THALES

On the Role of Trust and Explanation for AI Adoption in Industry.

May 16th, 2019

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

@freddylecue
https://tinyurl.com/freddylecue

Context

THALES


NETFLIX
AGENT.45
TAXI
bella ciao
KALIBER
hotel social
ZEN&ZACK
MADEMOISELLE
ZOMBIE COLONY
KENYA
ROCKET NINJA 99
OPIUM
JUILLET
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

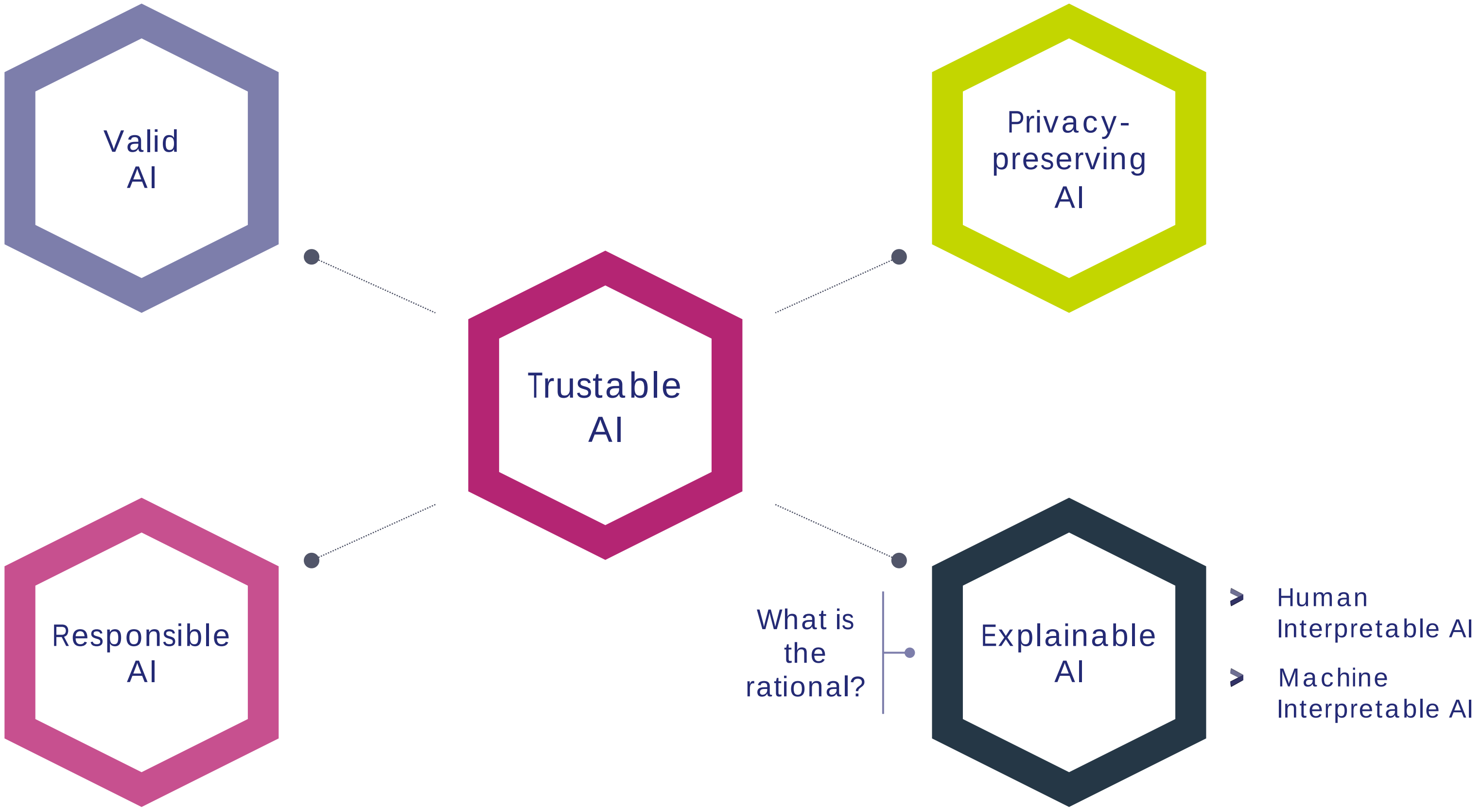
Trusted Partner For A Safer World

THALES

Trustable AI

THALES

AI Adoption: Requirements

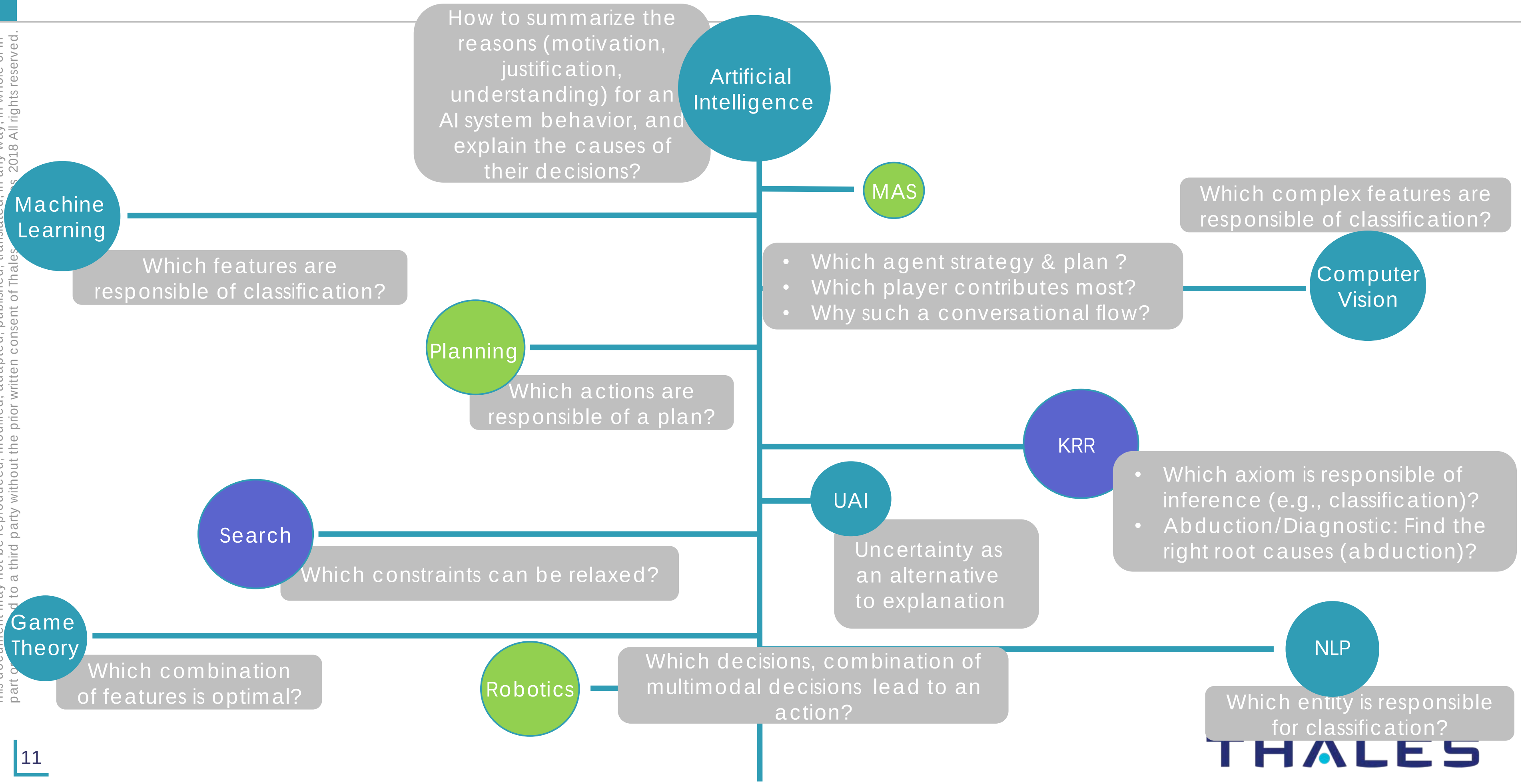


THALES

XAI in AI

THALES

XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



