Counterfactual Explanations Using Optimization with Constraint Learning

Ilker Birbil

UvA

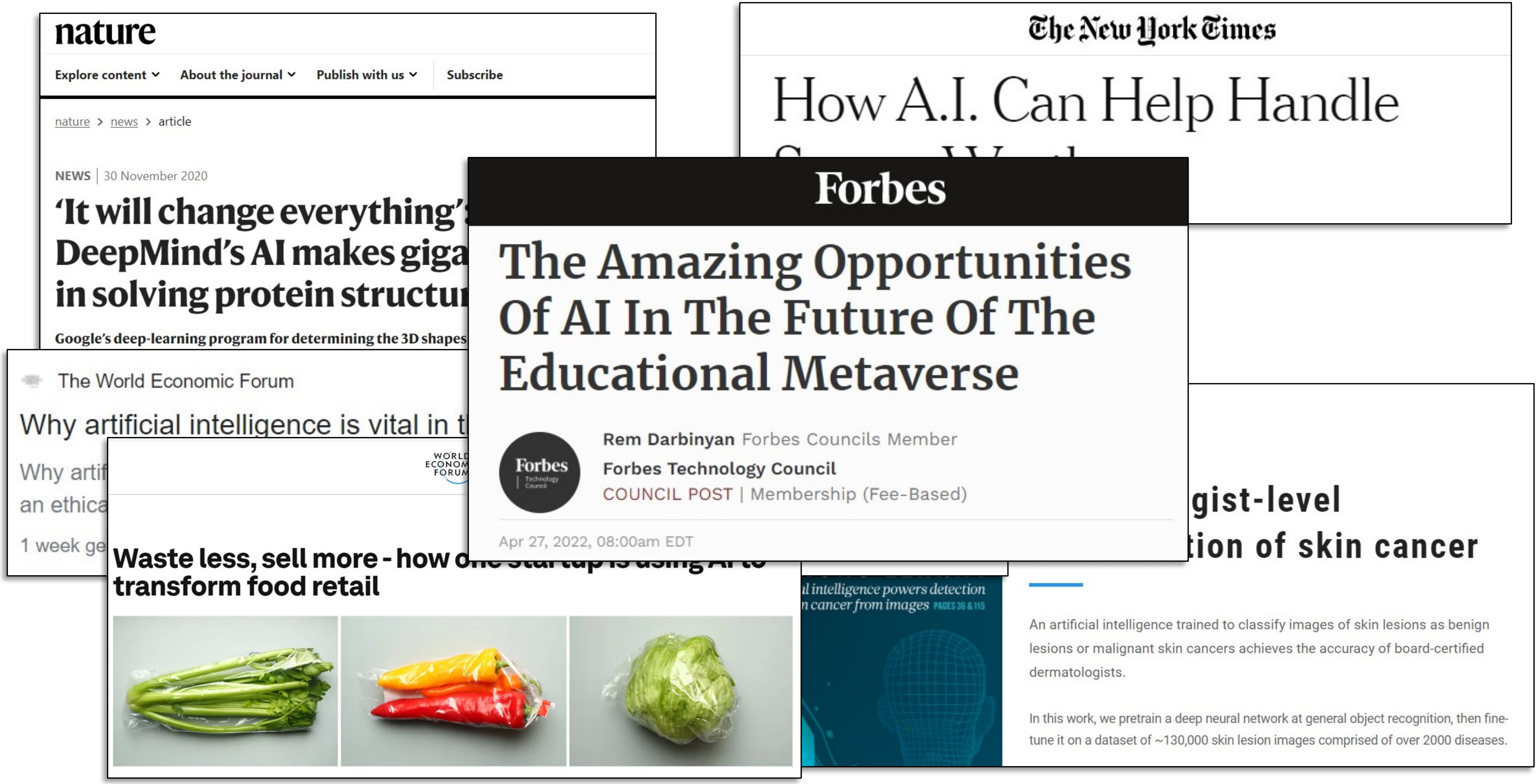
Donato Maragno, Tabea Röber, Dick den Hertog, Rob Goedhart

OPTIMAL

Optimization for and with Machine Learning

Motivation

UNIVERSITY OF AMSTERDAM
Amsterdam Business School



Motivation

DHH
@dhh
Follow

Creator of Ruby on Rails, Founder & CTO at Basecamp & HEY, NYT best-selling author, and Le Mans 24h class-winner. No DMs, no Twitter replies, email: dhh@hey.com
dhh.dk Joined April 2008
7 Following 442,3K Followers

DHH
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Traduci il Tweet

9:34 PM · 7 nov 2019 · Twitter for iPhone

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!".

64 696 4.553

Steve Wozniak
@stevewoz

I'm a current Apple employee and founder of the company and the same thing happened to us (10x) despite not having any separate assets or accounts. Some say the blame is on Goldman Sachs but the way Apple is attached, they should share responsibility.

