

# Overview of Data Science Process

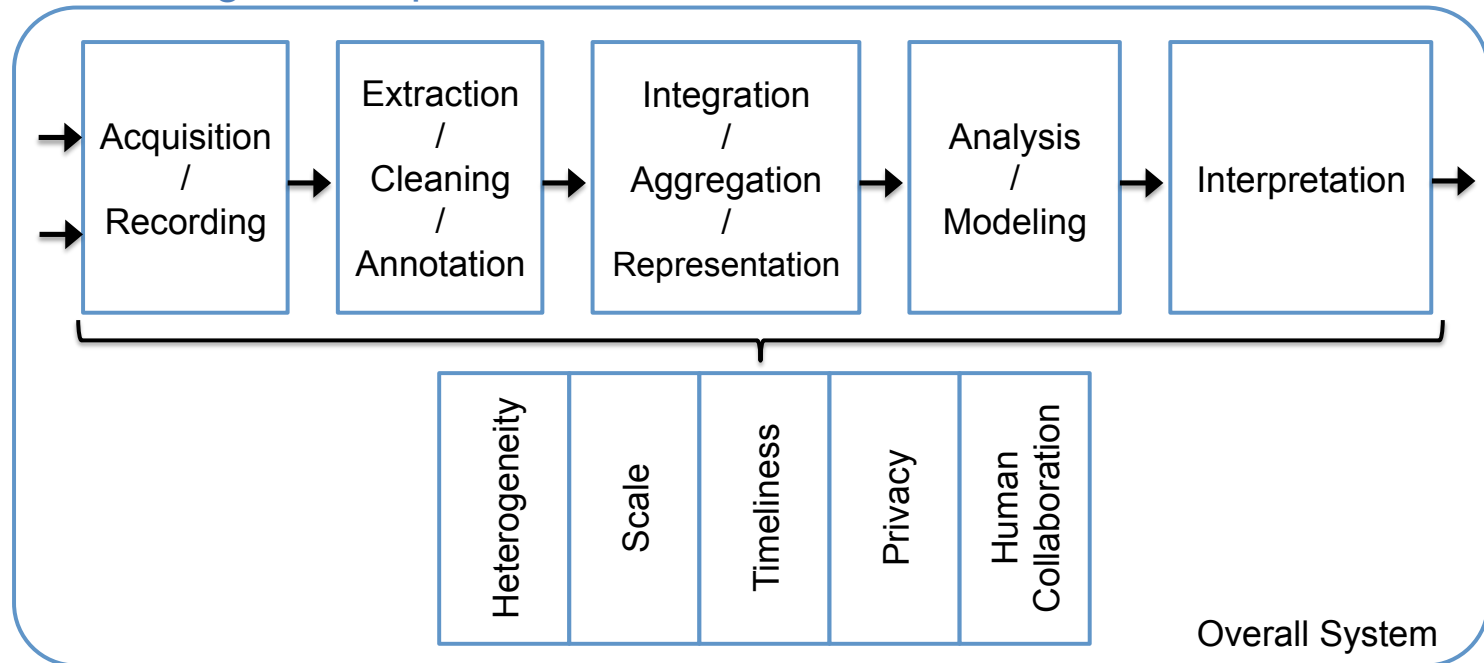
Cynthia Rudin

Machine Learning Course, Duke

# Historical Notes

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

