George Mason University

CSI 710: Scientific Databases

CSI 710 Final Exam

Author:

Bruce Goldfeder

January 4, 2018

Instructions

Submission Instructions: Submit your exam in either MS Word or PDF format. Please put your name in your file name LastName_FirstName_FinalExam.doc For example Borne_Kirk_FinalExam.doc or Borne_Kirk_FinalExam.pdf

Please note: This is an individual exam subject to the GMU Honor Code. You may not consult with any other person in completing this exam or consult with any other person to discuss this exam. We are interested in your own creative answers to the questions. In many cases there is no single ”correct” answer. Rather, we are interested in your ideas and approaches to problem-solving. We will use GMU's licensed and approved originality-verification SafeAssign service within Blackboard. Excessive copying from other sources, without proper references and bibliographic attribution to the original source material, is a serious offense and may be reported to the University.

Instructions: This Exam is designed to examine your knowledge, opinions, and ability to synthesize new ideas on the topic of Scientific Databases. There are no ”trick questions” and we expect concise answers in your own words. In some cases you may have to use the Internet to find the answers, but remember that we want answers in your own words – although you are permitted to quote small sections from other sources, if you give proper bibliographic reference to those sources. You may use and quote materials from your own research papers or from other course term papers that you have written. Answer the questions here within this document.

The total time for this take-home portion of the final exam is 18 hours. Please keep track of the time that you spend, and then stop after 18 hours. Please answer all 7 (seven) questions on the following pages. NOTE: If you are thinking about these questions, when you are at work or driving or eating dinner, those hours do not count toward your total allowed hours. But if you are searching the Internet or reading different materials or doing other active work in order to develop your answers, then that time spent does count toward your total allowed hours.

1 Question #1

(20 Points) Describe the application of data mining (machine learning) concepts to a specific scientific research discipline. In particular, answer the following questions as part of your response:

a) What is the science discipline that you have chosen?

I have chosen the discipline of Internet Cyber Infrastructure and Security. The scope and definition of this discipline is inspired by but not limited to the NIST Preliminary Cyber Security Framework[1] . The ontology and supporting data sets that will be part of this recently created specialty can run the gamut from internet topological information to industrial control systems to cyber warfare. My current research at DARPA is measuring and analyzing topological trends in the Internet landscape and how they change over time. In essence we are attempting to begin defining the “physical laws of the universe” for this domain space in and creating new human computer interfaces to convey this information such as the Oculus Rift[2].

This specific data set is well suited for a Big Data Scientific Database. The amount of data is beyond astronomical. In the year 2012, over 2.5 quintillion bytes of data were created every day, and 90% of the worlds data has been created in the past two years[5]. A recent paper by Cisco, The Zettabyte Era—Trends and Analysis estimates future internet traffic as follows, ”Annual global IP traffic will pass the zettabyte threshold by the end of 2015, and will reach 1.4 zettabytes per year by 2017.”[6]

b) Describe the role of data mining (machine learning) in the science discovery process within the science discipline that you have chosen.

Data Mining and machine learning at heretofore unforeseen scales are the basis of providing information and knowledge out of the chaos of the internet. Applying this knowledge to perform useful tasks such as resource allocation, behavior and usage predictions, as well as cyber security will require fast and meaningful analytical techniques that will need to be horizontally distributed to handle the volume, velocity, and variety of data of this discipline. Archives of full or partial packet captures from internet traffic correlated to aberrant events both accidental and intentionally malicious are available to be used as training sets. Two sets that are publicly available are the Shodan[3] vulnerabilities database and the Symantec World Intelligence Network Environment (WINE)[4]. These data sets provide or can be leveraged for both unsupervised and with some simple transformation labeled data sets.

