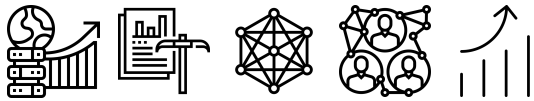


CSC 5741 (2020/21)
Data Mining and Warehousing
Lecture 1: Administrivia, Course Overview and Introduction

Lighton Phiri
Department of Library & Information Science
University of Zambia
http://lis.unza.zm/~lightonphiri

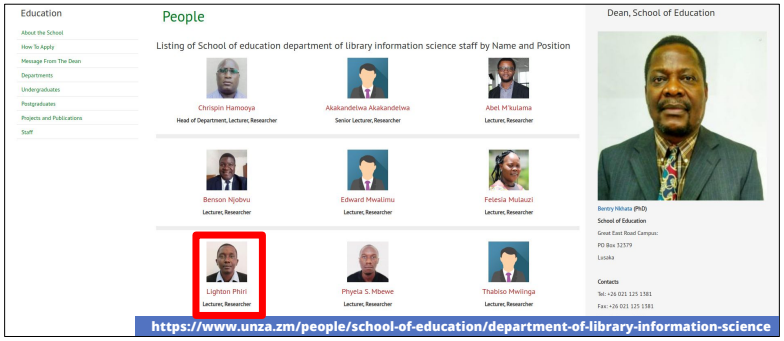
Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: How to Read a Paper
- Part IV: On Academic Activities
- Part V: About Next Week

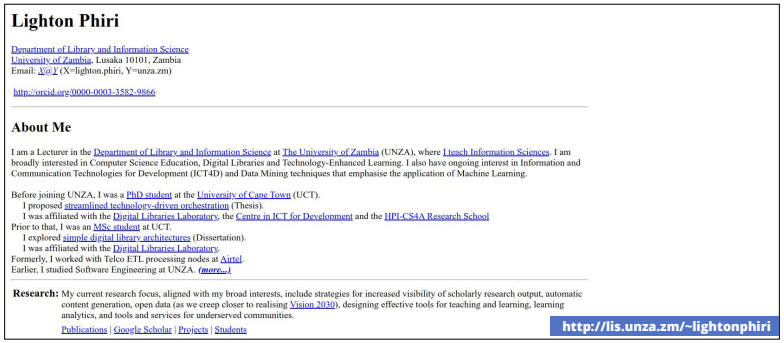
Lecture Series Outline

- Part I: Administrivia
  - Personal Introductions
  - Learning Outcomes
  - Course Structure
  - Prescribed Books
  - Tools and Services
  - Course Grading, Academic Dishonesty and Course Management
- Part II: Course Introduction
- Part III: How to Read a Paper
- Part IV: On Academic Activities
- Part V: About Next Week

Personal Introductions (1/5)



Personal Introductions (2/5)



Lighton Phiri

Department of Library and Information Science
University of Zambia, Lusaka 10101, Zambia
Email: X@Y (X=lighton.phiri, Y=unza.zm)

http://orcid.org/0000-0003-3582-9866

About Me

I am a Lecturer in the Department of Library and Information Science at The University of Zambia (UNZA), where I teach Information Sciences. I am broadly interested in Computer Science Education, Digital Libraries and Technology-Enhanced Learning. I also have ongoing interest in Information and Communication Technologies for Development (ICT4D) and Data Mining techniques that emphasise the application of Machine Learning.

Before joining UNZA, I was a PhD student at the University of Cape Town (UCT).
I proposed streamlined technology-driven orchestration (Thesis).
I was affiliated with the Digital Libraries Laboratory, the Centre in ICT for Development and the HPI-CS4A Research School
Prior to that, I was an MSc student at UCT.
I explored simple digital library architectures (Dissertation).
I was affiliated with the Digital Libraries Laboratory.
Formerly, I worked with Telco ETL processing nodes at Airtel.
Earlier, I studied Software Engineering at UNZA. ***(more...)***

Research: My current research focus, aligned with my broad interests, include strategies for increased visibility of scholarly research output, automatic content generation, open data (as we creep closer to realising Vision 2030), designing effective tools for teaching and learning, learning analytics, and tools and services for underserved communities.

Publications | Google Scholar | Projects | Students

http://lis.unza.zm/~lightonphiri

Personal Introductions (3/5)

Lighton Phiri

Department of Computer Science, University of Cape Town
Rondebosch 7701, Cape Town, South Africa
Email: X@Y (X=lphiri, Y=cs.uct.ac.za)

http://orcid.org/0000-0003-3582-9866

About Me

I was a PhD student in the Department of Computer Science at the University of Cape Town.
I explored Technology-driven Orchestration and was supervised by Hussein Suleman and Christoph Meinel.
I was affiliated with Digital Libraries Laboratory, Centre for ICT for Development and HPI-CS4A.

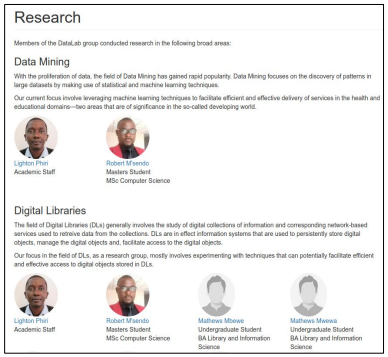
Prior to that, I was an MSc student in the Digital Libraries Laboratory. I investigated Simple Digital Libraries and was supervised by Hussein Suleman.
Formerly, I worked with Telco ETL processing nodes at Airtel.
Earlier, I studied Software Engineering at the University of Zambia.

Additional Information

Teaching: *CSC1015F: Computer Science 1015* – [2016] [notes] [code] [slides] [videos]
CSC1017F: Python for Engineers – [2015] [notes] [code] [slides] [videos]

https://people.cs.uct.ac.za/~lphiri

Personal Introductions (4/5)



Research

Members of the DataLab group conducted research in the following broad areas:

Data Mining

With the proliferation of data, the field of Data Mining has gained rapid popularity. Data Mining focuses on the discovery of patterns in large datasets by making use of statistical and machine learning techniques.

Our current focus involve leveraging machine learning techniques to facilitate efficient and effective delivery of services in the health and educational domains—two areas that are of significance in the so-called developing world.

Lighton Phiri
Academic Staff

Robert Mwendo
Masters Student
MSc Computer Science

Digital Libraries

The field of Digital Libraries (DLs) generally involves the study of digital collections of information and corresponding network-based services used to retrieve data from the collections. DLs are in effect information systems that are used to persistently store digital objects, manage the digital objects and, facilitate access to the digital objects.

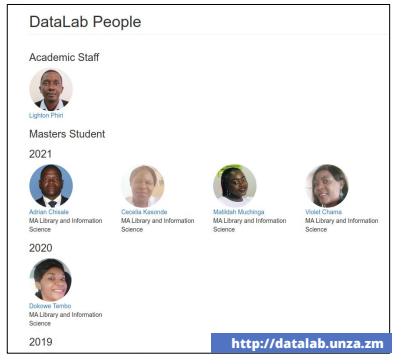
Our focus in the field of DLs, as a research group, mostly involves experimenting with techniques that can potentially facilitate efficient and effective access to digital objects stored in DLs.

