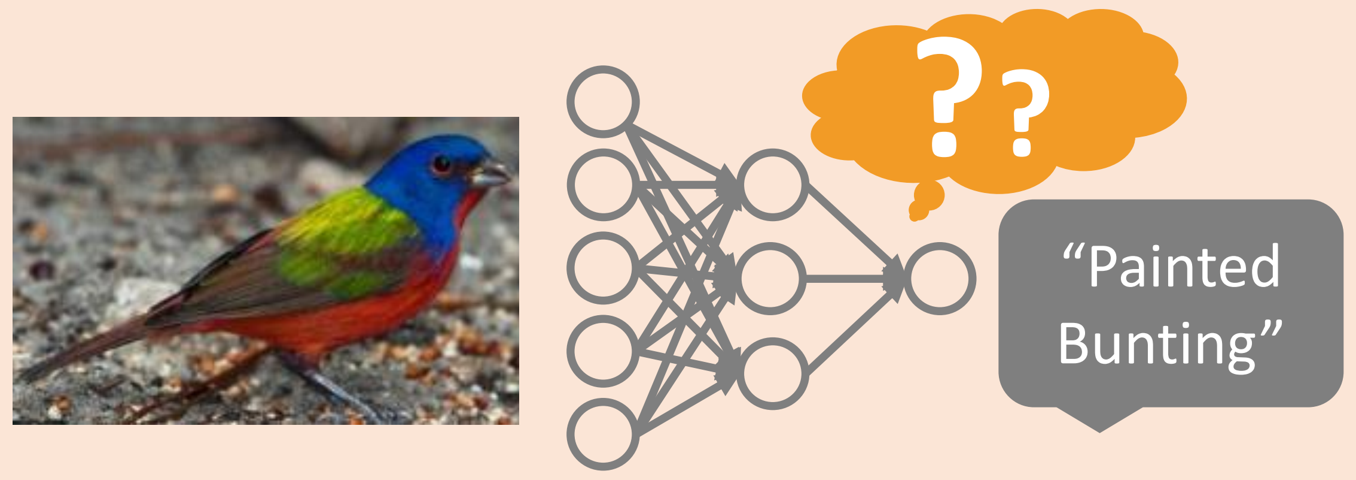
EXPLAINABLE AND INTERPRETABLE MACHINE LEARNING

MEIKE NAUTA

University of Twente, Enschede, The Netherlands & University of Duisburg-Essen, Germany

PhD Candidate, DMB group
✉ m.nauta@utwente.nl

FROM EXPLAINABLE AI TO INTERPRETABLE MACHINE LEARNING



What does a machine learning model actually learn?
Why is a classifier predicting the wrong class?
What is the reasoning behind a prediction?
Did the model learn the *intended* behaviour?

Daily life is increasingly governed by decisions made by algorithms due to the growing availability of big data sets. Most machine learning algorithms, especially deep neural networks, are black-box models. They cannot give insight into how they reach their outcomes. If we cannot understand the reasons for their decisions, how can we be sure that the decisions and the underlying reasoning are correct? What if the model is wrong or biased?

Explainable AI (XAI) aims to explain the decision-making process of machine learning models. Explanations can be heatmaps highlighting important regions in an image, visual prototypes, feature importance scores, text etc. Those explanations enable a user to check for correctness, fairness and potential biases. Explainable AI can also be useful for knowledge discovery to support scientific progress.

Most explainable methods are post-hoc: the generated explanations *approximate* an already trained model and are often incomplete. Instead, we should move towards the highly-demanded field of **intrinsically interpretable machine learning:** predictive models that are *explainable by design*, and truthfully show their reasoning in human-understandable terms.

Table 2. Our Co-12 explanation quality properties, grouped by their most prominent dimension: Content, Presentation or User.

From: Nauta et al. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI" (2022)

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Compositionality	Describes the format and organization of the explanation.
		Key idea: *How* something is explained.
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

EVALUATING EXPLAINABLE AI

What makes a good explanation?
How can we measure explanation quality?
How to quantify how good an explainable AI method is?

