ISA 480: Data Driven Security

04 - An Overiew of ETL Operations

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022

Outline

Quick Recap of Classes 01-03

Learning Objectives Discussed

- ☑ **Understand the structure and expectations of this course**
- ☑ **Define information security and its main goals**
- ☑ **Understand that breaches are frequent and target different industries**
- ☑ **Understand the basic operations in either R or Python**

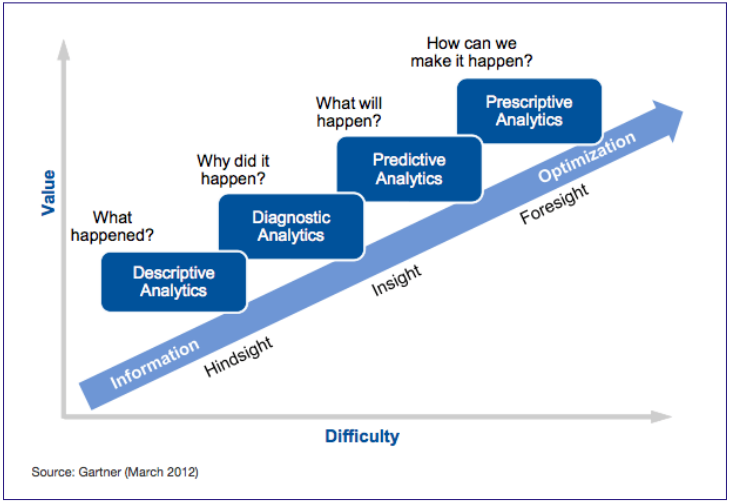
Plan for Today's Class

Learning Objectives

- Describe the data analytics process
- Understand the importance of ETL and data preprocessing
  - Explain what do we mean by technically correct data
  - Explain what do we mean by consistent data
- Perform basic ETL in R and/or Python

Outline

Big Picture: Levels of Data Analytics



Implementation Frameworks for ML/Data Analytics

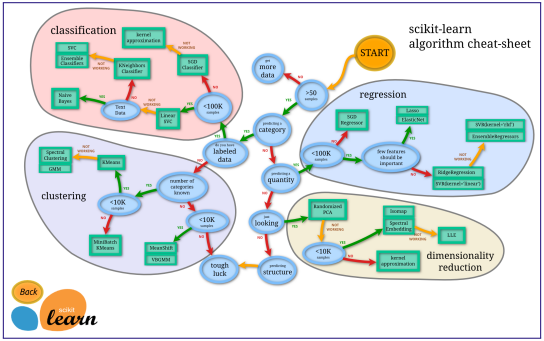
The KDD Process	SEMMA	CRISP-DM
Pre KDD	———————-	Business understanding
Selection	Sample	Data understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	———————-	Deployment

Table: A comparison of the three most commonly used Data Analytic/ Machine Learning Frameworks. Table adapted from Azevedo and Santos (2008).

Comment on the three Implementation Frameworks

Practical Issue with Existing Guidance

These three frameworks encourage iterating, but do not provide sufficient guidance on how.



Source: SciKit-Learn "Choosing the Right Estimator" (2022).

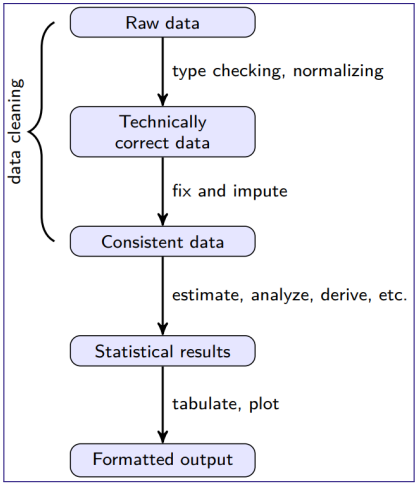
Outline

Data Analysis: A Crowd Sourced Definition from Wikipedia

Wikipedia "Data Analysis" (2022) defines **data analysis** to be the process of:

> *inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.*

Data Analysis Value Chain



Source: The figure is from De Jonge and Van Der Loo (2013). Please click on the image and read p. 7-8 from the reference.

Useful Operations/Functions (in R/Python)

Goal	R	pandas
structure/classes of variables	`str()` `dplyr::glimpse()`	`df.info()`
summaries for each col/variable	`summary()`	`df.describe()`
dimensions of object	`dim()`	`df.shape`

convert column types	`df$x1 = as.numeric(df$x1)`
	`df %>%`
	`mutate(x1 = as.numeric(x1))`
		`df["x1"] = df["x1"].astype(float)`
		`df = df.astype({"x2": int, "x3": complex})`

subset data frame by row numbers	`slice(df, 1:10)`	`df.iloc[:9]`

See the Panda's Official Comparison with R Libraries for more details.

