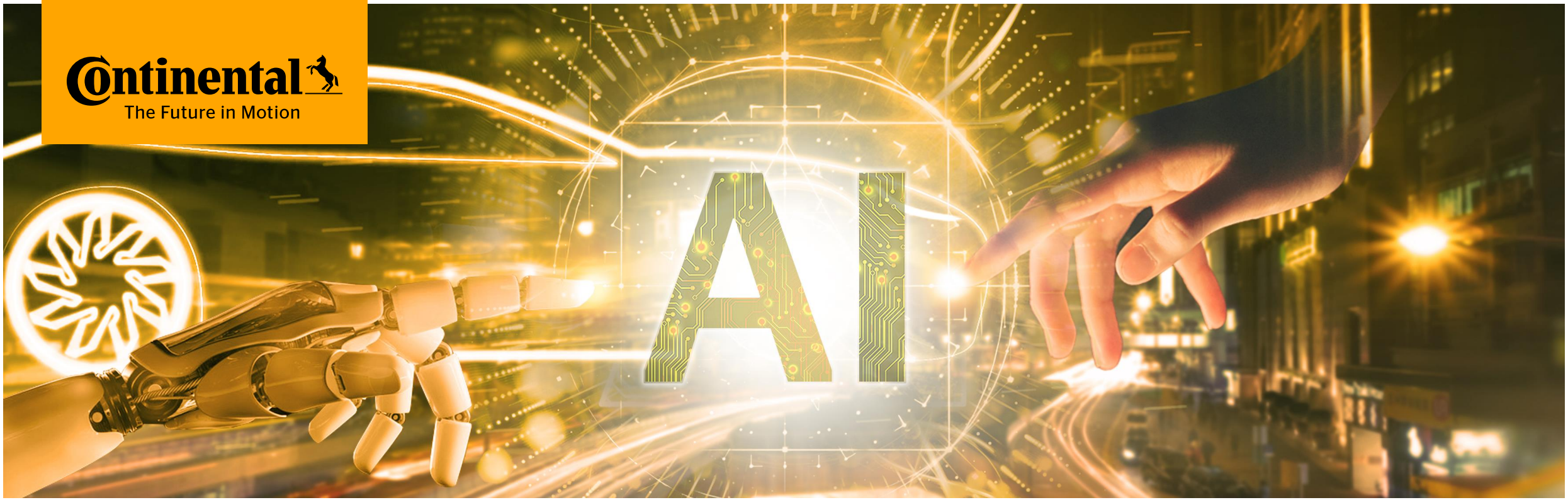


Diversity and Applications of Explainable Artificial Intelligence Methods

Gesina Schwalbe

www.continental.com

Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	Methods for Explaining ML Models	10

Agenda

1	**What is Explainability?**	**3**
2	Applications	5
3	A Method Taxonomy	8
4	Methods for Explaining ML Models	10

What Is Explainability?

Explainable decision system = There exists a

- **mechanism** providing an **explanation** (= explanator)
- to a **human** (= explainee)
- that allows them to **understand**
- one of (= explanandum)
  - the **model** resp. parts thereof,
  - evidence for a **model output**, or
  - the **context** of the system's reasoning.

XAI = lots of cognitive science!

Understanding = successful update of mental model; can be

- mechanistical = how it works, or
- functional = what is its purpose

Levels of **transparency** of a model (= mechanistic understanding):

- simulatable (= understandable as a whole)
- decomposable (into simulatable parts)
- algorithmically transparent (= mathematical understanding)

cf. (Schwalbe & Finzel 2022)

Agenda

1	What is Explainability?	3
2	**Applications**	**5**
3	A Method Taxonomy	8
4	Methods for Explaining ML Models	10

Use-cases

- **Compliance** with law and standards (e.g. GDPR); assessment e.g. wrt. safety, security, fairness, privacy

- Developer and expert users:
  - **Debugging** (robustness, trustworthiness)
  - **Knowledge retrieval**, transferability
  - **Assess compliance** with behavioral requirements

- Users:
  - (appropriate!) **Trust**, informed consent
  - Easily **getting familiar** with system
  - **Recourse** (e.g. in ranking systems)

Exemplary Application Fields

Wherever automated decisions influence human well-being!

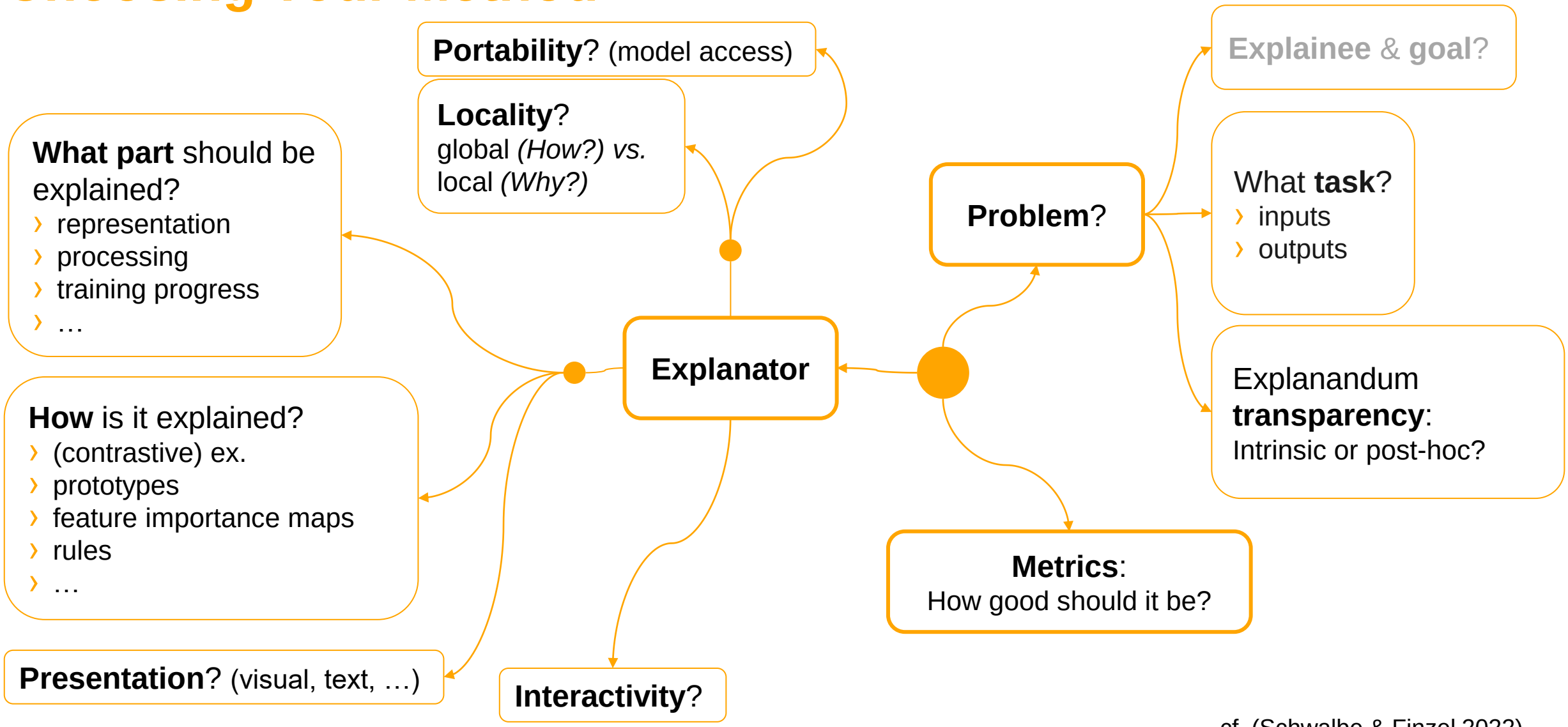
E.g.

