La Trobe Business School

Taking innovative machine learning research to industry

Damminda Alahakoon

SAS Analytics innovation Lab
Research Centre for Data Analytics and Cognition
La Trobe University

LA TROBE UNIVERSITY

La Trobe Business School

Presenter Bio

- *Damminda Alahakoon*
- Professor and Director, Research Centre for Data Analytics and Cognition
- La Trobe University, Melbourne, Australia

- Over 10 years' experience in the IT and finance industries, as a credit officer and Data Mining Specialist, in Sri Lanka, Australia and the Netherlands.

- 12 years as an academic

- Key research expertise in the areas of Data Mining, Predictive Analytics, Text Analytics, Machine Learning and Business Intelligence.

- Published over 100 research articles in data mining, clustering, neural networks, machine learning and cognitive systems.

LA TROBE UNIVERSITY

La Trobe Business School

Melbourne, Australia

2017 Most liveable cities in the world

1. Melbourne, Australia
2. Vienna, Austria
3. Vancouver, Canada
4. Toronto, Canada
5. Calgary, Canada
6. Adelaide, Australia
7. Perth, Australia
8. Auckland, New Zealand
9. Helsinki, Finland
10. Hamburg, Germany

LA TROBE UNIVERSITY

La Trobe Business School

La Trobe University

Established in 1964
36,000 students
1,500 academic and
1,700 admin staff

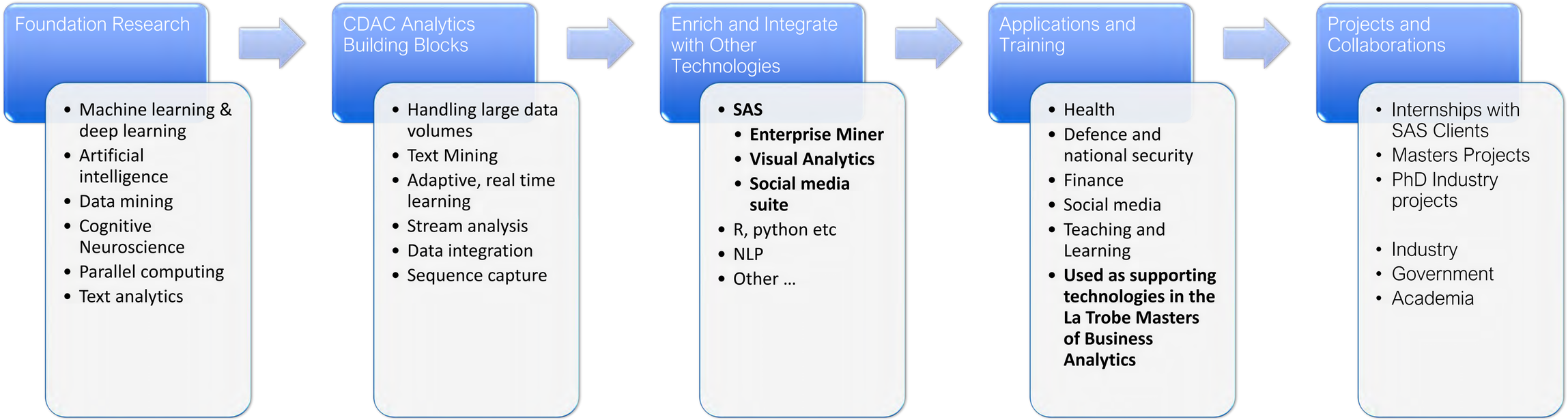
Named after the first lieutenant-governor of the state of Victoria

Faculties: Engineering, Computing, Business, Law, Health, Arts and Social Sciences

LA TROBE UNIVERSITY

La Trobe Business School

Research Centre for Data Analytics and Cognition
SAS Analytics Innovation Lab



LA TROBE UNIVERSITY

La Trobe Business School

La Trobe Masters of Business Analytics

Foundation	**BUS5SMM** Sustainable Marketing and Management	**BUS5IAF** Introduction to Accounting and Finance	**BUS5BIM** Business information management	**BUS5SBF** Statistics for business and finance
Core analytics	**BUS5PB** Principles of business analytics	**BUS5PA** Predictive analytics	**BUS5VA** Visual analytics	**BUS5WB** Data warehousing and big data
Advanced analytics	**BUS5CA** Customer analytics	**BUS5AP** Analytics in practice	**CSE5DWD** Data warehouse concepts and design	**CSE4DSS** Decision support systems
Specialisation	4x electives, can be used to complete a specialisation, e.g., marketing analytics, sports analytics, data science, etc.			

SAS as the core technology

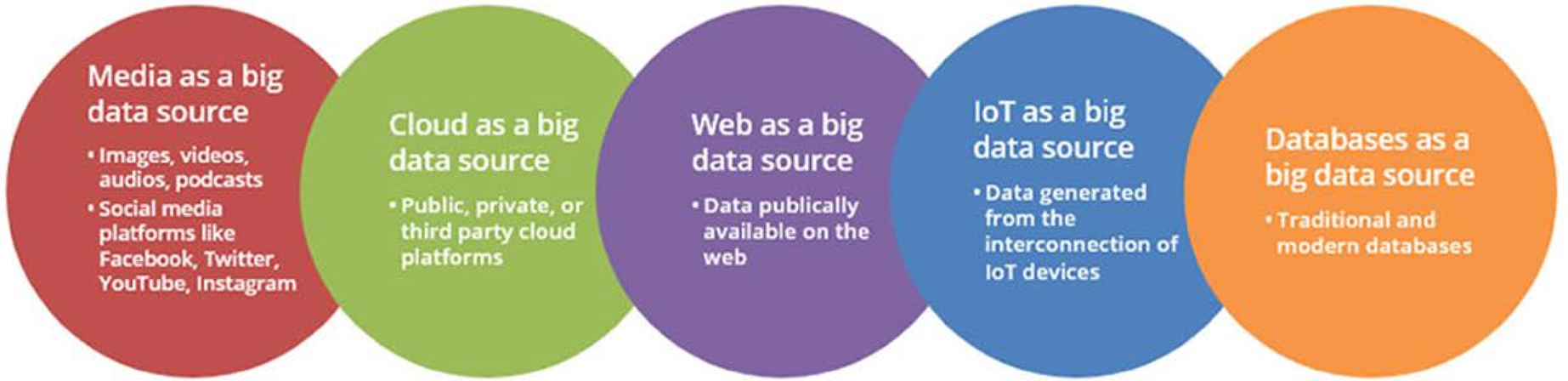
- *Guest lectures by SAS consultants*
- *Joint certification with SAS*
- *Internship and placement opportunities through SAS*

SAS voucher system (unique) – where students can access SAS industry training up to three years after graduation at highly discounted prices.

LA TROBE UNIVERSITY

La Trobe Business School

Data and Information in the Age of Big Data



https://www.allerin.com/blog/top-5-sources-of-big-data

- Volume
- Volatility
- Variety
- V…
- V…

Text, images, video..	• Semantics, meaning culture etc • Social media – sentiment, emotions, human behaviours and traits ..
Machine and process generated	• Can relate to behaviours undefined by humans
High frequency of generation	• Could be hundreds of readings per second (or more) • Patterns can exist in different levels of abstraction/aggregation

LA TROBE UNIVERSITY

La Trobe Business School

Difficulties for Machine Learning

What is data ? What does data represent ?

Traditionally

Data as a general concept refers to the fact that some existing information or knowledge is *represented* or *coded* in some form suitable for better usage or processing - Wikipedia

Can be labelled
Even when unstructured can transform in to structured form

Suitable for Machine Learning

Big Data (New environment)

Text, images, video..

