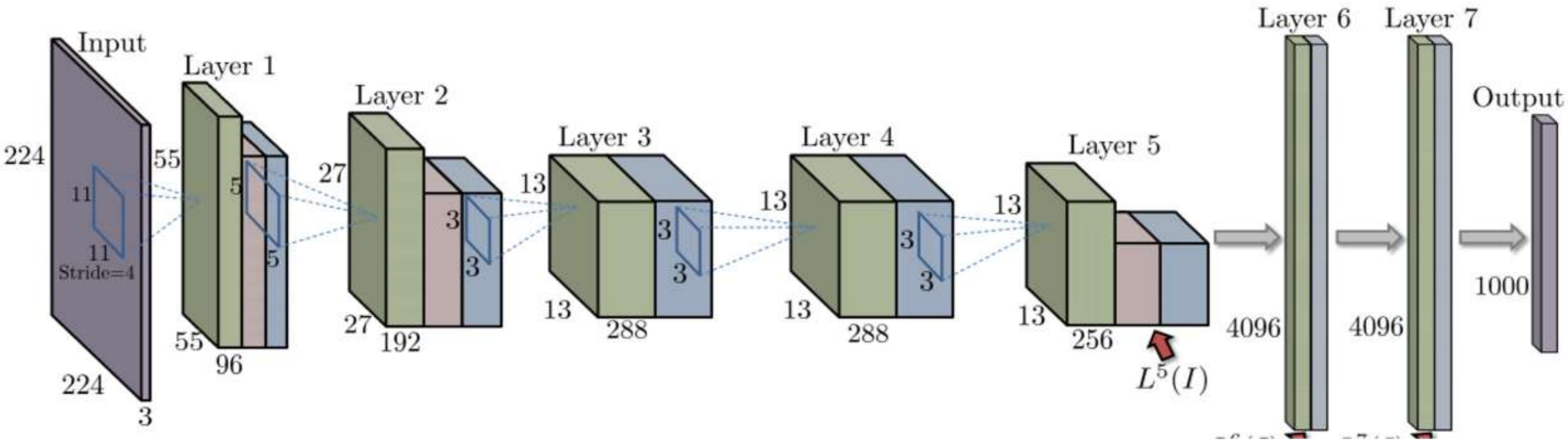
Class: Visualisation and Interpretation Methods for Image Classification: UMAP, Saliency Maps, Class Activation Maximization and Interpretability

Petia Radeva
Universitat de Barcelona

What accuracy did you get with your Neural Network?



Apple
Frog
Tiger
Lion
Heart
.
.
.

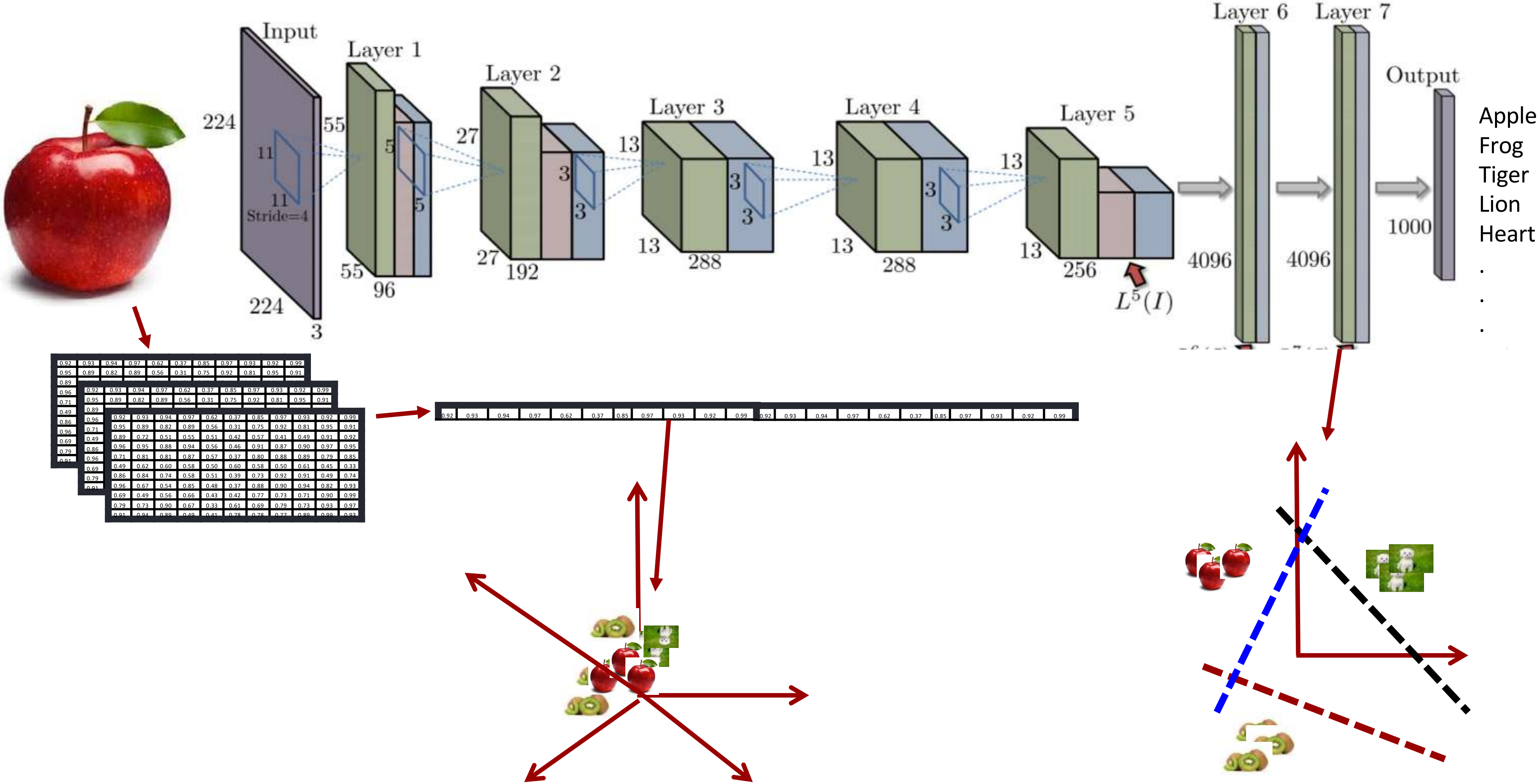
What to do when my NN is not learning?

Look at the model

Look at the data

01:51

What accuracy did you get with your Neural Network?



0.92	0.93	0.94	0.97	0.62	0.37	0.85	0.97	0.93	0.92	0.99
0.95	0.89	0.82	0.89	0.56	0.31	0.75	0.92	0.81	0.95	0.91
0.89	0.72	0.51	0.55	0.51	0.42	0.57	0.41	0.49	0.91	0.92
0.96	0.95	0.88	0.94	0.56	0.46	0.91	0.87	0.90	0.97	0.95
0.71	0.81	0.81	0.87	0.57	0.37	0.80	0.88	0.89	0.79	0.85
0.49	0.62	0.60	0.58	0.50	0.60	0.58	0.50	0.61	0.45	0.33
0.86	0.84	0.74	0.58	0.51	0.39	0.73	0.92	0.91	0.49	0.74
0.96	0.67	0.54	0.85	0.48	0.37	0.88	0.90	0.94	0.82	0.93
0.69	0.49	0.56	0.66	0.43	0.42	0.77	0.73	0.71	0.90	0.99
0.79	0.73	0.90	0.67	0.33	0.61	0.69	0.79	0.73	0.93	0.97
0.91	0.94	0.89	0.49	0.41	0.78	0.78	0.77	0.89	0.99	0.93

0.92	0.93	0.94	0.97	0.62	0.37	0.85	0.97	0.93	0.92	0.99	0.92	0.93	0.94	0.97	0.62	0.37	0.85	0.97	0.93	0.92	0.99

01:51

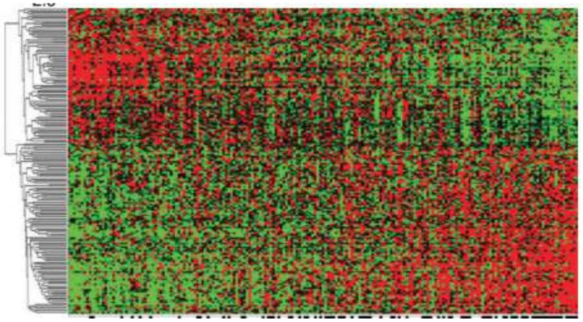
Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Problem:

How can humans visualize **high-dimensional** data?

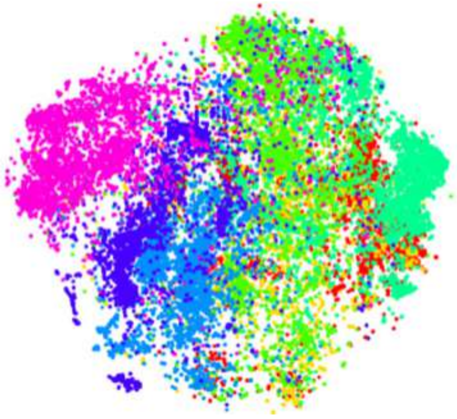
We can only see in 2D and 3D!



Visualization is important

- To see how mixed/separated are our data
- To see their relationships (if any)

- Check if any features are separating the data

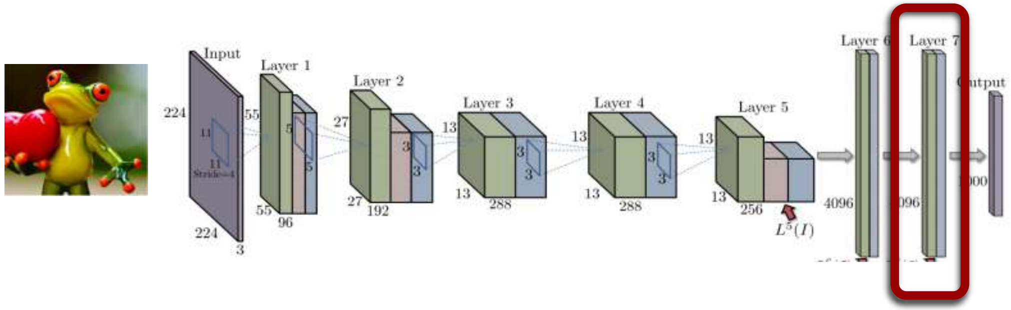


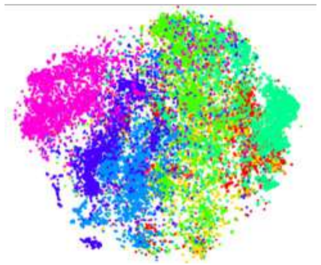
Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014, January). Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning (pp. 647-655). PMLR.

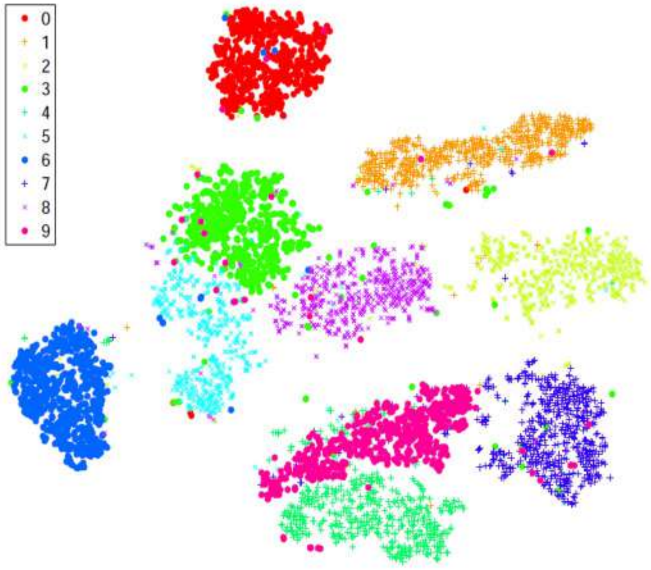
Dimensionality reduction

Original high-dimensional (high-d) space

Desired low-dimensional (low-d) space







(a) Visualization by t-SNE.

Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Principal Components Analysis

- Method to optimally summarise large multi-dimensional datasets

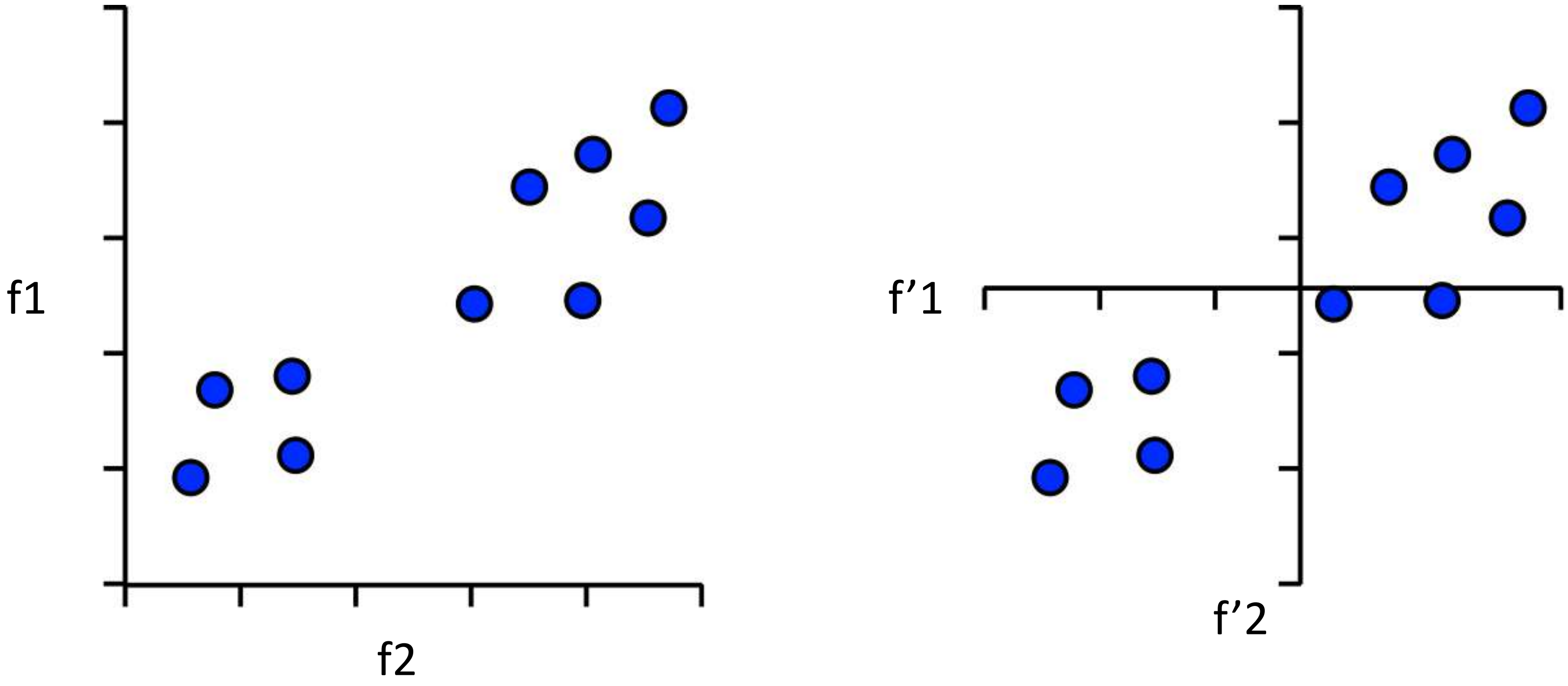
- **Goal**: find a smaller number of dimensions (ideally 2) which retain most of the useful information in the data

- Builds a recipe for converting large amounts of data into a single value, called a Principal Component (PC), eg:

$$PC = (f1*10)+(f2*3)+(f3*(-4))+(f4*(-20))\ldots$$

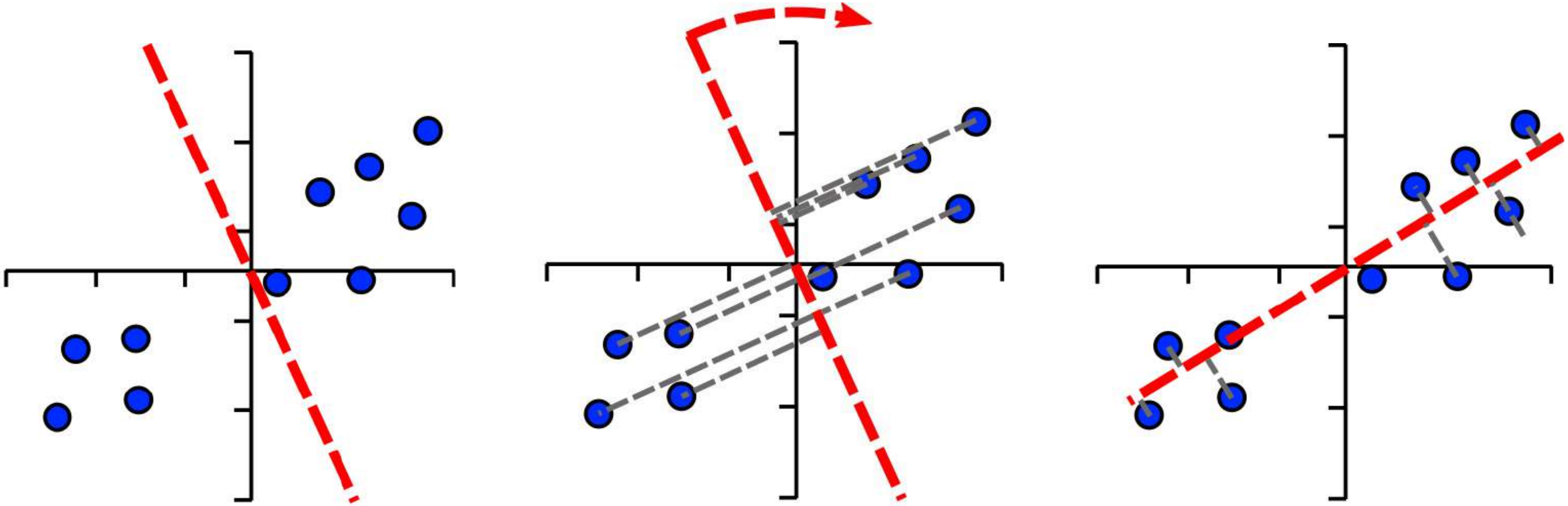
How does PCA work?

