# *What If?*
# Demystifying AI Decisions with Counterfactuals

David Martens

admantwerp.github.io

Universiteit Antwerpen

# Advanced AI in our lives


DHH
@dhh

Volgen

The @AppleCard is such a fucking sexist prog
retur
state
time.
think
does

12:34 - 7

9.664 ret

1.583

why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

69 542 4.253

**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

CVS just delivered its first prescriptions via drone

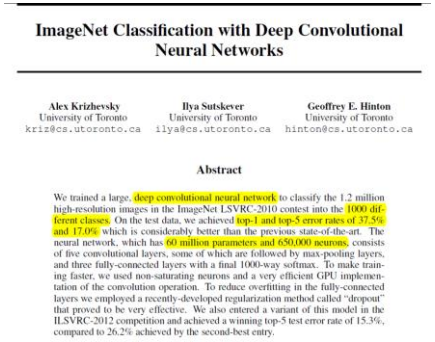
Financial Services is looking into
e Apple Card, which is administered by

edit card accounts or any
n Twitter, in reply to

rrection though. It's big tech

# Black Box?

- Deep learning: large artificial neural network with massive number of parameters
  - MobileNetV2: 4 million parameters
  - GPT-4: estimated 100 trillion parameters

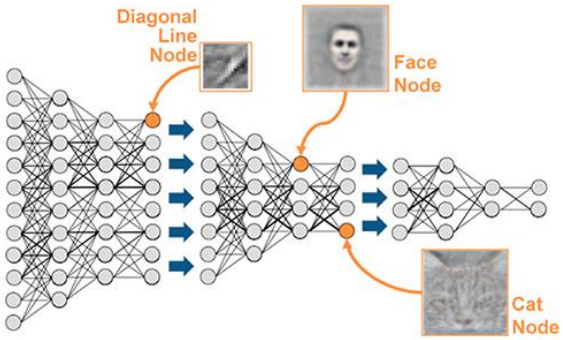


**ImageNet Classification with Deep Convolutional Neural Networks**

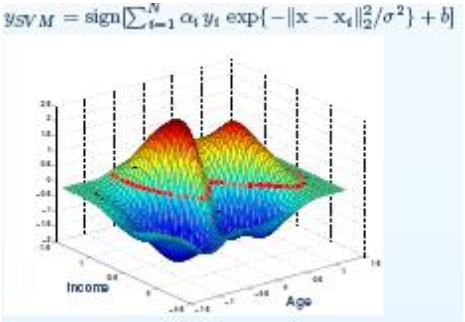
Alex Krizhevsky, University of Toronto, krizz@cs.utoronto.ca
Ilya Sutskever, University of Toronto, ilya@cs.utoronto.ca
Geoffrey E. Hinton, University of Toronto, hinton@cs.utoronto.ca

**Abstract**

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.




- Non-linear models



Complex AI models are creeping into our lives

Whose decisions are extremely difficult to explain

**But Explanations are needed if we want to trust AI**

# Explain individual predictions

**Counterfactual explanation:**

*What needs to change in your data, to reach a different decision?*

# Explain individual predictions

**Example**: credit scoring using sociodemo and financial data

User $x_i$: Sam, with income $ 32,000 and 39 years old.

Sam is denied credit

**WHY?**

**Counterfactual explanation:**

*What needs to change in your data, to reach a different decision?*

# Explain individual predictions

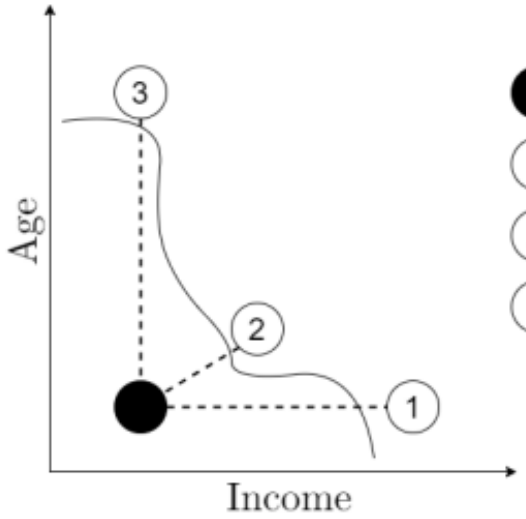
**Example**: credit scoring using sociodemo and financial data

User $x_i$: Sam

Sam is denied credit

**WHY?**

**IF** Sam would make 8,000$ more
**THEN** his predicted class would change from ***denied*** to ***granted***




| | Income | Age |
|---|---|---|
| ● | $ 32,000 | 39 |
| 1 | $ 40,000 | 39 |
| 2 | $ 36,000 | 44 |
| 3 | $ 32,000 | 140 |

Dieter Brughmans, Pieter Leyman, David Martens (2023) NICE : an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery* p. 1-39

