The AI/Data Science Professional

Transparency and Explainability

Acknowledgement of Country

The AI/Data Science Professional

Course Coordinator: Flora Salim

What's next...

RMIT UNIVERSITY

RMIT University acknowledges the people of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands we conduct the business of the University.

RMIT University respectfully acknowledges their Ancestors and Elders, past and present. RMIT also acknowledges the Traditional Custodians and their Ancestors of the lands and waters across Australia where we conduct our business.

Ngarara Place

The AI/Data Science Professional – Week 7

Transparency and Explainability

Transparency in AI

The ethical guidelines published by the EU Commission's High-Level Expert Group on AI (AI HLEG) in April 2019 states transparency as one of seven key requirements for the realisation of 'trustworthy AI', as also stated in the Commission's white paper on AI, published in February 2020.

"Transparency" is the single most common, and one of the key five principles emphasised in the vast number – a recent study counted 84 – of ethical guidelines addressing AI on a global level (Jobin et al., 2019).

Larsson, S. & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). https://doi.org/10.14763/2020.2.1469; https://policyreview.info/concepts/transparency-artificial-intelligence

Rooted in Trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Definition of Transparency (according to HLEG AI)

Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Transparency in AI: a multidisciplinary discussion

"fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections)" (AI HLEG, 2018, p. 6). Thus, the interplay between AI technologies and societal values – the applied ethics, social and legal norms – underscores the importance of combining social scientific and contributions from the humanities to computer scientifically based AI research (see Dignum, 2019).

Larsson, S. & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). https://doi.org/10.14763/2020.2.1469; https://policyreview.info/concepts/transparency-artificial-intelligence

The need for Transparency and Explainable AI (XAI)

- The problems of accountability as computing technologies becoming more complex and less intelligible (Helen Nissenbaum).
- The opacity in Machine Learning Systems (Jean Burrell, 2016) due to :
  - Trade secrets
  - Limited people with the knowledge of programming languages and ML
  - The complexity and high dimensionality of data for decision making no longer match human-scale reasoning
- Institutional transparency, public values, regulations
  - Customer's rights for explanations (GDPR Article 15(1))
  - Requirement for human in the loop (GDPR Article 22)
  - Requirement for algorithmic auditing (US Algorithmic Accountability act)

Source: Jake Goldenfein, 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019), Available at SSRN: https://ssrn.com/abstract=3445873, https://ovic.vic.gov.au/wp-content/uploads/2019/08/closer-to-the-machine-web.pdf p.45-65

Human in the loop

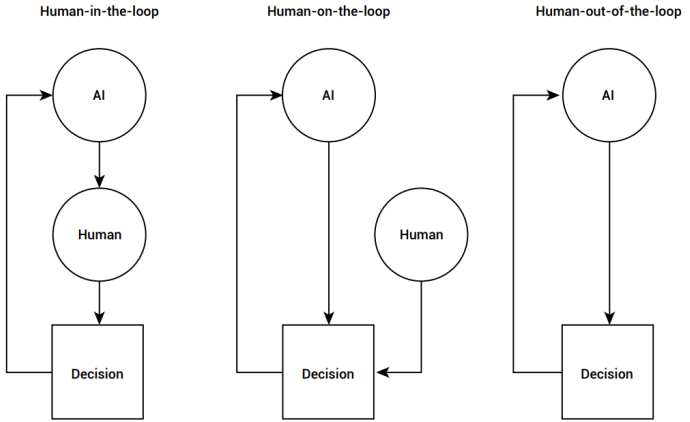


Figure 1

Three different kinds of human involvement with AI: human-in-the-loop – where an AI system provides information to a human in order for them to make a decision; human-on-the-loop – where a human supervises an AI system making a decision; and human-out-of-the-loop – where an AI system makes a decision without any human involvement.

Source: Jake Goldenfein, 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019), Available at SSRN: https://ssrn.com/abstract=3445873, https://ovic.vic.gov.au/wp-content/uploads/2019/08/closer-to-the-machine-web.pdf p.45-65

Transparency and Explainability

- Inscrutable autonomous decisions
- Policing and prediction
- AI in Healthcare and medicine
- Explainable and interpretable AI
- From Explainable AI to Human Centered AI

Inscrutable autonomous decisions

[CS: Nvidia Self Driving Car] In late 2016, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors. It showed the rising power of AI. The car didn't follow a single instruction provided by an engineer or programmer. Instead, it relied entirely on a Deep learning algorithm that had taught itself to drive by watching a human do it.

Nvidia AI Car Demo: https://youtu.be/-96BEoXJMs0

Discussion

- Why might people have concern about this AI system?
- What are the potential ethical issues with the design?
- Who would be held accountable for the ethical implications of the inscrutable algorithms they develop?

The Moral Machine experiment

The Game: https://www.moralmachine.net/

The Result: https://www.nature.com/articles/s41586-018-0637-6 . An online copy of the Nature article: https://www.americaninno.com/wp-content/uploads/2017/05/The-MM-Experiment.pdf

A debate on the paper: https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf

Inscrutable autonomous decisions

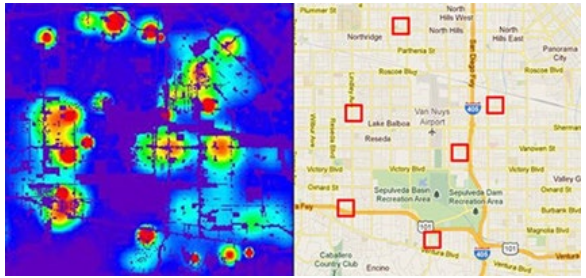
[CS: Public teacher employment evaluations] IA proprietary AI system was used by the Houston school district to assess the performance of their teaching staffs. The system used student test scores over time to assess the teachers' impact. The results were then used to dismiss teachers deemed ineffective by the system.

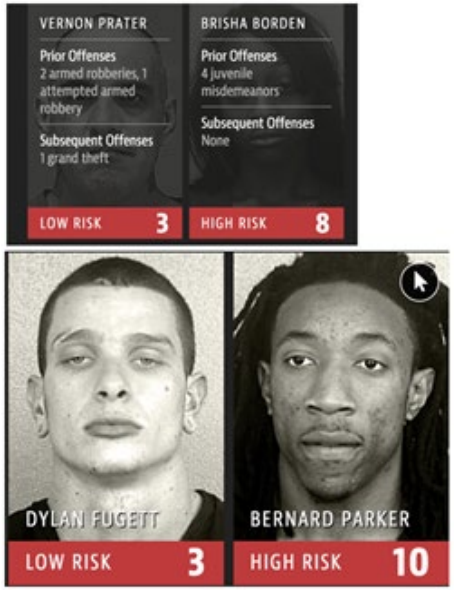
The teacher's union challenged the use of the AI system in court. As the algorithms used to assess the teacher's performance were considered proprietary information by the owners of the software, they could not be scrutinized by humans. This inscrutability was deemed a potential violation of the teachers' civil rights, and the case was settled with the school district withdrawing the use of the system.

Policing and prediction

Types of predictive policing

- Predicting crimes
- Predicting offenders
- Predicting perpetrators' identities
- Predicting victims of crime

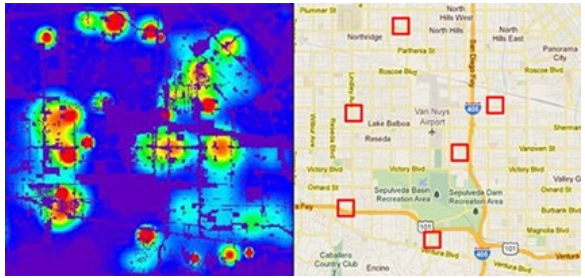




Policing and prediction

Data in use

- Historical Data
- Social Media Information
- Human mobility information



[CS: Predictive policing in the US] When data analysis company Palantir partnered up with the New Orleans police department, it began to assemble a database of local individuals to use as a basis for predictive policing. To do this it uses information gathered from social media profiles, known associates, license plates, phone numbers, nicknames, weapons, and addresses. Media reports indicated that this database covered around 1% of the population of New Orleans. After it had been in operation for six years there was a flurry of media attention over the “secretive” program. The New Orleans Police Department (NOPD) clarified that the program was not secret and had been discussed at technology conferences. But insufficient publicity meant that even some city council members were unaware of the program.

