UNIVERSIDAD POLITÉCNICA DE MADRID

HORIZON-MSCA-2023-DN-01-01 ID-101168344
MSCA Doctoral Networks 2023
TUAI Project

Deep Visual Analytics

Research workshop

Ángel Panizo-LLedot · Universidad Politécnica de Madrid

CC BY-NC 4.0

TUAI - D2.1

AI+DA

Who am I



Ángel Panizo LLedot

Universidad Politécnica de Madrid

angel.panizo@upm.es

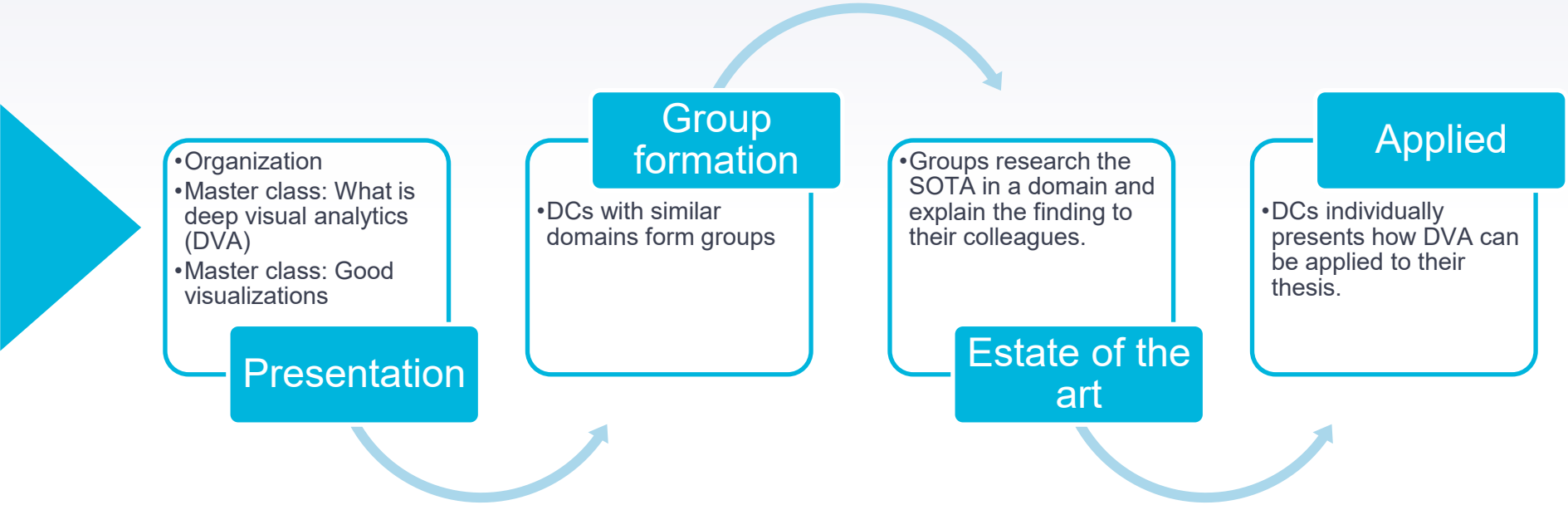
Research Workshop

Workshops developed over three months, featuring weekly or **biweekly** meetings in a face-to-face or **hybrid format**. They are based on **problem-based learning**, promoting a **student-centered** approach oriented toward solving real, open-ended challenges.

RW deep visual analytics

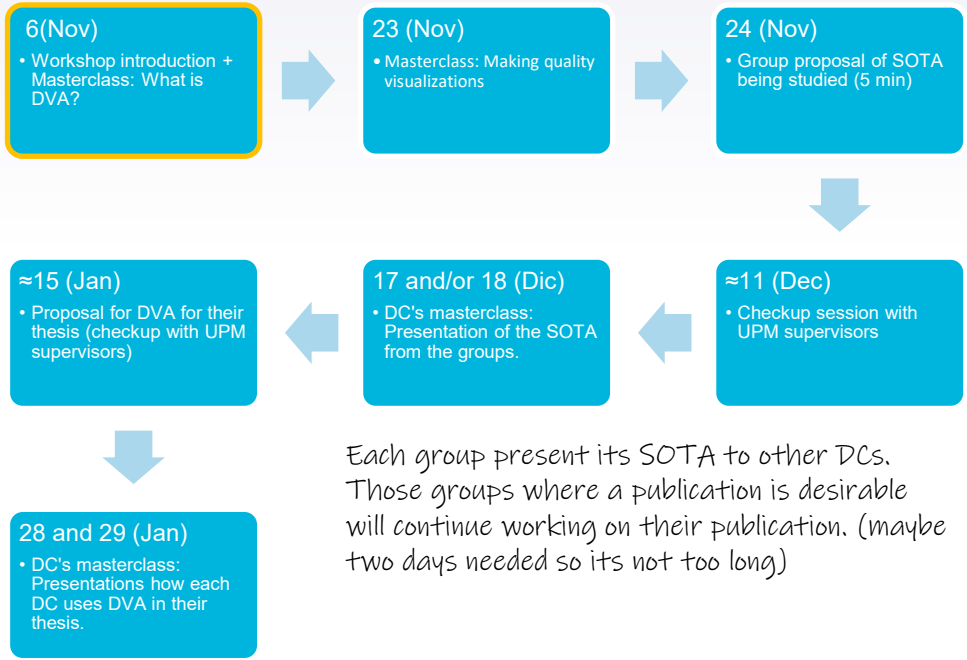
- Deep visual analytics is very wide field and dependent on the domain (image, graph, time-series … etc).
- The objective of the workshop is for the DCs to **delve deeper** into the deep visual analytics techniques of their **thesis domain** while acquiring a **general knowledge** of how to apply it to **other domains** (those of their colleagues).

Plan for the workshop



Organization

DCs must get into groups and decide which part of SOTA want to explore



Each group will have a UPM supervisor assigned. In the checkup session the work so far is evaluated. The focus here must be deciding if a publication is feasible/desirable

After the SOTA each DC must do individual presentation on how DVA is going to be used in their thesis. That is discussed in this session with their UPM supervisors. Publication related subject can be addressed here as well

Each group present its SOTA to other DCs. Those groups where a publication is desirable will continue working on their publication. (maybe two days needed so its not too long)

Final individual presentations on how DCs will use DVA in their thesis and discussion with colleagues with feedback. (two days for sure)

Today:
What is Deep Visual Analytics





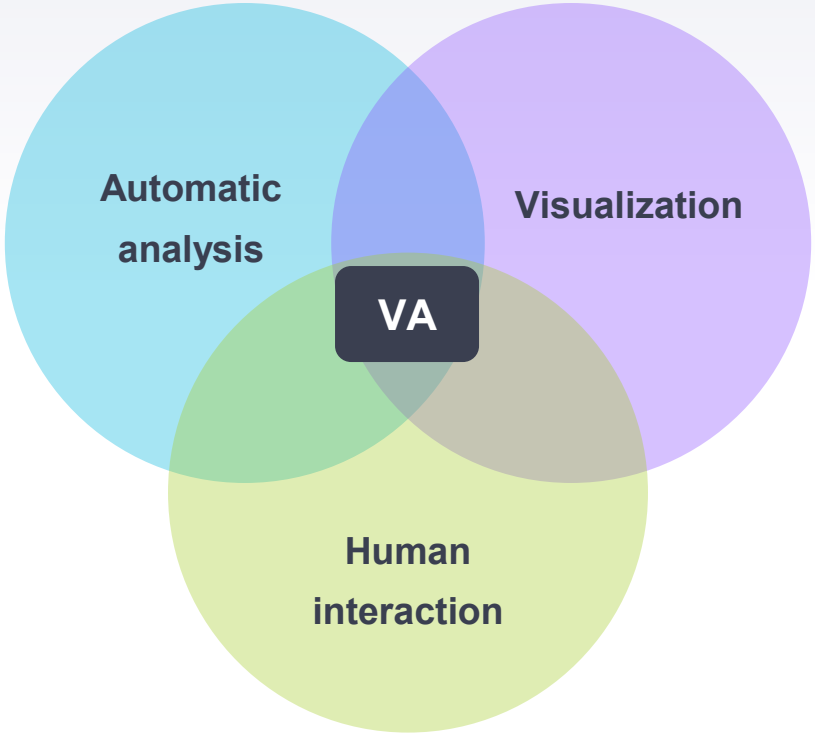
WHAT IS ~~DEEP~~ VISUAL ANALYTICS ?



The science of analytical reasoning facilitated by interactive visual interfaces

Thomas & Cook, 2005

Three ingredients

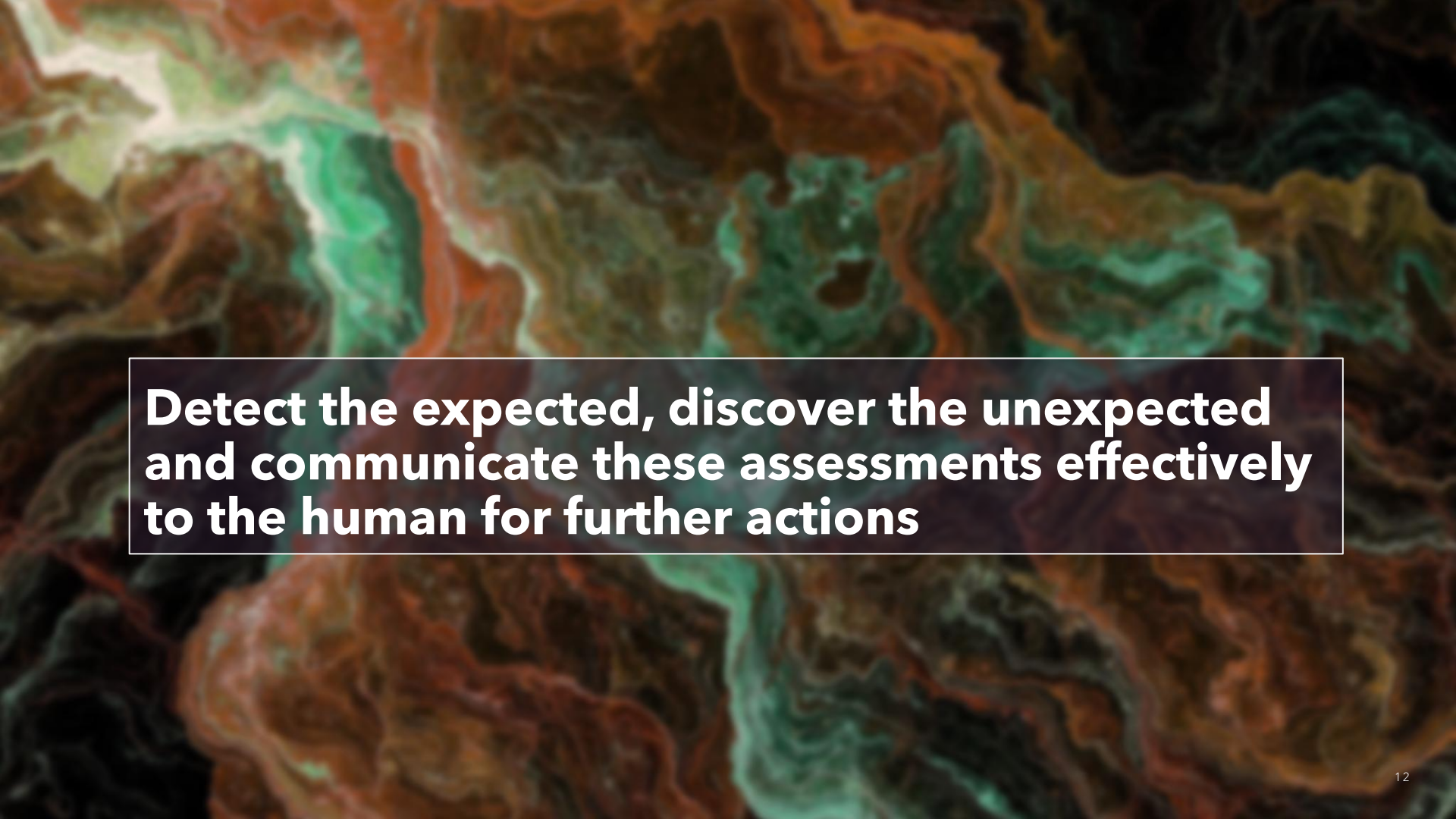


Three ingredients



The goal of VA is to turn complex data into actionable **insight** by combining the computational power of automated analysis with human perception, domain knowledge and judgment through interactive visual interfaces.

interaction


Detect the expected, discover the unexpected and communicate these assessments effectively to the human for further actions

Machine vs human

🤖 Only automatic

- **scales** to huge data
- fails when the problem is **ill-defined**

👤 human visualization

- uses **expertise** and judgment
- **doesn't scale**



Machine vs human



💡

The output of VA lives **in the analyst's head**.

Case Study:
Detecting Water Leaks with Visual Analytics

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

- Minimum Night Flow (MNF)
- MNF Trend
- Pressure Drop
- Early Morning Peak
- Summer/Winter Ratio



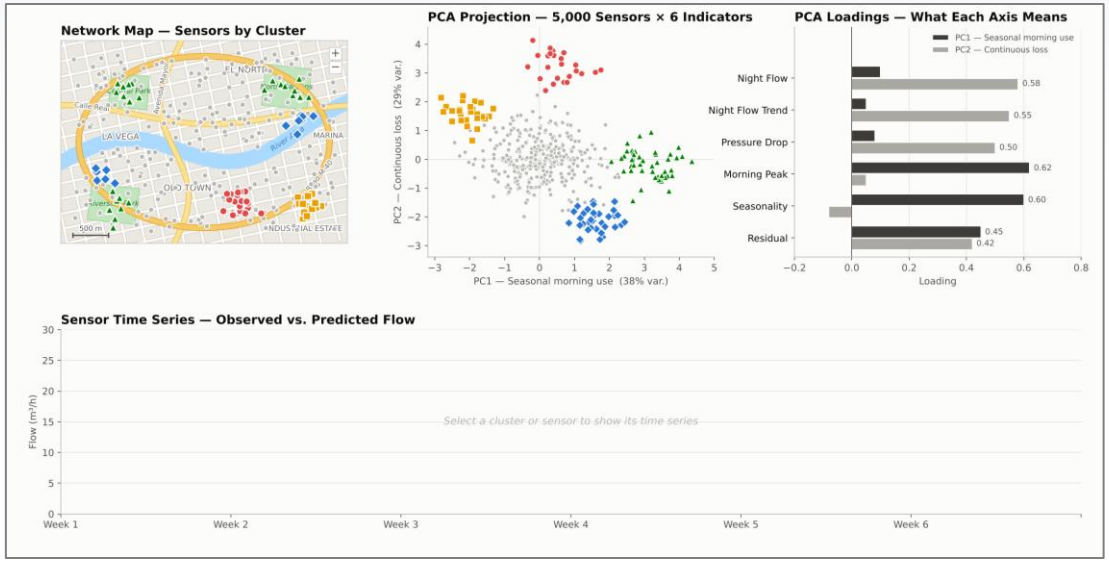
⚠️ **3,000 Alerts!**

😱 **Only 10 teams and needs to select which must be checked before**

Case Study:
Detecting Water Leaks with Visual Analytics

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

Human Operator

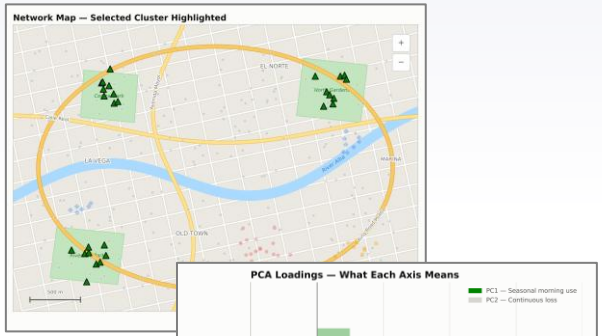


Human operator uses a control panel to check the alerts

Case Study:
Detecting Water Leaks with Visual Analytics

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

Human Operator





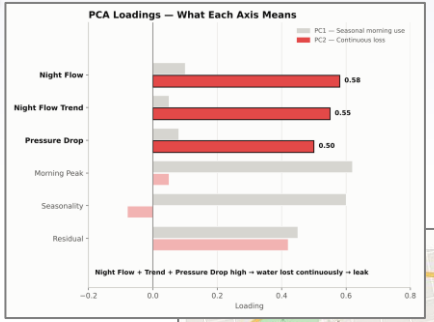
A cluster of alerts corresponds to green areas of the city. It looks that the ML algorithm is detecting the increase irrigation in summer as an anomaly. Model must be retrained

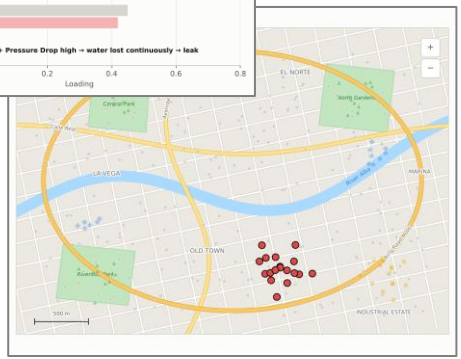
Human operator finds a cluster that corresponds to Parks and green areas

Case Study:
Detecting Water Leaks with Visual Analytics

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

Human Operator







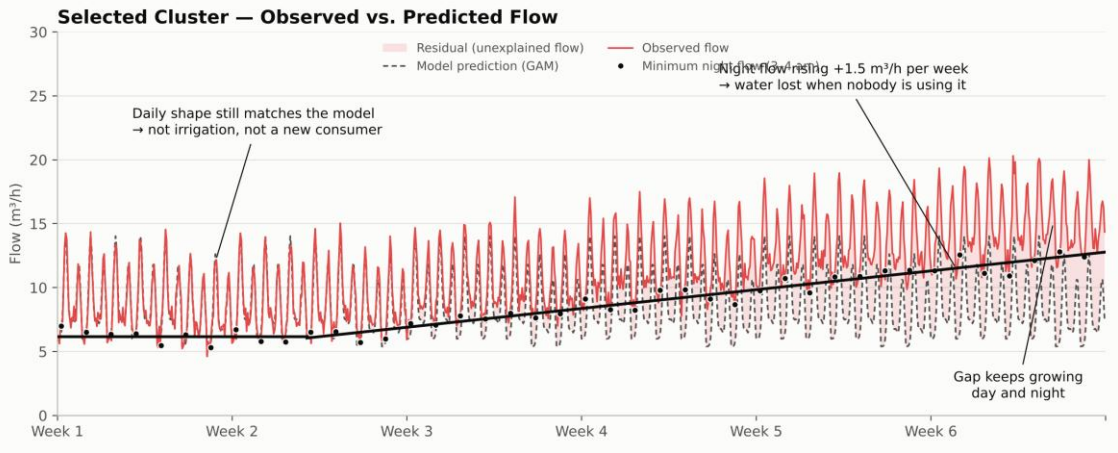
A cluster of alerts corresponds to old parts of the city. The PCA plots shows that pressure drops and night flow anomalous. It can correspond to a leakage.

Human operator finds a cluster that corresponds to al old part of town

Case Study:
Detecting Water Leaks with Visual Analytics

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

Human Operator



Human operator check the sensors plot of some sensors of the zone and see the leakage

Case Study: Conclusions

5,000 sensors · **400** districts · **500M** readings/year · **~20%** water lost

3,000
alerts from the model alone

−95% →

150
alerts after feedback

focus →

25
pipe sections for 10 teams

KEY INSIGHT

An alert is not a leak. The model flagged anomalies; only the human could explain them.

- **Park irrigation → false alarms,** fixed with the irrigation calendar
- **Rising night flow in 1960s pipes → hidden leaks,** now a renewal plan

The machine finds anomalies; the human finds the meaning.

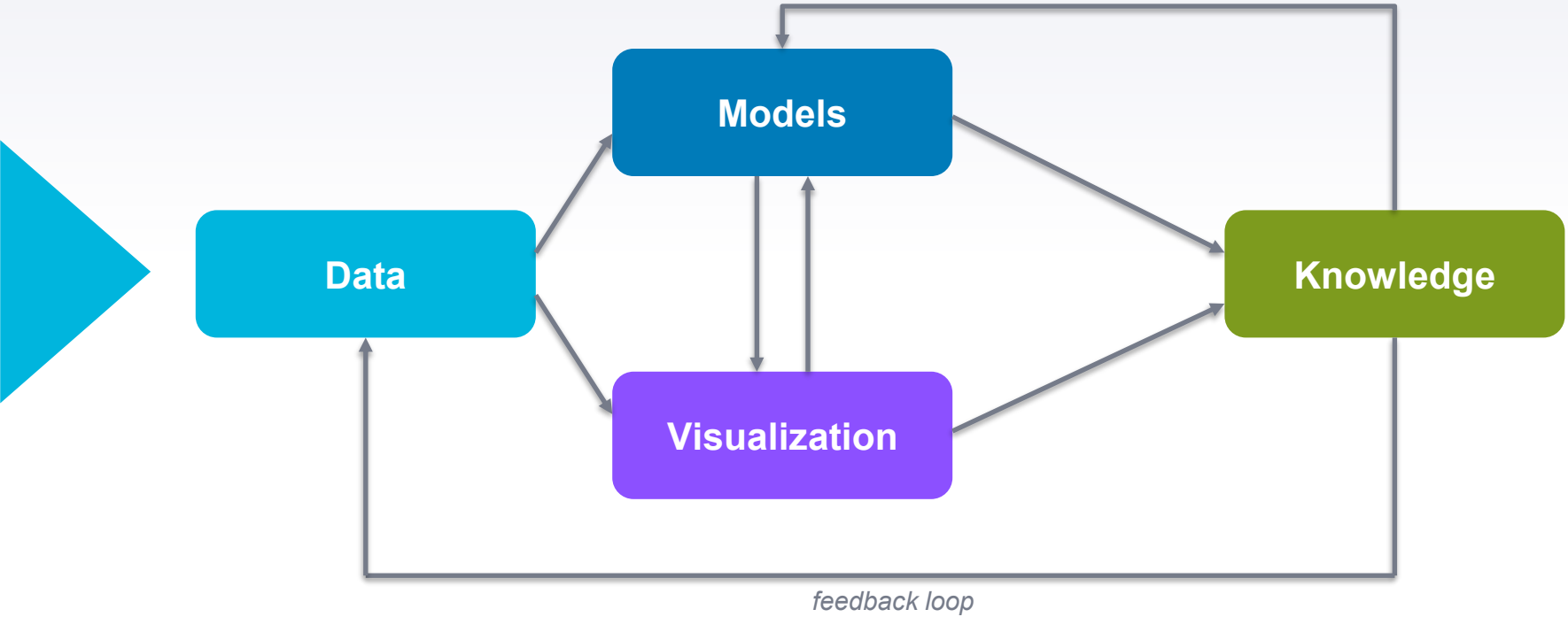
Machine vs human





VA lives where problems are **too big to just look** and **too ill-defined to just compute**.

VA process



Redrawn after Keim et al., 2008

🤔

IN CLASSIC VA, MODELS WERE TOOLS WE UNDERSTOOD. WHAT IF THE MODEL IS WHAT WE DON'T?

Vision model trained on Pascal VOC



ML Model

Is a Horse !

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

Vision model trained on Pascal VOC



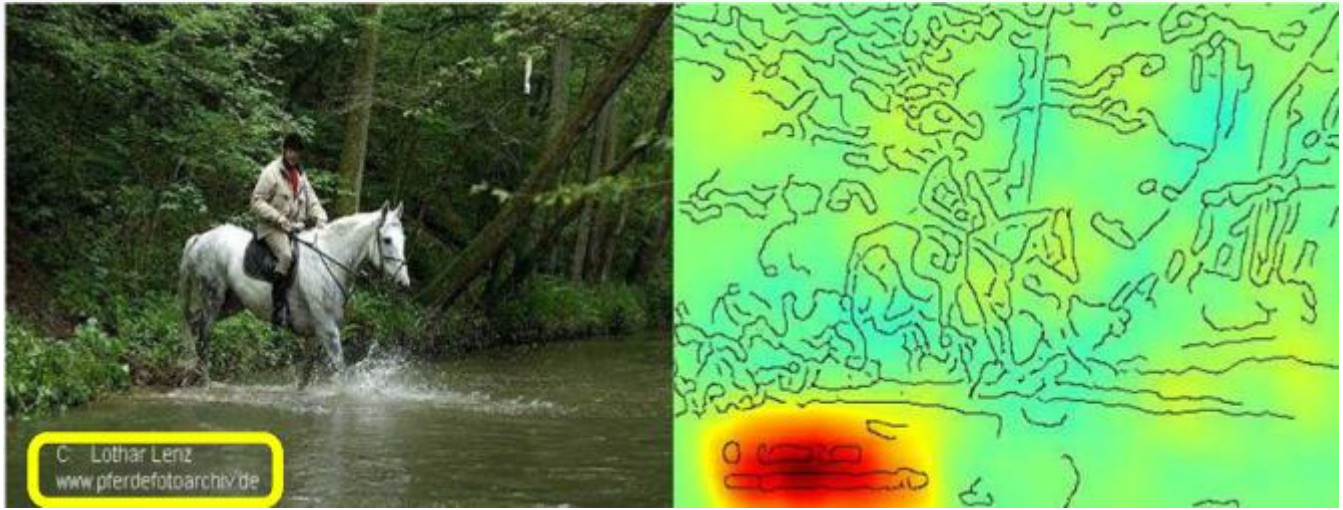
🤔

Your model get it right. But does it get it right for the **right reason**?

orse !

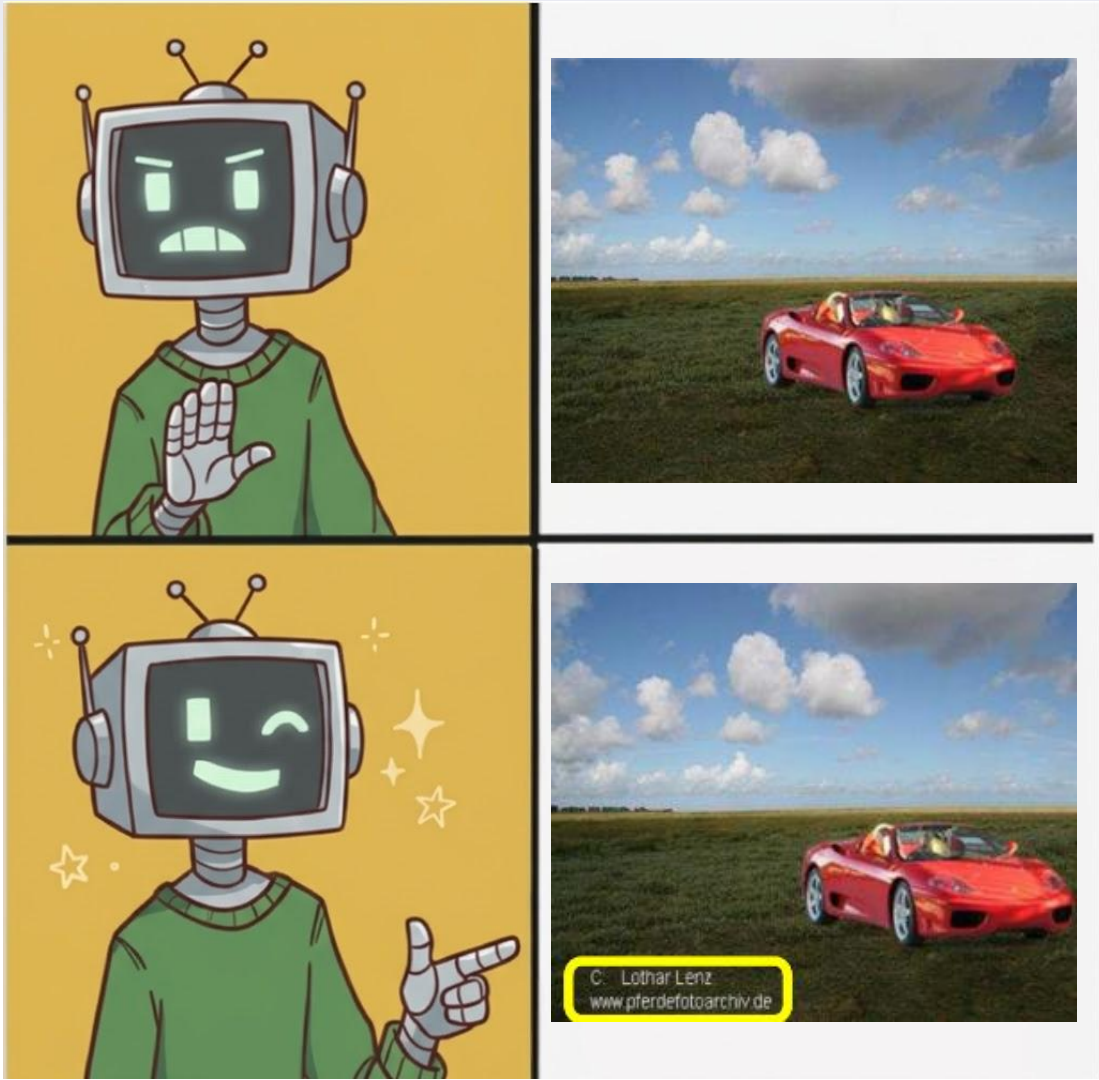
Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

What does the model look at?



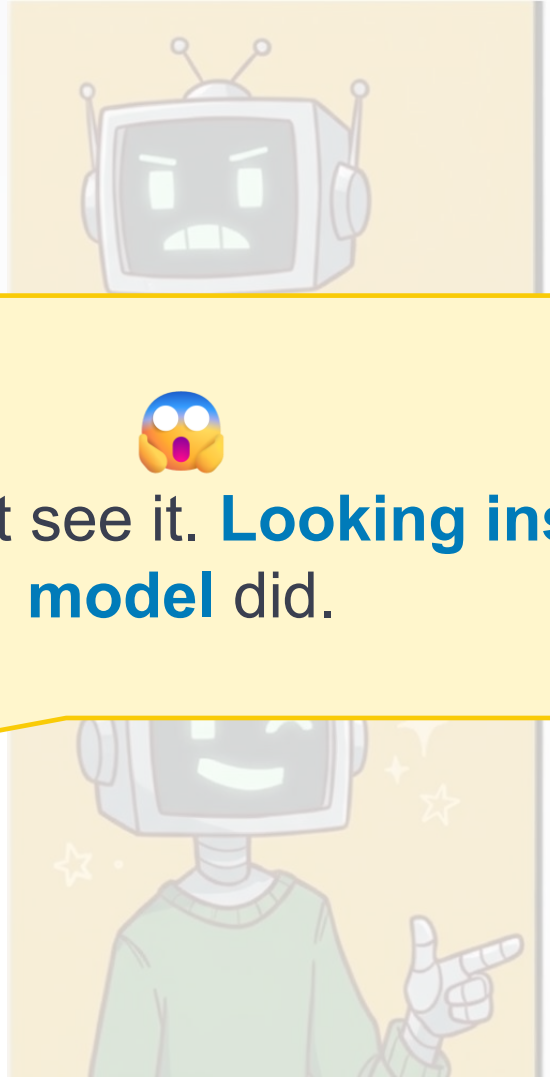
Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

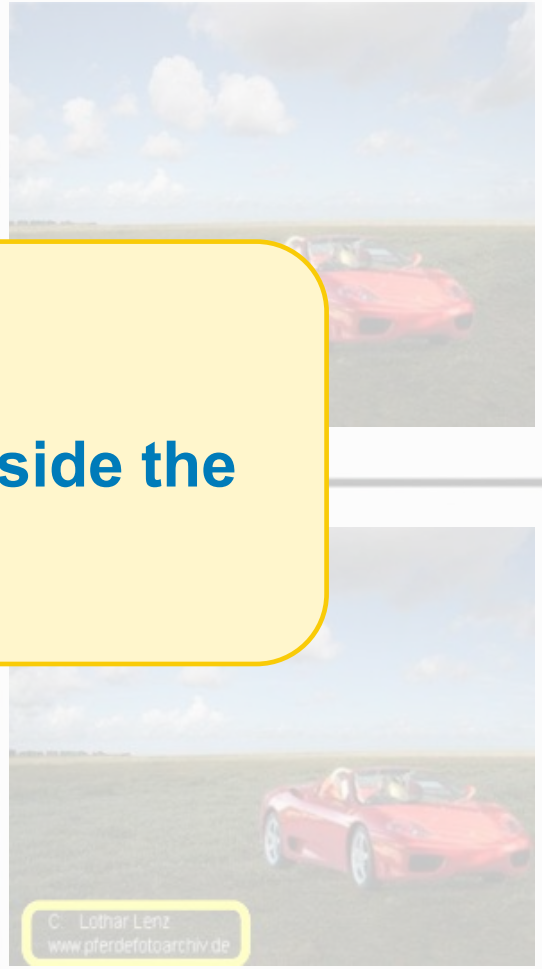
Is this a horse ?



Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

Is this a horse ?

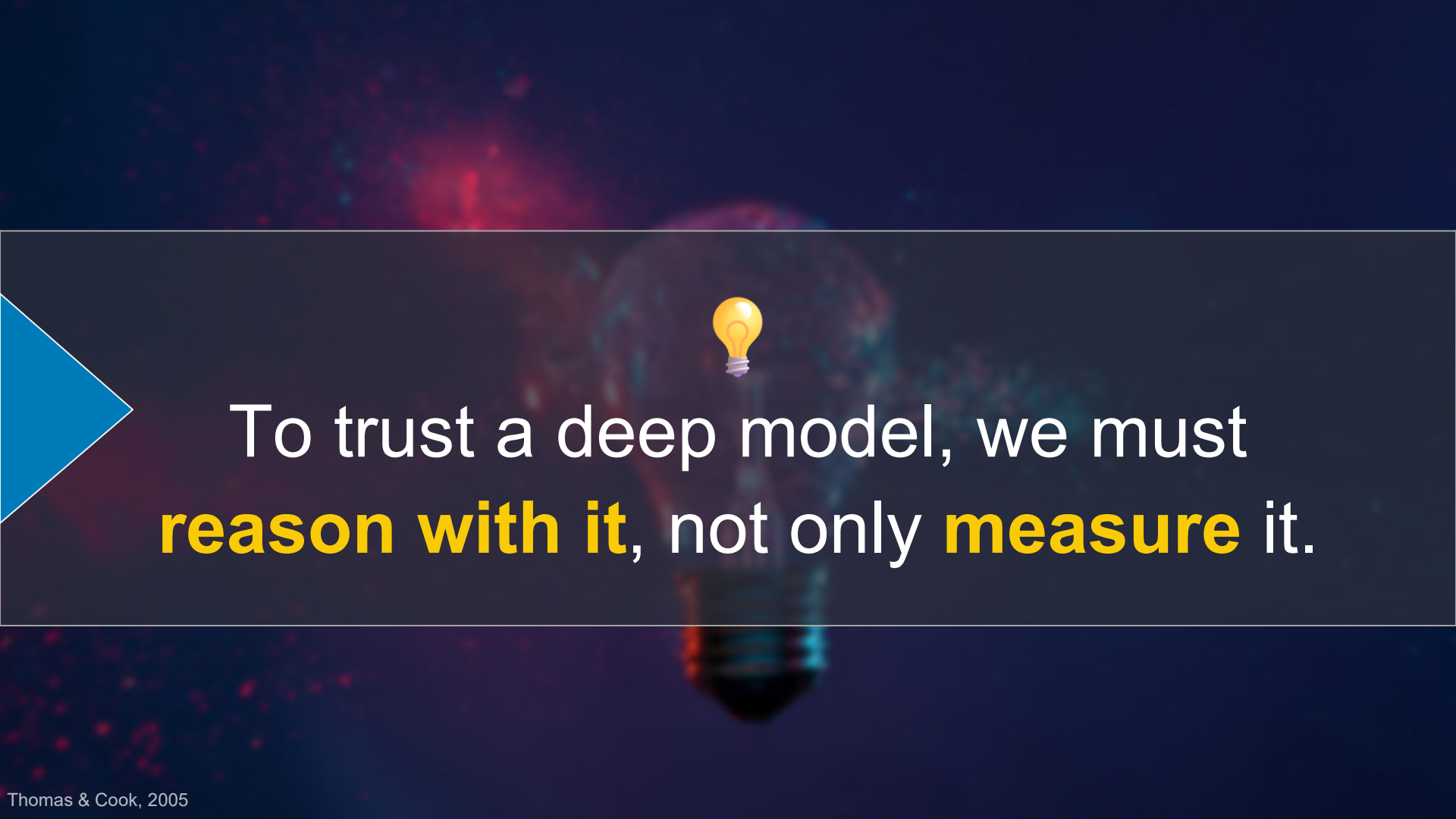




😱

Accuracy didn't see it. **Looking inside the model** did.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.



