Deep Layers

WORKSHOP / ARTIFICIAL INTELLIGENCE

20. - 21. 9 . 2022

Institute of Scientific Instruments of the Czech Academy of Sciences
Kralovopolska 147, Brno, Czech Republic

Registration (free, but mandatory) | www.isibrno.cz/deep

Deep Layers
a machine-learning forum

STRATEGIE AV21

brno.ai

AI MT

Průlomové technologie budoucnosti

ISI CAS

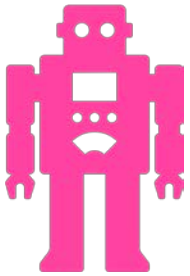
What is the AI (today)?

Filip Plesinger

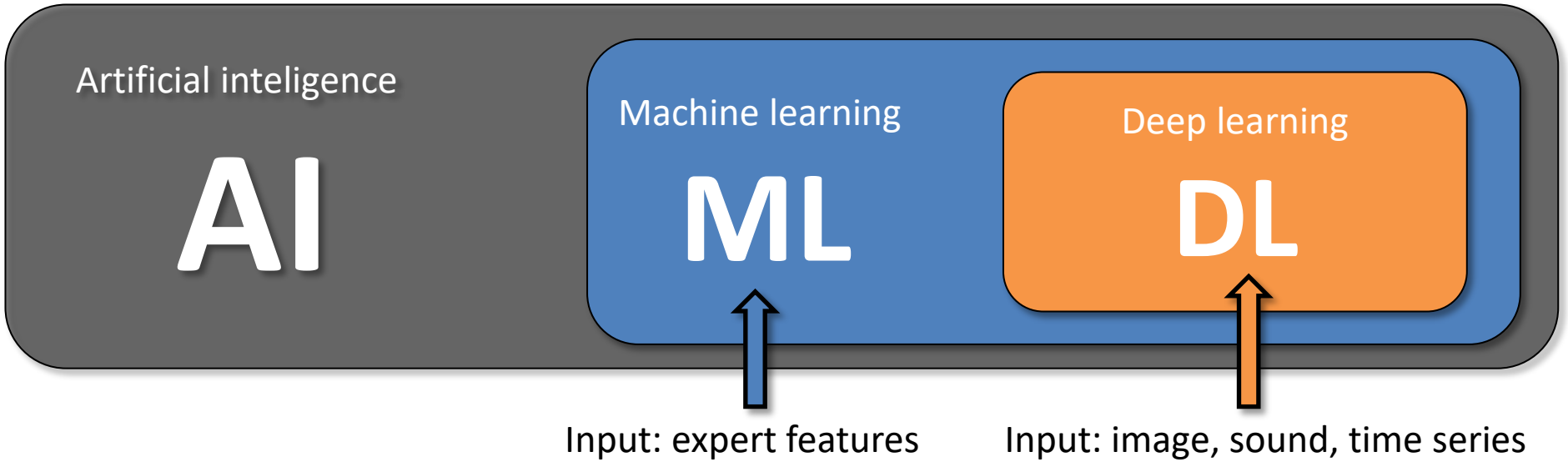
AIMT, ISI of the CAS, Brno, CZ

What is the AI (today)?

AI (Artificial Intelligence)



- AI imitates **human cognitive abilities**
- AI = machine learning / deep learning
- AI methods produce **a computational model**

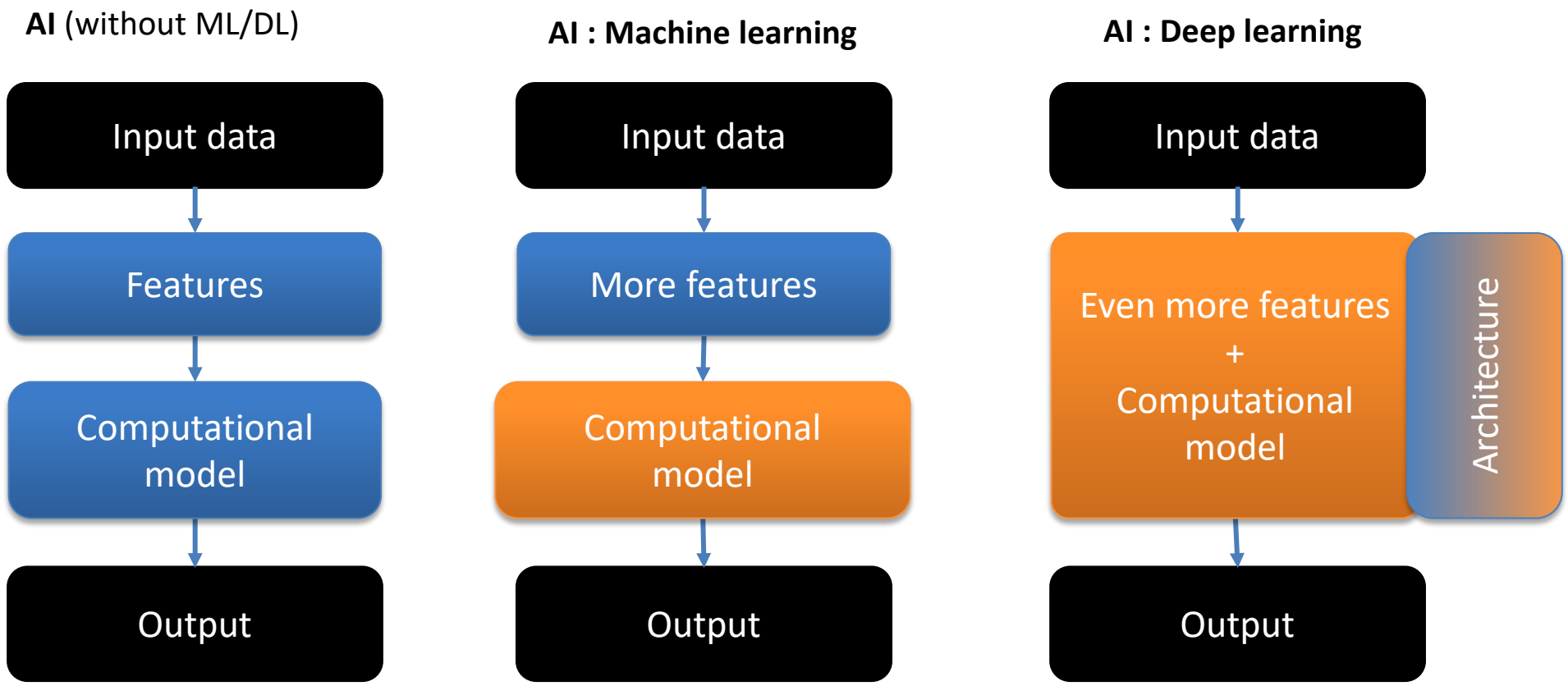


What is the AI (today)?

AI (Artificial Intelligence)

Designed by man

Designed by Machine



We have to perfectly know what is the input and expected output. If not ….

AI (Artificial Intelligence)

If the task is vaguely defined…

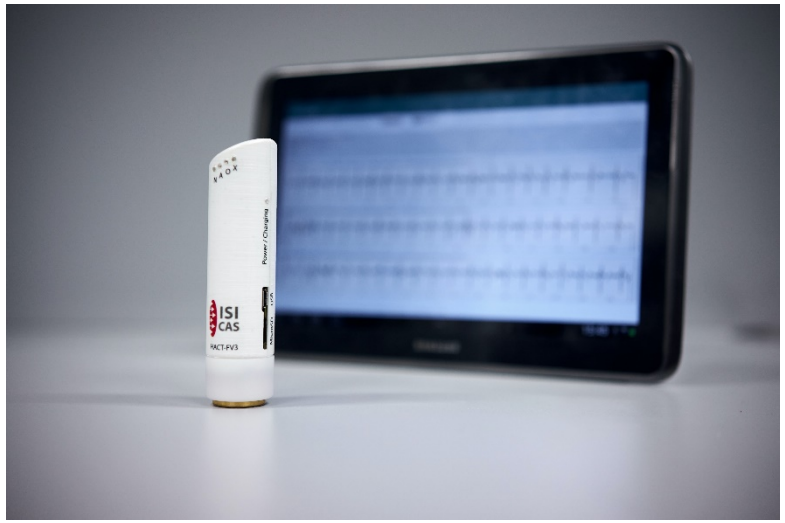
... we could receive the same answer as these guys.

https://www.imdb.com/title/tt1113230/mediaviewer/rm935773953/?ref_=tt_md_3

What is the AI (today)?

