

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

March 15, 2021

CSC 5741 (2020/21) L01 - 2

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)

Education

[About the School](#)
[How To Apply](#)
[Message From The Dean](#)
[Departments](#)
[Undergraduates](#)
[Postgraduates](#)
[Projects and Publications](#)
[Staff](#)

People

Listing of School of education department of library information science staff by Name and Position

	Chiriongo Hamoya Head of Department, Lecturer, Researcher		Akakanetwa Akakanetwa Senior Lecturer, Researcher		Abel Mwilaama Lecturer, Researcher
	Besson Njibvu Lecturer, Researcher		Edward Mwalamu Lecturer, Researcher		Francisca Mulonda Lecturer, Researcher
	Lighton Phiri Lecturer, Researcher		Phyela S. Mbewe Lecturer, Researcher		Thabiso Mvelenga Lecturer, Researcher

Dean, School of Education

Berry Nhata (PhD)
School of Education
Great East Road Campus
PO Box 33779
Loaka

Contacts

Tel: +26 021 221 1581
Fax: +26 021 221 1581

<https://www.unza.zm/people/school-of-education/department-of-library-information-science>

March 15, 2021

CSC 5741 (2020/21) L01 - 4

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 emphasize the application of Machine Learning.

Before joining UNZA, I was a PhD student at the University of Cape Town (UCT).

I proposed [grid-based technology-driven learning](#) (Thesis).

I was affiliated with the Digital Libraries Laboratory, the Centre in ICT for Development and the HPLC4A 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 [Artel](#).

Earlier, I studied Software Engineering at UNZA ([meecu](#)).

Research: My current research focus, along with my broad interests, include strategies for increased visibility of scholarly research output, automatic citation generation, open data for us to get 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>

March 15, 2021

CSC 5741 (2020/21) L01 - 5

Personal Introductions (3/5)

Lighton Phiri

Department of Computer Science, University of Cape Town
Rondebosch 7701, Cape Town, South Africa
Email: X@Y (X=lpfiri, 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 Suleiman](#) and [Christoph Meinel](#).

I was affiliated with Digital Libraries Laboratory, Centre for ICT for Development and HPLC4A.

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

Additional Information

Teaching: CSC101SF: Computer Science 101S – [2016] [[notes](#)] [[code](#)] [[slides](#)] [[videos](#)]
CSC101F: Python for Engineers – [2015] [[notes](#)] [[code](#)] [[slides](#)] [[videos](#)]

<https://people.cs.uct.ac.za/~lpfiri>

March 15, 2021

CSC 5741 (2020/21) L01 - 6

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.



Digital Libraries

The field of Digital Libraries (DLs) generally involves the study of digital collections of information and corresponding network-based systems used to manage data from these 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.



DataLab People

Academic Staff



Lighton Phiri

Masters Student

MSc Computer Science

Masters Student

2021



Adrian Chikale

MA Library and Information

Science

2020



Adeline Kassonde

MA Library and Information

Science

2019



Matilda Muchinga

MA Library and Information

Science

2019

Vilas Chama

MA Library and Information

Science

<http://datalab.unza.zm>

March 15, 2021

CSC 5741 (2020/21) L01 - 7

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

March 15, 2021

CSC 5741 (2020/21) L01 - 8

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

March 15, 2021

CSC 5741 (2020/21) L01 - 9

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

March 15, 2021

CSC 5741 (2020/21) L01 - 10

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 11

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 32279 | Lusaka, 10101 | Tel: +260 211 381 893
Fax: +260 1 253 952 | Email: comm.marketing@unza.zm | Website: www.unza.zm

PRESS STATEMENT

RE-OPENING 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 decided to review the University's decision 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 Students: <https://www.unza.zm/node/1794>

March 15, 2021

CSC 5741 (2020/21) L01 - 12

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-OPENING 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: <https://www.unza.zm/node/1794>

March 15, 2021 CSC 5741 (2020/21) L01 - 13

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-OPENING 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: <https://www.unza.zm/node/1794>

March 15, 2021 CSC 5741 (2020/21) L01 - 14

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-OPENING 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: <https://www.unza.zm/node/1794>

March 15, 2021 CSC 5741 (2020/21) L01 - 15

Course Structure (6/9)

MON	TUE	WED	THU	FRI	SAT	SUN
15	16	17	18	19	20	21
14:00						
15:00						
16:00						
17:00						
18:00	CSC 5741: Lecture 17:30 - 19:30 Classroom #3, Department of Computer Science.	17				
19:00						
20:00						
21:00						

March 15, 2021 CSC 5741 (2020/21) L01 - 16

Course Structure (7/9)

- Course Resources

- All course resources will be made available on Astria.

