Data Science using Python 101

Lunch and Learn agenda for today

- Outline of sessions to come
- Session 1: Data Science process overview
  - Why have a data science process
  - KDD
  - CRISP-DM
  - Others
  - Common theme in all
  - How to apply that to RPG, Python, any analysis
  - Make CRISP-DM templates
  - CookieCutter for Data Science
  - Q&A
  - Summary of URLs

Session Outline

Lunch and Learn outline

- Session 1: Data Science Process Overview

- Session 2: Aquire and Prepare
  - Curl, requests, liburl to get data from the web
  - About JSON
  - Pandas
- Session 3: Analyze
  - Exploratory Data Analysis
  - Supervised, Unsupervised
  - Classifying, Regression
  - Clustering, Outlier Detection
  - NLP
  - Model Evaluation

Lunch and Learn outline

- Session 4: Act
  - Predicting as part of a decision making process
  - Production considerations

- Session 5: Tools and platforms
  - Anaconda
  - Sk-learn, H2O.ai
  - Knime, RapidMiner, Dataiku, DataRobot, etc.

Session 1

Data Science Process Overview

Why have a data science process?

[60 Minute Segment: Deception at Duke]

From 4:44 to 5:21

Why did it fail and what can be done

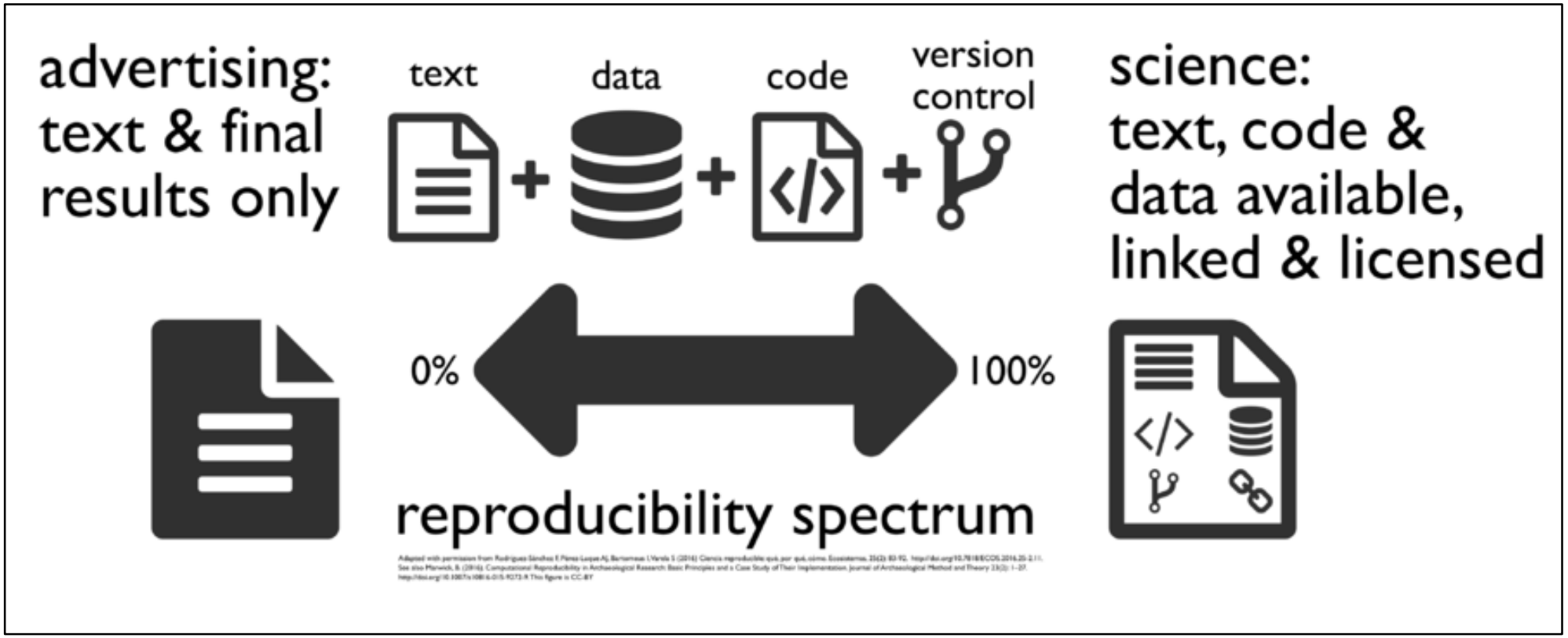
http://www.cbsnews.com/news/deception-at-duke-fraud-in-cancer-care/2/

Some Conclusion

- Verify work of others
- Others need to trust / understand the steps in your science
- Separation of roles demands clarity
- Hard to get errors corrected, especially good news
- Data Management so others can see the raw data

Reproducibility

- Know what level you are required to achieve



Understandability

- Often data science is a team sport
- Can you pass your work on to your team mate?

```
from itertools import permutations
n = 8
cols = range(n)
for vec in permutations(cols):
    if (n == len(set(vec[i]+i for i in cols))
          == len(set(vec[i]-i for i in cols))):
       print ("\n".join('#' * i + 'Q' + '#' * (n-i-1) for i in vec) + '\n')
```

- Can anyone tell me what this very pythonic piece of code does?

Efficiency / Performance

Sentiment Analysis Project

- Download 2M Facebook updates
- Remove non-English
- Filter related to our brands
- Apply sentiment analysis

Key Opinion Leader Project

- Download 2M Facebook updates
- Remove non-English
- Filter related to our brands
- Create social network of influencers

Manageability of resources and budget

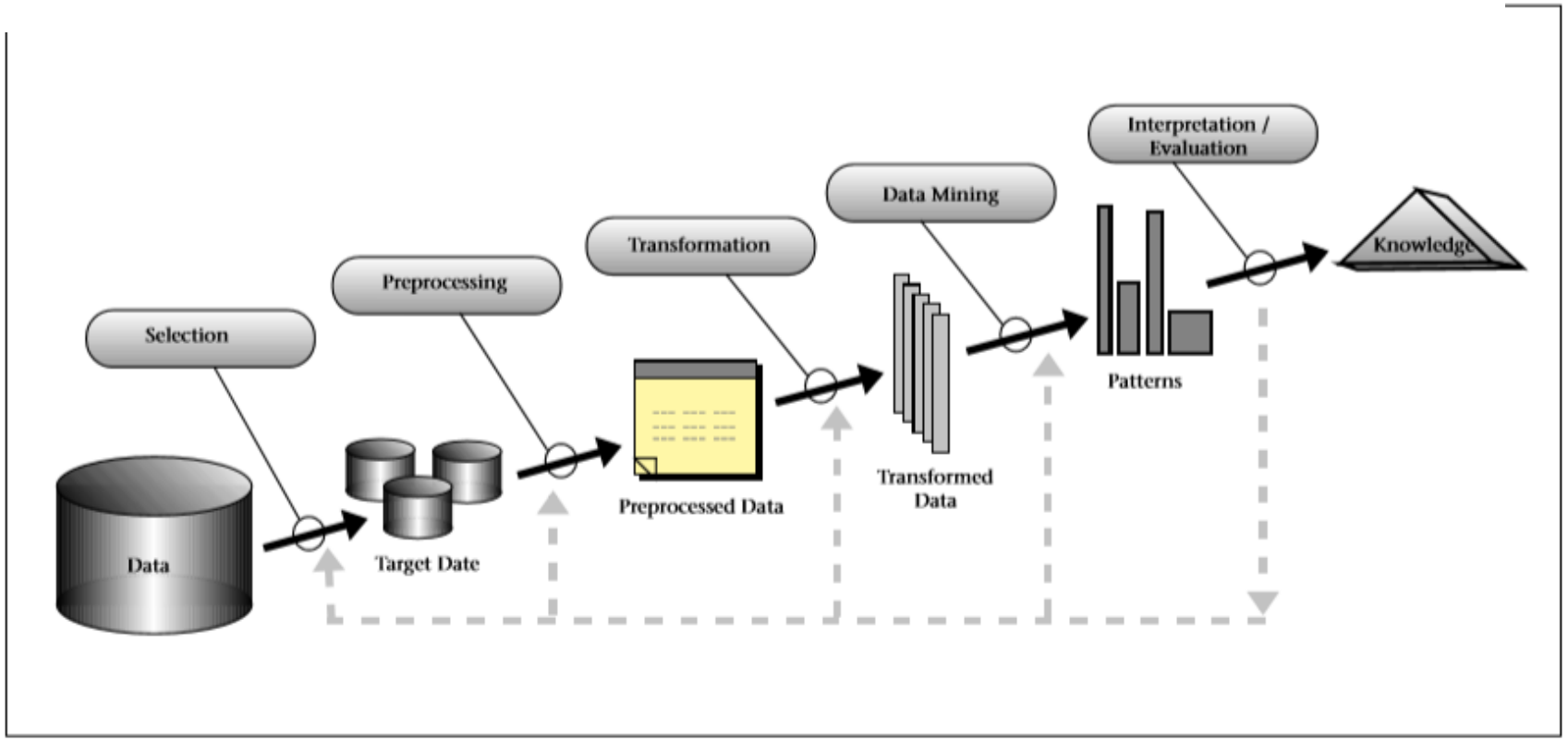
- Don’t need PhD to write SQL extract
- Need PhD to validate statistical significance of moving from normal to a gamma distribution for a variable
- Some tasks / resources can be applied to multiple projects at the same time, others can’t

- Work breakdown estimates rely on repeated tasks
- Smallest task of 2 days or less
- Standard process leads to comparable project estimates
- Ability to set and deliver to expectations

