

I Seminarium Python Geo Data Science Hel 18-22 marzec 2019



wine_data.txt Barendsburg.csv Alesund.csv Hopen.csv Hopen_NyAlesund_Barendsburg.csv iris_dane.csv midwest_filter.csv drinki.csv oecd_bli_2015.csv

```
42702 "NOE00134778", "NY ALESUND, NO", "2018-06-03", "0.0", "0.0", "1.8", "2.7", "0.4"
42703 "NOE00134778", "NY ALESUND, NO", "2018-06-04", "0.0", "0.0", "3.3", "4.9", "1.8"
42704 "NOE00134778", "NY ALESUND, NO", "2018-06-05", "0.0", "0.0", "2.8", "4.8", "1.9"
42705 "NOE00134778", "NY ALESUND, NO", "2018-06-06", "0.0", "0.0", "2.8", "4.1", "1.3"
42706 "NOE00134778", "NY ALESUND, NO", "2018-06-07", "0.0", "0.0", "2.3", "5.1", "-0.3"
42707 "NOE00134778", "NY ALESUND, NO", "2018-06-08", "0.0", "0.0", "-0.1", "0.8", "-0.7"
42708 "NOE00134778", "NY ALESUND, NO", "2018-06-09", "0.0", "0.0", "1.1", "2.2", "-0.9"
42709 "NOE00134778", "NY ALESUND, NO", "2018-06-10", "0.0", "0.0", "2.1", "2.6", "1.6"
```

Normal text file length : 4 893 718 lines : 72 183 Ln : 3 Col : 21 Sel : 0 | 0 Unix (LF) UTF-8 IN

C:\JACEK2\Hel19_ESRI\dane\Gdynia3_gps.csv - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?



wine_data.txt Barendsburg.csv Alesund.csv Gdynia3_gps.csv

```
1 ID, YY, XX
2 1, 54.4651031, 18.4819145
3 2, 54.4666366, 18.4797172
4 3, 54.4679451, 18.4777965
5 4, 54.4707565, 18.4738788
6 5, 54.4748802, 18.4719429
7 6, 54.4766655, 18.4722099
```

Normal text file

length : 5 888 908 lines : 200 001

```
wine_data.txt x Barenburg.csv x Alesund.csv x Gdynia3_gps.csv x PM_10_poland3.csv x
1 Code,Concentration,TimeSM,T,H%,Wind_dir,Wind_vel,UR200,WA200,GR200,
2 PL0496A,38.2,2015-12-31 23:00:00,-13.6,84.0,Wind blowing from the ea
3 PL0496A,107.599999999,2016-01-01 00:00:00,-14.5,84.0,Wind blowing fro
4 PL0496A,129.199999999,2016-01-01 01:00:00,-15.0,84.0,Wind blowing fro
5 PL0496A,49.0,2016-01-01 02:00:00,-15.1,85.0,Wind blowing from the ea
6 PL0496A,44.6,2016-01-01 03:00:00,-14.9,86.0,Wind blowing from the ea
7 PL0496A,32.6,2016-01-01 04:00:00,-14.6,84.0,Wind blowing from the ea
8 PL0496A,26.0,2016-01-01 05:00:00,-14.3,85.0,"Calm, no wind",0.0,62.9
9 PL0496A,29.6,2016-01-01 06:00:00,-12.2,87.0,Wind blowing from the ea
10 PL0496A,23.899999999,2016-01-01 07:00:00,-11.0,83.0,Wind blowing from
11 PL0496A,20.699999999,2016-01-01 08:00:00,-10.3,81.0,Wind blowing from
12 PL0496A,18.899999999,2016-01-01 09:00:00,-9.5,79.0,Wind blowing from
13 PL0496A,19.8,2016-01-01 10:00:00,-8.8,75.0,Wind blowing from the eas
14 PL0496A,16.000000000,2016-01-01 11:00:00,-8.6,74.0,Wind blowing from
Normal text file length: 548 480 137 lines: 2 203 479 Ln: 1 Col: 1 Sel: 0 | 0 Windows (CR LF) UTF-8 INS
```

