Gaining insight into AI systems on digital footprints

Yanou Ramon

PhD researcher @ Applied Data Mining

35th Data Science Leuven Meetup - March 2021

Universiteit Antwerpen

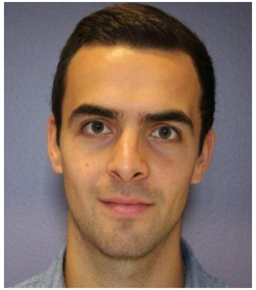
Research Foundation Flanders
Opening new horizons

Applied Data Mining

Prof. David Martens

Stiene Praet

Yanou Ramon



Tom Vermeire

Sofie Goethals

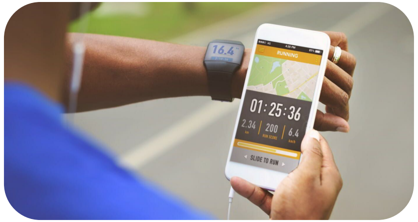
Raphael M.B. De Oliveira

Dieter Brughmans

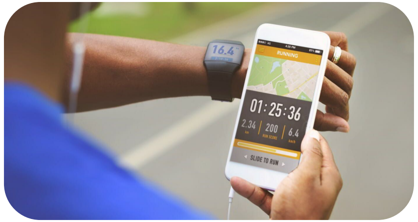
Overview

1. Interpretability issues of prediction models on behavioral data
2. Rule extraction with metafeatures
3. Empirical results
4. Implications
5. Key takeaways

Behavioral data



Behavioral data



Applications

Targeted advertising

Churn prediction

Fraud detection

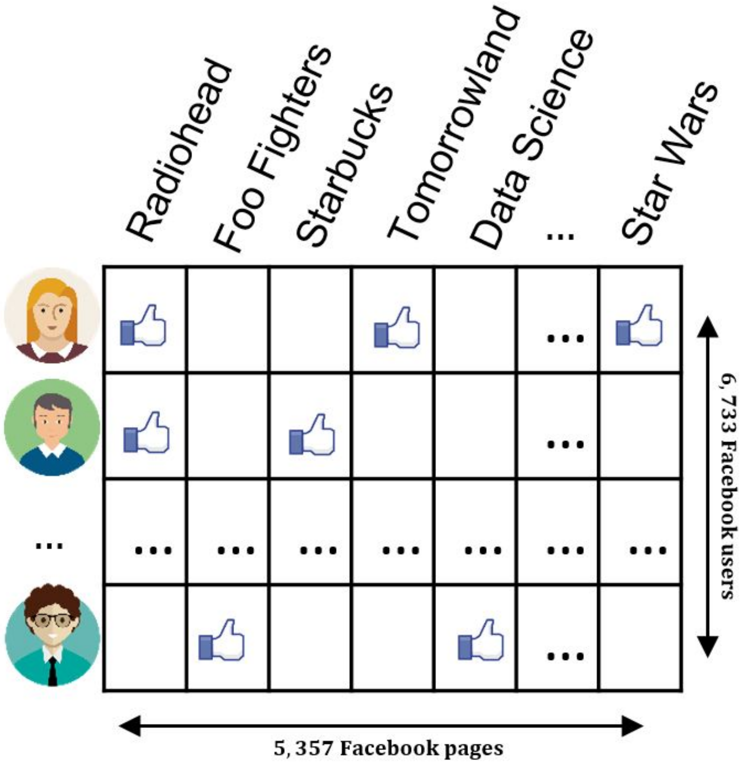
Credit scoring

Pricing

…

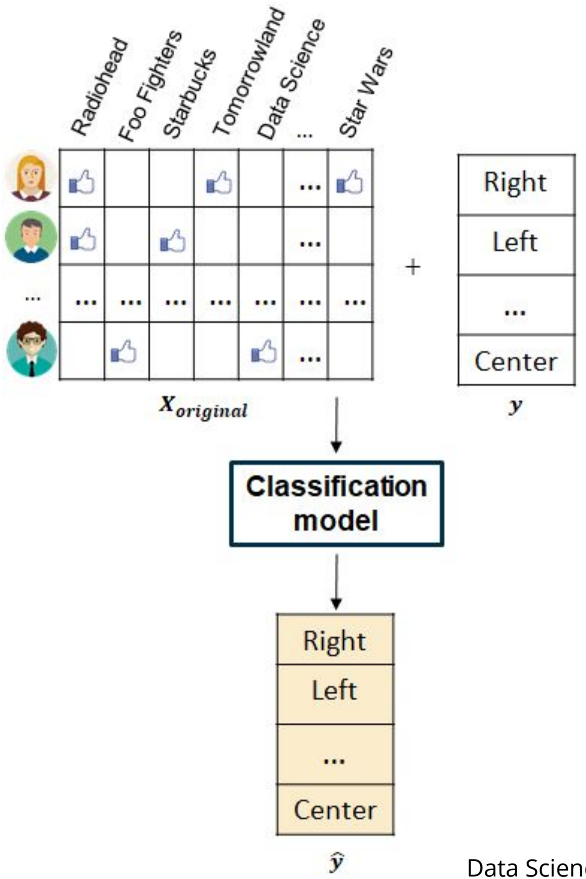
Prediction models on behavioral data *(Example: Facebook likes and political leaning)*

Reference: [Praet et al., 2018]



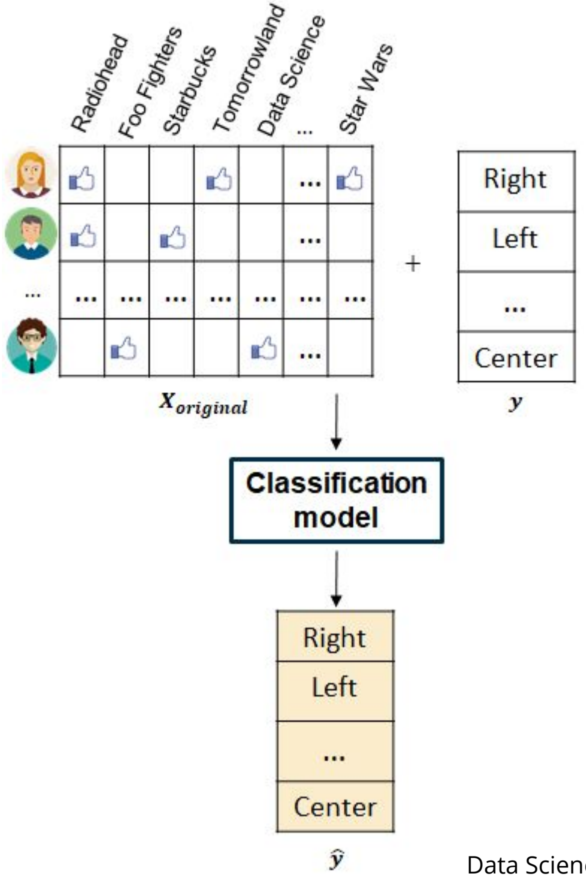
Prediction models on behavioral data *(Example: Facebook likes and political leaning)*

Reference: [Praet et al., 2018]



Prediction models on behavioral data *(Example: Facebook likes and political leaning)*

Reference: [Praet et al., 2018]



