

# **CSC 5741 (2020/21)**

# **Data Mining and Warehousing**

## **Lecture 1: Administrivia, Course Overview and Introduction**

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

# **Lecture Series Outline**

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

# Lecture Series Outline

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

# Personal Introductions (1/5)

## 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



Chrispin Hamooya



Akakandewla Akakandelwa



Abel M'kulama



Benson Njobvu



Edward Mwalmu



Felesia Mulaazi



Lighton Phiri  
Lecturer, Researcher



Phyela S. Mbewe  
Lecturer, Researcher



Thabiso Mwiinga  
Lecturer, Researcher

Dean, School of Education



Bentley Nkhata (PhD)

School of Education

Great East Road Campus:

PO Box 32379

Lusaka

Contacts

Tel: +26 021 125 1381

Fax: +26 021 125 1381

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

# Personal Introductions (2/5)

## Lighton Phiri

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

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

### About Me

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

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

I proposed [streamlined technology-driven orchestration](#) (Thesis).

I was affiliated with the [Digital Libraries Laboratory](#), the [Centre in ICT for Development](#) and the [HPI-CS4A Research School](#)

Prior to that, I was an [MSc student](#) at UCT.

I explored [simple digital library architectures](#) (Dissertation).

I was affiliated with the [Digital Libraries Laboratory](#).

Formerly, I worked with Telco ETL processing nodes at [Airtel](#).

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

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

[Publications](#) | [Google Scholar](#) | [Projects](#) | [Students](#)

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

# Personal Introductions (3/5)

## Lighton Phiri

[Department of Computer Science, University of Cape Town](#)

Rondebosch 7701, Cape Town, South Africa

Email: [X@Y](mailto:X@Y) (X=lphiri, Y=cs.uct.ac.za)

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

### About Me

I was a PhD student in the [Department of Computer Science](#) at the [University of Cape Town](#).

I explored [Technology-driven Orchestration](#) and was supervised by [Hussein Suleman](#) and [Christoph Meinel](#).

I was affiliated with [Digital Libraries Laboratory](#), [Centre for ICT for Development](#) and [HPI-CS4A](#).

Prior to that, I was an MSc student in the [Digital Libraries Laboratory](#). I investigated [Simple Digital Libraries](#) and was supervised by [Hussein Suleman](#).

Formerly, I worked with Telco ETL processing nodes at [Airtel](#).

Earlier, I studied Software Engineering at the [University of Zambia](#).

### Additional Information

**Teaching:** *CSC1015F: Computer Science 1015 – [2016] [\[notes\]](#) [\[code\]](#) [\[slides\]](#) [\[videos\]](#)*

*CSC1017F: Python for Engineers – [2015] [\[notes\]](#) [\[code\]](#) [\[slides\]](#) [\[videos\]](#)*

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

# Personal Introductions (4/5)

## Research

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

### Data Mining

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

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



Lighton Phiri  
Academic Staff



Robert M'sendo  
Masters Student  
MSc Computer Science

### Digital Libraries

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

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



Lighton Phiri  
Academic Staff



Robert M'sendo  
Masters Student  
MSc Computer Science



Mathews Mbewe  
Undergraduate Student  
BA Library and Information Science



Mathews Mwewa  
Undergraduate Student  
BA Library and Information Science

## DataLab People

### Academic Staff



Lighton Phiri

### Masters Student



Adrian Chisale  
MA Library and Information Science



Cecelia Kasonde  
MA Library and Information Science



Matildah Muchinga  
MA Library and Information Science



Violet Chama  
MA Library and Information Science

### 2020



Dokowe Tembo  
MA Library and Information Science

### 2019

<http://datalab.unza.zm>

# **Personal Introductions (5/5)**

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

# CSC 5741 Learning Outcomes

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

# CSC 5741 Desired Outcome

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

# **Course Structure (1/9)**

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

# Course Structure (2/9)

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



**THE UNIVERSITY OF ZAMBIA  
COMMUNICATION AND MARKETING DEPARTMENT**

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

---

**PRESS STATEMENT**

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

**09<sup>th</sup> February 2021** - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09<sup>th</sup> February 2021, has resolved to review the University Senate resolution of 27<sup>th</sup> 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 8<sup>th</sup> February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

**B. All Second Year S** [\*\*https://www.unza.zm/node/1794\*\*](https://www.unza.zm/node/1794)

# Course Structure (3/9)

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



## THE UNIVERSITY OF ZAMBIA COMMUNICATION AND MARKETING DEPARTMENT

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

### PRESS STATEMENT

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

**09<sup>th</sup> February 2021** - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09<sup>th</sup> February 2021, has resolved to review the University Senate resolution of 27<sup>th</sup> 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 8<sup>th</sup> February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S <https://www.unza.zm/node/1794>

# Course Structure (4/9)

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



## THE UNIVERSITY OF ZAMBIA COMMUNICATION AND MARKETING DEPARTMENT

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

### PRESS STATEMENT

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

**09<sup>th</sup> February 2021** - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09<sup>th</sup> February 2021, has resolved to review the University Senate resolution of 27<sup>th</sup> 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 8<sup>th</sup> February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

B. All Second Year S <https://www.unza.zm/node/1794>

# Course Structure (5/9)

- **Seminars**

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



**THE UNIVERSITY OF ZAMBIA  
COMMUNICATION AND MARKETING DEPARTMENT**

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

**PRESS STATEMENT**

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

**09<sup>th</sup> February 2021** - The University of Zambia (UNZA) Senate, at its special meeting held on Tuesday, 09<sup>th</sup> February 2021, has resolved to review the University Senate resolution of 27<sup>th</sup> 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 8<sup>th</sup> February 2021 will continue with blended learning for a period of three months starting February 2021 to April 2021.

**B. All Second Year S** [\*\*https://www.unza.zm/node/1794\*\*](https://www.unza.zm/node/1794)

# Course Structure (6/9)

MON	TUE	WED	THU	FRI	SAT	SUN
15 GMT+02	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,					
19:00						
20:00						
21:00						

# Course Structure (7/9)

- Course Resources

- All course resources will be made available on Astria.

The screenshot shows the Astria Learning Management System (LMS) interface for the course LIS 5310. The left sidebar contains navigation links: Home, Announcements (with 8 notifications), Modules, Assignments, Discussions, Grades, People, Pages, Files, Outcomes (with 2 notifications), Quizzes, Syllabus (with 2 notifications), Conferences, and Settings. The main content area displays the course title "Information Systems and Technologies in Information ..." and the University of Zambia logo. To the right, there are several action buttons: Choose Home Page, View Course Stream, Course Setup Checklist, New Announcement, Student View, and a To Do list with three items: Grade Discussion 1 - Information Management (due Mar 26 at 11:59pm), Grade Lab Exercise on Networks (due Mar 26 at 11:59pm), and Grade Lab Exercise on Databases (due Mar 2). At the bottom right, the URL <https://elearning.unza.zm> is displayed.

# Course Structure (8/9)

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

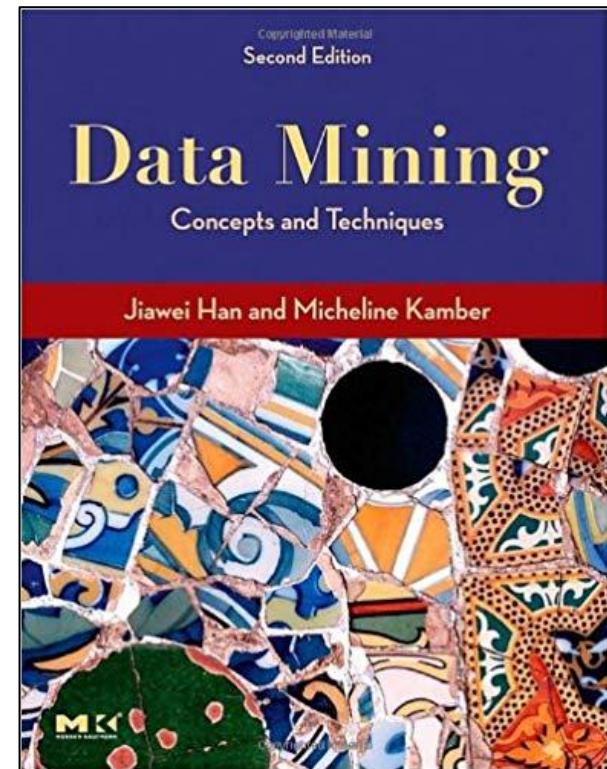
The screenshot shows the course structure for LIS 5310 on the Astria Learning Management System. The left sidebar contains a navigation menu with icons for Account, Dashboard, Courses, Calendar, Inbox, and eLibrary. The main content area shows the course title 'LIS 5310' and the 'Modules' section. Under 'INTRODUCTION', there are two items: 'Acknowledgement and Copyright' and 'Need Help?'. Under 'Module 1 - Introduction', there is one item: 'notes-unza19-lis5310-lecture-01-handout.pdf'. At the bottom right, a blue bar displays the URL <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)

