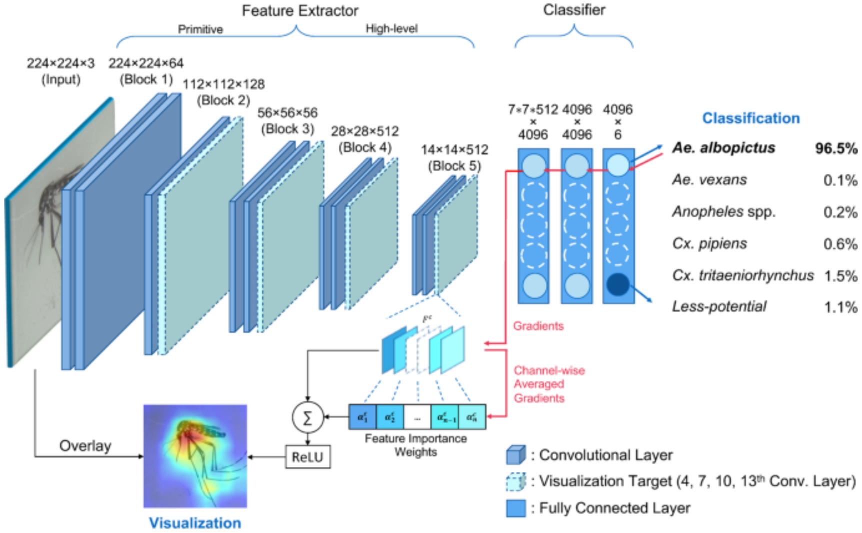
X-AI

Explainable Artificial Intelligence

Florence d'Alché-Buc

AI today



Park et al. Nature, 2020.

- significant advances in Statistical Machine Learning vs Symbolic Machine Learning
- spectacular results of Deep Neural Networks
- data-driven AI embedded in decision-making processes

Explainability in AI

« to describe the purpose, rationale and decision-making process of the AI tool in a way that can be understood by the average person »

Data scientist

Expert of the field (finance for instance)

User/customer

Regulator / lawyer /…

The lack of explainability in data-driven AI

1 Linked to the nature of statistical machine learning algorithms

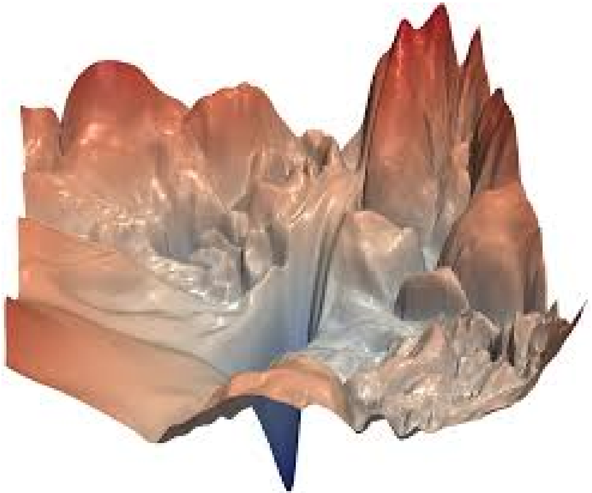
Learning is a complex optimization process that takes a training dataset and produces a predictive model

Bias in your data ?

Local Minima ?

Model selection ?

(© Matthieu Ferrand)



Visualization of a loss function,
Li et al. NeurIPS 2018.

TELECOM Paris
IP PARIS

The lack of explainability in data-driven AI

2 Linked to the objectives of machine learning algorithms

A learning algorithm attempts to define a predictive model by searching for input patterns correlated with the output variable based on a strong assumption about data: the i.i.d. assumption

Animal in the snow ?

Wolf ?

