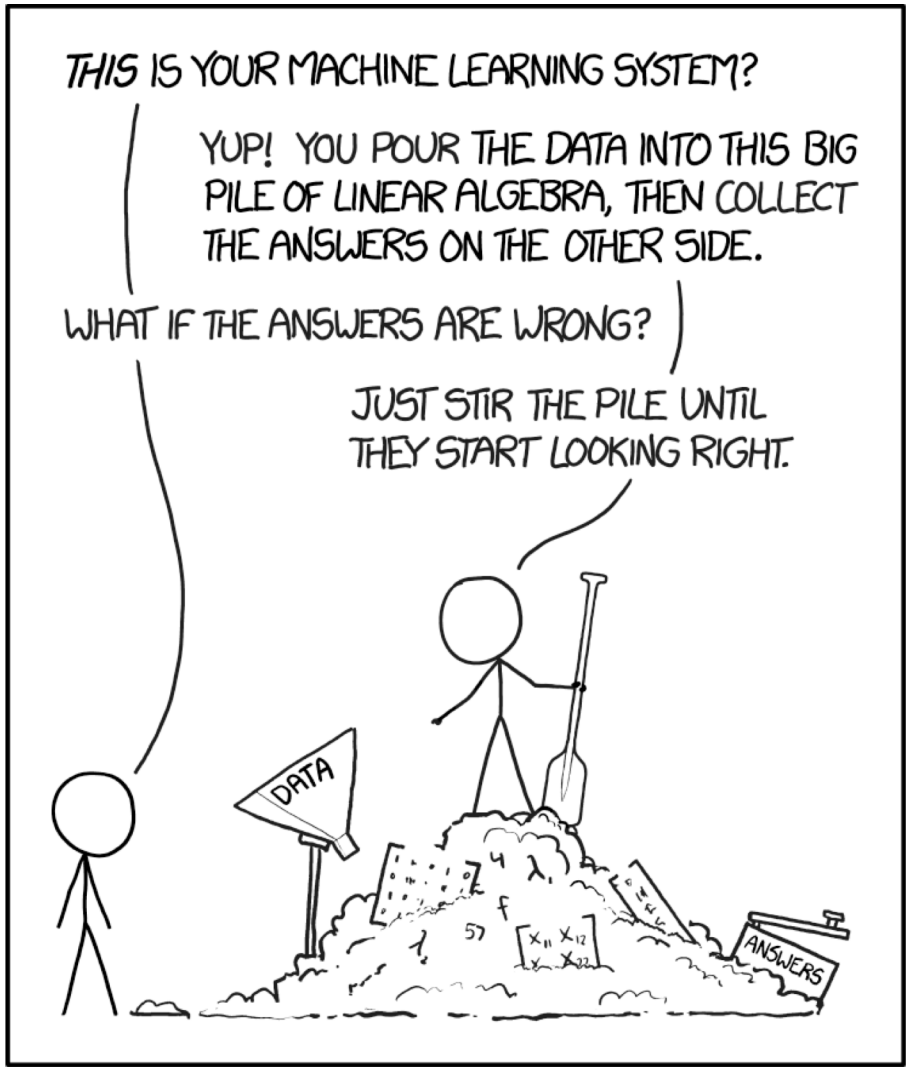
Explainable AI and Semantic Web

Md Kamruzzaman Sarker **Wright State University**

Outline

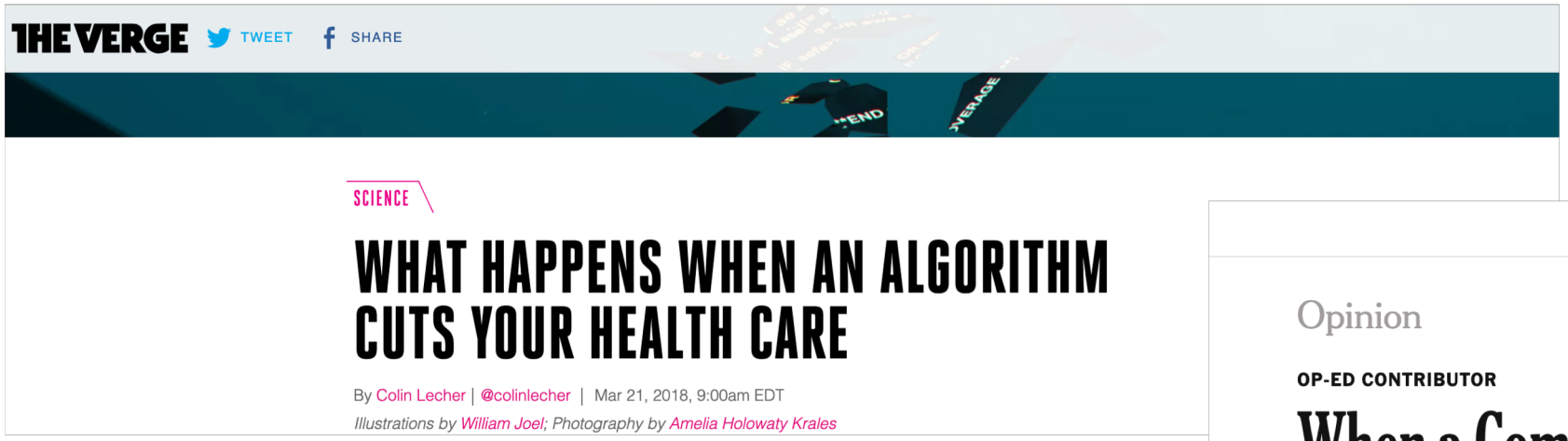
- Motivation
- What is explainable AI
- State of the art in different AI fields
- Limitation
- Scope of using Semantic web technologies

Our Machine Learning Model



https://imgs.xkcd.com/comics/machine_learning_2x.png

Consequences


THE VERGE TWEET SHARE

SCIENCE

WHAT HAPPENS WHEN AN ALGORITHM CUTS YOUR HEALTH CARE

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

Illustrations by William Joel; Photography by Amelia Holowaty Krales

https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy


The New York Times

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

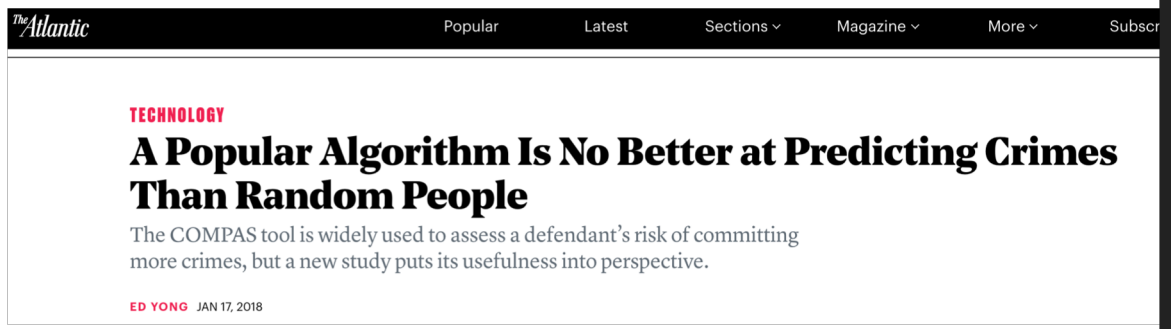
June 13, 2017

232

http://nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

- Artificial Intelligence is being used in Health Care, Judiciary Systems and has great impact on our life.

Consequences


The Atlantic Popular Latest Sections Magazine More Subscr

TECHNOLOGY

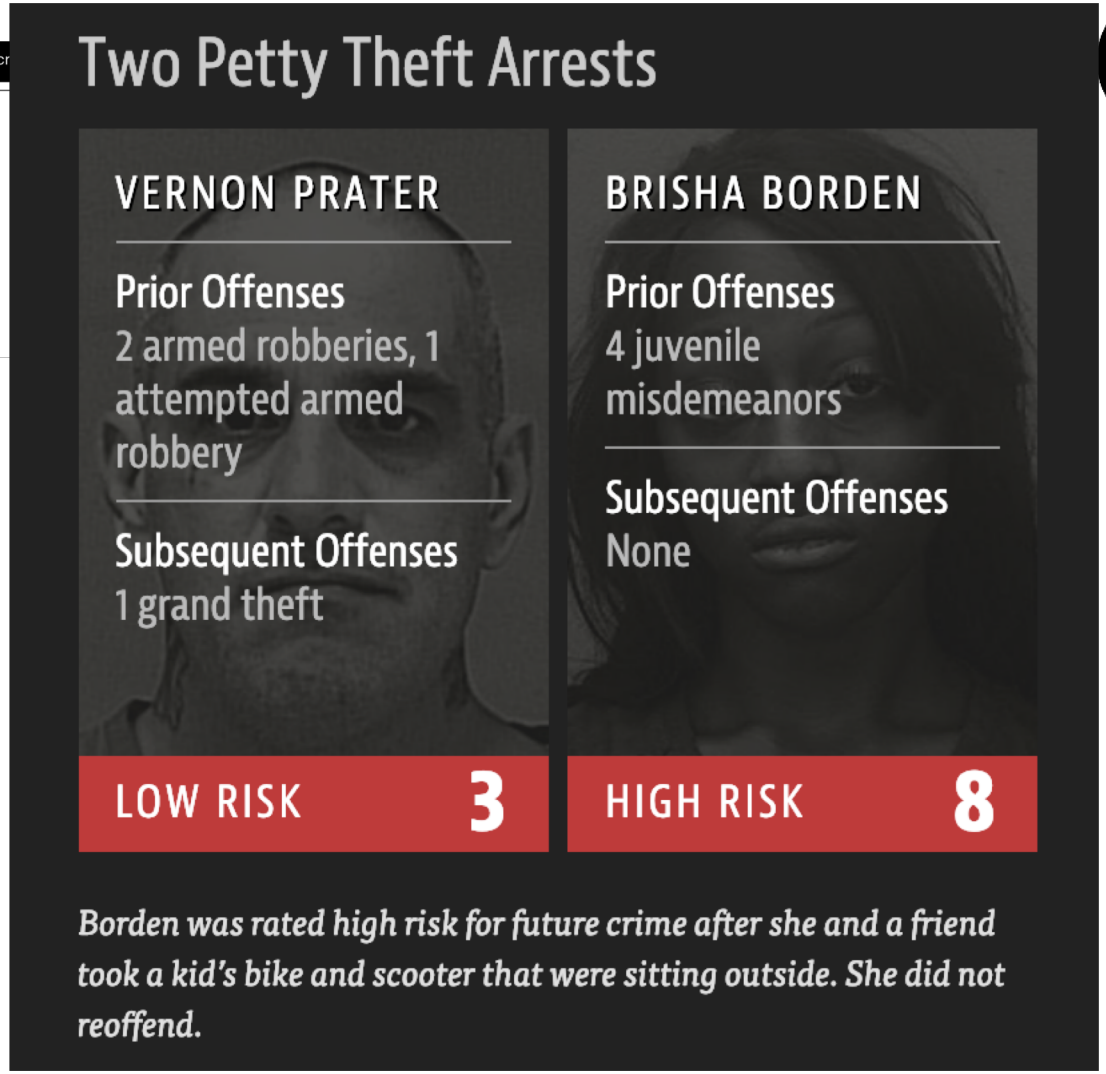
A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant's risk of committing more crimes, but a new study puts its usefulness into perspective.