- Simple example using 2 features and 10 images



How does PCA work?

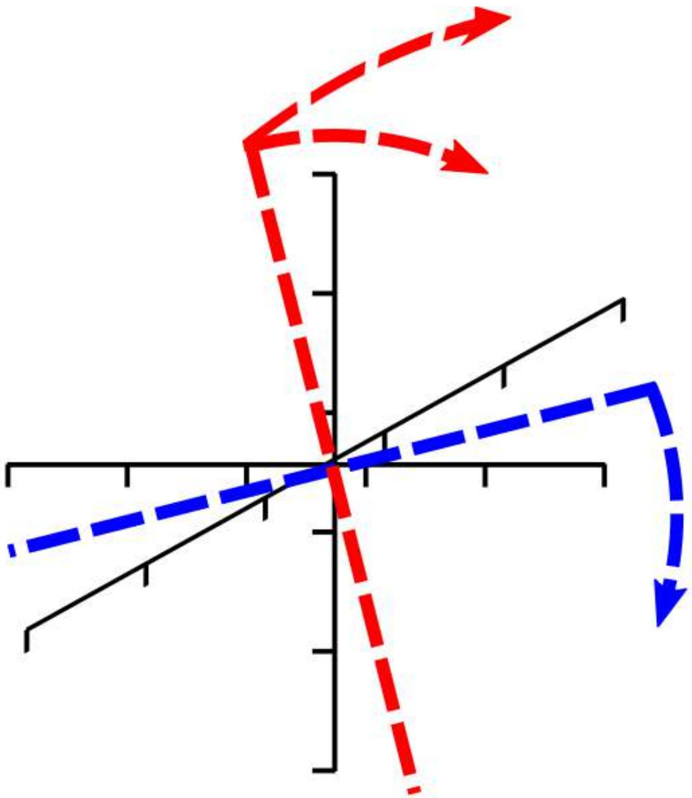
- Find line of best fit, passing through the origin



More dimensions

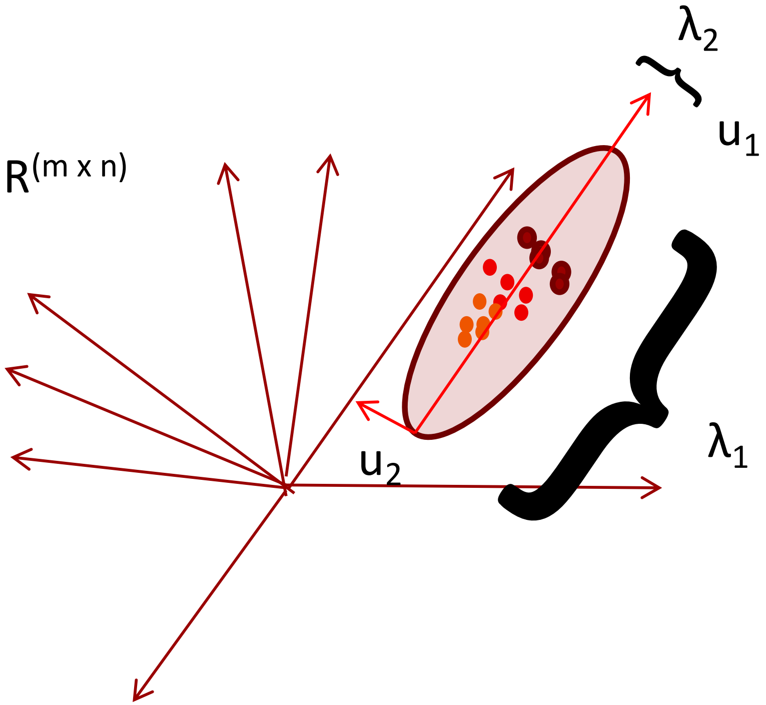
- The same idea extends to larger numbers of dimensions (n)

- Calculation of first PC rotates in (n-1) -dimensions
  - Next PC is perpendicular to PC2, but rotated similarly (n-2)
  - Last PC is remaining perpendicular (no choice)
  - Same number of PCs as features



Principal Component Analysis: the algorithm

According to the PCA, if we compute:

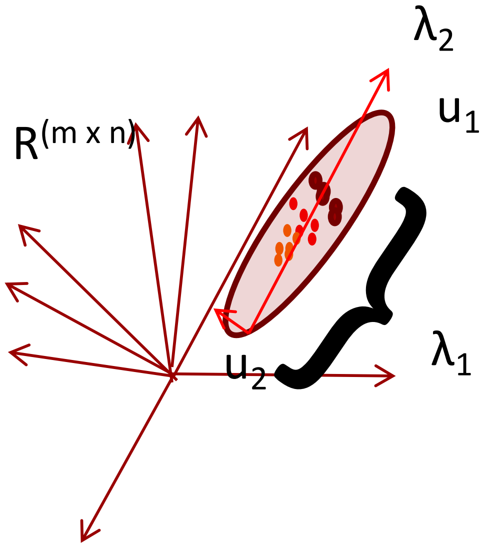


- the eigenvectors (e1, e2, ...) of the covariance matrix define the axis of maximum variance,
- and the eigenvalues give a measure of the variance of the data.

$$\sum = \tfrac{1}{M} \sum_{i=1}^{M} (X_i - \overline{X})(X_i - \overline{X})^T = AA^T;$$

Algorithm of PCA

•Eigenvectors (u_1, u_2, ...) of the covariance matrix Σ define the subspace that represents the data distribution.



•Remember: given any matrix B: a vector u is called eigenvector, iff:

$$Bu = \lambda u$$

•λ is the eigenvalue of the matrix.

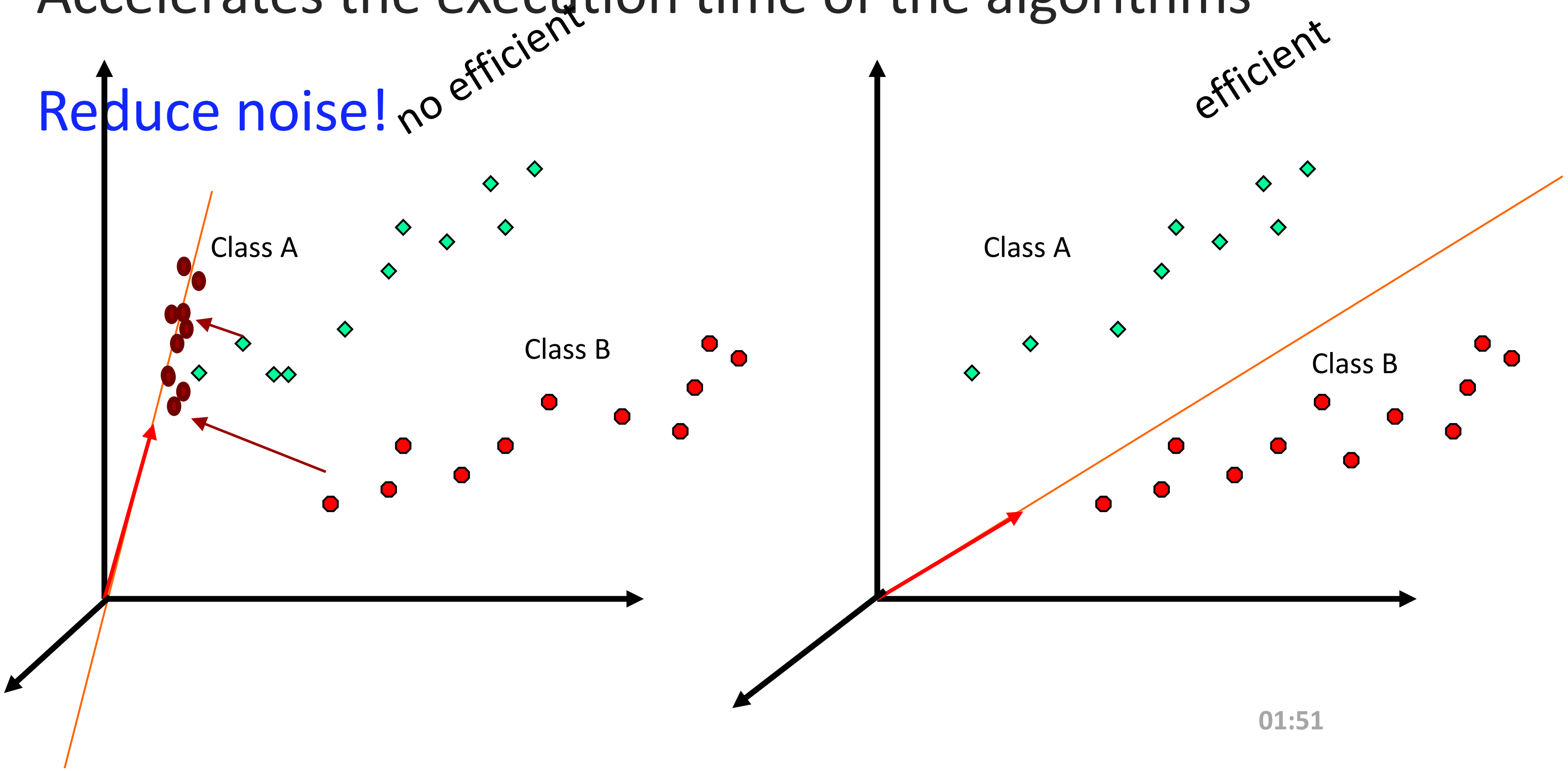
•Eigenvectors (u_1, u_2, ...) are orthogonal and form a basis.

•The original data once centered is projected into the eigenspace: X-> Y:

$$Y_i = (X_i - \overline{X})^T * (u_1, u_2, \ldots u_M)$$

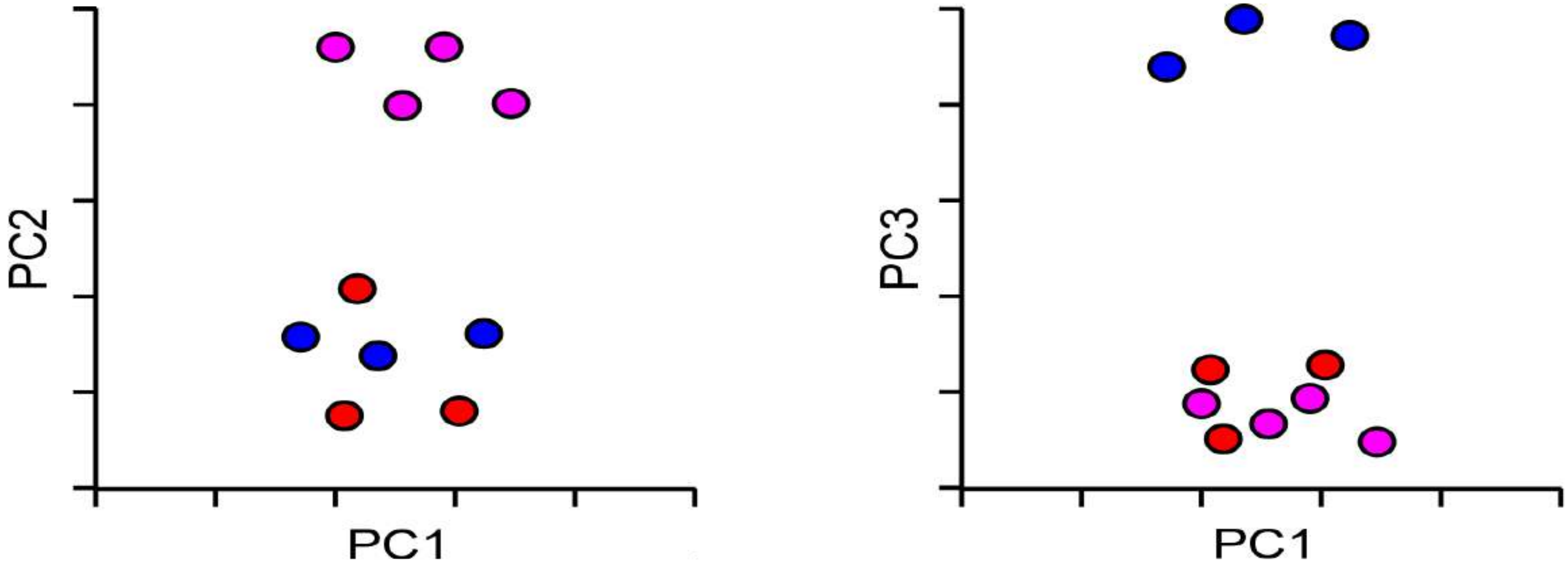
Properties of Principal Component Analysis

- Seeks directions efficient for representing data in all its variance
- Reduces the dimensions of data
- Accelerates the execution time of the algorithms
- Reduce noise!



01:51

So is PCA great always?



Not optimised for 2-dimensions

Disadvantages of PCA:

a) Linear decomposition
b) Orthogonal axes
c) Represents a transformation consisting of a rotation, affine rescaling and translation
d) Is not discriminative.

Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
  - Feature Visualization
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

tSNE to the rescue…

- T-Distributed Stochastic Neighbour Embedding
  - L.v.d. Maaten, G. Hinton. Visualizing data using t-SNE [PDF]
    Journal of Machine Learning Research, Vol 9(Nov), pp. 2579—2605. 2008.

- Aims to solve the problems of PCA
  - **Non-linear scaling** to represent changes at different levels
  - Optimal separation in 2-dimensions

    https://distill.pub/2016/misread-tsne/

What is t-SNE?

t-Distributed Stochastic Neighbor Embedding (t-SNE)

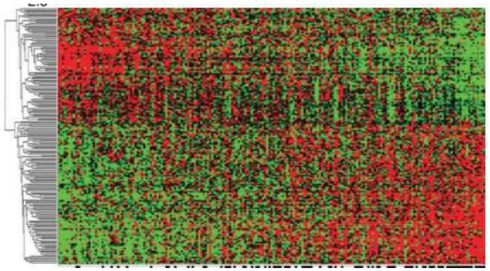
Unsupervised (does not use class labels)

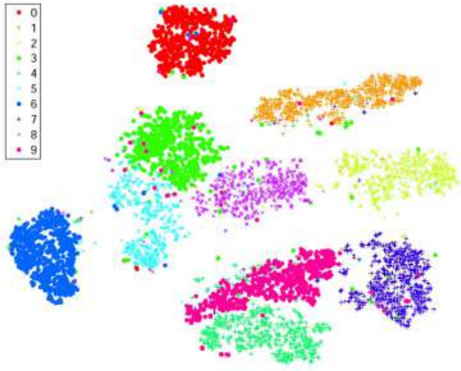
Models the probabilities that **high-d** points x_i, x_j are neighbors

Models the probabilities that the corresponding **low-d** points y_i, y_j are neighbors

If $(x_i,x_j) < (x_i,x_k)$, *for all k* => $(y_i,y_j) < (y_i,y_k)$, *for all k*

Goal: Finds a **low-d** representation that is faithful to the **high-d** model.



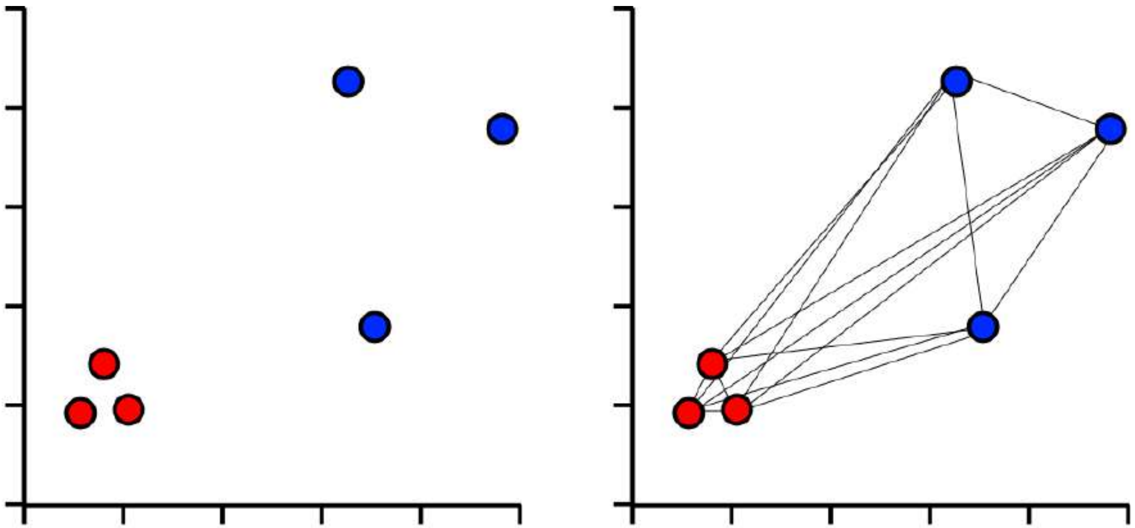


(a) Visualization by t-SNE.

How does tSNE work?

