



MeetContentWRO #36

Embeddings in AI and Content Anecdotes



with Jarosław Drapała and Paweł Martin

 6 PM | 22 April 2026

 Ocado Technology
Szewska 8 | 4th floor
Wrocław



<https://docs.google.com/forms/d/e/1FAIpQLScsCsfZIVXAjaR0u2g6myWTFR97ikLzcKb2QNn1R53RvQVgpQ/viewform>



Selekcja cech

Metody i narzędzia
Big Data

Selekcja cech

Przez filtrowanie

statystyczne własności cechy

Przez opakowywanie

Przez osadzanie

Selekcja cech

Filtrowanie – progowanie wariancji

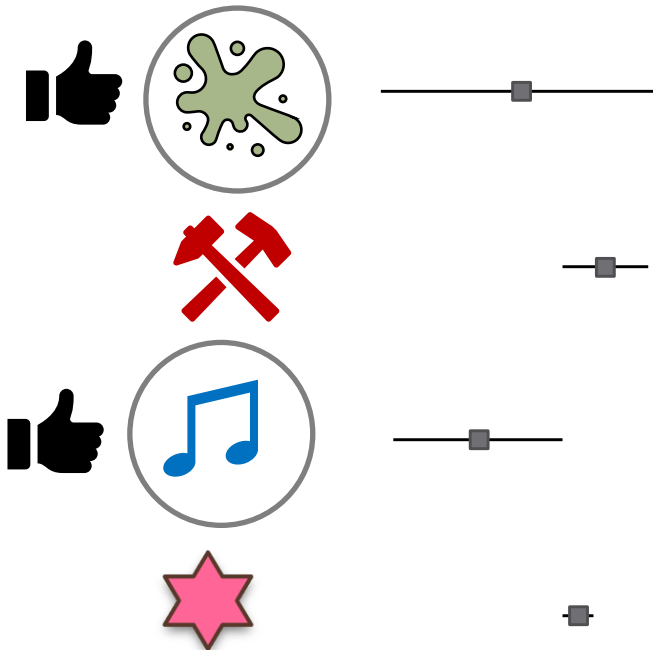
- Niska wariancja cechy → niska zawartość informacyjna
- https://scikit-learn.org/stable/modules/feature_selection.html
- Cecha liczbowa

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

- Cecha binarna

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

- Uwaga: nie standaryzować cech

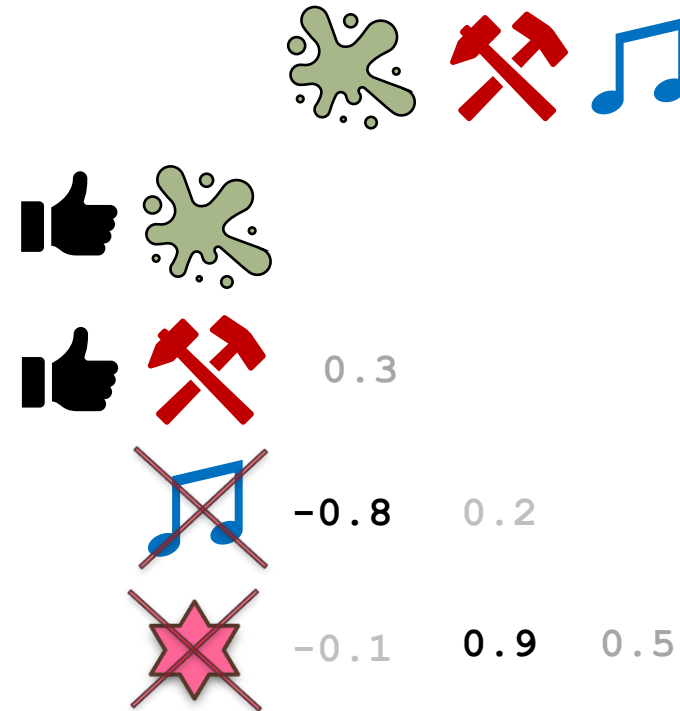


Selekcja cech

Filtrowanie – usunięcie skorelowanych cech



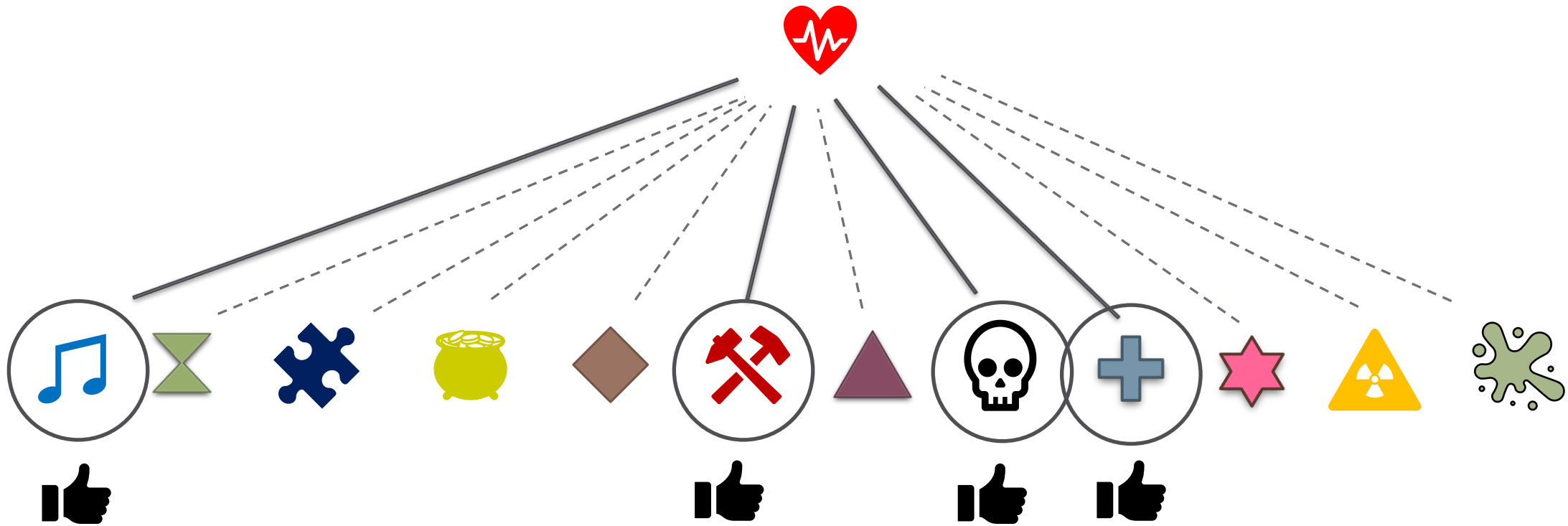
- Skorelowane cechy → podobna zawartość informacyjna
- Współczynniki korelacji:
 - Pearsona
 - Spearmana
 - Kendalla
 - U-Theila
- Test chi-kwadrat dla danych kategoryjnych
- Sprawdzać istotność korelacji
- Uwaga: czy celem jest
 - predykcja / klasyfikacja?
 - analiza zjawiska?



