

# Data Collection for Machine Learning

Week 1 · CS 203: Software Tools and Techniques for AI

Prof. Nipun Batra

*IIT Gandhinagar*

# Part 1: The Motivation

Why do we need to collect data?

# Imagine: You Work at Netflix

**NETFLIX** — Your Boss: *"We have \$500M budget for movie acquisitions. Which movies should we license?"*

**The Question:** Can we predict which movies will succeed?

**Your Role:** Data Scientist

**Your Mission:** Build a model to predict movie success

# The Problem Statement

**Goal:** Predict box office revenue based on movie attributes



**But wait...** What features? What data? Where does it come from?

# What We Need: The Target Dataset

Title	Year	Genre	Budget	Revenue	Rating	Director	Cast
Inception	2010	Sci-Fi	\$160M	\$836M	8.8	C. Nolan	DiCaprio
Avatar	2009	Action	\$237M	\$2.9B	7.9	Cameron	Worthington
The Room	2003	Drama	\$6M	\$1.9M	3.9	Wiseau	Wiseau
...	...	...	...	...	...	...	...

**We need 10,000+ movies** with complete information.

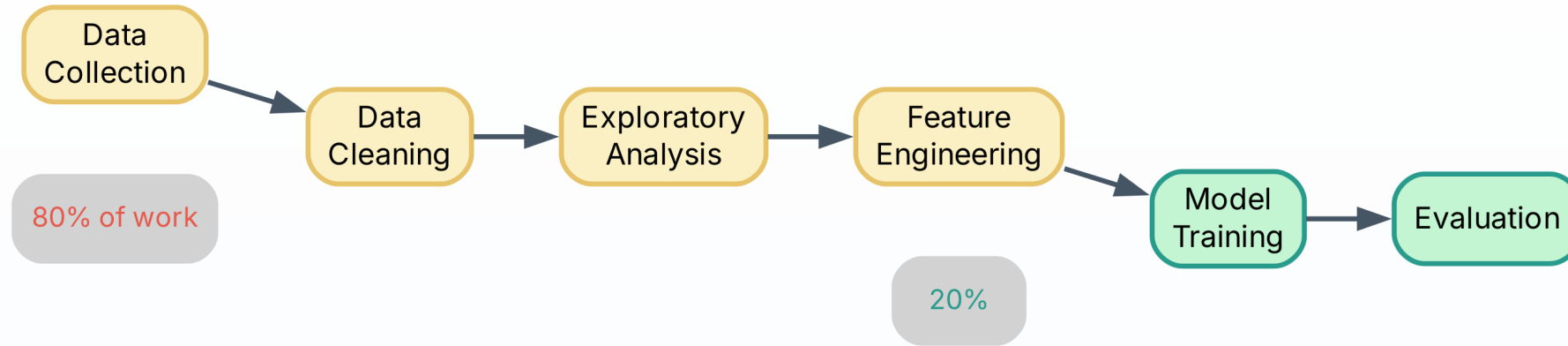
**Question:** Where does this data come from?

# The Reality Check

- This data doesn't exist in one place
- No single CSV file with everything
- Can't just "download" the dataset
- **We must BUILD the dataset ourselves**

**This is the real world of data science.**

# The ML Pipeline Reality



## The uncomfortable truth:

- 80% of ML work is data engineering
- Models are the easy part
- **Garbage In = Garbage Out**

# Why Is Data Collection So Hard?

**The Data Collection Paradox:** The data you need rarely exists in the form you need it.

Challenge	Example
Scattered sources	IMDb, Box Office Mojo, Rotten Tomatoes
Different formats	JSON, HTML, CSV
Missing values	Budget missing for 40% of movies
Inconsistent naming	"The Dark Knight" vs "Dark Knight, The"
Rate limits	Only 100 requests/day



# Today's Mission

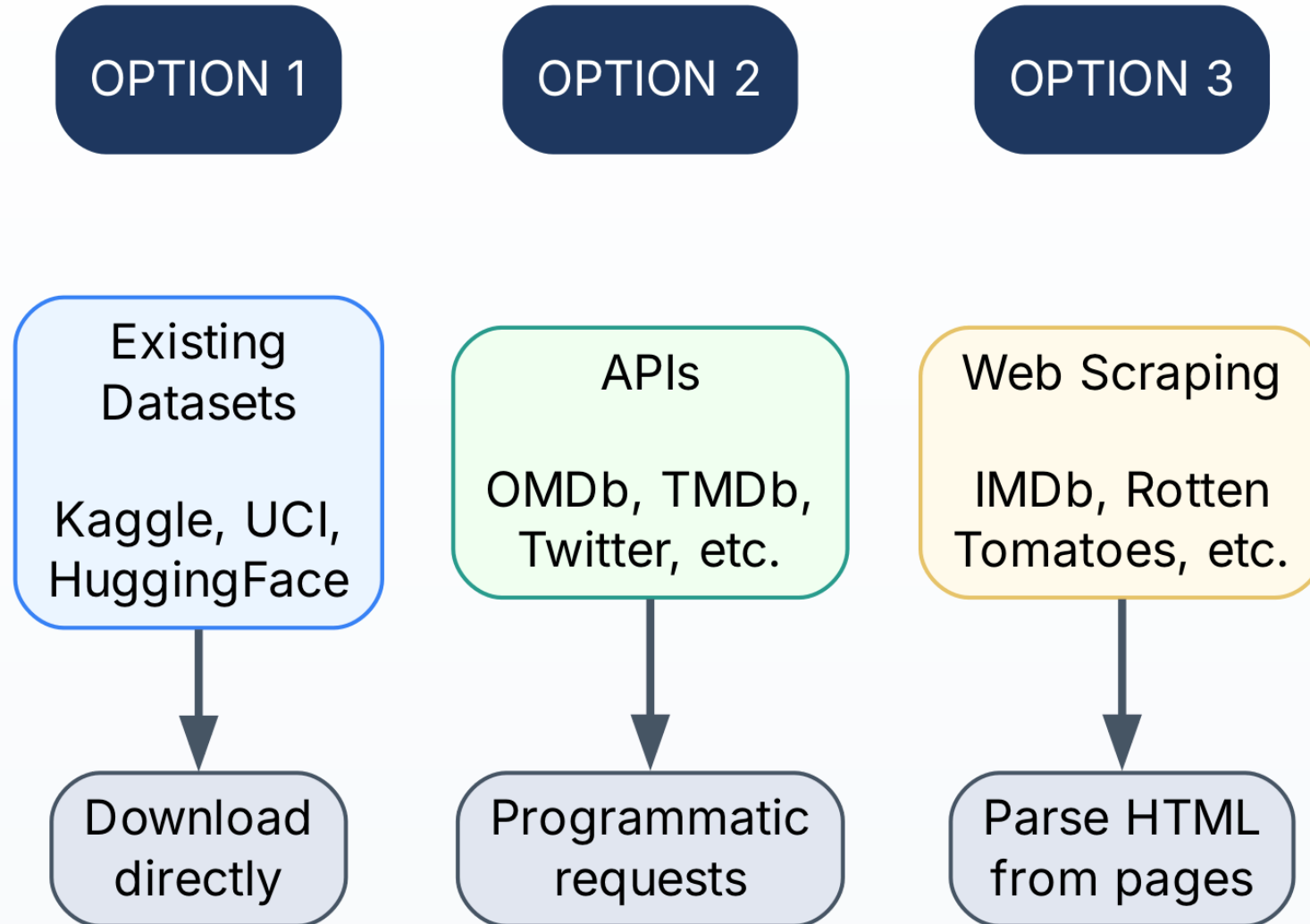
By the end of this lecture, you will know how to:

1. Find data sources for any project
2. Understand how the web works (HTTP)
3. Use Chrome DevTools to inspect network traffic
4. Make requests using curl from the command line
5. Write Python scripts with the requests library
6. Handle different data formats
7. Scrape websites when APIs don't exist

# Part 2: Where Does Data Come From?

*Finding the right sources*

# Three Ways to Get Data



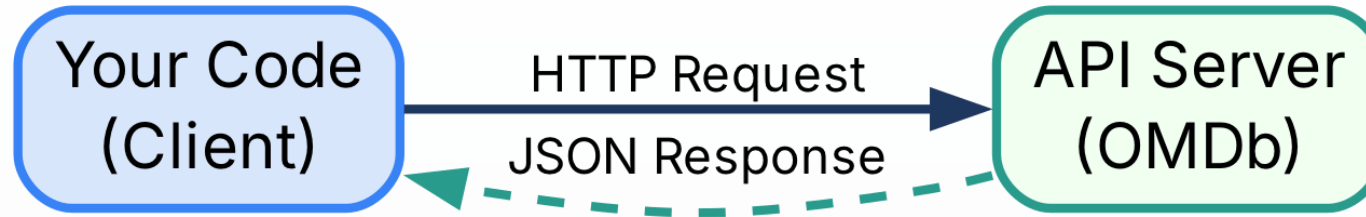
# Option 1: Pre-built Datasets

Where to find them:

Source	Example Datasets	Pros	Cons
Kaggle	Movies, Titanic, Housing	Ready to use, competitions	May be outdated
UCI ML Repository	Classic ML datasets	Well-documented	Academic focus
HuggingFace	NLP datasets, models	Easy loading	Specialized
Government Portals	Census, economic data	Authoritative	Limited scope

**Verdict:** Great starting point, but often not enough for real projects.

## Option 2: APIs (Application Programming Interface)



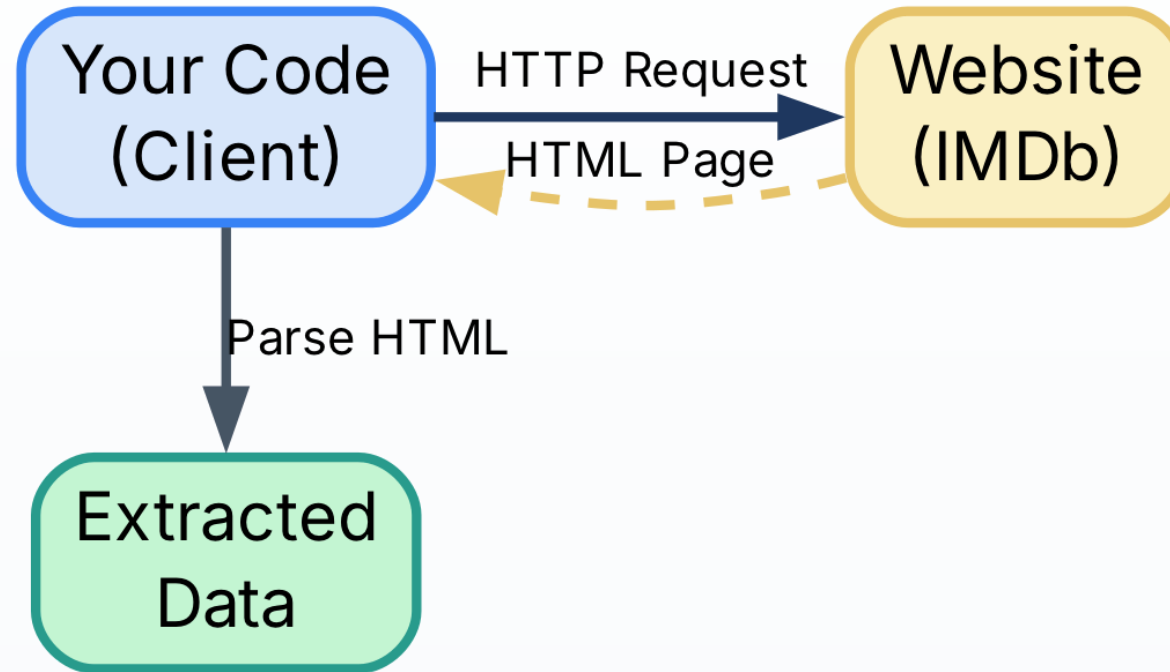
**APIs** = Structured way to request data from servers

**Examples for our Netflix project:**

- **OMDb API**: Movie metadata (title, year, ratings)
- **TMDb API**: Detailed movie info, cast, crew
- **Box Office Mojo**: Revenue data

# Option 3: Web Scraping

When APIs don't exist or don't have what you need:



**When to scrape:** Reviews, prices, content not in APIs.

# Our Strategy for Netflix Project

Data Needed	Source	Method
Movie titles, years	OMDb API	API calls
Ratings, genres	OMDb API	API calls
Budget, revenue	TMDb API	API calls
User reviews	IMDb website	Scraping
Critic reviews	Rotten Tomatoes	Scraping

**Today's focus:** Learn both API calls and scraping.

# Decision Tree: How to Get Data

Ask these questions in order:

1. Does a ready-made dataset exist?
  - YES: Download it (Kaggle, HuggingFace)
  - NO: Continue to step 2...
2. Does an official API exist?
  - YES: Is it free/affordable? → Use the API
  - NO: Continue to step 3...



# Decision Tree (continued)

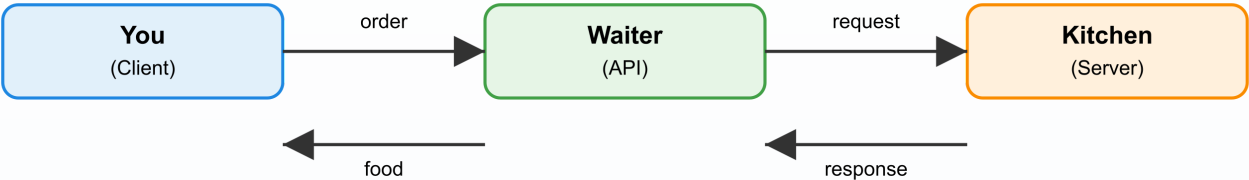
3. Can you scrape the website?
  - Check robots.txt and ToS first
  - YES: Scrape ethically
  - NO: Look for alternatives
4. None of the above?
  - Manual data collection
  - Partner with data owner
  - Reframe the problem

**Most real projects use a combination of all methods!**

# Part 3: What is an API?

*The contract between programs*

# API: A Restaurant Analogy



Restaurant	API
Menu	Documentation
Order	Request
Kitchen	Server
Food	Response

# API: The Formal Definition

## API (Application Programming Interface)

A defined set of rules and protocols for building and interacting with software applications.

```
# Without API (direct database access - dangerous!)  
SELECT * FROM movies WHERE title = 'Inception';  
  
# With API (safe, controlled access)  
GET /movies?title=Inception
```

## APIs provide:

- Security (no direct DB access)
- Rate limiting (fair usage)
- Versioning (backwards compatibility)
- Documentation (how to use it)

# Why Do APIs Exist?

**APIs are like a bank teller window.** You can't walk into the vault, but you can request transactions through a controlled interface.

Without APIs	With APIs
Anyone reads ALL data	Only expose what you want
Anyone can modify/delete	Validate every request
No tracking	Log and monitor usage
Server overwhelmed	Protect against abuse

# Reading API Documentation

Before making any API call, check the docs for:

1. **Base URL** - Where do requests go?
2. **Authentication** - API key? Where does it go?
3. **Endpoints** - What resources are available?
4. **Rate limits** - How many requests per day?

