

**CSC 5741**

## Lecture 1: Administrivia and Course Overview



Lighton Phiri <[lighton.phiri@unza.zm](mailto:lighton.phiri@unza.zm)>

Department of Library and Information Science  
University of Zambia

### Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week

March 15 2019

CSC 5741 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: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week

March 15 2019

CSC 5741 L01 - 3

### Personal Introductions

- Your full names and preferred reference (first name, Mrs./Ms.Mr. X)
  - I prefer that you refer to me using my first name: Lighton
- Your formal education background
- What you are presently upto (Think what you do for a living)
- What you hope to get from CSC 5741

March 15 2019

CSC 5741 L01 - 4

## CSC 5741 Learning Outcomes

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

March 15 2019

CSC 5741 L01 - 5

CSC 5741 L01 - 6

## CSC 5741 Desired Outcome

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

March 15 2019

## 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.
    - One class theory tests

March 15 2019

CSC 5741 L01 - 7

CSC 5741 L01 - 8

## 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 15 2019

### TERM I

Monday 28 <sup>th</sup> January, 2019 to Sunday 10 <sup>th</sup> March, 2019	On-line Registration
Monday 4 <sup>th</sup> February, 2019 to Friday 8 <sup>th</sup> February, 2019	Orientation of First Year students
Sunday 10 <sup>th</sup> February, 2019	Arrival of Returning Regular Students
Monday 18 <sup>th</sup> February, 2019 to Friday 31 <sup>st</sup> May, 2019	Lectures for Regular Students (15 weeks)
Monday 4 <sup>th</sup> March, 2019 to Thursday 7 <sup>th</sup> March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11 <sup>th</sup> March, 2019 to Sunday 24 <sup>th</sup> March, 2019	Late Registration Period (2 Weeks)
Friday 26 <sup>th</sup> April, 2019	Social and Cultural Day
Monday 13 <sup>th</sup> May, 2019 to Friday 26 <sup>th</sup> May, 2019	IDE Students School Experience (11 Weeks)
Monday 20 <sup>th</sup> May, 2019 to Thursday 2 <sup>nd</sup> June, 2019	Graduation Week (Second Graduation Ceremony)
Monday 27 <sup>th</sup> June, 2019 to Friday 7 <sup>th</sup> June, 2019	Study Break and Post Graduate Seminars Month

<http://www.unza.zm>

## Course Structure (3/9)

- Lecture sessions**
  - Basic introduction to core concepts
  - Practical walkthroughs

TERM I	
Monday 28 <sup>th</sup> January, 2019 to Sunday 10 <sup>th</sup> March, 2019	On-line Registration
Monday 4 <sup>th</sup> February, 2019 to Friday 8 <sup>th</sup> February, 2019	Orientation of First Year students
Sunday 10 <sup>th</sup> February, 2019	Arrival of Returning Regular Students
Monday 18 <sup>th</sup> February, 2019 to Friday 31 <sup>st</sup> May, 2019	Lectures for Regular Students (15 weeks)
Monday 4 <sup>th</sup> March, 2019 to Thursday 7 <sup>th</sup> March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11 <sup>th</sup> March, 2019 to Sunday 24 <sup>th</sup> March, 2019 Friday 26 <sup>th</sup> April, 2019	Late Registration Period (2 Weeks) Social and Cultural Day
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 Committee Week
<a href="http://www.unza.zm">http://www.unza.zm</a>	
CSC 5741 L01 - 9	

March 15 2019

## 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

TERM I	
Monday 28 <sup>th</sup> January, 2019 to Sunday 10 <sup>th</sup> March, 2019	On-line Registration
Monday 4 <sup>th</sup> February, 2019 to Friday 8 <sup>th</sup> February, 2019	Orientation of First Year students
Sunday 10 <sup>th</sup> February, 2019	Arrival of Returning Regular Students
Monday 18 <sup>th</sup> February, 2019 to Friday 31 <sup>st</sup> May, 2019	Lectures for Regular Students (15 weeks)
Monday 4 <sup>th</sup> March, 2019 to Thursday 7 <sup>th</sup> March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11 <sup>th</sup> March, 2019 to Sunday 24 <sup>th</sup> March, 2019 Friday 26 <sup>th</sup> April, 2019	Late Registration Period (2 Weeks) Social and Cultural Day
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 Committee Week
<a href="http://www.unza.zm">http://www.unza.zm</a>	
CSC 5741 L01 - 10	

