Overview of Data Science Process

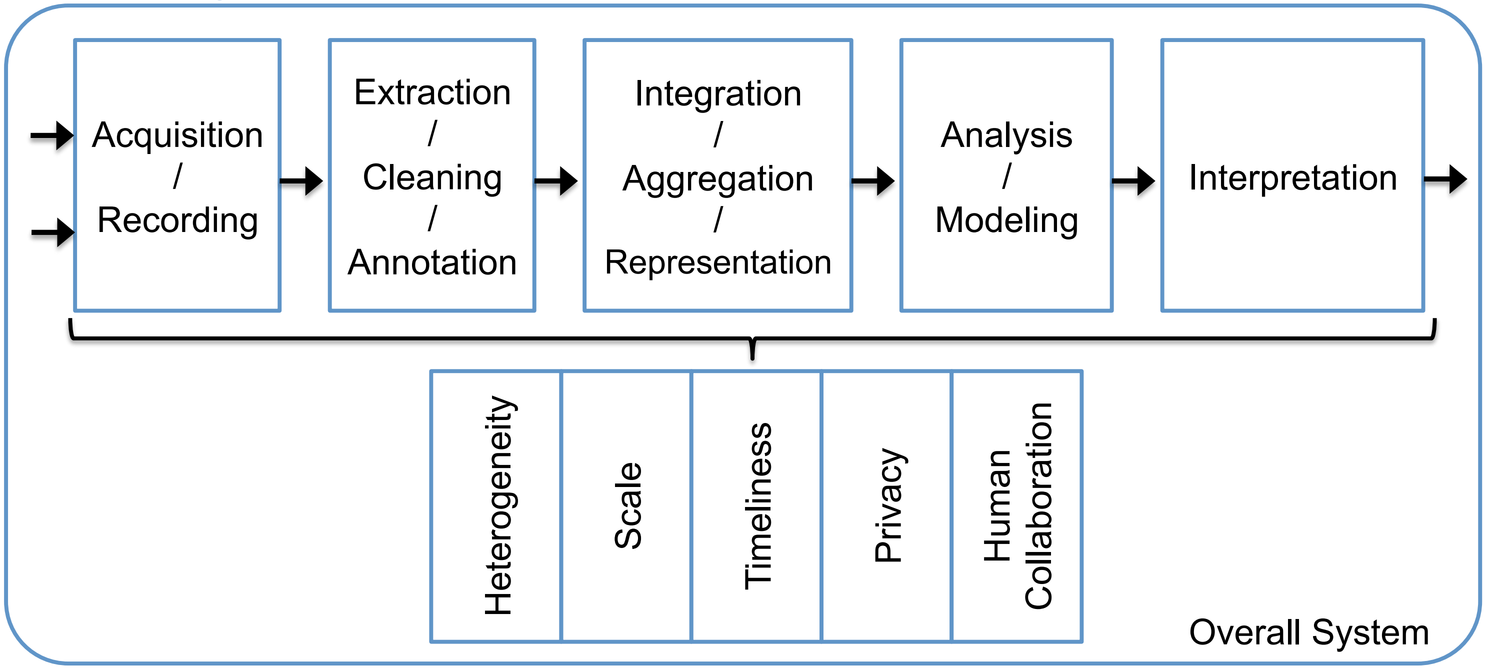
Cynthia Rudin

Machine Learning Course, Duke

Historical Notes

- Term “Big Data” coined by astronomers Cox and Ellsworth in 1997

