Counterfactual explanations for more transparency in AI

David Martens

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

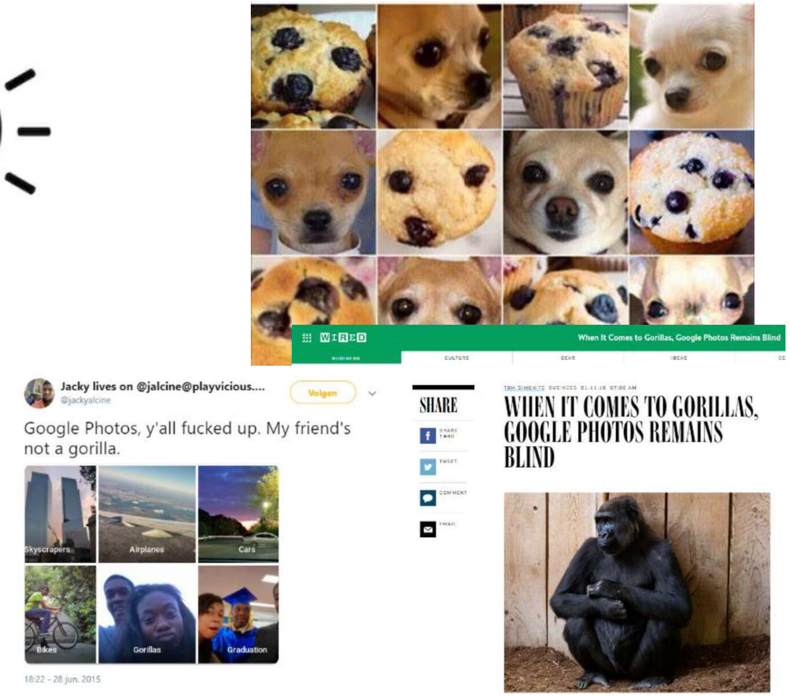
Overview

- The Need for Explanations
- Explainable AI
- The Counterfactual
- Open Issues

The Need for Explanations

- Why?

Trust	Insight	Improve

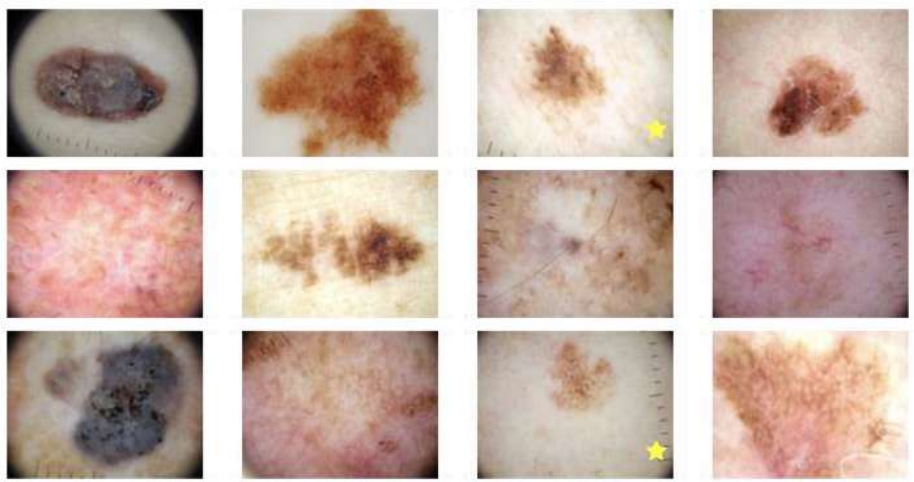


Trust

- Trust: *"Firm belief in the reliability, truth, or ability of someone or something."* (Oxford Dictionary)
  - Did the model learn the true pattern?
  - Is the model discriminating against sensitive groups?
- Test accuracy/AUC: already proxy, but issues:
  - In-lab versus real-life deployment
  - Summarizing performance in one number
- When users do not understand the workings, they will be skeptical and reluctant to use the model, even if the model is known to improve decision performance (Kayande et al, 2009)