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

Przez filtrowanie

Przez opakowywanie
ang. wrapper

Przez osadzanie

*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)*

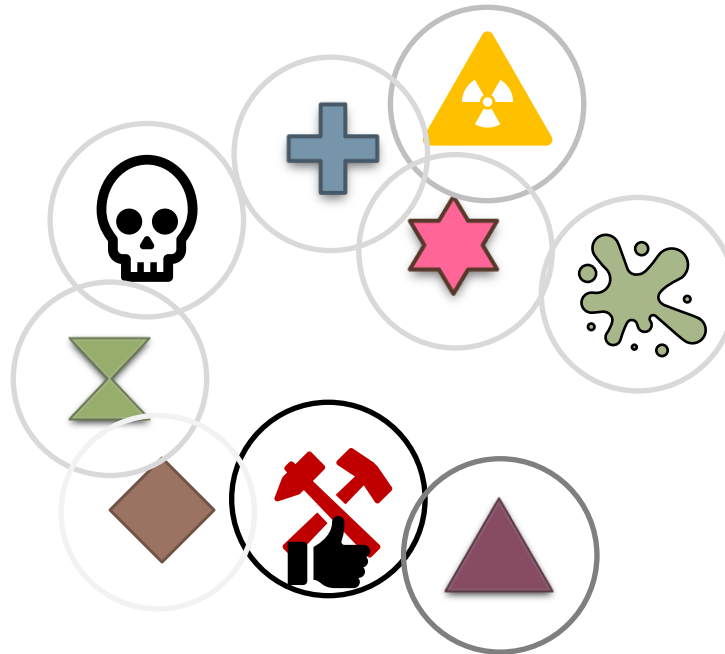
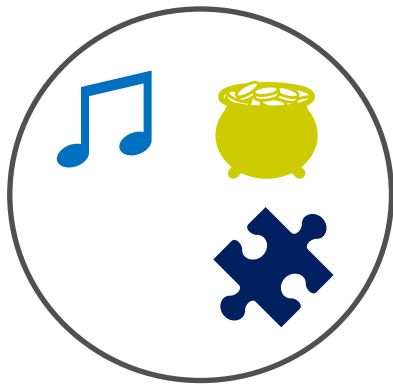


Selekcja cech

Sekwencyjny wybór cech



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

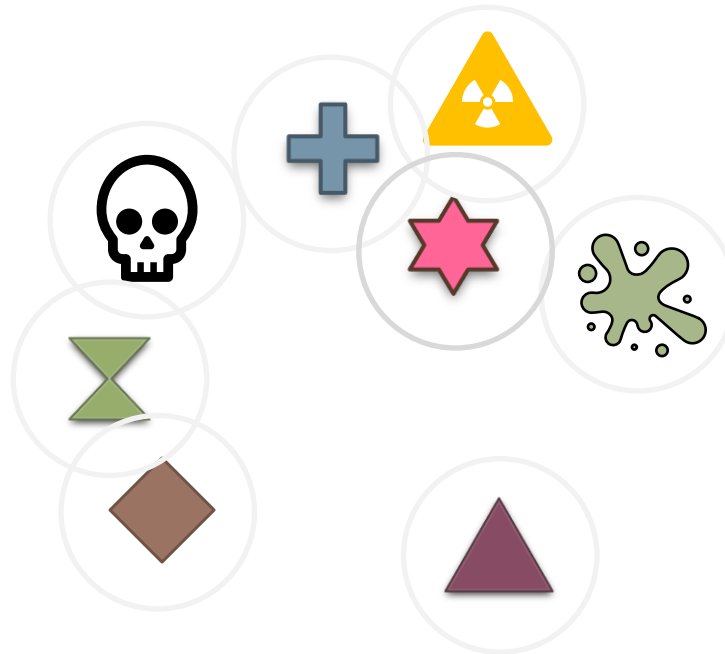
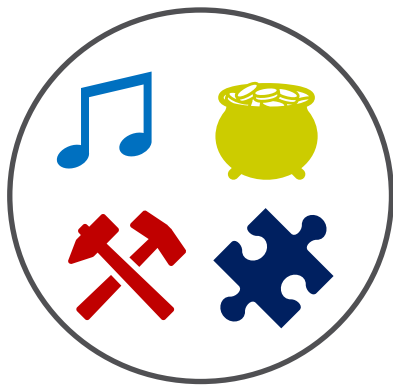