March 15 2019

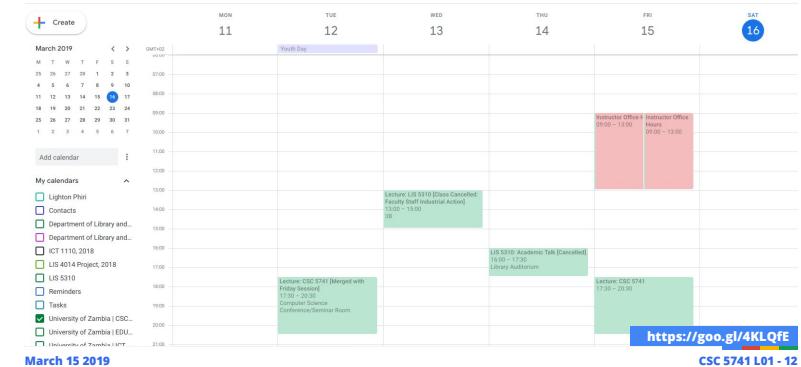
## Course Structure (5/9)

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

TERM I	
Monday 28 <sup>th</sup> January, 2019 to Sunday 10 <sup>th</sup> March, 2019	On-line Registration
Monday 4 <sup>th</sup> February, 2019 to Friday 8 <sup>th</sup> February, 2019	Orientation of First Year students
Sunday 10 <sup>th</sup> February, 2019	Arrival of Returning Regular Students
Monday 18 <sup>th</sup> February, 2019 to Friday 31 <sup>st</sup> May, 2019	Lectures for Regular Students (15 weeks)
Monday 4 <sup>th</sup> March, 2019 to Thursday 7 <sup>th</sup> March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11 <sup>th</sup> March, 2019 to Sunday 24 <sup>th</sup> March, 2019 Friday 26 <sup>th</sup> April, 2019	Late Registration Period (2 Weeks) Social and Cultural Day
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 Committee Week
<a href="http://www.unza.zm">http://www.unza.zm</a>	
CSC 5741 L01 - 11	

March 15 2019

## Course Structure (6/9)



March 15 2019

## Course Structure (7/9)

- **Course Resources**

- All course resources will be made available on The Moodle.



The screenshot shows the sign-in page of the University of Zambia Moodle LMS. It features a logo for 'The University of Zambia E-Learning Portal'. The form includes fields for 'Username' and 'Password', and a 'Log in' button. Below the form, there are links for 'Forgot your username or password?' and 'Some courses may allow guest access'. At the bottom, it says 'March 15 2019' and 'CSC 5741 L01 - 13'.

## Course Structure (8/9)

- **Course Resources**

- All course resources will be made available on The Moodle.



The screenshot shows the course page for 'LIS 5310 Information Systems and Technologies in Information Management, 2019' on the University of Zambia Moodle. The page includes a sidebar with links for 'Participants', 'Badges', 'Competencies', 'Grades', 'Dashboard', 'Site home', and 'Calendar'. The main content area displays information about 'Course Instructors' (Lighton Phiri) and his contact details. At the bottom, it says 'March 15 2019' and 'CSC 5741 L01 - 14'.

## Course Structure (9/9)

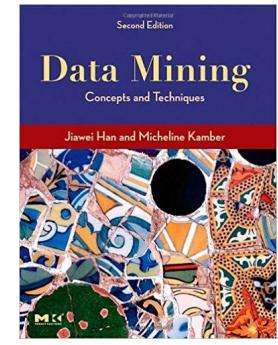
- Additionally, course resources will be disseminated as follows:

- Large files such as videos and software tools will be made available on the instructor's profile Website

## Prescribed and Recommended Textbooks (1/4)

- **Data Mining Concepts and Techniques**

- J. Han and M. Kamber (2011)

