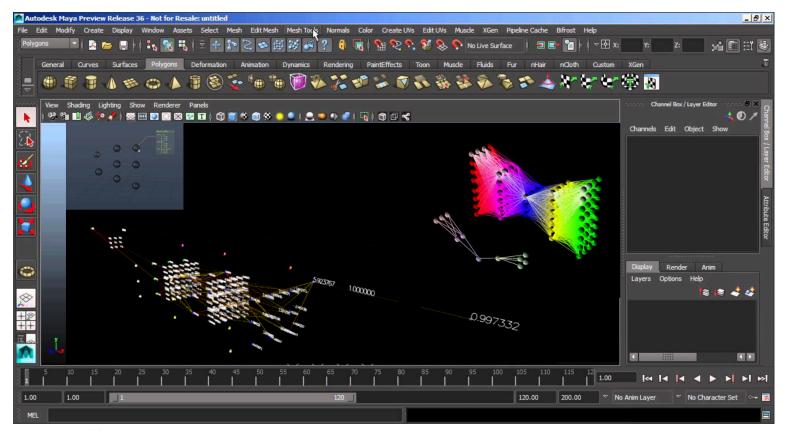
Applications

Debugging Artificial Neural Networks – Industry Agnostic



Challenge: Designing Artificial Neural Network architectures requires lots of experimentation (i.e., training phases) and parameters tuning (optimization strategy, learning rate, number of layers...) to reach optimal and robust machine learning models.

AI Technology: Artificial Neural Network

XAI Technology: Artificial Neural Network, 3D Modeling and Simulation Platform For AI



Explaining Visual Question Answering – Industry Agnostic

Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win? A: 197 Visual QA



Q: How symmetrical are the white bricks on either side of the building? A: very

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest guarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the guarterback who was 38 in Super Bowl XXXIII? A: John Elway

Challenge: What is the robustness of Visual Question Answering models? What is the impact of semantics?

AI Technology: Artificial Neural Networks.

XAI Technology: Integrated Gradients



Neural Programmer (2017) model **33.5%** accuracy on WikiTableQuestions

Kazemi and Elgursh (2017) model. 61.1% on VQA 1.0 dataset (state of the art = 66.7%)

Yu et al (2018) model. 84.6 F-1 score on SQuAD (state of the art)

Q: How symmetrical are the white bricks on either side of the building? A: very

Q: How asymmetrical are the white bricks on either side of the building? A: very

Q: How big are the white bricks on either side of the building? A: very

Q: How fast are the bricks speaking on either side of the building? A: very

What is the **man** doing? \rightarrow What is the **tweet** doing? How many **children** are there? → How many **tweet** are there?

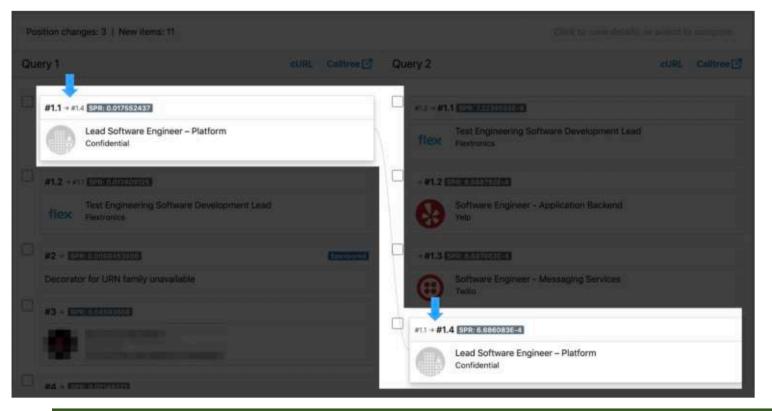
VQA model's response remains the same 75.6 of the time on questions that it originally answered correctly

Relevance Debugging and Explaining – Industry Agnostic

Challenge: A Machine Learning system can fail in many different points e.g., data features selection, construction, inconsistencies. How to debug bad performance in machine learning models and prediction?

AI Technology: Artificial Neural Networks.

