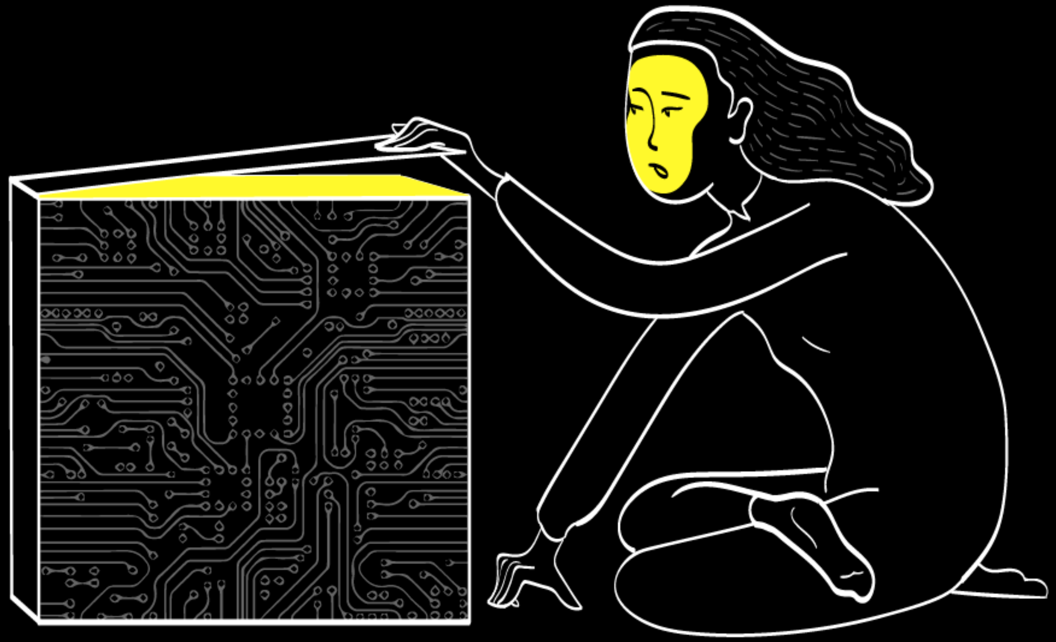
An overview of interpretability of Recommender Systems



Agenda

- Explainability
  - Why do we need to explain?
  - Notions of explainability
  - Meaningful explanations
- Explainability in Recommender Systems
  - Paradigms
  - Information Category
  - Model
  - Evaluation
- Summary

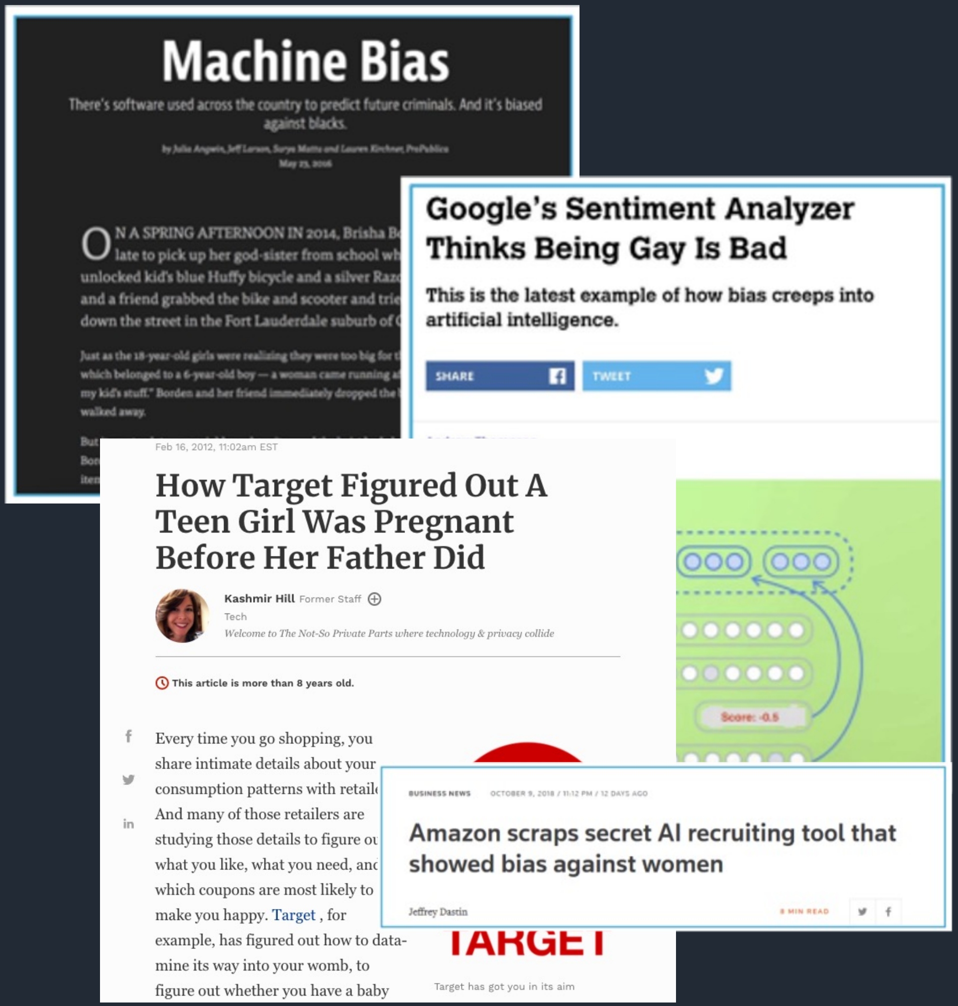
Why Explainability?

AI is now used in many high-stakes decision-making applications (credit, employment, admission, sentencing).

89% of consumers say…

“technology companies need to be more transparent.”

Technology companies need to comply with GDPR – “Right to Explanation.”