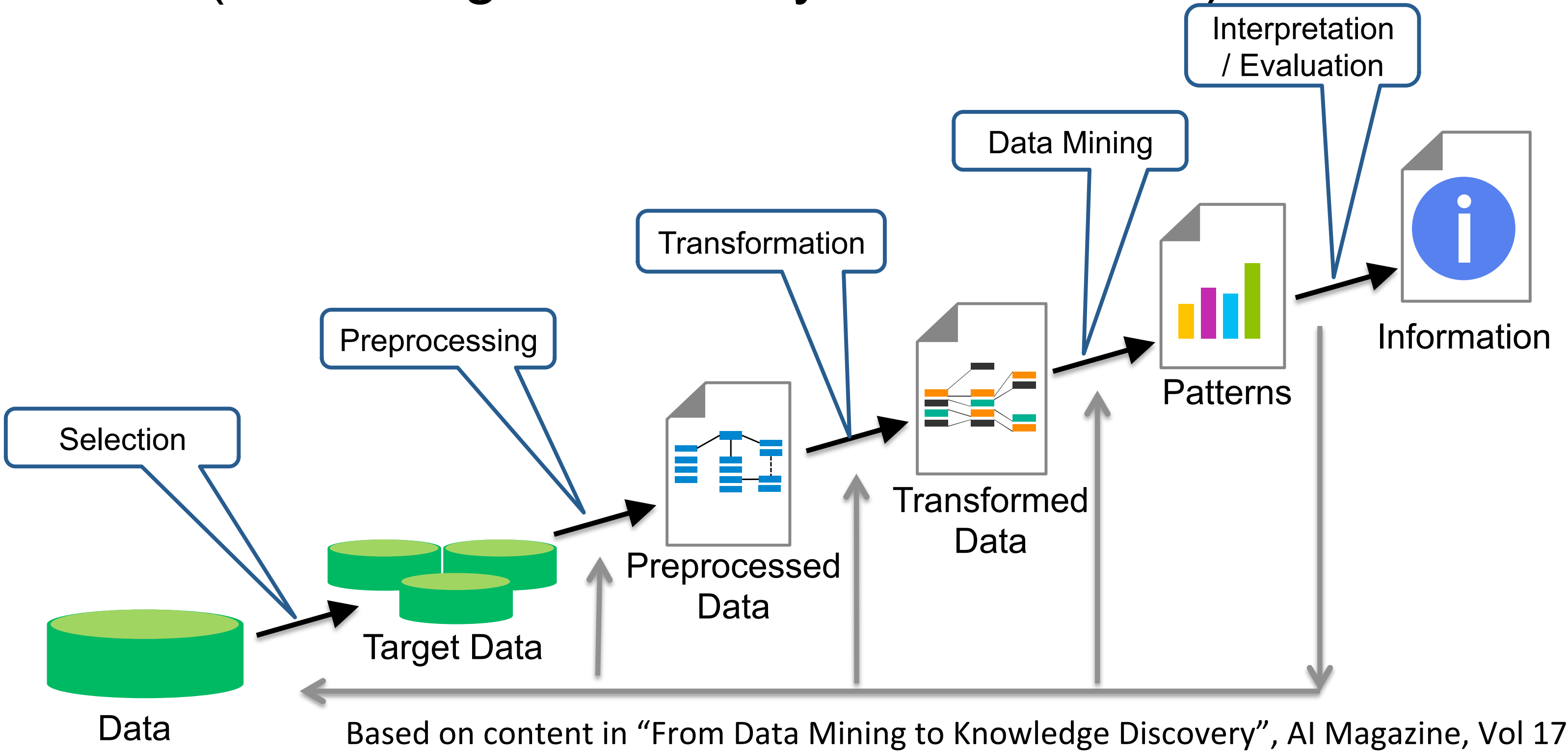
CCC Big Data Pipeline from 2012*



*From the Computing Community Consortium Big Data Whitepaper: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