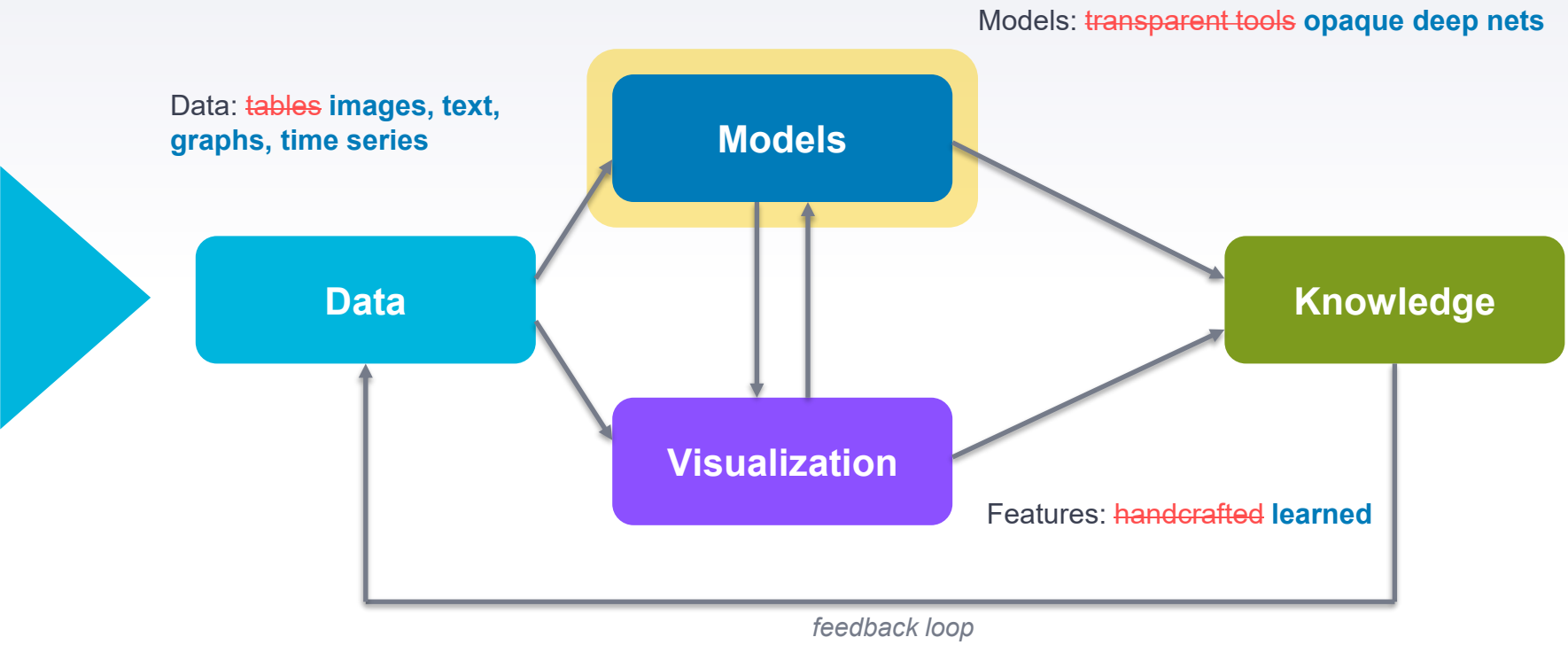
Thomas & Cook, 2005



From VA to DVA

WHEN THE MODEL BECOMES THE THING WE DO NOT UNDERSTAND

VA process


Models: ~~transparent tools~~ opaque deep nets
Data: ~~tables~~ images, text, graphs, time series
Models
Data
Knowledge
Visualization
Features: ~~handcrafted~~ learned
feedback loop

VA process

Models: ~~transparent tools~~ **opaque deep nets**

Data: ~~tables~~ **images, text, graphs, time series**

💡 **Key idea**

DVA = VISUAL ANALYTICS WHEN DEEP LEARNING IS THE PROTAGONIST, WITH THE HUMAN AT THE CENTER OF THE REASONING

Which human?

Architects

DL KNOWLEDGE

DL experts

ROLE

Develop new components or architectures, or adapt existing ones to new problems

Trainers

DL KNOWLEDGE

Know DL, don't research architectures

ROLE

Apply existing architectures with known recipes, tuning only hyperparameters and data. Also called practitioners

End users

DL KNOWLEDGE

Little or no DL knowledge

ROLE

Use pre-trained models in their domain: clinicians, domain experts, the general public

More DL expertise — Less DL expertise

Taxonomy from Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., & Rush, A. M. (2018). Seq2Seq-Vis: A visual debugging tool for sequence-to-sequence models. IEEE TVCG, 25(1), 353–363.

🧶 THREAD 1 · THE HUMAN IN DVA/XAI

An explanation is only worth something if it helps **a specific person do a specific thing**. This thread covers **how we evaluate, for whom, and how we communicate to non-ML experts**.

- Which **evaluation levels** exist?
- Which **quantitative metrics** exist and which can be computed without users?
- How do we balance **fidelity** against **comprehensibility** when simplifying for a clinician or an operator?
- How do we measure **trust**, and how do we detect overtrust and undertrust?
- What does **verbalization** add over a heatmap?
- How do we show the **uncertainty** of the model and of the explanation?

VIS4DL vs DL4VIS

🔬 VIS4DL

- DL is the **object**
- understand, debug, improve **models**

🛠️ DL4VIS

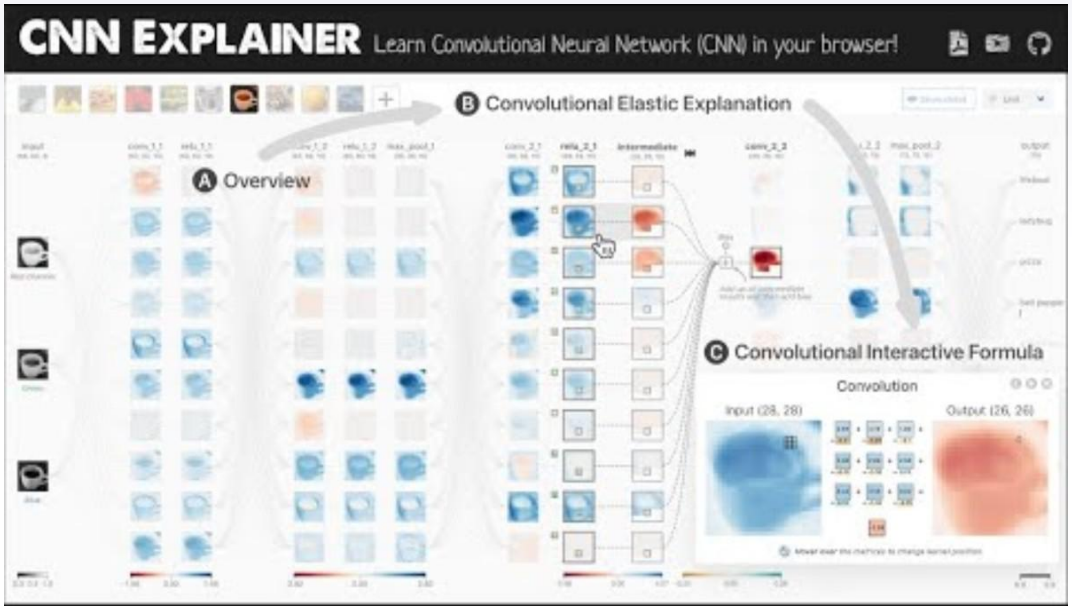
- DL is the **instrument**
- **recommend, generate, read** charts



Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... & Chau, D. H. P. (2020). CNN explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1396-1406.

VIS4DL: an example

CNN Explainer, an interactive tool to learn how a CNN classifies images

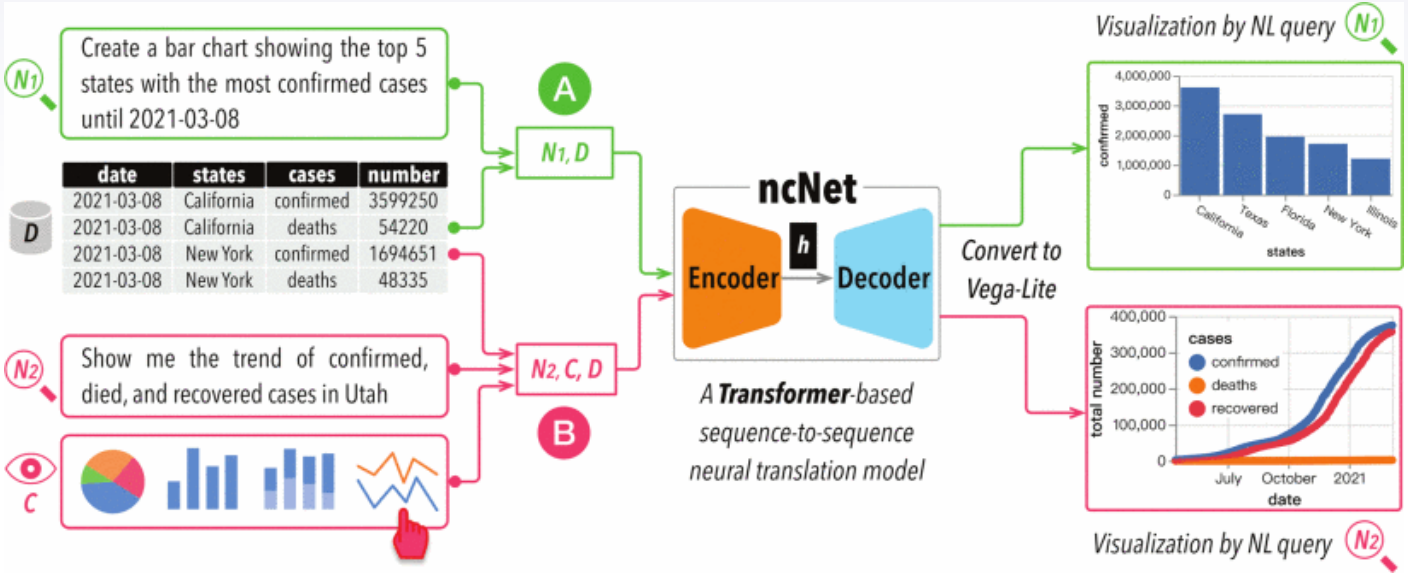


https://www.youtube.com/watch?v=HnWIHWFbuUQ

💡 **DL is the object: the visualization explains the model**

DL4VIS: an example

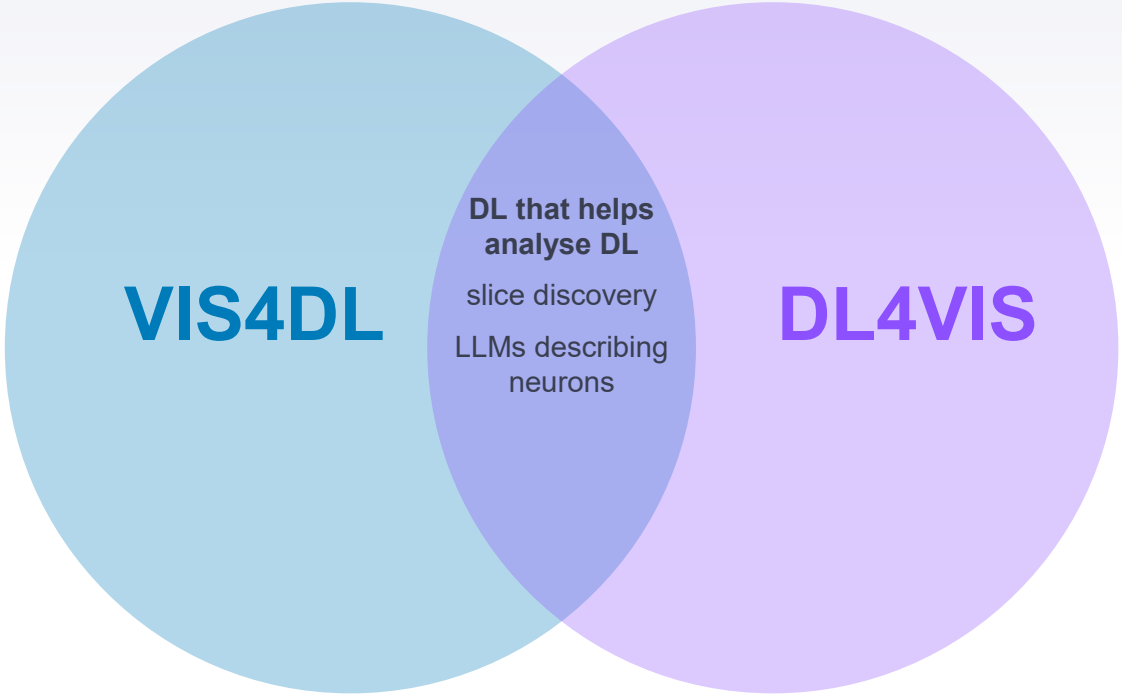
NL2VIS: from a question in natural language to a chart



💡 **DL is the instrument: the model builds the visualization**

Luo, Y., Tang, N., Li, G., Tang, J., Chai, C., & Qin, X. (2021). Natural language to visualization by neural machine translation. IEEE Transactions on Visualization and Computer Graphics, 28(1), 217-226.

Overlap


VIS4DL
DL that helps analyse DL
slice discovery
LLMs describing neurons
DL4VIS

Overlap


VIS4DL
DL that helps analyse DL
slice discovery
LLMs describing neurons
Today we focus on VIS4DL.

Overlap

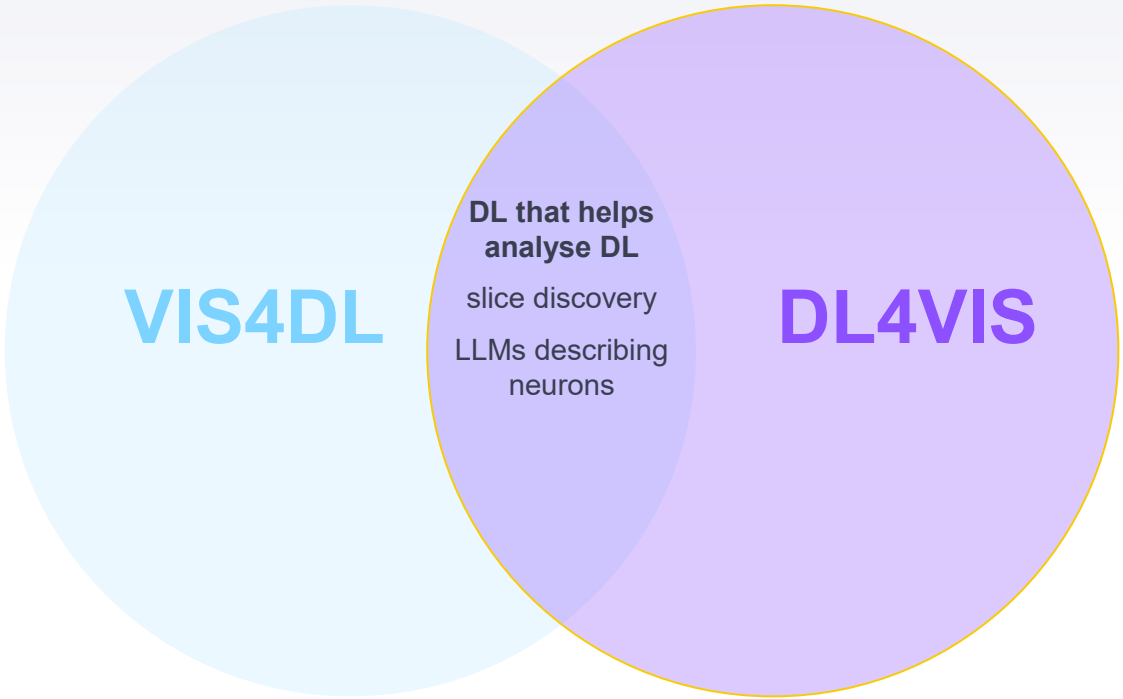

VIS4DL
VIS4XDL
No only where and when the model fails, but also why !

🧶 THREAD 2 · DL4VIS & LLMS

The reverse direction to this class: DL and LLMs as **tools to generate, recommend, read and summarize visualizations**. It includes the overlap (slice discovery, LLMs that describe neurons). If the model picks the view and writes the insight, **is the human still at the center?**

- Which **visualization tasks** does DL automate today? (AI4VIS and ML4VIS taxonomies.)
- What changes with **foundation models** (VIS4FM vs. FM4VIS)?
- What systems are already there that do this?
- What risks do **hallucinations** and **overreliance** bring, and how do we evaluate a generated chart?

Overlap


VIS4DL
DL that helps analyse DL
slice discovery
LLMs describing neurons
DL4VIS



ISN’T THIS JUST **XAI**?

XAI vs VA

XAI

- **methods**
- output: an **artifact** (heatmap, SHAP)
- about model **decisions**
- human as **receiver**



VA

- a **process**
- output: **knowledge**
- the whole **pipeline**
- human as **agent**

XAI vs DVA

XA

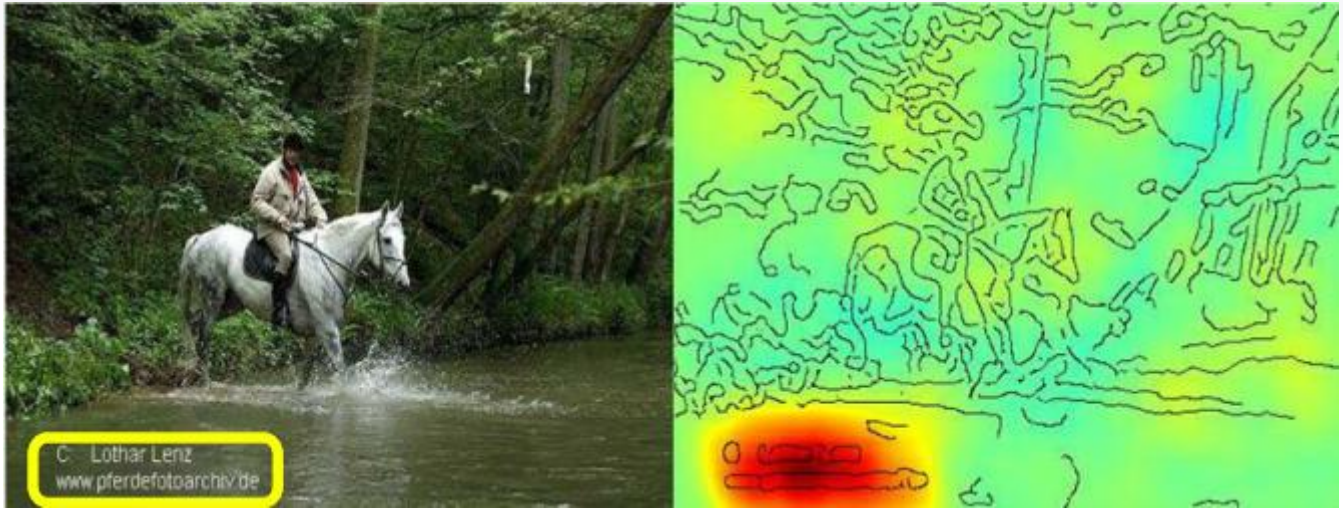
- human as



XAI gives VA **what to show**. VA gives XAI **aggregation, comparison, interaction, context**.

One explanation

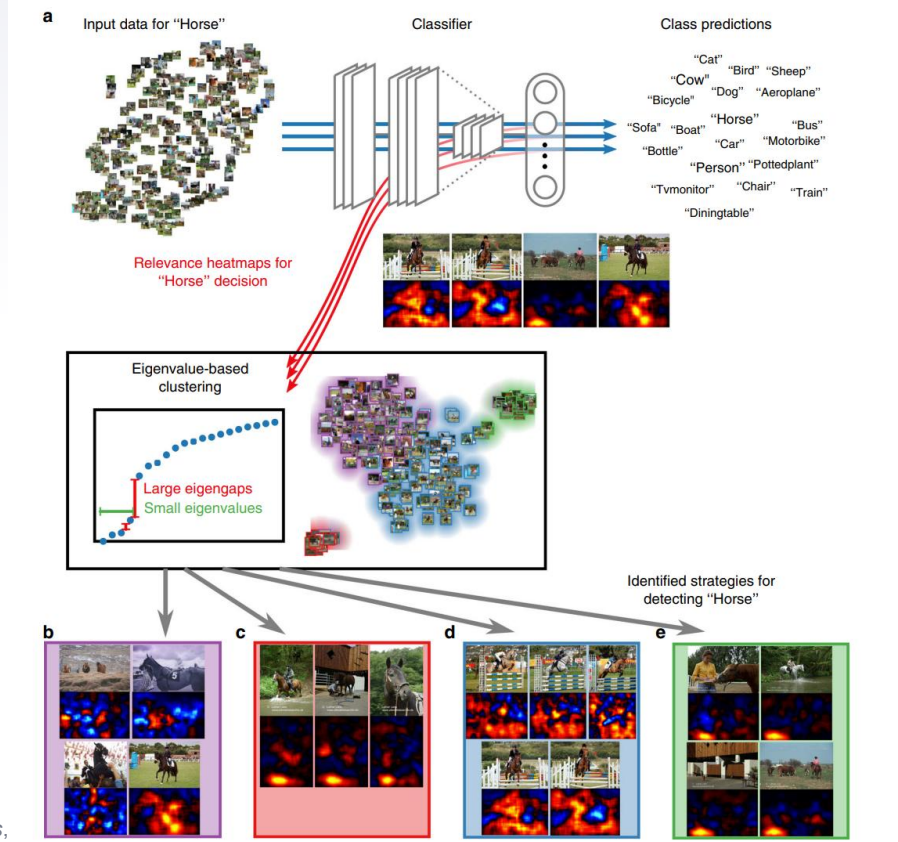
XAI



Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

Thousands of explanations

DVA



Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.

XAI vocabulary

🤝 Trustworthiness
Confidence that the model acts as intended. Trustworthy does not mean explainable

For: Domain experts, affected users

🔗 Causality
Hint at causal links to test later; models only see correlations

For: Domain experts, managers, regulators

🔄 Transferability
Understand the model's limits to reuse that knowledge elsewhere or improve it

For: Domain experts, data scientists

💡 Informativeness
Extract information about the problem and the model's inner relations to support decisions

For: All users

🎯 Confidence
Robustness and stability: know in which regime the model is reliable

For: Domain experts, developers, managers, regulators

⚖️ Fairness
Enable ethical analysis: detect bias in the data and prevent unfair uses

For: Affected users, regulators

🙋 Accessibility
Let non-experts take part in improving and developing the model

For: Product owners, managers, affected users

🎛️ Interactivity
Let users interact with the model and tune it

For: Domain experts, affected users

🔒 Privacy awareness
Know what the model has memorized, and whether the explanation itself leaks data

For: Affected users, regulators

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

XAI techniques

	 **Gradients** Internal quantity	 **Perturbations** Procedure	 **Activations** Internal quantity	 **Search algorithms** Procedure	 **Attention weights** Internal quantity
Uses	Derivative of output w.r.t. input: ∂o/∂x	Modified inputs and how the output changes	Neuron and layer values in the forward pass	Other samples from the dataset	Weights the model already computes
Model access	White-box, via backpropagation	Black-box: only needs to call f(x)	White-box	Varies: needs a similarity space, often latent	Transformers and attentive models only
Returns	A score per feature (pixel)	A score per feature, or an alternative input	Concept a neuron or layer encodes	Real examples from the dataset	Weights over parts of the input

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum, 42*(1), 319–355.

How the components differ

Computed extra, or already there?

Attention: already there, a by-product of inference

Gradients: one backward pass

Perturbations: many forward passes

Attention and gradients both end in a similar heatmap.

Inside the model, or only input and output?

White-box: gradients, activations, attention

Black-box: perturbations, so they work on any model

Same question as perturbations, but gradients measure an infinitesimal change analytically; perturbations, a real one empirically.

Explain with this input, or with other samples?

This input: gradients, perturbations, attention

The network: activations, what a neuron or layer learned

Other samples: search algorithms, “it looks like this”

Only activations explain the network rather than this prediction.

Three components are internal quantities (what you look at) and two are procedures (how you get the explanation). Most real methods combine several.

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum, 42*(1), 319–355.

Vocabulary of VA



Select
Mark something to follow



Explore
See other parts of data



Connect
See the same things in different views



Abstract/Elaborate
Summarize or show more detailed



Filter
Remove following a condition



Encode
Change representation



Reconfigure
Same data different shape

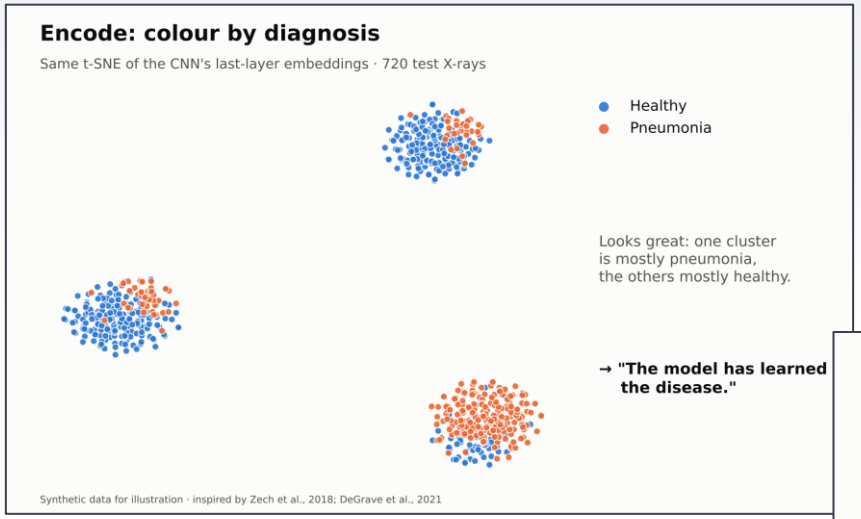


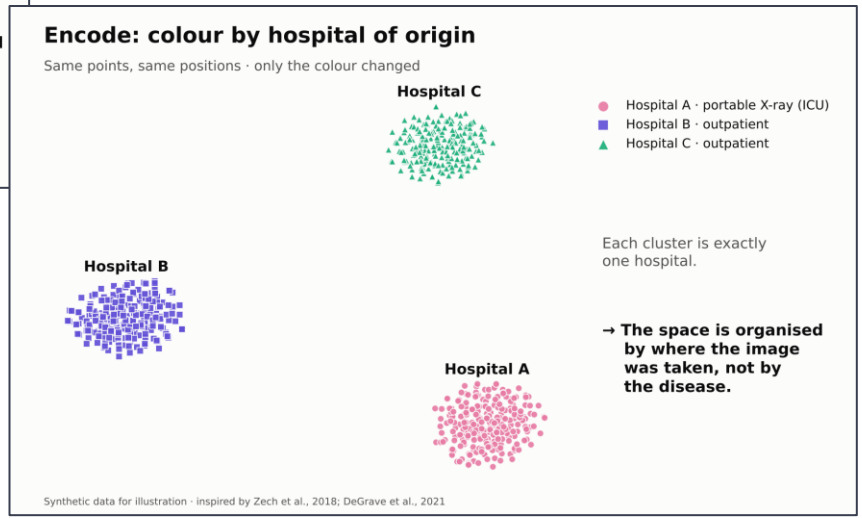
Guide
Propose what to do or look



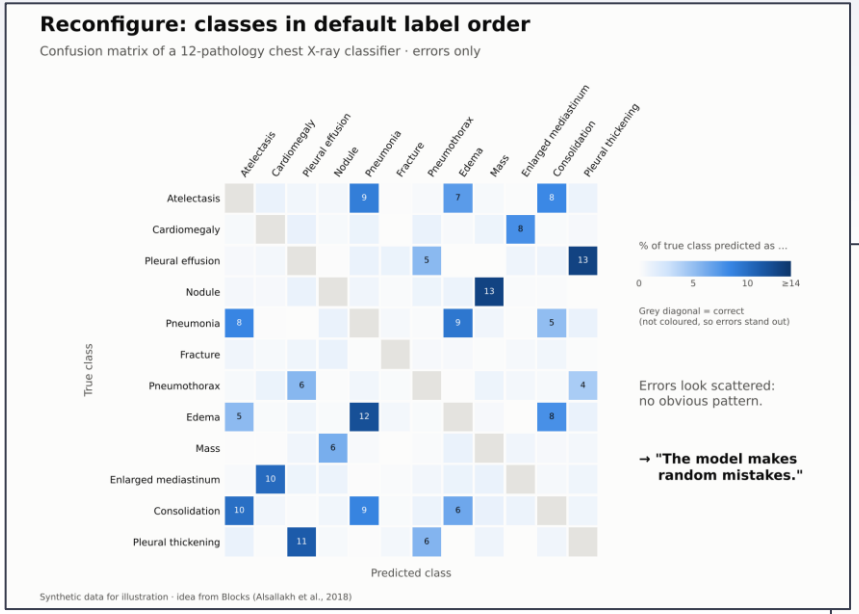
Verbalize
Explain in natural language

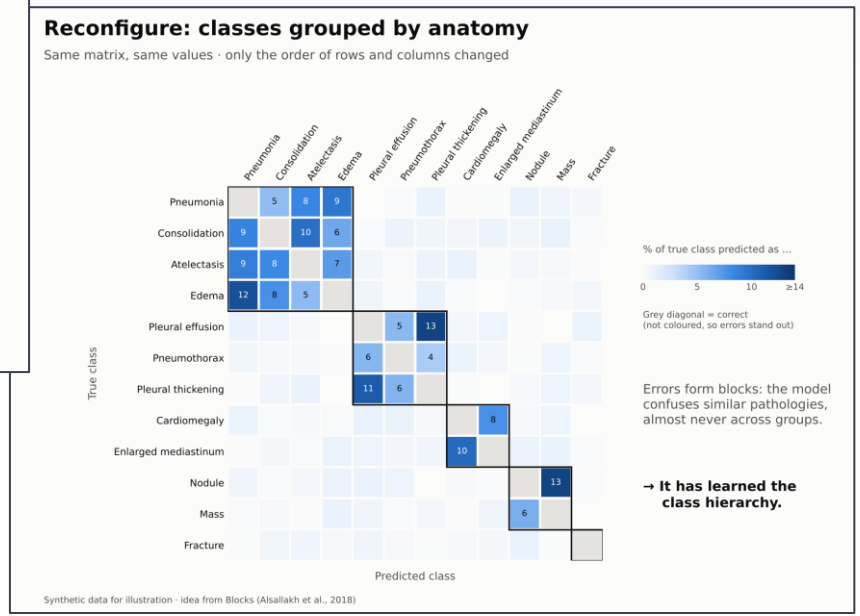
Encode


Encode: colour by diagnosis
Same t-SNE of the CNN's last-layer embeddings · 720 test X-rays
Healthy
Pneumonia
Looks great: one cluster is mostly pneumonia, the others mostly healthy.
→ "The model has learned the disease."
Synthetic data for illustration · inspired by Zech et al., 2018; DeGrave et al., 2021


Encode: colour by hospital of origin
Same points, same positions · only the colour changed
Hospital C
Hospital A · portable X-ray (ICU)
Hospital B · outpatient
Hospital C · outpatient
Hospital B
Each cluster is exactly one hospital.
Hospital A
→ The space is organised by where the image was taken, not by the disease.
Synthetic data for illustration · inspired by Zech et al., 2018; DeGrave et al., 2021

Reconfigure





VA techniques by data type

Data type	Technique	How it represents the data · refs. in the paper
Text	**Matrix & area charts**	Classic charts over word frequencies or attributes [KK15]
	Word clouds	Text size ∝ frequency or weight [AWS05, HLLE14]
	Word sequences	Node-link diagrams, glyph small multiples or in-text annotations [HDH*13]
Numbers	**Bar, pie, line & area charts**	Length, angle, position and area [AAF*20, KKEM10, TS20]
	Heatmaps	Color on a grid [AAF*20, KKEM10, TS20]
Distributions	**Histograms, box & violin plots**	Shape of the distribution: frequencies, quartiles, density [HN98]
Multi-dimensional	**Scatterplot (2D/3D)**	Each item is a point, in Cartesian or polar coordinates [SG18]
	Scatterplot + dim. reduction	PCA, MDS, t-SNE, ISOMAP or UMAP project to 2D/3D, keeping the original structure [EMK*21]
	Parallel coordinates (PCPs)	Each item is a polyline across one axis per dimension; clustering, line sampling and axis order help [ID90, HW12, JLJC05, BZP*20]
	RadViz & star coordinates	Radial 2D projection; dimensions are anchors on the circle [HGM*97, Kan00, RLSR21]
Hierarchical	**Treemaps**	Nested rectangles; area encodes a value, color or glyphs add more [Shn92, SW01, STLD20, ABC*19]
Networks	**Adjacency matrix**	Rows and columns are nodes, cells are edges; patterns depend on ordering [BBR*16]
	Node-link diagrams	Circles and lines; what you see depends on the layout algorithm [BBDW16]

Refs. as cited in La Rosa, B., et al. (2023). State of the art of visual analytics for explainable deep learning. Computer Graphics Forum, 42(1), 319–355.

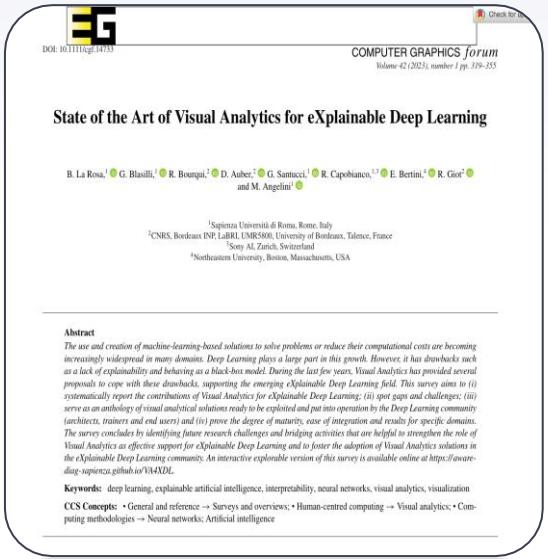
Today:
What is Deep Visual Analytics



Five ways to explain

VA SYSTEMS FOR XDL ORGANIZE AROUND THE KIND OF EXPLANATION THEY SUPPORT

Our map: La Rosa et al., 2023



DOI: 10.1111/cgf.14733

COMPUTER GRAPHICS *forum*

Volume 42 (2023), number 1 pp. 319–355

State of the Art of Visual Analytics for eXplainable Deep Learning

B. La Rosa,[1] G. Blasilli,[1] R. Bourqui,[2] D. Auber,[2] G. Santucci,[1] R. Capobianco,[1,3] E. Bertini,[4] R. Giot[2] and M. Angelini[1]

[1]Sapienza Università di Roma, Rome, Italy
[2]CNRS, Bordeaux INP, LaBRI, UMR5800, University of Bordeaux, Talence, France
[3]Sony AI, Zurich, Switzerland
[4]Northeastern University, Boston, Massachusetts, USA

Abstract

The use and creation of machine-learning-based solutions to solve problems or reduce their computational costs are becoming increasingly widespread in many domains. Deep Learning plays a large part in this growth. However, it has drawbacks such as a lack of explainability and behaving as a black-box model. During the last few years, Visual Analytics has provided several proposals to cope with these drawbacks, supporting the emerging eXplainable Deep Learning field. This survey aims to (i) systematically report the contributions of Visual Analytics for eXplainable Deep Learning; (ii) spot gaps and challenges; (iii) serve as an anthology of visual analytical solutions ready to be exploited and put into operation by the Deep Learning community (architects, trainers and end users) and (iv) prove the degree of maturity, ease of integration and results for specific domains. The survey concludes by identifying future research challenges and bridging activities that are helpful to strengthen the role of Visual Analytics as effective support for eXplainable Deep Learning and to foster the adoption of Visual Analytics solutions in the eXplainable Deep Learning community. An interactive explorable version of this survey is available online at https://aware-diag-sapienza.github.io/VA4XDL.

Keywords: deep learning, explainable artificial intelligence, interpretability, neural networks, visual analytics, visualization

CCS Concepts: • General and reference → Surveys and overviews; • Human-centred computing → Visual analytics; • Computing methodologies → Neural networks; Artificial intelligence

State of the Art report in *Computer Graphics Forum*

67 VA systems for XDL, 2015–2022

each paper read by ≥2 authors and coded on **38 dimensions**

interactive explorer: **aware-diag-sapienza.github.io/VA4XDL**

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733

Our map: La Rosa et al., 2023

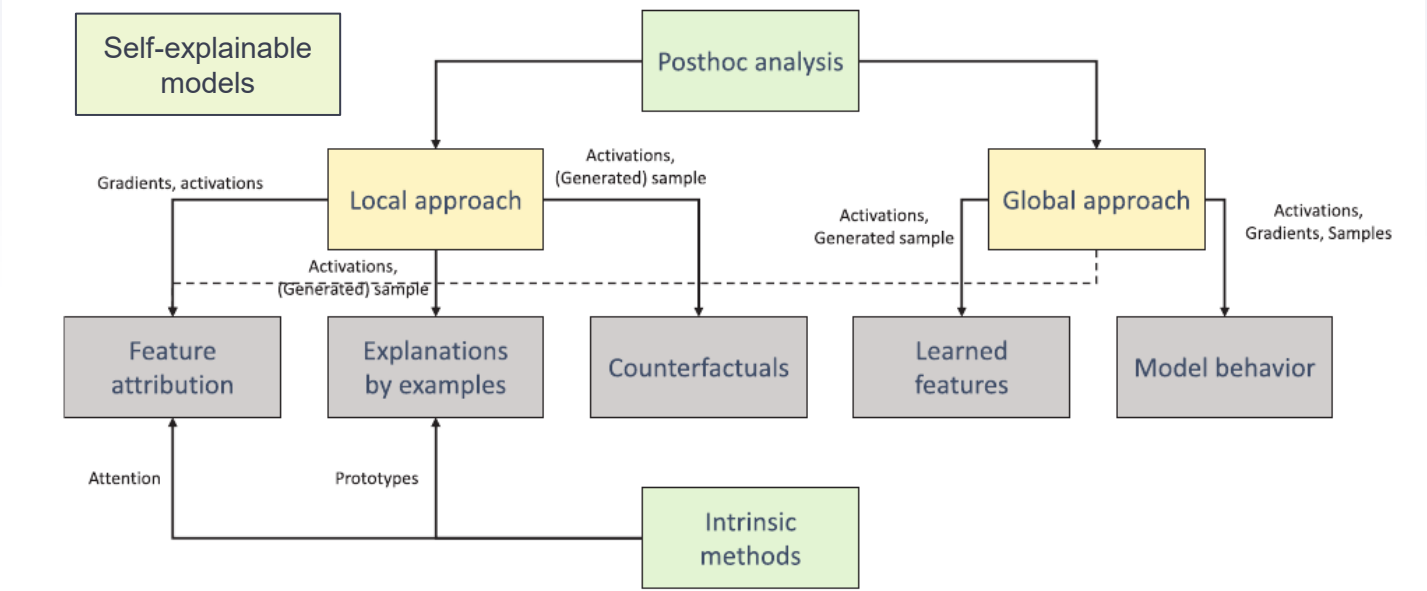


Why **this** taxonomy? It is built from what **VA systems actually support.** A **real VA system**, a **real DNN**, and the goal is to **explain** it. Performance monitoring is **out**…
we'll come back to it..

interactive explorer: **aware-diag-sapienza.github.io/VA4XDL**

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733

La Rosa et al. taxonomy



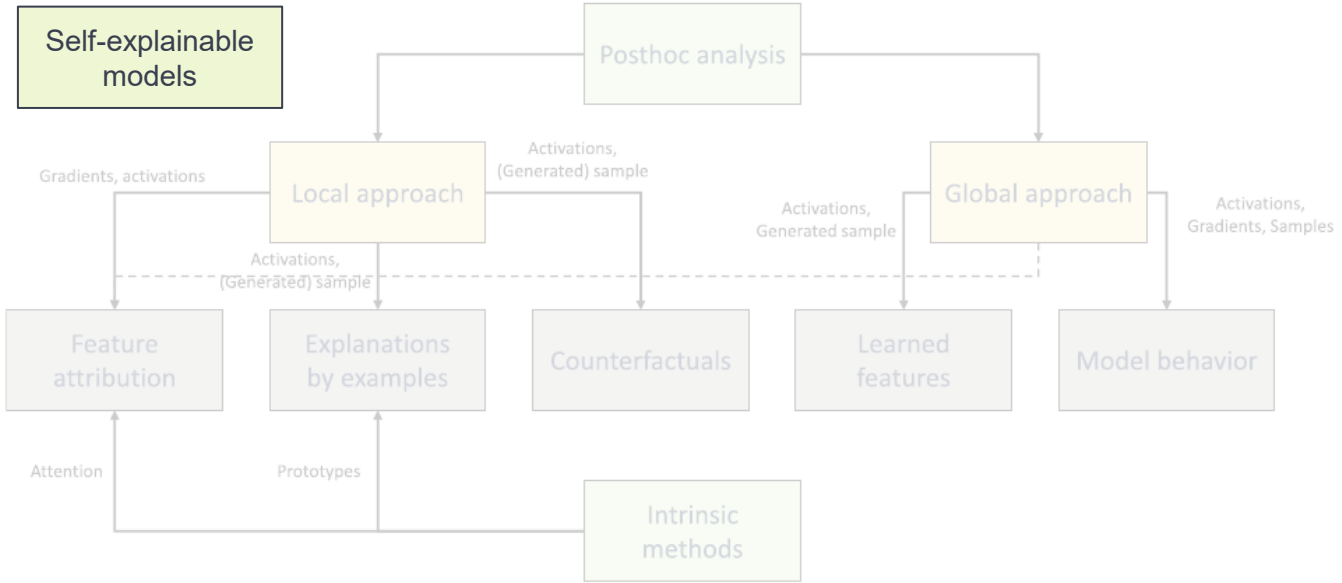
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§4.1, Fig. 6)

🧶 THREAD 3 · SELF-EXPLAINABLE MODELS

Models that **carry the explanation inside** (attention, prototypes, concept bottlenecks… etc). What matters for VA is **steering**: editing prototypes or running what-if analyses, exactly what La Rosa finds missing.

