Metody wizualizacji wyników uczenia maszynowego

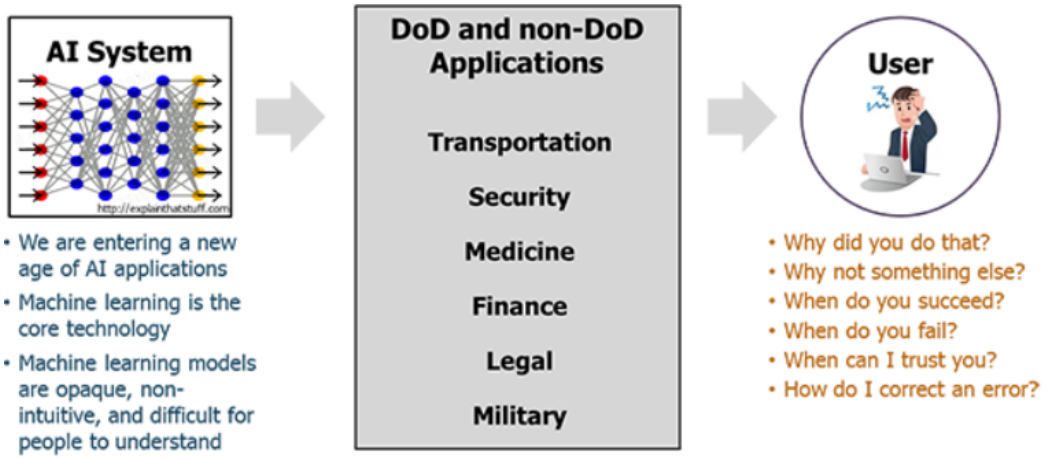
Stanisław Kaźmierczak

Agenda

- Wytłumaczalne AI
- Metody wyjaśniania modeli
- *LIME*
- *SP-LIME*
- Rezultaty

Wytłumaczalne AI

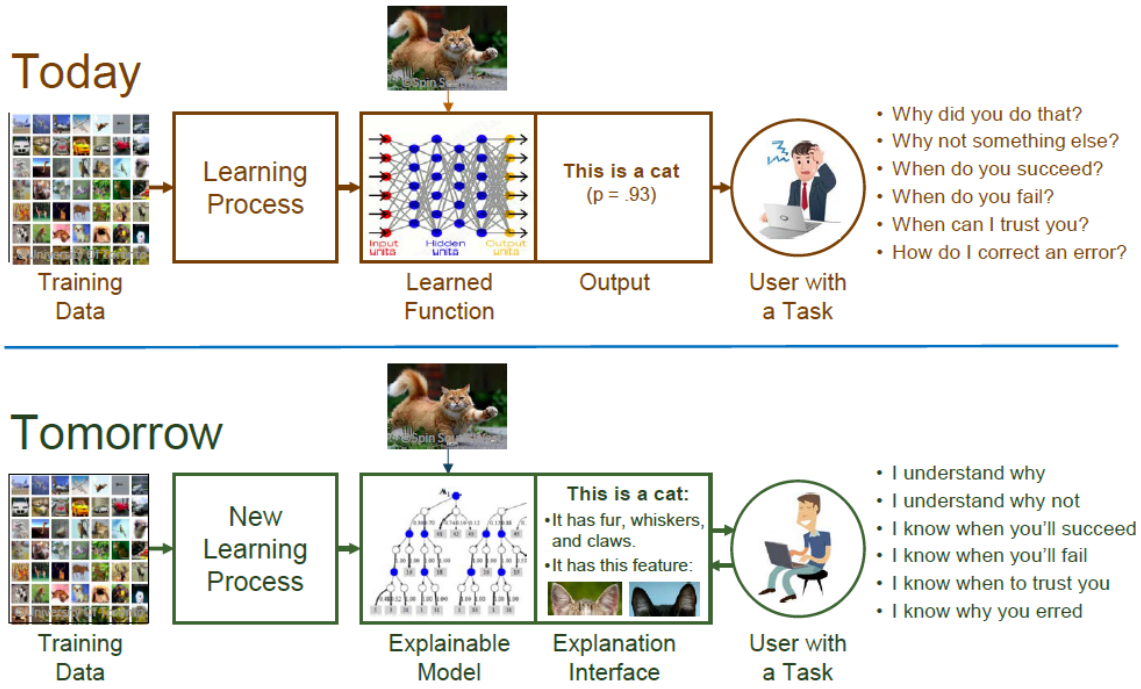
- Mimo wysokiej efektywności systemów AI, nie da się w prosty sposób wytłumaczyć użytkownikowi (człowiekowi), w jaki sposób została podjęta decyzja/działanie
- W wielu zastosowaniach intepretowalność procesu decyzyjnego jest mocno wskazana (np. systemy finansowe), a w niektórych wręcz konieczna (np. gdy w grę wchodzi ludzkie zdrowie i życie)
- Aspekty etyczne i prawne
- Wytłumaczalne/interpretowalne systemy AI stają się bardzo pożądane