8:06 AM · Nov 10, 2019

CNN BUSINESS

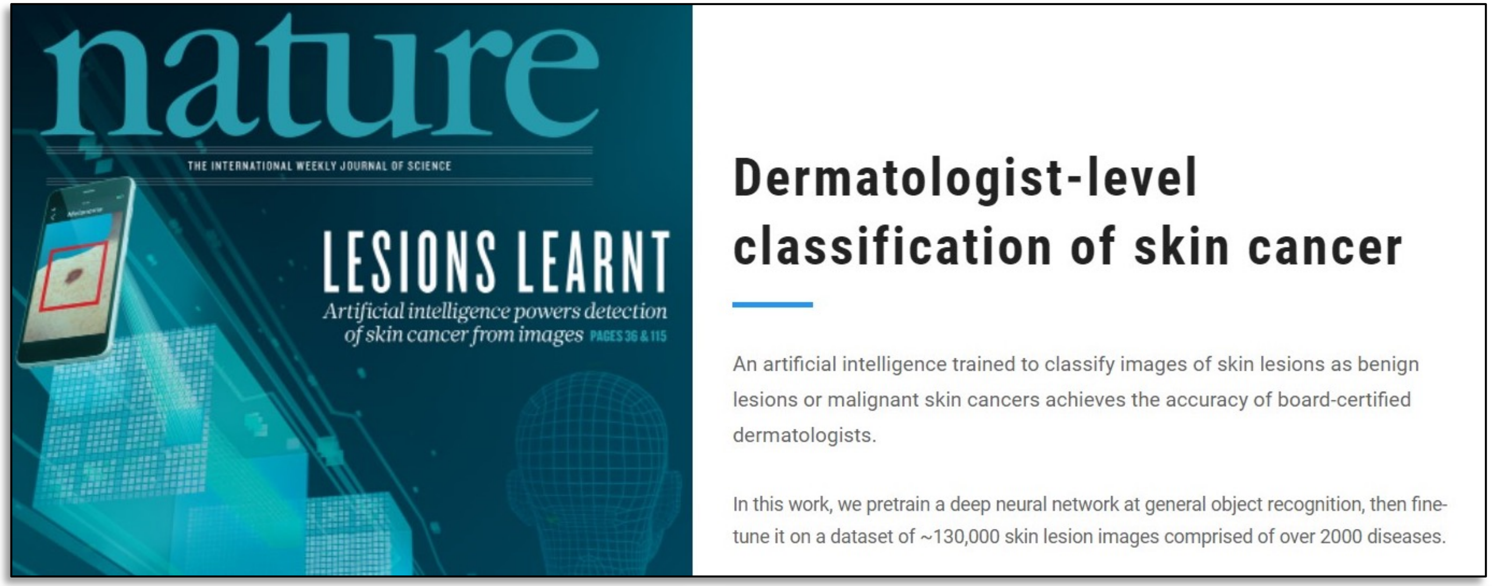
Apple co-founder Steve Wozniak says Apple Card discriminated against his wife

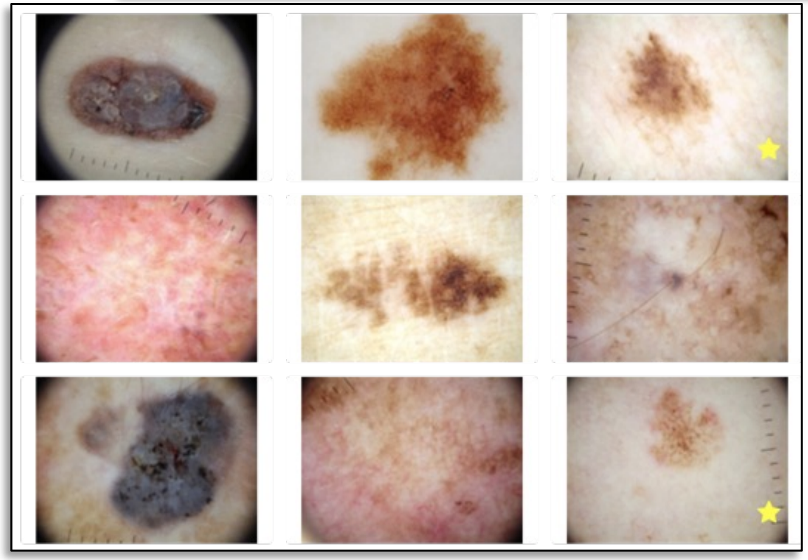
By Clare Duffy, CNN Business
Updated 1615 GMT (0015 HKT) November 11, 2019

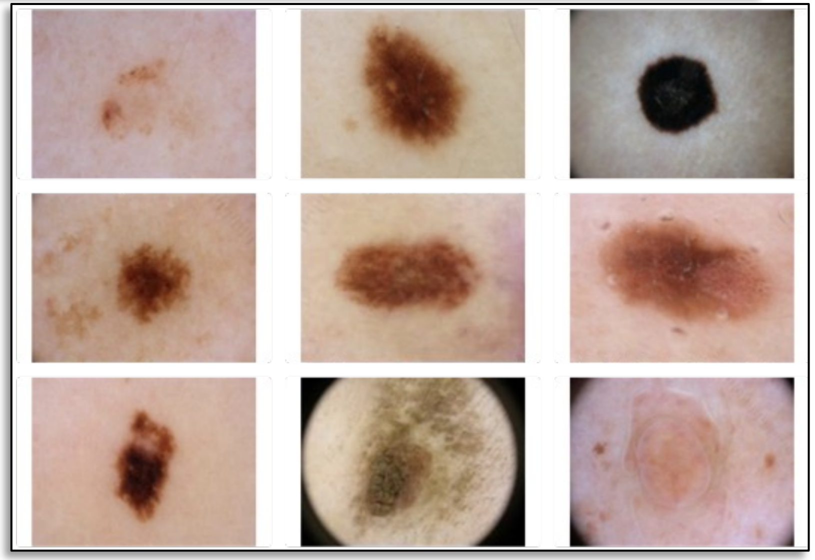
Motivation

Did the model learn the true pattern?
Is the model discriminating?

<u>Test accuracy</u>
In-lab vs. Real-life deployment







source

Motivation

"An AI system is **explainable** if the task model is intrinsically interpretable (here the AI system is the task model) or if the non-interpretable task model is complemented with an interpretable and faithful explanation (here the AI system also contains a post-hoc explanation)."

Markus et al. (2020)

Explainable Artificial Intelligence (XAI) Methods

- Model-based explanations: linear/logistic regression, decision trees, k-nearest neighbours.
- Post-hoc explanations: instance level vs global level
  - **Model-based explanations**: other interpretable models are used to explain the uninterpretable model.
  - **Attribution-based explanations**: features importance methods.
  - **Example-based explanations** → Counterfactual Explanations (CE)

