
Big Data, Small Data

Lecture 4

Elena Shirokova

Structure

- What is data science
 - Why python
 - Work with data
 - Instruments
 - Urban applications
-

What is data science?

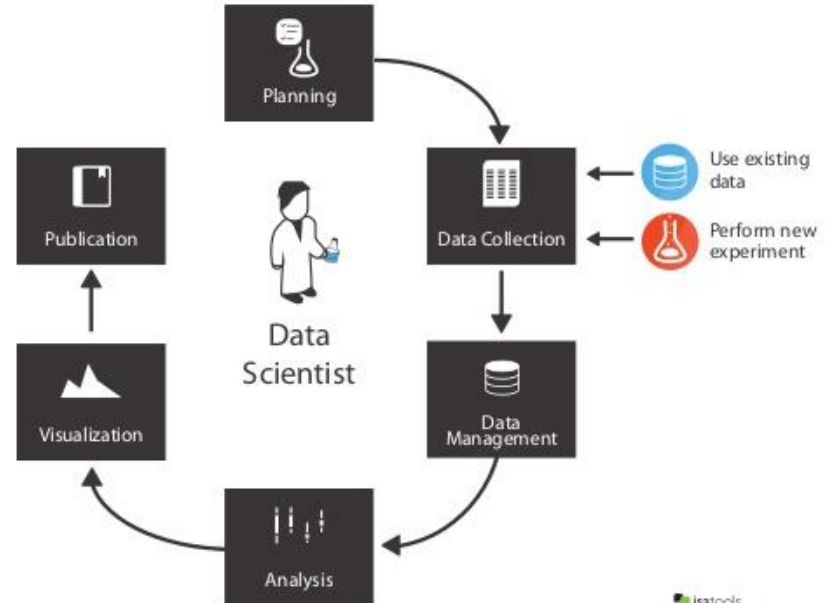
The ability to extract knowledge and insights from large and complex data sets

Data science includes:

- statistics and probability theory
 - linear algebra
 - machine learning
-

Data science workflow

The workflow of a “data scientist”...



Why data science

Social science applications

- People satisfaction level
- Communities detection

Urban studies

- Places collocation,
 - People transport behaviour
 - City zoning
 - Urban mobility, space development
-

Data

Numerical data

- time,
- station load
- temperature,
- pollution level

Categorical data

- City location
- District name

Spatial data

- Spatial polygons
 - lines
 - points with geo
-

Data terminology

Data aggregation

The act of collecting data from multiple sources for the purpose of reporting or analysis.

Data cleaning

Preprocessing data to remove duplicate entries, correct misspellings, add missing data, and provide more consistency.

API

An abbreviation for Application Program Interface, a set of programming standards and instructions for accessing or building web-based software applications.

Features

The measurements which represent the data

Scrapping

Automatically extracting data from websites

Data

Social media

- Twitter
- Instagram
- Foursquare

Review sites

- Yelp
- Tripadvisor
- Booking
- Airbnb

Images

- Satellite images
- Traffic cameras

Maps

- OSM
- Wikimapia
- Booking

Open city data

- NYC Open data
- London TFL

GPS data

- Navigation data
 - Mobile data
-

Data formats and sources

Range of data: tables, time series, texts, photos, graphs, etc.

Popular data formats:

- CSV - comma separated files
 - JSON
 - GeoJSON
 - Shapefile
-

Data extraction

Using open API

- Twitter API,
- Foursquare API,
- Google Places API

Download from open sites

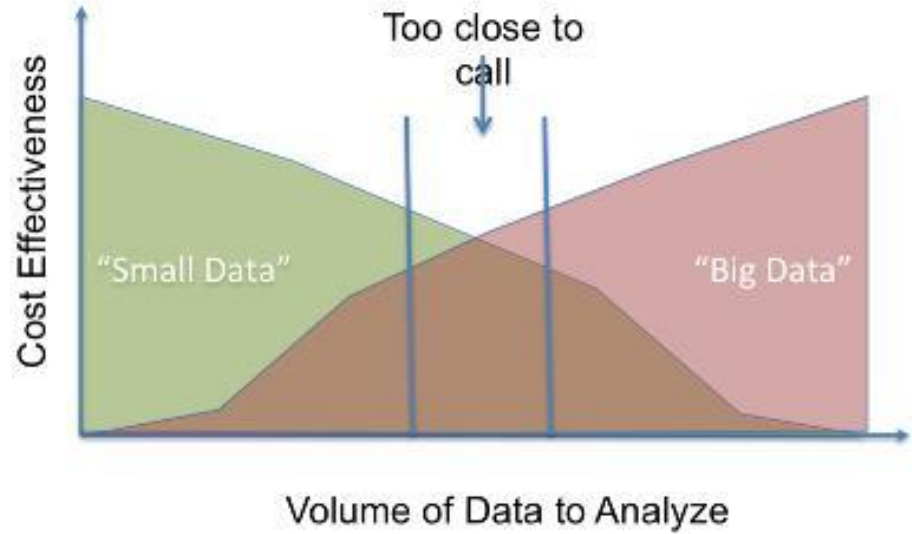
- OpenStreetMap,
- city open data

Site scraping

Big data vs small data

Small data is data in a volume and format that makes it accessible, informative and actionable.

Big Data are huge chunks of structured and unstructured data.



City open data

[London Datastore](#)

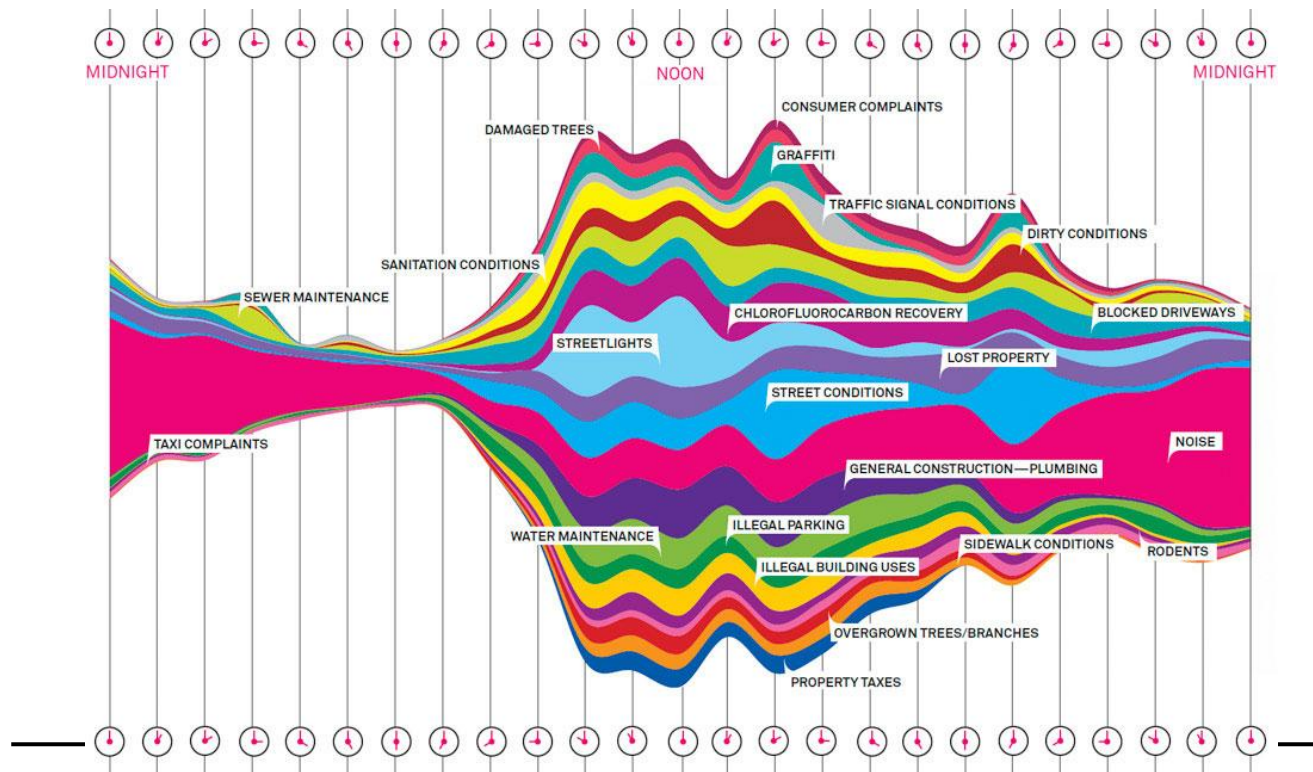
[NYC Open data](#)

[Data from the City of Chicago](#)

