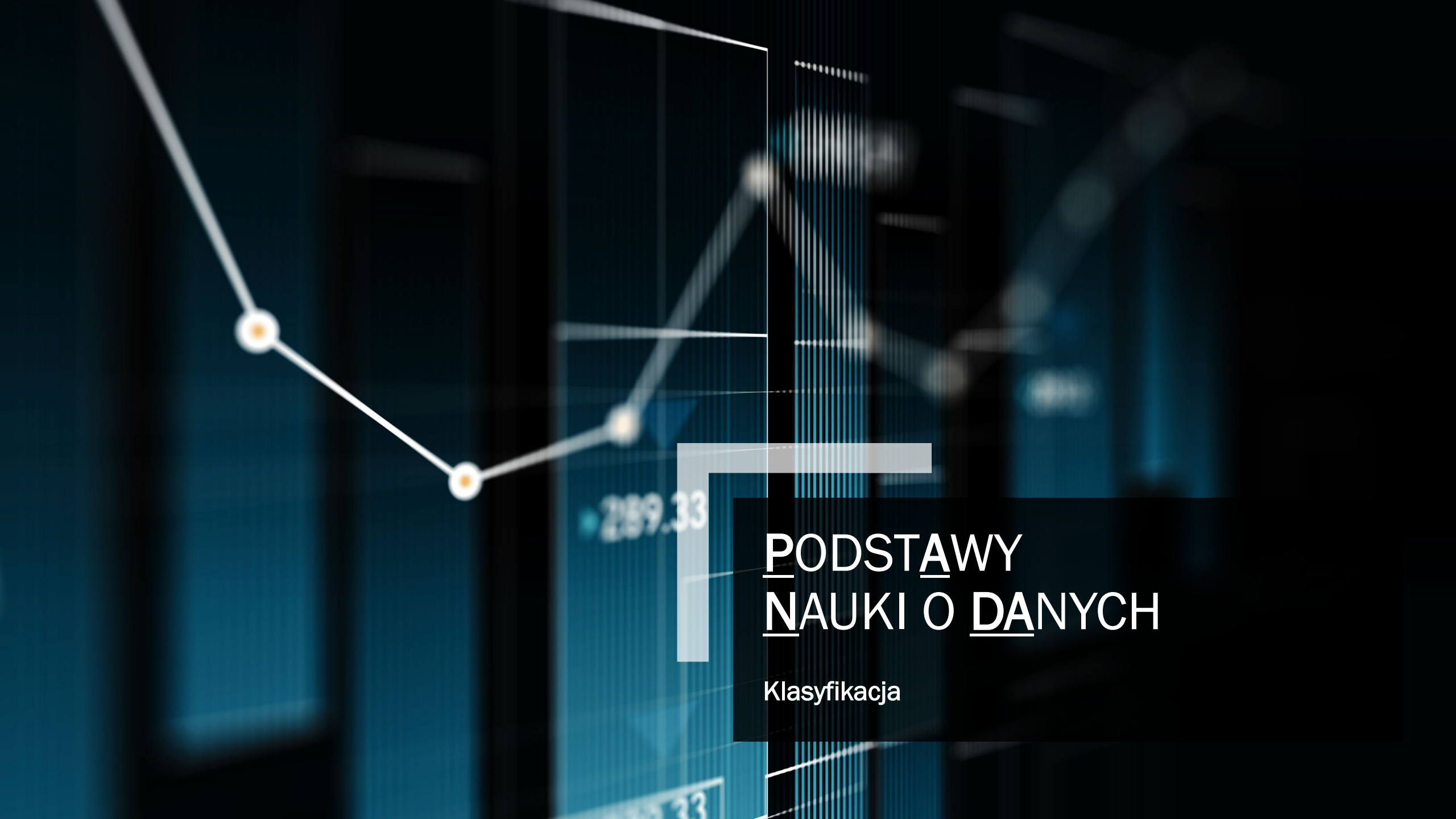
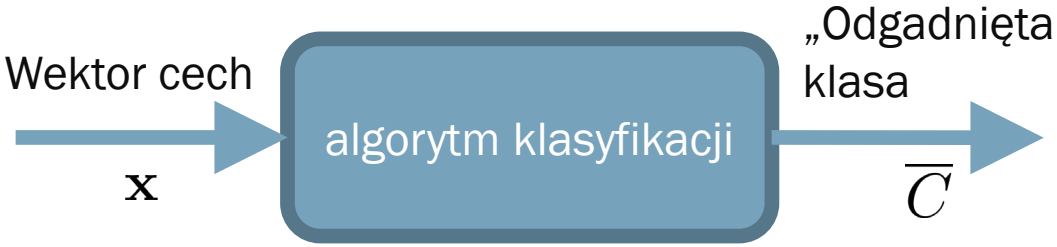


PODSTAWY NAUKI O DANYCH

Klasyfikacja

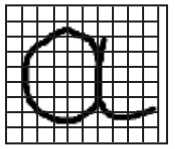
Klasyfikacja



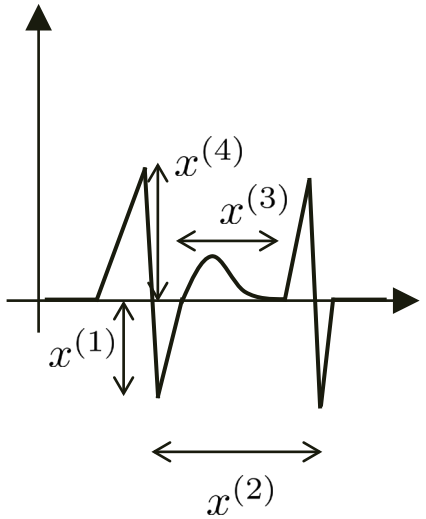
Dane pomiarowe

$$\left\{(\mathbf{x}_i, C_i)\right\}_{i=1}^{N}$$

- Cechy



$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \equiv \mathbf{x}$$



$$\begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(K)} \end{bmatrix} \equiv \mathbf{x}$$

- Przykład algorytmu klasyfikacji (modelu)

$$\overline{C} = \varphi(\theta^{\mathrm{T}}\mathbf{x})$$

$$\phi(w) = \begin{cases} 1 & \text{jeżeli } w > 0 \\ 0 & \text{jeżeli } w \leq 0 \end{cases}$$