Selekcja cech

Sekwencyjny wybór cech



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

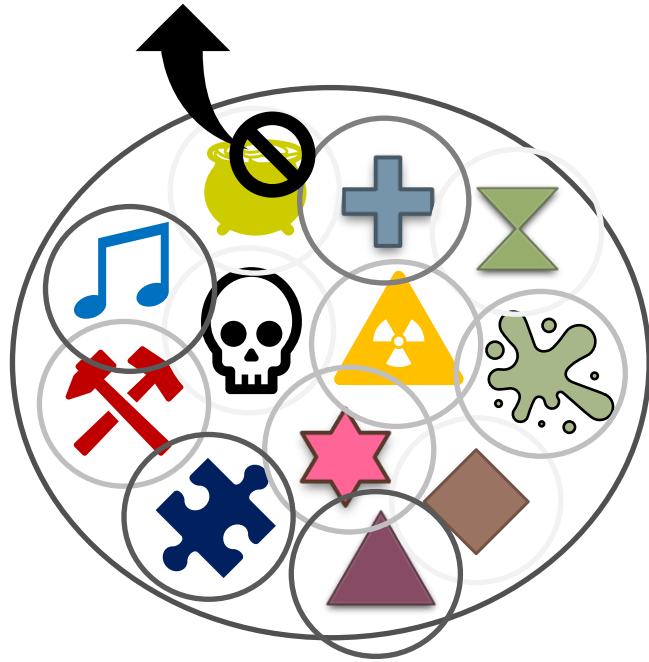


Selekcja cech

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)*

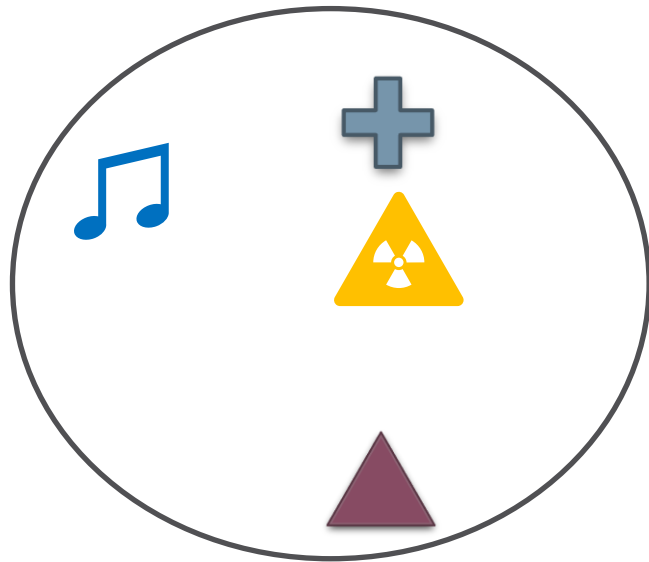


Selekcja cech

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

Przez filtrowanie

Przez opakowywanie

Przez osadzanie

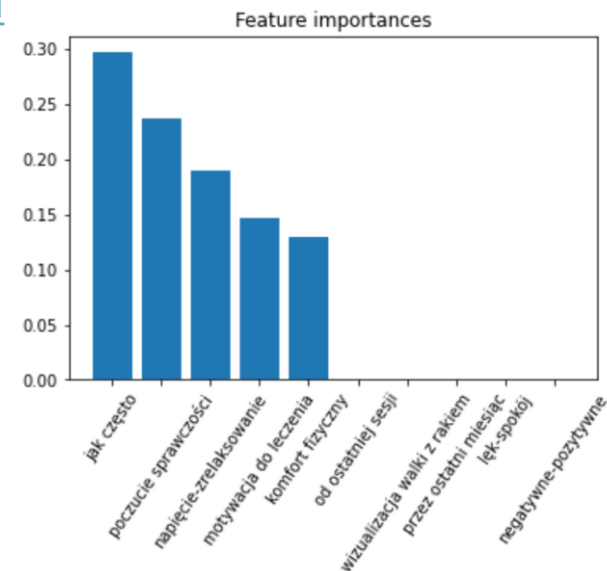
model ujawnia ważność cech

Selekcja cech

Metody osadzone (w modelu)

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

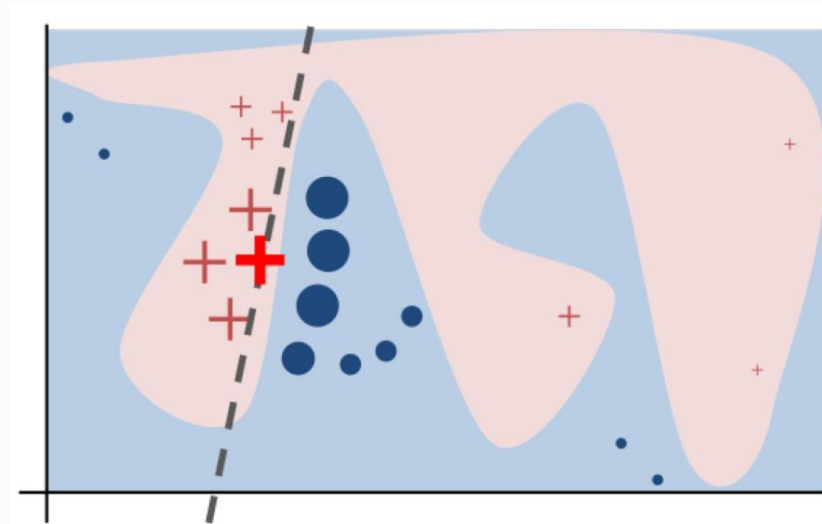
$$\bar{y} = \sum_{m=0}^M a_m \psi_m(\mathbf{x})$$



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](#).