# Example: OMDb API Docs

Base URL: `https://www.omdbapi.com/`

Auth: `apikey` parameter in URL

Parameters: `t` (title), `i` (IMDb ID), `y` (year)

Rate limit: 1,000 requests/day (free tier)

# Types of APIs

Type	Description	Example
<b>REST API</b>	HTTP-based, stateless, resource-oriented	OMDb, GitHub
<b>GraphQL</b>	Query language, get exactly what you need	GitHub v4, Shopify
<b>SOAP</b>	XML-based, enterprise	Legacy banking
<b>WebSocket</b>	Real-time, bidirectional	Chat apps, live data

For data collection, we focus on **REST APIs** (most common).



# REST API: Key Principles

**REST** = **RE**presentational **S**tate **T**ransfer

1. **Stateless**: Server doesn't remember previous requests
2. **Resource-based**: URLs represent things (nouns)
3. **HTTP Methods**: Standard verbs (GET, POST, PUT, DELETE)
4. **Standard formats**: JSON or XML responses

Good URL Design:

GET /movies	→ List all movies
GET /movies/123	→ Get movie with ID 123
POST /movies	→ Create new movie
PUT /movies/123	→ Update movie 123
DELETE /movies/123	→ Delete movie 123

# Anatomy of an API Call

```
https://api.omdbapi.com/?apikey=abc123&t=Inception&y=2010
```

Protocol  
(HTTPS)

Domain  
(server)

Path  
(endpoint)

Query Parameters  
(key=value pairs)

**Query Parameters** (after the `?`):

- `apikey=abc123` → Authentication
- `t=Inception` → Movie title
- `y=2010` → Year (optional filter)

Multiple parameters joined with `&`

# API Authentication

Most APIs require authentication to:

- Track usage
- Enforce rate limits
- Bill customers

## Common methods:

```
# 1. API Key in URL (simplest)
```

```
GET /movies?apikey=YOUR_KEY
```

```
# 2. API Key in Header
```

```
GET /movies
```

```
X-API-Key: YOUR_KEY
```

```
# 3. Bearer Token (OAuth)
```

```
GET /movies
```

```
Authorization: Bearer YOUR_TOKEN
```

# Rate Limiting

**Why?** Servers have limited resources.

Tier	Requests/Day
Free	100
Basic	1,000
Pro	10,000

**If you exceed:** HTTP 429 (Too Many Requests)

**Check headers:** `X-RateLimit-Remaining: 42`

# Dealing with Rate Limits

## Strategy 1: Simple delay

```
for movie in movies:  
    response = requests.get(api_url, params={"t": movie})  
    time.sleep(1) # Wait 1 second between requests
```

# Exponential Backoff

**Strategy 2:** Wait longer after each failure

```
wait_time = 1
while True:
    response = requests.get(url)
    if response.status_code == 429:
        time.sleep(wait_time)
        wait_time *= 2 # Double: 1, 2, 4, 8...
    else:
        break
```

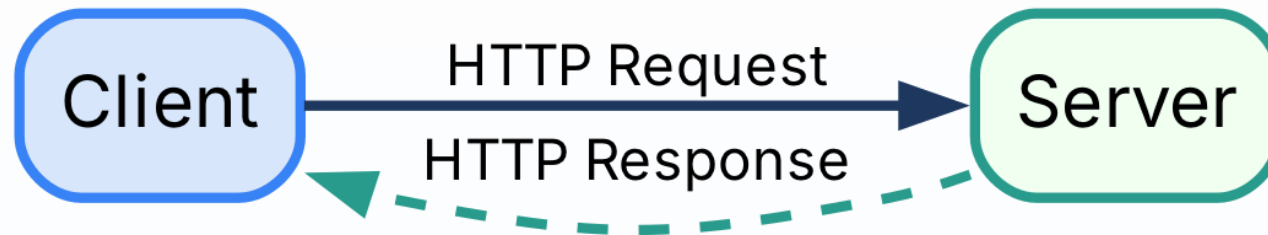
# Part 4: HTTP Fundamentals

*The language of the web*

# What is HTTP?

**HTTP** = **H**yper**T**ext **T**ransfer **P**rotocol

The foundation of data communication on the web.



## Key characteristics:

- **Stateless**: Each request is independent
- **Text-based**: Human-readable (mostly)
- **Port 80** (HTTP) or **Port 443** (HTTPS)



# Understanding "Stateless"

**The Goldfish Analogy:** Server forgets you after every request.

```
Request 1: "I'm Alice. Show me Inception." → "Here's data."
```

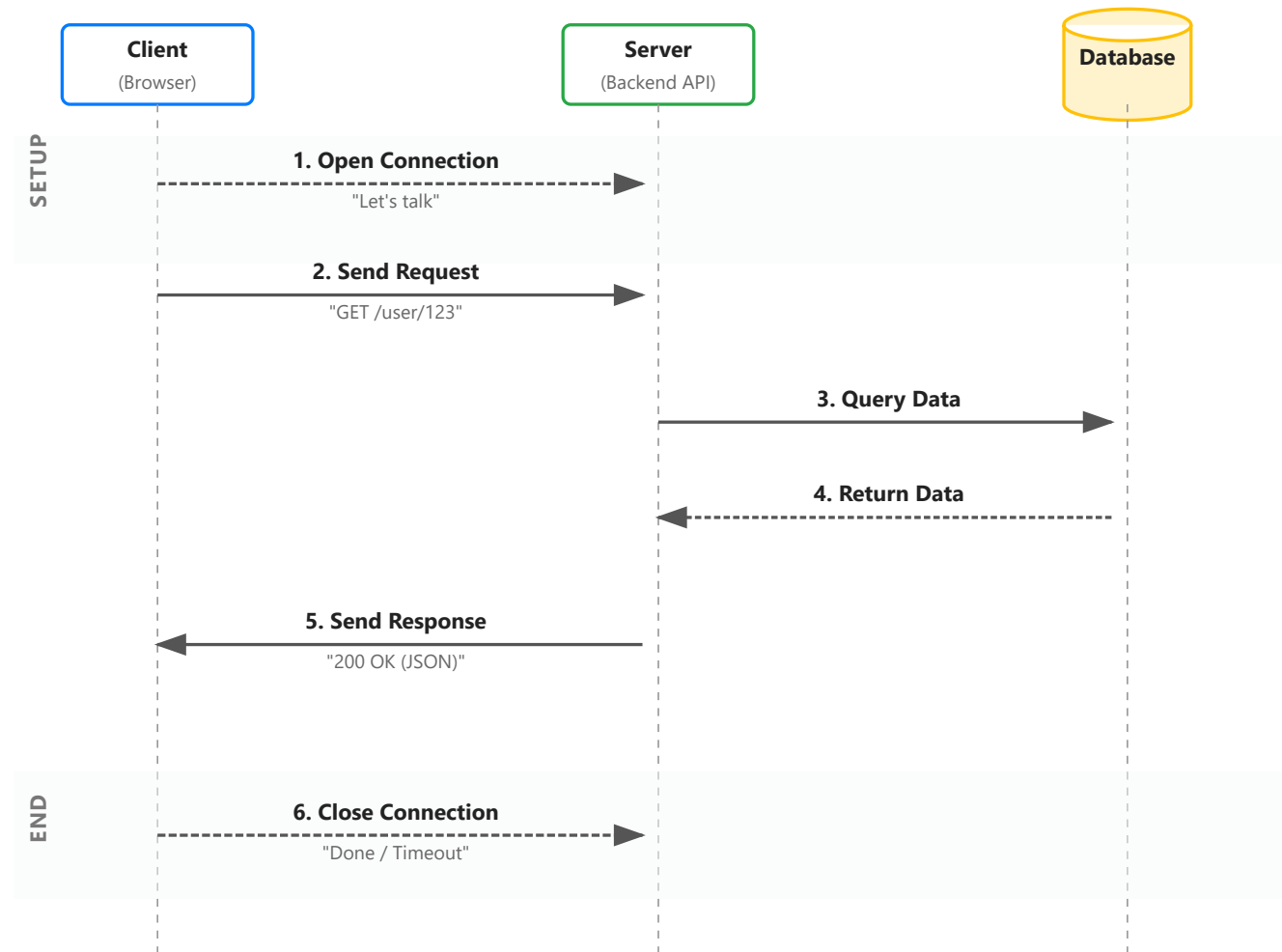
```
Request 2: "Now show me Avatar." → "Who are you?"
```

