Data Science: spojrzenie praktyczne na dane oraz podejmowany problem

Adam Pacławski, PhD
e-mail: adam.paclawski@uj.edu.pl

Cel oraz plan zajęć

- Zapoznanie z podstawowymi pojęciami z zakresu Data Science
- Czym jest Data Science oraz Data Scientist
- Dane – czyli co?
- Źródła danych: ogólne, n. farmaceutyczne
- Zarządzanie danymi
- Zdefiniowanie problemu - rozpuszczalność sl.
- Przygotowanie baz danych do późniejszych etapów pracy: skalowanie, normalizacja
- Transformacja regresja -> klasyfikacja

Zacznijmy od... przygotowania środowiska

Nowe środowisko wirtualne:

```
conda --version
conda create -n DataScience python=3.6
conda activate DataScience
conda install -c conda-forge scikit-learn
conda install numpy
conda install pandas
conda install matplotlib
conda install -c conda-forge jupyterlab
conda install -c mordred-descriptor/label/dev mordred
```

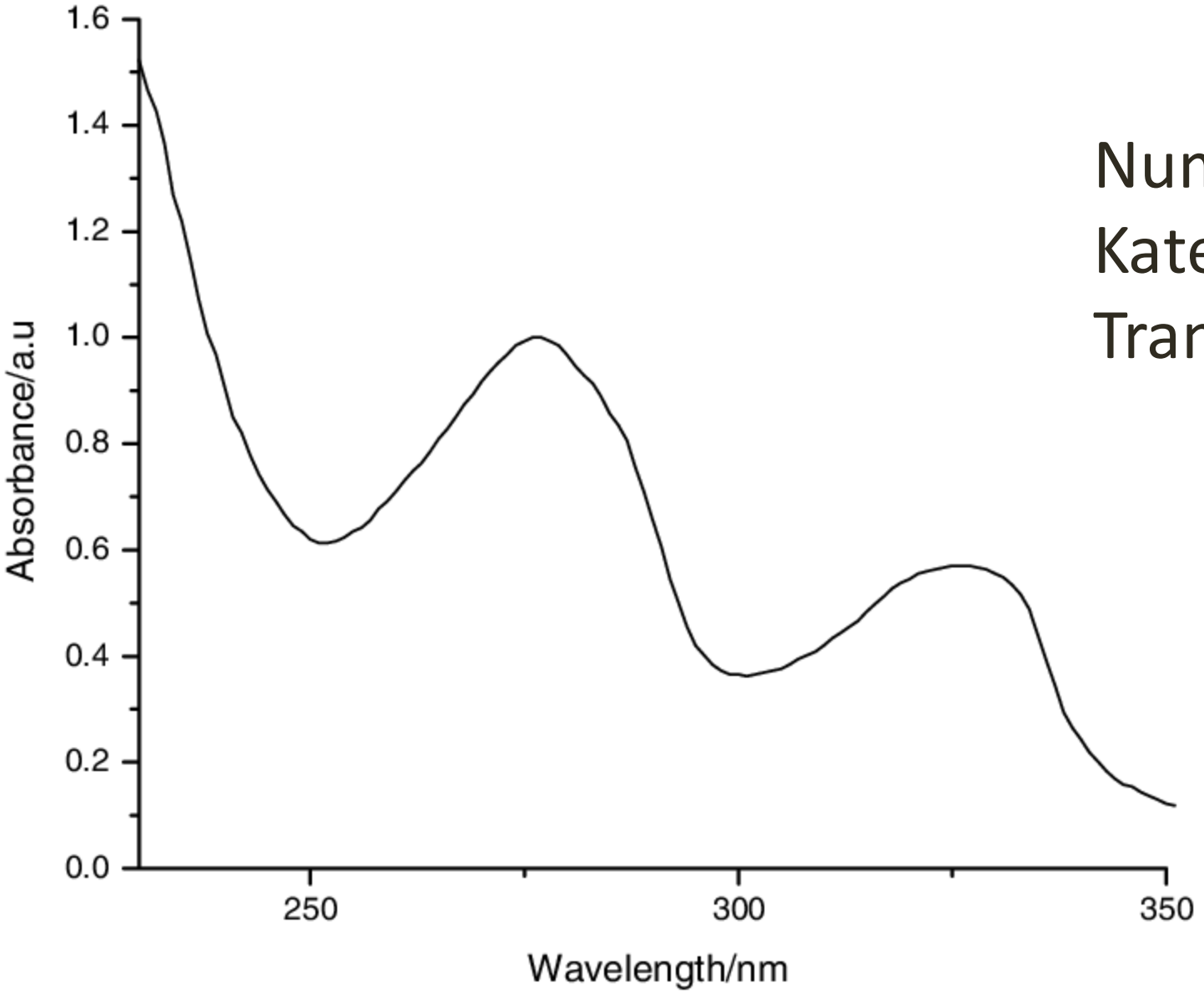
W niedalekiej przyszłości:

```
conda install -c conda-forge keras
```

Dane – czyli co?