⚠ Interpretability issues

→ *Complexity of model*

→ *High dimensionality + sparsity + many relevant predictors*

Rule extraction *(e.g., Baesens et al., 2003, Martens et al., 2007, Guidotti et al., 2018)*

- Explanation rules mimic predictions of model
- Limited complexity → small set of rules
- Model predictions are used as labels instead of ground-truth labels

Rule extraction *(e.g., Baesens et al., 2003, Martens et al., 2007, Guidotti et al., 2018)*

- Explanation rules mimic predictions of model
- Limited complexity → small set of rules
- Model predictions are used as labels instead of ground-truth labels

⚠️**Challenge:** *High dimensionality + sparsity + many relevant predictors*

Rule extraction *(e.g., Baesens et al., 2003, Martens et al., 2007, Guidotti et al., 2018)*

- Explanation rules mimic predictions of model
- Limited complexity → small set of rules
- Model predictions are used as labels instead of ground-truth labels

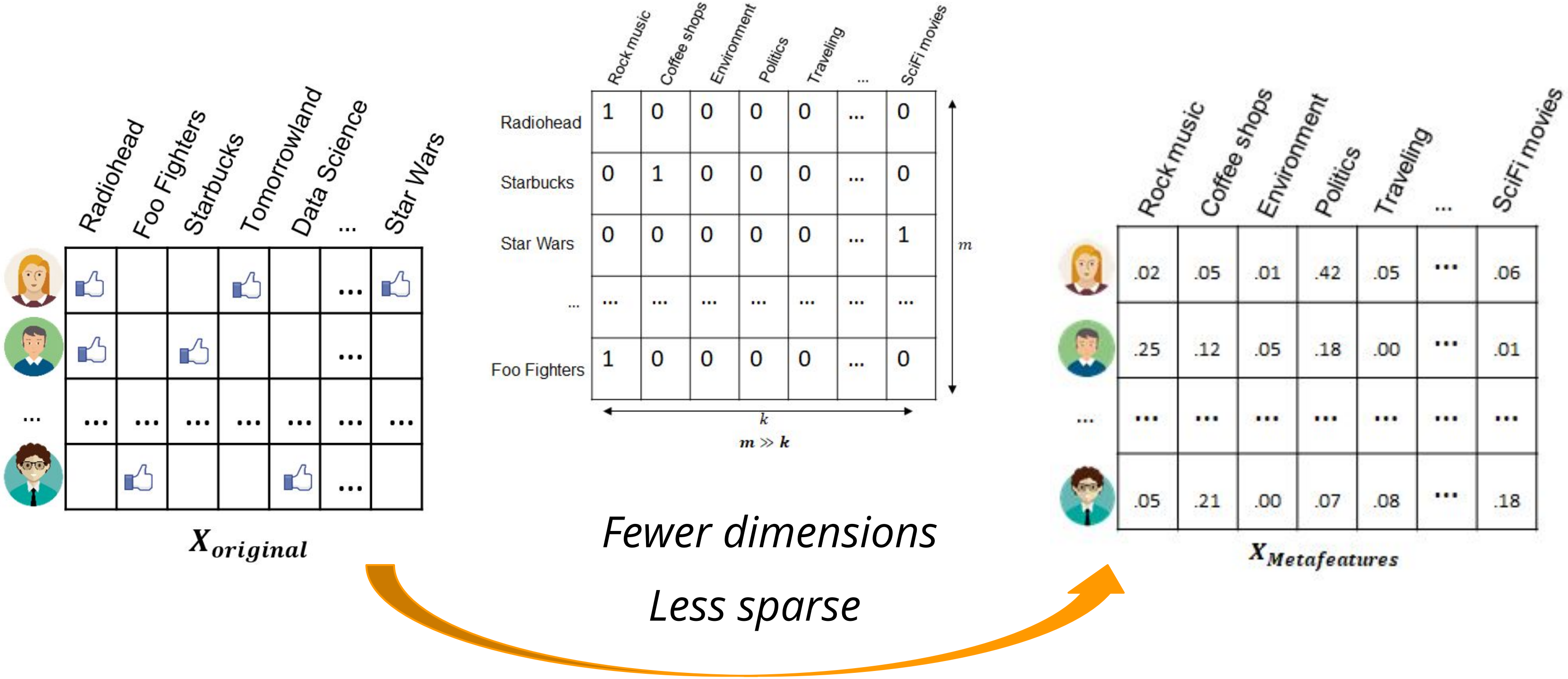
⚠️**Challenge:** *High dimensionality* + *sparsity* + *many relevant predictors*

⚠️Small explanation does not explain much of the model's behavior

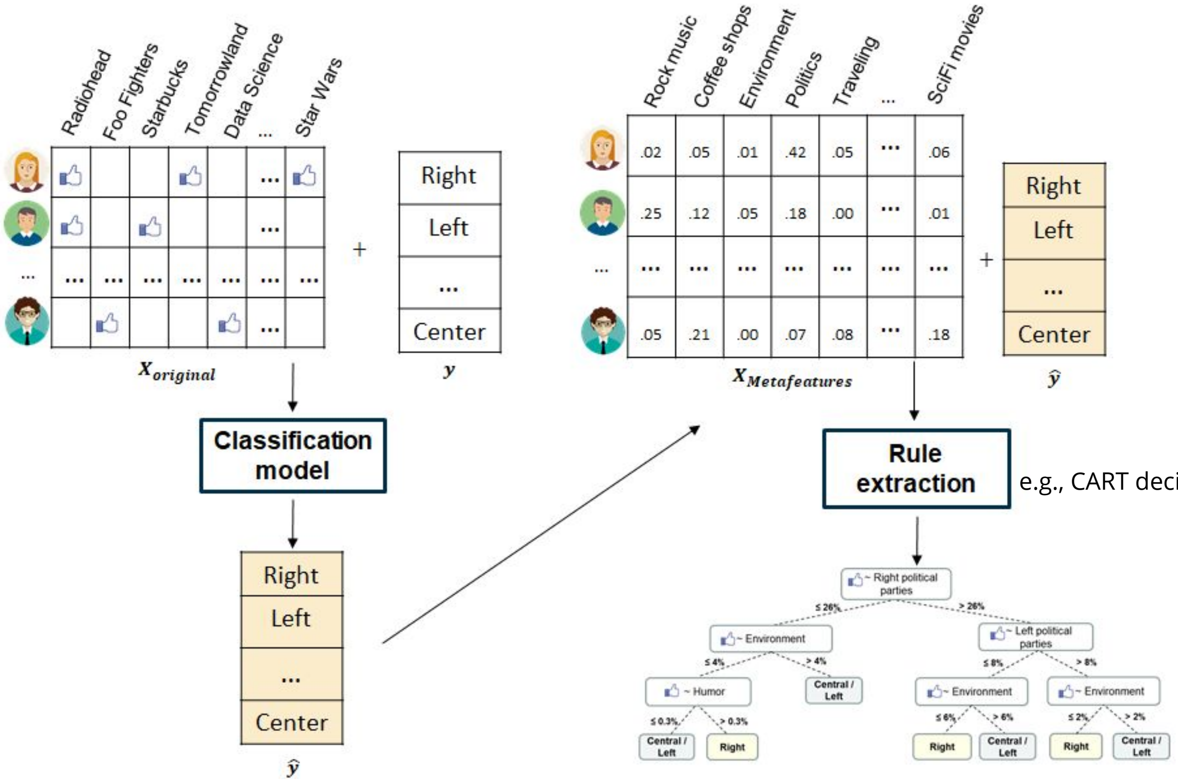
Contribution: Rule extraction with metafeatures

