I Seminarium Python Geodata Science

Hel 18-22 marzec 2019

GISCO: GEOGRAPHICAL INFORMATION AND MAPS

Overview
GEODATA
GISCO activities
Frequently asked questions (FAQ)

GEODATA

General information on geo-datafiles

All datasets are provided with their metadata in XML format. Geographic information system (GIS) software is needed to process the downloaded geodata.

Geographical datasets are provided in two formats: **Shapefile** and '**Personal File Geodatabase**'.

Shapefiles have a limitation in the attribute name length. For this reason, the attribute name might be truncated when compared with the name in the metadata or in the personal file geodatabase.

wine_data.txt | Barensburg.csv | Alesund.csv | Hopen.csv | Hopen_NyAlesund_Barensburg.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

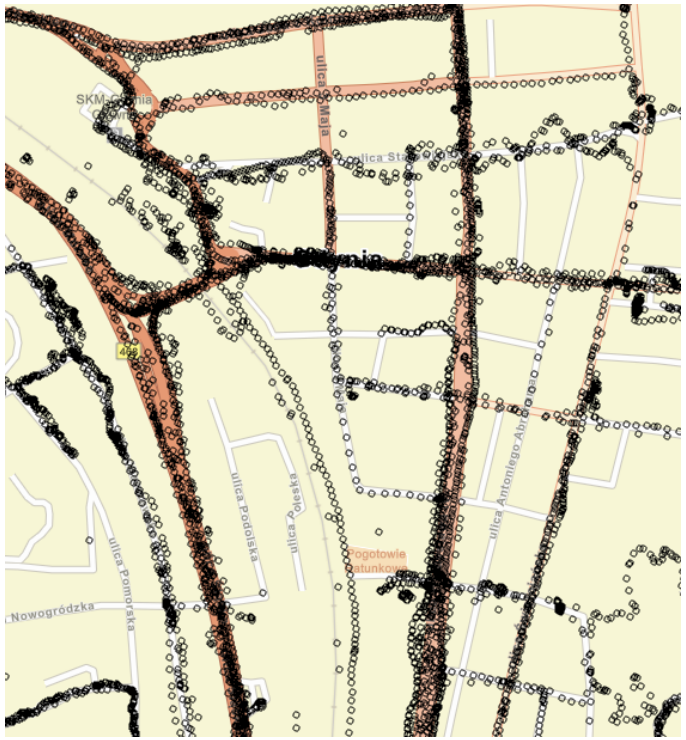
C:\JACEK2\Hel19_ESRI\dane\Gdynia3_gps.csv - Notepad++

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

wine_data.txt | Barensburg.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
```

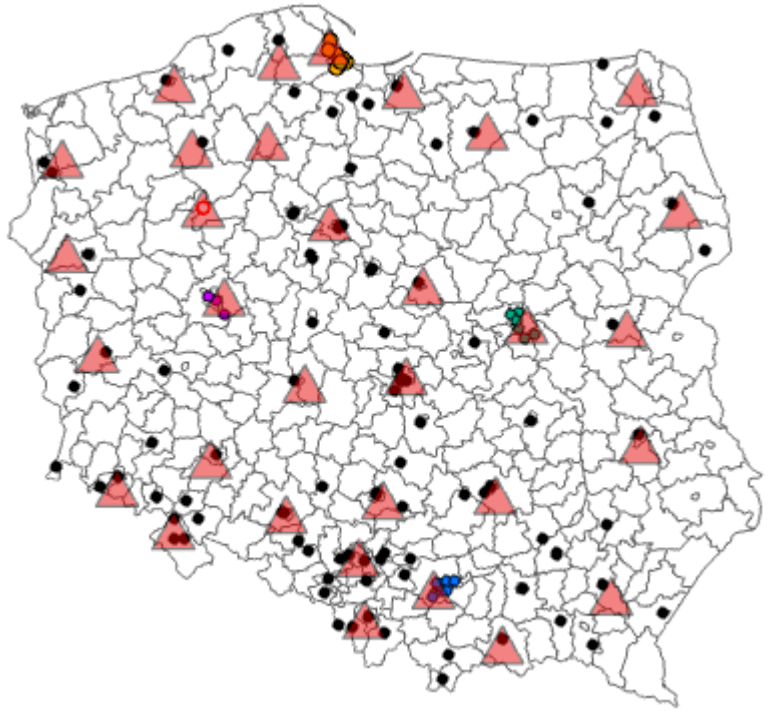
Normal text file | length : 5 888 903 | lines : 200 001



wine_data.txt | Barensburg.csv | Alesund.csv | Gdynia3_gps.csv | PM_10_poland3.csv