One of the roles of data mining (machine learning) in analyzing internet traffic is the classification of IP Traffic. Various papers are written on the subject, a recent one which I have reviewed is ”Comparative Analysis of Five Machine Learning Algorithms for IP Traffic Classification.”[7] This paper describes five methods of ML applied to captured IP traffic containing various types of traffic. These are classified into traffic correlated to seven internet applications and protocols, www, e-mail, web media, P2P, FTP data, instant messaging and VoIP. Five well known machine learning algorithms are applied to the data sets. These are as follow: Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF), C 4.5 Decision Tree Algorithm, Bayes Net Algorithm and Naïve Bayes Algorithm.[8] The experiments performed in this paper resulting in 94% accuracy for the Bays' Net and C4.5 algorithms. Identification of traffic, usage patterns over time, and discovery of unknown protocols serve an important role in managing quality of service, capital expenditures in networking, and security for organizations. This information can be assessed to provide knowledge into how your networks are being used.

c) Describe and differentiate these two generic data mining approaches: Supervised Learning (Classification) and Unsupervised Learning (Clustering).

Supervised learning such as classification makes use of training sets, similar to the Kaggle Acceleromter Project that our team worked on. Given this labelled training set, in our case, the "train" database, we were able to apply various methodologies including support vector machines and linear regression. These functions were then applied to the test set, in our paper, the "test" data set. The resulting classification was sent into kaggle for scoring.[11]

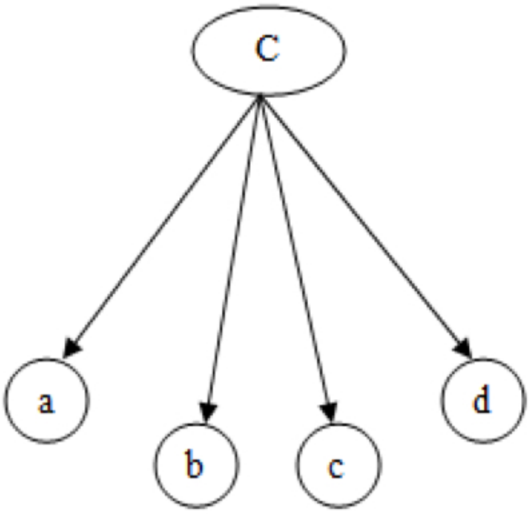
Unsupervised Learning work without the advantage of a training set and are often used in conjunction with supervised learning techniques in a discover and learn pattern. The difference is that there is no starting labelled data set with the "answers". The intent of these types of algorithms are to discover the structure within a set of data such as clustering. We also leveraged techniques of this kind in our Kaggle report during our feature generation phase. After cleaning the data set we applied a Hidden Markov Model in order to help discover features or covariants in our data set. The intent was that different activities such as walking or driving can be modelled as hidden states. In order to help with identifying these states we leveraged a k-means clustering algorithm for different numbers of states. While a good excersize and explanation of the use of unsupervised learning to support supervised learning, in our experiments we were unable to determine meaningful results from this[11].

d) Describe one specific detailed scientific example of the Supervised Learning approach (e.g., Classification, Regression, Bayes Networks, Markov Modeling) for the specific science discipline that you have chosen. Your example should describe the types of data, the availability of training data sets, the types of learning algorithms, and what specific scientific outcomes are expected.

In the Singh et al[7] paper, "Comparative analysis of five machine learning algorithms for IP traffic classification", one of the classifying techniques applied was Naive Bayesian (NB). This algorithm uses a class node as the parent node of all other nodes as depicted in the below figure 1.

Note the main class identified as C and the correlated attributes a thru b.