THALES

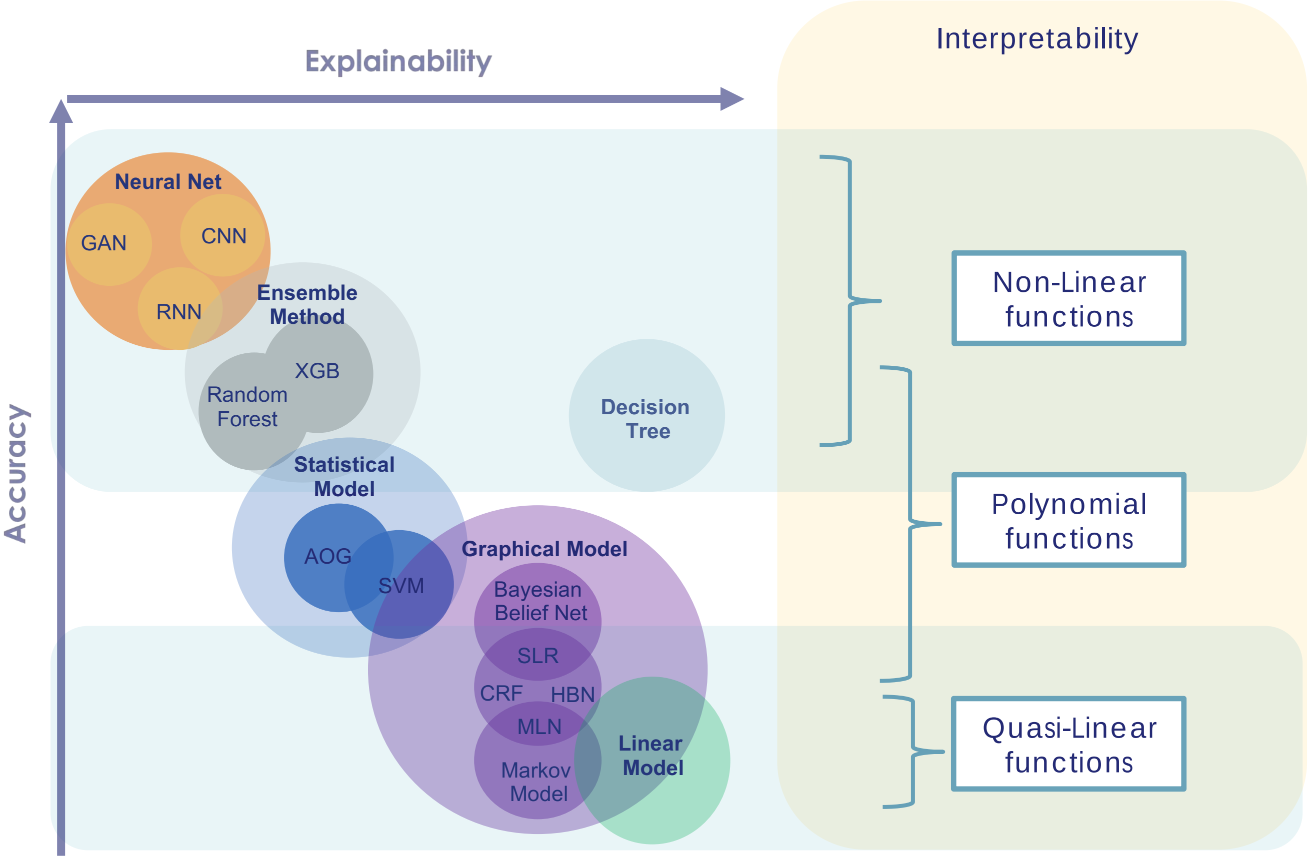
XAI in Machine Learning

THALES

How to Explain? Accuracy vs. Explanability

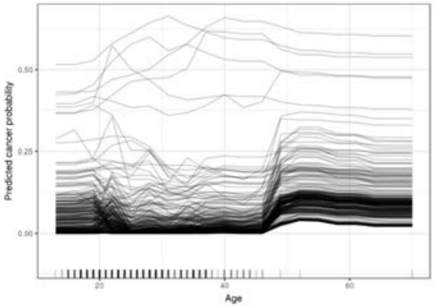
Learning

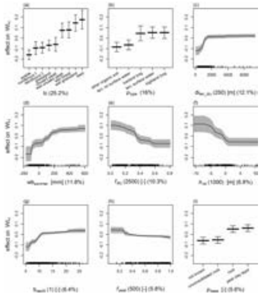
- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - Correlation
  - No causation

