



# **Klasteryzacja dla ekspresji genów w nowotworach mózgu**

---

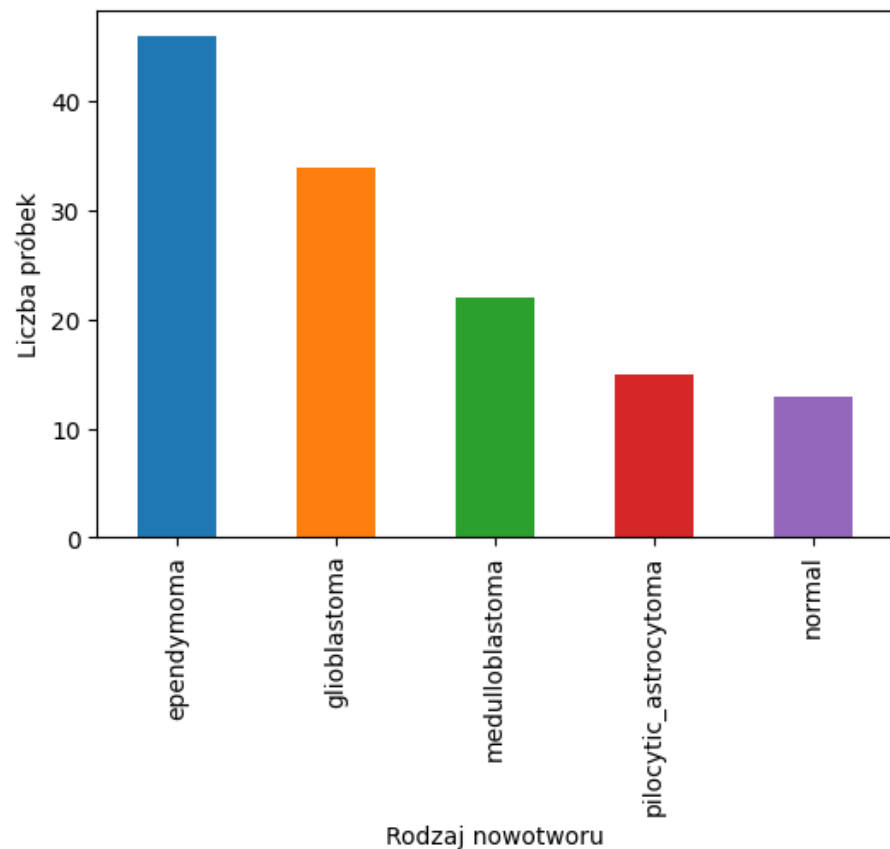
Julia Radacka, Liliana Sirko

# Cel biznesowy

- Optymalizacja projektów badań klinicznych: próbki z podobnym profilem mogą reagować na leczenie w podobny sposób.
- Dzięki klasteryzacji będziemy w stanie wyłonić grupy pacjentów o zbliżonych genach, co stworzy bardziej jednorodną grupę do badania i zwiększy szanse powodzenia próby klinicznej.