"From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI" (Nauta et al., 2023)

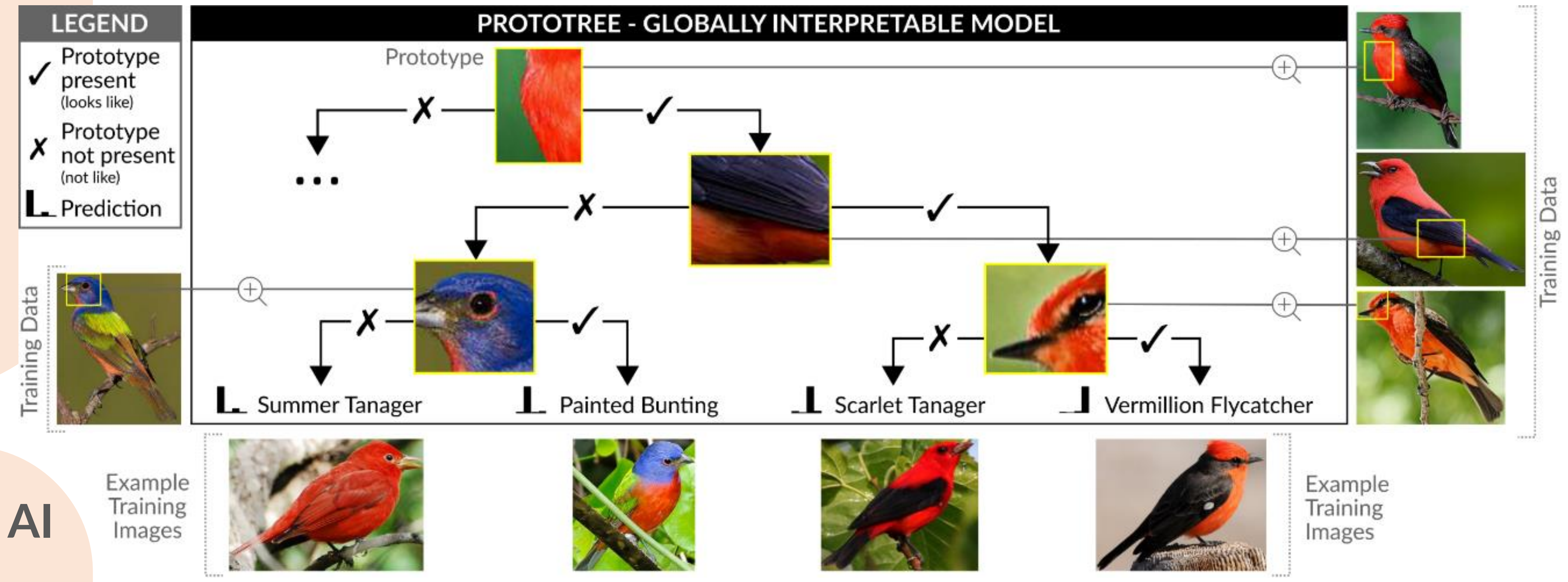
- Co-12: Twelve conceptual properties on explanation quality: from Correctness to Controllability.
- Systematic review of >300 papers on explainable AI. Main findings:
  - 1 in 5 papers evaluate with a user study
  - 1 in 3 papers evaluate only with anecdotal evidence
- Systematically identified 29 quantitative evaluation methods for explainable AI

Our collection of evaluation methods can be used by researchers and practitioners to quantitatively and objectively validate, benchmark and compare XAI methods.