Counterfactual Explanations

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

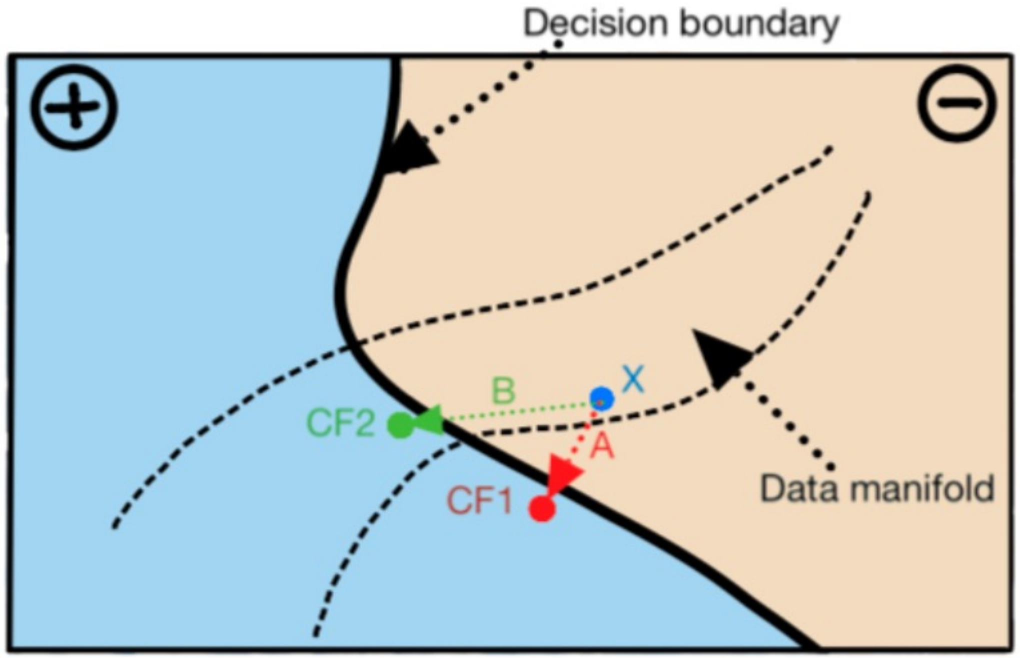
Harvard Journal of Law & Technology, 2018

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Artificial Intelligence, 2019



(Verma et al., 2020)

Counterfactual Explanations for Machine Learning: A Review

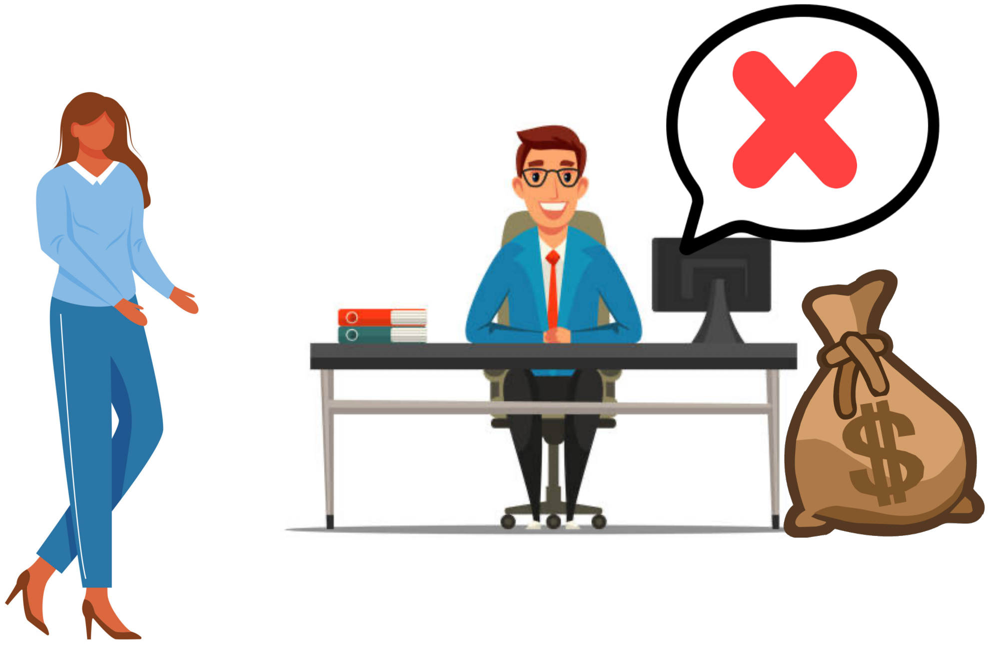
Sahil Verma	John Dickerson	Keegan Hines
University of Washington	Arthur AI	Arthur AI
Arthur AI	University of Maryland	keegan@arthur.ai
vsahil@cs.washington.edu	john@arthur.ai	

arXiv, 2020

Counterfactual Explanations

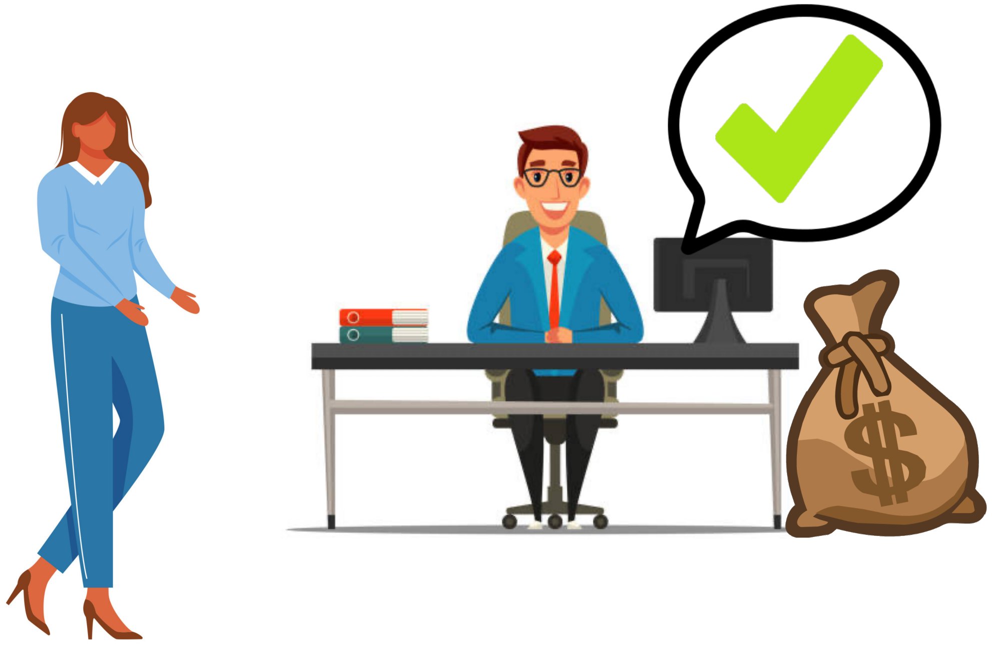
Factual Instance

Beatrice is 27yo
Full-time job: 45K $/y
Account balance: 50K $



Counterfactual

Beatrice is 27yo
Full-time job: **50K** $/y
Account balance: **60K** $



Counterfactual: set of features that should be changed in order to flip a model's prediction