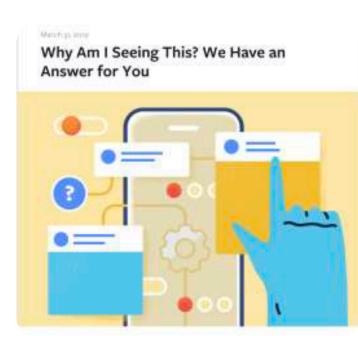
XAI Technology: Model / Prediction comparison

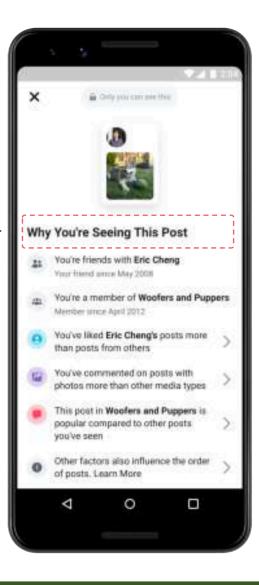


Source: Explainable AI in Industry. KDD 2019 Tutorial. Daniel Qiu, Yucheng Qian



Explaining Recommendation–Social Media





Challenge: How to establish trust between Social Media and their users? Explaining post / news recommendation is crucial for users to engage with content providers.

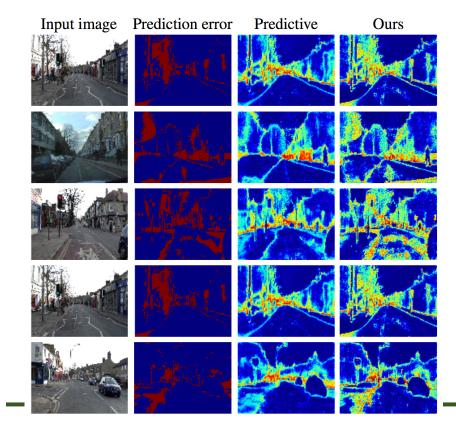
AI Technology: Artificial Neural Networks.

XAI Technology: Recommendation-based

KDD 2019 Tutorial on Explainable AI in Industry - 5https://sites.google.com/view/kdd19-explainable-ai-tutorial

Obstacle Identification Certification (Trust) - Transportation





THALES

Challenge: Public transportation is getting more and more selfdriving vehicles. Even if trains are getting more and more autonomous, the human stays in the loop for critical decision, for instance in case of obstacles. In case of obstacles trains are required to provide recommendation of action i.e., go on or go back to station. In such a case the human is required to validate the recommendation through an explanation exposed by the train or machine.

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and semantic segmentation.

XAI Technology: Deep learning and Epistemic uncertainty







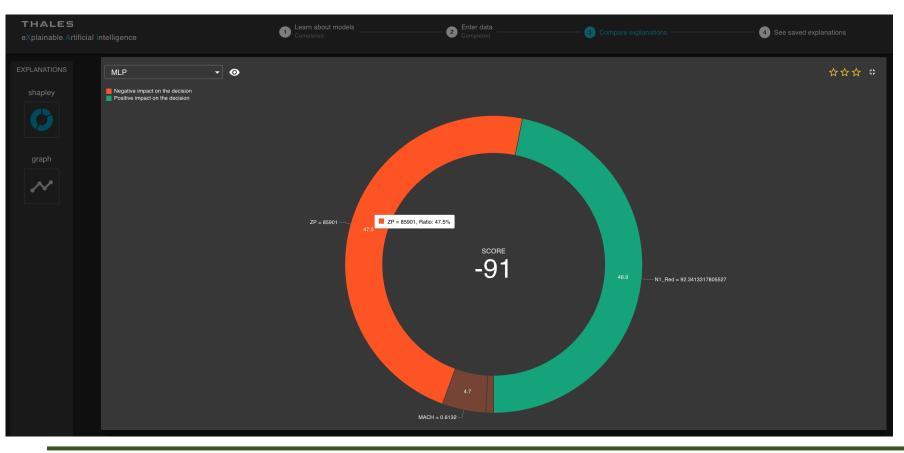
Explaining Flight Performance- Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values