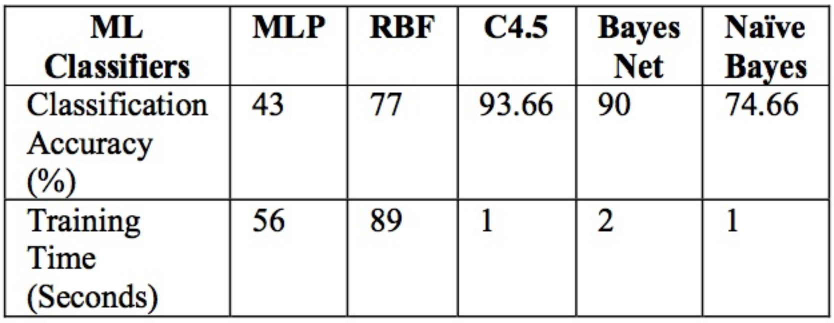
Figure 1: Representation of a Naive Bayesian Classifier



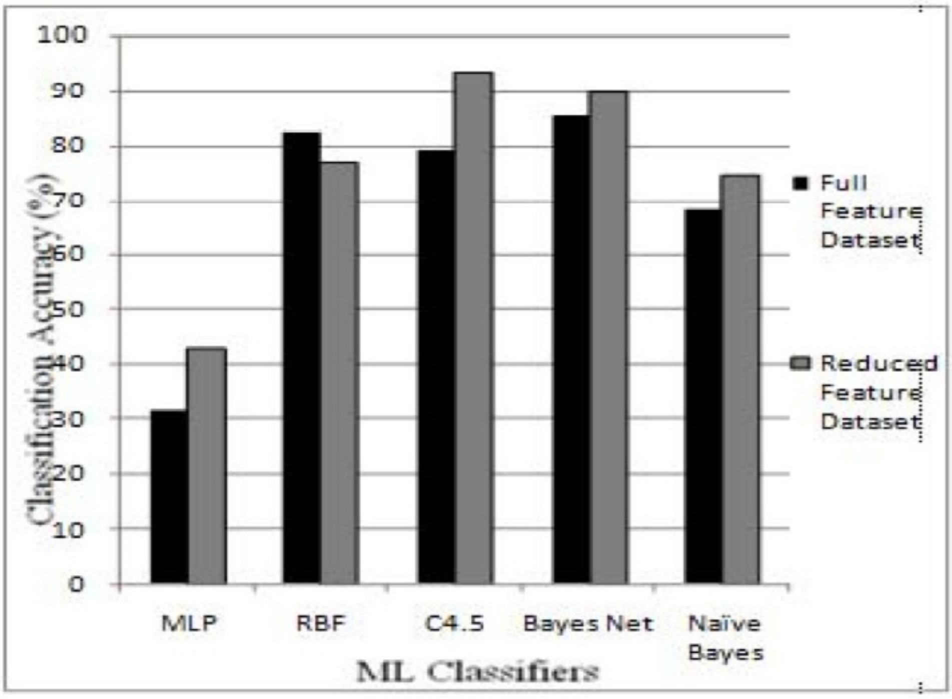
This simple structure is constrained to just this structure. This simplification makes this easier to apply removing the burden of identifying any other structure. Another simplification is that Naive Bayes assumes all the features are independent of each other.

This methodology was applied to two different datasets one comprised of a full featured data set with 261 features which mainly consist of minimum, maximum, mean, variance and total values of the number of IP packets, average packets per second, packet size. While reduced feature dataset is obtained from full feature dataset using cfsSubsetEval evaluator and Best First search in attribute selection filter of Weka tool[12] The results of the application of this algorithm on the dataset are presented in figure 2. along with the results of the four other techniques. The Naive Bayesian did not fare as well as the Bayes Net and the C4.5.[7]

Figure 2: Results of the five ML algorithms for Classifying IP Traffic[7]



ML Classifiers	MLP	RBF	C4.5	Bayes Net	Naïve Bayes
Classification Accuracy (%)	43	77	93.66	90	74.66
Training Time (Seconds)	56	89	1	2	1