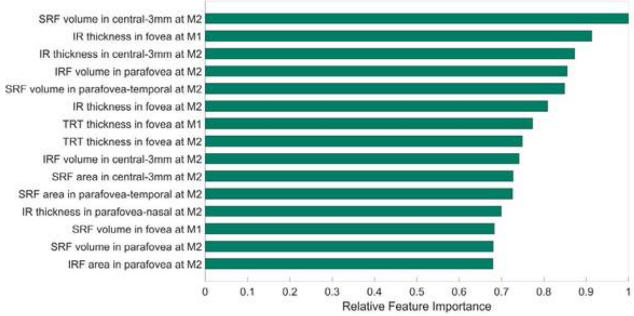


Overview of explanation in Machine Learning fields (1)

Machine Learning (except Artificial Neural Network)





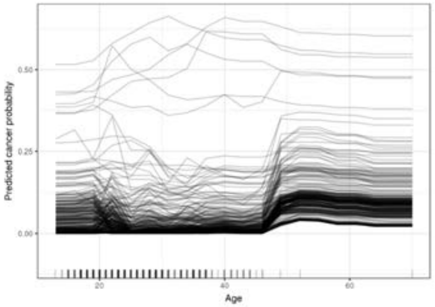


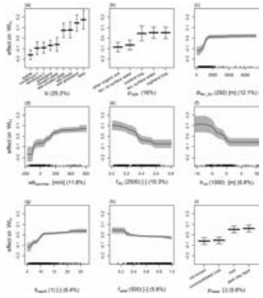
(a)

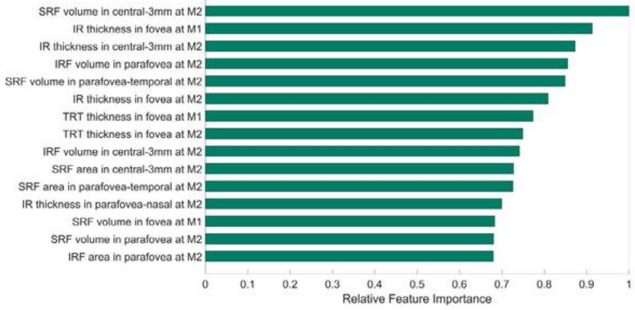
Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

Overview of explanation in Machine Learning fields (1)

Machine Learning (except Artificial Neural Network)

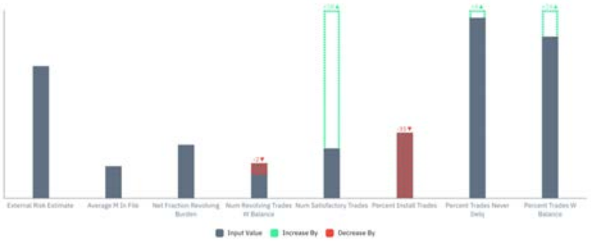






(a)

Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis



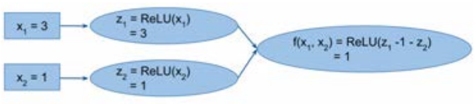
Counterfactual
What-if

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)

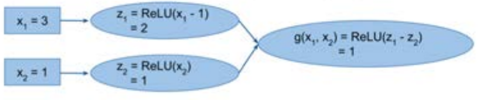
Overview of explanation in Machine Learning fields (2)

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 $x_1 = 3, x_2 = 1$

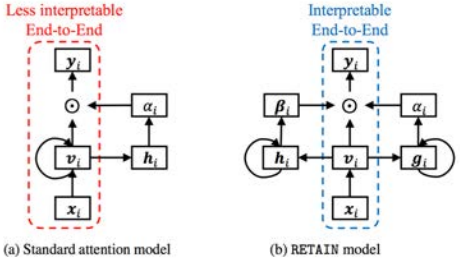
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

