RUTGERS
UNIVERSITY | NEW BRUNSWICK

Explainable AI: From Human to Nature

Yongfeng Zhang
Department of Computer Science
Rutgers University
yongfeng.zhang@rutgers.edu

Outline of the Talk

- Motivation and Background
- Our Research on Explainable AI

Applications →

Methods ↓

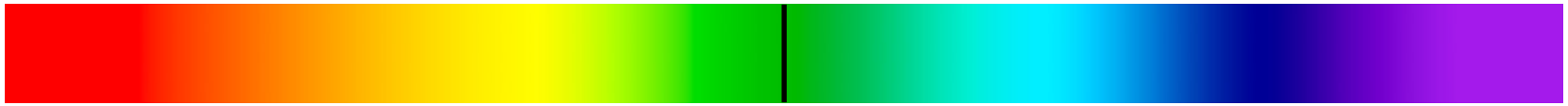
	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

- Summary and Future Directions

AI helps in many Research Areas

- A (very rough) spectrum of research discipline system

Human/Arts Nature/Science



… Arts Literature Sociology Education Economics CS/AI Engineering Biology Medical Chemistry Physics …

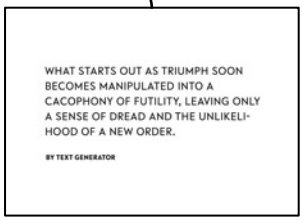
AI helps in many Research Areas

- A (very rough) spectrum of research discipline system

Human/Arts — Nature/Science

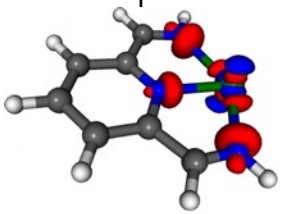
… Arts Literature Sociology Education Economics CS/AI Engineering Biology Medical Chemistry Physics …

Computer Painting



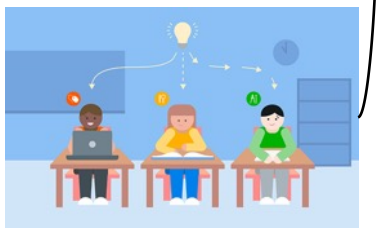
Computer-generated poems/computer writing

Computational Biology



Computational Chemistry

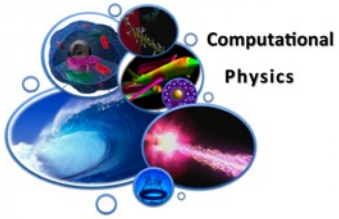
Computer Music Composing



Personalized education based on ML

Quantitative and Computational Econ

Computational drug discovery



Computational Physics

Explainable AI Matters in Both Directions

- In both Human-centered and Nature-oriented Areas
  - 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, why we should trust this decision)

- Even more important in **high-stake applications** related to health, safety, and law

Healthcare

Self Driving

Legal Assistants

  - Errors/bias may cause severe loss in life, money, and reputation
- Explainable AI helps humans to make better decisions
- Explainable AI helps scientists to make insightful science discoveries

Example I: Resume Ranking and Recommendation
– A Human-centered Task

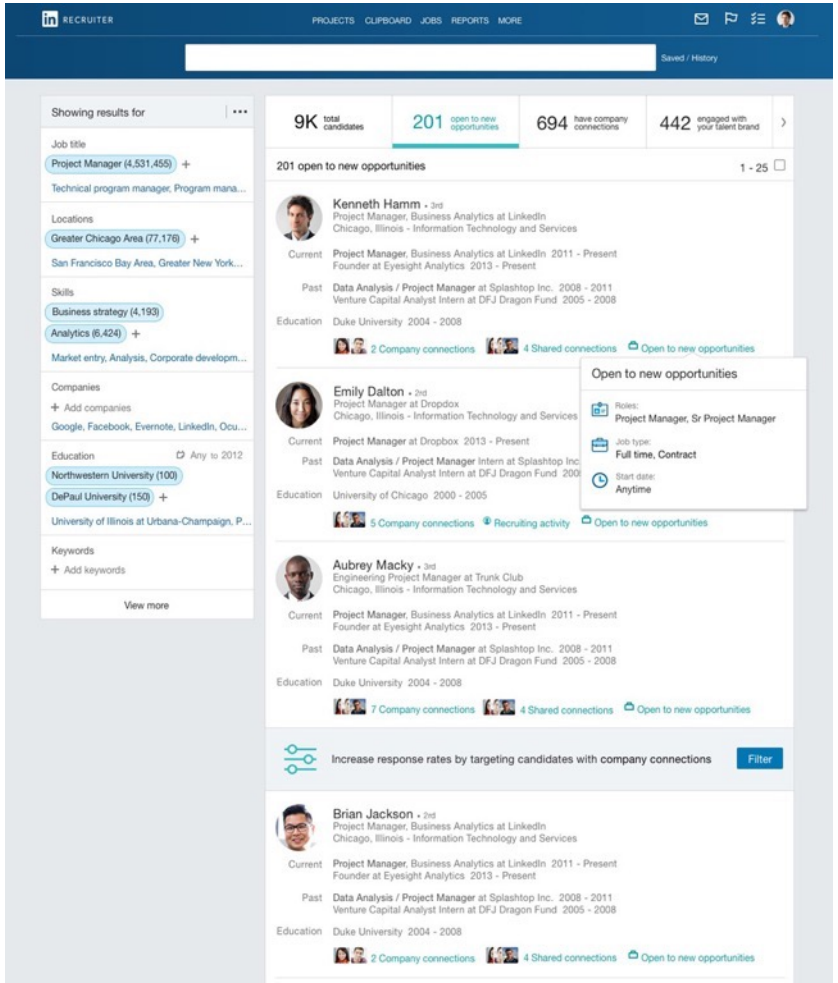


Figure 1: A (mocked) screenshot from the LinkedIn Recruiter (credit to [1])

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)

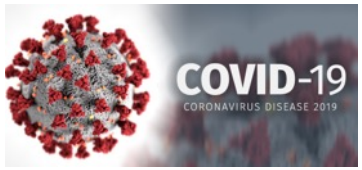
Problem:
From recruiter's perspective:
Why this candidate is a better fit than another?

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!

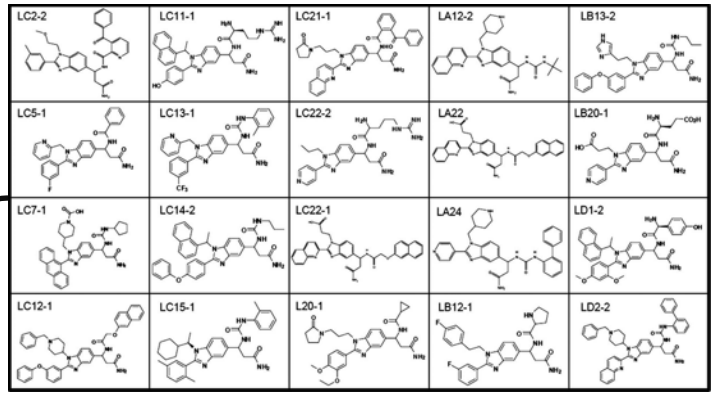
[1] https://engineering.linkedin.com/blog/2019/04/ai-behind-linkedin-recruiter-search-and-recommendation-systems

Example II: Drug Discovery
– A Nature-oriented Task



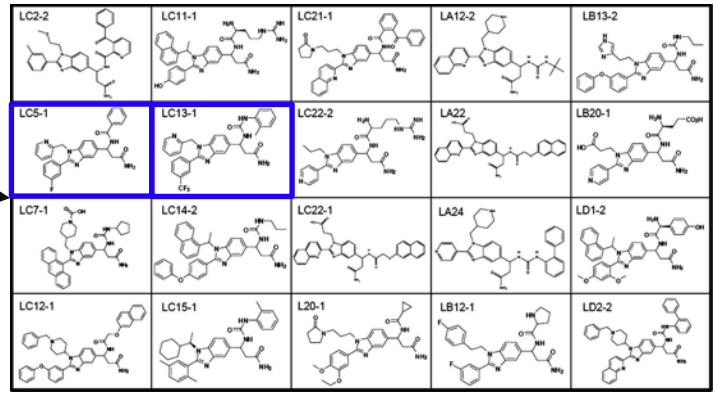
Step 1: Target a disease

Generative AI Models
e.g., Graph Generation Model

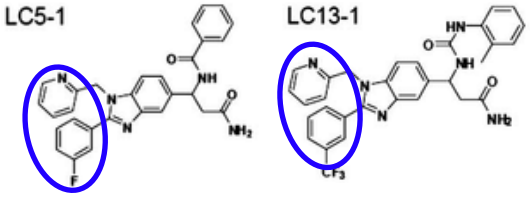


Step 2: There could be many AI-generated candidate drug molecular structures. Impossible to do clinical trial for all *.

Discriminative AI Models
e.g., Graph Classification based on Graph Neural Networks (GNN)



Step 3: Train a ML model (e.g., Graph Neural Network) for pre-selection. Selects only a few potentially effective drugs.



Step 4: Use Explainable AI to explain the results.
- **Why** the model believes the selected drugs are potentially effective (e.g., the special benzene structure can effectively bond with the virus).

Explainable AI Models
e.g., Explainable GNN

Step 5: The explanation helps scientists to analyze results, make decisions, do clinical trials, or refine the molecular structure.

* Image credit to [2] Fei et al. "Screening small-molecule compound microarrays for protein ligands without fluorescence labeling with a high-throughput scanning microscope." *Journal of biomedical optics* 15, no. 1 (2010): 016018.

Our Research on Explainable AI

Two dimensions: Method and Application

Applications

Methods

	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

Neural Logic and Neural Symbolic Reasoning

Two dimensions: Method and Application

Applications

Methods

	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

Neural Logic and Neural Symbolic Reasoning

- AI ≠ ML, AI ⊃ ML

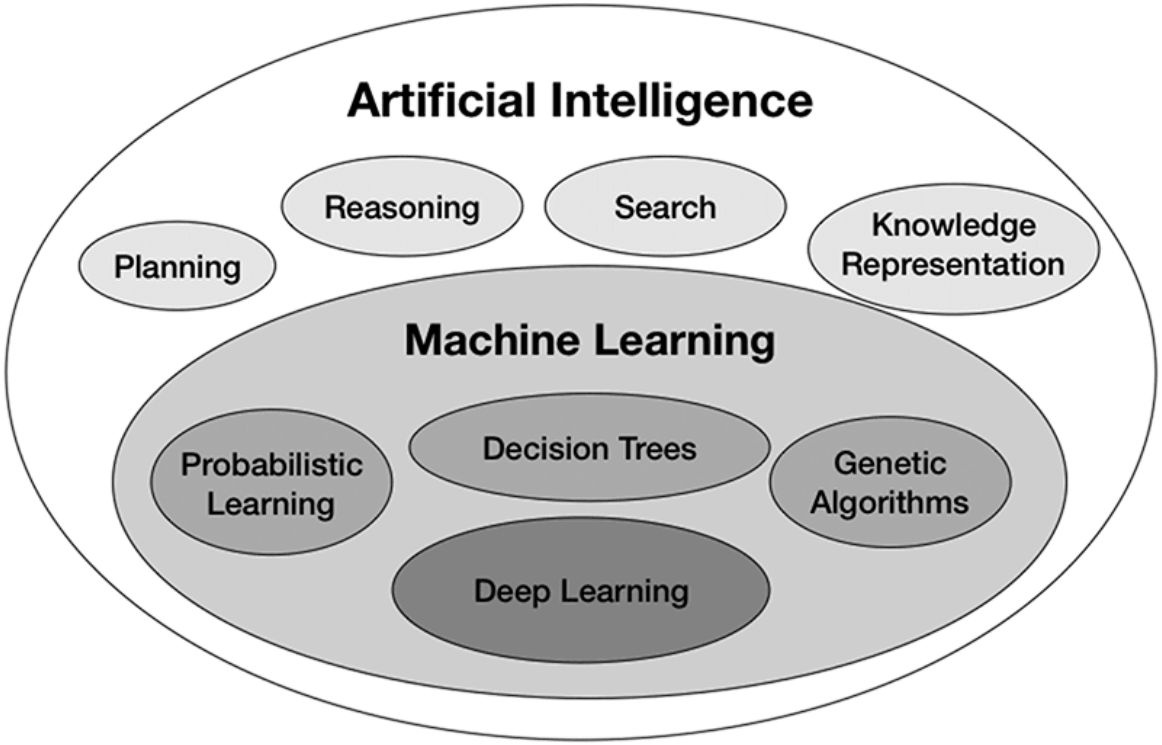
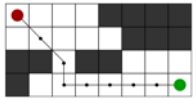


Image credit to [3] Marcus G, Davis E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon; 2019 Sep 10.

