Selekcja cech

Metody i narzędzia *Big Data*

Selekcja cech

statystyczne własności cechy

Selekcja cech

Filtrowanie – progowanie wariancji

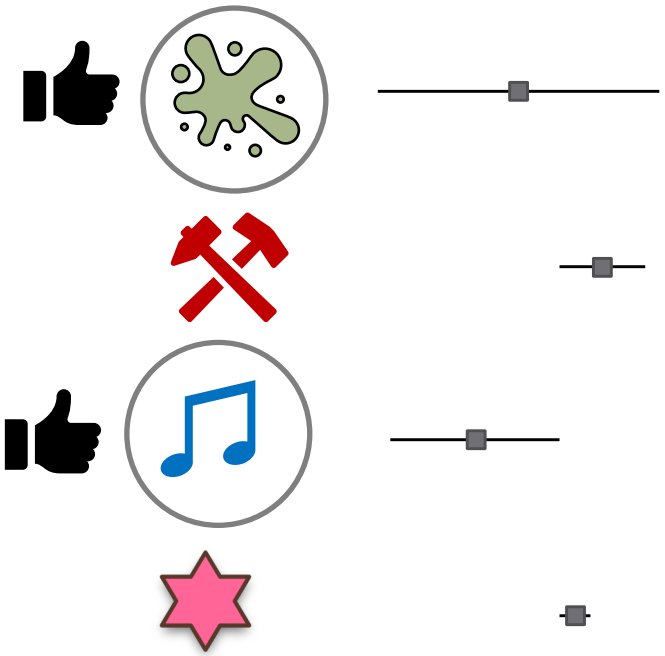
- Niska wariancja cechy → niska zawartość informacyjna
- https://scikit-learn.org/stable/modules/feature_selection.html
- Cecha liczbowa

