Explainable AI

Building a bridge between business and data science using Explainable AI

Warsaw, 21.03.2019

Agenda

01 What is Explainable AI – XAI ?

02 Overview of Current Development Areas of Explainable AI (XAI)

03 Exploration of AI adoption and the role of XAI

04 Final remarks

XAI in media.

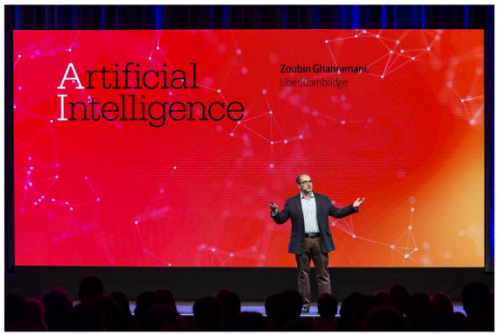
CIO JOURNAL.

Companies Grapple With AI's Opaque Decision-Making Process

Uber, Xerox's PARC, Capital One among organizations investigating how AI solves problems

By Sara Castellanos
May 2, 2018 2:15 pm ET

0 COMMENTS



Zoubin Ghahramani, chief scientist at Uber, speaks at an AI conference this week hosted by OReilly Media Inc. and Intel Corp.s AI division. PHOTO: TRICIA O'NEILL, COURTESY OF O'REILLY MEDIA

Recommended Videos

1. Elon Musk Appears to Smoke Marijuana
2. Powerful Political Messages Sent in Kavanaugh Hearings
3. Naked Body Scanner Reviewed by... Naked Cowboy
4. Three ways 5G technology will change your life

CIO JOURNAL.

Capital One Pursues 'Explainable AI' to Guard Against Bias in Models

The effort aims to better understand how a machine-learning model comes to a logical conclusion.

By Sara Castellanos
Dec 6, 2016 1:33 pm ET

1 COMMENTS

Adam Wenchel, vice president of data innovation at Capital One Financial Corp., at the AI Summit in New York on Dec. 1. PHOTO: SARA CASTELLANOS / WSJ

Most Popular Videos

1. Elon Musk Appears to Smoke Marijuana
2. How 'Fear' and New York Times Op-Ed Could Impact White House
3. Why Warren Buffett Said No to Lehman and AIG in 2008
4. Powerful Political Messages Sent in

XAI in media.

CIO JOURNAL.

Facing Growing Concern Over AI, Tech Firms Call for 'Responsible' Development

By *Steven Norton*
Oct 26, 2017 3:14 pm ET

0 COMMENTS

Employees work in front of computers at the Sinovation Ventures headquarters in Beijing, Aug. 15, 2017. PHOTO: GIULIA MARCHI/BLOOMBERG

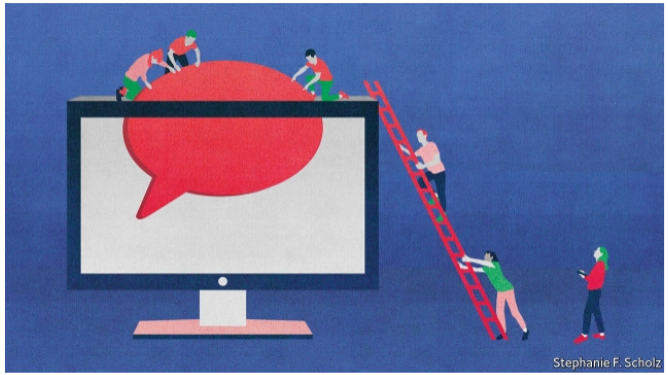
Recommended Videos

1. Elon Musk Appears to Smoke Marijuana
2. Three ways 5G technology will change your life
3. Powerful Political Messages Sent in Kavanaugh Hearings
4. Naked Body Scanner Reviewed by... Naked Cowboy

AI in society

For artificial intelligence to thrive, it must explain itself

If it cannot, who will trust it?



Print edition | Science and technology ›
Feb 15th 2018

AI applications that can make you worry.

The New York Times

Sent to Prison by a Software Program's Secret Algorithms

Chief Justice John G. Roberts Jr., center, recently said that the day of using artificial intelligence in courtrooms was already here, "and it's putting a significant strain on how the judiciary goes about doing things." Stephen Crowley/The New York Times

- Eric L. Loomis, who was sentenced to six years in prison based in part on a private company's proprietary software. Mr. Loomis says his right to due process was violated by a judge's consideration of a report generated by the software's secret algorithm, one Mr. Loomis was unable to inspect or challenge.

AI applications that can make you worry.

TECH | By Jordan Pearson | Feb 2 2017, 4:23pm

AI Could Resurrect a Racist Housing Policy

And why we need transparency to stop it.

SHARE TWEET

Data has always been a weapon. Between 1934 and 1968 the US Federal Housing Administration systematically denied loans to black people by using entire neighbourhoods, colour-coded by perceived risk factor, as their decision-making metric. Modern computer scientists might call this intentionally "coarse" data.

This practice, known as redlining, had damaging financial and social effects that spanned generations of black families. And now, experts worry that similar practices could return in the algorithms that make decisions about who poses a risk to their community, or, rather chillingly, who deserves to be granted a loan.

AI applications that can make you worry.

SCIENCE

WHAT HAPPENS WHEN AN ALGORITHM CUTS YOUR HEALTH CARE

By Colin Lecher | @colinlecher | Mar 21, 2018, 9:00am EDT

Illustrations by William Joel, Photography by Amelia Holowaty Krales

SHARE

For most of her life, Tammy Dobbs, who has cerebral palsy, relied on her family in Missouri for care. But in 2008, she moved to Arkansas, where she signed up for a state program that provided for a caretaker to give her the help she needed.