[1]

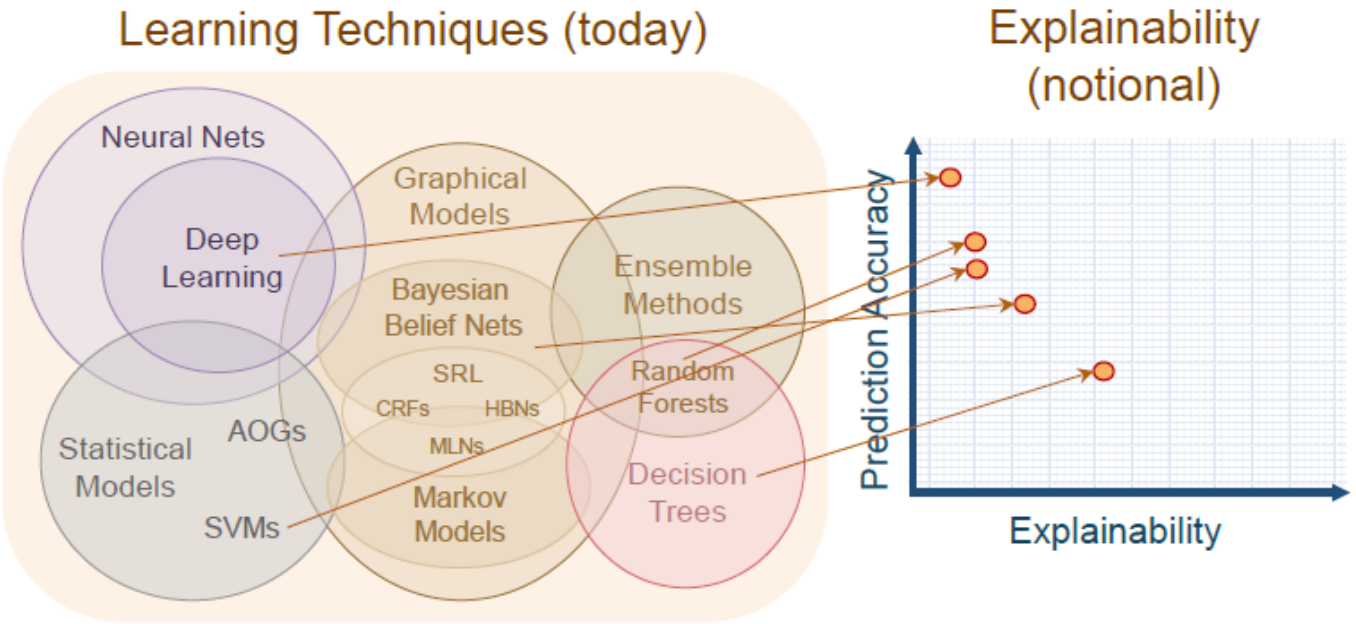
Wytłumaczalne AI

- Celem jest stworzenie modeli, które:
  - Są bardziej wyjaśnialne, a jednocześnie utrzymują wysoką efektywność
  - Pozwalają użytkownikowi rozumieć, ufać i skutecznie z nimi współpracować



