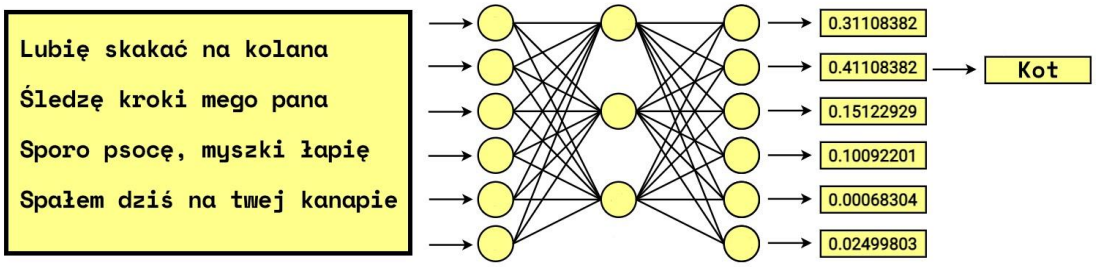
Wyjaśnialna sztuczna inteligencja – przegląd literatury

Stanisław Kaźmierczak

Agenda

1. Wprowadzenie
2. Interpretowalność a jakość działania
3. Po co wyjaśniać modele?
4. Modele interpretowalne
5. Rodzaje wyjaśniania
6. Taksonomia
7. LIME
8. SP-LIME

Przykład



[1]

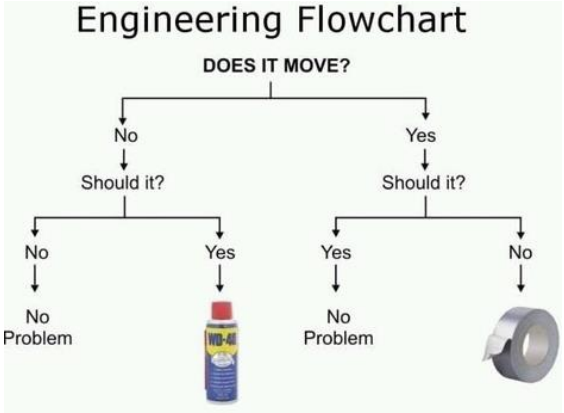
AI --> To jest kot, ponieważ:

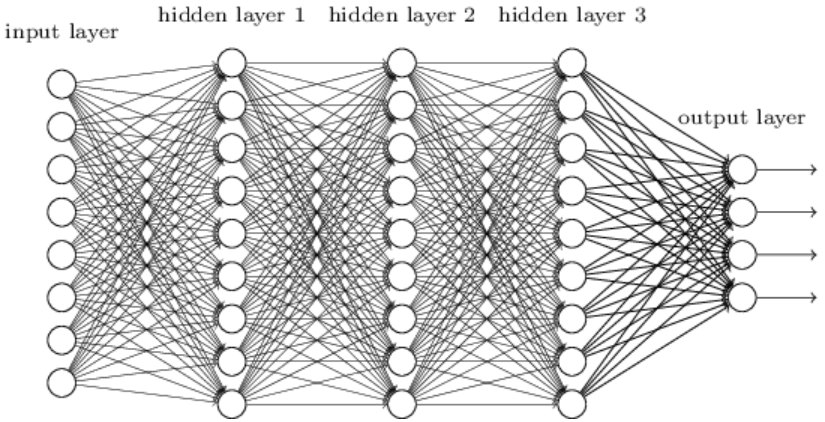
- wartość neuronu odpowiadającego kategorii „kot” w warstwie wyjściowej była najwyższa

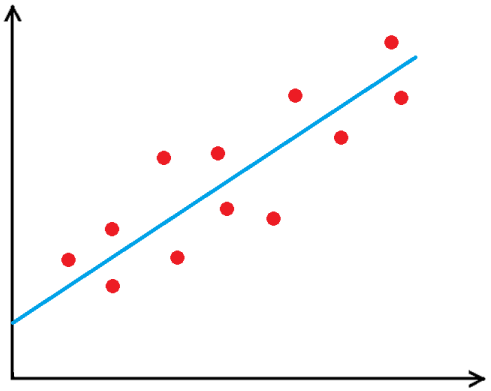
XAI (*Explainable AI*) --> To jest kot, ponieważ:

- Koty lubią skakać na kolana
- Koty są zwierzętami terytorialnymi
- Koty polują na myszy
- Koty dużo śpią (średnio 12 – 14 godzin na dobę)

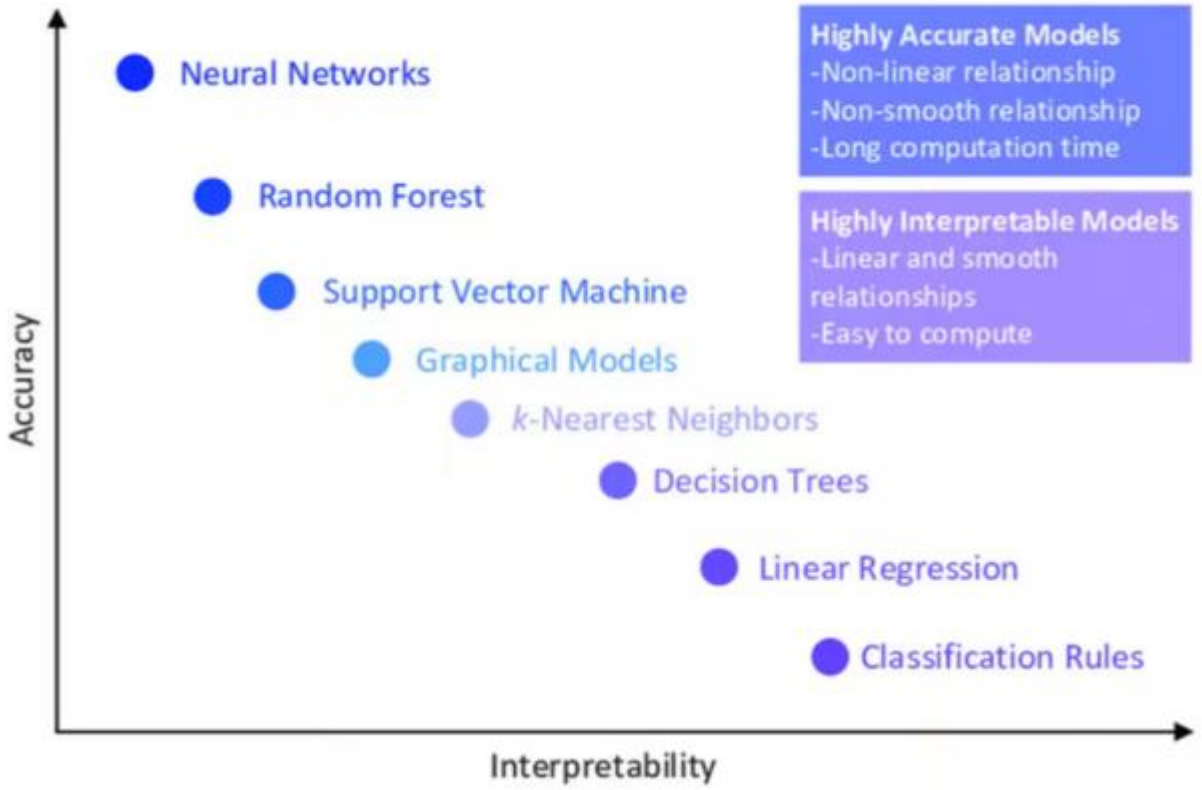
Interpretowalność a jakość działania (1)