Original features	Metafeatures
Social media (e.g., Facebook likes)	Categories
Financial transactions (e.g., Carrefour)	Spending categories (e.g., Grocery stores)
Location data (e.g., Starbucks)	Venue types (e.g., Coffee shops)
Movie viewing data	Movie genres
Text data (e.g., Google searches)	Topics
Browsing behavior	Website categories

Contribution: Rule extraction with metafeatures

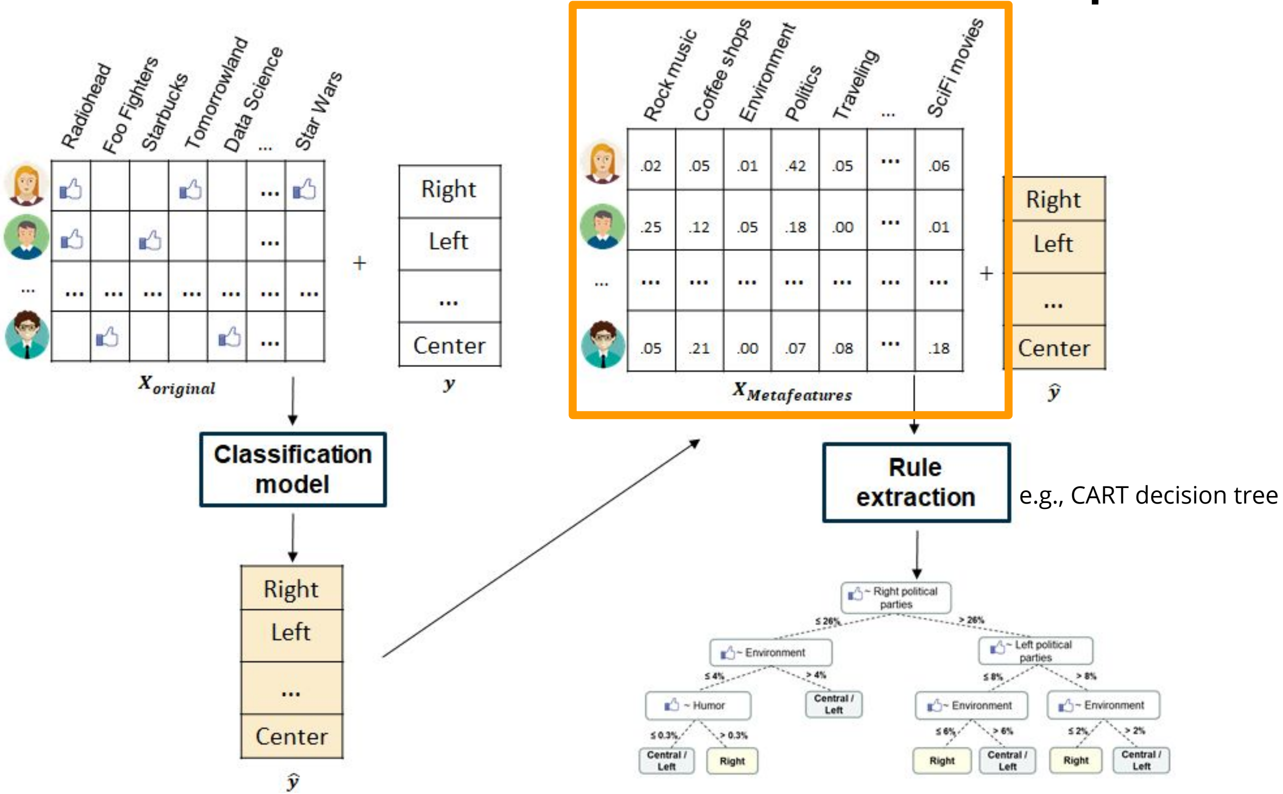


Contribution: Rule extraction with metafeatures

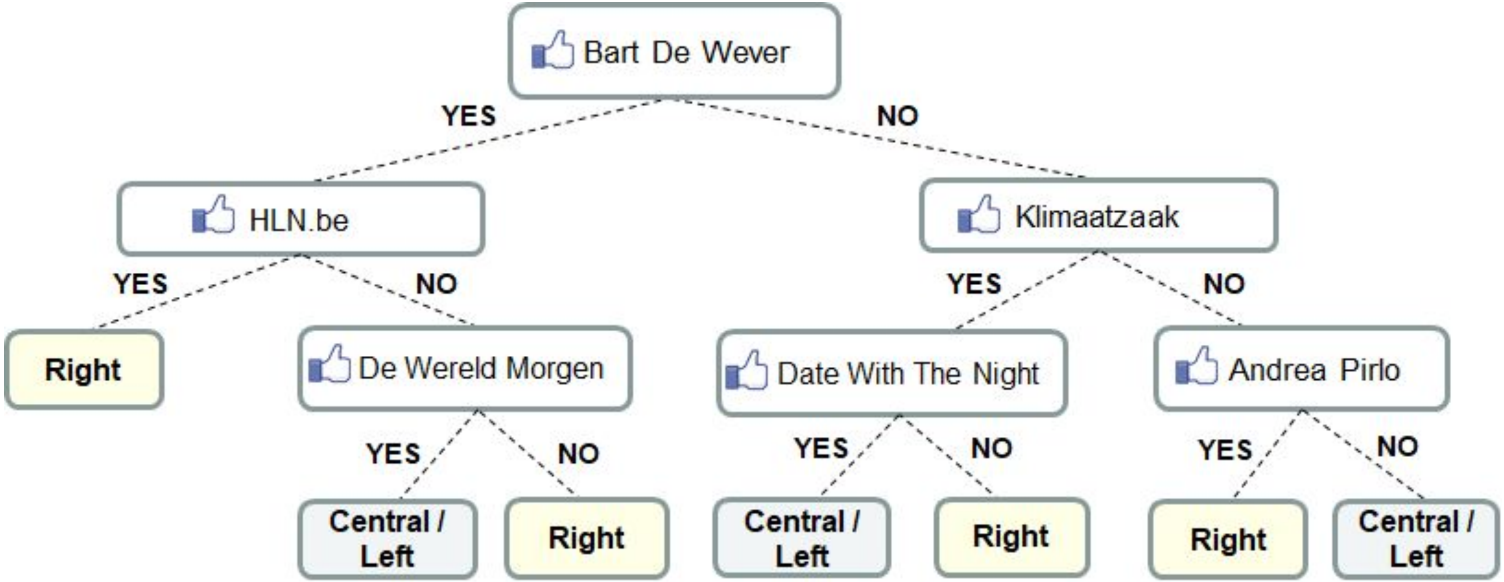


e.g., CART decision tree

Contribution: Rule extraction with metafeatures

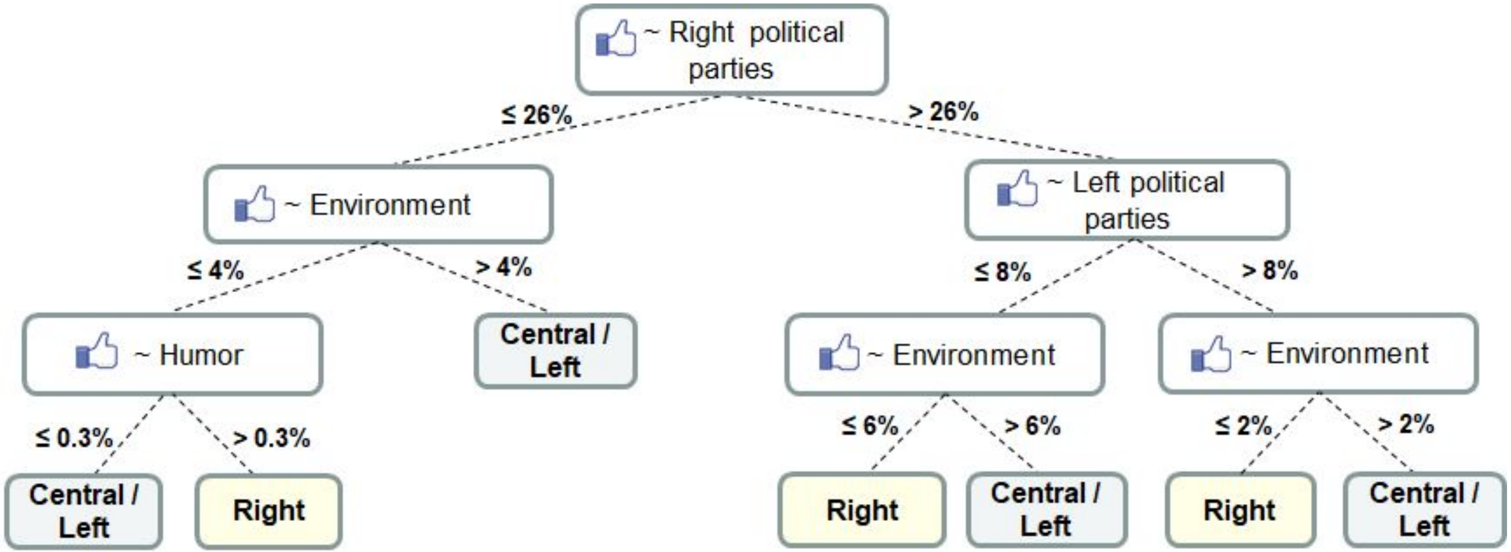


Example: Explanation rules with Facebook pages

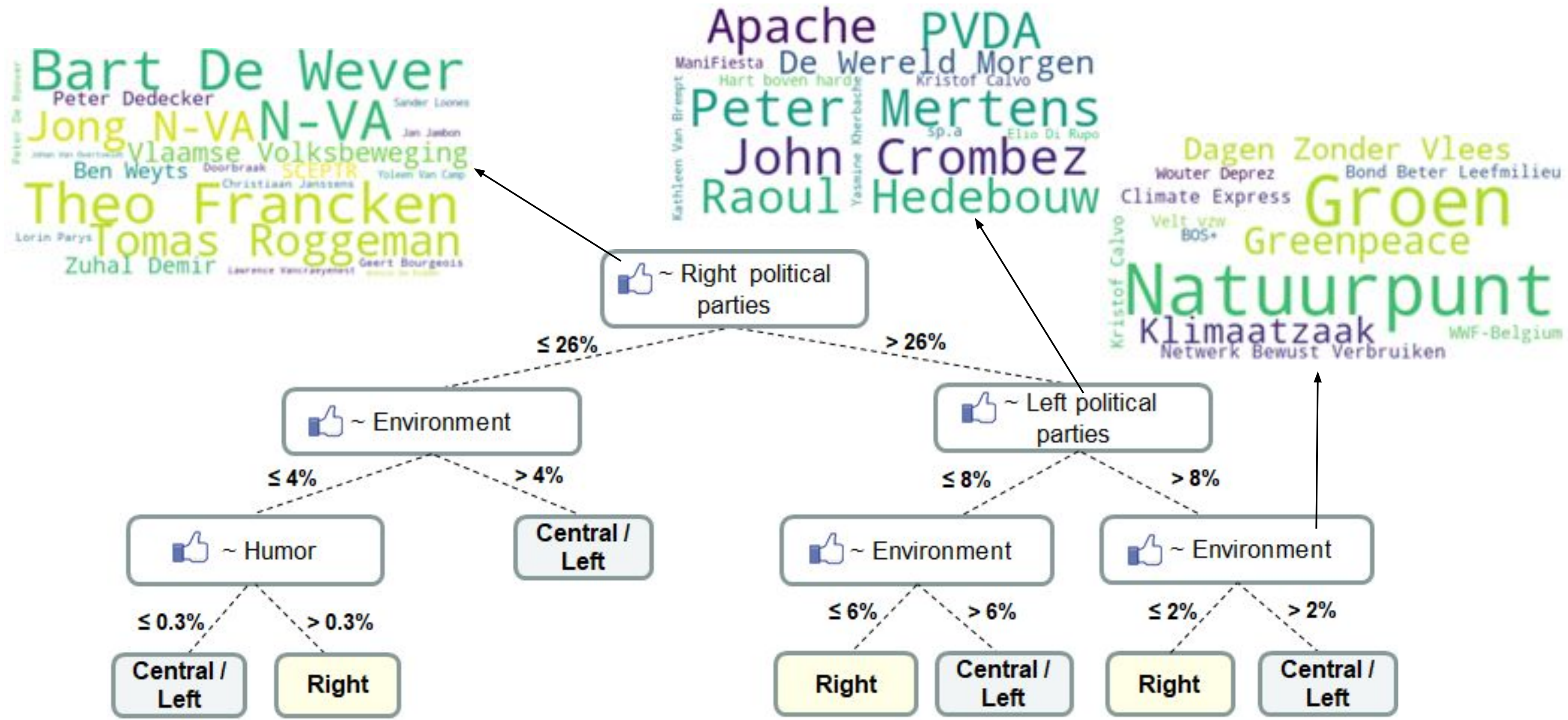


Note: Explanation rules for Logit model on *all* Facebook likes to predict political leaning.

Example: Explanation rules with metafeatures



Note: Explanation rules for Logit model on *all* Facebook likes. Data-driven metafeatures via non-negative matrix factorization (k=70).



Note: Explanation rules for Logit model on *all* Facebook likes. Data-driven metafeatures via non-negative matrix factorization (k=70).

Empirical question

Do explanation rules extracted with metafeatures result in better approximations of the model on behavioral data than explanations with the original features?

Empirical question

Do explanation rules extracted with metafeatures result in **better approximations** of the model on behavioral data than explanations with the original features?

