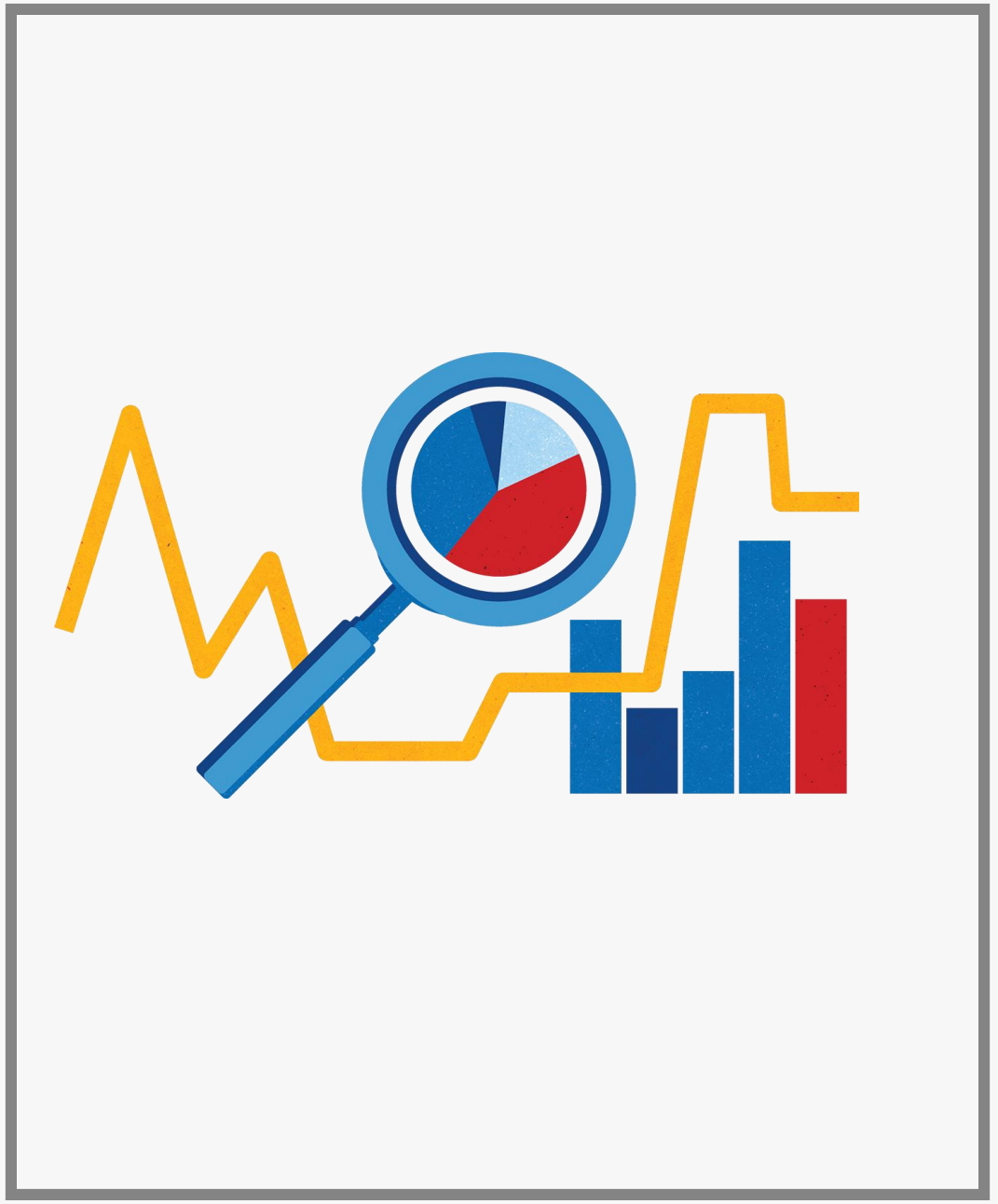
By
Dr. Hussein Hazimeh

Lebanese University
Faculty of Information

Data analysis

March - 2022



Agenda

- Whats is data?
- Data facts
- Data types
- Data sources
- Exploratory & confirmatory data analysis
- Predictive data analysis
- Text analytics
- Data integration
- Data analysis process

What is Data?

Whats is Data?

» What is the term "**data**" means?

- A collection of text, numbers or symbols in raw or unorganised form.
- Data therefore has to be processed, or provided with a context, before it can have meaning.