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



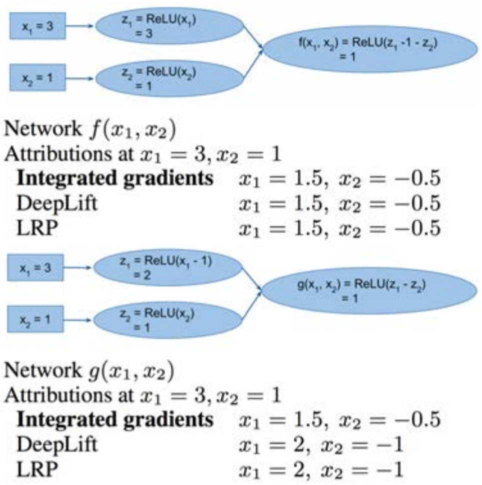
D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

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

THALES

Overview of explanation in Machine Learning fields (2)

Machine Learning (only Artificial Neural Network)

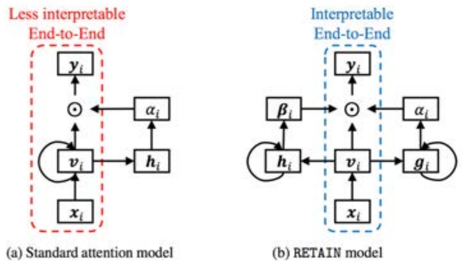


Attribution for Deep Network (Integrated gradient-based)

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

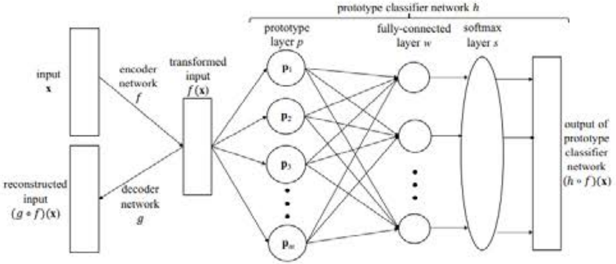
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

Attention Mechanism



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

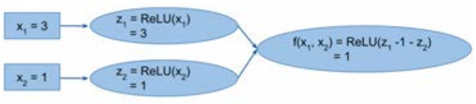


Auto-encoder

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

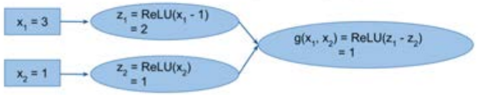
Overview of explanation in Machine Learning fields (2)

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 $x_1 = 3, x_2 = 1$

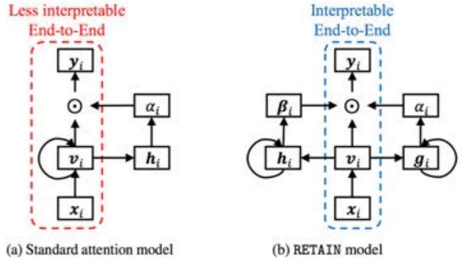
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

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.

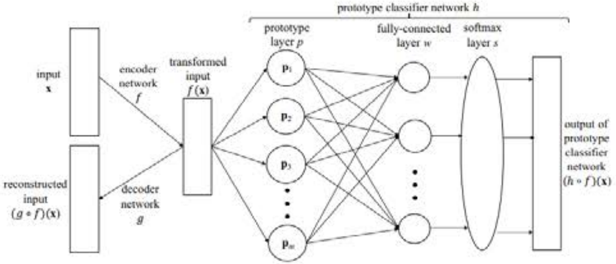
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

Attention Mechanism



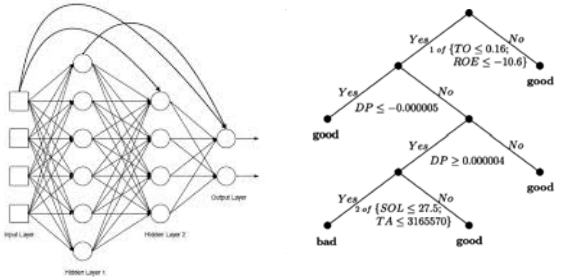
D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