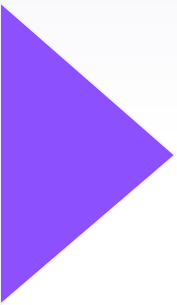
- Which **families** exist, and where do they fall?
- Is **attention** an explanation?
- Does the **accuracy–interpretability trade-off** really exist in DL?
- How do we check that a **prototype** or a **concept** means something?
- What can VA do with an **intrinsic** model that it cannot do with a **post hoc** one?

La Rosa et al. taxonomy



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§4.1, Fig. 6)

Five questions





Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

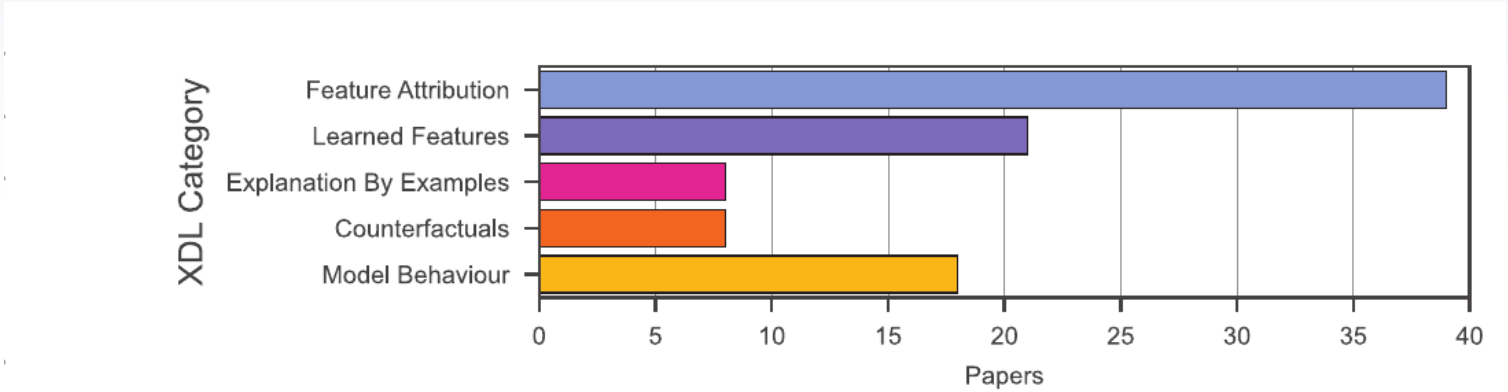
What must change to get another outcome?



Model behaviour

How does it react in a given situation?

Which one dominates?



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (Fig. 5)

Five questions



Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

What must change to get another outcome?



Model behaviour

How does it react in a given situation?



WHERE IS THE MODEL **FOCUSING** FOR **THIS** PREDICTION?

FA: what is it

What: a **score per input feature** for its impact on the prediction

Why: shows which inputs drive a decision, so shortcuts like the copyright tag become visible



Feature attribution

Where is the model focusing?

Local | Global

Post hoc | Self-explainable

How (1/2)



Gradients

How much does the output change if this input changes a tiny bit? ∂o/∂x

Saliency maps (Simonyan et al., 2014): raw gradient magnitude

Integrated Gradients (Sundararajan et al., 2017): gradients averaged on a path from a baseline

Grad-CAM (Selvaraju et al., 2017): gradients × feature maps of the last conv layer

Propagation

Redistribute the prediction backwards, layer by layer, with conservation rules

LRP (Bach et al., 2015): relevance is conserved from layer to layer

DeepLIFT (Shrikumar et al., 2017): contributions relative to a reference input

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319–3328).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), e0130140.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3145–3153).

How (2/2)

🔀 Perturbation & surrogates

Change the input and watch the output: works on any model (black-box)

Occlusion (Zeiler & Fergus, 2014): slide a grey patch over the image

LIME (Ribeiro et al., 2016): local linear surrogate fitted on perturbed samples

SHAP (Lundberg & Lee, 2017): Shapley values, the fair share of each feature

👁️ 🏛️ Attention (self-explainable)

Read the weights the model already computes during inference

Attention weights (Bahdanau et al., 2015)

Attention rollout / flow (Abnar & Zuidema, 2020): combine the weights across layers

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Springer.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135–1144).
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.
Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the ACL* (pp. 4190–4197).

Gradients vs propagation

Gradients

How sensitive is the score to this pixel?

A **slope**: what if it changed a bit

Saliency, Integrated Gradients, Grad-CAM

Propagation

How much of the score comes from this pixel?

A **share** of the score

LRP, DeepLIFT



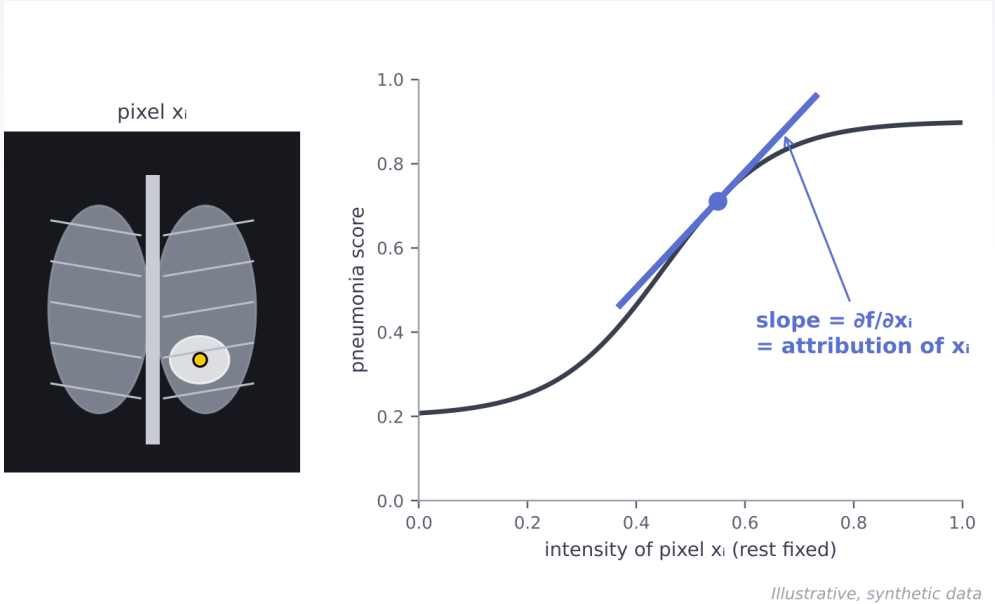
Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), e0130140.

Gradients: nudge a pixel

One backward pass of the **derivative** $\partial f/\partial x_i$

A **local slope**: what if this pixel changed **a tiny bit**?

The same backprop used in **training**



Illustrative, synthetic data

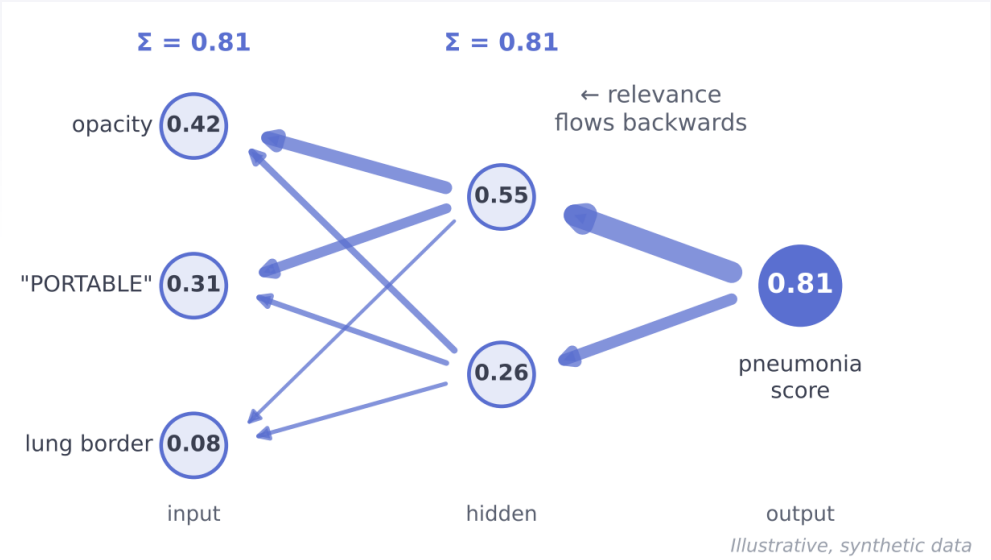
Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.
Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319–3328).

Propagation: share out the score

Start from the **score** (0.81) and **split** it backwards, layer by layer

Each neuron gets relevance **in proportion** to its contribution

Conservation: every layer adds up to **0.81**



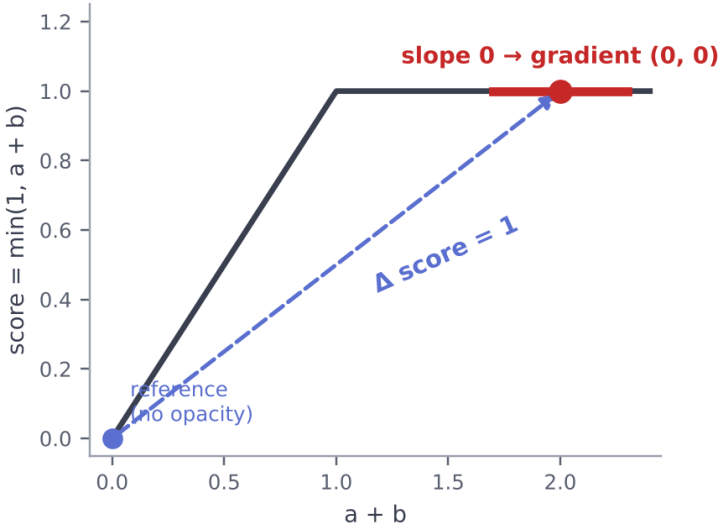
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), e0130140.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3145–3153).

When gradients fail: saturation

two opacities:
each one is enough





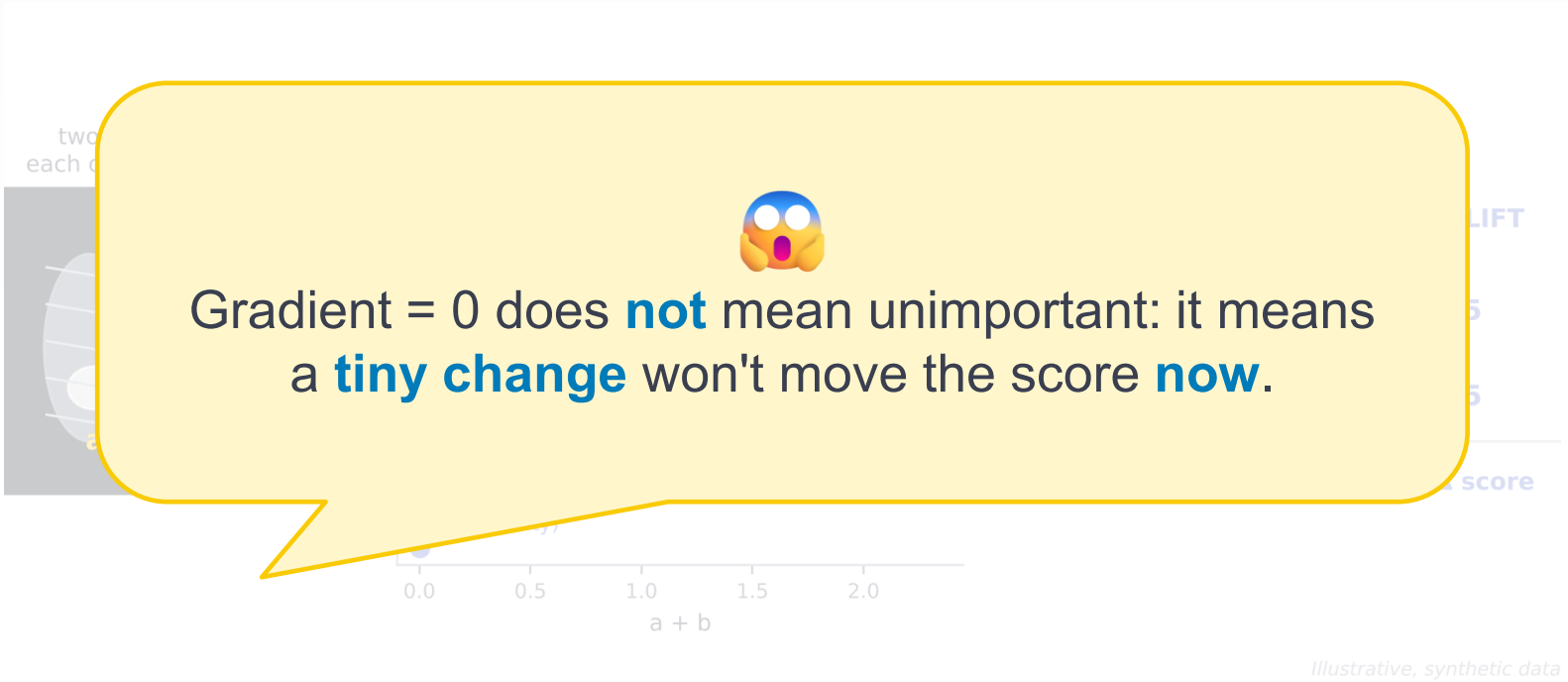
a = 1, b = 1

	Gradient	DeepLIFT
opacity a	0	0.5
opacity b	0	0.5
sum	0	1 = Δ score

Illustrative, synthetic data

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3145–3153).

When gradients fail: saturation



Gradient = 0 does **not** mean unimportant: it means a **tiny change** won't move the score **now**.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3145–3153).

Side by side

	Gradients	Propagation
Question	What if this pixel **changed**?	How much of the score is **its share**?
Computes	A **derivative** $\partial f/\partial x_i$	A **redistribution** of the score
Adds up to the score?	No	**Yes** (conservation)
Needs	Any **differentiable** model	**Rules** per layer type
Typical trap	**Saturation**, noisy maps	Results depend on the **rule** and reference

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), e0130140.
Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3145–3153).

Side by side

	Gradients	Propagation
Questi		share?
Compu		ore
Adds u		
Needs		
Typical trap	**Saturation**, noisy maps	Results depend on the **rule** and reference

In some cases both rules and gradients give the same results. Both categories are fuzzy.

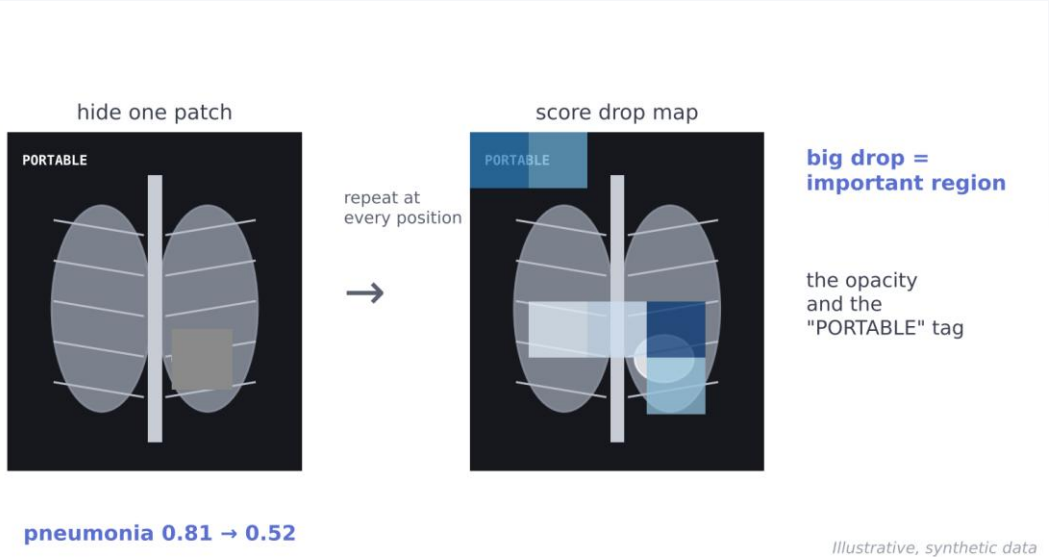
Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations (ICLR)*.

Perturbation: poke and watch

Change part of the input, **watch** the score

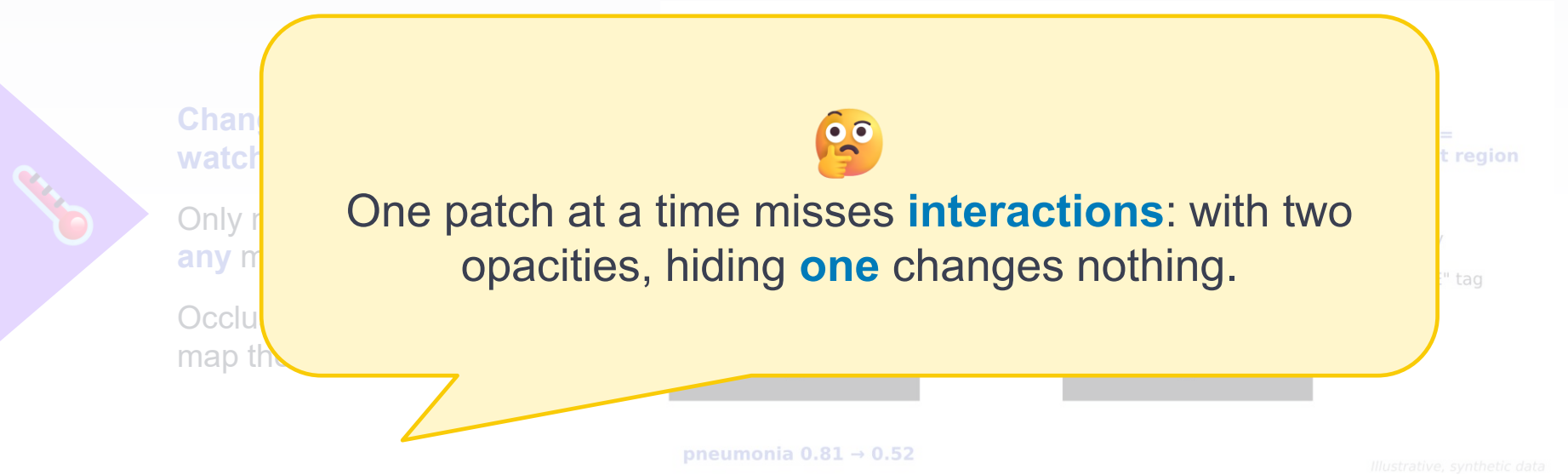
Only needs **predict()**: works on **any** model

Occlusion: slide a **grey patch**, map the **drop**



Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Springer.

Perturbation: poke and watch



🤔

One patch at a time misses **interactions**: with two opacities, hiding **one** changes nothing.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Springer.

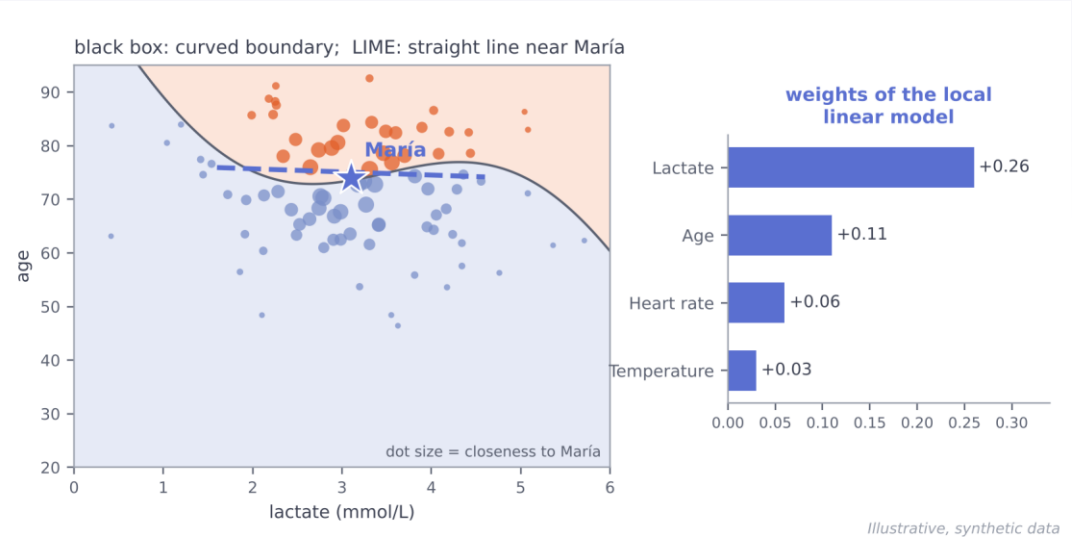
LIME: a simple local model

① **Perturb** María: many variants

② Ask the **black box** for each one

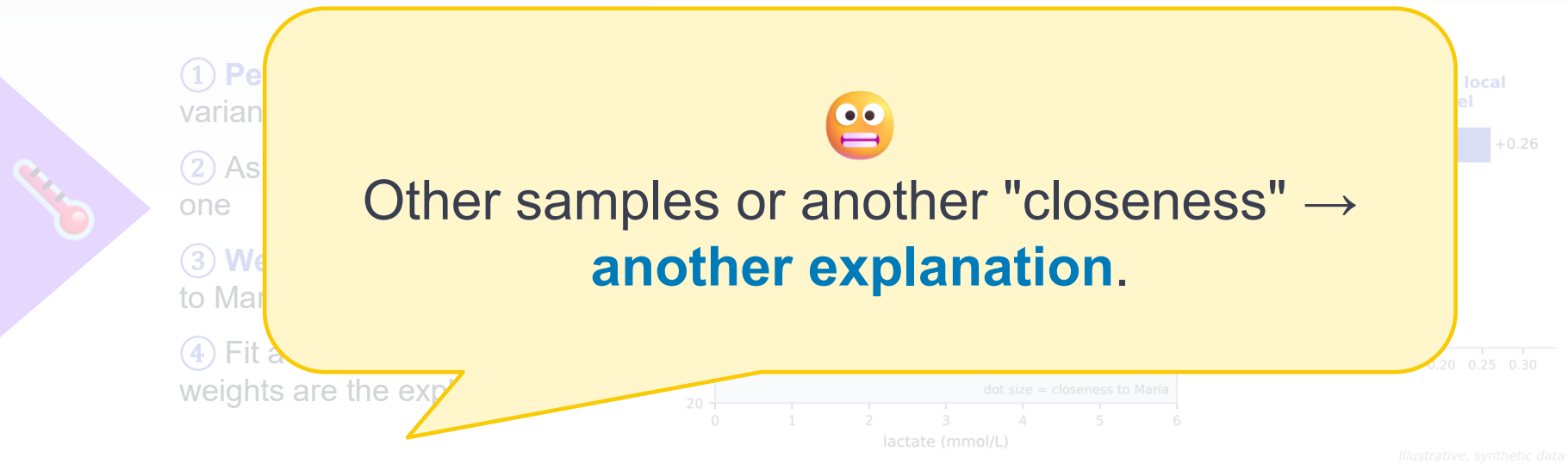
③ **Weight** them by closeness to María

④ Fit a **linear model**: its weights are the explanation



Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135–1144).

LIME: a simple local model



Other samples or another "closeness" → **another explanation**.

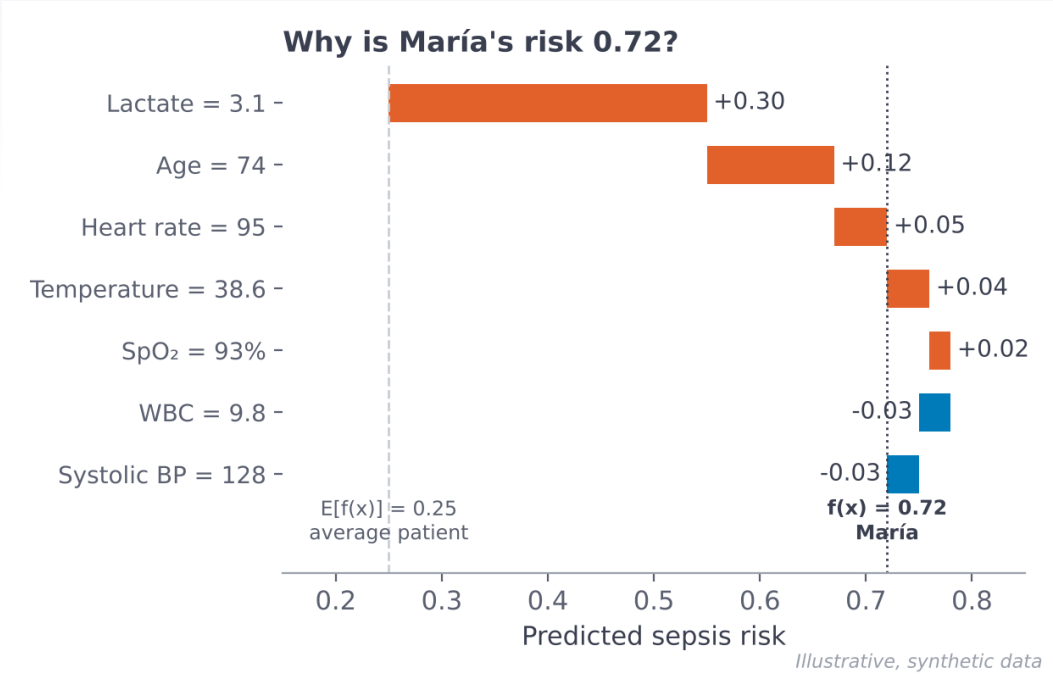
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135–1144).

SHAP: a fair share per feature

Features are **players**, the risk is the **prize**

Each gets its **average** extra contribution over all **orders**

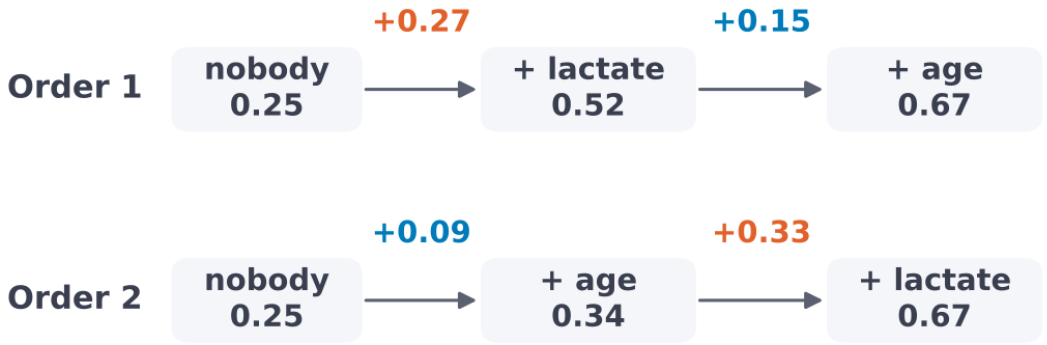
Shares **add up**: 0.25 + … = **0.72**



Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Computing a Shapley value

Only two features, to keep it small. "Absent" = filled with average-patient values



lactate: (0.27 + 0.33) / 2
= 0.30

age: (0.15 + 0.09) / 2
= 0.12

0.25 + 0.30 + 0.12 = 0.67 ✓

Same feature, different extra value depending on who came first → take the average over all orders

Illustrative, synthetic data

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

LIME vs SHAP

	🍋 LIME	🎲 SHAP
Idea	A **local linear** surrogate	A **fair share** (game theory)
Returns	Weights of the linear model	Contributions per feature
Adds up to the prediction?	No	**Yes**: base + shares = f(x)
Stability	Changes with **sampling** and kernel	**Unique** values (given the background)
Cost	Moderate	**Exponential**: approximated

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135–1144).
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

LIME vs SHAP

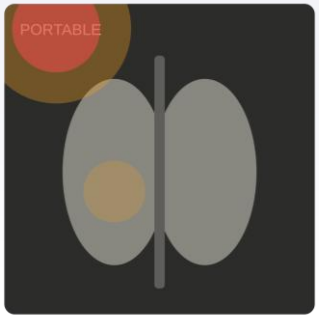
Idea		
Return		
Adds u… predict…		
Stabilit…		…ground)
Cost	Moderate	**Exponential**: approximated



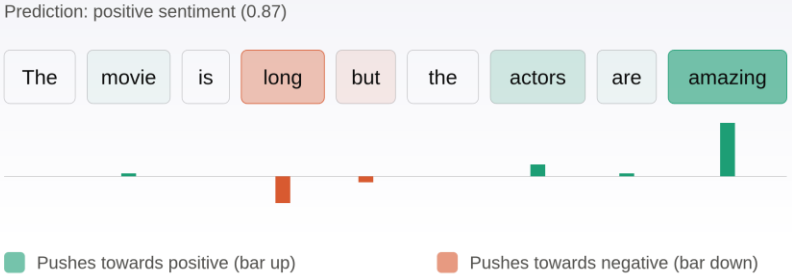
KernelSHAP is LIME with a special weighting: the right kernel turns LIME's weights into **Shapley values**.

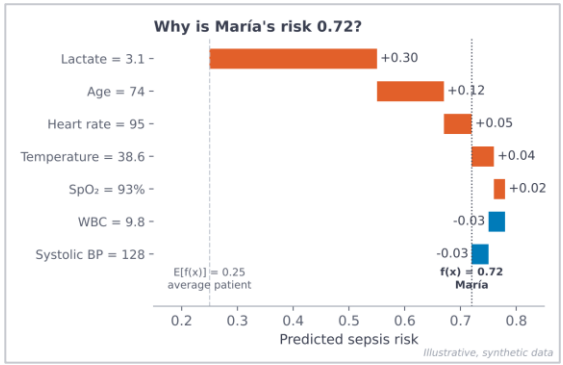
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135–1144).
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

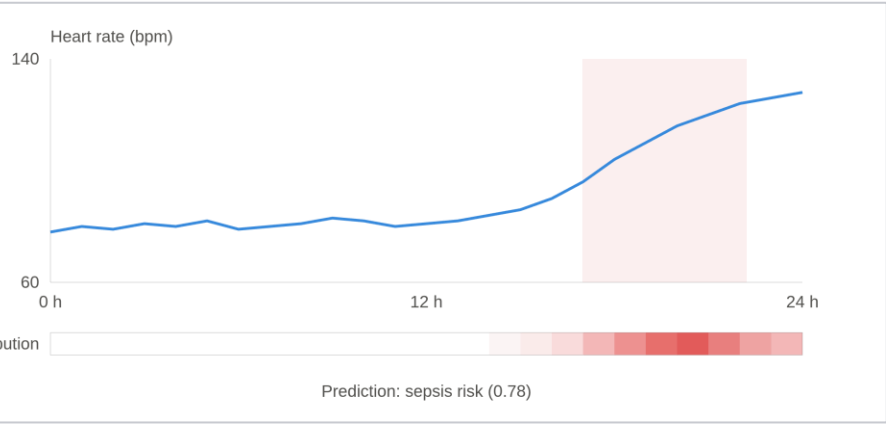
FA: Looks like (1/2)



Saliency map (Grad-CAM)
Prediction: pneumonia (0.91)







Five questions



Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

What must change to get another outcome?



Model behaviour

How does it react in a given situation?

🧶 THREAD 4 · FEATURE ATTRIBUTION

The most mature category… and the most treacherous. Dozens of methods that disagree, some fail sanity checks, and most VA systems hard-code just one.

- **Which method?** Which families (gradients, perturbations, attention… etc) respect your data's structure?
- **Can we trust it?** Which sanity checks and benchmarks exist, and which methods pass them?
- **Beyond heatmaps:** how do you show attribution on a series, a graph or a text?
- **Aggregate:** from a thousand local explanations to one global strategy.
- **Disagreement:** two methods, two answers. What do you show?
- **Generative:** the output is an image or a text, not a class. Which input (or training sample) produced which part of the output?



WHAT HAS THE MODEL **LEARNED** DURING TRAINING?

LF: what is it

What: associate **concepts** with neurons, groups of neurons or layers

Why: checks that the network learned meaningful concepts, not artefacts of the data

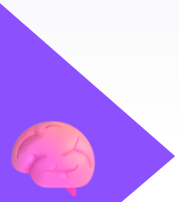


Learned features

What has it learned?

Local	Global
Post hoc	Self-explainable

How (1/2)



⚡ Activation maximization

Synthesize the input that most excites a neuron

Erhan et al., 2009

Feature visualization (Olah et al., 2017)

⚡🔍 Top-activating examples

Search the dataset for the inputs that activate it most

Zeiler & Fergus, 2014

⚡ Concept probing

Test a neuron or a direction against human-labelled concepts

Network Dissection (Bau et al., 2017)

TCAV (Kim et al., 2018)

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network* (Technical Report 1341). University of Montreal.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Springer.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE CVPR* (pp. 6541–6549).

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th ICML* (pp. 2668–2677).

How (2/2)

⚡ Clustering & projection

Group neurons or activations and project them to 2D

t-SNE (van der Maaten & Hinton, 2008)
UMAP (McInnes et al., 2018)

💬 Describe in words

Generate a natural-language label for each neuron

MILAN (Hernandez et al., 2022): next slide



📈 Activation inspection

Plot a unit's activation directly over the input and look for patterns

LSTMVis (Strobelt et al., 2018)
RNNVis (Ming et al., 2017)
Karpathy et al., 2016: character-level LSTM cells that track quotes and line length

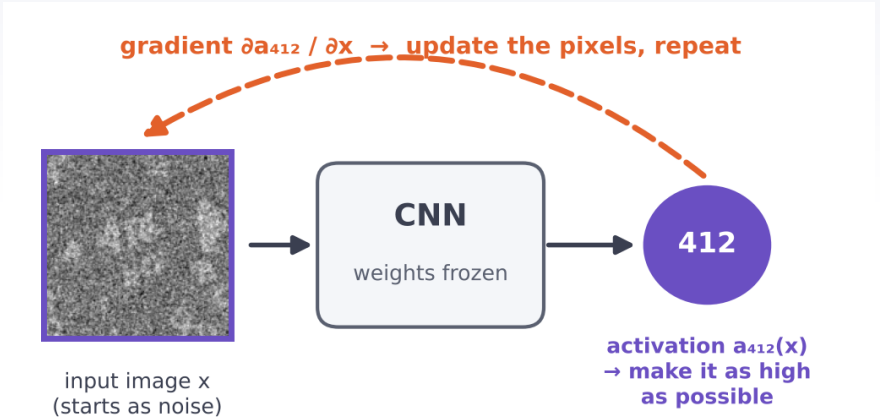
van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.
McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.
Strobelt, H., Gehrmann, S., Pfister, H., & Rush, A. M. (2018). LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 667–676.
Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., & Qu, H. (2017). Understanding hidden memories of recurrent neural networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 13–24).
Karpathy, A., Johnson, J., & Fei-Fei, L. (2016). Visualizing and understanding recurrent networks. In *International Conference on Learning Representations, Workshop Track*.

