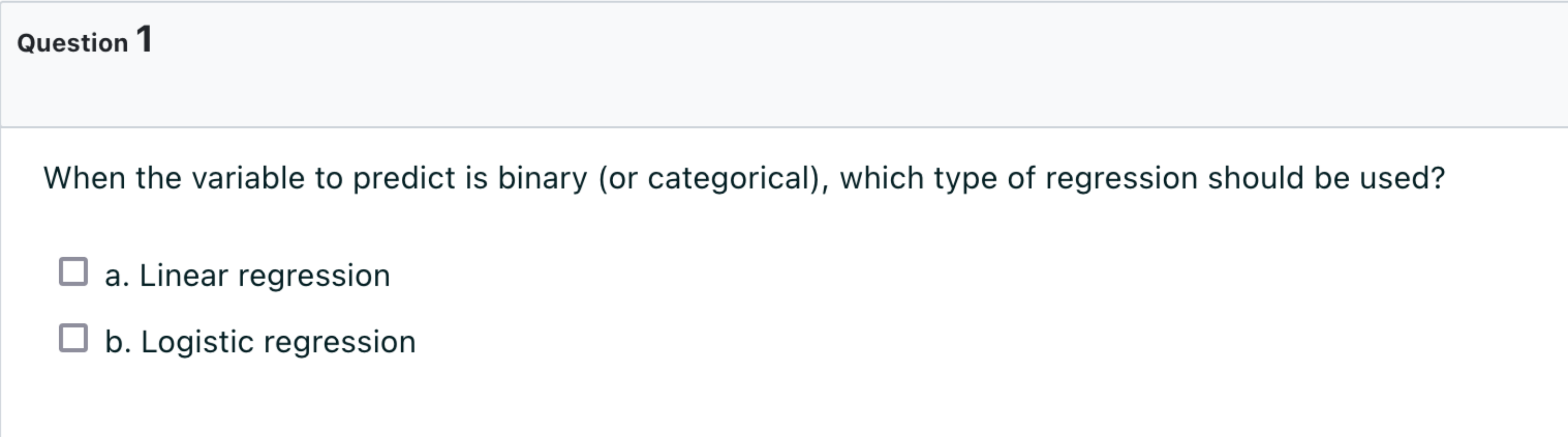
NBE-4070 : Basics of Biomedical Data Analysis

Stéphane Deny
Prof. in Neuroscience and Biomedial Engineering and Computer Science
Aalto University

Lecture 7: Non-linear methods for data analysis

Quiz

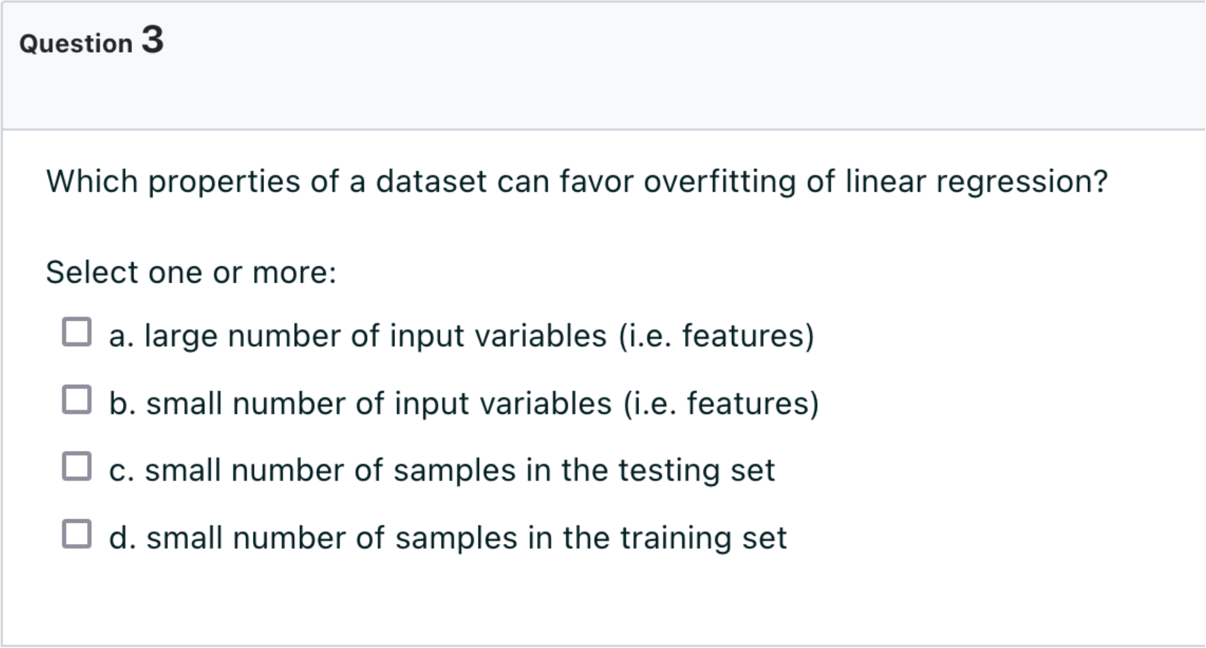


Question 1

When the variable to predict is binary (or categorical), which type of regression should be used?

□ a. Linear regression

□ b. Logistic regression



Question 3

Which properties of a dataset can favor overfitting of linear regression?

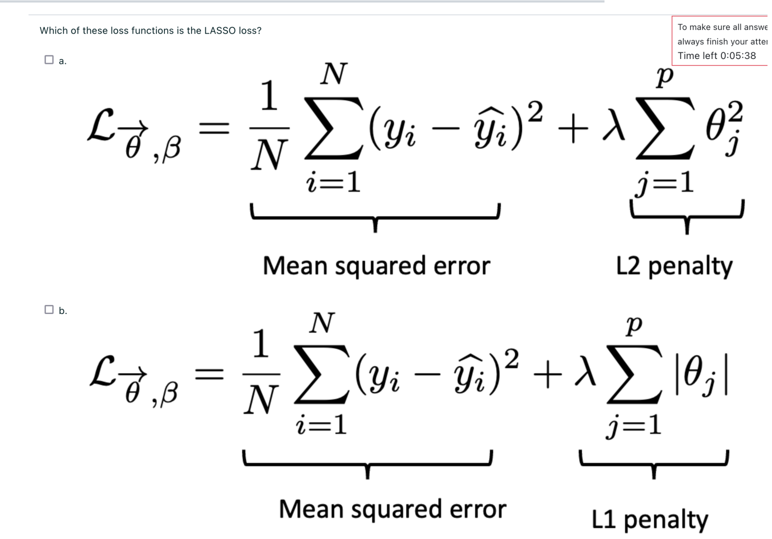
Select one or more:

□ a. large number of input variables (i.e. features)

□ b. small number of input variables (i.e. features)

□ c. small number of samples in the testing set

□ d. small number of samples in the training set



Which of these loss functions is the LASSO loss?

To make sure all answe
always finish your atte
Time left 0:05:38

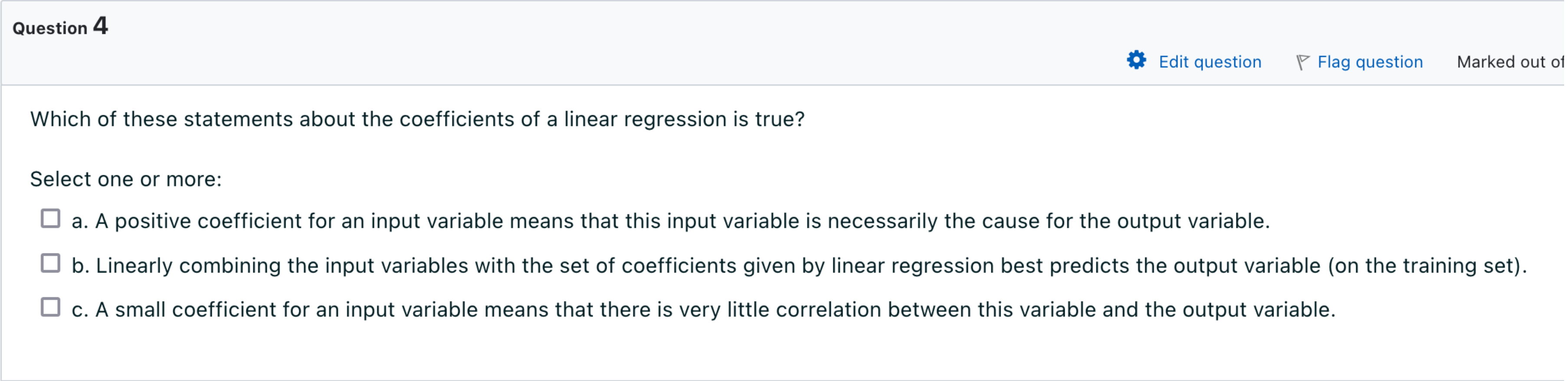
□ a.

$$\mathcal{L}_{\vec{\theta},\beta} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}_{\text{Mean squared error}} + \lambda \underbrace{\sum_{j=1}^{p}\theta_j^2}_{\text{L2 penalty}}$$

□ b.

$$\mathcal{L}_{\vec{\theta},\beta} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}_{\text{Mean squared error}} + \lambda \underbrace{\sum_{j=1}^{p}|\theta_j|}_{\text{L1 penalty}}$$

Quiz



Question 4

Marked out of

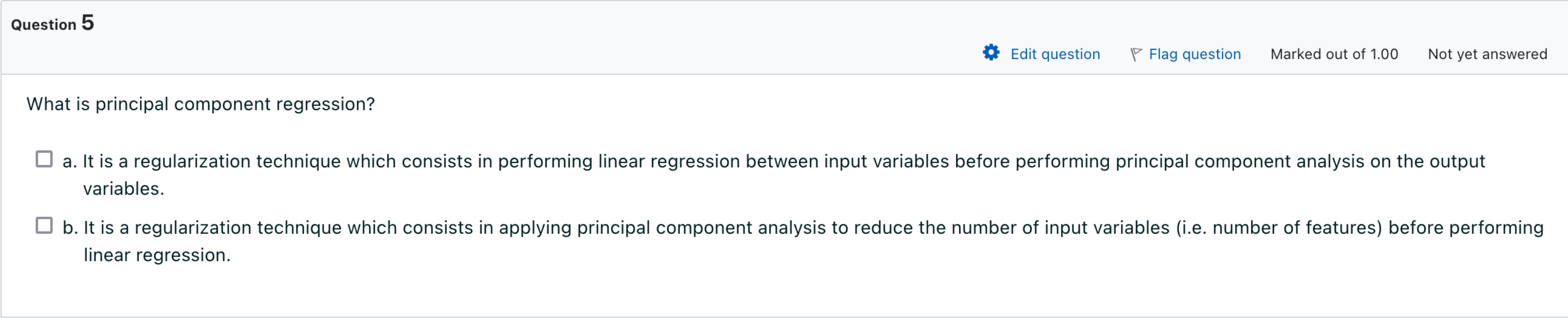
Which of these statements about the coefficients of a linear regression is true?

