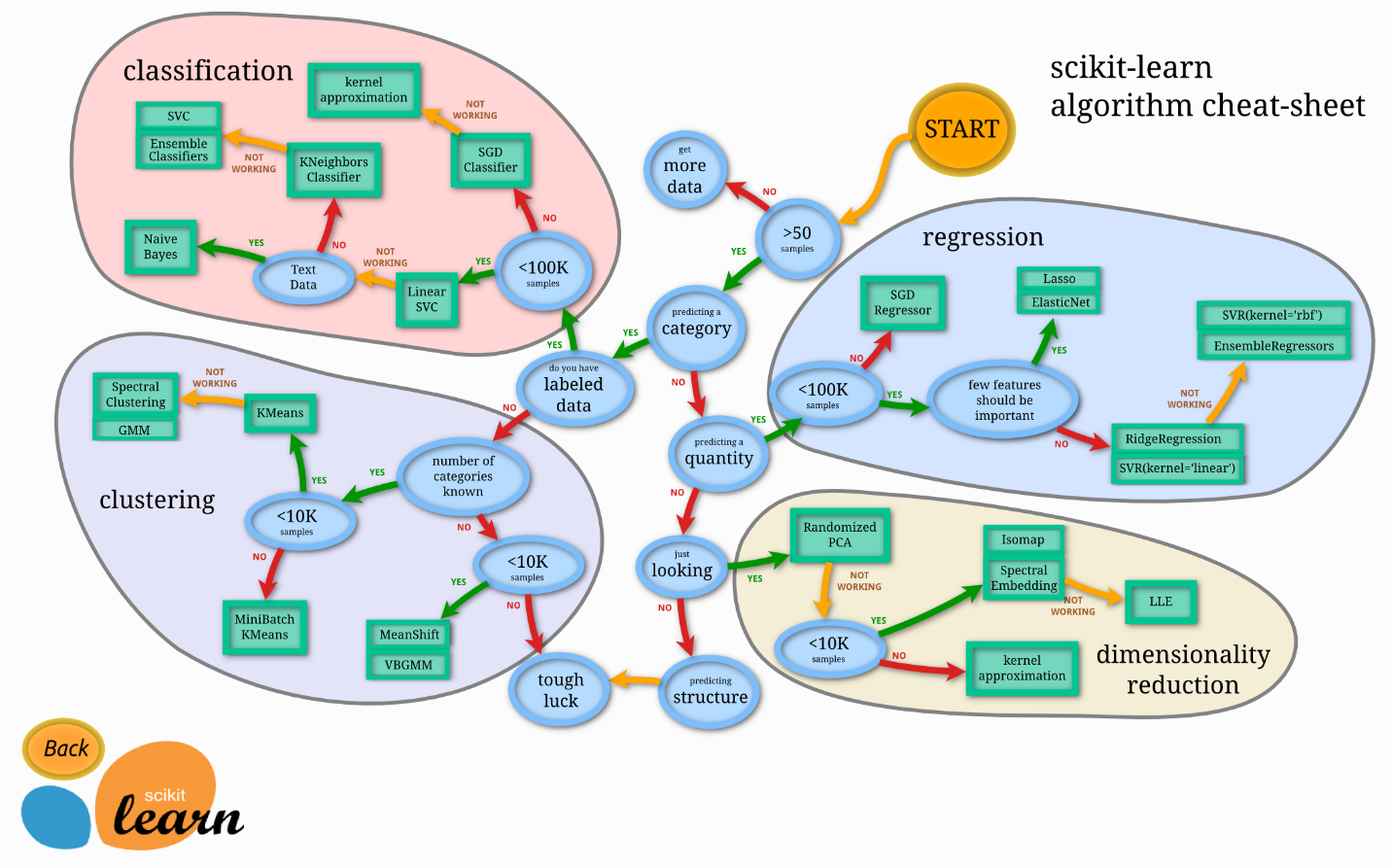
You now have a fairly substantial starting toolbox of supervised learning methods that you can use to tackle a host of exciting problems. To make sure all of these ideas are organized in your mind, please go through the list of problems below. For each, identify which supervised learning method(s) would be best for addressing that particular problem. Explain your reasoning and discuss your answers with your mentor.

1. Predict the running times of prospective Olympic sprinters using data from the last 20 Olympics.
2. You have more features (columns) than rows in your dataset.
3. Identify the most important characteristic predicting likelihood of being jailed before age 20.
4. Implement a filter to “highlight” emails that might be important to the recipient
5. You have 1000+ features.
6. Predict whether someone who adds items to their cart on a website will purchase the items.
7. Your dataset dimensions are 982400 x 500
8. Identify faces in an image.
9. Predict which of three flavors of ice cream will be most popular with boys vs girls.

I used the Scikit-learn cheat-sheet to choice the potential models:

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



1. Predict the running times of prospective Olympic sprinters using data from the last 20 Olympics.

Path: >50 samples -> not predicting categorical -> predicting a quantity -> less than 100k samples (not that many data from last 20 Olympics) -> then we can try:

Lasso (linear regression), ElasticNet (linear regression), Ridge (linear regression), SVR, Ensemble regressors (i.e. Random forest)

1. You have more features (columns) than rows in your dataset.

Feature engineering is more important than choosing a model. i.e. PCA (ref: <https://rodrigo.ai/feature-engineering/2018/2/10/more-features-than-observations>)

1. Identify the most important characteristic predicting likelihood of being jailed before age 20.

It is a classification problem. Also, we need to figure out which feature (most important characteristic) can best explain the outcome. So Gradient boosting model can be an option due to its interpretability to point out which feature is the most important. While black-box model like random forest might not fit the task.

1. Implement a filter to “highlight” emails that might be important to the recipient

It is a classification problem, and it seems require good performance over interpretability. So, random forest model, or SVM classifier can be candidates.

1. You have 1000+ features.

It is a very complex model. Feature engineering is important. Potential model is linear regression/classification model. Despite the simplicity of this algorithm, it works pretty well when you have thousands of features, for example, bag of words or n-gramms in text analysis. More complex algorithms suffer from overfitting many features and not huge datasets, while linear regression provides decent quality. (ref: <https://blog.statsbot.co/machine-learning-algorithms-183cc73197c>)

1. Predict whether someone who adds items to their cart on a website will purchase the items.

It is a binary classification problem. Normally logistic regression is a good candidate for such task.

1. Your dataset dimensions are 982400 x 500

Potentially to use linear regression or linear classifier, since they are useful on huge number of features where more complex algorithms suffer from overfitting.

1. Identify faces in an image.

Neural network (haven’t learnt yet). Since “If you’re working with images, convolutional deep neural networks show the great results. Nonlinearities are represented by convolutional and pooling layers, capable of capturing the characteristic features of images.”

1. Predict which of three flavors of ice cream will be most popular with boys vs girls.

This one is a bit confusing, I am not even sure this is a supervised learning problem.