Metody uczenia maszynowego w naukach przyrodniczych

Artur Kalinowski
Uniwersytet Warszawski,
Wydział Fizyki

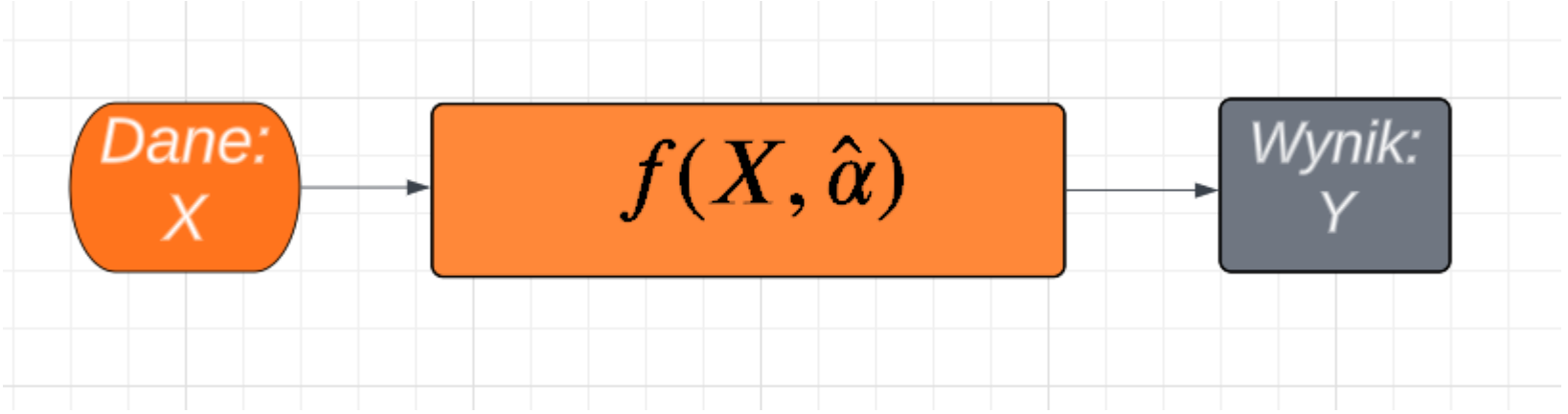
Ogólne zagadnienie w analizie danych

Dane:

$\mathbf{R^n}$ – n-wymiarowa przestrzeń danych wejściowych: X

$\mathbf{R^k}$ – k-wymiarowa przestrzeń danych wyjściowych: Y

Szukane:



Dopasowanie funkcji

- baza w której szukamy współczynników rozwinięcia funkcji opisującej dane

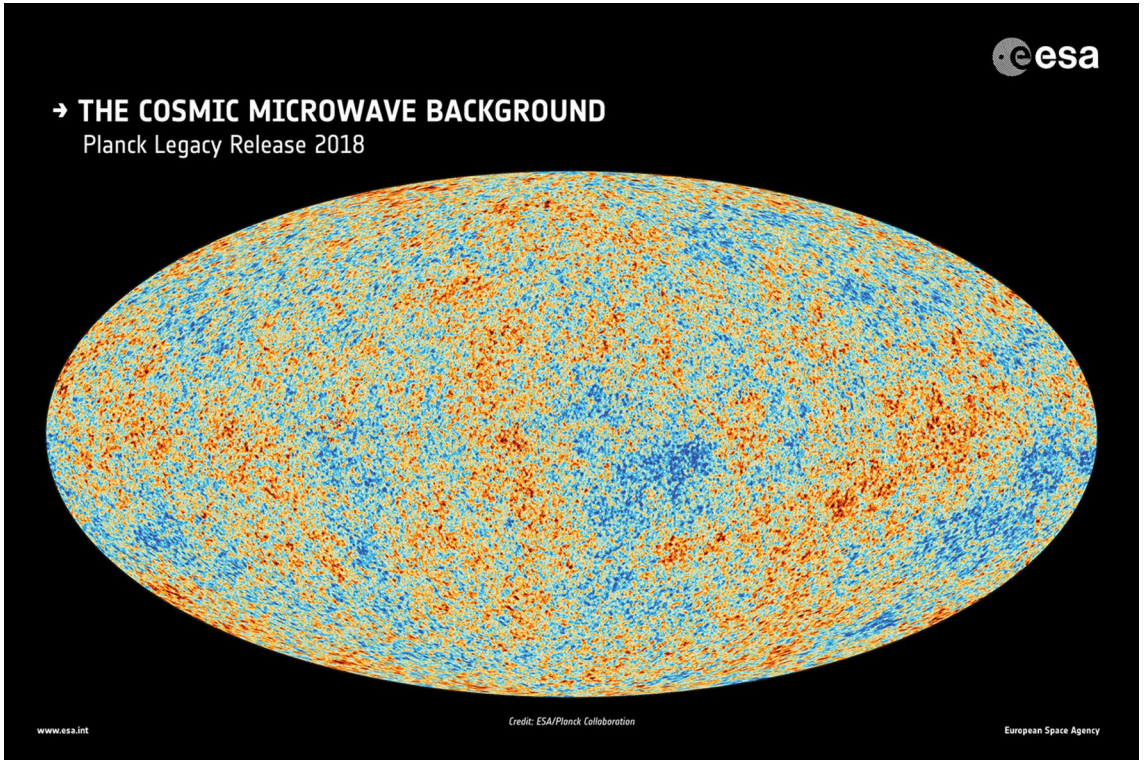
$$f(X, \alpha)$$

- funkcje bazy są określone przez parametry α

Przykład:

- harmoniki sferyczne - baza funkcji zdefiniowanych na sferze:

$$T(\theta, \varphi) = \sum_{l,m} a_{l,m} Y_l^m(\theta, \varphi)$$



ESA/Planck Collaboration

Dopasowanie funkcji

• baza w której szukamy współczynników rozwinięcia funkcji opisującej dane

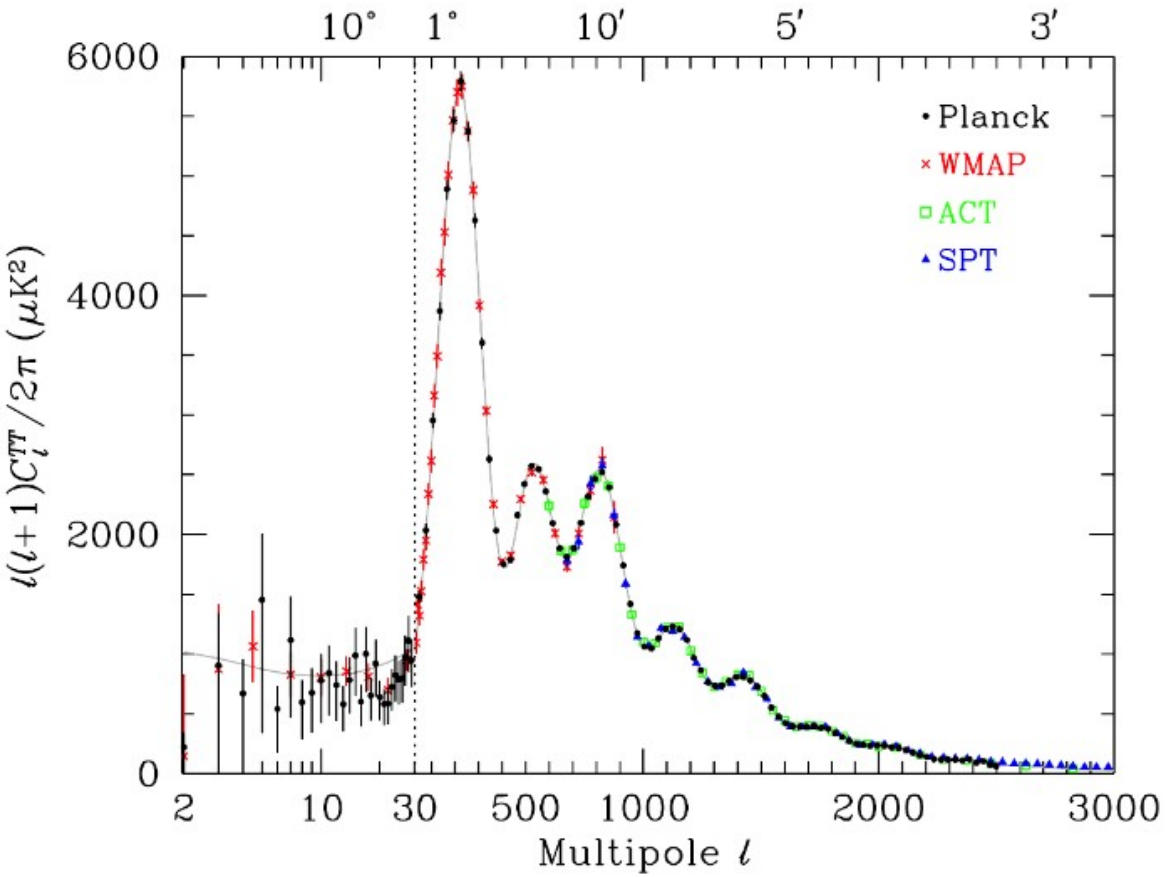
• funkcje bazy są określone przez parametry α

Przykład:

• harmoniki sferyczne - baza funkcji zdefiniowanych na sferze:

$$T(\theta,\varphi) = \sum_{l,m} a_{l,m} Y_l^m(\theta,\varphi)$$

$$f(X,\alpha)$$

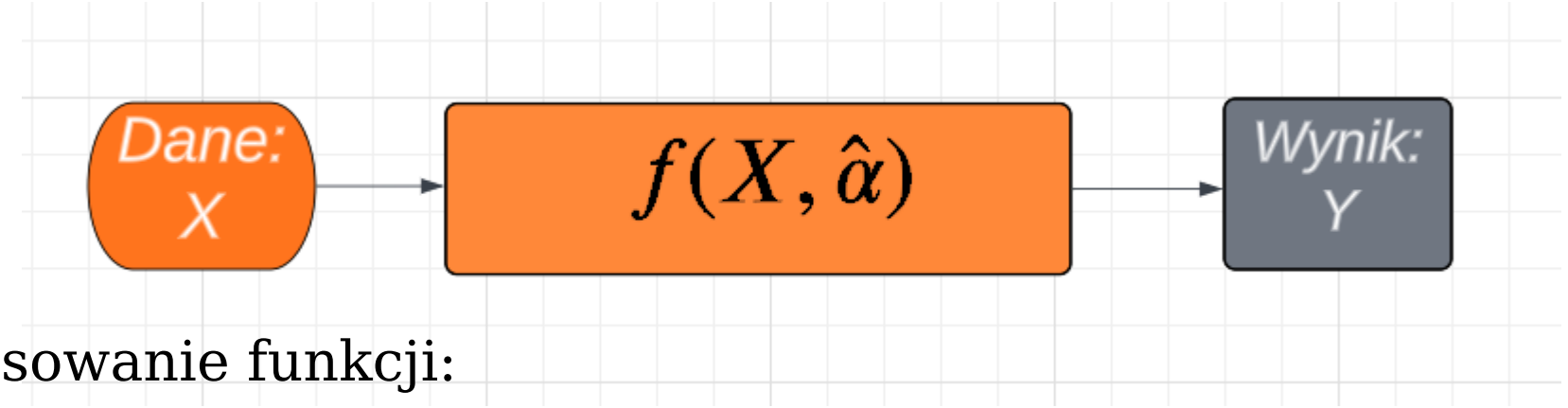


R.L. Workman et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2022, 083C01 (2022) and 2023 update

Ogólne zagadnienie w analizie danych

X: n-wymiarowa przestrzeń danych wejściowych

Y: k-wymiarowa przestrzeń danych wyjściowych



Dopasowanie funkcji:

- **n~2, k~1**
- **postać funkcji bazy zadana *explicite***
- **współczynniki rozwinięcia mogą być interpretowalne**

Ogólne zagadnienie w analizie danych

X: n-wymiarowa przestrzeń danych wejściowych

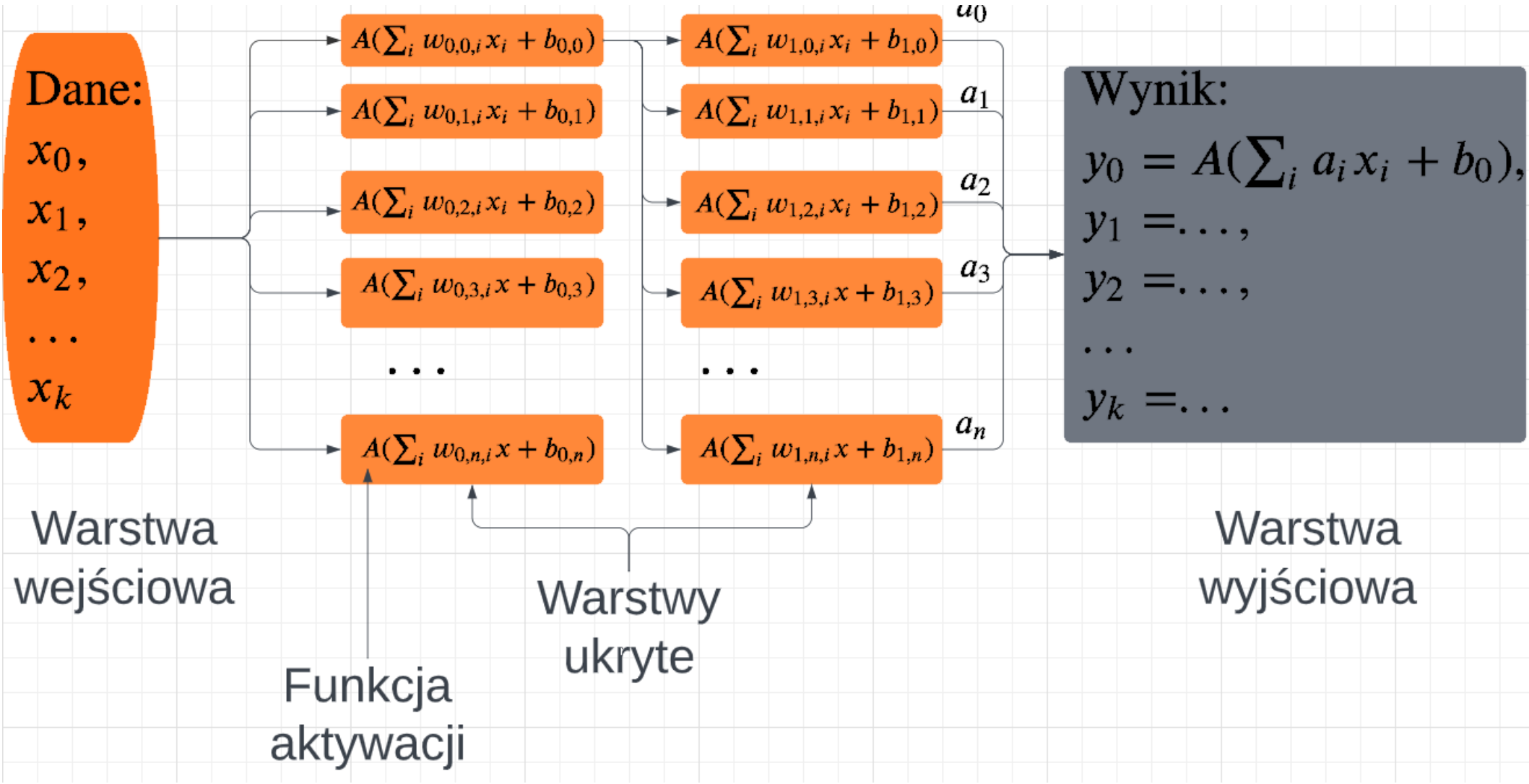
Y: k-wymiarowa przestrzeń danych wyjściowych

Uczenie maszynowe:

- **$n \sim 10^l$, $k \sim 10^m$, $l,m \sim 6$**
- **baza funkcji zdefiniowana *implicite* przez strukturę przepływu danych – architekturę**
- **współczynniki rozwinięcia nie są interpretowalne**

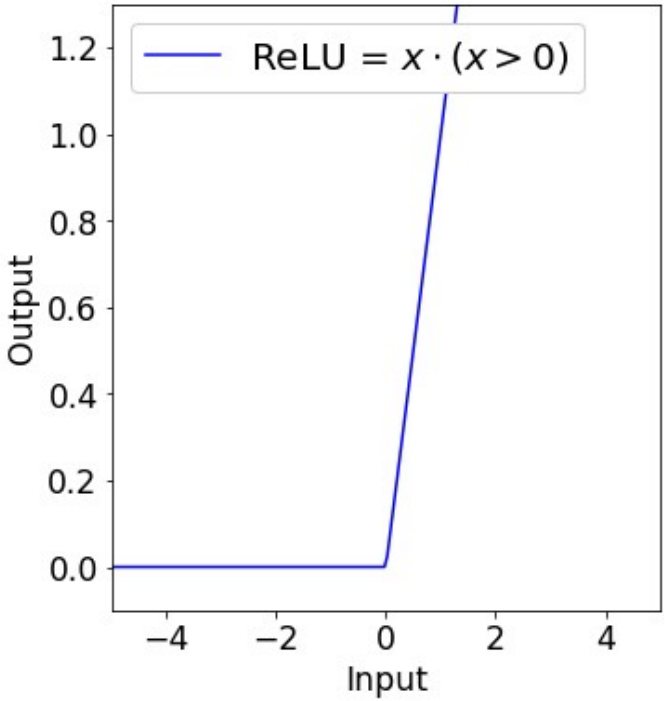
Architektura: „fully connected”

