BCS1520: Statistics

Lecture 03:
Machine Learning

Maastricht University | Department of Advanced Computing Sciences

Be a robot.

- Write instructions for a robot to pick up trash and put it in the garbage.

Prediction and Classification

- Statistical models are good for interpretation: we can use them to make causal claims and so to essentially do *science*
- We may alternatively have more practical desires, we want to make *decisions* (or just write our emails for us)
- For this, we want to either predict something about the outcome or classify to understand the outcome variable
- For example: Large language models like ChatGPT predict words.

Prediction and Classification

Definition

Prediction: Using the independent variables in some algorithm to provide an estimate of what the dependent variable will look like for new values of the independent variable – more relevant for continuous variables.

Definition

Classification: Using independent variables in some algorithm to tell us the category that is appropriate given the values of the independent variables.

Prediction and Classification

- Prediction may be useful for weather or the stock price or the next words of a story, or *translating people's brain activity into speech*
- Classification is useful for understanding a customer’s likes and dislikes, or for what an object is that is seen by a camera, or (terrifyingly) who a person is.
- Classification and prediction are two sides of the same coin

Speech Neuroprosthesis

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

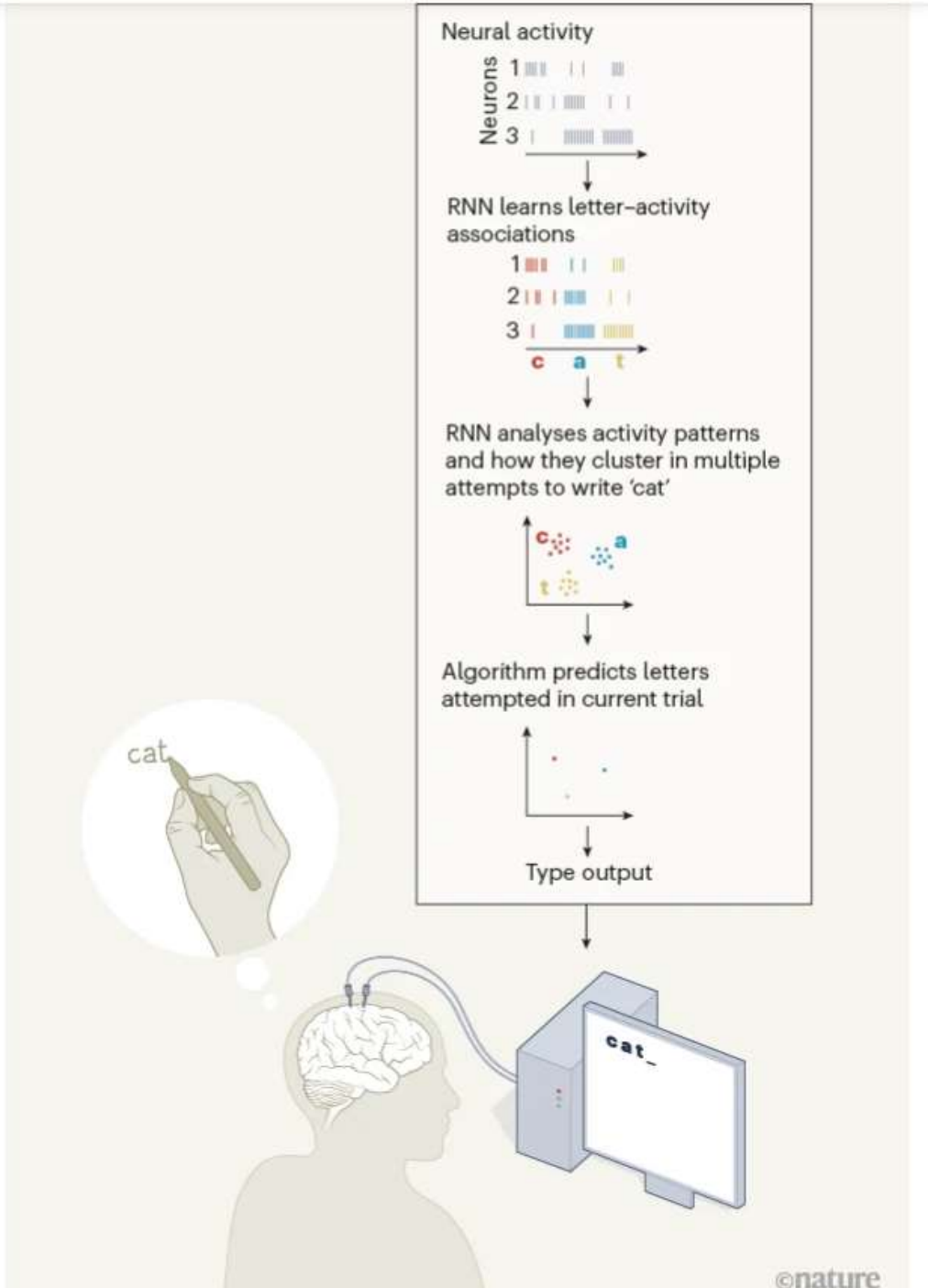
Article | Published: 23 August 2023

A high-performance neuroprosthesis for speech decoding and avatar control

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, R
Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn T
Karunesh Ganguly, Gopala K. Anumanchipalli & Edward F. Chang ✉

Nature **620**, 1037–1046 (2023) | Cite this article

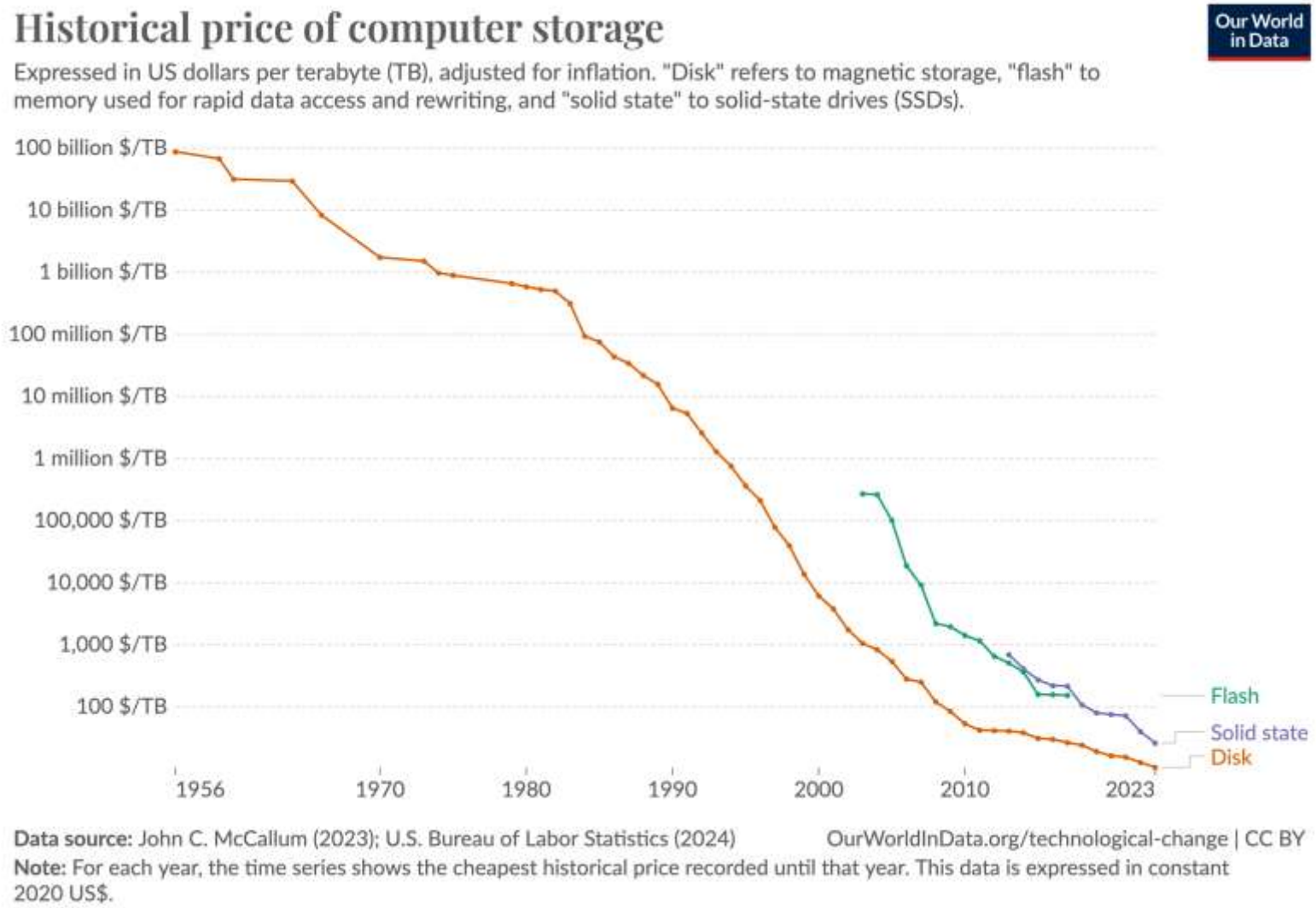
41k Accesses | **85** Citations | **3630** Altmetric | Metrics



What is your *n*?

- In data science, number of samples or number of data points is usually represented by *n*
- And then the number of parameters is represented by *p*
- How many parameters for the bell curve? $\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$

Big Data: *large n and large p*



Predicting who survived the Titanic

- What are the chances that an individual person was likely to survive the Titanic?
- Francis William Somerton left Ilfracombe in the UK to make his fortune in the USA on the Titanic
- He never made it.
- Can we predict how unlucky Francis was to not survive?

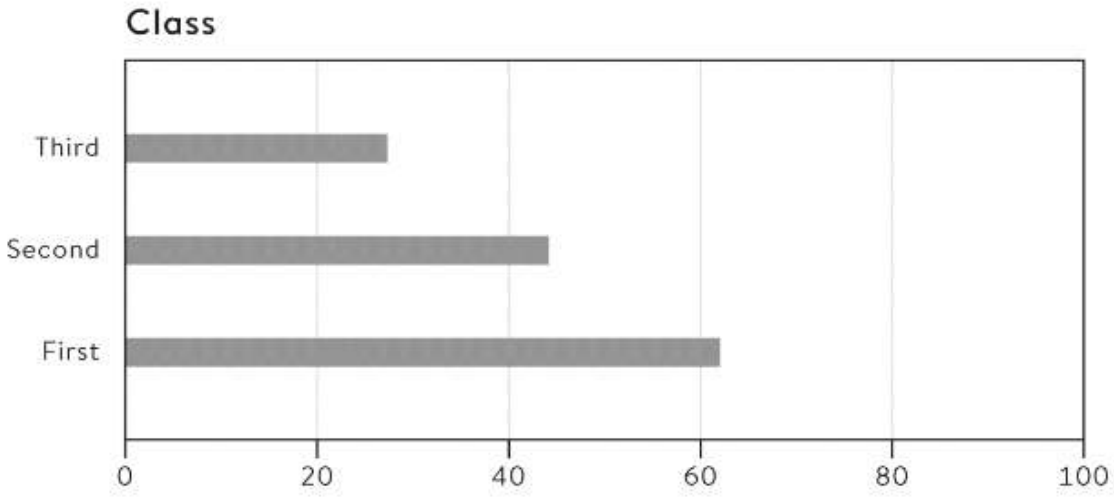
An aside on Kaggle

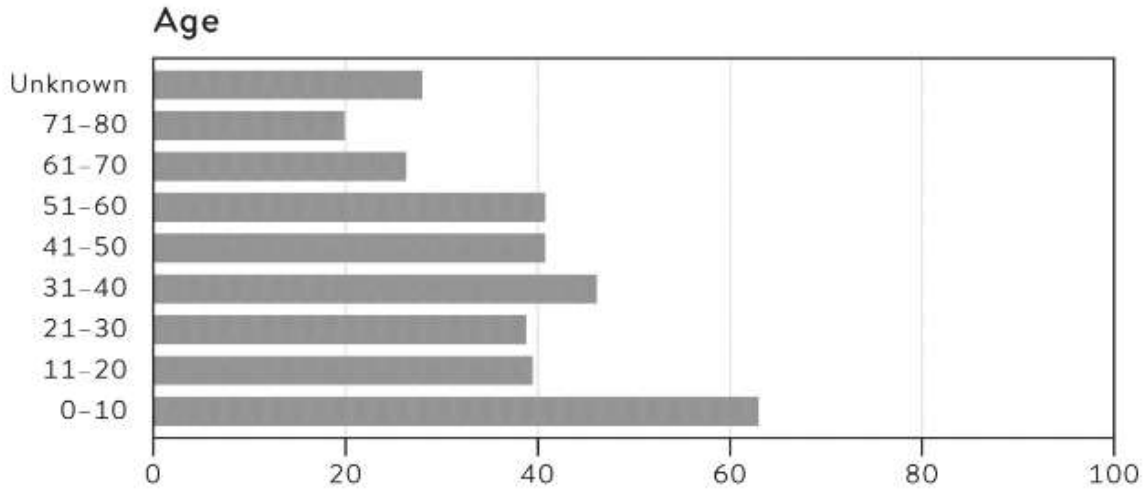
- Website with competitions on who can best build an algorithm that can predict or classify some data
- A company or academic institute or hospital puts up a dataset and asks a specific question
- People build different algorithms to win a prize
- Some data is hidden and is used to finally test the algorithm