Trust: lab-setting versus real-life

- Data: image of skin lesion
- Task: diagnose skin cancer
- High test accuracy, matching accuracy of 21 dermatologists, low accuracy when deployed in the field
- Issue: when dermatologist is concerned about lesion, a ruler is placed next to it in the picture
- Pattern learnt: if ruler then malignant



https://towardsdatascience.com/is-the-medias-reluctance-to-admit-ai-s-weaknesses-putting-us-at-risk-c355728e9028

Trust: lab-setting versus real-life

- Data: picture
- Task: predict if horse or not
- High test accuracy, low accuracy when deployed in the field
- Issue: horse pictres had watermark with copyright at bottom left
- Pattern learnt: if watermark then horse

https://arxiv.org/pdf/1902.10178.pdf

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

Trust

- Is the model not discriminating?

DHH @dhh · 9 nov. 2019
Als antwoord op @dhh
I wasn't even pessimistic to expect this outcome, but here we are: @AppleCard just gave my wife the VIP bump to match my credit limit, but continued to be an utter fucking failure of a customer service experience. Let me explain...

16 150 3K

DHH @dhh · 9 nov. 2019
She spoke to two Apple reps. Both very nice, courteous people representing an utterly broken and reprehensible system. The first person was like "I don't know why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

70 570 4,8K

CNN BUSINESS

Apple co-founder Steve Wozniak says Apple Card discriminated against his wife

By Clare Duffy, CNN Business
Updated 1615 GMT (0015 HKT) November 11, 2019

Trust

- Is the model not discriminating?
  (Cf. SyRI case, Calders and Van de Vijver, 2020)


REUTERS Business Markets World Politics TV More

TECHNOLOGY NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

Compliance

- Special case of trust
- In domains as credit scoring and medical diagnosis
- GDPR: Article 14.2.g: data subjects not only have the right to know that there is automated decision making, including profiling, but also that the data subject has then the right to obtain ***meaningful information about the logic involved***.
- EU's 2019 guidelines on ethics in AI: *"Another* ***great challenge*** *is to clarify how to implement the* ***requirement of explainability*** *in a context where the complexity of AI algorithms can make it difficult to provide a clear explanation and justification for a decision made by a machine (i.e. black box effect)."* (Madiega, 2019).
- Specifically for tax: see De Raedt, Martens and Brughmans (2021)

Insight

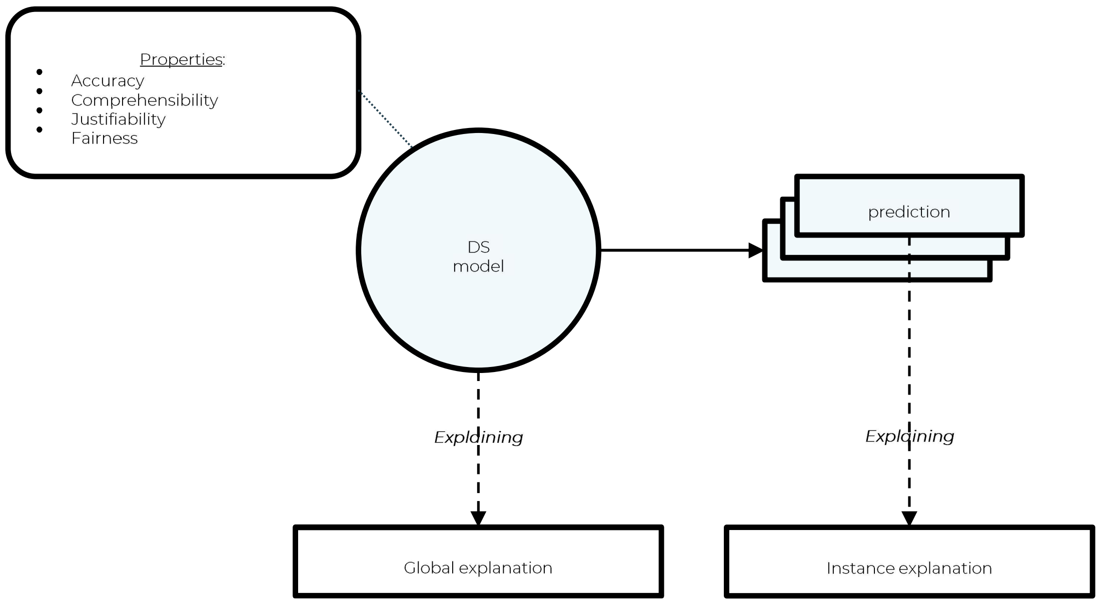
- What do we learn about
  - How the world works
  - How the model works

