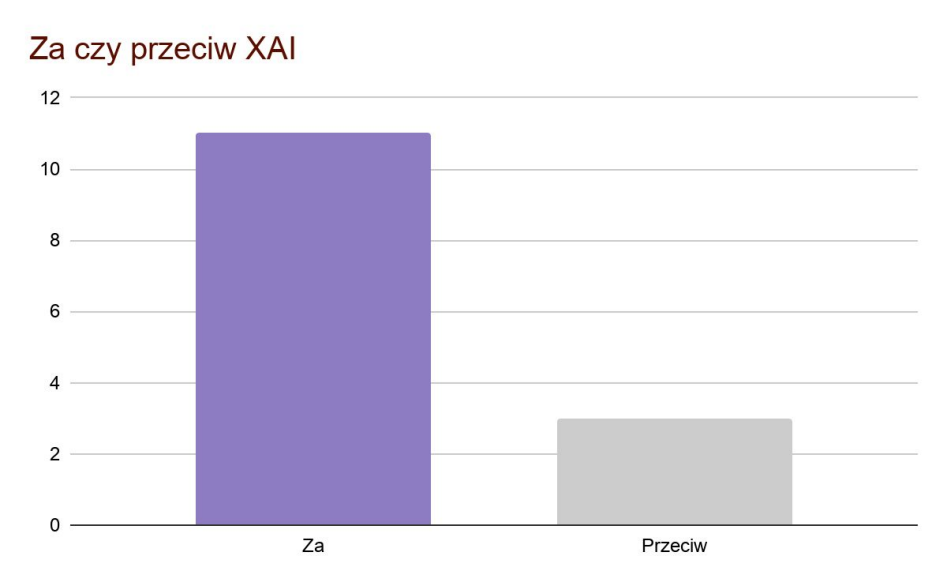
XAI : What's going on?

Kasia Woźnica Anna Kozak

Seminarium MI2DataLab, 30.03.2020

Jesteś za czy przeciw XAI?



Zachary C. Lipton

FOLLOW

GET MY OWN PROFILE

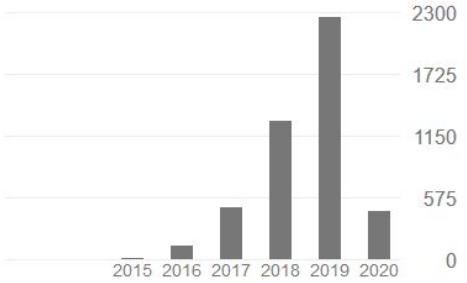
Assistant Professor at Carnegie Mellon University
Verified email at cmu.edu - Homepage

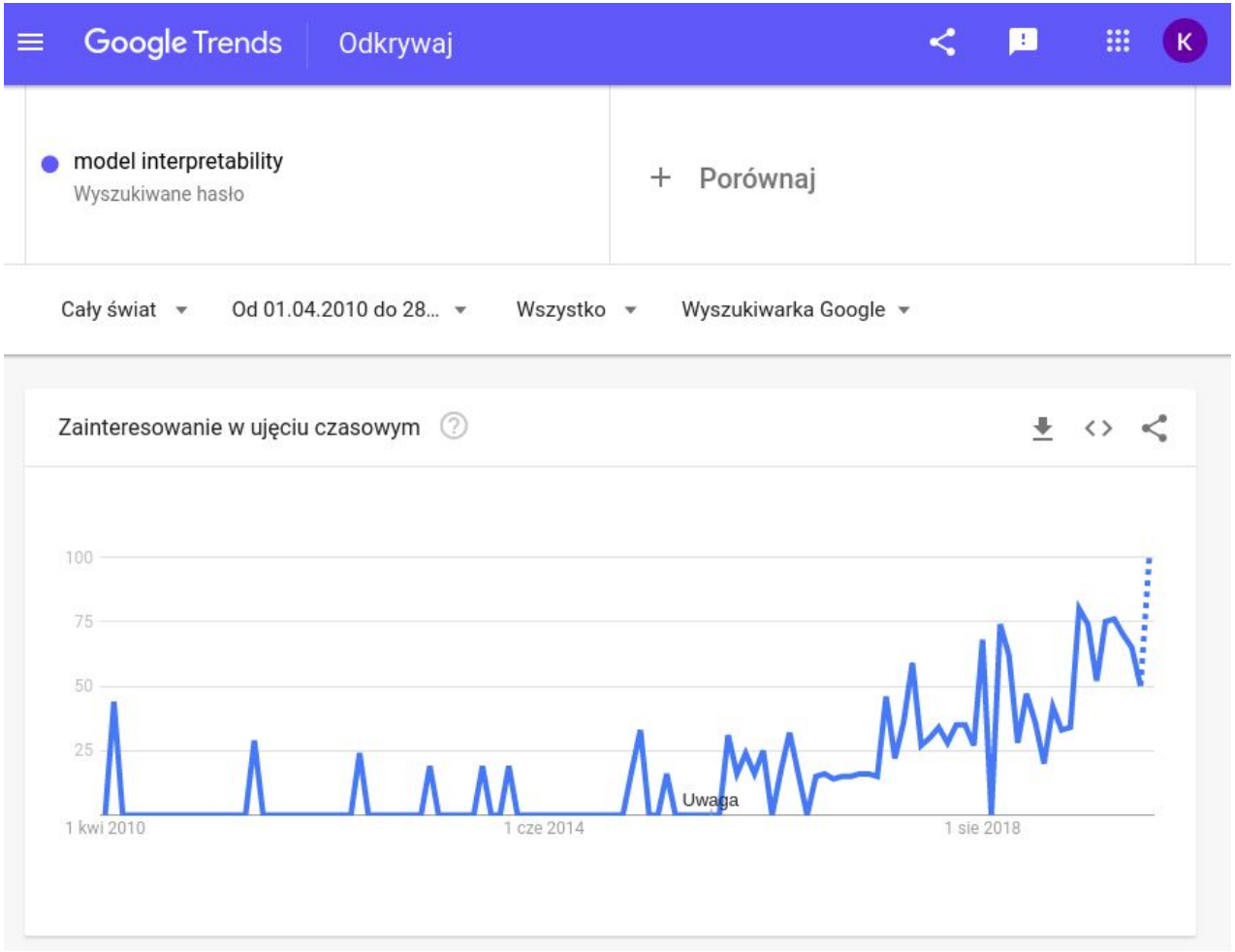
Artificial Intelligence Machine Learning Healthcare Technology & Society
Natural Language Processing