- [lime package](#)
 - Subpackages
 - Submodules
 - [lime.discretize](#) module
 - [lime.exceptions](#) module
 - [lime.explanation](#) module
 - [lime.lime_base](#) module
 - [lime.lime_image](#) module
 - [lime.lime_tabular](#) module
 - [lime.lime_text](#) module
 - [lime.submodular_pick](#) module
 - Module contents

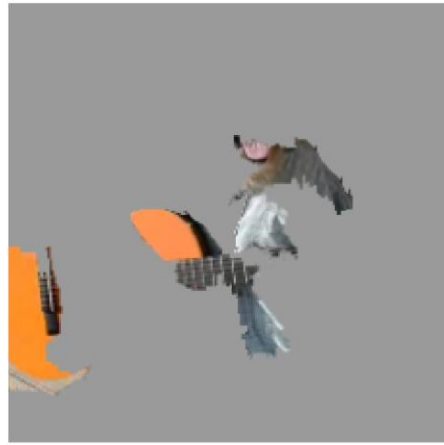




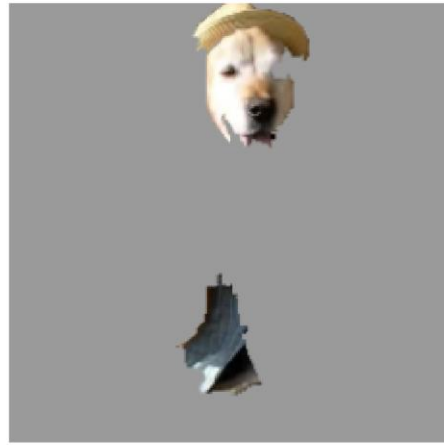
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*

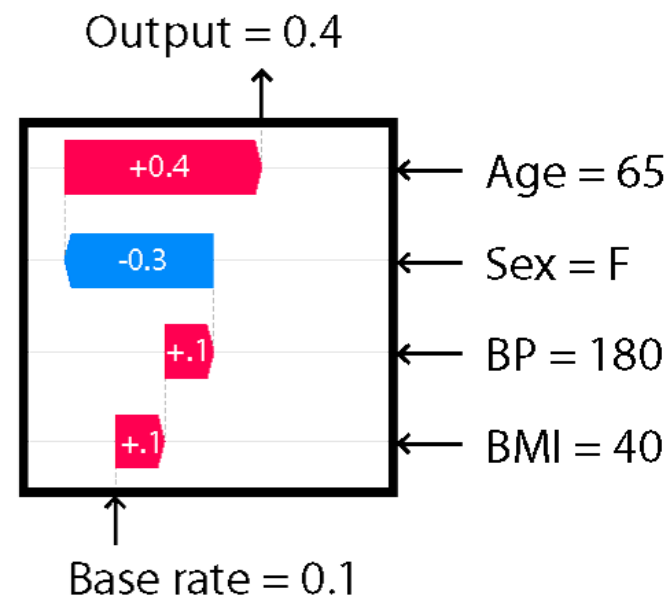
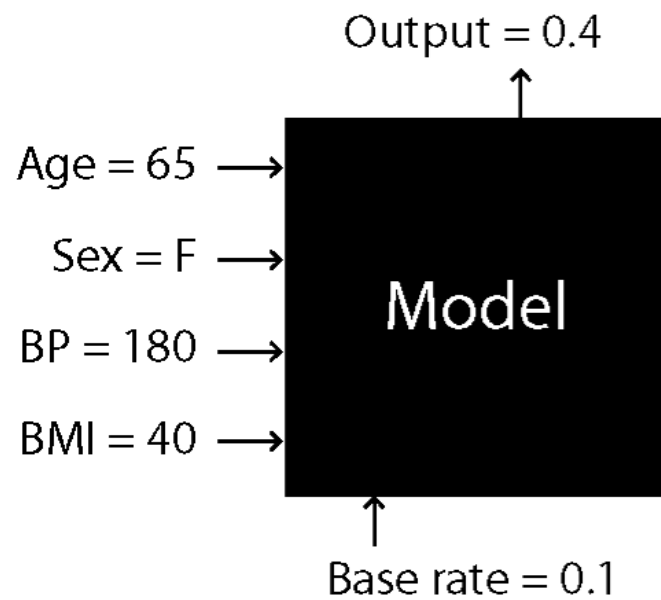


(d) Explaining *Labrador*

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” ($p = 0.21$)

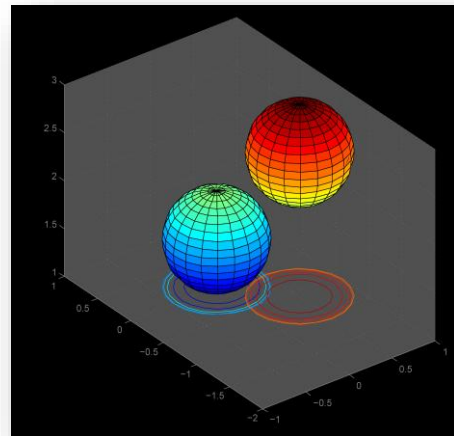


SHAP



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)

Selekcja cech poprzez **regularyzację**



Przez filtrowanie



Przez opakowywanie



Przez osadzanie

Selekcja cech

Regularyzacja

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

norma l_2 :
$$\|\theta\|_2 = \sqrt{\sum_{i=1}^M \theta_i^2} = \sqrt{\theta\theta^T}$$