Mathematical Model

$$\tilde{\boldsymbol{x}} = arg \min_{\boldsymbol{x} \in \mathbb{R}^n} d(\boldsymbol{x}, \hat{\boldsymbol{x}})$$

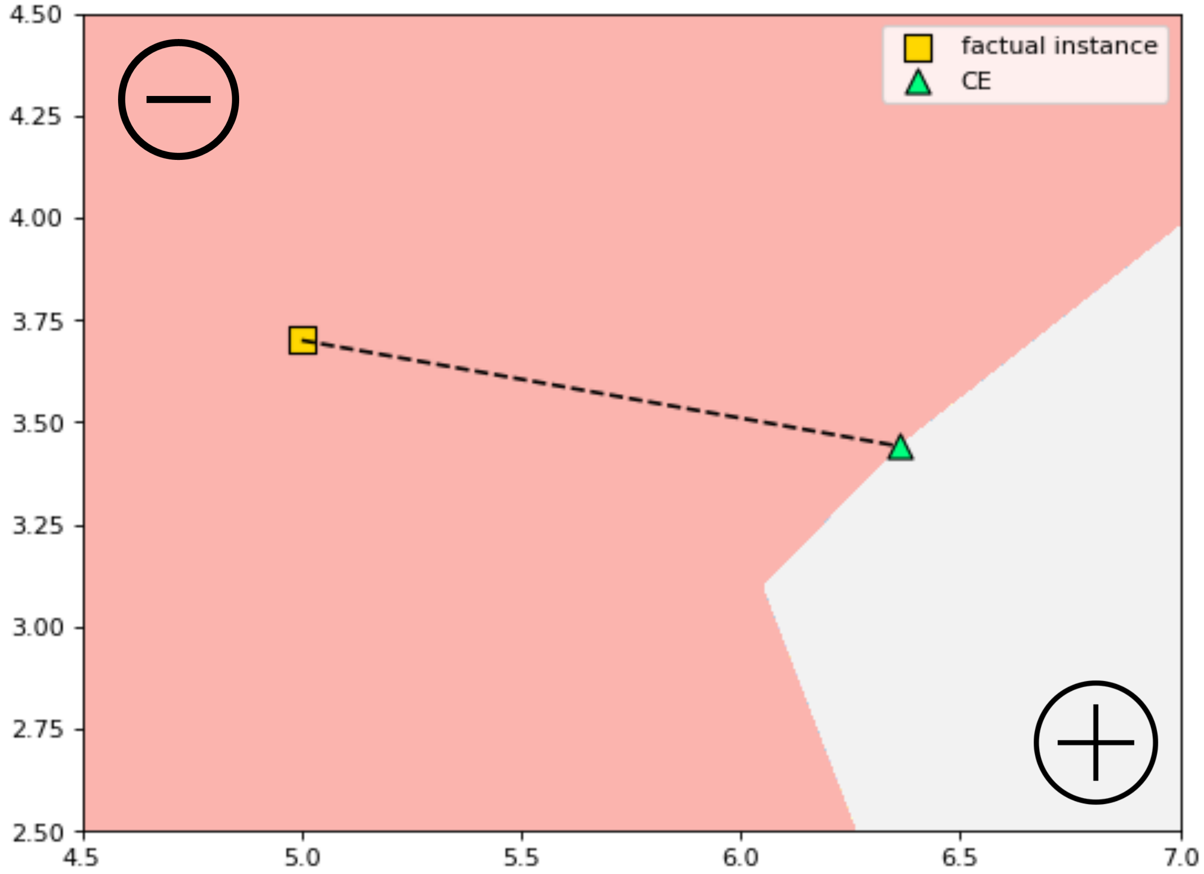
$$s.t. \;\; h(\boldsymbol{x}) = \tilde{y}$$

$\boldsymbol{d}(.,.)$: distance function
$\boldsymbol{h}(.,.)$: trained model
$\hat{\boldsymbol{x}}$: factual instance
$\tilde{\boldsymbol{x}}$: counterfactual explanation
$\tilde{y}$: desired outcome

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

Harvard Journal of Law & Technology, 2018



“Good” Counterfactual Explanations (CEs)

	Proximity	Sparsity	Coherence	Actionability	Data Manifold Closeness	Causality	Diversity
Russell [2019]	●	◐	●	-	-	-	●
Ustun et al. [2019]	●	●	●	●	-	-	-
Kanamori et al. [2020]	●	-	●	-	●	-	-
Mahajan et al. [2019]	●	-	●	◐	●	●	-
Karimi et al. [2021]	●	-	●	-	-	●	-
Kanamori et al. [2021]	●	●	●	●	-	●	●
Mothilal et al. [2020]	●	◐	●	●	-	◐	●
Karimi et al. [2020]	●	●	●	●	-	-	●
CE-OCL	●	●	●	●	●	●	●

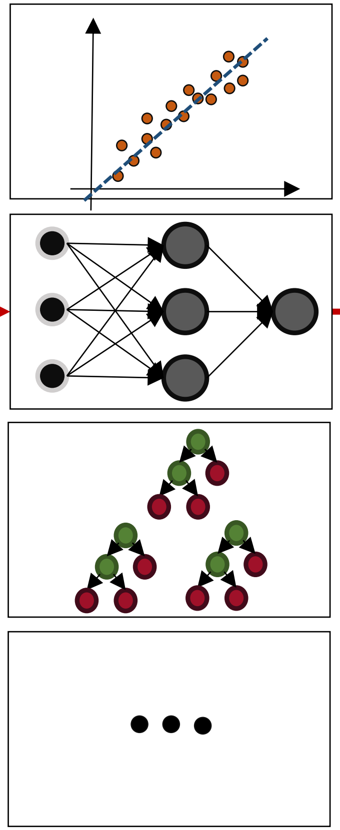
●: addressed; ◐: partially addressed; –: absent