What is the use of AI?

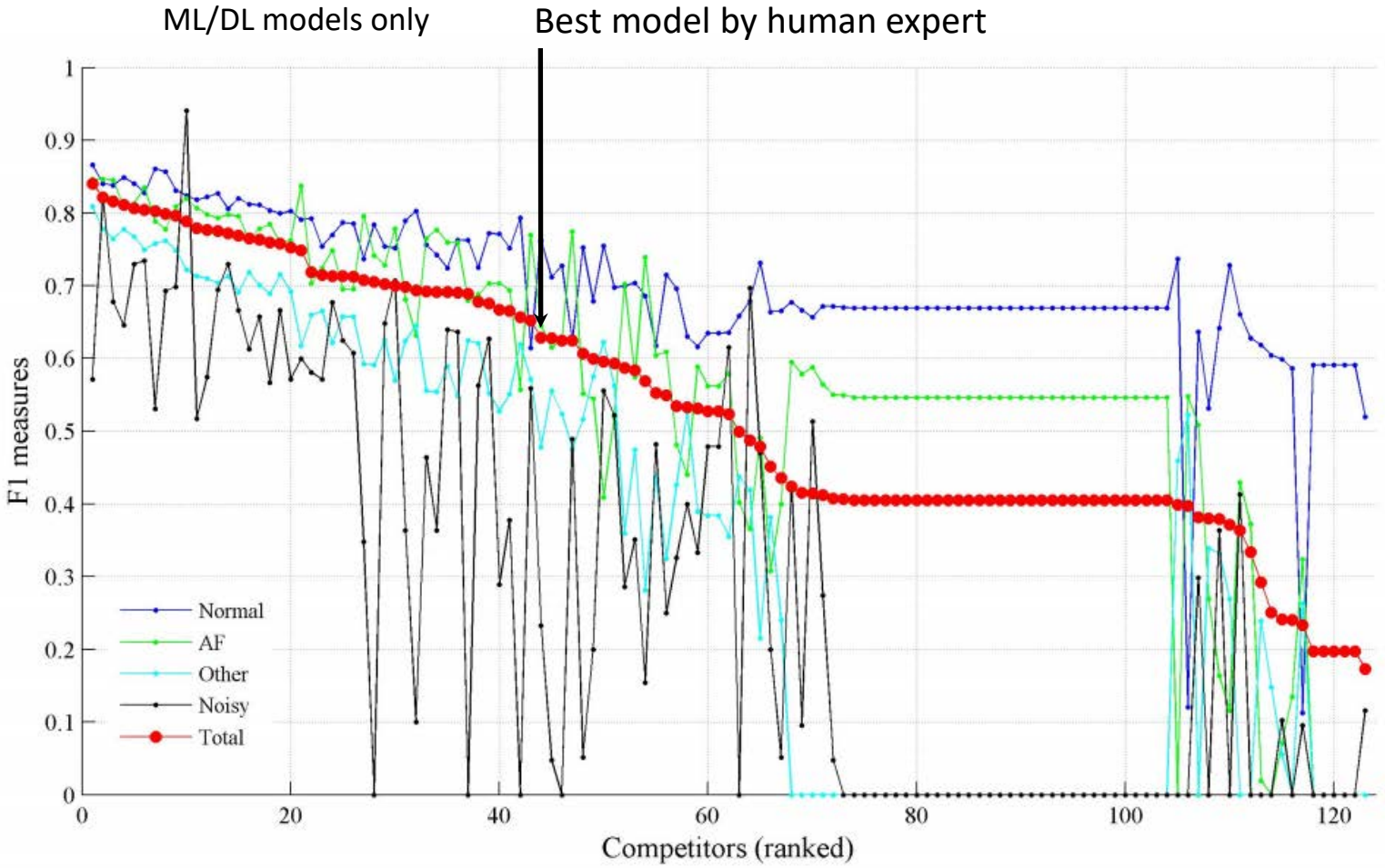
- **Classification tasks** (what is the pathology in this ECG signal?)
- **Regression tasks** (how long I will live?)
- Clustering (can we further separate our recordings by any clue?)
- Extending our own knowledge



Hand-held arrhythmia detector (AIMT-ISI of the CAS)
Supported by project TG03010046

What is the AI (today)?

Is the AI better than human expert?



PhysioNet Challenge 2017
Official round results

ECG classification into 4 classes

What is the AI (today)?

Is the AI better than human expert?

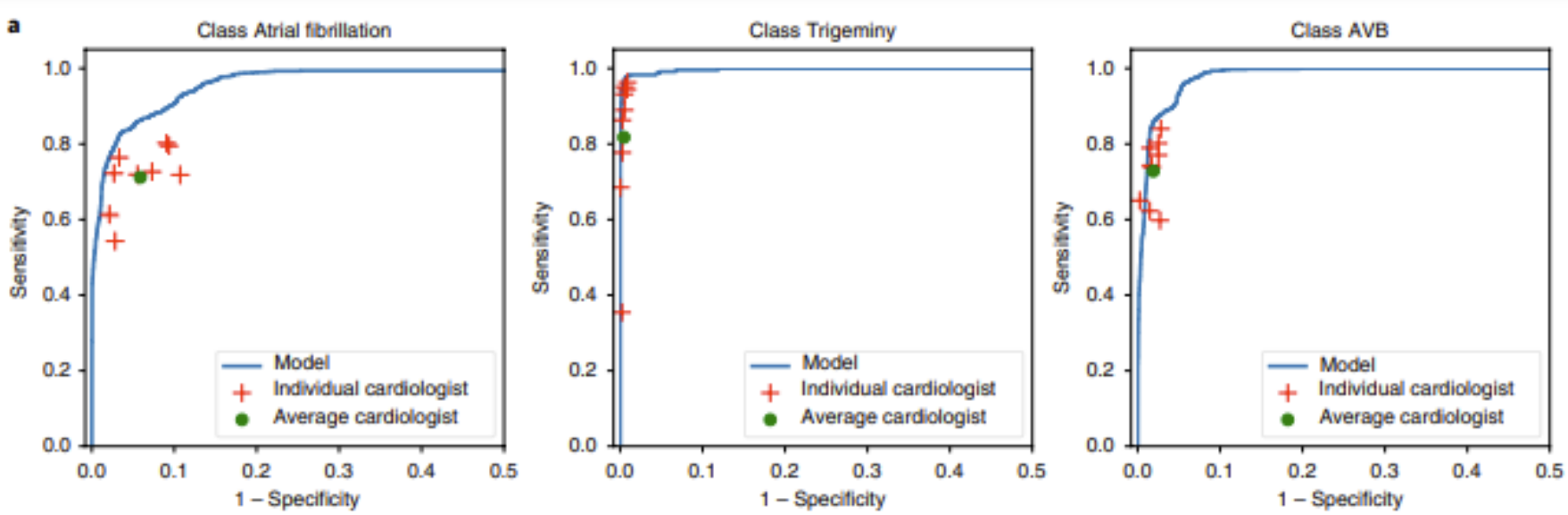
nature medicine

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun *, Pranav Rajpurkar , Masoumeh Haghpanahi, Geoffrey H. Tison , Codie Bourn, Mintu P. Turakhia and Andrew Y. Ng

Patients count:

53 549

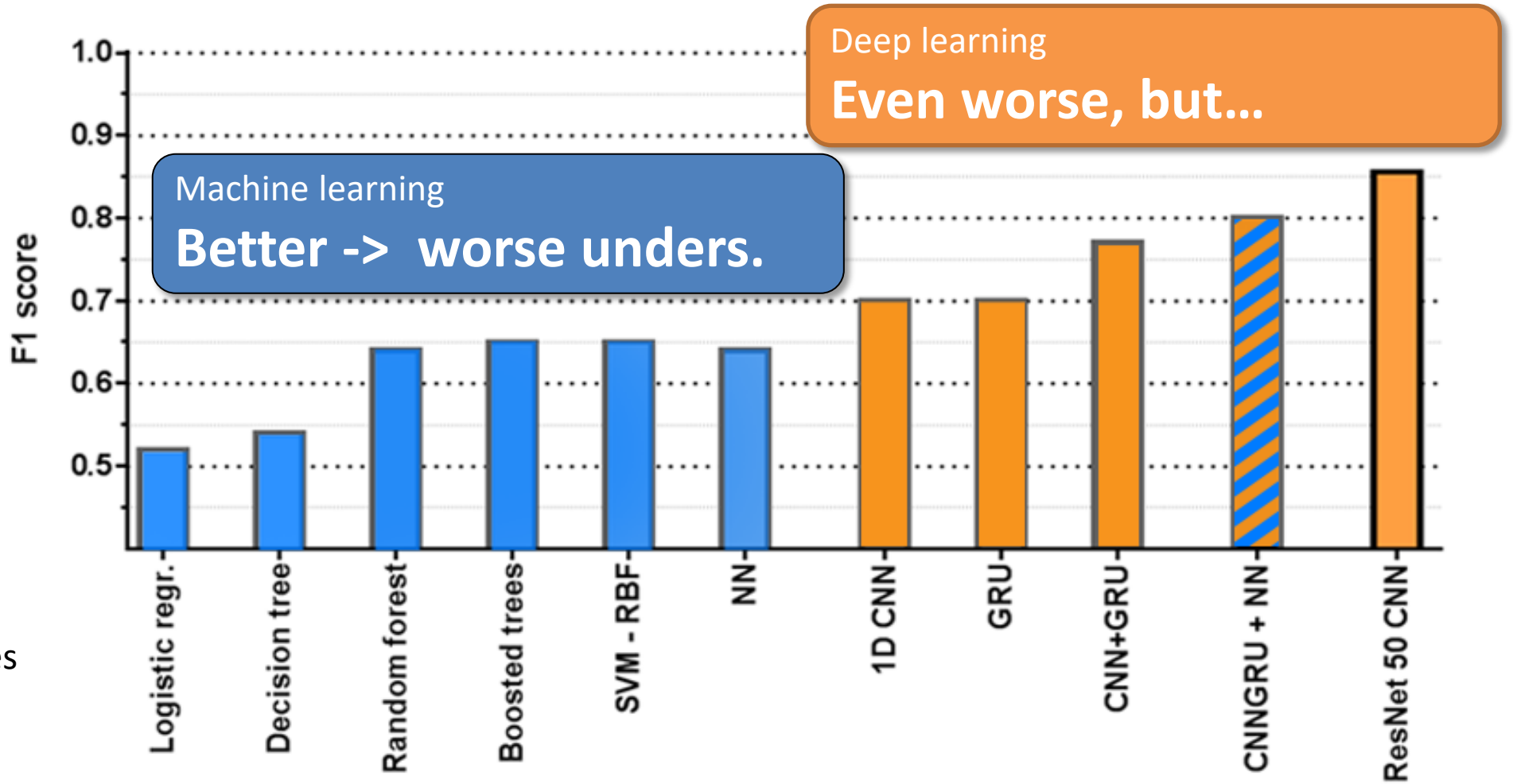


What is the AI (today)?

Can we see under the AI hood?

What is the AI (today)?

Understanding of AI depends on model type/complexity



$>10^{2\text{-}3}$ samples

$>10^{4,5\ldots}$ samples

What is the AI (today)?

… it can be explained in DL methods – Att. mechanism

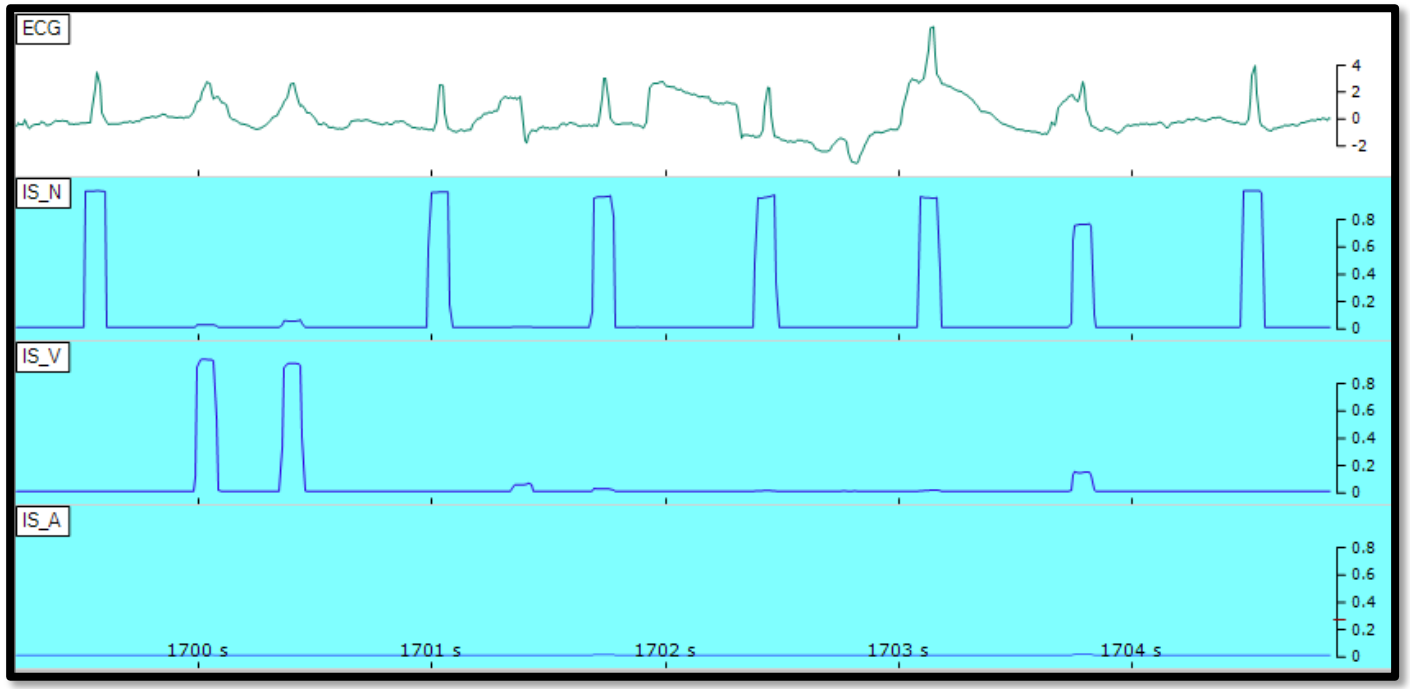
Baltruschat, I.M., Nickisch, H., Grass, M. *et al.* Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Sci Rep* **9,** 6381 (2019). https://doi.org/10.1038/s41598-019-42294-8

What is the AI (today)?

And what about you?

Who uses AI for:

- image processing?
- video processing?
- volumetric data?
- sound/speech analysis?
- other data analysis?



A. Ivora *et al.*, "QRS detection and classification in Holter ECG data in one inference step," *Sci. Reports* |, vol. 12, p. 12641, 2022.