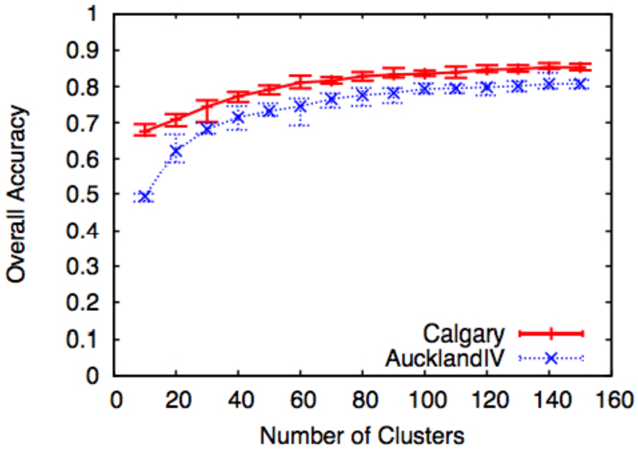


e) Describe one specific detailed scientific example of the Unsupervised Learning approach (e.g., Clustering, Association Analysis, Link Analysis, Correlation Analysis) for the specific science discipline that you have chosen. Your examples should describe the types of data, the type of analysis, the types of learning algorithms, and what specific scientific outcomes are expected.

A good description of using unsupervised learning within the IP traffic domain is presented in the Erman et al. paper, "Traffic Classification Using Clustering Algorithms."[13] This paper uses two unsupervised learning approaches, K-Means clustering[16] algorithm and DBSCAN[15]. This is similar to the previous supervised learning approach in that the intent is to also classify IP traffic. The difference is that this paper leverages unsupervised techniques in order to cluster the data into these different types. The results of these experiments were then compared to prior results that were gained using the AutoClass[14] algorithm using empirical internet traces. The experiments were run on two data sets defined as Auckland IV and Calgary.

Within the K-Means section of this paper, they identified the input parameter of K, the number of disjoint partitions, to represent each traffic class. Due to the diversity of the traffic within the classes it was expected that each class could correlate to more than one cluster. The K-Means algorithm was run using an initial value of k to be 10, and by adding 10 for each susequent run. Their results were that initially, when the number of clusters was small the overall the accuracy was at 49% for the Auckland IV data sets and 67% for the Calgary data sets. Their observation was that the overall accuracy steadily improved as the number of clusters increased. The continued increase in percentage proceeded steadily until K was 100, whereupon the increases tapered off. There was marginal increases to about 80% when K was 500 however this runs into the risk of overfitting. These results are depicted below in figure 3.

Figure 3: Results of K-Means Clustering Algorithm on IP Traffic[13]



2 Question #2

(15 Points) Describe the application of distributed computational science and distributed data science concepts in the new e-Science research paradigm. In particular, answer the following questions as part of your response:

a) What is e-Science?

E-Science is bringing data to knowledge and is the answer to the question, "How do we make large distributed heterogeneous data collections findable, accessible, and usable for scientists worldwide?" The advent of the Big Data age make transporting data sets continuously and redundantly around the internet infeasible. By embracing this concept of having data remain in situ and providing services to efficiently access this data and its corresponding metadata enable e-science. E-Science encompasses the tools of the internet like REST, XML, ontologies, and communication standards and protocols to enable this level of data sharing, access, and security.

- E-Science
  - Built on Web Services (eCommerce) paradigm
  - XMLbased
  - Distributed heterogeneous data are the norm
  - Data integration across projects & institutions
  - One-stop shopping: "The right data, right now"
  - Examples: Virtual Observatories, or FGDC Geospatial "One-Stop"

b) Describe the role of distributed federated data access (e.g., Virtual Observatories) within the e-Science scientific discovery paradigm.

In the age of Big Data and Data Sharing organizations and communities need to find mechanisms to share data while retaining the individualism and stewardship of their unique data sets. Federated databases such as the Virtual Observatories represent this ability for all to contribute, access, enhance, and even curate a distributed collection of data sets.

c) What are the key e-Science technologies that make distributed data access possible and successful?

