Data Mining
and
Information Retrieval

Data Preprocessing and Exploratory Data Analysis

DATA SCIENCE & ANALYTICS

Data Preprocessing

Data may come from a variety of sources

- Flat Files (CSV) or Structures (DB Tables)
- Non-Flat, hierarchical (JSON, HTML, XML, etc)

It is often critical to explore and pre-process data before feeding it into more advanced tools

- Dashboards, Statistical Models, Data Visualizations, Machine Learning, etc.

Reflection

What is are some of the most challenging data you have to deal with, and why?

Data Carpentry

Data Carpentry is often of critical importance

- Data Cleaning
- Data Transforming
- Data Restructures

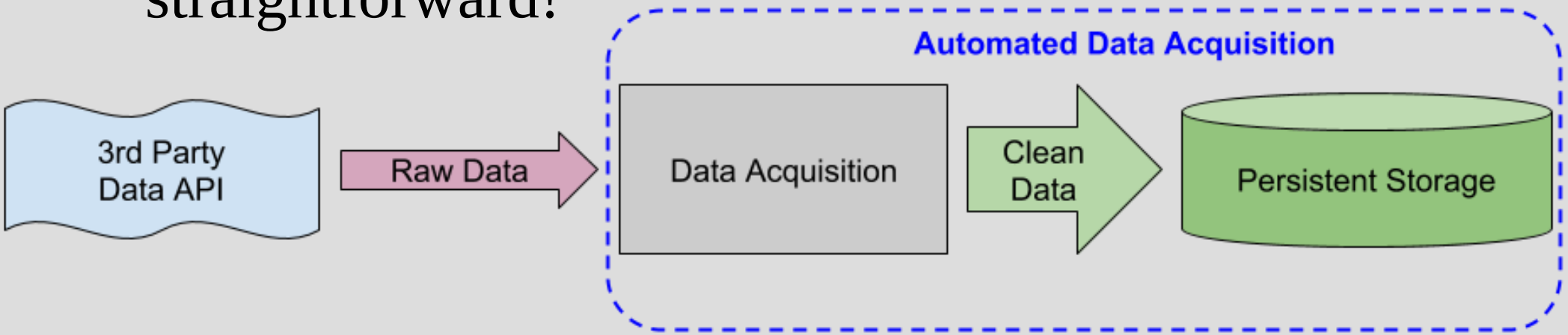
Data Extraction often the first step:

- Scraping sources (HTML, APIs, etc.)

Data Acquisition

Automated Data Acquisition from third parties through use of Web API is a powerful capability!

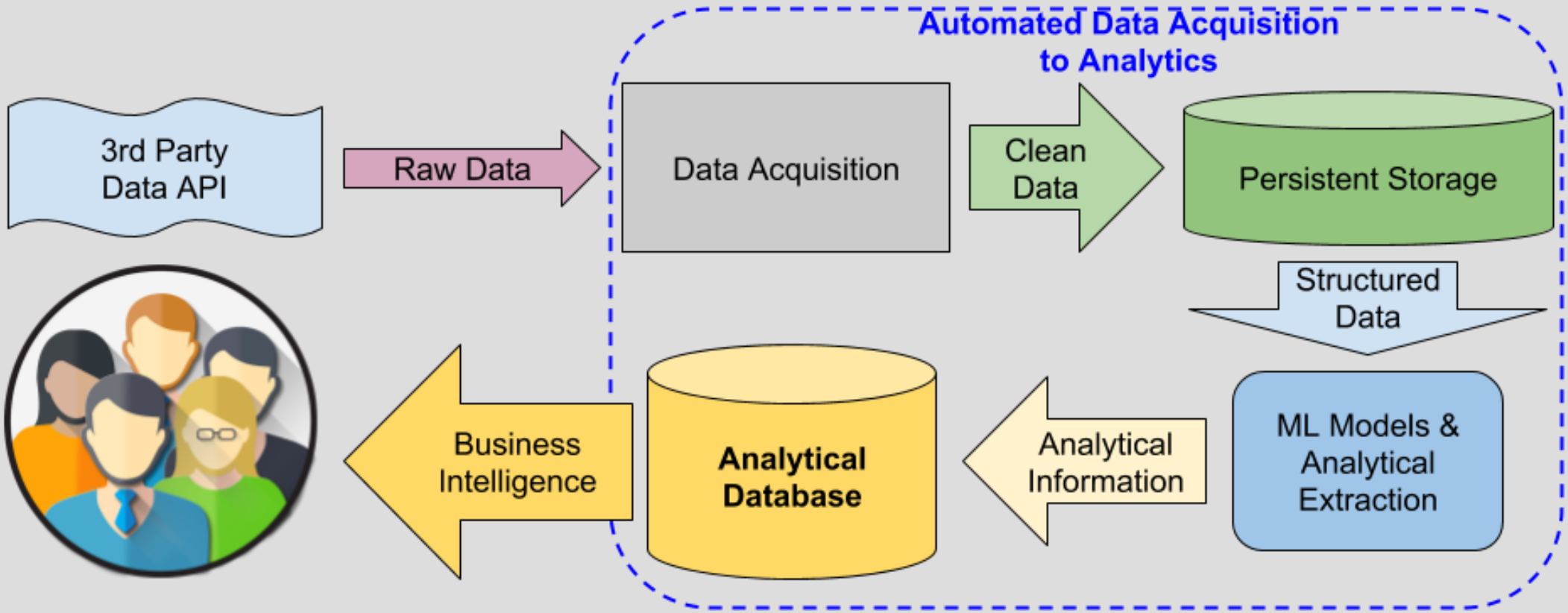
- Programming languages, such as Python, make this easy!
- Persistence using libraries (e.g., SQLAlchemy) also straightforward!



Persisted Data to Analytics

Data persisted becomes a repository!

How can this data be used for analytical purposes?



Module 1 Labs

Module 1 is a warm up to refresh some concepts and techniques you have seen previously:

- Data Cleaning
- Data Extraction
- Web Scraping
- Exploratory Data Analysis

Complete Labs, then reflect:

- What are some of key libraries leveraged within the labs?
- What are the key takeaways?

Module 1 Practice

Work through practice!

Please discuss and collaborate on ideas (not code) within Slack

Post-Practice Reflections:

- What was the challenging portion?
- How comfortable are you with web scraping, data cleaning, and EDA?

Module 1 Exercise

Cleaning and Exploring

NYC_death_causes

Please discuss and collaborate on ideas (not code) within Slack

Ensure you are giving yourself enough time to work on code and the brief annotation answers you must provide

DATA SCIENCE & ANALYTICS

Questions?