Dane tekstowe .csv

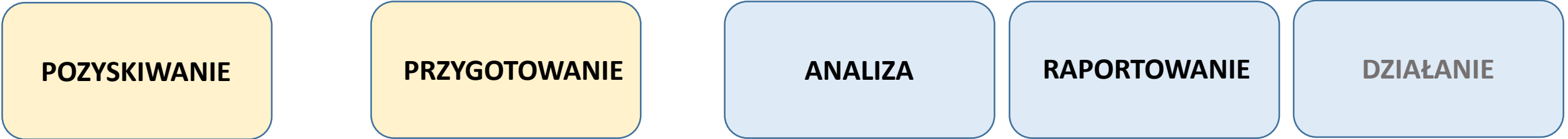


Zdjęcia satelitarne

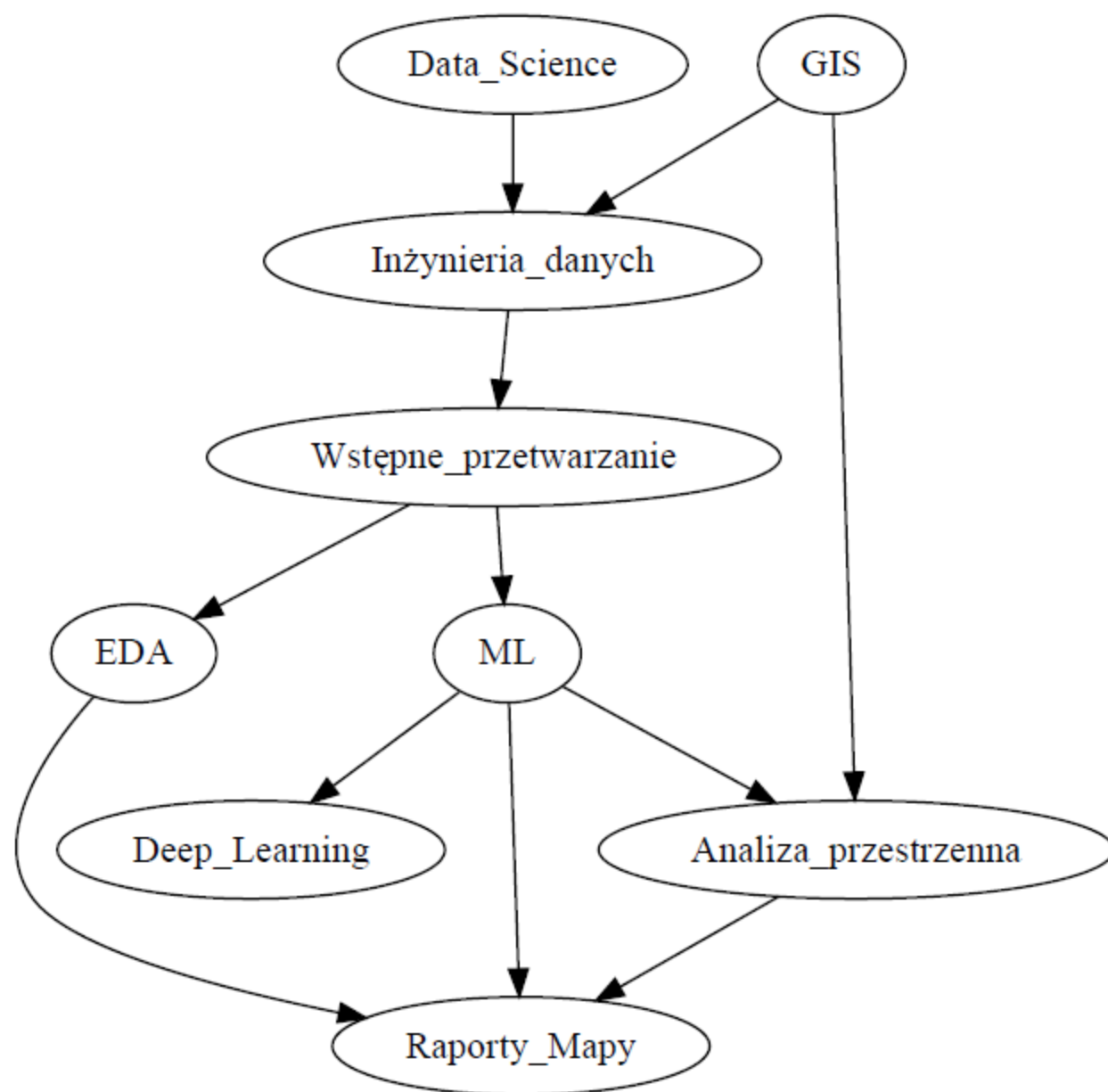
DATA SCIENCE

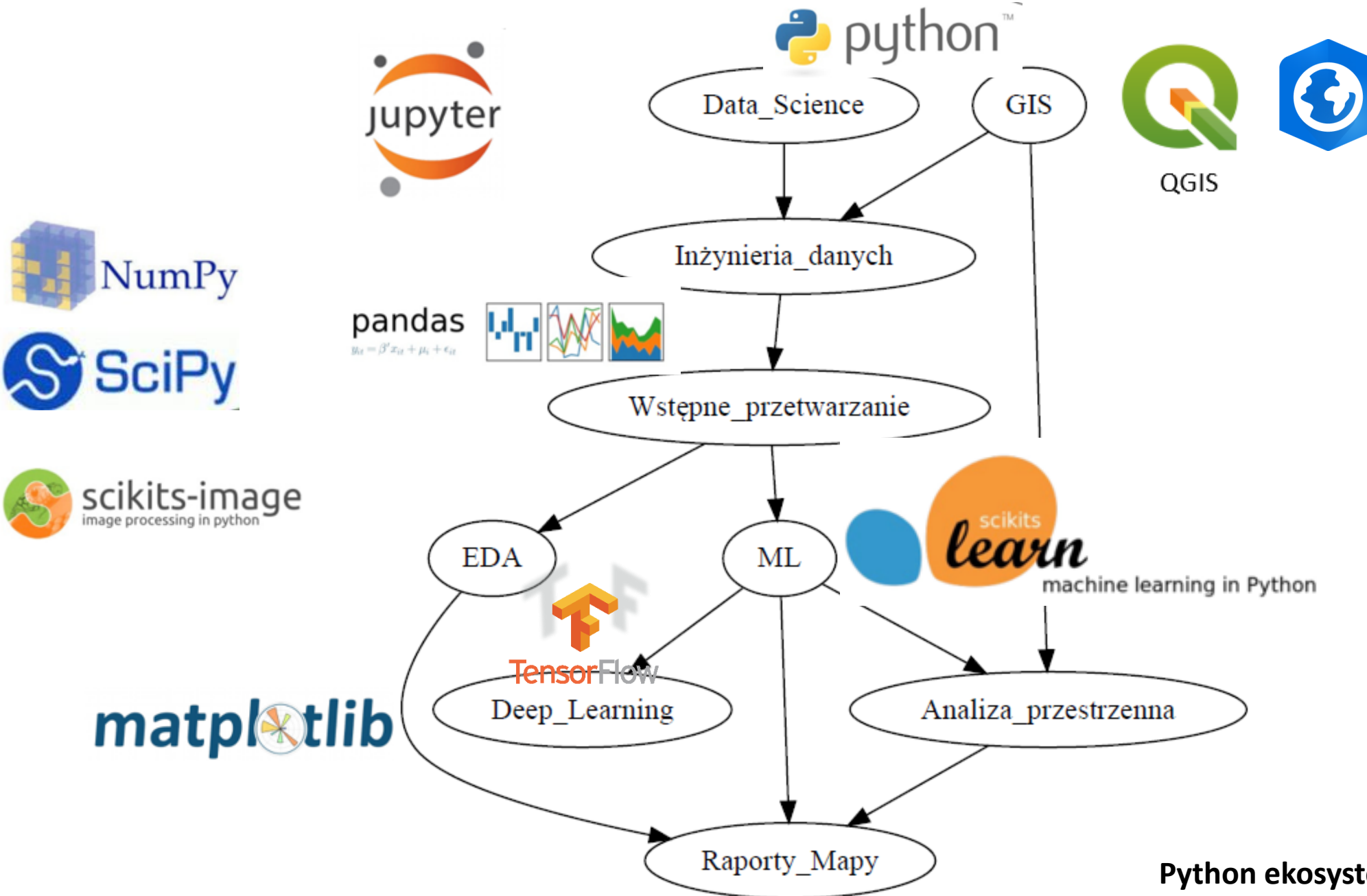
DATA ENGINEERING

COMPUTATIONAL DATA SCIENCE



	Eksploracja EDA	PREPROCESSING	
PLIKI TEKSTOWE			Classification
BAZY DANYCH			Regression
Web serwisy			Clustering
No SQL storage			Graph Analytics
	Korelacja	Czyszczenie
	Trend	Transformacja	
	Outliers	Duplicate	
	Statystyki	Inconsistent	
		Missing	
	Wizualizacja	Outlier	
	Histogramy		
	Boxplots		
	Line graph		





Python ekosystem (środowisko) dla (geo)data science

[Pull requests](#) [Issues](#) [Marketplace](#) [Explore](#)[urbanskigis](#) / [HEL-geodata-science](#)[Watch](#)

0

[★ Star](#)

0

[Fork](#)

0

[Code](#)[Issues](#) 0[Pull requests](#) 0[Projects](#) 0[Wiki](#)[Insights](#)[Settings](#)

Seminarium geoscience na Helu

[Edit](#)[Manage topics](#)


25 commits

1 branch

0 releases

1 contributor

Branch: master

[New pull request](#)[Create new file](#)[Upload files](#)[Find File](#)[Clone or download](#) urbanskigis Delete Wstęp do jupyter-kurs.ipynb

Latest commit 31d514b a minute ago

 dane