- Automated driving
- Medical assistant systems
- Ranking systems (social, credits, …)
- Military decision systems
- Recommendation systems
- Data science
- HMI in Production
- …

Agenda

1	What is Explainability?	3
2	Applications	5
3	**A Method Taxonomy**	**8**
4	Methods for Explaining ML Models	10

Choosing Your Method



cf. (Schwalbe & Finzel 2022)

Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	**Methods for Explaining ML Models**	**10**
4.1	Inherently Interpretable and Blended Models	11
4.2	Feature Importance	14
4.3	Explaining Representations	19
4.4	Explainable Surrogates	23

Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	**Methods for Explaining ML Models**	**10**
4.1	**Inherently Interpretable and Blended Models**	**11**
4.2	Feature Importance	14
4.3	Explaining Representations	19
4.4	Explainable Surrogates	23

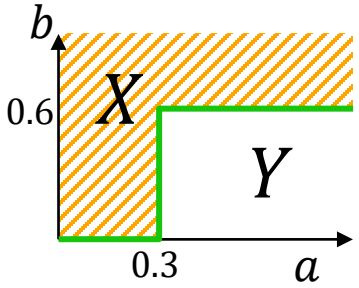
Inherently Interpretable Models

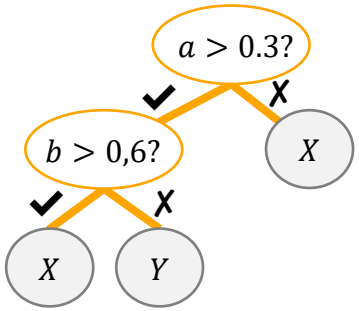
› Decision **Rules**

```
face(F) :- contains(F, A), isa(A, nose), contains(F, B), isa(B, mouth),
           top_of(A, B), contains(F, C), top_of(C, A).
```

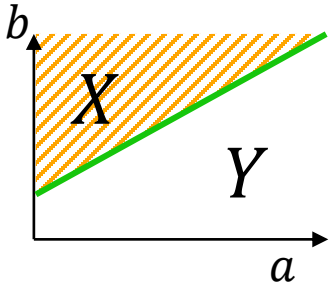
cf. (Rabold et al. 2020)

› Decision **Trees**





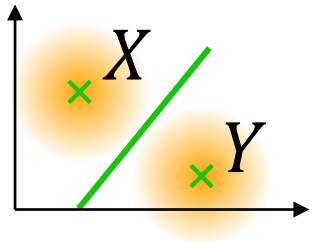
› **Linear** Models,
Generalized **Additive** Models



Linear: $f(x) = \alpha a + \beta b$

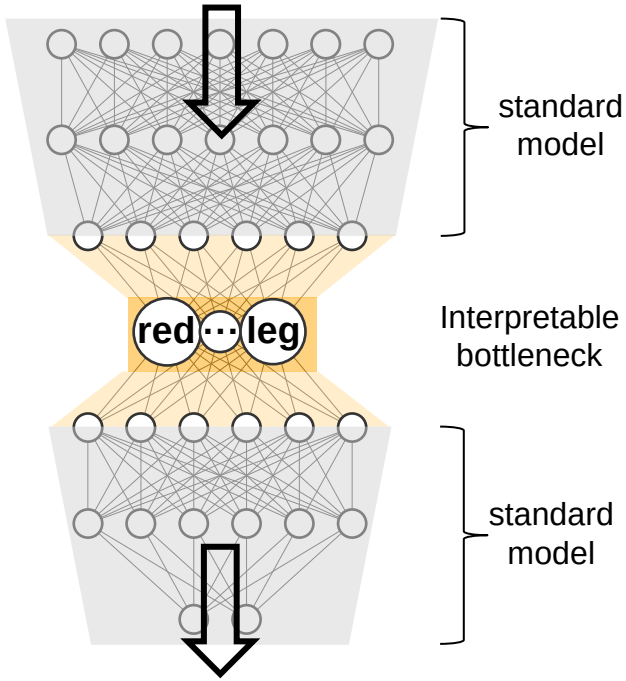
GAM: $f(x) = g^{-1}(f_a(a) + f_b(b))$

› **Clustering**

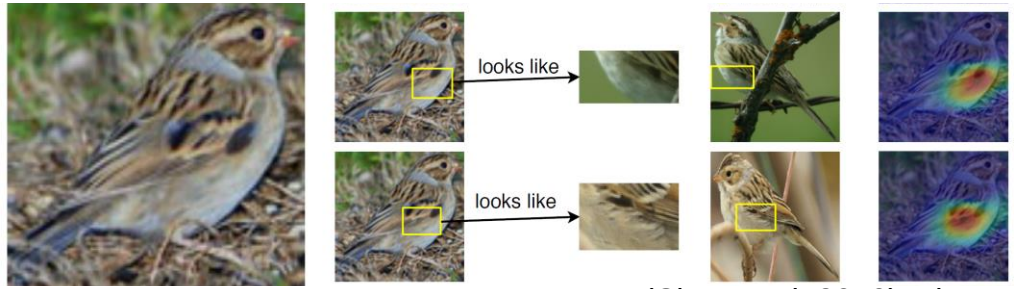


Blended and Self-explaining Models

Concept Bottlenecks (Losch et al. 2020), (Koh et al. 2020)
Disentangled Representations