Activation maximization: the idea

Training: fix the image, **change the weights**

Here: fix the weights, **change the image**

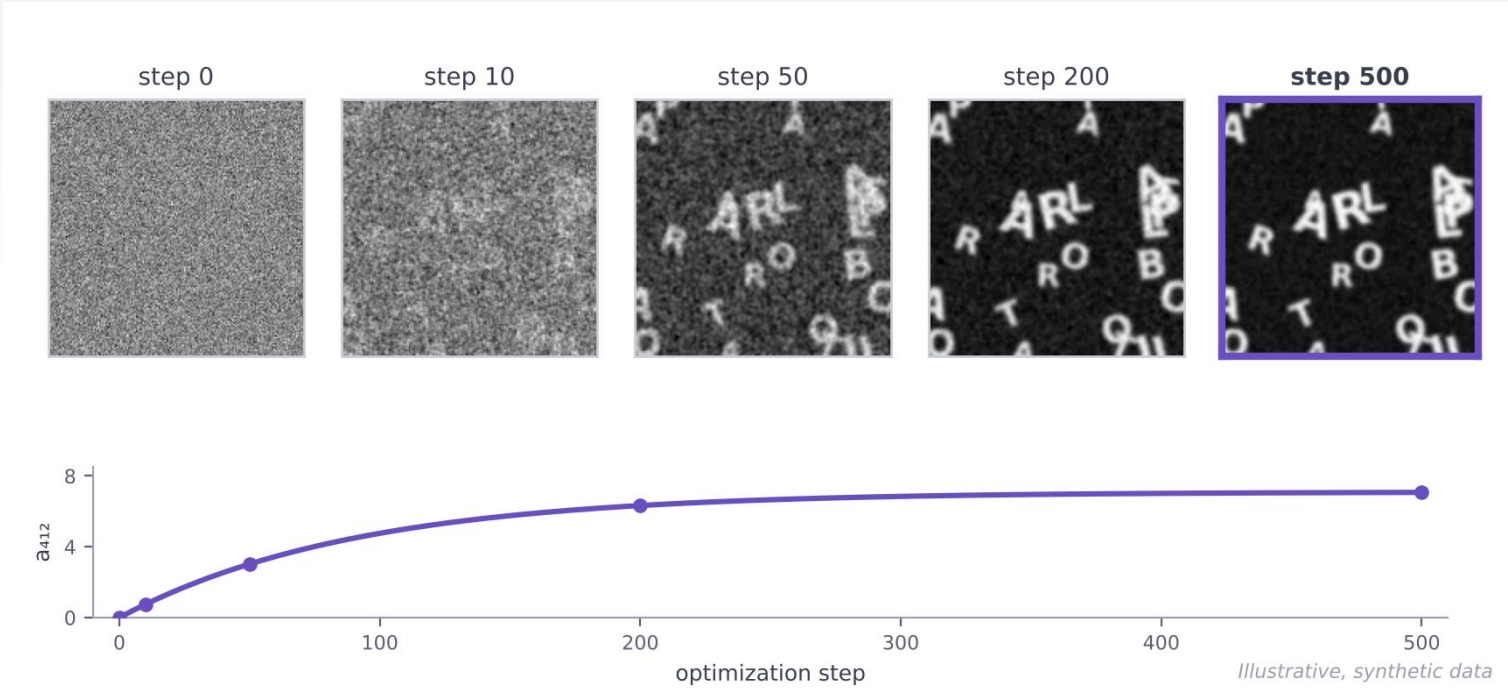
Start from **noise** and follow the gradient so that **neuron 412** fires as much as possible



$$x^* = \text{argmax}_x \; a_{412}(x) - \lambda \cdot R(x)$$

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network* (Technical Report 1341). Université de Montréal.
Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Neuron 412, step by step

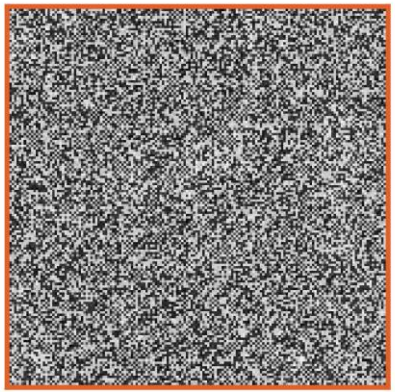


Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Regularization

No regularization

Pixels optimized freely: high-frequency noise



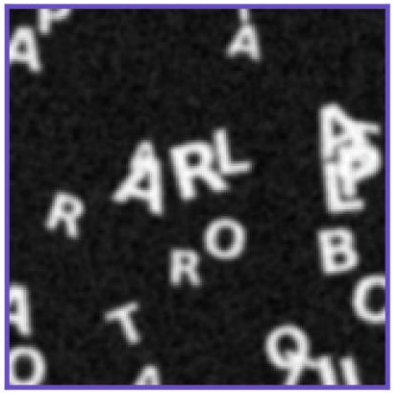
activation 9.8

Very high activation…
but nothing a human
can read

Illustrative, synthetic data

With jitter + blur

Shift and smooth the image at every step



activation 7.1

Lower activation,
but readable:
"printed letters"

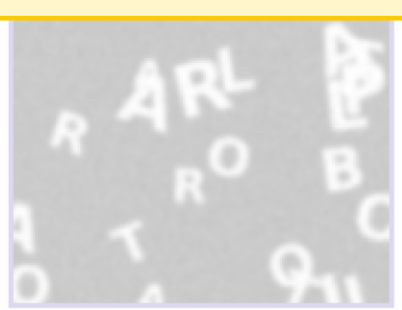
Illustrative, synthetic data

Regularization

No regularization

Pixels optimized freely: high-frequency noise





Lower activation,
but readable:
"printed letters"

Illustrative, synthetic data

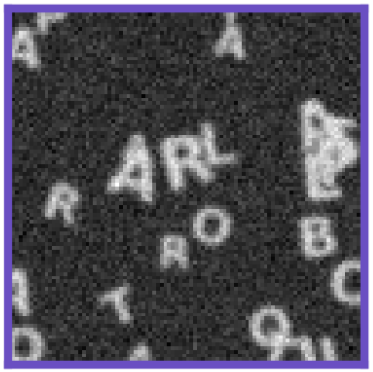
💡

Jitter and **blur** keep the dream natural: only what the neuron **really detects** survives.

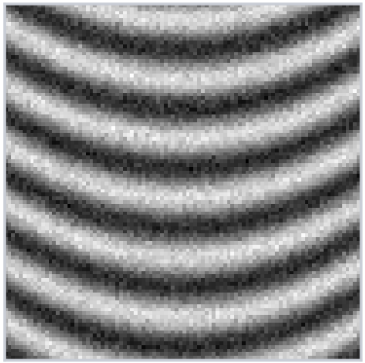
Reading the dreams

Activation maximization: the input each neuron "dreams" of

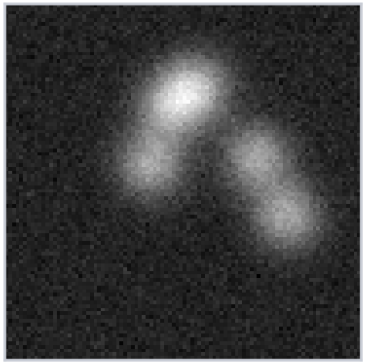
Start from noise, change the input to maximize one neuron (layer 4), then a human names the pattern



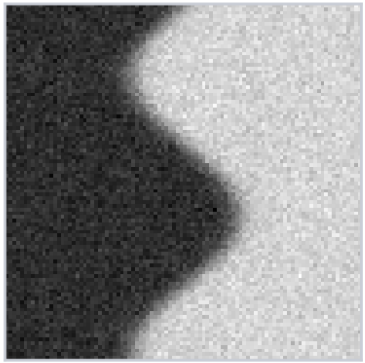
Neuron 412
"printed letters"



Neuron 87
"ribs / stripes"



Neuron 205
"round opacity"



Neuron 33
"lung border"

Illustrative, synthetic data

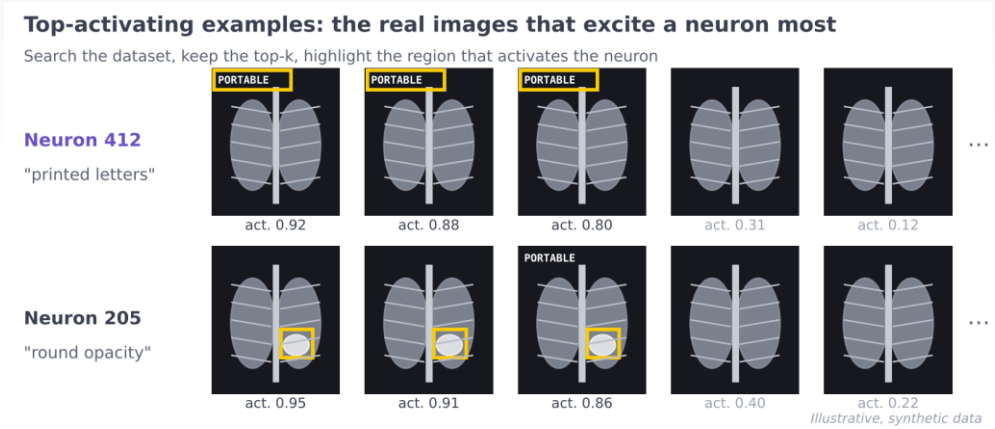
Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Check it on real images

A dream can **mislead**: compare it with the **top-activating** X-rays

All of neuron 412's top images carry the **"PORTABLE" tag**

Dream and real images **agree**: the neuron reads text



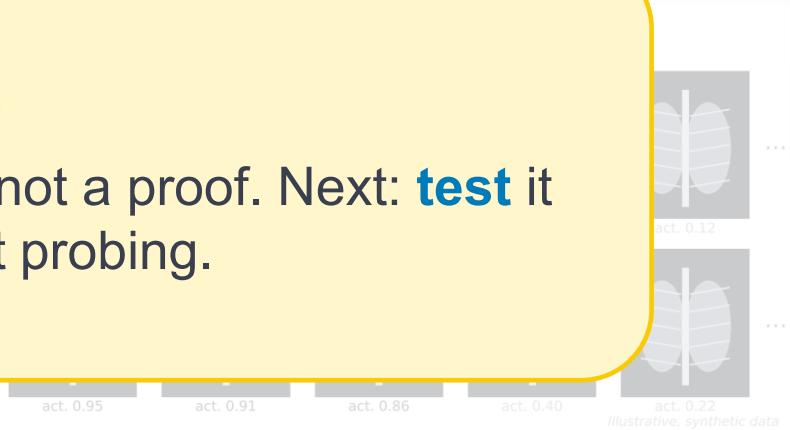
Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network* (Technical Report 1341). Université de Montréal.
Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Check it on real images

A drea
with th

All of t
carry

Dream
neuron



🤔

A dream is a **hypothesis**, not a proof. Next: **test** it with concept probing.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network* (Technical Report 1341). Université de Montréal.
Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007

Concept Probing: TCAV in 4 steps



① Concept examples

images with and without the concept

→

② Concept direction

a linear classifier gives the CAV

→

③ Sensitivity

does the class react to that direction?

→

④ Score + significance

how often, and is it chance?

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

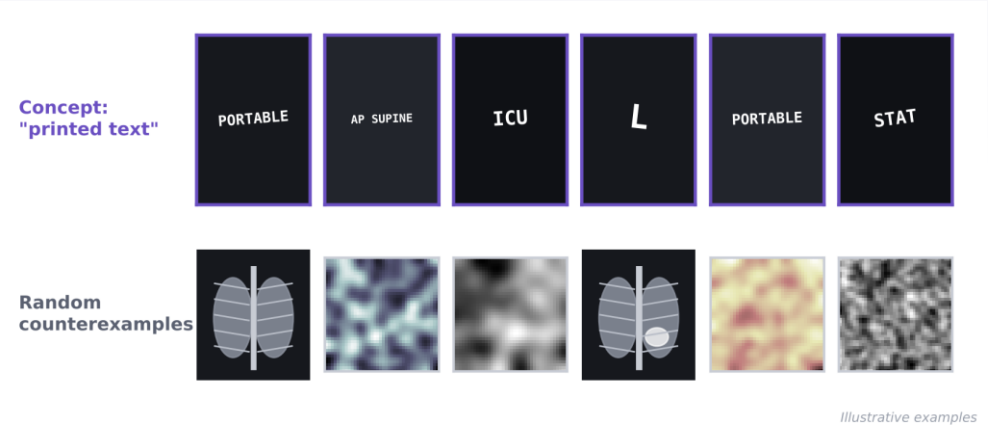
TCAV ①: concept examples

~50 images **with** the concept: printed text

~50 **random** images as counterexamples

They do **not** need to come from the training set

No class labels needed



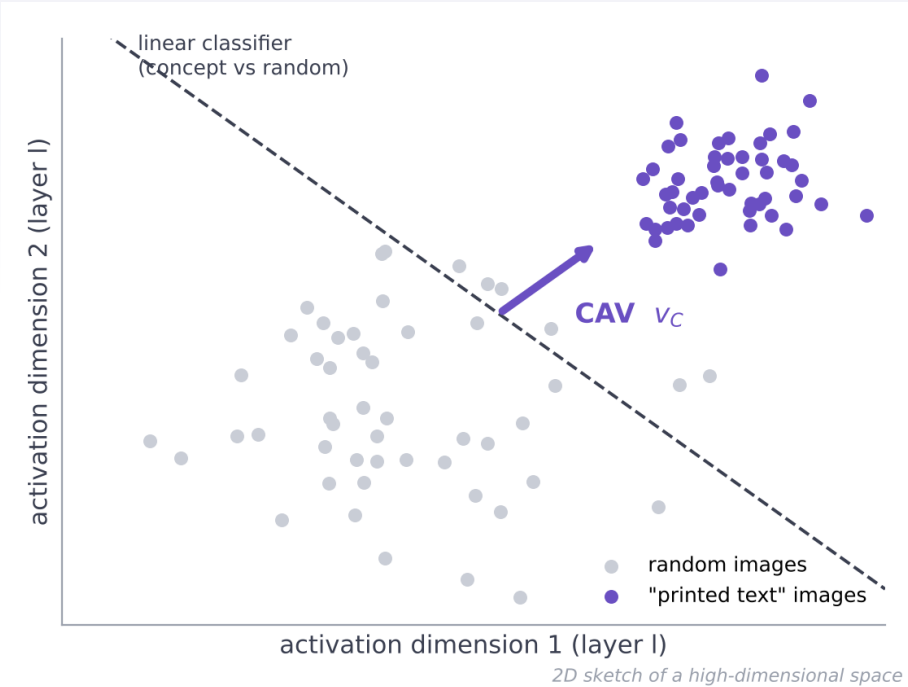
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV ②: the concept direction

Pass both sets through the network up to **layer l**

Train a **linear classifier**: concept vs random

The vector **orthogonal** to its boundary is the **CAV** v_C



2D sketch of a high-dimensional space

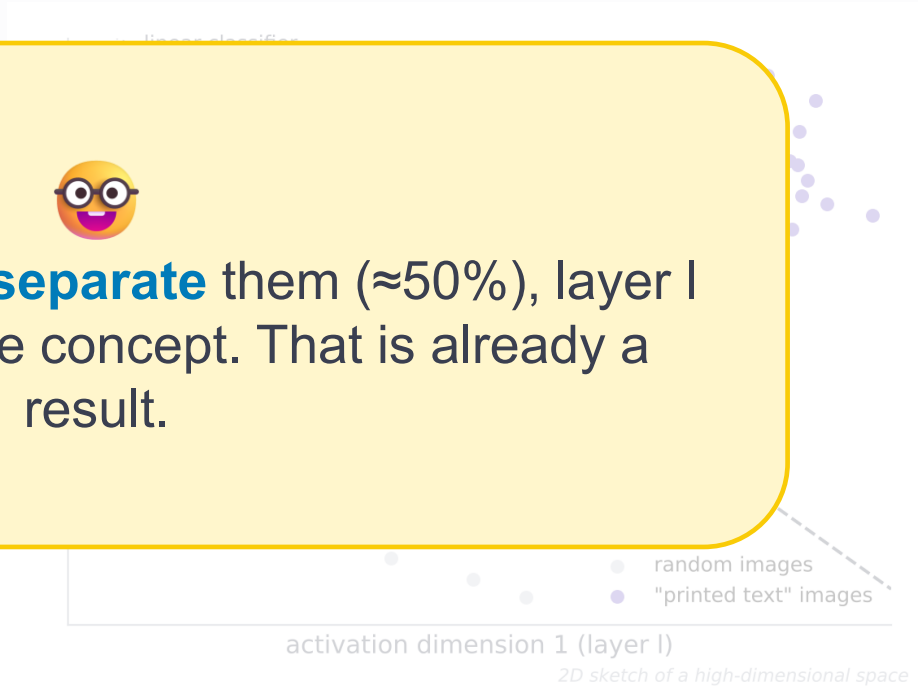
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV ②: the concept direction

Pass
to lay

Train
rando

The v
is the



If the classifier **can't separate** them (≈50%), layer l does not encode the concept. That is already a result.

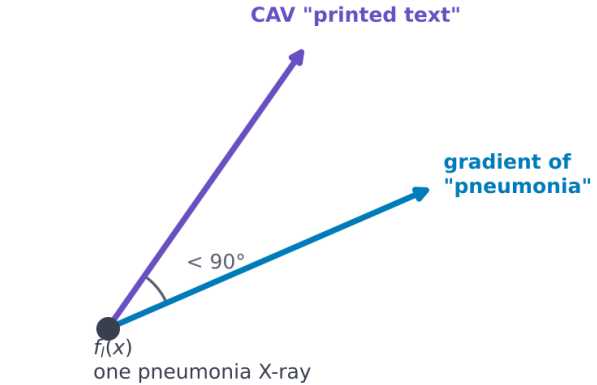
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV ③: sensitivity

$$S(x) = \nabla h_{pneumonia}(f_l(x)) \cdot v_C$$

∇h(f_l(x)): how the **pneumonia** score changes at layer l

v_C: the direction **"more printed text"**

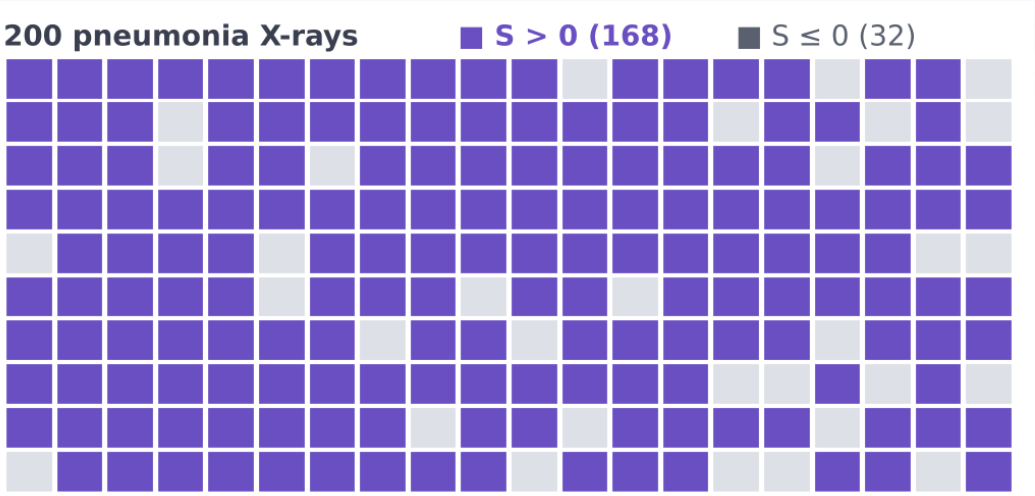


S > 0 → pushing towards "more text" **raises** the pneumonia score

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV ④: the score

TCAV score = fraction of pneumonia X-rays with $S(x) > 0$



168 / 200

= 0.84

It counts in how many cases the score goes up, not by how much

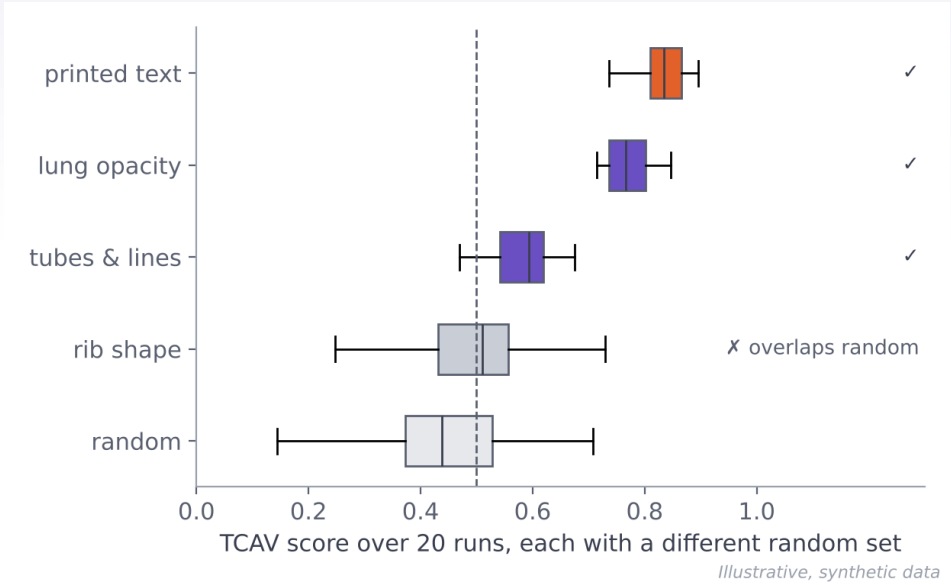
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV: Is it chance?

Repeat with **many random sets** → many CAVs

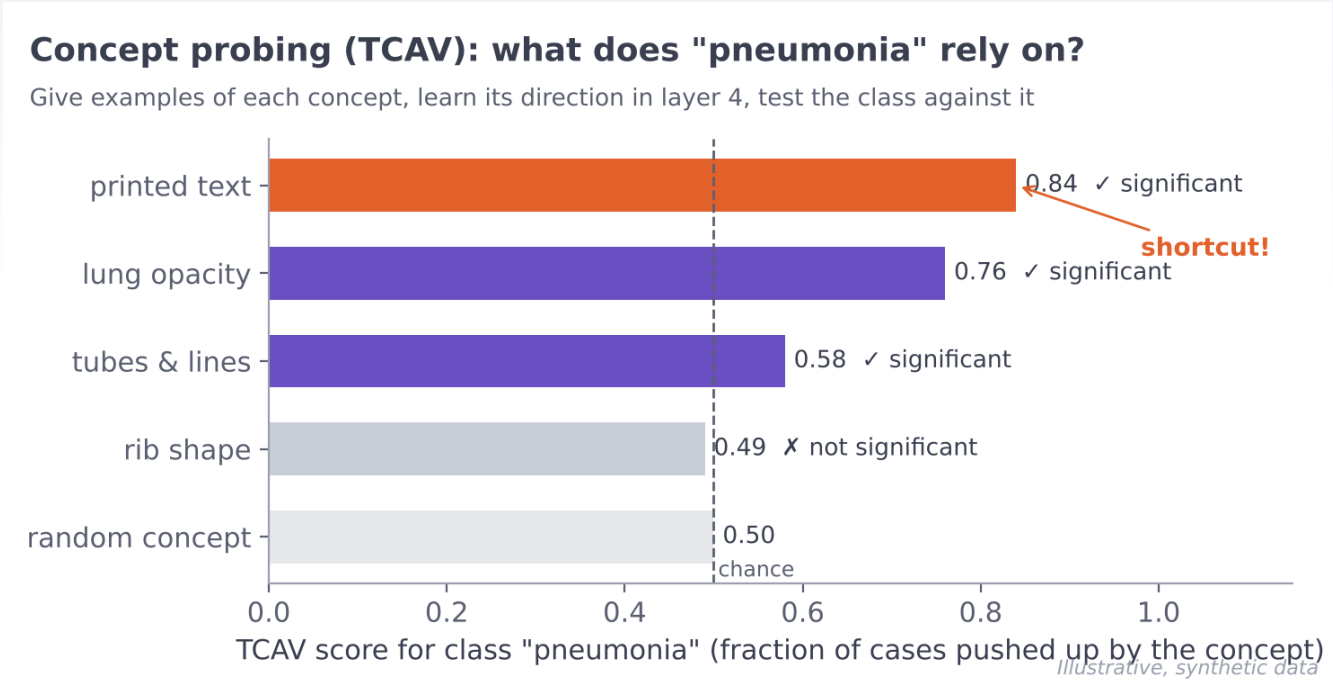
Compare with **random vs random**: around 0.5

Keep the concept only if it differs **significantly** (t-test)



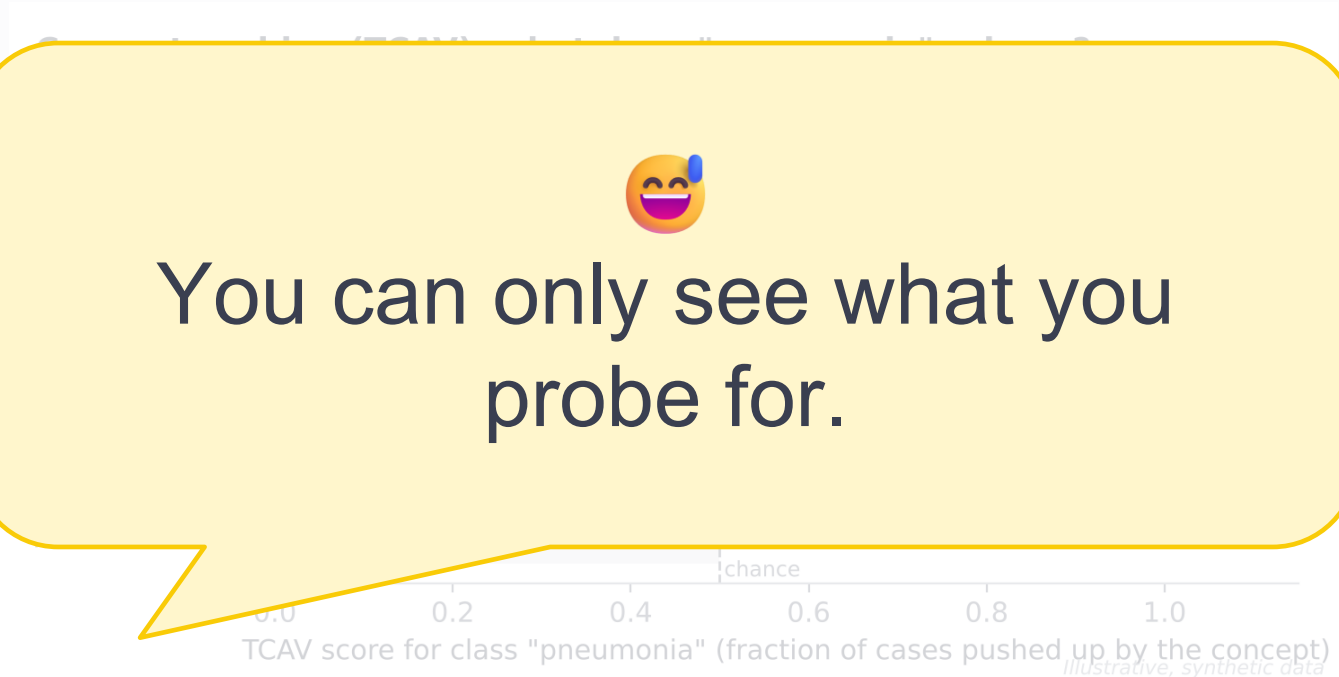
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

TCAV: the result



Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

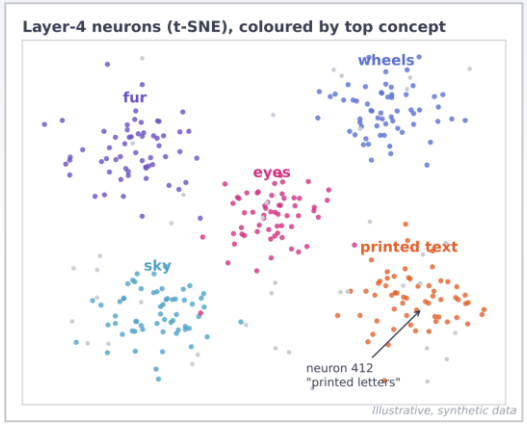
TCAV: the result

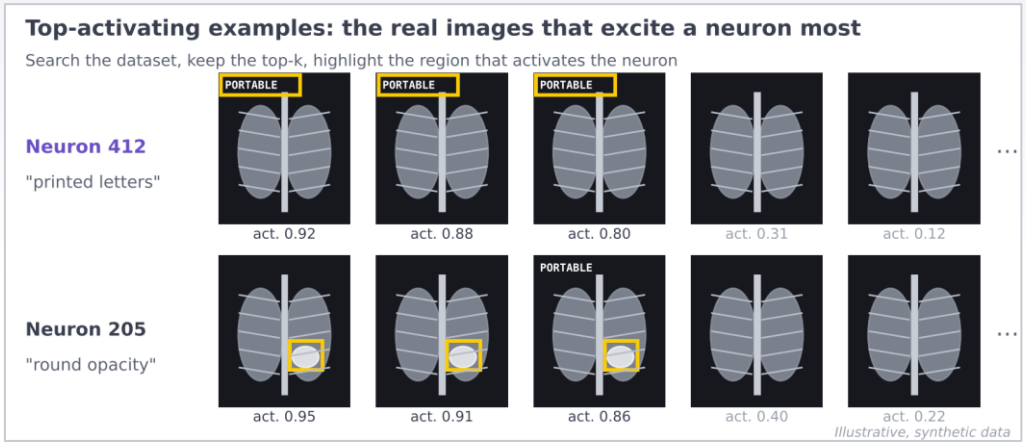


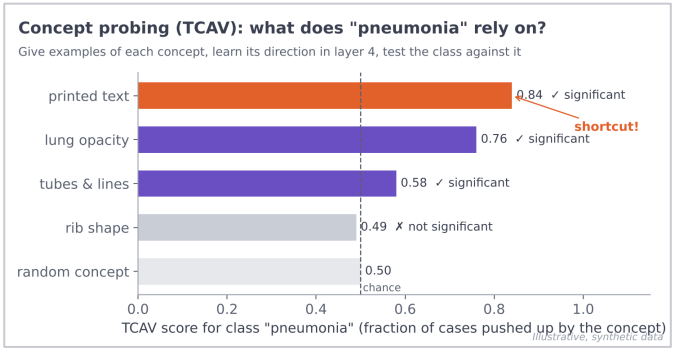
You can only see what you probe for.

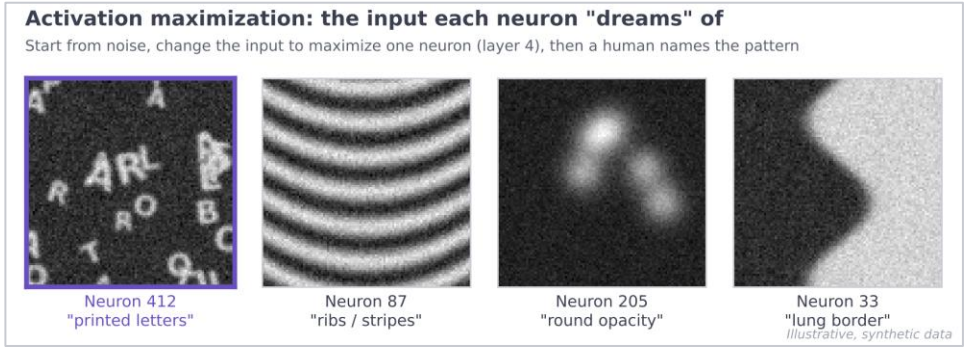
Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677).

LF: Looks like (1/2)

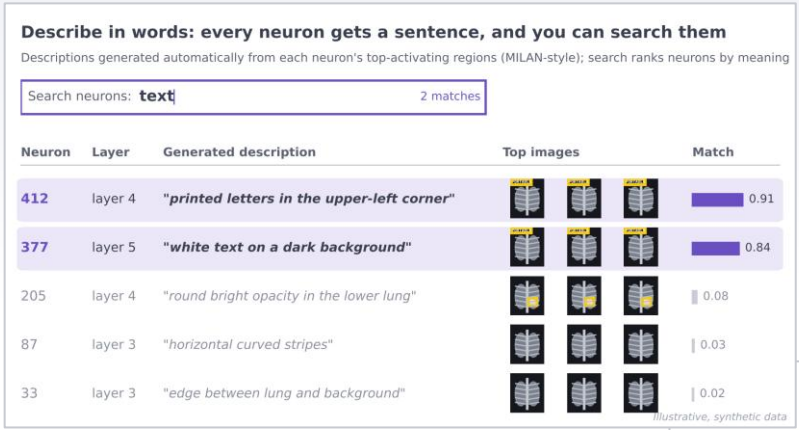


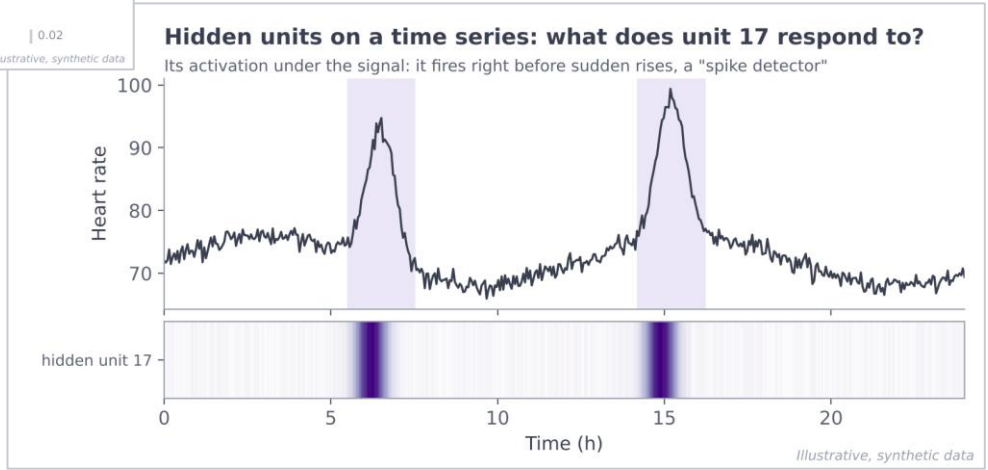






LF: Looks like (2/2)





🧶 THREAD 5 · LEARNED REPRESENTATIONS

In DL, features are learned and live in the latent space. We look at it by projecting to 2D (in MB, almost always t-SNE), yet La Rosa finds dimensionality reduction used off the shelf, not adapted to DL.

- **Taxonomy:** linear (PCA), distance (MDS, Isomap), neighborhood (t-SNE, UMAP), parametric (parametric UMAP, autoencoders). Local or global structure?
- **Reading:** effect of perplexity, n_neighbors, min_dist, init and seed? What must never be read from a projection, and how to report one?
- **Quality:** trustworthiness, continuity, neighborhood hit, stress, Shepard diagram. How to paint false and missing neighbors on the scatterplot?
- **What to project:** instances or neurons?
- **Scale:** from millions of points to something legible (sampling, density, Activation Atlas).
- **Comparison:** two spaces (after fine-tuning, client vs. global) without comparing the shapes of two drawings (Embedding Comparator).

Five questions



Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

What must change to get another outcome?



Model behaviour

How does it react in a given situation?



WHICH SAMPLES DOES THE MODEL CONSIDER **SIMILAR**?

EE: what is it

What: training samples used as the explanation

Why: people reason by analogy: examples need no technical vocabulary

Explanation by examples

Which samples does it consider similar?

Local | Global

Post hoc | Self-explainable

How

🔍 ⚡ Neighbours in latent space

Find the training samples whose representation is closest

Deep k-NN (Papernot & McDaniel, 2018)

🏛️ Prototypes (self-explainable)

The model classifies by comparing parts of the input with learned prototypes

ProtoPNet, "This looks like that" (Chen et al., 2019)

🧮 Influential training points

Which training samples changed this prediction most?

Influence functions (Koh & Liang, 2017)

Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th ICML* (pp. 1885–1894).

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49.

Similar vs responsible

🔍 Neighbours

Who looks like this case?

Distance in **latent space**

Same answer for any model that sees alike

🧮 Influential points

Who caused this prediction?