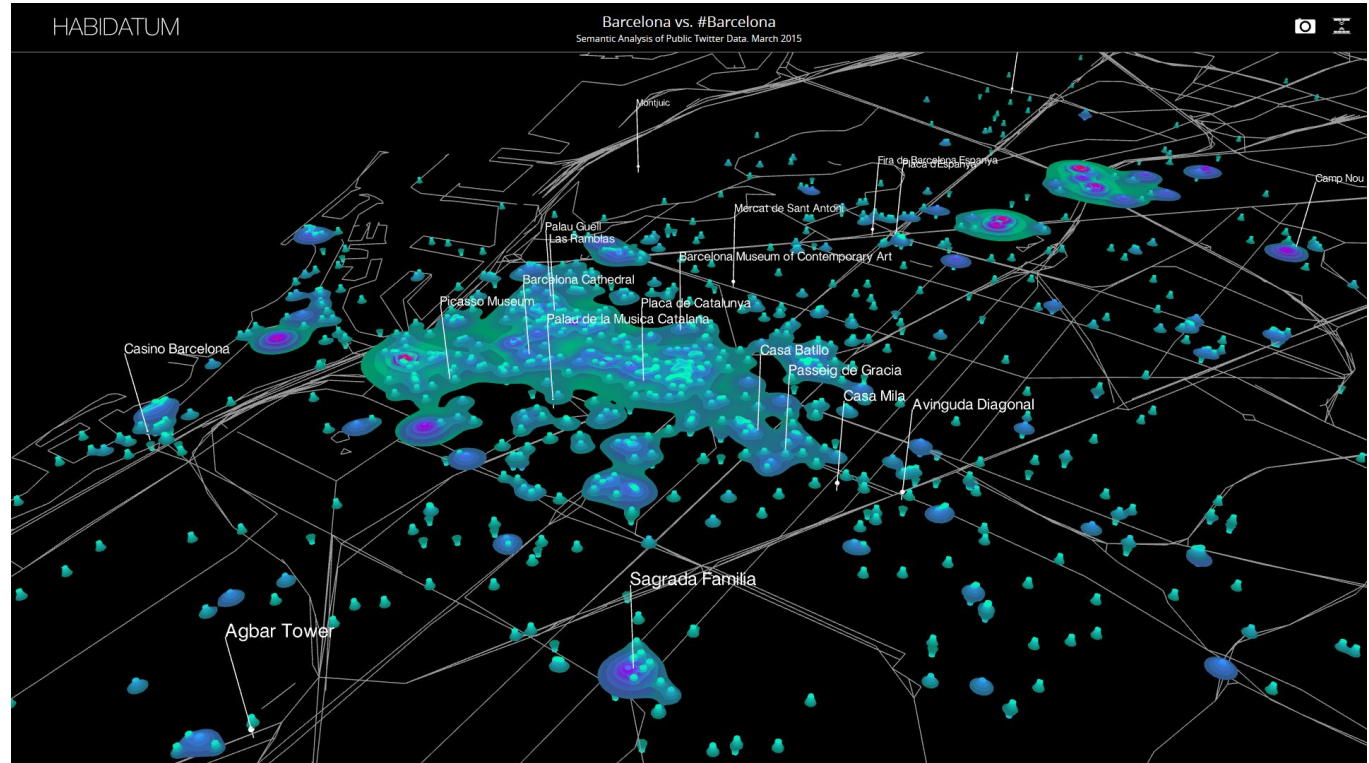
Relationship between data and city

311 complaint data from the NYC Open Data portal, published by the Department of Housing Preservation and Development (HPD)

This dataset represents all calls by tenants to the city since 2010 to report issues in their apartments or building-wide issues



