RUTGERS
UNIVERSITY | NEW BRUNSWICK

Towards Trustworthy Recommender Systems: From Shallow Model to Deep Model to Large Model

Yongfeng Zhang

Department of Computer Science, Rutgers University

yongfeng.zhang@rutgers.edu

http://yongfeng.me

Recommender Systems are Everywhere

- Influence our daily life by providing personalized services

E-commerce

Social Networks

News Feeding

Search Engine

Navigation

Travel Planning

Professional Networks

Healthcare

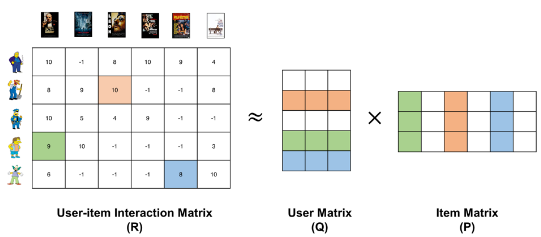
Online Education

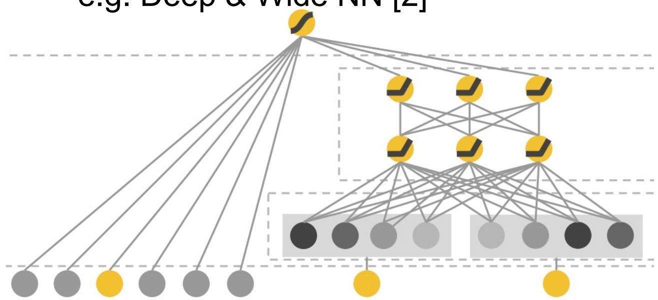
Technical Advancement of Recommender Systems

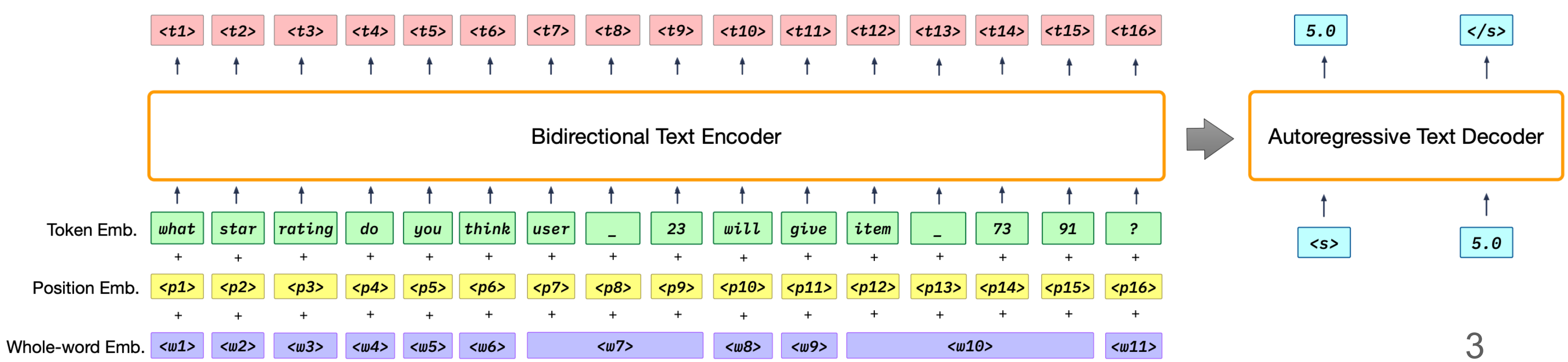
- From Shallow Model, to Deep Model, and to Large Model

Shallow Models → Deep Models → Large Models

e.g. Matrix Factorization [1] | e.g. Deep & Wide NN [2] | e.g. P5 [3]



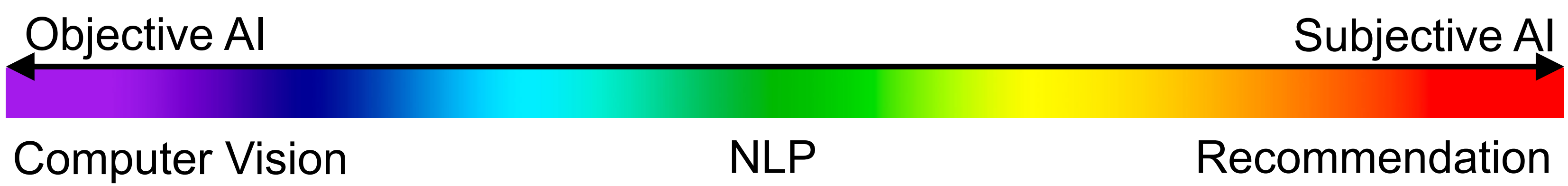




[1] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 42, no. 8 (2009): 30-37.
[2] Cheng, Heng-Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson et al. "Wide & deep learning for recommender systems." DLRS 2016.
[3] Geng, Shijie, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)." RecSys 2022.

Objective AI vs. Subjective AI

- Recommendation is unique in the AI family
  - Recommendation is most close to human among all AI tasks
  - Recommendation is a very representative Subjective AI
  - Thus, leads to many unique challenges in recommendation research



(Relatively) far from human.
Problems have exact answers.

Very close to human.
Problems have no absolute answers.

RUTGERS

Computer Vision: (mostly) Objective AI Tasks

Objective AI

Subjective AI

Computer Vision

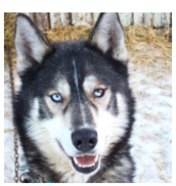
NLP

Recommendation

Image Classification

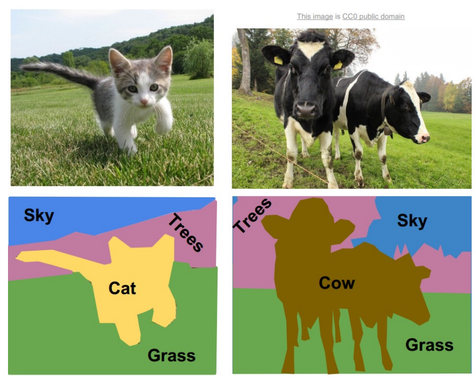
cat

dog

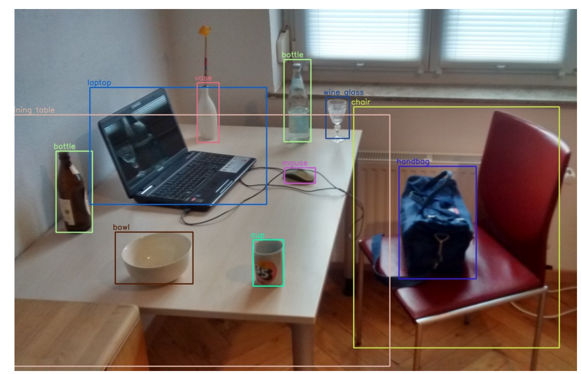


Husky like a wolf

Image Segmentation

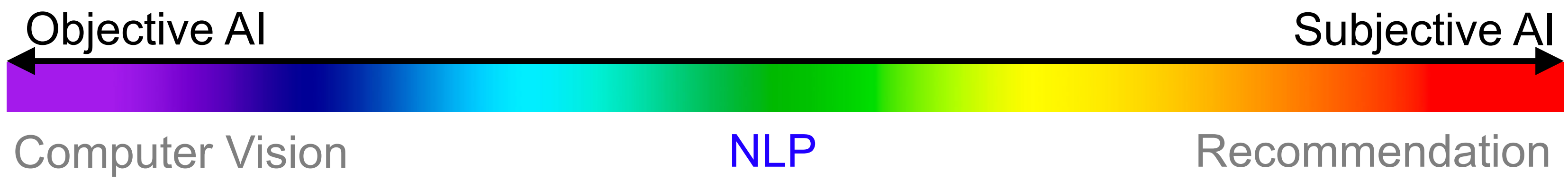


Object Detection

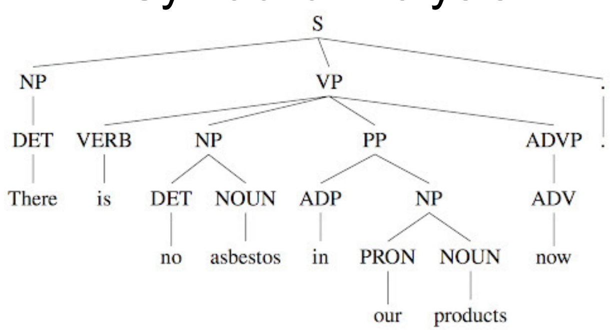


RUTGERS

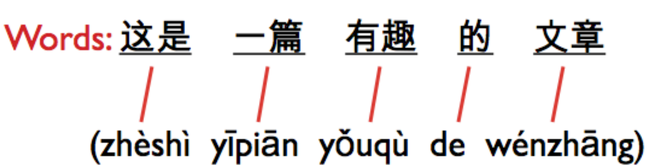
NLP: partly Objective, partly Subjective



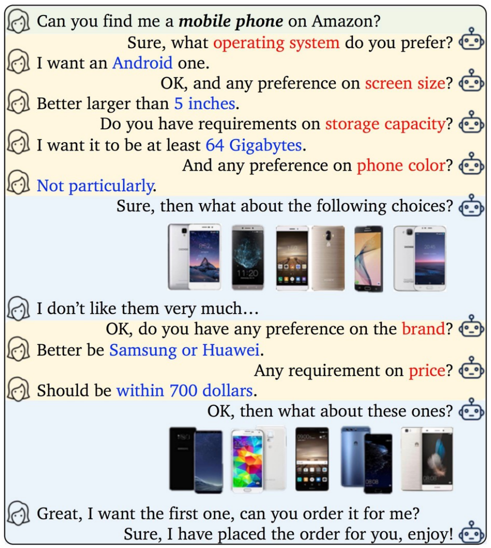
Syntactic Analysis



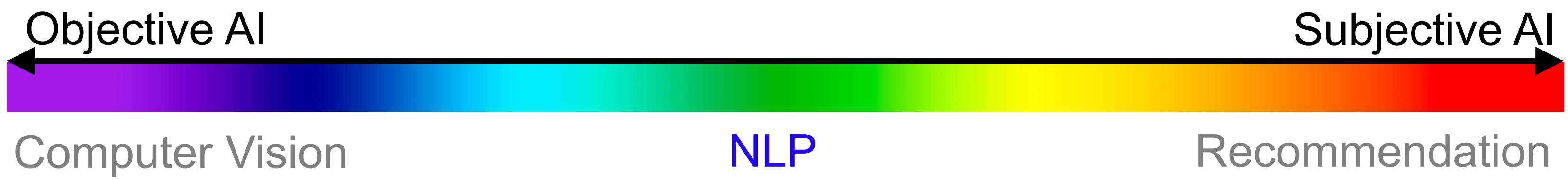
Word Segmentation



Dialog Systems



Recommendation: mostly Subjective AI Tasks



Movie Recommendation

Product Recommendation

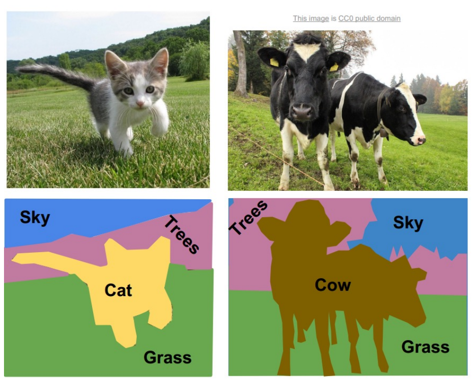
Subjective AI needs Explainability

- Objective vs. Subjective AI on Explainability

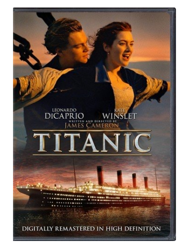
Objective AI
Human can directly identify if the AI-produced result is right or wrong

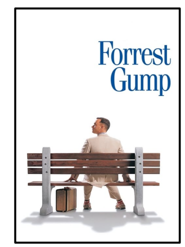

cat

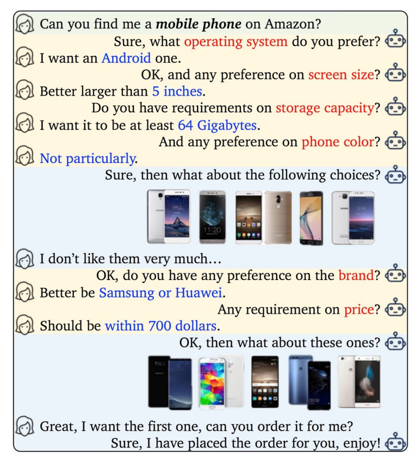

dog



Subjective AI
Human can hardly identify if the AI-produced result is right or wrong





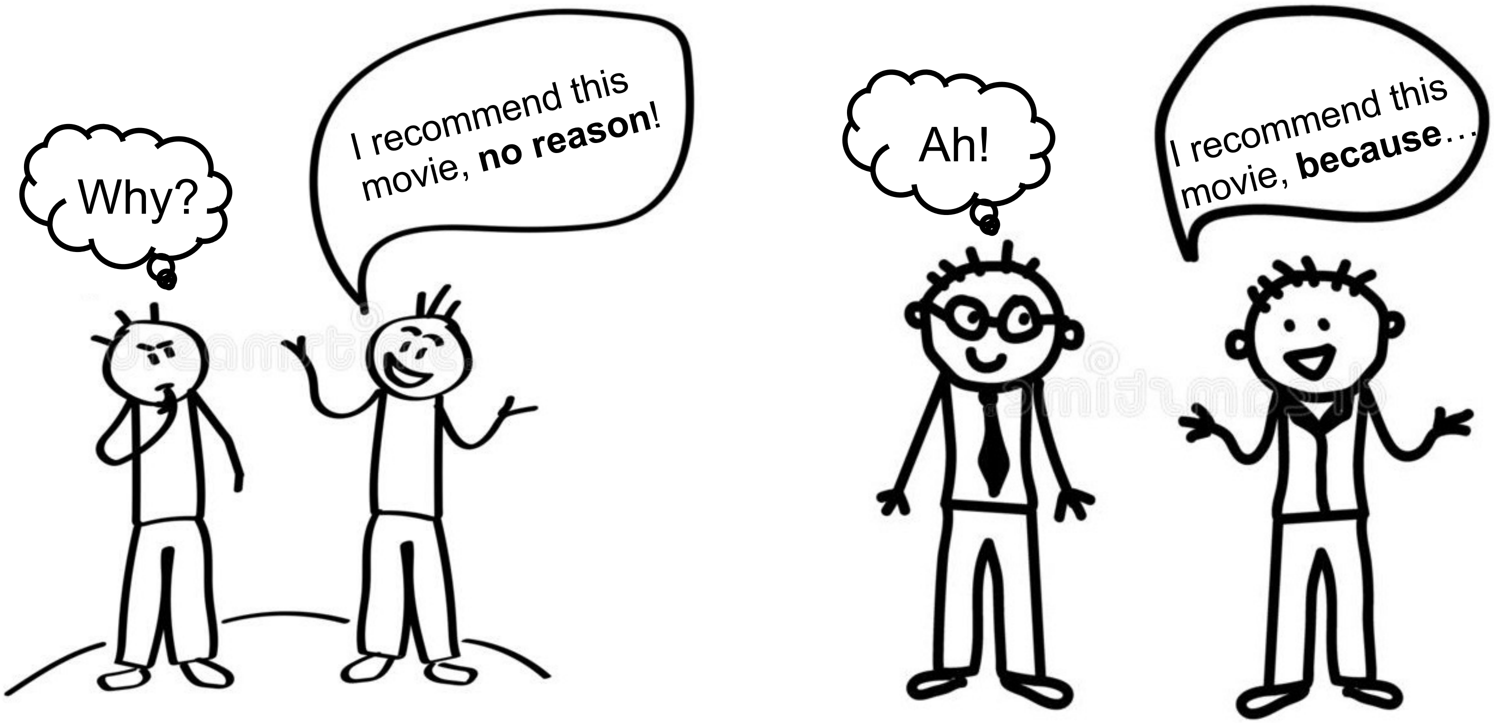


Nothing is definitely right or wrong.

Highly subjective, and usually personalized.