Gradients of the loss

Depends on **how the model learned**



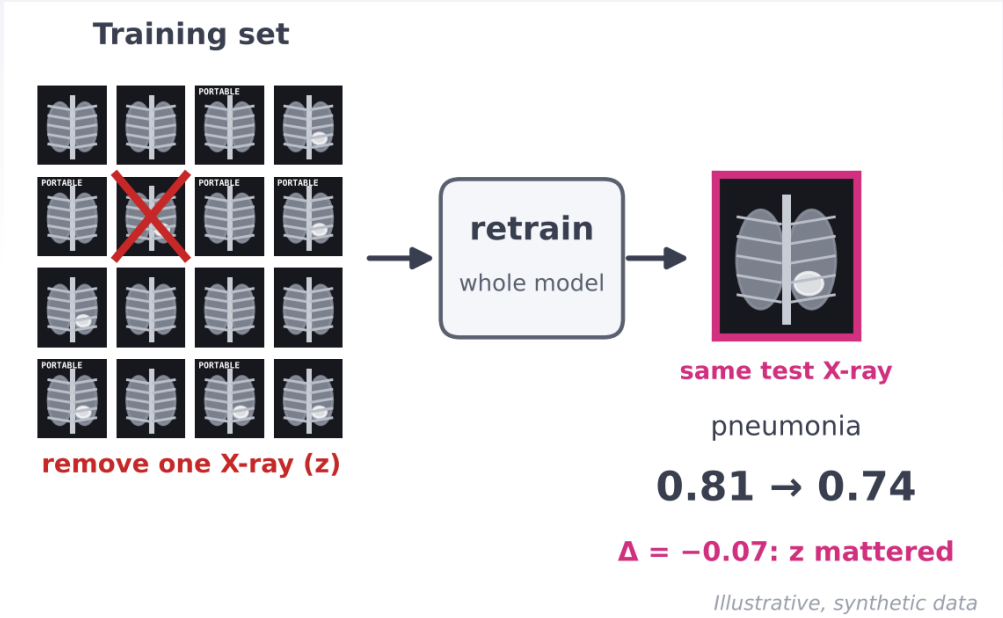
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Ideal experiment: leave one out

Remove **one** training X-ray

Retrain the model

Measure how the test prediction **changes**



Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Ideal experiment: leave one out

Training set

Remo

Retra

Meas
chan

ay

0.81 → 0.74

Δ = −0.07: z mattered

Illustrative, synthetic data



10,000 training images → **10,000 retrainings**.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Influence functions: no retraining

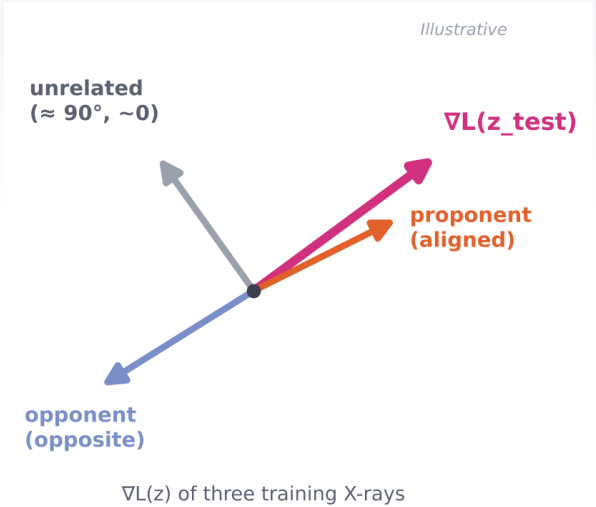
$$I(z, z_{test}) = -\nabla L(z_{test})^T \cdot H^{-1} \cdot \nabla L(z)$$

∇L(z_test): what matters for **this** prediction

H^{-1}: how the model would **readjust** (curvature)

∇L(z): what **z** taught the model

A **local** answer: computed for **this** test X-ray



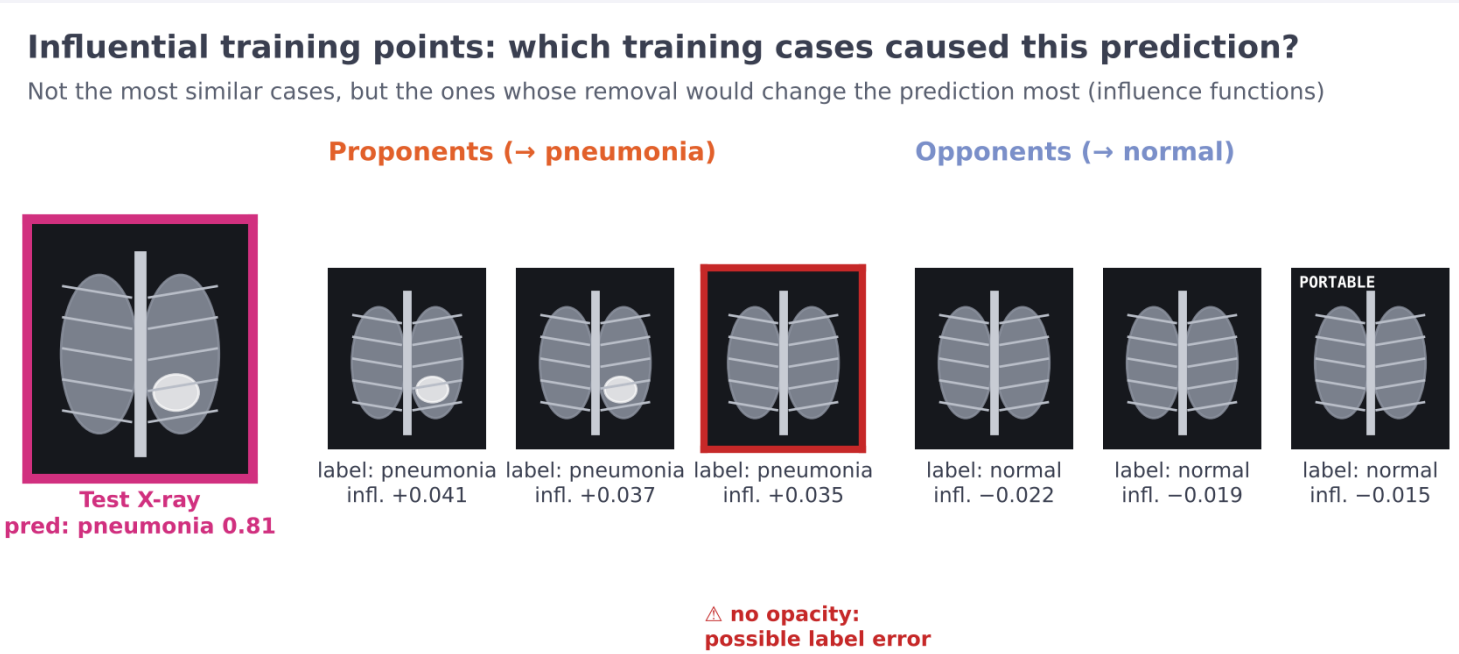
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Influence functions: no retraining

Gradients point the **same way** → that training case **pushed** this prediction. Estimated **without retraining**.

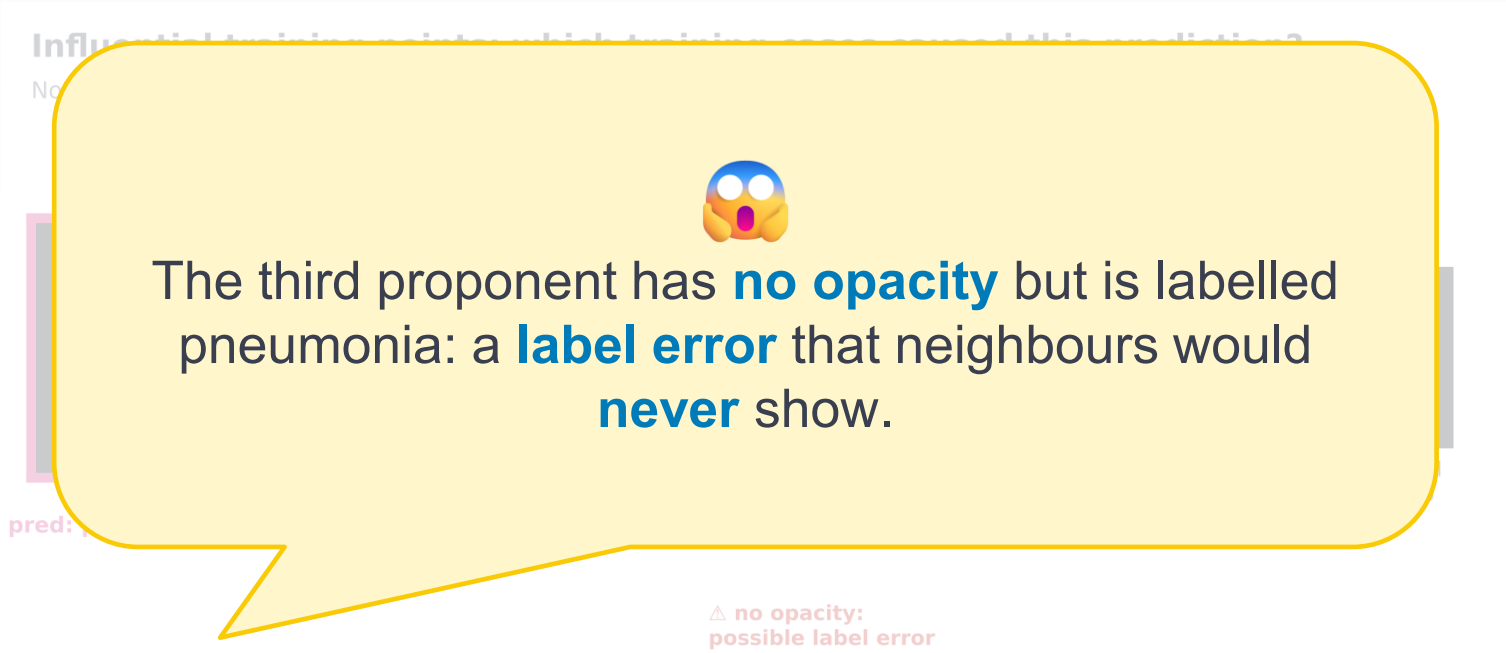
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Proponents and opponents



Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Proponents and opponents



The third proponent has **no opacity** but is labelled pneumonia: a **label error** that neighbours would **never** show.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

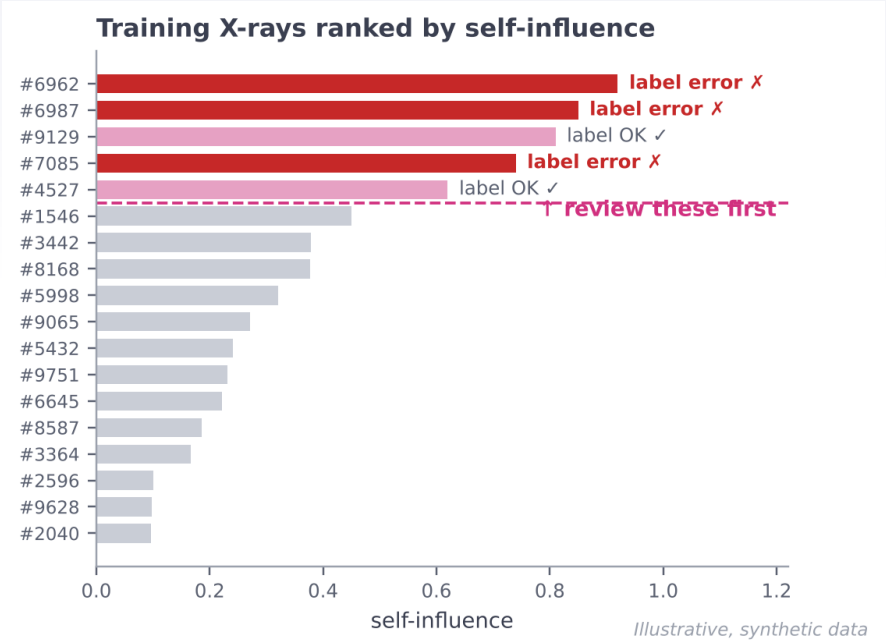
Hunting label errors

Rank **all** training cases by **self-influence**: how much each one props up its **own** prediction

The top of the list is full of **suspicious labels**

Review those first

Koh & Liang (2017, Sec. 5.4): spam emails, 10% flipped labels



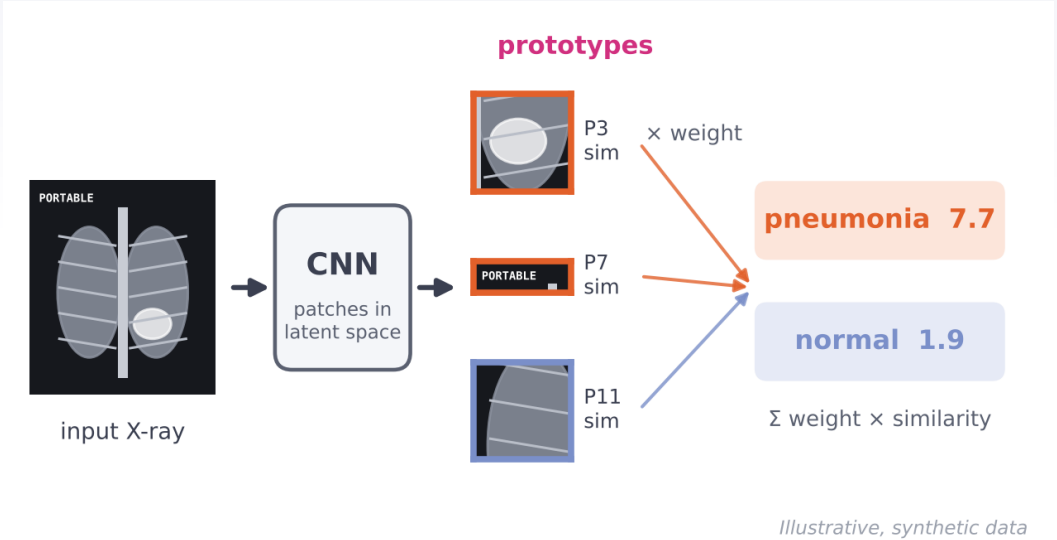
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1885–1894).

Prototypes: built-in explanations

The model **compares parts** of the X-ray with **prototypes** learned in training

The comparison **is** the decision

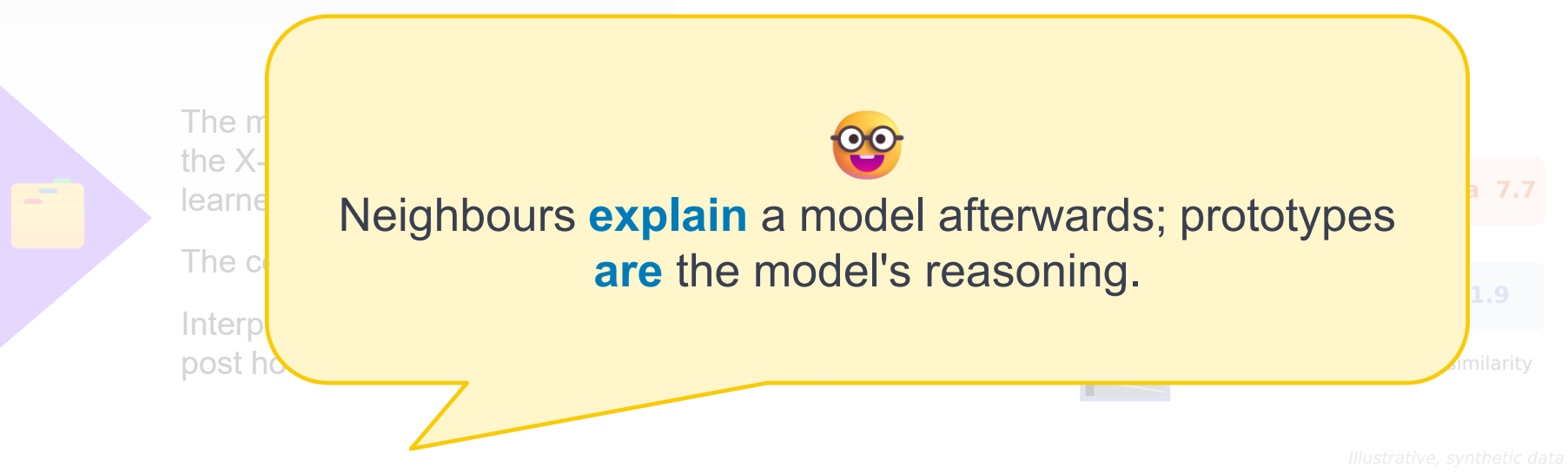
Interpretable **by design**, not post hoc



Illustrative, synthetic data

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

Prototypes: built-in explanations



Neighbours **explain** a model afterwards; prototypes **are** the model's reasoning.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

Who computes the prototypes?

The network itself, during training, in three phases



Illustrative, synthetic data

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

Who computes the prototypes?

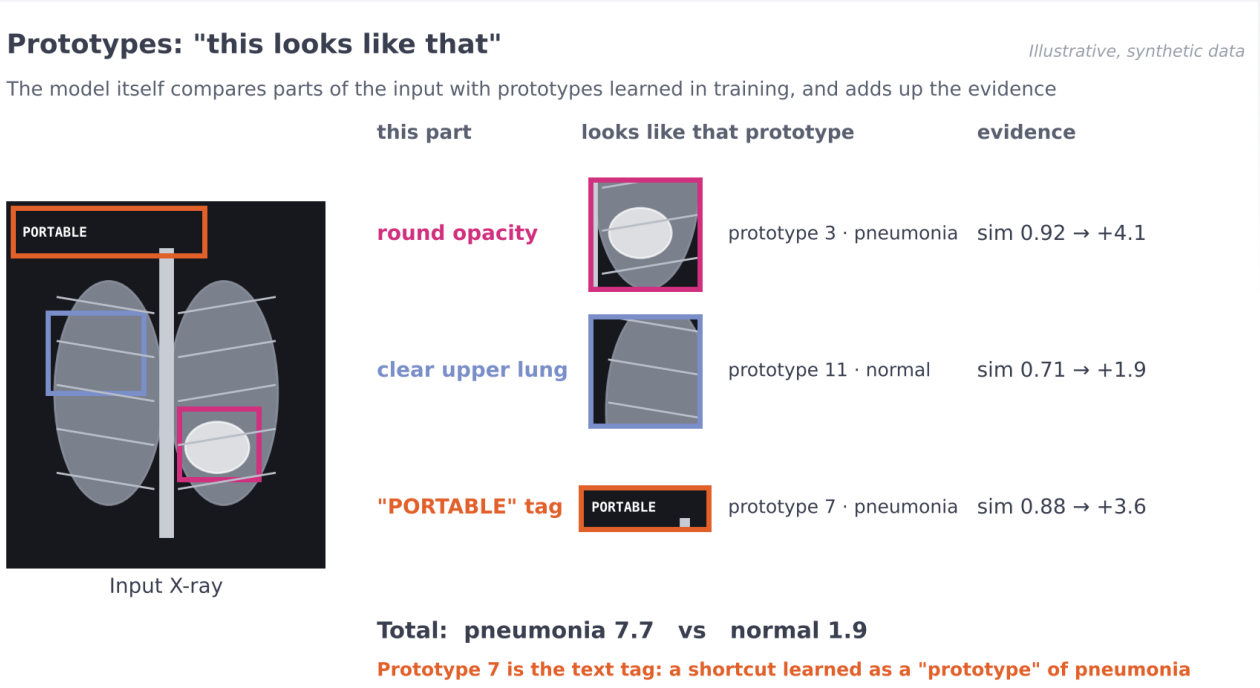
The network itself, during training, in three phases.



After the push, every prototype **is** a real patch from a training X-ray: we can **show** it.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

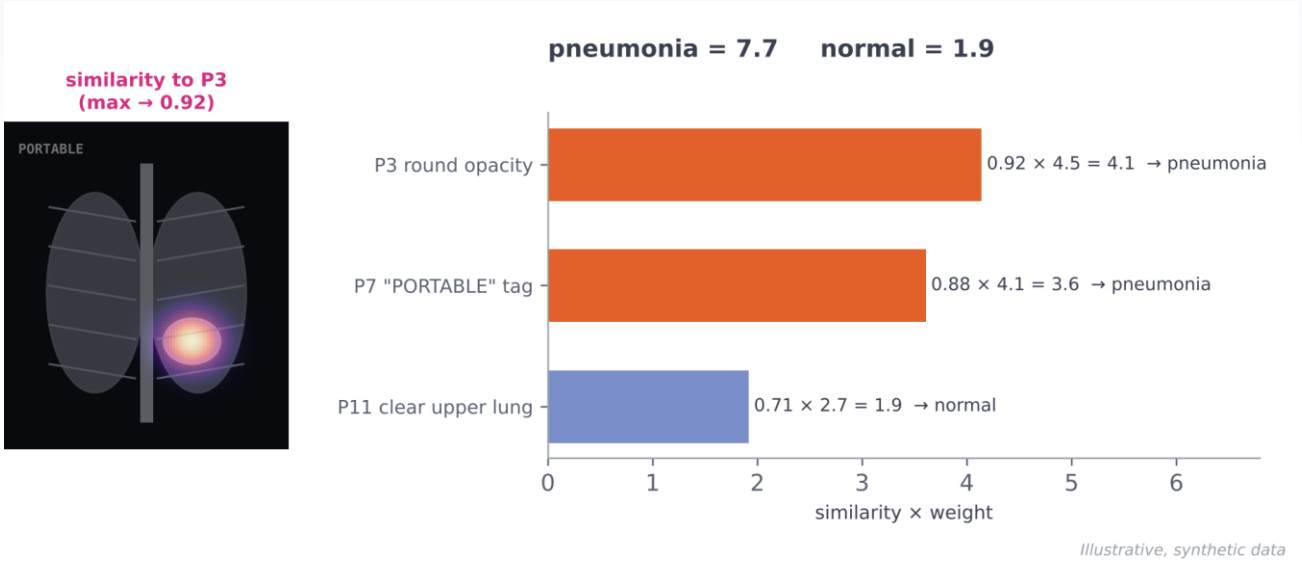
This looks like that



Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

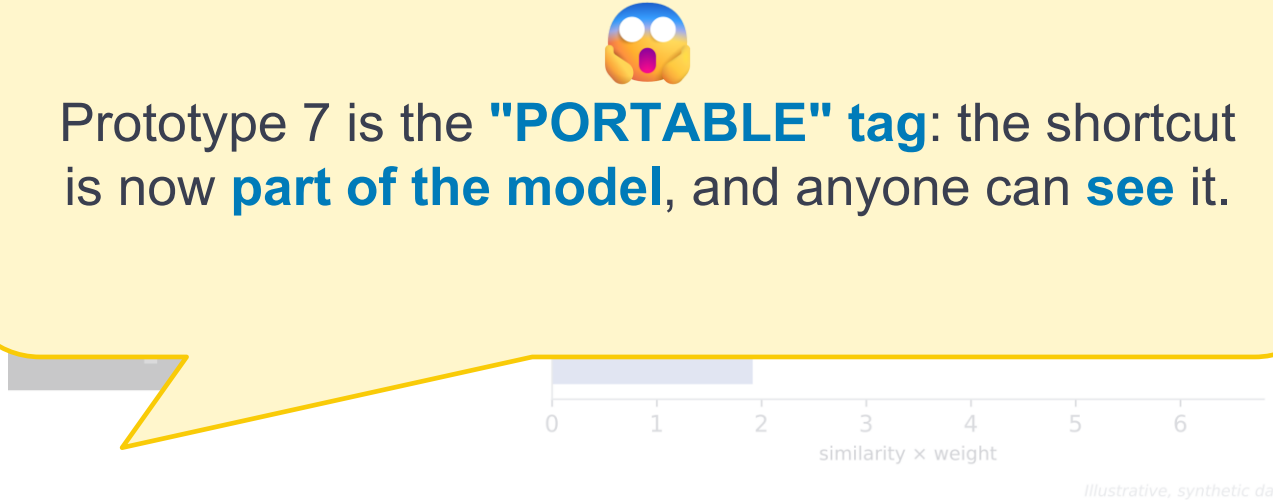
How the evidence adds up

$$sim_j = max_{patches} \log\left((d^2 + 1) / (d^2 + \varepsilon) \right) \qquad score_{class} = \Sigma_j \, w_j \cdot sim_j$$



Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

How the evidence adds up



Illustrative, synthetic data

Prototype 7 is the **"PORTABLE" tag**: the shortcut is now **part of the model**, and anyone can **see** it.

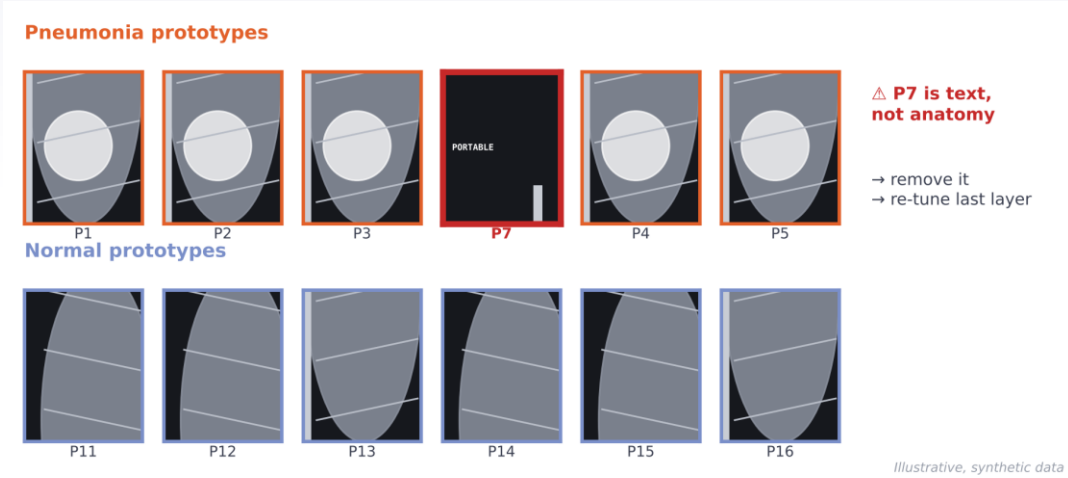
Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

Auditing the prototype gallery

A **global** view: all the prototypes of each class

A clinician checks they are **anatomy**

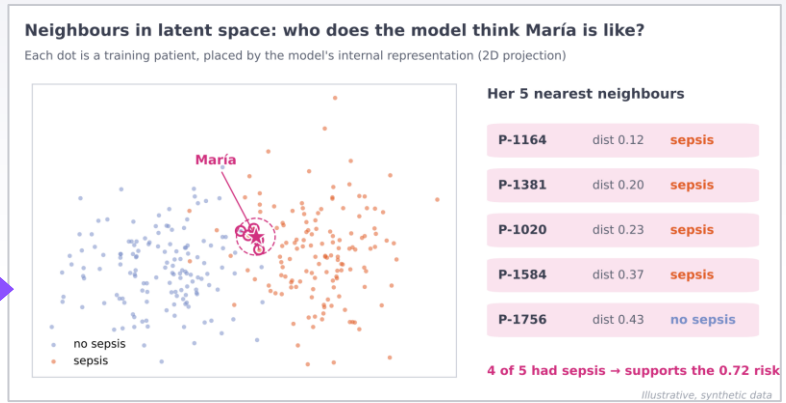
Remove a bad one and **re-tune** the last layer

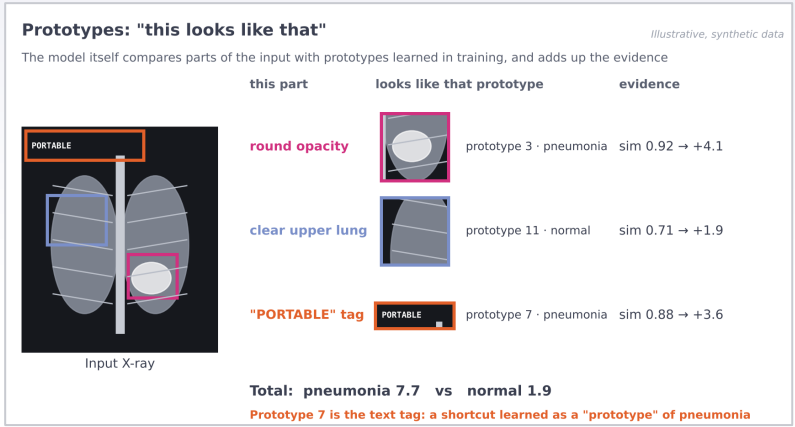


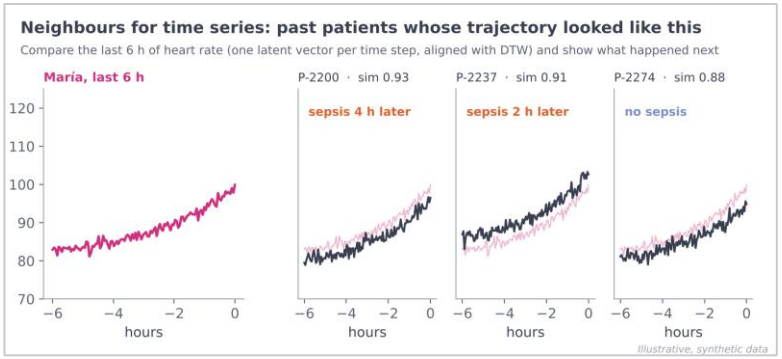
Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems, 32*.

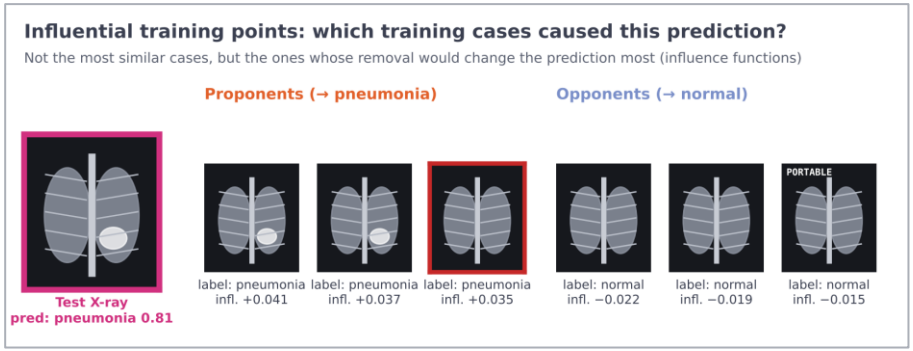
Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., & Rudin, C. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence, 3*(12), 1061–1070.

EE: Looks like

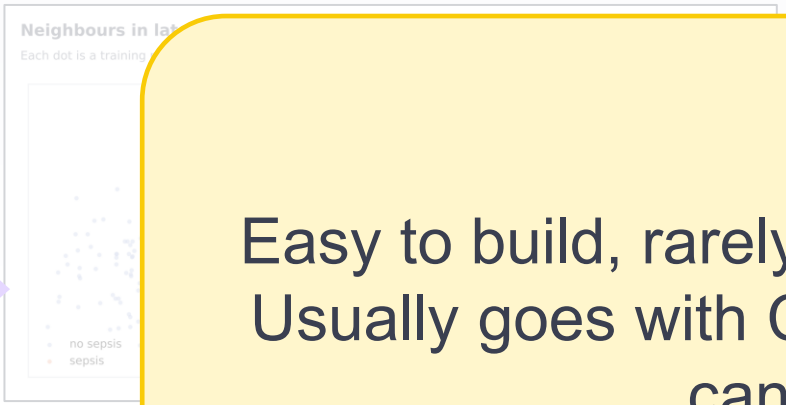


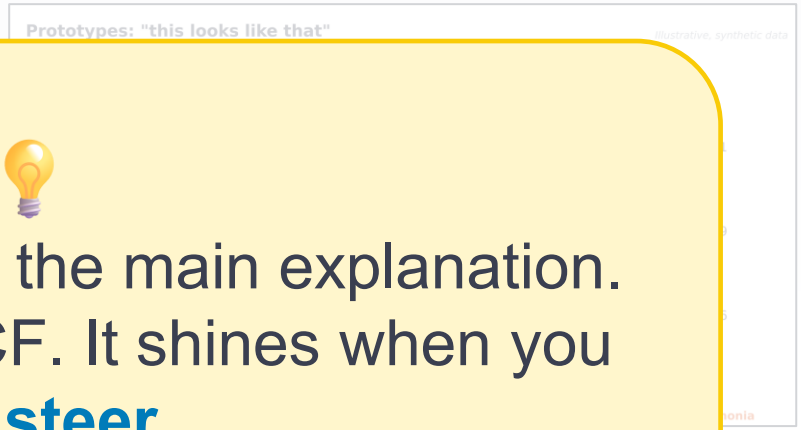


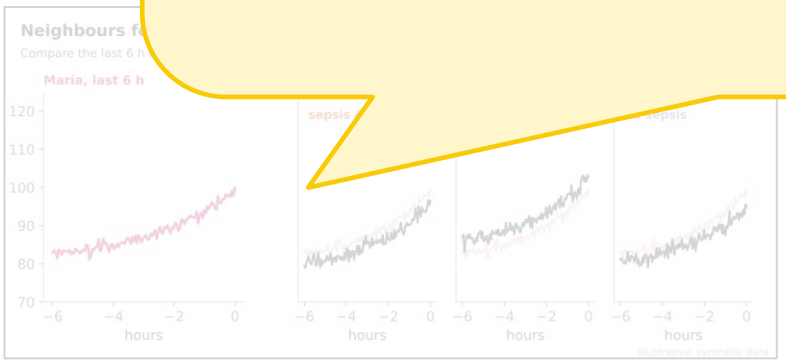


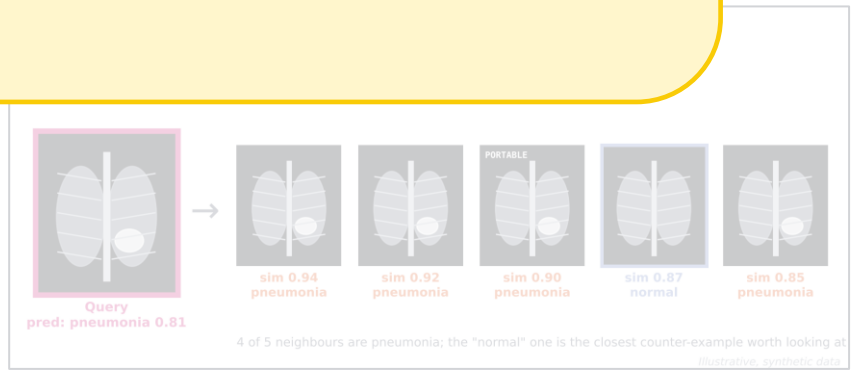


EE: Looks like









Easy to build, rarely the main explanation. Usually goes with CF. It shines when you can **steer**.

Five questions



Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

What must change to get another outcome?



Model behaviour

How does it react in a given situation?



WHAT DO I HAVE TO **CHANGE** TO GET A **DIFFERENT OUTCOME**?

CE: what is it

What: the **minimal change** to the input that flips the prediction

Why: recourse: how to reverse an unfavourable decision

Counterfactuals

What must change to get another outcome?

- Local
- Global
- Post hoc
- Self-explainable

CE: How

Optimization
Find the closest input with a different prediction

Wachter et al., 2017

Diversity
Several different counterfactuals, not just one

DiCE (Mothilal et al., 2020)

Actionability
Only change what the person can actually change

Actionable recourse (Ustun et al., 2019)

Retrieval
Take the nearest real sample with another prediction

What-If Tool (Wexler et al., 2020)

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617).

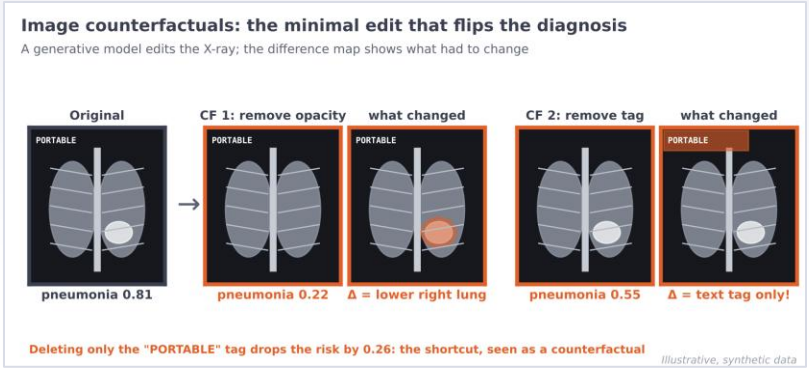
Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10–19).

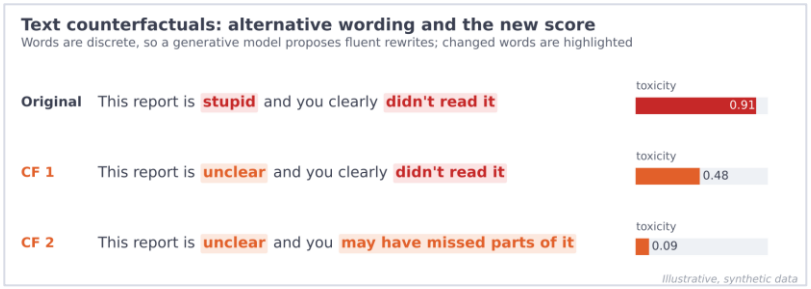
Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The What-If Tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 56–65.

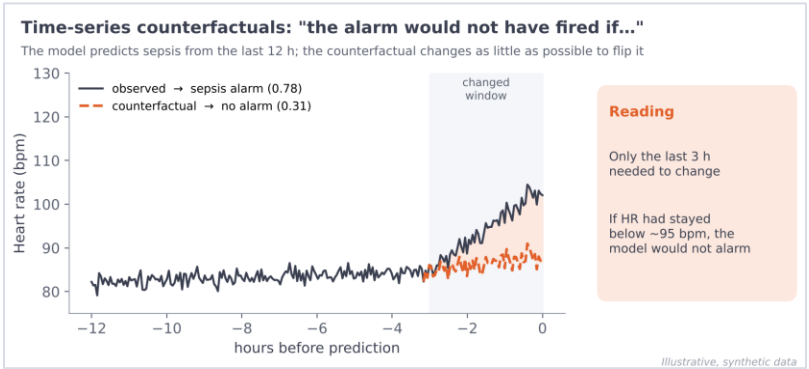
Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (2021). Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of ACL-IJCNLP 2021* (pp. 6707–6723).

CE: Looks like

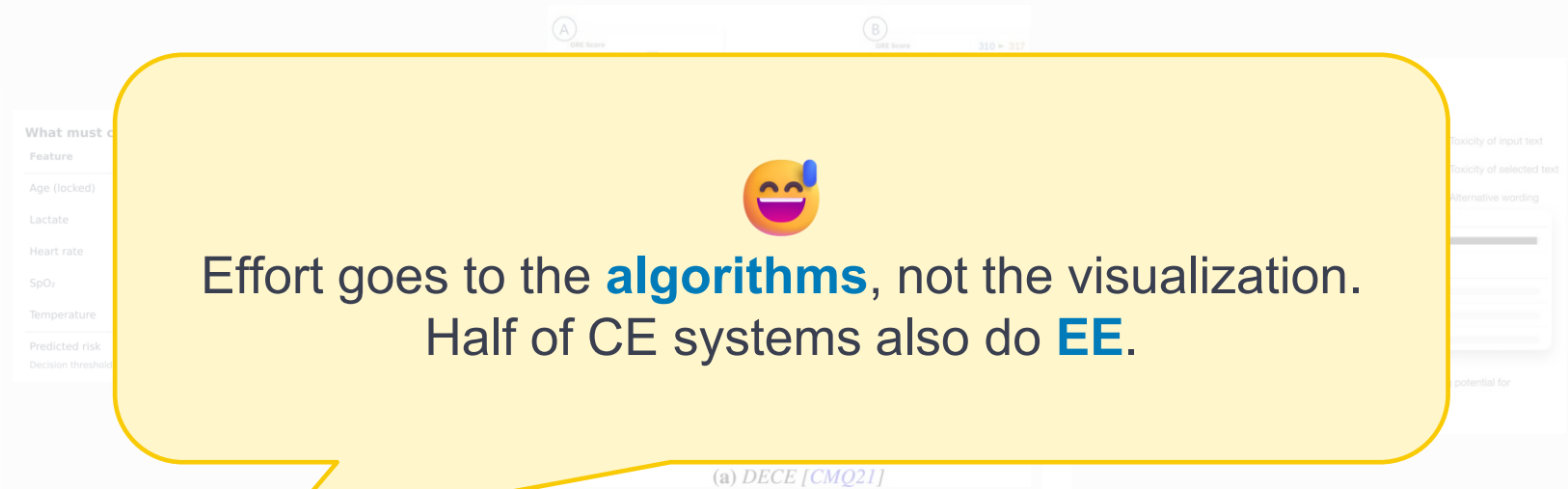








CE: Looks like



Effort goes to the **algorithms**, not the visualization.
Half of CE systems also do **EE**.

Tabular → original vs counterfactuals, changes highlighted

Multi-dimensional → parallel coordinates

Text → alternative words, inline

🧶 THREAD 6 · EXAMPLES & COUNTERFACTUALS

Explaining through similar cases and “what would have to change” is closest to expert reasoning and to regulation.

- **Taxonomy:** what kinds of example-based (EE) and counterfactual (CE) explanations exist?
- **A good counterfactual:** which properties should it have, and how do you visualize their trade-offs?
- **Constraints:** how does the user express domain constraints?
- **Beyond tabular:** what is a “minimal plausible change” in a time series, a graph or a text?
- **Similarity:** what is “similar” for the model, and what is it for the expert?
- **Recourse:** does the explanation give the affected person something they can do? (GDPR, AI Act)

Five questions



Feature attribution

Where is the model focusing?