KDD

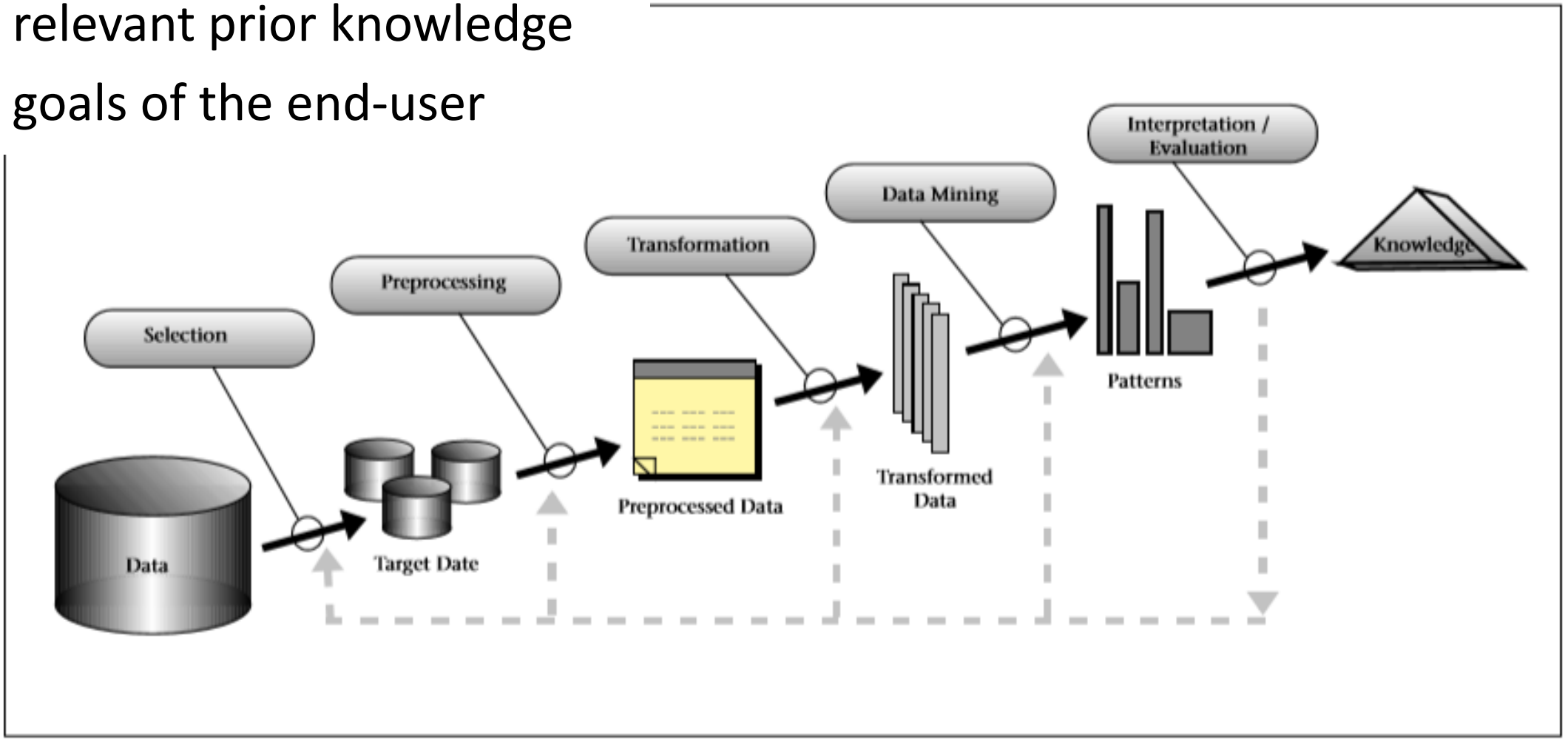
KDD - Knowledge Discovery and Data Mining

- ACM – SIGKDD formed in 1995.
- Evolved from conference series of AAAI
  (Association for the Advancement of Artificial Intelligence)
- http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf
- 9 Steps



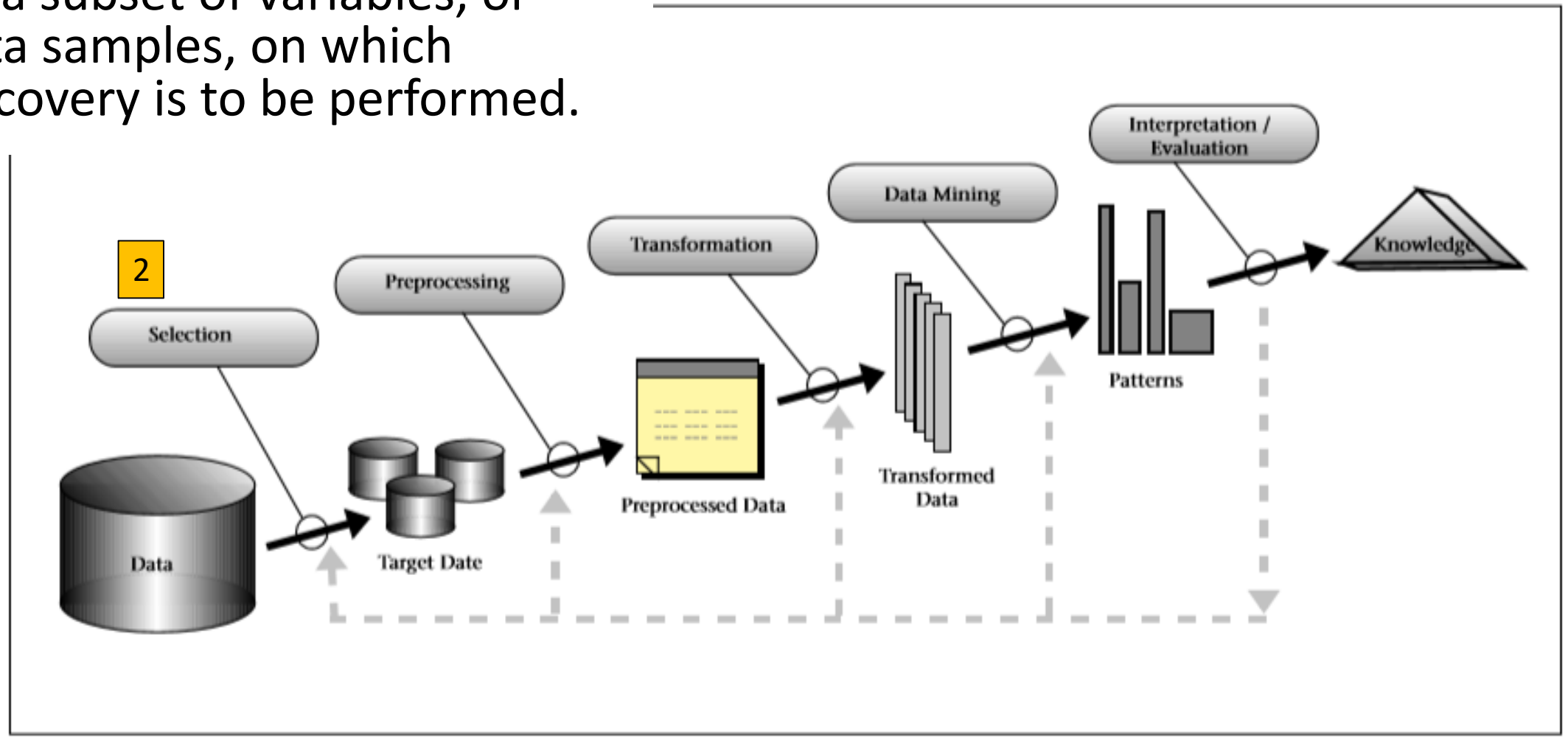
1. Developing an understanding

- the application domain
- the relevant prior knowledge
- the goals of the end-user



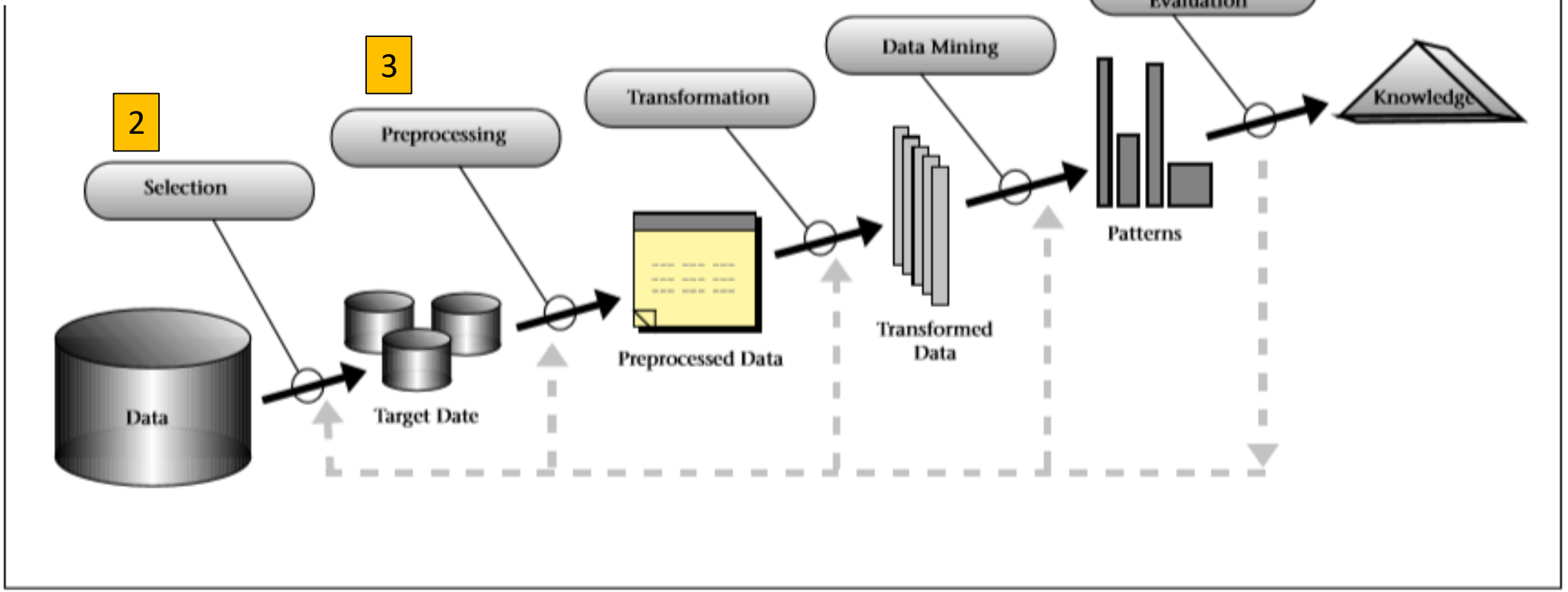
2. Creating a target data set:

- selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.



