Chapter 5

Who needs data?

Data Collection & Simulation/Generation

API & Web Scraping

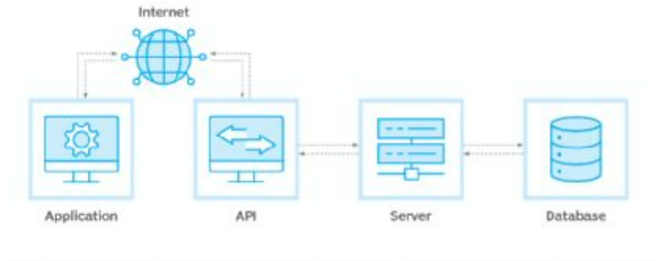
What is an API?

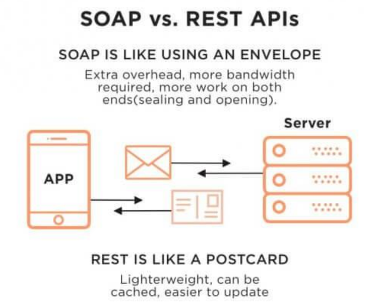
API: **A**pplication **P**rogramming **I**nterface.

Allows different software systems to communicate.

RESTful vs. SOAP APIs:

- **REST (Representational State Transfer):** Lightweight, uses JSON/XML.
- **SOAP (Simple Object Access Protocol):** More rigid, uses XML.



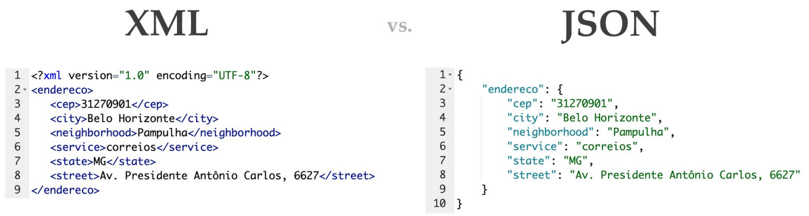

SOAP vs. REST APIs

SOAP IS LIKE USING AN ENVELOPE

Extra overhead, more bandwidth required, more work on both ends(sealing and opening).

REST IS LIKE A POSTCARD

Lighterweight, can be cached, easier to update



```xml
<?xml version="1.0" encoding="UTF-8"?>
<endereco>
   <cep>31270901</cep>
   <city>Belo Horizonte</city>
   <neighborhood>Pampulha</neighborhood>
   <service>correios</service>
   <state>MG</state>
   <street>Av. Presidente Antônio Carlos, 6627</street>
</endereco>
```

```json
{
    "endereco": {
        "cep": "31270901",
        "city": "Belo Horizonte",
        "neighborhood": "Pampulha",
        "service": "correios",
        "state": "MG",
        "street": "Av. Presidente Antônio Carlos, 6627"
    }
}
```

How APIs Work

HTTP Methods:

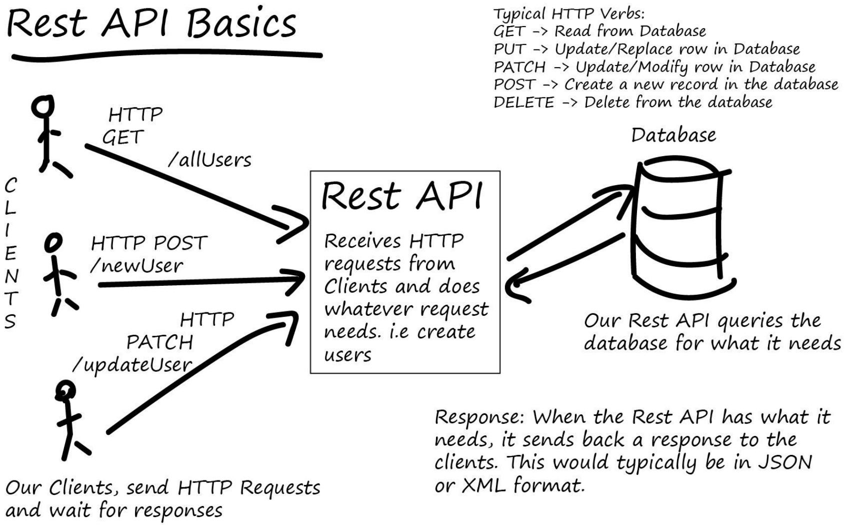
- **GET:** Retrieve data.
- **POST:** Submit data.
- **PUT:** Update data.
- **DELETE:** Remove data.

