Sztuczna inteligencja inna niż ją widzimy: granice zaufania do modeli SI

2026-03-11
Dni Otwarte WMII UJ

Mateusz Pyla
Tomasz Kuśmierczyk

POLONEZ BIS

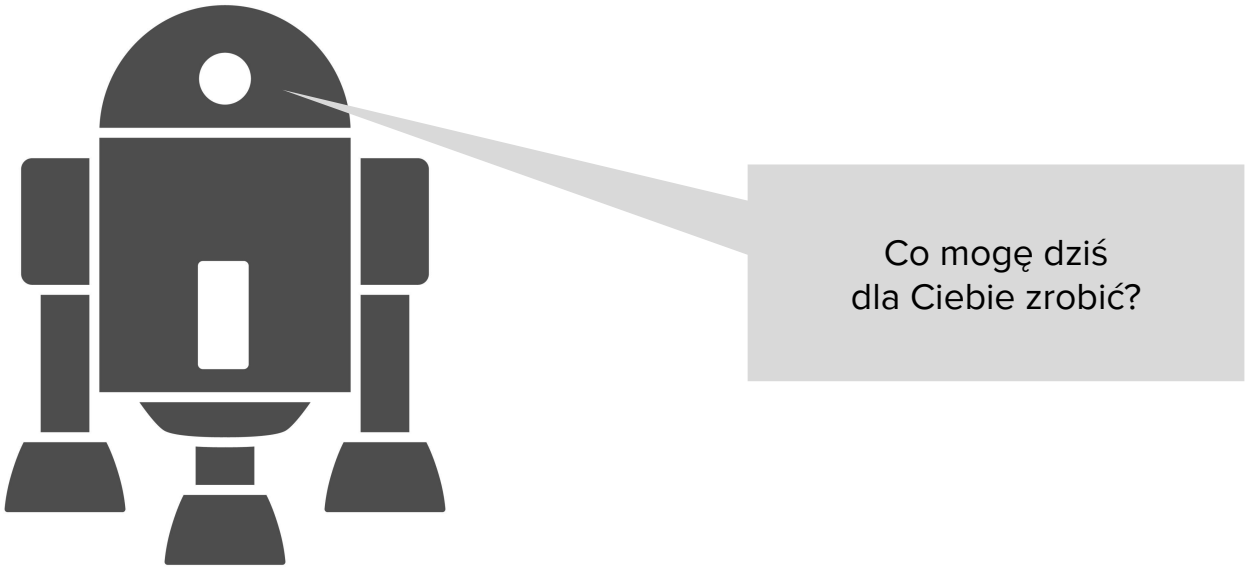
NARODOWE CENTRUM NAUKI

UNIWERSYTET JAGIELLOŃSKI W KRAKOWIE

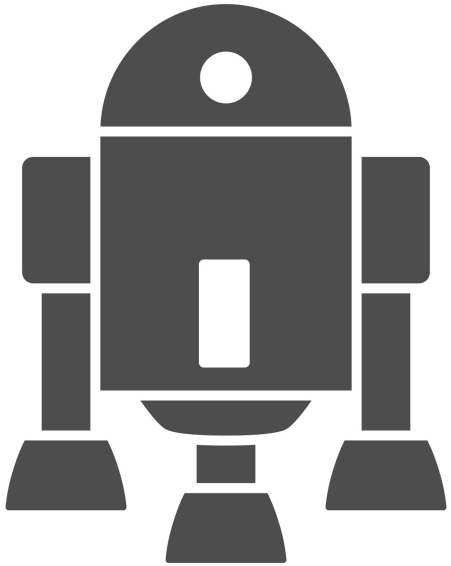
„All models are wrong, but some are useful."
George Box

group of machine gmum learning research

Uczymy modele



Uczymy modele



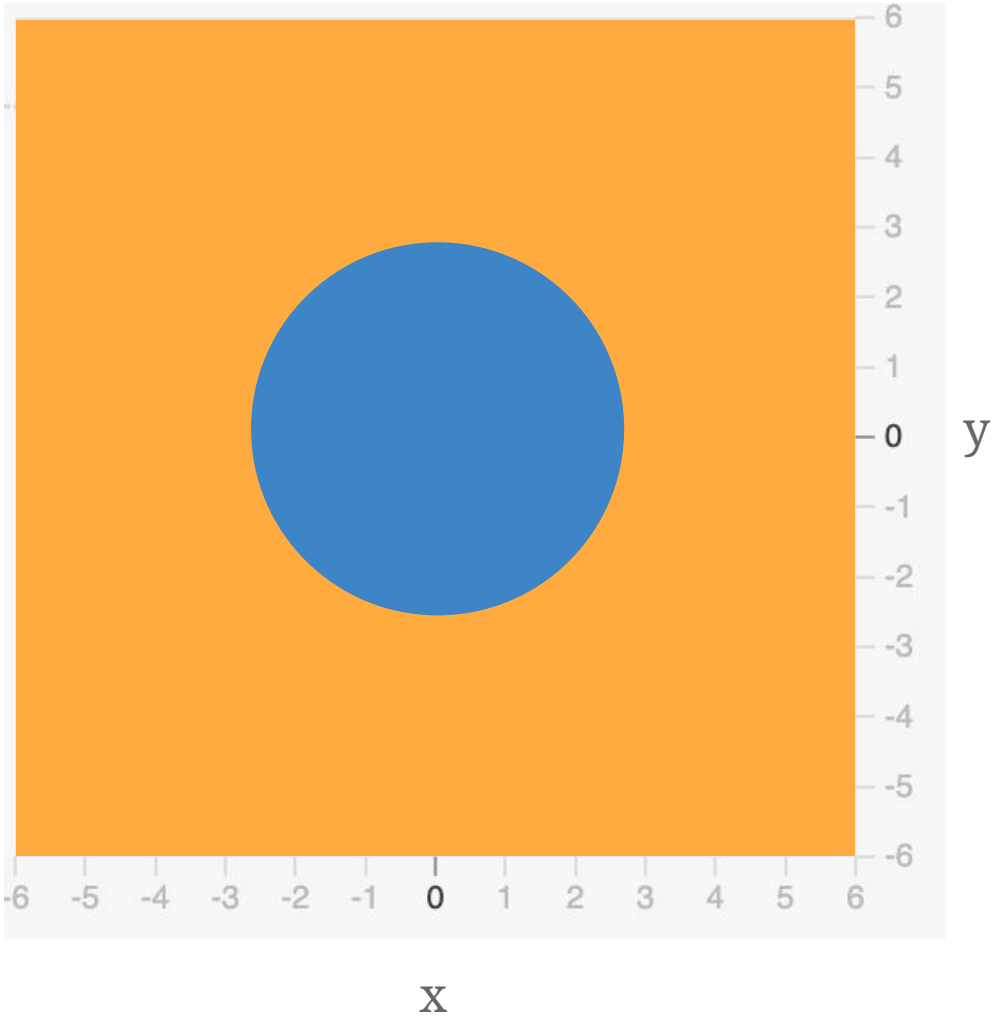
Chcę, żebyś nauczył się
rozpoznawać zwierzęta!
A ja chcę zrozumieć jak to robisz.





