THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

BUSINESS SCHOOL

Course Outline 2017
INFOSYS 722: Data Mining and Big Data (15 POINTS)

Semester 2 (1175)

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.

Programme and Course Advice

None

Goals of the Course

The goals of the course are to introduce students to:

1. Decision Making, Big Data, and Data Mining – foundational concepts.
2. Big Data and Data Mining Computing Environment – hardware, distributed systems and analytical tools.
3. Turning data into insights that deliver value - through methodologies, 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

By the end of this course it is expected that a student will be able to:

1. Understand foundational concepts of **decision making** and **decision support** from a variety of disciplines;
2. Understand fundamental principles of **Data Mining** and **Big Data**;
3. Compare, contrast and synthesise a **process for Data Mining**
4. Understand the key components of the computing environment for Big Data and Data Mining including **hardware**, **distributed systems**, and **analytical tools**;
5. 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;
6. 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;

7. Design and implement a prototypical **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;
8. 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.

Content Outline

Week - Date	Lectures (Tuesday 9 AM - 12 PM)
1 : 25 Jul	**Lecture:** Decision Making and Support. Intelligence Density. Big Data, Data Mining, and Machine Learning. Case studies from Marr 2016.
2 : 1 Aug	**Lecture:** Data Mining Processes (KDD, SEMMA, and CRISP-DM), Passive Data Mining (Browsing, Visualisation, Statistics, and Hypothesis testing)
3 : 8 Aug	**Lecture:** Active Data Mining (Neural Networks, Rule Induction, Regression) **Guest Lecture:** Professor Michael Myers (Writing Publishable Research Papers)
WORKSHOP 12th & 13th Aug 9 AM – 5 PM	**Objectives:** Determine the business questions, designing and filling the data warehouse, visualising and machine learning. **Resources:** Few 2006; Jensen et al 2010; Kaplan 2009.
4 : 15 Aug	**Guest Lecture:** Karen Hardie and colleagues from IBM on Advanced Data Mining using SPSS Modeller
5 : 22 Aug	**Lecture:** Overview of tools and technologies **Students Present:** Hardware, Distributed Systems & Analytical Tools (Chapters 1, 2, 3 - Dean 2014). Groups 1 – 3.
6 : 29 Aug	**Lecture:** Modelling **Students Present:** Predictive Modelling (Chapters 4, 5 – Dean 2014). Groups 4 – 6.
7 : 19 Sep	**Lecture:** Visualisation **Students Present:** Segmentation (Chapter 6 – Dean 2014). Groups 7 – 9.
8 : 26 Sep	**Lecture:** Interpretation **Students Present:** Incremental Response Modeling & Time Series Data Mining (Chapters 7, 8 - Dean 2014). Groups 10 – 12.
9 : 3 Oct	**Lecture:** Assessment, Evaluation, and Iteration **Students Present:** Text Analytics and Recommendation Systems (Chapters 10, 9 – Dean 2014). Groups 13 – 15.
10 : 10 Oct	**Lecture:** Action **Students Present:** Case Studies of Big Data Analytics (Chapters 11-16 of Dean 2014 and Marr 2016). Groups 16 – 18.
11 : 17 Oct	Conclusion
12 : 24 Oct	The five best **PechaKucha** presentations from each tutorial stream (15 in total) will be presented in class.

Week	Labs
1	Data Mining Basics: Steps[1] 1 - 9 using SPSS Modeller
2	Data Integrator (Kettle / Spoon)
3	Data Integrator (Kettle / Spoon)
	Workshop
4	SPSS Modeller
5	SPSS Modeller
6	Microsoft Stack Overview (SQL Server / Azure ML / Power BI)
	Mid-Semester Break
7	Microsoft Stack (Power BI)
8	Microsoft Stack (Azure ML)
9	Big Data (Hadoop with MapReduce and HDInsight)
10	Big Data (Hadoop with MapReduce and HDInsight)
11	Big Data (Hadoop with MapReduce and HDInsight)
12	Assignment Assistance

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 and the final exam.