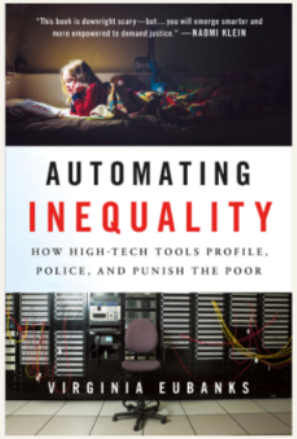
There, under a Medicaid waiver program, assessors interviewed beneficiaries and decided how frequently the caretaker should visit. Dobbs' needs were extensive. Her illness left her in a wheelchair and her hands stiffened. The most basic tasks of life — getting out of bed, going to the bathroom, bathing — required assistance, not to mention the trips to yard sales she treasured. The nurse assessing her situation allotted Dobbs 56 hours of home care visits per week, the maximum allowed under the program.

AI applications that can make you worry.

AUTOMATING INEQUALITY

How High-Tech Tools Profile, Police, and Punish the Poor

Virginia Eubanks
St. Martin's Press



READ AN EXCERPT →

BUY THE BOOK

g★★★★☆

Hardcover

St. Martin's Press
St. Martin's Press
01/23/2018
ISBN: 9781250074317
272 Pages

$26.99

Amazon | Barnes & Noble | Books-a-Million | IndieBound | Powells

But companies take steps.

Salesforce is hiring its first Chief Ethical and Humane Use officer to make sure its artificial intelligence isn't used for evil

Rosalie Chan Dec. 16, 2018, 3:10 PM

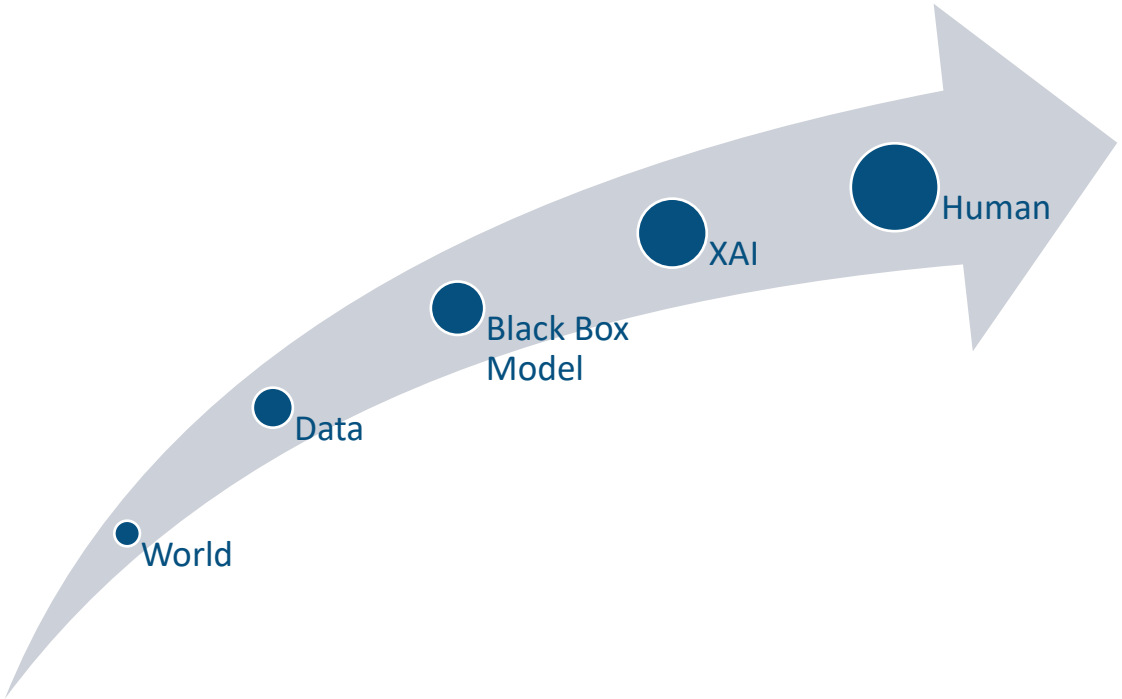
What is XAI ?

Explainable AI

- XAI aims to produce "glass box" models that are explainable to a "human-in-the-loop", without greatly sacrificing AI performance.
- Human users should be able to understand the AI's cognition (both in real-time and after the fact), and should be able to determine when to trust the AI and when the AI should be distrusted

Source https://en.wikipedia.org/wiki/Marketing_mix_modeling

XAI in action.



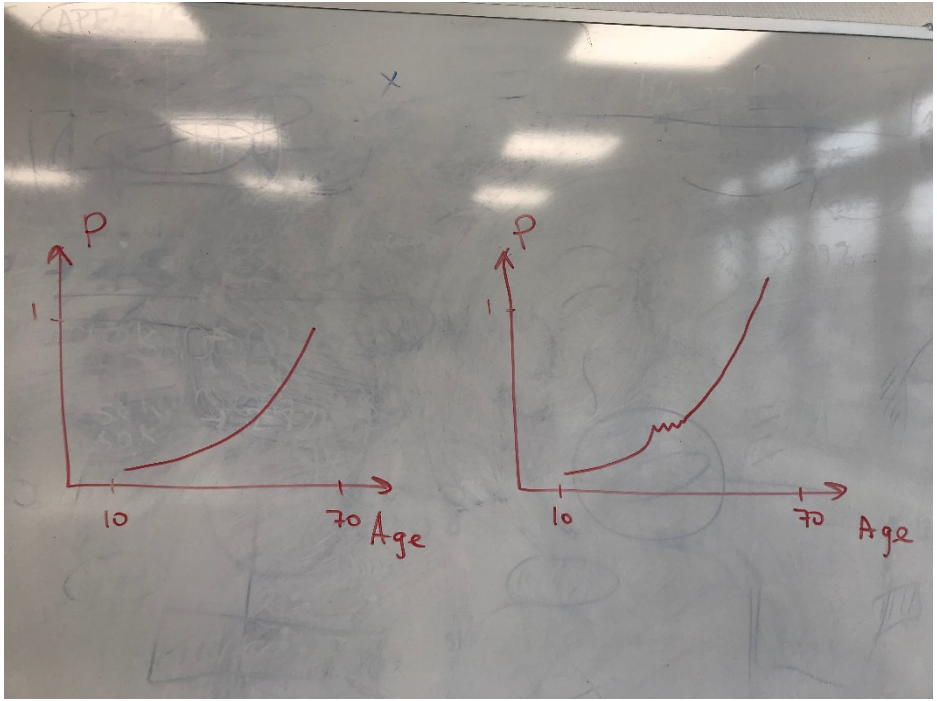
All the attention was here.