Overview

- The Need for Explanations
- **Explainable AI**
- The Counterfactual
- Open Issues

Comprehensible and Explaining

- *Comprehensibility*: a property of a model (and explanation)
  - Other terms used: interpretable, understandable, transparent, explainable, intelligible
- *Explaining*: an action

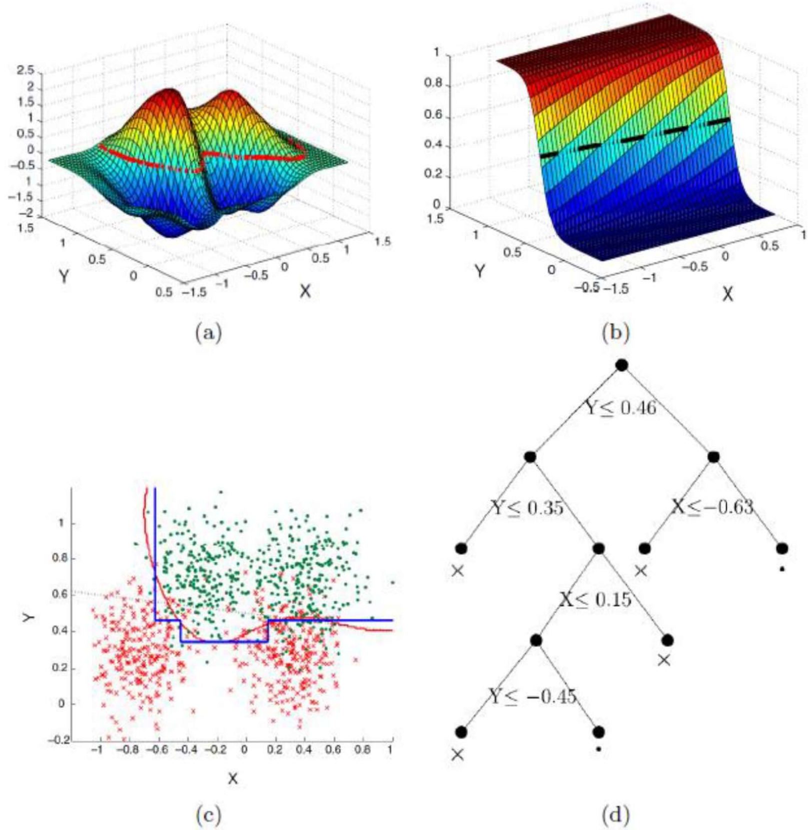


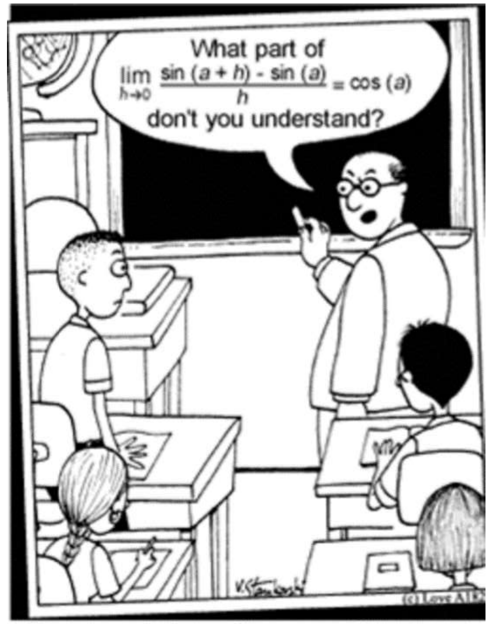
Comprehensible Models

- *"Each time one of our favorite ML [Machine Learning] approaches has been applied in industry, each time the comprehensibility of the results, though ill-defined, has been **a decisive factor of choice** over an approach by pure statistical means, or by neural networks."* Kodratoff (1994)
- What makes a model comprehensible?
  - Mainon and Rokach (2005): *"The comprehensibility criterion (also known as interpretability) refers to how well humans grasp the classifier induced. While the generalization error measures how the classier fits the data, comprehensibility measures the "**Mental Fit"** of that classifier."*
  - **Output type**: rule/tree-based > linear > non-linear
  - **Output size**: less > more (weights, nodes, rules, etc.)

Comprehensible Models

- What makes a model comprehensible
  - Output type: rule/tree-based > linear > non-linear
  - Output size: less > more (weights, nodes, rules, etc.)





DigiTax
Centre of Excellence

Explanations

- Global Explanations
  - Explain the model over the **complete dataset** as “good” as possible
  - “Good”: human understandable *and* high fidelity
  - Common approach: Rule extraction for non-linear models and top coefficients for linear models