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



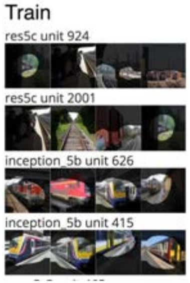
Surogate Model

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

THALES

Overview of explanation in Machine Learning fields (3)

Computer Vision

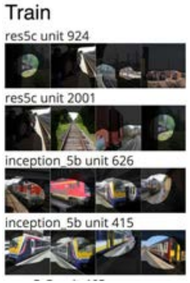


Interpretable Units

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

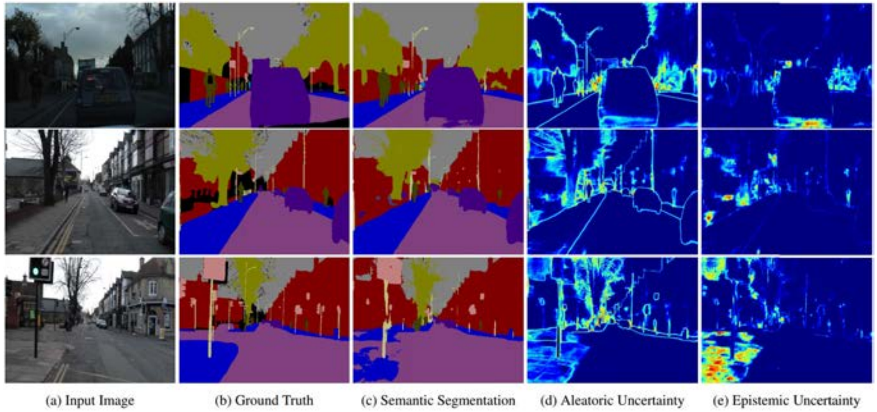
Overview of explanation in Machine Learning fields (3)

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

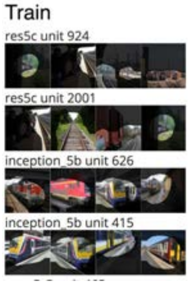


Uncertainty Map

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

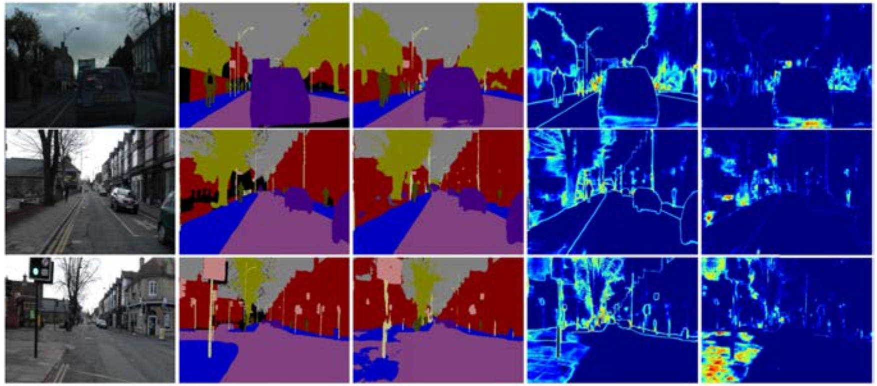
Overview of explanation in Machine Learning fields (3)

Computer Vision



Interpretable Units

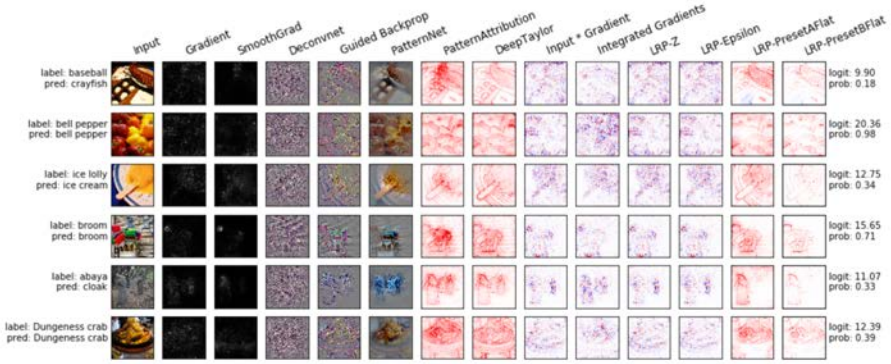
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