The underlying infrastructure to enable this such as Service Oriented Architecture (SOA) and common semantics, transport formats, and transport protocols are all enablers for this technology. Web Services, XML, Http and REST are examples of these technologies. All of these utilize the tenets of SOA, intified in the CSI Class SlideSet Week #11 as:

- SOA Characteristics
  - Logical View of Services
  - Message Orientation
  - Description Orientation

  - Granularity
  - Network Orientation
  - Platform Neutral

d) Describe the role, the strengths, and the weaknesses of XML within e-Science-enabled scientific databases. Provide specific information about each item (role, strengths, and weaknesses).

   The role of XML is a markup language that enables organizations and communities to create standards for the exchange of data. An early adopter of this technology was the chemistry community which came up with the Chemical Markup Language or CML, based on XML. The strength of this is that there is a well defined mechanism which chemists can convey all molecular information to each other. The weakness of this is when there are multiple standards and standards organizations including proprietary ones. Another weakness that is related to this is the maintenance of the standard and not making it to obtuse for the average user.

e) Describe the potential role of semantic technologies (e.g., ontologies, taxonomies, user annotations, tagging systems, or folksonomies) in the future of e-Science-enabled scientific research. Provide specific examples.

   Semantic technologies are critical for e-science just as a common lexicon is needed for anyone to communicate. The concepts of aggregating and abstracting complex scientific concepts into terms or other labels is difficult even within my own office at work, it easily becomes a tower of babel when applied worldwide amongst people of different backgrounds and references. A specific example is that we have a person is literally the ontologist on our program. There currently does not exist a lingua franca for describing cyber events that is common. This leads to mis-communications lack of communications, and fracturing of the community sometimes based on what open source or proprietary tool you are using. The logical view of services is lost and often times large, ambiguous files of unstructured or semistructured text is the result.

3 Question #3

(15 Points) Describe the applicability of Intelligent Archive concepts for scientific databases. In particular, answer the following questions as part of your response:

responses cite the CSI 710 slide set 14 as applicable

a) Briefly describe an application of Intelligent Archive concepts to a specific scientific database project.

An Intelligent Archive concepts enabling a scientific community is the Virtual Observitory. It is an application of the following concepts identified in the week 14 slideset:

- IA Concepts
  - All items stored to support "end-to-end" research
    - Data, information (metadata), and knowledge (ontologies)
    - Software and processing needed to manage holdings and improve self-knowledge
    - Interfaces to algorithms to help transform information into knowledge
  - Architecture is highly distributed
  - Evolved, advanced functions
  - Future looking and future proof
  - Highly adaptable

The Virtual Observatory contains accessible data sets that adhere to the standards and ontologies as defined by the International Virtual Observatory Alliance (http://www.ivoa.net/). This system of systems integrates observatories, archives, and users with VU Physics (code), VU Astrophysical Objects, and Virtual Telescope to research, process, and create simulations worldwide. The system architecture is service oriented and extensible.

b) What role does data mining have in Intelligent Archive projects?

Intelligent Archives enable Intelligent Web Mining which is an advancement above Web Searching in that applies analytic techniques as well as collaborative and individual preferences. One example of this is the website www.clusty.com which clusters results into semantic clusters. Examples of both collaborative and personal preferences are common on E-Commerce sites such as Amazon.com. These are presented in an intuitive manner as recommended purchases for you, and other folks that looked at this ended up buying this type of capabilities. These capabilities are very useful and have become somewhat commonplace, but the realization of these advanced capabilities rely on data mining techniques such as clustering, association, and link analysis.

c) What role do semantics and ontologies have in Intelligent Archive projects?

   Semantics and Ontologies provide the language with which heterogenous data, services, and communities can communicate. I would state that it is the key element required for distributed systems to interact.

d) What is the role of collaboration and data/knowledge-sharing in Intelligent Archive projects?