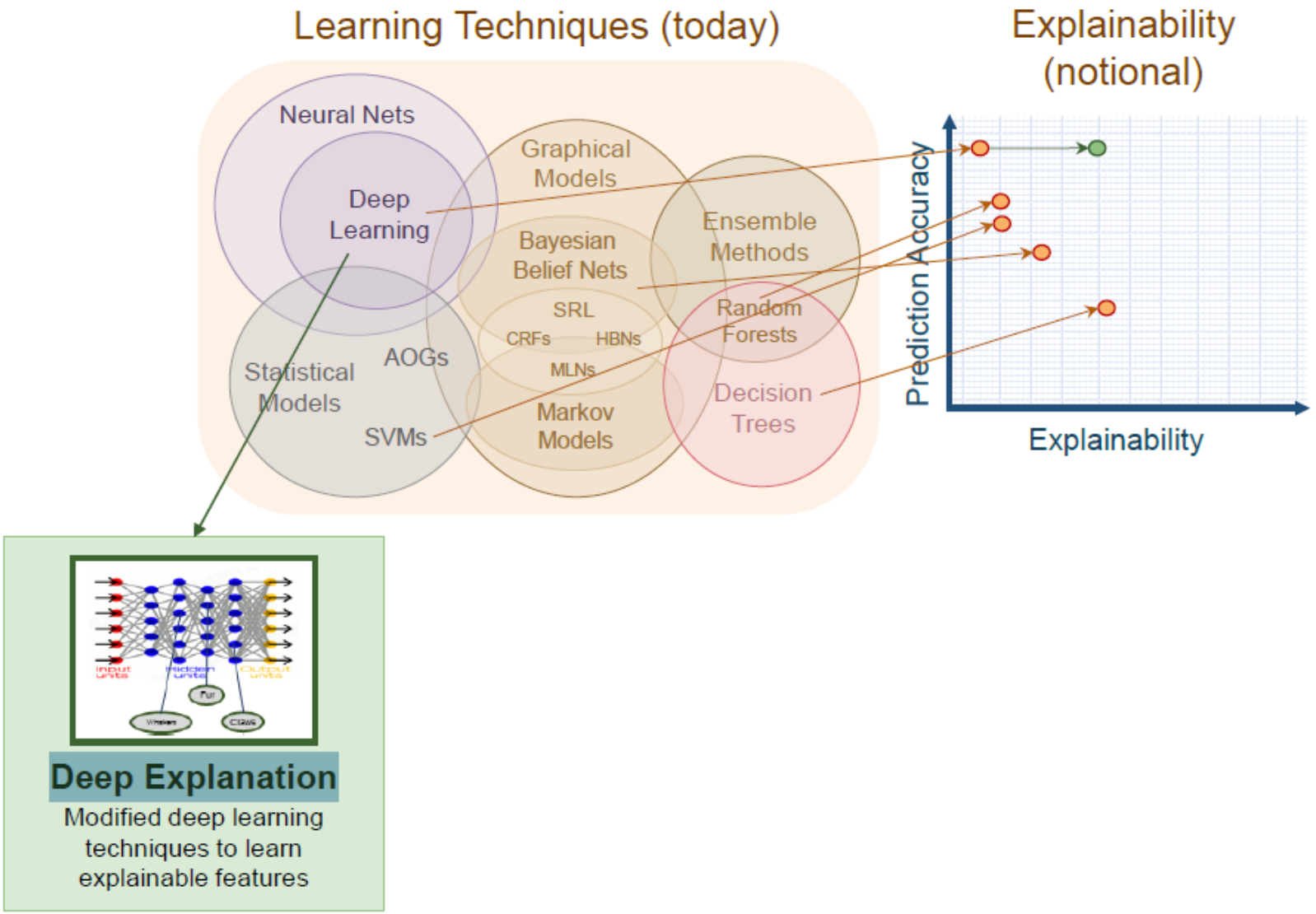
[1]