[illegible]

Optimization with Constraint Learning (OCL)

Maragno et al. (2021)

$D = \{(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{w}}_i, \bar{\boldsymbol{y}}_i)\}_{i=1}^{N}$



$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y})$$

$$\text{s.t.} \quad \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) \leq \boldsymbol{0}$$

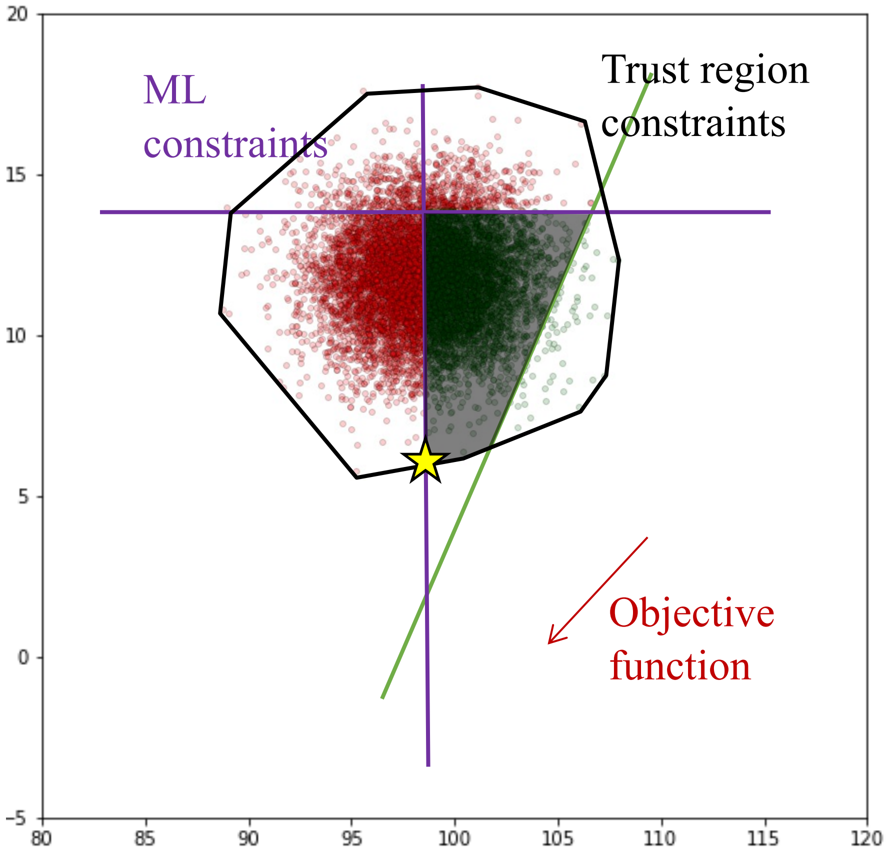
$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{w})$$

Trust region constraints

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$



OCL Model

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})}$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \le \boldsymbol{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \overline{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \ge 0, \qquad i = 1,\dots,N$$

Proximity $\min_{\boldsymbol{x}\in\mathbb{R}^n} d(\boldsymbol{x},\widehat{\boldsymbol{x}}) \rightarrow l_1, l_2, l_\infty - norms$

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\boxed{\begin{aligned} \text{s.t. } & \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0} \\ & \boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w}) \end{aligned}}$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \overline{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \dots, N$$

Validity $\quad h_{\mathcal{D}}(\boldsymbol{x}) = \tilde{y}$

Proximity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\boxed{\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0}}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_i^N \lambda_i \overline{\boldsymbol{x}}_i$$

$$\sum_i^N \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \dots, N$$

Proximity

Validity

Coherence $\sum_{i\in\mathcal{C}_j} x_i = 1, \qquad j = 1, \dots, k$

Sparsity $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_0 \leq K$

Actionability $x_i = \hat{x}_i, \qquad \forall i \in \mathcal{I}_{im}$

Causality

Data manifold closeness

Diversity

OCL Model

University of Amsterdam
Amsterdam Business School

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\boxed{\begin{aligned} \text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) &\leq \boldsymbol{0} \\ \boldsymbol{y} &= \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w}) \end{aligned}}$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \overline{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$

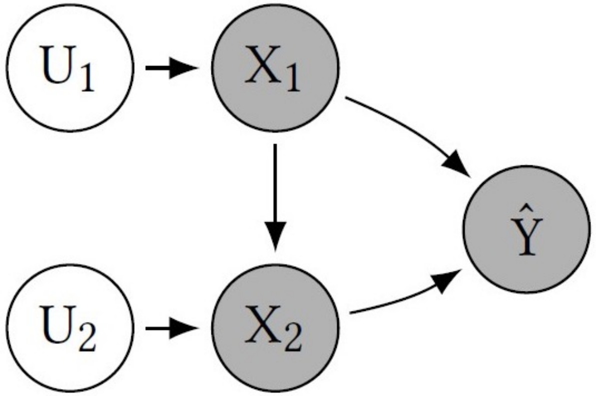
Proximity

Validity

Coherence

Sparsity

Actionability



Causality $\quad x_i = \hat{x}_i + c_i(\boldsymbol{p}_i) - c_i(\hat{\boldsymbol{p}}_i), \quad \forall i \in \mathcal{E}$

[Karimi et al. (2020)](#)