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{(j)} - \mu_j\right)^2}$$

- Cecha binarna

$$\sigma_j = p\,(1-p)$$

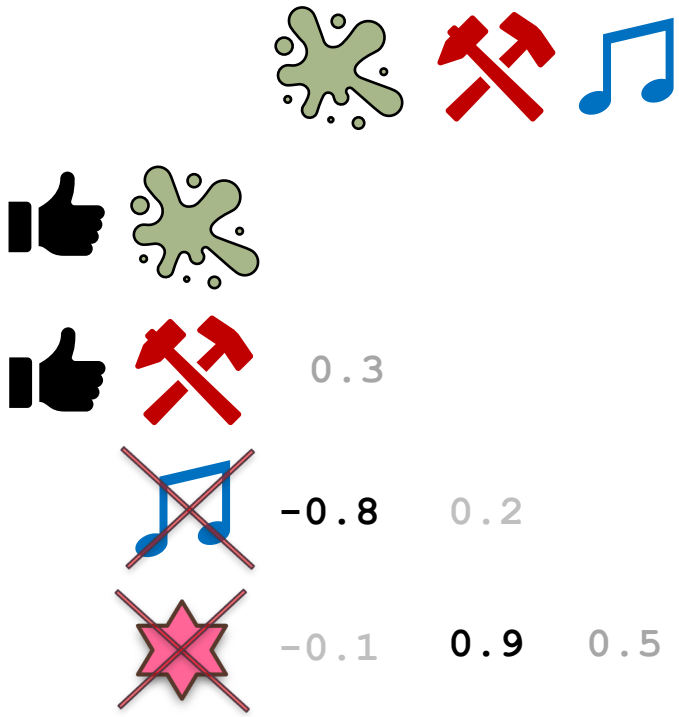
- Uwaga: nie standaryzować cech



Selekcja cech

Filtrowanie – usunięcie skorelowanych cech

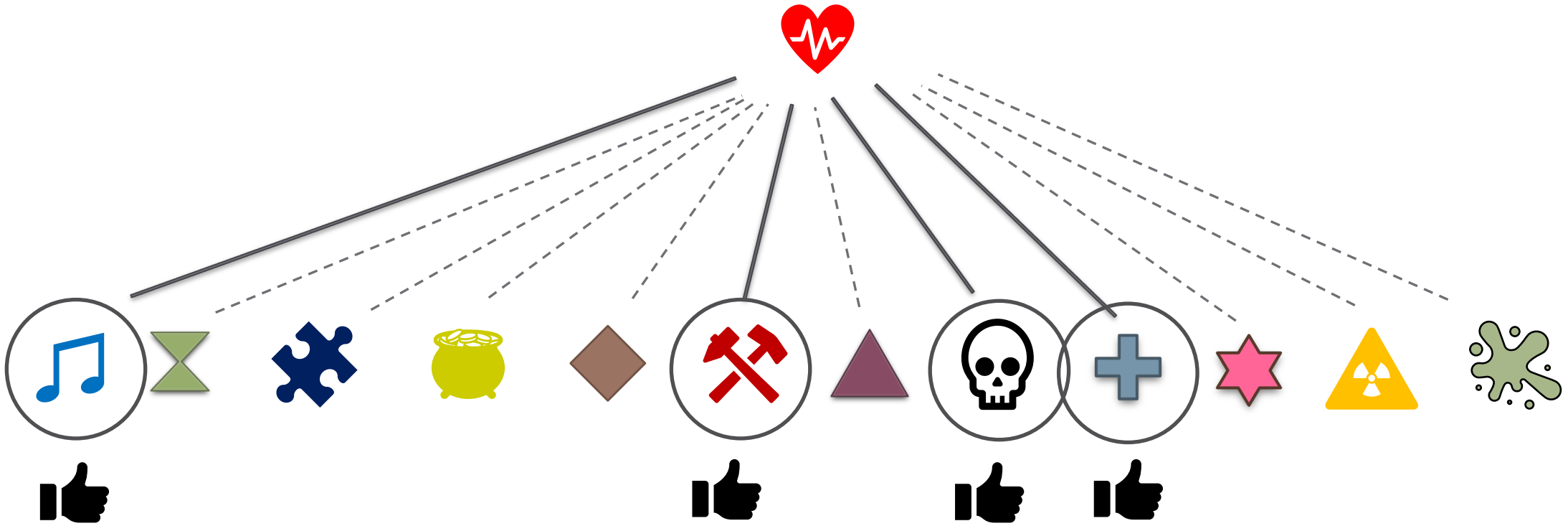
- Skorelowane cechy→ podobna zawartość informacyjna
- Współczynniki korelacji:
  - Pearsona
  - Spearmana
  - Kendalla
  - U-Theila
- Test chi-kwadrat dla danych kategorialnych
- Sprawdzać istotność korelacji
- Uwaga: czy celem jest
  - predykcja / klasyfikacja?
  - analiza zjawiska?



Selekcja cech

Testowanie statystyczne współzmienności cech i wyjść

- Selekcja jednoczynnikowa (ang. *Univariate Selection*)
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html



Selekcja cech

jaka jest jakość modelu
przy zadanym zbiorze cech

Selekcja cech

Sekwencyjny wybór cech

- **Sekwencyjna selekcja postępująca**, ang. *Sequential Forward Selection (SFS)*

