#ref #ret

%%For **each** of the four scenarios below, answer the following questions. Please explain where your targets/rewards would come from (#2), how you would make your inputs numerical (#3), and a bit of your reasoning on ethical issues (#6). Other questions do not need explanation.

1. What type of machine learning problem (regression, classification, clustering, generation, reinforcement learning) do you think this is?
2. If this is a supervised problem, what will you use as your targets (aka labels)? If it is reinforcement learning, what will you use as your rewards? If it is unsupervised, just say ”unsupervised”.
3. What processing do you need to do to your input data?
4. What type(s) of model(s) would you try? You may need to combine models. Remember to start with the simplest thing that might work! Types of models we’ve talked about are linear regression, decision trees, random forest, logistic regression, naive bayes, support vector machines, K-means, DBSCAN, hierarchical clustering, fully connected neural networks, convolutional neural networks, recurrent neural networks, generative adversarial networks, deep Q learning, and evolutionary methods.
5. What validation metric(s) would you use to decide how well you’re doing?
6. What ethical challenges do the data collection, creation, and/or use of this model create? If you feel there aren’t any, just say ”None”.%%

# 1 | Scenarios:

1. **You want your model to learn to play Frogger.**
   (a) Reinforcement learning
   (b) Points, as given by the game.
   (c) The agent would be passed some input vector with a grid representing the game, perhaps in addition to some extra normalized variables like time left and goals filled.
   (d) Perhaps deep Q learning with an RNN to be able to process different obstacle speeds.
   (e) Regret
   (f) Botting in video games, when the same techniques are applied to multiplayer?
2. **You would like a model to write tweets in the style of a particular author.**
   (a) Generation
   (b) Unsupervised
   (c) Padding, byte pair encoding, tokenization, ect.
   (d) GAN, RNN / LSTM for generator, RNN for discriminator
   (e) Human evaluation
   (f) Could be used to fake tweets or impersonate people which has many larger implications. Data harvesting might also be a problem.
3. *A company would like to be able to predict the next months’ sales for each of its products. You have a dataset that the company has collected for many years, with data for a particular product on a particular month in each row. Each row contains the number of sales for the month, the number of sales from the previous month, the average rating (1-5) of the product in the previous month, the number of reviews in the previous month, the product type (e.g. ”toaster”, ”coffee maker”, ”rice cooker”), its price the previous month, and its price for the current month.*

(a) Regression

(b) The number of sales for the month in the collected dataset

(c) Normalize average rating, OHE product type

(d) FNN

(e) $R^2$

(f) None

4. *You would like to predict the presence of a certain disease using chest x-ray data. You have a lot of x-ray images, a small amount of which have been labeled as having the disease or not having the disease. The rest of the images are unknown as to whether or not the person has the disease.*

   (a) Classification

   (b) Semi-supervised, using the provided labels as well as generated pseudo-labels

   (c) Down-sampling, grayscale?

   (d) CNN

   (e) F-score

   (f) Model explainability challenges, covered in our ethics presentation last semester in ML.

%%### Example:

Here is an example answer, taken from `http://archive.ics.uci.edu/ml/datasets/Abalone`:

Scenario: You want to predict the age of an abalone (a type of shellfish). You have a dataset that includes the age of the abalone, the sex ('M', 'F', and 'I'), the length, the diameter, the height, and the weight.

1. regression
2. age (included in dataset)
3. Length, diameter, height, and weight are numeric. I will scale them. For the sex feature, make it one-hot-encoded.
4. linear regression. maybe a fully connected neural network later
5. $R^2$%%