Components:

- **Endpoint:** URL of the API.
- **Headers:** Metadata about the request.
- **Parameters:** Additional data sent with the request.
- **Body:** Data sent with POST/PUT requests.

Authentication:

- **API keys:** Unique key for accessing the API.
- **OAuth:** Token-based authentication.



Rest API Basics

Typical HTTP Verbs:
GET -> Read from Database
PUT -> Update/Replace row in Database
PATCH -> Update/Modify row in Database
POST -> Create a new record in the database
DELETE -> Delete from the database

Receives HTTP requests from Clients and does whatever request needs. i.e create users

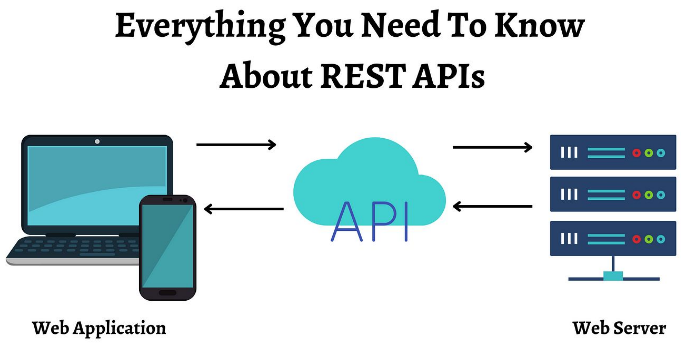
Our Rest API queries the database for what it needs

Our Clients, send HTTP Requests and wait for responses

Response: When the Rest API has what it needs, it sends back a response to the clients. This would typically be in JSON or XML format.

Practical Example:
API Request

Add python snippet



Use Cases of APIs

Integration of Third-Party Services:

- e.g., Payment gateways, social media APIs.

Automation of Tasks:

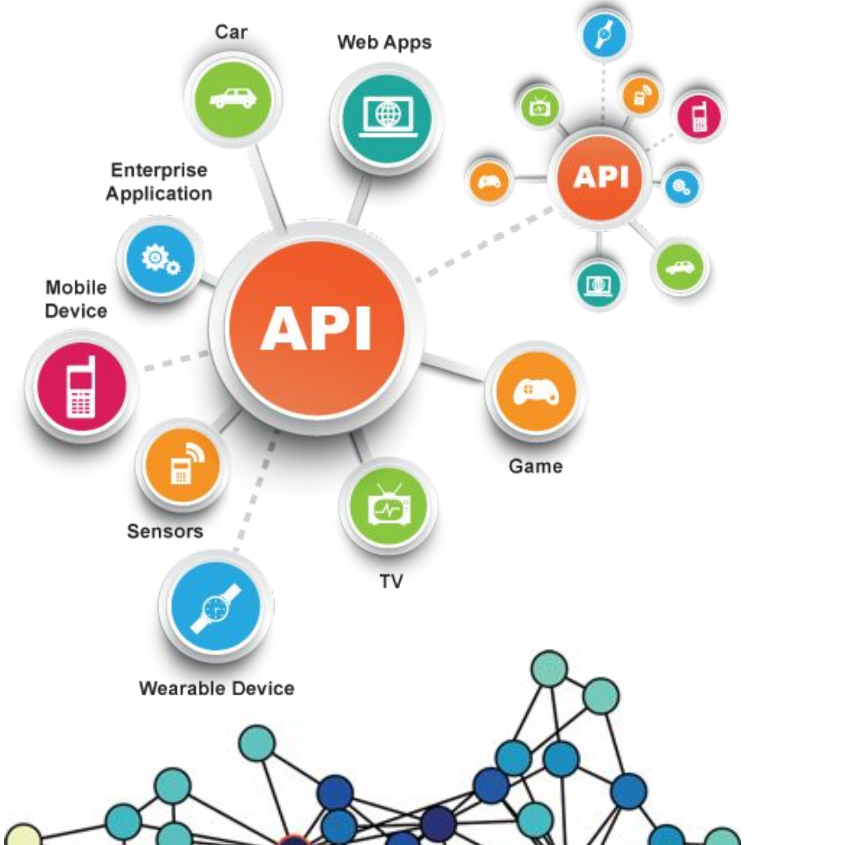
- e.g., Automated reporting, data synchronization.