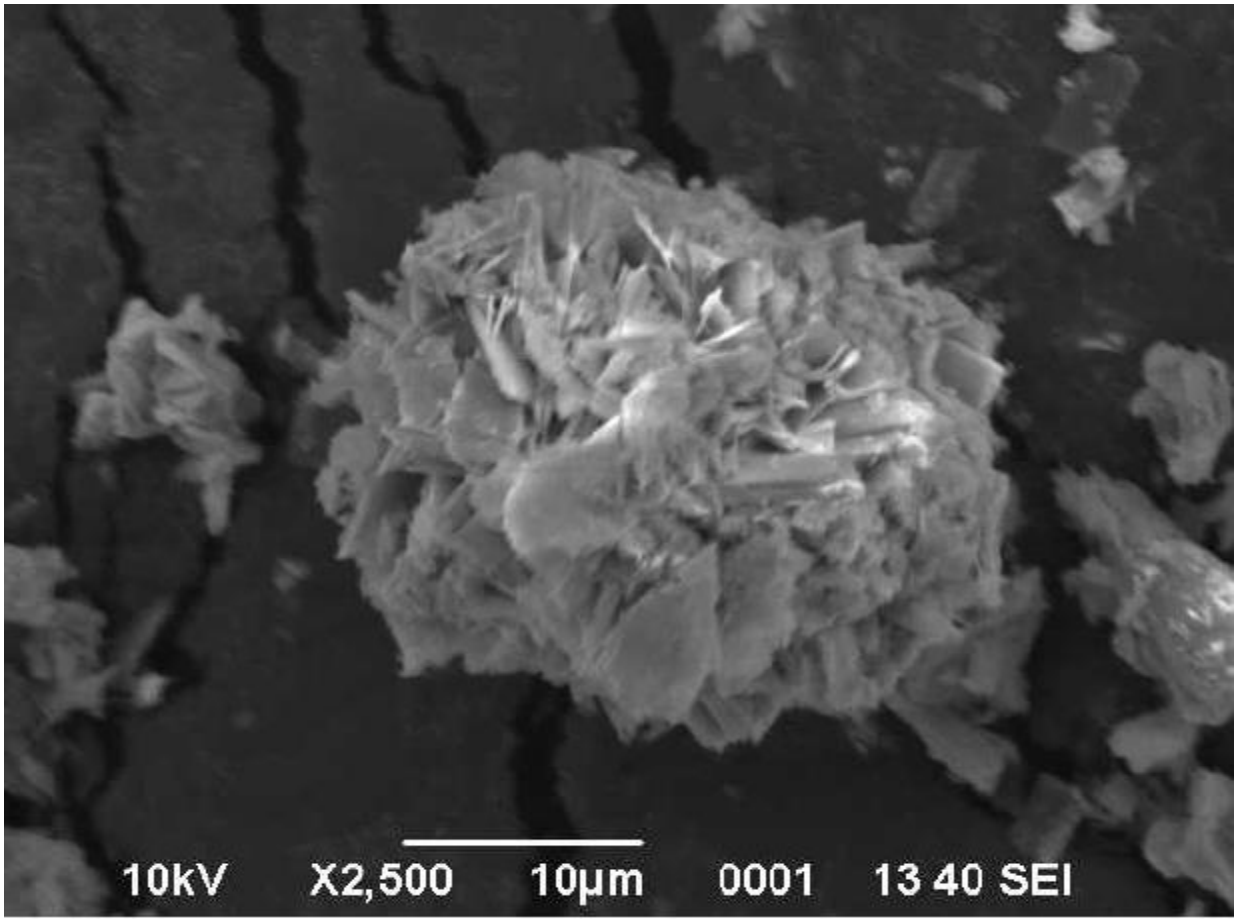
23.92572454	2.723722699	0	39.3513
18.96143814	0.911538135	18.23341436	39.3178
2.460356433	0.934373273	14.78773236	41.8636
6.671030211	1.387419415	16.14821464	41.8631
5.120731267	1.116554157	15.84881728	41.8626
0.569626668	0.708347924	13.63825524	41.8041
10.87166408	0.776468168	17.1663343	41.8038
10.53354234	2.441569931	16.85482801	41.8034
0	1.379761014	14.60317921	41.803
7.801037005	2.326071152	16.55407403	41.8025
14.76488891	3.087514214	16.8034056	41.8019
10.30362906	3.315200301	16.01791988	41.7435
8.30897104	0.671031356	14.82390775	41.7432
6.503045869	2.350649564	16.1276946	41.7428
2.910840667	2.304303988	15.97757088	41.7423
6.641688755	1.480053042	14.45093102	41.7418

Dane czyli co?