ED YONG JAN 17, 2018

https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

- The COMPAS tool is widely used to assess a defendant's risk of committing more crimes.
- But a new study puts its usefulness into perspective.

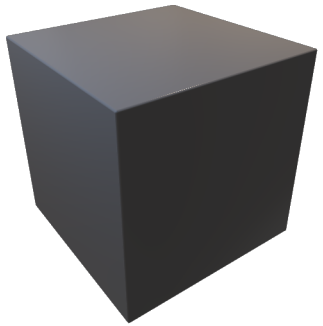


AI Model

$$Y = f(X)$$

Minimize loss is the objective, non-linear transformation is the usual tool.

AI Model



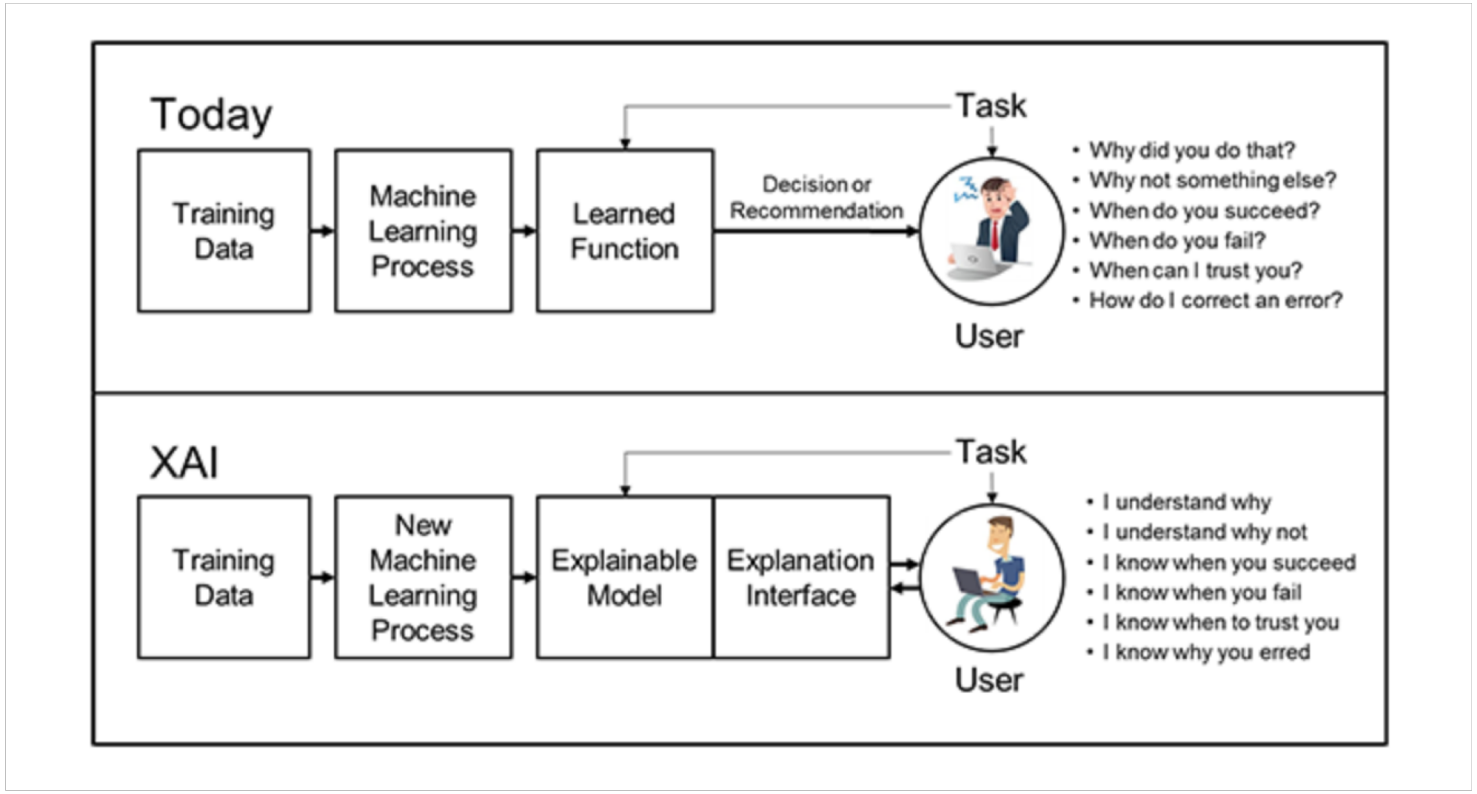
Warehouse

Why it's a warehouse?
Why it's not a market?
How it decided to call it warehouse?

We are using AI systems which we don't understand.
We call it black box.

Explainable Artificial Intelligence

- We need to understand what is going on.
- Explainable AI are the techniques to -
  - explain the decision
  - understand the decision
  - debug the decision



Explainable Artificial Intelligence

Mr. David Gunning, 2017, https://www.darpa.mil/program/explainable-artificial-intelligence

So what is Explainable AI

- No formal, technical, agreed upon definition!
- Comprehensive philosophical overview out of scope of the tutorial
- Explainability is an ill-defined term! [Lipton 2016]
- Different level of people need different type of explainability!
- Not limited to machine learning! [Lipton 2016, TomseL et al. 2018, Rudin 2018]

So what is Explainable AI

DaSe Lab

SSC

Ontology Design pattern for explanation

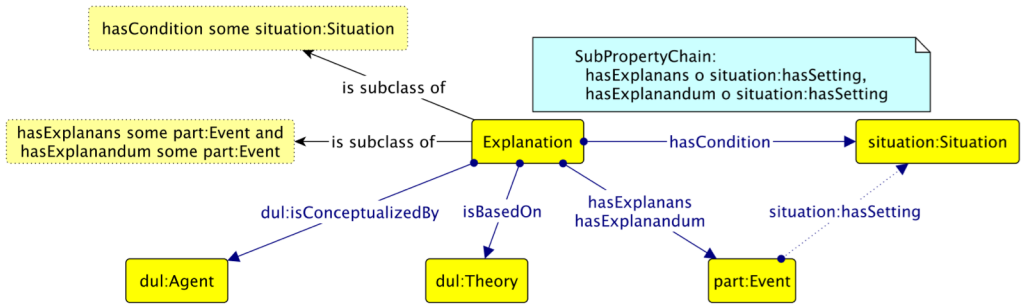


Figure 2: The Explanation Ontology Design Pattern.

[Ilaria 2015]

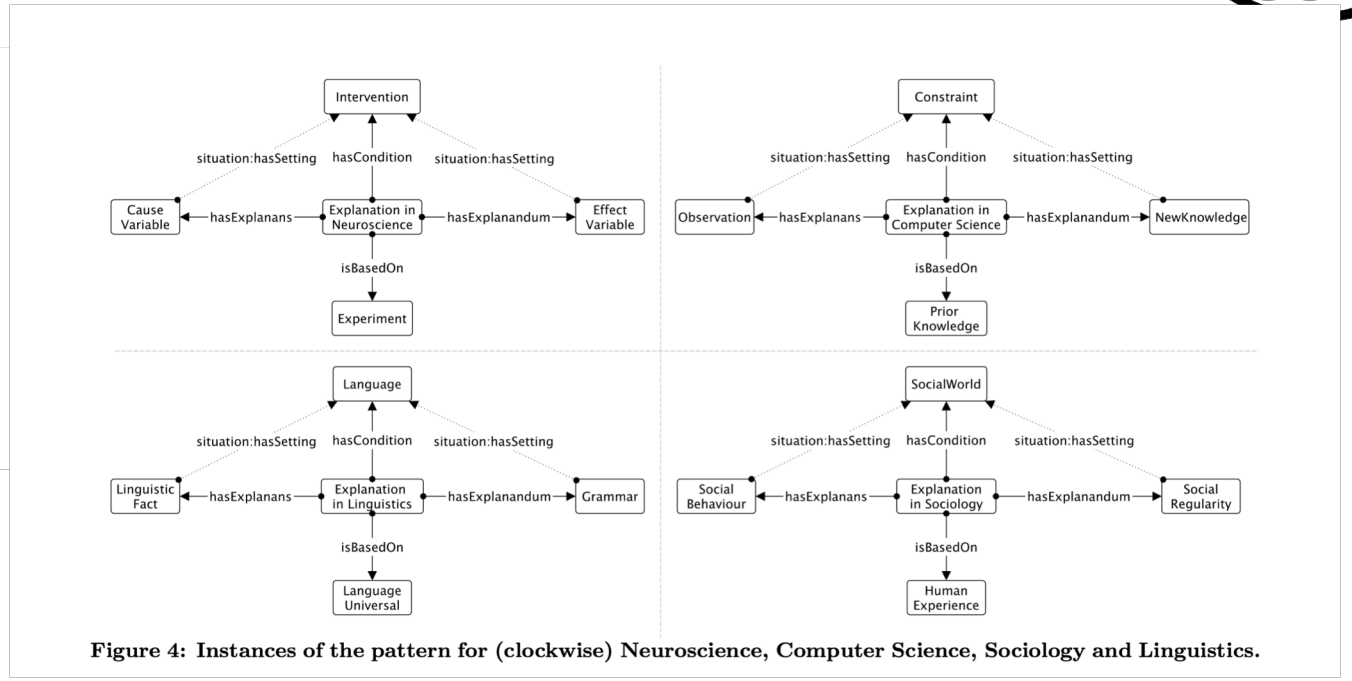


Figure 4: Instances of the pattern for (clockwise) Neuroscience, Computer Science, Sociology and Linguistics.

Explanation differs by field !!

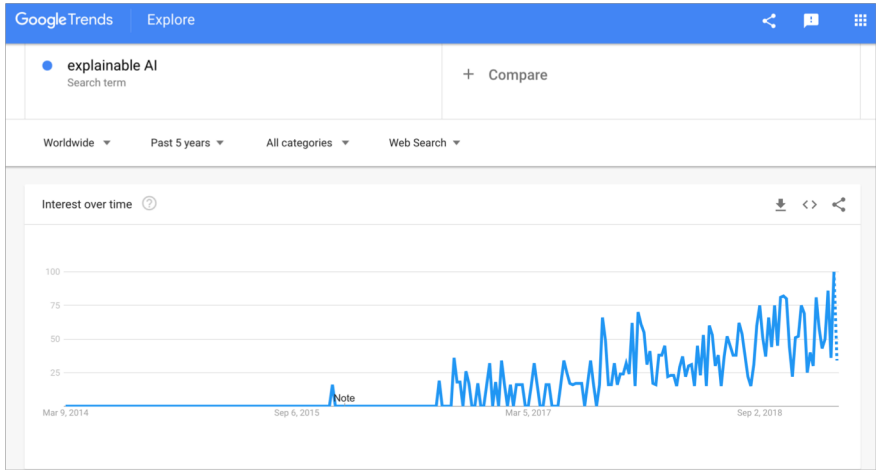
Our (arguable) definition

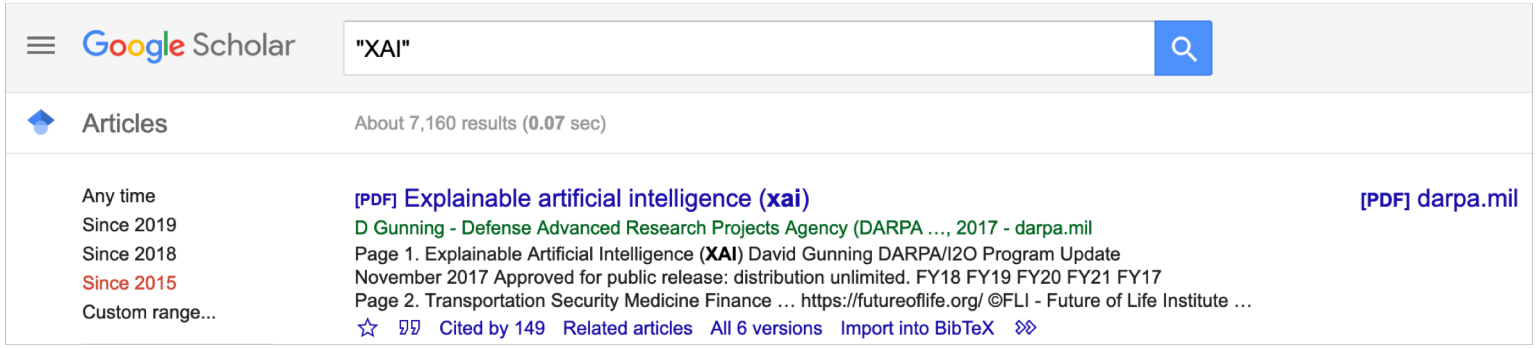
Explainable AI is a branch of AI which help to get explanation from the AI model.

