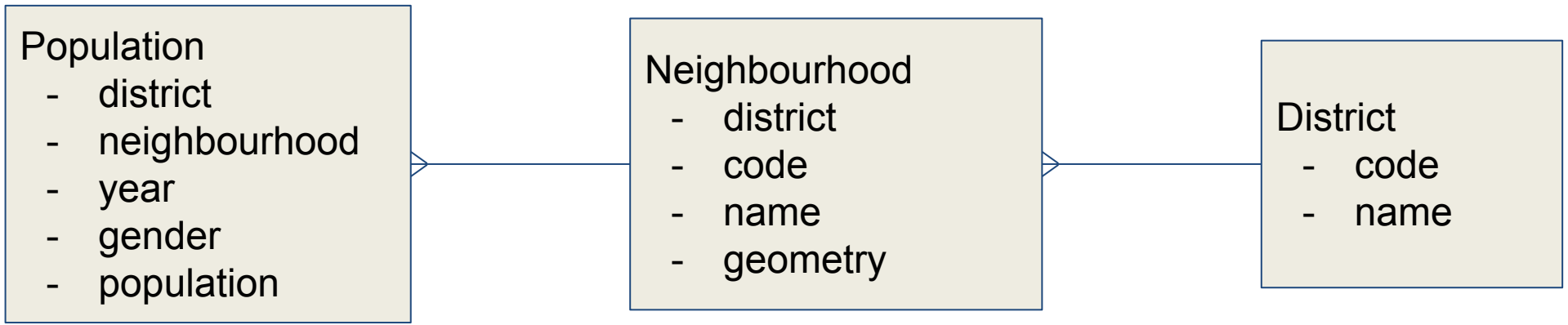
Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA - Second Term
2019/2020

FACULTY DIEGO PAJARITO

Big Data

Traditional data sources



Data source:

https://opendata-ajuntament.barcelona.cat/data/en/dataset/20170706-districtes-barris
https://opendata-ajuntament.barcelona.cat/data/en/dataset/est-ine-sexe

There are different steps to create a database.

Users, database and tables are created through scripts.

Geospatial information is created using GIS tools

There is a backup file if you want to set your own PostgreSQL server and load it life class.

Use pgadmin to get a connection with the database

Connection Name: your_name
Username: mact
Password: mact
Host: 192.168.0.21

Comment on the data structures and presentation

Use qgis to get a connection with the database.
Browser: Postgis/new connection
Menu: Database/DB Manager

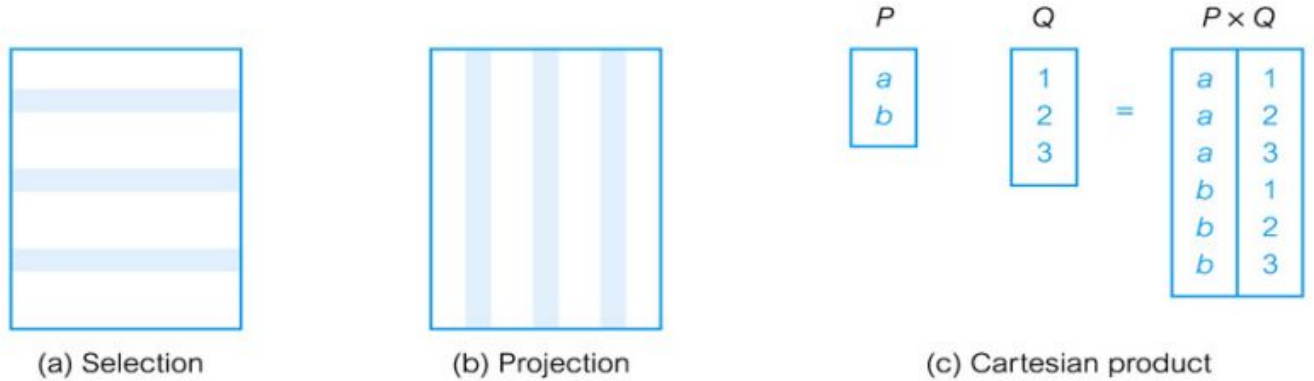
Connection Name: your_name
Database: example
Username: mact
Password: mact
Host: 192.168.0.21

Comment on the differences you see when using a GIS interface

Go through the following queries and find the general procedure to extract information from a database

- **What neighbourhoods are part of district X?**
- **What is the population per district?**
- **How many people live in district X?**
- **Order Barcelona districts using population**
- **Any other query?**

Map neighbourhoods per population using QGIS



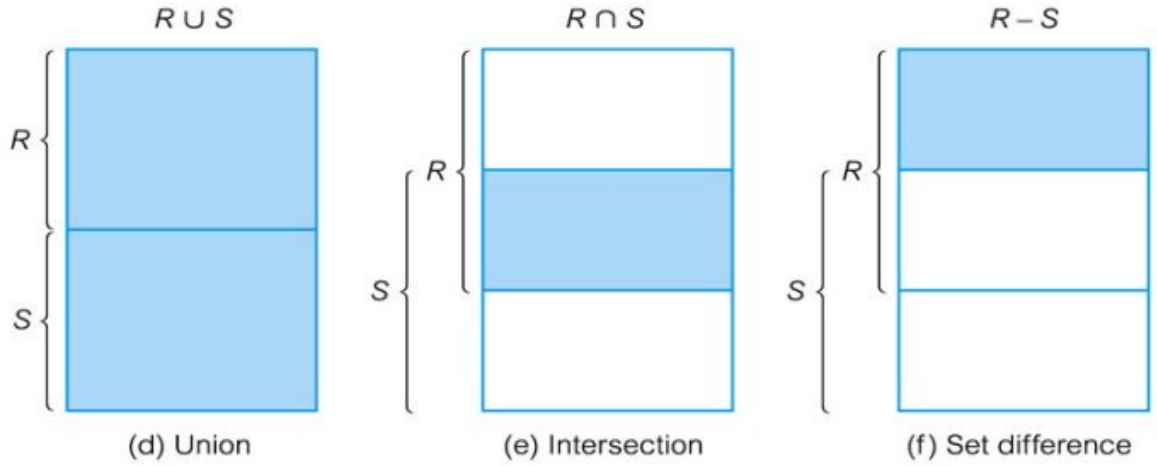


Illustration by Robert M. Siegfried, Ph. D. Department of Mathematics and Computer Science
Adelphi University, Garden City, NY 11530
Full presentation available at: https://home.adelphi.edu/~siegfried/cs443/443l9.pdf
More information about Relational Algebra available https://en.wikipedia.org/wiki/Relational_algebra

PostgreSQL also supports geospatial queries while enhancing the analysis options

- **What are the district areas?**
- **Spatially Aggregate neighbourhoods into districts**
- **What is the population density per district?**

Think about spatial relationships within the existing datasets

Visit different websites (e.g., tourism, local commerce, real estate or rentals) and try to imagine a new data set or structure to analyse urban activities

Discuss with your peers on the way these data sets add value or support urban design

Big Data

Web Data Sources - API

Find Restaurants in Bogotá
Trip Advisor, Lonely Planet, The fork, ViaMichelin, Yelp, The Bogotá Post

How many restaurants there are?
Which are located near the airport?
What are the areas having more restaurants?