TITLE	CITED BY	YEAR
The Mythos of Model Interpretability ZC Lipton Communications of the ACM (CACM) [Prev. ICML Workshop on Human …	1140	2016
A critical review of recurrent neural networks for sequence learning ZC Lipton, J Berkowitz, C Elkan arXiv preprint arXiv:1506.00019	1033	2015
Learning to Diagnose with LSTM Recurrent Neural Networks ZC Lipton, DC Kale, C Elkan, R Wetzell International Conference on Learning Representations (ICLR)	588	2015
Modeling Missing Data in Clinical Time Series with RNNs ZC Lipton, DC Kale, R Wetzel Machine Learning for Healthcare (MLHC)	175 *	2016

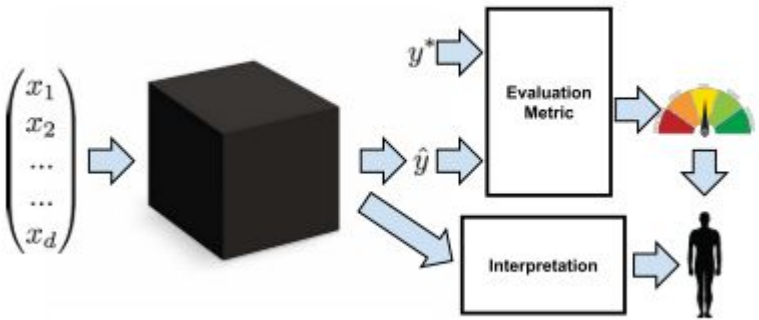
Cited by

	All	Since 2015
Citations	4712	4701
h-index	25	25
i10-index	39	39




Google Trends
Odkrywaj
K
model interpretability
Wyszukiwane hasło
+ Porównaj
Cały świat
Od 01.04.2010 do 28...
Wszystko
Wyszukiwarka Google
Zainteresowanie w ujęciu czasowym
100
75
50
25
Uwaga
1 kwi 2010
1 cze 2014
1 sie 2018

Why we need interpretable models?



It turns out that many situations arise when our real world objectives are difficult to encode as simple real-valued functions.

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.

The problem can also arise when the dynamics of the deployment environment differ from the training environment.

Interpretability - what does it mean?

Many papers propose interpretability as a means to engender **trust**. But what is trust? Does it refer to faith in a model's performance, robustness, or to some other property of the decisions it makes? Does interpretability simply mean a low-level mechanistic **understanding** of our models? If so does it apply to the features, parameters, models, or training algorithms? Other papers suggest a connection between an interpretable model and one which uncovers **causal structure** in data.

Glassbox/Black box

Some papers equate interpretability with understandability or interlligibility, i.e., that we can grasp how the model work. In these papers, understandable models are called transparent, while incomprehensible models are called black box. But what constitutes transparency? We might look to the algorithm itself. Will it converge? Does it produce a unique solution? Or we might look to its parameters: do we understand what each represents? Alternatively, we could consider the model's complexity. Is it simple enough to be examined all at once by a human?