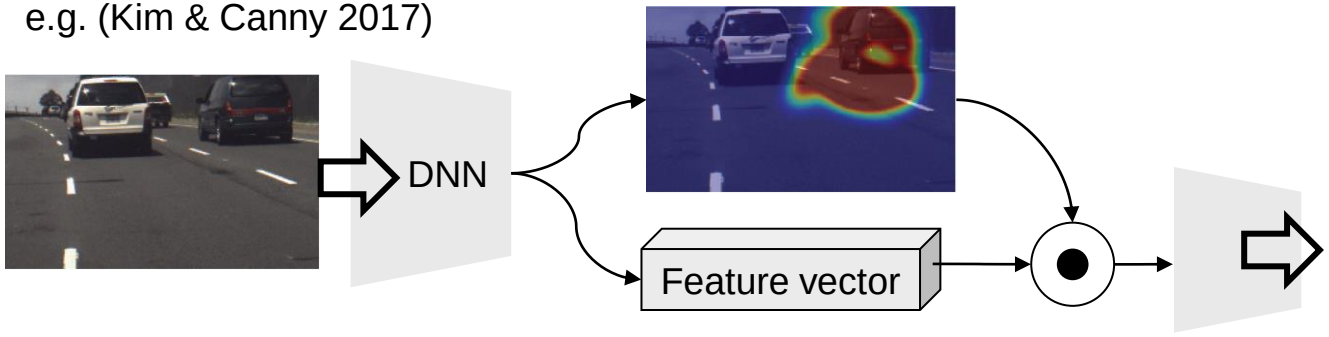
ProtoPNet (Chen et al. 2019)



(Chen et al. 2019), Fig. 1

Attention models
e.g. (Kim & Canny 2017)

(filtered) Attention



(Kim & Canny 2017, Fig. 5)

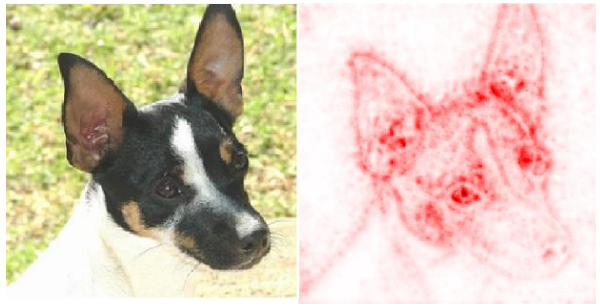
Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	**Methods for Explaining ML Models**	**10**
4.1	Inherently Interpretable and Blended Models	11
4.2	**Feature Importance**	**14**
4.3	Explaining Representations	19
4.4	Explainable Surrogates	23

Feature Saliency

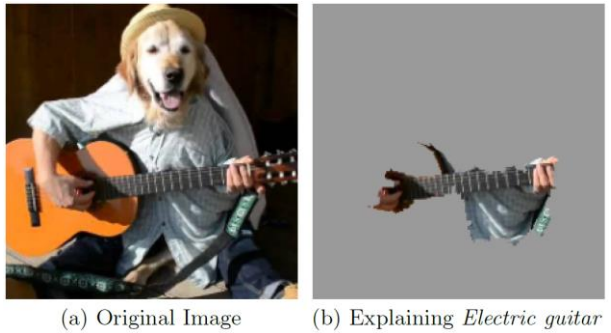
› What **input features** were (how) **important** to the **output**?

Works for features like …



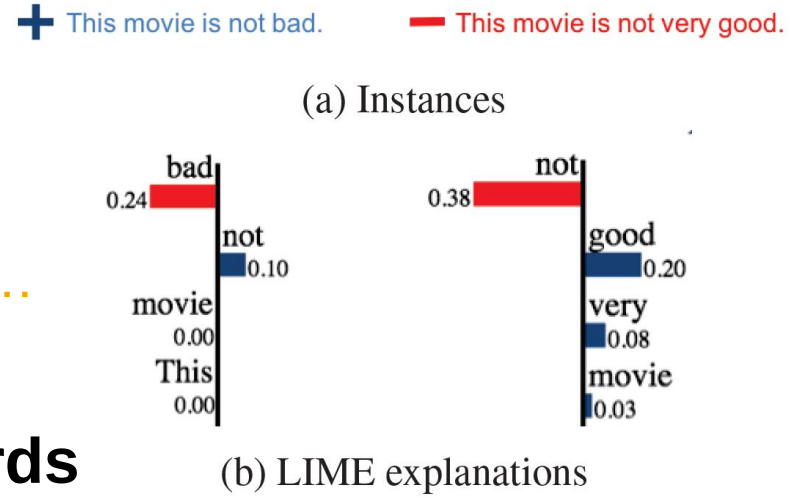
Pixels

(Kindermans et al. 2018), Fig. 6



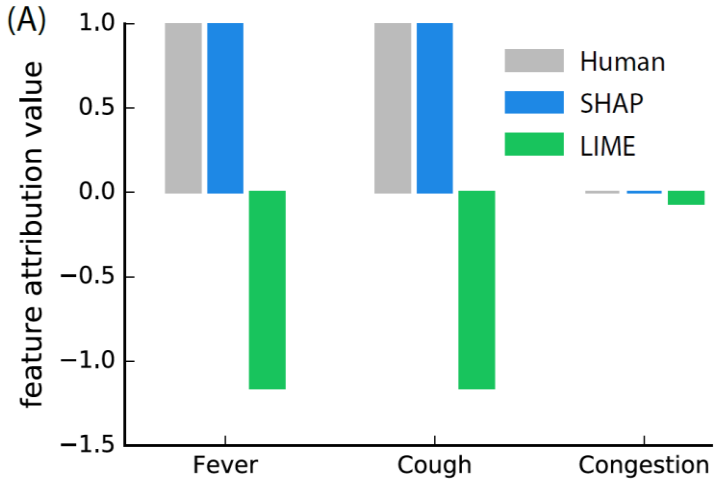
Super-pixels

(Ribeiro et al. 2016), Fig. 4



Words

(Ribeiro et al. 2018), Fig. 1



Tabular Features

(Lundberg & Lee 2017), Fig. 4

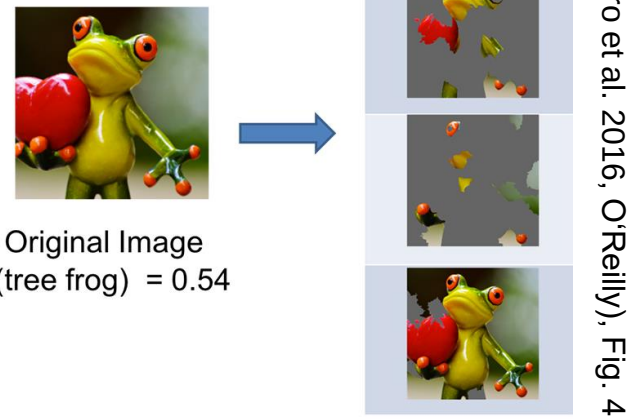
Feature Saliency: Types

› **Activation map**-based (Attention)

(cf. attention models)

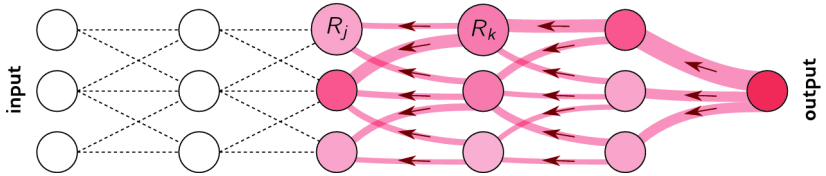
(Kim & Canny 2017, Fig. 5)

