

CSC 5741

Lecture 3: Data Mining and Data Processing

Lighton Phiri <lighton.phiri@unza.zm>

Department of Library and Information Science
University of Zambia



Lecture Series Outline

- **Part I: Data Mining**
- **Part II: Data Processing and Transformation**
- **Part III: Paper Reading Discussion**
- **Part IV: Academic Talk**

Lecture Series Outline

- **Part I: Data Mining**
 - Introduction
 - Data Mining Process
 - Data Mining Process Example
- **Part II: Data Processing and Transformation**
- **Part III: Paper Reading**
- **Part IV: Academic Talk**

Introduction

- **Similar to the motivation for using computer systems, data mining involves processing of input data to yield information**
 - Increasingly, massive amounts of data are being frequently produced
 - Raw data does not provide useful insight in comparison to information

CRISP-DM Open Standard (1/4)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
 - CRISP-DM segments a data mining project into six phases with no strict order of execution
 - Surveys conducted suggest CRISP-DM is the most widely used methodology



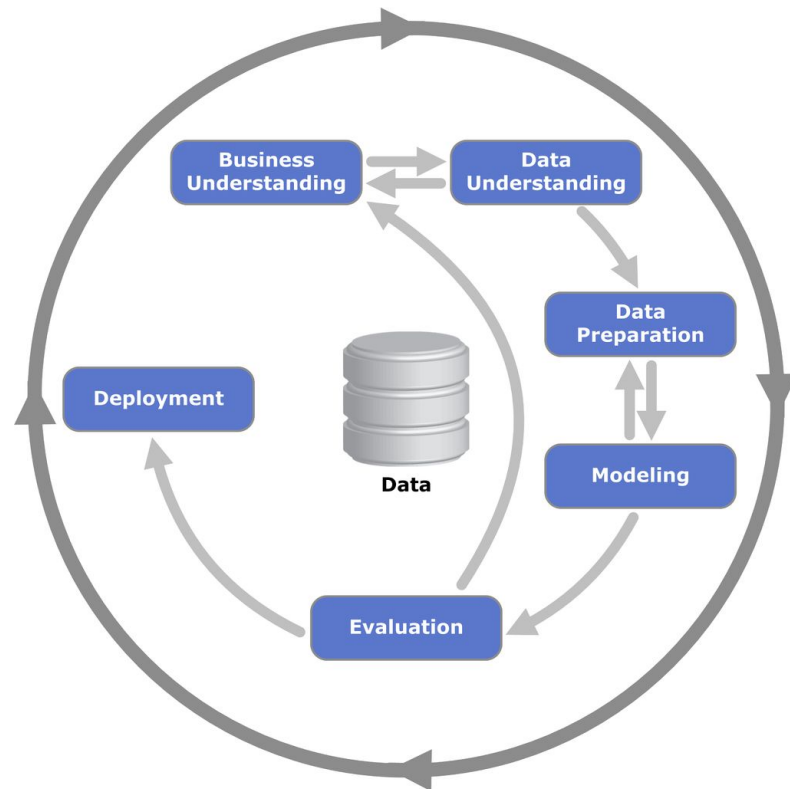
CRISP-DM Open Standard (2/4)

- **Business Understanding**
 - Situational analysis; problem definition, general and specific objectives objectives; **research question(s)** and general requirements analysis
- **Data Understanding**
 - Identification of data sources; familiarisation of data sources and initial data collection



CRISP-DM Open Standard (3/4)

- **Data Preparation**
 - Data preprocessing; data cleaning and feature selection
- **Modeling**
 - Creation of model—probably machine learning model—using data mining tools
- **Evaluation**
 - Evaluation results against goals
- **Deployment**
 - Deployment of models