- **Is it worth spending thousands of dollars to improve prediction accuracy by 0.001% ?**

All the attention was here.

- **Is it worth spending thousands of dollars to improve prediction accuracy by 0.001% ?**
- **When real business applications will round it all up.**

A possible situation



- Two models.
  - One with feature set A
  - The other one with an extended featue set

What Data Scientists get wrong about explainability.

01 Judge AIs as alternatives rather than aides

02 Expect stakeholders to “think more like me”

03 Optimize for model performance over enterprise utility

04 Value XAI only as a placebo

05 Believe what is said is what will be heard

06 Provide a single explanation for all audiences

07 Undervalue explanation friendly features

08 Fail to design for debugging

09 Assume rather than demonstrate generalizability

10 Think moonshots are the model

Source https://xai.world/2018/01/25/what-data-scientists-get-wrong-about-explainability/

DARPA program

XAI will enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

CIO JOURNAL.

The Morning Download: Darpa Orchestrates Effort to Make AI Explain Itself

Aug 11, 2017 7:56 am ET

0 COMMENTS

Darpa's David Gunning PHOTO: DEFENSE ADVANCED RESEARCH PROJECT AGENCY

Recommended Videos

1. iPhone XR, XS and XS Max: First Look
2. Jack Ma's Retirement Plans: In His Own Words
3. China, Russia Cement Ties Over Caviar, War Games

GDPR

- ✓ **GDPR Article 22 Paragraph 3 states that a data controller "shall implement suitable measures to safeguard…at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision", otherwise a person has "the right not to be subject to a decision based solely on automated processing" (Paragraph 1).**

GDPR

- ✓ GDPR Article 22 Paragraph 3 states that a data controller “shall implement suitable measures to safeguard…at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision”, otherwise a person has “the right not to be subject to a decision based solely on automated processing” (Paragraph 1).

- ✓ **There are different interpretations of GDPR „right for explanation“ .**

Why we need XAI?

01 Safety. We should build systems that make sound decisions.

02 Debugging. We should understand why a system does not work and how to fix it.

03 Science. We want to understand something new.

04 Mismatched Objectives. The system may not be optimized for the true objective.

05 Legal. Are we legally required to provided an explanation?

Source https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf

In other words

Optimize	• Model performance • Decision making
Retain	• Control • Safety
Maintain	• Trust • Ethics
Comply	• Accountability • Regulation

What are good explanations?

01 Contrastive explanations. „Why P rather than Q?“

02 Social attribution.

03 Causal connection

04 Explanation selection

05 Simple, general, and more coherent.

Main focus areas

01 Interpretable models

02 Post-hoc explanation

Overview of methods and software.

01 **Explainable models**

02 Prediction explanation

- **Linear regression.**
  - Numerical feature: For an increase of the numerical feature x_j by one unit, the estimated outcome changes by β_j. An example of a numerical feature is the size of a house.
  - Extension: lasso, GLM, GAM
  - Represent only linear relationship.
  - Software: widely available.

Overview of methods and software.

01 **Explainable models**

02 Prediction explanation

- **Decision trees.**
  - DTs split the data set into regions and provide explanations about the logic.
  - There are a lot of methods:
    - CART
    - ID3
    - Conditional Inference Trees
    - Etc.
  - Software:
    - ctree, rpart, RWeka (R)
    - sklearn (Python)
  - Advantages : cover interactions, natural visualization, good explanation.
  - Disadvantages: lack of smoothness, unstable.

Overview of methods and software.

01 **Explainable models**

02 Prediction explanation

- **Decision rule lists and decision rule sets.**
  - Produce explainable output.
  - A lot of methods appreared in the last 5 years
    - (Scalable) Baysian Rule Lists (Rudin)
    - Falling Rule Lists (Rudin)
    - Association Rule classification: CBA, QCBA, etc.
    - Certifiably Optimal Rule Lists (Rudin)
    - Interpretable Desicion Sets (Lakkaraju)
    - Etc.
  - Software:
    - sbrl (R)
    - Skater (Python)
    - Mostly code on github

Overview of methods and software.