Relationship between data and city

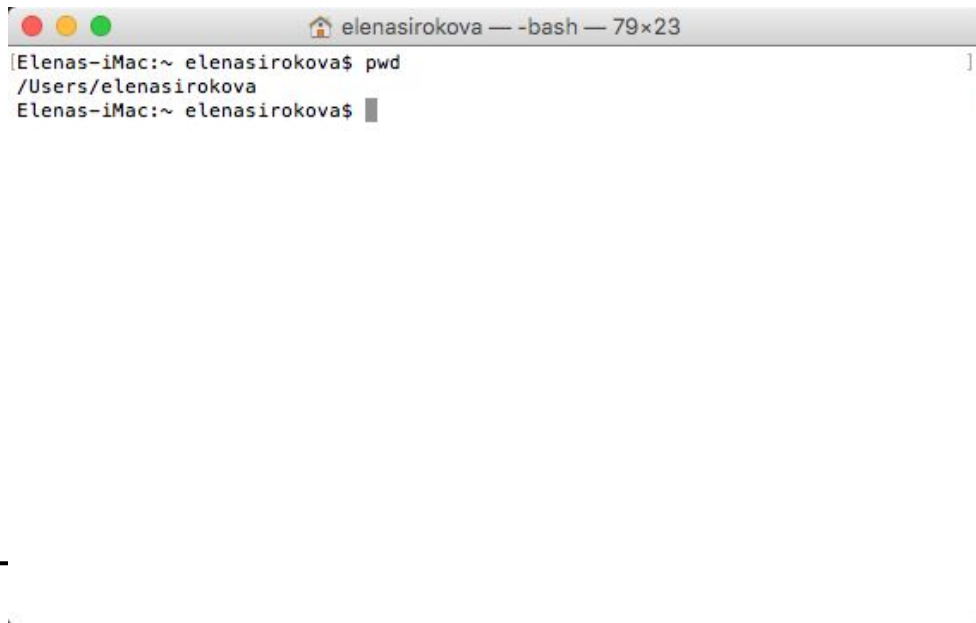


Data science instruments

- Programming language - Python, Bash
 - Version control systems - Git
 - Data storage - MongoDB, Postgres, MySQL
 - Visualization tools - Carto, Tableau, Matplotlib
-

Bash

Bash is a program on your computer like any other, but designed to be easy for you to talk to

A screenshot of a macOS terminal window. The title bar shows a home icon, the username 'elenasirokova', and the shell type '-bash' with window dimensions '79x23'. The terminal content shows the prompt 'Elenas-iMac:~ elenasirokova\$' followed by the command 'pwd', which returns the output '/Users/elenasirokova'. The prompt then returns to 'Elenas-iMac:~ elenasirokova\$' with a cursor. A small '1' is visible in the top right corner of the terminal window.

```
elenasirokova — -bash — 79x23
Elenas-iMac:~ elenasirokova$ pwd
/Users/elenasirokova
Elenas-iMac:~ elenasirokova$
```

Bash commands

List all files: **ls -a**
ls -l

Make directory: **mkdir DIR**

Change directory: **cd** PATH . current dir

cd PATH .. parent dir

cd PATH ~ home dir

Print working directory: **pwd**

Create empty file: **touch** FILE

Copy: **cp** FROM TO

Rename or move files: **mv** FROM TO

Ctrl-key commands

Remove: **rm** [OPT] File

-r recursively remove directories

Concatenate and print files: **cat** FILES

View file: **less** FILE

Display first lines: **head** FILE

Display last lines: **tail** File

Kill process: **Ctrl + C**

Stop process: **Ctrl + Z**

I/O redirections

Redirect stdout to a file:

command > FILE overwrite

command >> FILE append

Redirect stdin to a file:

command < FILE

Redirect the output of one command as input to the next one:

command1 | command2 | command3

Additional commands

* - any number of characters

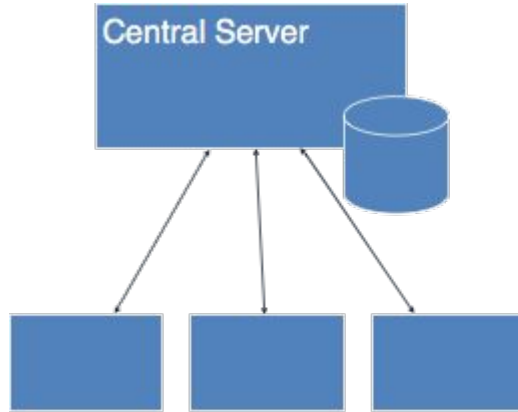
? - any single character

history - print list of previous commands

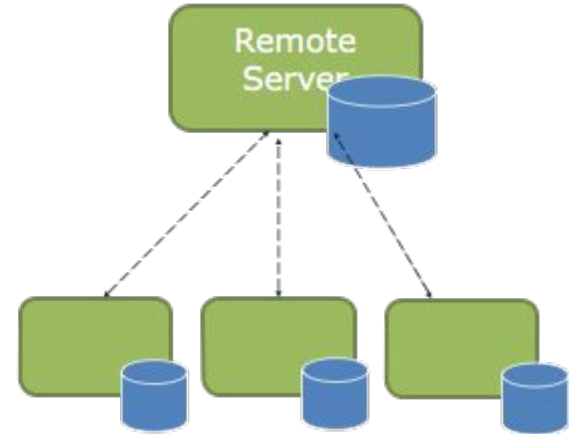
top - display top cpu process

Version control systems

Centralized VC



Distributed VC



What is Git?

