

## CSC 5741 (2021/22)

# 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 21, 2021

CSC 5741 (2021/22) 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

	Chiriong Hamoya Head of Department, Lecturer, Researcher
	Akakanetwa Akakanetwa Senior Lecturer, Researcher
	Abel Mwalema Lecturer, Researcher
	Besson Njibvu Lecturer, Researcher
	Edward Mwalemu Lecturer, Researcher
	Francis 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 21, 2021

CSC 5741 (2021/22) 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 emphasise 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 orchestration](#) (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 [Airtel](#).

Earlier, I studied Software Engineering at UNZA ([unza.ac.zm](#)).

**Research:** My current research interests, alongside 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 21, 2021

CSC 5741 (2021/22) 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 [Airtel](#).

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 21, 2021

CSC 5741 (2021/22) L01 - 6

## Personal Introductions (4/5)

- The DataLab research group at The University of Zambia is composed of faculty staff and students—undergraduate and postgraduate—working in three main areas
  - Data Mining
  - Digital Libraries
  - Technology-Enhanced Learning

**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 education domains—two areas that are of significance in the so-called developing world.

 **Lighton Phiri**  
Academic Staff

 **Robert Wendo**  
Masters Student  
MSc Computer Science

**Digital Libraries**  
The field of Digital Libraries (DL) generally involves the management of digital collections of information and corresponding network-based services that enable retrieval and delivery of collections. DL are often referred to as digital object systems that are used to permanently 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.

 **Mathew Mbewe**  
Undergraduate Student  
MSc Computer Science

 **Mathew Mbewe**  
Undergraduate Student

<http://datalab.unza.zm>

March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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

March 21, 2021

CSC 5741 (2021/22) L01 - 12

SESSIONAL DATES FOR 2021/22 ACADEMIC YEAR	
TERM I	
Monday 14 <sup>th</sup> February, 2022 to Monday 14 <sup>th</sup> March, 2022	On-line Registration
Monday 14 <sup>th</sup> February, 2022 to Friday 18 <sup>th</sup> February, 2022 20 <sup>th</sup> February, 2022	Orientation of First Year students
Monday 21 <sup>st</sup> February, 2022 to Friday 3 <sup>rd</sup> June, 2022 15 <sup>th</sup> March, 2022 to Sunday 22 <sup>nd</sup> March, 2022	Arrival of Returning Regular Students
Monday 25 <sup>th</sup> to Friday 29 <sup>th</sup> April, 2022	Lectures for Regular Students Start (15 weeks)
	Late Registration
	Graduation Week

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

## Course Structure (3/9)

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

SESSIONAL DATES FOR 2021/22 ACADEMIC YEAR	
TERM I	
Monday 14 <sup>th</sup> February, 2022 to Monday 14 <sup>th</sup> March, 2022	On-line Registration
Monday 14 <sup>th</sup> February, 2022 to Friday 18 <sup>th</sup> February, 2022 20 <sup>th</sup> February, 2022	Orientation of First Year students
Monday 21 <sup>st</sup> February, 2022 to Friday 3 <sup>rd</sup> June, 2022 15 <sup>th</sup> March, 2022 to Sunday 22 <sup>nd</sup> March, 2022 Monday 25 <sup>th</sup> to Friday 29 <sup>th</sup> April, 2022	Arrival of Returning Regular Students Lectures for Regular Students Start (15 weeks) Late Registration Graduation Week

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

March 21, 2021

CSC 5741 (2021/22) 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

SESSIONAL DATES FOR 2021/22 ACADEMIC YEAR	
TERM I	
Monday 14 <sup>th</sup> February, 2022 to Monday 14 <sup>th</sup> March, 2022	On-line Registration
Monday 14 <sup>th</sup> February, 2022 to Friday 18 <sup>th</sup> February, 2022 20 <sup>th</sup> February, 2022	Orientation of First Year students
Monday 21 <sup>st</sup> February, 2022 to Friday 3 <sup>rd</sup> June, 2022 15 <sup>th</sup> March, 2022 to Sunday 22 <sup>nd</sup> March, 2022 Monday 25 <sup>th</sup> to Friday 29 <sup>th</sup> April, 2022	Arrival of Returning Regular Students Lectures for Regular Students Start (15 weeks) Late Registration Graduation Week

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

March 21, 2021

CSC 5741 (2021/22) L01 - 14

## Course Structure (5/9)

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

SESSIONAL DATES FOR 2021/22 ACADEMIC YEAR	
TERM I	
Monday 14 <sup>th</sup> February, 2022 to Monday 14 <sup>th</sup> March, 2022	On-line Registration
Monday 14 <sup>th</sup> February, 2022 to Friday 18 <sup>th</sup> February, 2022 20 <sup>th</sup> February, 2022	Orientation of First Year students
Monday 21 <sup>st</sup> February, 2022 to Friday 3 <sup>rd</sup> June, 2022 15 <sup>th</sup> March, 2022 to Sunday 22 <sup>nd</sup> March, 2022 Monday 25 <sup>th</sup> to Friday 29 <sup>th</sup> April, 2022	Arrival of Returning Regular Students Lectures for Regular Students Start (15 weeks) Late Registration Graduation Week

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

March 21, 2021

CSC 5741 (2021/22) L01 - 15

## Course Structure (6/9)

Today	<	>	March 2022	MON	TUE	WED	THU	FRI	SAT	SUN
				21	22	23	24	25	26	27
GMT+02										
14:00										
15:00										
16:00										
17:00										
18:00			CSC 5741: Lecture 17:30 - 19:30 Classroom #3, Department of Computer Science.							
19:00										
20:00										
21:00										

March 21, 2021

CSC 5741 (2021/22) L01 - 16

## Course Structure (7/9)

- Course Resources

- All course resources will be made available on Astria.

