eXplainable Deep Learning

Romain Giot romain.giot@u-bordeaux.fr
15/09/2022

Deep Learning: Generalities

- Since 2012
  - Scientific and industrial revival on the use of Deep Learning
  - Computing resources (GPU) / Big data (storage) / End of manual extraction of characteristics
- Numerous applications and uses
  - Segmentation / Instance segmentation / Classification / Clustering / Dimension reduction / Data generation/ …
- Almost data agnostic
  - Any tensor (tabular data, images, videos, ...) / graphs / structured business data / text /

https://www.francaisauthentique.com/usine-a-gaz/

Deep Learning: Limitations

We consider 3 main limitations

- Technical limitations
- Legal limitations
- Acceptance limitation

Deep Learning: Technical Limitations

Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

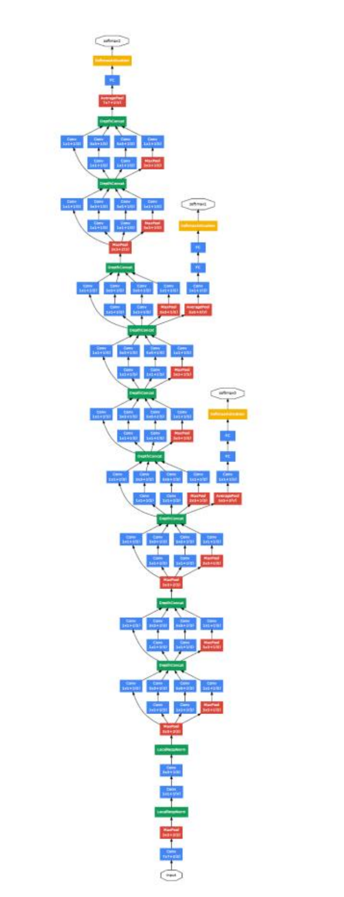
- Sensitive to data bias and attacks in operational environments
  - Do not always generalize
  - Easily attacked
- Need to have masses of data
  - The amount of data needed to learn a new model is astronomical / beyond the reach of a research lab
- The black box effect
  - Models are often oversized for their use

Deep Learning: legal limitations

Black box effect

- LGPD
  - **Whenever requested to do so, the controller shall provide clear and adequate information regarding the criteria and procedures used for an automated decision, subject to commercial and industrial secrecy.**
- Loi pour une république numérique
  - Art. L. 311-3-1. – Sous réserve de l’application du 2o de l’article L. 311-5, une décision individuelle prise sur le fondement d’un traitement algorithmique comporte une mention explicite en informant l’intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en oeuvre sont communiquées par l’administration à l’intéressé s’il en fait la demande. «Les conditions d’application du présent article sont fixées par décret en Conseil d’État.»
  - Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique
- GDPR
  - Article 13 RGPD. Informations à fournir lorsque des données à caractère personnel sont collectées auprès de la personne concernée /2/f/ l’existence d’une prise de décision automatisée, y compris un profilage, visée à l’article 22, paragraphes 1 et 4, et, au moins en pareils cas, des informations utiles concernant la logique sous-jacente, ainsi que l’importance et les conséquences prévues de ce traitement pour la personne concernée. https://gdpr-text.com/fr/read/article-13/

Deep Learning: Acceptance limitations



Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

- The black box effect
  - Models too deep/complicated to understand exactly what they do

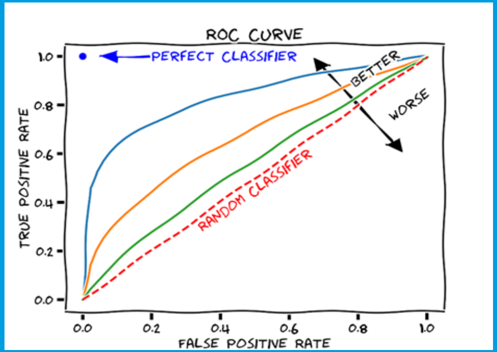
- Acceptance & Criticism of decisions
  - No guarantees offered on predictions
  - Adoption for health, safety, autonomous vehicles, etc. difficult

Deep Learning: Which solutions

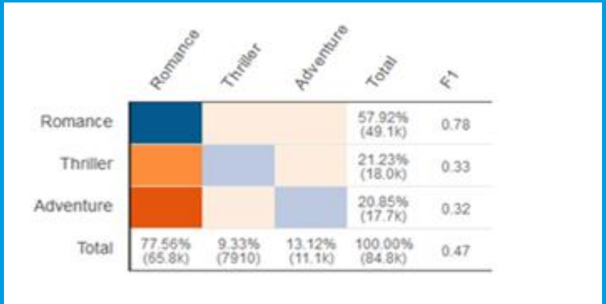
- **Technical limitations**
  - **Sensitive to data bias and attacks:** Improved learning, data collection and labeling methods, Creation of more robust models. (e.g. [AFGG17]), Addition of model checking mechanism (e.g. [G*19]), etc
  - **Need to have masses of data:** data augmentation, pre-learning + specialization, one-shot/few-shot learning, ...
  - **The black box effect: XDL**
- **Legal limits**
  - **The black box effect: XDL**
- **Acceptance limits**
  - **The black box effect and criticality of decisions: XDL**

[AFGG17] Aung, A. M., Fadila, Y., Gondokaryono, R., & Gonzalez, L. (2017). Building robust deep neural networks for road sign detection. arXiv:1712.09327.
[G*19] Goel, Akhil, et al. "DeepRing: Protecting deep neural network with blockchain", proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

eXplainable Deep Learning (XDL) goes beyond standard evaluation methods



Language	Precision	Recall	F-score	Accuracy
English	0.63	0.87	0.73	0.64
French	0.59	0.86	0.70	0.6
German	0.63	0.90	0.74	0.64
Spanish	0.58	0.89	0.70	0.60
Japanese	0.73	0.88	0.80	0.74

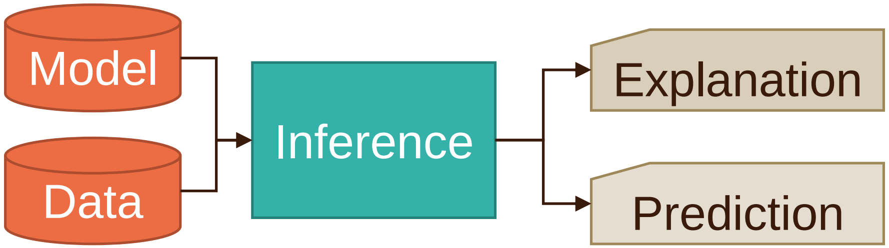


- Several aspects of importance
  - **Explainability**, interpretability, transparency,… [BA*20]
- Provide (visual) tools to interpret various aspects
  - Model, dataset, sample, …
- Intrinsic explainable model vs Posthoc analysis

[BA*I20] A. Barredo Arrieta *et al.*, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, Pages 82-115, 2020.

Self interpretable models can be a solution

- **Simple non deep learning models**
  - As replacement or student model
  - Decision Tree, Decision Rules, Linear classifier
- **Interpretable deep learning**
  - Modification of training procedure
  - Addition of explainable modules



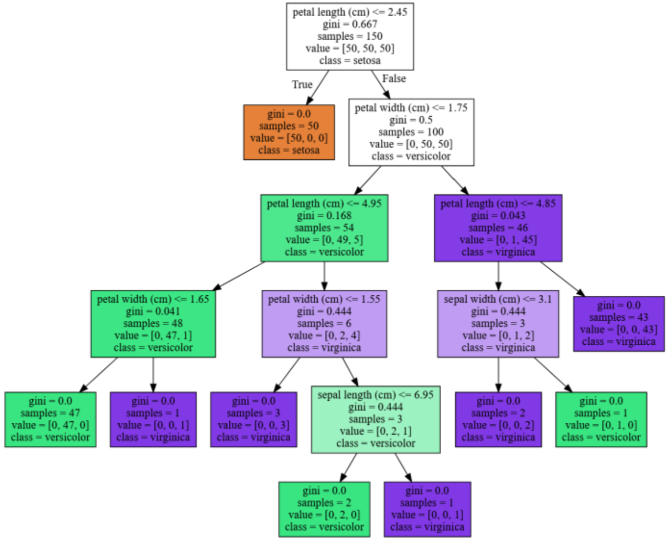
Interpretability: standard association rules representation

	Rules (supp, conf)
1	gill-attachment-f ⇒ veil-type-p (0.97415,1.0)
2	gill-spacing-c ⇒ veil-type-p (0.8385,1.0)
3	veil-color-w ⇒ veil-type-p (0.97538,1.0)
4	ring-number-o ⇒ veil-type-p (0.92171,1.0)
5	gill-attachment-f,veil-color-w ⇒ veil-type-p (0.97317,1.0)
6	gill-attachment-f,ring-number-o ⇒ veil-type-p (0.89808,1.0)
7	gill-spacing-c,veil-colo-w ⇒ veil-type-p (0.81487,1.0)
8	gill-attachment-f,gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81265,1.0)
9	veil-color-w,ring-number-o ⇒ gill-attachment-f,veil-type-p (0.8971,1.0)

Xu, Y., Li, Y., & Shaw, G. (2011). Reliable representations for association rules. *Data & Knowledge Engineering*, *70*(6), 555-575.

- **Idea**
  - **Depiction of the rules in a list**
  - **Can be accompanied of metrics**
- **Limits**
  - **Do not scale with the number of rules…**
  - **… but some strategies can help to reduce their amount**

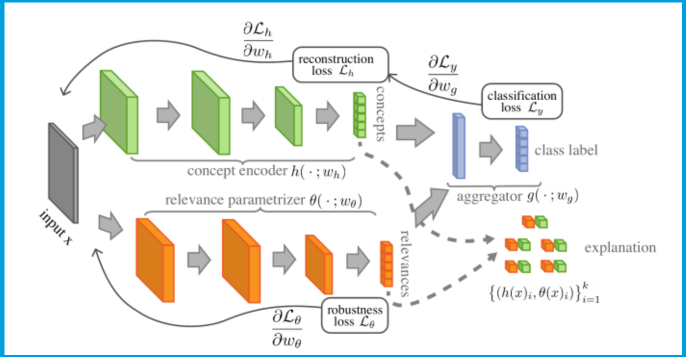
```
|--- feature_2 <= 2.45
|   |--- class: 0
|--- feature_2 >  2.45
|   |--- feature_3 <= 1.75
|   |   |--- feature_2 <= 4.95
|   |   |   |--- feature_3 <= 1.65
|   |   |   |   |--- class: 1
|   |   |   |--- feature_3 >  1.65
|   |   |   |   |--- class: 2
|   |   |--- feature_2 >  4.95
|   |   |   |--- feature_3 <= 1.55
|   |   |   |   |--- class: 2
|   |   |   |--- feature_3 >  1.55
|   |   |   |   |--- feature_0 <= 6.95
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_0 >  6.95
|   |   |   |   |   |--- class: 2
|   |--- feature_3 >  1.75
|   |   |--- feature_2 <= 4.85
|   |   |   |--- feature_1 <= 3.10
|   |   |   |   |--- class: 2
|   |   |   |--- feature_1 >  3.10
|   |   |   |   |--- class: 1
|   |   |--- feature_2 >  4.85
|   |   |   |--- class: 2
```