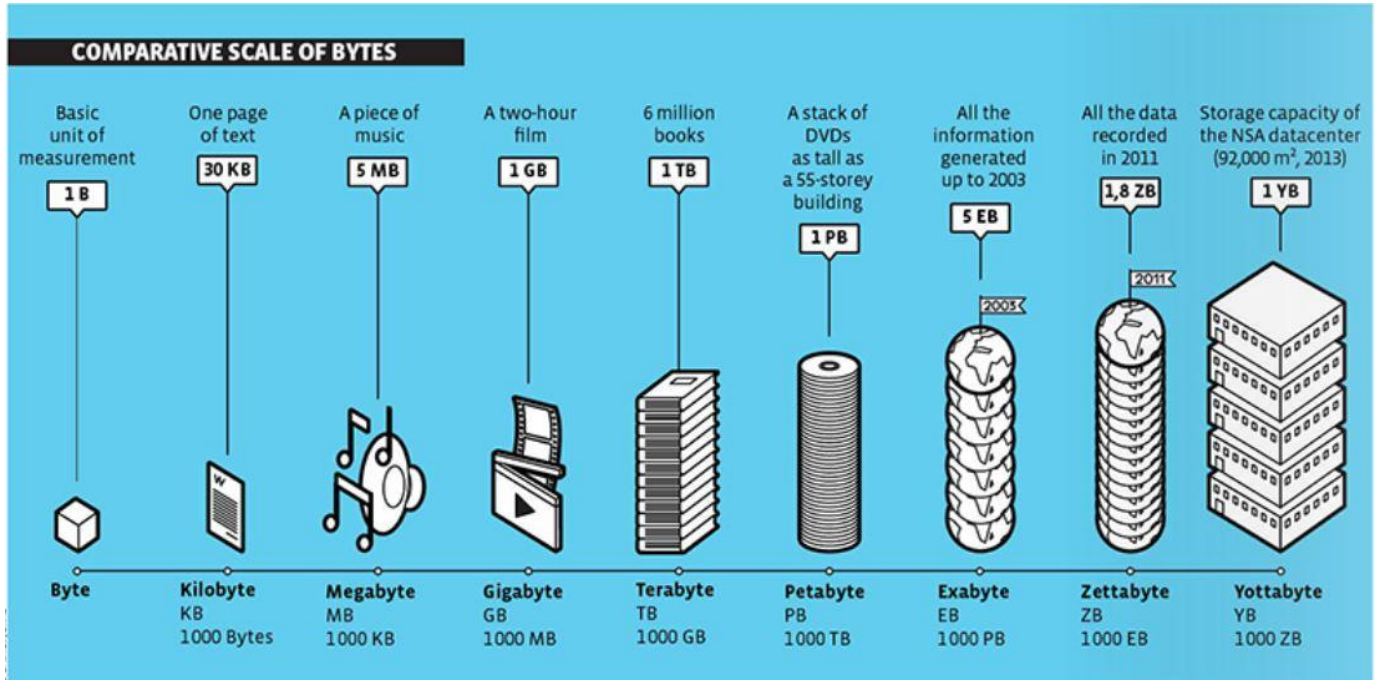
Example
- 3, 6, 9, 12
- cat, dog, gerbil, rabbit, cockatoo
- 161.2, 175.3, 166.4, 164.7, 169.3

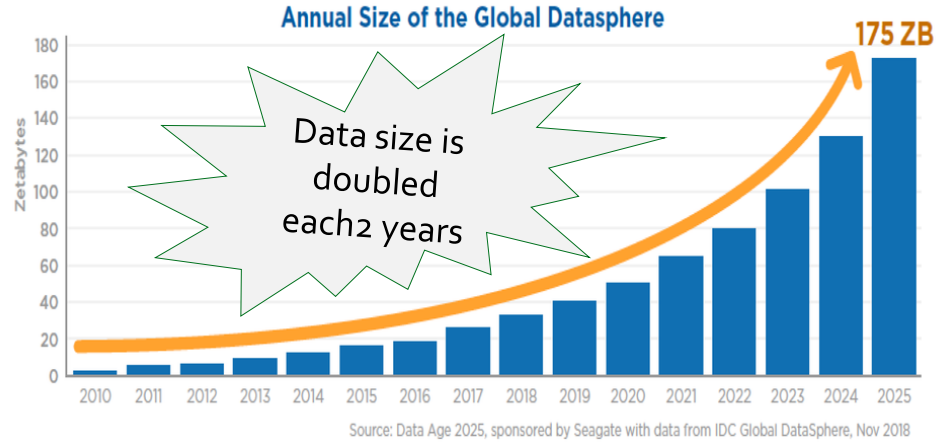
These are meaningless sets of data. They could be the first four answers in the 3 x table, a list of household pets and the heights of 15-year-old students but without a context we don't know.

Data facts

- How is data size is measured?
  - In Computer systems, data is measured in bits & bytes
- the size of the data in increasing.