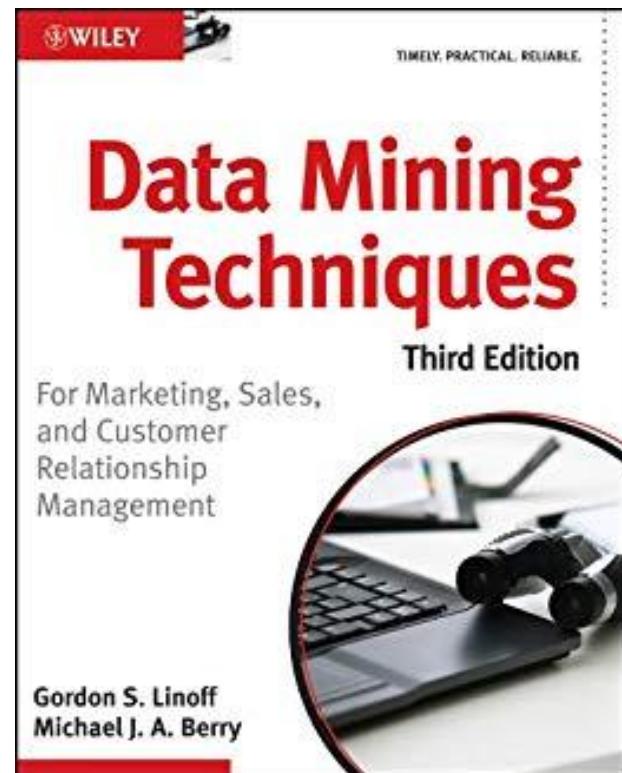
# Prescribed & Recommended Textbooks (1/4)

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



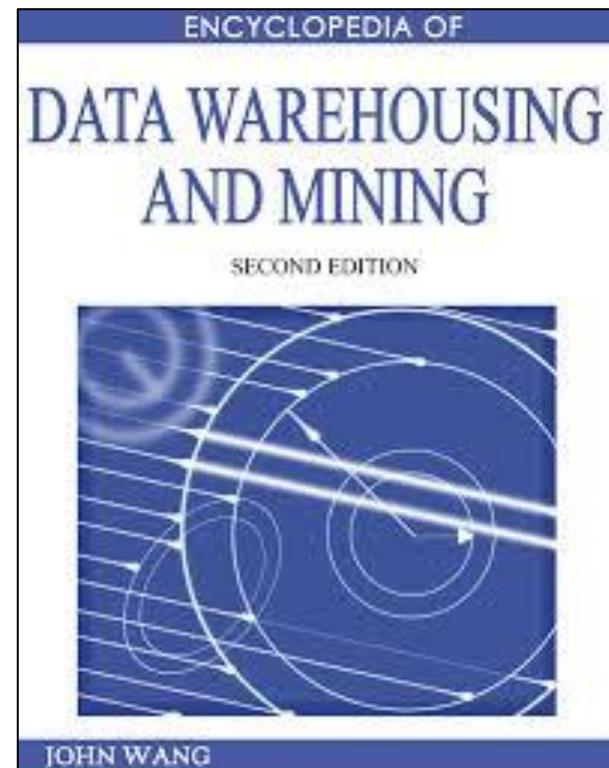
# Prescribed & Recommended Textbooks (2/4)

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



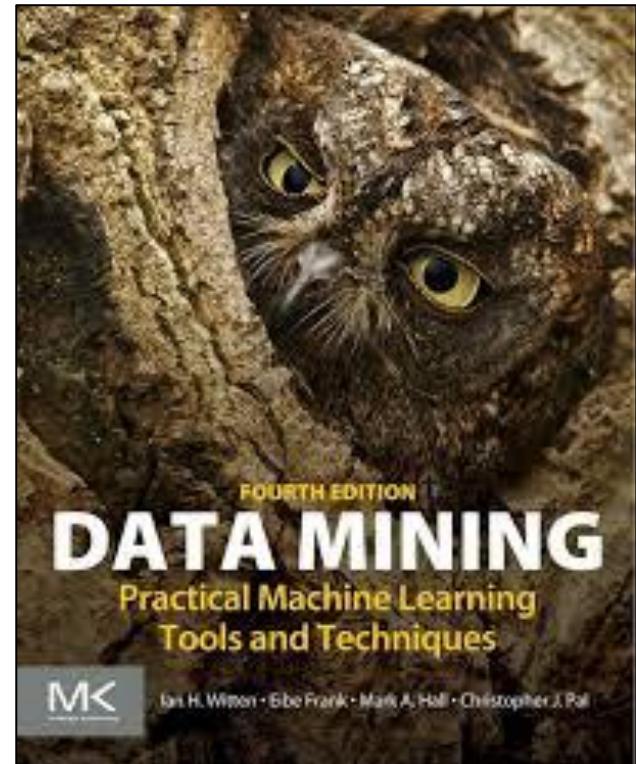
# Prescribed & Recommended Textbooks (3/4)

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



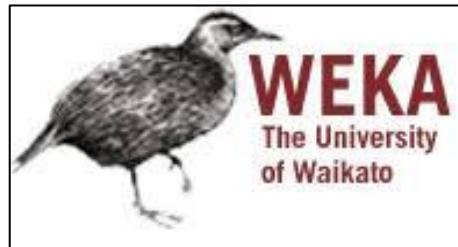
# Prescribed & Recommended Textbooks (4/4)

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



# Tools and Services (1/6)

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



# Tools and Services (2/6)

- **scikit-learn**

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

The screenshot shows the official scikit-learn website. At the top, there's a navigation bar with links for Home, Installation, Documentation, Examples, Google Custom Search, and a search icon. Below the navigation is a banner featuring a grid of 25 small plots illustrating various machine learning concepts like clustering and classification. To the right of the banner, the text "scikit-learn Machine Learning in Python" is displayed, followed by a bulleted list of features: Simple and efficient tools for data mining and data analysis; Accessible to everybody, and reusable in various contexts; Built on NumPy, SciPy, and matplotlib; Open source, commercially usable - BSD license.

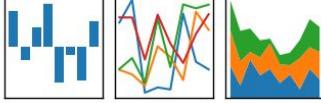
Classification	Regression	Clustering
Identifying to which category an object belongs to. <b>Applications:</b> Spam detection, Image recognition. <b>Algorithms:</b> SVM, nearest neighbors, random forest, ... — Examples	Predicting a continuous-valued attribute associated with an object. <b>Applications:</b> Drug response, Stock prices. <b>Algorithms:</b> SVR, ridge regression, Lasso, ... — Examples	Automatic grouping of similar objects into sets. <b>Applications:</b> Customer segmentation, Grouping experiment outcomes <b>Algorithms:</b> k-Means, spectral clustering, mean-shift, ... — Examples
Dimensionality reduction	Model selection	Preprocessing
Reducing the number of random variables to consider. <b>Applications:</b> Visualization, Increased efficiency <b>Algorithms:</b> PCA, feature selection, non-negative matrix factorization. — Examples	Comparing, validating and choosing parameters and models. <b>Goal:</b> Improved accuracy via parameter tuning <b>Modules:</b> grid search, cross validation, metrics. — Examples	Feature extraction and normalization. <b>Application:</b> Transforming input data such as text for use with machine learning algorithms. <b>Modules:</b> preprocessing, feature extraction. — Examples

<https://scikit-learn.org>

# Tools and Services (3/6)

- Pandas
  - Python data analysis library

**pandas**

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$


[home](#) // [about](#) // [get pandas](#) // [documentation](#) // [community](#) // [talks](#) // [donate](#)

**Python Data Analysis Library**

*pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](#) programming language.

*pandas* is a [NumFOCUS](#) sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to [donate](#) to the project.

A Fiscally Sponsored Project of

**NUMFOCUS**  
OPEN CODE = BETTER SCIENCE

v0.23.4 Final (August 3, 2018)

This is a minor bug-fix release in the 0.23.x series and includes some regression fixes, bug fixes, and performance improvements. We recommend that all users upgrade to this version.

The release can be installed with conda from conda-forge or at

**<https://pandas.pydata.org>**

**VERSIONS**

<b>Release</b>
0.24.2 - March 2019
<a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>

<b>Development</b>
0.25.0 - April 2019
<a href="#">github</a> // <a href="#">docs</a>

<b>Previous Releases</b>
0.24.1 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.24.0 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.23.4 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.23.3 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.23.2 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.23.1 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.23.0 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.22.0 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.21.1 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.21.0 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>
0.20.3 - <a href="#">download</a> // <a href="#">docs</a> // <a href="#">pdf</a>

# Tools and Services (4/6)

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

The 2019 SciPy John Hunter Excellence in Plotting Contest is accepting submissions until June 8th!

# matplotlib

Version 3.0.3

[home](#) | [examples](#) | [tutorials](#) | [API](#) | [docs](#) »

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.



Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the [sample plots](#) and [thumbnail gallery](#).