Distributed version control system

Why Git is cool?

- Everyone has the complete history
 - Everything is done offline
 - No central authority
-

Git terminology

Directory

A folder used for storing multiple files.

Repository

A directory where Git has been initialized to start version controlling your files.

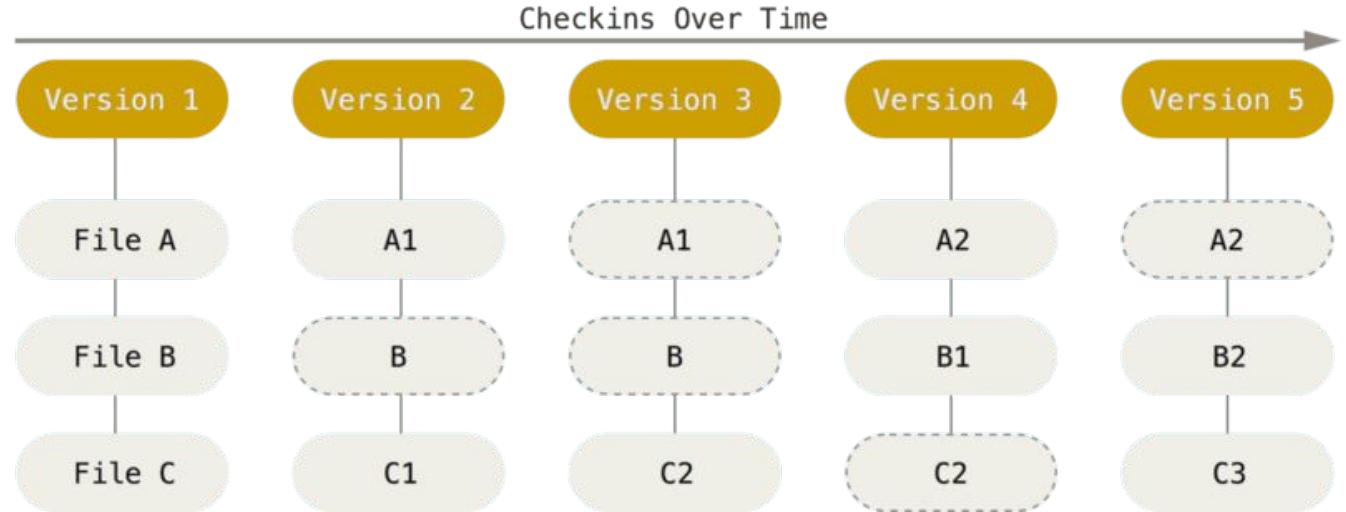
Master

The default development branch.

Git

Git stores not file versions but **difference** between files.

The difference are called **snapshots** of project without having to copy the whole directory each time.



Git terminology

gitfile

A plain file `.git` at the root of a working tree that points at the directory that is the real repository.

branch

A "branch" is an active line of development

commit

The action of storing a new snapshot of the project's state in the Git history

push

Send your commits to a remote repository

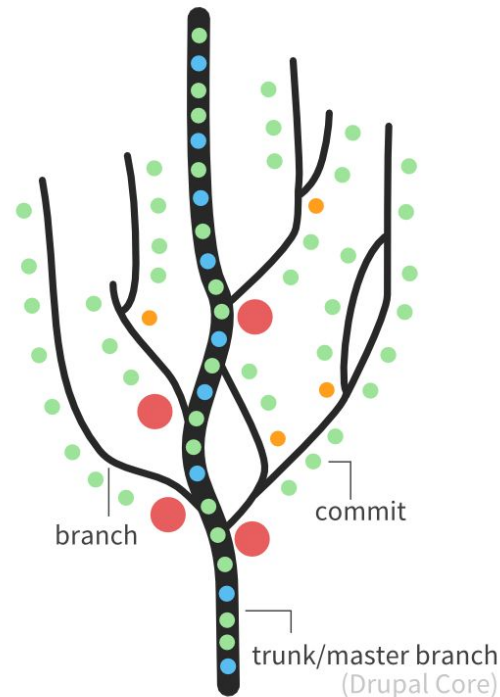
Git structure

Your git repository looks like a **tree** with the main master branch and additional developer branches