Interpretowalność a jakość działania (2)



[2]

Gartner Hype Cycle

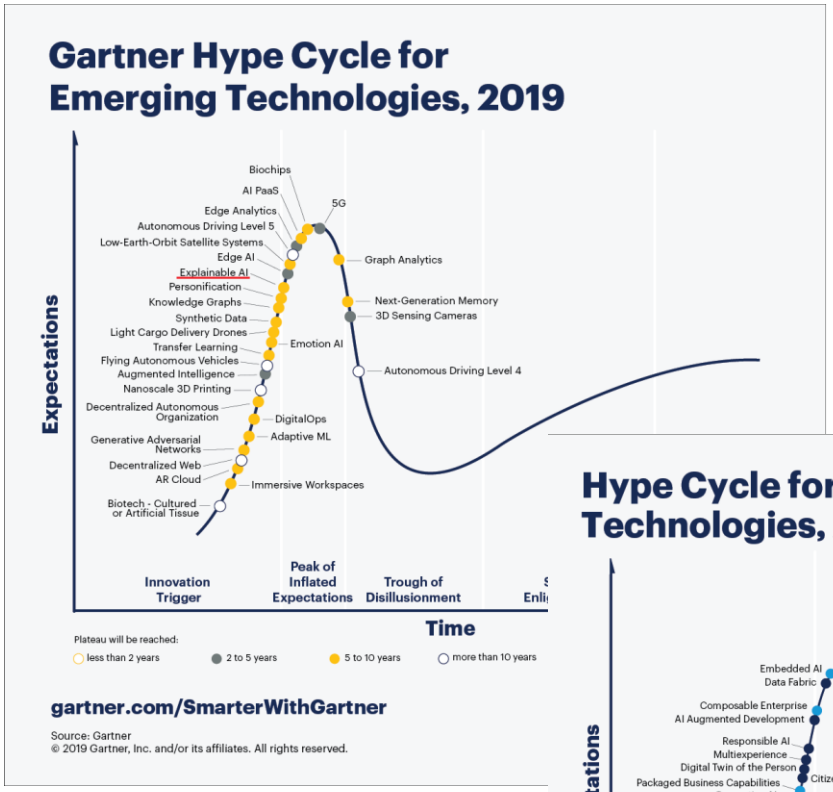
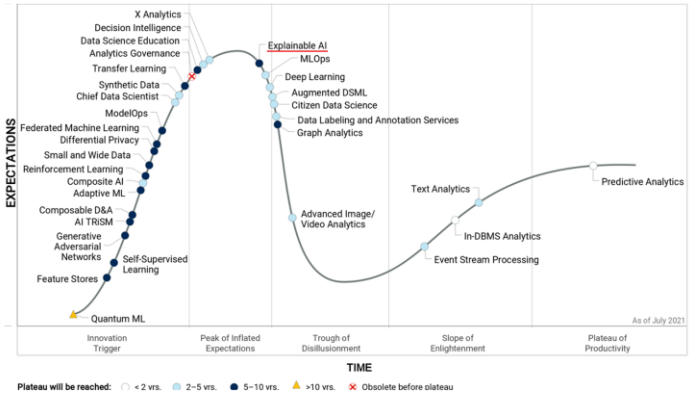
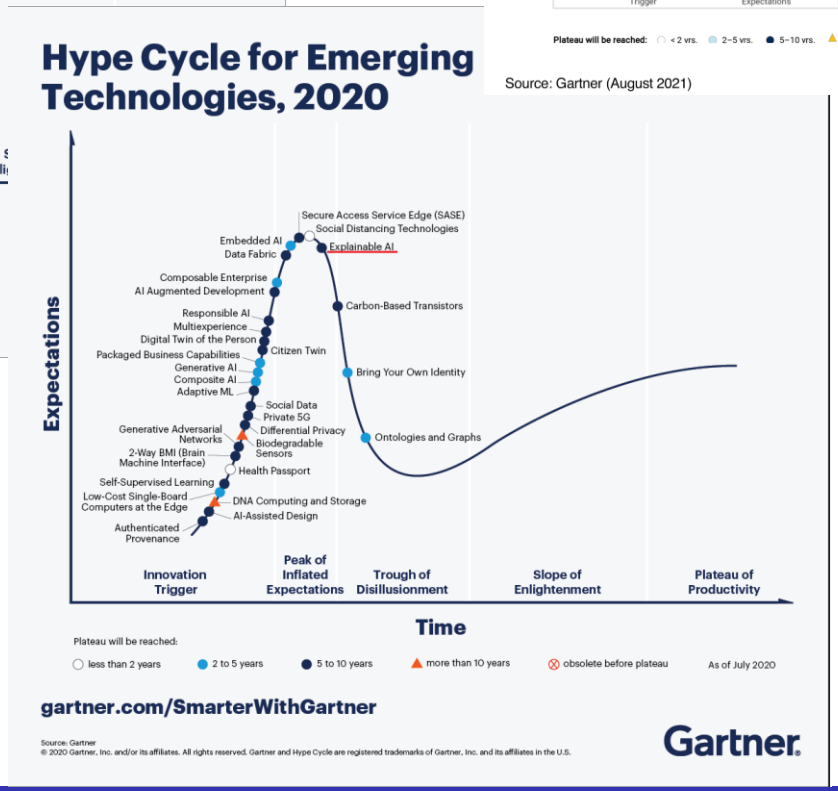


Figure 1: Hype Cycle for Data Science and Machine Learning, 2021



Source: Gartner (August 2021)

Gartner

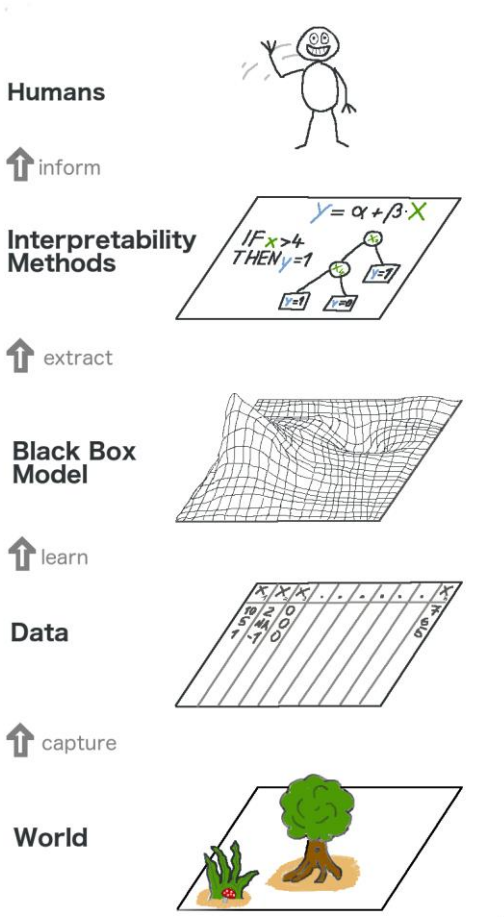


Po co wyjaśniać modele?