How Cops Are Using Algorithms to Predict Crimes

Discussion

- What are the insights and lessons learned from predictive policing?
- What are the potential ethical issues that arise with the advent of predictive policing?
- What are the potential positive and negative impact on the citizens and law enforcement bodies?

[CS: Predictive policing in Brisbane] One predictive policing tool has already been modelled to predict crime hotspots in Brisbane. Using 10 years of accumulated crime data, the system used 70% of the data to predict crime, with the researchers seeing if its predictions correlated with the remaining 30%. The results proved more accurate than existing models, with an improvement of 16% accuracy for assaults, 6% more accuracy for predicting unlawful entry, 4% better accuracy for predicting drug offences and theft, and 2% better for fraud. The Brisbane study used information from location-based app foursquare, and incorporated information from Brisbane and New York.

Extracted from CSIRO & Data61, Innovation, Department of Industry, and Australian Government Science. 2019. Artificial Intelligence: Australia's Ethics Framework [a Discussion Paper]

https://www.brisbanetimes.com.au/national/queensland/app-data-predicts-when-where-brisbane-criminals-will-strike-next-20181030-p50cui.html

S K Rumi, K Deng, F D Salim, Crime Event Prediction with Dynamic Features, EPJ Data Science, 2018, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0171-7

https://www.itnews.com.au/news/real-time-social-media-location-data-used-to-predict-crime-514419

LATEST NEWS | Next stop Uber: Transport Workers Union chief | Origin Energy to send electricity demand shifting AI bot national | Digital health agency gets its first CDO | Finance departments face new challenges | Qantas tackles double the number of IT projects

itnews | GOVERNMENT IT | SECURITY | FINANCE IT | TELCO | BENCHMARK AWARDS | LOG IN | SUBSCRIBE

Real-time social media location data used to predict crime

By Matt Johnston
Oct 24 2018
11:06AM

0 Comments

RMIT uses Foursquare check-ins to predict drug busts.

Researchers from RMIT have developed algorithms they say can comb through social media data to more accurately predict when and where different kinds of crime will occur.

10/24/2018 Crime stopped before it happens

THE AUSTRALIAN

Crime stopped before it happens

By **DAVID SWAN**, REPORTER
1:33PM OCTOBER 24, 2018 - 18 COMMENTS

Computers can now predict the time and place of crimes before they happen, thanks to new research by computer scientists from Melbourne's RMIT University.

In scenes that could have been lifted from the Tom Cruise film Minority Report, researchers took over 20,000 check-ins from location app Foursquare in Brisbane, and nearly 230,000 check-ins in New York City, and used the data to predict assaults, drug offences, theft and fraud.

brisbane times

NATIONAL QUEENSLAND CRIME

App data predicts when, where Brisbane criminals will strike next

By Toby Crockford
31 October 2018 — 10:15pm

Researchers have used historical crime figures, location app data and weather conditions to predict when and where Brisbane's criminals will strike next.

Melbourne university analysts used artificial intelligence to create an algorithm based on the check-ins from location app Foursquare and other historical Brisbane information.

TODAY'S TOP STORIES

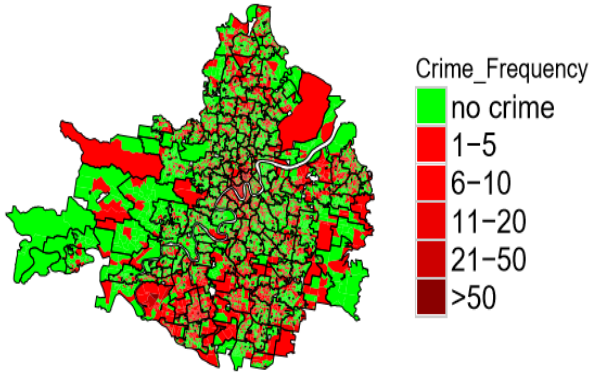
SECURITY
China uses the cloud to step up spying on Australian business
31 minutes ago

QUEENSLAND RAIL
Queensland Rail train drivers pocket tens of thousands in overtime

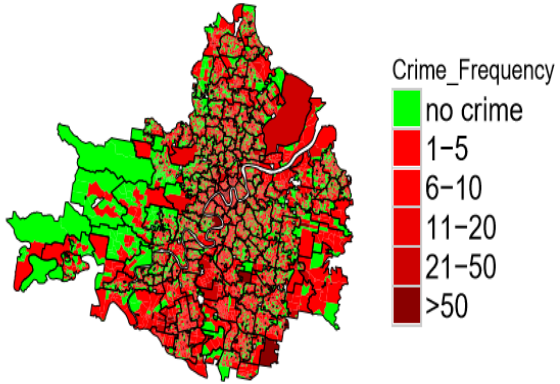
SUPERANNUATION
'Life-changing': New measures to help women leaving abusive relationships
49 minutes ago

CRIME
Homeless parents arrested over

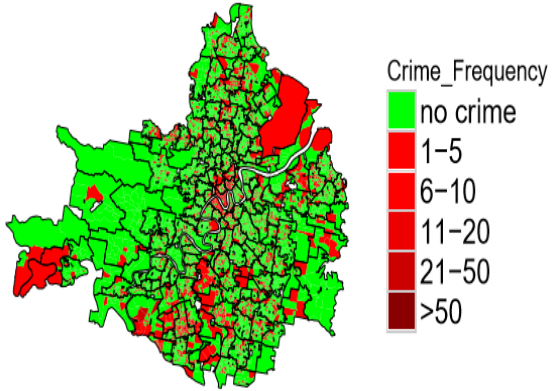
Brisbane Crime Hotspot Analysis



Drug Offence



Theft

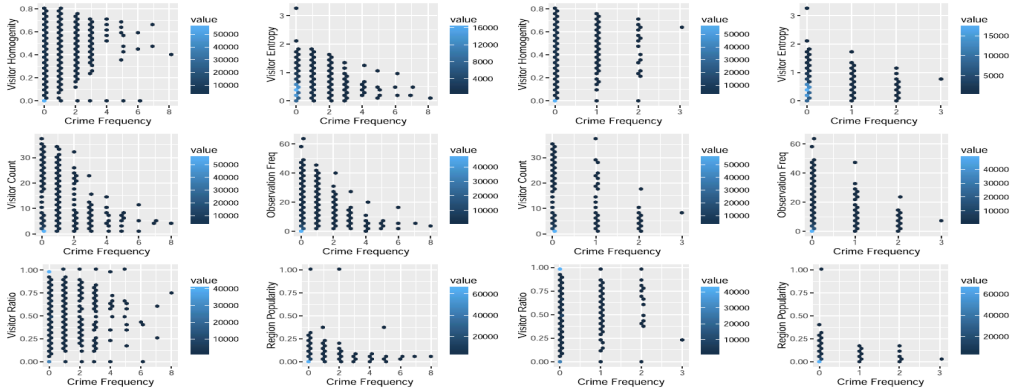


Assault

Rumi, S.K., Deng, K., Salim, F.D. (2018), Crime event prediction with dynamic features, EPJ Data Science (2018) 7: 43.

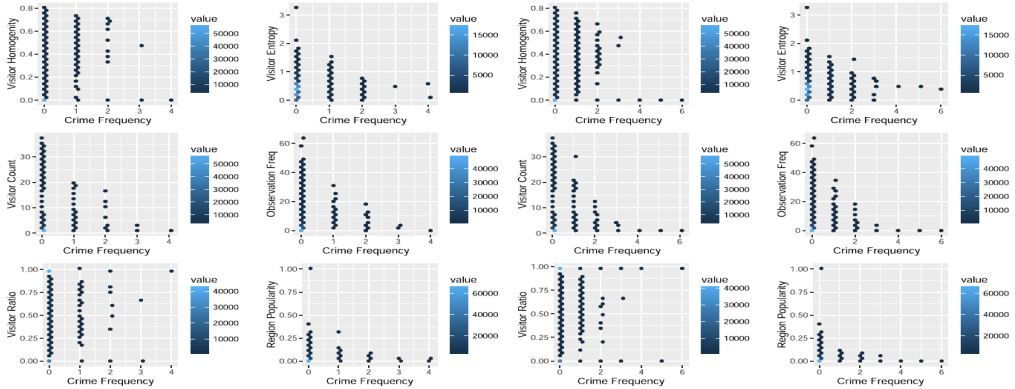
Dynamic feature analysis

Hexbin Plot between dynamic features & frequency of crime events



(a) Dynamic features vs. Theft.