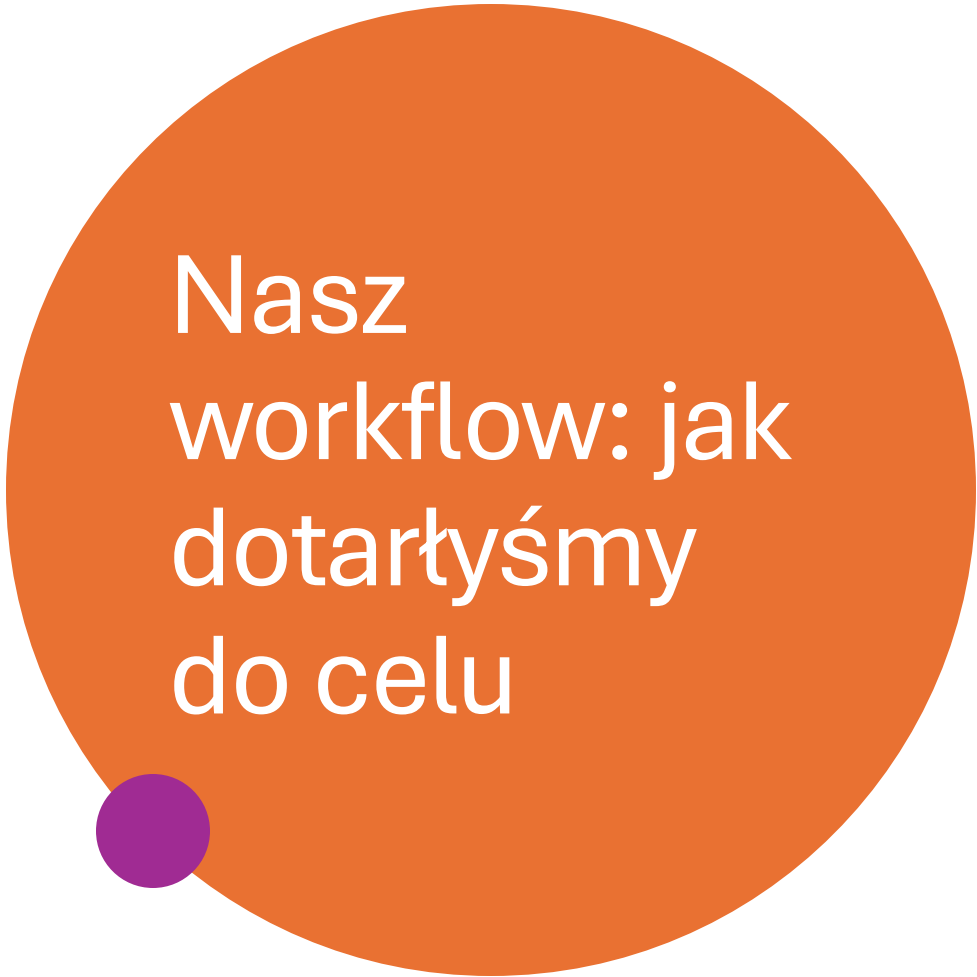
# Ramka danych: CUMIDA

Ekspresja genów: wskazuje jak bardzo aktywne są dane geny



130 wierszy, 54677 kolumn

	samples	1007_s_at	1053_at	117_at	121_at
count	130.000000	130.000000	130.000000	130.000000	130.000000
mean	898.500000	12.276393	8.769583	7.722634	9.160209
std	37.671829	0.790160	0.673396	1.037339	0.615369
min	834.000000	10.156207	6.627878	6.222515	8.044421
25%	866.250000	11.679721	8.378760	7.007678	8.595505
50%	898.500000	12.502518	8.786242	7.521674	9.194487
75%	930.750000	12.883374	9.211098	8.249157	9.707397
max	963.000000	13.655639	10.716003	12.054143	10.407136



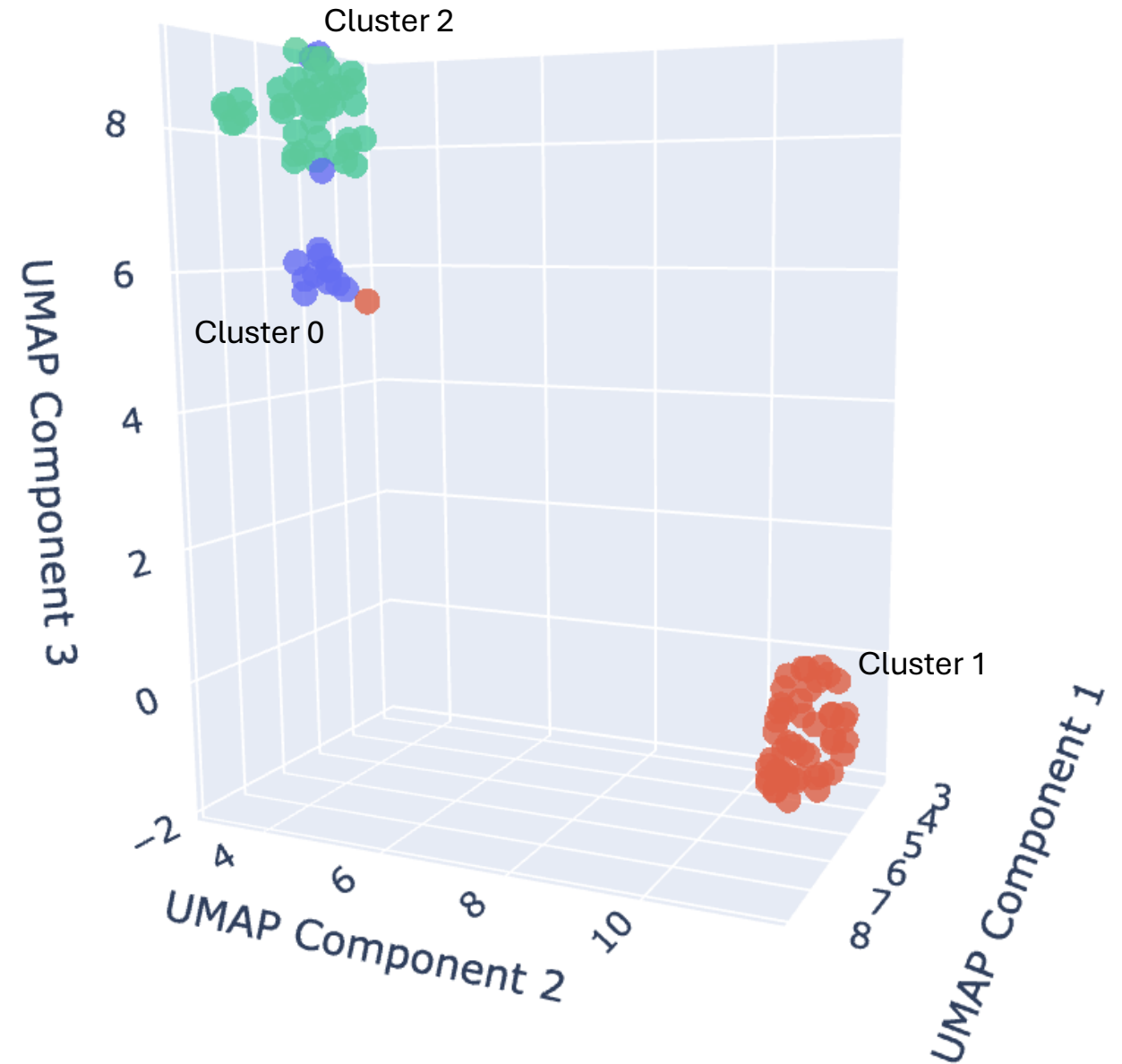
Nasz  
workflow: jak  
dotarliśmy  
do celu

