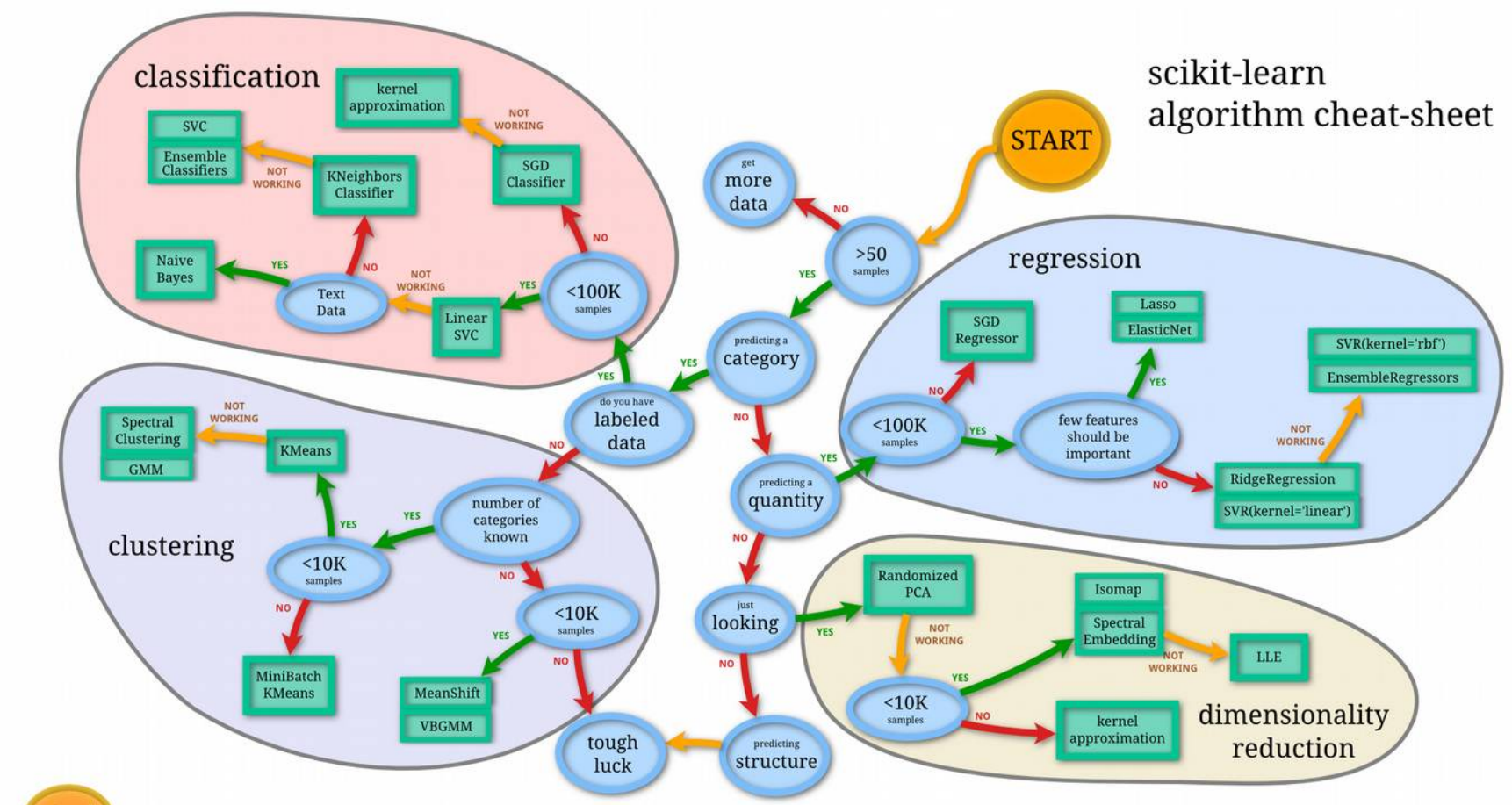
Critical Thinking about Machine Learning



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Goals of this section

‘Here is a thing I want ML to do, is it reasonable?’

- Is the problem well-posed for ML? (goal and task alignment)
- Is the problem possible? (data, etc)

‘I am reading about an ML result that claims something, should I believe it?’

- Evaluation criteria, baselines, identifying sources of improvement
- Data leakage
- Correlation vs causation

Here is a thing I want to do
Is it reasonable?

Scientific problems

Scientific problems often take the form of:

- Want to know **what is true**/test hypotheses
- Want to relate **causes** to **effects**
- Want to **explain** or **understand** observations

Unless very carefully constructed, ML doesn't:

- Distinguish between what is true and what works
- Distinguish correlation from causation
- Work via reference to an established framework of concepts

Well-formed problems

General principle: you cannot safely assume ML approaches will do anything beyond exactly optimizing the function you've asked to be optimized.