Explanation is the ability of the model to satisfy users understandability of the decision and model.

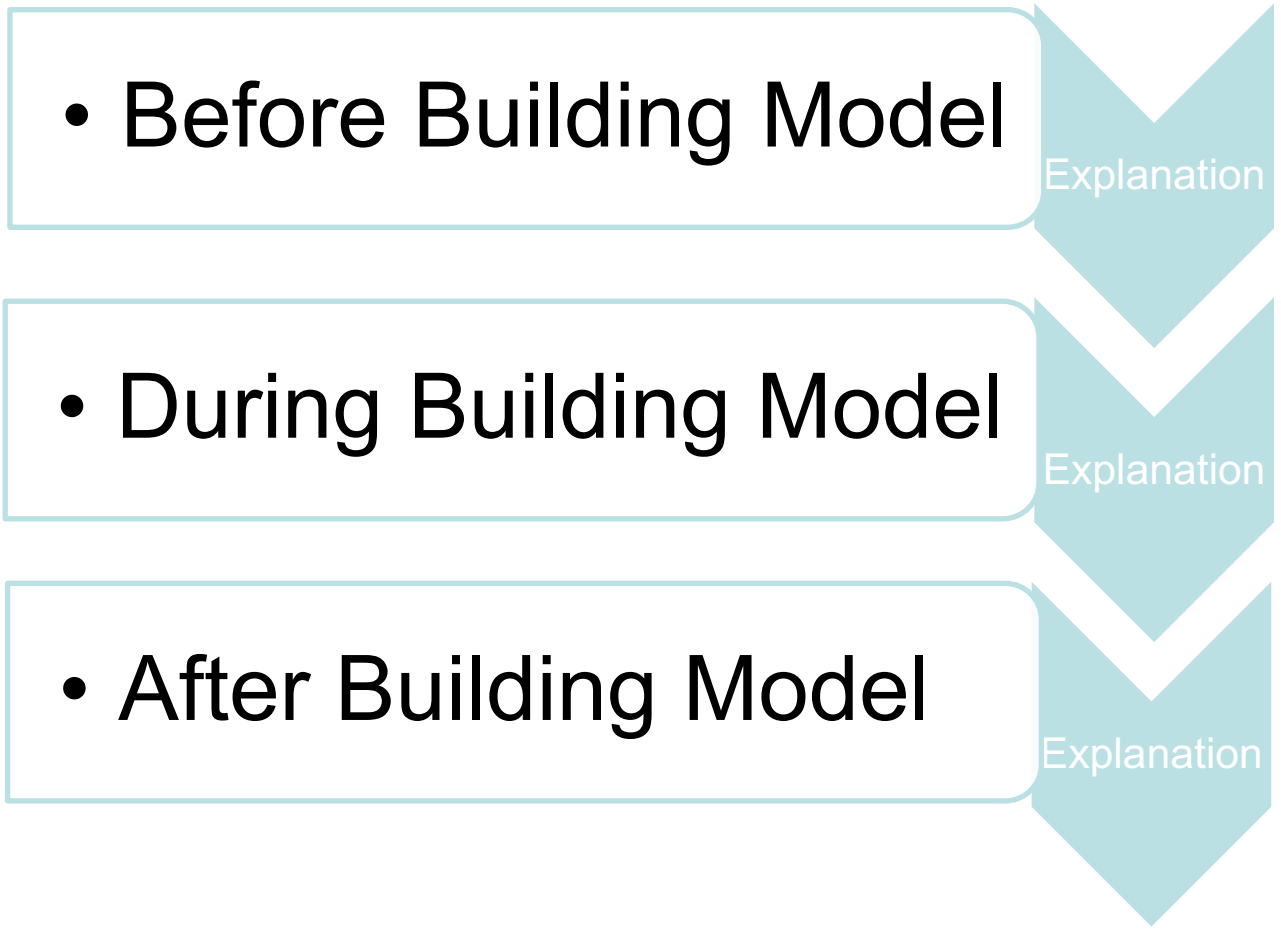
Explainable AI

- Moving from the definition we need to explain the AI models.
- Trend on google shows large interest for this.
- Researchers are also working hard to find ways.
- More than 7000 papers related to XAI since 2015.





Options for Explanation



Before Building Model

- Feature Visualization[1]
- Exploratory data analysis[2]
  - Some features may be redundant
  - Some features may not help the model

- Correlation can be used to identify redundant features. [The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie Robert Tibshirani Jerome Friedman]

How much this help for explanation?

[1,2]
[Viégas and Wattenberg '07]
[Maaten et al. '08]
[Amershi et al. '09]
[Patel et al. '10]
[Varshney et al. '12]
[Tukey 77]

Feature Visualization & Exploratory data analysis

DaSe Lab

SSC

- Word Cloud for 20 news group dataset for different classes.
- Helps to get idea for which class has which type of texts.

http://qwone.com/~jason/20Newsgroups/

How much this help for explanation?

- Give some idea about the dataset.
- Different classifier work differently, so not giving explanation about the model.
- If features are not understandable (numeric) this only helps to get cluster of group, but does not help for human understanding of the classifier decision or model.

WRIGHT STATE UNIVERSITY

- Before Building Model — Explanation
- During Building Model — Explanation
- After Building Model — Explanation

Not satisfying us, let's go to next step

During Building Model

- **Choosing interpretable model**
  - Common belief that some models are explainable.
    - Rule Based model
    - KNN
    - Linear Regression
    - Decision tree
  - Are they really interpretable?

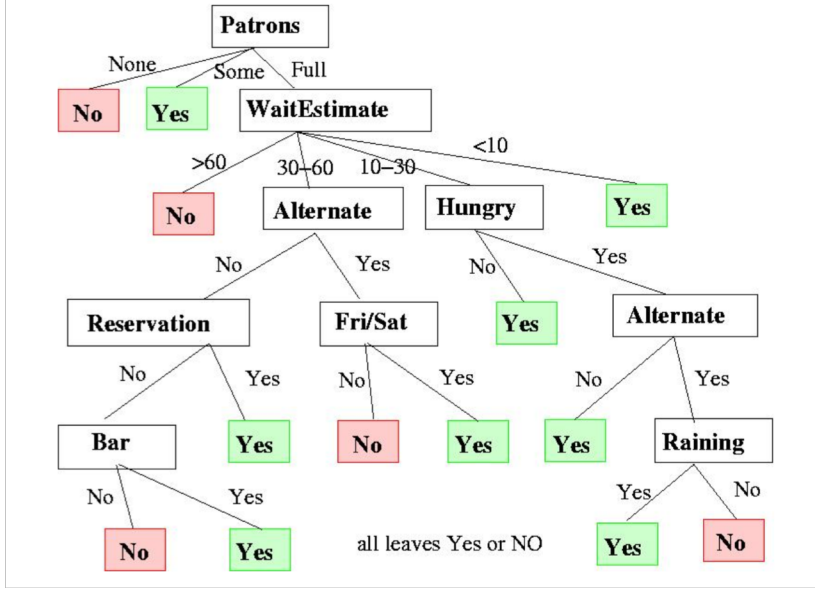
During Building Model

DaSe Lab

SSC

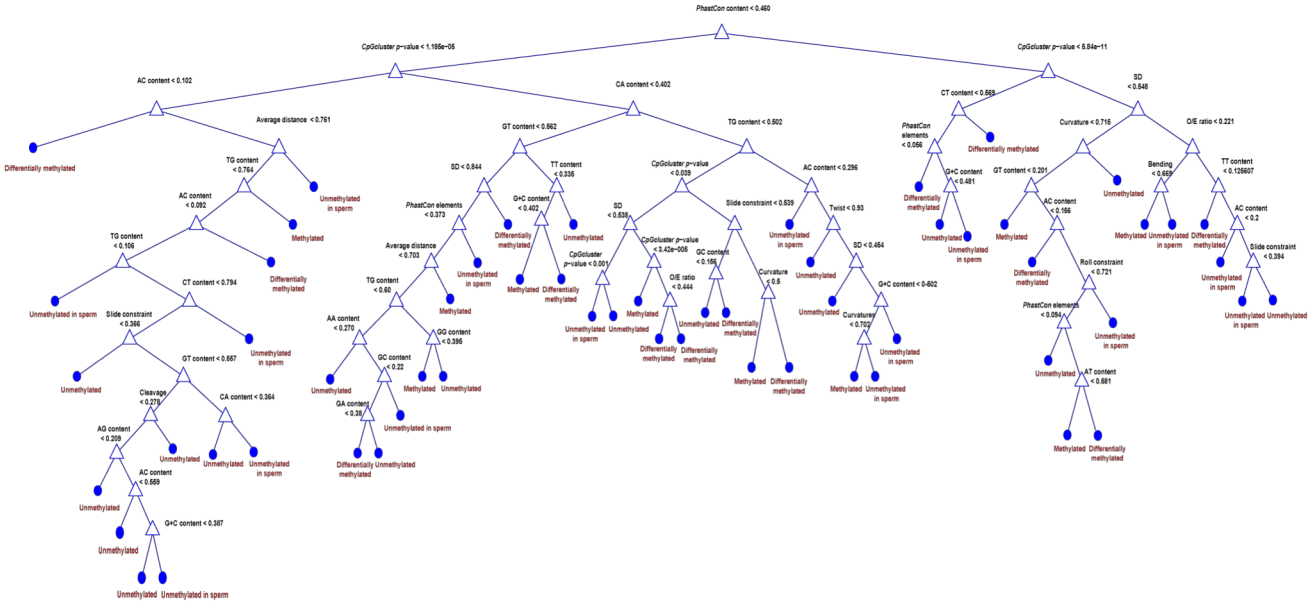
Decision Tree : Example

- Problem to wait or not for a table at a restaurant?
- Understandable as features are understandable.



http://www.cs.bham.ac.uk/~mmk/Teaching/AI/l3

- Problem is to predict the class from four CGI methylation classes.
- Hard to understand and even hard to visualize.