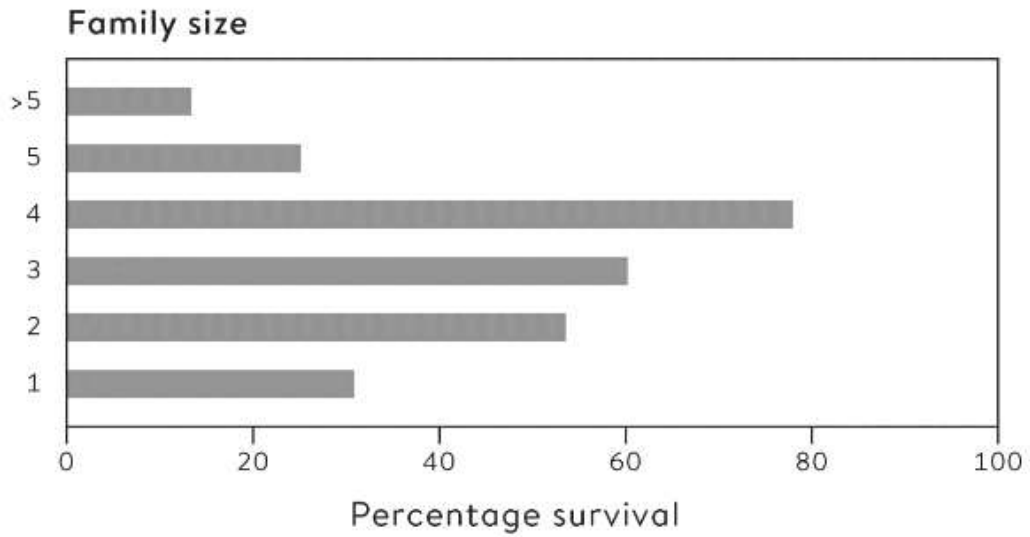
Predicting who survived the Titanic

- Plan to get all the data we have on the passengers and try different algorithms
- We know the name, title, gender, age, class of travel (1^{st}, 2^{nd} ,3^{rd}) fare paid, if they were part of a family and where they boarded of 1309 passengers
- The response variable indicates survival.

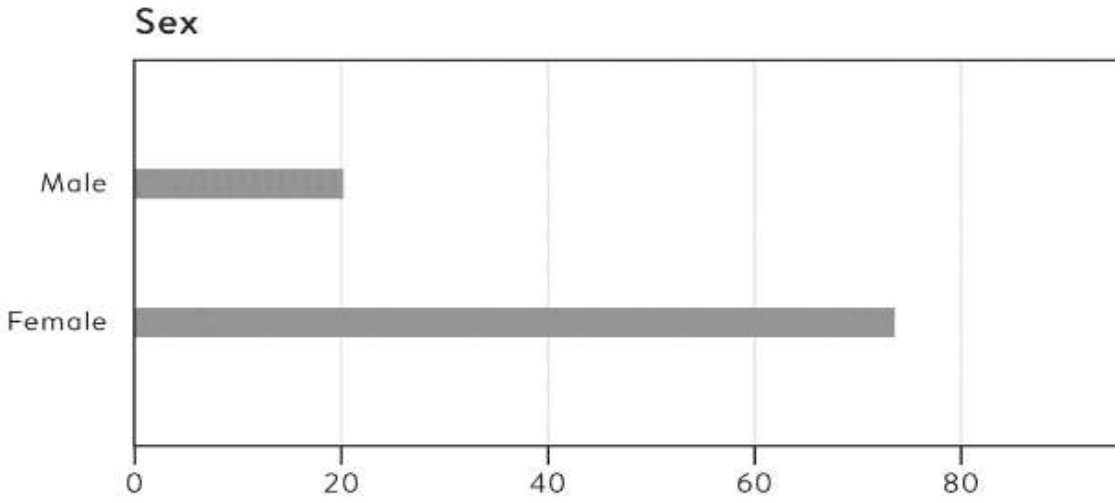
Exploratory Data Analysis

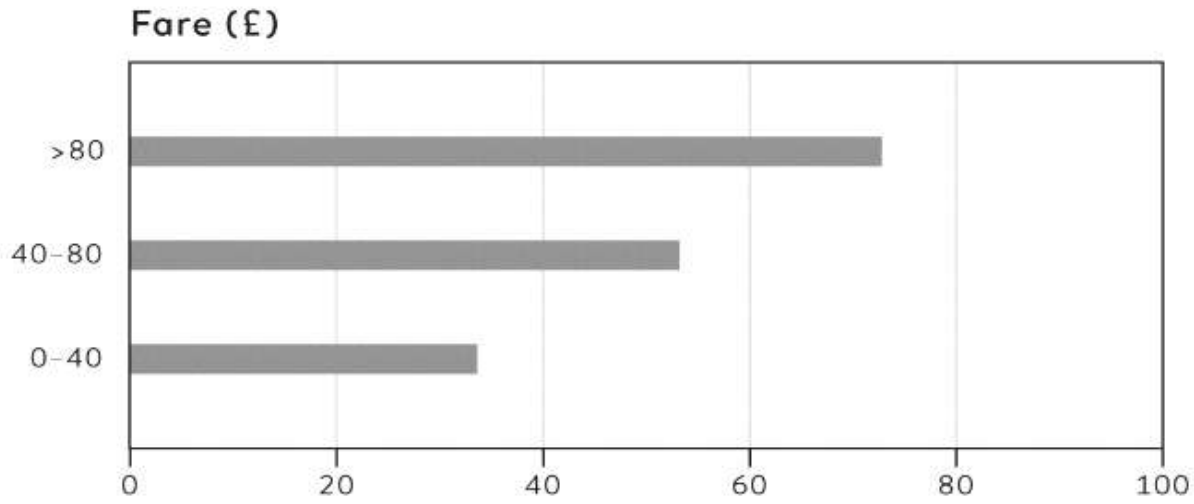


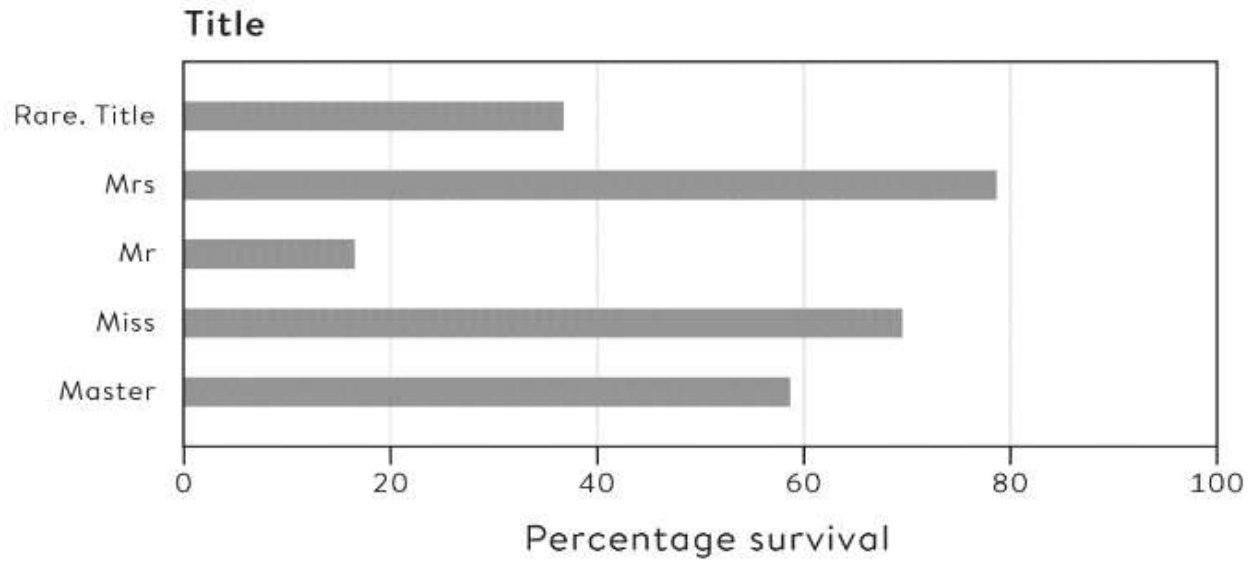




Exploratory Data Analysis







Predicting who survived the Titanic

- To make sure we don't fool ourselves and to enable generalizability (for the first ship to Mars?), we want to separate the data into training and testing data
- This idea ensures that we don't end up biased by the data we are looking at, and have a more realistic estimate of the accuracy of our algorithm
- Let's split data into 2/3rd training and 1/3rd test: this is arbitrary, but there can be more principled ways to do this depending on the features of the data.