A (very rough) History of AI Research

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

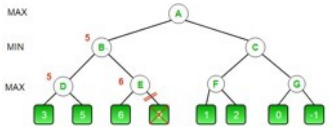
Example Methods:



A* Search

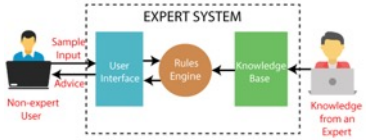
Knowledge Representation and Reasoning

Production Rules



Alpha-Beta Pruning

Example Systems:

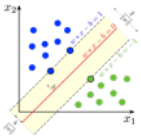


Expert systems
(If-Then production rules)

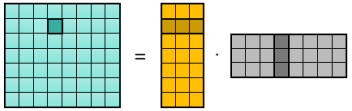
Chess AI
(IBM Deep Blue)

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

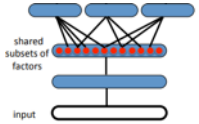
Example Methods:



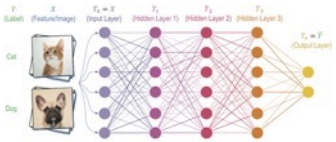
Support Vector Machine



Matrix Factorization

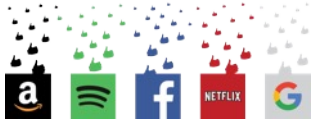


Representation Learning



Deep Neural Networks

Example Systems:



Recommender Systems

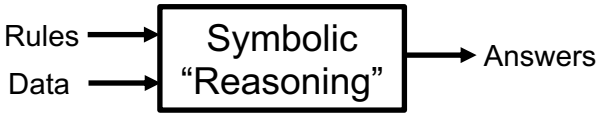
Image and Language Processing

Symbolism vs Connectionism - A comparison

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

Symbolism/Rationalism

A top-down design approach



Advantages:

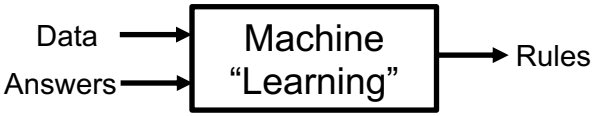
- Accurate decision
- Highly explainable & human readable

Disadvantages:

- Extensive expert human efforts
- Difficult to handle noisy data

Connectionism/Empirism

A bottom-up design approach



Advantages:

- Less human efforts
- Better at working with noisy data

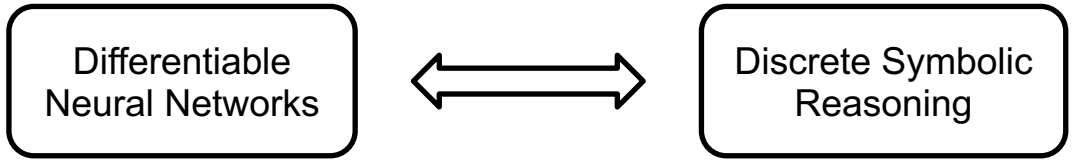
Disadvantages:

- 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



SAT and Solving Logical Equations

- The Propositional Satisfiability (SAT) Problem
  - Determining if there exists an interpretation that satisfies a given Boolean formula

$$\neg a \wedge b \wedge c = T \ ? \qquad \neg a \wedge a \wedge c = T \ ?$$

  - SAT is the first problem proven to be NP-complete

- Extension: Solving logical equations
  - Also an NP-complete problem

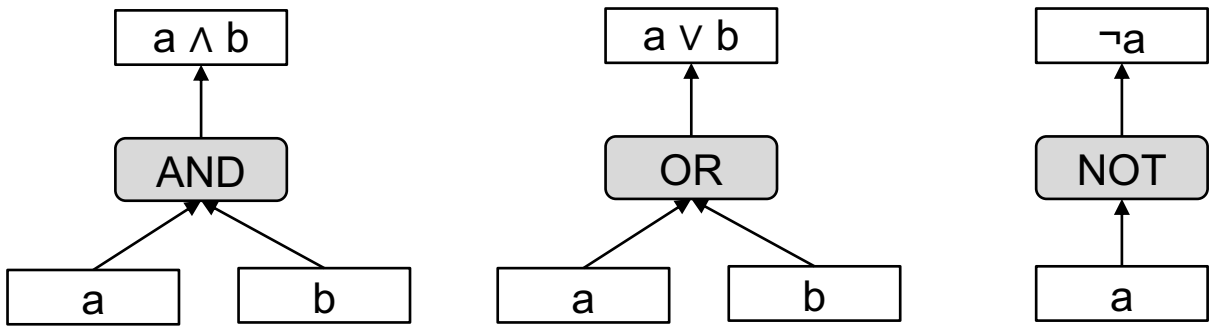
$$\begin{aligned} \neg a \wedge b &= T \\ (a \wedge b) \vee c &= T \\ a \vee b \vee \neg c &= F \end{aligned} \quad \Longrightarrow \quad \begin{aligned} &a = ?, b = ?, c = ? \\ &a \wedge (\neg b \vee c) = ? \end{aligned}$$

- Logical expressions are highly transparent
  - E.g., $\neg a \wedge b \wedge c = T$ is because $a = F$ and $b = c = T$
  - Many applications can be formulated as logical reasoning problems
  - Help to enhance explainable AI in many applications

- Can we use machine learning to
  - Solve logical equations?
  - Predict the True/False value of new logical equations?

Neural Logic Reasoning [4]

- Key idea
  - 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

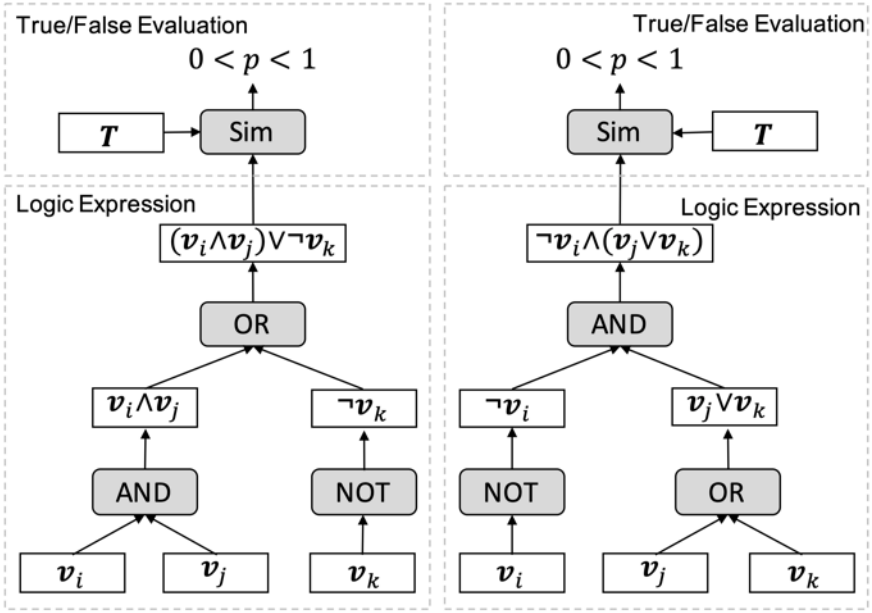
$$\mathrm{AND}(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{H}_{a2} f(\mathbf{H}_{a1}(\mathbf{w}_i \oplus \mathbf{w}_j) + \mathbf{b}_a) \quad \mathrm{NOT}(\mathbf{w}) = \mathbf{H}_{n2} f(\mathbf{H}_{n1}\mathbf{w} + \mathbf{b}_n)$$

[4] S. Shi, H. Chen, W. Ma, J. Mao, M. Zhang, and Y. Zhang. "Neural Logic Reasoning." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1365-1374. 2020.

Logic-Integrated Neural Network (LINN)

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

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



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

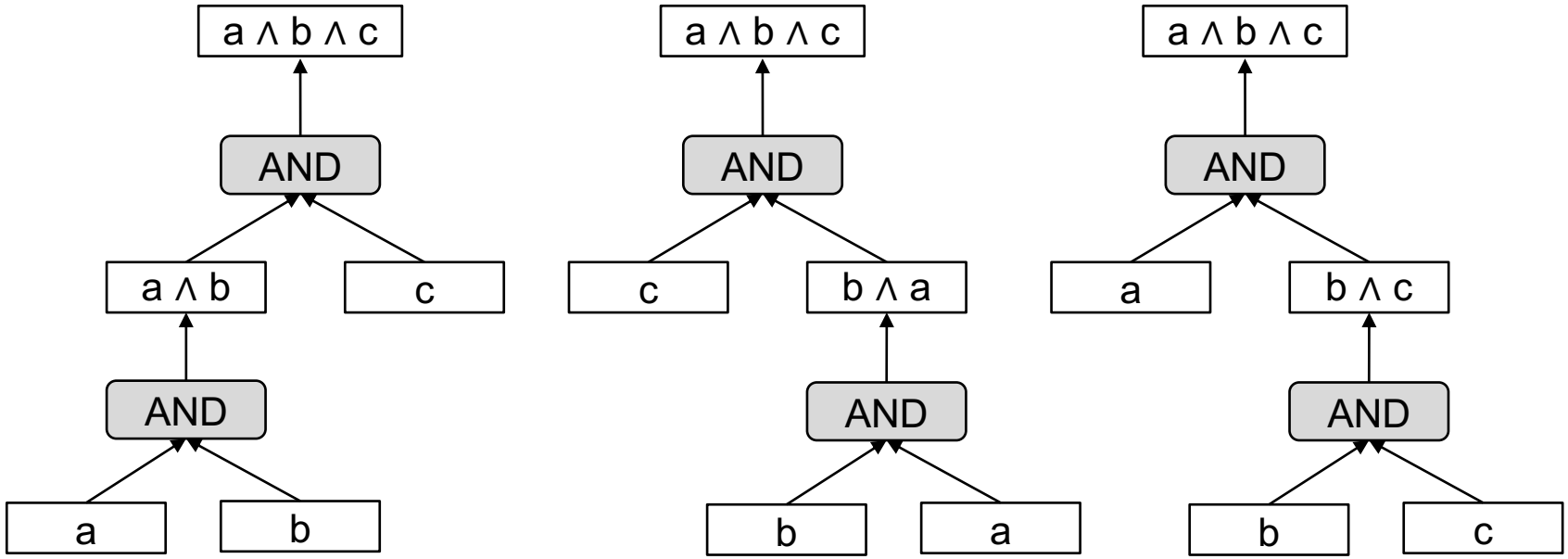
	Logical Rule	Equation	Logic Regularizer r_i
NOT	Negation	$\neg T = F$	$r_1 = \sum_{w \in W \cup \{T\}} Sim(\text{NOT}(\mathbf{w}), \mathbf{w})$
	Double Negation	$\neg(\neg w) = w$	$r_2 = \sum_{w \in W} 1 - Sim(\text{NOT}(\text{NOT}(\mathbf{w})), \mathbf{w})$
AND	Identity	$w \wedge T = w$	$r_3 = \sum_{w \in W} 1 - Sim(\text{AND}(\mathbf{w}, \mathbf{T}), \mathbf{w})$
	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 $L_1 = L_t + \lambda_l R_l = L_t + \lambda_l \sum_i r_i$

Logical Regularization over Neural Modules

- Some laws can not be easily represented as regularizers
  - Commutative Law: $a \wedge b = b \wedge a, a \vee b = b \vee a$
  - Associative Laws: $(a \wedge b) \wedge c = a \wedge (b \wedge c),\ (a \vee b) \vee c = a \vee (b \vee c)$

- Solution: Randomly shuffling variables for model training

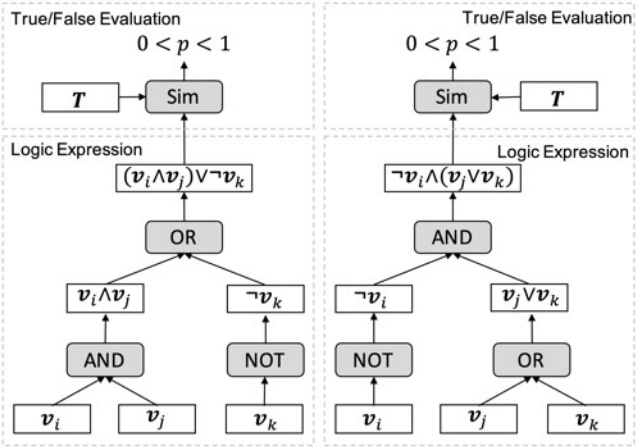


Final Model

- Input

$$(\neg v_{80} \wedge v_{56} \wedge v_{71}) \vee (\neg v_{46} \wedge \neg v_7 \wedge v_{51} \wedge \neg v_{47} \wedge v_{26}) \vee v_{45} \vee (v_{31} \wedge v_{15} \wedge v_2 \wedge v_{46}) = T$$
$$(\neg v_{19} \wedge \neg v_{65}) \vee (v_{65} \wedge \neg v_{24} \wedge v_9 \wedge \neg v_{83}) \vee (\neg v_{48} \wedge \neg v_9 \wedge \neg v_{51} \wedge v_{75}) = F$$
$$\neg v_{98} \vee (\neg v_{76} \wedge v_{66} \wedge v_{13}) \vee v_{97}(\wedge v_{89} \wedge v_{45} \wedge v_{83}) = T$$
$$v_{43} \wedge v_{21} \wedge \neg v_{53} = F$$

- Model
  - Randomly initialize variable embeddings and logical operator modules
  - Model Training $L_1 = L_t + \lambda_l R_l = L_t + \lambda_l \sum_i r_i$



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

- Output
  - The learned variable embeddings
  - The learned logical operator modules (AND, OR, NOT)

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

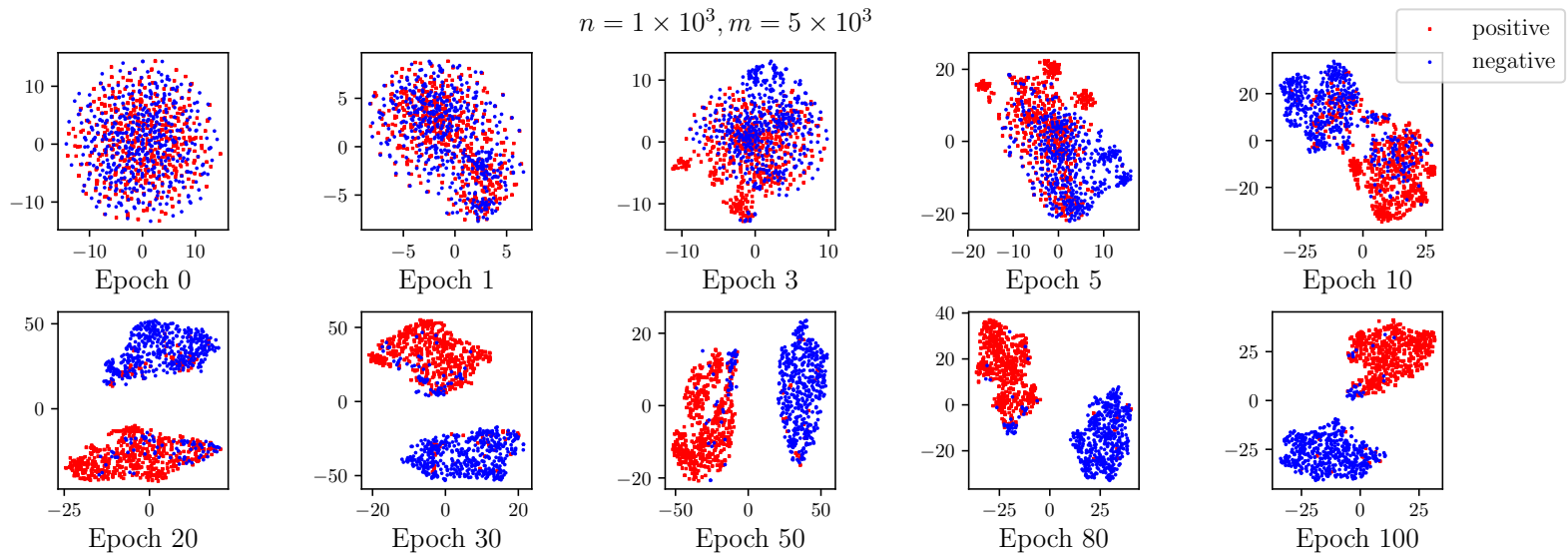
$$(v_{65} \wedge \neg v_{24} \wedge v_9 \wedge \neg v_{83}) \vee (\neg v_{48} \wedge \neg v_9 \wedge \neg v_{51} \wedge v_{75}) = ?$$

	$n = 1 \times 10^3, m = 3 \times 10^3$		$n = 1 \times 10^4, m = 3 \times 10^4$	
	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.