[Previti 2009]

During Building Model

- **Choosing interpretable model**
  - Develop rule based model which has theoretical underpinning to the solution.
    - [Breiman, Friedman, Stone, Olshen 84]
    - [Rivest 87]
    - [Muggleton and De Raedt 94]
    - [Wang and Rudin 15]
  - Using rule set user can trace how the solution is being made.
  - For large systems making rule set is not feasible.
    - Face recognition system with thousand of features!!!!!
  - For some systems making rule set is not even practical.
    - Music classifier where input comes as waveform.

During Building Model

- Rule Based model
- KNN
- Linear Regression
- Decision tree

Rule based model are not scalable.

Others are interpretable as long as the model is small.

- Before Building Model

Explanation

Not satisfying us, let's go to next step

- During Building Model

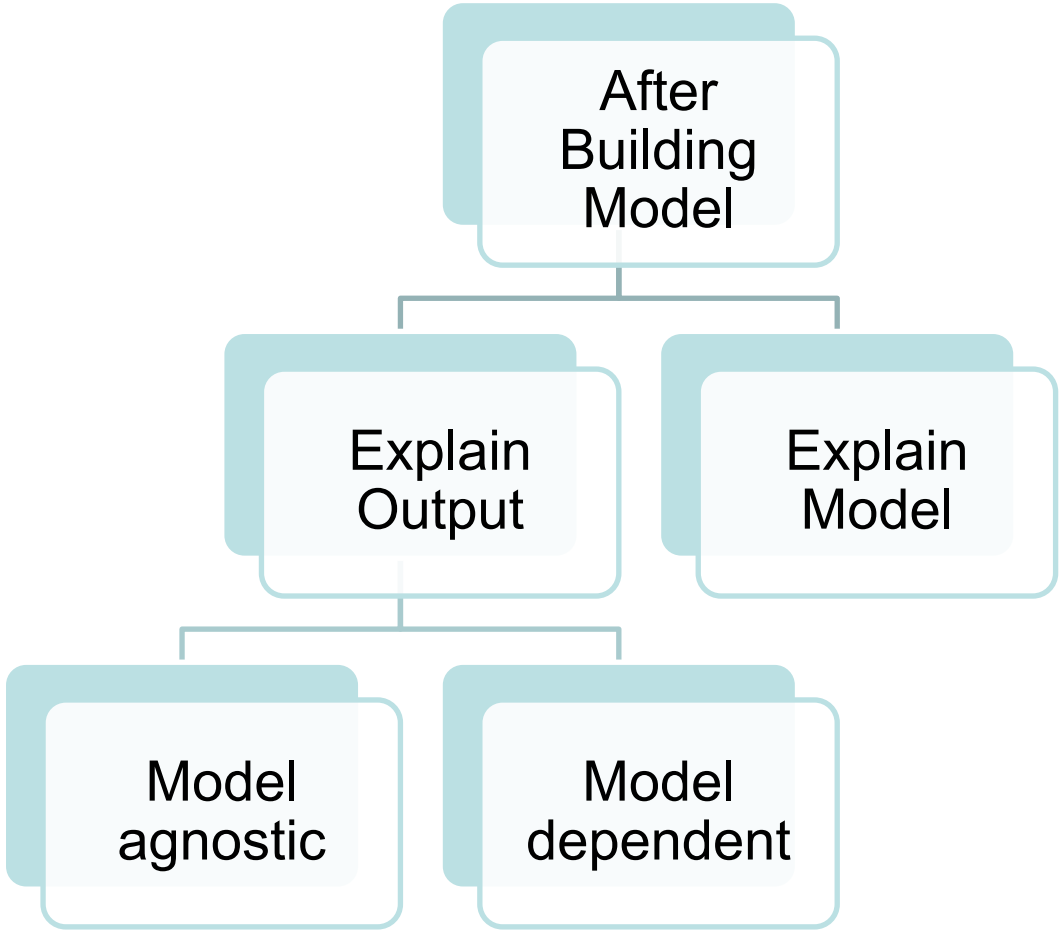
Explanation

Not satisfying us, let's go to next step

- After Building Model

Explanation

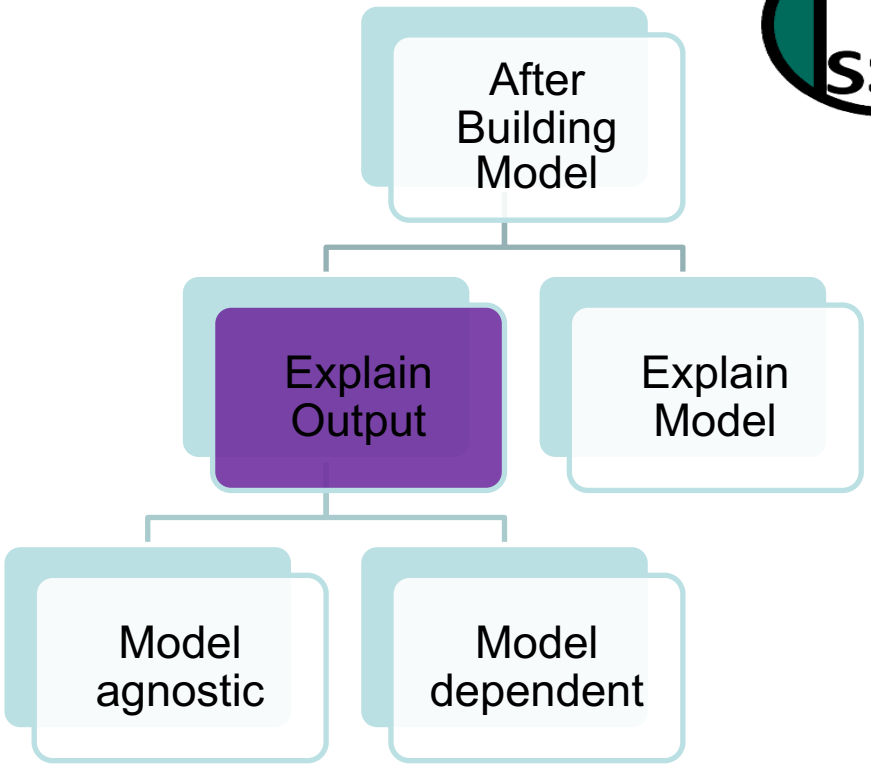
After Building Model



Explain Output

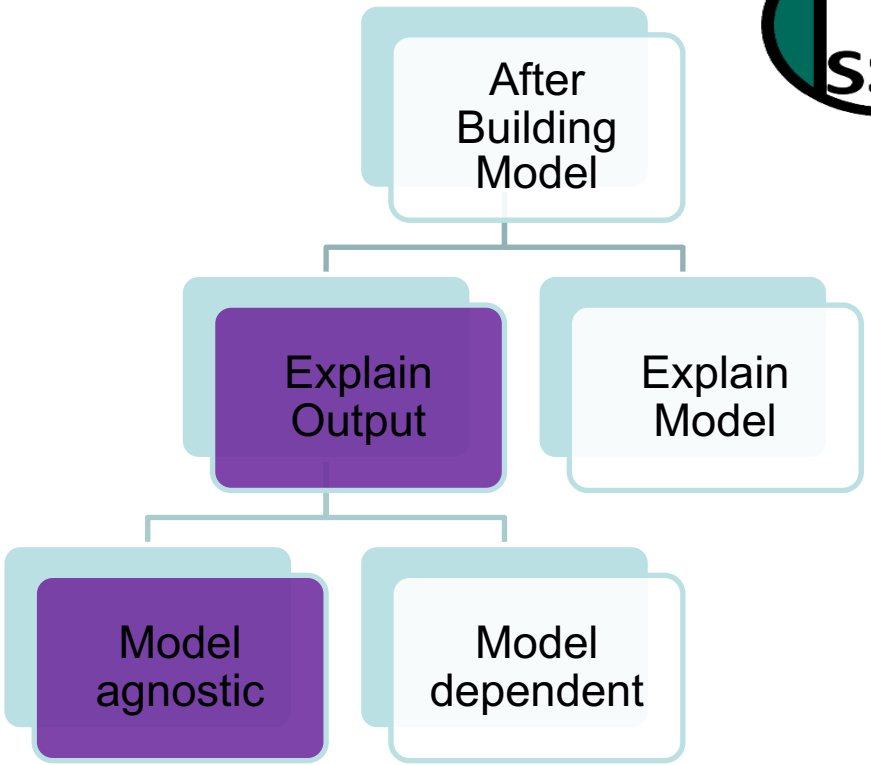
- Some explanation are model dependent and some are not. [Riccardo, 2018]

- Large class of AI models
  - Deep Neural Network
  - Convolutional Neural Network
  - LSTM
  - Recurrent Neural Network
  - SVM



Model Agnostic

- These techniques try to explain any AI model.
- $y = f(x)$
- f is any AI model, it can be
  - Neural Network
  - Support Vector Machine
  - Random Forest
  - ……



Model Agnostic

LIME : Local Interpretable Model-Agnostic Explanations[1]
Anchors: High-Precision Model-Agnostic Explanations[2]

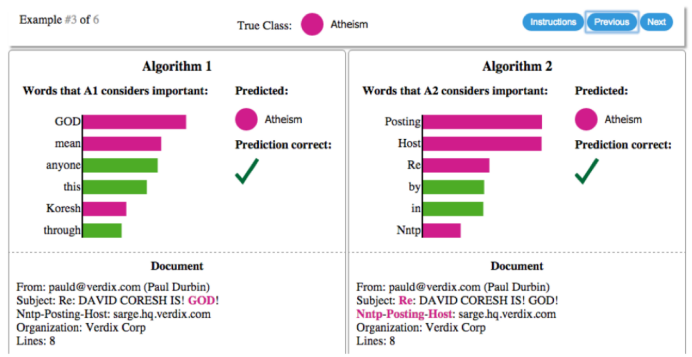


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

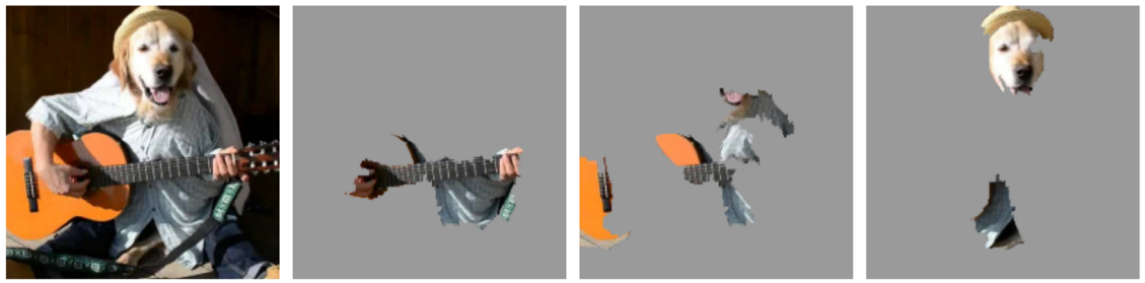
What animal is featured in this picture ?	**dog**
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

Figure 3: Anchor Explanations for Image Classification and Visual Question Answering (VQA)



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

- Explain models by presenting representative individual predictions
- Perturb the input data around the prediction

[1][Ribeiro 2016], [2][Ribeiro 2018]