Historical Notes

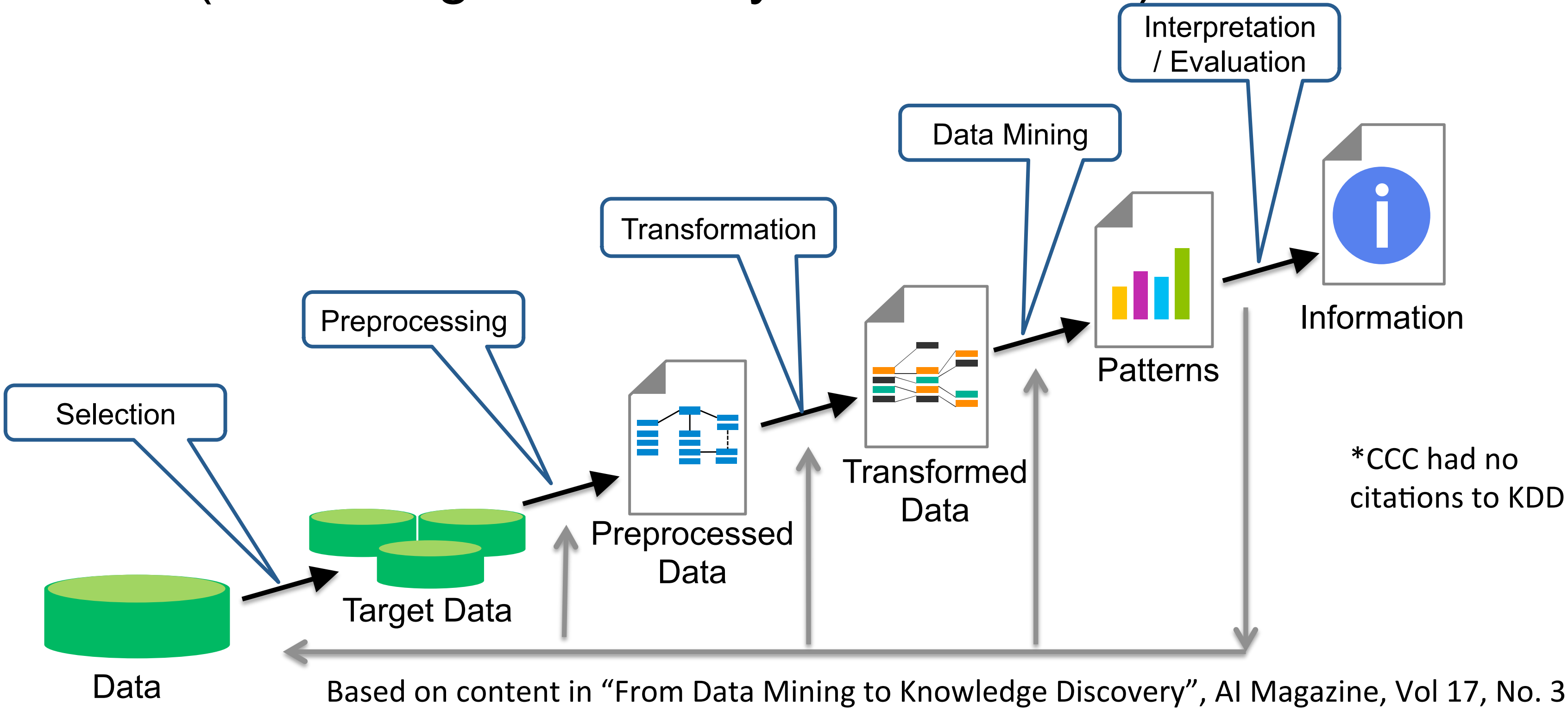
- KDD (Knowledge Discovery in Databases) Process



Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Historical Notes

- KDD (Knowledge Discovery in Databases) Process



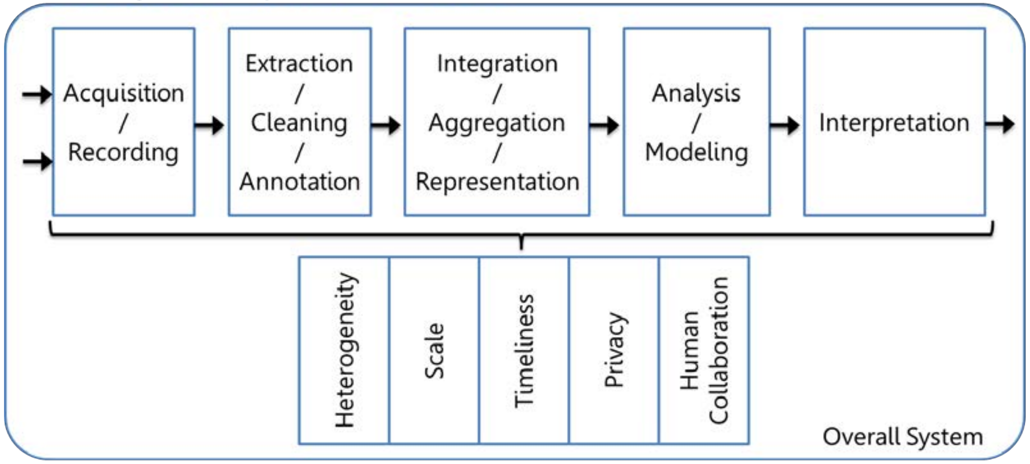
*CCC had no citations to KDD

1996

Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

CCC 2012

CCC Big Data Pipeline from 2012*



Acquisition / Recording

Extraction / Cleaning / Annotation

Integration / Aggregation / Representation

Analysis / Modeling

Interpretation

Heterogeneity

Scale

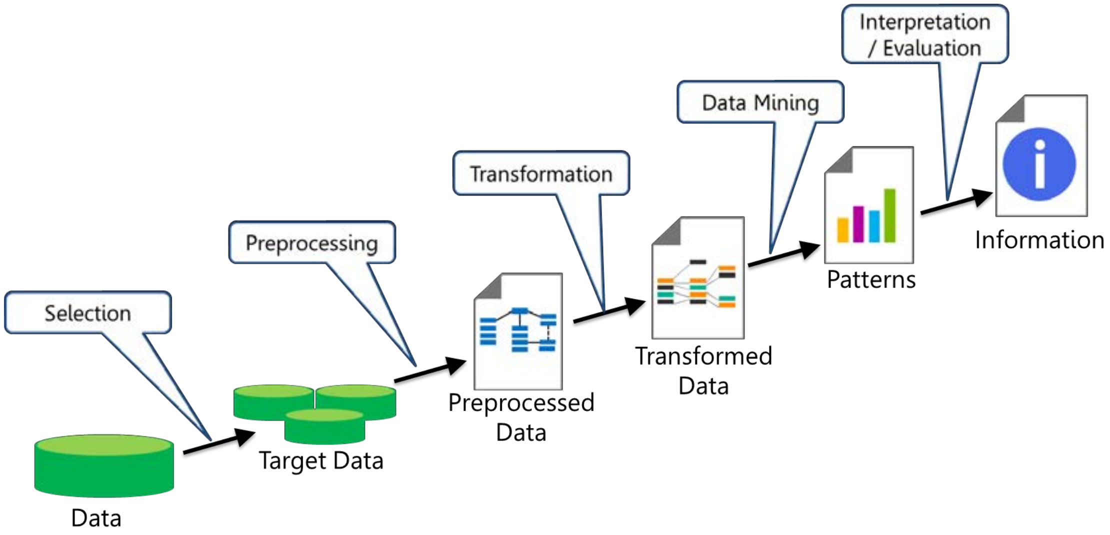
Timeliness

Privacy

Human Collaboration