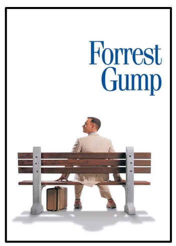
Subjective AI needs Explainability

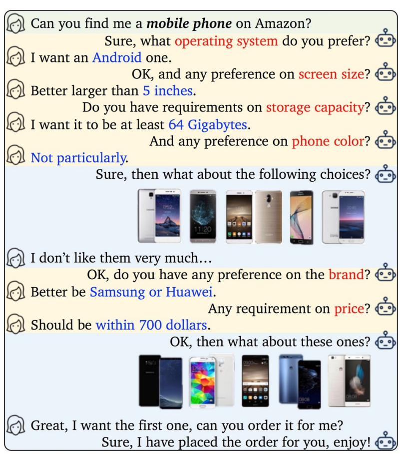
- In many cases, it doesn't matter what you recommend, but how you explain your recommendation
- How do humans make recommendation?



Subjective AI needs Fairness

- Users cannot easily identify if something is right or wrong
  - They have to take the recommendations as is
  - Users are very vulnerable
  - Users could be manipulated, utilized or even cheated by the system





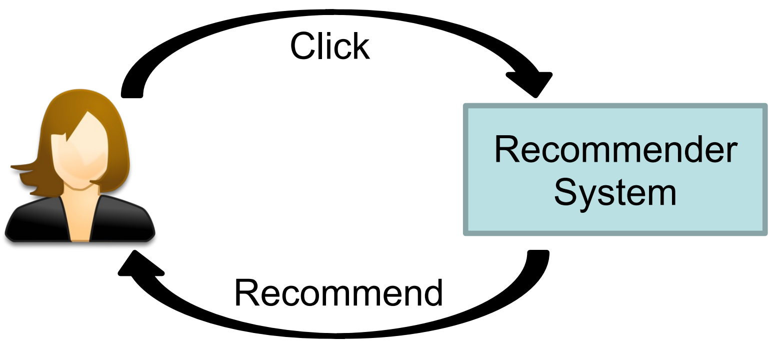
Nothing is definitely right or wrong.

Highly subjective, and usually personalized.

Users need to be treated fairly.

Subjective AI leads to Echo Chambers

- Users don't know which recommendations are "right" and which are "wrong", they just click. [5]
- Lack of explanation makes the problem worse.

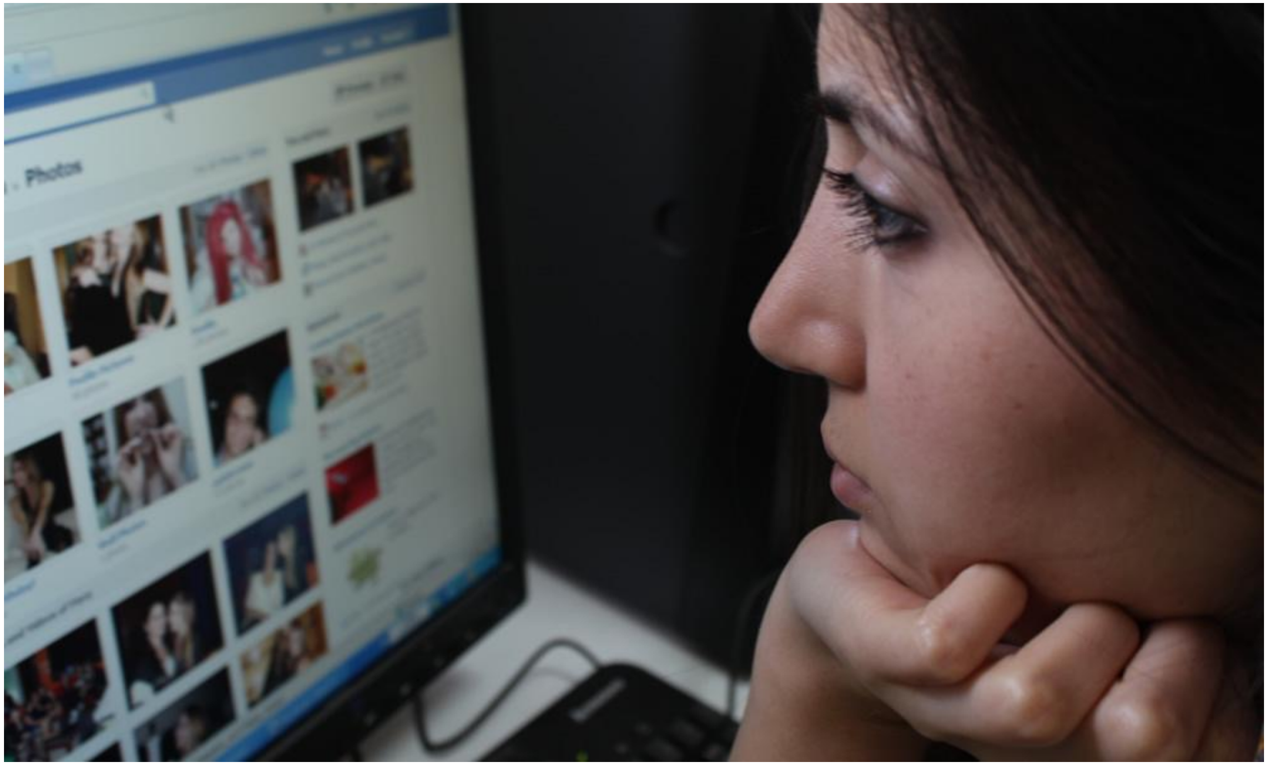


The more you like something, the more RS will recommend similar things, and thus you like them even more.

[1] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, Y. Zhang. Understanding Echo Chambers in E-commerce Recommender Systems. SIGIR'20.

Subjective AI needs Controllability

- Users almost have no control of their recommender system
  - They can only passively receive recommendations

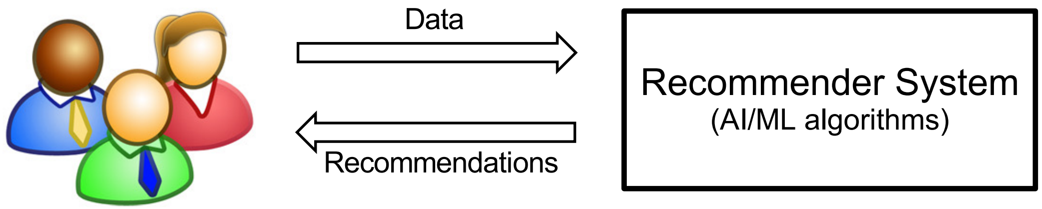


Trustworthy and Responsible Recommendation

- Explainability, Fairness, Echo Chambers, Controllability
- And many more …
  - Robustness, Accountability, Privacy, etc.

RecSys as a Human-centered AI task

- Recommender System (RS) is a representative Human-centered AI task
  - Naturally involves human-in-the-loop
  - Influences human decision making everyday and everywhere



- A wide scope of applications
  - E-commerce (product recommendation)
  - Smart and Connected Communities (driving route recommendation / passenger recommendation)
  - Social Networks (friend/tweet recommendation)
  - Sharing Economy (house recommendation)
  - Search Engines (personalized search / advertising)
  - Travel and Planning Services (ticket and hotel recommendation)
  - Professional Networks (job recommendation)

Even some high-stake application scenarios

- Financial Services (financial / investment recommendation)
- Medial Services (doctor recommendation, patient-doctor matching)
- Legal Services (parole decision recommendation)

Example: Resume Ranking and Recommendation
– Explainability for Responsible AI

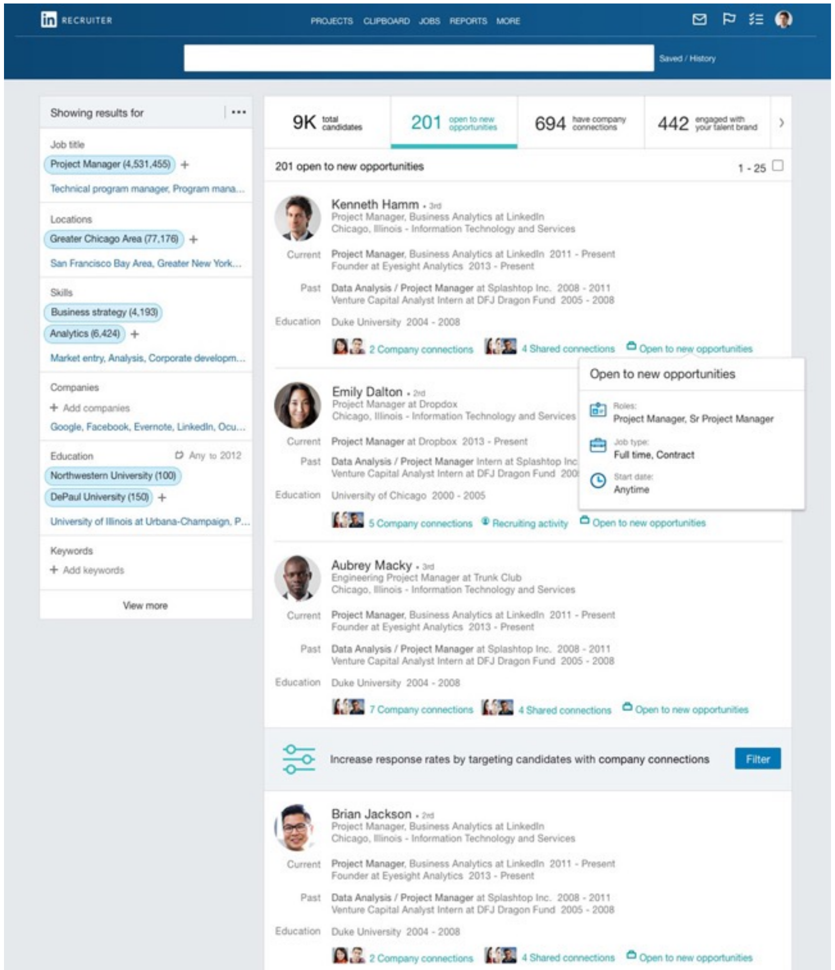


Figure 1: A (mocked) screenshot from the LinkedIn Recruiter (credit to [1])

Background: Many companies use automated tools such as LinkedIn for recruiting

When a job is posted, could receive thousands of applications -- impossible for HR to manually screen every candidate's resume

Solution: Use ML to rank the candidates based on some "matching score" between resume and job description.

You only have a chance of interview if the algorithm ranks your resume at top positions (e.g., top-10)

Problem:
From recruiter's perspective:
Why this candidate is a better fit than another?

From applicant's perspective:
Why should I trust the algorithm?
Why should my whole career be decided by a machine?

To answer these **WHY** questions, we need Explainable AI!

https://engineering.linkedin.com/blog/2019/04/ai-behind-linkedin-recruiter-search-and-recommendation-systems

Human-centered Explainable, Fair and Controllable AI

- AI in Human-centered Tasks
  - We not only want to know a model works (e.g., make accurate predictions)
  - We also want to know **why** it works (e.g., why the model makes this decision, is it fair, and why we should trust this decision)
  - Human controls AI, rather than AI controls human
- Even more important in **high-stake applications** related to health, safety, and law

Healthcare

Financial Assistants

Legal Assistants

  - Errors/bias may cause severe loss in life, money, and reputation

- Explainable AI helps humans to make better decisions

RUTGERS

The Scope of AI

- AI ≠ ML, AI ⊃ ML

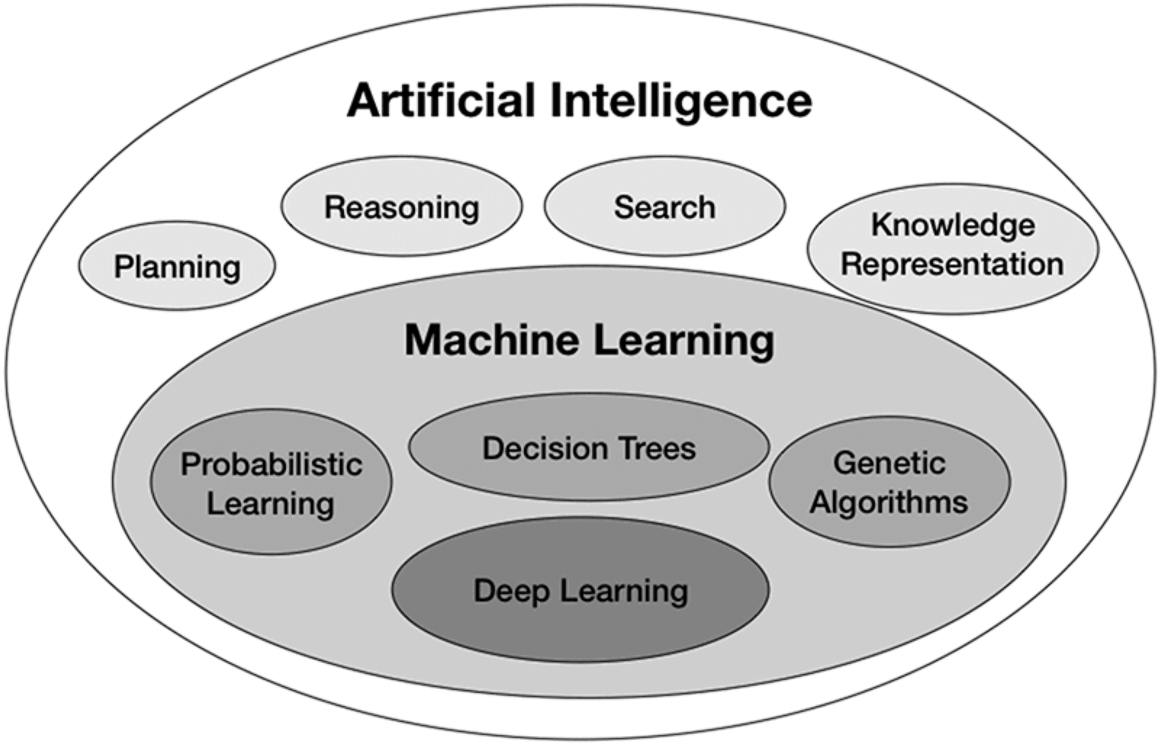
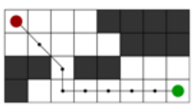


Image credit to Marcus G, Davis E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon; 2019 Sep 10.

A (very rough) History of AI Research

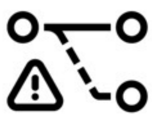
- Symbolic Reasoning Approach to AI
  - Mid-1950s to late 1980s

Example Methods:

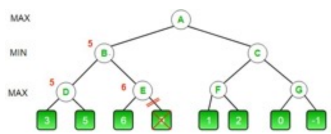


A* Search

Knowledge Representation and Reasoning

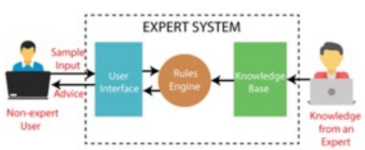


Production Rules



Alpha-Beta Pruning

Example Systems:

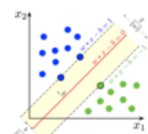


Expert systems
(If-Then production rules)

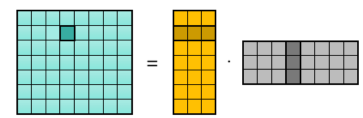
Chess AI
(IBM Deep Blue)

- Machine Learning Approach to AI
  - Early 1990s to date

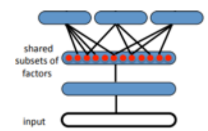
Example Methods:



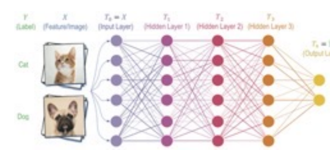
Support Vector Machine



Matrix Factorization

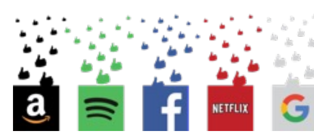


Representation Learning



Deep Neural Networks

Example Systems:



Recommender Systems

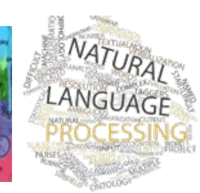



Image and Language Processing

Symbolism vs Connectionism - A comparison

- a.k.a. Rationalism vs Empiricism approaches to AI

Symbolism/Rationalism

A top-down design approach

Advantages:

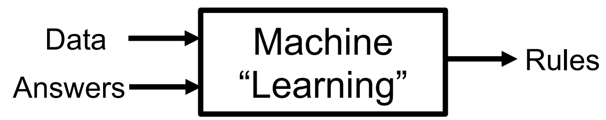
- Accurate decision
- Highly explainable & human readable

Disadvantages:

- Extensive expert human efforts
- Difficult to handle noisy data

Connectionism/Empiricism

A bottom-up design approach



Advantages:

- Less human efforts
- Better at working with noisy data

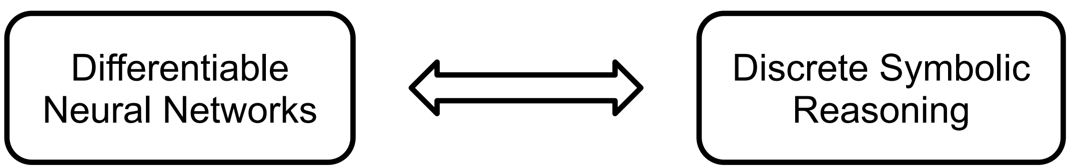
Disadvantages:

- Decisions are usually approximate
- Difficult to explain (black-box model)

Bridge the best of two Worlds?