Klasyczna sieć neuronowa

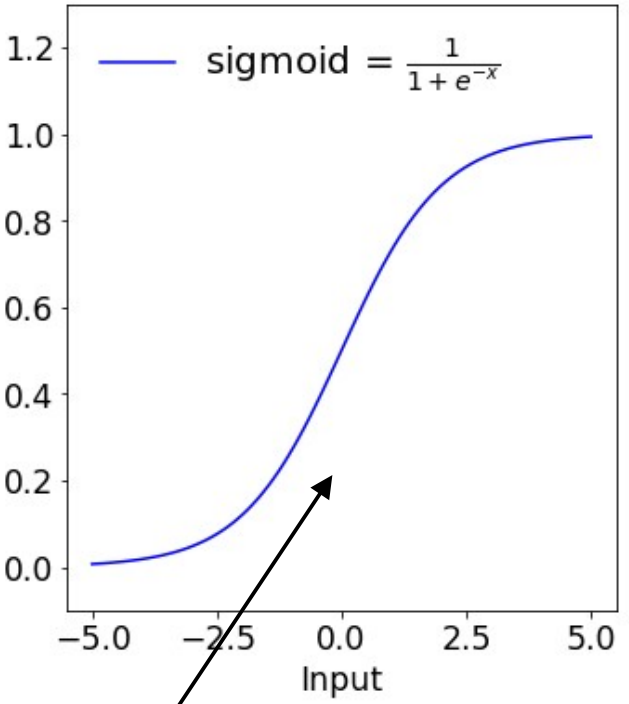


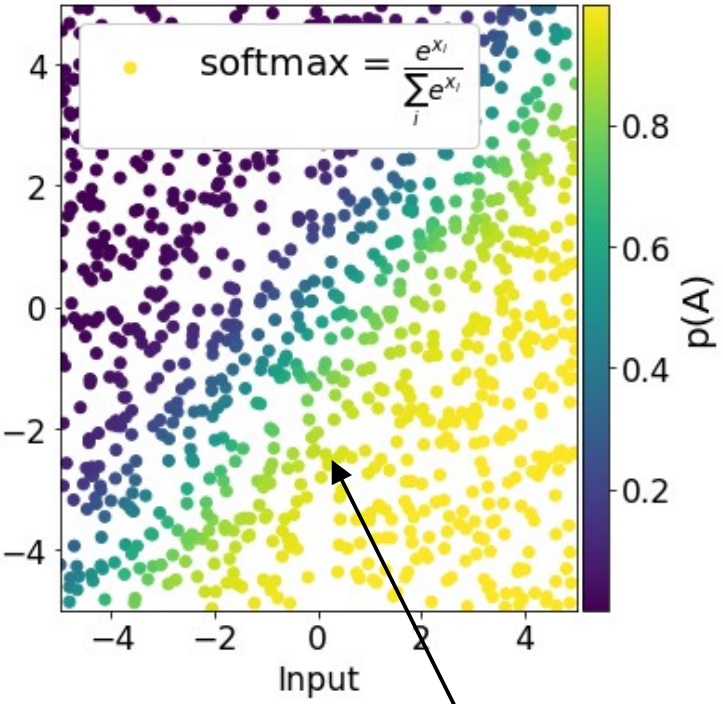
Funkcja aktywacji

standard dla warstw ukrytych



standard dla warstwy wyjściowej:



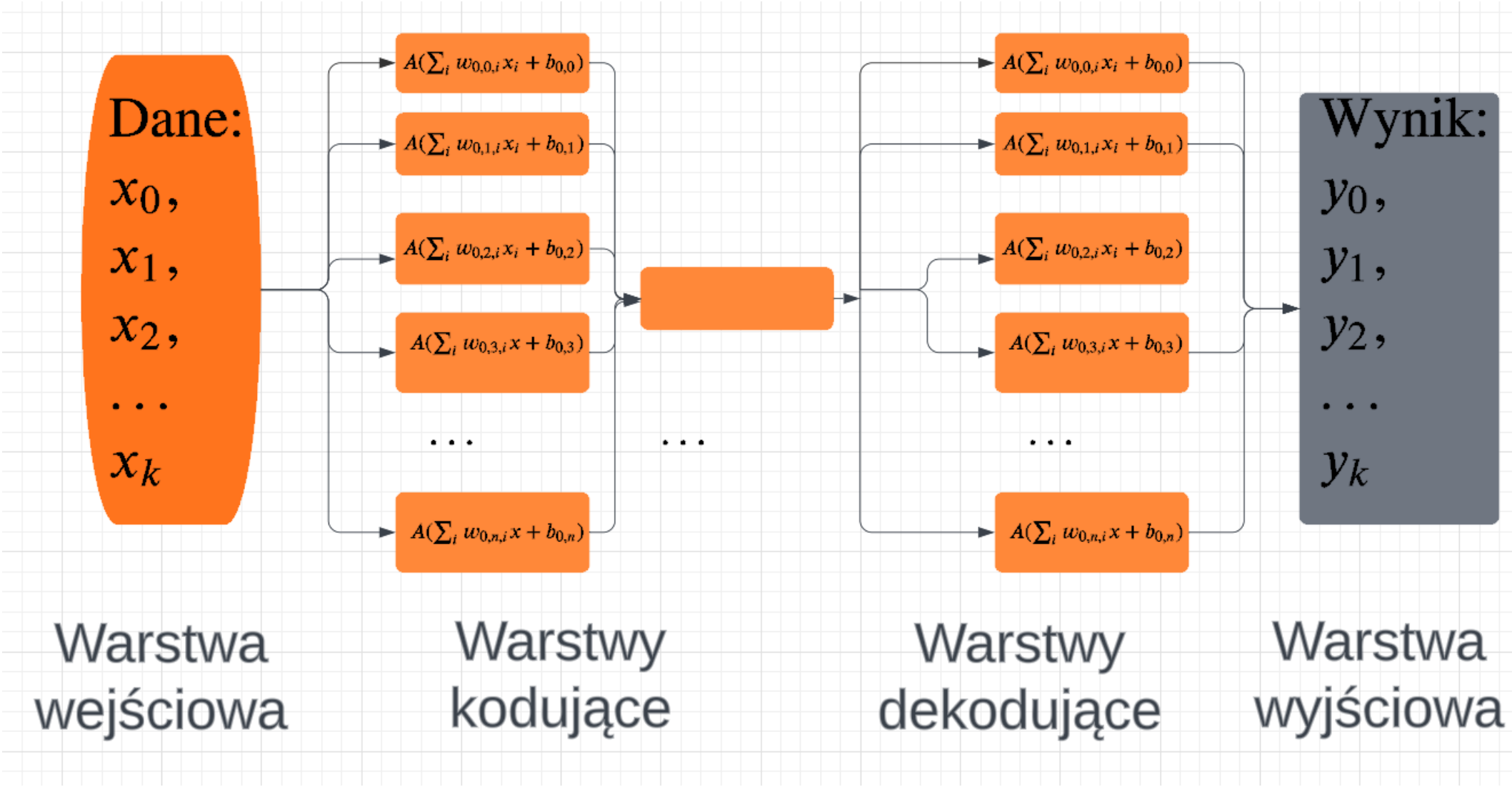


pojedyncze prawdopodobieństwo

prawdopodobieństwo dla wielu wariantów

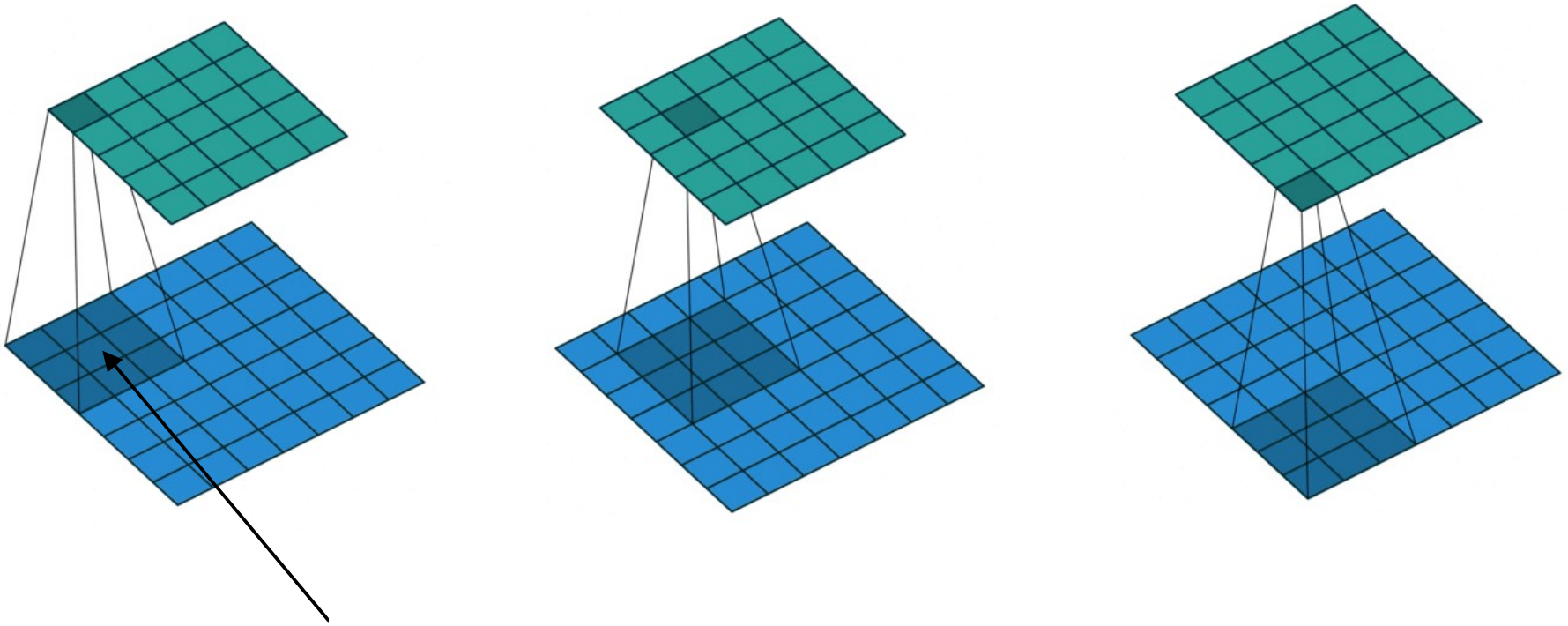
Architektura: „autoencoder”

Sieć kodująca



Architektura: „convolution”

Sieć neuronowa ze splotem



dane z okienka 3x3 są sumowane z tymi samymi wagami dla calego obrazu

3_0	3_1	2_2	1	0
0_2	0_2	1_0	3	1
3_0	1_1	2_2	2	3
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

3	3_0	2_1	1_2	0
0	0_2	1_2	3_0	1
3	1_0	2_1	2_2	3
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0

arXiv:1603.07285 [stat.ML]