(transition to the next lesson)

Machine Learning Elements I – Data treatment

How to explore, understand, clean and prepare your data for ML

Filip Plesinger

ISI of the CAS, Brno, CZ

You are welcome to experiment with dataset during the lesson.

QR code link to COLAB NOTEBOOK :
(Or through https://www.isibrno.cz/deep/)

1. Understanding our data & our task

Link to jup. notebook

Our dataset is about … **… music!**

We will **predict** song **popularity** from its features (Spotify data – years 1957-2020)

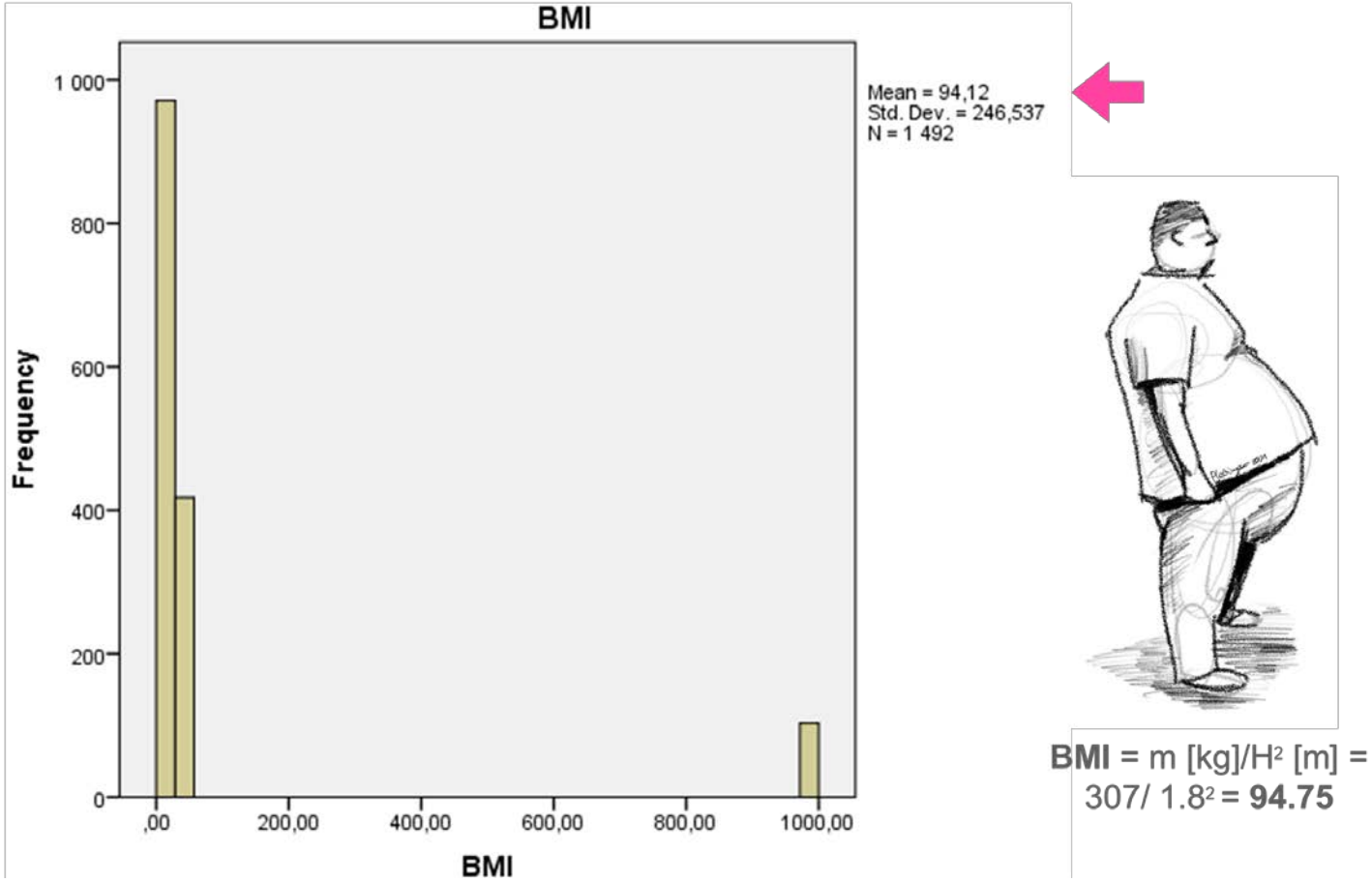
1. Exploring our dataset

- What features we have?
- How many **samples** we have?
- What is the **output**?
- What is the meaning of all **features?** What is **their** range?
- What is **their** distribution?
- How many **missing values** are there?
- How many **unique values** each feature has?
- What is the **type of each feature?** (categorical/ordinal/continuous)
- Are features correlated?

2. Cleaning the dataset

Removing missing values & checking distributions

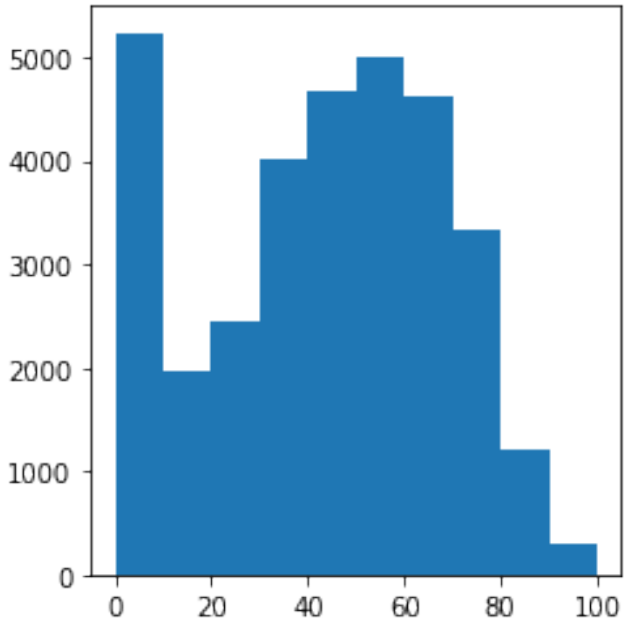
	Hospital	Gender	Age	BMI
1	A	Female	51	26,73
2	A	Male	52	24,98
3	B	Female	57	27,28
4	C	Female	47	23,59
5	C	Female	56	28,06
6	A	Female	61	20,76
7	A	Female	60	27,04
8	B	Female	62	27,17
9	C	Female	69	31,89
10	B	Female	58	25,02
11	C	Female	73	27,41
12	C	Female	69	27,78
13	B	Female	43	26,83
14	C	Female	76	26,78
15	B	Female	77	27,45
16	C	Female	80	25,08