Green circles are **commit** to the branches

Right circles are **start** of new branch

Branches can be **merged** to master or split off for new features



What is Github?

www.github.com

Largest web-based git repository

Allows code collaboration

Extra functionality
documentation, bug tracking, feature
requests, pull requests



Git resources

<https://git-scm.com/>

<https://try.github.io>

Why python?

- easy to learn
- wide range of applications
- wide online community
- main language for data science

Python can be used for almost everything ranging from software or web development to scientific applications

Python installation

Use python 3

- [Python official](#) - Python official site
 - [Anaconda](#) - open easy to install Python distribution
-

PyData stack

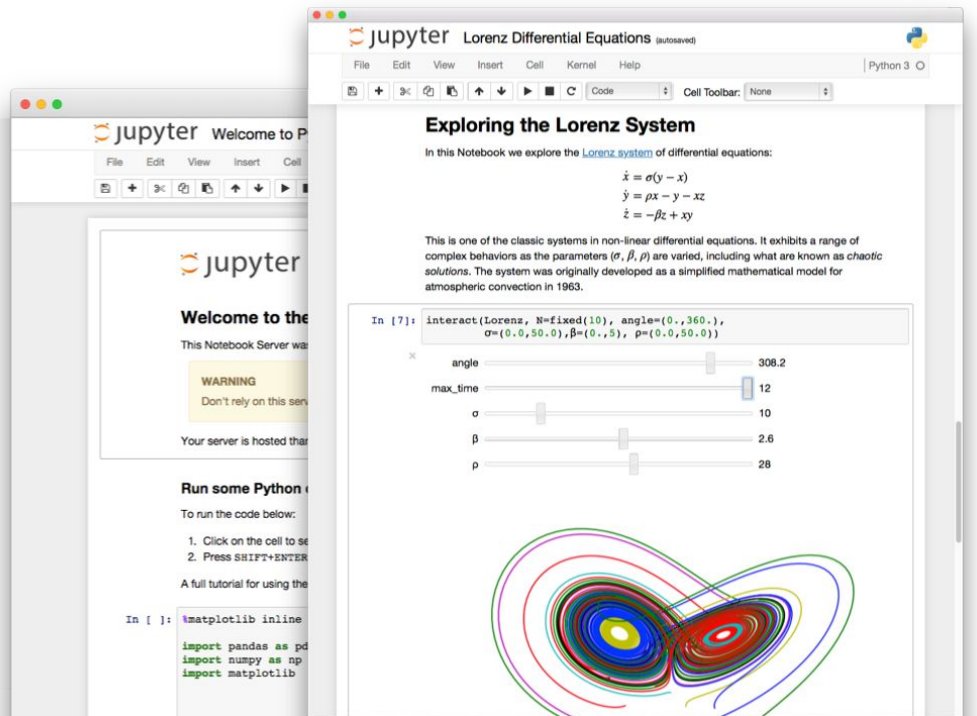
- numpy and scipy - scientific computing
 - pandas - data analysis and manipulation
 - matplotlib - visualization
 - scikit-learn - machine learning
 - geopandas, shapely, pysal - spatial analysis
-

Jupyter notebook

Flexible web-based tool for interactive computing and data analysis

- Keeps code, plots, images, comments
 - Supports various programming languages (Julia, Python, R)
 - Runs both on local machine and remote server
-

Jupyter notebook



Python resources

[Python style guide](#)

[Stackoverflow](#)

[Python cheat sheets](#)

[Gallery of Jupyter notebooks](#)

Course materials

Course repository

https://github.com/Casyfill/AUD_Metropolitan_Data

Clone repo:

git clone

https://github.com/Casyfill/AUD_Metropolitan_Data.git

Python dzen

Beautiful is better than ugly.

Explicit is better than implicit.

Simple is better than complex.

Complex is better than complicated.

Developer principles

KISS

Keep it Simple Stupid - The simplest explanation tends to be the right one

DRY

Don't Repeat Yourself - A basic strategy for reducing complexity to manageable units is to divide a system into pieces.