Obliczenie jakości klasyfikatora / predyktora

Selekcja cech

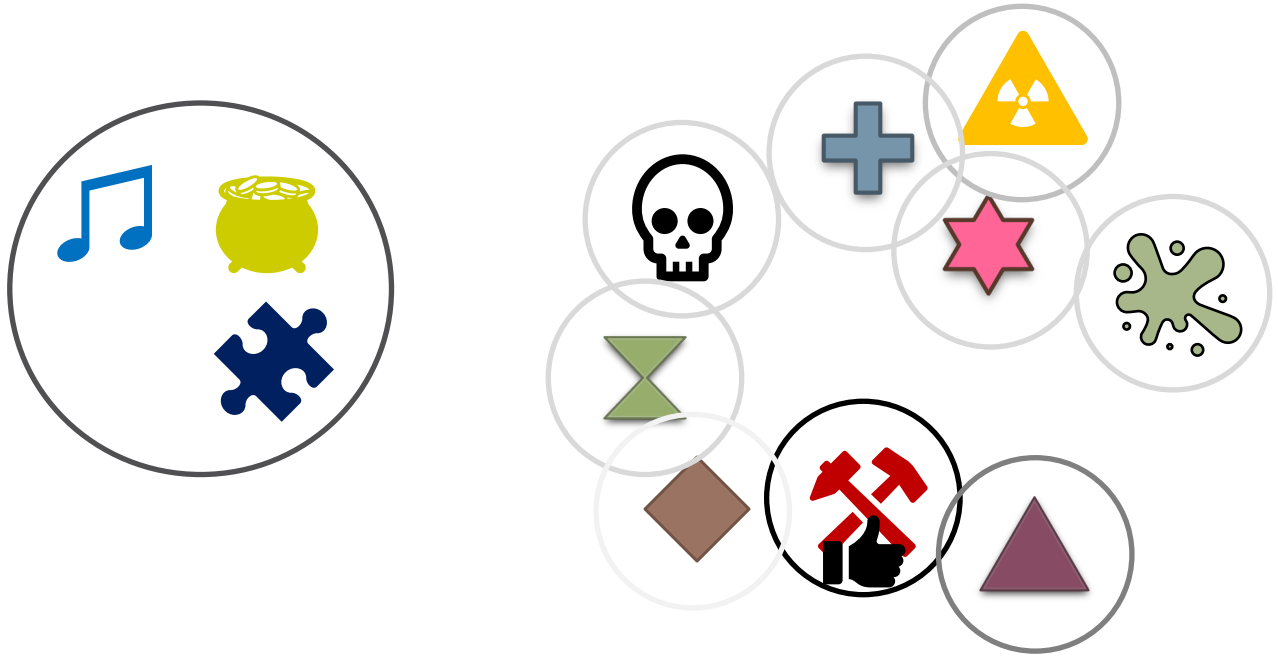
Sekwencyjny wybór cech

- **Sekwencyjna selekcja postępująca**, ang. *Sequential Forward Selection (SFS)*

Selekcja cech

Sekwencyjny wybór cech

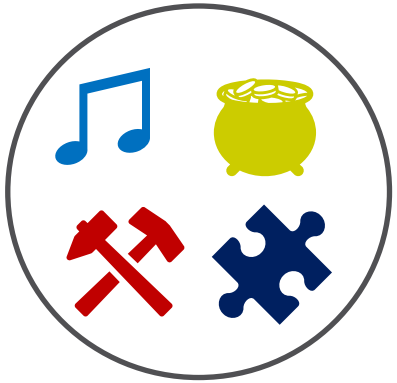
- **Sekwencyjna selekcja postępująca**, ang. *Sequential Forward Selection (SFS)*

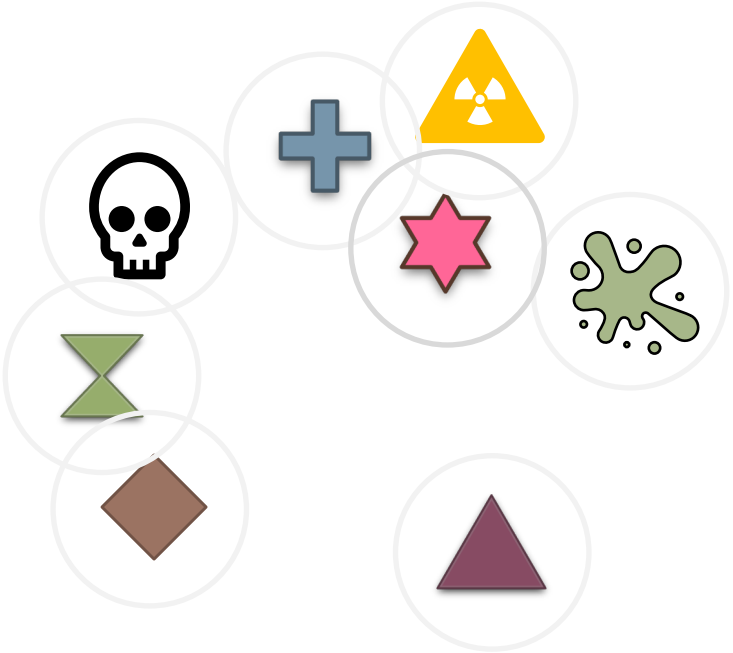


Selekcja cech

Sekwencyjny wybór cech

- **Sekwencyjna selekcja postępująca**, ang. *Sequential Forward Selection (SFS)*

