ALMAWAVE

L'interpretabilità dei modelli di Machine Learning come parametro di qualità: L'esperienza Datalake Giustizia

R. Bizzoni, **M. Borriello**, A. Favalli, C. Giannone,

D. Preti, R. Romagnoli, F. Wolenski

Ital-IA 2023

Use Case - Datalake Giustizia

Datalake

Analisi del patrimonio informativo del MdG

Dati

Estrazione del dataset di addestramento

Giustizia Predittiva

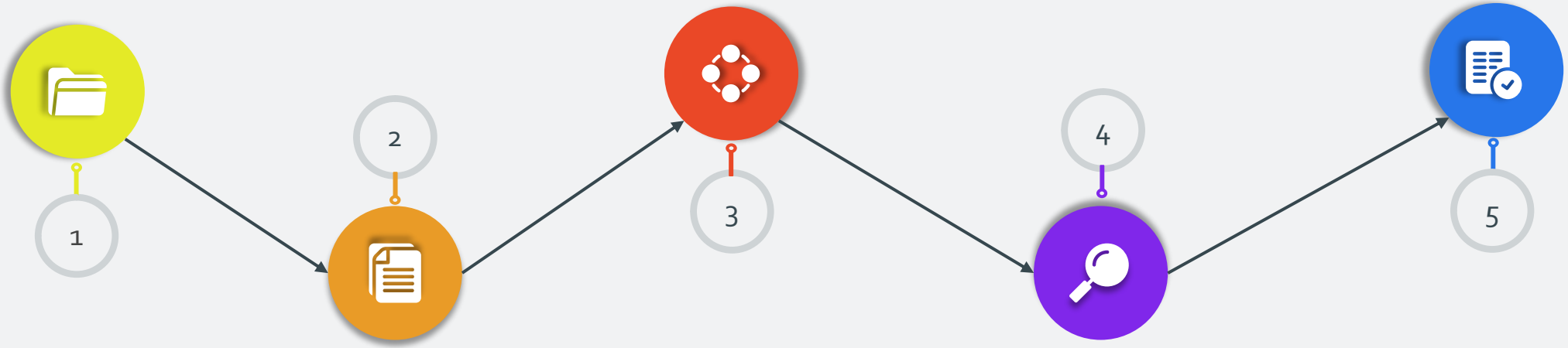
Addestramento del modello di ML-DL

Qualità

Analisi della qualità delle tecniche di IA

Mitigazione

Trade-off tra *performance* e *qualità*



ALMAWAVE

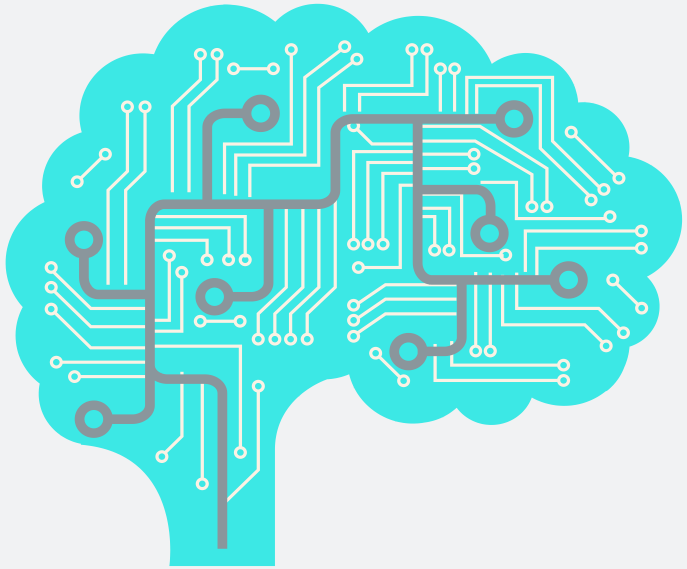
XAI - Explainable AI

Spiegabilità

Capacità di spiegare dal punto di vista umano una meccanica interna al modello; creare modelli **trasparenti**

Interpretabilità

Capacità di comprendere le relazioni causa-effetto di modelli **parametrici** complessi



Metodi *Gradient*-based

I metodi basati su **gradienti** consistono nell'interpretare la rete neurale tramite i gradienti dei suoi pesi

Metodi *Perturbation*-based

I metodi **perturbation-based** si basano su perturbazioni degli esempi sui quali si vuole ottenere una *explanation*

LIME - Local Interpretable Model-agnostic Explanations

- Algoritmo basato su **perturbazioni** delle predizioni del modello per interpretare i modelli «black box»
- Le ***explanations*** fornite da LIME si basano su un'approssimazione lineare del modello originale attorno ad una predizione
- Nel caso di entità multi-token nella nostra implementazione le ***explanations*** non sono direttamente legate alle entità, ma alla loro annotazione **BIO**

```
Nigeria B-Loc
'S O
President B-Per
Olusegun I-Per
Obasanjo I-Per
expressed O
his O
condolences O
, O
nothing O
```

formato CONLL 2002

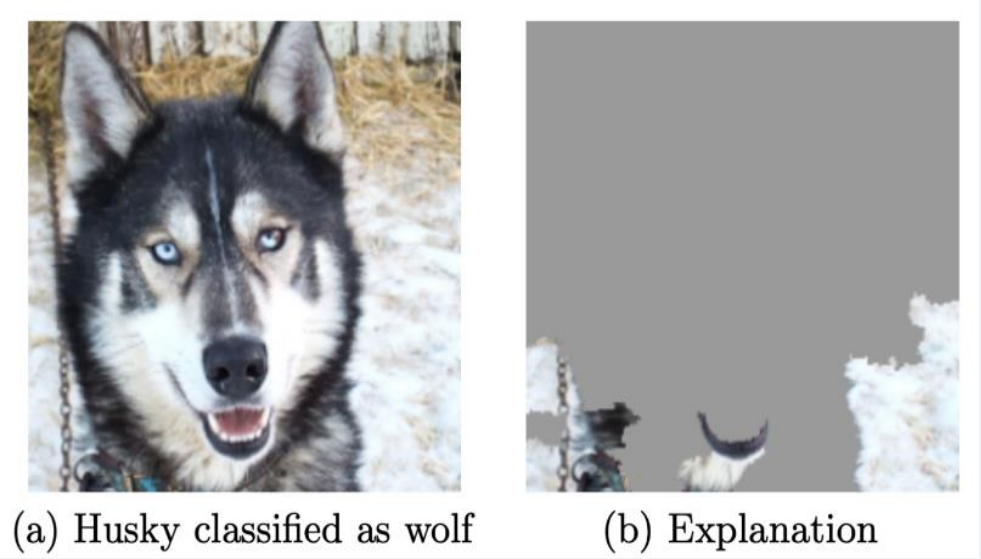
NER - LIME

L'**explainer** LIME applicato a un modello di **NER** basato su *Transformer* (**Bert**):

- ❑ *Word-pieces* come tokens
- ❑ Rappresentazione contesuale
- ❑ Perturbazioni applicate anche ai token che fanno parte delle *entities*

Perché interpretare i modelli?

- Velocizzare il processo di *evaluation*
- Istaurare un rapporto di fiducia con il modello
- Avere coscienza dell'impatto del ML sulla vita degli individui in ambiti come la giustizia predittiva



(a) Husky classified as wolf (b) Explanation

ALMAWAVE

Grazie per l’attenzione