- Decyzje bazujące na AI muszą być wyjaśnialne, aby można im zaufać (84% respondentów, 22nd Annual Global CEO Survey)
- Bezpieczeństwo w krytycznych obszarach, np. diagnozy medyczne, autonomiczne pojazdy
  - Aby móc zaudytować model, musi być on interpretowalny
- Regulacje prawne, np. konieczność uzasadnienia decyzji o odmowie udzielenia kredytu
- Lepszy efekt współpracy człowiek (ekspert) + AI
- Akceptacja społeczna
  - Część społeczeństwa boi się AI, ponieważ nie wie, jak działa
- Debugowanie, np. wykrycie, że model podejmuje decyzje na podstawie mało istotnych detali/szumu w danych
- Ludzka ciekawość

Wyjaśnialne AI

- Terminy „wyjaśnialny” i „interpretowalny” są pojęciami dość miękkimi
- Przykłady wyjaśnień
  - Wskazanie zbioru cech, które są kluczowe z punktu widzenia danego problemu predykcyjnego (wyjaśnienie globalne)
  - Wskazanie powodu, dlaczego dana instancja została sklasyfikowana w taki, a nie inny sposób (wyjaśnienie lokalne)

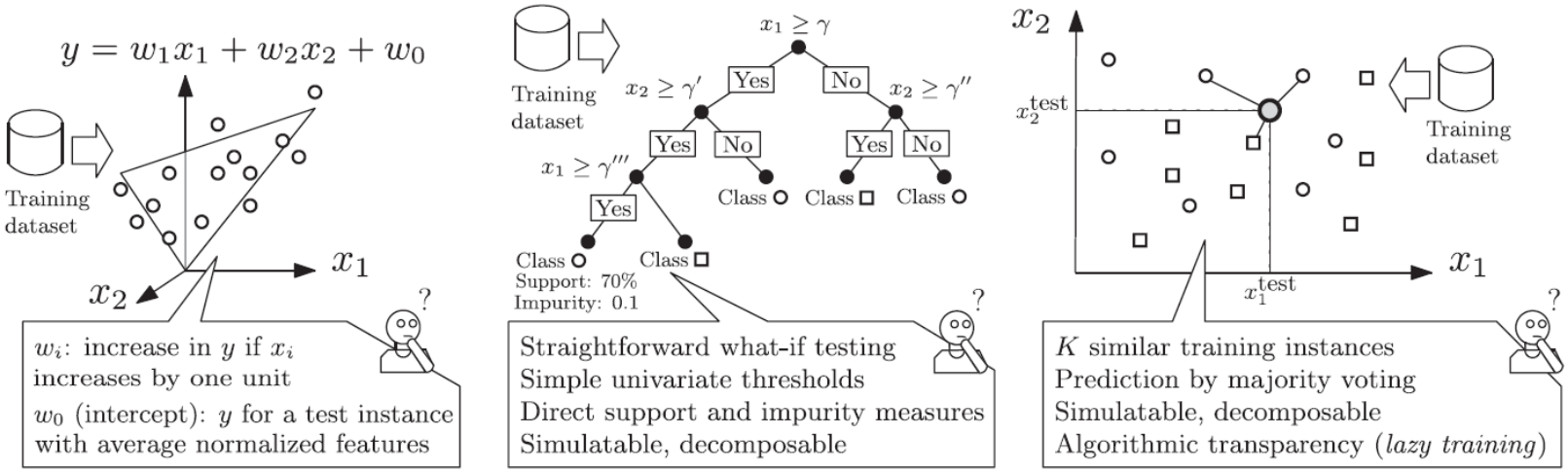


[3]

Modele interpretowalne

Niektóre modele są ze swej natury wyjaśnialne, np.

- Modele regresyjne
- Drzewa decyzyjne
- K najbliższych sąsiadów
- Reguły decyzyjne, np. One rule
- Modele bayesowskie



[4]

Literatura

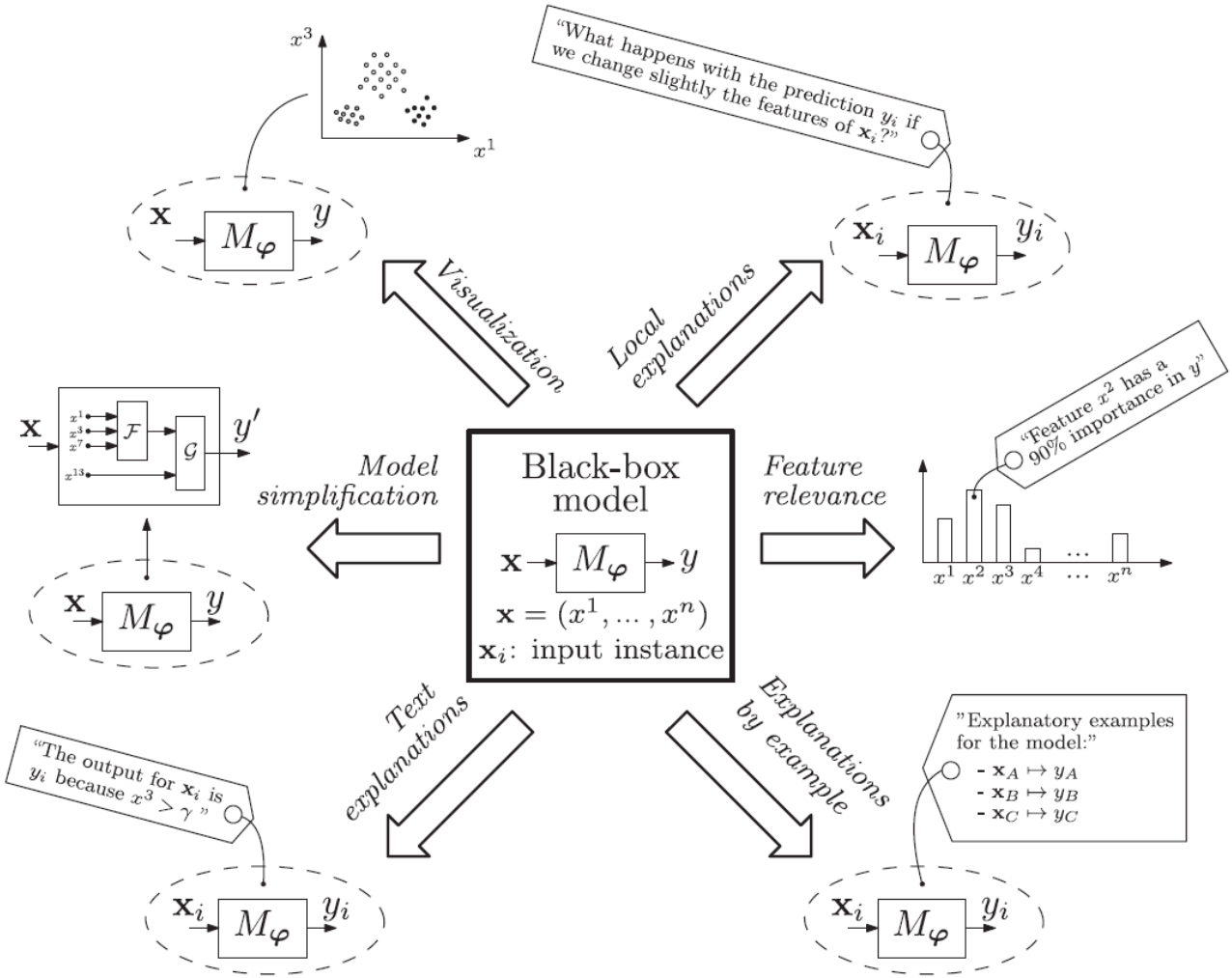
Prace przeglądowe (survey):

