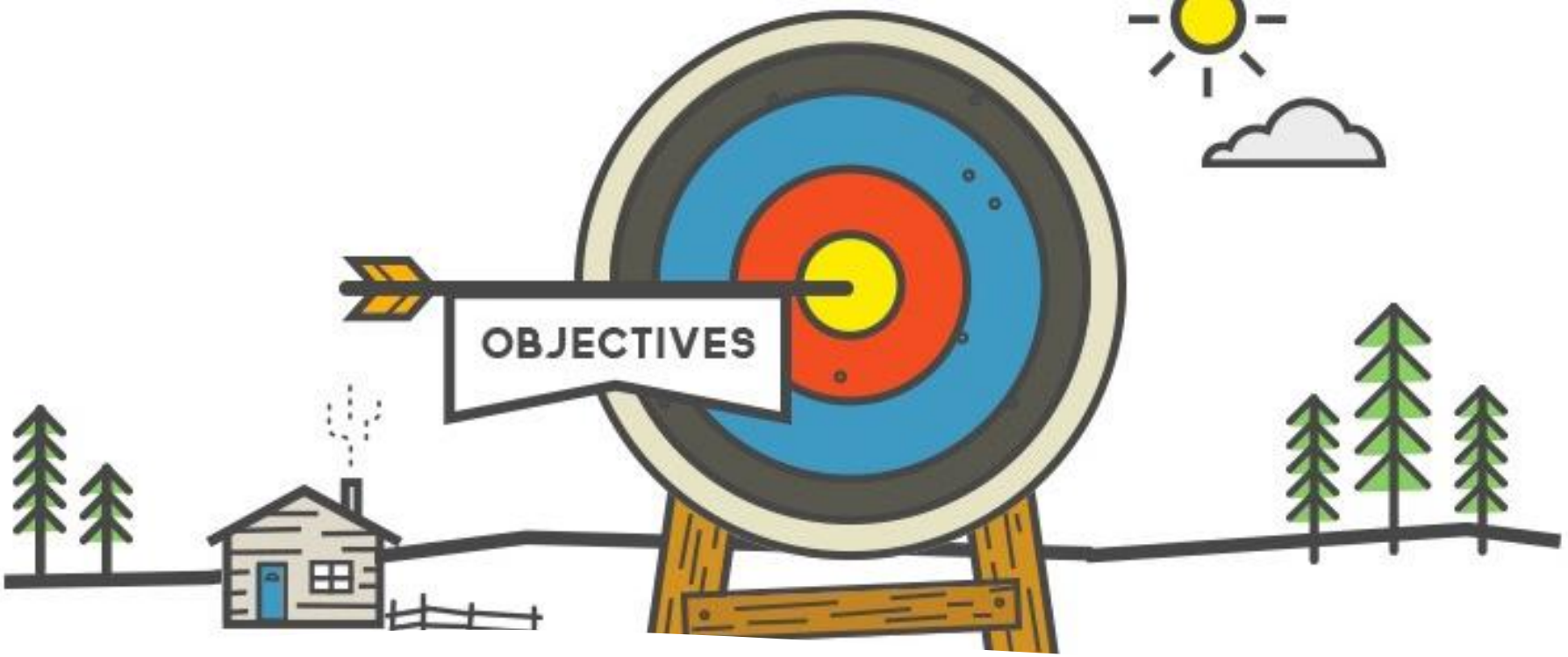
Data Collection and Management

Prof. Carlos J. Costa, PhD



Learning Goals

- Understand different methods of data collection
- Apply real-world data collection techniques
- Distinguish between structured and unstructured data storage
- Design basic data storage strategies
- Evaluate governance, ethics, and sustainability issues

PART 1: DATA COLLECTION

What is Data Collection?

- Systematic process of gathering data
- Foundation for analytics and decision-making
- Poor data leads to poor insights ("garbage in, garbage out")

Real-World Example

Example: E-commerce Company

- Tracks clicks, purchases, browsing time
- Uses data to recommend products

Types of Data Sources

- Primary Data (surveys, experiments)
- Secondary Data (reports, APIs)
- Internal (CRM, ERP)
- External (social media, open datasets)

Surveys (Concept)

- Structured questionnaires
- Online tools (e.g. Google Forms, Qualtrics)

Surveys (Practical Example)

- Example: Customer satisfaction survey
- Q1: Rate experience (1–5)
- Q2: Would you recommend us?
- Output: Structured dataset

https://github.com/masterfloss/FakeNewsData

Hands-on Mini Exercise

Design 3 survey questions for:

- A sustainability-focused app

Web Scraping (Concept)

- Extracting website data automatically
- Used in price tracking, sentiment analysis

Web Scraping (Code Example)

```
import requests
from bs4 import BeautifulSoup


url = "https://example.com"
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")


for title in soup.find_all("h2"):
print(title.text)
```

Web Scraping (Code Example)

```
import requests
from bs4 import BeautifulSoup

url = "https://quotes.toscrape.com"
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")

# Find all quote div elements
quotes = soup.find_all('div', class_='quote')

# Loop through each quote and extract text and author
for quote in quotes:
    text = quote.find('span', class_='text').text
    author = quote.find('small', class_='author').text
    print(f"\"" + text + "\" - " + author)
```

Web Scraping Use Case

Example: Competitor price monitoring

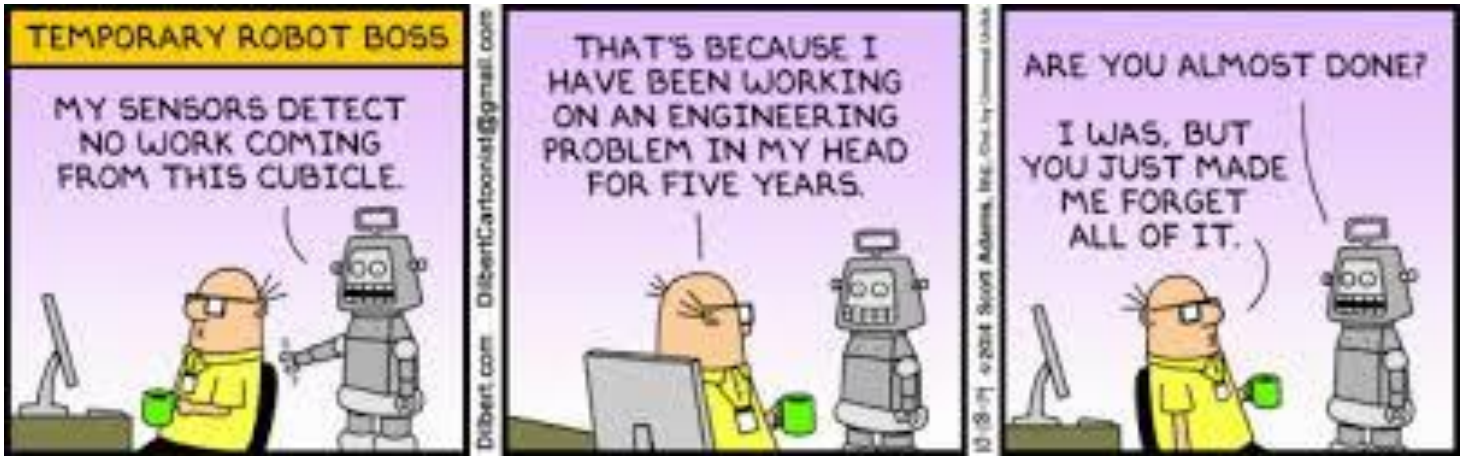
- Collect product prices daily
- Analyze trends

Sensors & IoT

- Devices collect real-time data
- Examples: temperature, traffic, pollution

IoT Example (Sustainability)

- Smart city sensors measure air quality
- Data used to reduce pollution



API-Based Data Collection

- Access structured data via APIs
- Example: weather APIs, Twitter API

Exercise 2 – Collect Data via API

```
import requests
import pandas as pd

url = "https://jsonplaceholder.typicode.com/posts"
response = requests.get(url)
data = response.json()

df = pd.DataFrame(data)
df.to_csv("posts.csv", index=False)
print(df.head())
```

API types: Based on Architecture / Style

- **REST (Representational State Transfer)**
  - Most popular;
  - uses HTTP methods (GET, POST, PUT, DELETE).
- **SOAP (Simple Object Access Protocol)**
  - XML-based and more strict;
  - used in enterprise systems.
- **GraphQL**
  - Allows clients to request only the data they need.
- **gRPC**
  - High-performance,
  - uses Protocol Buffers (commonly used in microservices).

API types: Based on Functionality

- **Web APIs**
  - Operate over the internet (HTTP/HTTPS).
- **Library APIs**
  - Provided by programming libraries (e.g., Java API).
- **Operating System APIs**
  - Allow apps to interact with the OS (e.g., Windows API).
- **Database APIs**
  - Used to communicate with databases (e.g., JDBC).