01 **Explainable models**

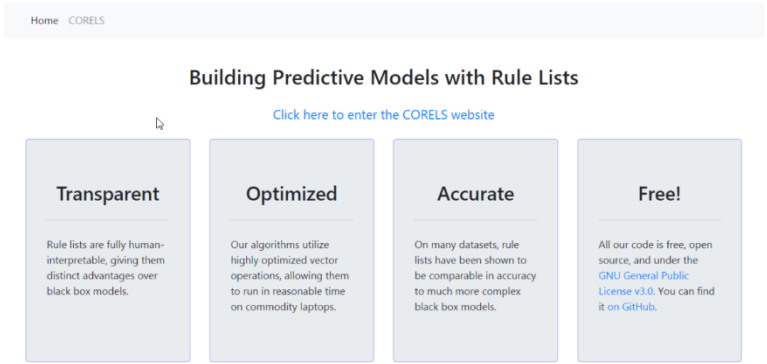
02 **Prediction explanation**

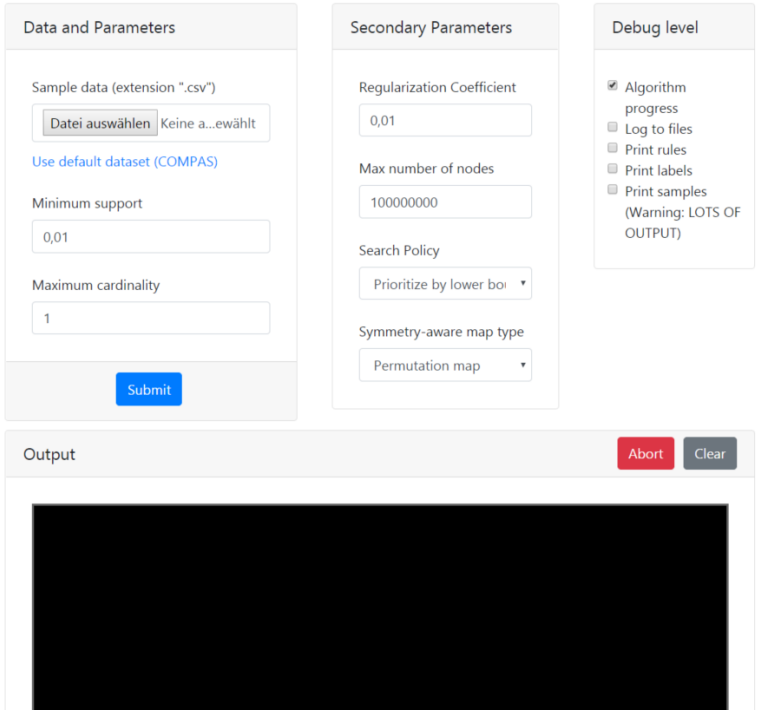
- **Decision rule lists and decision rule sets.**
  - Advantages.
    - High level of interpretability
    - Good performance (for certain data sets)
    - Sample output

      If (city is London AND hour is 12) then prob= 0.6
      Else if (city is Berlin AND hour 15) then prob= 0.4
      Else prob = 0.1

  - Disadvantages:
    - Data need to be discretized before using such methods.

Try CORELS at https://corels.eecs.harvard.edu/index.html



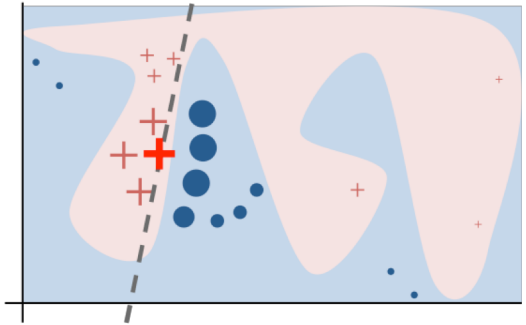


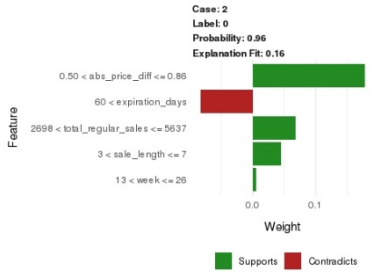
Overview of methods and software.

01 Explainable models

02 **Prediction explanation**

- **LIME**
  - It is a local surrogate model
  - It works for tabular, text, and image data
  - Software
    - Iml, lime, DALEX (R)
    - Skater, eli5 (Python)



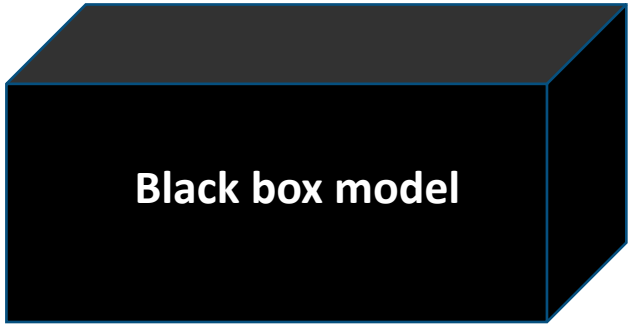


Prediction explanation?

Machine learning on a single slide

Input

Click	City	Hour
1	London	12
0	London	13
0	London	14
1	Berlin	15



Output

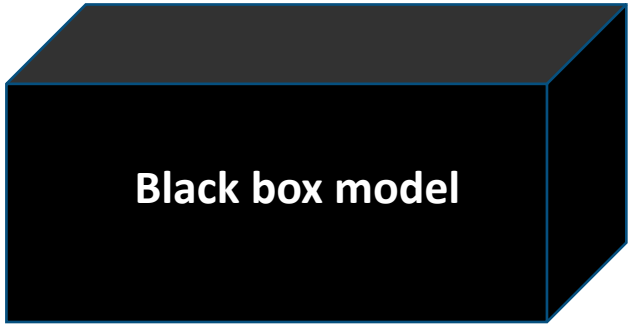
- Decision yes/no

Prediction explanation?

Machine learning on a single slide

Input

Click	City	Hour
1	London	12
0	London	13
0	London	14
1	Berlin	15



Output

- Decision yes/no
- Also generrate why yes/no

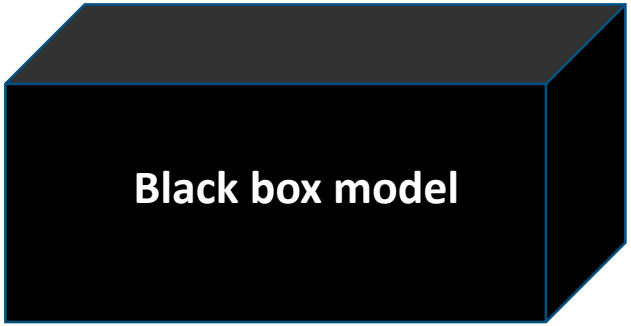
If (city is London AND hour is 12) then Yes

Prediction explanation?