- **Preprocessing:** Standaryzacja, PCA
- **Wizualizacja danych:** Umap
- **Metryki:** Silhouette score, Davies-Bouldin Index, Dunn Index
- **Modelowanie:** Kmeans, modele aglomeratywne, DBSCAN, GMM...
- **Interpretacja klastrów**

# Klasteryzacja

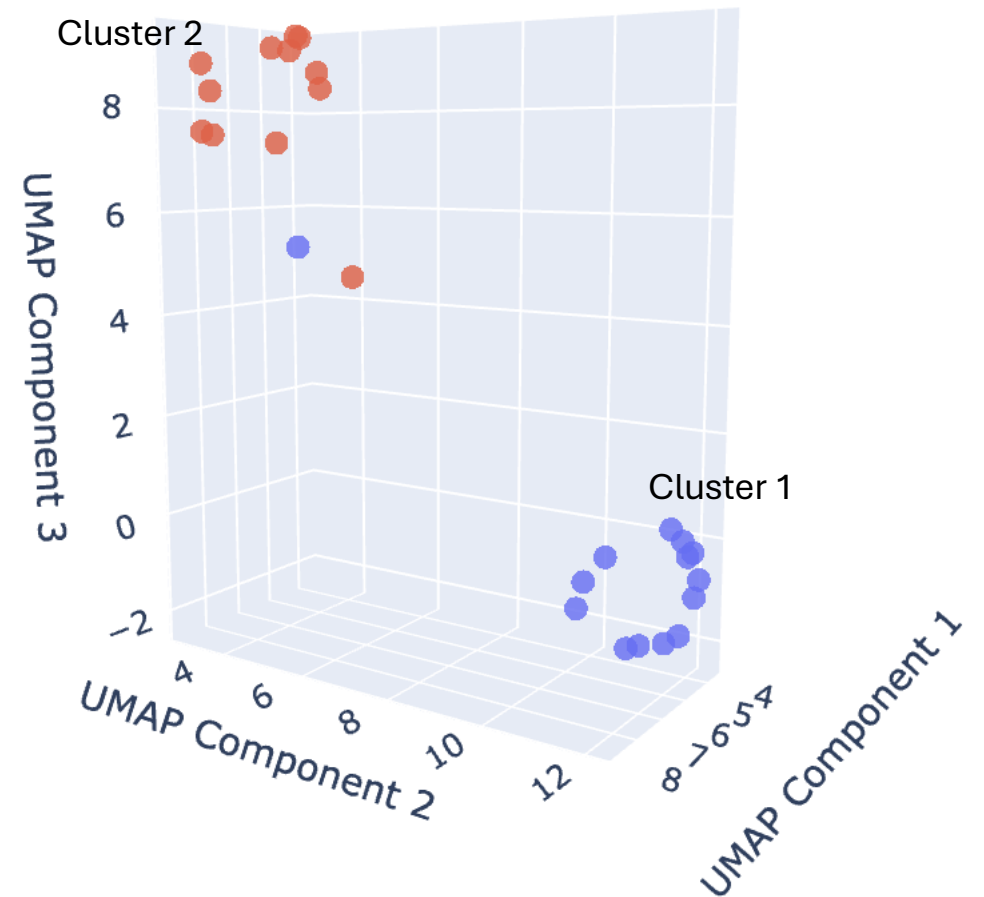
## Najlepszy model:

Agglomerative, complete linkage,  
3 klastry



# Zbiór walidacyjny: jak to działa dla niezależnych danych

Accuracy on test set: 0.85  
(klasyfikacja z RandomForestClassifier)





# Interpretacja klastrow

---

Co oznaczają te geny? Jak może nam to pomóc?