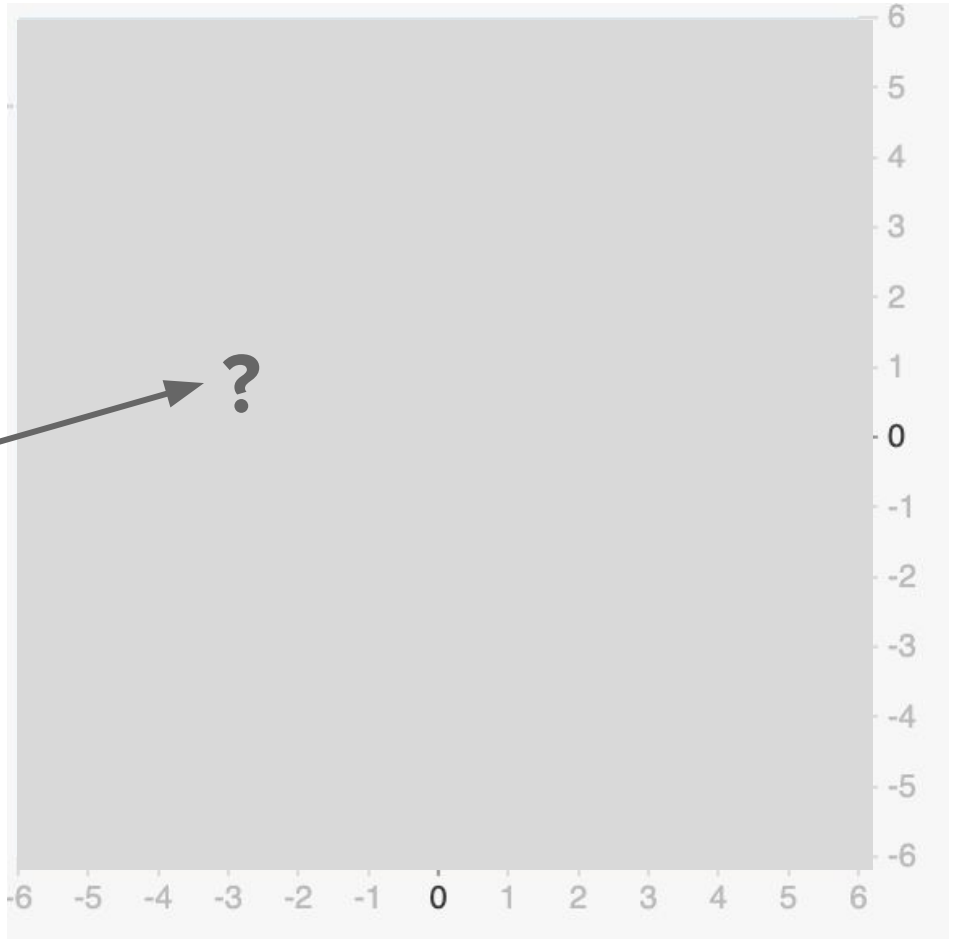


Zacznijmy od czegoś prostszego:
Na podstawie położenia przewiduj kolor

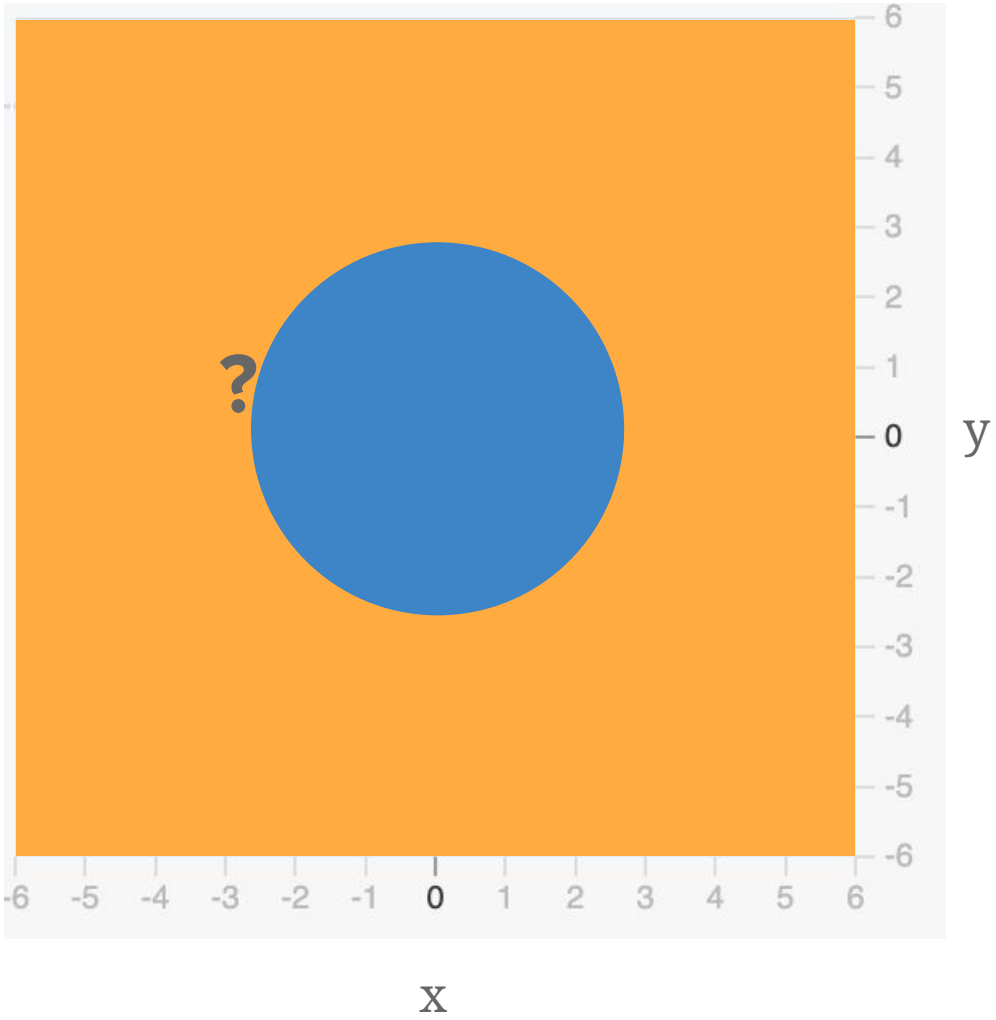


Uczenie sieci prostego zadania:
Na podstawie położenia przewiduj kolor

Jaki kolor tutaj?

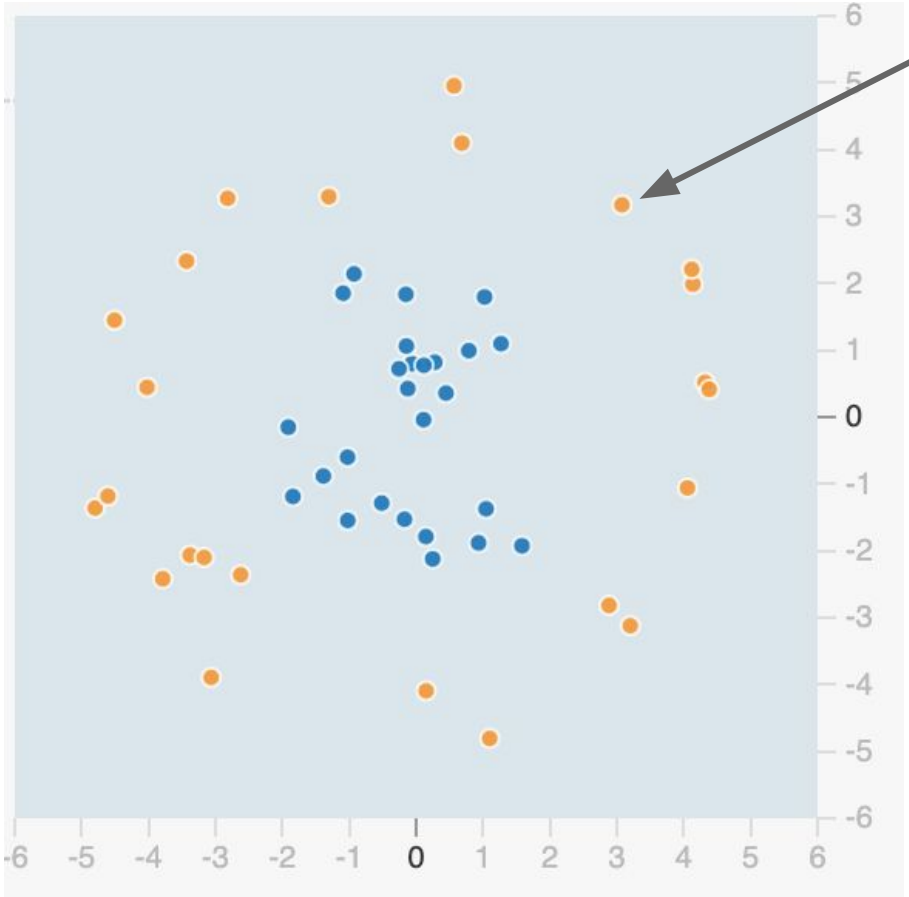


Uczenie sieci prostego zadania:
To przecież banalne!

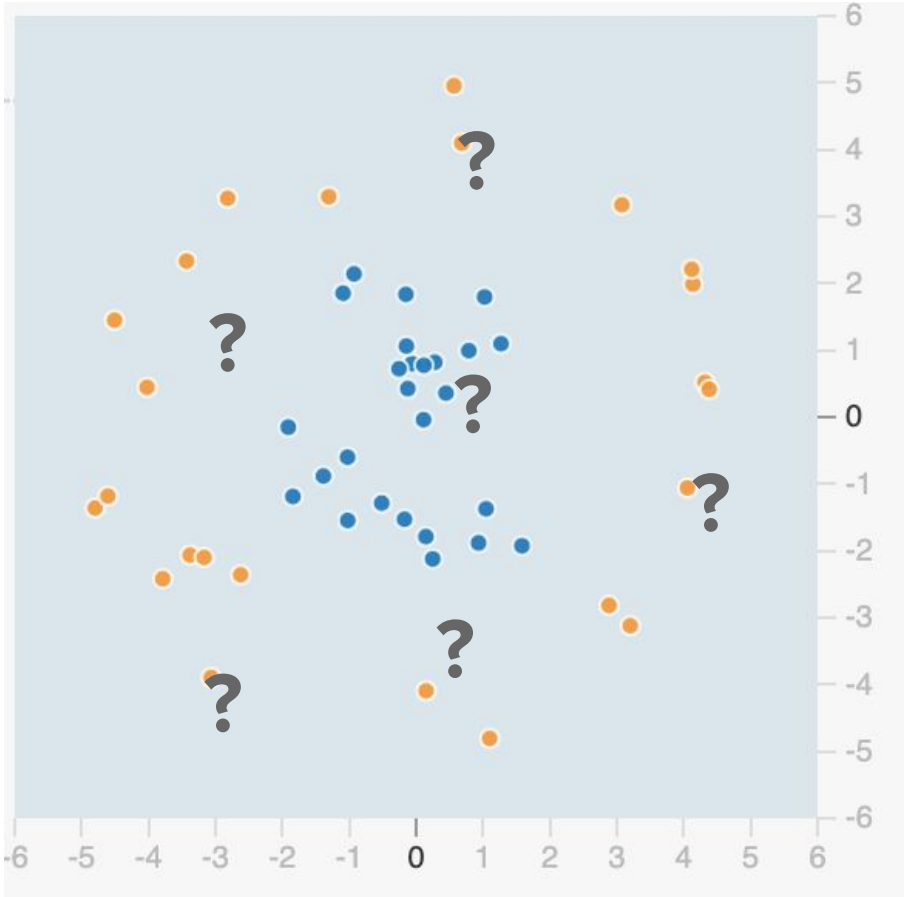


Dane: W rzeczywistości nie znamy całego problemu

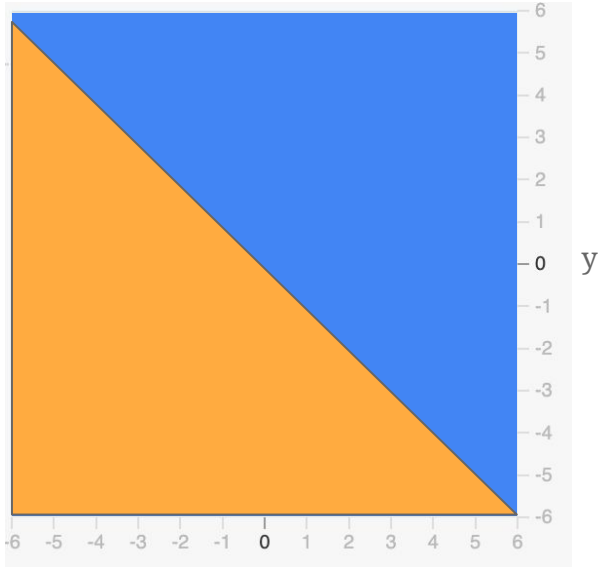
(= nie widzieliśmy wszystkich możliwości)