THALES



Explainable On-Time Performance - Transportation

KLM / Transavia Flight Delay Prediction

PLANE INFO	ARRIVAL			TURNAROUND			DEPARTURE					
Status / Aircraft	Flight	ETA	Status	Delay Code	Gate	Slot	Progress	Milestones	Flight	ETA	Status	Delay Code
🕑 urtwet 🗸	4567	18:30	Scheduled	-	345345	1			5678	19:00	Scheduled	-
\rm idsfew 🗸	4567	18:30	Delayed	ABC, DEF, GHI	345345	1			5678	19:00	Delayed	ABC, DEF, GHI
🕑 pssjdb 🐱	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
Ø kshdbs ✓	4567	-	Cancelled	ABC, DEF, GHI	-	-			5678	-	Cancelled	ABC, DEF, GH
9 wwwdfs∨	4567	18:35	Delayed	ABC, DEF, GHI	345345	1			5678	19:00	Delayed	ABC, DEF, GH
0 pdigbs v	4567	18:30	Delayed	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
😔 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🛛 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🕑 <u>aedbsc</u> 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🕑 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🛛 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🕑 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🛛 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🕑 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH
🛛 aedbsc 🗸	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1			5678	19:00	Scheduled	ABC, DEF, GH

Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, Huajun Chen: Knowledge-Based Transfer Learning Explanation. KR 2018: 349-358

Nicholas McCarthy, Mohammad Karzand, Freddy Lecue: Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines: AAAI 2019 **Challenge:** Globally 323,454 flights are delayed every year. Airlinecaused delays totaled 20.2 million minutes last year, generating huge cost for the company. Existing in-house technique reaches 53% accuracy for **predicting flight delay**, does not provide any time estimation (in <u>minutes</u> as opposed to True/False) and is unable to capture the underlying reasons (explanation).

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain delays by comparing with historical cases.

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs



Model Explanation for Sales Prediction - Sales



Challenge: How to predict and explain upsell / churn for a company?

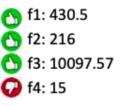
AI Technology: Artificial Neural Networks.

XAI Technology: Features importance (contribution, influence), LIME.

Company: CompanyX



Top Feature Contributor



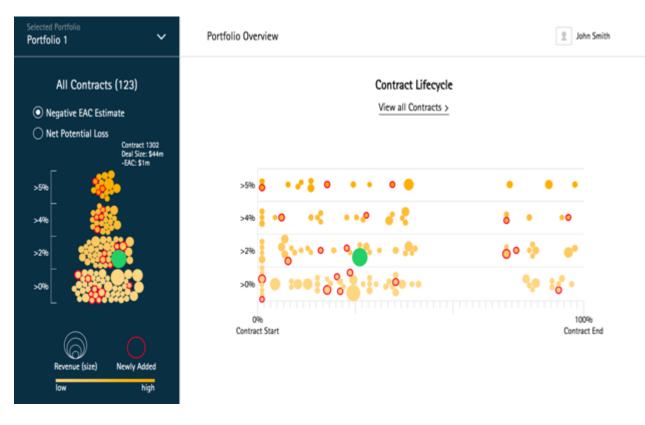
Top Feature Influencer (Positive)

Top Feature Influencer (Negative)

f5: 0 → 5.4,	<i>(</i> ² 0.03
f6: 168 러 0,	0.03
f7: 0 ⇒ 0.24,	0.02

f1: 430.5 \rightarrow 148.7, \checkmark 0.20 f2: 216 \rightarrow 0, \checkmark 0.17 f8: 423 \rightarrow 146.0, \checkmark 0.07

Explainable Risk Management - Finance



Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

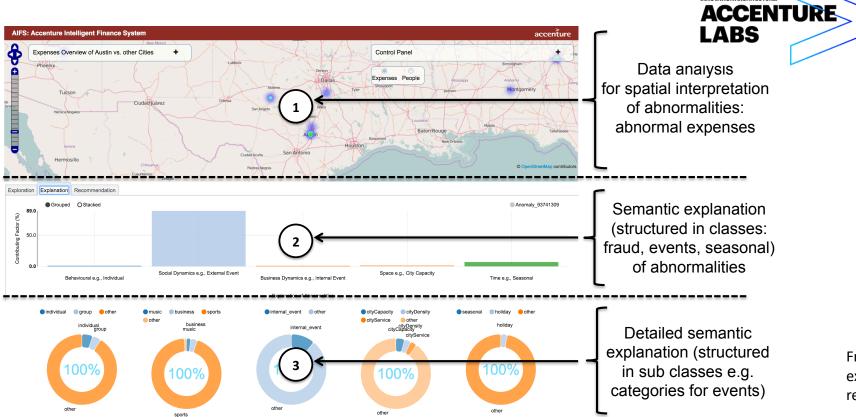


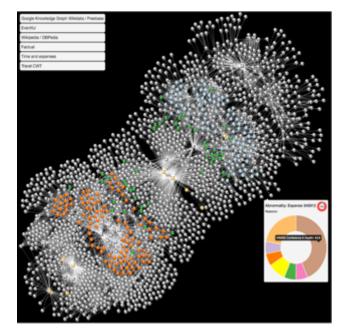
Challenge: Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of \$34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to assess and control, and (2) the existing non-systematic assessment process.