Lighton Phiri
Academic Staff

Robert Mwendo
Masters Student
MSc Computer Science

Mathews Mbewe
Undergraduate Student
BA Library and Information Science

Mathews Mwewa
Undergraduate Student
BA Library and Information Science



DataLab People

Academic Staff

Lighton Phiri

Masters Student

2021

Adrian Chisale
MA Library and Information Science

Cecelia Kasonde
MA Library and Information Science

Mallitah Muchinga
MA Library and Information Science

Violet Chama
MA Library and Information Science

2020

Dokowe Tembo
MA Library and Information Science

2019

http://datalab.unza.zm

Personal Introductions (5/5)

- **Your full names and preferred reference (first name, Mrs./Ms.Mr. X)**
- **Your formal education background**
- **What you are presently upto (THINK: what you do for a living)**
- **What you hope to get from CSC 5741**

CSC 5741 Learning Outcomes

- Identify the key processes of data mining, data warehousing and knowledge discovery process
- Describe the basic principles and algorithms used in practical data mining and understand their strengths and weaknesses
- Apply data mining techniques to solve problems in other disciplines in a mathematical way
- Apply data mining methodologies with information systems and generate results which can be immediately used for decision making in well-defined business problems

CSC 5741 Desired Outcome

- Desired outcome, for me, is to ensure we are all in a position to successfully undertake a Data-driven Research Project.
- [...]
- Data Mining "Research" Project
- Practical Knowledge
- Experimentation
- Evaluation Strategies
- Ethics and Bias
- [...]
- We will need to read and discuss what others have done

Course Structure (1/9)

- CSC 5741 is a half course
- CSC 5741 will be run using a seminar session
  - One three hour-long lecture session per week
    - One seminar every fortnight.
    - Paper reading sessions every fortnight.
    - Formal lecture session with theory and practical walkthroughs.

Course Structure (2/9)

- Tentative Lecture series and session structure
  - Lecture session (120 minutes)
  - Paper discussion (30 minutes)
  - Seminar session (30 minutes)
- We will tentatively spend two weeks on each CSC 5741 theme

THE UNIVERSITY OF ZAMBIA
COMMUNICATION AND MARKETING DEPARTMENT

Great East Road Campus| P.O Box 32379| Lusaka, 10101| Tel: +260 211 251 593|
Fax: +260 1 253 952 | Email: comm.marketing@unza.zm | Website: www.unza.zm

PRESS STATEMENT

RE-OPENNING OF THE UNIVERSITY OF ZAMBIA IN A PHASED APPROACH

09th February 2021 - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09th February 2021, has resolved to review the University Senate resolution of 27th January 2021 regarding the phased re-opening of the University of Zambia for students in their first and second year of study. Further, the University Senate has also reviewed the phased re-opening for students in the schools of Veterinary Medicine, Health Sciences, Nursing Sciences and students in the School of Public Health.

The resolutions for the decisions of Senate above were as follows:

A. All First Year Students:
ALL First Year Students that reported for face-to-face learning on 8th February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S

https://www.unza.zm/node/1794

Course Structure (3/9)

- **Lecture sessions**
  - Basic introduction to core concepts. Theory + a little math
  - Practical walkthroughs

THE UNIVERSITY OF ZAMBIA
COMMUNICATION AND MARKETING DEPARTMENT
Great East Road Campus| P.O Box 32379| Lusaka, 10101| Tel: +260 211 251 593|
Fax: +260 1 253 952 | Email: comm.marketing@unza.zm | Website: www.unza.zm

PRESS STATEMENT

RE-OPENNING OF THE UNIVERSITY OF ZAMBIA IN A PHASED APPROACH

09th February 2021 - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09th February 2021, has resolved to review the University Senate resolution of 27th January 2021 regarding the phased re-opening of the University of Zambia for students in their first and second year of study. Further, the University Senate has also reviewed the phased re-opening for students in the schools of Veterinary Medicine, Health Sciences, Nursing Sciences and students in the School of Public Health.

The resolutions for the decisions of Senate above were as follows:

A. All First Year Students:
ALL First Year Students that reported for face-to-face learning on 8th February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S

https://www.unza.zm/node/1794

Course Structure (4/9)

- **Paper discussions**
  - Explore problems tackled by other researchers
  - Implicitly look at aspects that will not be explicitly discussed, e.g. ethics and experimentation

THE UNIVERSITY OF ZAMBIA
COMMUNICATION AND MARKETING DEPARTMENT
Great East Road Campus| P.O Box 32379| Lusaka, 10101| Tel: +260 211 251 593|
Fax: +260 1 253 952 | Email: comm.marketing@unza.zm | Website: www.unza.zm

PRESS STATEMENT

RE-OPENNING OF THE UNIVERSITY OF ZAMBIA IN A PHASED APPROACH

09th February 2021 - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09th February 2021, has resolved to review the University Senate resolution of 27th January 2021 regarding the phased re-opening of the University of Zambia for students in their first and second year of study. Further, the University Senate has also reviewed the phased re-opening for students in the schools of Veterinary Medicine, Health Sciences, Nursing Sciences and students in the School of Public Health.

The resolutions for the decisions of Senate above were as follows:

A. All First Year Students:
ALL First Year Students that reported for face-to-face learning on 8th February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S

https://www.unza.zm/node/1794

Course Structure (5/9)

- **Seminars**
  - Academic talks by current and former students
  - Industry talks from entities that employ data mining techniques

THE UNIVERSITY OF ZAMBIA
COMMUNICATION AND MARKETING DEPARTMENT
Great East Road Campus| P.O Box 32379| Lusaka, 10101| Tel: +260 211 251 593|
Fax: +260 1 253 952 | Email: comm.marketing@unza.zm | Website: www.unza.zm

PRESS STATEMENT

RE-OPENNING OF THE UNIVERSITY OF ZAMBIA IN A PHASED APPROACH

09th February 2021 - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09th February 2021, has resolved to review the University Senate resolution of 27th January 2021 regarding the phased re-opening of the University of Zambia for students in their first and second year of study. Further, the University Senate has also reviewed the phased re-opening for students in the schools of Veterinary Medicine, Health Sciences, Nursing Sciences and students in the School of Public Health.

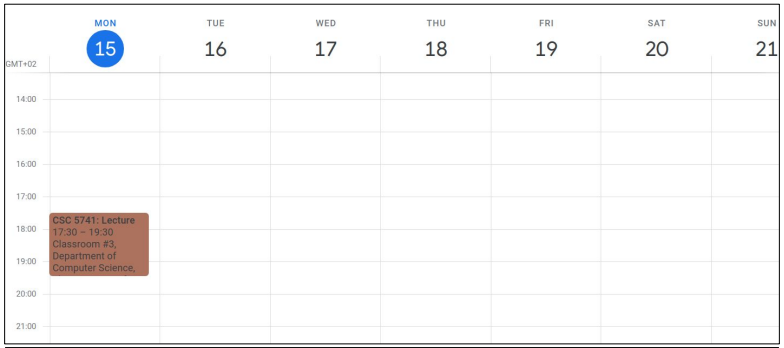
The resolutions for the decisions of Senate above were as follows:

A. All First Year Students:
ALL First Year Students that reported for face-to-face learning on 8th February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S

https://www.unza.zm/node/1794

Course Structure (6/9)

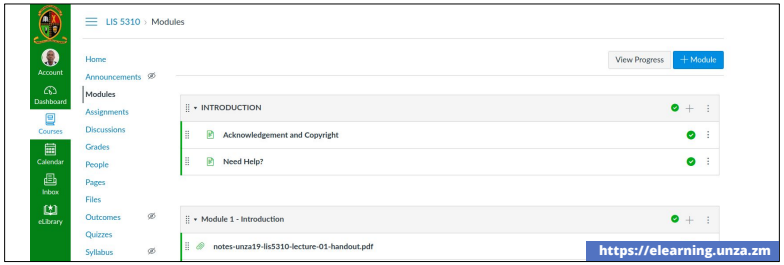


Course Structure (7/9)

- **Course Resources**
  - All course resources will be made available on Astria.

