On the Role of Knowledge Graphs for the adoption of Machine Learning Systems in Industry

May 7th, 2019

Freddy Lecue
Chief AI Scientist, CortAIx, Thales, Montreal – Canada
Inria, Sophia Antipolis - France

@freddylecue
https://tinyurl.com/freddylecue

Context


NETFLIX
AGENT.45
TAXI
bella ciao
KALIBER
hotel social
ZEN&ZACK
MADEMOISELLE
ZOMBIE COLONY
KENYA
ROCKET NINJA
OPIUM
dharma yoga


Gary Chavez added a photo you might be in.
about a minute ago ·


BNSF

Markets we serve

Aerospace

Space

Ground Transportation

Defence

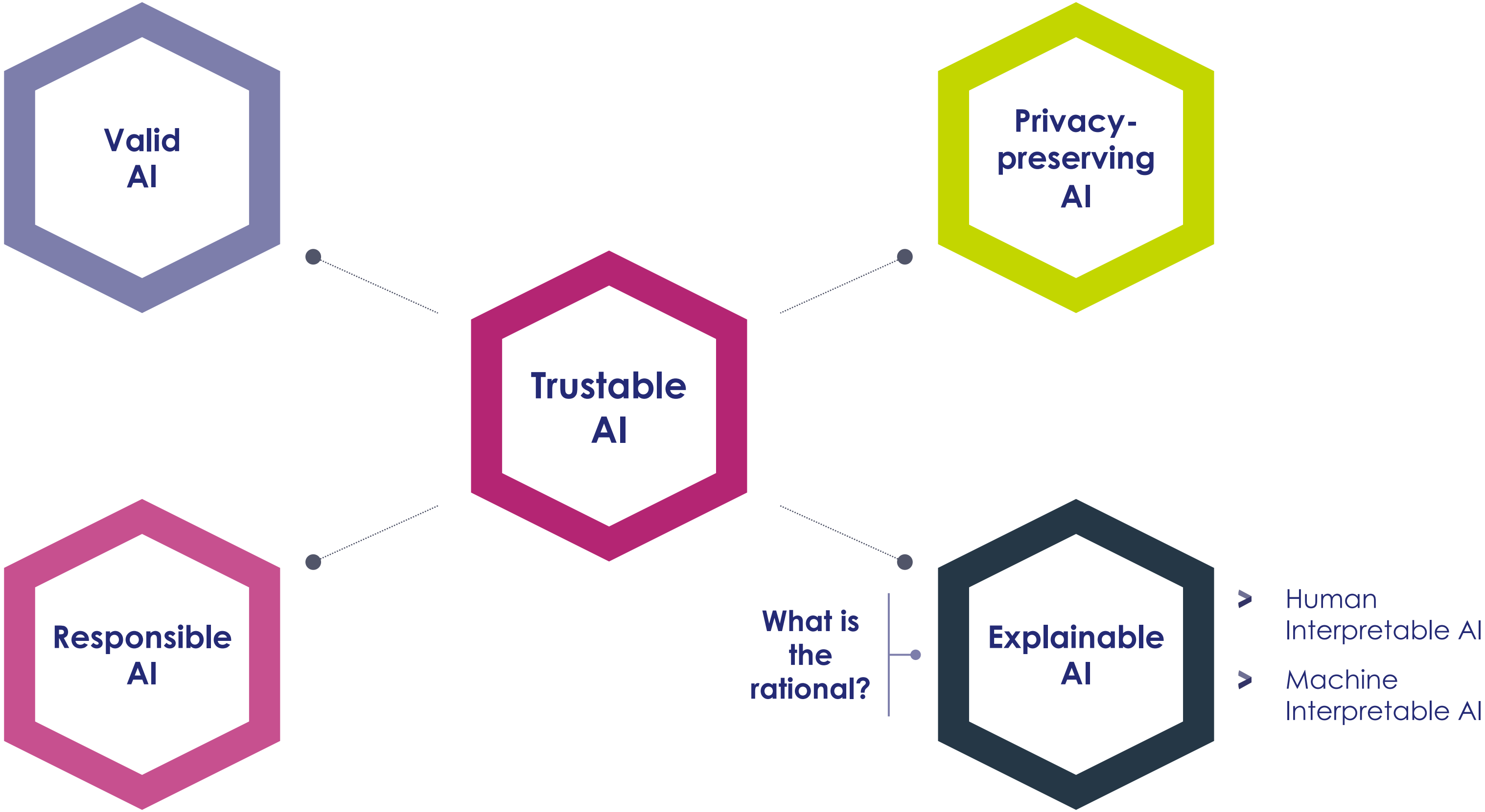
Security

Trusted Partner For A Safer World

THALES

Trustable AI