**Why stateless?** Scalability - any server can handle any request.

**Workaround:** Cookies, tokens, session IDs (sent with every request)

# The Client-Server Model

## The Complete Request Lifecycle



# HTTP Request Structure

Every HTTP request has three parts:

## 1. REQUEST LINE

```
GET /movies?t=Inception HTTP/1.1
```

## 2. HEADERS

```
Host: api.omdbapi.com
```

```
User-Agent: Python/3.9
```

```
Accept: application/json
```

```
Authorization: Bearer abc123
```

## 3. BODY (optional, for POST/PUT)

```
{"title": "New Movie", "year": 2024}
```

# HTTP Response Structure

Every HTTP response has three parts:

## 1. STATUS LINE

HTTP/1.1 200 OK

## 2. HEADERS

Content-Type: application/json

Content-Length: 1234

X-RateLimit-Remaining: 99

## 3. BODY (the actual data)

```
{"Title": "Inception", "Year": "2010", ...}
```

# URL Anatomy

`https://api.omdbapi.com:443/v1/movies?t=Inception&y=2010`



Component	Description	Example
Protocol	<code>http://</code> or <code>https://</code>	<code>https://</code>
Host	Domain or IP	<code>api.omdbapi.com</code>
Port	Usually implicit	443 (HTTPS)
Path	Resource location	<code>/v1/movies</code>
Query	<code>key=value</code> pairs	<code>?t=Inception</code>

# Common HTTP Headers

**Request Headers** (what client sends):

Header	Purpose	Example
Host	Target server	api.omdbapi.com
User-Agent	Client identification	Mozilla/5.0
Accept	Preferred response format	application/json
Authorization	Authentication	Bearer token123
Content-Type	Body format (POST)	application/json

# Common HTTP Headers (Response)

**Response Headers** (what server sends):

Header	Purpose	Example
Content-Type	Body format	application/json
Content-Length	Size in bytes	1234
Cache-Control	Caching rules	max-age=3600
X-RateLimit-Remaining	API quota left	99
Set-Cookie	Session cookie	session=abc123

# Part 5: HTTP Methods - GET and POST

*The two most important verbs*



# HTTP Methods Overview

Method	Purpose	Has Body?	Safe?	Idempotent?
<b>GET</b>	Retrieve data	No	Yes	Yes
<b>POST</b>	Create/submit data	Yes	No	No
PUT	Replace resource	Yes	No	Yes
PATCH	Partial update	Yes	No	No
DELETE	Remove resource	No	No	Yes

**For data collection:** 90% GET, 10% POST

# What Do "Safe" and "Idempotent" Mean?

**Safe** = "Looking doesn't change anything" (like window shopping)

**Idempotent** = "Doing it twice has the same effect as once"

Action	Safe?	Idempotent?
Reading a book	Yes	Yes
Ordering pizza	No	No
Setting thermostat to 72°	No	Yes

## Why this matters:

- GET can be cached and retried safely
- POST should not be auto-retried (double-charge risk!)

# GET Request: Retrieving Data

**Purpose:** Fetch data without modifying anything.

```
GET /movies?t=Inception&y=2010 HTTP/1.1
Host: api.omdbapi.com
Accept: application/json
```

## Characteristics:

- Parameters in URL (query string)
- No request body
- **Safe:** Doesn't change server state
- **Idempotent:** Same request = same result
- **Cacheable:** Responses can be cached

# POST Request: Sending Data

**Purpose:** Submit data to create or process something.

```
POST /api/feedback HTTP/1.1
Host: example.com
Content-Type: application/json

{"movie_id": 123, "rating": 5, "review": "Great!"}
```

## Characteristics:

- Data in request body (not URL)
- **Not safe:** Modifies server state
- **Not idempotent:** Multiple POSTs create multiple resources
- **Not cacheable**

# GET vs POST: When to Use Which

Scenario	Method	Why
Fetching movie details	GET	Retrieving data
Searching for movies	GET	Query in URL
Submitting a review	POST	Creating new data
Uploading an image	POST	Sending binary data
User login	POST	Sensitive data in body
Listing all movies	GET	No modification

**Data Collection = Mostly GET**

**Data Submission = POST**

# HTTP Status Codes

Status codes are grouped by category:

Range	Category	Meaning
1xx	Informational	Request received, processing
2xx	Success	Request succeeded
3xx	Redirection	Further action needed
4xx	Client Error	Your fault
5xx	Server Error	Their fault

# Common Status Codes

Code	Meaning	When
200 OK	Success	Request succeeded
201 Created	Created	POST created resource
400 Bad Request	Client error	Malformed request
401 Unauthorized	Auth needed	Missing credentials
403 Forbidden	Denied	Not allowed
404 Not Found	Missing	Resource doesn't exist
429 Too Many Requests	Rate limit	Slow down!
500 Internal Error	Server crash	Their fault

# Status Code Intuition

**First digit = who's to blame:** 2xx = OK, 4xx = your fault, 5xx = their fault

```
if response.status_code == 200:
    data = response.json()      # Success!
elif response.status_code == 404:
    print("Not found")         # Bad ID
elif response.status_code == 429:
    time.sleep(60)             # Rate limited
elif response.status_code >= 500:
    time.sleep(5)              # Server error
```



# Part 6: Response Formats

*Same data, different representations*

# Why Different Formats?

Same movie data can be represented in different formats:

Format	Full Name	Use Case
JSON	JavaScript Object Notation	APIs, Web apps
XML	eXtensible Markup Language	Enterprise, Legacy
CSV	Comma Separated Values	Spreadsheets, ML
HTML	HyperText Markup Language	Web pages
Protobuf	Protocol Buffers	High-performance

**Content-Type header** tells you the format:

- `application/json` → JSON
- `application/xml` → XML
- `text/html` → HTML
- `text/csv` → CSV

# Format 1: JSON

The most common API format today.

```
{
  "title": "Inception",
  "year": 2010,
  "genres": ["Sci-Fi", "Action", "Thriller"],
  "director": {
    "name": "Christopher Nolan",
    "nationality": "British"
  },
  "rating": 8.8,
  "in_production": false
}
```

**Pros:** Human-readable, lightweight, native to JavaScript

**Cons:** No schema validation, no comments

# JSON Data Types

```
{  
  "string": "Hello World",  
  "number": 42,  
  "decimal": 3.14159,  
  "boolean": true,  
  "null_value": null,  
  "array": [1, 2, 3],  
  "object": {  
    "nested": "value"  
  }  
}
```

**Only 7 data types:** string, number, boolean, null, array, object

**Note:** No native date type! Dates are typically strings: "2010-07-16"

# JSON Gotchas

```
# Numbers might be strings!
data = {"year": "2010"}      # String, not int!
year = int(data["year"])     # Must convert

# Missing keys crash your code
data["director"]             # KeyError!
data.get("director", "Unknown") # Safe!
```

# More JSON Gotchas

```
# null becomes None in Python
data = {"budget": None}
if data["budget"]:           # This is False!
    print("Has budget")

# Empty string vs null vs missing
{"rating": ""}              # Empty string
{"rating": None}            # Null
{}                          # Missing key
```