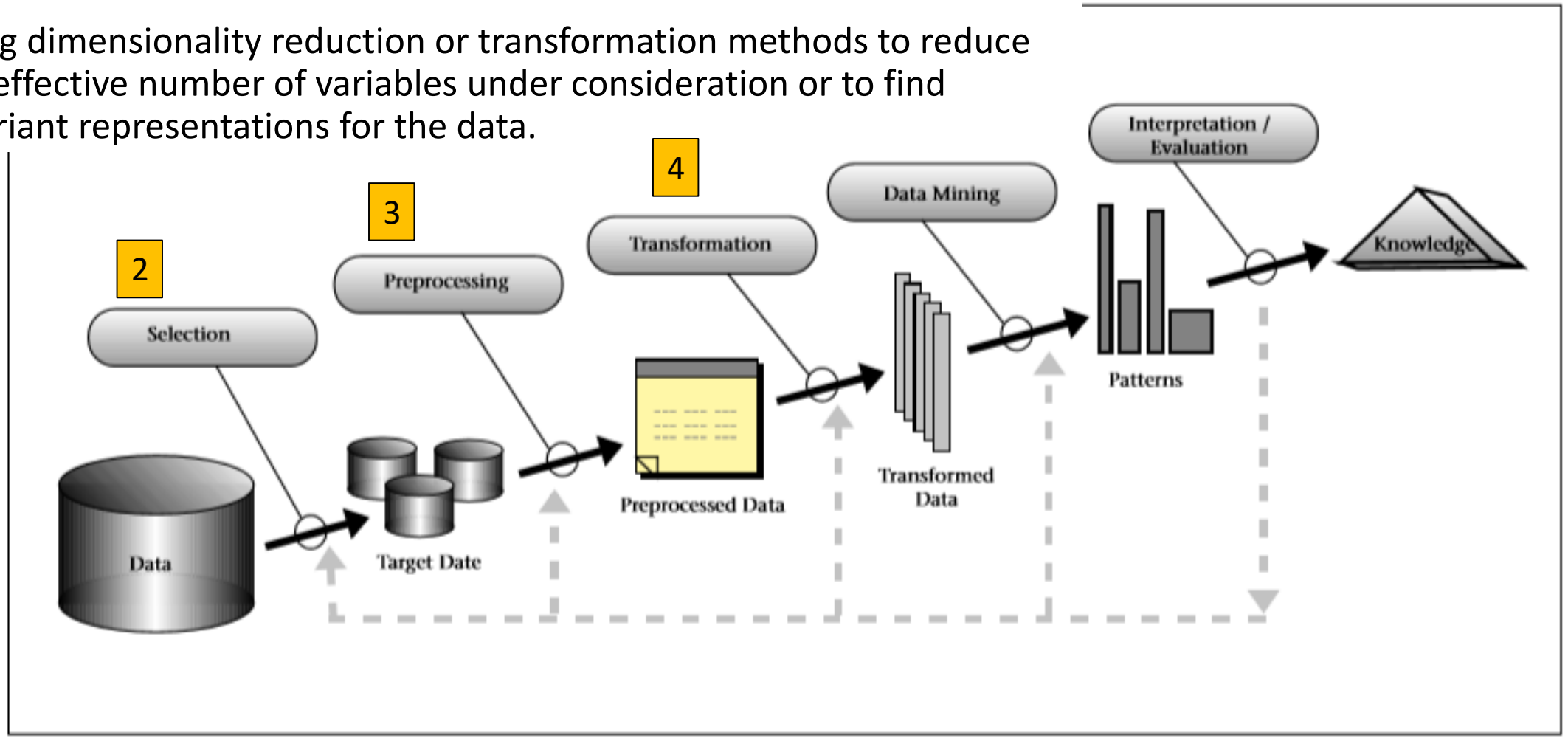
3. Data cleaning and preprocessing.

- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- Strategies for handling missing data fields.
- Accounting for time sequence information and known changes.



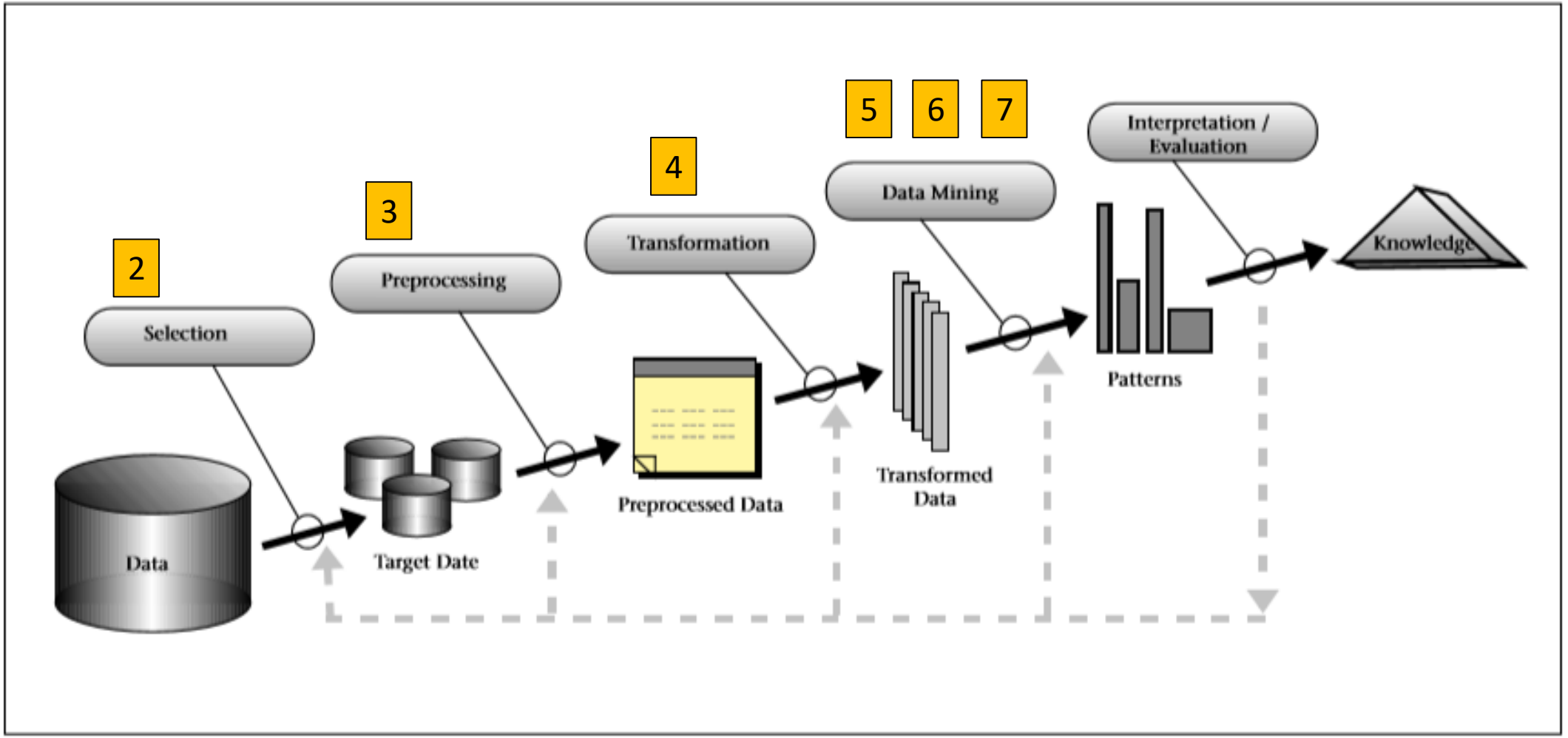
4. Data reduction and projection.

- Finding useful features to represent the data depending on the goal of the task.
- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

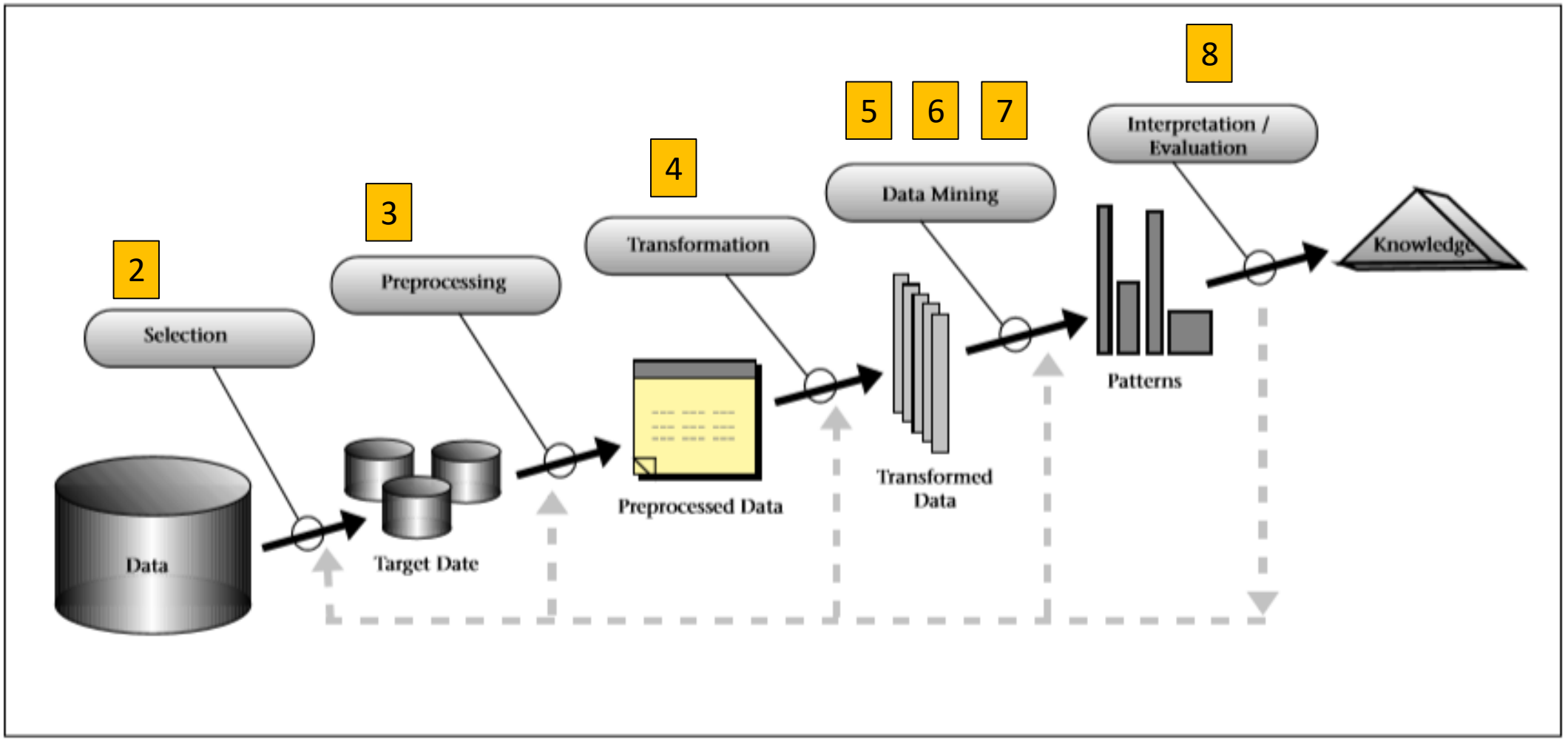


5. Choosing the data mining task.
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm(s).
   - Selecting method(s) to use for searching for patterns in data.
   - Deciding which models and parameters may be appropriate.
   - Matching a particular data mining method with the overall criteria of the KDD process.
7. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