Application 1: Solving Logical Equations

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

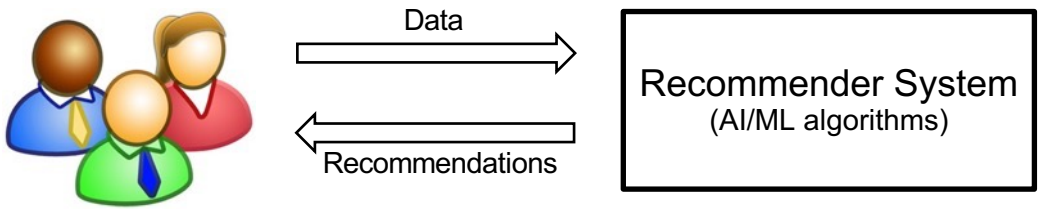


Accuracy of variable solving: 96%

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

Application 2: Explainable Recommendation

- Recommender System (RS) is a representative Human-centered AI task
  - Naturally involves human-in-the-loop



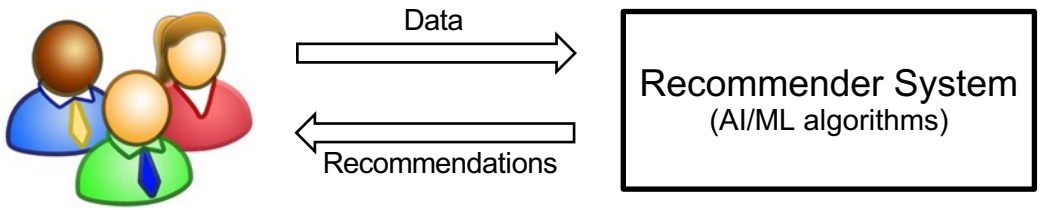
- A wide scope of applications
  - E-commerce (product recommendation)
  - Smart and Connected Communities (driving route recommendation / passenger recommendation)
  - Social Networks (friend/tweet recommendation)
  - Sharing Economy (house recommendation)
  - Search Engines (personalized search / advertising)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

Application 2: Explainable Recommendation

- Recommender System (RS) is a representative Human-centered AI task
  - Naturally involves human-in-the-loop



- Explanation is extremely important in recommender systems
  - Because the system is eventually making suggestions or decisions for humans
  - People naturally want to know the **Why**
  - Improve transparency, trustworthiness and reliability
  - An important approach towards Responsible AI

Responsible AI

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 Ranking Loss

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

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

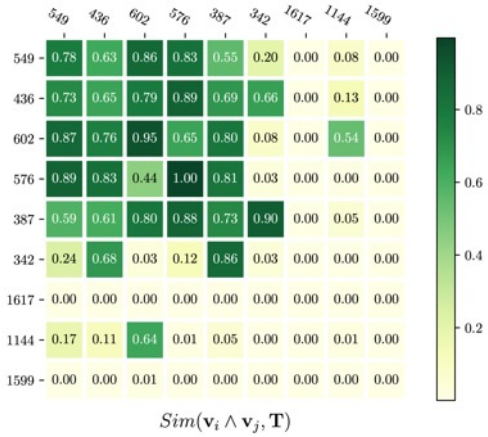
$$L = -\sum_{e^+} \log\left(sigmoid(p(e^+) - p(e^-))\right) + \lambda_l \sum_i r_i + \lambda_\ell \sum_{w \in W} \|\mathbf{w}\|_F^2 + \lambda_\Theta \|\Theta\|_F^2$$

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.



Knowledge Graph Reasoning

Two dimensions: Method and Application

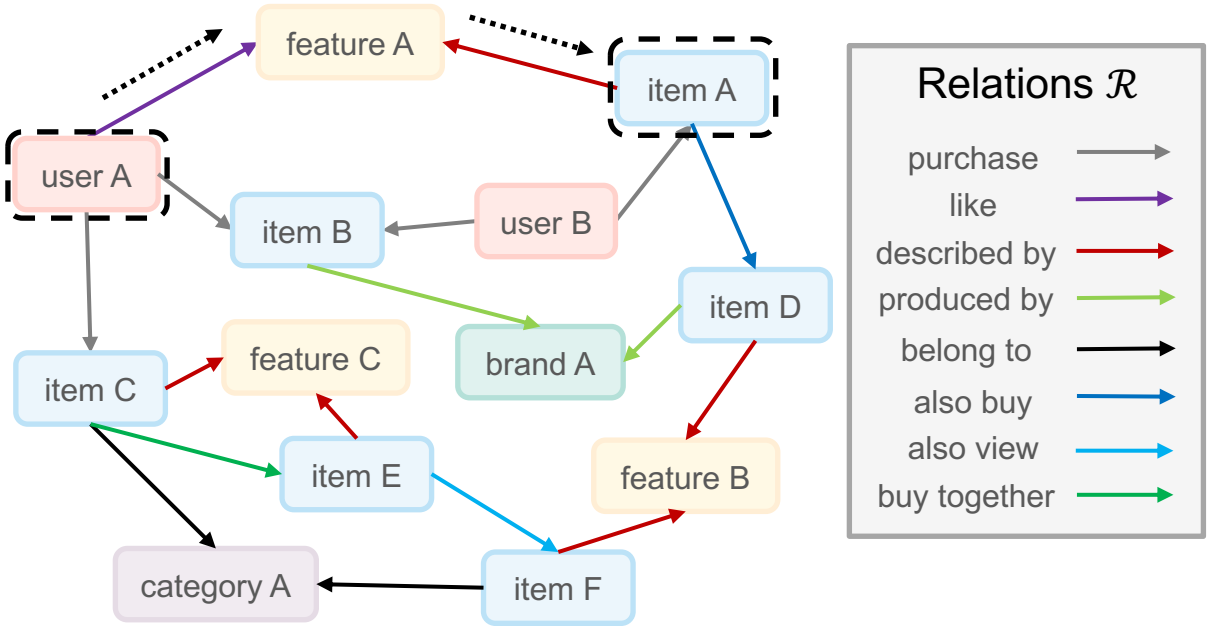
Applications

Methods

	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

Knowledge Graph Reasoning for Explainable AI

- Knowledge Graph (KG) contains rich and structured information
  - A knowledge graph $\mathcal{G}$ with entity set $\mathcal{E}$ and relation set $\mathcal{R}$ is defined as $\mathcal{G} = \{(e_h, r, e_t) | e_h, e_t \in \mathcal{E}, r \in \mathcal{R}\}$, where each triplet (e_h, r, e_t) represents a fact of the relation r from head entity e_h to tail entity e_t.
  - Facilities many applications: search, recommendation, dialog systems
- Path-reasoning in KG
  - Provides transparent decisions and the reasoning path services as explanations

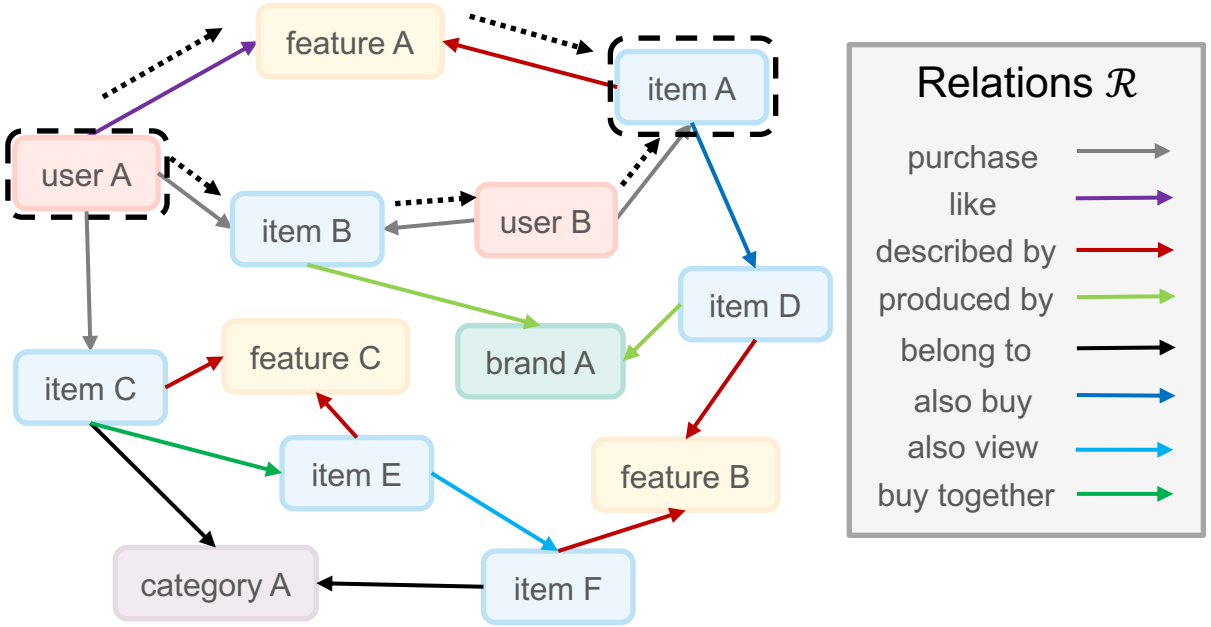


Reinforcement Knowledge Graph Reasoning [5]

Input: KG, u="user A", maximum exploration depth (e.g., 3)

Output:
(1) A final decision (e.g., to recommend item A)
(2) Path-based explanations for the decision:

user A → purchase → item B → purchased_by → user B → purchase → item A
user A → like → feature A → described_by → item A



Challenge:
(1) Unknown target: the target node is NOT known before path reasoning
(2) Huge search space

[5] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang. "Reinforcement knowledge graph reasoning for explainable recommendation." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 285-294. 2019.

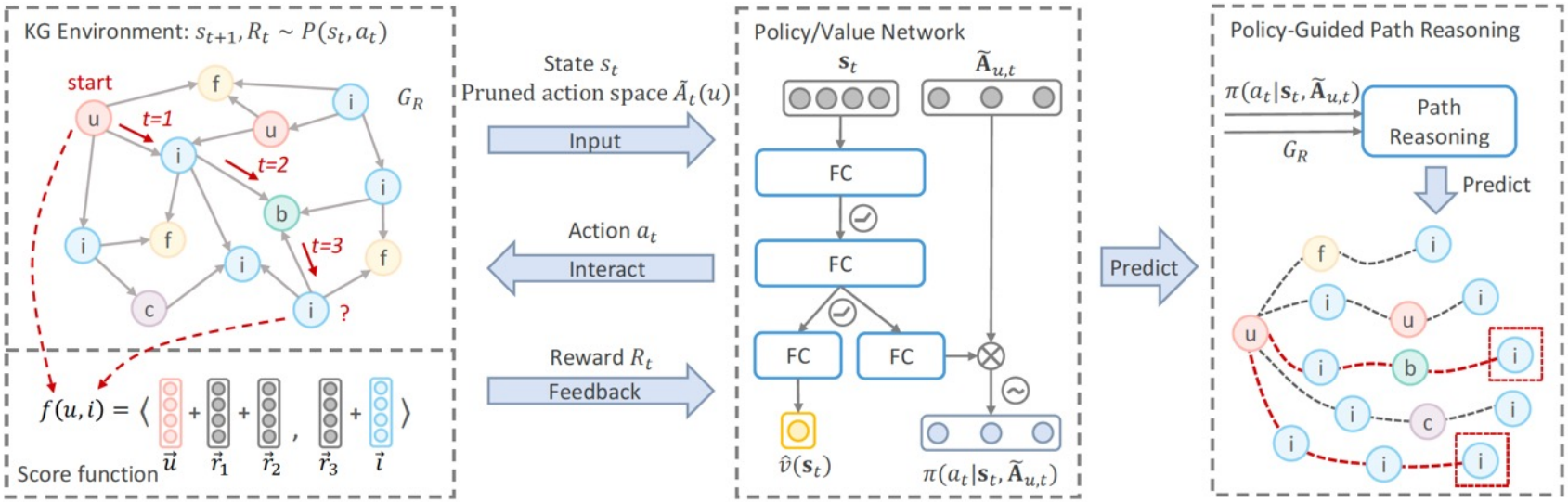
Reinforcement Knowledge Graph Reasoning

Overview of the RL-based KG reasoning architecture

(1) Markov Decision Process (MDP) over KG

(2) Policy/Value Network

(3) Policy-Guided Path Reasoning



State $s_t = (u, e_t, h_t)$

u: user, e_t current position

h_t: history positions

$$u \xrightarrow{r_1} e_1 \xrightarrow{r_2} \cdots \xrightarrow{r_{t-1}} e_{t-1} \xrightarrow{r_t} e_t$$

Action Space: Adjacent entities

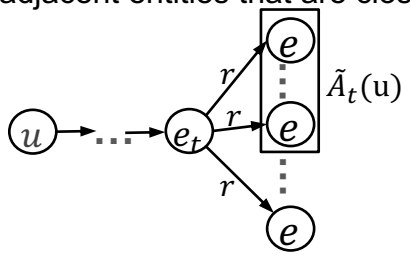
$A_t = \{(r,e) | (e_t, r, e) \in \mathcal{G}, e \notin \{e_0, \ldots, e_{t-1}\}\}$

Reduced Action Space: action pruning

$\tilde{A}_t(u) = \{(r,e) | rank(f((r,e)|u)) \leq \alpha, (r,e) \in A_t\}$

$f((r,e)|u)$ is some heuristic function

Select adjacent entities that are close enough



Reward

$$R_T = \begin{cases} \max\left(0, \dfrac{f(u, e_T)}{\max_i f(u,i)}\right), & if\ e_T \in I \\ 0, & otherwise \end{cases}$$

Indicates how "good" is the action

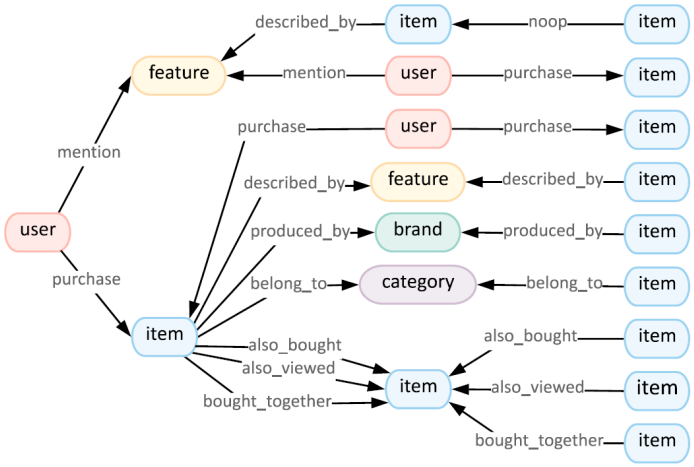
A good action leads to an item that the user likes w.h.p.