- Based around all-vs-all table of pairwise image to image distances

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2 / 2\sigma^2\right)}$$

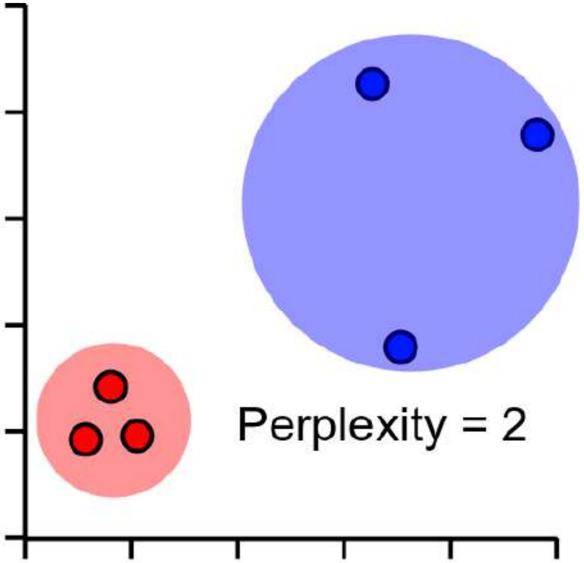


	🟥	🟥	🟥	🟦	🟦	🟦
🟥	0	10	10	295	158	153
🟥	9	0	1	217	227	213
🟥	1	8	0	154	225	238
🟦	205	189	260	0	23	45
🟦	248	227	246	44	0	54
🟦	233	176	184	41	36	0

Distance scaling and perplexity

- **Perplexity** = expected number of neighbours within a cluster
  - Prefixed parameter, used to compute the variance of data

- Distances scaled relative to perplexity neighbours



	🟥	🟥	🟥	🟦	🟦	🟦
🟥	0	4	6	586	657	836
🟥	4	0	4	815	527	776
🟥	9	3	0	752	656	732
🟦	31	28	29	0	4	7
🟦	31	24	25	4	0	7
🟦	40	37	32	8	8	0

- Compute the distances in the low-dimensional space

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

How t-SNE works

The t-SNE algorithm finds the similarity measure between pairs of instances in higher and lower dimensional space.

It tries to optimize two similarity measures in three steps:

1. t-SNE models a point being selected as a neighbour of another point in both higher and lower dimensions.
    1. It starts by calculating a pairwise similarity between all data points in the high-dimensional space using a Gaussian kernel.
    2. The points that are far apart have a lower probability of being picked than the points that are close together.

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2 / 2\sigma^2\right)}$$

How t-SNE works

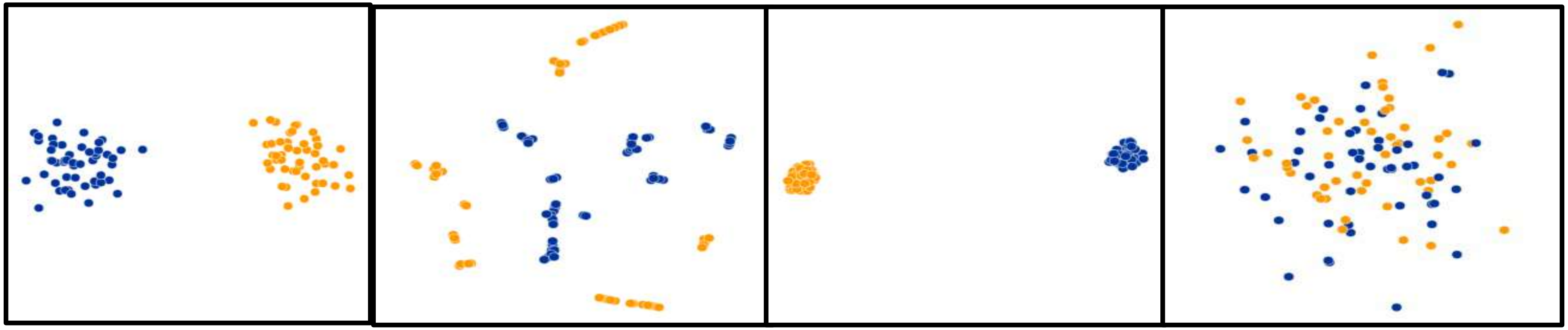
- Then, the algorithm tries to map higher dimensional data points onto lower dimensional space while preserving the pairwise similarities.
  1. It is achieved by minimizing the divergence between the probability distribution of the original high-dimensional and lower-dimensional.
  2. The algorithm uses gradient descent to minimize the divergence.
  3. The lower-dimensional embedding is optimized to a stable state.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

- The optimization process allows the creation of clusters and sub-clusters of similar data points in the lower-dimensional space
  - visualized to understand the structure and relationship in the higher-dimensional data.

Perplexity Settings Matter

https://distill.pub/2016/misread-tsne/



Original

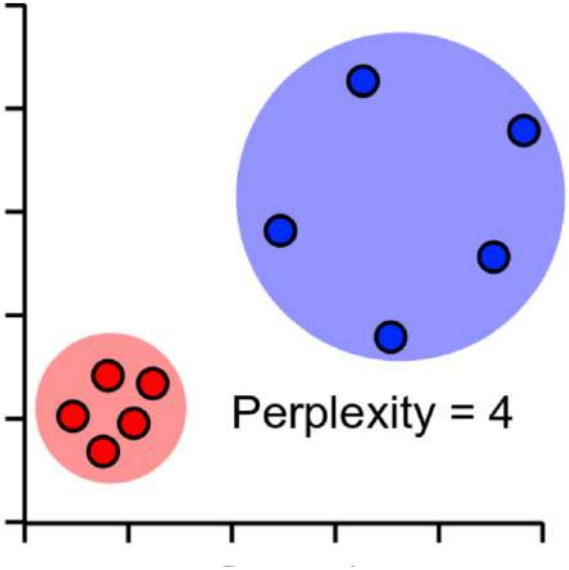
Perplexity = 2

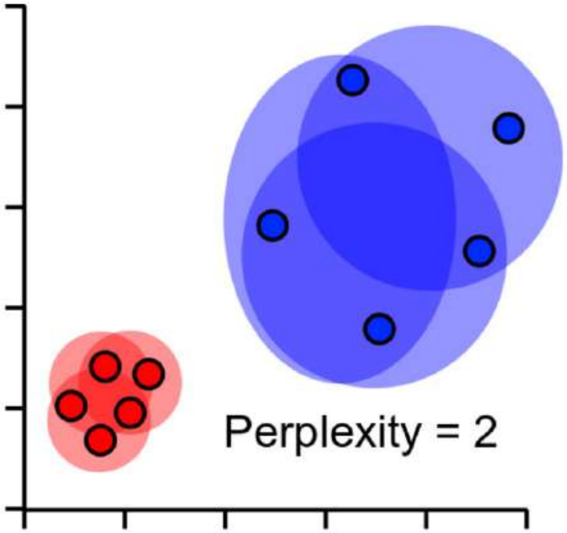
Perplexity = 30

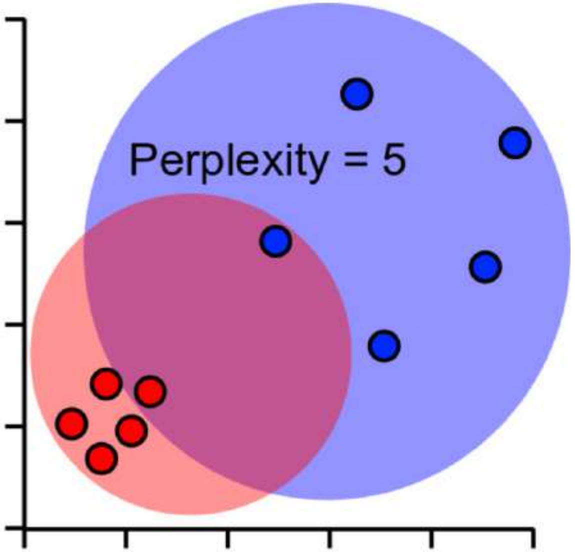
Perplexity = 100

- X and Y don’t mean anything (unlike PCA)
- Distance doesn’t mean anything (unlike PCA)
- Close proximity is highly informative
- Distant proximity isn’t very interesting
- Can’t rationalise distances, or add in more data

Perplexity robustness







Still not perfect

- PCA
  - Requires more than 2 dimensions
  - Expects linear relationships

- tSNE
  - Can’t cope with noisy data
  - Loses the ability to cluster

Answer: Combine the two methods, get the best of both worlds

- PCA
  - Good at extracting signal from noisy data
  - Extracts informative dimensions

- tSNE
  - Can reduce to 2D well
  - Can cope with non-linear scaling

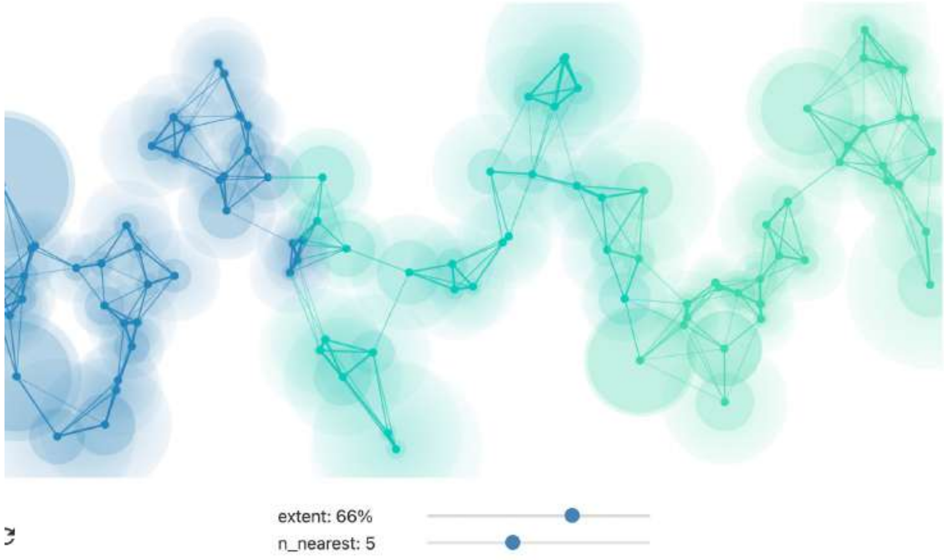
Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
  - Feature Visualization
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

How UMAP Works?

- In order to construct the initial high-dimensional graph, UMAP builds something called a "fuzzy simplicial complex".
  - This is really just a representation of a weighted graph, with edge weights representing the likelihood that two points are connected.

- To determine connectedness, UMAP extends a radius outwards from each point, connecting points when those radii overlap.
  - Choosing this radius is critical - too small a choice will lead to small, isolated clusters, while too large a choice will connect everything together.
  - UMAP overcomes this challenge by choosing a radius locally, based on the distance to each point's *n*th nearest neighbour.



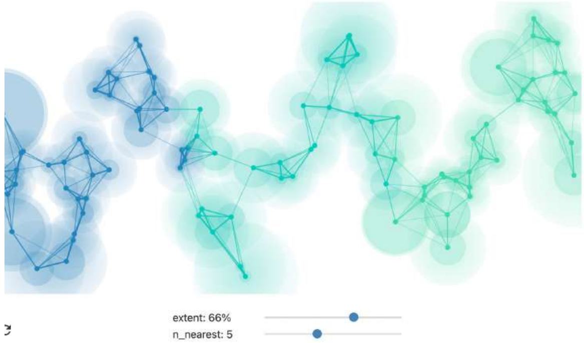
How UMAP Works?

- UMAP then makes the graph "fuzzy" by decreasing the likelihood of connection as the radius grows.

- By stipulating that each point must be connected to at least its closest neighbour, UMAP ensures that local structure is preserved in balance with global structure.

- Once the high-dimensional graph is constructed, UMAP optimizes the layout of a low-dimensional analogue to be as similar as possible essentially the same as in t-SNE.

- The key to effectively using UMAP lies in understanding the construction of the initial, high-dimensional graph.
  - Though the ideas behind the process are very intuitive, the algorithm relies on some advanced mathematics to give strong theoretical guarantees about how well this graph actually represents the data.



UMAP differences

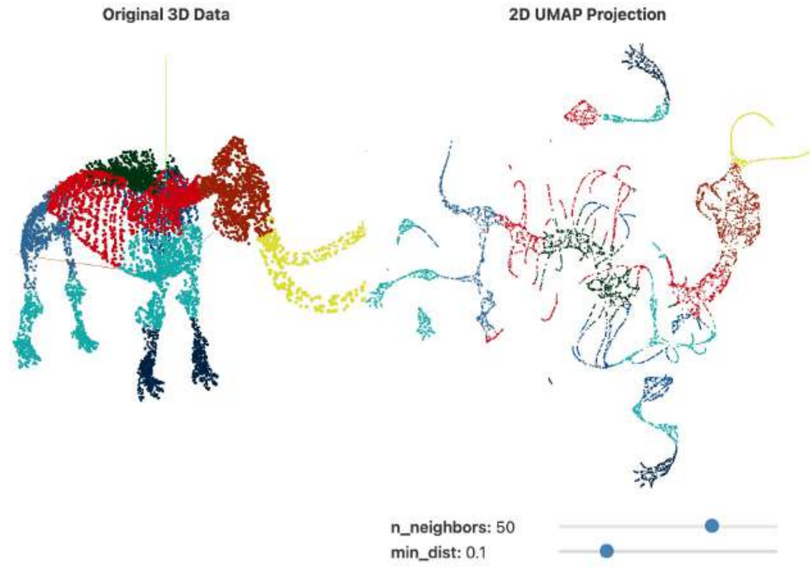
- Instead of the single perplexity value in tSNE, UMAP defines

- Nearest neighbours: the number of expected nearest neighbours – basically the same concept as perplexity

- Nearest neighbours will affect the influence given to global vs local information.

- **Minimum distance**: how tightly UMAP packs points which are close together

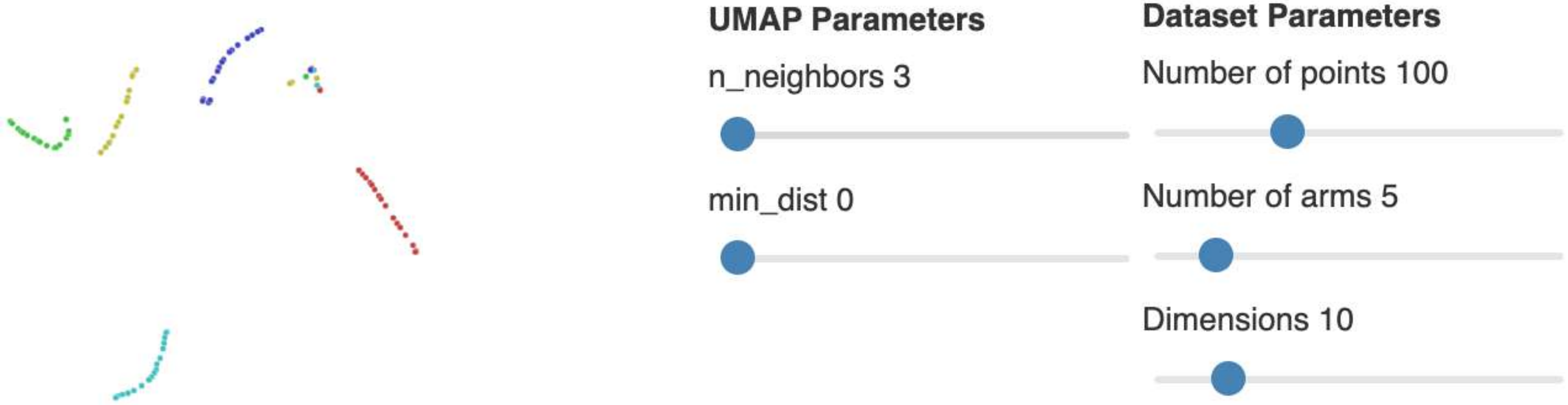
- Min_dist will affect how compactly packed the local parts of the plot are.
- As the min_dist parameter increases, UMAP t points, leading to decreased clustering of the structure.



n_neighbors: 50
min_dist: 0.1

UMAP to the rescue!

- UMAP is a replacement for tSNE to fulfil the same role
- Conceptually very similar to tSNE, but with a couple of relevant (and somewhat technical) changes.
- Based on Topological Data Analysis



UMAP Parameters

n_neighbors 3

min_dist 0

Dataset Parameters

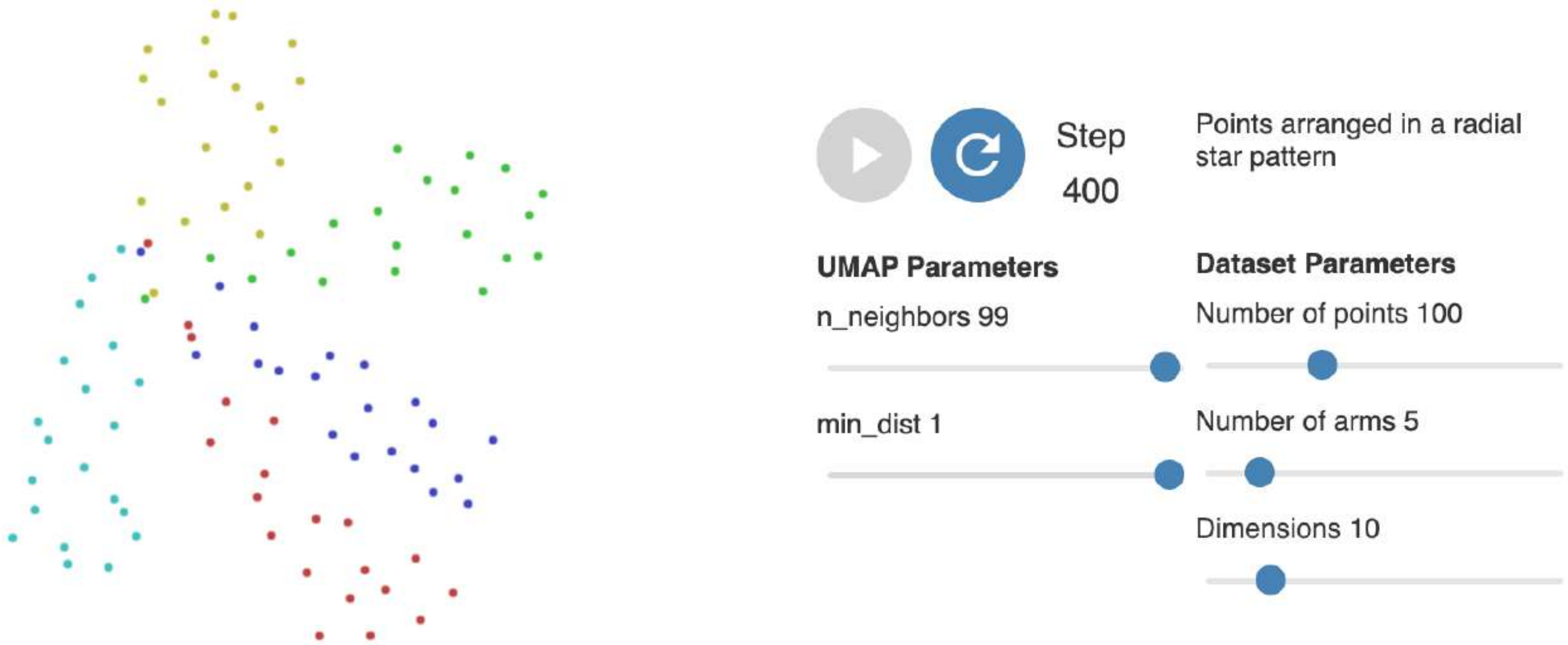
Number of points 100

Number of arms 5

Dimensions 10

https://pair-code.github.io/understanding-umap/

UMAP to the rescue!