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

## Installation

Visit the [Matplotlib installation instructions](#).

## Documentation

This is the documentation for Matplotlib version 3.0.3.

<https://matplotlib.org>

# Tools and Services (5/6)

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

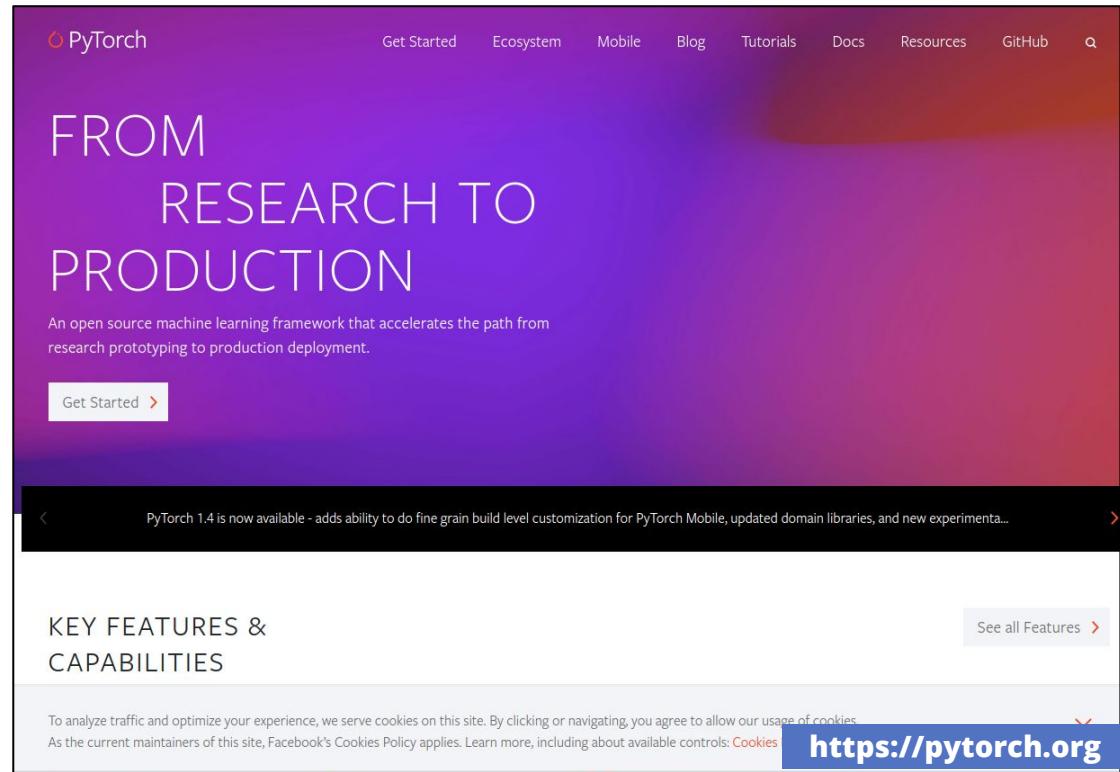
The screenshot shows the Google Colab interface. On the left, there's a sidebar titled "Table of contents" with links to "Getting started", "Data science", "Machine learning", "More Resources", and "Machine Learning Examples". The main area displays a section titled "What is Colaboratory?" which includes a bulleted list: "Zero configuration required", "Free access to GPUs", and "Easy sharing". Below this, a "Getting started" section is shown with a code cell containing Python code to calculate seconds in a day:

```
[ ] 1 seconds_in_a_day = 24 * 60 * 60  
2 seconds_in_a_day  
86400
```

The URL <https://colab.research.google.com> is visible at the bottom of the browser window.

# Tools and Services (6/6)

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



# **Course Grading (1/2)**

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

# **Course Grading (2/2)**

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

# Course Grading—Paper Readings

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		

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

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

# Course Grading—Seminars

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

**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 participants	✓ 0/1	✓ 1/1	✓ 1/1	✓ 1/1	✓ 0/1	✓ 1/1	✓ 1/1
Enter your name	●	●	●	●	●	●	●
Lillian Mzyece						✓	
Andreya Kumwenda				✓			
Friday C. Chazanga		✓					
Soft Mulizwa	✓						
Francis Chulu						✓	

# 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

	(c) NDLTD: Cluster analysis of TD subjects	1 (d) NDLTD: Cluster analysis of ETD by region	2 (a) NETD: Cluster analysis by publication date	2 (b) NETD: Classification of universities based on ETD output	2 (c) ana by
4 participants	✓ 0/1	✓ 0/1	✓ 0/1	✓ 1/1	
Enter your name	●	●	●	●	
Inonge Lamaswala					
Kaumba Mutende					
David Mulenga				✓	
Tasha Shamane					

# Course Grading—MiniProject (2/3)

#	Mask	Implementation					Technical Report							Presentation					Presentation Total	Grand Total		
		Data	Code/Scripts	Novelty	Relevance	Demo	Implementation Total	Abstract	Aim/Problem	Implementation	Dataset	Experiment	Results	Quality	Report Total	Content	Quality	Visualisations	Comprehensive	Q/A		
1	Elastic Net	30	30	15	10	10	95	6	6	8	10	18	20	18	86	18	20	20	18	19	95	18.28
2	Linear SVC					10	10	4	5	5	10	0	15	15	54	10	20	20	10	18	78	8.24
3	SGD Classifier					10	10	4	3	5	6	15	15	15	63	15	18	20	15	18	86	9.28
4	Kernal Approximation					10	10								0						0	0.8
5	Lasso						0								0						0	0
6	Naive Bayes					10	10	6	5	5	10	15	10	10	61	10	20	20	15	20	85	9.08
7	Ensemble Classifiers	30	30	5	10	10	85	4	6	0	5	0	5	5	25	10	10	15	10	15	60	11.2
8	Spectral Clustering	30	30	10	10	10	90	0	4	10	10	17	17	15	82	18	20	1				

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

# Course Grading—MiniProject (3/3)

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

## CSC 5741: Mini Practical Projects

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

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

# Course Grading—Class Participation

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

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

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

# Course Grading—Class Theory Tests

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

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

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

# Course Grading—Final Examination

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

THE UNIVERSITY OF ZAMBIA  
SCHOOL OF NATURAL SCIENCES

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

---

**MARKS: 100**

**TIME: THREE (3) HOURS**

**INSTRUCTIONS:**

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

## Question 1

It was recently reported<sup>1</sup> that the Government of The Republic of Zambia (GRZ) is working

# Course Grading—CA (1/2)

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

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

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

# Course Grading—CA (2/2)

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

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

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

# Course Grading—CA (2/2)

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

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

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

# Course Grading—CA (2/2)

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

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

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

# Course Grading Thresholds

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

# Course Management (1/2)

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

# Course Management (2/2)

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

The screenshot shows a web-based course management system interface. At the top, there is a navigation bar with a logo, a search bar labeled "Search for messages", and various administrative icons. Below the navigation bar, a sidebar on the left lists "Groups", "My groups", "Home", "My discussions", "Starred", "Favorites" (with a note to click a star icon to add it), and "Recently viewed" (listing CSC 5741, LIS 4014, LIS 5310, and ICT 1110). The main content area displays a group page for "CSC 5741: Data Mining and Warehousing". The page header includes a "NEW TOPIC" button, a "C" icon, "Mark all as read", "Actions", "Filters", and user management buttons. A yellow banner at the top of the content area states: "In May of 2019, we'll be merging and deprecating some of our settings to make group management easier." with a "Learn more" link. The group description for CSC 5741 mentions it is a graduate-level course in Computer Science addressing data warehousing and mining concepts. It lists several bullet points about the course's objectives and provides links for managing the group, members, and about information. At the bottom of the page, there are buttons for "Edit welcome message" and "Clear welcome message", and a reminder message: "Reminder: We Have Class at 17H30 T". The URL of the page is displayed as a blue link at the bottom right: <https://groups.google.com/a/unza.zm/d/forum/csc5741>.