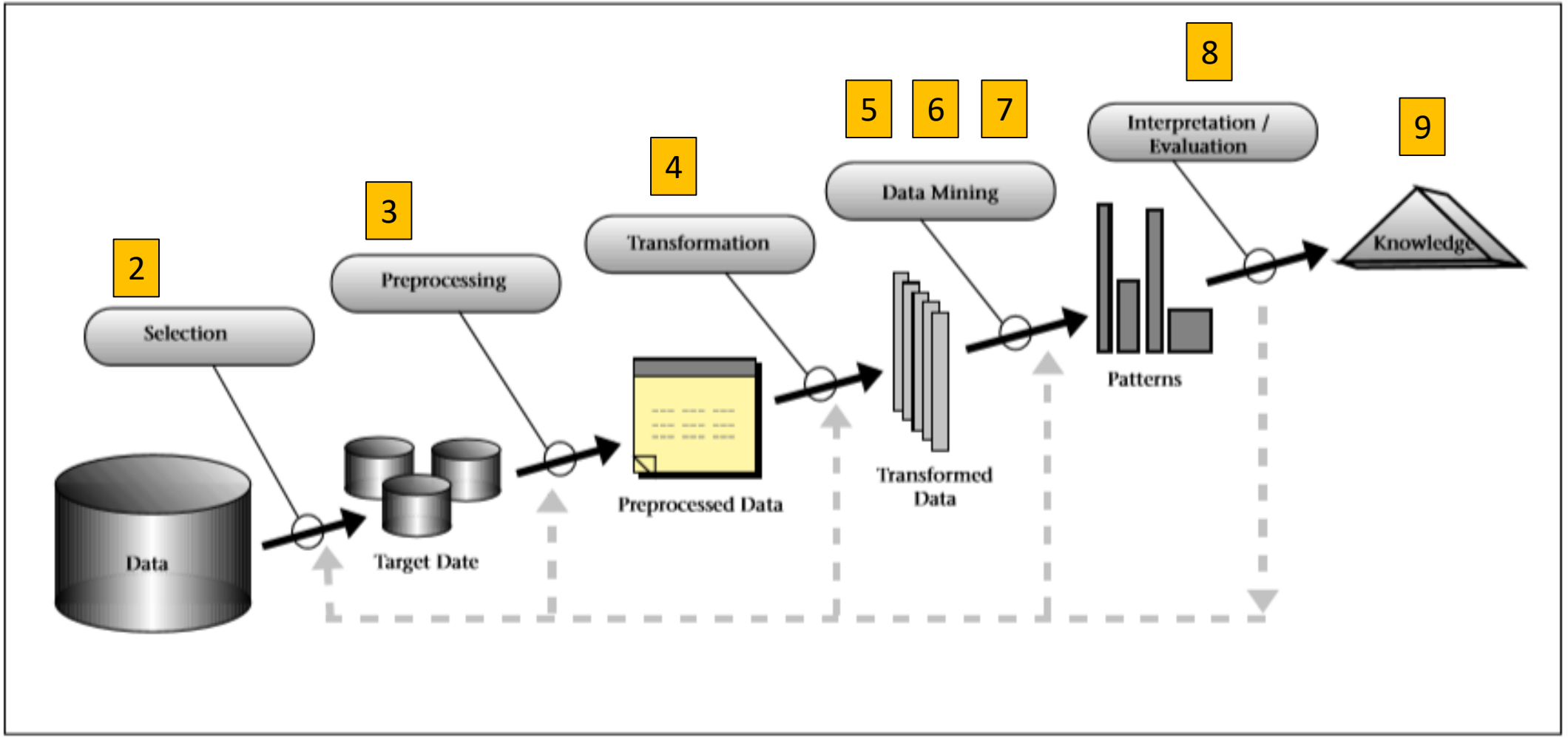
5. 6. 7. Data Mining



8. Interpreting mined patterns.



9. Consolidating discovered knowledge.



Often referenced, not much support

- SIGKDD is active SIG of ACM – I'm a member myself
- KDD Nuggets is great resource
- KDD Cup is yearly prestigious data science competition
- But
- Lack of additional documentation
- No tool support

CRISP-DM

CRISP-DM
Cross Industry Standard Process for Data Mining

- Originated in 1996 with Daimler-Chrysler, SPSS and NCR
- Became a SIG in 1997
- Many supporters
- CRISP-DM 1.0 released in 2000
- Support through UI of SPSS tool
- CRISP-DM v1.0