Machine learning on a single slide

Input

Click	City	Hour
1	London	12
0	London	13
0	London	14
1	Berlin	15



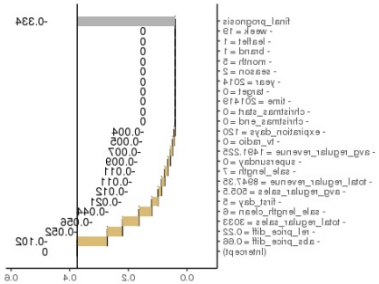
Output

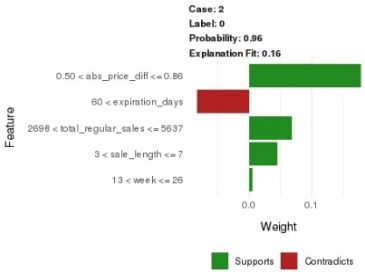
- Decision yes/no
- Also generrate why yes/no

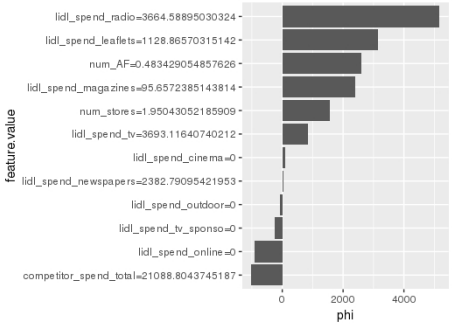
If (city is London AND hour is 13) then No

Methods for tabular data

- LIME
- SHAP
- Breakdown
- Anchor
- MAPLE
- Contrastive Explanations
- Etc.

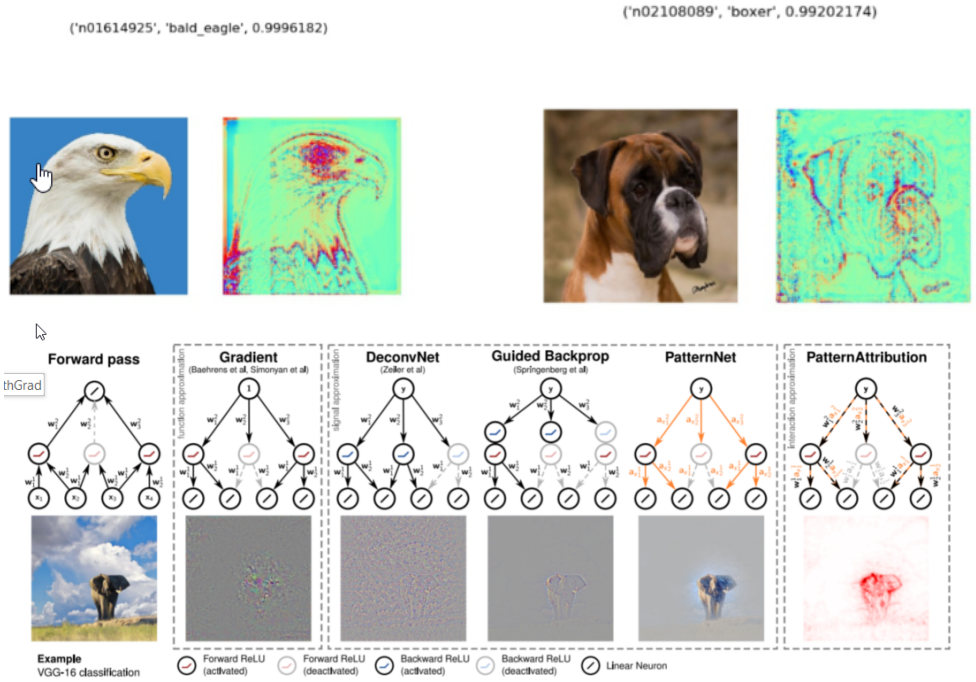




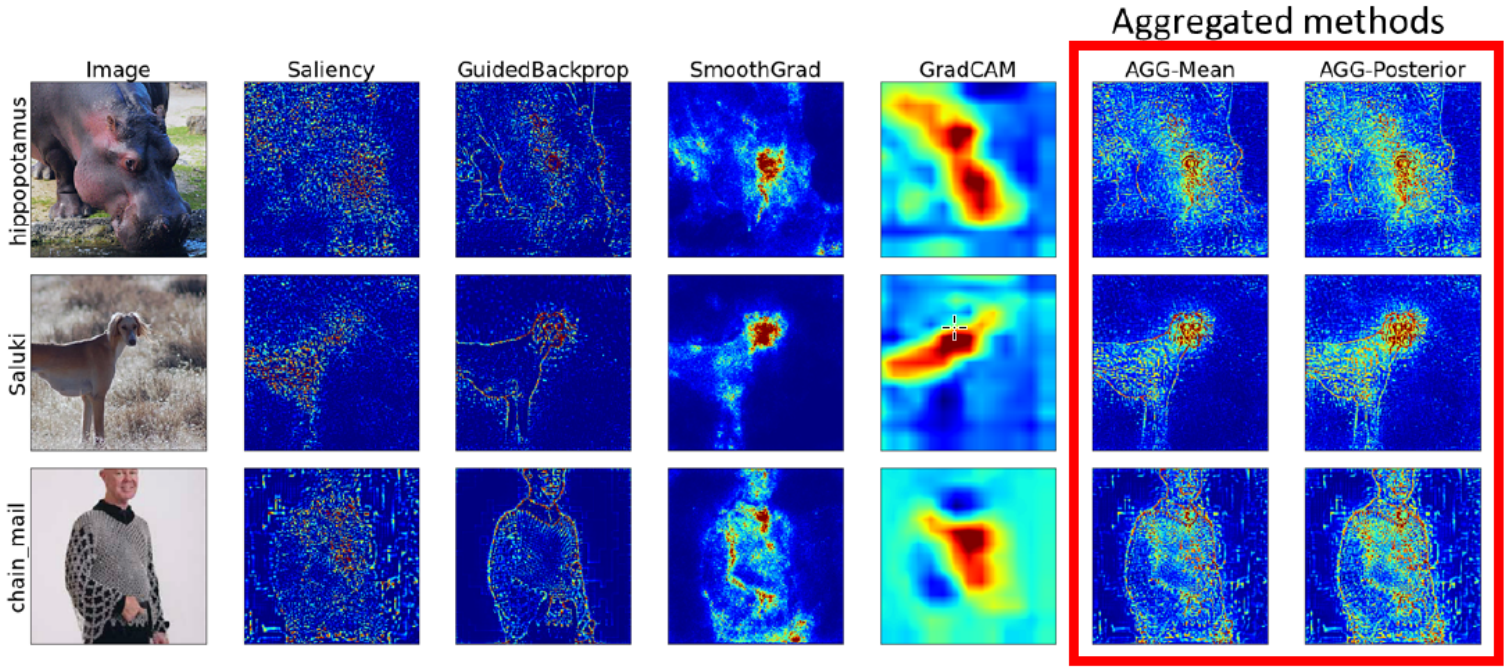


Methods for images. Deep learning.