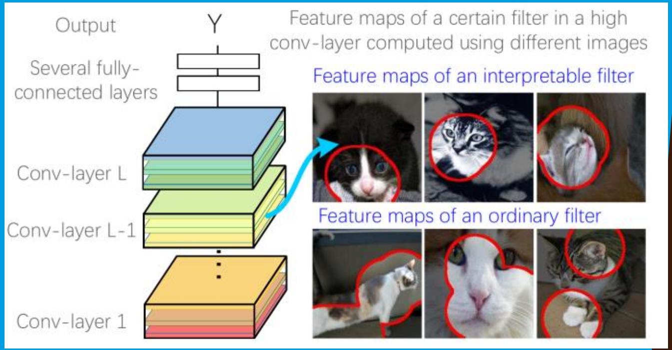
https://mljar.com/blog/visualize-decision-tree/

Interpretability: standard decision tree representation

- **Idea**
  - **Print the tree architecture …**
  - **.. or Draw the tree in a node-link diagram**
  - **Include additional information**
- **Limits**
  - **Do not scale with the size of the tree**



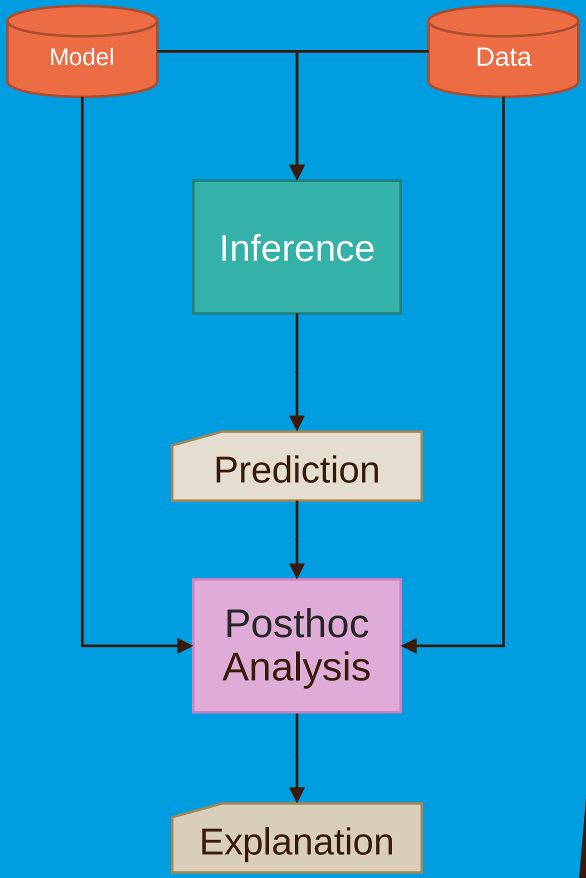
Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, *31*.



Zhang, Q., Wu, Y. N., & Zhu, S. C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8827-8836).

Interpretability: Interpretable deep neural networks

- **Idea**
  - **The network learns concepts while learning to solve the task**
  - **These concepts support the decision making and can be presented to the user**
- **Limitations**
  - **When automatically computed, concepts may be hard to be named**



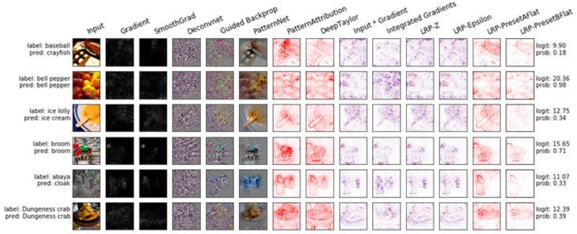
Posthoc Analysis

- **Local approaches**
  - **Feature attribution**
  - **Explanations by examples**
  - **Counterfactuals**
- **Global approaches**
  - **Feature attribution**
  - **Learned features**
  - **Model behavior**

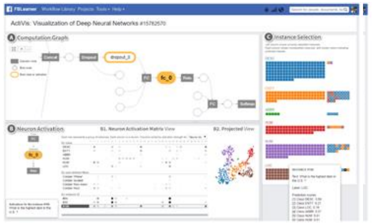
eXplainable Deep Learning (XDL)

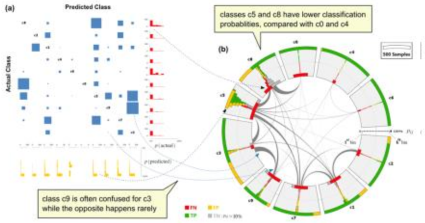
Several communities involved

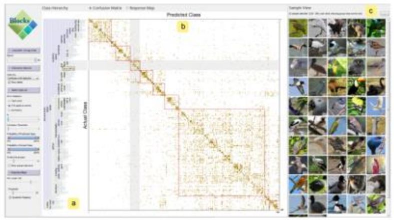
- Machine Learning



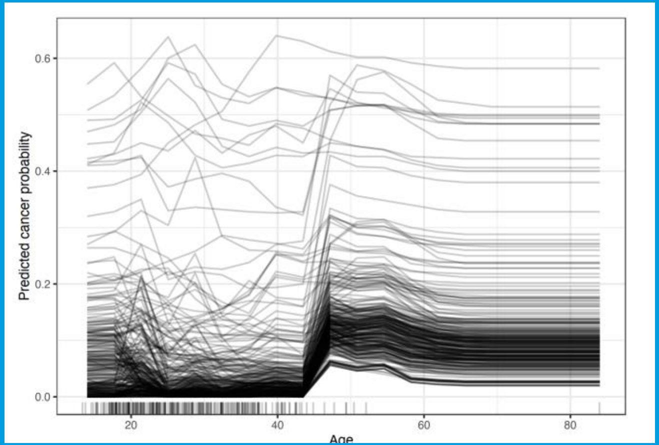
- Information Visualization







Feature Attribution: Individual Conditional Expectation



Input: the unique predictor values $x_{11}, x_{12}, \ldots, x_{1k}$;

Output: the estimated partial dependence values $\bar{f}_1(x_{11}), \bar{f}_1(x_{12}), \ldots, \bar{f}_1(x_{1k})$.

for $i \in \{1, 2, \ldots, k\}$ **do**

(1) copy the training data and replace the original values of x_1 with the constant x_{1i};

(2) compute the vector of predicted values from the modified copy of the training data;

~~(3) compute the average prediction to obtain~~ $\cancel{\bar{f}_1(x_{1i})}$.

end

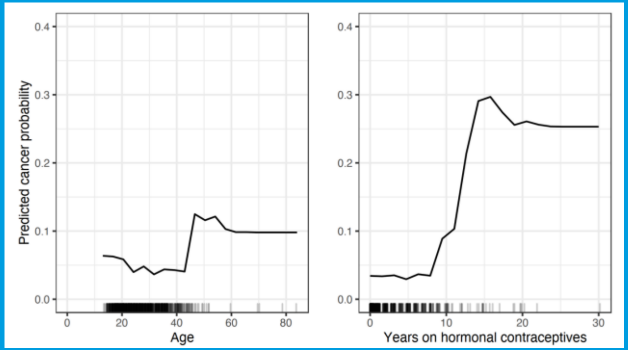
Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

- **Black box / global**
- **Aim**
  - **Show marginal effect of a feature on the predicted outcome of a model**
- **Idea**
  - **Iteratively replace, for all samples, each feature value by one of the domain**
- **Limitations**
  - **Number of selected features must be small (e.g. constrained to tabular data)**
  - **Features must be uncorrelated**

Feature attribution: Partial Dependency Plot

Input: the unique predictor values $x_{11}, x_{12}, \ldots, x_{1k}$;
Output: the estimated partial dependence values $\bar{f}_1(x_{11}), \bar{f}_1(x_{12}), \ldots, \bar{f}_1(x_{1k})$.

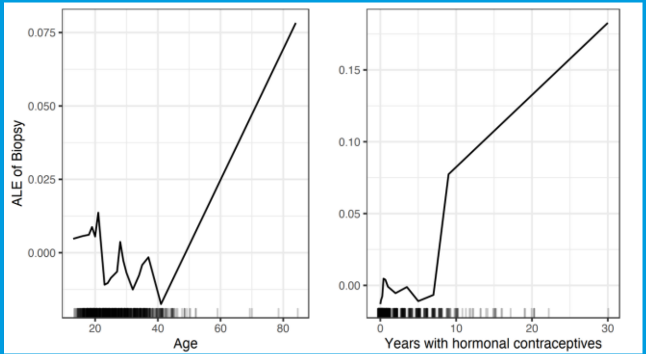
for $i \in \{1, 2, \ldots, k\}$ **do**
  (1) copy the training data and replace the original values of x_1 with the constant x_{1i};
  (2) compute the vector of predicted values from the modified copy of the training data;
  (3) compute the average prediction to obtain $\bar{f}_1(x_{1i})$.
end



Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

- Black box / local
- Similar to Individual Conditional Expectation
  - BUT depict the average instead of all samples

Feature attribution: Accumulated Local Effect Plot

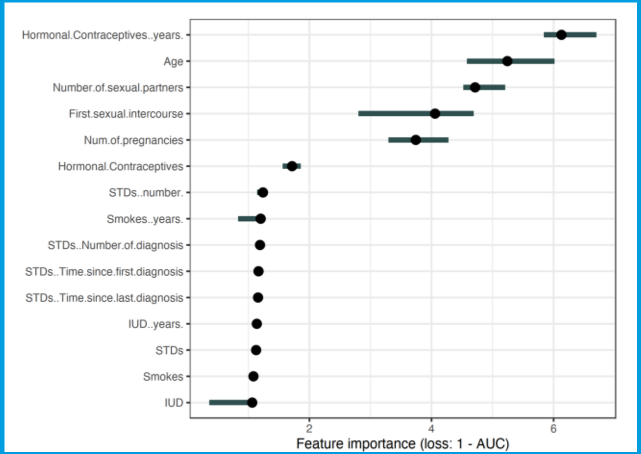


https://christophm.github.io/interpretable-ml-book/ale.html

- **Black box / global**
- **Aim**
  - **Describe how features influence the prediction of a machine learning model on average**
- **Idea**
  - **Compute the output difference when replacing a feature by its local extremums**
- **Limitations**
  - **Number of selected features must be small (e.g. constrained to tabular data)**
  - **Quantiles are used to discretize feature space (bins are of different width)**

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059-1086.