Evaluation

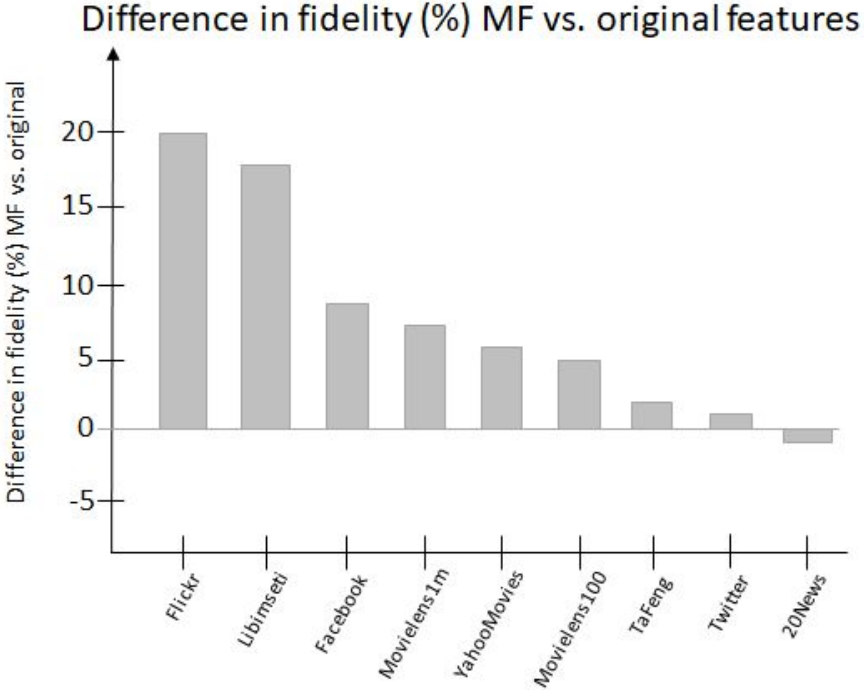
Fidelity → How well do explanation rules approximate the model?

Ground-truth labels	Predictions of model	Predictions of rules
1	1	1
1	0	0
0	1	1
1	1	0
...	...	...

Accuracy
of prediction model

Fidelity
of explanation rules

Results: fidelity ↑ when using metafeatures to explain



Note: Explanation rules for Logit on all features. Similar results for explanations of Random Forest model.

Implications

Implications

- **Validation and insight**

Implications

- Validation and insight
- **Ethics**

Amazon scraps secret AI recruiting tool that showed bias against women

Reference: Reuters, 2018

Hiring algorithms are being put to the test

Reference: MIT Technology Review, 2021

Algorithms drive online discrimination, academic warns

Reference: Financial Times, 2019

Implications

- Validation and insight
- Ethics
- **Improvement**

Implications

- Validation and insight
- Ethics
- **Improvement**

Example: Document classification

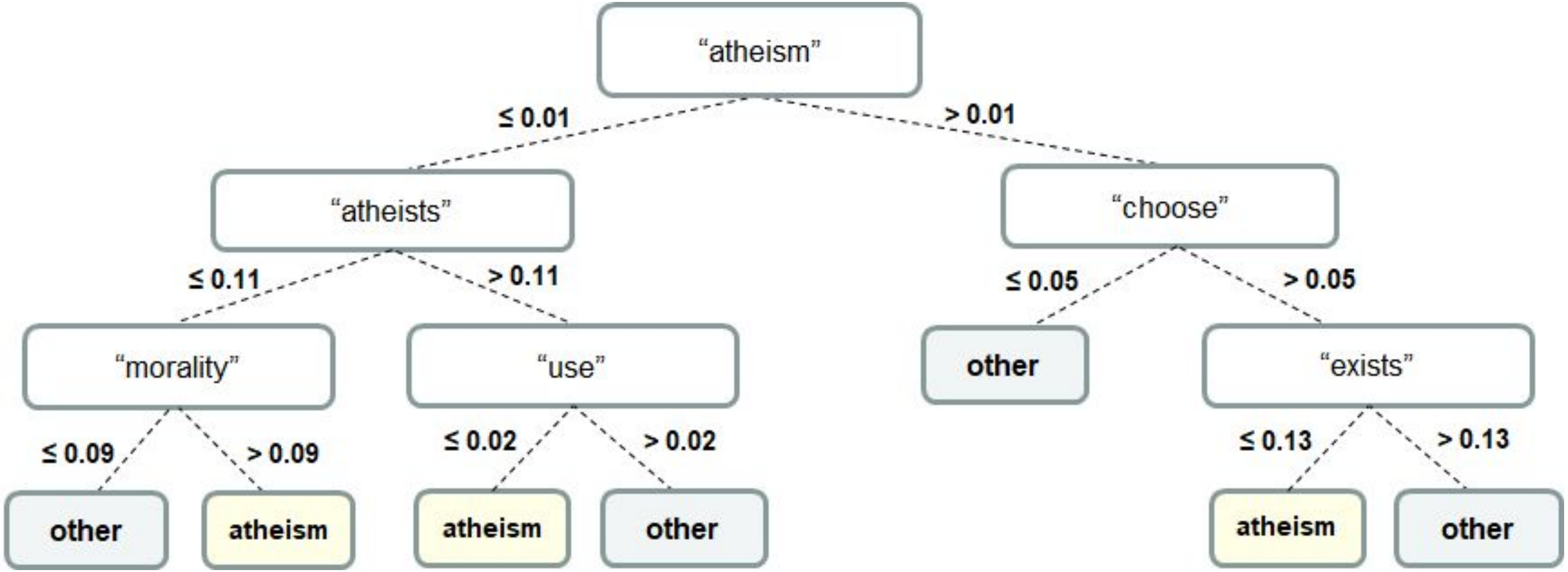
Dimension of the evidence pool: 37,685

Documents

Doc ID	"atheists"	"psilink"	"p00261"	...	"God"	# evidence present?	*Model decision: Atheism topic?*
...	...	...	...	...	...	...	...
01	1	1	1	...	1	28	*YES*
02	0	0	1	...	0	50	*NO*
03	1	0	0	...	1	98	*YES*
...	...	...	...	...	...	...	...