- Saliency maps
- Layer-wise Relevance Propagation
- Gradient based methods
- TCAV
- Etc.



Another example



Methods for text data

- Layer-wise Relevance Propagation (LRP)
- Learning to explain (L2X)
- SHAP
- LIME
- Etc.

true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
		6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliches ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Another example. Sentiment prediction.

Truth	Model	Key words
positive	positive	Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.
negative	negative	Sorry to go against the flow but I thought this film was unrealistic, boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the crisis she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value. I wish I would have skipped it.
negative	positive	This movie is chilling reminder of Bollywood being just a parasite of Hollywood. Bollywood also tends to feed on past blockbusters for furthering its industry. Vidhu Vinod Chopra made this movie with the reasoning that a cocktail mix of deewar and on the waterfront will bring home an oscar. It turned out to be rookie mistake. Even the idea of the title is inspired from the Elia Kazan classic. In the original, Brando is shown as raising doves as symbolism of peace. Bollywood must move out of Hollywoods shadow if it needs to be taken seriously.
positive	negative	When a small town is threatened by a child killer, a lady police officer goes after him by pretending to be his friend. As she becomes more and more emotionally involved with the murderer her psyche begins to take a beating causing her to lose focus on the job of catching the criminal. Not a film of high voltage excitement, but solid police work and a good depiction of the faulty mind of a psychotic loser.

Applications

- Healthcare
- Autonomous vehicles
- Drug discovery
- Legal domain
- Asset management and finance
- etc.

XAI and AI adoption

- People do not trust AI
- Workers have to justify their decisions

Situation 1. Planning in retail.

- Decision
  - Yes/No
  - Numbers

Situation 1. Planning in retail.

- Information is aggregated and rules are formed

Situation 1. Planning in retail.

- Information is aggregated and rules are formed

It is a primitive communication

- 1.80 < **abs_price_diff** < 3.11 **AND** 327.75 < **avg_regular_sales** < 662.83
- **expiration_days** = 6 **AND sale_length_clean** = 3 **AND** 4689.8 < **total_regular_revenue** < 11219.6

- The model predicted 5.48 instead of more than 5.48 because **abs_price_diff** <= 1.796 **and year** > 2015
- The model predicted 4.54 instead of more than 4.54 because **total_regular_revenue** > 6129.841 **and sale_length** > 6.463

Situation 2. Marketing Mix Model Sample (fake) Data

Revenue	Radio spend	TV spend	Cinema spend	Outdoor spend
111111	234	125	100	111
121212	345	122	200	167
131213	134	111	139	89

- One can try to build a blackbox model in order to predict revenue based on different marketing chennels spends.

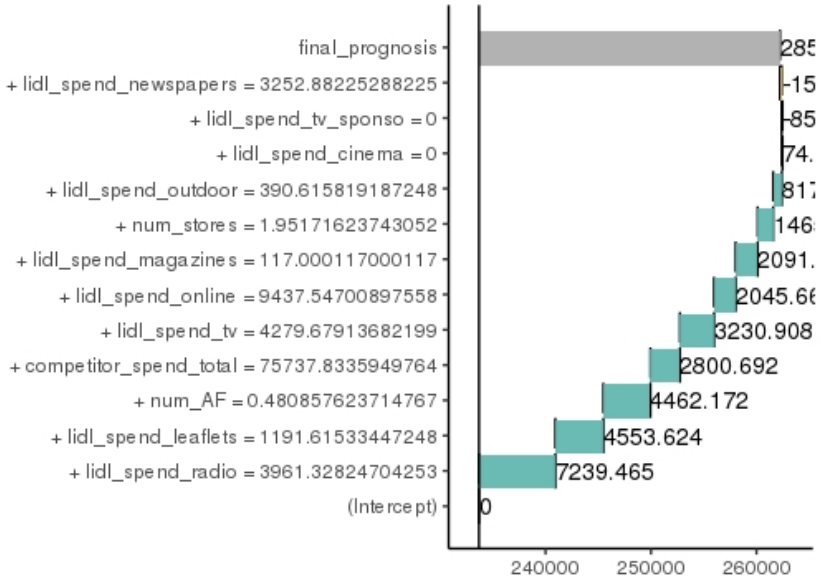
Situation 2. Marketing Mix Model Sample (fake) Data

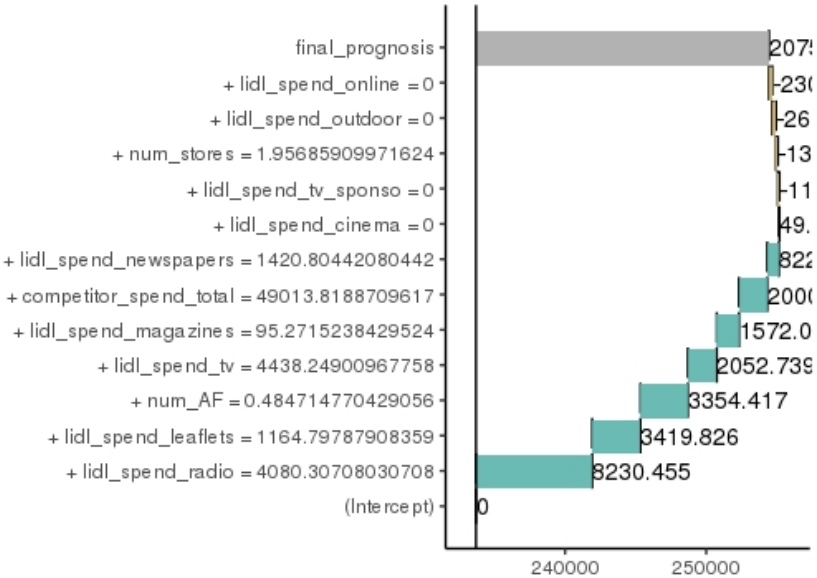
Revenue	Radio spend	TV spend	Cinema spend	Outdoor spend
111111	234	125	100	111
121212	345	122	200	167
131213	134	111	139	89

- One can try to build a blackbox model in order to predict revenue based on different marketing chennels spends.