Name	Equal to:	Size in Bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terrabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176





Data types

» **Data** is divided into 4 types

1. Nominal
2. Ordinal
3. Interval
4. Ratio

» **Nominal**

- Nominal data has no distance between its values, it can not be ordered or measured. Hence, we can not calculate its mean or median.
- Nominal data can be analyzed using the grouping method. The variables can be grouped together into categories, and for each category, the frequency or percentage can be calculated. The data can also be presented visually such as by using a pie chart.
- E.g. eye color, gender, player number

» **Ordinal**

- Dislike nominal data, ordinal data can be ordered, however, alike nominal data it has no distance between values.
- We can calculate the median and mode, but not the mean.
- E.g. ranking, winners in a race

» **Interval**

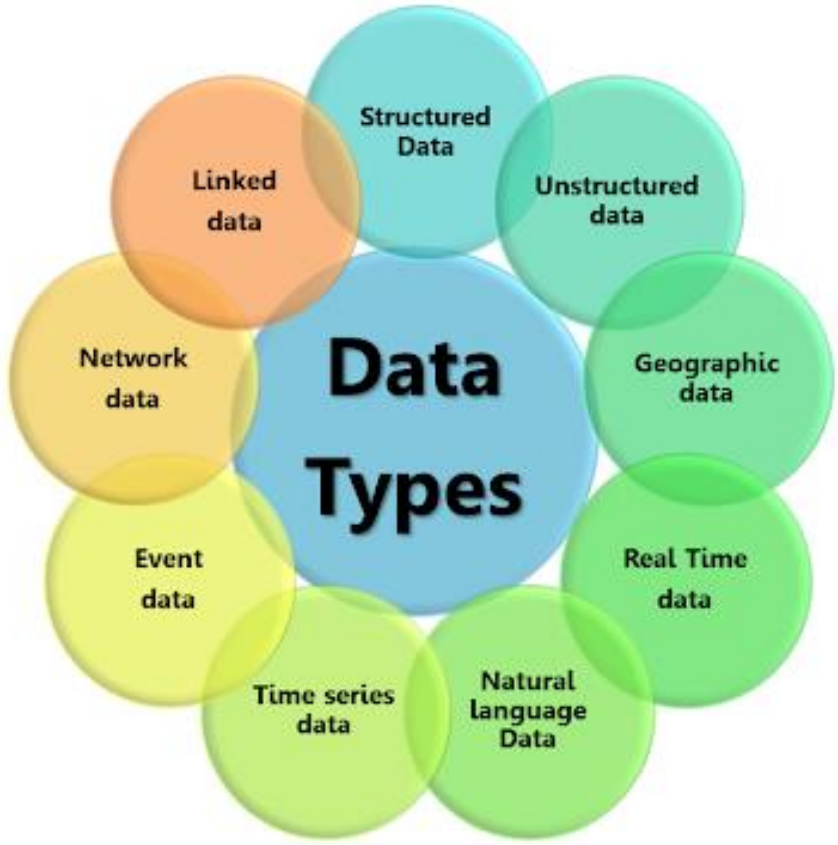
- Interval data has an order between its values and has a distance. They don't have true zero.
- We can calculate the mean, mode, median, and standard deviation.
- E.g. temperature

» **Ratio**

- Ratio data are interval but they also have an absolute zero.
- We can calculate the mean, mode, median, and standard deviation.
- E.g. weight, age, length.

Contextual data types

- Contextual data is any relevant facts from the environment.
- This can include data from customer interactions, social media, weather, news, broader market changes, internet of things (IoT) devices, and geography.
- **E.g.**
  - Weather data
  - Time series data
  - Geographical data
  - Medical data
  - Event data

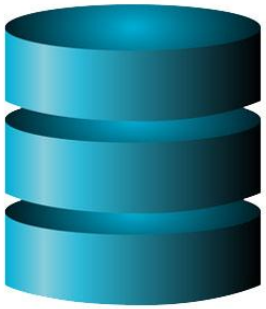


Where do the data come from?

- Sources of the **data**:
  - Database
  - Observations
  - Interviews
  - Surveys
  - Sensors
  - Social media
- **Data** can be generated from different resources