(© Matthieu Ferrand)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

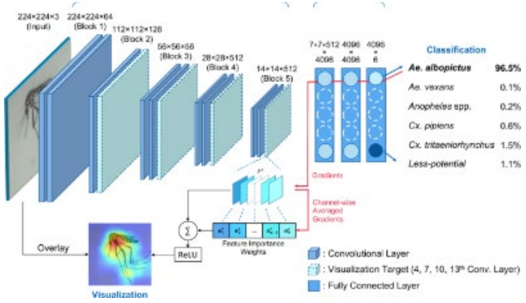
Guestrin et al. 2016

The lack of explainability in data-driven AI

3 Due to the nature of the predictive models

- *Some models are more explainable than others*: sparse linear models, decision trees, probabilistic graphical models, random forests, …

- deep neural networks exhibit a very high level of complexity (millions of parameters)

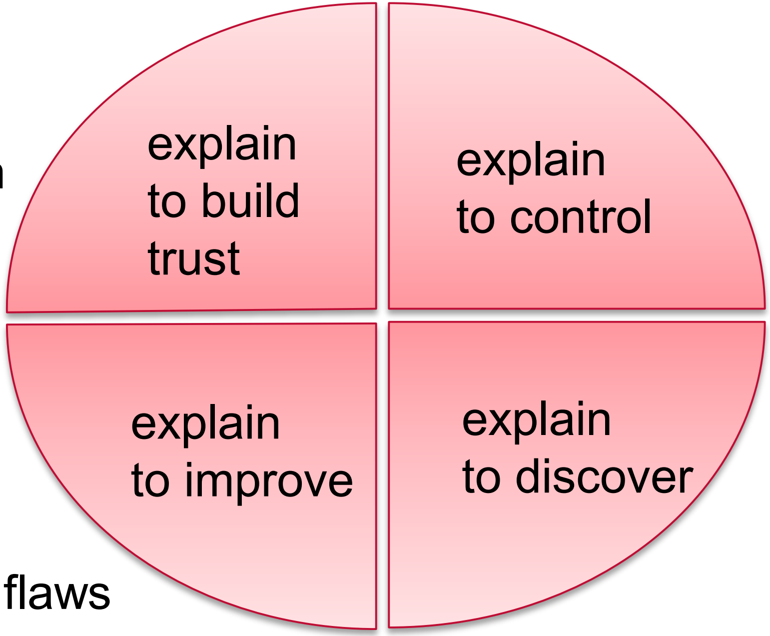


Pb : performance is often associated to very complex models & ability to tackle massive training datasets

The need for Explainability

Human- readable justification of a decision

Compliance to legislation
“Right to explanation”



Identification of systems flaws

Information extraction

XAI, a compound of trustworthy AI

Explainability in data-driven AI

Focus on local explainability: provide an “explanation” of the predictive model’s decision

What is an « explanation » ? For whom ? a data scientist, an expert of the field a user, the regulator ?

Main factors that led to that prediction

High level concepts that are activated when the prediction is given

Counterfactual reasoning: if I change this feature value, does the prediction change ?

Explanations also depend on the nature of data

Multivariate hand-defined features

Image / Audio / Video

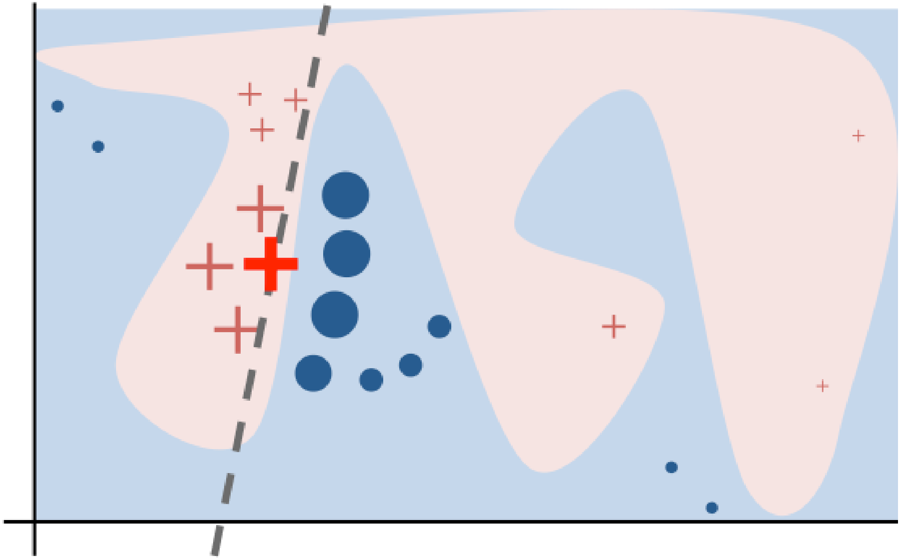
Natural Language processing

Post-hoc Approaches: local linear proxy

- **LIME (Ribeiro et al. 2016)**

- Model-agnostic approach that builds a sparse linear **proxy model** to get insights on a local decision once the whole model is learned.