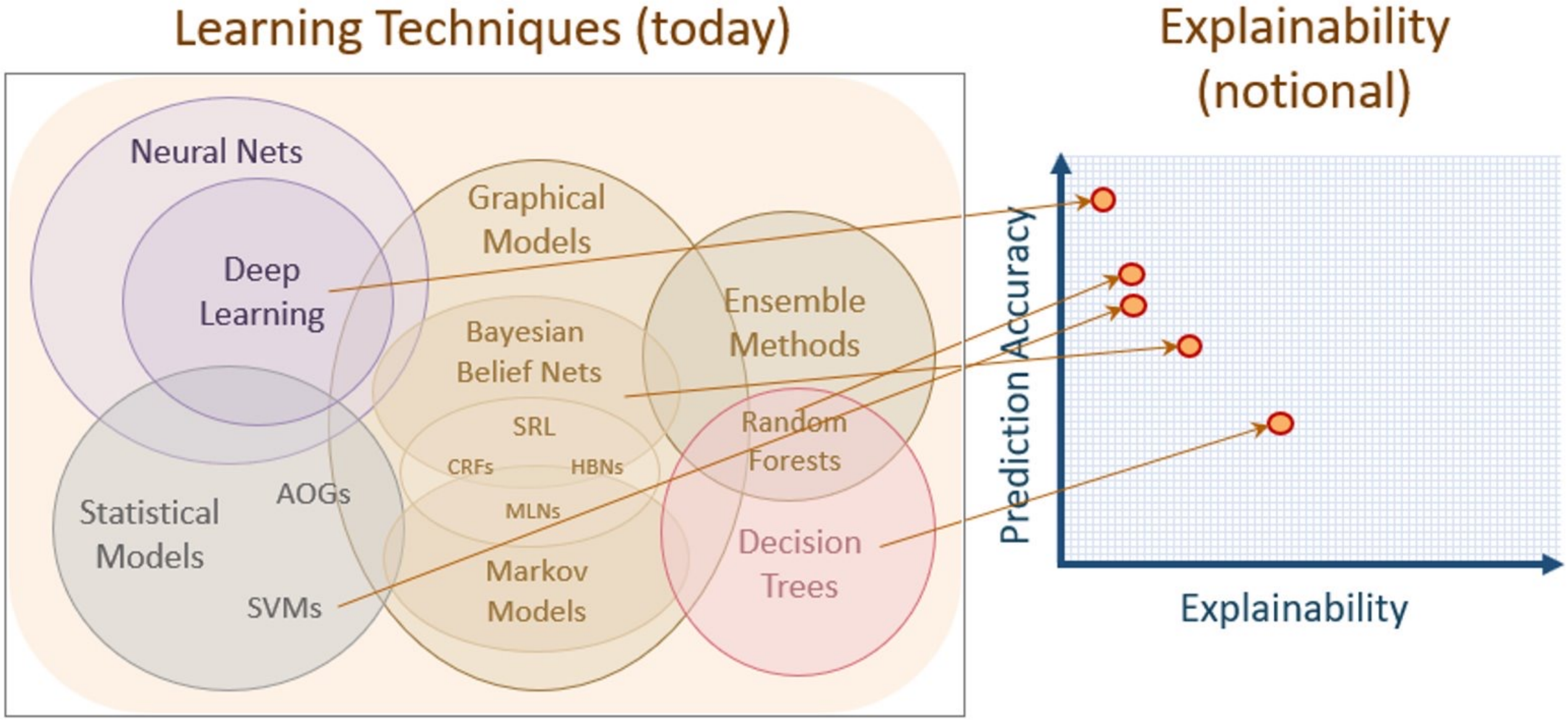
Explainability across industries

ADVERTISING
The reputational risk of placing adverts alongside content that doesn't 'fit' the brand

HEALTH
49% of physicians in the US are anxious or uncomfortable with AI

FINANCE
Explaining why automated decision-making rejects loan applications

The most effective algorithms are the hardest to explain



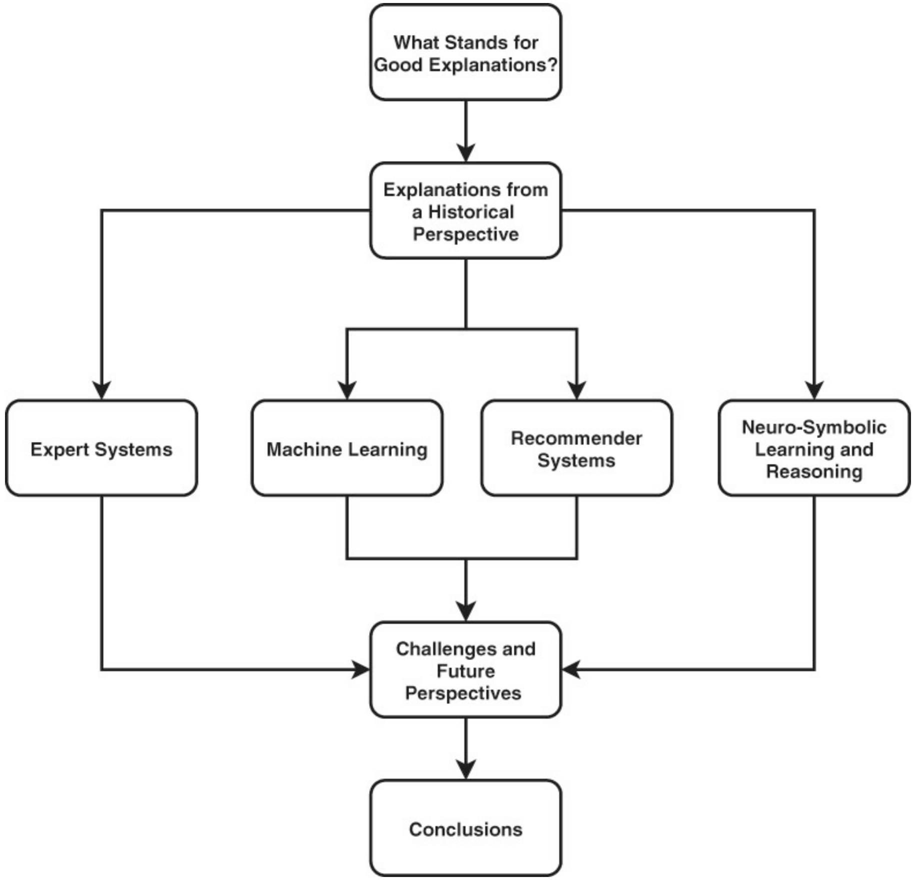
DARPA. Explainable AI - program, 2016. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Why Explainability is a challenge?

Different notions

Different requirements

Plethora of approaches



Confalonieri, R., Coba, L., Wagner, B., and Besold, T.. A historical perspective of explainable artificial intelligence. WIREs Data Mining and Knowledge Discovery, 11(1), 2021. doi: https://doi.org/10.1002/widm.1391

Explainability - Notions

Interpretable Systems

A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs.

E.g. regression models, support vector machines, decision trees, ANOVAs, and data clustering (assuming a kernel that is itself interpretable)

Comprehensible Systemes

A system that emits symbols along with its output that allow the user to relate properties of the inputs to their output. The user is responsible for compiling and comprehending the symbols, relying on her own implicit form of knowledge and reasoning about them.

E.g. High dimensional data visualizations like t-SNE and receptive field visualization on convolutional neural networks

Opaque systems

A system where the mechanisms mapping inputs to outputs are invisible to the user

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. 2071).

Explainability - Notions

Interpretable Systems

A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs.

E.g. regression models, support vector machines, decision trees, ANOVAs, and data clustering (assuming a kernel that is itself interpretable)

Comprehensible Systems

A system that emits symbols along with its output that allow the user to relate properties of the inputs to their output. The user is responsible for compiling and comprehending the symbols, relying on her own implicit form of knowledge and reasoning about them.

E.g. High dimensional data visualizations like t-SNE and receptive field visualization on convolutional neural networks

Opaque systems

A system where the mechanisms mapping inputs to outputs are invisible to the user

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. 2071).

Explainability - Notions

Interpretable Systems

- A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs.
- E.g. regression models, support vector machines, decision trees, ANOVAs, and data clustering (assuming a kernel that is itself interpretable)