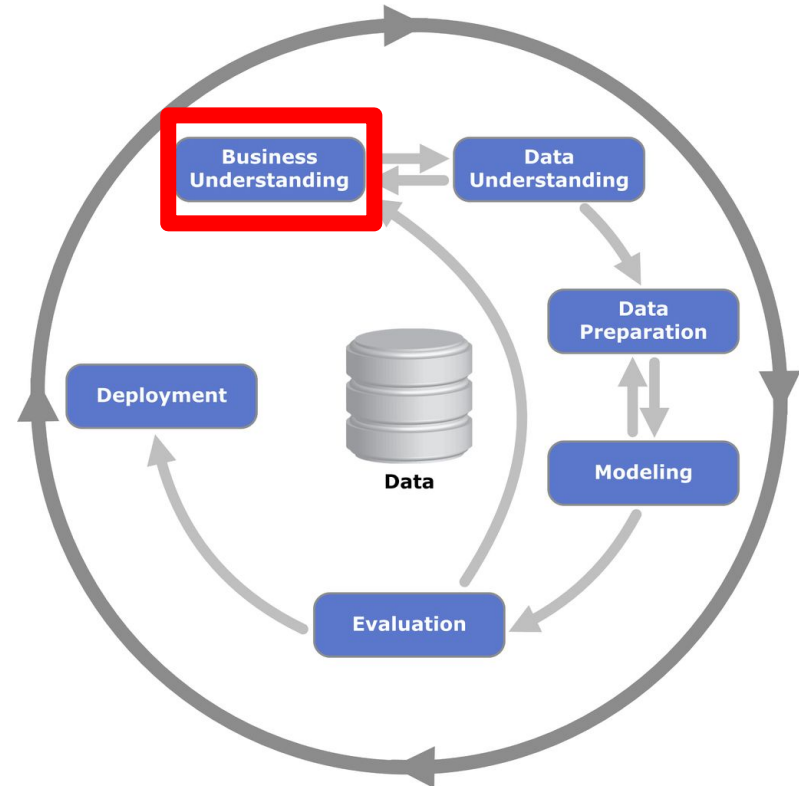
CRISP-DM Open Standard (4/4)

- While CRISP-DM is the most widely used model [2], other data mining process models that exist include:
 - Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM)
 - Sample, Explore, Modify, Model, and Assess (SEMMA)



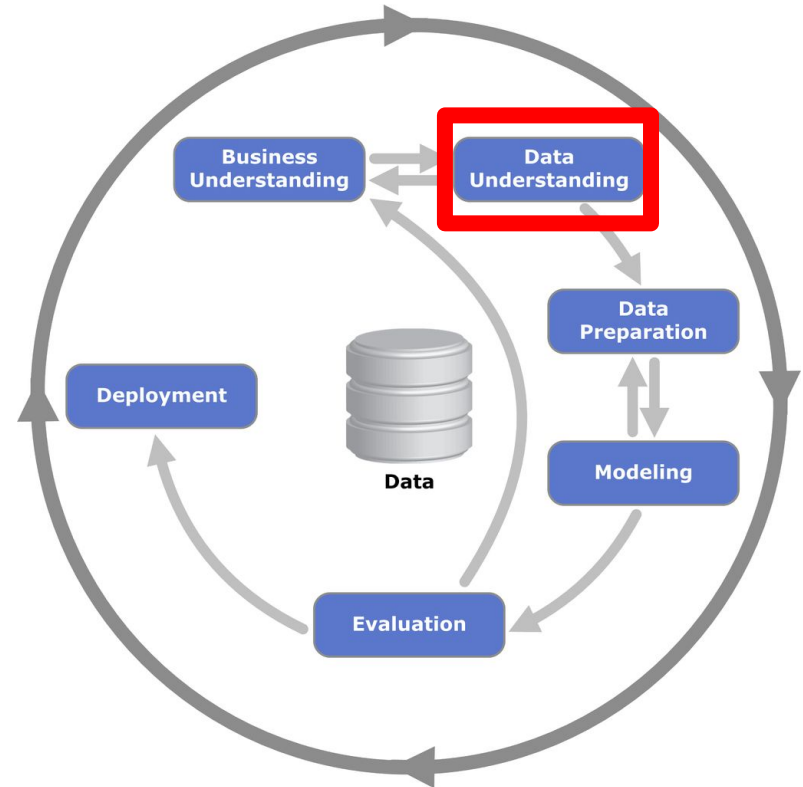
CRISP-DM—Business Understanding (1/)

- Outline business and data mining goals and objectives
- Conduct a situational analysis to identify how problem is current resolved
- Prepare an overall project plan



CRISP-DM—Data Understanding (1/)

- Identify data sources
- Extract/collect required data
- Described and explore the data collected to gain some sense of what insights to derive
- Ascertain quality of data collected



CRISP-DM—Data Preparation (1/)

- Select data required for modeling process/phase
- Clean the data
- Reconstruct the data and derive necessary attributes
- Merge different data sources
- Reformat the data
 - e.g. Naming conventions



