February 3rd,2016

Predicting Trends in Scientific Research

Danish Khan
[redacted]
SFWR ENG 2XB3-L04

1. Abstract

The following project proposes to analyze the NSERC grant database to find a dependable model that maps the number of grants awarded, size of grants and various other parameters to a field of research over time. The program is intended to produce a time-dependent graph of this data based on a given institute and will output future estimates of the size of a research field, amount and number of expected of grants awarded to that institute.

2. Objectives and scope of the project

Description, context and history of the problem.
According to the report, *Climate Change and Infrastructure, Urban Systems and Vulnerabilities*, released by the US. Department of Energy in February of 2012, Universities like Yale, Harvard, MIT, Stanford, Waterloo, and York are building remote modules almost every decade to meet the technological needs of current and future research projects. Furthermore, infrastructure within these Universities is quickly becoming obsolete as more and more researchers are finding that their current technological resources are becoming “museum-ware” too soon, in their research endeavours, especially in engineering and science. Moreover, the coupling of future research projects, software engineering and the computational power of supercomputer is now greater than ever and only increasing. Thus, the cost of maintaining, renovating and reallocating the current infrastructural, human and institutional resources is now higher than ever. With both, the governments of Canada and the US., two of the largest administrations in North America funding scientific research, swimming in massive debt, it is now more important than ever to consider the problem of finding a precision in resource allocation and curb the “innovation gap” that occurs between funding and scientific research.

Assumptions and background study.
The problem of resource allocation is essentially a problem of future prediction. Finding arguments for formulating investment policies for the research technologies of today requires a precise and prescient knowledge of where research is expected to be tomorrow. According to the Department of Industry and Science, the most important influential parameter the affects the direction in which science leads is perhaps unquestionably industry. Building upon this assumption, it is essential then, that one examines the behaviour of this relationship, between research and industry, to predict the future state of research.

Proposed solution.
The largest institution in Canada, one that was built upon this industry and scientific research relationship, is the NSERC (National Science and Engineering Research Council of Canada). This solution proposes to examine the grant and research approval data of the NSERC from 1992 to the present. The paramount objective of this project will be to find reliable projections into the future of the behaviour of scientific research in it's various fields, and nuance it's construction and development into the future in hopes of allowing the policy makers of today to better formulate their policies for investment of tomorrow.

Goals, tasks involved and motivation.
The goal of this project will be to find a model that describes the relationship between research grant awards and the specific field of research to which that grant was allocated, to find a relationship between the number of grants awarded in a particular field of study over time, and how a particular field of research total funding overtime changes as a function of grants and time. The motivation for this project is to reduce the amount of financial waste that occurs as a result of misallocation of resources in the field scientific research and development.

3. Input & Output

Expected input.
The NSERC grant awards database contains the following information:

CLE	Unique identifier of current grant entry.
Name	Full name of researcher.
Department	Department to which the researcher belongs.
Organization ID	Researcher's organization number.
Institution	Researcher's organization name
Province	Researcher's current provincial location.
Country	Researcher's current country.
Fiscal Year	Year in which the grant was awarded.
Competition Year	Competition year in which grant was awarded.
Award Amount	Amount of grant in CAD.
Program ID	Researcher's current program unique identifier.
Program Name	Researcher's current program.
Group	Researcher's current grouping.
Committee Code	Researcher's current committee code.
Committee Name	Researcher's current committee name.
Area of Application Code	Unique identifier of area of application.
Area of Application Group	Name of area of application group.
Area of Application	Name of area of application.
Research Subject Code	Unique identifier for research subject.
Research Subject Group	Name of group of research subject.
Research Subject	Name of research subject.
Instalment Number	Instalment amount.
Application Title	Title of research on application.
Keyword	Used for source data.
Application Summary	If available.

Expected output.
This database will serve as the input for our algorithm. The planned output for this program will consists of a graph connecting each participating University in the database to all the available fields of research in the database over annual time intervals. This would rank each University by the type of research most associated and conducted within it, allowing policy makers to see each University's development in research over time. The second output will be contain each universities grant amount changing over time in the various fields of research adjusted for inflation. The final output would be a total grant amount in each field of research listed over time and adjusted for inflation. This output will also show the program's expected projection into the next five years.

4. Algorithmic challenges of this project.

Data set challenges.
The greatest challenge in this project will be sort, analyze and graph an average 25,000 points of data for each of the 23 years of data. In total, this is approximately 575,000 entries. Another challenge will be generating models from the data that reliably predict the state of an area of research in the future.

Quick sort and Linear Regression algorithm.
The data set in the condition provided by the government is relatively random with regards to any particular category of grant. Thus, it expected that the quick sort algorithm will perform most efficiently and will be employed in the first iteration of the program. To render and analyze the future projections, it is likely that this project will employ a linear regression algorithm to predict a future for a particular field of research.

5. Project Timetable

Time-line.

Phase Name	Tasks & Objectives	Expected time for completion
1. Data Extraction & Compilation	Build an extraction pipeline to capture and compile all of the relevant information from the entries found within the proposed NSERC grant database and compile as necessary.	(1 Week) February 22, 2016 - February 28th, 2016
2. Data Organization	Sort the data by institution, field of research, researcher etc. And remove any fault entries that may appear.	(1 Week) February 29h, 2016 - March 6th, 2016
3. Preliminary Research	Plot data, build models of representation, use statistics libraries to find trends in the data. Conclude preliminary research. Design and complete a working demo.	(2 Weeks) March 7th, 2016 - March 21th, 2016

4. Collaborate with Statistics department, Faculty of Engineering & Professor Karakostas for Algorithm refinement support.	Discuss preliminary findings with faculty, members of the Mathematics department of Statistics. Ventilate ideas, permeate opinions, etc. Refine demo.	(2 Weeks) March 22nd, 2016 - April 3th, 2016
5. Refine analytic code. Complete model.	Test, refine and complete the working demo and publish for a web-based UI interface.	(6 days) April 4th, 2016 - April 10th, 2016
6. Render findings on UI.	Complete the UI design.	(2 days) April 11th /12th, 2016
7. Submit project	Submit project for critique.	April 12th, 10:00 pm.

References

Wilbanks, T., & Fernandez, S. (n.d.). *Climate Change and Infrastructure, Urban Systems and Vulnerabilities. Technical Report to the US. Department of Energy in Support of the National Climate Assessment* (pp. 7-54, Rep.). Oak Ridge: US. Department of Energy Office of Science

Review of premises and standards. (n.d.). Department of Industry, Innovation and Science. Retrieved February 10, 2016, from http://www.industry.gov.au/industry/IndustrySectors/buildingmetalsandconstruction/Pages/PremisesStandardsReview.aspx