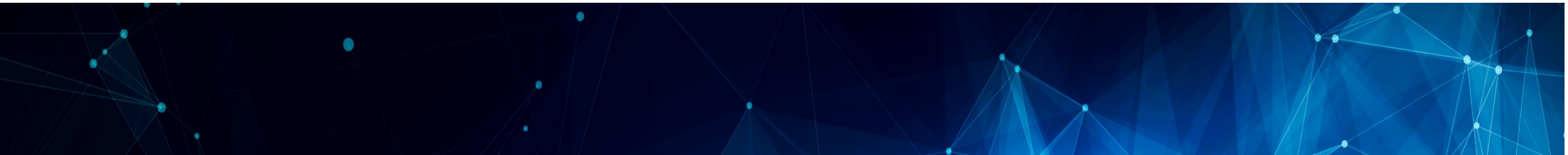
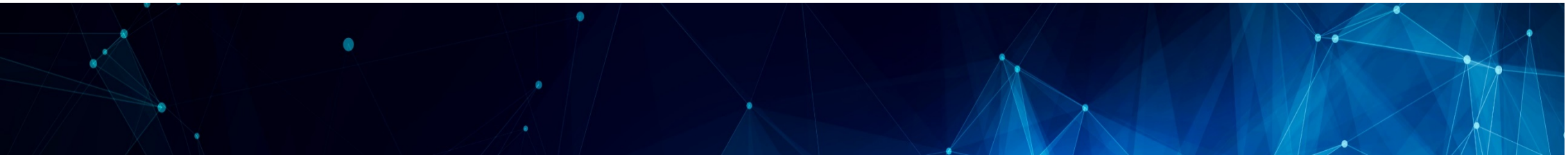
# Fundamentals of Data Science

## Data collection and management



Southampton Data Science Academy

# Data collection: sources of data



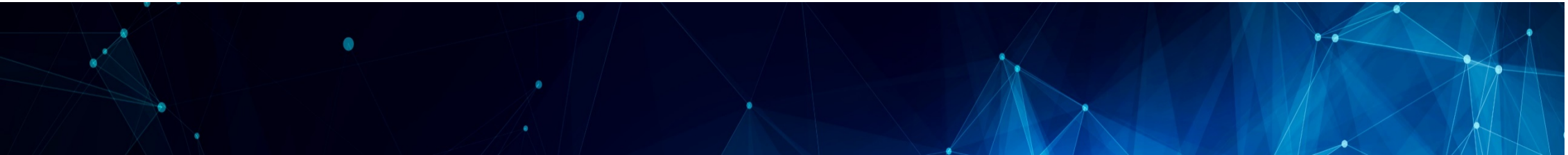
Southampton Data Science Academy

# Data collection

Types and sources of data

Collecting data using RESTful APIs

Collecting data using page scraping



Southampton Data Science Academy

# Big Data?

Data science is often associated with “big data” which is large scale, unstructured, and timely

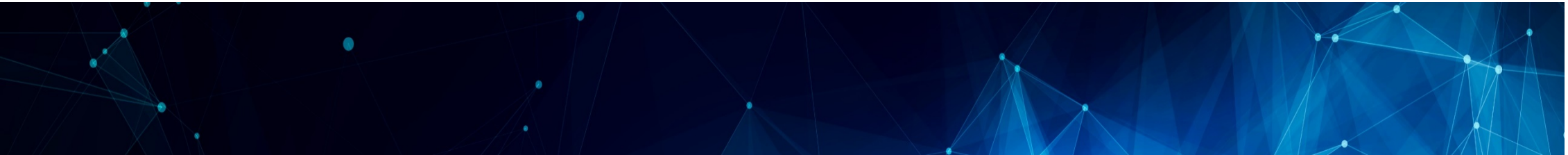
Often referred to in the context of four “Vs”: Volume, Variety, Velocity and Veracity

Large scale data helps eliminate randomness, and allows additional insights to be found

Big data is a subset of data science

Often “small” data is enough to solve the problem or business need

It will add extra complexity to any data processing operation, so should only be used if there’s a business need



Southampton Data Science Academy

# Possible Data Sources

Sensors on hardware

- E.g., boats, planes, cars, …

Digital networks

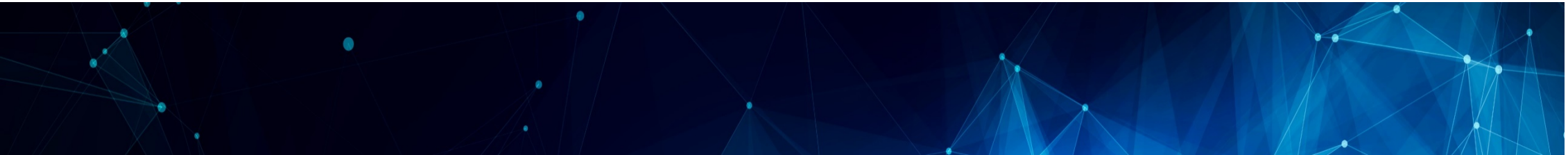
- Web, Internet, …

Physical devices

- Phones, credit card, GPS, …

Finance

- Stock market, currency exchange, …



Southampton Data Science Academy

# Getting the data

Two types of data (can be complementary)

- Static datasets, e.g., database dumps
- Streaming data, e.g., real-time stream of GPS coordinates

In this course, we will be focusing on static data

# Getting data – RESTful APIs

The preferred way to obtain data on the Web is to use a RESTful API where a company makes a service available on their server

A RESTful API identifies a resource on the server, which will have one of the HTTP verbs applied to it from the request

- Examples: GET, POST, PUT, DELETE

This is just an example of the exact same technology as Web browsing

The server can return the data in any format, although JSON or XML are the most common

- Example: GET /api/user/42