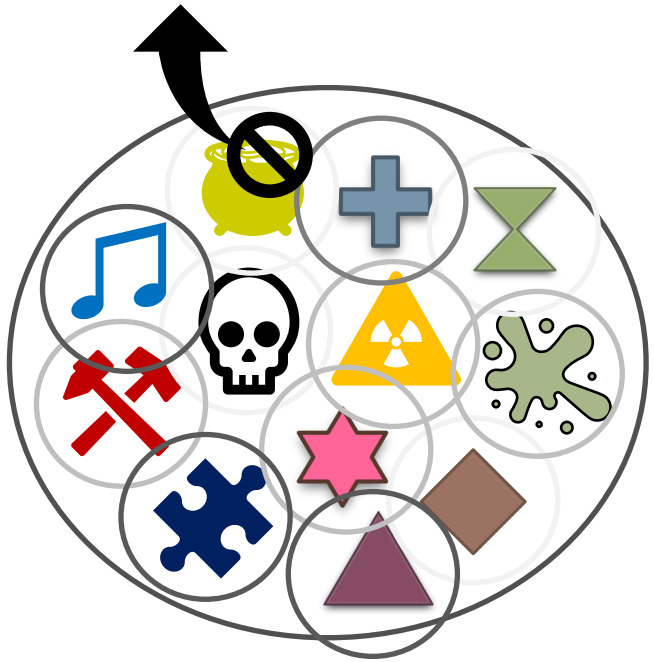




Selekcja cech

Sekwencyjny wybór cech

- **Sekwencyjna selekcja wsteczna**, ang. *Sequential Backward Selection (SBS)*



Selekcja cech

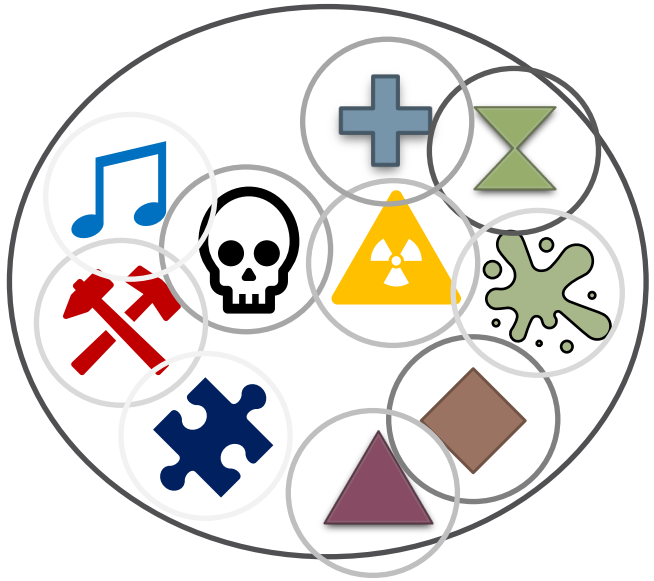
Sekwencyjny wybór cech

- **Sekwencyjna selekcja wsteczna**, ang. *Sequential Backward Selection (SBS)*

Selekcja cech

Sekwencyjny wybór cech

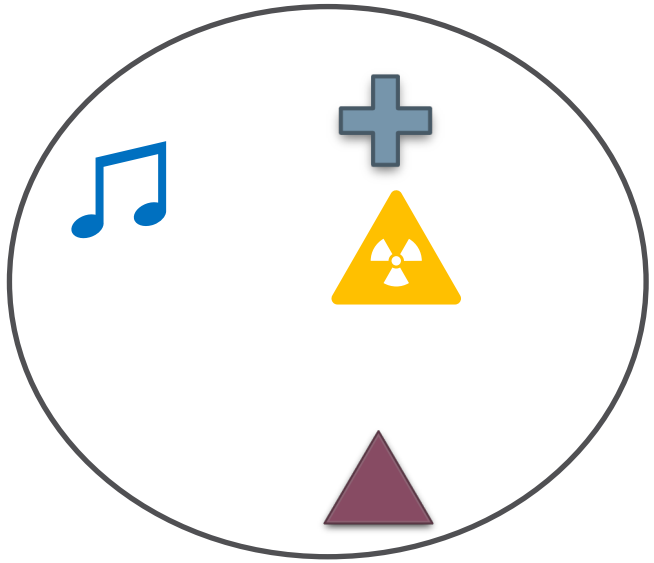
- **Sekwencyjna selekcja wsteczna**, ang. *Sequential Backward Selection (SBS)*



Selekcja cech

Sekwencyjny wybór cech

- **Sekwencyjna selekcja wsteczna**, ang. *Sequential Backward Selection (SBS)*



Selekcja cech

Sekwencyjny wybór cech

- **Sekwencyjna selekcja postępująca / wsteczna z nawrotami**, ang. *Sequential Forward / Backward Selection (SFFS / SFBS)*
- Wady metod zachłannych

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html

https://sebastianraschka.com/pdf/software/mlxtend-latest.pdf

http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

Selekcja cech

model ujawnia ważność cech

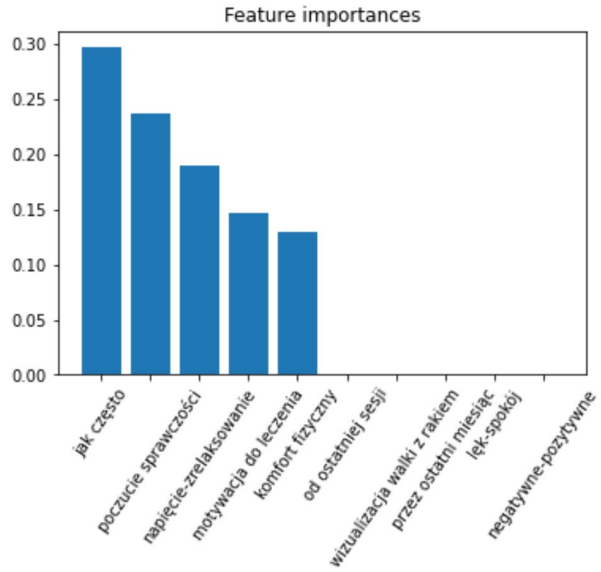
Selekcja cech

Metody osadzone (w modelu)