# Format 2: XML

The enterprise standard (still used in SOAP APIs).

```
<?xml version="1.0" encoding="UTF-8"?>
<movie>
  <title>Inception</title>
  <year>2010</year>
  <genres>
    <genre>Sci-Fi</genre>
    <genre>Action</genre>
  </genres>
  <director nationality="British">
    Christopher Nolan
  </director>
  <rating>8.8</rating>
</movie>
```

**Pros:** Schema validation (XSD), attributes, widespread support

**Cons:** Verbose, heavier than JSON

# JSON vs XML: Same Data

Aspect	JSON	XML
Syntax	<code>{"name": "Inception"}</code>	<code>&lt;name&gt;Inception&lt;/name&gt;</code>
Structure	Curly braces <code>{}</code>	Tags <code>&lt;tag&gt;&lt;/tag&gt;</code>
Size	Lighter (~30% smaller)	More verbose
Attributes	Not supported	Supported
Arrays	<code>[1, 2, 3]</code>	Repeated elements
Usage	Modern APIs	Legacy/Enterprise



# Format 3: CSV

The data scientist's friend.

```
title,year,genre,director,rating
Inception,2010,Sci-Fi,Christopher Nolan,8.8
Avatar,2009,Action,James Cameron,7.9
```

## Pros:

- Opens in Excel/Google Sheets
- Easy to load into pandas: `pd.read_csv("movies.csv")`
- Very compact

## Cons:

- Flat structure only (no nesting)
- No data types (everything is text)
- Escaping issues with commas in data

# Format 4: HTML

What you get when scraping websites.

```
<div class="movie-card">
  <h2 class="title">Inception</h2>
  <span class="year">2010</span>
  <ul class="genres">
    <li>Sci-Fi</li>
    <li>Action</li>
  </ul>
  <p class="rating">Rating: 8.8/10</p>
</div>
```

Not designed for data exchange!

- Mixed with presentation (CSS, layout)
- Need to parse and extract relevant data
- Structure varies by website

# Format 5: Protocol Buffers (Protobuf)

Google's high-performance binary format.

```
// movie.proto (schema definition)
message Movie {
  string title = 1;
  int32 year = 2;
  repeated string genres = 3;
  float rating = 4;
}
```

Binary on the wire (not human-readable):

```
0a 09 49 6e 63 65 70 74 69 6f 6e 10 da 0f ...
```

**Pros:** 10x smaller, 100x faster parsing

**Cons:** Need schema, binary format, requires tooling

# Format Comparison: Same Movie

Format	Size	Readability	Use Case
JSON	150 bytes	High	REST APIs
XML	200 bytes	Medium	Enterprise
CSV	50 bytes	High	Data exchange
HTML	300 bytes	Low	Web pages
Protobuf	30 bytes	None	High-perf APIs

**For this course:** Focus on JSON and HTML

# Part 7: Chrome DevTools

*Your window into HTTP traffic*

# Why Chrome DevTools?

DevTools lets you see:

- Every HTTP request your browser makes
- Request headers, body, timing
- Response headers, body, status codes
- Copy requests as curl commands!

This is how you learn what APIs a website uses.

# Opening DevTools

## Three ways to open:

1. **Keyboard:** `F12` or `Ctrl+Shift+I` (Windows/Linux) / `Cmd+Option+I` (Mac)
2. **Right-click:** Right-click on page → "Inspect"
3. **Menu:** Chrome menu → More Tools → Developer Tools

## Navigate to the "Network" tab

# The Network Tab

Network				
[*] Preserve log    [ ] Disable cache    [Filter]				
Name	Status	Type	Size	Time
api/movies	200	fetch	1.2 KB	45 ms
styles.css	200	css	5.4 KB	23 ms
logo.png	200	image	15 KB	67 ms
analytics.js	200	script	8.1 KB	89 ms

Every row = one HTTP request/response



# Filtering Requests

Filter by type:

Filter	Shows
All	Everything
Fetch/XHR	API calls (AJAX) ← Most useful!
Doc	HTML documents
CSS	Stylesheets
JS	JavaScript files
Img	Images

Click "Fetch/XHR" to see only API calls

# Inspecting a Request

Click on any request to see details:

Headers	Preview	Response	Timing	Cookies
General:				
Request URL: <code>https://www.omdbapi.com/?i=tt3896198&amp;apikey=[KEY]</code>				
Request Method: GET				
Status Code: 200 OK				
Response Headers:				
content-type: application/json				
x-ratelimit-remaining: 99				
Request Headers:				
authorization: Bearer eyJhbGc...				
user-agent: Mozilla/5.0...				

# The Preview & Response Tabs

**Preview Tab:** Formatted JSON viewer

```
{  
  "title": "Inception",  
  "year": 2010,  
  "rating": 8.8  
}
```

**Response Tab:** Raw response body

```
{"title":"Inception","year":2010,"rating":8.8}
```

# Copy as curl

## The most powerful feature!

1. Right-click on any request
2. Select "Copy" → "Copy as cURL"
3. Paste into terminal

```
curl "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]" \  
-H "accept: application/json" \  
--compressed
```

Now you can replay the exact request from your terminal!

# Demo: Finding Hidden APIs

Many websites use hidden APIs. Here's how to find them:

1. Open DevTools → Network tab
2. Filter by "Fetch/XHR"
3. Interact with the website (search, click, load more)
4. Watch for API calls appearing in the list
5. Click on interesting requests to inspect them
6. Copy as curl to test in terminal

**Example:** Search on IMDb and watch for API calls...

# DevTools Pro Tips

**Preserve log:** Keep requests when navigating between pages

**Disable cache:** See fresh requests every time

**Search:** `Ctrl+F` to search in all requests

**Filter by URL:** Type in filter box to match URLs

**Clear:** Click the clear icon to clear all requests

**Throttling:** Simulate slow networks (3G, offline)

# Part 8: Making Requests with curl

*The command-line HTTP client*

# What is curl?

**curl** = "Client URL" - a command-line tool for transferring data.

```
# Your first curl command  
curl "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]"
```

## Why learn curl?

- Universal (works everywhere)
- Quick debugging
- Foundation for understanding HTTP
- Copy from DevTools, paste and run



# curl: Basic Syntax

```
curl [options] [URL]
```

## Common options:

Option	Meaning	Example
-X	HTTP method	-X POST
-H	Add header	-H "Accept: application/json"
-d	Send data (body)	-d '{"key": "value"}'
-o	Output to file	-o movie.json
-I	Headers only	-I
-v	Verbose output	-v
-s	Silent mode	-s

# curl: GET Request

```
# Simple GET request  
curl "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]"
```

## Output:

```
{"Title":"Inception","Year":"2010","Rated":"PG-13",  
"Released":"16 Jul 2010","Runtime":"148 min",...}
```

**Important:** Quote the URL! (prevents shell interpretation of `&` )

# curl: Adding Headers

```
curl "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]" \  
  -H "Accept: application/json" \  
  -H "Authorization: Bearer YOUR_TOKEN" \  
  -H "User-Agent: MyApp/1.0"
```

## Common headers to add:

- `Accept: application/json` - Request JSON response
- `Authorization: Bearer TOKEN` - Authentication
- `Content-Type: application/json` - When sending JSON

# curl: Viewing Response Headers

```
# Show only response headers (no body)
curl -I "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]"
```

## Output:

```
HTTP/1.1 200 OK
Content-Type: application/json; charset=utf-8
Content-Length: 1024
Cache-Control: public, max-age=86400
X-RateLimit-Remaining: 999
```

# curl: Verbose Mode

```
curl -v "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]"
```

Shows everything (request AND response):

```
> GET /?apikey=demo&t=Inception HTTP/2
> Host: api.omdbapi.com
> User-Agent: curl/7.79.1
> Accept: */*
>
< content-length: 1024
<
{"Title":"Inception"...}
```