Feature attribution: Permutation Feature Importance



https://christophm.github.io/interpretable-ml-book/feature-importance.html

- **Black box / global**
- **Aim**
  - **Measures the increase in the prediction error of the model after permutation of the feature's values**
- **Idea**
  - **Permutes feature value over samples and compute feature importance by comparing obtained error rate with initial one**
- **Limitations**
  - **Number of selected features must be small (e.g. constrained to tabular data)**

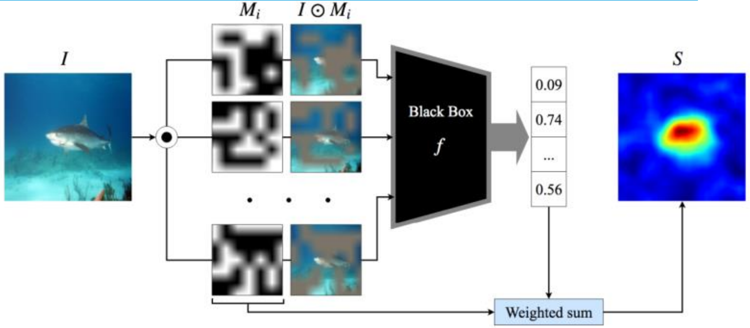
Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res., 20(177), 1-81.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Feature attribution: Lime

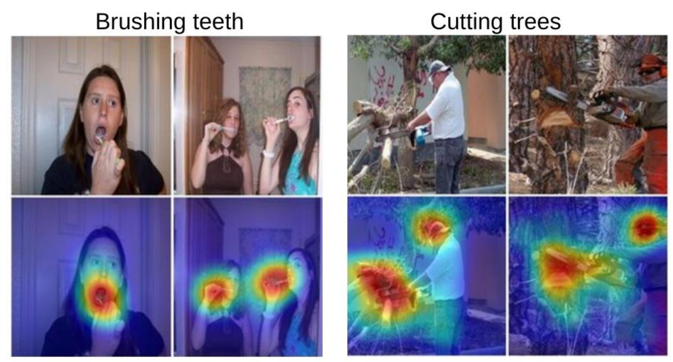
- Black box / local
- Idea
  - Learns a local surrogate (and interpretable) model to explain a specific instance
  - Relies on the removal of input features to generate neighbors
  - Scales on images by using superpixels
- Limitations
  - Removal of input features is not well defined
  - Method is not stable

Feature attribution: RISE

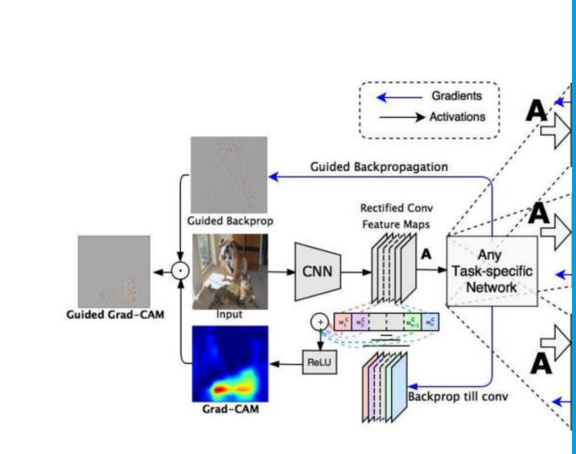


- Black box / local
- Idea
  - Relies on random masks to generate neighbors
  - Scales by generating low resolution masks
  - Masks contribution depends on model output
- Limitations
  - Limited to image classification

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.



Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

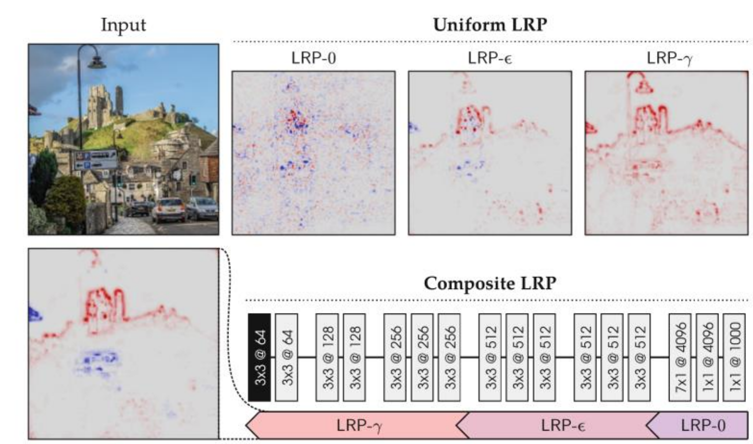


Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Feature attribution: Class Activation Mapping-based methods

- **White box / local**
- **Idea**
  - **Focuses on the last pooling layer before the first fully connected layer**
  - **Relies on activations (and eventually gradients)**
- **Limitation**
  - **CAM requires a network modification (not GRADCAM)**
  - **Fails to localize full object**
  - **Explanation resolution is low**

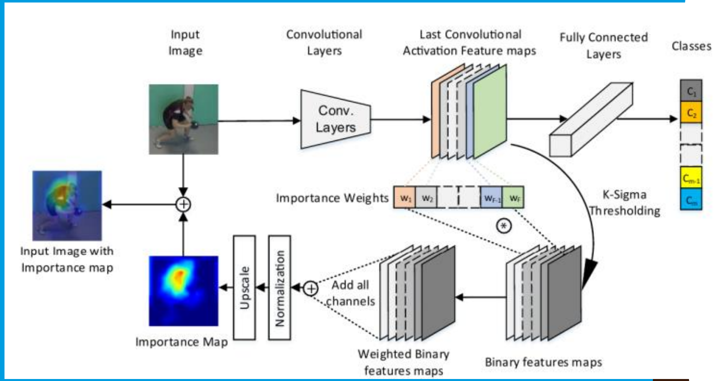
Feature attribution: Layerwise Relevance Propagation



- **White box / local**
- **Idea**
  - **Rule-based system to propagate relevance from output to input**
  - **Heterogeneous rules can be combined**
  - **Explanation resolution is high**
- **Limitations**
  - **Architecture has to be compatible with rules**
  - **Rules can be complex to configure**

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209.

Feature attribution: FEM

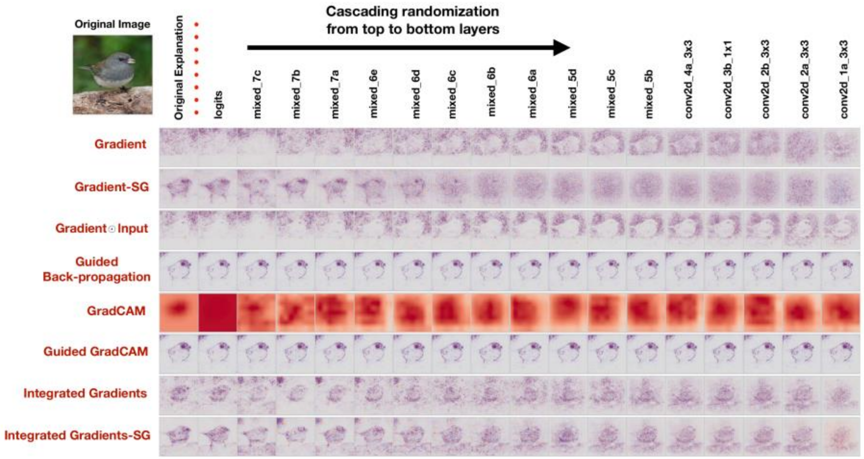


- **White box / local**
- **Idea**
  - **Focuses on the latest convolutional result**
  - **Relies ONLY on activation values**
  - **Evaluation shows a better correspondence with gaze fixation density maps than gradcam**
- **Limitations**
  - **Low resolution (a workaround is to use Multi-Layered FEM)**

Fuad, K. A. A., Martin, P. E., Giot, R., Bourqui, R., Benois-Pineau, J., & Zemmari, A. (2020, November). Features understanding in 3D CNNS for actions recognition in video. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.

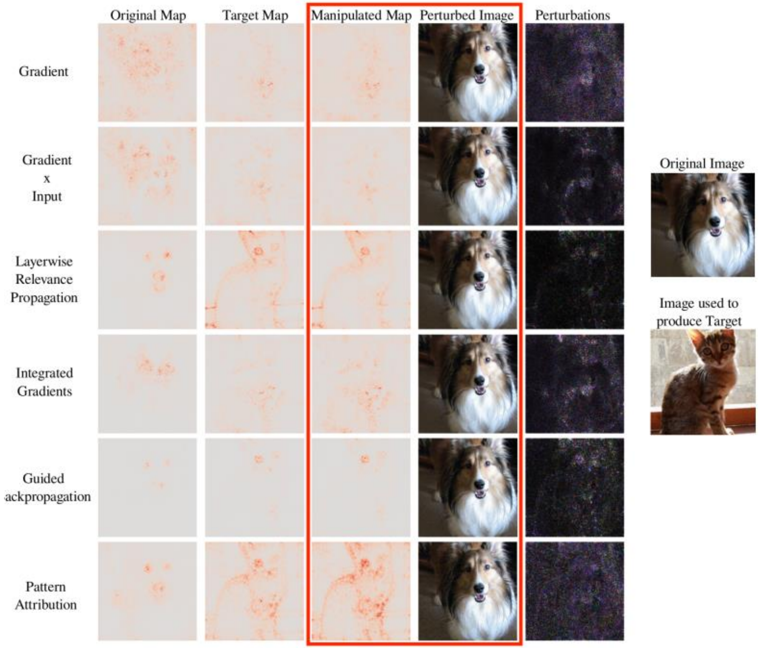
Feature attribution has still some limitations

The modification of the network may have few impact on the explanation



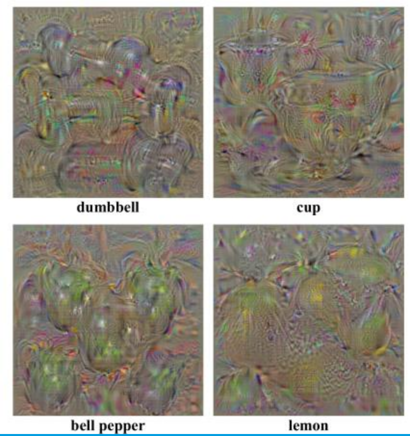
Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Explanations can be forged

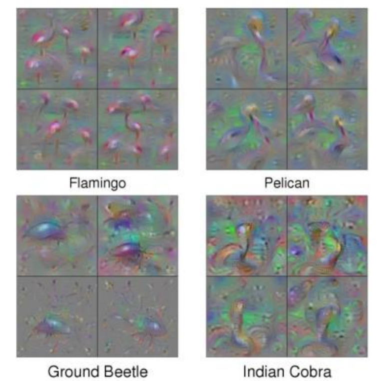


Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, *32*.

Learned Features: Class-score Maximization



K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. International Conference on Learning Representations Workshop, 2014.

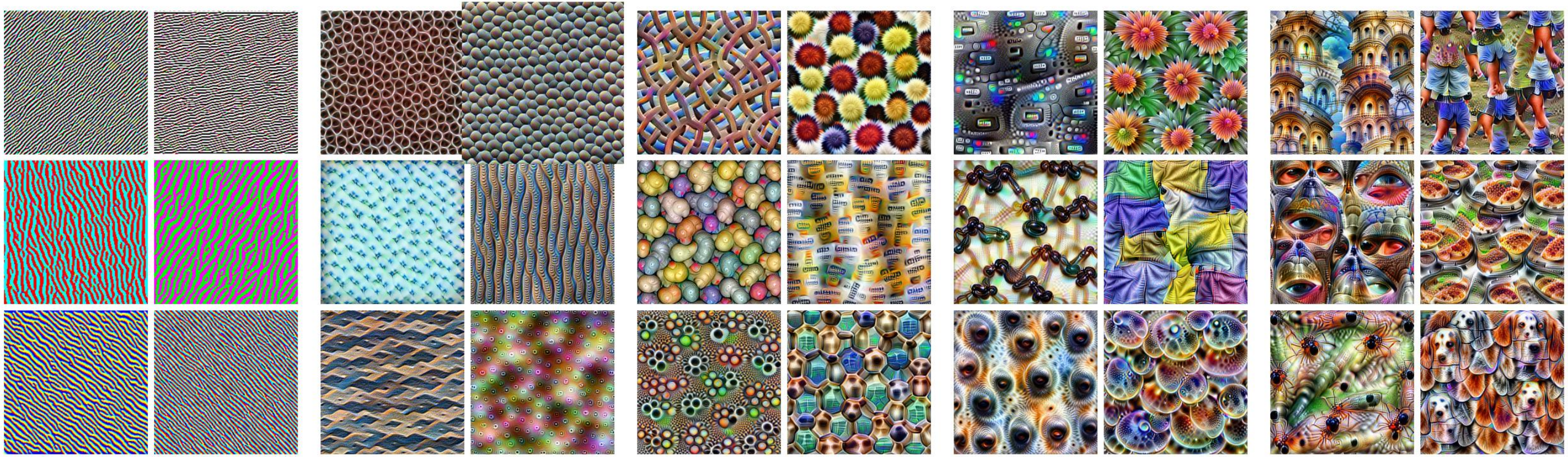


J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson. Understanding neural networks through deep visualization. International Conference on Machine Learning Deep Learning Workshop, 2015.