(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

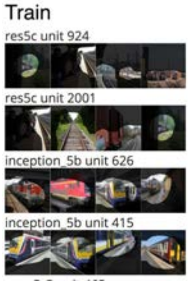


Saliency Map

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

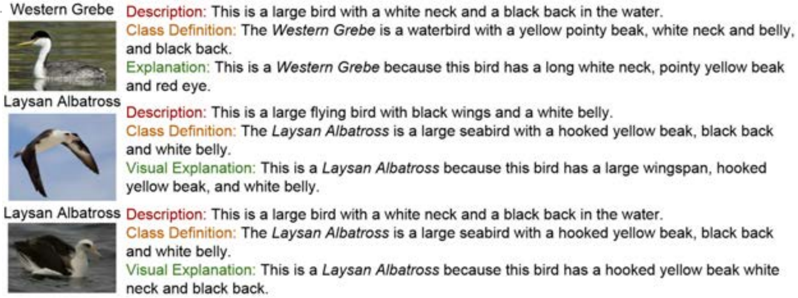
Overview of explanation in Machine Learning fields (3)

Computer Vision



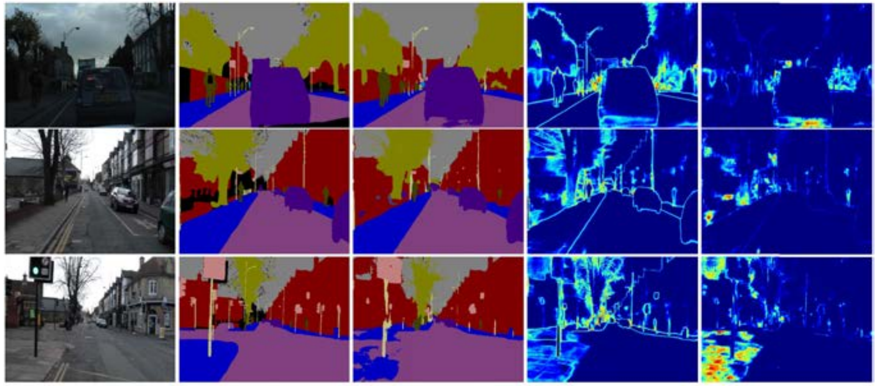
Interpretable Units

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



Visual Explanation

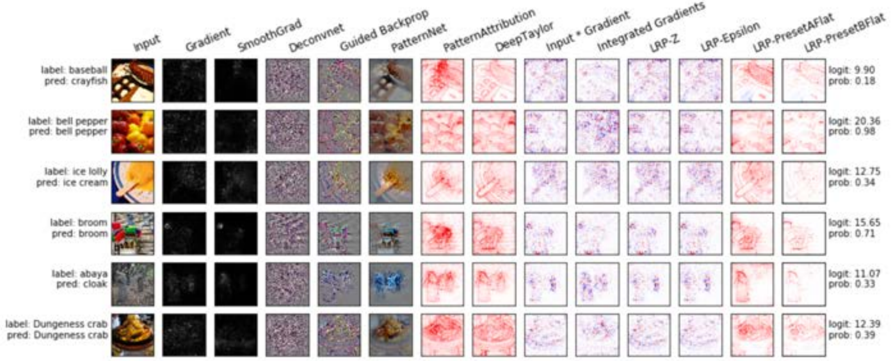
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