Data manifold closeness

Diversity

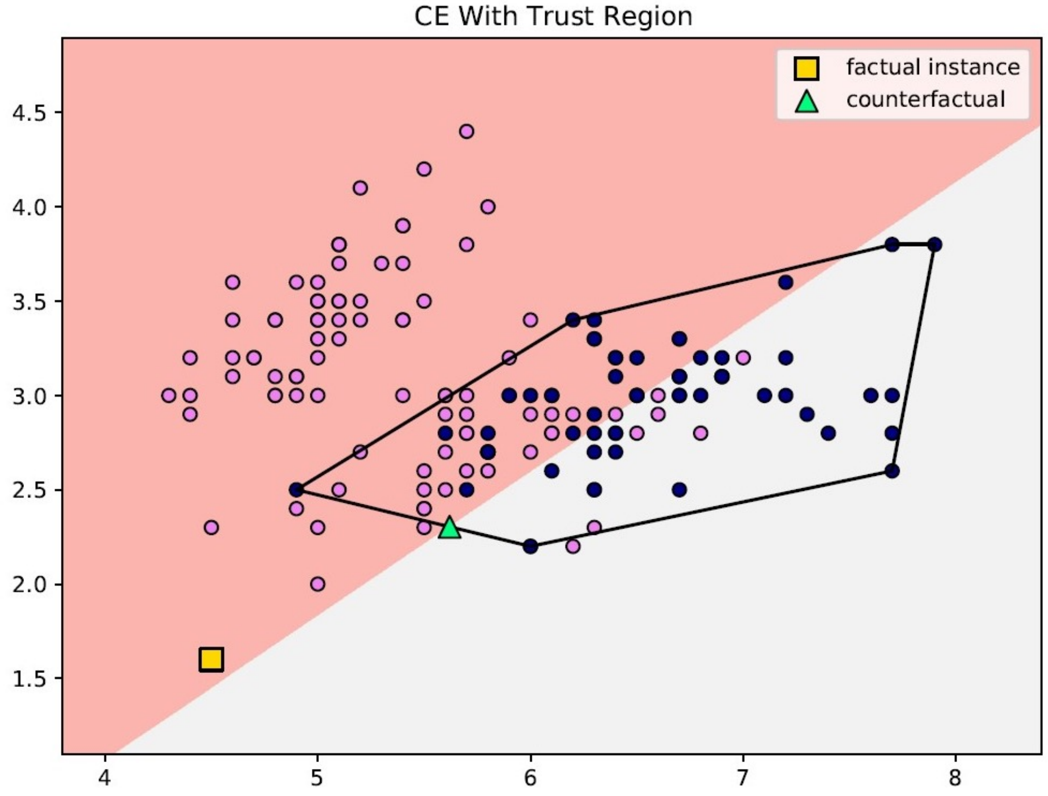
OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \mathbf{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boxed{\begin{aligned} &\boldsymbol{x} = \sum_i^N \lambda_i \overline{\boldsymbol{x}}_i \\ &\sum_i^N \lambda_i = 1 \\ &\lambda_i \geq 0, \qquad i = 1, \dots, N \end{aligned}}$$



Data manifold closeness

Proximity

Validity

Coherence

Sparsity

tionability

Causality

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0}$$

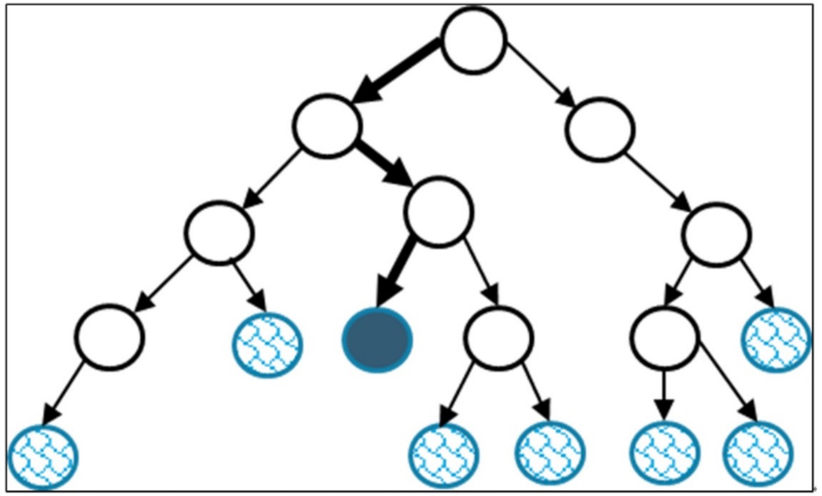
$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \overline{\boldsymbol{x}}_i$$

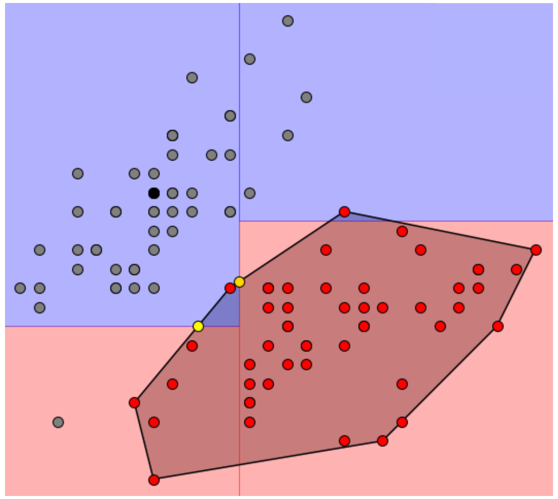
$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$

Incumbent solutions



source



Diversity

Proximity

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Case Study

University of Amsterdam
Amsterdam Business School

OptiCL:

A Python Package for Optimization with Constraint Learning

Codes and Examples

https://github.com/hwiberg/OptiCL

Label	Variable name
F1	duration
F2	credit_amount
F3	instalment_commitment
F4	age
F5	residence_since
F6	existing_credits

y: good or bad credit risk?