API Example

```
import requests

# API request
url = "https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/une_rt_m"
params = {
    "geo": "PT",
    "unit": "PC_ACT",
    "sex": "T"
}
data = requests.get(url, params=params).json()
# Extract data
values = data["value"]
times = data["dimension"]["time"]["category"]["index"]
# Print results
print("--- Unemployment Data ---")
for time, idx in times.items():
    value = values.get(str(idx))
    if value:
        print(f"{time}: {value}")
```

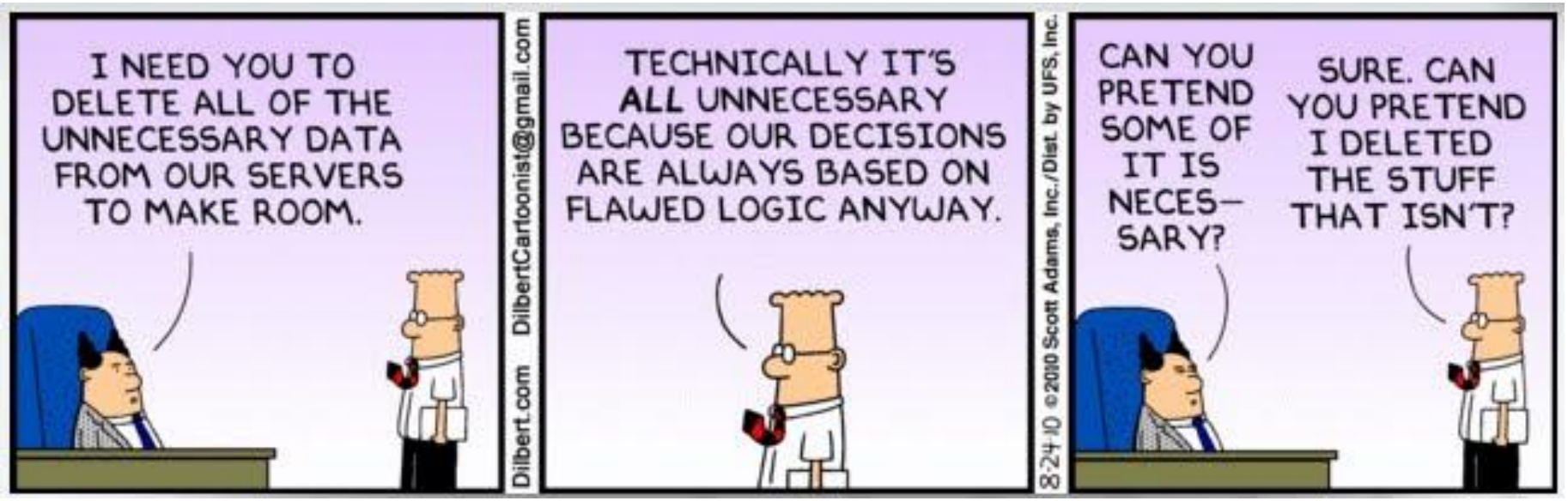
FINALLY! AFTER ALL THOSE YEARS
I FINALLY FOUND
THE SOURCE OF THE DATA!

Activity (15 min)

- Choose a scenario:
  - Smart city
  - E-commerce
  - Healthcare
- Define:
  - Data source
  - Collection method

"How can you say we're not behaving like a team?
We're all wearing the same color shirts, aren't we?"

PART 2: DATA STORAGE & MANAGEMENT

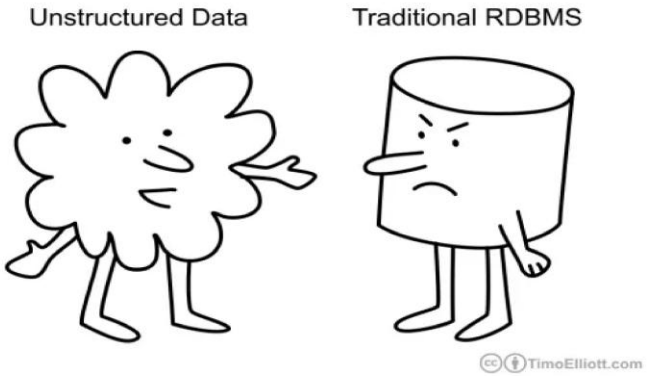


Why Storage Matters

- Enables retrieval and analysis
- Must handle scale, speed, and variety

Structured Data

- Tables, rows, columns
- Example: sales database



"I'm sorry, I'm just not that into you..."