Other papers investigate so-called post-hoc interpretations.These interpretations might explain predictions without elucidating the mechanisms by which models work. Examples of post-hoc interpretations include the verbal explanations produced by people or the saliency maps used to analyze deep neural networks. Thus, humans decisions might admit post-hoc interpretability despite the black box nature of human brains, revealing a contradiction between two popular notions of interpretability.

Trust in model?

Is it simply confidence that a model will perform well?

Trust might also be defined subjectively.

- a person might feel more at ease with **a well-understood model,** even if this understanding served no obvious purpose (we might call a model transparent if a person can contemplate the entire model at once).

- when the training and deployment objectives diverge, trust might denote confidence that the **model will perform well with respect to the real objectives and scenarios.**

Trust in model?

Another sense in which we might trust a machine learning model might be that we feel comfortable relinquishing control to it. In this sense, we might care not only about how often a model is right but also **for which examples it is right.**

Another important issue is fairness.

researchers have expressed concern that we must produce interpretations for the purpose of assessing whether decisions produced automatically by algorithms conform to ethical standards

Informativeness?

While the machine learning objective might be to reduce error, the real-world purpose is to provide useful information.

An interpretation may prove informative even without shedding light on a model's inner workings. For example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision.

Causality?

Although supervised learning models are only optimized directly to make associations, researchers often use them in the hope of inferring properties or generating hypotheses about the natural world.

One might hope, however, that by interpreting supervised learning models, we could generate hypotheses that scientists could then test experimentally.

The associations learned by supervised learning algorithms are not guaranteed to reflect causal relationships.

Transparency?

Informally,transparency is the opposite of opacity or blackboxness. It connotes some sense of understanding the mechanism by which the model works. We consider transparency at the level of the entire model (**simulatability**), at the level of individual components (e.g. parameters)(**decomposability**), and at the level of the training algorithm(**algorithmic transparency**).

However, in order to get comparable performance, linear models often must operate on heavily hand-engineered features.

Post-hoc interpretability

Post-hoc interpretability presents a distinct approach to extracting information from learned models. While post- hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning.

- Text explanations
- Visualizations
- Local Explanations
- Explanation by example

Post-hoc interpretability

The weights of a linear model might seem intuitive, but they can be fragile with respect to feature selection and preprocessing.

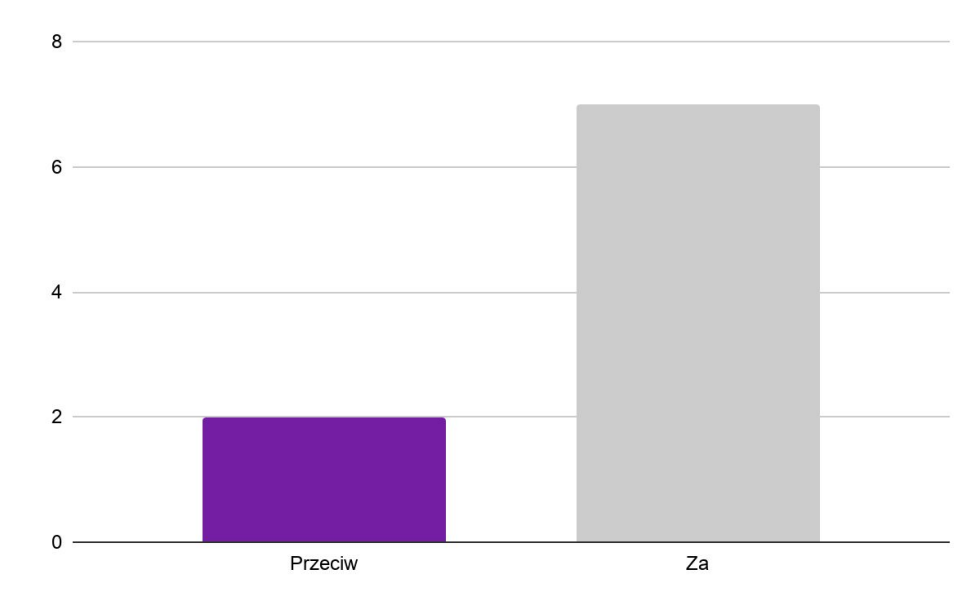
… explanations of what a model is focusing on may be misleading. The saliency map is a local explanation only. Once you move a single pixel, you may get a very different saliency map.

Globalne wyjaśnienia też są jakimś uproszczeniem działania modelu, więc porównywanie rzeczy zawsze jest obarczone jakąś niedokładnością.

MI

W “niebie” używają XAI ;)

Jesteś za czy przeciw XAI?



Dziękujemy za uwagę!