Teaching Staff

David Sundaram (Lecturer)
Office: OGGB Room 476
Office Hour: Tuesdays 12-1 PM
Email: d.sundaram@auckland.ac.nz Phone: 09 – 923 5078 Fax: 09-373-7430

Course Coordinator and Tutors
Shohil Kishore (Course Coordinator)
Office: OGGB Room 428
Office Hour: Wednesday 1-2 PM
Email: s.kishore@auckland.ac.nz

Shahab Bayati (Tutor)
Email: s.bayati@auckland.ac.nz

Jose Ortiz (Tutor)
Email: j.ortiz@auckland.ac.nz

Roshan Jonnalagadda (Tutor)
Email: jros093@aucklanduni.ac.nz

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

Learning Resources

Course Material

There are two primary textbooks used for the course. These text books 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

SPSS – IBM – SPSS Modeller Solution.
MSAS – Microsoft Analytics Solution – Microsoft SQL Server, SQL Server BI, & Azure Machine Learning.
OSAS – Open Source Analytics Solution – MySQL, Workbench, Kettle/Spoon, Tableau, & Weka.
BDAS – Big Data Analytics Solutions – Hadoop, MapReduce, and/or HDInsight.

Assessment	**Name**	**Marks**	***Due Date***
1.	Group Presentations – Dean 2014	5	Weeks 5 - 10
2. Iteration 1	Proposal (Steps[2] 1 – 2)	0	Week 2 – 31st Jul – 5pm
3. Iteration 2	SPSS (Steps 1 – 8)	20	Week 5 – 25th Aug – 5pm
4. Iteration 3	MSAS or OSAS (Steps 1 – 5)	15	Week 7 – 22nd Sep – 5pm
5. Iteration 4	MSAS or OSAS (Steps 6 – 8)	20	Week 10 – 13th Oct – 5pm
6. Iteration 5	BDAS (Steps 6 – 8)	20	Week 12 – 24th Oct – 9am
7. Paper	Research Paper (Details of Steps 1 – 9)	20	Week 12 – 27th Oct – 5pm

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

Plussage applies between Iterations 2-5. That is if you re-submit Iterations 2-4 along with Iteration 5 then we will remark them and if you score a better mark we will take the better mark as your mark. You will get a **bonus of 7 marks** if you implemented Iterations 3 and 4 in MSAS as well as OSAS!

Learning Outcome	Assessment
1	1,2,3,4,5,6,7
2	1,2,3,4,5,6,7
3	1,2,3,4,5,6,7
4	1,2,3,4,5,6,7
5	1,2,3,4,5,6,7
6	1,2,3,4,5,6,7
7	1,2,3,4,5,6,7
8	1,2,3,4,5,6,7

Inclusive Learning

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

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.

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).

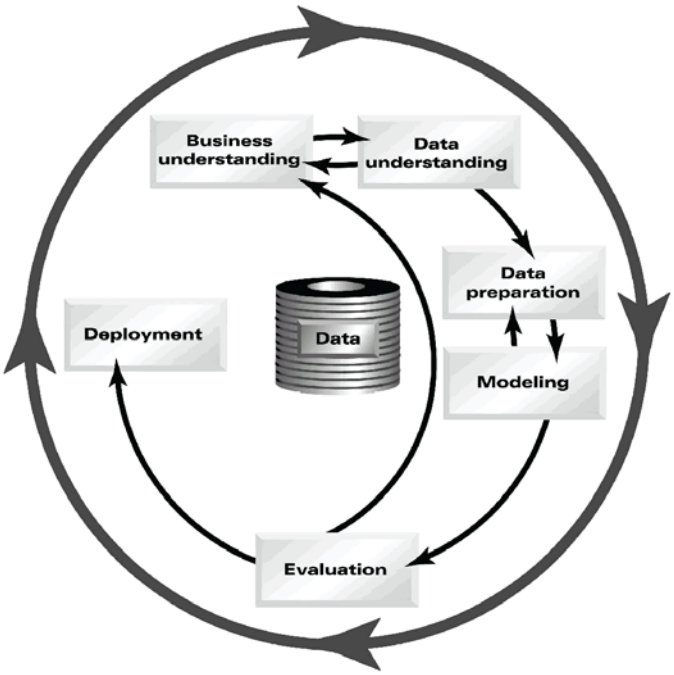

Figure 1: CRISP DM Process (SPSS, 2007)

