

# FOREWORD

---

This is the ninth in the series of highly successful international workshops on the Teaching, Learning and Assessment of Databases (TLAD 2011), which once again is held as a workshop of BNCOD 2011 - the 28th British National Conference on Databases. TLAD 2011 is held on the 11<sup>th</sup> July at Manchester University, just before BNCOD, and hopes to be just as successful as its predecessors.

The teaching of databases is central to all Computing Science, Software Engineering, Information Systems and Information Technology courses, and this year, the workshop aims to continue the tradition of bringing together both database teachers and researchers, in order to share good learning, teaching and assessment practice and experience, and further the growing community amongst database academics. As well as attracting academics from the UK community, the workshop has also been successful in attracting academics from the wider international community, through serving on the programme committee, and attending and presenting papers.

Due to the healthy number of high quality submissions this year, the workshop will present eight peer reviewed papers. Of these, six will be presented as full papers and two as short papers. These papers cover a number of themes, including: the teaching of data mining and data warehousing, databases and the cloud, and novel uses of technology in teaching and assessment. It is expected that these papers will stimulate discussion at the workshop itself and beyond. This year, the focus on providing a forum for discussion is enhanced through a panel discussion on assessment in database modules, with David Nelson (of the University of Sunderland), Al Monger (of Southampton Solent University) and Charles Boisvert (of Sheffield Hallam University) as the expert panel.

We would like to thank members of the programme and steering committees for their reviews and their continuing support of this workshop. Many members have been involved in the workshops since the first TLAD, thus showing the strength of the database teaching community both within the UK and overseas. We would also like to thank the BNCOD steering and programme committees, who have ensured that once again TLAD has its place in BNCOD this year. Finally, we express our appreciation to the Higher Education Academy, especially Karen Fraser, for her continuing support and efforts in assisting with the organization of the workshop past and present, and for the continued support by the HEA which is crucial to the success and future growth of TLAD.

Petra Leimich  
Workshop Chair

## STEERING COMMITTEE

Petra Leimich (University of Abertay Dundee) Workshop Chair

David Nelson (University of Sunderland)

Richard Cooper (University of Glasgow)

Anne James (Coventry University)

Alastair Monger (Southampton Solent University)

## PROGRAMME COMMITTEE

Fang Fang Cai (London Metropolitan University)

Jackie Campbell (Leeds Metropolitan University)

Karen Davis (University of Cincinnati)

Mary Garvey (University of Wolverhampton)

Alasdair Gray (University of Manchester)

Petra Leimich (University of Abertay Dundee)

Alastair Monger (Southampton Solent University)

David Nelson (University of Sunderland)

Clare Stanier (University of Staffordshire)

## **HIGHER EDUCATION ACADEMY**

Karen Fraser (University of Ulster)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

# CONTENTS

---

## FULL PAPERS:

- Teaching Oracle Data Miner Using Virtual Machine. 1  
*Qicheng Yu, Preeti Patel (London Metropolitan University)*
- Data Mining Project: A critical element in Teaching Learning and assessment of a data mining module. 9  
*Hongbo Du (University of Buckingham)*
- The teaching of: Cloud Computing & Databases. 21  
*Mark Dorling (Langley Grammar School)*
- Inquiry Based Learning Applications For Computing And Computing Forensic Students. 31  
*Jackie Campbell, Sanela Lazarevski (Leeds Metropolitan University)*
- Using Web-Enabled Mobile Phones for Audience Participation In Database Lectures. 37  
*Craig McCreath, Petra Leimich (University of Abertay Dundee)*
- Using Video to Provide Richer Feedback to Database Assessment. 45  
*Howard Gould (Leeds Metropolitan University)*

## SHORT PAPERS AND EXTENDED ABSTRACTS:

- Teaching Database for the Cloud(s). 49  
*Clare Stanier (Staffordshire University)*
- Making Data Warehousing Accessible. 55  
*Tony Valsamidis (University of Greenwich)*

## PANEL DISCUSSION:

- Assessment in Database Modules. 59  
*David Nelson (University of Sunderland), Al Monger (Southampton Solent University), Charles Boisvert (Sheffield Hallam University)*



# TEACHING ORACLE DATA MINER USING VIRTUAL MACHINE

Qicheng Yu  
Faculty of Computing  
London Metropolitan University  
166-220 Holloway Road London  
[q.yu@londonmet.ac.uk](mailto:q.yu@londonmet.ac.uk)

Preeti Patel  
Faculty of Computing  
London Metropolitan University  
166-220 Holloway Road London  
[p.patel@londonmet.ac.uk](mailto:p.patel@londonmet.ac.uk)

---

## ABSTRACT

*Oracle is a popular commercial database product which has been widely adopted in the teaching of a database related curriculum. Different Oracle packages or versions are required by various courses and other applications within the university environment. However, due to the complexity of Oracle package installation and maintenance, any attempt to install new versions of the Oracle package together with old versions and other software packages could easily cause massive conflicts in the network and operating system. Information Systems and Services (ISS) departments have been unable to provide adequate support for installing multiple or new versions of the Oracle package, which are required to teach students cutting edge practical skills for future employment.*

*We report on teaching experiences of using Virtual Machine to overcome constraints posed by a centralised Oracle system administered by ISS. We also present three of the most widely available Virtual Machine platforms and highlight the benefits and issues of using Virtual Machine.*

## Keywords

*Virtual machine, data mining, Oracle database, teaching*

## 1. INTRODUCTION

London Metropolitan University is one of the largest universities in London and currently offers 400 courses. Oracle is a popular commercial database product and it has been widely adopted in the teaching of a database related curriculum at the university. In the Faculty of Computing, we make use of various Oracle packages to support our teaching, which include Oracle SQL\*Plus, Oracle Designer, Oracle Forms and Reports, Oracle Data Miner, Oracle Warehouse Builder. The Oracle DBA at the university uses a centralised Oracle system to provide Oracle support for teaching as well as some of the University's management (MIS) applications.

Data mining as a specialised course continues to play an important part in the Computing curriculum. In order to teach and equip students with modern professional skills for future employment, there is a strong demand for newer versions of Oracle Data Miner packages. However, due to the fact that older versions of the Oracle system are being used by other departments and other courses and also by the university's management applications, the Information Systems and Services (ISS) department is not able to support the new version of Oracle Data Miner so as to avoid the ensuing conflicts with other packages and applications.

### 1.1 Oracle Data Miner

Our motivation for exploring the use of VM is the ever-increasing need to teach and equip students with relevant professional skills in preparation for future employment. The following is a brief description of the functionality of various versions of Oracle Data Miner and is intended to highlight the necessity of using the most current versions of packages whenever possible.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

Oracle Data Miner (ODM) is a set of data mining tools embedded in the Oracle Database that enable users to discover new insights hidden in data. It helps find patterns in data, identify key attributes, discover new clusters and associations, and discover insights which aid prediction and recommendation. It provides a collection of in-database data mining algorithms that solve a wide range of business problems. As it is built on Oracle technology, it has been well integrated with Oracle SQL, Oracle Data Warehouse Builder, and Oracle Business Intelligence. Also since the data, models and results remain in the Oracle Database, security is maximised, data movement is eliminated and information latency is minimised.

Oracle Data Miner was first released in 2002 and became one of the leading data mining tools in 2004 based on MET Aspectrum<sup>SM</sup> evaluation [1]. To maintain this leading position, Oracle Data Miner has undergone several developments. In the first version, only two mining algorithms were supported. In 2004, Oracle Data Miner 10gR1 supported eight algorithms and provided a graphical user interface (GUI). In 2006, Oracle Data Miner 10gR2 supported ten algorithms and provided an enhanced GUI. The 2007 version introduces Automatic Data Preparation (ADP) and two new Generalised Linear Models (GLM) packages.

More recently, Oracle released Oracle Data Miner 11gR2, in which Oracle Data Miner GUI is an extension to Oracle SQL Developer 3.0 that enables data analysts to work directly with data inside the database, explore the data graphically, build and evaluate multiple data mining models, apply Oracle Data Mining models to new data and deploy Oracle Data Mining's predictions and insights throughout the enterprise. Oracle Data Miner work flows capture and document the user's analytical methodology and can be saved and shared with others to automate advanced analytical methodologies. [8]

## 1.2 Virtual Machine

Virtual Machine (VM) allows the running of multiple operating systems or platforms at the same time on the same physical computer. This means that it is possible to use VM to run any version of Oracle Data Miner independently without worrying about the side effects on other packages or systems.

The concept of VM has existed since the 1960s [3], however, it has recently become increasingly popular as well as being utilised more innovatively. This popularity is due to the maturity of the technology, the advancements in computing power and freely available products from trusted vendors. There are literally dozens of VM products available on the market to choose from, examples include Virtual PC from Microsoft [7], VirtualBox from Oracle [12], and VMware Player from VMWare [13].

VM is commonly used to test system and application software under multiple operating systems or operating system versions on a single physical machine. VM has been increasingly attracting attention in the academic computing community and has been used successfully for the teaching of Computer Networking and Operating Systems [2][3][4]. In the Faculty, we have used VM for research and software development since 2008.

In this paper, we report our teaching experience of using VM in a formal Data Mining course. The issues we have tackled during the platform preparation and the subsequent teaching period are highlighted. In the conclusion, key benefits and issues of using VM are summarised and some recommendations of using VM are outlined. Also included are the features needed for future VM usage.

## 2. TEACHING EXPERIENCE OF USING VIRTUAL MACHINE

Data Mining is one of several specialised areas in the postgraduate Computing portfolio of courses. The course aims to equip students with technical, analytical and personal skills necessary for the graduate to demonstrate a professional attitude and work successfully in a wide range of industries and organisations. Oracle Data Miner is one of the data mining tools covered in the course for students to gain these necessary practical skills in the field.

Within a competitive employment market, educators are compelled to always attempt to teach students with the latest technologies and skills to enhance graduate employability. To overcome the resource confliction issues of providing new versions of Oracle Data Miner in the university's central system, we proposed to use VM for teaching purposes. The proposal was supported by the Faculty of Computing as well as ISS.

## 2.1 Management issues of Virtual Machine platforms

Before system installation could commence, some management issues of the VM ware environment were raised:

### 1) Operating system license

A virtual machine environment is viewed as an individual machine and it does not have an embedded operating system. Therefore, in order to run a particular operating system within VM, a separate license covering the operating system may be required for each virtual machine. As a VM could be easily cloned and used outside campus, the campus license may not be able to cover the use of the operating system in VM.

Although we used to teach the Oracle Data Miner tools within the university standard Windows XP platform, to ensure that no license agreement would be breached, we changed our initial plan to use Window XP as the guest operating system for the virtual machine and instead opted for open source Linux.

### 2) Oracle Data Miner license

Oracle tools are generally available for educational purposes of development and testing but not for deployment. The latest version of Oracle Data Miner has been included in the Oracle SQL Developer 3.0 which is one of Oracle's Free and Open Source Software - software which is available without charge for use, modification and distribution. [9][11] This alleviates the need to secure permissions from students for Oracle software usage.

### 3) Storage for the virtual machine

A virtual machine requires a considerable amount of disk space, as it needs to include all necessary software packages such as OS, Oracle database with example tables, the actual data mining tools, as well as the working area. In our case, Oracle Linux OS, Oracle database with data mining example tables, Oracle Data Miner, Oracle Warehouse Builder (optional), and 5-10GB working area needed to be included in the virtual machine. Therefore the total amount of disk space required is approximately 30GB-40GB.

In order to allow students to work on any PC, the disk space needed to be allocated in a network shared area. However, it would be too costly for the university to reserve such large space in a network shared area and it would also increase network traffic which could adversely impact other users and applications.

To tackle the problem, we suggested using portable hard drives for the VM storage requirements. The device costs around £35 with an ample capacity of 250GB. After consulting with students, all students in the class were willing to purchase the devices themselves as it would be helpful for them to continue studying and practicing beyond the taught sessions. An alternative would have been for the Faculty to purchase a batch of devices which could then be loaned out and reused appropriately.

### 4) Other resource issues

Although the VM could be run in a 1 GHz or faster processor and 1GB RAM, more CPU power and memory are needed for data mining tools, which deal with large amounts of data, to achieve better performance.

We experienced problems with a lab of older PCs which had only 1GB memory. With the university's support, we were able to change to a lab with higher specification PC, which has 3GB memory and Dual Core 2.4Khz CPU. As a result, the system on the virtual machine is running very smoothly. As VM only takes memory locally, the data mining performance is much better than that which we used to experience running with the centralised Oracle system. The memory space is released as soon as VM is shut down.

### 5) Virtual Machine platform is not portable

In general, a Virtual Machine is movable with a portable hard drive, however, the Virtual Machine platform is not. This means that we need all PCs in a lab to run Virtual Machine with the Virtual Machine platform installed. Although the installation of the Virtual Machine platform is a relatively straightforward process, the fact that software installation permissions are only granted to ISS staff, we have to request ISS to help us to setup the platform. Although the VirtualBox has a portable version of VM platform, our testing showed that

administrator rights and the installation of some system files is still a requirement. It would be extremely helpful if the Virtual Machine platform vendor could make the product fully portable in the future.

## 2.2 Virtual Machine Platform Options

VM has become increasingly popular recently due to the maturity of the technology, the advancement of computing power and freely available products offered by vendors. There are more than 50 products available to choose from. As Windows XP is the client platform used for students in our university, the choices are focused on the three most popular, freely accessible Windows XP supported VM platforms: Virtual PC from Microsoft, VirtualBox from Oracle, and VMware Player from VMWare.

Virtual PC is a virtualization program for Microsoft Windows [7], which allows running multiple operating systems at the same time on the same physical computer. It is free for personal/academic use from Microsoft and is easy to install. Unfortunately, the latest version of the software supports only Windows 7. In order to use the software in Windows XP, a downgraded version Virtual PC 2007/2004 is needed. The main weakness of this version is the lack of ability to redirect USB devices to the guest machine and copy and paste between the host and the guest machine.

VirtualBox was created by Sun and now owned by Oracle [12]. Similar to Virtual PC, VirtualBox is also capable of running multiple operating systems at the same time on the same physical computer and is free for personal/academic use. Oracle also offers Oracle VM VirtualBox Pre-built Database App Development package which includes Oracle Linux 5, Oracle Database 11g, Oracle SQL Developer, and many other tools. However, the VirtualBox can be difficult to use is not particularly user friendly. It supports USB devices, shared folders, and copy and paste between the host and the guest machine, but they are not working smoothly.

VMware Player is provided by VMWare [13]. It is another alternative for running multiple operating systems at the same time on the same physical PC. VMWare has made it free for personal/academic use since version 3.0. It is the most user friendly VM platform with good performance. It fully supports USB, share folder and copy and paste between the host and the guest machine.

All three virtual machine platforms could be used for teaching and learning purposes. The VM could easily convert among three virtual machine platforms. VirtualBox could be a better option for Oracle Data Miner if its usability were improved. Overall, VMware Player was considered the best, and we chose it over the others for our teaching requirements.

## 2.3 Software installation

We need to install the following software packages in the VM for the delivery of our course: Oracle Linux OS, Oracle database with data mining example tables, Oracle Data Miner, Oracle Warehouse Builder (optional).

Oracle Linux OS is quite straightforward process when following the Oracle-on-Linux Installation Guides [10]. However, the Oracle database system (Oracle 11g) installation on Oracle Linux OS, is not as easy as installing Oracle 11g on Windows XP. It would be very helpful if Oracle were to integrate their products in this context better in the future. Nevertheless, the installation was completed with help of some online guidelines [6]. Oracle Warehouse Builder (optional) is part of Oracle 11g package and will be included if it is not explicitly excluded.

As Oracle Data Miner is a Java (version 1.5 or above) based software package, installation is as simple as unzipping a folder. However, we did experience problems initially because the default Java version of Oracle Linux OS is older than the latest recommended version of Java (which is freely downloadable from Oracle [5]). It is important to mention that the Java installation will automatically overwrite any older Java version and not allow both versions to co-reside in the system. Therefore it is vital to check versions, as shown in Figure 1, and issue the following command to select the correct version:

```
sudo update-alternatives --config java
```



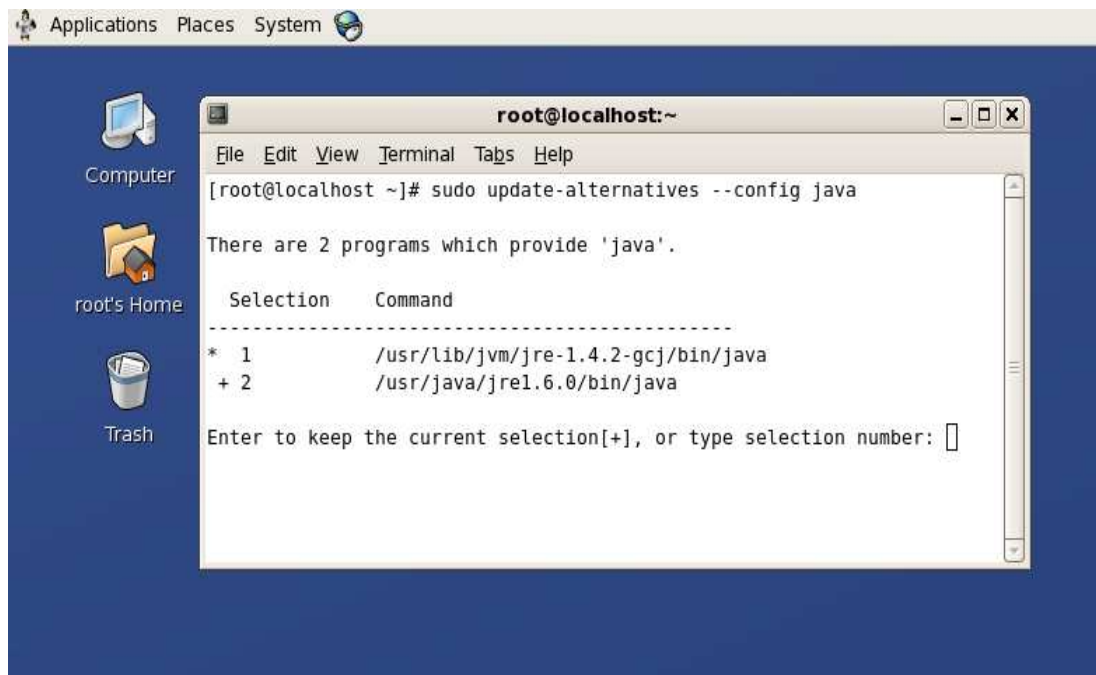


Figure 1. Selecting the correct version in the VM

## 2.4 Running Oracle Data Miner in VM

Once the installation is complete, Oracle Data Miner in VM runs extremely smoothly and the performance is even better than when it runs in the centralised Oracle system. In addition, there is no problem of Oracle memory shortage which we used to have when mining a large data set. As VM is able to remember the environment state when it is suspended, it is convenient for students to work continuously, whilst still being able to take breaks when necessary. Figure 2 shows the Oracle Data Miner's explain results screen on Oracle Linux in the VM.

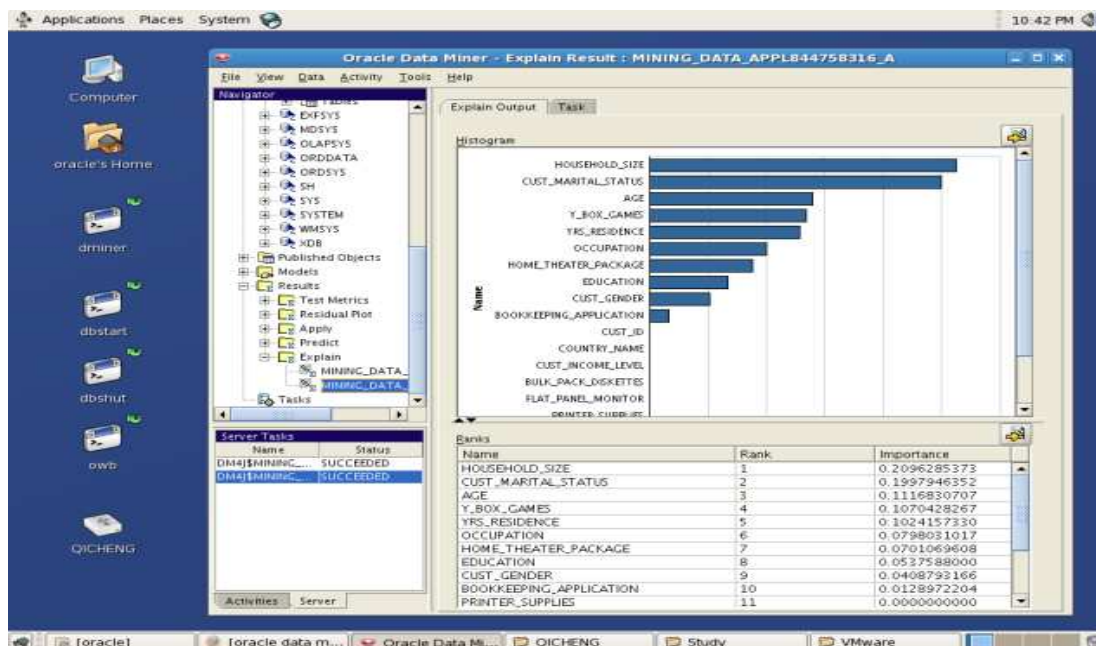


Figure 2. Data Miner - explain results

However, when the VM is completely powered off, it will automatically force the Oracle database and its listener to shutdown. It is therefore essential to start up the Oracle database and listener successfully before running Oracle Data Miner.

- **Start up Database**

```
$ sqlplus / as sysdba  
SQL> startup;  
SQL> exit
```

- **Start Listener**

```
$ lsnrctl start
```

During the teaching period, a few students had the problem of losing control of the mouse within a VM screen when the VM restarted. As a result, the VM had to be powered off and restarted to regain mouse control. In this case, students had to use the above code to restart the Oracle database and listener.

### 3. OUTCOME

After some initial trial and error, the course teaching team was able to make available to all students a VM environment.

#### 3.1 Student Feedback

For many students in the class, this was their first experience of using VM. The feedback received from students is enthusiastic and positive. They found that the VM assists their study as they are able to have full control of the Oracle system, which helps them to gain a better understanding of the Oracle database management system. Students appreciated the fact that they were able to practice and have exposure to the latest version of Oracle data mining tools. Although learning Linux OS was beyond the scope of the course, students felt that it was good to know the basics of Linux OS as it might be useful for their future careers. Students also enjoyed the VM feature of easy backup and convenient suspend function, which allows them to work continuously after restarting VM at any time. All these positive affects would not be possible under a centralised Oracle system which is under the management of ISS. Within the VM, students are able to control and manage all aspects of the environment themselves, thereby acquiring additional knowledge.

Currently the main disadvantage of VM is that the system is not portable. The students were constrained by not being able to work in those labs where the VM platform was not available. Additionally the system appears not to be robust enough, and the students were frustrated when they lost mouse control on several occasions.

#### 3.2 Benefits

The Faculty of Computing, the course teaching team and the students are able to gain many benefits from the use of VM platforms. These benefits include:

- Resolution of the resource conflicts issue to enable the use of up to date tools for teaching purposes. We have already planned to use this mode to teach an Oracle DBA certification course to allow students to install and manage the Oracle system themselves.
- Time-constrained students gain greater flexibility as they can readily replicate the environment any where.
- Teaching teams are more autonomous as reliance on ISS support is reduced. This could also significantly reduce the workload of ISS.
- System maintenance is enhanced, as once system installation is satisfactorily completed, it is possible to clone a new system simply by copying the existing VM. Since the installation of Oracle packages is not a trivial task, the cloning is very helpful. The same method could also be used to backup the system.
- Opens up the possibility of utilising and comparing of tools on different operating systems; something which would not be possible without the ability of VM to run multiple operating systems at once.

