Interpretable machine learning

by

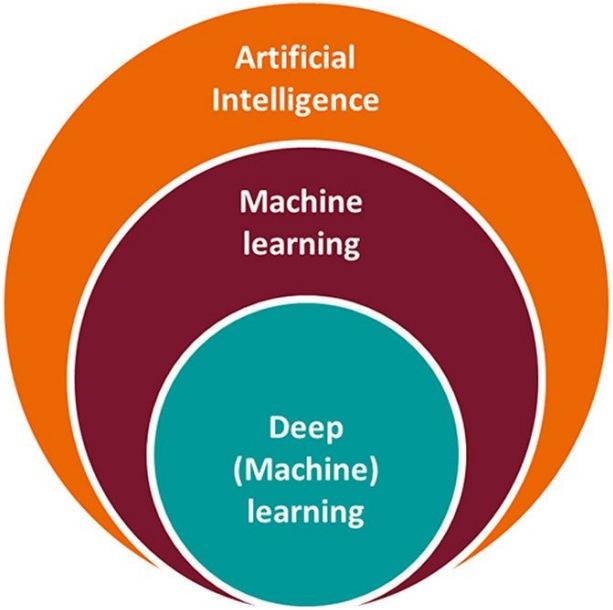
Juri Zach

30.10.2018

Table of content

- Introduction to machine learning
  - Different kind of machine learning
  - Work-flow of a machine learning application
  - Machine learning methods
  - Complexity of machine learning algorithms
- Interpretation of machine learning models
  - What is interpretability?
  - Why is interpretability important?
  - Methods to interpret machine learning models
- Conclusion

Introduction to machine learning



"Machine Learning is concerned with
computer programs that automatically
improve their performance through
Experience"

(Herbert Simon)

www.datasciencecentral.com/profiles/blogs/the-artificial-neural-networks-handbook-part-1

Different kind of machine learning

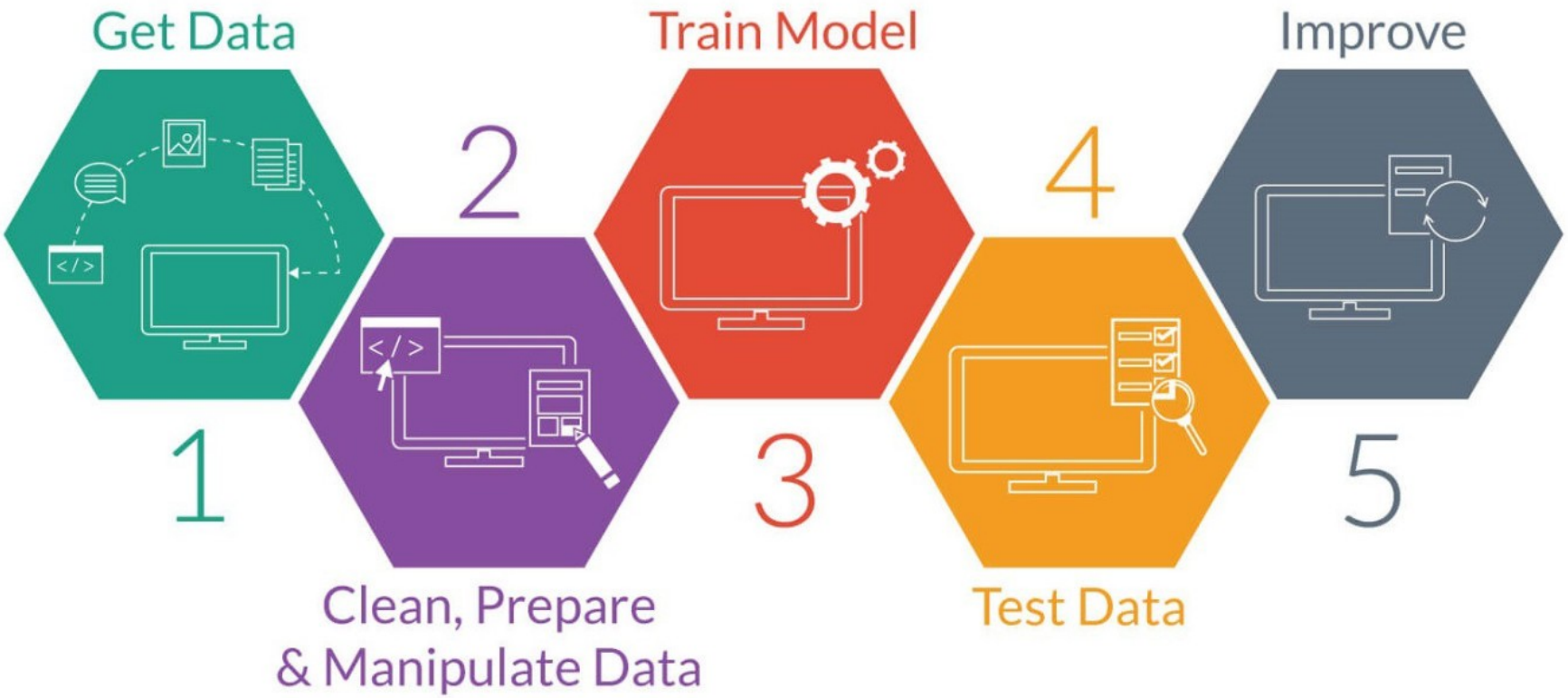
Supervised Learning

Unsupervised Learning

Reinforcement Learning

www.kickstarter.com/projects/1311831077/learn-real-world-machine-learning-by-building-proj

Machine learning work-flow

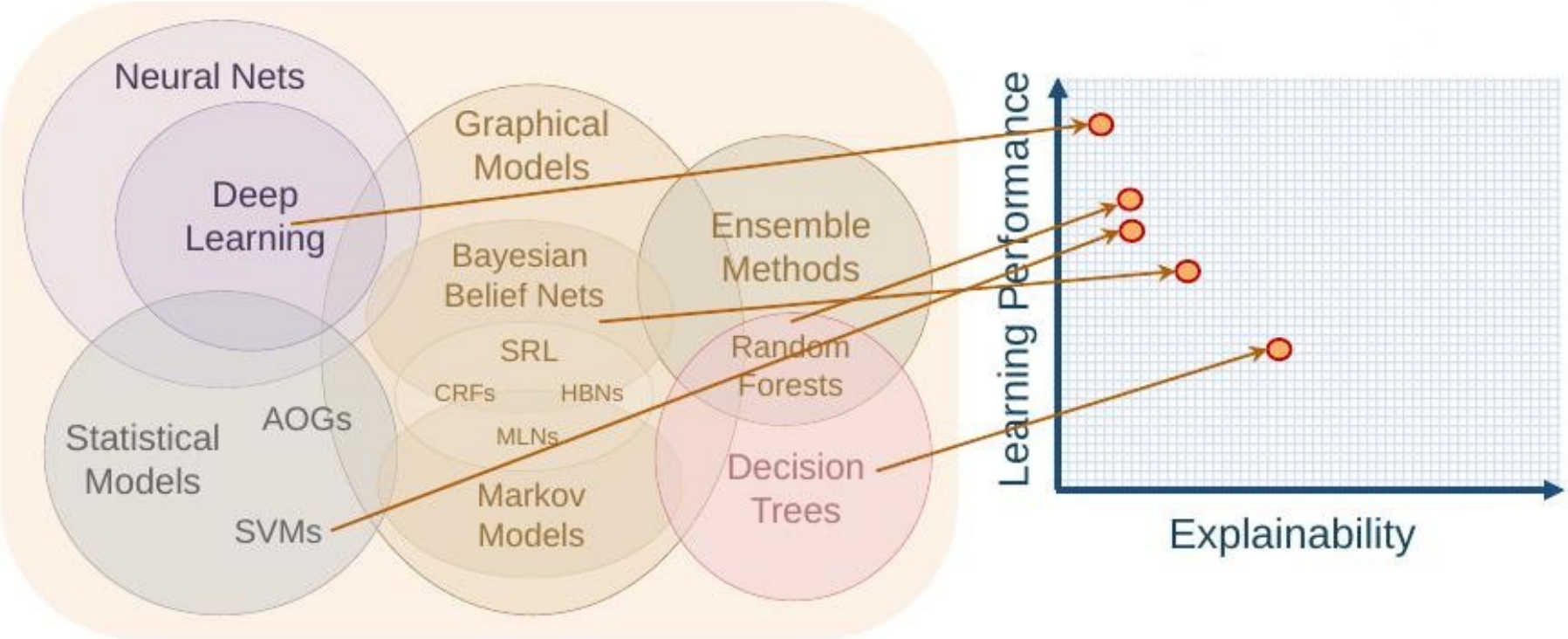


www.machinelearning-blog.com/2017/11/19/fsgdhfju/

Machine learning methods

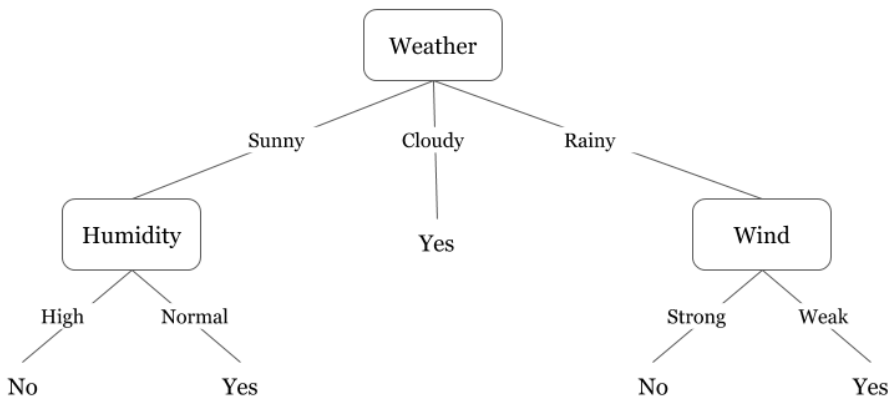
- Decision trees
- Linear models
- Logistic regression
- Bayesian nets
- Random forest
- Deep neural networks

Complexity of machine learning models

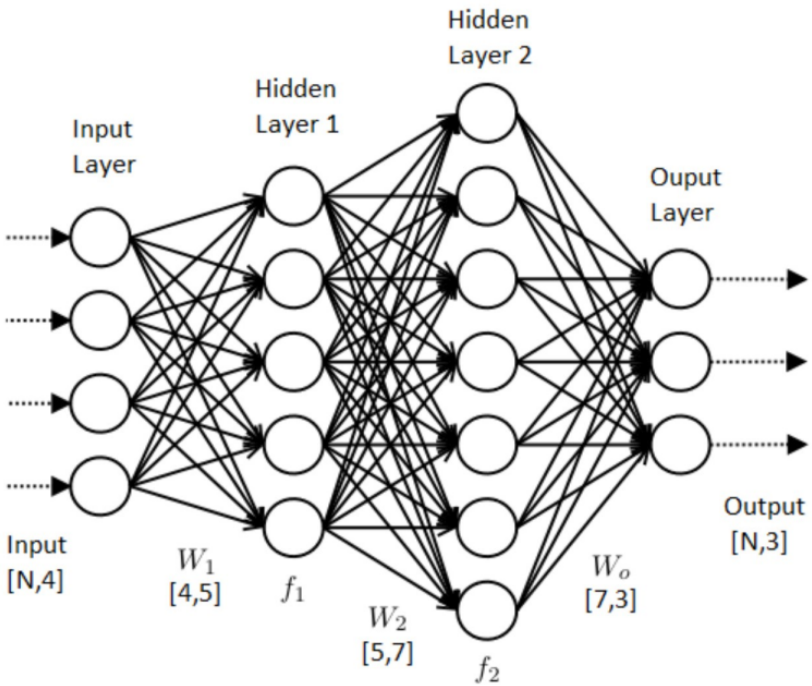


David Gunning (2017) Explainable Artificial Intelligence (XAI)

Decision tree vs neuronal net



Should I play Badminton?



www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/
www.datasciencecentral.com/profiles/blogs/the-artificial-neural-networks-handbook-part-1

Interpretation of machine learning models

What is interpretability?

"The term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character"

(Zachary C. Lipton)

Lipton (2017) The Mythos of Model Interpretability

What is interpretability?

"we define interpretability as the ability to explain or present in understandable terms to humans"

(F. Doshi-Velez and B. Kim)

F. Doshi-Velez and B. Kim, (2017) Towards A Rigorous Science of Interpretable Machine Learning

Why is interpretability important?

- Debugging
- Safety
- Subconscious biases
- Gaining trust
- Scientific understanding
- General Data Protection Regulation

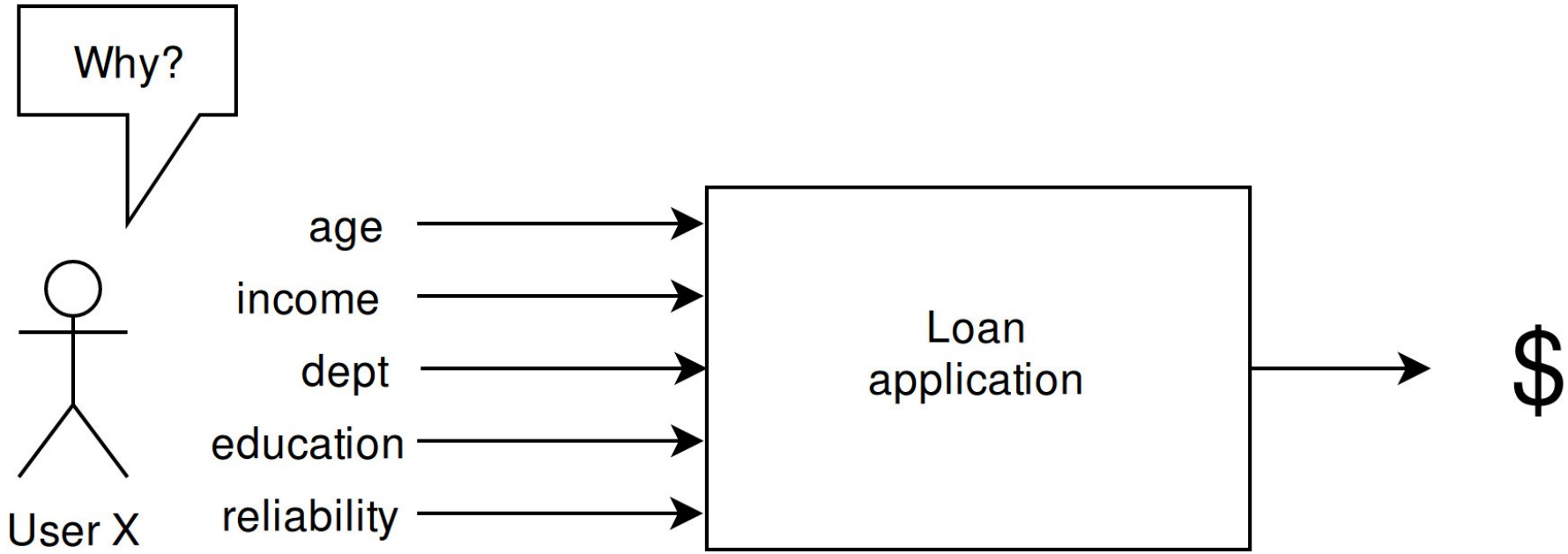
Different methods to interpret machine learning

- Designed interpretable models vs post-hoc interpretation
- Model level interpretation vs prediction (instance) level interpretation
- Model agnostic vs model specific method

Counterfactual explanations

- Properties
  - Post hoc interpretation
  - Prediction level interpretation
  - Model agnostic method
- Causal explanation to describe event (prediction) by its cause (features)
- Describes small change to the feature values that change prediction of predefined output

Counterfactual explanations example



Transparency by Design (TbD-net)

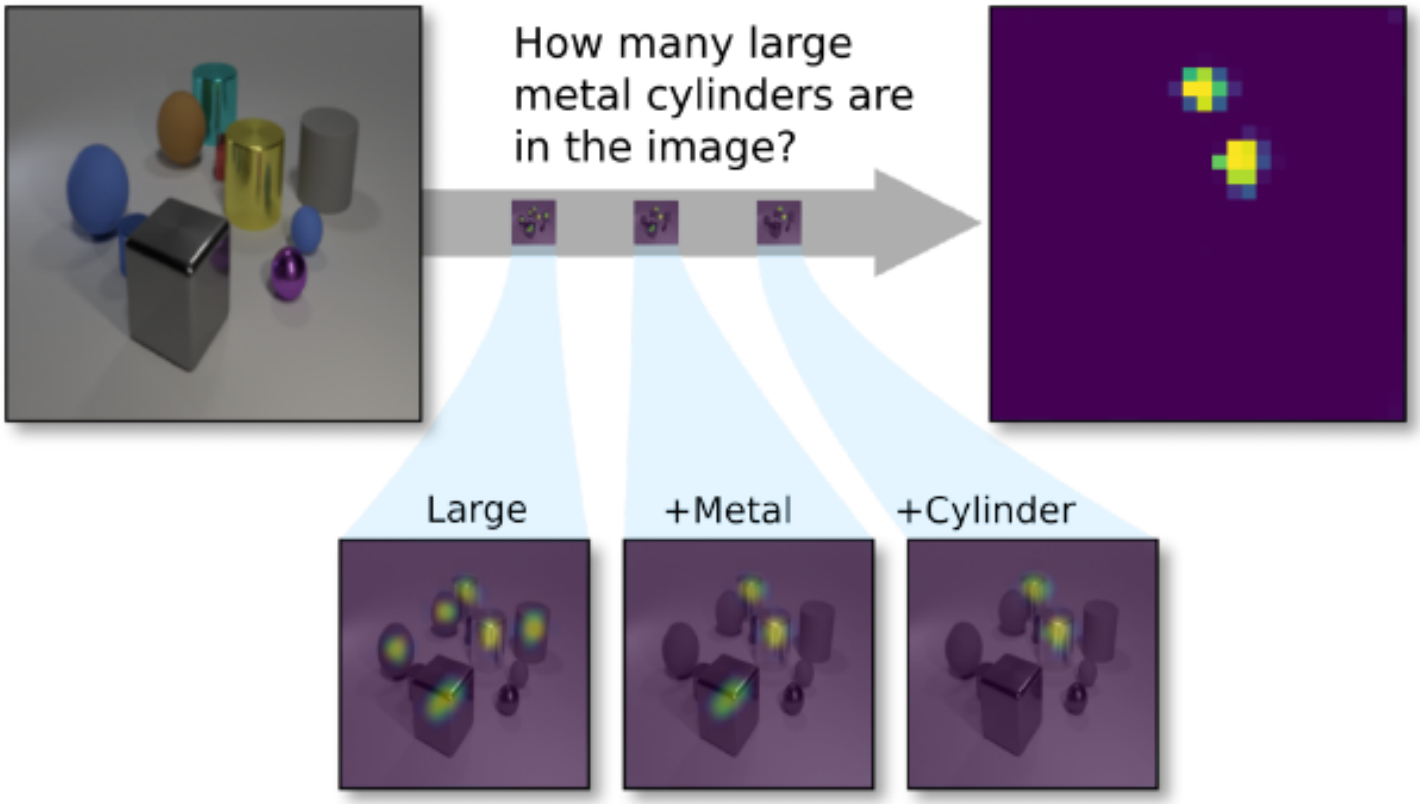
- Properties
  - Interpretable model design
  - Model interpretation & prediction interpretation
  - Model specific method

- Designed for visual question answering (VQA)

TbD-net Architecture

- Natural language component to parse question into series of logical operations
- Module network to perform operations in image
  - Module is a small neural network performing a given logical step
  - Complex chain of reasoning is broken down in smaller problems which can be solved independently

TbD-net example



D. Mascharaka, P. Tran i.a. (2018) Transparency by Design: Closing the Gap Between Performance andInterpretability in Visual Reasoning

Conclusion

Technical conclusion

- Extremely fast evolving scientific field
- Undefined patchwork science
- Research only, no practical application yet

Social conclusion

- Important to adapt model to application by evaluating many objectives like ethic, legality, safety et al.
- Great to extract complex knowledge from huge amounts of data
- Potential in improving human computer interaction for AI to target a future where humans and computers work together to solve problems

Reference

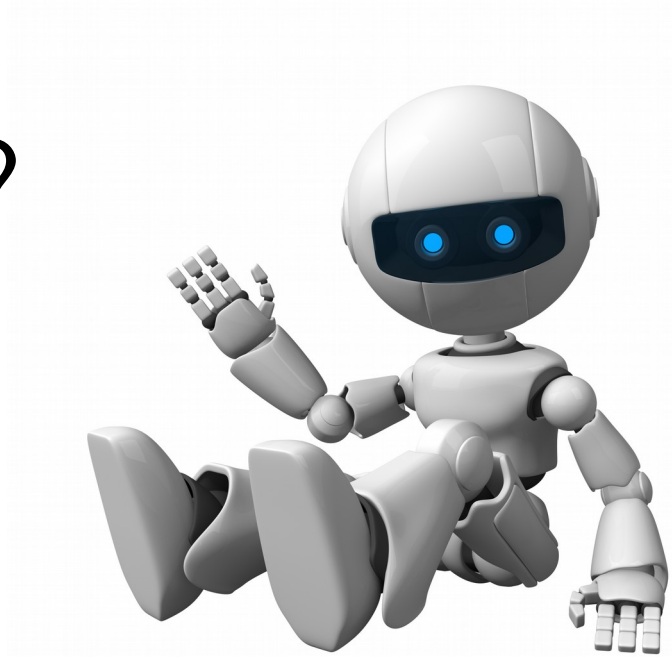
- David Gunning (2017) Explainable Artificial Intelligence (XAI)
- Lipton (2017) The Mythos of Model Interpretability
- F. Doshi-Velez and B. Kim, (2017) Towards A Rigorous Science of Interpretable Machine Learning
- M. Du, N. Liu, X. Hu (2018) Techniques for Interpretable Machine Learning
- D. Mascharaka, P. Tran i.a. (2018) Transparency by Design: Closing the Gap Between Performance andInterpretability in Visual Reasoning

Images Reference

- www.datasciencecentral.com/profiles/blogs/the-artificial-neural-networks-handbook-part-1
- www.kickstarter.com/projects/1311831077/learn-real-world-machine-learning-by-building-proj
- www.machinelearning-blog.com/2017/11/19/fsgdhfju/
- www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/
- www.datasciencecentral.com/profiles/blogs/the-artificial-neural-networks-handbook-part-1

Thanks for your attention!

Any Questions?



www.vaslui.robo-stiri.ro/2018/10/07/noutati-in-cea-mai-recenta-versiune-robo-stiri-vaslui/