Model-Agnostic Counterfactual Explanations of Recommendations

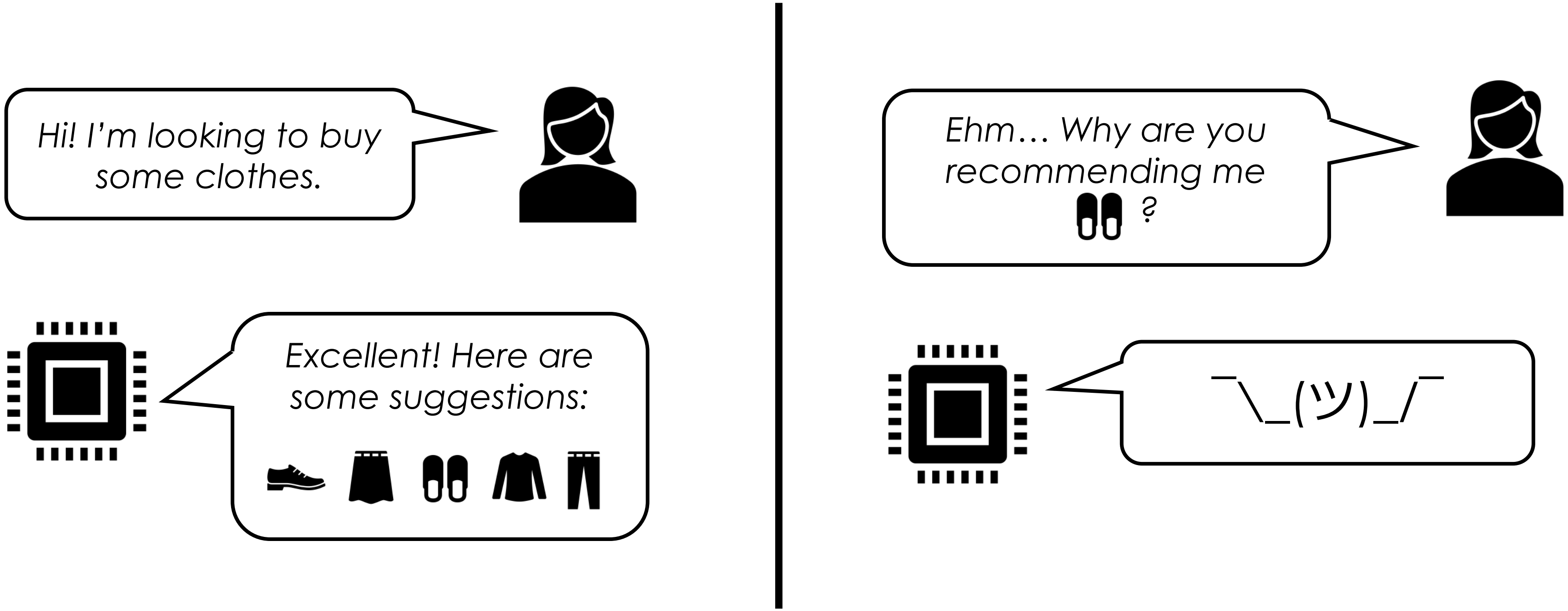
Vassilis Kaffes[1], **Dimitris Sacharidis**[2], Giorgos Giannopoulos[1]