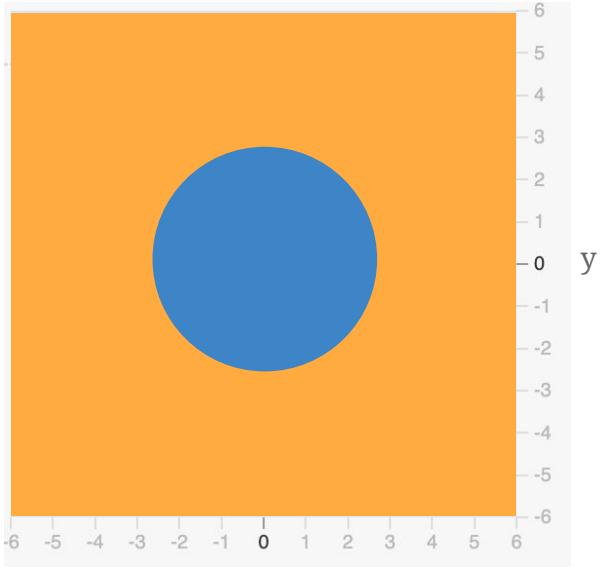


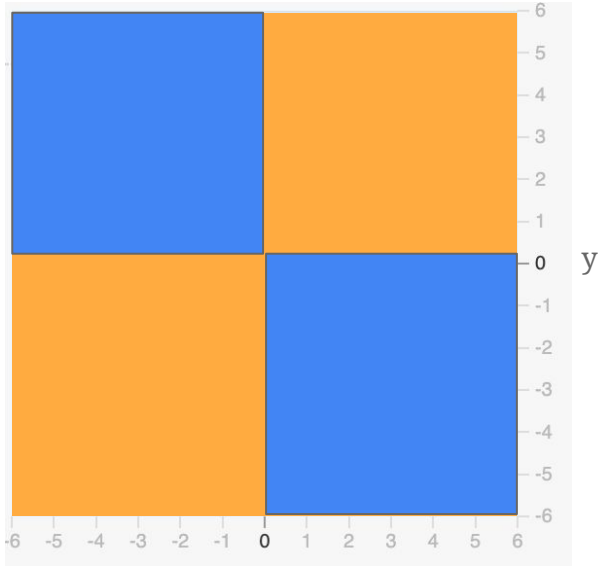
Ale model ma działać ogólnie:
nie wiemy jakie będą zapytania

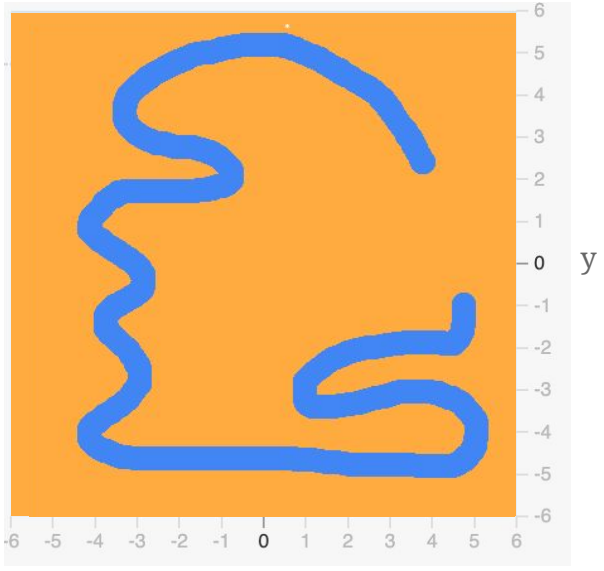


Inne kolorowania

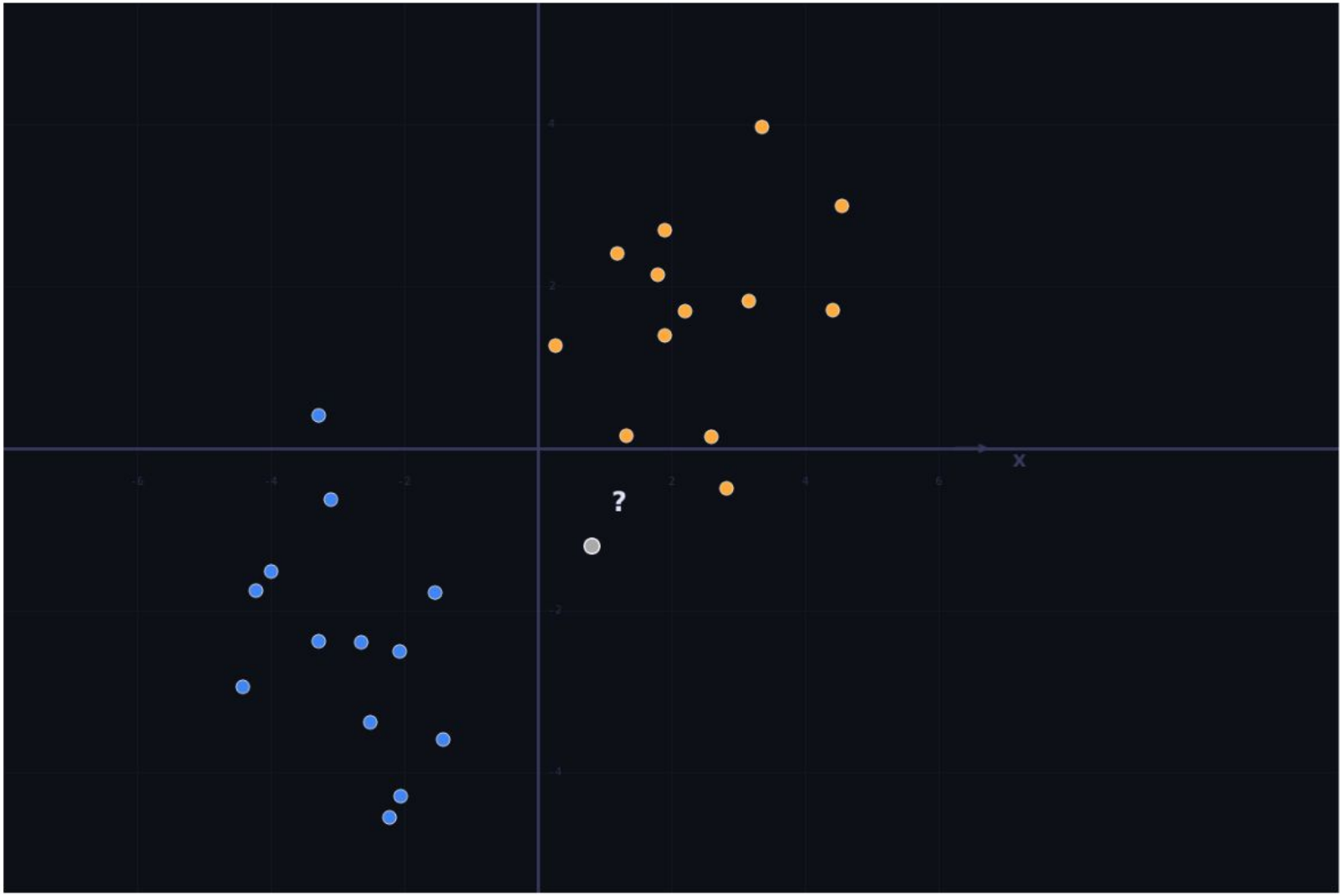




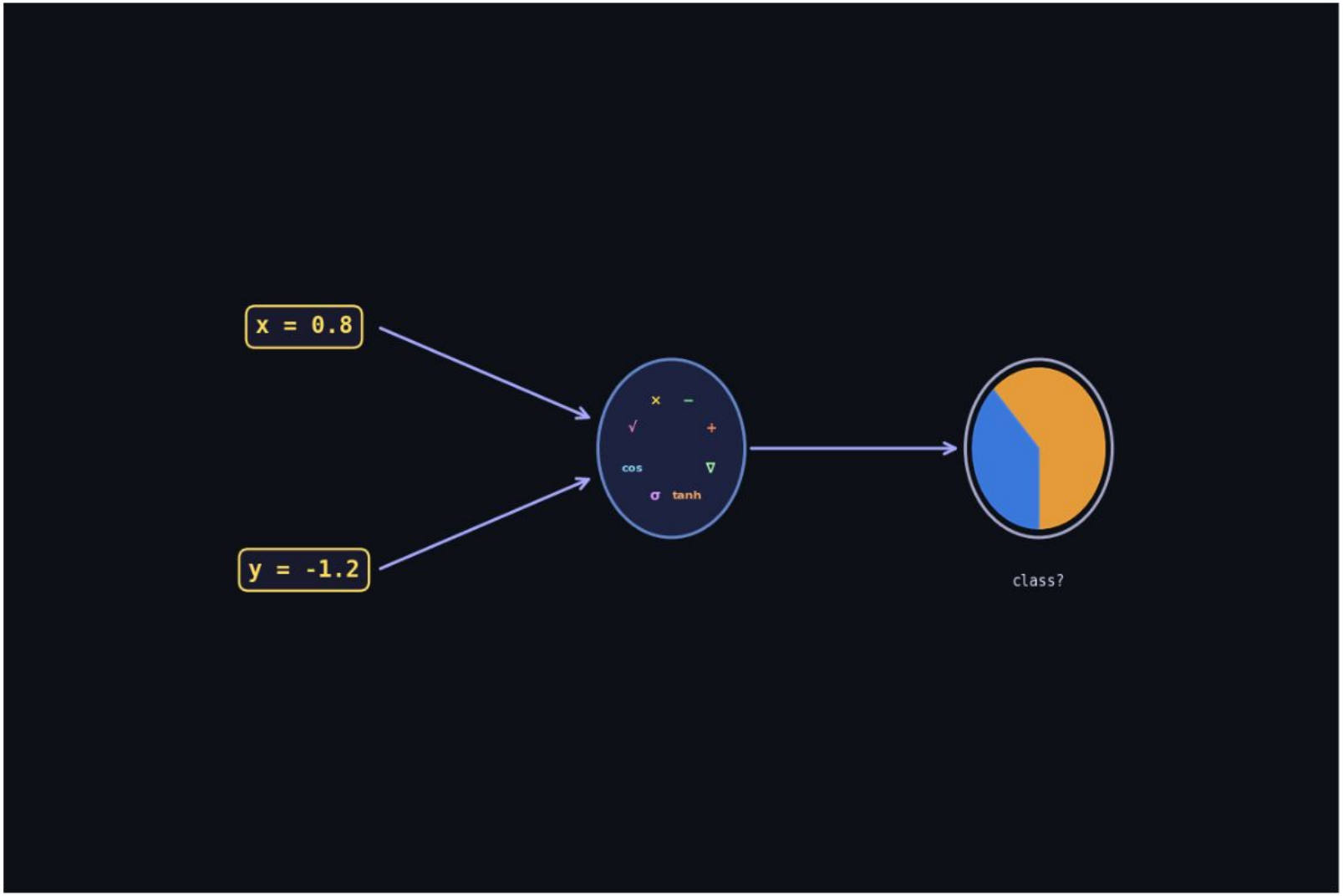




Sformalizowanie problemu:
Toż przecież funkcja!

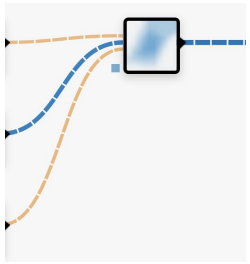


Sformalizowanie problemu:
Toż przecież funkcja!