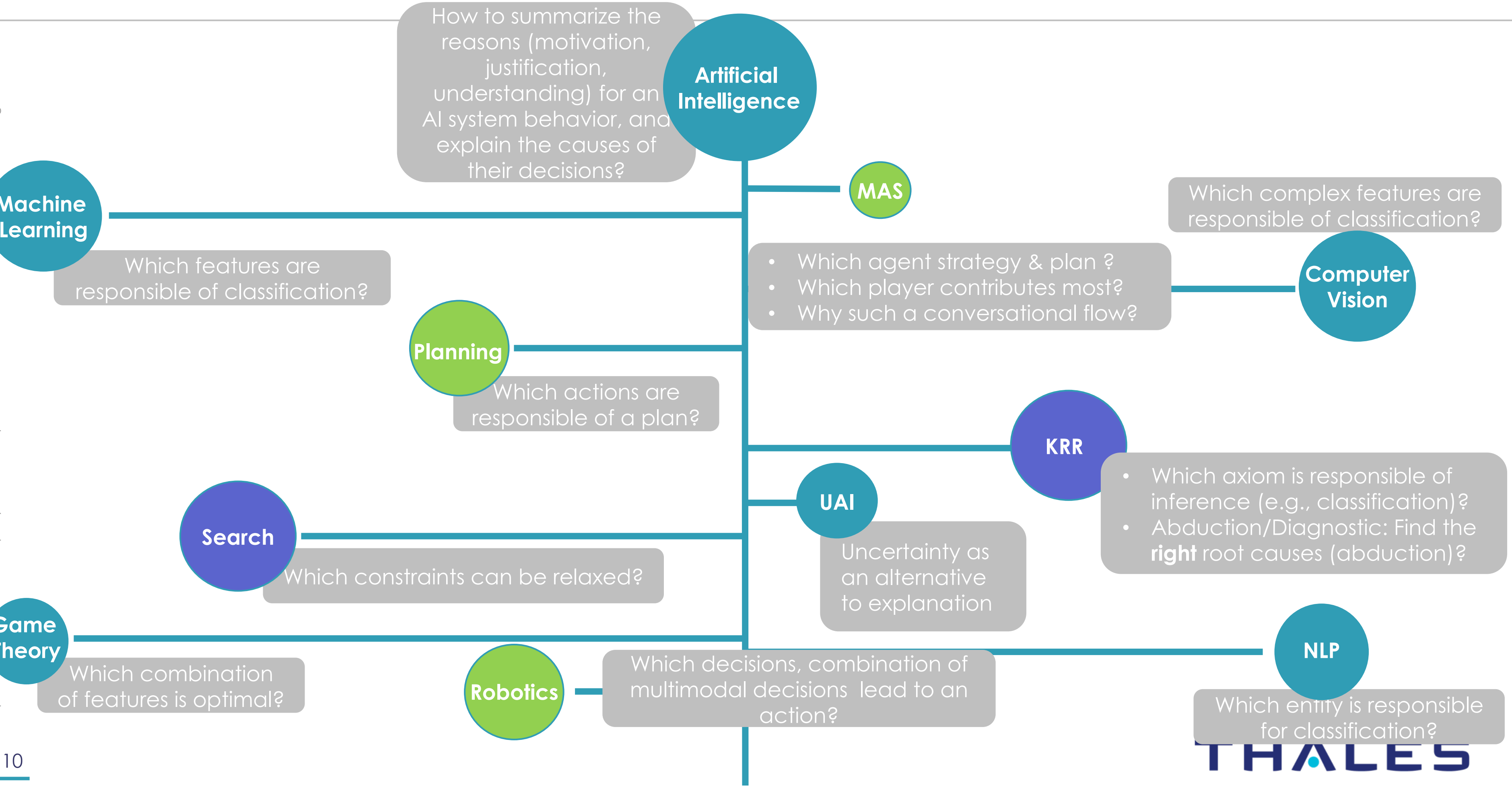
AI Adoption: Requirements



THALES

XAI in AI

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

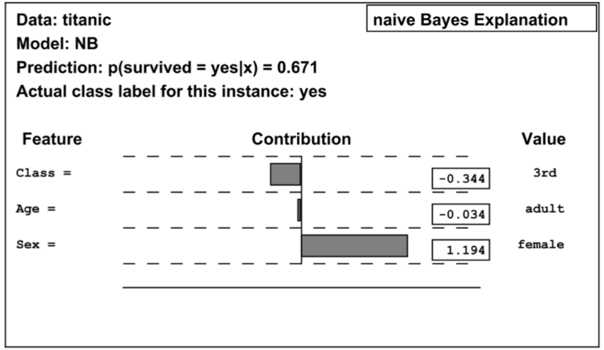


XAI in Machine Learning

Overview of explanation in different AI fields (1)

Machine Learning (except Artificial Neural Network)

Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs



Naive Bayes model

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

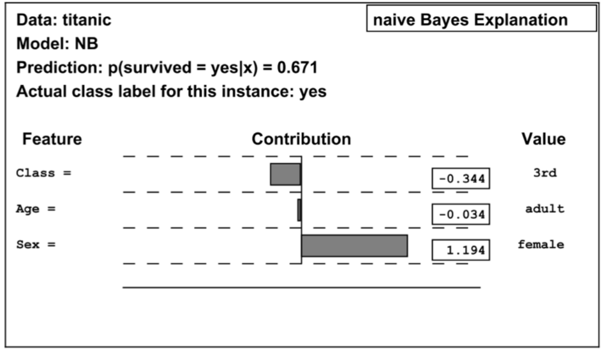
THALES

Overview of explanation in different AI fields (1)

Machine Learning (except Artificial Neural Network)

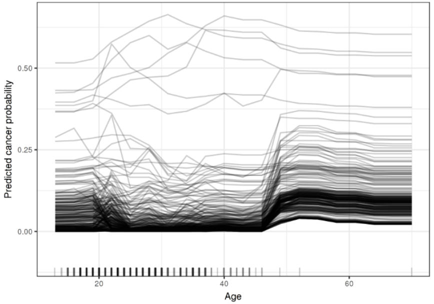
Interpretable Models:

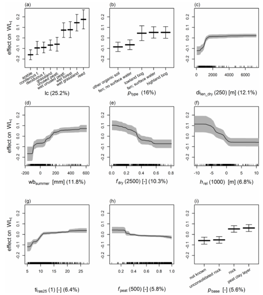
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs

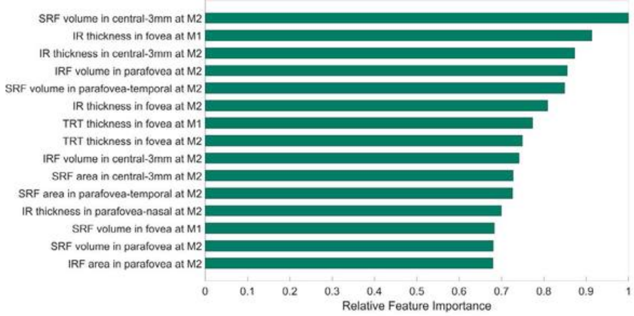


Naive Bayes model

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.







(a)

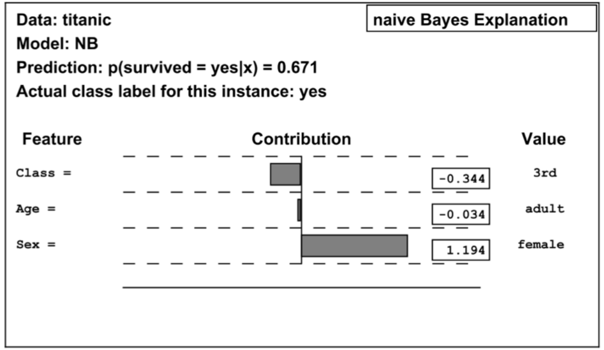
Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

THALES

Overview of explanation in different AI fields (1)

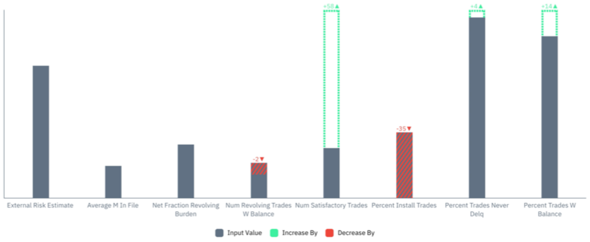
Machine Learning (except Artificial Neural Network)

Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs



Naive Bayes model

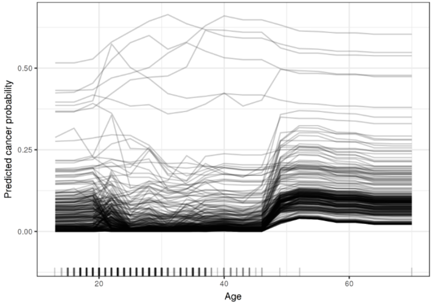
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

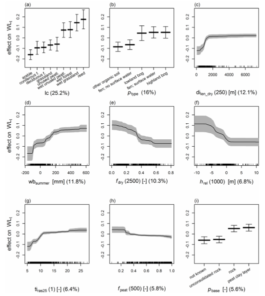


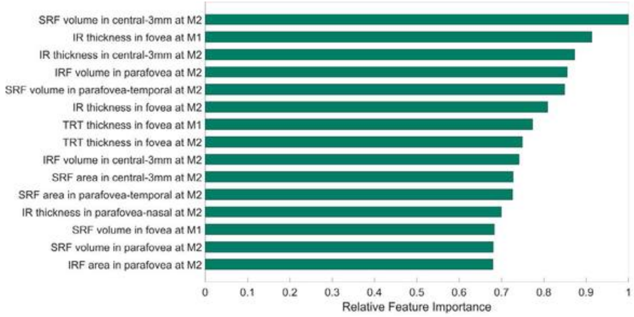
Counterfactual What-if

Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. FAT 2019: 279-288

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. CoRR abs/1811.05245 (2018)







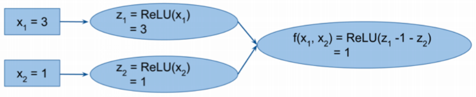
(a)

Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

THALES

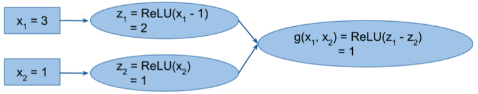
Overview of explanation in different AI fields (2)

Machine Learning (only Artificial Neural Network)



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

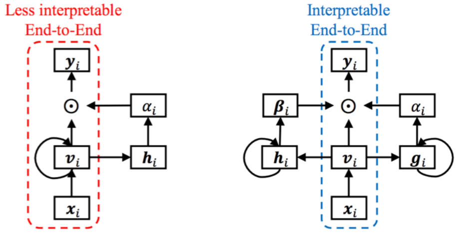
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

Attention Mechanism



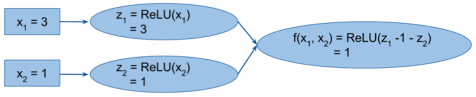
(a) Standard attention model (b) RETAIN model

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512

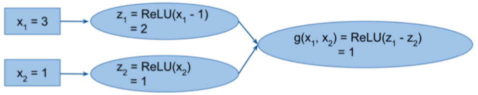
Overview of explanation in different AI fields (2)

Machine Learning (only Artificial Neural Network)