Wydajność vs wytłumaczalność



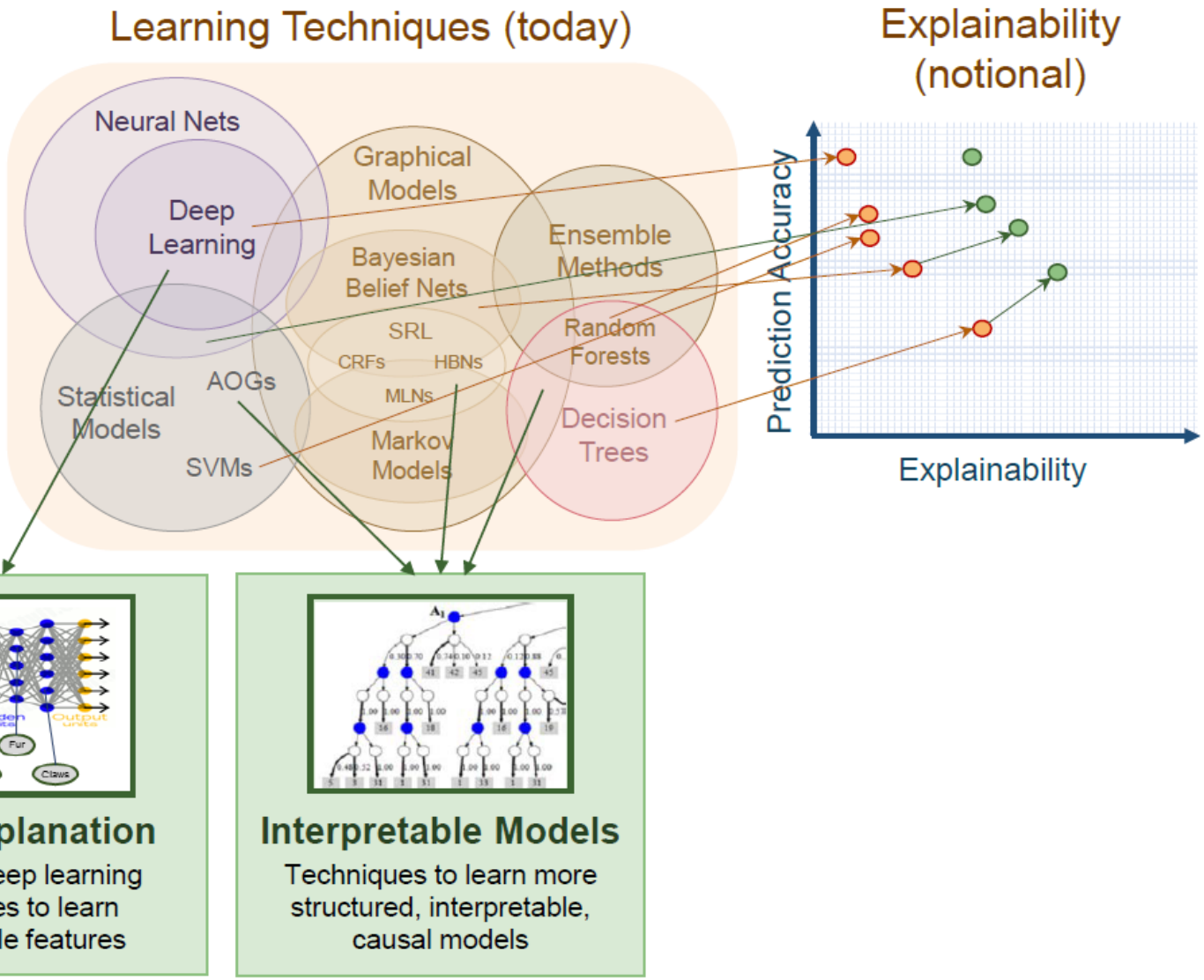
[1]