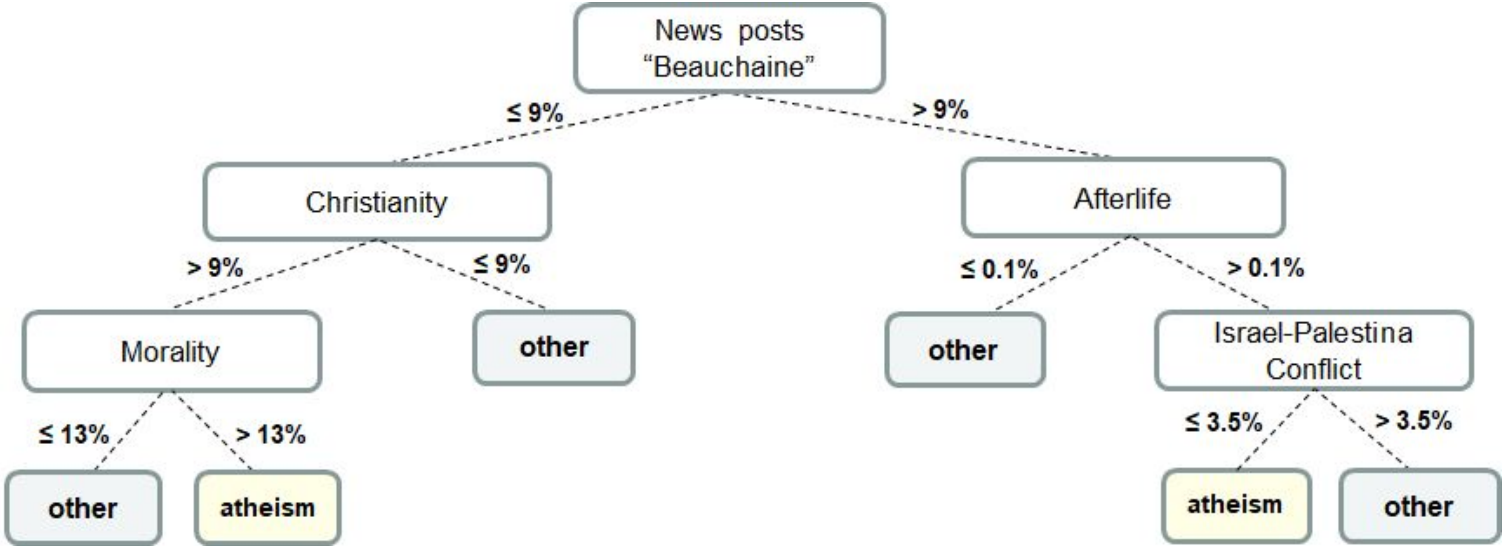
Figure 3a: 20 Newsgroups data to predict "Atheism" topic

Example: Explanation rules with words

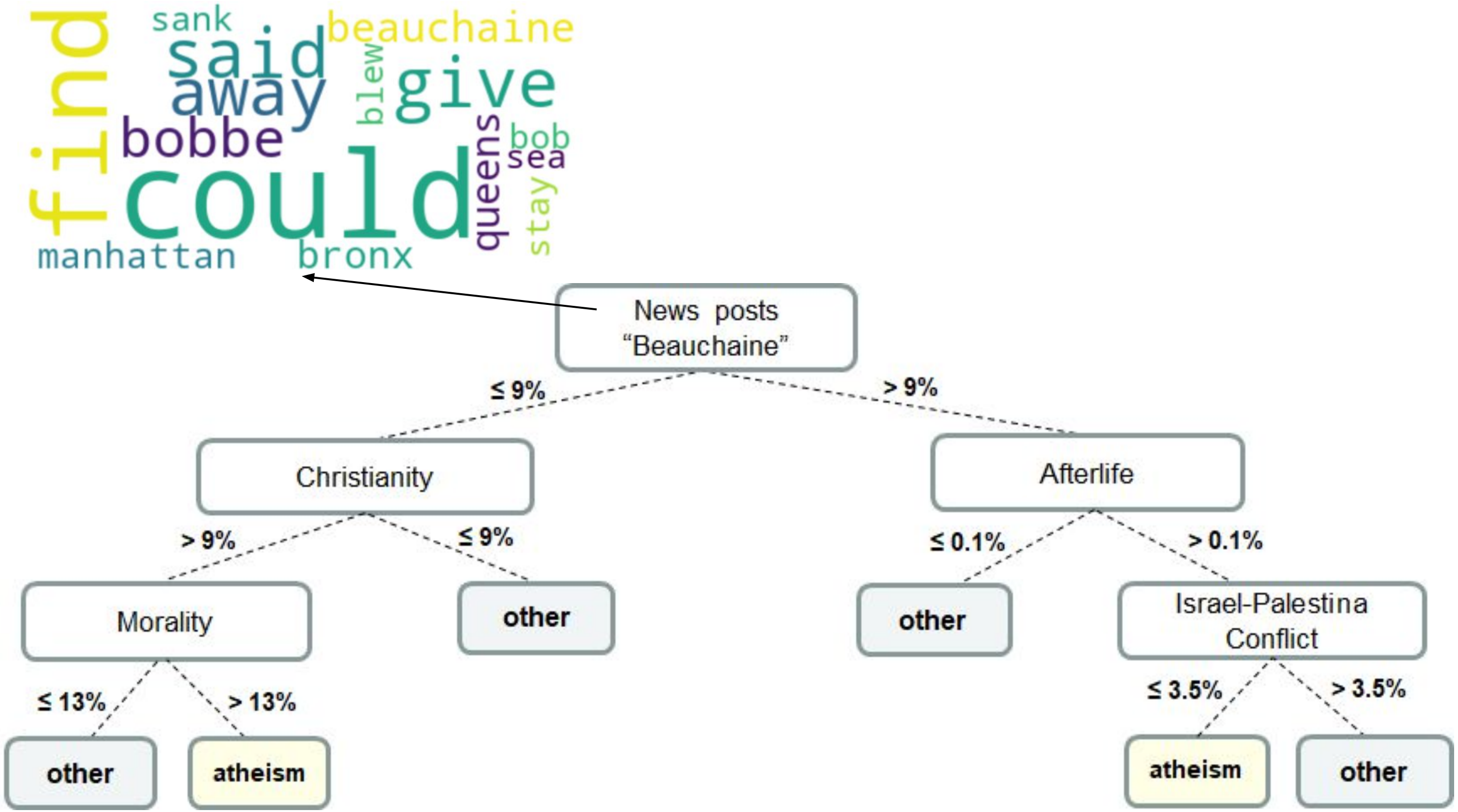


Note: Explanation rules for Logit model on *all* words in *20news* data (*tf-idf* representation) to predict topic "Atheism". *Reference: Ramon et al., 2020*

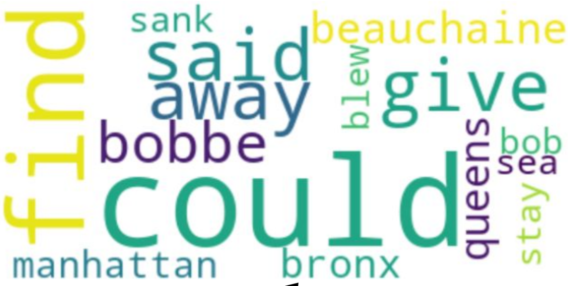
Example: Explanation rules with metafeatures



Note: Explanation rules for Logit model on *all* words. Data-driven metafeatures via non-negative matrix factorization (k=30). *Reference: Ramon et al., 2020*