A Primer on Technically Correct and Consistent Data

Let us discuss the concepts of **technically correct** and **consistent** data based on the `KDDTrain+.csv` dataset. Note that the data was created by Tavallaee et al. (2009) (click here to access their paper).

Comments:

- I have intentionally picked a large dataset so we can think about how we can design a process to check for technically correct and consistent data.
- How many rows, variables and unique values for the `label` column do we have?
- Can you report summary statistics for the data?
- How can we visualize such data quickly in both R (`DataExplorer::create_report()`) and Python (`pandas-profiling`)?

Outline

Developing an intuition [1]

```
system("ping google.com", intern = T)
```

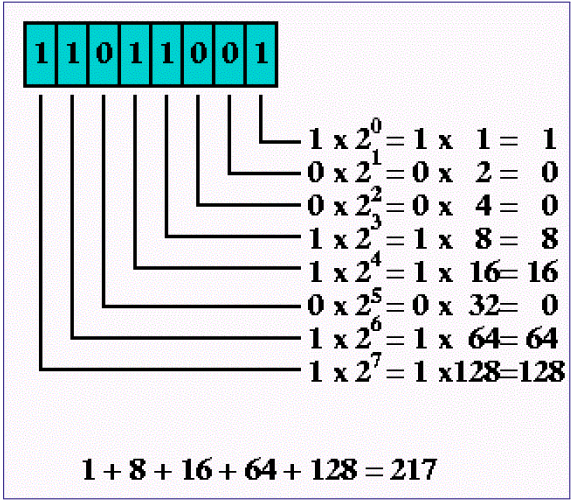
```
##  [1] ""
##  [2] "Pinging google.com [172.217.0.174] with 32 bytes of data:"
##  [3] "Reply from 172.217.0.174: bytes=32 time=30ms TTL=114"
##  [4] "Reply from 172.217.0.174: bytes=32 time=27ms TTL=114"
##  [5] "Reply from 172.217.0.174: bytes=32 time=28ms TTL=114"
##  [6] "Reply from 172.217.0.174: bytes=32 time=26ms TTL=114"
##  [7] ""
##  [8] "Ping statistics for 172.217.0.174:"
##  [9] "    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),"
## [10] "Approximate round trip times in milli-seconds:"
## [11] "    Minimum = 26ms, Maximum = 30ms, Average = 27ms"
```

Developing an intuition [2]

Dissecting the IP Address

- IP addresses is one of the most fundamental security domain elements
- IPV4 is comprised of four bytes, and is commonly stored in a dotted-decimal notation
  - A byte (i.e. 8-bits) can range in values from 0 to 255
  - The total number of IPv4 addresses is equal to: $(2^8)^4 = 2^{32} = 4,294,967,296$
  - Thus, any ip address can be converted to/from a 32-bit integer value
- The conversion from a 32-bit value can be used to save space since each string character is stored in one byte
  - Thus, the storage cost is reduced from $15 \times 8 = 120$ to $4 \times 8 = 32$ bits per IP address (assuming most efficient storage for both object types)

Developing an intuition [3]



Converting the dotted-decimal notation IP to a 32-Bit Integer [1]

To understand how the binary notation can be extended to IP addresses, let us convert the following IP addresses to their 32bit Integer number.

- Q1: "0.0.0.1"
- Q2: "0.0.0.3"
- Q3: "0.0.1.0"
- Q4: "0.0.1.240"
- Q5: "0.0.100.240"

Converting the dotted-decimal notation IP to a 32-Bit Integer [2]

You can check your answer using the iptools package function titled: `ip_to_numeric`.

```
print("Conversion of IP addresses using R")
```

```
## [1] "Conversion of IP addresses using R"
```

```
pacman::p_load(iptools)

ip_to_numeric("0.0.100.240")
```

```
## [1] 25840
```

```
ip_to_numeric("0.0.100.240") %>% numeric_to_ip()
```

```
## [1] "0.0.100.240"
```

Converting the dotted-decimal notation IP to a 32-Bit Integer [3]

```
print('Python Conversion of IP addresses')
```

```
## Python Conversion of IP addresses
```