March 21, 2021 CSC 5741 (2021/22) L01 - 17

## Course Structure (8/9)

- Course Resources

- All course resources will be made available on Astria.

March 21, 2021 CSC 5741 (2021/22) L01 - 18

## 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)

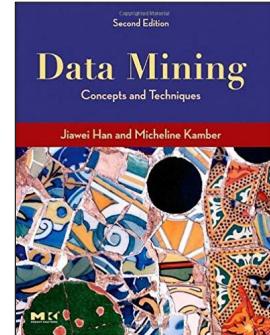
March 21, 2021

CSC 5741 (2021/22) L01 - 19

## Prescribed & Recommended Textbooks (1/4)

- Data Mining Concepts and Techniques

- J. Han and M. Kamber (2011)

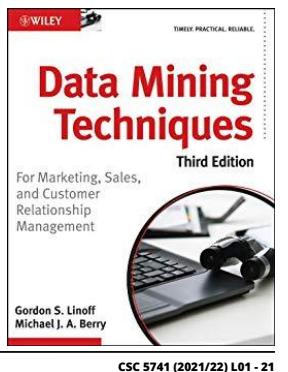


March 21, 2021

CSC 5741 (2021/22) 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)

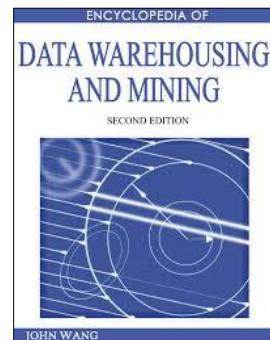


March 21, 2021

CSC 5741 (2021/22) L01 - 21

## Prescribed & Recommended Textbooks (3/4)

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

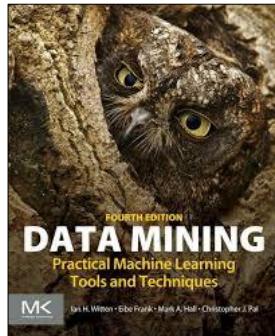


March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 main page of the scikit-learn website. It features a header with navigation links for Home, Installation, Documentation, Examples, and a search bar. Below the header, there's a banner with the text "Machine Learning in Python". The main content area is divided into several sections:

- Classification:** Describes identifying which category an object belongs to. Includes applications like spam detection, image recognition, and algorithms like SVM, nearest neighbors, random forests, etc.
- Regression:** Describes predicting a continuous-valued attribute associated with an object. Includes applications like drug response, stock prices, and algorithms like OVR, ridge regression, Lasso, etc.
- Clustering:** Describes automatic grouping of similar objects into sets. Includes applications like customer segmentation, document retrieval, and algorithms like k-means, spectral clustering, mean-shift, etc.
- Dimensionality reduction:** Describes reducing the number of random variables to consider. Includes applications like visualization, increased efficiency, and algorithms like PCA, feature selection, non-negative matrix factorization, etc.
- Model selection:** Describes comparing, validating, and choosing parameters and models. Includes a goal of improved accuracy via parameter tuning.
- Preprocessing:** Describes feature extraction and normalization. Includes applications like transforming input data such that it can be used with machine learning models.

<https://scikit-learn.org>

March 21, 2021

CSC 5741 (2021/22) L01 - 25

## Tools and Services (3/6)

- **Pandas**
  - Python data analysis library

The screenshot shows the pandas Python Data Analysis Library website. It has a header with links for Home, About, Get Pandas, Documentation, Community, and Talks. The main content area includes:

- pandas**: The logo and the equation  $y_t = \beta^T x_t + \mu_t + \epsilon_t$ .
- Versions**: A table showing versions from 0.24.1 to 0.25.0, with download links for each.
- Development**: Information about the project being a NumFOCUS sponsored project.
- A Fiscally Sponsored Project of NUMFOCUS**: The text "OPEN CODE + BETTER SCIENCE".
- v0.23.4 Final (August 3, 2018)**: A note about minor bugfixes and performance improvements.
- Upgrades**: A note about upgrading to v0.23.4.
- GitHub**: A link to the GitHub repository.
- Links**: A section with links to GitHub, Issues, Pull Requests, and Releases.
- https://pandas.pydata.org**: The URL of the site.

March 21, 2021

CSC 5741 (2021/22) L01 - 26

## Tools and Services (4/6)

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

The screenshot shows the Matplotlib website. It features a header with links for Home, Examples, Tutorials, API, and Docs. The main content area includes:

- Matplotlib**: The logo and version 3.0.3.
- What is Matplotlib?**: A brief description of Matplotlib as a 2D plotting library.
- Installation**: Instructions for installing Matplotlib.
- Documentation**: The documentation for Matplotlib version 3.0.3.
- Sample plots**: A section showing various types of plots like line plots, scatter plots, heatmaps, and 3D surface plots.
- API**: A link to the Matplotlib API documentation.
- https://matplotlib.org**: The URL of the site.