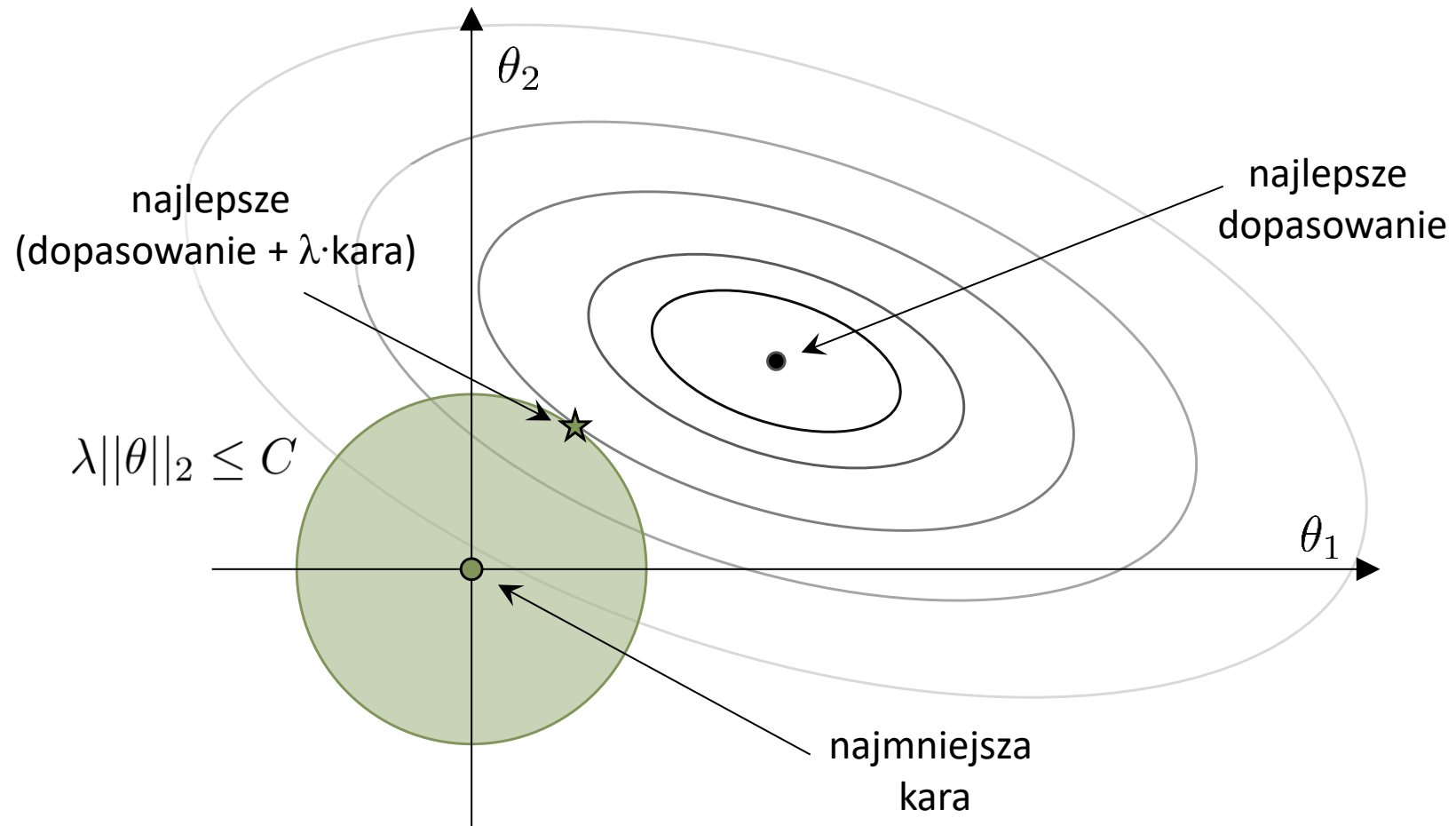
norma l_1 :
$$\|\theta\|_1 = \sum_{i=1}^M |\theta_i|$$