Wyjaśnianie głębokiego uczenia



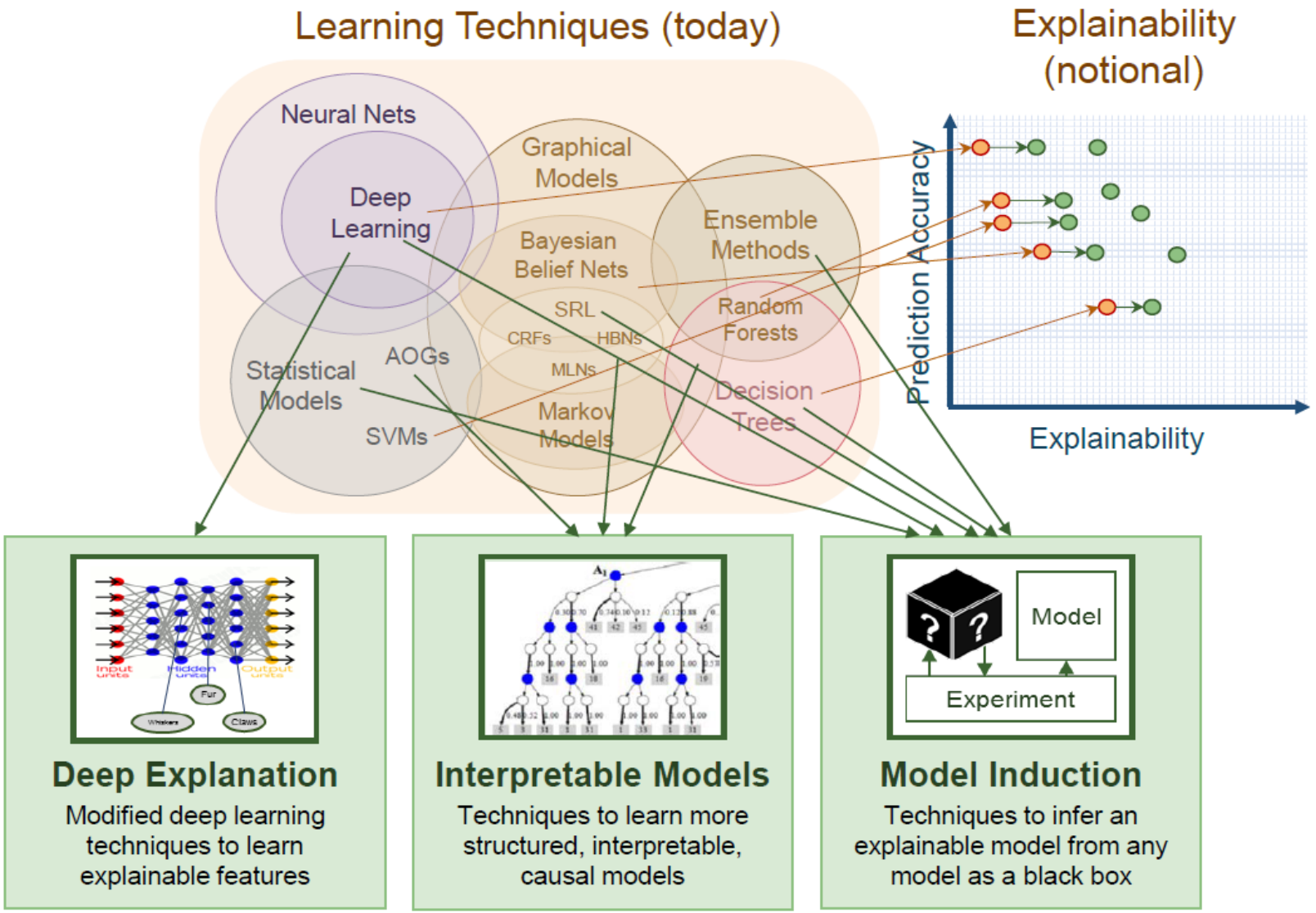
[I]