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

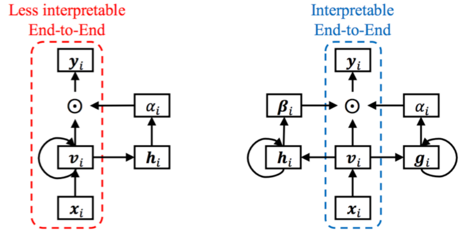
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

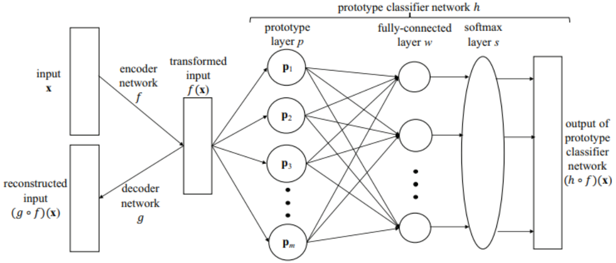
Attention Mechanism



(a) Standard attention model (b) RETAIN model

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512



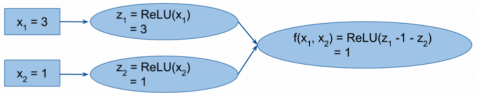
Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537

THALES

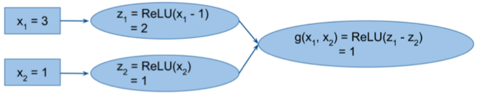
Overview of explanation in different AI fields (2)

Machine Learning (only Artificial Neural Network)



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

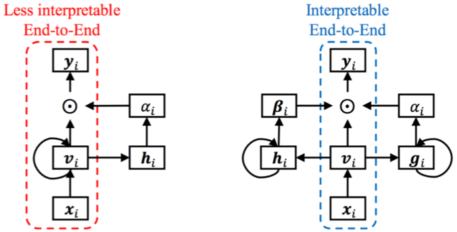
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

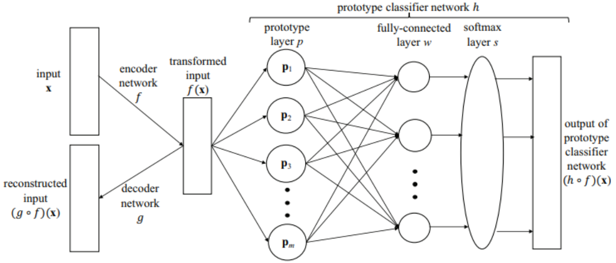
Attention Mechanism



(a) Standard attention model (b) RETAIN model

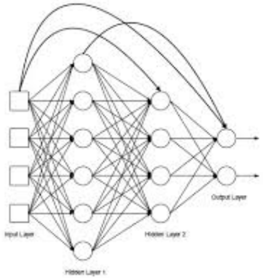
D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

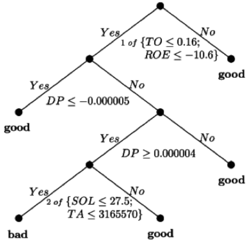
Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512



Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537





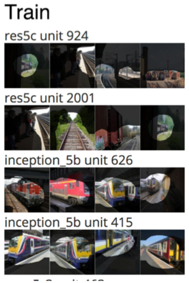
Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

THALES

Overview of explanation in different AI fields (3)

Computer Vision

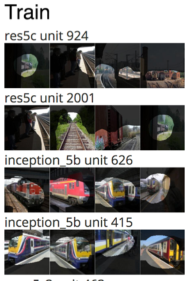


Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327

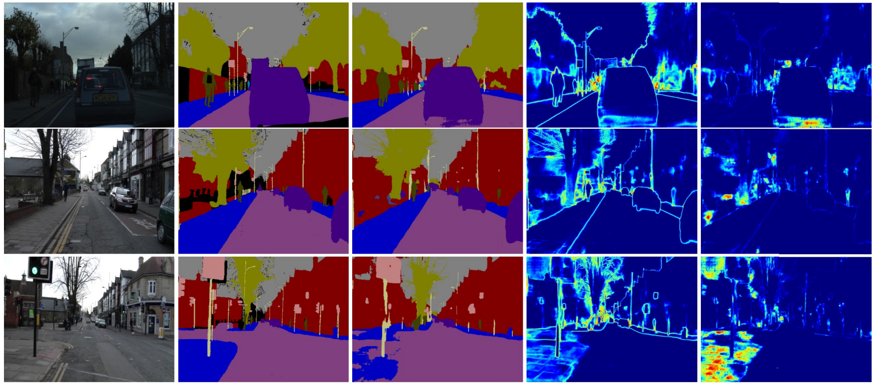
Overview of explanation in different AI fields (3)

Computer Vision



Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

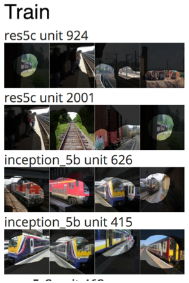
Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

THALES

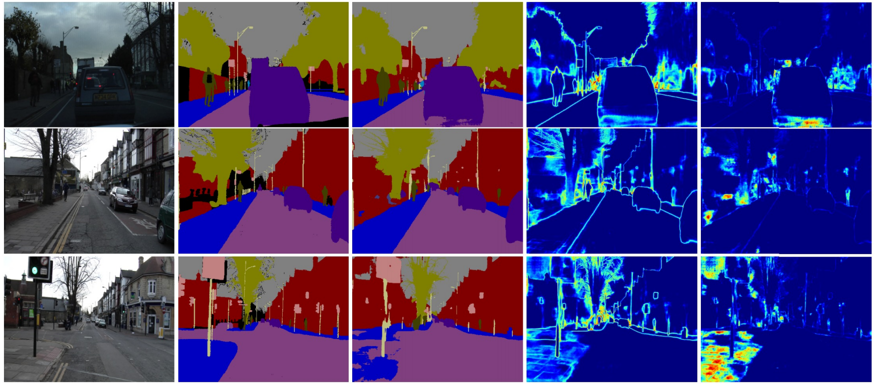
Overview of explanation in different AI fields (3)