Machine and process generated

Machine and process generated

May not relate to actual (known) objects or events
May represent emotions, behaviours without clear classifications
May be highly granular data points which require aggregation to be ‘meaningful’

Difficult or impossible to label
Unstructured data may contain emotions and individuality which are difficult to capture

Not ideal for traditional Machine Learning

LA TROBE UNIVERSITY

La Trobe Business School

Our Approach in New Machine Learning Algorithms

To address problems due to: Unlabelled, unstructured and highly granular data

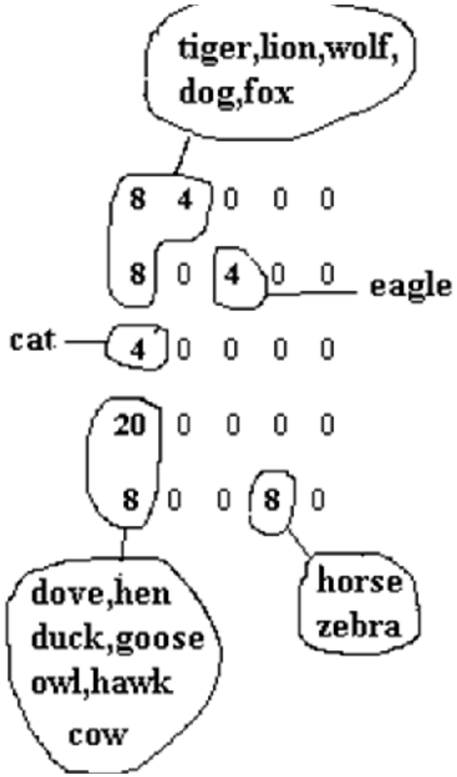
Unsupervised machine learning	Self structuring algorithms	Situation capture from unstructured data
• Self organising maps	• Models which adapt structure and evolve to represent data	• Enrich traditional ‘data’ • Cognitive psychology • Emotion • Behavioural traits