Select one or more:

□ a. A positive coefficient for an input variable means that this input variable is necessarily the cause for the output variable.

□ b. Linearly combining the input variables with the set of coefficients given by linear regression best predicts the output variable (on the training set).

□ c. A small coefficient for an input variable means that there is very little correlation between this variable and the output variable.



Question 5

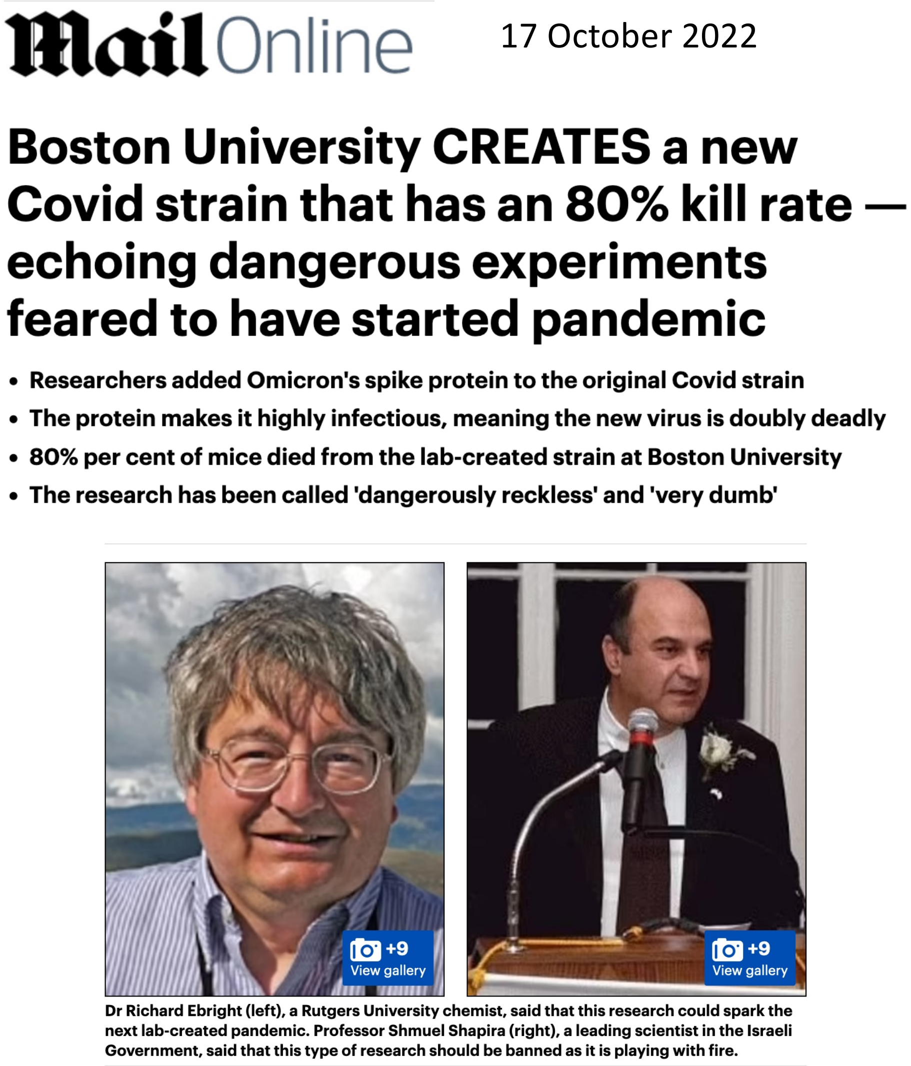
Marked out of 1.00

Not yet answered

What is principal component regression?

□ a. It is a regularization technique which consists in performing linear regression between input variables before performing principal component analysis on the output variables.

□ b. It is a regularization technique which consists in applying principal component analysis to reduce the number of input variables (i.e. number of features) before performing linear regression.



MailOnline

17 October 2022

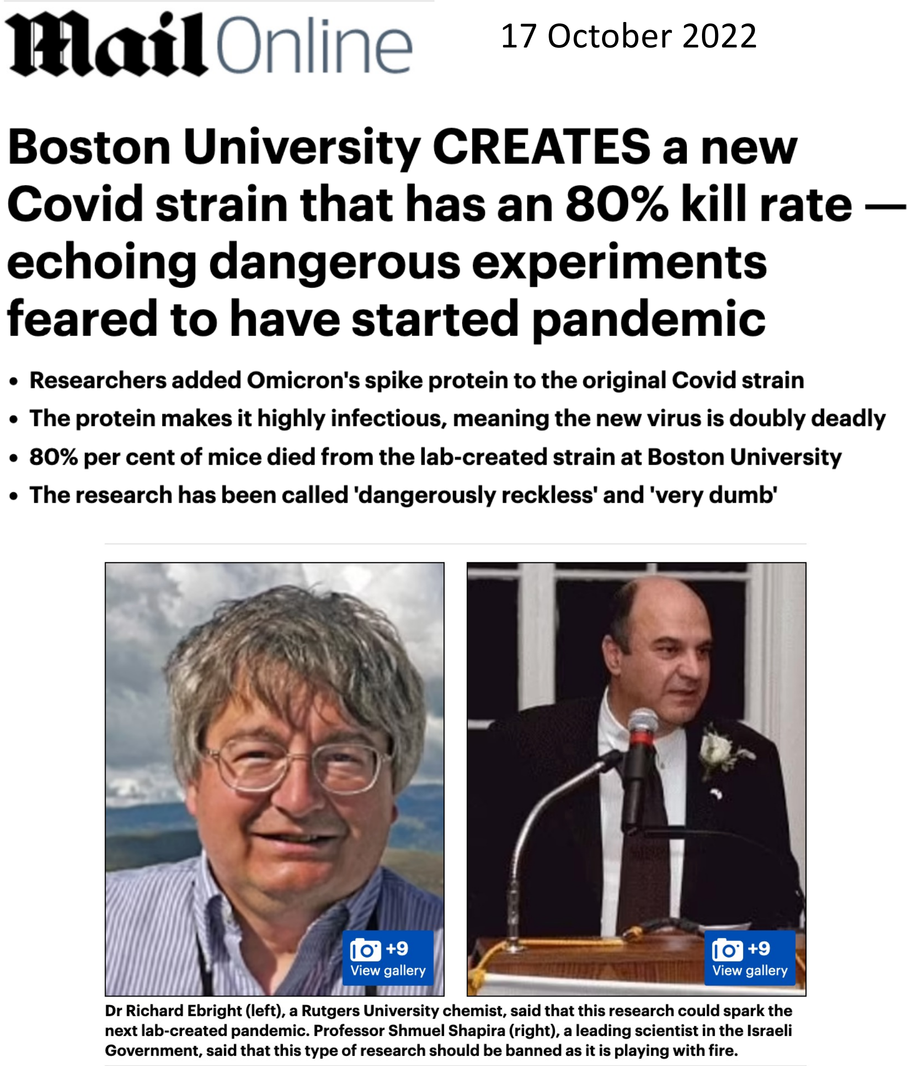
Boston University CREATES a new Covid strain that has an 80% kill rate — echoing dangerous experiments feared to have started pandemic

- Researchers added Omicron's spike protein to the original Covid strain
- The protein makes it highly infectious, meaning the new virus is doubly deadly
- 80% per cent of mice died from the lab-created strain at Boston University
- The research has been called 'dangerously reckless' and 'very dumb'

Dr Richard Ebright (left), a Rutgers University chemist, said that this research could spark the next lab-created pandemic. Professor Shmuel Shapira (right), a leading scientist in the Israeli Government, said that this type of research should be banned as it is playing with fire.

Breaking news!

source: https://www.dailymail.co.uk/health/article-11323677/Outrage-Boston-University-CREATES-Covid-strain-80-kill-rate.html



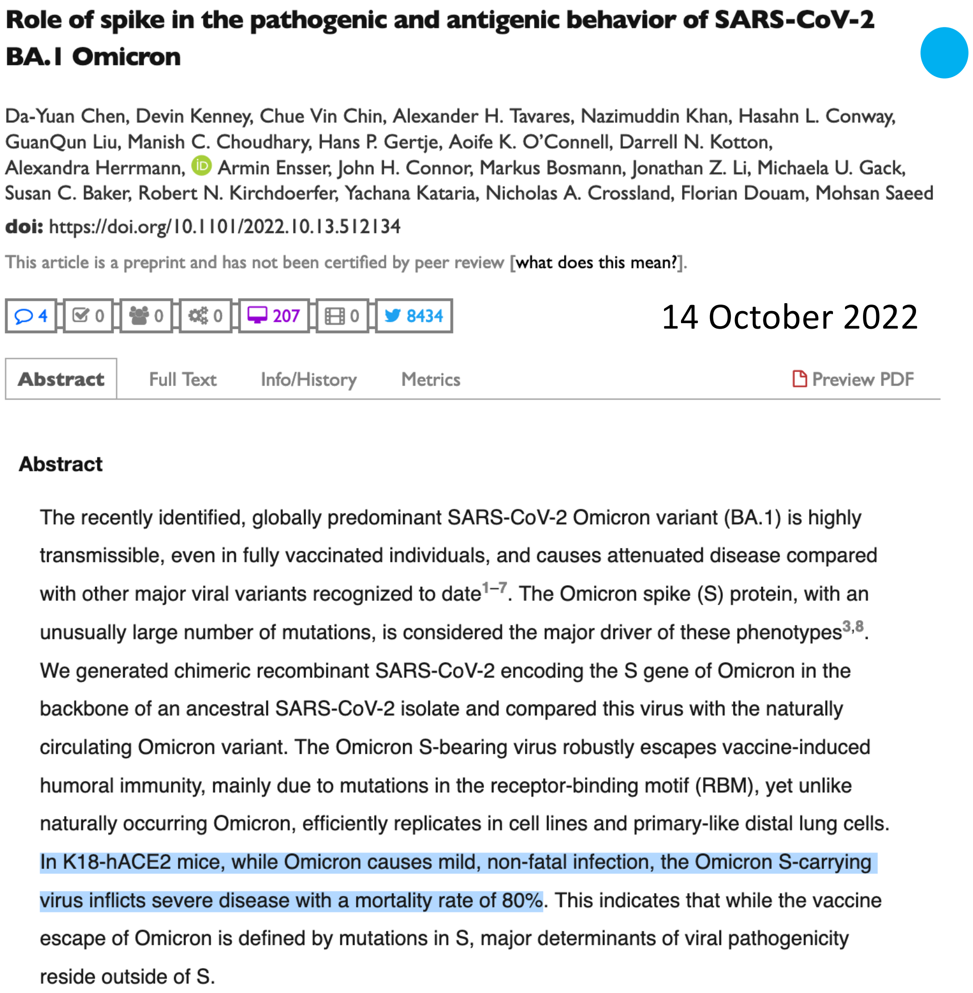
MailOnline

17 October 2022

Boston University CREATES a new Covid strain that has an 80% kill rate — echoing dangerous experiments feared to have started pandemic