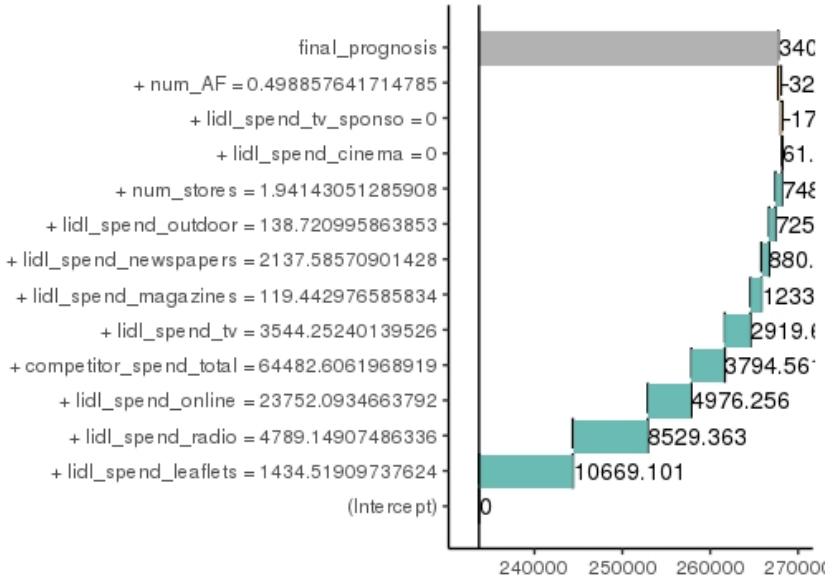
- How can one make sense out of that blackbox?

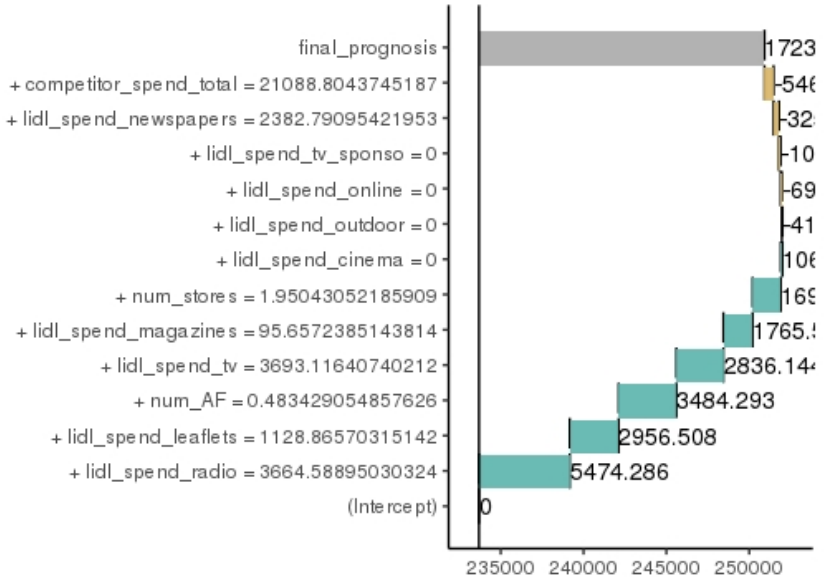
Situation 2. Marketing Mix Model. Channel contribution.