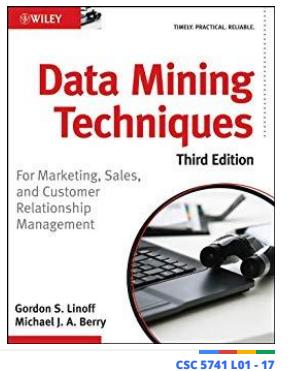


March 15 2019

CSC 5741 L01 - 16

## Prescribed and Recommended Textbooks (2/4)

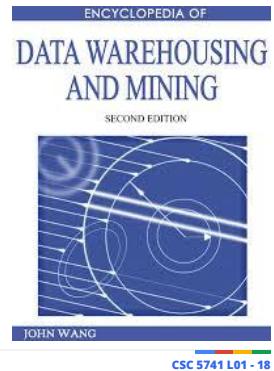
- Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management
  - G. S. Linoff and M. J. Berry (2011)



March 15 2019

## Prescribed and Recommended Textbooks (3/4)

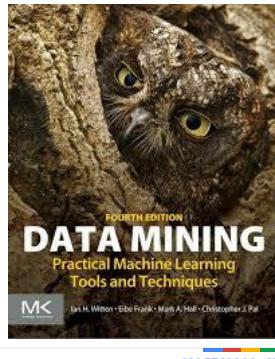
- Encyclopedia of Data warehousing and Mining
  - J. Wang (2005)



March 15 2019

## Prescribed and Recommended Textbooks (4/4)

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



March 15 2019

## Tools and Services (1/4)

- Tools and services
  - VirtualBox for creating virtual environments for running Ubuntu 18.04.
  - Ubuntu 18.04 for running all practical-oriented activities and tasks.
  - Python for scripting
  - TensorFlow
  - Weka



March 15 2019

CSC 5741 L01 - 20

## Tools and Services (2/4)

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

The screenshot shows the scikit-learn homepage. It features a main banner with a grid of small images related to machine learning. Below the banner, there are several sections: 'Classification' (describing identifying categories), 'Regression' (describing predicting continuous values), 'Clustering' (describing grouping similar objects), 'Dimensionality reduction' (describing reducing variables), 'Model selection' (describing choosing models), and 'Preprocessing' (describing feature extraction). Each section includes a brief description, applications, algorithms, and examples. At the bottom, there is a link to <https://scikit-learn.org> and a footer with the text 'CSC 5741 L01 - 21'.

March 15 2019

## Tools and Services (3/4)

- **pandas**
  - Python data analysis library



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 finance its growth.

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 the default channel:

conda install pandas

<https://pandas.pydata.org>

CSC 5741 L01 - 22

March 15 2019

## Tools and Services (4/4)

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

The screenshot shows the Matplotlib homepage. It features a large logo at the top. Below the logo, there are sections for 'Installation', 'Documentation', and 'API'. There are also several sample plots and a 'User's Guide'. At the bottom, there is a link to <https://matplotlib.org> and a footer with the text 'CSC 5741 L01 - 23'.

March 15 2019

## Course Grading (1/4)

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

March 15 2019

CSC 5741 L01 - 24

## Course Grading (2/4)

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

March 15 2019

CSC 5741 L01 - 25

## Course Grading (3/4)

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

March 15 2019

CSC 5741 L01 - 2

## Course Grading (3/4)

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

Examination:														
1	Student ID	N	Paper #1 [1%]	N	Paper #2 [1%]	N	Paper #3 [1%]	N	Paper #4 [1%]	N	Paper #5 [1%]	N	[...]	Paper #10 [1%]
2	XXXXXXXXXX	N	N	N	N	N	N	N	N	N	N	N	N	Paper Total [10%]
3	YYYYYYYYYY	N	N	N	N	N	N	N	N	N	N	N	N	Seminar #1 [1%]
														Seminar #2 [1%]
														Seminar #3 [1%]
														Seminar #4 [1%]
														Seminar #5 [1%]
														[...]
														Seminars Total [10%]
														Z Mini Project [20%]
														Z Class. Test [20%]
														N Continuous Assessment [60%]
														N Final Examination [40%]

March 15 2019

CSC 5741 L01 - 27

## Course Grading (4/4)

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

March 15 2019

CSC 5741 L01 - 2

## **Readings and Paper Summaries (1/5)**

The screenshot shows the Mendeley Desktop application window. The menu bar includes File, Edit, View, Tools, Help, and a Mendeley logo. Below the menu is a toolbar with icons for Add, Folders, Related, Sync, and Help. The main area has tabs for My Library and Creating a National Electron... The left sidebar shows sections for 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 Title, Year Published, and In. The right sidebar includes Details, Notes, and Contents tabs, and a Type filter set to Conference Proceedings. A blue box highlights the 'Creating a National Africa' section and the 'Mendeley Desktop' tab.