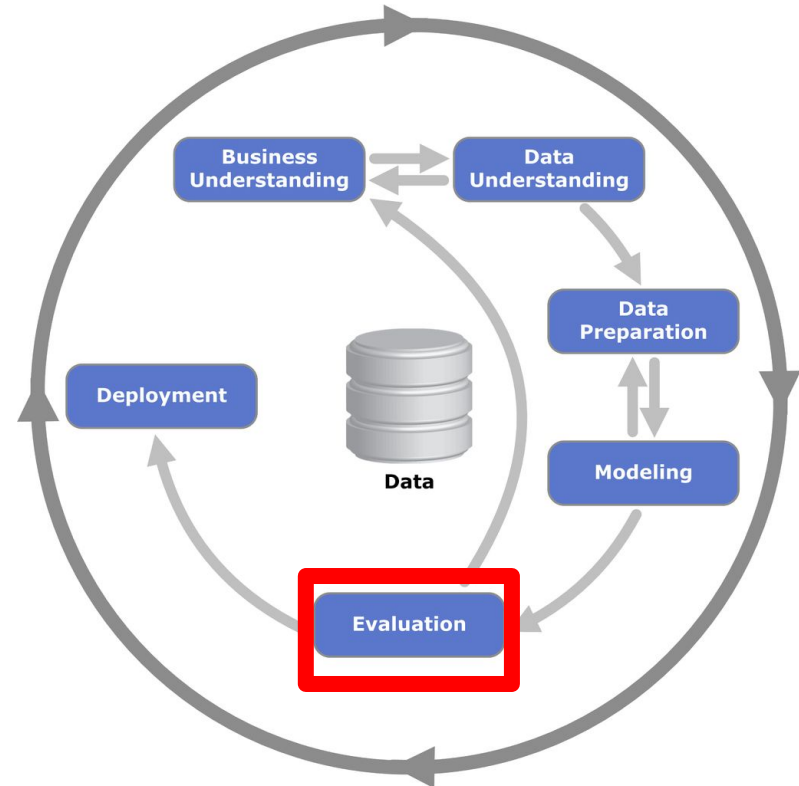
CRISP-DM—Modeling (1/)

- Define the model components, features, how it behaves and how to interpret it
- Evaluate the various alternative techniques that can be integrated with the model
 - e.g. Evaluate different classification algorithms



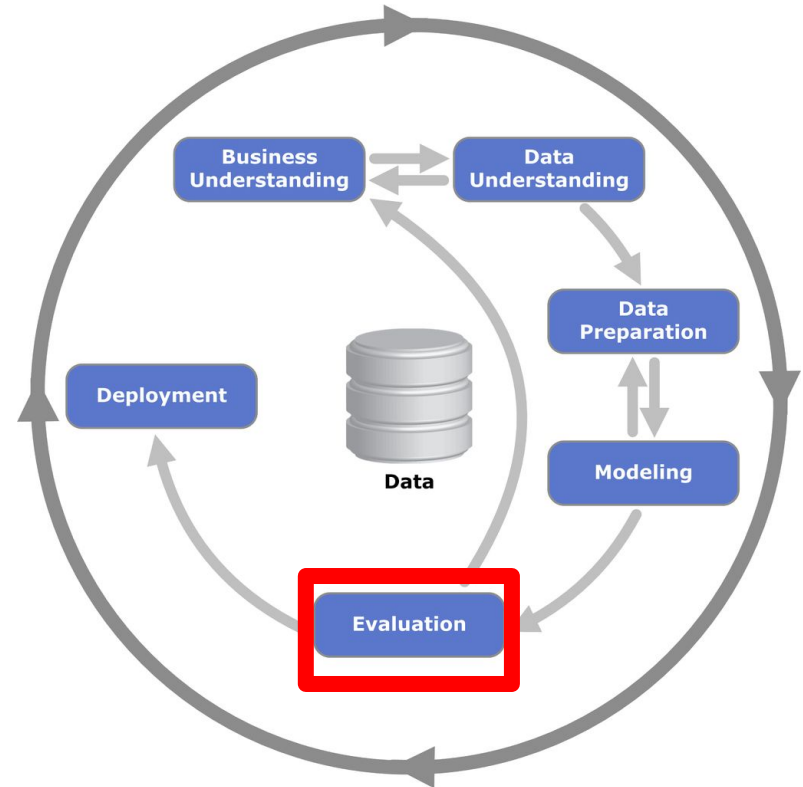
CRISP-DM—Evaluation (1/)

- Devise evaluation techniques to be used
 - Efficiency vs effectiveness/efficacy
- Interpret model results to ascertain if model should be deployed
- Review the process if necessary



CRISP-DM—Deployment (1/)

- Determine how the model results will be presented to end users
- Identify end user that will need to use the model results