Course Structure (8/9)

- **Course Resources**
  - All course resources will be made available on Astria.

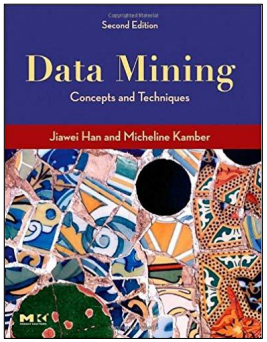


Course Structure (9/9)

- **Additionally, course resources will be disseminated as follows:**
  - Large files such as videos and software tools will be made available via Google Drive and YouTube (recorded sessions)

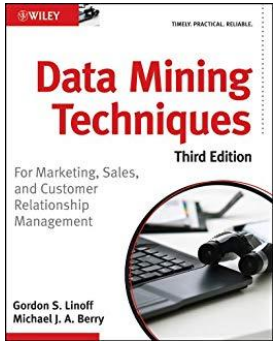
Prescribed & Recommended Textbooks (1/4)

- **Data Mining Concepts and Techniques**
  - J. Han and M. Kamber (2011)



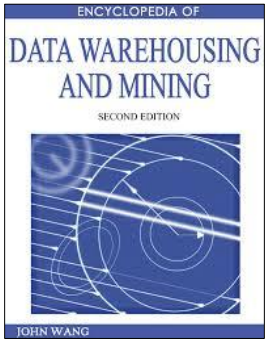
Prescribed & Recommended Textbooks (2/4)

- **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**
  - G. S. Linoff and M. J. Berry (2011)



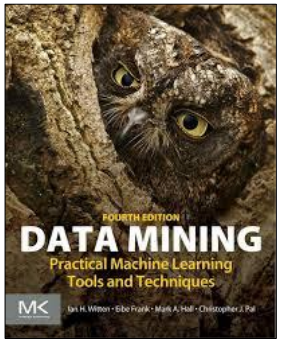
Prescribed & Recommended Textbooks (3/4)

- **Encyclopedia of Data warehousing and Mining**
  - J. Wang (2005)



Prescribed & Recommended Textbooks (4/4)

- **Data Mining: Practical Machine Learning Tools and Techniques**
  - I. H. Witten, E. Frank and M. A. Hall

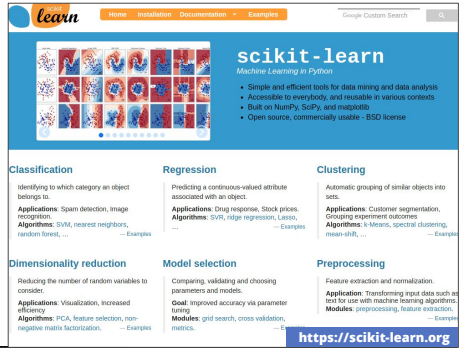


Tools and Services (1/6)

- **Tools and services**
  - VirtualBox for creating virtual environments for running Ubuntu 20.04.
  - Ubuntu 20.04 for running all practical-oriented activities.
  - Python 3
  - Pandas
  - Jupyter Notebook and Google Colab
  - TensorFlow, Keras and Pytorch

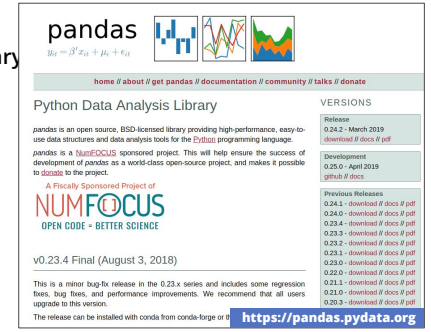
Tools and Services (2/6)

- **scikit-learn**
  - Python machine learning library
  - Implements most of the algorithms we will be exploring



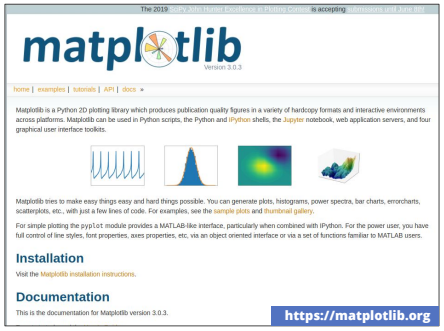
Tools and Services (3/6)

- **Pandas**
  - Python data analysis library



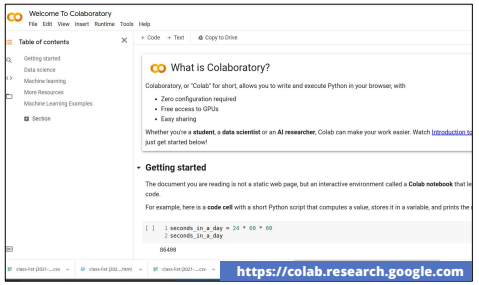
Tools and Services (4/6)

- **Matplotlib**
  - Graphical representation during EDM and analysis



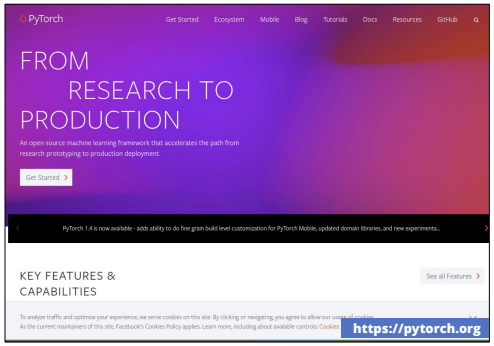
Tools and Services (5/6)

- **Jupyter Notebook and Google Colab**
  - Sharing code used in modules



Tools and Services (6/6)

- **PyTorch**
  - Python machine learning library
  - To be optionally used for deep learning lecture series



Course Grading (1/2)

- **10% Paper readings**
  - Paper summaries of peer-reviewed publications.
- **5% Seminar presentations**
  - Questions and discussions during seminars and reading sessions. Marks awarded for participation.
- **5% Class participation**
  - Discussion in class.
- **20% Practical Projects**
  - Hands-on practical project assignments that will involve a project deliverable.

Course Grading (2/2)

- **20% Class Theory Test**
  - One 90 minutes-long class test will be held towards the end of Term #1
- **40% Final Examination**
  - The final examination is based on the entire course outline.

