BUSINESS SCHOOL

Course Outline 2018

INFOSYS 722: DATA MINING AND BIG DATA (15 POINTS)

Semester 2 (1185)

Course prescription

Data mining and big data involves storing, processing, analysing and making sense of huge volumes of data extracted in many formats and from many sources. Using information systems frameworks and knowledge discovery concepts, this project-based and research-oriented course uses latest published research and cutting-edge business intelligence tools for data analytics.

Course advice

None

Goals of the course

The goals of the course are to introduce students to:

1. Decision Making, Big Data, Machine Learning, and Data Mining – foundational concepts.
2. Big Data and Data Mining Computing Environment – hardware, software, distributed systems and analytical tools.
3. Turning data into insights that deliver value - through methodologies, processes, algorithms and approaches for big data analytics.
4. Big Data and Data Mining in Practice – how the world's most successful companies use big data analytics to deliver extraordinary results.
5. Apply the knowledge gained through the design and implementation of a prototype.

Learning outcomes (LO)

#	Learning outcome	Graduate profile capability*
LO1	Understand foundational concepts of **Decision Making** and **Decision Support** from a variety of disciplines.	1. Disciplinary knowledge and practice 4a. Communication (Oral) 4c. Engagement
LO2	Understand fundamental principles of **Data Mining Machine Learning,** and **Big Data**.	1. Disciplinary knowledge and practice 4a. Communication (Oral) 4c. Engagement
LO3	Compare, contrast and synthesise a **process for Data Mining.**	2. Critical thinking 4a. Communication (Oral) 3. Solution seeking
LO4	Understand the key components of the computing environment for Big Data and Data Mining including **hardware**, **software**, **distributed systems**, and **analytical tools**.	1. Disciplinary knowledge and practice 3. Solution seeking 4c. Engagement
LO5	Understand the process of turning **data into insights that deliver value** using predictive modelling, segmentation, incremental response modeling, time series data mining, text analytics, and recommendations.	1. Disciplinary knowledge and practice 2. Critical thinking 3. Solution seeking
LO6	Understand, discuss, and reflect on how successful companies have applied big data and data mining **methodologies, algorithms,** and **enabling technologies** to deliver extraordinary results and value.	2. Critical thinking 3. Solution seeking 4c. Engagement
LO7	Design and implement a prototypical **Big Data Analytics Solution** to address one of the **17 Sustainable Development Goals** (https://sdgsinaction.com/) of the UN or a **decision making situation** facing an organization of your choice.	3. Solution seeking 6a. Social responsibilities 6b. Environmental responsibilities
LO8	Write a **research paper** that details (a) the practical problem (b) the research problem (c) the research objectives (d) the literature that explores potential solutions and methodologies that addresses your objectives (e) the research methodology adopted (f) the design of the processes that converts data into insights and (g) the description of the implementation using various algorithms and enabling technologies (h) your interpretation of the patterns and results and (i) your proposed actions based on the discovered knowledge.	2. Critical thinking 4b. Communication (Written) 5a. Independence

* See the graduate profile this course belongs to at the end of this course outline.