Ocena modelu = błąd dopasowania + λ · kara za złożoność



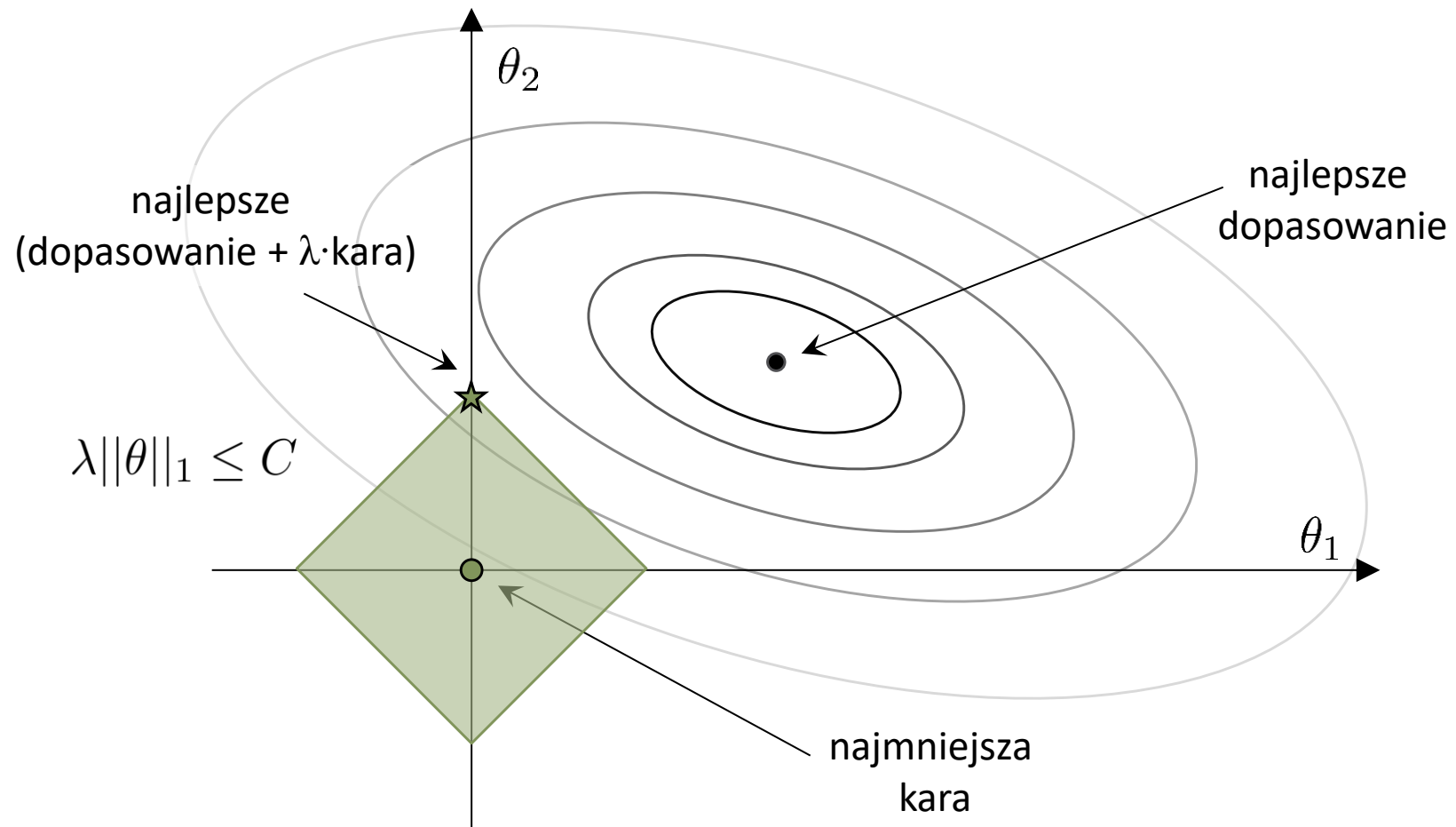
Selekcja cech

Regularyzacja – dlaczego nie l_2



Selekcja cech

Regularyzacja – dlaczego akurat l_1



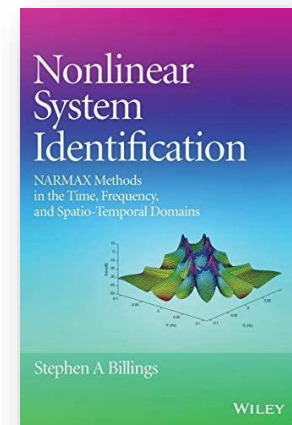
Wydobywanie cech
w uczeniu maszynowym i
uczeniu głębokim

Czym jest uczenie maszynowe

- ang. *machine learning* *learning* ≠ uczenie

I've learned that you are a confident person

- modelowanie statystycznych regularności w obserwacjach / pomiarach
- modelowanie danych vs modelowanie procesu
- dopasowanie modelu do danych (zadanie estymacji)
- **Def.** Proces automatycznego poszukiwania lepszej **reprezentacji danych** w ramach zdefiniowanej przestrzeni możliwości na podstawie sygnału **informacji zwrotnej**
- „pokrewny” termin: identyfikacja systemów (ang. *system identification*)



Czym nie jest uczenie maszynowe

- naśladowanie ludzkich metod i mechanizmów uczenia się
- bo maszyna to

Czym nie jest uczenie maszynowe

- naśladowanie ludzkich metod i mechanizmów uczenia się
- bo maszyna to $f(x)$

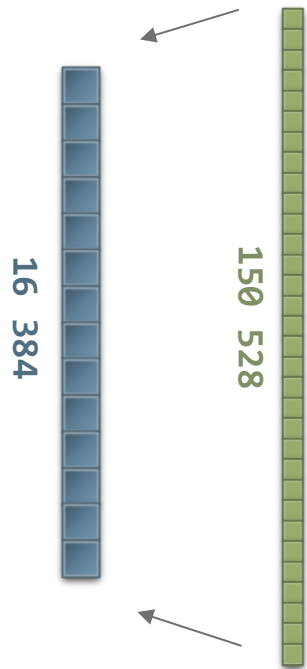
X

150 528



$$\begin{aligned} 224 \times 224 \times 3 \\ = \\ 150\ 528 \end{aligned}$$

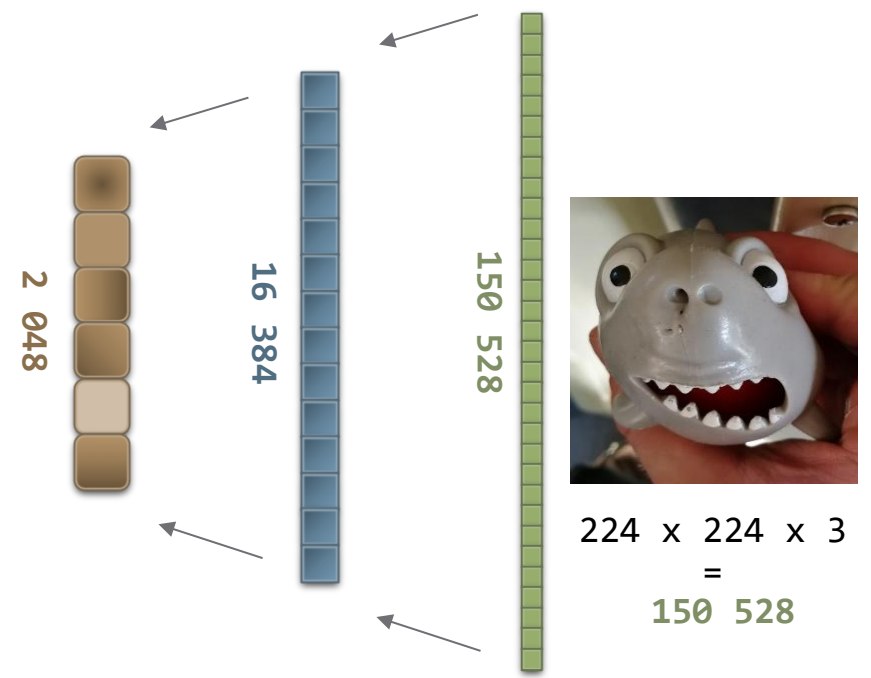
$$\varphi_1(\mathbf{W}_1 \cdot \mathbf{x}) \rightarrow 16\,384 \times 150\,228$$