A well-formed problem:

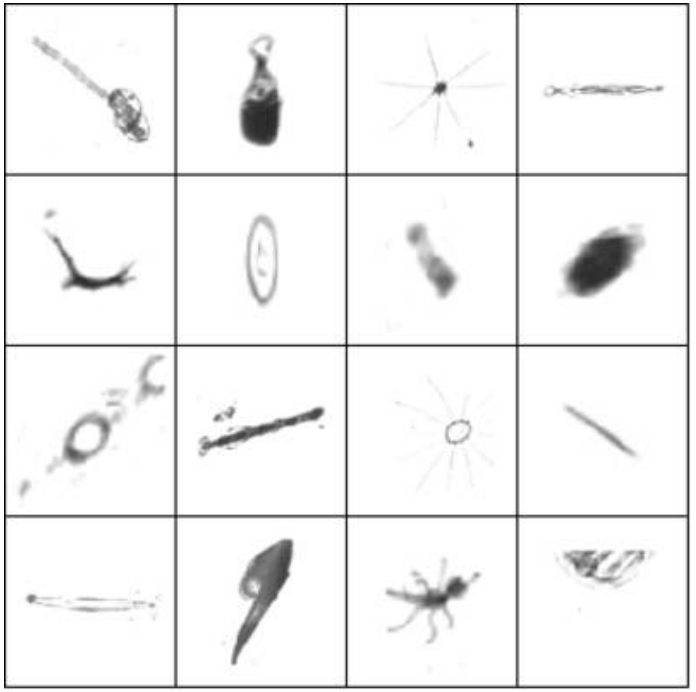
- The specified task is exactly and **entirely** the thing you care about, such that the method is irrelevant
- You have a strong way of verifying the thing you actually want **in the way it will be deployed**
- Approximate solutions are acceptable

Positive example: Plankton counts

https://www.kaggle.com/c/datasciencebowl
Hatfield Marine Science Center at Oregon State University.

Plankton population estimation:

- Drag a camera through the water
- Graduate students hand-label and count examples



http://benanne.github.io/2015/03/17/plankton.html

Neither students nor ML are perfect

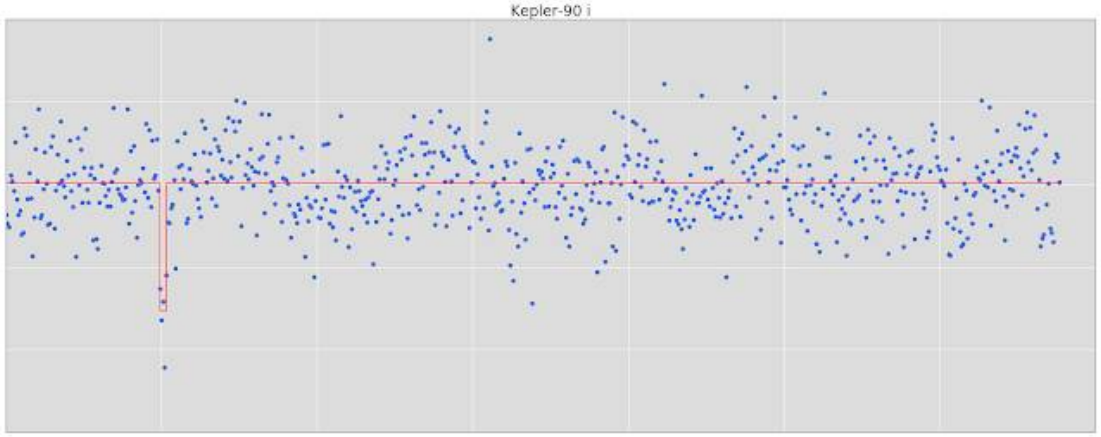
Explanation of individual plankton classifications erased in aggregate:

Both are effectively able to act as a black box process

Positive example: Pre-filtering

Sometimes must sift through large numbers of negative examples to find positive examples, which are then studied and interpreted in detail.