Model Agnostic

- **LORE : LOcal Rule-based Explanations**
  - Uses a genetic algorithm approach to generate a decision tree as interpretable classifier
  - Explanation consists in a rule derived from the decision tree classifier.
  - LORE also returns a set of counterfactual rules, suggesting the changes in the instance's features of x that may lead to a different outcome.

$$
\begin{aligned}
r = \quad & (\{credit\_amount \leq 836, housing = own, other\_debtors = \\
& \quad none, credit\_history = critical\ account\} \rightarrow decision = 0) \\
\Phi = \quad & \{\ (\{credit\_amount > 836, housing = own, other\_debtors = \\
& \quad none, credit\_history = critical\ account\} \rightarrow decision = 1), \\
& (\{credit\_amount \leq 836, housing = own, other\_debtors = \\
& \quad none, credit\_history = all\ paid\ back\} \rightarrow decision = 1)\ \}
\end{aligned}
$$

[Guidotti, 2018]

Model Agnostic

DaSe Lab

SSC

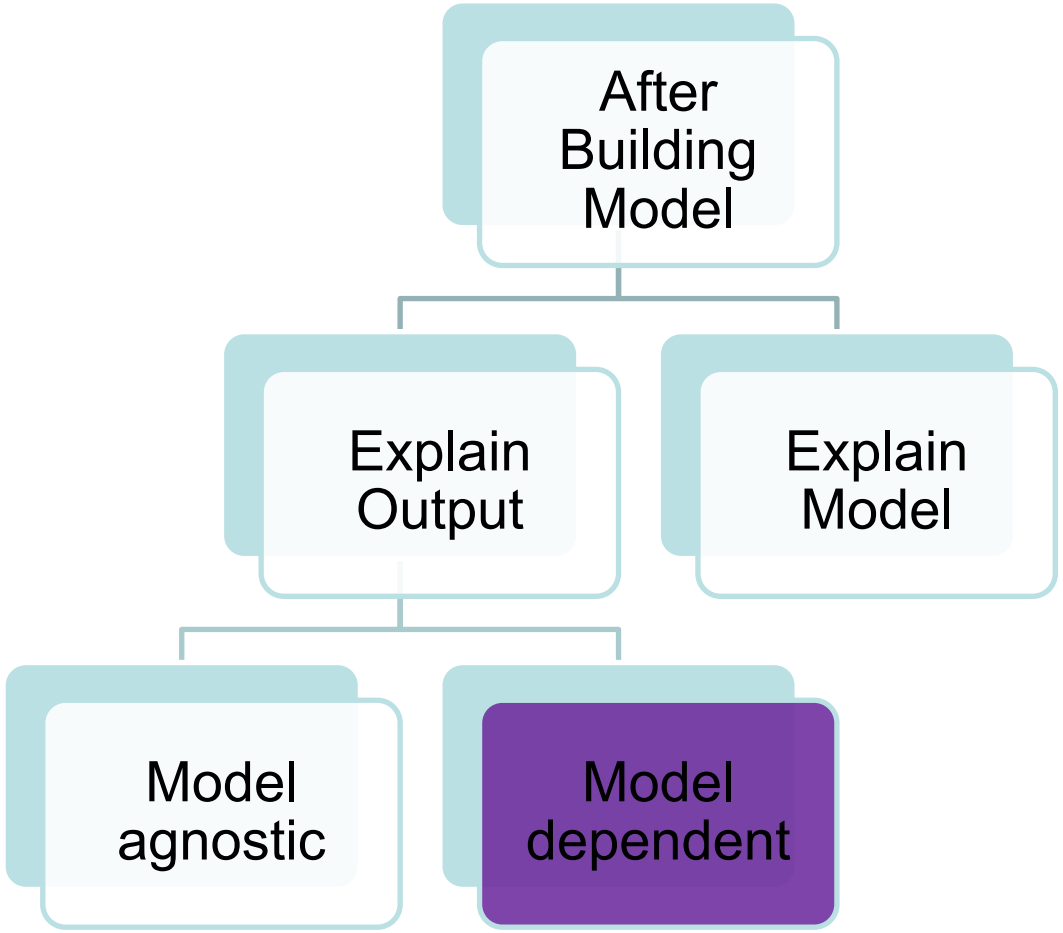
Limitation

- Explanation are not stable.
- Same model m, same input x can generate different explanation.
- Explanation only faithful locally.
- If input x come from different distribution explanation are terrible.

After Building Model

- Model dependent explanations are mostly for neural networks.

- Mainly two approaches are used.
  1. **Investigate the hidden layers**
  2. **Build another model to create explanation**



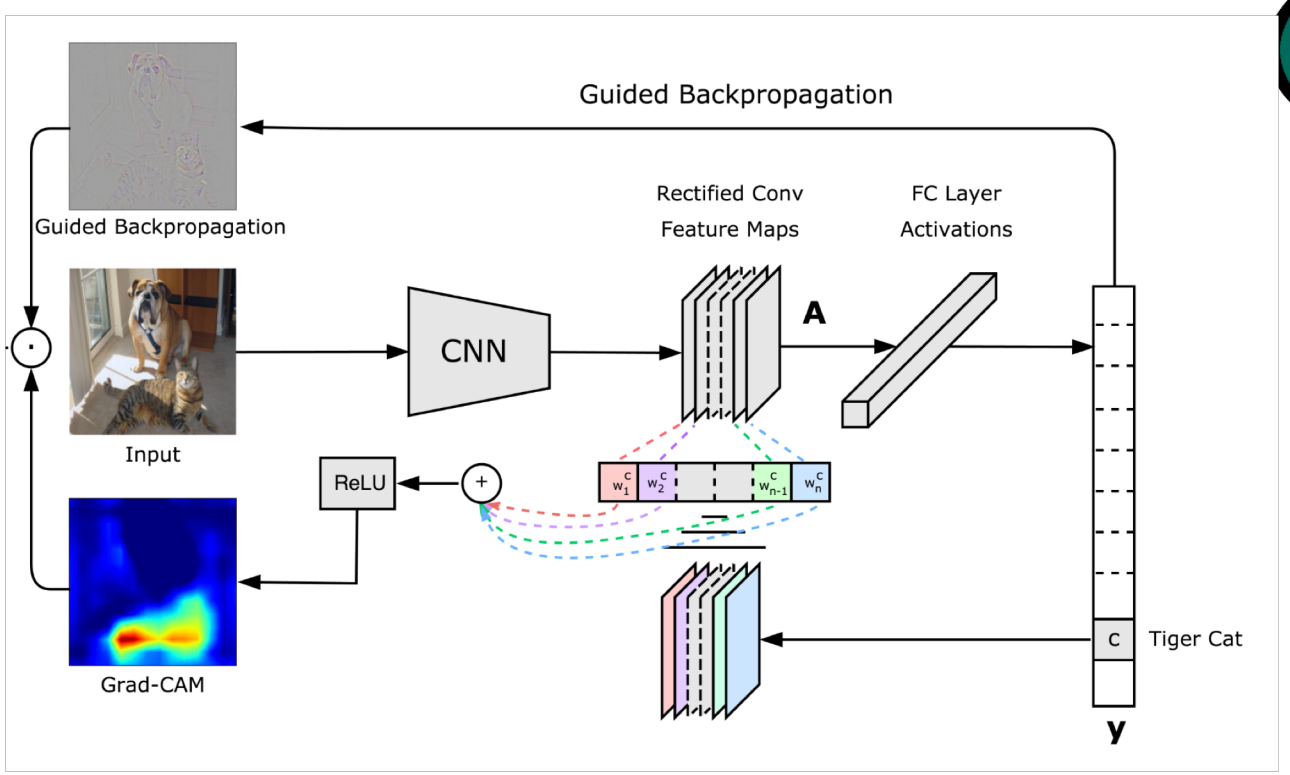
Model Dependent

Investigate the hidden layers

Visualizing the regions of the input that are "important" for predictions.

Common Strategy:

1. Backpropagate the output/loss/gradient to the input layer.
2. Compares the activation of each neuron to its 'reference activation'
3. Show the difference.



Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", https://arxiv.org/abs/1610.02391

Model Dependent

Investigate the hidden layers

Showing important part of the input
which contributes more to the class decision.

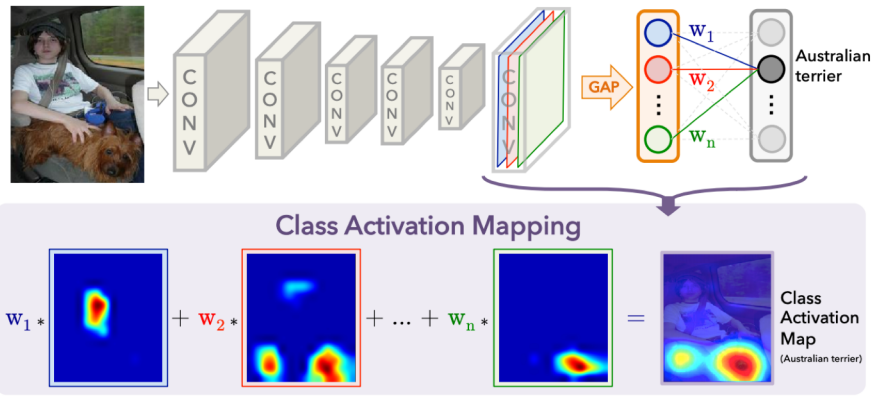


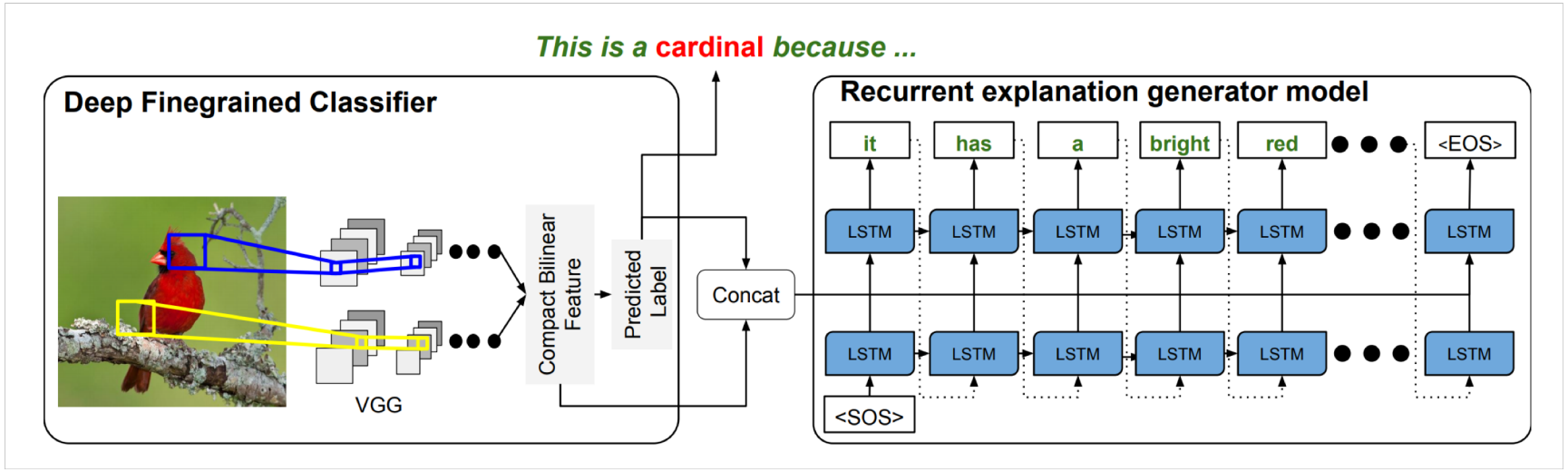
Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba,
Learning Deep Features for Discriminative Localization, https://arxiv.org/abs/1512.04150

Model Dependent

Build another model to create explanation

Train a black box and then **again train another black box** to produce explanation!!