AI Technology: Integration of AI technologies i.e., Machine Learning, Reasoning, Natural Language Processing for building a robust model which can (1) predict revenue loss, (2) recommend corrective actions, and (3) explain why such actions might have a positive impact.

XAI Technology: Knowledge graph embedded Random Forrest

Explainable Anomaly Detection – Finance (Compliance)





Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

Challenge: Predicting and explaining abnormally employee expenses (as high accommodation price in 1000+ cities).

AI Technology: Various techniques have been matured over the last two decades to achieve excellent results. However most methods address the problem from a statistic and pure data-centric angle, which in turn limit any interpretation. We elaborated a web application running live with real data from (i) travel and expenses from Accenture, (ii) external data from third party such as Google Knowledge Graph, DBPedia (relational DataBase version of Wikipedia) and social events from Eventful, for explaining abnormalities.

Counterfactual Explanations for Credit Decisions (1) - Finance

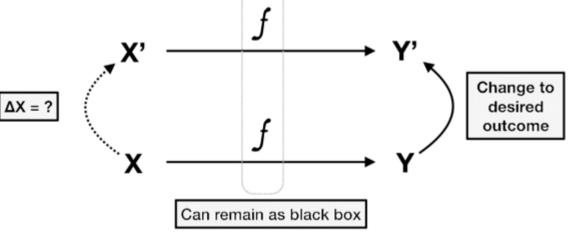
- Local, post-hoc, contrastive explanations of black-box classifiers
- Required minimum change in input vector to flip the decision of the classifier.
- Interactive Contrastive
 Explanations THALES



Challenge: We predict loan applications with off-the-shelf, interchangeable black-box estimators, and we explain their predictions with counterfactual explanations. In counterfactual explanations the model itself remains a black box; it is only through changing inputs and outputs that an explanation is obtained.

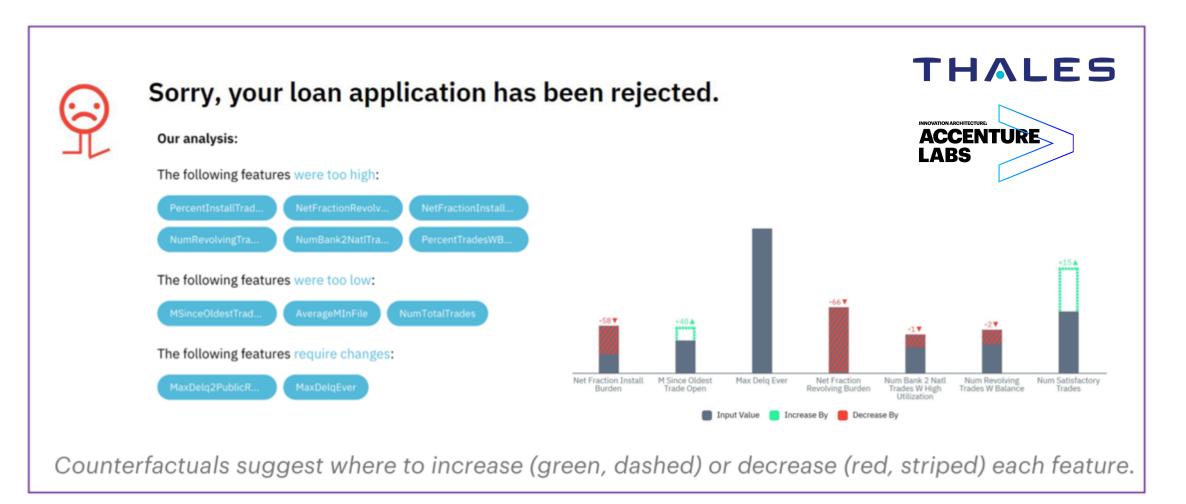
Al Technology: Supervised learning, binary classification.