	Title	Year Published	In
1	Willinsky, John - Open Journal Systems	2005	Library Hi...
2	Akandabwila, Ak...	2009	African Jo...
3	Kulyambarino, C...	2016	

- You MUST develop the habit of using a bibliographic manager
    - Mendeley, Zotero, JabRef, RefWorks

March 15 2019

CSC 5741 L01 - 29

## **Readings and Paper Summaries (2/5)**

The screenshot shows a Google Scholar search results page. The search query is "data mining and machine learning". The results are filtered by "Articles". There are 2,040,000 results found in 0.12 seconds. The first result is a citation for "Data Mining: Practical machine learning tools and techniques" by Ian H. Witten, E. Frank, and M. A. Hall Jr., published in 2016. The second result is a preprint titled "Distributed GraphLab: a framework for machine learning and data mining in the cloud" by Y. Low, J. Bickson, C. Gonzalez, and C. Guestrin, presented at the 2012 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The third result is a book chapter titled "Business data mining—a machine learning perspective" by I. Bosscher and R.K. Mahapatra, published in 2001. The fourth result is a citation for "Machine learning software: an update" by M. Hall, E. Frank, G. Holmes, B. Pfahringer, and P.M. Clark, published in 2009.

March 15 2019

[PDF] researchgate.net  
<https://scholar.google.com>  
CSC 5741 L01 - 30

## **Readings and Paper Summaries (3/5)**

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	Int. Sci.	62.00	
Computer Network	3	ACM Knowledge Discovery and Data Mining	KDD	56.00	
Network and Information Security	4	IEEE Transactions on Knowledge and Data Engineering	TKDE	53.00	
Software Engineering	5	ACM International Conference on Web Search and Data Mining	WSDM	50.00	
Database and Data Mining	6	International Conference on Research an Development in Information Retrieval	SIGIR	47.00	
Theoretical Computer Science	7	Journal of the American Society for Information Science and Technology	JASIST	42.00	
Computer Graphics and Image Processing	8	IEEE International Conference on Data Engineering	ICDE	40.00	
Computer Vision and Pattern Recognition	9	ACM International Conference on Information and Knowledge Management	CIKM	38.00	
Machine Learning	10	IEEE International Conference on DataMining	ICDM	33.00	
Robotics and Mechatronics	11	Journal of Web Semantics	J. Web Sem.	33.00	
Information Systems	12	Knowledge and Information Systems	KAIS	31.00	
Geoinformatics	13	International Journal of Geographical Information Science	IJGIS	31.00	

March 15 2019

CSC 5741 L01 - 31

## **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 CV's) so please email me if you notice any errors or omissions.

