Introduction to Data Science

Data Science

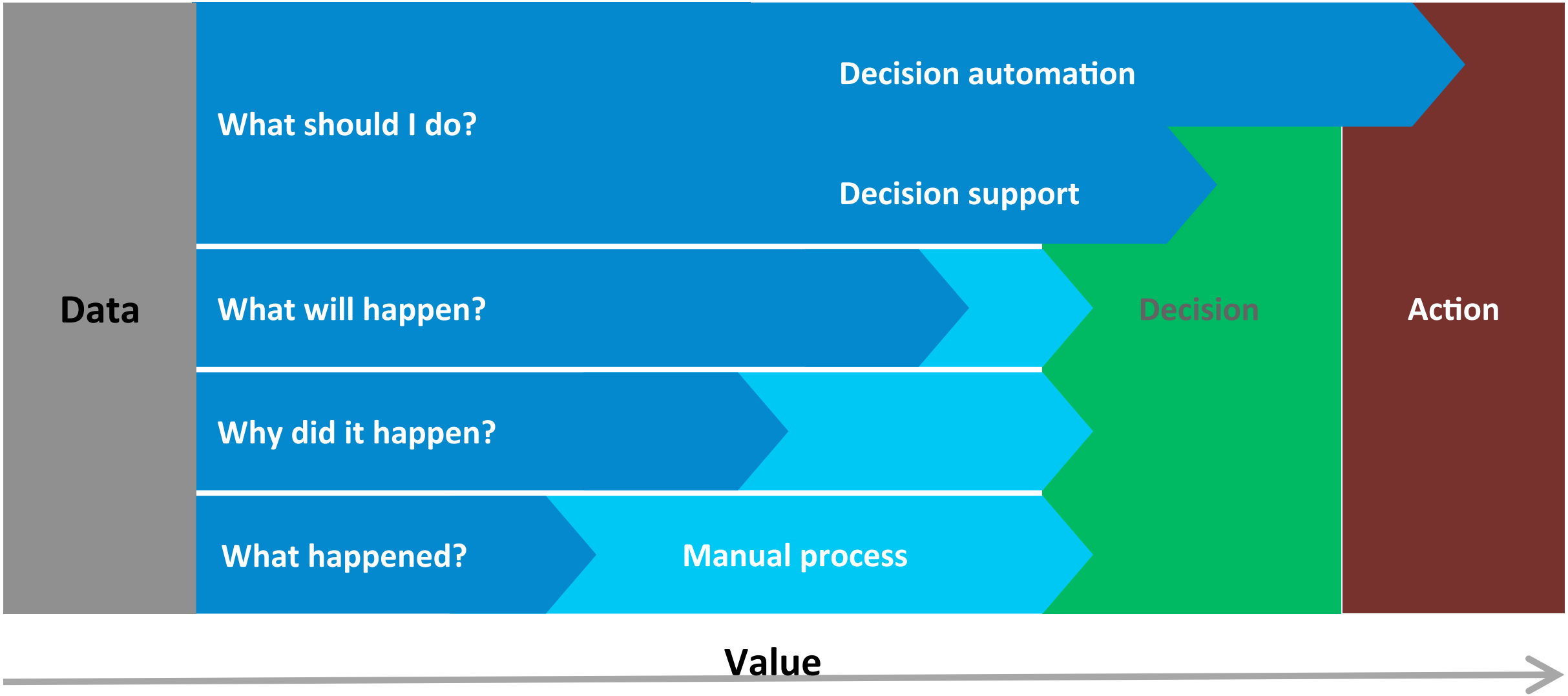
This course aims to provide:

- The skills you need to need to convincingly explain your evidence.
- The probability and statistics background you need to solidify your story.
- Programming experience and knowledge on state-of-the-art data science software.

What is Data Science?

Data science is the exploration and quantitative analysis of all available structured and unstructured data to develop understanding, extract knowledge, and formulate actionable results.

Data → Decisions → Actions



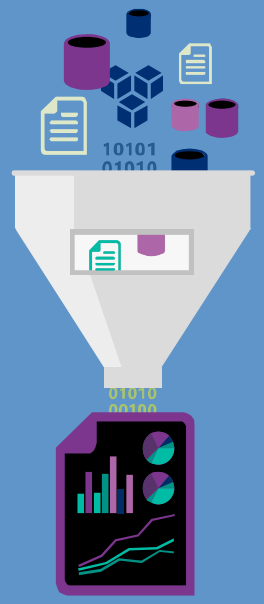
Data Scientists are Obsessed with Data

- Finding data sources
- Acquiring data
- Cleaning and transforming data
- Understanding relationships in data
- Delivering value from data
- Visualizing the result

What Type of Analytics?

Retrospective analytics

Real-time analytics



Predictive analytics

Intelligent SaaS apps

Predictive vs Prescriptive Analysis

- Predictive analytics calibrated on past data, tells us what to expect
- Prescriptive analysis tells what actions to take

Data Analytic Thinking

Data Analytic Thinking: Make decisions based on analysis of data.

- Replace intuition with data driven analytical decisions
- Extract value from data assets
- Increase pace of actions

Data Analytic Example

Medical treatment:

- Formerly waited for patient to show symptoms
- Classify patients with high risk; take preventative action
- Similar to predictive maintenance applications

Data Analytic Example

Detect risking payment accounts:

- Formally performed manual retrospective analysis
- Real-time decision on account
- Classification widely used for anomaly detection

Data Analytic Example

Bicycle Rental Demand:

- Under stocking or over-stocking costly
- Manual forecasting difficult
- Forecast demand to optimize inventory
- Forecasting widely used

Data Analytic Thinking

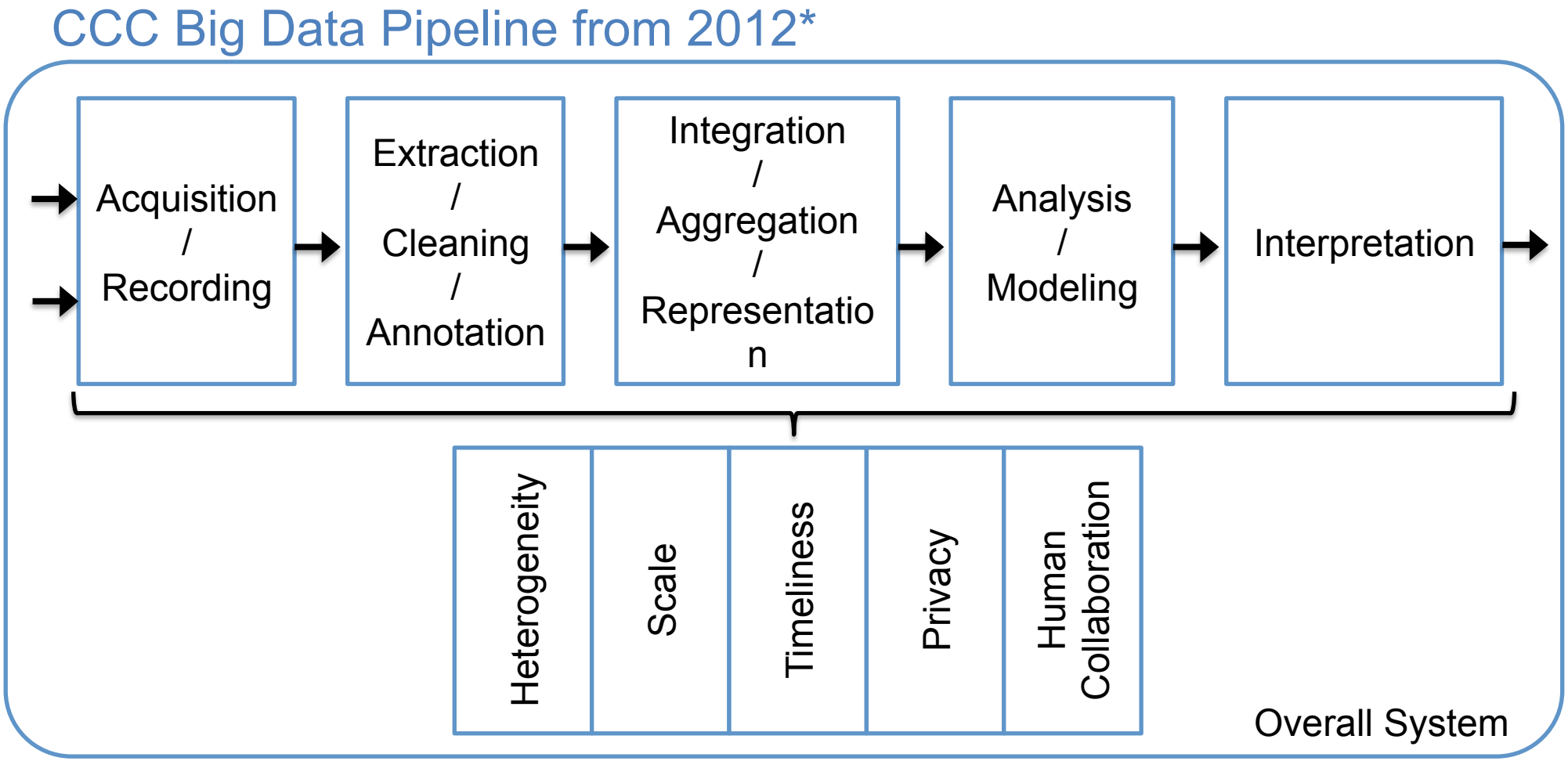
Data Analytic Thinking:

- Replace intuition with data driven analytical decisions
- Transform raw data to valuable asset
- Increase pace of action

Historical Notes on KDD, CRISP-DM, Big Data and Data Science and their relationship to Data Mining and Machine Learning

Historical Notes

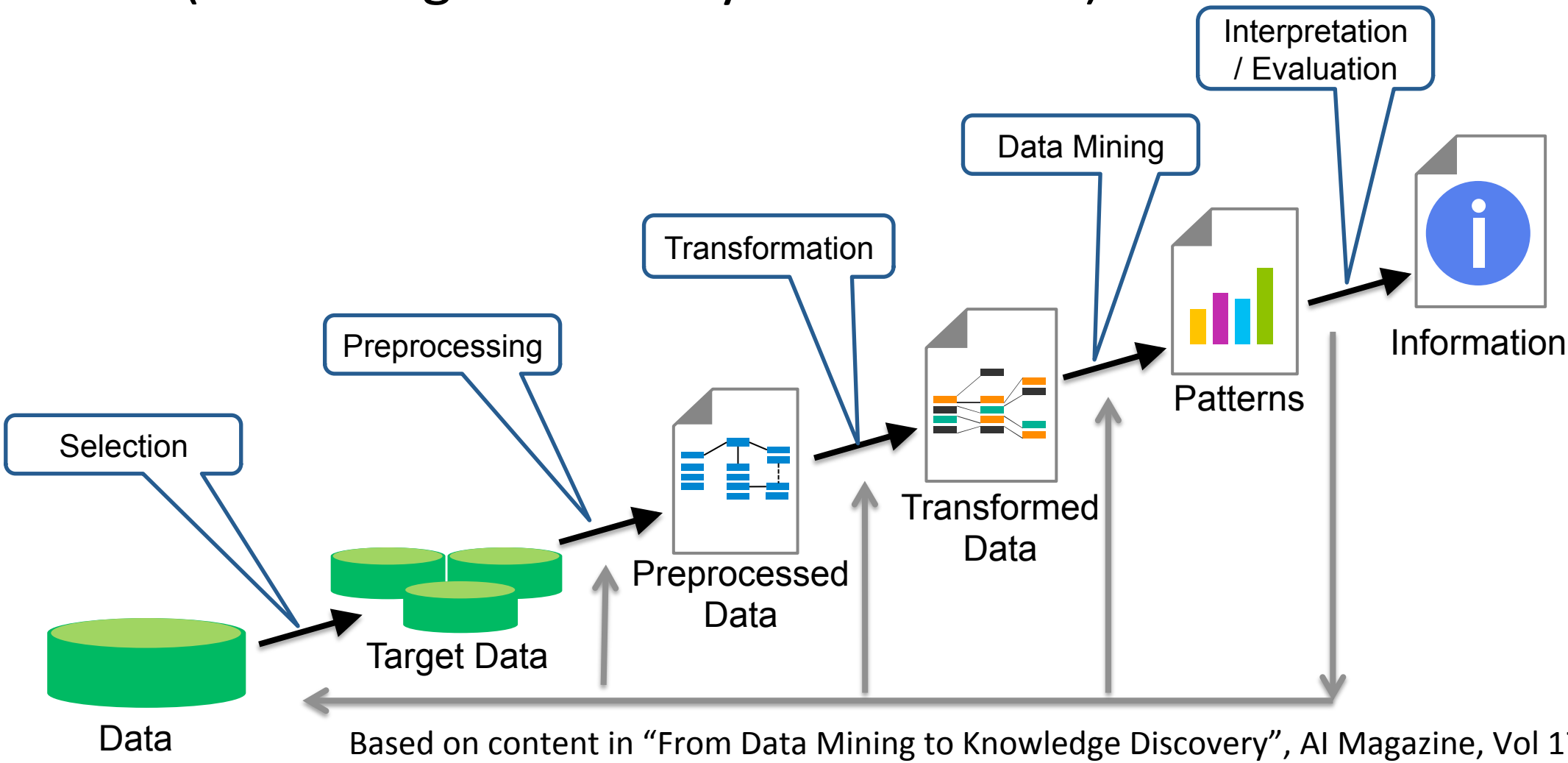
- Term “Big Data” coined by astronomers Cox and Ellsworth in 1997



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

Historical Notes

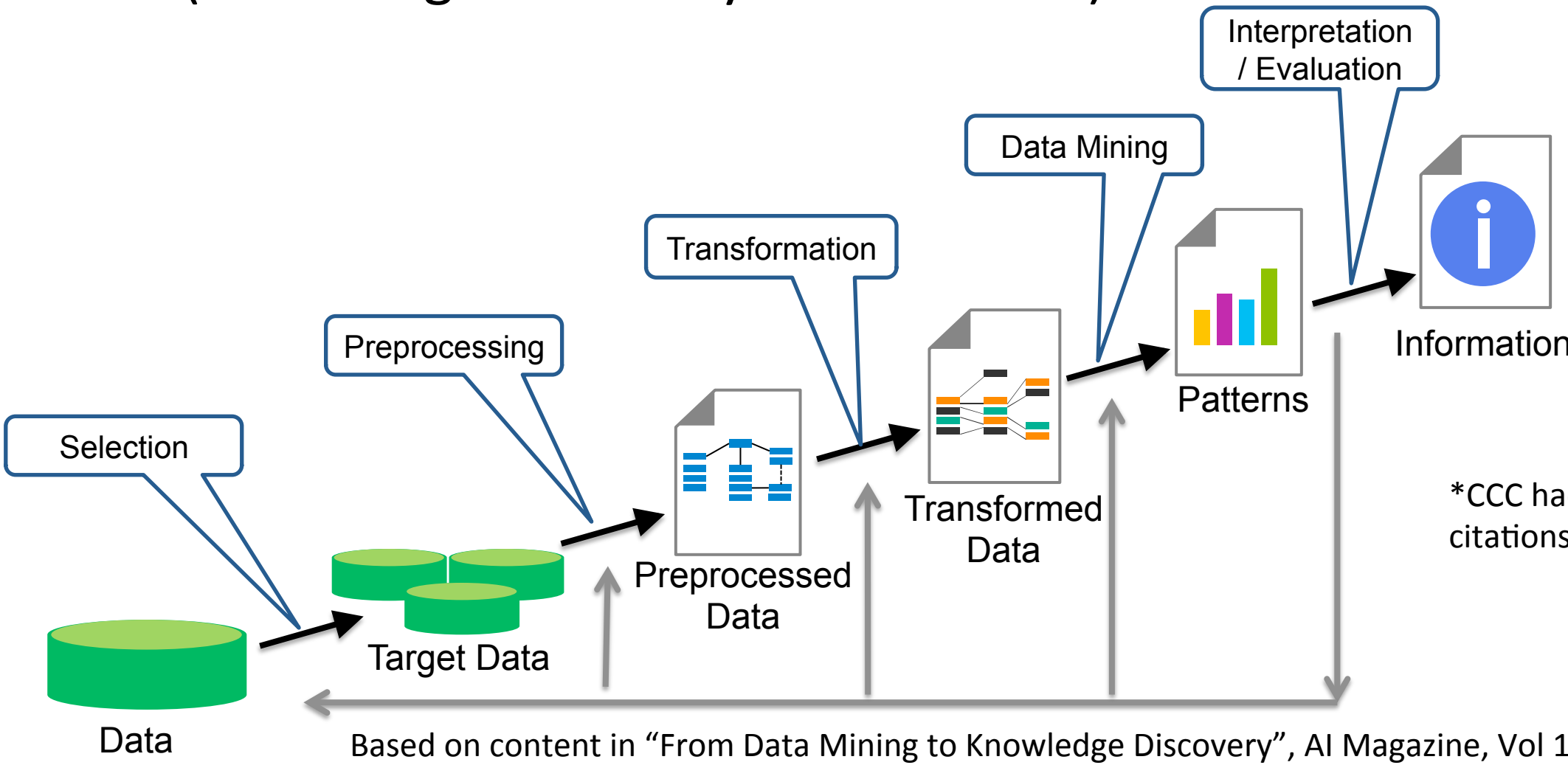
- KDD (Knowledge Discovery in Databases) Process