# **Academic Dishonesty**

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

# **Q & A Session**

- Comments, concerns and complaints?**

# Lecture Series Outline

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

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

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

AI 'outperforms' doctors diagnosing breast cancer

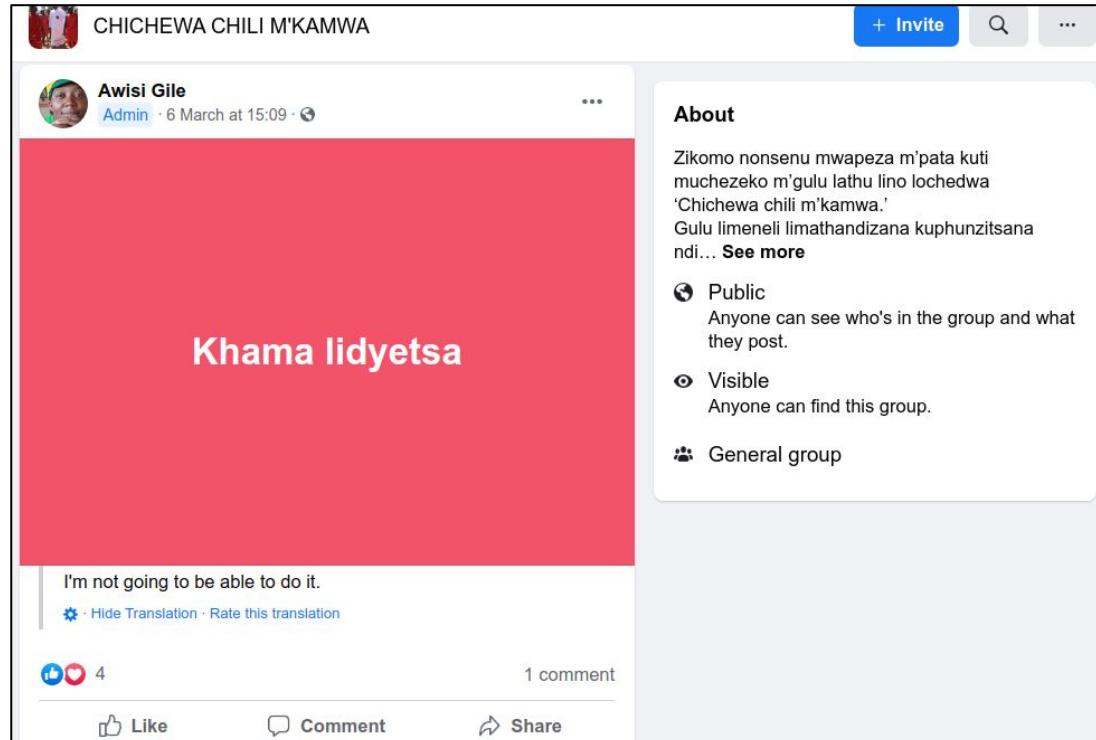
Fergus Walsh  
Medical correspondent  
[@BBCFergusWalsh](#)

2 January 2020



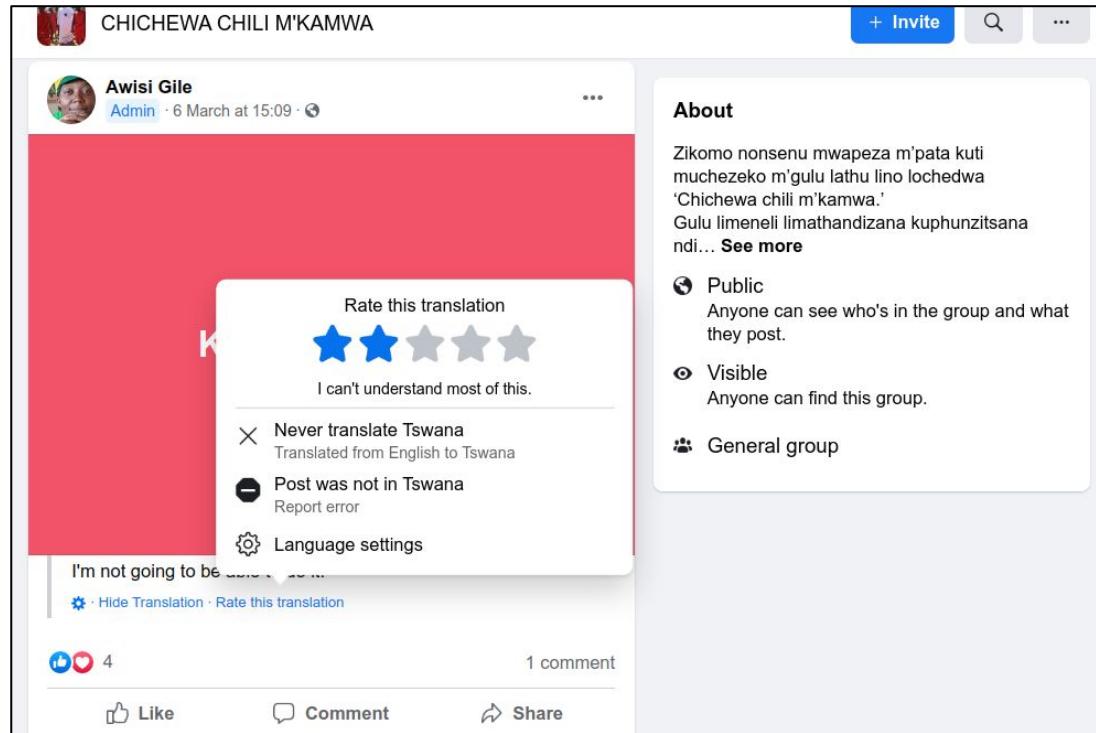
# Contextualising Data Mining & Warehousing: Everyday Examples (2/5)

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



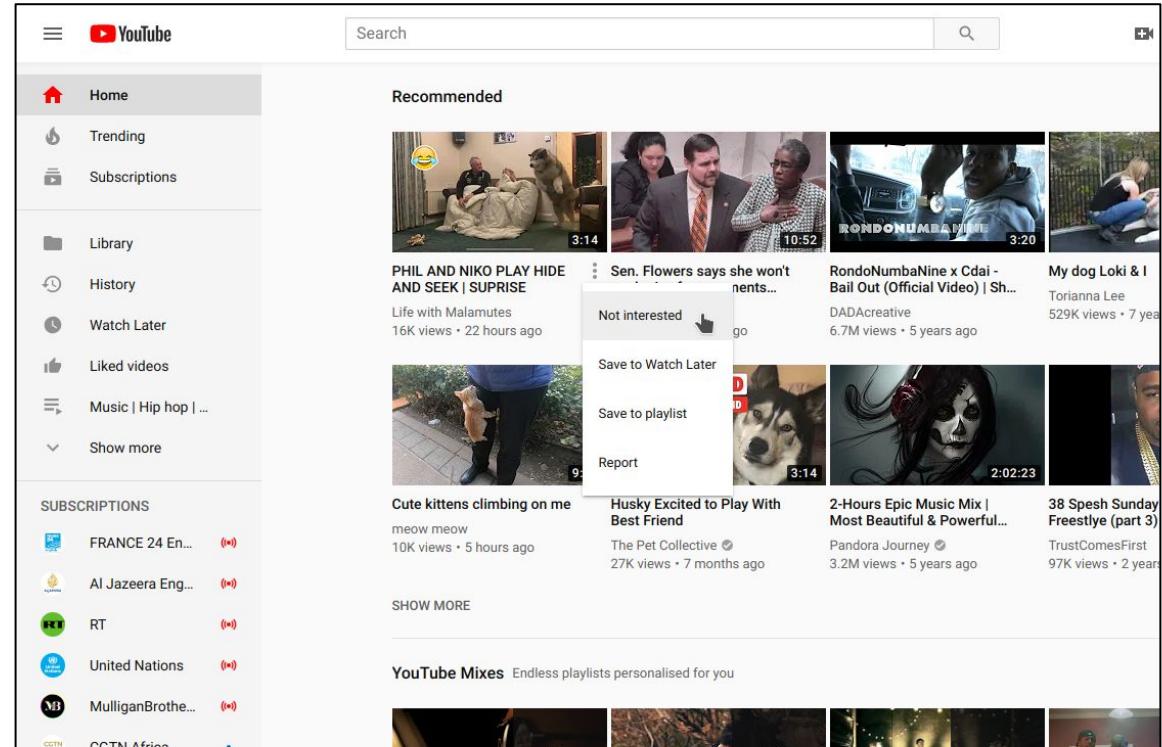
# Contextualising Data Mining & Warehousing: Everyday Examples (2/5)

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



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

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



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

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

The screenshot shows a video editing interface. On the left, a sidebar lists 'Channel content' with 'Course Resources (1/4)' and 'Your video'. Below these are buttons for 'Details' (selected), 'Analytics', 'Editor', 'Comments', and 'Subtitles'. The main area displays a video thumbnail titled 'Course Resources (1/4)' showing a presentation slide with logos for Lucidchart, Project Libre, and git. The video duration is 2:09:46. To the right, under 'Video details', is a text box containing information about the recording setup and links to various tools used:

Video recording/screencasting was done using ffmpeg [3] and SmartRecorder [4] on a OnePlus 3T connected with a Boya BY-M1 microphone [5].

