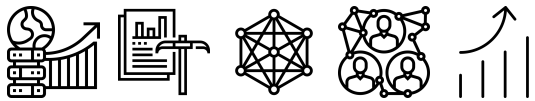


CSC 5741
Lecture 3: Data Mining and Data Processing

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)

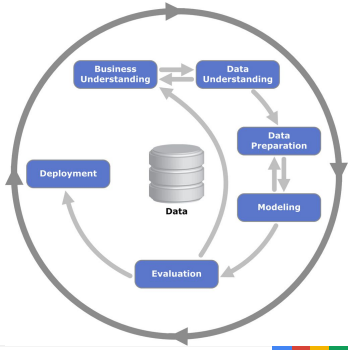
- **The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining**
  - CRISP-DM segments a data mining project into six phases with no strict order of execution
  - Surveys conducted suggest CRISP-DM is the most widely used methodology

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

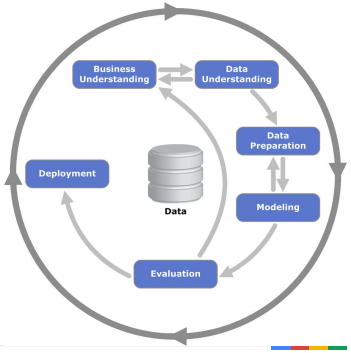
CRISP-DM Open Standard (3/4)

- **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)

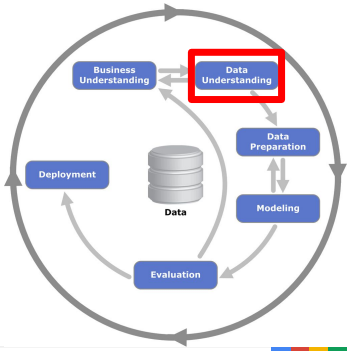


CRISP-DM—Business Understanding (1/)

- 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

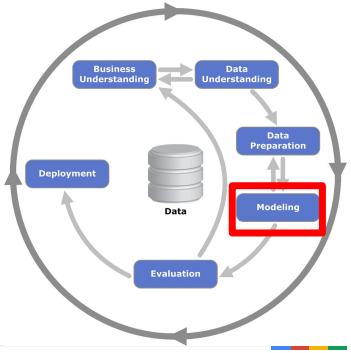


CRISP-DM—Data Preparation (1/)

- 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

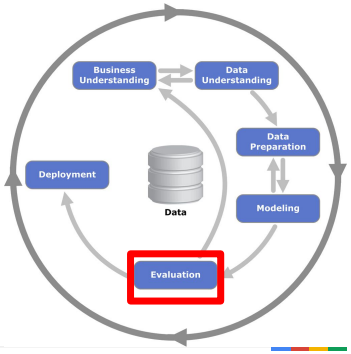
CRISP-DM—Modeling (1/)

- 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



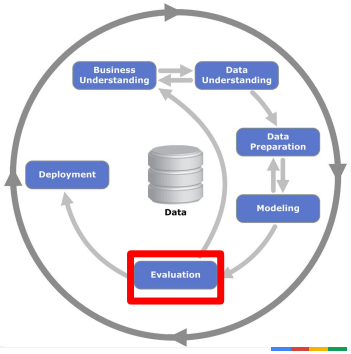
CRISP-DM—Evaluation (1/)

- 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

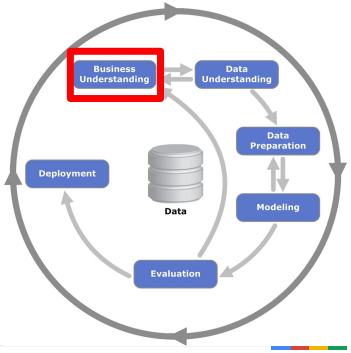


CRISP-DM—Random Example (1/)