Comprehensible Systemes

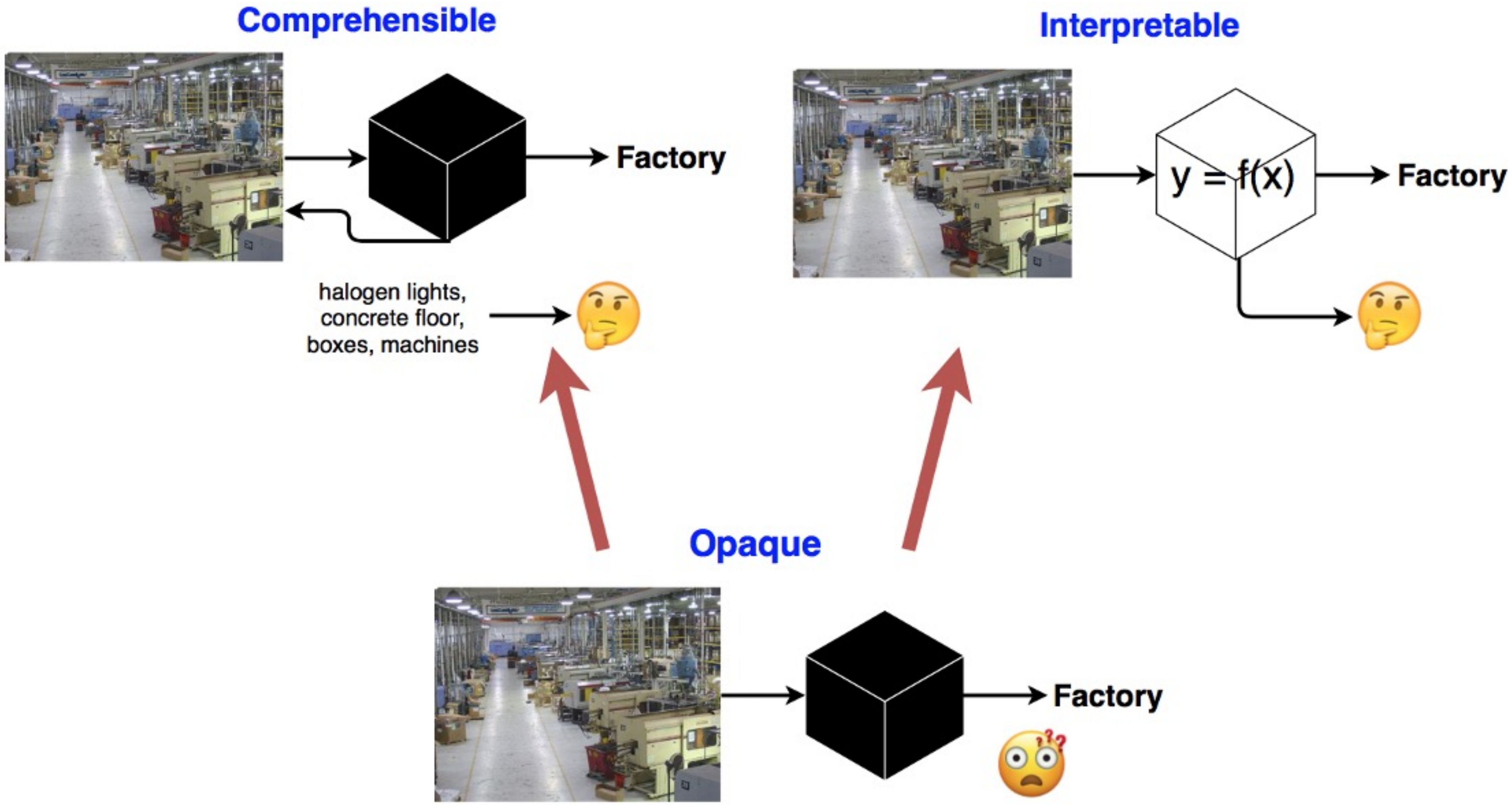
- A system that emits symbols along with its output that allow the user to relate properties of the inputs to their output. The user is responsible for compiling and comprehending the symbols, relying on her own implicit form of knowledge and reasoning about them.
- E.g. High dimensional data visualizations like t-SNE and receptive field visualization on convolutional neural networks

Opaque systems

- A system where the mechanisms mapping inputs to outputs are invisible to the user

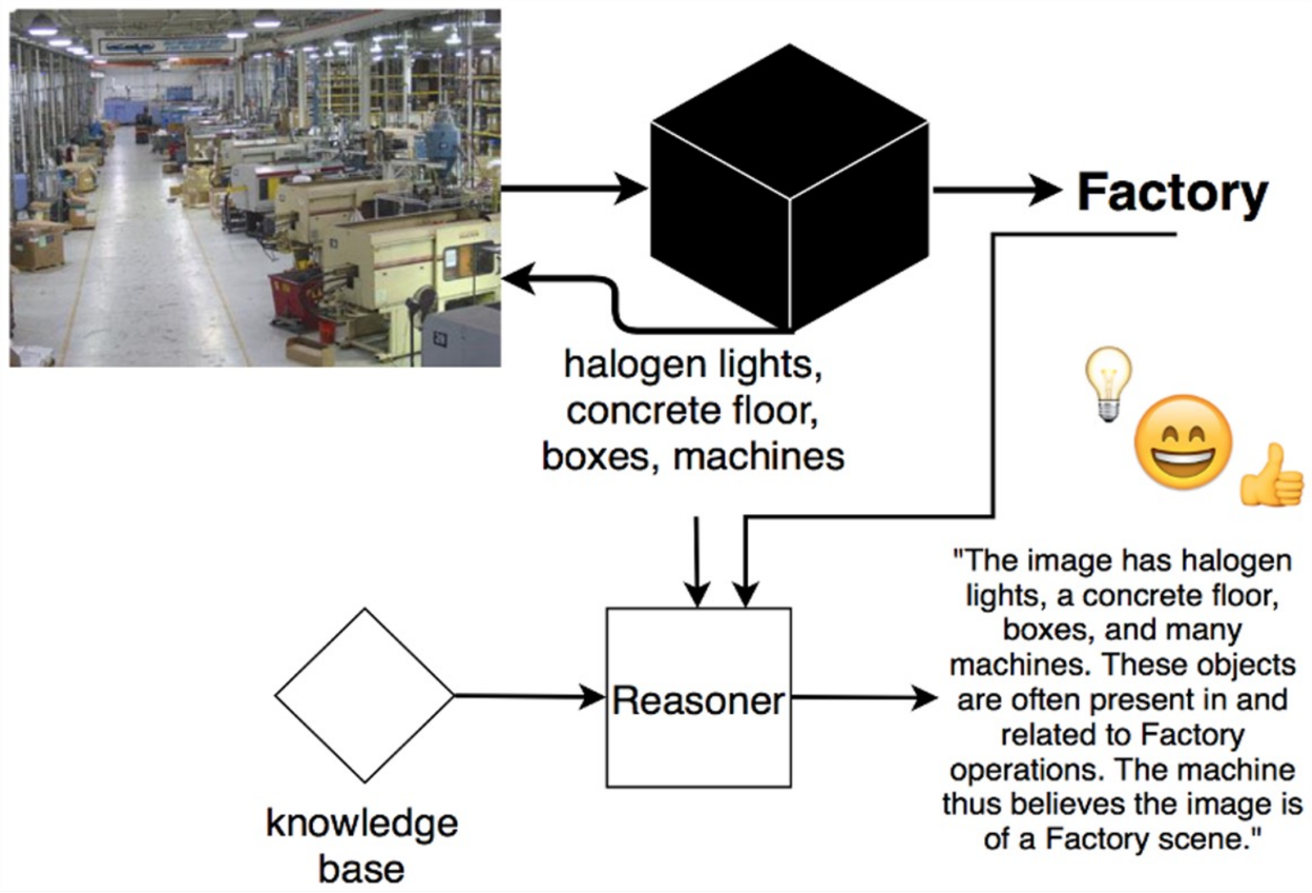
Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. 2071).