(b) Dynamic features vs. Drug Offence.



(c) Dynamic features vs. Fraud.

(d) Dynamic features vs. Assault.

AUC comparison in Ensemble model

Time interval	Theft		Drug Offence		Assault		Fraud		Property Damage		Unlawful Entry		Traffic Offences		Unlawful use of Motorvehicle	
	wod	wd	wod	wd	wod	wd	wod	wd	wod	wd	wod	wd	wod	wd	wod	wd
[9 – 12)	0.72	0.72	0.51	0.6	0.5	0.5	0.74	0.75	0.5	0.55	0.5	0.5	0.53	0.54	0.5	0.5
[12 – 15)	0.7	0.71	0.59	0.57	0.5	0.52	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.51
[15 – 18)	0.64	0.68	0.61	0.66	0.5	0.55	0.52	0.7	0.5	0.55	0.51	0.54	0.53	0.53	0.53	0.53
[18 – 21)	0.52	0.59	0.55	0.52	0.5	0.56	0.61	0.73	0.5	0.53	0.5	0.5	0.5	0.58	0.5	0.5
[21 – 24)	0.58	0.64	0.65	0.74	0.61	0.64	0.5	0.5	0.52	0.6	0.5	0.51	0.5	0.5	0.56	0.6

wod: without dynamic features; wd: with dynamic features.

To learn more, read the paper: https://rdcu.be/9Esr

AI in Healthcare: The Current Status and How

- The state of artificial intelligence in medicine ---- by Stanford Medicine

- How doctors can help AI to revolutionize medicine ---- by Greg Corrado, Co-founder of Google Brain and Principal Scientist at Google

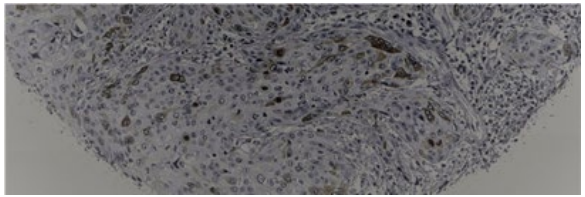
AI in Healthcare and Medicine

Common applications:

- Diagnose and reducing error
- Developing new medicines
- Streamlining patient experience
- Mining and managing medical data
- Robot-assisted surgery

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

Diagnose and reducing error

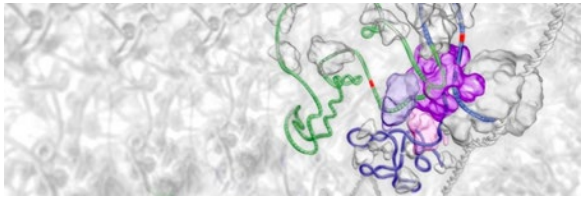


Examples:

- PATHAI: more accurate cancer diagnosis
- BUOY HEALTH: an intelligent symptom checker
- ENLITIC: deep learning for actionable insights
- FREENOME: earlier cancer detection
- BETH ISRAEL DEACONESS MEDICAL CENTER
- ZEBRA MEDICAL VISION: radiology assistant

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

Developing new medicines

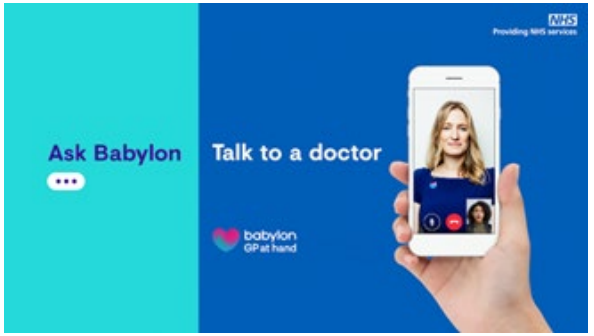


Examples:

- BIOXCEL THERAPEUTICS: biopharmaceutical development
- BERG HEALTH: treating rare disease
- XTALPI: ai, cloud-based digital drug discovery
- ATOMWISE: neural network for clinical trials
- DEEP GENOMICS: finding better candidates for developmental drugs
- BENEVOLENTAI: deep learning for targeted treatment

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

Streamlining patient experience



Examples:

- OLIVE: automating healthcare's most repetitive processes
- QVENTUS: real-time patient flow optimization
- Babylon Health: increasing access to healthcare
- CLOUDMEDX: using ml for a better patient journey
- Cleveland Clinic: personalized healthcare plans
- Johns Hopkins Hospital: faster hospital visits

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

Mining and managing medical data.

Examples:

- TEMPUS: personalized health
- KENSCI: hospital risk prediction
- PROSCIA: medical image
- H2O.AI: the health system
- Google Deepmind Health: alerting doctors
- ICARBONX: data and the 'digital life'

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

Robot-assisted surgery

examples:

- Vicarious Surgical: virtual reality-enabled robotics
- Auris Health: endoscopy
- ACCURAY: cancer treatment
- INTUITIVE: pioneering robotic surgery
- Carnegie Mellon University: heart therapy
- Microsure: improving surgical precision
- Mazor Robotics: spinal surgery

Source from https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

[CS: Predicting coma outcomes] A program in China that analyses the brain activity of coma patients was able to successfully predict seven cases in which the patients went on to recover, despite doctor assessments giving them a much lower chance of recovery. This AI system took examples of previous scans, and was able to detect subtle brain activity and patterns to determine which patients were likely to recover and which were not. One patient was given a score of just seven out of 23 by human doctors, which indicated a low probability of recovery, but the AI system gave him over 20 points. He subsequently recovered. His life may have been saved by the AI system.

If this AI system lives up to its potential, then this kind of tool would be of immense value in saving human lives by spotting previously hidden potential for recovery in coma patients—those given high scores can be kept on life support long enough to recover.

Discussion: what about people with low scores?

AI in Insurance

Source from https://towardsdatascience.com/how-are-insurance-companies-implementing-artificial-intelligence-ai-aaf845fce6a7

ML Interpretability

What's next...

Interpretability: concepts

Interpretability is the degree to which a human can understand the cause of a decision. [1]

Interpretability is the degree to which a human can consistently predict the model's result. [2]

[1] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).
[2] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

If ML works well, why don't we just trust the model?

"The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks."
-- (Doshi-Velez and Kim 2017)

The Machine Learning Black Box

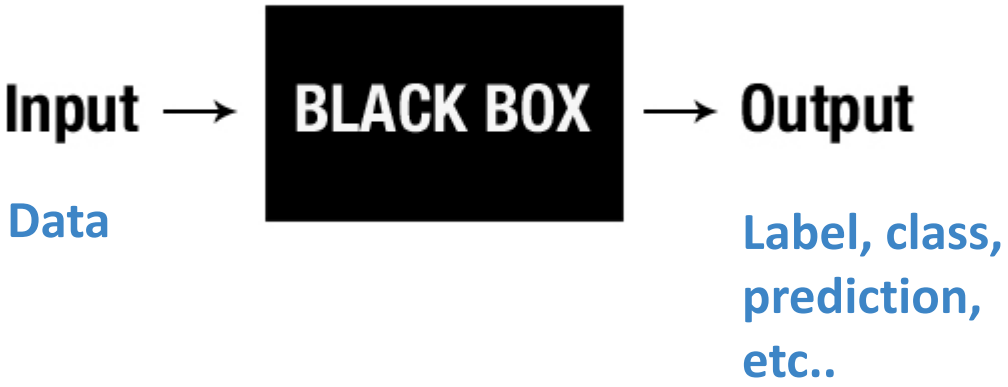


Image source from https://www.kdnuggets.com/2017/04/ai-machine-learning-black-boxes-transparency-accountability.html