- Związane z konkretnym modelem
- Uogólniony model liniowy: współczynniki modelu reprezentują „ważność” cechy

$$\overline{\mathbf{y}} = \sum_{m=0}^{M} a_m \psi_m(\mathbf{x})$$

- Drzewa decyzyjne (lasy losowe): ważność cechy (ang. *feature importance*) jako średni spadek zanieczyszczeń (ang. *impurity*) dla wszystkich drzew
  - https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
- UWAGA: jeżeli cechy są powiązane to jedna oceniona jest wysoko a druga nisko (przy predykcji OK, ale nie przy interpretacji istotności cech)



Selekcja cech

Regularyzacja

- Regularyzacja obniża złożoność modelu
- Użycie normy l_1 prowadzi do rzadkich wektorów cech

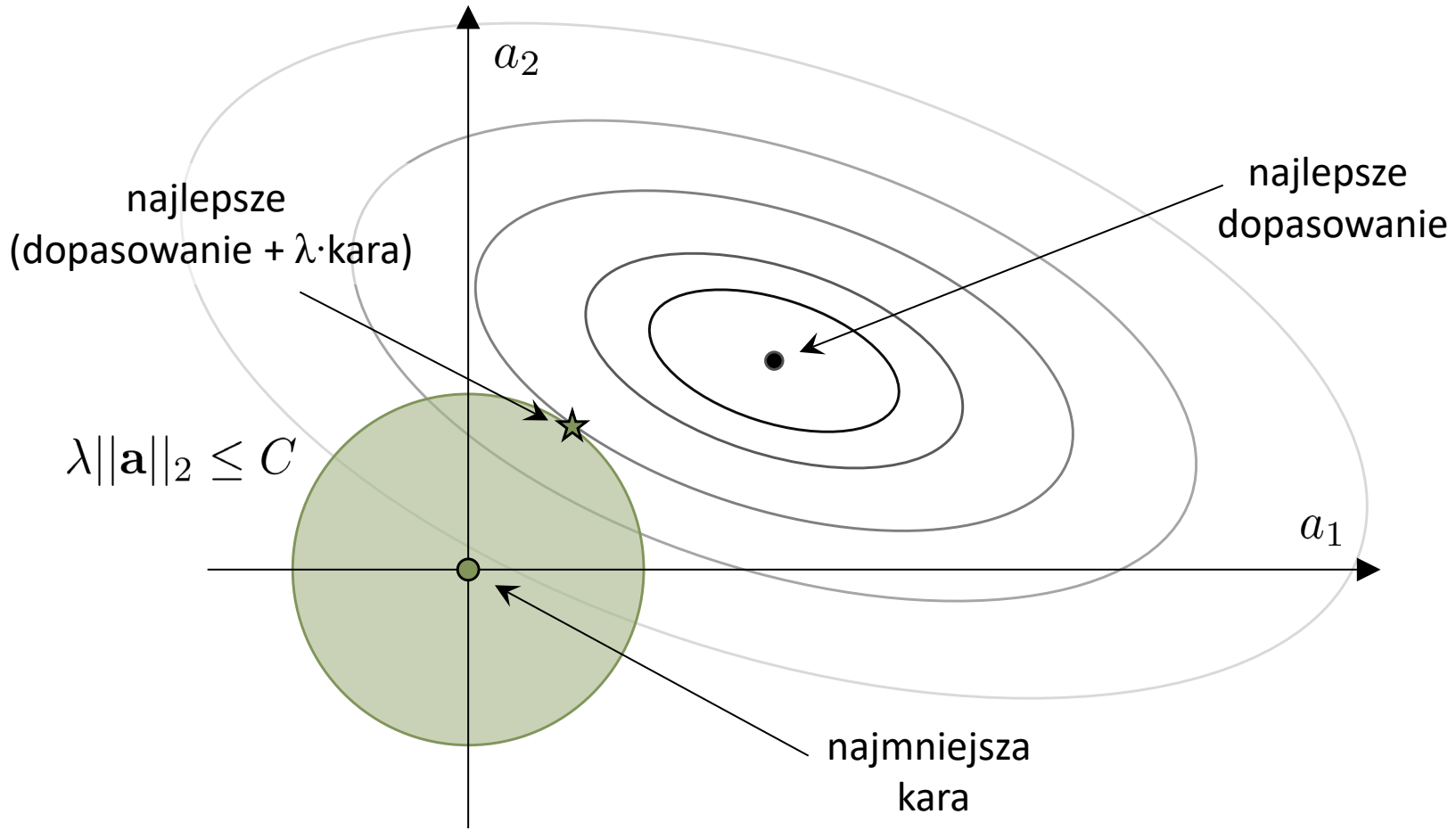
$$\text{norma } l_2: \quad ||\mathbf{a}||_2 = \sqrt{\sum_{i=1}^{M} a_i^2} = \sqrt{\mathbf{a}\mathbf{a}^T}$$

