# Text Retrieval for biomimicry function identification in a corpus of biology papers

Brandon Ruffridge – br24 (Captain)

Christian Ortiz - caortiz3

Jay Kim – jk73

Identify biology papers that describe various biomimicry functions from a known list of 100 functions and rank them in order of relevance to a particular biomimicry function. An example would be a paper describing how geckos feet enables them to cling to surfaces, should be given a high relevancy rank for the biomimicry function, “attach temporarily”.

This open-source information retrieval and ranking system is intended to be used by designers and engineers who seek nature’s solutions to their design and engineering problems, as well as by biologists who seek to extend the application of their scientific discoveries. The goal is to integrate it into the [open-source](https://github.com/nasa-petal) PeTaL [project](https://www1.grc.nasa.gov/research-and-engineering/vine/petal/).

While the details of the project tasks will become clearer following the tech review of various approaches, at a high level the project will involve the following.

Python will be the primary programming language used for this project.

Gather Data (5 hours): We currently have about 1200 biology papers that have been labelled by biomimicry experts with the biomimicry functions described by the paper. We also have access through Microsoft Academic Graph (MAG) API to ~13 million biology papers. Paper metadata that may be useful for our task include title, abstract, authors, references, venue, and MAG topics.

Preprocess (20 hours): Data preprocessing may include removing stopwords, stemming, lemmatization, and removing papers with short or empty abstracts.

Train (40 hours): We will need to train model parameters to assign relevancy scores to papers for each topic in our set of 100 biomimicry topics. Models we may explore include retrieval models, topic models, and text classification models.

Rank (20 hours): We will run the trained model on new biology papers to produce relevancy scores for each topic.

Evaluate (20 hours): Due to the lack of relevancy judgements we will have to rely primarily on human evaluation of the ranked lists for each topic. We may be able to gain some insight into how well our ranking model performs by generating nDCG scores comparing the model’s ranked list to the ideal ranked list produced by the 1200 papers that have known labels.

Feedback (20 hours): If there is enough time, we will also implement an implicit feedback mechanism based on clickthroughs to improve the ranking algorithm.