Exoplanet light curves



https://ai.googleblog.com/2018/03/open-sourcing-hunt-for-exoplanets.html

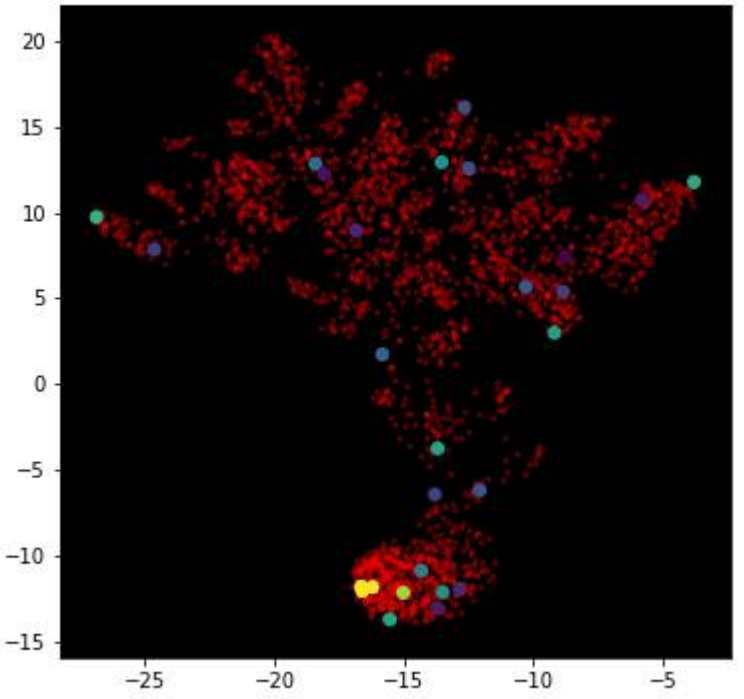
Caveat: Same as instrumental considerations, ML may introduce bias in the form of differential sensitivity.

Difficult case: Exploratory analysis

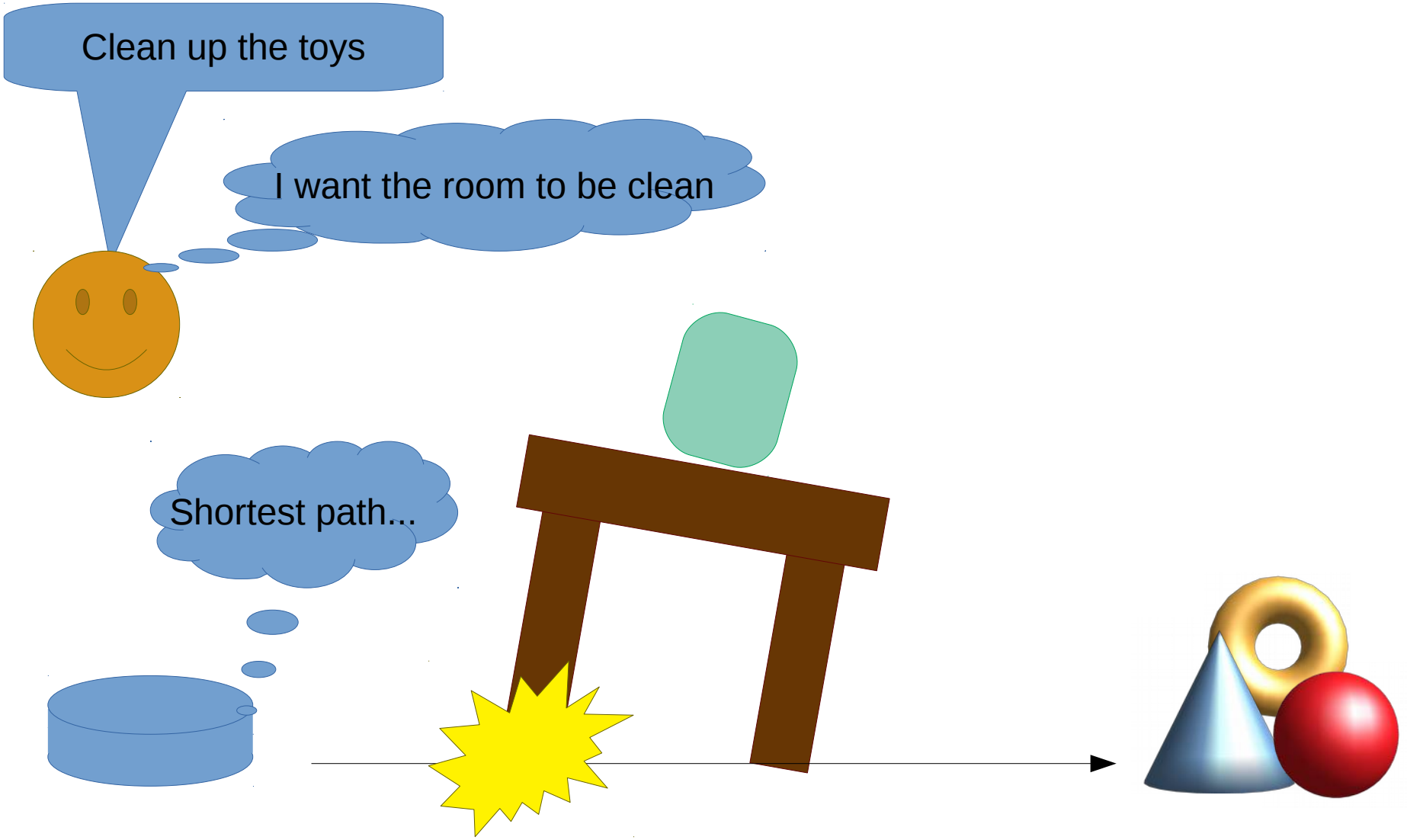
Sometimes ML can be used to discover that there are features which need explanation in a dataset that is otherwise opaque.