March 15, 2021

CSC 5741 (2020/21) L01 - 17

<https://elearning.unza.zm>

Course Structure (8/9)

- Course Resources

- All course resources will be made available on Astria.

March 15, 2021

CSC 5741 (2020/21) L01 - 18

<https://elearning.unza.zm>

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)

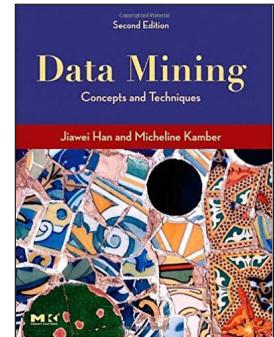
March 15, 2021

CSC 5741 (2020/21) L01 - 19

Prescribed & Recommended Textbooks (1/4)

- Data Mining Concepts and Techniques

- J. Han and M. Kamber (2011)

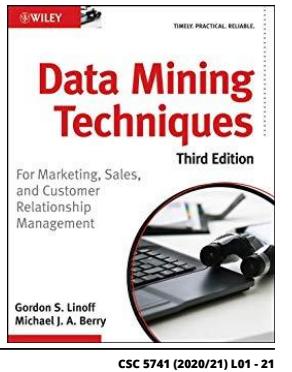


March 15, 2021

CSC 5741 (2020/21) L01 - 20

Prescribed & Recommended Textbooks (2/4)

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

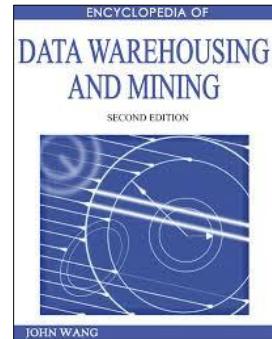


March 15, 2021

CSC 5741 (2020/21) L01 - 21

Prescribed & Recommended Textbooks (3/4)

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

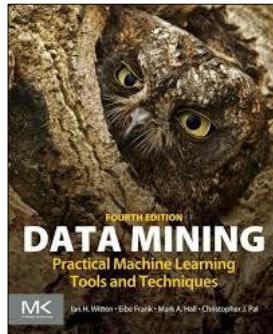


March 15, 2021

CSC 5741 (2020/21) L01 - 22

Prescribed & Recommended Textbooks (4/4)

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



March 15, 2021

CSC 5741 (2020/21) L01 - 23

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



March 15, 2021

CSC 5741 (2020/21) L01 - 24

Tools and Services (2/6)

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

The screenshot shows the scikit-learn homepage. It features a main banner with the text "Machine Learning in Python" and several sub-sections: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing. Each section includes a brief description, applications, and algorithms. At the bottom is a navigation bar with links like Home, Installation, Documentation, Examples, and a search bar.

<https://scikit-learn.org>

March 15, 2021

CSC 5741 (2020/21) L01 - 25

Tools and Services (3/6)

- **Pandas**
 - Python data analysis library

The screenshot shows the pandas website. It has a header with the pandas logo and a search bar. Below the header are sections for Versions, Release Notes, and Development. The main content area features a banner for NumFOCUS, followed by a release note for v0.23.4 Final (August 3, 2018). The page also includes a link to the GitHub repository and a download section.

<https://pandas.pydata.org>

March 15, 2021

CSC 5741 (2020/21) L01 - 26

Tools and Services (4/6)

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

The screenshot shows the Matplotlib website. It features a large "matplotlib" logo at the top, followed by a "Version 3.0.3" header. Below this are sections for Home, Examples, Tutorials, API, and Docs. The main content area displays several sample plots including a line plot, a histogram, a heatmap, and a scatter plot. A "Documentation" section at the bottom provides links to the Matplotlib version 3.0.3 documentation.

<https://matplotlib.org>

March 15, 2021

CSC 5741 (2020/21) L01 - 27

Tools and Services (5/6)

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

The screenshot shows the Google Colab interface. It features a top navigation bar with File, Edit, View, Insert, Runtime, Tools, and Help. Below this is a "Welcome To Colaboratory" message and a "Table of contents" sidebar. The main area contains a Jupyter Notebook cell with code, output, and a "What is Colaboratory?" sidebar. At the bottom is a code editor and a "https://colab.research.google.com" URL.