- **White box / global**
- **Idea**
  - **Generation of an artificial input image**
  - **Optimization of the class score**
- **Limitation**
  - **Final result is completely out of distribution**

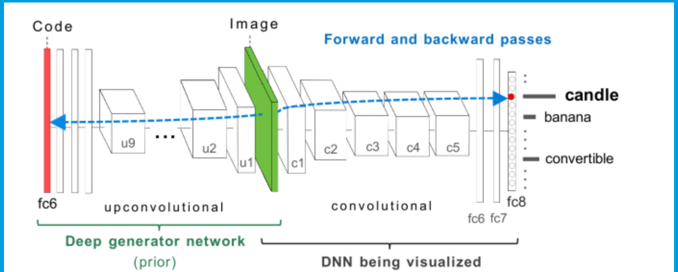
Learned Features: Feature Visualization

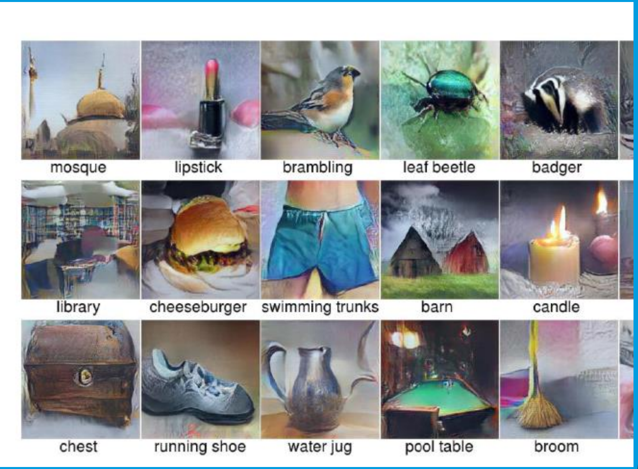
A step beyond



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.



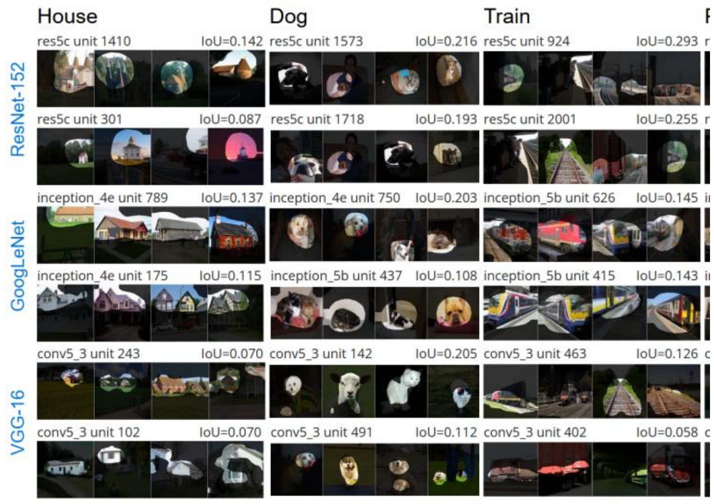


Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, *29*.

Learned features: data synthesis

- **White box / global**
- **Idea**
  - **Relies on deep generator to improve realism of generated samples**
  - **Optimization of the embedding of a trained network**

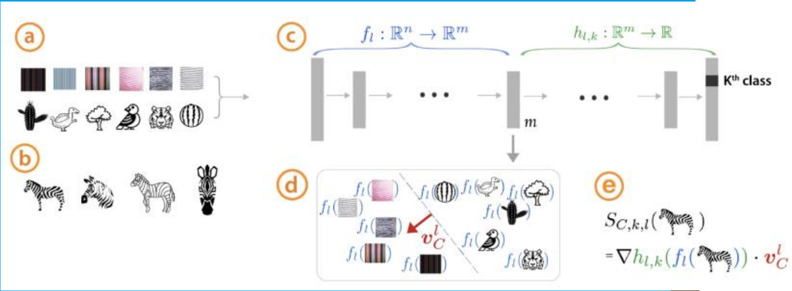
Learned Features: Network dissection



- **White box / global**
- **Idea**
  - **Identifies human-labeled visual concepts**
  - **Gathers hidden variables response to known concepts**
  - **Qualifies alignment of hidden variable/concept pairs**
- **Limitations**
  - **Limited to convolutional layers**

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541-6549).

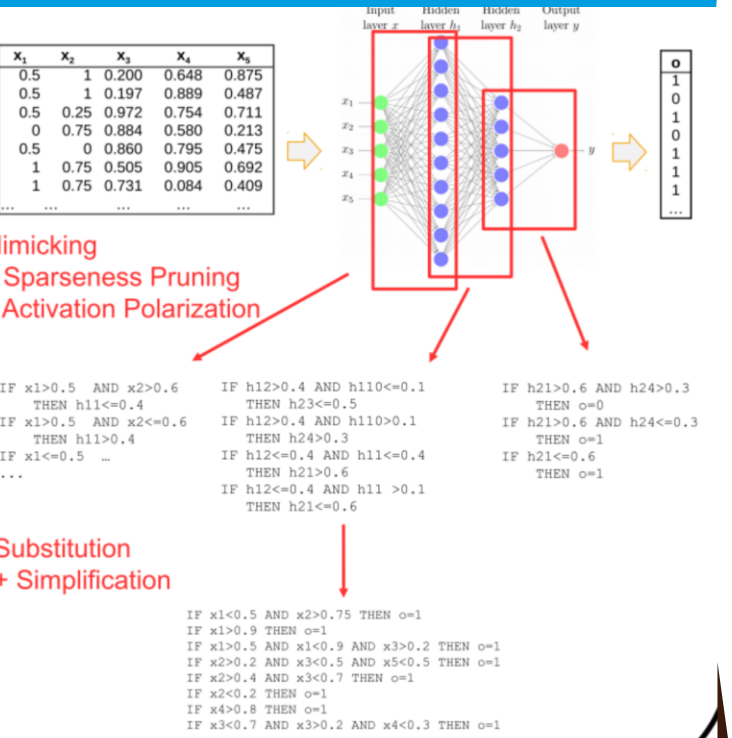
Learned Features: Concept Activation Vectors



- White box / local
- Idea
  - Identifies concepts related to classes
  - Computes relative proximity between samples and concepts
- Limitations
  - Limited to concepts explicitly trained (e.g. properly defined and with training data)

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668-2677). PMLR.

Learned features : rules extraction



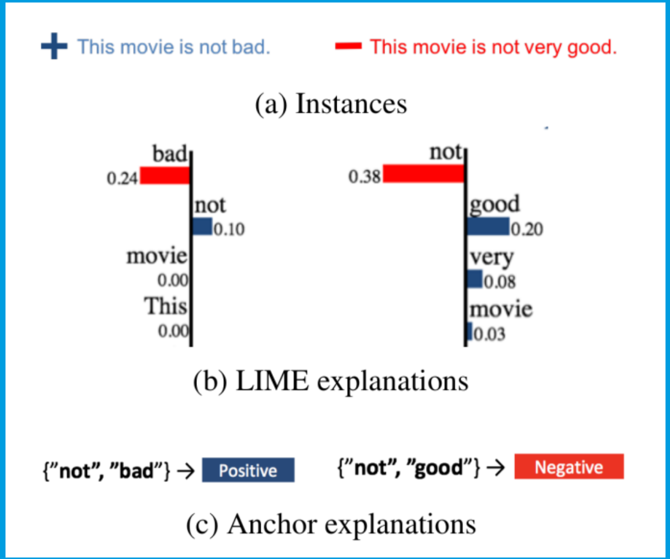
- **White box / global**
- **Idea**
  - **Extract global rules from the network**
  - **Display them**
- **Limitations**
  - **Few proposals in the literature.**
  - **Seem to focus on tabular data with few features**

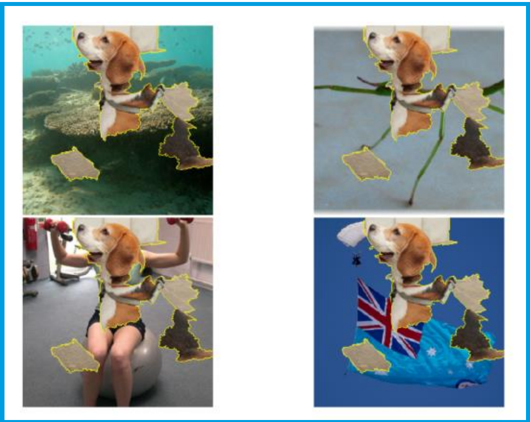
Zilke, J. R., Loza Mencía, E., & Janssen, F. (2016, October). Deepred–rule extraction from deep neural networks. In *International conference on discovery science* (pp. 457-473). Springer, Cham.

Explanation by example: Influential instances

- Black box / global
- Idea
  - Influential instances used for training have a strong impact on the model performance
  - We expect to have few influential instances in the training set to trust the model
  - Strategy: remove samples from training data and observe difference in retrained model

https://christophm.github.io/interpretable-ml-book/influential.html

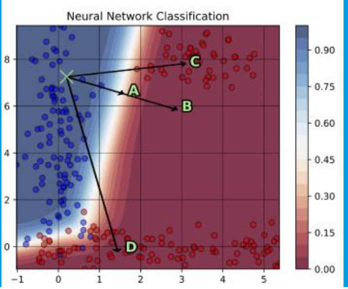




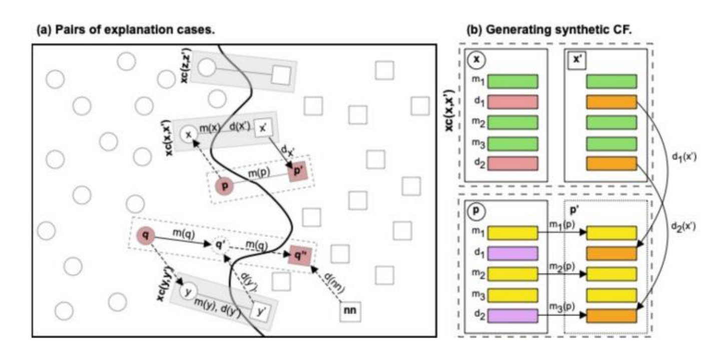
Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. "Anchors: high-precision model-agnostic explanations". AAAI Conference on Artificial Intelligence (AAAI), 2018

Explanation by example: anchors

- Black box / local
- Idea
  - If-then rules
  - Modification of other features of the anchor has no impact on the prediction
- Limitations
  - Lots of parameters
  - Computationally intensive
  - Anchors at the boundary decision are complex
  - To compute the domain distribution may be complex



Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020, February). FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 344-350).



Keane, M. T., & Smyth, B. (2020, June). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In International Conference on Case-Based Reasoning (pp. 163-178). Springer, Cham.