   Collaboration and data/knowledge sharing are critical to Intelligent Archive projects. The intent of transforming data to information to knowledge implies both algorithmic and human involvement. U-Science techniques such as crowd-sourcing can mitigate complex analytics such as image interpretation in the GalaxyZoo project (www.galaxyzoo.org). Data sharing either implicit on an E-Commerce site or explicit by open sourcing code and data enable the aggregation of data to create new discoveries, patterns, and knowledge.

4 Question #4

(25 Points) Describe the key features, attributes, and characteristics that distinguish scientific databases from other types of databases (e.g., commercial). In particular, answer the following questions as part of your response:

a) Describe at least 5 features of scientific databases that are not common features of commercial databases.

- Features of Scientific Databases
  - Measurement Errors Scientific data generally come with error e.g. +/ .5mm
  - Equipment and Experiments Scientific data depend on the techniques and tools used
  - Calibration Techniques requires validation against independent samples
  - Levels of Abstraction raw, calibrated, extracted information, knowledge, published. For example you collect data, express in a spreadsheet, write in words in a paper, and derive a formula
  - Metadata Metadata are a critical component of a scientific database

b) How are the data and metadata types found in these database systems different (scientific data/metadata versus commercial data/metadata)?

Metadata is the description of the data. While in a commercial database this may or may not be relevant or it may have limited context, such as the ISO listing of country names to abreviations. In a scientific database metadata is critical to accomplish the following as described on slide 11 CSI 710 Slide Set Week #2:

- Administrative: who, where
- Content: what (data)
- Structural: how (syntax; for data use and re-use)
- Data pedigree (Provenance): when (and how)
- Semantic (Context): why and what (domain)

These correlate to attribution of who created and modified the data, how they modified it, when, and why. The structural portions can apply to ontologies or identifying units and allow the data to be understood and used by the community. Withing our group paper we leveraged the Dublin Core for identifying many of these attributes.

Commercial databases are usually transaction in nature in well defined paradigms. With few exceptions, scientific databases are more inclined to be in the Big Data, Petabyte+ range than a typically commercial database.

c) How are the types of analyses performed on scientific data different than the analyses performed on commercial data?

Commercial databases are concerned more with transactional data sets. This being the case the analytics are typically aggregations and rollups such as monthly sales reports. Scientific databases deal with highly complex datasets that involve correspondingly complex analysis. A sampling of this is Numerical Weather Prediction, numerical simulations for fluid dynamics, and gene sequencing. The amount of data will be typically in the Big Data range of PB+. Sensors or measurements such as those at CERN can be in the billions or trillions. The data content can be complex such as images, spatial data, temporal sequences. Different levels of work, such as the Nasa levels of abstraction require provenance to track and reproduce intermediate data products. Statistical analysis and tools for assurance are typically used and need to be integrated into a scientific database system. Parameter or category data categorize the measured data, measurement data are the data that will be summarized and analyzed

d) What is database normalization? Why is it used?

Originally introduced by E.F. Codd in his canonical papers in the early 1970s as a way of configuraing the fields and tables in a relational structure in order to minimize dependendencies and redundancies. The forms were identified by both Codd and Raymond Boyce as a series of Normal Forms, 1 NF, 2NF, 3NF, Boyce-Codd NF, additional forms exist 4 NF, and 5 NF but are not typically used in standard practice. The purpose of this organization was to enable logical functioning of a unified querying language called Structured Query Language (SQL). It also serves to preserve the integrity of the database keeping foregin keys intact. Prevents inappropriate or unwanted insertions and deletions accidentally. It serves to distinguish functionally distinct entities to remove ambiguity of entities. It also serves to seperate functionally independent attributes. The three major functions are to elim

e) Describe why or why not database normalization is used in scientific databases. Provide the scientific rationale (i.e., science-based explanation) for your answer.

Integrity of a data set and maintaining appropriate correlation to metadata such as provenance is critical in scientific databases. At a minimum 3rd normal form needs to be maintained, Boyce-Codd preferred in order to accurately insert, update, and delete data from within the database. Otherwise data integrity and quality issues could quickly eliminate the utility of a data set. In a modern federated system, you will quickly observe that your data will not be accessed and begin to ”rust.”