Enhancement of Application Functionality:

- e.g., Adding weather updates, maps.

Data collection (?)

- Social media (X/Twitter, Meta, Last.fm...)



What is Web Scraping?

Extracting data from websites.

Automating data collection from web pages.

Difference between web scraping and web crawling:

- **Web Scraping:** Extract specific data.
- **Web Crawling:** Index entire websites.

How Web Scraping Works

HTML Structure Basics

- **Tags:** <html>, <head>, <body>, <div>, <p>, etc.
- **Attributes:** id, class, href, src, etc.

Tools and Libraries (Python)

- **Beautiful Soup:** Parsing HTML and XML.
- **Scrapy:** Web scraping framework.
- **Selenium:** Browser automation tool.

HTML WEBSITES
WEB SCRAPING
DATA

Practical Example:
Web Scraping

Add python snippet (BeautifulSoup?)

Challenges and Best Practices

Handling Dynamic Content

- Javascript-driven websites
- Selenium can render JavaScript

Managing Request Limits

- Avoiding bans
- Implementing delays, rotating IP addresses

Ethical Considerations

- Respecting ToS
- Impact on server performances

Legal Considerations

- If it is out there it doesn’t meant you have the right to collect it!

Legal Aspects and Ethical Considerations

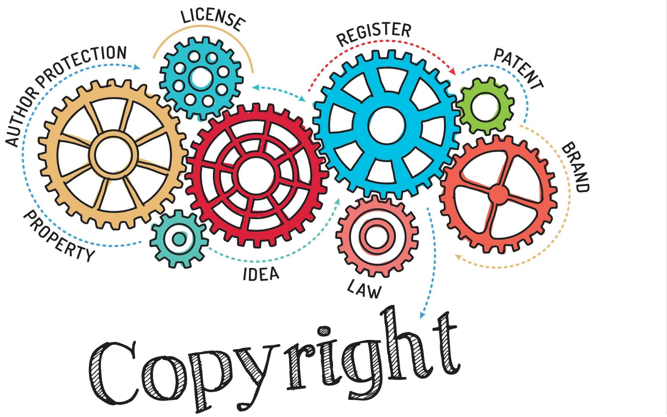
Intellectual Property and Copyright issues

When collecting data through scraping/API:

- consider intellectual property rights, and
- copyright issues.

Data and content on the web are often protected by copyright laws.

Example: Extracting and reusing content from a website without permission could infringe on the owner's intellectual property rights.