The Machine Learning Black Box

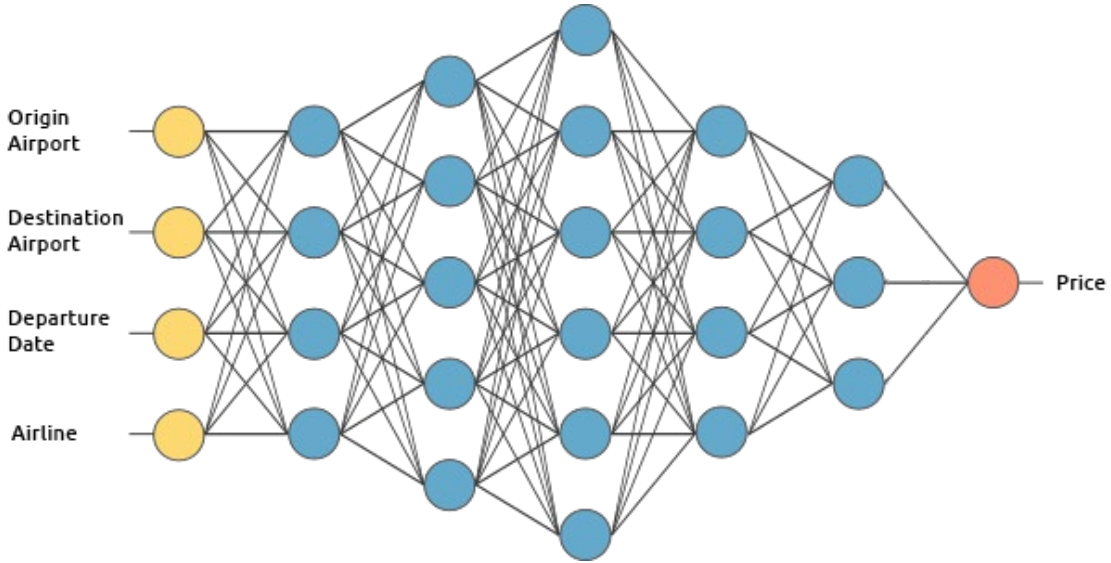


Image source from https://www.houseofbots.com/news-detail/11733-4-know-how-deep-learning-works-heres-a-quick-guide-for-all-engineer

The Machine Learning Black Box

Sneeze
Weight
Age
Headache
No fatigue
...

The patient is having a flu

The ML technology is getting better, but ……

Trust:
How can we trust the predictions are correct?

Interpret:
How can we understand and predict the behavior?

Detect & Improve:
How do detect errors in the model and improve it?

Interpretability: importance

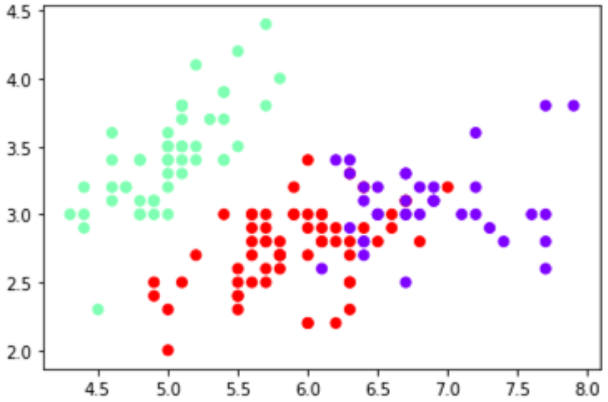
- human curiosity and learning
- finding meaning in the world
- the goal of science
- safety measures
- detecting bias
- manage social interactions
- debugged and audited

Taxonomy of Interpretability

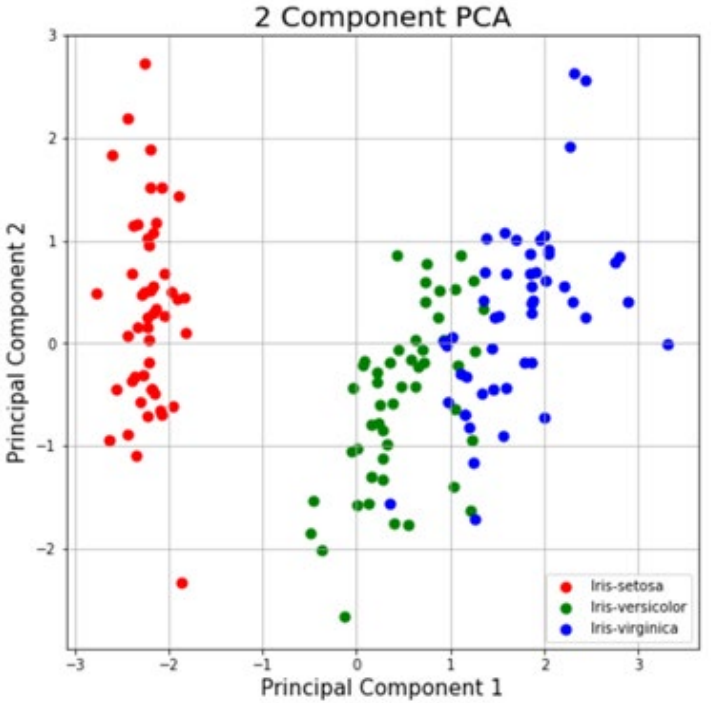
- Interpretable model (Intrinsic)
- Model specific vs. model agnostic
- Local vs. Global
- Type/style of the interpretations