5 Question #5

(15 points) Spend some time examining and exploring the Internet Movie Database (IMDB: http://www.imdb.com/), then suggest a similar solution for a science DB project. In particular, answer the following questions as part of your response:

a) What is the field of science that you have chosen?

I have chosen the field of cyber security as the field of science to represent in my science database project.

b) Describe your own idea for something similar to "IMDB" for your science discipline.

The IMDB database and website provide a "one-stop-shopping" location for you to find out everything you would need to know about the entertainment industry. The architecture is not dissimilar in logical functionality to that of the Virtual Observitory in that it integrates numerous sensors, databases, archive, functions, collabortive capabilities (e.g. ratings) in a worldwide accessible database that extends to applications on mobile devices. The visualization is very user friendly and presents the relevent information in a portal environment at an abbreviated or high level fashion enabling the user to drill down for greater detail. This paradigm would be perfect to overlay onto internet and cyber technologies. By open sourcing collected information on internet traffic along with algorithms to analyze the data and visualization to help understand the data would provide the basis for unifying what is now a fractured community. Just as scientific data in the past was kept hidden, internet information is still in the realm of nation states and proprietary companies. Creating a universally accessible and easy to use web site with federated databases would be a compelling leverage point to open up this scientific area.

c) Outline the requirements (e.g., list the "10 questions") for your "IMDB"-like system.

(a) What is the average daily traffic of the internet?

(b) How does todays traffic relate to yesterdays?

(c) How many identified vulnerabilities were identified in New Jersey last month?

(d) What transport protocol is the most used by data size?

(e) What transport protocol is the most common by number of connections?

(f) How much email transited into and out of Virginia last month?

(g) What state has the highest average internet speeds?

(h) What percentage of internet traffic by size is used by Netflix?

(i) How many internet outages are observed in California per month last year?

(j) What is the average latency for downloading YouTube videos?

d) Summarize the key features of your system (in terms of e-Science, database, and data-related concepts that we studied in this class).

The key features of the system are to create a federated systems of internet traffic, services, and security. The combined data set would easily reach into the Big Data range even without package capture requiring distributed computing techniques in order to analyze. An agreed upon ontonlogy with overlays into XML would be required for data storage, transport, and usability. Provenance of the data sets will need to be enabled and maintained in order to de-duplicate data, ensure integrity, and quality confidence for using the data. Collaboration on various data sets will enable open monitoring and open standards enabling data to more rapidly move from information to knowledge.

e) Describe at least one new feature that you would like to see as part of your system that is not part of the existing IMDB. What are the benefits of this new feature (for either the scientist end-users or the scientist providers of your system)?

Data Mining and Machine learning as exposed applications would need to be integrated into the system I would be creating. This is behind the scenes (no pun intended) on IMDB, but within my internet and cyber system this would need to be exposed as a first class capability for not just my personal data sets, but also to have the ability to perform distributed mining of data across distant and possibly heterogeneous data sources.

6 Question #6

(5 points) The New Data-Intensive Era of Big Data Science: Describe the single most influential topic (or concept) that you studied in CSI 710 this semester, and describe how it will affect your future work (your graduate course selections, or your research, or your PhD thesis topic, or your career plans). Specifically, you should provide a description of the topic, what you found to be most important about that topic, and how has it influenced your future plans.

The single most influential topic that I studied in CSI 710 has been that of the Virtual Observatories. The differences between distributed data minin (not so good) and the mining distributed mining of data (efficiently using metadata) are a common issue within the data sets I have worked with in the past. I have already briefed to my colleagues on the concepts of metadata, standardizations and ontologies that enabled data to remain in situ and be mined efficiently using the indexes on the metadata. Within the realm of ontologies and standards, we are researching the use of the VOEvent[9] format as well as the VOEvent Transport Protocol (VTP)[10] for use in describing and transmitting cyber events.