Learned features

What has it learned?



Explanation by examples

Which samples does it consider similar?



Counterfactuals

What must change to get another outcome?



Model behaviour

How does it react in a given situation?



HOW DOES THE MODEL **REACT** IN A GIVEN **SITUATION**?

MB: what is it

What: common patterns of the model behaviour, linked to inputs, outputs or internals

Why: summarizes how the whole model behaves when no single explanation is enough



Model behaviour

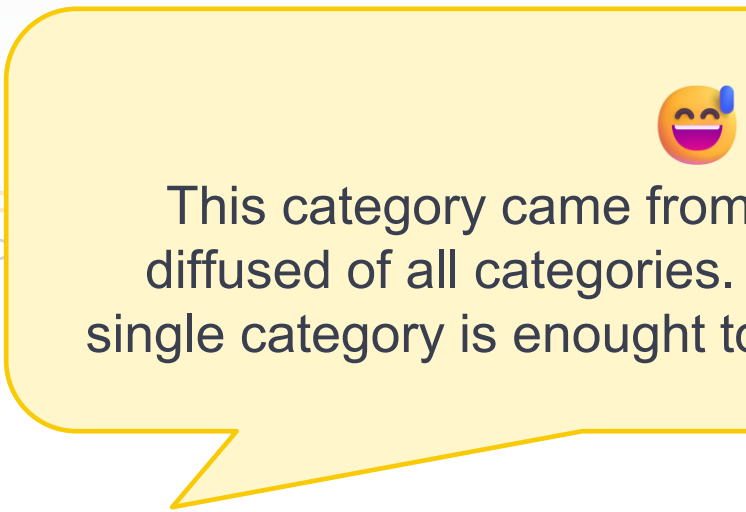
How does it react in a given situation?

Local	Global
Post hoc	Self-explainable

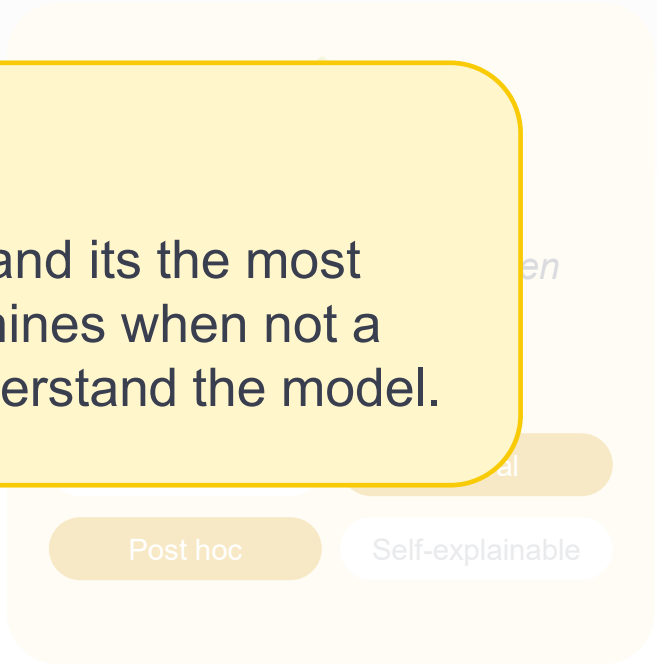
MB: what is it

What
linked

Why:
when



This category came from VA and its the most diffused of all categories. Its shines when not a single category is enought to understand the model.



en

Post hoc

Self-explainable

MB: How

Global surrogates

Train a simple model that mimics the network

TREPAN (Craven & Shavlik, 1995)

Soft decision tree (Frosst & Hinton, 2017)

Group situations by internal state

Find recurring patterns in activations, states or actions

e.g. **SpRAy** (Lapuschkin et al., 2019)

Mine behaviour patterns

Project internal states and link them to inputs and outputs

e.g. **DQNViz** (Wang et al., 2019) for RL agents

Average effects

How the output responds, on average, to one input

Partial dependence (Friedman, 2001)

Craven, M. W., & Shavlik, J. W. (1995). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems, 8*.

Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Wang, J., Gou, L., Shen, H.-W., & Yang, H. (2019). DQNViz: A visual analytics approach to understand deep Q-networks. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 288–298.

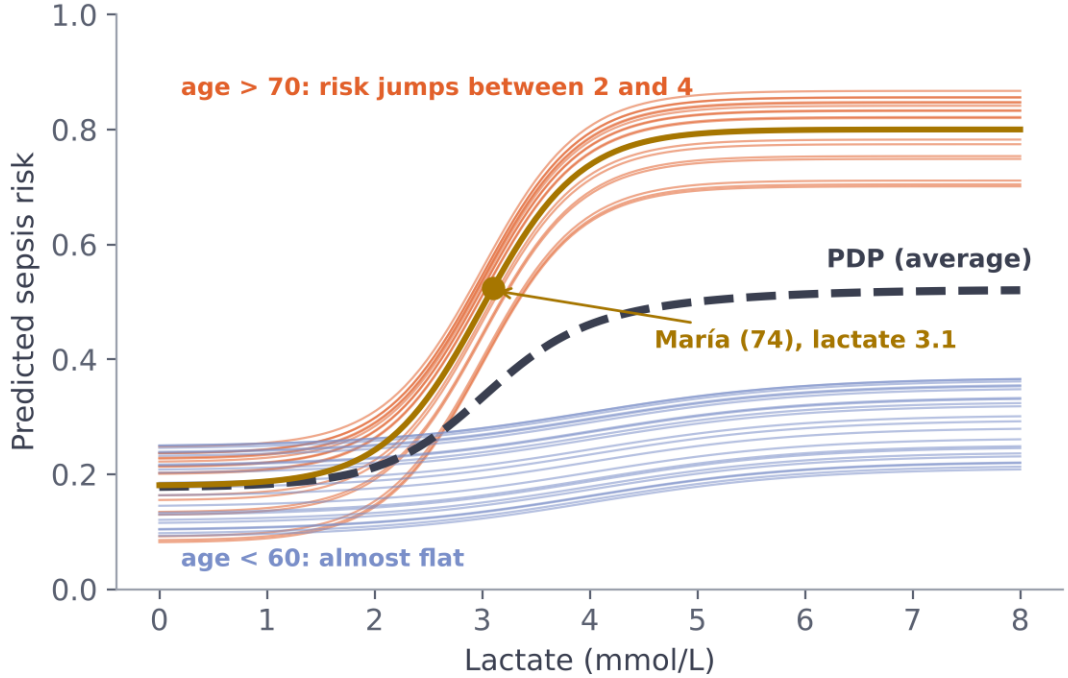
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications, 10(1), 1096.

MB: Looks like

Feature effects: average behaviour (PDP) and case-by-case curves (ICE)

How the predicted risk responds to lactate for 40 patients; the thick line is their average



How to read it

Each thin line: one patient, moving only their lactate (ICE)

Dashed line: the average (PDP) suggests a mild, smooth effect...

...but it describes nobody: two behaviours by age. Colouring the ICE curves reveals the pattern

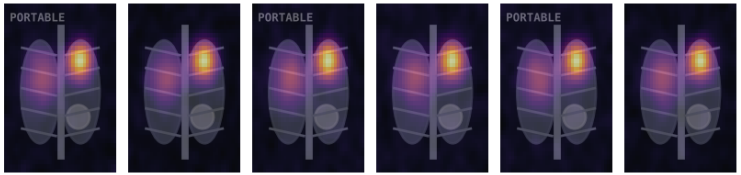
Illustrative, synthetic data

MB: Looks like

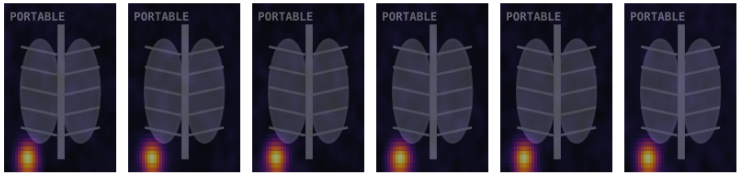
Clustering explanations (SpRAy-style): which strategies does the model use?

Compute an attribution heatmap for every "pneumonia" X-ray, then cluster the heatmaps, not the images

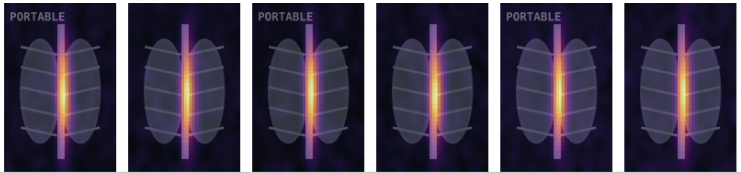
Strategy A · looks at the lung opacity

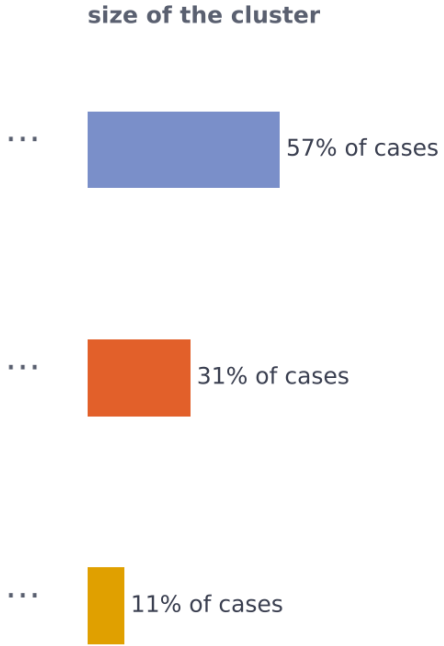


Strategy B · looks at the "PORTABLE" tag



Strategy C · looks at tubes and lines





How to read it

Raw material: one FA heatmap per case

Cluster the heatmaps → each group is a strategy of the model

Strategy B (31%) is the shortcut: found without looking at images one by one

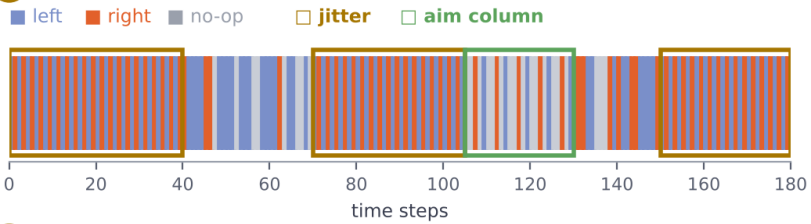
Illustrative, synthetic data

MB: Looks like

Mining behaviour patterns in an RL agent (Breakout) + a Markov chain as surrogate

No need to open the network: record what the agent does, mine recurring patterns, summarise the policy as phases + transitions

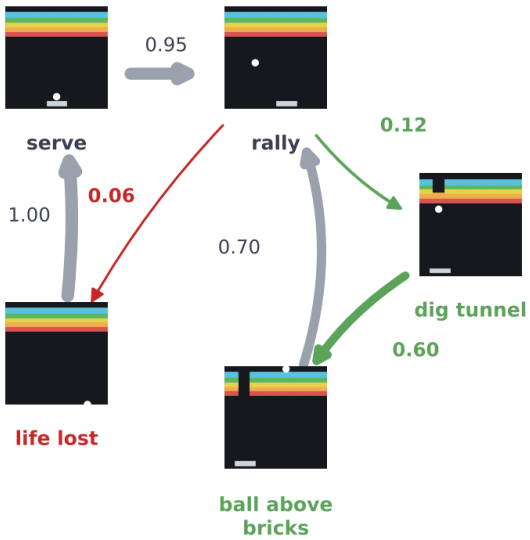
1 record actions over many episodes (one shown)



2 mine frequent action patterns and their context

Pattern	Share	When it happens
`L R L R ("jitter")`	38%	ball far away, moving up
track ball → hit	41%	ball falling towards paddle
`N N R N L ("aim")`	9%	one column almost cleared
miss after long jitter	4%	ball returns fast after jitter

3 Markov chain as surrogate: phases of the policy



Phases = clusters of game states · arrows = transition prob.
Green loop = the "tunnel" strategy · fidelity 0.87

Illustrative, synthetic data · inspired by DQNViz and Zahavy et al. (2016)

🧶 THREAD 7 · SURROGATE MODELS

A surrogate explains the surrogate. A tree, rules or a linear model trained on the network's predictions, not the labels. VA must show its logic, its fidelity and the data each rule covers; with low fidelity where it matters, it misleads while looking transparent.

- **Taxonomy:** global (TREPAN, soft decision trees, CNN2DT) vs. local (LIME, Anchors); pedagogical vs. decompositional extraction.
- **Fidelity:** measured overall, per region, per class? How to balance and show fidelity, coverage and complexity together?
- **When it misleads:** high global but low local fidelity, instability under resampling, correlated features, explaining the data instead of the model.
- **Visualization:** How does it scale to hundreds of rules, and how does the user tune complexity?
- **The DL question:** on which interpretable features, if pixels or series points mean nothing? (Metadata, concepts, neurons, series features.)
- **Error analysis:** can rules describe the subgroups where the model fails?
- **Beyond classification:** RL policies, GNNs, generative methods … etc.

Case studies

REAL SYSTEMS MIX SEVERAL CATEGORIES

How to read the cases

Categories

FA
Feature attribution

LF
Learned features

EE
Explanation by examples

CE
Counterfactuals

MB
Model behaviour

Users

Architect
designs new DL architectures

Trainer
applies known recipes

End user
little or no DL knowledge

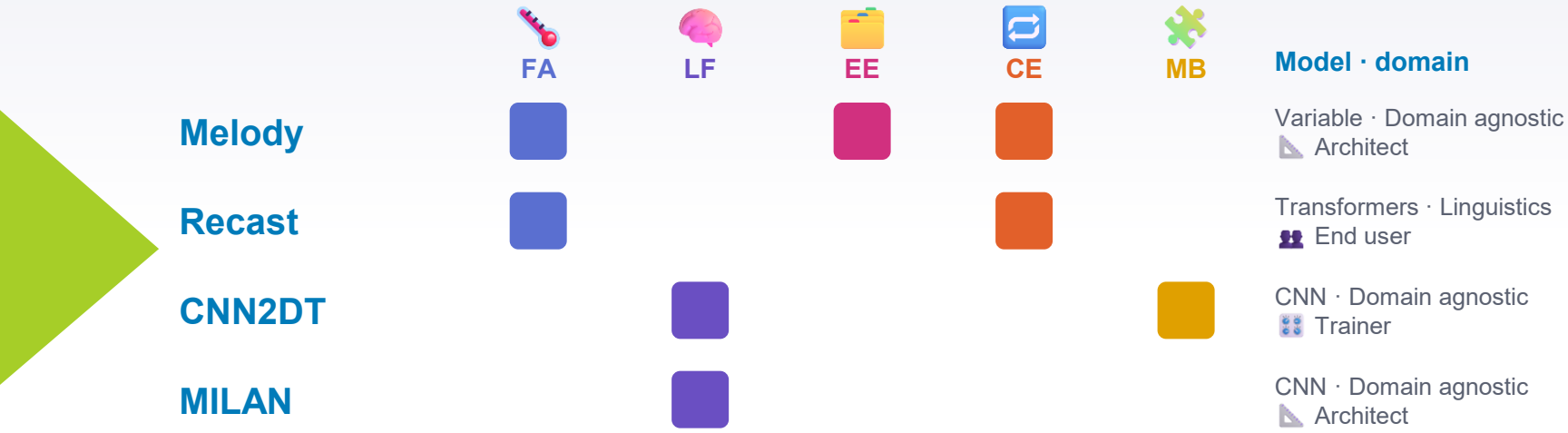
Interactivity

Passive
only looks

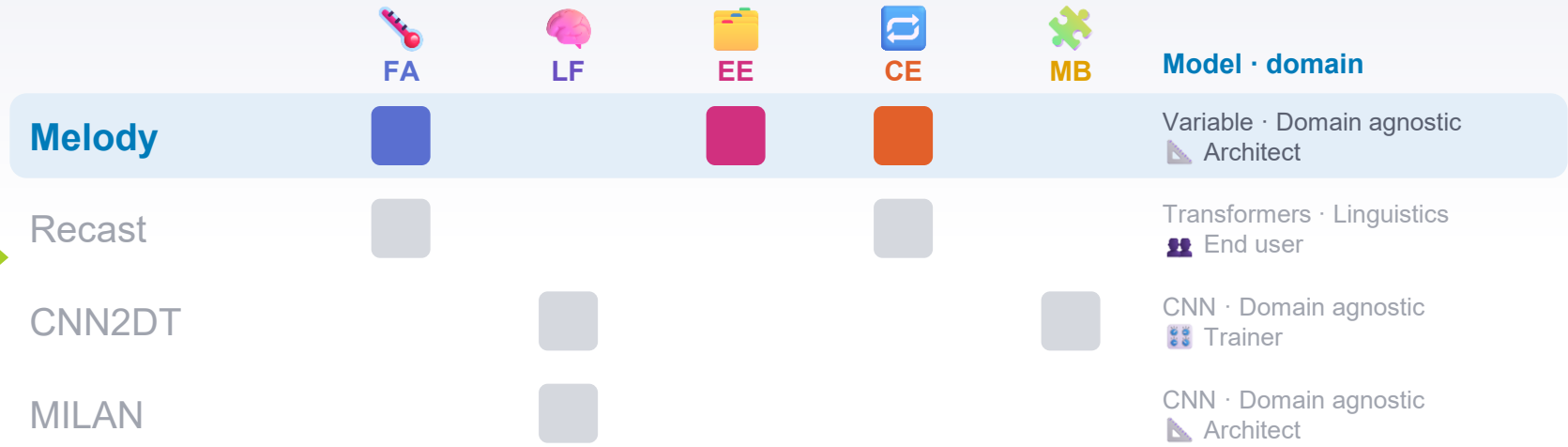
Input
edits the input

Model
edits the internals

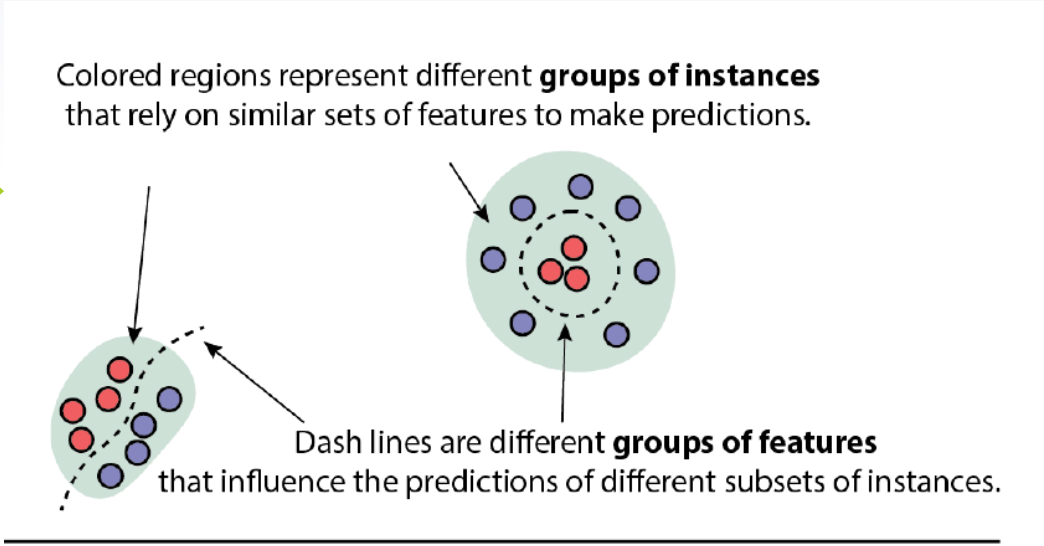
Category fingerprint



Category fingerprint



Melody: the idea



FA | LF | EE | CE | MB

Problem: local explanations are faithful but come **one at a time**

Idea: group the **instances** that are explained alike…

…and, at the same time, the **features** that explain them

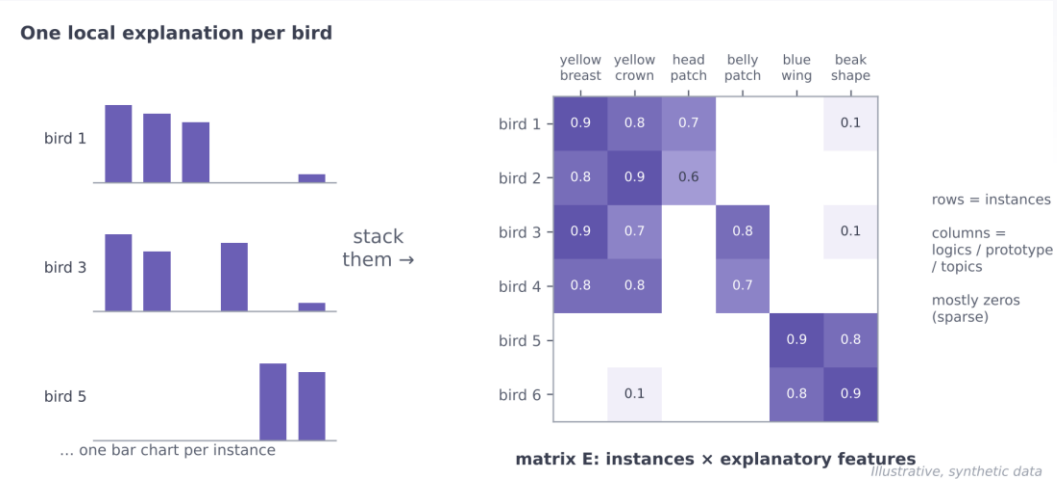
Result: a compact **global** view built from local pieces

Architect

Variable · Domain agnostic

Chan, G. Y.-Y., Bertini, E., Nonato, L. G., Barr, B., & Silva, C. T. (2020). Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *arXiv preprint arXiv:2007.10614*.

Step 1: the explanation matrix



FA | LF | EE | CE | MB

Any local method: SHAP, LIME, Integrated Gradients, prototypes…

One **row per instance**, one **column per explanatory feature**

Features depend on the data: **logics** (tabular), **prototypes** (image), **topics** (text)

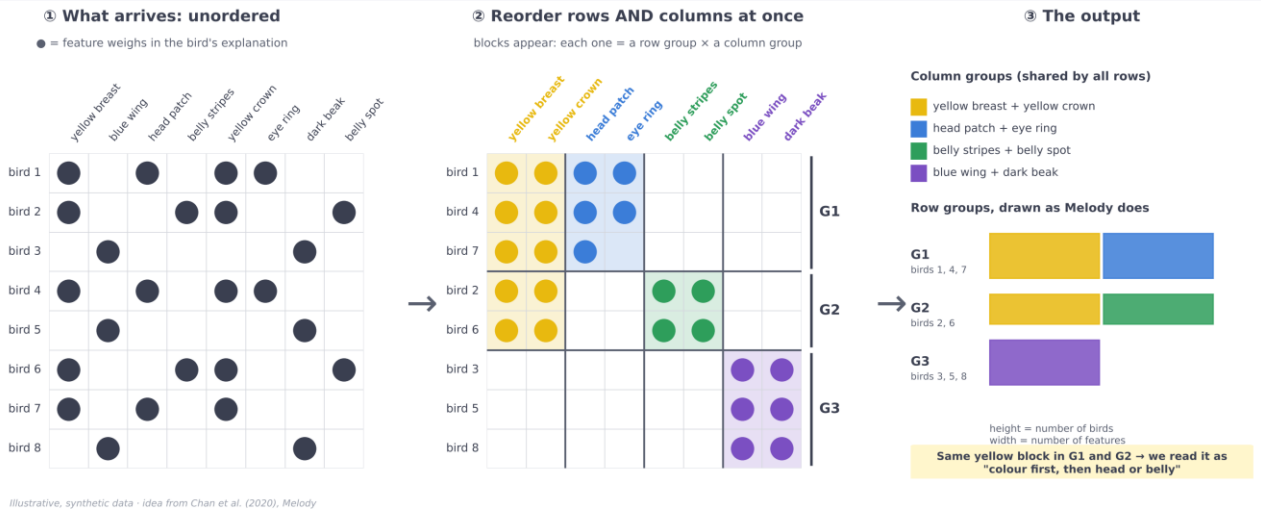
Normalized, non-negative and **sparse**

Architect

Variable · Domain agnostic

Chan, G. Y.-Y., Bertini, E., Nonato, L. G., Barr, B., & Silva, C. T. (2020). Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *arXiv preprint arXiv:2007.10614*.

Step 2: co-clustering



FA | LF | EE | CE | MB

Group **rows and columns** at once: each block is a **co-cluster**

The summary is the **mass** of each block

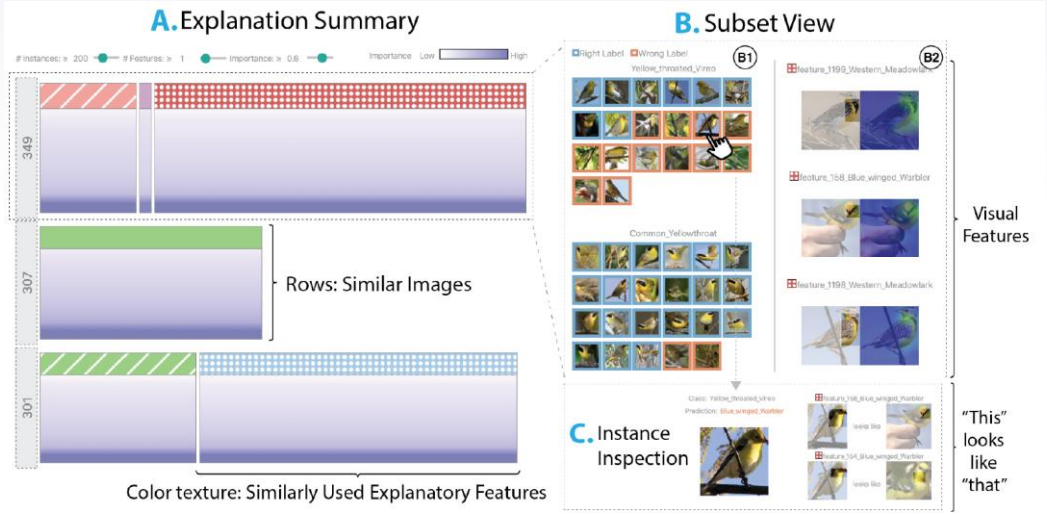
Rebuild a matrix from the blocks

📐 Architect

Variable · Domain agnostic

Chan, G. Y.-Y., Bertini, E., Nonato, L. G., Barr, B., & Silva, C. T. (2020). Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *arXiv preprint arXiv:2007.10614*.

Global → subset → local



FA | LF | EE | CE | MB

Click a row → **subset view**: its instances and their features

Click an image → **instance view**: "this looks like that"

Click a patch → a **query** that filters the summary

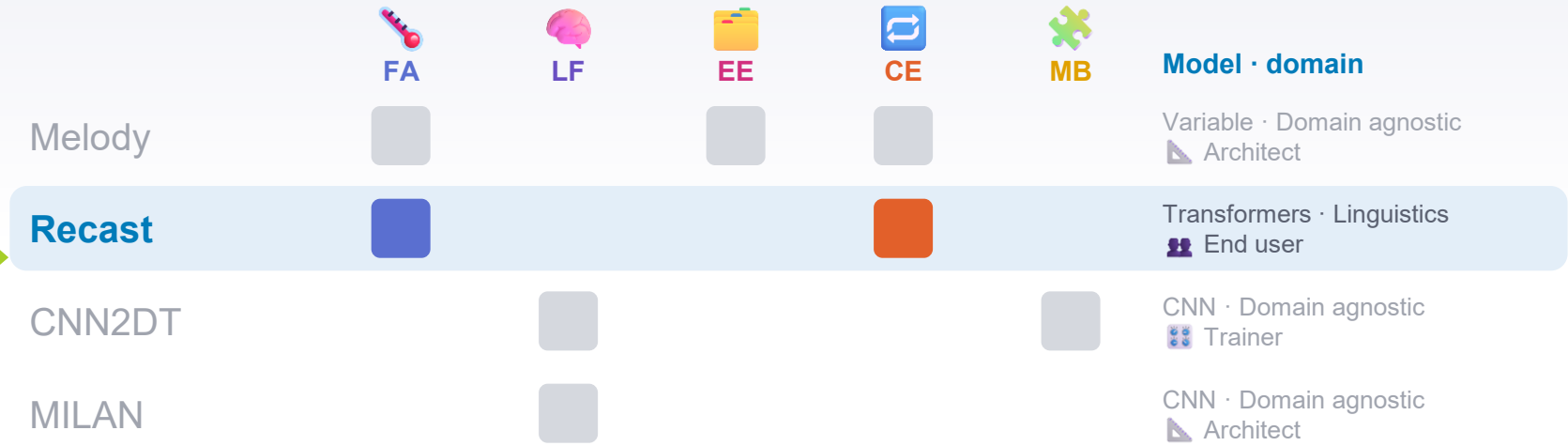
The paper calls it **"the loop"**

Architect

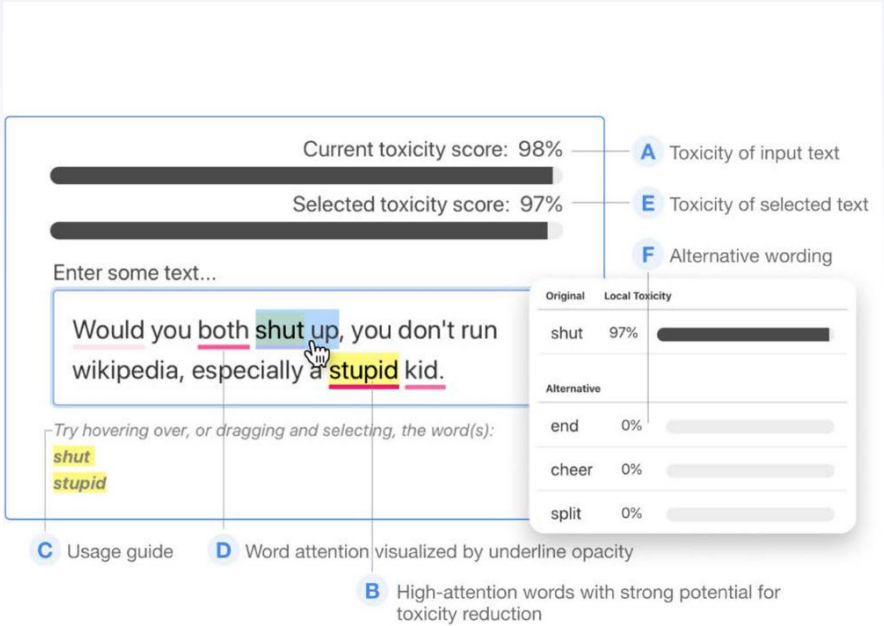
Variable · Domain agnostic

Chan, G. Y.-Y., Bertini, E., Nonato, L. G., Barr, B., & Silva, C. T. (2020). Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *arXiv preprint arXiv:2007.10614*.

Category fingerprint



Case: Recast



(b) *Recast [WSP*21]*

FA | LF | EE | CE | MB

Task: help people write **less toxic** comments

FA: high-attention words are underlined

CE: alternative wording and the new toxicity score

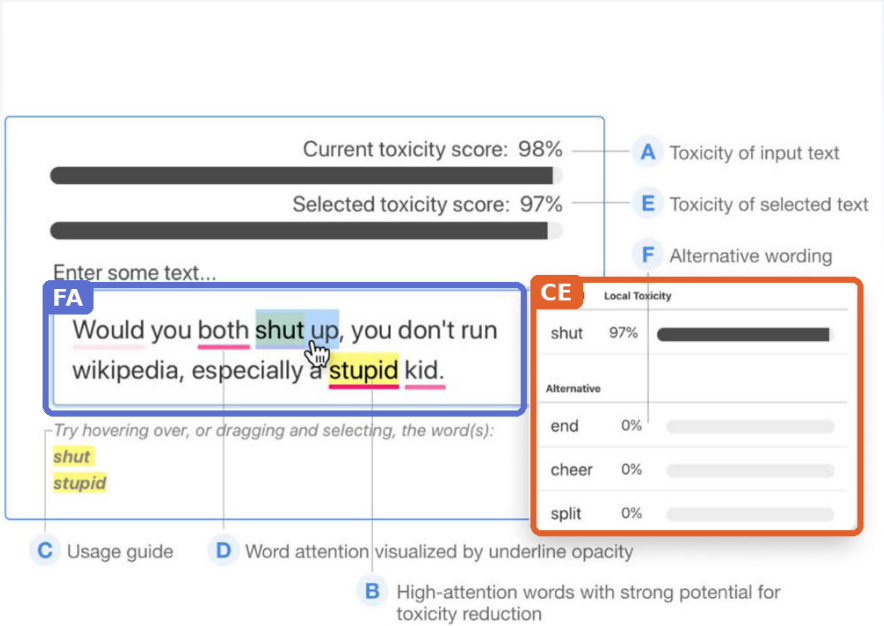
Built for **end users**: looks like a spell-checker

End user

Transformers · Linguistics

Wright, A. P., Shaikh, O., Park, H., Epperson, W., Ahmed, M., Pinel, S., Chau, D. H. P., & Yang, D. (2021). RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–26.

Recast: FA + CE



(b) *Recast [WSP*21]*

FA | LF | EE | CE | MB

FA (blue): which words drive the score

CE (orange): what to write instead

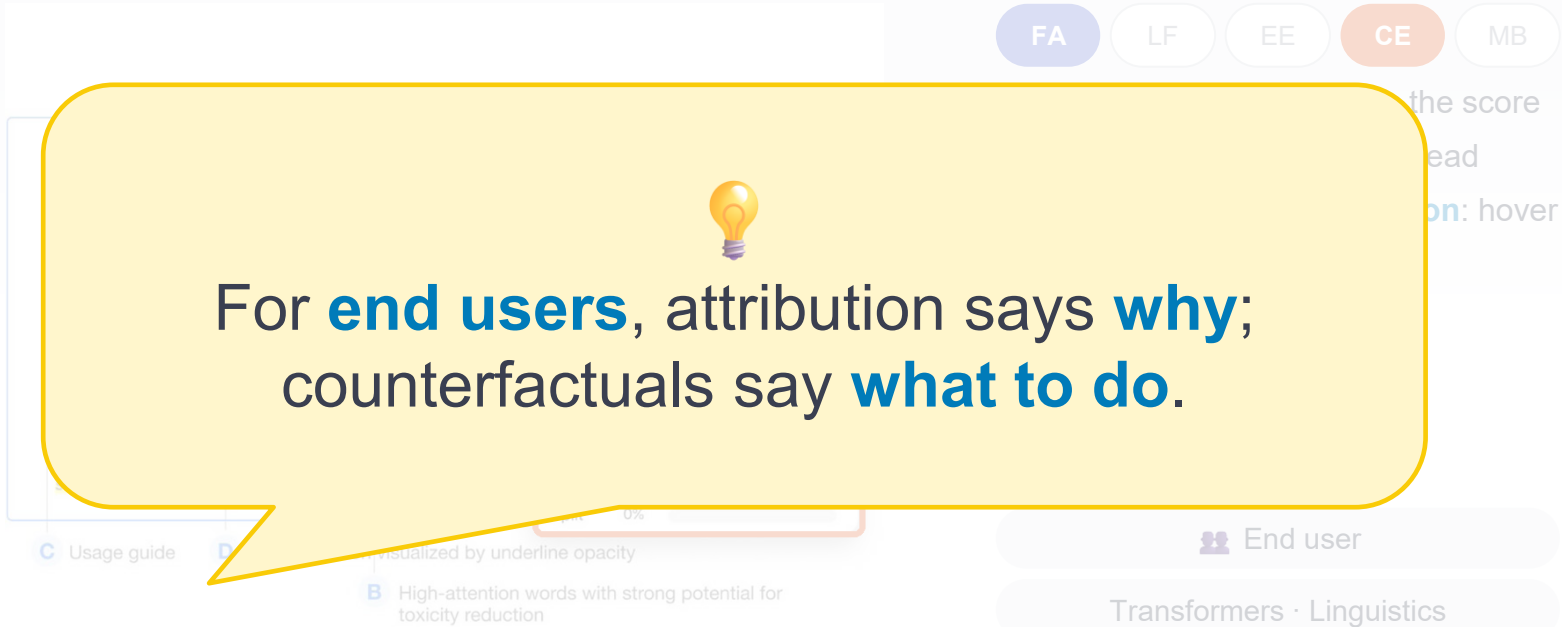
Two categories, **one interaction**: hover a word

End user

Transformers · Linguistics

Wright, A. P., Shaikh, O., Park, H., Epperson, W., Ahmed, M., Pinel, S., Chau, D. H. P., & Yang, D. (2021). RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–26.

Recast: FA + CE

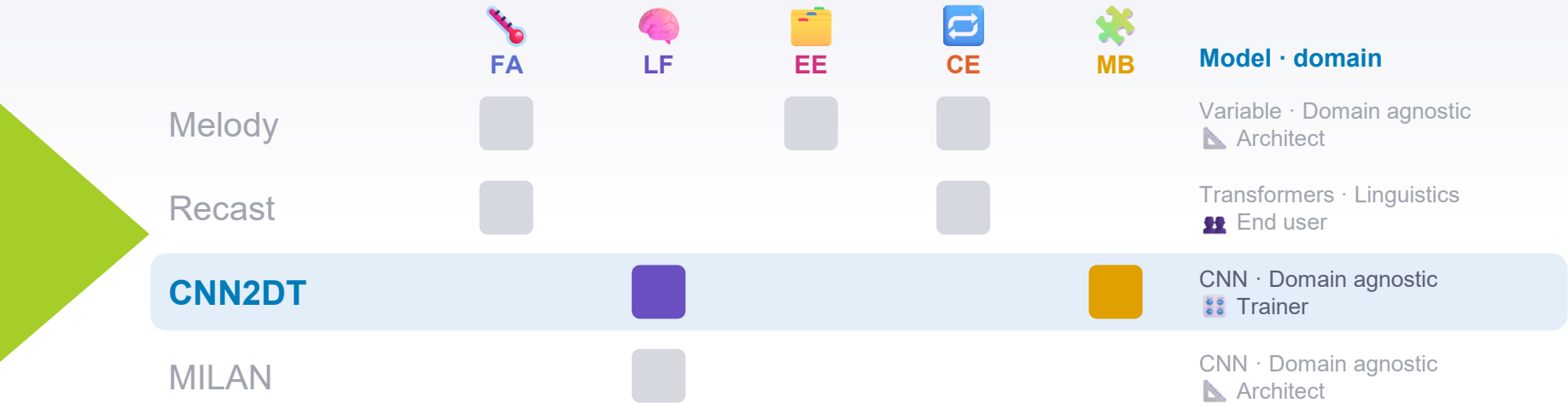


(b) *Recast [WSP*21]*

For **end users**, attribution says **why**; counterfactuals say **what to do**.