# Explain individual predictions

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

User $x_i$: Sam

Sam watched 120 movies
Sam is predicted as male

**WHY?**

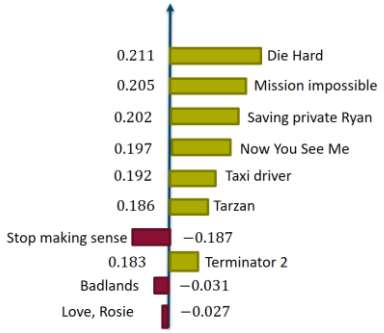
Counterfactual Explanation

**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

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

admantwerp.github.io

LIME: **L**inear **I**nterpretable **M**odel-Agnostic **E**xplainer (k=10)




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

Ramon Y., Martens D., Provost F., Evgeniou T. (2019) Instance-level explanation algorithms SEDC, LIME, SHAP for behavioral and textual data: a counterfactual-oriented comparison, *Advances in Data Analysis and Classification* 14:4, p. 801-819.

Complex AI models are creeping into our lives

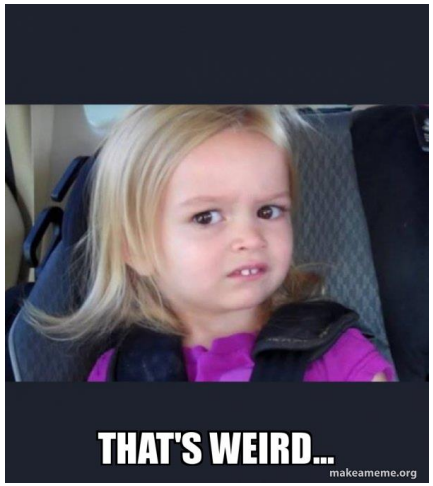
Whose decisions are extremely difficult to explain

But explanations are needed if we want to trust AI

**Counterfactuals explain a decision of a model for an instance**

# Counterfactual generating algorithms

- Yet more algorithms to explain complex algorithms




- **Ill-defined**, measures: sparse/short, near, diverse, actionable, feasible, plausible, justified, fast, model-agnostic, data-agnostic, truthful, complete, stable, robust, …

200+ methods

| Algorithm | Formulation | | | | | | | | | | | | | | Solution | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model | | | | Actionability | | Plausibility | | | Extra | | Data types | | | Tools | Access | Properties | | | Code |
| | TB | KB | DF | OT | uncond. | cond. | dom. | dens. | proto. | diver. | spar. | ▦ | 📷 | 📄 | | | opt. | cov. | rtm. | |
| (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

Complex AI models are creeping into our lives

Whose decisions are extremely difficult to explain

But Explanations are needed if we want to trust AI

Counterfactuals explain a decision of a model for an instance

**What these should look like is ill-defined and open research**

# The Counterfactual Explains **Wrong** Decisions

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

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine de Oliveira, David Martens (2022)
Explainable Image Classication with Evidence Counterfactual, Pattern Anal Applic 25, 315–335

# The Counterfactual Explains **Wrong** Decisions

- 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

**(a)** Original
class: *missile*

# The Counterfactual Explains **Wrong** Decisions

- 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, resembling exhaust plume behind missiles, indicates missile

**(a)** Original class: *missile*

**(b)** Counterfactual explanation

**(c)** Counterfactual class: *beacon*

Complex AI models are creeping into our lives

Whose decisions are extremely difficult to explain

But Explanations are needed if we want to trust AI

Counterfactuals explain a decision of a model for an instance

What these should look like is ill-defined and open research

**Very useful to explain (incorrect or unfair) decisions**

# Imagine a world with CF explanations


**Apple Card** @AppleCard

Well, if your wife would also have had a 20+ year relationship with our bank, and would have been regarded as Premium customer at some point in time, she would also receive a 20x credit limit

**Apple Card** @AppleCard

Well, if your wife's relationship status would have been "husband" instead of "wife", she would also receive a 20x credit limit

We clearly messed up, we're updating our models now.

**DHH** @dhh

Ah, ok, thanks for the additional feedback!

**Steve Wozniak** @stevewoz

Glad you found this and react responsibly.
It's how big tech should be in the 21st century.

# Q&A

- Play around with CF expanations: admantwerp.github.io

| Data | Algorithm | Reference |
|---|---|---|
| TEXT | SEDC | Martens and Provost (2014) |
| | LIME-C, SHAP-C | Ramon et al (2020) |
| BEHAVIORAL | SEDC | Martens and Provost (2014) |
| | LIME-C, SHAP-C | Ramon et al (2020) |
| IMAGE | SEDC(-T) | Vermeire and Martens (2020) |
| STRUCTURED | NICE | Brughmans and Martens (2021) |

- Book on Data Science Ethics: www.dsethics.com

David Martens (2022)
*Data Science Ethics: Concepts, Techniques and Cautionary Tales*
Oxford University Press, 272 pages.