Continuous monitoring and incremental learning algorithms

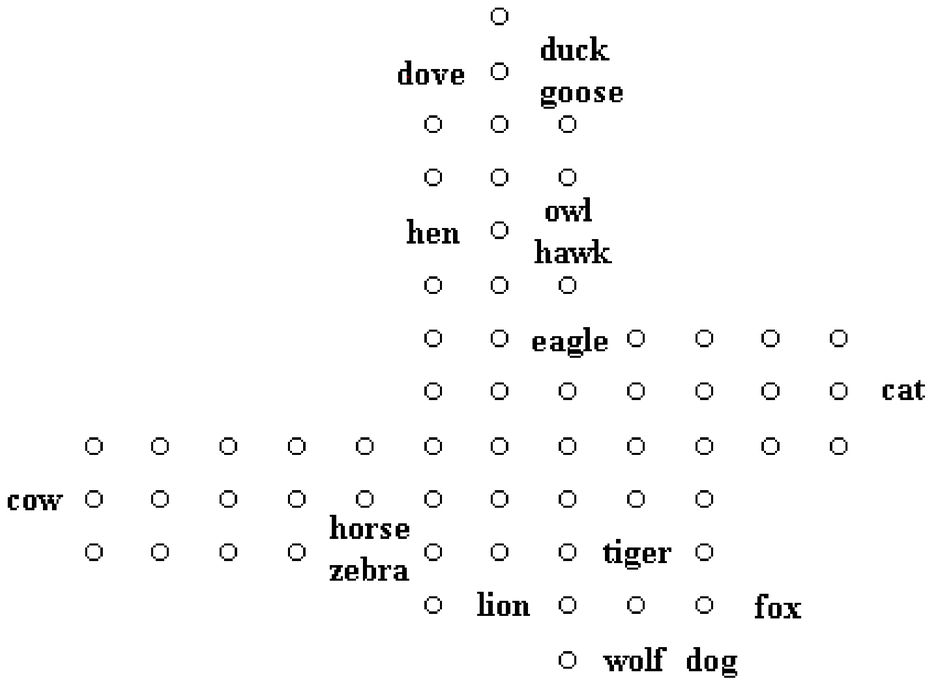
LA TROBE UNIVERSITY

1. Self structuring and unsupervised learning

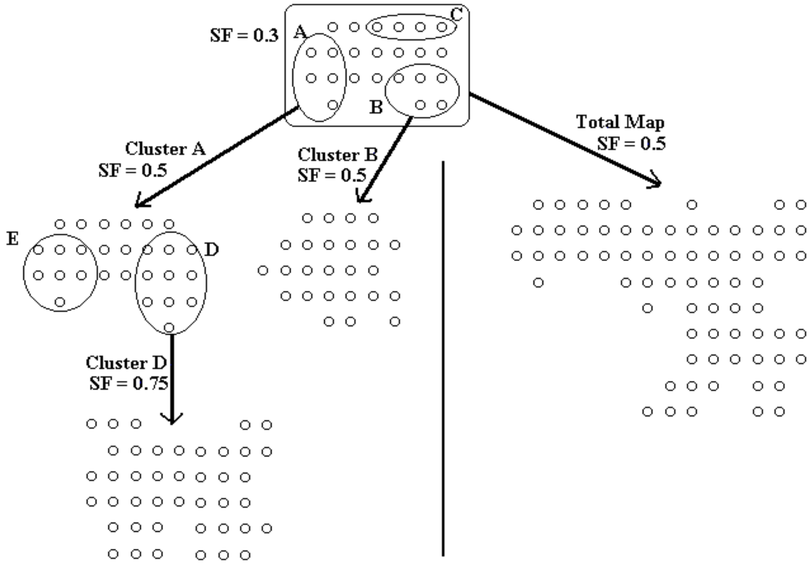
Growing SOM Algorithm



Self Organizing Map (SOM) – fixed structure



Growing SOM – self structuring

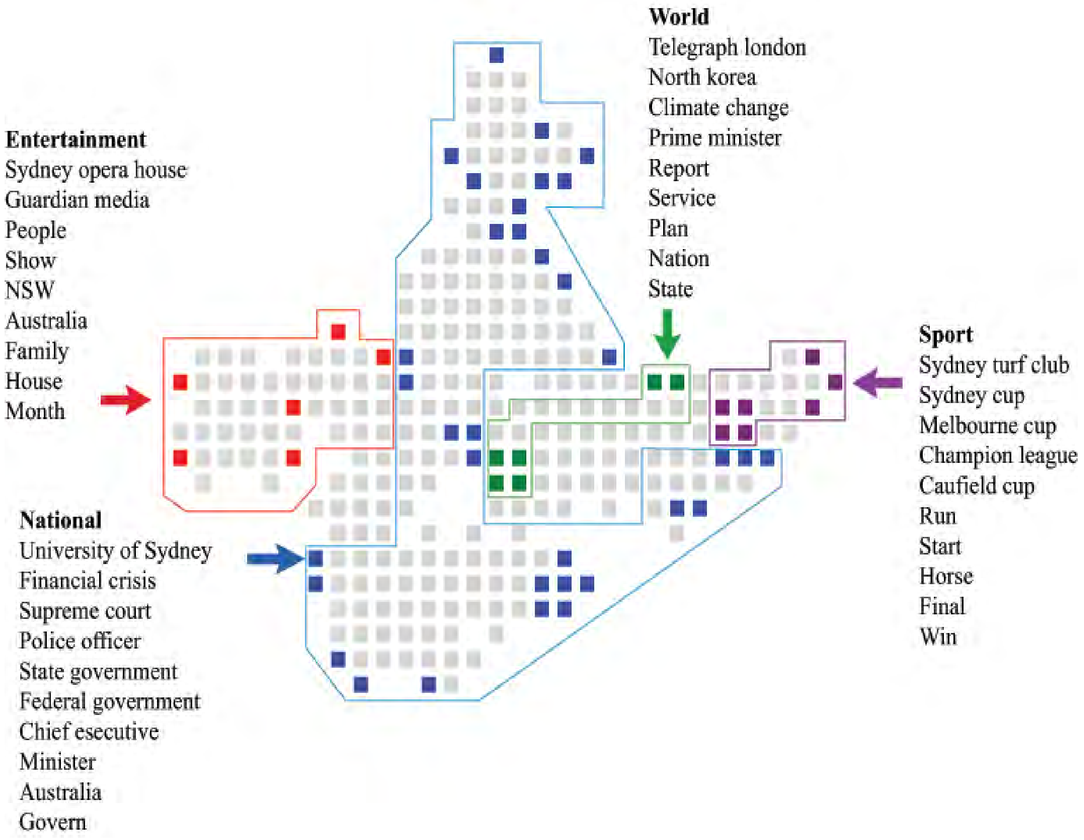


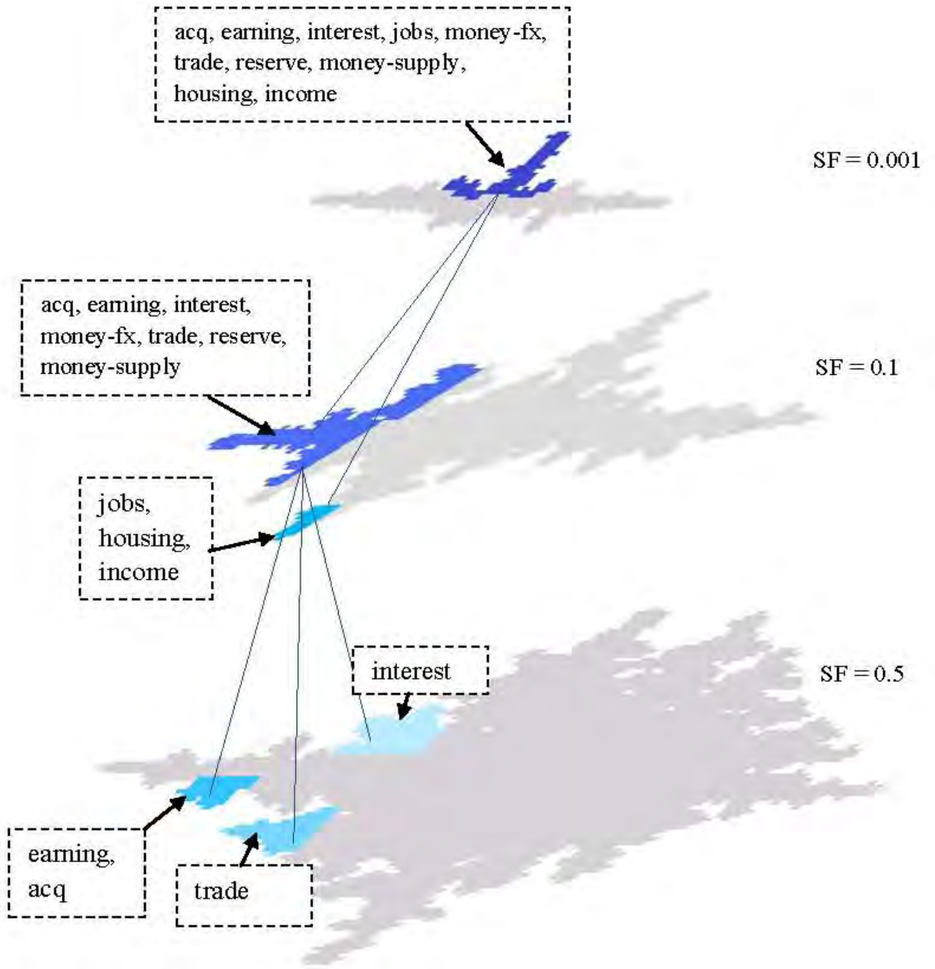
Growing SOM – Generating cluster hierarchies