Computer Vision



Interpretable Units

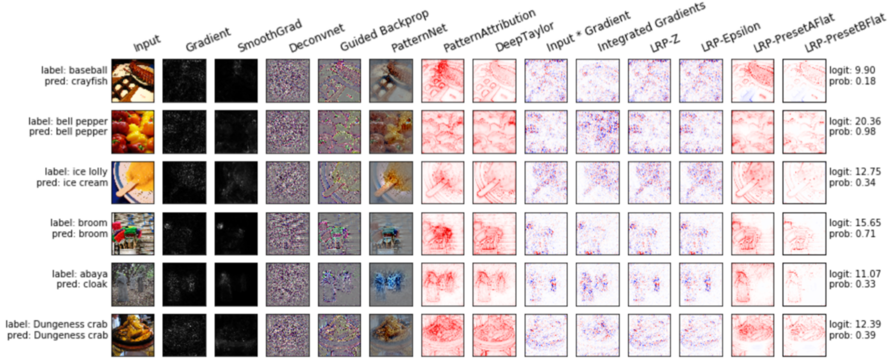
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

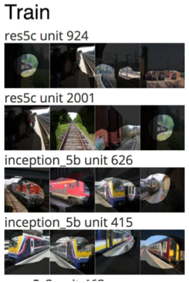


Saliency Map

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

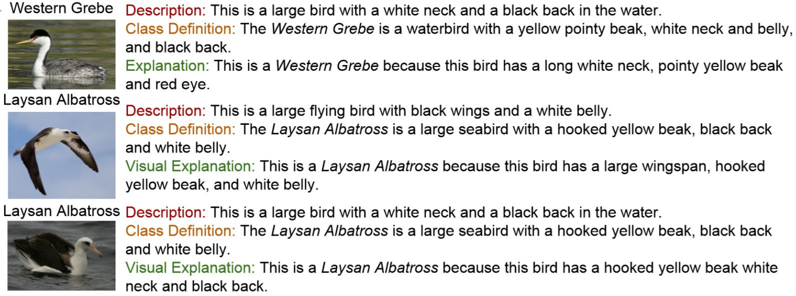
Overview of explanation in different AI fields (3)

Computer Vision



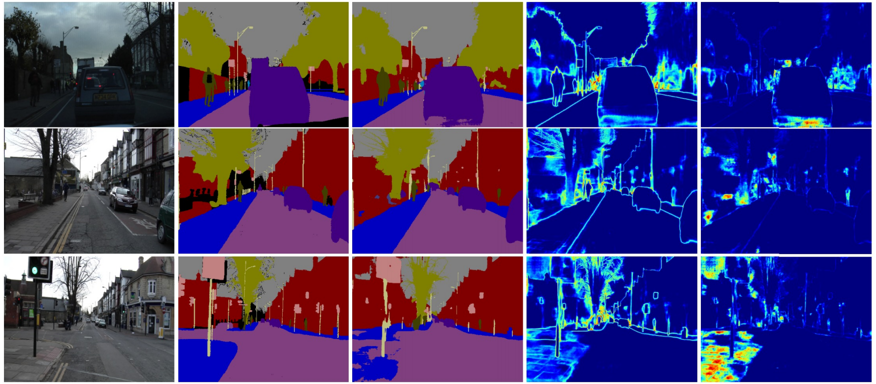
Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



Visual Explanation

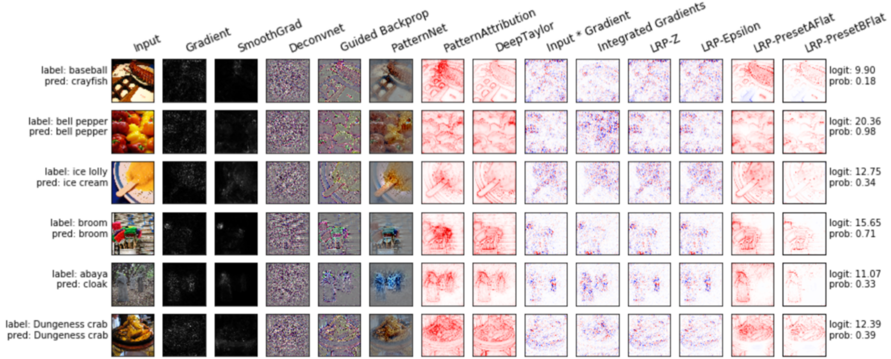
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590