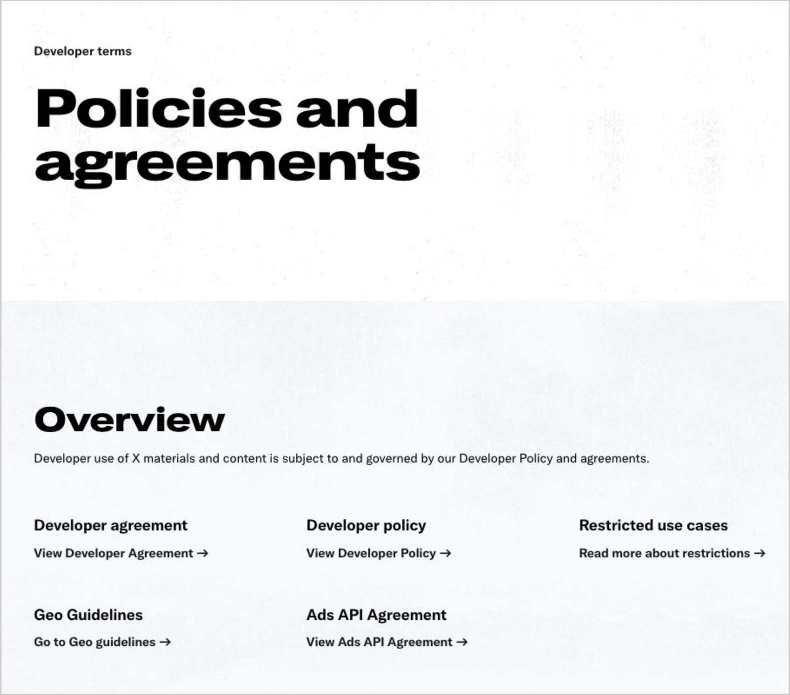


Terms of Service (ToS) Violations

Websites and APIs often have ToS that explicitly:

- prohibit scraping or automated access;
- regulate data usage;
- regulate collection/sharing procedures.

Example: "LinkedIn's ToS, prohibit the use of bots to scrape user data. Violating these terms can lead to legal action, as seen in the case of LinkedIn vs. hiQ."



Developer terms

Policies and agreements

Overview

Developer use of X materials and content is subject to and governed by our Developer Policy and agreements.

Developer agreement

View Developer Agreement →

Developer policy

View Developer Policy →

Restricted use cases

Read more about restrictions →

Geo Guidelines

Go to Geo guidelines →

Ads API Agreement

View Ads API Agreement →

We are in EU: GDPR

Data protection law enacted by the European Union to protect the personal data and privacy of individuals within the EU and the European Economic Area (EEA).

It also addresses the transfer of personal data outside the EU and EEA areas.

GDPR impact
API usage & Web Scraping

Key principles to be ensured:

Lawfulness, Fairness, and Transparency:
Data must be processed lawfully, fairly, and in a transparent manner in relation to individuals.

Purpose Limitation:
Data must be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

Data Minimization:
Data collection should be limited to what is necessary in relation to the purposes for which they are processed.

Accuracy:
Personal data must be accurate and, where necessary, kept up to date. Inaccurate data should be corrected or deleted promptly.

Storage Limitation:
Data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

Integrity and Confidentiality:
Data must be processed in a manner that ensures appropriate security, including protection against unauthorized or unlawful processing and against accidental loss, destruction, or damage.

Accountability:
The data controller is responsible for and must be able to demonstrate compliance with these principles.

We are in EU: DSA

Online platforms must:

- clearly disclose their data collection practices
- inform users about how their data is collected and used.

DSA mandates algorithmic accountability: platforms must explain the functioning of algorithms used in content curation and targeted advertising

The DSA recognizes the importance of **research**.

- Platforms may be required to provide data to researchers under certain conditions.

For Platform related Data-Driven stiìudies:

- Researchers must clearly disclose data collection methods and purposes.
- Data collection should be limited to what is necessary for the research purpose.
- Researchers must implement robust data protection measures.

Anonymization vs GDPR Compliant Pseudonymisation

Anonymization

NO Linkability

Inability to link back to individuals even if authorized, or if it would produce beneficial results, or if people's lives depend on it.

NO Auditability

No data lineage means unable to depict flows and transformations among data sources required for accountable data governance.

Degraded Accuracy

Data fidelity suffers when unable to rely on accuracy, consistency, and completeness of data over time.

Uncontrolled Liability

The risk of liability from unauthorized re-identification extends beyond your organization.

Pseudonymisation

Complete Linkability

Ability to link back to individuals for all authorized purposes.

Complete Auditability

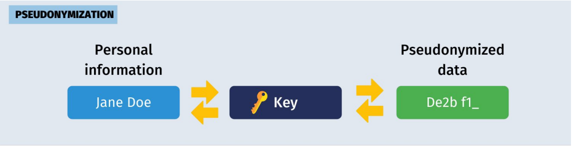
Data lineage enables tracing of flows and transformations among data sources required for accountable data governance.