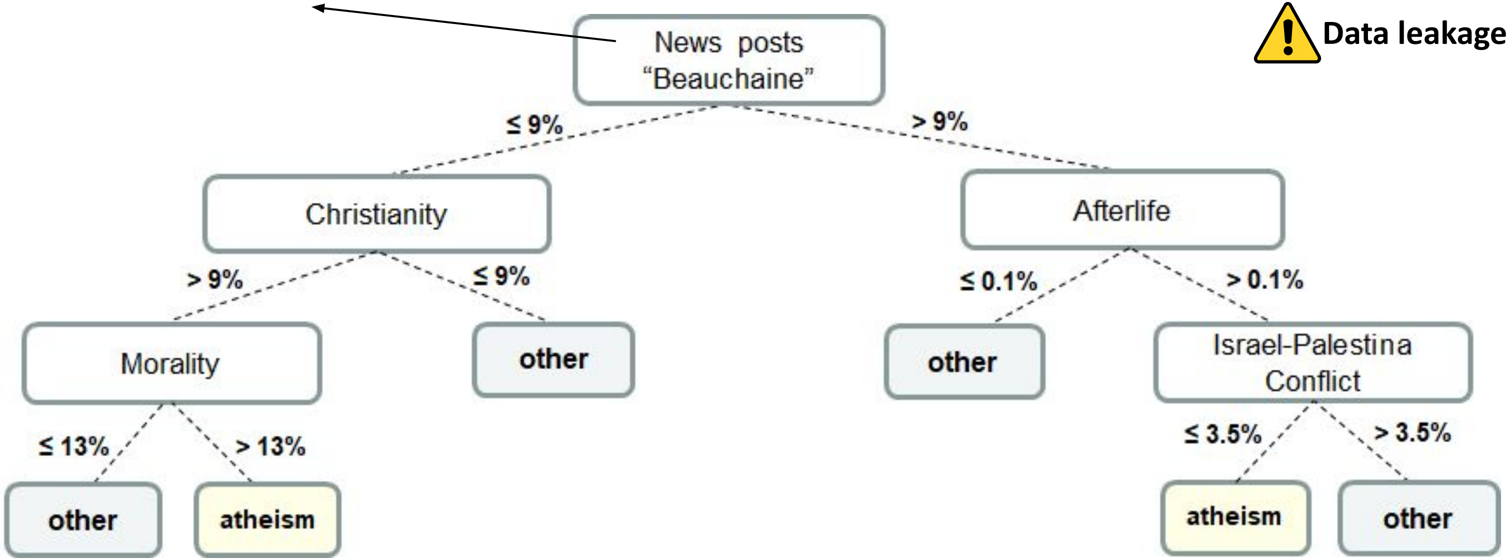


Note: Explanation rules for Logit model on *all* words. Data-driven metafeatures via non-negative matrix factorization (k=30). *Reference: Ramon et al., 2020*



'In the Old testement, Satan is RARELY mentioned, if at all.
Huh? Doesn't the SDA Bible contain the book of Job?
This is why there is suffering in the world, we are caught inthe crossfire.
(…)
Bob Beauchaine
They **said** that **Queens could stay**, they blew the **Bronx away**,
and **sank Manhattan** out at sea.
(…)'

Data leakage



Note: Explanation rules for Logit model on *all* words. Data-driven metafeatures via non-negative matrix factorization (k=30). *Reference:* [Ramon et al., 2020]

Key takeaways

To gain insight into prediction models on behavioral data

Key takeaways

To gain insight into prediction models on behavioral data

→ use higher-level, less-sparse “metafeatures” to explain the model

Key takeaways

To gain insight into prediction models on behavioral data

→ use higher-level, less-sparse “metafeatures” to explain the model

Why?

Key takeaways

To gain insight into prediction models on behavioral data

→ use higher-level, less-sparse "metafeatures" to explain the model

Why?

→ better approximation of model than explanations with original features

Key takeaways

To gain insight into prediction models on behavioral data

→ use higher-level, less-sparse “metafeatures” to explain the model

Why?

→ better approximation of model than explanations with original features

→ different types of information about model’s behavior

Thanks!

Yanou Ramon

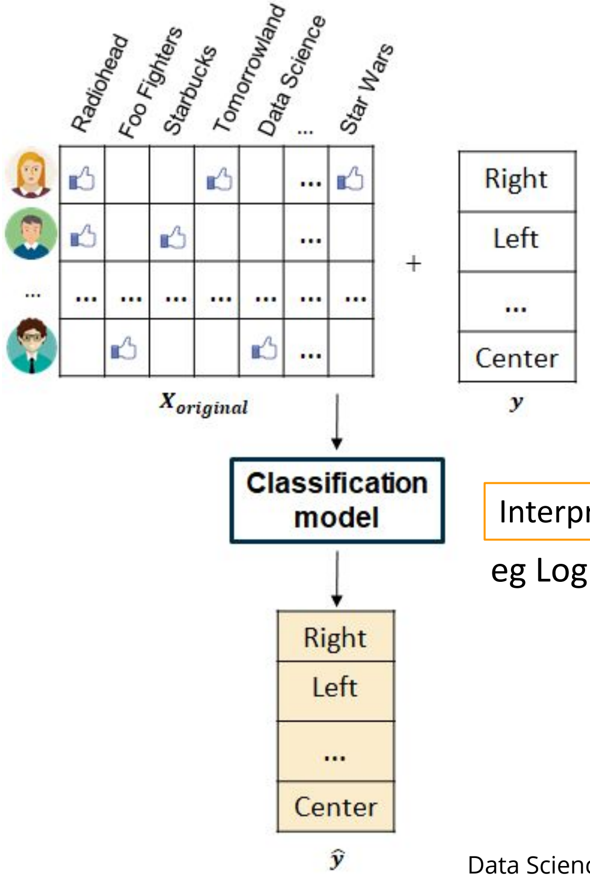
Prof. David Martens

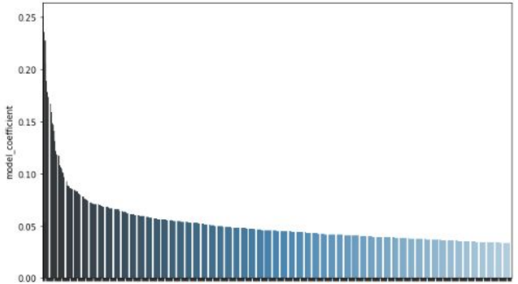
Prof. Theodoros Evgeniou

Stiene Praet

Ramon Y, Martens D, Evgeniou T, Praet S, Metafeatures-based rule extraction for classifiers on behavioral and textual data, 2020, preprint: https://arxiv.org/abs/2003.04792

Extra: Prediction models on behavioral data



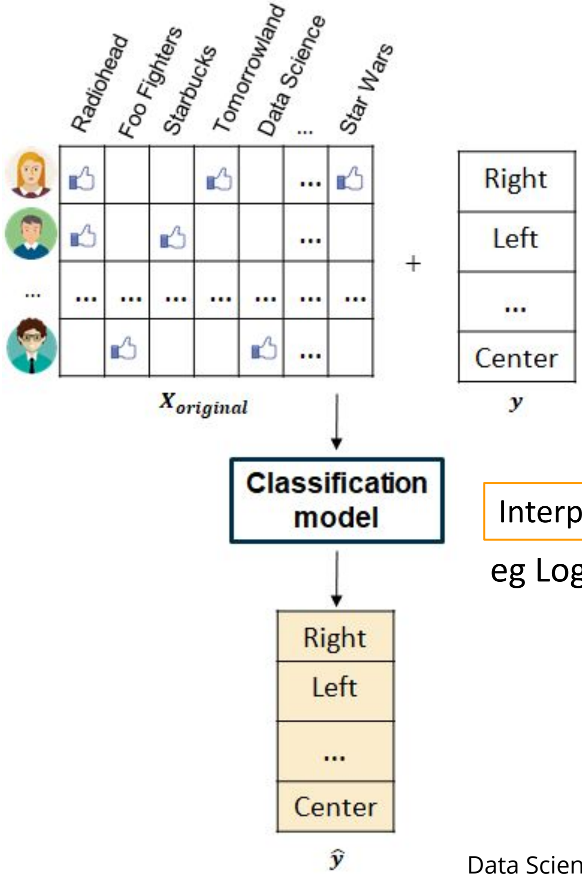


inspect coefficients

Interpretability issues!

