4.2 Evaluating machine learning models: Measuring generalization

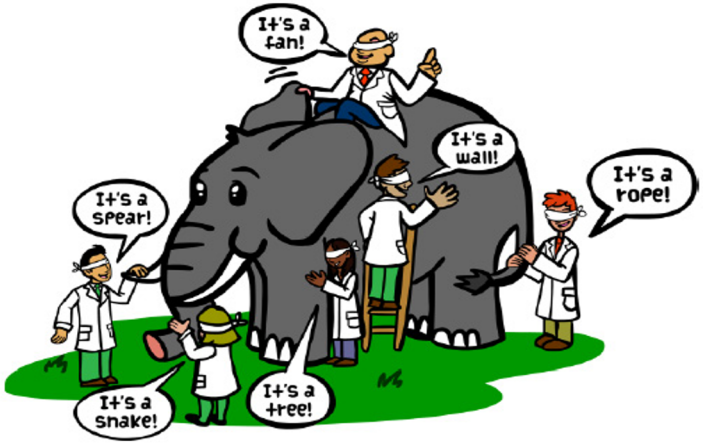
Optimization: It's a "spear + fan + wall + **rope** + tree + snake"

Generalization: It's an elephant (one side is snake like and the other side rope like, etc.)

Evaluating a machine learning model is "measuring generalization"

In ML, the goal is to achieve models that "generalize"

that perform well on never-before-seen data



4.2.2 How to choose an evaluation protocol?

1. Data representativeness
   - You want both your training set and test set to be representative of the data at hand
   - On the Pima Indian Diabetes dataset, what can happen if we don't shuffle the data?
2. The arrow of time
   - If you're trying to predict the future given the past (for example, tomorrow's weather), you should not randomly shuffle your data before splitting it, because doing so will create a temporal leak:
     - Our model will effectively be trained on data from the future
     - Make sure all data in your test set is **posterior** to the data in the training set (you cannot use future data in order to predict past events)
3. Redundancy in data
   - If some data points in your data appear twice (fairly common with real-world data), then shuffling the data and splitting it into a training set and a validation set will result in redundancy between the training and validation sets
   - In effect, you'll be testing on part of your training data!
   - Make sure your training set and validation set are disjoint

Evaluation metrics

Classification

- Accuracy, Precision, Recall, F1-score, AUC
- Baseline accuracy is 50%
- Baseline precision/recall can be 100%

Multi-class classification

- Baseline accuracy depends on the number of classes

Regression

- MAE, MSE, log of MAE
- Scatter/histogram plot the distribution of MAE
- Scatter plot of ‘true label vector’ vs ‘prediction vector’