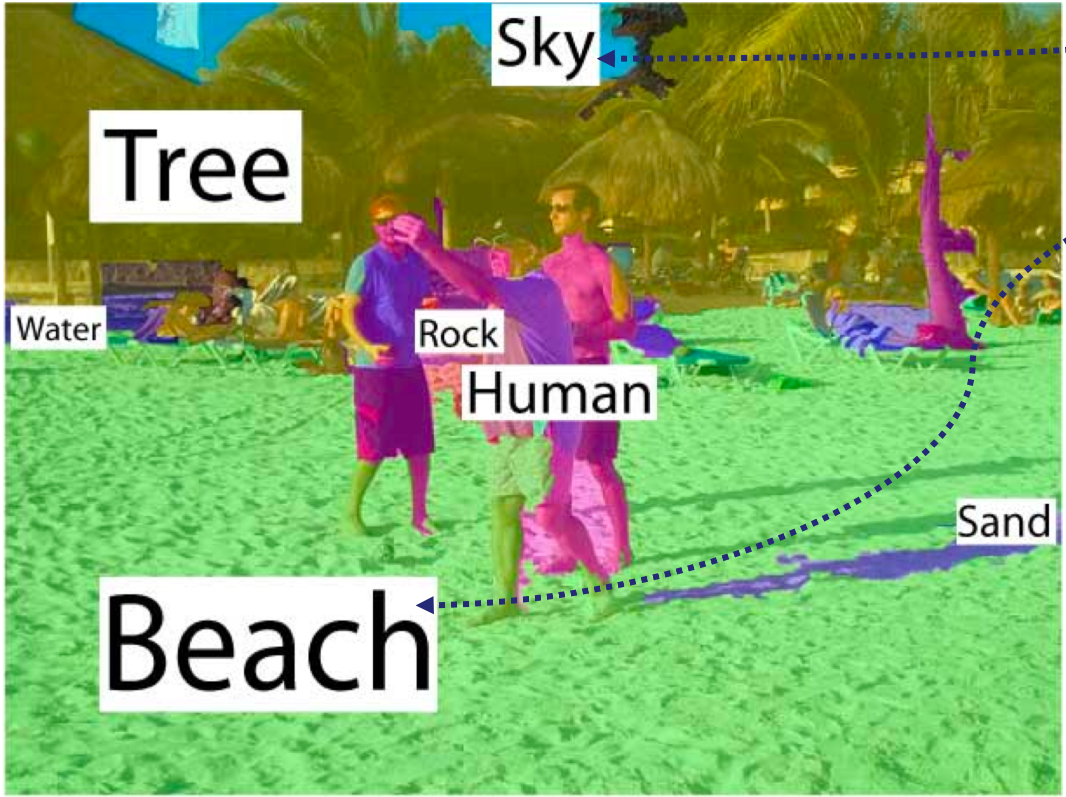
Saliency Map

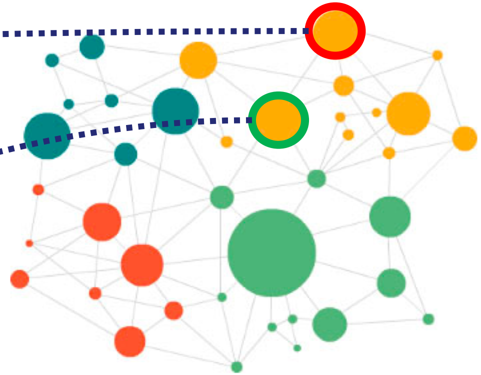
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