Content outline

Week / Module	Lectures	Labs	Assessment due this period
1 17 Jul	**Lecture:** Decision Making and Support. Intelligence Density. Big Data, Data Mining, and Machine Learning. Case studies from Marr 2016.	Data Mining Basics: Nine Step Process[1] using ISAS: Introduction and Application	
2 24 Jul	**Lecture:** Data Mining Processes (KDD, SEMMA, and CRISP-DM), Passive Data Mining (Browsing, Visualisation, Statistics, and Hypothesis testing)	ISAS - Advanced Use Case	Iteration 1 – Proposal - Steps 1 – 2 (Due 5pm, Wed, 25th July)
3 31 Jul	**Lecture:** Active Data Mining (Neural Networks, Rule Induction, Regression)	ISAS - Advanced Use Case	
4 7 Aug	**Lecture:** Overview of tools and technologies **Students Present:** Hardware, Distributed Systems & Analytical Tools (Chapters 1, 2, 3 - Dean 2014). Groups 1 – 3.	MSAS - Azure Machine Learning	Iteration 2 – ISAS - Steps 1 – 8 (Due 5pm, Wed, 8th Aug)
5 14 Aug	**Lecture:** Modelling **Students Present:** Predictive Modelling (Chapters 4, 5 – Dean 2014). Groups 4 – 6.	MSAS - Azure Machine Learning Collaboration and Power BI	
6 21 Aug	**Presenter:** Guest Lecturer **Objectives:** Understanding how to apply advanced data mining to complex real-world business problems.	OSAS - Introduction to Python for Machine Learning	
WORKSHOP 25th and 26th of August 10 AM – 6 PM Presenter: Dr. Chintan Amrit (University of Amsterdam, Netherlands).		**Objectives:** Apply a variety of OSAS tools to an advanced use case utilising the 9 step data mining process. **Resources:** Few 2006; Jensen et al 2010; Kaplan 2009.	

[1] Refer to the nine steps of the assignment specification at the end of this document

Week / Module	Lectures	Labs	Assessment due this period
7 11 Sep	**Lecture:** Visualisation **Students Present:** Segmentation (Chapter 6 – Dean 2014). Groups 7 – 9.	OSAS - Python Machine Learning Workflow	
8 18 Sep	**Lecture:** Interpretation **Students Present:** Incremental Response Modeling & Time Series Data Mining (Chapters 7, 8 - Dean 2014). Groups 10 – 12.	BDAS - Introduction to GitHub, Amazon Web Services, Jupyter, Secure Shell (SSH), PySpark and Spark	Iteration 3 – MSAS/OSAS - Steps 1 – 8 (Due 5pm, Mon, 17th Sept)
9 25 Sep	**Lecture:** Assessment, Evaluation, and Iteration **Students Present:** Text Analytics and Recommendation Systems (Chapters 10, 9 – Dean 2014). Groups 13 – 15.	BDAS - PySpark, Spark, DataFrames and Data Cleaning Application	
10 2 Oct	**Lecture:** Action **Students Present:** Case Studies of Big Data Analytics (Chapters 11-16 of Dean 2014 and Marr 2016). Groups 16 – 18.	BDAS - PySpark Algorithms and Visualisations	
11 9 Oct	Conclusion	BDAS – Recap on Tools and Technology and Assignment Help	Iteration 4 – BDAS - Steps 1 – 8 (Due 5pm, Fri, 12th Oct)
12 16 Oct	The five best PechaKucha presentations from each tutorial stream (15 in total) will be presented in class.	Assignment Help	Resubmission of Iteration 2 and 3 (Due 5pm, Fri, 19th Oct)
Research Paper - Details of Steps 1 – 9 (Due 5pm, Fri, 26th Oct)			

Learning and teaching

The class will meet for three hours each week. Class time will be used for a combination of lectures and discussions. In addition to attending classes, students should be prepared to spend at least about another ten hours per week on activities related to this course. These activities include carrying out the required readings, labs and research relevant to this course, and preparing for assignments.

150 hours learning over a single semester including:

- 36 contact hours through lectures
- 22 contact hours through laboratories/tutorials
- 16 contact hours through technical workshop
- 76 hours of reading, programming, and self-study

Teaching staff

Course Director and Lecturer

David Sundaram Office: OGGB Room 476 Office Hour: Tuesdays 1-2 PM
Email: d.sundaram@auckland.ac.nz Phone: 09 – 923 5078 Fax: 09-373-7430

Coordinators and Tutors

Shahzad Khan (Course Coordinator and Tutor) Office Hour: 4-5 PM Thursday Room: 4102
Email: shahzad.khan@auckland.ac.nz

Diego Aparicio (Tutor) Office Hour: 4-5 PM Thursday Room: 4102
Email: denr517@aucklanduni.ac.nz

Shohil Kishore (Technical Coordinator)
Email: s.kishore@auckland.ac.nz

Learning resources

Course Material

There are two primary textbooks used for the course. These textbooks can be downloaded free of cost from the University of Auckland library.