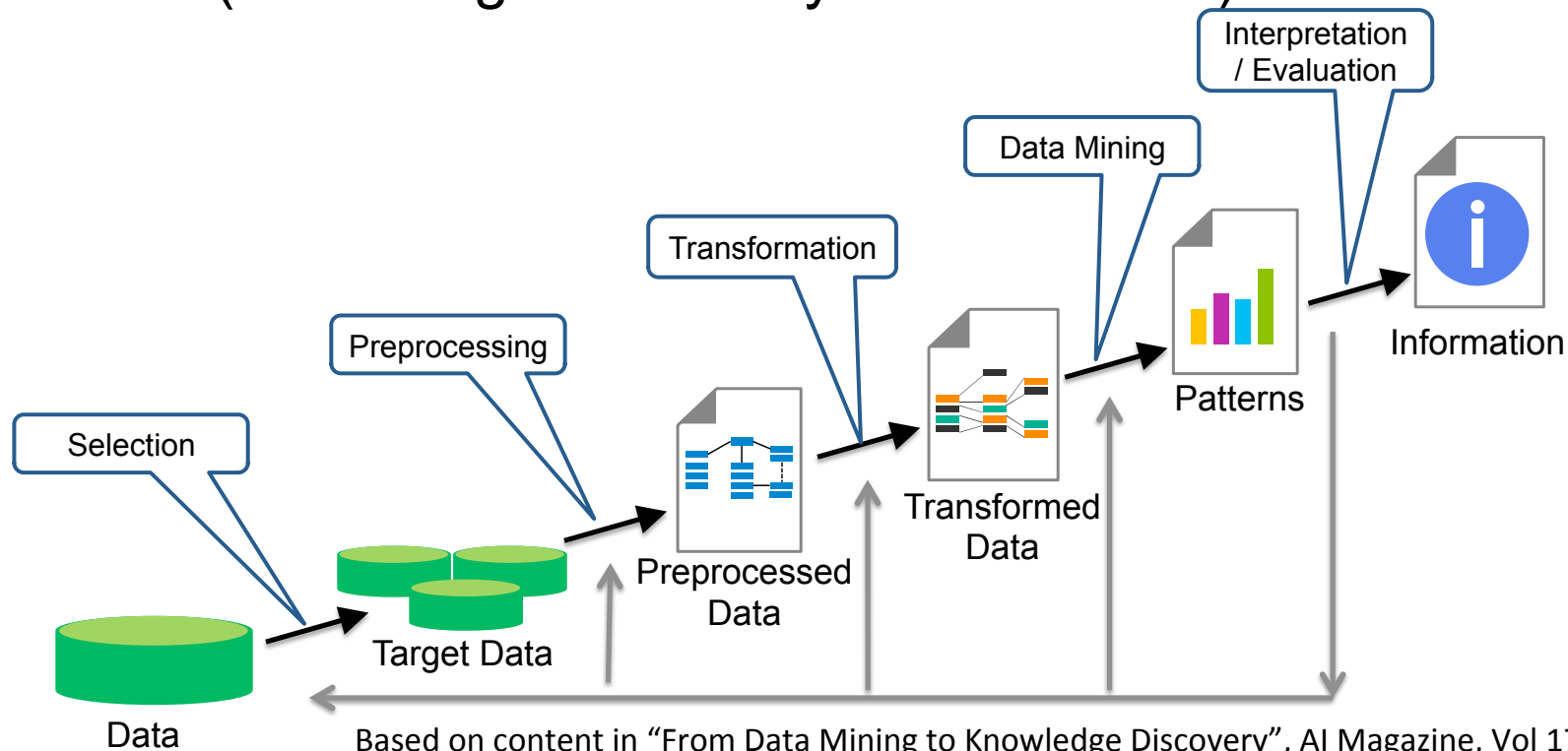
## CCC Big Data Pipeline from 2012\*



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

# Historical Notes

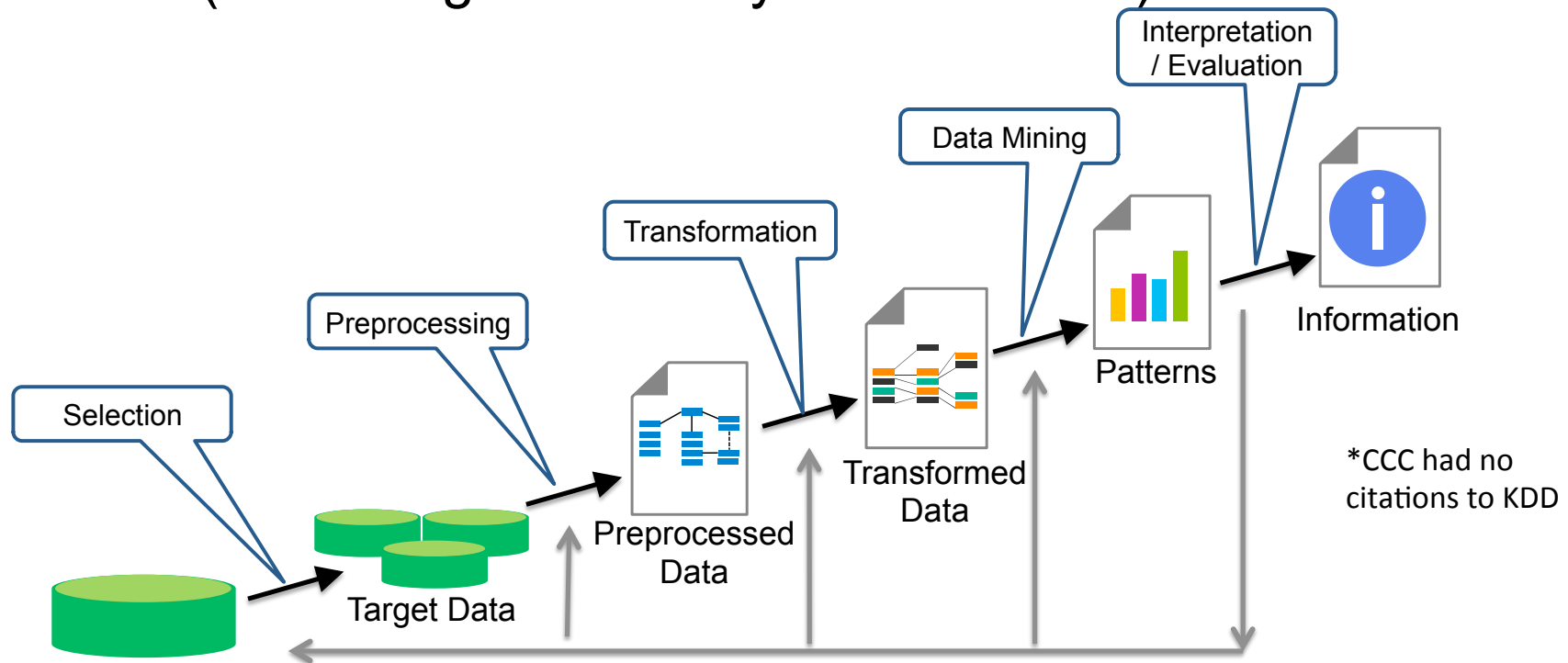