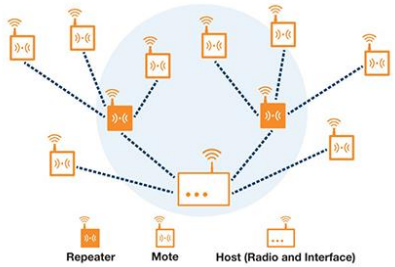
Emergent Data Sources

Social networks



Databases

Cameras

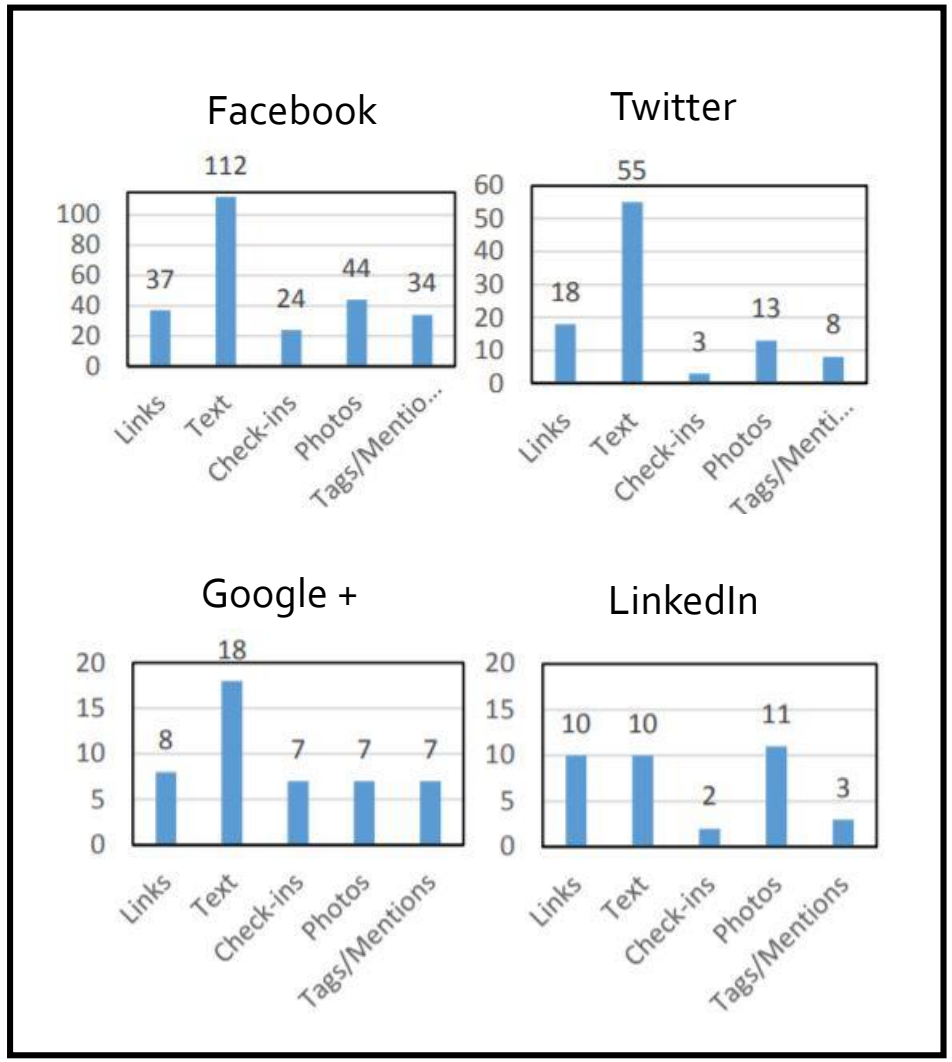


WSN

IoT

Exploratory & confirmatory data analysis (EDA & CDA)

- Is the very first step in a data project.
- **EDA** is important because it allows the investigator to make critical decisions about what is interesting to follow up on and what probably isn't worth pursuing because the data just don't provide the evidence (and might never provide the evidence, even with follow up).
- **E.g.** analysing data shared on different social media platforms.
- **CDA** is the task of confirming certain hypothesis if it is true or false.
- https://bookdown.org/rdpeng/exdata/managing-data-frames-with-the-dplyr-package.html#data-frames



Predictive data analysis

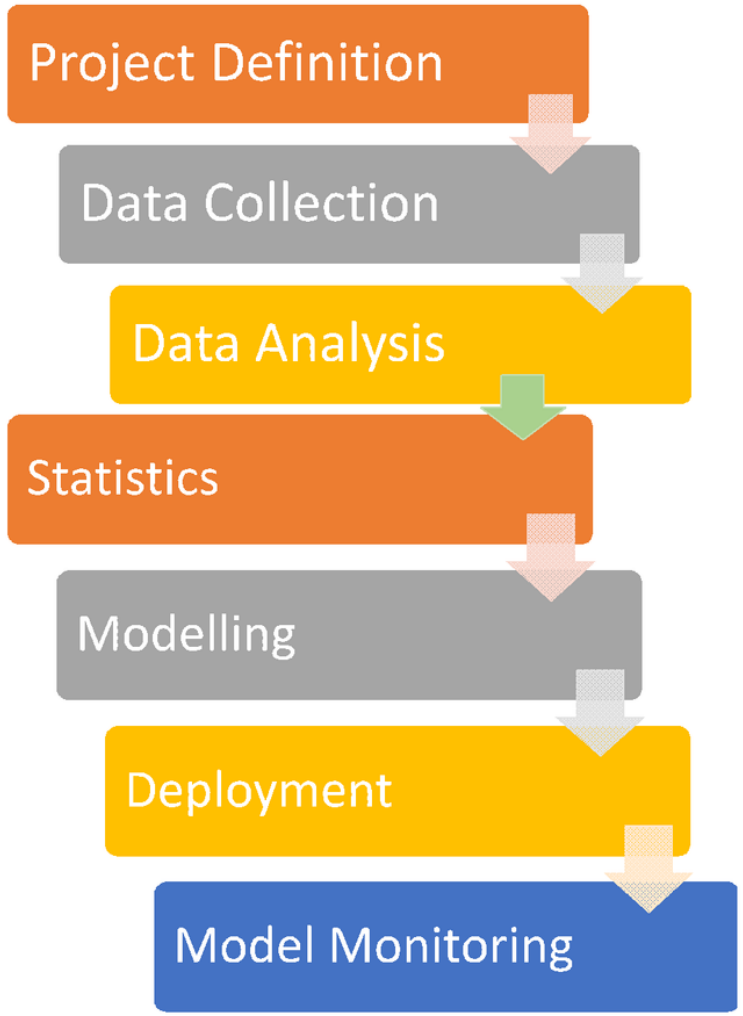
» **Prediction** is the science of estimating, or predicting a future value given the current and old facts of an object.

