Appendix A The Project Proposal form

CI301 The Individual Project

LEVEL THREE COMPUTING PROJECT PROPOSAL 2015/2016

The deadline for submitting this form is Thursday, October 22, 2015. If you cannot meet this deadline for an acceptable reason you should ask your award leader for an extension. Your proposal will be marked on a Pass/Fail basis and failure to hand it in by the deadline will result in a Fail.

Name Ibraheem Ajibola Ganiyu.....

Title of project

Which course are you on? Please tick the one that applies:

☐ BA (Hons) Business Information Systems

☐ BSc (Hons) Business Computer Systems

☐ BSc (Hons) Computer Science

☐ BSc (Hons) Computer Science (Games)

☐ BSc (Hons) Digital Media Development

☐ BSc (Hons) Internet Computing (UCH)

- BSc (Hons) Digital Games development (UCH)

☐ BA (Hons) Digital Media

☑ BSc (Hons) Software Engineering

☐ BSc Computing

Are you:

☑ Full-time

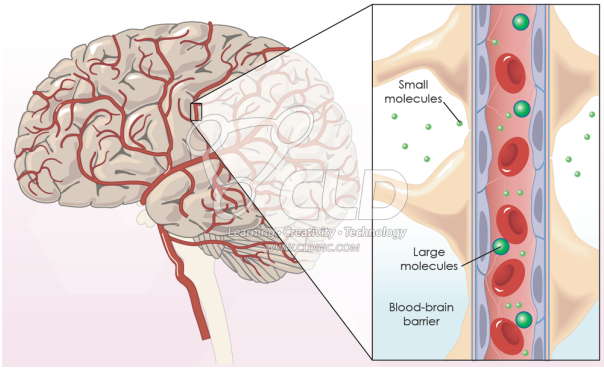
☐ Part-time

Your Proposed Project Machine Learning in Drug discovery and design: Ensemble techniques for predicting blood brain barrier penetration of drugs

Please outline what you intend to do and indicate the main stages or elements in the project. Insert another sheet if necessary.

Problem Statement:

The blood brain barrier (BBB) is a membrane that separates blood flowing through the body from the brain extracellular fluid. It also serves to protect the brain from foreign substances amongst many other functions. It is very selective about what can pass through and this is a problem when designing drugs especially for the Central Nervous systems as the drug needs to be able to pass through this membrane to bind with its target receptor(William M. Pardridge, 2005)



In Silico models i.e computer modelling in virtual screening, especially machine learning techniques have had a found profound use case in drug discovery in the recent years as they can drastically reduce the time-consuming process of searching through a library of drugs, as a typical molecular database usually contain about molecules ranging from thousands to millions

Drug discovery and design is a very expensive process, which also opens up opportunities for many creative solutions to help reduce this costly process. One way to reduce the cost of drug discovery is pouring resources into tasks that would yield the most returns. In reference to the blood brain barrier problem, If it is possible to know with confidence the probability of a drug passing through this barrier, resources can be spent developing these drugs as opposed to developing an entire library of drugs and realising at a later stage that they cannot pass through the barrier.

Project Proposal:

This project aims to produce a piece of software powered by machine learning techniques to predict the probability of a drug-like molecule passing through the blood brain barrier.

A similar solution already exists online at http://b3pp.lasige.di.fc.ul.pt , however, having worked with one of the researchers on this project for over a year, I have come to realise that there is still room for improvements on this project in terms of efficiency and even exploring other possible approaches that were not explored in the accompanying publication "A Bayesian Approach to *in Silico* Blood-Brain Barrier Penetration Modelling" and also distributed computing techniques to improve the prediction engine.

Main Stages and deliverables:

- Predict blood brain barrier permeability using bayesian techniques on select Machine learning classifiers (Random Forests, SVMs, Neural networks etc) ~ November, 2016
- Apply Ensemble Techniques on these classifiers running on different Apache Spark clusters to improve training efficiency and using ensemble techniques to improve prediction results ~ January, 2017