Interpretowalne modele



[I]

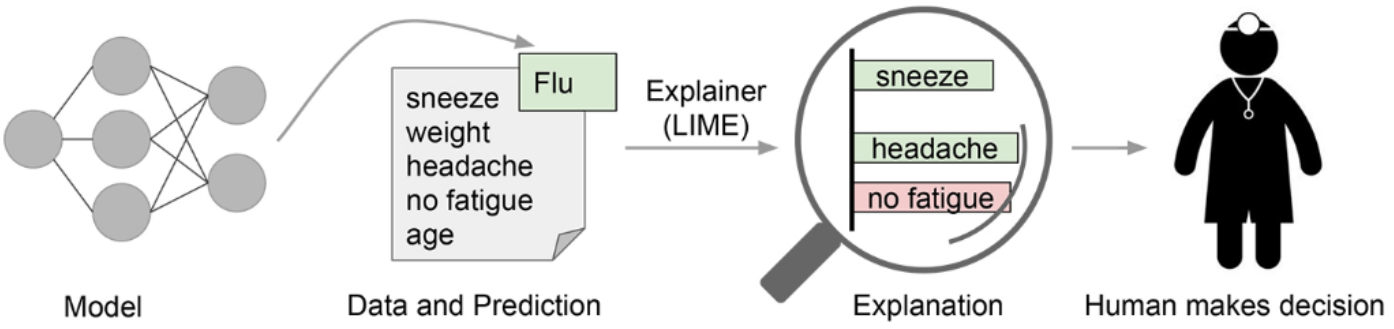
Indukcja modelu



[1]

LIME

- *Local Interpretable Model-Agnostic Explanations* [2]
  - *Local* – używa lokalnie ważonej regresji liniowej
  - *Model-Agnostic* – *bez wiedzy* o modelu (czarna skrzynka)
- Przykład: predykcja choroby
  1. Model przewiduje, że pacjent ma grypę.
  2. LIME zaznacza, które symptomy w historii pacjentów prowadziły do takiej predykcji
  3. Kichanie i ból głowy przyczyniły się do predykcji grypy, brak znużenia jest argumentem przeciw
  4. Ostateczną decyzję podejmuje lekarz



[2]

LIME – zasada działania

- Pojedynczy wejściowy obraz jest modyfikowany na wiele sposobów i sprawdzane jest jak zmieni się predykcja
- Jest to kluczowe w kontekście interpretowalności nowego modelu, ponieważ na wejściu można zmieniać elementy czytelne dla człowieka, np. wyrazy czy fragmenty obrazu

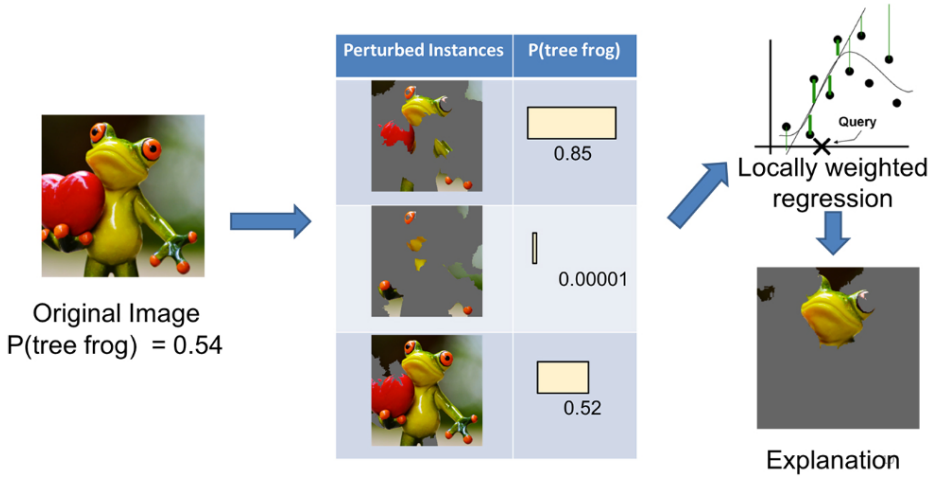
Original Image

Interpretable Components

[4]