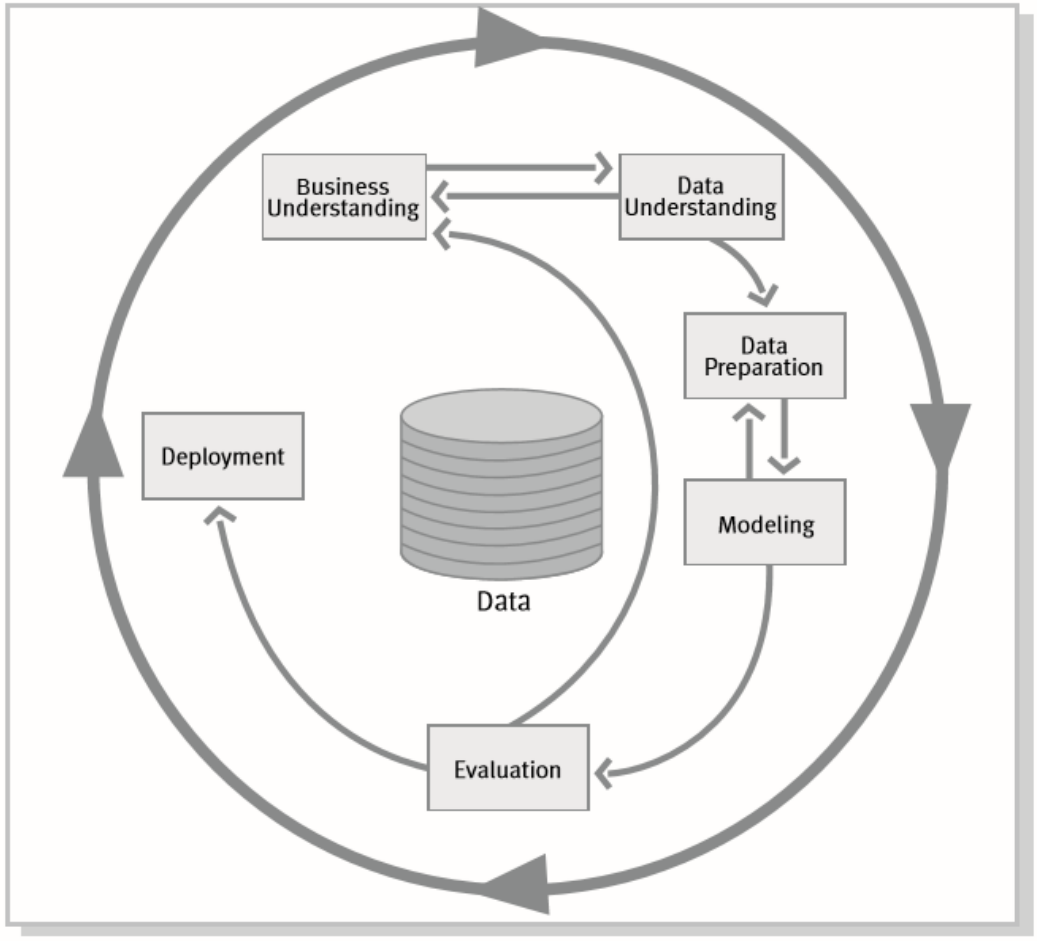


Figure 2: Phases of the CRISP-DM reference model

Business Understanding

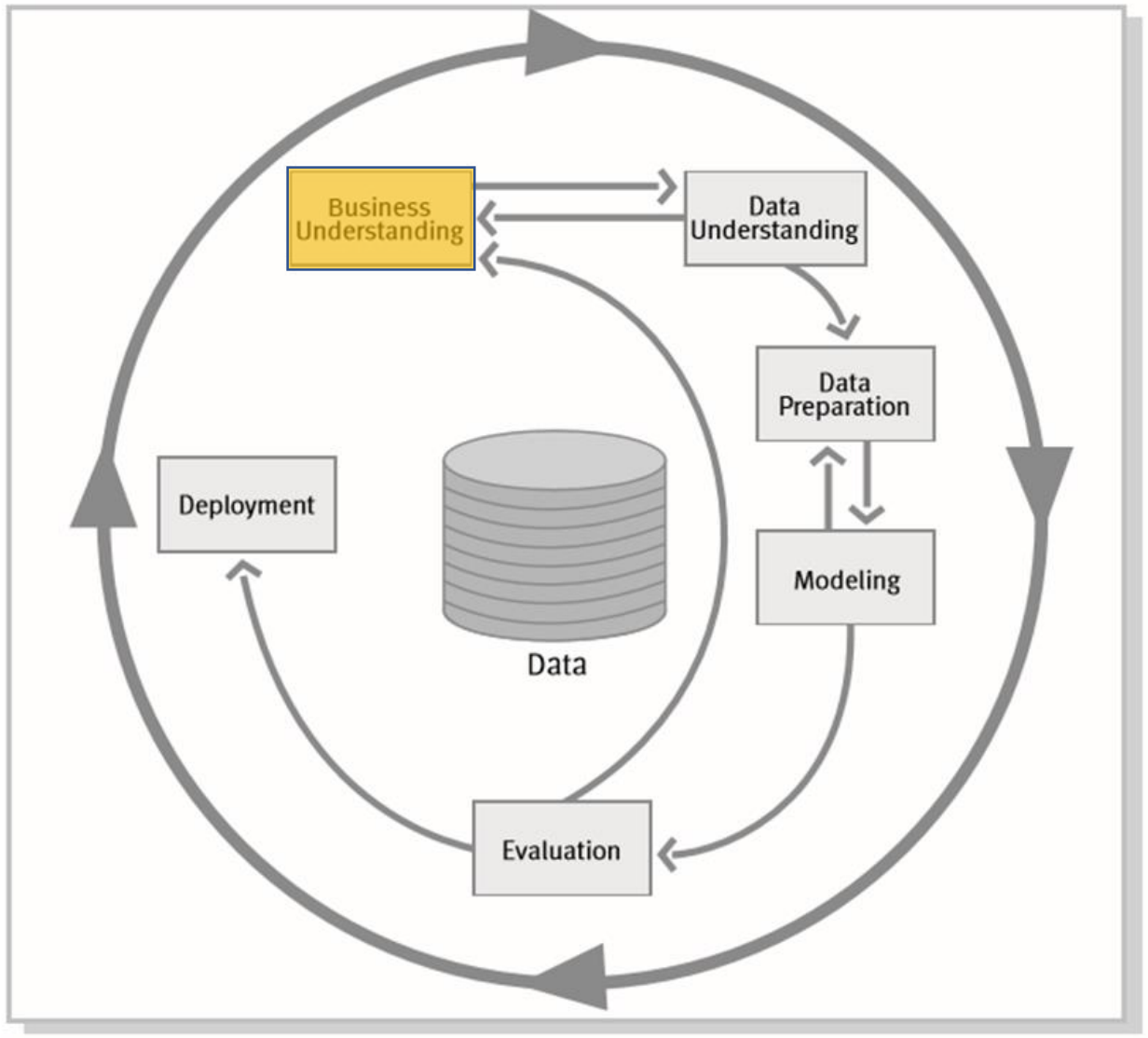


Figure 2: Phases of the CRISP-DM reference model

- This initial phase focuses on understanding the project objectives and requirements from a business perspective,
- then converting this knowledge into a data mining problem definition and
- a preliminary plan designed to achieve the objectives.

Data Understanding

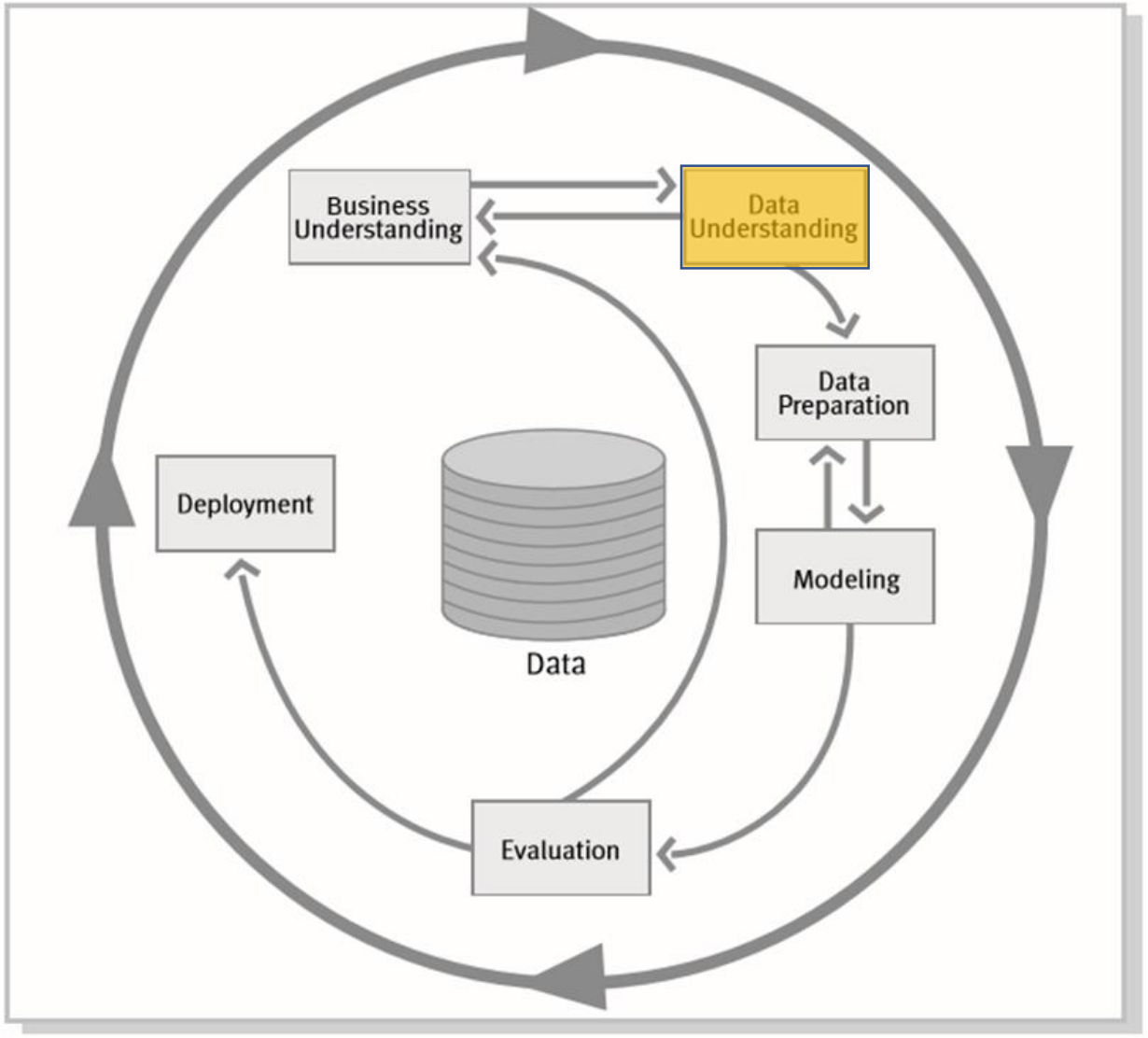


Figure 2: Phases of the CRISP-DM reference model

- Initial data collection
- Become familiar with the data
- Identify data quality problems
- Discover first insights into the data, and/or
- Detect interesting subsets to form hypotheses regarding hidden information.

Data Preparation

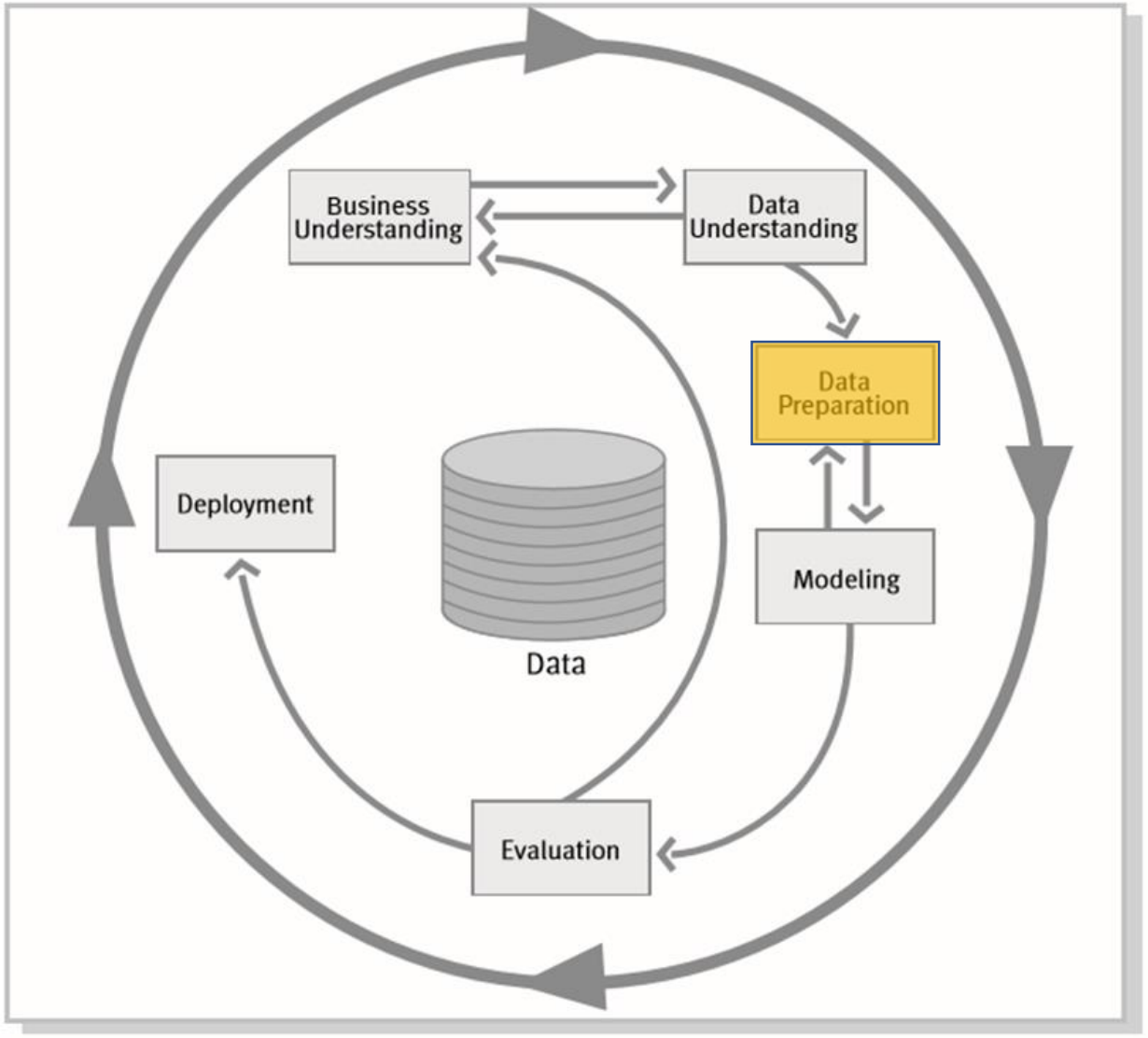


Figure 2: Phases of the CRISP-DM reference model

- All activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data.
- Data preparation tasks are likely to be performed multiple times and not in any prescribed order.
- Tasks include table, record, and attribute selection
- Transformation and cleaning of data for modeling tools