ACM UMAP 2021

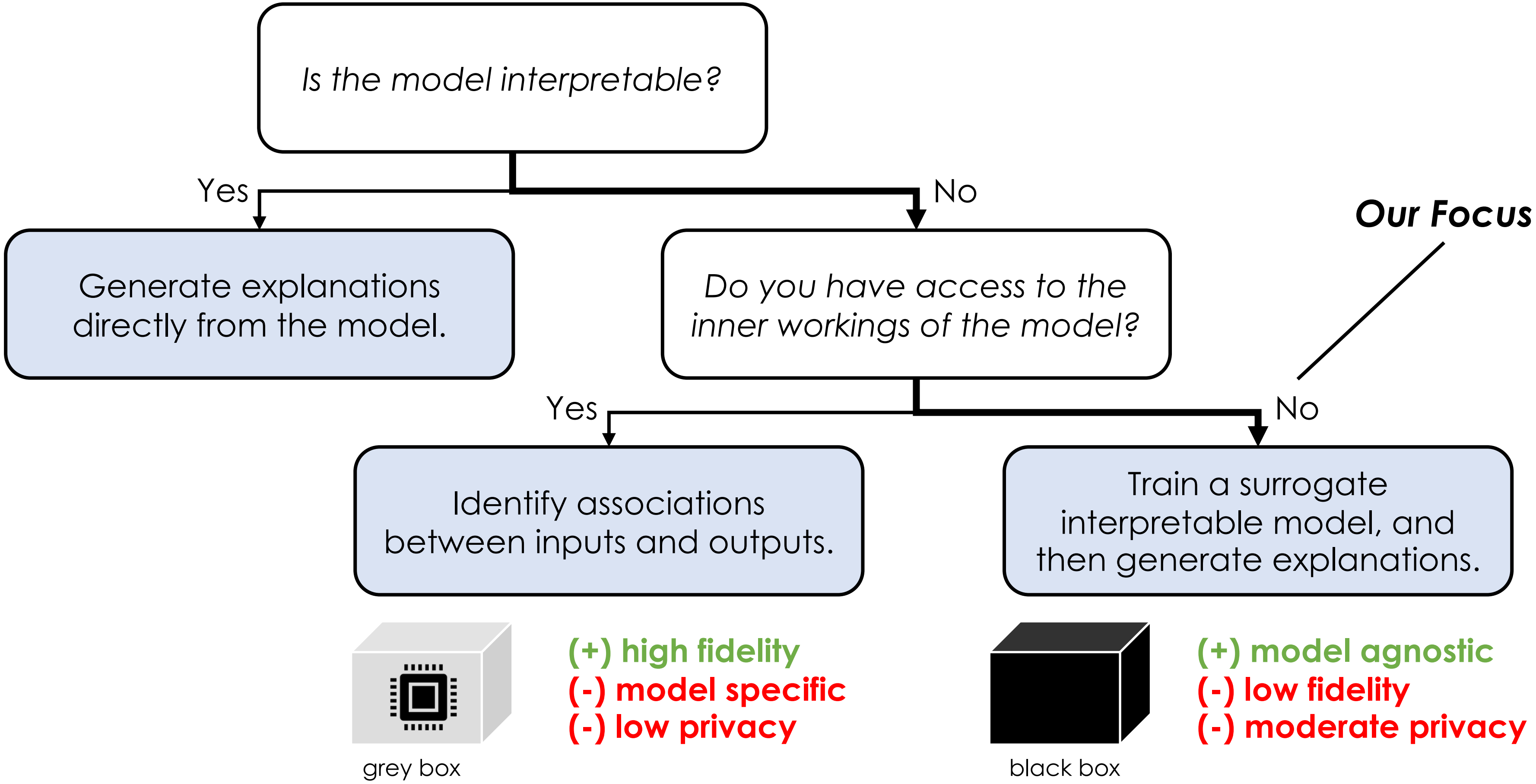
(1) ATHENA Research & Innovation Information Technologies