However, much like visualization techniques, the ML components of the process are generally **not the primary scientific result**.

This can be very useful to guide science, but in the end, the interpretation usually should not rest directly on the ML output itself.



Unintentional side-effects



How much data is big data?

Modern computer vision networks are trained on millions of images. Language models on hundreds of billions of words. Speech recognition on years of audio data.

How much do you need?

For classical approaches, need data **roughly** linear in # of features.

Linear problem

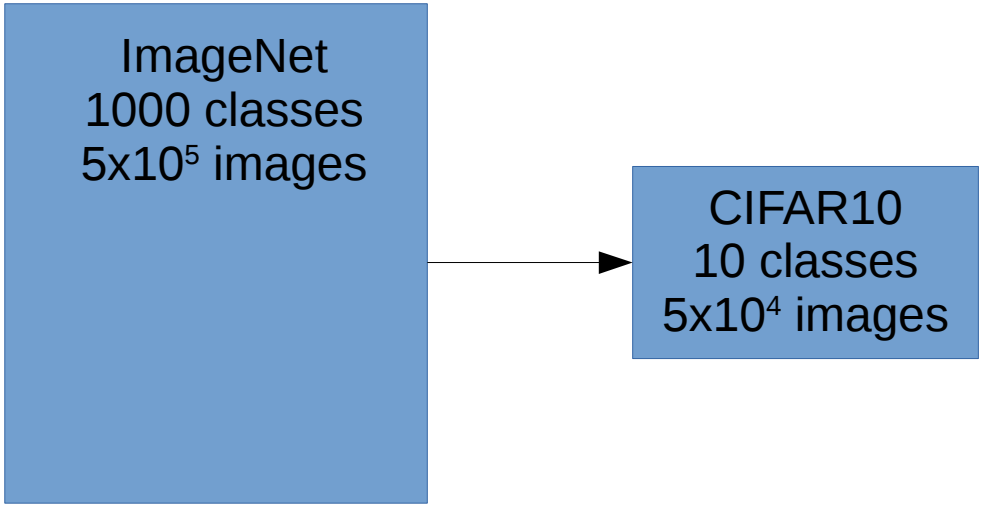
Fully non-linear problem

How much data is big data?

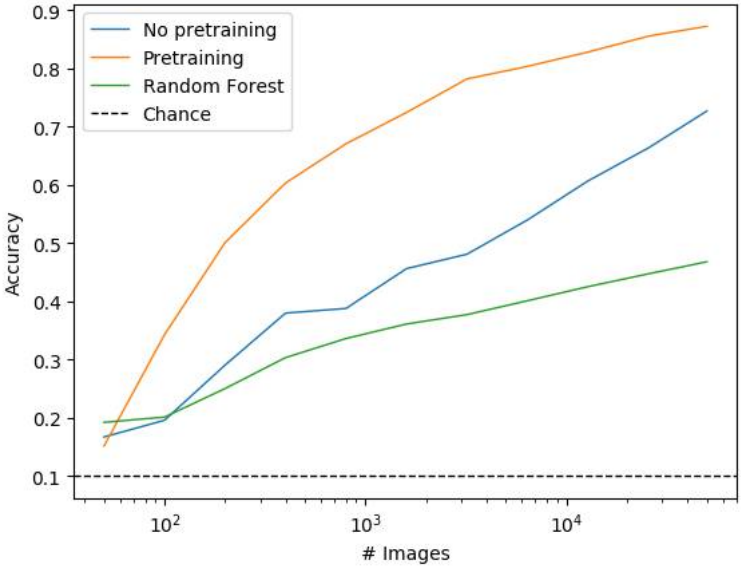
Neural networks need more data to get started, but scale better. <1000 data points is difficult. Exceptions:

- **Pretraining:** If you have a related set of problems, a network trained on those problems needs less data on new problems. Needs task generator, or well-established domain (images)

- **Data Augmentation**: Train on perturbed examples to increase data diversity. Helps a bit, but much less than pretraining.



CIFAR-10, VGG-19 architecture



Will more data help?

In image, text, and audio domains:

- Continued increases in accuracy in the ~10^9 data point range if the network is scaled appropriately.
- However, returns are **logarithmic**.

Have an idea of what level of performance would be ‘good enough’ for your scientific goals **before you start.**

Impossible problems

The most successful ML problems are ones where (we expect) than an expert human could achieve near 100% performance:

We know that it is solvable.

For novel problems, the peak performance possible may actually be very low.