Taxonomy of Interpretability

- Pre-Model
- In-Model
- Post-Model



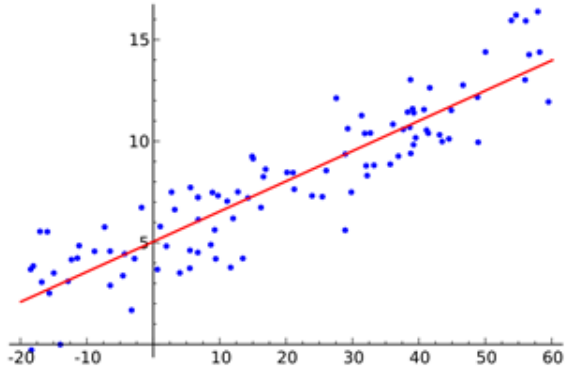
Clustering - Iris data



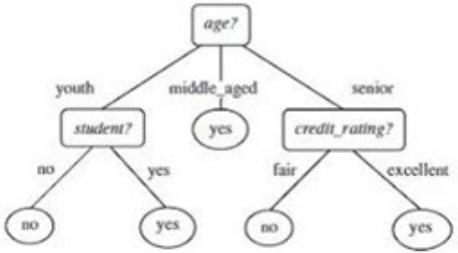
PCA - Iris data

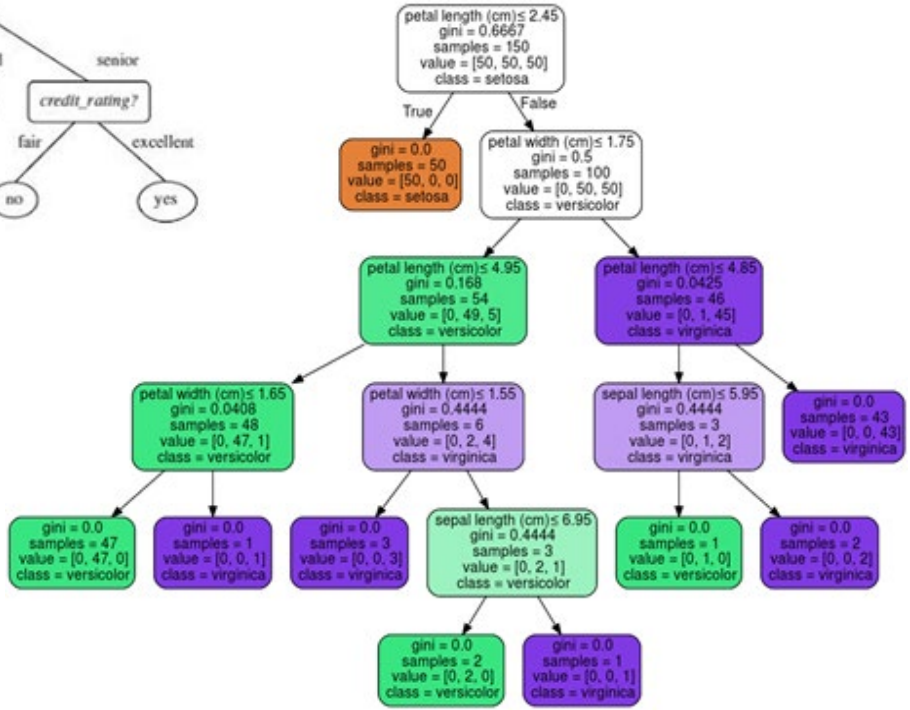
Taxonomy of Interpretability

- Pre-Model
- In-Model (Intrinsic)
- Post-Model



Simple linear regression

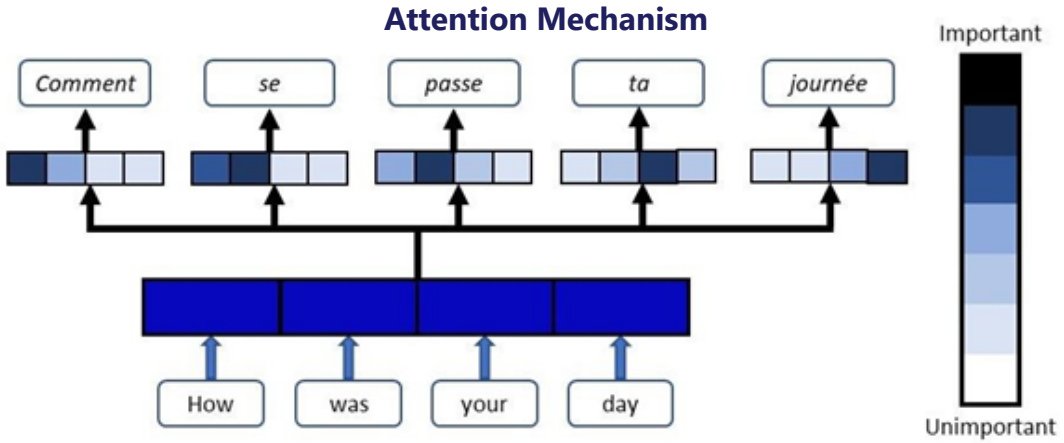


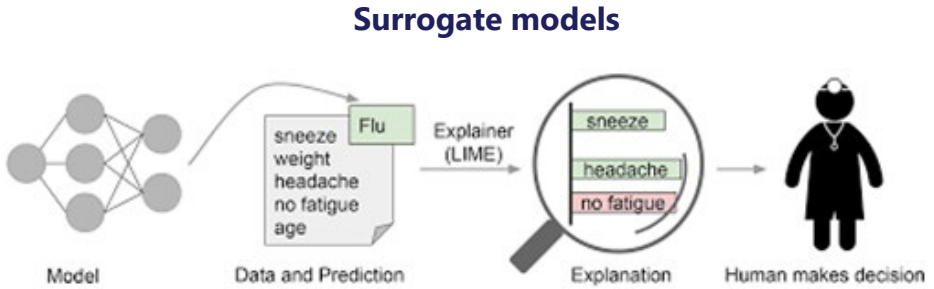


Decision tree and Regression tree

Taxonomy of Interpretability

- Pre-Model
- In-Model
- Post-Model (Post Hoc)
  - Model-specific
  - Model-Agnostic





Forms of Explanation

- Feature summary, e.g., feature importance
- Model internals, e.g., weights in linear models
- Data points, e.g., data point examples
- Surrogate intrinsically interpretable model

Scope of Interpretability

- Algorithm Transparency
- Global Model Interpretability
  - Holistic level
  - Modular level
- Local Model Interpretability
  - Single prediction
  - Group of predictions

What is a good explanation?

Evaluation of Interpretability

- Application level evaluation (real task)
- Human level evaluation (simple task)
- Function level evaluation (proxy task)

Properties of Explanations Methods

- **Expressive Power**: the language structure it generates (e.g. if/then/else, decision trees, etc)
- **Translucency**: how much the explanation method relies on looking into the machine learning model, like its parameters (e.g. linear regression vs. generative models)
- **Portability**: the range of machine learning models with which the explanation method can be used. Methods with a low translucency have a higher portability because they treat the machine learning model as a black box (e.g. surrogate models)
- **Algorithmic Complexity**: describes the computational complexity of the method that generates the explanation

Properties of Individual Explanations

- Accuracy: How well does an explanation predict unseen data?
- Fidelity: How well does the explanation approximate the prediction of the black box model?
- Consistency: How much does an explanation differ between models that have been trained on the same task and that produce similar predictions?
- Stability: How similar are the explanations for similar instances (with the same model)?
- Comprehensibility: How well do humans understand the explanations?
- Certainty: Does the explanation reflect the certainty of the machine learning model?
- Importance: How well does the explanation reflect the importance of features or parts of the explanation?
- Novelty: Does the explanation reflect whether a data instance to be explained comes from a "new" region far removed from the distribution of training data?
- Representativeness: How many instances does an explanation cover?

Human-friendly Explanations

- **Explanations are contrastive**
- **Explanations are selected**
- **Explanations are social**
- **Explanations focus on the abnormal**
- **Explanations are truthful**
- **Good explanations are consistent with prior beliefs of the explainee**
- **Good explanations are general and probable**

Discussion

Assume we have a dataset of test situations between teachers and students. Students attend a course and pass the course directly after giving a presentation. The teacher has the option to additionally ask the student questions to test their knowledge. Students who cannot answer these questions will fail the course. Students can have different levels of preparation, which translates into different probabilities for correctly answering the teacher's questions (if they decide to test the student). We want to predict whether a student will pass the course and explain our prediction. The chance of passing is 100% if the teacher does not ask any additional questions, otherwise the probability of passing depends on the student's level of preparation and the resulting probability of answering the questions correctly.

Scenario 1: The teacher usually asks the students additional questions (e.g. 95 out of 100 times). A student who did not study (10% chance to pass the question part) was not one of the lucky ones and gets additional questions that he fails to answer correctly. **Why did the student fail the course?**

Scenario 2: The teacher rarely asks additional questions (e.g. 2 out of 100 times). For a student who has not studied for the questions, we would predict a high probability of passing the course because questions are unlikely. Of course, one of the students did not prepare for the questions, which gives him a 10% chance of passing the questions. He is unlucky and the teacher asks additional questions that the student cannot answer and he fails the course. **Why did the student fail the course?**

Interpretable models

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

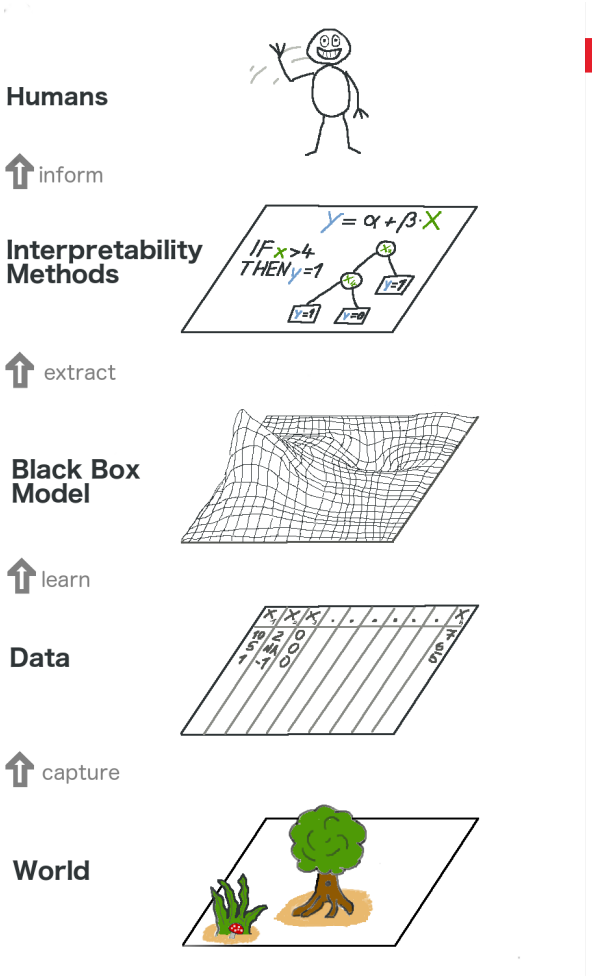
An overview of the interpretable model types and their properties.

- A model is **linear** if the association between features and target is modelled linearly.
- A model with **monotonicity** constraints ensures that the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature: An increase in the feature value either always leads to an increase or always to a decrease in the target outcome.
- **Monotonicity** is useful for the interpretation of a model because it makes it easier to understand a relationship.
- Some models can automatically include **interactions** between features to predict the target outcome.
- Some models handle only regression, some only classification, and still others both. From this table, you can select a suitable interpretable model for your task, either regression (regr) or classification (class).