- **Arrieta, A. B. et al. (2020). *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*.**
- Vilone, G., & Longo, L. (2020). *Explainable artificial intelligence: a systematic review*.
- Islam, S. et al. (2021). *Explainable Artificial Intelligence Approaches: A Survey*.

Monografie:

- Molnar, C. (2020). *Interpretable machine learning*.
- Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models*.

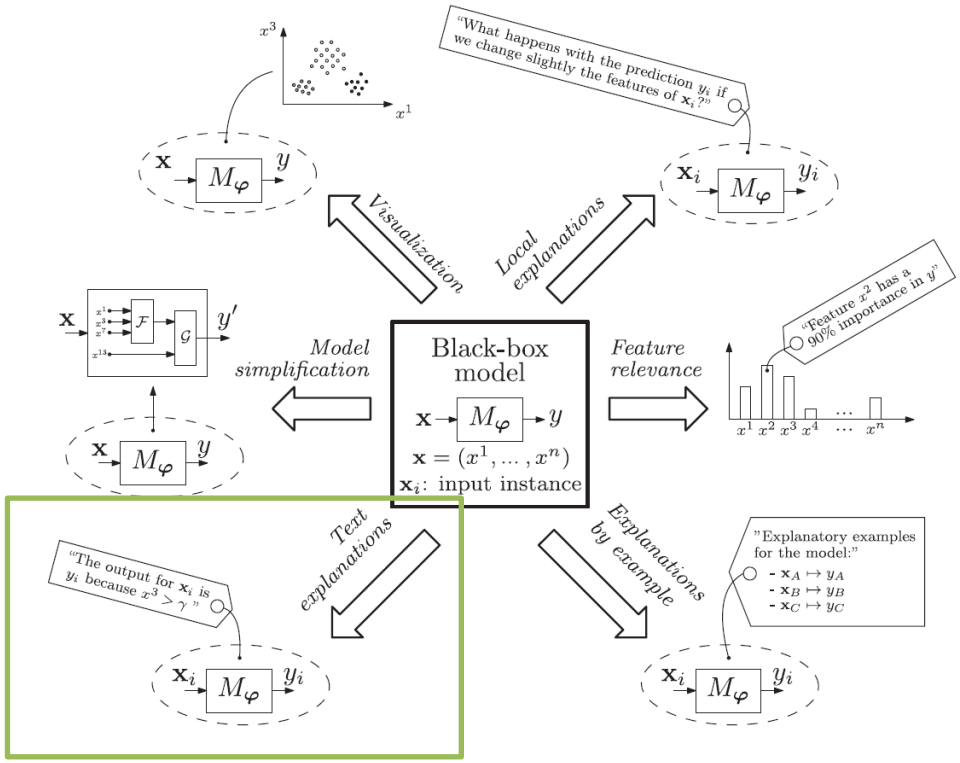
Metody wyjaśniania



Wyjaśnienie tekstem

Wyjaśnienie tekstem

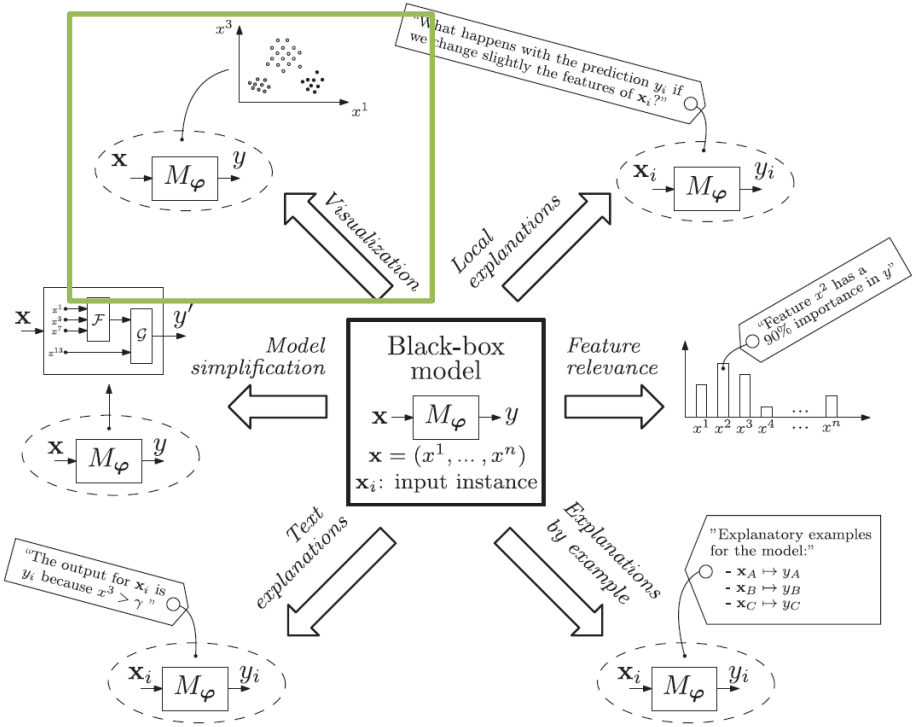
- Językiem naturalnym
- Symbolami reprezentującymi model