Step 400

Points arranged in a radial star pattern

UMAP Parameters

n_neighbors 99

min_dist 1

Dataset Parameters

Number of points 100

Number of arms 5

Dimensions 10

Practical outcome is:

- UMAP is quite a bit quicker than tSNE
- UMAP can preserve more global structure than tSNE
- UMAP can run on raw data without PCA preprocessing
- UMAP can allow new data to be added to an existing projection

https://pair-code.github.io/understanding-umap/

How to (mis)read UMAP

- 1. Hyperparameters really matter
- 2. Cluster sizes in a UMAP plot mean nothing
- 3. Distances between clusters might not mean anything
- 4. Random noise doesn’t always look random.
- 5. You may need more than one plot (UMAP is stochastic)

UMAP vs tSNE Conclusions

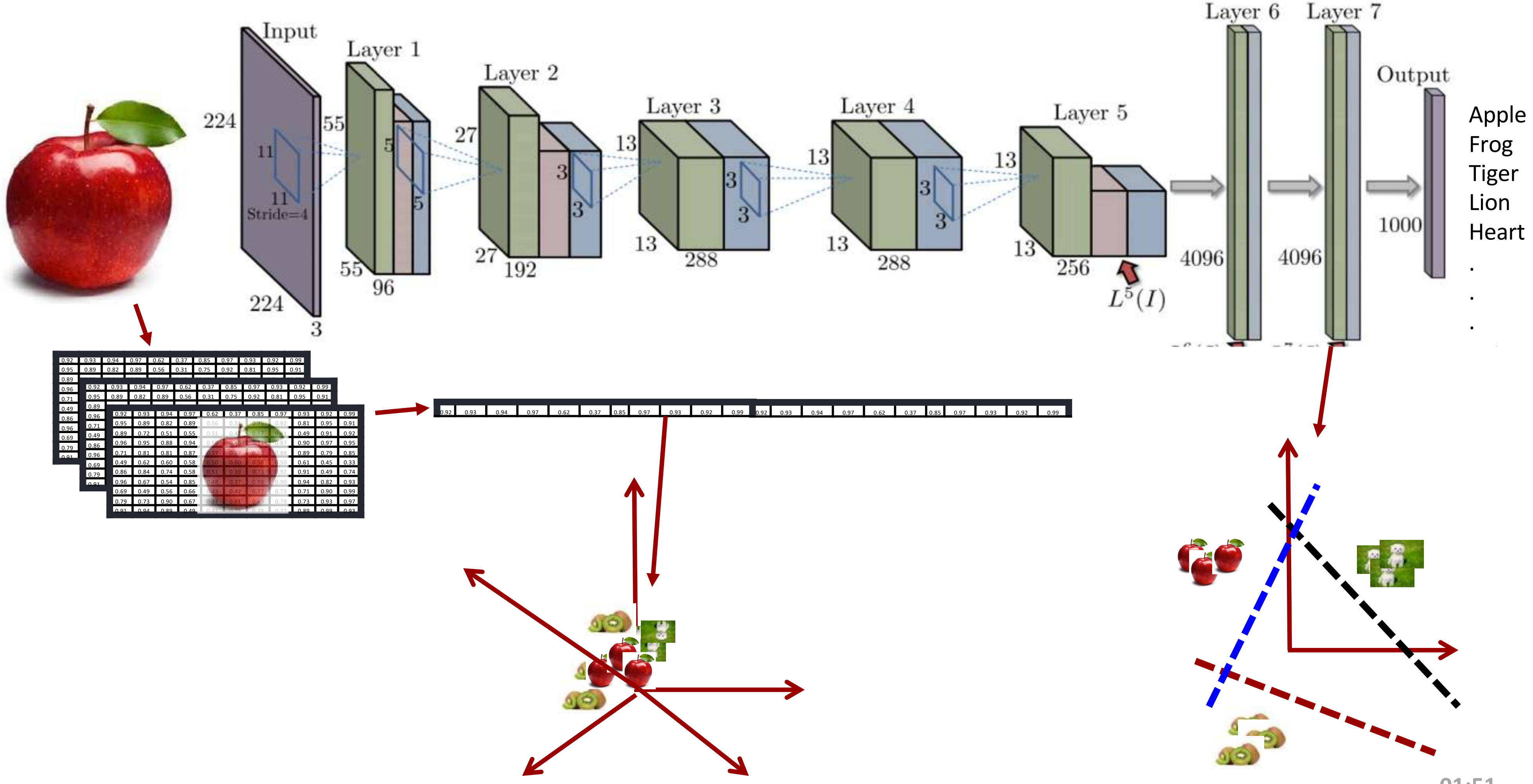
- UMAP is an incredibly **powerful tool** in the data scientist's arsenal, and offers a number of **advantages over t-SNE**.

- While both UMAP and t-SNE produce somewhat similar output, the **increased speed, better preservation of global structure**, and more **understandable parameters** make UMAP a more effective tool for visualizing high dimensional data.

- **No dimensionality reduction technique is perfect** - by necessity, we're distorting the data to fit it into lower dimensions - and UMAP is no exception.

- By building up an intuitive understanding of how the algorithm works and understanding how to tune its parameters, we can more effectively use this powerful **tool to visualize and understand large, high-dimensional datasets**.

Can we get insight about the Neural Network?



01:51

Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Saliency Maps with DeconvNet

Zeiler and Fergus proposed a method able to recognize what features in the input image an intermediate layer of the network is looking for.

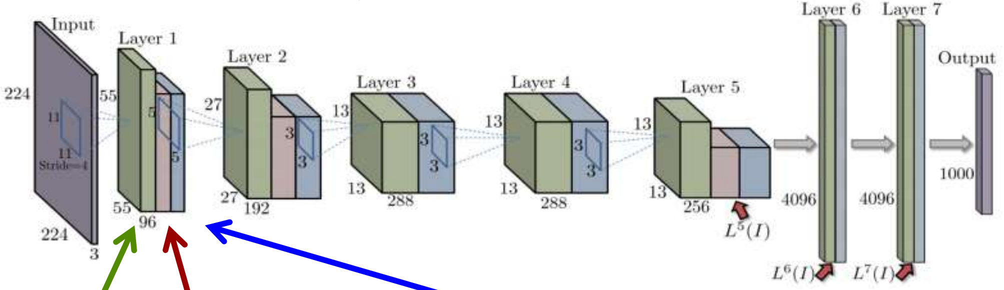
The method is known as **Deconvolutional Network (DeconvNet),** because it inverts the operations from a particular layer to the input using the activation of that specific layer.

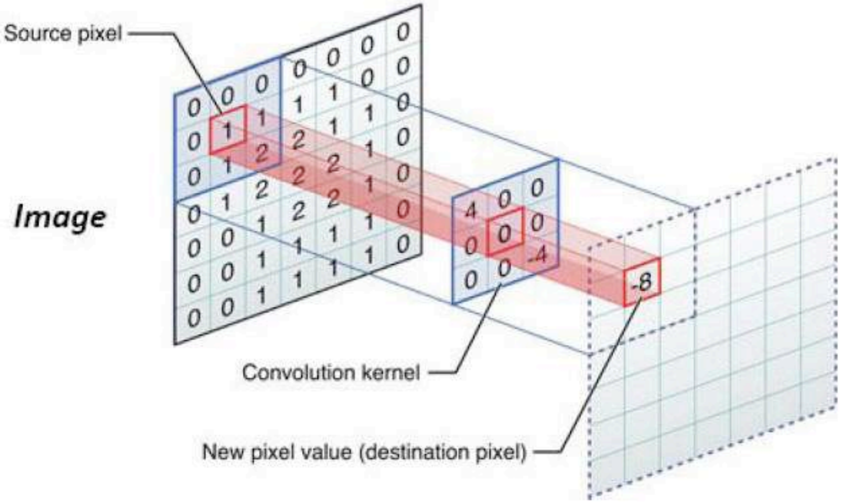
Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, 2013, arXiv:1311.2901

Layers of a CNN

Spatial info

No spatial info

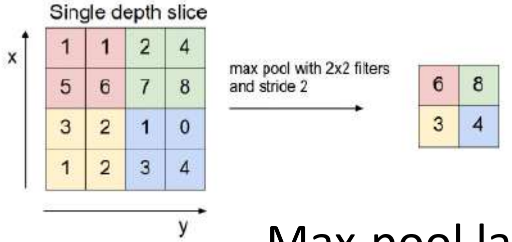




Filter Response

ReLU max(0,x)

Convolutional layer



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Max-pool layer

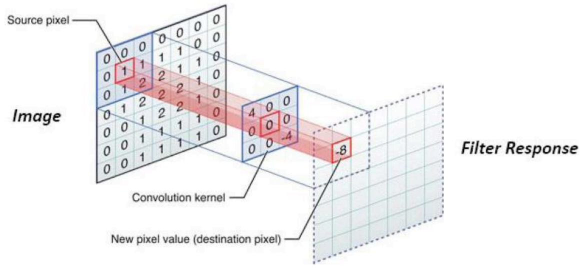
Saliency Maps with DeconvNet

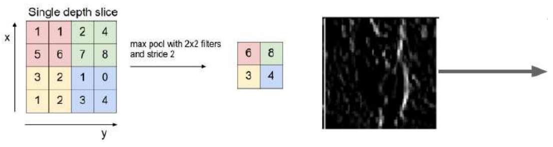
To invert the process the authors used:

- **Deconvolution** is used as the inverse of convolution (with the transposed version of the same filters),

- **Unpooling** as the inverse of pooling,

- **Inverse ReLU** removing the negative values as the inverse of itself.





1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

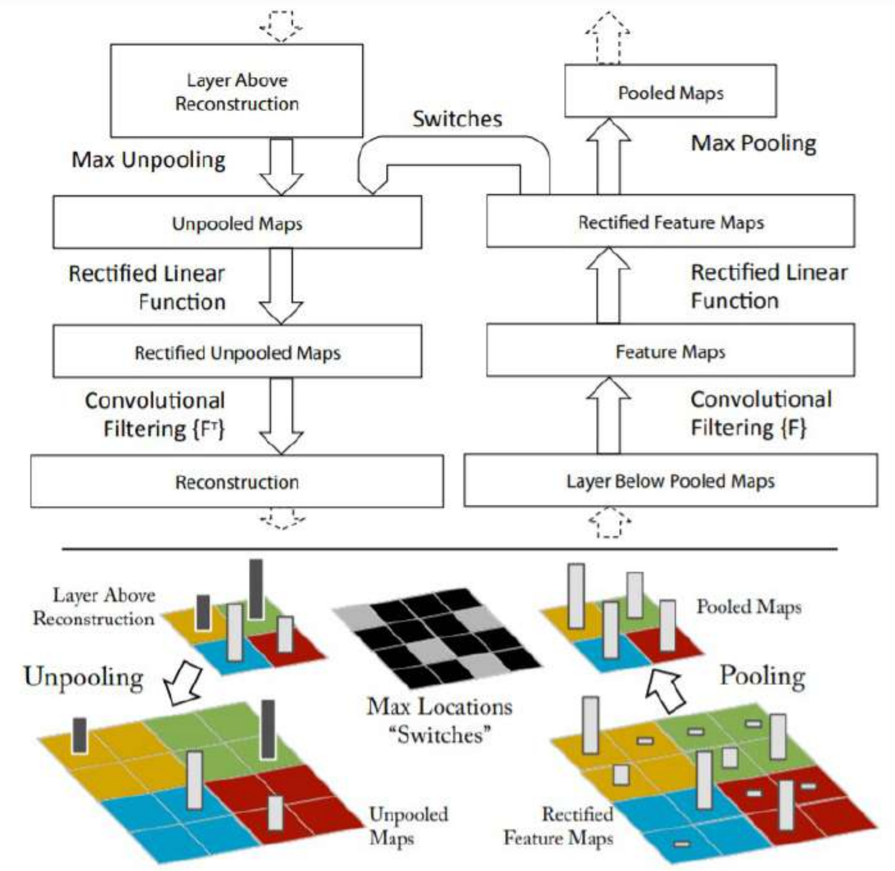
ReLU max(0,x)

Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, 2013, arXiv:1311.2901

Saliency Maps with DeconvNet <cont>

The whole process is illustrated in the figure:

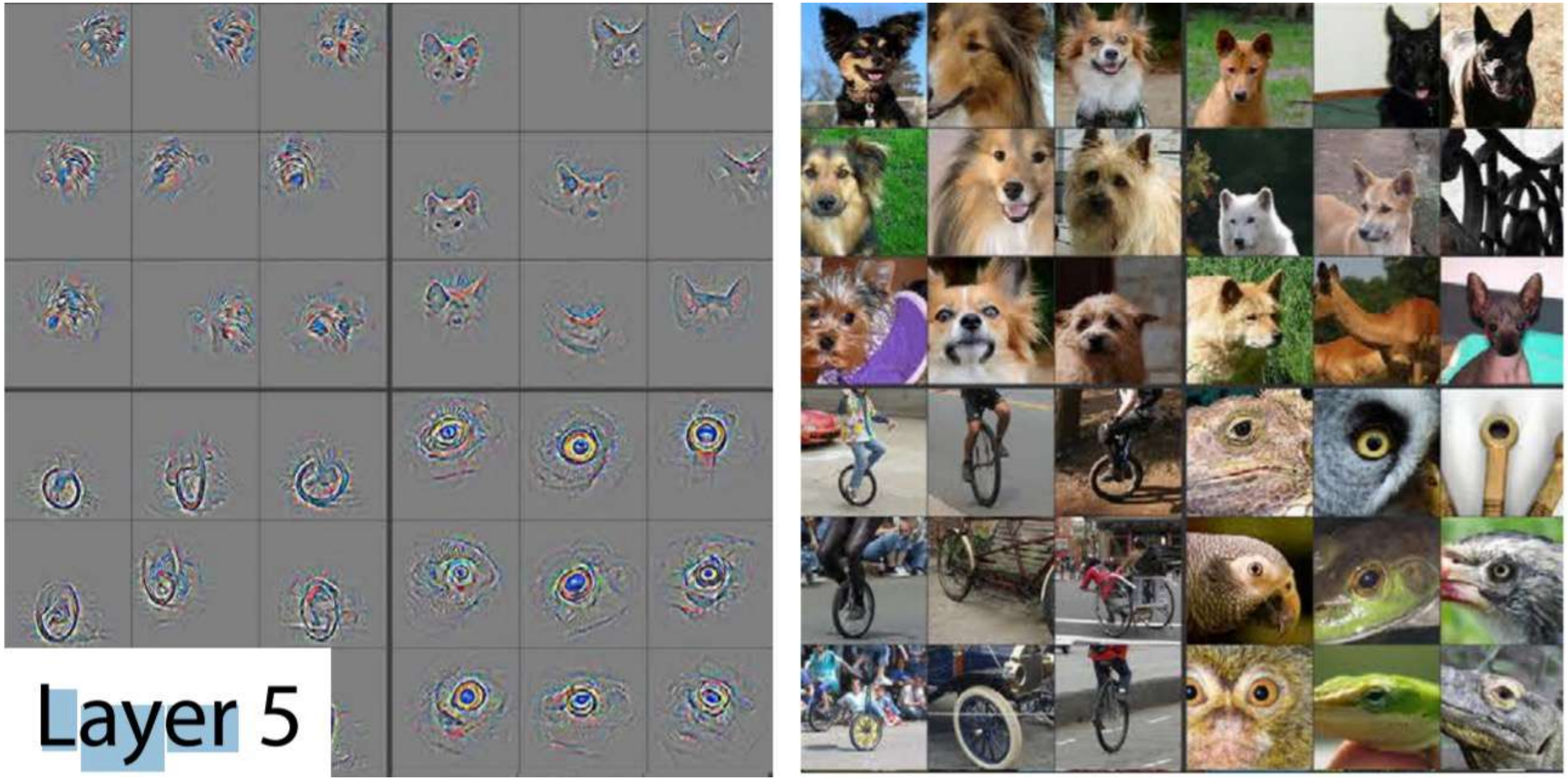
- **Maxpooling**: The authors used a module called switch to recover the positions of maxima in the forward pass because the pooling operation is noninvertible.



Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, 2013, arXiv:1311.2901

Saliency Maps with DeconvNet

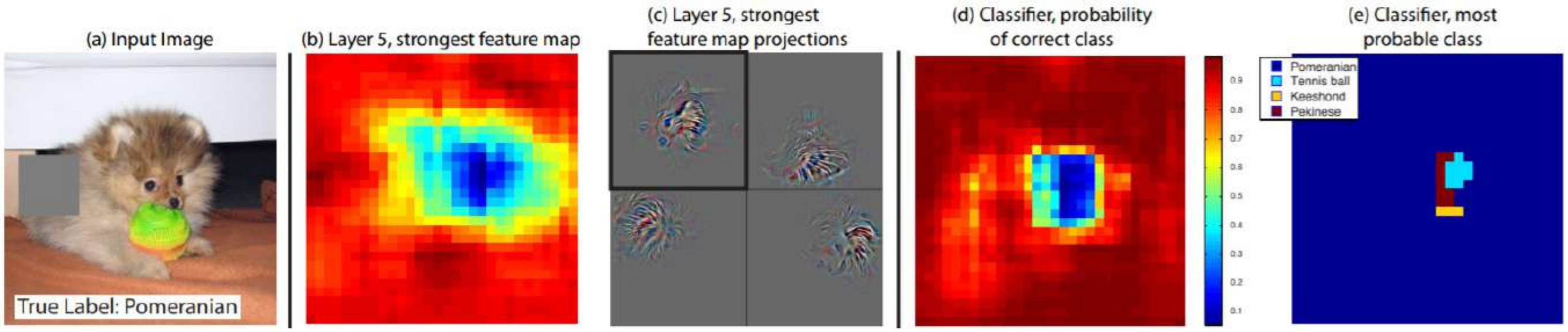
- Results of the reconstruction of the input image from the activations of the fifth layer.



Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, 2013

Saliency Maps with DeconvNet <cont>

- Is the model truly identifying the location of the object in the image, or just using the surrounding context?
- The examples show the model is localizing the objects within the scene, as the probability of the correct class drops when the object is occluded.



Zeiler and Fergus, Visualizing and Understanding Convolutional Networks, 2013

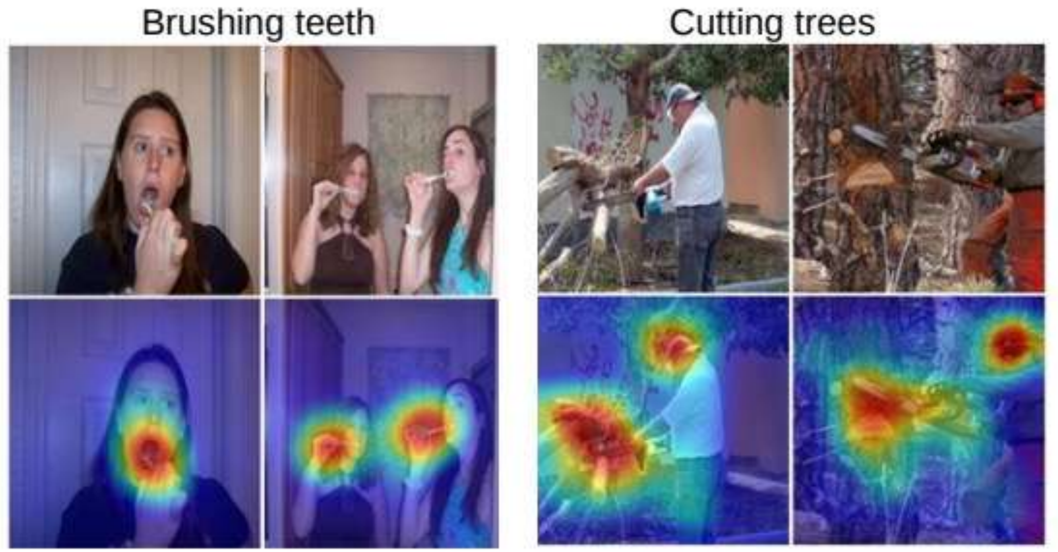
Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Class Activation Maximization (CAM)

Class Activation Maximization is another explanation method for interpreting convolutional neural networks (CNNs) introduced by Zhou et al. 2016.

They proposed a network where the fully connected layers at the very end of the model has been replaced by a layer named Global Average Pooling (GAP) and combined with a c

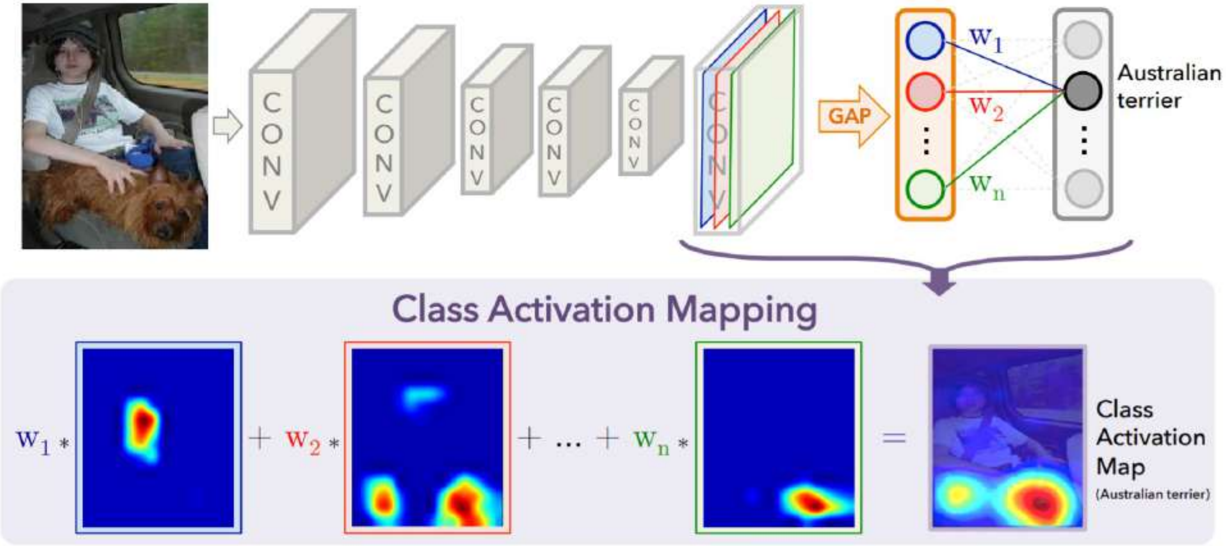


- A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category

Class Activation Maximization (CAM)

- The GAP averages the activations of each feature map, concatenates and outputs them as a vector.

- Then, a weighted sum of the resulted vector is fed to the final softmax loss layer.

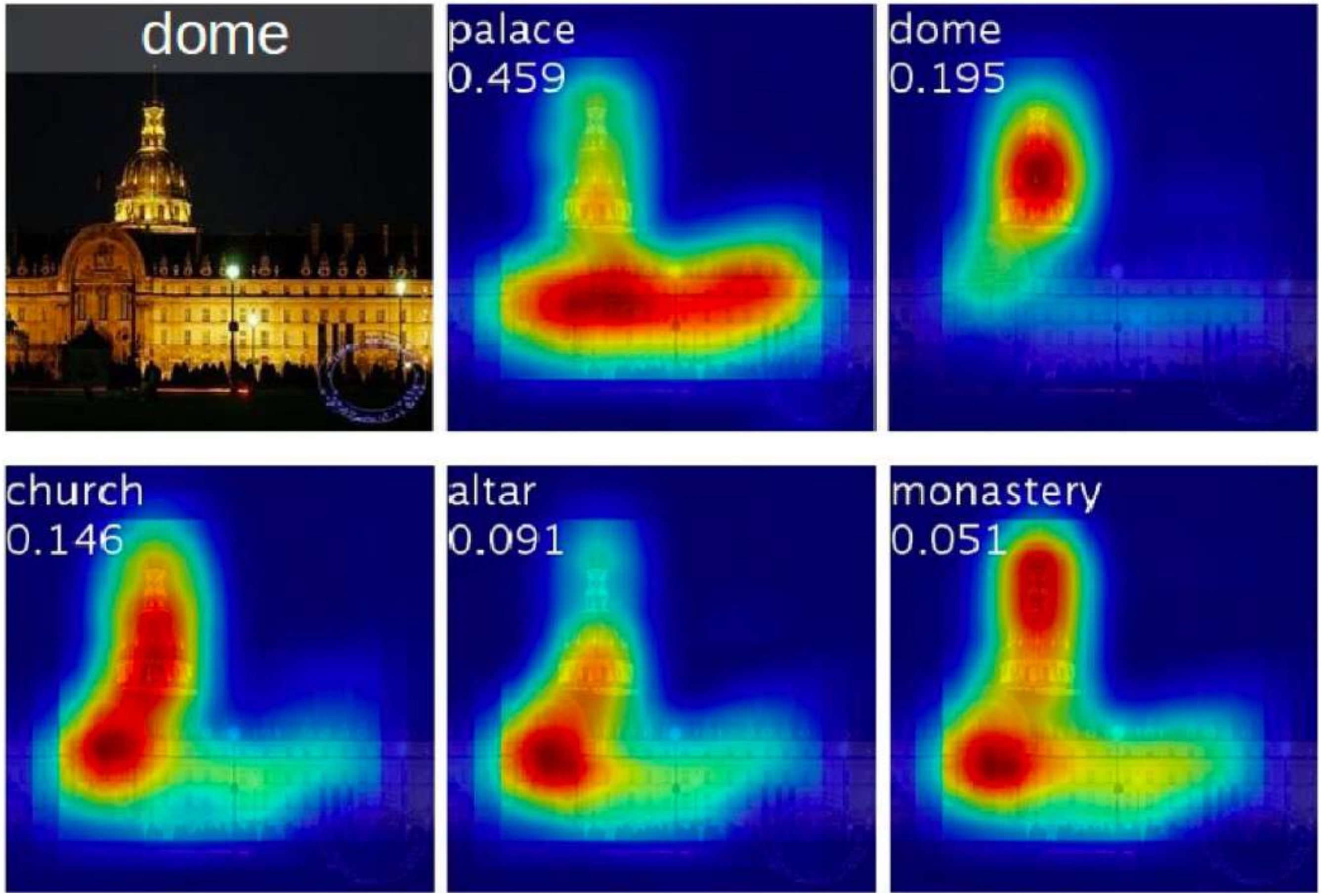


- **Compute a weighted sum of the feature maps of the last convolutional layer to obtain the class activation maps**

Zhou et al., Learning Deep Features for Discriminative Localization, 2016

01:51

CAM



Examples of the CAMs generated from the top 5 predicted categories with ground-truth as dome.
The predicted class and its score are shown above each class activation map.
The highlighted regions vary across predicted classes e.g., dome activates the upper round part while palace activates the lower flat part of the compound.

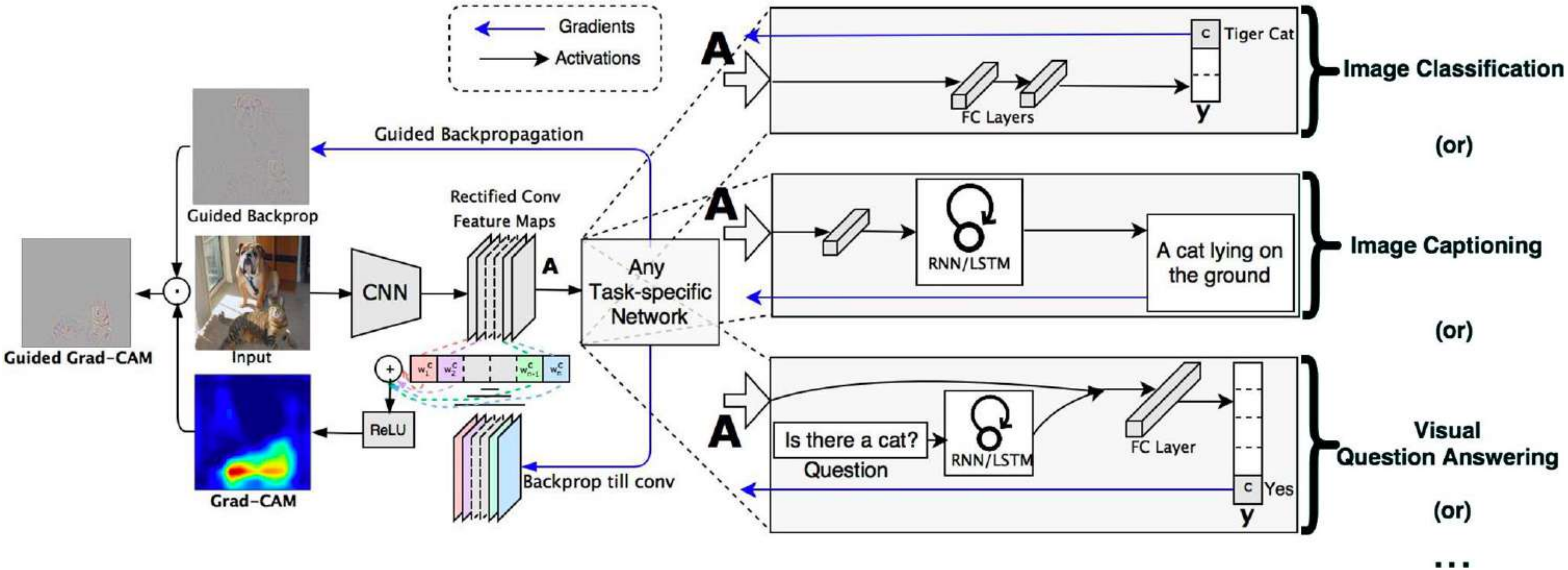
More on Class Activation Maximization

Other approaches to Class Activation Maximization have been developed by Selvaraju et al. 2016:

- **Grad-CAM**: is a more versatile version of CAM that can produce visual explanations for any arbitrary CNN, even if the network contains a stack of fully connected layers too (e.g. the VGG networks);

- **Guided Grad-CAM**: by adding an element-wise multiplication of guided-backpropagation visualization.

Grad-CAM



- Guided Backpropagation visualizes gradients with respect to the image where negative gradients are suppressed when backpropagating through ReLU layers.
- Intuitively, this aims to capture pixels detected by neurons, not the ones that suppress neurons.

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017. (cited 17727)

Class Activation Maximization (Grad-CAM)

- Similarly to saliency maps:

1. we let the gradients of any target concept score flow into the final convolutional layer;

2. compute an importance score based on the gradients;

- produce a coarse localization map highlighting the important regions in the image for predicting that concept.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

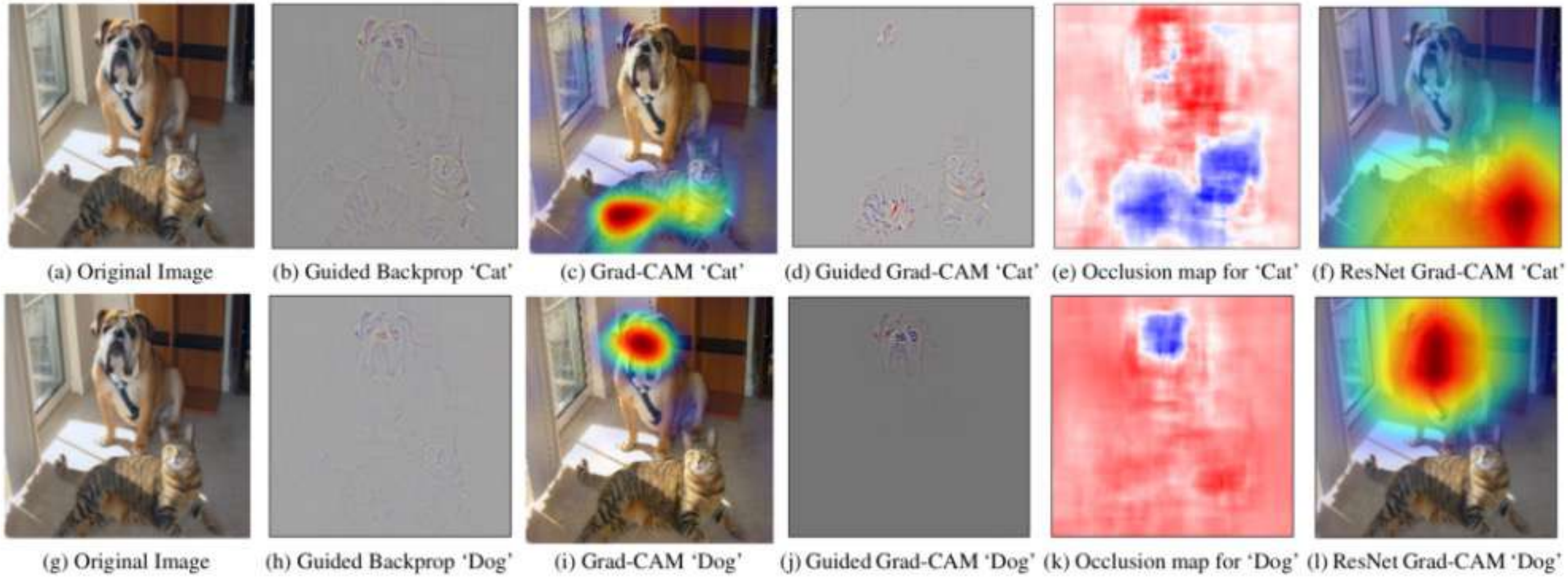
More formally, at first, the gradient of the logits of the class c w.r.t the activations maps of the final convolutional layer is computed and then the gradients are averaged across each feature map to give us an importance score.

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

k: index of the activation map in the last convolutional layer;
c: is the class of interest;
α: computed above shows the importance of feature map k for the target class c.

Class Activation Maximization Recap

- Both CAM and Grad-CAM are local backpropagation-based interpretation methods.
- They are model-specific as they can be used exclusively for interpretation of convolutional neural networks.