AAAI (Artificial Intelligence)	
2018	Memory-Augmented Monte Carlo Tree Search
2017	Label-Free Supervised Neural Networks with Physics and Domain Knowledge
2016	Bidirectional Search That Is Guaranteed to Meet in the Middle
2015	From Non-Negative to General Operator Cost Partitioning
2014	Recovering from Selection Bias in Causal and Statistical Inference
2013	HC-Search: Learning Heuristics and Cost Functions for Structured Prediction
2012	SMILE: Shuffling Multiple-Instance Learning
2011	Learning SVM Classifiers with Indefinite Kernels Document Summarization Based on Multi-Data Reconstruction Dynamic Programming Action in Continuous Reinforcement Learning Complexity of an Algorithm for Berlekamp-Massey
2010	How Incomplete Is Your Semantic Web Reasoner? Systematic Analysis of the Completeness of Query Ans. A Novel Transition-Based Encoding Scheme for Planning as Satisfiability
2009	How Good Is Almost Perfect?
2008	Optimal False-Name-Proof Voting Rules with Costly Voting
2007	PLOW: A Collaborative Task Learning Agent Thresholded Rewards Acting Optimally in Timed, Zero-Sum Games
	Chenjun Xiao, University of Alberta; et al. Russell Stewart & Stefano Ermon, Stanford University Robert C. Holte, University of Alberta; et al. Florian Pommerehne, University of Basel; et al. Elias Bareinboim, University of California Los Angeles; et al. Janardhan Rao Doppa, Oregon State University; et al. Gary Domon & Soumya Ray, Case Western Reserve University Suicheng Gu & Yuhong Tang, Temple University Zhenling He, Zhejiang University; et al. Dmitrii Likhachev, University of Illinois Urbana-Champaign; et al. Jesús de la Cosa, University of Toronto; et al. Georges Stolas, Oxford University; et al. Ruoyun Huang, Washington University in St. Louis; et al. Matte Helmert & Gabriele Röger, Albert-Ludwigs-Universität Freiburg Lisa Wagner & Vincent Conitzer, Duke University James Allen, Institute for Human and Machine Cognition; et al.
	<a href="http://jeffhuang.com/best_paper_awards.html">http://jeffhuang.com/best_paper_awards.html</a>

March 15 2019

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

FSC E741 L01 33

## Readings and Paper Summaries (5/5)



Zambia ICT Journal Announcements Current Archives About ▾

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

### Announcements

Call for paper for Volume 3 Issue 2 (June 2019)

□ 2019-03-08

The Zambia ICT Journal wishes to call for original research papers containing new research findings which have not been submitted concurrently to any other publication to be published in Volume 3 Issue 2 (June 2019) in any of the

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

CSC 5741 L01 - 33

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



University of Cape Town

My Author Page My Binders SIGN OUT Lighton Phiri

SEARCH

### How to read a paper

Full Text: [PDF](#)

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

Published in:  
- Newsletter  
ACM SIGCOMM Computer Communication Review archive

### 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  
Volume 41 Issue 2, June 2009  
Pages 138-140



Bibliometrics

### Tools and Resources

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

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

[TOC Service](#)

[Email](#) [RSS](#)

### Tools and Resources

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

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

[TOC Service](#)

[Email](#) [RSS](#)

<https://dl.acm.org>

CSC 5741 L01 - 34

March 15 2019

## 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

March 15 2019

CSC 5741 L01 - 36

March 15 2019

CSC 5741 L01 - 35

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

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

March 15 2019

CSC 5741 L01 - 37

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

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

March 15 2019

CSC 5741 L01 - 38

## Mini Project Overview (1/2)

- Ideally, everyone will work on an individual project, however, the themes will be similar
  - Mini Project descriptions
  - Grade distribution
  - Open date and deadline
- Deliverables
  - Properly curated datasets
  - Well-organised scripting code (Github/Bitbucket)
  - Technical report



March 15 2019

CSC 5741 L01 - 39

## Mini Project Overview (2/2)

- The 20% score allocated to the Mini Project will be distributed
  - Implementation of chosen problem
  - Presentation of implementation
  - Technical report based on implementation
- Individual problem sets
- Open date: March 29, 2019
- Due date: May 17, 2019

ASPECT	SCORE
1. IMPLEMENTATION (data collection, data cleaning, functionality)	8%
2. TECHNICAL REPORT (presentation, layout, style, language, evaluation, analysis)	8%
3. PRESENTATION (Q&A, style, presentation slides)	4%

March 15 2019

CSC 5741 L01 - 40

## Academic Dishonesty

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

March 15 2019

CSC 5741 L01 - 41

## 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>)

March 15 2019

CSC 5741 L01 - 42

## Course Management (2/2)

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

The screenshot shows a Moodle course mailing list interface. At the top, there's a search bar and navigation buttons for 'Groups', 'New topic', 'Mark all as read', 'Actions', and 'Filters'. Below this, a message from '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 main area displays a group named 'CSC 5741: Data Mining and Warehousing' which is shared privately. It shows 2 of 2 topics. A message from 'CSC 5741: Data Mining and Warehousing' discusses the course's objectives and provides links to welcome messages and a reminder about class times. The URL at the bottom is <https://groups.google.com/a/unza.zm/d/forum/csc5741>.