XAI Technology: Post-hoc explanation, Local explanation, Counterfactuals, Interactive explanations



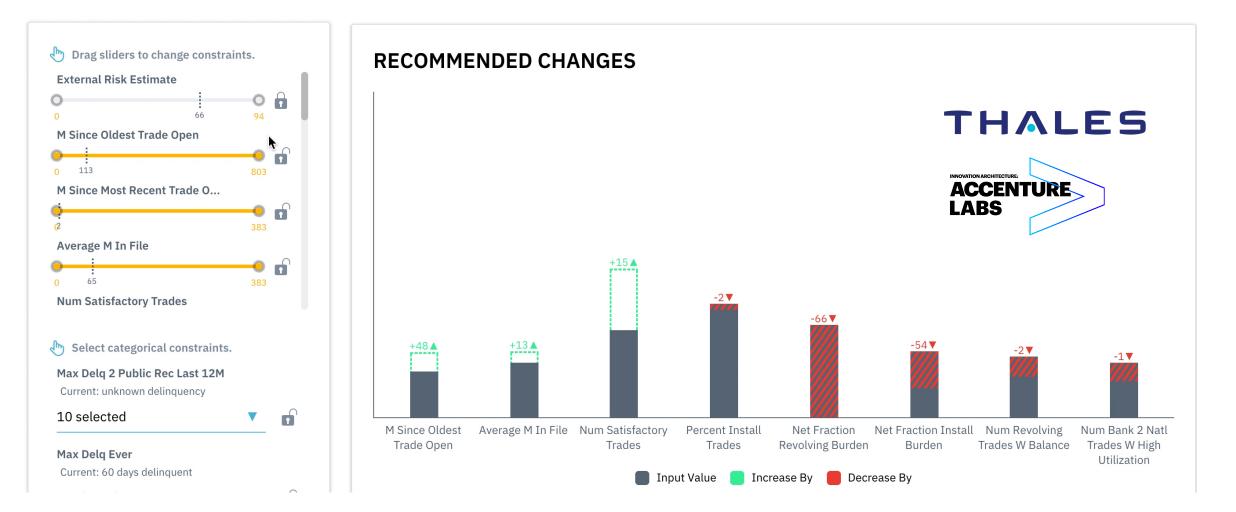
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Counterfactual Explanations for Credit Decisions (2) - Finance



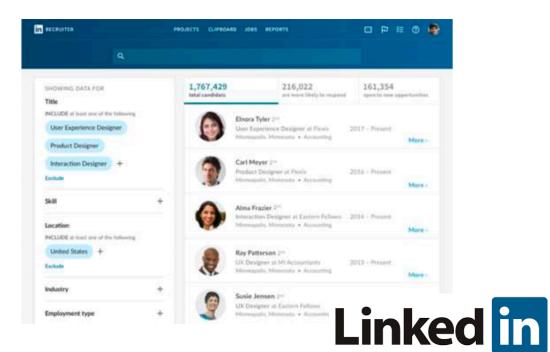
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Counterfactual Explanations for Credit Decisions (3) - Finance



Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Explaining Talent Search Results – Human Resources

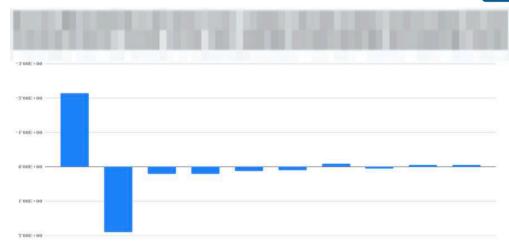


Challenge: How to rationalize a talent search for a recruiter when looking for candidates for a given role. Features are dynamic and costly to compute. Recruiters are interested in discriminating between two candidates to make a selection.

AI Technology: Generalized Linear Mixed Models, Artificial Neural Networks, XGBoost

XAI Technology: Generalized Linear Mixed Models (inherently explainable), Integrated Gradient, Features Importance in XGBoost

Feature	Description	Difference (1 vs 2)	Contribution
Feature	Description	-2.0476928	-2.144455602
Feature	Description	-2.3223877	1.903594618
Feature	Description	0.11666667	0.2114946752
Feature	Description	-2.1442587	0.2060414469
Feature	Description	-14	0.1215354111
Feature	Description	1	0.1000282466
Feature	Description	-92	-0.085286277
Feature	Description	0.9333333	0.0568533262
Feature	Description	-1	-0.051796317
Feature	Description	-1	-0.050895940



Explanation of Medical Condition Relapse – Health







Challenge: Explaining medical condition relapse in the context of oncology.

