IITB-Monash Research Academy

Indian Institute of Technology Bombay

MONASH University

An Indian-Australian research partnership

Project Title:	**Big Data Mining**	
Project Number	**IMURA0395**	
Monash Supervisor(s)	Prof. Kai Ming Ting Prof. Geoff Webb Prof. Kevin Korb	*Full names and titles*
Monash Primary Contact:	KaiMing.Ting@monash.edu	*Email, phone*
Monash Head of Department:	Graham Farr	*Full name, email*
Monash Department:	Clayton School of IT	*Full name*
Monash ADRT:	Kai Ming Ting	*Full name, email*
IITB Supervisor(s)	Prof. Ganesh Ramakrishnan	*Full names and titles*
IITB Primary Contact:	ganesh@cse.iitb.ac.in	*Email, phone*
IITB Head of Department:	S Sudarshan	*Name, Email,*
IITB Department:	Computer Science and Engineering	*Full name*

Research Academy Themes:

Highlight which of the Academy's Theme(s) this project will address?
(*Feel free to nominate more than one. For more information, see www.iitbmonash.org)*

1. Advanced computational engineering, simulation and manufacture
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research

The research problem

Define the problem

The amount of data, published and stored in databases and the internet, is growing at an ever increasing rate. This phenomenon is now widely called 'big data'. Smart use of this data requires efficient and effective automatic methods, commonly called data mining, for the timely extraction of previously unknown, valid, and actionable information. Mining big data is a now big business, so the ability to deal faster with orders of magnitudes more data is potentially worth billions of dollars.

The existing data mining paradigm has two constraining weaknesses: (a) it relies on density

estimation as their core modelling mechanism which is computationally expensive in terms of runtime and memory space requirements; (b) it must use some form of similarity metric to compute the similarity between any two instances. As such, density-based algorithms cannot be applied to (i) big data, including data streams which potentially have infinite data, and (ii) high dimensional problems. Existing research has focused on reducing the runtime and memory space requirements from quadratic in the input data size to near linear; and that is the limit it can achieve within the existing paradigm.

The aim of this project is to enable big data mining by overcoming the underlying constraining weaknesses of current data mining approaches.

The expected outcomes of this project includes

- **Create a new paradigm that enables big data mining**
- **Build new algorithms to solve big data problems in predictive modelling, clustering, anomaly detection, information retrieval and other data mining tasks.**
- **Provide insights into new approaches to data modelling that is superior to existing approaches in terms of task-specific performance measure, time and space requirements.**
- **Determine the relative strengths and weaknesses of existing approaches and new approaches, specifically in relation to the curse of dimensionality and big data.**

Project aims

Define the aims of the project

Expected outcomes

Highlight the expected outcomes of the project

How will the project address the Goals of the above Themes?

Describe how the project will address the goals of one or more of the 6 Themes listed above.

Capabilities and Degrees Required

List the ideal set of capabilities that a student should have for this project. Feel free to be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.

Proficiency in programming in either C, Java, or Matlab.
Have background in algorithmic methodology in data mining or machine learning (but not a must)