- *Perturbation-based approach*



Post-hoc Approaches: saliency maps

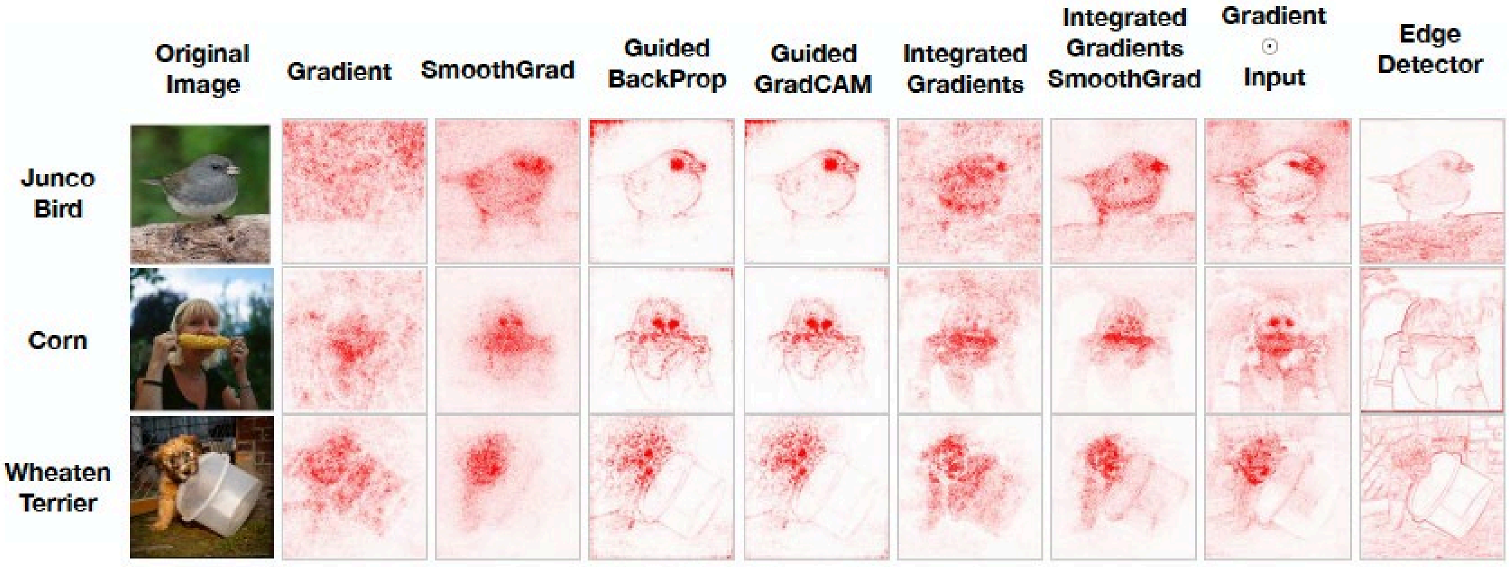
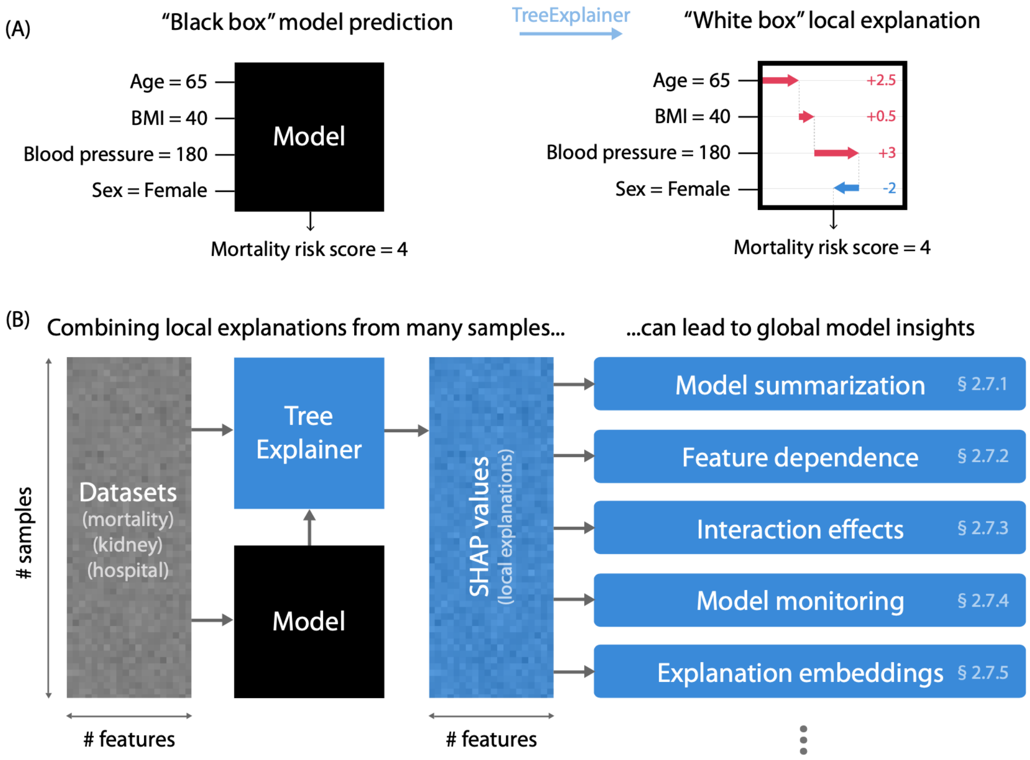