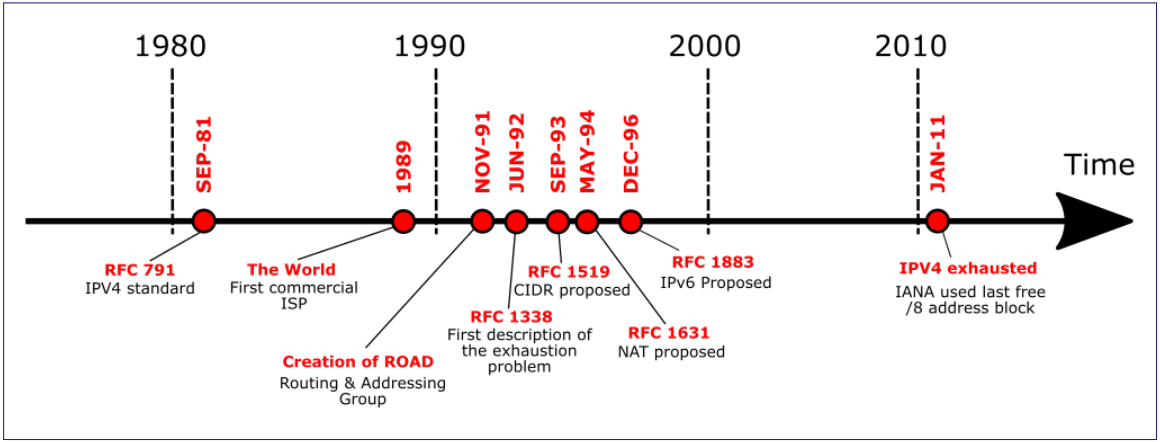
```
import ipaddress

int(ipaddress.ip_address('0.0.100.240'))
```

```
## 25840
```

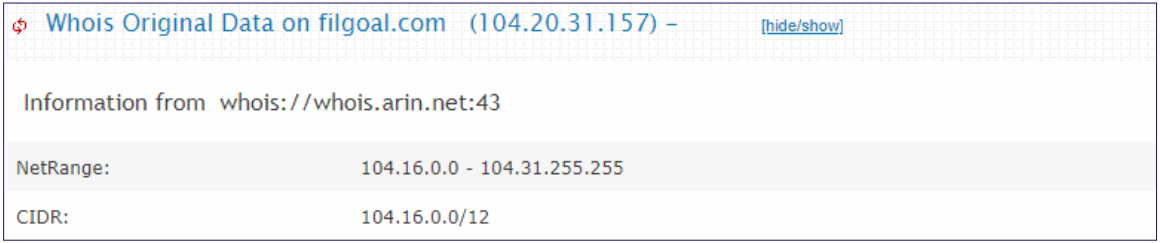
```
str(ipaddress.ip_address(25840))
```

```
## '0.0.100.240'
```

Classless Inter-Domain Routing (CIDR) Blocks/ Ranges [1]



Classless Inter-Domain Routing (CIDR) Blocks/ Ranges [2]



Number of IP addresses in picture: $= 2^{32-12} = 2^{20} = (31 - 16 + 1) * 2^8 * 2^8 = 1,048,576$

```
iptools::ips_in_cidrs("104.16.244.12","104.16.0.0/12") # R
```

```
import ipaddress as ip
ip.ip_address('104.16.244.12') in ip.ip_network('104.16.0.0/12') # Python
```

```
## True
```

Live Demo: Dissecting IP addresses

We will explore working with IP addresses. Specifically, we will work with IP addresses since they are commonly used in Info Security to highlight suspicious domains. Please check AT&T Cybersecurity (2022) and SANS Internet Storm Center (2022) for more details.

By the end of the demo, you should be able to do the following:

- Explain the basics of the IPV4 structure
- Convert IP addresses to their corresponding 32-bit integer representation
- Be able to use online APIs to extract relevant information for a given API/ list of APIs

Prior to our live demo, let us examine some online Miami University IP address information to introduce you to the IP Stack API. You will be required to create an account and utilize the **free** API.

Outline

What I have tried to accomplish in today's class

Learning Objectives

- Describe the data analytics process
- Understand the importance of ETL and data preprocessing
  - Explain what do we mean by technically correct data
  - Explain what do we mean by consistent data
- Perform basic ETL in R and/or Python

References [1]

AT&T Cybersecurity. 2022. "What Is IP/Domain Reputation?" https://cybersecurity.att.com/resource-center/videos/what-is-ip-domain-reputation.

Azevedo, Ana Isabel Rojão Lourenço, and Manuel Filipe Santos. 2008. "KDD, SEMMA and CRISP-DM: A Parallel Overview." *IADS-DM*.

"Choosing the Right Estimator." 2022. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

"Data Analysis." 2022. Wikipedia. https://en.wikipedia.org/wiki/Data_analysis.

De Jonge, Edwin, and Mark Van Der Loo. 2013. *An Introduction to Data Cleaning with r*. Statistics Netherlands Heerlen.

SANS Internet Storm Center. 2022. "Suspicious Domains." https://isc.sans.edu/suspicious_domains.html.

Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. 2009. "A Detailed Analysis of the KDD CUP 99 Data Set." In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–6. IEEE.

ISA 480: Data Driven Security

04 - An Overiew of ETL Operations

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022