Course Grading—Paper Readings

#	Mask	Paper #01 (Mgala & Mbogo)	Paper #02 (Caragea et al.)	Paper #03 (Félix et al.)	Paper #04 (Silva and Azevedo)	Pa
1	Elastic Net	68	70	70	90	
2	Linear SVC	70	60	60	75	
3	SGD Classifier	60	45	60	65	
4	Kernal Approximation	50	40	55	65	
5	Lasso	60	0	0	80	
6	Naive Bayes	60	55	60	80	
7	Ensemble Classifiers	49	45	50	75	
8	Spectral Clustering	60	60	65	80	
9	Mean Shift	60	60	80	90	
10	K Neighbors	55	53	65	65	
11	SGD Regressor	75	60	70	70	

http://bit.ly/2Jg6Glk

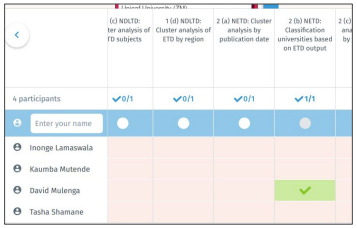
- **The 10% score allocated to the paper readings will be distributed equally amongst the readings**

Course Grading—Seminars

- **The 5% score allocated to the seminars will be distributed amongst all talks**
  - Marks awarded for attendance and participation
- **Invited speakers to be announced soon**

Course Grading—MiniProject (1/3)

- **The 20% score allocated to the Mini Project will be distributed**
  - Implementation of chosen problem
  - Presentation of implementation
  - Technical report based on implementation



Course Grading—MiniProject (2/3)

		Implementation					Technical Report								Presentation						
#	Mask	Data	Code/Scripts	Novelty	Relevance	Demo	Implementation Total	Abstract	Aim/Problem	Implementation	Dataset	Experiment	Results	Quality	Report Total	Content	Quality	Visualisations	Comprehensive	Q/A	Presentaion Total	Grand Total
1	Elastic Net	30	30	15	10	10	95	6	6	8	10	18	20	18	86	18	20	20	18	19	95	18.28
2	Linear SVC					10	10	4	5	5	10	0	15	15	54	10	20	20	10	18	78	8.24
3	SGD Classifier					10	10	4	3	5	6	15	15	15	63	15	18	20	15	18	86	9.28
4	Kernal Approximation					10	10								0						0	0.8
5	Lasso						0								0						0	0
6	Naive Bayes					10	10	6	5	5	10	15	10	10	61	10	20	20	15	20	85	9.08
7	Ensemble Classifiers	30	30	5	10	10	85	4	6	0	5	0	5	5	25	10	10	15	10	15	60	11.2

http://bit.ly/2Jg6Glk

Course Grading—MiniProject (3/3)

- **Wide range of problems and techniques explored**
  - Different problem domains
  - Different ML techniques—classification and clustering

CSC 5741: Mini Practical Projects

#	Student(s)	Project Topic/Title
1.	John Daka	Scholarly Output Classification
2.	Inonge Lamaswala	Advertisements Classification
3.	Mubanga Mubanga	Web Search Classification
4.	Nonde Mukuma	NETD Portal ETD Publication Date Clustering
5.	David Mulenga	NETD Portal Institution Ranking
6.	Memory Mumbi	NETD Portal ETD Subject Clustering
7.	Kaumba Mutende	YouTube Comments Classification
8.	Justin Nongola	Scholarly Output Clustering
9.	Anthony Sampa	YouTube Video Recommender
10.	Tasha Shamane	Blogposts Classification
11.	Mweemba Sikuyuba	Random YouTube Video Classification

http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741

Course Grading—Class Participation

1.	John Daka	Using data mining for bank direct marketing: an application of the CRISP-DM methodology
2.	Inonge Lamaswala	Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment
3.	Mubanga Mubanga	A Novel Position-based Sentiment Classification Algorithm for Facebook Comments
4.	Nonde Mukuma	Speeding up Support Vector Machines
5.	David Mulenga	Mining Educational Data to Analyze Students' Performance
6.	Memory Mumbi	Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics
7.	Kaumba Mutende	Classification of Diabetes patient by using Data Mining Techniques
8.	Justin Nongola	Educational Data Mining & Studentsâ€™ Performance Prediction
9.	Anthony Sampa	TIGER POPULATION GROWTH PREDICTION
10.	Tasha Shamane	A System to Filter Unwanted Messages from OSN User Walls
11.	Mweemba Sikuyuba	Educational Data Mining Rule based

http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741

- **The 5% score allocated to the participation will be distributed equally amongst the talks**

Course Grading—Class Theory Tests

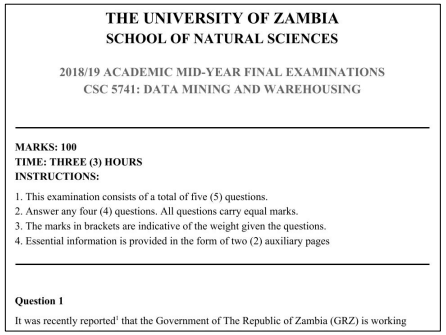
- **The 20% score allocated to class theory tests is distributed equally amongst the tests**
  - Typically two class theory tests

#	Mask	Class Theory Test #1	Class Theory Test #2	Grand Total
1	Elastic Net	62	76	13.8
2	Linear SVC	34	54	8.8
3	SGD Classifier	39	48	8.7
4	Kernal Approximation	6	11	1.7
5	Lasso			0
6	Naive Bayes	26	30	5.6
7	Ensemble Classifiers	9	16	2.5
8	Spectral Clustering	58	64	12.2
9	Mean Shift	56	58	11.4
10	K Neighbors	35	38	7.3
11	SGD Regressor	46	54	10
12	K Means	33	44	7.7
13	MiniBatch K Means	8	20	2.8

http://bit.ly/2Jg6Glk

Course Grading—Final Examination

- **The final examination accounts for 50% of the course weighting**
  - Three hour-long closed examination
  - Content covered in the course



THE UNIVERSITY OF ZAMBIA
SCHOOL OF NATURAL SCIENCES

2018/19 ACADEMIC MID-YEAR FINAL EXAMINATIONS
CSC 5741: DATA MINING AND WAREHOUSING

MARKS: 100
TIME: THREE (3) HOURS
INSTRUCTIONS:

1. This examination consists of a total of five (5) questions.
2. Answer any four (4) questions. All questions carry equal marks.
3. The marks in brackets are indicative of the weight given the questions.
4. Essential information is provided in the form of two (2) auxiliary pages

Question 1

It was recently reported[1] that the Government of The Republic of Zambia (GRZ) is working

Course Grading—CA (1/2)

- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.

#	Mask	Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	Required Exam Score to Pass Course [%]
1	Elastic Net	18.28	13.8	2.5	5	8	47.58	79	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	81
5	Lasso	0	0	1	0	5	6	10	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	59	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	48
11	SGD Regressor	16.52	10	3.5	5	8	43.02	72	17
12	K Means	16	7.7	3	2.5	6	35.2	59	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	102

http://bit.ly/2Jg6Glk

Course Grading—CA (2/2)

- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.

#	Mask	Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	Required Exam Score to Pass Course [%]
1	Elastic Net	18.28	13.8	2.5	5	8	47.58	79	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	81
5	Lasso	0	0	1	0	5	6	10	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	59	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	48
11	SGD Regressor	16.52	10	3.5	5	8	43.02	72	17
12	K Means	16	7.7	3	2.5	6	35.2	59	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	102

http://bit.ly/2Jg6Glk

Course Grading—CA (2/2)

- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.

#	Mask	Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	Required Exam Score to Pass Course [%]
1	Elastic Net	18.28	13.8	2.5	5	8	47.58	79	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	81
5	Lasso	0	0	1	0	5	6	10	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	59	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	48
11	SGD Regressor	16.52	10	3.5	5	8	43.02	72	17
12	K Means	16	7.7	3	2.5	6	35.2	59	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	102

http://bit.ly/2Jg6Glk

Course Grading—CA (2/2)

- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.