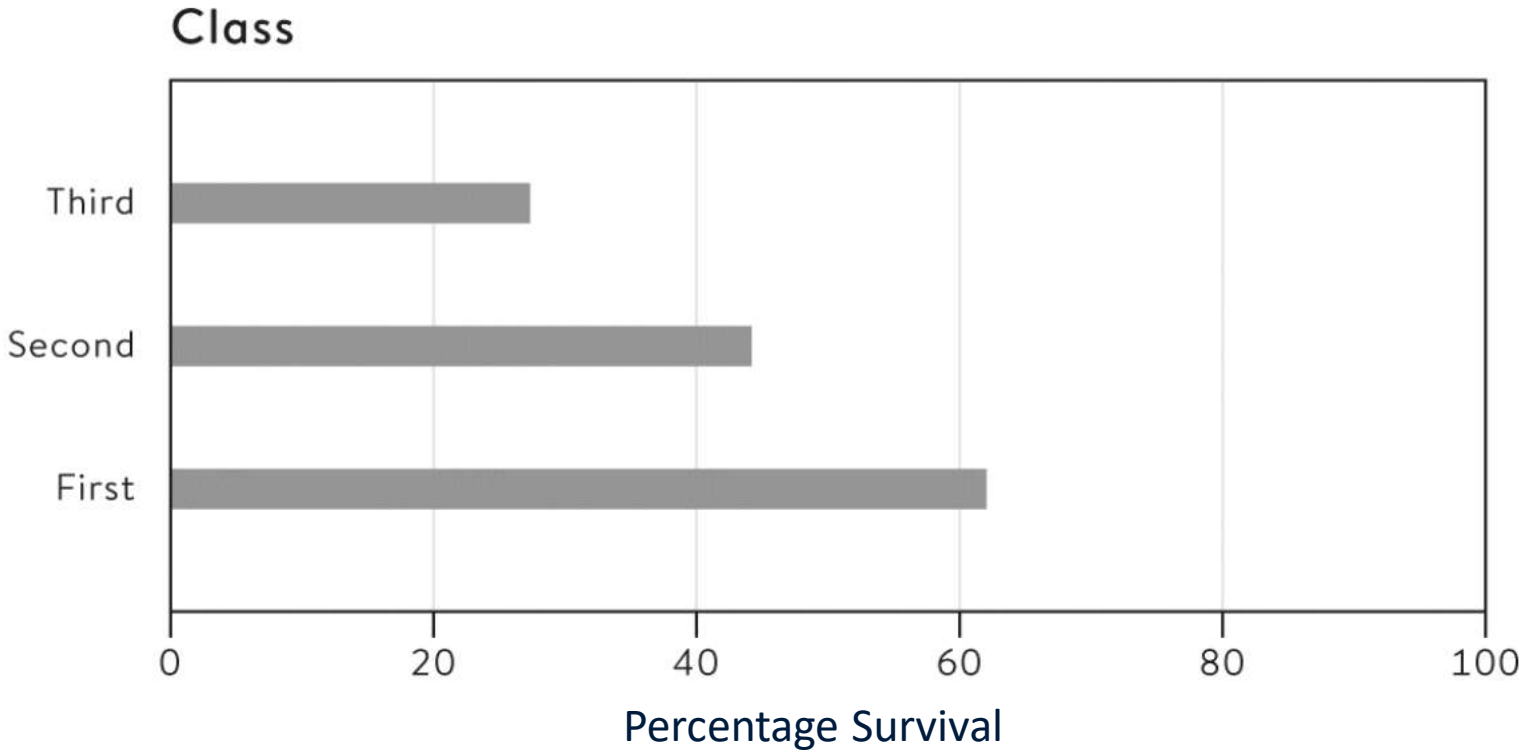
Pre-processing data

- Some people have missing information: assume they paid the median fare
- Combine number of siblings and parents to give a family size variable
- Simplify titles and if its weird, name it a “rare title”

- This is arbitrary, and alternative ways may provide better data for the algorithm we build

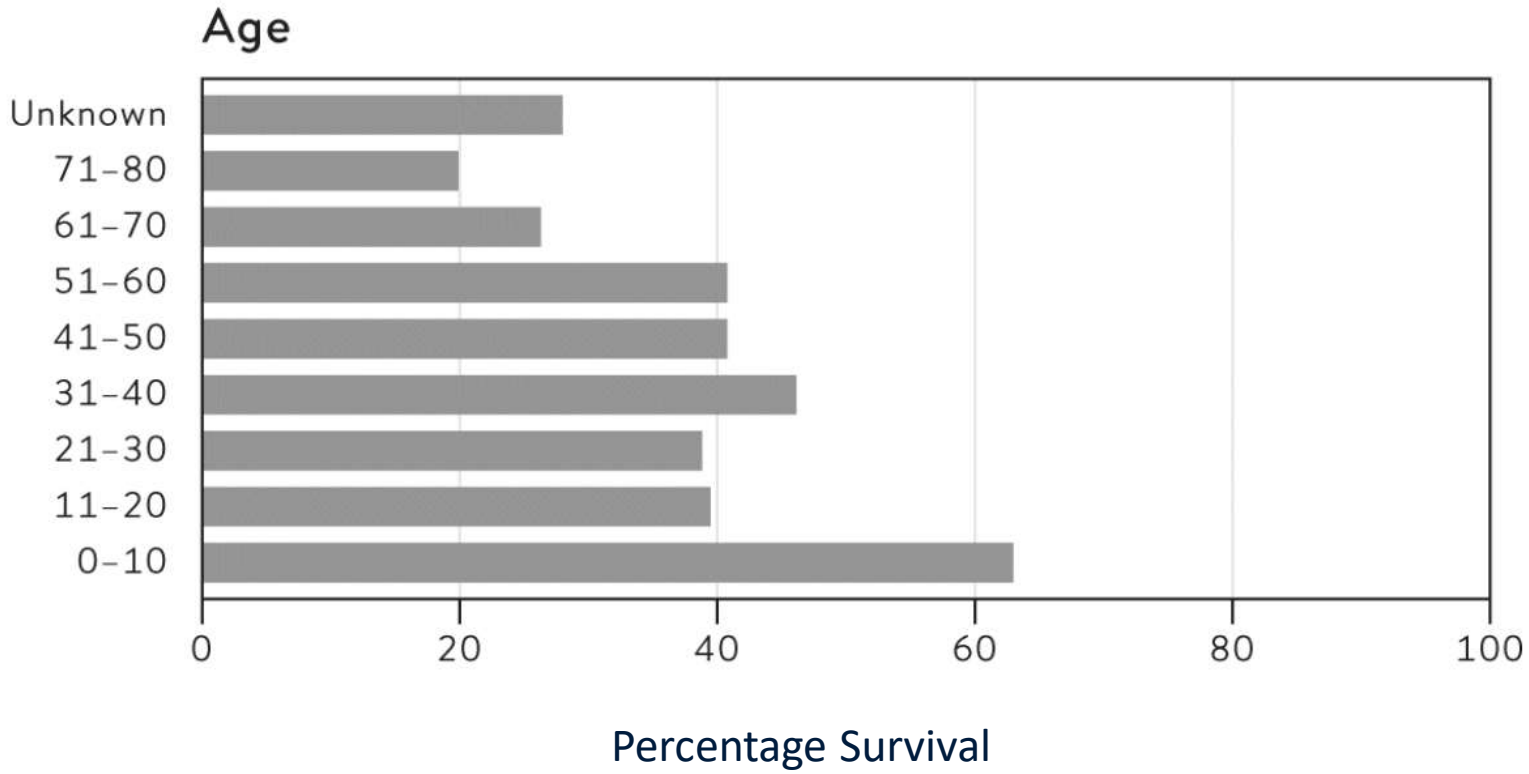
Predicting who survived the Titanic

- Decision stump on class:



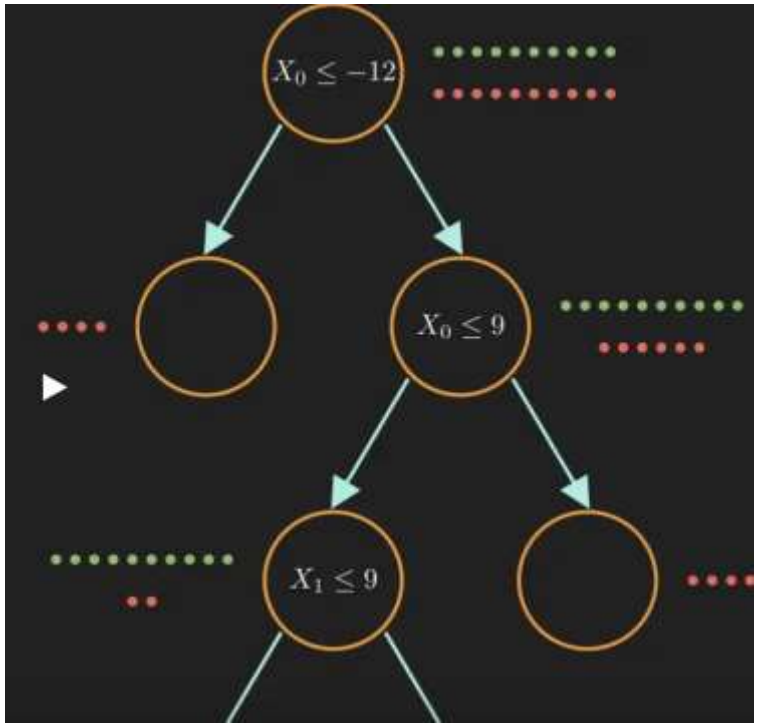
Predicting who survived the Titanic

- Decision stump on age:



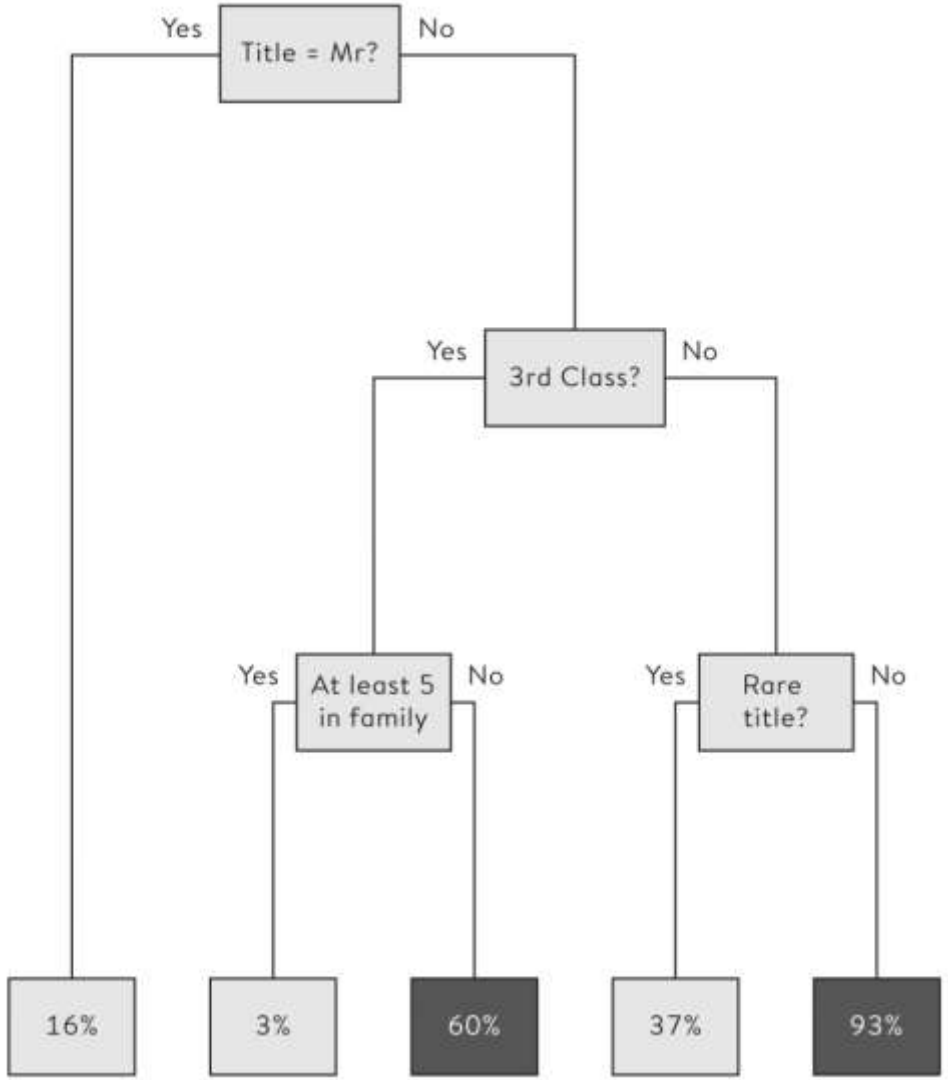
Predicting who survived the Titanic

- What could be a way we use this data to make a prediction?



Classification or decision tree

- One possible option
- But we could do this in multiple ways, as you know
- First. Let's think about ways to assess accuracy.