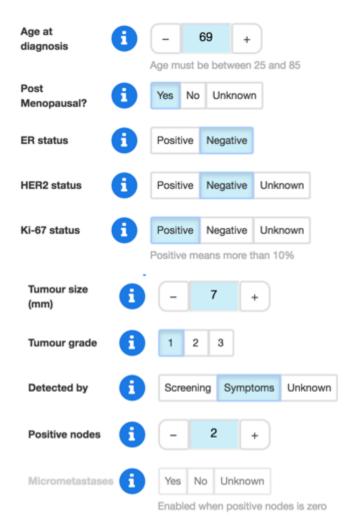
Al Technology: Relational learning

XAI Technology: Knowledge graphs and Artificial Neural Networks

Knowledge graph parts explaining medical condition relapse

Breast Cancer Survival Rate Prediction - Health predict breast cancer

.....



Results										
Table Curves	Chart ling	Texts	lcons							
These results are for women who have already had surgery. This table										
shows the percentage	of wome	en who sui	vive at lea	ast 5	10	15	years			
after surgery, based on the information you have provided.										
Treatment	Addit	ional Ben	efit O	verall	Surv	ival	%			
							/0			
					Juiv	- Tai	/0			
Surgery only	-			2%	Juiv		,.			
	- 0%		72							
Surgery only			72	2% 2%						

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote predict.nhs.uk/tool

Challenge: Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

Al Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

More on XAI

(Some) Tutorials, Workshops, Challenge

Tutorial:

- AAAI 2019 Tutorial on On Explainable AI: From Theory to Motivation, Applications and Limitations (#1) https://xaitutorial2019.github.io/
- ICIP 2018 / EMBC 2019 Interpretable Deep Learning: Towards Understanding & Explaining Deep Neural Networks (#2) http://interpretable-ml.org/icip2018tutorial/ http://interpretable-ml.org
- ICCV 2019 Tutorial on Interpretable Machine Learning for Computer Vision (#2) https://interpretablevision.github.io/
- KDD 2019 Tutorial on Explainable AI in Industry (#1) https://sites.google.com/view/kdd19-explainable-ai-tutorial

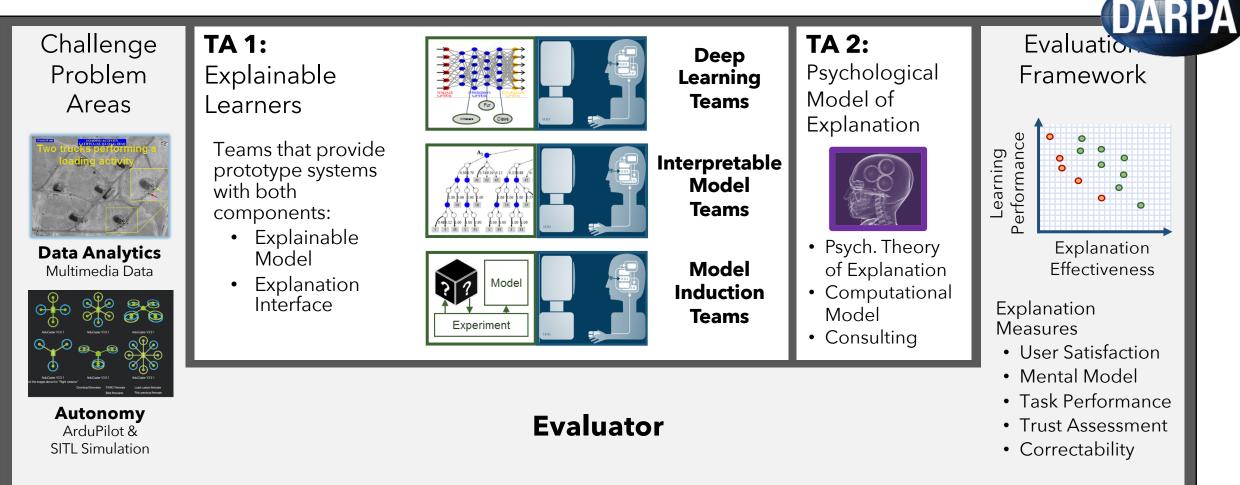
Workshop:

- ISWC 2019 Workshop on Semantic Explainability (#1) http://www.semantic-explainability.com/
- IJCAI 2019 Workshop on Explainable Artificial Intelligence (#3) https://sites.google.com/view/xai2019/home 55 paper submitted in 2019
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) https://www.doc.ic.ac.uk/~kc2813/OXAI/
- SIGIR 2019 Workshop on Explainable Recommendation and Search (#2) https://ears2019.github.io/
- ICAPS 2019 Workshop on Explainable Planning (#2)- https://kcl-planning.github.io/XAIP-Workshops/ICAPS 2019 23 papers submitted in 2019 <a href="https://brancipla
- KDD 2019 Workshop on Explainable AI for fairness, accountability, and transparency (#1) <u>https://xai.kdd2019.a.intuit.com</u>
- ICCV 2019 Workshop on Interpreting and Explaining Visual Artificial Intelligence Models (#1) http://xai.unist.ac.kr/workshop/2019/
- NeurIPS 2019 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy https://sites.google.com/view/feap-ai4fin-2018/
- CD-MAKE 2019 Workshop on Explainable AI (#2) https://cd-make.net/special-sessions/make-explainable-ai/
- AAAI 2019 / CVPR 2019 Workshop on Network Interpretability for Deep Learning (#1 and #2) http://networkinterpretability.org/ https://explainai.net/
- IEEE FUZZ 2019 / Advances on eXplainable Artificial Intelligence (#2) https://sites.google.com/view/xai-fuzzieee2019
- International Conference on NL Generation Interactive Natural Language Technology for Explainable Artificial Intelligence (EU H2020 NL4XAI; #1) https://sites.google.com/view/nl4xai2019/
 Challenge:
- 2018: FICO Explainable Machine Learning Challenge (#1) https://community.fico.com/s/explainable-machine-learning-challenge

(Some) Software Resources

- DeepExplain: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. github.com/marcoancona/DeepExplain
- iNNvestigate: A toolbox to iNNvestigate neural networks' predictions. <u>github.com/albermax/innvestigate</u>
- SHAP: SHapley Additive exPlanations. github.com/slundberg/shap
- Microsoft Explainable Boosting Machines. <u>https://github.com/Microsoft/interpret</u>
- GANDissect: Pytorch-based tools for visualizing and understanding the neurons of a GAN. https://github.com/CSAILVision/GANDissect
- ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. github.com/TeamHG-Memex/eli5
- Skater: Python Library for Model Interpretation/Explanations. github.com/datascienceinc/Skater
- Yellowbrick: Visual analysis and diagnostic tools to facilitate machine learning model selection. github.com/DistrictDataLabs/yellowbrick
- Lucid: A collection of infrastructure and tools for research in neural network interpretability. github.com/tensorflow/lucid
- LIME: Agnostic Model Explainer. https://github.com/marcotcr/lime
- Sklearn_explain: model individual score explanation for an already trained scikit-learn model. https://github.com/antoinecarme/sklearn_explain
- Heatmapping: Prediction decomposition in terms of contributions of individual input variables
- Deep Learning Investigator: Investigation of Saliency, Deconvnet, GuidedBackprop and more. https://github.com/albermax/innvestigate
- Google PAIR What-if: Model comparison, counterfactual, individual similarity. <u>https://pair-code.github.io/what-if-tool/</u>
- Google tf-explain: <u>https://tf-explain.readthedocs.io/en/latest/</u>
- IBM AI Fairness: Set of fairness metrics for datasets and ML models, explanations for these metrics. https://github.com/IBM/aif360
- Blackbox auditing: Auditing Black-box Models for Indirect Influence. <u>https://github.com/algofairness/BlackBoxAuditing</u>
- Model describer: Basic statiscal metrics for explanation (visualisation for error, sensitivity). <u>https://github.com/DataScienceSquad/model-describer</u>
- AXA Interpretability and Robustness: <u>https://axa-rev-research.github.io/</u> (more on research resources not much about tools)

(Some) Initiatives: XAI in USA



TA1: Explainable Learners

> Explainable learning systems that include both an explainable model and an explanation interface

TA2: Psychological Model of Explanation

> Psychological theories of explanation and develop a computational model of explanation from those theories

(Some) Initiatives: XAI in Canada



System Robustness

- To biased data
- Of algorithm
- To change
- To attacks

Certificability

- Structural warranties
- Risk auto evaluation
- External audit

Explicability & nterpretability

Privacy by design

- Differential privacy
- Homomorphic coding
- Collaborative learning
- To attacks

(Some) Initiatives: XAI in EU



Conclusion

Why do we Need XAI by the Way?