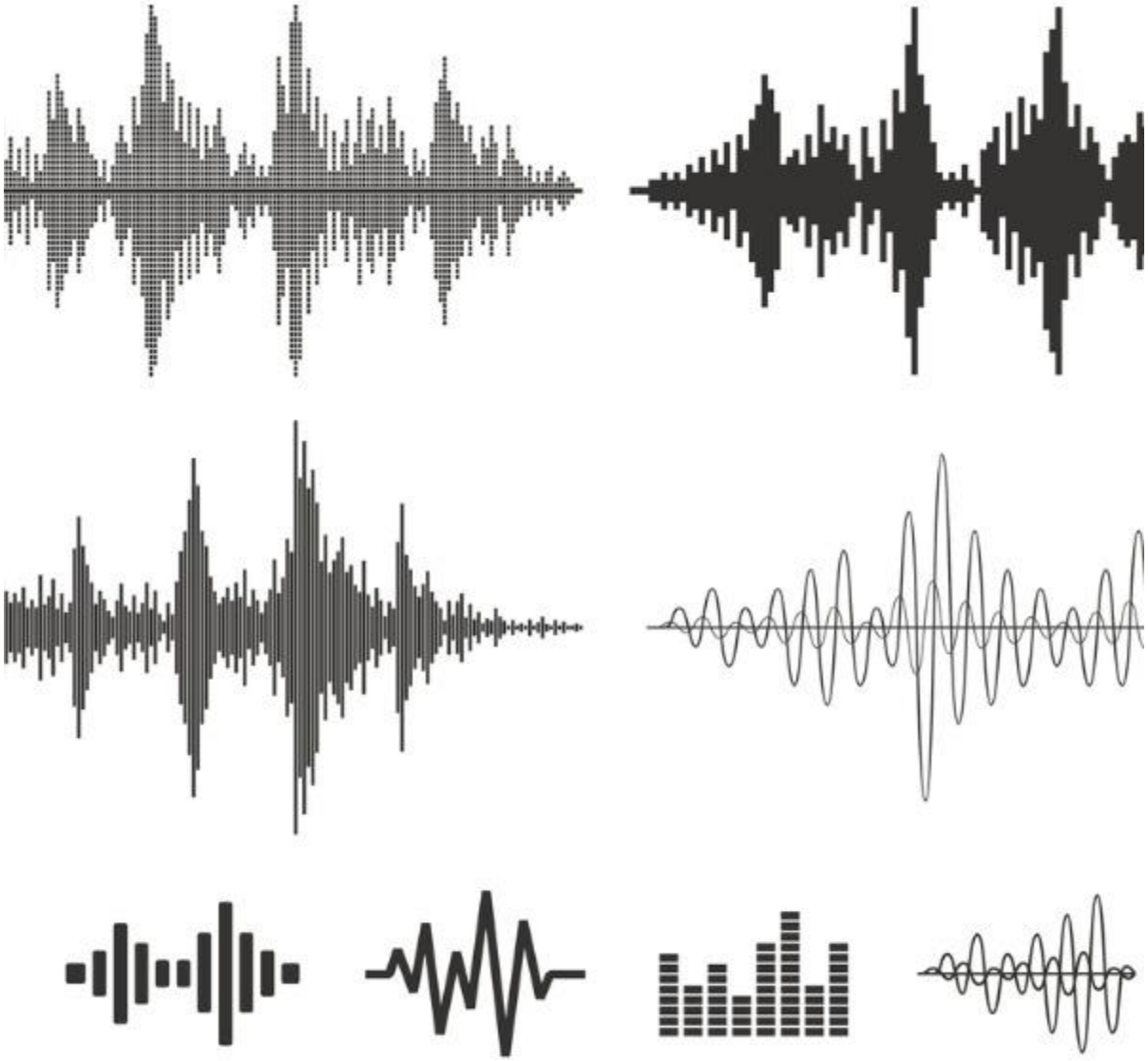


Numeryczne?
Kategoryczne?
Transformacja?