$$\text{norma } l_1: \quad ||\mathbf{a}||_1 = \sum_{i=1}^{M} |a_i|$$

Ocena modelu = błąd dopasowania + $\lambda \cdot$ kara za złożoność

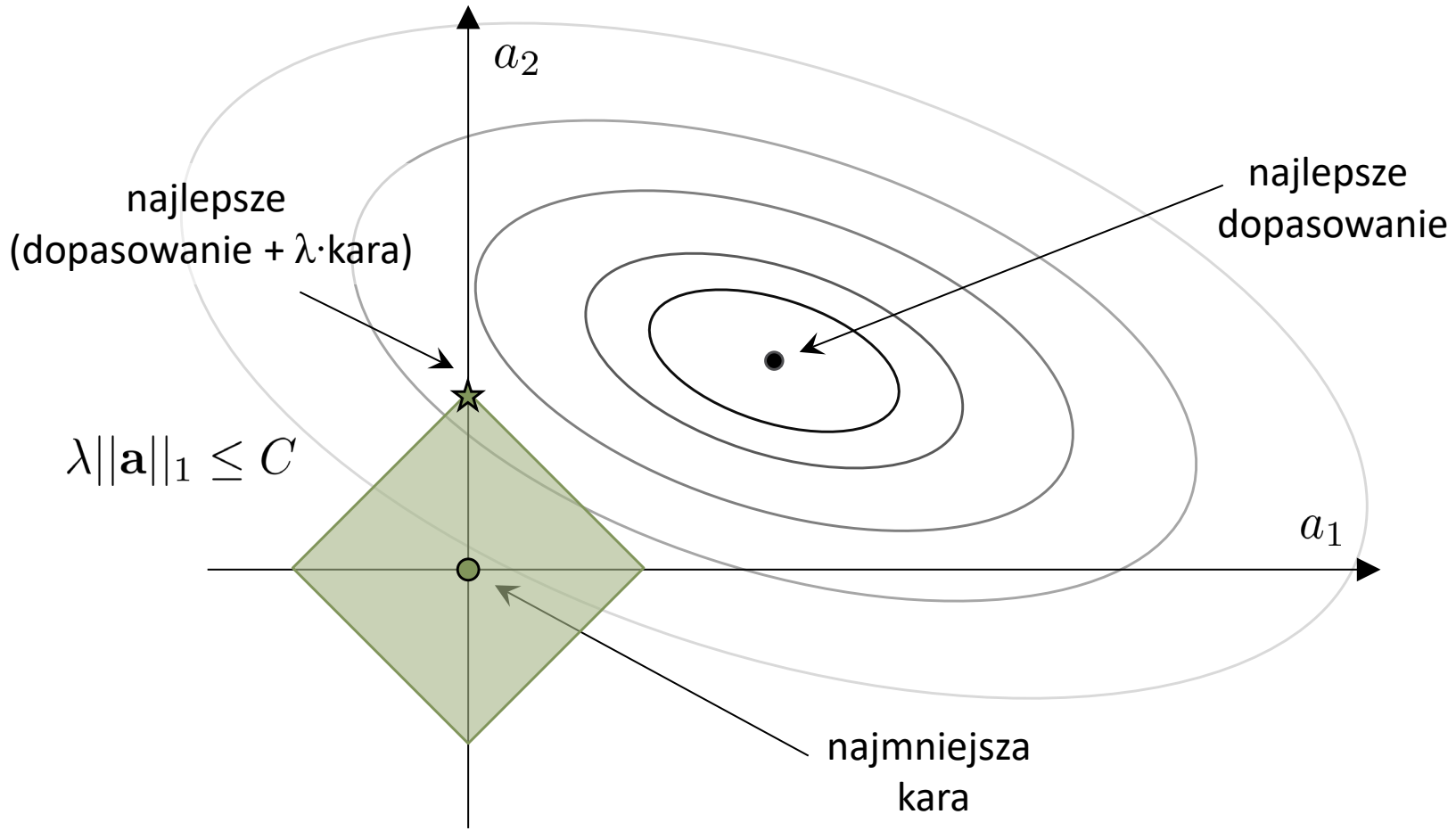
Selekcja cech