In those cases, **more data will not help**.

Human intuition doesn't predict ML performance

Quite common for people to claim 'this is impossible for ML' or 'this shouldn't work' and be proven wrong immediately after.

With ML, you don't have to imagine how to solve a problem before trying it. Things that seem inelegant or arbitrary or nonsensical often work.

But **even** if you can imagine a way to solve it, a specific ML technique may consistently fail.

Listen to the results

And take care about what they actually mean

Should I believe it?

Accuracy vs AUC

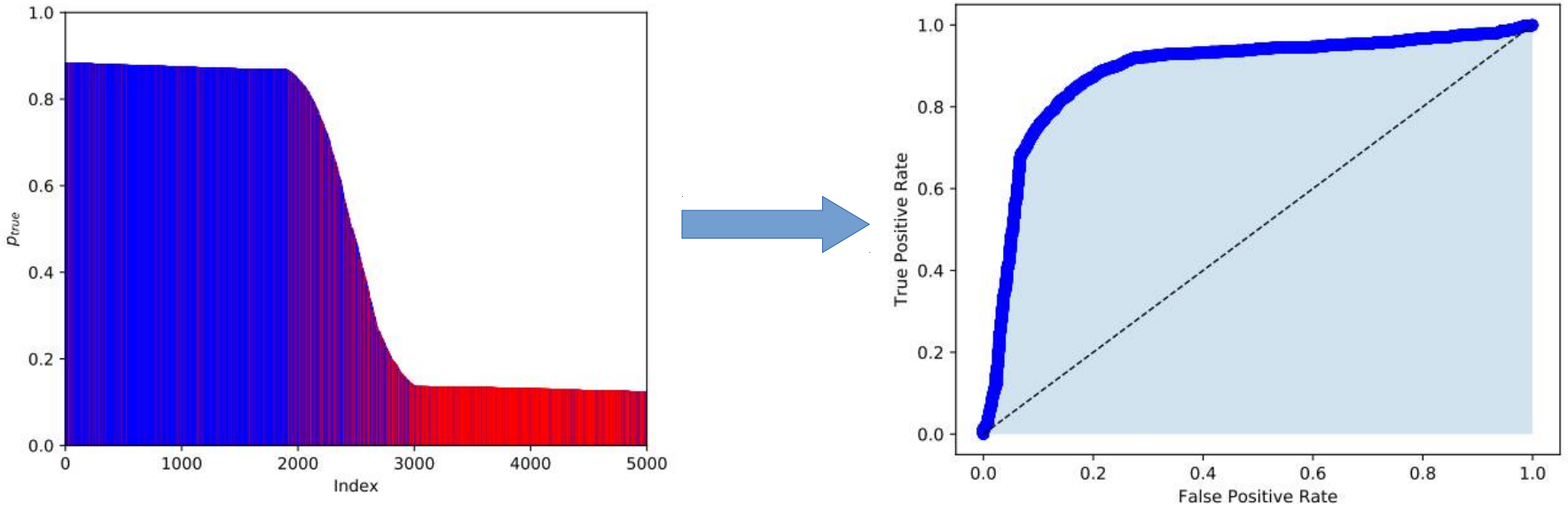
If someone trains a network to analyze a medical imaging test and reports that ‘the diagnostic **accuracy** is 90%’, is that high?

Often, data are **unbalanced**. If 90% of the cases are ‘false’, then 90% accuracy is chance level.

Accuracy vs AUC

Area Under Curve (AUC):

- Sort all test cases by predicted probability of the label in descending order
- Take the top X%, varying X. Plot fraction of total true positives captured vs fraction of false positives included.
- What is the area under that curve?