Wyjaśnienie wizualizacją

Wyjaśnienie wizualizacją

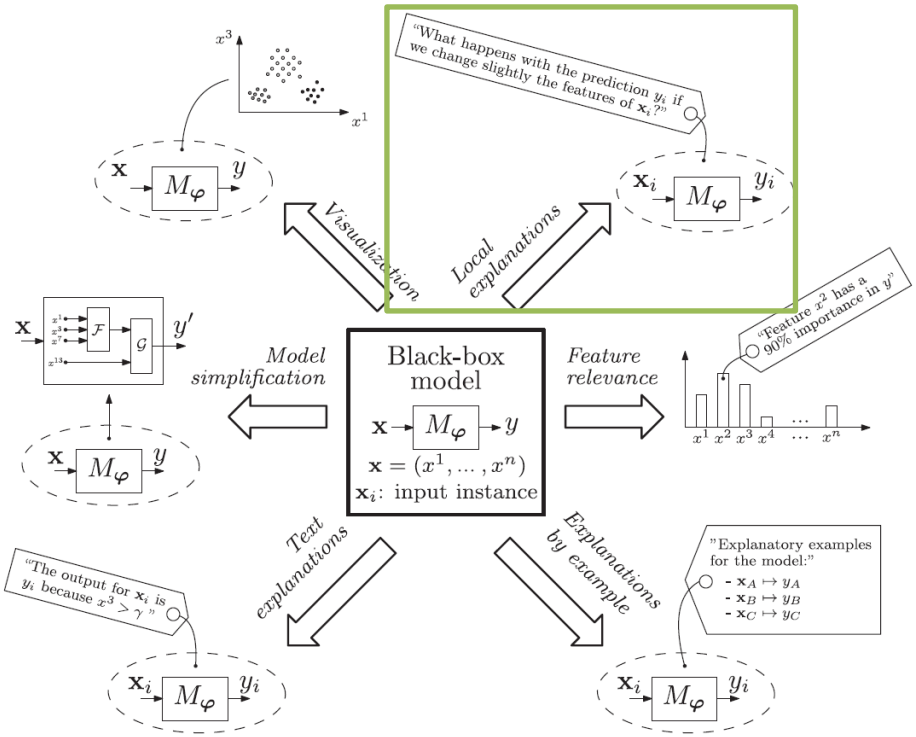
- Duża część metod bazuje na technikach redukcji wymiarowości do 2/3 wymiarów
- Alternatywnie poprzez wybór zestawów 2/3 cech



Lokalne wyjaśnienie

Lokalne wyjaśnienie

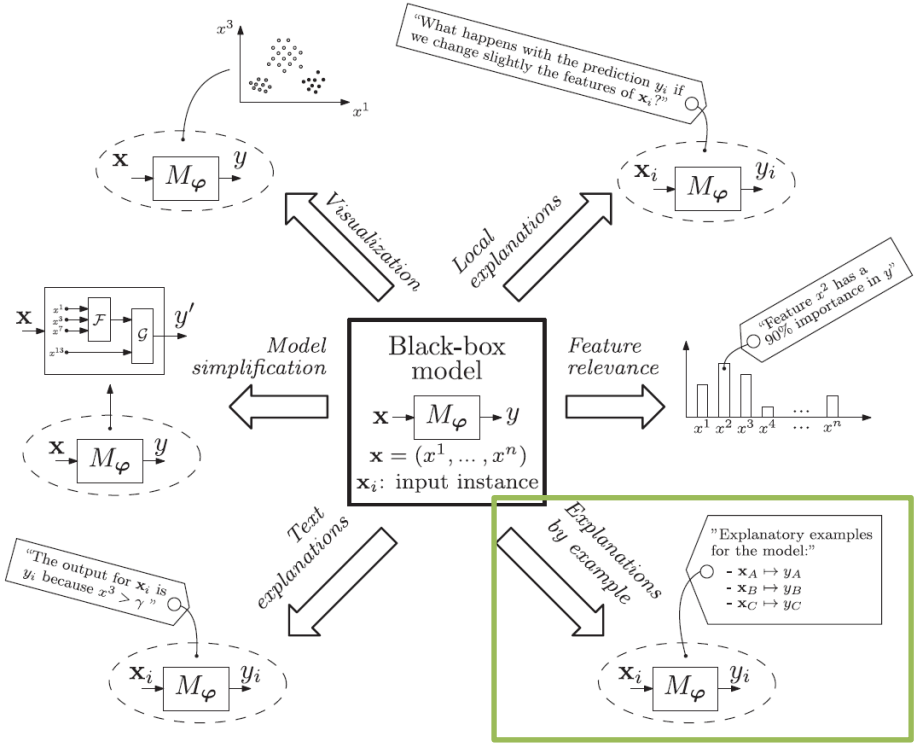
- Segmentacja przestrzeni rozwiązań
- Oddzielne wyjaśnienia dla mniej złożonych podprzestrzeni rozwiązań



Wyjaśnienie przez przykład

Wyjaśnienie przez przykład

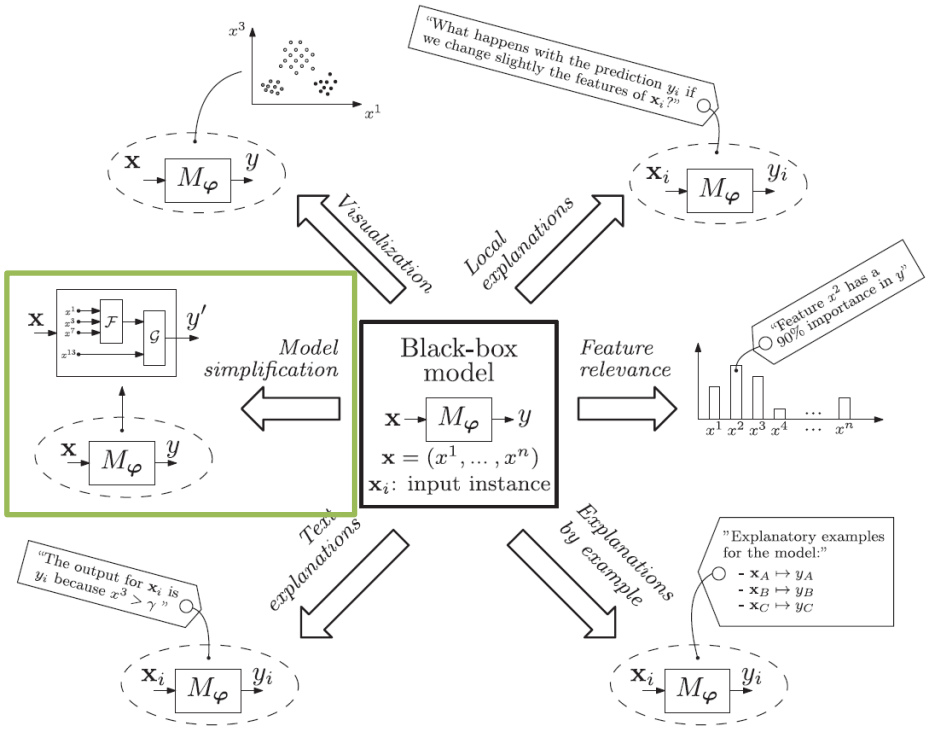
- Selekcja reprezentatywnych przykładów uchwytujące relacje znalezione przez model
- Podobieństwo do ludzkiego „Wyjaśnijmy to na przykładzie”