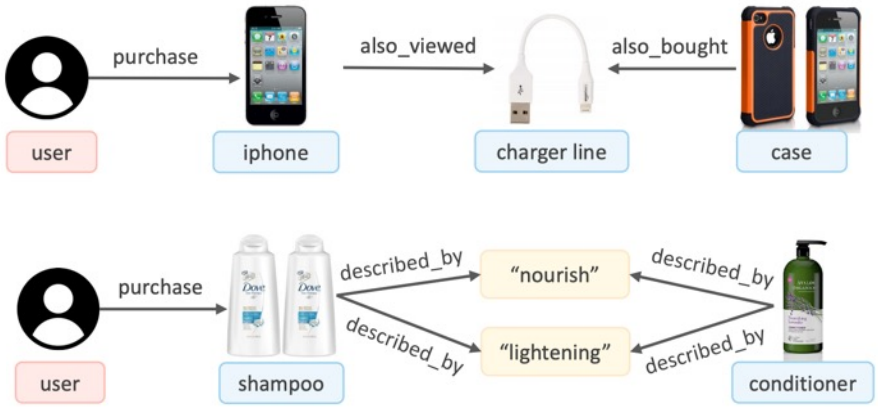
Better Recommendation and Better Explanation

- Provides better recommendations in E-commerce

Dataset	CDs & Vinyl				Clothing				Cell Phones				Beauty			
Measures (%)	NDCG	Recall	HR	Prec.	NDCG	Recall	HR	Prec.	NDCG	Recall	HR	Prec.	NDCG	Recall	HR	Prec.
BPR	2.009	2.679	8.554	1.085	0.601	1.046	1.767	0.185	1.998	3.258	5.273	0.595	2.753	4.241	8.241	1.143
BPR-HFT	2.661	3.570	9.926	1.268	1.067	1.819	2.872	0.297	3.151	5.307	8.125	0.860	2.934	4.459	8.268	1.132
VBPR	0.631	0.845	2.930	0.328	0.560	0.968	1.557	0.166	1.797	3.489	5.002	0.507	1.901	2.786	5.961	0.902
TransRec	3.372	5.283	11.956	1.837	1.245	2.078	3.116	0.312	3.361	6.279	8.725	0.962	3.218	4.853	0.867	1.285
DeepCoNN	4.218	6.001	13.857	1.681	1.310	2.332	3.286	0.229	3.636	6.353	9.913	0.999	3.359	5.429	9.807	1.200
CKE	4.620	6.483	14.541	1.779	1.502	2.509	4.275	0.388	3.995	7.005	10.809	1.070	3.717	5.938	11.043	1.371
JRL	5.378*	7.545*	16.774*	2.085*	1.735*	2.989*	4.634*	0.442*	4.364*	7.510*	10.940*	1.096*	4.396*	6.949*	12.776*	1.546*
PGPR (Ours)	**5.590**	**7.569**	**16.886**	**2.157**	**2.858**	**4.834**	**7.020**	**0.728**	**5.042**	**8.416**	**11.904**	**1.274**	**5.449**	**8.324**	**14.401**	**1.707**

- Intuitive explanations for the recommendations





Extension 1: Explainable Conversational AI

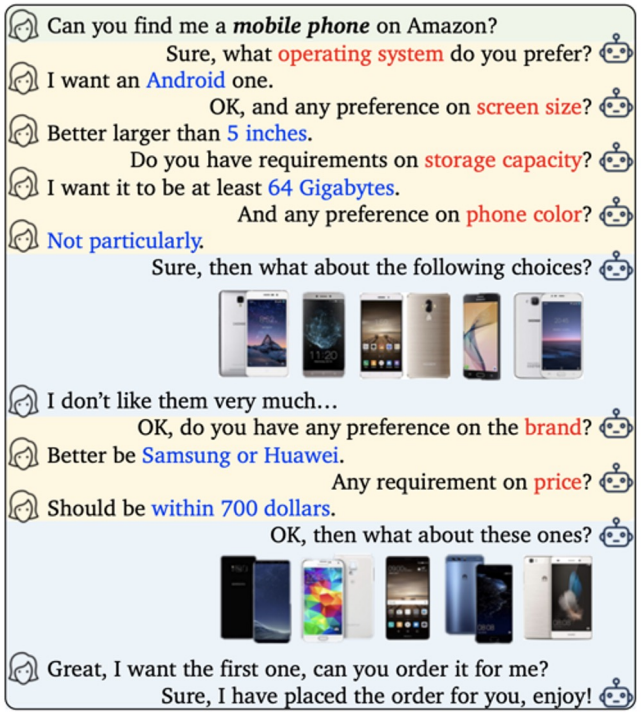
- Conversational Recommendation [6]
  - Conversational recommendation provides personalized recommendations through natural language dialog with users
  - Usually through voice-based interfaces

Siri

Alexa

Apple HomePod

  - Promising for smart home services

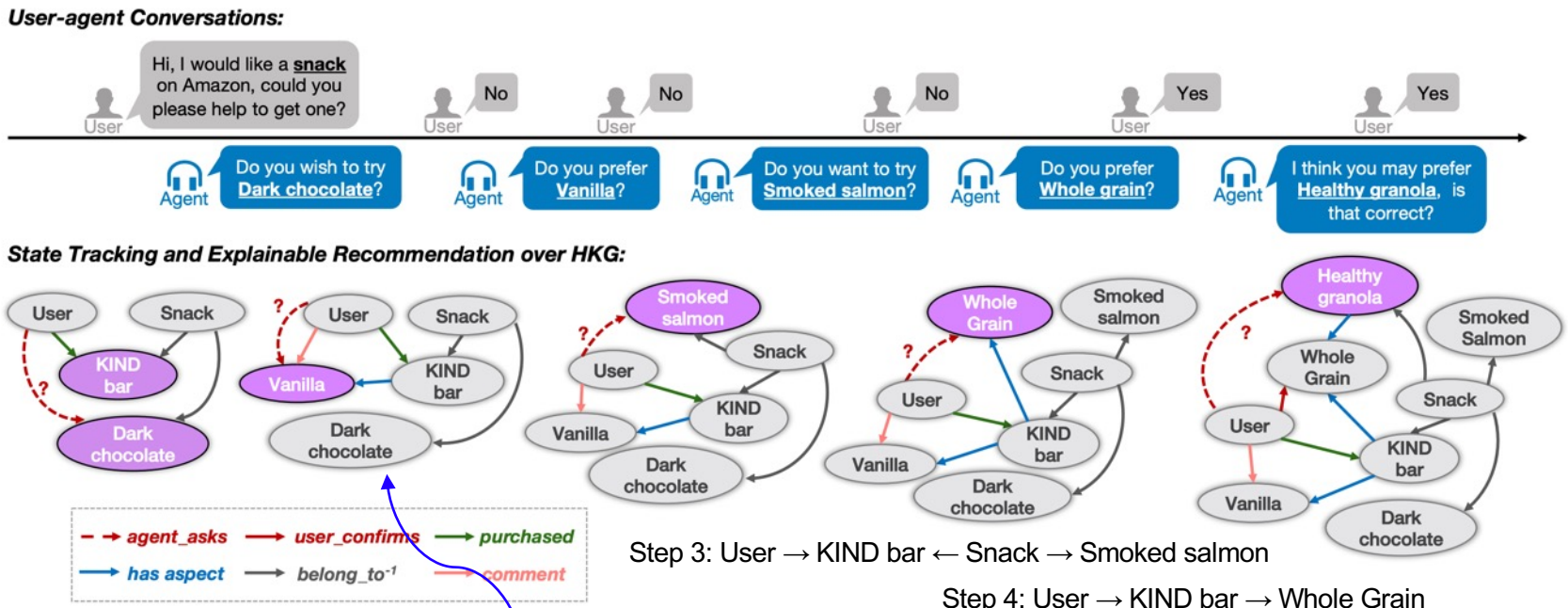


- Explanation is even more important
  - Since system usually can only provide one single recommendation through voice channel
    - E.g., a piece of music recommendation
  - Different from a "list" of recommendations in traditional visual interfaces
  - Important to explain why the single item is recommended

[6] Z. Fu, Y. Xian, Y. Zhang, Y. Zhang. "Tutorial on Conversational Recommendation Systems." In 14th ACM Conference on Recommender Systems. 2020.

KG Reasoning for Explainable Conversation

- Explainable Conversational Recommendation over KG Reasoning [7,8]
  - Transparent dialog state tracking over knowledge graphs



Step 1: User → KIND bar ← Snack → Dark chocolate

Step 2: User → KIND bar → Vanilla

Step 3: User → KIND bar ← Snack → Smoked salmon

Step 4: User → KIND bar → Whole Grain

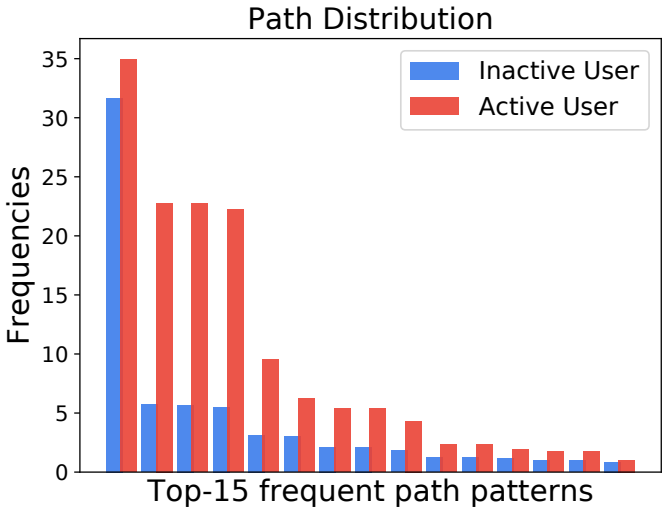
Step 5: User → KIND bar → Whole Grain ← Healthy granola

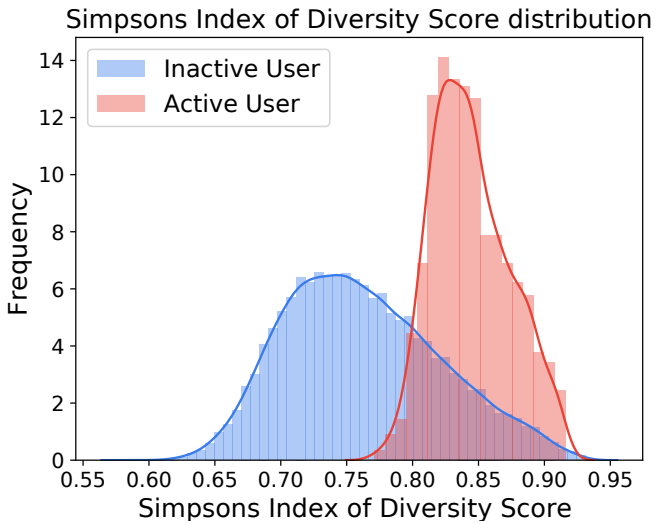
Explanation: Because you liked the KIND bar produce before and it was vanilla flavor.

[7] Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. de Melo, Y. Zhang. "HOOPS: Human-in-the-Loop Graph Reasoning for Conversational Recommendation." SIGIR 2021.
[8] Z. Fu, Y. Xian, S. Geng, G. de Melo, Y. Zhang. " Popcorn: Human-in-the-loop Popularity Debiasing in Conversational Recommender Systems." CIKM 2021.

Extension 2: Fairness of Explanations

- Unfairness issues in KG-based Explainable Recommendation [9,10]
  - Observation: Top 5% active users have quantitatively more and qualitatively more diverse data in KG
    - Top 5% active users may be economically advantaged than 95% inactive users, more money and time in EC





  - Consequence: Active users' data will dominate a ML algorithm
    - The majority 95% users' recommendation accuracy is much worse than the 5% active users

Dataset	CDs & Vinyl								Clothing							
Measures (%)	Overall		Inactive Users		Active Users		GRU		Overall		Inactive Users		Active Users		GRU	
	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1
HeteroEmbed	6.992	3.576	6.526	3.373	15.843	7.429	9.317	4.056	3.221	1.404	3.121	1.348	**5.130**	2.461	2.009	1.113

[9] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, G. de Melo. "Fairness-Aware Explainable Recommendation over Knowledge Graphs." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.*
[10] Y. Li, H. Chen, Z. Fu, Y. Ge and Y. Zhang. "User-oriented Group Fairness In Recommender Systems." WWW 2021.