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

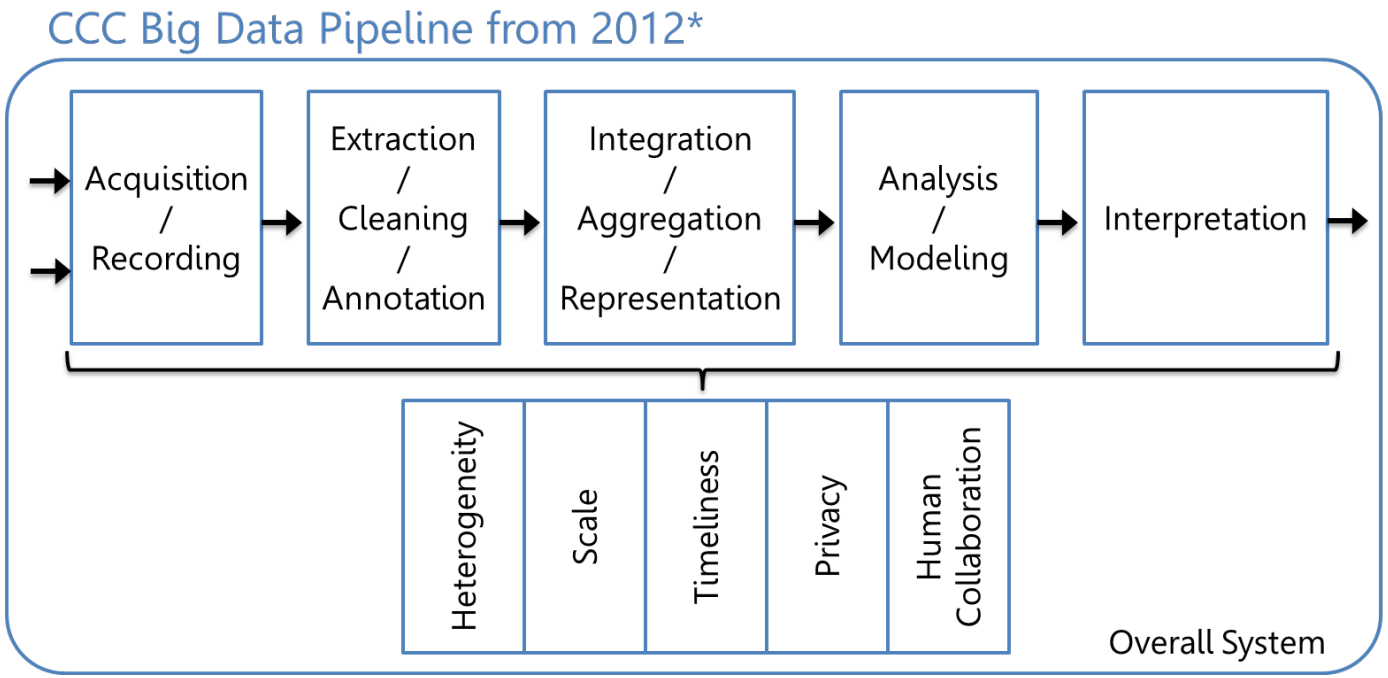
Historical Notes

- KDD (Knowledge Discovery in Databases) Process

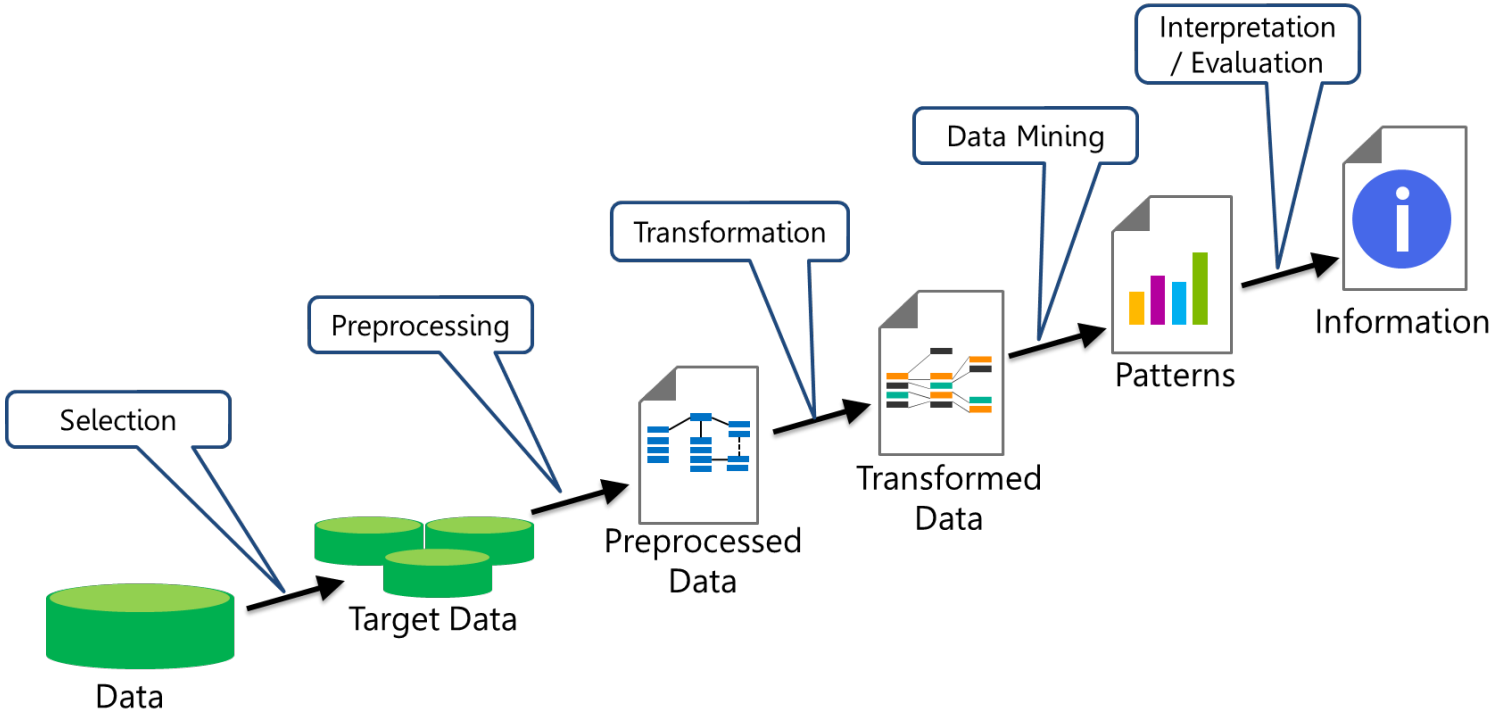


Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

CCC 2012

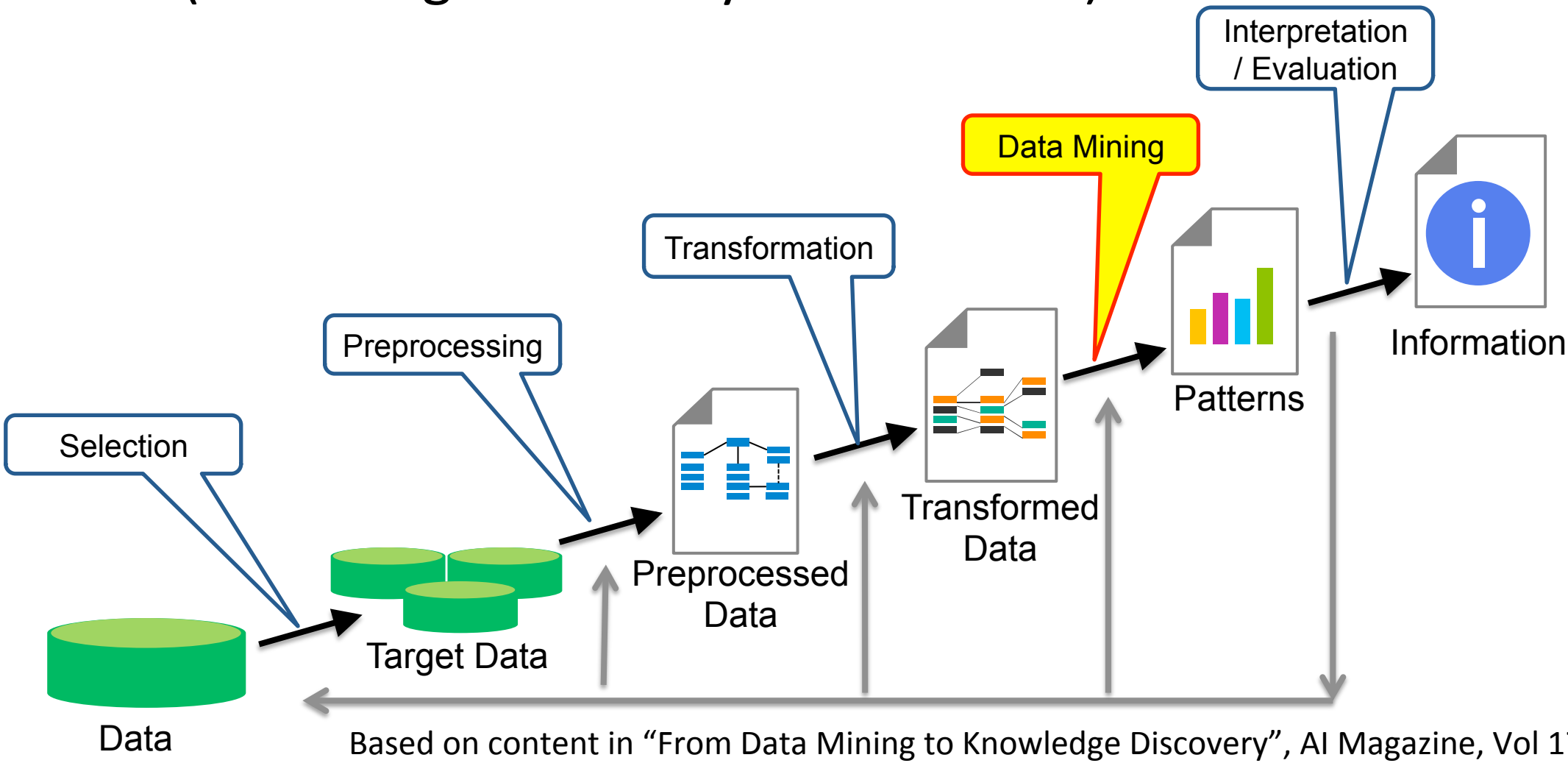


KDD 1996



Historical Notes