Dean, J., 2014. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons.

Marr, B., 2016. Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. John Wiley & Sons.

Workshop Material

Few, S., 2006. Information Dashboard Design: The Effective Visual Communication of Data.

Jensen, C.S., Pedersen, T.B. and Thomsen, C., 2010. Multidimensional databases and data warehousing. *Synthesis Lectures on Data Management*, *2*(1), pp.1-111.

Kaplan, R.S., 2009. Conceptual foundations of the balanced scorecard. *Handbooks of management accounting research*, *3*, pp.1253-1269.

Other readings and supplemental material will be distributed in class as needed. Students are also advised to take advantage of the extensive software resources made available for this course.

Assessment information

ISAS – IBM Software Analytics Solution – SPSS Modeller.

MSAS – Microsoft Software Analytics Solution – Microsoft SQL Server, Azure Machine Learning and Power BI.

OSAS – Open Source Analytics Solution – Python, Jupyter, MySQL, MySQL Workbench, Kettle/Spoon, Tableau, & Weka.

BDAS – Big Data Analytics Solutions – GitHub, AWS (EC2/AMI), Jupyter, PySpark, Spark.

Assessment task	Weight %	Group and/or individual	Submission
1. Presentations – Dean (2014)	5	Group	Weeks 4 – 10
2. Iteration 1 Proposal (Steps 1 – 2)	0	Individual	Week 2 – 5pm, Wed, 25th July
3. Iteration 2 ISAS (Steps 1 – 8)	20	Individual	Week 4 – 5pm, Wed, 8th Aug
4. Iteration 3 MSAS/OSAS (Steps 1 – 8)	25	Individual	Week 8 – 5pm, Mon, 17th Sept
5. Iteration 4 BDAS (Steps 1 – 8)	30	Individual	Week 11 – 5pm, Fri, 12th Oct
6. Resubmission of Iterations 2 and 3		Individual	Week 12 – 5pm, Fri, 19th Oct
7. Research Paper (Details of Steps 1 – 9)	20	Individual	5pm, Fri, 26th Oct

Pass requirements

Plussage applies for Iterations 2 and 3. That is if you **re-submit Iterations 2 and 3 in Week 12** then we will remark them and if you score a better mark we will take the better mark as your mark. You will also get a **bonus of 5 marks** if you implemented Iteration 3 in **MSAS as well as OSAS and hand it in Week 12.**

Description of assessment tasks

Assessment task	Learning outcome to be assessed
1. Presentations – Dean (2014)	LO1 – LO6
2. Iteration 1 Proposal (Steps 1 – 2)	LO7
3. Iteration 2 ISAS (Steps 1 – 8)	LO1 – LO7
4. Iteration 3 MSAS/OSAS (Steps 1 – 8)	LO1 – LO7
5. Iteration 4 BDAS (Steps 1 – 8)	LO1 – LO7
6. Resubmission of Iteration 2 and 3	LO1 – LO7
7. Research Paper (Details of Steps 1 – 9)	LO8

Inclusive learning

Students are urged to discuss privately any impairment-related requirements face-to-face and/or in written form with the courses convenor/lecturer and/or tutor.

Academic integrity

The University of Auckland will not tolerate cheating, or assisting others to cheat, and views cheating in coursework as a serious academic offence. The work that a student submits for grading must be the student's own work, reflecting his or her learning. Where work from other sources is used, it must be properly acknowledged and referenced. This

requirement also applies to sources on the worldwide web. A student's assessed work may be reviewed against electronic source material using computerised detection to provide an electronic version of their work for computerised review.