(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



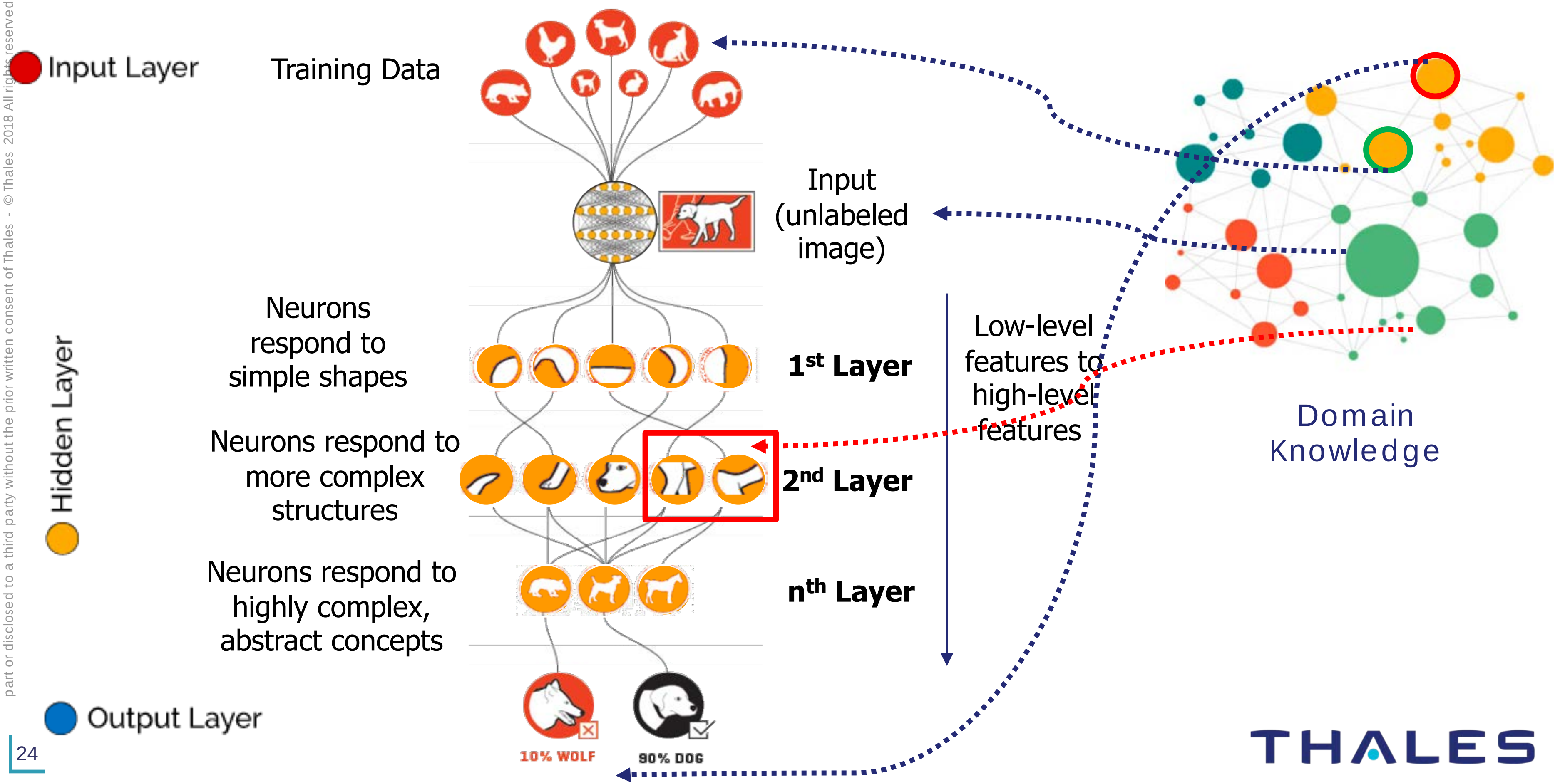
Saliency Map

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

XAI MUST-HAVE in INDUSTRY

THALES

XAI Must-Have in Industry: On Neural Network Architecture



THALES

XAI Must-Have in Industry: On Outputs

Description 0: Two trains

THALES

XAI Must-Have in Industry: On Outputs

Description 0: Two trains

Description 1: This is a train accident including a orange train

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

XAI Must-Have in Industry: On Evaluation

Comprehensibility	Succinctness	Actionability	Reusability	Accuracy	Completeness
How much effort for correct human interpretation?	How concise and compact is the explanation?	What can one action, do with the explanation?	Could the explanation be personalized?	How accurate and precise is the explanation?	Is the explanation complete, partial, restricted?

Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

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

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, statistic
- 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