- Wektor parametrów klasyfikatora

$$\theta \in \Re^K$$

UJĘCIE PROBABILISTYCZNE

Pełna informacja probabilistyczna

wektor cech

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(K)} \end{bmatrix}$$

etykiety klas

$$\{C_1, C_2, \cdots, C_M\}$$

$p(\mathbf{x}, C)$

$p(\mathbf{x})$ – rozkład cech

$p(C)$ – rozkład klas

$p(\mathbf{x}|C)$ – rozkład cech w klasach

$$\begin{bmatrix} p\left(x^{(1)}|C\right) \\ \vdots \\ p\left(x^{(K)}|C\right) \end{bmatrix}$$

$p(\mathbf{x}|C)$ – model **generatywny**

$p(C|\mathbf{x})$ – model **dyskryminatywny**

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \cdot p(C)}{p(\mathbf{x})}$$

Optymalny algorytm Bayesa

$$\widehat{m} = \arg\max_m p(C_m|\mathbf{x})$$

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \cdot p(C)}{p(\mathbf{x})}$$

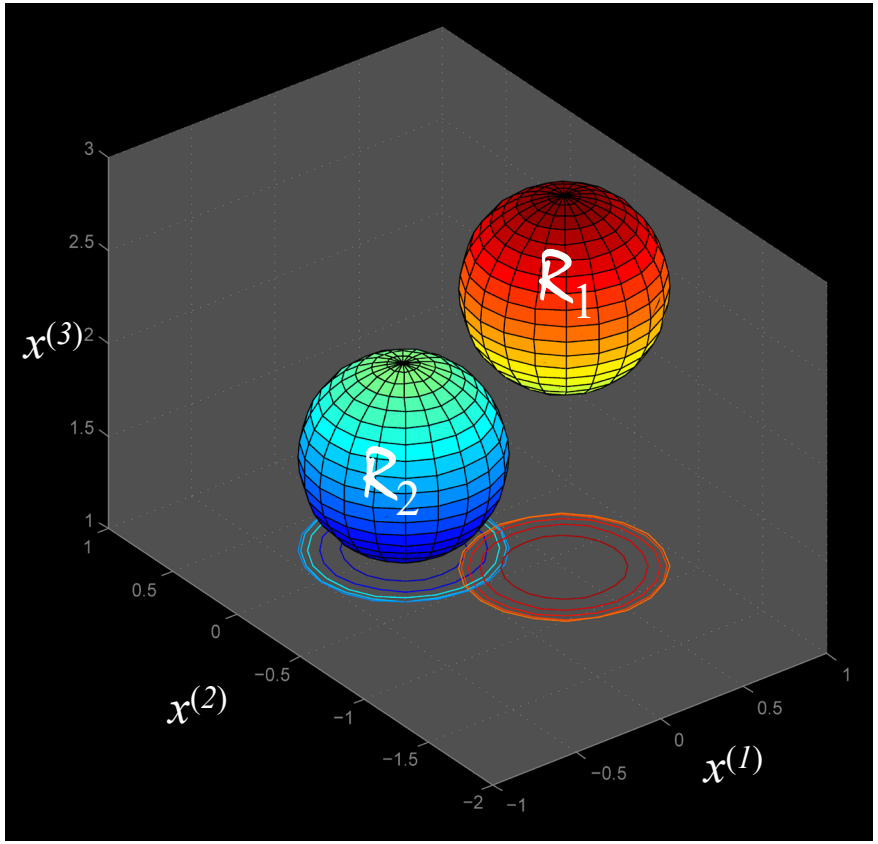
$$p(C_m|\mathbf{x}) \propto p(\mathbf{x}|C_m) \cdot p(C_m)$$

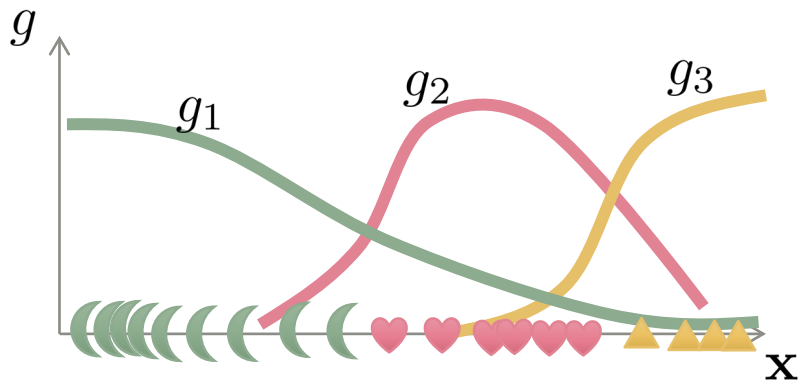
$$\log p(C|\mathbf{x}) = \log p(\mathbf{x}|C) + \log p(C) - \log p(\mathbf{x})$$

funkcja dyskryminująca : $g_m(\mathbf{x}) = \log p(\mathbf{x}|C_m) + \log p(C_m)$

powierzchnia rozdzielająca
/ granica decyzyjna : $s_{rj}(\mathbf{x}) = g_r(\mathbf{x}) - g_j(\mathbf{x}) = \log \frac{p(\mathbf{x}|C_r)}{p(\mathbf{x}|C_j)} + \log \frac{p(C_r)}{p(C_j)} = 0$