(2) ULB UNIVERSITÉ LIBRE DE BRUXELLES

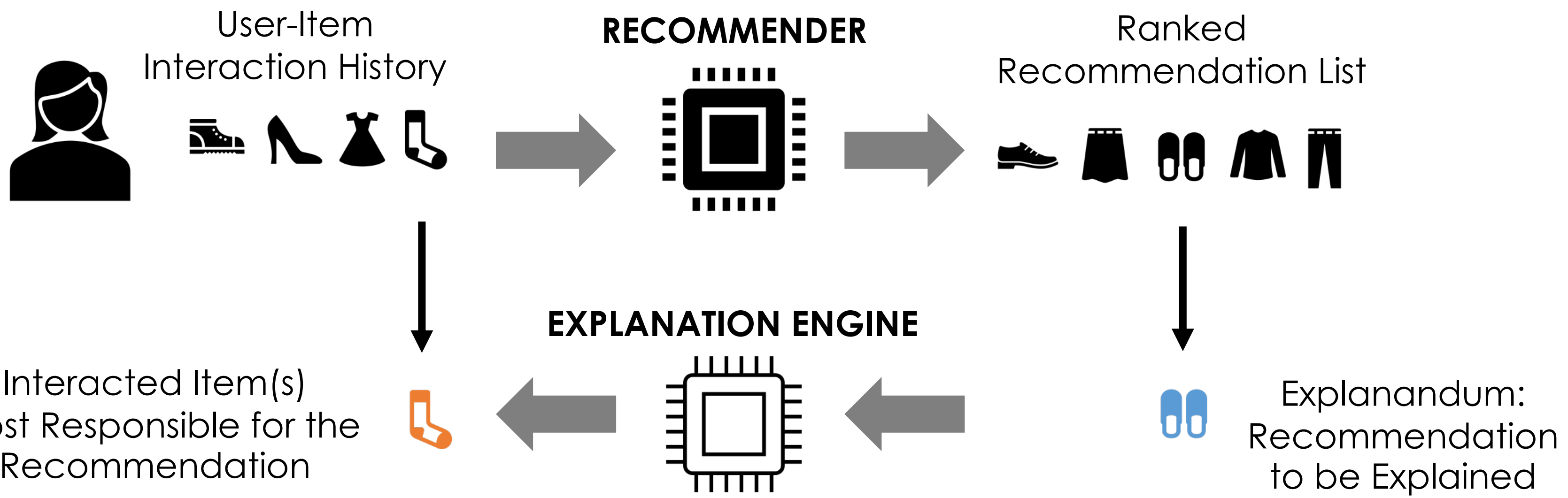
Explanations of Recommendations



How to Explain Recommendations



Counterfactual Explanations



COUNTERFACTUAL EXPLANATION

"Had you not interacted with [sock], you would not be recommended [slippers]."

(+) model agnostic
(+) high fidelity
(+) high privacy

Counterfactual Explanations

EXPLANANDUM

FACTUAL

COUNTERFACTUAL

EXPLANATION

When is an **explanation** E good?

- When its **counterfactual** is similar to the **factual**, i.e., the **explanation** consists of few interacted items.
  - measured by the **normalized length** $l(E) = {|E|}/{|I|}$.
    - I is the interaction history, i.e., the **factual**
- When it causes the recommender to rank the **explanandum** low.
  - measured by the **impotence** $i(E) = \max\left\{0, \frac{m - rank(t;E) + 1}{m}\right\}$.
    - m is the desired low rank; $rank(t;E)$ is the rank of the **explanandum** t given E.
    - an explanation is called **valid** when it has zero impotence (moves t beyond rank m)

PROBLEM DEFINITION

Find an explanation with **low normalized length** and **low impotence**.

Finding Counterfactual Explanations

The search space of possible explanations, is the **powerset** of the **interaction** history. Very expensive to explore exhaustively.

We propose three efficient search strategies:

Breadth First Search (BFS): greedily looks for a valid explanation, and then tries to improve on its length.

Priority Search (Pri): drives the search using a priority queue; each explanation is given a score (a convex combination of $l(E)$ and $i(E)$); upon dequeuing E, its neighborhood is examined and enqueued.

Hybrid Search (Hyb): it first exhaustively examines all short explanations (of length two), and then switches to priority search.