Overall System

KDD 1996



Selection

Preprocessing

Transformation

Data Mining

Interpretation / Evaluation

Data

Target Data

Preprocessed Data

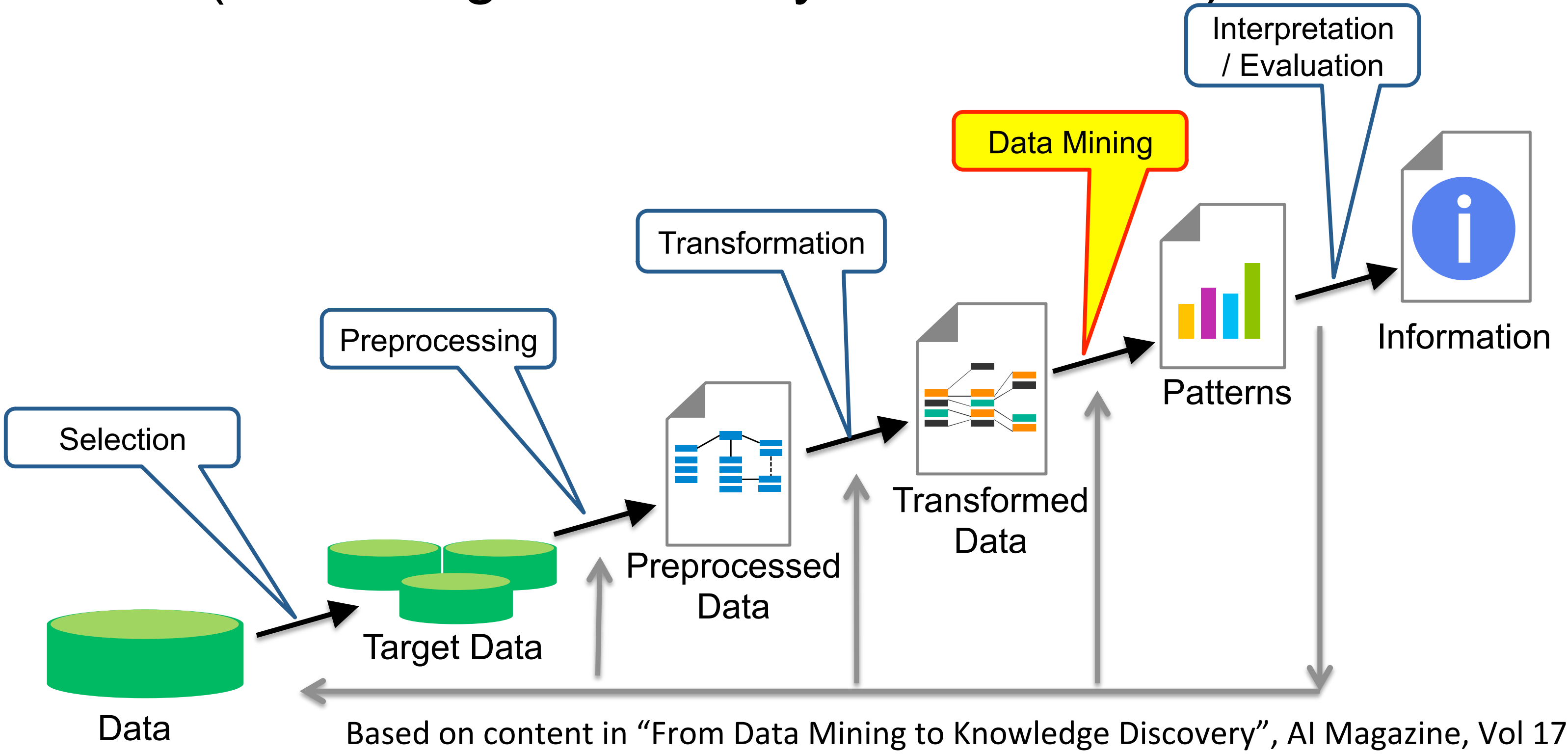
Transformed Data

Patterns

Information

Historical Notes

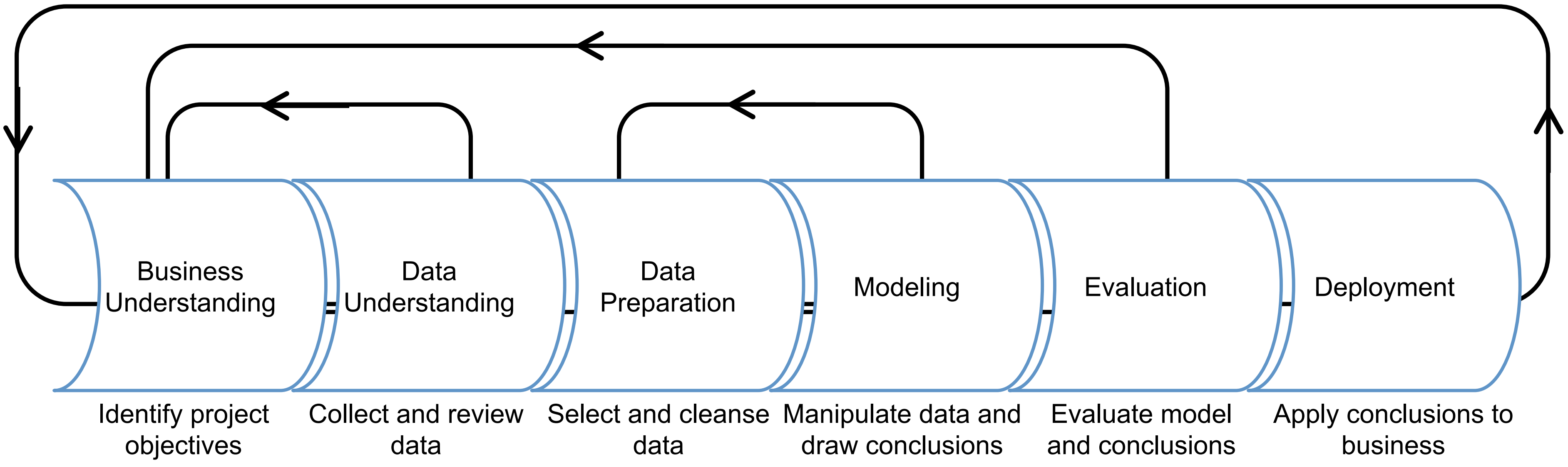
- KDD (Knowledge Discovery in Databases) Process



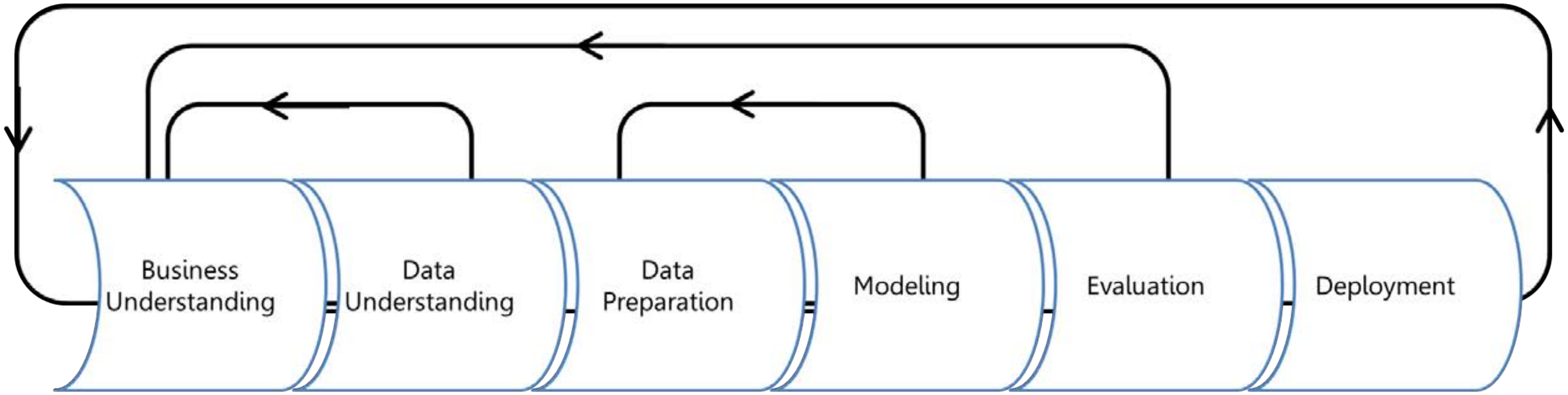
Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Historical Notes

CRoss Industry Standard Process for Data Mining (CRISP-DM)



From 2000, 77 pages



Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

Identify project objectives

Collect and review data

Select and cleanse data

Manipulate data and draw conclusions

Evaluate model and conclusions

Apply conclusions to business

Historical Notes

- The stages are basically the same no matter who invents or reinvents the (knowledge discovery / data mining / big data / data science) process. You may not always need all the stages.
- Data science is an iterative process.
  - Backwards arrows on most process diagrams.