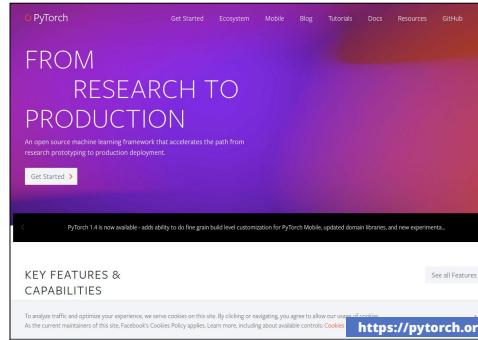
<https://colab.research.google.com>

March 15, 2021

CSC 5741 (2020/21) L01 - 28

Tools and Services (6/6)

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



March 15, 2021

CSC 5741 (2020/21) L01 - 29

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 30

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 31

Course Grading—Paper Readings

1	#	Mask	Paper #01 (Mgala & Mbogo)	Paper #02 (Caragea et al.)	Paper #03 (Félix et al.)	Paper #04 (Silva and Azevedo) Pa
2	1	Elastic Net	68	70	70	90
3	2	Linear SVC	70	60	60	75
4	3	SGD Classifier	60	45	60	65
5	4	Kernal Approximation	50	40	55	65
6	5	Lasso	60	0	0	80
7	6	Naive Bayes	60	55	60	80
8	7	Ensemble Classifiers	49	45	50	75
9	8	Spectral Clustering	60	60	65	80
10	9	Mean Shift	60	60	80	90
11	10	K Neighbors	55	53	65	65
12	11	SGD Regressor	75	60	70	70

<http://bit.ly/2Jg6Gik>

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

March 15, 2021

CSC 5741 (2020/21) L01 - 32

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

CSC 5741 Invited Talks Slots

by Lighton Phiri 4 days ago · Print

University of Zambia

All times displayed in Africa/Lusaka

Table Calendar

Apr 2 TUE	Apr 9 TUE	Apr 16 TUE	Apr 23 TUE	Apr 30 TUE	May 7 TUE	May 14 TUE	
5:30 PM 6:30 PM	5:30 PM 6:30 PM	5:30 PM 6:30 PM	5:30 PM 6:30 PM	5:30 PM 6:30 PM	5:30 PM 6:30 PM	5:30 PM 6:30 PM	
5 participants	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1 ✓/1	
Enter your name							
Lillian Myece Andrea Kumwenda Friday C. Chazanga Soft Mulenga Francis Chulu							

CSC 5741 (2020/21) L01 - 33

March 15, 2021

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

	(G) NDTB: Cluster analysis of ETD subjects	1 (D) NDTB: Cluster analysis of ETD by region	2 (A) NDTB: Cluster analysis of publication date	2 (B) NDTB: Classification of universities based on ETD output	2 (C) NDTB: Classification of universities based on ETD output
4 participants	✓/0/1	✓/0/1	✓/0/1	✓/0/1	✓/1/1
Enter your name	● ● ● ●				
Inonge Lamaswala Kaumba Mutende David Mulenga Tasha Shamane					✓

March 15, 2021

CSC 5741 (2020/21) L01 - 34

Course Grading—MiniProject (2/3)

# Mask	Data	Implementation			Technical Report			Presentation			Report Total	Content	Quality	Visualisations	Comprehensive	Q/A	Presentational Total	Grand Total	
		Code/Scripts	Novelty	Relevance	Demo	Abstract	Aim/Problem	Implementation	Dataset	Experiment									
1 Elastic Net	30	30	15	10	10	95	6	6	8	10	18	20	18	18	20	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 Kernel 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	15 60 11.2	
8 Convex Clustering	30	30	10	10	10	90	0	4	10	10	17	17	15	92	10	20	20	0 0	

<http://bit.ly/zjg66lk>

March 15, 2021

CSC 5741 (2020/21) L01 - 35

Course Grading—MiniProject (3/3)

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

#	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://iis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741>

March 15, 2021

CSC 5741 (2020/21) L01 - 36

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. Nondi Mukuma	Speeding up Support Vector Machines
5. David Mulenga	Mining Educational Data to Analyze Students' Performance
6. Memory Munibi	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 & Students' 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://iis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741

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

March 15, 2021

CSC 5741 (2020/21) L01 - 37

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/2Jg6Gik>

March 15, 2021

CSC 5741 (2020/21) L01 - 38

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 to 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

March 15, 2021

CSC 5741 (2020/21) L01 - 39

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	6	47.58	70	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.64	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/2Jg6Gik>

March 15, 2021

CSC 5741 (2020/21) L01 - 40

Course Grading—CA (2/2)

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

#	Mask	E	F	G	H	I	J	K	L	M	Required Exam Score to Pass Course (%)
		Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	CA %	Required Exam Score to Pass Course (%)	
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	6	41
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	7	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	7	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	6	81
5	Lasso	0	0	1	0	5	6	10	6	6	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	7	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	6	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	7	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	8	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	6	48
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	6	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	6	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	5	102