La Trobe Business School

1. Self structuring and unsupervised learning

Growing SOM Algorithm

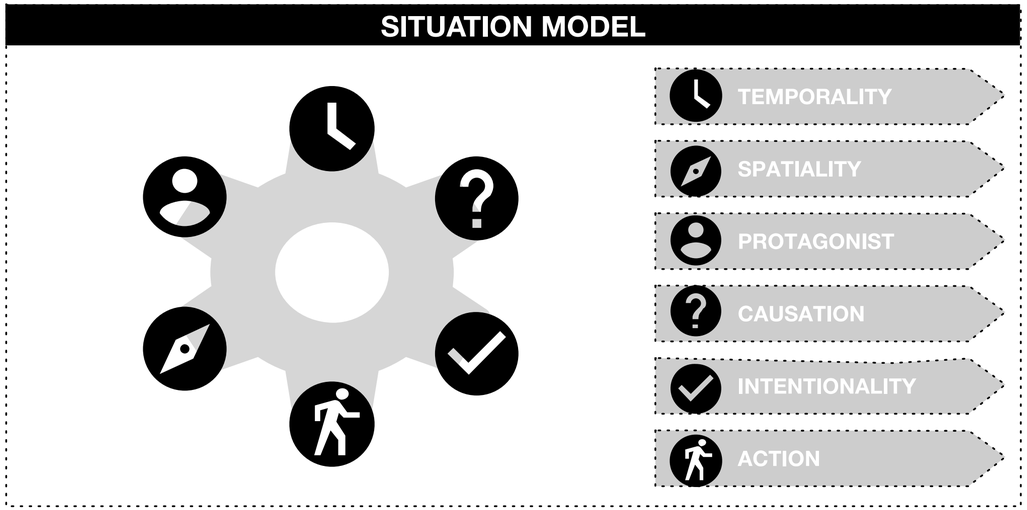


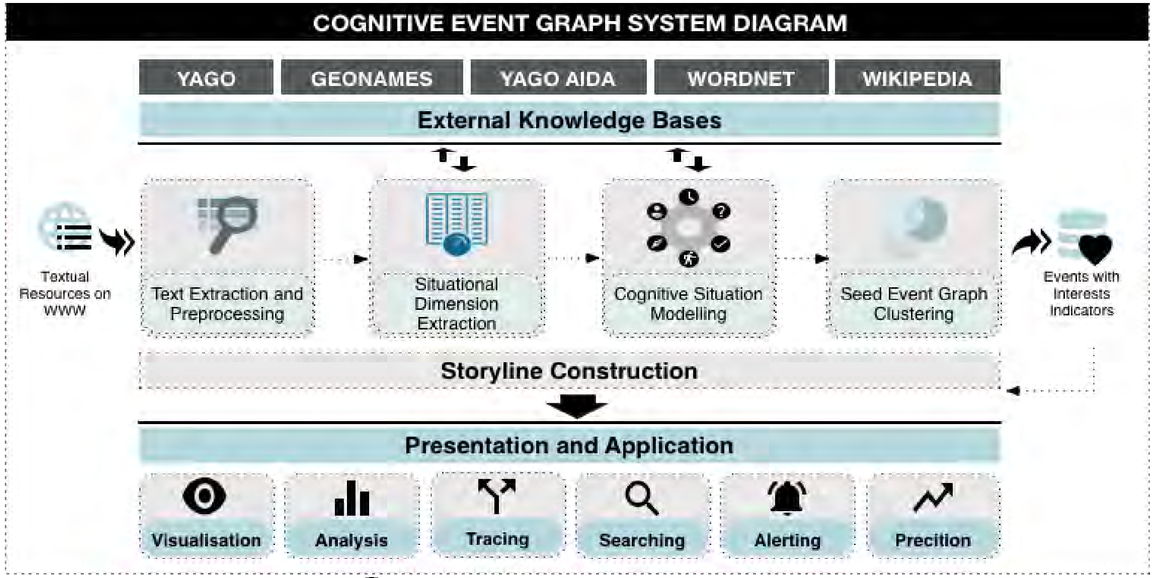


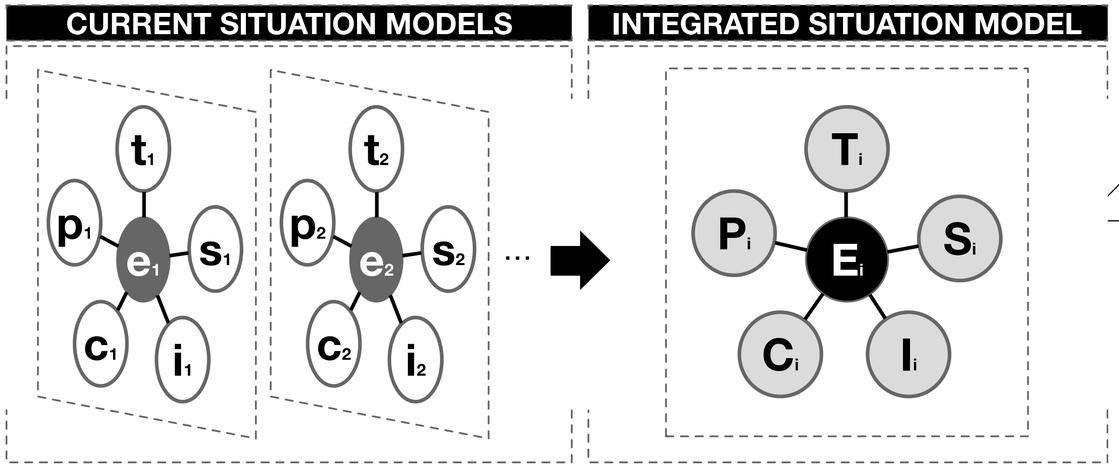
LA TROBE UNIVERSITY

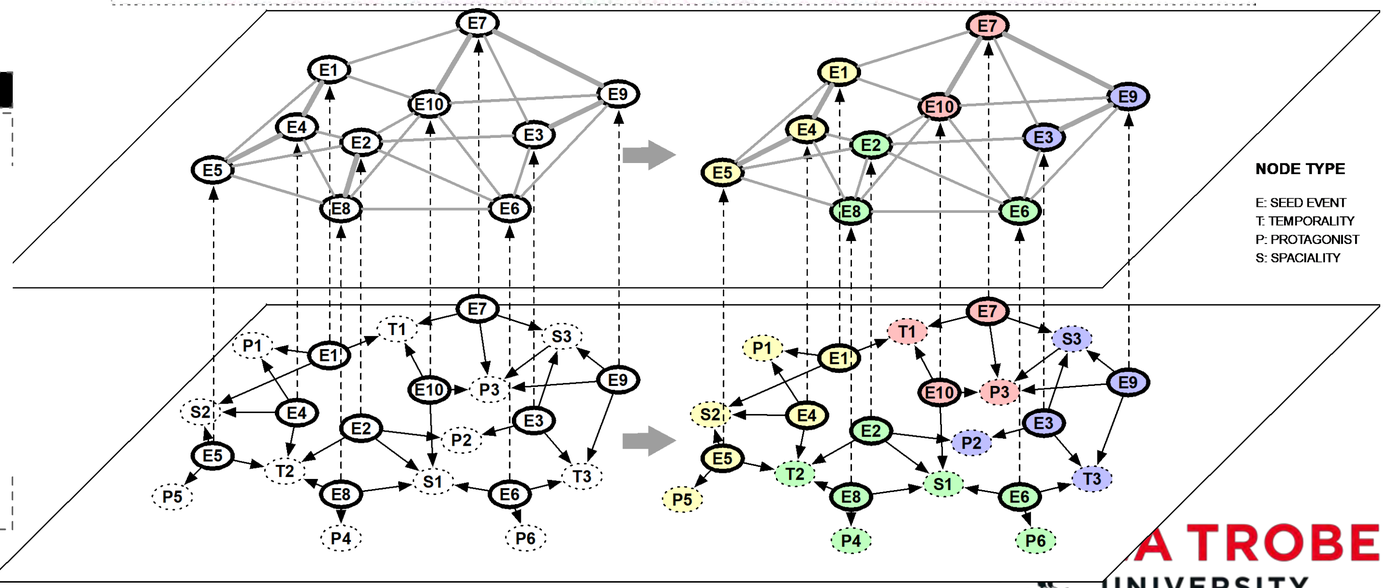
La Trobe Business School

2. Cognitive models for situation capture







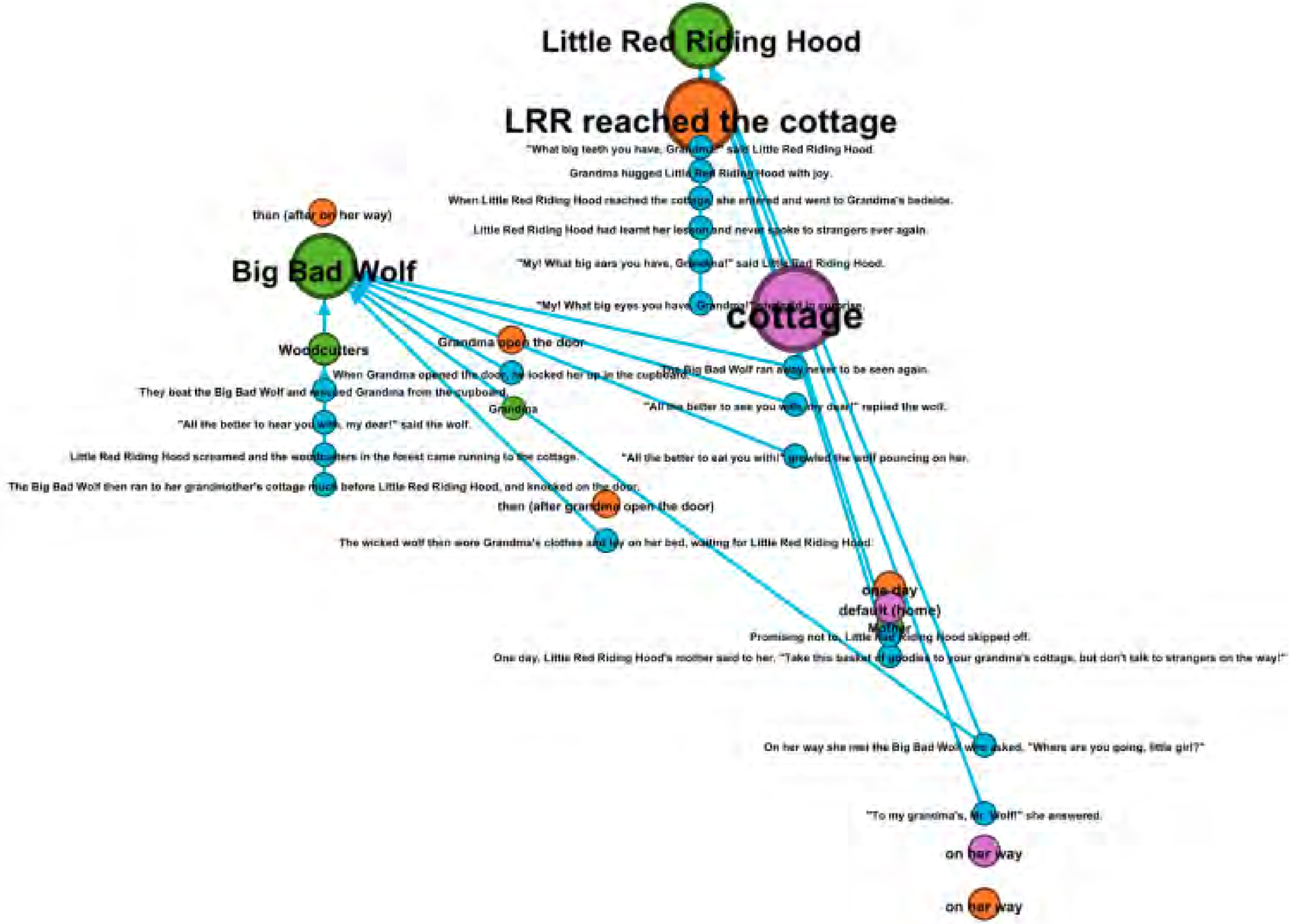


Event indexing model based base event integration

LA TROBE UNIVERSITY

La Trobe Business School

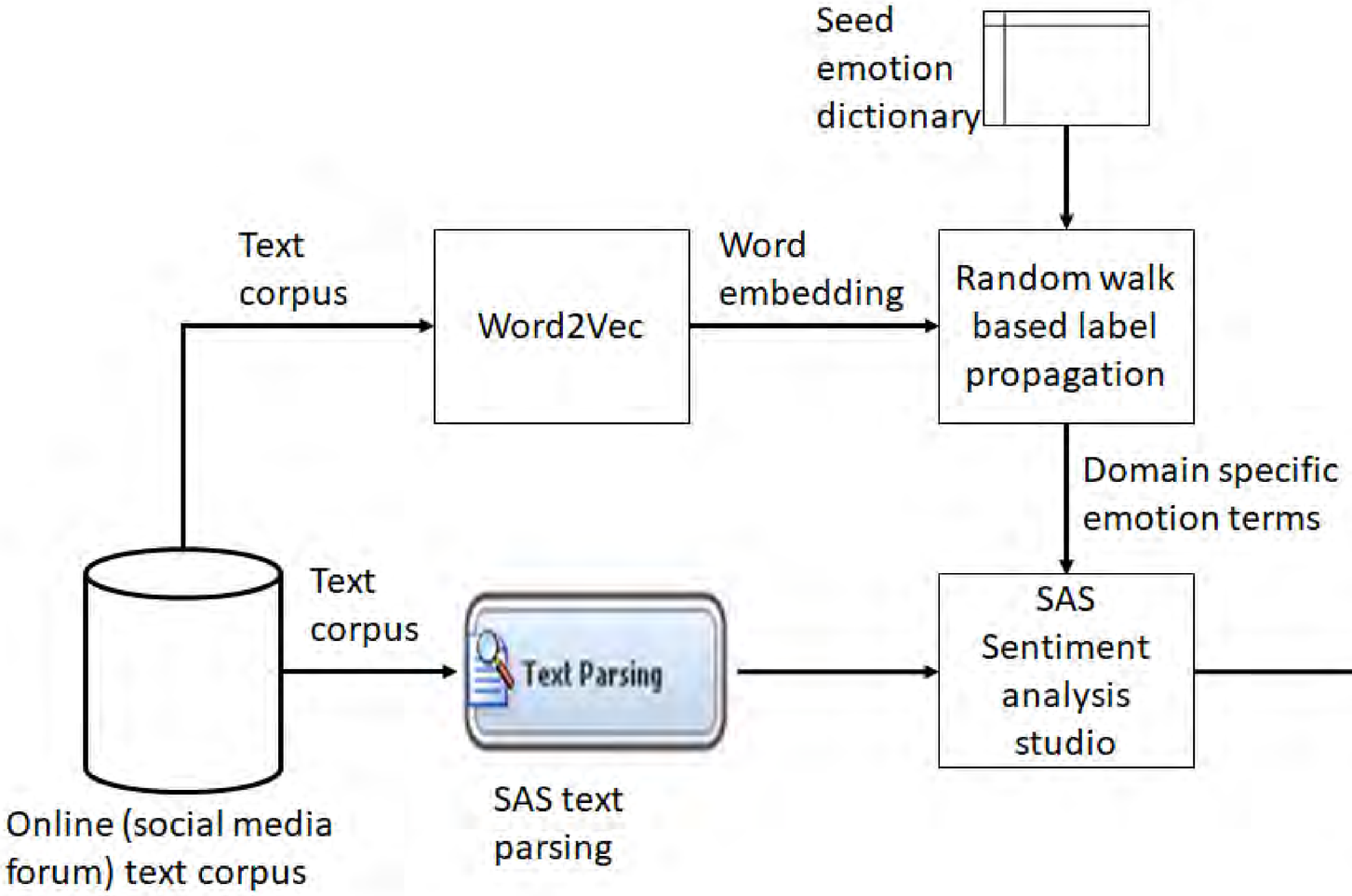
2. Cognitive models for situation capture



LA TROBE UNIVERSITY

La Trobe Business School

3. Capturing emotion and personality traits

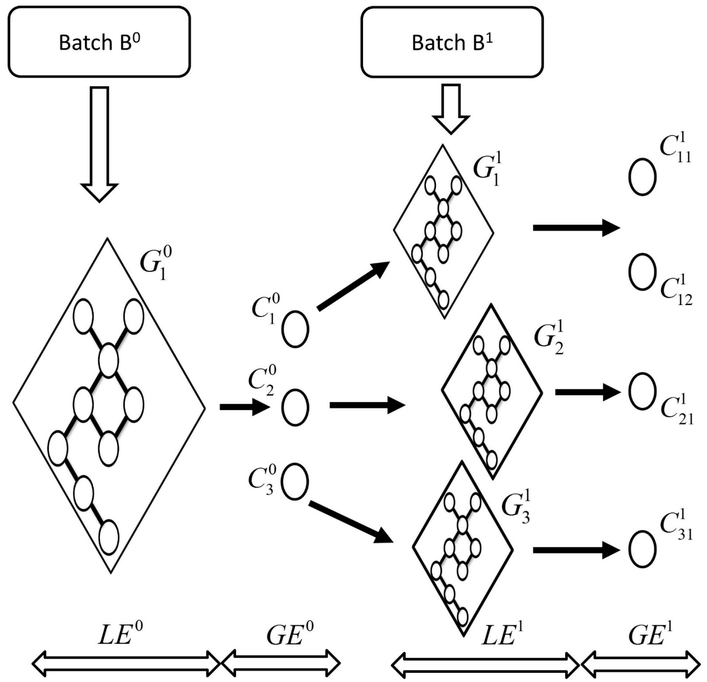


	Sentence	Positive	Negative
1	I hope this doesn't offend anyone.	Optimism: 0.5000	