› **Perturbation**-based

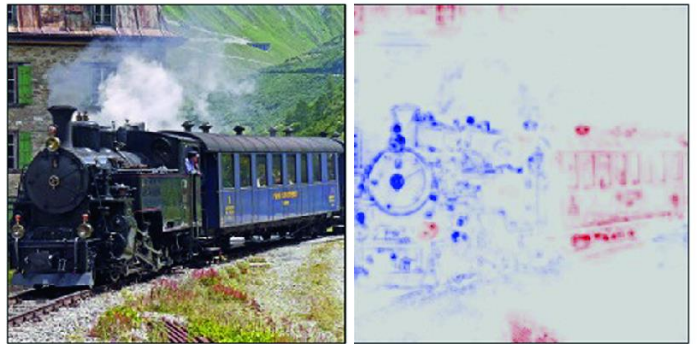


(Ribeiro et al. 2016, O'Reilly), Fig. 4

› **Backpropagation-** or **Gradient**-based



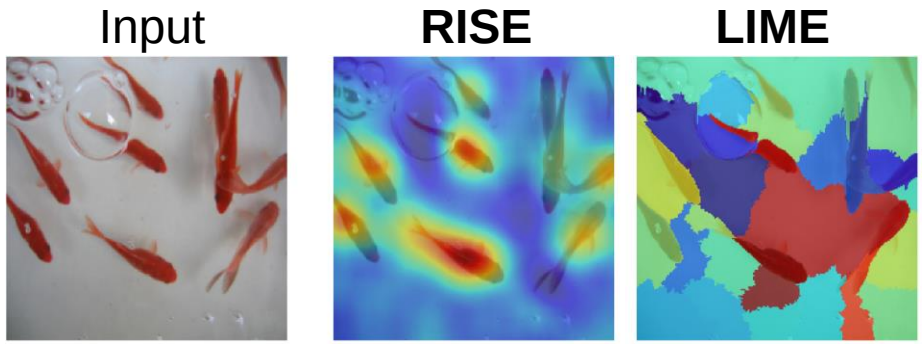
(Montavon et al. 2019), Fig. 10.2



(Montavon et al. 2019), Fig. 10.5

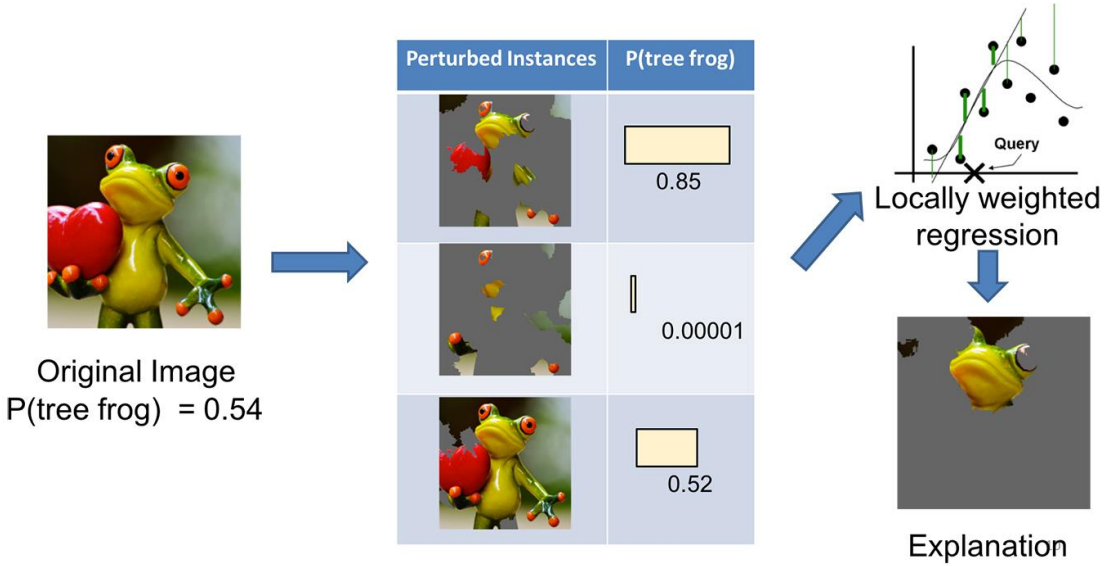
Feature Saliency: Perturbation-based

- *Idea:* “Remove” features and observe effect
- Often **model-agnostic**
- **Several inferences** needed
- *Differences:*
  - Definition of **features** & feature **removal**, *e.g.*

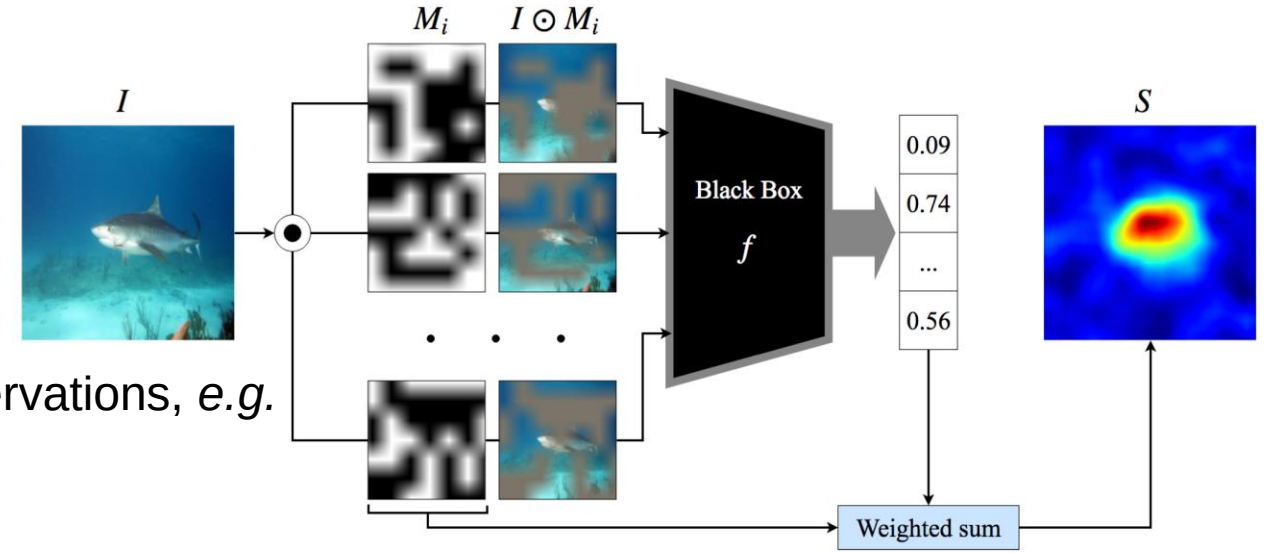


(Petsiuk et al. 2018), Fig. 2

  - **Calculation** of feature importance from observations, *e.g.* linear regression in LIME (Ribeiro et al. 2016) vs. gaming theory in SHAP (Lundberg & Lee 2017)



