniques**

(Local) interpretability issues
➔ Counterfactual explanations

COUNTERFACTUAL EXPLANATIONS

- Instance-level
- Causality within the model
- Output is a rule: minimal set of features such that the predicted class changes when removing them (setting values to zero)
- Intuitive and valuable for humans → contrastive: *"Why X rather than not-X?"* (Miller, 2017)

COUNTERFACTUAL EXPLANATIONS

Example: Tourist prediction using NYC location data

Anna visited 120 places last month
Anna was predicted as “tourist”

COUNTERFACTUAL EXPLANATIONS

Example: Tourist prediction using NYC location data

Anna visited 120 places last month
Anna was predicted as "tourist"

Why?

COUNTERFACTUAL EXPLANATIONS

Example: Tourist prediction using NYC location data

Anna visited 120 places last month
Anna was predicted as "tourist"

		Columbia University	Time Square	DUMBO	...	Chelsea Market	Target $\hat{y}$ Tourist
$\mathbf{x}$	Anna	1	1	1	...	0	*1*
$\mathbf{z}_1$	Anna (perturbed)	*1*	*0*	*0*	...	*0*	*0*

IF Anna would not have visited **{Time Square, DUMBO}**,
THEN the predicted class changes from "tourist" to "NY citizen"


COUNTERFACTUAL
ALGORITHMS

DESIDERATA

- Model-agnostic
- Find **minimum-sized** counterfactual explanation $\boldsymbol{E}$ for a single model prediction of instance $\mathbf{x}$

DESIDERATA

- Model-agnostic
- Find **minimum-sized** counterfactual explanation $\boldsymbol{E}$ for a single model prediction of instance $\mathbf{x}$

More comprehensible (~cognitive limitations)

More actionable: e.g., "cloak" fewer online traces to get a desired outcome (not be targeted with ads of gay bars)

DESIDERATA

tourist

NY citizen

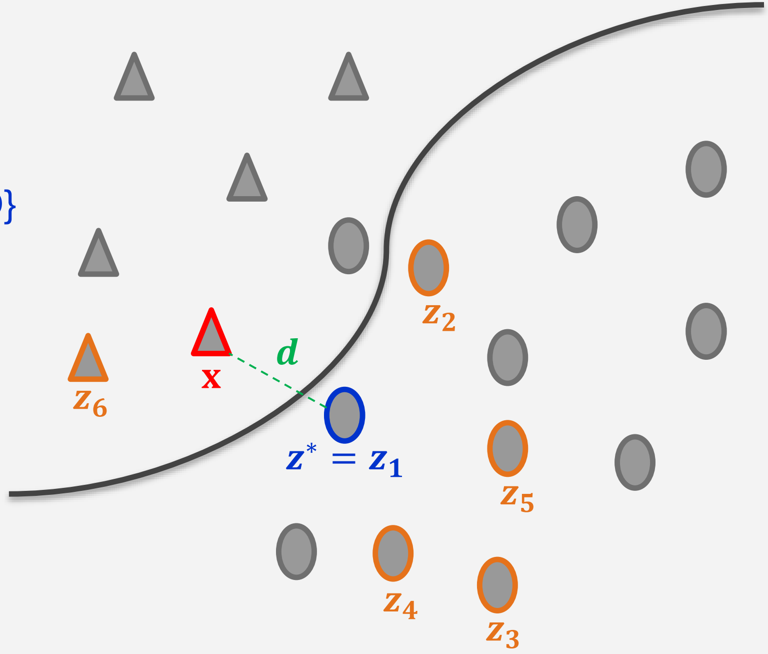
$z^* = x \backslash$ {Time Square, DUMBO}

$E^* =$ {Time Square, DUMBO}

original instance

perturbed instances

d distance



WHY COMPLETE SEARCH FAILS

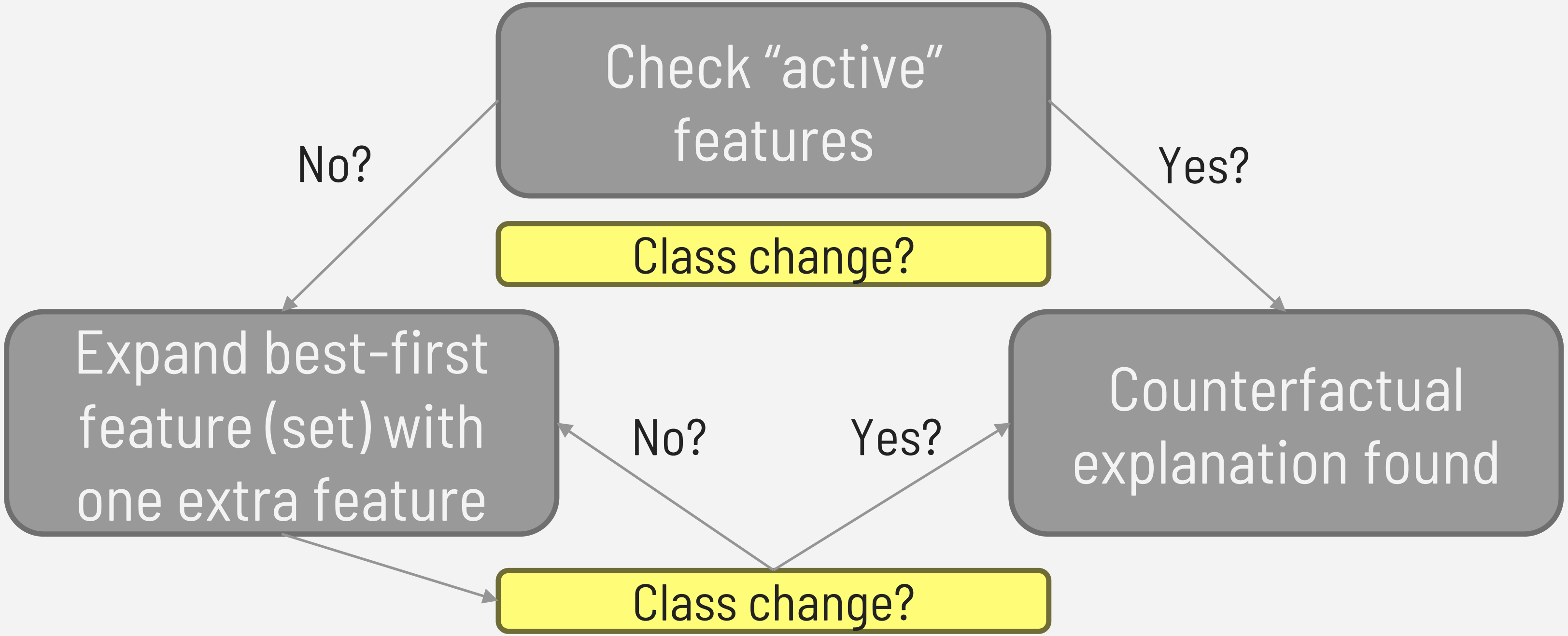
- Start with removing one feature and increase number of features in the subset until the predicted class changes
- Scales exponentially with active features m and required number of features k to be removed