[1] <https://moodle.unza.zm/course/edit.php?id=2413>  
[2] <http://www.unza.zm>  
[3] <https://www.ffmpeg.org>  
[4] <https://play.google.com/store/apps/details?id=com.andrwq.recorder&hl=en>  
[5] <http://www.boyamic.com/lavaliermicrophones/BY-M1.html>

**Thumbnail**  
Select or upload a picture that shows what's in your video. A good thumbnail stands out and draws viewers' attention.  
[Learn more](#)

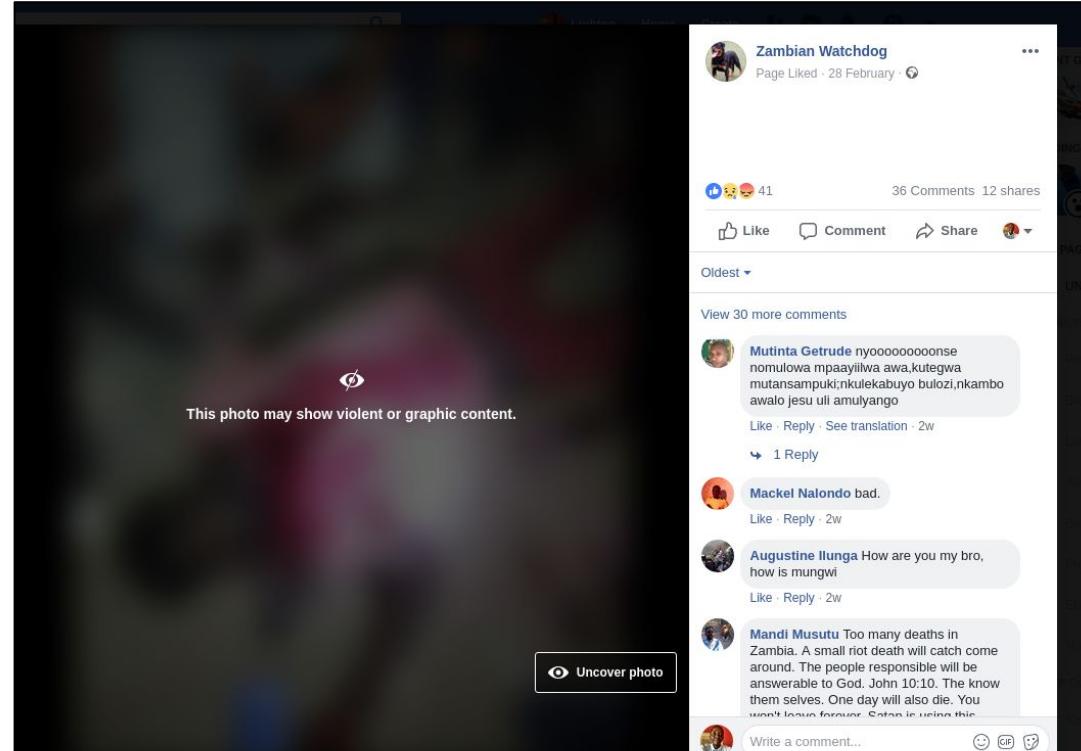
Upload thumbnail

**Playlists**  
Add your video to one or more playlists. Playlists can help viewers to discover your content faster. [Learn more](#)

Playlists

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

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



# **Contextualising Data Mining & Warehousing: Postgraduate Projects**

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

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

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

```
<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/
  instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  instance">
    <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, School of
    </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>
```

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

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

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

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

```
1###0###1###Moodle at University of Zambia###moodle#####0###site###1###3###0###0###0###0###0###
91769##2016-11-22 11:33:45
2###25###4940006###VMM 7802 Health Economics, Policy, Monitoring and Evaluation###VMM 7802###V
#0########
#1479915665###1525439465###0###1###0###1536578236##2016-11-23 17:41:05
3###25###4940009###VMM 7501 Principles of Epidemiology and Biostatistics###VMD 7501###VMD 7501
#####
#1479916418###1530255329###0###1###0###1535391769##2016-11-23 17:53:38
4###25###4940013###Socio-Anthropology###VMM7412###VMM7412#####1###topics###1###2###1479852000
#1###0###1535391769##2016-11-23 17:53:39
6###25###4940001###Applied Environmental Health, Water and Sanitation###VMM 7312###VMM 7312###
#####
#1479916423###1480061951###0###1###0###1535391769##2016-11-23 17:53:43
7###25###4940002###Applied Food Microbiology and Nutritional Toxicology###VMM 7120###VMM 7120#
#####
#1479916432###1480060975###0###1###0###1535391769##2016-11-23 17:53:52
9###25###4940010###VMM 8901 Research Methodology###VMM 8901###VMM 8901#####1###topics###1###5
5439517###0###0###0###1535391769##2016-11-23 17:57:44
11###25###4940003###Ethics in Food Safety Practice###VMM 8911###VMM 8911#####1###topics###1###
492093217###0###1###0###1535391769##2016-11-23 17:59:57
14###25###4940005###Food Safety Managemnt###VMM 7501###VMM 7501#####1###topics###1###4###1543
###0###1###0###1535391769##2016-11-23 18:04:53
16###25###4940012###VMM 8201 Risk Analysis and Surveillance###VMM 8201###VMM 8201#####1###top
917386###1525439576###0###0###0###153550221##2016-11-23 18:09:46
17###220###5120004###VMD 6800 Veterinary Public Health###VMD 6800###VMD 6801#####1###topics##
###1524427683###0###0###0###1535391769##2016-11-23 21:19:13
19###25###4940007###Health Promotion, Education and Communication###VMM8711###VMM8711#####1###
79970494###1480062724###0###1###0###1535391769##2016-11-24 08:54:54
20###25###4940014###Zoonotic Diseases and Infections###VMM 7610#####1###topics###1###2###1
890###0###1###0###1535391769##2016-11-24 10:41:59
\1###25###4940011###Research Paper###VMM 8900###VMM 8900##<p><br></p><p><strong>
\</strong><span lang="EN">The RESEARCH PROJECT is an important component of these programme and
ded the degree of Master of Science in One Health Food Safety. The project is not only importa
erves as the final test of students' capability to work independently and think critically. It
he sense that researchers attempt to build on and improve upon previous work' <i>(Johnson 199
p;that will result in writing a research paper that will be evaluated and graded by your super
```

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

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



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

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

```
enter the examination period at: C:\R\R-3.6.1\
```