2	I've come to observe that it is easier on people (without the disease) to cope with people (with cancer) to die quickly so they can process it, deal with it and move on.	Optimism: 0.2120	Sad: 0.0568
3	I'm not sure people know what to do with we (us) who are dying slowly.		Apprehension: 0.3336, Sad: 0.1336
4	It's a long slow support process, we're still here years after they heard we were ill - what do they do?		Apprehension: 0.1904, Sad: 0.1492
5	what do they say?		
	Total	**0.7120**	**0.8608**

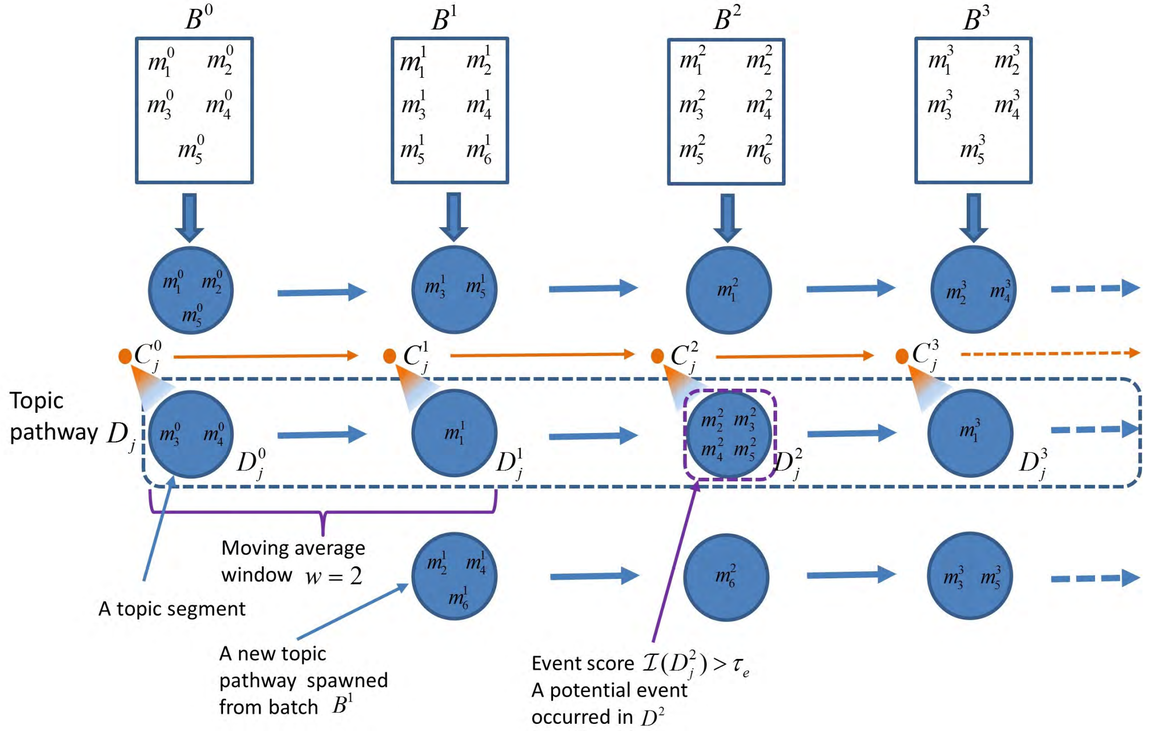
LA TROBE UNIVERSITY

La Trobe Business School

4. Monitor behaviour over time



GSOM as a base building block to capture data movement over time (IKASL algorithm)
Generates ‘cluster pathways’



Detecting potential events using changes in data volume and sentiment (text/social media application)
Captures events in cluster pathways