March 15 2019

CSC 5741 L01 - 43

## Q & A Session

- Comments, concerns and complaints?

March 15 2019

CSC 5741 L01 - 44

## Lecture Series Outline

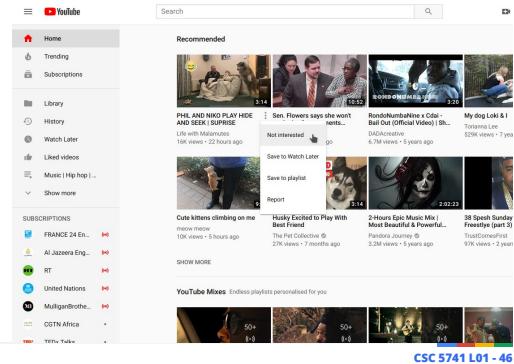
- **Part I: Administrivia**
- **Part II: Course Introduction**
  - Contextualising Data Mining and Warehousing
  - CSC 5741 Themes and Topics
- **Part III: On Academic Activities**
- **Part IV: Getting Started With Python**
- **Part V: About Next Week**

March 15 2019

CSC 5741 L01 - 45

## Contextualising Data Mining & Warehousing (1/14)

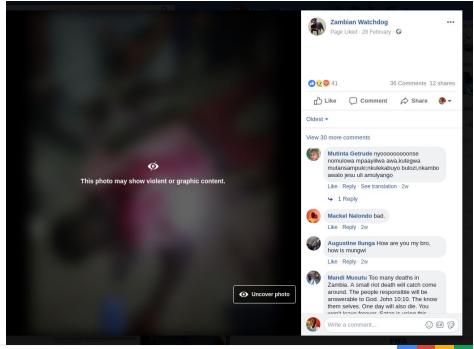
- **With the ever increasing amount of data being generated, application of data mining techniques are increasing.**



March 15 2019

## Contextualising Data Mining & Warehousing (2/14)

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



March 15 2019

CSC 5741 L01 - 47

## Contextualising Data Mining & Warehousing (3/14)

- **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
- **Current CS@ UNZA Dissertations**
  - Simon Hawatichke Chiwamba (2019—). Machine Learning Automated Image Capture and Identification of Fall Armyworm
  - Francis Chulu (2019—). Automatic identification and early warning and monitoring web based system of fall Armyworm

March 15 2019

CSC 5741 L01 - 48

## Contextualising Data Mining & Warehousing (4/14)

- **Automatic classification of scholarly research**
    - Automatic generation of metadata
    - Automatic reclassification of digital 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/oai_dc.xsd" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:title>
      Customer service management in the retail
    </dc:title>
    <dc:creator>Atanga, Muyenga</dc:creator>
    <dc:subject>Banks</dc:subject>
    <dc:subject>Retail banking</dc:subject>
    <dc:subject>MBA THESIS</dc:subject>
    <dc:subject>Customer service</dc:subject>
    <dc:description>v.116p.</dc:description>
    <dc:description>Copperbelt University, Sch</dc:description>
    <dc:date>2011-07-19T14:32:14Z</dc:date>
    <dc:date>2011-07-19T14:32:14Z</dc:date>
    <dc:date>2011-07-19</dc:date>
    <dc:type>Thesis</dc:type>
  </oai_dc:dc>
</metadata>
```

March 15 2019

CSC 5741 L01 - 49

## Contextualising Data Mining & Warehousing (6/14)

- LMS Log Mining
    - Moodle usage logs

March 15 2019

CSC 5741 L01 - 51

## Contextualising Data Mining & Warehousing (5/14)

- University of Zambia Ranking Committee Research Report

- Mining for scholarly output on the Web

March 15 2019

CSC 5741 L01 - 50

## Contextualising Data Mining & Warehousing (7/14)

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



March 15 2019

CSC 5741 L01 - 52

## Contextualising Data Mining & Warehousing (8/14)

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