» **Real-world cases**:

- Fraud detection
- Inventory prediction
- Project risk management
- Reaction prediction in social media

» **Techniques**

- Regression analysis (linear and logistic)
- Time series models
- Decision trees
- Neural networks
- Machine learning (SVM, Naïve Bayes, KNN)



Natural text analytics

» **Text analytics** or text mining is the task of automatically extracting high-quality and useful information from natural text.

» **Typical examples** can be websites, emails, books, and articles.

» **Traditional operations** for text analytics process include:

- Tagging
- Named entity recognition (http://nlp.stanford.edu:8080/ner/)
- Stemming
- Coreference resolution

» **Advanced topics**

- Sentiment analysis (http://nlp.stanford.edu:8080/sentiment/rntnDemo.html)
- Topic extraction
- Word Sense Disambiguation

» **Algorithms** used in text analytics

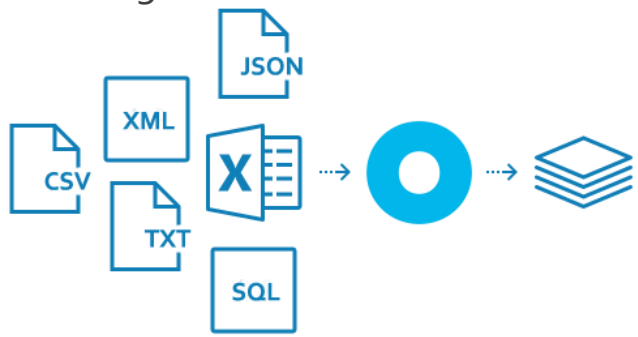
- Latent Dirichlet Allocation (**LDA**)
- **PageRank**
- Vector Space Model (**VSM**)
- **N-grams**
- Conditional Random Fileds (**CRF**)
- Hidden Markov Models (**HMM**)

» **N.B.** Text analytics and data extraction from database are two totally different subjects.

Data integration

- **Data integration** is the problem of combining data residing at different sources, and providing the user with a unified view of these data.
- **Challenges**: integration across different types of resources. No unified format, lacking to portability, code reuse problems, inconsistency of resources.
- **Data integration** task requires a pre-task called data matching, and data merging.
- **Data matching** is the task of testing the similarity across different data values, and validate the conformity of them.

- **Applications**
  - Search engines
  - Booking websites
  - Scholarly data integration
- **Algorithms** for data matching:
  - Similarity models
    - Cosine, tf-idf, Jaccard, RGB Histogram, etc.
  - Topic similarity
    - LDA
  - SQL matching



Data sets

What is a dataset?

- A dataset is a collection of data. E.g. A database.
- A dataset is composed of a collection of variables, and values for each one of them.
- For e.g. dataset listing the degree and years of experience of university instructors.
- Each dataset to be ready for use it must be clean and with high quality.
- Typical dataset extensions: txt, csv, sql, kml, json, xml, etc.
- http://www.kaggle.com

```
<when>2017-05-03T21:38:13Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:38:14Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:38:33Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:38:48Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:39:03Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:39:18Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:39:33Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:39:48Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:40:03Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:40:18Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:40:33Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:40:51Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T21:56:24Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T22:05:02Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
<when>2017-05-03T22:05:39Z</when>
<gx:coord>7.1290217 46.7994157 0</gx:coord>
```