Add files via upload

3 days ago

 plus

Delete pp2.txt

3 days ago

 Data Science start.pdf

Add files via upload

3 days ago

 LaTeX1.ipynb

Add files via upload

3 days ago

 Matplotlib_podstawy.ipynb

Add files via upload

3 minutes ago

 Numpy - algebra liniowa.ipynb

Add files via upload

3 minutes ago

 Podstawowe operacje Numpy.ipynb


Add files via upload

3 minutes ago

 README.md


Update README.md

11 days ago

 Wstęp do jupyter.ipynb

Add files via upload

3 minutes ago

 README.md

HEL-geodata-science

Seminarium geodata science na Helu 18-22 marca 2019

Struktura danych - Tidy data

Tidy data is the data obtained as a result of a process called data tidying. It is one of the important cleaning processes during big data processing and is a recognized step in the practice of [data science](#). Tidy data sets have structure and working with them is easy; they're easy to manipulate, model and visualize. Tidy [data sets](#) are arranged such that each variable is a column and each observation (or case) is a row.^{[1][2]}

Tidy data provide standards and concepts for [data cleaning](#), and with tidy data there's no need to start from scratch and reinvent new methods for data cleaning.

Jeff Leek in his book *The Elements of Data Analytic Style* summarizes the characteristics of tidy data as the points:^[3]

1. Each variable you measure should be in one column.
2. Each different observation of that variable should be in a different row.
3. There should be one table for each "kind" of variable.
4. If you have multiple tables, they should include a column in the table that allows them to be linked.

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	2059360
Brazil	1999	31737	17206362
Brazil	2000	84488	174504898
China	1999	212258	1272015272
China	2000	213766	1280428583

variables

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	2059360
Brazil	1999	31737	17206362
Brazil	2000	84488	174504898
China	1999	212258	1272015272
China	2000	213766	1280428583

observations

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	2059360
Brazil	1999	31737	17206362
Brazil	2000	84488	174504898
China	1999	212258	1272015272
China	2000	213766	1280428583

values

<https://towardsdatascience.com/>

<https://dataelixir.com/newsletters/>

<https://statquest.org/video-index/>