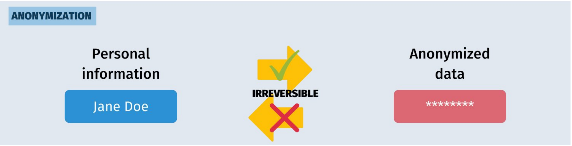
Superior Accuracy

Data fidelity preserved due to accuracy, consistency, and completeness of data over time.

Controlled Liability

The risk of liability is controlled by your organization since it holds the additional information (keys) required for re-identification.





Synthetic Realities:
Simulations and Digital Twins

Have you heard about the **AP(I)-pocalypse**?

(Thanks, Elon, it wasn’t needed though…)

What are Model-Based Simulations?

Definition:
Simulations using computational models to study and predict the behavior of complex systems.

Purpose:
Analyze dynamics that are difficult to observe directly.

Benefits:
Allows testing of hypotheses, prediction of future trends, and analysis of "what-if" scenarios.

Example:
Opinion Dynamics

Study of how opinions form and evolve within a network of individuals.

Key Concepts:
Consensus, polarization, and influence.

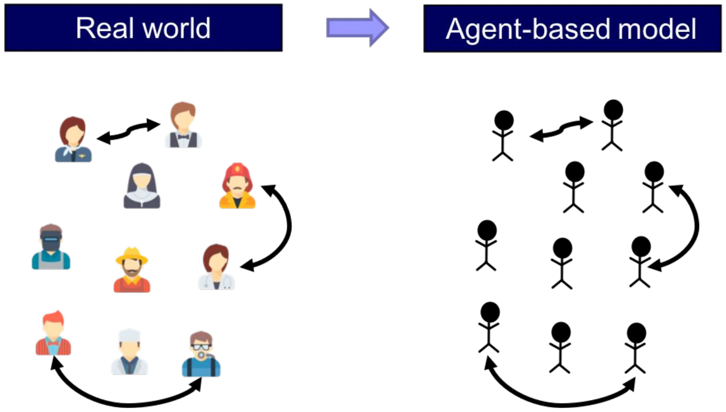
Applications:
Political campaigns, marketing, social movements, etc.

Simulation Techniques

Agent-Based Simulation:
Simulates interactions of autonomous agents to assess their effects on the system.

Monte Carlo Simulation:
Uses random sampling to understand the behavior of a system.

Mean Field Approximations:
Simplifies a large network to study the average behavior of the system.



Challenges and Limitations

Model Complexity:
Simplified models may not capture all real-world complexities.

Computational Resources:
Large-scale simulations require significant computational power.

Validation:
Difficult to validate models against real-world data.

Research Directions

Integration with Real Data:
Using social media and other real-world data to enhance model accuracy.

Hybrid Models:
e.g., Combining different models to capture various aspects of opinion dynamics.

Machine Learning:
e.g., Incorporating machine learning techniques to predict opinion changes.

Policy Implications:
e.g., Simulations to inform policy decisions on issues like misinformation and public opinion management.

Digital Twins

A Digital Twin is a virtual representation of a physical entity or system.

Purpose:
To simulate, predict, and optimize the performance of the physical counterpart.

Applications:
Used in various fields such as manufacturing, healthcare, urban planning, and social network analysis.

Integrating Digital Twins in SNA

Creating a Digital Twin of a social network to mirror real-world social interactions and dynamics.

Benefits:

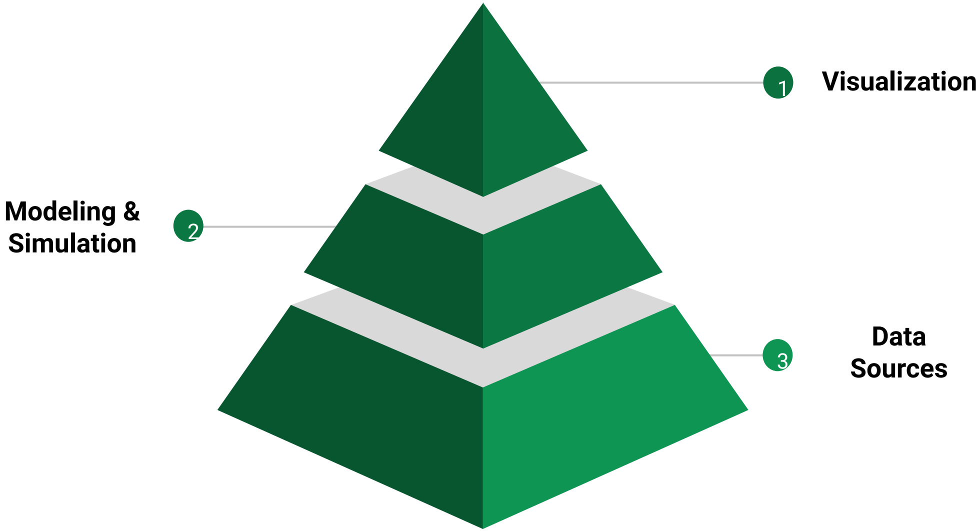
- Enhanced analysis and prediction capabilities.
- Real-time monitoring and intervention.

Component of a Social Network Digital Twin

Data Sources (to fit the simulation)**:**
Social media, surveys, communication logs, etc.

Modeling and Simulation:
Using algorithms to replicate users' behavior (social and contents).

Visualization:
Graphical representation of the digital twin for analysis.



1 Visualization

Modeling & Simulation 2

3 Data Sources

Benefits of Digital Twins in SNA

Enhanced Predictive Analytics:
Ability to forecast social trends and behaviors.

Real-time Monitoring:
Continuous updates provide current insights.

Scenario Testing:
Test various scenarios and interventions before implementation.

Personalization:
Tailored insights for different groups or individuals.

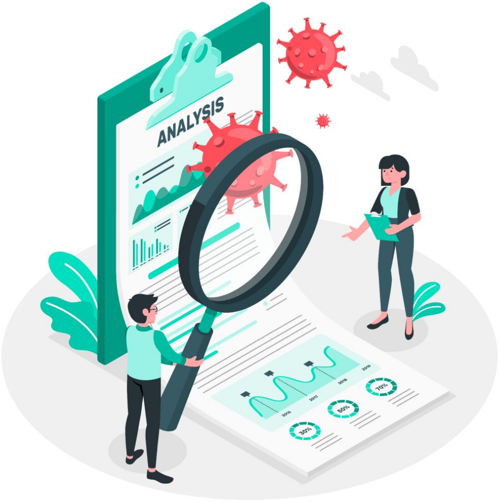
Challenges and Considerations

Data Integration:
Combining data from diverse sources.

Model Accuracy:
Ensuring the digital twin accurately mirrors real-world dynamics.

Population size scalability:
Few detailed agents vs. OSN-like population size

Computational Complexity:
Agents can be particularly complicated entities
(e.g. LLMs)