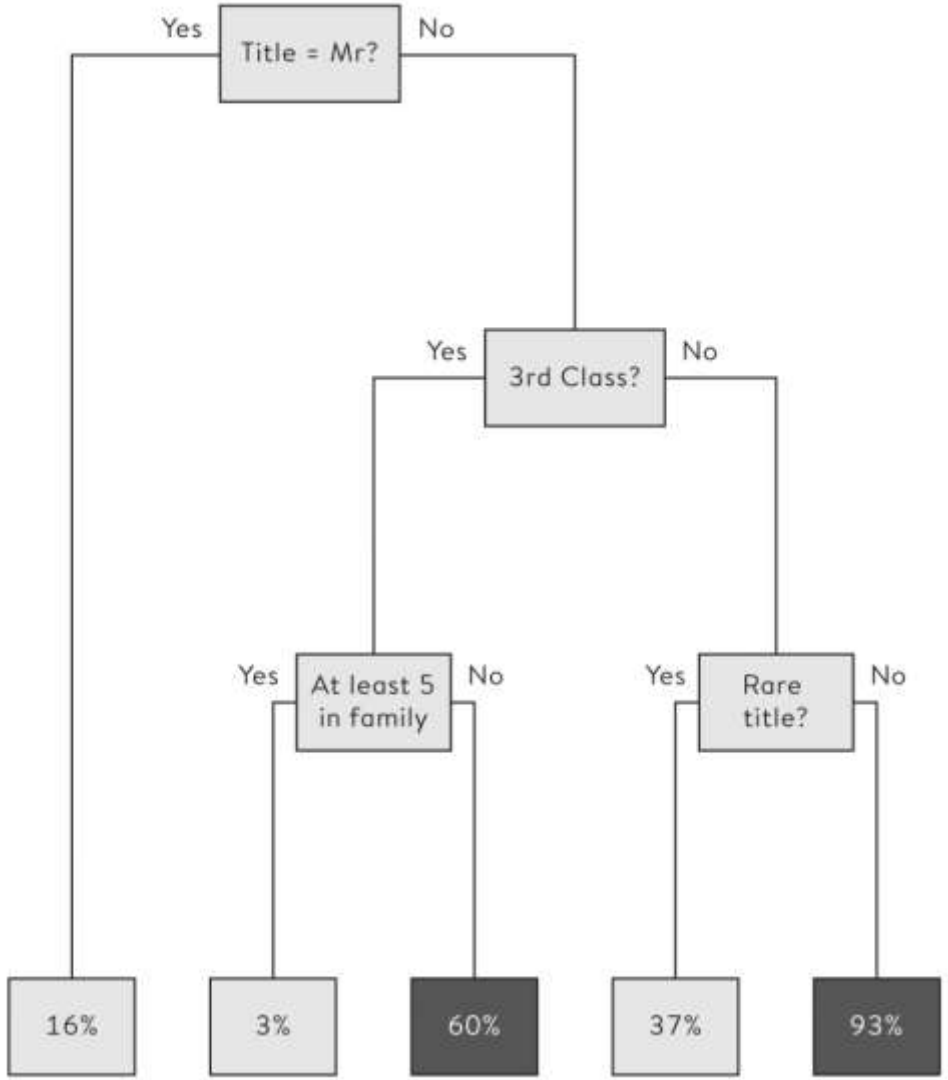


Assessing performance of an algorithm

- First, the response data is categorical: lived or died.
- This means that we can assess accuracy: how often did the algorithm we just looked at suggest death when the person died and vice versa?
- If we always say that a person died, then we have accuracy of 61%: that's how many people died of 1309 people.

Performance of this tree

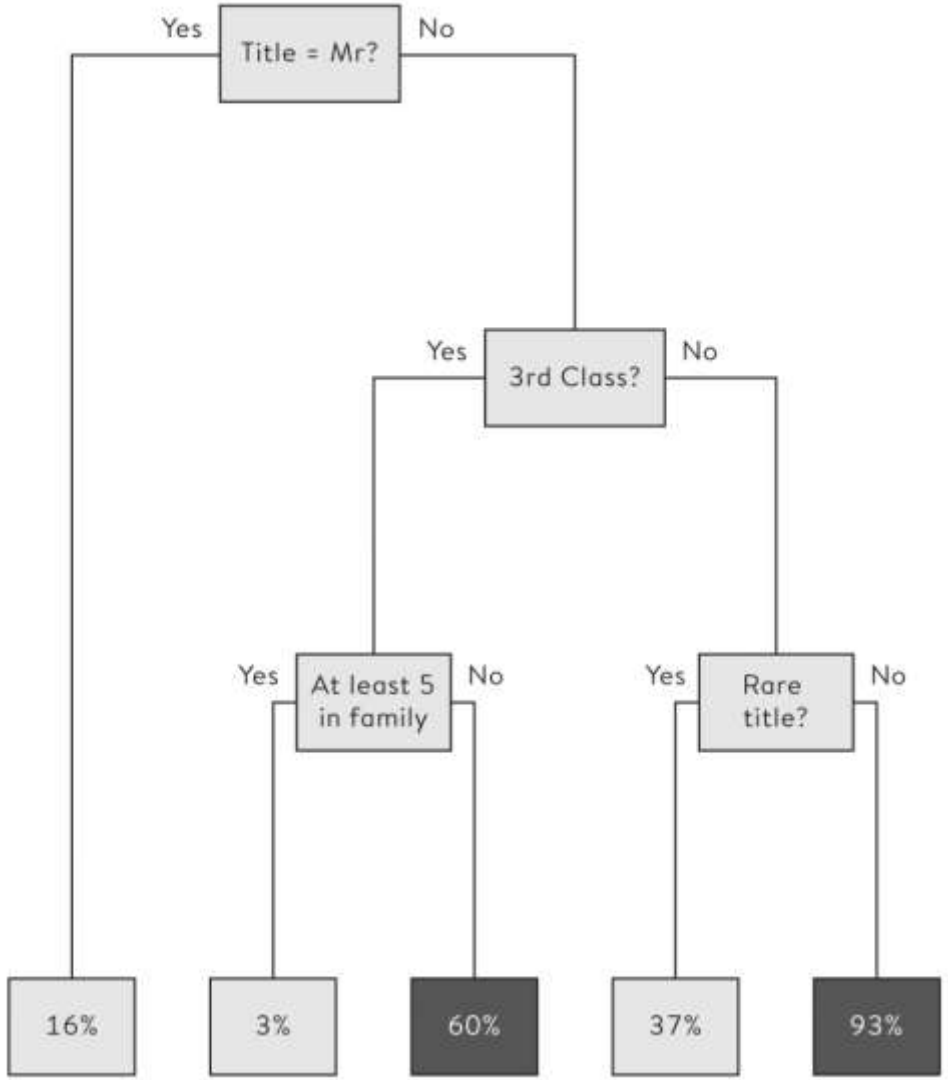
- This tree gives 82% accuracy on the training set and 81% on the test: this is expected.
- Algorithms do worse on the test set mostly, because they are likely overfitted to the training set.



Performance of this tree

- But accuracy only tells me of the times that I was right, is there a pattern to how this algorithm was wrong?

- For that we want a confusion matrix.



Confusion Matrix

- Percentage of true survivors correctly predicted is called sensitivity.
- Percentage of true non-survivors correctly predicted is called specificity
- This is especially useful when there is a large bias: what if there is 80% dead and 20% alive? Then we want to especially check the sensitivity.

Confusion Matrix

	TRAINING SET			TEST SET		
	Predicted not to survive	Predicted to survive		Predicted not to survive	Predicted to survive	
Did not survive	475	93	568	228	45	273
Survived	71	258	329	35	104	139
	546	351	897	263	149	412

Accuracy
$= (475 + 258) / 897 = 82\%$

Sensitivity
$= 258 / 329 = 78\%$

Specificity
$= 475 / 568 = 84\%$

Accuracy
$=$ [redacted] $= 81\%$

Sensitivity
$=$ [redacted]

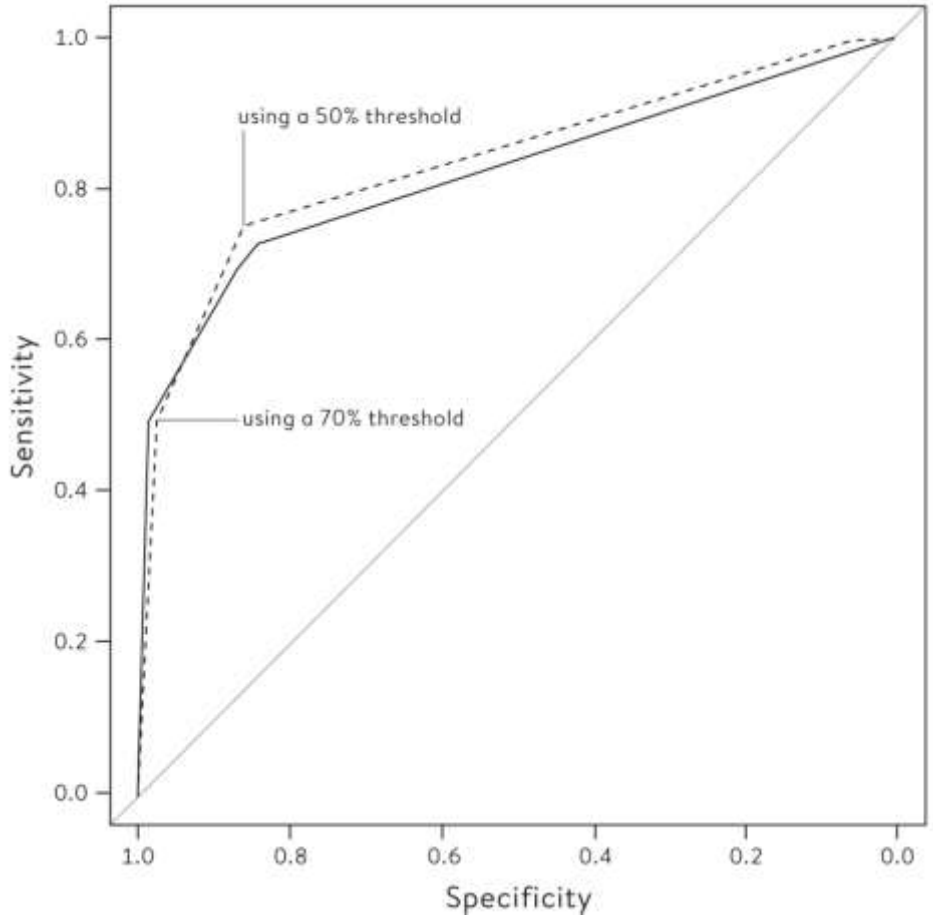
Specificity
$=$ [redacted]

Assessing performance: Using Probability

- If we only make the confusion matrix, we have dropped a crucial bit of information we received from the algorithm: the probability of survival, since we thresholded it at 50%
- Alternatively, we can see how the sensitivity and specificity change when we change that threshold
- Making a plot of sensitivity vs specificity at different thresholds give a Receiver Operating Characteristic or ROC curve (which is how everyone talks of it)