Jak to działa?
Sztuczne sieci neuronowe

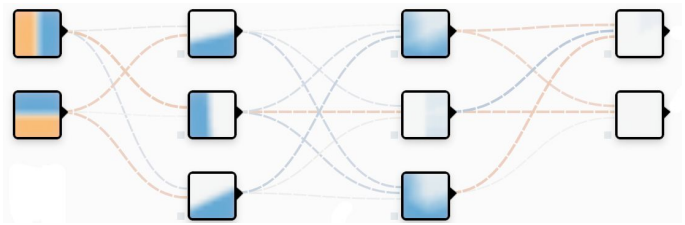
neuron



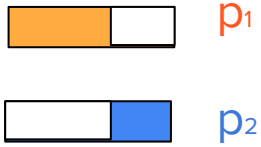
wejścia:

x

y

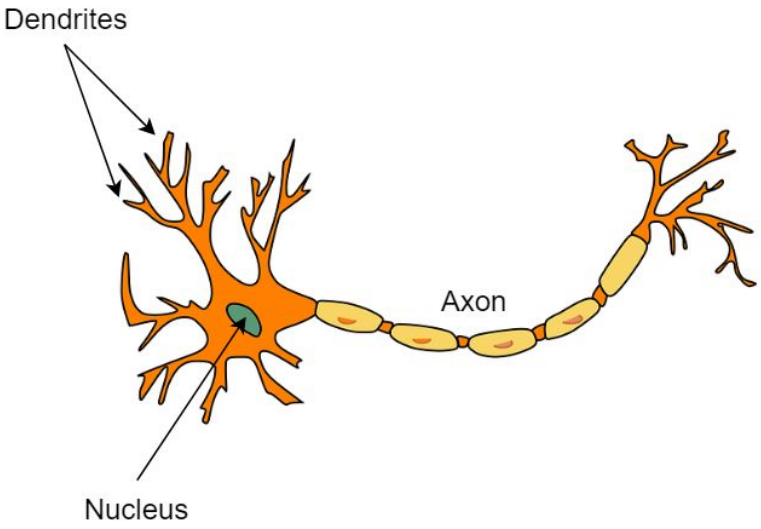


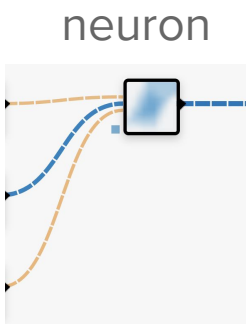
prawdopodobieństwa wyjściowe:



sieć neuronowa
to
skomplikowana funkcja

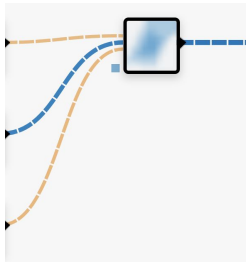
Neurony
Te w głowie i te w komputerach





Jak to działa?
Sztuczne sieci neuronowe

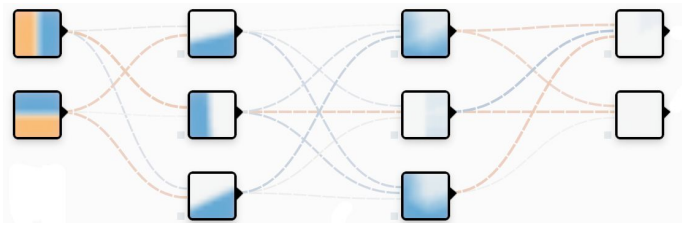
neuron



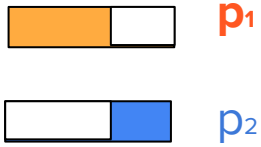
wejścia:

x

y



prawdopodobieństwa wyjściowe:



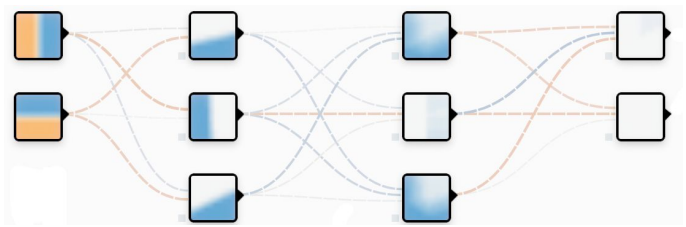
sieć neuronowa
to
skomplikowana funkcja

Jak to działa?
Wsteczna propagacja błędu

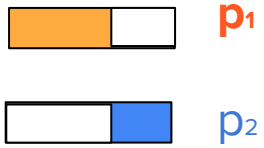
wejścia:

x

y



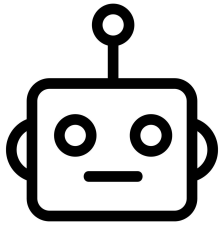
prawdopodobieństwa wyjściowe:



poprawna etykieta:

błąd!

korekta parametrów

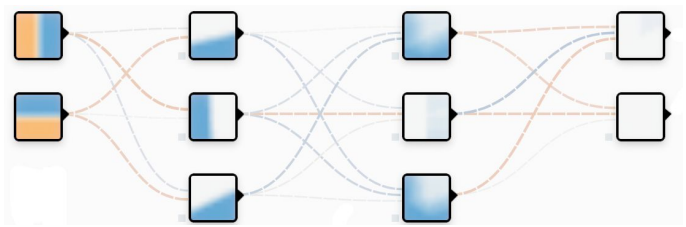


Jak to działa?
Wsteczna propagacja błędu

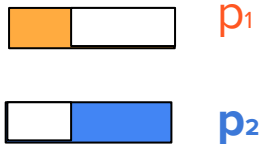
wejścia:

x

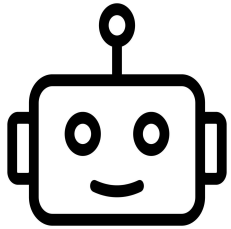
y



prawdopodobieństwa wyjściowe:



poprawna etykieta:



Sztuczne sieci neuronowe
Rozmiar, liczba warstw, liczba neuronów

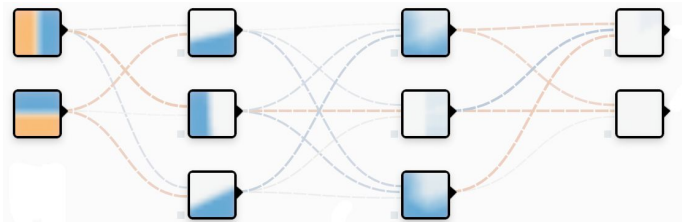
- DEMO: Trenowanie małej sieci neuronowej
- DEMO: Trenowanie dużej sieci neuronowej

Dla zainteresowanych: https://pair-code.github.io/what-if-tool/demos/image.html

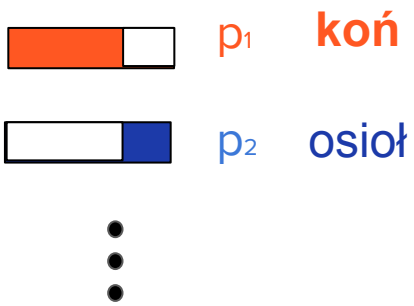
Model *dyskryminatywny*

wejście

model
(sieć neuronowa)

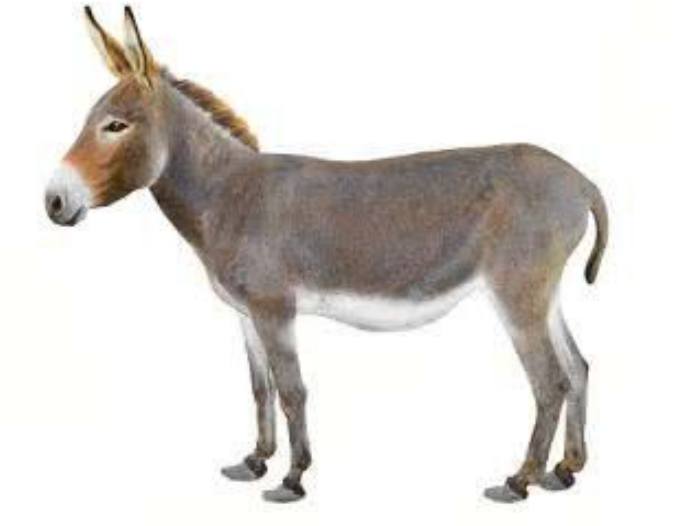


wyjście
(predykcja
prawdopodobieństwa)



dla uproszczenia póki co tylko 2 zwierzątka

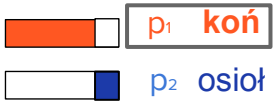
Gra w konia i osła:
Zasady



Gra w konia i osła
Punktacja

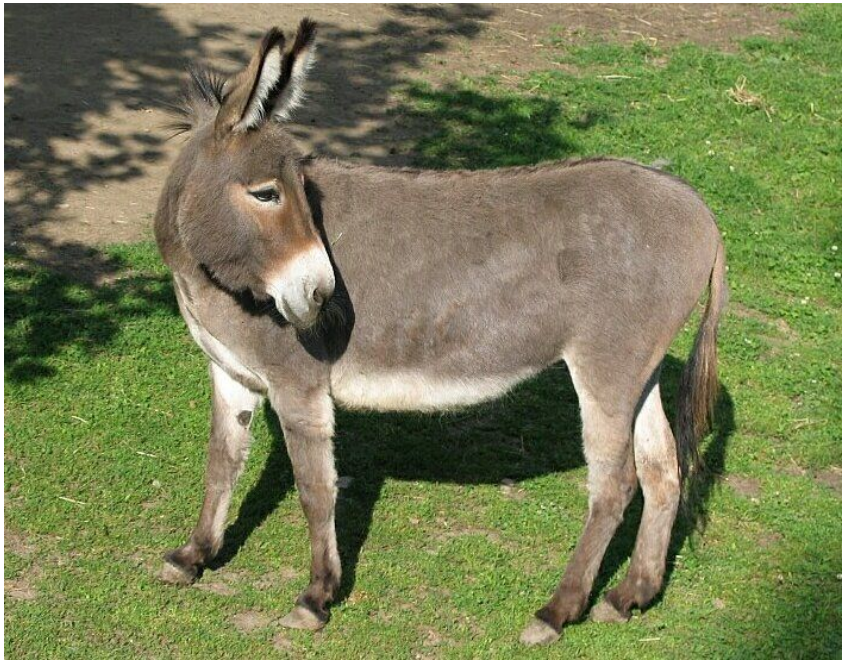
1. Zobaczymy obrazek
2. Ustalamy prawdopodobieństwo: koń vs. osioł
3. Sprawdzamy błędy
4. Oceniamy predykcje:

Jakość = prawdopodobieństwo prawdziwej etykiety



Gra: w konia i osła
Jak sobie publiczność poradzi?