<http://bit.ly/2Jg6GIk>

March 15, 2021

CSC 5741 (2020/21) L01 - 41

Course Grading—CA (2/2)

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

#	Mask	E	F	G	H	I	J	K	L	M	Required Exam Score to Pass Course (%)
		Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	CA %	Required Exam Score to Pass Course (%)	
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	6	41
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	7	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	7	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	6	81
5	Lasso	0	0	1	0	5	6	10	6	6	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	7	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	6	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	7	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	8	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	6	48
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	6	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	6	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	5	102

<http://bit.ly/2Jg6GIk>

March 15, 2021

CSC 5741 (2020/21) L01 - 42

Course Grading—CA (2/2)

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

#	Mask	E	F	G	H	I	J	K	L	M	Required Exam Score to Pass Course (%)
		Mini Project [20%]	Theory Tests [20%]	Seminar Presentations [5%]	Class Participation [5%]	Paper Assignments [10%]	CA [60%]	CA %	CA %	Required Exam Score to Pass Course (%)	
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	6	41
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	7	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	7	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	6	81
5	Lasso	0	0	1	0	5	6	10	6	6	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	7	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	6	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	7	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	8	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	6	48
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	6	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	6	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	5	102

<http://bit.ly/2Jg6GIk>

March 15, 2021

CSC 5741 (2020/21) L01 - 43

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

March 15, 2021

CSC 5741 (2020/21) L01 - 44

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>)

March 15, 2021

CSC 5741 (2020/21) L01 - 45

Course Management (2/2)

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

The screenshot shows a Google Groups interface. At the top, there's a search bar and various navigation buttons like 'NEW TOPIC', 'Mark all as read', 'Actions', and 'Filters'. Below the header, there are sections for 'Groups' (My groups, Home, My discussions, Started), 'Favorites' (with a note to click on a group's star icon to add it to your favorites), and 'Recently viewed' (listing CSC 5741: Data Mining and Warehousing, LIS 4012: Research in Devel..., and LIS 5310: Information Syst...). A message at the top right says: 'In May of 2019, we'll be merging and deprecating some of our settings to make group management easier.' A link to 'Learn more' is provided. On the right side, there are links for 'Manage group', 'Manage members', 'Members', and 'About'. At the bottom, there are buttons for 'Edit welcome message' and 'Clear welcome message', followed by a reminder: 'Reminder: We Have Class at 17H00 T' and a link: <https://groups.google.com/a/unza.zm/d/forum/csc5741>.