Student feedback

Student feedback is important to us and has been used to improve the course from semester to semester. This semester you may be asked to complete evaluations on the teaching of the course, both in lectures and in tutorials. Please note that you do not have to wait until these evaluations are conducted in order to provide feedback. If there is something that you think we could improve then please let us know (via email or in person) as soon as possible.

In the event of an unexpected disruption

We undertake to maintain the continuity and standard of teaching and learning in all your courses throughout the year. If there are unexpected disruptions, the University has contingency plans to ensure that access to your course continues and your assessment is fair, and not compromised. Some adjustments may need to be made in emergencies, In the event of a disruption, the University and your course coordinators will make every effort to provide you with up to date information via Canvas and the University website.

Graduate profile for Master of Commerce

The following six themes represent the capabilities that the Business School seeks to foster in all of its graduates. The development of these capabilities does not come all at once, but rather is expected to build from year to year. Each course is not expected to contribute to all capabilities, but each course will have its own goals and learning outcomes that relate to the overall development of this profile.

Graduate Profile	
1.	Disciplinary knowledge and practice Graduates will be able to apply highly specialised knowledge within the discipline to demonstrate an advanced awareness and understanding in a global context.
2.	Critical thinking Graduates will be able to analyse and evaluate the relevant literature, and design and develop scholarly arguments that demonstrate advanced and diverse thinking.
3.	Solution seeking Graduates will be able to creatively research and analyse complex issues, and develop innovative solutions.
4.	Communication and engagement Graduates will be able to engage, communicate, and collaborate with diverse groups using multiple formats and effectively address a range of professional and academic audiences.
5.	Independence and integrity Graduates will be able to demonstrate advanced independent thought, self-reflection, ethics, and integrity.
6.	Social and environmental responsibility Graduates will consider, in relation to their discipline, the potential significance of the principles underpinning both the Treaty of Waitangi and sustainability.

INFOSYS 722 – Assignment Specification

Design and implement a prototypical ***Data Mining*** and ***Big Data Analytics Solution*** to address one of the **17 Sustainable Development Goals of the UN** or a **decision making situation facing an organization of your choice**.

The assignment follows a sequence of steps that is a synthesis of the Cross-Industry Standard Process for Data Mining (CRISP-DM) process (SPSS, 2007) and the KDD process (Fayyad et al., 1996).

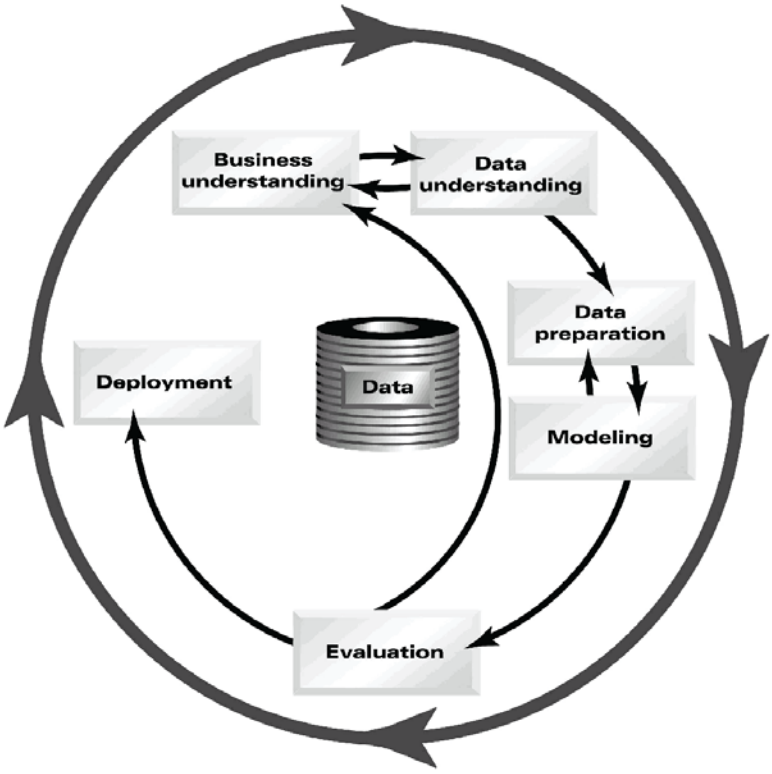

Figure 1: CRISP DM Process (SPSS, 2007)