Wyjaśnienie przez uproszczenie

Wyjaśnienie przez uproszczenie

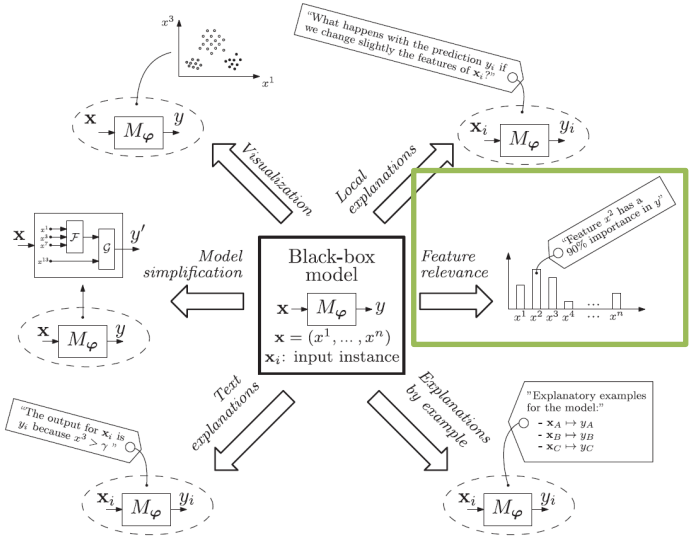
- Wymaga zbudowania nowego systemu
- Cel: utrzymanie podobieństwa w działaniu i podobnej jakości działania przy zmniejszonej złożoności



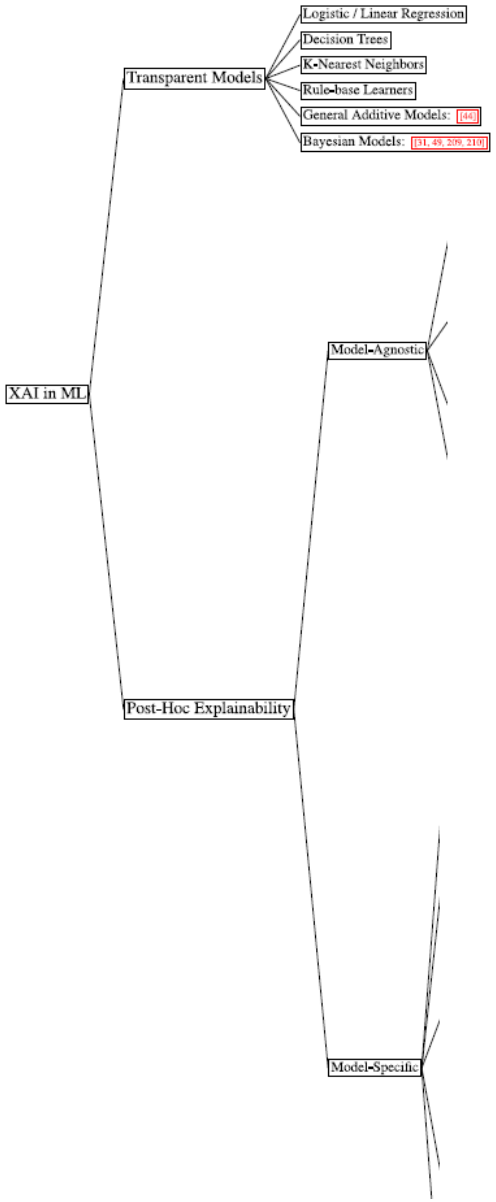
Wyjaśnienie istotności cech

Wyjaśnienie istotności cech

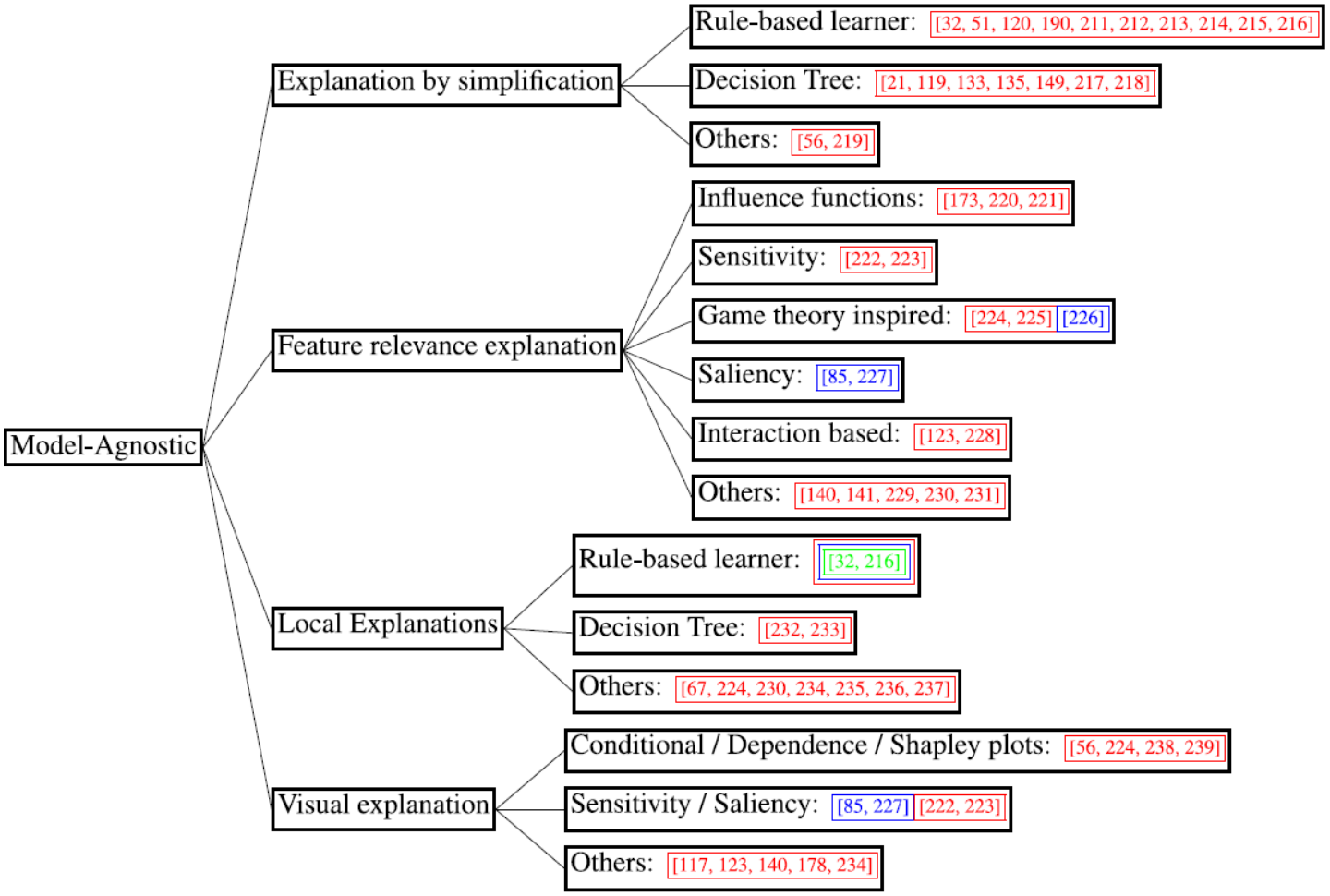
- Dla każdej cechy zwracany jest konkretny wynik liczbowy
- Przykłady:
  - współczynniki w regresji liniowej
  - wagi w pierwszej warstwie sieci neuronowej
  - Współczynnik Giniego w drzewie decyzyjnym/lesie losowym
  - Liczba wystąpień cechy w węzłach drzewa/lasu