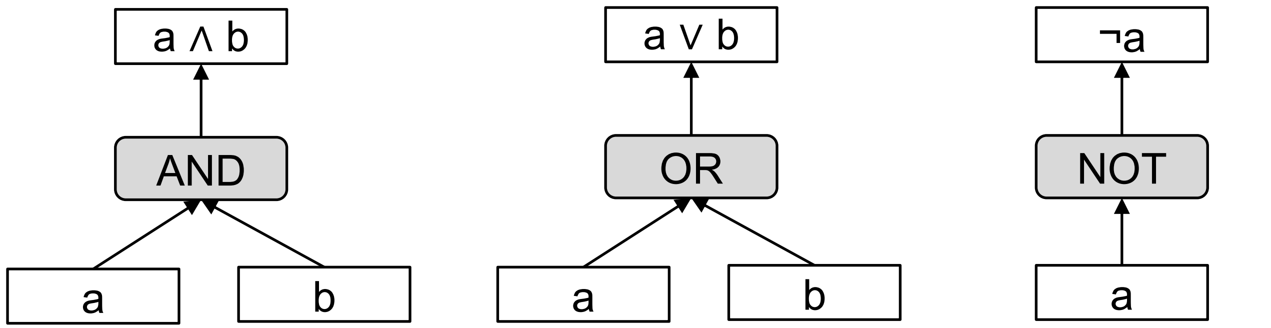
- Neural Symbolic Machine Learning
  - Grant learning systems with reasoning ability
  - Improve decision accuracy
  - Improve decision transparency

- Key Challenge
  - How to bridge differentiable neural networks and discrete symbolic reasoning in shared architecture for optimization and inference



Neural Logic Reasoning

- Key idea [4-8]
  - Learning logical variables as vectors in logical embedding space
  - Learning logical operations as neural modules in the latent space



In our implementation, AND(*,*), OR(*,*), NOT(*) are simple 2-layer neural networks

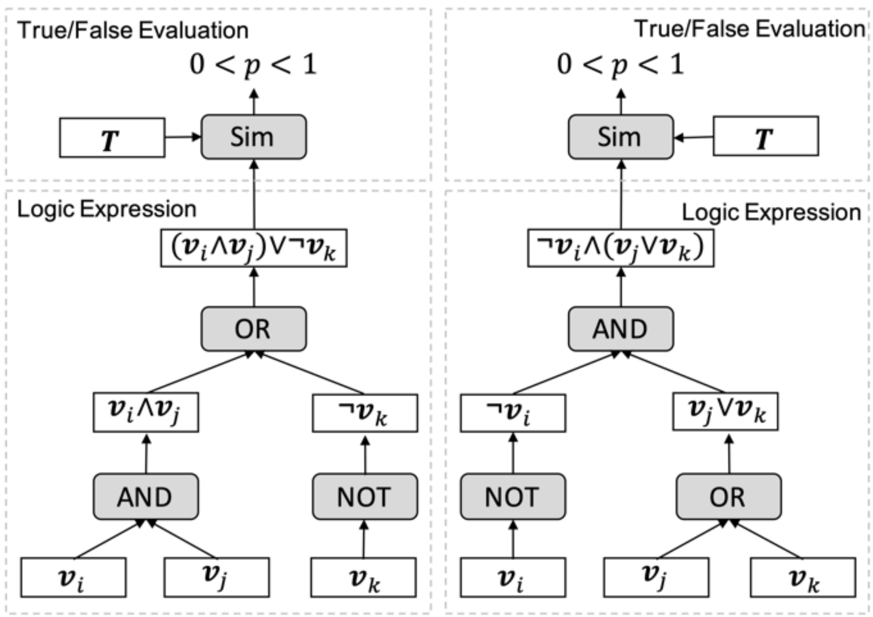
$$\mathrm{AND}(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{H}_{a2} f(\mathbf{H}_{a1}(\mathbf{w}_i \oplus \mathbf{w}_j) + \mathbf{b}_a) \quad \mathrm{NOT}(\mathbf{w}) = \mathbf{H}_{n2} f(\mathbf{H}_{n1}\mathbf{w} + \mathbf{b}_n)$$

[4] Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. "Neural Logic Reasoning", CIKM 2020.
[5] Hanxiong Chen, Shaoyun Shi, Yunqi Li and Yongfeng Zhang. “Neural Collaborative Reasoning”, WWW 2021.
[6] Hanxiong Chen, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu and Yongfeng Zhang. “Graph Collaborative Reasoning”, WSDM 2022.
[7] Jianchao Ji, Zelong Li, Shuyuan Xu, Max Xiong, Juntao Tan, Yingqiang Ge, Hao Wang, Yongfeng Zhang. “Counterfactual Collaborative Reasoning”, WSDM 2023.
[8] Wenyue Hua and Yongfeng Zhang. “System 1 + System 2 = Better World: Neural-Symbolic Chain of Logic Reasoning”, EMNLP 2022.

Logic-Integrated Neural Network (LINN)

- Any logical expression can be dynamically assembled into a neural structure

$(v_i \wedge v_j) \vee \neg v_k = T$ $\neg v_i \wedge (v_j \vee v_k) = F$



(a) $(v_i \wedge v_j) \vee \neg v_k$ **(b)** $\neg v_i \wedge (v_j \vee v_k)$

Optimize with task dependent loss, e.g.,:

Cross-Entropy Loss:

$$L_t = L_{ce} = -\sum_{e_i \in E} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Pair-wise Ranking Loss:

$$L_t = L_{bpr} = -\sum_{e^+} \log\left(sigmoid(p(e^+) - p(e^-))\right)$$

Logical Regularization over Neural Modules

- How do we know the AND(*,*) module is really doing logical AND?
  - And also, for OR(*,*) and NOT(*)?

- Logical Regularization
  - Logical operators should satisfy a set of basic requirements

	Logical Rule	Equation	Logic Regularizer r_i
NOT	Negation	$\neg T = F$	$r_1 = \sum_{w \in W \cup \{T\}} Sim(\text{NOT}(\mathbf{w}), \mathbf{w})$
	Double Negation	$\neg(\neg w) = w$	$r_2 = \sum_{w \in W} 1 - Sim(\text{NOT}(\text{NOT}(\mathbf{w})), \mathbf{w})$
AND	Identity	$w \wedge T = w$	$r_3 = \sum_{w \in W} 1 - Sim(\text{AND}(\mathbf{w}, \mathbf{T}), \mathbf{w})$
	Annihilator	$w \wedge F = F$	$r_4 = \sum_{w \in W} 1 - Sim(\text{AND}(\mathbf{w}, \mathbf{F}), \mathbf{F})$
	Idempotence	$w \wedge w = w$	$r_5 = \sum_{w \in W} 1 - Sim(\text{AND}(\mathbf{w}, \mathbf{w}), \mathbf{w})$
	Complementation	$w \wedge \neg w = F$	$r_6 = \sum_{w \in W} 1 - Sim(\text{AND}(\mathbf{w}, \text{NOT}(\mathbf{w})), \mathbf{F})$
OR	Identity	$w \vee F = w$	$r_7 = \sum_{w \in W} 1 - Sim(\text{OR}(\mathbf{w}, \mathbf{F}), \mathbf{w})$
	Annihilator	$w \vee T = T$	$r_8 = \sum_{w \in W} 1 - Sim(\text{OR}(\mathbf{w}, \mathbf{T}), \mathbf{T})$
	Idempotence	$w \vee w = w$	$r_9 = \sum_{w \in W} 1 - Sim(\text{OR}(\mathbf{w}, \mathbf{w}), \mathbf{w})$
	Complementation	$w \vee \neg w = T$	$r_{10} = \sum_{w \in W} 1 - Sim(\text{OR}(\mathbf{w}, \text{NOT}(\mathbf{w})), \mathbf{T})$

- Logical Regularized Loss $L_1 = L_t + \lambda_l R_l = L_t + \lambda_l \sum_i r_i$

Application 1: Solving Logical Equations

- 10k logical variables, 30k randomly generated logical equations
  - In Disjunctive Normal Form (DNF)

$$(\neg v_{80} \wedge v_{56} \wedge v_{71}) \vee (\neg v_{46} \wedge \neg v_{7} \wedge v_{51} \wedge \neg v_{47} \wedge v_{26}) \vee v_{45} \vee (v_{31} \wedge v_{15} \wedge v_{2} \wedge v_{46}) = T$$
$$(\neg v_{19} \wedge \neg v_{65}) \vee (v_{65} \wedge \neg v_{24} \wedge v_{9} \wedge \neg v_{83}) \vee (\neg v_{48} \wedge \neg v_{9} \wedge \neg v_{51} \wedge v_{75}) = F$$
$$\neg v_{98} \vee (\neg v_{76} \wedge v_{66} \wedge v_{13}) \vee v_{97}(\wedge v_{89} \wedge v_{45} \wedge v_{83}) = T$$
$$v_{43} \wedge v_{21} \wedge \neg v_{53} = F$$

  - Expressions: training (80%), validation (10%), and test (10%) sets.
  - Task: Predict the T/F value for expressions in test sets

$$(v_{65} \wedge \neg v_{24} \wedge v_{9} \wedge \neg v_{83}) \vee (\neg v_{48} \wedge \neg v_{9} \wedge \neg v_{51} \wedge v_{75}) = ?$$

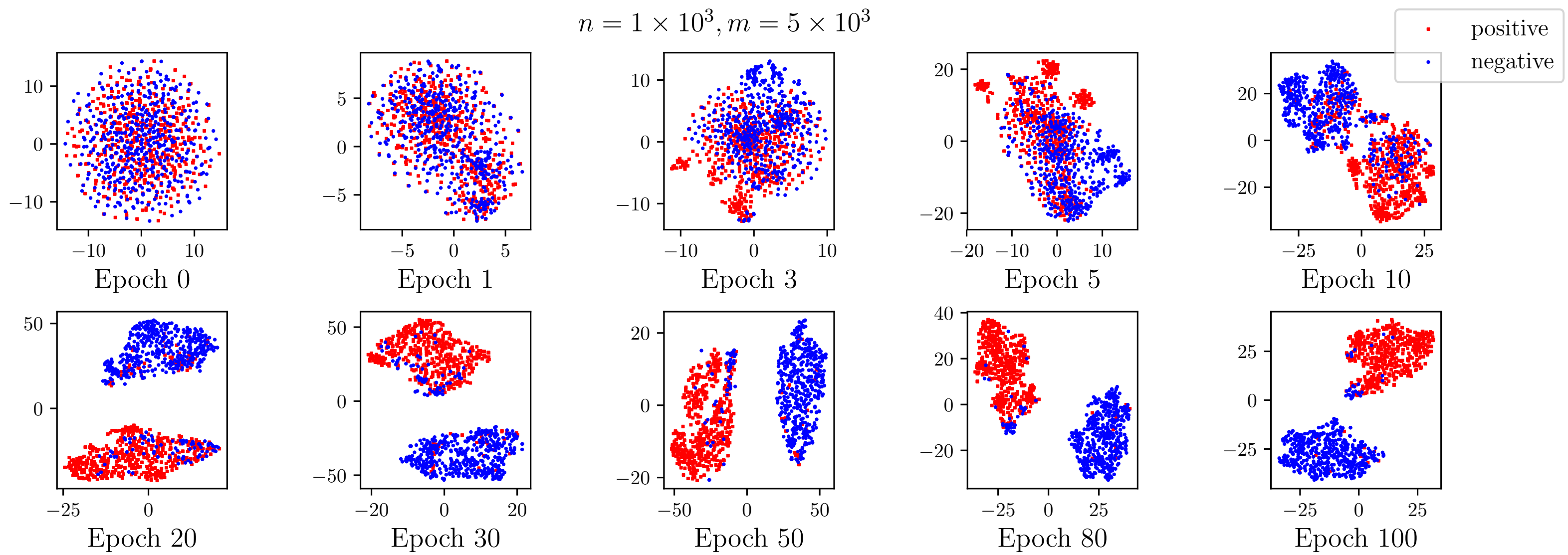
	$n = 1 \times 10^3, m = 3 \times 10^3$		$n = 1 \times 10^4, m = 3 \times 10^4$	
	Accuracy	RMSE	Accuracy	RMSE
Bi-RNN [32]	0.6493 ± 0.0102	0.4736 ± 0.0032	0.6942 ± 0.0028	0.4492 ± 0.0009
Bi-LSTM [11]	0.5933 ± 0.0107	0.5181 ± 0.0162	0.6847 ± 0.0051	0.4494 ± 0.0020
CNN [19]	0.6380 ± 0.0043	0.5085 ± 0.0158	0.6787 ± 0.0025	0.4557 ± 0.0016
LINN-R_l	*0.8353 ± 0.0043*	*0.3880 ± 0.0069*	*0.9173 ± 0.0042*	*0.2733 ± 0.0065*
LINN	**0.9440 ± 0.0064***	**0.2318 ± 0.0124***	**0.9559 ± 0.0006***	**0.2081 ± 0.0018***

LINN outperforms traditional (non-logical) neural networks.
RNN/LSTM/CNN does not model the compositional logical structure.
Logical regularization is important.

RUTGERS

Application 1: Solving Logical Equations

- t-SNE visualization of logical variable embeddings
  - LINN can finally separate the True and False variables



Accuracy of variable solving: 96%

We can use machine learning to (approximately) solve NP-complete problems

Agnostic to small errors and noise in data.

Application 2: Explainable Recommendation

Neural Logic Reasoning for Explainable Recommendation

- Logic expressions help to model item relationships in recommendation
  - Complimentary: iPhone ∧ iPhone case = T
  - Substitutive: (Coke ∧ ¬ Pepsi) ∨ (¬ Coke ∧ Pepsi) = T
  - Irrelevant: iPhone ∧ Android data line = F.