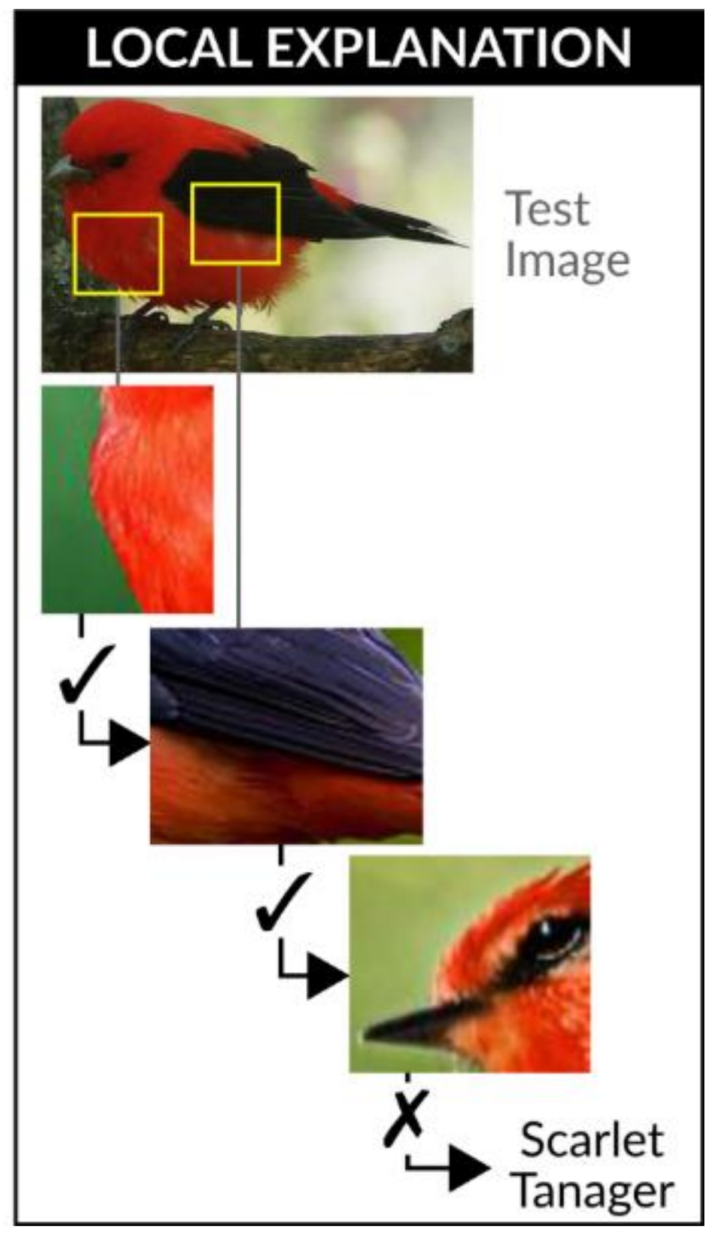
NEURAL PROTOTYPE TREE

ProtoTree (Nauta et al., CVPR 2021) is an **intrinsically interpretable method for fine-grained image recognition.** ProtoTree uses the expressiveness of deep learning to learn prototypes and incorporates interpretability by structuring these prototypes in a built-in decision tree. Leaves of the ProtoTree learn class distributions. Hence:

- ProtoTree **truthfully shows its full reasoning**
- A path through the tree concisely explains one prediction
- Back-tracking of misclassifications is easy
- Prototypes can reveal learned biases (*e.g.* leaf instead of bird)
- An ensemble with 5 ProtoTrees achieves similar accuracy as state-of-the-art black boxes



Example of a ProtoTree. ProtoTree is a **globally interpretable model** truthfully explaining its entire behaviour.



Transparent reasoning process for a **single prediction**.

PROTOTREE RESULTS & CONCLUSION

ProtoTree is highly accurate, while enabling a human to interpret the whole model. ProtoTree is also the first accurate model to allow retraceable decisions for single images in a **human-comprehensible number of steps**. An ensemble of 5 ProtoTrees approximates the accuracy of uninterpretable state-of-the-art. As a result, our novel work questions the existence of a so-called accuracy-interpretability trade-off.

APPLY PROTOTREE TO YOUR DATA:
https://github.com/M-Nauta/ProtoTree

TRY IT OUT

Nauta et al. "Neural prototype trees for interpretable fine-grained image recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.