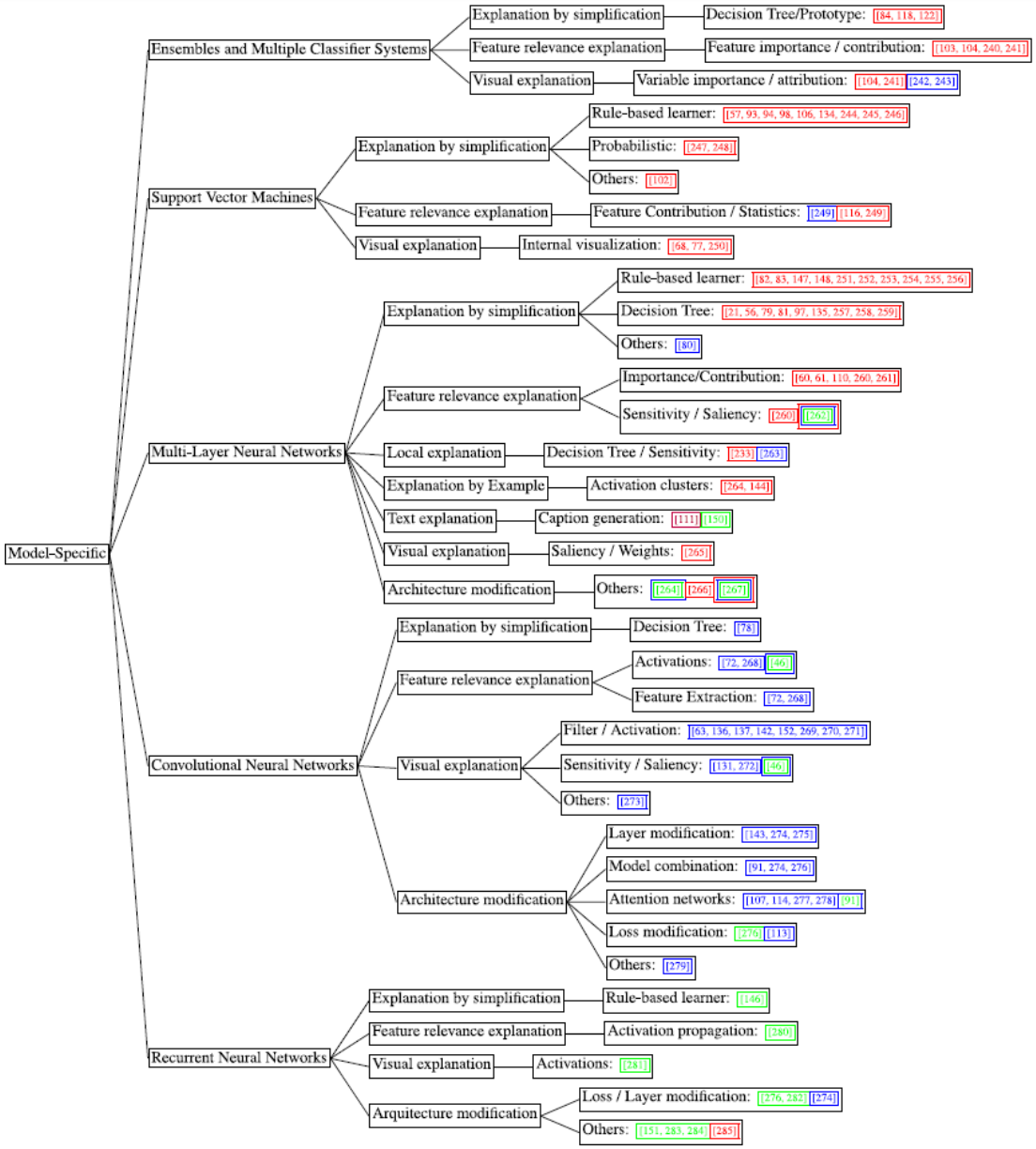
Taksonomia – najwyższy poziom



Taksonomia – metody Model-Agnostic



Taksonomia – metody Model-Specific



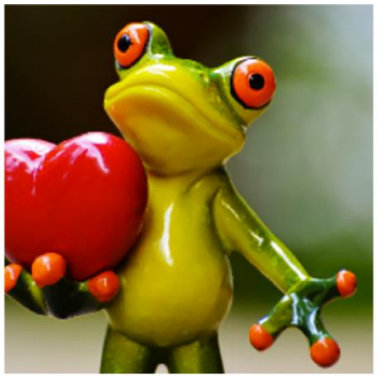
LIME

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

- *Local Interpretable Model-Agnostic Explanations*
  - Lokalna interpretacja (dla pojedynczych instancji)
  - Niezależny od modelu (model bazowy – czarna skrzynka)
- Ogólna idea: zaburz wejście i sprawdź, jak zmieni się predykcja
- Działa dla różnego rodzaju danych (tabelaryczne, obrazowe, językowe)

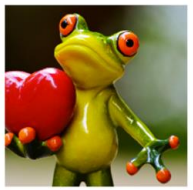
LIME dla obrazów (1)

- Podział obrazu na interpretowalne komponenty (*contiguous superpixels*)
- *Wyłączenie* (ustawienie na kolor szary) niektórych komponentów
- Dla każdej z zaburzonych instancji otrzymujemy prawdopodobieństwo poprawnej klasy

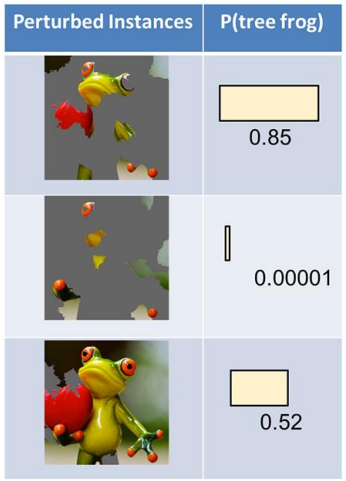


Original Image

Interpretable Components



Original Image
P(tree frog) = 0.54



[5]

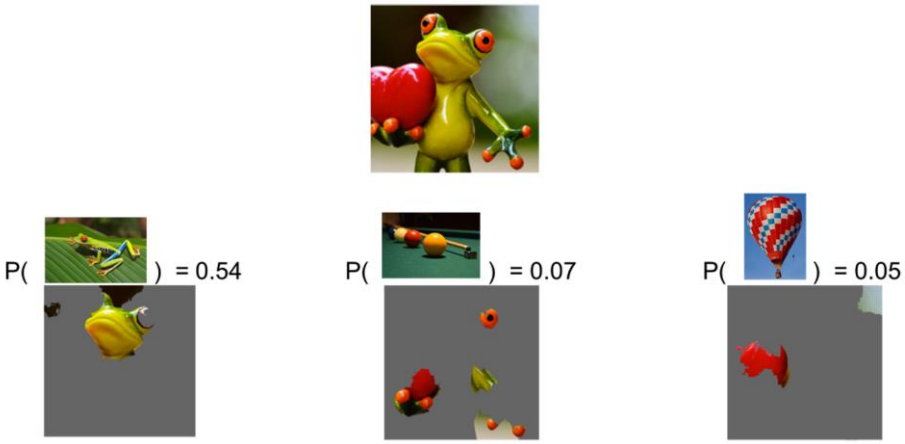
LIME dla obrazów (2)