- KDD (Knowledge Discovery in Databases) Process



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

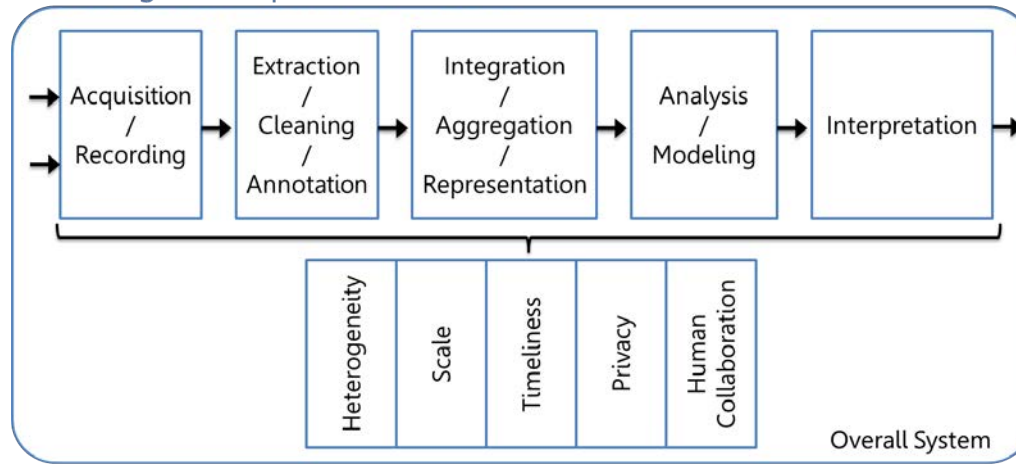
- KDD (Knowledge Discovery in Databases) Process