Explanation by example: counterfactuals

- White and black box / local
- Idea
  - Answers the question: given a classifier and an observation, what is the closest sample with another groundtruth
  - Generation of a random sample that minimize a loss (low distance with sample to explain + expected prediction)
  - Case Base Reasoning
- Limitation
  - Sensitive to the Rashomon effect (many different counterfactuals can be generated)
  - Out Of Distribution

Explanation by example: counterfactuals

- **Criteria for a good counterfactual**
  - **(reduced) Prolixity (search for the minimal changes)**
  - **Sparsity (few features modified)**
  - **Plausibility (data points in the domain)**

Model behavior: Rewriting a model

- **Idea**
  - **Generators are composed of rules with a specific semantic**
  - **To rewrite a model needs to understand the rules**
  - **think of the layer as a memory that associates keys to values.**
- **Limitation**
  - **Approach for generative models only**
  - **Explanation is not the key of approach**

Bau, D., Liu, S., Wang, T., Zhu, J. Y., & Torralba, A. (2020, August). Rewriting a deep generative model. In *European conference on computer vision* (pp. 351-369). Springer, Cham.

Model behavior: generalities

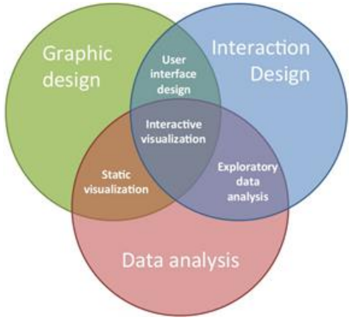
In opposite to static data and visualization from previous slides

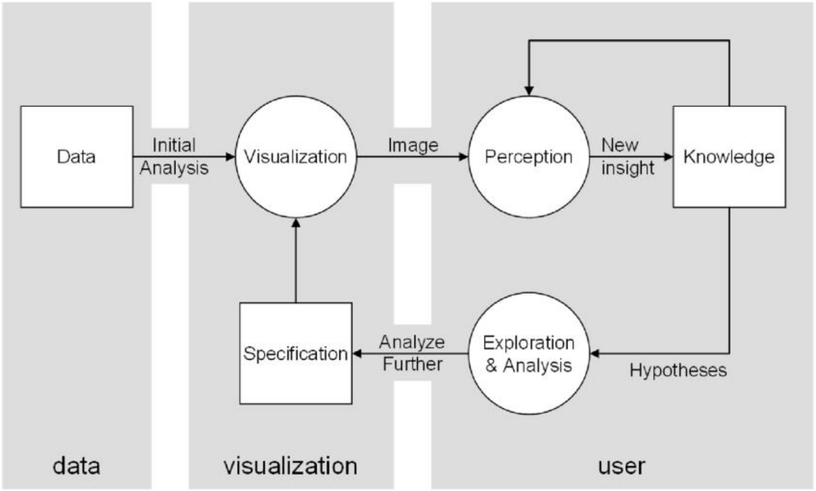
- In fact most approaches rely on Visual Analytics
- So they are presented a bit later in the presentation

Information Visualization ?

- “Visualization can be described as the mapping of data to visual form that supports human interaction in a workspace for visual sense making” [C*99]

- Use at best the Visual and cognitive capacities of the user

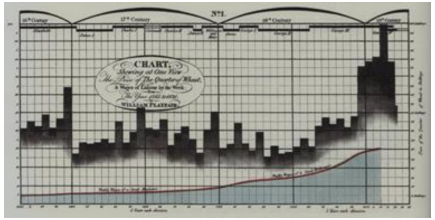


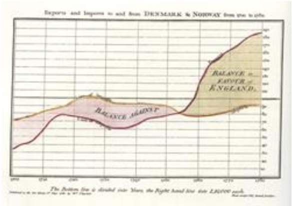


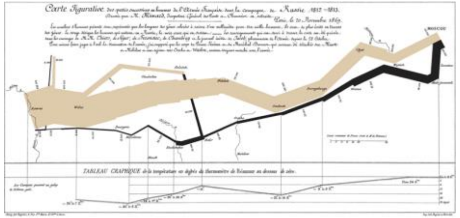
[C*99] S. Card *et al.*, Readings in information visualization: using vision to think, Morgan Kaufmann.

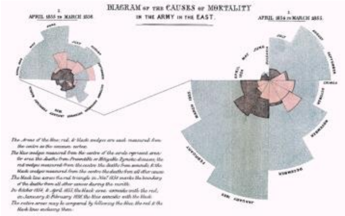
This is a short story in human history

- **First visual representations dates from the end of the 18th**

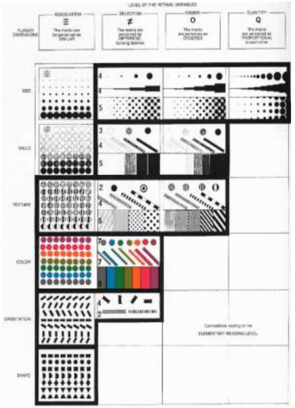


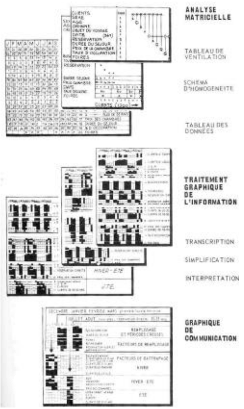




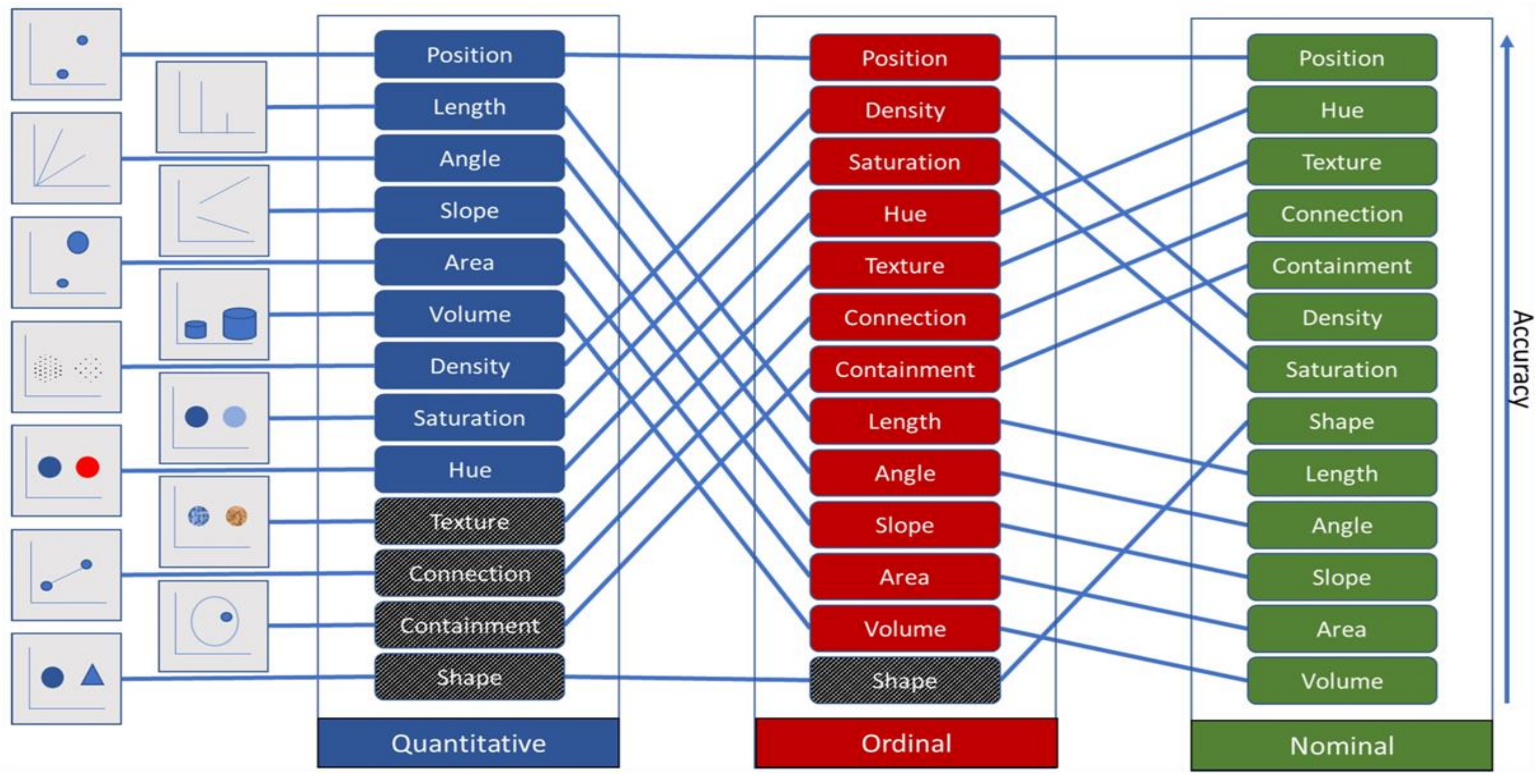


- **Updated with the book *Sémiologie Graphique* (1967) by J. Bertin**



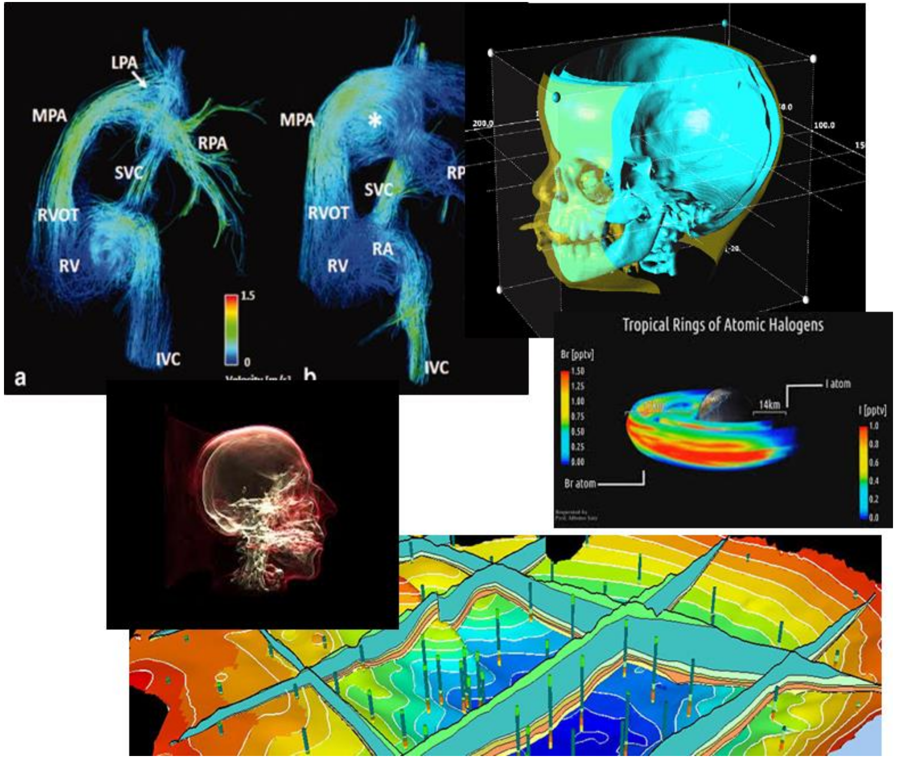


Computing ressources allow to design interactive and complex visualization. But some rules remains