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

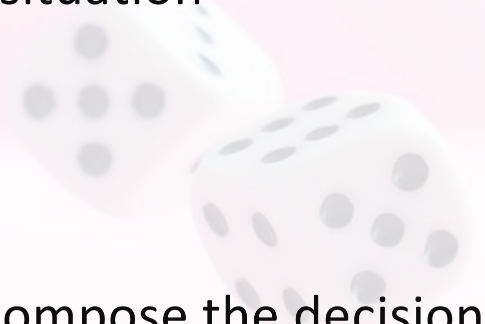
- (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation: highlights all contributing features. (c, f) Grad-CAM: localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualisations.
- Interestingly, the localizations achieved by Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer.
- Note that in (c, f, i, l), red regions correspond to high score for class, while in (e, k), blue corresponds to evidence for the class.

Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques

Why Care About Interpretability?

- 1. Help building trust:
  - • Humans are reluctant to use ML for critical tasks
  - • Fear of unknown when people confront new technologies
- 2. Promote safety:
  - Shift in distributions
  - • Explain model’s representation (i.e. important feature) providing opportunities to remedy the situation
- 3. Allow for contestability:
  - • Black-box models don't decompose the decision into submodels or illustrate a chain of reasoning



theGradient.pub, Interpretability in Machine Learning: An Overview, 2020

Importance

Until recently, humans had a monopoly on agency in society

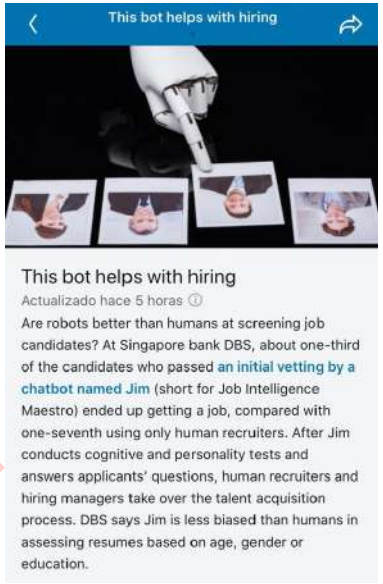
Today, we leave more and more the AI algorithms to take decisions.

If you applied for a job, loan, or bail, a human decided your fate.

If you went to the hospital, a human would attempt to categorize your malady and recommend treatment.

The *reasons* for a decision often matter.

However, today's predictive models are not capable of reasoning at all.



This bot helps with hiring

Actualizado hace 5 horas ⓘ

Are robots better than humans at screening job candidates? At Singapore bank DBS, about one-third of the candidates who passed an initial vetting by a chatbot named Jim (short for Job Intelligence Maestro) ended up getting a job, compared with one-seventh using only human recruiters. After Jim conducts cognitive and personality tests and answers applicants' questions, human recruiters and hiring managers take over the talent acquisition process. DBS says Jim is less biased than humans in assessing resumes based on age, gender or education.

FDA permits marketing of AI software that autonomously detects diabetic retinopathy

By Dave Muoio | April 12, 2018 | 11:53 am

SHARE

The FDA has granted diagnostics company IDx's De Novo request to market its AI-based software system for the autonomous detection of diabetic retinopathy in adults who have diabetes, called IDx-DR.

This decision represents the first AI-based diagnostic system authorized by the FDA for commercialization in the US that can provide a screening decision without the need for

AI does not give the reasons for the decision and it can matter.

- For example if I know what are the risk factors to avoid a heart attack taking into account I'm in a high risk of it, probably I can get prevention measures.
- A doctor's explanation might educate you about your condition.

01:51

Interpretability and Explainability

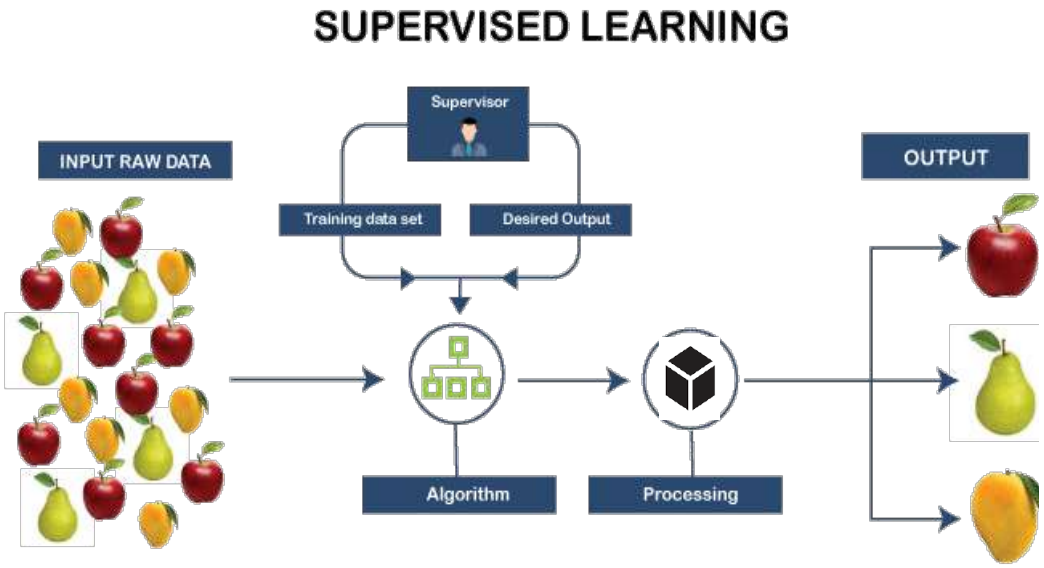
- Interpretability: to establish a cause and effect relationship
- Explainability: to explain the internal mechanics of ML or DL in human terms
- Most of the time, the terms are interchangeable
  - Interpretability, explainability, explainable AI, XAI

Machine learning

ML algorithm is trained to take some input and predict the corresponding output

ML-based systems do not know why a given input should receive some label, only that certain inputs are correlated with that label.

As ML penetrates critical areas such as medicine, the inability of humans to understand these models seems problematic.



Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Desiderata of Interpretability Research: Transparency vs black boxes

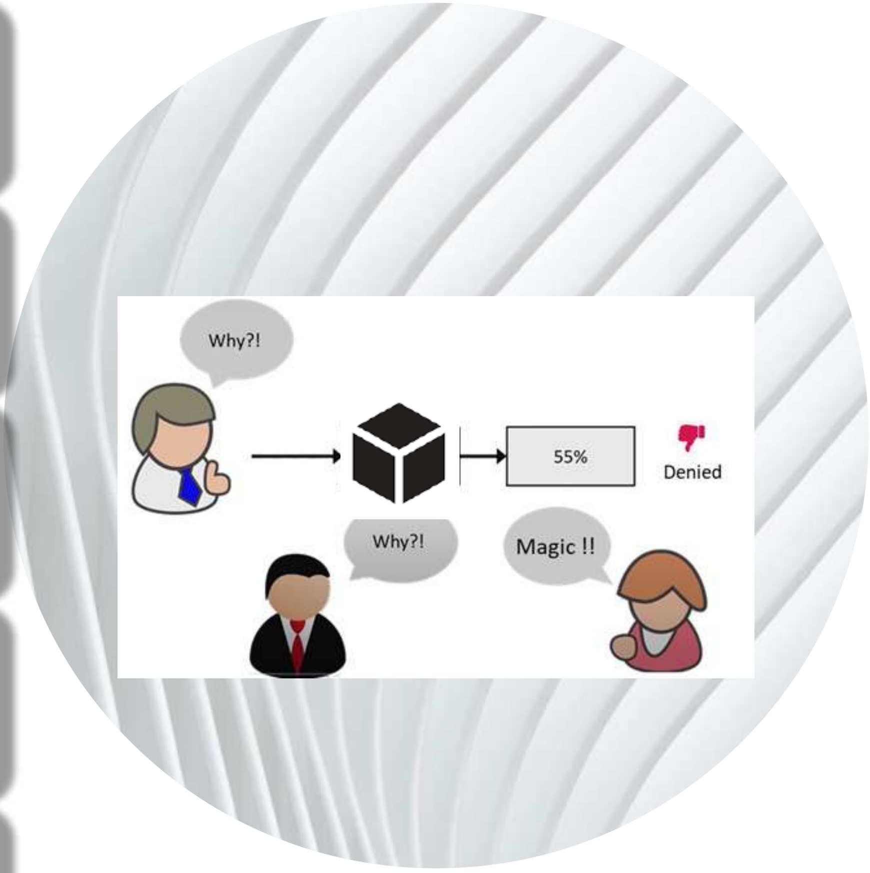
But what constitutes transparency?

Will it converge?

Does it produce a unique solution?

Do you understand what each parameter represents?

Or model's complexity: Is it simple enough to be examined all at once by a human?

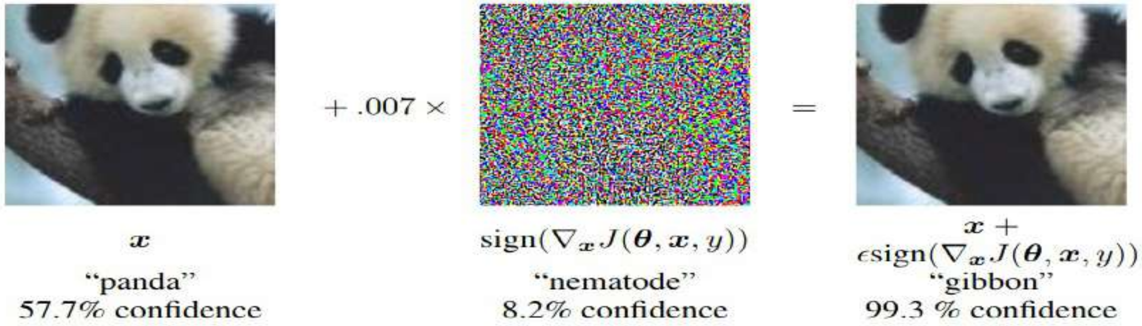


Desiderata of Interpretability Research: Transferability

Typically, training and test data are chosen by randomly partitioning examples from the same distribution.

A model's generalization error is then judged by the gap between its performance on training and test data.

Humans exhibit a far richer capacity to generalize, transferring learned skills to unfamiliar situations.


$\boldsymbol{x}$
"panda"
57.7% confidence
$+ .007 \times$
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$=$
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Images By *Goodfellow et al, ICLR 2015.* *Explaining and Harnessing Adversarial Examples*

Desiderata of Interpretability Research: Transparency

Debugging

Understand why a system doesn’t work

- Model has good accuracy in validation set,

but it is not performing well in the wild

Science

Comprehend how the system took that conclusion and learn with it

- explain all factors, understand the model, learn new relations

Desiderata of Interpretability Research: Informativeness

Informing feature engineering

Create new features using transformations of your raw data or other features to improve model accuracy

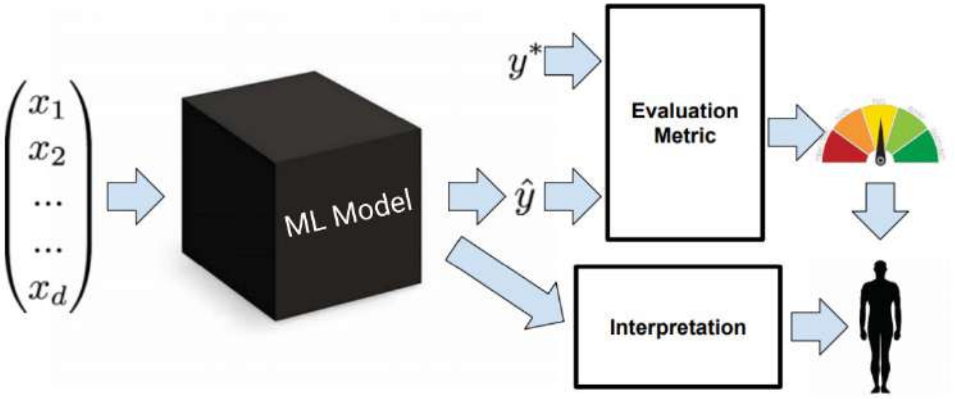
Directing future data collection

Understand the value of the features that you already have and, therefore, decide what new values will be more helpful

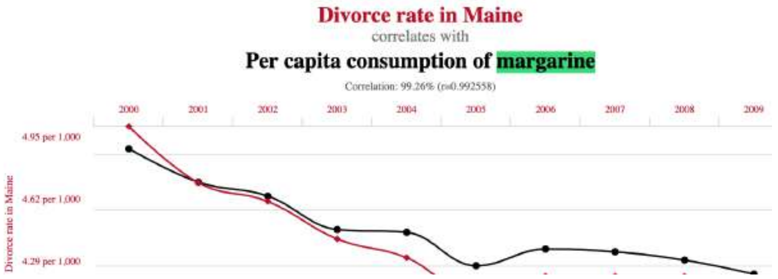
Desiderata of Interpretability Research: Causality

Standard ML problem formulations, e.g. maximizing accuracy on a set of hold-out data for which the training data is perfectly representative, are imperfectly matched to the complex real-life tasks they are meant to solve. (e.g. medical longitudinal studies).

An interpretable model is desirable because it might help uncover causal structure in observational data (e.g. smoking and lung cancer)



Causality <> correlation



Desiderata of Interpretability Research: Trust

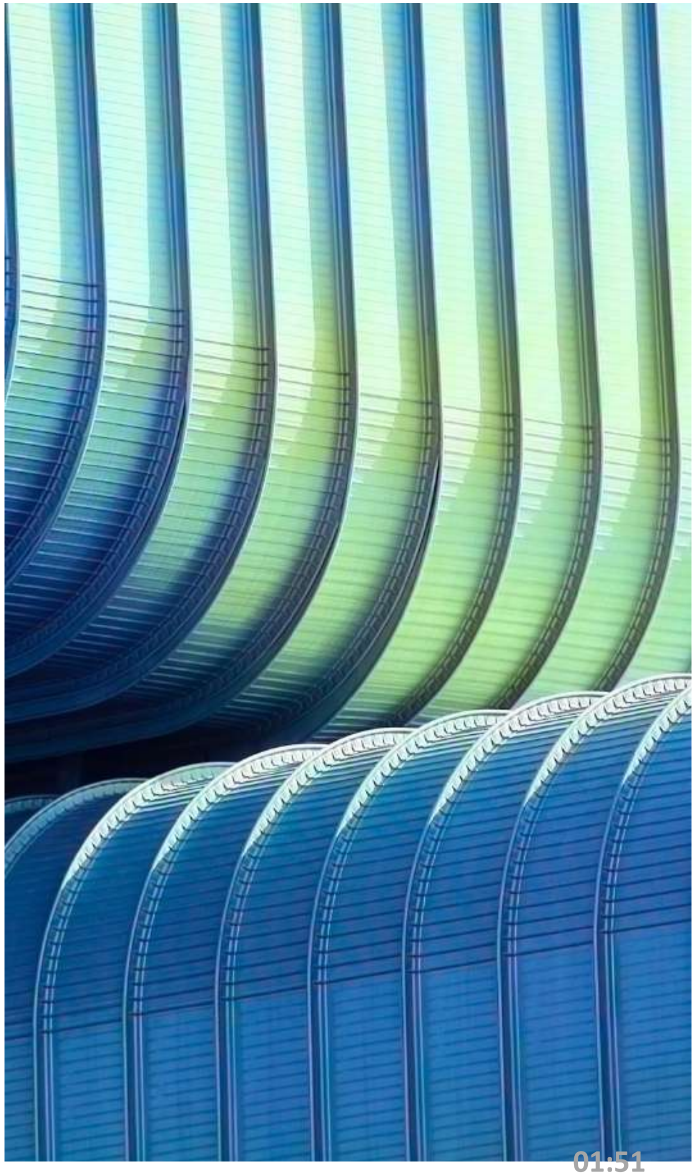
Is it simply confidence that a model will perform well?

If so, a sufficiently accurate model should be demonstrably trustworthy, and interpretability would serve no purpose.

The user is comfortable with relinquishing control to it.

- care not only about *how often* a model is right,
- but also *for which examples* it is right.

Trust in models that make mistakes where human do.



01:51

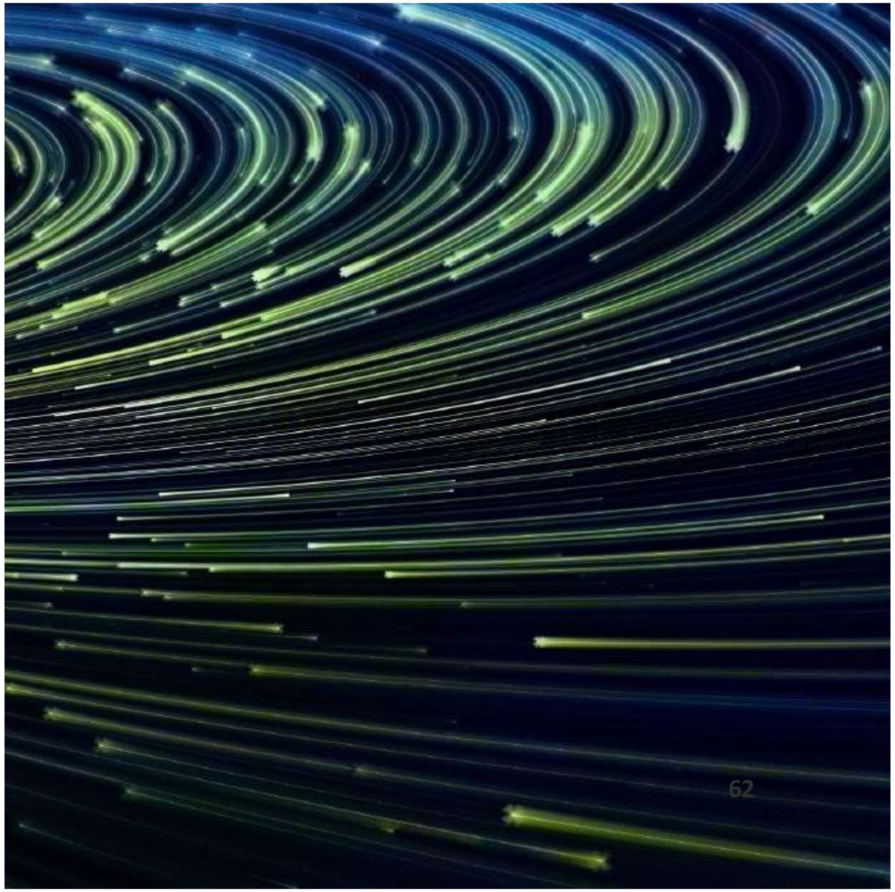
Desiderata of interpretability research: Ethics

REAL-WORLD OBJECTIVES ARE DIFFICULT TO ENCODE AS SIMPLE MATHEMATICAL FUNCTIONS

AN ALGORITHM FOR MAKING HIRING DECISIONS SHOULD SIMULTANEOUSLY OPTIMIZE PRODUCTIVITY, ETHICS, AND LEGALITY.