Co się dzieje w przestrzeni cech





Przypominajka statystyczna

$$p(AB) = p(A) \cdot p(B|A)$$

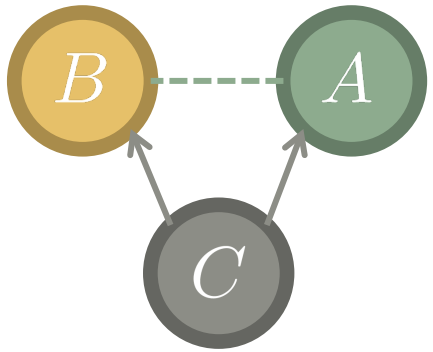
- Bezwarunkowa niezależność

$$B \perp A$$

$$\Downarrow$$

$$p(AB) = p(A) \cdot p(B)$$

$$p(AB|C) = p(A|C) \cdot p(B|AC)$$



- Warunkowa niezależność

$$A \perp\!\!\!\perp B \mid C$$

$$\Downarrow$$

$$p(AB|C) = p(A|C) \cdot p(B|C)$$

Warunkowa niezależność cech

$$
\begin{aligned}
p(\mathbf{x}) &= p\left(x^{(1)}, \ldots, x^{(K)}\right) = \\
&= p\left(x^{(2)}, \ldots, x^{(K)} | x^{(1)}\right) \cdot p\left(x^{(1)}\right) = \\
&= p\left(x^{(3)}, \ldots, x^{(K)} | x^{(1)}, x^{(2)}\right) \cdot p\left(x^{(2)} | x^{(1)}\right) \cdot p\left(x^{(1)}\right) = \\
&= \prod_{k=1}^{K} p\left(x^{(k)} | x^{(2)}, \ldots, x^{(k-1)}\right)
\end{aligned}
$$

$$p\left(x^{(i)} | x^{(j)}\right) = p\left(x^{(i)}\right)$$

$$p(\mathbf{x}) = \prod_{k=1}^{K} p\left(x^{(k)}\right)$$

Naiwny algorytm Bayesa

Optymalny algorytm Bayesa: $\widehat{m} = \arg\max_m p(C_m|\mathbf{x}) = \arg\max_m p(\mathbf{x}|C_m) \cdot p(C_m)$

Naiwnie zakładamy niezależność cech w klasach

$$p(\mathbf{x}|C_m) = p(x^{(1)}|C_m) \cdot p(x^{(2)}|C_m) \cdot \ldots \cdot p(x^{(K)}|C_m)$$

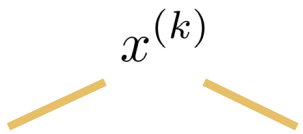
$$\widehat{m} = \arg\max_m \; p(C_m) \prod_{k=1}^{K} p(x^{(k)}|C_m)$$

jaki rozkład cech w klasach?

`GaussianNB`

`BernoulliNB`

`MultinomialNB`



dyskretne ciągłe

	wiek	dochód	studiuje	kupił sztabkę złota
1	młody	wysoki	nie	nie
2	młody	wysoki	nie	nie
3	produkcyjny	wysoki	nie	tak
4	senior	przeciętny	nie	tak
5	senior	niski	tak	tak
6	senior	niski	tak	nie
7	produkcyjny	niski	tak	tak
8	młody	przeciętny	nie	nie
9	młody	niski	tak	tak
10	senior	przeciętny	tak	tak
11	młody	przeciętny	tak	tak
12	produkcyjny	przeciętny	nie	tak
13	produkcyjny	wysoki	tak	tak
14	senior	przeciętny	nie	nie

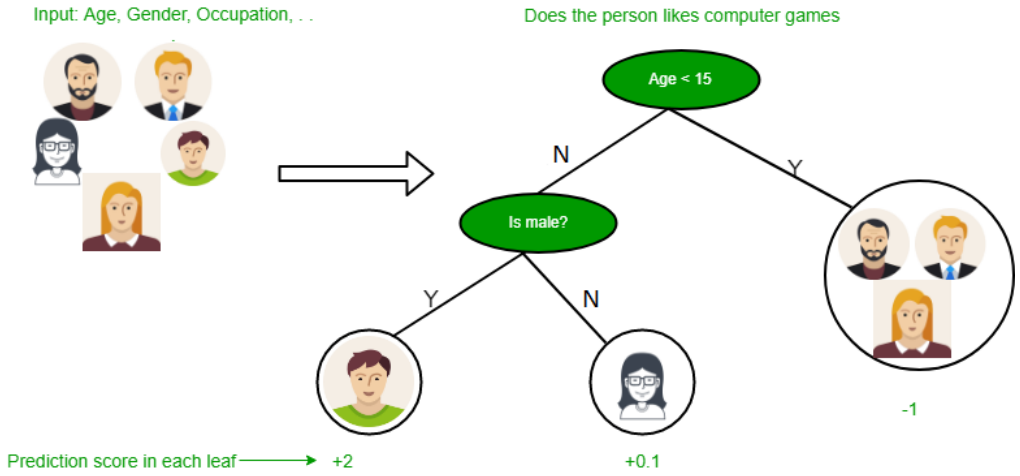
$$P(\text{wiek} = \textit{młody} \mid \text{kupił} = \textit{tak})$$

Algorytm *k*-najbliższych sąsiadów

- stały dostęp do bazy danych
- drzewo poszukiwań

Drzewa decyzyjne

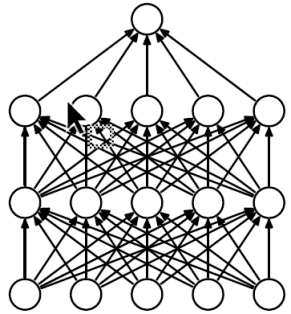
- Drzewa decyzyjne – dzielą dane na bloki
- Lasy losowe – dekorelacja drzew przez losowy wybór wierszy i kolumn (pakowanie w worki, ang. *bagging*) + uśrednianie predykcji
- Drzewa wzmacniane gradientowo – sekwencyjne uczenie na błędach poprzednika