Model Agnostic methods

Desirable aspects of a model-agnostic explanation system are (Ribeiro, Singh, and Guestrin 2016):

- **Model flexibility:** The interpretation method can work with any machine learning model, such as random forests and deep neural networks.
- **Explanation flexibility:** You are not limited to a certain form of explanation. In some cases it might be useful to have a linear formula, in other cases a graphic with feature importances.
- **Representation flexibility:** The explanation system should be able to use a different feature representation as the model being explained. For a text classifier that uses abstract word embedding vectors, it might be preferable to use the presence of individual words for the explanation.



Model Agnostic methods

Examples:

- LIME (Local interpretable model-agnostic explanations)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

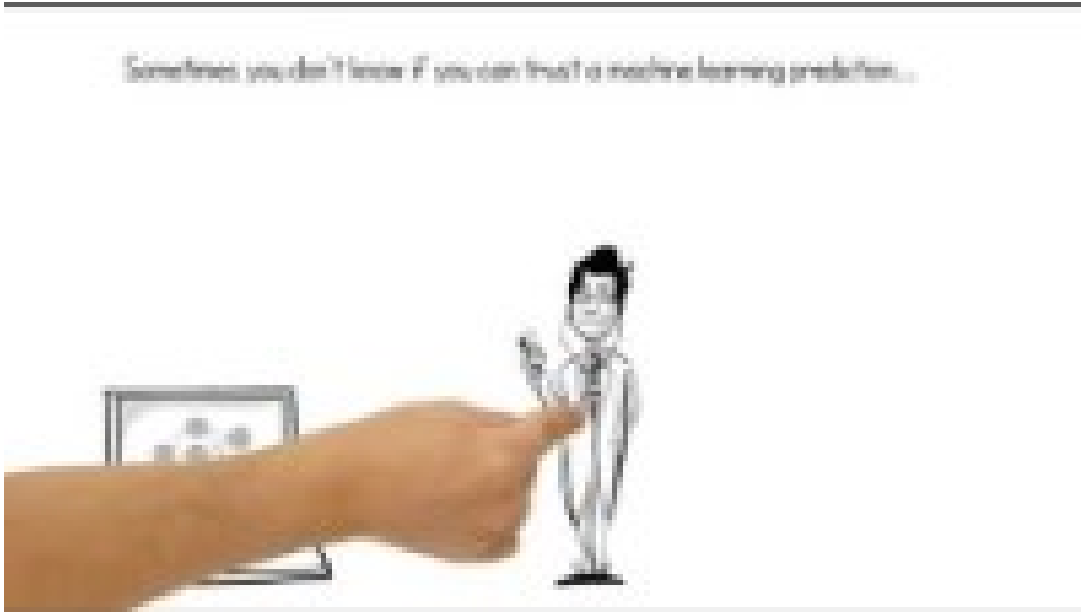
- SHAP (SHapley Additive exPlanations)

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017

LIME (Local Interpretable Model-Agnostic Explanations)

An algorithm that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model.

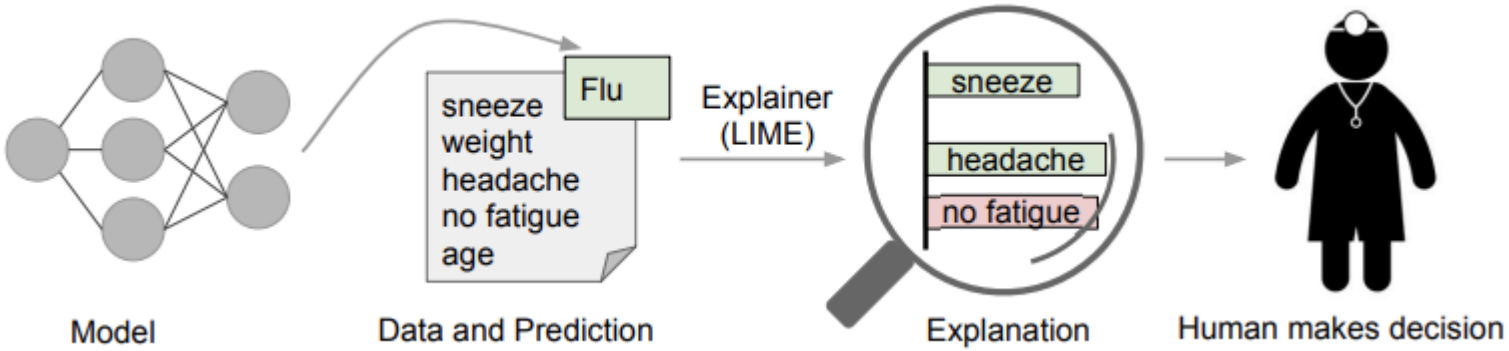
LIME



Video: https://youtu.be/hUnRCxnydCc
https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/
Paper: http://arxiv.org/abs/1602.04938

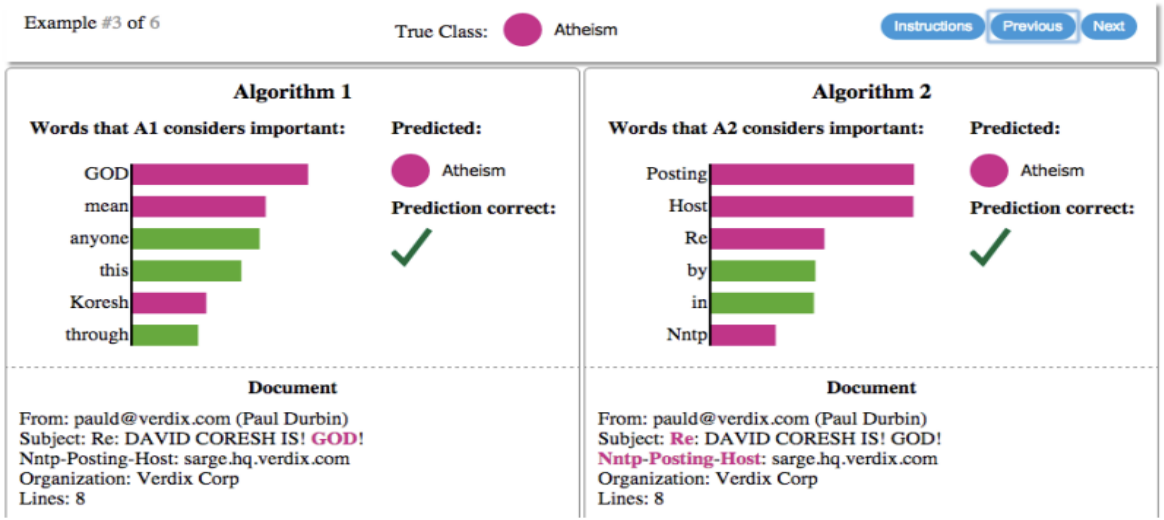
How a local explanation can help

To aid decision making

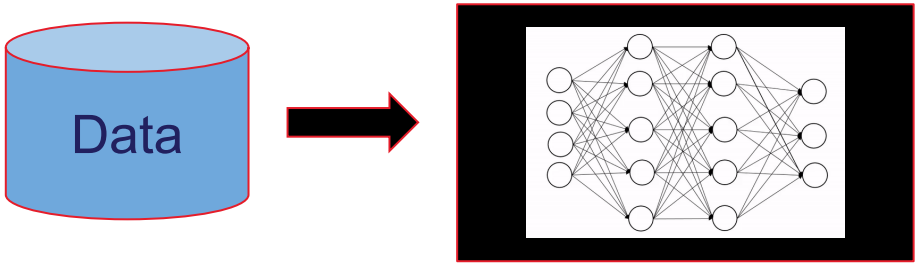


How a local explanation can help

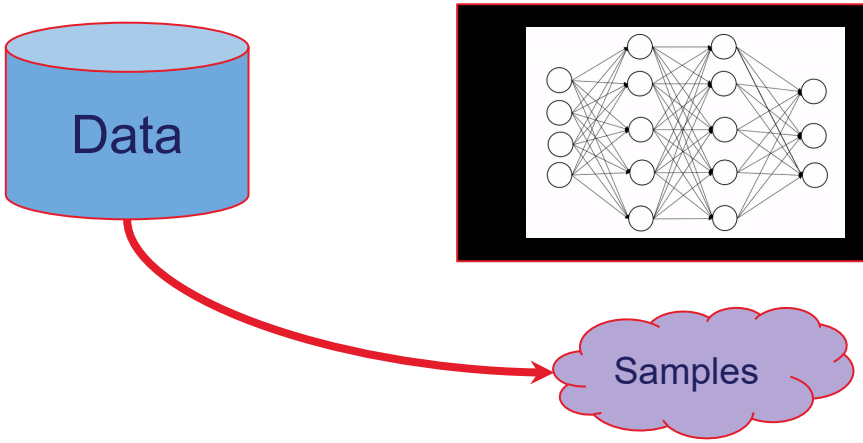
Which model to trust?



