DATA COLLECTION TECHNIQUES IN DATA SCIENCE

AGENDA

- Why Data Collection is Critical?
- Types of Data Sources: APIs, Webpages, Databases, Files, Sensors
- Ethical Considerations (Consent, Licensing, Fair Use)
- Overview of Tools: APIs, Web Scraping (BeautifulSoup), Webhooks, etc.
- Techniques in Python for data collection

WHY DATA COLLECTION IS CRITICAL?

WHY DATA COLLECTION IS CRITICAL

- "Garbage In, Garbage Out" — Quality of data defines the quality of your analysis.
- Data is the foundation for Machine Learning, Analytics, and AI.
- Good data enables better decision-making and product improvements.
- 80% of a data scientist's time is spent on collecting and cleaning data.

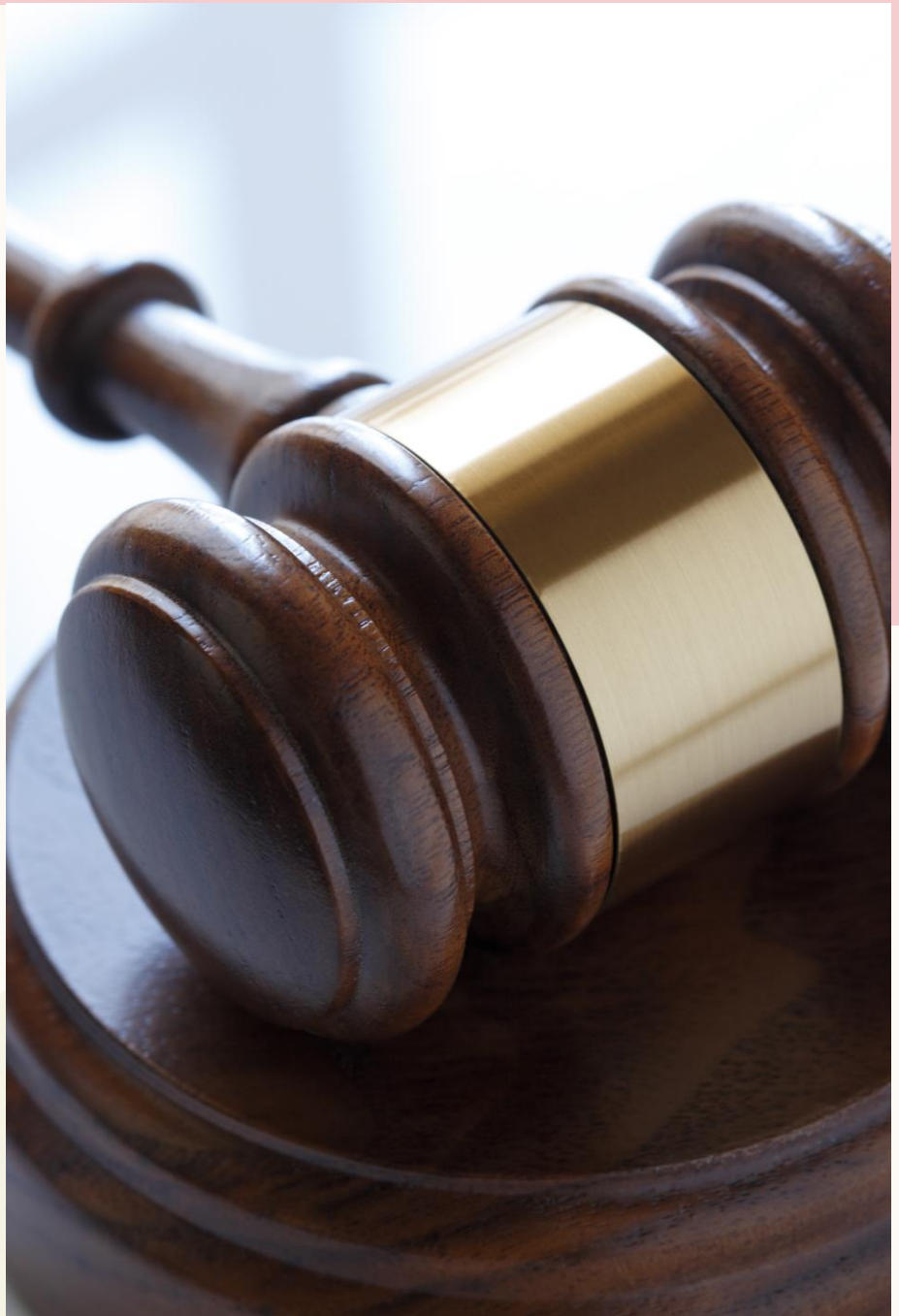
TYPES OF DATA SOURCES

TYPES OF DATA SOURCES

- **APIs:** Pre-organized data access (e.g., weather API, stock API)
- **Databases:** SQL, NoSQL systems (e.g., MySQL, MongoDB)
- **Webpages:** HTML content, blogs, product pages
- **Files:** CSV, JSON, XML, Excel sheets
- **Sensors/Logs:** IoT devices, server logs, mobile app events