- Overcoming barriers such as the potential side effect on other systems, not being able to install in a standard environment and students needing additional rights to control the system environment.
- It is scalable across a range of subjects, including non-Computing subjects.
- Possibilities for varied assessment materials.

## 4. CONCLUSIONS

We have utilised VM for Oracle Data Mining teaching successfully and our students have had an enhanced learning experience from it. Our experience has highlighted several short-term and long-term benefits which can have a favourable impact not only on database teaching but also on the institution as a whole.

However, to utilise a VM platform, some important issues mentioned in the previous sections should not be ignored. Here is the list to be considered:

- Licensing issues – check the permission of software licenses,
- Where to store VM and who will bear the cost,
- PCs used for VM need to be powerful enough,
- VM platform needs to be installed in advance.

We believe that the flexibility that VM gives can be harnessed and utilised for the teaching of many database technologies, particularly if VM platform vendors were to make their platforms fully portable.

## REFERENCES

- [1] Berger, C., "Oracle Data Mining 11g Know More, Do More, Spend Less," in Oracle White Paper, June 2007.
- [2] Bullers, J., Burd, S. and Seazzu, A. F., "Virtual machines - an idea whose time has returned: application to network, security, and database courses," in *ACM SIGCSE Bulletin*, vol. 38, pp. 102–106, 2006.
- [3] Stockman, M., "Creating remotely accessible "virtual networks" on a single PC to teach computer networking and operating systems," in *Proceedings of the 4th conference on Information technology curriculum*, pp. 67–71, 2003.
- [4] Wang, X., Hembroff, G. C. and Yedica, R., "Using VMware VCenter lab manager in undergraduate education for system administration and network security," in *Proceedings of the 2010 ACM conference on Information technology education*, pp. 43–52, 2010.
- [5] "Download Free Java Software." <http://www.java.com/en/download/index.jsp>. [Accessed: 04-Apr-2011].
- [6] "Installing Oracle 11g Release 2 Standard Edition on Oracle Enterprise Linux 6.0 (x86\_64) within 60 minutes." [https://www.oratoolkit.ch/knowledge/howto/installation/seSrv-11g-R2-on-OEL-6.0/x86\\_64.php?cookies=1](https://www.oratoolkit.ch/knowledge/howto/installation/seSrv-11g-R2-on-OEL-6.0/x86_64.php?cookies=1). [Accessed: 04-Apr-2011].
- [7] "Microsoft Virtual PC: Virtual PC 2007." <http://www.microsoft.com/windows/virtual-pc/support/virtual-pc-2007.aspx>. [Accessed: 04-Apr-2011].
- [8] "Oracle Data Miner." <http://www.oracle.com/technetwork/database/options/odm/dataminerworkflow-168677.html>. [Accessed: 06-May-2011].
- [9] "Oracle Free and Open Source Software." <http://www.oss.oracle.com/>. [Accessed: 06-May-2011]
- [10] "Oracle-on-Linux Installation Guides." <http://www.oracle.com/technetwork/topics/linux/index-099698.html>. [Accessed: 04-Apr-2011].
- [11] "Oracle Technology Network Developer License Terms." <http://www.oracle.com/technetwork/licenses/standard-license-152015.html>. [Accessed: 06-May-2011].
- [12] "Oracle VM VirtualBox | Virtualization | Oracle." <http://www.oracle.com/us/technologies/virtualization/oraclevm/061976.html>. [Accessed: 04-Apr-2011].
- [13] "VMware Player: Run Windows 7, Chrome OS - Free Download for a Virtual PC." <http://www.vmware.com/products/player/>. [Accessed: 04-Apr-2011].



# Data Mining Project: a Critical Element in TEACHING, LEARNING AND ASSESSMENT OF A DATA MINING MODULE

Hongbo Du  
Department of Applied Computing  
University of Buckingham  
Buckingham MK18 1EG  
hongbo.du@buckingham.ac.uk

---

## ABSTRACT

*Data mining has been introduced into undergraduate and postgraduate computing curricula. A module on data mining should emphasise not only the technical but also the practical sides of the subject. This paper stresses the importance of using a data mining project as a critical element of the coursework. The paper outlines the intended learning outcomes and the expectations from students undertaking such a project. The paper proposes a framework for project administration and assessment. By using a number of past projects as case studies, the paper demonstrates the project work involved and summarises good and bad experiences in running the project. The paper highlights the uncertain nature of data mining, consequent challenges and difficulties. The paper is intended to contribute towards a wider discussion over the best practices in teaching, learning and assessment of data mining.*

## Keywords

*Data mining, module project, learning, teaching, assessment.*

## 1. INTRODUCTION

Data mining is a popular and interesting subject in computing, and has started to appear in undergraduate and postgraduate computing curricula, either in the form of a full module unit or as a part of a module on business intelligence or advanced databases ([3], [7], [10]). Because of the diversity in student backgrounds and module intensions, different approaches and methods in teaching, learning and assessment have been practised.

Data mining involves not only theories and techniques of computation but also processes, tasks and trade-offs of discovery concerns. Learning the subject is not only about knowledge and understanding but also about experience and practical skills. This balance should be reflected in the intended learning outcomes, the teaching and learning strategies, and the assessment criteria for the module. The argument naturally leads to the idea of using a data mining project as a major component of the coursework. It is felt that a data mining project should play a critical role for a data mining module in the same way as a database design project does for a database module. Using data mining project for teaching, particularly in the UK, is rare although Rob and Ellis briefly mentioned using two types of projects in data warehousing and data mining [9].

Similar attempts by this author before 2000 ran into various difficulties due to lack of useful software tools, relevant guidelines and good cases of reference. In recent years, the teaching environment for data mining has been greatly improved. An increasing number of software tools have become available for tutors to select. Among them are free downloadable tools, such as Weka [12], that offer a wide range of data mining solutions with a basic but effective graphical user interface. In 2000, the Cross Industry Standard Process for Data Mining (CRISP-DM) was published [2]. For the first time, a rigorous step-by-step industrial standard methodology has been introduced to and endorsed by data mining practitioners. The methodology provides students with a complete lifecycle to mimic and a guideline for detailed actions and tasks to follow. Besides, more and more cases of successful data mining have been reported ([1], [5], [6]). These cases become good references for students in preparation for their own projects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

A data mining project is harder than a database design project due to the uncertain nature of data mining, and therefore faces its own difficulties and challenges. This paper is intended to share the author's experience in this regard. The paper is a follow-up of an early work regarding the design of a data mining module for an undergraduate computing programme [3]. A data mining project is intended as a major part of the coursework for the module.

The rest of the paper is organised as follows. Section 2 outlines a specification of a data mining mini-project. The paper then addresses related issues arising from the specification, and proposes a framework for administering and assessing various aspects of the project. In section 3, the paper uses a number of selected projects from the author's own classes as case studies and measures their successes according to the proposed framework. In section 4, the paper highlights the uncertain natures of data mining as well as the challenges and the difficulties involved, and summarises some useful lessons learnt.

## 2. DATA MINING MINI-PROJECT: A SPECIFICATION

### 2.1 Project Aim, Objectives and Scope

The data mining mini-project, referred to as *the project* hereafter, is concerned with discovering possible hidden patterns from a given data set by using a data mining software tool. The purpose of the project is to provide students with an opportunity to experience the complete lifecycle of data mining. In particular, students are required to follow the principles of the CRISP-DM methodology, define and undertake relevant tasks, exercise judgement and make justifiable decisions over relevant issues throughout the whole data mining process.

The key word here is *experience*. The project makes the students go through the mining process in practice and face the real challenge of making decisions in uncertain situations. It is unrealistic, however, to treat the project as real-life data mining and expect students handling it as professionals. Data mining is an art that requires a lot of practice to mature and master. Consequently, the usefulness of the discovered patterns are far less important than what the students learn through their experience in applying knowledge obtained from classrooms to practice. Again, an analogy to a database design project can be drawn: we are more interested in the process of developing a database than the final database product.

Because of the discovery nature of data mining tasks, the project scope must be controlled carefully. First, the project should be a joint work by a group of 2 or 3 students taken over a period of 5 to 6 weeks. The group project enables sharing of workload and at the same time encourages debates and discussions over related issues. Second, the project should normally involve *one* of three main types of data mining, i.e. classification, cluster analysis and association discovery. In the cases when more than one type is required for the intended business purpose, the number of mining tasks for each type should be limited. Doing one thing properly is always better than attempting many superficially. Comparing to real-life data mining, this project is indeed a small scale mini-project in every way.

The project suits a full module in the final year (FHEQ level 6) worth 15 to 20 units of credit. A specification for the module and its pre-requisite are presented in [3], where some elementary knowledge in probability and statistics is assumed.

### 2.2 Project Content and Deliverables

The project should concentrate on the following main stages of the data mining process:

1. Data understanding. This stage involves activities in studying the data and data backgrounds, understanding related business activities from which the data are collected, conducting exploratory data summary and outlining possible directions for discovery.
2. Data preparation. This part of the project includes tasks in preparing and formatting data records, pre-processing the data (such as discretisation, transformation, attribute selection, sampling, etc.) and if possible improving data quality for the mining purposes.
3. Data modelling/mining. This stage is concerned with selecting suitable data mining solutions, setting appropriate parameters for the solutions, observing mining results and deciding whether alternative mining solutions should be attempted, and whether any further data preparation is required before another round of mining begins.
4. Post processing. This stage of the project involves collecting results, evaluating the patterns for their objective statistical significance, attempting to interpret the patterns, and evaluating their goodness of fit to the purposes outlined in 1.

Two phases of the CRISP-DM standard shown in figure 1, i.e. business understanding and deployment, have not been mentioned. This does not mean that these two stages are not important. It is because the true and complete business context of a selected data set may not be available, and hence difficult for the project to *mock* the business reality. However, the tutor and students should seek maximal amount of information about the data set background from limited sources. Students should make effort in considering possible deployments of useful patterns given the limited understanding of the application.

The deliverables for the project include a written report and an oral presentation from each project group. The report documents details of the project work and rationales behind them at each stage of the data mining process. The oral presentation aims to outline main issues with the data set, highlight major project tasks and key findings, justify any decisions taken, and defend the project work.

## 2.3 Related Issues

A number of issues regarding the project must be addressed. First, a suitable data set should be located. Such data sets were hard to find in the early years. Since 2005, increasing numbers of data sets from the public domain and commercial sources have become available online [8]. Tutors and students should now have little trouble in finding one. The project may require use of a single data set chosen by the tutor for all project groups, or allow the groups to search and obtain a data set of their own choice. Given the time constraint and the complexity of the project, one data set should normally be used by each project group. It should not be too large in size or very high in dimensionality. A data set with hundreds or even thousands of records and tens of variables would be ideal.

One potential concern regarding data sets is permission for data use. Most data sets from the public domain come with such permission. If the data set is from the tutor's own research and consultancy, the tutor must ensure that the permission for data use for teaching purposes is given. If students locate a data set from their own sources, both students and the tutor must check if the permission has been granted.

Data mining software is another issue to be addressed before starting the project. Data mining software can be categorised into commercial systems and free downloadable tools. The commercial systems, e.g. Oracle Data Mining and SPSS Clementine, are built to cope with the workload of real-life data sets of large sizes and high dimensionality. However, these systems are often cumbersome to learn and operate and limited with few choices of alternative mining solutions. On contrast, the free downloadable tools are often lightweight, easy to learn and operate. Although many free tools cannot cope with data sets of extremely large sizes and high dimensionality, they should be sufficient for the kind of data sets for the project. Weka [12] is a free downloadable tool that has already been widely used. Its Explorer module has a simple graphical user interface through which small-scale data mining and data exploration can be performed. The Knowledge Flow module can be used for a more serious piece of data mining through carefully designed task flows. The Experimenter module allows the operator to compare performances of classification methods. Its free license allows students to install it on their own computers, overcoming a potential constraint for group work.

A single data mining tool may not always meet all the requirements of the project. A number of related tools may also be used. For example, some data summary and pre-processing may be better done using tools such as Microsoft Excel before the data set is loaded into a mining tool like Weka. All practical knowledge regarding using the mining tool is acquired through purposely designed practical classes. Practical knowledge about any other software tools can be obtained either via added sessions of the practical classes for the module or through transferable skills from early modules.

## 2.4 Administration of the Project

The administration of the project follows the data mining lifecycle as described in section 2.2. At the beginning, the tutor provides students with a specification document and perhaps gives a presentation about the project. This is followed by a process of data selection and group formation by the students. Each group has a leader. The leader role may be taken by a specific group member or played in rotation by all members

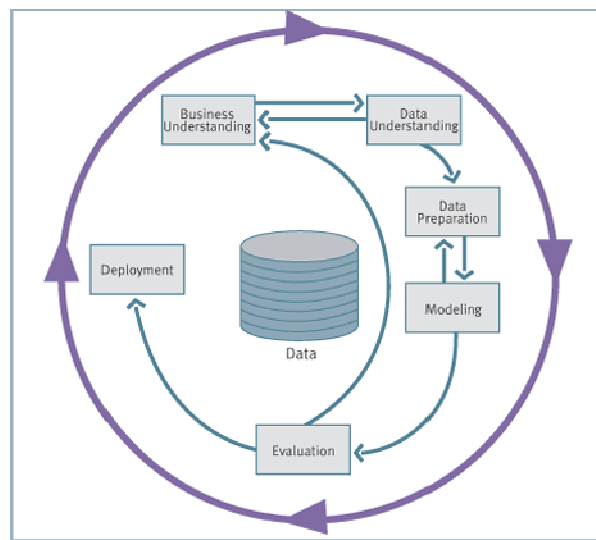


Figure 1: Phases of the CRISP-DM Process Model [2]

of the group. Each group should then arrange a start-off meeting with the tutor to gain more understanding about the background of the chosen data set and to present a project plan. During the project period, each group should hold regular short meetings with the tutor to report the project progress and discuss issues relating to the project. The tutor should by no means intervene in project decisions and activities. The tutor should play the roles of a progress monitor, a critic and a fictional client. The tutor monitors student progress closely through a sequence of small progress deliverables such as verbal reports, demonstrations, etc. By the end of the project, the reports from all groups may be compiled into a single proceeding and shared among all students of the class before the oral presentation is held.

The coverage of the key topics of the module should ideally coincide with the progression of the project. At the beginning when the project is given out to the students, the module introduces data mining concepts, principles and methodologies. The module then proceeds on topics about data exploration, pre-processing, mining and modelling techniques, and evaluation of patterns, followed by application issues towards the end. This approach will enable the students to gain knowledge on the principles and methods and then apply the knowledge in practice.

## 2.5 Assessment Framework of the Project

The project is assessed on the basis of quality of work at each stage of the data mining process. Correctness, completeness and soundness of judgement are the main factors for determining the marks. Table 1 presents a general framework for project assessment. The matrix highlights the assessment focuses by each factor at each stage. Further details for assessment can be found in the CRISP-DM methodology document [2].

| <b>Phases<br/>Factors</b>                                  | <b>Data Understanding<br/>&amp; Exploration (20%)</b>  | <b>Data Preparation &amp;<br/>Pre-processing (25%)</b>         | <b>Data Modelling &amp; Mining<br/>of Patterns (25%)</b>                                      | <b>Evaluation of Result<br/>Patterns (20%)</b>   |
|--|--|--|---|--|
| <b>Correctness</b>   | Correct understanding of data characteristics and features                                   | Correct data pre-processing and preparation operations         | Correct mining tasks, correct choices of modelling solutions. Sensible setting of parameters. | Correct understanding of evaluation metrics and correct interpretation of patterns                   |
| <b>Completeness</b>  | Coverage of aspects of data features such as data types, distributions, missing values, etc. | Sufficiency of the operations for the purpose of discovery     | Use of alternative solutions. Alternative setting of parameters. Comparison of solutions.     | Complete collection, summarisation and categorisation of patterns. Evaluation of both pros and cons. |
| <b>Soundness of Judgement</b>                              | Needs for data pre-processing  | Justification for the operations and their relevance to mining | Justification for selection of mining operations and parameter settings                       | Need for further mining. Selection/Identification of interesting patterns.                           |
| Project Planning, Execution, Management and Teamwork (10%) |  |  |   |  |

**Table 1: Assessment Framework**

Given the importance of all the stages, marks should be evenly divided. Because of the complexity and amount of time required, the Data Preparation and Pre-processing stage and the Data Modelling/Mining stage may take a marginally larger share of the total mark than the other stages. A certain small percentage of the total mark may be given to the successful planning, effective execution and management of the project, and collaborative teamwork. The distribution of percentage marks under the stage headings in table 1 is an example scheme practised by this author.

According to the framework, a project can be broadly classified into one of the following four categories:

1. **Unsatisfactory.** This kind of projects normally has major flaws in project activities at certain stage(s) of data mining. Without a clear discovery aim, random decisions are made and random actions are taken. Data are not well examined and prepared before mining. Mining solutions with default parameter settings are chosen without good reasons. The final result is a collection of irrelevant patterns with little or no interpretation. Little attention is paid to evaluation of significance of the result patterns. The poor quality of work reflects no serious attempt. The total percentage mark awarded for this category should be below the bare pass mark (e.g.40%).
2. **Fair.** This kind of projects normally produces some positive results in terms of project experience. However, the project work is not well planned. It involves either too many trial-and-error tasks or limited project activities. Students show shallow understanding of knowledge, and make decisions



without thorough consideration of the issues concerned. Some directly “copy-and-paste” style references to seen examples are made without questioning the relevance. Limited understanding about evaluation of result patterns is evident. The total percentage mark for this category is between the pass and a lower 2.II (e.g. between 40% and 54%).

3. Good. This kind of projects shows sufficient understanding and good application of knowledge in practice. The objectives are reasonably clear. The project activities are targeted rather than randomly decided. Decisions and actions taken in preparing data are thoroughly thought with good supporting arguments. The selection of mining methods is sensible. The mining tasks are well designed. Settings of relevant parameters for mining solutions are well thought. Alternative mining solutions or alternative parameter settings are attempted with justification. There are clear evidences of appreciation of evaluation of result patterns. Sensible interpretations and implications for course of actions are drawn from the mining results. The total percentage mark for this category is between a higher 2.II and a high 2.I (e.g. 55% and 69%).
4. Excellent. This kind of projects shows all the merits of the category above and furthermore demonstrates excellent performance throughout the entire data mining process. A sense of critical analysis and critical evaluation is demonstrated at every stage. There is a well-thought reasoning from a business objective to data mining tasks. The project work shows sufficient data understanding before mining. All decisions and actions in data preparation and pre-processing are supported by sensible supporting arguments. Data mining tasks are well defined and relevant to the aim. Each trial of data mining serves a clear purpose. The evaluation of resulting patterns is thorough and appropriate, and influences the selection of useful patterns for potential use. The total percentage mark for this category is a clear first (e.g.  $\geq 70\%$ ).

In practice, the marking can be done via a top-down approach from a category to percentage marks or a bottom-up approach from percentage marks to a category. In the bottom-up approach, detailed marks are first given to each stage of the project according to the factors outlined in the assessment framework. The total mark for the whole project should then reflect the appropriate category for the project. In the top-down approach, the assessor first classifies the project into one of the categories according to the category descriptors, and then fine tune the percentage marks for each stage to reflect the description of the category.

The report and the oral presentation should not be marked independently. It is suggested that a provisional mark is first given to the work described in the report. This provisional mark is either confirmed or adjusted accordingly at the oral presentation.

### 3. CASE STUDIES: THE GOOD, THE BAD AND THE UGLY

#### 3.1 Overview

The project was first introduced into a final-year module on data mining in the author's department in 2001. The project has become an established part of the coursework ever since. The weight for the project in the coursework has increased from a mere 40% at the beginning to 60% later and to the current 100% in 2010. In 2007, the author helped creating a data mining module at Sarajevo School of Science and Technology (SSST). The project was and still is used as a significant part (60%) of the coursework for that module.

Between 2001 and 2005, the author mainly relied on data sets from the UCI Machine Learning Repository [11] and a few real-life data sets from his consultancy work for the industry with the permission of use granted. Since then, data source was no longer a major concern. Before 2006, IBM Intelligent Miner for Data was used as the data mining tool. Since 2006, Weka has been used for its rich solutions, simplicity and availability for students. Microsoft Excel has also been used for assisting data exploration and data pre-processing.

Figure 2 shows the distribution of the project works in the four categories for the 35 project groups collected between 2002 and 2010 for the purpose of this paper. It is true that most of the classes are fairly small in size, and the biggest class size is 34 students (11 groups). Majority of project results fall into Fair and Good categories while Unsatisfactory and Excellent projects are in the minority. This seemingly good result reflects

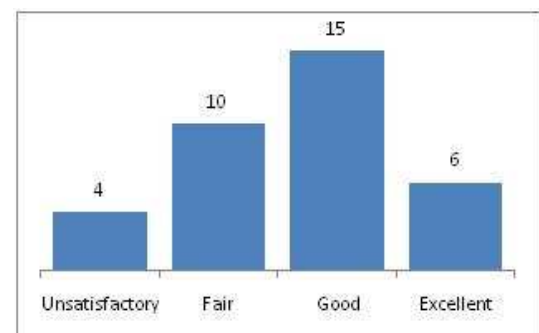


Figure 2: Distribution of Project Categories

the practice that we mark more on the project experience and what students learn from it than the final findings. This paper selects three typical projects from the 35 group projects as case studies, in order to show the nature of the tasks involved, student works produced and assessment of their works.

## 3.2 Project 1: The Bad

### The Data Set

The data set for this project is about insurance purchases from an anonymous source. It contains 14,845 recordings about purchased product type (4 types of products such as health, car, etc.), regional code (4 numeric codes plus 999 for unknown), profession (a set of categorical descriptors such as farmer, employee, etc.), age (numerical values between 1 and 95), first rate paid (Boolean values yes/no), contracting sum (numerical between 1,000 and 425,000) and closing date (month and year between 1990 and 1996). Potentially interesting patterns include a classification model regarding which type of product is purchased by what kinds of customers.

### The Student Work

The project was undertaken by a group of two students in 2010. Details of their work are listed as follows:

- Data understanding. Besides simple facts about the attribute domains and the data set size already known, virtually nothing more on data understanding was done. Some anomaly records (farmers of three years old and car insurance buyers under the legal minimum age for driving) were spotted. No discovery objectives were mentioned.
- Data preparation and pre-processing. Regional codes were replaced with nominal labels (A, B, C and D). The unknown regional code 999 was converted to unknown symbol recognisable by Weka. The anomaly records were removed from the data set. The justification given was that 21 anomalies count only 0.14% of the total number of records. The values for age attribute were discretized using an unsupervised equal-length method with default parameter setting (10 bins). The values for the contracting sum attribute were also discretized into 8 bins using the same method. Figure 3(a) shows the result of the discretisation for the contracting sum attribute.
- Data modelling/mining. One classification model using a decision tree method was obtained. The tree has an overall accuracy of 75%. Figure 3(b) presents the evaluation details of the tree. 8 quantitative association rules with support of 10% and confidence of 91% were also discovered. No explanation was given about the selection of the rules. Redundancy exists between two of the rules.
- Post-processing. Little attempt regarding evaluation of patterns was made. No clear interpretation of the patterns was given.