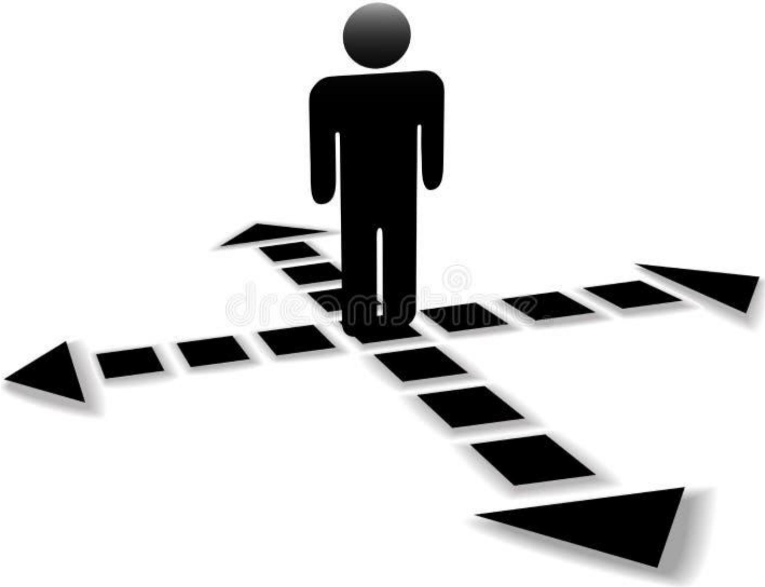
Research Directions

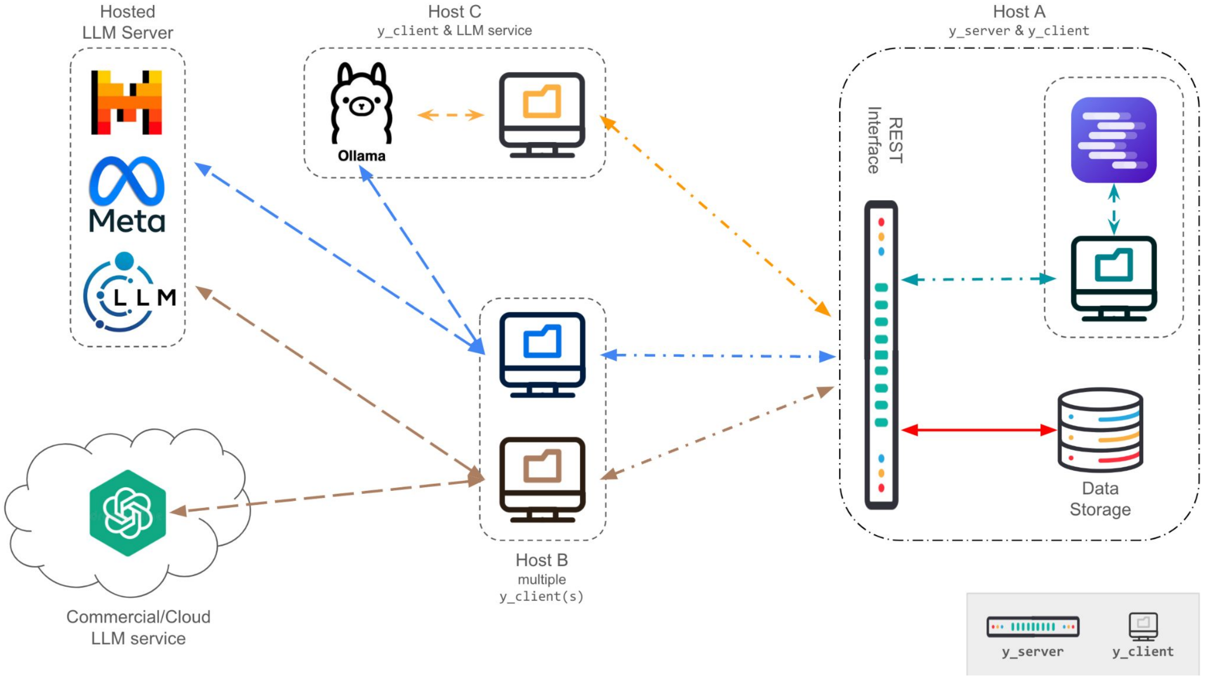
Integration with AI and Machine Learning:
Enhancing predictive capabilities and automation (e.g., LLMs).

Scaling Up:
Applying digital twins to larger and more complex social networks.

Cross-disciplinary Applications:
Integrating insights from psychology, sociology, and data science.