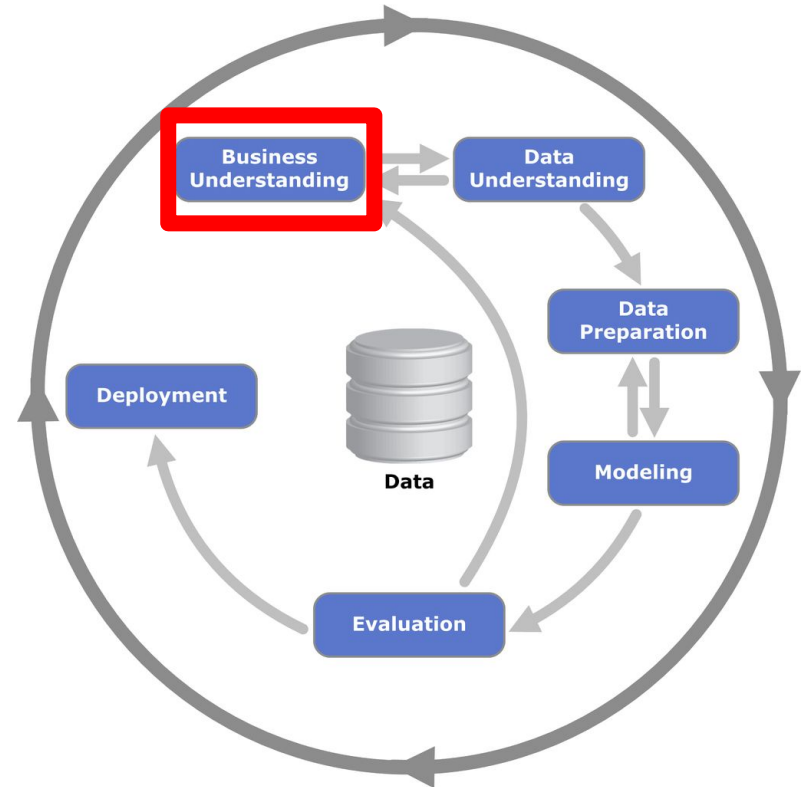


CRISP-DM—Random Example (1/)

- **ICT 1110 performance is bad. The poor performance transcends all the various assessments written by students: quizzes, tests and practical programming questions.**

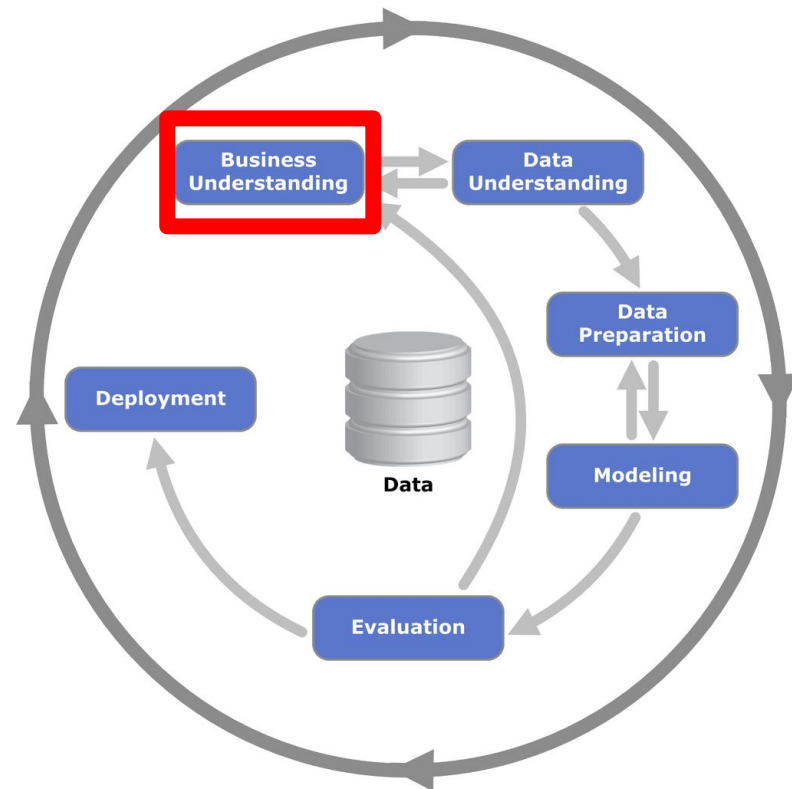
CRISP-DM—Random Example (2/)

- **Outline business and data mining goals and objectives**
 - Monitor student performance to prevent poor performance
 - Identify at risk students and devise corrective measures
- **Conduct a situational analysis to identify how problem is current resolved**
 - How are at-risk students currently identified