Structured, and unstructured data

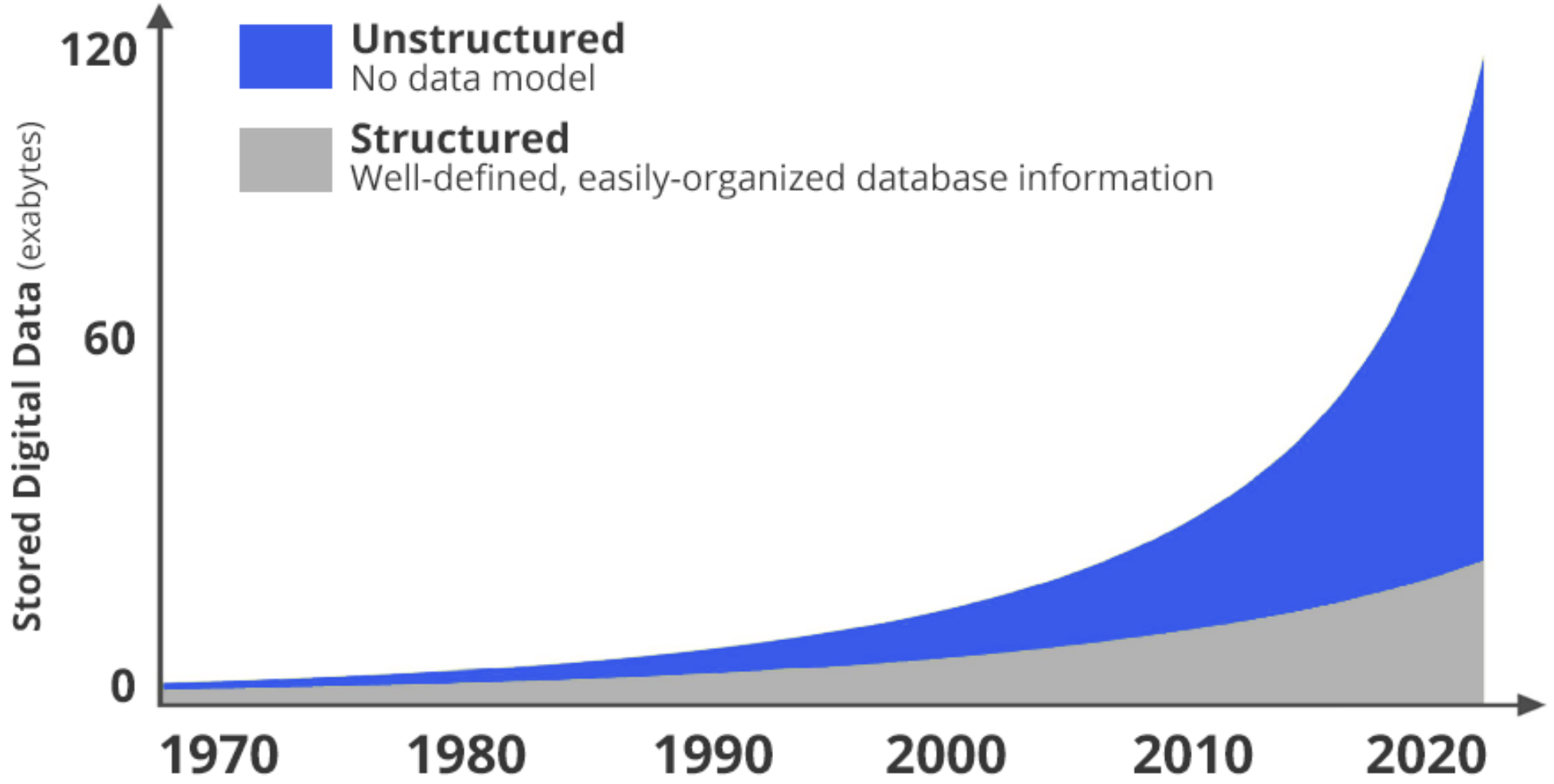
Structured data

- Structured data — or quantitative data — is the type of data that fits nicely into a relational database. It's highly organized and easily analyzed. Most IT staff are used to working with structured data.
- E.g. DOB, ZIP codes, mobile numbers, age.
- Stored usually inside: database, json files, xml files, **excel sheets**.

Unstructured data

- Unstructured data — or qualitative data — is just the opposite. It doesn't fit nicely into a spreadsheet or database. It can be textual or non-textual. It can be human- or machine-generated.
- E.g. media, audio, social media data, chats, sms, calls, etc.
- Those examples are largely human-generated, but machine-generated data can also be unstructured: satellite images, scientific data, surveillance images and video, weather sensor data.

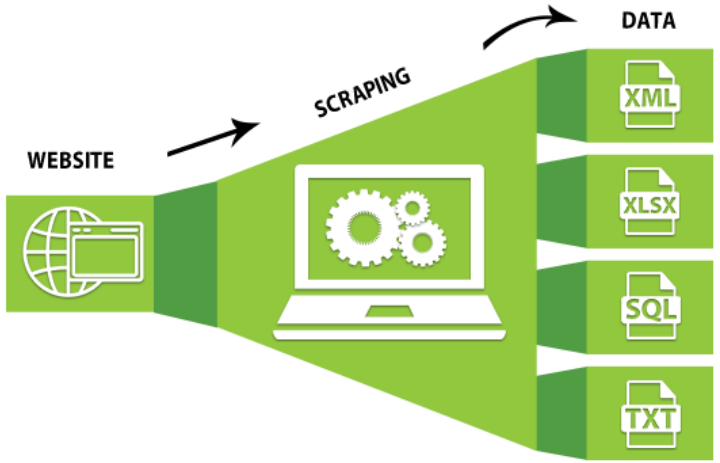
Structured, and unstructured data (cont.)

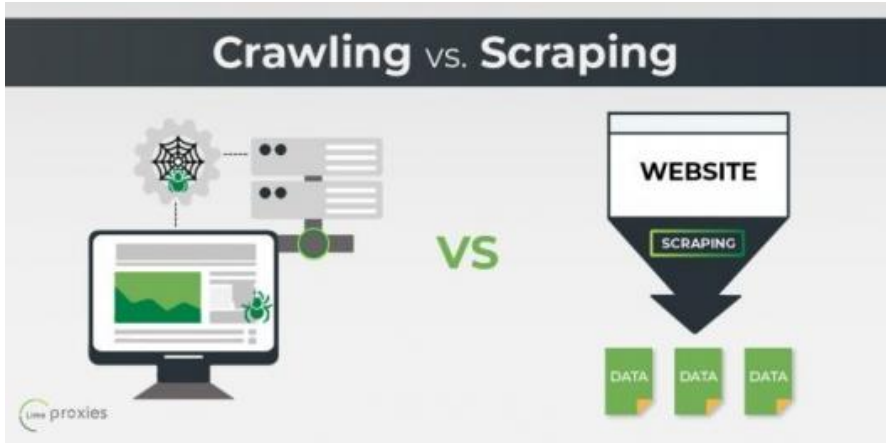


What is Data Scraping?