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

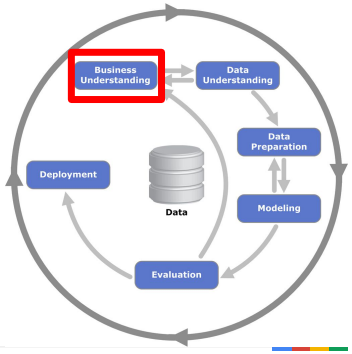
CRISP-DM—Random Example (2/)

- 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



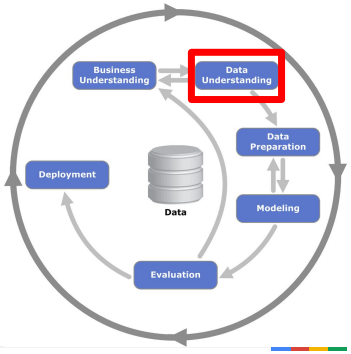
CRISP-DM—Random Example (3/)

- **Prepare an overall project plan**
  - Timeline of project execution



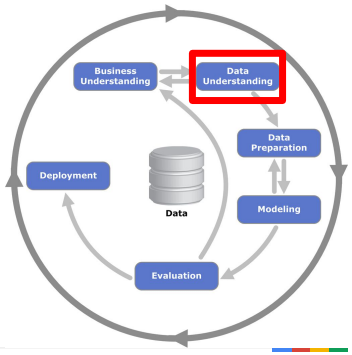
CRISP-DM—Random Example (4/)

- **Identify data sources**
  - (i) assessment results (ii) student demographics (iii) student past experience (iv) Moodle interaction logs (v) tutorial attendandance (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**

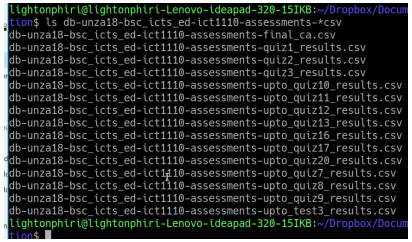


CRISP-DM—Data Preparation (1/)

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

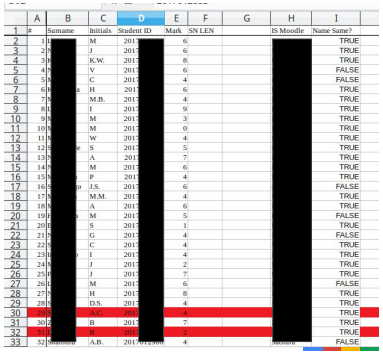
CRISP-DM—Random Example: Data Sources (1/)

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



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

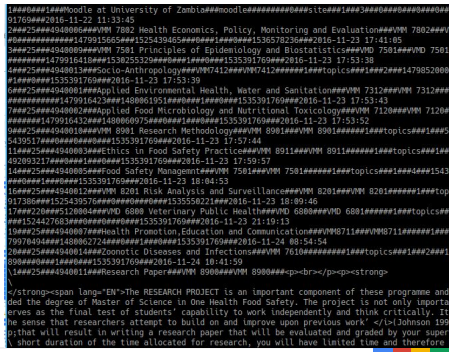
- **Assessment results broken down by question**
  - Concepts associated with question
  - Topics associated with question



	A	B	C	D	E	F	G	H	I
1	#	Surname	Initials	Student ID	Mark	SN LEN		IS Moodle	Name Same?
2	1	[redacted]	M	2013	6			[redacted]	TRUE
3	2		J	2013	6				TRUE
4	3		K.W.	2013	8				TRUE
5	4		V	2013	6				FALSE
6	5		C	2013	4				FALSE
7	6		H	2013	6				TRUE
8	7		M.B.	2013	4				TRUE
9	8		I	2013	9				TRUE
10	9		M	2013	3				TRUE
11	10		M	2013	0				TRUE
12	11		W	2013	4				TRUE
13	12		S	2013	5				TRUE
14	13		A	2013	7				TRUE
15	14		M	2013	6				TRUE
16	15		P	2013	4				TRUE
17	16		J.S.	2013	6				FALSE
18	17		M.M.	2013	4				TRUE
19	18		A	2013	6				TRUE
20	19		M	2013	5				FALSE
21	20		S	2013	1				TRUE
22	21		G	2013	4				FALSE
23	22		C	2013	4				TRUE
24	23		I	2013	4				TRUE
25	24		J	2013	2				TRUE
26	25		J	2013	7				TRUE
27	26		M	2013	6				FALSE
28	27		H	2013	8				TRUE
29	28		D.S.	2013	4				TRUE
30	30		A.C.	2013	4				TRUE
31	30		B	2013	7				TRUE
32	31		R	2013	7				TRUE
33	32	[redacted]	A.B.	2013	4			[redacted]	FALSE