- User's interaction history can be represented as logical expressions
  - Suppose user purchased item v_3 after several history interactions { v_1 = T (likes), v_2 = F (dislikes) }
  - Training example: $(v_1 \wedge v_3) \vee (\neg v_2 \wedge v_3) \vee (v_1 \wedge \neg v_2 \wedge v_3) = T$
  - This is a noisy reasoning problem: different users' equation may conflict

- Pair-wise Contrastive Ranking Loss

$$e^+ = (\cdot \wedge v^+) \vee \cdots \vee (\cdot \wedge v^+)$$
$$e^- = (\cdot \wedge v^-) \vee \cdots \vee (\cdot \wedge v^-)$$

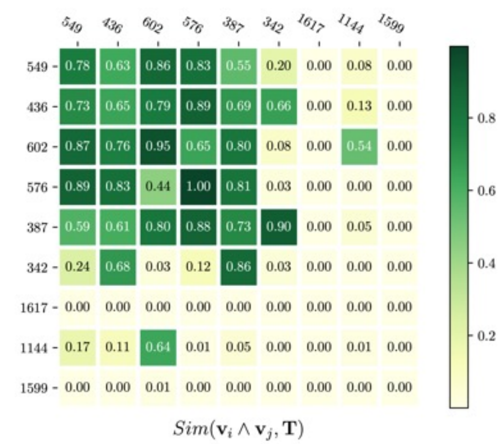
$$L = -\sum_{e^+} \log\left(sigmoid(p(e^+) - p(e^-))\right) + \lambda_l \sum_i r_i + \lambda_\ell \sum_{w \in W} \|\mathbf{w}\|_F^2 + \lambda_\Theta \|\Theta\|_F^2$$

Application 2: Explainable Recommendation

- Recommendation Performance
  - LINN makes significant improvements on Movie and E-commerce recommendation

	ML-100k		Amazon Electronics	
	nDCG@10	Hit@1	nDCG@10	Hit@1
BPRMF [31]	0.3664 ± 0.0017	0.1537 ± 0.0036	0.3514 ± 0.0002	0.1951 ± 0.0004
SVD++ [21]	0.3675 ± 0.0024	0.1556 ± 0.0044	0.3582 ± 0.0004	0.1930 ± 0.0006
STAMP [25]	0.3943 ± 0.0016	0.1706 ± 0.0022	0.3954 ± 0.0003	0.2215 ± 0.0003
GRU4Rec [16]	0.3973 ± 0.0016	0.1745 ± 0.0038	0.4029 ± 0.0009	0.2262 ± 0.0009
NARM [24]	*0.4022 ± 0.0015*	0.1771 ± 0.0016	0.4051 ± 0.0006	0.2292 ± 0.0005
LINN-R_l	0.4022 ± 0.0027	*0.1783 ± 0.0043*	*0.4152 ± 0.0014*	*0.2396 ± 0.0019*
LINN	**0.4064 ± 0.0015***	**0.1850 ± 0.0053***	**0.4191 ± 0.0012***	**0.2438 ± 0.0014***

- Extracting Explanations for the Recommendations
  - The AND module extracts complimentary item explanations
  - E.g., iPhone ∧ iPhone case = T
  - Explanation: We recommend this iPhone case is because you have purchased an iPhone.



Neural Collaborative Reasoning

- Personalize the Reasoning Process

- Reasoning with Implicit Feedback
  - User u, items $\{v_1, v_2, \ldots, v_r\}$

    Horn Clause: $I(u, v_1) \wedge I(u, v_2) \wedge \cdots \wedge I(u, v_r) \rightarrow I(u, v_x)$

  - $I(u, v_i)$ is an encoding function showing user u interacted with an item
  - $I(u, v_i)$ can be a simple neural network

- Reasoning with Explicit Feedback
  - User u, items $\{v_1, v_2, \ldots, \neg v_r\}$, where $\neg v_r$ represents a user has negative feedback

    Horn Clause: $L(u, v_1) \wedge L(u, v_2) \wedge \cdots \wedge \neg L(u, v_r) \rightarrow L(u, v_x)$

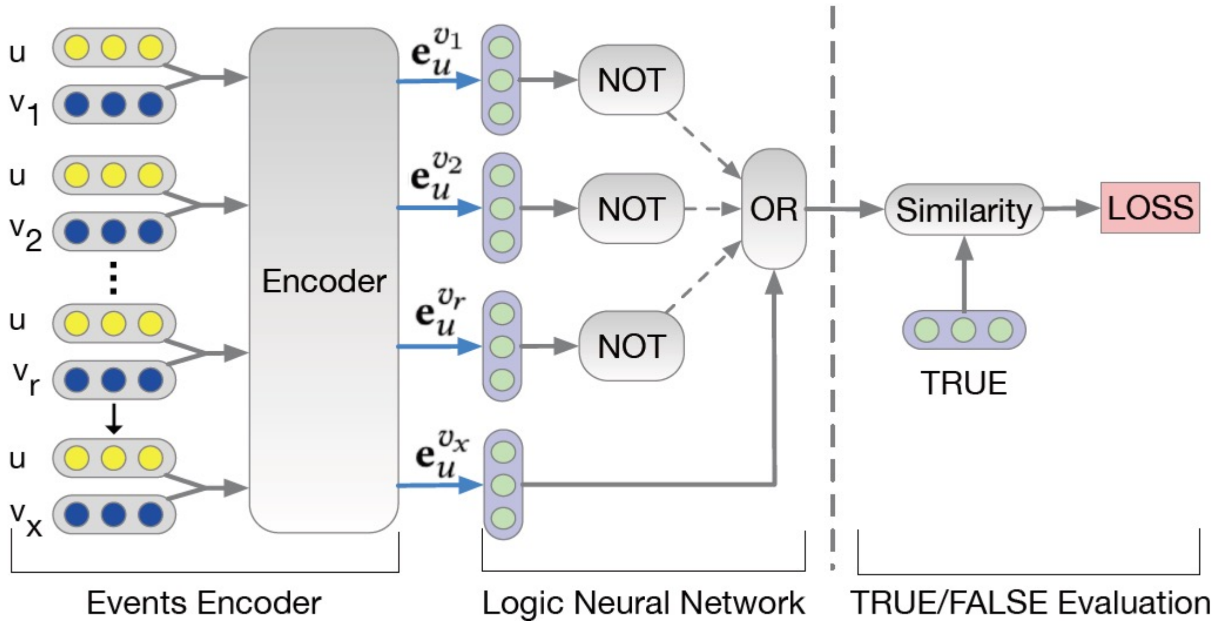
  - $L(u, v_i)$ is an encoding function showing user likes an item

Collaborative Reasoning Architecture

Horn Clause: $I(u, v_1) \wedge I(u, v_2) \wedge \cdots \wedge I(u, v_r) \rightarrow I(u, v_x)$

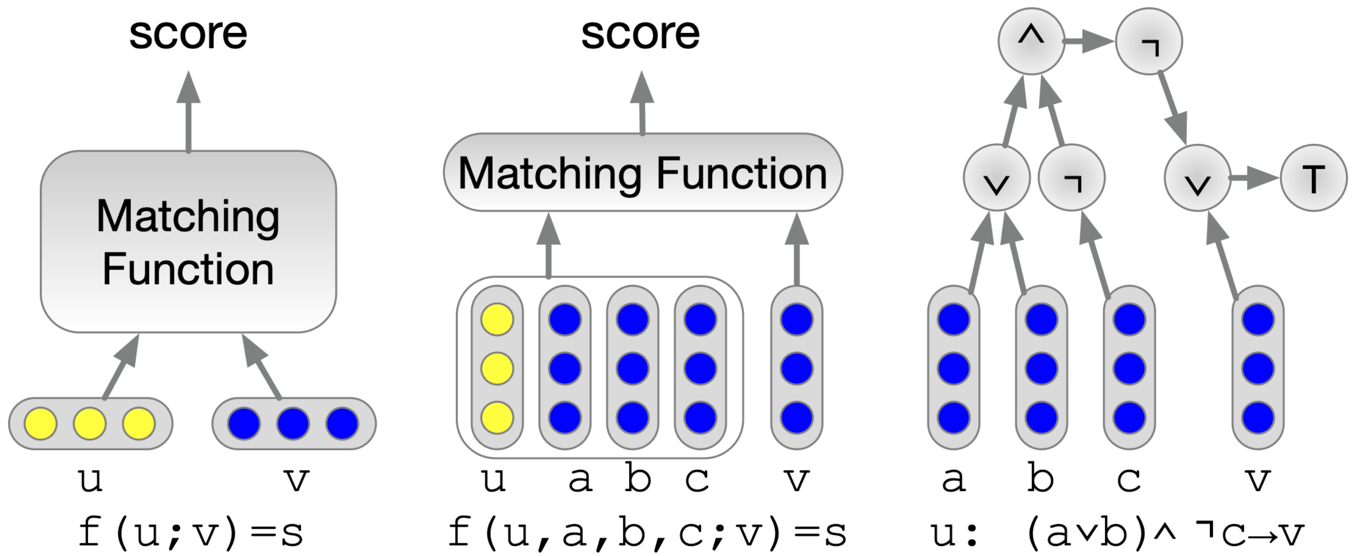
$$e_u^{v_1} \wedge e_u^{v_2} \cdots \wedge e_u^{v_r} \rightarrow e_u^{v_x} \Leftrightarrow \neg e_u^{v_1} \vee \neg e_u^{v_2} \cdots \vee \neg e_u^{v_r} \vee e_u^{v_x}$$

$$p \rightarrow q \Leftrightarrow \neg p \vee q$$



From Learning to Reasoning for AI

- From Perception to Cognition
- From System 1 to System 2

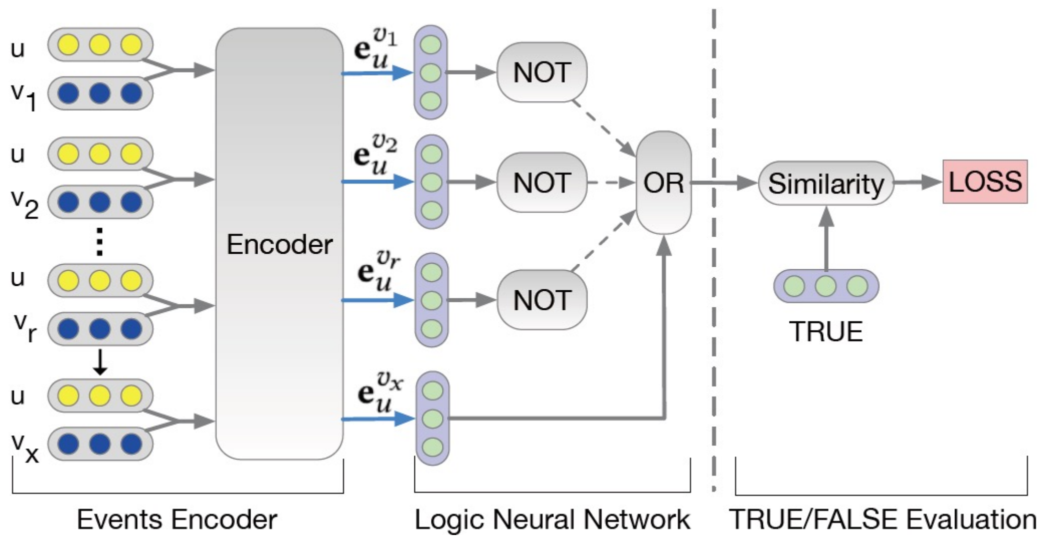


Matching Models Sequential Models Reasoning Models

From Learning to Reasoning

- From System 1 to System 2 for AI

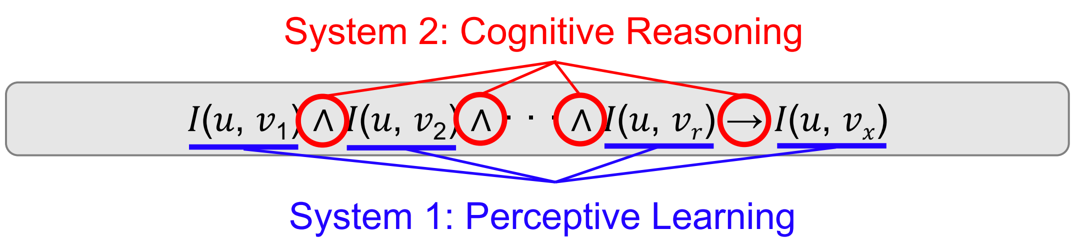
$$I(u, v_1) \wedge I(u, v_2) \wedge \cdots \wedge I(u, v_r) \rightarrow I(u, v_x)$$

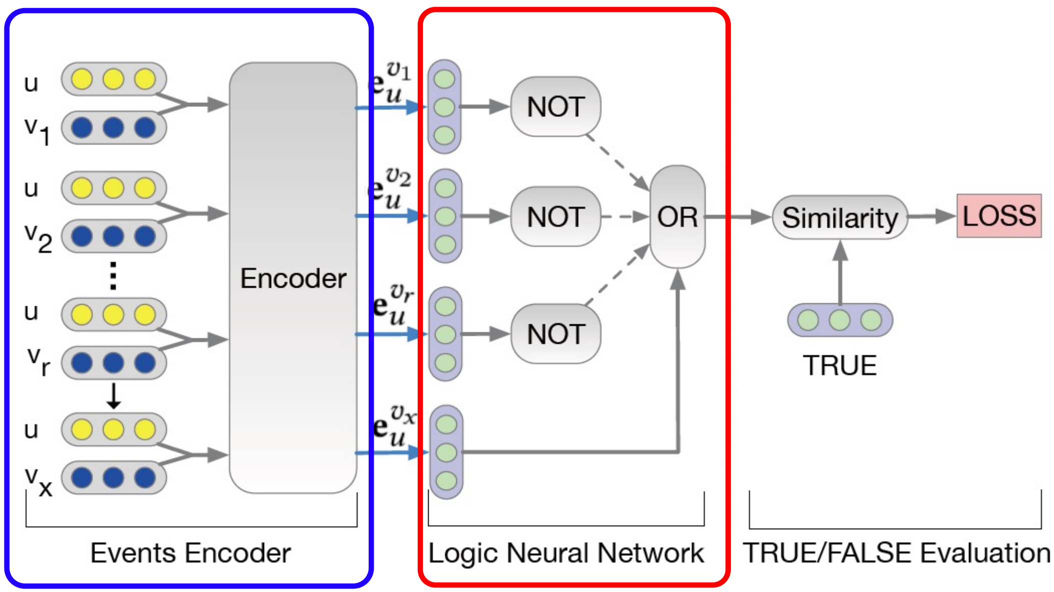


RUTGERS

From Learning to Reasoning

- From System 1 to System 2 for AI





System 1: Perceptive Learning System 2: Cognitive Reasoning

Results

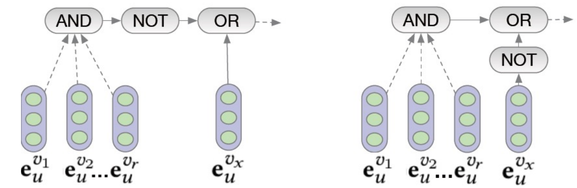
	ML100k				Movies and TV				Electronics			
	N@5	N@10	HR@5	HR@10	N@5	N@10	HR@5	HR@10	N@5	N@10	HR@5	HR@10
BPR-MF	0.3024	0.3659	0.4501	0.6486	0.3962	0.4392	0.5346	0.6676	0.3092	0.3472	0.4179	0.5354
SVD++	0.3087	0.3685	0.4586	0.6433	0.3918	0.4335	0.5224	0.6512	0.2775	0.3172	0.3848	0.5077
DMF	0.3023	0.3661	0.4480	0.6450	0.4006	0.4455	0.5455	0.6843	0.2775	0.3143	0.3783	0.4922
NeuMF	0.3002	0.3592	0.4490	0.6316	0.3791	0.4211	0.5134	0.6429	0.3026	0.3358	0.4031	0.5123
GRU4Rec	0.3564	0.4122	0.5134	0.6856	0.4038	0.4459	0.5287	0.6688	0.3154	0.3551	0.4284	0.5511
STAMP	0.3560	0.4070	0.5159	0.6730	0.3935	0.4366	0.5246	0.6577	0.3095	0.3489	0.4196	0.5430
NLR	0.3602	0.4151	0.5102	0.6795	0.4191	0.4591	0.5506	0.6739	0.3475	0.3852	0.4623	0.5788
NCR-I	0.3697	0.4219	0.5265	0.6890	0.4152	0.4550	0.5479	0.6709	0.3226	0.3604	0.4331	0.5500
NCR-E w/o **LR**	0.3671	0.4219	0.5180	0.6890	0.4126	0.4535	0.5444	0.6705	0.3272	0.3649	0.4377	0.5544
NCR-E	**0.3760****	**0.4240****	**0.5456****	**0.6943****	**0.4255****	**0.4670****	**0.5611****	**0.6891**	**0.3499***	**0.3878***	**0.4639***	**0.5812***
Improvment[1]	5.50%	2.86%	5.76%	1.27%	5.37%	4.73%	2.86%	0.70%	10.94%	9.21%	8.29%	5.46%
Improvment[2]	4.39%	2.14%	6.71%	2.66%	1.53%	1.72%	1.91%	2.26%	0.69%	0.67%	0.35%	0.41%

- NCR-I: Reasoning with Implicit Feedback
- NCR-E: Reasoning with Explicit Feedback