> = What you sent (request)

< = What you received (response)

# Pretty Printing with jq

Raw JSON is hard to read. Pipe to `jq` for formatting:

```
curl -s "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]" | jq .
```

Output (formatted):

```
{
  "Title": "Inception",
  "Year": "2010",
  "Rated": "PG-13",
  "Runtime": "148 min",
  "Genre": "Action, Adventure, Sci-Fi"
}
```

# jq: Extracting Specific Fields

```
# Get just the title
```

```
curl -s ... | jq '.Title'
```

```
# Output: "Inception"
```

```
# Get multiple fields as new object
```

```
curl -s ... | jq '{title: .Title, year: .Year, rating: .imdbRating}'
```

```
# Output: {"title": "Inception", "year": "2010", "rating": "8.8"}
```

```
# Get first element of array
```

```
curl -s ... | jq '.Search[0]'
```

```
# Get all titles from array
```

```
curl -s ... | jq '.Search[].Title'
```

# curl: Saving to File

```
# Save response to file
```

```
curl "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]" \  
-o inception.json
```

```
# Silent mode (no progress bar)
```

```
curl -s "https://www.omdbapi.com/?t=Inception&apikey=[API_KEY]" -o output.json
```

```
# Save with pretty formatting
```

```
curl -s ... | jq . > formatted.json
```



# curl: POST Request

```
curl -X 'POST' \  
  'https://nipun-api-testing.hf.space/items' \  
  -H 'accept: application/json' \  
  -H 'Content-Type: application/json' \  
  -d '{  
    "name": "Laptop",  
    "price": 999.99,  
    "quantity": 1,  
    "description": "A powerful laptop"  
  }'
```

## Components:

- `-X POST` - Use POST method
- `-H "Content-Type: application/json"` - Tell server we're sending JSON
- `-d '...'` - The data (request body)

# curl: POST with Form Data

```
curl -X POST "https://nipun-api-testing.hf.space/form/contact" \  
  -H "Content-Type: application/x-www-form-urlencoded" \  
  -d "name=Alice" \  
  -d "email=alice@example.com" \  
  -d "subject=Hello" \  
  -d "message=Nice API!"
```

# curl: File Upload

```
# Upload a file  
curl -X POST "https://nipun-api-testing.hf.space/upload/file" -F "file=@dummy.txt"
```

**-F** = multipart form data (for file uploads)

**@** = read from file

# curl: Useful Options

# Retry on failure

```
curl --retry 3 "https://www.omdbapi.com/data"
```

# Set timeout (seconds)

```
curl --max-time 10 "https://www.omdbapi.com/slow"
```

# Follow redirects

```
curl -L "https://short.url/abc"
```

# Fail silently on HTTP errors

```
curl -f "https://www.omdbapi.com/notfound"
```

# (exits with error code instead of showing error page)

# Part 9: Python requests Library

*Programmatic data collection*

# Why Python requests?

curl is great for testing, but for automation you need Python.

```
# Install  
pip install requests
```

## Benefits over curl:

- Loop over many URLs
- Parse JSON automatically
- Handle errors gracefully
- Store data in variables
- Integrate with pandas, ML pipelines

# requests: Simple GET

```
import requests

# Make a GET request to OMDb API
response = requests.get(
    "https://www.omdbapi.com/",
    params={
        "apikey": "demo",    # replace with your real API key
        "t": "Inception"
    }
)

# Check HTTP status code
print(response.status_code) # 200 means OK

# Parse JSON response
data = response.json()

# Access fields from JSON
print(data["Title"]) # Inception
print(data["Year"])  # 2010
```

# requests: Using params

Don't manually build query strings!

```
# Bad (manual string building)
url = "https://www.omdbapi.com/?apikey=demo&t=Inception&y=2010"

# Good (use params dict)
response = requests.get(
    "https://www.omdbapi.com/",
    params={
        "apikey": "demo",
        "t": "Inception",
        "y": 2010
    }
)
```

Python handles URL encoding automatically!



# requests: Adding Headers

```
import requests

response = requests.get(
    "https://httpbin.org/headers",
    headers={
        "Authorization": "Bearer test-token-123",
        "Accept": "application/json",
        "User-Agent": "MyApp/1.0"
    }
)

print(response.status_code)
print(response.json())
```

# requests: Response Object

```
import requests
response = requests.get(
    "https://www.omdbapi.com/",
    params={
        "apikey": "demo",      # replace with real key for reliability
        "t": "Inception"
    }
)
print(response.status_code)    # 200

print(response.headers["Content-Type"])

print(response.text) # Body as raw text (string)

# Body parsed as JSON (Python dict)
print(response.json())
print(response.ok) # True for 2xx status codes
```

# requests: POST with JSON

```
import requests

response = requests.post(
    "https://nipun-api-testing.hf.space/items",
    json={
        "name": "Laptop",
        "price": 999.99,
        "quantity": 1,
        "description": "A powerful laptop"
    }
)

print(response.status_code)
print(response.json())
```

# requests: POST with Form Data

```
import requests

url = "https://nipun-api-testing.hf.space/form/contact"

data = {
    "name": "Alice",
    "email": "alice@example.com",
    "subject": "Hello",
    "message": "Nice API!"
}

response = requests.post(url, data=data)

print(response.status_code)
print(response.json())
```

## Remember:

- `json=` → Content-Type: application/json
- `data=` → Content-Type: application/x-www-form-urlencoded

# requests: Error Handling

```
url = "https://www.omdbapi.com/"
params = {
    "apikey": "demo",    # replace with real key if needed
    "t": "Inception"
}
try:
    response = requests.get(url, params=params, timeout=10)
    response.raise_for_status() # Raise exception for 4xx / 5xx responses

    data = response.json()
    print(data["Title"], data["Year"])

except requests.exceptions.Timeout:
    print("Request timed out")
```

## Key points:

- Always set `timeout` to avoid hanging
- `raise_for_status()` converts bad status codes to exceptions

# requests: Looping Over Multiple Items

```
movies = ["Inception", "Avatar", "The Matrix", "Interstellar"]
results = []

url = "https://www.omdbapi.com/"
params_base = {
    "apikey": "demo"    # replace with your real key for reliability
}

for title in movies:
    response = requests.get(
        url,
        params=**params_base, "t": title,
        timeout=10
    )

    if response.ok:
        data = response.json()
        if data.get("Response") == "True": # Application-level success check
            results.append(data)
            print(f"Got: {title}")
        else:
            print(f"Not found: {title} ( {data.get('Error')} )")
    else:
        print(f"HTTP error for {title} : {response.status_code}")
print(f"Collected {len(results)} movies")
```

# requests: Session for Multiple Requests

```
import requests

# Create a session (persists headers, cookies, connections)
session = requests.Session()

# Set default headers once
session.headers.update({
    "Authorization": "Bearer test-token-123",
    "User-Agent": "MyApp/1.0"
})

# All requests now reuse headers + connection
r1 = session.get("https://httpbin.org/headers")
r2 = session.get("https://httpbin.org/headers")
r3 = session.get("https://httpbin.org/headers")

print(r1.status_code)
print(r1.json()["headers"]["User-Agent"])
print(r1.json()["headers"]["Authorization"])
```

# requests: Practical Example

```
def fetch_movie_data(titles, api_key):
    movies = []

    for title in titles:
        response = requests.get(
            "https://www.omdbapi.com/",
            params={"apikey": api_key, "t": title}, timeout=10
        )

        if response.ok: # Check HTTP + API-level success
            data = response.json()
            if data.get("Response") == "True":
                movies.append(data)
            else:
                print(f"Movie not found: {title}")
        else:
            print(f"HTTP error for {title}: {response.status_code}")

    return pd.DataFrame(movies) # Convert to pandas DataFrame

df = fetch_movie_data(["Inception", "Avatar"], "demo") # demo API key
print(df[["Title", "Year", "Genre"]])
```