Extension 2: Fairness of Explanations

- Fairness-aware Explainable Recommendation
  - Two groups of users: G_1 top 5% active users, G_2 the other 95% inactive users.
  - $\boldsymbol{Q_i}$: the recommendation result for user $\boldsymbol{i}$
  - $\mathcal{F}(\boldsymbol{Q_i})$: Any evaluation measure over the recommendation result

Group-level Recommendation Unfairness (GRU):

$$GRU(G_1, G_2, \boldsymbol{Q}) = \left| \frac{\sum_{i \in G_1} \mathcal{F}(\boldsymbol{Q_i})}{G_1,} - \frac{\sum_{i \in G_2} \mathcal{F}(\boldsymbol{Q_i})}{G_2,} \right|$$

Group-level Explanation Diversity Unfairness (GEDU):

$$GEDU(G_1, G_2, \boldsymbol{Q}) = \left| \frac{\sum_{i \in G_1} \boldsymbol{f}(\boldsymbol{Q_i})}{G_1,} - \frac{\sum_{i \in G_2} \boldsymbol{f}(\boldsymbol{Q_i})}{G_2,} \right|$$

Fairness constrained optimization

$$\max_{Q_{ij}} \mathcal{R} = \sum_{i=1}^{m} \mathcal{R}_{rec}(\boldsymbol{Q_i})$$

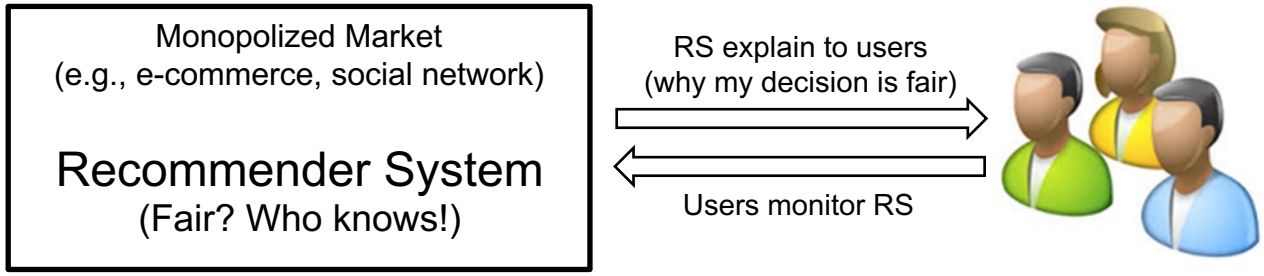
s.t. $GRU(G_1, G_2) < \varepsilon_1, GREU(G_1, G_2) < \varepsilon_2$

Dataset	CDs & Vinyl								Clothing							
Measures (%)	Overall		Inactive Users		Active Users		GRU		Overall		Inactive Users		Active Users		GRU	
	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1	NDCG	F_1
HeteroEmbed	6.992	3.576	6.526	3.373	15.843	7.429	9.317	4.056	3.221	1.404	3.121	1.348	**5.130**	2.461	2.009	1.113
Fair HeteroEmbed	**8.094**	**4.019**	**7.674**	**3.820**	**16.074**	**7.801**	**8.400**	**3.981**	**3.494**	**1.536**	**3.484**	**1.482**	3.691	**2.556**	**0.207**	**1.074**

Active users' recommendation keep stable or slightly decrease
Inactive users' recommendation improve
Overall recommendation improves, since inactive user takes the majority (95%)

Discussion: Transparency and Fairness

- Relationship between Transparency and Fairness
  - Two important aspects of Responsible AI
- Recent legal regulations to responsible AI
  - EU General Data Protection Regulation (GDPR), The California Privacy Act of 2018
  - Emphasize the fairness of algorithmic decisions in AI systems
- However, law enforcement is a big challenge
  - How does the law enforcement department know if a system is fair?
  - Check the source code? Unrealistic
    - (1) Modern AI systems are very complicated, (2) Confidential assets of the private entities
  - A system being fair on one aspect does not imply fairness on another aspect
- Solution: Fairness through Transparency
  - Require the system to explain to users why its decision is fair
  - Rely on each every user to monitor the system
    - User will report if he/she does not accept the explanation



A healthy virtuous cycle between user and system, and thus a healthy digital world

Natural Language Explanations

Two dimensions: Method and Application

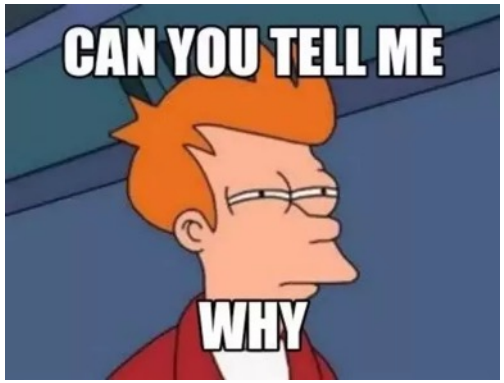
Applications

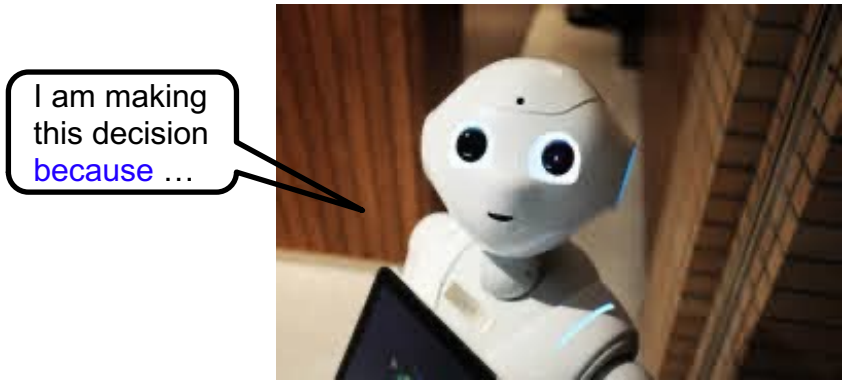
Methods

	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

Natural Language Explanations

- Natural language sentence is the most human-friendly way of explanation
  - 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

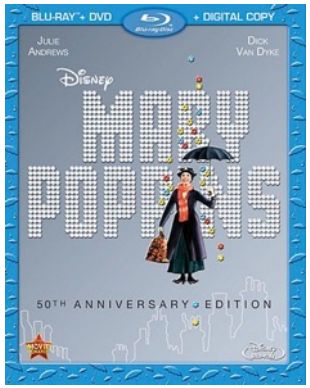
- Explainable Recommendation as Natural Language Generation
  - Recommendation is a very suitable task for developing natural language explanation models
  - High quality ground-truth explanations from humans

★★★★★ **Trip Saver**

Reviewed in the United States on October 22, 2017

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

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

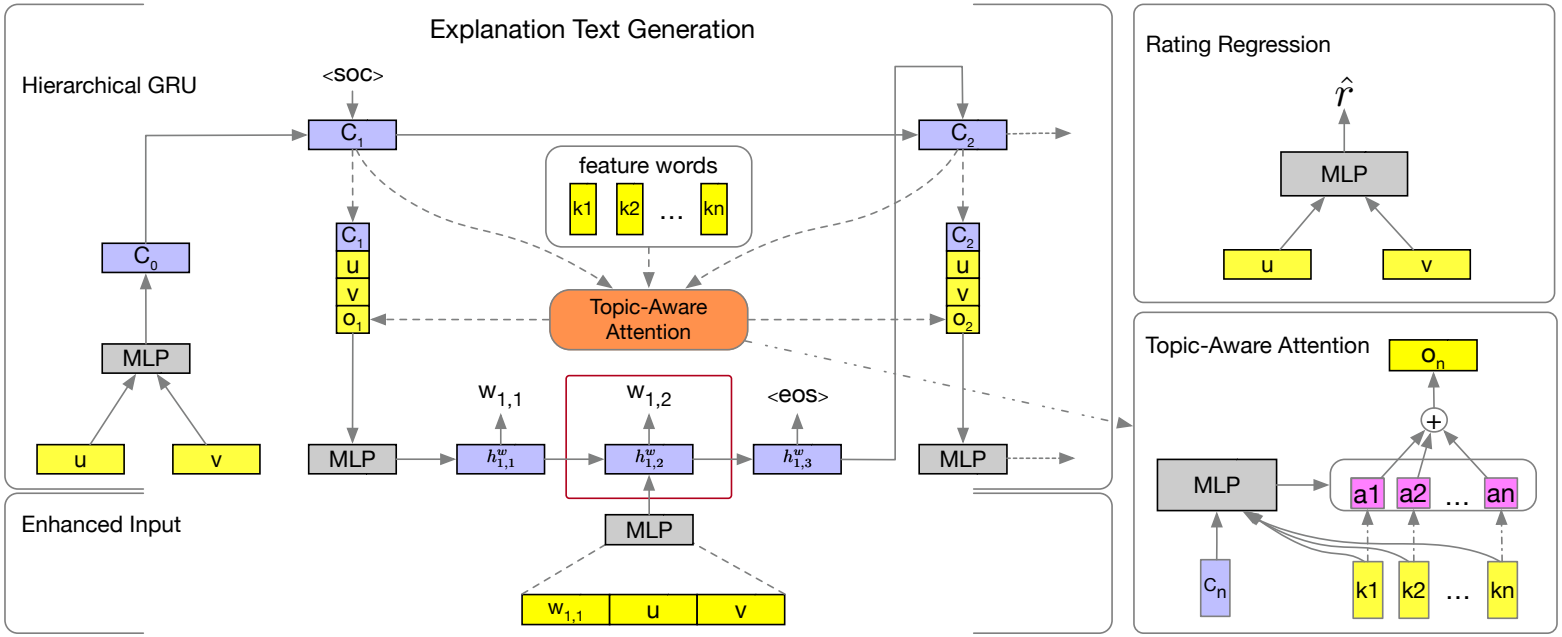
The movie holds up well as a glorious musical. The acting, singing, choreography, staging, special effects are all great. The plot still works. This is a movie my wife and I love watching over and over. The blu ray version is beautiful. The quality of the image shows itself during the extreme close-ups 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. 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

Natural Language Generation

- Hierarchical Sequence-to-Sequence (HSS) Generation [11]



Overview of our HSS model. Two major modules—explanation text generation module and rating regression module

[11] H. Chen, X. Chen, S. Shi, Y. Zhang. "Generate natural language explanations for recommendation." In Proceedings of SIGIR'19 Workshop on Explainable Recommendation and Search (EARS). ACM. 2019.

Natural Language Generation

- Some examples of the generated explanations

Cases	Item	Generated Explanation Sentences
Good Cases	Perfume	The **bottle** is very **light** and the **smell** is very **strong.**
	Headphone	The **price** is great. The **sound quality** is great.
Bad Cases	Shampoo	The **scent** is a good **product**. I have to use this product.
	Wireless Router	The **price** is a great. The **sound** is great.

Observations:

1. Some explanations are not talking about the product features.
2. All explanations tend to be positive (the machine is trying to persuade users no matter what item!)

We believe a responsible AI system should be honest:
say good about good items and say bad about bad ones.

Neural Template Generation -- Better Controllability

- Explanation generation is very different from free-text generation
  - Explanation is "purposed-text generation": the explanation needs to meet certain purposes
  - Required more controllability of the generation model
    - A key problem of current natural language generation (NLG) models

- Controllability
  - 1) What feature to discuss about, and 2) Sentiment of the generated sentence

Rating	Feature	Explanation
3.90	bathroom	The **bathroom** was large and had a separate shower.
	tub	The bathroom had a separate shower and **tub**.
	rooms	The **rooms** are large and comfortable.
2.76	airport	It is not close to the **airport**.

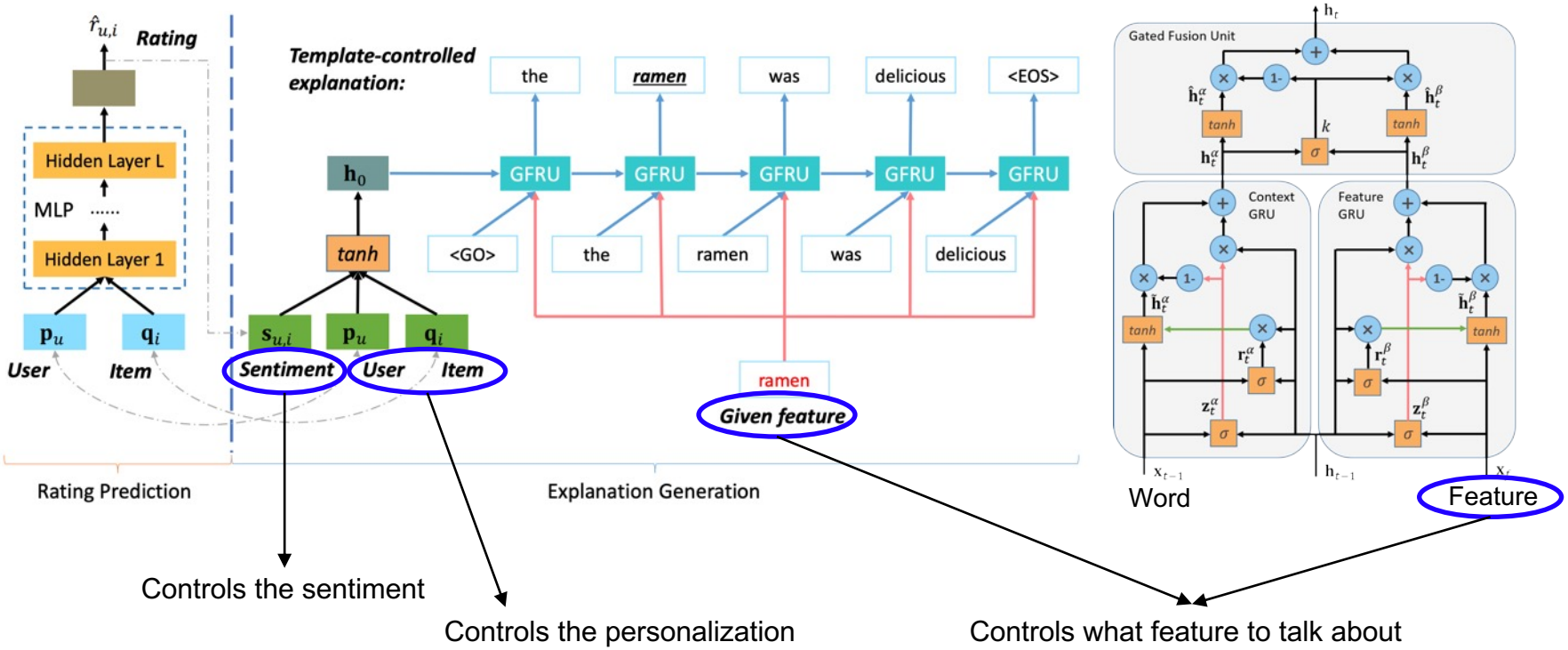
Able to talk about different features

Able to control the sentiment

- Solution: Neural Template Generation
  - Template-based sentence generation: Highly controllable, but less flexible
    - E.g., We recommendation because the product is good on [*feature*]
  - Free-text sentence generation: Highly flexible, but less controllable
    - E.g., Sequence-to-Sequence models, BERT, GPT, etc.

Neural Template Generation -- Better Controllability

- The Model: Gated Fusion Recurrent Unit (GFRU) [12]



[12] L. Li, Y. Zhang, L. Chen. "Generate neural template explanations for recommendation." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 755-764. 2020.
[13] L. .Li, Y. Zhang, L. Chen. “Personalized Transformer for Explainable Recommendation”. ACL 2021.
[14] S. Geng, Z. Fu, Y. Ge, L. Li, G. de Melo, Y. Zhang. “Improving Personalized Explanation Generation through Visualization”. ACL 2022.