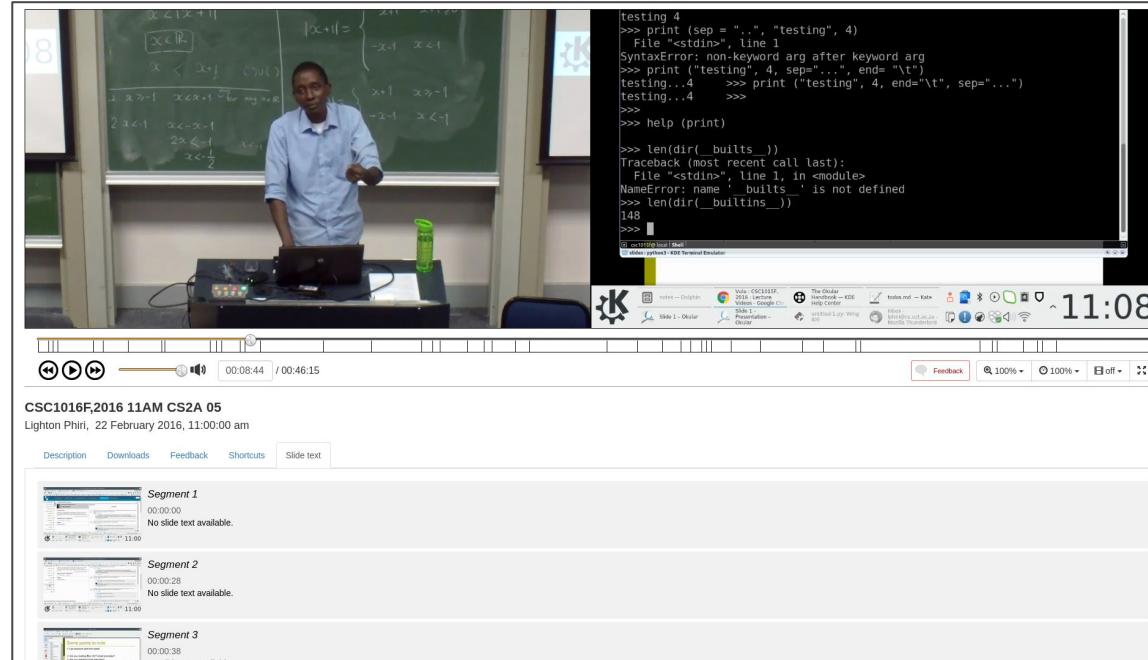
```
> colnames(dataset_cfs0405_crop)
[1] "PROV"   "DIST"   "CONST"  "WARD"   "REGION" "CSA"
[7] "SEA"    "HHNUM"  "CROP"   "ID009"  "S1AFIELD" "S1ACF01"
[13] "S1ACF02" "S1ACF03" "S1ACF04" "S1ACF05" "S1ACF06" "S1ACF07"
[19] "S1ACF08" "S1ACF09" "S1ACF10" "S1ACF11" "S1ACF12" "S1ACF13"
[25] "S1ACF14" "S1ACF15"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

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

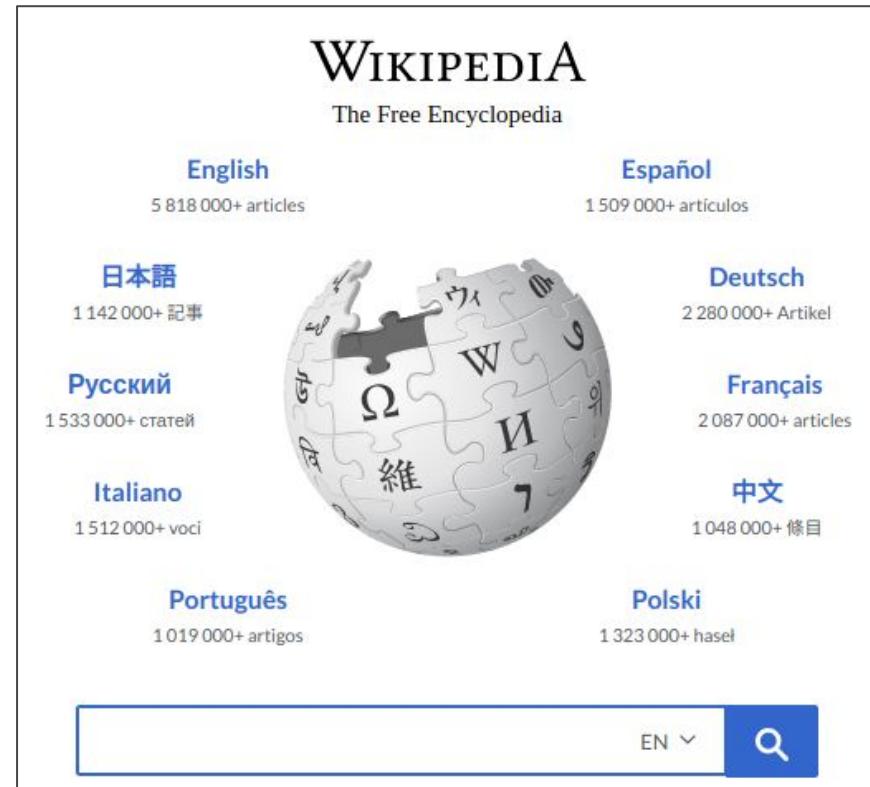
# Contextualising Data Mining & Warehousing: Some Ongoing Projects (6/8)

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



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

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



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

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



# **Contextualising Data Mining & Warehousing: Zambia Centric Projects**

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

# Contextualising Data Mining & Warehousing: Endless Possibilities

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

Break Into AI: Building a Career in Machine Learning with Andrew Ng

December 4, 2018

The screenshot shows a video player interface. At the top, it says "Break Into AI: Building a Career in Machine Learning with Andrew Ng" and "December 4, 2018". Below that is a thumbnail image of a person's head. To the right of the thumbnail are three icons: a blue square with a white 'E', a white cloud-like shape, and a vertical ellipsis. A table of contents box is overlaid on the video player. The box has a header "TABLE OF CONTENTS" and "Quality: High" with a close button. It lists ten items with their titles and durations:

1 Break Into AI: A Q&A with Andrew Ng on Building a...	00:27	2 ACM Highlights	01:21
3 "Housekeeping"	01:55	4 Talk Back	03:08
5 Welcome	03:36	6 Introduction	04:50
7 T-Shaped Individuals	06:32	8 Dirty Work	11:16
9 Lifelong Learning	12:53	10 The Best Opportunities: Outside the Software...	15:16
11 Q&A	16:29	12 Graduate Programs	17:44

At the bottom of the table of contents box is a yellow bar with the text "▶ PLAY VIDEO".

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

# 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 Squ

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

# 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 1,314 views 13

[Facebook](#) [Twitter](#) [Google+](#) [Pinterest](#) [LinkedIn](#) [Email](#) [Print](#)



Government Spokesperson, Dora Siliya, who is also Information and Broadcasting Minister, speaking when she featured on the local Breeze FM Radio station

Minister of Information and Broadcasting Services Dora Siliya says the









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

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



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

PICTURE: BUTTYSON KANDIMBA

## ZRA drones land on 7 trucks

### From page 1

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

has also confiscated 26 heavy-duty trucks and earth-moving machinery that were suspiciously imported through the misapplication of the value-added tax (VAT) deferment scheme.  
"ZRA will not relent in impounding

miscalculations, under-declarations, and under-valuations is a serious offence and offenders will be dealt with in accordance with the law," he said.  
Mr Sikalinda urged all citizens to always pay taxes directly to ZRA

in 26 trucks being impounded and will continue till all the trucks that have not paid the taxes pay what is due to the government," Mr Sikalinda said.  
Mr Sikalinda said so far, ZRA has published about 1,000 trucks in the

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

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

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



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

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



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

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



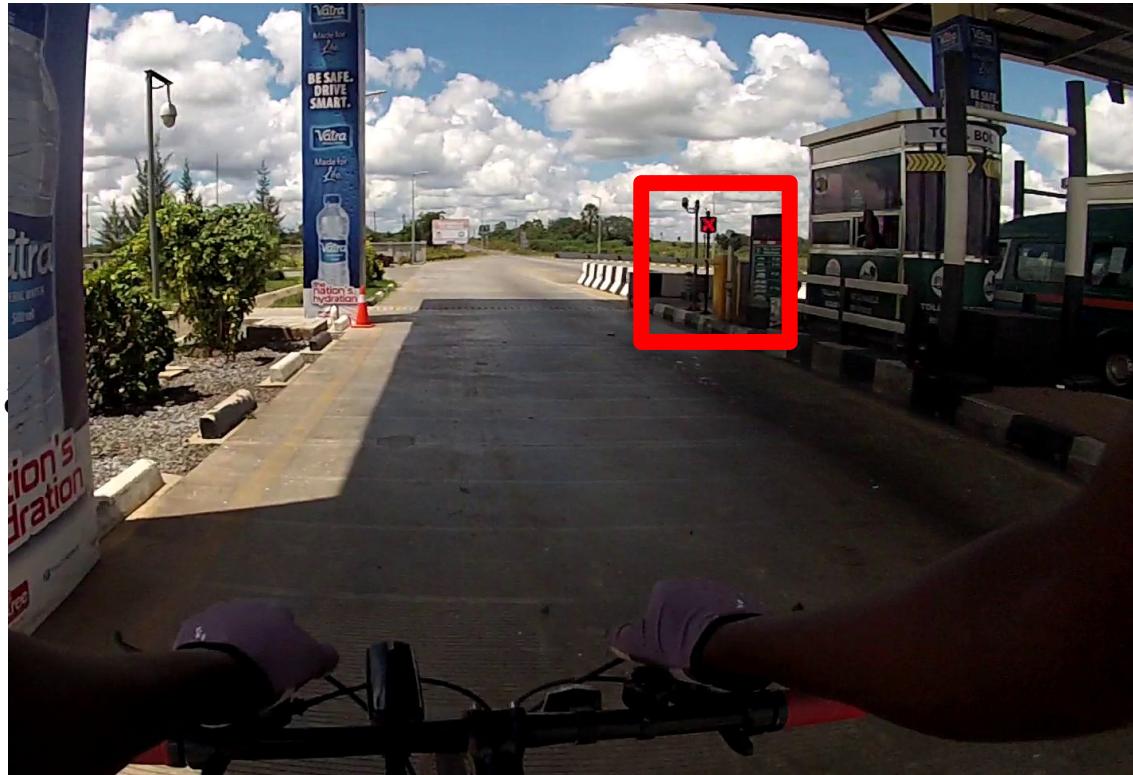
# Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

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



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

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



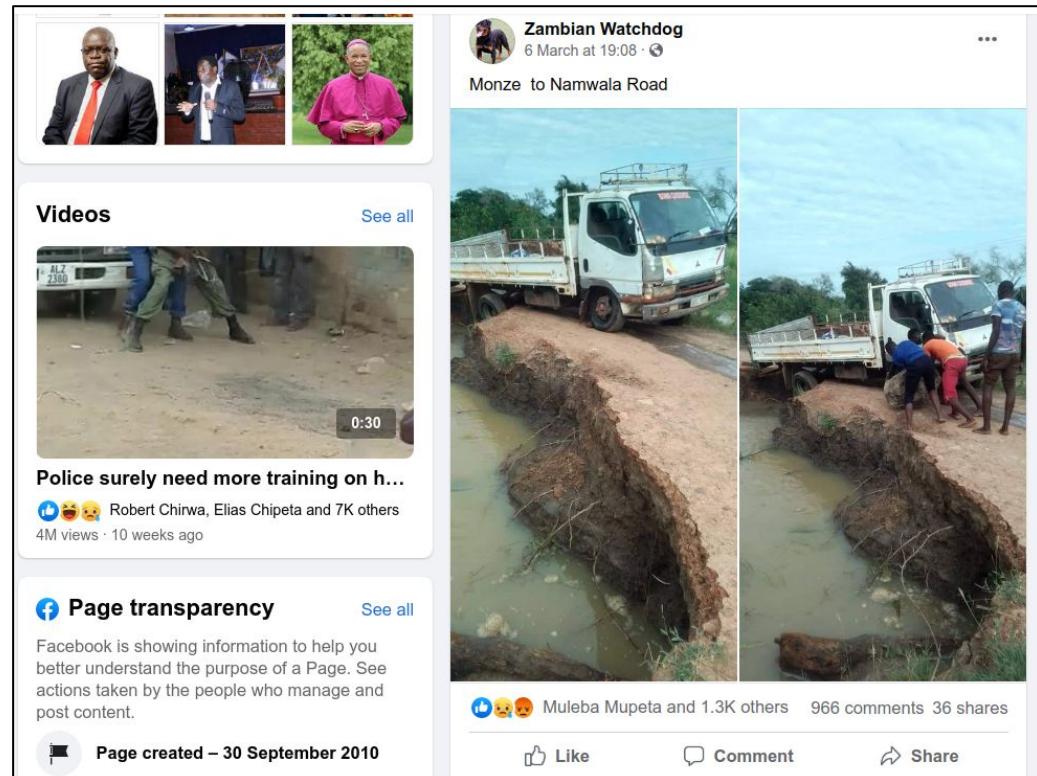
# Contextualising Data Mining & Warehousing: Curiosity vs Impact (5/6)

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



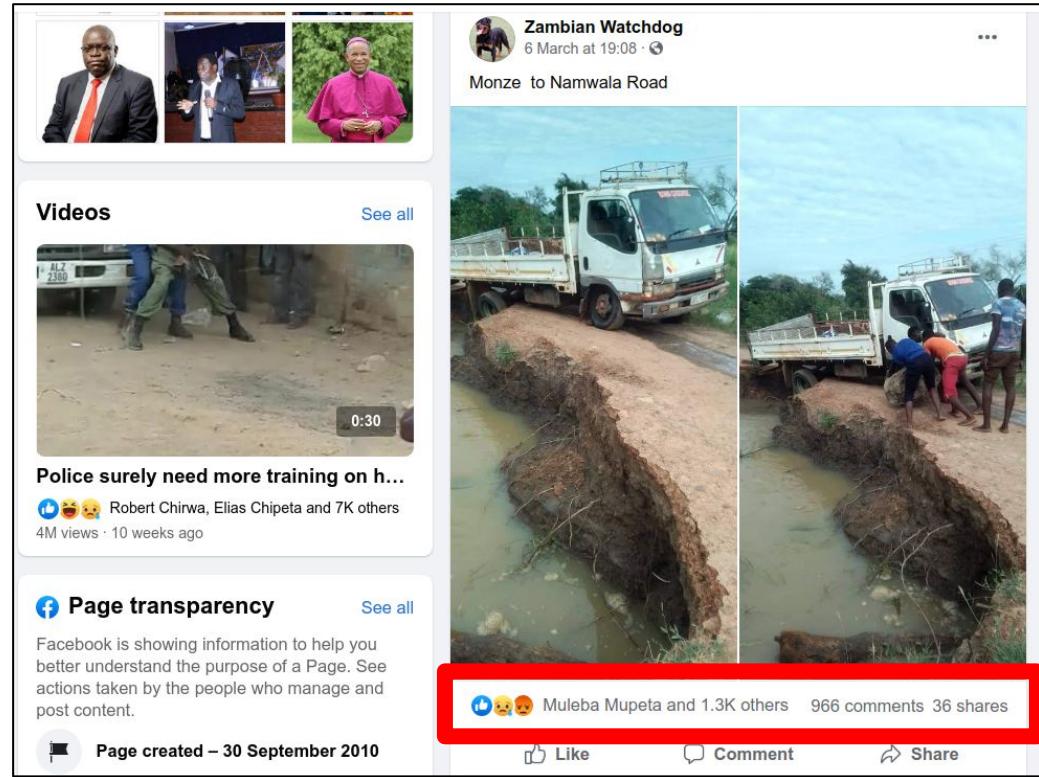
# Contextualising Data Mining & Warehousing: Curiosity vs Impact (6/6)

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



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

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



# Data is Key to Data Mining

State House Press Office - Zambia  
28 October at 10:41 · 52 shares

PHOTO FOCUS. MEMORIAL SERVICE OF ZAMBIA'S LATE PRESIDENT SATA

Memorial service of Zambia's late 5th President, His Excellency, Mr Michael Chilufya Sata taking place at the embassy park in Lusaka. In attendance are President of the Republic of Zambia, His Excellency Dr. Edgar Chagwa Lungu and the 1st Lady Mrs Esther Lungu... [See more](#)

+2

471 105 comments 52 shares

Charli... It always day a Like ↗ 1 Weluv I rem truly maj... propo again more Like ↗ 1 Jackie May restin Like ↗ 2 Sam... Great Like Reply 3 w ↗ 3 replies Peter Nkomba Snr I remember Mr Michael action,a hard working

### Course Resources (1/4)

- Software
  - Dia
  - Git
  - ProjectLibre
- Software applications are just tools
  - You are allowed to use alternatives to the tools that will be used in the course

March 1, 2021 ICT 3020 (2020/21) L01 - 7

**Lucidchart**  
**Project Libre™**  

Lighton Phiri

An Introduction to MLOps  
AIEngineering  
3.2K views • Streamed 2 days ago  
New

SACRIFICE - Powerful Vocal Music Mix | Epic Cinematic...  
Premium Music HQ  
1.1M views • 1 year ago

Mind Of  