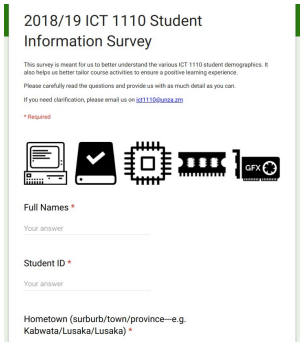
CRISP-DM—Random Example: Data Sources (2/)

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



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?)



CRISP-DM—Random Example (6/)

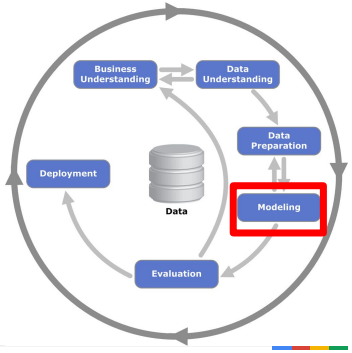
- **Select data required for modeling process/phase**
  - Will all the data sources be used?
- **Clean the data**
  - Normalise student names names
  - Normalise their demographic details (e.g. Home Towns)

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

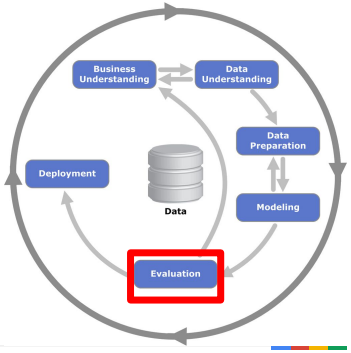
CRISP-DM—Random Example (7/)

- **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



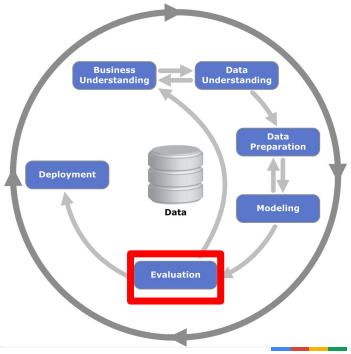
CRISP-DM—Evaluation (8/)

- **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?**

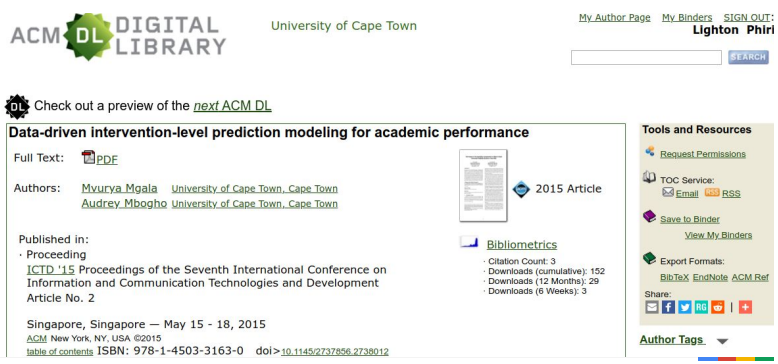
Lecture Series Outline

- **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**

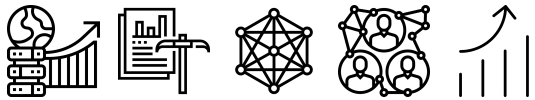
Paper Reading Session



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 Musilek, 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