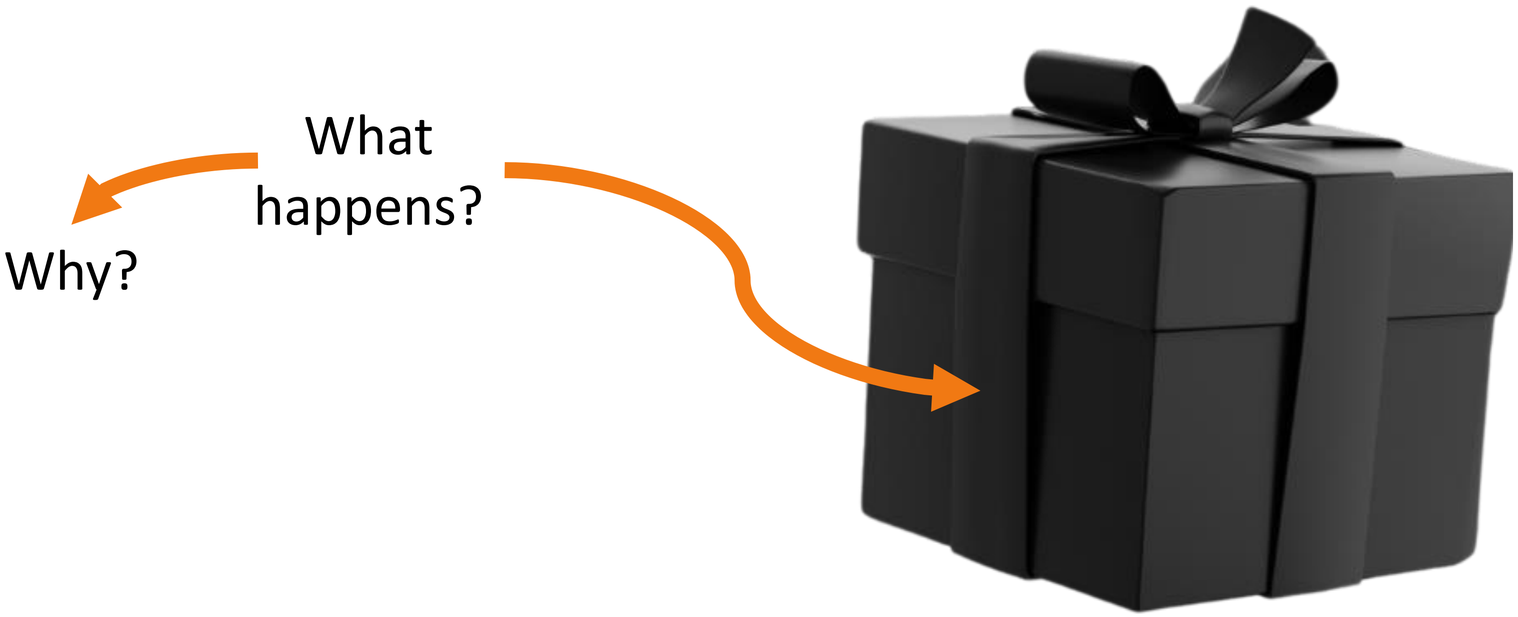
LIME

Intuition

LIME - intuition

We have a black box model with complex boundaries, and we want to understand how it makes some predictions.