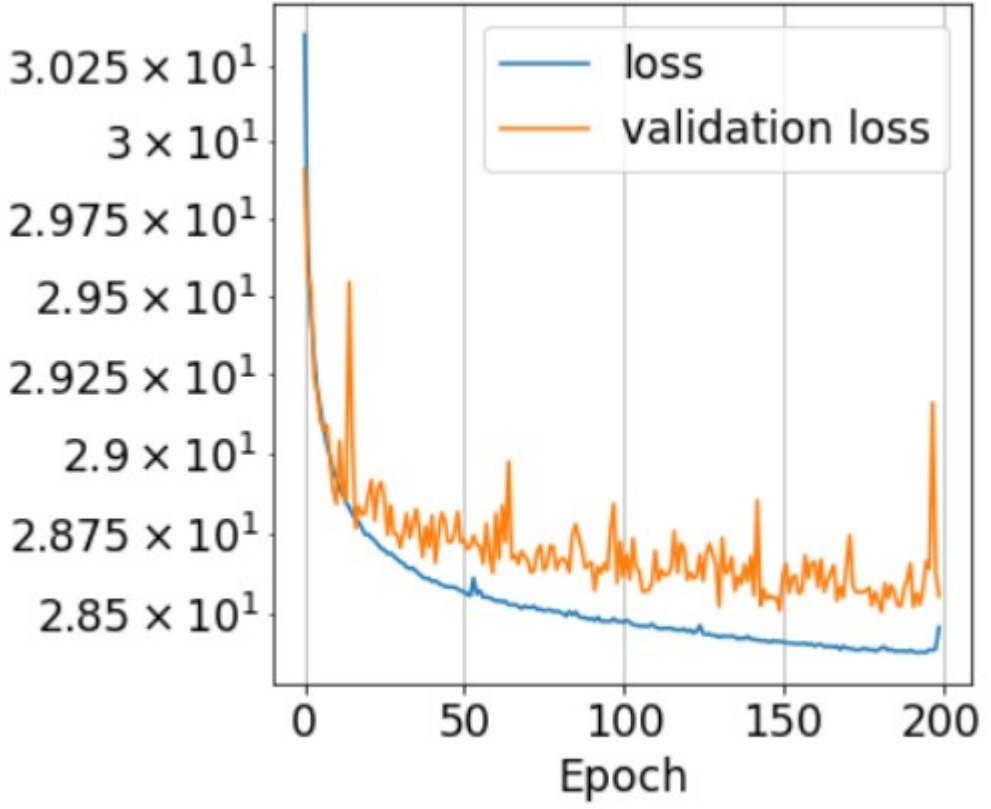
Trening modelu

- parametry znajdowane w iteracyjnym procesie minimalizacji **funkcji kosztu:**

$$L\left(f\left(X,\alpha\right),Y\right)$$

$$\alpha = argmin_{\alpha} L\left(f\left(X,\alpha\right),Y\right)$$

- znajdowanie parametrów modelu jest określane mianem **treningu**



Funkcje straty

- **zagadnienie regresji:** f(X) zwraca dowolną, ciągłą, wartość

$$L\left(f\left(X,\alpha\right),Y\right)=\frac{1}{N}\sum_i\left(f\left(X,\alpha\right)-Y\right)^2+\lambda J\left(\alpha\right)$$

- **zagadnienie klasyfikacji:** f(X) zwraca prawdopodobieństwo

$$L\left(f\left(X,\alpha\right),Y\right)=$$
$$\frac{1}{N}\sum_i\log\left(f_{poprawna\,klasa}\left(X,\alpha\right)\right)+\ldots$$

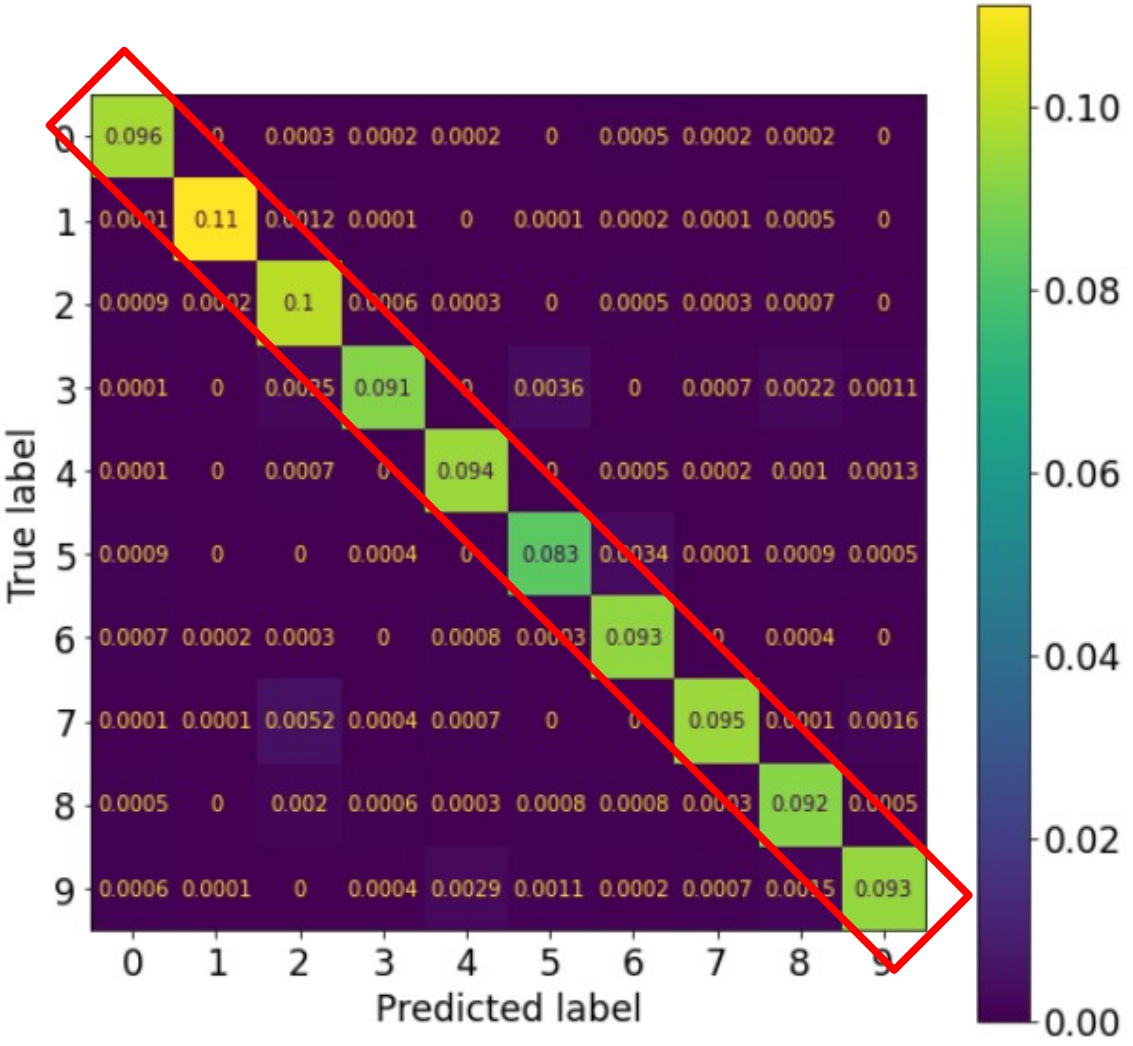
Metryki

• poprawność modeli klasyfikacyjnych jest szacowana przy użyciu macierzy pomyłek, oraz **metryk –** liczb opisujących zgodność modelu z danymi, np.:

$$\epsilon = \frac{poprawne}{normalizacja}$$

• **accuracy:**

normalizacja - wszystkie przykłady



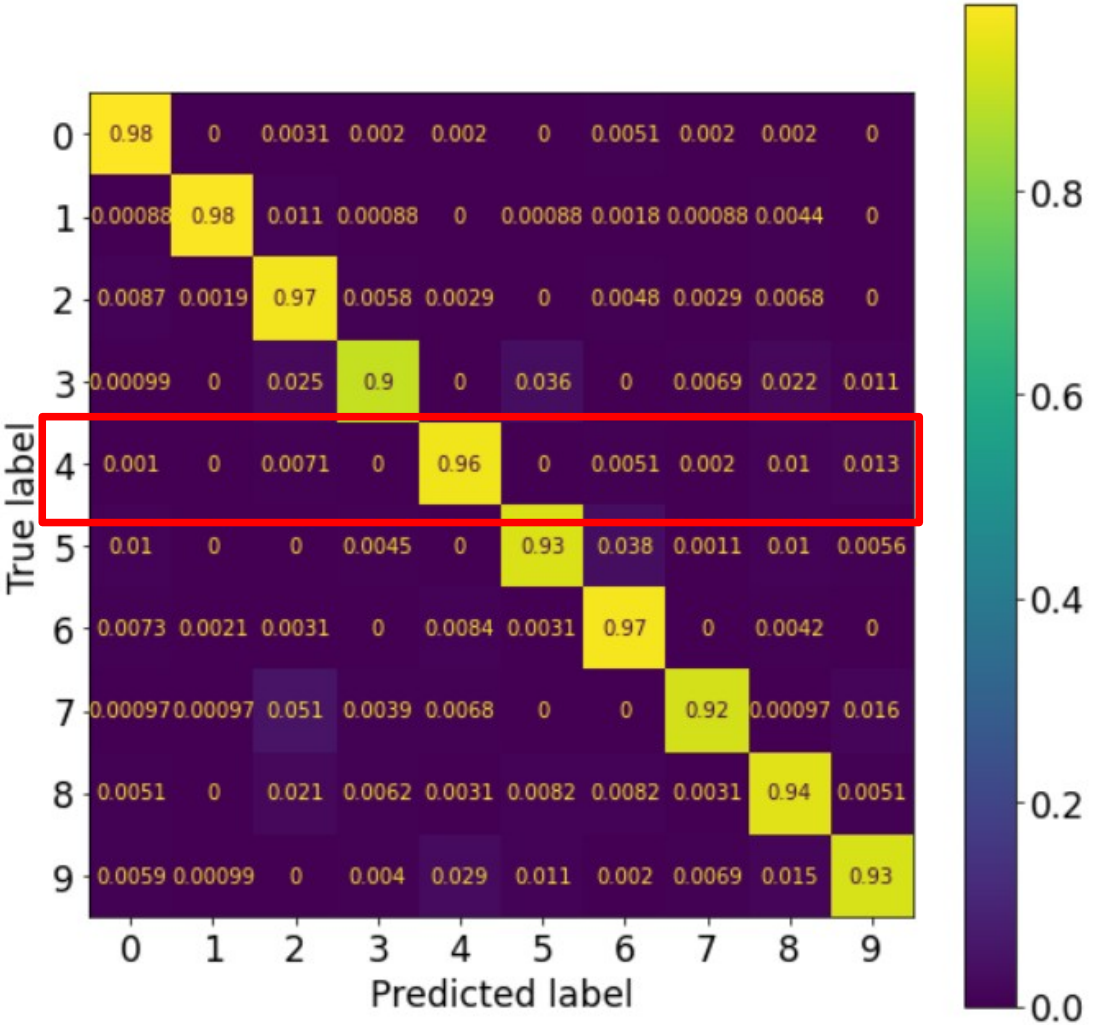
Metryki

• poprawność modeli klasyfikacyjnych jest szacowana przy użyciu macierzy pomyłek, oraz **metryk –** liczb opisujących zgodność modelu z danymi, np.:

$$\epsilon = \frac{poprawne}{normalizacja}$$

• **recall:**

normalizacja - przykłady z danej klasy



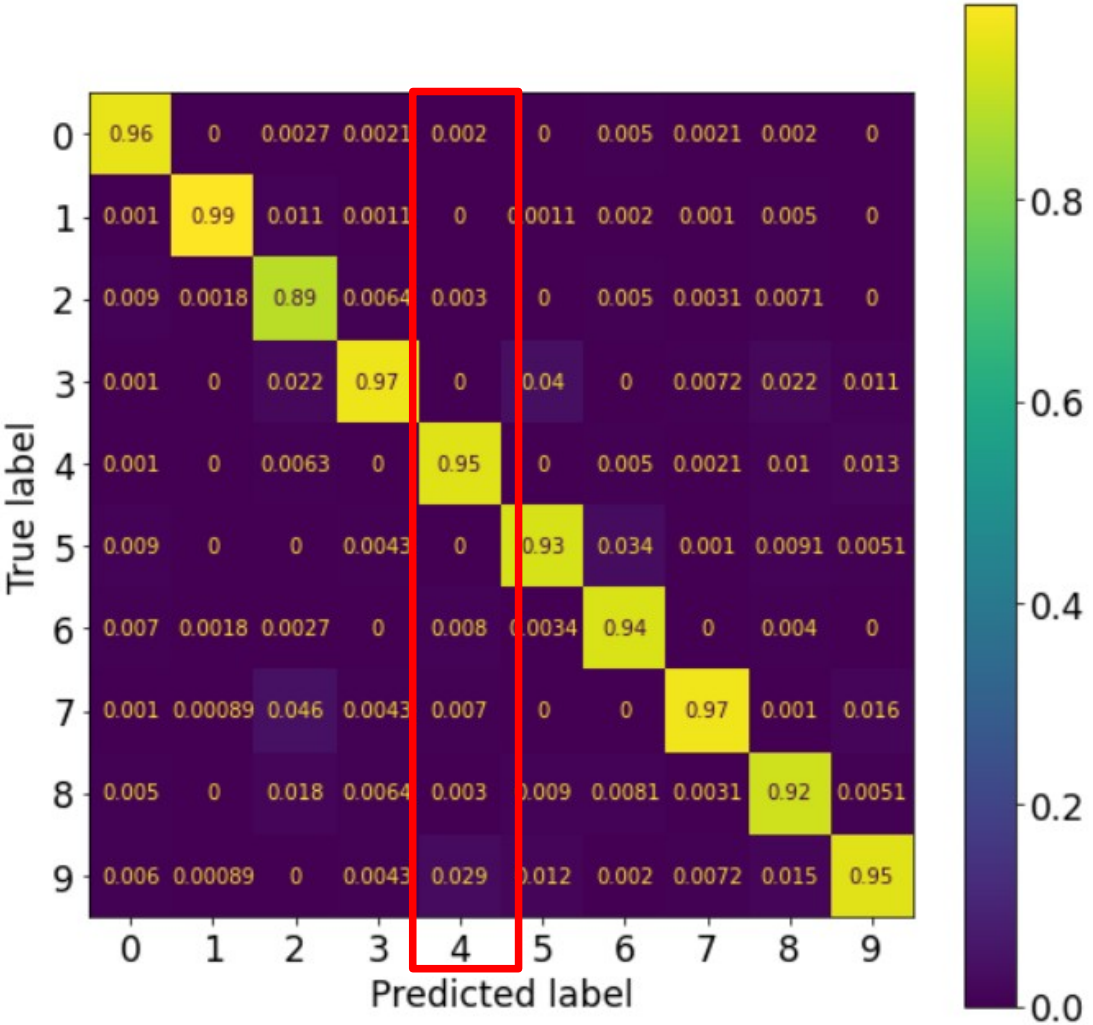
Metryki

• poprawność modeli klasyfikacyjnych jest szacowana przy użyciu macierzy pomyłek, oraz **metryk –** liczb opisujących zgodność modelu z danymi, np.:

$$\epsilon = \frac{poprawne}{normalizacja}$$

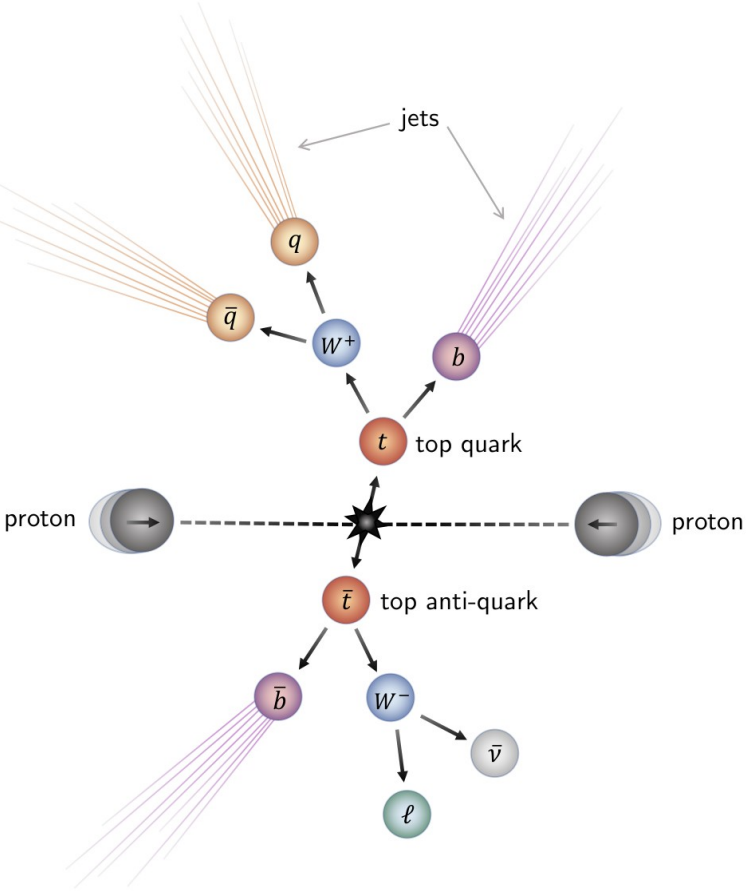
• **precision:**

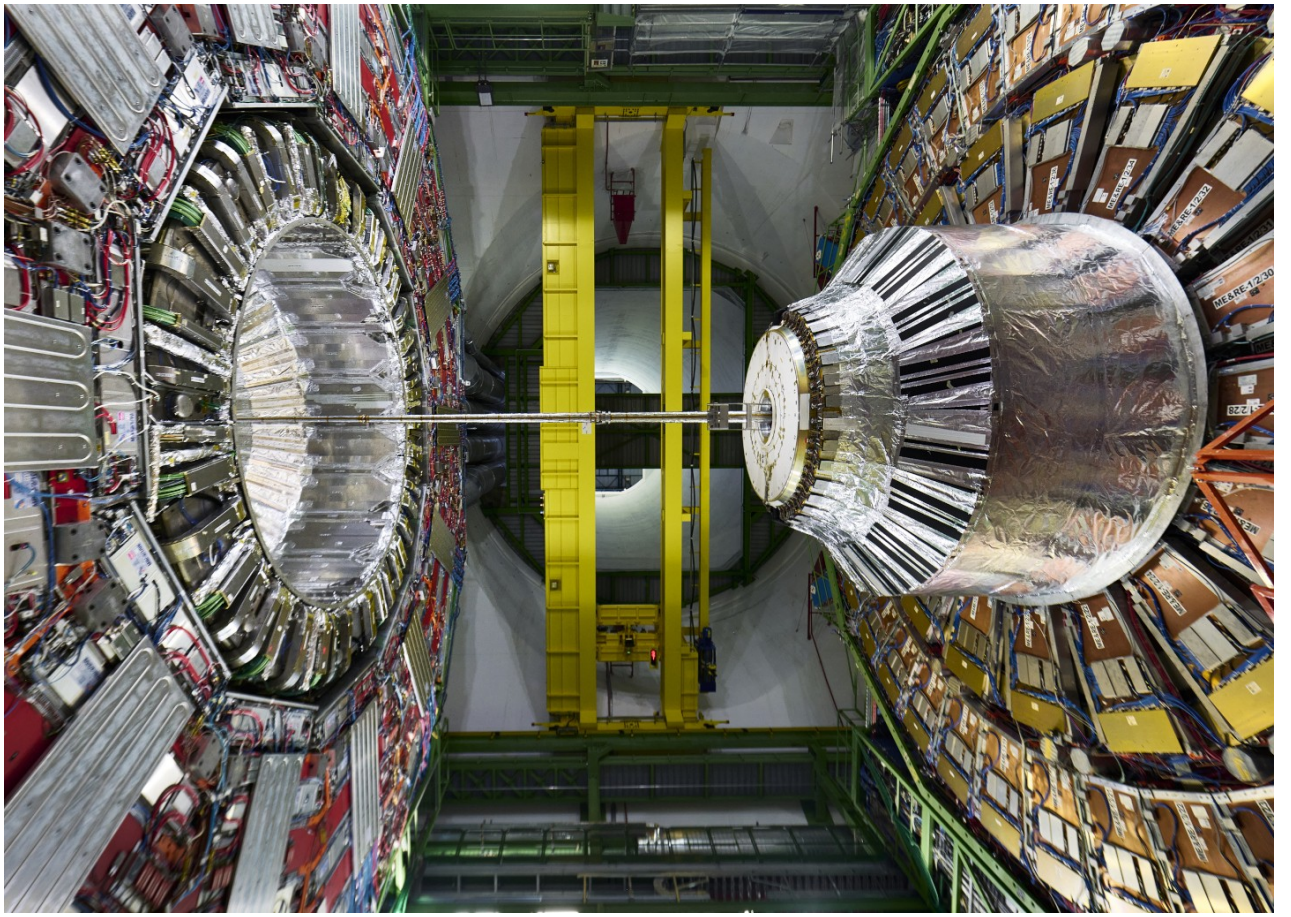
normalizacja - przykłady którym model przyznał daną klasę



Przykład: fizyka cząstek elementarnych

Eksperyment CMS działający przy LHC (CERN)



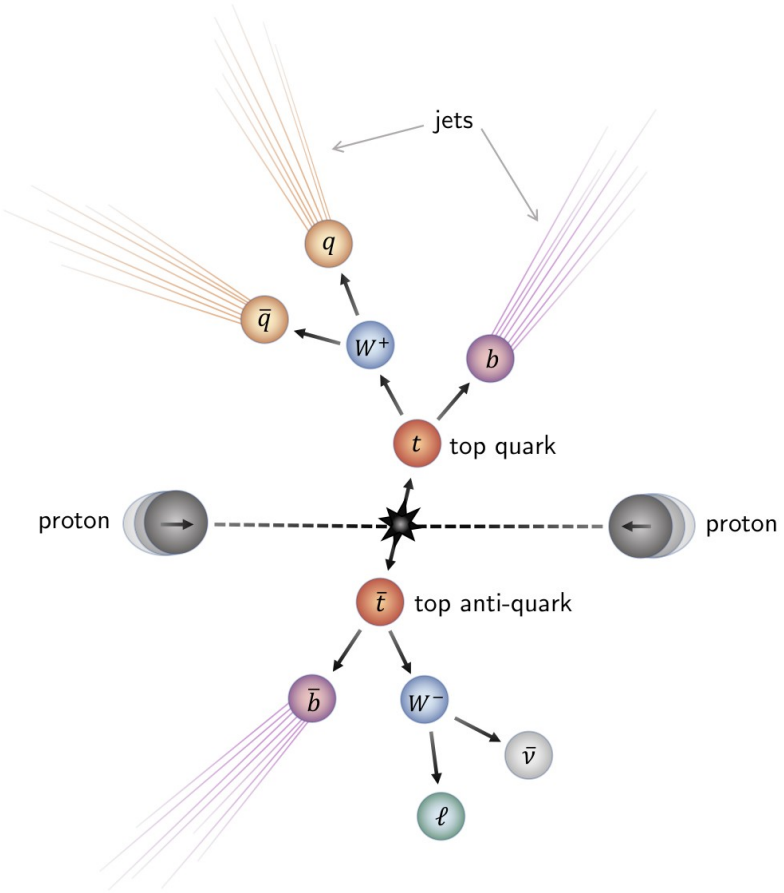


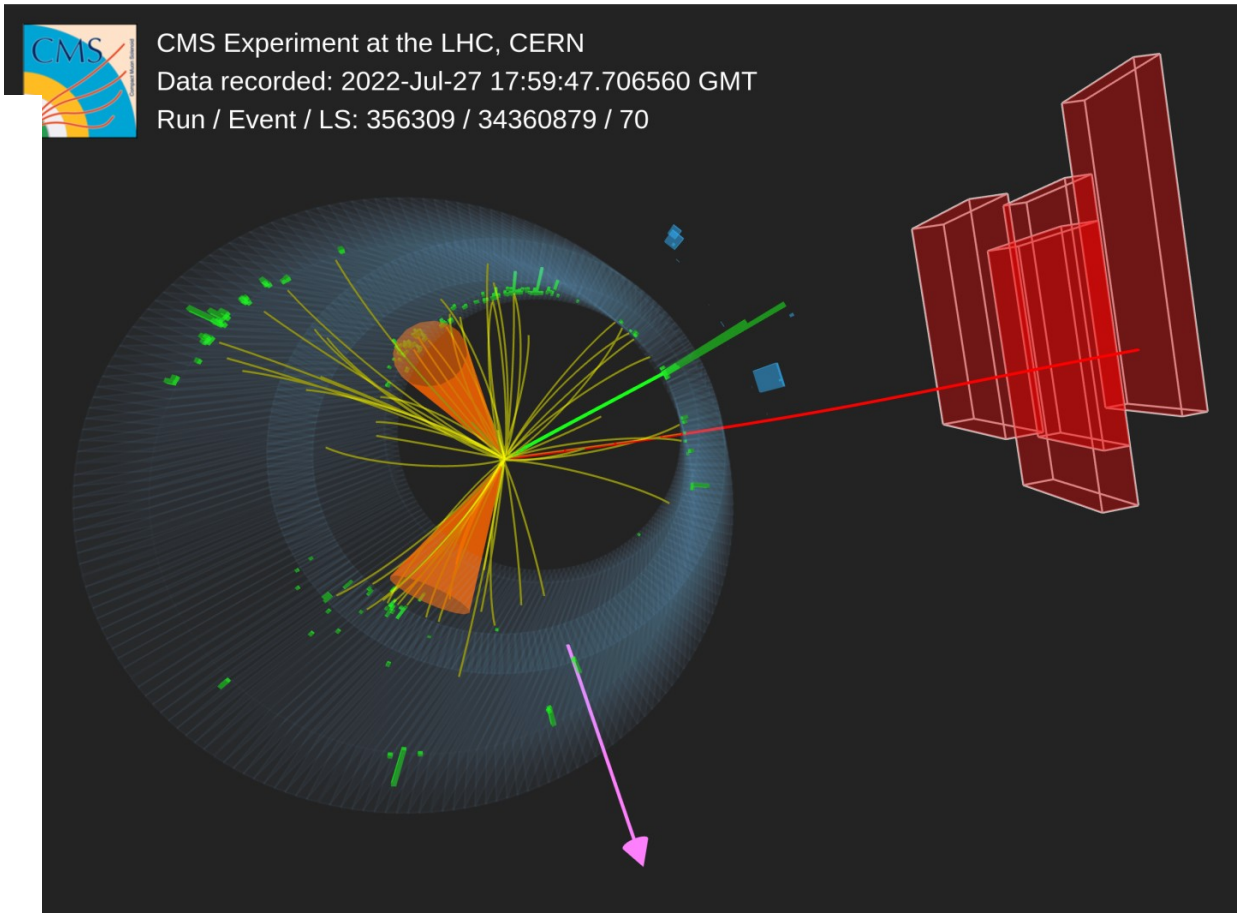
CERN/CMS Collaboration

Schemat zderzenia proton-proton

Przykład: fizyka cząstek elementarnych

Eksperyment CMS działający przy LHC (CERN)