  - Use: explain to manager before deploying

- Instance-based Explanations
  - Explain an **individual** prediction
  - Common approaches: LIME and Evidence Counterfactual
  - Use: explain to data scientist, data subject

Instance-based explanations

- Why instance-level
  - Often interested only in explanation for one data instance (customer/article/company/etc.)
  - Global models are too complex or limited fidelity

- Main approaches:
  - Feature Importance
    - LIME, SHAP: linear approximation with coefficients indicating feature importance
    - Input: instance + model + ***prediction score***
  - Counterfactuals
    - Minimal set of evidence present in the data instance, when removed, changes the decision
    - Input: instance + model + ***decision***

Overview

- The Need for Explanations
- Explainable AI
- **The Counterfactual**
- Open Issues

The Counterfactual

- Example: gender prediction using movie viewing data
  - Sam watched 120 movies
  - Sam is predicted as male
  - Why?
  - 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

Yanou Ramon, David Martens (2019) *Instance-based Explanations: Motivation, Overview, and the Evidence Counterfactual Approach*, European Conference on Data Analysis, Bayreuth, Germany, 18-20 March 2019

Explain individual predictions

EDC: EviDence Counterfactual

E is an explanation for $C_M(D) = c$

1. $E \subseteq W_D$ (the words are in the document),
2. $C_M(D \setminus E) \neq c$ (the class changes), and
3. $\nexists E' \subset E : C_M(D \setminus E') \neq c$ (E is minimal).

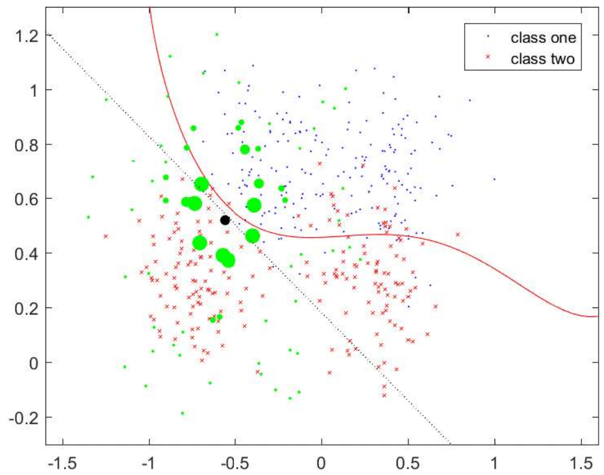
$D \setminus E$ denotes the result of removing the words in E from document D.