March 15, 2021

CSC 5741 (2020/21) L01 - 46

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 47

Q & A Session

- **Comments, concerns and complaints?**

March 15, 2021

CSC 5741 (2020/21) L01 - 48

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

March 15, 2021

CSC 5741 (2020/21) L01 - 49

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

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

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh
2 January 2020

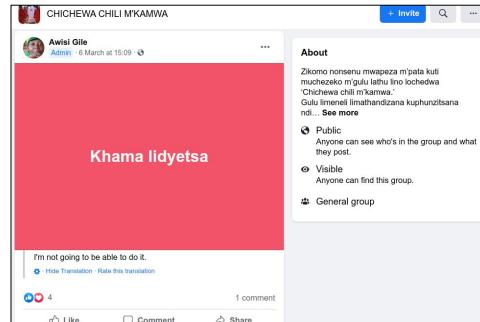


CSC 5741 (2020/21) L01 - 50

March 15, 2021

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

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

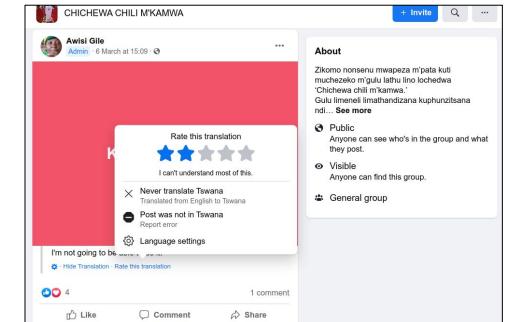


March 15, 2021

CSC 5741 (2020/21) L01 - 51

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

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

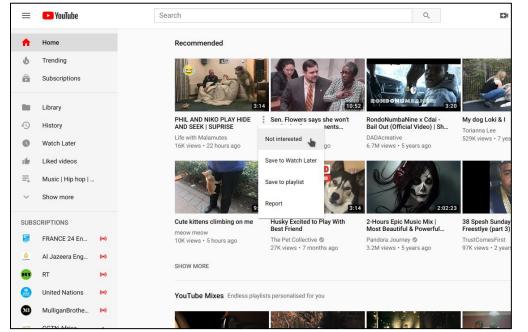


March 15, 2021

CSC 5741 (2020/21) L01 - 52

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

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



March 15, 2021

CSC 5741 (2020/21) L01 - 53

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

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

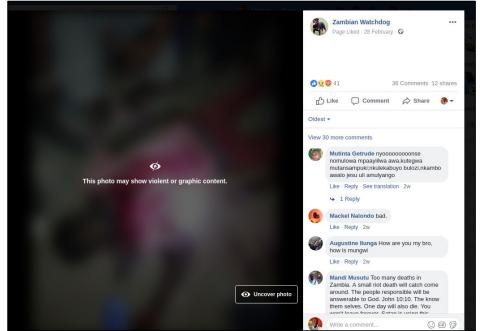
A screenshot of a YouTube channel content page titled "Your video". It shows a thumbnail for a video titled "2020/21 ICT 3020 | Lecture Series #...". Below the thumbnail are sections for "Details", "Analytics", "Editor", "Comments", and "Subtitles". On the right side, there's a "Thumbnail" section with a placeholder message, a "Playlists" section with a link to "Live Lectures | 2020/21 ICT 3020 | Screencasts", and a "Video details" sidebar with links to various resources.

March 15, 2021

CSC 5741 (2020/21) L01 - 54

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



March 15, 2021

CSC 5741 (2020/21) L01 - 55

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

March 15, 2021

CSC 5741 (2020/21) L01 - 56

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

March 15, 2021

CSC 5741 (2020/21) L01 - 57

```
<header>
  <identifier>oai:dspace.cbu.ac.zm:123456789/6
  <datestamp>2011-08-18T08:59:44Z</datestamp>
  <setSpec>hdl_123456789_23</setSpec>
</header>
<metadata>
  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:title>
      Customer service management in the retail
    <dc:title>
    <dc:creator>Atanga, Muyenga</dc:creator>
    <dc:subject>Banks</dc:subject>
    <dc:subject>Retail banking</dc:subject>
    <dc:subject>MBA THESIS</dc:subject>
    <dc:subject>Customer service</dc:subject>
    <dc:description>v.116p.</dc:description>
    <dc:description>Copperbelt University, Sch</dc:description>
    <dc:date>2011-07-19T14:32:14Z</dc:date>
    <dc:date>2011-07-19T14:32:14Z</dc:date>
    <dc:date>2011-07-19</dc:date>
    <dc:type>Thesis</dc:type>
  </oai_dc:dc>
</metadata>
```

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

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

March 15, 2021

CSC 5741 (2020/21) L01 - 58

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

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

March 15, 2021

CSC 5741 (2020/21) L01 - 59

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



March 15, 2021

CSC 5741 (2020/21) L01 - 60

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

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

```
> colnames(dataset_cfs0405_crop)
[1] "PROV"   "DIST"  "CONST" "WARD"  "REGION" "CSA" 
[7] "SEAT"   "HHNUM" "CROP"  "ID009"  "S1ACF01"
[13] "S1ACF02" "S1ACF03" "S1ACF04" "S1ACF05" "S1ACF06" "S1ACF07"
[19] "S1ACF08"  "S1ACF09"  "S1ACF10"  "S1ACF11"  "S1ACF12"  "S1ACF13"
[25] "S1ACF14"  "S1ACF15"  "S1ACF16"  "OTHARV"  "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 1 1 1 388.84409
```

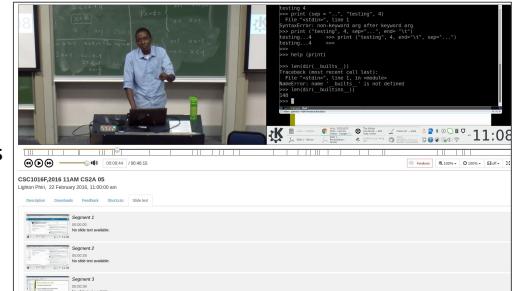
March 15, 2021

CSC 5741 (2020/21) L01 - 61

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

- Open Matterhorn
Video Segmentation Analysis

- Seeking to points of interest



March 15, 2021

CSC 5741 (2020/21) L01 - 62

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

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



March 15, 2021

CSC 5741 (2020/21) L01 - 63

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



March 15, 2021

CSC 5741 (2020/21) L01 - 64

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

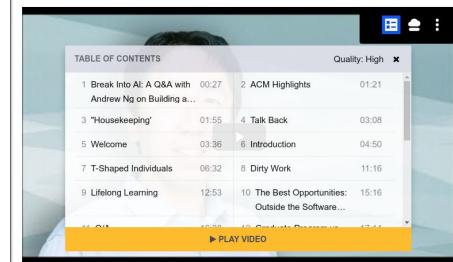
March 15, 2021

CSC 5741 (2020/21) L01 - 55

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?
 - [...]
 - [...]