	F1	F2	F3	F4	F5	F6	F7	F8*	F9*	F
$\hat{x}$	24.0	1371.26	4.0	25.0	4.0	1.0	1.0	A	A	1
Part A: validity, proximity, coherence										
(a)	16.48	-57.75	3.88	26.71						
Part B: validity, proximity, coherence, sparsity										
(a)	7.12	–	–	–						
Part C: validity, proximity, coherence, sparsity, diversity										
(a)	7.12	–	–	–						
(b)	–	-3346.67	–	–						
(c)	–	–	–	–						
Part D: validity, proximity, coherence, sparsity, diversity, actionability										
(a)	7.12	–	–	–						
(b)	–	–	1.96	26.63						
(c)	–	–	–	75.52						
Part E: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness										
(a)	22.0	1283.52	–	–	–	–	–	B	–	4
(b)	–	1965.12	–	42.0	–	2.0	–	C	B	–
(c)	12.0	1893.04	–	29.0	–	–	–	–	–	–
Part F: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness, causality										
(a)	–	–	–	–						
(b)	22.0	990.51	–	–						
(c)	26.83	1910.28	–	–						

Next Steps: Enlarged Trust Region

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

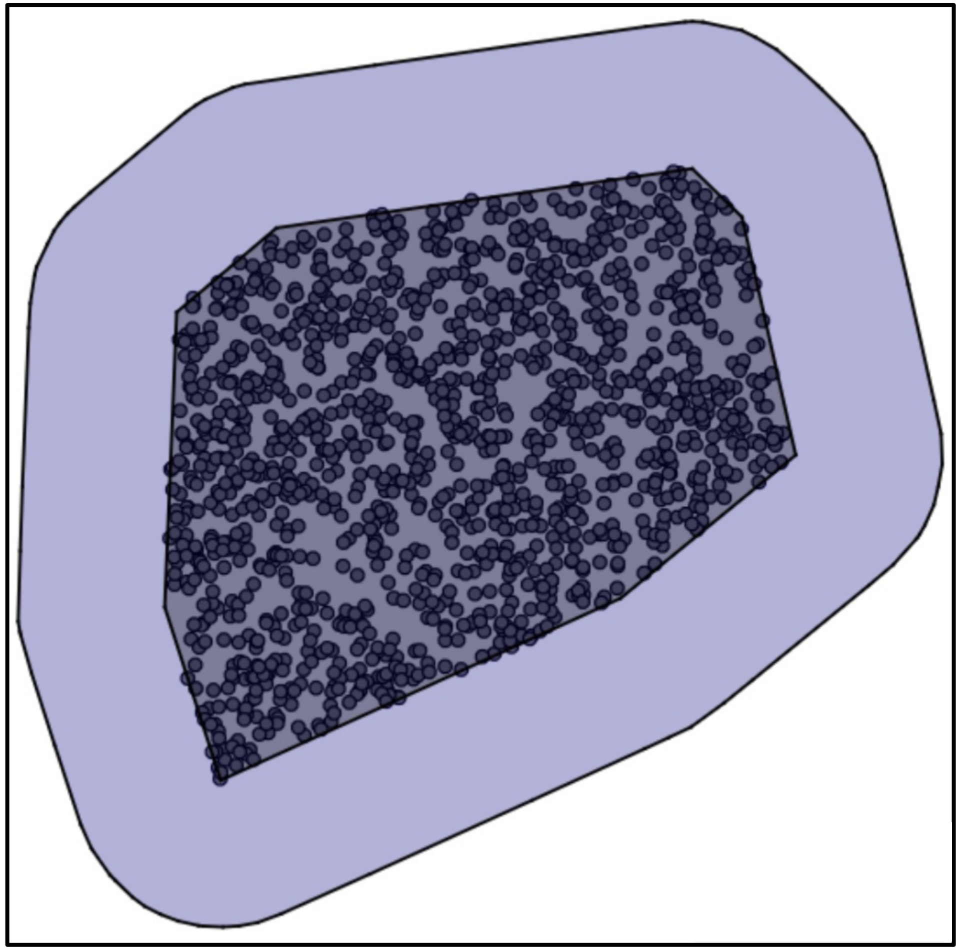
$$\lambda_i \geq 0, \qquad i = 1, \dots, N$$

$$\boldsymbol{x} + \boldsymbol{s} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \dots, N$$

$$\|\boldsymbol{s}\|_p \leq \epsilon$$



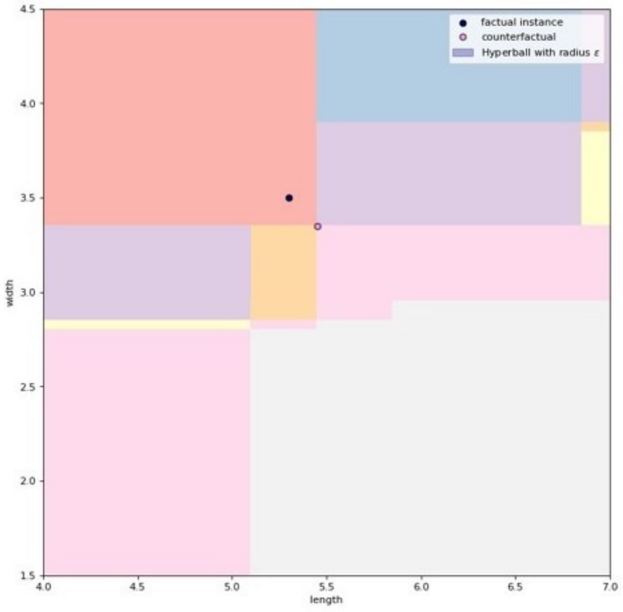
Next Steps: Intervals of CEs

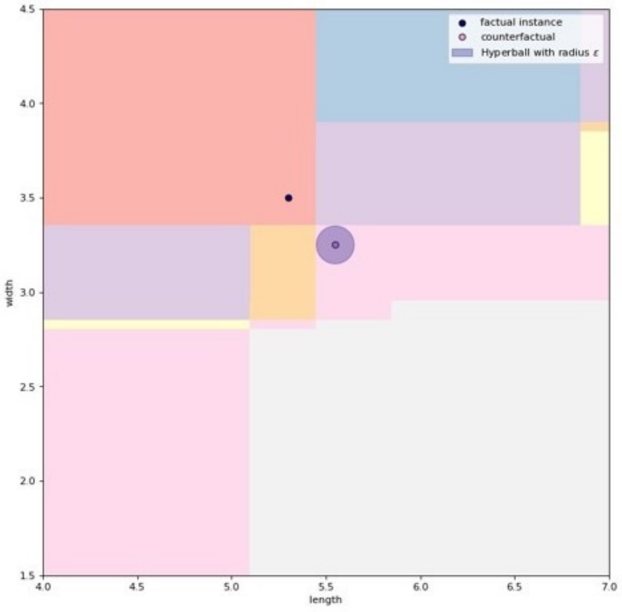
$$\widetilde{\boldsymbol{x}} = arg \min_{\boldsymbol{x} \in \mathbb{R}^n} d(\boldsymbol{x}, \widehat{\boldsymbol{x}})$$
$$s.t. \;\; h(\boldsymbol{x}) = \tilde{y}$$