Neural Template Generation -- Better Controllability

- Examples of the generated explanations

- Good linguistic quality
  - Learn templates from data e.g., "__ was large/comfortable"

- Good controllability
  - Generate more targeted explanations for specific features
  - Consistent with the sentiment of the predicted ratings

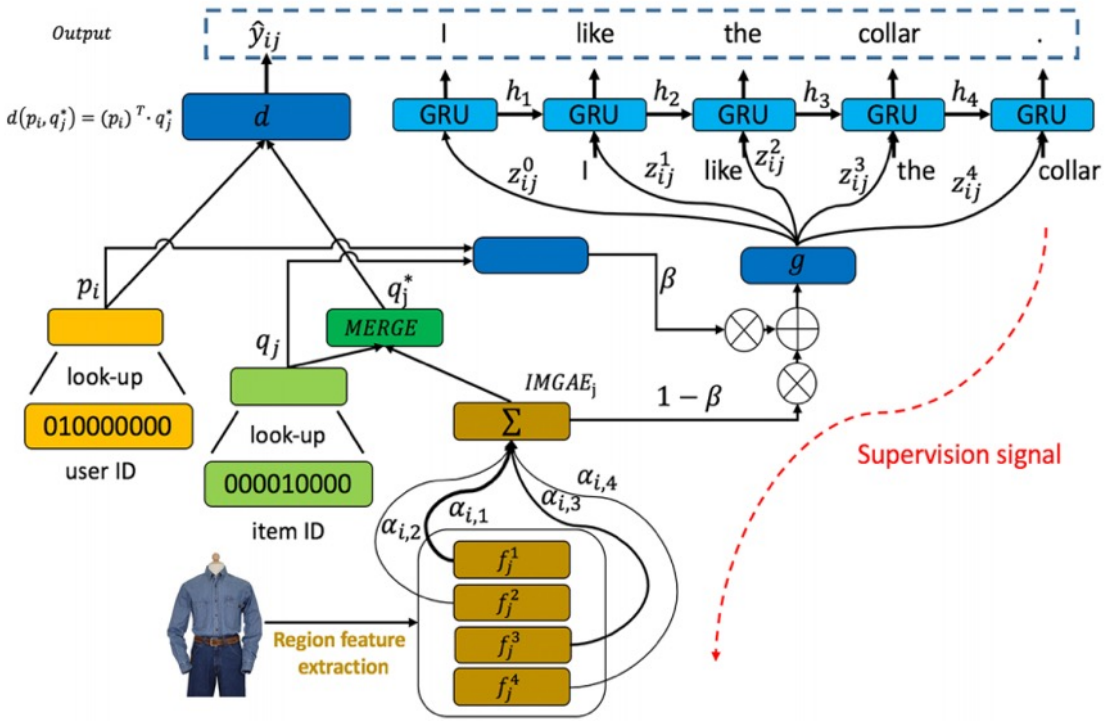
Rating	Feature	Explanation
4		***The rooms are spacious and the bathroom has a large tub.***
3.90	bathroom	The **bathroom** was large and had a separate shower.
	tub	The bathroom had a separate shower and **tub**.
	rooms	The **rooms** are large and comfortable.
4		***The rooms are brilliant and ideal for business travellers.***
4.13	rooms	The **rooms** are very spacious and the **rooms** are very comfortable.
2		***The broken furniture and dirty surfaces are a dead giveaway.***
2.96	furniture	The **furniture** is worn.
4		***Ideal for plane spotters and very close to the airport.***
2.76	airport	It is not close to the **airport**.

Vision-Language co-Learning

- Users comment about image features in the review text

Vision-Language co-Learning

- Visually Explainable Collaborative Filtering [13]

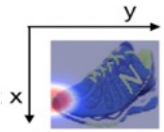
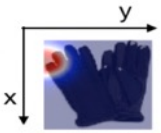
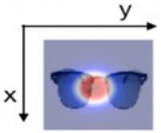
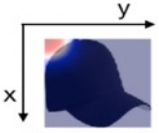
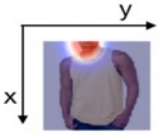


$$l_2 = \delta \sum_{(i,j)} \sum_{t=1}^{l_{ij}} \log\, p(w_{ij}^t | \boldsymbol{w}_{ij}^{1:t-1}, \boldsymbol{z}_{ij}^{t-1}) + (1-\delta)\left(\sum_{i \in \boldsymbol{u}} \sum_{j \in \boldsymbol{v}_+^i} \log \hat{y}_{ij} + \sum_{i \in \boldsymbol{u}} \sum_{j \in \boldsymbol{v}/\boldsymbol{v}_+^i} \log(1-\hat{y}_{ij}) \right) - \lambda ||\Theta||_F^2$$

[15] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha. "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation." In *ACM SIGIR*, pp. 765-774. 2019.

Vision-Language co-Learning

- Examples of visual and textual explanations

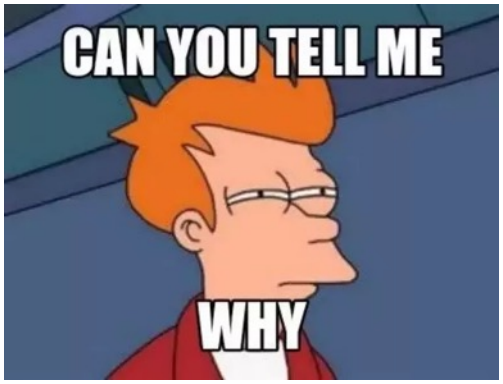
Good Cases		I loved about the previous generation and ***expanded the toe box a little to improve the fit.*** great buy, highly recommended.	
		They fit my stubby fingered hand pretty well. I bought the large and my hand measured 9.25&34 at the knuckles.	
		These sunglasses fit well and ***I like the design around the nose;*** they sit rather than dig like most other glasses can. The included pouch is great for keeping your glasses safe and scratch free.	
Bad Cases		The cap, which is made of a fairly heavy fabric, makes the head feel hot when worn for several hours in a warm gym or outside on a warm day. I, therefore, tend to wear it only when it is cold outside . -bi	
		The fabric is amazingly soft and the fit is perfect. I own several items from next level and will continue to add to my collection with different colors and styles. Amazing company, Amazing product.	

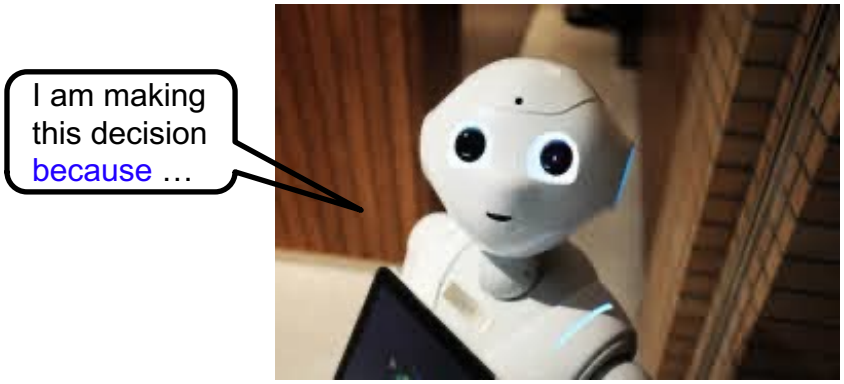
Some features are global and thus can not be aligned to a region of the image

Brief Summary

- Natural Language Explanation
  - An important way to foster human-machine understanding

- More controllable generation of natural language explanations
- Vision-Language alignment for generating explanations





Explainable AI for Nature-oriented Tasks

Two dimensions: Method and Application

Applications

Methods

	Human-centered Tasks	**Nature-oriented Tasks**
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

Why Explainable AI for Science Discovery

Motivating Example: Kepler's Law of Planetary Motion

We can **Obverse** it!

We can **Predict** it!

Tycho Brahe (1546-1610)
Demark astronomer

Good at astro-observation

Observed and recorded a lot of data about Mars movement.

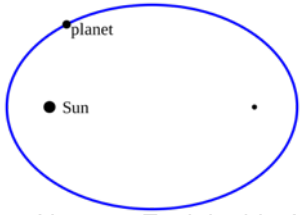
Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

Analyzed Tycho's data, and discovered the rules hidden in the data.
The "Kepler's laws of planetary motion":

$$\frac{\tau^2}{r^3} = K$$

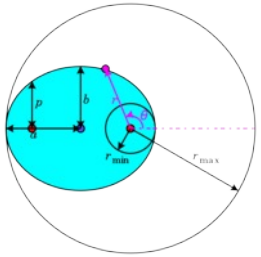
τ: period of circling around the sun, *r*: radius

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)



Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

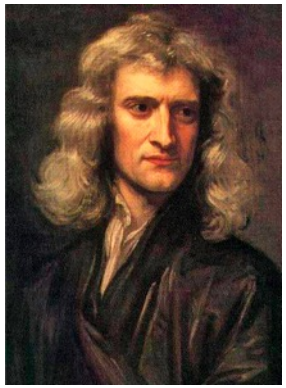
$$\frac{\tau^2}{r^3} = K$$



[16] Z. Li, J. Ji, Y. Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." arXiv:2111.12210.

Is the Story Over? No!

We can **Predict** it!



We **Understand** it!
We know **Why**!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

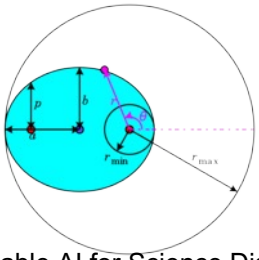
Analyzed Tycho's data, and discovered the rules hidden in the data.
The "Kepler's laws of planetary motion":

$$\frac{\tau^2}{r^3} = K$$

τ: period of circling around the sun, *r*: radius

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

$$\frac{\tau^2}{r^3} = K$$



Isaac Newton (1643-1727)
English mathematician, physicist, astronomer, theologian, and author.

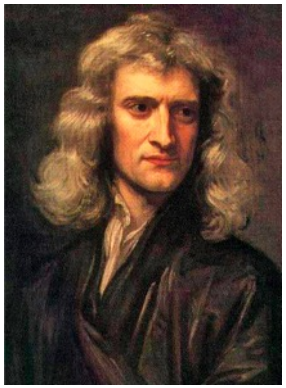
Proposed the Newton's law of universal gravitation + differential calculus:

Naturally derives out the Kepler's laws of planetary motion!

$$\frac{\tau^2}{r^3} = K \quad \text{is because} \quad F = G\frac{m_1 m_2}{r^2}$$

[16] Z. Li, J. Ji, Y. Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." arXiv:2111.12210.

Three Key Roles in AI/ML



Tycho Brahe (1546-1610)

Johannes Kepler (1571-1630)

Isaac Newton (1643-1727)

Data Collection

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

Model Learning

$$\frac{\tau^2}{r^3} = K$$

Model Interpretation

$$F = G\frac{m_1 m_2}{r^2}$$

[16] Z. Li, J. Ji, Y. Zhang. ”From Kepler to Newton: Explainable AI for Science Discovery.” arXiv:2111.12210.

What if Kepler had DL in the 16-17th Century?

We can **Obverse** it!

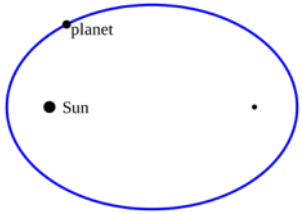
We can **Predict** it!

Tycho Brahe (1546-1610)
Demark astronomer

Good at astro-observation

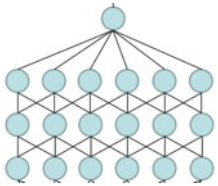
Observed and recorded a lot of data about Mars movement.

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)



Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

There could be some rules underlying the data.
I don't know what it is, but NN can fit any function.
So I'm going to train a NN to fit the data!



It fits the data pretty well!
I can make predictions!

$$\tau = some\ NN(r)$$

But wait: can this be called science discovery?
Science is not only about know HOW, but also know WHY!

Challenges in Modern Science Research

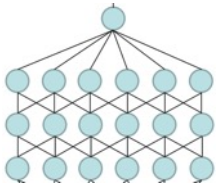
We can
Predict it!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

There must be some rules underlying the data.

I don't know what it is, but NN can fit any function.

So I'm going to train a NN to fit the data!



It fits the data pretty well!

I can make predictions!

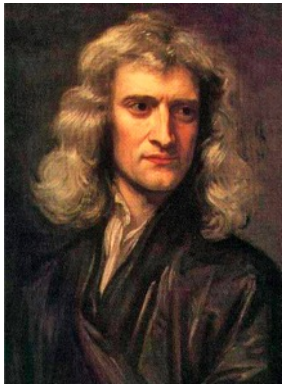
$$\tau = some\ NN(r)$$

But wait: can this be called science discovery?
Science is not only about know HOW, but also know WHY!

- However, manually analyzing data like Kepler did is almost impossible in modern science research
  - Since amount of data is huge
  - e.g., produced by astronomical telescope and particle colliders
- We need AI for automatic data processing and model learning

Challenges in Modern Science Research

We can **Predict** it!



We **Understand** it!
We know **Why**!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

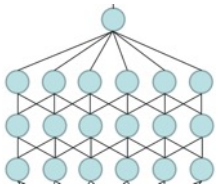
Isaac Newton (1643-1727)

Explainable AI (XAI) plays the role of Newton

There must be some rules underlying the data.

I don't know what it is, but NN can fit any function.

So I'm going to train a NN to fit the data!



It fits the data pretty well!

I can make predictions!

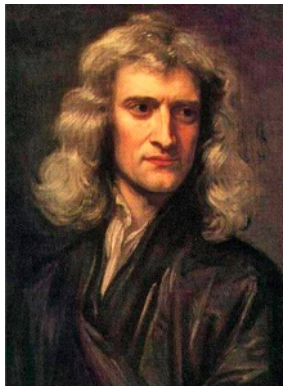
$$\tau = some\ NN(r)$$

Interpret and **explain** the learned (black-box) model, derive insightful discoveries/theories.

Help us better understand the nature.

But wait: can this be called science discovery?
Science is not only about know HOW, but also know WHY!

A New Paradigm of (AI-assisted) Science Discovery



Tycho Brahe (1546-1610)

Johannes Kepler (1571-1630)

Isaac Newton (1643-1727)

Data Collection

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

Almost automatic

Model Learning

$$\frac{\tau^2}{r^3} = K$$

Many available methods

Model Interpretation (XAI)

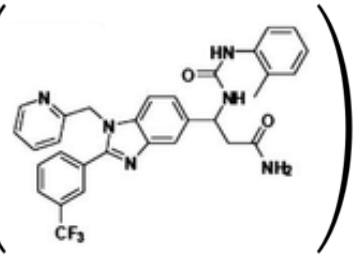
$$F = G\frac{m_1 m_2}{r^2}$$

Still needs much exploration

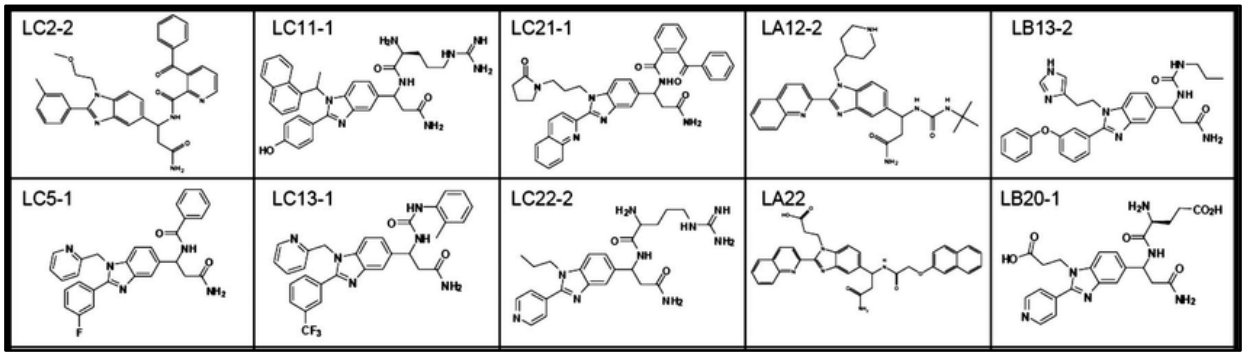
[16] Z. Li, J. Ji, Y. Zhang. ”From Kepler to Newton: Explainable AI for Science Discovery.” arXiv:2111.12210.