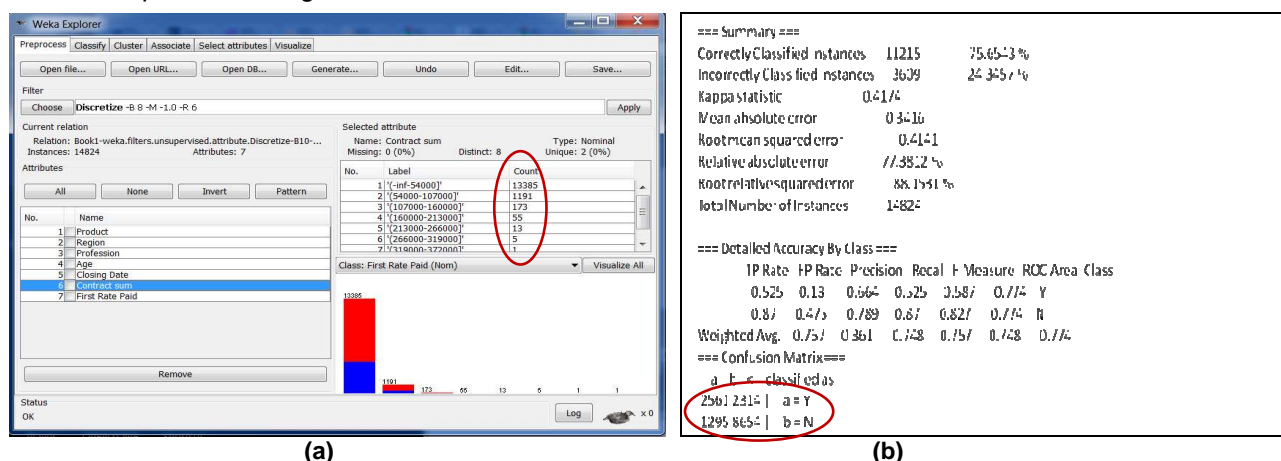


Figure 3: Some Results from Project 1

### The Assessment

The project does not outline possible directions of discovery, making the mining decisions largely a matter of trial-and-error. Understanding of data characteristics is limited, which results in the random decision over discretisation. The use of unsupervised discretisation methods is inappropriate for classification. No reasons why 10 and 8 bins are appropriate for discretising the age and contracting sum attributes were given. A clear sign of problem as indicated by the circle in figure 3(a) was ignored. Some credits should be given for the identification and removal of the anomalies. The replacement of regional codes is also plausible. The project shows serious weaknesses in the data modelling/mining stage. Only one trial of decision tree induction was

attempted without justification. The purpose of the association rules is not clear. The weakest point of the whole project is post-processing. Little attention was paid to the performances of the tree and the association rules. The students did not recognise that the tree is almost useless in classifying paying the first rate as indicated by the confusion matrix in figure 3(b). Overall, the break-down of marks is as follows: 5 out of 20 to Data Understanding, 12 out of 25 to Data Preparation and Pre-processing, 10 out of 25 to Modelling/Mining, and 4 out of 20 to Post-processing. Because of the disorganised approach to work, 3 out of 10 marks were given to the project management. With the total mark of 34%, the project is considered unsatisfactory.

### 3.3 Project 2: The Good

#### The Data Set

This project used a public domain data set about heart diseases donated by Cleveland Clinic Foundation. The data set has 303 records and 14 attributes. The attributes represent patient age (numerical), patient gender (male/female), and a range of clinic test results. The result measurements include chest pain type (4 nominal labels), resting blood pressure (numerical), amount of cholesterol (numerical), fasting blood sugar greater than 120 (true/false), resting electrocardiographic result (3 nominal labels), maximum heart rate (numerical), presence of exercise-induced angina (true/false), ST depression (numerical), slope of peak exercise ST segment (3 ordinal labels), number of major vessels coloured by fluoroscopy (numerical), thalassaemia (3 nominal labels), and angiographic disease status (healthy/disease). The number of data records is limited. Students are expected to demonstrate good use of the limited data. Potentially interesting patterns would be classification models regarding the presence of heart disease.

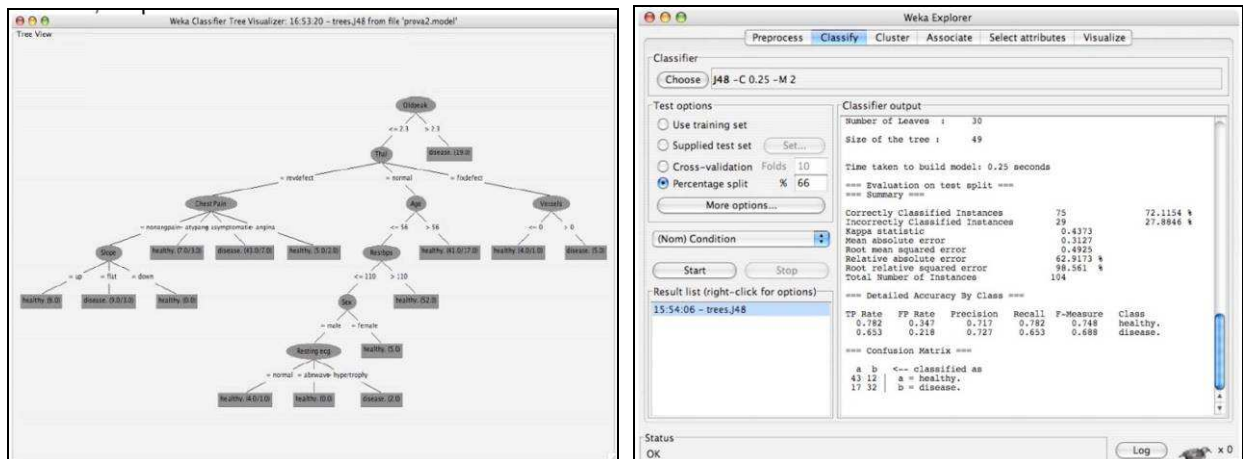
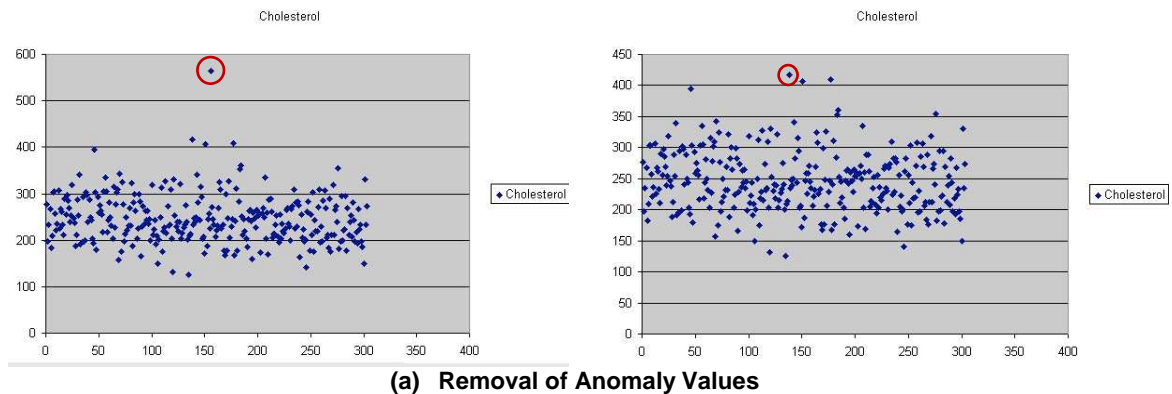
#### The Student Work

The project was undertaken by a group of three students in 2007. Details of the work are listed as follows:

- Data understanding. From the very start, the group outlined a clear business objective, i.e. finding patterns relating to the presence or absence of the heart disease. The group conducted operations such as collecting and formatting data, exploring domain types and values, obtaining descriptive statistics (only for numerical attributes unfortunately), and assessing data quality. In the end, a data description report, a data exploration report and a data quality report were produced.
- Data preparation and pre-processing. The group focused on data cleaning by removing outliers and replacing missing values with sensible alternatives. For both purposes, the students first converted ordinal and nominal values into discrete integers. To deal with missing values, the group decided to find the record's nearest neighbour and use the attribute value of the neighbour to replace the missing value. To deal with outliers, the students first plotted the data records as points in a scatter plot and manually located those anomaly values. An anomaly value was considered as being wrongly entered and hence replaced by the value of its nearest neighbour. Figure 4(a) shows the presence of an anomaly for attribute Cholesterol (the figure on the left), and the replacement of the anomaly by its nearest neighbour value (the figure on the right). After the data cleaning operations, the discrete integers for ordinal and nominal attributes were converted back to the original labels.
- Data modelling/mining. The group conducted two main data mining tasks: classification that builds a model to classify if a patient is healthy or having the disease, and clustering to profile the patients in both classes. For classification, the group used J4.8 decision tree method with different parameter settings and 2/3-1/3 split of training-testing examples as the test option. A number of possible trees with overall accuracy rates from 72% to 79% were obtained. The students realised that pruning improves the accuracy. Figure 4(b) shows one of the trees and the performance summary of one of the trees. To consolidate the finding, a similar classification task was also attempted by using the tree induction method of another tool (RDS). Some similarities in the resulting trees were found. For clustering, the k-means method was used with tweaking of different k values for good cluster quality, and eventually the optimal value for k was set to 4. Figure 4(c) presents the summary of the clusters (NB: the top line represents means or modes and the line below represents standard deviations).
- Post-processing. Both the decision trees and the clusters were evaluated when different parameter settings for the tree induction and different values for k were attempted. The students also converted the trees into rules to assist understanding. By consulting external medical experts, some of the rules made good medical sense, and supported recommendations for certain people to avoid having heart disease. The interpretation of clustering results was attempted via cluster summary and through visualising membership of the clusters.

## The Assessment

This project has paid detailed attention to every task at every stage of the data mining process. The project adheres to the CRISP-DM guideline and the tasks are performed in a systematic manner. The business objective of discovery is outlined and related to the data mining goals. Data characteristics are carefully studied and understood, but data summary is done only for numeric data. The methods for cleaning the data are sensible and justifiable. The conversion of nominal data to numerical in order to identify anomaly values is not really necessary. In order to perform data mining tasks thoroughly, the group decided to limit data mining to classification and clustering only, which is sensible. The project has shown repeated attempts for both mining tasks in order to obtain the best results. Such controlled trial-and-error activities are appropriate. However, the group did not attempt alternative classification techniques, and did not conduct a comparative study among the alternative approaches. It is also questionable whether using a 2/3-1/3 split is the most appropriate way of making effective use of the limited data. The group considered evaluation seriously, and used the evaluation results to determine a better model. However, the group did not give the presence of the heart disease higher priority and find models with better true positive performance for the presence of the disease. Consequently, 15 marks were given to Data Understanding, 18 to Data Preparation and Pre-processing, 15 to Data Modelling/Mining, and 14 to Post-processing. Because of the systematic organisation and very good documentation of work, 7 out of 10 marks were given to the project management. With the total mark of 69%, the project is standing at the border between good and excellent. The project is not considered as a clear first because of the limitations in modelling/mining and evaluation.



|                           |          |          |         |             |          |         |        |        |        |           |          |
|---------------------------|----------|----------|---------|-------------|----------|---------|--------|--------|--------|-----------|----------|
| 56.9574 male asymptomatic | 134.6383 | 252.4362 | FALSE   | hypertrophy | 133.0745 | TRUE    | 1.7883 | flat   | 1.2553 | revdefect | disease. |
| 8.1833 N/A                | N/A      | 17.9473  | 52.5968 | N/A         | N/A      | 19.9592 | N/A    | 1.1965 | N/A    | 1.0363    | N/A      |
| 53.4299 female nonangpain | 129.4766 | 250.1495 | FALSE   | normal      | 156.7944 | FALSE   | 0.515  | up     | 0.3178 | normal    | healthy. |
| 9.6789 N/A                | N/A      | 15.1902  | 53.0364 | N/A         | N/A      | 19.0707 | N/A    | 0.6951 | N/A    | 0.6813    | N/A      |
| 49.3276 male asymptomatic | 126.8448 | 226.3621 | FALSE   | normal      | 163.3793 | FALSE   | 0.3983 | up     | 0.3966 | revdefect | healthy. |
| 8.3064 N/A                | N/A      | 16.0469  | 37.3789 | N/A         | N/A      | 17.6425 | N/A    | 0.5652 | N/A    | 0.7479    | N/A      |
| 57.75 male nonangpain     | 136.1591 | 242.9091 | FALSE   | hypertrophy | 150.9318 | FALSE   | 1.375  | flat   | 0.6591 | revdefect | disease. |
| 7.0187 N/A                | N/A      | 20.1828  | 32.9685 | N/A         | N/A      | 19.8663 | N/A    | 1.0634 | N/A    | 0.9387    | N/A      |

(c) Clustering Result Summary

Figure 4: Some Results from Project 2

### 3.4 Project 3: The Ugly

#### The Data Set

The data set for this project is the same insurance data set as the one used for project 1.

#### The Student Work

The project was also undertaken in 2010. The student did not want to work with others and undertook the entire project alone against the tutor's advice. Key points of the project work are summarised as follows:

- Data understanding. No directions of discovery were outlined. The student decided to carry out a brute force bottom-up discovery of any potential patterns. At this stage, the student used Weka and Excel to gain understanding about domain types, value distributions of attributes, and unknown values. Unlike the group for project 1, this student did not identify any anomalies, but spotted that values for the contracting sum attribute were extremely skewed towards the lower end.
- Data preparation and pre-processing. Like project 1, regional codes were replaced with nominal labels (A, B, C, D and E, where E for unknown). The age attribute was discretized into more natural age groups such as child, teenager, young, adult and senior. Because of the skew of the contracting sum values, the student decided to apply logarithm transformation on the original values so that the levels of magnitude of contracting sums were used instead of the actual figures.
- Data modelling/mining. The student decided to do almost everything: classification, clustering and association mining. The student laboriously tried four methods for clustering, ten methods for classification, and two methods for association rule discovery. For clustering, different values of  $k$  were attempted for the  $k$ -means and the EM methods. For classification, the student attempted to induce classification models for product types, and used the Experimenter module in Weka to compare performances among the classification methods. Little explanation was given regarding the setting of parameters. One could only assume that the default settings might have been used. For association rule discovery, confidence and accuracy were the criteria for selecting the top 10 rules. Figure 5 shows one of the classification models as a decision tree and a snapshot of top 10 association rules discovered using the Apriori method.
- Post-processing. On clustering, the student was conscious of the value for  $k$ , and used the evaluation of cluster quality to determine the optimal value for  $k$ . However, except the performance analysis using Experimenter, very little attention was paid to the detailed performance evaluation of different classification models shown in confusion matrices. The student did not seem to notice the strength of JRip method in classifying life insurance buyers. The student did not pay enough attention to the evaluation of association rules and the meaning of the rules at all.

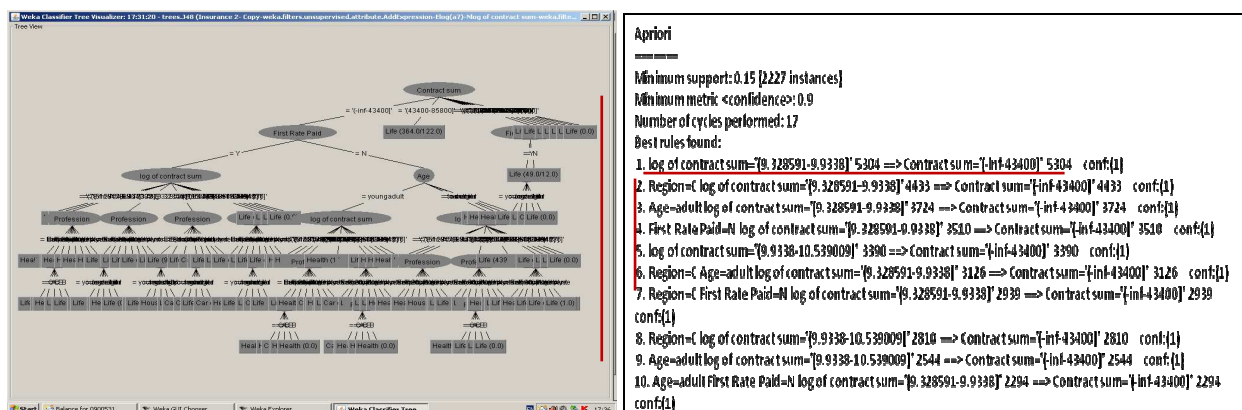


Figure 5: Some Results from Project 3

#### The Assessment

The project is a showcase of trial-and-error gone to the extreme. Many trials were made and many patterns were discovered, but collectively these patterns are not carefully examined. There are gems of good ideas here and there in some individual tasks, such as the logarithm transformation for the contracting sum attribute, the attempt to create models to predict the types of insurance product types, comparative study of techniques of classification, etc., but the project as a whole is not a piece of coherent work. The student did not realise the complexity of most decision trees, and totally ignore the inappropriate associations with contracting sums and their logarithm-transferred values appearing on both sides of the rules (marked out by lines in figure 5).



The project can only be classified as fair with a total percentage mark of 48% (10 for Data Understanding, 15 for Data Preparation and Pre-processing, 10 for Data Modelling/Mining, 8 for Post-processing and 5 for project management).

## 4. DISCUSSIONS

### Uncertainty and Difficulty

Uncertainty is the most noticeable feature of data mining projects. Given the same data set and the same objectives, there can be various ways of preparing and mining the data. Because of the uncertainty, trials of alternatives are unavoidable. The exercise of sound judgement is therefore essential. The sense of sound judgement depends on levels of understanding of knowledge, analytic skills and experience. This means that data mining projects can be seen as time consuming and difficult for beginners. However, the difficulty should not come as a surprise, nor should it suggest that the project is unsuitable for undergraduate students. Data mining should be taught in the final year of an undergraduate programme. By that time, a greater degree of academic maturity in terms of analytic skills, logical reasoning and sense of judgement is expected upon the students. With advices on directions and close monitoring of progress from the tutor, the difficulty can be overcome.

There is a genuine difficulty, i.e. the absence of domain experts from the data mining lifecycle. *Domain experts* are those who know the application domain well, and can judge which patterns are potentially useful to the business practice and which are not. The experts are in fact present throughout the data mining process in most, if not all, real-life data mining projects [4]. It is not realistic to expect the tutor to play such a role for various domains of application for all kinds of data sets. This particular difficulty could be avoided by the tutor intervening in the data set selection and only allowing students to select a data set that the tutor is familiar with the application domain. By doing so, however, the motivation of the students may be affected.

### Roles of Case Studies

Because of the uncertainty and difficulties, it is very useful and beneficial for students to study cases of reported data mining projects. These case studies bring the data mining process alive and provide students an opportunity to observe how data mining is done before they try it themselves. The value of case studies cannot be underestimated. Good case studies were rare and difficult to get. Indeed, industry and commerce often consider data mining as a closely guarded secret. However, the situation appears changing. It is now possible to find successful data mining cases both commercial and non-commercial [5]. It is recommended that a successful case be presented in teaching sessions so that every aspect of the case can be discussed thoroughly under the tutor's supervision. The author has used a good data mining project conducted by Gary Saarevirta of IBM as a case study in his recent textbook [4].

The author has also used an assignment as a way of getting good case studies. Students are asked to search for a case in a designated application area from the published sources. They are then required to study the case, present it and comment on it. This work has been proved very useful for students, and can be taken as a part of assessment. Such an assignment may also be used as a part of the coursework element for modules that only briefly cover the data mining topic.

### Levels of Expectation and Scope of the Project

The level of expectation must reflect the aim of the project, i.e. to provide students with an opportunity to *experience* data mining. It is necessary to emphasise again that the success or failure of a project should not be judged on applicability of the result patterns, but on the tasks and actions taken and their justifications.

The amount of work should justify the amount of time allocated. The project specification given in section 2 is meant for a full module on data mining in a single-honours computing programme. The scope of the project must be adjusted accordingly if the module has a different emphasis or it is for a different programme. For instance, a module on advanced databases with a significant part on data mining may require a much smaller scope. Everything within the project, such as the data set dimensionality, the amount of data preparation work and the number of data mining tasks, should all be limited. The project group size may even be increased to 4 students rather than 2 or 3 to further spread the workload.

### Student Feedback

The author has not yet conducted a systematic survey of student feedbacks on the project mainly because of small class sizes. However, from the annual student feedback questionnaires on the module where a question about the coursework is asked, comments on the use of the project are overall positive. Most students consider that the project experience is enjoyable and has helped enhancing their understanding of the subject. At the same time, many students consider data mining projects harder than other practical projects they have

done, e.g. database design project and software engineering project. Weka as a tool for supporting the project has also received good feedback comments for its simplicity and ease of use.

Some cautious feedback comments are also obtained. Due to the lack of business context and short of domain experts, some students cannot see how useful result patterns can be, which in turn affect the level of their confidence towards the patterns they have discovered. The biggest difficulty is how to deal with the uncertainty, which means they have to understand and decide what options are sensible to explore. This could be a cultural shock for computing students who have worked largely towards deterministic solutions. Some students also have difficulties in data understanding and pattern evaluation due to lack of training and understanding of basic statistics.

A small minority of students have very few clues about the project. They tend to make bad judgement and rely heavily on trial and error or hand-on guidance from the tutor. Their comments to the project are quite negative. Effort is certainly needed to find out how to minimise the size of this group.

#### Resource Implications

Running the project may require additional resources in terms of tutor's hours. Enhancing student learning experience has cost. Indeed, no pain, no gain. In the author's experience, however, the extra hours spent on administrating/supervising the project is only marginally more than that for a database design project, thanks to the academic maturity of the final-year students. The good students usually need very little guidance. They can obtain the required knowledge from lectures and reference materials. Weak and fair students in fact take up most of the consultation time. Whether the project idea can scale up to classes of hundreds of students is yet to be verified.

## **5. CONCLUDING REMARKS**

This paper argues that a data mining mini-project should become a vital part of the coursework for a module on data mining. The project should benefit student learning experience of this interesting subject in the same way as a database design project does for databases. Data mining may be more challenging to do for students and harder to manage for the tutor than a database design project, and hence might require more resources in terms of tutor's hours. Despite all kinds of difficulties, the paper argues and demonstrates that such a project should still be feasible to do by final year students of computing major. With the ambitious goal of practising data mining, the expectation of the project must be realistic: it is more about experience than discovery results.

The assessment framework and project administration proposed in this paper are only ideas that have been practised and refined by this author. It undoubtedly needs the scrutiny of the data mining teaching community. It is hoped that the criticism will lead to improvements over the management and assessment of such an important element.

## **6. ACKNOWLEDGEMENT**

The author wishes to thank all his students for undertaking the data mining mini-project in the past years, particularly those students, whose names are kept anonymous, for using their works as case studies.