Martens D, Provost F. (2014) *Explaining Data-Driven Document Classifications.* MIS Quarterly 38(1):73-99.

```
Python: github.com/YanouRamon/edc
Matlab: www.applieddatamining.com/cms/?q=software
```

Ramon Y., Martens D., Provost F., Evgeniou T. (2021) *Instance-level explanation algorithms SEDC, LIME, SHAP for behavioral and textual data: a counterfactual-oriented comparison*, Machine Learning.

LIME: Linear Interpretable Model-Agnostic Explainer



M.T. Ribeiro, S. Singh, C. Guestrin (2016) *Model-Agnostic Interpretability of Machine Learning*
2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY

github.com/marcotcr/lime

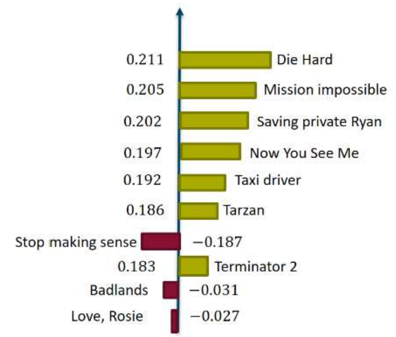
Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

Explain individual predictions

EDC: EviDence Counterfactual	LIME: Linear Interpretable Model-Agnostic Explainer

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



Issues:

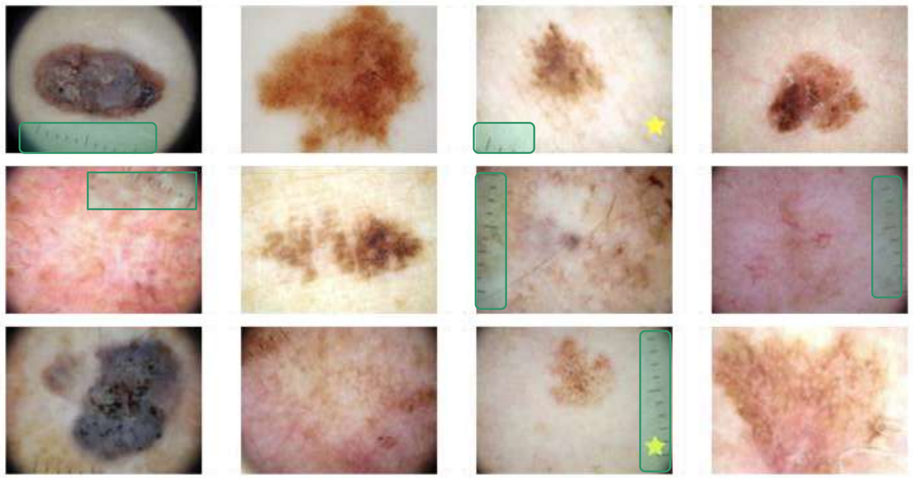
- What is proper value for k?
- How accurate is linear approximation?
- Stability: run twice, two different explanations
- Usefulness of negative evidence
- Rather slow
- Explains prediction score

The Counterfactual

- To provide insight
- Fictitious example: **fraud detection** using invoicing (listing) data
  - Company FraudACME transacted with 56 businesses
  - Predicted by black box model to be fraudulent
  - Why?
  - "if FraudACME would not have transacted with *CasinoOostende*, *KnownFraudsterUS* and *GarageDodgy* then the prediction would change to non-fraudulent"

The Counterfactual

- Image data



The Counterfactual

- To improve the predictive performance of the model
- Example
  - Data: image
  - Task: predict if missile in image

The Counterfactual

- To improve the predictive performance of the model
- Example
  - Data: image
  - Task: predict if missile in image
  - Mainly interested in improving misclassifications
  - Issue: Lighthouse wrongly classified as missile
  - Pattern learnt: line of smoke indicates missile