```

> colnames(dataset_cfs0405_crop)
[1] "PROV" "DIST" "CONST" "WARD" "REGION" "CSA"
[7] "SEA" "HHNUM" "CROP" "ID009" "SIAFIELD" "SIAFC01"
[13] "SIAFC02" "SIAFC03" "SIAFC04" "SIAFC05" "SIAFC06" "SIAFC07"
[19] "SIAFC08" "SIAFC09" "SIAFC10" "SIAFC11" "SIAFC12" "SIAFC13"
[25] "SIAFC14" "SIAFC15" "SIAFC16" "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 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 14 1 1 388.84409

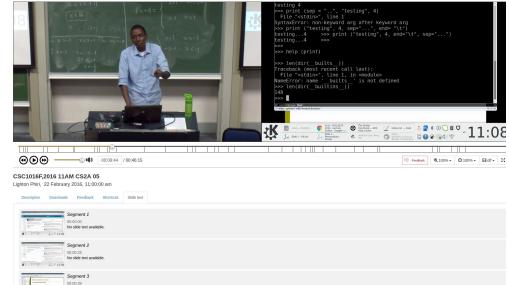
```

March 15 2019

CSC 5741 L01 - 53

## Contextualising Data Mining & Warehousing (9/14)

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



March 15 2019

CSC 5741 L01 - 54

## Contextualising Data Mining & Warehousing (10/14)

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



March 15 2019

CSC 5741 L01 - 55

## Contextualising Data Mining & Warehousing (11/14)

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

March 15 2019

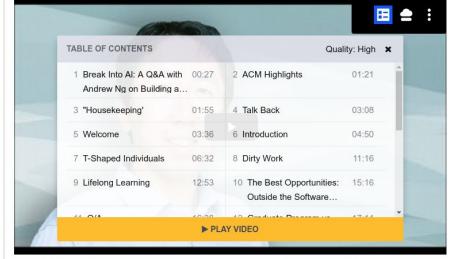
CSC 5741 L01 - 56

## Contextualising Data Mining & Warehousing (12/14)

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

March 15 2019

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



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

CSC 5741 L01 - 57

## Contextualising Data Mining & Warehousing (13/14)

- Curiosity-driven research
  - Puzzles
  - Games

March 15 2019

### 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 Sq

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

CSC 5741 L01 - 58

## Introduction (1/2)

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

March 15 2019

## Contextualising Data Mining & Warehousing (14/14)

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

March 15 2019

Government wants help with monitoring content from Radio and TV stations in Zambia-Siliya

March 18, 2019 1:24 pm 11



Minister of Information and Broadcasting Services Dora Siliya says the

CSC 5741 L01 - 59

## Introduction (2/2)

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

March 15 2019

CSC 5741 L01 - 61

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

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

March 15 2019

CSC 5741 L01 - 62

## Theme #2: Data Warehousing

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

March 15 2019

CSC 5741 L01 - 63

## Theme #3: Classification

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

March 15 2019

CSC 5741 L01 - 64

## Theme #4: Associative Rule Mining

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

March 15 2019

CSC 5741 L01 - 65

## Theme #5: Clustering Analysis

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

March 15 2019

CSC 5741 L01 - 66

## Closing CSC 5710 Remarks

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

March 15 2019

CSC 5741 L01 - 67

## Q & A Session

- Comments, concerns and complaints?

March 15 2019

CSC 5741 L01 - 68

## Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
  - Public Talks
  - Public Oral Examinations
  - DRGS Organised Events
- Part IV: Getting Started With Python
- Part V: About Next Week

March 15 2019

CSC 5741 L01 - 69

## Public Talks

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

March 15 2019

2nd Seminar in the Colloquium Series for 2019 [Inbox X](#)

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

GIP Wed, Feb 27, 3:30 PM

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 Chihale  
Manager, Public Relations

\*\*\*



CSC 5741 L01 - 70

## Public Oral Examinations

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



### SCHOOL OF AGRICULTURAL SCIENCES SEMINAR SERIES

#### PhD Public Defence

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

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

All students to attend

DATE: Thursday, 7<sup>th</sup> March, 2019

TIME: 12:00-13:00 hrs.

VENUE: VET LT

March 15 2019

CSC 5741 L01 - 71

## 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 15 2019

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 (Senate and IDE Panels)

Monday 14<sup>th</sup> October, 2019 to  
Friday 18<sup>th</sup> October, 2019

Study Break and Post Graduate Seminar Week

Monday 22<sup>nd</sup> October, 2019 to  
Friday 15<sup>th</sup> November, 2019

Final Examinations (25 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 L01 - 72

## Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
  - Installation and Setup
  - Basics
  - Data Structures
  - Flow Control
  - Functions and Modules
- Part V: About Next Week

March 15 2019

CSC 5741 L01 - 73

## Getting Started With Python (1/2)