The Molecule Classification Problem

- Predicting the function of molecules
  - A fundamental problem in many chemistry/biological/medical research tasks, e.g., drug discovery
- Mathematically, molecule is a graph
  - Current approaches use Graph Neural Networks (GNN) for prediction
  - E.g., Predict if a molecule can pass the Blood-Brain Barrier (BBB)
  - A binary classification problem

GNN  = Yes / No

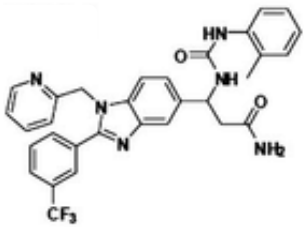
- However, we want to know **why** the model believes in the classification result

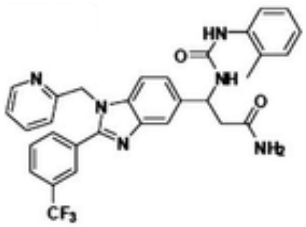


[17] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y, Li, Y. Zhang. " Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning." WWW 2022.

Explainable Graph Neural Networks

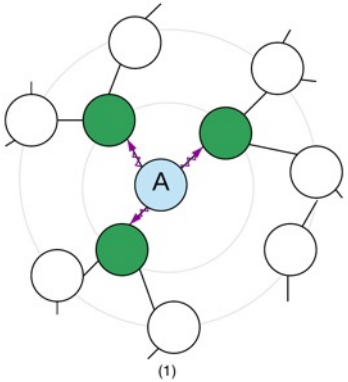
- Our goal is to develop Explainable Graph Neural Networks (XGNN)

GNN () = Yes / No

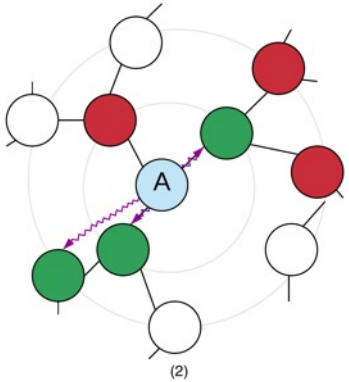
XGNN () = Yes / No + Explanation

Explainable Graph Attention Networks (XGAT)

- Explainable Graph Attention Networks (XGAT) [14]
- Key Idea
  - For each node A, identify its neighbors that are most important in predicting A's label
  - Those neighbors serve as the explanation of node A's prediction



Standard GNN: Each node aggregates information of its immediate neighbors



XGAT: Each node aggregates information from the relevant nodes (green) beyond just the immediate neighbors while ignoring irrelevant nodes (red)

$$\mathbf{h}_v^k = \sigma\left(\left[\mathbf{A}_k \cdot \mathrm{AGG}(\{\mathbf{h}_u^{k-1}, \forall u \in N(v)\}), \mathbf{B}_k \mathbf{h}_v^{k-1}\right]\right)$$

$$\min_{G_s} H(Y|G = G_s) = E_{Y|G_s}[\log P_\Phi(Y|G = G_s)]$$

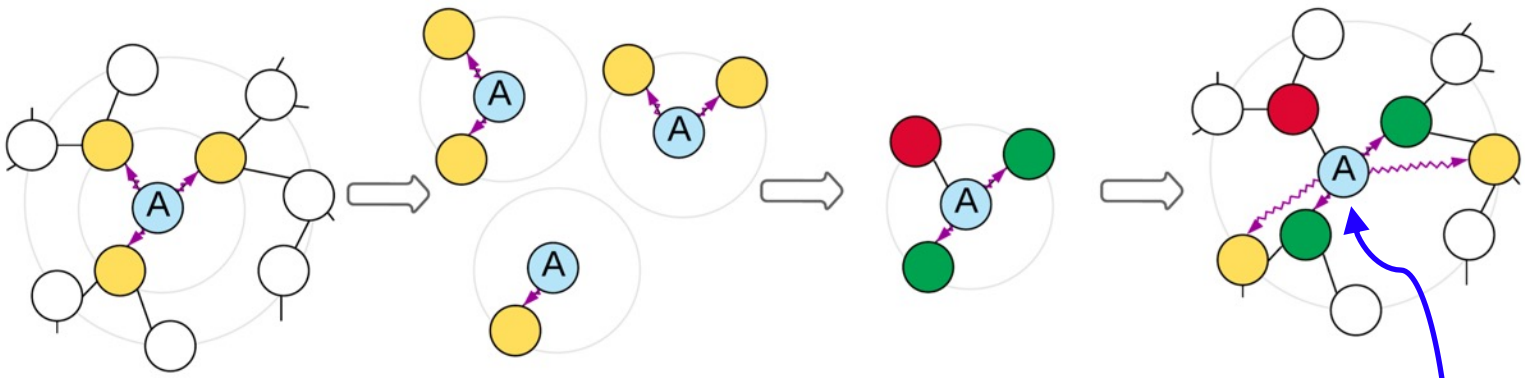
Intuition: Find a subgraph G_s, so that we are most sure about the model's prediction of A's label by aggregating information from nodes in G_s. G_s is the explanation, and "most sure about" is reflect by H – the entropy of prediction.

[17] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y, Li, Y. Zhang. " Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning." WWW 2022.
[18] D. Pham and Y. Zhang. "Explainable Graph Attention Networks". LOD 2021.

Explainable Graph Attention Networks (XGAT)

- Extracting Per-Node Explanations

```
for i = 1 to K: //loop for each hop of neighbors
    Run the following four steps;
```




```
Step 1: Aggregate information from current neighbors;
Step 2: Sample different neighbors of A and compute entropy
```

$$H(Y|G = G_s) = E_{Y|G_s}[\log P_\Phi(Y|G = G_s)]$$

```
Step 3: Keep neighborhood that minimizes entropy
```

$$\min_{G_s} H(Y|G = G_s) = E_{Y|G_s}[\log P_\Phi(Y|G = G_s)]$$

```
Step 4: Aggregate information from new neighborhood
```

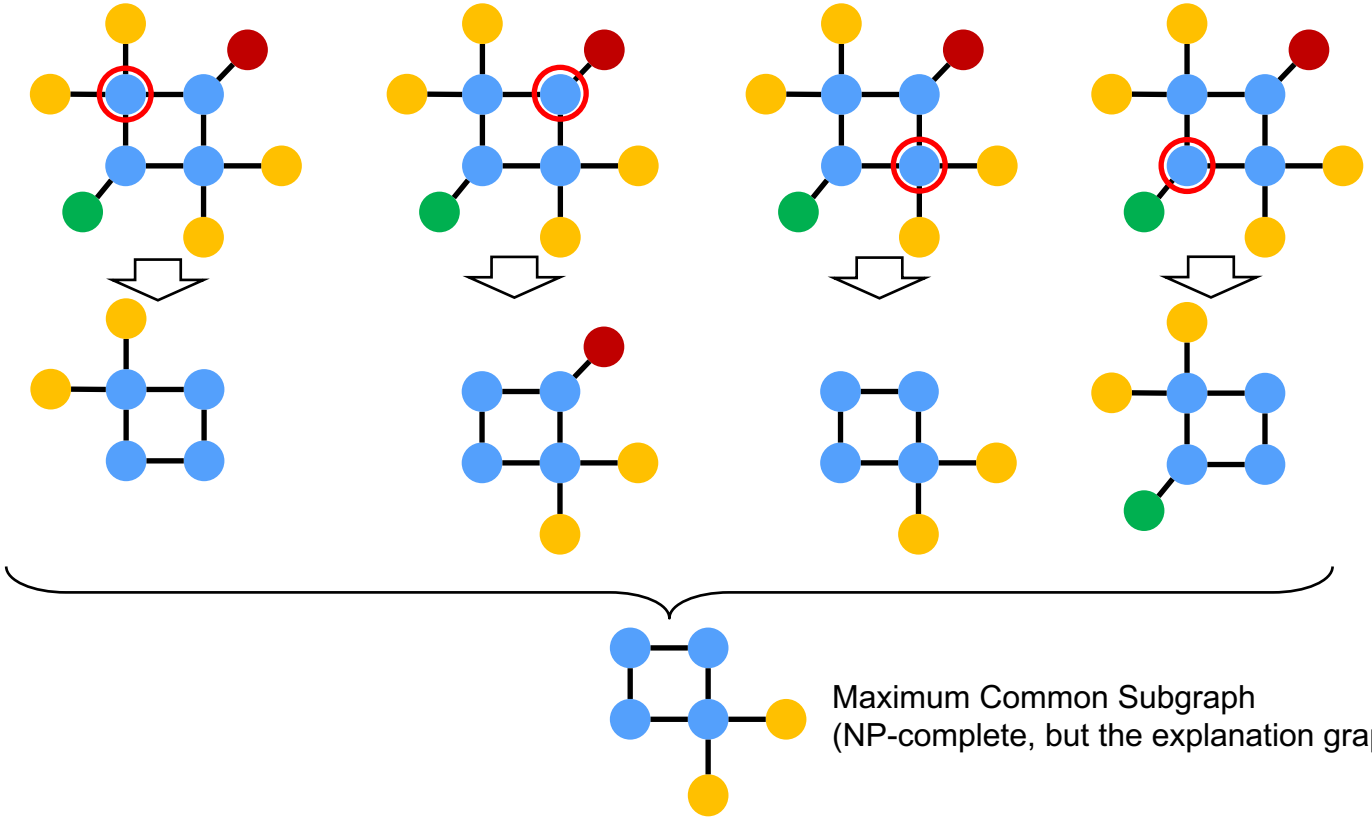
$$agg(\cdot) := h_v^k \leftarrow \sigma(MEAN(h_v^{k-1} \cup E(u,v) \cdot h_u^{k-1}, \forall u \in N(v))$$

Percentage of times that a neighbor node appeared in a sample graph G_s that reduced entropy, thus positively contribute to the prediction of node A

Neighbor nodes are weighted by Degree of Explainability $E(u,v) = \frac{p(u,v)}{M}$

Explainable Graph Attention Networks (XGAT)

- Extracting Graph-level Explanation
  - Explanation for the overall graph classification result
  - If XGAT(Graph) = Y, consider all nodes whose label predicted as Y
  - Extract per-node explanation



Maximum Common Subgraph
(NP-complete, but the explanation graphs are small)

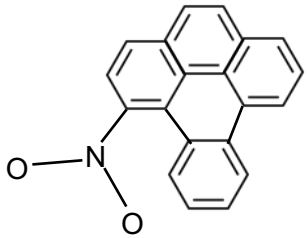
Explainable Graph Attention Networks (XGAT)

- Experiment on MUTAGS dataset
  - Task: Predict if a molecule is mutagenic (binary classification: mutagenic or non-mutagenic)
  - 4,337 molecule graphs labeled according to their mutagenic effect
  - Labeled ground-truth explanation: which part of the graph makes it mutagenic
    - Evaluate on node recall

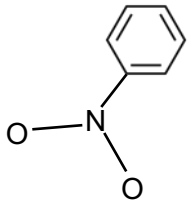
	Dataset		
Method	BA-SHAPES	BA-COMMUNITY	MUTAGS
Attention-based	87.5%	81.2%	76.9%
Gradient-based	89.0%	72.1%	80.5%
XGAT	**93.9%**	**88.5%**	**81.2%**

Our model also works on other dataset since the model is general purpose

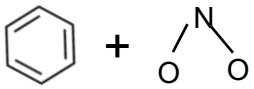
Case study:



Molecule Graph



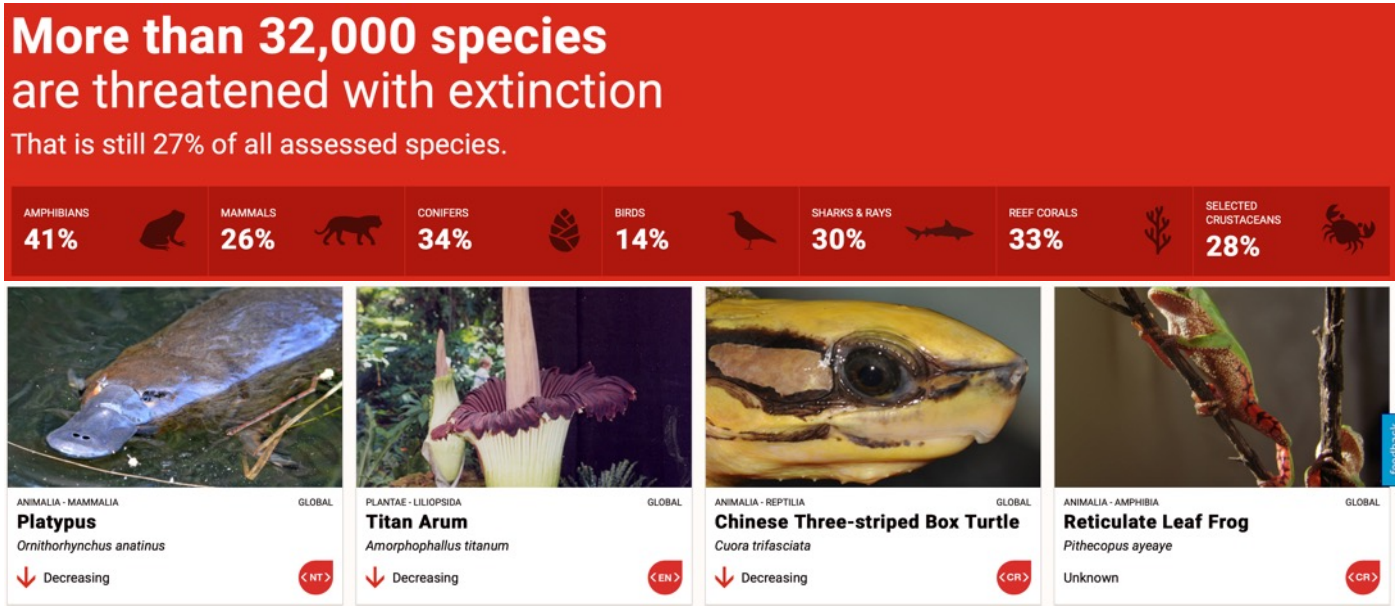
Explanation



Ground-truth: Benzene and NO_2 chemical groups (toxic)

Explainable AI for Biodiversity Conservation

- Task: Predict if a species is endangered or not [15]
  - An important nature-oriented task
  - A dynamic task: species that were not endangered may become endangered now, and vice versa
  - Needs dynamic monitoring and fast reaction
    - E.g., IUCN Red List maintains the status for animal species
    - Critically Endangered, Endangered, Extinct, Extinct in the Wild, Least Concern, Low Risk, Threatened, Vulnerable



From the IUCN (International Union for Conservation of Nature) Red List of Threatened Species https://www.iucnredlist.org

[19] M. Mukadam, M. Jayaram, and Y. Zhang. “A Representation Learning Approach to Animal Biodiversity Conservation”. COLING 2020.