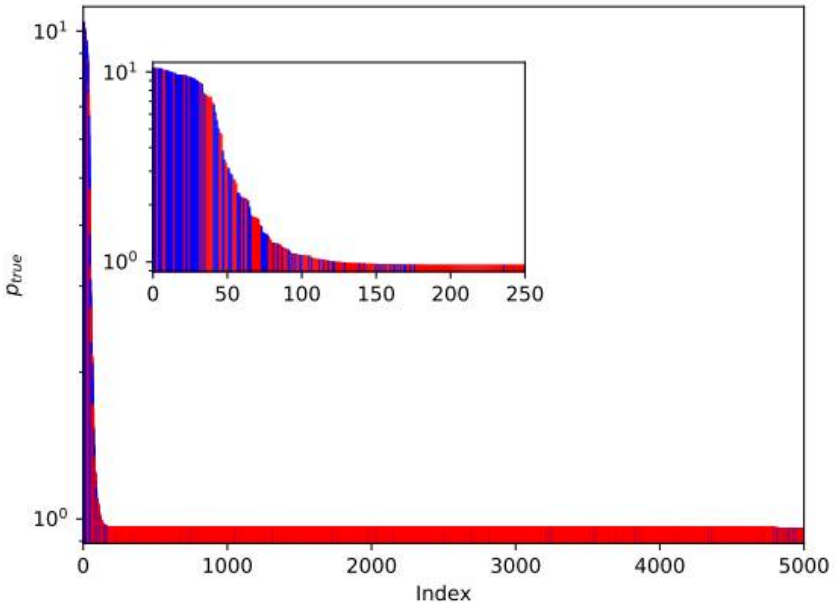


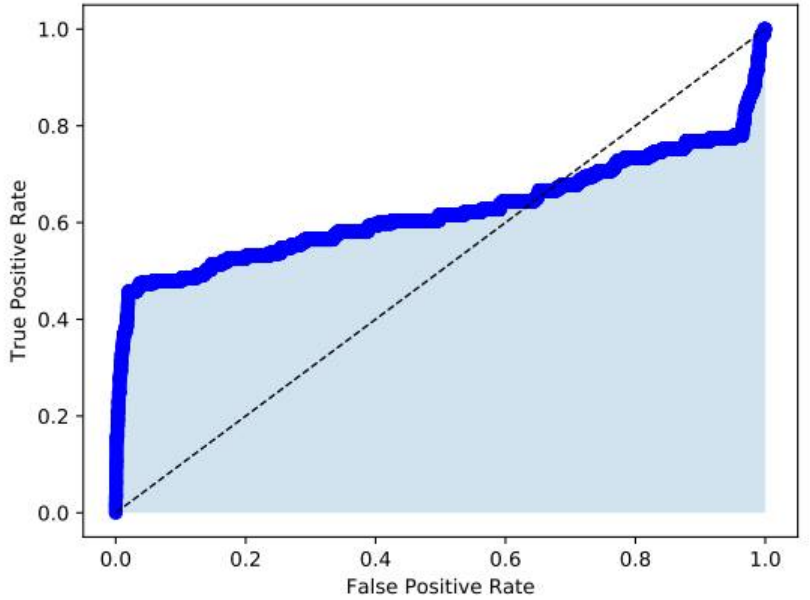
Accuracy vs AUC

Unbalanced case: 24/25 label 0, 1/25 label 1

Accuracy is 97%, but chance level is 96%

AUC is 0.622, chance level is 0.5





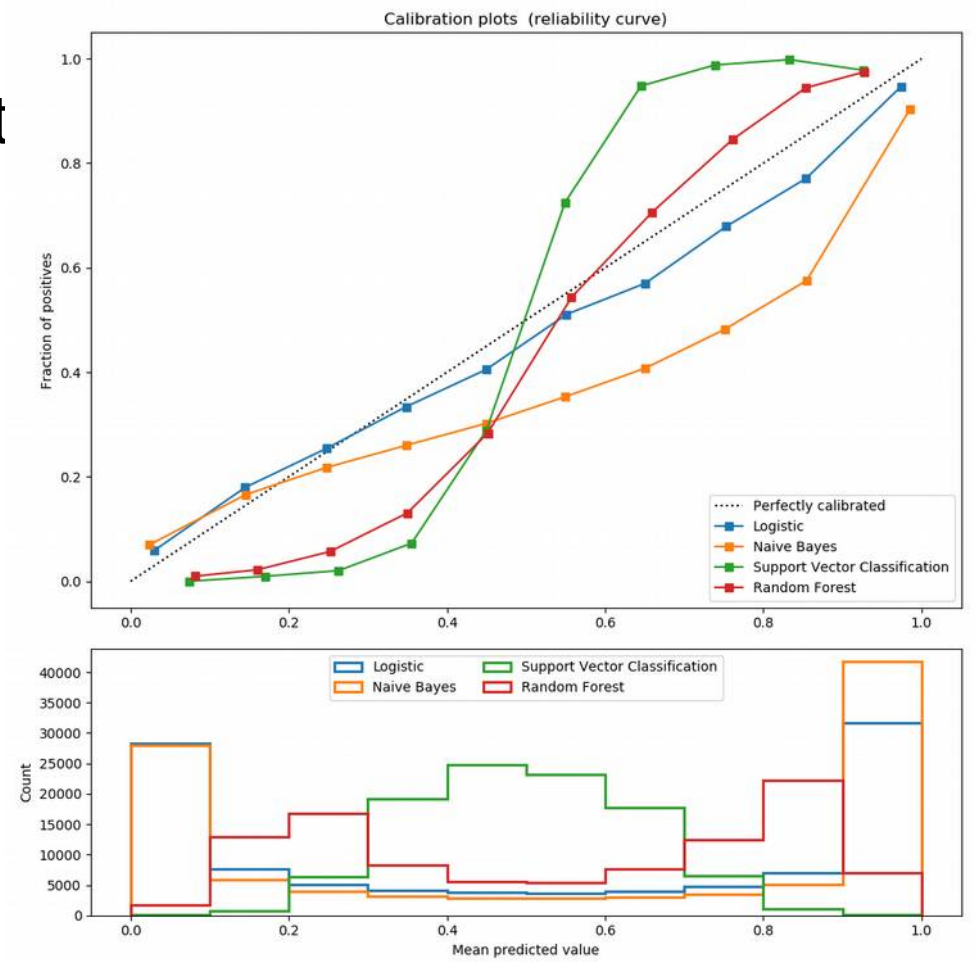
Probability Calibration

Often it would be useful to interpret the probability output of a model as if it were an actual probability, but justification for this is tenuous