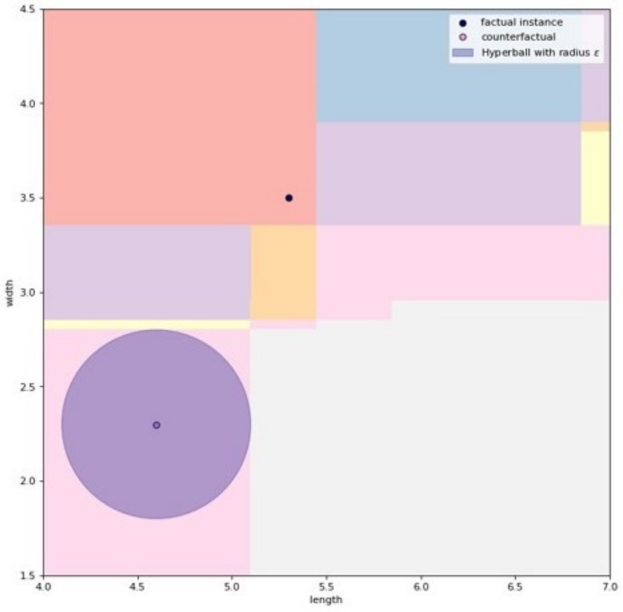
$$\widetilde{\boldsymbol{x}} = arg \min_{\boldsymbol{x} \in \mathbb{R}^n} d(\boldsymbol{x}, \widehat{\boldsymbol{x}})$$
$$s.t. \;\; h(\boldsymbol{x} + \boldsymbol{s}) = \tilde{y}, \quad \boldsymbol{s} \in \boldsymbol{S}$$

"Uncertainty Set"

Example: $\boldsymbol{S} = \{\boldsymbol{s} \in \mathbb{R}^n : \|s\|_p \leq \epsilon\}$





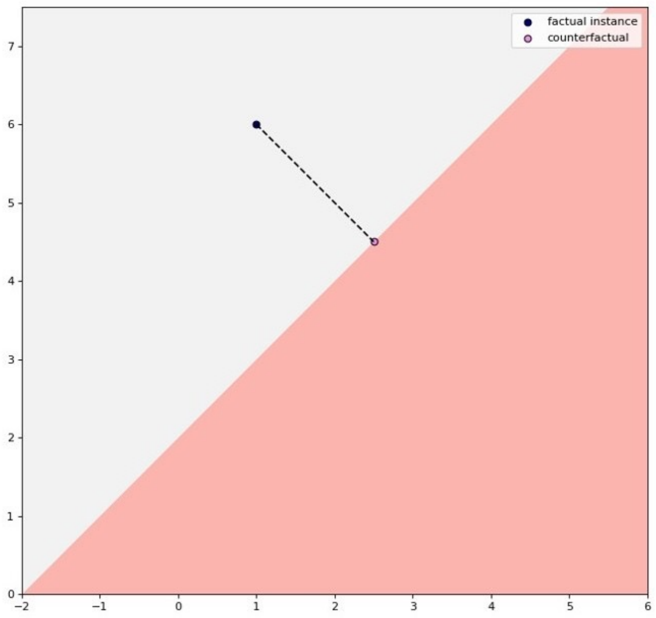


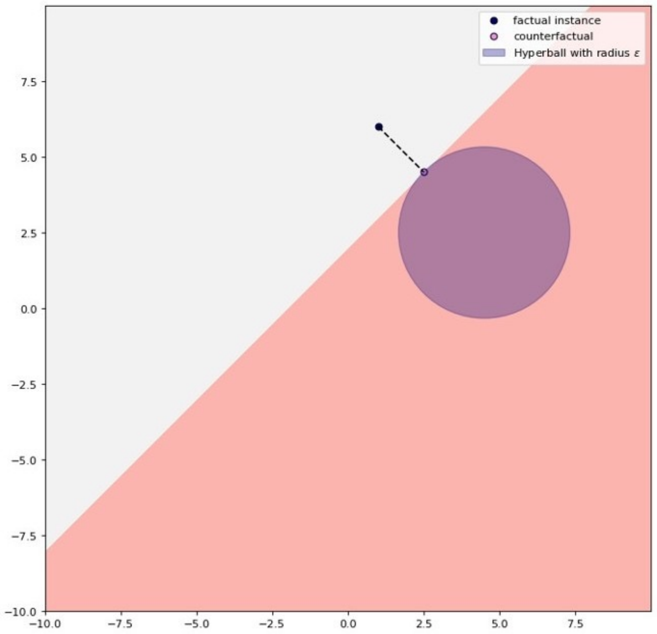
Next Steps: Intervals of CEs

$$h(\boldsymbol{x}) = \tilde{y}$$

$$h(\tilde{\boldsymbol{x}} + \boldsymbol{w}) = \tilde{y}, \quad \boldsymbol{w} \in \boldsymbol{W}$$

$$\max_{\boldsymbol{W}} g(\boldsymbol{W})$$
$$s.t. \quad h(\tilde{\boldsymbol{x}} + \boldsymbol{w}) = \tilde{y}, \quad \boldsymbol{w} \in \boldsymbol{W}, \quad \tilde{\boldsymbol{x}} \in \boldsymbol{W}$$





DHH
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Traduci il Tweet

9:34 PM · 7 nov 2019 · Twitter for iPhone

Apple Card
@AppleCard

Well, if your wife would have had a 20+ years 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.

source

Conclusion

- A general framework for CEs
- ϵ-convex hull for data manifold closeness
- Intervals for infinitely-many CEs

- Additional case studies
- Comparison against other methods
- User study via dedicated webpage

- Food for thought: Adversarial attacks