LA TROBE UNIVERSITY

La Trobe Business School

Text Mining and Sentiment Extraction Tools Used

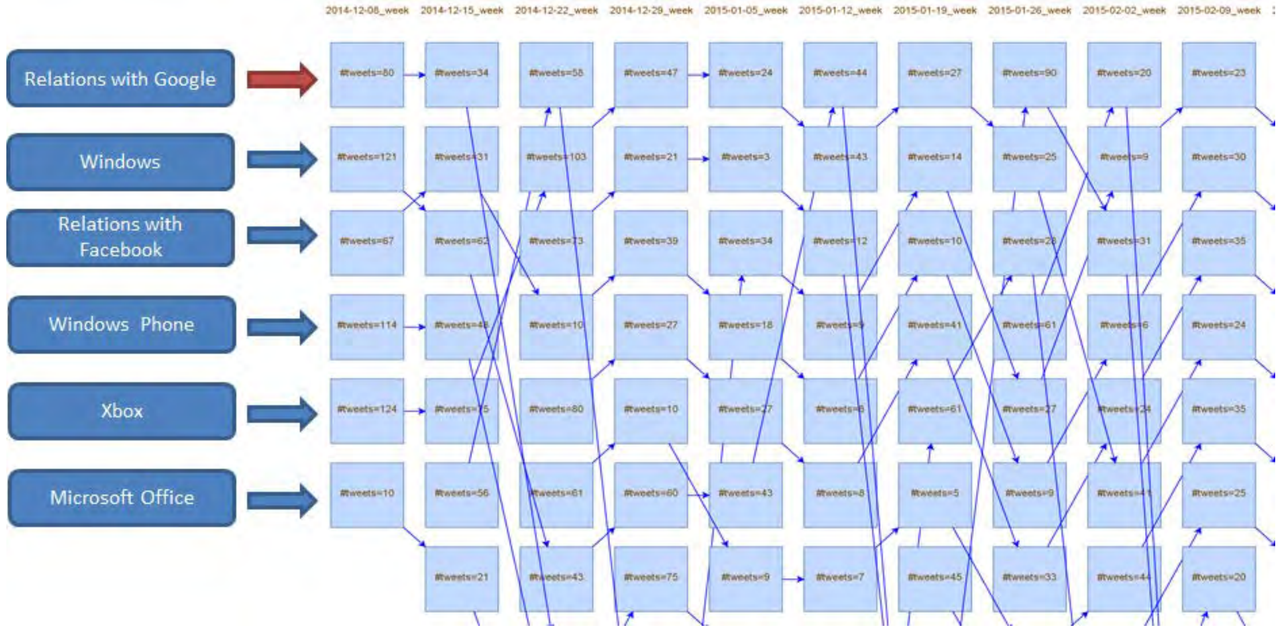
Text pre-processing	Text feature representation	Entity mining	Topic mining	Opinion mining
Text Filtering: • SAS Text Filter • NLTK python library • Text Mining (tm) R package • Weka	**Document frequency based:** • SAS Text Parser • NLTK TfidfVectorizer • Weka StringToWordVector • JATE (Java Automatic Term Extraction toolkit)	**Entity recognition:** • Stanford NER • OpenNLP NER	**Text clustering:** • SAS Text clustering • Weka java library • Sklearn python library	**Sentiment analysis:** • SAS sentiment analysis workbench • Stanford NLP sentiment analysis • SentiWordnet • SentiStrength • ANEW
Text parsing: • SAS Text Parser • NLTK text tokenizer • OpenNLP sentence parser, tokenizer • Stanford NLP sentence parser, tokenizer	**Word-embedding based:** • Word2Vec • GloVe	**Coreference resolution:** • Stanford NLP Coreference Resolution • spaCy • BART	**Topic modelling:** • SAS Text topic • Mallet • Gensim • Sklearn	**Emotion analysis:** • WordNet-Affect • EmoLex • DeepMoji

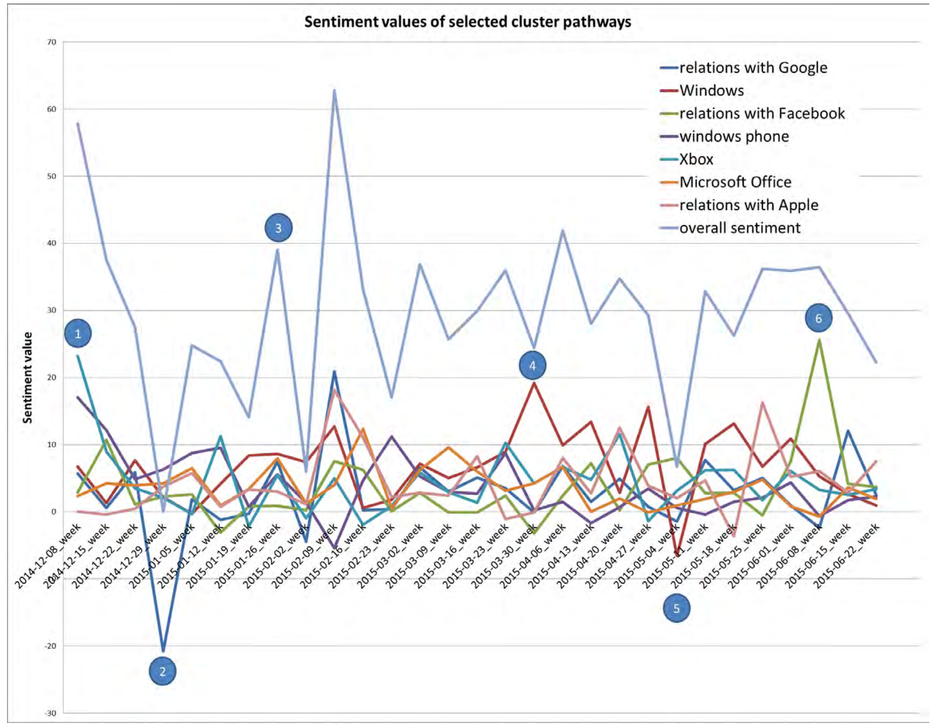
LA TROBE UNIVERSITY

La Trobe Business School

Example Application 1: Cluster pathway based event detection from tweets

topic pathways of #Microsoft dataset





Topic Pathway	Frequent Terms	Key focus of the Topic Pathway
$TP^{1}_{Microsoft}$	Google, android, amazon, apple, cyanogen, ibm, work, sony, app, window	Relations with Google
$TP^{2}_{Microsoft}$	Xbox, ps4, update, windows 10, xboxone, game, Cortana, sony, bitcoin, microsoft xbox	Microsoft Xbox related products and services
$TP^{3}_{Microsoft}$	microsoft office, ipad, android, mac, android tablet gmail, windowsphone, google, onenote, microsoft tech, office	Microsoft Office and its compatibility in different devices
$TP^{4}_{Microsoft}$	Windows, version, windows 10, os news, future, microsoft windows, upgrade, microsoft windows10 Week, update	Microsoft Windows
$TP^{5}_{Microsoft}$	Facebook, bing, google, microsoft bing, zdnet Apple, favour, monumental deal, search result, virtual reality	Relations with Facebook
$TP^{6}_{Microsoft}$	windows phone, app, office, android, bitcoin, onedrive, update, bitcoin payment, tablet, ipad	Windows Phone
$TP^{7}_{Microsoft}$	Apple, google, cloud, amazon, Samsung, microsoft store, bigdata, fight, device, patent	Relations with Apple

1. Microsoft accepts Bitcoin to buy Xbox games and Windows apps.
2. Google published a Windows 8.1 vulnerability.
3. Microsoft Consumer Preview on new products
4. Microsoft is thinking to make Windows an open source operating system in coming years.
5. Microsoft mentioned that Windows 10 would be the "last version" of Windows
6. Facebook and Microsoft announced a partnership on Virtual Reality project

LA TROBE UNIVERSITY

Example Application 2: Detecting depression from social media posts

Online Support Group	Depression themed	General
Healthboards	23,528	-
Healingwell	14,984	-
Patientinfo	6,413	193,098
Beyondblue	2,696	-
DailyStrength	1,498	-
	49,119	193,098

P_0 **I don't know**

I really don't know if I'm depress … whatever is wrong … or it's just a stage in life … I was in a … and I tried committing suicide … I cause a car accident … My head was so clouded …

P_1 I am with you, you. I feel you… Please look for some professional help … do not give up!