March 21, 2021

CSC 5741 (2021/22) L01 - 27

## Tools and Services (5/6)

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

The screenshot shows the Google Colab website. It features a header with links for Welcome To Colaboratory, File, Edit, View, Insert, Runtime, Tools, Help, and a search bar. The main content area includes:

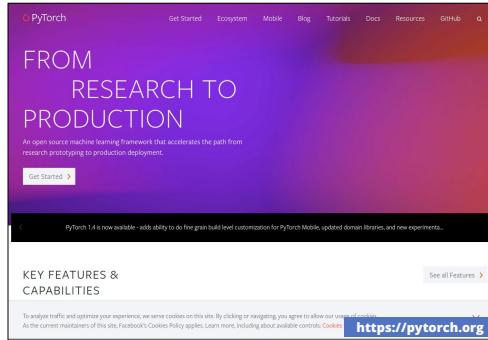
- Table of contents**: A sidebar with sections like Getting started, Machine learning, More resources, and Machine Learning Examples.
- What is Colaboratory?**: A brief description of Colaboratory as a Python environment for writing and executing code in a browser.
- Getting started**: A section with tips like "Zero configuration required", "Free access to GPUs", and "Easy sharing".
- Code cell**: A code cell containing Python code that calculates the value of a variable based on time.
- https://colab.research.google.com**: The URL of the site.

March 21, 2021

CSC 5741 (2021/22) L01 - 28

## Tools and Services (6/6)

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



March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) 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	<a href="http://bit.ly/2Jg6Gik">http://bit.ly/2Jg6Gik</a>

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

March 21, 2021

CSC 5741 (2021/22) 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/Zambia

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 Mulanza	Francis Chulu	Inonge Lamaswala

CSC 5741 (2021/22) L01 - 33

March 21, 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 21, 2021

CSC 5741 (2021/22) 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	20	19	95 <b>18.28</b>	
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		

<http://bit.ly/zjg60k>

March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) L01 - 36

## Course Grading—Class Participation

1. John Daka	<a href="#">Using data mining for bank direct marketing: an application of the CRISP-DM methodology</a>
2. Inonge Lamaswala	<a href="#">Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment</a>
3. Mubanga Mubanga	<a href="#">A Novel Position-based Sentiment Classification Algorithm for Facebook Comments</a>
4. Nondi Mukuma	<a href="#">Speeding up Support Vector Machines</a>
5. David Mulenga	<a href="#">Mining Educational Data to Analyze Students' Performance</a>
6. Memory Munibi	<a href="#">Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics</a>
7. Kaumba Mutende	<a href="#">Classification of Diabetes patient by using Data Mining Techniques</a>
8. Justin Nongola	<a href="#">Educational Data Mining &amp; Students' Performance Prediction</a>
9. Anthony Sampa	<a href="#">TIGER POPULATION GROWTH PREDICTION</a>
10. Tasha Shamane	<a href="#">A System to Filter Unwanted Messages from OSN User Walls</a>
11. Mweemba Sikuyuba	<a href="#">Educational Data Mining Rule based</a> <a href="http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741">http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741</a>

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

March 21, 2021

CSC 5741 (2021/22) 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	<b>62</b>	<b>76</b>	<b>13.8</b>
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 21, 2021

CSC 5741 (2021/22) 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

<b>THE UNIVERSITY OF ZAMBIA</b> SCHOOL OF NATURAL SCIENCES  2018/19 ACADEMIC MID-YEAR FINAL EXAMINATIONS CSC 5741: DATA MINING AND WAREHOUSING
<b>MARKS: 100</b> <b>TIME: THREE (3) HOURS</b> <b>INSTRUCTIONS:</b> 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
<b>Question 1</b> It was recently reported <sup>1</sup> that the Government of The Republic of Zambia (GRZ) is working

March 21, 2021

CSC 5741 (2021/22) 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	<b>47.58</b>	<b>70</b>	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 21, 2021

CSC 5741 (2021/22) 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 [60%]	CA %	Required Exam Score to Pass Course (%)
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	41	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	41	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	38	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	81	81
5	Lasso	0	0	1	0	5	6	10	6	110	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	47	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	60	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	10	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	36	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	49	49
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	17	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	37	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	102	102

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

March 21, 2021

CSC 5741 (2021/22) 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 [60%]	CA %	Required Exam Score to Pass Course (%)
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	41	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	41	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	38	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	81	81
5	Lasso	0	0	1	0	5	6	10	6	110	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	47	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	60	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	10	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	36	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	49	49
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	17	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	37	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	102	102