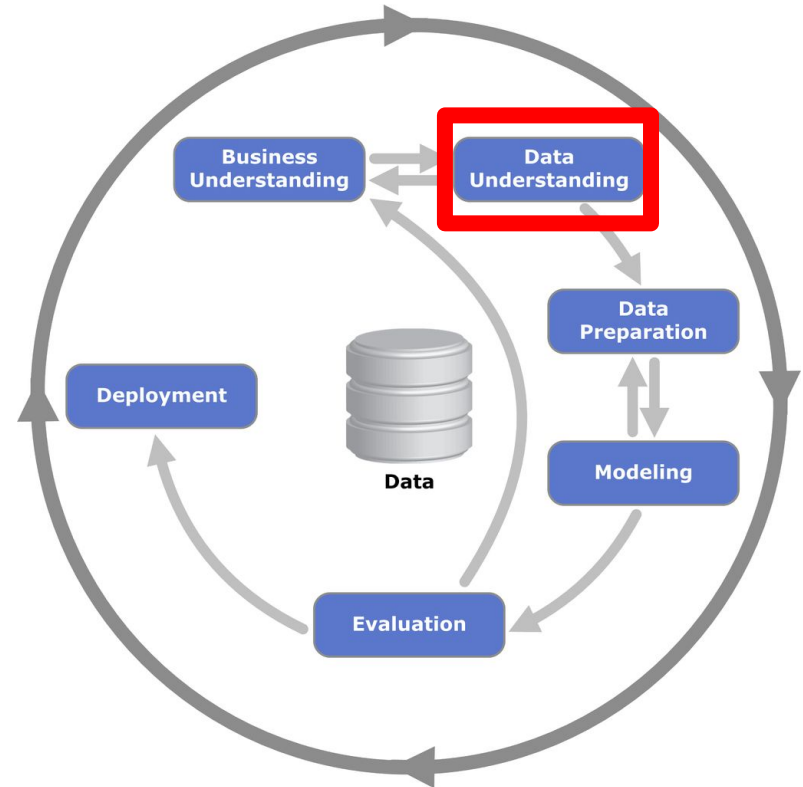
CRISP-DM—Random Example (3/)

- Prepare an overall project plan
 - Timeline of project execution



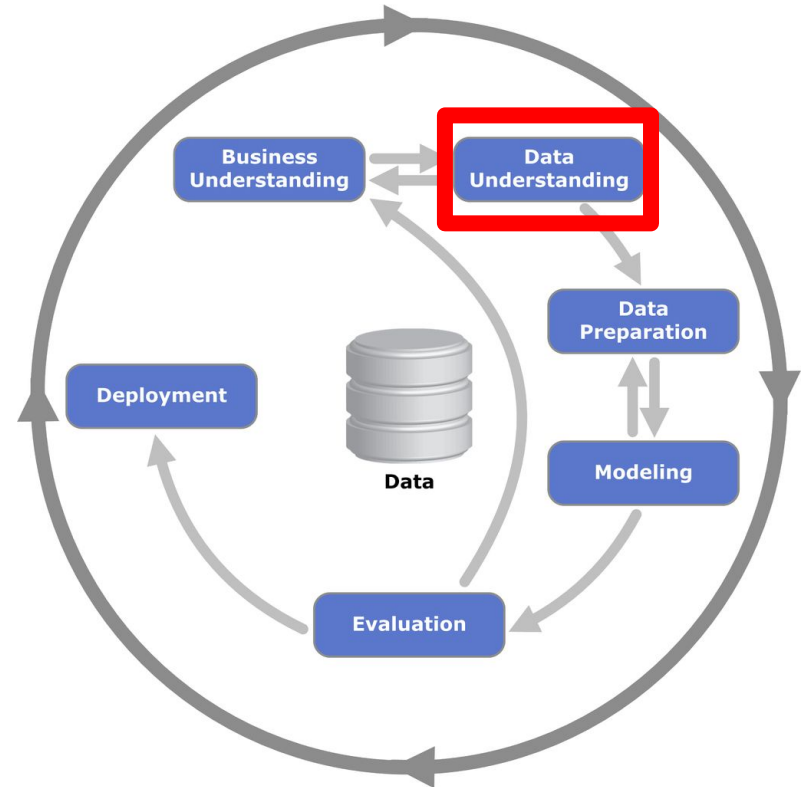
CRISP-DM—Random Example (4/)

- **Identify data sources**
 - (i) assessment results (ii) student demographics (iii) student past experience (iv) Moodle interaction logs (v) tutorial attendance (vi) lecture attendance (vii) tutor feedback
- **Extract/collect required data**
 - (i) assessment results (ii) SIS extraction (iii) questionnaire??



CRISP-DM—Random Example (5/)

- Described and explore the data collected to gain some sense of what insights to derive
- Ascertain quality of data collected



CRISP-DM—Data Preparation (1/)

- Some information for this basic example is not yet available for 2017/18
 - Lecture attendance
 - Tutorial attendance



CRISP-DM—Random Example: Data Sources (1/)

- **Assessment results broken down by question**
 - Concepts associated with question
 - Topics associated with question

```
lightonphiri@lightonphiri-Lenovo-ideapad-320-15IKB:~/Dropbox/Docum
tion$ ls db-unza18-bsc_icts_ed-ict1110-assessments-*csv
db-unza18-bsc_icts_ed-ict1110-assessments-final_ca.csv
db-unza18-bsc_icts_ed-ict1110-assessments-quiz1_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-quiz2_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-quiz3_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz10_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz11_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz12_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz13_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz16_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz17_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz20_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz7_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz8_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_quiz9_results.csv
db-unza18-bsc_icts_ed-ict1110-assessments-upto_test3_results.csv
lightonphiri@lightonphiri-Lenovo-ideapad-320-15IKB:~/Dropbox/Docum
tion$
```