- To empower individual against undesired effects of automated decision making
- To reveal and protect new vulnerabilities
- To implement the "right of explanation"
- To improve industrial standards for developing AI-powered products, increasing the trust of companies and consumers
- To help people make better decisions
- *To align* algorithms with human values
- To preserve (and expand) human autonomy
- To scale and industrialize AI

Conclusion

- Explainable AI is motivated by **real-world applications in AI**
- Not a new problem a reformulation of past research challenges in AI
- Multi-disciplinary: multiple AI fields, HCI, social sciences (multiple definitions)
- In AI (in general): many interesting / complementary approaches
- Many industrial applications already crucial for Al adoption in critical systems

- There is *no agreement* on *what an explanation is*
- There is *not a formalism* for *explanations*
- There is *no work* that seriously addresses the problem of *quantifying* the grade of *comprehensibility* of an explanation for humans
- Is it possible to join *local* explanations to build a *globally* interpretable model?
- What happens when black box make decision in presence of latent features?
- What if there is a *cost* for querying a black box?



Future Challenges

- Creating awareness! Success stories!
- Foster multi-disciplinary collaborations in XAI research.
- Help shaping industry standards, legislation.
- More work on transparent design.
- Investigate symbolic and sub-symbolic reasoning.
- Evaluation:
 - We need benchmark Shall we start a task force?
 - *We need an XAI challenge -* Anyone interested?
 - *Rigorous, agreed upon, human-based* evaluation protocols

Wherever safety and Security are Critical, Thales c build smarter solutions. Everywhere.

- Strong knowledge of Machine Learning foundations
- Strong development skills with Machine Learning frameworks e.g., Scikit-learn, Tensoflow, PyTorch, Theano
 - Knowledge of mainstream Deep Learning architectures (MLP, CNN, RNN, etc).
 - Strong Python programming skills
 - Working knowledge of Linux OS
 - Eagerness to contribute in a team-oriented environment
 - Demonstrated leadership abilities in school, civil or business organisations
 - Ability to work creatively and analytically in a problem-solving environment
 - Proven verbal and written communication skills in English (talks, presentations, publications, etc.)

Basic Qualifications

- Master's degree in computer science, engineering or mathematics fields
- Prior experience in artificial intelligence, machine learning, natural language processing, or advanced analytics

working structured and unstructured data (SQL, Cassandra, MongoDB, Hive, etc.)

• A track record of outstanding AI software development with Github (or similar) evidence

Minimum 3 years of analytic experience Python with interest in artificial intelligence with

paced projects.

Professional Skill Requirements

- Chief AI Scientist, CortAlx, Thales, Montreal Canada

@freddvlecue https://tinyurl.com/freddylecue Freddy.lecue.e@thalesdigital.io

AUGUST 28TH, 2019

Freddy Lecue

• Demonstrated interest in Explainable AI and or relational learning

Demonstrated abilities in designing large scale AI systems

- Work experience with programming languages such as C, C++, Java, scripting languages (Perl/Python/Ruby) or similar
- Hands-on experience with data visualization, analytics tools/languages
- Demonstrated teamwork and collaboration in professional settings
- Ability to establish credibility with clients and other team members

- Job Openings is a global technology leader for the Defendence of the Combined expertises of the Combin
 - nave made Thales a key player in keeping the pub protecting the national security interests of count

Established in 1972, Thales Canada has over 1,800 Toronto and Vancouver working in Defence, Avior

This is a unique opportunity to play a key role on t Technology (TRT) in Canada (Quebec and Montrea applied R&T experts at five locations worldwide. 1 intelligence technologies. Our passion is imagining cutting edge AI technologies. Not only will you joi network, but this TRT is also co-located within Cor Intelligence eXpertise) i.e., the new flagship progr to work.

Job Description

An AI (Artificial Intelligence) Research and Techno developing innovative prototypes to demonstrate intelligence. To be successful in this role, one mos what's new, and a strong ability to learn new tech hand-on technical skills and be familiar with latest will contribute as technical subject matter experts and its business units. In addition to the implement Preferred Qualifications individual will also be involved in the initial projec thinking, and team work is also critical for this role

As a Research and Technology Applied AI Scientist

- Good foundation in mathematics, statistic