## Top geny wpływające na PC:

241291_at	0.033121
222974_at	0.029052
205590_at	0.028633
221111_at	0.028348
244724_at	0.028040
1553254_at	0.027517
1556166_x_at	0.027425
237351_at	0.027291
1566469_at	0.026977
241764_at	0.026563
1566772_at	0.026516
226942_at	0.026408
1562049_at	0.026405
223975_at	0.026404
1561055_at	0.025990

# Interpretacja klastrow cd...

Cechy PCA mające największy wpływ na predykcję

	Feature	Mean_SHAP
0	PC1	0.037588
1	PC2	0.023995
46	PC47	0.011260
42	PC43	0.008821
2	PC3	0.007545
32	PC33	0.005898
69	PC70	0.005023
16	PC17	0.005009
35	PC36	0.004350
18	PC19	0.004020
13	PC14	0.003832
8	PC9	0.003747
14	PC15	0.003634
52	PC53	0.003501
22	PC23	0.003426

Średnia ekspresja genów w klastrach

	241291_at	222974_at	205590_at	221111_at	244724_at	1553254_at	\
cluster							
0	-0.302717	0.265828	1.358599	0.117547	1.089405	0.549944	
1	0.075950	-0.129927	-0.399265	-0.240005	0.206419	-0.011140	
2	0.014106	0.059879	0.003436	0.235109	-0.599042	-0.170583	

	1556166_x_at	237351_at	1566469_at	241764_at	1566772_at	\
cluster						
0	-0.196101	-0.118658	0.615706	-0.101596	1.342552	
1	0.396476	0.085926	0.346728	0.018078	-0.021520	
2	-0.387748	-0.058648	-0.601496	0.013205	-0.422924	

	226942_at	1562049_at	223975_at	1561055_at
cluster				
0	-0.686195	0.306848	0.271953	0.701664
1	-0.273371	-0.147955	0.336479	0.126758
2	0.541155	0.066809	-0.475199	-0.378755



# Interpretacja klastrow cd...

- **205590\_at** (RASGRP1) - regulator cyklu komórkowego, często aktywny w agresywnych guzach.

	241291_at	222974_at	205590_at	221111_at	244724_at	1553254_at	\
cluster							
0	-0.302717	0.265828	1.358599	0.117547	1.089405	0.549944	
1	0.075950	-0.129927	-0.399265	-0.240005	0.206419	-0.011140	
2	0.014106	0.059879	0.003436	0.235109	-0.599042	-0.170583	

	1556166_x_at	237351_at	1566469_at	241764_at	1566772_at	\
cluster						
0	-0.196101	-0.118658	0.615706	-0.101596	1.342552	
1	0.396476	0.085926	0.346728	0.018078	-0.021520	
2	-0.387748	-0.058648	-0.601496	0.013205	-0.422924	

	226942_at	1562049_at	223975_at	1561055_at
cluster				
0	-0.686195	0.306848	0.271953	0.701664
1	-0.273371	-0.147955	0.336479	0.126758
2	0.541155	0.066809	-0.475199	-0.378755

# Nasze wnioski

- **Klaster 0:** (wysoka ekspresja genów RASGRP1 i kilku innych) -> reprezentuje agresywny rodzaj nowotworu (prawdopodobnie **glioblastoma** lub **medulloblastoma**), gdzie aktywne geny napędzają proliferację, cykl komórkowy i potencjalnie migrację komórek nowotworowych.
- **Klaster 1:** (ekspresja bliska zeru) -> grupa **normalnych** lub zdrowych próbek, bez wyraźnych zmian w genach.
- **Klaster 2:** (obniżona ekspresja większości genów) -> może reprezentować mniej agresywny lub inny typ nowotworu (np. **ependymoma, pilocytic astrocytoma**), który nie korzysta z aktywacji wymienionych szlaków.

Dziękujemy za uwagę 😊

# Bibliografia

- [https://dom-pubs.onlinelibrary.wiley.com/doi/10.1111/dom.15123?](https://dom-pubs.onlinelibrary.wiley.com/doi/10.1111/dom.15123)