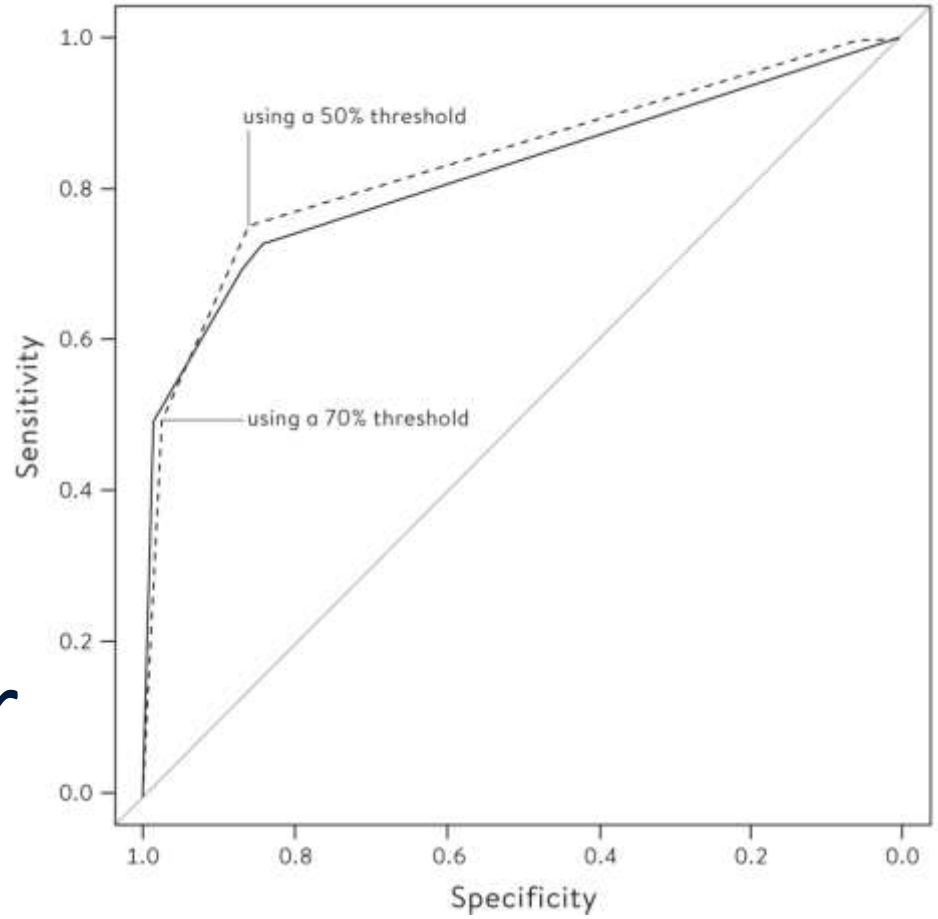
Assessing performance: ROC curve

- Solid line: training data; dashed line: test data
- Now we need a way to summarize this curve
- We can get the area under the ROC curve = 0.82



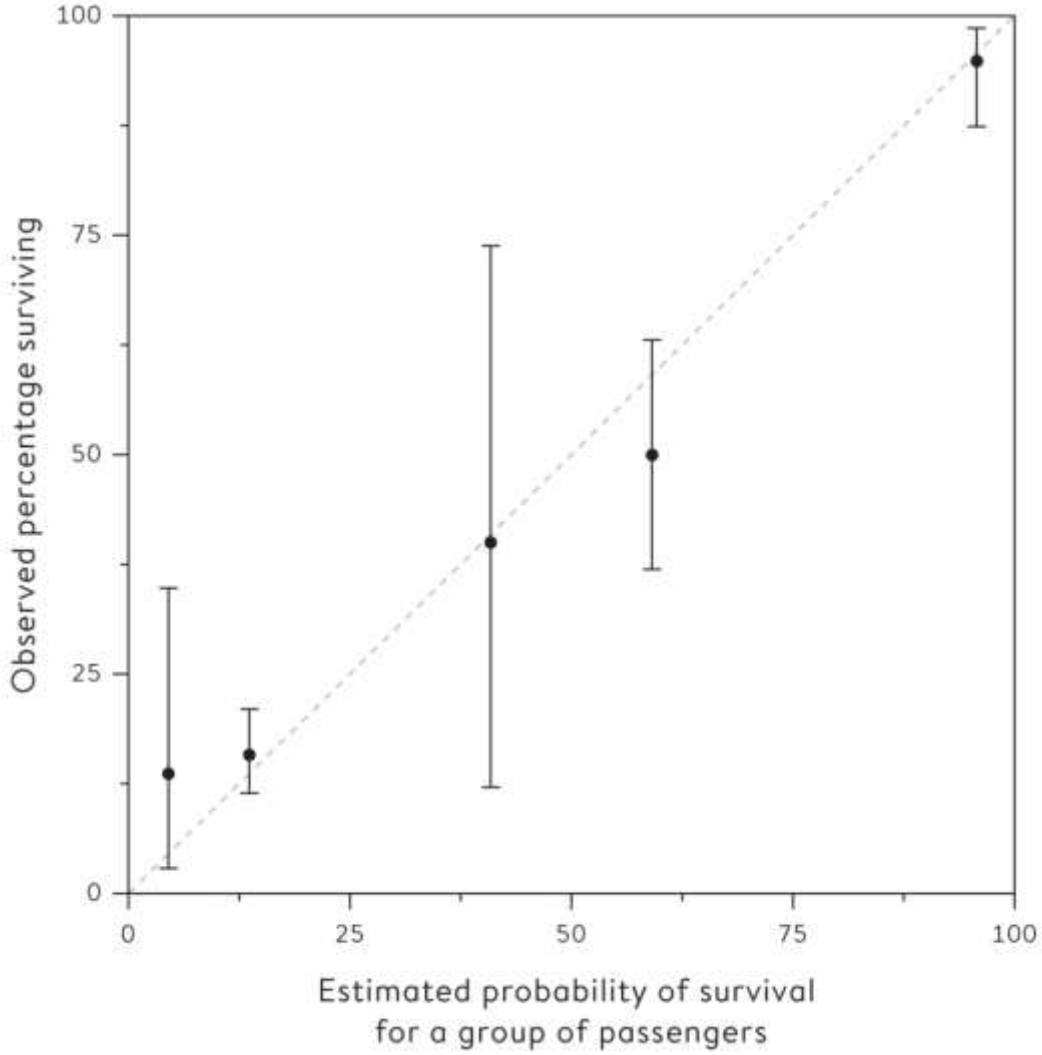
Assessing performance: ROC curve

- We can get the area under the ROC curve = 0.82
- Interpretation: If we pick a true survivor and true non-survivor at random, our algorithm will pick the survivor with 82% chance to have a higher probability of survival



Calibration Plot

- Calibration plot is a visual tool to assess the agreement between predictions and observations in different percentiles (mostly deciles) of the predicted values.



Assessing performance: Brier Score

- An alternative strategy to measure performance comes from linear regression where we measure error between our predicted value and true value and square that.
- But how do we measure that for probability?
- Basically, the exact same way: we let the response be 1 or 0 but its prediction will be a probability between 0 and 1

Assessing performance: Brier score

$$BS = \frac{1}{N}\sum(p_i - o_i)^2$$

Where p_i is the probability given our model, o_i is the true outcome (0 or 1) and N is the number of samples present.

Assessing performance: Brier Score

- So, if the probability is 0.1 and true value is 0, then the Brier score (the RSS for probabilities) will be (0

Assessing performance: Brier Score

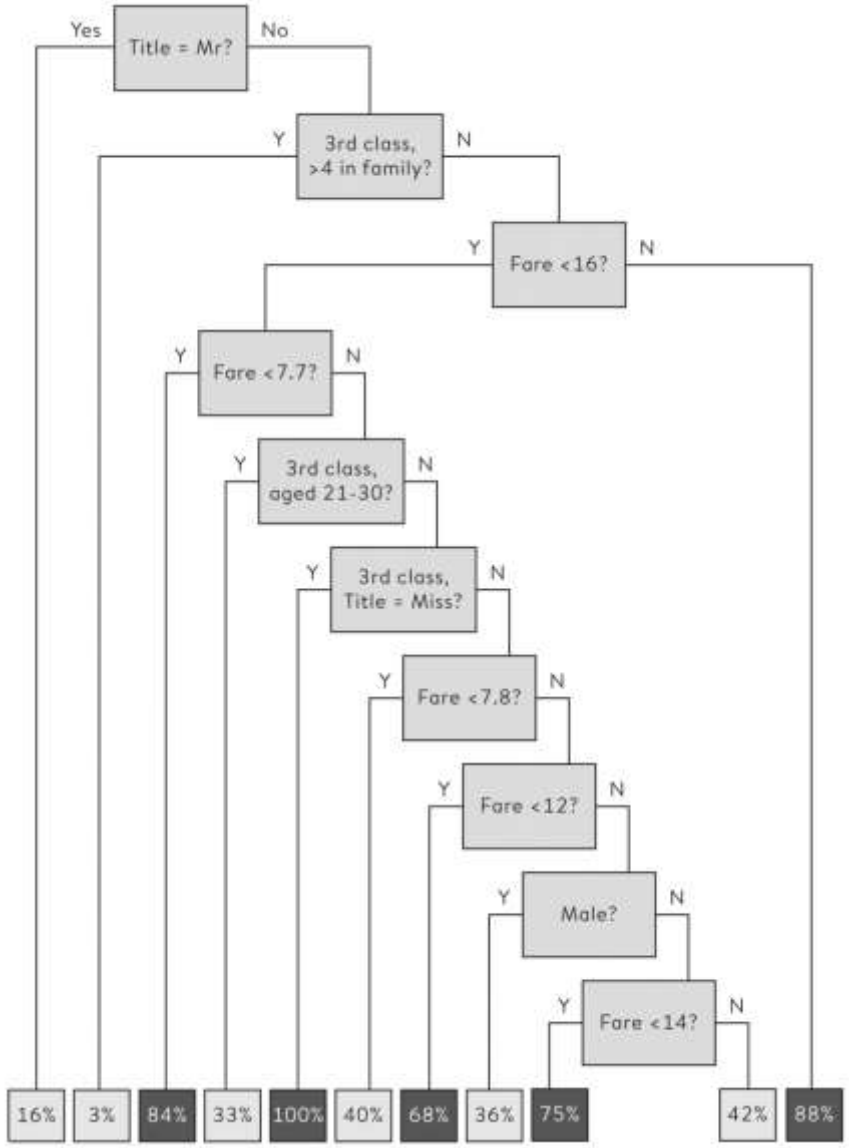
	Monday	Tuesday	Wednesday	Thursday	Friday	Mean-squared-error (Brier Score)
'Probability of precipitation'	0.1	0.2	0.5	0.6	0.3	
Did it actually rain?	No	No	Yes	Yes	No	
True response	0	0	1	1	0	
Error	–0.1	–0.2	0.5	0.4	–0.3	
Squared error	0.01	0.04	0.25	0.16	0.09	B = 0.54 / 5 = 0.11
Probability from climate	0.2	0.2	0.2	0.2	0.2	
Climate error	–0.2	–0.2	0.8	0.8	0.2	
Squared climate error	0.04	0.04	0.64	0.64	0.04	BC = 1.4 / 5 = 0.28

Assessing performance: Brier score

- For the Brier score, this means we estimate the average Brier score when we assigned a probability of 39% for survival
- Brier Score for basic prediction = 0.232
- Brier Score for decision tree = 0.139
- This means we had a 0.139/0.232 = 0.4 implying a 40% reduction in the error