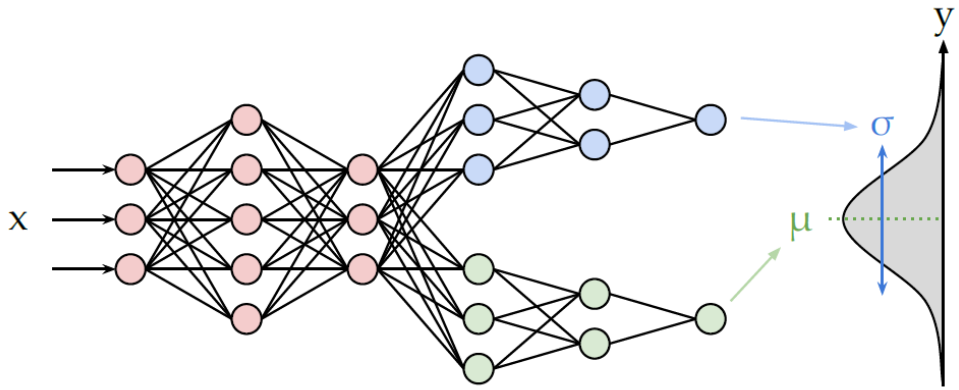


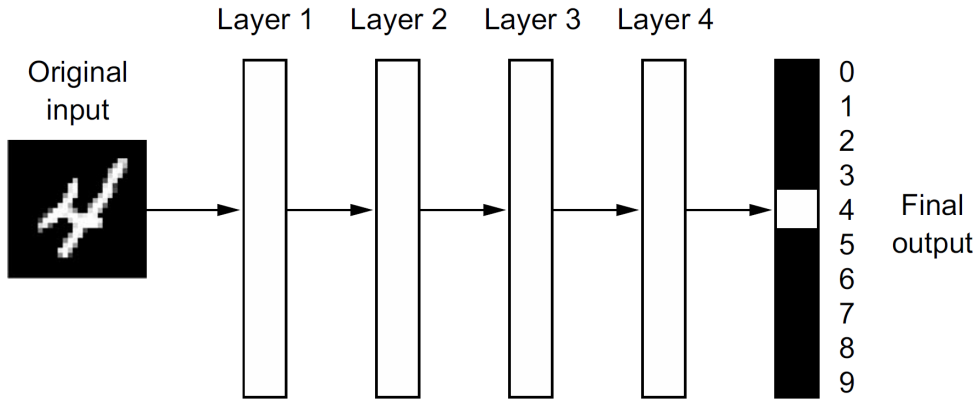
Sieci neuronowe

Model warstwowy: sekwencja (mnożenie macierzy → `nieliniowa transformacja`), DAG

„Neuron”: model liniowy + nieliniowa funkcja (tzw. *funkcja aktywacji*)







TensorFlow Playground

Epoch 000,000 | Learning rate 0.03 | Activation Tanh | Regularization None | Regularization rate 0 | Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1

X_2

X_1^2

X_2^2

X_1X_2

$\sin(X_1)$

$\sin(X_2)$

2 HIDDEN LAYERS

4 neurons | 2 neurons

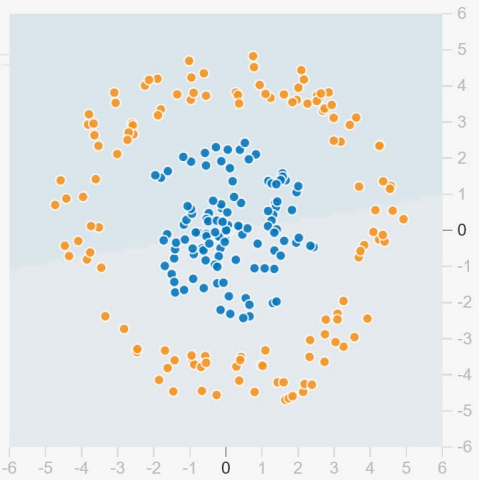
This is the output from one neuron. Hover to see it larger.

The outputs are mixed with varying weights, shown by the thickness of the lines.

OUTPUT

Test loss 0.503

Training loss 0.507

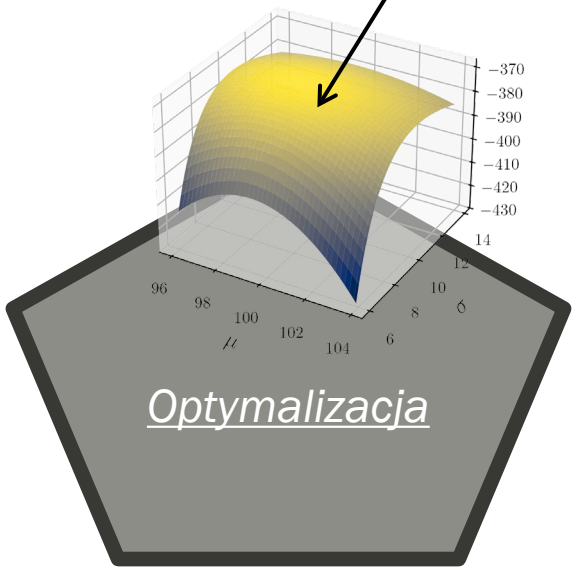


Colors shows data, neuron and weight values

-1 0 1

Uczenie maszynowe

jako



Optymalizacja

jako

Estymacja

Ewaluacja modelu

Dopasowanie do danych
z uwzględnieniem wiedzy

Kryterium jakości dla klasyfikatora wieloklasowego

Kodowanie gorąco-jedynkowe (ang. *one-hot encoding*): $\mathbf{y}_i = \begin{bmatrix} y_i^{(1)} \\ y_i^{(2)} \\ \vdots \\ y_i^{(M)} \end{bmatrix}$

Dla obiektu z klasy C_m tylko $y_i^{(m)} = 1$ a pozostałe są zerowe.

$C_m(\mathbf{x}_i)$ – etykieta klasy numer m, do której należy $\mathbf{x}_i$

Metoda maksymalnej wiarygodności

$$L(\theta) = \prod_{i=1}^{N} \prod_{m=1}^{M} p\big(\mathbf{x}_i | C_m(\mathbf{x}_i)\big)^{y_i^{(m)}} \Rightarrow \quad \log L(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_i^{(m)} \log p\big(\mathbf{x}_i | C_m(\mathbf{x}_i)\big)$$

kategorialna entropia krzyżowa
(ang. *categorical cross-entropy*)

Kryterium jakości dla klasyfikatora binarnego

Dla obiektu z klasy C przyjmuje się $y = 1$, w przeciwnym przypadku $y = 0$.