ETHICAL CONSIDERATIONS

(Consent, Licensing, Fair Use)

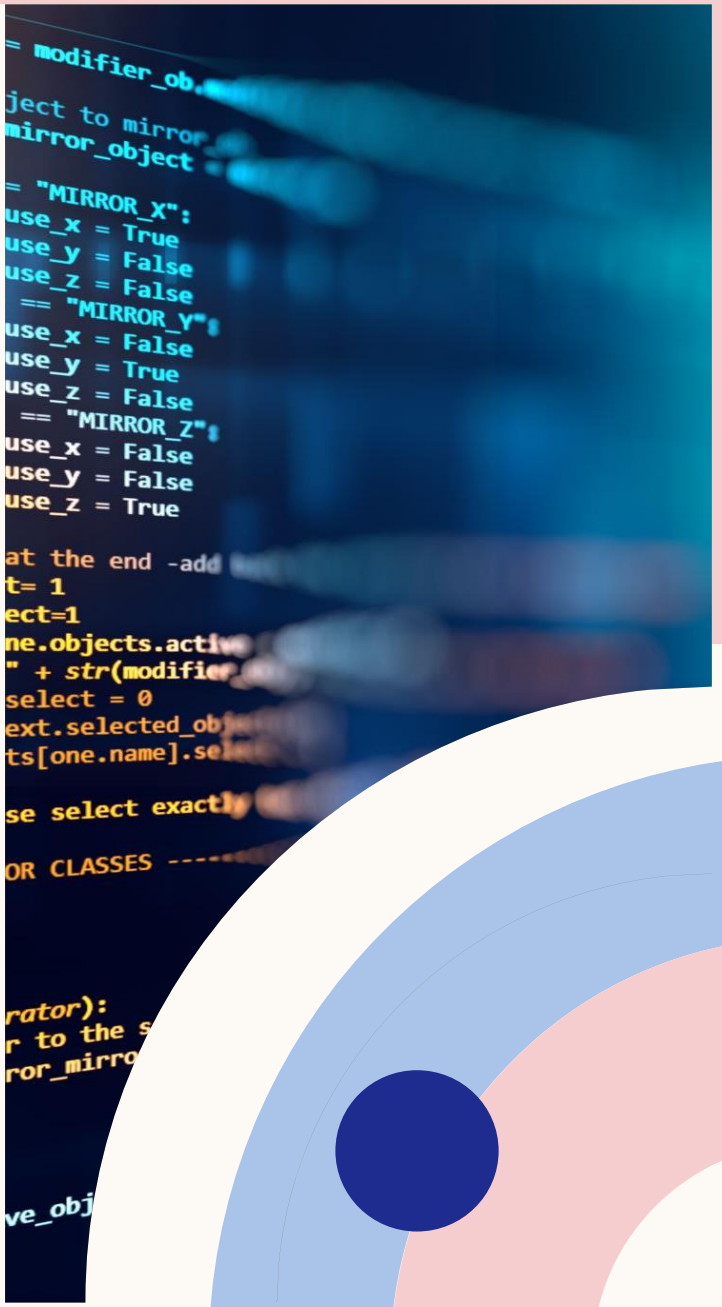


ETHICAL CONSIDERATIONS

- Always check Terms of Service before scraping or collecting data.
- Respect robots.txt files when scraping websites.
- Attribution and Licensing: Can you reuse this data? Give credit where required.
- Consent: If collecting personal data, informed consent is necessary.
- Data collection must comply with laws like GDPR, CCPA.

OVERVIEW OF TOOLS FOR DATA COLLECTION

1. APIs (REST, GraphQL) — Standardized, structured data access
2. Web Scraping: Scraping HTML content (BeautifulSoup, Scrapy, Selenium)
3. Webhooks: Event-driven data collection (e.g., payment success notification)
4. Manual Entry: Forms, surveys (crowd-sourced data)
5. File Handling: Reading local or cloud files



TECHNIQUES IN PYTHON FOR DATA COLLECTION

APIs:

- Use requests library to send HTTP requests and parse JSON/XML responses.

Web Scraping:

- BeautifulSoup + requests for parsing HTML
- Selenium for scraping JavaScript-heavy websites

File Systems:

- Pandas, csv, json libraries to read/write files

Databases:

- Use sqlite3, SQLAlchemy, pymongo for connecting to SQL/NoSQL databases

FINAL TIPS & TAKEAWAYS

- Data collection is the first and most important step in Data Science.
- Know your source, respect ethics, and choose the right tool.
- Python provides powerful libraries for all major types of data collection.

THANK YOU