BUT HOW WOULD YOU GO ABOUT WRITING A FUNCTION THAT MEASURES ETHICS OR LEGALITY?

HOW TO ENCODE ROBUSTNESS TO CHANGES IN THE DYNAMICS BETWEEN THE TRAINING AND DEPLOYMENT ENVIRONMENTS?



Desiderata of Interpretability Research: Fairness and ethics

Fairness: Concern that interpretations must be produced for assessing whether decisions produced automatically by algorithms conform to ethical standards

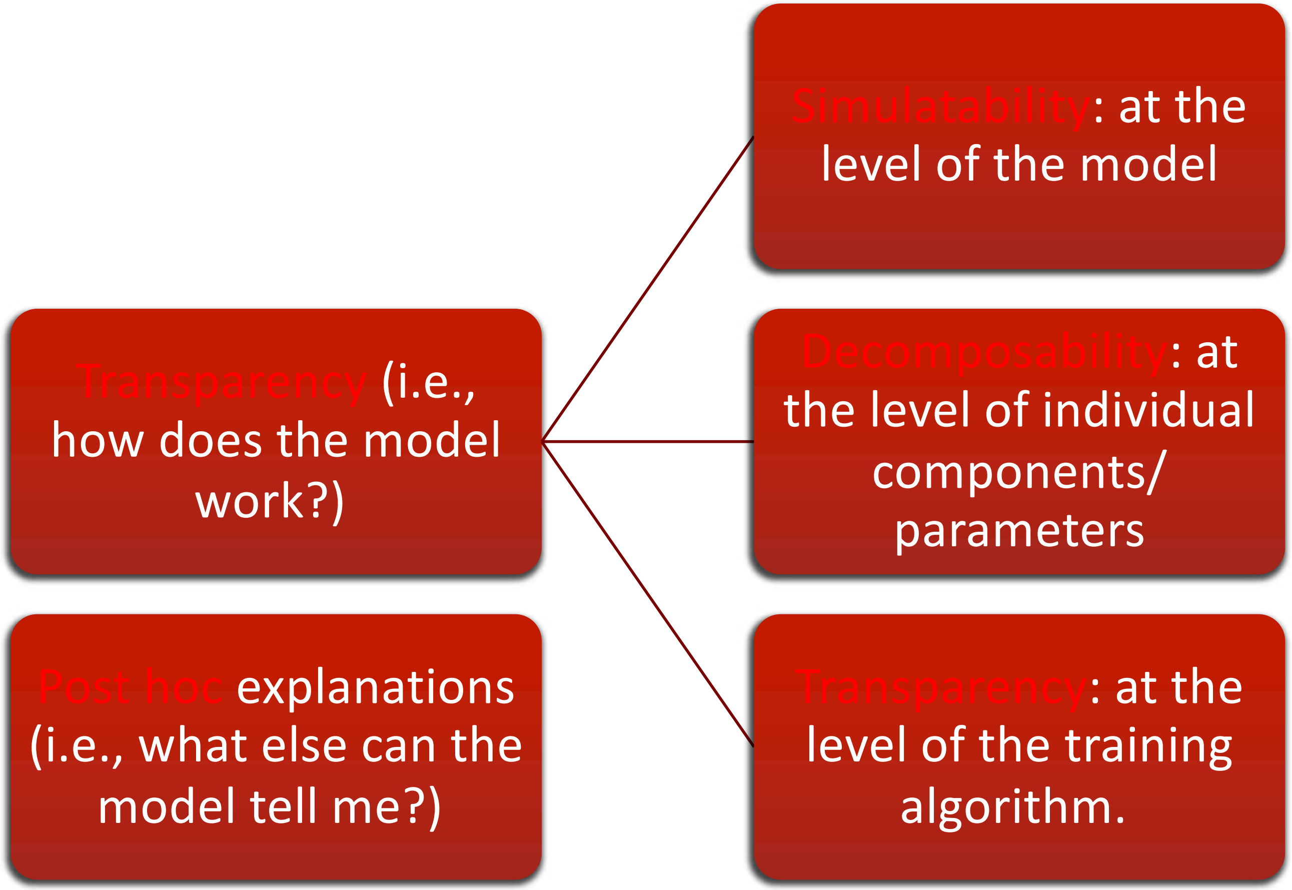
- How can you be sure that predictions do not discriminate on the basis of race (skin cancer)?

Safety: Make sure the system is taking safe decisions

Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Types of Interpretability techniques



Transparency (i.e., how does the model work?)

Post hoc explanations (i.e., what else can the model tell me?)

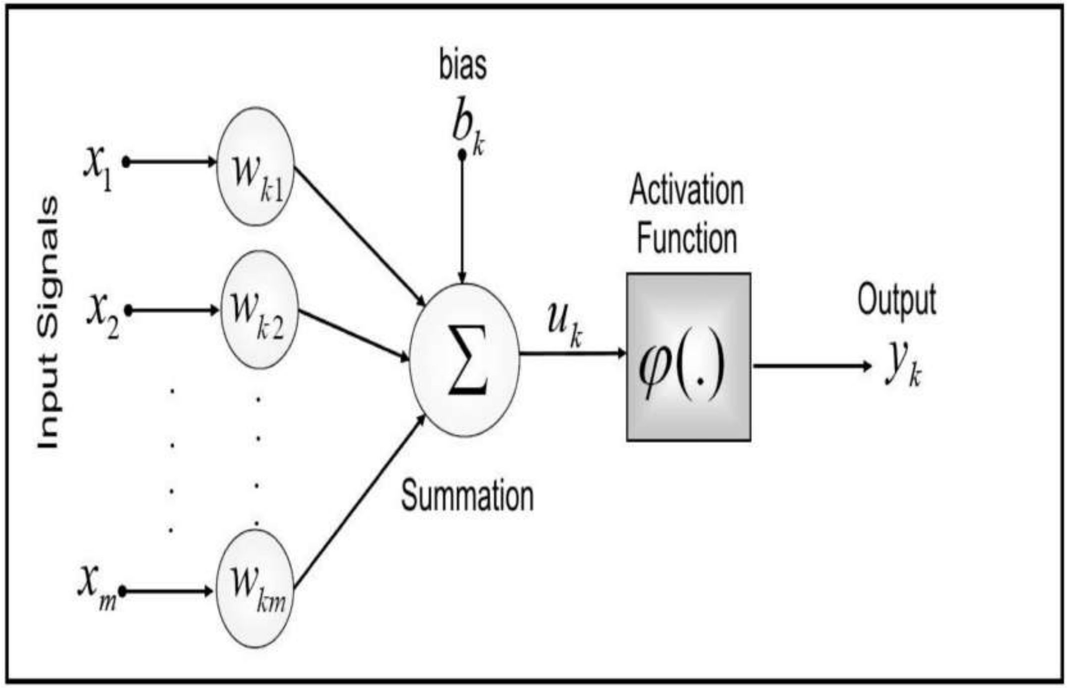
Simulatability: at the level of the model

Decomposability: at the level of individual components/ parameters

Transparency: at the level of the training algorithm.

Simulatability

- Transparent model: if a person can contemplate the entire model at once
  - -> simple model.
  - Additive model
  - Model easily visualized

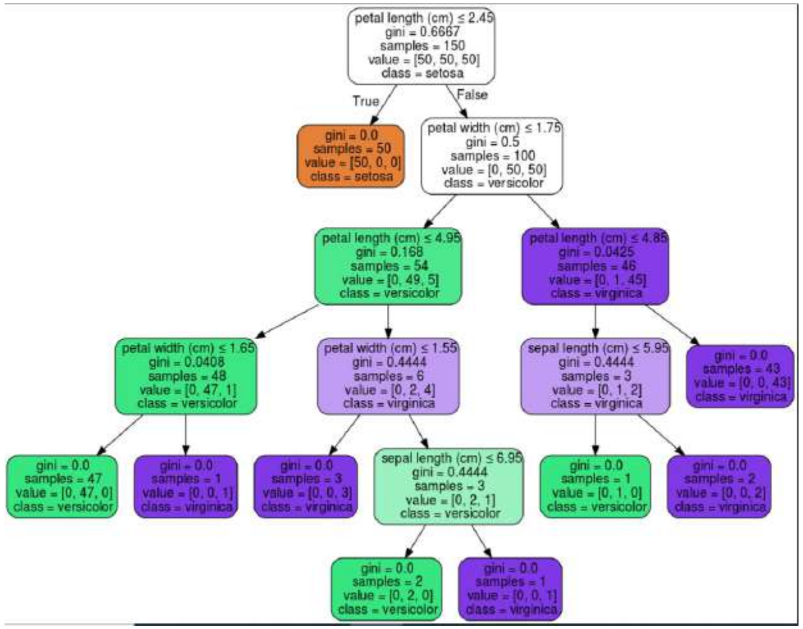


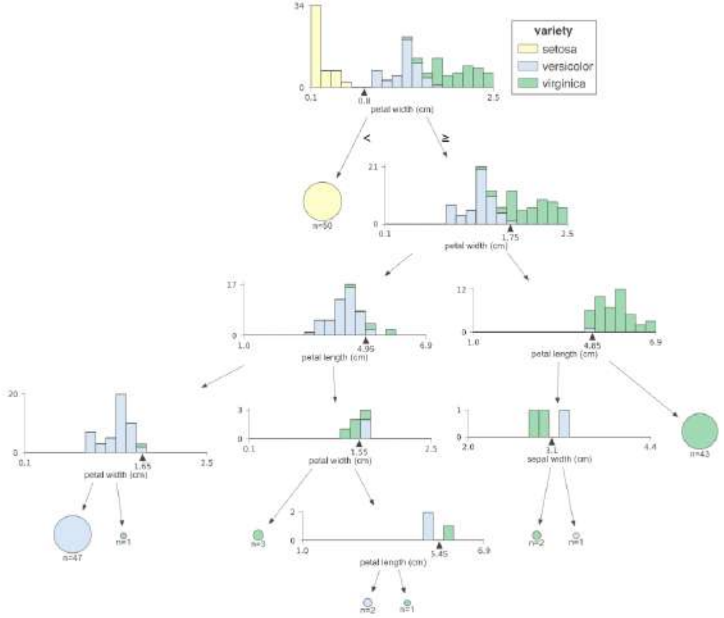
Defining Interpretability: Decomposability

- What is it? This property addresses what the model is doing at each step.

- Decision trees are the most interpretable. How can we make it clear at each step?

Decision Trees on Iris Dataset



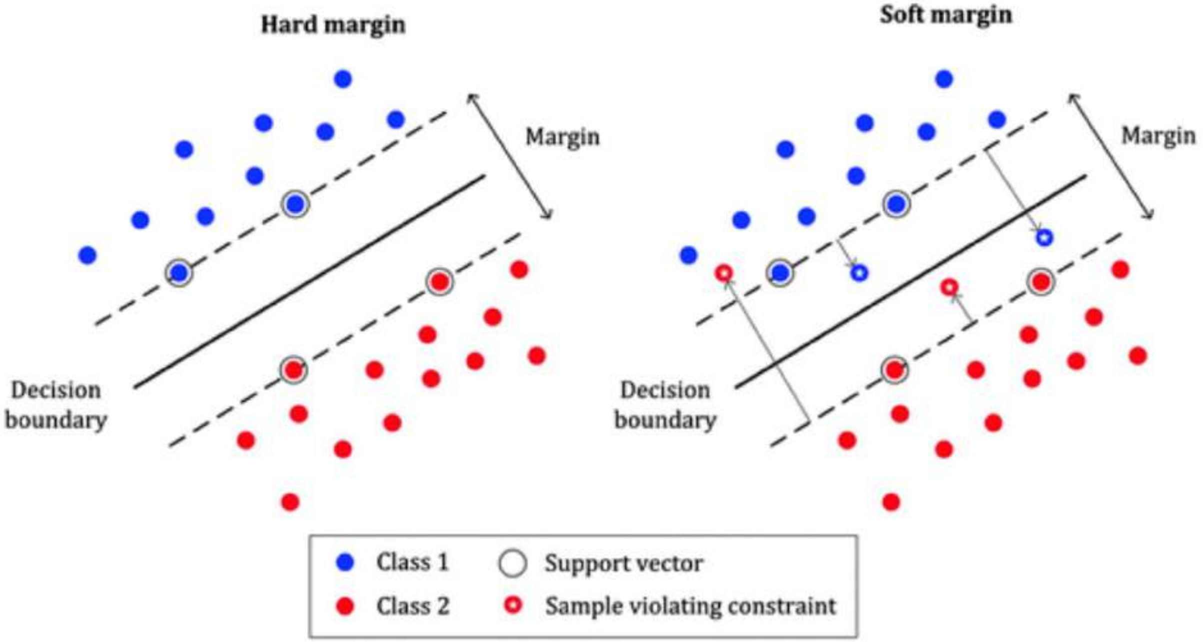


explained.ai, How to Visualize Decision Trees, sklearn versus dtreeviz

01:51

Defining Interpretability: Algorithmic Transparency

- What is it? This property addresses whether the model has any desirable properties which are easy to understand. Good reparameterization, what else?



- Bottom line, only some models are informative from a simulatable perspective.

theGradient.pub, Interpretability in Machine Learning: An Overview, 2020

Post hoc interpretability

Post hoc interpretations do not elucidate precisely how a model works

They confer useful information for practitioners and end users of ML.

natural language explanations,

visualizations of learned representations or models, and

explanations by example

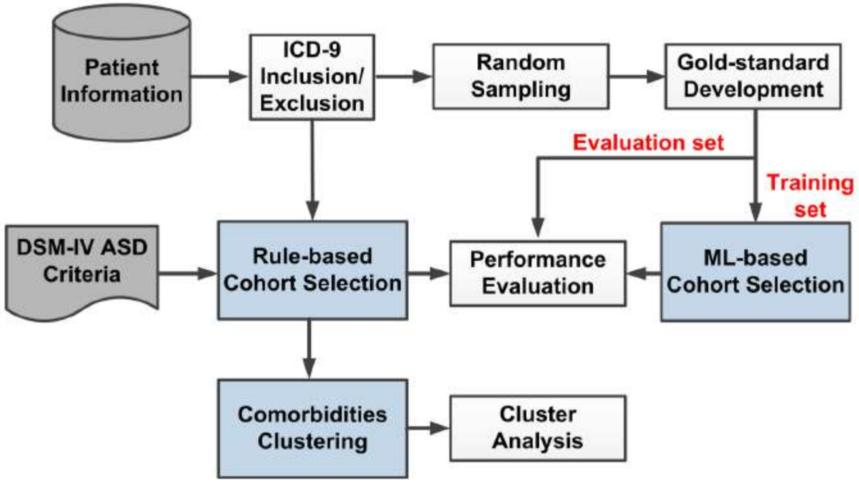
Opaque models can be interpreted, without sacrificing performance.

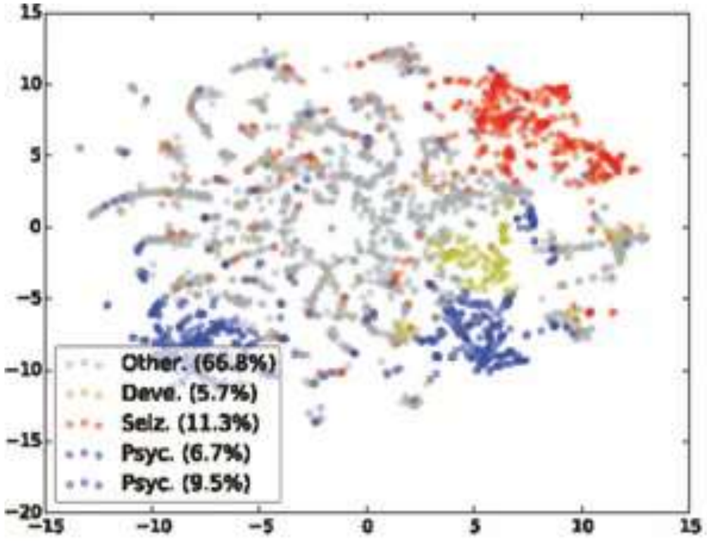
Outline

- Visualization of High-dimensional data
  - PCA
  - tSNE
  - UMAP
- Interpretation Methods for Image Classification
  - Saliency Maps
  - Class Activation Maximization (CAM)
- Interpretability and explainability of NNs
  - Desiderata of interpretable models
  - Types of interpretable models
  - Practical techniques and examples

Visualization

- t-SNE (t-distributed stochastic neighbour embedding), a technique that renders 2D visualizations in which nearby data points are likely to appear close together.