Metoda maksymalnej wiarygodności

$$L(\theta) = \prod_{i=1}^{N} p\big(\mathbf{x}_i|C(\mathbf{x}_i)\big)^{y_i} \left[1 - p\big(\mathbf{x}_i|C(\mathbf{x}_i)\big)\right]^{1-y_i}$$

$$\Downarrow$$

$$L(\theta) = \sum_{i=1}^{N} \Big\{ y_i \log p\big(\mathbf{x}_i|C(\mathbf{x}_i)\big) + (1 - y_i) \log \left[1 - p\big(\mathbf{x}_i|C(\mathbf{x}_i)\big)\right] \Big\}$$

binarna entropia krzyżowa (ang. *binary cross-entropy*)

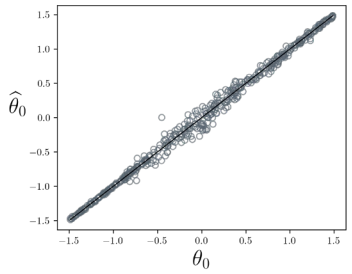
Uczenie maszynowe

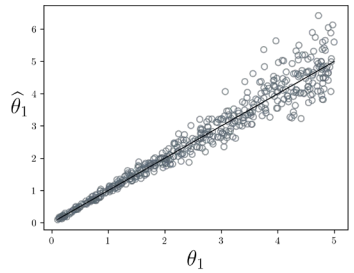
jako

jako

Optymalizacja

Estymacja





Dopasowanie do danych z uwzględnieniem wiedzy

Ewaluacja modelu

Błędy decyzyjne

Fałszywie dodatni

Błąd I rodzaju

- **Hipoteza zerowa** jest **prawdziwa**, ale ją odrzucam
- Wykrywam efekt, gdy go nie ma
- **Fałszywy alarm** (ang. *False Positive*)
- $P(błąd\ I\ rodz.)$ to ***poziom istotności***

- ➢ *Mylnie stwierdzona choroba*
- ➢ *Dobry lek nie dopuszczony do obrotu*
- ➢ *Ukarano niewinną osobę*

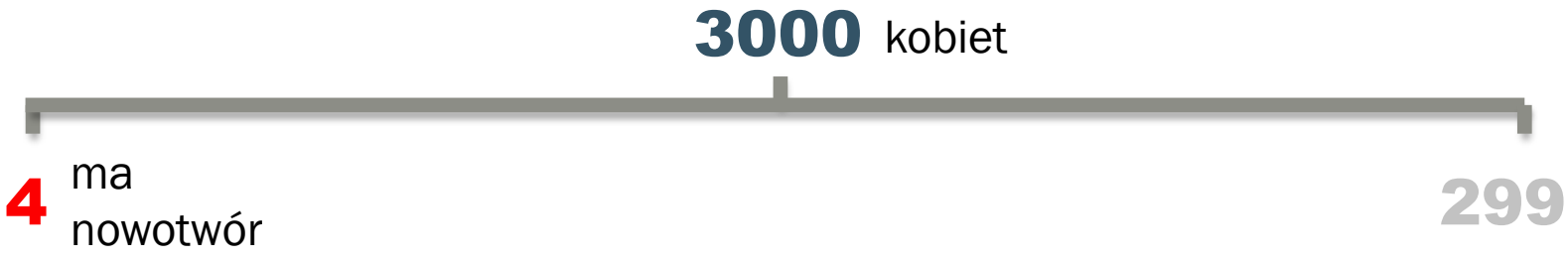
Fałszywie ujemny

Błąd II rodzaju

- **Hipoteza zerowa** jest **fałszywa**, ale ją akceptuję
- Nie zauważam efektu
- **Przeoczenie** (ang. *False Negative*)
- $1 - P(błąd\ II\ rodz.)$ to ***moc testu***

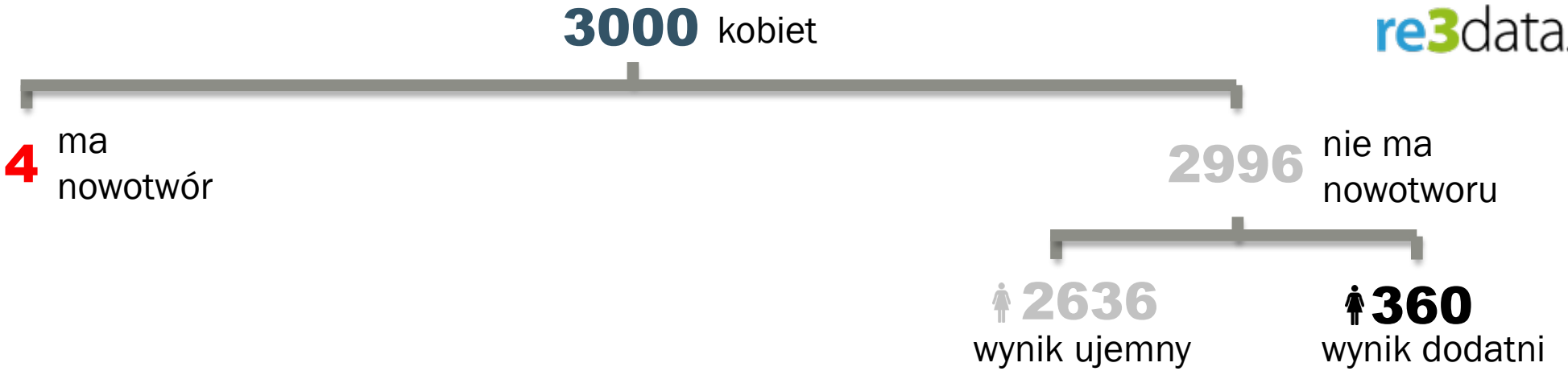
- ➢ *Nie rozpoznana choroba*
- ➢ *Zły lek dopuszczony do obrotu*
- ➢ *Uniewinniono sprawcę*