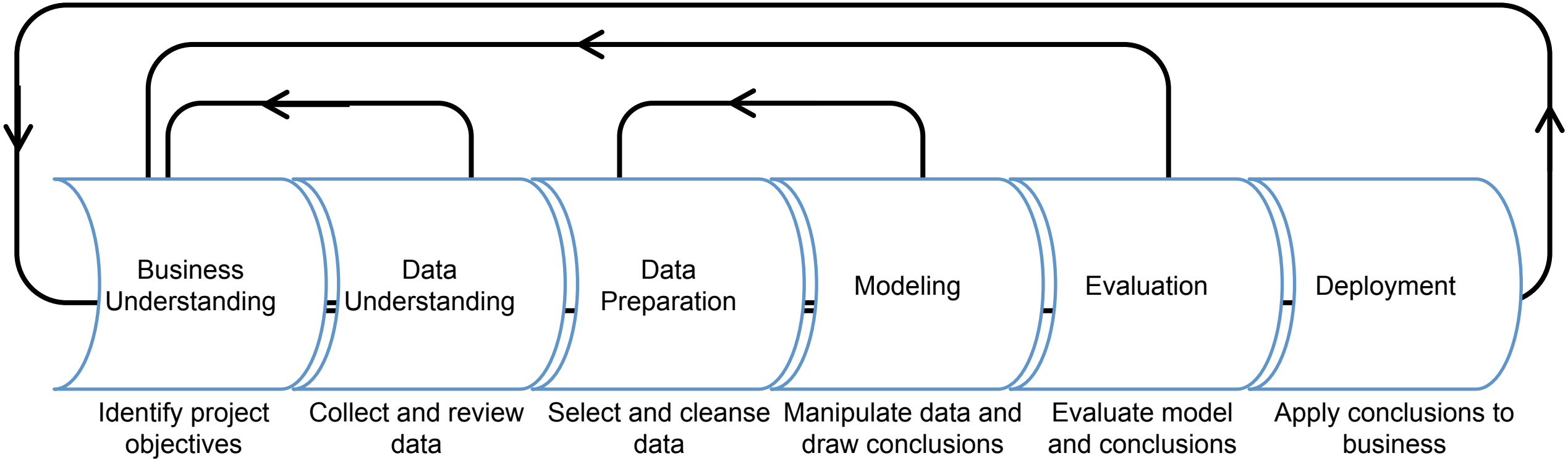
- KDD (Knowledge Discovery in Databases) Process



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.

Knowledge Discovery Process Example

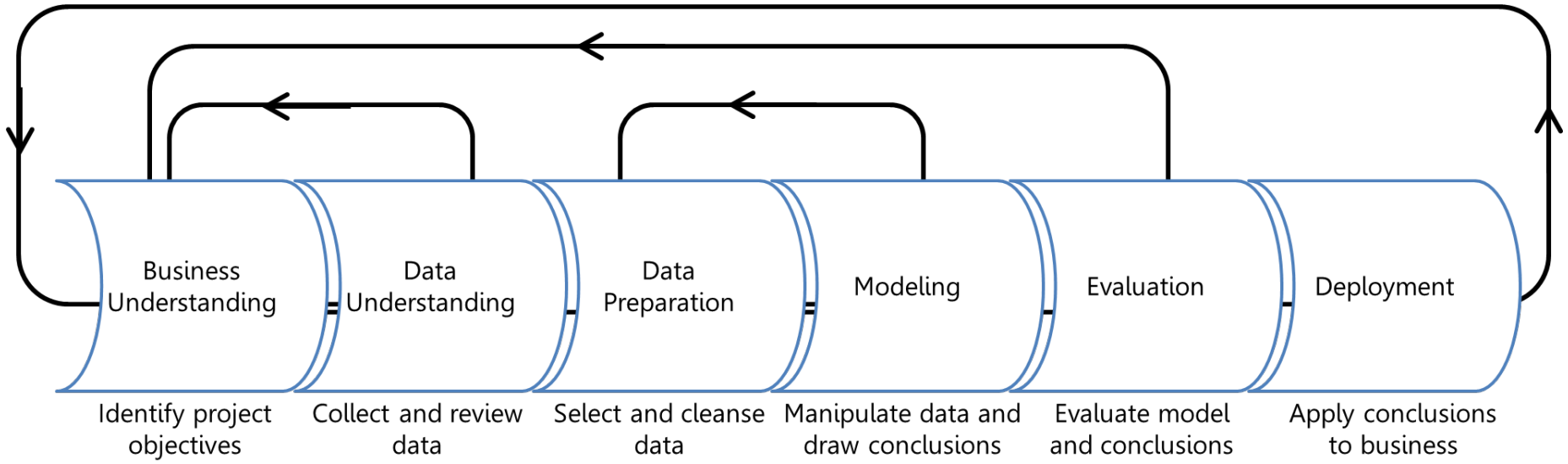
- I’ll walk you through the knowledge discovery process with an example – the process of predicting power failures in Manhattan.