#	Mask	Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	Required Exam Score to Pass Course [%]
1	Elastic Net	18.28	13.8	2.5	5	8	47.58	79	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	81
5	Lasso	0	0	1	0	5	6	10	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	59	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	48
11	SGD Regressor	16.52	10	3.5	5	8	43.02	72	17
12	K Means	16	7.7	3	2.5	6	35.2	59	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	102

http://bit.ly/2Jg6Glk

Course Grading Thresholds

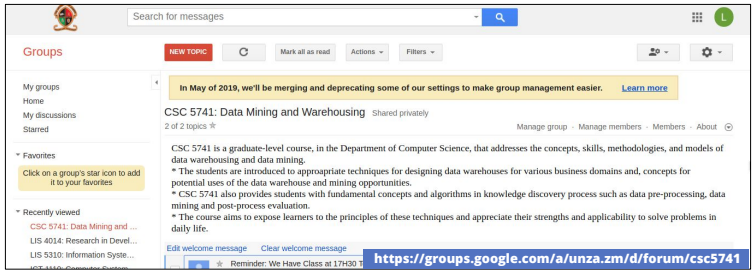
GRADE	DESCRIPTION	SCORE RANGE	GRADE POINT
A+	DISTINCTION	86–100	5
A	DISTINCTION	75–85	4
B+	MERITORIOUS	70–74	3.5
B	CREDIT	65–69	3
C+	CREDIT	55–64	2.37
C	PASS	50–54	1.5
D	FAIL	<49	0

Course Management (1/2)

- **Instructor: Lighton Phiri and TBA (possible Invited Talks)**
- **Email: lighton.phiri@unza.zm**
- **Office: Room 515, Fifth Floor, School of Education Building**
- **Office hours: Friday 09H00–13H00**
  - Alternatively, schedule an appointment via email after checking free/busy slots on my calendar (https://goo.gl/6kHrnA)

Course Management (2/2)

- **Communication exclusively done electronically**
  - The Moodle, Course Mailing List and Email



Academic Dishonesty

- **Every assessment submitted must be your own work. Academic dishonesty of any form is considered very seriously.**
  - NOTE: Any form of academic dishonesty (plagiarism, copying, cheating etc) will result in a ZERO mark for the entire continuous assessment score.

Q & A Session

- **Comments, concerns and complaints?**

Lecture Series Outline

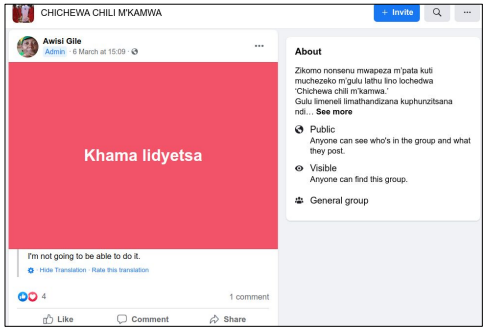
- **Part I: Administrivia**
- **Part II: Course Introduction**
  - Contextualising Data Mining and Warehousing
  - CSC 5741 Themes and Topics
- **Part III: How to Read a Paper**
- **Part IV: On Academic Activities**
- **Part V: About Next Week**

Contextualising Data Mining & Warehousing: Everyday Examples (1/5)

- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**

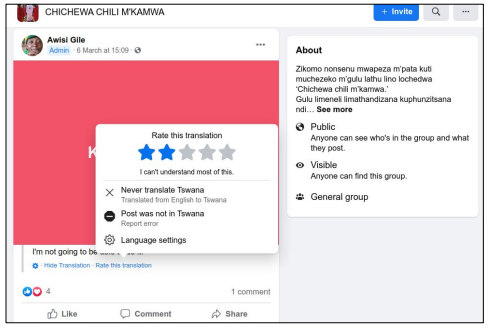
Contextualising Data Mining & Warehousing: Everyday Examples (2/5)

- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**



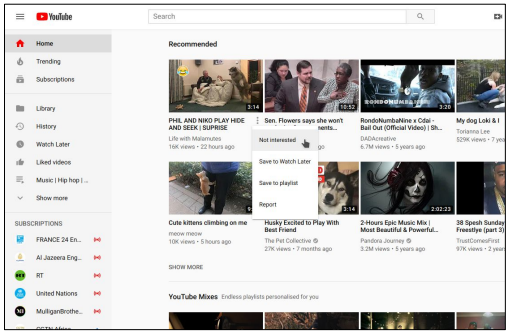
Contextualising Data Mining & Warehousing: Everyday Examples (2/5)

- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**



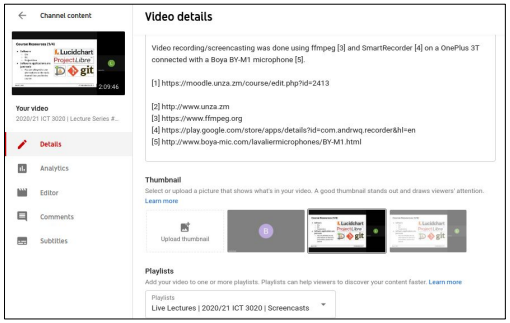
Contextualising Data Mining & Warehousing: Everyday Examples (3/5)

- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**



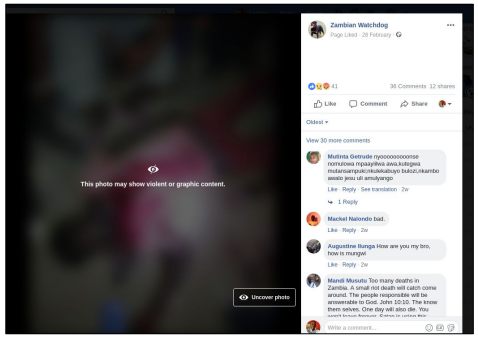
Contextualising Data Mining & Warehousing: Everyday Examples (4/5)

- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**



Contextualising Data Mining & Warehousing: Everyday Examples (5/5)

- **Effective ways are needed to automatically make sense out of digital content.**
  - Relevance
  - Recommendation
  - Restricted and obscene materials

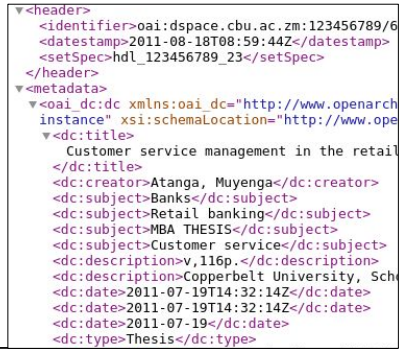


Contextualising Data Mining & Warehousing: Postgraduate Projects

- **Past CS@ UNZA Dissertations**
  - Lillian Muzyece (2019). Automatic Weather Prediction
  - Soft Mulizwa (2019). Automatic Customer Segmentation for effective Targeted Campaigns
  - Friday Chazanga (2019). Automatic Number Plate Recognition
  - Francis Chulu (2020). Automatic identification and early warning and monitoring web based system of fall Armyworm
  - Knox Kamusweke (2020). Data Mining for Fraud Detection
- **Current CS@ UNZA Dissertations**
  - Simon Hawatichke Chiwamba (2019—). Machine Learning Automated Image Capture and Identification of Fall Armyworm

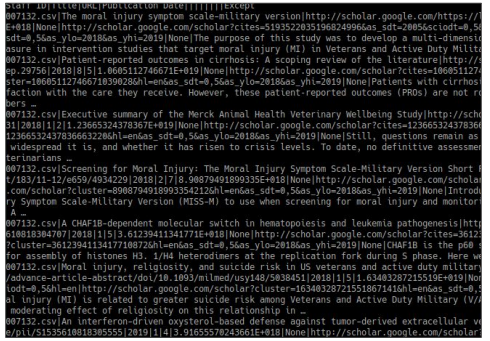
Contextualising Data Mining & Warehousing: Some Ongoing Projects (1/8)

- **Automatic classification of scholarly research**
  - Automatic generation of metadata
  - Automatic reclassification of digital objects
  - Project #1: Automatic Classification of ETDs
  - Project #2: Automatic Classification of IR objects