P_2 this is definitely not a stage in life everyone goes through…wishing to die and actively killing yourself? no …

Each sentence in the social media post is represented as feature vector using three sentence representation techniques.

1. A psycholinguistic representation that employs features delineated in LIWC.
2. A deep emotion based sentence representation which is captured using DeepMoji.
3. A word embedding based technique that uses a RNN to capture and encode relevant semantic and topical aspects of depression related discourse.

Top three posts with highest probability of being depression related in *Pizza-request* dataset

Forgive me if there's a better place for this, but I thought I'd try here. I'm a divorced dad with limited funds. The ex is remarried and in a better financial place. Not trying to compete with her, but I hate always telling the kids we can't go out to eat. …
today's been a pretty sh***y day. I've been suffering from depression all semester long, and it's really affected my school work. I also had a really sh***y therapy session today that mainly resulted in a lot of crying. …
I just moved to Indianapolis while I take a semester off of college from Indiana State and I have no food or drink. …. I'm just in a bad spot tonight so if I could get a pizza bro'd to me that would be great. …

Example Application 2: Discussion

Comments from computer science/AI conference reviewers

Extracts from Review 1

This is also a problem since the supervision is given by the discussion topic (explicitly listed as depression), while everything not in there is supposed to be from the negative class, as if people never show signs of depression in other topics.
This would need to be proven, or evaluated, especially to confirm the distribution of positive/negative instances

Extracts from Review 2

I think the critical point in depression recognition is to determine whether or not the target person/sample is a depressed patient. The person who post on the forum may not truly suffer the depressive disorder according to a clinical criterion. In the paper, you should figure out how to ensure or filter out the samples came from true depressed individuals

Early signs of depression are traditionally diagnosed in primary care and assessed using Depression Screening Questionnaires.
However, recent studies have found that this approach only has a positive case recognition rate of 36%-56%

Extracts from Review 2

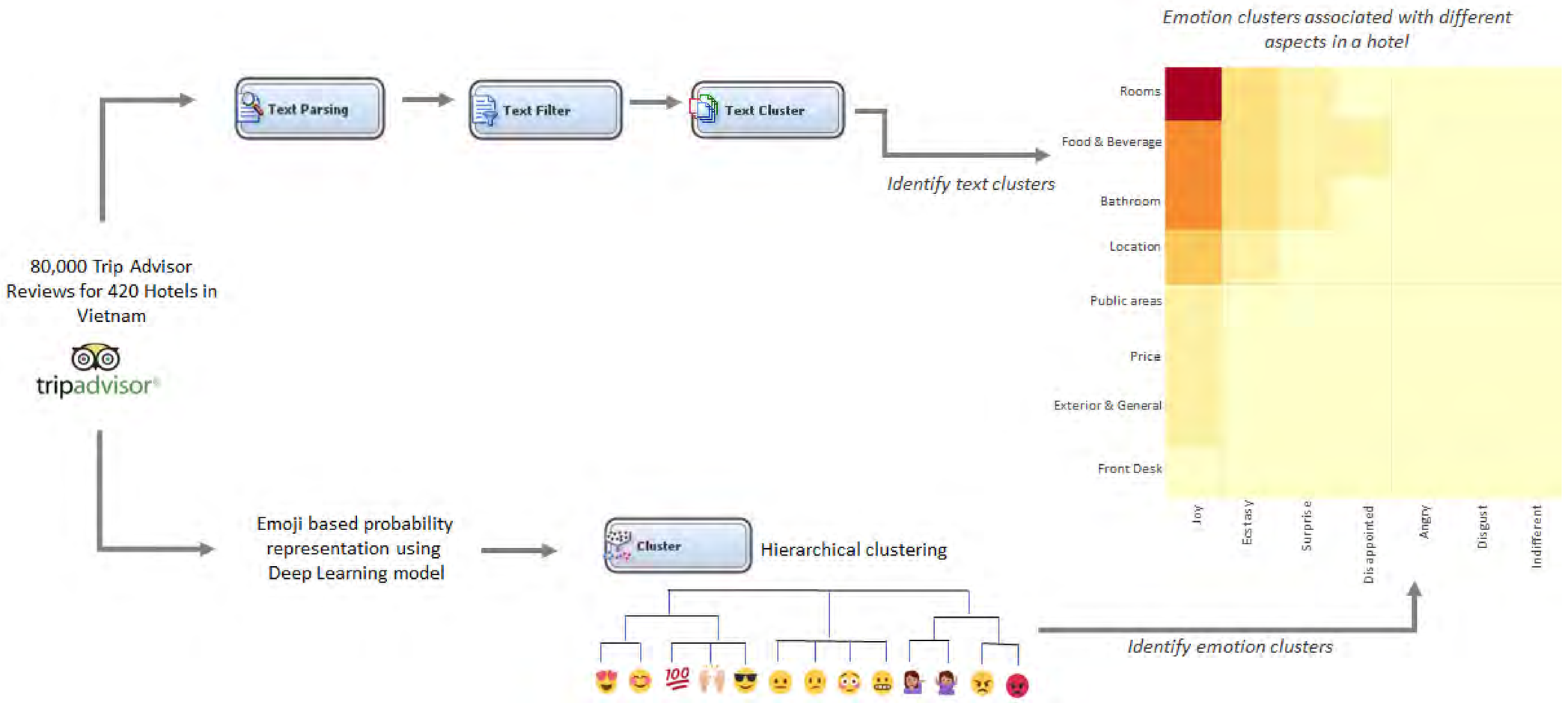
The only representation of the three that one would use is pretrained word2vec embeddings. The conclusion is that LIWC and DeepMoji are just useless, what is the point of using them? It is a negative result, and it is not surprisingly that pretrained embedding are good at this (we already know that BiLSTM work)

DeepMoji feature	Description	% AUC drop
😔	Pensive face: involved or engaged in deep serious thought	6.31
😖	Confounded face: frustrated, confused, failed or bewildered	5.82
😢	Crying face: crying with an attitude that other person has caused it	4.69
😷	Face with a medical mask: sick or do not want to talk	1.79
😂	Grinning face: smiling	1.41

LIWC feature	Sample terms	% AUC drop
Biological: Body	body, face, head, heart, muscle,	7.34
Affective: Sadness	sad, depress, cry, abandon, broke	6.82
Biological: Health	addict, bleed, heal, asthma, cramp	2.33
Regular verbs	ask, brought, call, care, gave	2.12
Auxiliary verbs	are, be, can, must, will	2.11

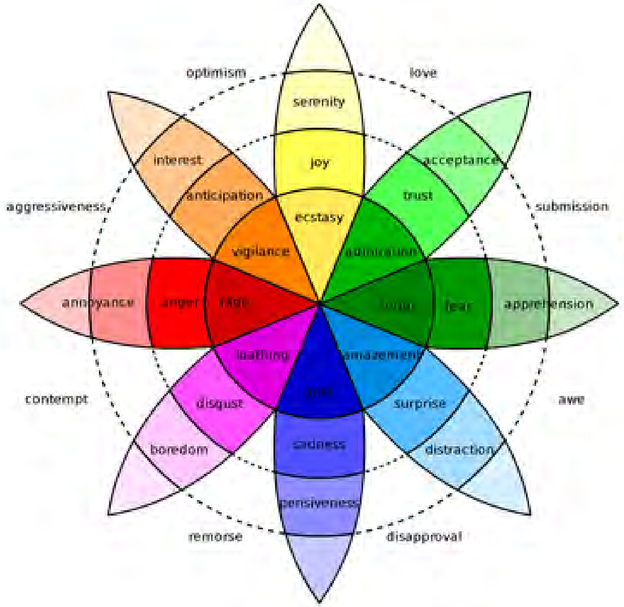
La Trobe Business School

Example Application 3: Emotion analysis of hotel reviews using deep learning and SAS