88%



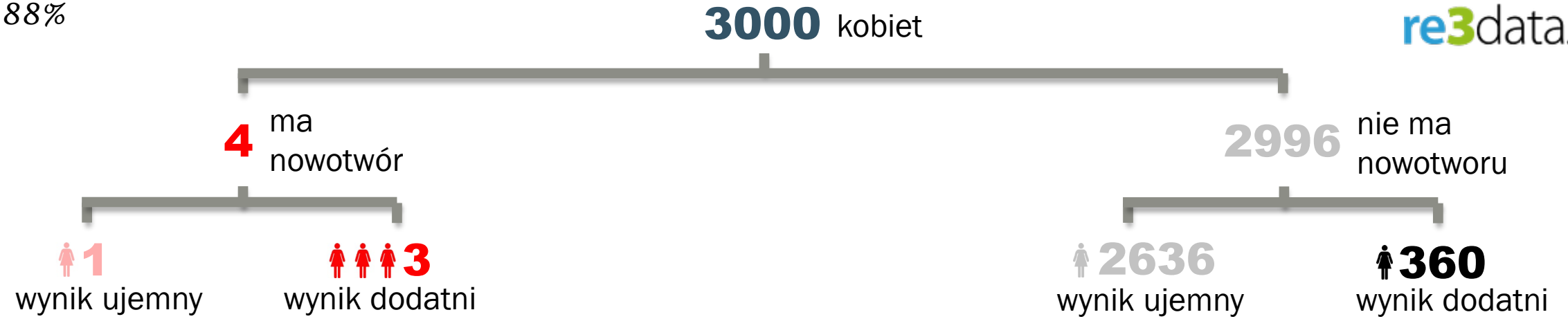
re3data.org

88%



re3data.org

88%



re3data.org

88%

re3data.org

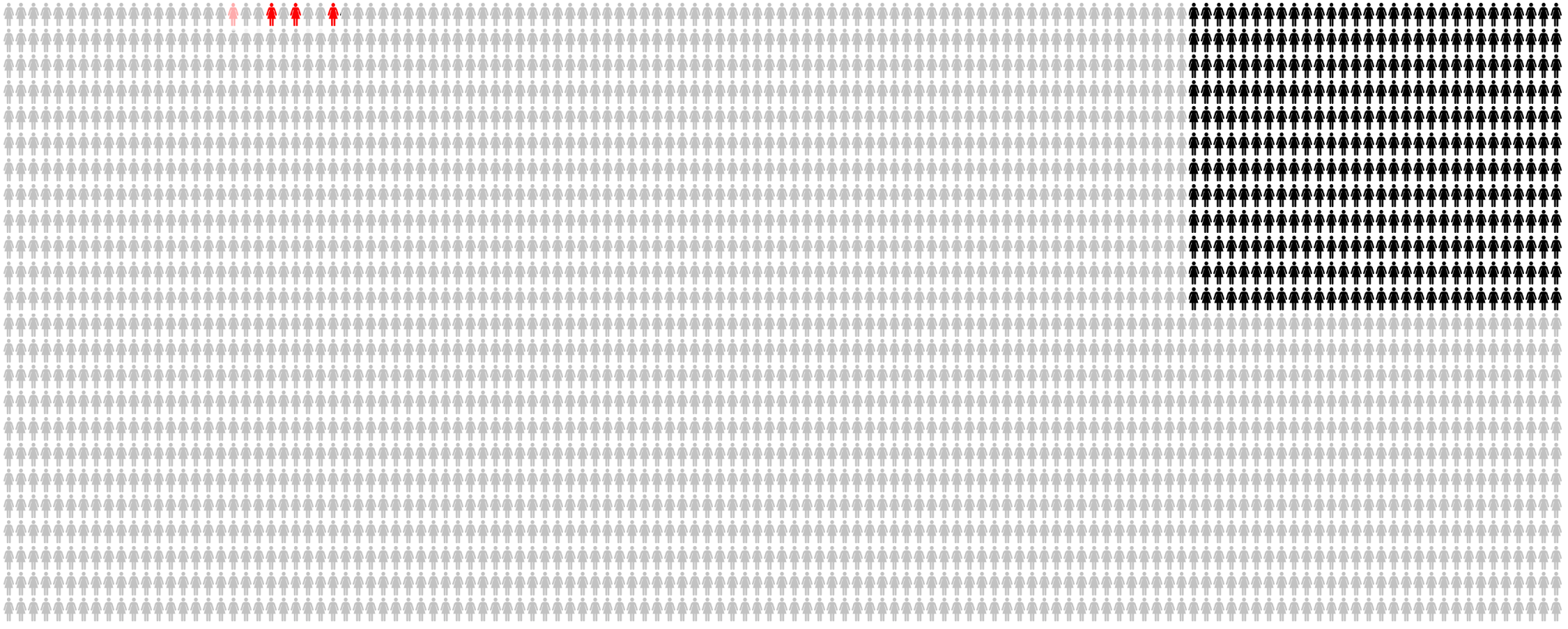
3000 kobiet

4 ma nowotwór

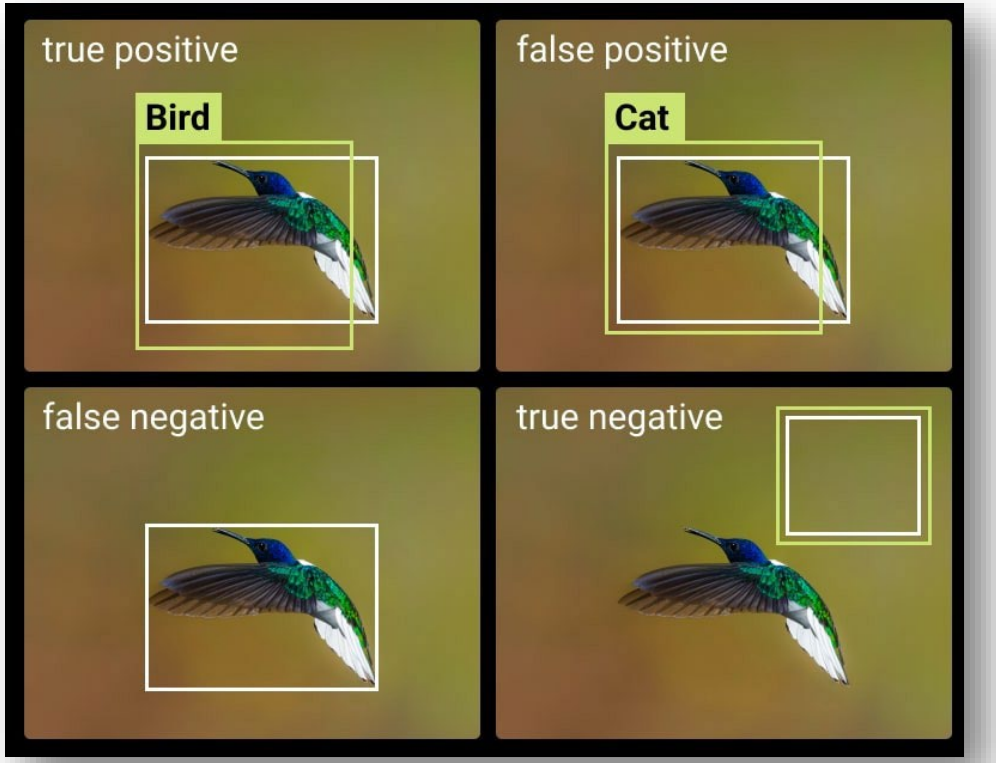
- **1** wynik ujemny
- **3** wynik dodatni