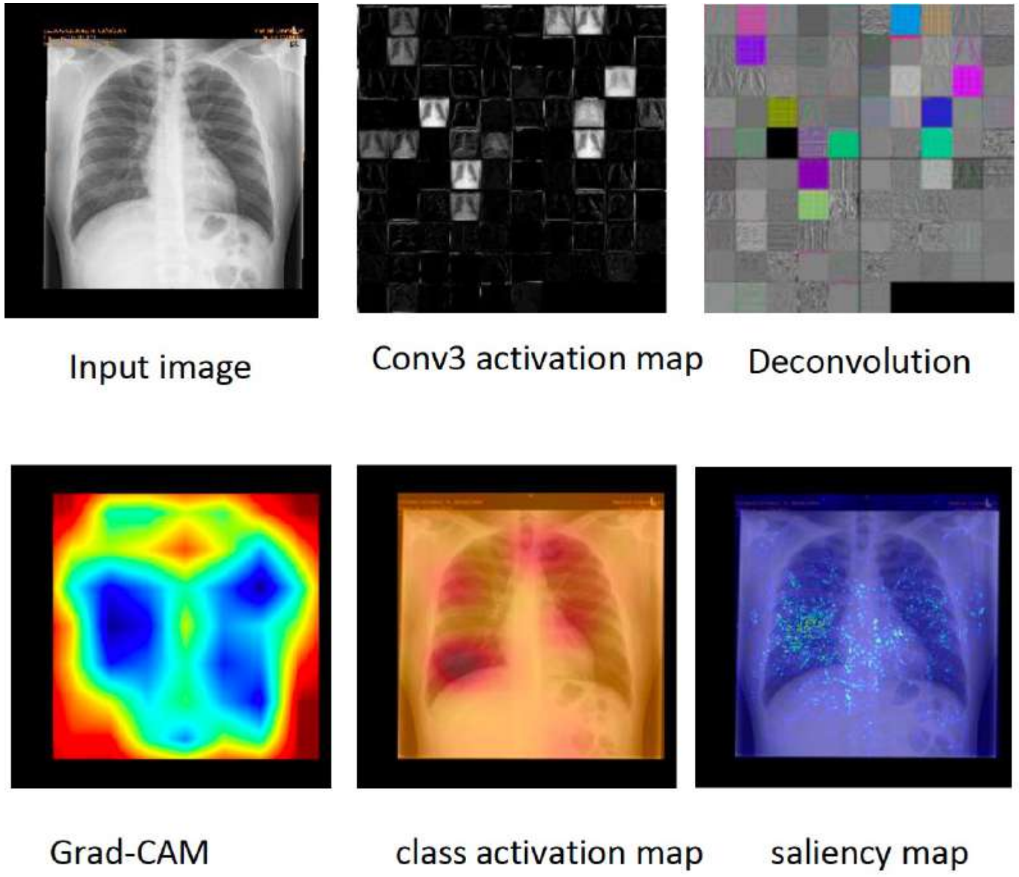




Lingren, Todd, et al. "Electronic health record based algorithm to identify patients with autism spectrum disorder." PloS one 11.7 (2016): e0159621.

Local explanations

- Saliency map: take the gradient of the output corresponding to the correct class with respect to a given input vector.
  - It may be misleading. The saliency map is a local explanation only
- Discriminative Localization / Attention
  - Class Activation Map
  - Gradient-Class Activation Map
  - Deconvolution



Input image

Conv3 activation map

Deconvolution

Grad-CAM

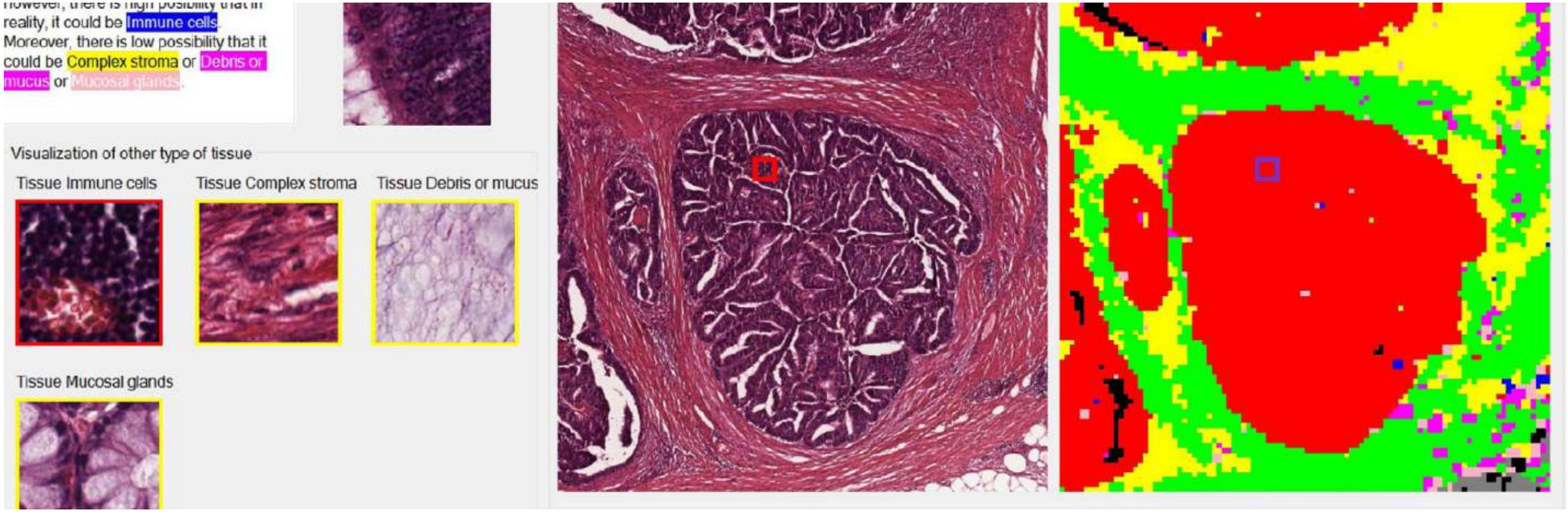
class activation map

saliency map

https://www.reasonfieldlab.com/post/a-complete-guide-on-computer-vision-xai-libraries

https://mrsalehi.medium.com/a-review-of-different-interpretation-methods-in-deep-learning-part-1-saliency-map-cam-grad-cam-3a34476bc24d

Explanation by example



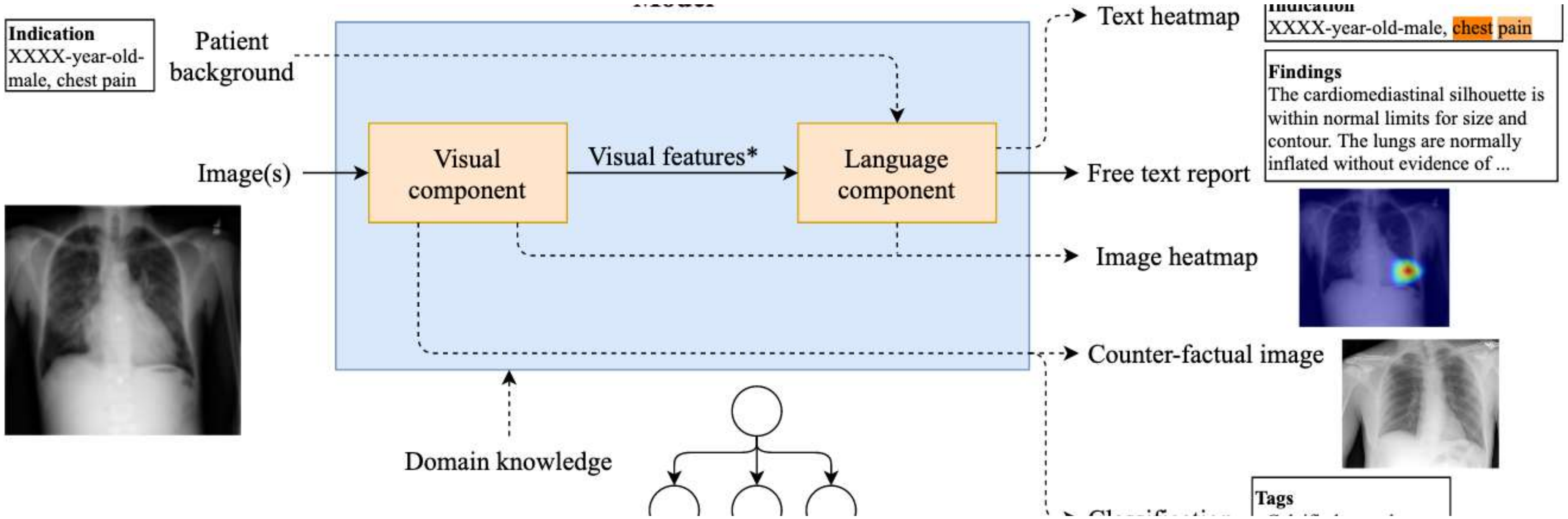
however, there is high possibility that in reality, it could be Immune cells. Moreover, there is low possibility that it could be Complex stroma or Debris or mucus or Mucosal glands.

Case-based reasoning approaches for interpreting generative models.

01:51

Sabol, Patrik, et al. "Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images." *Journal of Biomedical Informatics* 109 (2020): 103523.

Text explanations



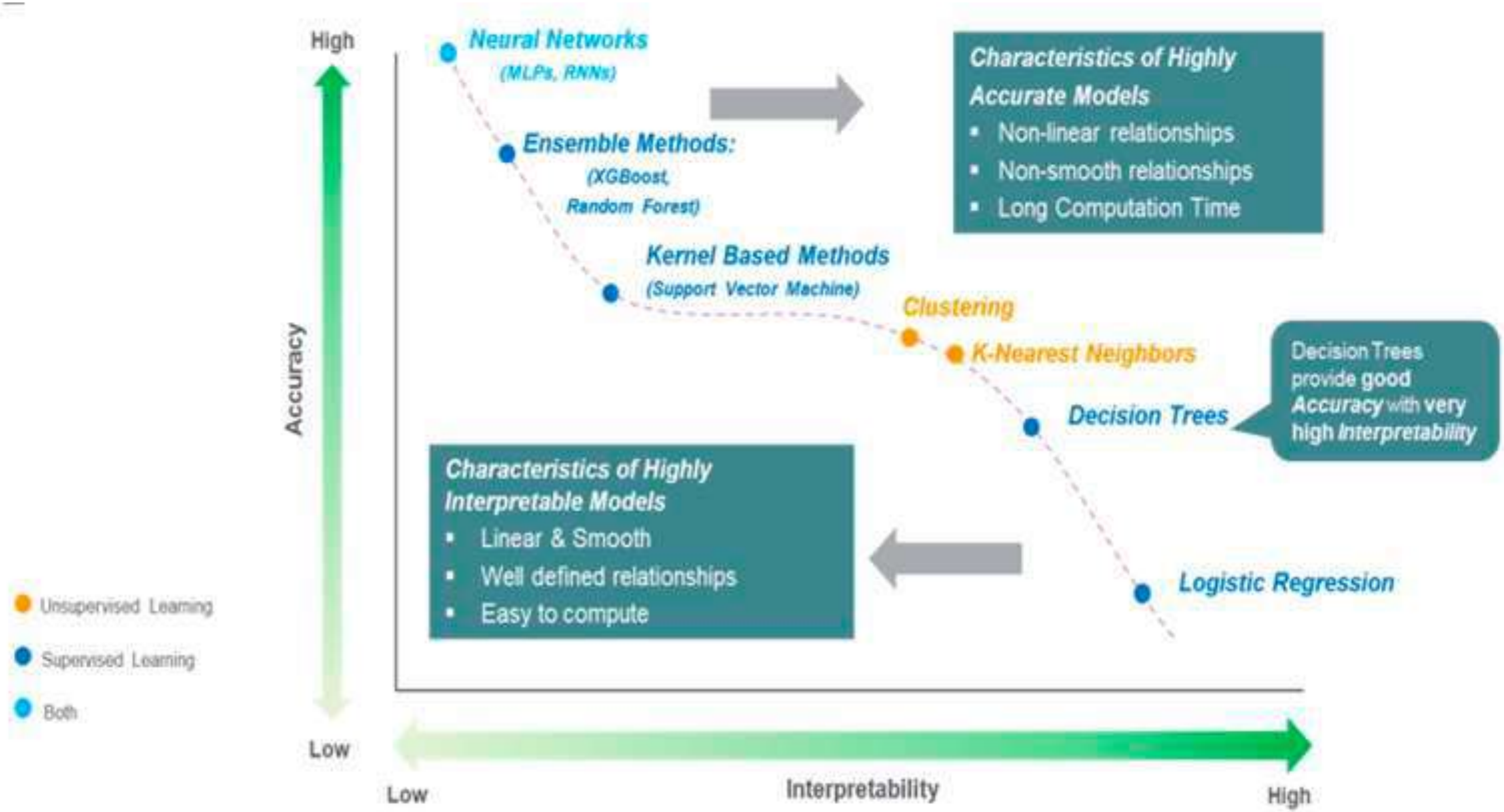
Findings
The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of ...

- One model might be trained to generate predictions, and a separate model, such as a recurrent neural network language model, to generate an explanation.

- Neural image captioning in which the representations learned by a discriminative CNN are coped by a second model to generate captions.

Messina, Pablo, et al. "A Survey on Deep Learning and Explainability for Automatic Image-based Medical Report Generation." *arXiv preprint arXiv:2010.10563* (2020)

Summary



Characteristics of Highly Accurate Models

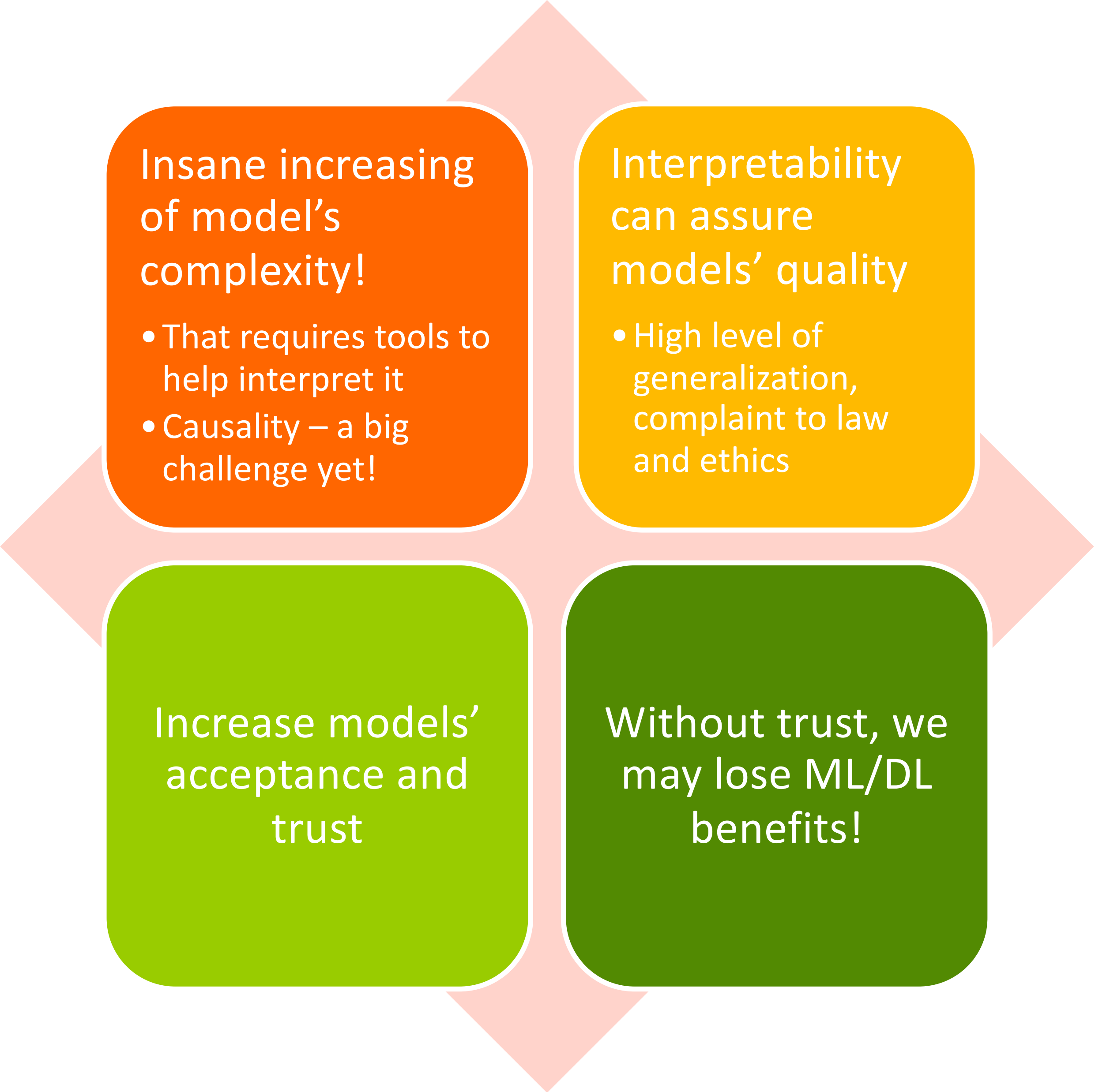
- Non-linear relationships
- Non-smooth relationships
- Long Computation Time

Decision Trees provide good **Accuracy** with very high ***Interpretability***

Characteristics of Highly Interpretable Models

- Linear & Smooth
- Well defined relationships
- Easy to compute

Conclusions



Insane increasing of model's complexity!

- That requires tools to help interpret it
- Causality – a big challenge yet!

Interpretability can assure models' quality

- High level of generalization, complaint to law and ethics

Increase models' acceptance and trust

Without trust, we may lose ML/DL benefits!

Questions?

UNIVERSITAT DE BARCELONA