$$224 \times 224 \times 3 = 150\,528$$

$$\varphi_2(\mathbf{W}_2 \cdot \varphi_1(\mathbf{W}_1 \cdot \mathbf{x}))$$

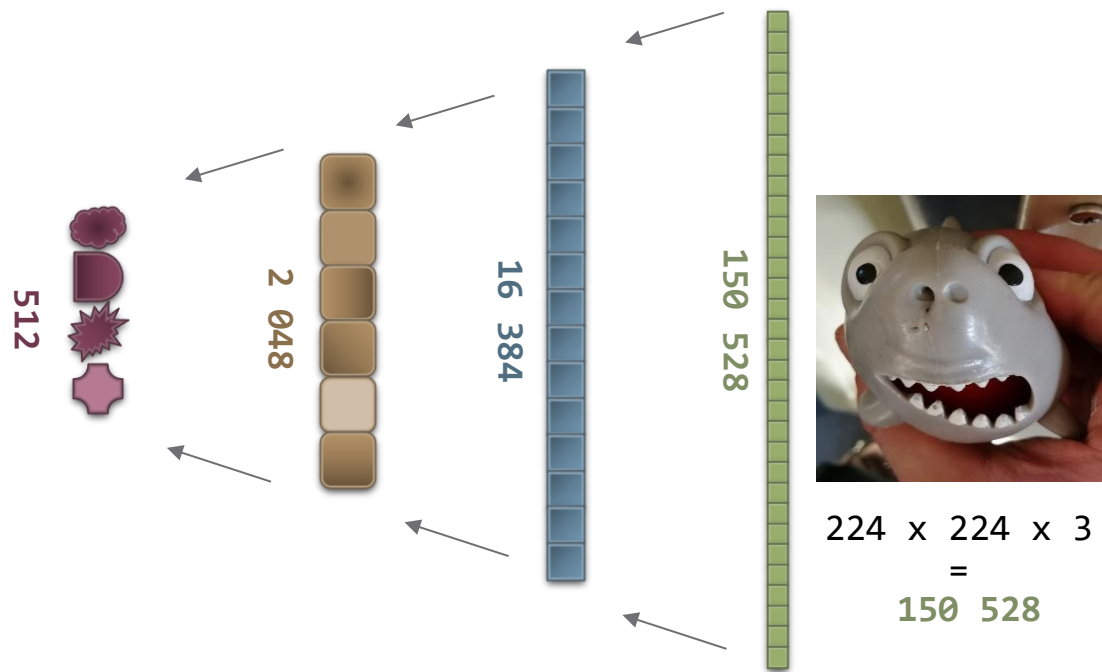
$\rightarrow 2\,048 \times 16\,384$
 $\rightarrow 16\,384 \times 150\,528$