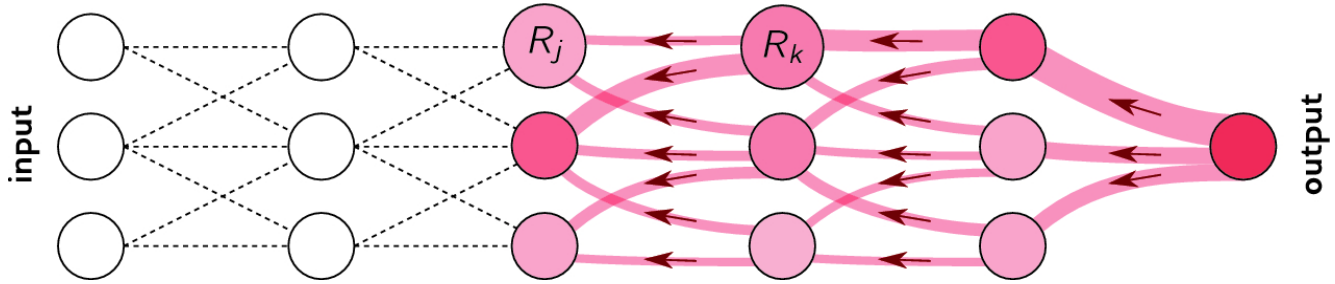
LIME
(Ribeiro et al. 2016, O'Reilly), Fig. 4



RISE
(Petsiuk et al. 2018), Fig. 3

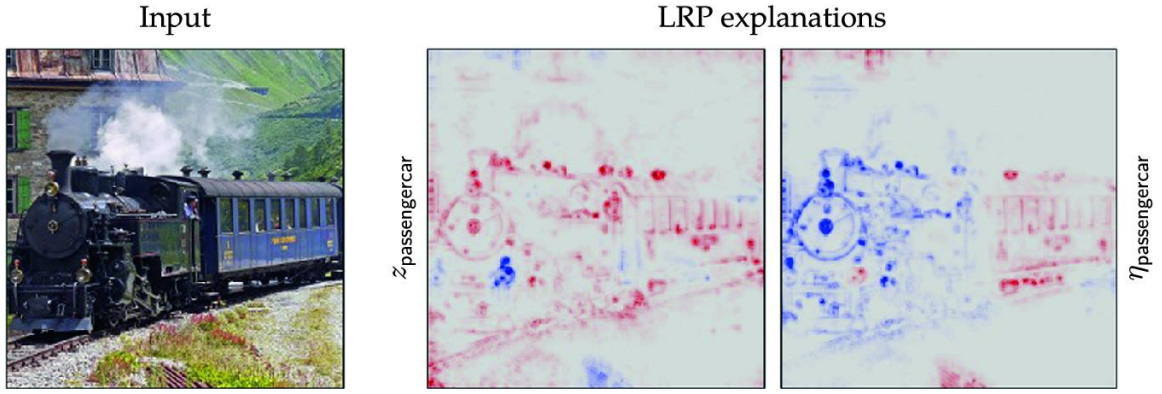
Feature Saliency: Backpropagation-based

- *Idea:*
  - **Trace back influence** (=*Relevance*) of activations from output to input
  - Total relevance within a layer l stays constant:
    $f(x) = \cdots = \sum_i R_i^{(l-1)} = \sum_i R_i^l$
- One additional **backwards-pass**
- Requires **access to model** internals
- Backpropagation functions **must be chosen carefully** wrt. layer type and question



(Montavon et al. 2019), Fig. 10.2

- Special case applicable to general differentiable models: Use gradients!



(Montavon et al. 2019), Fig. 10.5

Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	**Methods for Explaining ML Models**	**10**
4.1	Inherently Interpretable and Blended Models	11
4.2	Feature Importance	14
4.3	**Explaining Representations**	**19**
4.4	Explainable Surrogates	23

Latent Space Analysis: Feature Visualization

› *What does a network unit/part (e.g. neuron, channel) encode?* Use e.g.

Examples
activating unit strongly

DeepDream Prototypes
= starting image optimized to activate unit strongly

Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

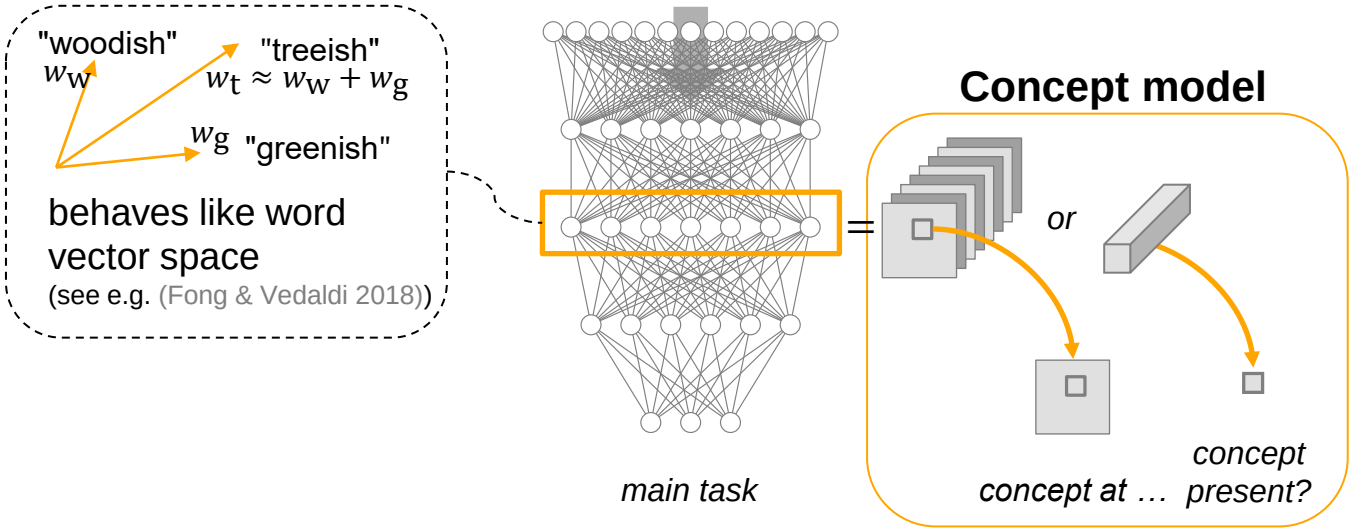
Clouds—or fluffiness? *mixed4a, Unit 453*