Break Into AI: Building a Career in Machine Learning with Andrew Ng
December 4, 2018



Andrew Ng will share tips and tricks on how to break into AI. He will discuss some of the most valuable

CSC 5741 (2020/21) L01 - 66

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

- Curiosity-driven research
 - Puzzles
 - Games

The rise of machine learning in astronomy
September 4, 2018, Particle

The SKA will have over 2000 radio dishes and 2 million low-frequency antennas once finished. Credit: The Square Kilometer Array.

When mapping the universe, it pays to have some smart programming. Experts say.

March 15, 2021

CSC 5741 (2020/21) L01 - 67

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

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

Government wants help with monitoring content from Radio and TV stations in Zambia-Silaya
March 18, 2019, AllAfrica

Government Spokesperson, Dora Silaya, who is also Information and Broadcasting Minister, launching a new research on the role of Radio and Television stations in monitoring content.

Minister of Information and Broadcasting Services Dora Silaya says the

THE 125TH CANTON FAIR SPRING, 2019

skip Ad

CSC 5741 (2020/21) L01 - 68

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: GOMI JAKY/AFRIQUE MEDIAS

ZRA drones land on 7 trucks

From page 1

The impounded trucks were general loads in the bed during a routine road safety operation by ZRA authorities in Zambia.

He also confiscated 20 heavy-duty trucks and earth-moving machinery, and under valuation is a series of vehicles worth K100 million. "We will take appropriate action in accordance with the law," he said. "The ZRA has been given a mandate by the government," Ms Shikabala said. "We have been given a mandate to ensure that we have no illegal goods in Zambia."

ZRA authorities are investigating

the matter and will release the vehicles.

Shikabala said the ZRA has so far seized

over 100 trucks since the beginning of the year.

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

CSC 5741 (2020/21) L01 - 69

March 15, 2021

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

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



CSC 5741 (2020/21) L01 - 71

March 15, 2021

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

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



CSC 5741 (2020/21) L01 - 70

March 15, 2021

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

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



CSC 5741 (2020/21) L01 - 72

March 15, 2021

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

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



March 15, 2021

CSC 5741 (2020/21) L01 - 73

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

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



March 15, 2021

CSC 5741 (2020/21) L01 - 74

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

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

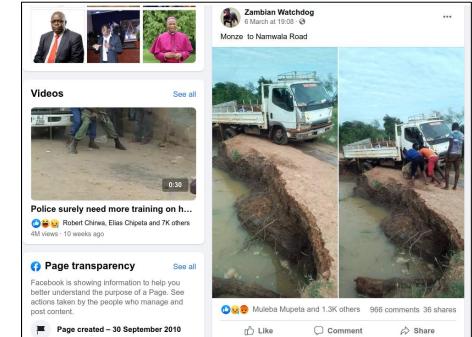


March 15, 2021

CSC 5741 (2020/21) L01 - 75

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

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

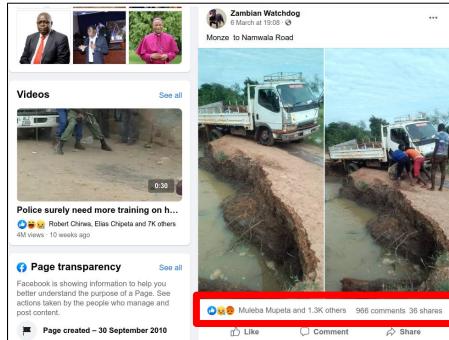


March 15, 2021

CSC 5741 (2020/21) L01 - 76

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

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



March 15, 2021

CSC 5741 (2020/21) L01 - 77

Data is Key to Data Mining

A screenshot of a LinkedIn course page titled 'Data is Key to Data Mining'. The page shows course resources, including software like Dia, Gantt, ProjectLibre, and Git. It also features a video thumbnail of a memorial service and a comment section.

March 15, 2021

CSC 5741 (2020/21) L01 - 78

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

March 15, 2021

CSC 5741 (2020/21) L01 - 79

Introduction (2/2)

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

March 15, 2021

CSC 5741 (2020/21) L01 - 80

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

March 15, 2021

CSC 5741 (2020/21) L01 - 81

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

March 15, 2021

CSC 5741 (2020/21) L01 - 82

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 83

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

March 15, 2021

CSC 5741 (2020/21) L01 - 84

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 85

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 86

Q & A Session

- Comments, concerns and complaints?