Regularyzacja – dlaczego nie l_2



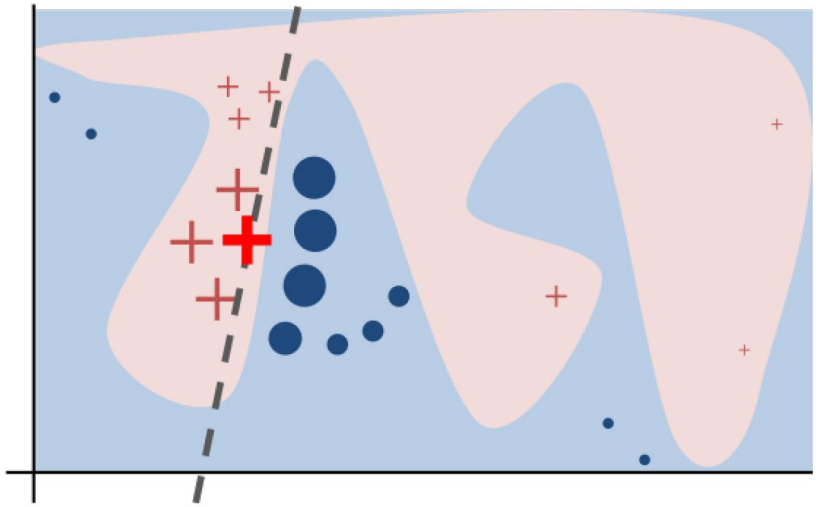
Selekcja cech

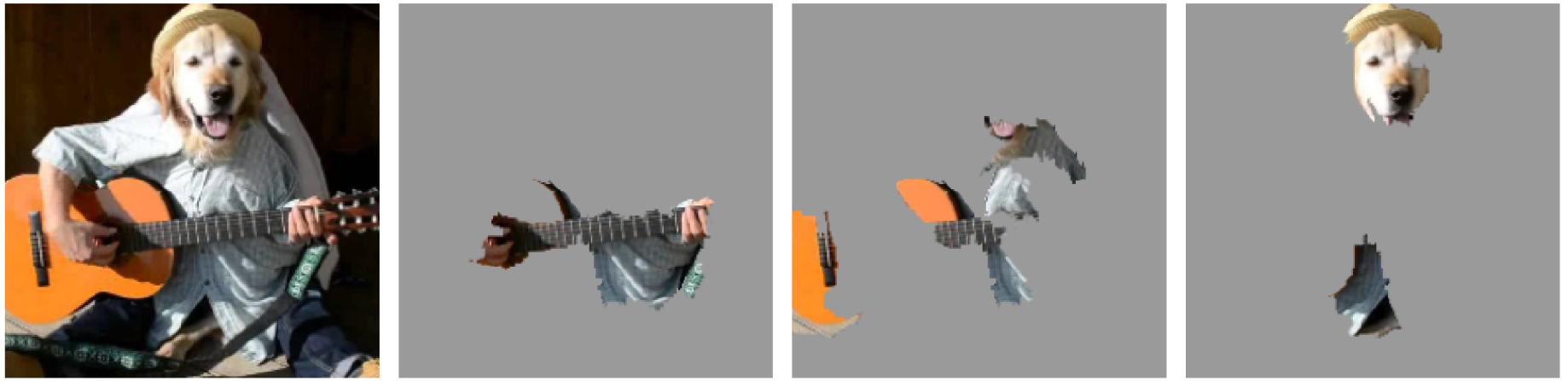
Regularyzacja – dlaczego akurat l_1



Local Interpretable Model-Agnostic Explanations (lime)