- Model is Partly Explainable

The Importance of Causal-Consistent Reasoning

- EqModel (causally consistent):
  - $\boldsymbol{e}_u^{v_1} \wedge \boldsymbol{e}_u^{v_2} \cdots \wedge \boldsymbol{e}_u^{v_r} \rightarrow \boldsymbol{e}_u^{v_x} \Leftrightarrow \neg \boldsymbol{e}_u^{v_1} \vee \neg \boldsymbol{e}_u^{v_2} \cdots \vee \neg \boldsymbol{e}_u^{v_r} \vee \boldsymbol{e}_u^{v_x}$ (1)
  - $\boldsymbol{e}_u^{v_1} \wedge \boldsymbol{e}_u^{v_2} \cdots \wedge \boldsymbol{e}_u^{v_r} \rightarrow \boldsymbol{e}_u^{v_x} \Leftrightarrow \neg (\boldsymbol{e}_u^{v_1} \wedge \boldsymbol{e}_u^{v_2} \cdots \wedge \boldsymbol{e}_u^{v_r}) \vee \boldsymbol{e}_u^{v_x}$ (2)
- CMPModel (causally inconsistent):
  - $\boldsymbol{e}_u^{v_x} \rightarrow \boldsymbol{e}_u^{v_1} \wedge \boldsymbol{e}_u^{v_2} \cdots \wedge \boldsymbol{e}_u^{v_r} \Leftrightarrow \neg \boldsymbol{e}_u^{v_x} \vee (\boldsymbol{e}_u^{v_1} \wedge \boldsymbol{e}_u^{v_2} \cdots \wedge \boldsymbol{e}_u^{v_r})$ (3)



(a) EqModel (b) CMPModel

	ML100k				Movies and TV				Electronics			
	N@5	N@10	HR@5	HR@10	N@5	N@10	HR@5	HR@10	N@5	N@10	HR@5	HR@10
GRU4Rec	0.3564	0.4122	0.5134	0.6856	0.4038	0.4459	0.5287	0.6688	0.3154	0.3551	0.4284	0.5511
NLR	0.3529	0.4066	0.5113	0.6763	0.4191	0.4591	0.5506	0.6739	0.3475	0.3852	0.4623	0.5788
[1]EqModel	0.3664	0.4224	0.5318	**0.7070**	0.4105	0.4521	0.5429	0.6686	0.3249	0.3626	0.4355	0.5518
[2]CMPModel	0.3551	0.4144	0.5106	0.6932	0.4100	0.4506	0.5417	0.6670	0.3165	0.3541	0.4252	0.5416
[3]**NCR-E**	**0.3760**	**0.4240**	**0.5456**	0.6943	**0.4255**	**0.4670**	**0.5611**	**0.6891**	**0.3499**	**0.3878**	**0.4639**	**0.5812**
p-value[1,3]	0.0825	0.0606	0.1073	0.0547	0.0156*	0.0230*	0.0212*	0.0197*	0.0015*	0.0021*	0.0010*	0.0009*
p-value[2,3]	0.0099*	0.0250*	0.0258*	0.4668	0.0108*	0.0103*	0.0057*	0.0048*	0.0022*	0.0019*	0.0023*	0.0018*

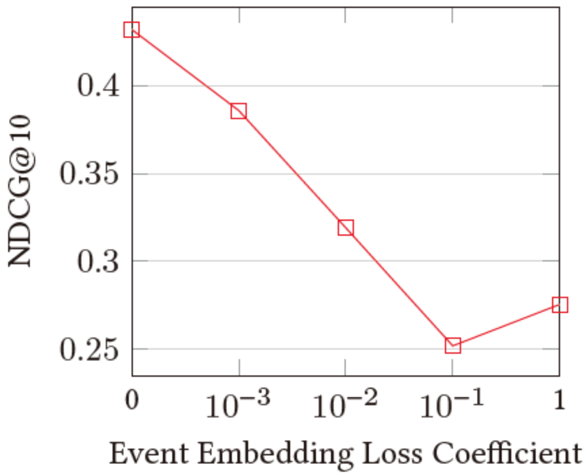
Causally consistent models are comparable
Causally consistent models are better than causally inconsistent models

The Importance of Neural-Symbolic Reasoning (compared with Pure-Symbolic Reasoning)

- Boolean logic constraint:

$$\mathcal{L}_{event} = \sum_{u} \sum_{v \in \mathcal{V}_u^+} \mathrm{MSE}(\mathbf{e}_u^v, \mathbf{G})$$

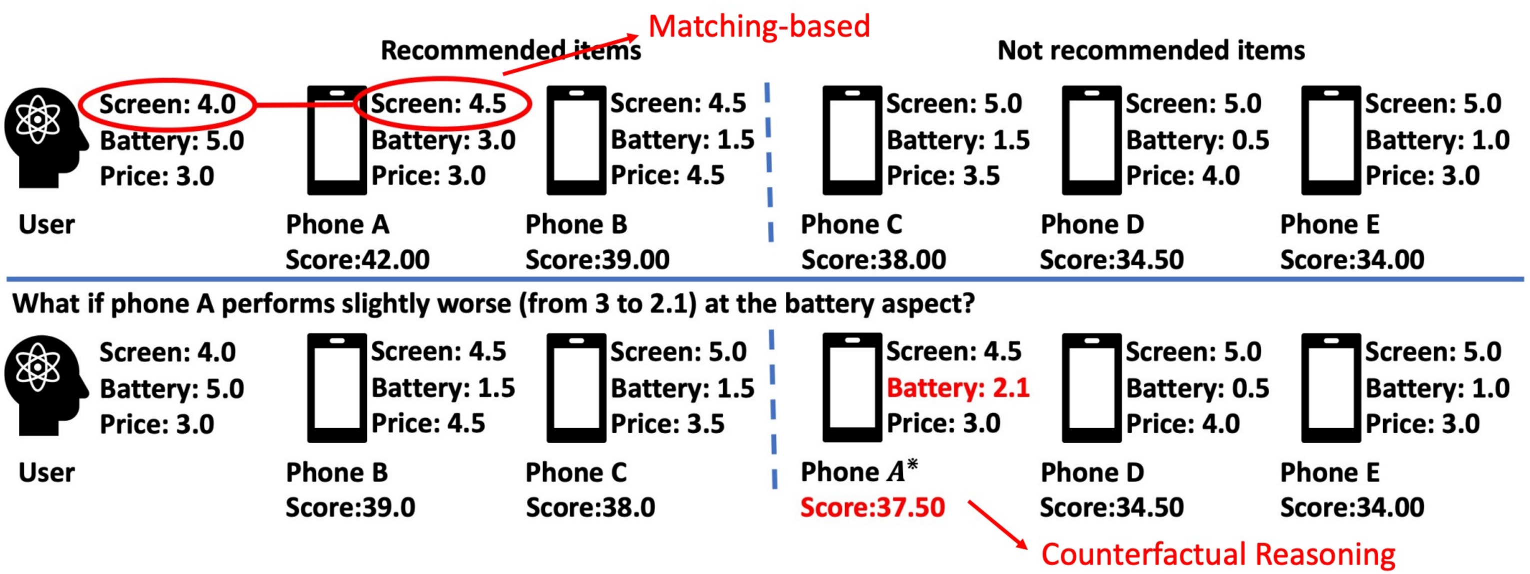
  - G is for ground-truth vector, which is either **T** or **F**;
  - MSE() is mean square error.



- Neural-Symbolic Reasoning is better than Pure Boolean Logic Reasoning
  - We leverage both Learning and Reasoning abilities

Counterfactual Explanations

- Associative vs. Causal/Counterfactual Reasoning



Counterfactual Explanation:

If the item had been slightly worse on [aspect(s)], then it would not have been recommended.

[9] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen and Yongfeng Zhang. “Counterfactual Explainable Recommendation”, CIKM 2021.
[10] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li and Yongfeng Zhang. “Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning”, WWW2022.

Simple and Effective Explanations

- Occam's Razor Principle
  - If two explanations are equally effective in explaining the results, we prefer the simpler explanation than the complex one.

- To character Simpleness
  - Explanation Complexity $C(\Delta) = \gamma||\Delta||_0 + ||\Delta||_2^2$

How many aspects are used to generate explanations

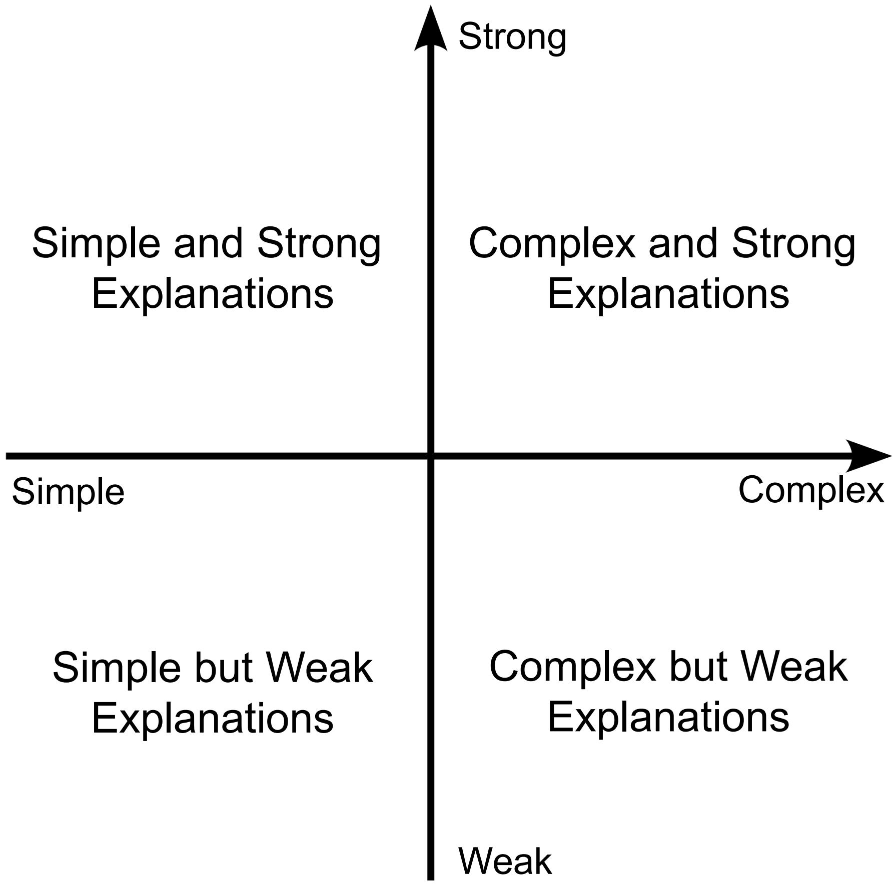
How many changes need to be applied on these aspects

- To character Effectiveness
  - Explanation Strength $S(\Delta) = \underline{s_{i,j} - s_{i,j_\Delta}}$

The decrease of V_j's ranking score in user U_i's recommendation list after applying Δ

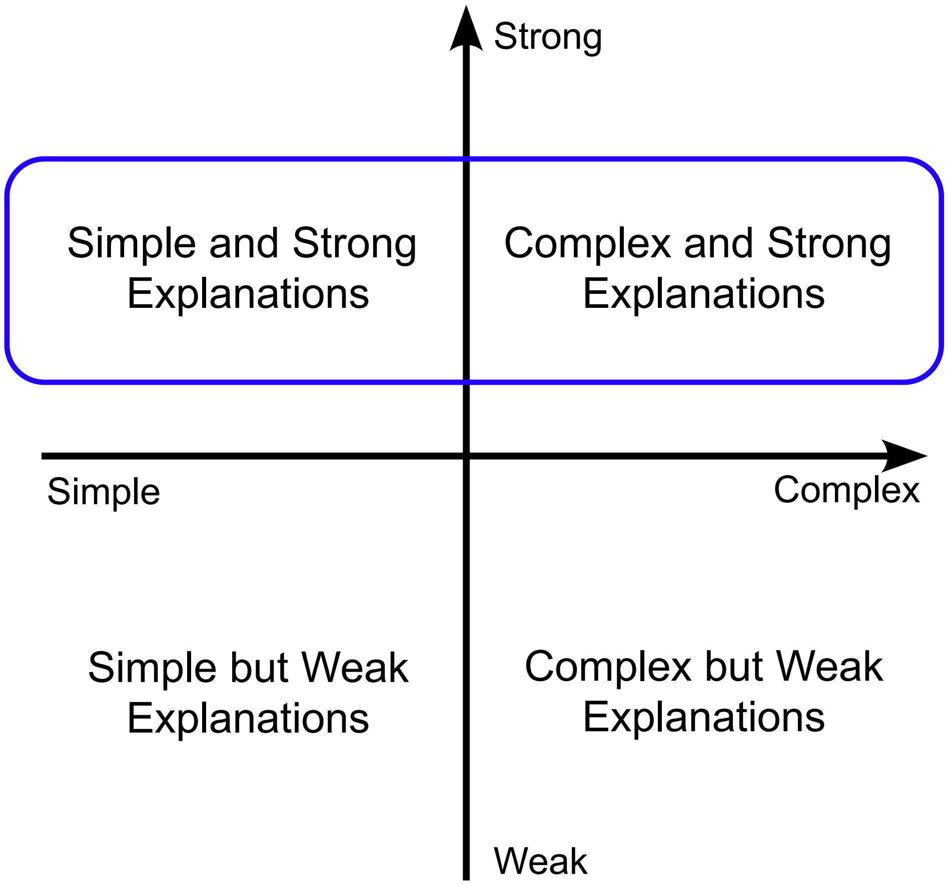
Complexity vs. Strength

- Two orthogonal concepts



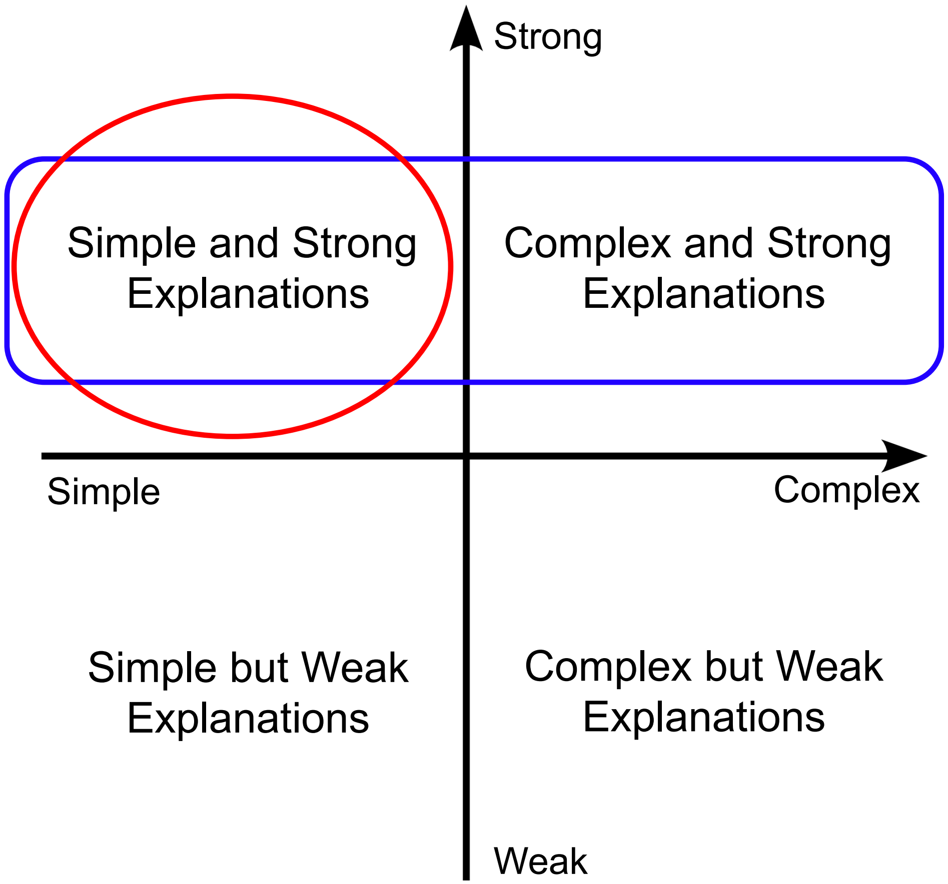
Complexity vs. Strength

- Two orthogonal concepts



Complexity vs. Strength

- Two orthogonal concepts



Counterfactual Learning and Reasoning

- Seek simple and effective explanations

$$\begin{aligned} &\text{minimize} \;\; \text{Explanation Complexity} \\ &\quad \text{s.t.,} \;\; \text{Explanation is Strong Enough} \end{aligned} \quad \Rightarrow \quad \begin{aligned} &\text{minimize} \;\; C(\Delta) = \|\Delta\|_2^2 + \gamma\|\Delta\|_0 \\ &\quad \text{s.t.,} \;\; S(\Delta) = s_{i,j} - s_{i,j_\Delta} \geq \epsilon \end{aligned}$$