```
Code,Concentration,TimeSM,T,H%,Wind_dir,Wind_vel,UR200,WA200,GR200,
PL0496A,38.2,2015-12-31 23:00:00,-13.6,84.0,Wind blowing from the ea
PL0496A,107.59999999,2016-01-01 00:00:00,-14.5,84.0,Wind blowing fro
PL0496A,129.19999999,2016-01-01 01:00:00,-15.0,84.0,Wind blowing fro
PL0496A,49.0,2016-01-01 02:00:00,-15.1,85.0,Wind blowing from the ea
PL0496A,44.6,2016-01-01 03:00:00,-14.9,86.0,Wind blowing from the ea
PL0496A,32.6,2016-01-01 04:00:00,-14.6,84.0,Wind blowing from the ea
PL0496A,26.0,2016-01-01 05:00:00,-14.3,85.0,"Calm, no wind",0.0,62.
PL0496A,29.6,2016-01-01 06:00:00,-12.2,87.0
PL0496A,23.89999999,2016-01-01 07:00:00,-11
PL0496A,20.69999999,2016-01-01 08:00:00,-10
PL0496A,18.89999999,2016-01-01 09:00:00,-9.
PL0496A,19.8,2016-01-01 10:00:00,-8.8,75.0,
```

Normal text file | length : 548 480 137 | lines : 2 203 479 | Ln : 1 | Col : 1 | Sel



Dane tekstowe .csv

Zdjęcia satelitarne

DATA SCIENCE

DATA ENGINEERING

COMPUTATIONAL DATA SCIENCE