Visual explanations [Hendricks et al. '16]

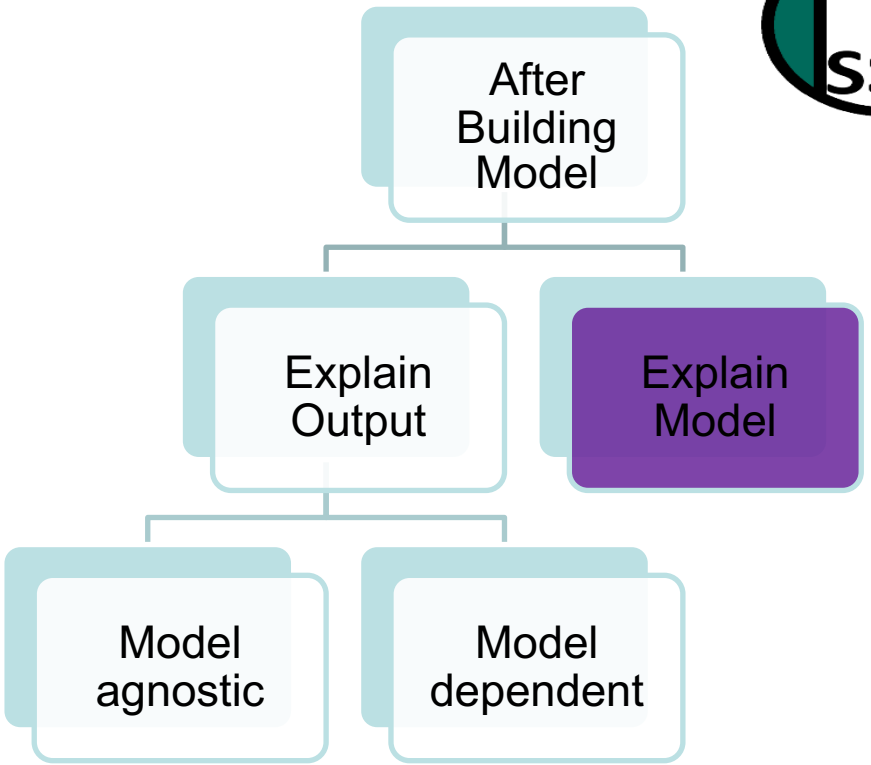
What will happen if the explanation module produces wrong explanation ??

Limitation

- Some time explanation are hard to understand.
- Single bit change in the input (even if the if input is large) may give different explanation.
- Newly created explanation module may give wrong answers.

Model Explanation

- Need to understand how the model is producing output.
- Is the model generalized enough so it will not be fooled? (Adversarial attack) [Chakraborty, 2018]
- Mainly two approaches:
  - Rule extraction from neural network.
  - Looking the hidden layers



Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay Adversarial Attacks and Defences: A Survey, https://arxiv.org/abs/1810.00069

Model Explanation

Rule extraction from neural network.

[Sebastian Thrun, 1995]
[Tameru Hailesilassie, 2016]
[Mark W. Craven, 1996]

- Create rule/decision tree from trained neural network.
- Works for small number of features
- and where the features are understandable.

Model Explanation

- **Rule extraction from neural network**
  - Showing rule set for classifying electronics out of 4 class (Space', 'Medicine', 'Electronics' and 'Cryptography' from the 20-newsgroups dataset for text classification).
  - If several features frequently co-occur and infrequently occur with-out each other, this technique may find only one of them.
  - Not scalable for large models
  - Is it easy to understand the rules?

if *(just = -1) and (use = 1)* $\implies$ *electronics* (24/24)
elif *(circuit = 1)* $\implies$ *electronics* (32/32)
elif *(just = -1) and (don = 1)* $\implies$ *electronics* (11/11)
elif *(people = 0) and (used = 1) and (key = 0) and (don = 0) and (use = 0) and (edu = 0) and (medication = 0) and (concept = 0) and (did = 0)* $\implies$ *electronics* (36/43)
elif *(electronics = 1)* $\implies$ *electronics* (17/18)
elif *(battery = 1)* $\implies$ *electronics* (23/23)
elif *(radio = 1) and (shack = 1)* $\implies$ *electronics* (9/9)
elif *(people = 0) and (thanks = 1) and (advance = 1)* $\implies$ *electronics* (12/14)
elif *(signal = 1)* $\implies$ *electronics* (13/15)
elif *(people = 0) and (company = 1) and (just = 0)* $\implies$ *electronics* (13/18)
elif *(pc = 1)* $\implies$ *electronics* (16/19)
elif *(people = 0) and (use = 1) and (just = 0) and (good = 0) and (clipper = 0) and (probably = 0) and (center = 0) and (unless = 0) and (18084tm = 0) and (algorithms = 0)* $\implies$ *electronics* (29/33)
elif *(appreciated = 1) and (time = 0)* $\implies$ *electronics* (11/16)
elif *(voltage = 1)* $\implies$ electronics (8/8)
elif *(program = -1)* $\implies$ electronics (10/15)
else: others (1134/1281)

Figure 2: Set of if-then-else rules that explain the predictions of the neural network for the 'electronics' class when using mutual information feature selection. Here the discrete test value 1 means a positive correlation between a feature value (which we can approximate to relative frequency due to the use of TF-IDF vectors) and the probability of the class, -1 means a negative correlation, and 0 shows an absence of a feature. The values (a/b) mean that a of b instances covered by the rule are correct. The rules with lower values of a/b are less trustworthy, and the rules with lower value of b, especially in the first few conditions, are less generalized.

[Sushil 2018]

Model Explanation

Looking the hidden layers

- Visualize the hidden conv layers
- This approach mostly suited for image classification.
- Some times hidden pictures are not understandable.

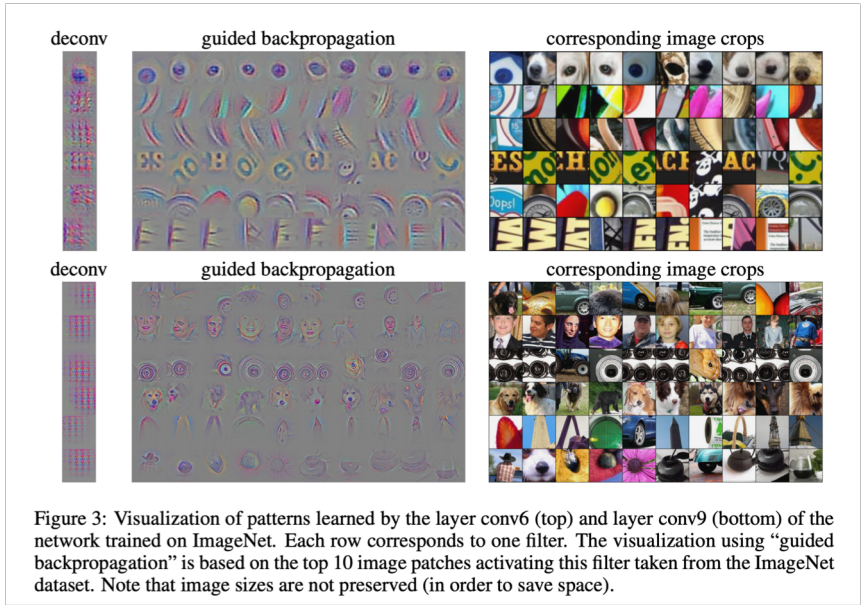
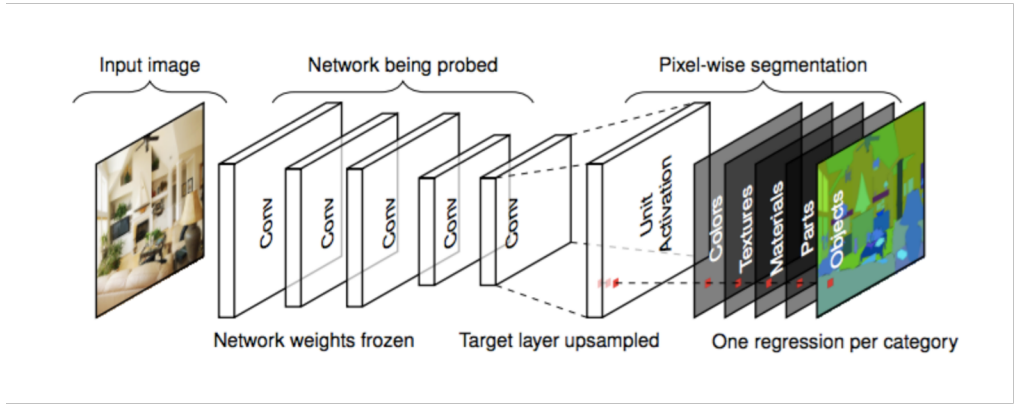
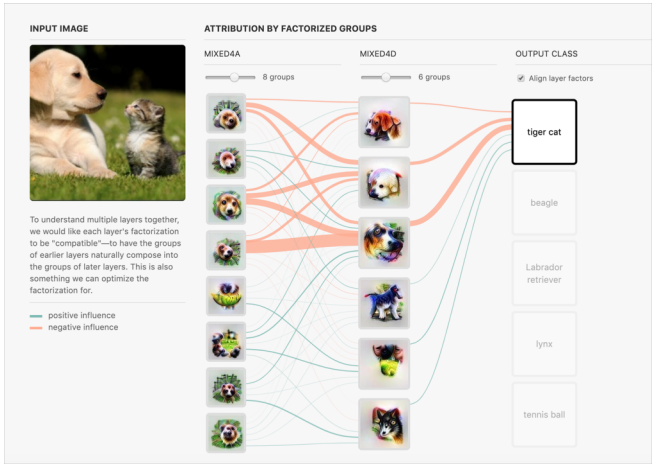


Figure 3: Visualization of patterns learned by the layer conv6 (top) and layer conv9 (bottom) of the network trained on ImageNet. Each row corresponds to one filter. The visualization using "guided backpropagation" is based on the top 10 image patches activating this filter taken from the ImageNet dataset. Note that image sizes are not preserved (in order to save space).

[Springenberg, 2015]





https://distill.pub/2019/activation-atlas/

Limitation

- Ruleset are hard to understand for big models.
- Visualization only helps for image classification
- Parts of the pictures don’t always make understandable object.
- Human understands hierarchical relation easily but, no relationship between the objects are maintained.