Motivation for Example

- In NYC the peak demand for electricity is rising.
- The infrastructure dates back to the 1880's from the time of Thomas Edison.
- Power failures occur fairly often (enough to do statistics) and are expensive to repair
- We want to determine how to prioritize manhole inspections in order to reduce the number of manhole events (fires, explosions, outages) in the future.
- This is a real problem.

Stages in the knowledge discovery process

- Opportunity Assessment & Business Understanding
- Data Understanding & Data Acquisition
- Data Preparation, including Cleaning and Transformation
- Model Building
- Policy Construction
- Evaluation, Residuals and Metrics
- Model Deployment, Monitoring, Model Updates



Opportunity Assessment & Business Understanding

What do you really want to accomplish and what are the constraints? What are the risks? How will you evaluate the quality of the results?

- For manhole events the general goal was to “predict manhole fires and explosions before they occur.” We made it more precise:
  - Goal 1: Assess predictive accuracy for predicting manhole events in the year before they happen.
  - Goal 2: Create a cost-benefit analysis for inspection policies that takes into account the cost of inspections and manhole fires. Determine how often manholes need to be inspected.

Data Understanding & Data Acquisition

Data were:

- Trouble tickets – free text documents typed by dispatchers documenting problems on the electrical grid.
- Records of information about manholes
- Records of information about underground cables
- Electrical shock information tables
- Extra information about serious events
- Inspection reports
- Vented cover data

How do Big Data sets include "Variety" and "Veracity"

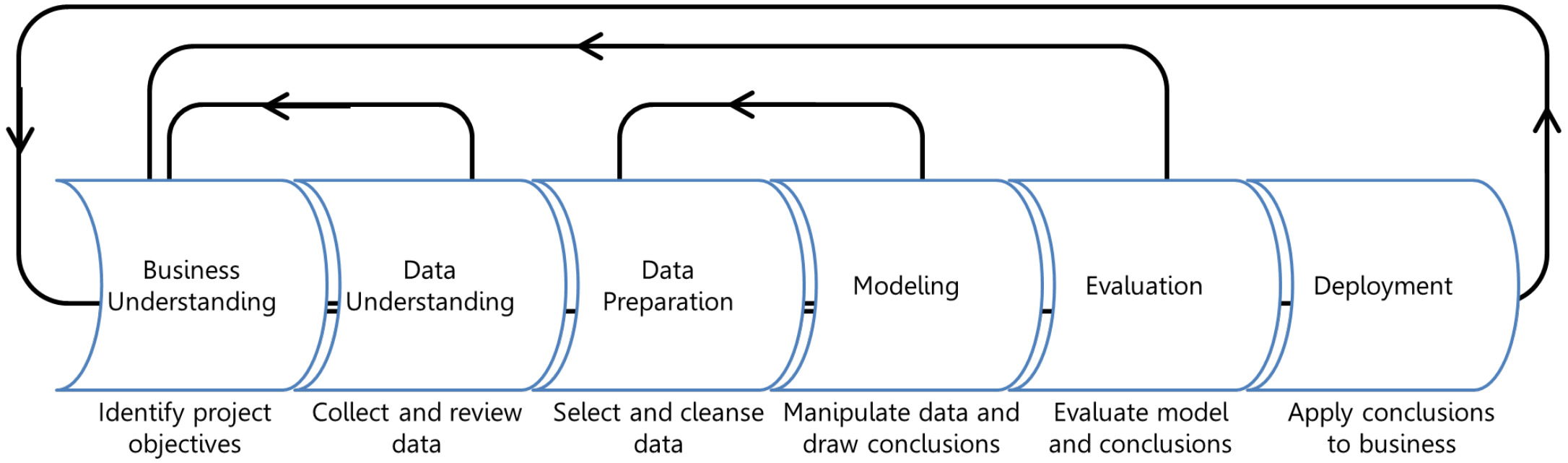
Data Cleaning and Transformation

- Sometimes 99% of the work

Data Cleaning and Transformation

- Turn free text into structured information:
  - Trouble tickets turned into a vector like:
    - Serious / Less Serious / Not an Event
    - Year
    - Month
    - Day
    - Manholes involved
    - …
- Try to integrate tables (create unique identifiers):
  - If you join manholes to cables, half of the cable records disappear

Knowledge Discovery Process



Model Building

- Often predictive modeling, meaning machine learning or statistical modeling
- If you want to answer a yes/no question, this is classification.
  - For manholes, will the manhole explode next year? Y/N
- If you want to predict a numerical value, this is regression.
- If you want to group observations into similar-looking groups, this is clustering.
- If you want to recommend someone an item (e.g., book/movie/product) based on ratings data from customers, this is a recommender system.
- Note: There are many other machine learning problems.

Policy Construction

- How will your model be used to change policy?
  - For manholes, how should we recommend changing the inspection policy based on our model?
  - Consider using social media and customer purchase data to determine customer participation if Starbucks moves into New City. After the model is created, how to optimize where the shops are located, how big they are, and where the warehouses are located.
- Model building is predictive, Policy Construction is prescriptive.

Evaluation

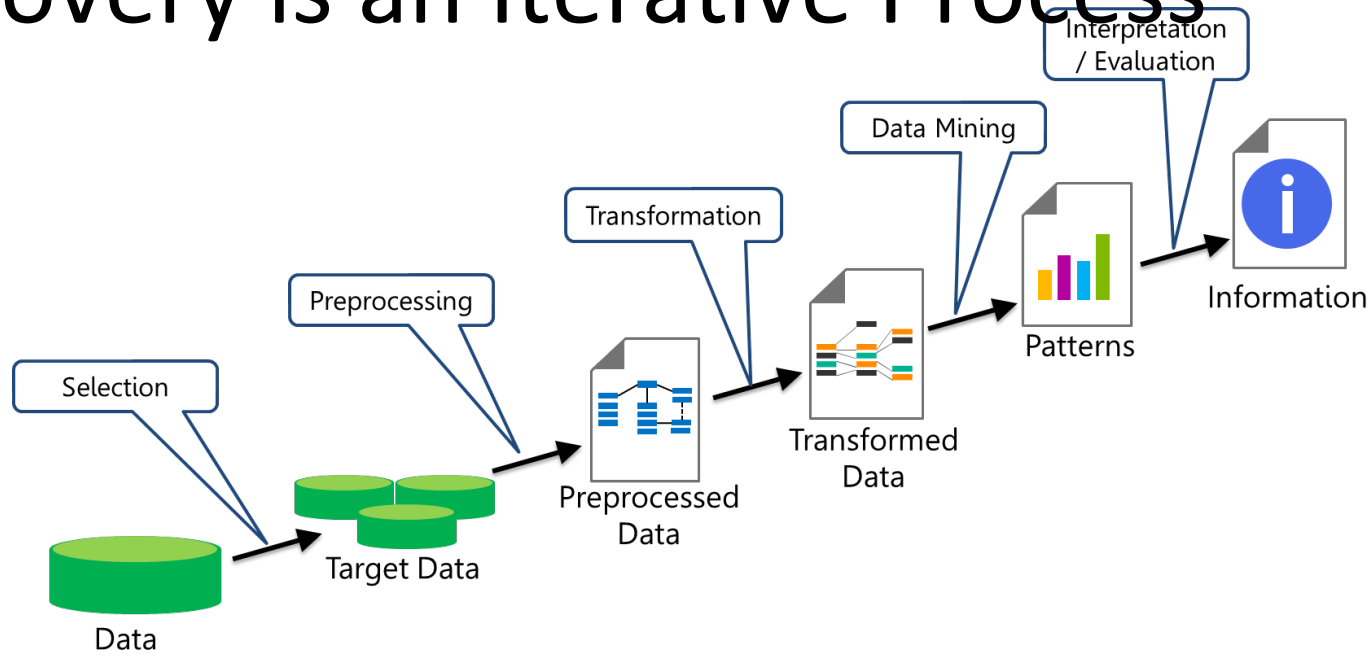
- How do you measure the quality of the result? Evaluation can be difficult if the data do not provide ground truth.
- For manhole events, we had engineers at Con Edison withhold high quality recent data and conduct a blind test.