LIME – zasada działania

- Ukryty model jest aproksymowany przy pomocy modelu czytelnego który podejmując decyzję bazuje na modyfikacjach oryginalnego wejścia poprzez np. usunięcie pewnych słów lub zakrycie części obrazu
- Model czytelny jest modelem lokalnym
- Na końcu, jako cechy wyjaśniające, brane są te, które mają podobne wartości zarówno dla badanej instancji jak i dla rekordów zbioru treningowego zbliżonych do niej



[4]

Lokalnie ważona regresja liniowa

- Aby wyjaśnić konkretną decyzję wystarczy rozumieć jak model zachowuje się w niewielkim lokalnym otoczeniu
- Analogia do różniczkowania – całościowy kształ krzywej może być bardzo złożony, natomiast obliczenie gradientu w danym punkcie jest często łatwym zadaniem
- Regresja liniowa

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

[3]

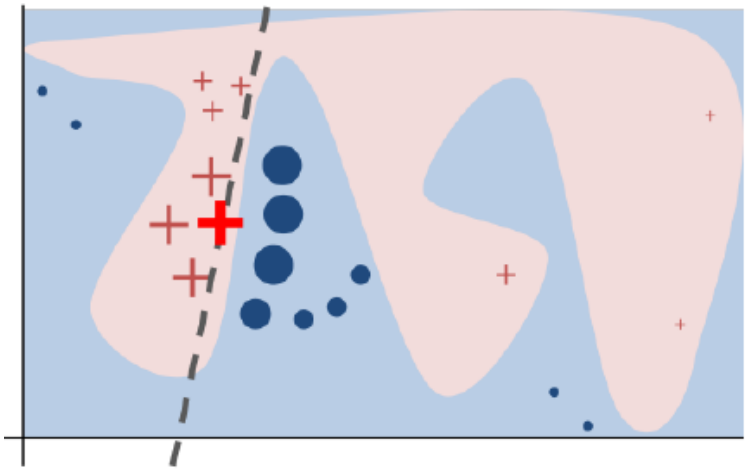
- Lokalnie ważona regresja liniowa

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

[3]

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

Lokalnie ważona regresja liniowa



[4]

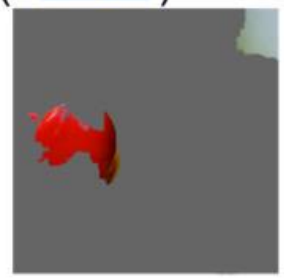
- Złożona funkcja decyzji modelu ukrytego *f* (oznaczona jest na niebiesko/różowo) nie może być ona dobrze aproksymowana przez model liniowy
- Czerwony duży krzyż jest analizowaną instancją
- Próbkowane rekordy są predykowane przy pomocy *f*, a następnie ważone odległością do badanej instancji
- Kreskowana linia jest granicą decyzji oraz definiuje wyjaśnienie, które jest lokalnie (tylko lokalnie) wiarygodne

Lime – działanie

P() = 0.54

P() = 0.07

P() = 0.05

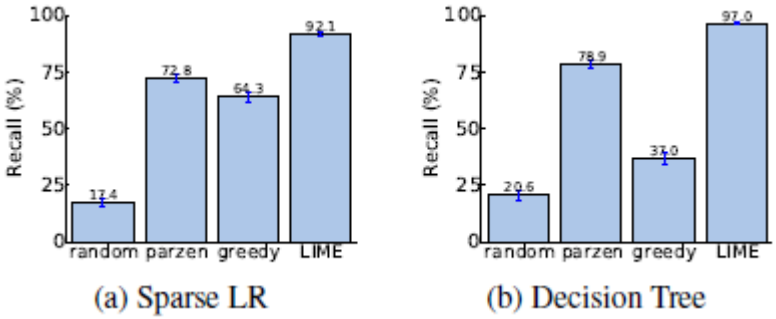


[4]