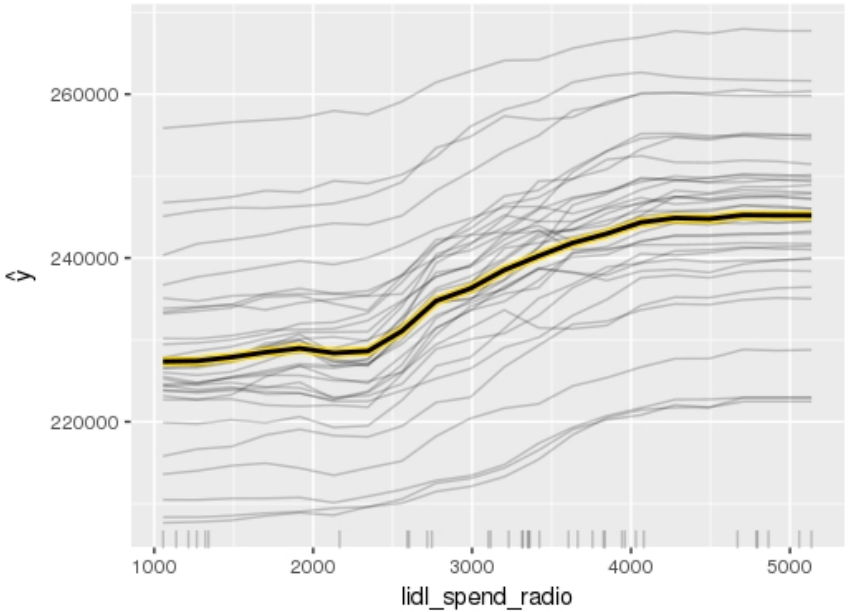


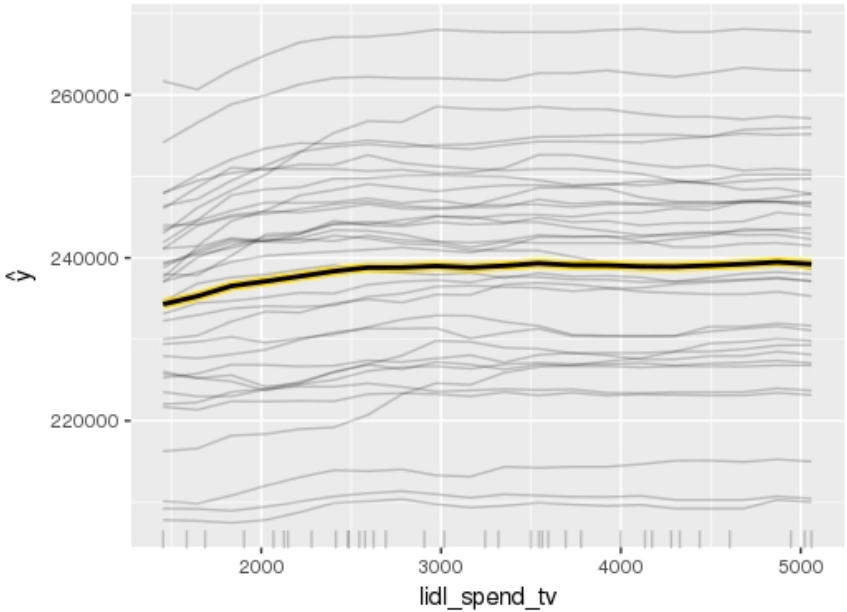
Situation 2. Marketing Mix Model. Channel contribution.



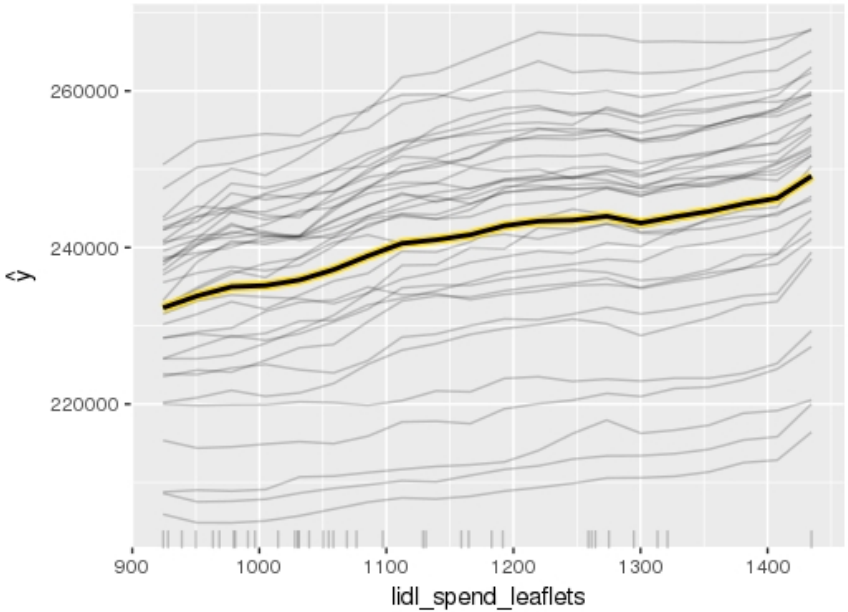


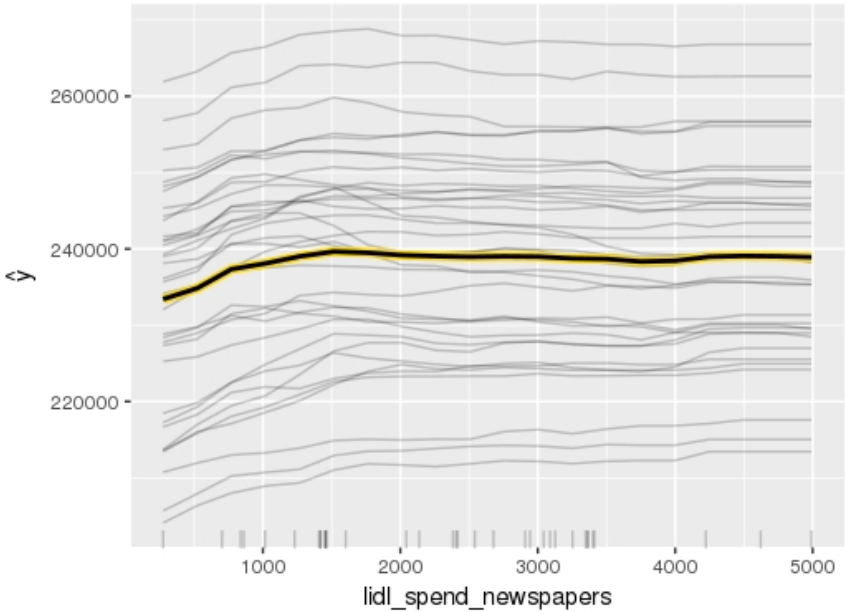
Situation 2. Channel contribution





Situation 2. Channel contribution.





Looking forward.

- XAI will gain more and more popularity in industry since
  - stakeholders and business leader demand explainability of ML models and results
  - doctors, lawyers will be use it to justify their own decisions
  - Fair ML and XAI will go hand in hand.
  - applications to image and test data

Future of XAI

- Accurate Models
- Trustworthy Models
- Natural Language Explanation
- Adversarial Use (misuse)
- Collaboration with Machines

Join the club

- https://www.linkedin.com/groups/8672810/ - my LinkedIn group where I post XAI related article.
- You can find here a lot of useful links and community interested in XAI.
- You can also contact me directly on LinkedIn.

Software

Python

- aequitas
- anchor
- ContrastiveExplanation (Foil Trees)
- eli5
- fairml
- L2X
- lime
- pyBreakDown
- PDPbox
- PyCEbox
- shap
- Skater
- tensorflow/model-analysis
- themis-ml
- treeinterpreter

R

- ALEPlot
- breakDown
- DALEX
- ExplainPrediction
- ICEbox
- iml
- lime
- live
- xgboostExplainer
- lightgbmExplainer

Thank you!