- Before Building Model

Explanation

Not satisfying us, let's go to next step

- During Building Model

Explanation

Not satisfying us, let's go to next step

- After Building Model

Explanation

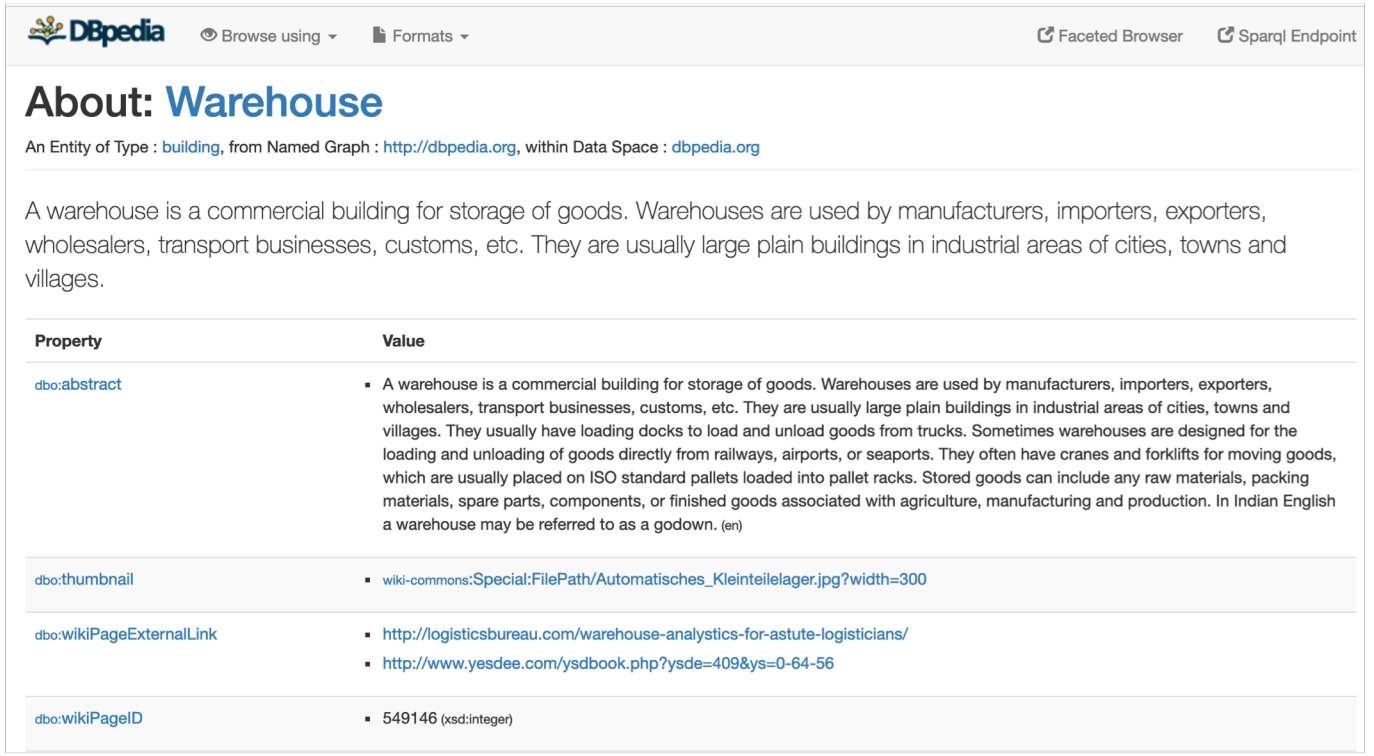
Somewhat satisfying but not fully.

SSC

- **After failing to understand the explanation using XAI what can we do?**
  - **More Research using statistics**
  - **Can we use semantic web techniques?**

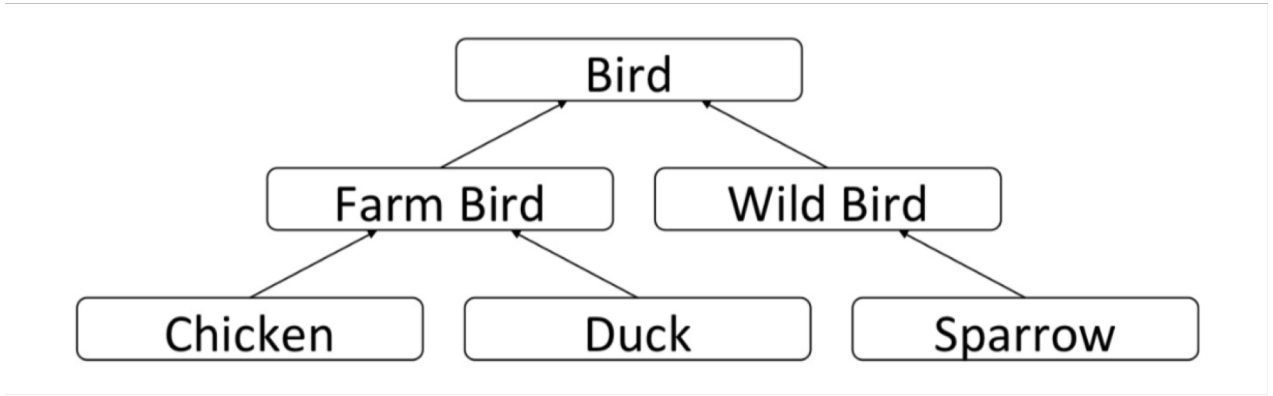
Using Semantics

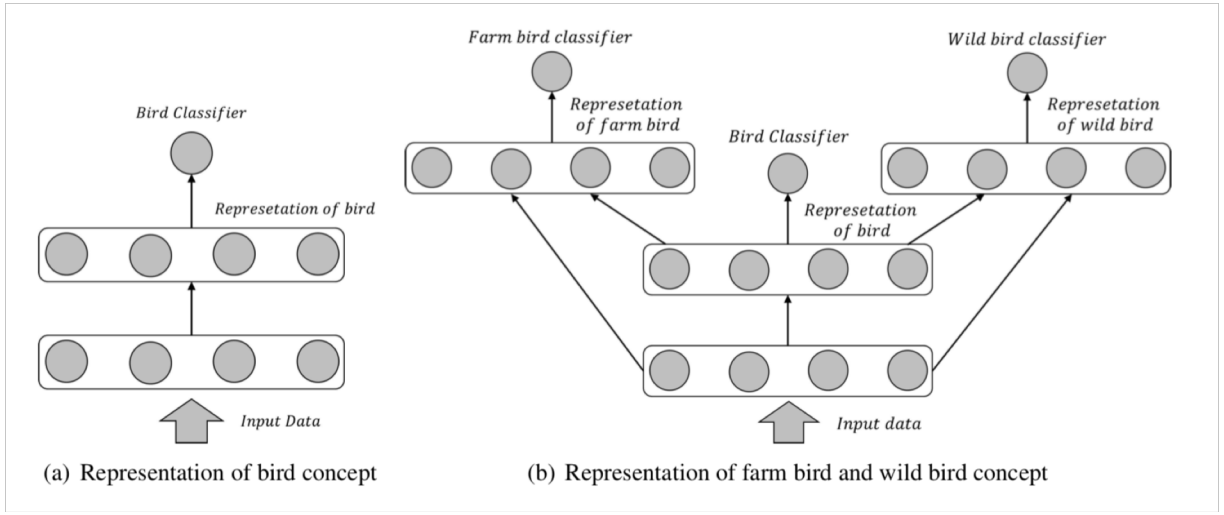
- Semantic web has millions of structured triples.
- And those triples are related to each other.
- Using those related triples may enhance our understandability on what's going inside.



Using Semantics

- **Set the goal of classification as hierarchical classification.**
  - Easily understandable by human.
  - But the problem is optimization gets even harder.

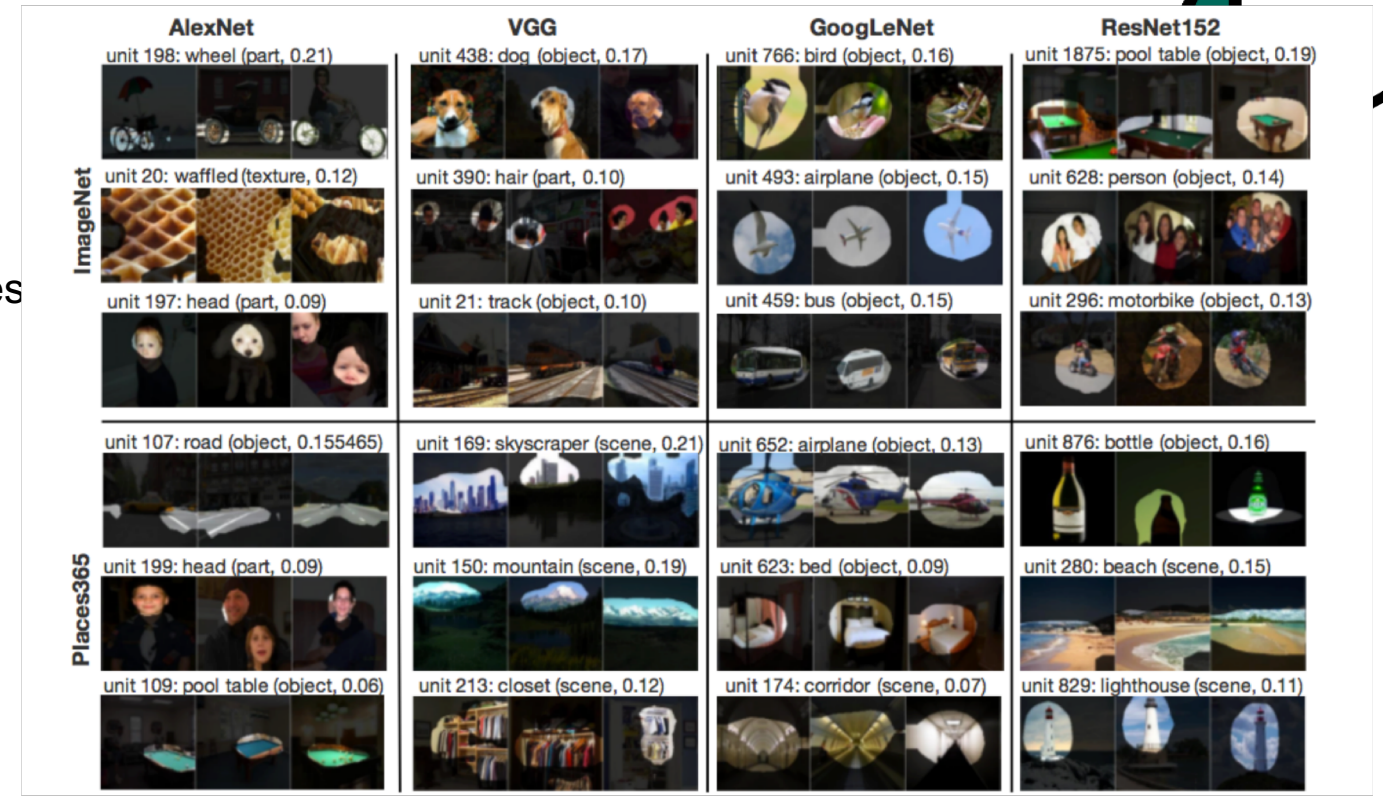




[Wang 2015]