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

March 21, 2021

CSC 5741 (2021/22) 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 [60%]	CA %	Required Exam Score to Pass Course (%)
1	Elastic Net	18.28	13.8	2.5	5	0	47.58	70	6	41	6
2	Linear SVC	8.24	8.8	4.5	5	7	33.54	56	5	41	41
3	SGD Classifier	9.28	8.7	5	5	7	34.98	58	5	38	38
4	Kernal Approximation	0.8	1.7	4	5	6	17.5	29	5	81	81
5	Lasso	0	0	1	0	5	6	10	6	110	110
6	Naive Bayes	9.08	5.6	4.5	5	7	31.18	52	5	47	47
7	Ensemble Classifiers	11.2	2.5	4	2.5	6	26.2	44	6	60	60
8	Spectral Clustering	17.32	12.2	4.5	5	7	46.02	77	10	10	10
9	Mean Shift	9.52	11.4	4	2.5	8	35.42	58	5	36	36
10	K Neighbors	10.6	7.3	3	5	5	30.9	52	5	49	49
11	SGD Regressor	16.82	10	3.5	5	8	43.02	72	17	17	17
12	K Means	16	7.7	3	2.5	6	35.2	58	5	37	37
13	MiniBatch K Means	0	2.8	1.5	0	5	9.3	16	5	102	102

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

March 21, 2021

CSC 5741 (2021/22) L01 - 43

March 21, 2021

CSC 5741 (2021/22) L01 - 44

## 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	<50	0

## Course Management (1/2)

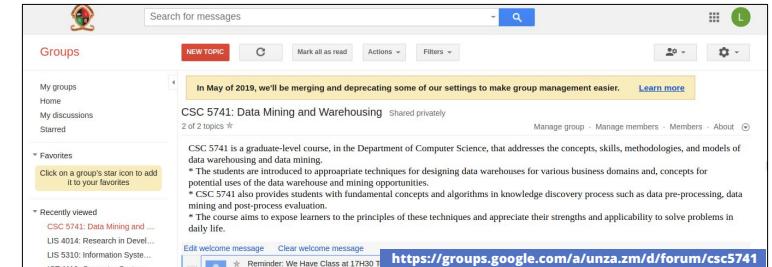
- Instructor: Lighton Phiri and TBA (possible Invited Talks)
- Email: [lighton.phiri@unza.zm](mailto:lighton.phiri@unza.zm)
- Office: Room 515, Fifth Floor, School of Education Building
- By appointment
  - Schedule an appointment via email after checking free/busy slots on my calendar (<https://goo.gl/6kHrnA>)

March 21, 2021

CSC 5741 (2021/22) L01 - 45

## Course Management (2/2)

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



In May of 2019, we'll be merging and deprecating some of our settings to make group management easier. [Learn more](#)

Groups

My groups Home My discussions Started

Favorites Click on a group's star icon to add it to your favorites

Recently viewed CSC 5741: Data Mining and... LIS 4012: Research in Devel... LIS 5310: Information Syste... HOT ALLOYS FOR HIGH TEMPERATURE

Manage group · Manage members · Members · About

2 of 2 topics \*

CSC 5741: Data Mining and Warehousing Shared privately

CSC 5741 is a graduate level course, in the Department of Computer Science, that addresses the concepts, skills, methodologies, and models of data warehousing and data mining.

- \* The students are introduced to appropriate techniques for designing data warehouses for various business domains and, concepts for potential uses of the data warehouse and mining opportunities.
- \* CSC 5741 also provides students with fundamental concepts and algorithms in knowledge discovery process such as data pre-processing, data mining and post-process evaluation.
- \* The course aims to expose learners to the principles of these techniques and appreciate their strengths and applicability to solve problems in daily life.

Edit welcome message Clear welcome message <https://groups.google.com/a/unza.zm/d/forum/csc5741>

March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) L01 - 47

## Q & A Session

- Comments, concerns and complaints?

March 21, 2021

CSC 5741 (2021/22) 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 21, 2021

CSC 5741 (2021/22) L01 - 49

## Contextualising Data Mining & Warehousing: Everyday Examples (1/7)

Android 9 Pie: Powered by AI for a smarter, simpler experience that adapts to you

Aug 06, 2018 - 5 min read

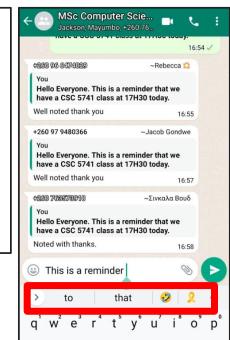
 Sameer Samat  
VP of Product Management, Android & Google Play



Share

- Applications of data mining techniques have become mainstream.

March 21, 2021



MSc Computer Scie... Jackson Mayombo +260 76... has been added to this conversation 16:54 ✓  
0880 00 0202009 ~Rebecca  
You Hello Everyone. This is a reminder that we have a CSC 5741 class at 17H30 today.  
Well noted thank you 16:55  
<260 97 4480366 ~Jacob Gordwe  
You Hello Everyone. This is a reminder that we have a CSC 5741 class at 17H30 today.  
Well noted thank you 16:57  
<260 70942009 ~Zwetska Boué  
You Hello Everyone. This is a reminder that we have a CSC 5741 class at 17H30 today.  
Noted with thanks. 16:58  
This is a reminder  
> to that 

