Department of Mathematics
University of Calabria

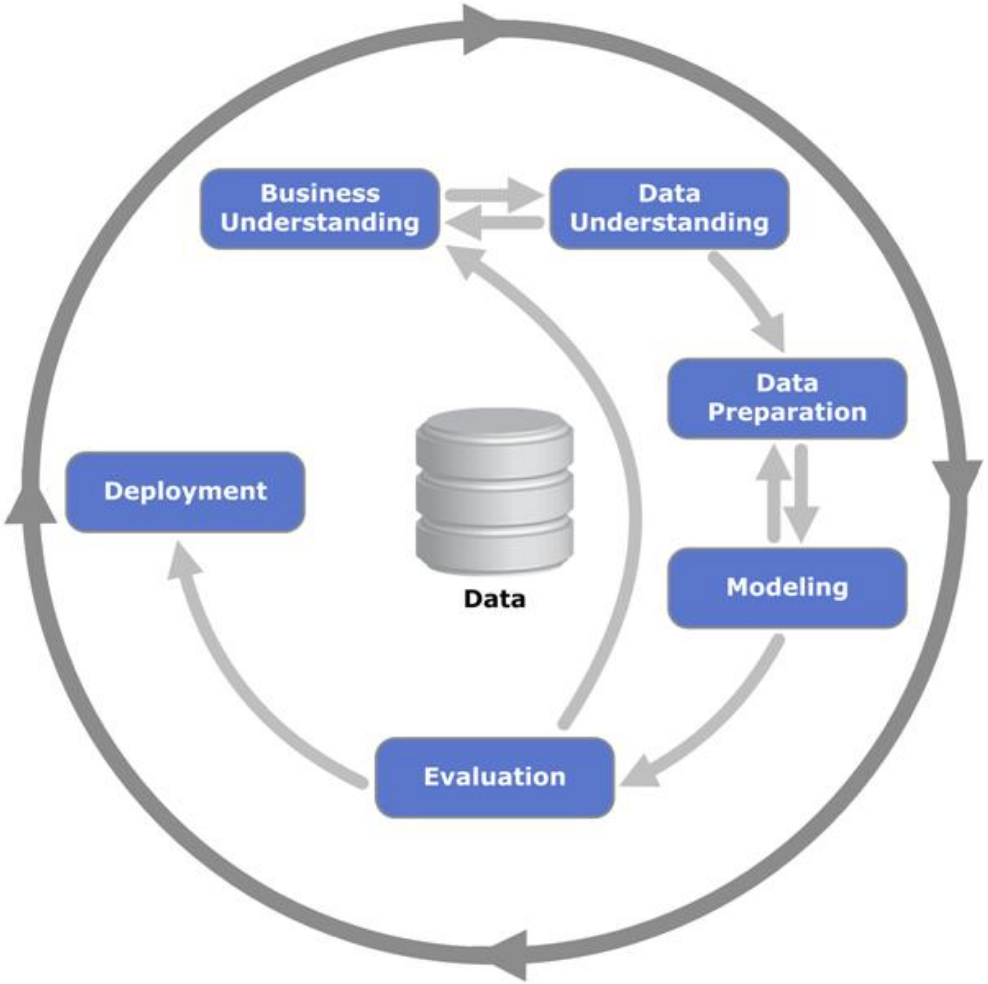
# Data Warehouse and Data Mining

## Module II – Data Mining

# Study case – Intrusion Detection

Ph.D. Ettore Ritacco

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

# About this lesson

- The proposed study case is a simplification of the KDD Cup 1999 Challenge
- KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining, the leading professional organization of data miners.
- http://www.sigkdd.org/kddcup/index.php

# Business Understanding

- Software to detect network intrusions protects a computer network from unauthorized users, including perhaps insiders.

- The intrusion detector learning task is to build a predictive model capable of distinguishing between “bad” connections, called intrusions or attacks, and “good” normal connections.

# Business Understanding

- Moreover, we want that predictive model is capable of distinguishing the attack types
- Attacks:
  - DOS: denial-of-service, e.g. syn flood;
  - R2L: unauthorized access from a remote machine, e.g. guessing password;
  - U2R: unauthorized access to local superuser (root) privileges, e.g., various "buffer overflow" attacks;
  - PROBE: surveillance and other probing, e.g., port scanning.

# Data Understanding

| feature name | description | type |
|---|---|---|
| duration | length (number of seconds) of the connection | continuous |
| protocol_type | type of the protocol, e.g. tcp, udp, etc. | discrete |
| service | network service on the destination, e.g., http, telnet, etc. | discrete |
| src_bytes | number of data bytes from source to destination | continuous |
| dst_bytes | number of data bytes from destination to source | continuous |
| flag | normal or error status of the connection | discrete |
| land | 1 if connection is from/to the same host/port; 0 otherwise | discrete |
| wrong_fragment | number of ``wrong'' fragments | continuous |
| urgent | number of urgent packets | continuous |

# Data Understanding

| feature name | description | type |
|---|---|---|
| hot | number of "hot" indicators | continuous |
| num_failed_logins | number of failed login attempts | continuous |
| logged_in | 1 if successfully logged in; 0 otherwise | discrete |
| num_compromised | number of "compromised" conditions | continuous |
| root_shell | 1 if root shell is obtained; 0 otherwise | discrete |
| su_attempted | 1 if "su root" command attempted; 0 otherwise | discrete |
| num_root | number of "root" accesses | continuous |
| num_file_creations | number of file creation operations | continuous |
| num_shells | number of shell prompts | continuous |
| num_access_files | number of operations on access control files | continuous |
| num_outbound_cmds | number of outbound commands in an ftp session | continuous |
| is_hot_login | 1 if the login belongs to the "hot" list; 0 otherwise | discrete |
| is_guest_login | 1 if the login is a "guest" login; 0 otherwise | discrete |

# Data Understanding

| feature name | description | type |
|---|---|---|
| count | number of connections to the same host as the current connection in the past two seconds | continuous |
| | Note: The following features refer to these same-host connections. | |
| serror_rate | % of connections that have "SYN" errors | continuous |
| rerror_rate | % of connections that have "REJ" errors | continuous |
| same_srv_rate | % of connections to the same service | continuous |
| diff_srv_rate | % of connections to different services | continuous |
| srv_count | number of connections to the same service as the current connection in the past two seconds | continuous |
| | Note: The following features refer to these same-service connections. | |
| srv_serror_rate | % of connections that have "SYN" errors | continuous |
| srv_rerror_rate | % of connections that have "REJ" errors | continuous |
| srv_diff_host_rate | % of connections to different hosts | continuous |