# Background

## Grants

ExPORTER is a tool provided by the NIH which allows direct access to all funding data and related publications. It is free and open to everyone with a simple to use, if not always responsive, web portal. The files it provides are comma delimited, though they contain a complex quoting system due to the variety of titles and descriptions provided by PIs.

Though ExPORTER provides a large amount of information, funding information is non-existent before 2000, most likely due to institutional policies in the NIH. This causes some issues when trying to look at historical grant data, as funding is a very important part of the overall picture.

The lack of funding data wouldn't be too insurmountable a problem if not for the large amount of missing data both during this pre-2000 period as well as at the end of every year. Lack of budget date information, term information, and especially funding exists near the end of every grant cycle, making data cleanup difficult and time-consuming.

# Method

## Grant clustering on terms

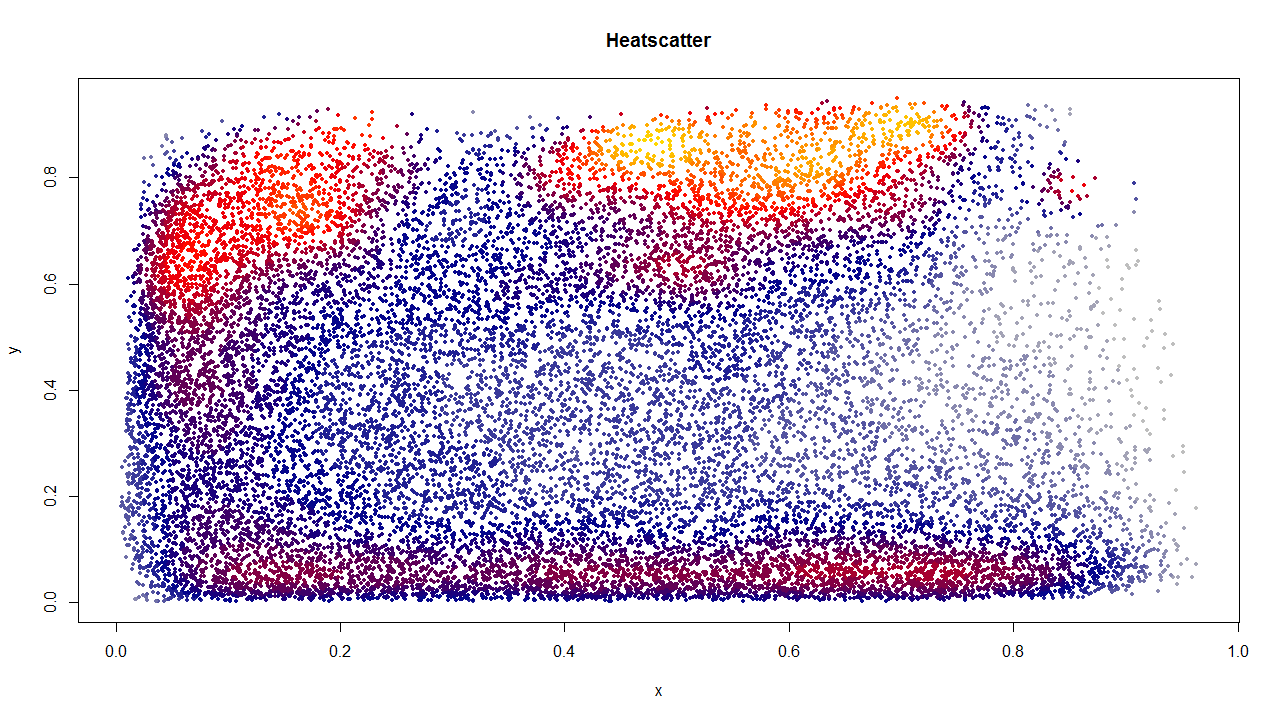
In order to get a better understanding of the terms used to describe the grants we use feature reduction for clustering. In order to do this, we utilize stacked autoencoders as proposed by Hinton[1]. Following Hinton's lead, we build a two layer neural network with a mirrored input and output, with a hidden layer size smaller than that of the input. Minimizing the reconstruction error we are able to build a set of abstracted features which can represent the input with high accuracy. These abstracted features, the activation/output of the hidden layer, can now be used as input in place of the raw input. These autoencoders can then be stacked in order to reduce the feature space to any level the user desires. By reducing the feature space to two dimensions it is possible to plot the grants in order to view possible clusters.

Before we can do this, however, we have to reduce the feature size from 117,000, the number of unique terms contained in the grant terms column. We do this by choosing only those terms which existes in at least 1 in every 100 grants, meaning 20,000 occurrences or more. This reduces our features from 117,000 to 438. After this it is just a matter of reducing the feature space from 438 to 2. Staying with the concept of a deep network, we reduce the feature space from 438 to 200, to 100, to 50, then finally 25 and 2.

# Results

## Grant clustering on terms

We were able to successfully reduce the feature space from 438 to 2 by using the methods described previously: 438-200-100-50-25-2. By doing this we were able to successfully reduce the dimensionality of the grant terms, visualizing it below:



This was only possible by using a distribution of 1/50th of the data, however, due to inconsistencies in the data as mentioned before. Multiple sets of data were tried with varying results, with 1/50th being the best. More analysis of the data would lead to a better understanding of why this could be the case, anything from a poor distribution of terms outside of the first 1/50th to the issues stated previously about missing and incomplete data depending on how late in the year the data is collected from.

# Conclusions

When it comes to terms describing the data, it is apparent that the 438 chosen are not entirely sufficient in order to do accurate clustering on the entire dataset. Expansion of the terms used may be able to counteract this issue, as would filtering of the grants before clustering.

Along those lines, more preprocessing of the data could be performed as we currently use all data available with no filtering or normalization. This should yield higher quality results and better insight into the grant process as a whole.

# Future work

One possible future analysis of the data is looking at term importance over time, allowing us to see trends in terms which have received funding. This would be the first move toward the prediction of both term popularity as well as grant numbers and amounts based on the terms used. This information can be useful for both helping direct future grants as well as audit grants and grant committees, which is important when it comes to selecting committees as well as taking the temperature of scientific progress as a whole.

# Citation

1. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786 (2006): 504-507.