Module specification

1. Factual information			
Module title	**TM351: Data management and analysis**	**Level**	**3**
Module tutor	TBA	**Credit value**	**30**
Module type	Taught	**Notional learning hours**	**8**

2. Rationale for the module and its links with other modules

Data management and analysis (TM351) – an overview of the concepts, techniques, and tools of modern data management and analysis. The requirements of data management continually evolve. Recently those requirements have surpassed the capabilities of traditional data management. So, in order to better prepare our graduates for the new demands of the job market, it is necessary to address both the traditional concepts of data management as well as the increasingly important area of data analytics.

3. Aims of the module

This module aims to address some of the key concepts required for the traditionally important area of data management, and the increasingly important area of data analytics. The module will compare traditional relational databases with an alternate model (a NoSQL database), and will enable students to choose between the alternatives to select an appropriate means of storing and managing data, depending on the size and structure of a particular dataset and the use to which that data will be put. Students will be introduced to preliminary techniques in data analysis, starting from the position that data is used to answer a question, and introduced to a range of data visualisation and visual analysis techniques that will instil an understanding of how to start exploring a new data set.

To ensure that students are comfortable with handling datasets, they will explore a range of openly licensed real-world datasets (either downloaded from their host websites, or provided as snapshots) to illustrate the key concepts in the course. Sources such as data.gov.uk, the World Bank, and a range of other national and international agencies will be used to provide appropriate data. The module will aim to divide approximately equally between issues in data management (technical and socio-legal issues in storing and maintaining datasets), and issues in data analytics (using data to answer questions). Students are not expected to have a background in statistics, but should be comfortable working with mathematical concepts and will need to be competent programmers.

The module will be framed around a narrative that looks at how to manage and extract value and insight from a range of increasingly large data collections. At each stage, a comparison will be drawn between different ways of representing the data (for example, using different sorts of charts or geographical mapping techniques), and limitations of the mechanisms presented. To enable students to get a feel for the use of data, each stage will also include an overview of some data analysis techniques, including summary reporting and exploratory data visualisation.

3. Aims of the module

The module will be driven by Richard Hamming's famous quote: The purpose of computing is insight, not numbers.

Some of the key ideas are:

- Introducing data analysis. Starting with a text based data file such as comma separated variable (CSV) document, this unit will provide a brief introduction to some basic operations on simple data files. This will give an opportunity to provide an outline of the key ideas in the module, to ensure that the students have installed the module software correctly, and to begin to familiarise themselves with that software.
- Concepts in data management. The module will look at three key areas in data management: data architectures and data access (CRUD), data integrity, and transaction management (ACID). Each of these will be illustrated using a relational database, and one non-relational alternative, and the advantages and limitations of each model discussed.
- Legal and ethical issues. The module will consider the legal and ethical issues involved in managing data collections. Students will be required to obtain and read (parts of) the Data Protection Act and the Freedom of Information Act, and demonstrate how these apply to issues in data management. They will also consider privacy, ownership, intellectual property and licensing issues in data collection, management, retrieval and reuse.
- Concepts in data analytics. These sections will focus on using data to answer a real question; the focus will be on exploratory techniques (such as visualisation) and formulating a question into a form which can realistically be answered using the data that is available. Issues in processing techniques for large and real-time streamed data collections will also be addressed along with techniques and technologies (such as mapreduce) for handling them. This part will use a statistical package such as the python scientific libraries and/or ggplot to visualise the data and carry out appropriate analyses. It is not anticipated that students will need to understand statistical methods in depth.

4. Pre-requisite modules or specified entry requirements

Students should have completed Module M269 & M251 before taking this module.

5. Intended learning outcomes	
A. Knowledge and understanding	**Learning and teaching strategy**
Upon completing this course, students will be able to: **A1.** Discuss and describe the similarities and differences between at least two different database models, and how they are used to manage data collections. **A2.** Identify and explain the legal issues surrounding data collection, usage and retention. **A3.** Explain the stages and process of database design	• Lectures • face-to-face teaching • Take-home assignments • Final project

B. Cognitive skills	**Learning and teaching strategy**
Upon completing this course, students will be able to: **B1.** Select an appropriate database model for a data collection. **B2.** Use data to answer a practical question. **B3.** Analyse a simple scenario to produce a conceptual model.	• Lectures • face-to-face teaching • Take-home assignments • Final project

C. Practical and professional skills	**Learning and teaching strategy**
Upon completing this course, students will be able to: **C1.** Use a query language to extract information from a database.	• Lectures • face-to-face teaching • Take-home assignments

C. Practical and professional skills	Learning and teaching strategy
C2. Use a statistical package to explore a data set **C3.** Present an analysis of a dataset to a variety of audiences.	• Final project

D Key transferable skills	Learning and teaching strategy
Upon completing this course, students will be able to: **D1.** Write a report detailing a systematic approach to analysing a data set. **D2.** Gain Active listening to the stakeholders regarding their data analysis needs **D3.** Communicate the results of data analysis to stakeholders at appropriate level	• Lectures • face-to-face teaching • Take-home assignments • Final project

6. Indicative content.

- Introducing data management and analysis
- Acquiring and representing data
- Data preparation
- Data analysis
- Presentation: telling the story
- With data comes responsibility
- When spreadsheets fail
- Databases
- Relational databases I
- Normalisation
- Relational databases II
- Data integrity

6. Indicative content.

- Alternatives to relational databases
- Document databases
- Complex queries and analysis
- Scaling out: replication and sharding
- Distributed transactions
- Sample data investigation and report
- Data warehousing
- Data mining I: classification tasks
- Data mining II: clustering tasks
- Data mining III: semi-supervised techniques for investigating a data warehouse
- With data comes responsibility II: keeping data safe
- Linked Data on the Semantic Web
- Querying and manipulating linked data
- Applications of Semantic Web data
- End-of-module project: data investigation and report
- Finish and submit project

7. Assessment strategy, assessment methods and their relative weightings

One TMA (could be broken into a number of quizzes and/or computer-based exercises) – 20%

One MTA – 30%

One Final (in the form of a project) – 50%

8. Mapping of assessment tasks to learning outcomes

	Learning Outcomes											
Assessment tasks	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**D1**	**D2**	**D3**
TMA	x	x	x	x	x		x	x	x			
MTA			x	x	x	x	x				x	x
Final			x	x	x		x	x	x	x		

9. Teaching staff associated with the module
Name and contact details
TBA

10. Key reading list				
Author	**Year**	**Title**	**Publisher**	**Location**
Open University Course Materials for TM351	2016	M351: data management and analysis	Open University	UK

11. Other indicative text (e.g. websites)

- Download Oracle VM VirtualBox
- Technical Articles About Oracle Virtualization Technologies
- Documentation (pdf)
- VirtualBox page on OTN
- How to Look Smart At Cocktail Parties
- Welcome to Python.org
- The IPython Notebook — IPython