Modeling

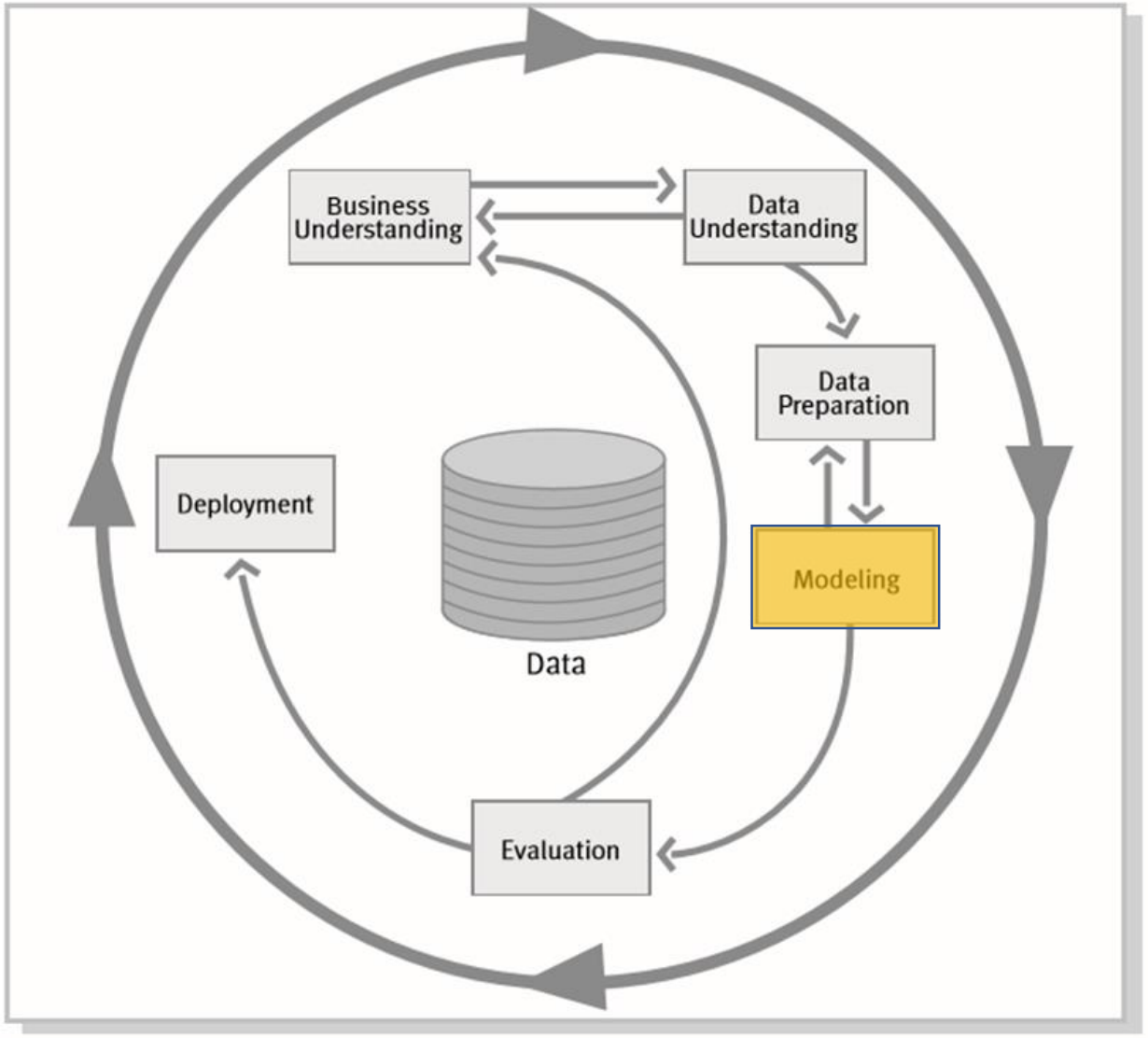


Figure 2: Phases of the CRISP-DM reference model

- Various modeling techniques are selected and applied
- Their parameters are calibrated to optimal values.
  - Typically, there are several techniques for the same data mining problem type.
  - Some techniques have specific requirements on the form of data.
  - Therefore, going back to the data preparation phase is often necessary

Evaluation

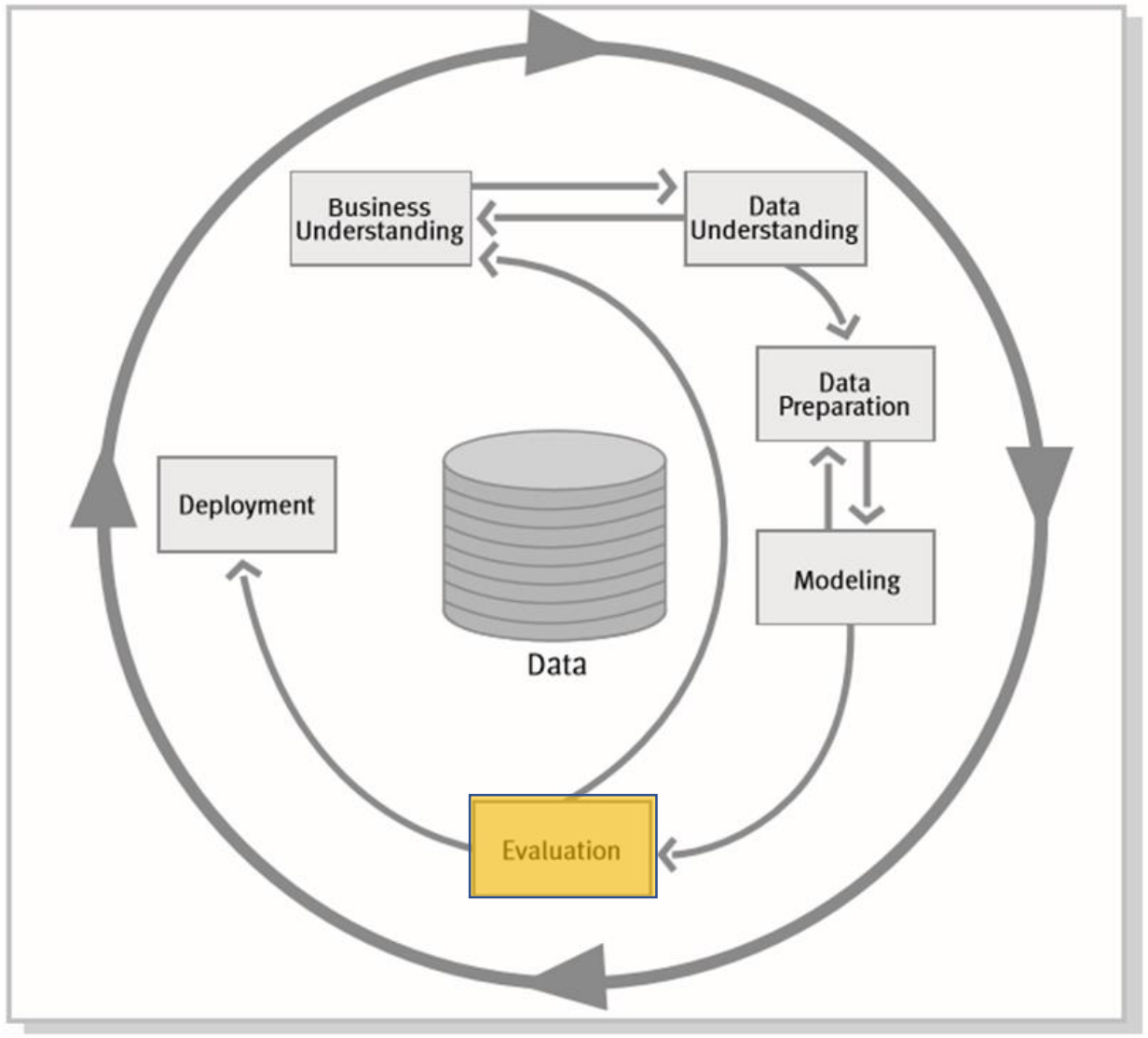


Figure 2: Phases of the CRISP-DM reference model

- At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective.
- Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives.
- A key objective is to determine if there is some important business issue that has not been sufficiently considered.
- At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