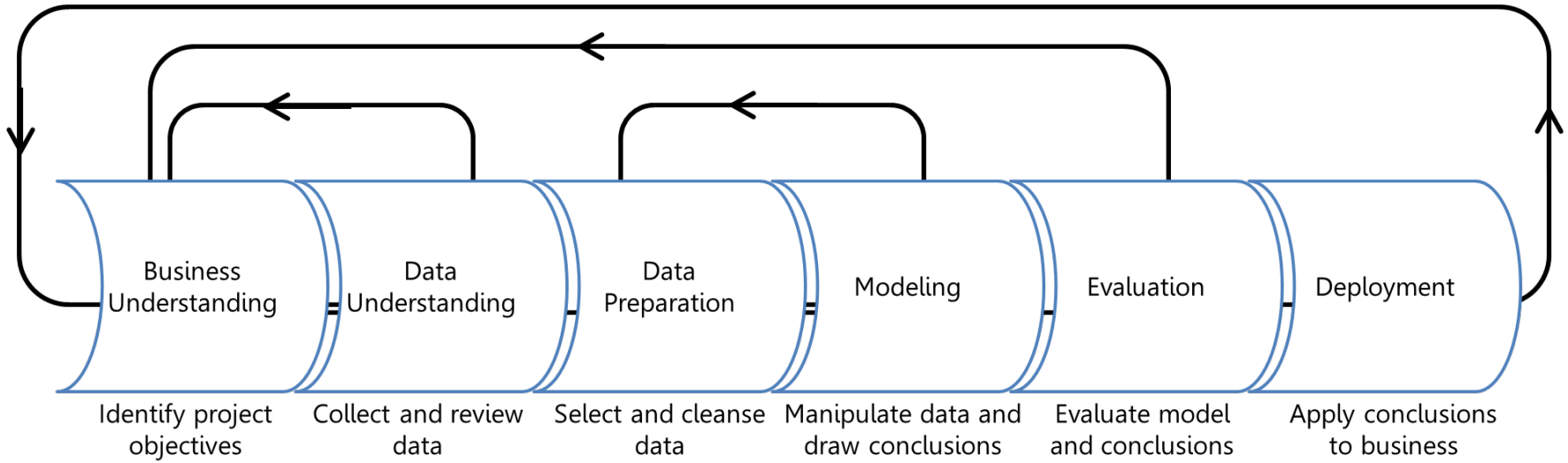
Deployment

- Getting a working proof of concept deployed stops 95% percent of projects.
- Don’t bother doing the project in the first place if no one plans to deploy it.*
- Keep a realistic timeline in mind. Then add several months.
- While the model is deployed it will need to be updated and improved.

* Unless it’s fun.

Knowledge Discovery is an Iterative Process





Summary

- Several attempts to make the process of discovering knowledge scientific
  - KDD, CRISP-DM,CCC Big Data Pipeline
- All have very similar steps
  - Data Mining is only one of those steps (but an important one)

Tools for Data Science

- Open source tools: R and Python
- Azure Machine Learning
  - Easy to use
  - Easy to deploy
- Data science stacks

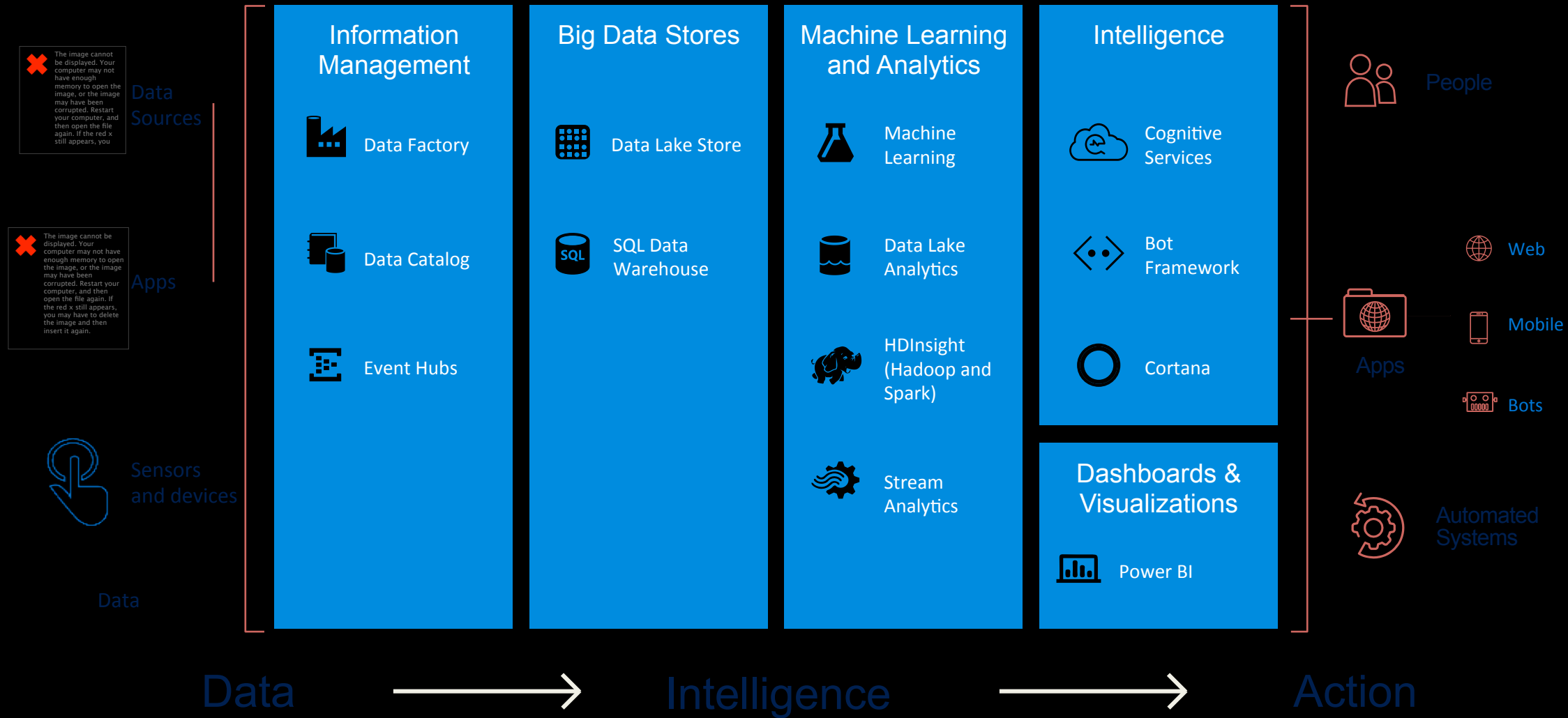
Why Open-Source Tools?

- R and Python widely used in data science
- Highly interactive
- Good visualization
- Vast packages (libraries) of utilities and algorithms
- Excellent development environments
- Jupyter notebooks

R or Python?

- R and Python are widely used in data science
- Powerful open-source data science tools
- Both have advantages and disadvantages
- Python tends to be more systematic and faster
- R contains wider range of packages and capabilities
- R offers grammar of graphics

Cortana Intelligence Suite



Why Azure ML?