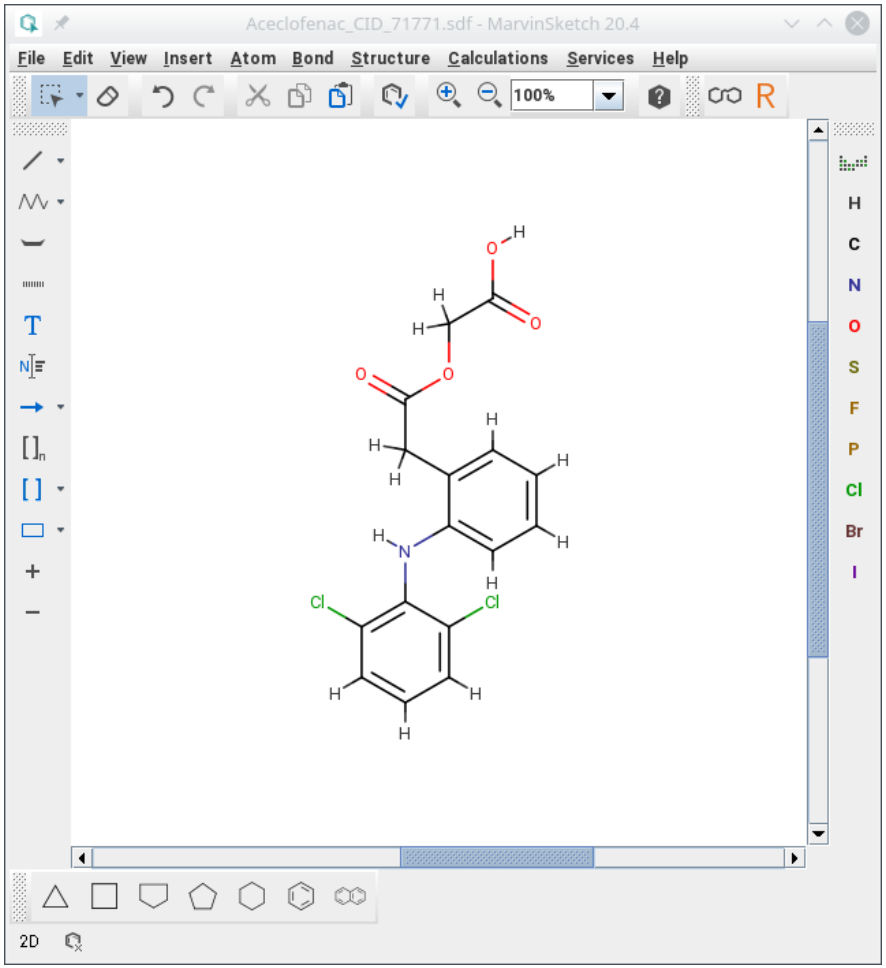
Dane czyli co?



Dane czyli co?



Dane czyli co?



Źródła danych

- Dane własne!
- Poszukiwanie danych w dostępnych źródłach
- Gotowe bazy danych
- Ogólne:

https://github.com/awesomedata/awesome-public-datasets

Kaggle:
https://www.kaggle.com/datasets

Czym jest Data Science?

- Podejście naukowe do analizy danych, systematyczna analiza, zastosowanie metodyki naukowej
- Dlaczego jest to istotne?
- Kim jest Data Scientist? Ekspertem w danej dziedzinie? Programistą? Statystykiem?

Zdefiniujmy problem

- Rozpuszczalność sl. w wodzie!
- Dlaczego?
- Co wiemy o lekach: LADME
- Dlaczego w wodzie?
- Czy istnieją jakieś rozwiązania?

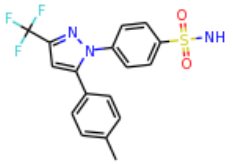
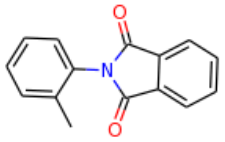
- ChemAxon
https://disco.chemaxon.com/calculators/demo/plugins/solubility/
- ALOGPS 2.1
http://www.vcclab.org/lab/alogps/

Bazy danych w farmacji

- OCHEM
- https://ochem.eu/home/show.do
- Dane oraz platforma do budowy modeli
- QSAR () Np.. Odniżenie ciśnienia, Aktywnosc p/wirusowa (SARS-CoV2),
- QSPR () np. Rozpuszczalność w wodzie!

Bazy danych w farmacji

- ZINC
- https://zinc.docking.org
- Dane o oaktywności biologicznej

9.15 (0.49)	50	ZINC2570895 Celebrex 
9.00 (0.70)	1	ZINC50522 

Proces B&R nowego leku

- Synteza
- Badania lab. W zakresie właściwości fiz-chem i biologicznych
- Badania na zwierzętach
- Badania na ludziach
- Wprowadzenie do terapii (Wycofanie b. kosztowne)

- Koszt wzrasta z etapem zaawansowania prac B&R

Materiały do zajęć

- Plik z deskryptorami generowanymi z zast. pakietu mordred

https://drive.google.com/file/d/1Kkep6FD24Z3ttyM8qIY6ESqjWjDeqLe4/view

- Repozytorium GitHub

https://github.com/adamPaclawski/AGH_DataScience_2021