  e.g., for an instance with m features, a combination of k features requires $\frac{m!}{(m-k)!k!}$ evaluations

BEST-FIRST SEARCH (SEDC)

- Explaining document classifications (Martens & Provost, 2013)
- Model-agnostic algorithm: heuristic best-first search
- Optimal for linear models

BEST-FIRST SEARCH (SEDC)



NOVEL HYBRID ALGORITHMS

Additive Feature Attribution (AFA) methods:

- LIME: Local Model-agnostic Explainer (Ribeiro et al., 2016)
- SHAP: Shapley Additive Explanations (Lundberg et al., 2018)

Output: Importance-ranked list

NOVEL HYBRID ALGORITHMS

Novelty: importance rankings may be an “intelligent” starting point for computing counterfactuals

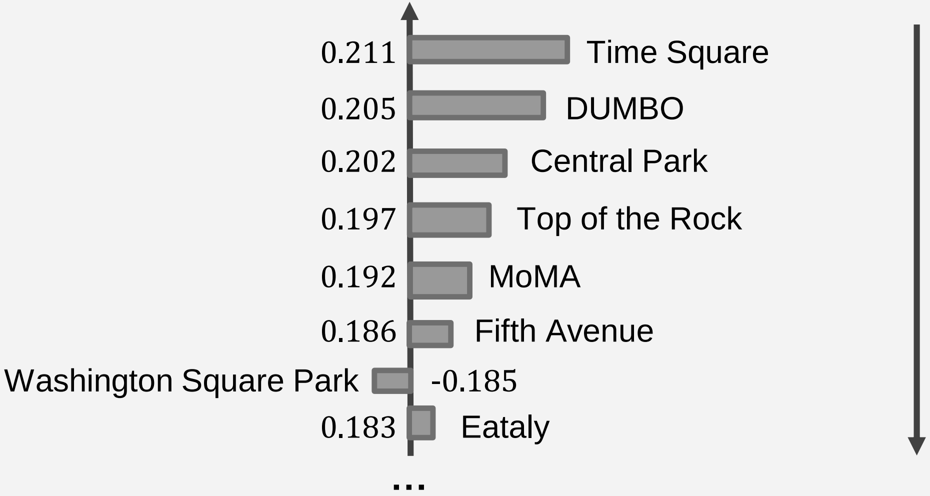
⇒ **LIME–C / SHAP-C**

⇒ Addresses open problem: how to select complexity of LIME/SHAP for models on behavior/text?

NOVEL HYBRID ALGORITHMS

LIME-C / SHAP-C

Example: Tourist prediction using NYC location data



Remove features with positive importance weight until the class changes

RESULTS & CONCLUSION

PERCENTAGE EXPLAINED

Table 2 Percentage explained (counterfactuals smaller than 30 features)

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

For stochastic *LIME-C/SHAP-C*, 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 (Fagerland et al. 2013)

PERCENTAGE EXPLAINED

Table 2 Percentage explained (counterfactuals smaller than 30 features)

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

For stochastic *LIME-C/SHAP-C*, 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 (Fagerland et al. 2013)

PERCENTAGE EXPLAINED

Table 2 Percentage explained (counterfactuals smaller than 30 features)

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

For stochastic *LIME-C/SHAP-C*, 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 (Fagerland et al. 2013)

PERCENTAGE EXPLAINED

Table 2 Percentage explained (counterfactuals smaller than 30 features)

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

For stochastic *LIME-C/SHAP-C*, 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 (Fagerland et al. 2013)

CONCLUSION

- **SEDC** most efficient and effective for small data instances, however:
  - weakness of best-first search for some nonlinear models

- **SHAP-C** overall good performance, however:
  - problems with highly unbalanced data
  - computation time more sensitive to # active features than LIME-C

- **LIME-C**: suitable alternative to SEDC for large data instances:
  - good effectiveness results for all data and models
  - low computation times
  - efficiency least sensitive to size of explanation

! Addresses open issue of LIME/SHAP: setting complexity parameter

CODE & TUTORIALS

Algorithms implemented with Python

SEDC: https://github.com/yramon/edc

LIME-C: https://github.com/yramon/LimeCounterfactual

SHAP-C: https://github.com/yramon/ShapCounterfactual

THANKS!

Further questions?

Mail: yanou.ramon@uantwerp.be
Website: https://yramon.github.io/
www.linkedin.com/in/yanou-ramon
www.applieddatamining.com