CSC 5741 (2021/22) L01 - 50

## Contextualising Data Mining & Warehousing: Everyday Examples (2/7)

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



AI 'outperforms' doctors diagnosing breast cancer

 Fergal Walsh  
Medical correspondent @BBCFergalWalsh  
2 January 2020

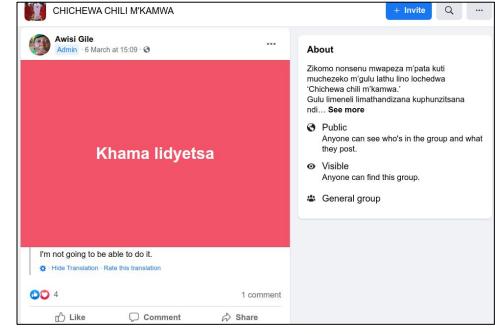
R-CC L-CC

March 21, 2021

CSC 5741 (2021/22) L01 - 51

## Contextualising Data Mining & Warehousing: Everyday Examples (3/7)

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



CHICHEWA CHILI MKAMWA

 Awisi Gile  
Admin - 6 March at 15:09

About

Zikomo nosenso mivapeza m'pata kuli mucchezeko m'gulu latu lino lochewda 'Chichewa chili m'kamwa'. Ondi ndi kulelo kudzidzana kupharuselena ndi... See more

Public Anyone can see who's in the group and what they post.

Visible Anyone can find this group.

General group

Khama lidyetsa

I'm not going to be able to do it.

How Translation Rate this translation

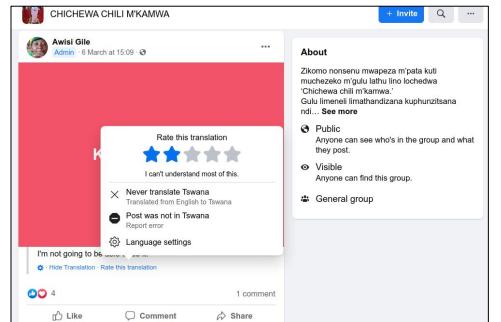
4 1 comment Like Comment Share

March 21, 2021

CSC 5741 (2021/22) L01 - 52

## Contextualising Data Mining & Warehousing: Everyday Examples (4/7)

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

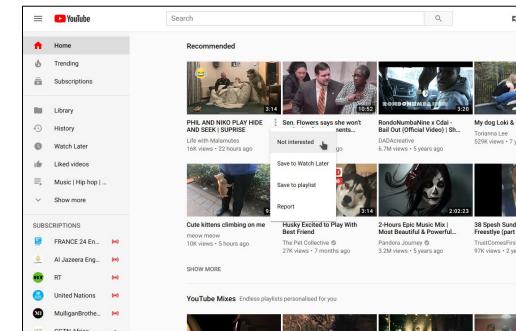


March 21, 2021

CSC 5741 (2021/22) L01 - 53

## Contextualising Data Mining & Warehousing: Everyday Examples (5/7)

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

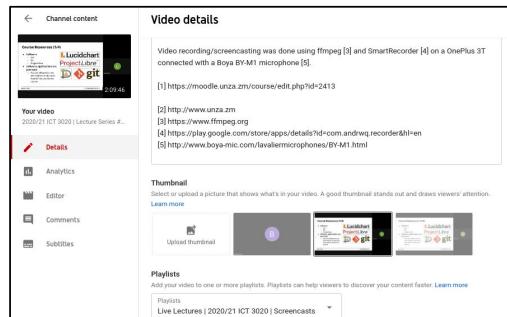


March 21, 2021

CSC 5741 (2021/22) L01 - 54

## Contextualising Data Mining & Warehousing: Everyday Examples (6/7)

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

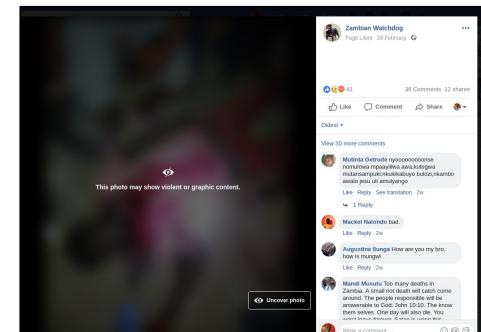


March 21, 2021

CSC 5741 (2021/22) L01 - 55

## Contextualising Data Mining & Warehousing: Everyday Examples (7/7)

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



March 21, 2021

CSC 5741 (2021/22) L01 - 56

# **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 21, 2021

CSC 5741 (2021/22) L01 - 57

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (2/9)**

- 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 21, 2021

---

CSC 5741 (2021/22) L01 - 59

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (1/9)**

## ACKNOWLEDGEMENT

My sincere gratitude goes to the almighty God for the opportunity, strength and courage he gave me to undertake this research. I would also like to acknowledge my supervisor Dr. Jackson Phiri PERSON for his continued support, patience, dedication and guidance he gave me in the course of my study.

His guidance, commitment and motivation helped me a lot during the course of the research. My acknowledgement also goes to Prof. **Philip O. Y. Nkunika PERSON** for his guidance during the research.