- Easy to use
- Quickly deploy production solutions as web services
- Models run in a highly scalable cloud environment
- Secure cloud environment for data and code
- Powerful, efficient built-in algorithms
- Extensible with, SQL, Python, and R
- Integration with Jupyter notebooks

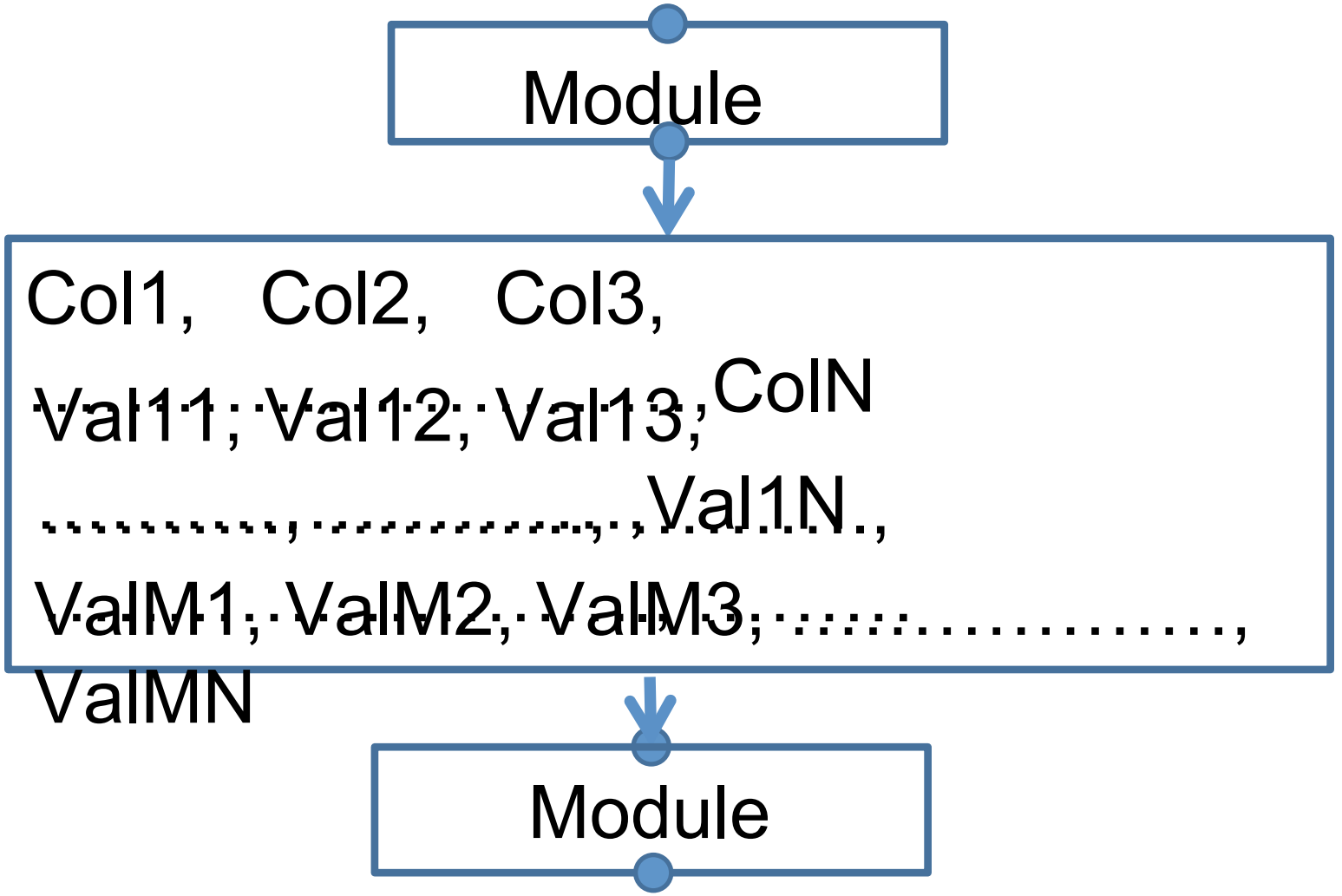
Azure ML Free Tier Account

- Free Tier Account
- Unlimited time, with restricted priority
- Paid account provides full performance

Azure ML Studio

- Experiments contain workflow
- Experiments constructed of modules
- Experiments in sharable workspace
- Modules transform data, compute models, score models, and evaluate models
- Create custom modules with SQL, R and Python
- Deploy solutions as web services

Data Passed from Module to Module in Azure ML Tables



Rectangular table

N Columns

M Rows

Equal length columns

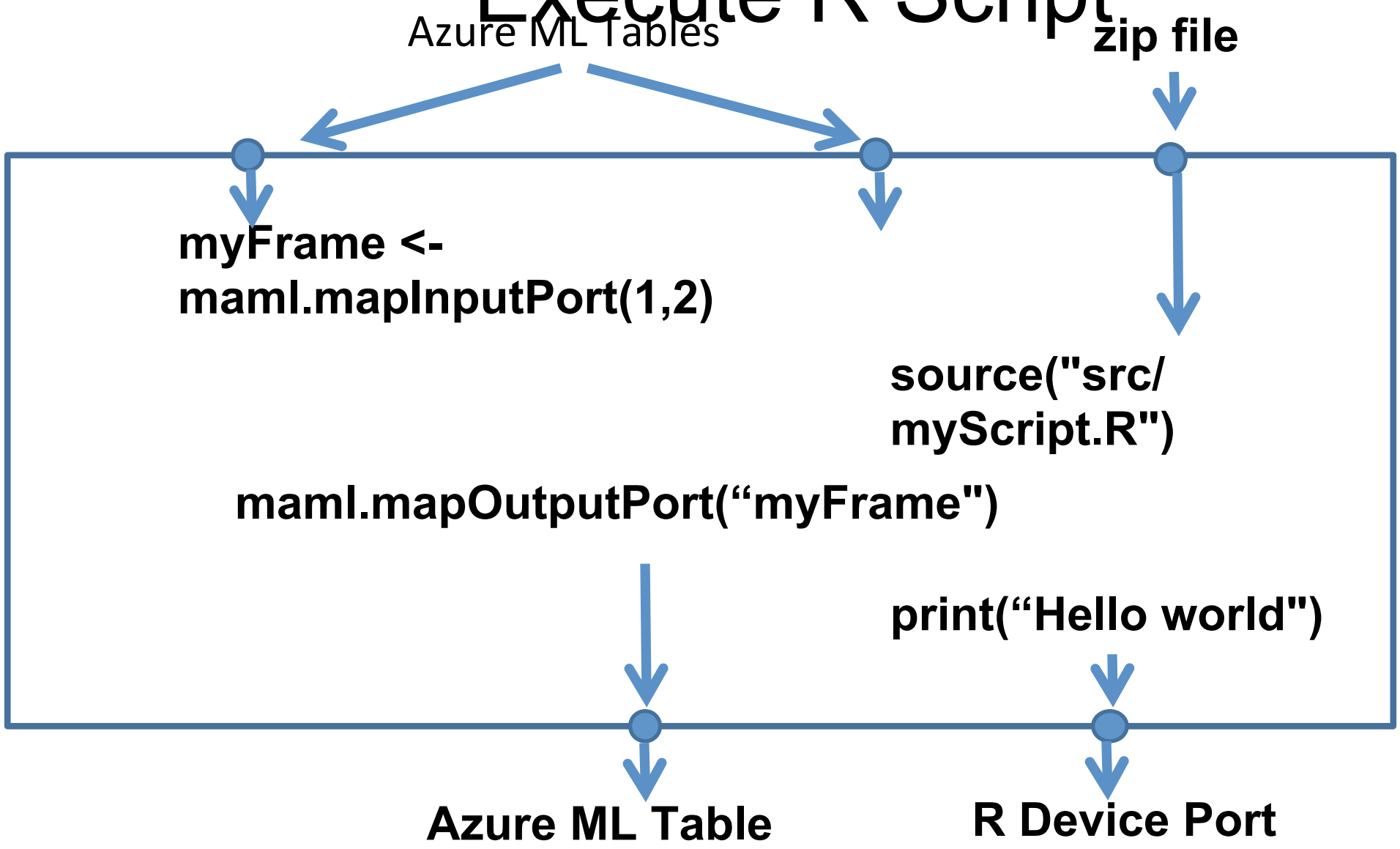
Azure ML Table Data Types

- Numeric; Floating Point
- Numeric: Integer
- Boolean
- String
- Categorical
- Date-time
- Time-Span
- Image