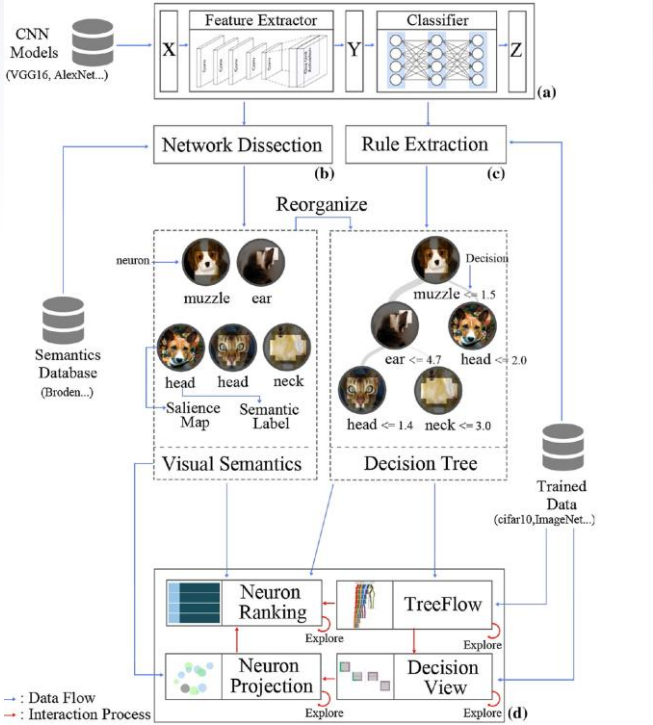
Wright, A. P., Shaikh, O., Park, H., Epperson, W., Ahmed, M., Pinel, S., Chau, D. H. P., & Yang, D. (2021). RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–26.
Image: La Rosa et al. (2023), *Computer Graphics Forum, 42*(1), Fig. 10b. Frames added.

Category fingerprint



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (Table 1)

CNN2DT: the idea



FA **LF** EE CE **MB**

Surrogate trees are readable, but a tree of a whole CNN splits on **pixels** and is **huge**

Idea: cut the CNN in two and extract the tree **only from the classifier**

Name each input of the tree with the concept its neuron detects

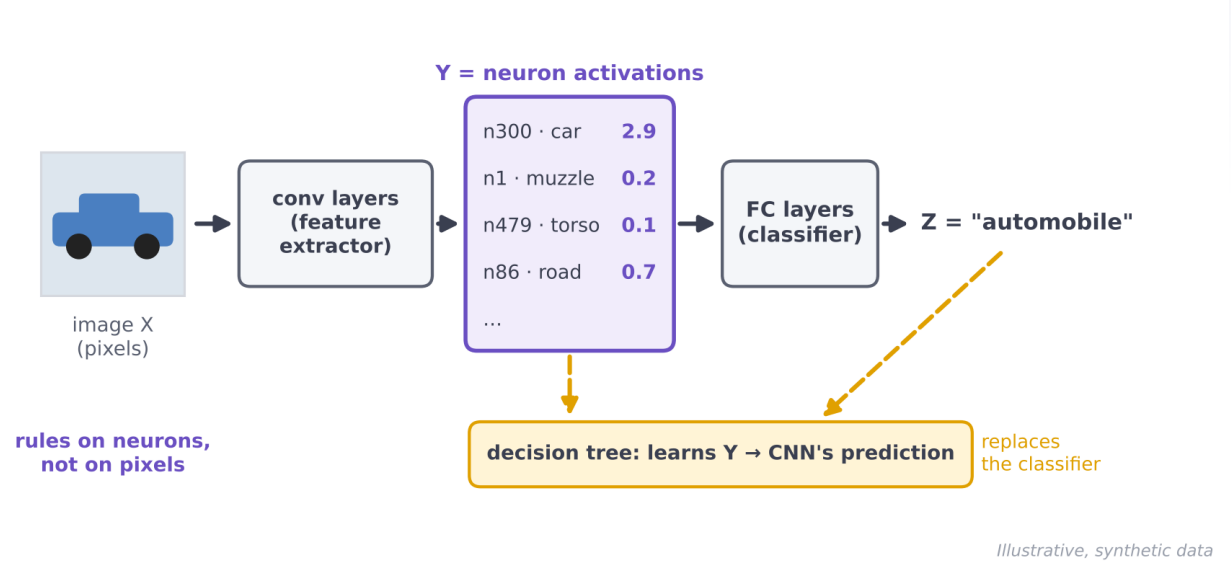
Result: rules like *"car(300) ≤ 2.53"*

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

Step 1: split the CNN



FA **LF** EE CE **MB**

Feature extractor (conv layers): image → **activations of the last conv layer**

Classifier (FC layers): activations → class

The tree learns **activations → the CNN's prediction**

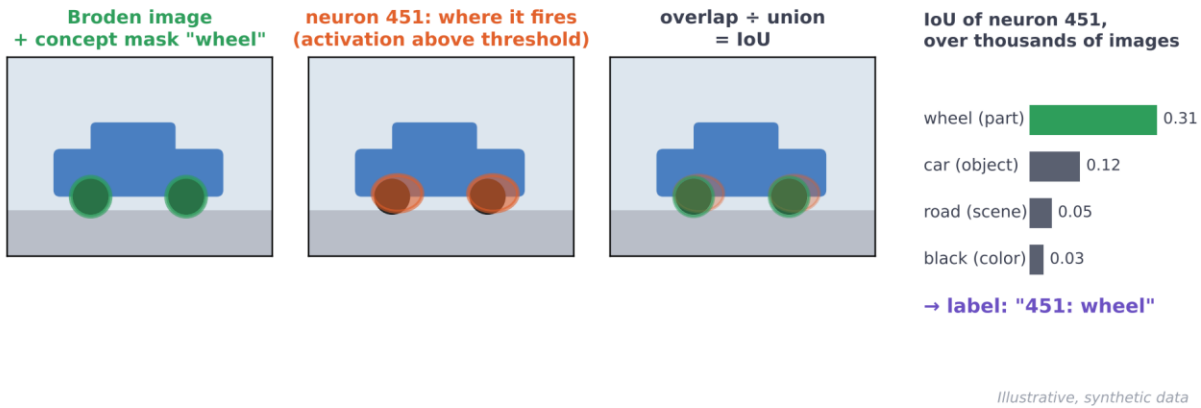
So its rules talk about **neurons**, not pixels

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156..

Step 2: name the neurons



FA **LF** EE CE **MB**

Network Dissection with **Broden**: images with pixel-level concept masks

Compare where the neuron **fires** with each concept's **mask** (IoU)

Each neuron gets the concept with the **highest IoU**

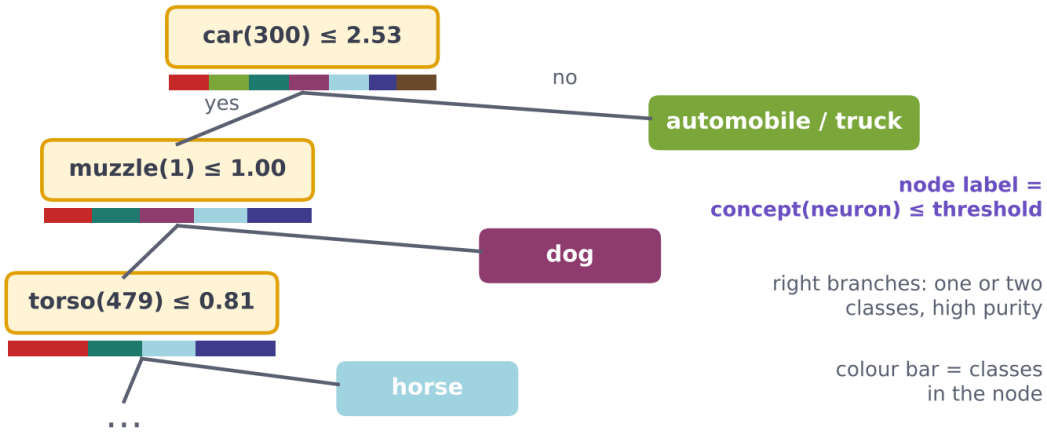
6 categories: object, part, scene, material, texture, colour

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

Step 3: grow the surrogate tree



Illustrative, based on Jia et al. (2020), Fig. 9

FA | **LF** | EE | CE | **MB**

CART (Gini) on the CNN's predictions: pick the neuron and threshold that best **purify** the data

Node label = **concept(neuron) ≤ threshold**

Right branches: images that **activate** the neuron, often one or two classes

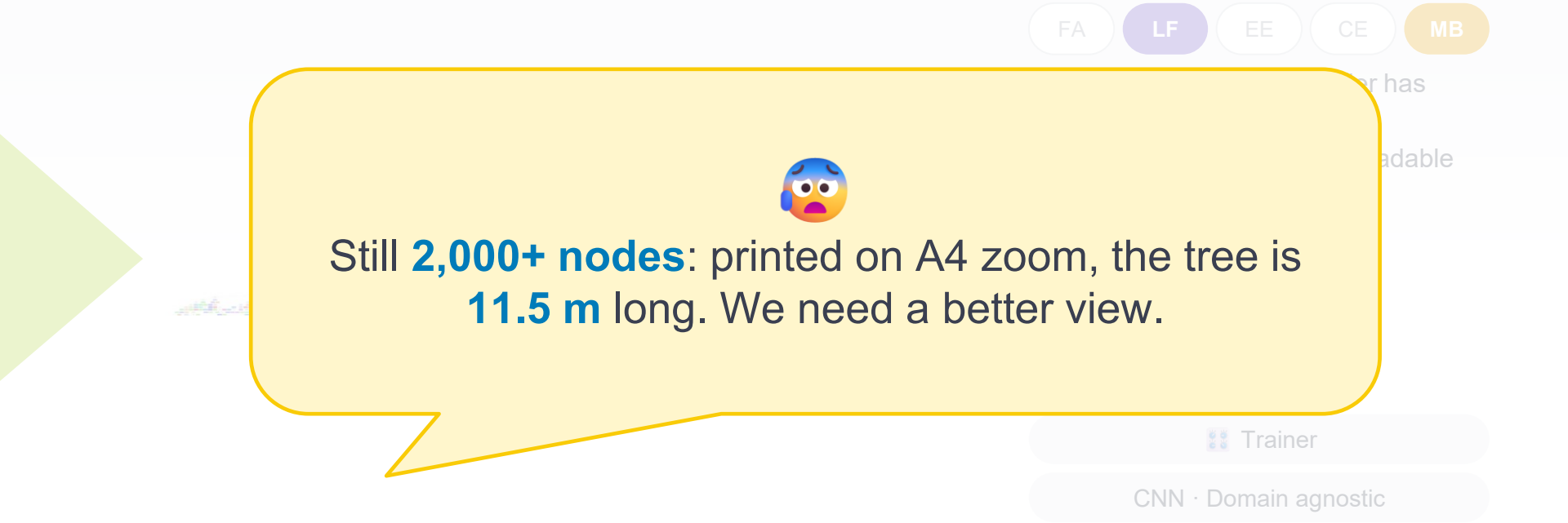
Fidelity ≈ 75%; D(n) ranks neurons by how **high** in the tree they appear

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

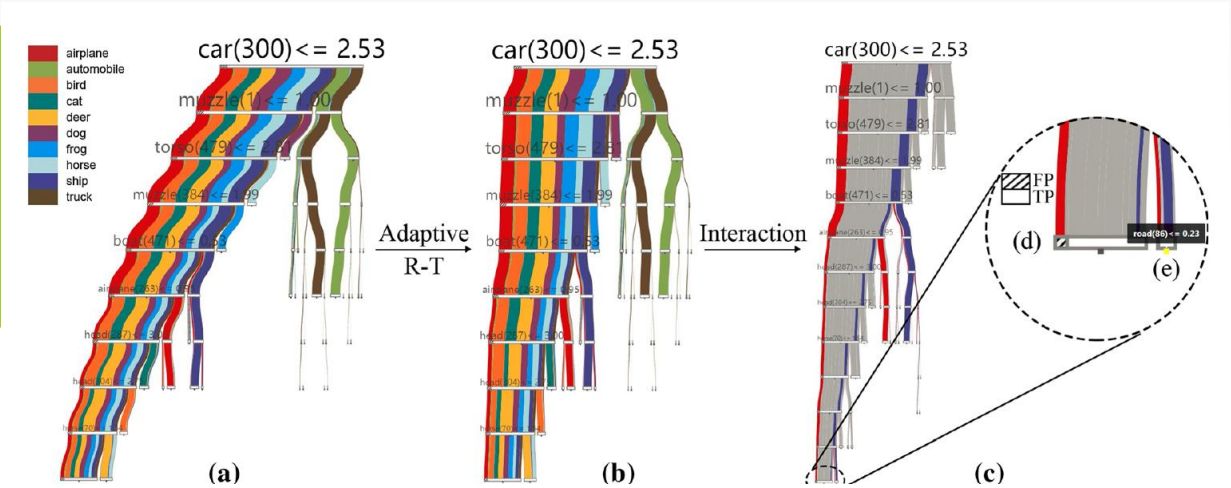
Why a visualization?



Still **2,000+ nodes**: printed on A4 zoom, the tree is **11.5 m** long. We need a better view.

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.
Image: Jia et al. (2020), Fig. 3.

TreeFlow



FA | **LF** | EE | CE | **MB**

Bands = data flowing down the tree: width = samples, colour = class

Label = **concept + threshold**; font shrinks with depth

Collapse nodes, **switch classes** off, cut thin branches

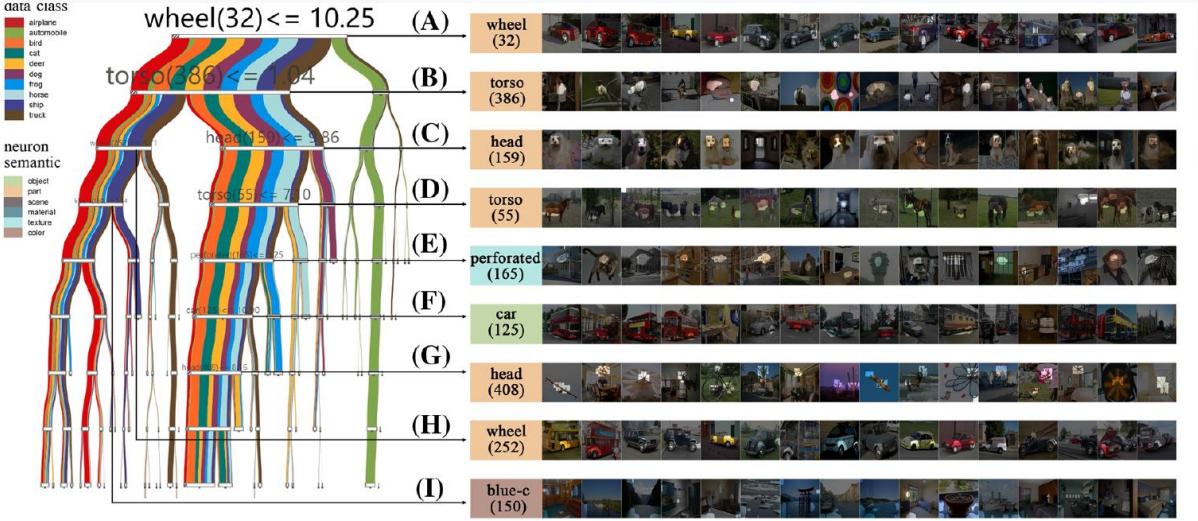
Layout keeps **left child left, right child right** (adaptive R-T)

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

VGG16 vs AlexNet



FA | **LF** | EE | CE | **MB**

Same pipeline, **two CNNs** on CIFAR-10 (VGG16 89.5%, AlexNet 85%)

VGG16: object and part neurons (car, muzzle, torso), a tidy tree

AlexNet: more, deeper, less pure branches; **texture and colour** neurons

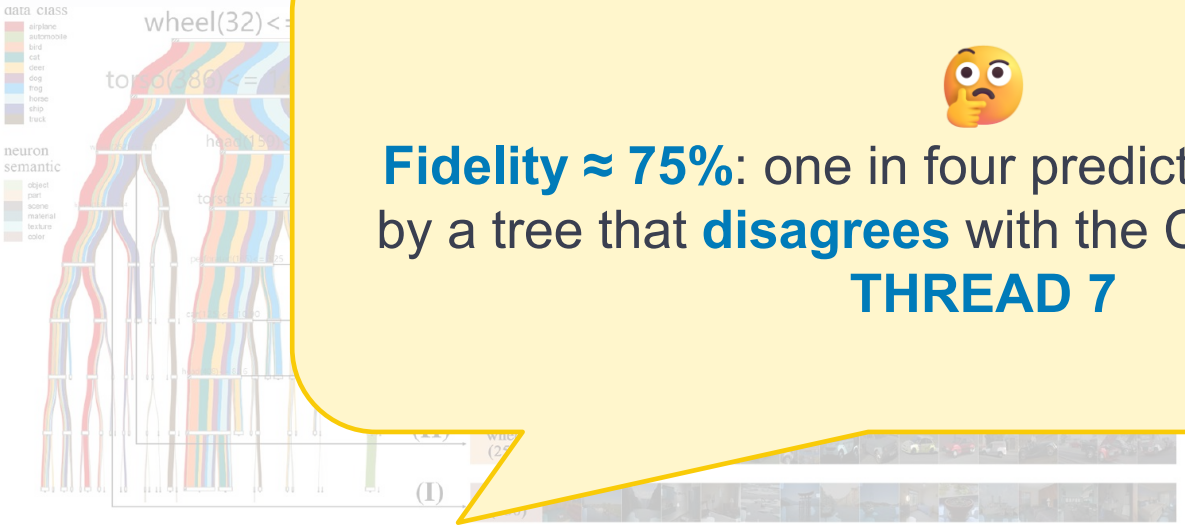
AlexNet separates ships with **"blue"**: works here, risky elsewhere

Trainer

CNN · Domain agnostic

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

VGG16 vs AlexNet

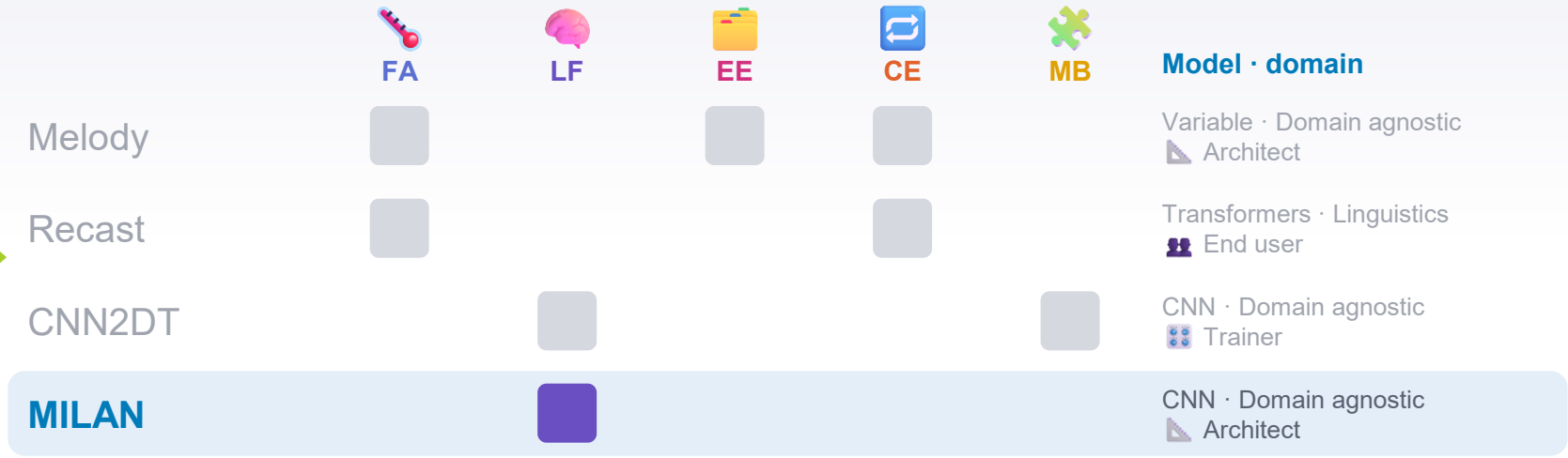


🤔

Fidelity ≈ 75%: one in four predictions is explained by a tree that **disagrees** with the CNN. Enough? → **THREAD 7**

Jia, S., Lin, P., Li, Z., Zhang, J., & Liu, S. (2020). Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization, 23*(1), 141–156.

Category fingerprint



Case: MILAN

FA | **LF** | EE | CE | MB

Architect

CNN · Domain agnostic



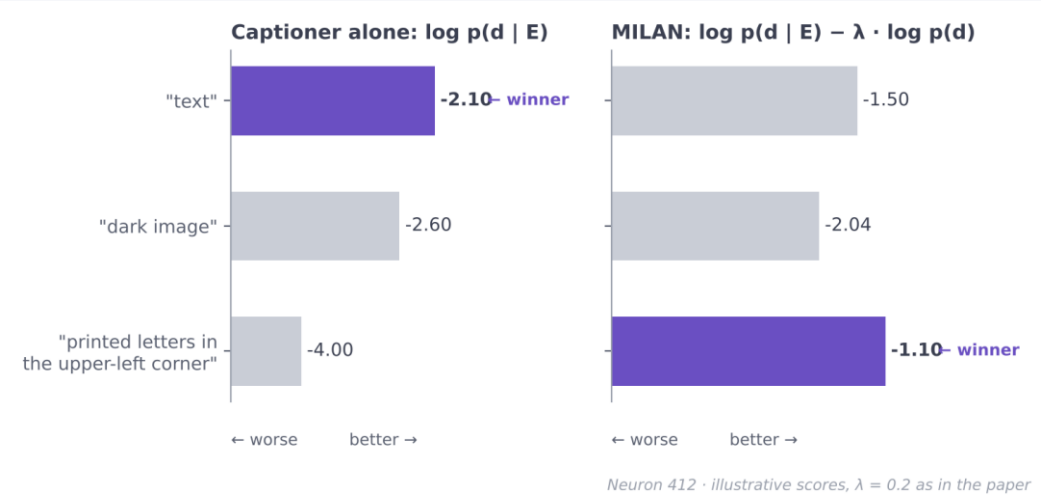
Task: describe **every neuron** of a vision model **in words**

How: 15 top-activating images + masks → a captioner trained on **60k human descriptions** → keep the **most specific** sentence

Output: neuron 412 → *"printed letters in the upper-left corner"*

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.

MILAN: picking the words



Captioner p(d | E): does the sentence fit *these* regions?

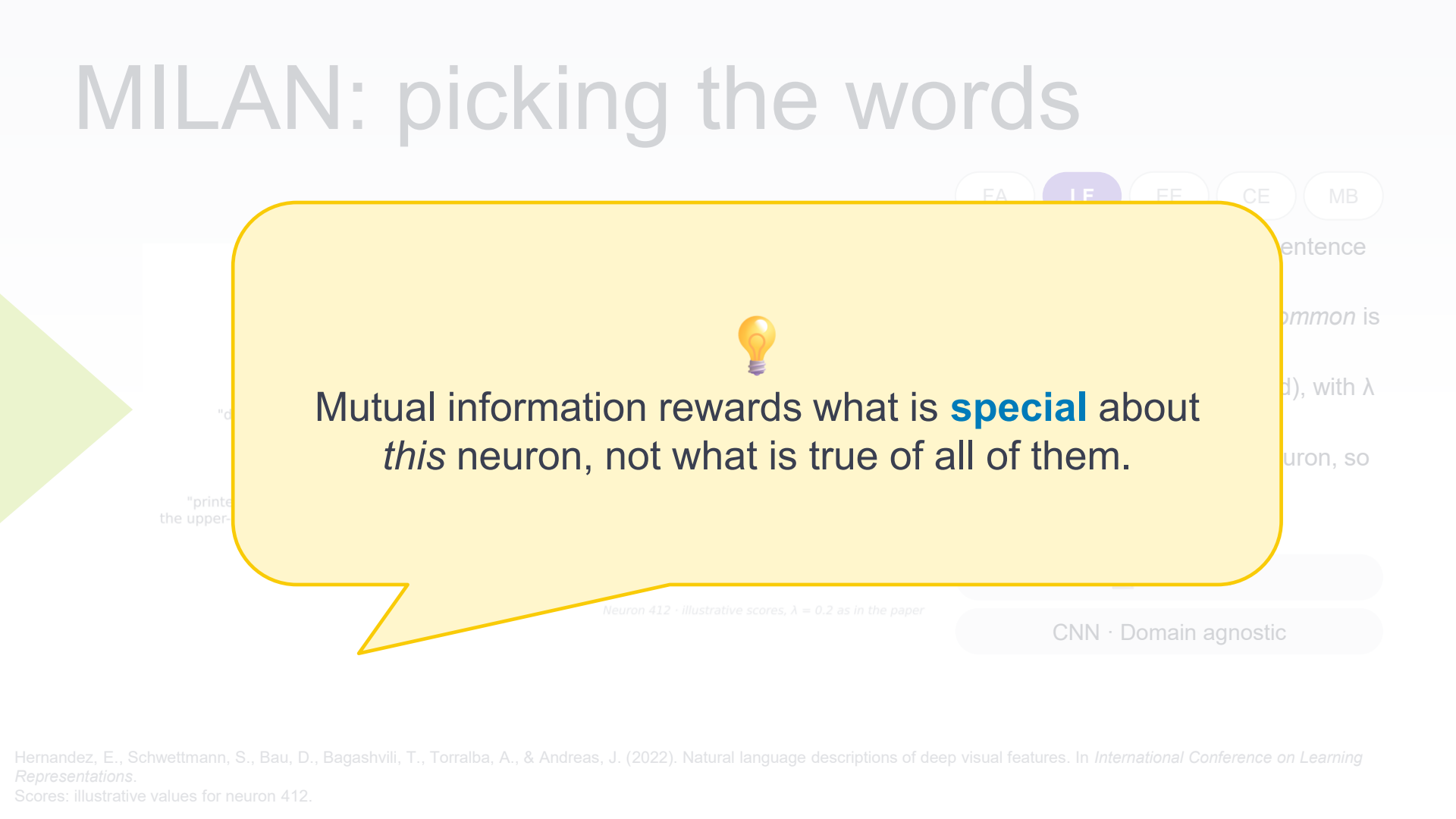
Language model p(d): how *common* is the sentence?

Score: log p(d | E) − λ · log p(d), with λ = 0.2

Generic sentences fit every neuron, so they **lose**

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.

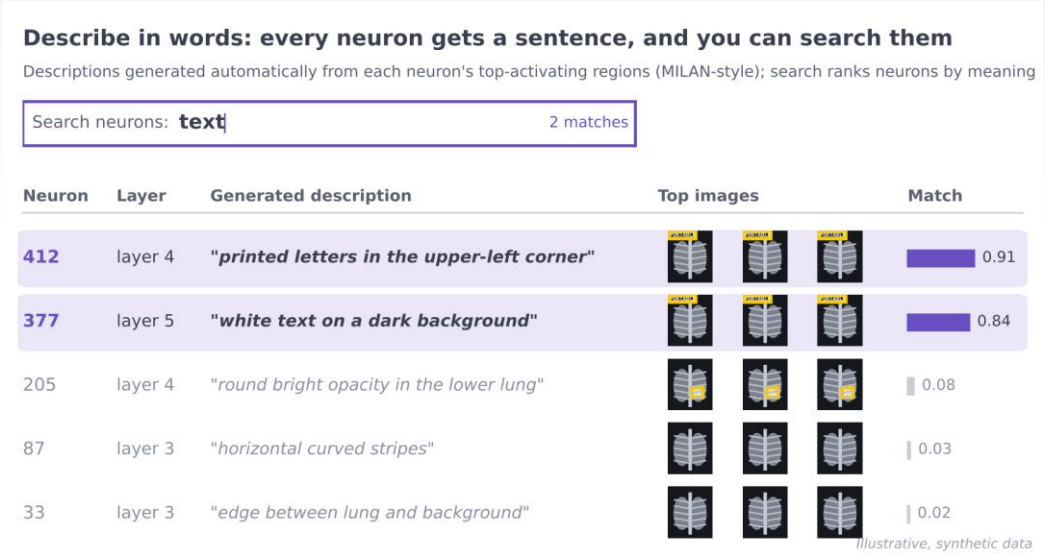
MILAN: picking the words



Mutual information rewards what is **special** about *this* neuron, not what is true of all of them.

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.
Scores: illustrative values for neuron 412.

MILAN: search and uses



Search: type *"text"* and get the neurons that detect it

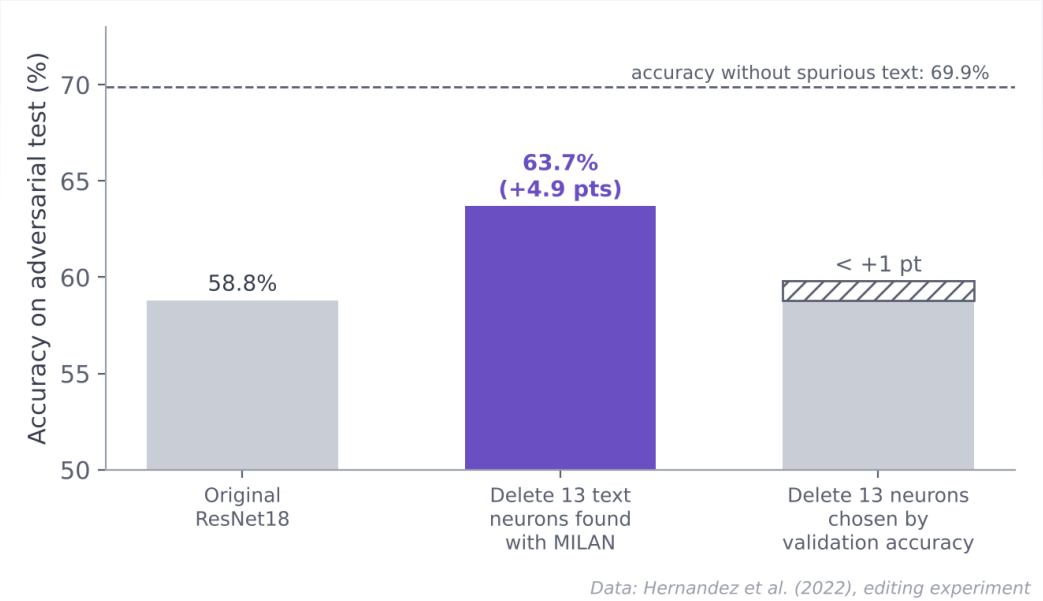
Analysis: which kinds of features each layer learns

Auditing: **213** face neurons in ImageNet models, **142** even with faces **blurred**

Editing: delete the neurons behind a shortcut (next slide)

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.

MILAN: deleting a text shortcut



Data: Hernandez et al. (2022), editing experiment

Setup: ResNet18 trained on images with **spurious text** that predicts the class

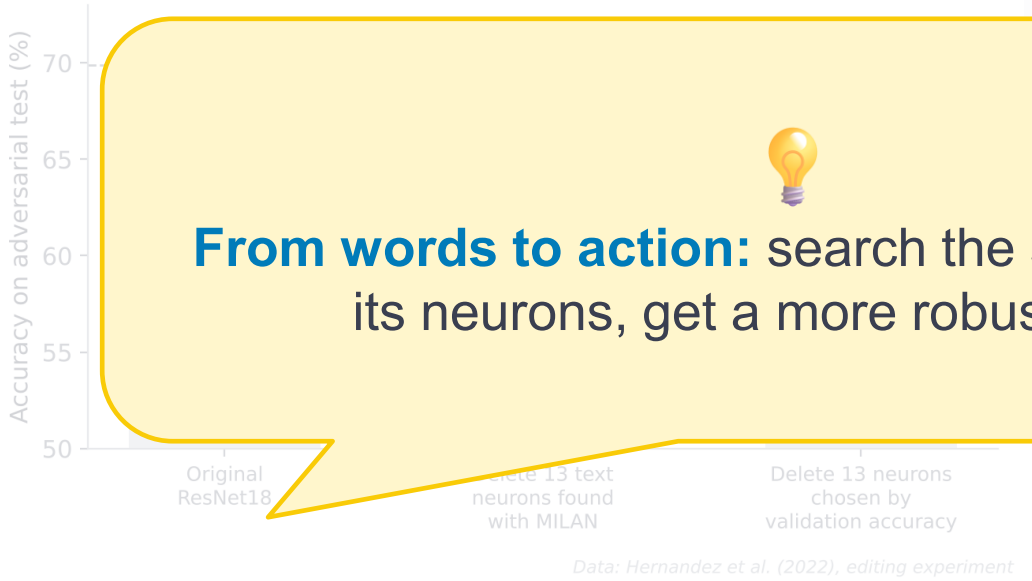
MILAN flags **300 of 1024** units as *"text"*

Deleting **13** of them: **+4.9 pts** on the adversarial test (−12% error)

Choosing neurons by validation accuracy: **< 1 pt**

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.

MILAN: deleting a text shortcut



From words to action: search the shortcut, delete its neurons, get a more robust model.

FA LF EE CE MB

images

s the

as *"text"*

on the

Architect

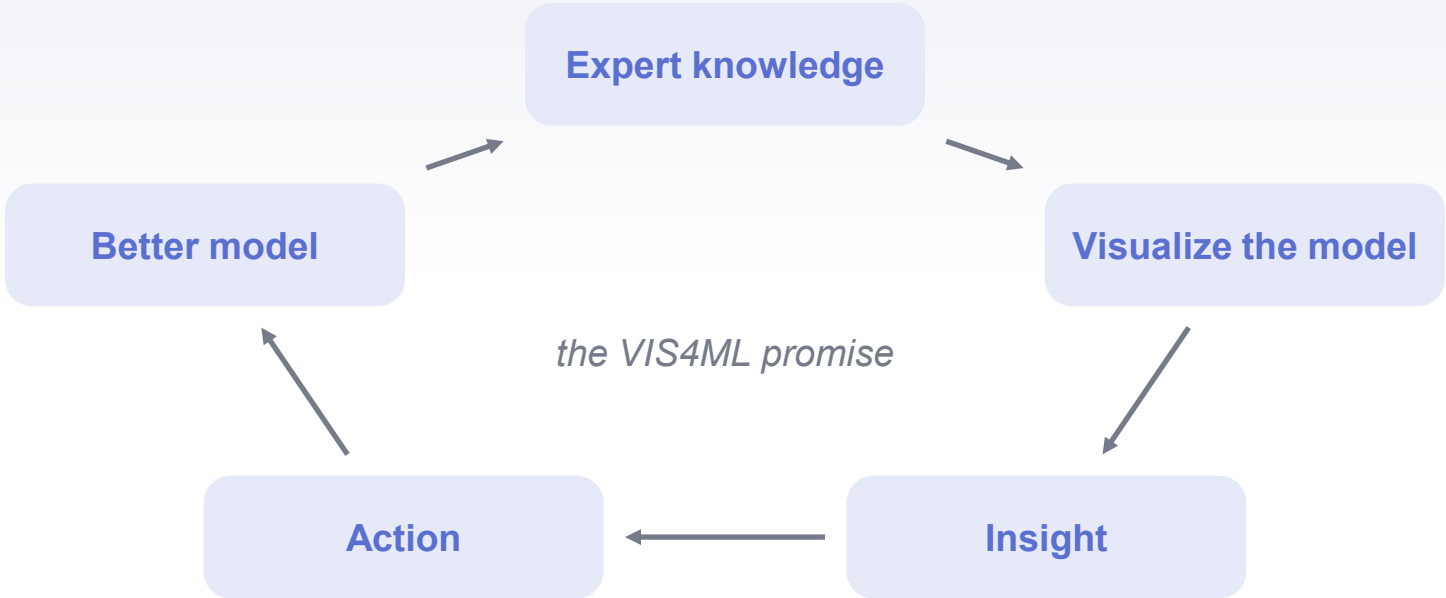
CNN · Domain agnostic

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.
Data from the editing experiment of the paper.

🤔

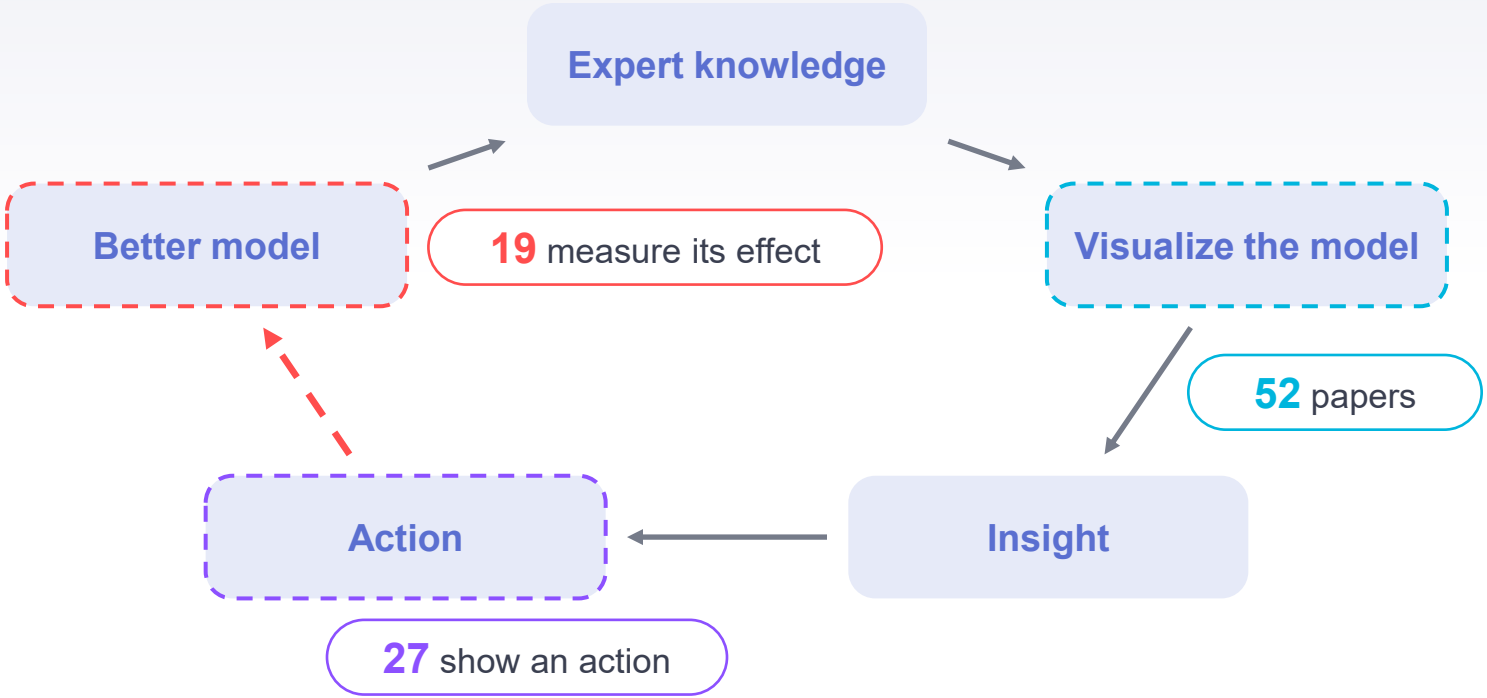
ALL THESE SYSTEMS LOOK GREAT… BUT DID ANY OF THEM **ACTUALLY IMPROVE** A MODEL?