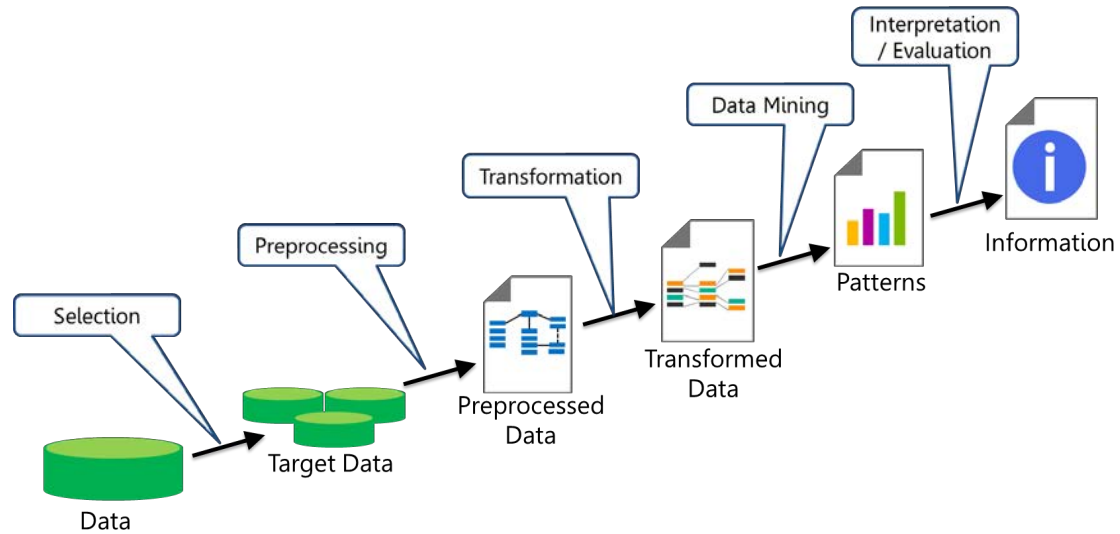
Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)  
<http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>