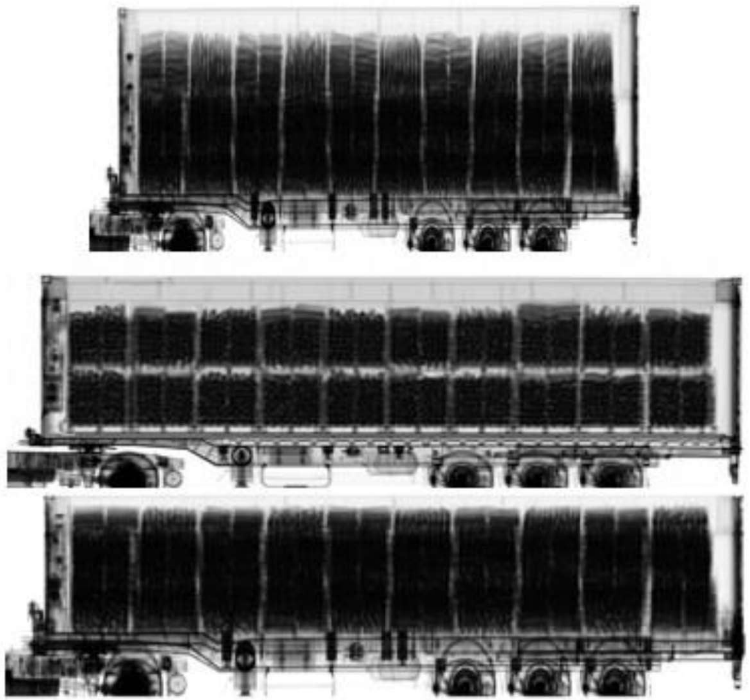
(a) Original
class: *missile*

(b) Counterfactual
explanation

(c) Counterfactual
class: *beacon*

Explainable Image Classication with Evidence Counterfactual
Tom Vermeire, David Martens (2020)

The Counterfactual



https://www.openaccessgovernment.org/automated-comparison-x-ray-images-cargo-scanning-acxis/25162/

The Counterfactual

- Example
  - ECB communication
  - Task: predict if market IR go up or down, based on communication
  - Predicted to lead to hawkish response from market (IR go up)
  - Why?

Figure 8.11: Extract from the introductory statement of 4 July 2002. The media perceived the statement as hawkish. On 5 December 2002, the ECB started a 6-months period of monetary policy easing. The words that should be removed from this extract in order to change the predicted perception to dovish are indicated in bold.

Turning to price developments, Eurostat's flash estimate indicates that annual HICP inflation fell from 2.0% in May to 1.7% in June. However, it is too early to **interpret** this fall as a sign of receding **upward pressure** on prices, given that HICP inflation excluding the more volatile items of energy and unprocessed **food prices** has remained high throughout the first half of this year, reflecting in particular trends in services prices. Moreover, it is to be expected that overall HICP inflation rates will fluctuate around 2% in the coming months. Overall, the strengthening of the euro exchange rate is a new factor suggesting a potential for lower inflation rates. However, other factors - in particular monetary developments and wage **trends** - do not indicate a moderation in price pressures. Monetary policy therefore needs to remain vigilant as regards the key factors determining the outlook for price stability over the medium term.

Tobback Ellen, Nardelli Stefano, Martens David (2017)
Between hawks and doves: measuring central bank ommunication
European Central Bank,39 p.(Working paper)

Overview

- The Need for Explanations
- Explainable AI
- The Counterfactual
- **Open Issues**

Counterfactual Generating Algorithms

- Started with SEDC in 2014 (Martens and Provost, 2014)
- Over 60 (!) up till 2020
- Benchmarking study
- Mazzine and Martens (2021): "it depends"