# Data Collection Best Practices

1. **Save raw responses** - Save the full JSON, not just extracted fields
2. **Log everything** - Track successes, failures, and why
3. **Use checkpoints** - Resume after crashes
4. **Handle edge cases** - Missing budgets, directors, etc.
5. **Validate as you go** - Check data types early

**Why?** Don't re-collect 10,000 movies because you missed a field!

# curl vs requests: Comparison

Aspect	curl	Python requests
Use case	Quick testing	Automation
Learning	Interactive exploration	Production code
Looping	Bash scripts	Native Python
JSON parsing	Needs jq	Built-in .json()
Error handling	Exit codes	Exceptions
DevTools	Copy as curl (yes)	Convert from curl

**Workflow:** DevTools → Copy as curl → Test → Convert to Python

# Part 10: Web Scraping

*When APIs don't exist*

# When to Scrape?

## DO scrape when:

- No API available
- API doesn't have the data you need
- API is too expensive
- Public information on public websites

## DON'T scrape when:

- robots.txt disallows it
- Terms of Service prohibit it
- Data is behind login (personal data)
- It would harm the website

# API vs Scraping Comparison

Aspect	API	Scraping
Reliability	Stable	Fragile (HTML changes)
Speed	Fast	Slower
Data Format	Structured JSON	Unstructured HTML
Rate Limits	Documented	Unknown
Legality	Clear TOS	Gray area
Maintenance	Low	High

**Rule:** Always prefer APIs when available.

# HTML Structure Basics

HTML = Nested elements forming a tree (DOM)

```
<!DOCTYPE html>
<html>
  <head>
    <title>Movie Database</title>
  </head>
  <body>
    <div class="movie" id="movie-123">
      <h2 class="title">Inception</h2>
      <span class="year">2010</span>
      <p class="plot">A thief who steals...</p>
    </div>
  </body>
</html>
```

# The DOM Tree

```
      html
     /  \
    head body
    |    |
    title div.movie
         /  |  \
        h2.title span.year p.plot
         |        |        |
        "Inception" "2010" "A thief..."
```

**DOM** = Document Object Model

**Scraping** = Navigating this tree to extract data

# CSS Selectors: Finding Elements

Selector	Meaning	Example Match
<code>div</code>	Element type	<code>&lt;div&gt;...&lt;/div&gt;</code>
<code>.movie</code>	Class name	<code>&lt;div class="movie"&gt;</code>
<code>#main</code>	Element ID	<code>&lt;div id="main"&gt;</code>
<code>div.movie</code>	Tag with class	<code>&lt;div class="movie"&gt;</code>
<code>.movie .title</code>	Nested element	<code>.title</code> inside <code>.movie</code>
<code>a[href="/movies"]</code>	Attribute value	<code>&lt;a href="/movies"&gt;</code>



# BeautifulSoup: Setup

```
pip install beautifulsoup4 requests
```

```
# Fetch the hosted sample movie page
url = "https://nipunbatra.github.io/stt-ai-teaching/html/sample-movie-website.html"
response = requests.get(url)
html = response.text

# Parse it
soup = BeautifulSoup(html, 'html.parser')

# Now we can search and extract elements
print(soup.title.string) # "My Movie Library"
```

# BeautifulSoup: Finding Elements

```
html = """
<div class="movie">
    <h2 class="title">Inception</h2>
    <span class="year">2010</span>
    <span class="rating">8.8</span>
</div>
"""

soup = BeautifulSoup(html, 'html.parser')

# Find single element
title = soup.find('h2', class_='title')
print(title.text) # "Inception"

all_movies = soup.find_all('div', class_='movie') # Find all elements (if multiple movies)
print(f"Found {len(all_movies)} movie(s)")
```

# BeautifulSoup: CSS Selectors

```
soup = BeautifulSoup(html, 'html.parser')

# Select first match
title = soup.select_one('.movie .title')
print(title.text) # "Inception"

# Select all matches
all_titles = soup.select('.movie .title')
for t in all_titles:
    print(t.text)

# Example: all links starting with "/movies/"
links = soup.select('a[href^="/movies/"]')
for link in links:
    print(link.get('href'))
```

# BeautifulSoup: Extracting Data

```
# Get text content
element = soup.select_one('.title')
print(element.text)           # "Inception"
print(element.get_text())     # "Inception"
print(element.get_text(strip=True)) # Remove extra whitespace

# Get attributes
link = soup.select_one('a')
print(link.get('href'))       # "/movies/123"
print(link['href'])           # "/movies/123"
print(link.attrs)             # {'href': '/movies/123', 'class': ['btn']}
```

# Scraping Example: Movie List

```
# Scraping Example: Movie List

url = "https://nipunbatra.github.io/stt-ai-teaching/html/sample-movie-website.html"
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')
movies = []

for card in soup.select('.movie-card'):
    movie = {
        'title': card.select_one('.title').text.strip(),
        'year': card.select_one('.year').text.strip(),
        'genre': card.select_one('.genre').text.strip(),
        'rating': card.select_one('.rating').text.strip(),
        'plot': card.select_one('.plot').text.strip()
    }
    movies.append(movie)

for m in movies:
    print(f"{m['title']} ({m['year']}) - {m['genre']} - Rating: {m['rating']}") # Print results
```

# Handling Pagination

```
base_url = "https://example.com/movies?page="
all_movies = []

for page in range(1, 11): # Pages 1-10
    response = requests.get(f"{base_url}{page}")
    soup = BeautifulSoup(response.text, 'html.parser')

    movies = soup.select('.movie-card')
    if not movies:
        break # No more pages

    for m in movies:
        all_movies.append(m.select_one('.title').text)

    time.sleep(1) # Be polite!
```

# Scraping Ethics & Best Practices

```
headers = {'User-Agent': 'MyBot/1.0 (contact@example.com)'}

for url in urls:
    response = requests.get(url, headers=headers)
    time.sleep(1) # Wait between requests
```

## Rules:

1. Check `robots.txt` first
2. Add delays between requests
3. Identify yourself (User-Agent)
4. Cache responses when possible
5. Respect rate limits

# Common Scraping Mistakes

Mistake	Solution
No delays	Add <code>time.sleep(1)</code>
Hardcoded selectors	Handle missing elements
No error handling	Wrap in try/except
Ignoring encoding	Check <code>response.encoding</code>
Not saving raw HTML	Save before parsing



# Defensive Scraping Pattern

```
try:
    title = card.select_one('.title')
    movie['title'] = title.text.strip() if title else "Unknown"
except Exception as e:
    logging.error(f"Failed to parse: {url}, error: {e}")
```

**Always handle missing elements gracefully!**

# Checking robots.txt

```
curl https://www.imdb.com/robots.txt
```

```
User-agent: *  
Disallow: /search/  
Disallow: /ap/  
Allow: /title/  
Crawl-delay: 1
```

## Meaning:

- Cannot scrape `/search/` pages
- Can scrape `/title/` pages (movie pages)
- Wait 1 second between requests