BLUME  
1.3M views • 2 years ago

Insam 6 Million Dollar Airport Scammers Prank Call: Part 2 ...  
The Hoax Hotel  
200K views • 4 years ago

Resilience | Deep Chill Music Mix  
Fluidified  
454K views • 6 months ago

F A HERO - Epic Music Mix | Powerful...  
Music HQ  
1.5M views • 1 year ago

JUL BATTLE - Epic Battle Music Mix ...  
Music HQ  
1.5M views • 2 months ago

# **Introduction (1/2)**

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

# **Introduction (2/2)**

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

# **Theme #1: Data Mining and Data Pre-processing**

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

# **Theme #2: Data Warehousing**

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

# Theme #3: Classification

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

# **Theme #4: Associative Rule Mining**

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

# **Theme #5: Clustering Analysis**

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

# Closing CSC 5710 Remarks

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

# **Q & A Session**

- Comments, concerns and complaints?**

# Lecture Series Outline

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

# Readings and Paper Summaries (1/5)

The screenshot shows the Mendeley Desktop application window. The 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 area has tabs for My Library and All Documents, with All Documents selected. A sub-menu "Creating a National Electron..." is open. On the left, a sidebar lists Mendeley, Literature Search, My Library (with All Documents selected), Recently Added, Recently Read, and Favorites. The central pane displays a table of documents with columns for Authors, Title, Year, Published In, and Added. Three documents are listed:

Authors	Title	Year	Published In	Added
Willinsky, John	Open Journal Systems	2005	Library Hi...	06/08/18
Akakandelwa, Ak...	Author Collaboration and Productivit...	2009	African Jo...	06/08/18
Kulyambanino, C...	Faculty Productivity at The Universit...	2016		06/08/18

On the right, there are buttons for Details, Notes, and Contents. A search bar says "Type: Conference Proceedin". A large blue banner at the bottom right reads "Creating a National Africa Mendeley Desktop".

# Readings and Paper Summaries (2/5)

The screenshot shows the Google Scholar search interface. The search query 'data mining and machine learning' is entered in the search bar. The results page displays several academic papers with their titles, authors, publication details, and citation counts. The results are filtered by 'Articles' and sorted by relevance. The interface includes search filters for time range (Any time, Since 2019, etc.), sorting options (Sort by relevance, Sort by date), and checkboxes for including patents and citations. A 'Create alert' button is also present.

Google Scholar

data mining and machine learning

Articles About 2,040,000 results (0.12 sec)

Any time [CITATION] Data Mining: Practical machine learning tools and techniques [PDF] fue.edu.eg

Since 2019 IH Witten, E Frank, MA Hall, CJ Pal - 2016 - Morgan Kaufmann

Since 2018 ☆ 99 Cited by 34827 Related articles All 38 versions

Since 2015

Custom range... Distributed GraphLab: a framework for machine learning and data mining in the cloud [PDF] arxiv.org

Y Low, D Bickson, J Gonzalez, C Guestrin... - Proceedings of the ..., 2012 - dl.acm.org

While high-level data parallel frameworks, like MapReduce, simplify the design and implementation of large-scale data processing systems, they do not naturally or efficiently support many important data mining and machine learning algorithms and can lead to ...

☆ 99 Cited by 1554 Related articles All 29 versions

Sort by relevance

Sort by date

include patents

include citations

Create alert

Business data mining—a machine learning perspective

I Bose, RK Mahapatra - Information & management, 2001 - Elsevier

The objective of this paper is to inform the information systems (IS) manager and business analyst about the role of machine learning techniques in business data mining. Data mining is a fast growing application area in business. Machine learning techniques are used for ...

☆ 99 Cited by 367 Related articles All 5 versions

The WEKA data mining software: an update

M Hall, E Frank, G Holmes, B Pfahringer... - ACM SIGKDD ..., 2009 - dl.acm.org

<https://scholar.google.com>

# Readings and Paper Summaries (3/5)

The screenshot shows the AMiner search interface with a search bar containing "Whatever comes to your mind". The results are filtered by "Computer Science" and "All". The table lists 12 conferences with their full names, short names, and H5-Indexes.

Computer Science	Rank	Conference (Full Name)	Short Name	H5-Index
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
Software Engineering	7	Journal of the American Society for Information Science and Technology	JASIST	42.00
Database and Data Mining	8	IEEE International Conference on Data Engineering	ICDE	40.00
Theoretical Computer Science	9	ACM International Conference on Information and Knowledge Management	CIKM	38.00
Database and Data Mining	10	IEEE International Conference on DataMining	ICDM	33.00
Theoretical Computer Science	11	Journal of Web Semantics	J. Web Sem.	33.00
	12	Knowledge and Information Systems	KAIS	

<https://aminer.org>

# 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](#) [PLDI](#) [PODS](#) [S&P](#) [SIGCOMM](#) [SIGIR](#) [SIGMETRICS](#) [SOSP](#)

### 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. 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 1st place" (e.g. NeurIPS).

### AAAI (Artificial Intelligence)

2018	Memory-Augmented Monte Carlo Tree Search	Chenjun Xiao, University of Alberta; <a href="#">et al.</a>
2017	Label-Free Supervision of Neural Networks with Physics and Domain Knowledge	Russell Stewart & Stefano Ermon, Stanford University