- Tworzymy model regresyjny uczony na zaburzonych instancjach
  - Jest lokalnie ważony – zaburzone instancje, które bardziej przypominają obraz oryginalny mają większą wagę

1. Fit θ to minimize $\sum_i w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2$.
2. Output $\theta^T x$.

  - Cechami są superpiksele
  - Superpiksele z największymi współczynnikami wyznaczonymi przez algorytm regresji są krytyczne z punktu widzenia predykcji

LIME – interpretacja sieci Inception

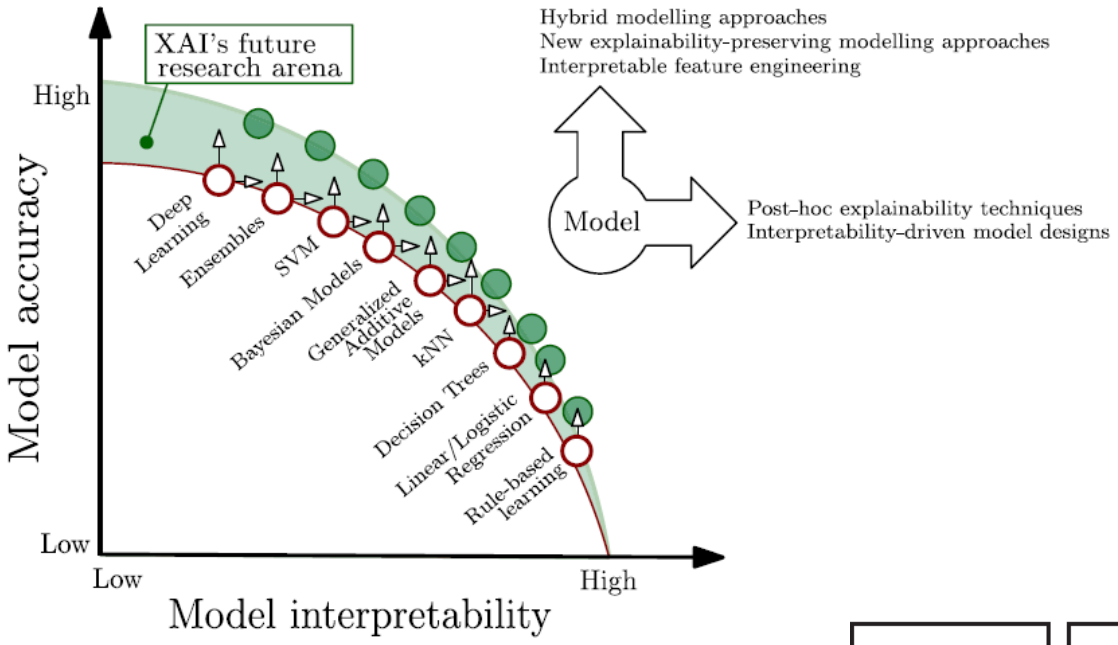


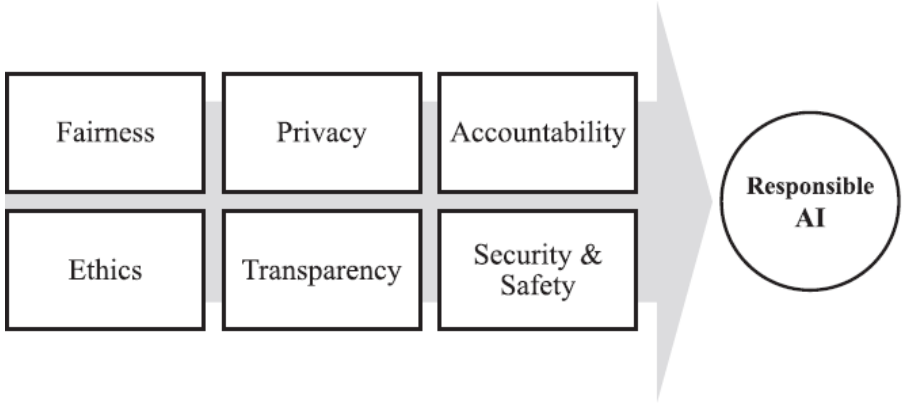
- Dla podanego zdjęcia przedstawione są 3 klasy, dla których sieć zwróciła największe prawdopodobieństwo
  - Ręce i oczy żaby rzekotki przypominają nieco kule bilardowe
  - Serce przypomina część balonu

SP-LIME

- Celem jest selekcja istotnych cech i stworzenie modelu globalnego
- Analizując zbiór instancji, widzimy które cechy są istotne lokalnie przy predykcji owych instancji
- Cechy, które są lokalnie istotne dla wielu instancji, są też globalnie istotne; pozostałe cechy nie są dalej rozpatrywane
- Następnie wybierany jest zbiór instancji, które pokrywają przestrzeń globalnie istotnych cech, a jednocześnie nie są redundantne między sobą w kontekście wyjaśniania

Podsumowanie (1)



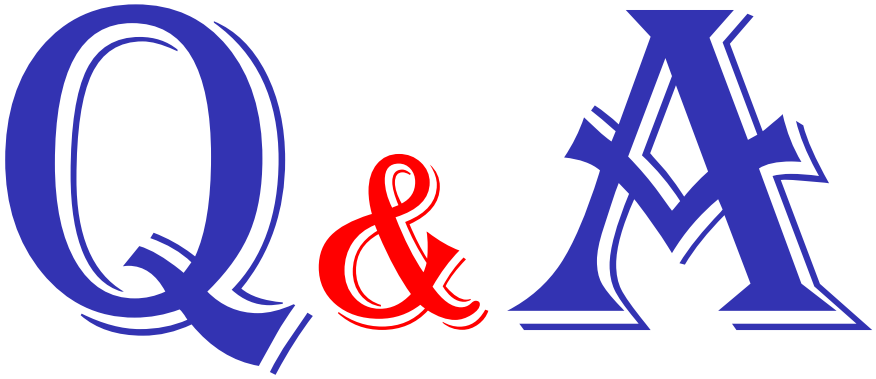


Podsumowanie (2)

- Niektórych ludzkich decyzji nie da się w sposób prosty i wyczerpujący wytłumaczyć
- Są to na przykład (błyskawiczne) decyzji oparte na intuicji (która z kolei tworzona jest na podstawie doświadczenia)
- Wydaje się więc, że w niektórych obszarach należy na pewnym etapie postawić kropkę i nie próbować wyjaśniać modeli za wszelką cenę

Źródła

1. https://blog.edrone.me/pl/xai-pl/
2. https://ichi.pro/pl/wyjasnialna-sztuczna-inteligencja-xai-100919187492012
3. Molnar, C. (2020). *Interpretable machine learning*
4. Arrieta, A. B. et al. (2020). *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*.
5. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/


Q&A