Dr. Mayumbo Nyirenda PERSON and the entire Department ORG for the support rendered during the period of doing the research. I also acknowledge my colleague Mr. Simon Chiwamba PERSON for working closely

with me for the entire period of the research. I would also like to thank the Food Agriculture Organization of the United Nations (FAO) for the necessary

- Automatic generation of descriptive metadata

- Natural Language Processing
  - Named Entity Recognition

March 21, 2021

CSC 5741 (2021/22) L01 - 58

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (3/9)**

- University of Zambia Ranking Committee Research Report

- Mining for scholarly output on the Web

March 21, 2021

CSC 5741 (2021/22) L01 - 60

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (4/9)**

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

March 21, 2021

CSC 5741 (2021/22) L01 - 61

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (5/9)**

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



March 21, 2021

CSC 5741 (2021/22) L01 - 62

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (6/9)**

- **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" "HNHUM" "CROP" "ID009" "SIAFIELD" "SIACF01"
[13] "SIACF02" "SIACF03" "SIACF04" "SIACF05" "SIACF06" "SIACF07"
[19] "SIACF08" "SIACF09" "SIACF10" "SIACF11" "SIACF12" "SIACF13"
[25] "SIACF14" "SIACF15" "SIACF16" "TOTHARV" "WEIGHT" "HA_HARV"
[31] "convert" "HA_PLANT"
> head(dataset_cfs0405_crop)
   PROV DIST CONST WARD REGION CSA SEA HNHUM          CROP ID009
1 Central Chilombo 1 1 14 1 55 Maize 1
2 Central Chilombo 1 3 1 2 2 77 Maize 1
3 Central Chilombo 1 3 1 2 2 33 Other crops (Specify) 2
4 Central Chilombo 2 12 1 2 3 96 Maize 3
5 Central Chilombo 2 12 1 2 3 96 Groundnuts 3