Explainability - Notions



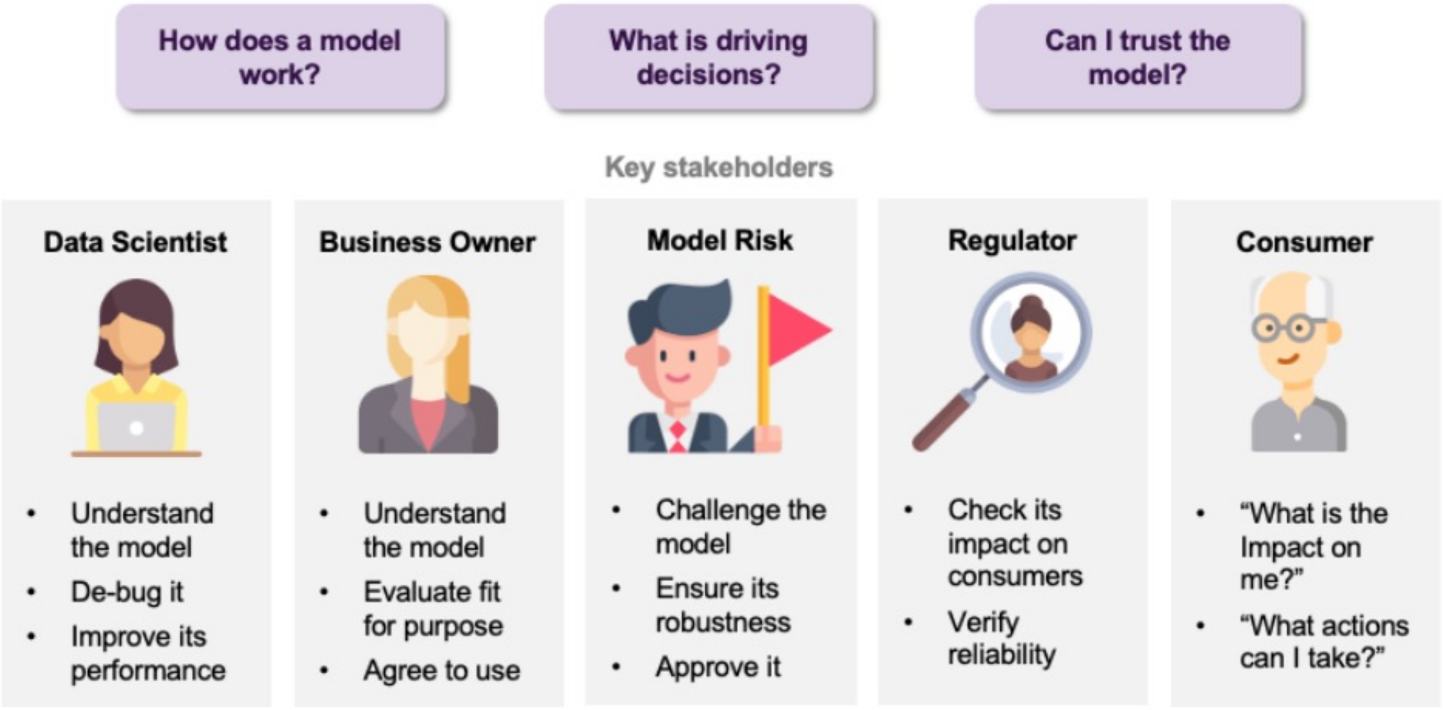
Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. 2071).

Explainability - Notions



Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. 1st International Workshop on Comprehensibility and Explanation in AI and ML Colocated with AI*IA 2017 (Vol. 2071).

Meaningful explanations depend on the stakeholder!



Belle, V., & Papantonis, I. (2020). Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*.