Contextualising Data Mining & Warehousing: Some Ongoing Projects (2/8)

- **University of Zambia Ranking Committee Research Report**
  - Mining for scholarly output on the Web



Contextualising Data Mining & Warehousing: Some Ongoing Projects (3/8)

- **LMS Log Mining**
  - Moodle usage logs
  - Project #1: Predicting students at-risk of performing poorly

Contextualising Data Mining & Warehousing: Some Ongoing Projects (4/8)

- **Mwabu Tablet Usage Analysis**
  - Android app usage and interaction logs
  - Interaction patterns for learners and educators

Contextualising Data Mining & Warehousing: Some Ongoing Projects (5/8)

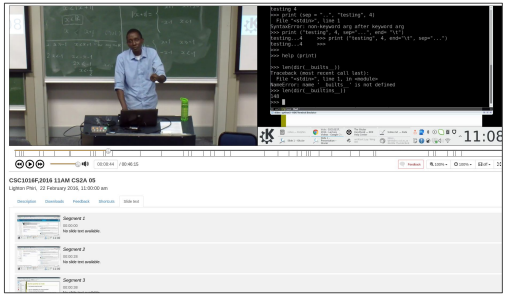
- **Effectiveness of FISP Programme Using 'Tripple Effect' Method**
  - Collaboration with two economists

```
> colnames(dataset_cfs0405_crop)
 [1] "PROV"     "DIST"     "CONST"    "WARD"     "REGION"   "CSA"
 [7] "SEA"      "HHNUM"    "CROP"     "ID009"    "S1AFIELD" "S1ACF01"
[13] "S1ACF02"  "S1ACF03"  "S1ACF04"  "S1ACF05"  "S1ACF06"  "S1ACF07"
[19] "S1ACF08"  "S1ACF09"  "S1ACF10"  "S1ACF11"  "S1ACF12"  "S1ACF13"
[25] "S1ACF14"  "S1ACF15"  "S1ACF16"  "TOTHARV"  "WEIGHT"   "HA_HARV"
[31] "convert"  "HA_PLANT"
> head(dataset_cfs0405_crop)
    PROV    DIST CONST WARD REGION CSA SEA HHNUM                CROP ID009
1 Central Chibombo    1    1    1 14  1   55               Maize    1
2 Central Chibombo    1    3    1  2  2   77               Maize    1
3 Central Chibombo    1    3    1  2  2   33 Other crops (specify)    2
4 Central Chibombo    2   12    1  2  3   96               Maize    3
5 Central Chibombo    2   12    1  2  3   96          Groundnuts    3


> colnames(dataset_cfs2004_2005_weight)
[1] "ID001"  "ID002"  "ID003"  "ID004"  "ID005"  "ID006"  "ID007"  "ID009"
[9] "WEIGHT"
> head(dataset_cfs2004_2005_weight)
    ID001   ID002 ID003 ID004 ID005 ID006 ID007 ID009    WEIGHT
1 Central Chibombo    1    1    1   14    1    1 388.84409
```


Contextualising Data Mining & Warehousing: Some Ongoing Projects (6/8)

- **Open Matterhorn Video Segmentation Analysis**
  - Seeking to points of interest



Contextualising Data Mining & Warehousing: Some Ongoing Projects (7/8)

- **Automatic Content Generation**
  - Underrepresentation on platforms like Wikipedia
  - We have VERY few Textbooks!!!

Contextualising Data Mining & Warehousing: Some Ongoing Projects (8/8)

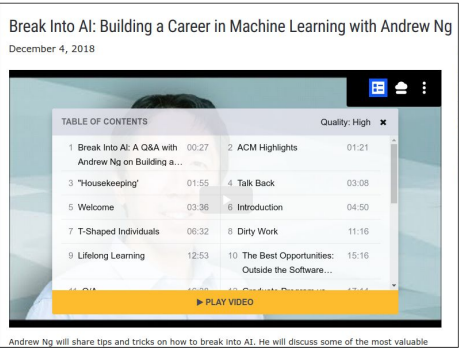
- **Working with Radiologists at UTHs. Requirements**
  - Software and hardware designers
  - Private entities and entrepreneurs to develop cost effective tools and provide local solutions
  - Govt's WAN
  - Political will

Contextualising Data Mining & Warehousing: Zambia Centric Projects

- There is more out there [...]
  - Parliament TV? Video and audio analysis
  - Tollgates! Automatic detection of vehicles
  - Automatic prediction of learning outcomes
  - [...]
  - [...]
  - Sentiment analysis: Popular Zambian Facebook pages, Twitter
  - Opinion mining from social media
  - What are people discussing on platforms like WhatsApp?
  - What if we harvested articles written in mainstream newspaper articles

Contextualising Data Mining & Warehousing: Endless Possibilities

- At the rate data is being generated, we will have an endless list of data mining problems to work on.
  - What problems to work on?
  - [...]
  - [...]



Contextualising Data Mining & Warehousing: Curiosity vs Impact (1/6)

- Curiosity-driven research
  - Puzzles
  - Games

Contextualising Data Mining & Warehousing: Curiosity vs Impact (2/6)

- Impact-driven research/studies
  - Education
  - Health
  - So-called ICT for development perhaps?

Contextualising Data Mining & Warehousing: Curiosity vs Impact (3/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps?

OVER 200 trucks transporting various goods are marooned at Kipushi border following an impasse between clearing agents and authorities in the Democratic Republic of Congo. PICTURE: BUTTYSON KANDIMBA

ZRA drones land on 7 trucks

From page 1

The impounded trucks were spotted hidden in the bush during a test flight by ZRA "pilots", a statement issued yesterday by ZRA corporate

has also confiscated 26 heavy-duty trucks and earth-moving machinery that were suspiciously imported through the misapplication of the value-added tax (VAT) deferment scheme.

misclassifications, under-declarations, and under-valuations is a serious offence and offenders will be dealt with in accordance with the law," he said.

in 26 trucks being impounded and will continue till all the trucks that have not paid the taxes pay what is due to the government," Mr Sikalinda said.

Zambia Daily Mail | August 18, 2019 | Volume 22 No. 033

Contextualising Data Mining & Warehousing: Curiosity vs Impact (4/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps

Contextualising Data Mining & Warehousing: Curiosity vs Impact (4/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps

Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps?

Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps?

Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

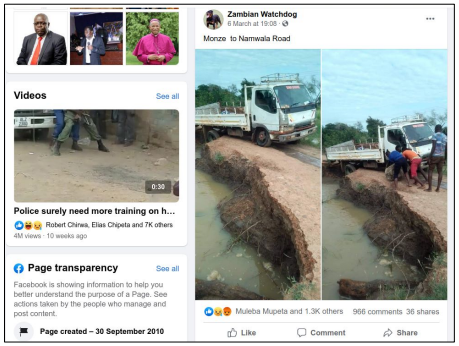
- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perh

Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perh

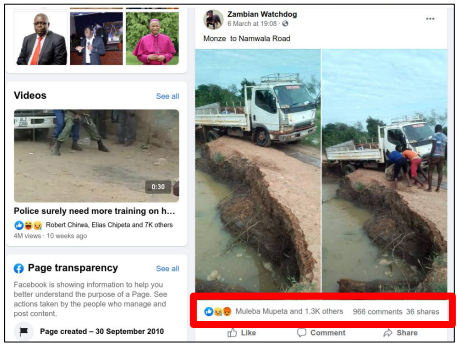
Contextualising Data Mining & Warehousing: Curiosity vs Impact (6/6)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps?