## **REFERENCES**

- [1] Berry, M. J. A. and Linoff, G., *Mastering Data Mining: The Art and Science of Customer Relationship Management*, John Wiley & Sons (2000)
- [2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., *CRISP-DM 1.0: Step-by-step Data Mining Guide*, SPSS (2000)
- [3] Du, H., *Teaching Undergraduate Students Data Mining: Ideas, Experience and Challenges*, 8<sup>th</sup> International Workshop in Teaching, Learning and Assessment of Databases (TLAD), University of Abertay Dundee (2010)
- [4] Du, H., *Data Mining Techniques and Applications, An Introduction*, Cengage Learning: Andover (2010)
- [5] Kitts, B. Melli, G. and Rexer K. (eds), *Data Mining Case Studies*, Proceedings of the First International Workshop on Data Mining Case Studies, 2005 IEEE International Conference on Data Mining, Huston, USA (2005)

- [6] Luan, J. and Zhao, C-M.(eds), *Data Mining in Action: Case Studies of Enrolment Management*, Wiley Periodicals Inc (2006)
- [7] Mrdalj, S., *Teaching An Applied Business Intelligence Course*, Issues in Information Systems, Vol. VIII, No. 1, pp134-138 (2007)
- [8] Piatetsky-Shapiro, G., <http://www.kdnuggets.com/>, Accessed 30 March 2011
- [9] Rob, M. A. and Ellis, M. E., Case Projects in Data Warehousing and Data Mining, Issues in Information Systems, Vol. VIII, No.1 (2007)
- [10] The Quality Assurance Agency for Higher Education, *Subject Benchmark Statements, Computing* (2007)
- [11] The University of California at Irvine, *UCI Machine Learning Repository*, <http://archive.ics.uci.edu/ml/about.html>, assessed on 30 March 2011
- [12] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, 3<sup>rd</sup> Ed. Morgan Kaufmann Publishers (2011)



# THE TEACHING OF: CLOUD COMPUTING & DATABASES

Mark Dorling  
Langley Grammar School's Digital  
Schoolhouse  
Langley, SL3 7QS  
markdorling@lgs.slough.sch.uk  
www.digitalschoolhouse.org.uk

---

## ABSTRACT

*The Digital Schoolhouse Database Detectives lesson is aimed at Key Stage 2 pupils and based on the book Certain Death by Tanya Landman. Before completing the database lesson, the class teacher is encouraged to read the book (except the last chapter) and complete a series of numeracy puzzle challenges loosely based on the book, the answers providing pupils with the clues to question the database and identify the murderer.*

*Pupils use cloud computing technology e.g. Google Documents: Spreadsheets, to collaboratively input data about the suspects from profile cards based on the book. Pupils then perform verification on their neighbour's data entry before downloading the spreadsheet and importing it into Microsoft Access. After importing the data, pupils first use the filter tool to solve the murder using the answers from the numeracy challenges, then create a report for the Court based on a query identifying the murderer.*

*The class teacher is also provided with Literacy, PE and Art tasks to complete before reading the final chapter of the book to see if their database work is correct.*

## Keywords

*databases, spreadsheets, cloud computing, transition, creative curriculum, pedagogy.*

## 1. INTRODUCTION

### 1.1 Background

The Digital Schoolhouse (DSH) is a transition project established by Langley Grammar School (LGS) to offer predominately Year 6 pupils from local primary schools the opportunity to visit LGS for a day of specialist teaching in a dedicated ICT and Computing environment. Pupils work with a primary trained teacher who also has secondary ICT teaching and industry experience to learn new skills and concepts with a focus on how they are deployed in secondary education, the world of work and business.

There is no cost for the teaching provided at the DSH; the only cost incurred by the visiting school is the transportation to and from LGS and, where applicable, a small contribution towards the cost of any materials used. In circumstances where the visiting primary school has limited funds or a high number of free school meals, the DSH Trust would hope to support the school by contributing towards their transportation costs.

### 1.2 Supporting transition

The first priority for the project was to meet with Head Teachers, Subject Specialists, Special Educational Needs (SEN) and Gifted and Talented (G&T) Coordinators from local primary and secondary schools to gain a holistic insight into the challenges of transition from primary to secondary school. This valuable feedback helped shape the structure, content, and range of lessons e.g. databases, multimedia, etc. offered by the project.

Feedback from visiting class teachers indicated that they would like to be more involved in the lessons and use them as opportunities for their own professional development. Class teachers have been provided with the post-lesson resources and team teaching opportunities; the effect is that teachers feel confident to deliver similar lessons in school, and use future DSH booking to experience a different lesson focus the following year.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

TLAD2011

### 1.3 Resources

The DSH room is a dedicated ICT and computing suite located in the main building at LGS. The room has 31 networked computers and an interactive white board (IWB) connected to the teacher's computer at the front of the class. Also connected to the IWB is a Visualiser (Digital version of an OHP) that enables the teacher to model learning and allow pupils to display their paper-based work.

The DSH is always looking at how it can deploy new technologies to enhance the learning of pupils who attend the project. This year the DSH has introduced pupils to cloud computing to promote collaboration using Google Documents: Spreadsheets, and solve a murder mystery using a relational database software [7].

### 1.4 Structure of the day

At the start of each academic year the DSH writes to primary schools within the specified radius of LGS to invite their pupils to attend an ICT and computing day.

The DSH is open two days per week during term time. These days are set at the beginning of the year, days are chosen based on the time tabling capacity at LGS. The day starts at around 10:00am and finishes at 2:30pm. However, the DSH always tries to accommodate the needs of visiting schools.

Each DSH lesson is designed to split into three sessions which last approximately an hour to an hour and a half; between each session is a break. The DSH schedule has been deliberately matched with LGS's timetable to give pupils the opportunity to experience secondary school life.

### 1.5 The Digital Schoolhouse approach

#### 1.5.1 Creative curriculum

In September 2008, the DSH began by developing multimedia focused lessons that identified cross curricular opportunities and provided visiting schools with learning resources after their visits. However, based on lesson feedback and the proposed changes to the national curriculum in the Rose Report [6], the DSH made the decision to design any future ICT focus lesson to meet the creative curriculum needs of schools.

Since this decision, all new lessons cover a wide range of creative curriculum links to the core and foundation subjects including: literacy, numeracy, science, geography and history, art, physical education (PE), and philosophy for children.

All the cross curricular and creative curriculum links have been developed through collaboration with LGS subject specialist teachers and the library resource centre. The purpose of this collaboration was to identify possible activities and tasks that would reinforce the ICT and computing teaching whilst making the learning relevant and engaging to pupils.

#### 1.5.2 Transition

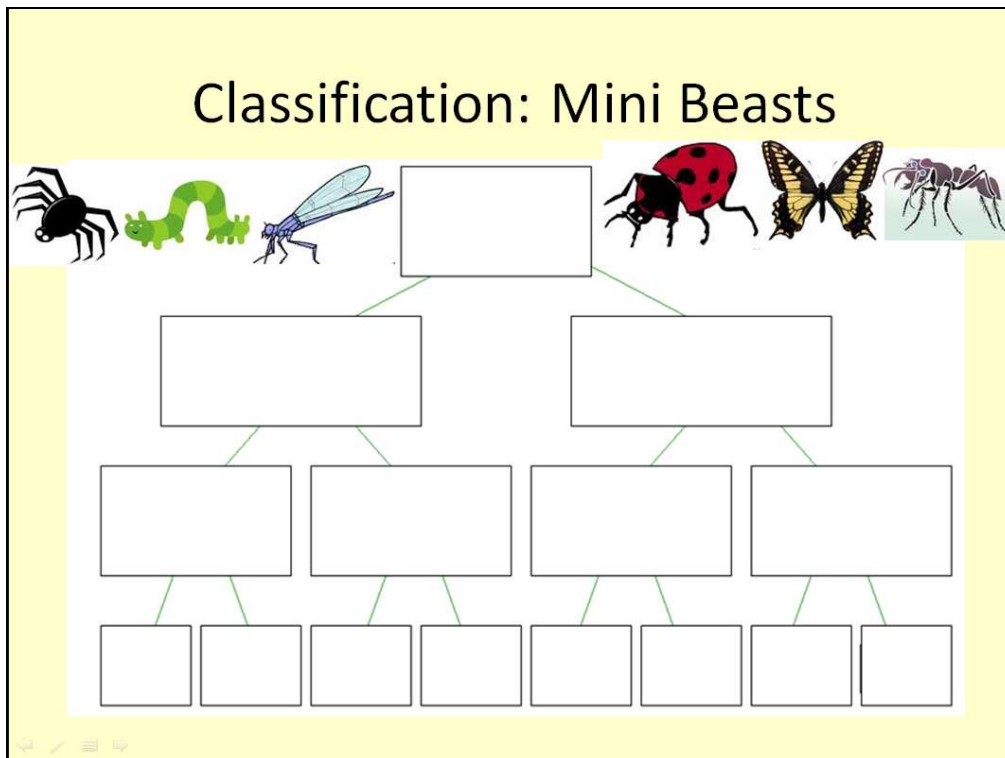
To ensure a solid foundation of ICT skills and concepts, the DSH revisits relevant previous ICT learning from Years 3 – 5. Once consolidation has taken place, the teacher develops the learning to focus on the expected objectives and outcomes from the QCA ICT Year 6 unit of work. Where applicable, the teacher builds on the learning by introducing pupils to ICT Key Stage 3 content. Merging the Key Stage 2 and Key Stage 3 curriculum frameworks provides exciting transitional opportunities, which are highlighted for the class teacher in the relevant QCA units of work downloadable from the Digital Schoolhouse website [8, 9, 10, 11, 12].

## 2. FORGET THE TECHNOLOGY, IT'S THE PEDAGOGY THAT MATTERS

### 2.1 Consolidating learning

Many of the students who visit the DSH have very little prior knowledge of databases. Therefore, we begin the day by discussing what the word *data* and *base* means to them. Most of the responses given by pupils are *data is information* which enables the DSH teacher to develop pupils' understanding of the difference between *data* and *information* through careful questioning. This key questioning focuses on pupils' understanding of the *real* world for each of the words. When this consolidation has taken place, the DSH teacher extends the pupils by asking them to combine together the meanings of each word to gain a greater understanding of what is a *database*.

The DSH teacher briefly revisits *branching databases* using the topic Mini-beasts because most schools use this Science topic as a branching databases activity i.e. *categorising* and grouping insects based on their physical features (*Figure 1*) in Key Stage 1 and Key Stage 2 [5]. The purpose of this learning is to focus the pupils on what physical features does each insect have that makes it different from others, and to ask questions where the answer is either *yes* or *no* [2].



**Figure 1: Branching database of Mini-beasts.**

Once consolidation has taken place, the teacher encourages pupils to work together using personal whiteboards to think firstly of data that we can see about each of them (mirroring what we did during the Mini-beasts exercise) and then what questions we could ask of each other. Pupils are introduced to key vocabulary *fields* being the same as column headings in a spreadsheet e.g. the different data we can see about them, and *records* being each person i.e. a row in a spreadsheet.

When pupils have completed their list of the different types of data we can see about one another e.g. eye colour, the DSH teacher asks them to consider appropriate questions with *yes* or *no* answers. The DSH teacher then asks all pupils to line up outside the classroom. The DSH teacher introduces the concept of the classroom being the database with each of them being a record in the database. Pupils are then *imported* into the database and stand behind their chairs.

## 2.2 Human database activity

The DSH teacher then questions the human database using the pupils' questions from the previous activity, pupils move around the classroom to stand in groups depending on the answer to the question. This activity helps pupils to understand that data becomes information by adding meaning through asking a question.

Boolean searches are integrated into the activity by asking the pupils how we could make searching quicker and more efficient – pupils tend to reply by combining questions so the DSH teacher asks them what connective words they use in Literacy to do this e.g. AND, OR and NOT. The DSH teacher then continues this human database activity but this time extending the pupils by using the Boolean connectors to combine questions. To get pupils really thinking about the question being asked, the DSH teachers asks for *male AND female* which usually involves a few pupils standing in the wrong place, which provides an excellent discussion point! To conclude the learning on Boolean searches the DSH teacher asks pupils to consider how the use of AND, OR and NOT affects the number of search results e.g. AND reduces number of pupils, with OR increases the number of pupils.

The human database activity is concluded with pupils developing an understanding how reports work by sorting themselves into an order based on any of the fields identified previously. This activity starts by asking pupils to sort themselves into register, then first name order, followed by date of birth, etc. To extend the pupils the DSH teacher often asks questions to group pupils e.g. gender, before repeating the sorting activity within the smaller groups (Figure 2).

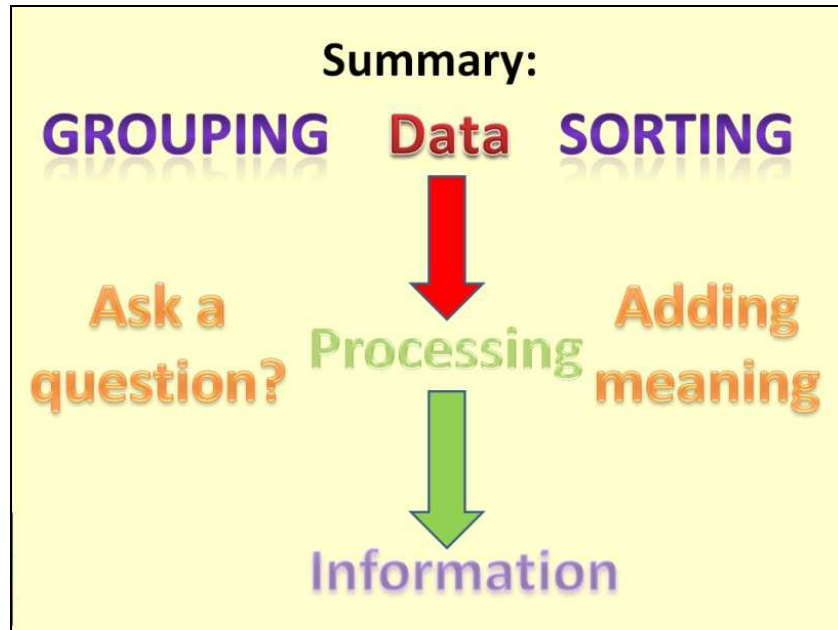


Figure 2: Summary of learning from the human databases activity.

### 3. ENHANCING PEDAGOGY WITH TECHNOLOGY

#### 3.1 Using the cloud

To get pupils up to speed on the benefits of cloud computing technology and how it works, the DSH teacher shows them a short video clip which uses simple language to explain how cloud computing works [1]. Pupils then familiarise themselves with the characters from the book *Certain Death* [4].

Pupils are given a suspect profile card (Figure 3) based on the characters in the book. Each suspect profile card has a unique identifier i.e. *suspect id* with pupils being briefly question about the benefits of using a *suspect id*.

| Suspect database   |              | Suspect number: |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
|--|--------------|-----------------|------------|-------|---------|------|------|----|-------------|--------------|---------|-------|--------|--------|------------|----|----------|---|--------------|------|--------------|------|--------------|-----------|-------------|-------|------------|---|---|
| <table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 30%;">First name:</td><td>Yuri</td></tr> <tr><td>Last name:</td><td>Mehic</td></tr> <tr><td>Gender:</td><td>Male</td></tr> <tr><td>Age:</td><td>55</td></tr> <tr><td>Occupation:</td><td>Sharpshooter</td></tr> <tr><td>Height:</td><td>Short</td></tr> <tr><td>Build:</td><td>Medium</td></tr> <tr><td>Shoe size:</td><td>11</td></tr> <tr><td>Glasses:</td><td>N</td></tr> <tr><td>Hair colour:</td><td>Dark</td></tr> <tr><td>Hair length:</td><td>Long</td></tr> <tr><td>Facial hair:</td><td>Moustache</td></tr> <tr><td>Eye colour:</td><td>Brown</td></tr> <tr><td>Ear rings:</td><td>Y</td></tr> </table> | First name:  | Yuri            | Last name: | Mehic | Gender: | Male | Age: | 55 | Occupation: | Sharpshooter | Height: | Short | Build: | Medium | Shoe size: | 11 | Glasses: | N | Hair colour: | Dark | Hair length: | Long | Facial hair: | Moustache | Eye colour: | Brown | Ear rings: | Y | <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">           Photograph:           <div style="border: 1px solid black; width: 150px; height: 100px; margin: 5px auto; position: relative;"> <span style="position: absolute; top: -10px; left: 50%; transform: translateX(-50%);">Paperclip</span> </div> </div> <div style="border: 1px solid black; padding: 5px;">           Finger print:           <div style="border: 1px solid black; width: 100px; height: 80px; margin: 5px auto; position: relative;"> <span style="position: absolute; top: -10px; left: 50%; transform: translateX(-50%);">Right index finger:</span> </div> </div> |
| First name:  | Yuri         |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Last name:   | Mehic        |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Gender:  | Male         |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Age:   | 55           |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Occupation:  | Sharpshooter |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Height:  | Short        |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Build:   | Medium       |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Shoe size:   | 11           |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Glasses:   | N            |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Hair colour:   | Dark         |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Hair length:   | Long         |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Facial hair:   | Moustache    |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Eye colour:  | Brown        |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |
| Ear rings:   | Y            |                 |            |       |         |      |      |    |             |              |         |       |        |        |            |    |          |   |              |      |              |      |              |           |             |       |            |   |   |

Figure 3: Example suspect profile card.

Pupils then select the row number in the spreadsheet that matches their *suspect id*. Pupils then begin collaboratively inputting data about the suspects from profile cards into a Google Docs: Spreadsheet [13]. The DSH teacher always tell the pupils at the start of the day that “*good computing should appear as though it is magic*” and the pupils are never disappointed to see this technology in action e.g. watching their entries and the entries of their colleagues magically appearing on the DSH teachers IWB (Figure 4).

|    | A                   | B         | C      | D  | E                | F     | G        | H  | I    | J     | K        |
|----|---------------------|-----------|--------|----|------------------|-------|----------|----|------|-------|----------|
| 1  | Yuri                | Mehic     | Male   | 55 | Sharpshooter     |       |          | 11 | N    | Dark  | Long     |
|    |                     | Bellini   | Female | 26 | Acrobat          |       |          | 5  | N    | Dark  | Long     |
|    | Marco               | Bellini   | Male   | 29 | Acrobat          |       |          | 8  | N    | Dark  | Short    |
|    | Carlotta            | Mancini   | Female | 24 | Circus Artist    |       |          | 6  | N    | Dark  | Long     |
| 5  | brady               | sparkle   | Male   | 59 | ringmaster /     | short | fat      | 9  | N    | bald  | n        |
| 6  | Italo flying        | Bellini   | Male   | 27 | Acrobat          | tall  | slim     | 10 | Y    | dark  | long     |
| 7  | Angelo              | Bellini   | Male   | 22 | Acrobat          | Tall  | Slim     | 8  | N    | Dark  | Short    |
| 8  | Draco               | Blade     | male   | 59 | knife thrower    | short | fat      | 8  | Y    | dark  | long     |
| 9  |                     |           |        |    | knife thrower's  |       |          |    |      |       |          |
|    | Ruby                | Diamond   | Female | 34 | assistant        | short | medium   | 4  | Y    | dark  | long     |
| 10 | Majestic            | Whizzbang | Male   | 62 | Magician         | Tall  | Fat      | 12 | Y    | Light | Short    |
| 11 |                     |           |        |    | Roller skating   |       |          |    |      |       |          |
|    | Zippo               | Marx      | Male   | 35 | juggler          | Med   | Athletic |    | Y    | Red   | Short    |
| 12 | Alfonso "The brave" | Romero    | Male   | 47 | Lion tamer       | Tall  | Medium   |    |      | 10    | Y        |
| 13 |                     |           |        |    | Lion Tamer's     |       |          |    |      |       |          |
|    | Alisa               | Romero    | Female | 41 | Assistant        | Short | Slim     | 5  | N    | Light | Shoulder |
| 14 | Emilie              | Trapeze   | Female | 19 | highwire artiste | tall  | slim     | 8  | none | dark  | Long     |
| 15 | Pianiot             | Petrov    | male   | 32 | Clown            | tall  | athletic | 11 | N    | red   | short    |
| 16 | Columbine           | Petrova   | Female | 32 | Clown            | Tall  | Medium   | 10 | Y    | Dark  | Long     |
| 17 | Rudi                | Petrova   | Male   | 37 | Clown            | short | fat      | 6  | Y    | bald  | N        |

Figure 4: Example of collaborative data entry from the suspect profile cards.

When this is completed many of the pupils acknowledge that they have made a mistake in entering the data so the DSH teacher plays a short game with the pupils to illustrate the importance of data verification and that computing specialists use what they know about the *real* world when verifying data in a database.

Pupils perform data verification on their neighbour's data entry by swapping profile cards and correcting any data incorrectly entered. The final stage in using cloud computing involves downloading the spreadsheet into their user area.

### 3.2 Combining the cloud with de-facto software

The DSH teacher takes the pupils through the importing process in Microsoft Access. After successfully importing the data [14], pupils work with the DSH teacher to use the clues from the book with the filter tool to identify the murderer [15].

If there is time and to encourage independent learning, the DSH teacher shows pupils how to use pre-prepared independent learning videos (Figure 5) e.g. play, pause, rewind and fast forward. These videos model the tasks for *creating a query* [16] and then *creating a report based on a query*.

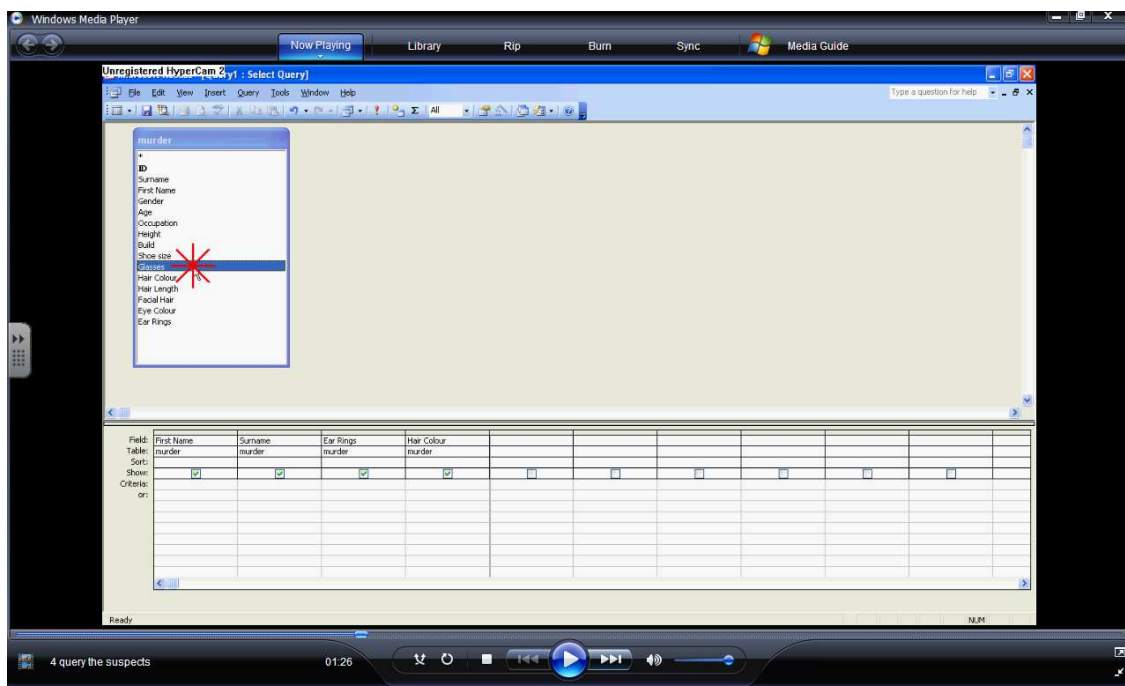


Figure 5: Example of an independent learning video.

When the pupils return to school, the visiting class teacher is encouraged to read the last chapter of the book to see if their Maths and ICT have led them to correctly solving the murder.

## 4. FEEDBACK FROM THE LESSON

The DSH values feedback received from all visitors. Pupils use a free online audio-enabled questionnaire consisting of 12 closed questions and a comments box. This feedback provides the DSH with a valuable insight into the views of the pupils and allows us to focus on the areas for development.

The teacher's questionnaire is a side of A4 paper and currently includes ten open questions. These questionnaires are given to each member of the teaching staff at the start of the day. If the visiting class teacher is an experienced member of staff or a subject specialist, he or she is encouraged to perform a slightly more formal observation during the day.