Approaches

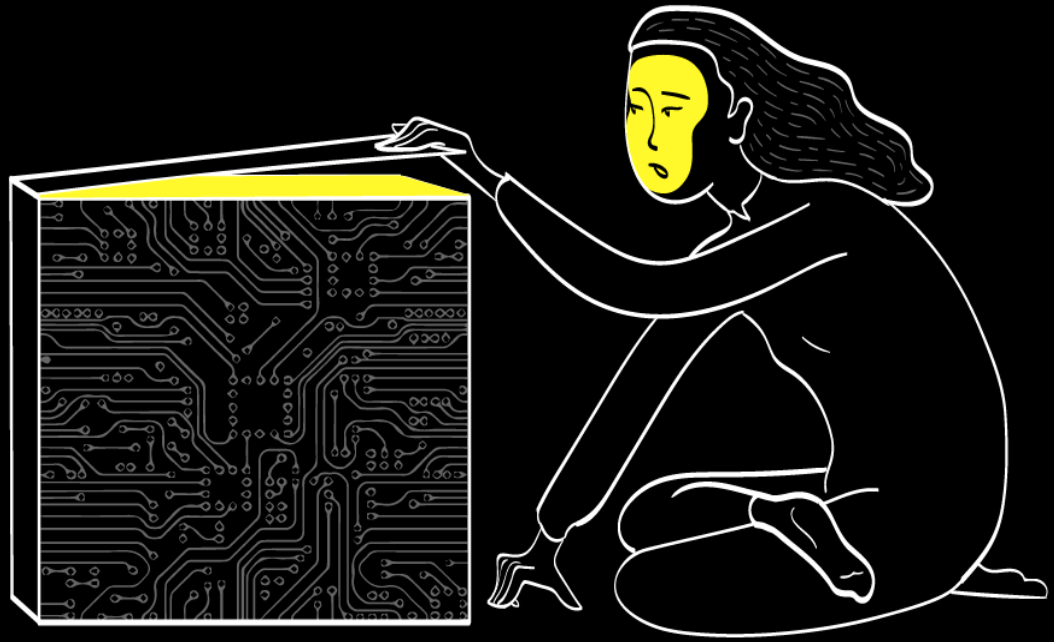
Expert Systems

Machine Learning

Neuro-symbolic Learning and Reasoning

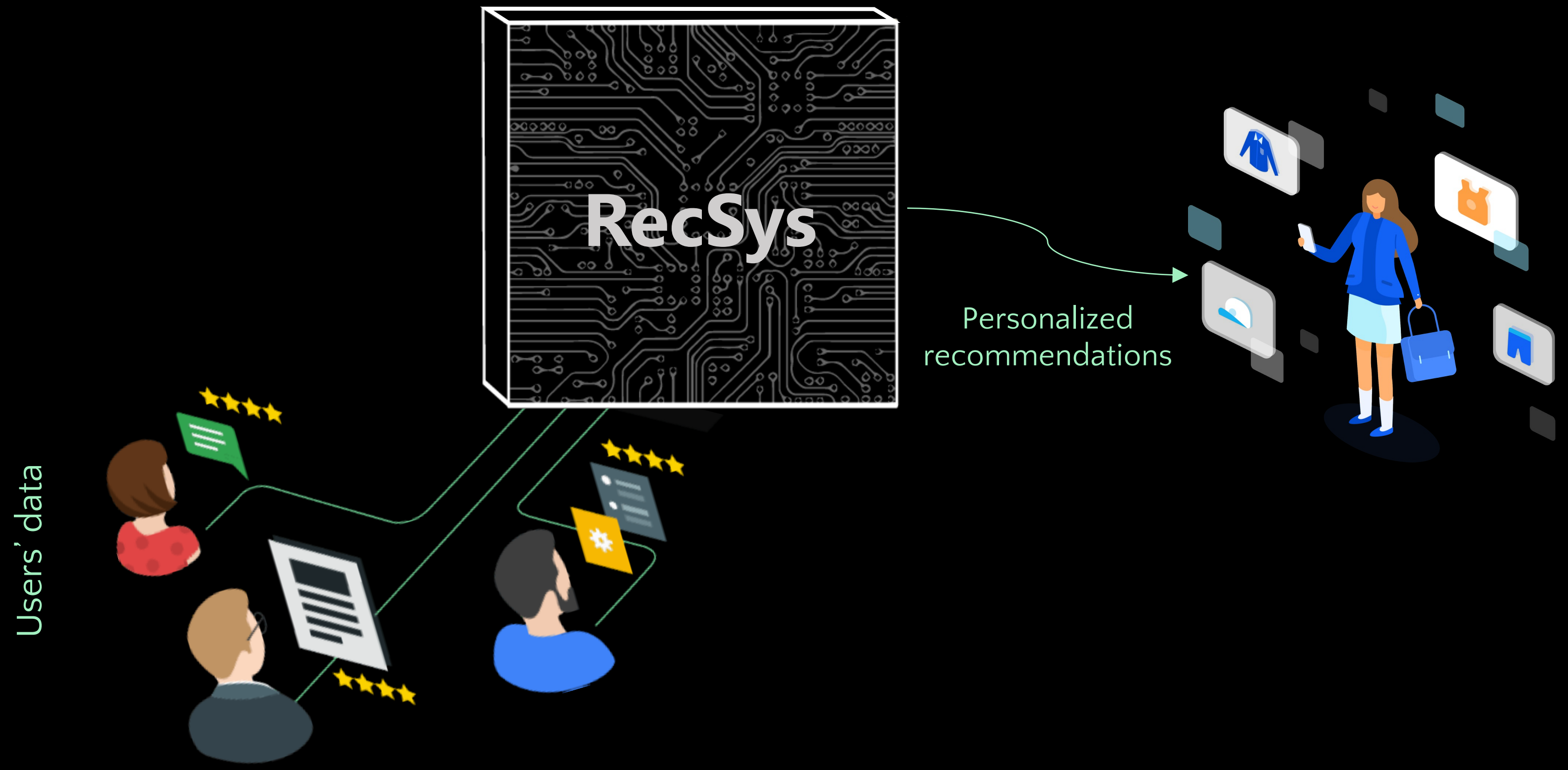
Recommender Systems

Recommender Systems



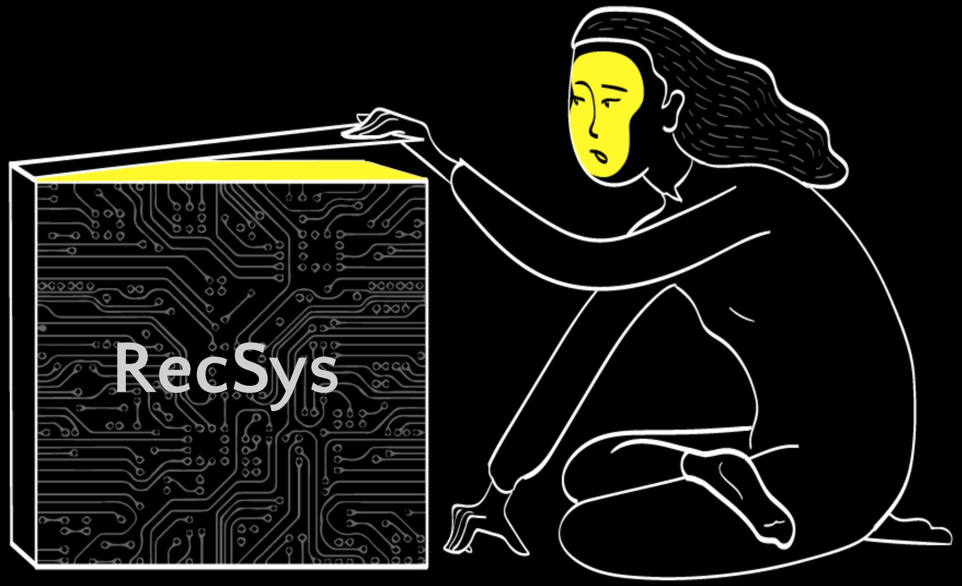
Coba, L.

Recommender Systems



Coba, L.

Explainability



Transparency

Scrutability

Trust

Effectiveness

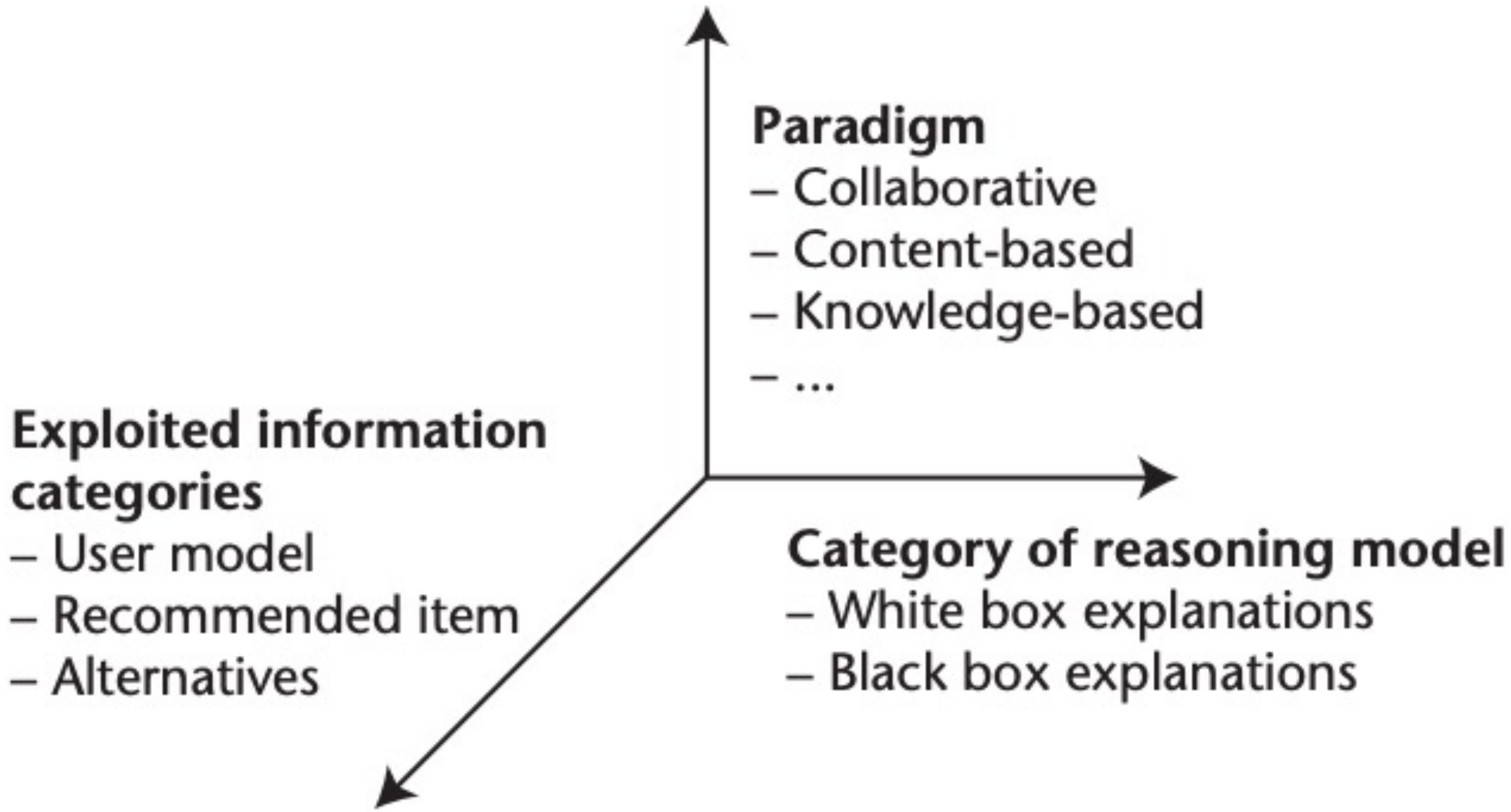
Persuasiveness

Efficiency

Satisfaction

Tintarev, Nava, and Judith Masthoff. "Designing and evaluating explanations for recommender systems." Recommender systems handbook. Springer, Boston, MA, 2011. 479-510.

Explainability

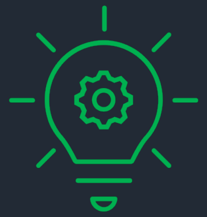


Coba, L.