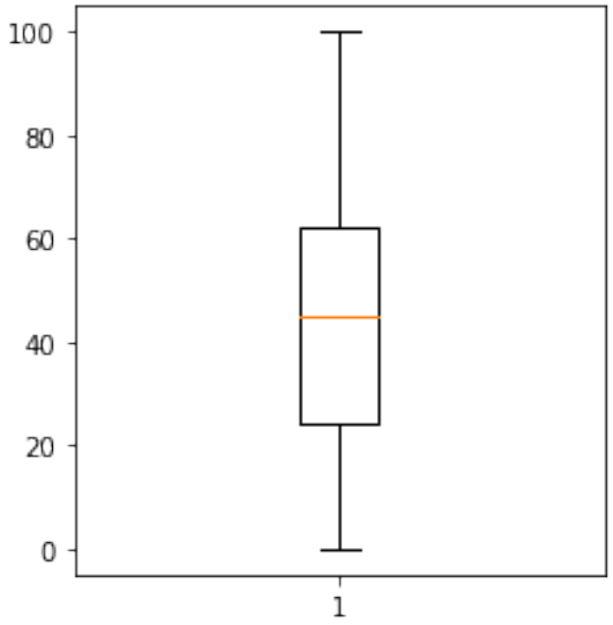


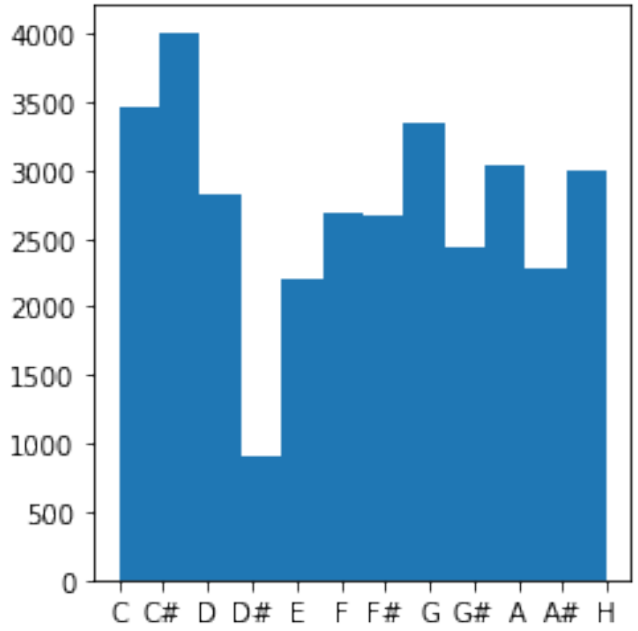
BMI = m [kg]/H² [m] = 307/ 1.8² = **94.75**

2. Distribution & feature meaning (1)

Histograms and boxplots are usefull. Do not rely on boxplots alone.



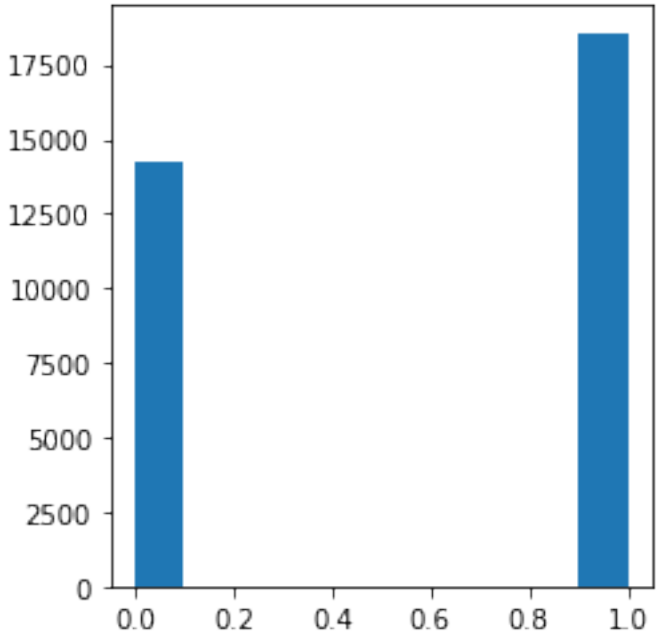


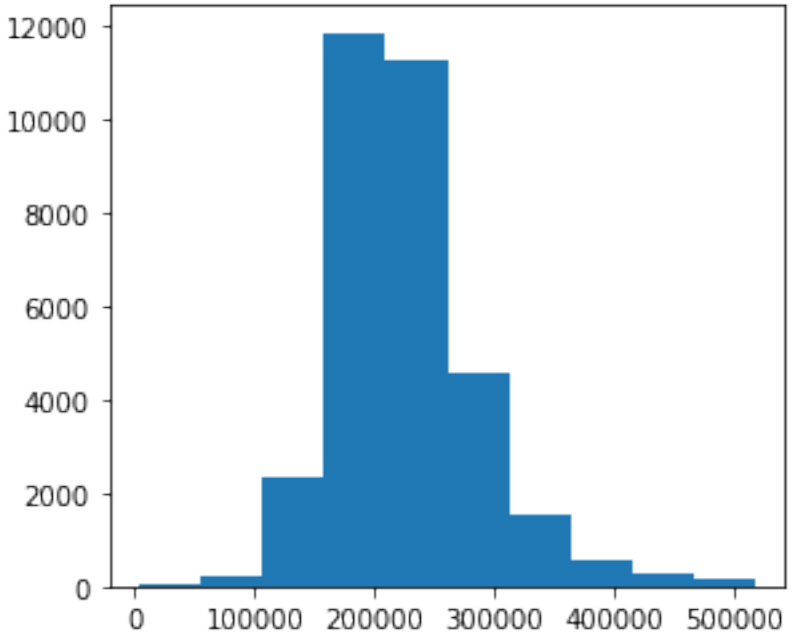


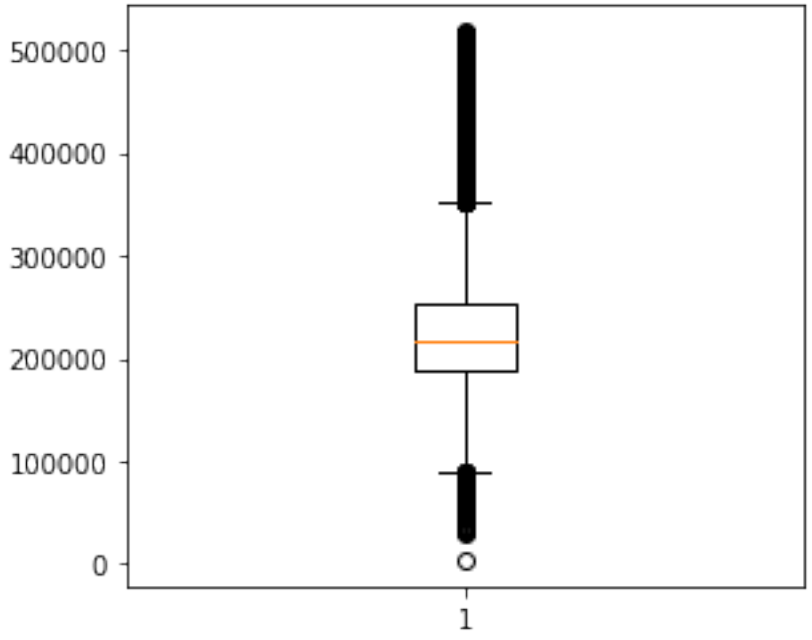
Does a feature makes a sense?

2. Distribution & feature meaning (2)

Some features are Binary, as MODE (minor/major).