Algorithm	Tools	Access
(2014.03) SEDC [129]	heuristic	query
(2015.08) OAE [51]	ILP	white-box
(2016.05) HCLS [110, 112]	grad opt/heuristic	gradient/query
(2017.06) Feature Tweaking [186]	heuristic	white-box
(2017.11) CF Expl. [196]	grad opt	gradient
(2017.12) Growing Spheres [114]	heuristic	query
(2018.02) CEM [55]	FISTA	class prob.
(2018.02) POLARIS [209]	heuristic	gradient
(2018.05) LORE [80]	gen alg + heuristic	query
(2018.06) Local Foil Trees [190]	heuristic	query
(2018.09) Actionable Recourse [189]	ILP	white-box
(2018.11) Weighted CFs [77]	heuristic	query
(2019.01) Efficient Search [175]	MILP	white-box
(2019.04) CF Visual Expl. [76]	greedy search	white-box
(2019.05) MACE [99]	SAT	white-box
(2019.05) DiCE [145]	grad opt	gradient
(2019.05) CERTIFAI [179]	gen alg	query
(2019.06) MACEM [56]	FISTA	query
(2019.06) Expl. using SHAP [165]	heuristic	query
(2019.07) Nearest Observable [201]	brute force	dataset
(2019.07) Guided Prototypes [191]	grad opt/FISTA	gradient/query
(2019.07) REVISE [95]	grad opt	gradient
(2019.08) CLEAR [202]	heuristic	query
(2019.08) MC-BRP [123]	heuristic	query
(2019.09) FACE [162]	graph + heuristic	query
(2019.09) Equalizing Recourse [83]	ILP/heuristic	white-box/query
(2019.10) Action Sequences [163]	program synthesis	class prob.
(2019.10) C-CHVAE [156]	grad opt + heuristic	query + gradient
(2019.11) FOCUS [124]	grad opt + heuristic	white-box
(2019.12) Model-based CFs [127]	grad opt	gradient
(2019.12) LIME-C/SHAP-C [164]	heuristic	query
(2019.12) EMAP [41]	grad opt	dataset/query
(2019.12) PRINCE [71]	graph + heuristic	query
(2019.12) LowProFool [18]	grad opt	gradient
(2020.01) ABELE [79]	gen alg + heuristic	query + data
(2020.01) SHAP-based CFs [66]	heuristic	query
(2020.02) CEML [11–13]	grad opt/heuristic	gradient/query
(2020.02) MINT [100]	SAT	white-box
(2020.03) ViCE [74]	heuristic	query
(2020.03) Plausible CFs [22]	grad opt + gen alg	dataset
(2020.04) SEDC-T [193]	heuristic	query
(2020.04) MOC [52]	gen alg	query
(2020.04) SCOUT [199]	grad opt	gradient
(2020.04) ASP-based CFs [28]	answer-set prog.	query
(2020.05) CBR-based CFs [103]	heuristic	query + data
(2020.06) Survival Model CFs [106]	gen alg	query
(2020.06) Probabilistic Recourse [101]	grad opt/brute force	gradient/query
(2020.06) C-CHVAE [155]	grad opt	gradient
(2020.07) FRACE [210]	grad opt	gradient
(2020.07) DACE [96]	MILP	white-box
(2020.07) CRUDS [60]	grad opt	gradient/data
(2020.07) Gradient Boosted CFs [5]	heuristic	data
(2020.08) Gradual Construction [97]	heuristic	class prob.
(2020.08) DECE [44]	grad opt	gradient
(2020.08) Time Series CFs [16]	heuristic	query
(2020.08) PermuteAttack [87]	gen alg	query
(2020.10) Fair Causal Recourse [195]	grad opt/brute force	gradient/query
(2020.10) Recourse Summaries [167]	itemset mining alg	query
(2020.10) Strategic Recourse [43]	Nelder-Mead	query
(2020.11) PARE [172]	grad opt + heuristic	query

Karimi et al (2021) A survey of algorithmic recourse: contrastive explanations and consequential recommendations https://arxiv.org/pdf/2010.04050.pdf

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

What do Users Want?