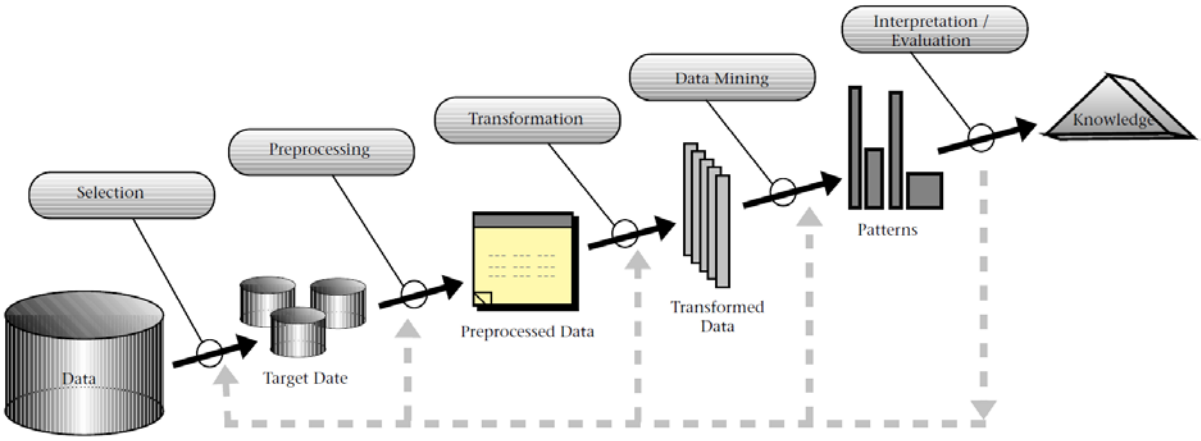

Figure 2: KDD Process (Fayyad et al., 1996)

1. **Business and/or Situation understanding.** "First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint." (Fayyad et al., 1996)

1.1 Identify the objectives of the business and/or situation
1.2 Assess the situation
1.3 Determine data mining objectives, and
1.4 Produce a project plan

2. **Data understanding.** Data provides the "raw materials" of data mining. This phase addresses the need to understand what your data resources are and the characteristics of those resources. "Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed." (Fayyad et al., 1996)

   2.1 Collect initial data
   2.2 Describe the data
   2.3 Explore the data, and
   2.4 Verify the data quality

3. **Data preparation.** After cataloguing your data resources, you will need to prepare your data for mining. "Third is data cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes" (Fayyad et al., 1996)

   3.1 Select the data
   3.2 Clean the data
   3.3 Construct the data
   3.4 Integrate various data sources
   3.5 Format the data as required

4. **Data transformation:** "Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found." (Fayyad et al., 1996)

   4.1 Reduce the data
   4.2 Project the data

5. **Data-mining method(s) selection:** "Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as in Fayyad, Piatetsky-Shapiro, and Smyth (1996)." (Fayyad et al., 1996)

   5.1 Match and discuss the objectives of data mining (1.1) to data mining methods
   5.2 Select the appropriate data-mining method(s) based on discussion

6. **Data-mining algorithm(s) selection:** "Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities)." (Fayyad et al., 1996)

   6.1 Conduct exploratory analysis and discuss
   6.2 Select data-mining algorithms based on discussion
   6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

7. **Data Mining:** “Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.” (Fayyad et al., 1996) This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data.