RecSys paradigmes

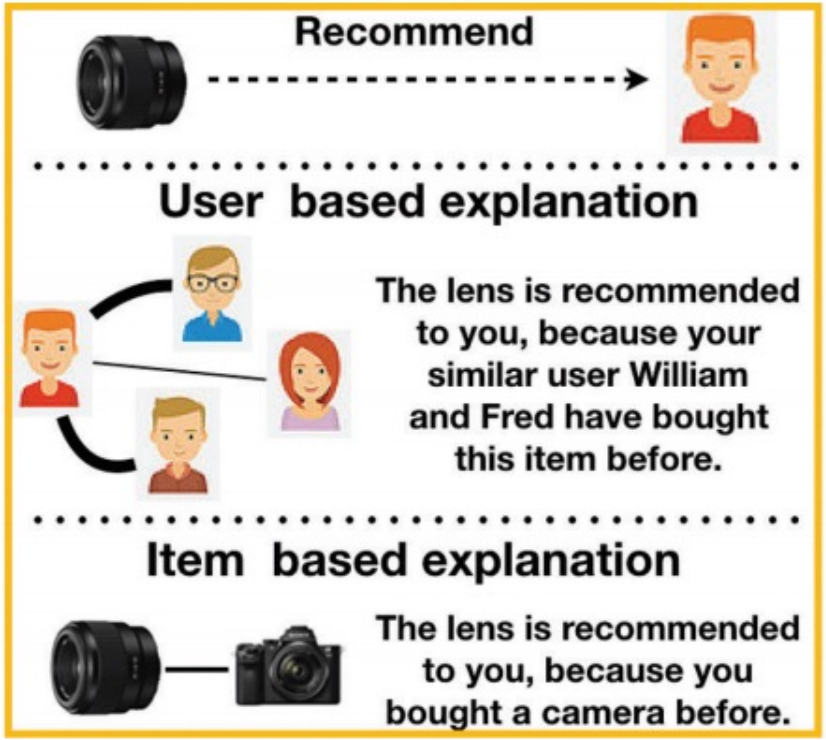
Collaborative Filtering

Content-based



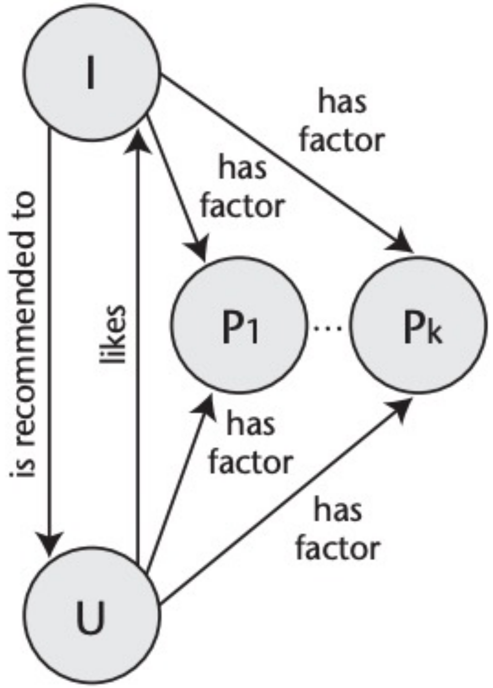
Knowledge-based

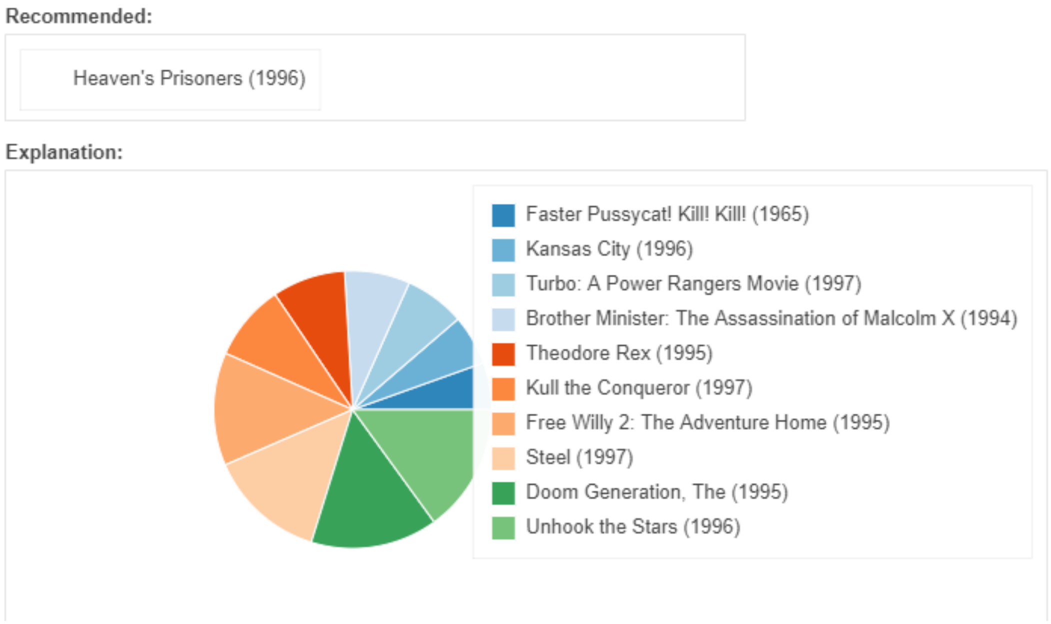
Collaborative Filtering



Coba, L., Rook, L., Zanker, M., & Symeonidis, P. (2019, March). Decision making strategies differ in the presence of collaborative explanations: two conjoint studies. IUI 19.

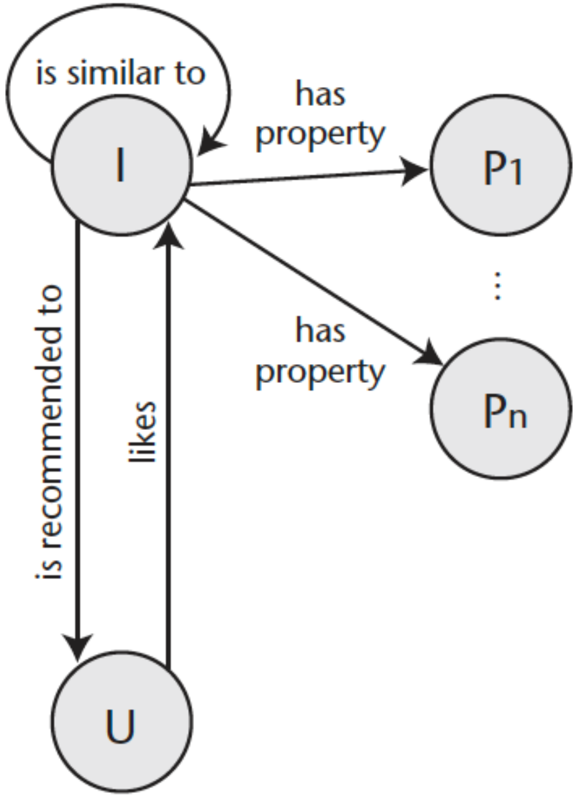
Collaborative Filtering

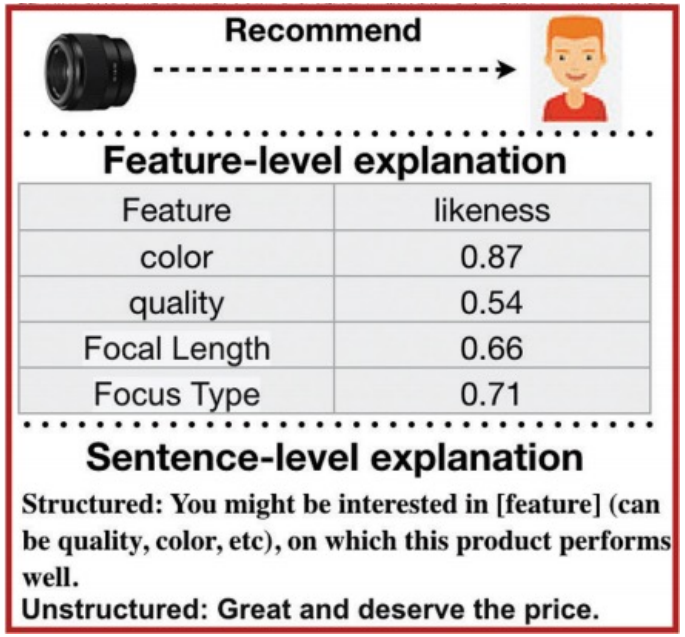




Coba, L., Confalonieri, R., Zanker, M. (2021). RECOXPLAINER: An Extensible Toolkit for Explainable Recommender Systems. AAAI 21.