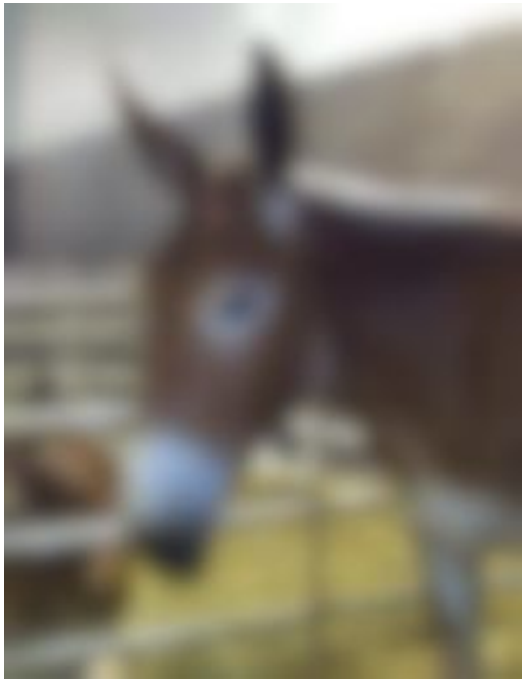
Gra: w konia i osła
Jak sobie publiczność poradzi?



Gra: w konia i osła
Jak sobie publiczność poradzi?

Gra: w konia i osła
Jak sobie publiczność poradzi?

Gra: w konia i osła
Jak sobie publiczność poradzi?



Gra: w konia i osła
Jak sobie publiczność poradzi?

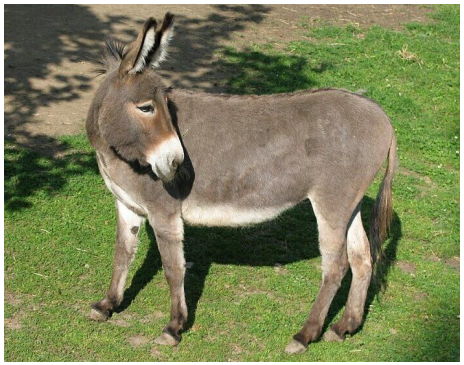
Gra: w konia i osła

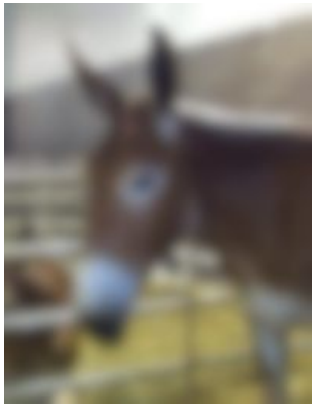
Wyniki

koń

koń

osioł





muł

koń Przewalskiego

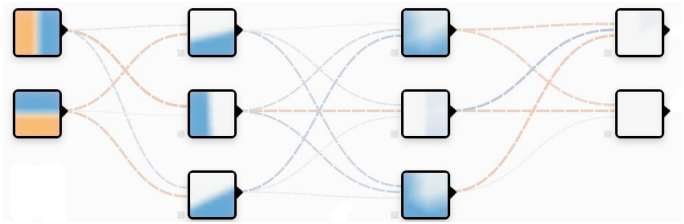
muł

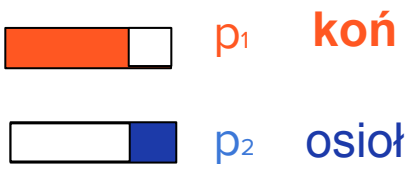
Model *dyskryminatywny*
Na przykładzie jednego zdjęcia

wejście

model
(sieć neuronowa)

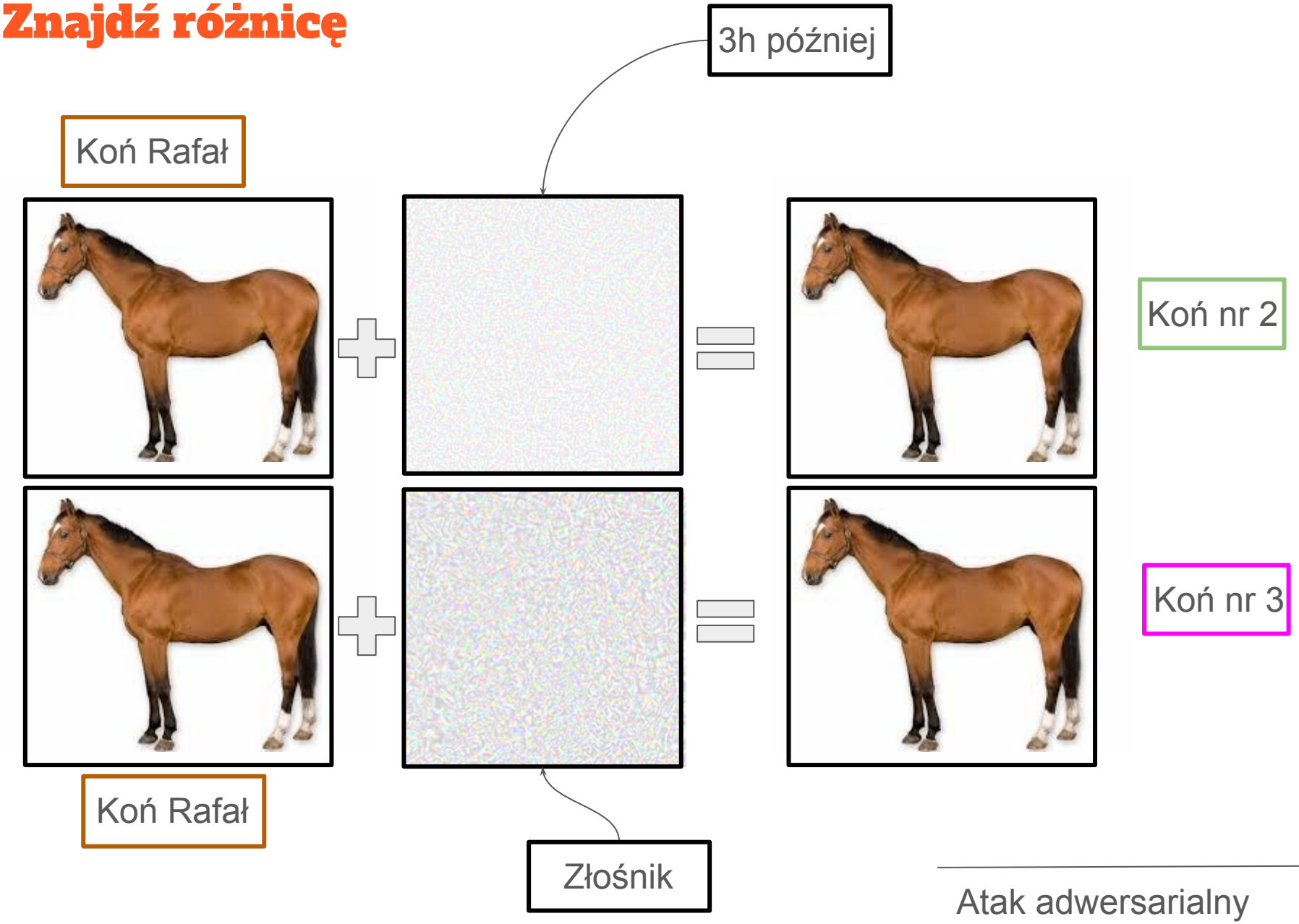
wyjście
(predykcja
prawdopodobieństwa)