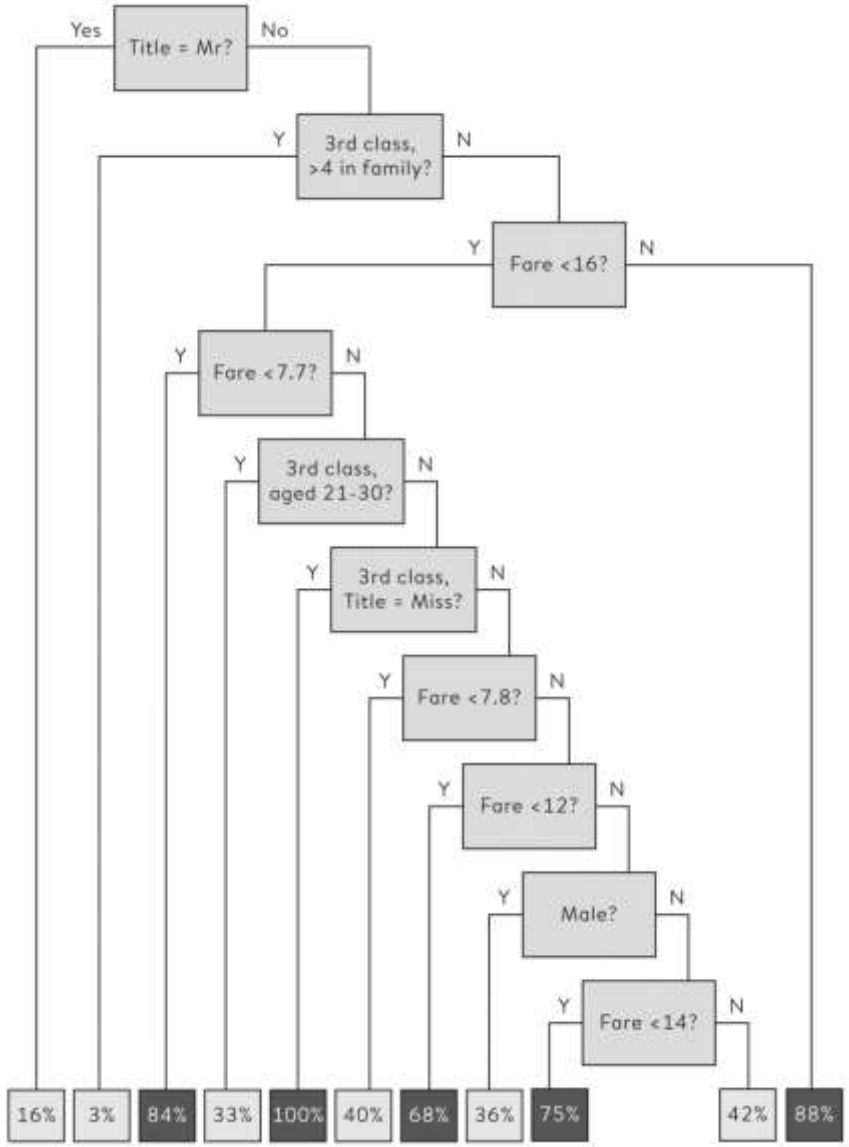
Assessing performance: Risk of Overfitting

- Consider this slightly excessive tree
- This tree gets an accuracy of 83% , better than the earlier tree
- But on the test set its accuracy is 81 % and Brier score is 0.15



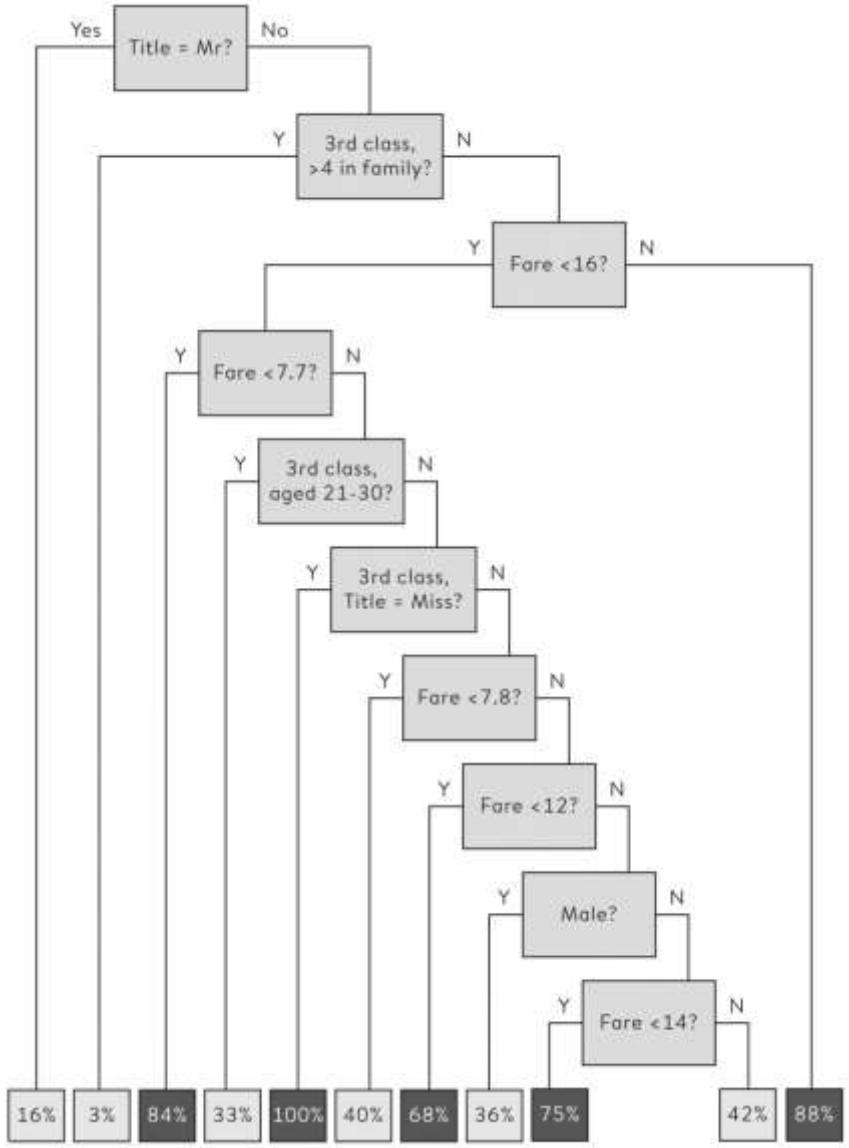
Assessing performance: Risk of Overfitting

- So, overfitting to the training data gives us worse results: why?
- When overfitting, we have adapted too much to local information
- This means less bias: that is, we are closer to the mean



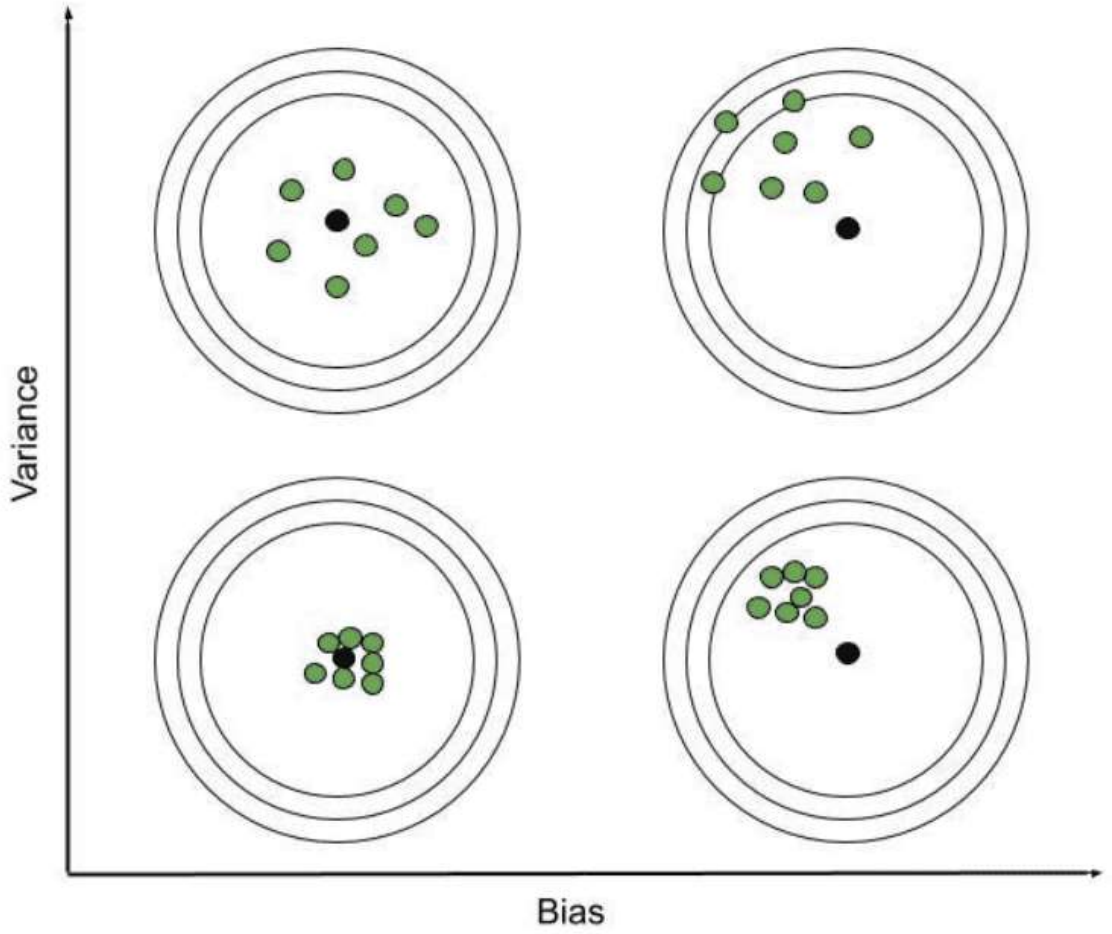
Assessing performance: Risk of Overfitting

- Overfitting causes reduced bias: algorithm is closer to the data
- But also, as we get closer to the data, the less data there is to inform our parameters, making variance or uncertainty in the parameters go up



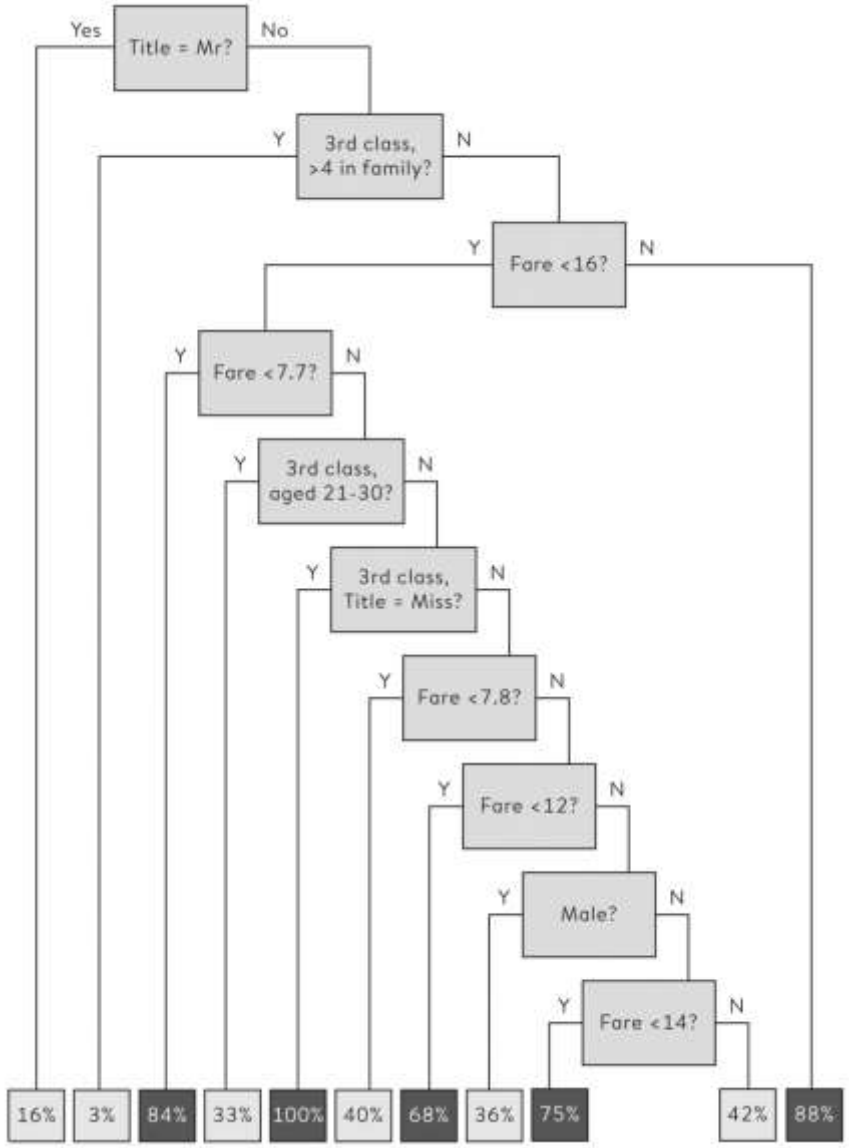
Bias and Variance

- Bias: How close to the sample mean are you with your model?
- Variance: How high is the variance of predicted response with your model?