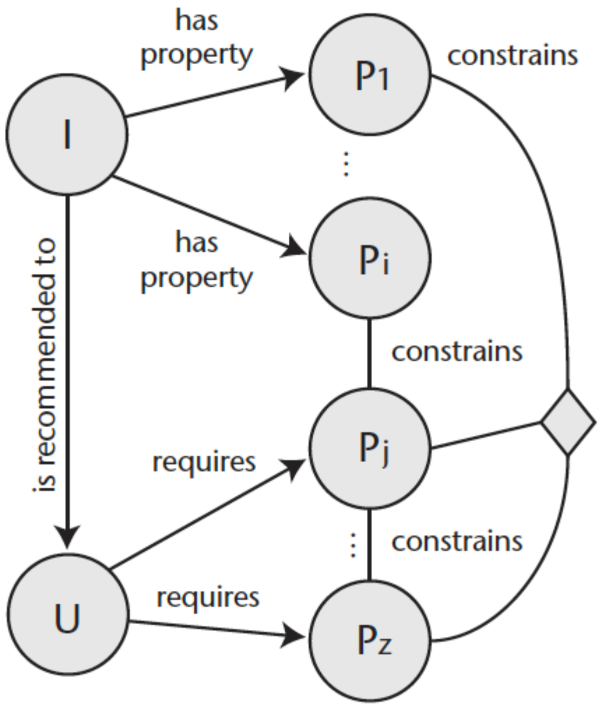
Content-based





Dominguez, V., Messina, P., Donoso-Guzmán, I., & Parra, D. (2019, March). The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. IUI 2019.

Knowledge-based



Information Categories

User Model
Are explanations tailored to the system's beliefs about a given user?

Recommendation Features
Is the recommendation dependent on the specific recommended item?

Alternatives
Do explanations argue in favour or against alternatives to the recommended item?

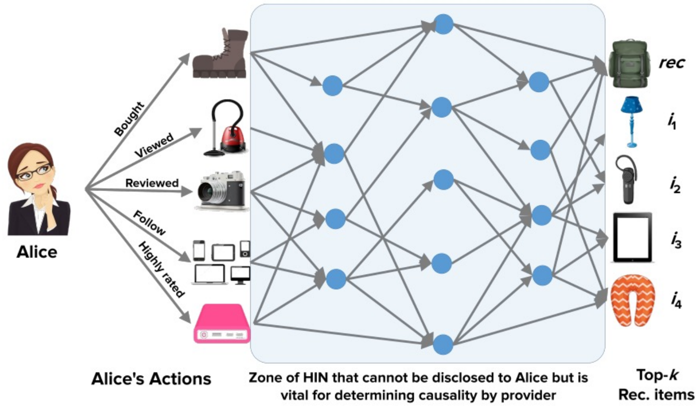
Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019, March). Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 379-390).

User Model

Recommendation Features

- From your MovieLens profile it seems that you prefer movies tagged as space, this movie takes you in space and it feels claustrophobic to be there. It keeps you on the edge of your seat the whole time.
- From your MovieLens profile it seems that you prefer movies tagged as visual, Gravity is unlike what we have seen on a cinema screen before and arguably it has one of the best uses of 3D in a movie.
- From your MovieLens profile it seems that you prefer movies tagged as intense, the movie a pretty intense ninety minutes, with Bullock's character constantly battling one catastrophe after another, and all of it is amazing to see.

Alternatives



Alice: Why did I receive this recommendation "Jack Wolfskin backpack"?

PRINCE: You **bought** "Adidas Hiking Shoes";
You **reviewed** "Nikon Coolpix Camera" with "Sleek! Handy on hikes!";
You **rated** "Intenso Travel Power Bank" highly.

If you **had not** done these actions:
"iPad Air" **would have replaced** "Jack Wolfskin backpack".

[User Model] Coba, L., Zanker, M., Rook, L., & Symeonidis, P. (2018). Exploring Users' Perception of Collaborative Explanation Styles. CBI 18.
[Rec Features] Chang, S., Harper, F. M., & Terveen, L. G. (2016, September). Crowd-based personalized natural language explanations for recommendations. RecSys 16.
[Alternatives] Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020). PRINCE: provider-side interpretability with counterfactual explanations in recommender systems. WSDM 20.

Reasoning model

White-box
How did the system derive a recommendation.

Black-box
What justifies the recommendation in the eyes of its recipient.

Coba, L., Confalonieri, R., Zanker, M. (2021). RecoXplainer: An Extensible Toolkit for Explainable Recommender Systems, AAAI 2021.

White-box example: Association-rules

Association rule mining algorithms

Detect rules of the form X → Y (e.g., beer → diapers) from a set of transactions T = {t1, t2, … tn} over a catalogue I

Measure quality by means of support, confidence used as a threshold to cut off unimportant rules

Pros: Interpretable by design
Cons: The model is not flexible

Explaining Black-box models

Model-Based Explanations are obtained by constraining the loss function

- Pros: No interpretable proxies needed
- Cons: Model loses flexibility

Post-Hoc Explanations are obtained by means of an interpretable proxy

- Pros: No under-the-hood reworking of the black-box
- Cons: Additional training step, not complete; Accuracy-interpretability trade-off

Coba, L., Confalonieri, R., Zanker, M. (2021). RecoXplainer: An Extensible Toolkit for Explainable Recommender Systems, AAAI 2021.

Model-based example: EMF

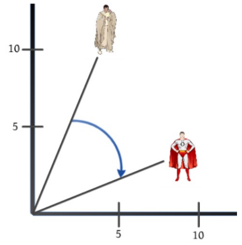
Desired Explanation style:

We recommend you Movie 1 because your neighbours' ratings for this movie are the following:

Rating	Number of Neighbours
★	0
★★	0
★★★	0
★★★★	10
★★★★★	23

How it works:

1



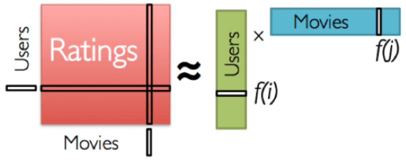
Building the neighborhood (NN) of the users

2

$$E_{u,i} = \sum_{\substack{\forall r \in R \\ r \geq P_\tau}} r * |NN^k(u)_{i,r}|$$

Determine the explainability of an item *i* by measuring in the identified neighborhood how frequently item has been highly rated

3



$$\sum_{i,j \in R} (r_{ij} - u_i v_j^T)^2 + \frac{\beta}{2}(\|u_i\|^2 + \|v_j\|^2) + \lambda\|u_i - v_j\|^2 E_{ij}$$