Fig. *Adebayo et al. NeurIPS 2018.*

f is the predictive model, x is the input: $\dfrac{\partial f}{\partial x}$

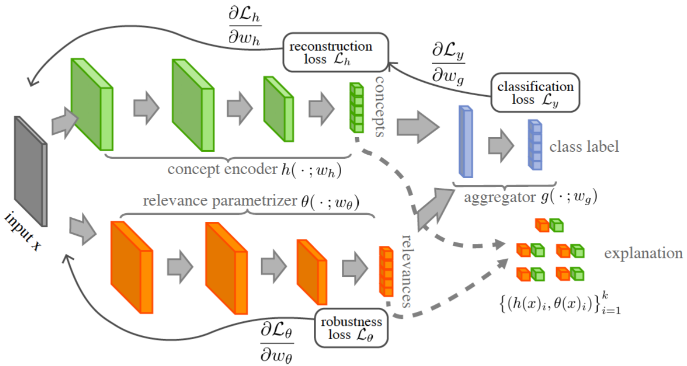
Refs: Werbos 1982, Pridy et al. 1993, Steppe & Bauer 1997, Simonyan et al. 2013, Springenberg et al. 2014, Smilkov et al. 2017, Selvajaru et al. 2017…

Post-hoc approaches: Tree explainer
(Lundberg et al. 2019)



Explainability By Design

- Identify a (specific) neural network to a set of logical rules (hybrid networks)
- Modify the architecture of a network to make it interpretable (Self-explainable Networks, Alvarez-Melis & Jaakola 2018)



SENN

Explainability by design

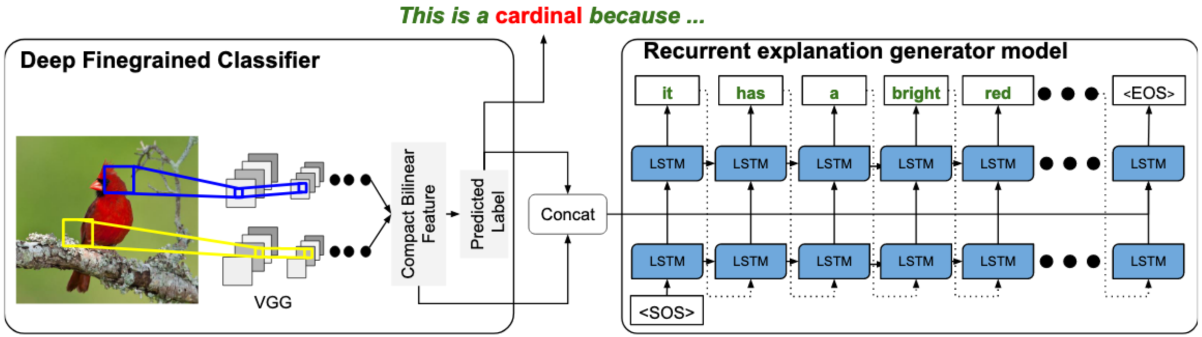
- Impose some properties that an interpretable neural network should satisfy, (d'Alché-Buc et al. 1994, Alvarez-Melis et al. 2018, Plumb et al. 2019)
  - (logical) consistency: non contradictory rules
  - Completeness
  - Fidelity of the « explanations » to the model's output
  - Sparsity of high level concepts
  - Stability of explanations