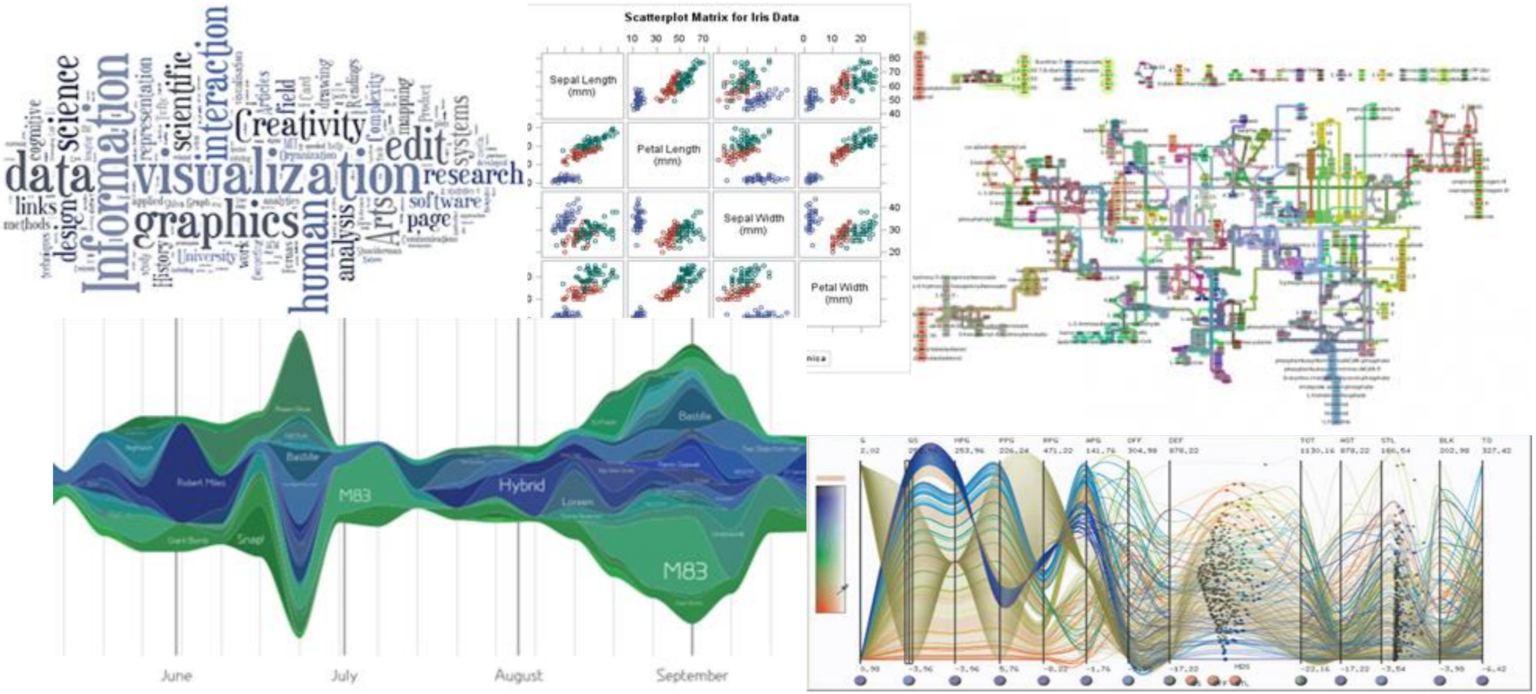


Mackinlay. Automating the design of graphical presentations of relational information. ACM Trans Graph, 5(2):110–141, April 1986. ISSN 0730-0301. doi: 10.1145/22949.22950.

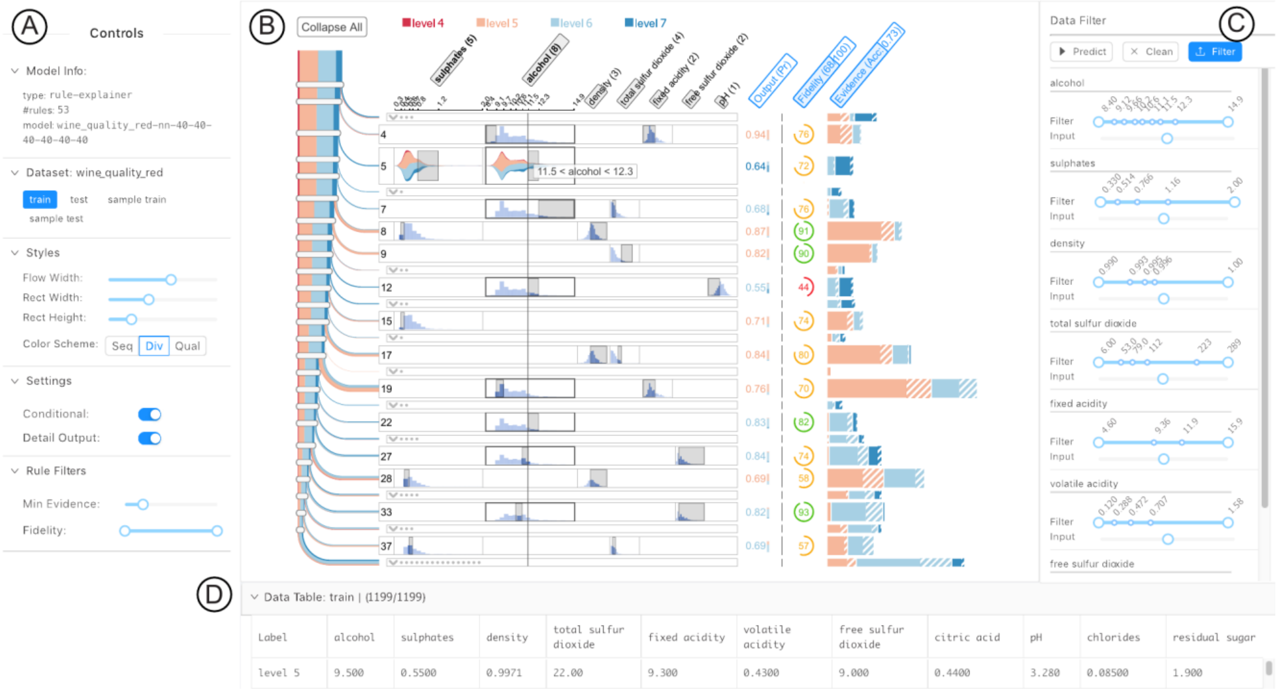
Scientific visualization concerns concrete data



Information visualization concerns abstract data

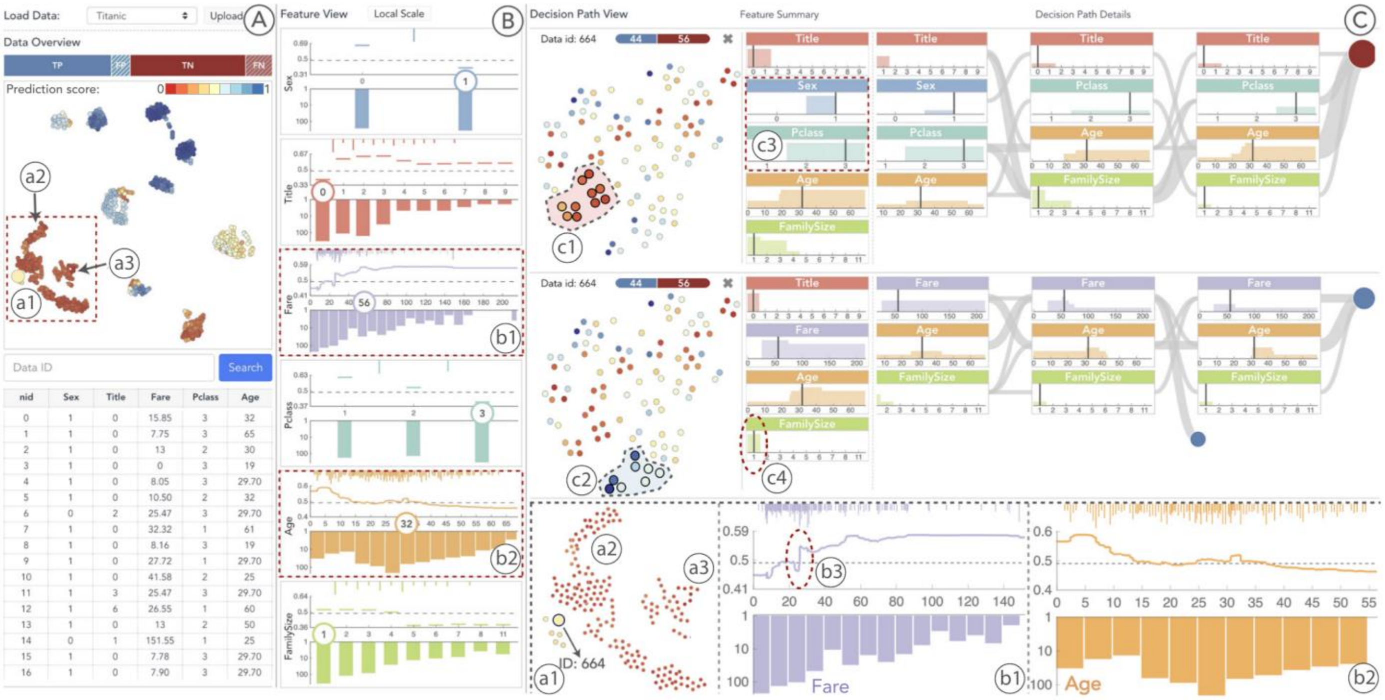


Rulematrix



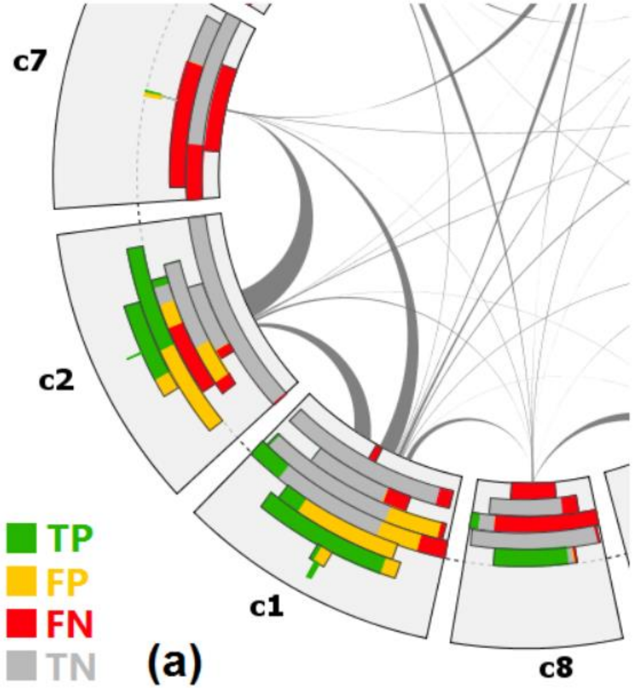
Ming, Y., Qu, H., & Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, *25*(1), 342-352.

iForest



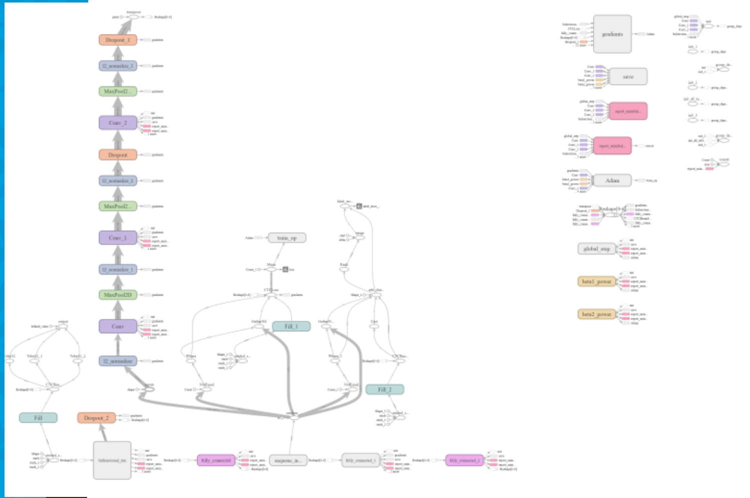
Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2018). iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics*, *25*(1), 407-416.