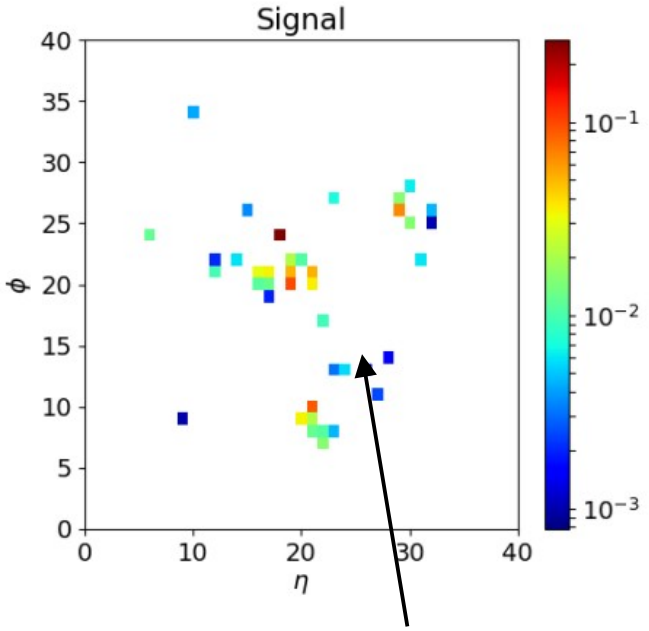
CERN/CMS Collaboration

Schemat zderzenia proton-proton

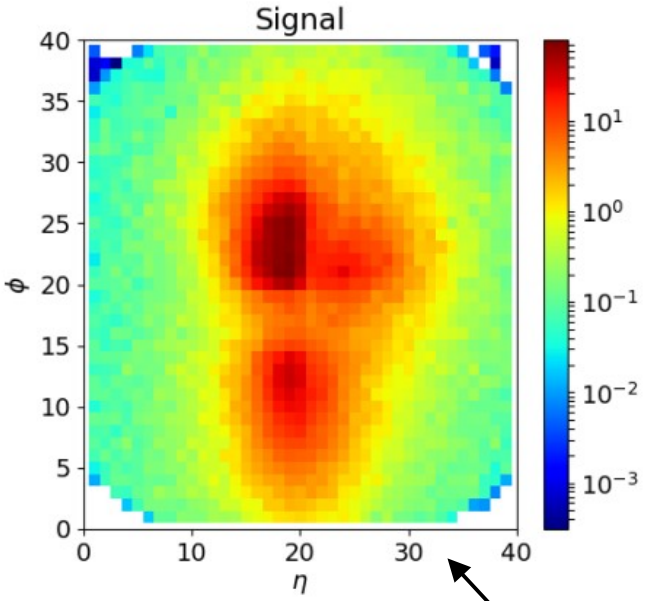
Przykład: fizyka cząstek elementarnych

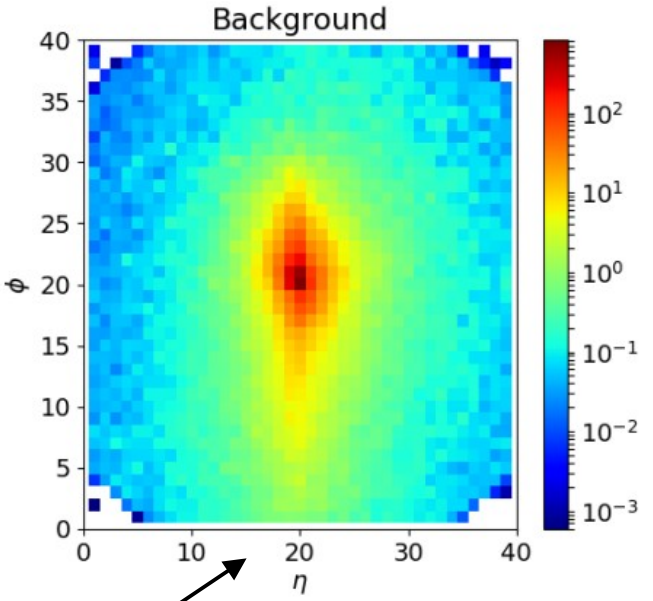
Zadanie: klasyfikacja “obrazów” zderzeń: kwark top/coś innego

Dane wejściowe: obrazy dżetów



pojedyncze zderzenie proton-proton



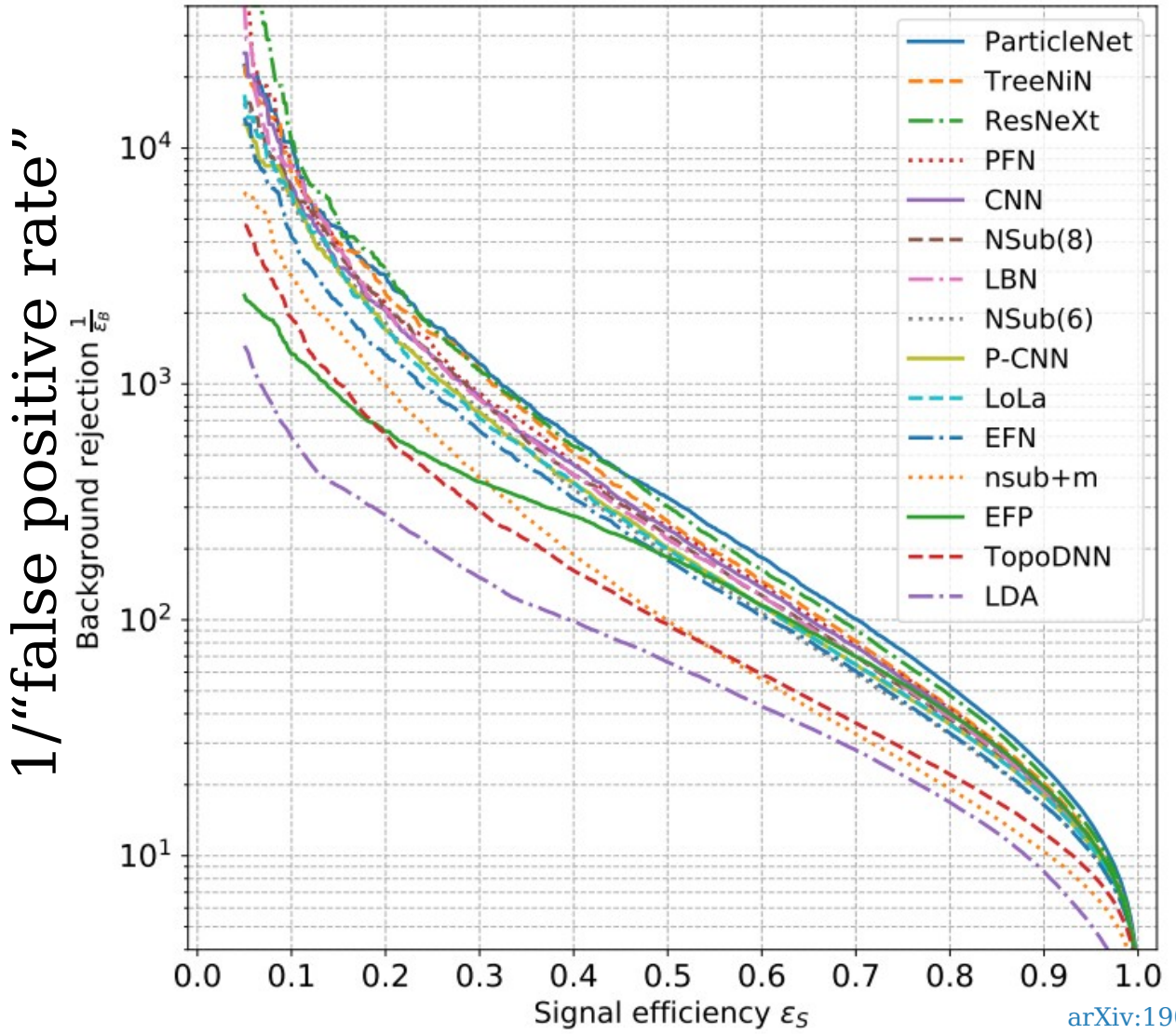


arXiv:1902.09914 [hep-ph]

Obrazy uśrednione po 10 000 zderzeń proton-proton

Przykład: fizyka cząstek elementarnych

Zadanie: klasyfikacja “obrazów” zderzeń: kwark top/coś innego



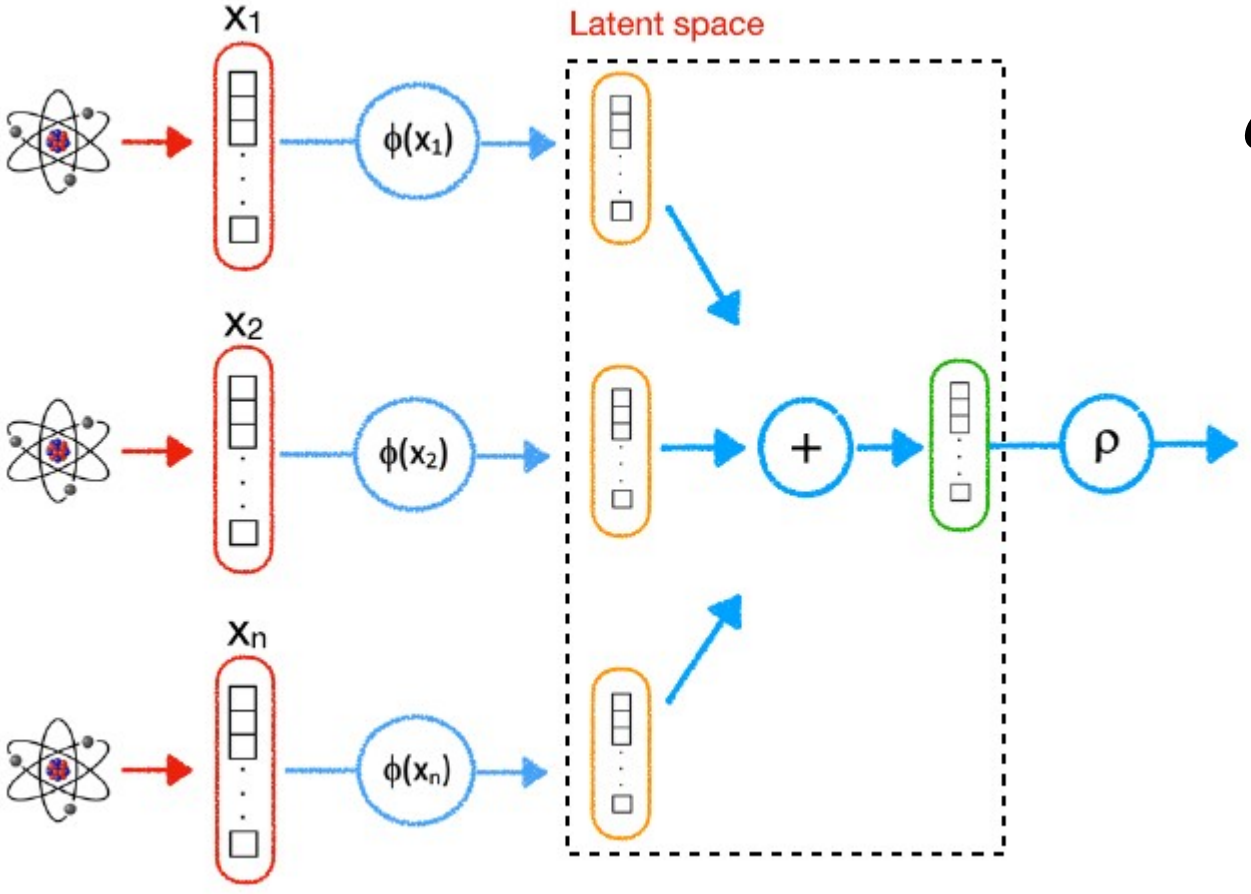
arXiv:1902.09914 [hep-ph]