Based on his browsing activity,
Sam was shown the following ad:

If you were Sam, which of the explanations that explain why Sam is seeing this advertisement, would you prefer?

IF you would not have visited *{uantwerpen.be, student.be, scholar.google.com}* **THEN** this ad would not be shown to you

(a)

(b)

Ramon Yanou, Vermeire Tom, Toubia Olivier, Martens David, Evgeniou Theodoros (2021) Understanding consumer preferences for explanations generated by XAI algorithms
https://arxiv.org/abs/2107.02624

What do Users Want?

- Own study looks at impact of format, complexity, specificity and users' cognitive styles for advertising and credit scoring
  - Preference for feature importance methods
  - If negative outcome: preference for counterfactuals
  - How specific explanation should be depends on user's cognitive style

Ramon Yanou, Vermeire Tom, Toubia Olivier, Martens David, Evgeniou Theodoros (2021) Understanding consumer preferences for explanations generated by XAI algorithms
https://arxiv.org/abs/2107.02624

What Do Users Want?

- Tax fraud detection, many roles
  - **Data Scientist**: improve, trust or insight into model?
  - **Manager**: global model?
  - **Selection Officer**: agree or overrule?
  - **Investigator**: need explanation? Keep him/her sharp vs efficient
  - **End Users**: business or person right to an explanation?
    Don’t rock the boat vs GDPR/transparency

➔ Surely different cognitive styles,
different view of negative outcome

Advantages

- Advantages of **instance-based explanations** LIME/SHAP/EDC
  1. No limitation on complexity of BB model
  2. Avoids disclosing the model
  3. Automate task of generating explanations

- Additional advantages of **the Counterfactual**
  1. **Concrete justification** for a decision decision-making model
     - provide grounds to ***contest*** adverse decisions, and
     - to understand what could be ***changed*** to receive a desired result in the future, based on the current
  2. Comply with **GDPR** requirements on this matter
     - "meaningful information about the logic involved" GDPR 13.2 (f)
     - "The controller should find simple ways to tell the data subject about the **rationale behind**, or the criteria relied on in **reaching the decision**. … The information provided should be sufficiently comprehensive for the data subject to understand the reasons for the decision." (Working Party 29, 2018).
  3. Stable, Fast, No assumptions

Barocas et al (2020), Wachter et al (2018)

Open Issues

- Challenges of instance-based explanations LIME/SHAP/EDC
  1. Which explanation method to use?
  2. How to choose among explanations: new power to businesses, moral hazard
  3. Should all features be part of an explanation (e.g. gender, marital status)
  4. What explanations do users want?
  5. Who should get access to explanations?

Open Issues

- Challenges of instance-based explanations LIME/SHAP/EDC ***for fraud detection***
  1. Which explanation method to use?
  2. How to choose among explanations
     New power to tax administration
  3. Should all features be part of an explanation
     (e.g. secret or actionable features)
  4. What explanations do users want?
     Different goals and cognitive ability
  5. Who should get access to explanations?
     Companies, investigators, selection officers?

Ongoing research in collaboration with Belgian customs
(work by Dieter Brughmans)

Conclusion

- Ability to understand a model and prediction are important
  - To **trust**: key driver for acceptance of model
  - To obtain **insight**: what can we learn from the world and model
    (for example spot new fraud pattern or drive investigation)
  - To **improve** the model: wrong labels or wrong patterns learnt
- Explanations to various **roles**
  - Manager: global explanations
  - Data scientist: global and instance-level explanations
  - End user (customer or business): instance-level explanations
- Explainable AI
  - has become **focus point in recent research** and **applications**, still much work to do
  - Just one part of Ethical Data Science
    (FAT: Fair, Accountable, Transparent)
- Consider explainability when using predictive model


Q&A
OXFORD
Data Science Ethics
Concepts, Techniques and Cautionary Tales
DSE TECHNIQUES
UTILITY
ETHICS
ETHICS
ETHICS
DAVID MARTENS
desethics.com