LA TROBE UNIVERSITY

Example Application 4: Emotion analysis from online cancer forums



Plutchik's wheel of emotions

Emotion dictionary using LIWC

Comprehensive natural language processing techniques

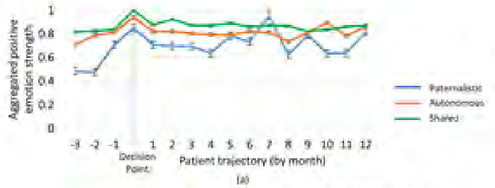
Online forum posts

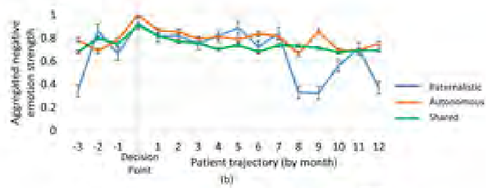
"I'm not happy about my condition now but I'm terrified about the pain after the surgery, I have heard it can be unbearable.. "

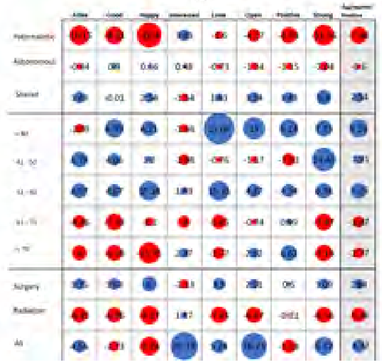
Emotion Analysis

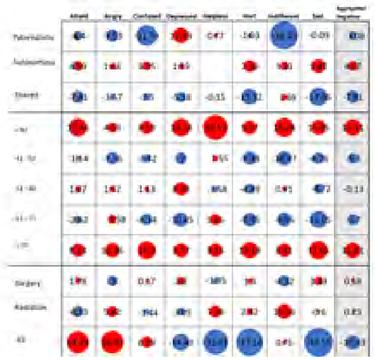
"I'm not happy about my condition now but I'm terrified about the pain after the surgery, I have heard it can be unbearable.. "

Emotions: SAD, AFRAID







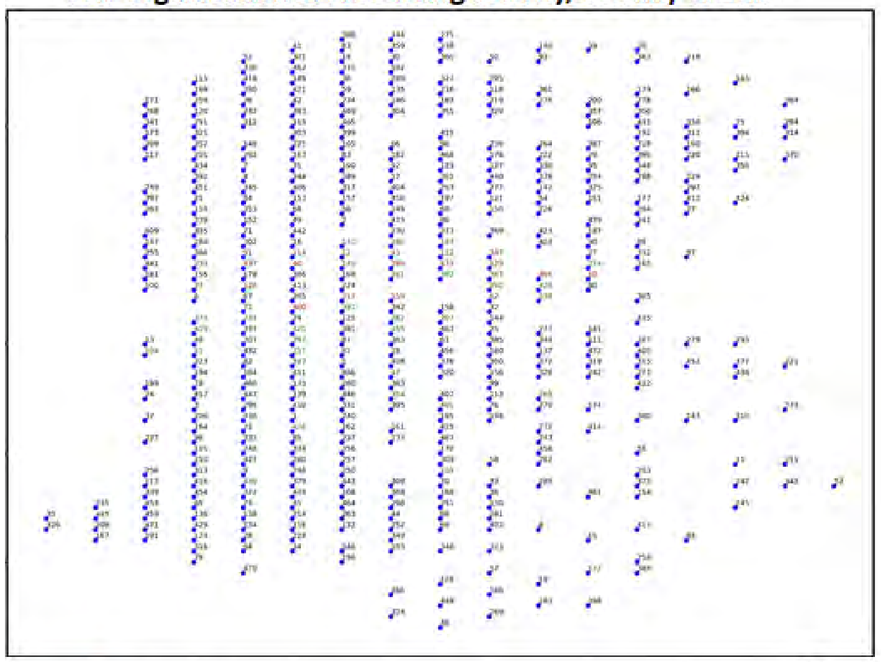


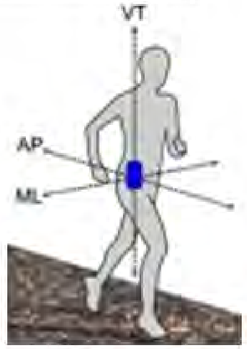
Case study: Different emotions detected in Prostate cancer patients using online forums

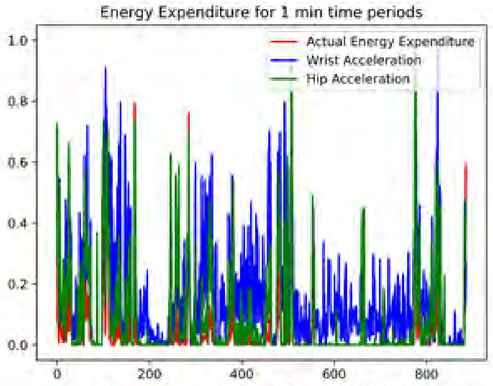
LA TROBE UNIVERSITY

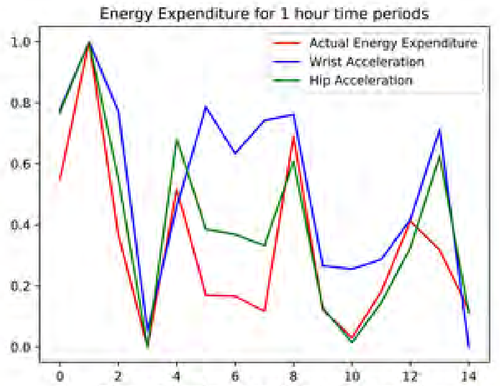
La Trobe Business School

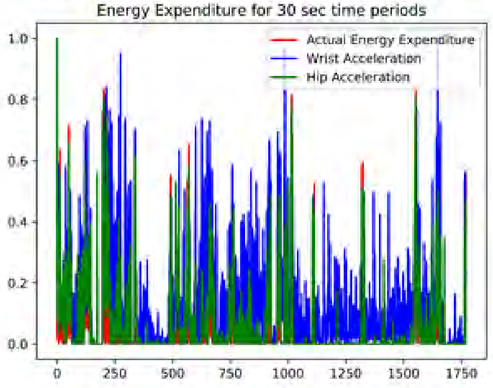
Example Application 5: Analysing data from ‘free living’ accelerometry data

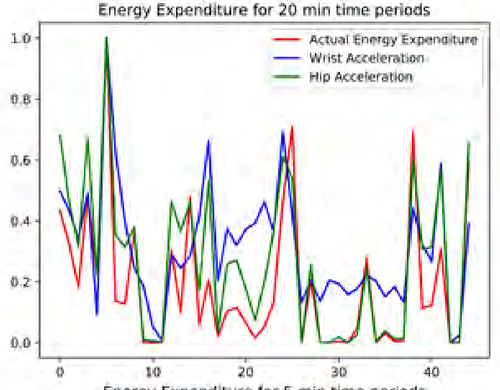


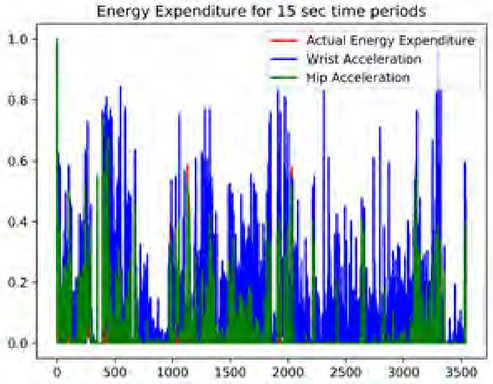


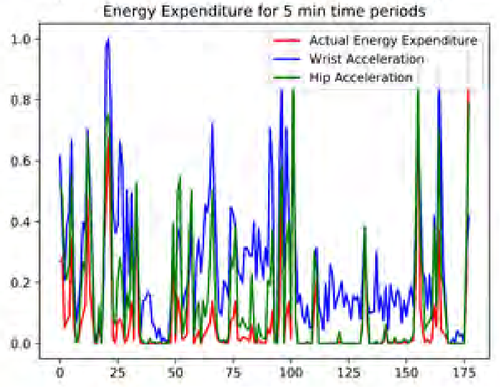












LA TROBE UNIVERSITY

La Trobe Business School

Thank You

LA TROBE UNIVERSITY