On the role of Knowledge Graphs in Explainable Machine Learning

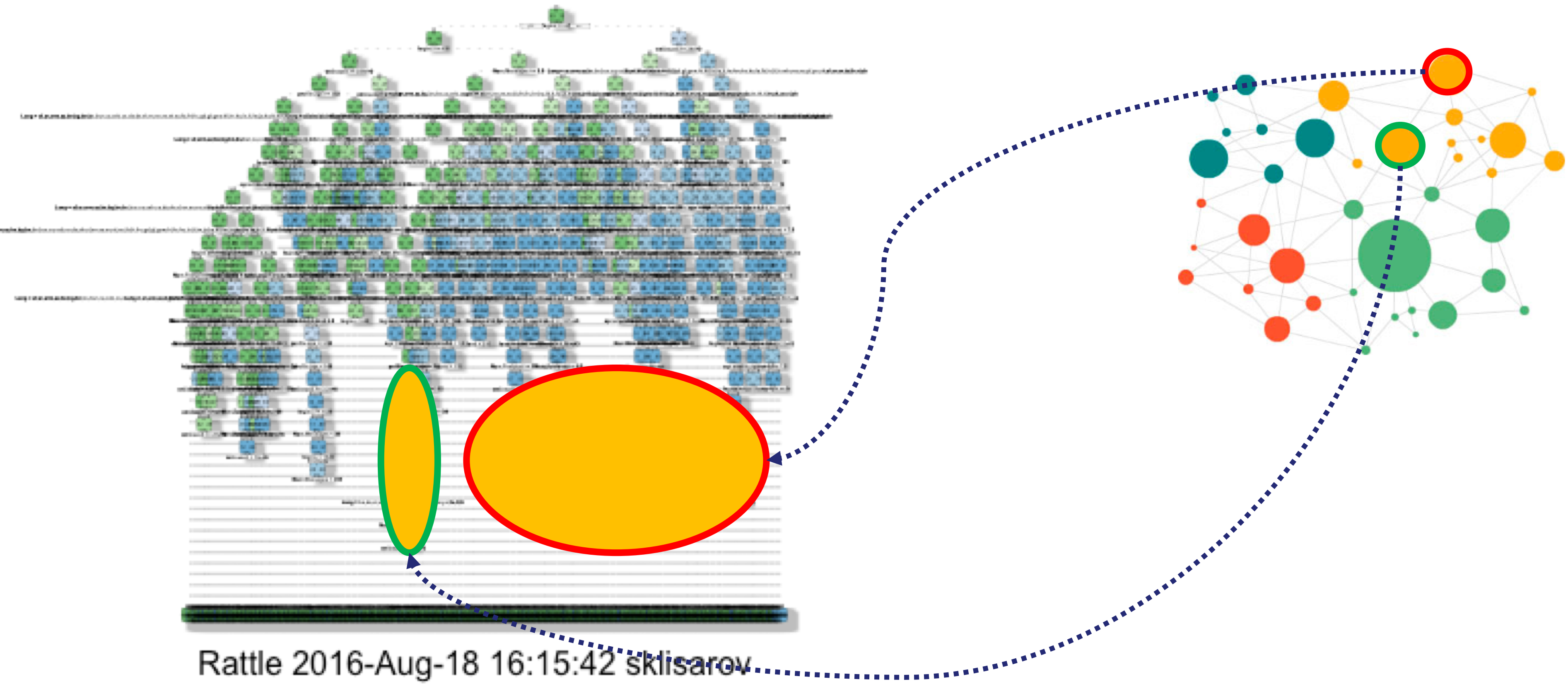
THALES

Knowledge Graph Embeddings in Machine Learning



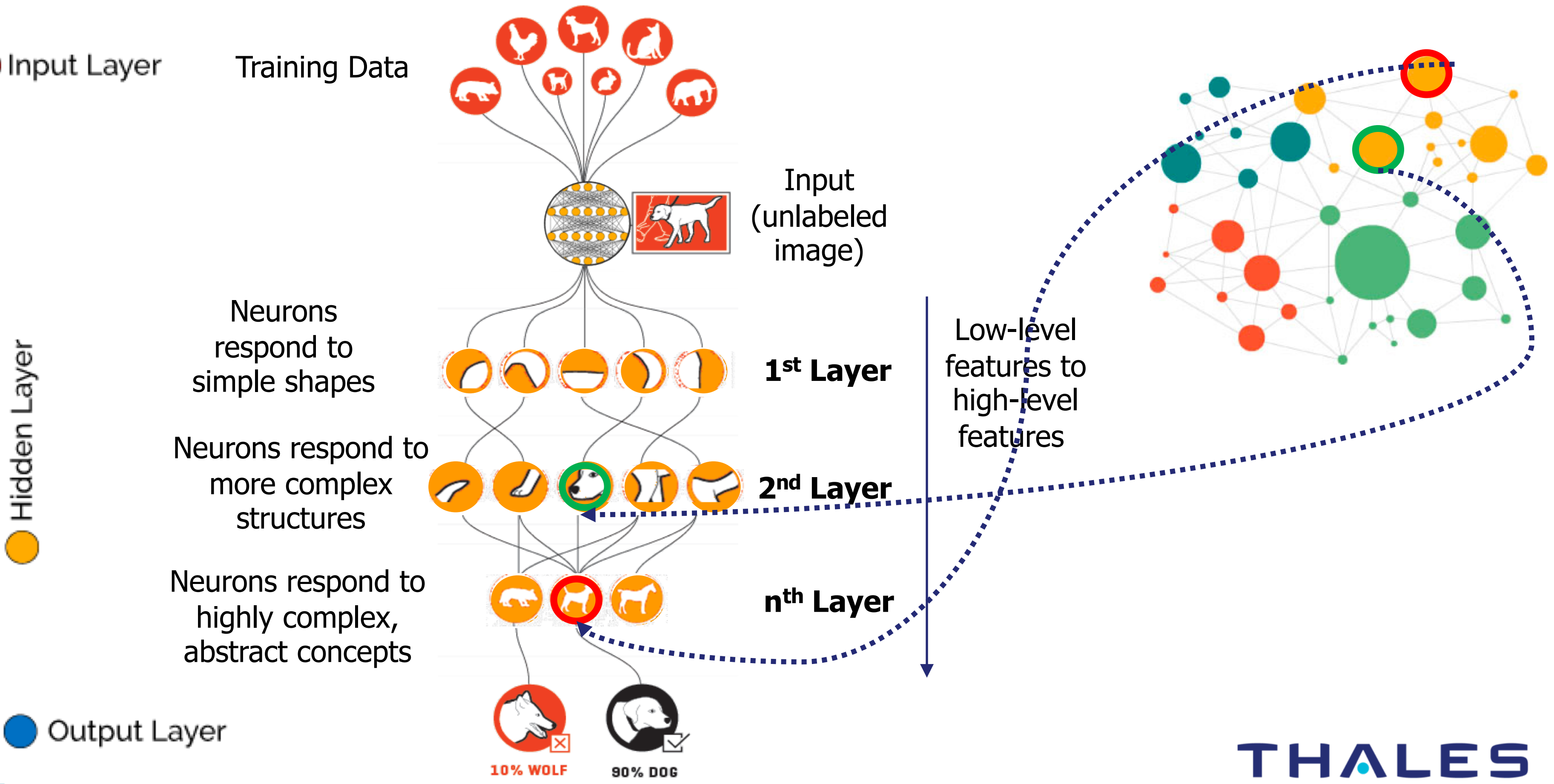


Knowledge Graph for Decision Trees



https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

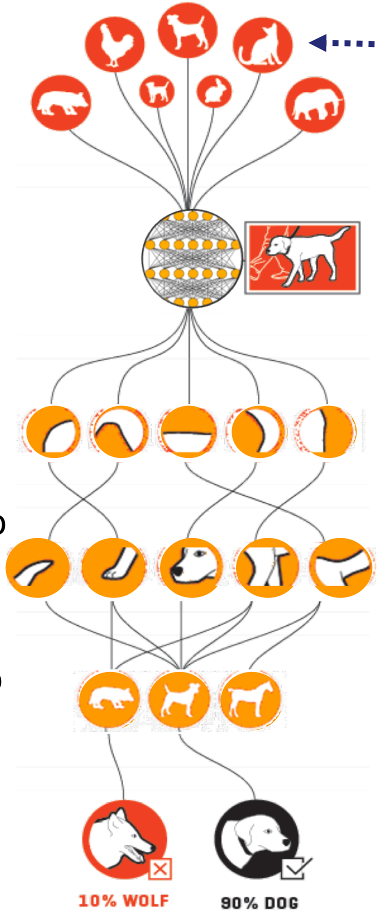
Knowledge Graph for Deep Neural Network (1)



Knowledge Graph for Deep Neural Network (2)

Input Layer

Training Data



Input (unlabeled image)

Hidden Layer

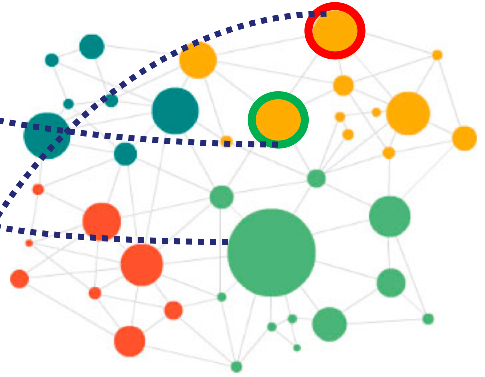
Neurons respond to simple shapes — 1st Layer

Neurons respond to more complex structures — 2nd Layer

Neurons respond to highly complex, abstract concepts — nth Layer

Low-level features to high-level features

Output Layer



What is the causal relationship between the input / hidden / output layers

THALES

Knowledge Graph for Personalized XAI

Description 1: This is an orange train accident

Description 2: This is an train accident between two speed merchant trains of characteristics X43-B and Y33-C in a dry environment

Description 3: This is a public transportation accident

THALES

"How to explain transfer learning with appropriate knowledge representation?

Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)

Knowledge-Based Transfer Learning Explanation

Jiaoyan Chen
Department of Computer Science
University of Oxford, UK

Freddy Lecue
INRIA, France
Accenture Labs, Ireland

Jeff Z. Pan
Department of Computer Science
University of Aberdeen, UK

Ian Horrocks
Department of Computer Science
University of Oxford, UK

Huajun Chen
College of Computer Science, Zhejiang University, China
Alibaba-Zhejian University Frontier Technology Research Center

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/embc2019tutorial/

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
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) - https://www.doc.ic.ac.uk/~kc2813/OXAI/
- ICAPS 2019 Workshop on Explainable Planning (#2)- https://kcl-planning.github.io/XAIP-Workshops/ICAPS_2019
- 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/

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. github.com/albermax/innvestigate
- SHAP: SHapley Additive exPlanations. github.com/slundberg/shap
- 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. https://pair-code.github.io/what-if-tool/
- 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. https://github.com/algofairness/BlackBoxAuditing
- Model describer: Basic statistical metrics for explanation (visualisation for error, sensitivity). https://github.com/DataScienceSquad/model-describer

THALES