- **Researchers added Omicron's spike protein to the original Covid strain**
- **The protein makes it highly infectious, meaning the new virus is doubly deadly**
- **80% per cent of mice died from the lab-created strain at Boston University**
- **The research has been called 'dangerously reckless' and 'very dumb'**

Dr Richard Ebright (left), a Rutgers University chemist, said that this research could spark the next lab-created pandemic. Professor Shmuel Shapira (right), a leading scientist in the Israeli Government, said that this type of research should be banned as it is playing with fire.



Role of spike in the pathogenic and antigenic behavior of SARS-CoV-2 BA.1 Omicron

Da-Yuan Chen, Devin Kenney, Chue Vin Chin, Alexander H. Tavares, Nazimuddin Khan, Hasahn L. Conway, GuanQun Liu, Manish C. Choudhary, Hans P. Gertje, Aoife K. O'Connell, Darrell N. Kotton, Alexandra Herrmann, Armin Ensser, John H. Connor, Markus Bosmann, Jonathan Z. Li, Michaela U. Gack, Susan C. Baker, Robert N. Kirchdoerfer, Yachana Kataria, Nicholas A. Crossland, Florian Douam, Mohsan Saeed

doi: https://doi.org/10.1101/2022.10.13.512134

This article is a preprint and has not been certified by peer review [what does this mean?].

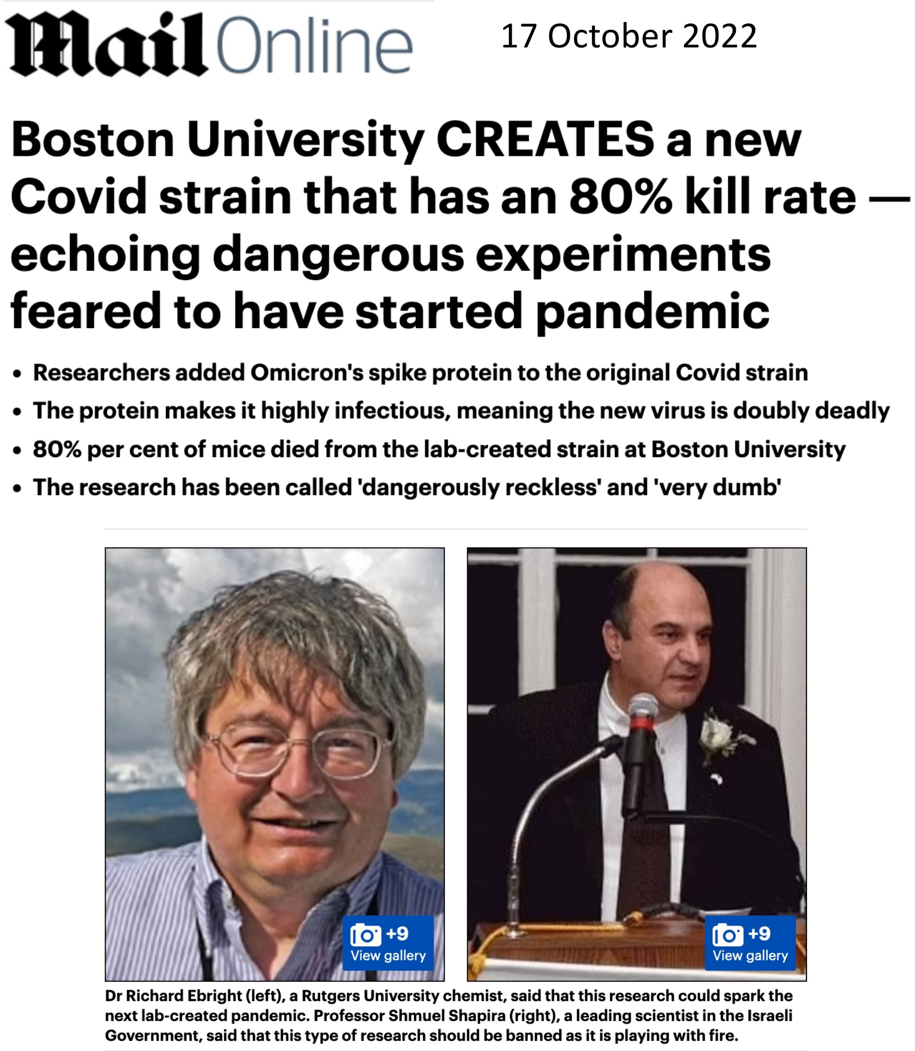
14 October 2022

Abstract

The recently identified, globally predominant SARS-CoV-2 Omicron variant (BA.1) is highly transmissible, even in fully vaccinated individuals, and causes attenuated disease compared with other major viral variants recognized to date[1–7]. The Omicron spike (S) protein, with an unusually large number of mutations, is considered the major driver of these phenotypes[3,8]. We generated chimeric recombinant SARS-CoV-2 encoding the S gene of Omicron in the backbone of an ancestral SARS-CoV-2 isolate and compared this virus with the naturally circulating Omicron variant. The Omicron S-bearing virus robustly escapes vaccine-induced humoral immunity, mainly due to mutations in the receptor-binding motif (RBM), yet unlike naturally occurring Omicron, efficiently replicates in cell lines and primary-like distal lung cells. In K18-hACE2 mice, while Omicron causes mild, non-fatal infection, the Omicron S-carrying virus inflicts severe disease with a mortality rate of 80%. This indicates that while the vaccine escape of Omicron is defined by mutations in S, major determinants of viral pathogenicity reside outside of S.

source: https://www.dailymail.co.uk/health/article-11323677/Outrage-Boston-University-CREATES-Covid-strain-80-kill-rate.html

17 October 2022

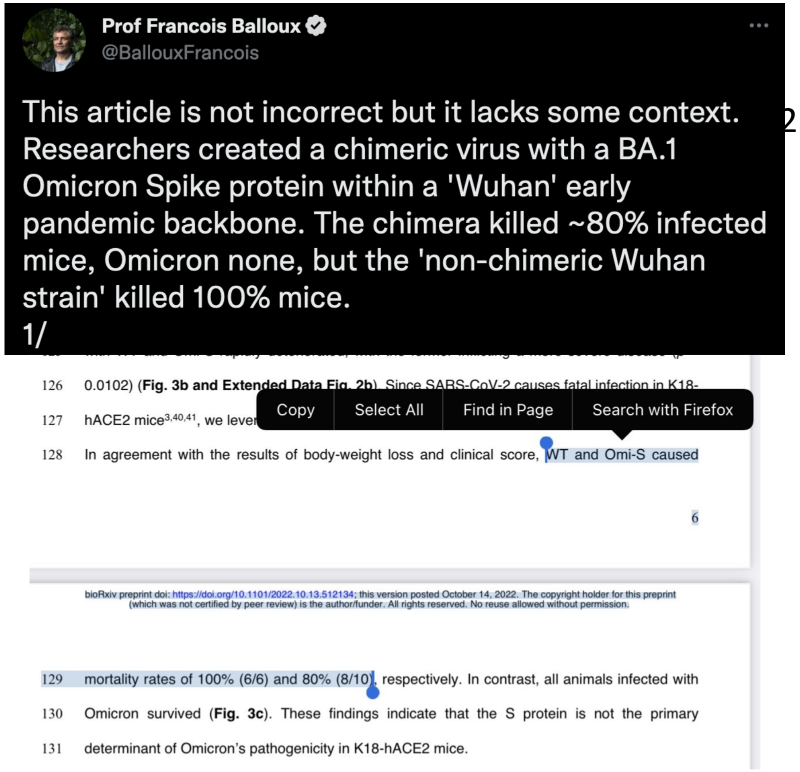


Boston University CREATES a new Covid strain that has an 80% kill rate — echoing dangerous experiments feared to have started pandemic

- **Researchers added Omicron's spike protein to the original Covid strain**
- **The protein makes it highly infectious, meaning the new virus is doubly deadly**
- **80% per cent of mice died from the lab-created strain at Boston University**
- **The research has been called 'dangerously reckless' and 'very dumb'**

Dr Richard Ebright (left), a Rutgers University chemist, said that this research could spark the next lab-created pandemic. Professor Shmuel Shapira (right), a leading scientist in the Israeli Government, said that this type of research should be banned as it is playing with fire.

In mice!