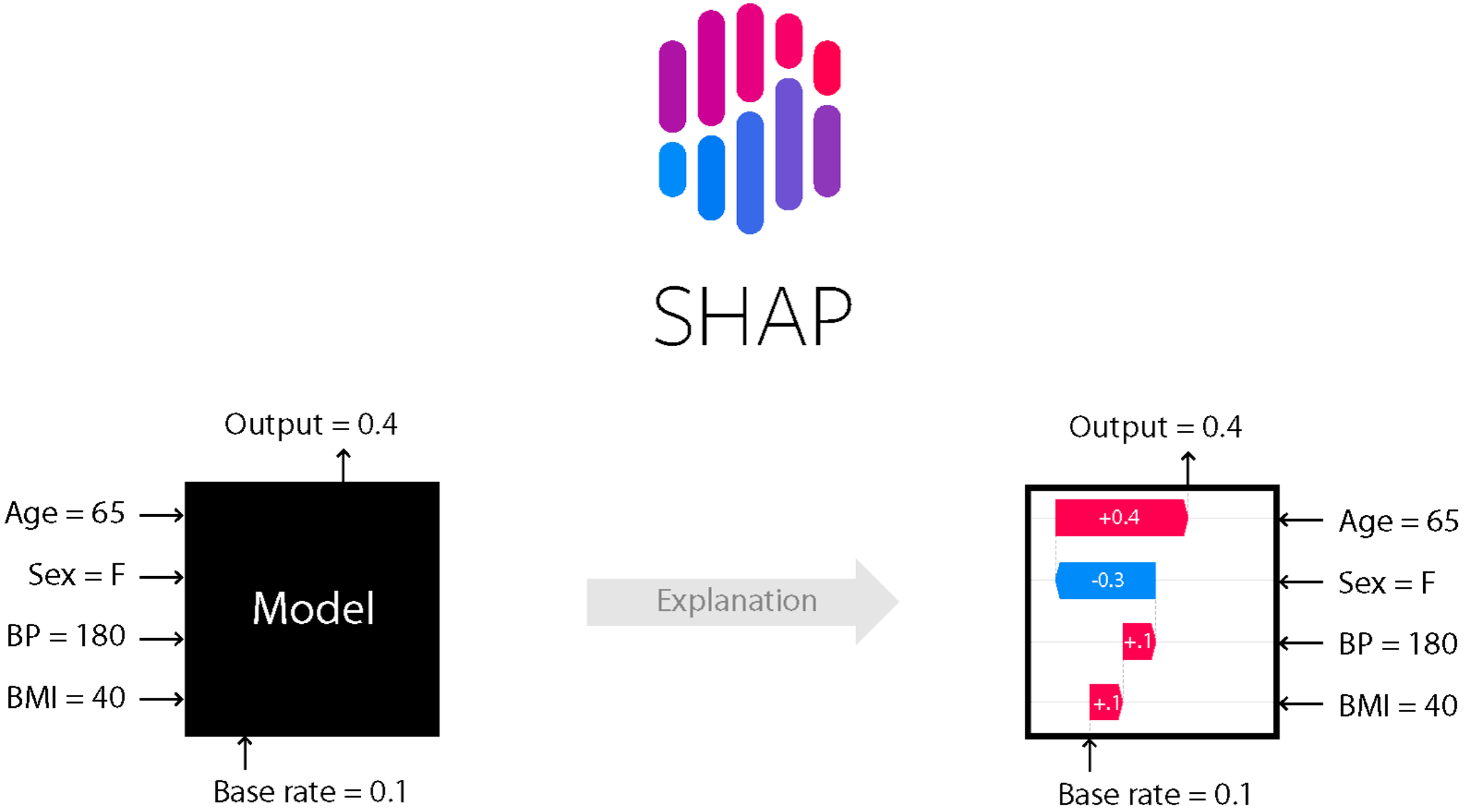
In this page, you can find the Python API reference for the lime package (local interpretable model-agnostic explanations). For tutorials and more information, visit the github page.





(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

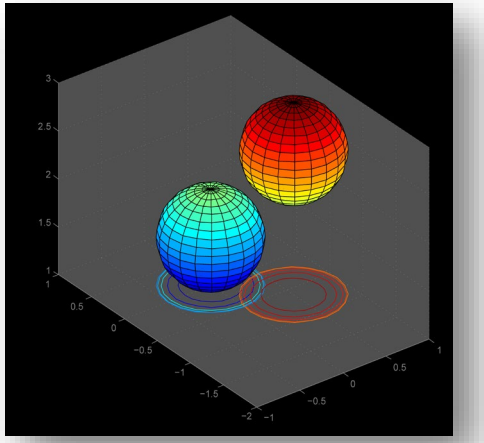
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)


SHAP
Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1

Selekcja cech

ang. *feature selection*

- Wyznaczyć **podzbiór** najbardziej przydatnych cech
- Można zrezygnować z niewiele wnoszących pomiarów



Ekstrakcja cech

ang. *feature extraction*

- Utworzyć nową **podprzestrzeń** cech
- **Kompresja danych** pod kątem zachowania jak największej ilości użytecznej informacji
- Niespodziewanie może **poprawić jakość predykcji** (klątwa wymiarowości i korelacje)


The End