Przykład: fizyka materiałowa

Zadanie: klasyfikacja związków: nadprzewodzący? **TAK/NIE**

Dane wejściowe:

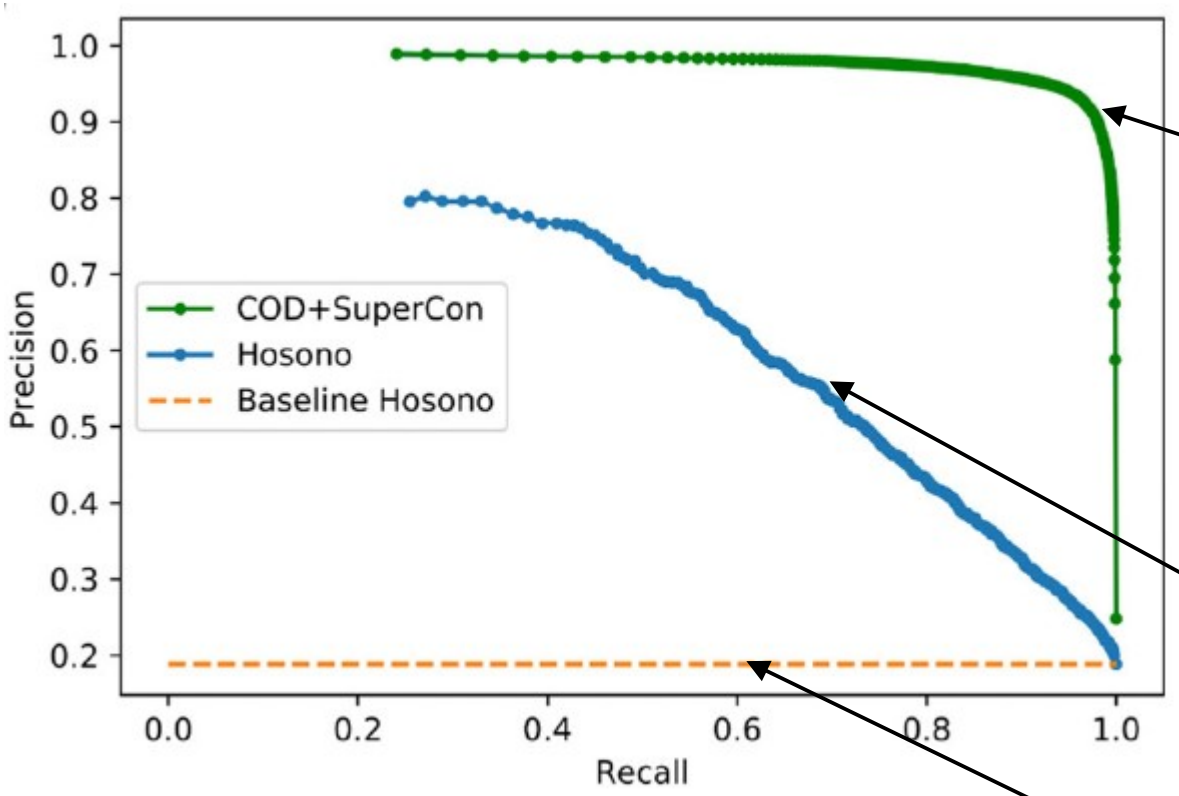
- zestaw pierwiastków tworzących analizowany związek,
  - 22 cechy na pierwiastek + współczynnik stochiometryczny



$\phi(x), \rho(x)$ - sieci neuronowe

K., Guizouarn, T. et al. From individual elements to rials: in search of new superconductors via machine put Mater 9, 71 (2023). 1038/s41524-023-01023-6

Przykład: fizyka materiałowa



zbiór o strukturze identycznej ze zbiorem użytym do treningu treningu

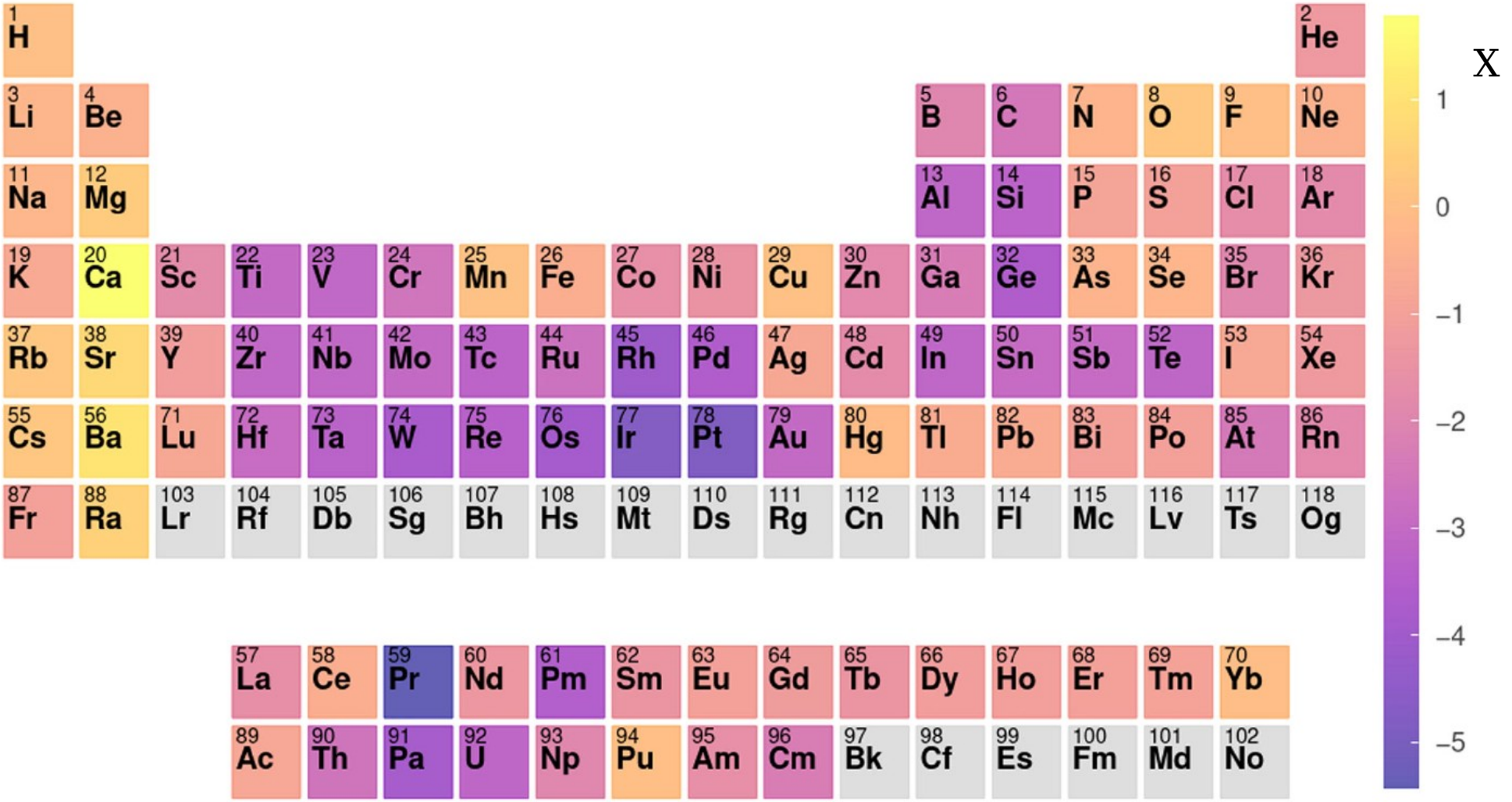
niezależny zbiór, nie używany do trenowania

zbiór Hosono zawiera 20% nadprzewodników

Pereti, C., Bernot, K., Guizouarn, T. et al. From individual elements to macroscopic materials: in search of new superconductors via machine learning. npj Comput Mater 9, 71 (2023). https://doi.org/10.1038/s41524-023-01023-6

Przykład: fizyka materiałowa

Mapa wpływu na zwiększenie temperatury krytycznej:
duże X → zwiększona T_c



Pereti, C., Bernot, K., Guizouarn, T. et al. From individual elements to macroscopic materials: in search of new superconductors via machine learning. npj Comput Mater 9, 71 (2023). https://doi.org/10.1038/s41524-023-01023-6

Podsumowanie

Współczesne uczenie maszynowe ma swoje korzenie w metodach statystycznej analizy danych rozwijanych w drugiej połowie XX wieku.

Błyskawiczny rozwój metod obliczeniowych pozwala na zastosowanie znanego podejścia – „dopasowanie funkcji” w całkiem nowych kontekstach