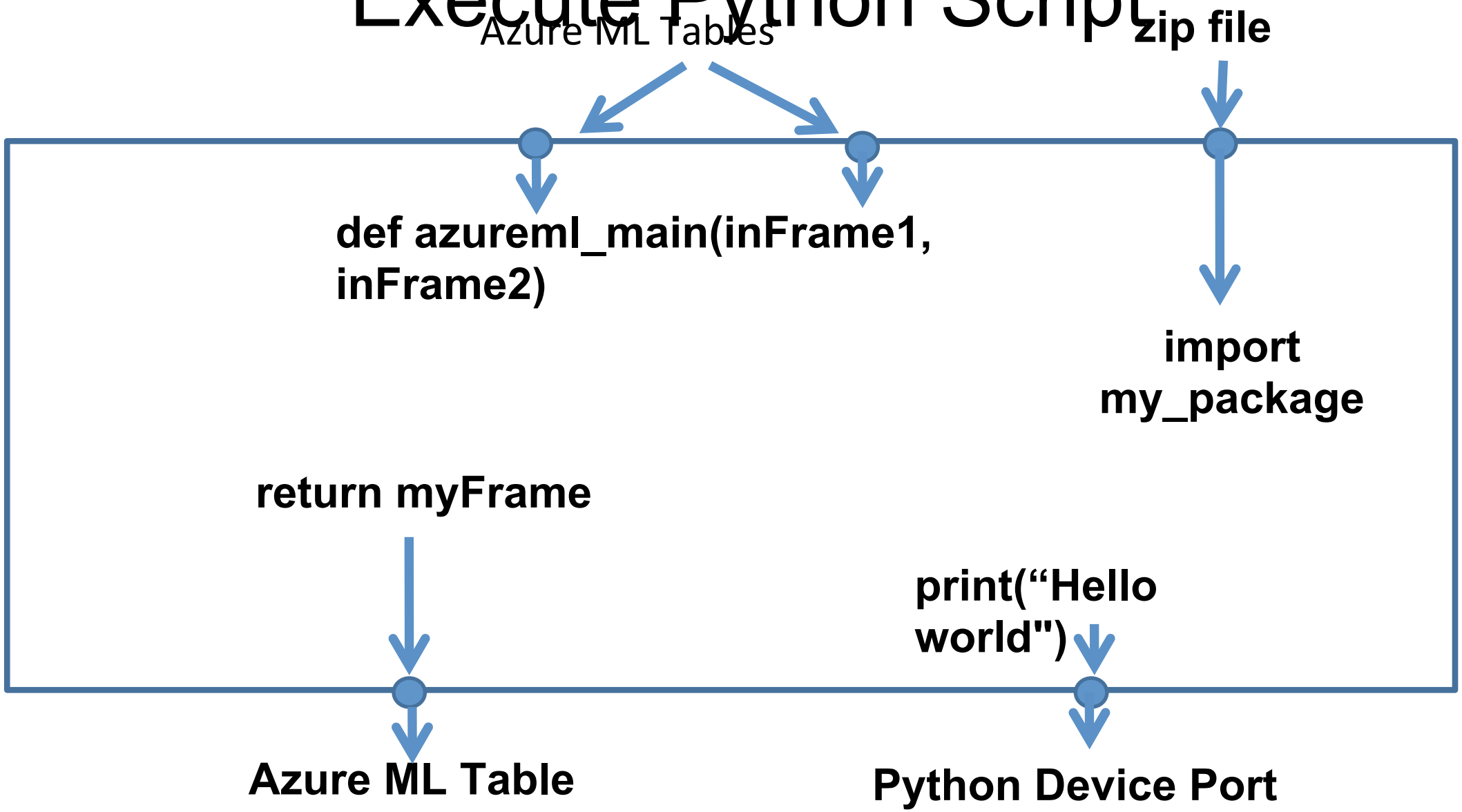
Developing and testing R and Python

- Azure ML is a production environment
- Interactively develop and test in IDE
- Subset data as needed – download as .csv
- IDE has powerful editor and debugger
- Cut and paste code into Execute R/Python Script module to test in Azure ML
- Jupyter notebooks

Execute R Script



Execute Python Script



Debugging R and Python in Azure ML

- Code tested in IDE should run in Azure ML, but…….
- If error occurs look at the error.log or output.log
- From R use print() function
- From Python use sys.stderr.write() from sys

SQL in Azure ML

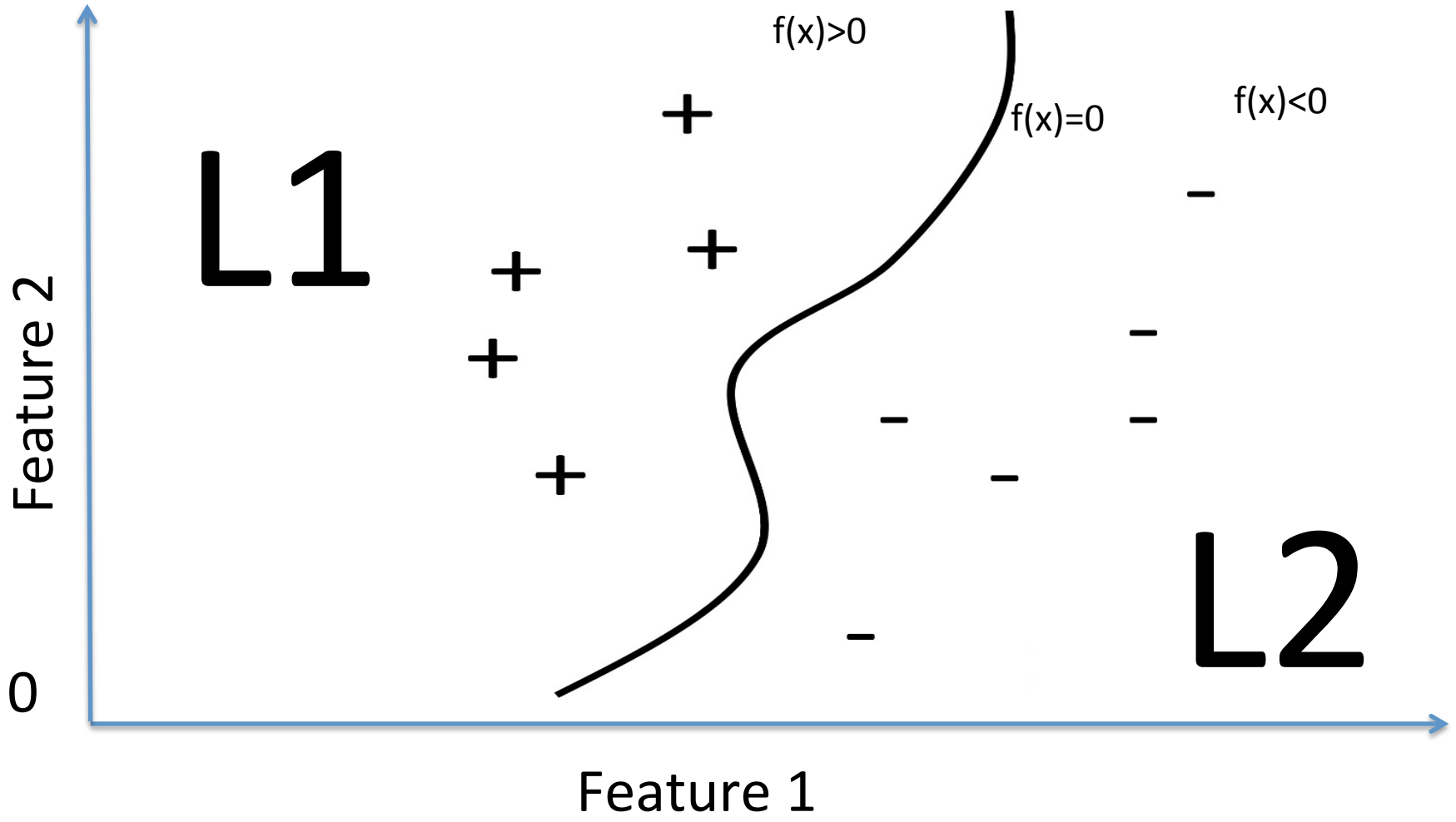
- SQL data I/O; Reader and Writer modules
- SQL transformation module
- SQL resources

Querying with Transact-SQL, Graeme Malcom and Geoff Allix, https://www.edx.org/course/querying-transact-sql-microsoft-dat201x-0

Classification

Formally, given training set (x_i, y_i) for i=1…n, we want to create a classification model f that can predict label y for a new feature values x.

f(x) = function(features)



Machine learning workflow