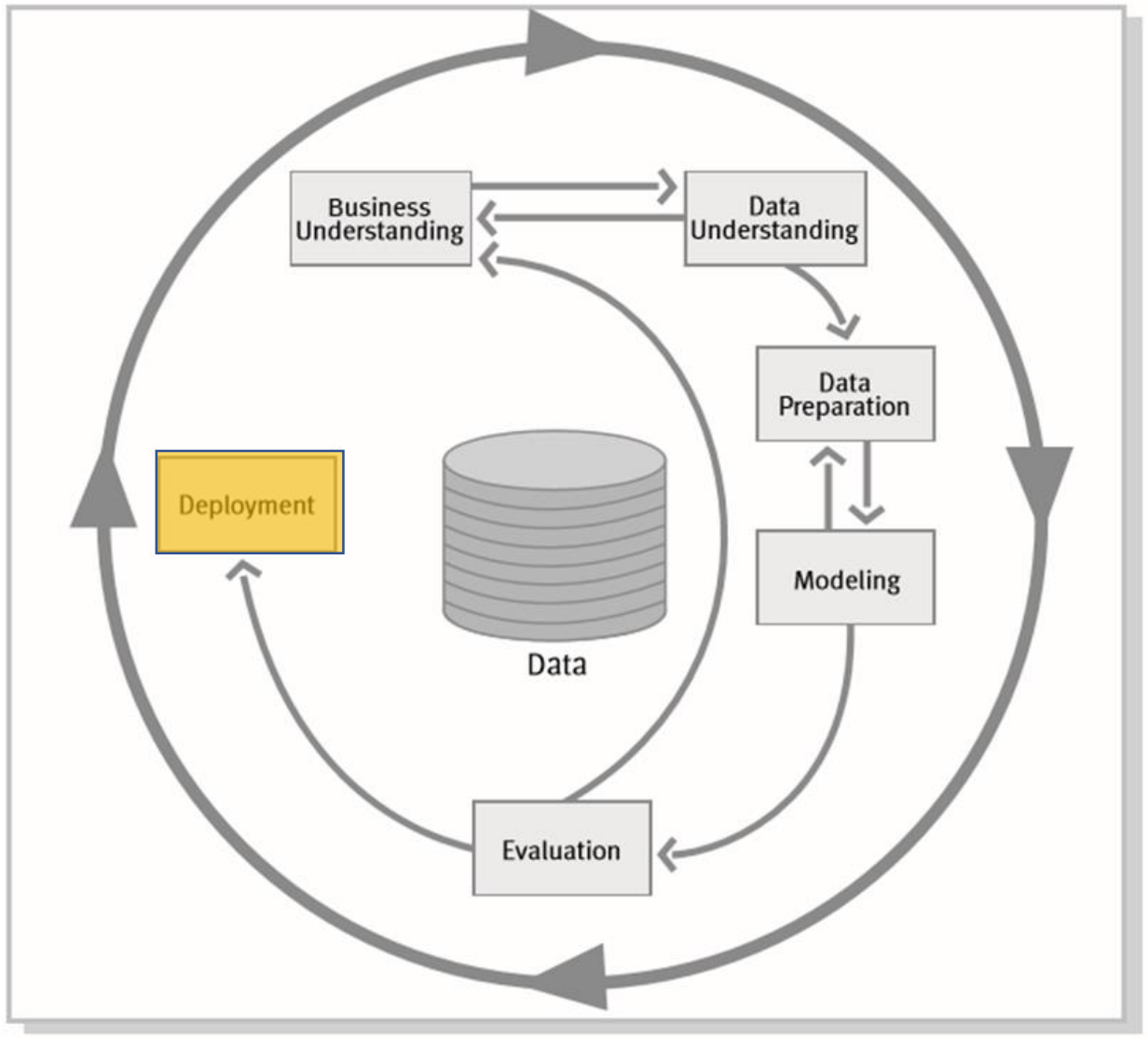
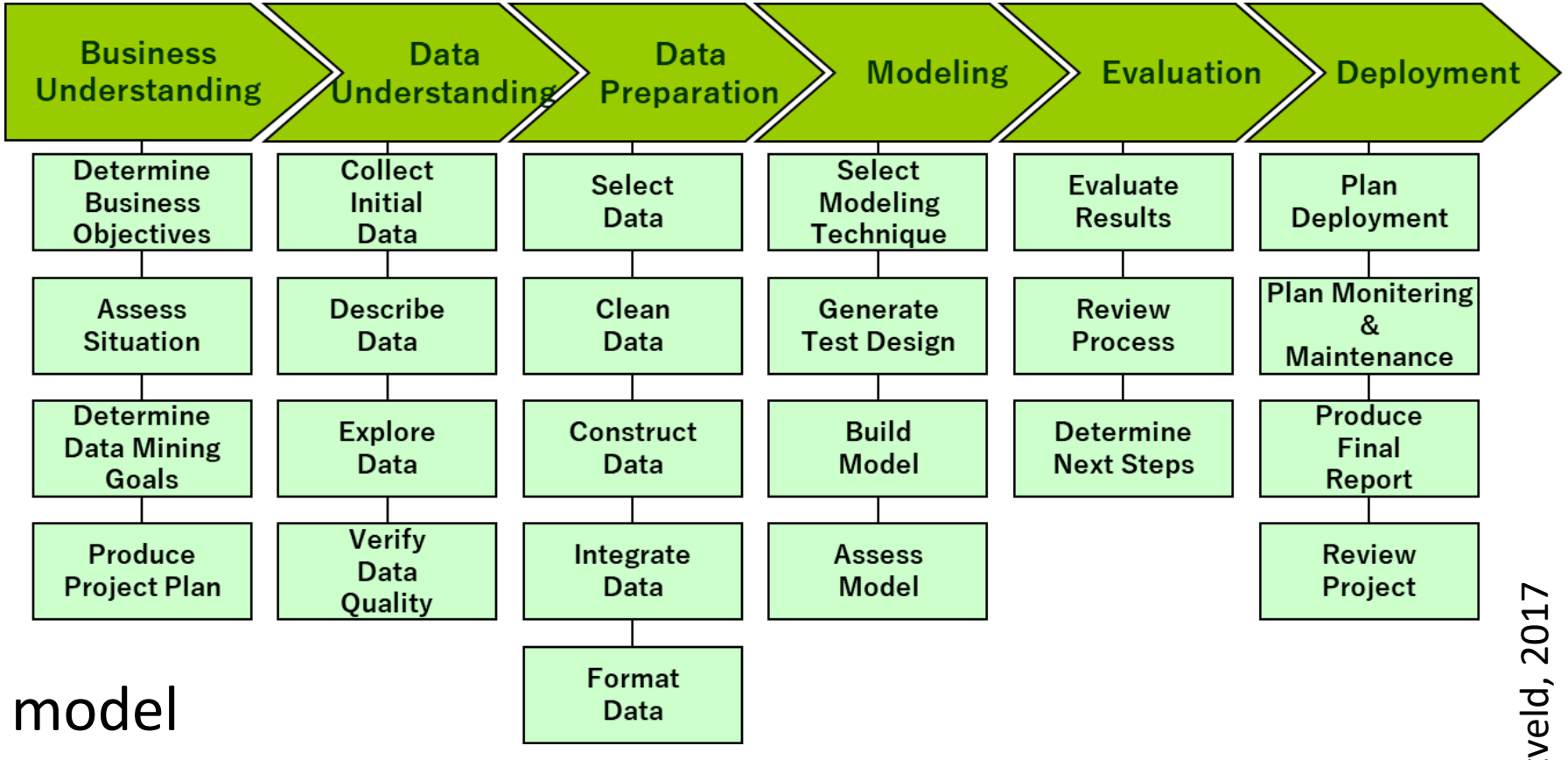


Figure 2: Phases of the CRISP-DM reference model

- Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.
- In many cases, it is the customer, not the data analyst, who carries out the deployment steps.
- However, even if the analyst will carry out the deployment effort, it is important for the customer to **understand up front what actions need to be carried out in order to actually make use of the created models.**

Six Phases Broken Down Into Tasks



- Detailed reference model
- Each task has one or more outputs defined
- User guide with Activities for each output
- Serves as template for documentation

Example Reference Model

1.3 Determine data mining goals

Task

Determine data mining goals

A business goal states objectives in business terminology. A data mining goal states project objectives in technical terms. For example, the business goal might be "Increase catalog sales to existing customers." A data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item."

Outputs

Data mining goals

Describe the intended outputs of the project that enable the achievement of the business objectives.

Data mining success criteria

Define the criteria for a successful outcome to the project in technical terms—for example, a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

Example User Guide

Activities	■ Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified). ■ Specify data mining problem type (e.g., classification, description, prediction, and clustering). For more details about data mining problem types, see Appendix 2.
Activities	■ Specify criteria for model assessment (e.g., model accuracy, performance and complexity) ■ Define benchmarks for evaluation criteria ■ Specify criteria which address subjective assessment criteria (e.g., model explain ability and data and marketing insight provided by the model)
Beware!	Remember that the data mining success criteria are different than the business success criteria defined earlier.

More useful. Most popular

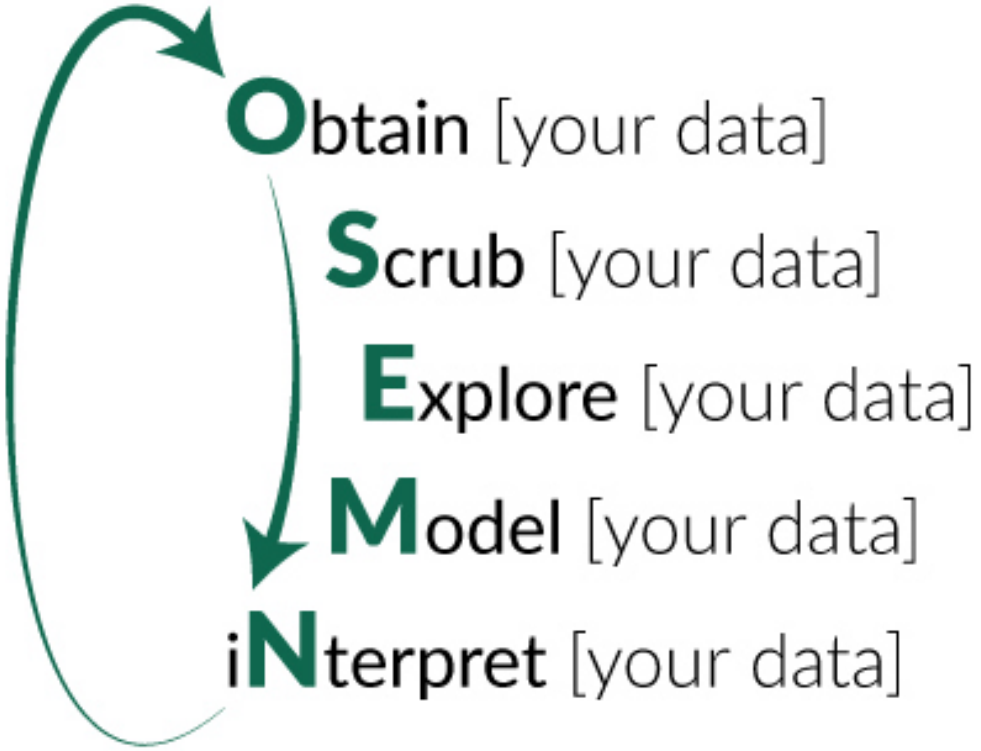
- Although Website has been off-line for years, still most popular
- Document can be used to create documentation templates
- Some tools specifically follow model

- Superseded in 2015 by ASUM - Analytics Solutions Unified Method
- ASUM however is licensed only to be used with IBM tools.