7 Question #7

(5 points) Peer Evaluation: Describe your role and interaction with other members of your group on your Group Homework Assignments and your Group Project. Please address the following questions when peer-evaluating your group members' performance: Did all members perform well? Did anyone in your group not contribute their fair share of the group's work load? Please explain. You may specify the contribution of each team member to the group homework and to the group project separately (by person's name) by using either their percentage contribution (total 100%) or by providing qualitative assessment: "[Well] [Above/Below] Average".

All in all we had a really good time working on the project. Each of us had different backgrounds and contributed where we could. I helped by putting together the team website, www.gmukaggle2.com and used it for storing and collaborating on files. Along with John, I put together the SQL database, created the ingest scripts, the logical and physical models, and created SQL queries and stored procedures for about half of the 10 questions, wrote up the data and data management sections for both the presentation and the paper.

Christine Harvey was very helpful in putting together the presentation and created our intial format of the paper on WriteLatex.com (for the paper we collaborated on this site). she completed the kaggle research and presentation.

Hillary Dennison was helpful in putting together the 10 questions and assisting John and I in addressing them. She did the data leaks analysis and presented this and did the write up in the paper.

John Riddles was by far the most prodigious member of the group. He has a very extensive knowledge of statistical analytics that far exceeded the other members. He performed the more complex analytics and provided the formal submission to the Kaggle project.

I look forward to being classmates and working with each of my teammates in the future.

References

[1] NIST. *Preliminary Cybersecurity Framework*, available at `http://www.nist.gov/itl/upload/preliminary-cybersecurity-framework.pdf`.

[2] Oculus VR, *Virtual Reality Headset*, available at `http://www.oculusvr.com/`.

[3] Shodan *Exposing Online Devices*, available at `http://www.shodanhq.com`.

[4] WINE. *Symantec World Intelligence Network Environment*, available at `http://www.symantec.com/about/profile/universityresearch/sharing.jsp`.

[5] Science Daily. *Big Data, for Better or Worse: 90% of World's Data Generated Over Last Two Years*, available at `http://www.sciencedaily.com/releases/2013/05/130522085217.htm`.

[6] Cisco. *The Zettabyte Era—Trends and Analysis*, available at `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/VNI_Hyperconnectivity_WP.html`.

[7] Singh, K. ; Univ. Inst. of Eng. & Technol., Panjab Univ., Chandigarh, India ; Agrawal, S. *Comparative analysis of five machine learning algorithms for IP traffic classification*, available at `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5958481`.

[8] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques, 2th edition*, Morgan Kaufmann Publishers, San Francisco, CA, 2005.

[9] IVOA. *Sky Event Reporting Metadata Version 2.0*, available at `http://www.ivoa.net/documents/VOEvent/index.html`

[10] IVOA, *VOEvent Transport Protocol Version 1.1*, available at `http://www.ivoa.net/documents/Notes/VOEventTransport/`

[11] GMU CSI 710, *Kaggle Team #2*, available at `https://mymasonportal.gmu.edu/webapps/assignment/uploadAssignment\?content\_id=\_3657343\_1\&course\_id=\_223063\_1\&assign\_group\_id=\&mode=view\#`

Weka website. Available: http:// www.cs.waikato.ac.nz/ml/weka/

[12] University of Waikato, *Weka 3: Data Mining Software in Java*, available at `http://www.cs.waikato.ac.nz/ml/weka/`

[13] Erman et al, *Traffic Classification Using Clustering Algorithms*, available at `http://mars.cs.kent.edu/~peyravi/Bibs/TE/erman06.pdf`

[14] P. Cheeseman and J. Strutz. *Bayesian Classification Theory and Results. In Advances in Knowledge Discovery and Data Mining* AAI/MIT Press, USA, 1996.

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu. *A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.* In 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD 96), Portland, USA, 1996.

[16] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, USA, 1988.