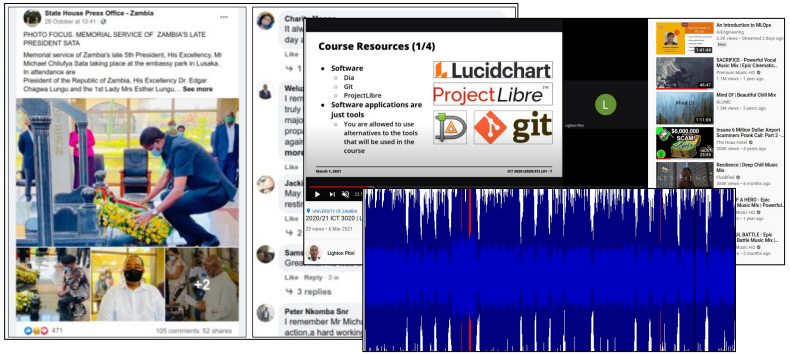


Contextualising Data Mining & Warehousing: Curiosity vs Impact (6/6)

- Impact-driven research/studies
  - Education
  - Health
  - So-called ICT for development perhaps?



Data is Key to Data Mining



Introduction (1/2)

- Identify the key processes of data mining, data warehousing and knowledge discovery process
- Describe the basic principles and algorithms used in practical data mining and understand their strengths and weaknesses
- Apply data mining techniques to solve problems in other disciplines in a mathematical way
- Apply data mining methodologies with information systems and generate results which can be immediately used for decision making in well-defined business problems

Introduction (2/2)

- Data Mining and Data Pre-processing
- Data Warehousing
- Classification
- Associative Rule Mining
- Clustering Analysis

Theme #1: Data Mining and Data Pre-processing

- Data Mining vs. Statistics
- Knowledge discovery process
- Machine learning
- Pattern recognition
- Data cleaning
- Data integration
- Data selection
- Data transformation
- Pattern evaluation
- Knowledge presentation

Theme #2: Data Warehousing

- Decision support system
- Data warehouse architecture
- Online transaction processing
- Online analytical processing
- Star schema, Snowflake schema
- Fact constellation
- Dimension Tables and Fact tables
- Data Granularity
- Data cube
- Pivot, slice and dice, roll-up and drill down

Theme #3: Classification

- Decision Tree; Hunt's Algorithm; C4.5; Tree Induction; Binary split and Multi-way
- split; Measures of Impurity: Gini Index, Entropy and Misclassification error; Rule-
- Based Classifier; Coverage and Accuracy; Mutually exclusive and exhaustive rules;
- Ripper; Rule Pruning; Instance-Based Classifiers; Nearest neighbour classification;
- Probabilistic classifier; Naïve Bayes classifier.

Theme #4: Associative Rule Mining

- Rule Evaluation Metrics: Support and confidence
- Frequent Itemsets, Maximal
- Frequent Itemset, Closed Frequent Itemsets
- Brute-force approach
- Apriori principle
- Frequent-Pattern Tree
- Prefix paths, Conditional FP-Tree
- Rule Generation

Theme #5: Clustering Analysis

- **Intra-cluster distances, Inter-cluster distances**
- **Partitional clustering**
- **K-means**
- **Centroid; Sum of Squared Error**
- **Hierarchical clustering**
- **Agglomerative and divisive**
- **Dendrogram**
- **Single linkage, complete linkage and group average**
- **Ward's Method.**

Closing CSC 5710 Remarks

- **Beyond CSC 5741**
  - Research focus
  - Vision 2030
- **About assessments**
  - Ensure all assessments are attempted
- **Academic dishonesty**
  - NOTE: Any form of academic dishonesty (plagiarism, copying, cheating etc) will result in a ZERO mark for the entire continuous assessment score.

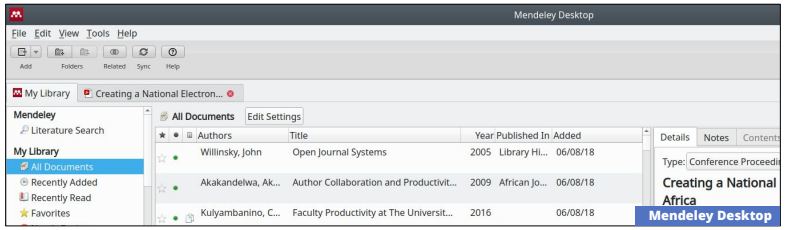
Q & A Session

- **Comments, concerns and complaints?**

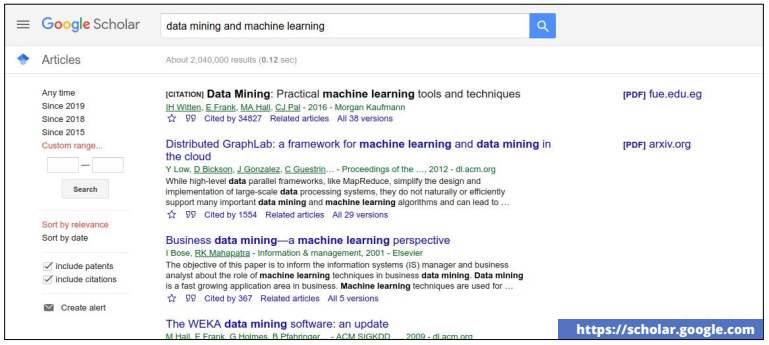
Lecture Series Outline

- **Part I: Administrivia**
- **Part II: Course Introduction**
- **Part III: How to Read a Paper**
  - Bibliographic Management Software
  - Reputable Publication Venues
  - How to Read a Paper: Keshav's Three-Pass Approach
- **Part IV: On Academic Activities**
- **Part V: About Next Week**

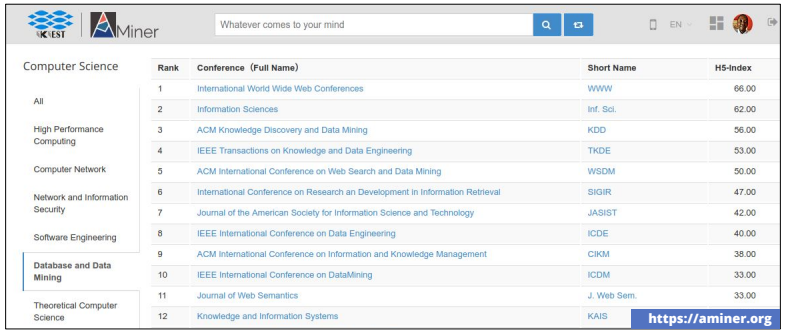
Readings and Paper Summaries (1/5)



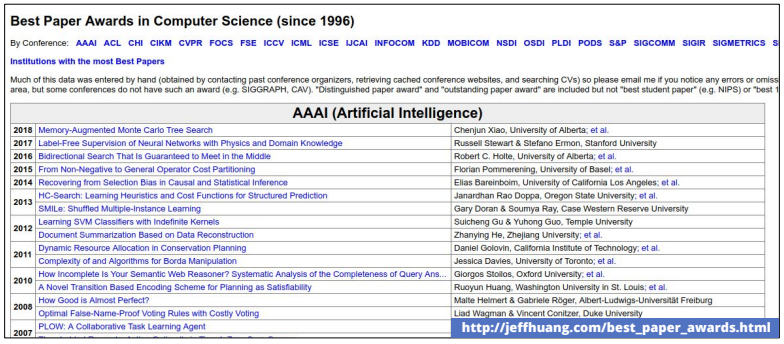
Readings and Paper Summaries (2/5)