March 15, 2021

CSC 5741 (2020/21) L01 - 87

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**

March 15, 2021

CSC 5741 (2020/21) L01 - 88

Readings and Paper Summaries (1/5)

The screenshot shows the Mendeley Desktop application interface. The top menu bar includes File, Edit, View, Tools, and Help. Below the menu is a toolbar with icons for Add, Folders, Related, Sync, and Help. The main window title is "Mendeley Desktop". A sidebar on the left lists "My Library" sections: Literature Search, My Library (selected), Recently Added, Recently Read, and Favorites. The main content area displays a table titled "All Documents" with columns for Authors, Title, Year Published, and In Added. The table contains three entries: Willinsky, John (Open Journal Systems, 2005); Akakandewla, Ak... (Author Collaboration and Productivit..., 2009); and Kulyambaniso, C... (Faculty Productivity at The Universit..., 2016). To the right of the table are tabs for Details, Notes, and Content. A status bar at the bottom right shows "Creating a National Africa" and "Mendeley Desktop".

March 15, 2021

CSC 5741 (2020/21) L01 - 89

Readings and Paper Summaries (2/5)

The screenshot shows the Google Scholar search interface. The search bar at the top contains the query "data mining and machine learning". Below the search bar, there are two tabs: "Articles" (selected) and "Books". The "Articles" tab displays a list of search results. The first result is a citation for a book titled "Data Mining: Practical machine learning tools and techniques" by Ian H. Witten, Eibe Frank, and Mark A. Hall, published in 2016. The second result is a paper titled "Distributed GraphLab: a framework for machine learning and data mining in the cloud" by Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, et al., presented at the Proceedings of the ... 2012. Both results show metrics: the book has 99 citations and 34 versions; the paper has 99 citations and 29 versions. Below these results, there is a section titled "Business data mining – a machine learning perspective" which includes a citation for a book by I. Bošnjak and R.K. Matzatza from 2001. At the bottom of the page, there is a link to "The WEKA data mining software" and the URL "https://scholar.google.com".

March 15, 2021

CSC 5741 (2020/21) L01 - 90

Readings and Paper Summaries (3/5)

		Whatever comes to your mind			
		Rank	Conference (Full Name)	Short Name	H5-Index
Computer Science	All	1	International World Wide Web Conferences	WWW	66.00
	High Performance Computing	2	Information Sciences	Inf. Sci.	62.00
	Computer Network	3	ACM Knowledge Discovery and Data Mining	KDD	56.00
	Network and Information Security	4	IEEE Transactions on Knowledge and Data Engineering	TKDE	53.00
	Software Engineering	5	ACM International Conference on Web Search and Data Mining	WSDM	50.00
	Database and Data Mining	6	International Conference on Research an Development in Information Retrieval	SIGIR	47.00
	Theoretical Computer Science	7	Journal of the American Society for Information Science and Technology	JASIST	42.00
		8	IEEE International Conference on Data Engineering	ICDE	40.00
		9	ACM International Conference on Information and Knowledge Management	CIKM	38.00
		10	IEEE International Conference on DataMining	ICDM	33.00
		11	Journal of Web Semantics	J. Web Sem.	33.00
		12	Knowledge and Information Systems	KAIS	https://aminer.org

March 15, 2021

CEC E741 (2020/21) | 01 - 81

Readings and Paper Summaries (4/5)