Prof Francois Balloux
@BallouxFrancois

This article is not incorrect but it lacks some context. Researchers created a chimeric virus with a BA.1 Omicron Spike protein within a 'Wuhan' early pandemic backbone. The chimera killed ~80% infected mice, Omicron none, but the 'non-chimeric Wuhan strain' killed 100% mice.
1/

0.0102) (**Fig. 3b and Extended Data Fig. 2b**). Since SARS-CoV-2 causes fatal infection in K18-hACE2 mice[3,40,41], we lever

In agreement with the results of body-weight loss and clinical score, WT and Omi-S caused

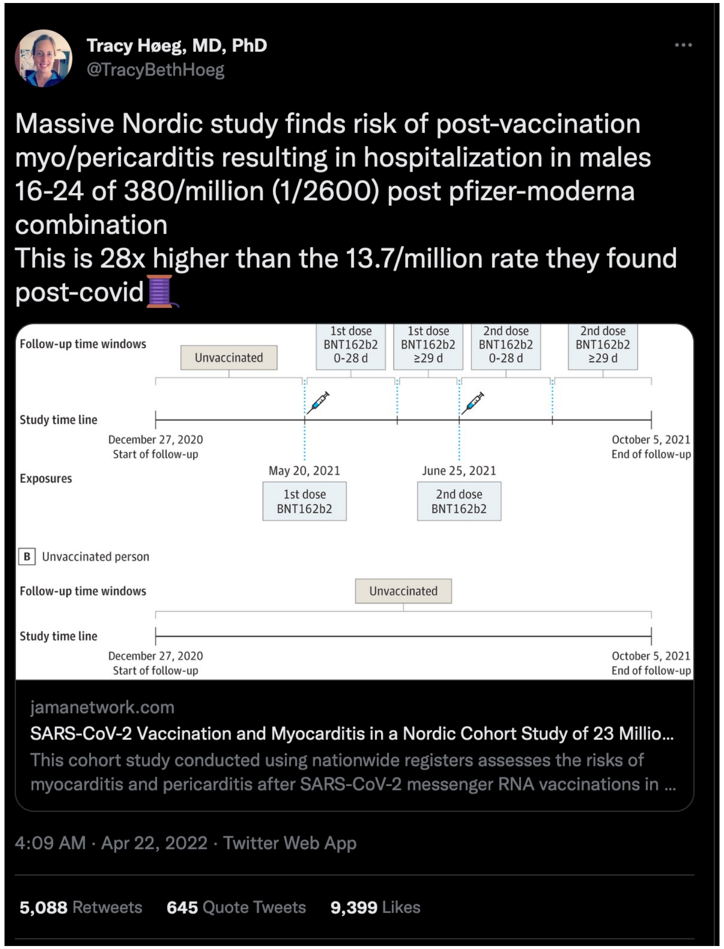
6

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.13.512134; this version posted October 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mortality rates of 100% (6/6) and 80% (8/10), respectively. In contrast, all animals infected with Omicron survived (**Fig. 3c**). These findings indicate that the S protein is not the primary determinant of Omicron's pathogenicity in K18-hACE2 mice.

more at: https://twitter.com/justsaysinmice

source: https://www.dailymail.co.uk/health/article-11323677/Outrage-Boston-University-CREATES-Covid-strain-80-kill-rate.html



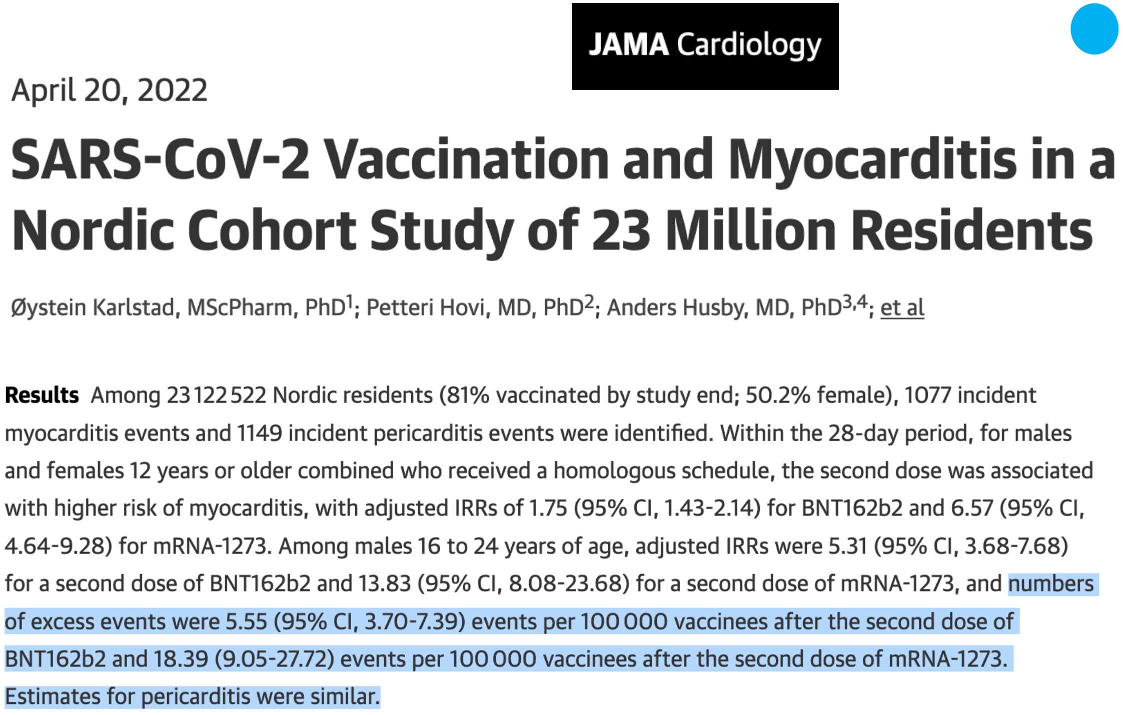
Tracy Høeg, MD, PhD
@TracyBethHoeg

Massive Nordic study finds risk of post-vaccination myo/pericarditis resulting in hospitalization in males 16-24 of 380/million (1/2600) post pfizer-moderna combination
This is 28x higher than the 13.7/million rate they found post-covid🧵

jamanetwork.com
SARS-CoV-2 Vaccination and Myocarditis in a Nordic Cohort Study of 23 Millio...
This cohort study conducted using nationwide registers assesses the risks of myocarditis and pericarditis after SARS-CoV-2 messenger RNA vaccinations in ...

4:09 AM · Apr 22, 2022 · Twitter Web App

5,088 Retweets **645** Quote Tweets **9,399** Likes



JAMA Cardiology

April 20, 2022

SARS-CoV-2 Vaccination and Myocarditis in a Nordic Cohort Study of 23 Million Residents

Øystein Karlstad, MScPharm, PhD[1]; Petteri Hovi, MD, PhD[2]; Anders Husby, MD, PhD[3,4]; et al

Results Among 23 122 522 Nordic residents (81% vaccinated by study end; 50.2% female), 1077 incident myocarditis events and 1149 incident pericarditis events were identified. Within the 28-day period, for males and females 12 years or older combined who received a homologous schedule, the second dose was associated with higher risk of myocarditis, with adjusted IRRs of 1.75 (95% CI, 1.43-2.14) for BNT162b2 and 6.57 (95% CI, 4.64-9.28) for mRNA-1273. Among males 16 to 24 years of age, adjusted IRRs were 5.31 (95% CI, 3.68-7.68) for a second dose of BNT162b2 and 13.83 (95% CI, 8.08-23.68) for a second dose of mRNA-1273, and numbers of excess events were 5.55 (95% CI, 3.70-7.39) events per 100 000 vaccinees after the second dose of BNT162b2 and 18.39 (9.05-27.72) events per 100 000 vaccinees after the second dose of mRNA-1273. Estimates for pericarditis were similar.

A multiplied risk means very little without the baseline risk as context:
Relative risk increase: 2800%
Absolute risk: 18.39/100000 = 0.018%

Myocarditis is inflammation of the heart muscle.
Pericarditis is inflammation of the outer lining of the heart.

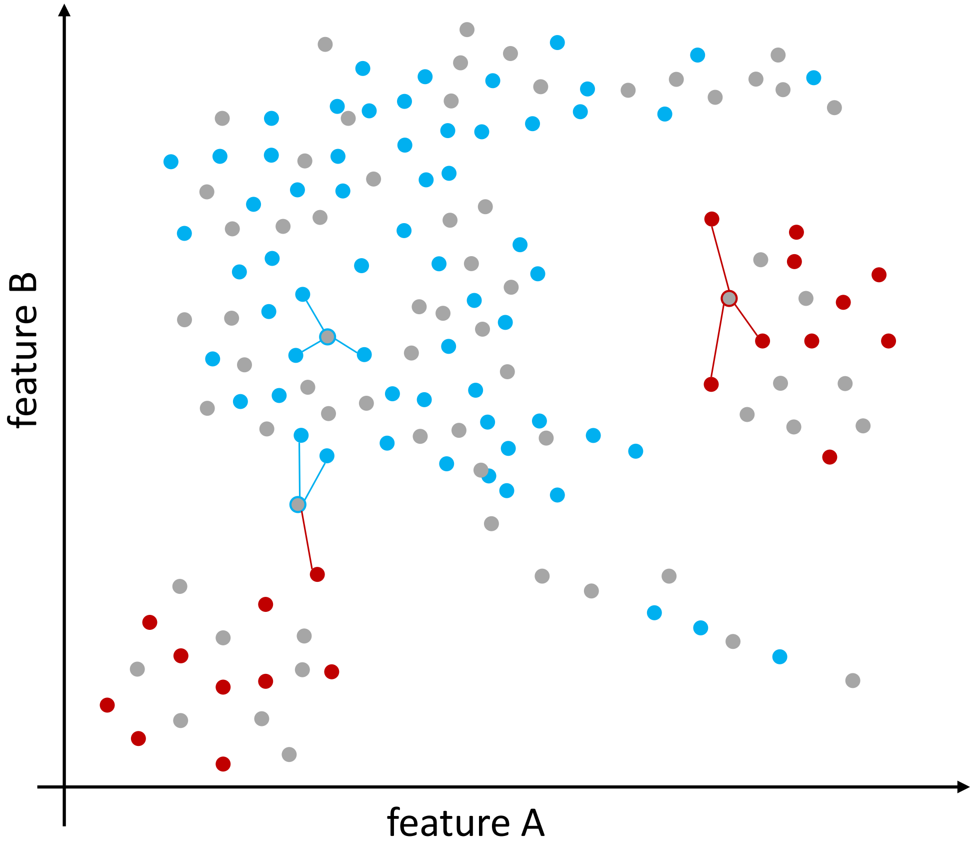
source: https://twitter.com/justsaysrisks

Outline of the course

1. Mean, Standard Deviation, Standard Error, Confidence Intervals, T-test
2. Fourier Transform, Wavelet Transforms, Spectrograms, High-pass, Low-pass filters
3. Covariance and Principal Component Analysis (PCA)
4. Clustering Methods
5. Pearson Correlation, PCA and SVD
6. Linear Regression / Logistic Regression
7. Non-linear Methods: k-NN, random forest, t-SNE, deep nets
8. Oral exam preparation / Invited lectures from the biomedical industry