2016	Bidirectional Search That Is Guaranteed to Meet in the Middle	Robert C. Holte, University of Alberta; <a href="#">et al.</a>
2015	From Non-Negative to General Operator Cost Partitioning	Florian Pommerening, University of Basel; <a href="#">et al.</a>
2014	Recovering from Selection Bias in Causal and Statistical Inference	Elias Bareinboim, University of California Los Angeles; <a href="#">et al.</a>
2013	HC-Search: Learning Heuristics and Cost Functions for Structured Prediction	Janardhan Rao Doppa, Oregon State University; <a href="#">et al.</a>
	SMILE: Shuffled Multiple-Instance Learning	Gary Doran & Soumya Ray, Case Western Reserve University
2012	Learning SVM Classifiers with Indefinite Kernels	Suicheng Gu & Yuhong Guo, Temple University
	Document Summarization Based on Data Reconstruction	Zhanying He, Zhejiang University; <a href="#">et al.</a>
2011	Dynamic Resource Allocation in Conservation Planning	Daniel Golovin, California Institute of Technology; <a href="#">et al.</a>
	Complexity of and Algorithms for Borda Manipulation	Jessica Davies, University of Toronto; <a href="#">et al.</a>
2010	How Incomplete Is Your Semantic Web Reasoner? Systematic Analysis of the Completeness of Query Ans...	Giorgos Stolios, Oxford University; <a href="#">et al.</a>
	A Novel Transition Based Encoding Scheme for Planning as Satisfiability	Ruoyun Huang, Washington University in St. Louis; <a href="#">et al.</a>
2008	How Good is Almost Perfect?	Malte Helmert & Gabriele Röger, Albert-Ludwigs-Universität Freiburg
	Optimal False-Name-Proof Voting Rules with Costly Voting	Liad Wagman & Vincent Conitzer, Duke University
2007	PLOW: A Collaborative Task Learning Agent	<a href="http://jeffhuang.com/best_paper_awards.html">http://jeffhuang.com/best_paper_awards.html</a>

# Readings and Paper Summaries (5/5)



Zambia ICT Journal    Announcements    Current    Archives    About ▾

The Zambia ICT Journal (ISSN: 2616-2156) 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

<http://ictjournal.icict.org.zm>

Make a Submission  
View Pending Submissions  
Author Guidelines  
Article Template  
Publications Fees  
Editorial Team  
Special Issue

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

**ACM DL DIGITAL LIBRARY** University of Cape Town [My Author Page](#) [My Binders](#) [SIGN OUT: Lighton Phiri](#)

[SEARCH](#)

**How to read a paper**

Full Text: [!\[\]\(ee3a2ba73b81fb637b53af969977ebce\_img.jpg\) PDF](#)

Author: [S. Keshav](#) [University of Waterloo](#)

Published in:

 · Newsletter  
ACM SIGCOMM Computer Communication Review archive

 2007 Article

 [Bibliometrics](#)

**Tools and Resources**

 [Buy this Article \(PRINT\)](#)

 [Recommend the ACM DL to your organization](#)

 TOC Service: [Email](#) [!\[\]\(503e17a186b997f9181b122945ac7956\_img.jpg\) RSS](#)

**Reading a computer science research paper**

Full Text: [!\[\]\(4784533c9bf23c1cf316483b2b57527a\_img.jpg\) PDF](#)

Author: [Philip 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

**Tools and Resources**

 [Buy this Article \(PRINT\)](#)

 [Recommend the ACM DL to your organization](#)

 TOC Service: [Email](#) [!\[\]\(816a19f6850b2fd6423b8af51f33e14b\_img.jpg\) RSS](#)

<https://dl.acm.org>

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

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

# On How to Read a Paper (3/5)

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

# On How to Read a Paper (4/5)

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

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

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

# **Lecture Series Outline**

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

# Public Talks

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

2nd Seminar in the Colloquium Series for 2019 Inbox ×

Print Compose

**Public Relations UNZA** <public@unza.zm>  
to unza ▾

Wed, Feb 27, 3:30 PM Star Reply More

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

Please see attachment for details.

You are all welcome to attend.

Regards.

Damaseke Chibale  
Manager, Public Relations

\*\*\*

  
THE UNIVERSITY OF ZAMBIA  
SCHOOL OF HUMANITIES AND SOCIAL SCIENCES  
DEPARTMENT OF MEDIA AND COMMUNICATION STUDIES  
PRESENTS THE SECOND COLLOQUIUM

# Public Oral Examinations

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

 THE UNIVERSITY OF ZAMBIA  
THE UNIVERSITY OF ZAMBIA  
TECHNICAL COLLEGE

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, 7<sup>th</sup> March, 2019

TIME: 12:00-13:00 hrs.

VENUE: VET LT

# DRGS Organised Events

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

Monday 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

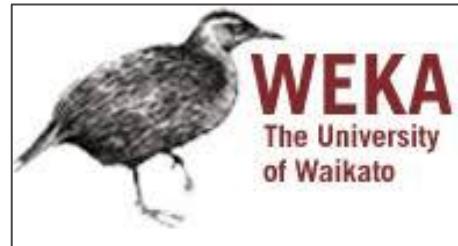
Wednesday 2 <sup>nd</sup> October, 2019	Senate Curriculum and Examinations Committee Meeting (Considering IDE Results)
Monday 14 <sup>th</sup> October, 2019 to Friday 18 <sup>th</sup> October, 2019	Study Break and Post Graduate Seminar Week
Monday 21 <sup>st</sup> October, 2019 to Friday 15 <sup>th</sup> 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

# Lecture Series Outline

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

# Getting Started with Python, SciKit-learn & Pandas

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



# Paper Reading List [Trial]

- [1] **S. Keshav (2007) "How to Read a Research Paper"**  
<https://doi.org/10.1145/1273445.1273458>
- [2] **P. W. L. Fong (2004) "How to Read a CS Research Paper?"**  
<https://doi.org/10.1145/1595453.1595493>
- [3] **L. Phiri (2018) "Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories"**  
<https://doi.org/10.1504/IJMSO.2020.112804>

# Academic Talk: L. Phiri [Trial]

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

# **Q & A Session**

- Comments, concerns and complaints?**

# Bibliography

[1] 2020/21 CSC 5741 Syllabus

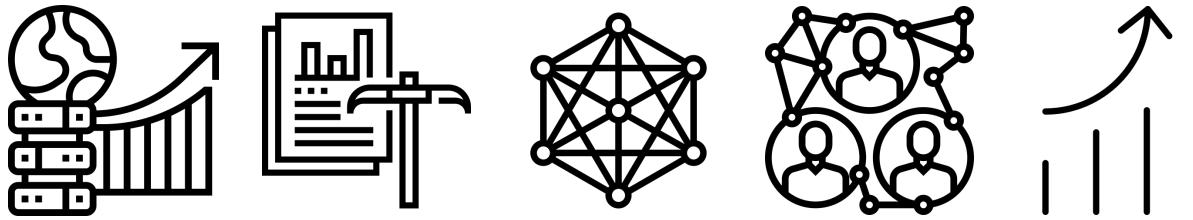
<http://bit.ly/30Pdm85>



✉ @ [csc5741@unza.zm](mailto: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>**