**Data science methodology**

* KDD (knowledge discovery in database)
  + selection
  + pre-processing
  + transformation
  + data mining
  + interpretation/evaluation
* SEMMA (sample, explore, modify, model, assess)
* CRISP-DM (cross-industry standard process for DM)
* business understanding
  + data understanding
  + data preparation
  + modelling
  + evaluation
  + deployment

**data science workflow:**

Diagram

Description automatically generated

**data science lifecycle:**

* business understanding
* data mining
* data cleaning
* data exploration
* future engineering
* predictive modelling
* data visualisation

**dirty data:**

* incorrect
  + value must adhere to its domain
* inaccurate
  + eg. Sydney, VIC (also inconsistent)
* business rule violation:
  + a start date always precedes finish date
* inconsistent data:
  + uncontrolled data redundancy results in inconsistency
  + eg. Mary Smith, Maria Louise Smith, Mary L. Smith
* incomplete data

**tidy data:**

* each variable is in its own column
* each observation in its own row
* each value in its own cell