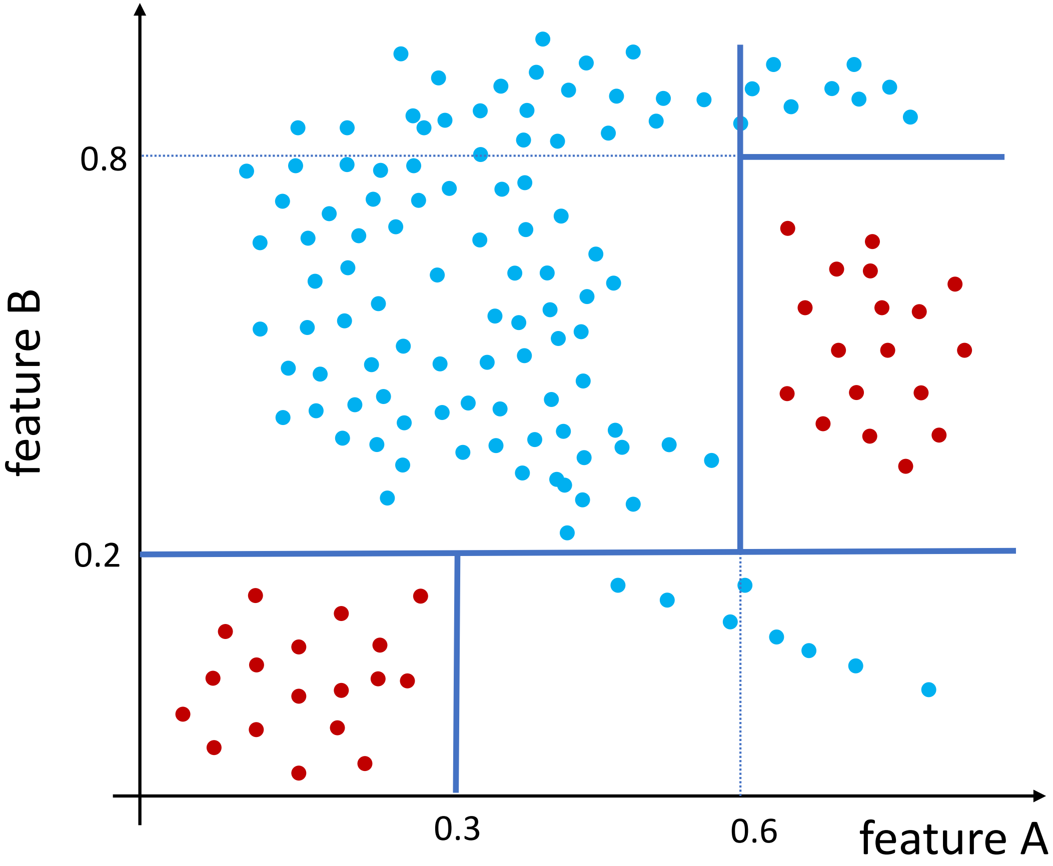
K-nearest neighbor algorithm: definition

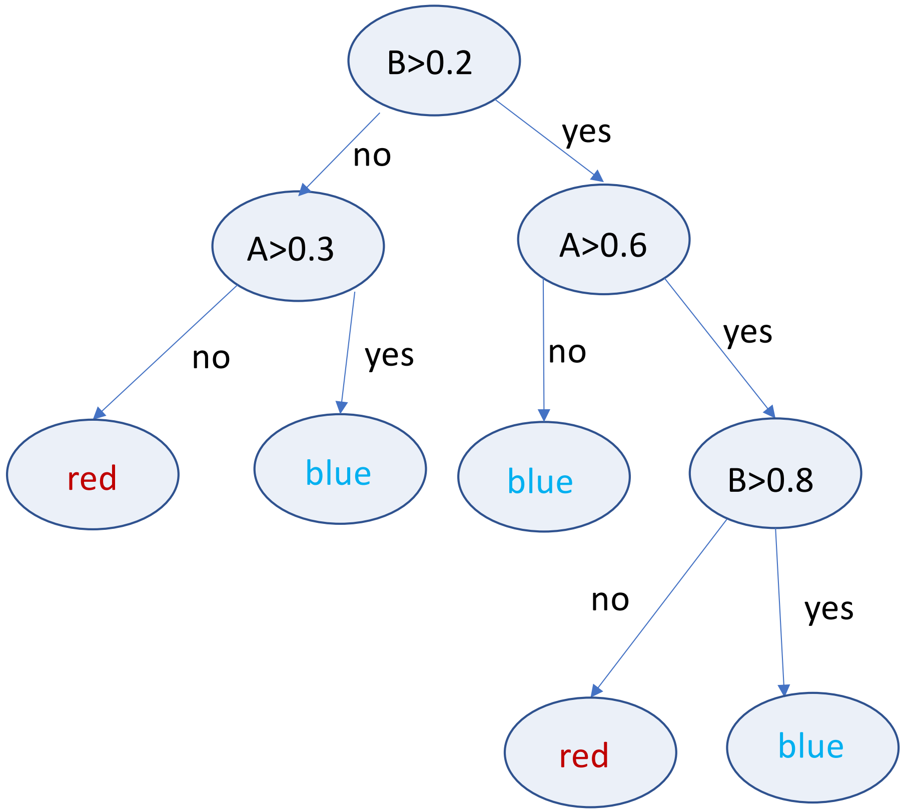
A data point is classified by a vote of its neighbors, with the point being assigned to the class most common among its ‘k’ nearest neighbors (‘k’ is a positive integer, typically small).



Decision tree: definition

A decision tree is built by splitting a dataset into subsets recursively. Each split operates on one feature of the dataset, and tries to maximize the separation of classes. The recursion is completed when the subset of points at any given node all have the same value.

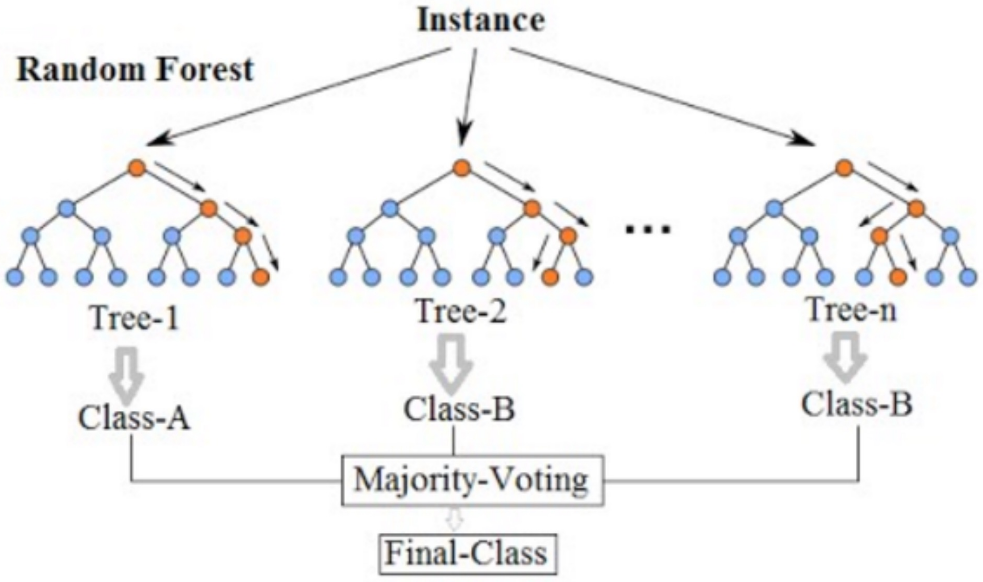




source: https://en.wikipedia.org/wiki/Decision_tree_learning

Random Forest: definition

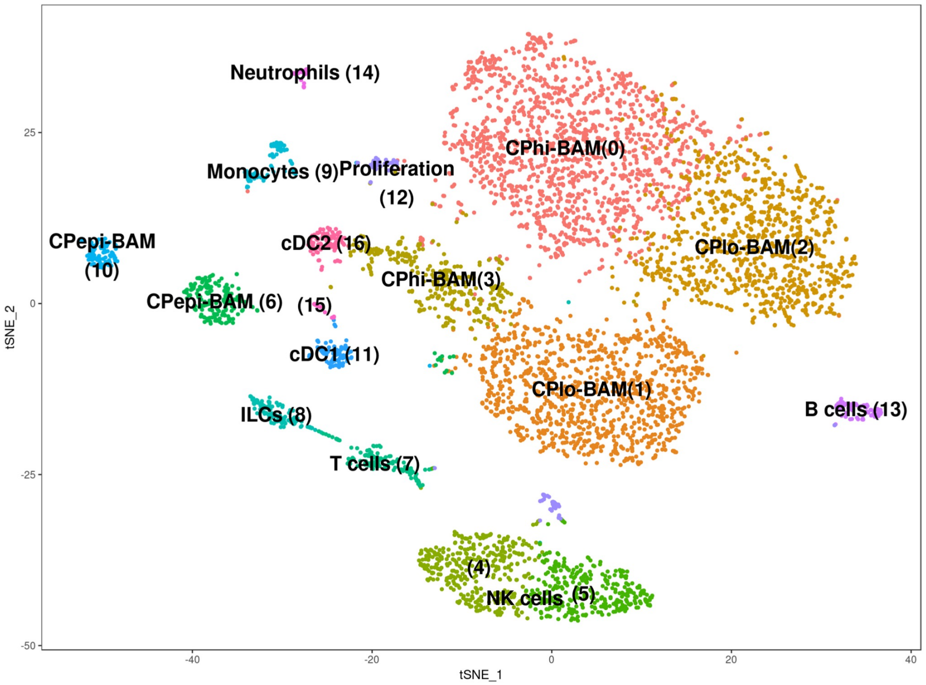
Decision trees are prone to overfitting. A random forest operates by constructing a multitude of decision trees at training time. The output of the random forest is the class selected by most trees. Random forests generally outperform decision trees. Random forest belongs to the class of "ensemble learning" algorithms.



source: https://en.wikipedia.org/wiki/Random_forest break

t-SNE data visualization: definition

t-distributed stochastic neighbor embedding (t-SNE) is a method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.