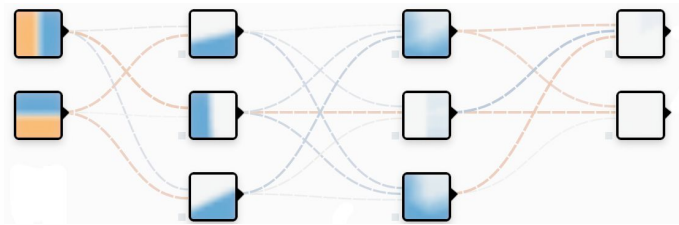
Koń Rafał

Znajdź różnicę

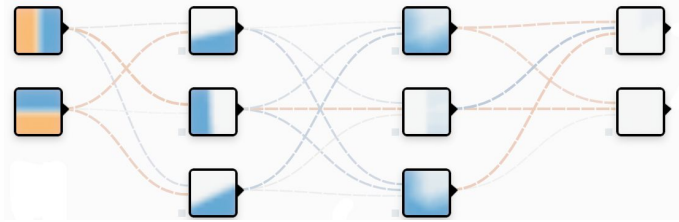


Choroba sieci neuronowych
Pewny, ale błędny

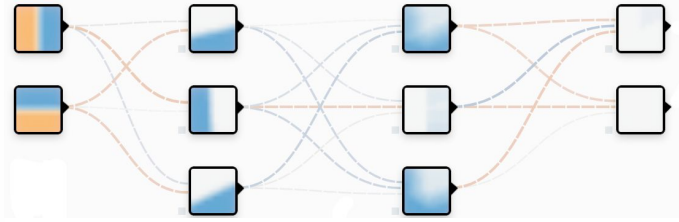
Koń Rafał

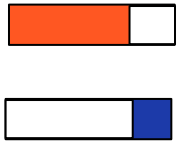


Koń nr 2



Koń nr 3



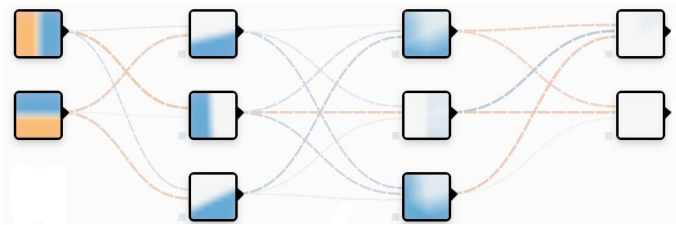


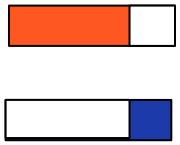
p_1 koń

p_2 osioł

Choroba sieci neuronowych
Pewny, ale błędny

Koń Rafał

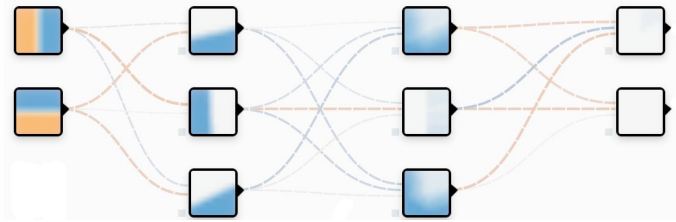




p_1 **koń**

p_2 osioł

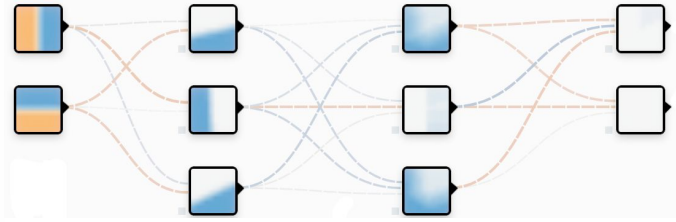
Koń nr 2



p_1 koń

p_2 osioł

Koń nr 3

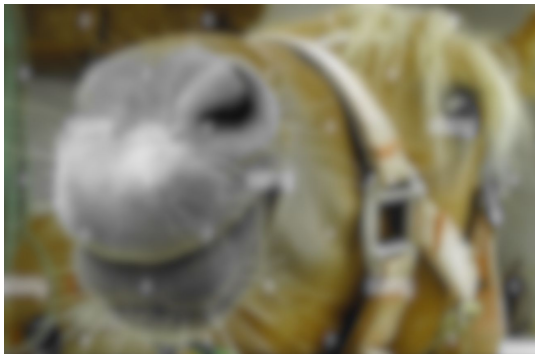


p_1 koń

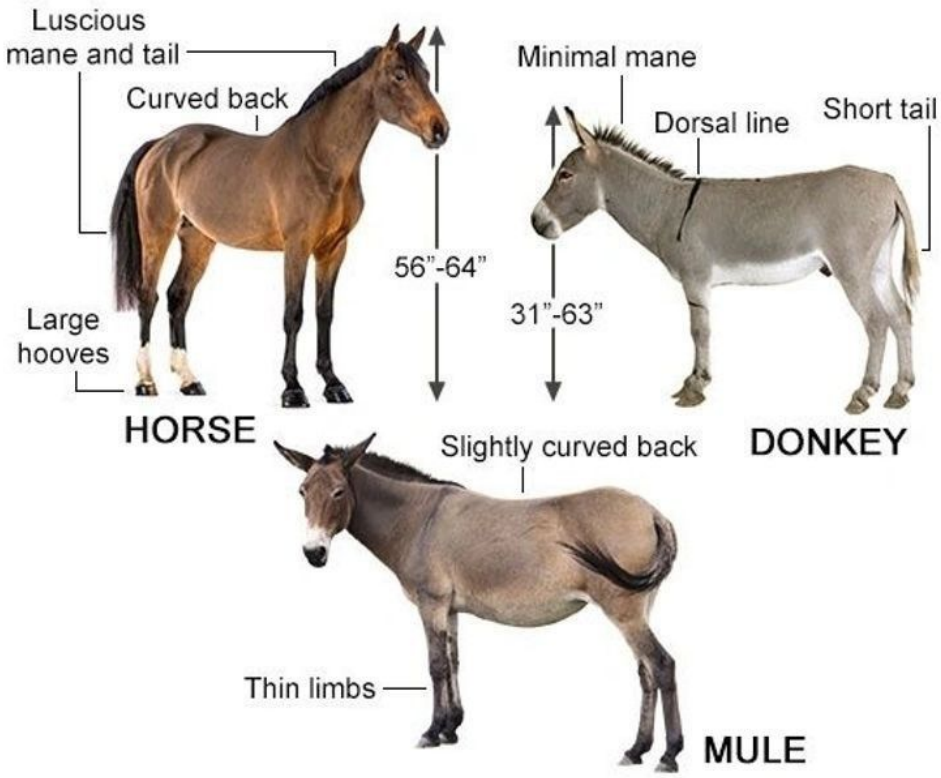
p_2 **osioł**

Ale czym jest 50%-50%?

Opcja 1: **nie wiem**



Opcja 2: **muł** (pół koń, pół osioł)



https://thebritishmulesociety.com/All-About-Mules/

Co z anomaliami?

Opcja 3: **nie znam**

Dlaczego (nie)pewność jest ważna?

- Diagnozy medyczne
- Porady finansowe
- Edukacja
- Decyzje prawne

Nie tylko modele dyskryminatywne:
LLM to modele przewidujące następne słowo

wejście

model

wyjście
(predykcja)

I need you to

help

dodaj wyjście na
koniec wejścia

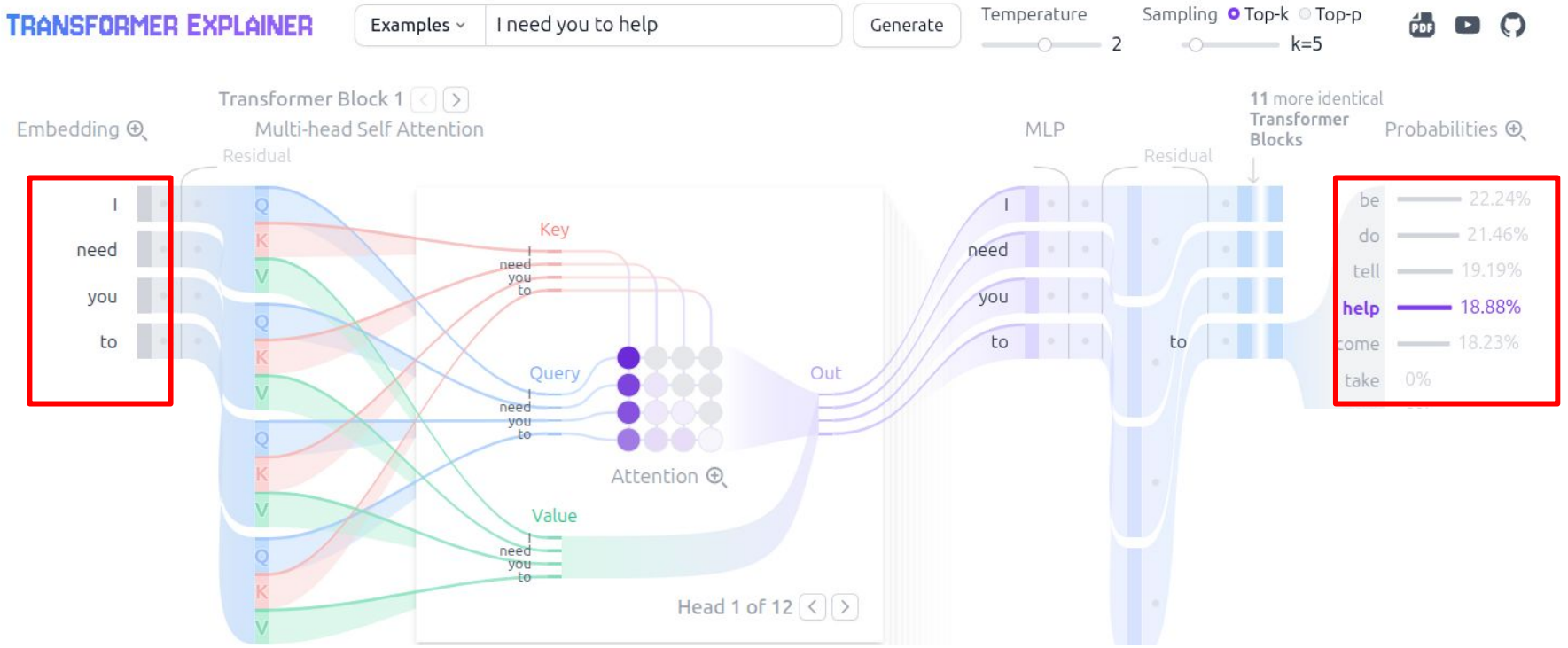
LLM (ang. Large language models)

Gemini

Claude

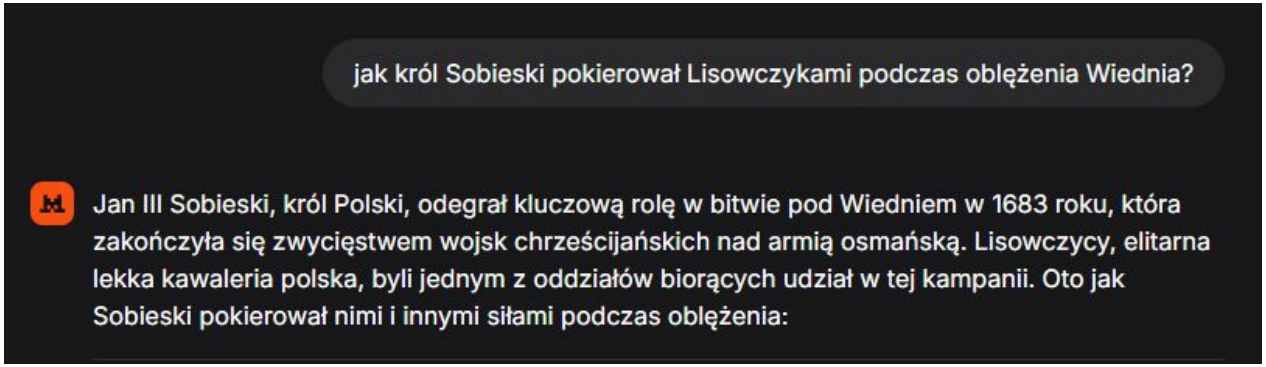
MISTRAL AI_

LLM to modele przewidujące następne słowo



https://poloclub.github.io/transformer-explainer/

Nie tylko modele dyskryminatywne...



Nie tylko modele dyskryminatywne...

jak król Sobieski pokierował Lisowczykami podczas oblężenia Wiednia?

Jan III Sobieski, król Polski, odegrał kluczową rolę w bitwie pod Wiedniem w 1683 roku, która zakończyła się zwycięstwem wojsk chrześcijańskich nad armią osmańską. Lisowczycy, elitarna lekka kawaleria polska, byli jednym z oddziałów biorących udział w tej kampanii. Oto jak Sobieski pokierował nimi i innymi siłami podczas oblężenia:

Lisowczyk – mal. Juliusz Kossak (ok. 1860–1865; kopia obrazu Rembrandta *Jeździec polski*)

Historia	
Państwo	I Rzeczpospolita
Sformowanie	1607
Rozformowanie	1636

[Lisowczycy – Wikipedia, wolna encyklopedia]

Niepewność epistemiczna i aleatoryczna
Gra w cukierki

Będziemy przewidywać jakie cukierki wyciągniemy z czapki

Potrzebuję dwójkę wolontariuszy!

Gra w cukierki
Runda 1: 1 cukierek kumulatywnie

Jaki będzie następny cukierek?

Gra w cukierki
Runda 2: 2 cukierki kumulatywnie

No dobra, a jaki cukierek będzie teraz?

Gra w cukierki
Runda 3: 3 cukierków kumulatywnie

Czy coś wiemy więcej?

Gra w cukierki
Runda 4: 4 cukierków kumulatywnie

Ostatnia runda pierwszego zawodnika.

Gra w cukierki:
Rundy 5-8: 5-8 cukierków kumulatywnie

Czy Wasze przewidywania się zmieniły?