Probability calibration:

Take all the points for which the model gives a certain probability of a label:

Is the model's error rate for those points equal to the probability given by the model?



https://scikit-learn.org/stable/modules/calibration.html

Distribution of Errors

Often only the overall test error rate is reported. But errors can be distributed non-uniformly or in ways that correlate with other variables.

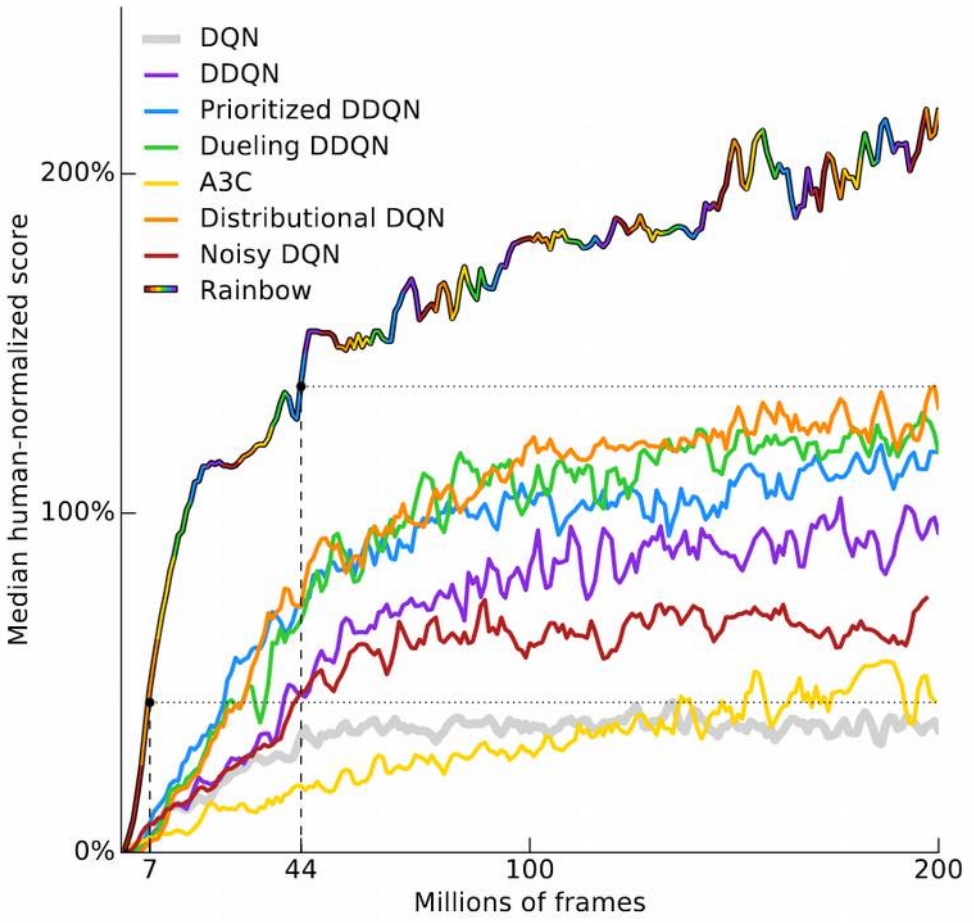
This can have a large impact on the suitability of the model.

Baselines

Does this technique improve a method?

Evidence is generally a comparison between the new model score with existing methods used as a baseline.

Some domains have accepted baseline tasks. In others it's not that simple.



https://www.alexirpan.com/2018/02/14/rl-hard.html

Baselines

In practice, many things can be responsible for a difference in scores:

- Different amount of computational resources used to find the best hyperparameters.
- Subtle implementation or pre-processing details (e.g. data augmentation is standard, but details differ from paper to paper)
- Highly non-Gaussian distribution of outcomes means ‘average error’ or ‘best error’ can be misleading.

Look for ablation studies where the new elements are turned off one by one, with everything else held constant.