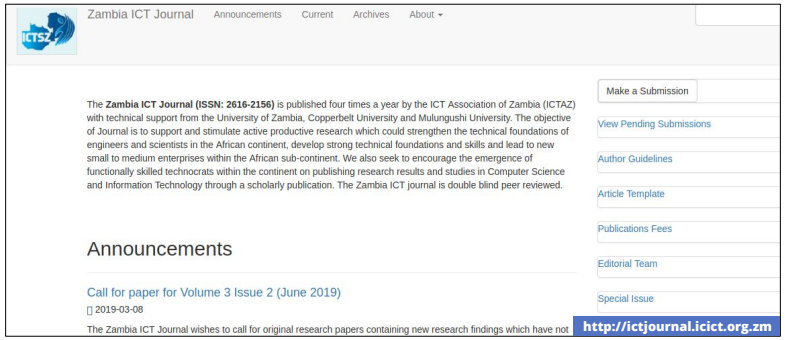
Readings and Paper Summaries (3/5)



Readings and Paper Summaries (4/5)



Readings and Paper Summaries (5/5)



On How to Read a Paper (1/5)

On How to Read a Paper (2/5)

- Title
- Abstract
- Introduction
- Related Work
- Implementation
- Evaluation
- Discussion
- Conclusion
- References

On How to Read a Paper (3/5)

- Keshav's Three Pass Approach is very helpful when initially getting started.
  - Pass #1
    - Title -> Abstract -> Introduction
    - Sections and subsections -> Conclusion -> References
    - Outcome of pass: paper classification, context, correctness, contributions, clarity
  - Pass #2
  - Pass #3

On How to Read a Paper (4/5)

- **Keshav's Three Pass Approach is very helpful when initially getting started.**
  - Pass #1
  - Pass #2
    - Analyse floats
    - Note key references not read
    - Outcome: Firm understanding of paper
  - Pass #3

On How to Read a Paper (5/5)

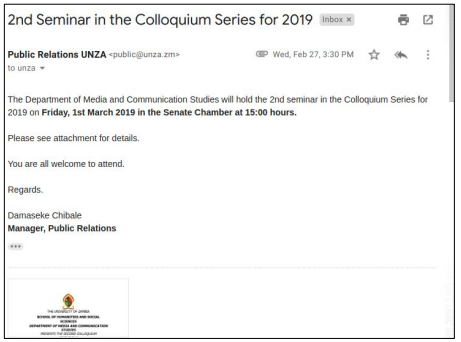
- **Keshav's Three Pass Approach is very helpful when initially getting started.**
  - Pass #1
  - Pass #2
  - Pass #3
    - Outcome: Identify potential flaws with experimental designs and analyses.

Lecture Series Outline

- **Part I: Administrivia**
- **Part II: Course Introduction**
- **Part III: How to Read a Paper**
- **Part IV: On Academic Activities**
  - Public Talks
  - Public Oral Examinations
  - DRGS Organised Events
- **Part IV: About Next Week**

Public Talks

- **Make time to attend public academic talks irrespective of whether it is computing related**
  - Inspiration for potential topics next year
  - Potential collaboration


2nd Seminar in the Colloquium Series for 2019 Inbox ×

Public Relations UNZA <public@unza.zm> Wed, Feb 27, 3:30 PM

to unza

The Department of Media and Communication Studies will hold the 2nd seminar in the Colloquium Series for 2019 on **Friday, 1st March 2019 in the Senate Chamber at 15:00 hours.**

Please see attachment for details.

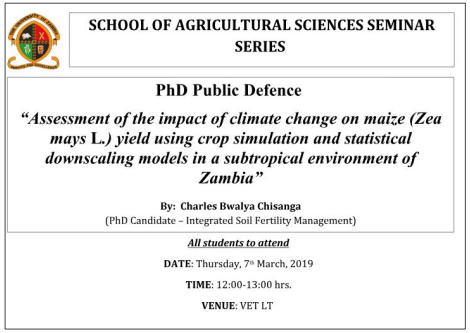
You are all welcome to attend.

Regards.

Damaseke Chibale
Manager, Public Relations

Public Oral Examinations

- Make time to attend public oral examinations so you have an idea what to expect.



SCHOOL OF AGRICULTURAL SCIENCES SEMINAR SERIES

PhD Public Defence

"Assessment of the impact of climate change on maize (Zea mays L.) yield using crop simulation and statistical downscaling models in a subtropical environment of Zambia"

By: Charles Bwalya Chisanga
(PhD Candidate – Integrated Soil Fertility Management)

All students to attend

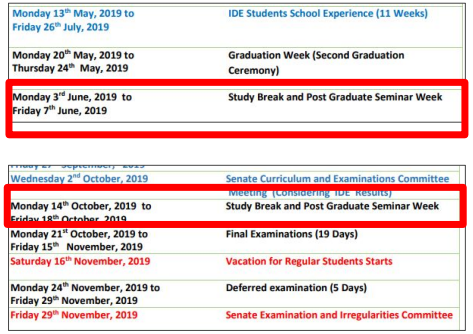
DATE: Thursday, 7th March, 2019

TIME: 12:00-13:00 hrs.

VENUE: VET LT

DRGS Organised Events

- You want to attend important postgraduate events in order to gain a sense of what is expected
  - Announcements are sent through to your official UNZA-assigned email addresses.



Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week
Wednesday 2nd October, 2019	Senate Curriculum and Examinations Committee Meeting (Considering IDE Results)
Monday 14th October, 2019 to Friday 18th October, 2019	Study Break and Post Graduate Seminar Week
Monday 21st October, 2019 to Friday 15th November, 2019	Final Examinations (19 Days)
Saturday 16th November, 2019	Vacation for Regular Students Starts
Monday 24th November, 2019 to Friday 29th November, 2019	Deferred examination (5 Days)
Friday 29th November, 2019	Senate Examination and Irregularities Committee

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: How to Read a Paper
- Part IV: On Academic Activities
- Part V: About Next Week
  - Getting Started: Jupyter Notebook, scikit-learn, pandas
  - Paper Reading List [Trial]
  - Academic Talk: L. Phiri [Trial]

Getting Started with Python, SciKit-learn & Pandas

- Tools installation and configuration
- Common commands
- SciKit-learn
- Pandas
- Sample datasets

Paper Reading List [Trial]

[1] S. Keshav (2007) "How to Read a Research Paper" https://doi.org/10.1145/1273445.1273458

[2] P. W. L. Fong (2004) "How to Read a CS Research Paper?" https://doi.org/10.1145/1595453.1595493

[3] L. Phiri (2018) "Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories" https://doi.org/10.1504/IJMSO.2020.112804

Academic Talk: L. Phiri [Trial]

[1] L. Phiri (2020) "Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories" https://doi.org/10.1504/IJMSO.2020.112804

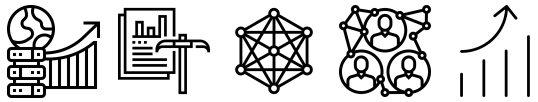
Q & A Session

- Comments, concerns and complaints?

Bibliography

[1] 2020/21 CSC 5741 Syllabus http://bit.ly/30Pdm85

csc5741@unza.zm
http://bit.ly/39HTdTK
http://bit.ly/2kK2ZkA



CSC 5741 (2020/21)
Data Mining and Warehousing
Lecture 1: Administrivia, Course Overview and Introduction

Lighton Phiri
Department of Library & Information Science
University of Zambia
http://lis.unza.zm/~lightonphiri