The Database Detective lesson was first trialed by Year 6 staff and pupils from St. Peter's Church of England Combined School in Burnham; some of their comments have been recorded below.

### 4.1 Teacher feedback

*"I was completely inspired by the day, I couldn't believe how much could be covered in just one day! Our children went from having never seen a database to inputting data, using search criteria in certain fields and picking up the correct computer vocabulary. Their behaviour all day was immaculate which was a big surprise as this particular class had a lot of challenging behavioral issues. The murder theme was very engaging for both the boys and girls, all of the children were actively involved (and completely fooled as they thought they were playing games) in the learning. The build up throughout the day kept all of our kids on the edge of the seats. The climax was definitely finding out who the murderer was using clues to refine their searches. It was a great way to teach a boring piece of software like Microsoft Access. Not only did the class learn a massive amount, I definitely picked up a thing or two that even to this day (nearly a year later) I can still clearly recall. A truly memorable day delivered by a super ICT teacher."* Mrs Spicksley, ICT Leader



## 4.2 Pupil feedback

*"I loved it when we did the human database and moved ourselves around the room depending on if our answer was yes or no to the question."* Aimee

*"I learnt a lot like searching databases. The fun bit was typing all of the data together onto one big worksheet and seeing it grow."* Adam

*"My most favorite part was when we filled out the spreadsheet and imported some of the information into a database."* Caitlin

*"My favourite bit was to find out who the murderer was. I found it hard when we had to do things like changing the field in the database but I got it and it was great fun."* Jason

## 5. CREATIVE CURRICULUM OPPORTUNITIES

### 5.1 Post lesson resources

Teachers are encouraged to continue the work started at the DSH when they return to their own school. The DSH has made it a priority to support visiting schools, both before and after visits. This is achieved through the uploading of all the ICT lesson resources (including the Independent Learning Videos) and creative curriculum activities. For example, the database lesson has a series of pre-visit math challenges loosely based on the book consisting of measuring the area of a footprint, code breaking using coordinates and cipher wheels (*Figure 6*). Visiting schools are also provided with access to philosophy for children resources, including THUNKS such as 'Who owns your finger print?' [3]. The post visit resources include a variety of activities including:

- Literacy: to rewrite the final chapter of the book as the Digital Schoolhouse detective.
- Art: to draw a facial e-fit of the murderer based on the profile card and descriptions in the book (*Fig. 7*).
- PE: to extend catching and throwing schemes of work by learning to juggle three balls.

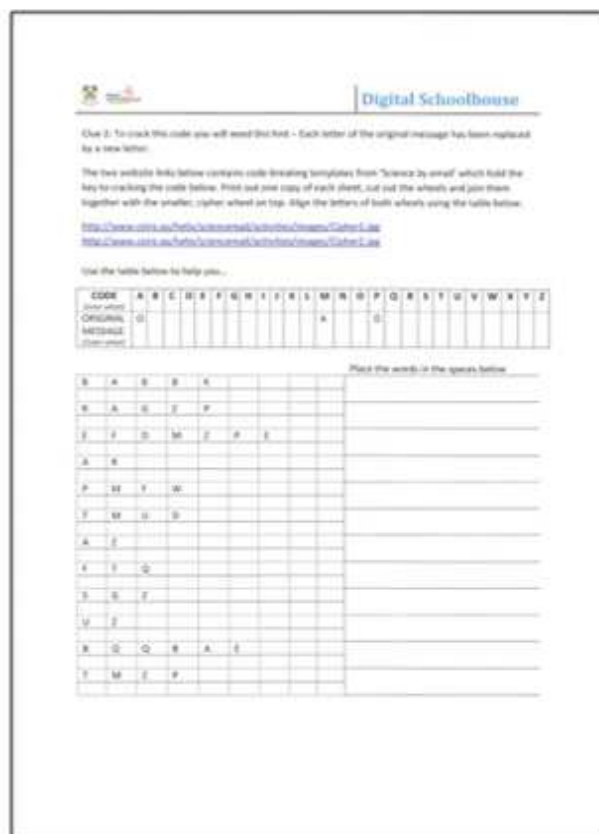


Figure 6: Cipher wheels Maths challenge.



Figure 7: Cross curricular activity.

## 5.2 The teaching of habitats using databases

To complement this Database lesson and to meet the needs of local primary schools, the DSH has also collaborated with the Jungle Fortress Animal Sanctuary to develop an Animal Detectives lesson which works on the same principles described above but the scenario is based on a forest mystery.

In addition to the Animal Detective lesson, the DSH has worked with the animal sanctuary to develop a workshop which is delivered by their education team and sets pupils the task of looking for clues to help them to solve the DSH forest mystery whilst having a hands on animal encounter. This workshop is delivered to pupils at their school for a small donation to the animal sanctuary.

## 6. IMPACT AT SECONDARY LEVEL

### 6.1 Transitional challenges

In Year 7, pupils at LGS receive two hours per week of ICT teaching. The school adopts an accelerated ICT curriculum by covering the Key Stage 3 curriculum in one year. In Years 8 and 9, pupils take their GCSE ICT. To prepare pupils for their GCSE, they are given a target of level 6 by the end of Year 7. This can be particularly challenging for the ICT teaching staff because LGS has links with up to 80 different feeder primary schools where pupils will have a vast range of knowledge and experience.

### 6.2 Using the pedagogy at Langley Grammar School

To help address these transitional challenges faced by the ICT & Computing department, the DSH teacher has worked with his colleagues to investigate how best the principles from the database lesson could be adapted for Key Stage 3 pupils. In particular, year 6 pupils are introduced to cloud computing on their year 7 induction day as a tool to help them to get to know their new classmates. Essentially, pupils enter non personal or sensitive data into the online spreadsheet e.g. first name, tutor group, hobby, favourite subject and colour etc. As soon as the data entry task is complete and it has been downloaded by pupils to their user area, the class teacher deletes it from the cloud. By using the spreadsheet filter tool, pupils play an electronic game of *Guess Who* which the ICT & Computing department calls the *Langley Grammar School Friend Finder*.

The following year, when pupils are in Year 7, the same lesson principles described in the Key Stage 2 lesson are employed but pupils are extended to focus on using primary keys, applying data types to fields, creating queries with the *greater than* or *less than* criteria, and creating reports that are grouped, sorted and generated from a query. This is achieved through replacing of the murder suspect profiles with a Geography topic (stats about countries) or Literacy topic (Harry Potter) based on the Top Trumps playing cards. Alternatively, you could work with your English department to create your own fictional superhero (*Figure 8*) with the pupils deciding the qualities required to be a Superhero and agreeing on the minimum and maximum values for each category. Pupils then add an additional column to the database to allocate superheroes or Top Trumps cards to the respective players. Pupils play the game as usual, the only difference being now it is electronic e.g. pupils ask questions based on the categories using *greater than* and *less than*, with one of the players keeping a tally chart to record wins and losses.





Figure 8: Superheroes work produced by a Year 7 student at LGS.

### 6.2.1 Teacher feedback

This approach to teaching databases has shown a particular improvement in the assessment results of less able pupils in classes where it was trialled, and the ICT staff have commented how well pupils have been able to “recall and use the learning from Year 7 during the preparation tasks” for their first GCSE Controlled Assessment.

### 6.2.2 Pupil feedback

*“When I joined the school I had absolutely no clue what a database was and how to use Microsoft Access. I really like how Mr Dorling used us pupils as his data in his human database because by role-playing Data it made things a lot easier to understand, in particular, creating complex queries and reports.”* Chinmayee in Year 7

## 6.3 Opportunities for creative curriculum days

LGS will further integrate the ICT and Computing curriculum into other subjects by a running a creative curriculum day where Year 7 pupils will be off time table and asked to apply their learning from a range of subjects to solve a murder. It is intended that the curriculum links for the day will focus on the teaching of forensic science and use of databases. However, it will be extended further through links to other subjects such as Maths, Modern Foreign Languages and Drama.

### 6.3.1 The teaching of forensic science using databases

Later this academic year, the Digital Schoolhouse and Royal Holloway, University of London, will be launching a CSI kit with a range of science experiments aimed at Key Stage 2 pupils to encourage primary to be more creative in their teaching of forensic science using databases. It is intended that the kit will be loaned to schools for a small (returnable) deposit.

## REFERENCES

- [1] Common Craft, *Google Docs in Plain English*.  
[http://www.youtube.com/watch?v=eRqUE6IHTEA&feature=player\\_embedded](http://www.youtube.com/watch?v=eRqUE6IHTEA&feature=player_embedded)
- [2] Crickweb, *Minibeasts classification*. <http://www.crickweb.co.uk/ks2science.html>
- [3] Gilbert I., *THE LITTLE BOOK OF THUNKS*. Crown House Publishing Limited. 81 (2007)
- [4] Landman T., *Certain Death (Poppy Fields Murder Mystery)*, Walker (2010)
- [5] National Curriculum, *Minibeasts tree diagram*. <http://curriculum.qcda.gov.uk/key-stages-1-and-2/assessment/nc-in-action/items/science/4/1431.aspx>

- [6] Qualifications and Curriculum Development Agency, *Independent Review of the Primary Curriculum: Final Report*. (2009)  
[https://www.education.gov.uk/publications/standard/Download?DownloadPublicationReference=DCSF-00499-2009&DownloadItemReference=Independent%20Review%20of%20the%20Primary%20Curriculum%3A%20Final%20Report%20PDF\(DfES%20Online%20Store\)&DocumentType=PDF&Url=%2Fpublications%2FOrderingDownload%2FPrimary\\_curriculum\\_Report.pdf](https://www.education.gov.uk/publications/standard/Download?DownloadPublicationReference=DCSF-00499-2009&DownloadItemReference=Independent%20Review%20of%20the%20Primary%20Curriculum%3A%20Final%20Report%20PDF(DfES%20Online%20Store)&DocumentType=PDF&Url=%2Fpublications%2FOrderingDownload%2FPrimary_curriculum_Report.pdf)
- [7] Smart Learning, *YR7 SMART SKILLS BUILDER ICT STUDENT'S BOOK*, 61-70 Alden Press (2006)
- [8] The Digital Schoolhouse, *Lesson links to the QCA: itx3c – Introduction to databases*.  
[http://www.digitalschoolhouse.org.uk/attachments/071\\_KS2\\_Y3\\_itx3c%20-%20Introduction%20to%20databases.pdf](http://www.digitalschoolhouse.org.uk/attachments/071_KS2_Y3_itx3c%20-%20Introduction%20to%20databases.pdf)
- [9] The Digital Schoolhouse, *Lesson links to the QCA: itx4c – Branching databases*.  
[http://www.digitalschoolhouse.org.uk/attachments/071\\_KS2\\_Y4\\_itx4c%20-%20Branching%20databases.pdf](http://www.digitalschoolhouse.org.uk/attachments/071_KS2_Y4_itx4c%20-%20Branching%20databases.pdf)
- [10] The Digital Schoolhouse, *Lesson links to the QCA: itx5b – Analysing data and asking questions*.  
[http://www.digitalschoolhouse.org.uk/attachments/071\\_KS2\\_Y5\\_itx5b%20-%20Analysing%20data%20and%20asking%20questions.pdf](http://www.digitalschoolhouse.org.uk/attachments/071_KS2_Y5_itx5b%20-%20Analysing%20data%20and%20asking%20questions.pdf)
- [11] The Digital Schoolhouse, *Lesson links to the QCA: Introduction to spreadsheets*.  
[http://www.digitalschoolhouse.org.uk/attachments/071\\_KS2\\_Y5\\_itx5d%20-%20Introduction%20to%20spreadsheets.pdf](http://www.digitalschoolhouse.org.uk/attachments/071_KS2_Y5_itx5d%20-%20Introduction%20to%20spreadsheets.pdf)
- [12] The Digital Schoolhouse, *Lesson links to the QCA: y7 ict5 – Data*.  
[http://www.digitalschoolhouse.org.uk/attachments/071\\_KS3\\_Y7\\_ict5%20-%20Data.pdf](http://www.digitalschoolhouse.org.uk/attachments/071_KS3_Y7_ict5%20-%20Data.pdf)
- [13] The Digital Schoolhouse, *Suspects Google Docs Spreadsheet*.  
<https://spreadsheets.google.com/ccc?key=0Ao96ldHP4M1-dFdmU25VNmQxVW1kemhMOXZTNnVZLVE&hl=en&authkey=COHBxtYO>
- [14] The Digital Schoolhouse, *Importing suspects google docs spreadsheet*.  
<http://www.youtube.com/watch?v=rHW-mnbKYfo>
- [15] The Digital Schoolhouse, *Filtering the suspects to solve the murder*.  
<http://www.youtube.com/watch?v=IQqFrPAoYWM>
- [16] The Digital Schoolhouse, *Creating a query as evidence for the court*.  
<http://www.youtube.com/watch?v=nGVHkrWDeBg>

# INQUIRY BASED LEARNING DATABASE ORIENTATED APPLICATIONS FOR COMPUTING AND COMPUTING FORENSIC STUDENTS

Jackie Campbell  
Faculty of Arts, Environment and Technology  
Leeds Metropolitan University  
[J.Campbell@leedsmet.ac.uk](mailto:J.Campbell@leedsmet.ac.uk)

Sanela Lazarevski  
Faculty of Arts, Environment and Technology  
Leeds Metropolitan University  
[S.lazarevski@leedmet.ac.uk](mailto:S.lazarevski@leedmet.ac.uk)

---

## ABSTRACT

*The purpose of this paper is to describe and discuss the use of specifically developed inquiry based learning materials for Computing and Forensic Computing students in the database area. Two of the applications developed are discussed here. The first 'SQL Quiz' is designed to aid students in their learning of SQL and to appreciate the difference between simple and complex queries in meeting the requirements of a 'report'. The second 'Application Development – First Job' requires the students to use investigative methods to test, discover and address issues within the application. Additionally this application intends to simulate industrial work such as evaluation, testing and fixing of software. The initial feedback is that students like the applications; they receive quick feedback and can move forward. It has raised their awareness of data quality, data integrity and improved their confidence to question results from reports and queries.*

## Keywords

*Inquiry based learning, investigative learning, problem based learning, e-learning, databases, forensics.*

## 1. INTRODUCTION

In this paper, we describe the teaching applications created to aid students with their learning in the database area at Leeds Metropolitan University. Inquiry based learning can provide a meaningful context to the learning [7] and so motivate the learners and technology enhanced learning can allow students to become more self directed in their own studies [11]. The teaching applications developed in this study fall into both of these categories. The first: **'The SQL Quiz'** aims to teach SQL by setting the students SQL queries to write and test. It allows students to set their own pace and level of learning, they receive instant feedback and can move to easier or more challenging tasks as they require.

The second **'Application Development – First Job'** is intended to give students an experience of the kind of job market they may enter. They 'test' the product and find issues with the data and code which they are then required to investigate and fix. The discovery of the issue is relatively straight forward, the investigation and fixing of the issue raises challenges of varying degrees and again feedback is clear and instant.

At LeedsMet the number of students taking the forensic computing based courses has increased over the last few years. As a result of this we had increased number of forensic students on the database modules, some of whom did not appreciate how the database skills were relevant in the field of forensics or felt they had learnt enough via the web-based modules [5]. Additionally we identified that many of the skills developed in the forensic courses such as investigative skills, problem solving and analysis are key skills to develop and encourage in all computing and database students. As a result the database curriculum at LeedsMet has been developed to include inquiry based learning, investigative learning and forensic-based case studies on the database curriculum, student numbers have increased on the optional database modules and feedback for databases has improved overall.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

In support of this initiative these are two of the electronic learning materials developed to be used in database teaching. Both applications were developed in Oracle Apex (version 3.1.2) which is the development tool used for teaching at LeedsMet.

The objectives of the applications are:

- To improve students learning, investigation and problem solving skills by introducing self-lead inquiry based learning;
- To give students exposure to 'real life' work;
- To show students applications developed in Oracle Apex.

This paper discusses how the 'SQL Quiz' and the 'Application Development – First Job' applications have been used added within the database curriculum and the initial findings on their evaluation.

## 2. APPLICATION DESIGN AND IMPLEMENTATION

This section describes the applications developed.

### 2.1 SQL Quiz: An application to develop skills in Structured Query Language (SQL) code.

Technology offers students the ability to learn at their own pace [2]. The 'SQL Quiz' application was introduced to students as part of their teaching on SQL. Teaching staff found that students struggled to relate a business reason with an SQL query and also were poor at evaluating the complexity (basic, intermediate, advanced) of the SQL query they had written. It seemed this could be because they had 'got so far' with the workbook and then 'given up'. This application was based entirely on a paper-based exercise. Yet we found they engaged much better with the 'online' activity that they had with a booklet.

The application is based around the Scott database case study (as provided by oracle). The user selects a 'level' (basic, challenge basics, using functions, group functions, simple joins, other joins and advanced), they are then given a query to write (of that level). They are able to run their code and the output is displayed. The application gives the opportunity to check the SQL code and the output with a suggested solution. The Entity-Relationship-Diagram for the case study can be viewed via the application.

SQL Quiz: All Questions are based on the scott Database

[View Database Diagram](#) Previous  
Next

Question Number: 15 Level: Using Functions Operator: All

Challenge: List the employee name and salary increased by 15% and expressed as a whole number.

Enter your SQL:

Run >

[hide <== Suggested Solution](#)

Suggested Solution:

```
select deptno, ename, round(sal*1.15) pctsal
from emp;
```

Run >

| DEPTNO | ENAME  | PCTSAL |
|--------|--------|--------|
| 20     | SMITH  | 920    |
| 30     | ALLEN  | 1840   |
| 30     | WARD   | 1438   |
| 20     | JONES  | 3421   |
| 30     | MARTIN | 1438   |
| 30     | BLAKE  | 3278   |
| 10     | CLARK  | 2818   |
| 20     | SCOTT  | 3450   |
| 10     | KING   | 5750   |
| 30     | TURNER | 1725   |
| 20     | ADAMS  | 1265   |

Figure 1: Screen shot of the 'SQL Quiz'

This is aimed at students in their second year of database study. It supports the learning objective to develop SQL skills. Examples of these reports are:

1. A 'basic' SQL statement: 'List the employees and their salary', this is a one table SELECT statement
2. An 'simple join' statement: 'Show the names of employees and the departments they work in', this is a two table join.
3. An 'advanced' statement: 'Show the names of employees and their managers', this is a one table self join. Another example is: 'Display information on the tasks which have gone over budget and the employee responsible for the table', this is a three table join which also requires a sub-query.

### 3.2 Application Development – First Job: An application to develop Oracle Apex and investigation skills.

This application is based around a PROMOs case study. The case study monitors the 'Promotional packs' for 'Accounts' – as allocated by 'Account Managers'. Students are told they need to test the application, initially they are asked to test on their own initiative, later they are given a test plan to test against. Once they have identified the 'bugs' they are required to fix them. The issues in the application are:

- A 'List Of Values' (LOV) that retrieves the incorrect list (it displays names instead of cities);
- An incorrectly labelled field (labelled 'account name' but contains the address);
- Incorrect SQL in a report (the report states that it shows the 'total sales value' per month, it actually displays a 'count' of the sales);
- Incorrect page navigation (the page link states 'Account Managers' the link takes the user to 'Accounts').

| Edit                 | Kit Name                 | Mgr Name        | Date Allocated To Mgr | Qty To Mgr |
|----------------------|--------------------------|-----------------|-----------------------|------------|
| <a href="#">Edit</a> | Bacardi Christmas Kit    | Carissa Shepard | 03-FEB-2011 09:51:59  | -10        |
| <a href="#">Edit</a> | Bacardi Pack             | Reese Robbins   | 08-MAY-2010 00:00:00  | 0          |
| <a href="#">Edit</a> | Southern Comfort Kit     | Robert Burt     | 11-FEB-2011 13:24:28  | 1          |
| <a href="#">Edit</a> | Guinness Halloween Kit   | Carissa Shepard | 24-APR-2010 00:00:00  | 4          |
| <a href="#">Edit</a> | Jack Daniels Kit         | Chantale Lewis  | 11-FEB-2011 13:43:34  | 22         |
| <a href="#">Edit</a> | Fosters Christmas Kit    | Gillian Cotton  | 08-FEB-2010 00:00:00  | 30         |
| <a href="#">Edit</a> | Budweiser Easter Kit     | Grady Bradley   | 05-JUN-2010 00:00:00  | 33         |
| <a href="#">Edit</a> | Carlsberg Christmas Kit  | Ishmael Ochoa   | 08-OCT-2010 00:00:00  | 61         |
| <a href="#">Edit</a> | Carlsberg Halloween Kit  | Melinda Pitts   | 15-SEP-2010 00:00:00  | 90         |
| <a href="#">Edit</a> | Carling Halloween Kit    | Paul Conway     | 23-MAY-2010 00:00:00  | 132        |
| <a href="#">Edit</a> | World Cup Kit            | Unity McDaniel  | 14-JUN-2010 00:00:00  | 152        |
| <a href="#">Edit</a> | Carlsberg Valentines Kit | Ahmed Oneil     | 06-JUL-2010 00:00:00  | 168        |
| <a href="#">Edit</a> | Jamesons Easter Kit      | Vera Rowe       | 12-SEP-2010 00:00:00  | 170        |
| <a href="#">Edit</a> | Seignoff Easter Kit      | Delana Black    | 18-JUL-2010 00:00:00  | 170        |

Figure 2: Screen shot of the 'First Job' Application

This is aimed at students in their second year of database study. They will have worked through a number of tutorials designed to teach them how to build applications using Oracle Apex. The teaching team found that students were capable of developing applications by following the instructions yet many of them struggled to apply the techniques to their own development. Additionally students tended to assume that if code compiled it was correct – this application contained working (but incorrect) code designed to make them question the functionality of the system. Again the exercise intends to give them some 'hands-on' experience of investigating and requires them to use appropriate tools in doing so.

### 3. METHODOLOGY

An empirical study consisting of paper-based survey and focus group data collection methods were undertaken to critically evaluate learning approach when using enquiry based activities. The sample used in the study is within Arts, Environment and Technology Faculty on Computing and Forensics award. Ideally, we would have liked larger sample of questionnaire sent, however due to time constraints this was not possible, hence, a further sample of this population was necessary. To restrict the population sample further, the researcher decided to disseminate survey questionnaires only to those individuals currently on L5 award, studying databases on Computing and Computing Forensic awards.

In surveys performed to ascertain student well being by UNIQoLL (UNiversity Quality of Life and Learning) by Audin, Davy & Bar "It was found that student response rates were higher in departments who chose the lecture completion method and therefore this was encouraged for the second survey." [1] One major limitation of the survey approach adopted is that it will "...only give a snapshot of students' experiences, ignoring variations over time." [ibid] Hence, this research made the decision to disseminate the paper-based survey during tutorial sessions, where we would be present if additional information (more than that included on the questionnaire cover-sheet) was requested. The week selected to conduct this survey was last week of teaching before Christmas break. This meant that students had all their teaching; however this also meant that many students had gone home for the break.

As the figures of individuals that would attend the tutorial session that week were not available, therefore we opted for non-probability sampling, where the sample reasoning can be based on a number of factors. [14] In an attempt to gain a balance between an appropriate sample size and time constraints, we decided to focus the primary research over one week period, each tutorial group once.

The questionnaire was constructed to represent the two main groups of people important to this study, students and teaching staff. We were limited to those who have used the applications and therefore able to provide us with feedback. The questionnaire contained open questions in order to get a true representation of respondents meaning without influencing their reply[10]. A difficulty with the use of open questions is they are notoriously difficult to code and use statistically due to variance in responses. In the design of questions we designed questions around the first three levels of learning of Bloom's taxonomy (Bloom et al., 1956), that is 1) 'Recall of Knowledge', 2) 'Comprehension' and 3 'Application', i.e. the ability to apply learning to a new or novel task. Questions on both questioners for respective applications were very similar. This is because of the differences in the nature of applications, but still looking to identify recollection of knowledge, comprehension and application. The major difference was that 'Application Development – First Job' application questionnaire included a summary of findings using the application, and asked students to comment.