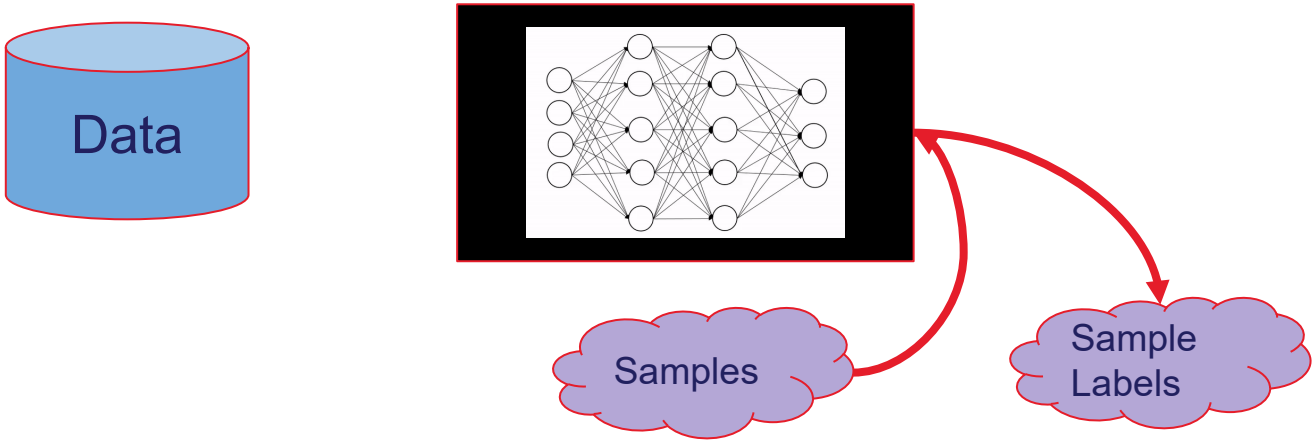
General Surrogate Model



General Surrogate Model

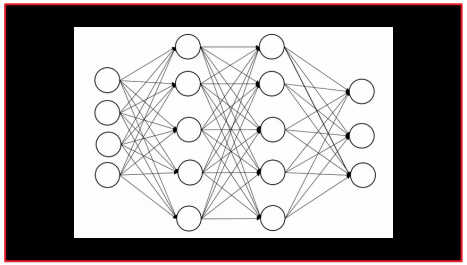


General Surrogate Model



General Surrogate Model

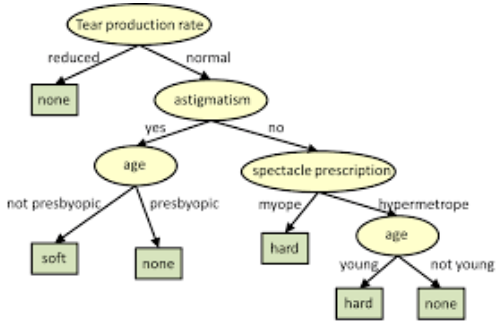
Data



New dataset

Samples

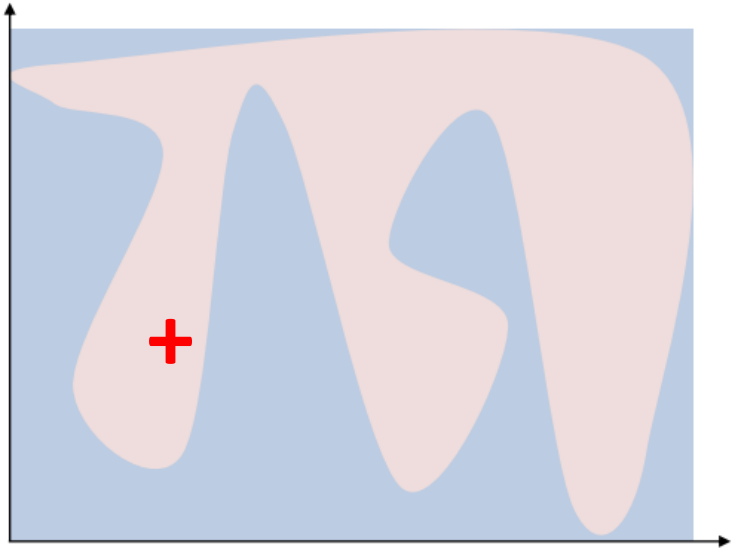
Sample Labels



How does LIME work?

1. Pick a model class interpretable by humans
2. Use that to locally model/learn/approximate the global Blackbox model

Simple linear models, Logistic regression, decision tree, etc..

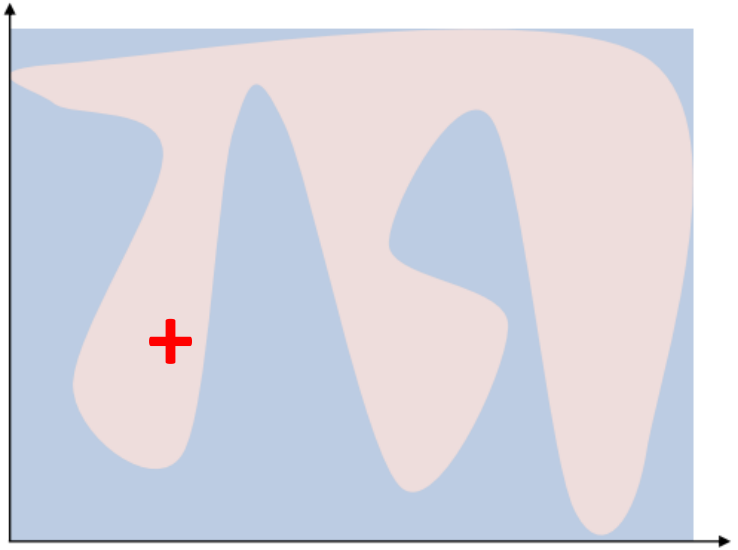


How does LIME work?

1. Pick a model class interpretable by humans
2. Use that to locally model/learn/approximate the global Blackbox model

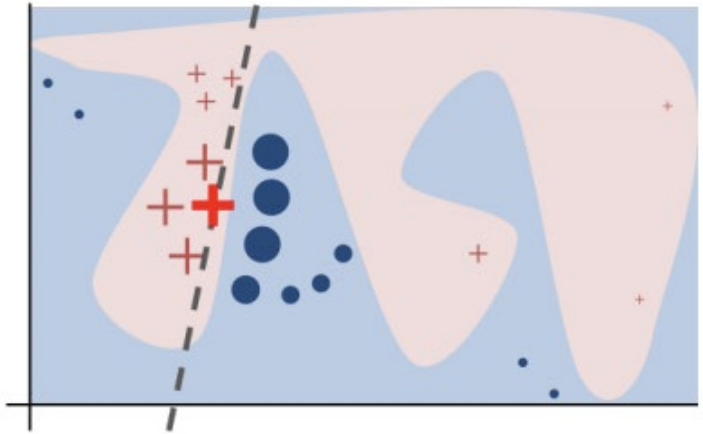
Simple model could be globally bad, but as long as it's locally good

Simple linear models, Logistic regression, decision tree, etc..



How does LIME work?