  - Idea: Find minimal changes to an item's features so that the item can be kicked out of the recommendation list

- Related Optimization for model learning

$$\underset{\Delta}{\text{minimize}} \; \|\Delta\|_2^2 + \gamma\|\Delta\|_1 + \lambda L(s_{i,j_\Delta}, s_{i,j_{K+1}})$$

$$\text{where } s_{i,j_\Delta} = f(X_i, Y_j + \Delta \mid Z, \Theta), \;\; s_{i,j_{K+1}} = f(X_i, Y_{j_{K+1}} \mid Z, \Theta)$$

$$L(s_{i,j_\Delta}, s_{i,j_{K+1}}) = \max(0, \alpha + s_{i,j_\Delta} - s_{i,j_{K+1}})$$

Sufficiency and Necessity of Explanations

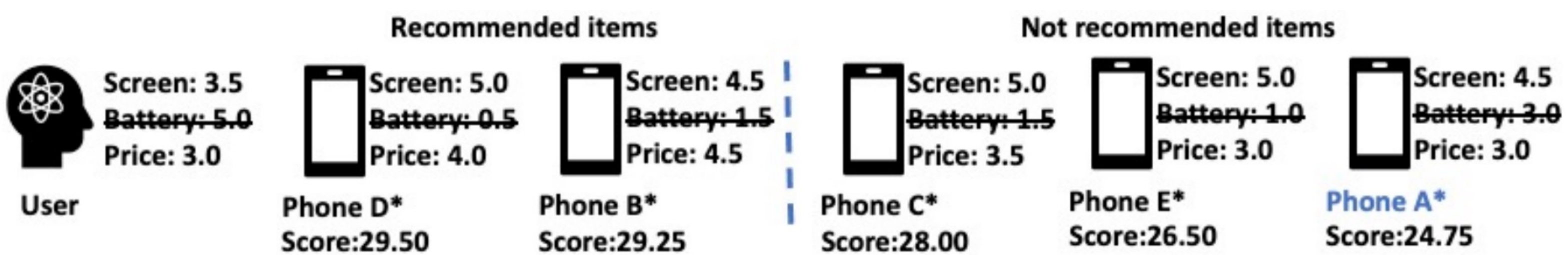
- $S \Rightarrow N$: S is a sufficient condition for N
- $\neg N \Rightarrow \neg S$: N is a necessary condition for S

Sufficiency and Necessity of Explanations

- S ⇒ N: S is a sufficient condition for N
- ¬N ⇒ ¬S: N is a necessary condition for S

CountER: If this phone had been slightly worse on [Battery], then it will not be recommended.

- Probability of Necessity (PN): If in a counterfactual world, the aspects in the explanation did not exist in the system, then what is the probability that the item would not be recommended.



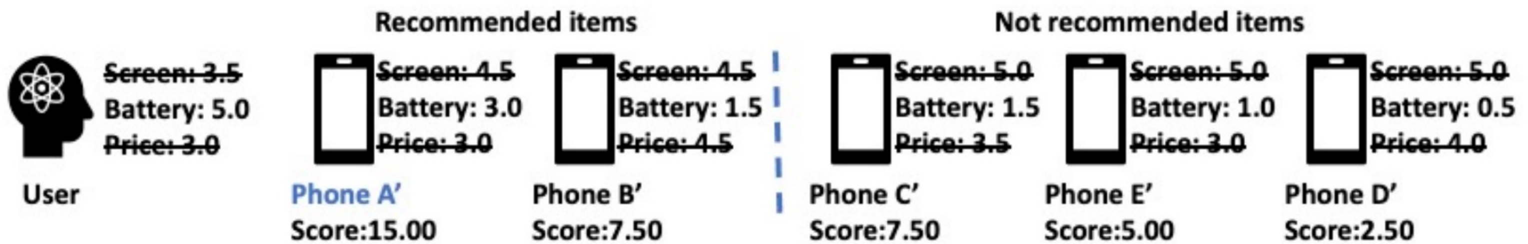
$$\mathrm{PN} = \frac{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} \mathrm{PN}_{ij}}{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} I(\mathcal{A}_{ij} \neq \emptyset)}, \text{ where } \mathrm{PN}_{ij} = \begin{cases} 1, & \text{if } v_j^* \notin R_{i,K}^* \\ 0, & \text{else} \end{cases}$$

Sufficiency and Necessity of Explanations

- S ⇒ N: S is a sufficient condition for N
- ¬N ⇒ ¬S: N is a necessary condition for S

CountER: If this phone had been slightly worse on [Battery], then it will not be recommended.

- Probability of Sufficiency (PS): If in a counterfactual world, the aspects in the explanation were the only aspects existed in the system, then what is the probability that the item would still be recommended.



$$\mathrm{PS} = \frac{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} \mathrm{PS}_{ij}}{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} I(\mathcal{A}_{ij} \neq \emptyset)}, \text{ where } \mathrm{PS}_{ij} = \begin{cases} 1, & \text{if } v'_j \in R'_{i,K} \\ 0, & \text{else} \end{cases}$$

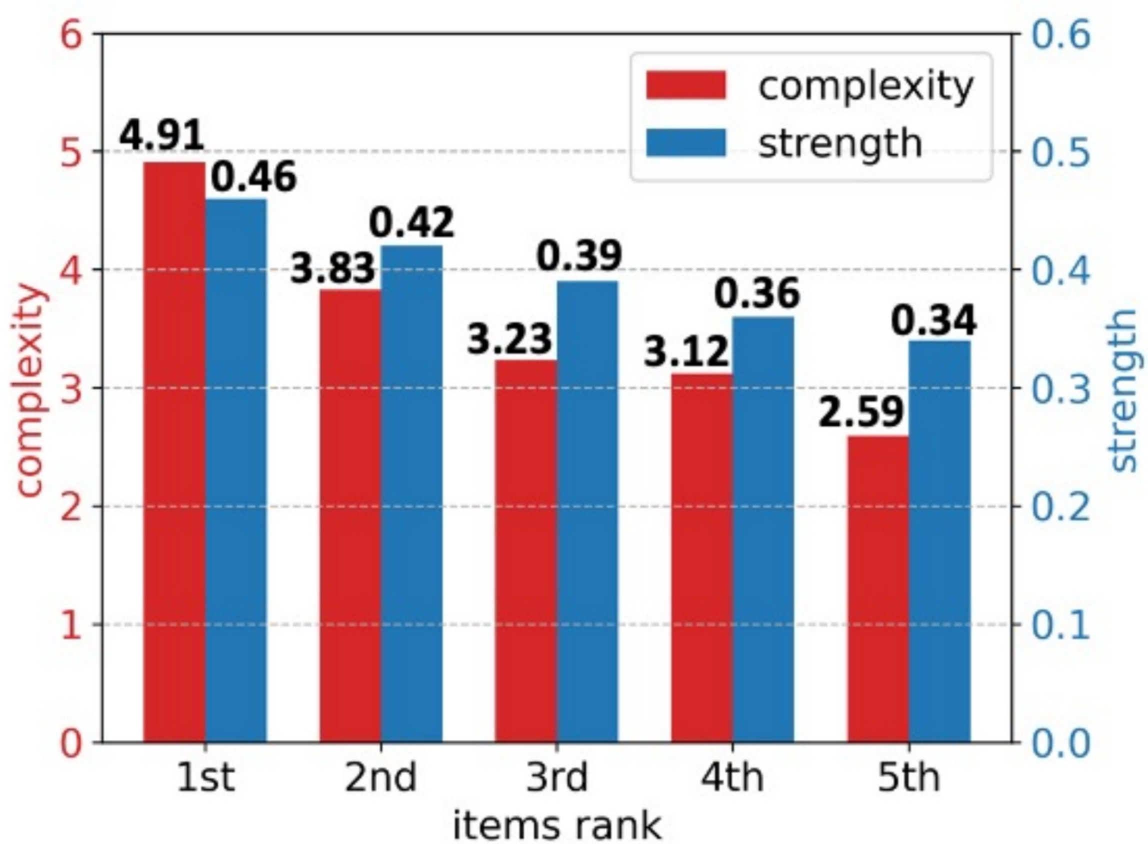
RUTGERS

Counterfactual Reasoning gives Better Explanations

	Single Aspect Explanation														
	Electronic			Cell Phones			Kindle Store			CDs and Vinyl			Yelp		
	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%
Random	2.05	2.10	2.07	3.39	3.50	3.44	3.16	2.75	2.94	1.58	2.03	1.78	7.52	10.68	8.82
EFM[50]	8.41	41.13	13.96	32.31	**82.09**	46.37	6.01	**73.84**	11.12	10.15	42.63	16.39	5.87	61.06	10.71
A2CF[9]	41.45	**77.60**	54.03	36.82	78.68	50.17	25.66	65.53	36.88	25.41	**84.51**	39.07	17.59	**96.92**	29.78
CountER	**65.54**	68.28	**66.83**	**74.03**	63.30	**68.25**	34.37	41.50	37.60	49.62	54.72	52.04	**65.26**	53.25	**58.64**
CountER (w/ mask)	56.73	62.03	59.26	70.11	54.71	61.46	**35.39**	46.91	**40.34**	**75.17**	49.18	**59.46**	58.52	52.56	55.38
	Multiple Aspect Explanation														
	Electronic			Cell Phones			Kindle Store			CDs and Vinyl			Yelp		
	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%
Random	2.24	4.90	3.08	6.25	10.13	7.73	5.80	7.80	6.65	3.22	7.65	4.53	13.84	12.92	13.36
EFM[50]	29.65	84.67	43.92	52.66	87.98	65.88	51.72	**96.42**	67.33	47.65	87.35	61.66	16.76	81.68	27.81
A2CF[9]	59.47	81.66	68.82	56.45	80.97	66.52	52.48	87.59	65.64	49.12	**91.52**	63.93	41.38	**98.28**	58.24
CountER	**97.08**	**96.24**	**96.66**	**99.52**	**98.48**	**99.00**	**64.00**	79.20	**70.79**	**80.89**	88.60	**84.57**	**99.91**	94.12	**96.93**
CountER (w/ mask)	77.96	89.26	83.23	86.62	91.78	89.13	60.70	80.10	69.06	72.47	67.72	70.01	96.73	94.39	95.55

Interesting Observations

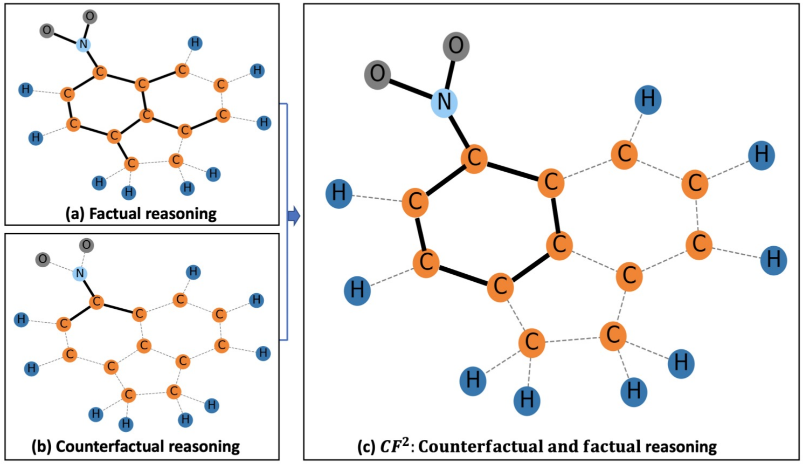
- Top-ranked items need to be backed by stronger and more complex explanations



PN & PS based Evaluation is Usable

- PN/PS metrics are highly correlated with ground-truth based metrics

$$F_{NS} = \frac{2 \cdot \mathrm{PN} \cdot \mathrm{PS}}{\mathrm{PN} + \mathrm{PS}}$$



(a) Factual reasoning

(b) Counterfactual reasoning

(c) CF^2: Counterfactual and factual reasoning

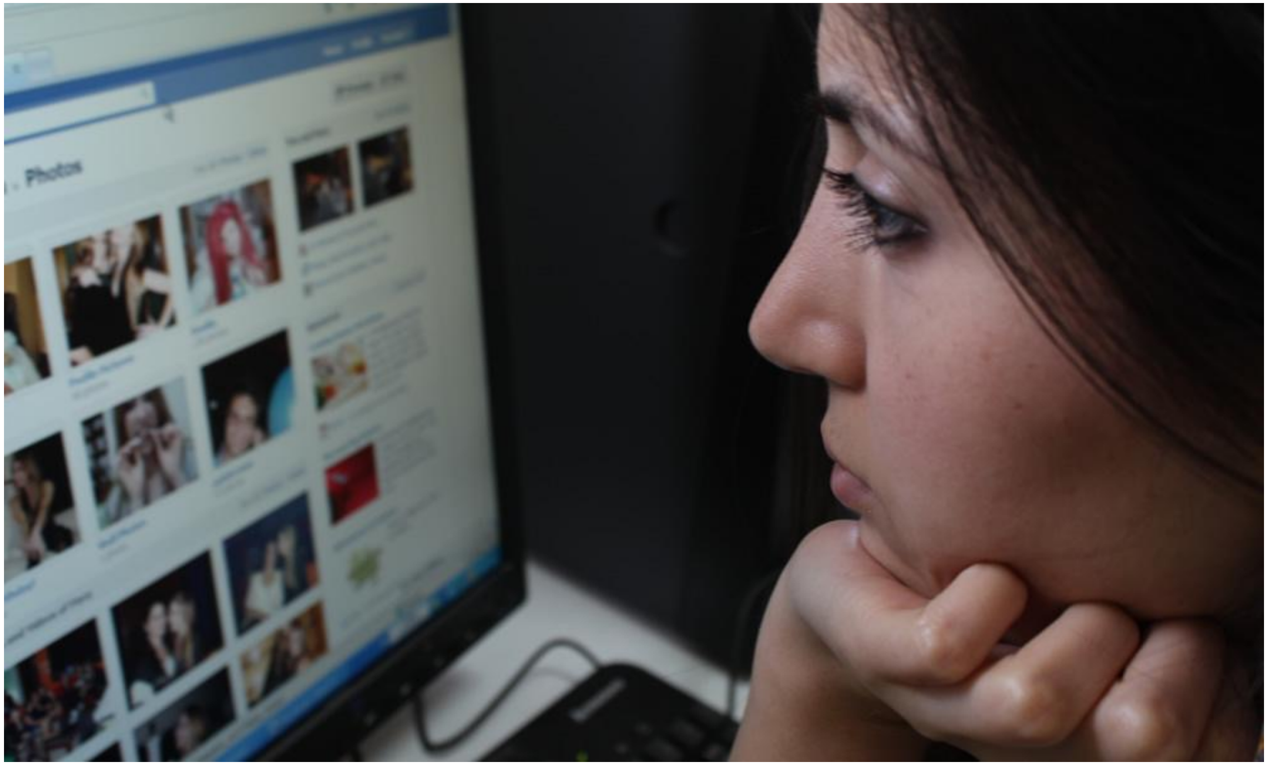
Kendall's τ and Spearman's ρ correlation

Table 7: Correlation between PN/PS-based evaluation and ground-truth evaluation.