Assessing performance: Risk of Overfitting

- How can we then try all the algorithms we want to on the training data without fear of over-fitting?
- We can turn to another clever trick: cross-validation

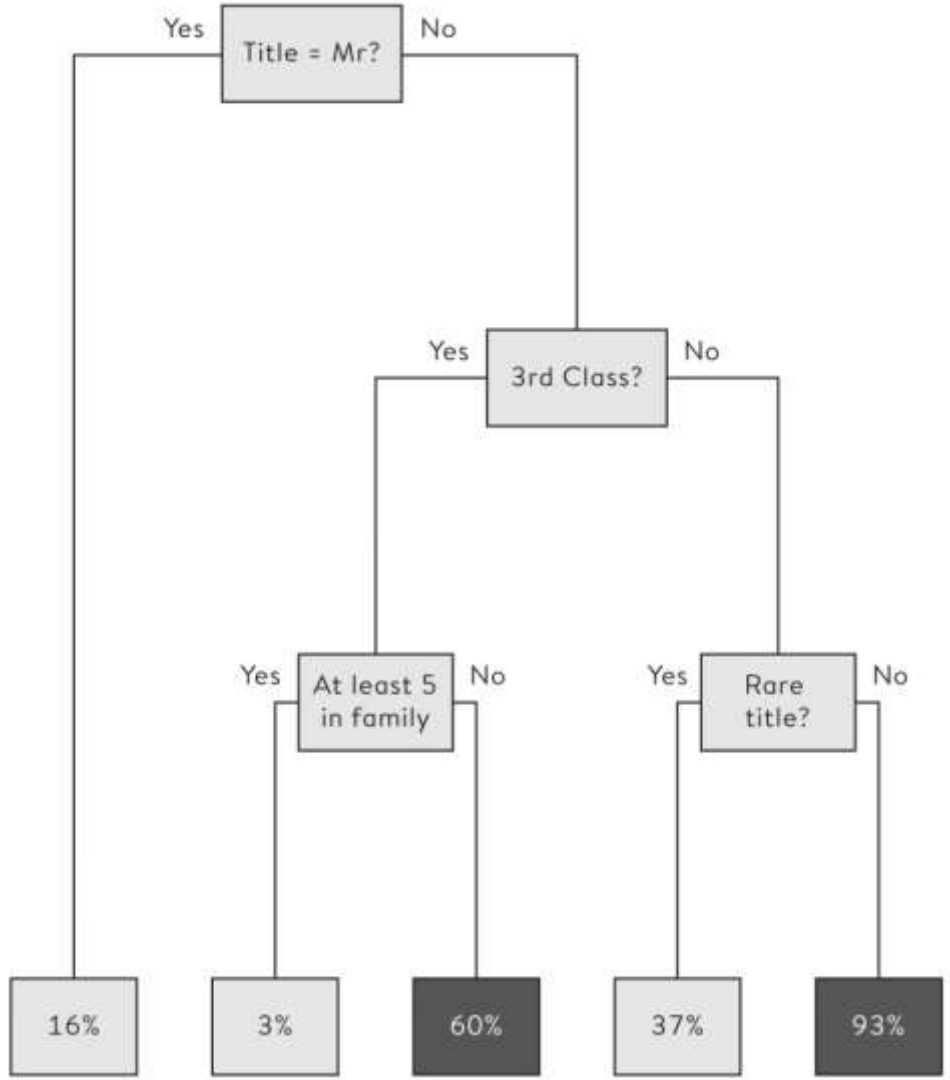


Improving performance: Cross-validation

- While we have already split training and test, we can further sub-divide the training so that we can reduce over-fitting
- Then we fit the tree on say 80% of the training data and check how good the tree works on the left out 20%, then repeat this with a new 80-20 split several times: this is called K-fold cross-validation

Improving performance: Cross-validation

- 10-fold cross validation was used to produce the tree to the right here.
- But, this was testing variations on one algorithm, could we try a different algorithm?



Cross-validation for Hyperparameter Selection

Definition

Hyperparameters: Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.

- Hyperparameters for decision trees:
  - Depth of tree
  - Stopping condition
  - Method for deciding optimal splits (note: not technically a hyper*parameter* but still can be chosen this way)
- Grid-search over hyperparameter values under cross-validation

Examine another classification tree.

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

Alternative algorithms: Regression

- We can build linear regression models and their generalization.
- Here, we can work with one such generalization: logistic regression, where the response has 0 or 1 as values only.
- $Prob(Y) = logit(X'\beta)$ where the logit function is:

$$\exp(X'\beta)/(1 + \exp(X'\beta))$$

- Brier score = 0.146 -> worse than the tree

Alternative algorithms: Regression

- Similar to decision tree, but now it is additive.
- A young man in third class will have a score of what?
- -1.91 which passed through the logit function (e(x)/1+e(x)) gives a probability of ?

Characteristic	Score
Starting score	3.20
Third class	−2.30
'Mr'	−3.86
Male in third class	+1.43
Rare Title	−2.73
Aged 51–60 in second class	−3.62
Each member of family	−0.38

Alternative algorithms

Method	Accuracy (high is good)	Area under ROC curve (high is good)	Brier score (low is good)
Everyone has a 39% chance of surviving	0.639	0.500	0.232
All females survive, all males do not	0.786	0.578	0.214
Simple classification tree	**0.806**	0.819	**0.139**
Classification tree (over-fitted)	**0.806**	0.810	0.150
Logistic regression	0.789	0.824	0.146
Random forest	0.799	**0.850**	0.148
Support Vector Machine (SVM)	0.782	0.825	0.153
Neural network	0.794	0.828	0.146
Averaged neural network	0.794	0.837	0.142
K-nearest-neighbour	0.774	0.812	0.180

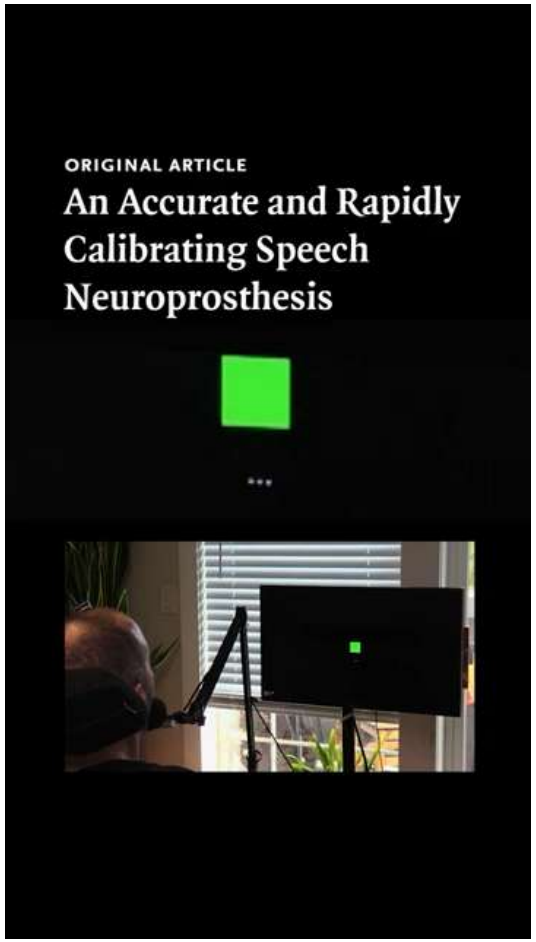
Challenges of ML Algorithms

- Lack of robustness
- Doesn’t account for statistical variability
- Implicit bias
- Lack of transparency

And yet…

But, algorithms exist that can clearly make life better.

https://www-nejm-org.mu.idm.oclc.org/doi/full/10.1056/NEJMoa2314132



A man unable to speak uses a neuroprosthesis that relies on machine learning algorithms to chat with his daughter.

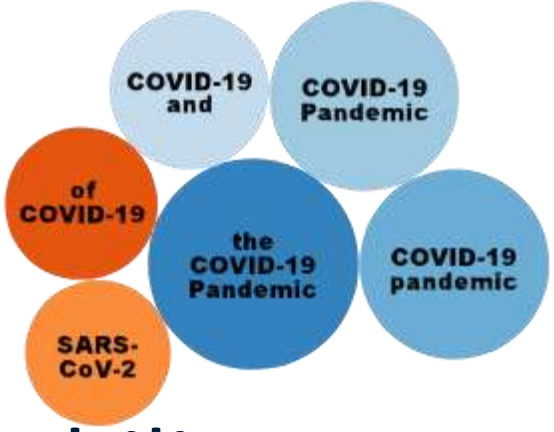
A now ubiquitous algorithm: ChatGPT

- An example of *unsupervised* learning, part of the broader topic of natural language processing
- GPT = Generative Pre-trained Transformer
- Generative: generates new words
- Pre-trained: Uses training data to create the model, and could be improved with more data
- Transformer: A certain structure of model within the space of deep learning - a subset of neural networks

A now ubiquitous algorithm: ChatGPT

- What is the general way that ChatGPT works?
- Let's consider the idea of context.
- What is the context of this class? What is the larger narrative in which this class sits?
- ChatGPT does this but with only one '*modality*' which is text (images, sound are different modalities)
- ChatGPT is based on the idea that if you train billions of parameters to understand the best-fit '*line' (you will be able to compare this to linear regression)* of context then you get pretty close to human language

N-grams



- Word: 1-gram with high probability
- Priority seating: 2-gram with high probability
- This is a sentence: 4-gram with high probability
- Note: A sentence can be a 1-gram in a higher-level schema, where a set of sentences may be a 5-gram representing a paragraph. Or a character, ‘a’ can be a 1-gram instead. Here we’ll focus on words.

N-grams

- Asdas: 1-gram with low, but not 0, probability (exists on the internet because of how keyboards are built)
- Prob loa: 2-gram with low probability
- Asd trwgr dfvv eqq: 4-gram with low probability

175 billion parameters to determine the probabilities of 13000-grams

- GPT3 has trained 175 billion parameters to model the probabilities of all possible 13000-grams found in its dataset
- This means that when you enter the lines: ‘Write me a fairy tale with a dolphin.” GPT3 finds the optimal 13000 gram that sentence fits into and generates text for you.
- For more details see: *For more see* *https://www.3blue1brown.com/lessons/gpt*

The idea of scale

- Central to the success of transformers with language and with images.
- The idea was if you fit billions and billions of parameters with billions and billions of samples of data, you get something impressive. That’s it.
- This is why every new giant-AI model claims to have billions of parameters more than the last one.