# Part 11: Putting It All Together

*Back to our Netflix mission*

# Remember Our Goal?

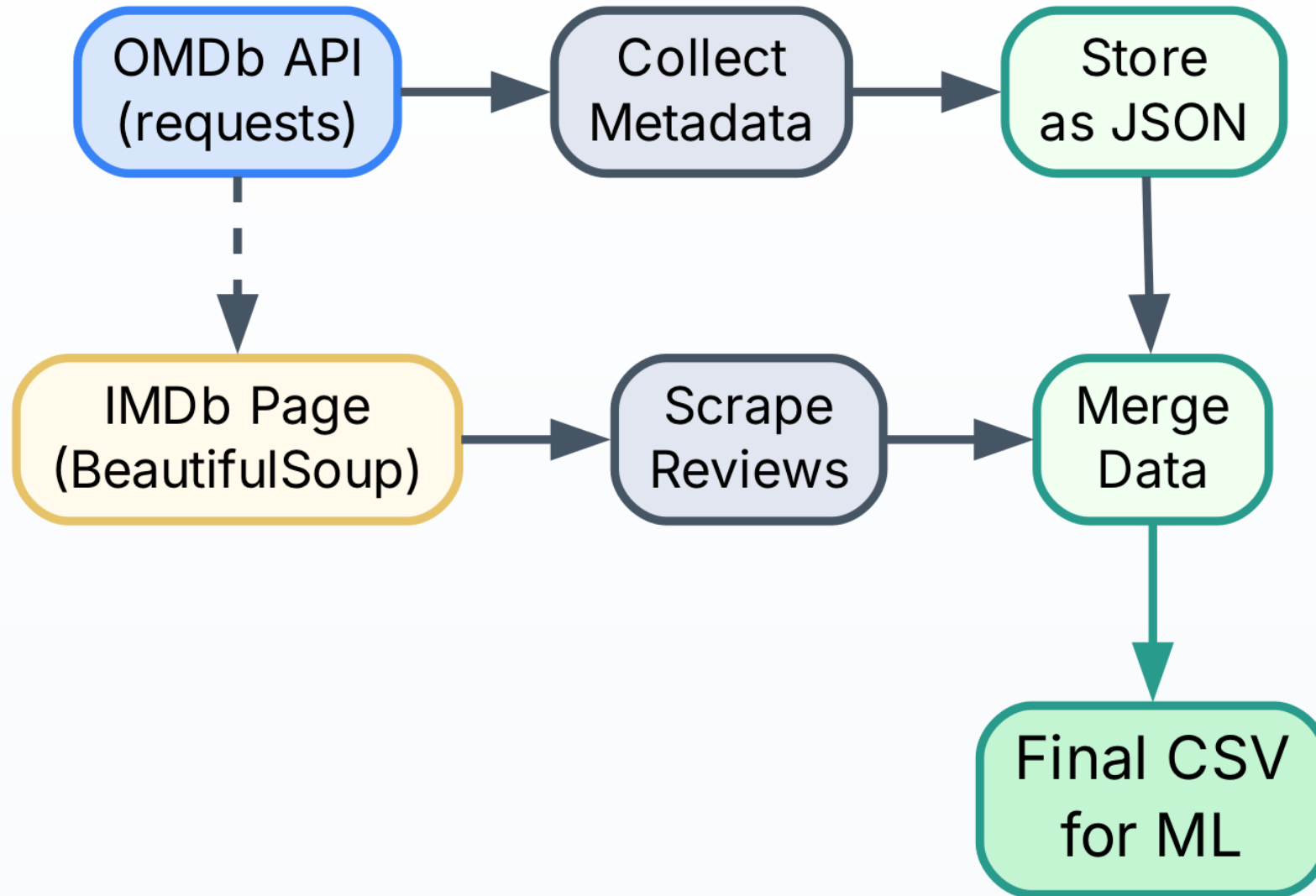
Build a dataset for movie success prediction:

Title	Year	Genre	Budget	Revenue	Rating	Director
?	?	?	?	?	?	?

We now have the tools!

- **DevTools** to find APIs
- **curl** to test requests
- **requests** to automate collection
- **BeautifulSoup** for scraping

# Our Data Collection Pipeline



# Step 1: Collect from API

```
API_KEY = "your_omdb_key" # Replace with your actual OMDb API key
movies_to_fetch = ["Inception", "Avatar", "The Matrix"]
results = []

for title in movies_to_fetch:
    response = requests.get(
        "https://www.omdbapi.com/",
        params={"apikey": API_KEY, "t": title},
        timeout=10 # Prevent hanging requests
    )

    if response.ok: # Check HTTP-level success
        data = response.json()

        # Check API-level success
        if data.get("Response") == "True":
            results.append(data)
            print(f"Fetched: {title}")
        else:
            print(f"Movie not found: {title} ({data.get('Error')})")
    else:
        print(f"HTTP error for {title}: {response.status_code}")

print(f"\nCollected {len(results)} movies")
```

## Step 2: Extract Relevant Fields

```
movies = []

for data in results:
    movie = {
        "title": data.get("Title"),
        "year": data.get("Year"),
        "genre": data.get("Genre"),
        "director": data.get("Director"),
        "rating": data.get("imdbRating"),
        "votes": data.get("imdbVotes"),
        "runtime": data.get("Runtime"),
        "imdb_id": data.get("imdbID")
    }
    movies.append(movie)
```

## Step 3: Save to CSV

```
import pandas as pd

# Convert to DataFrame
df = pd.DataFrame(movies)

# Clean data
df['year'] = pd.to_numeric(df['year'], errors='coerce')
df['rating'] = pd.to_numeric(df['rating'], errors='coerce')
df['votes'] = df['votes'].str.replace(',', '').astype(float)

# Save
df.to_csv('netflix_movie_data.csv', index=False)

print(df.head())
```



# The Result

```
   title  year      genre      director  rating
0  Inception  2010  Action, Adventure... Christopher Nolan    8.8
1    Avatar  2009  Action, Adventure...   James Cameron    7.9
2 The Matrix  1999  Action, Sci-Fi    Lana Wachowski...    8.7
```

Now ready for ML modeling!

# What We Learned: Three Tools

Tool	When to Use	Key Commands
Chrome DevTools	Discover APIs, inspect requests	Network tab, Copy as curl
curl	Test requests quickly	<code>-X</code> , <code>-H</code> , <code>-d</code> , `
Python requests	Automate collection	<code>.get()</code> , <code>.post()</code> , <code>.json()</code>

Plus BeautifulSoup for scraping when needed!

# Part 12: Looking Ahead

*Lab preview and next week*

# This Week's Lab

## Hands-on Practice:

1. **Chrome DevTools** - Inspect API calls on real websites
2. **curl exercises** - Making API requests from terminal
3. **OMDb API** - Collecting movie metadata
4. **Python requests** - Building a data collection script
5. **BeautifulSoup** - Scraping a sample website

**Goal:** Build a working data collection pipeline.

# Lab Environment Setup

```
# Install dependencies
pip install requests beautifulsoup4 pandas

# Get your API keys
# OMDb: https://www.omdbapi.com/apikey.aspx (free tier)

# Verify installation
python -c "import requests; print('Ready!')"
```

# Next Week Preview

## Week 2: Data Validation & Cleaning

- Schema validation with Pydantic
- Handling missing data
- Type conversion and normalization
- Data quality checks
- Building validation pipelines

The data we collect today needs cleaning tomorrow!

# Key Takeaways

1. **Data collection is 80% of ML work** - don't underestimate it
2. **DevTools reveals hidden APIs** - always check before scraping
3. **curl for quick testing** - then convert to Python
4. **requests for automation** - handle loops, errors, storage
5. **Scraping is plan B** - use when APIs don't exist
6. **Be ethical** - respect robots.txt, rate limits, ToS

# Resources

## Documentation:

- curl: <https://curl.se/docs/>
- requests: <https://requests.readthedocs.io/>
- BeautifulSoup: <https://beautiful-soup-4.readthedocs.io/>

## Free APIs for Practice:

- JSONPlaceholder: <https://jsonplaceholder.typicode.com/>
- OMDb: <https://www.omdbapi.com/>
- Public APIs list: <https://github.com/public-apis/public-apis>



# Questions?

# Thank You!

See you in the lab!