SP-LIME

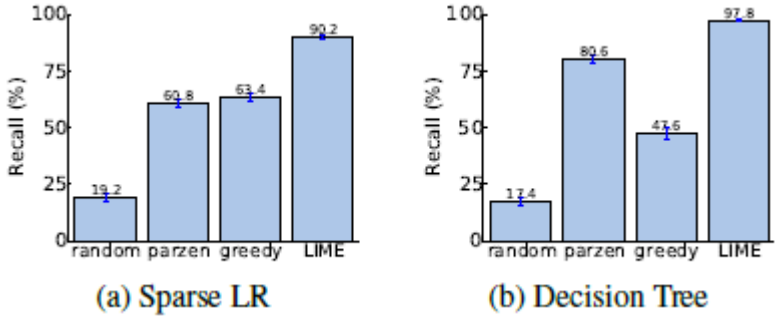
- Celem jest selekcja istotnych cech i stworzenie modelu globalnego
- Analizując zbiór instancji, widzimy które cechy są istotne lokalnie przy predykcji owych instancji
- Cechy, które są lokalnie istotne dla wielu instancji, są też globalnie istotne; pozostałe cechy nie są dalej rozpatrywane
- Następnie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

Rezultaty



(a) Sparse LR (b) Decision Tree

Zbiór książek



(a) Sparse LR (b) Decision Tree

Zbiór płyt DVD

[2]

Rezultaty

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

Uśredniona wartość miary F_1

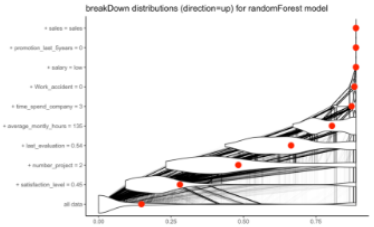
[2]

SmarterPoland.pl

Szukaj …

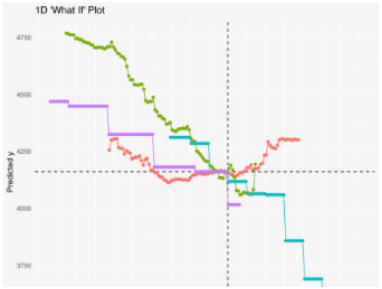
OSTATNIE WPISY

RODO + DALEX, kilka słów o moim referacie na DSS



W przyszły piątek (8 czerwca) na wydziale MiNI PW odbędzie się konferencja Data Science Summit.
W sali 107 pomiędzy 10:50 a 11:20 ma miejsce mój referat **Wyjaśnij! Jak budować wyjaśnialne modele ML / AI i jak to się ma do RODO?**, na który serdecznie zapraszam.

Planuję opowiedzieć o temacie, który wciąga mnie coraz bardziej, czyli wyjaśnialnym AI (XAI). Jak to się ma do RODO i o co chodzi z pogłoskami o „prawie do wyjaśnienia”?



To będzie techniczny referat (sorry, żadnych zdjęć kotów czy psów, być może jakieś zdjęcia robotów). Pokażę jak konstruować i używać wykresy breakDown (i powiem dlaczego są lepsze niż LIME czy wartości Shapleya), będzie też mowa o najnowszym wyniku naszego zespołu, czyli wykresach What-If.

Bibliografia

1. D. Gunning, *Explainable Artificial Intelligence*, DARPA/I2O
2. M. Tulio Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 10.18653/v1/N16-3020
3. D. Boneh, A. Ng, *CS229: Machine Learning. Autumn 2017*, Stanford, Lecture notes, http://cs229.stanford.edu/syllabus.html
4. https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime