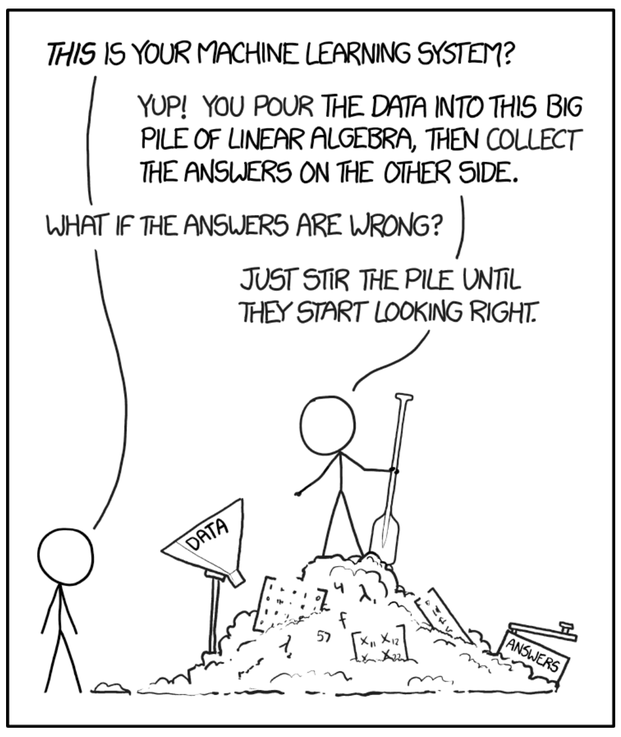
Interpretable Machine Learning

Accessing the Potential of Machine Learning in Data Engineering through a better Understanding

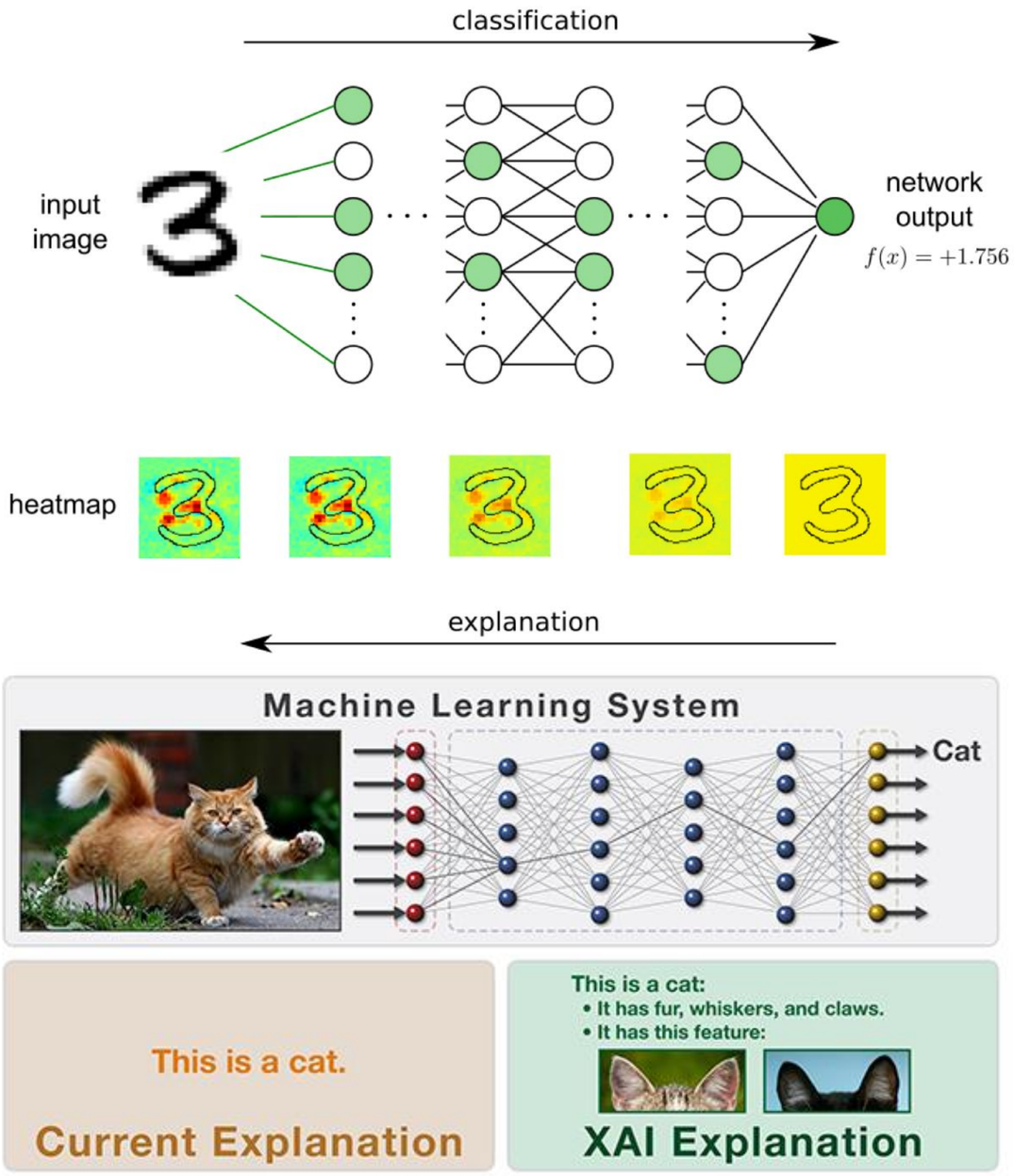
In the broader context of Data Engineering, Machine Learning Systems are often used to create value by enabling humans to take data driven decision. As we saw in some of the talks of this ring lecture, apart from metrics such as predictive performance or runtime, the notion of interpretability of the output from a ML systems for a given input is important as well. This was emphasized by the example of enersis' coal mining talk, where the customer requested to use an approach that sacrificed predictive performance for a higher degree of interpretability for its results. This pattern also applies to many other use cases, where end users or legal regulation prohibit black box behavior.

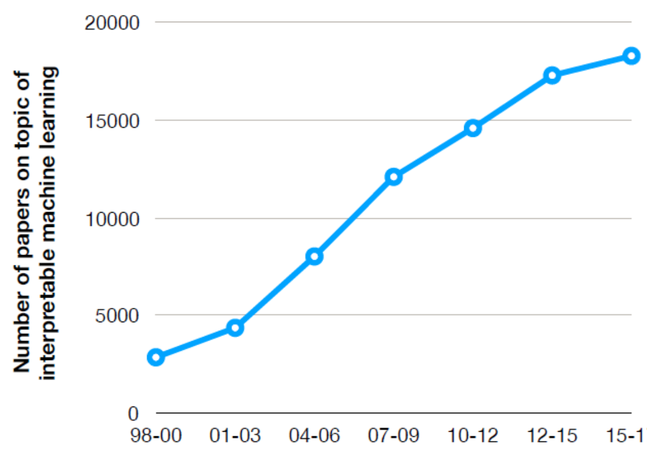
Recent advances in areas such as deep neuronal networks, have led to an increasing amount (see Graph[1]) of research to propose new ways for a better reasoning how a model infers a results for a given input. In the context of Artificial Intelligence Systems, similar approaches are titled under the Term Explainable Artificial Intelligence (XAI [4]). This poster shows some examples of possible ways how this problem might be overcome.

For many ML techniques it is not only problematic to interpret their results, but also to compare their results with each other and to reproduce them. To solve this problem, Dr. Ing. Sebastian Schelter proposed a solution of storing the meta data, like the parameters and feature transforms of model as well as the commit for the code base to solve these problems.



This comic illustrates light heartedly possible shortcomings in Machine Learning Systems[2]





Graph [1] this plot shows the increase in publications for interpretable ML in recent years

The images above demonstrate two approaches how the interpretability of the output from a neural network classifier could be improved. On the top [3] a heatmap is generated, that highlights which regions of the input have the strongest impact on the classification. The image on the bottom [4] shows a baseline output in textual form as well as an improved output that states the classified label plus additional information about the features of the input in textual and visual form to increase the understanding of the user about the reasoning of the neural network.

This poster was created for the ring lecture on "Data Engineering in practice" in the Winter Term 2017/18 by

Jan Selke
Master student IT-Systems Engineering
Hasso Plattner Institute, Potsdam, Germany

E-Mail: jan.selke@student.hpi.de

References:

[1] https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf slide 7
[2] https://xkcd.com/1838/
[3] https://www.hhi.fraunhofer.de/en/departments/vca/research-groups/machine-learning/research-topics/interpretable-machine-learning.html
[4] https://www.darpa.mil/program/explainable-artificial-intelligence

Hasso Plattner Institut