

# LIME

Emilia Majerz

15.10.2024

# Local Interpretable Model-agnostic Explanations (LIME)

- Świetnie nadaje się do interpretacji modeli nauczonych na danych z dużą liczbą zmiennych (w przeciwieństwie do np. SHAPa).
- Zaproponowana w artykule [“Why Should I Trust You?” Explaining the Predictions of Any Classifier](#) (2016).
- Lokalna aproksymacja modelu "czarnej skrzynki" przez prostszy, łatwiejszy w interpretacji model.
- Nadaje się zarówno do klasyfikacji, jak i regresji.

# Intuicja

- Prostszy model reprezentowany przez przerywaną linię.
- "Wyjaśnia" on zachowanie modelu czarnej skrzynki naokoło analizowanej próbki.
- Typowo wykorzystywane modele: regresja liniowa, drzewa.

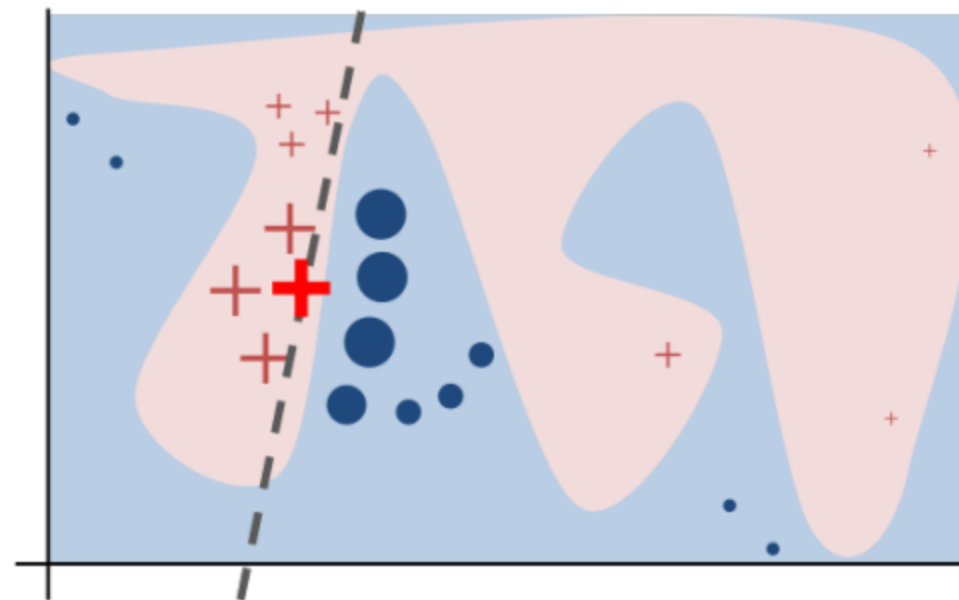


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function  $f$  (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using  $f$ , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

# Reprezentacja danych

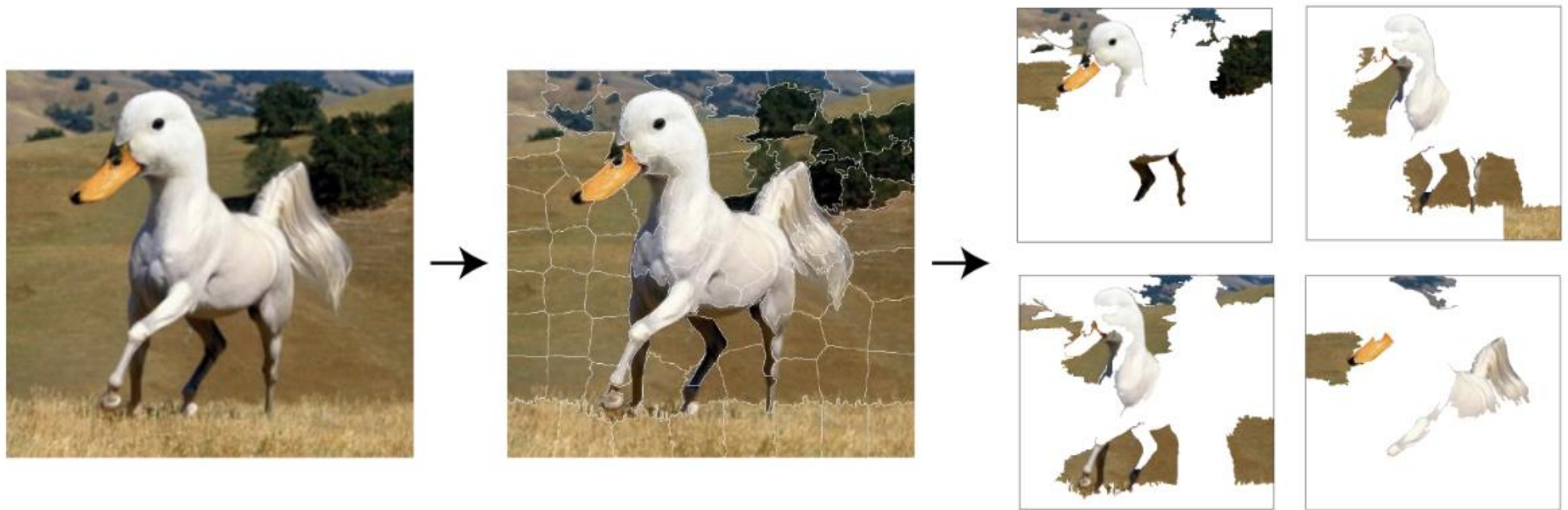
- Interpretowalna reprezentacja.
- Ważna szczególnie dla danych wielowymiarowych, jak obrazy.
- Piksele -> superpiksele (włączane bądź wyłączane).



# Generowanie danych dla potrzeb interpretacji

- Naokoło analizowanego przykładu.
- Nowe dane, a nie z oryginalnego zbioru.
  - W oryginalnym zbiorze może ich być za mało.
- Nowe dane: analizowany przykład poddawany perturbacjom.
  - Dane binarne: zamiana 0 na 1 i na odwrót.
  - Dane ciągłe: różne podejścia, np. dokładanie szumu, kwantyzacja i perturbacja tak uzyskanych danych binarnych.
  - Obrazy: włączanie/wyłączanie superpikseli.

# Generowanie danych – obrazy – przykład



# Trening lokalnego modelu

- Trening prostego (zwykle liniowego) modelu na wygenerowanych próbkach.
- Po prawej: 15 najważniejszych superpikseli dla predykcji pudła i gęsi.

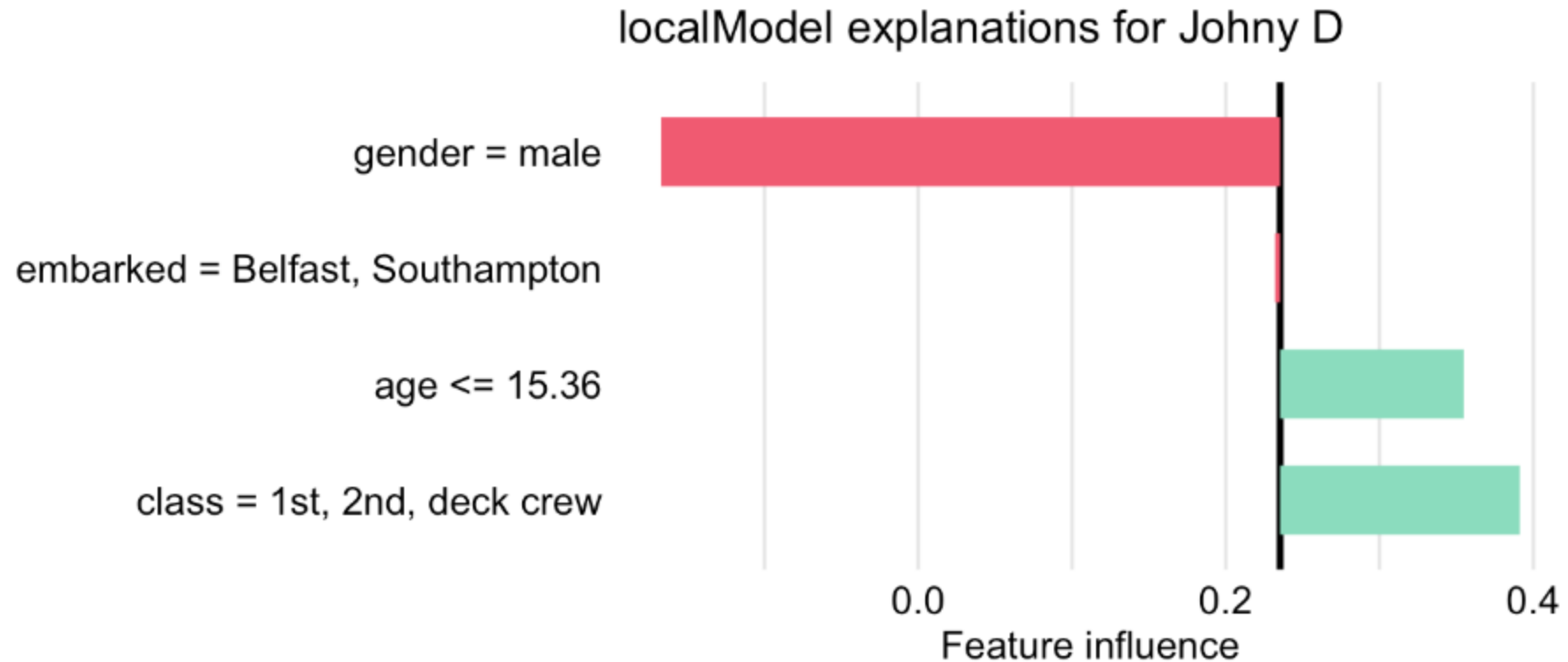
Label: standard poodle  
Probability: 0.18  
Explanation Fit: 0.37



Label: goose  
Probability: 0.15  
Explanation Fit: 0.55



# Przykład - Titanic



Łatwiejsze w interpretacji dyskretne cechy.



# Materialy

- <https://ema.drwhy.ai/LIME.html>