Models	BA-Shapes		Tree-Cycles		Mutag_0	
	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$
F_{NS} & F_1	1.00	1.00	1.00	1.00	1.00	1.00
F_{NS} & Acc	0.66	0.79	1.00	1.00	0.66	0.79

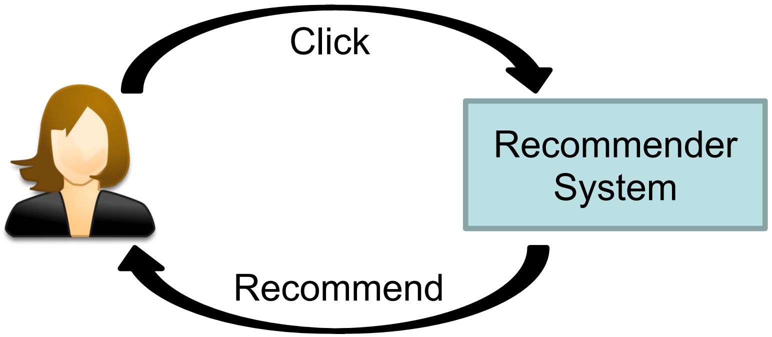
Towards User Controllable Recommender Systems

- Users almost have no control of their recommender system
  - They can only passively receive recommendations



Towards User Controllable Recommender Systems

- Users almost have no control of their recommender system
  - They can only passively receive recommendations
- This causes many problems, e.g., echo chamber

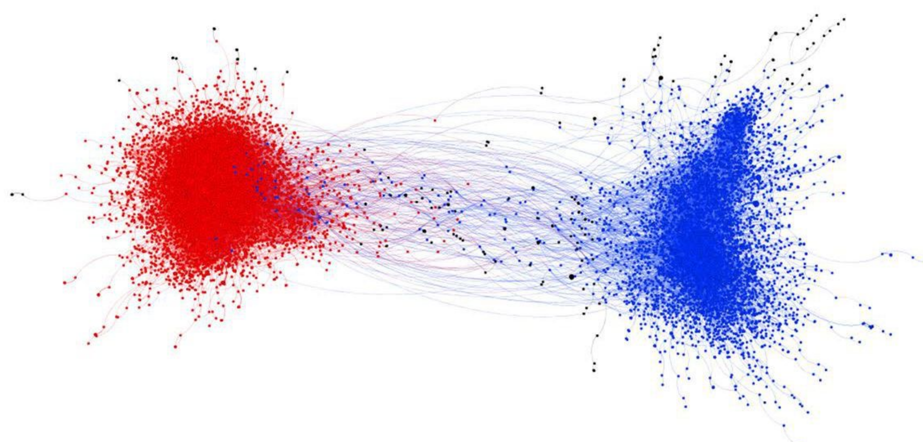


The more you like something, the more RS will recommend similar things, and thus you like them even more.

[11] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, Yongfeng Zhang. Understanding Echo Chambers in E-commerce Recommender Systems. SIGIR'20.

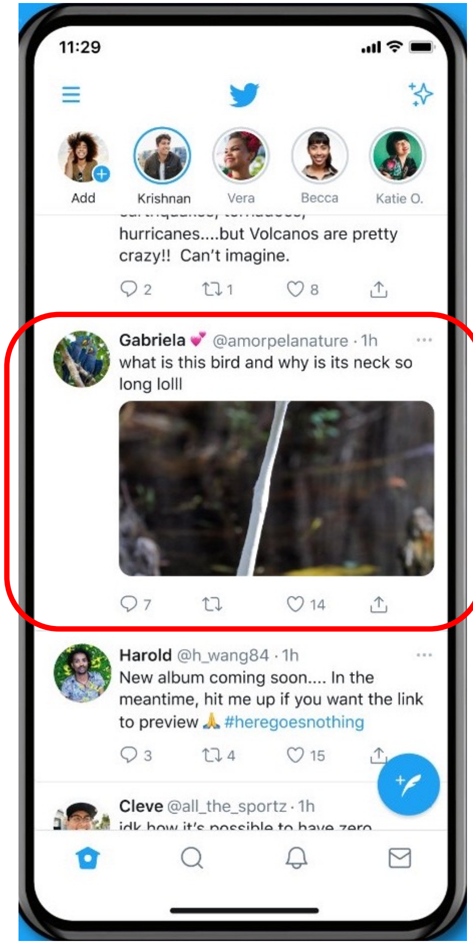
Towards User Controllable Recommender Systems

- Users almost have no control of their recommender system
  - They can only passively receive recommendations
- The Social Echo Chamber
  - Makes all your connections like-minded persons as you
  - Makes all your news feed similar as what you already liked
  - Makes it difficult to explore new ideas and opinions different from yours
  - May even reinforce people’s extreme ideas



RUTGERS

User Control based on Counterfactual Explanations



Counterfactual Retrospective Explanation:

We recommend this video X because you previously liked videos A and B, if you didn't like them, then we would not have recommended this video X.

Counterfactual Prospective Explanation:

If you click "like" on this newly recommended video X, then we will recommend videos such as D and E in the future.

Help users know the consequences of their behaviors so that they can take informed actions.
Users can control their recommendation by invoking or revoking certain actions.

Bridging Explainability and Fairness

- Counterfactual Explanation is a flexible framework
  - As long as the explanation target can be quantified, counterfactual framework can explain it
  - How changes in the input influences the output
- Explainable Fairness is important in Recommendation
  - Hundreds, thousands or even more features
  - System designers:
    - Want to know which feature(s) cause unfairness
  - Users:
    - Want to know how to intervene unfair results to make it more fair

[12] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li and Yongfeng Zhang. Explainable Fairness in Recommendation. SIGIR'22.

Counterfactual Explainable Fairness

- Too many features in RecSys, manually analysis is almost impossible
  - Automatic explainable fairness is needed.
  - E.g., top-5 features that lead to exposure unfairness

Method	Feature-based Explanations
Pop-User	food, service, chicken, prices, hour
Pop-Item	food, service, prices, visit, hour
EFM-User	store, patio, dishes, dish, rice
EFM-Item	flavor, decor, dishes, inside, cheese
SV	server, size, pizza, food, restaurant
CEF	meal, cheese, dish, chicken, taste

Table 5: Top-5 feature-based explanations on Yelp dataset.

Counterfactual Explainable Fairness

- Counterfactual Explainable Fairness framework

$$\text{min. Explanation Complexity}$$
$$s.t., \text{Model Unfairness } \leq \delta$$

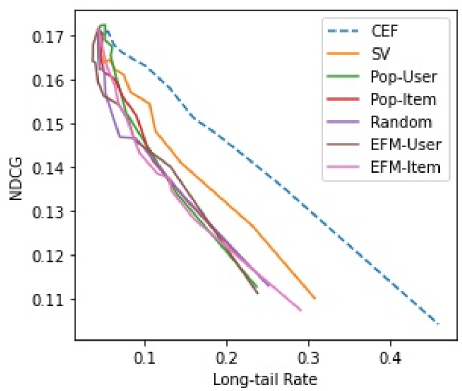
$$\min \|\Psi^{cf}\|_2^2 + \lambda\|\Delta\|_2$$

$$\Psi_{DP} = |\mathcal{G}_1| \cdot \text{Exposure}\,(\mathcal{G}_0|\mathcal{R}_K) - |\mathcal{G}_0| \cdot \text{Exposure}\,(\mathcal{G}_1|\mathcal{R}_K)$$

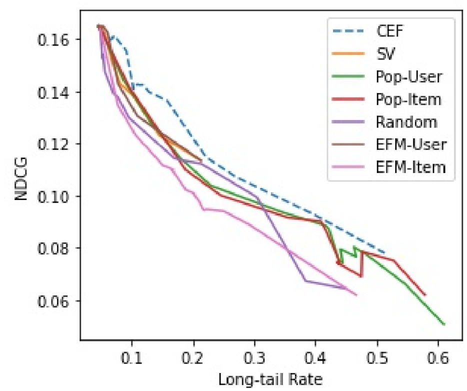
Fairness definition: equal opportunity fairness
Can be any other definition

Counterfactual Explainable Fairness

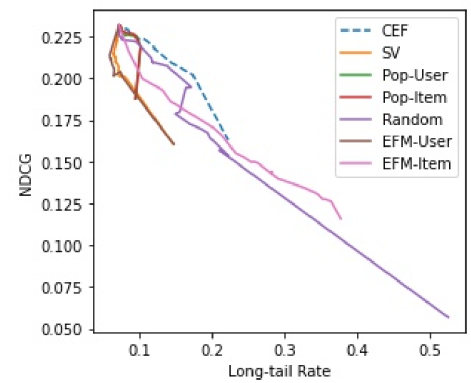
- Better Fairness-Utility Trade-off



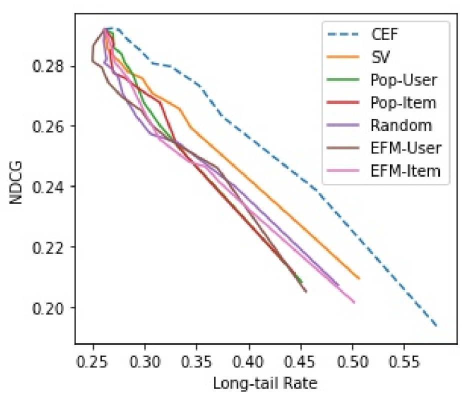
(a) NDCG@5 vs Long-tail Rate@5 on Yelp



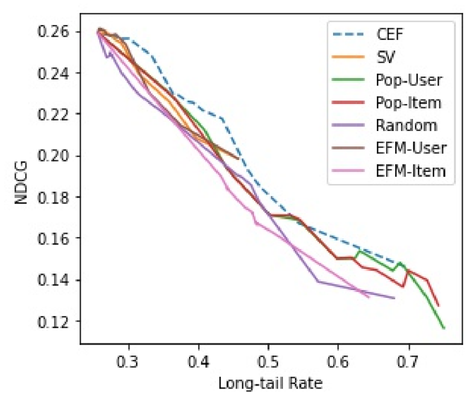
(b) NDCG@5 vs Long-tail Rate@5 on Electronics



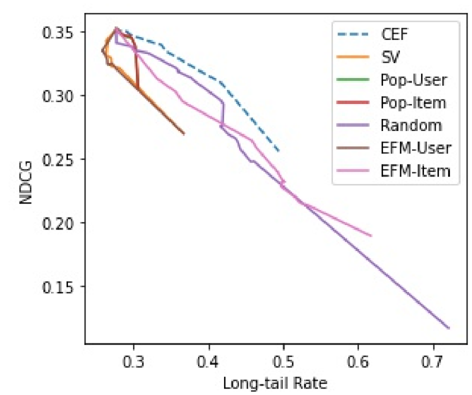
(c) NDCG@5 vs Long-tail Rate@5 on CDs&Vinyl



(d) NDCG@20 vs Long-tail Rate@20 on Yelp



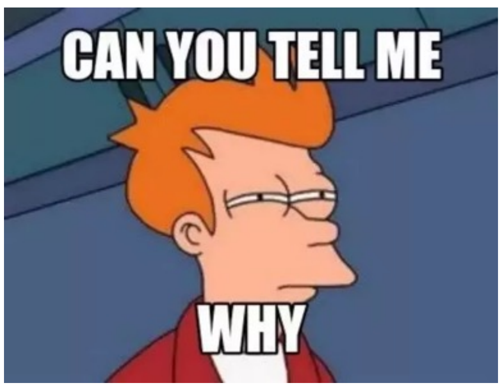
(e) NDCG@20 vs Long-tail Rate@20 on Electronics

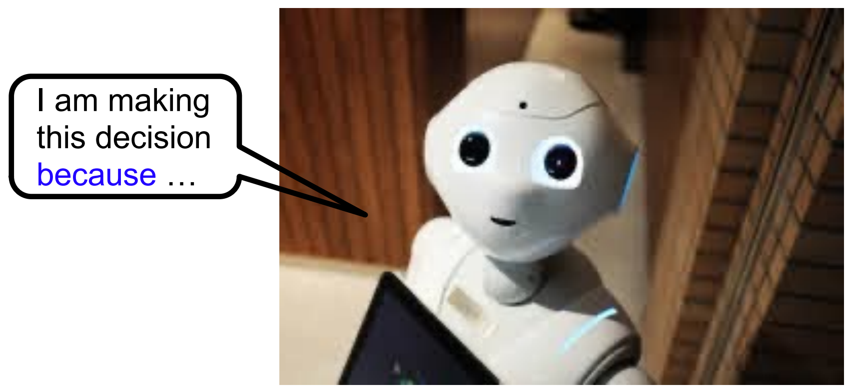


(f) NDCG@20 vs Long-tail Rate@20 on CDs&Vinyl

Natural Language Explanations

- Natural language sentence is the most human-friendly way of explanation
  - Human and machine will inevitably collaborate with each other in future jobs
  - We believe future machines should be able to explain themselves through natural language
  - Better understanding, collaboration and trust between human and machines





Natural Language Explanation in Recommendation

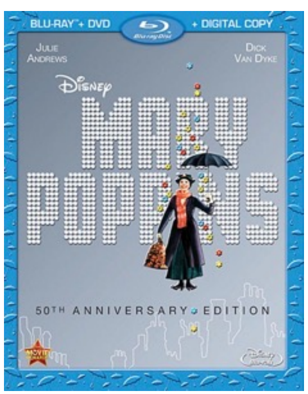
- Explainable Recommendation as Natural Language Generation
  - Recommendation is a very suitable task for developing natural language explanation models
  - High quality ground-truth explanations from humans

Trip Saver

Reviewed in the United States on October 22, 2017

Style: With Lifetime Maps and Traffic (USA) | **Verified Purchase**

Perfect. Lots of features... <u>accurate for finding upcoming restaurants, gas stations and community services.</u> We drive cross country every summer and updated our older GPS to this. <u>Worked through all states, even in low-service areas through the desert.</u> I like being able to search ahead for hotels and restaurants. <u>The battery lasted a long time and there wasn't a lot of screen glare.</u> We also purchased the weighted holder which we really liked.



A classic musical that is still entrancing and fun to watch

Reviewed in the United States on November 7, 2017

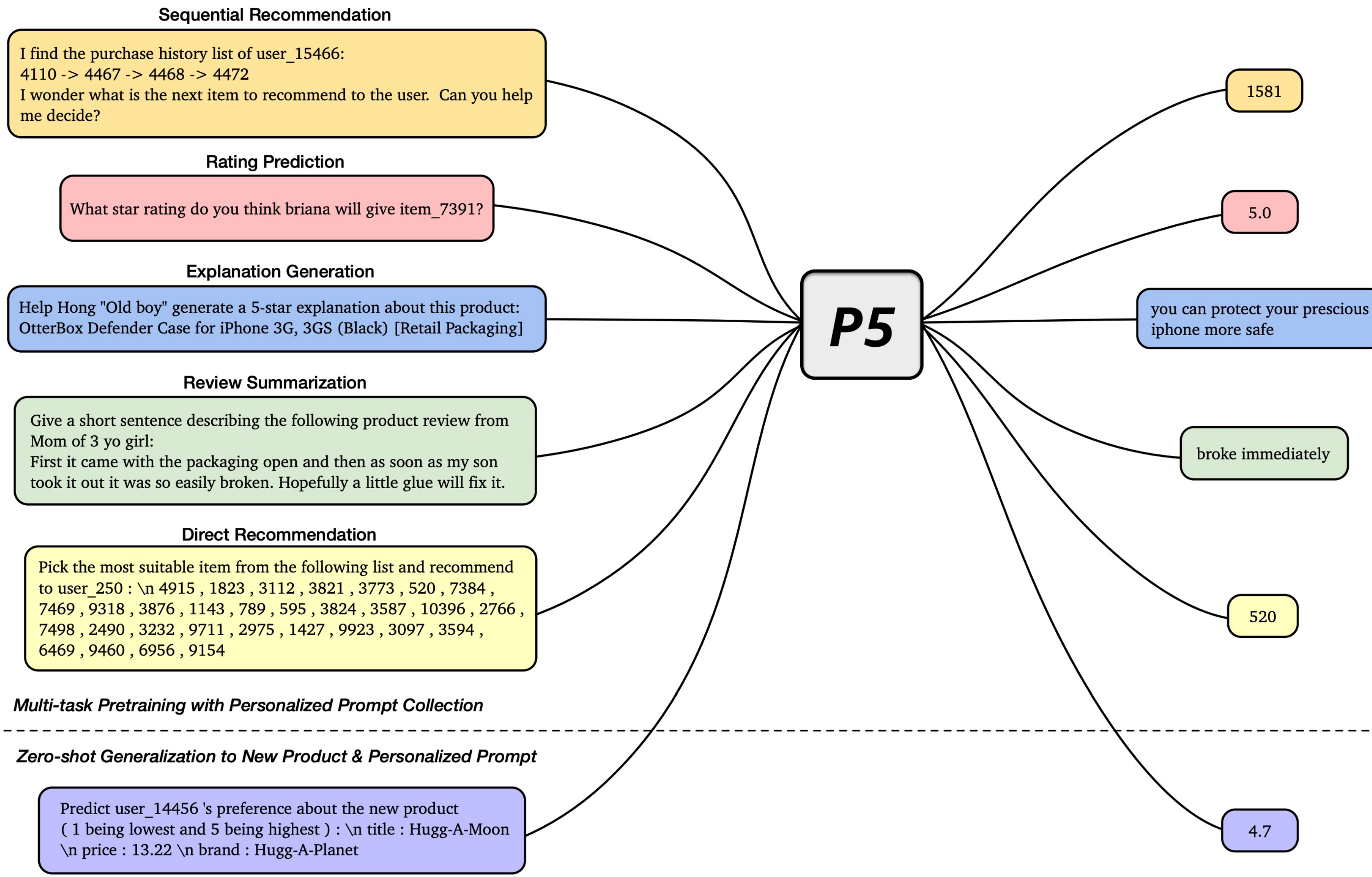
Verified Purchase

<u>The movie holds up well as a glorious musical. The acting, singing, choreography, staging, special effects are all great. The plot still works.</u> This is a movie my wife and I love watching over and over. The blu ray version is beautiful. <u>The quality of the image shows itself during the extreme close-ups of Julie Andrews and Dick Van Dyke -- the images are crystal clear with no blurring on a high quality 53 inch LCD HDTV. The sound is excellent.</u> The DVD authoring is a little idiosyncratic. Though "resume play" is activated but is only present after a lengthy video introduction and it is hard to bypass the "previews." There is a paucity of extras.

22 people found this helpful

https://www.amazon.com/Garmin-Navigator-Lifetime-Directions-Foursquare/dp/B01A1HL9Z6/ref=cm_cr_arp_d_product_top?ie=UTF8
https://www.amazon.com/Mary-Poppins-Anniversary-Julie-Andrews/dp/B004LLDN3A

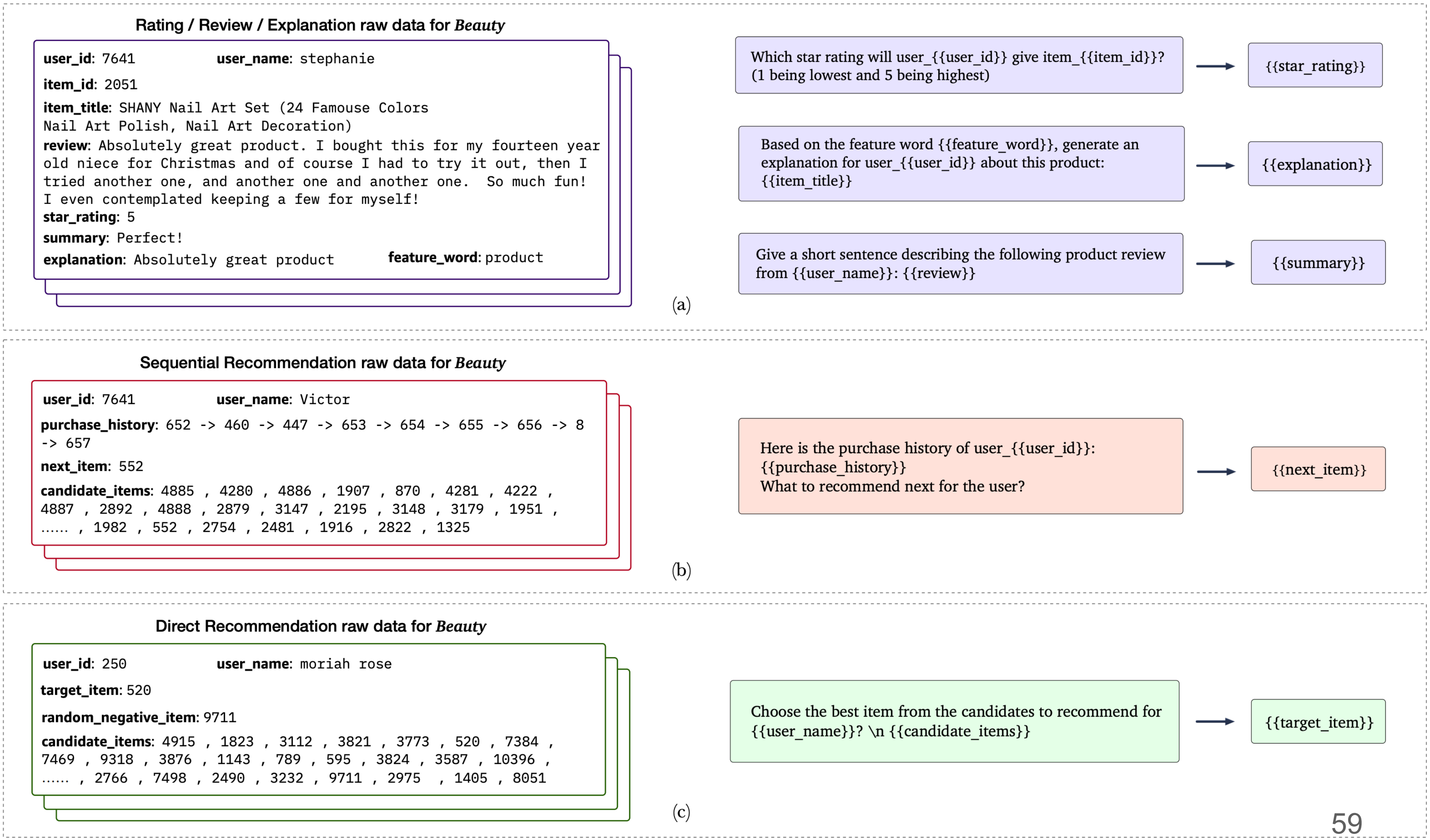
Large Recommendation Models (LRM) for Universal Recommendation Engine



[3] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)." RecSys 2022.

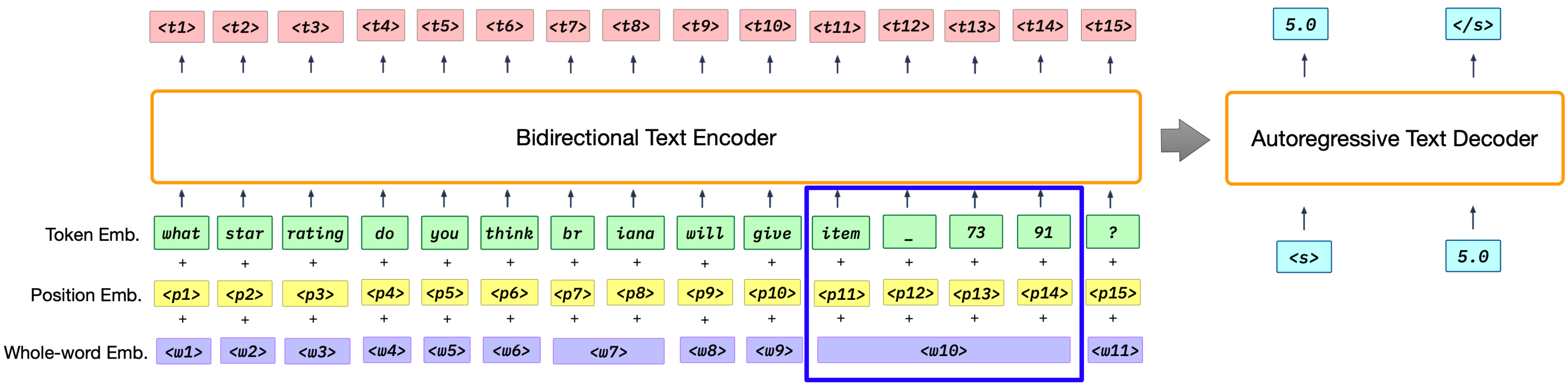
RUTGERS

Pretrain, Personalized Prompt & Predict Paradigm (P5)



The P5 Architecture

- P5 Architecture



ID tokenization is critically important

Keep a constant and manageable amount of tokens

Better Recommendation Accuracy

Table 2: Performance comparison on rating prediction.

Methods	Sports		Beauty		Toys	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
MF	**1.0234**	0.7935	**1.1973**	0.9461	**1.0123**	0.7984
P5-S (1-6)	1.0594	**0.6639**	1.3114	**0.8434**	1.0605	0.7142
P5-B (1-6)	1.0357	0.6813	1.2843	0.8534	1.0866	**0.6957**
P5-S (1-10)	1.0522	0.6698	1.3001	0.8444	1.0805	0.7057
P5-B (1-10)	1.0292	0.6864	1.2862	0.8530	1.0843	0.7007

Table 3: Performance comparison on sequential recommendation.

Methods	Sports				Beauty				Toys			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
Caser	0.0116	0.0072	0.0194	0.0097	0.0205	0.0131	0.0347	0.0176	0.0166	0.0107	0.0270	0.0141
HGN	0.0189	0.0120	0.0313	0.0159	0.0325	0.0206	0.0512	0.0266	0.0321	0.0221	0.0497	0.0277
GRU4Rec	0.0129	0.0086	0.0204	0.0110	0.0164	0.0099	0.0283	0.0137	0.0097	0.0059	0.0176	0.0084
BERT4Rec	0.0115	0.0075	0.0191	0.0099	0.0203	0.0124	0.0347	0.0170	0.0116	0.0071	0.0203	0.0099
FDSA	0.0182	0.0122	0.0288	0.0156	0.0267	0.0163	0.0407	0.0208	0.0228	0.0140	0.0381	0.0189
P5-S (2-3)	0.0272	0.0169	0.0361	0.0198	0.0508	**0.0385**	**0.0668**	**0.0436**	**0.0385**	**0.0269**	**0.0499**	**0.0305**
P5-B (2-3)	0.0364	0.0296	0.0431	0.0318	**0.0515**	0.0381	0.0664	0.0429	0.0363	0.0257	0.0457	0.0287
P5-S (2-13)	0.0258	0.0159	0.0346	0.0188	0.0502	0.0378	0.0656	0.0428	0.0370	0.0260	0.0471	0.0293
P5-B (2-13)	**0.0387**	**0.0312**	**0.0460**	**0.0336**	0.0499	0.0366	0.0651	0.0415	0.0346	0.0244	0.0444	0.0276

RUTGERS

Better Explanation Quality

Table 4: Performance comparison on explanation generation.

Methods	Sports				Beauty				Toys			
	BLUE4	ROUGE1	ROUGE2	ROUGEL	BLUE4	ROUGE1	ROUGE2	ROUGEL	BLUE4	ROUGE1	ROUGE2	ROUGEL
Attn2Seq	0.5305	12.2800	1.2107	9.1312	0.7889	12.6590	1.6820	9.7481	1.6238	13.2245	2.9942	10.7398
NRT	0.4793	11.0723	1.1304	7.6674	0.8295	12.7815	1.8543	9.9477	1.9084	13.5231	3.6708	11.1867
PETER	**0.7112**	12.8944	1.3283	9.8635	1.1541	14.8497	2.1413	11.4143	**1.9861**	14.2716	3.6718	11.7010
P5-S (3-3)	0.5902	**60.8892**	17.7514	18.0010	2.6533	61.6557	21.6574	25.6646	0.3787	**56.7474**	17.1475	16.7914
P5-B (3-3)	0.6213	58.7260	**18.5533**	**18.4670**	**3.1474**	**62.2778**	**21.9762**	**27.1758**	0.5652	56.4732	**17.7930**	**18.3364**
PETER+	2.4627	24.1181	5.1937	18.4105	3.2606	25.5541	5.9668	19.7168	4.7919	28.3083	9.4520	22.7017
P5-S (3-9)	**7.2129**	**67.4004**	**36.1417**	**30.8359**	5.4136	67.9526	36.5097	30.7446	8.2721	69.4591	39.9955	33.6941
P5-B (3-9)	3.5598	64.7683	34.0162	26.3184	6.5551	**68.2939**	36.7586	31.8136	**9.5411**	**69.6964**	**40.3364**	**34.7272**
P5-S (3-12)	5.8446	66.5976	35.5160	29.2766	5.5760	68.1710	**36.7876**	30.8561	7.5790	69.2164	39.9065	33.1177
P5-B (3-12)	4.6977	65.4562	34.9379	27.7223	**7.0183**	68.1908	36.7262	**32.2162**	8.2461	69.2331	39.9456	34.0081

RUTGERS

Zero-Shot Generalization to Items in New Domains

Table 9: Performance on zero-shot domain transfer.

Directions	Z-1 & Z-4	Z-2 & Z-3	Z-5 & Z-7		Z-6	
	Accuracy	MAE	BLUE4	ROUGE1	BLUE4	ROUGE1
Toys -> ***Beauty***	0.7922	0.8244	1.8869	61.1919	5.4609	66.4931
Toys -> ***Sports***	0.8682	0.6644	0.7405	60.9575	2.2601	62.0353
Beauty -> ***Toys***	0.8073	0.7792	0.0929	41.3061	11.8046	64.8701
Beauty -> ***Sports***	0.8676	0.6838	0.0346	39.7191	6.6409	66.9222
Sports -> ***Toys***	0.8230	0.7443	0.0687	42.9310	13.3408	69.7910
Sports -> ***Beauty***	0.8057	0.8102	0.0790	41.0659	13.1690	66.7687

Prompt ID: Z-1

```
Input template: Given the facts about the new product, do you think
user {{user_id}} will like or dislike it? title: {{item_title}}
brand: {{brand}} price: {{price}}

Target template: {{answer_choices[label]}} (like/dislike) – like
(4,5) / dislike (1,2,3)
```

Prompt ID: Z-2

```
Input template: Here are the details about a new product: title:
{{item_title}} brand: {{brand}} price: {{price}} What star will
{{user_desc}} probably rate the product?
-1 -2 -3 -4 -5

Target template: {{star_rating}}
```

Prompt ID: Z-5

```
Input template: Generate a possible explanation for {{user_desc}} 's
preference about the following product: title: {{item_title}} brand:
{{brand}} price: {{price}}

Target template: {{explanation}}
```

Prompt ID: Z-6

```
Input template: Based on the word {{feature_word}}, help
user_{{user_id}} write a {{star_rating}}-star explanation for this
new product: title: {{item_title}} price: {{price}} brand: {{brand}}

Target template: {{explanation}}
```

Summary

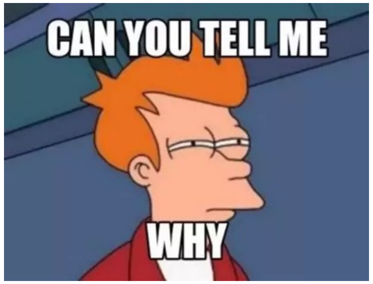
- Trustworthy and Responsible Recommendation
  - Explainability, Fairness, Echo Chambers, Controllability
  - Many other perspectives: Robustness, Accountability, Privacy, etc.

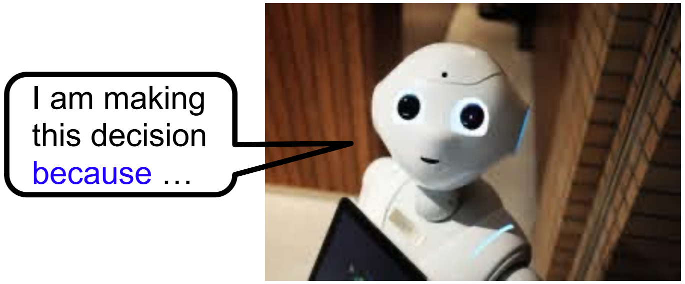
Objective AI　　　　Subjective AI

Computer Vision　　　　NLP　　　　Recommendation

Methods

	Human-centered Tasks
Counterfactual Reasoning	Counterfactual Explainable Recommendation
Counterfactual Fairness	Counterfactual Explainable Fairness
Human-controllable AI	User Controllable Recommendation
Large Recommendation Models	Multi-task Learning, Natural Language Explanation





Yongfeng Zhang
Department of Computer Science, Rutgers University
yongfeng.zhang@rutgers.edu
http://yongfeng.me