CRISP-DM—Random Example: Data Sources (2/)

- Assessment results broken down by question
 - Concepts associated with question
 - Topics associated with question

	A	B	C	D	E	F	G	H	I	
1	#	Surname	Initials	Student ID	Mark	SN LEN		IS Moodle	Name Same?	
2	1		M	2017	6				TRUE	
3	2		J	2017	6				TRUE	
4	3		K.W.	2017	8				TRUE	
5	4		V	2017	6				FALSE	
6	5		C	2017	4				FALSE	
7	6		H	2017	6				TRUE	
8	7		M.B.	2017	4				TRUE	
9	8		I	2017	9				TRUE	
10	9		M	2017	3				TRUE	
11	10		M	2017	0				TRUE	
12	11		W	2017	4				TRUE	
13	12		S	2017	5				TRUE	
14	13		A	2017	7				TRUE	
15	14		M	2017	6				TRUE	
16	15		P	2017	4				TRUE	
17	16		J.S.	2017	6				FALSE	
18	17		M.M.	2017	4				TRUE	
19	18		A	2017	6				TRUE	
20	19		M	2017	5				FALSE	
21	20		S	2017	1				TRUE	
22	21		G	2017	4				FALSE	
23	22		C	2017	4				TRUE	
24	23		I	2017	4				TRUE	
25	24		J	2017	2				TRUE	
26	25		J	2017	7				TRUE	
27	26		M	2017	6				FALSE	
28	27		H	2017	8				TRUE	
29	28		D.S.	2017	4				TRUE	
30	29		A.C.	2017	4				TRUE	
31	30		B	2017	7				TRUE	
32	31		R	2017	2				TRUE	
33	32	Shahmoradian	A.B.	2017012960	4			Shahmoradian	FALSE	

CRISP-DM—Random Example: Data Sources (2/)

- LMS interaction logs
 - How often do students access Moodle (login attempts)
 - Which Moodle features are being accessed (GradeBook, Messaging)
 - Time spent on Moodle

```
1###0###1###Moodle at University of Zambia###moodle#####0###site###1###3###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 Management###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###1535550221###2016-11-23 18:09:46
17###220###5120004###VMD 6800 Veterinary Public Health###VMD 6800###VMD 6800#####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
\ short duration of the time allocated for research, you will have limited time and therefore
```


CRISP-DM—Random Example: Data Sources (2/)

- **ICT 1110 information survey to capture information not available in SIS**
 - Experience with computers
 - Motivation for taking the course
 - Specific location where student lives (although this can be inferred from next of kin address perhaps?)

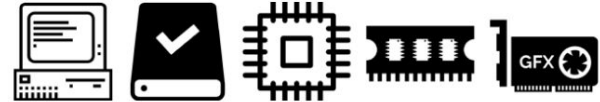
2018/19 ICT 1110 Student Information Survey

This survey is meant for us to better understand the various ICT 1110 student demographics. It also helps us better tailor course activities to ensure a positive learning experience.

Please carefully read the questions and provide us with as much detail as you can.

If you need clarification, please email us on ict1110@unza.zm

* Required



Full Names *

Your answer

Student ID *

Your answer

Hometown (suburb/town/province—e.g.
Kabwe/Lusaka/Lusaka) *

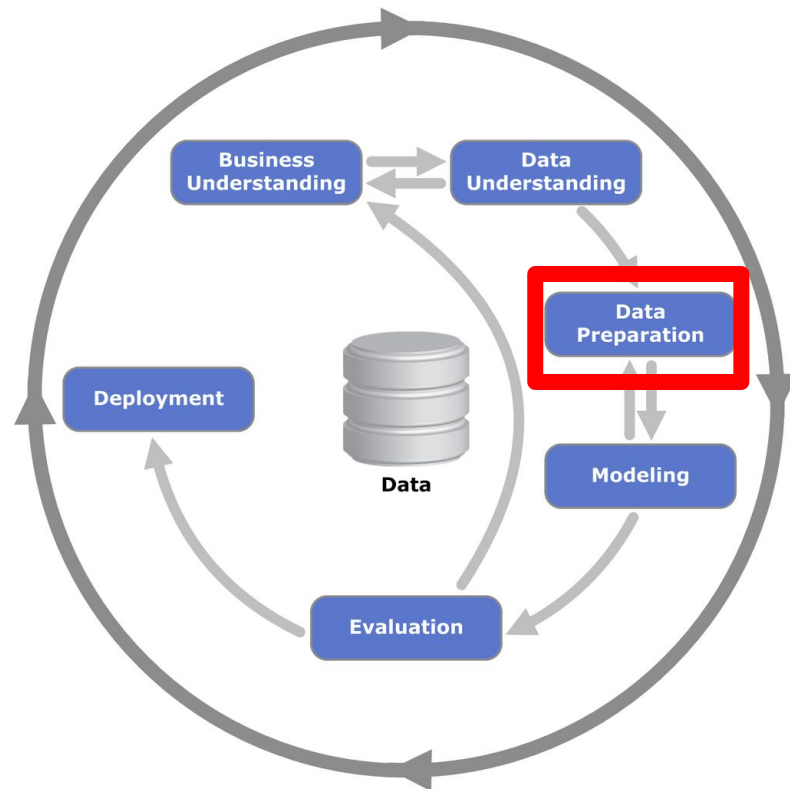
CRISP-DM—Random Example (6/)