2996 nie ma nowotworu

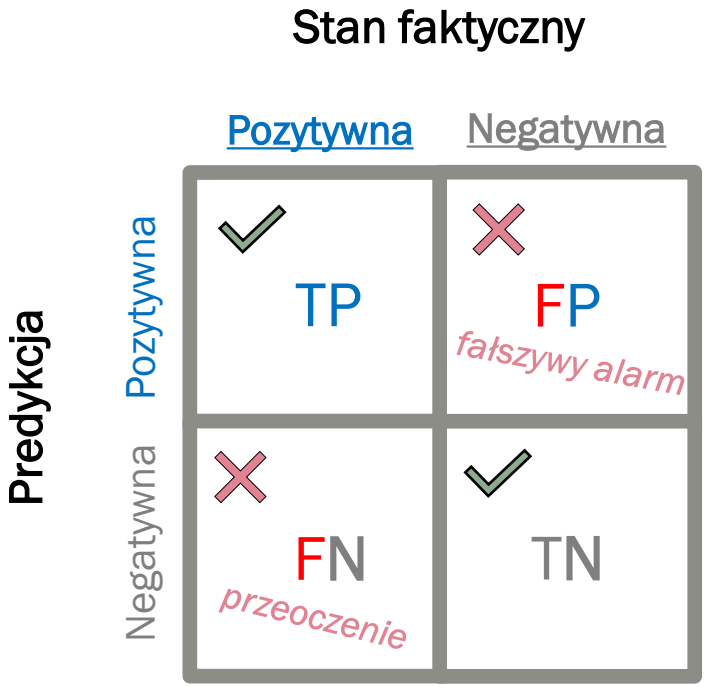
- **2636** wynik ujemny
- **360** wynik dodatni



Ocena jakości klasyfikacji



https://labelyourdata.com/articles/object-detection-metrics



Macierz pomyłek
(ang. *confusion matrix*)

Ocena jakości klasyfikacji

$$\text{Precyzja} = \frac{TP}{FP + TP}$$

Jak często przewidywania P są trafne?
Czy poprawnie wykrywa?

$$\text{Pełność} = \frac{TP}{FN + TP}$$

(ang. *recall*)

Jak często P są trafnie przewidziane?

$$F_1 = \frac{2}{\frac{1}{\text{Precyzja}} + \frac{1}{\text{Pełność}}}$$

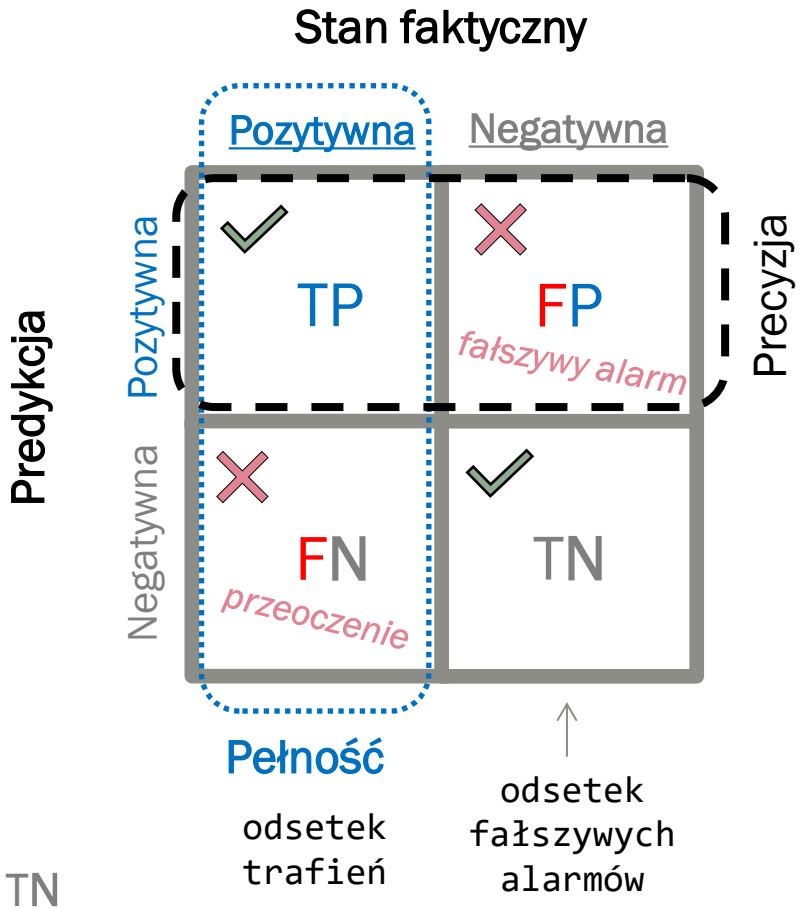
?wzór?

Czy poprawnie ignoruje?

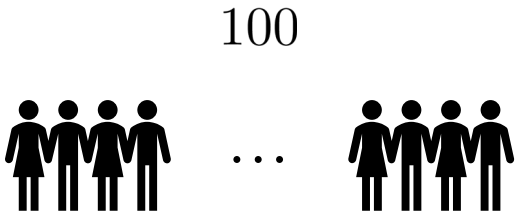
$$\text{Dokładność} = \frac{TP + TN}{TP + TN + FP + FN}$$

(ang. *accuracy*)

Odsetek prawidłowych decyzji



Przykład



- Czy klient zrezygnuje z subskrypcji?