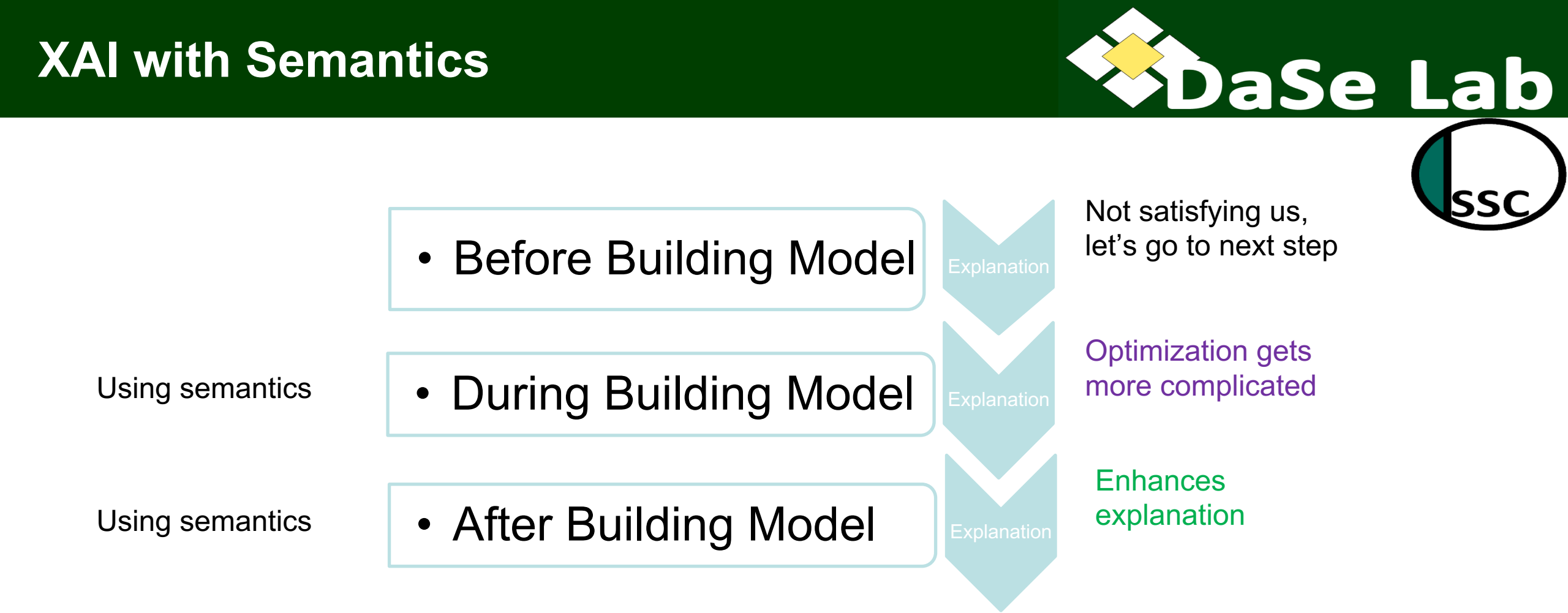
Using Semantics

- [Zhou 2018] is showing hidden layers objects and corresponding labels.
- Give intuition of what is going on inside, but it requires large semantic annotations of the objects.
- Currently one object is not related with other objects.
- **Need collaboration between semantic web community and deep learning community.**

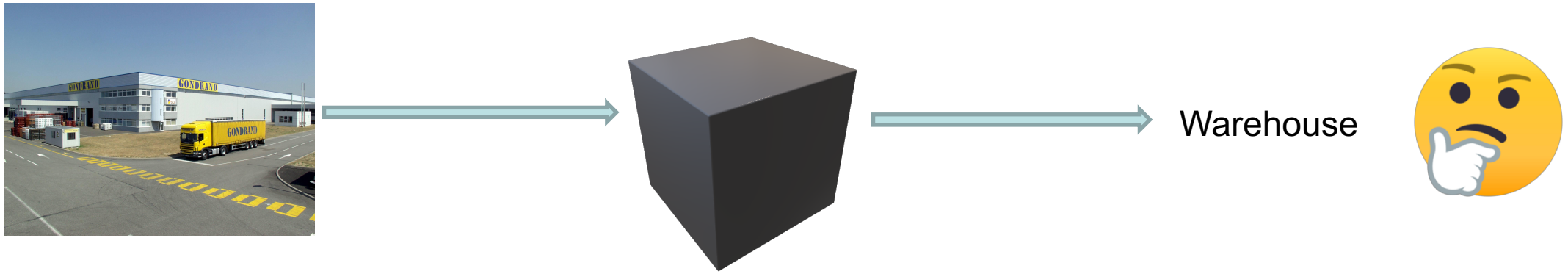


[Zhou 2018]

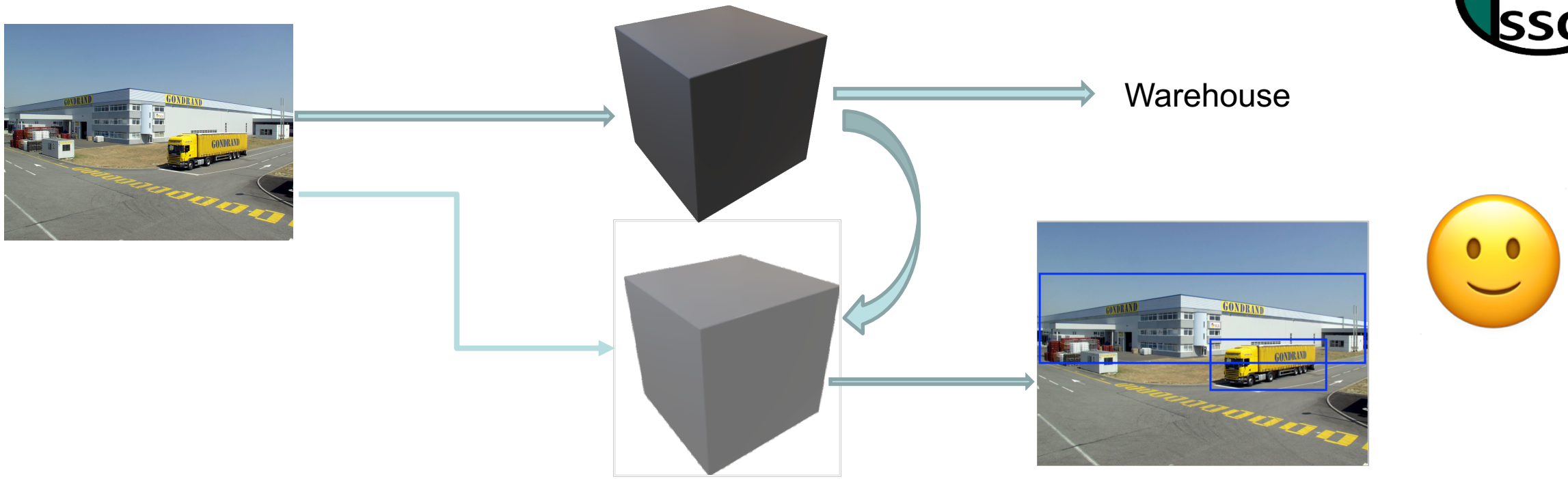
XAI with Semantics



Original AI Model



AI Model with some explanation



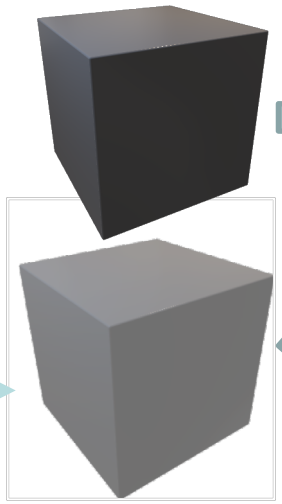
AI Model with some explanation

AI Model using semantics

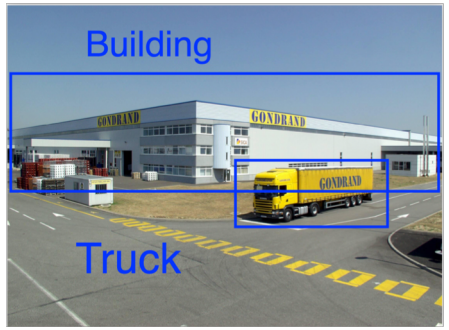
DaSe Lab

SSC

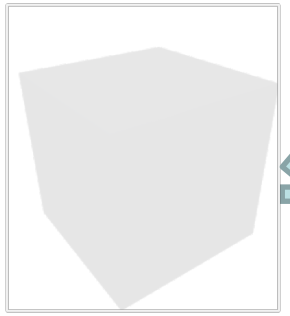
Enhanced explanation using semantic web technology



Warehoue



External Knowledge Base



Symbolic Learning System

This image contains building, truck, door, window which are usually found in warehouse. So it seems this picture is a Warehouse picture.[1]

[1]Derek, 2018

References

- [David Gunning 2017] Mr. David Gunning, 2017, https://www.darpa.mil/program/explainable-artificial-intelligence
- [Lipton 2016] Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.
- [Rudin 2018] Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv:1811.10154 (2018)
- [Ilaria 2015] Tiddi, Ilaria; d'Aquin, Mathieu and Motta, Enrico (2015). "An Ontology Design Pattern to Define Explanations". In: Proceedings of the 8th International Conference on Knowledge Capture, ACM, article no. 3.
- [Riccardo 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). "A survey of methods for explaining black box models". ACM Computing Surveys (CSUR), 51(5), 93.
- [Ribeiro 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?: Explaining the predictions of any classifier". KDD.
- [Ramprasaath 2016] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", https://arxiv.org/abs/1610.02391
- [Zhou 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, "Learning Deep Features for Discriminative Localization", https://arxiv.org/abs/1512.04150
- [Shrikumar 2017] Avanti Shrikumar, Peyton Greenside, Anshul Kundaje, "Learning Important Features Through Propagating Activation Differences", https://arxiv.org/abs/1704.02685

References

- [Hendricks 16] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell, "Generating Visual Explanations", https://arxiv.org/abs/1603.08507
- [Chakraborty 2018] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay "Adversarial Attacks and Defences: A Survey", https://arxiv.org/abs/1810.00069
- [Thrun, 1996] Sebastian Thrun, "Extracting Rules from Artificial Neural Networks with Distributed Representations", NIPS 1996
- [Hailesilassie, 2016] Tameru Hailesilassie, "Rule Extraction Algorithm for Deep Neural Networks: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No. 7, July 2016
- [Craven, 1996] Mark W. Craven, "Extracting Comprehensible Models from Trained Neural Networks". PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison, 1996
- [Doshi-Velez, 2017] Finale Doshi-Velez, Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning", https://arxiv.org/abs/1702.08608
- [Springenberg, 2015] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, "Striving for Simplicity: The All Convolutional Net" https://arxiv.org/abs/1412.6806
- [Zhou 2018] Bolei Zhou, David Bau, Aude Oliva, Antonio Torralba, "Interpreting Deep Visual Representations via Network Dissection" https://arxiv.org/abs/1711.05611
- [Sushil 2018] Madhumita Sushil, Simon Suster and Walter Daelemans, "Rule induction for global explanation of trained models", https://arxiv.org/abs/1808.09744

- [Wang 2015] Hao Wang, "Semantic Deep Learning", https://pdfs.semanticscholar.org/988d/1295ec32ce41d06e7cf928f14a3ee079a11e.pdf
- [Previti 2009] Christopher Previti, Oscar Harari ,Igor Zwir and Coral del Val, "Profile analysis and prediction of tissue-specific CpG islandmethylation classes", https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-116
- [Doran 2018] Derek Doran, "Okay but Really... What is Explainable AI? Notions and Conceptualizations of the Field", http://ontologforum.org/index.php/ConferenceCall_2018_11_28

Questions

- Other ideas how explanation can be enhanced?
- Next steps to go from here?

Thank you