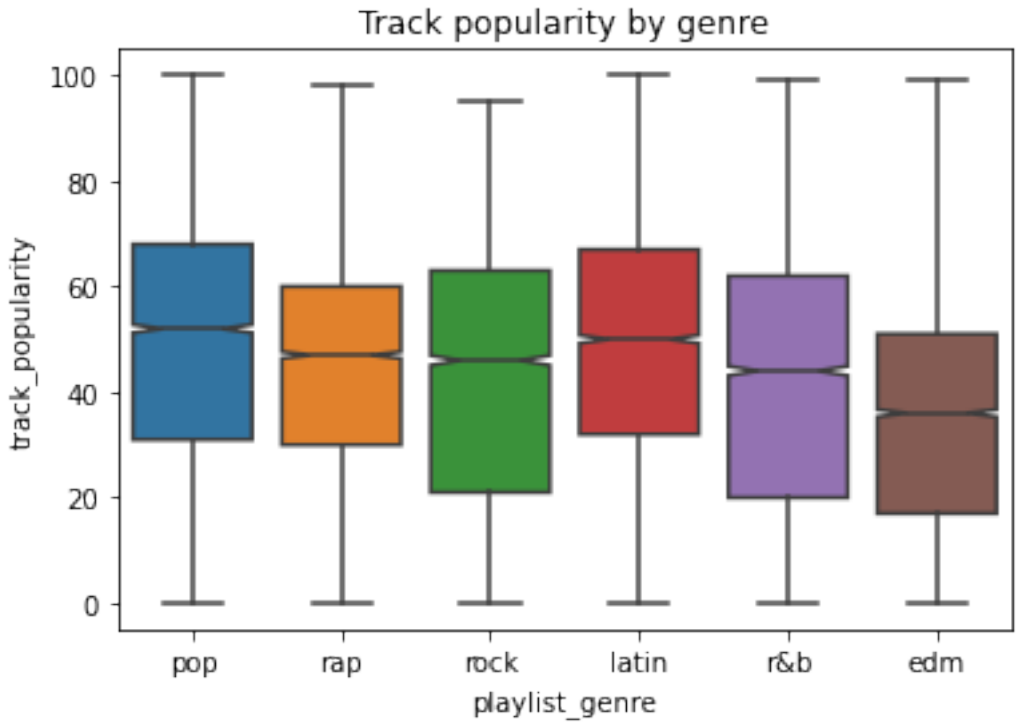


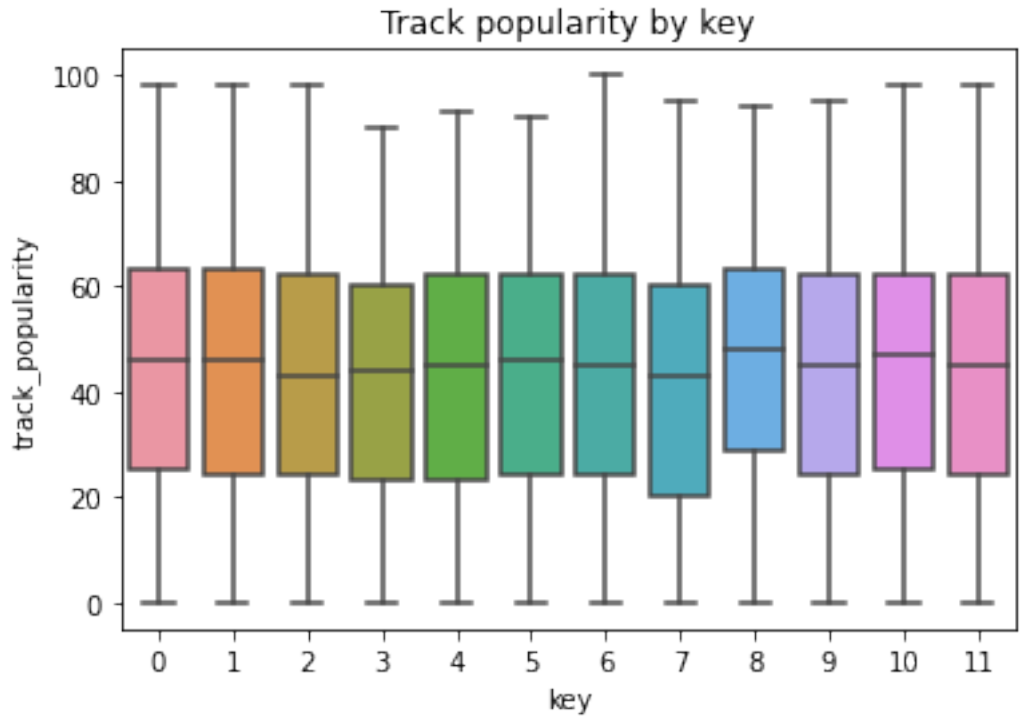




Some features are in very different scale

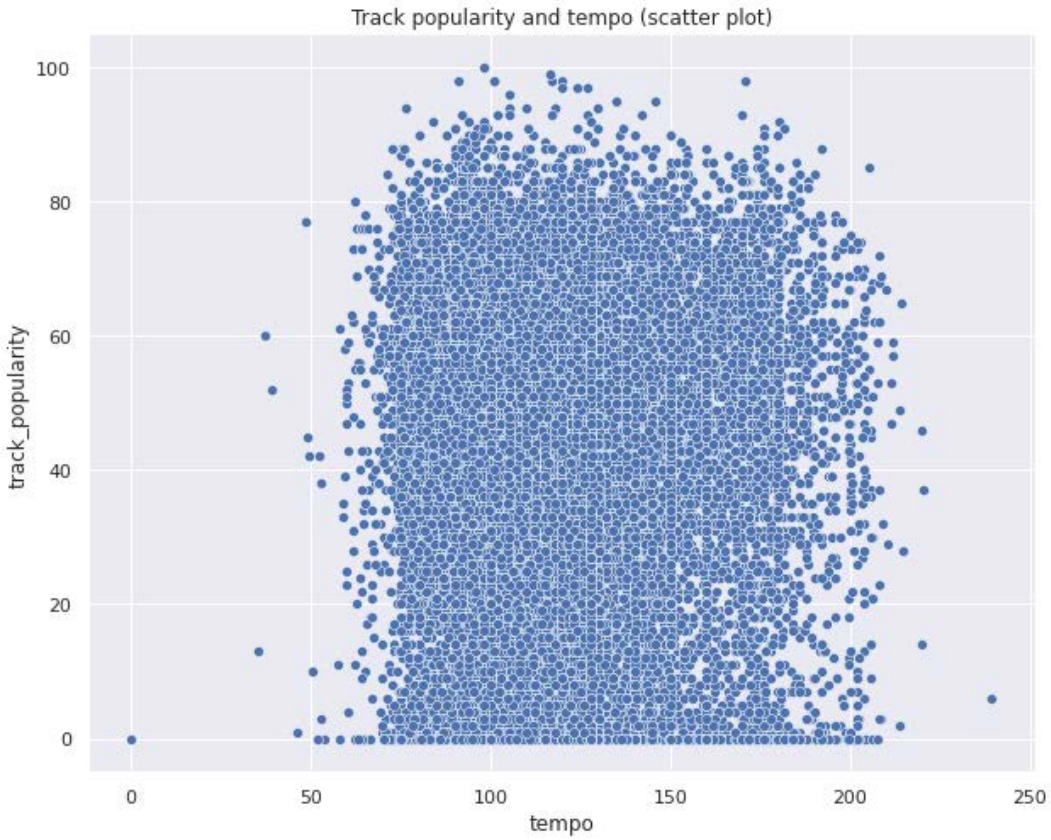
3. Explore connections to the outcome (1)