The loop



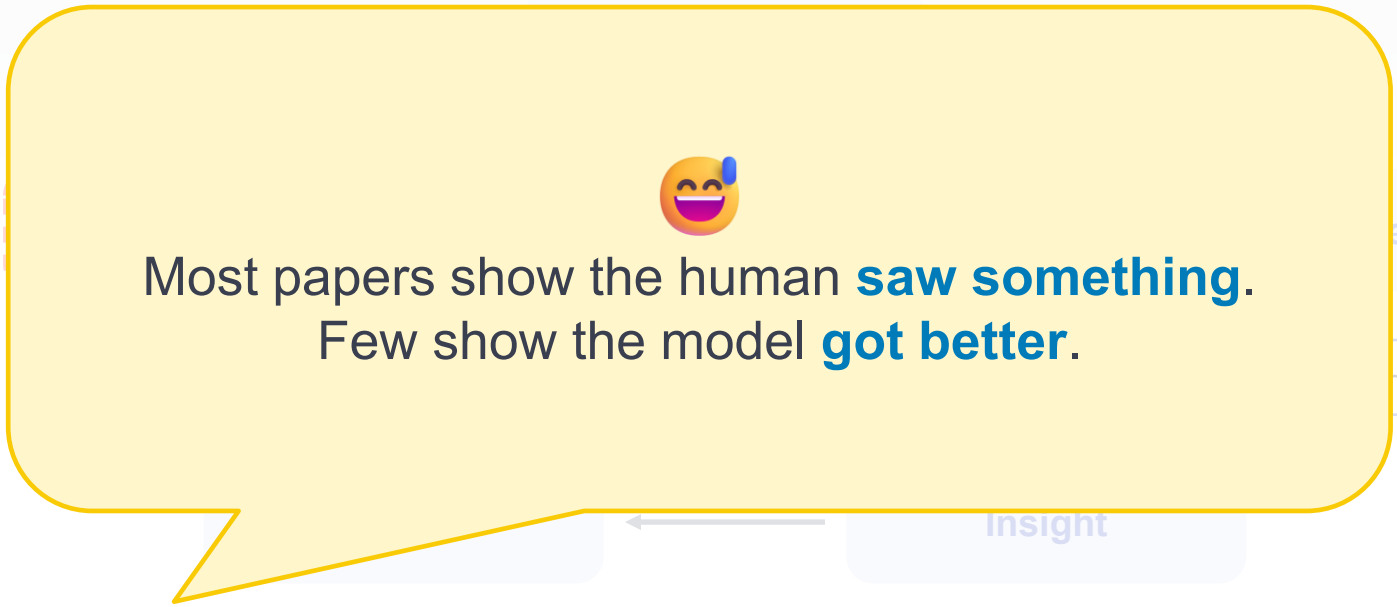
Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Closing the loop?



Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Closing the loop?



Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Who was in the loop?



50%

of participants are **ML experts**



5 / 52

papers combine **ML and domain experts**



17%

of papers use **ImageNet**, then CIFAR-10 and MNIST

Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Who was in the loop?



5

of p

exp

NIST



Existence proofs, not evidence: a handful of experts, toy datasets and **t-SNE in 1 of 3 papers**.

Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Before you believe a paper

1 · Who?

The users, and what they already knew

2 · Which data?

The datasets, and how far it scales

3 · What action?

What was changed in the pipeline

4 · What changed?

Model metrics before and after

5 · At what cost?

Time and effort to find the insight

6 · What claim?

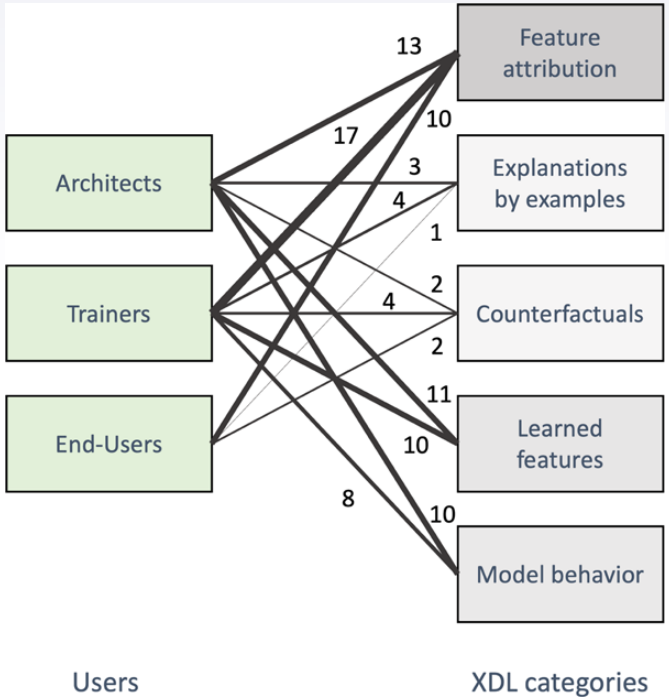
A design study, not a general law

Subramonyam, H., & Hullman, J. (2024). Are we closing the loop yet? Gaps in the generalizability of VIS4ML research. *IEEE Transactions on Visualization and Computer Graphics, 30*(1). arXiv:2308.06290

Some numbers

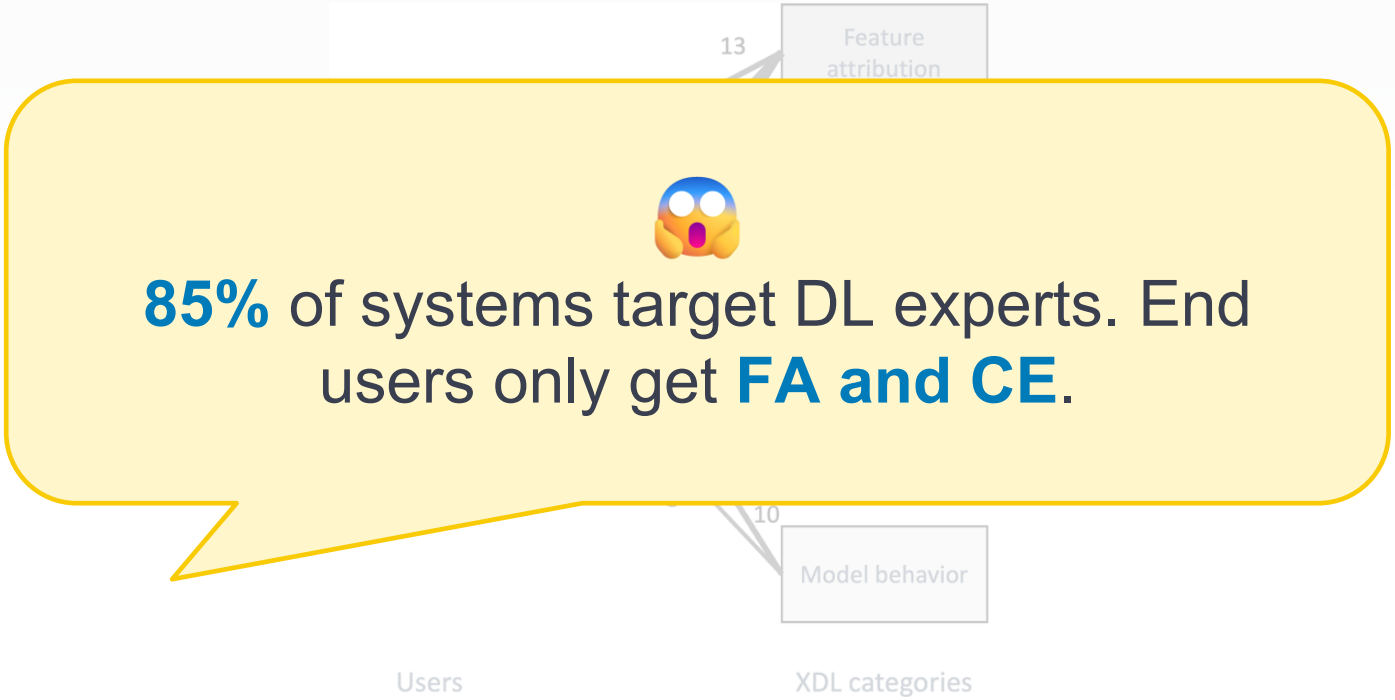
HOW IS THE SOTA

Who is it for?



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.2, Fig. 13)

Who is it for?

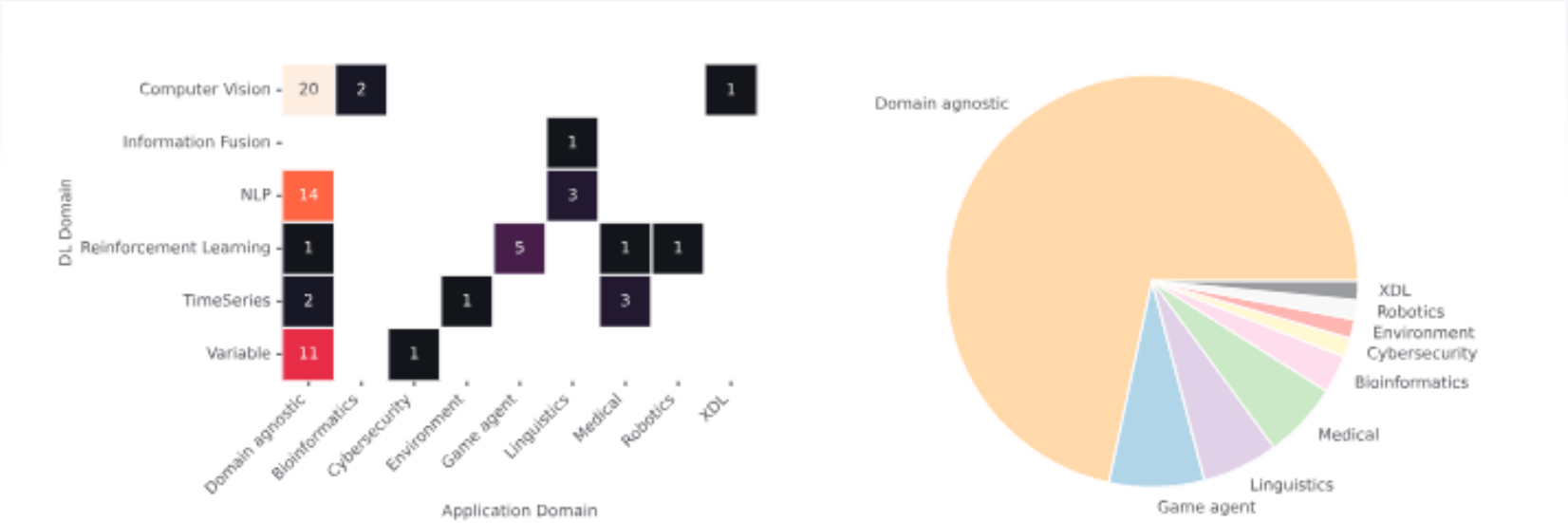


😱

85% of systems target DL experts. End users only get **FA and CE**.

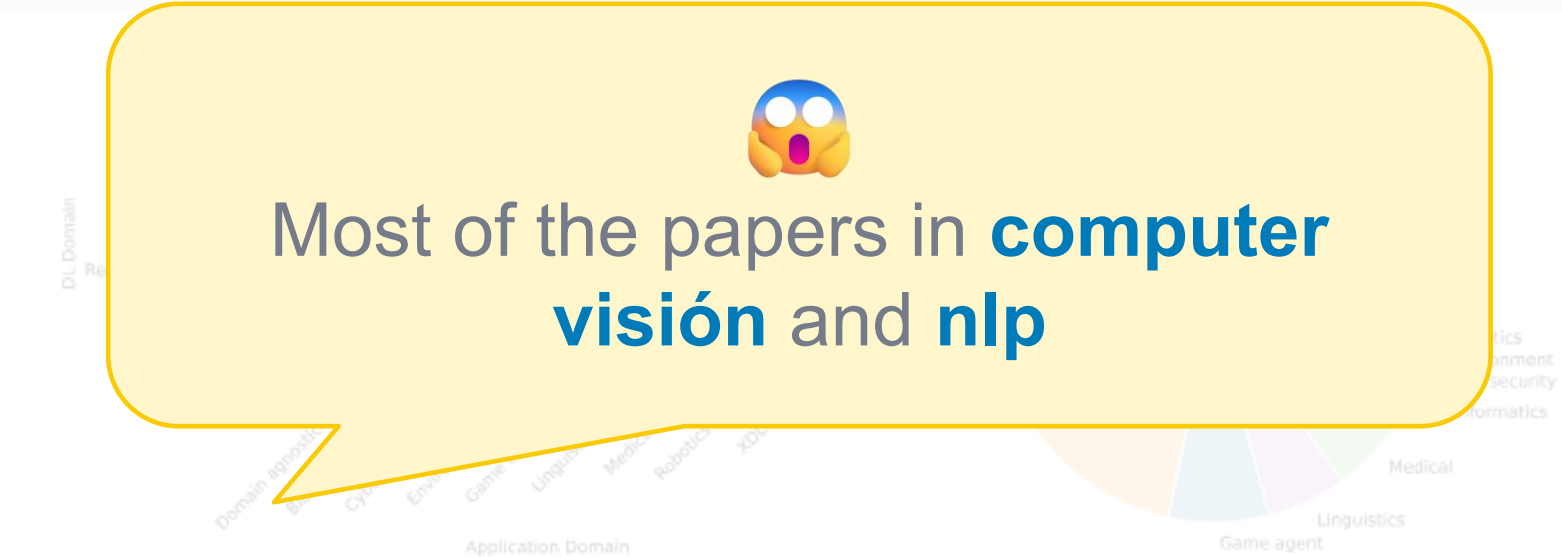
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.2, Fig. 13)

Where



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

Where

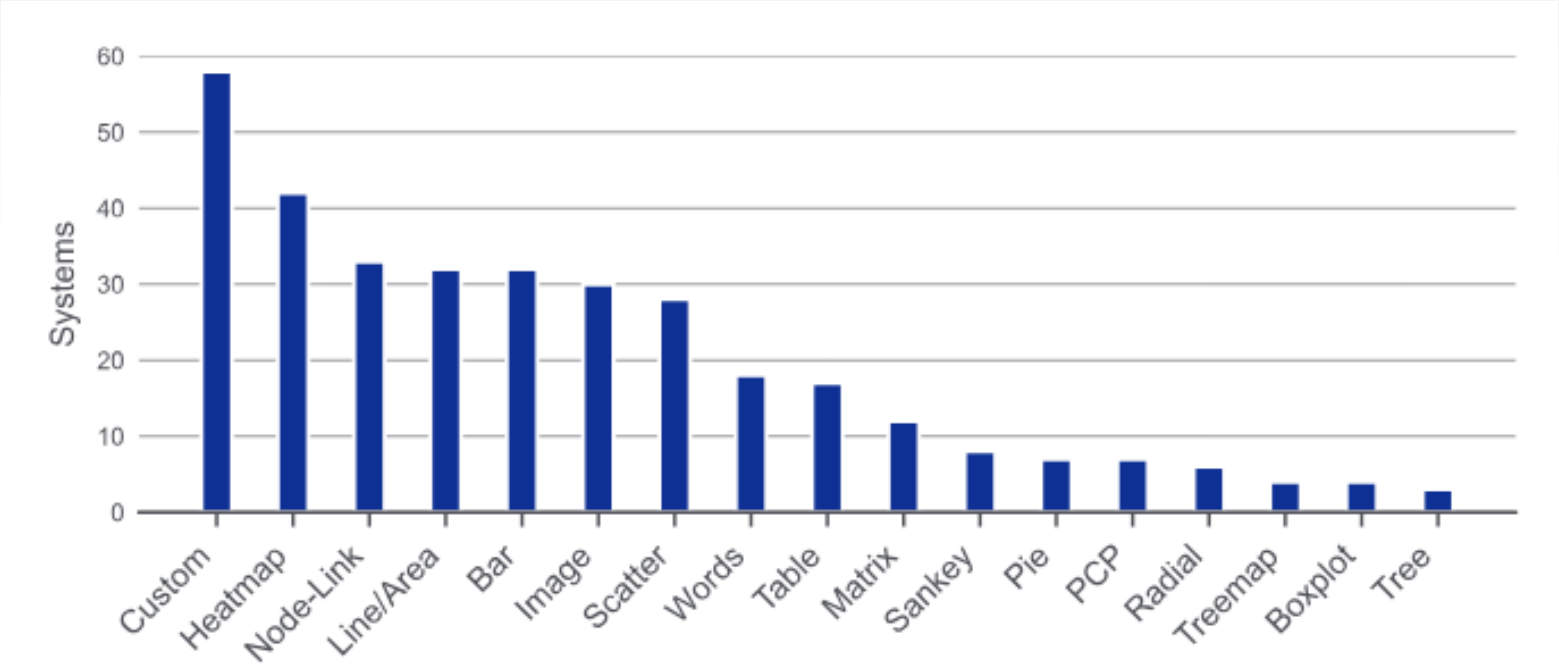


😱

Most of the papers in **computer visión** and **nlp**

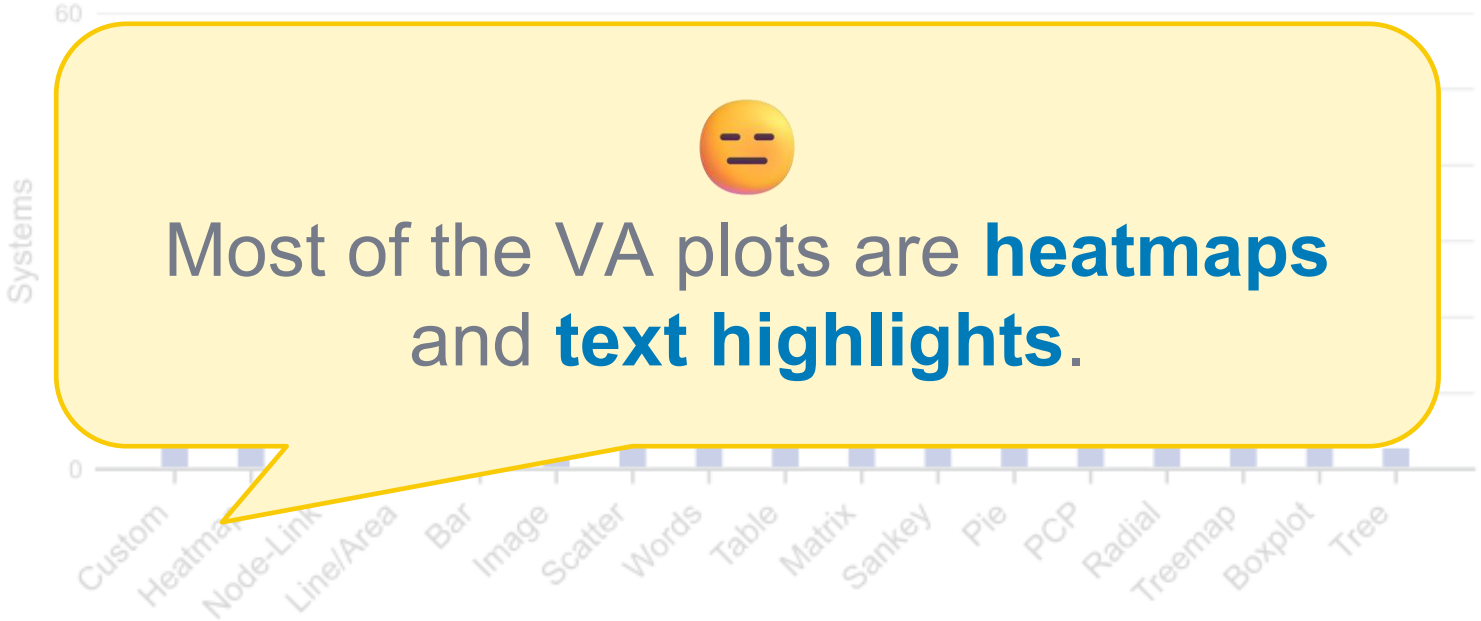
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

VA Techniques



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

VA Techniques

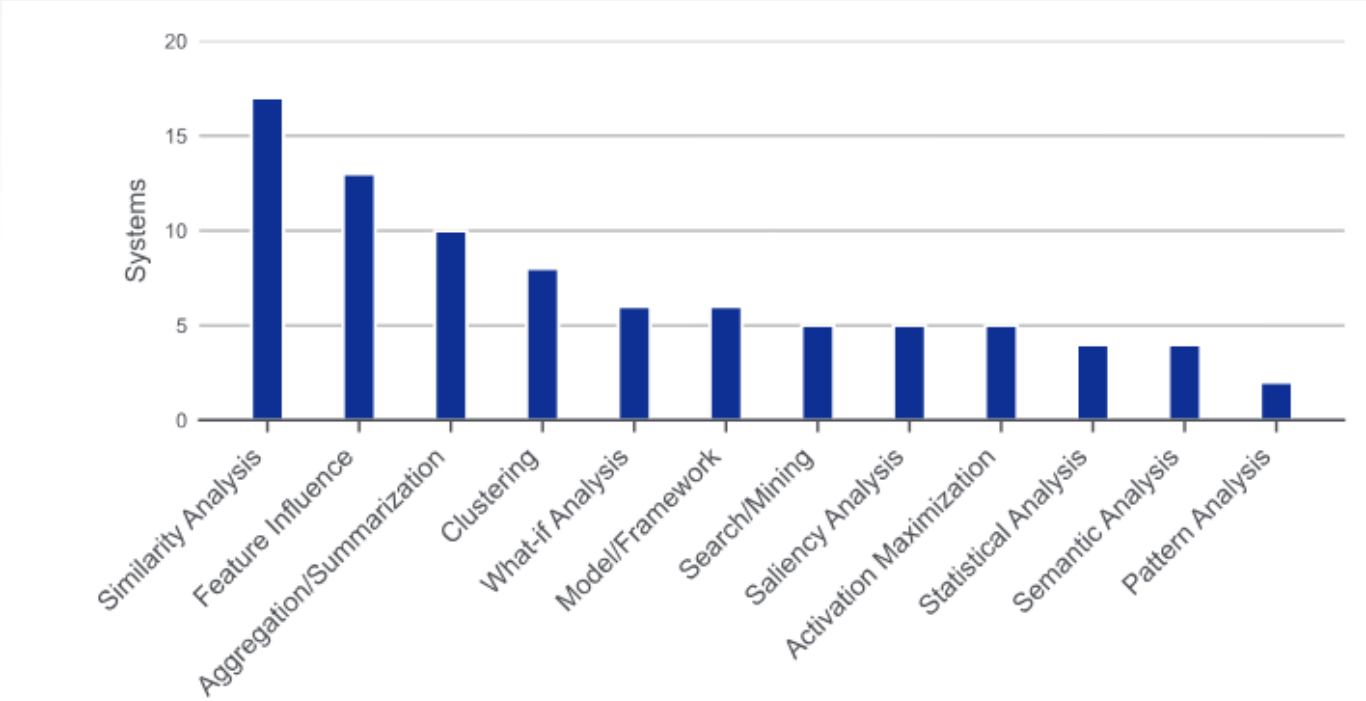


😑

Most of the VA plots are **heatmaps** and **text highlights**.

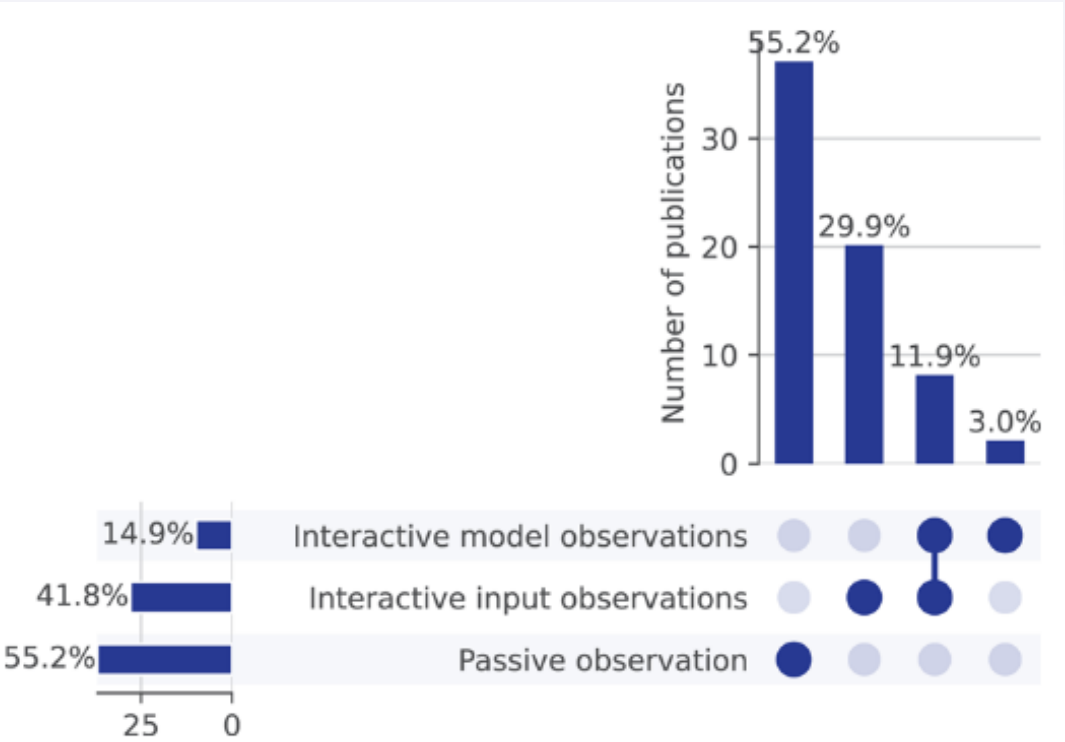
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

ML Techniques



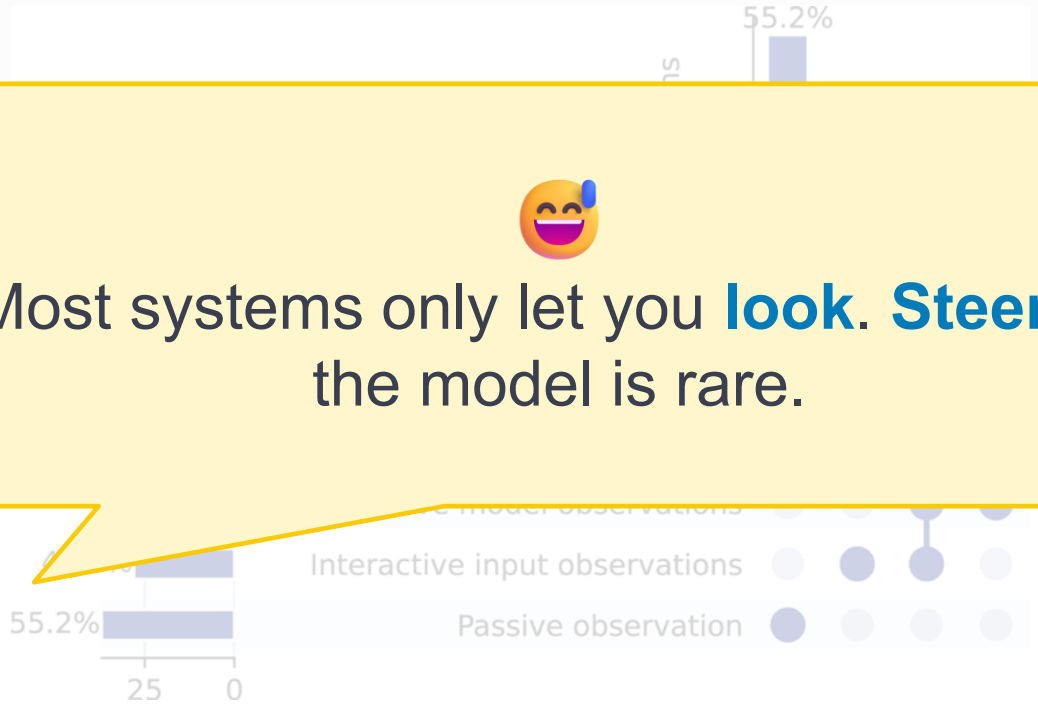
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

How interactive



La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

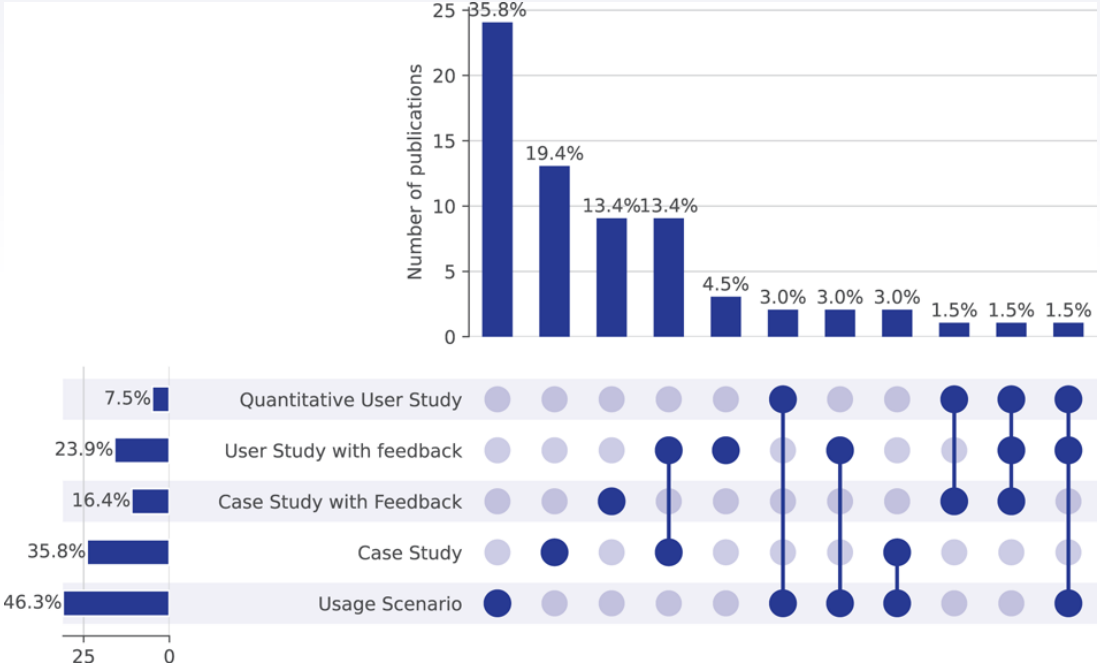
How interactive?



Most systems only let you **look**. **Steering** the model is rare.

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.3.3, Fig. 18)

How evaluated?



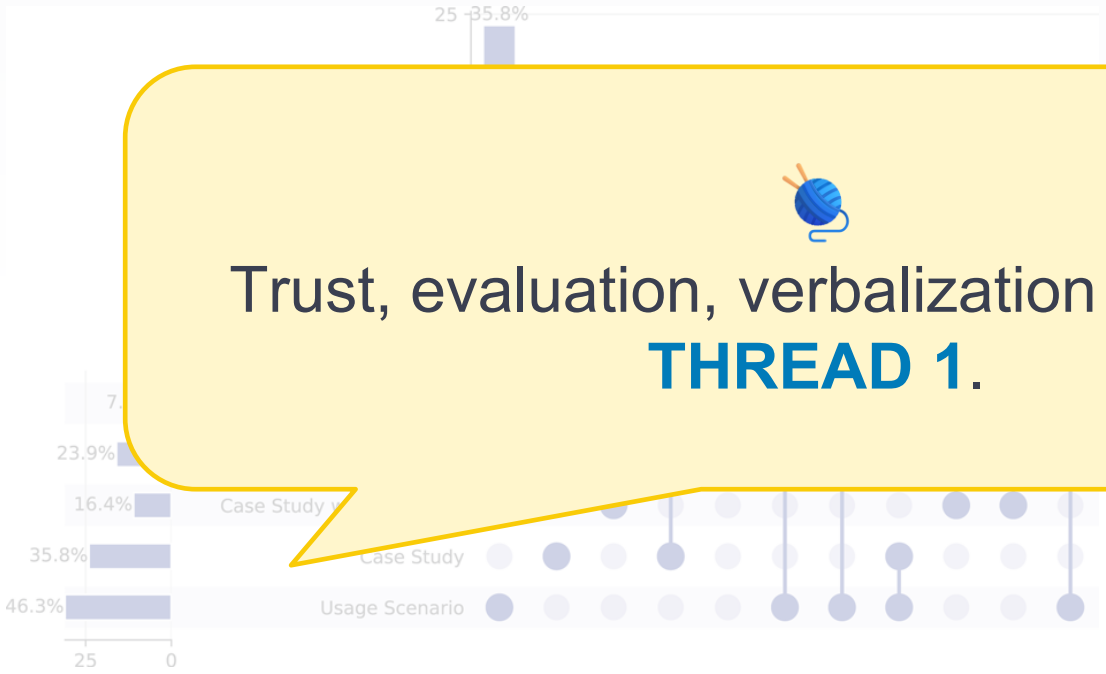
46%
usage scenario only

7.5%
quantitative user study

33%
release source code

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.4, §5.5, Fig. 19)

How evaluated?



Trust, evaluation, verbalization → back to **THREAD 1**.

33%
release source code

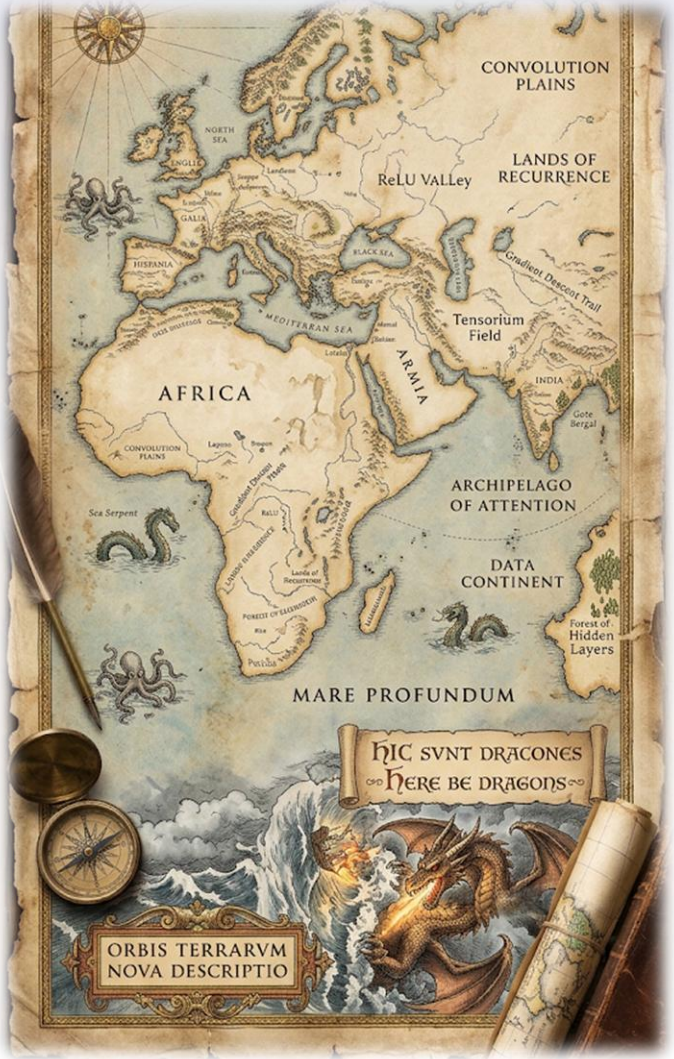
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., & Angelini, M. (2023). State of the art of visual analytics for eXplainable deep learning. *Computer Graphics Forum, 42*(1), 319–355. https://doi.org/10.1111/cgf.14733 (§5.4, §5.5, Fig. 19)

What La Rosa leaves out

Data: shortcuts, bias, wrong labels, drift

Training monitoring, also federated

Embeddings **not computed by DL**



🧶 THREAD 8 · DATA & TRAINING

The problem may come before the model (shortcuts, bias, wrong labels, drift) or during training, federated learning included.

- **Shortcuts and bias:** how do you detect them?
- **Wrong labels:** how do you find them?
- **Drift:** how do you monitor it?
- **Synthetic data:** how do you assess whether it resembles the real data?
- **Training:** what should you look at beyond the loss curve?
- **Federated learning:** what can you visualize without breaking privacy?



The 8 threads

1 · The human in DVA/XAI

An explanation counts only if it helps someone do something.

2 · DL4VIS & LLMs

DL and LLMs that generate, read and summarize visualizations.

3 · Self-explainable models

The explanation is built in; VA should let users steer it.

4 · Feature attribution

Mature, yet methods disagree and fail sanity checks.

5 · Learned representations

Projecting latent spaces, and what each projection distorts.

6 · Examples & counterfactuals

Similar cases and actionable changes; little VA work so far.

7 · Surrogate models

Readable mimics, only as good as their fidelity.

8 · Data & training

Problems before the model or during training; mostly VA, not XAI.