Buildings—or sky? *mixed4a, Unit 492*

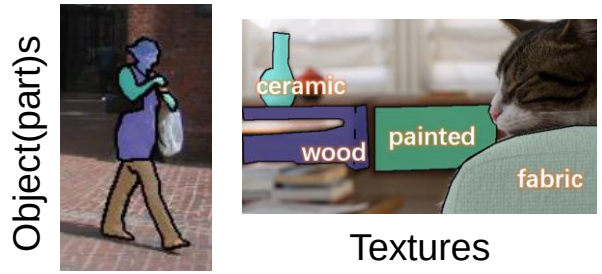
(Olah et al. 2017), Fig. 5

Latent Space Analysis: Concept Embedding Analysis

- *Goal: Associate semantic concept w/ latent space vector / subspace*
- *Idea:* Vector as parameters of simple predictor for concept (*concept model*)

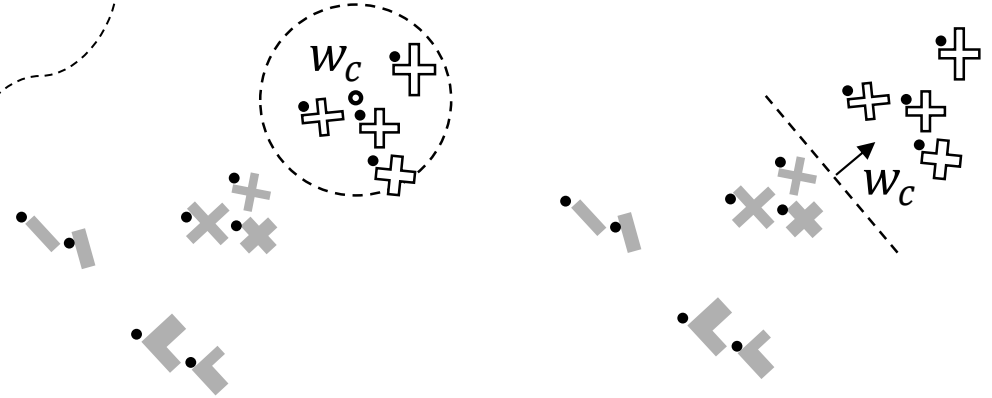


Some concept types



(Xiao et al. 2018), p. 6, Fig. 3

Main types of concept models:



Clustering & matrix factorization
e.g. ACE (Ghorbani et al. 2019), NCAV (Zhang et al. 2021)

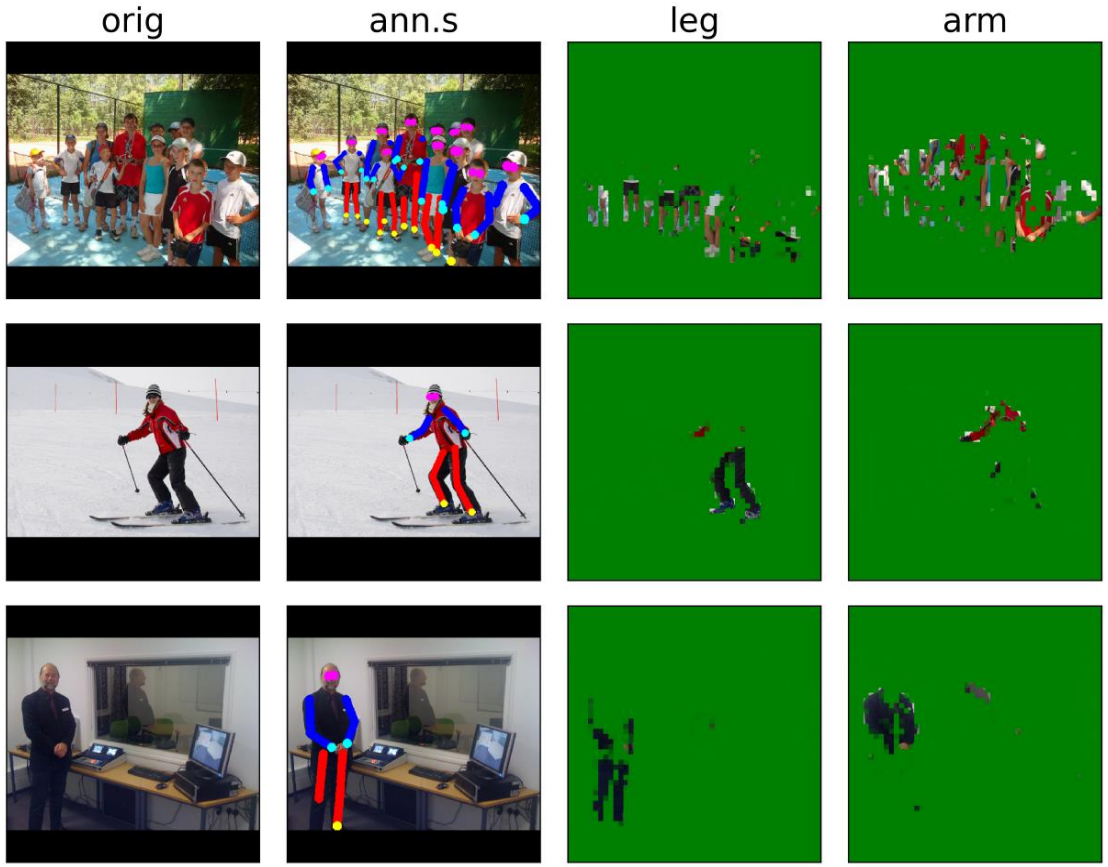
Linear
$f_c = \cdot \circ w_c + t_c$
e.g. TCAV (Kim et al. 2018), Net2Vec (Fong & Vedaldi 2018)