1. Sample around the instance we need to explain
2. Query the blackbox for the prediction of each sample
3. Weighted samples according to each of its distance to the instance
4. Choose and learn new interpretable (simple) model on weighted samples (and the corresponding prediction from the blackbox)
5. Use the learned simple model to explain



How does LIME work?

https://www.youtube.com/watch?v=A0JS_GP_mxY

LIME Pros and cons

Pros:

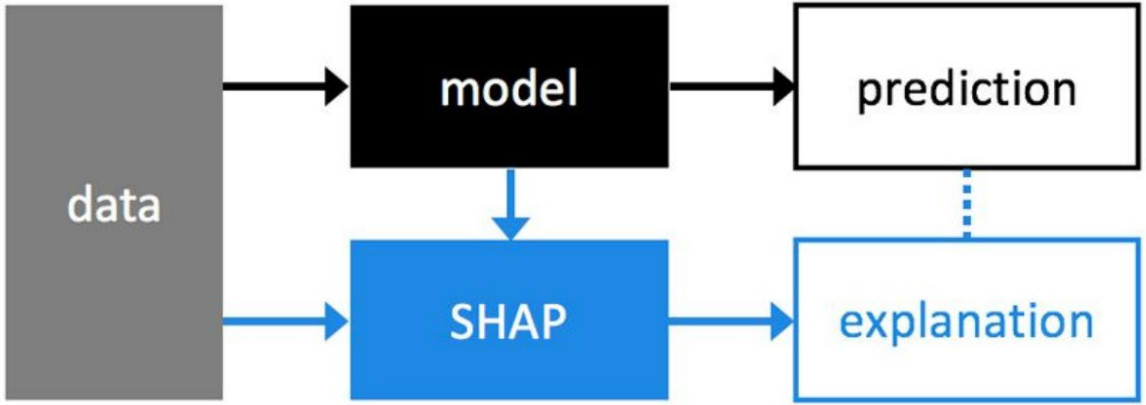
- Model agnostic
- A general framework
- Computation complexity

Cons:

- Instability of explanation (Explanations of similar examples might be totally different)
- Definition of neighbourhood
- Low fidelity with different sampling and label/data shift
- Hyperparameter setting (Explanation depends on the choice of LIME hyperparameters)
- Doesn’t work out on all models

SHAP (SHapley Additive exPlanations)

- To unify various explanation methods: model-agnostic or model-specific
- Based on the game theory, *Shapley Values*, by Scott Lundberg



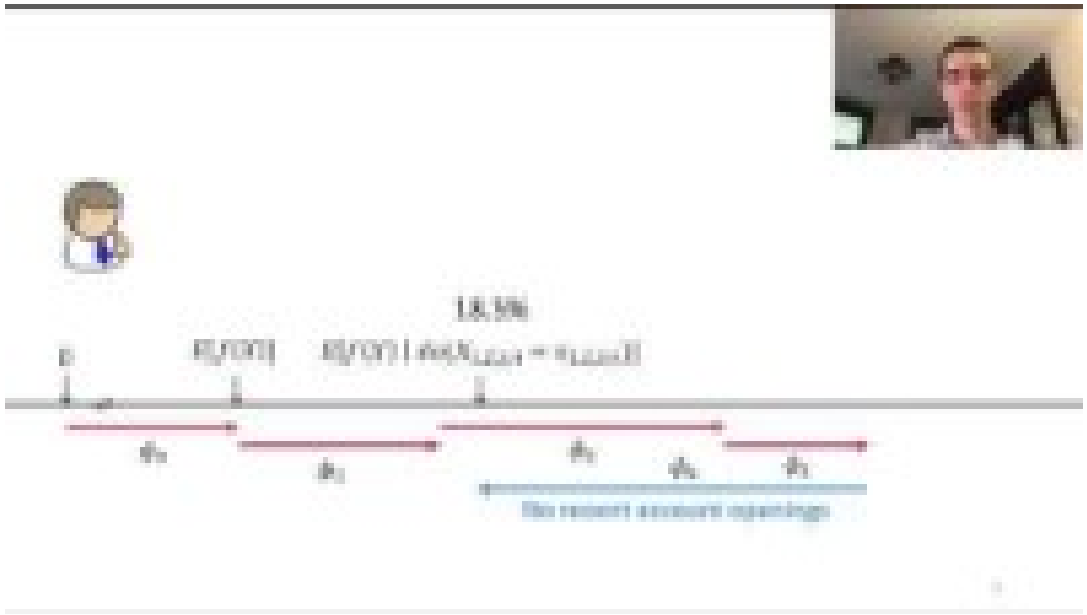
Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017

https://github.com/slundberg/shap

SHAP (SHapley Additive exPlanations)

- Produce explanations at the level of individual inputs
- Traditional feature importance tells us features that are most important across the population
- Individual-level SHAP pinpoints the features that are most impactful for each instance / object / user

SHAP (SHapley Additive exPlanations)



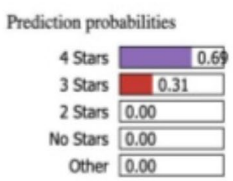
https://youtu.be/-taOhqkiuIo

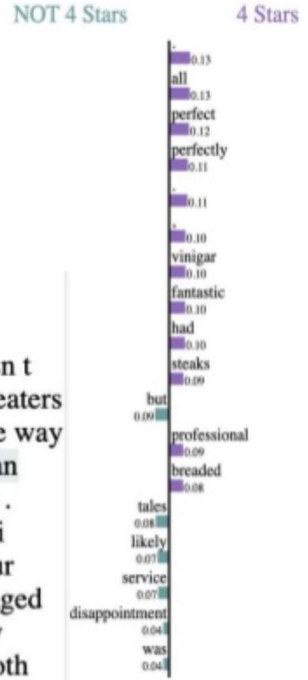
Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017

LIME vs SHAP

SHAP explainer: CNN

LIME: NLP Fasttext





Text with highlighted words

we had a late 8:30pm reservation . the restaurant wasn t busy but got busier at about 10pm , likely after the theaters let out . bonnie , our server was fantastic . loved the way she said y all ! great attentive service , professional an efficent . started out with a bowl of the lobster bisque . big chunks of lobster , hot and yummy . the calamari wasn t great . it was breaded but soggy . that was our only disappointment . i had the porcini rib eye with aged balsamic vinigar which was perfect . delicious . my husband had the tenderloin with two lobster tales . both steaks were perfectly cooked . sides of wild mushrooms and aspagus . cheesecake for dessert was yummy and nicely presented .

LIME vs SHAP

LIME

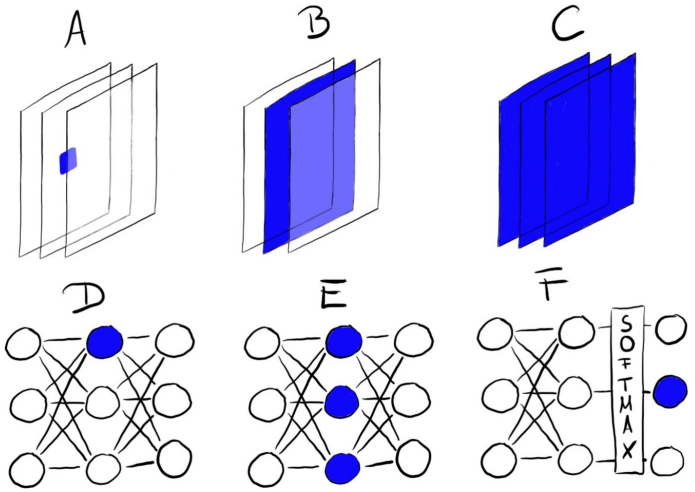
- Faster
- Use LIME for single prediction explanation

SHAP

- Slower, as more exhaustive
- Use SHAP to obtain summary plots and dependence plots
- Use SHAP for entire model (or single variable) explanation

NN interpretation

- Feature visualization
- Disentangled learning
- Adversarial Learning



Example based explanations

- Counterfactual explanations tell us how an instance has to change to significantly change its prediction. By creating counterfactual instances, we learn about how the model makes its predictions and can explain individual predictions.
- Adversarial examples are counterfactuals used to fool machine learning models. The emphasis is on flipping the prediction and not explaining it.
- Prototypes are a selection of representative instances from the data and criticisms are instances that are not well represented by those prototypes.
- Influential instances are the training data points that were the most influential for the parameters of a prediction model or the predictions themselves. Identifying and analysing influential instances helps to find problems with the data, debug the model and understand the model's behavior better.
- k-nearest neighbours model: An (interpretable) machine learning model based on examples.

What to do this week?

- Read materials
- Work on Task 1 (Case Study), due Monday 3pm

References

1.Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

2.Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017).

3.Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. “Examples are not enough, learn to criticize! Criticism for interpretability.” Advances in Neural Information Processing Systems (2016).

4. Sokol, K. and Flach, P., 2020, January. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 56-67). https://arxiv.org/ftp/arxiv/papers/1912/1912.05100.pdf