**Summary:**
This is a computer vision project, which aims to build a classifier to distinguish between control and drug treated samples. Data comes from automated image processing pipeline, which corrects, normalizes and segments collected images to automatically detect and describe structures of interest with 200-1000+ attributes. To date, the dataset holds quantification data from 30 000+ images of both control and experimentally treated samples. The project will involve loading both images and corresponding quantification data into one database and testing various supervised learning algorithms for accurate binary classification. Unsupervised approaches will also be used to gain additional insights as to which features might more strongly contribute to the formation of two clusters and/or to find additional structure in the data. If successful, the project can be built upon by using existing image datasets from molecular treatments affecting other molecular pathways and incorporating established hierarchical ontologies to both name the classes and maintain known links between them. All in all, the project aims to lay a groundwork for applying data science methods for biomedical data to accelerate drug discovery.

**Problem:**
Classification of experimental imaging data is currently done manually. This leads to two problems. First, it introduces researcher-dependent bias in what is called as normal/ perturbed. Second, it is incredibly time consuming and therefore, cost-intensive. Both of these problems could be alleviated by automation of image classification.

**Hypothesis:**
Machine learning classification approaches can be used for accurate and fast classification of experimentally treated samples to allow for binary “perturbed/ unperturbed” classification.

**Data set:**
Original dataset consists of 512x512-pixel RGB images with the signal in Red and Blue channel. They come from multiple independent experiments testing an effect of particular experimental perturbation. These are all run through an automated image processing pipeline, which corrects, normalizes and segments submitted images to automatically detect and describe structures of interest with 200-1000+ attributes. The data set for this proposal consists of multiple folders of two csv files per sample (1 per used RGB channel) with 111 features quantified for R ed channel and 80 for Blue channel. These have to be merged into one database. In addition, the data set still requires adding labels. Before feeding in the data into learning algorithms, I will fill-in missing values with randomly sampled values for respective features. Following that, I will scale and normalize features. In complementary approach, images will be cropped to a padded dimension of the largest object, converted to vectors of intensity values and used as such for classification. I will also experiment with dimensionality reduction approaches for this version of data set (non-negative matrix factorization).

**Algorithms:**

When using features derived from quantification (CSV version of the data set), I will start from using logistic regression. This should help me with feature selection. Following that I will explore SVMs and neural networks (with logistic regression and/or perceptron). I will also explore neural networks for actual image-based learning. I will evaluate performance of different methods using accuracy and confusion matrix (I will also do it when tuning hyper parameters in cross-validation). I will also plot ROC curves for all and derive AUC (trying to find the largest one).

**Statistical approaches:**
Frequentist statistics methods will be applied on the features extracted in the image analysis pipeline assuming Gaussian distribution for continuous explanatory variables (usually true for biological samples), Poisson - for count data. The outcome of classification will be evaluated using Bayesian statistics (Bernoulli binomial).

**Applications:**
The rising costs of developing new medicines make pharmaceutical business close to unsustainable for some companies (~12 billion $ per drug for AstraZeneca)[1]. This is mostly caused by high failure rate in getting FDA approvals for new drugs. This situation is exacerbated by the time it takes for a new drug to hit the market (~12 years in the US) [2]. Thus, any way these two issues can be addressed will have a positive effect on the pharmaceutical business bottom line. One of the reasons so many clinical trials fail is the biased way the pre-clinical data is analyzed (researcher wants to see an effect as only 'significantly different' result can be published). If proven successful, the approach described here will contribute to alleviating both of the described issues. It will remove bias in classifying data and will lead to reductions in time and men cost by shortening the amount of time between experiment (~one day) and result (currently done manually: ~one week).

1. http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/
2. http://www.medicinenet.com/script/main/art.asp?articlekey=9877

**GitHub repository:**
https://github.com/emmaggie/CompVisProject

**Multivariate plot for features with highest standard deviation.**