Unstructured Data

- Text, images, videos
- Example: social media posts

Semi-Structured Data

- JSON, XML
- Flexible schema

Practical Example (Data Types)

Data	Type
Excel file	Structured
Tweets	Unstructured
JSON API	Semi-structured

Relational Databases

- SQL-based
- Example: MySQL, PostgreSQL

```
SELECT * FROM customers
WHERE country = 'USA';
```

NoSQL Databases

- Flexible schema
- Example: MongoDB, Cassandra

```
{
"name": "Alice",
"age": 25,
"purchases": ["book",
"laptop"]
}
```

Data Warehouses

- Optimized for analytics
- Clean, structured data

Data Lakes

- Store raw data
- Flexible and scalable

Warehouse vs Lake Example

- Warehouse - Business reports
- Lake - Raw IoT sensor data

Cloud Storage

- AWS S3, Google Cloud Storage
- Scalable and distributed

Activity (10 min)

- Given datasets:
  - Images
  - Transactions
  - Logs
- Choose best storage solution

PART 3: GOVERNANCE, ETHICS & SUSTAINABILITY

- GOVERNANCE,
- ETHICS &
- SUSTAINABILITY

"They say that good governance is like a recipe, so we invited an expert."

What is Data Governance?

- Framework for managing data
- Ensures quality and security

Governance Example

- Bank ensures:
  - Accurate transactions
  - Secure customer data

Data Quality Issues

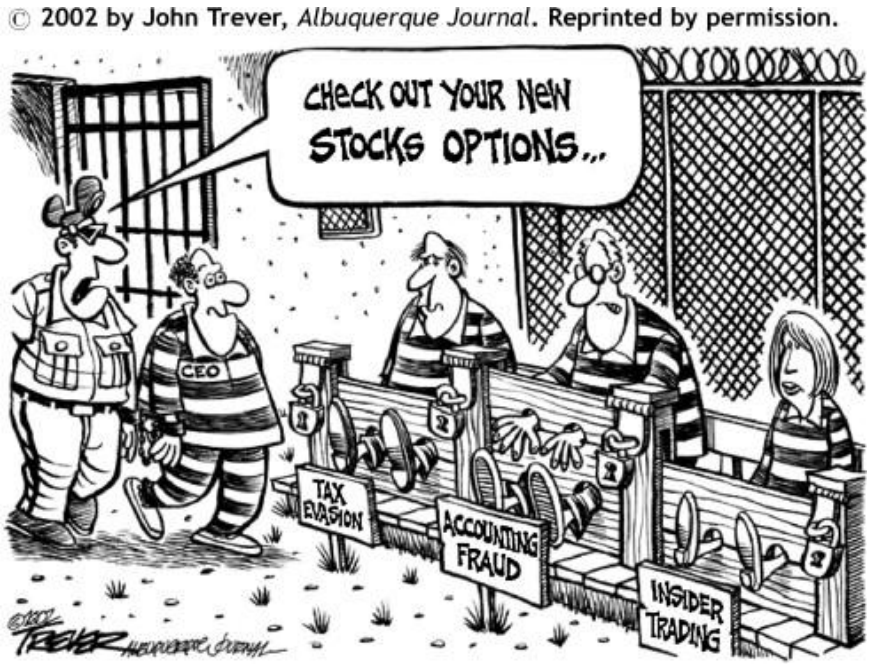
- Missing data
- Duplicate records
- Inconsistent formats

Data Privacy

- Personal data protection
- Regulations: GDPR

Ethical Issues

- Bias in data
- Misuse of personal information



Ethical Case Example

- Social media platform selling user data
- Consequences: loss of trust

Environmental Impact

- Data centers consume energy
- Carbon emissions impact sustainability

Green Data Practices

- Efficient storage
- Renewable energy usage
- Data minimization

Real Case (Sustainability)

- Company reduces storage by 30%
- Saves energy and costs

PART 4: PRACTICAL SESSION (20 MIN)

- Teamwork and Dicussion

Practical Exercise

- Dataset scenario:
  - Online store OR IoT system
- Tasks:
  - Define data collection method
  - Choose storage type
  - Identify ethical risks

Group Discussion

- Present your solution
- Compare approaches

Summary

- Multiple data collection methods exist
- Storage depends on data type and scale
- Governance ensures reliability and ethics
- Sustainability is increasingly important

Assignment

- Find a real dataset
- Describe:
  - How it was collected
  - Where it is stored
  - Ethical & sustainability concerns