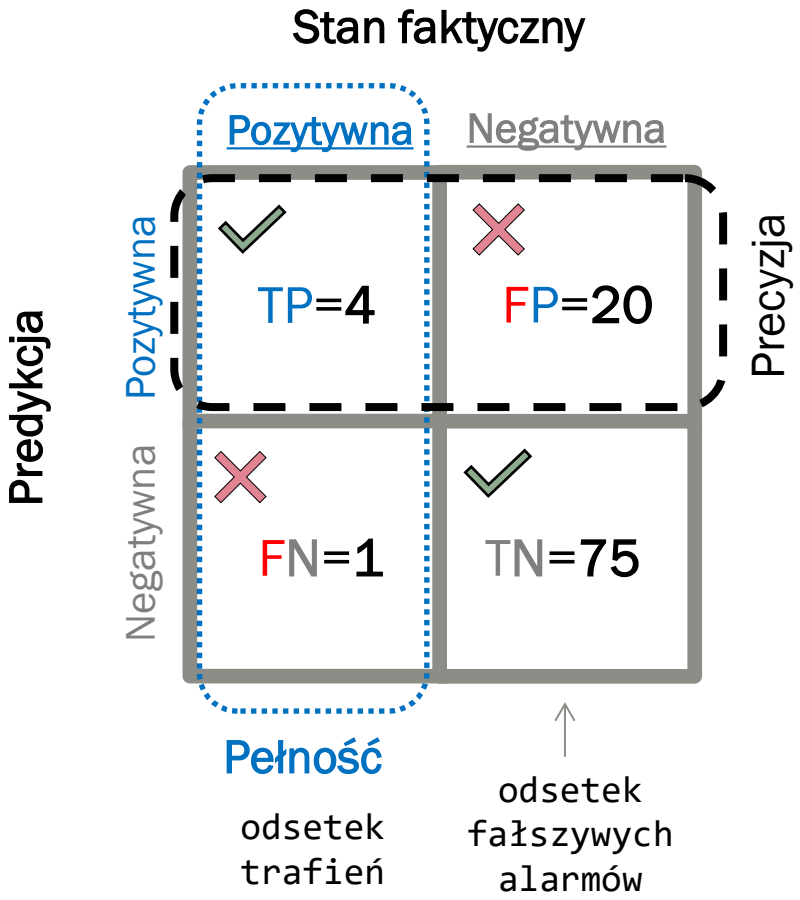
5 zrezygnowało, model wykrył 4 z nich

20 innych model wskazał, ale jednak nie zrezygnowali

Precyzja = ? $\frac{4}{4+20}$

Pełność = ? $\frac{4}{4+1}$

Dokładność = ? $\frac{4+75}{4+75+20+1}$



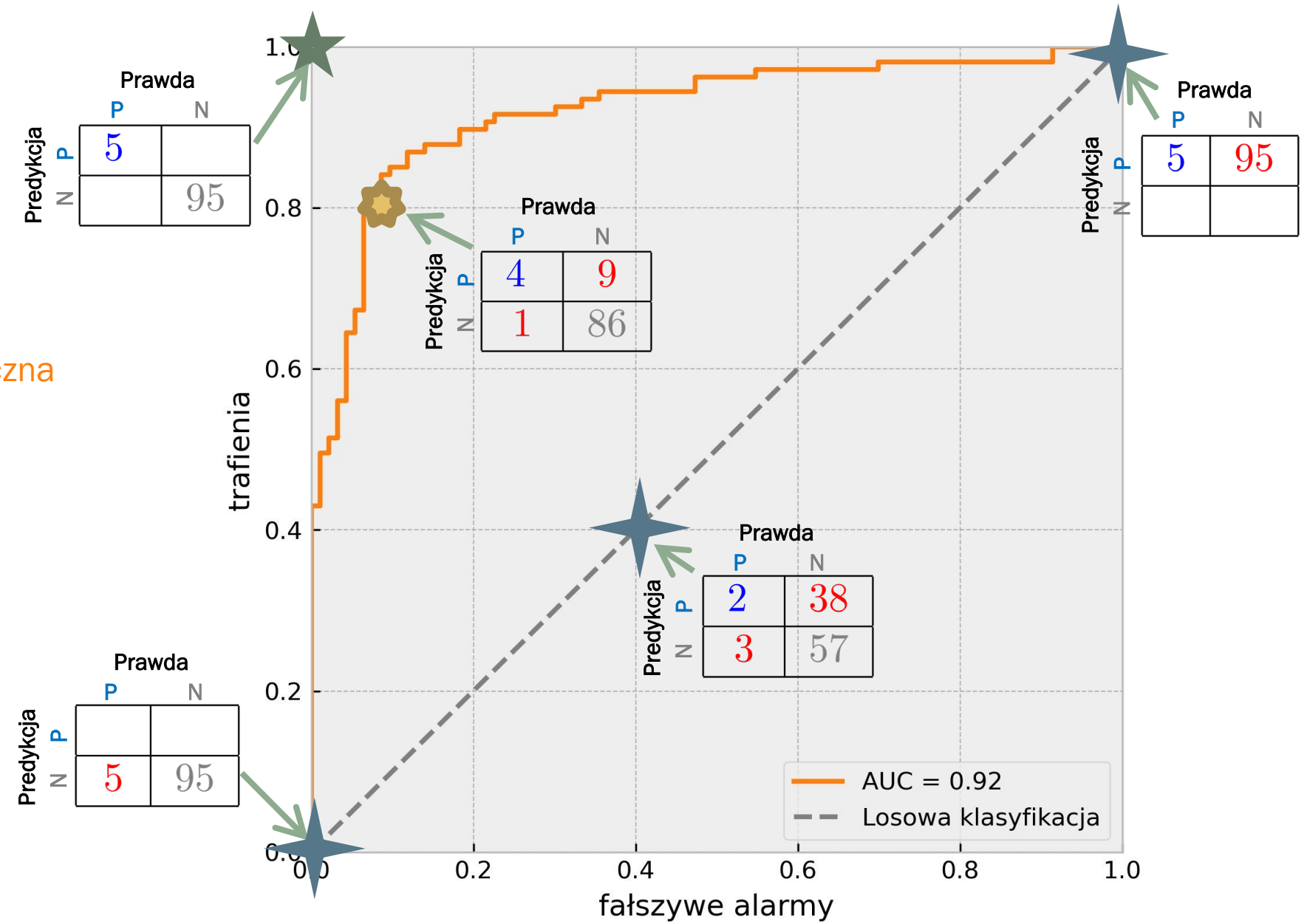
ROC

Receiver
Operating
Characteristic

krzywa
operacyjno-charakterystyczna

AUC

Area
Under the
Curve




The End