CCC 2012

CCC Big Data Pipeline from 2012\*

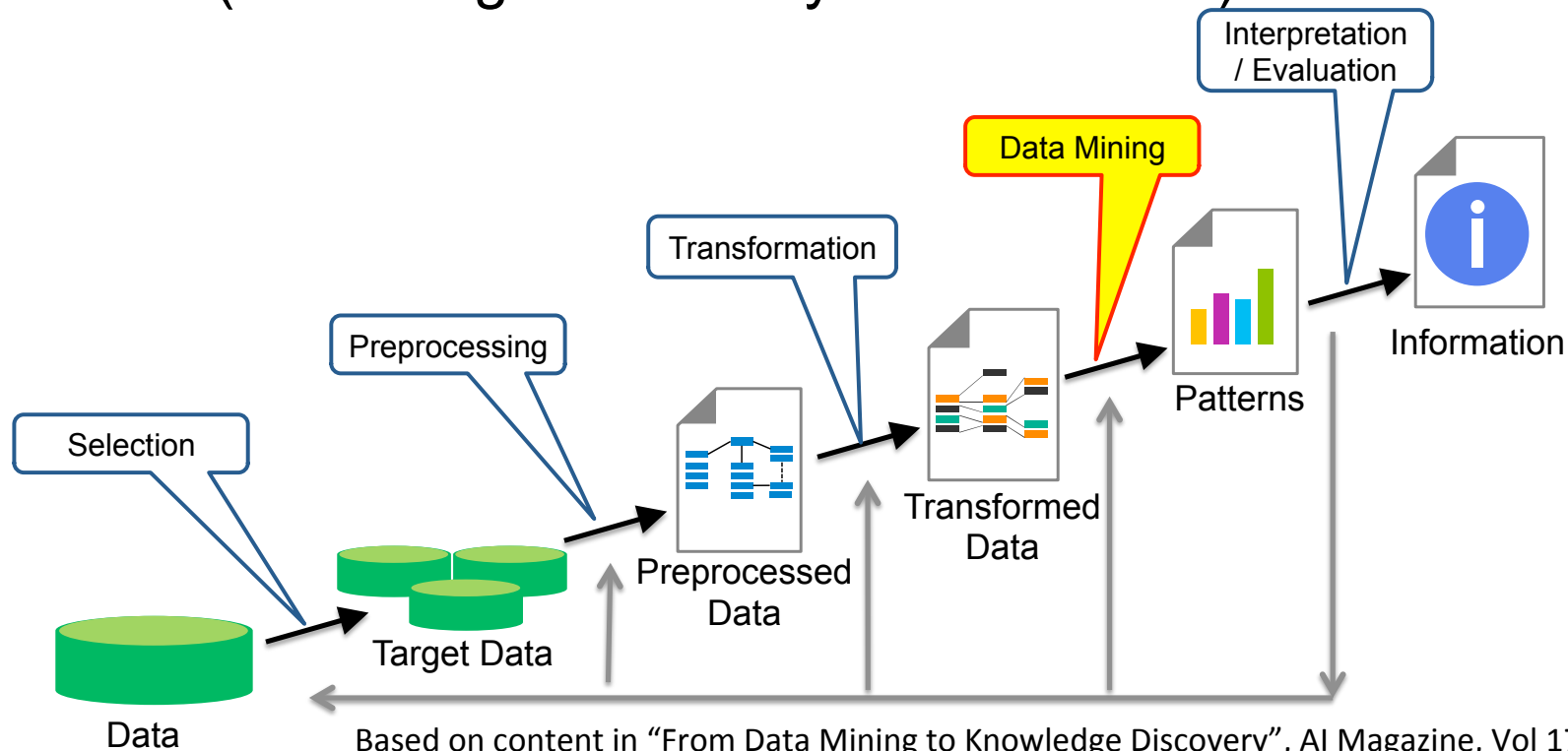


KDD 1996



# Historical Notes

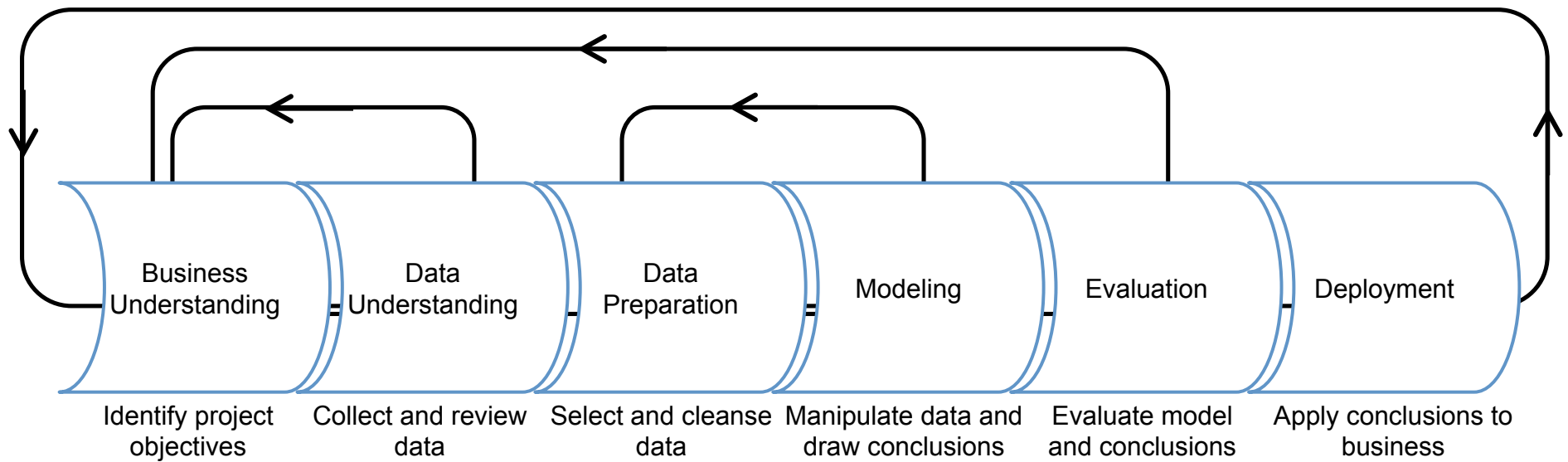
- KDD (Knowledge Discovery in Databases) Process



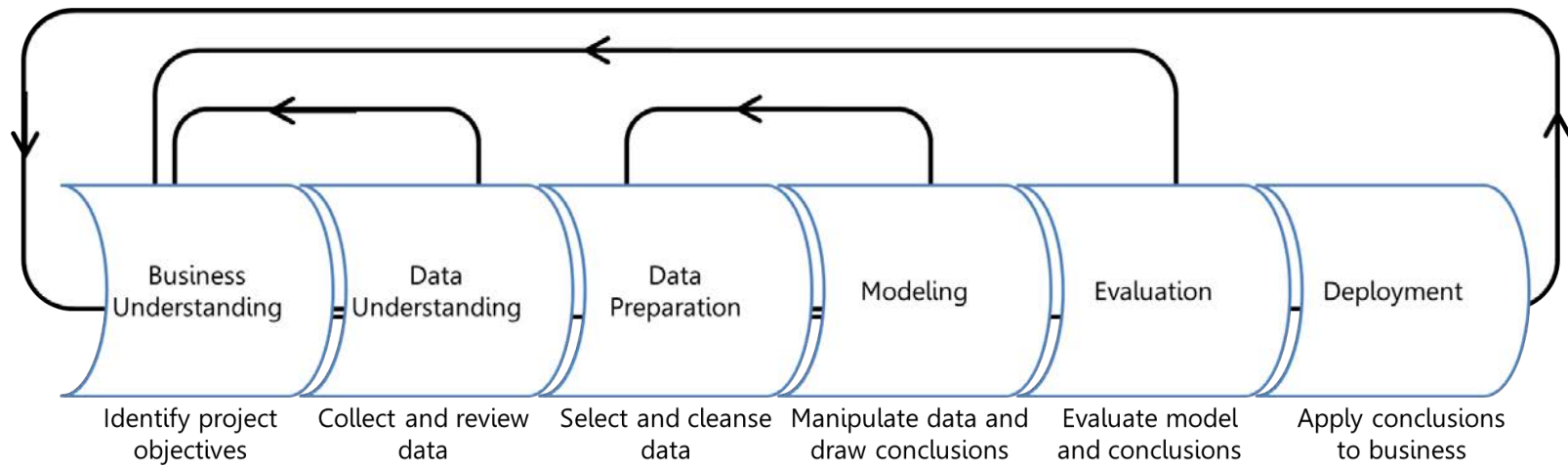
Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)  
<http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>

# Historical Notes

## Cross Industry Standard Process for Data Mining (CRISP-DM)



From 2000, 77 pages





# 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.