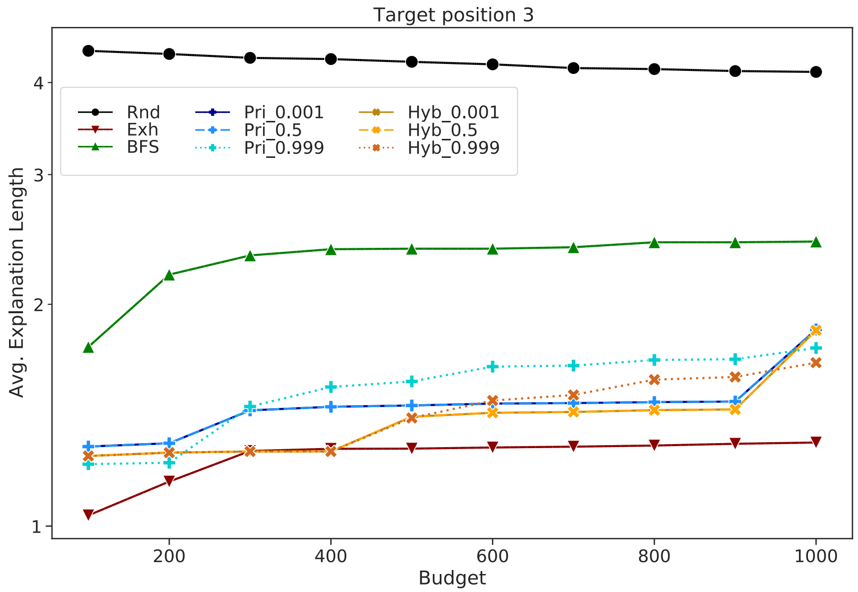
Evaluation of Strategies

Evaluation Protocol

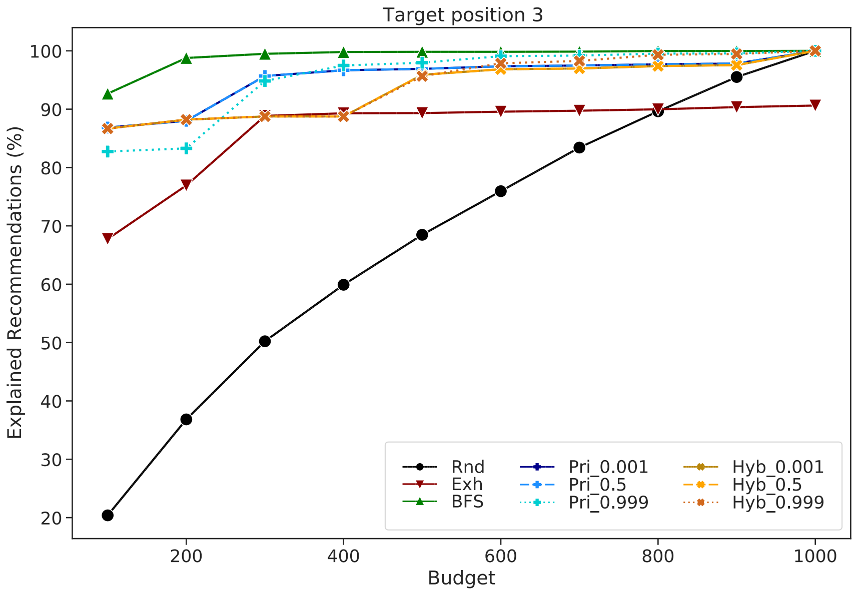
1. Train a session-based recommender with MovieLens 100K.
2. Repeat for 100 users selected at random.
3. Feed the user's interaction history, and request **top-20 recommendations**.
4. Select the **3rd ranked item** as the **explanandum**.
5. Given a **budget** (number of recommendation requests), search for a counterfactual that moves the explanandum **beyond rank 20**.

Evaluation of Strategies

(lower is better)



(higher is better)



Exhaustive search (Exh) identifies short explanations but in less than 90% of the cases.

Random search (Rnd) identifies long explanations in all cases.

Our strategies identify short explanations in all cases and are highly budget conscious.

Hybrid search (Hyb) exhibits the **best trade-off** between **speed** (budget spent) and **quality** (explanations' length, % of explanations given).

Thank you!

Model-Agnostic Counterfactual Explanations of Recommendations

Vassilis Kaffes[1], **Dimitris Sacharidis**[2], Giorgos Giannopoulos[1]

(1) ATHENA Research & Innovation Information Technologies

(2) ULB UNIVERSITÉ LIBRE DE BRUXELLES