POZYSKIWANIE | **PRZYGOTOWANIE** | **ANALIZA** | **RAPORTOWANIE** | **DZIAŁANIE**

PLIKI TEKSTOWE

BAZY DANYCH

Web serwisy

No SQL storage

Eksploracja EDA

Korelacja
Trend
Outliers
Statystyki

Wizualizacja

Histogramy
Boxplots
Line graph

PREPROCESSING

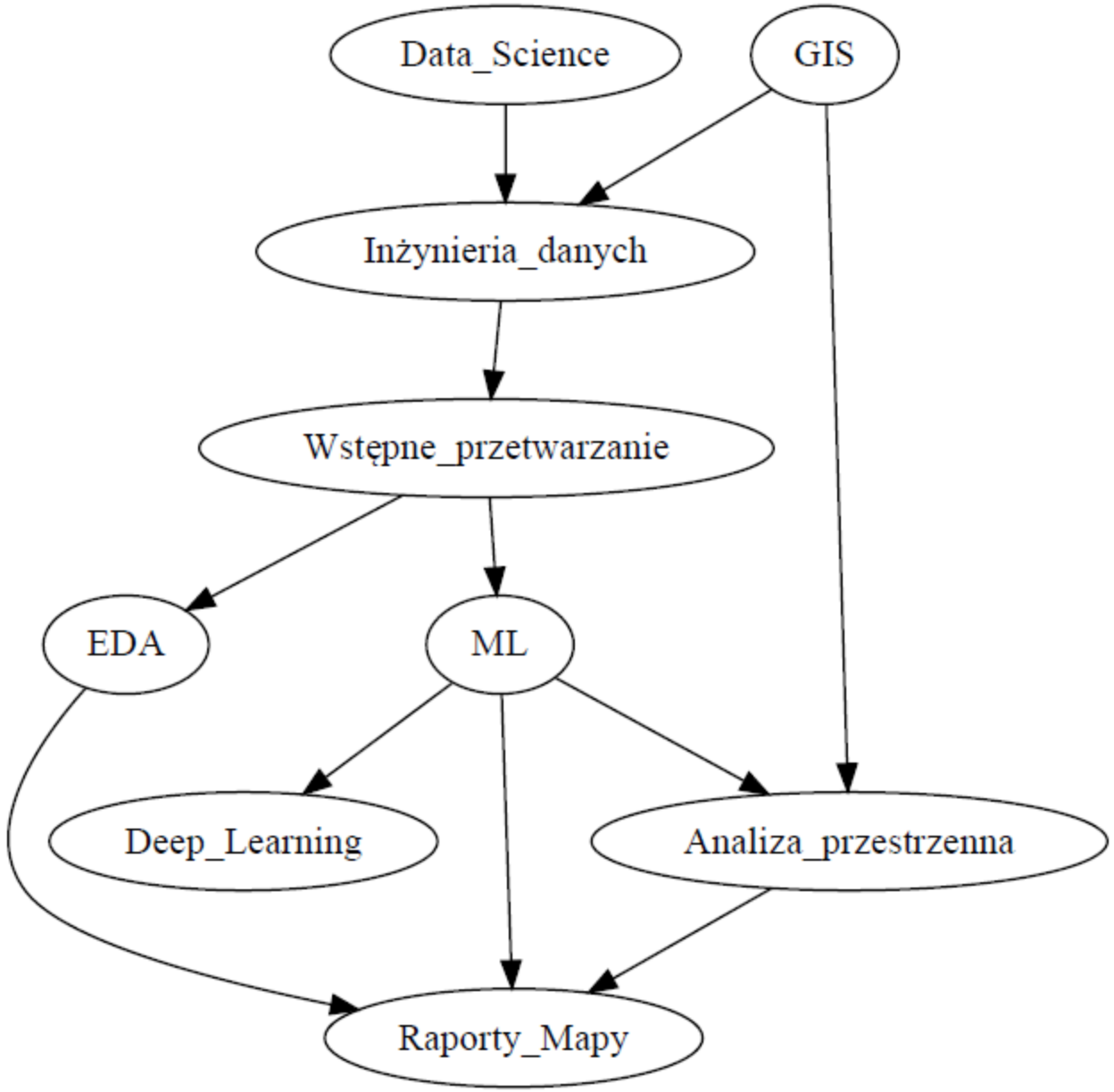
Czyszczenie

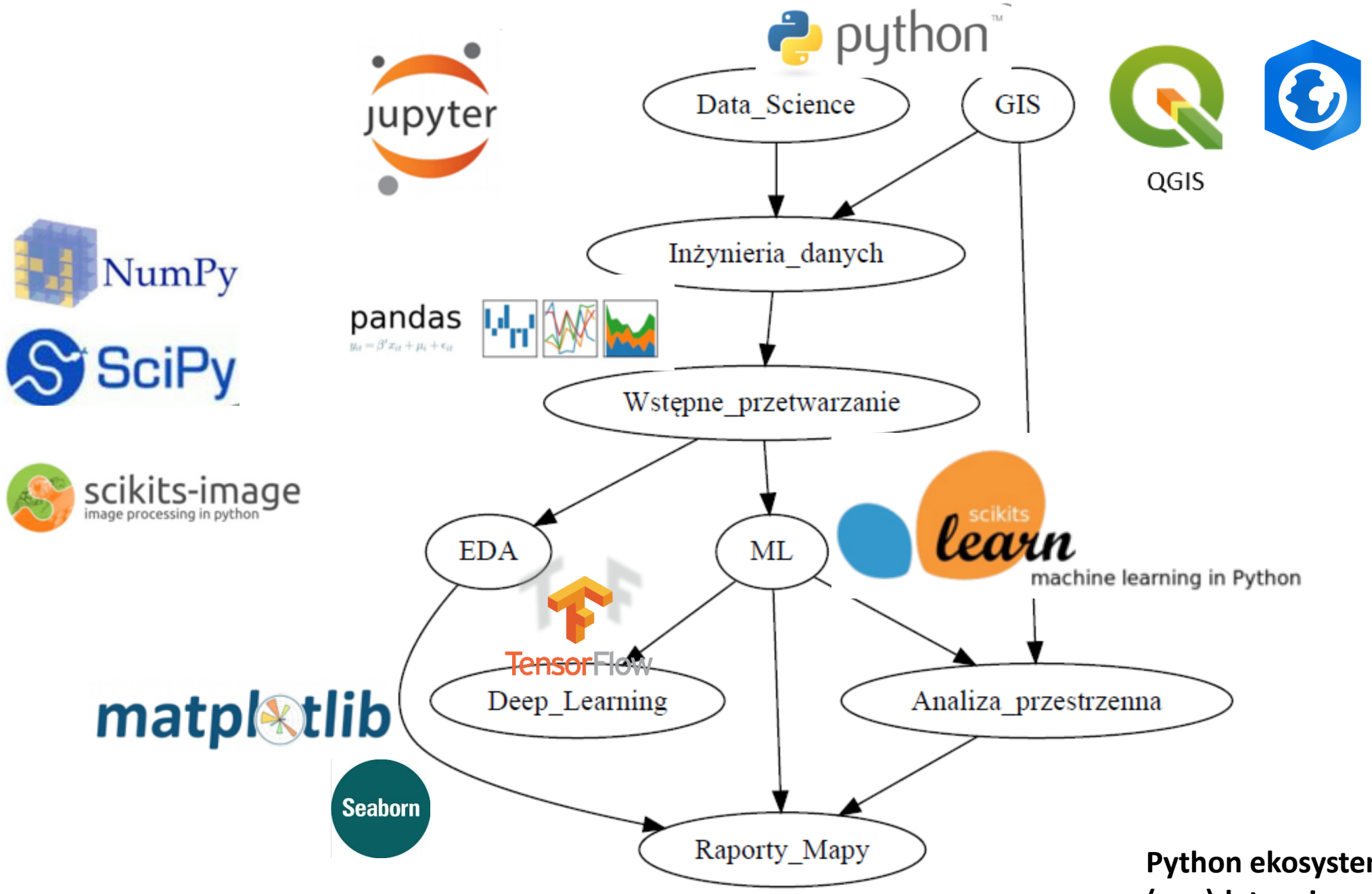
Duplicate
Inconsistent
Missing
Outlier

Transformacja

Scailing
Transformation
Feature selection
Dimensions PCA

Classification
Regression
Clustering
Graph Analytics
……..


Data_Science
GIS
Inżynieria_danych
Wstępne_przetwarzanie
EDA
ML
Deep_Learning
Analiza_przestrzenna
Raporty_Mapy

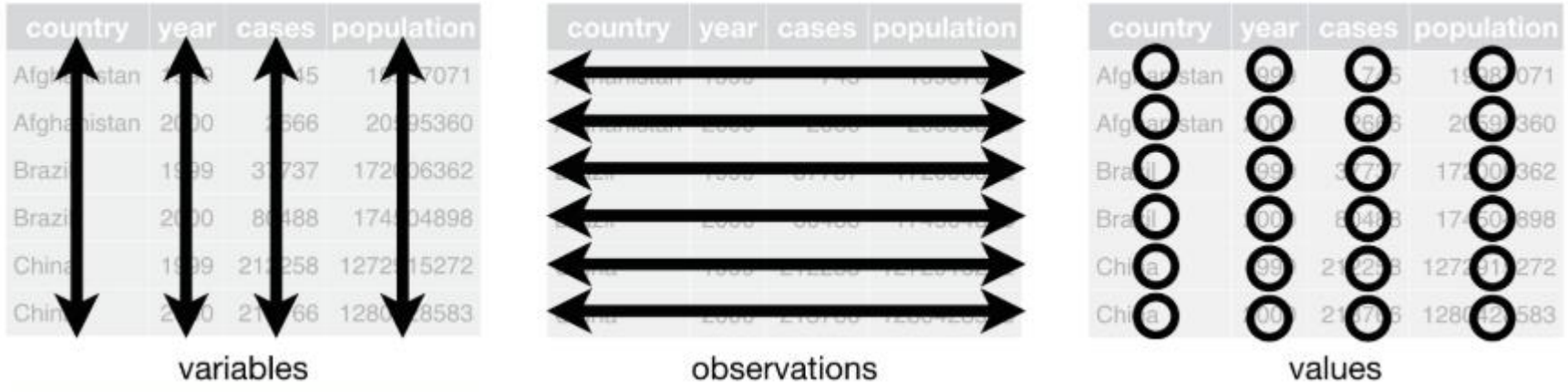


Python ekosystem (środowisko) dla (geo)data science

Struktura danych - **Tidy data (uprządkowane dane)**

Tidy data dane otrzymane w rezultacie procesu nazywanego data tidying. Jest to pierwszy podstawowwy krok poprzedzający **data cleaning** i jest istotnym krokiem w praktyce **data science.** Takie (tidy) zbiory danych mają określoną strukturę dzięki której praca z nimi jest prosta, są proste do przetwarzania, modelowania i wizualizacji. Dzięki temu można do nich stosować wiele generycznych algorytmów pozwalających na ich przetwarzanie. Tidy **data sets** są zorganizowane w taki sposób, że zmienne (variable, atrybuty, features) tworzą kolumny a każda obserwacja (observation, *case, wektor*) jest rzędem. Tidy data tworzą standardy i koncepcje do **data cleaning**, pozwala to na wykorzystanie standardowych metod czyszczenia danych.

1. Każda zmienna mierzona powinna tworzyć jedną kolumnę.
2. Każda jej obserwacja powinna być w innym rzędzie.
3. Powinna być jedna tablica dla każdego rodzaju zmiennych.
4. Jeżeli jest wiele tablic, powinny zawierać kolumnę, która pozwoli na ich łączenie.



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

https://towardsdatascience.com/

https://dataelixir.com/newsletters/

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