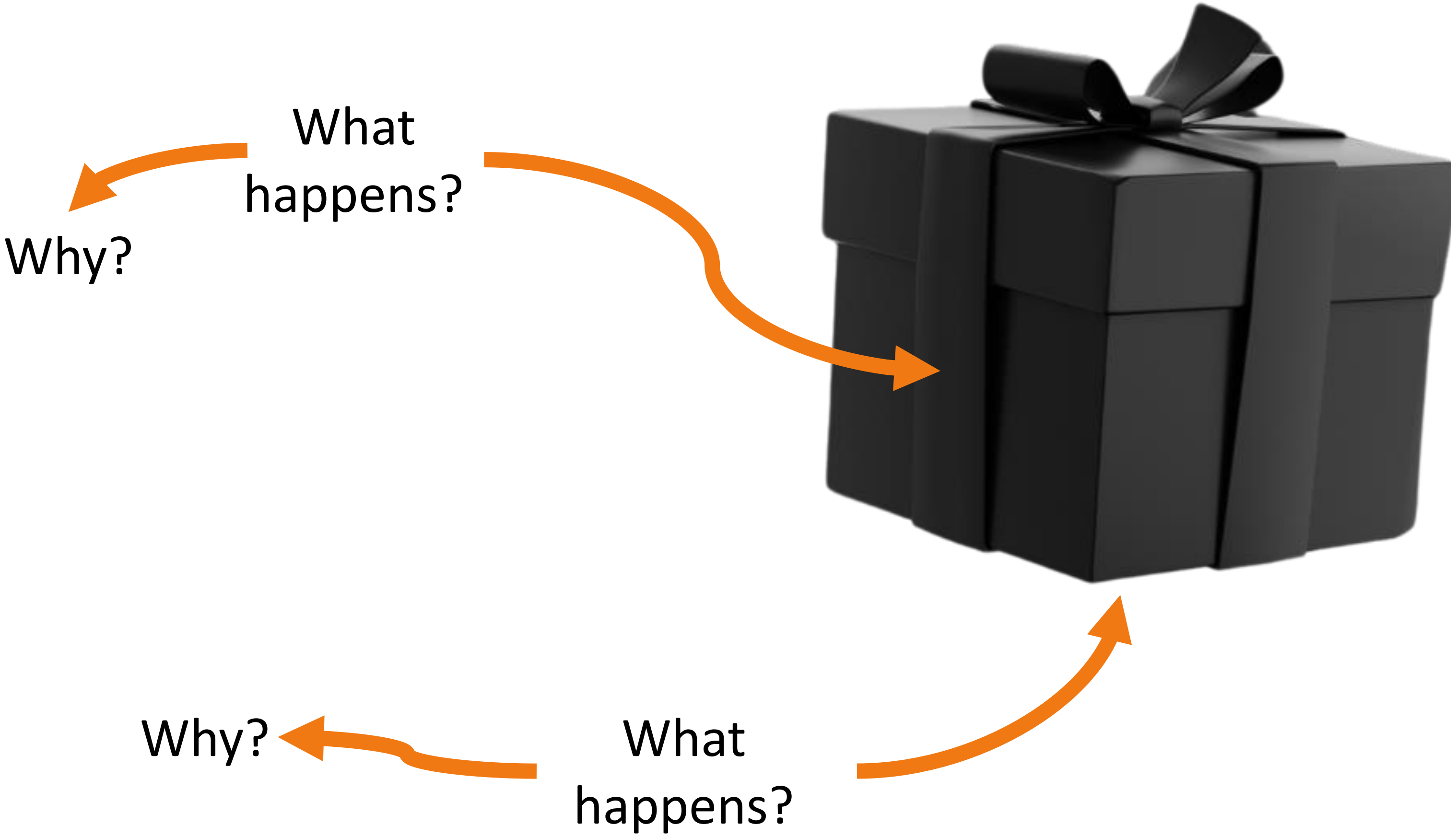
LIME - intuition



What happens?

Why?

LIME - intuition



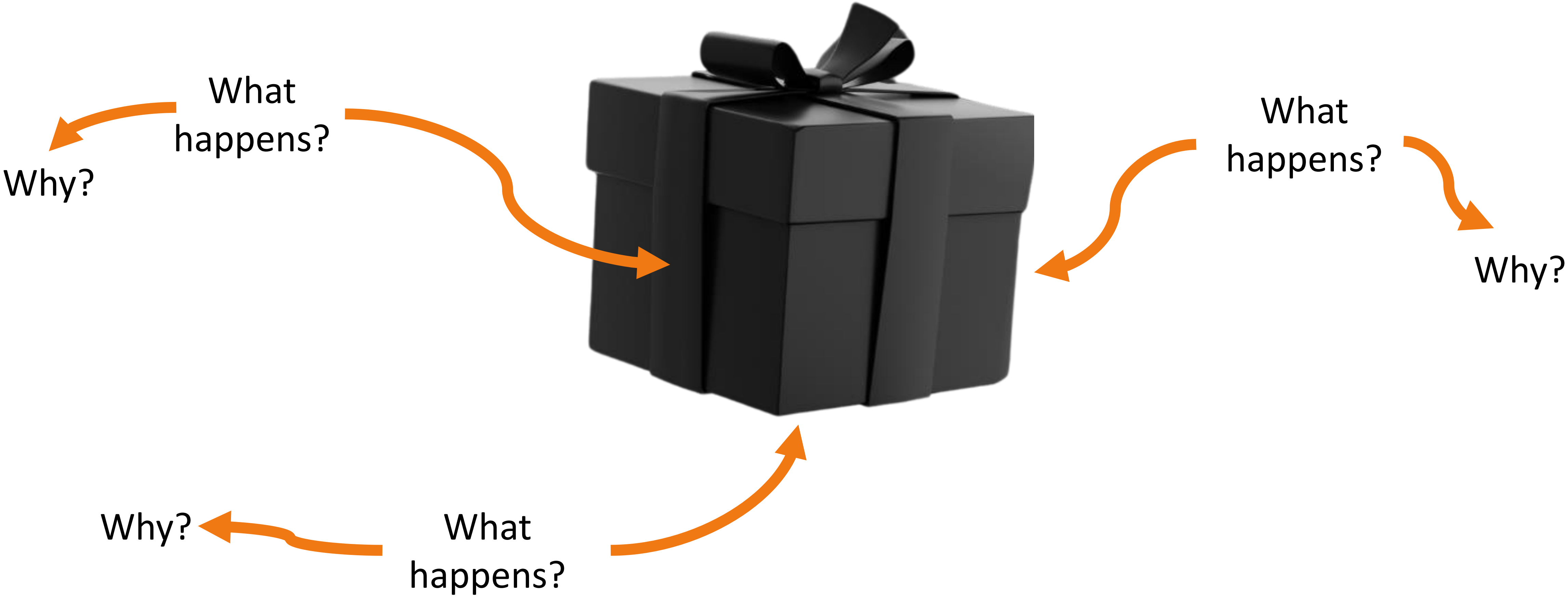
What happens?

Why?

What happens?

Why?

LIME - intuition



What happens?

Why?

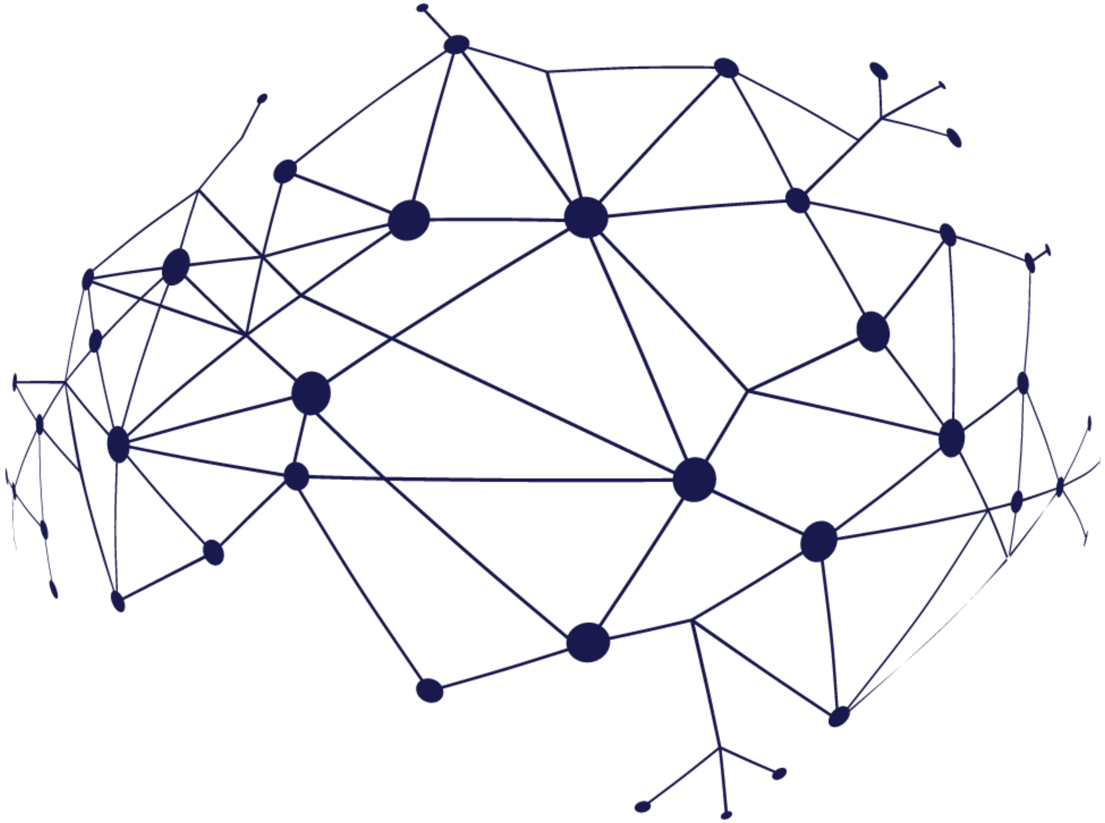
What happens?

Why?

What happens?

Why?

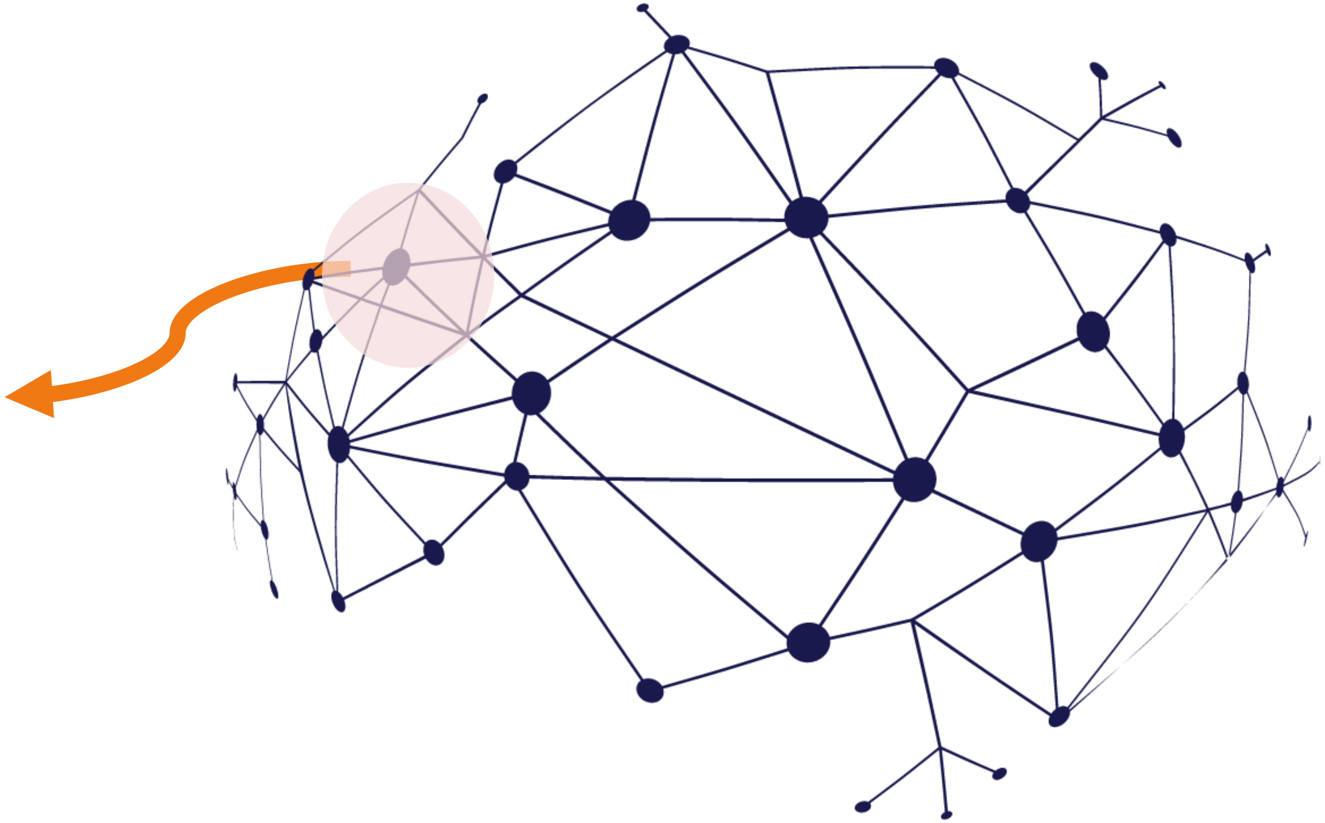
LIME - intuition



Black box

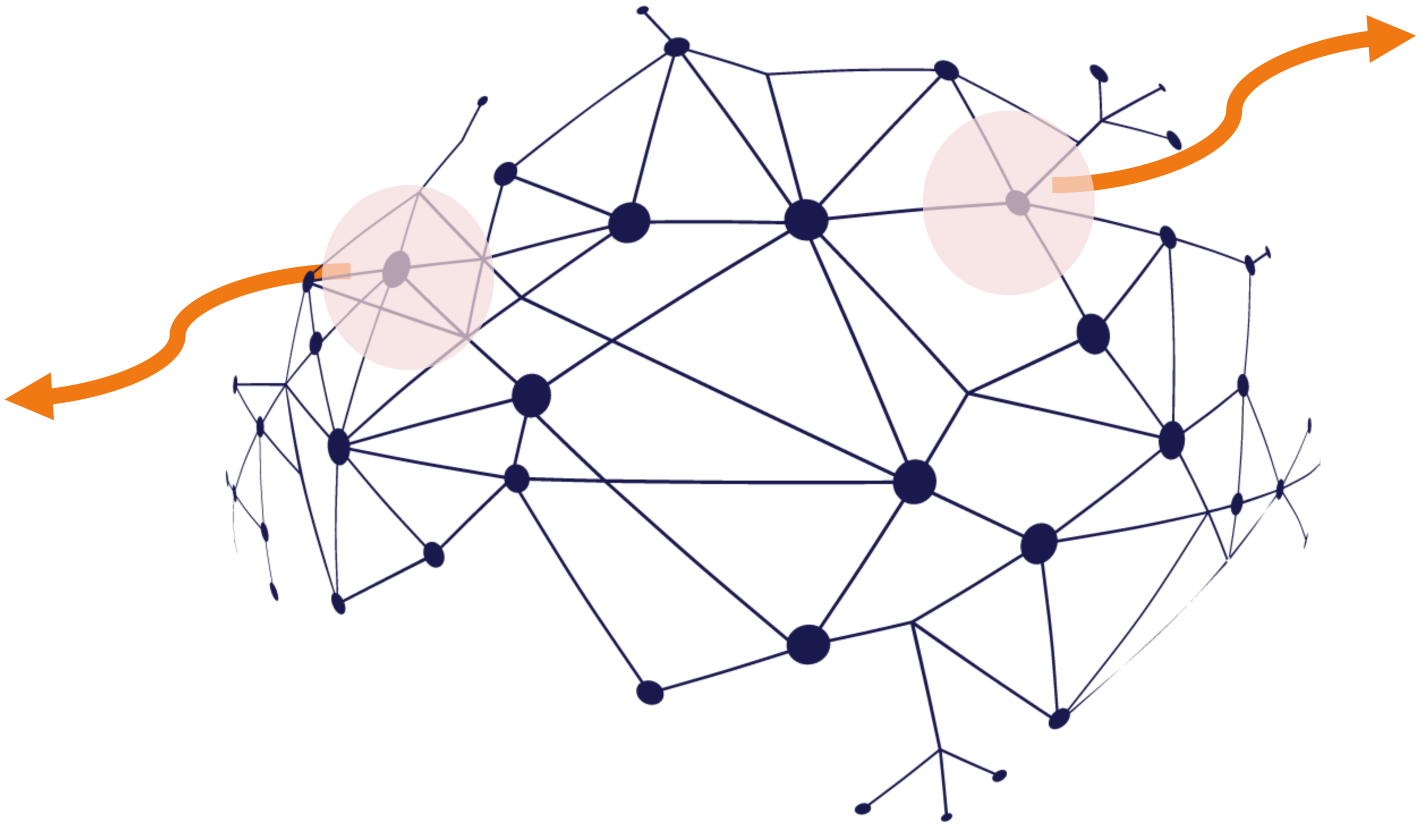
LIME - intuition

Why does it make this prediction?



Black box

LIME - intuition

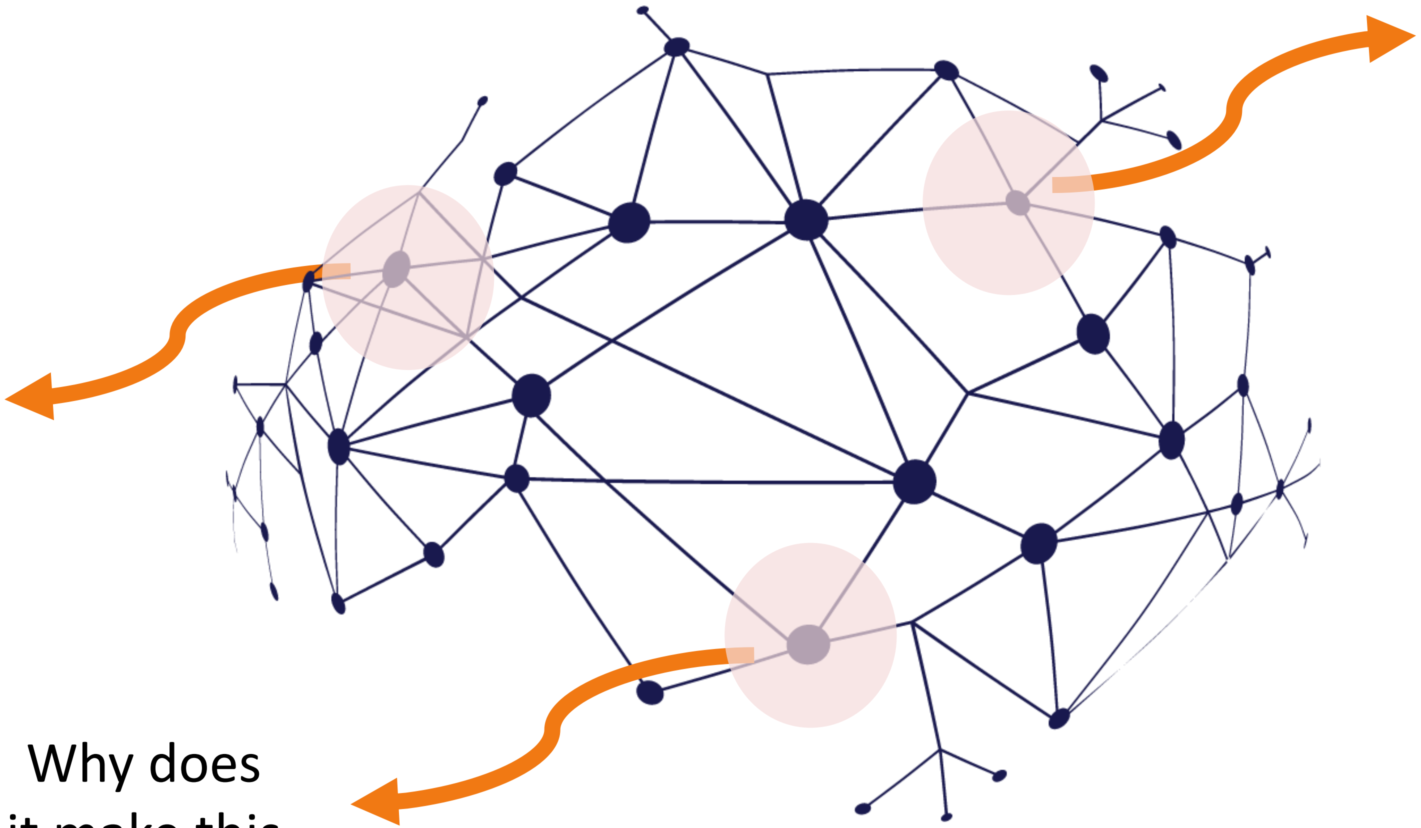


Why does it make this prediction?

Why does it make this prediction?

Black box

LIME - intuition



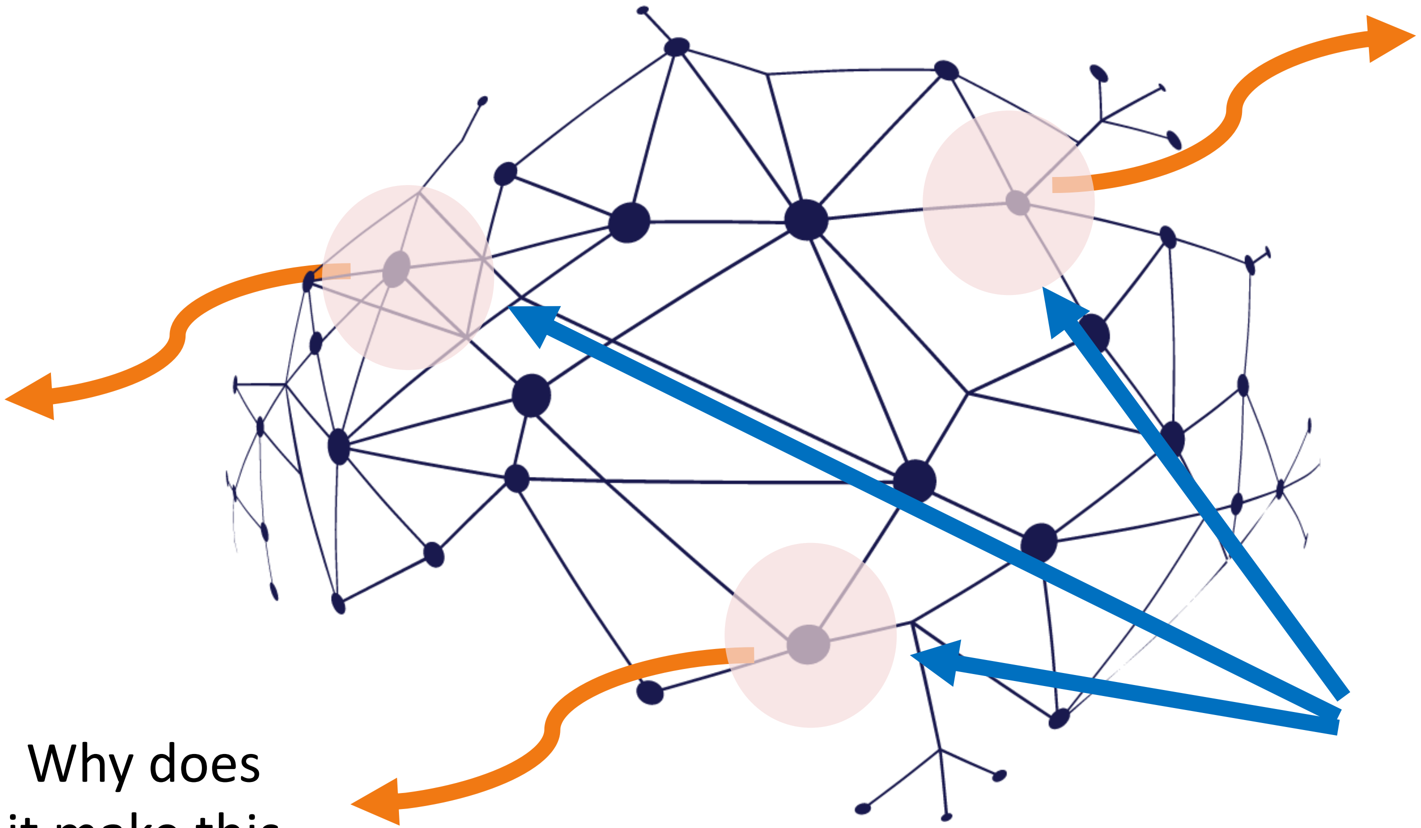
Why does it make this prediction?

Why does it make this prediction?

Why does it make this prediction?

Black box

LIME - intuition



Why does it make this prediction?

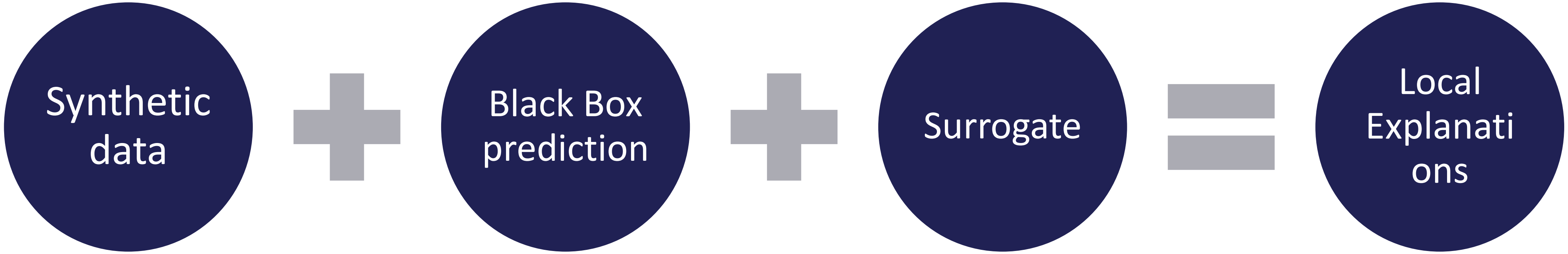
Why does it make this prediction?

Why does it make this prediction?

Local surrogates

Black box

LIME

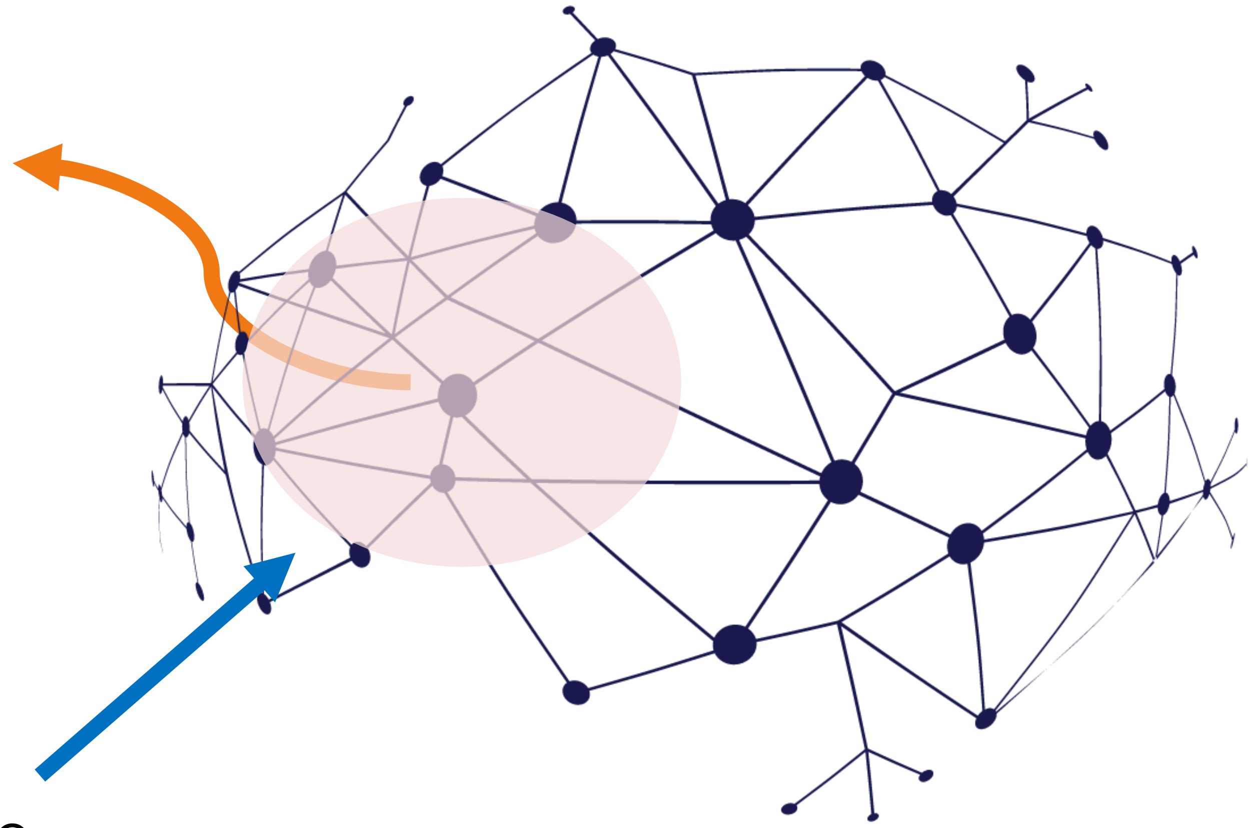


Synthetic data + Black Box prediction + Surrogate = Local Explanations

1. Creates synthetic data in the proximity of the observation
2. Obtains black box predictions for that data
3. Trains a surrogate
4. Explains the surrogate

Locality

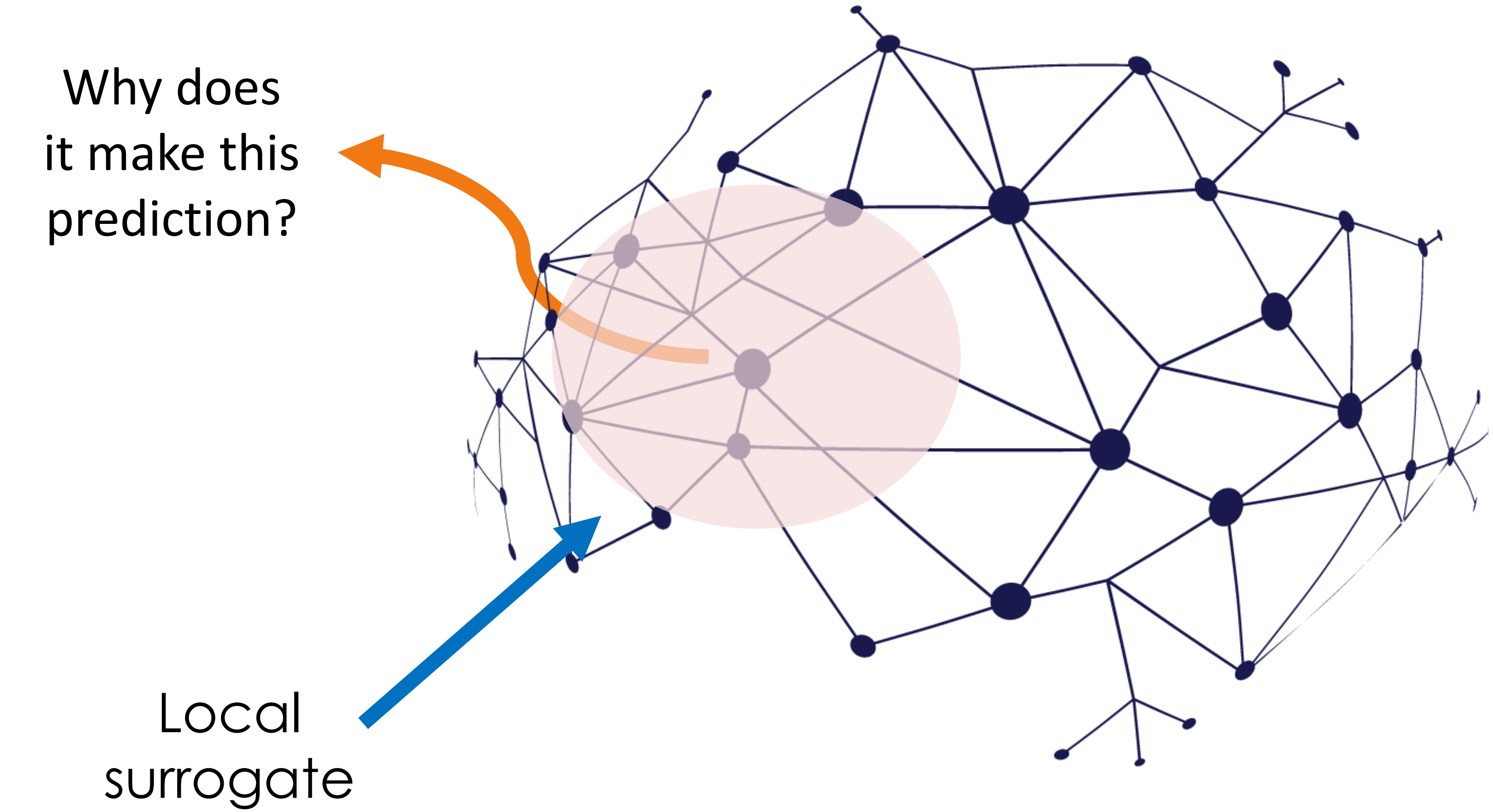
Why does it make this prediction?



Local surrogate

The further away from the data point to explain, the less accurate the local model may be.

Locality



The further away from the data point to explain, the less accurate the local model may be.

➔ Weight contributions based on distance to the data point.

LIME - mechanism

1. Choose the data point to explain.
2. Generate synthetic data in its proximity.
3. Obtain the black box predictions for the data from 2.
4. Obtain the distance between synthetic data and original data point.
5. Train a white box with the perturbed data (2) to predict the black box predictions (3), weighted by their locality (4).
6. Interpret the white box.

LIME - mechanism

1. How do we generate the synthetic data?
2. Which explainable model should we use?
3. How do we calculate the distances?

Train In Data

THANK YOU

www.trainindata.com