Schedule

Today

Midterm assignment revealed

17th Nov

Midterm assignment due

11th Jan

Exam

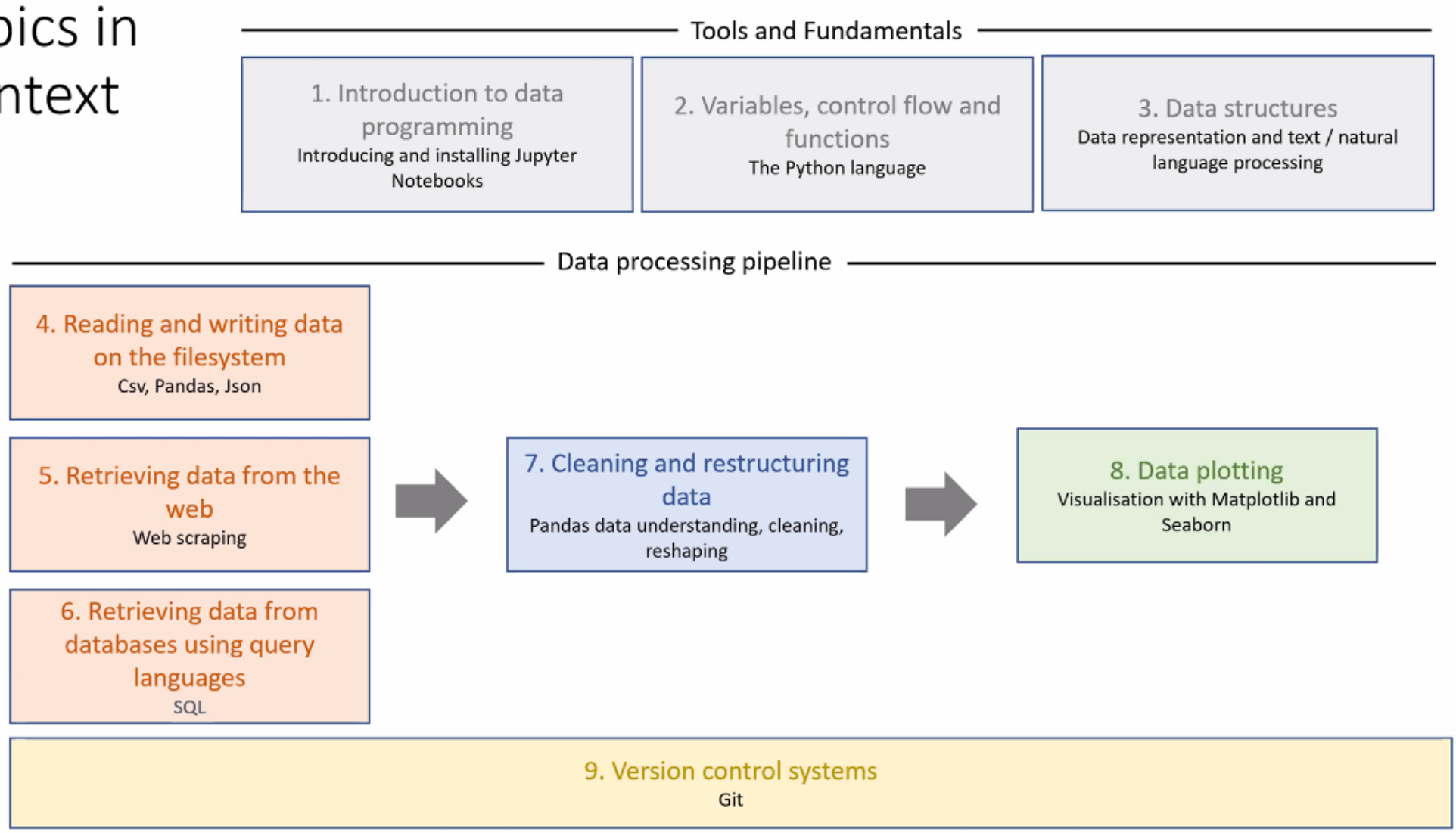
15th Mar

Week	Start Date	Topic	Title
Week 1	12-Oct-20	1	Introduction to data programming
Week 2	19-Oct-20		
Week 3	26-Oct-20	2	Variables, control flow and functions
Week 4	02-Nov-20		
Week 5	09-Nov-20	3	Data structures
Week 6	16-Nov-20		
Week 7	23-Nov-20	4	Reading and writing data on the filesystem
Week 8	30-Nov-20		
Week 9	07-Dec-20	5	Retrieving data from the web
Week 10	14-Dec-20		
Week 11	21-Dec-20	6	Retrieving data from databases using query languages
Week 12	28-Dec-20		
Week 13	04-Jan-21	7	Cleaning and restructuring data
Week 14	11-Jan-21		
Week 15	18-Jan-21		
Week 16	25-Jan-21		
Week 17	01-Feb-21	8	Data plotting
Week 18	08-Feb-21		
Week 19	15-Feb-21	9	Version control systems
Week 20	22-Feb-21		
Week 21	01-Mar-21		Exam Prep
Week 22	08-Mar-21		