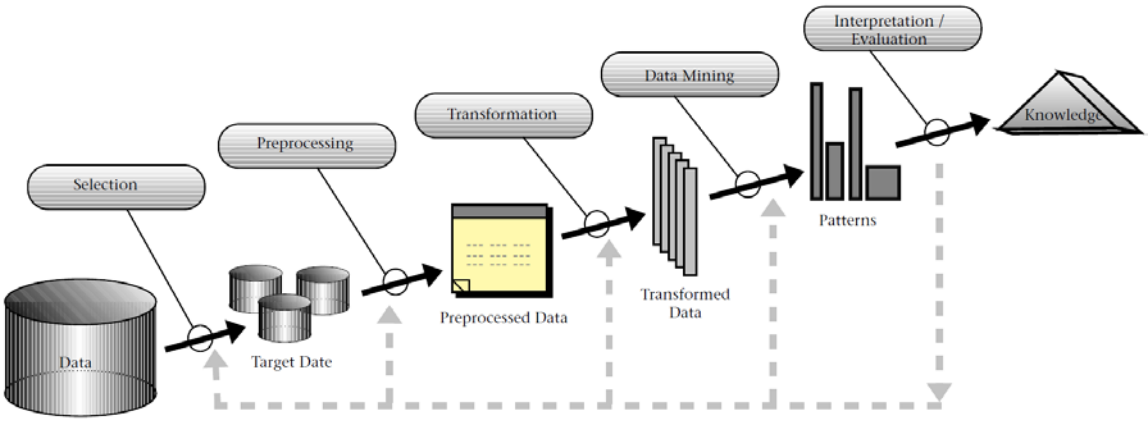

Figure 2: KDD Process (Fayyad et al., 1996)

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 goals, 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 the data
   3.5 Format the data

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 the goal of data mining to data mining methods
   5.2 Select appropriate data-mining method(s)

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
   6.2 Select data-mining algorithms
   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 test designs
7.2 Conduct data mining – classify, regress, cluster, etc.
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 the mined patterns
   8.2 Visualize the data, models, and patterns
   8.3 Interpret the patterns
   8.4 Assess and evaluate models
   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 Plan the deployment
   9.2 Implement the plan
   9.3 Monitor the implementation
   9.4 Maintain the implementation
   9.5 Produce a final report
   9.6 Review the project

––––––––––––––––––

INFOSYS 722 – Lecture and Lab Readings, Videos and Materials

Week 1	**Data Mining Basics: Steps 1 - 9 using SPSS Modeller**
	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.
	SPSS Modeller User Guide
	SPSS Modeller CRISP-DM Guide
	Clementine User Guide
	Microsoft Course on Data Science Fundamentals

Week 2 **Week 3**	**Data Integrator (Kettle / Spoon)**
	Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.
	What is LAMP?
	Kettle Fundamentals
	MySQL Workbench Fundamentals
	Iteration 1: Proposal Due (31st of July)

Two-Day Workshop (Kettle / Spoon / MySQL / MySQL Workbench / Tableau)

Week 4 **Week 5**	**SPSS Modeller**
	Building a Data Mining Model
	Predictive Analytics on SPSS Modeller / Constructing a Predictive Model
	Building a Data Visualisation Model
	Connecting SQL Server with SPSS Modeller
	Iteration 2: SPSS Iteration Due (25th of August)

Week 6	**Microsoft Stack Overview (SQL Server / Azure ML / Power BI)**
	Little, J. D. (2004). Models and managers: the concept of a decision calculus. Management science, 50(12_supplement), 1841-1853.
	Getting Started with Microsoft Azure
	Microsoft Course on Azure Data Factory
	What is Microsoft Azure SQL Server? / Data Storage on Azure
	Using Machine Learning and SQL Server

Mid-Semester Break

Week 7	**Microsoft Stack (Power BI)**
	Advanced Course on Power BI
	Iteration 3: MSAS/OSAS Iteration Due (22nd of September)

Week 8	**Microsoft Stack (Azure ML)**
	Machine Learning Overview
	Azure ML Basics
	Practical Azure ML Experiment / Comparing Regressors on Azure ML

Week 9 **Week 10** **Week 11**	**Big Data (Hadoop with MapReduce and HDInsight)**
	What is Hadoop? / What is Hortonworks Sandbox?
	What is MapReduce? / Basic MapReduce Tutorial
	What is HDInsight?
	Microsoft Course on Big Data Analytics with HDInsight
	Iteration 4: MSAS/OSAS Iteration Due (13th of October)

Week 12	**Assignment Assistance**
	Iteration 5: BDAS Iteration (24th of October) **AND** **Research Paper Due (27th of October)**