Confusion wheel



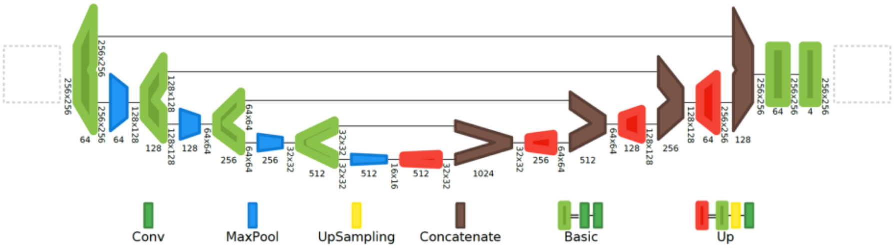
Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics*, *20*(12), 1703-1712.

To visualize model architecture is necessary but not sufficient



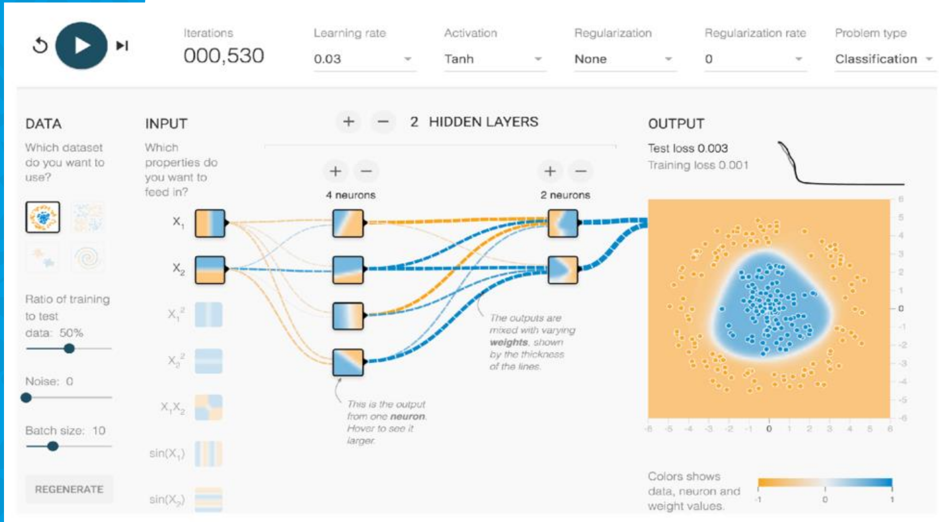
https://stackoverflow.com/questions/49812478/getting-input-names-from-frozen-tensorflow-estimator-graph

Wongsuphasawat, K.; Smilkov, D.; Wexler, J.; Wilson, J.; Mané, D.; Fritz, D.; Krishnan, D.; Viégas, F. B. & Wattenberg, M. Visualizing dataflow graphs of deep learning models in TensorFlow IEEE transactions on visualization and computer graphics, IEEE, 2018, 24, 1–12

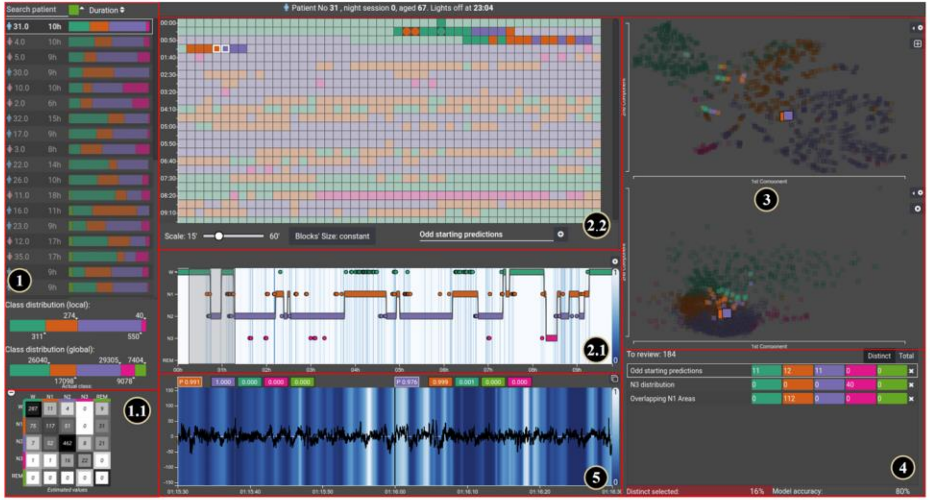


Bauerle, A., Van Onzenoodt, C., & Ropinski, T. (2021). Net2Vis-A Visual Grammar for Automatically Generating Publication-Ready CNN Architecture Visualizations. *IEEE Transactions on Visualization and Computer Graphics*.

Expectations of users diverge among applications

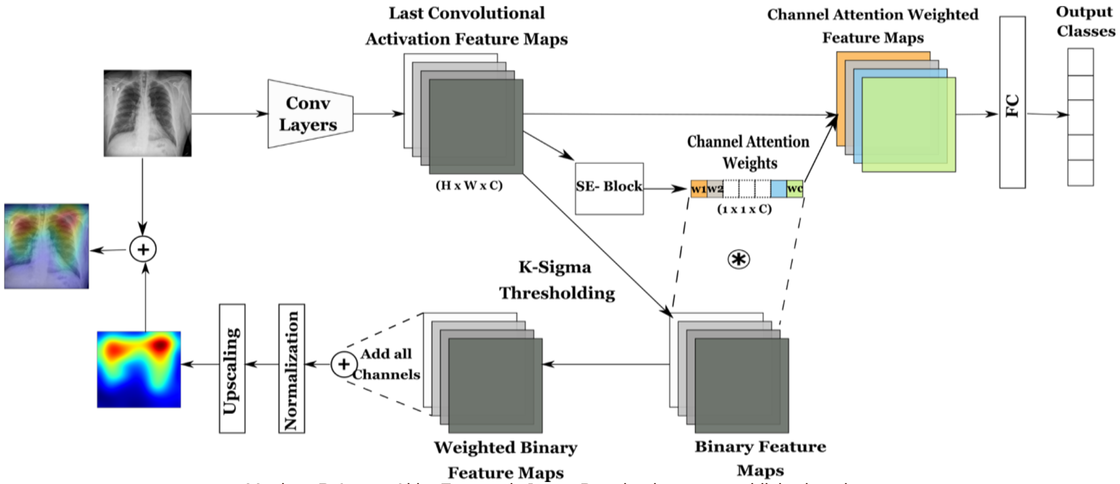


Smilkov, Daniel, et al. "Direct-manipulation visualization of deep networks." *arXiv preprint arXiv:1708.03788* (2017).

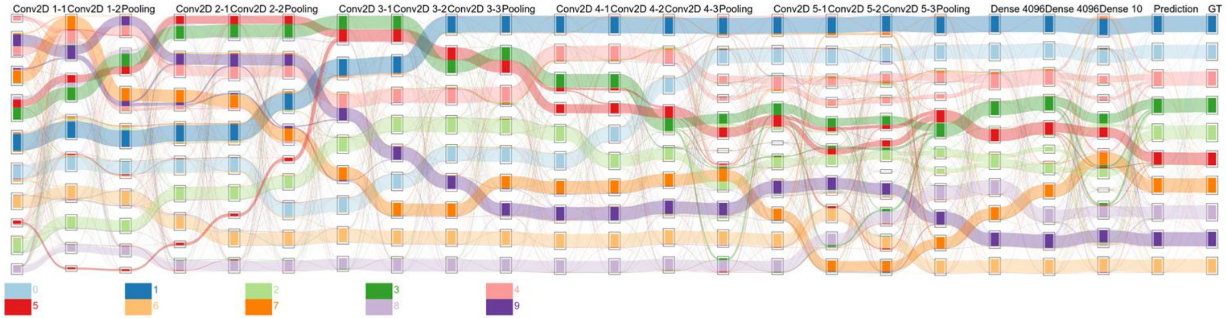


Garcia Caballero, H. S.; Westenberg, M. A.; Gebre, B. & van Wijk, J. J. V-Awake: A Visual Analytics Approach for Correcting Sleep Predictions from Deep Learning Models *Computer Graphics Forum,* 2019, *38*, 1-12

Local and global approaches are complimentary



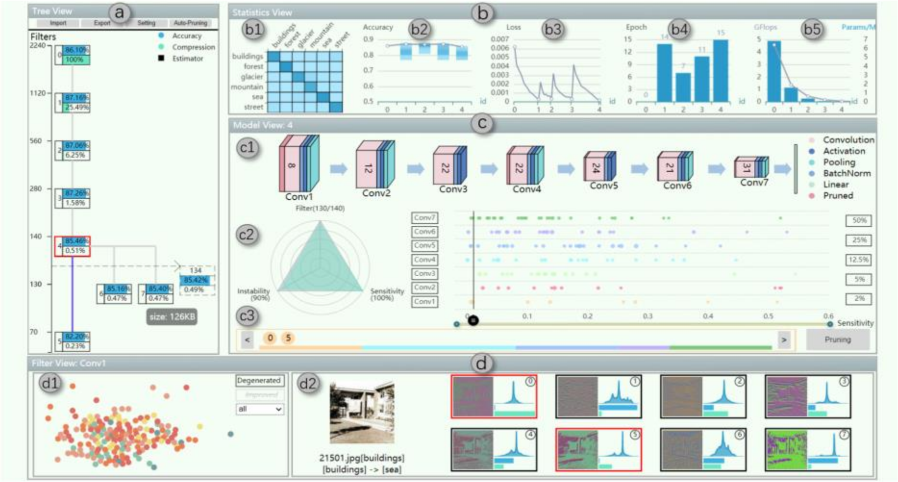
Meghna P Ayyar, Akka Zemmari, Jenny Benois-pineau, unpublished work



Halnaut, Adrien, et al. "Deep Dive into Deep Neural Networks with Flows." *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020): IVAPP*. Vol. 3. 2020.