A very small number of responses were received for both applications surveyed, (in total 3 staff and 17 students); hence, analysis performed may be weak and non-representative. It is generally preferable to have a sample of 30 or more[8]. An additional issue in relation to the sample size is that the measurements used in analysis are not as effective when applied to small amounts of data.

The questionnaires have been analysed manually for content and used to explain possible patterns where a strong relationship is established, or alternatively to suggest reasons for independence between variables. [14]

### 4. FINDINGS AND RESULTS

The 'SQL Quiz Application' received feedback from 2 Staff and 10 students, the 'Application Development – First Job' 2 staff and 7 students. While there was one member of staff giving feedback for both applications, none of students completed feedback for both applications. We decided to break down feedback for each application, but not to breakdown outputs given by staff and students.

#### Feedback for 'SQL Quiz' Application

First question was how they found instructions on how to install the application into their Apex account. Responses given were "fine", "easy", "clear", "clear enough is you follow instructions", "it could be default application within the application section, like sample form", "ok". On the question if the application's purpose was clear, comments included "it was straight forward", "clear", "don't know", "will see when assessment is

*done”, “would like more of applications like this”, “as self-lead learning tool for SQL is very good”, “if students use it for active and not passive learning, i.e not only as an electronic version of tutorial documents, copy, past and modify”.*

In response to did you find the application useful/valuable in your SQL learning responses were: “YES”, “different way to test my knowledge”, “helped to identify possible SQL assessment statements”, “my SQL understanding improved”, “I learned advanced SQL statements”. One point was to maybe have “more than one suggested solution for each question”, this will be considered for further application development.

The next question was about how useful the application was in their understanding for application interface building, and if they realized it was built in Oracle Apex. This was asked as one of the assessments parts is to build front end for the database using Apex. Some of the responses were “No”, “How was it done?”, “Very good”, “I learned techniques that could be implemented within the application”, “Kind off”, “Very good example for students to see”. Many students do not realize that this type of applications can be developed using Apex. Two students have responded on ‘Any other comments’ question that they would have “liked to have instruction manual for the application”. Another general comment by student regarding SQL quiz application was “to have save file feature”, to be able to revisit learning. Both comments will be taken into consideration for next year, when this application will be used as part of teaching and learning materials.

#### Feedback for ‘Application Development – First Job’

The first question was to whether they found the application purpose clear, responses included “it was straight forward”, “Yes”, “clear”, “purpose difficult to understand”, “extremely useful”, “valuable in understanding case study organization”, “with teachers support, purpose was clearer”

In response to ‘Did the application help you to learn/understand purpose of constraints, validation, triggers, etc’. Comments included: “extremely useful”, “great guidance for application development potential”, “great assessment support learning”, “Yes”, “takes time to understand how much we have learned using the application”, “in the beginning it’s very frustrating using it, as it looks like having loads of errors not realizing it’s designed like that for purpose of learning”

Though only ‘Application Development – First Job’ used investigative methods in identifying bugs in the application, both applications were used without instructions (they had been verbal briefed about the application) and it was up to the user to identify its usefulness and relevance. Therefore our next question was how useful the application was in their understanding for application interface building, and if they realized it was built in Oracle Apex. Some of the responses were: “I developed debugging skills”, “I learned ways to implemented within the application”, “No”.

We believe that students had used their previous knowledge to further their learning using these applications, (as was intended). Referring back to Blooms Taxonomy students were required to 1) ‘Recall Knowledge’, and 2) ‘Comprehend’. Some students did not see how their knowledge gained would be applied (3) ‘Application’).

If we are to conduct this survey again, the timing would have to be more considered as students required more time to use the applications. It could be useful to wait until after the assessment has taken place and then gather feedback. Also, groups of students to survey would have been selected differently, too those who finished studying database module, and those who are still studying.

## **5. CONCLUSION**

In conclusion we believe that we have achieved what we set out to do. Referring back to the initial objectives of the applications:

- To improve students learning, investigation and problem solving skills by introducing self-lead inquiry based learning;
- To give students exposure to ‘real life’ work;
- To show students applications developed in Oracle Apex.



It is felt that these were all met to some extent. The applications were appreciated by the students and their responses show that they enjoyed this kind of learning experience, found it valuable and want to learn and to be challenged.

We will encourage students at year 3(L6) to produce more of this type of application for database learning, the intention being to create a 'bank of applications' addressing different learning objectives to be used by other students.

## REFERENCES

- [1] Audin, Davy, J.& Barkham, M. (2003). UNiiversity Quality of Life and Learning (UNIQoLL): An approach to student well-being, satisfaction and institutional change. *Journal of Further and Higher Education*, 27, 365-382
- [2] A. W. Bates and Gary Poole (2003) *Effective Teaching with Technology in Higher Education*, Jossey-Bass Publishers, San Francisco CHECK
- [3] Bell, J (2005) *Doing you research project* 4th edition. Maidenhead, Open University.
- [4] Bloom, Benjamin, et al. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*. McKay, New York.
- [5] Campbell J, Lazarevski S. (2010) Introducing a forensic Flavour to teaching Databases at Level 6. *Proceedings of the 6th Annual International Conference on Computer Science & Information Systems* 25-28 June 2010
- [6] Campbell J, Gorra A, Lazarevski S. (2007) *An Account of using Integrated assessment for Students in the Area of databases at Level 2*. BNCOD 2007.
- [7] Daniel C. Edelson, Douglas N. Gordin, Roy D. Pea (1999) *Addressing the Challenges of Inquiry-Based Learning through Technology and Curriculum Design*, *Journal of the Learning Sciences*.
- [8] Denscombe M. (1998). *The Good Research Guide for Small-Scale Social Research Projects*. Buckingham, UK: Open University Press.
- [9] Devlin, M., Drummand, S., Phillips, C. Marshall, L. (2008) *Improving Assessment in Software Engineering Student Team Projects*. *Proceedings of the 9th Annual HE Academy - ICS conference*, Liverpool Hope University, 26th Aug. - 28th Aug. 2008, pp. 127-133.
- [10] Foddy, W. 1993: *Constructing questions for interviews and questionnaires: theory and practice in social research*. Cambridge: Cambridge University Press
- [11] Higher Education Funding Council for England (HEFCE) on behalf of JISC (2007) *Effective practice with e-Assessment: An overview of technologies, policies and practice in further and higher education*. Available from: <<http://www.jisc.ac.uk/assessment.html>>.
- [12] Jenkins, T. (2005) *How do they think they're doing?* *Proceedings of the 5th Annual HE Academy - ICS conference*, University of York, 30th Aug. - 1st Sept. 2005, pp/ 39-43.
- [13] Low O'Sullivan, M. (2005) *Authentic individual self-assessment, praxis and community*. 28. *Proceedings of the 2005 HERDSA Annual Conference*. Sydney Australia 3 - 6 July.
- [14] Saunders, Mark., Lewis, Philip., & Thornhill, Adrian. (2009) *Research Methods for Business Students*. 5<sup>th</sup> ed. Harlow, Prentice Hall.
- [15] Tapp, L (2003) *Computer- based training*. *Industrial Safty Hygeniene News*, Vol 37 issue 6, P35.



# USING WEB-ENABLED MOBILE PHONES FOR AUDIENCE PARTICIPATION IN DATABASE LECTURES

Craig McCreath

University of Abertay Dundee  
School of Computing and Engineering Systems  
Dundee DD1 1HG  
[craig@craigmccreath.com](mailto:craig@craigmccreath.com)

Petra Leimich

University of Abertay Dundee  
School of Computing and Engineering Systems  
Dundee DD1 1HG  
[p.leimich@abertay.ac.uk](mailto:p.leimich@abertay.ac.uk)

---

## ABSTRACT

*This paper evaluates the possible benefits of integrating audience response systems through mobile devices. Typically, existing clicker systems use proprietary radio remotes to send votes to a computer and display this information on screen. While proven to be a useful tool in lectures, adoption has been limited due to their up-front costs. Using mobile devices as a replacement is discussed as an alternative method by making it less intrusive and rootless. A prototype system was implemented to test these assumptions and was evaluated in a typical lecture with undergraduate students studying 'Database Fundamentals'. The students chose between this system and a typical clicker system and were asked to answer identical quizzes. While their feedback shows little benefit of the prototype over traditional systems, the mobile system offers tangible benefits to the lecturer in speeding up preparation of quizzes and setup time.*

## Keywords

*Audience response system, class feedback, mobile devices.*

## 1. INTRODUCTION

The ability to gauge student understanding is an important aspect of higher learning. If done well a "learning-conversation is constructed between the students and teacher," [1] allowing teaching to be at its best. However, the growing numbers of students per classes, some ranging to well over a hundred, causes this dialogue to become detached; teachers can no longer gauge the effectiveness of their lectures.

ARS (Audience Response Systems) have been around for over a decade with the first system, ClassTalk, becoming commercial in 1992 [2]. This system adapted graphing calculators, wiring them into a computer that received the responses. These systems show what answers participants selected to a multiple-choice question. The current market still uses similar systems with responses collated by a single computer. The clickers are no longer wired, instead often using wireless radio technology to send the responses. While more accessible to students as a whole, they continue to have a considerable overhead for the institution. A popular system such as TurningPoint can cost \$2,800 USD (£1,731 GBP as of 07/02/2011) for 24 response clickers—around £70 each.

Many studies have shown that mobile devices as a method of student participation can be an effective method of gauging students' understanding. So far, however, there has been little discussion about using mobile phones for this task despite more students than ever having access to web-enabled phones and mobile devices. According to Hopke [3], over 98% of students own a mobile phone. Yet many in the academic community regard their use as a distraction during lectures and have enforced policies to stop students using them. The opposite view should be considered; used correctly these devices could promote learning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

## 2. MOBILE-BASED AUDIENCE RESPONSE SYSTEM

### 2.1 Implementation

A web-based audience response system where students would use their mobile phone to participate, MARS (Mobile Audience Response System), was developed and tested within a university setting. The requirements of this application were not extensive but required much discourse between student needs and that of the teachers who may use this.

The operating environment was a significant challenge and moved the operation of this system to a web server instead of being a standalone product like others. While a change of practice from most of the products used in this environment it only follows the trend of typically stand-alone products making it a 'cloud service' e.g. Google Docs.

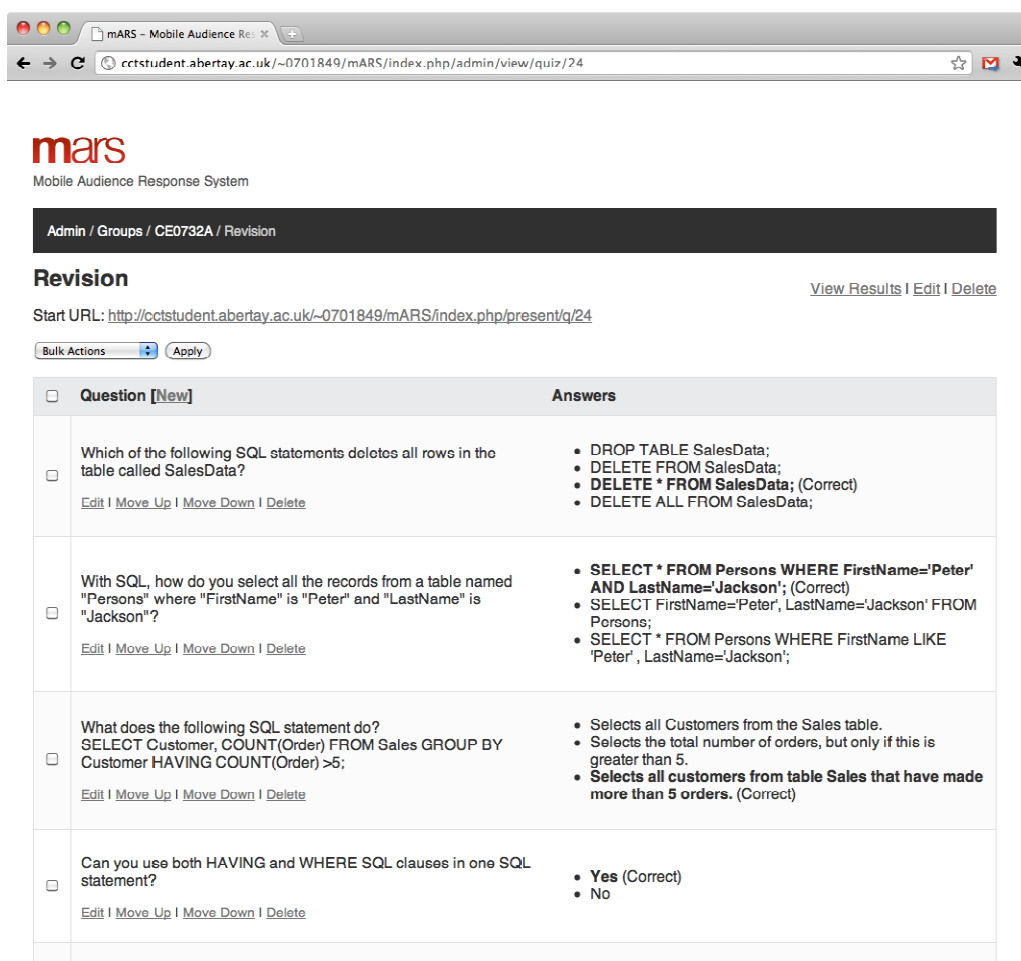


Figure 1 – A typical interface for presenters to edit quizzes before presentation.

For MARS, three distinct component applications were developed. First, and perhaps most significant, was an application that would allow entry and editing of a quiz split into distinctive groups for differentiation as shown in Figure 1. Secondly, a method of presenting the quiz questions and the results during a class was also required (Figures 2 and 3). This component also stores the results (anonymously) for later review by the lecturer. The third component allows participants to enter their vote on their device (Figure 4).

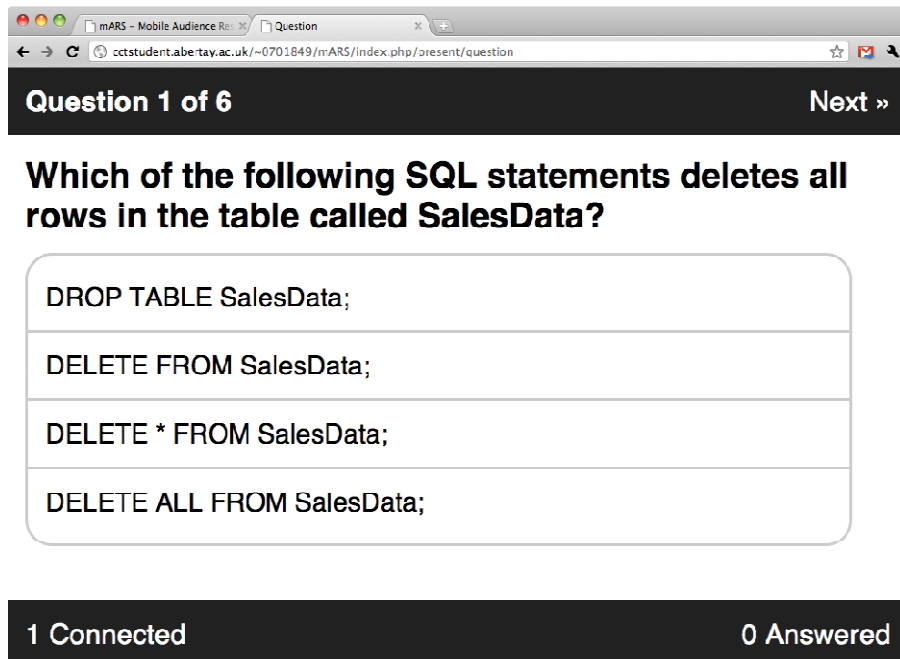


Figure 2 – Interface displayed on the presenter's display & projector.

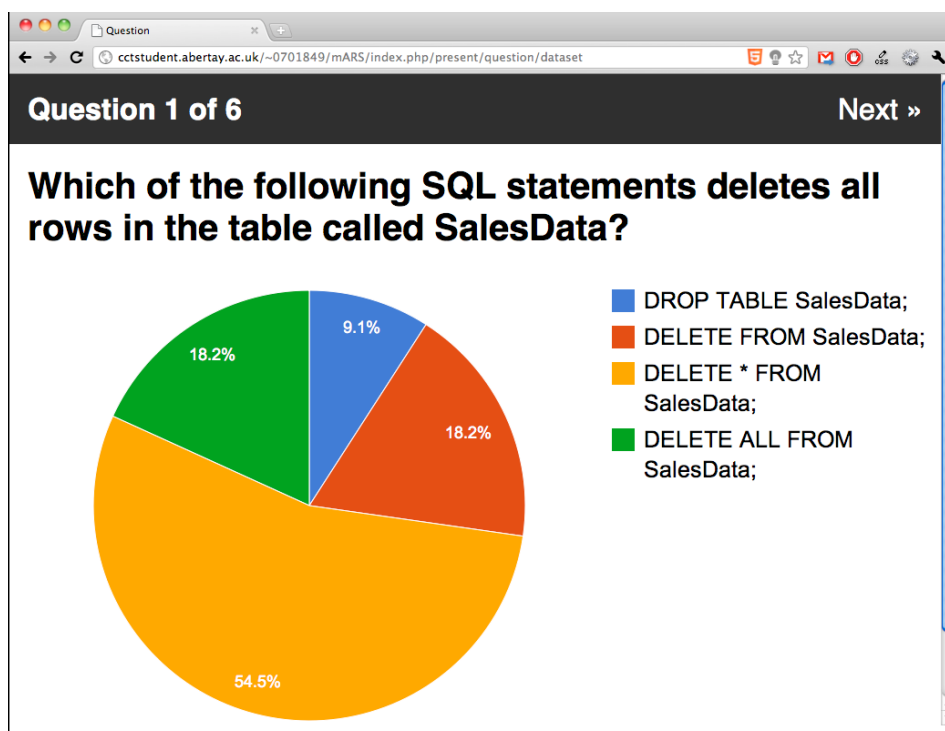


Figure 3 – Results screen shown to presenter and participants after each question.

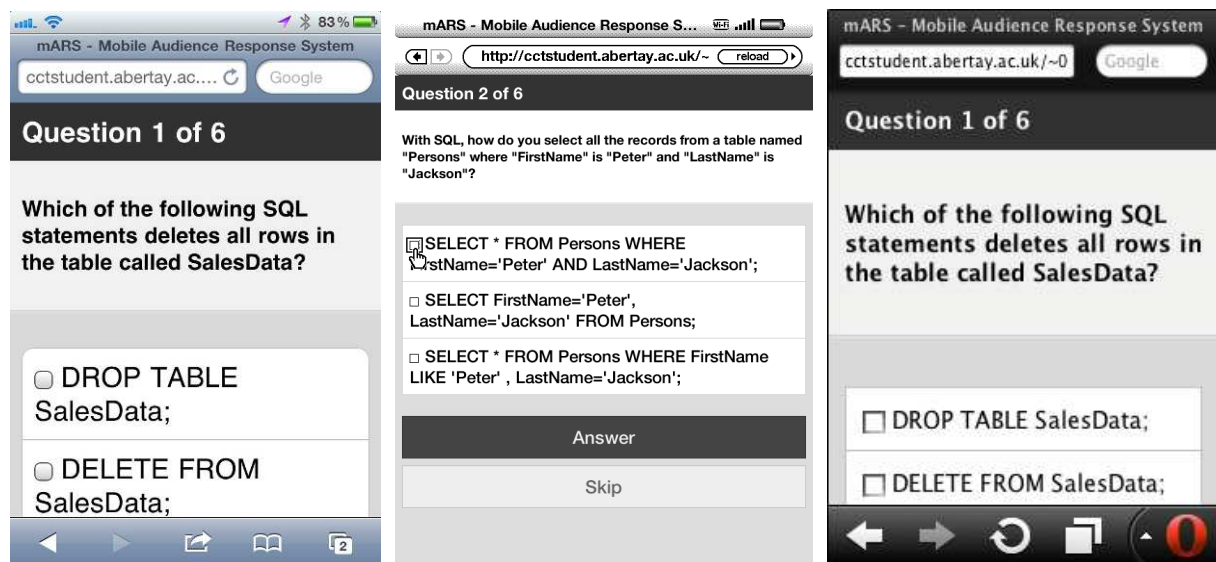


Figure 4 – Voting interface as seen from various mobile phones.

## 2.2 Evaluation

Evaluation of the mobile system focused initially on two aspects. The student participants' ability to use the device was notably the most important. The myriad of differences between phone devices and makes alter possible speed and determine technologies in use. Teachers themselves have to use the system and the need for speed, efficiency and clarity is of great significance at each step of the process.

The system was evaluated against the TurningPoint audience response system as a baseline. To evaluate the system from the student perspective a sample group was invited to participate during a lecture. This group consisted of approximately 25 first year undergraduate students studying "Database Fundamentals", containing typically male students with high computer literacy. Each student was given a choice between using TurningPoint with a borrowed clicker or the MARS system with their own mobile device. 56% of the participants chose to use their own devices, mainly mobile phones but also one or two laptops.

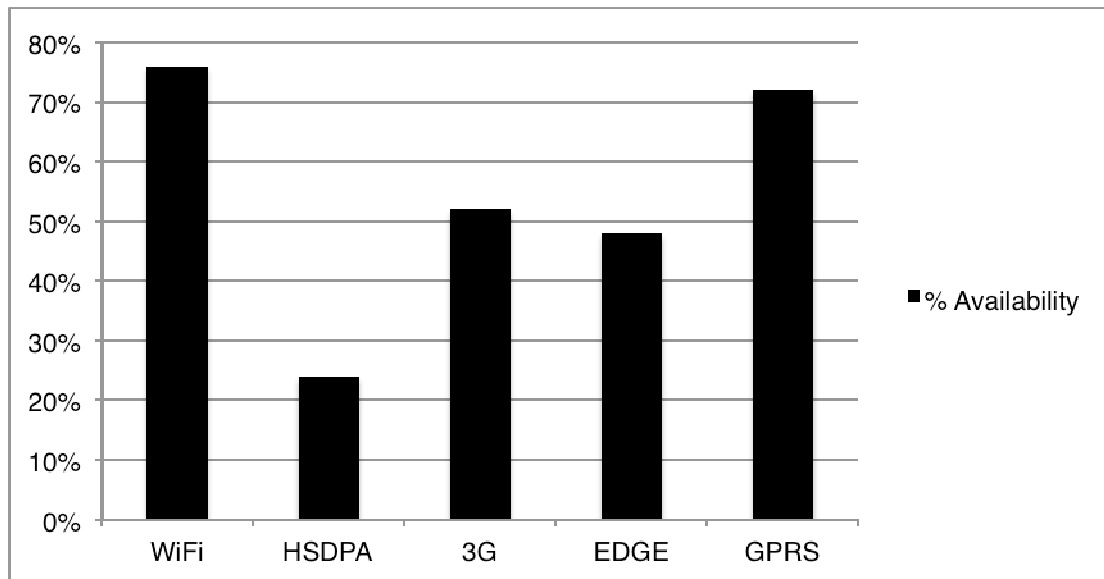
The two systems were enabled simultaneously with the lecturer switching between the two throughout. The questions and answers on these devices were kept identical and had been chosen by the lecturer as appropriate to the lecture content. During the test, the responses remained similar for the majority of the questions.