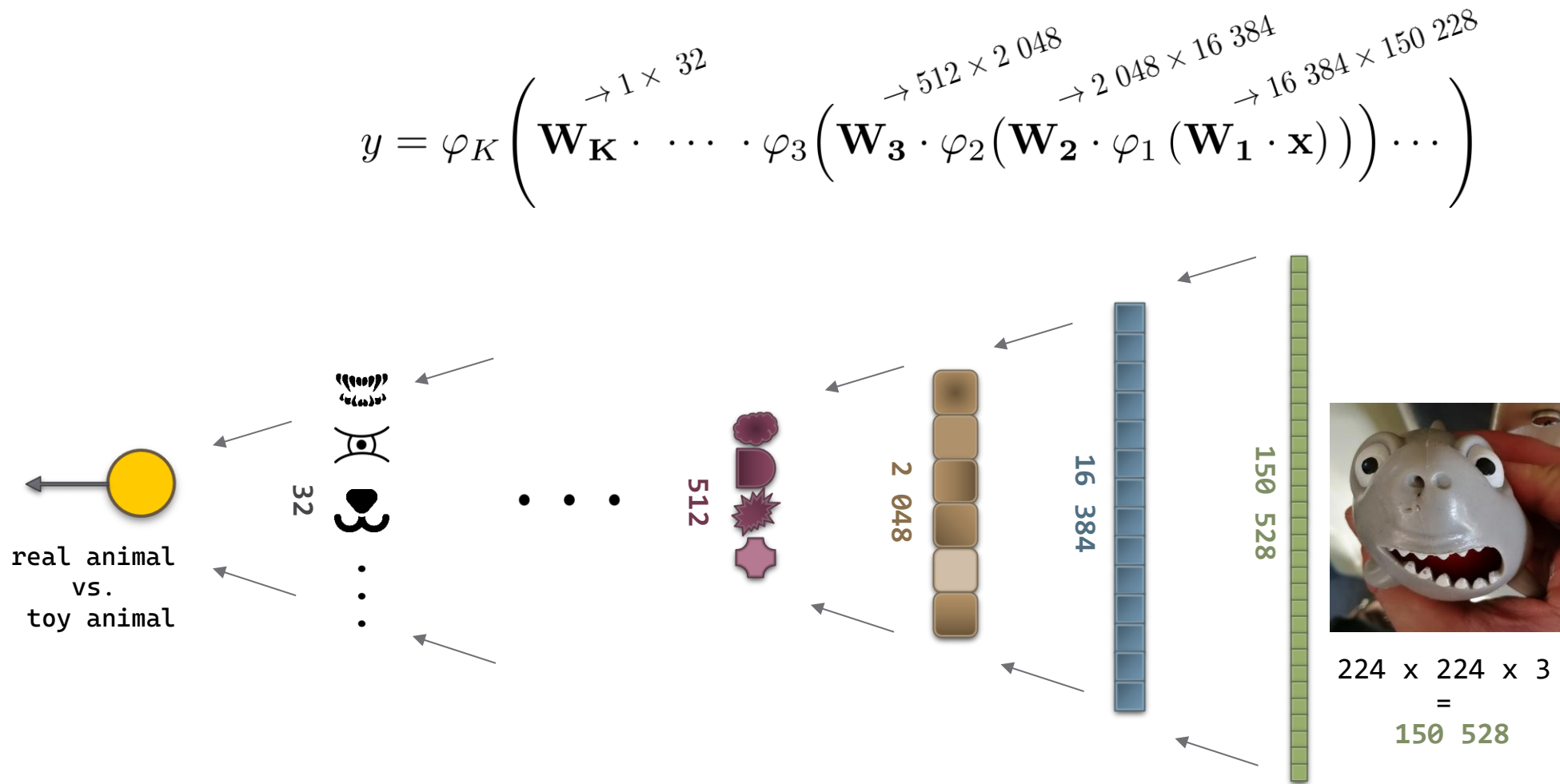


$$224 \times 224 \times 3 = 150\,528$$

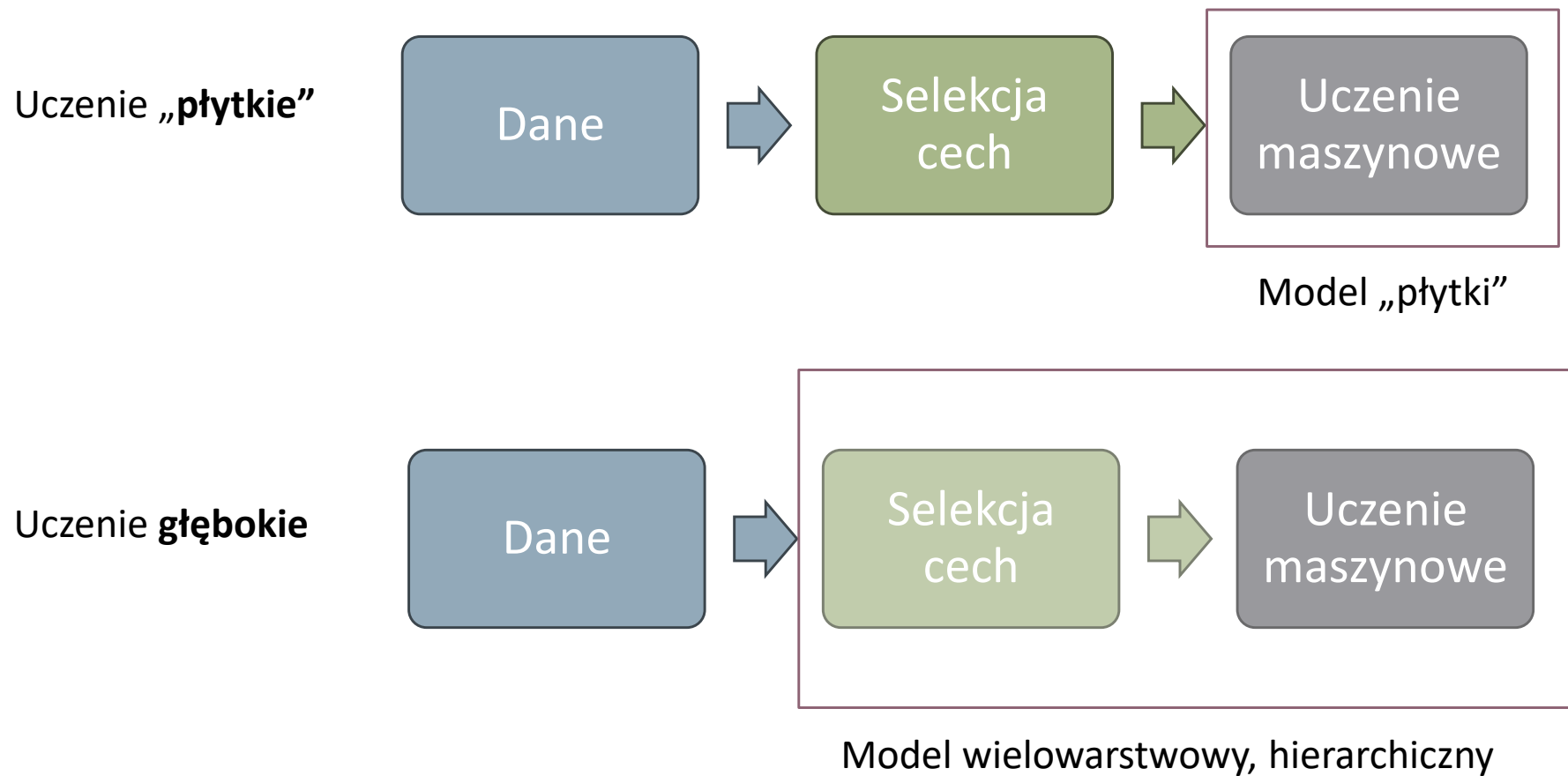
$$\varphi_3 \left(\mathbf{W}_3 \cdot \varphi_2 \left(\mathbf{W}_2 \cdot \varphi_1 \left(\mathbf{W}_1 \cdot \mathbf{x} \right) \right) \right)$$

$\rightarrow 512 \times 2\,048$ $\rightarrow 2\,048 \times 16\,384$ $\rightarrow 16\,384 \times 150\,228$





Uczenie głębokie





THE END