   7.1 Create and justify test designs
   7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute)
   7.3 Search for patterns

8. **Interpretation:** “Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.” (Fayyad et al., 1996) We assess and evaluate the models and the results and their reliability. “You are ready to evaluate how the data mining results can help you to achieve your objectives.” (SPSS, 2007)

   8.1 Study and discuss the mined patterns
   8.2 Visualize the data, results, models, and patterns
   8.3 Interpret the results, models, and patterns
   8.4 Assess and evaluate results, models, and patterns
   8.5 Iterate prior steps (1 – 7) as required

9. **Action:** “Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.” (Fayyad et al., 1996) “Now that you’ve invested all of this effort, it’s time to reap the benefits. This phase focuses on integrating your new knowledge into your everyday business processes to solve your original business problem and/or situation.” (SPSS, 2007)

   9.1 Discuss how you would apply the knowledge and deploy the implementation
   9.3 Discuss how you would monitor the implementation
   9.4 Discuss how you would maintain the implementation
   9.5 How could you enhance the solution in the future?

———————————

INFOSYS 722 – Lecture and Lab Readings, Videos and Materials

Week 1 **Week 2** **Week 3**	**Data Mining Basics: Steps 1 - 9 using ISAS – Introduction and Advanced Use Case**
	Langley, A., Mintzberg, H., Pitcher, P., Posada, E., & Saint-Macary, J. (1995). Opening up decision making: The view from the black stool. organization Science, 6(3), 260-279. https://www.nmbu.no/download/file/fid/15127 Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37. https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131
	SPSS Modeller User Guide
	SPSS Modeller CRISP-DM Guide
	Clementine User Guide
	Building a Data Mining Model
	Predictive Analytics on SPSS Modeller / Constructing a Predictive Model
	Building a Data Visualisation Model
	Changes to Computer Thinking
	Microsoft Course on Data Science Fundamentals (SL)
	Iteration 1 – Proposal (Due 5pm, 25th July)

Week 4 **Week 5**	**MSAS – Azure Machine Learning and Power BI**
	Getting Started with Microsoft Azure
	Machine Learning Overview
	AML Basics
	Practical AML Experiment / Comparing Regressors on AML
	Power BI Basics (SL) / Power BI Analytics (SL)
	From CSV to Power BI Dashboard
	SQL Server Data Tools Tutorial / Microsoft SSAS Tutorials
	The Beauty of Data Visualisation
	Using the Cloud to Forecast the Weather
	The Turing Test
	Big Data Is Better Data
	Iteration 2 – ISAS (Due 5pm, 8th Aug)

Week 6	**OSAS – Python for Machine Learning**
	'Hello World' for Machine Learning (SL)
	Python for Non-Programmers (SL)
	Python Semantics (SL) / Python Fundamentals (SL)
	What is Anaconda?
	What is Jupyter Notebooks? / What is Spyder?

Mid-Semester Break
OSAS Workshop 25th and 26th of August

Week 7	**OSAS – Machine Learning Workflow**
	Kettle Fundamentals
	Weka Fundamentals Course
	Tableau Fundamentals
	MySQL Fundamentals / MySQL Workbench
	Neural Networks that Change Everything
	Iteration 3 – MSAS/OSAS (Due 5pm, 17th Sept)

Week 8 **Week 9** **Week 10** **Week 11**	**BDAS – Introduction and Application**
	AWS EC2 Documentation / EC2 Masterclass
	Git Crash Course
	What is Spark? / Spark 101
	PySpark Introduction / PySpark with Big Data Course
	Can Google Predict the Stock Market?
	Iteration 4 – BDAS (Due 5pm, 12th Oct)

Week 12	**Assignment Assistance**
	Resubmission of Iteration 2 and 3 (Due 5pm, 19th Oct)

Research Paper (Due 5pm, 26th Oct)

SL = Self-Learning. Anything marked with 'SL' is a course where you may gain a certification. You can complete these courses at your discretion.