After this field test, students were asked to complete a questionnaire, which included technical details regarding their mobile phone capabilities and their evaluation of the system they used. The questionnaire was worded slightly differently depending on the system used, though the questions remained almost identical. This questionnaire asked information regarding their educational background and understanding, to confirm the mismatch discussed in the literature, as well as their opinion of either system. No data from the developed system was used for the evaluation as it is system performance, not student performance, which was to be evaluated.

## 3. RESULTS

### 3.1 Phone Properties

A large amount of credence in this research depends on the wide availability of mobile phones in the classroom. This research backs up claims made by Hopke [3] with only one participant (4%) not owning a mobile device. However there are issues with the number of students with a data connection, dropping down to 68%.



**Figure 5. Data capability of participant mobile phones. Horizontal axis ordered from fastest to slowest relative data speeds.**

Figure 5 shows the data capabilities of those with mobile phones during the study. The speeds of the mobile phone are the relative speeds which the user would experience during the voting process in a lecture. The ubiquity of WiFi capability reduces the worry of those who may not have a data connection, and provides a fast alternative to other data methods. GPRS however is significantly slower and the results show that participants who used GPRS did not enjoy using the system as much as others.

There is still a small disconnect between those who have a phone who do not have a fast enough connection to use this system and as such there is a concern. This minority may feel excluded from its use in classrooms due to their inability to participate. While other web devices could be used, such as laptops and other Wi-Fi capable devices, these students would have to own one to participate unless some could be loaned to these students.

### 3.2 Participant Backgrounds

Typically the participants did not report general difficulties with their classes, but around 32% had issues with paying attention to lectures and a similar percentage of students found it difficult to gauge how others understood the material. Just over half the class admitted they found it difficult to ask questions during lectures.

The audience response system engaged these students. While a large number had previously noted that they had no issue concentrating in class, 80% believed that they had paid more attention than in a typical lecture. Johnson [4] states similar rates of engagement, with roughly 88% of their respondents claiming that clickers kept them engaged during class.

### 3.3 System Comparison

Speed is an important factor in these systems on both student and lecturer applications. While student speed has already been mentioned, the time required for setting up questions to presenting the quiz should also be tested.

Both systems were timed as the questions from a source document were copied into each application. TurningPoint incurred a heavy overhead as it had to be installed on the computer before writing quizzes and required the hardware to be connected before any presentations could be created. Creation of the test quiz used in this study took roughly twenty minutes. While a quarter of that was time spent installing the software

itself onto a new machine, the other fifteen minutes consisted inputting information and formatting that information into a typical Microsoft PowerPoint slide. MARS took significantly less time. Text could be copied and entered into the system without worry of formatting and in full, for the example six-question quiz, took roughly three minutes to complete. No installation was required.

The critical number of TurningPoint clickers was not reached during the test but the limitation was apparent. The number of voters using the system is limited by the clickers available, and only one quiz per system could run at one time. Using MARS could theoretically eliminate this issue, possibly allowing hundreds of simultaneous quizzes (in different locations) with thousands of respondents, limited only by the web server's capabilities.

Results were displayed similarly across both devices. The resulting data was displayed differently - the TurningPoint system uses a bar chart, while it was decided early on to use a pie chart for MARS - but did produce similar results. Initially, a drawback with MARS was that this system only showed the answers which had received votes while TurningPoint showed all values, even if null. This issue was easily overcome after the experiment through a workaround, so that null-result answers are now displayed beside the graph for reference.

Delivering results to participants after a quiz becomes carefree. Presenters often provide slides before a lecture. With TurningPoint, these slides have to be exported and saved again to include the results of the votes. On the other hand, results from MARS are stored automatically, and are available to all participants right away through the same link in the presentation. Presenters have this and extra functionality, as they can see results each time they use the quiz individually and together with little effort.

Student preference between the two systems was surprisingly fixed with participants keeping preference of their input system after the test was completed. A small proportion of the participants using the mobile system would prefer to use a clicker instead, with those respondents experiencing the slower data connection speeds. 8% of total participants would use either system.

The large number of students who were capable of running MARS on their device yet choosing TurningPoint was unexpected. Reasons for this remain unclear. While it may be due to the perceived professionalism of the chosen ARS over a student-built MARS, this is merely speculative. Participants rated MARS on par with TurningPoint and welcomed either into their education, as they believed both could assist in their learning. Not knowing why participants chose either system was a significant weakness of this research that was overlooked. If such information were known, it would help to improve future systems.

Notably there was a divide between participants over the questions themselves being displayed not only on the lecture monitor (TurningPoint) but also on their input devices (MARS). Those who had access to this feature found it overall a useful addition, with 71% agreeing that it was either very or relatively useful. On the other hand the majority of TurningPoint users (63%) did not miss this feature and believed it to be of little or no use.

#### **4. CONCLUSION & FUTURE WORK**

The support for ARS by the student participants in this study confirmed reports by Beatty [1] and others that students enjoy using such systems. However, the ability for mobile use of these systems may not yet be as widespread as hoped with 20% of participants not having a data connection. While this could reduce the ability to participate in such quizzes, the ability to observe is not hindered; non-participating students could still engage with the question but not have a casting vote.

For the lecturer, the availability of an easy-to-use web-based system is a major benefit. Inputting data not only became more convenient but extremely rapid, reducing lead-time significantly, and perhaps more importantly the setup time during any classes where it may be used was much reduced.

Results became easier to access for students. If the link to the quiz is included in the presentation, participants can view their resulting dataset quickly and easily. With TurningPoint this takes several additional steps, requiring exporting and uploading of slides manually.

While this study shows that using an ARS may be better than none at all, there is a caveat. Although use of a system would be widely accepted—using it as a method of monitoring progress of individual students may significantly reduce adoption. Also there was a drop between system performance and affordability for MARS and TurningPoint. The possibility of making an ARS integrate with a MARS to allow interaction on both devices may yield better results if at all possible.

These systems build a method of giving a presenter the means of retrieving feedback from an audience. In a higher education setting, this holds promise of reducing issues in student retention and understanding. It does not however solve one issue that may prove essential—expanding class sizes make it difficult for students to ask questions directly. If the system could also provide the audience a method to ask their own questions, perhaps along the ideas of the Live Questions tool [5] this would reconnect the ‘learning-conversation’ [1] to a fuller extent.

This research was conducted on a typically computer-centric student group; their high computer literacy may deliver different outcomes of evaluation. If this was applied to other diverse and larger groups a different picture may be found.

## REFERENCES

- [1] Russel, M. 2008. Using an electronic voting system to enhance learning and teaching. *Engineering Education: Journal of the Higher Education Academy Engineering Subject Centre*. 3(2).
- [2] Beatty, I. 2004. Transforming Student Learning with Classroom Communication Systems. *EDUCAUSE Center for Applied Research*. 2004(3).
- [3] Hopke, K. 2009. Student Cell Phone Use in College Classrooms. *National Undergraduate Research Clearinghouse*. [Online] Available from: <http://www.webclearinghouse.net/volume/12/HOPKE-StudentCel.php>.
- [4] Johnson, M. and Robson, D. 2009. Clickers, Student Engagement and Performance in an Introductory Economics Course: a Cautionary Tale *CHEER*. 20.
- [5] Wallenberg Hall at Stanford University. Live Question Tool. <http://questionanswer.stanford.edu>.





# USING VIDEO TO PROVIDE RICHER FEEDBACK TO DATABASE ASSESSMENT

‘A PICTURE IS WORTH A THOUSAND WORDS, A VIDEO BRINGS THEM ALIVE’

Howard Gould  
Faculty of Arts, Environment & Technology  
Leeds Metropolitan University  
Leeds LS6 3QS  
h.gould@leedsmet.ac.uk

---

## ABSTRACT

*Computing Masters Degree students in the Faculty of Arts Environment & Technology at Leeds Metropolitan University studying databases took part in a trial to see if video feedback enhanced understanding of Entity Relationship Modelling. This paper discusses the reasons for taking this approach, its potential benefits, and some of the operational issues that have arisen during the trial.*

## Keywords

*Entity Relationship Modelling, Video feedback.*

## 1. INTRODUCTION

Driven by recent surveys including the National Student Survey and other internal surveys, Leeds Metropolitan University has, as part of its Assessment Learning and Teaching strategy, made improving assessment feedback a priority.

Student assessment feedback is a vital part of the learning process; however there are a number of factors which impact on the effectiveness of conventional feedback methods.

Often students do not collect their submitted work which has written feedback on, and even when they do they are often dissatisfied with the nature and quality of the feedback, as the Higher Education Academy Assessment and feedback video illustrates [1]. Many feel that feedback lacks clarity [2], particularly if they are not able to have one to one sessions with their tutor to discuss their work.

Additionally as the demands on staff time increase it becomes more onerous to write or type extensive feedback which is sometimes ignored or lacks sufficient detail to really assist the student in gaining a greater understanding. Some types of assessment do not lend themselves to written feedback; for example data models in the form of Entity Relationship Diagrams are particularly difficult to comment on in written form, especially when trying to explain cardinality/optionality and fan and chasm trap errors. However, if the shortcomings in a diagram can be pointed out diagrammatically, with an explanation, this significantly improves understanding. For example the direction of relationships can be indicated. Although smart boards and screen annotation products are available for capturing on-screen actions these often prove expensive and time consuming to set up and use and may present a staff training issue. Whilst the use of voice only feedback may provide a quicker and simpler solution for some types of feedback, it is limited in scope and effectiveness.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

As students become ever more immersed in the world of Web 2.0 technologies, conventional student feedback in written form or even voice recordings now appears to lack sufficient impact to encourage engagement with the feedback process. It was intended that the use of video messaging which can show the student's paper-based or PC-based work with live audio-visual annotation would lead to a more enriched form of feedback from which the students would benefit.

The use of video in teaching is well established and many institutions of Higher Education now make lecture materials available using their Virtual Learning Environments (VLEs) and Internet based technologies such as YouTube[3], as discussed in [4]. So it seemed a natural step to see if the use of low cost video technologies could be of assistance with assessment feedback. Alternative feedback approaches using technology were investigated including the Sounds Good Project [5]; this project focussed on just using voice recordings for student feedback.

A project bid to investigate the use of low cost video cameras for assessment feedback from the University's Teaching Quality Enhancement Fund was successful. The funding allowed for the purchase of the necessary equipment.

The main aim of the project was to see if using video feedback would benefit students. A secondary aim was to see if it would ease the assessment load on academic staff and an additional aim was to evaluate the suitability and effectiveness of the chosen technologies.

## 2. TECHNOLOGY AND PROJECT METHODOLOGY

Initial consideration had been given to using screen recording and annotation software such as Camtasia Studio 7 [6]. However as it would only be suitable for work submitted in electronic format this would prevent work submitted on paper being directly assessable and it was felt that this was still an important consideration.

The following equipment was purchased: Kodak Zi8 pocket video camera, Hama mini tripod and Dell Ultra Sharp 2709W LCD monitor. It was intended that the student assessment could be videoed from either the student's printed work or directly from the LCD monitor. The mini tripod provided a stable hands free method for directing the camera (see Figure 1). A large monitor was chosen to ensure that the displayed images would be viewable when recorded.

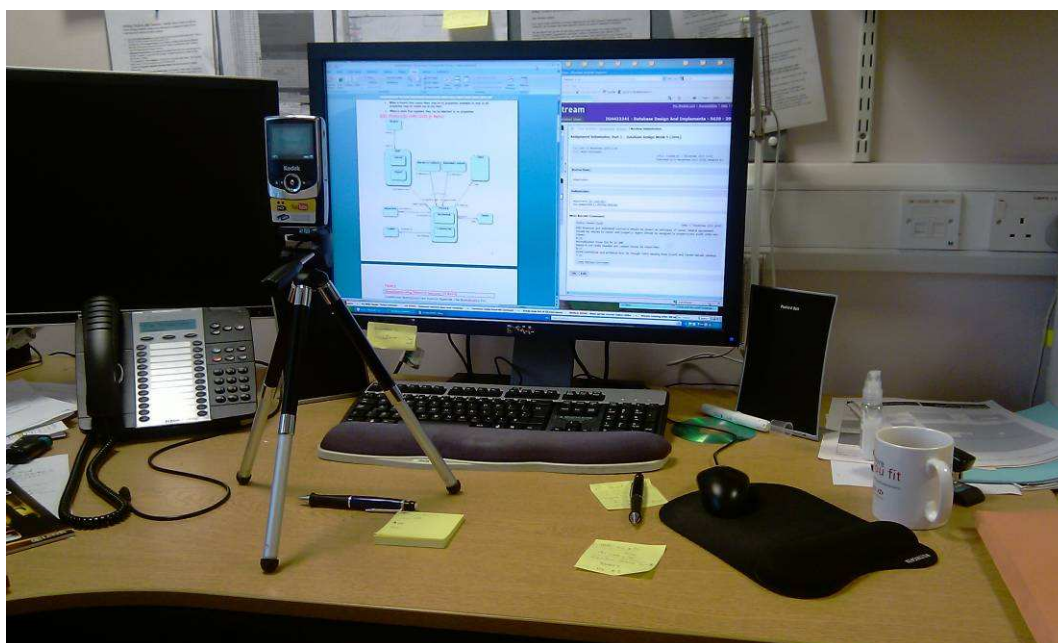


Figure 1

A group of 22 MSc students studying a Database Design and Implementation module was selected for the project and 15 offered to take part. An assessment was chosen for the video feedback which involved the students producing an Extended Entity Relationship Diagram consisting of approximately 10 entities, the normalised version of a document form and a set of table definitions for the modelled system. This particular assessment was submitted electronically rather than on paper – no paper submissions happened to be available during the project schedule. Students submitted their assessment work to the university's VLE in the form of a single Microsoft Word document.

Initially the intention had been to produce the video feedback in real time whilst actually viewing and marking the assessment work. However, this proved difficult as often there were lengthy quiet pauses whilst the work was being considered and the video camera did not have a pause button. Also in some cases comments were made that later needed to be retracted, so it was decided that the work would be marked conventionally first with brief text feedback being entered directly into the VLE and hand written notes being made to assist with the video feedback. A short while later, the feedback videos were produced for those taking part in the project and uploaded to the VLE.

The video feedback process adopted involved displaying the student's submitted assessment work on the LCD monitor which was connected to a normal staff PC. The work was opened in a suitable zoom size to enable the image to be recorded effectively by the camera, which was mounted on the tripod directly in front of the monitor. Having a larger screen allowed for the assessed work to be displayed, with written feedback comments to remain open alongside it for memory jogging, though it was also found to be helpful to have additional brief key notes written on paper, which were referred to whilst recording the feedback. This allowed the videos to be produced with minimal pauses in speaking. The majority of the video feedback concentrated on the assessor using a finger to point to parts of the work on screen, in particular the ERD relationship errors, whilst describing the problem. Often it was necessary to point to the student's normalisation template document to indicate errors in transforming the attributes between normal forms. The individually recorded videos were typically of 5 minutes duration and no post recording editing was undertaken.

### **3. PRELIMINARY FINDINGS**

All participants were contacted to ensure that they had received the video feedback and all confirmed they had viewed the feedback. An anonymous survey was prepared and although only 7 of the 15 students that took part in the project completed the survey, all the respondents were positive about the video feedback they received on their assessment, rating it as helpful or very helpful, and the majority found that video was a good addition to individual written feedback. The majority of the students found it easy to watch their video feedback though most did this at home on their own PC's. However, one commented on a few technical issues which appeared to relate to the size of the video file which caused a sound lag at times. One of the main issues related to the quality of the video: three found it to be poor, three acceptable, with only one rating it as good. This was due primarily to the monitor display scan lines appearing as light moving bars on the video image. Another interesting finding was that all would like video feedback for the other modules they were studying.

Further work could involve evaluating how much of an impact the video feedback had made on a student's performance and whether written feedback could be replaced entirely with video feedback, without any detrimental effect.

#### **3.1 Academic Issues**

As the secondary project aim was to see if academic assessment loads could be reduced, it was disappointing that time was not saved, as it was discovered that giving high quality video feedback on the fly would not easily be possible, though with more practice perhaps this could be achieved. However, the process of having to consider the work and make notes before recording the video feedback led to a more considered and detailed set of points to be fed back than was normally achieved by just writing on the work. It was felt that the quality of the feedback was improved. With further practice with this approach it may be possible to remove the need to produce separate written feedback for the student, in which case there may be a time saving.

#### **3.2 Technology Issues**

The video files were recorded to an SDHC card in a .MOV format and the average size was 250MB each. The recordings typically took in the region of 6-8 minutes to upload to or download from the VLE and this could not be done as a background task. An alternative approach which may avoid this problem is to upload the videos to a university media streaming server and provide a direct link to the video in the VLE.

The major issue relates to the viewable roll bar banding when videoing directly from the monitor. This is caused by the camera speed not being synchronised with the scanning rate of the monitor. Whilst it may be possible to avoid this problem by using monitors with higher refresh rates or more sophisticated cameras it was not possible to verify this. In view of the shimmer or flicker effect, care should be taken to ensure that anyone viewing the videos is not unduly sensitive; in particular people who suffer from photosensitive epilepsy may be at increased risk of a seizure. Video-recording feedback on printed work presents no such problem.

A major drawback of using low cost pocket video cameras of the type used is that they do not have a pause button. This is a serious limitation as it prevents pausing the recording whilst giving the feedback. Whilst it would be possible to edit a number of videos together, this would significantly increase the assessor's

workload and would not be practical in most cases. An alternative would be to consider using different types of cameras with pause facilities, though this may have cost implications.

There is a need for a suitable camera mounting that could be attached to the monitor to avoid the need for a tripod. This would free up the desk space in front of the monitor to allow easier pointing to the screen.

Possible further work could be undertaken to evaluate different screen and video camera technologies to see if the roll bar effect could be eliminated.

Finally there are potential issues with VLEs not being able to handle large numbers of video files or large individual files so care should be taken to check this before embarking on wide scale use.

## **4. CONCLUSION**

Though the current low cost technology is not ideal for the purpose, using video to enhance assessment feedback indicates interesting possibilities for the future. Lecturers may find this approach useful in particular for areas such as database modelling where feedback on a diagram, its structure and the relationships between its component parts is helpful to the student and is most comprehensively captured through video. Further research is needed to compare video recording with screen capture technologies.

## **REFERENCES**

- [1] The Higher Education Academy Assessment and Feedback Video.  
<http://www.heacademy.ac.uk/resources/audioandvideo/assessment>
- [2] Park, C. and Kulej, G. Postgraduate Taught Experience Survey 2009 Report (2009).  
[http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/surveys/PTES\\_2009\\_report\\_v5\\_3\\_22\\_Oct09.pdf](http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/surveys/PTES_2009_report_v5_3_22_Oct09.pdf)
- [3] Leeds Metropolitan University, Leeds Met Uni's channel YouTube.  
<http://www.youtube.com/user/LeedsMetUni#p/c/3C7567A67D126863/21/fmWKf5L0mfA>
- [4] Ronchetti, M. Using video lectures to make teaching more interactive. International Journal Of Emerging Technologies In Learning (IJET), 5(2) (2010). <http://online-journals.org/i-jet/article/view/1156/1404>
- [5] Leeds Metropolitan University Sounds Good: Quicker, better assessment using audio feedback, final project report (2009). <http://www.jisc.ac.uk/whatwedo/programmes/usersandinnovation/soundsgood.aspx>
- [6] Camtasia studio. <http://www.techsmith.com/camtasia/>

# TEACHING DATABASE FOR THE CLOUD(S)

Clare Stanier  
Faculty, Computing, Engin. & Tech  
Staffordshire University  
Staffordshire, ST18 0AD, UK  
c.stanier@staffs.ac.uk

---

## ABSTRACT

*Cloud computing is extending the options for database development, configuration and management. These changes will affect the teaching and learning of databases and the student experience. This paper discusses the ways in which databases in the cloud will impact on the traditional database curriculum and describes the initial experience of introducing this topic to database modules at Staffordshire University. The paper discusses the opportunities and challenges which cloud technology presents for database teaching and proposes a teaching and assessment strategy.*

## Keywords

*Database, Cloud, Database Curriculum*

## 1. INTRODUCTION

Cloud computing is now an established technology but the implications for database development are still unclear. Different strategies may be needed for costing and database administration in a cloud environment [1], for querying and optimisation [2] and for transaction management [3]. There are acknowledged concerns about security of data and there is a debate about the significance of the business model [4]. As mainstream database vendors move into the cloud, there is a need to review the delivery of the database curriculum in order to ensure that students have the knowledge and skills needed to prepare them to work in this field. A further motivation for discussing cloud databases is that, based on some work done at Staffordshire University in 2010/11, (many) students find this an attractive, state of the art, technology and are keen to engage with it. Since cloud database development requires an understanding of concepts such as transaction management and data models, cloud technology can be used to help bring alive some of the more theoretical aspects of the database curriculum.

The rest of this paper is structured as follows: section 2 defines cloud database concepts in this context; section 3 discusses the implications for the database curriculum; section 4 explores the teaching, learning and assessment issues; section 5 presents conclusions and suggestions for future developments.

## 2. THE CONTEXT

Data in the cloud has typically been associated with very large data volumes and non-relational data models but now also includes cloud versions of traditional relational database management systems (DBMS). The focus in this paper is on relational – which is here understood to include object-relational – cloud databases. There is some disagreement as to what constitutes a cloud database or for that matter, any cloud application. One view is that a cloud database is any database hosted in the cloud; another is that a database only qualifies as a cloud database if it has been 'built from the ground-up optimally for the cloud environment, providing 'natural' and unlimited elasticity by using only cloud resources' [5]. This paper uses the less restrictive, hosted in the cloud, definition, partly because the second definition, though possibly more correct, is difficult to apply without detailed understanding of the development process for each DBMS. The paper title refers to *Database For the Cloud(s)* because the single term Cloud does not reflect the fact that there are a number of different cloud offerings and that the variations are significant for the teaching and learning of databases. Some elements, such as elasticity, resource pooling and usage based charging, are common to all types of cloud. For present purposes, private and public clouds are the most significant deployment models. Private clouds are single tenant, meaning that the cloud serves a single organisation or part of an organisation, even though the cloud may be provided by a 3<sup>rd</sup> party. Private clouds imply an institutional

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

commitment and although private clouds for Higher Education (HE) are developing in the USA, they are not at present common in the UK HE sector and are unlikely to be available to support database teaching in the near future. Public clouds, such as Amazon EC2 (Elastic Compute Cloud) are multitenant with the infrastructure belonging to the organisation providing the cloud. Virtual private clouds are private clouds within a public cloud, providing features such as isolation to ensure security. UK HE institutions in general are just starting to look at cloud based services [6] and the attraction of public clouds, from a teaching standpoint, is that they can be used on an ad hoc basis. EC2, for example, emphasises that there is no long term commitment. The business model means that account management is particularly important in a student environment hosted in a public cloud – one study describes a requirement to ensure that students close their connections to avoid charges for an instance which is running but not being used [7]. There are three service models, Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The best definition is probably that provided by NIST [8], the US National Institute of Standards and Technology (with thanks to Peter Garraghan of Leeds University for suggesting this). Figure 1 summarises the implications of the different models for student use, and is based on the NIST definition.

| Service Model | Cloud/User Role   | Possible Student Use  |
|---------------|---|---|
| IaaS          | Infrastructure is provided through the cloud but the user can control elements such as operating systems and storage, can deploy and manage applications and may have some influence over network elements.                 | One study used Oracle with AWS (Amazon Web Services) to provide off-campus access for students [7]. From a teaching standpoint, developing the system is an advanced project which extends beyond purely database issues to include networking and other aspects. |
| PaaS          | The cloud provides the infrastructure including operating system but the user develops and deploys the applications, using languages and tools supported by the cloud provider. The user controls the deployed application. | Allows students to develop and deploy complete applications. The level of skill required, and the support required varies between providers but it is possible to deploy a complete enterprise DBMS.  |
| SaaS          | The cloud provides the user with access to applications but the user does not control the infrastructure or the application although the user may be able to configure application settings.                                | Provides access to the application, supporting off site practical work. Applications may be offered 'out of the box'. Ease of access makes this suitable for beginners.   |

**Figure 1 Service Models and Database Teaching (based on NIST definitions [8])**

### 3. IMPLICATIONS FOR THE DATABASE CURRICULUM

Database teaching covers a wide range of topics but there is consensus that the curriculum should cover the practical elements and underpinning theory needed to develop, manage and administer databases [9]; core topics usually include database design, the relational model, SQL, concurrency, transaction management, database performance and tuning and security [10]. Hansen [11] noted that changing technologies, particularly the influence of the web, has created pressure to expand the curriculum. The 2009/10 UK Database Disciplinary Commons project [12] illustrated both the focus on core elements and the wide range of topics taught. If cloud databases are to be added to an already stretched curriculum, then it is necessary to show that the technology is credible in database terms and from a teaching standpoint, that there are demonstrable benefits in terms of making teaching more relevant to the way in which databases are used, or increasing student engagement or deepening student understanding of core concepts, or preferably all three.