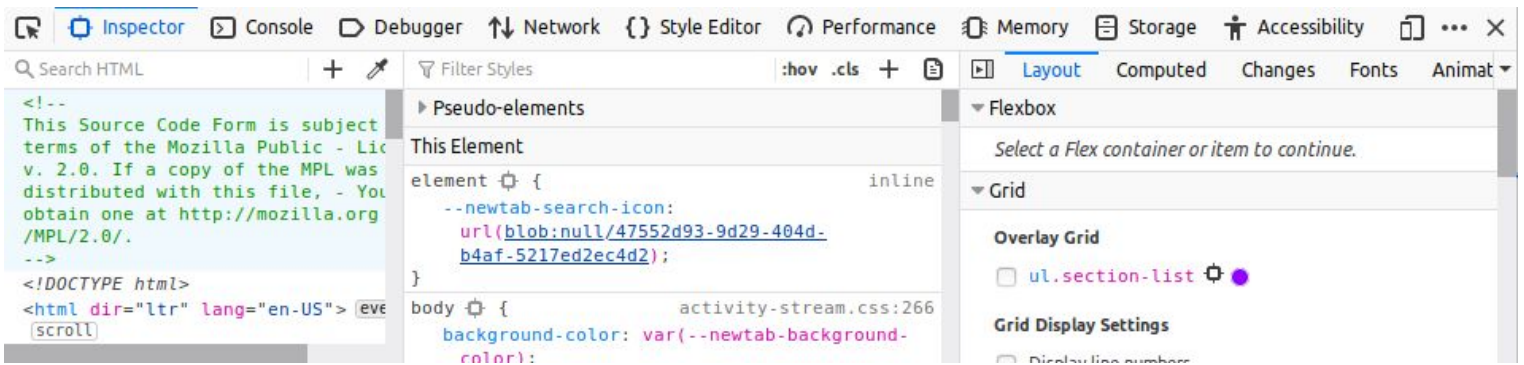
After finding the restaurants, try to create a suitable dataset for the analysis

Table, dataframe, csv...

Restaurants
- name
- address
- rating
- price
- city

Get a closer view of the websites using either the “F12” key the “View Source Code” option.

Does it look like a data set?
Network, inspector, sources...



Foursquare offers a search API

https://developer.foursquare.com/docs/api/venues/search

Get the credentials to use it and try to get information from a city of interest

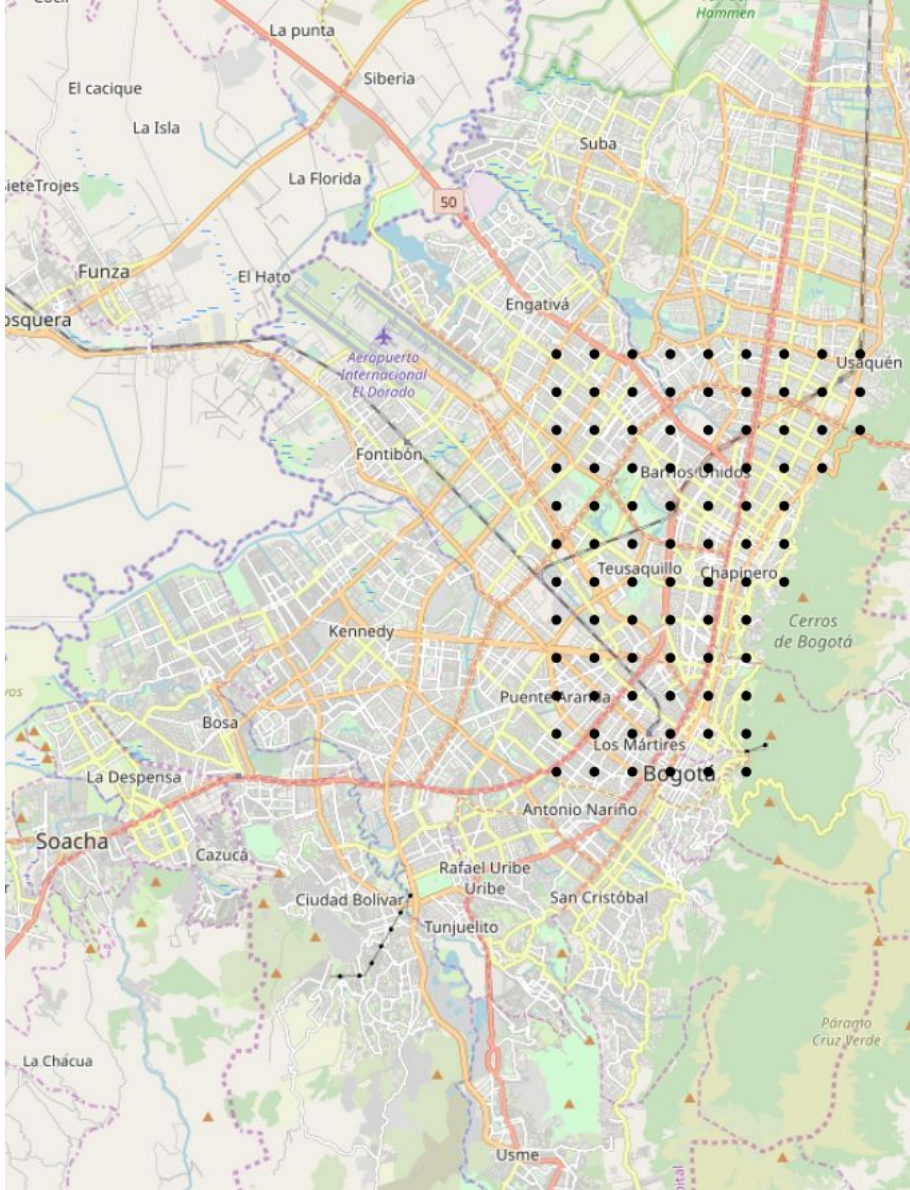
Analyse the use of regular grids to feed requests