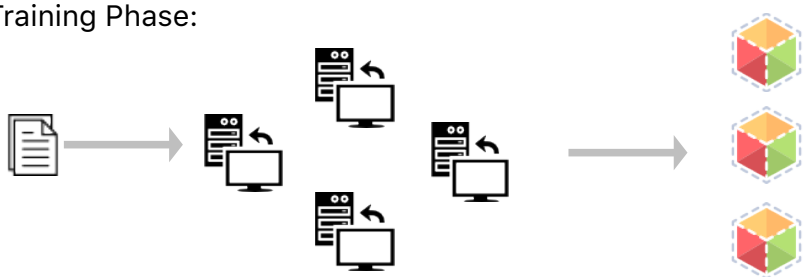
Extensions (Heavily Research based):

- Explore Deep learning techniques as these have shown promising results in virtual screening ~ March, 2017
- Prediction optimisations
  - Feature engineering on the current datasets
  - Techniques for handling imbalanced datasets
  - Improvements to the Random Forests classifier
  - Improvements to the training technique for Support Vector Machines
  - Effects of Boosting the Machine learning classifiers on overall predictions

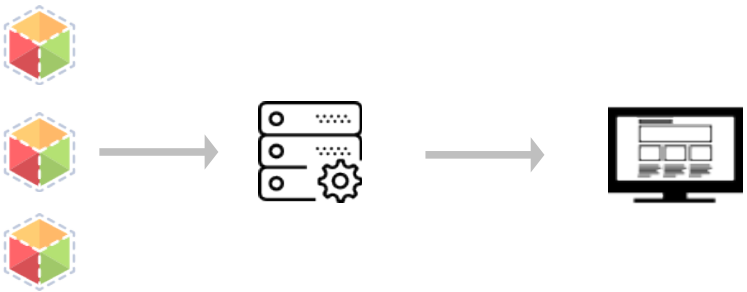
Architecture and Environment:
The prediction engine would primarily be written in the Python programming language, and would represent roughly 80% of the code base. The entire product will be based on a client-server architecture, with the prediction engine residing on a server and a client facing web app to query the prediction engine. During the training phase of the prediction method, Apache Spark would serve as our distributed computing framework, to speed to up the training of the classifiers.
The following libraries will be used

- AngularJS for the client web app
- NodeJS on the server side
- Python + Scikitlearn for the prediction engine
- Apache Spark for managing the clusters

Training Phase:



The training process can take a long time, and training numerous models also can take even longer, therefore, Apache spark would be used to reduce the time taken to train the models .
Production Phase:



The trained models are loaded into a python script on the server and passing the results to a NodeJS server which then passes the results to the client via HTTP/REST.

Implementation Issues:

- Domain knowledge about Chemical molecules
- Getting a good enough prediction is ok but Increasing the accuracy of the prediction engine will be another implementation issue
- Communicating between the Python process and the NodeJS server via unix sockets as the prediction scripts are written in python and the Node server needs to carry the prediction results to the client
- Cleaning and preprocessing the dataset to be used for training the models as some SMILE formats are not being parsed properly

Conclusion:
The BBB prediction problem is a very interesting problem with real life impact. The choice of this project was heavily inspired by my placement year at AstraZeneca, a biopharmaceutical company. I hope to produce software that can predict with high confidence the probability of a drug passing through the blood brain barrier.

References

1. A Bayesian Approach to *in Silico* Blood-Brain Barrier Penetration Modelling. Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao.
Journal of Chemical Information and Modeling 2012 52 (6), 1686-1697
DOI: 10.1021/ci300124c
2. The Blood Brain Barrier: Bottleneck in Brain Drug Development. William M. Pardridge, NeuroRx: The Journal of the American Society for Experimental NeuroTherapeutics, 2005.

The member of staff who has agreed to supervise this project you must have discussed this with the member of staff

Your supervisor's name (please print) ..... Aidan Delaney .....

If there are any major hardware/software resources required from the University, have you checked that they will be available for your use?

☐ Yes

☐ No

☑ Not applicable

If the project is for an outside client or organisation, we need a letter saying they approve of your proposed project and will support it as far as is necessary:

☐ Letter attached

☐ Awaiting letter

☑ Not applicable

Date ..... Friday, 7th October, 2016 .....

From time to time we may contact you by email to send you updated information and reminders about the project. Please check your University mailbox regularly and make sure that it does not go over the limit. You can forward your university email to your preferred email address by changing your personal settings on studentcentral. You should also keep an eye on the CI301 space on studentcentral for updates and announcements.

Jane Challenger Gillitt, Project Coordinator