Best-fit Line Analogy

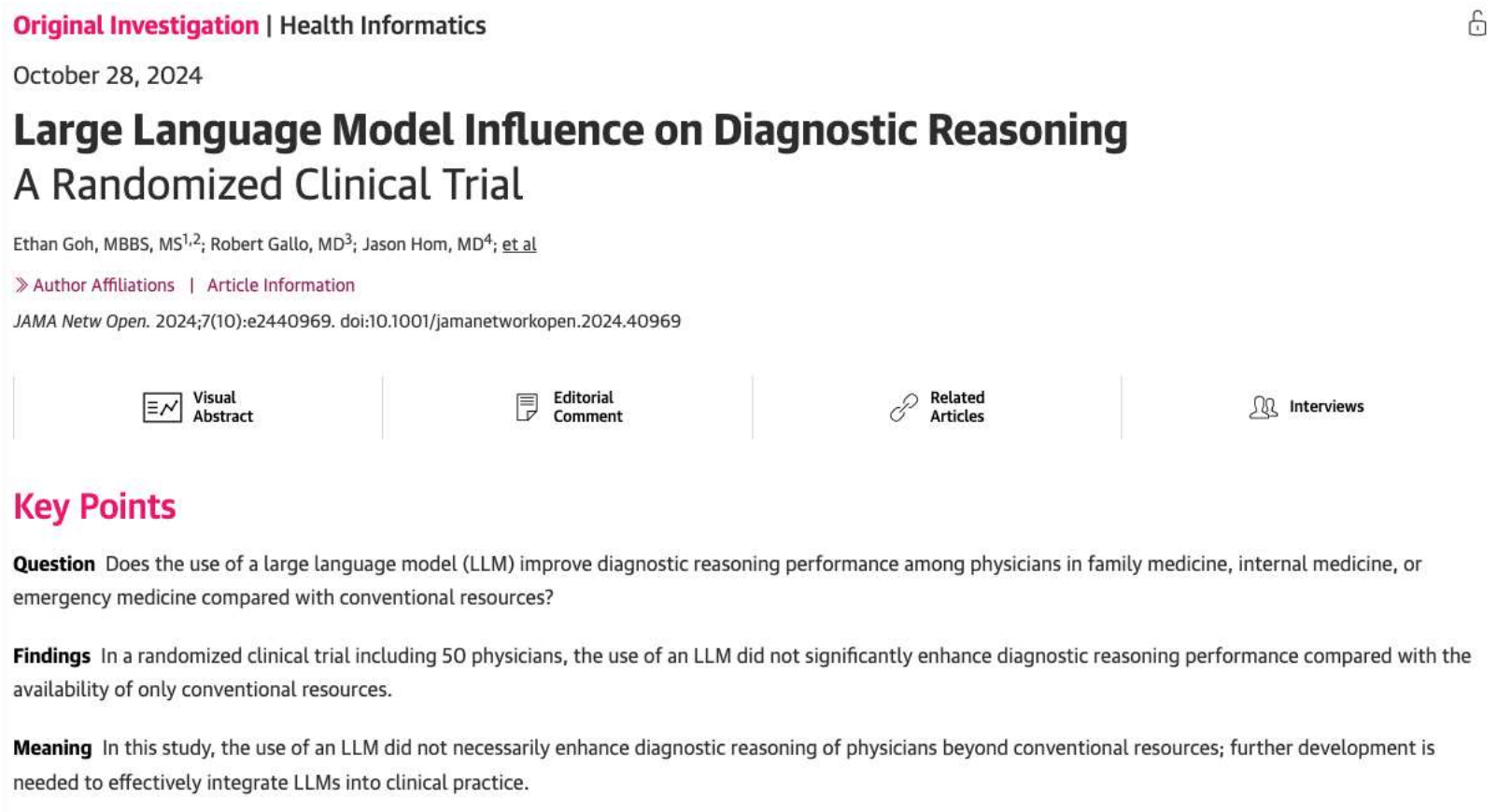
- GPT, Claude, DeepSeek, etc. all are neural networks with billions of parameters fit to the data.
- Essential operation is to predict new text with those billions of parameters.
- Any prediction exists within the fitted line of the model. We don't get anything truly *new*. ***Everything exists within the manifold (high-dimensional line) created on 12,728 dimensions***.
- Lot of room for surprising combinations (which may be useful), this is why we feel surprise when playing with these models.

Not hallucinate - *bullshit*

- The false information produced by LLMs is called *hallucinations.*
- This misses the heart of the issue though.
- To hallucinate, you go beyond the reality in some way.
- LLMs have no model of reality that is tested against the real world i.e. they are indifferent to *truth.*
- Instead of hallucinations, they *bullshit* because what they say is unlinked from reality (no interventional effects).

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38.

Do LLMs help? No, not really in healthcare so far.



Original Investigation | Health Informatics

October 28, 2024

Large Language Model Influence on Diagnostic Reasoning
A Randomized Clinical Trial

Ethan Goh, MBBS, MS[1,2]; Robert Gallo, MD[3]; Jason Hom, MD[4]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

Visual Abstract | Editorial Comment | Related Articles | Interviews

Key Points

Question Does the use of a large language model (LLM) improve diagnostic reasoning performance among physicians in family medicine, internal medicine, or emergency medicine compared with conventional resources?

Findings In a randomized clinical trial including 50 physicians, the use of an LLM did not significantly enhance diagnostic reasoning performance compared with the availability of only conventional resources.

Meaning In this study, the use of an LLM did not necessarily enhance diagnostic reasoning of physicians beyond conventional resources; further development is needed to effectively integrate LLMs into clinical practice.

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395

Do LLMs help? Maybe for research.

Article | Open access | Published: 27 November 2024

Large language models surpass human experts in predicting neuroscience results

Xiaoliang Luo ✉, Akilles Rechardt, Guangzhi Sun, Kevin K. Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O. Cohen, Valentina Borghesani, Anton Pashkov, Daniele Marinazzo, Jonathan Nicholas, Alessandro Salatiello, Ilia Sucholutsky, Pasquale Minervini, Sepehr Razavi, Roberta Rocca, Elkhan Yusifov, Tereza Okalova, Nianlong Gu, Martin Ferianc, Mikail Khona, Kaustubh R. Patil, Pui-Shee Lee, ... Bradley C. Love ✉ + Show authors

Nature Human Behaviour (2024) | Cite this article

86k Accesses | **1186** Altmetric | Metrics

Abstract

Scientific discoveries often hinge on synthesizing decades of research, a task that potentially outstrips human information processing capacities. Large language models (LLMs) offer a solution. LLMs trained on the vast scientific literature could potentially integrate noisy yet interrelated findings to forecast novel results better than human experts. Here, to evaluate this possibility, we created BrainBench, a forward-looking benchmark for predicting neuroscience results. We find that LLMs surpass experts in predicting experimental outcomes. BrainGPT, an LLM we tuned on the neuroscience literature, performed better yet. Like human experts, when LLMs indicated high confidence in their predictions, their responses were more likely to be correct, which presages a future where LLMs assist humans in making discoveries. Our approach is not neuroscience specific and is transferable to other knowledge-intensive endeavours.

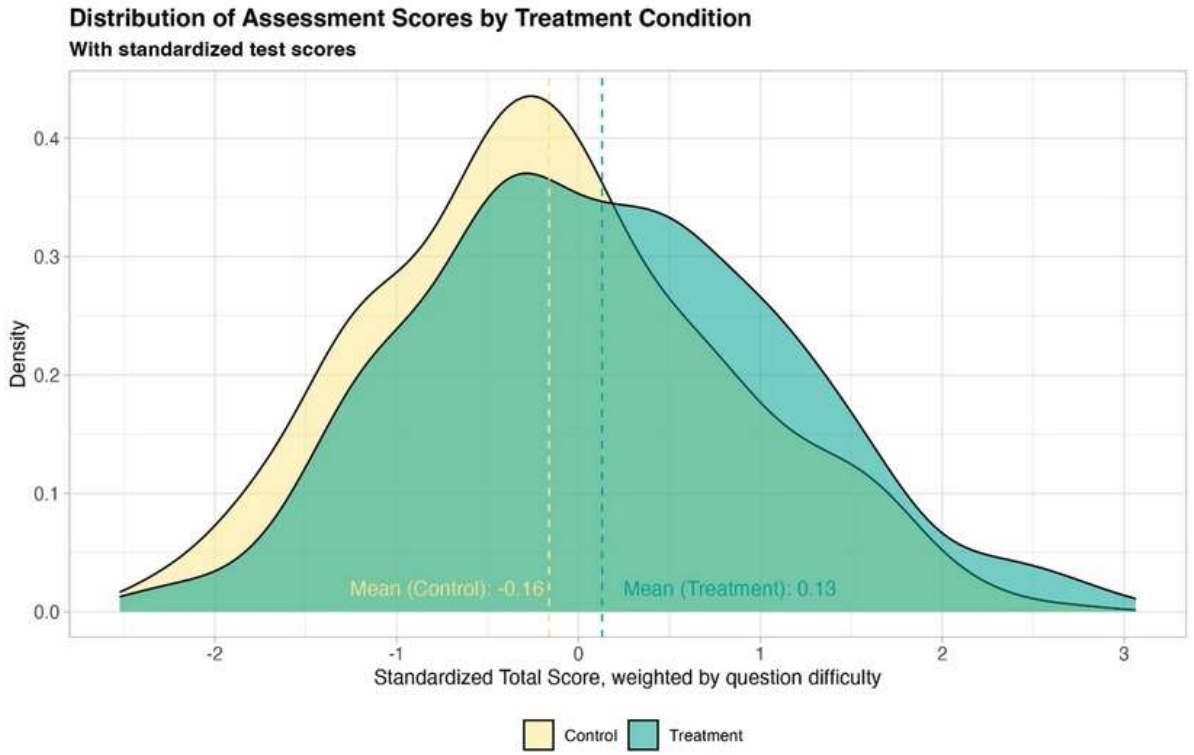
Do LLMs help? Yes, as virtual tutors in education.

From chalkboards to chatbots: Transforming learning in Nigeria, one prompt at a time

MARTÍN E. DE SIMONE, FEDERICO TIBERTI, WURAOLA MOSURO, FEDERICO MANOLIO, MARIA BARRON & ELIOT DIKORU | JANUARY 09, 2025

This page in: English

Students participate in an AI after-school program in Edo, Nigeria. Copyright: SmartEdge/World Bank



Campbell-Goodhart law

- Fitting to the measure instead of to the original intended goal.
- Major problem of current AI, leads to the feeling of shifting goalposts (ok, NOW we have General AI because it can do X, but no, we need it to do Y)
- The issue is that there simply isn't a good metric of intelligence.
- Also, a problem more generally for statistics.

Campbell-Goodhart law

1. ”Teaching to the test”: When schools are measured on standardized testing to get funding, teachers just help students get better at test taking.
2. Policing and crime statistics: An easy way to reduce crime is to stop reporting it.
3. The Big One: If we only focus on CO2 emissions, companies within a rich country will offload their carbon emission to developing nations.

A Personal Opinion

- The idea of *Metis*, also called implicit learning, and many other terms, defined as local, time-dependent, informal (no protocols, no math), decentralized knowledge.
- Essentially, this is how we build competence without comprehension for an infinite parameter reality that is only stationary on the truly macroscopic (planets) and microscopic scales (quantum scale).
- Depends critically on causal, interventional navigation of the world.
- No current AI is capable of this, and it will hit a fundamental barrier as a result. Though they still have remarkable competence without comprehension that remains to be explained.

If you want to read more on AI…

- An old public-facing article from me: https://archive.is/VN8r5
- Articles from Melanie Mitchell: https://aiguide.substack.com/p/the-llm-reasoning-debate-heats-up
- Nature of Intelligence podcast series from the Santa Fe Institute: https://www.santafe.edu/culture/podcasts
- Stuff from Ethan Mollick: https://www.oneusefulthing.org/p/thinking-like-an-ai (funny that he says: "I came to the conclusion that the advice in my book is still the advice I would give: just use AI to do stuff that you do for work or fun, for about 10 hours, and you will figure out a remarkable amount" since it suggests AI is just another tool, not a magic bullet.)

Data has no meaning. We give data meaning.