Data Leakage

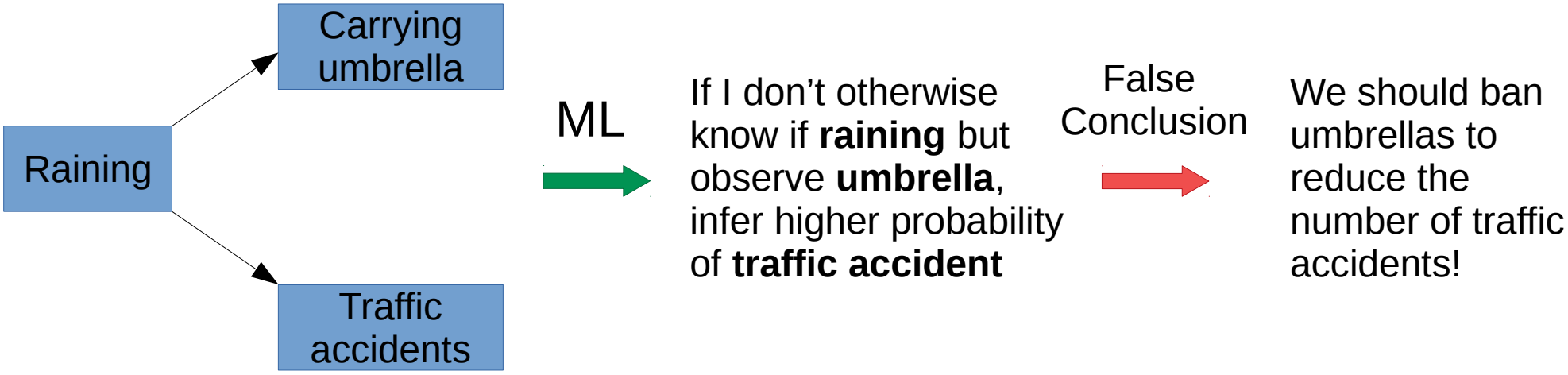
Sometimes a model may do very well because information about the test set has **leaked** into the training set.

Hypothetical case:

- Dataset is EEG readings for 6 patients taken over multiple months.
- Data is chopped into 1 minute long fragments, task is to predict if a seizure will occur within 5 minutes
- 20% of the fragments are reserved for the test set, other 80% for training.
- **Uh-oh...**

Correlation vs Causation

Machine learning models do not inherently extract the causes of relationships between variables, they simply detect that a relationship exists.



Example

Wu and Zhang, “Automated Inference on Criminality using Face Images” (2016)

Purported Goal: Detect if someone committed a crime (criminality)

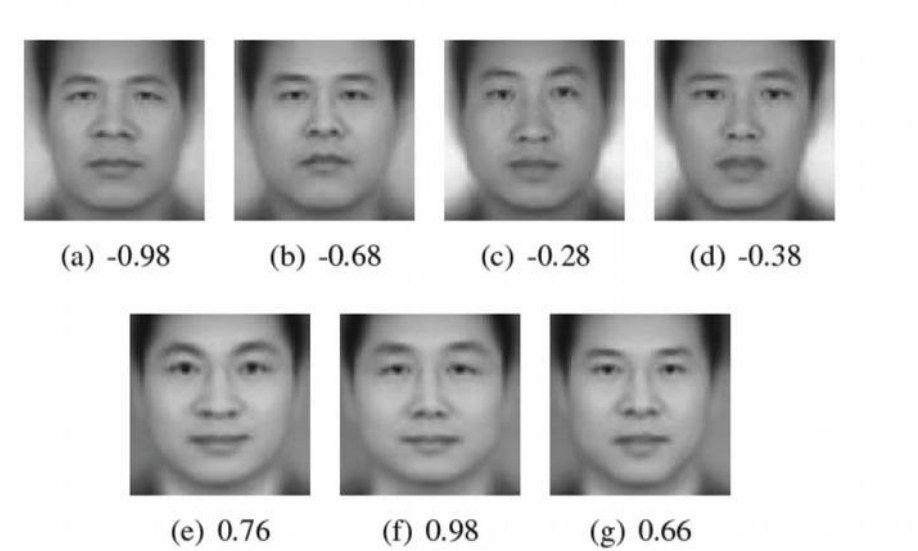
Input: Face images

Labels: ID photos of people vs photos used in wanted posters

Resulting accuracy: ~90% (high!)

What happened?

- Demographic factors unrelated to specific individuals
- Different kinds of photos are used for ID than for wanted posters (smiling, etc)



Accessed from https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html

Correlation vs Causation: Possible?

Can machine learning ever extract causes, not just correlations?

Yes, but it requires a different kind of data: **interventions**

Example:

- Random factor **z** that is known or constructed to have *no prior cause*, then correlations with **z** are evidence of causation by **z**.

Can do more than just this, but in general need constraints on how things influence each-other.

See e.g. Judea Pearl's 'do calculus' formalism