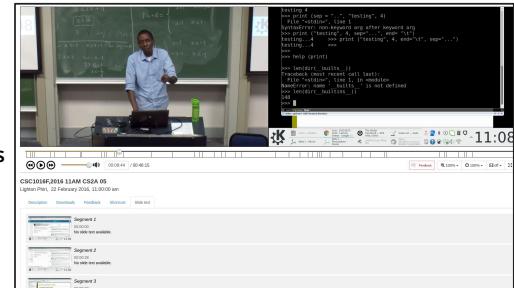
> colnames(dataset_cfs2004_2005_weight)
 [1] "ID001" "ID002" "ID003" "ID004" "ID005" "ID006" "ID007" "ID009"
 [9] "WEIGHT"
> head(dataset_cfs2004_2005_weight)
ID001 ID002 ID003 ID004 ID005 ID006 ID007 ID009 WEIGHT
1 Central Chilombo 1 1 14 1 1 388.84409
```

March 21, 2021

CSC 5741 (2021/22) L01 - 63

## **Contextualising Data Mining & Warehousing: Some Ongoing Projects (7/9)**

- Open Matterhorn Video Segmentation Analysis
    - Seeking to point of interest



March 21, 2021

---

CSC 5741 (2021/22) L01 - 64

## Contextualising Data Mining & Warehousing: Some Ongoing Projects (8/9)

- Automatic Content Generation

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



March 21, 2021

CSC 5741 (2021/22) L01 - 65

## Contextualising Data Mining & Warehousing: Some Ongoing Projects (9/9)

- 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



CSC 5741 (2021/22) L01 - 66

## 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 21, 2021

CSC 5741 (2021/22) L01 - 67

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

The screenshot shows a video player interface with a list of contents for a video titled "Break Into AI: Building a Career in Machine Learning with Andrew Ng". The contents are as follows:

Content	Duration
1 Break Into AI: A Q&A with Andrew Ng on Building a...	00:27
3 "Housekeeping"	01:55
5 Welcome	03:36
7 T-Shaped Individuals	06:32
9 Lifelong Learning	12:53
2 ACM Highlights	01:21
4 Talk Back	03:08
6 Introduction	04:50
8 Dirty Work	11:16
10 The Best Opportunities: Outside the Software...	15:16

March 21, 2021

CSC 5741 (2021/22) L01 - 68

## 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 Project

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

CSC 5741 (2021/22) L01 - 69

March 21, 2021

## 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-Siliya  
March 18, 2019



Minister of Information and Broadcasting Services Dora Siliya says the

CSC 5741 (2021/22) L01 - 70

March 21, 2021

## 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: BUTYUNA KAMBA

ZRA drones land on 7 trucks

From page 1  
The impounded trucks were parked in the bush during a routine inspection by ZRA officials. They found misplacement of the value-added tax (VAT) documents and other violations in some vehicles.  
ZRA will not relent in recovering VAT from traders.

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

March 21, 2021

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 72

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 73

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 74

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 75

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 76

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

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



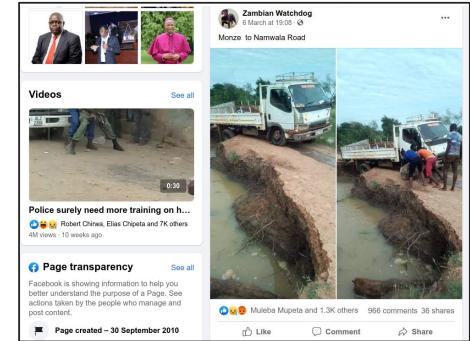
---

March 21, 2021

CSC 5741 (2021/22) L01 - 77

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

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



---

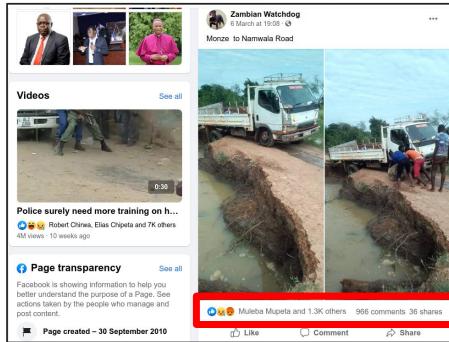
March 21, 2021

---

CSC 5741 (2021/22) L01 - 78

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

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



March 21, 2021

CSC 5741 (2021/22) L01 - 79

# **Data is Key to Data Mining**



March 21, 2021

---

CSC 5741 (2021/22) L01 - 80

## **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 21, 2021

CSC 5741 (2021/22) L01 - 81

## **Introduction (2/2)**

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

March 21, 2021

CSC 5741 (2021/22) L01 - 82

## **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 21, 2021

CSC 5741 (2021/22) L01 - 83

## **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 21, 2021

CSC 5741 (2021/22) L01 - 84

## 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 21, 2021

CSC 5741 (2021/22) L01 - 85

## 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 21, 2021

CSC 5741 (2021/22) L01 - 86

## 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 21, 2021

CSC 5741 (2021/22) L01 - 87

## Closing CSC 5741 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 21, 2021

CSC 5741 (2021/22) L01 - 88

## Q & A Session

- Comments, concerns and complaints?

March 21, 2021

CSC 5741 (2021/22) L01 - 89

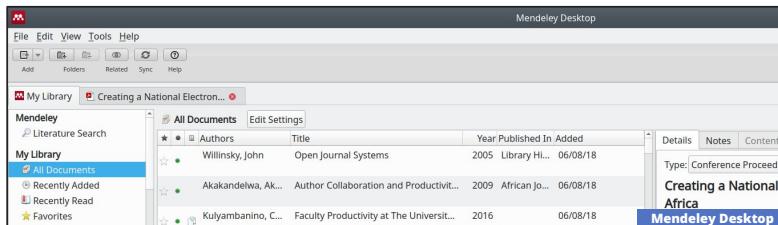
## Lecture Series Outline

- 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 21, 2021

CSC 5741 (2021/22) L01 - 90

## Readings and Paper Summaries (1/5)



March 21, 2021

CSC 5741 (2021/22) L01 - 91

## Readings and Paper Summaries (2/5)

A screenshot of a Google Scholar search results page. The query 'data mining and machine learning' is entered in the search bar. The results section shows a list of articles. The first result is '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 'Distributed GraphLab: a framework for machine learning and data mining in the cloud' by Yann Low, David Bickson, and Jure Leskovec. The third result is 'Business data mining—a machine learning perspective' by I. Bose and R.K. Mahapatra. At the bottom right, a link to 'https://scholar.google.com' is visible.

March 21, 2021

CSC 5741 (2021/22) L01 - 92

## Readings and Paper Summaries (3/5)

Whatever comes to your mind

EN | DE

Computer Science

Rank	Conference (Full Name)	Short Name	H5-Index
1	International World Wide Web Conferences	WWW	66.00
2	Information Sciences	Inf. Sci.	62.00
3	ACM Knowledge Discovery and Data Mining	KDD	56.00
4	IEEE Transactions on Knowledge and Data Engineering	TKDE	53.00
5	ACM International Conference on Web Search and Data Mining	WSDM	50.00
6	International Conference on Research an Development in Information Retrieval	SIGIR	47.00
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 21, 2021 CSC 5741 (2021/22) L01 - 93

## Readings and Paper Summaries (4/5)

### Best Paper Awards in Computer Science (since 1996)

By Conference: AAAI ACL CHI CIKM CVPR FOCS FSE ICCV ICML ICSE IJCAI INFOCOM KDD MOBICOM NSDI OSDI PODS S&P SIGCOMM SIGIR SIGMETRICS SODA TREC VLSI