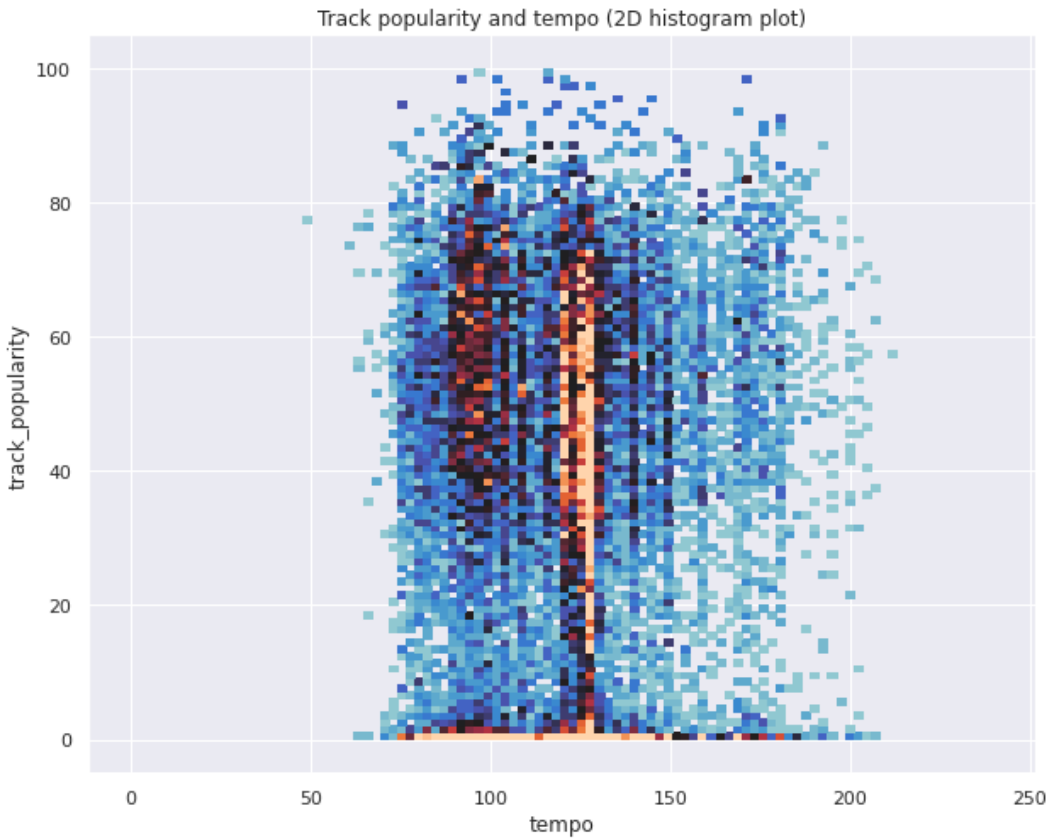




Features seem to be connected to the outcome differently

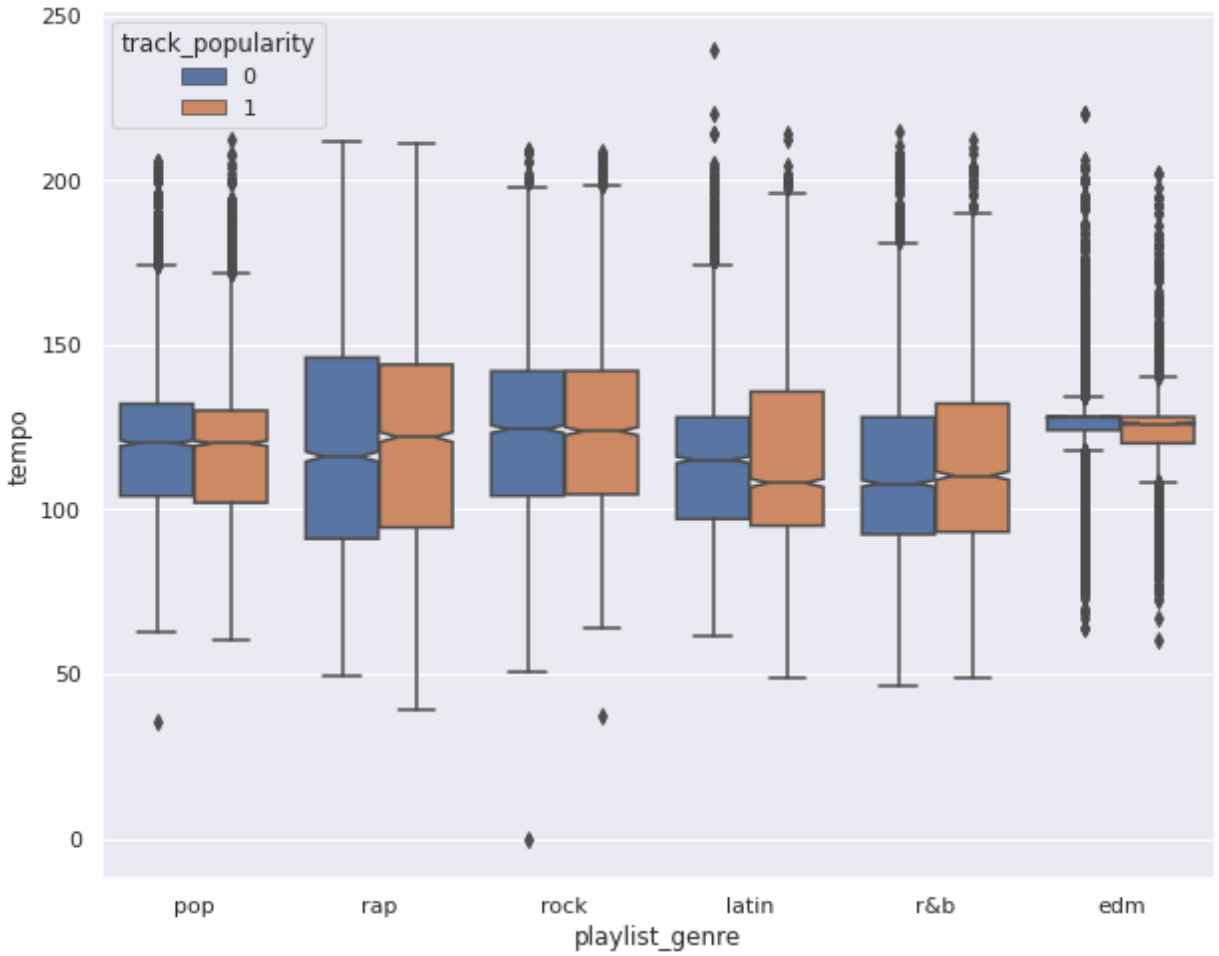
3. Explore connections to the outcome (2)





Exploring density can reveal hidden details

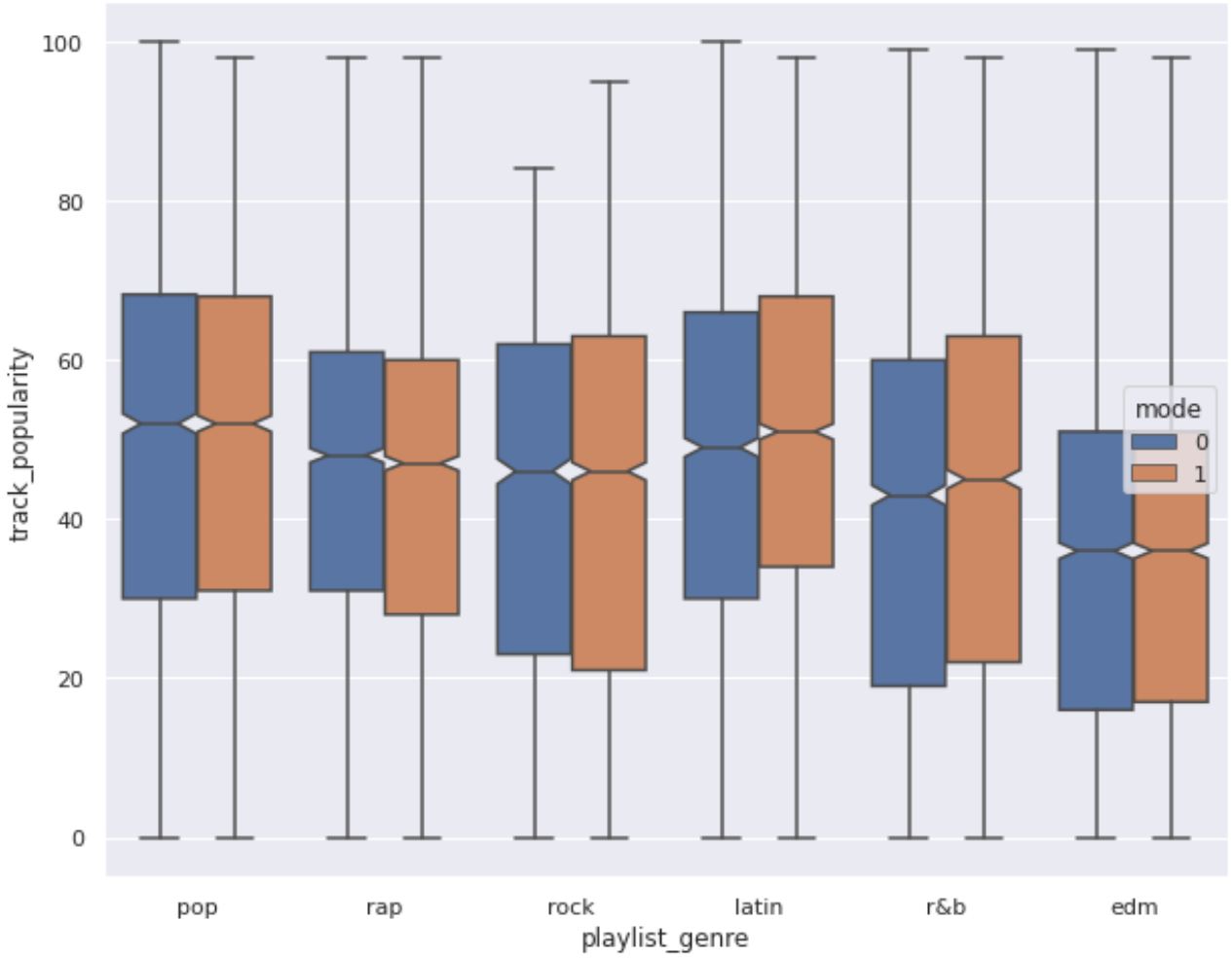
3. Explore connections to the outcome (3)