Gra w cukierki
Rozwiązanie

Dostarczono nam pewną liczbę *N* cukierków różnego typu.
Nieznane nam: *a*, *b*, *c* (liczba krówek, mieszanek i michałków).
Interesują nas proporcję.

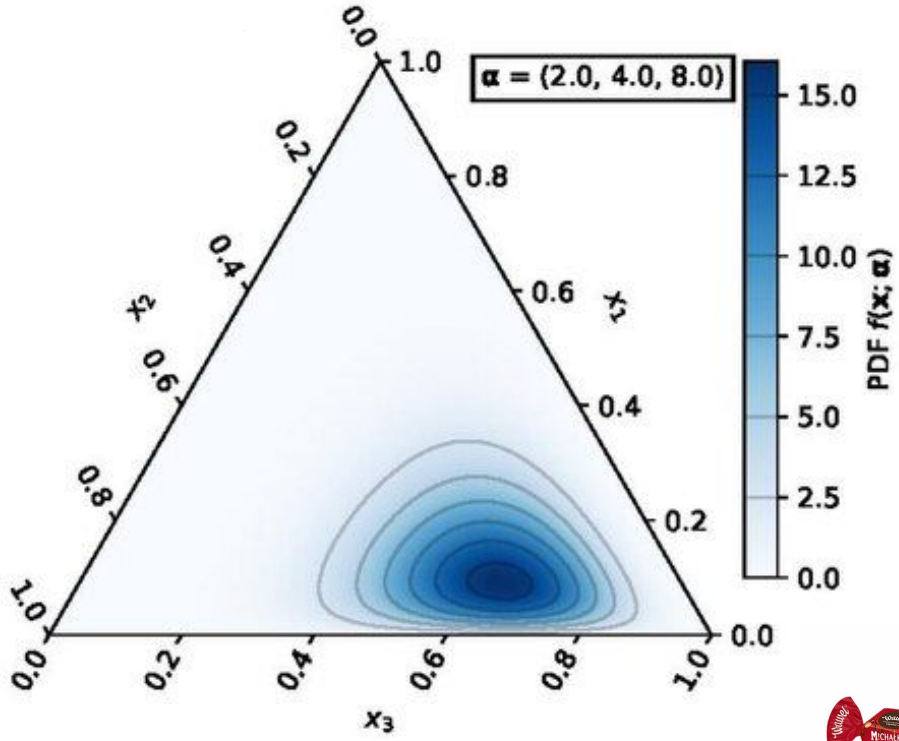
Dopiero gdy wylosujemy wszystkie cukierki poznamy dokładnie *a, b, c*
poznamy dokładnie udział każdego typu cukierków

Ale czy możemy powiedzieć coś wcześniej?

No tak! Przecież już wcześniej mieliście przypuszczenie, że nie ma Michałków.

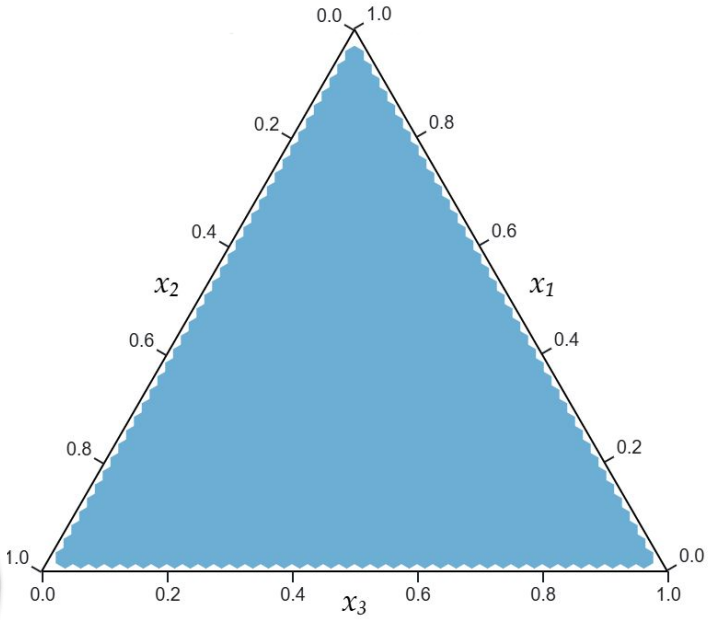
W rzeczywistości nie mamy wszystkich danych

Narysujmy to!
Rozkład Dirichleta



Online demo: Gra z cukierkami

[Dirichlet Distribution / Herb Susmann | Observable]

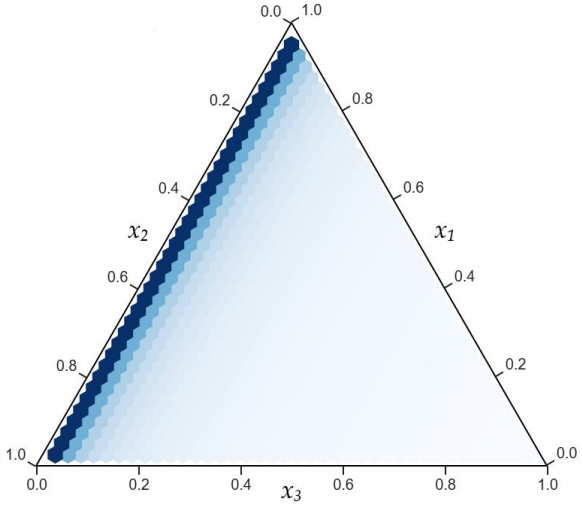


Bardziej informatywny start

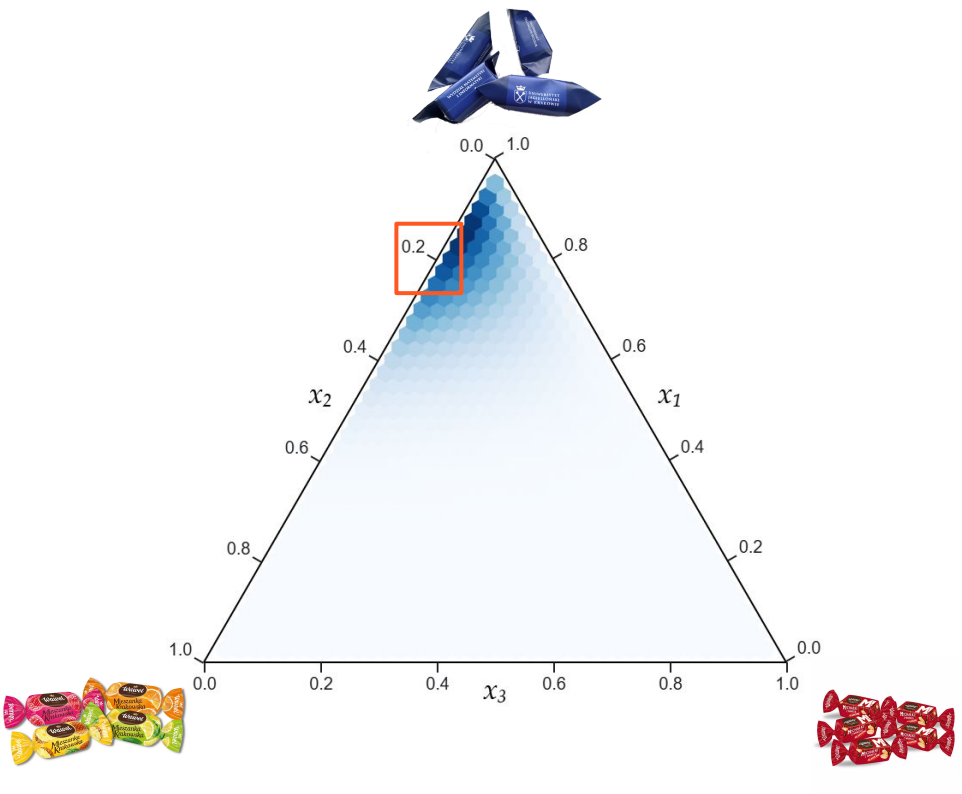
A co jeżeli wiemy od samego początku,
że są **tylko dwa typy** cukierków w torbie?

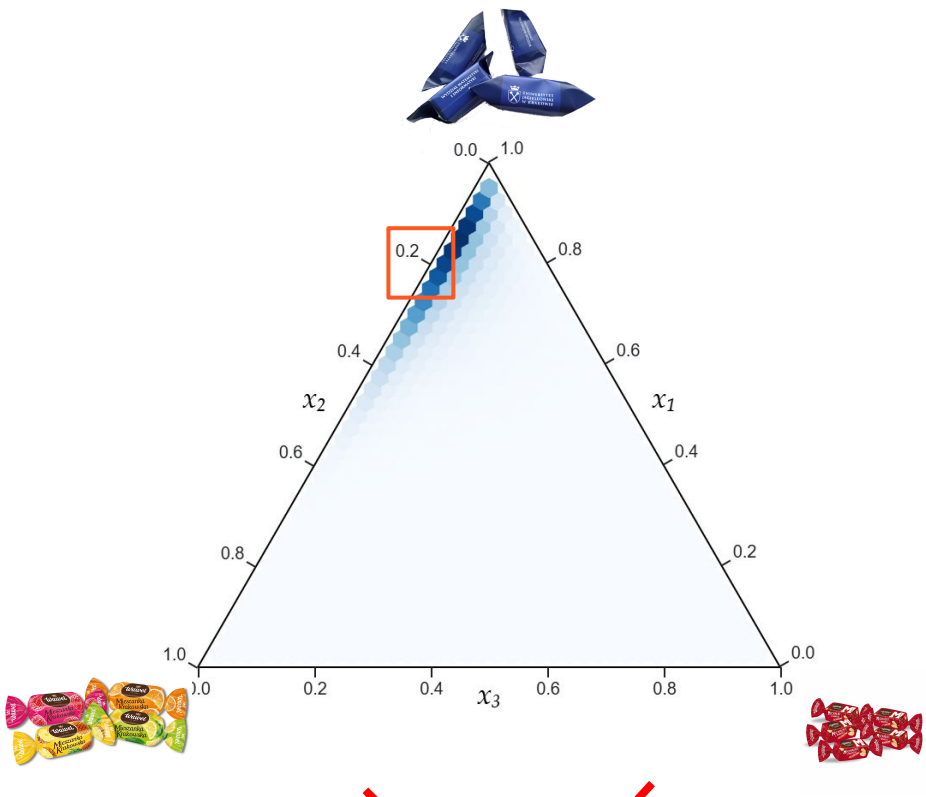
Zmieniamy *początkowe kolorowanie tak*,
żeby niemożliwy jest trzeci typ cukierka.