#### 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 omissions area, but some conferences do not have such an award (e.g. SIGGRAPH, CAN)."Distinguished paper award" and "outstanding paper award" are included but not "best student paper" (e.g. NIPS) or "best"

#### AAAI (Artificial Intelligence)

2018	Memory-Augmented Monte Carlo Tree Search	Chenjun Xiao, University of Alberta; et al.
2017	Label-Free Supervision of Neural Networks with Physics and Domain Knowledge	Russell Stewart & Stefano Ermon, Stanford University
2016	Bi-directional Search 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 Pommerehning, 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	HC-Search: Learning Heuristics and Cost Functions for Structured Prediction	Janardhan Rao Doppa, Oregon State University; et al.
2012	SMILE: Shuffled Multiple-Instance Learning	Gary Doran & Soumya Ray, Case Western Reserve University
2011	Learning SVM Classifiers with Indefinite Kernels	Suicheng Gu & Yuhong Guo, Temple University
2010	Document Summarization Based on Data Reconstruction	Zhangyu He, Zhejiang University; et al.
2009	Dynamic Resource Allocation in Conservation Planning	Daniel Golovin, Carnegie Mellon University of Technology; et al.
2008	On the Completeness of and Approximation to Boolean Manipulation	Jean-Pierre Bousquet & Thomas J. Schatzki; et al.
2007	How Complete Is Your Semantic Web Reasoner? Systematic Analysis of the Completeness of Query Ans...	Georgios Stassis, Oxford University; et al.
2006	A Novel Translation Based Encoding Scheme for Planning as Satisfiability	Ruyun Huang, Washington University in St. Louis; et al.
2005	How Good Is Almost Perfect?	Matte Helmert & Gabriele Röger, Albert-Ludwigs-Universität Freiburg
2004	Optimal False-Name-Proof Voting Rules with Costly Voting	Liad Wagman & Vincent Conitzer, Duke University
2003	PLOW: A Collaborative Task Learning Agent	
2002		

[http://jeffhuang.com/best\\_paper\\_awards.html](http://jeffhuang.com/best_paper_awards.html)

March 21, 2021

CSC 5741 (2021/22) L01 - 94

## 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 and Mzumbe 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 before. The papers should be submitted online at <http://ictjournal.ictict.org.zm>.

March 21, 2021 CSC 5741 (2021/22) L01 - 95

## On How to Read a Paper (1/5)

University of Cape Town My Author Page My Editors SIGN OUT Light/Dark Print SEARCH

ACM 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: Philip W.L. Fong University of Calgary\_Calgary\_Alberta\_Canada Published in: Newsletter ACM SIGCSE Bulletin archive 2009 Article Bibliometrics Citation Count: 4

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

March 21, 2021 CSC 5741 (2021/22) L01 - 96

## On How to Read a Paper (2/5)

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

March 21, 2021

CSC 5741 (2021/22) L01 - 97

## 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 21, 2021

CSC 5741 (2021/22) L01 - 98

## 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 21, 2021

CSC 5741 (2021/22) L01 - 99

## 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 21, 2021

CSC 5741 (2021/22) L01 - 100

## 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 21, 2021

CSC 5741 (2021/22) L01 - 101

## 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 21, 2021

SCHOOL OF AGRICULTURAL SCIENCES  
SEMINAR SERIES



### *Precision-Farming Opportunities in Zambia*

Ms. Sheila Zulu

AGRONOMIST & FARM SOLUTIONS SPECIALIST – AGCO CORPORATION - ZAMBIA

All students should attend

DATE: Thursday, 24<sup>th</sup> March, 2022

TIME: 12:00-13:00 hrs

VENUE: VET LECTURE THEATRE (VLT)

## Public Oral Examinations

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

PhD Public Defense

Title: Title: Uranium Exposure and Health Impact Assessment Among Communities in the Vicinity of a Uranium Mine In Siavonga District, Zambia

Candidate: Titus Haakonde, PhD in Environmental Toxicology



Supervisor: Dr. John Yabe  
Date: 28<sup>th</sup> February 2022  
Time: 14:15  
Meeting Platform: [meet.google.com/ljp-xqz-qbw](https://meet.google.com/ljp-xqz-qbw)  
Examiners: Dr. V. Wepener (North West University), Dr. K. Muzandu (UNZA), Dr. E. Mkandawire (UNZA)

March 21, 2021

CSC 5741 (2021/22) L01 - 103

## 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.

March 21, 2021

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

CSC 5741 (2021/22) L01 - 104

## 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 21, 2021

CSC 5741 (2021/22) L01 - 105

## Getting Started with Python, SciKit-learn & Pandas

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



March 21, 2021

CSC 5741 (2021/22) L01 - 106

## 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] Chipangila, B., Liswaniso, E., Mawila, A., Mwanza, P., Nawila, D., M'sendo, R., Nyirenda, M., & Phiri, L. (2021, September). Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 100-109). IEEE.  
<https://doi.org/10.1109/JCDL52503.2021.00022>

March 21, 2021

CSC 5741 (2021/22) L01 - 107

## Academic Talk: L. Phiri [Trial]

- [1] Chipangila, B., Liswaniso, E., Mawila, A., Mwanza, P., Nawila, D., M'sendo, R., Nyirenda, M., & Phiri, L. (2021, September). Improved Discoverability of Digital Objects in Institutional Repositories Using Controlled Vocabularies. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 100-109). IEEE.  
URL: <https://doi.org/10.1109/JCDL52503.2021.00022>

March 21, 2021

CSC 5741 (2021/22) L01 - 108

## Q & A Session

- Comments, concerns and complaints?

March 21, 2021

CSC 5741 (2021/22) L01 - 109

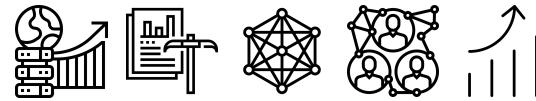
## Bibliography

- [1] 2021/22 CSC 5741 Syllabus

[https://drive.google.com/file/d/1t8xXwtksD7ITXZA\\_jRcmCTktj04tXwEi](https://drive.google.com/file/d/1t8xXwtksD7ITXZA_jRcmCTktj04tXwEi)



✉ [csc5741@unza.zm](mailto:csc5741@unza.zm)  
☞ <http://bit.ly/39HTdTK>  
▷ <http://bit.ly/2kK2ZkA>



# CSC 5741 (2021/22) 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>