- Learn jointly two models: one for prediction, the other for explanation

(Hendricks et al. 2016, Dong et al. 2017, Parekh et al. 2020)

Explainability by design

- **Generating visual explanations: Hendrycks et al. 2016**



XAI in its infancy

- Tools on the shelves: mainly post-hoc approaches – good for existing black box models currently in production, but with some flaws: provide an explanation but may be not the one « used » by the model

- Formal work on interpretations/explanations in ML, re-think machine learning/AI at the lense of explainability for **a next generation AI tools**

- « **Can biologist fix a radio ?** » (Lazebnik, 2002) the celebrated paper in quantiative biology in 2000's applies somehow here. Can a statistician provides an explanation ?

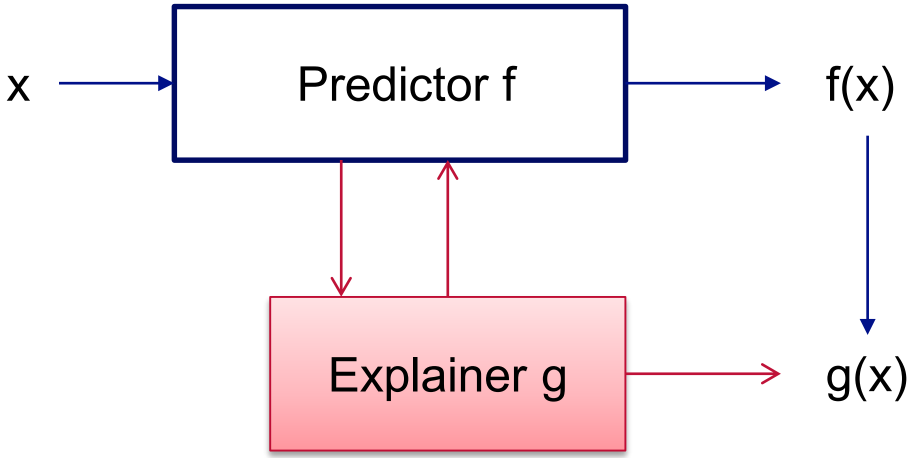
  - Explanations are currently more interpretations than explanations: what link with reasoning ? What link with logics ? What link with knowledge ?

  - Making a predictive model explainable belongs more to symbolic AI and calls for automated reasoning, knowledge representation etc… a lot to borrow from years of AI.

- Other ways of thinking: counterfactual reasoning, intervention, Bayesian approaches, probabilistic programming, knowledge graph and automated reasoning

FLINT: a framework for learning intepretable network

Usage 1:
Joint learning of f and g,
Mutual benefits,
g can even the final predictor



Usage 2:
Post-hoc/reverse engineering of a pre-defined network f

Parekh et al. 2020.

References

- Beaudouin et al., Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach Arxiv, 2020.
- A recent review: J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas and J. Wilson, "The What-If Tool: Interactive Probing of Machine Learning Models," in IEEE Tr. on Visualization and Computer Graphics, vol. 26, no. 1, pp. 56-65, Jan. 2020.
- Christoph Molnar’s online book on interpretable machine learning (link)
- Adebayo et al. Sanity Checks for Saliency Maps, NeurIPS 2018
- Alvarez-Melis & Jaakola, Self-Explaining Neural networks, NeurIPS 2018 (link)
- Hendricks, Akata, …, Darrell, Generating visual explanations, ECCV 2016 (link)
- Plumb, Al-Shedivat, Xing, Talwalkar:Regularizing Black-box Models for Improved Interpretability. CoRR abs/1902.06787 (2019)
- Ribeiro, Singh, Guestrin, Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD 2016.
- Platform, common tools:
  - What If tools (Google), Captum (Pytorch/Facebook), 360xAI (IBM) https://aix360.mybluemix.net/, iml R package

TELECOM Paris

IP PARIS