eg Logistic Regression

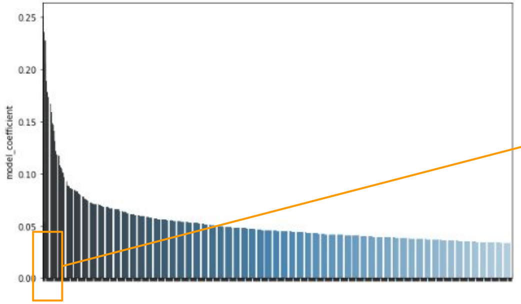
Extra: Prediction models on behavioral data



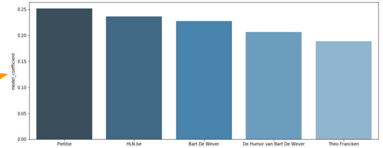
Interpretability issues!

eg Logistic Regression

inspect coefficients



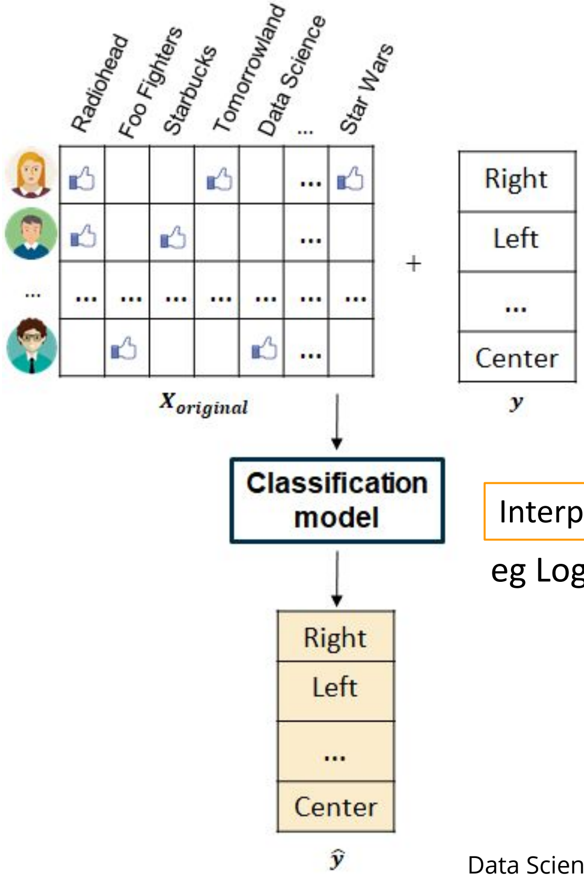
Very sparse data!



Facebook page	Coverage
“HLN.be”	21.49%
“Bart De Wever”	9.21%
“De humor van Bart De Wever”	7.41%
“Theo Francken”	9.02%
“Werkbond”	5.08%

avg coverage top-100 = 5.52%

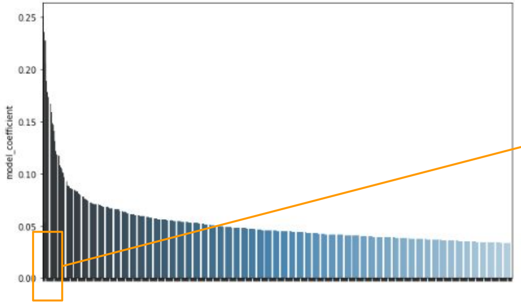
Extra: Prediction models on behavioral data



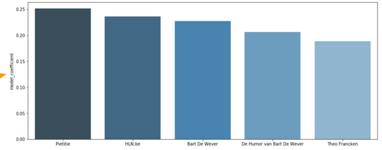
Interpretability issues!

eg Logistic Regression

inspect coefficients



Very sparse data!

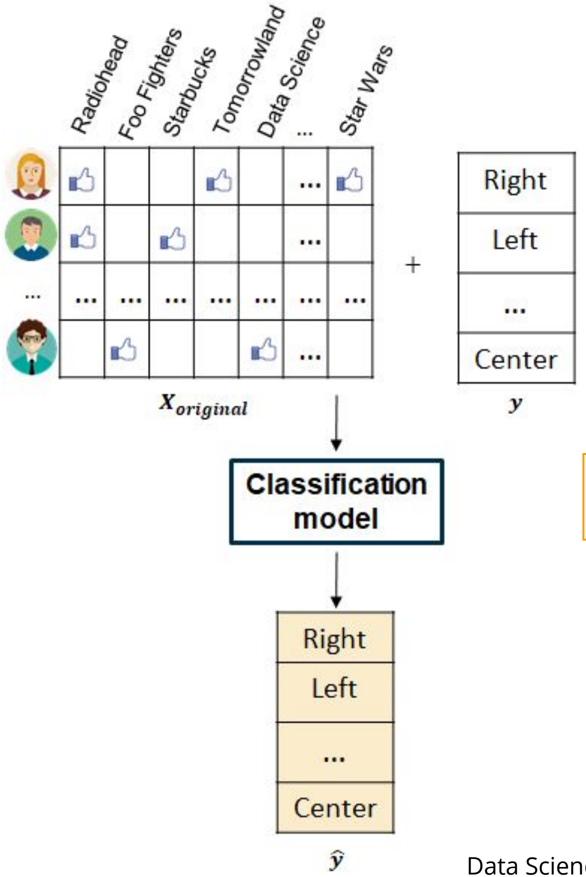


Facebook page	Coverage
"HLN.be"	21.49%
"Bart De Wever"	9.21%
"De humor van Bart De Wever"	7.41%
"Theo Francken"	9.02%
"Werkbond"	5.08%

avg coverage top-100 = 5.52%

Kosinski et al., 2013
user likes 0.3% of all pages

Extra: Prediction models on behavioral data



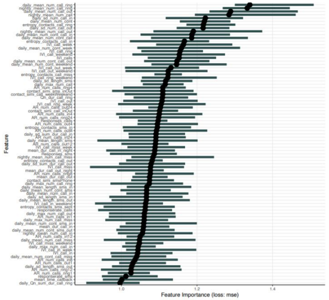
Dimensionality reduction → worse predictive performance
(e.g., [Junqué de Fortuny, 2015](), [Clark & Provost, 2019]())

Extra: Gaining insight into prediction models on behavioral data

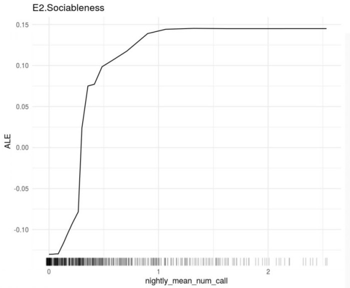
How?

Feature importance
(e.g., *Breiman, 2001*)

Linear model approximation
(e.g., *Ribeiro et al., 2016*)



Visual explanation
(e.g., *partial dependence plots*)



Rule extraction
(e.g., *Baesens et al., 2003*, *Martens et al., 2007*, *Guidotti et al., 2018*)