Best Paper Awards in Computer Science (since 1996)	
By Conference:	AAAI ACL CHI CIKM CVPW FOCS FSE ICCV ICML ICSE IJCAI INFOCOM KDD MOBICOM NSDI OSDI PLDI PODS S&P SIGCOMM SIGIR SIGMETRICS SIIS
Institutions with the most Best Papers	
Much of this data was entered by hand (obtained by contacting past conference organizers, retrieving cached conference websites, and searching CVs) so please email me if you notice any errors or omission area, but some conferences do not have such an award (e.g. SIGGRAPH, CAV). "Distinguished paper award" and "outstanding paper award" are included but not "best student paper" (e.g. NIPS) or "best 1	
AAAI (Artificial Intelligence)	
2018	Memory-Augmented Monte Carlo Tree Search Cherjun Xiao, University of Alberta; et al.
2017	Labeled-SuperVision of Neural Networks with Physics and Domain Knowledge Russell Stewart & Stefano Ermon, Stanford University
2016	Bidirectional Selection That is Guaranteed to Meet in the Middle Robert C. Holte, University of Alberta; et al.
2015	From Non-Negative to General Operator Cost Partitioning Florian Pommerehne, University of Basel; et al.
2014	Recovering from Selection Bias in Causal and Statistical Inference Elias Bareinboim, University of California Los Angeles; et al.
2013	RC-Search: Learning Heuristics and Cost Functions for Structured Prediction Janardhan Rao Doppa, Oregon State University; et al.
2012	SMiLE: Shuffled Multi-Instance Learning Gary Dorn & Sourya Ray, Case Western Reserve University
2011	Document Summarization Based on Data Reconstruction Zhenyang He, Zhejiang University; et al.
2010	Dynamic Resource Allocation in Conservation Planning Daniel Golovin, California Institute of Technology; et al.
2009	Complexity of and Algorithms for Borda Manipulation Jessica Davies, University of Toronto; et al.
2008	How Incomplete Is Your Semantic Web Reasoner? Systematic Analysis of the Completeness of Query Ans. Georgios Stolas, Oxford University; et al.
2007	A Novel Transition Based Encoding Scheme for Planning as Satisfaction Ruoyun Huang, Washington University in St. Louis; et al.
2006	How Good is Almost Perfect? Mark Helmert & Gabriele Röder, Albert-Ludwigs-Universität Freiburg
2005	Optimal False-Name-Proof Voting Rules with Costly Voting Lisa Wagner & Vincent Conitzer, Duke University
2004	PLOW: A Collaborative Task Learning Agent http://jehuanz.com/best_paper_awards.html

March 15, 2021

CSEC EZ41 (2020/21) L01 82

Readings and Paper Summaries (5/5)

Zambia ICT Journal Announcements Current Archives About ▾

The Zambia ICT Journal (ISSN: 2016-2156) is published four times a year by the ICT Association of Zambia (ICTAZ) with technical support from the University of Zambia, Copperbelt University and Mulungushi University. The objective of Journal is to support and stimulate active productive research which could strengthen the technical foundations of engineers and scientists in the African continent, develop strong technical foundations and skills and lead to new small to medium enterprises within the African sub-continent. We also seek to encourage the emergence of functionally skilled technocrats within the continent on publishing research results and studies in Computer Science and Information Technology through a scholarly publication. The Zambia ICT journal is double blind peer reviewed.

Announcements

Call for paper for Volume 3 Issue 2 (June 2019)
2019-03-08
The Zambia ICT Journal wishes to call for original research papers containing new research findings which have not been published elsewhere.
<http://ictjournal.icict.org.zm>

March 15, 2021 CSC 5741 (2020/21) L01 - 93

On How to Read a Paper (1/5)

University of Cape Town My Author Page My Binders SIGN OUT Lighten Print SEARCH

ACM DL DIGITAL LIBRARY Tools and Resources Buy this Article (PRINT) Recommend the ACM DL to your organization TOC Service Email RSS https://dl.acm.org

How to read a paper Full Text: PDF Author: S.Keshav University of Waterloo Published in: Newsletter ACM SIGCOMM Computer Communication Review archive

2007 Article Bibliometrics

Reading a computer science research paper Full Text: PDF Author: Phillip W.L. Fong University of Calgary,Calgary, Alberta, Canada Published in: Newsletter ACM SIGCSE Bulletin archive Volume 41 Issue 2, June 2009

2009 Article Bibliometrics Citation Count: 4

March 15, 2021 CSC 5741 (2020/21) L01 - 94

On How to Read a Paper (2/5)

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

March 15, 2021

CSC 5741 (2020/21) L01 - 95

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

March 15, 2021

CSC 5741 (2020/21) L01 - 96

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

March 15, 2021

CSC 5741 (2020/21) L01 - 97

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.

March 15, 2021

CSC 5741 (2020/21) L01 - 98

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

March 15, 2021

CSC 5741 (2020/21) L01 - 99

Public Talks

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

March 15, 2021



CSC 5741 (2020/21) L01 - 100

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

March 15, 2021

CSC 5741 (2020/21) L01 - 101

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

March 15, 2021

CSC 5741 (2020/21) L01 - 102

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]

March 15, 2021

CSC 5741 (2020/21) L01 - 103

Getting Started with Python, SciKit-learn & Pandas

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



March 15, 2021

CSC 5741 (2020/21) L01 - 104

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>

March 15, 2021

CSC 5741 (2020/21) L01 - 105

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>

March 15, 2021

CSC 5741 (2020/21) L01 - 106

Q & A Session

- Comments, concerns and complaints?

March 15, 2021

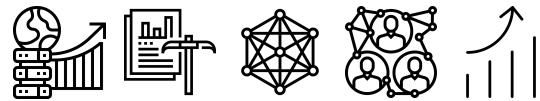
CSC 5741 (2020/21) L01 - 107

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>