Opinions differ as to the relevance of cloud computing. There is a minority opinion that cloud computing is little more than a new name for existing technology – in 2010 the CEO of Oracle complained that “all the cloud is, is computers and networks .. they change a term and they think they’ve invented a technology” [13] - but Oracle itself has invested heavily in cloud applications. The Gartner 2010 survey [14] forecast a shift towards cloud computing; total market share is still small but there is evidence that cloud applications are becoming a mainstream technology and most significantly, from the database perspective, Oracle, IBM and Microsoft all have cloud versions of their mainstream databases; there are versions of MySQL and Postgres (Postgres-XC) for the cloud and there is a range of non relational cloud databases. The database industry may not have fully made up its mind, but it is certainly hedging its bets. To be complete, any discussion of current and evolving database technologies and architectures needs to include the study of databases in the cloud.

## 4. DATABASE TEACHING AND LEARNING

### 4.1 Database Theory and Practice

In 2010/11, cloud databases were introduced as a lecture topic into the database curriculum at Staffordshire University in two final year undergraduate modules and a masters level module. The topic generated real interest and debate, particularly at masters level, and a number of students went beyond the taught material. However, there were also some issues. Most students had had experience of installing an enterprise database, usually Oracle 11g or SQL Server Enterprise 2008/2010. A number of students accessed an SaaS cloud database discussed in the lecture sessions and were enthusiastic about the ease with which they could create an account and build a working database without having to install software. As they were using the database on a trial basis, they were not encountering real world issues or any of the limitations of the SaaS service model or being exposed to other service models. Some students felt that cloud database computing would mean there was no longer a requirement for optimisation and transaction management and argued that it spelt the end of traditional database administration, summed up by the comment 'So we'll all be out of a job'.

From a teaching perspective, the lessons drawn were that cloud databases are seen by students as a relevant and state of the art technology and one with which they are keen to engage. It was also concluded that cloud databases are not a topic which can be discussed in isolation, or without a practical element, because to do so overstates the attractiveness of cloud computing, which can appear as a simple answer to all configuration problems, and under represents the issues of managing data in the cloud.

Based on this experience, it was felt that a more holistic approach to the teaching of cloud databases was needed and that cloud databases, rather than being discussed as a separate topic, should be integrated into the teaching of advanced database concepts. There are two ways in which cloud database issues can be embedded in taught material. Firstly, in relation to the specialist focus of modules: one module, for example, has data management as a major topic and for this module, the cloud business model is relevant; another module has database security as a major topic and for this module cloud storage and deployment models are relevant. Secondly, in relation to core database concepts: this will provide context - cloud databases are one strand of database provision, not necessarily the whole future - and is also expected to enhance students' understanding and interest by connecting database theory to an attractive and current technology. The proposed teaching strategy is illustrated by examples based on ACID (Atomicity, Consistency, Isolation, Durability) and Query Optimisation. Other areas where cloud database concepts would be relevant include design, architectures, data models, distributed data, database tuning and transaction management.

- As ACID properties are enforced by DBMS, students in a teaching environment are unlikely ever to encounter a breach of the ACID protocol with the exception that some DBMS permit dirty reads. Given the current debate about how far ACID should be enforced in distributed and web based systems [15], advanced students are expected to understand the theory that underpins ACID, to understand strong and eventual consistency and the implementation issues. Cloud databases exemplify the problem well since some databases used in the cloud do not fully implement ACID due to the overhead [1,16] while it has been suggested that systems such as transaction processing which require strong consistency may be less suited to migration to the cloud [3]. This can be used to link the discussion of ACID and transaction management to issues presented by a state of the art technology.
- Students on advanced modules are taught a number of techniques for query optimisation such as indexing, cost based optimisation, query redesign, performance heuristics and similar. To demonstrate the effect of optimisation techniques, they are asked to populate tables with between 500000 and 3 million rows, using automatic data generation. This volume of data enables students to see the effect of the different techniques but does not address the question as to why optimisation is important. It is sometimes difficult for students to appreciate why measures other than time are significant; the benefits of reducing CPU usage, for example, are not obvious in a student environment. In a cloud environment, where the pricing model is partly based on the amount of data returned – EC2 charges data transfer out by GB/month - or on CPU cycles, there is an incentive to reduce the elements which cost most. A possible tutorial exercise is to give students a notional pricing schedule and ask them to design a query to execute as cheaply as possible in cash terms or to ask them to optimise the same query for different environments.

## 4.2 Practical Work with Cloud Databases

Practical work in this context – final year and masters level students - presents a number of challenges and opportunities. Cloud databases are only a small part of the Staffordshire University database curriculum and this limits teaching time and resources. The technology is changing very rapidly but at the time of writing, some cloud databases are only available in Beta version and some lack full functionality. As the technology evolves, teaching strategies will need to evolve to reflect this. Standardising on any one cloud database excludes all the others and limits student experience of an evolving technology. Relational databases delivered through the cloud are intended to provide users with an experience which is very similar to that provided by a traditional DBMS and it may be difficult, particularly with an SaaS service model, for students to recognise the significance of the cloud paradigm. This was one of the lessons drawn from current students' experience with an SaaS cloud database. Advanced students are required to understand the cloud business, deployment and service models but it is difficult to provide hands on experience of the different models if students are working with one, tutor selected, system. Advanced students can be expected to go beyond taught material and carry out some research themselves. Based on the work done in the current academic year, students are willing to engage with cloud database technology.

Given these factors, it is proposed for the short term, defined as the next academic year and a half, to build practical work on cloud databases around a collaborative approach which treats students as partners in the exploration of cloud technology; the aim is to build a developer/database administrator rather than an end user understanding of cloud databases. Spread over three academic semesters and a number of advanced database modules, this process will involve students in evaluating, commissioning, deploying and then re-evaluating a cloud database offering for use with advanced database modules. Figure 2 illustrates how the process is designed to operate:

| Semester 1 (2011/12)  | Semester 2 (2011/12)  | Phase 3 (2012/13)  |
|---|---|--|
| One element of the assessment for an advanced database module will be an evaluation of cloud databases. Students and tutors will jointly develop evaluation criteria and each student will be asked to evaluate one or more cloud databases chosen from a tutor supplied list which will include relational and non relational databases. The module has a technical focus and students will be expected to discuss relevant technical issues and to look at different data models. | In the companion modules in semester 2, which have a more business focus, students will be given the results of the previous cohort's evaluations and as part of the assessment, will be asked to review and update the evaluations, to extend them to cover elements such as costing, security and compliance and to recommend a cloud database for use in taught modules. Students will also be required to recommend a deployment model. | Subject to funding, students will be asked as part of the assessment to deploy a database in the cloud and to evaluate the outcomes. Given the cloud database pricing model, it is expected that a number of different cloud databases will be deployed. |

**Figure 2 Cloud Database Practical Work and Assessment**

The rationale for this approach is that students will be given the opportunity to take part in a real life procurement and deployment exercise since the intention is that the cloud databases selected will, subject to funding, be commissioned for use in the faculty, probably through a public cloud. The process is spread across several modules because as cloud databases are only a small part of the curriculum, the exercise cannot be allocated a larger share of any one module, but also because the cloud database environment is changing very rapidly and this approach builds in time to review the original recommendations. In the longer term, as cloud databases become more established and students become used to working with the technology, a different assessment strategy will be required.

## 4.3 Teaching Resources

Resources to support the teaching of databases in the cloud include textbooks, most usually as chapters in texts on aspects of cloud computing rather than as cloud database textbooks, academic papers, vendor and user material, and in many cases, free-to-use and demo versions of relational and non relational cloud databases, although sometimes with licensing issues. A list of resources is available at [www.fcet.staffs.ac.uk/cs8](http://www.fcet.staffs.ac.uk/cs8).



## 5. CONCLUSIONS AND FUTURE DEVELOPMENTS

Teaching and assessment for a new technology is always problematic; a similar paper written in the early 1990s would have assumed a wholesale shift to object-oriented databases. There are, however, solid grounds for believing that the cloud is likely to be a permanent feature of the database landscape. The technology is already established, there is support from the major database vendors and the business model is proving popular. What is uncertain is the impact that cloud technology will have on the database curriculum. The teaching and assessment approach discussed in this paper was designed for advanced students working with an evolving technology and the focus was on adoption and transaction management issues. This meant that other aspects of the curriculum, such as data warehousing, distributed database systems and database design, were not considered although these are all areas where cloud technology may affect practice and this in turn will influence teaching and learning. As the technology matures, it will become easier to separate implementation issues from underpinning concepts and the significance for the database curriculum will become clearer. This paper discusses the start of a process of teaching and learning.

## REFERENCES

- [1] Kraska T., Hentschel M., Alonso G. and Kossmann D., Consistency Rationing in the Cloud: pay only when it matters *J. Proceedings of the VLDB* **2:1**, 253-264 (2009)
- [2] Zhao J., Hu X. and Meng X., ESQP: An efficient SQL query processing for cloud data management *Proceedings of 2<sup>nd</sup> International Workshop on Cloud Data Management CIKM '10* Toronto Canada (2010)
- [3] Abadi D.J., Data Management in the Cloud: Limitations and Opportunities *IEEE Bulletin of the Technical Cttee on Data Engineering* **32:1**, 3-12 (2009)
- [4] Durkee D., Why Cloud Computing Will Never Be Free *Communications of the ACM* **53:5**, 63-69 (2010)
- [5] <http://blog.xeround.com/2011/04/newsq-cloud-database-as-a-service> accessed 20/5/11
- [6] Sultan N., Cloud Computing for Education: A New Dawn *International Journal of Information Management* **30**, 109-114 (2010) available on line at [www.elsevier.com/locate/ijinfomgt](http://www.elsevier.com/locate/ijinfomgt)
- [7] Holden E.P., Kang J.W., Bills D.P. and Ilyassov M., Databases in the Cloud: a work in progress *Proceedings of 10<sup>th</sup> ACM Conference on SIG-Information Technology Education SIGITE '09* (2009)
- [8] Mell P. and Grance T., The NIST Definition of Cloud Computing (Draft) (2011) *National Institute of Standards and Technology* [http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145\\_cloud-definition.pdf](http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf) accessed 21/5/11
- [9] Quality Assurance Agency for Higher Education Subject Benchmark Statement Computing (2007) <http://www.qaa.ac.uk/academicinfrastructure/benchmark/statements/computing07.pdf> accessed 23/5/11
- [10] Murray M. and Guimaraes M., Expanding the database curriculum *Journal of Computing Sciences in Colleges* **23:3** 69-75 (2008)
- [11] Hansen D.M., A sin of omission: database transactions *Journal of Computing Sciences in Colleges* **21:1** 225-230 (2005)
- [12] Database Disciplinary Commons 2009/10 <http://www.cs.kent.ac.uk/people/staff/saf/dbdc/index.html> accessed 23/5/11
- [13] <http://www.virtualization.tv/channels/fun/> accessed 23/5/11
- [14] <http://www.gartner.com/it/page.jsp?id=1438813> published 22 September 2010; accessed 23/5/11
- [15] Florescu D. and Klossmann D., Rethinking Cost and Performance of Database Systems *ACM Sigmod Record* **38:1** 43-48 (2009)
- [16] Leavitt N., Will NoSQL databases live up to their promise *Computer* **43:2** 12-14 (2010)



# MAKING DATA WAREHOUSING ACCESSIBLE

Tony Valsamidis  
University of Greenwich  
Park Row  
London SE10 9LS  
A.Valsamidis@gre.ac.uk  
www.cms.gre.ac.uk

---

## ABSTRACT

*This short paper describes how the area of data warehousing can be made accessible to a class of final year undergraduates on an information systems degree programme.*

## Keywords

*Data warehousing, Accessibility, Teaching and Learning*

## 1. INTRODUCTION

Data warehousing can be a daunting subject to approach at undergraduate level. The demands of large data sets, unfamiliar data models, unwieldy query languages and complex scripts can make the subject very inaccessible. For a cohort of not highly technical undergraduates in a one-semester business intelligence course, it is proposed that the main principles of data warehousing can be put across using a simple but realistic case study comprising of a data set of limited size implemented in Microsoft Access. An introductory reading that students are given echoes the data warehouse design process to be followed in a business context [1].

## 2. CASE STUDY

The case study used is a variant of the Wine Club, a standard text book example [2] [3]. The product data, comprising about 700 records, used to populate this data warehouse is gathered from a local wine dealer's website and initially put into a spreadsheet (Figure 1).

|     | A     | B            | C       | D             | E     |   |
|-----|-------|--------------|---------|---------------|-------|---|
| 1   | ID    | Category     | Vintage | Region        | Price | Description                             |
| 155 | 10008 | Red Wine     | 2006    | South America | £8,60 | Preludio Tannat 2006                    |
| 156 | 3037  | White Wine   | 2009    | Bordeaux      | £8,60 | Chateau La Rose du Pin Blanc 2009       |
| 157 | 82208 | Red Wine     | 2008    | Australia     | £8,60 | De Bortoli VAT 10 Pinot Noir 2008       |
| 158 | 21399 | Sundries     |         | Italy         | £8,70 | Zero36 Extraverigin Olive Oil, Sicily   |
| 159 | 45764 | Dessert wine | 2002    | Greece        | £8,70 | Samos Nectar 2002 37.5cl                |
| 160 | 50505 | Red Wine     | 2005    | Bordeaux      | £8,80 | Chateau La Rose du Pin Rouge 2005       |
| 161 | 85364 | Red Wine     | 2009    | Burgundy      | £8,90 | Manoir du Carra Beaujolais Village 2009 |
| 162 | 91621 | Red Wine     | 2004    | Other France  | £8,90 | Chateau Moulin Caresse 2004 Bergerac    |
| 163 | 29317 | Red Wine     | 2008    | Australia     | £8,90 | Four Sisters Merlot 2008                |
| 164 | 7168  | Red Wine     | 2006    | Australia     | £8,90 | Four Sisters Shiraz 2006                |

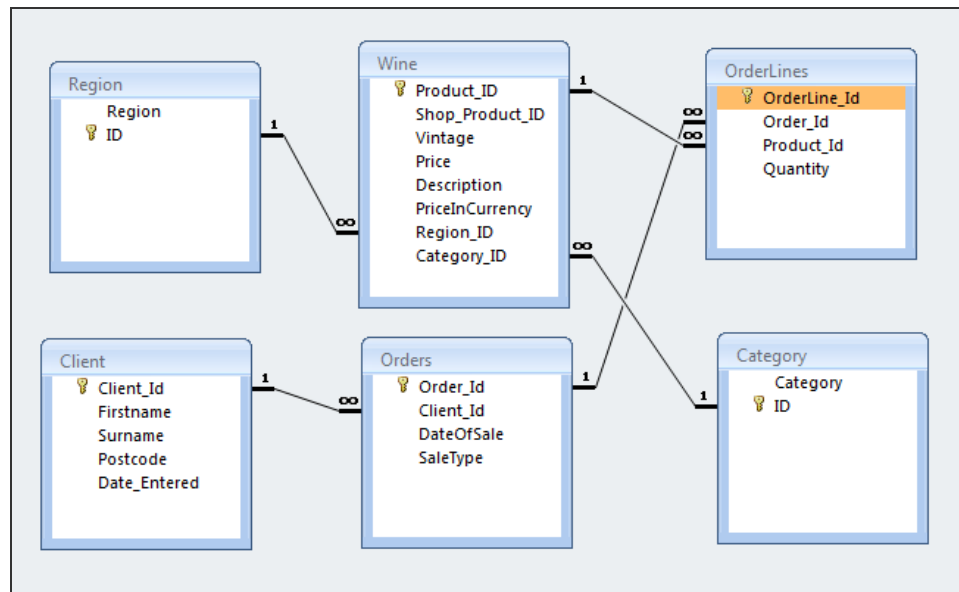
Figure 1: Wine product data in spreadsheet

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

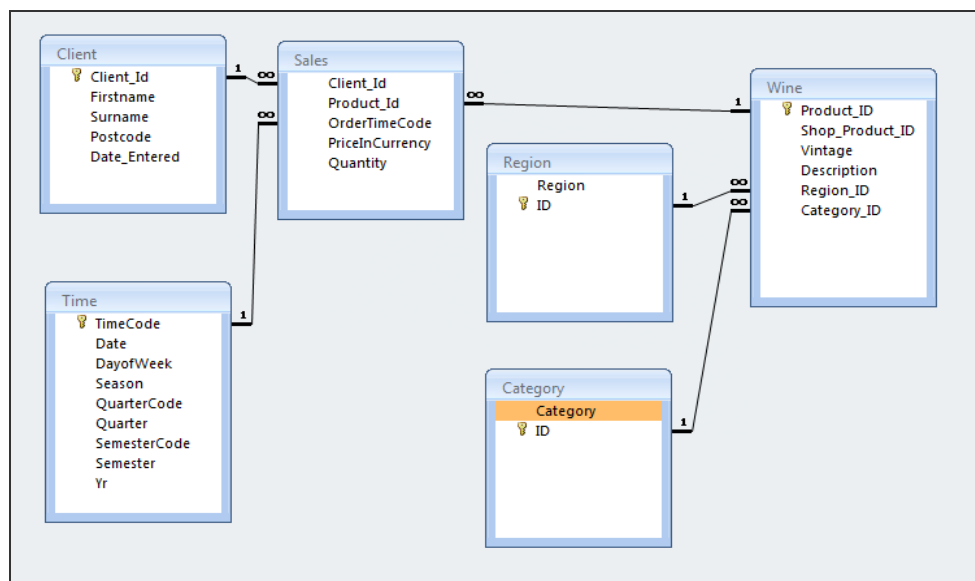
© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences

The product data is then merged with sales data that is modified from a separate database of anonymised customer order records. The customer orders online transaction processing (OLTP) system can be built quite easily from these elements using the import functions of Access (Figure 2). The product details derived from the spreadsheet contain a number of features which allow for the consideration and treatment of data cleansing concepts as well as format and semantic integration. This OLTP environment can then be transformed using simple rules into a star or snowflake schema model (Figure 3).



**Figure 2: OLTP Access Relationship Diagram**



**Figure 3: OLTP Access Relationship Diagram (snowflake schema)**

The data from the OLTP system can then transferred to the schema using appropriate Query by Example (QBE) queries to make the required tables using the Make Table function. In addition, a web-based generator is used for the time dimension data [4]. A set of directors' questions, such as "Which are our best selling wines at Christmas?", can then be taken through the interpretation process and converted into appropriate queries, which are expressed in Query By Example (QBE) form and executed to produce appropriate reports. This practical example is a useful animation of the basic model the textbook provides.

### 3. DISCUSSION

The implementation throws up an unexpected consequence in the treatment of keys in the warehouse model; this highlights the importance of examining the source data carefully when implementing such systems. In particular, the Sales fact table in Figure 3 does not have a unique key. This is a consequence of the data that is used, because the combination of the Client\_Id, Product\_Id and OrderTimeCode is not unique. This can form a useful discussion point that students can attempt to find solutions to.

Format integration and cleansing can be shown by some simple text recordset processing code to transform the price column from the Wine table from text type, e.g. "£8,60" to an appropriate numerical type which will allow for arithmetic and sorting operators. The other main transformation in the transfer of data from the OLTP system to the OLAP one is to convert the OrderDate into the appropriate OrderTimeCodes that reference the Time dimension table. This can once again be done by using recordset text processing and type conversion.

The chosen example can then be used as a basis for a coursework assessment which expects students to build a similar data warehouse, and to interpret the same kind of directors' questions in the same way in a different domain. This approach seemed quite successful and built on the students' prior knowledge of the development environment acquired in prerequisite courses.

An extension that has been considered but not yet tried involves the personalization of assessments through the randomization of the product keys before the data merger. Randomising the keys in this way may lead to less meaningful results, but provides a uniqueness in the results that can be a deterrent for plagiarism.

A drawback of this approach is that it does not really allow for the full exploration of performance issues, which do need large amounts of data. However, we feel that this approach is a good way of introducing the main principles of data warehousing to a non specialist audience.

### REFERENCES

- [1] Albescu, F., Pugna, I., Paraschiv, D. Business Intelligence & Knowledge Management – Technological Support for Strategic Management in the Knowledge Based Economy *Revista Informatica Economica*, **4(48)**, (2008).
- [2] Todman C. *Designing a Data Warehouse - Supporting Customer Relationship Management*. Prentice Hall (2000).
- [3] Open University, *M358 Relational Databases* (1998).
- [4] Regnecentralen/National Danish Center for Advanced Calculations (NDCAC), Time Dimension Generator/Builder, [http://www.regnecentralen.dk/time\\_dimension\\_generator.html](http://www.regnecentralen.dk/time_dimension_generator.html) [Last accessed 14 July 2011]



# ASSESSMENT IN DATABASE MODULES: PANEL DISCUSSION

David Nelson  
University of Sunderland  
david.nelson@sunderland.ac.uk

Al Monger  
Southampton Solent University  
Al.Monger@solent.ac.uk

Charles Boisvert  
Sheffield Hallam University  
c.boisvert@shu.ac.uk

---

## ABSTRACT

Database assessment is a perennial "hot topic" and has stimulated discussion at many previous TLAD workshops. It is hoped that this panel discussion will provide a coherent forum for discussing many aspects of database assessment.

The three panellists will introduce the topic briefly before inviting questions and contributions from the workshop participants.

Some of the topics envisaged are:

- Novel assessment strategies
- Effective and efficient online methods for providing formative and summative assessment feedback
- The role of examinations / projects / presentations / group work / etc.
- Plagiarism detection and designing assessment to avoid plagiarism
- Assessment strategies for different stages
- Departmental / institutional frameworks and constraints

Examples of questions include:

- Is there anything special about database assessment (compared with other subjects)?
- Does the method of assessment determine how and/or what you teach?
- We know after an assessment what was wrong with it, but how can we get it right?
- Can / should a subject like SQL be assessed by exam?
- Can plagiarism be "designed out"? How can plagiarism be detected in SQL? In a database application?
- Should projects require a database application to be built "from start to finish" (design to implementation)?
- Is it possible to determine a grade by marking exercises in-class?
- Do we over-assess?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2011 Higher Education Academy

Subject Centre for Information and Computer Sciences