cf. (Schwalbe 2022), Fig. 3

Latent Space Analysis: Concept Embedding Analysis

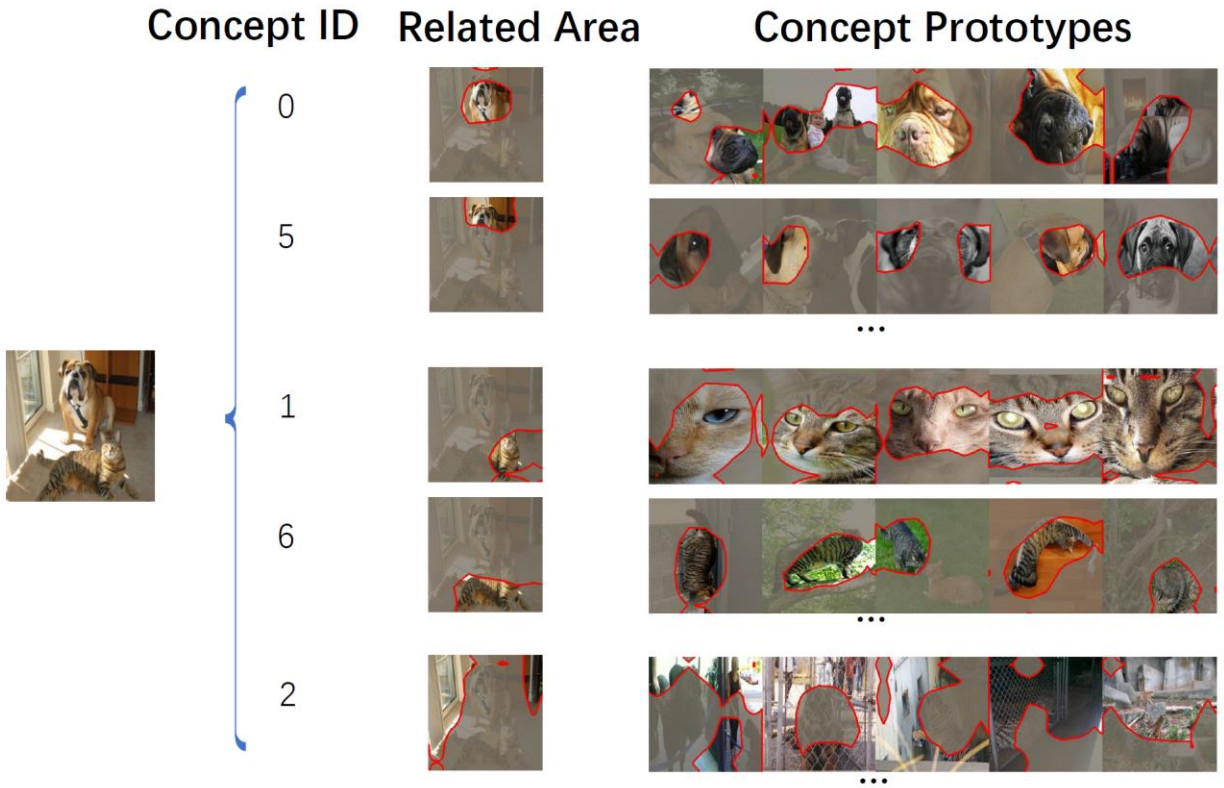
Examples

› **Supervised** using simplified Net2Vec on object detection DNN:



(Schwalbe 2021), Fig. 5

› **Unsupervised** using NCAV on classification DNN:



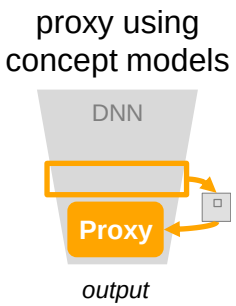
(Zhang et al. 2021), Fig. 1

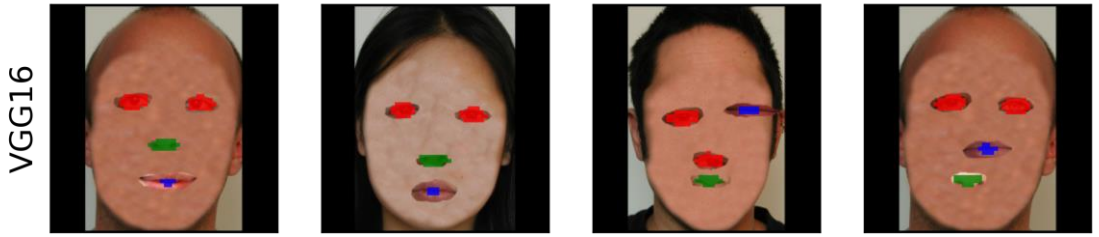
Agenda

1	What is Explainability?	3
2	Applications	5
3	A Method Taxonomy	8
4	**Methods for Explaining ML Models**	**10**
4.1	Inherently Interpretable and Blended Models	11
4.2	Feature Importance	14
4.3	Explaining Representations	19
4.4	**Explainable Surrogates**	**23**

Explainable Surrogates

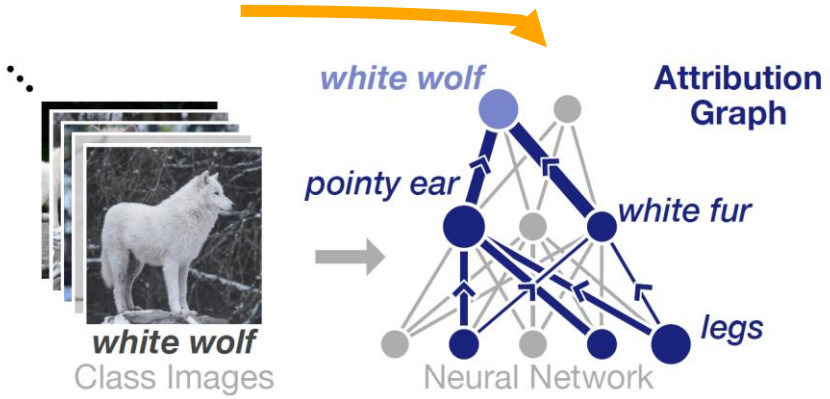
- *Idea:* Approximate DNN (parts) by interpretable model
- Example surrogate types:
  - (Locally) **linear**, *e.g.* LIME (Ribeiro et al. 2016)
  - Local or global **decision tree or rules**, *e.g.* **CA ILP** (Rabold et al. 2020) *(local)*
  - Dependency / flow **graphs**, *e.g.* **SUMMIT** (Hohman et al. 2020)



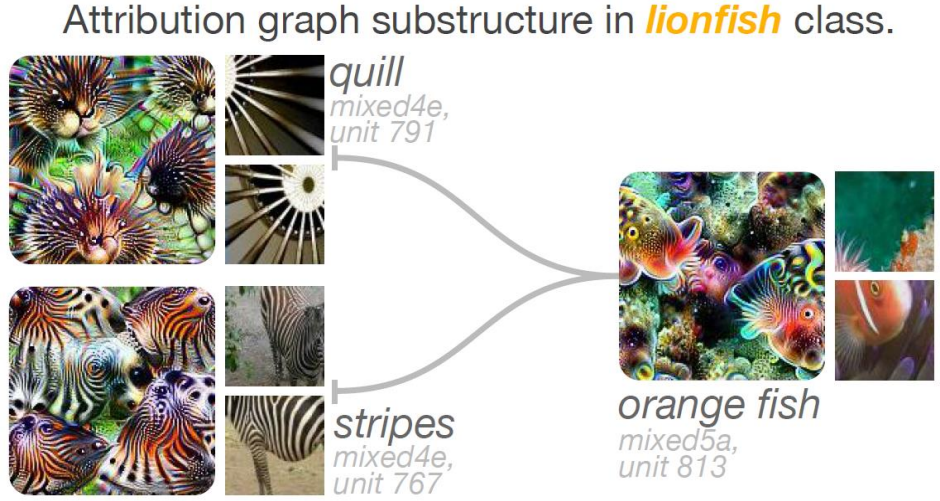


(Rabold et al. 2020), Fig. 4

```
face(F) :- contains(F, A), isa(A, nose), contains(F, B), isa(B, mouth),
           top_of(A, B), contains(F, C), top_of(C, A).
```



(Hohman et al. 2020), Fig. 2



(Hohman et al. 2020), Fig. 7

What Makes “Good” Explanations?