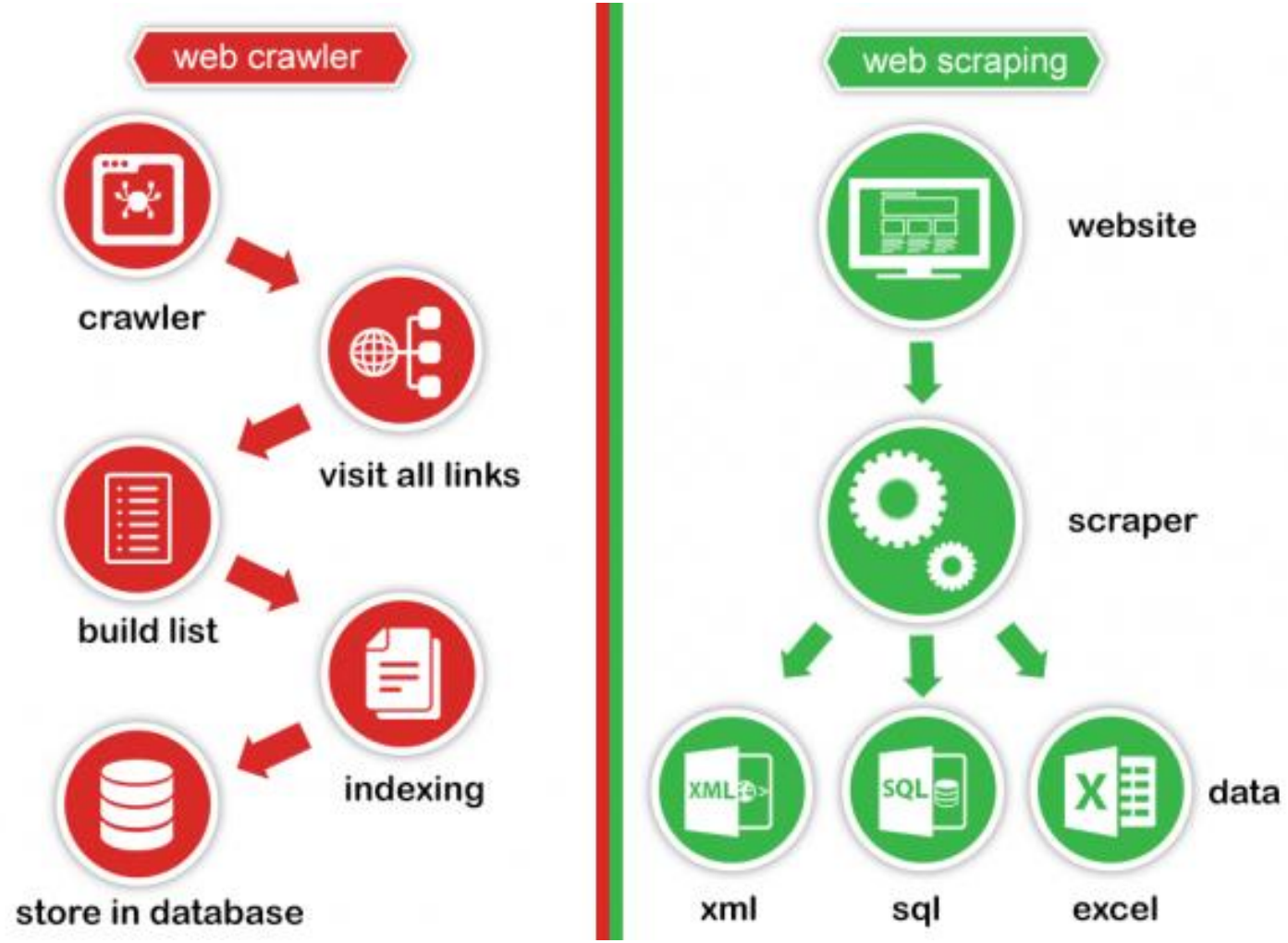
Whats is Data Scraping?

- **Data scraping** is the process of collecting pieces of data from one or multiple data sources.
- A typical example is the extraction of data from web pages.
- **Web scraping** is the part of data scraping that is oriented to web page data collection.
- **Data crawling** is different from data scraping.
- It is the process of navigating and visiting web pages, typically via links.
- Data scraping is a part of the data crawling process.
- **Data scrapping results** can be saved in different formats such as excel, json, xml, txt, etc.