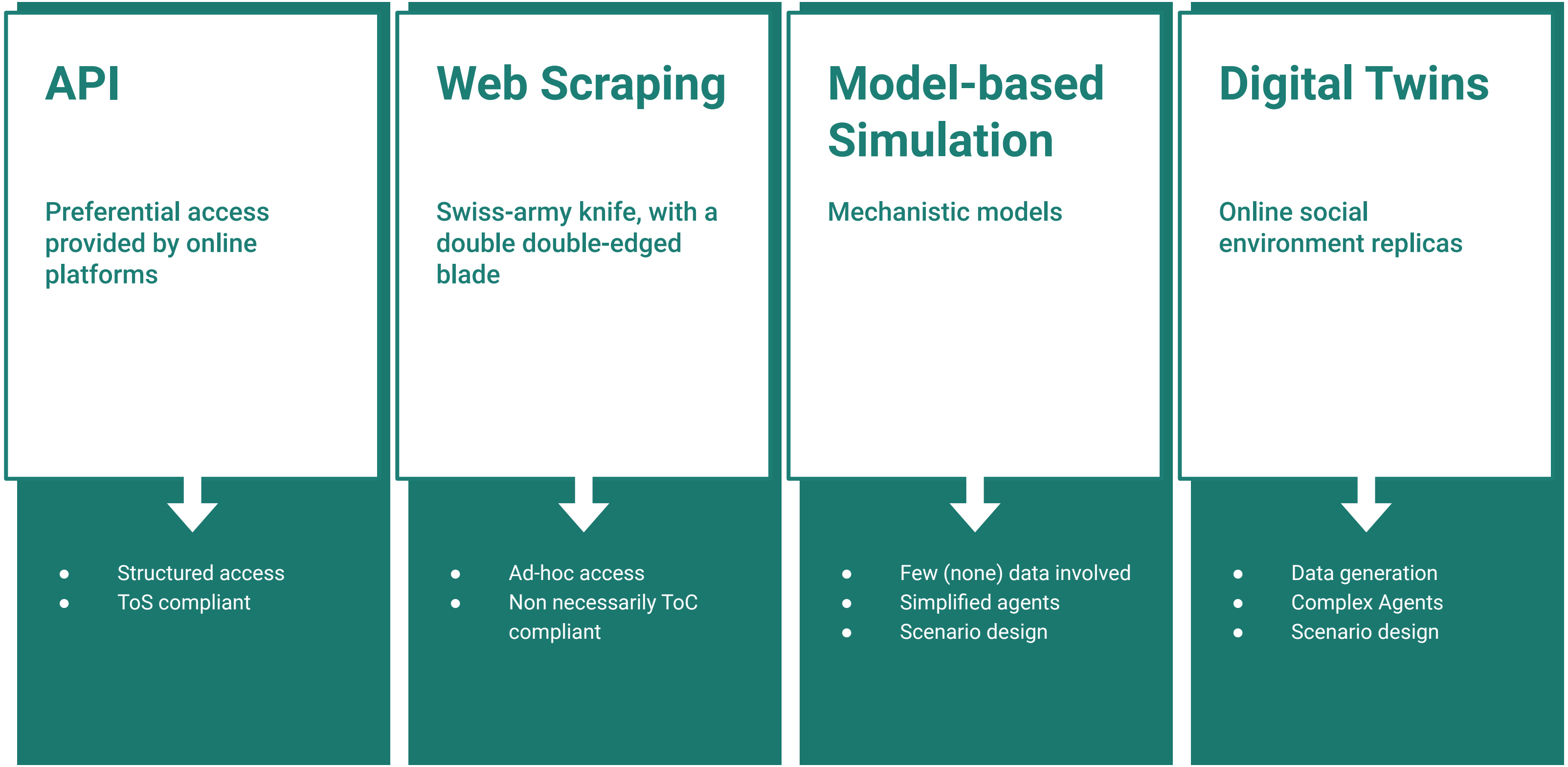
Policy and Decision Support:
Using digital twins to inform policy and organizational decisions.





YSocial

https://ysocialtwin.github.io



API

Preferential access provided by online platforms

- Structured access
- ToS compliant

Web Scraping

Swiss-army knife, with a double double-edged blade

- Ad-hoc access
- Non necessarily ToC compliant

Model-based Simulation

Mechanistic models

- Few (none) data involved
- Simplified agents
- Scenario design

Digital Twins

Online social environment replicas

- Data generation
- Complex Agents
- Scenario design

Chapter 5

Conclusion

Take Away Messages

1. There are multiple ways to collect data
2. Each methodology has its limitations: technical and legal

What's Next

Chapter 6: Graph Transformation