- **Select data required for modeling process/phase**
 - Will all the data sources be used?
- **Clean the data**
 - Normalise student names names
 - Normalise their demographic details (e.g. Home Towns)



CRISP-DM—Random Example (7/)

- **Reconstruct the data and derive necessary attributes**
 - Student ages perhaps?
- **Merge different data sources**
- **Reformat the data**
 - e.g. Naming conventions



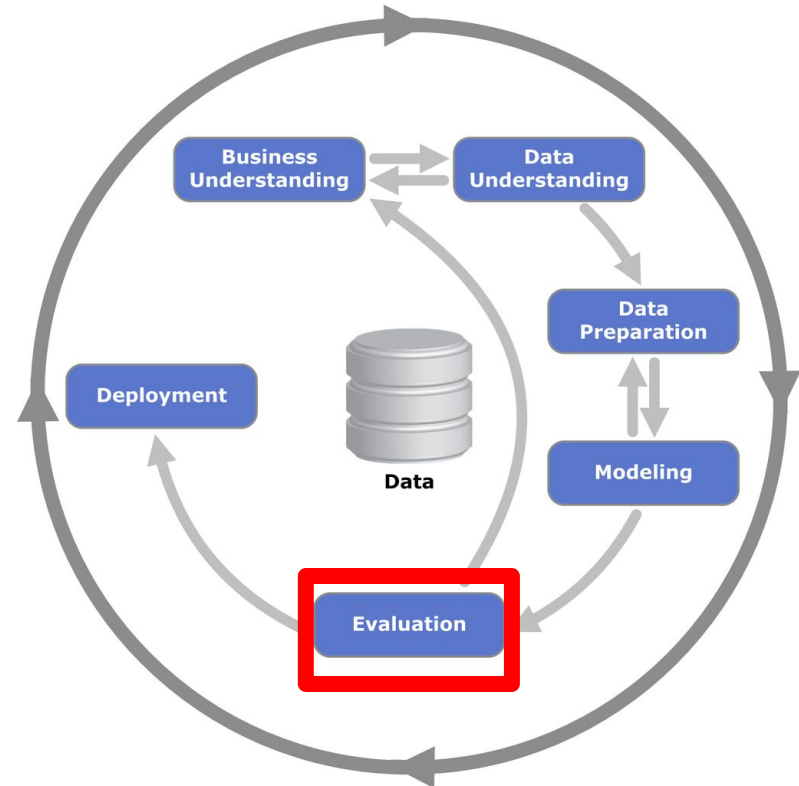
CRISP-DM—Random Example (7/)

- Define the model components, features, how it behaves and how to interpret it
- Evaluate the various alternative techniques that can be integrated with the model
 - Evaluate potentially useful classification algorithms