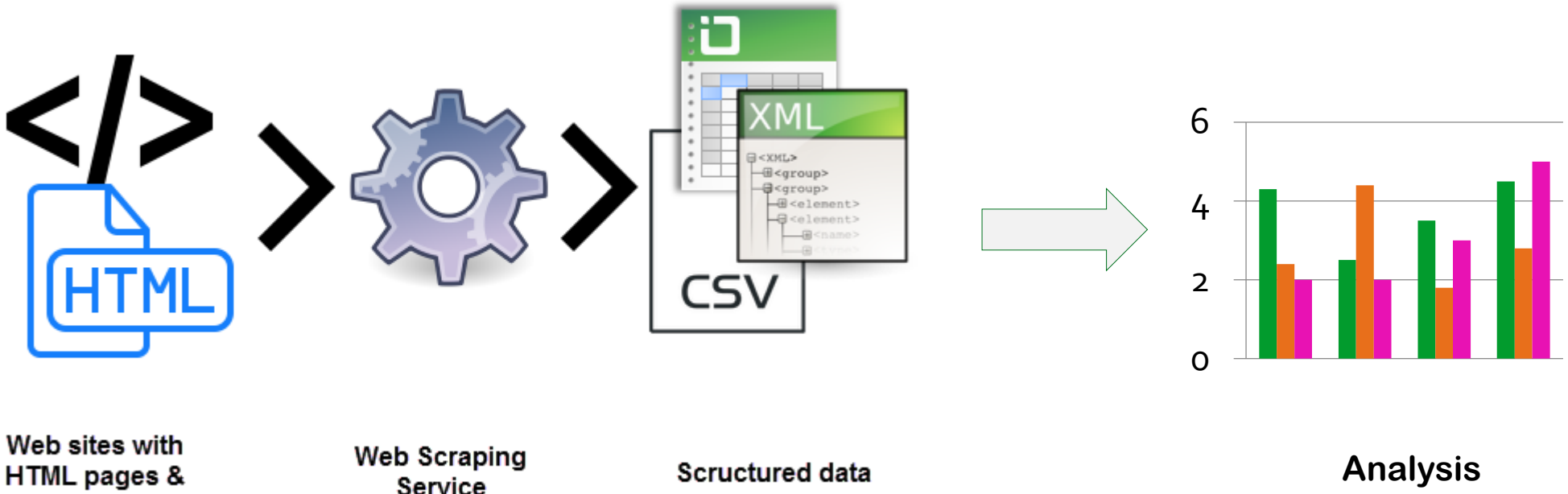


Data Scraping vs Data Crawling?



Data Scraping vs Data Analysis?

» Data scraping is a part of the data analysis process.



Data Scraping techniques?

- **Copy-pasting**
- **DOM Parsing**
- **HTTP Programming**: Using socket programming, posting HTTP requests can help one retrieve dynamic as well as static web page information.
- **Web scraping Software**: It can automatically retrieve the information off the web page, convert it into recognizable information, and store it in a local database.
  - Selenium [https://www.selenium.dev/]

Data Scraping in R?

- R provides multi-puropose libraries for extracting data from different formats.
  - Databases
  - Web pages
  - Text documents
  - Json files
  - XML files

- It provides also very useful libraries for extracting web data
  - RSelenium
- The libraries can be simply implemented.

What is Selenium?

- **Selenium** primarily it is for automating web applications for testing purposes, but is certainly not limited to just that.
- Boring web-based administration tasks can (and should) also be automated as well.

- **Selenium IDE**: Selenium IDE is a complete integrated development environment (IDE) for Selenium tests. It is implemented as a Firefox Add-On and as a Chrome Extension. It allows for recording, editing and debugging of functional tests.

- **Selenium WebDriver:** is the successor to Selenium RC. Selenium WebDriver accepts commands (sent in Selenese, or via a Client API) and sends them to a browser. This is implemented through a browser-specific browser driver, which sends commands to a browser and retrieves results.

- **Selenium Grid**: is a server that allows tests to use web browser instances running on remote machines. With Selenium Grid, one server acts as the hub. Tests contact the hub to obtain access to browser instances.

What is RSelenium?

- The goal of **RSelenium** is to make it easy to connect to a Selenium Server/ Remote Selenium Server from within R.
- RSelenium provides R bindings for the Selenium Webdriver API.
- **RSelenium** allows you to carry out unit testing and regression testing on your webapps and webpages across a range of browser/OS combinations.
- This allows us to integrate from within R testing and manipulation of popular projects such as shiny, sauceLabs.
- Installation:
  - install.packages("RSelenium")

Rselenium tutorial

How do I connect to a running server?

`RSelenium` has a main reference class named `remoteDriver`. To connect to a server you need to instantiate a new `remoteDriver` with appropriate options.

```
# RSelenium::startServer() if required
require(RSelenium)
remDr ← remoteDriver(remoteServerAddr = "localhost"
                     , port = 4444
                     , browserName = "firefox"
                     )
```

Rselenium web data extraction

Finding web elements by name:

```
con$navigate("http://www.google.com/ncr")
webElem <- con$findElement(using = "name", value = "q")
webElem$getElementAttribute("name")
```