Dla lepszego *prioru* uczenie przebiega szybciej
to znaczy, że wcześniej będziemy w pobliżu
poprawnego rozwiązania



Po rundzie 4 (łącznie 10 cukierków)



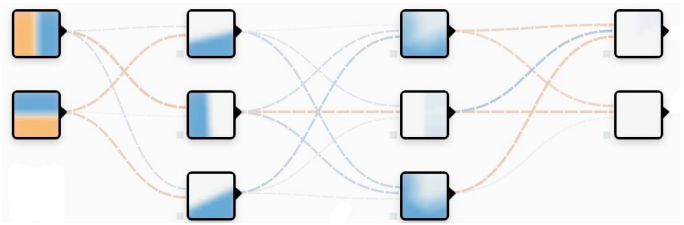


Wiele typów niepewności

Niepewność aleatoryczna

- naturalna losowość zjawiska
- ***ML:*** *Model nigdy nie „zgadnie” tego idealnie, bo to zmienność wpisana w rzeczywistość.*
  →Nawet gdy dokładnie znasz skład torebki,
  wynik pojedynczego losowania pozostaje losowy.

Model punktowy vs. rozkład



parametry θ

rozkład nad parametrami

$p(\theta|\mathcal{D})$

niepewność modelu!

Wiele typów niepewności

Tę modelują standardowe modele SI

Niepewność aleatoryczna

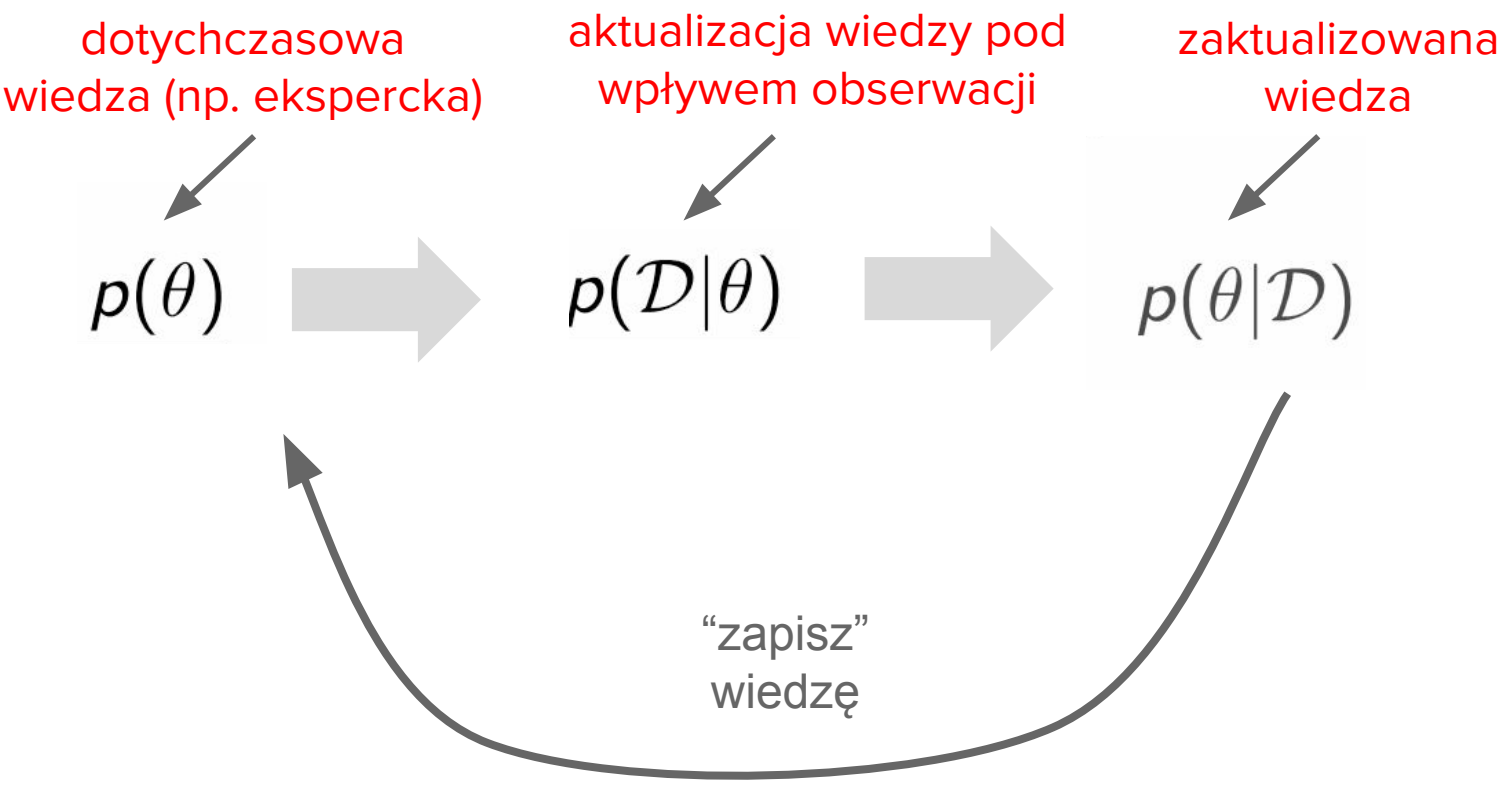
- naturalna losowość zjawiska
- ***ML:*** *Model nigdy nie „zgadnie” tego idealnie, bo to zmienność wpisana w rzeczywistość.*
  →Nawet gdy dokładnie znasz skład torebki, wynik pojedynczego losowania pozostaje losowy.

Tej nie modelujemy dla parametrów modeli

Niepewność epistemiczna

- wynika z braku wiedzy
- ***ML:*** *Możesz ją zniwelować, douczając model nowymi obserwacjami*
  →Jeśli zaczniemy losować wiele cukierków, możemy oszacować proporcje.

Uczenie Bayesowskie



Twierdzenie Bayesa

wiedza “ekspercka”

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta)}{p(\mathcal{D})}$$

niepewność modelu!

Lepsze modele!

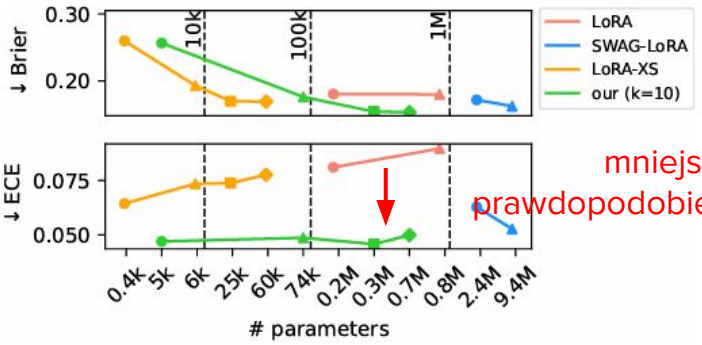
Minimal Ranks, Maximum Confidence: Parameter-efficient Uncertainty Quantification for LoRA

Patryk Marszałek* **Klaudia Bałazy** **Jacek Tabor** **Tomasz Kuśmierczyk***†

Jagiellonian University

Abstract

Low-Rank Adaptation (LoRA) enables parameter-efficient fine-tuning of large language models by decomposing weight updates into low-rank matrices, significantly reducing storage and computational overhead. While effective, standard LoRA lacks mechanisms for uncertainty quantification, leading to overconfident and poorly calibrated models. Bayesian variants of LoRA address this limitation, but at the cost of a significantly increased number of trainable parameters, partially offsetting

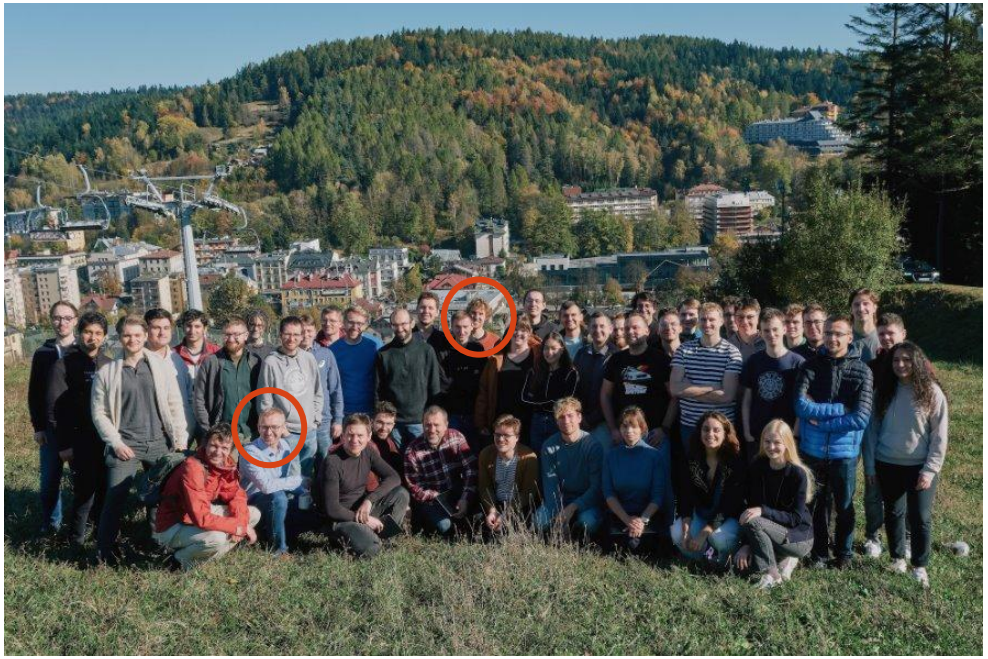


mniejszy błąd prawdopodobieństw (kalibracji)

Figure 1: Performance averaged over multiple GLUE datasets (individual results in Fig. 3). Our method achieves superior calibration (ECE) and competitive

group of machine
gmum
learning research

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE



Dzięki za uwagę!

Olimpiada AI

Zainteresuj się!

+ Kółka przygotowujące w czwartki na UJ
+ Pierwszy kontakt z badaniami naukowymi

Rank	
1	Poland Team 1
2	Russia Team 1
3	Poland Team 2
4	Hungary Team 1
5	Vietnam Team 2
6	IOAI TEAM

https://bayes.ii.uj.edu.pl

Badania zrealizowane w ramach projektu nr ***2022/45/P/ST6/02969*** *współfinansowanego ze środków Narodowego Centrum Nauki oraz programu ramowego Unii Europejskiej w zakresie badań naukowych i innowacji Horyzont 2020 na podstawie umowy nr 945339 w ramach działań „Marie Skłodowska-Curie”. Dla celów Open Access autor udostępnia na licencji CC-BY (Creative Commons – uznanie autorstwa) każdą wersję AAM, która może powstać w oparciu o niniejszy manuskrypt;*

POLONEZ BIS

NARODOWE CENTRUM NAUKI