Recommendations:

On local explanations (=*“Why this decision?”*) (Miller 2019):

- Use **contrastive** explanations (=“Why this action *instead of another*?”)
- **Causal links** > probabilities
- **Explainee background** matters!
- **Context** matters for explanation selection! (Why is the explanation needed?)

On global explanations (=“How does it work?”):
Use interpretable models where possible! (Rudin 2019)

Metric examples:

- Functional level:
  - Faithfulness
  - Coverage
  - Accuracy
  - Scalability
  - …
- User level:
  - Comprehensibility
  - Improvement of human-AI system
  - …

Tradeoffs necessary
(no one-fits-all method)

Conclusion

- Explanation here means “*help a human understand an aspect of a model*”
  → *intrinsically cognitive!*

- Field of XAI:
  - very **diverse**
  - **needed** for many applications

- The **why**, **who**, **what**, and **how** matter!
- **Tradeoffs** require careful choice of method(s)

- More research needed!

Thanks for listening! Questions?

Contact: Gesina.Schwalbe@continental-corporation.com

Some further reading by us:

Schwalbe, Gesina, and Bettina Finzel. 2021. “A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts.” https://arxiv.org/abs/2105.07190.

Schwalbe, Gesina. 2022. “Concept Embedding Analysis: A Review.” http://arxiv.org/abs/2203.13909.

References (I)

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” PLOS ONE 10 (7): e0130140. https://doi.org/10.1371/journal.pone.0130140.

Bruckert, Sebastian, Bettina Finzel, and Ute Schmid. 2020. “The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions.” Frontiers in Artificial Intelligence 3: 507973. https://doi.org/10.3389/frai.2020.507973.

Chen, Chaofan, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. 2019. “This Looks like That: Deep Learning for Interpretable Image Recognition.” In Advances in Neural Information Processing Systems 32, edited by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, 32:8928–39. Vancouver, BC, Canada. https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

Fong, Ruth, and Andrea Vedaldi. 2018. “Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks.” In Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition, 8730–38. Salt Lake City, UT, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00910.

Ghorbani, Amirata, James Wexler, James Y. Zou, and Been Kim. 2019. “Towards Automatic Concept-Based Explanations.” In Advances in Neural Information Processing Systems 32, edited by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, 9273–82. Vancouver, BC, Canada. http://papers.nips.cc/paper/9126-towards-automatic-concept-based-explanations.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. “Explaining Explanations: An Overview of Interpretability of Machine Learning.” In Proc. 5th IEEE Int. Conf. Data Science and Advanced Analytics, edited by Francesco Bonchi, Foster J. Provost, Tina Eliassi-Rad, Wei Wang, Ciro Cattuto, and Rayid Ghani, 80–89. Turin, Italy: IEEE. https://doi.org/10.1109/DSAA.2018.00018.

Hohman, Fred, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2020. “Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations.” IEEE Transactions on Visualization and Computer Graphics 26 (1): 1096–1106. https://doi.org/10.1109/TVCG.2019.2934659.

Karimi, Amir-Hossein, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2021. “A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects.” ArXiv:2010.04050 [Cs, Stat], March. http://arxiv.org/abs/2010.04050.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. “Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).” In Proc. 35th Int. Conf. Machine Learning, 80:2668–77. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm, Sweden: PMLR. http://proceedings.mlr.press/v80/kim18d.html.

References (II)

Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. “Concept Bottleneck Models.” In *Proc. 2020 Int. Conf. Machine Learning*, 5338–48. PMLR. http://proceedings.mlr.press/v119/koh20a.html.

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. “What Do We Want from Explainable Artificial Intelligence (XAI)? -- A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research.” *Artificial Intelligence*, February, 103473. https://doi.org/10.1016/j.artint.2021.103473.

Losch, Max, Mario Fritz, and Bernt Schiele. 2019. “Interpretability beyond Classification Output: Semantic Bottleneck Networks.” In *Proc. 3rd ACM Computer Science in Cars Symp. Extended Abstracts*. Kaiserslautern, Germany. https://arxiv.org/pdf/1907.10882.pdf.

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Long Beach, CA, USA: Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Miller, Tim. 2019. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” *Artificial Intelligence* 267 (February): 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. “Layer-Wise Relevance Propagation: An Overview.” In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 193–209. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_10.

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. 2017. “Feature Visualization.” *Distill* 2 (11): e7. https://doi.org/10.23915/distill.00007.

Páez, Andrés. 2019. “The Pragmatic Turn in Explainable Artificial Intelligence (XAI).” *Minds and Machines* 29 (3): 441–59. https://doi.org/10.1007/s11023-019-09502-w.

Petsiuk, Vitali, Abir Das, and Kate Saenko. 2018. “RISE: Randomized Input Sampling for Explanation of Black-Box Models.” In *Proc. British Machine Vision Conf.*, 151. Newcastle, UK: BMVA Press. http://bmvc2018.org/contents/papers/1064.pdf.

Rabold, Johannes, Gesina Schwalbe, and Ute Schmid. 2020. “Expressive Explanations of DNNs by Combining Concept Analysis with ILP.” In *KI 2020: Advances in Artificial Intelligence*, edited by Ute Schmid, Franziska Klügl, and Diedrich Wolter, 148–62. Lecture Notes in Computer Science. Bamberg, Germany: Springer International Publishing. https://doi.org/10.1007/978-3-030-58285-2_11.

References (III)

Ribeiro, Marco Túlio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 1135–44. KDD '16. San Francisco, California, USA: ACM. https://doi.org/10.1145/2939672.2939778.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction." O'Reilly Media. August 12, 2016. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/.

———. 2018. "Anchors: High-Precision Model-Agnostic Explanations." In *Proc. AAAI Conf. Artificial Intelligence*. Vol. 32. https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Schwalbe, Gesina. 2021. "Verification of Size Invariance in DNN Activations Using Concept Embeddings." In *Artificial Intelligence Applications and Innovations*, edited by Ilias Maglogiannis, John Macintyre, and Lazaros Iliadis, 374–86. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-79150-6_30.

———. 2022. "Concept Embedding Analysis: A Review." *ArXiv:2203.13909 [Cs, Stat]*, March. http://arxiv.org/abs/2203.13909.

Schwalbe, Gesina, and Bettina Finzel. 2021. "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts." *ArXiv:2105.07190 [Cs]*, May. https://arxiv.org/abs/2105.07190.

Zhang, Ruihan, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. 2021. "Invertible Concept-Based Explanations for CNN Models with Non-Negative Concept Activation Vectors." In *Proc. 35th AAAI Conf. Artificial Intelligence*, 35:11682–90. virtual: AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/17389.

Zilke, Jan Ruben, Eneldo Loza Mencía, and Frederik Janssen. 2016. "DeepRED – Rule Extraction from Deep Neural Networks." In *Proc. 19th Int. Conf. Discovery Science*, 457–73. Lecture Notes in Computer Science. Bari, Italy: Springer International Publishing. https://doi.org/10.1007/978-3-319-46307-0_29.