(Some) Initiatives: XAI in Canada

DEEL (Dependable Explainable Learning) Project 2019-2024

- Research institutions
- Industrial partners
- Academic partners
  - Science and technology to develop new methods towards Trustable and Explainable AI

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

Certificability
- Structural warranties
- Risk auto evaluation
- External audit

Explicability & Interpretability

Privacy by design
- Differential privacy
- Homomorphic coding
- Collaborative learning
- To attacks

Conclusion

- Not a new problem – a reformulation of past research challenges in AI
- Explainable AI is motivated by real-world applications in AI
- Explainable AI is a strong requirement for adoption of AI in industry
- Lots of approaches for eXplainable Machine Learning… but no semantics attached
- Need more work on joint learning and reasoning systems
- In AI (in general): many interesting / complementary approaches

THALES

Job Openings

Research and Technology Applied AI (Artificial Intelligence) Scientist

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

hnology leader for the Defen
ogy, the combined expertise c
have made Thales a key player in keeping the pub
protecting the national security interests of count

Established in 1972, Thales Canada has over 1,80
Toronto and Vancouver working in Defence, Avio

This is a unique opportunity to play a key role on t
Technology (TRT) in Canada (Quebec and Montre
applied R&T experts at five locations worldwide. T
intelligence technologies. Our passion is imagining
cutting edge AI technologies. Not only will you joi
network, but this TRT is also co-located within Co
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 impleme
individual will also be involved in the initial projec
thinking, and team work is also critical for this rol

As a Research and Technology Applied AI Scientist
paced projects.

Professional Skill Requirements

- Good foundation in mathematics, statisti
- 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

Preferred Qualifications

- Minimum 3 years of analytic experience Python with interest in artificial intelligence with working structured and unstructured data (SQL, Cassandra, MongoDB, Hive, etc.)
- A track record of outstanding AI software development with Github (or similar) evidence
- Demonstrated abilities in designing large scale AI systems
- Demonstrated interest in Explainable AI and or relational learning
- 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

MAY 7TH, 2019

Freddy Lecue
Chief AI Scientist, CortAIx, Thales, Montreal – Canada

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