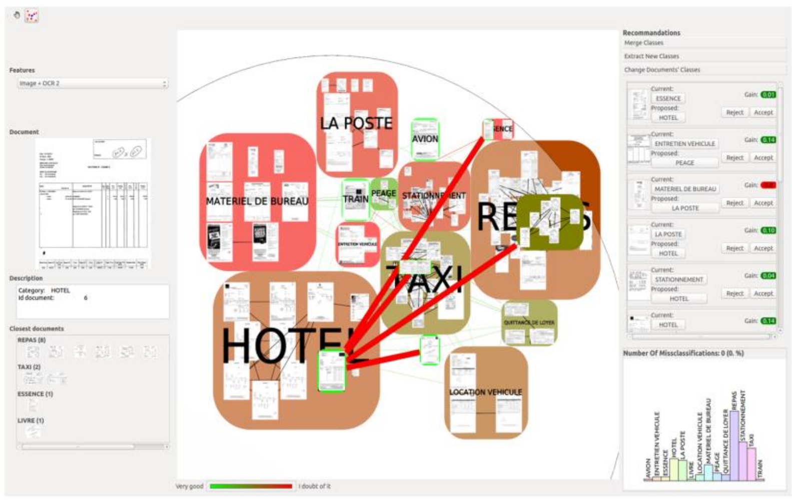
Visual analytics can help to build simpler models

Pezzotti, N.; Höllt, T.; Van Gemert, J.; Lelieveldt, B. P.; Eisemann, E. & Vilanova, A. DeepEyes: Progressive visual analytics for designing deep neural networks. *IEEE transactions on visualization and computer graphics, IEEE,* 2018, *24*, 98–108

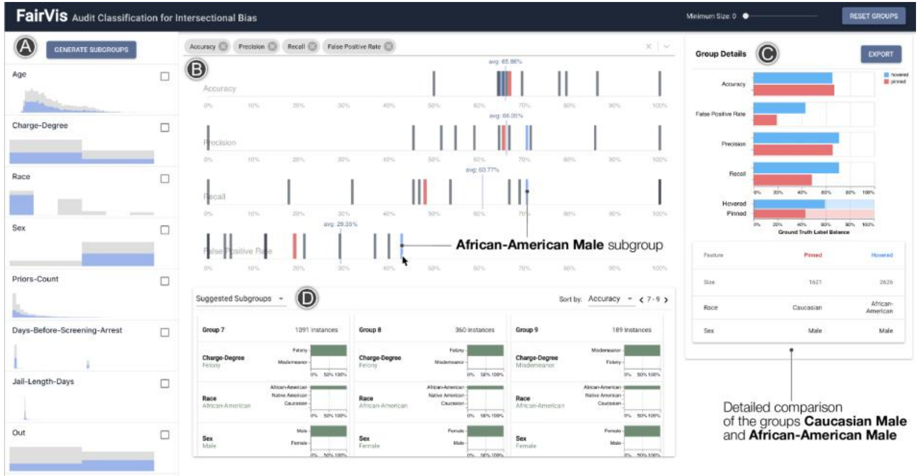


Li, Guan, et al. "CNNPruner: Pruning Convolutional Neural Networks with Visual Analytics." *IEEE Transactions on Visualization and Computer Graphics* (2020).

Performance understanding is strongly linked to dataset understanding

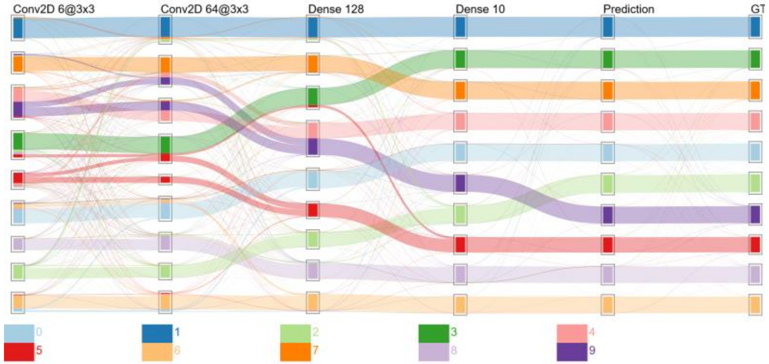


Romain Giot, Romain Bourqui, Nicholas Journet, Anne Vialard, "Visual Graph Analysis for Quality Assessment of Manually Labelled Documents Image Database", 13th International Conference on Document Analysis and Recognition (ICDAR 2015), pp 1136–1140, 2015.



Cabrera, Ángel Alexander, et al. "Fairvis: Visual analytics for discovering intersectional bias in machine learning." *2019 IEEE Conference on Visual Analytics Science and Technology (VAST).* IEEE, 2019.

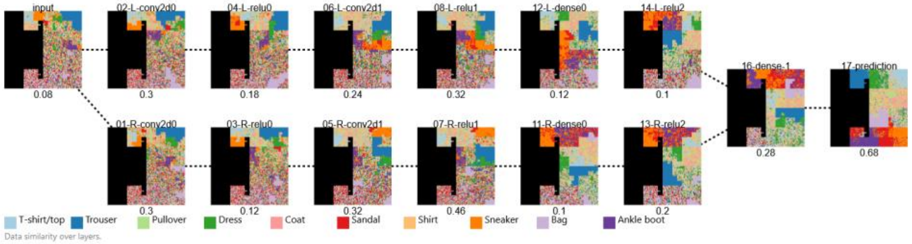
Some XAI works in my lab



Flow analysis of data over network

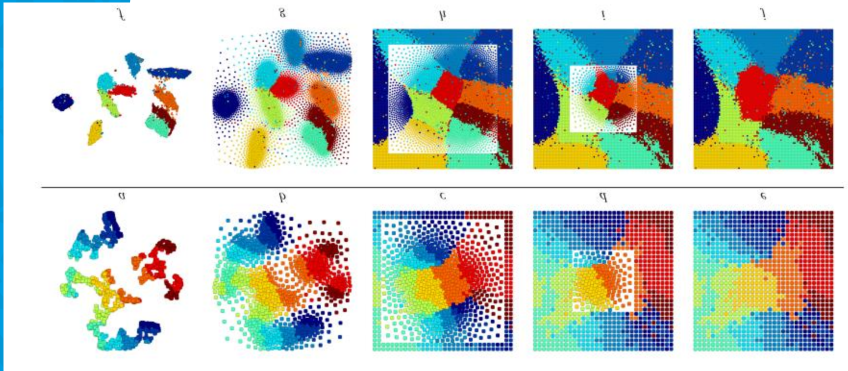
A. Halnaut, R. Giot, R. Bourqui, D. Auber. Deep Dive into Deep Neural Networks with Flows. *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020): IVAPP, Feb 2020, Valletta, Malta. pp.231-239.*

A. Halnaut, R. Giot, R. Bourqui, D. Auber. Samples Classification Analysis Across DNN Layers with Fractal Curves. *ICPR 2020's Workshop Explainable Deep Learning for AI, Jan 2021, Milano (virtual), Italy.*



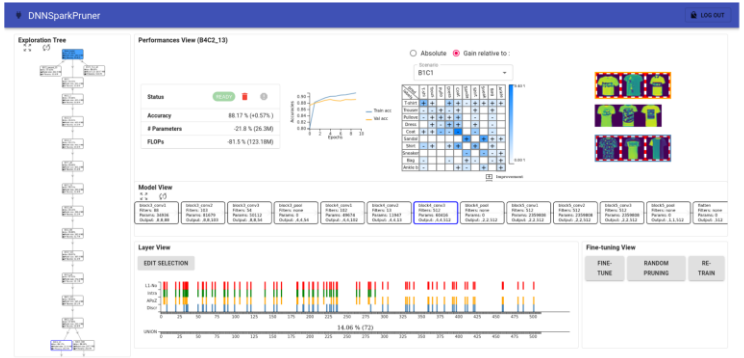
Network behavior over layers

Some XAI works in my lab



Pixel oriented data projection

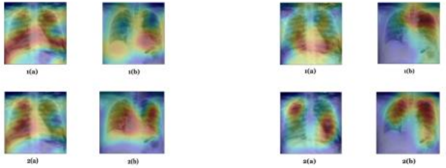
A. Halnaut, R. Giot, R. Bourqui, and D. Auber, "VRGrid: Efficient Transformation of 2D Data into Pixel Grid Layout," *Proceedings of the 26th International Conference Information Visualisation (IV2022)*, 2022.



Deep neural network simplification

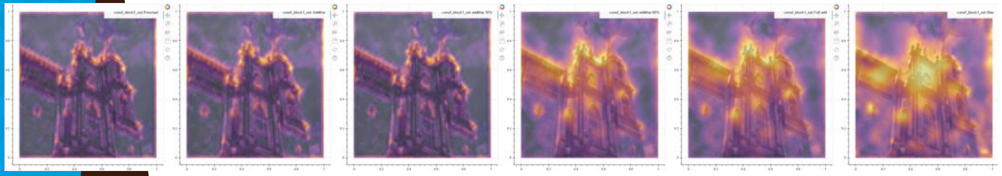
FEM

Modified FEM



ResNet50 + SE

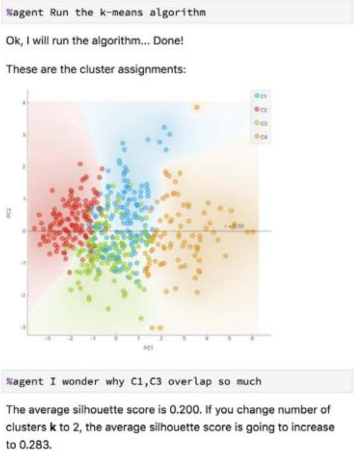
InceptionV3 +SE

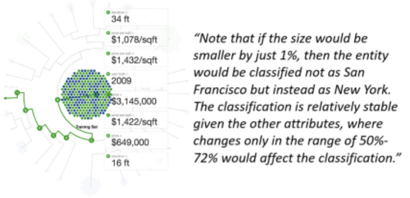


Improvment of feature attribution methods

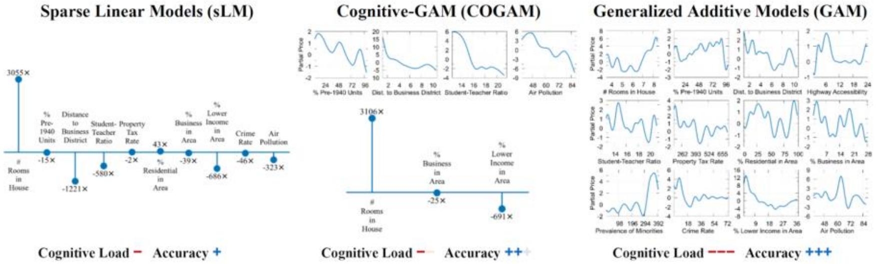
The future of interpretability

- Verbalization should help for understanding
  - Direct / no interpretation / no learning
- Self and posthoc explanability and should co-occur
- There are opportunities to create simpler explanations, easier to understanding even if less true





Sevastjanova, R.; Beck, F.; Ell, B.; Turkay, C.; Henkin, R.; Butt, M.; Keim, D. A. & El-Assady, M. Going beyond visualization: Verbalization as complementary medium to explain machine learning models
Workshop on Visualization for AI Explainability at IEEE VIS, 2018



Abdul, A.; Weth, C. V. D.; Kankanhalli, M. & Lim, B. Y. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations
Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020

The future of explainability

- Collaborations between machine learning & visualization communities need to be strengthened
  - data experts / model experts / data viz experts
  - Call for papers should be opened to both communities
- Semantic information has to be provided to interpretations
  - Feature importance do not bring enough information
  - Deciders needs to take decision on semantic
  - Interpretability has to be trustworthy

eXplainable Deep Learning

Romain Giot <romain.giot@u-bordeaux.fr>