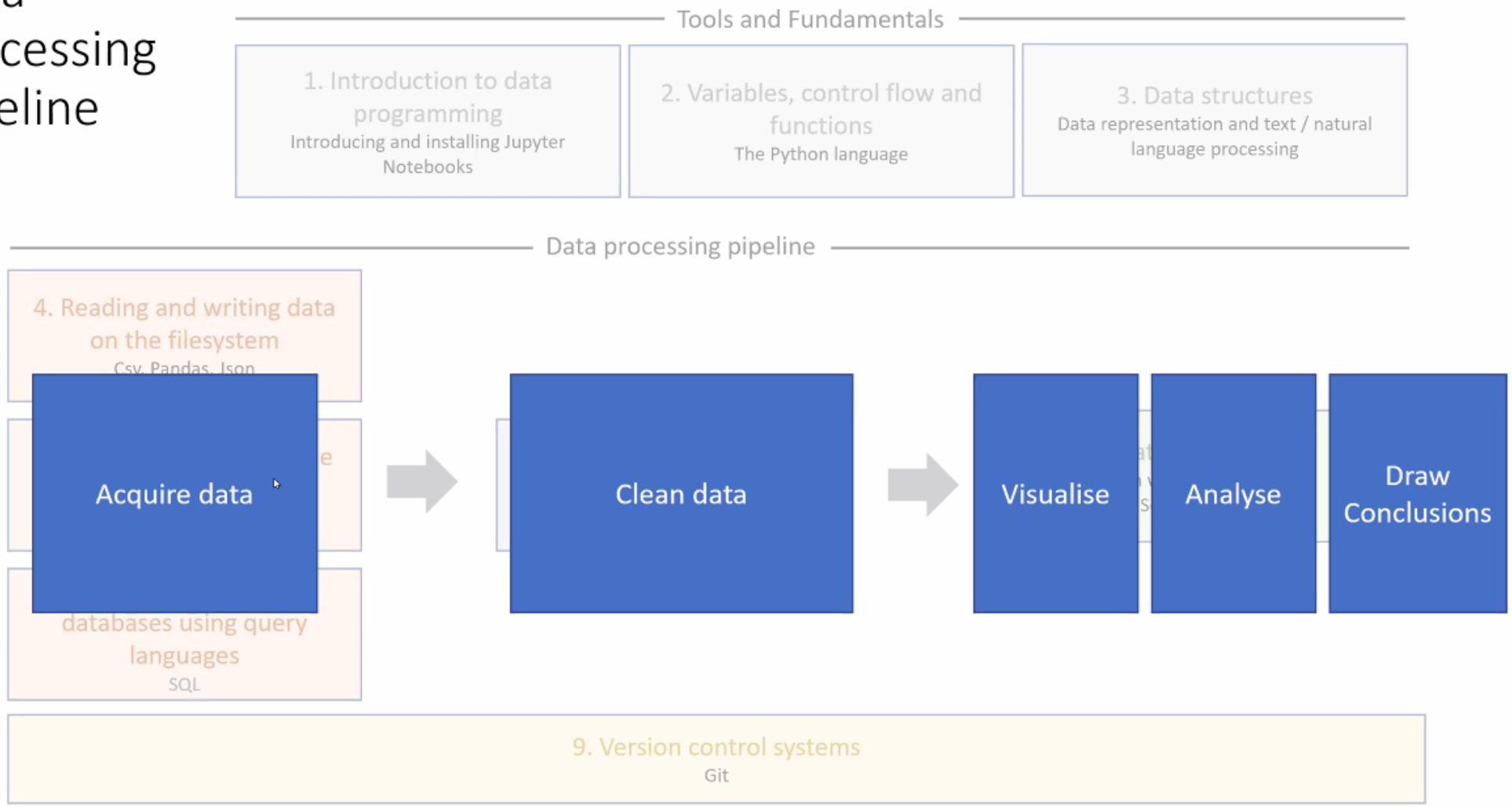
Peer review

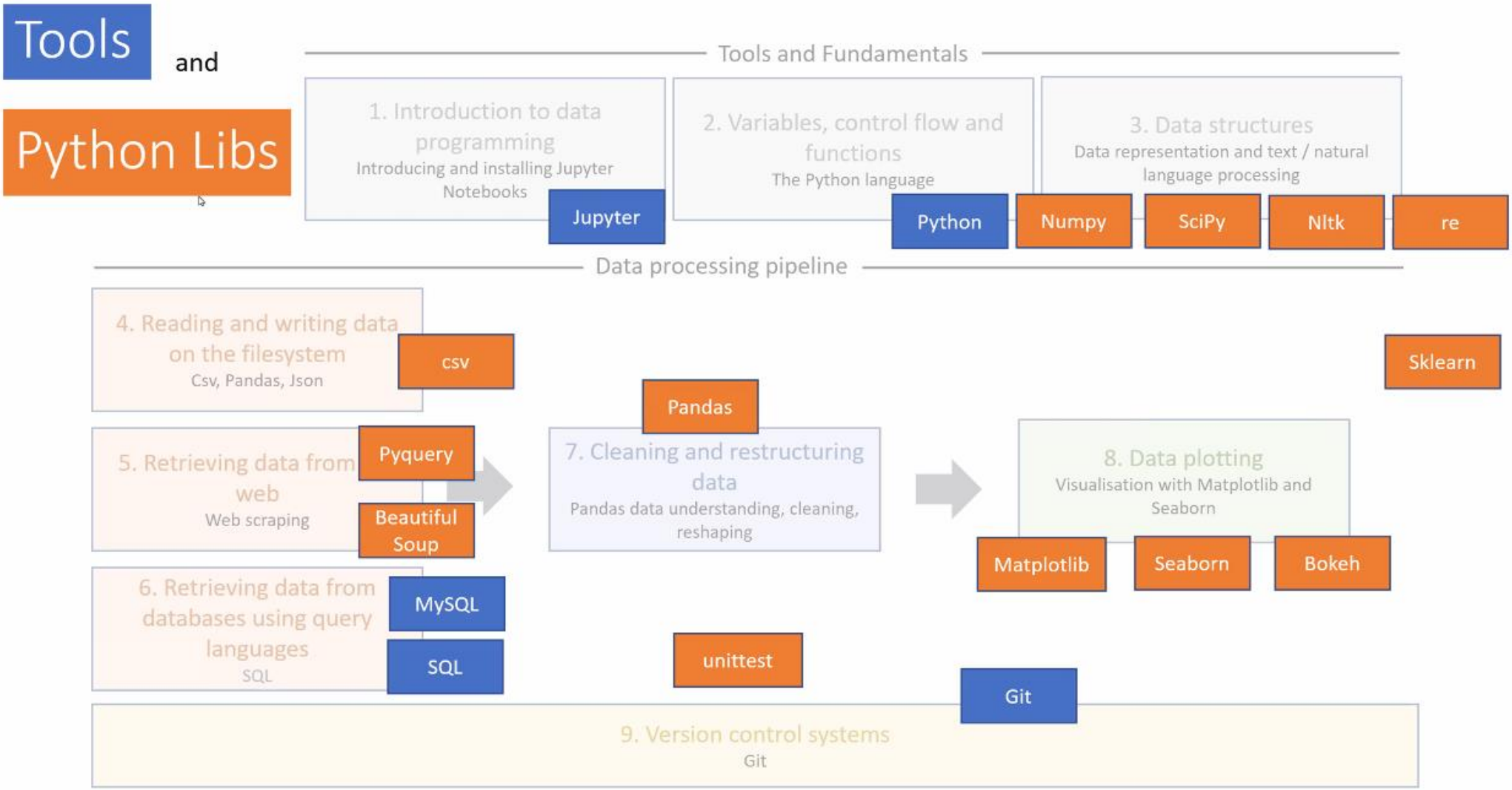
Peer review

Topics in Context



Data Processing Pipeline




Tools
and
Python Libs
Tools and Fundamentals
1. Introduction to data programming
Introducing and installing Jupyter Notebooks
Jupyter
2. Variables, control flow and functions
The Python language
Python
3. Data structures
Data representation and text / natural language processing
Numpy
SciPy
Nltk
re
Data processing pipeline
4. Reading and writing data on the filesystem
Csv, Pandas, Json
csv
Sklearn
Pandas
5. Retrieving data from web
Web scraping
Pyquery
Beautiful Soup