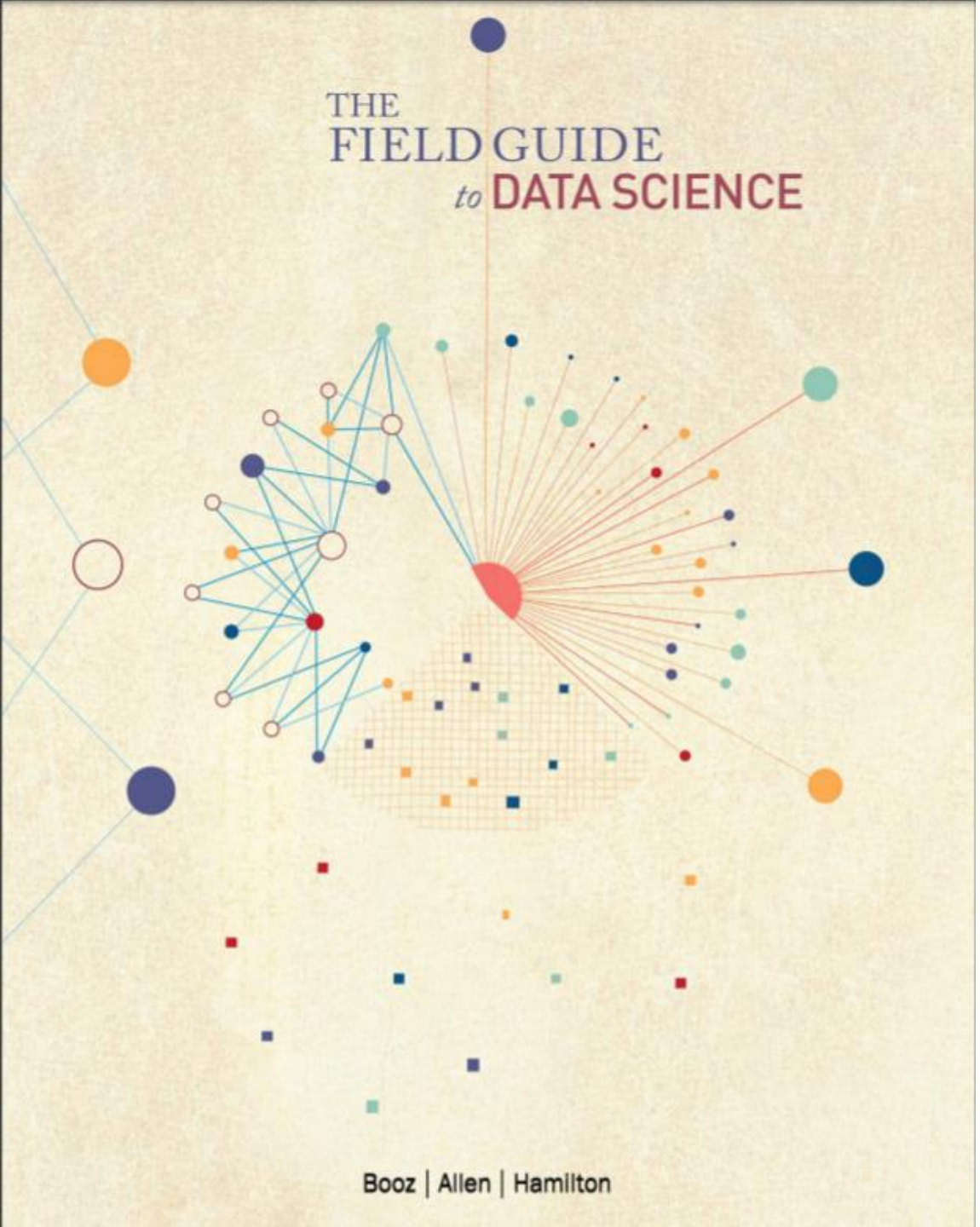
Others

OSEMN-
Obtain, Scrub, Explore, Model, and iNterpret

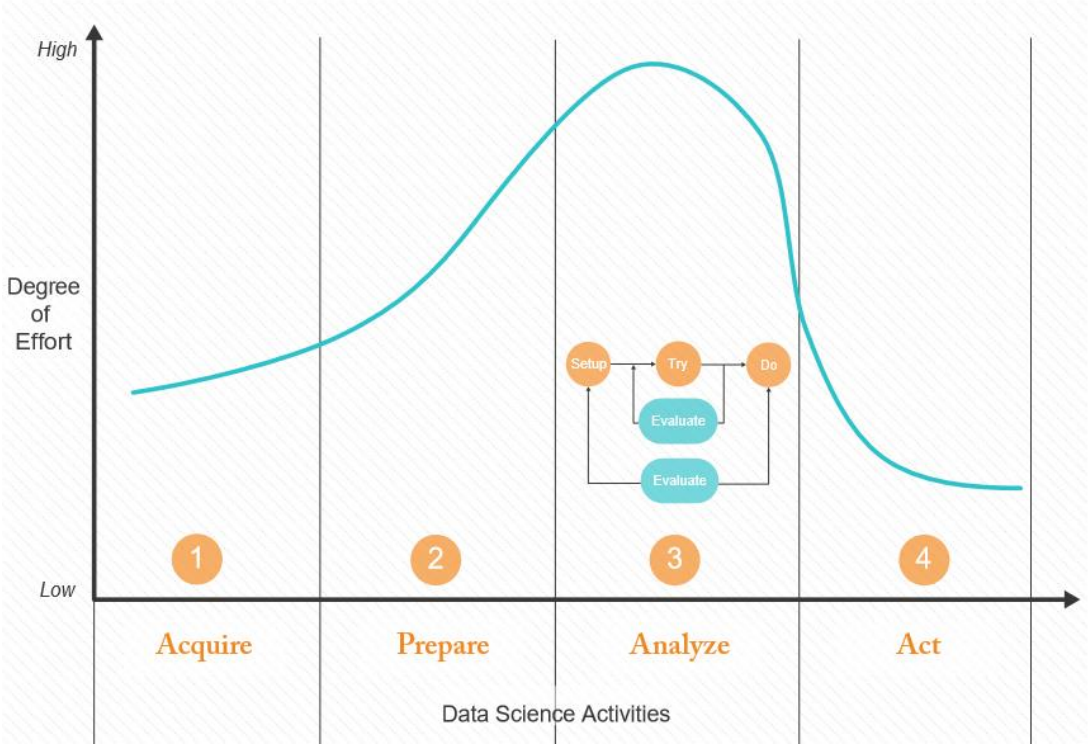
Dataist.com - Hilary Mason



http://www.dataists.com/2010/09/a-taxonomy-of-data-science/



Acquire-Prepare-Analyze-Act



Breakdown often seen in structure of training programs & lunch and learns

Contains many other pieces of advice

https://www.boozallen.com/content/dam/boozallen_site/sig/pdf/publications/2015-field-guide-to-data-science-160211215115.pdf

Common Theme

- Understand Business and Understand Data!
- Process is iterative
- Documentation of knowledge gained throughout process is crucial
- Actually deploying is optional depending on goal and outcome
- Phases use different skillset

How To Apply

How to apply that to RPG, Python, other

- Pick a process model as a team
- Agree to stick with it
- Before you even start a project
  - Document your business processes
  - Document your data
  - Know what decisions impact top and bottom line
- Separate the code and deliverables of each step
  - Allows you to cache intermediate results
  - Recap at end of each step and make minor fixes
  - Enables work breakdown

How to apply that to RPG, Python, other

- Be disciplined about the process
  - Continue the cycle when unexpected insights appear.
  - Come to the conclusion that the first question may not have an answer.
  - Don’t answer an alternate version of the question without starting a new cycle by verifying with business owners if there is any value there
- Link data, code and documentation
- Use version control
- Use Documentation Templates
- Use code templates:
  - ProjectTemplate for R
  - CookieCutter for Python

CookieCutter for Data Science

- CookieCutter is a Python library that removes tedious setup of new project (creating folders, initialize setup files, creating connections) https://drivendata.github.io/cookiecutter-data-science/
- You can create templates for any type of project
- CookieCutter for Data Science is a template for,

  wait for it,….

  Data Science

- `pip install cookiecutter`
- `cookiecutter https://github.com/drivendata/cookiecutter-data-science`
- `cookiecutter cookiecutter-data-science`

CookieCutter for Data Science

Standard folders for

- Documentation
- Exploratory analysis
- References
- Data
  - Raw
  - External
  - Interim
  - Processed

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
```

Cc

CookieCutter for Data Science

- Structured source code
- Standard filenames

```
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.testrun.org
```

CookieCutter for Data Science for RPG?

- You can fork the template and change the SRC folder and templated code to suit RPG
- Create template code for
  - make_dataset
  - build_features
  - train_model
  - predict_model
  - visualize
- If you are ambitious perhaps you can port CookieCutter to RPG

CookieCutter for Data Science for RPG

Why use a Data Science process in combination with CookieCutter?

- Other people will thank you!
  - Collaborate more easily with you on this analysis
  - Learn from your analysis about the process and the domain
  - Feel confident in the conclusions at which the analysis arrives
- You will thank you!

- https://drivendata.github.io/cookiecutter-data-science/

Q&A

Questions I have for you!

- Your suggestions will drive the next session!
  - Session 2: Aquire and Prepare
    - Curl, requests, liburl to get data from the web
    - About JSON
    - Pandas
- Are these topics relevant?
- Want a hands on session?
- Discuss pros and cons after introduction?

Questions for me?

- ….

Summary of URLs

- Duke Scandal
  Official location for CBS 60 Minute Segment on Duke
  YouTube Links to video segments
  60 Minute Segment: Deception at Duke
  CBS Background interview
- KDD Materials
  AI Magazine 1996, Fayyad article
  Summary of steps
- CRISP-DM Materials
  CRISP-DM 1.0
- Others
  Booz Allen Hamilton: Field Guide to Data Science
  Dataist: OSEMN
- Coding Templates
  CookieCutter Data Science (Python)
  Project Template (R)