⇒ Faster rap is more popular then slower rap

⇒ Faster latin is less popular then faster latin

3. Explore connections to the outcome (4) – stat.tests

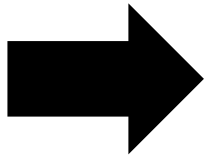


Statistical test (Mann-Whitney-U test):

```
Genre: edm p-value:          0.854
Genre: latin p-value:        0.024 *
Genre: pop p-value:          0.657
Genre: r&b p-value:          0.005 **
Genre: rap p-value:          0.011 *
Genre: rock p-value:         0.915
```

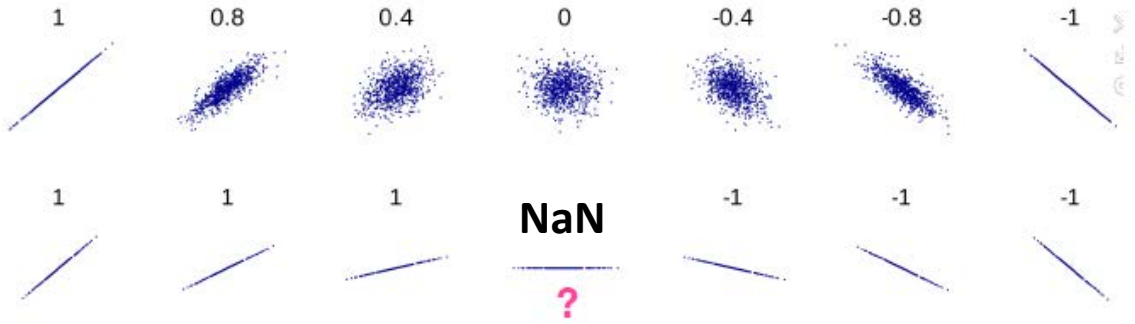
4. Dealing with categorical features: one-hot-encoding

playlist_genre
pop
pop
pop
pop
pop
...
edm
edm
edm
edm
edm



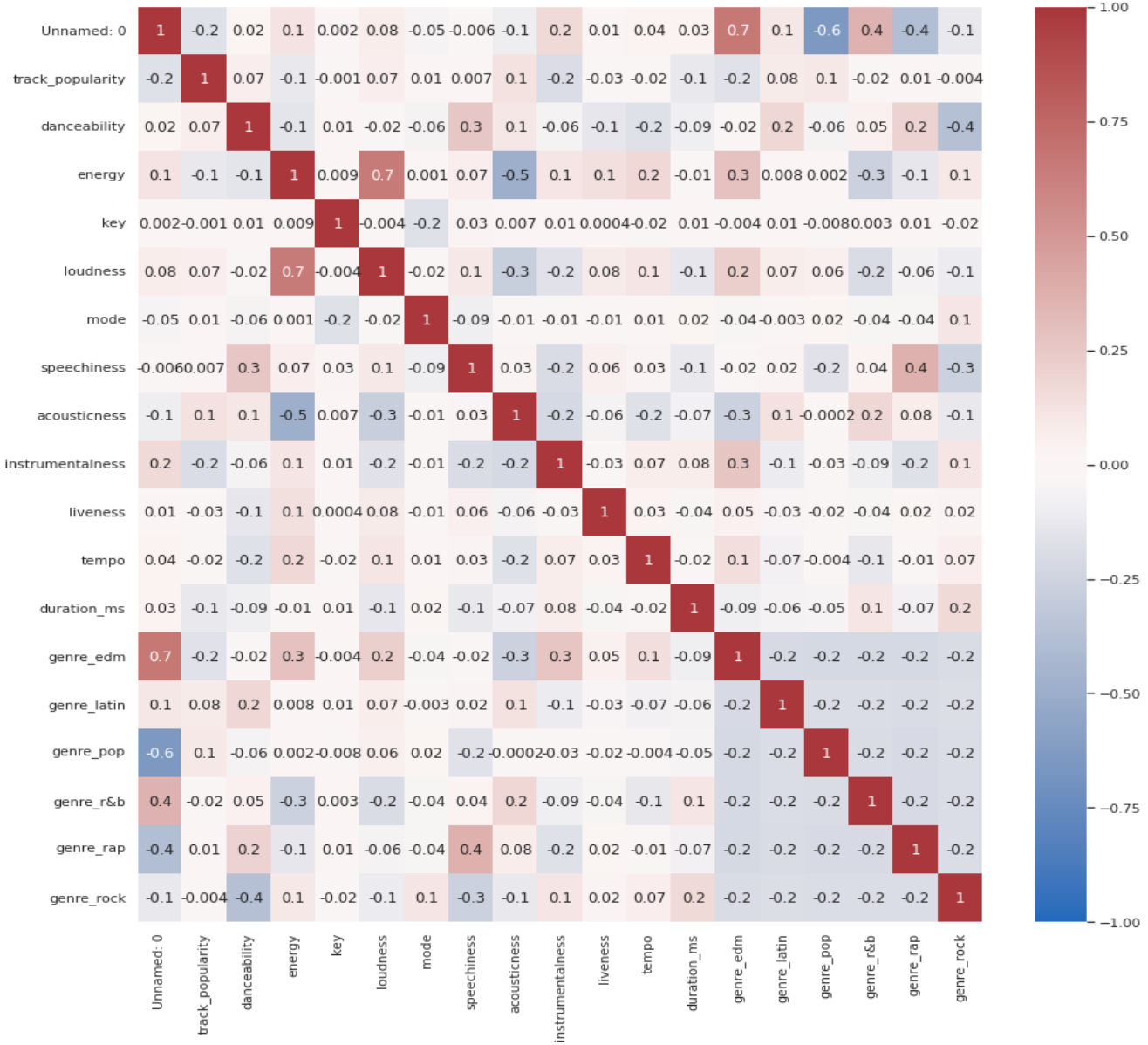
genre_edm	genre_latin	genre_pop	genre_r&b	genre_rap	genre_rock
0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0
...	...	...	...	...	...
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0

5. Exploring colinearities



https://cs.wikipedia.org/wiki/Korelace#/media/Soubor:Correlation_examples2.svg

$$r_{x,y} = \frac{cov(x,y)}{std(x) \cdot std(y)}$$



Important points

- **Do not imply** that the dataset contains only **valid data**
- Take your time to **understand** the **meaning** of each feature
- Do not leave **NaN**s inside. Look out for constant columns
- **Correlated** features may debase your effort (depending on a model)
- Do not forget to **encode categorical features**

Thank you for your attention

Filip Plesinger (fplesinger@isibrno.cz)

Do you have any questions?

Our further activities:

5.10.2022 – ICRC Academy (15:00, here)
Umělá inteligence pro analýzu poruch srdeční činnosti
https://akademie.fnusa.cz/?p=1311

8.11.2022 – SignalPlant workshop (the whole day, here)
signal analysis and processing
www.signalplant.org