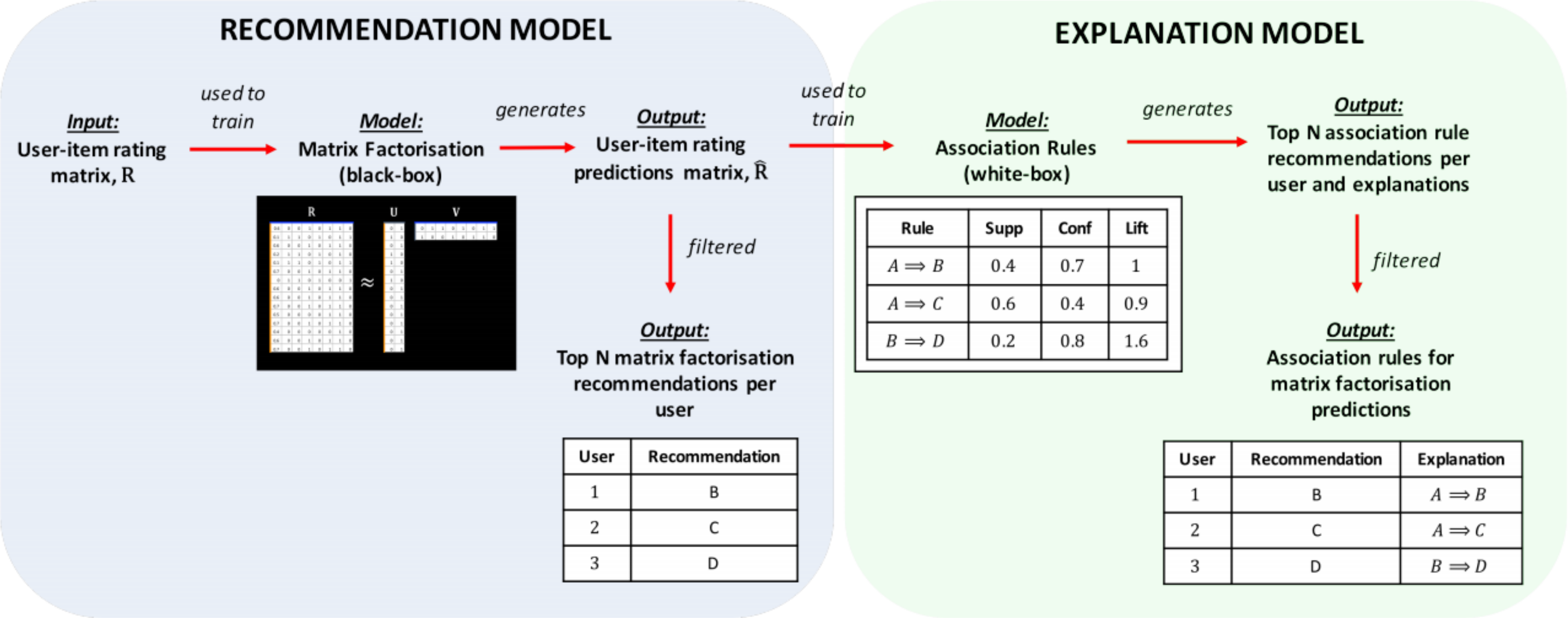
Soft constraint

We extend the traditional matrix factorization to recommend explainable items

Popularity bias

Coba, L.

Post-hoc example: Association-rules proxy



Rule	Supp	Conf	Lift
$A \Longrightarrow B$	0.4	0.7	1
$A \Longrightarrow C$	0.6	0.4	0.9
$B \Longrightarrow D$	0.2	0.8	1.6

User	Recommendation
1	B
2	C
3	D

User	Recommendation	Explanation
1	B	$A \Longrightarrow B$
2	C	$A \Longrightarrow C$
3	D	$B \Longrightarrow D$

Peake, G., & Wang, J. (2018). Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. KDD 18.

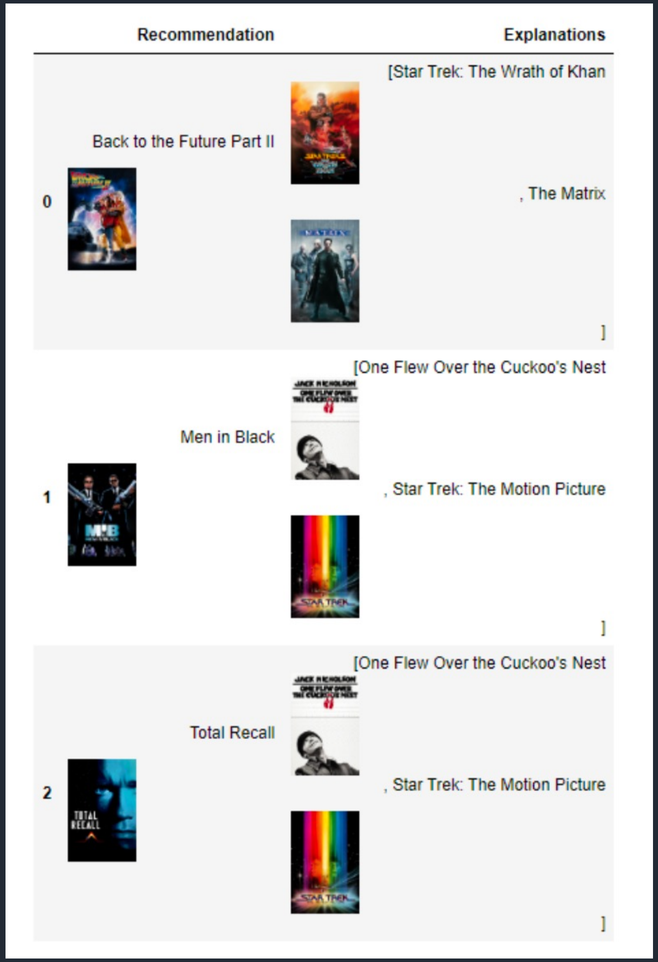
Post-hoc example: Association-rules proxy

Association rules to generate post-hoc explanations

Mine association rules on the generated predictions from a black-box RS

For each user filter the learned transactions such that antecedents are in the training set and consequents are unseen or non-interacted items

The resulting subset is ranked by support/confidence/lift. We keep the top-D consequents



Evaluation

Offline evaluation:

Based on a mathematical understanding of the user.

Examples: Model Fidelity, Mean Explainable Precision, E-nDCG

Online/user studies:

Require feedback from users and are specific to the goal of the explanation.

Examples: standardised psycolagiacal scales measuing trust, efficientcy, etc.

Evaluation

Offline evaluation:

Is based on a mathematical understanding of the user.

Examples: Model Fidelity, Mean Explainable Precision, E-nDCG

Online/user studies:

Requires feedback from users and are specific to the goal of the explanation.

Examples: standardised psycolagiacal scales measuing trust, efficientcy, etc.

Evaluation

Offline evaluation:

Is based on a mathematical understanding of the user.

Examples: Model Fidelity, Mean Explainable precision, E-nDCG

Online/user studies:

Requires feedback from users and are specific to the goal of the explanation.

Examples: standardised psycolagiacal scales measuing trust, efficientcy, etc.

(Bonus) Evaluate the Recommender:

Enable developers to undestand the reasoning and the quality of the recommender

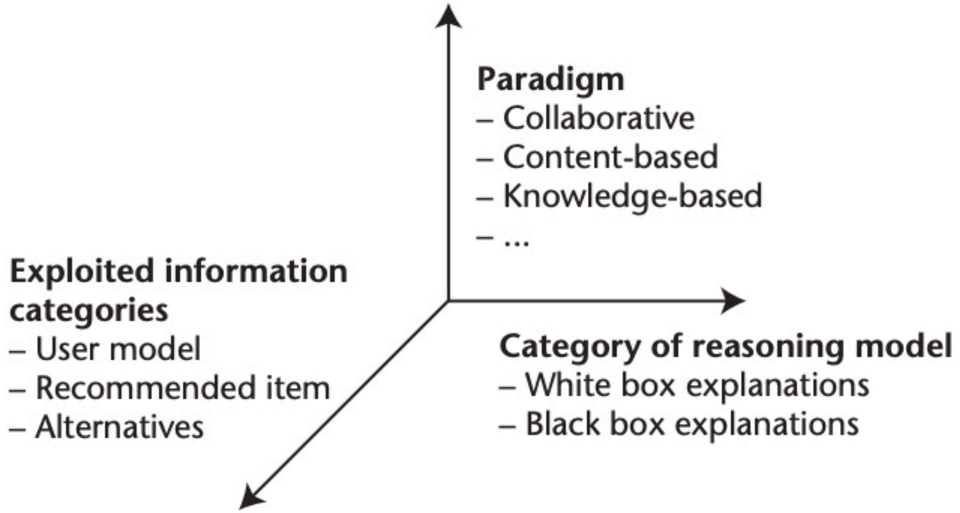
Summary

In RecSys explanations are designed based on the paradigm, the information category and the reasoning model

Be aware of the interpretability-accuracy trade-off

Advanteges as trust, persuasivness, efficiency, scrutiny, etc

Offline evaluation via proxy, online evaluation via standardised scales.



Thank you!