Explainable AI for Biodiversity Conservation

- Machine learning may help as an assistive tool
  - Why Machine Learning may work?
  - Intuition: Species become endangered mostly because habitat destruction due to human activities
  - If we know one species in a habitat is endangered, other species in the same habitat may too
- Habitat (and other useful information) can be found in Wikipedia
  - Information is dynamic/up-to-date due to real-time edits

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit

Not logged i

Article Talk Read Edit View histo

White-bellied hummingbird

From Wikipedia, the free encyclopedia

The **white-bellied hummingbird** (*Elliotomyia chionogaster*) is a species of hummingbird in the family Trochilidae. It is found at forest edge, woodland, scrub and gardens in the Andes, ranging from northern Peru south through Bolivia to north-western Argentina. There are also lowland populations in Santa Cruz, Bolivia, and Mato Grosso, Brazil.

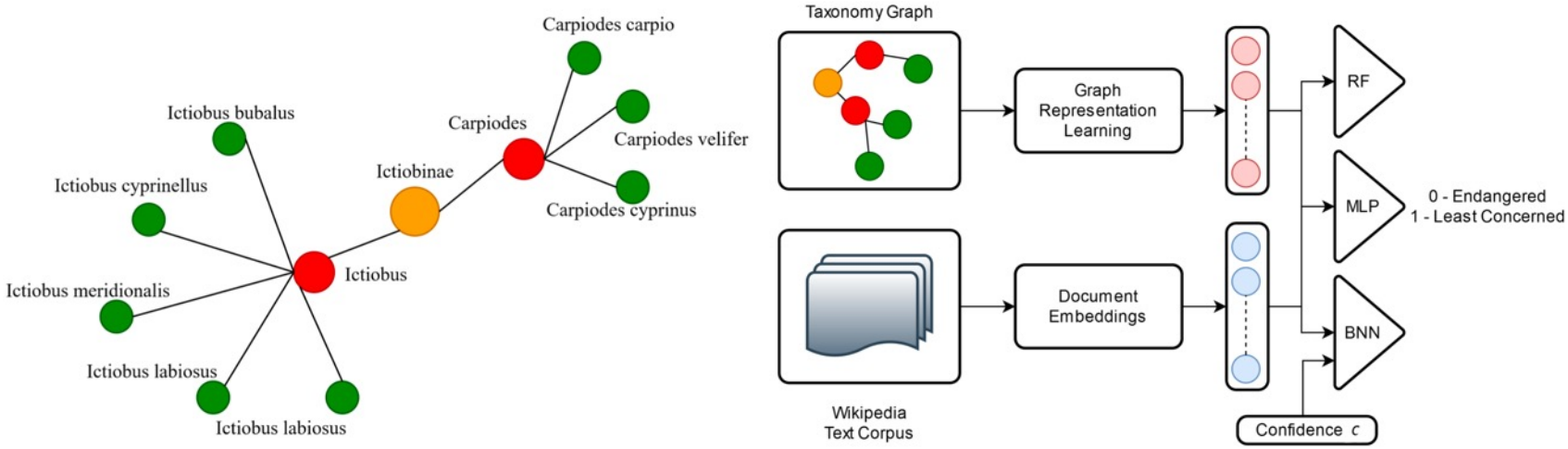
Description [edit]

Its upperparts are green and its underparts are white. Unlike other similar hummingbirds in its range (for example the green-and-white hummingbird), the basal half of the inner webs of the rectrices are white, but this is typically only visible from below.

Explainable AI for Biodiversity Conservation

- Wikipedia text is not enough
  - Due to Interspecific Competition, one species get endangered may imply another competitor species get more populated

- Solution: Graph-Text Co-Learning for Animal Biodiversity Conservation [15]
  - Animal taxonomy graph shows the relationship between species



[19] M. Mukadam, M. Jayaram, and Y. Zhang. “A Representation Learning Approach to Animal Biodiversity Conservation”. COLING 2020.

Explainable AI for Biodiversity Conservation

Metric	Value
Total animal species	45,170
Endangered animal species	10,947
Least-concern animal species	27,053
Data-Deficient species	7,170
Average length of Wikipedia documents in training corpus (number of words)	146
Documents with length more than average length	13,970
Documents with length less than average length	31,200
Documents that explicitly contain Red List status information	14,253
BNN training data points	14,083
BNN test data points ($c \geq 0.75$)	3,521

Dataset statistics (data collected from Wikipedia, IUCN, and ITIS)

Model	F1 Score	-AUC
RF w/ node2vec	0.862	0.784
RF w/ doc2vec	**0.869**	0.820
RF w/ node2vec + doc2vec	0.860	**0.827**
MLP w/ node2vec	0.843	0.729
MLP w/ doc2vec	**0.886**	0.864
MLP w/ node2vec + doc2vec	0.885	**0.873**
BNN w/ node2vec + doc2vec + $c \geq 0.75$	0.856	0.868
BNN w/ node2vec + doc2vec + $c \geq 0.9$	**0.889**	**0.911**

Prediction accuracy

Nycticryphes semicollaris - The South American painted-snipe (Nycticryphes semicollaris), or lesser painted-snipe, is a shorebird in the family Rostratulidae. There are two other species in its family, the Australian painted-snipe and the greater painted-snipe. Measurements: 19–23 cm in length; 65–86 g in weight. Vocalizations: A hoarse, hissing “wee-oo” has been recorded from birds in captivity. Distribution and habitat: The species is found in the southern third of South America, from southern Brazil, Paraguay, and Uruguay to Chile and Argentina. It inhabits lowland freshwater wetlands, including wet grasslands. Breeding: South American painted-snipes are monogamous and breed semi-colonially. The nest is a shallow cup on the ground in a wetland, with a clutch of 2-3 eggs. Breeding has been recorded mainly from July to February. Feeding: The South American painted-snipe is omnivorous, feeding by probing in mud and shallow water for small animals and seeds, often at dusk.

Attention-based Explanation

IUCN: International Union for Conservation of Nature (for ground-truth label)
ITIS: Integrated Taxonomic Information System (for Animal taxonomy graph) https://www.itis.gov

Summary

- Explainable AI
  - Two dimensions: Methods and Application

Applications →

Methods ↓

	Human-centered Tasks	Nature-oriented Tasks
Neural Logic Reasoning	Solving Equations, Recommendation	
Neural Symbolic Reasoning	Recommendation	
Knowledge Graph Reasoning	Search, Recommendation, Dialog, Fairness	
Natural Language Generation	Recommendation	
Explainable GNN		Molecular Classification
Graph-Language Alignment		Biodiversity Conservation

  - Human-centered tasks
    - Explainable AI enhanced better communication and understanding between human and machine
  - Nature-oriented tasks
    - Explainable AI assists scientists for more efficient and informed science discoveries

Some Other Research

- Fairness of Machine Learning
  - Towards Long-term Fairness in Recommendation [20]
  - User-oriented Fairness in Recommendation [10]
  - Towards Personalized Fairness based on Causal Notion [35]
  - Fairness-aware Explainable Recommendation over Knowledge Graphs [9]

- Economics-Inspired Machine Learning for Recommender Systems
  - Learning user's Risk Preferences based on Prospect Theory [22]
  - Learning Weighted Expected Utility for Recommendation [23]
  - Learning and Maximizing Marginal Utility for Recommendation [24]

- Echo Chambers
  - Analyzing the Feedback Loop and Echo Chambers in Recommender Systems [25]

- From System 1 to System 2 AI based on Neural-Symbolic Learning and Reasoning
  - Neural Logic Reasoning [4]
  - Neural Collaborative Reasoning [28]
  - Graph Collaborative Reasoning [29]

Yongfeng Zhang
Department of Computer Science, Rutgers University
yongfeng.zhang@rutgers.edu

References

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

[2] Fei et al. "Screening small-molecule compound microarrays for protein ligands without fluorescence labeling with a high-throughput scanning microscope." *Journal of biomedical optics* 15, no. 1 (2010): 016018.

[3] Marcus G, Davis E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon; 2019 Sep 10.

[4] S. Shi, H. Chen, W. Ma, J. Mao, M. Zhang, and Y. Zhang. "Neural Logic Reasoning." In CIKM, pp. 1365-1374. 2020.

[5] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang. "Reinforcement knowledge graph reasoning for explainable recommendation." In SIGIR, pp. 285-294. 2019.

[6] Z. Fu, Y. Xian, Y. Zhang, Y. Zhang. "Tutorial on Conversational Recommendation Systems." In 14th ACM Conference on Recommender Systems. 2020.

[7] Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. de Melo, Y. Zhang. "HOOPS: Human-in-the-Loop Graph Reasoning for Conversational Recommendation." SIGIR 2021.

[8] Z. Fu, Y. Xian, S. Geng, G. de Melo, Y. Zhang. " Popcorn: Human-in-the-loop Popularity Debiasing in Conversational Recommender Systems." CIKM 2021.

[9] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, G. de Melo. "Fairness-Aware Explainable Recommendation over Knowledge Graphs." In SIGIR. *2020.*

[10] Y. Li, H. Chen, Z. Fu, Y. Ge and Y. Zhang. "User-oriented Group Fairness In Recommender Systems." WWW 2021.

[11] H. Chen, X. Chen, S. Shi, Y. Zhang. "Generate natural language explanations for recommendation." In Proceedings of SIGIR'19 Workshop on Explainable Recommendation and Search (EARS). ACM. 2019.

[12] L. Li, Y. Zhang, L. Chen. "Generate neural template explanations for recommendation." In CIKM, pp. 755-764. 2020.

[13] L. .Li, Y. Zhang, L. Chen. "Personalized Transformer for Explainable Recommendation". ACL 2021.

[14] S. Geng, Z. Fu, Y. Ge, L. Li, G. de Melo, Y. Zhang. "Improving Personalized Explanation Generation through Visualization". ACL 2022.

[15] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha. "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation." In *ACM SIGIR*, pp. 765-774. 2019.

[16] Z. Li, J. Ji, Y. Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." arXiv:2111.12210.

[17] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y, Li, Y. Zhang. " Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning." WWW 2022.

[18] D. Pham and Y. Zhang. "Explainable Graph Attention Networks". LOD 2021.

[19] M. Mukadam, M. Jayaram, and Y. Zhang. "A Representation Learning Approach to Animal Biodiversity Conservation". COLING 2020.

[20] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, and Y. Zhang. "Towards Long-term Fairness in Recommendation". In WSDM 2021.

[21] G. Wang, Y. Zhang, Z. Fang, S. Wang, F. Zhang, D. Zhang. “FairCharge: A Data-Driven Fairness-Aware Charging Recommendation System for Large-Scale Electric Taxi Fleets”. UbiComp.

[22] Y. Ge, S. Xu, S. Liu, Z. Fu, F. Sun and Y. Zhang. “Learning Personalized Risk Preferences for Recommendation”. In SIGIR, 2020.

[23] Z. Xu, Y. Han, Y. Zhang, Q. Ai. “E-commerce Recommendation with Weighted Expected Utility”. In CIKM, 2020.

[24] Y. Ge, S. Xu, S. Liu, S. Geng, Z. Fu and Y. Zhang. “Maximizing Marginal Utility per Dollar for Economic Recommendation”. In WWW 2019.

[25] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou and Y. Zhang. “Understanding Echo Chambers in E-commerce Recommender Systems”. SIGIR 2020.

[26] Y. Zhang. “Problem Learning: Towards the Free Will of Machines”. arXiv:2109.00177.

[27] S. Xu, Y. Ge, Y. Li, Z. Fu, X. Chen, Y. Zhang. “Causal Collaborative Filtering”. arXiv:2102.01868.

[28] H. Chen, S. Shi, Y. Li and Y. Zhang. “Neural Collaborative Reasoning”. WWW 2021.

[29] H. Chen, Y. Li, S. Shi, S. Liu, H. Zhu and Y. Zhang. “Graph Collaborative Reasoning”. WSDM 2022.

[30] S. Geng, Z. Fu, J. Tan, Y. Ge, G. de Melo and Y. Zhang. “Path Language Modeling over Knowledge Graphs for Explainable Recommendation”. WWW 2022.

[31] Y. Li, H. Chen, J. Tan and Y. Zhang. “Causal Factorization Machine for Robust Recommendation”. JCDL 2022.

[32] S. Liu, F. Sun, Y. Ge, C. Pei and Y. Zhang. “Variation Control and Evaluation for Generative Slate Recommendations”. WWW 2021.

[33] Z. Li, J. Ji, Z. Fu, Y. Ge, S. Xu, C. Chen and Y. Zhang. “Efficient Non-Sampling Knowledge Graph Embedding”. WWW 2021.

[34] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen and Y. Zhang. “Counterfactual Explainable Recommendation”. CIKM 2021.

[35] Y. Li, H. Chen, S. Xu, Y. Ge and Y. Zhang. “Towards Personalized Fairness based on Causal Notion”. SIGIR 2021.

[36] S. Liu, S. Xu, W. Yu, Z. Fu, Y. Zhang and A. Marian. “FedCT: Federated Collaborative Transfer for Recommendation”. SIGIR 2021.

[37] Y. Zhang and X. Chen. ”Explainable Recommendation: A Survey and New Perspectives”. Foundations and Trends in Information Retrieval. 2022.

[38] Y. Xian, Z. Fu, H. Zhao, Y. Ge, X. Chen, Q. Huang, S. Geng, Z. Qin, G. de Melo, S. Muthukrishnan, Y. Zhang. “CAFE: Coarse-to-Fine Neural Symbolic Reasoning for Explainable Recommendation”. CIKM 2020.