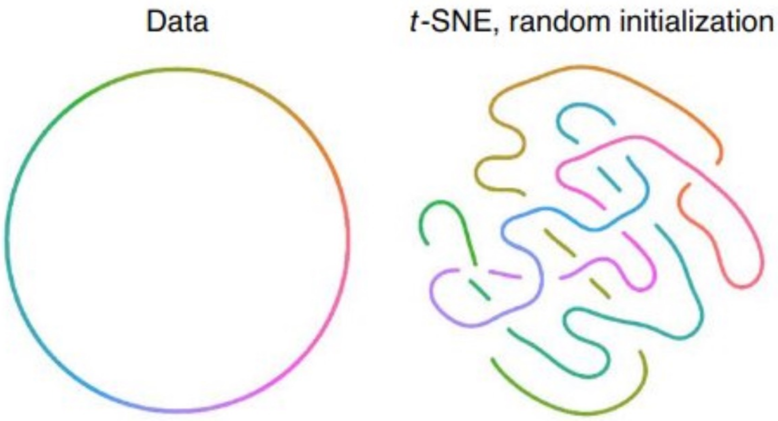


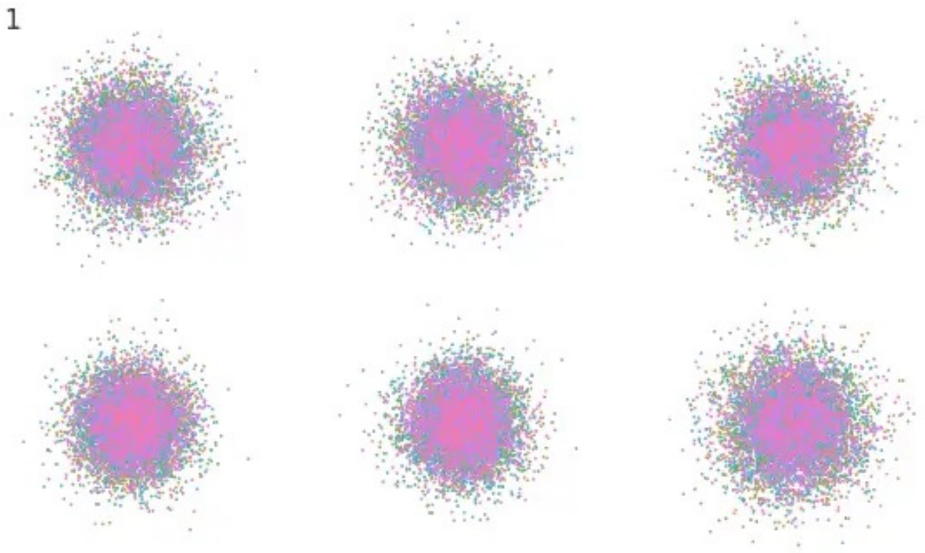
Single-cell RNA sequencing datasets of brain immune cells, projected into 2 dimensions using t-SNE

source: https://www.brainimmuneatlas.org/index.php

Over-interpretation risks with t-SNE

- Local structure is preserved by t-SNE embeddings, but global structure is not!
- t-SNE is a stochastic algorithm, it never converges twice to the same mapping
- t-SNE can be misleading as it can “hallucinate” structure that is not present in the data.





source: https://twitter.com/hippopedoid/status/1356977421287911434

Deep networks: definition

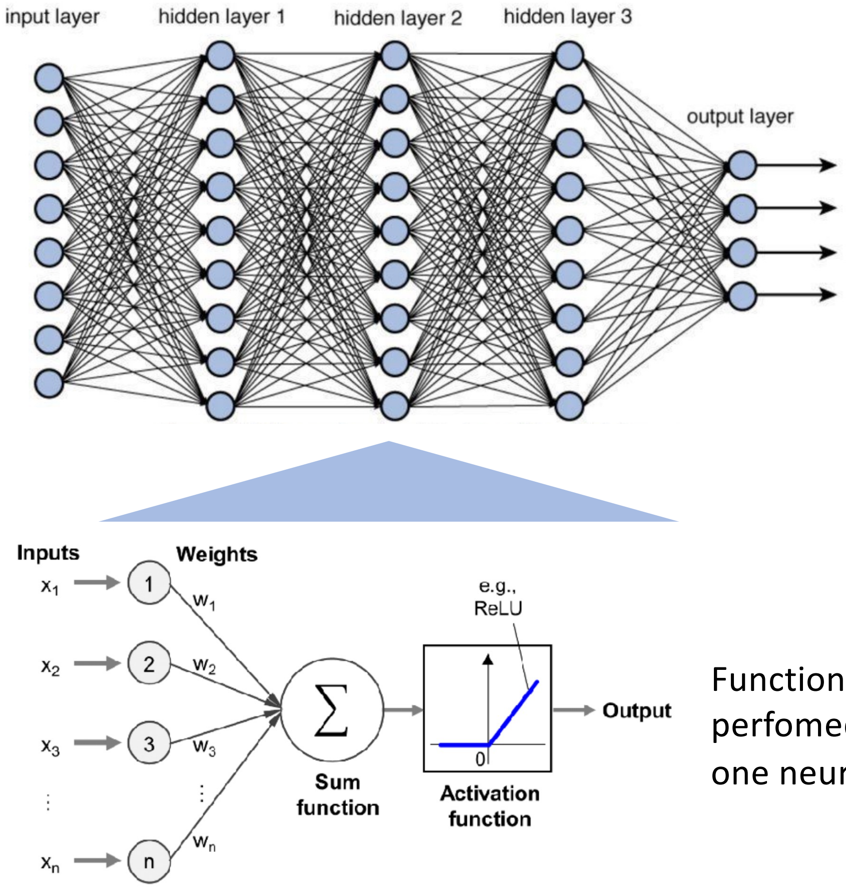
A deep neural network is an artificial neural network with multiple layers between the input and output layers. The network takes as an input a data sample ‘x’ and predicts as its output the data label ‘y’:

The network output is given by:

$$\widehat{y} = net(x)$$

where net() is composed of layers defined by:

$$a_i^l = ReLU\left[\sum_j w_{ij} a_j^{l-1}\right]$$



Function perfomed by one neuron

Deep networks: training with backpropagation

The network output is given by:

$$\widehat{y} = net(x)$$

where net() is composed of layers defined by:

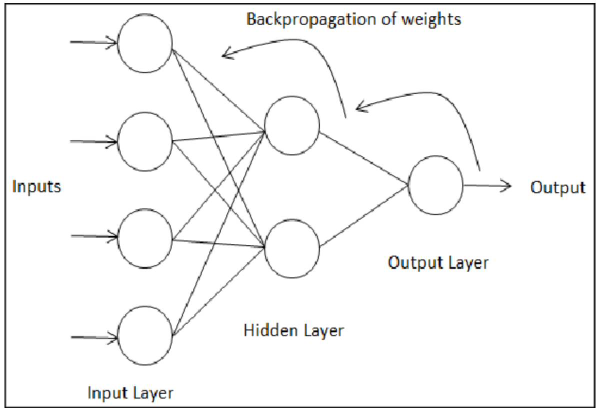
$$a_i^l = ReLU\left[\sum_j w_{ij} a_j^{l-1}\right]$$

We define the loss function ('y' is the true label):

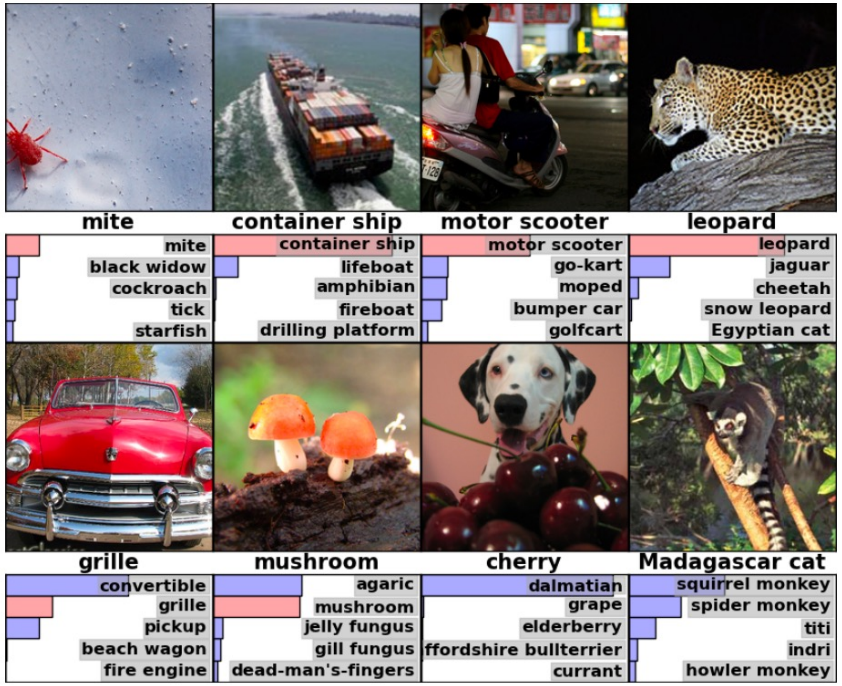
$$\mathcal{L} = ||\widehat{y} - y||_2$$

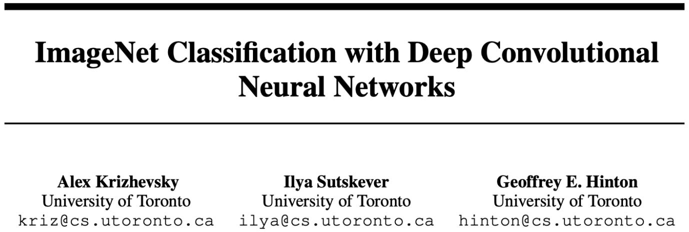
The backpropagation update rule is given by:

$$w_{ij}^{t+1} = w_{ij}^t - \epsilon * \frac{d\mathcal{L}}{dw_{ij}^t}$$



Success of deep networks in image recognition





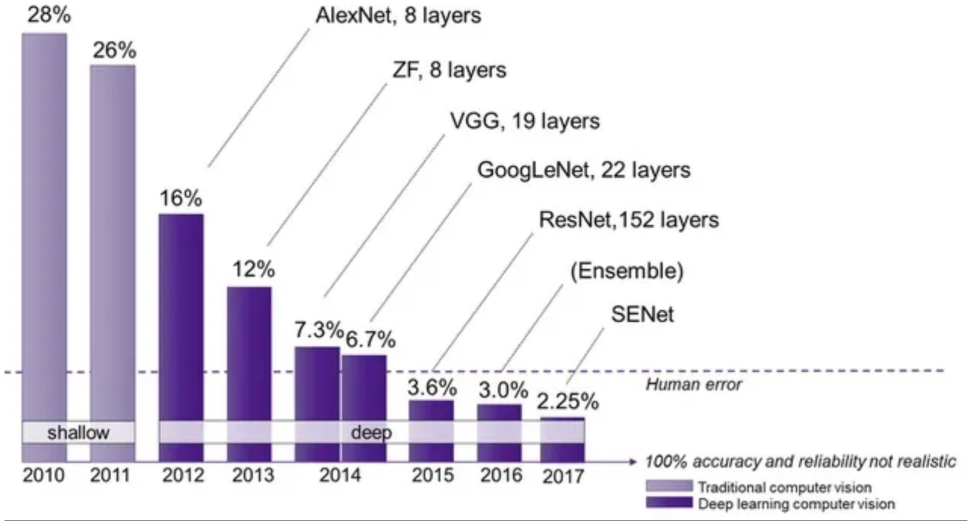
ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

2012

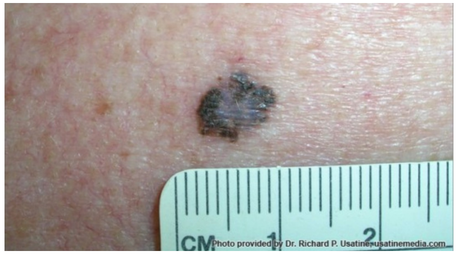


Caveats of using deep networks in medicine

- Deep networks need very large dataset to be trained.

- Deep networks do not necessarily generalize to datasets slightly different than those seen during training (e.g. x-ray scans taken from a different machine brand, with different lighting conditions, etc.)

- Sometimes deep networks can “cheat” and use unrelated factors to make a decision (e.g. a ruler in the image). They do not give an explanation for their decision, so they are difficult to trust.

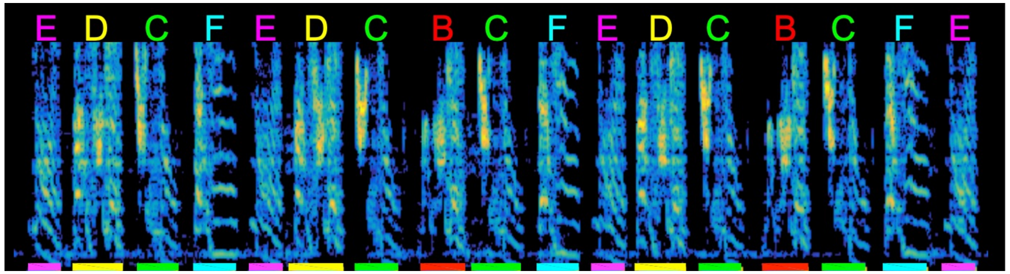


- A doctor’s clinical impression and diagnosis is based on contextual factors beyond direct image inspection.

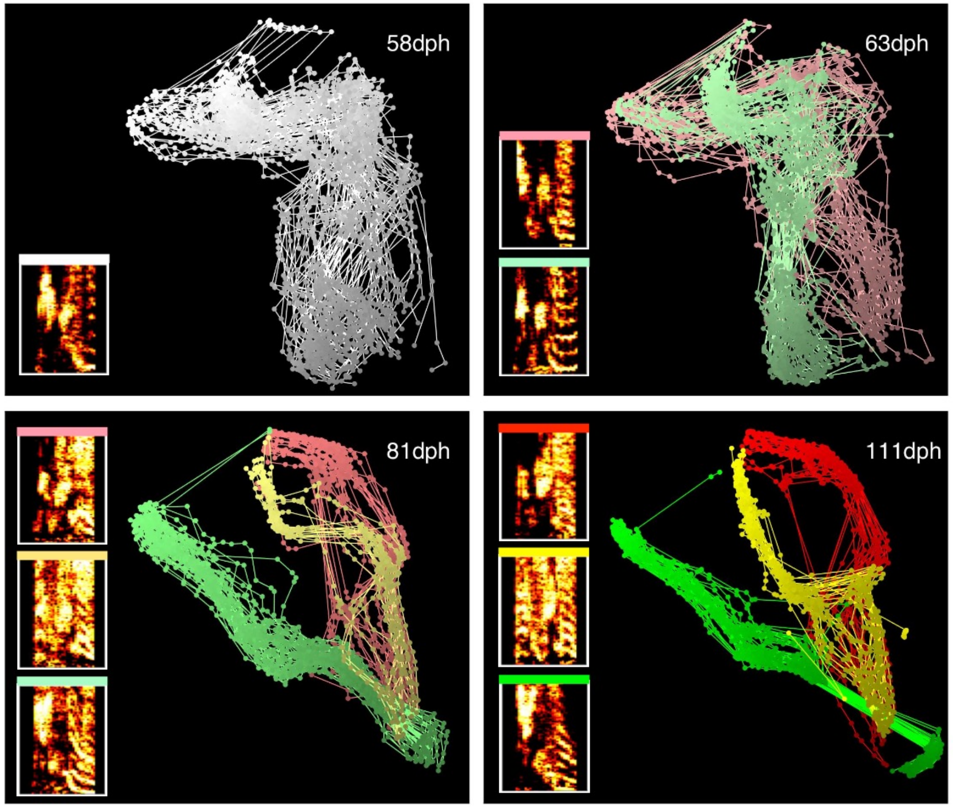
Case study 1: tracking song development in the song bird with t-SNE

Zebra finch

Spectrogram of zebra finch song



t-SNE vizualization of syllables through development. dph: days post hatch



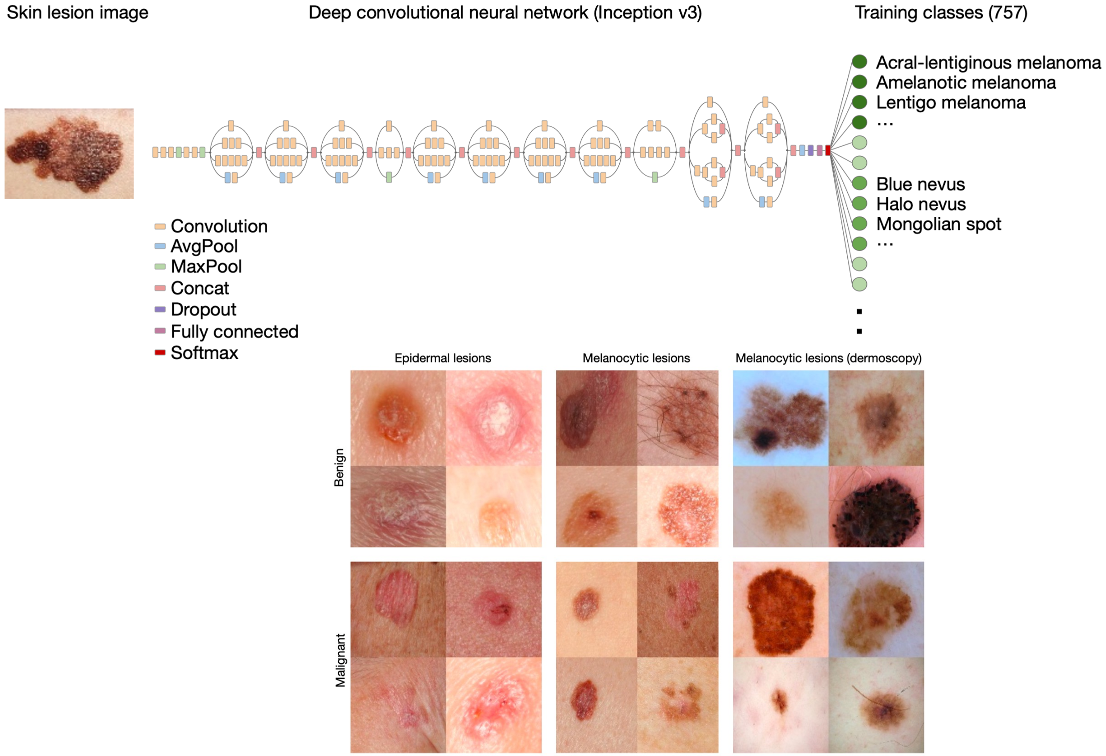
Deny & Mackevicus et al., ICLR 2016

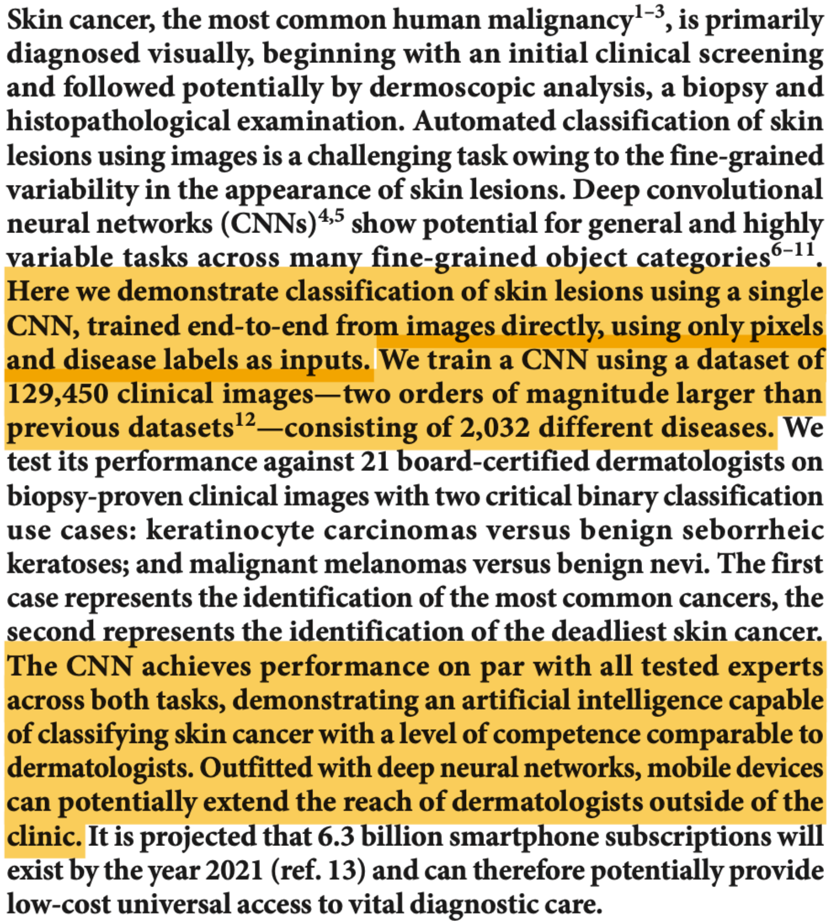
Case study 2: diagnosing skin cancers with deep nets

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Nature 2016





Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

Next lecture

Oct 31:

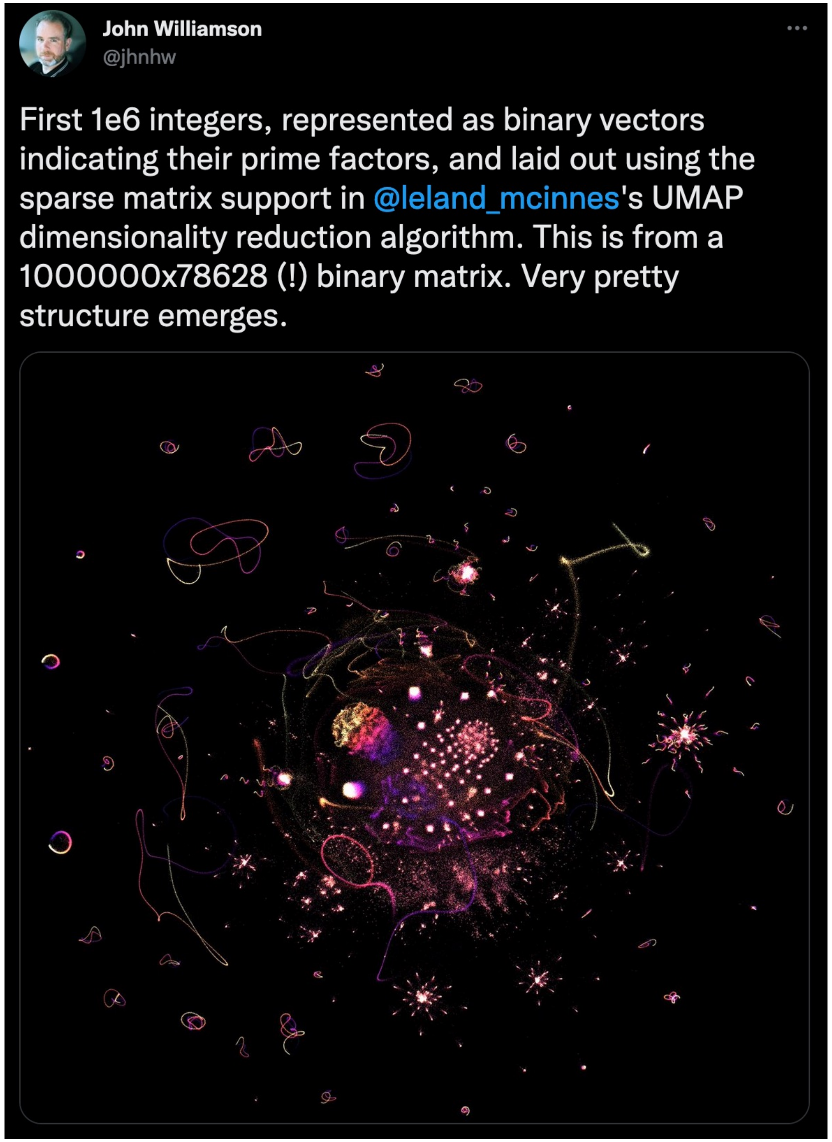
- Guest lecture from Dr. Karita Salo – Biostatistician at Nordic Healthcare Group
  Personalized medicine and drug development

- Preparation of the oral exam with an exercise from last year

Supplementary Material

How t-SNE can be misleading

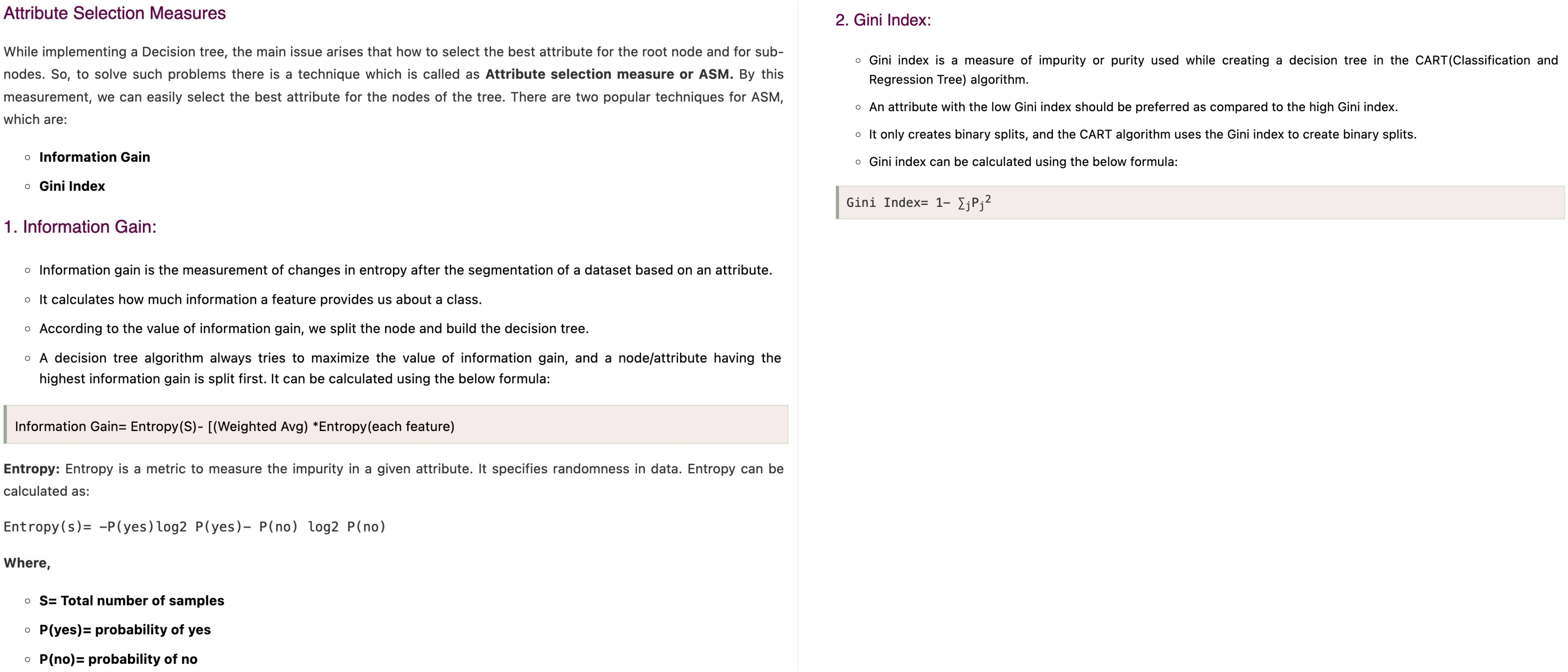
- Explanation here: https://twitter.com/hippopedoid/status/1318917878364672001



John Williamson
@jhnhw

First 1e6 integers, represented as binary vectors indicating their prime factors, and laid out using the sparse matrix support in @leland_mcinnes's UMAP dimensionality reduction algorithm. This is from a 1000000x78628 (!) binary matrix. Very pretty structure emerges.

Decision tree: typical criteria for the splits



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**
- **Gini Index**

1. Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

> Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= −P(yes)log2 P(yes)− P(no) log2 P(no)

Where,

- **S= Total number of samples**
- **P(yes)= probability of yes**
- **P(no)= probability of no**

2. Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

> Gini Index= $1- \sum_j P_j^2$

source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

Abstract

We present a new technique called "t-SNE" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales. This is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints. For visualizing the structure of very large data sets, we show how t-SNE can use random walks on neighborhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed. We illustrate the performance of t-SNE on a wide variety of data sets and compare it with many other non-parametric visualization techniques, including Sammon mapping, Isomap, and Locally Linear Embedding. The visualizations produced by t-SNE are significantly better than those produced by the other techniques on almost all of the data sets.

Keywords: visualization, dimensionality reduction, manifold learning, embedding algorithms, multidimensional scaling

For nearby datapoints, $p_{j|i}$ is relatively high, whereas for widely separated datapoints, $p_{j|i}$ will be almost infinitesimal (for reasonable values of the variance of the Gaussian, σ_i). Mathematically, the conditional probability $p_{j|i}$ is given by

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}, \tag{1}$$

where σ_i is the variance of the Gaussian that is centered on datapoint x_i. The method for determining the value of σ_i is presented later in this section. Because we are only interested in modeling pairwise similarities, we set the value of $p_{i|i}$ to zero. For the low-dimensional counterparts y_i and y_j of the high-dimensional datapoints x_i and x_j, it is possible to compute a similar conditional probability, which we denote by $q_{j|i}$. We set[2] the variance of the Gaussian that is employed in the computation of the conditional probabilities $q_{j|i}$ to $\frac{1}{\sqrt{2}}$. Hence, we model the similarity of map point y_j to map point y_i by

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

Again, since we are only interested in modeling pairwise similarities, we set $q_{i|i} = 0$.

If the map points y_i and y_j correctly model the similarity between the high-dimensional datapoints x_i and x_j, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. Motivated by this observation, SNE aims to find a low-dimensional data representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$. A natural measure of the faithfulness with which $q_{j|i}$ models $p_{j|i}$ is the Kullback-Leibler divergence (which is in this case equal to the cross-entropy up to an additive constant). SNE minimizes the sum of Kullback-Leibler divergences over all datapoints using a gradient descent method. The cost function C is given by

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{2}$$

in which P_i represents the conditional probability distribution over all other datapoints given datapoint x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i. Because the Kullback-Leibler divergence is not symmetric, different types of error in the pairwise distances in the low-dimensional map are not weighted equally. In particular, there is a large cost for using widely separated map points to represent nearby datapoints (i.e., for using

Blog posts

- https://medium.com/swlh/random-forest-ac5227dabb08
- https://serokell.io/blog/random-forest-classification
- https://towardsdatascience.com/random-forest-regression-for-continuous-well-log-prediction-61d3ec1c683a
- https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
- https://medium.com/analytics-vidhya/https-medium-com-shashi-kiran-ai-decision-tree-intuition-92f708f13f33