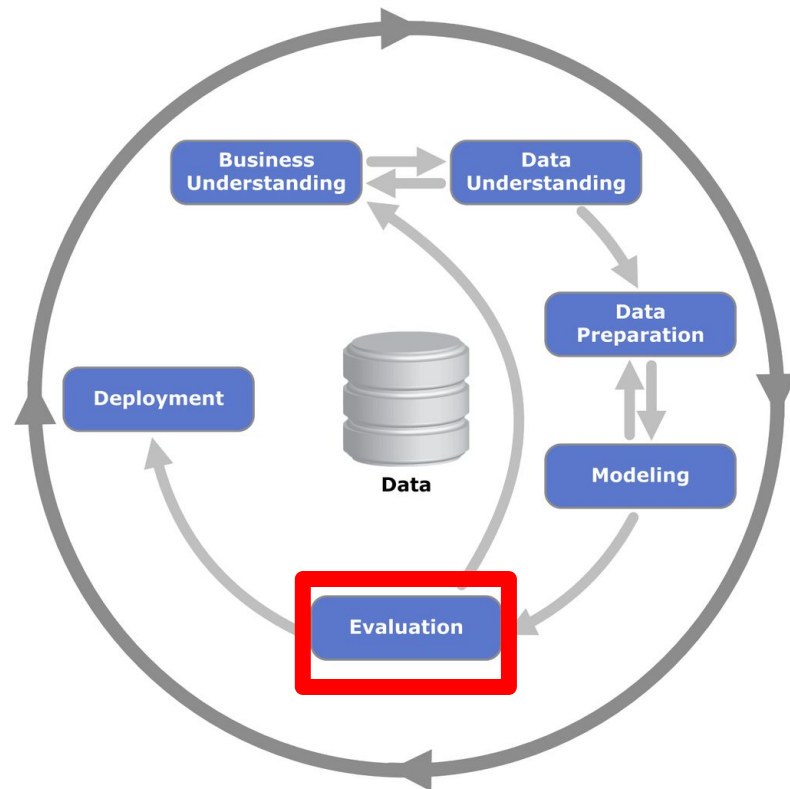
CRISP-DM—Evaluation (8/)

- Devise evaluation techniques to be used
 - Evaluate classification model's effectiveness
 - Notice that efficiency is not important here
- Interpret model results to ascertain if model should be deployed
- Review the process if necessary



CRISP-DM—Deployment (1/)

- **Determine how the model results will be presented to end users**
 - Implement a Web application that will be used to determine and present at-risk students
- **Identify end user that will need to use the model results**
 - Figure out if lecturers and tutors will have access to results. Perhaps HoD as well?



Q & A Session

- **Comments, concerns and complaints?**

Lecture Series Outline

- **Part I: Data Mining**
- **Part II: Data Processing and Transformation**
 - Data Collection and Cleaning
 - Transforming and merging data
- **Part II: Paper Reading Discussion**
- **Part IV: Academic Talk**

Lecture Series Outline

- **Part I: Data Mining**
- **Part II: Data Processing and Transformation**
- **Part II: Paper Reading Discussion**
 - M. Mgala and A. Mbogho (2015) “Data-driven intervention-level prediction modeling for academic performance”
- **Part IV: Academic Talk**

Paper Reading Session



University of Cape Town

[My Author Page](#) [My Binders](#) [SIGN OUT:](#)

Lighton Phiri

Check out a preview of the [next ACM DL](#)

Data-driven intervention-level prediction modeling for academic performance

Full Text: [PDF](#)

Authors: [Mvurya Mgala](#) [University of Cape Town, Cape Town](#)
[Audrey Mbogho](#) [University of Cape Town, Cape Town](#)

Published in:

· Proceeding

[ICTD '15](#) Proceedings of the Seventh International Conference on Information and Communication Technologies and Development
Article No. 2

Singapore, Singapore — May 15 - 18, 2015

[ACM](#) New York, NY, USA ©2015

[table of contents](#) ISBN: 978-1-4503-3163-0 doi>[10.1145/2737856.2738012](#)



2015 Article



[Bibliometrics](#)

- Citation Count: 3
- Downloads (cumulative): 152
- Downloads (12 Months): 29
- Downloads (6 Weeks): 3

Tools and Resources

[Request Permissions](#)



TOC Service:

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



[Save to Binder](#)

[View My Binders](#)



Export Formats:

[BibTeX](#) [EndNote](#) [ACM Ref](#)

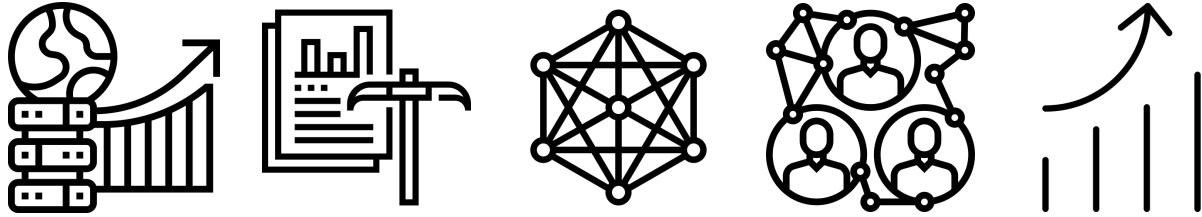
Share:



[Author Tags](#) ▼

Bibliography

- [1] **Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 1**
<https://www.cs.waikato.ac.nz/ml/weka/book.html>
- [2] **Kurgan, L. A. and Muselek, P. (2006) A Survey of Knowledge Discovery and Data Mining Process Models**
<https://doi.org/10.1017/S0269888906000737>



CSC 5741

Lecture 3: Data Mining and Data Processing

Lighton Phiri <lighton.phiri@unza.zm>

Department of Library and Information Science
University of Zambia