```
Python 3.6.7 (default, Oct 22 2018, 11:32:17)
[GCC 8.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import this
The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
```

March 15 2019

CSC 5741 L01 - 74

## Getting Started With Python (2/2)

- Python is an interpreted language
- Python is a scripting language
- Python is a general purpose language
- Python is an Object Oriented language
- [...]
- [...]
- We recommend using Python 3

March 15 2019

CSC 5741 L01 - 75

## Installation and Setup

- [...]

March 15 2019

CSC 5741 L01 - 76

## Installation and Setup

- Python statements can be executed directly from the interpreter
- Python scripts can be executed as shell commands

March 15 2019

CSC 5741 L01 - 77

## Basics

- No need to specify data types on variable declaration
- Indentation is important

March 15 2019

CSC 5741 L01 - 78

## Data Structures

- Tuple
  - var = (1, 2, 3, 4, 5)
- List
  - var = [1, 2, 3, 4, 5]
- Dictionary
  - var = {"one":1, "two":2, "three":3, "four":4, "five":5}
- Set
  - var = {1, 2, 3, 4, 5}

March 15 2019

CSC 5741 L01 - 79

## Loops

```
for i in [1,2,3]:  
    print(i)
```

```
while i < 5:  
    i += 1  
    print(i)
```

- No curly braces or "end for"
- Structure is derived from level of indentation
- One statement per line
- No semicolons required

March 15 2019

CSC 5741 L01 - 80

## Loops

```
def csc5741(x, y='Y', z='Z'):
    print(x + ' ' + y)
return 0

csc5741('Xxxx', 'Yyyyy')
```

- All arguments are named
- Naming useful for optional arguments
- Return is optional

March 15 2019

CSC 5741 L01 - 81

## Modules (1/2)

- Modules facilitate extensibility and reusability
- [...]
- [...]
- from math import sqrt
- import math

March 15 2019

CSC 5741 L01 - 82

## Modules (2/2)

- Modules come with documentation
- Module documentation also accessible via interpreter

March 15 2019

CSC 5741 L01 - 83

## Installing pandas, scikit-learn and matplotlib

- pip

March 15 2019

CSC 5741 L01 - 84

## Installation and Setup

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
  - Installation and Setup
  - Basics
  - Data Structures
  - Flow Control
  - Functions and Modules
- Part V: About Next Week

March 15 2019

CSC 5741 L01 - 85

## Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week
  - Data Mining and Data Processing
  - Getting Started: SciKit-learn, Pandas
  - Mini Project Descriptions
  - Paper Reading List [Trial]
  - Academic Talk: L. Phiri [Trial]

March 15 2019

CSC 5741 L01 - 86

## Theme #1: Data Mining and Data Processing

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

March 15 2019

CSC 5741 L01 - 87

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

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



March 15 2019

CSC 5741 L01 - 88

## 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) "Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia"  
[http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online\\_visibility.pdf](http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online_visibility.pdf)

March 15 2019

CSC 5741 L01 - 89

## Academic Talk: L. Phiri [Trial]

- [1] L. Phiri (2018) "Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia"  
[http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online\\_visibility.pdf](http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online_visibility.pdf)

March 15 2019

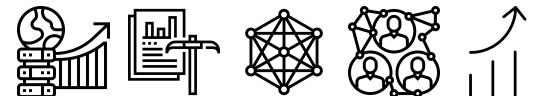
CSC 5741 L01 - 90

## Bibliography

- [1] CSC 5741 Syllabus  
<http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741>

March 15 2019

CSC 5741 L01 - 91



## CSC 5741 Lecture 1: Administrivia and Course Overview

Lighton Phiri <[lighton.phiri@unza.zm](mailto:lighton.phiri@unza.zm)>  
Department of Library and Information Science  
University of Zambia

CSC 5741 L01 - 91