7. Cleaning and restructuring data
Pandas data understanding, cleaning, reshaping
8. Data plotting
Visualisation with Matplotlib and Seaborn
Matplotlib
Seaborn
Bokeh
6. Retrieving data from databases using query languages
SQL
MySQL
SQL
unittest
Git
9. Version control systems
Git

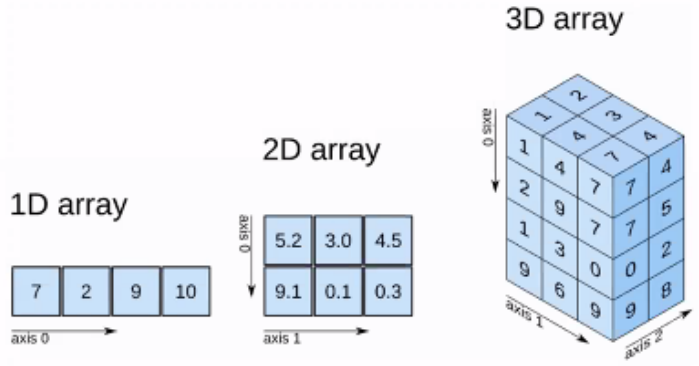
Tip

Get a good handle on Numpy and Pandas

Numpy

Arrays

Super-fast manipulation of arrays, including multi-dimensional arrays.
Much faster than Python lists with a huge range of functions for manipulation and analysis.



Pandas

Dataframes

Like an Excel spreadsheet, with named/indexed rows and columns.
Can do database-like operations on them, e.g. join, merge, etc.

	country	continent	year	lifeExp	pop	gdpPercap
0	Afghanistan	Asia	1952	28.801	8425333	779.445314
1	Afghanistan	Asia	1957	30.332	9240934	820.853030
2	Afghanistan	Asia	1962	31.997	10267083	853.100710
3	Afghanistan	Asia	1967	34.020	11537966	836.197138
4	Afghanistan	Asia	1972	36.088	13079460	739.981106

Skills learnt can be used in

- Data science
- Machine learning
- Image processing
- Computer vision
- Audio processing
- Scientific analysis
- …

Mid-Term Programming Assignment

Expect to use knowledge gained from topics 1-7 to work with a data set, acquire and clean the data and draw conclusions from that data.