Regular grid generated using Qgis

API location using lat,long

General arrangement for optimisation

Gather data for Barcelona and two more cities.



Big Data

Web Data Sources - Scrapy

A library to get data from web resources

Using Anaconda environment...

```
pip install Scrapy
```

... after, within the script

```
import scrapy
```

Follow the scrapy tutorial here: https://docs.scrapy.org/en/latest/intro/install.html

Scrapy Tutorial

https://docs.scrapy.org/en/latest/intro/tutorial.html

Follow this tutorial to test the tool and extract data from a web site.

Although “Scrapy” is a python-based tool, it is mainly controlled through the terminal. Navigate to this folder

```
../b_Non.../Trip_Advisor_scrapy/
```

Create a scrapy project

```
scrapy startproject tripadvisor
```

Run a “spider”

```
scrapy crawl your_spider_name
```

Follow the scrapy tutorial here: https://docs.scrapy.org/en/latest/intro/tutorial.html

There are tools to search and extract HTML object using the ‘css’ and ‘xpath’ commands.

Compare the browser structure and the query tasks for

```
scrapy shell https://www.tripadvisor.com/Restaurants-g294074-Bogota.html
```

GitHub repository: https://github.com/IaaC/MACT19.20_Digital_tools_Big_Data_part_2

Using the following commands, try to find HTML commands such as Title, Hyperlinks, Texts, Tables and “next” button

```
response.css('title').get()

response.css('type.class::text').get()

response.attrib['href']

response.xpath("//div[@class='search_for_class']").get()

response.xpath("//a[@class='search_for_class']").get()
```

GitHub repository: https://github.com/IaaC/MACT19.20_Digital_tools_Big_Data_part_2

Find the spider name and use it to run (crawl) the spider

Run a “spider”

```
scrapy crawl tripadvisor_a
```

```
scrapy crawl tripadvisor_a -o output.json
```

Analyse the results either in the screen or in the file

GitHub repository: https://github.com/IaaC/MACT19.20_Digital_tools_Big_Data_part_2

‘Parse’ function manages data extraction

```
def parse(self, response):
  for restaurant in response.css("a._15_ydu6b"):
    yield {
      'name': restaurant.get(),
      'url': restaurant.attrib['href']
    }
```

Try saving the output data (yield) into a json file.

```
scrapy crawl spider_name -o file_name.json
```

We can also follow links within the existing html file ()

```
restaurant_url = response.urljoin(restaurant.attrib['href'])

yield scrapy.Request(restaurant_url, callback=self.parse_restaurant)
```

Which websites would you like to scrap?

Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA - Second Term
2019/2020

FACULTY DIEGO PAJARITO