NLP for Recommender Systems

CPSC 677, December 9, 2021

Sumner Magruder

Overview

Recommender System Crash Course

Graph Theory Crash Course

Required Papers

Additional Papers

Goals

If my talk is successful you will walk away with:

- Conceptual understanding of common approaches to Recommender Systems (RecSys)
- Understanding of two of the most popular types of RecSys (Content-based and Collaborative Filtering)
- Approaches for improving RecSys (augmentation via metadata - often text)
- Conceptual understanding of graph curvature analysis (since RecSys is primarily based around User-Interaction-Item events)

Recommender System Crash Course

Required Paper 2 and Additional Papers 3&4

What is a Recommender System (RecSys)

Beel[1] et al.

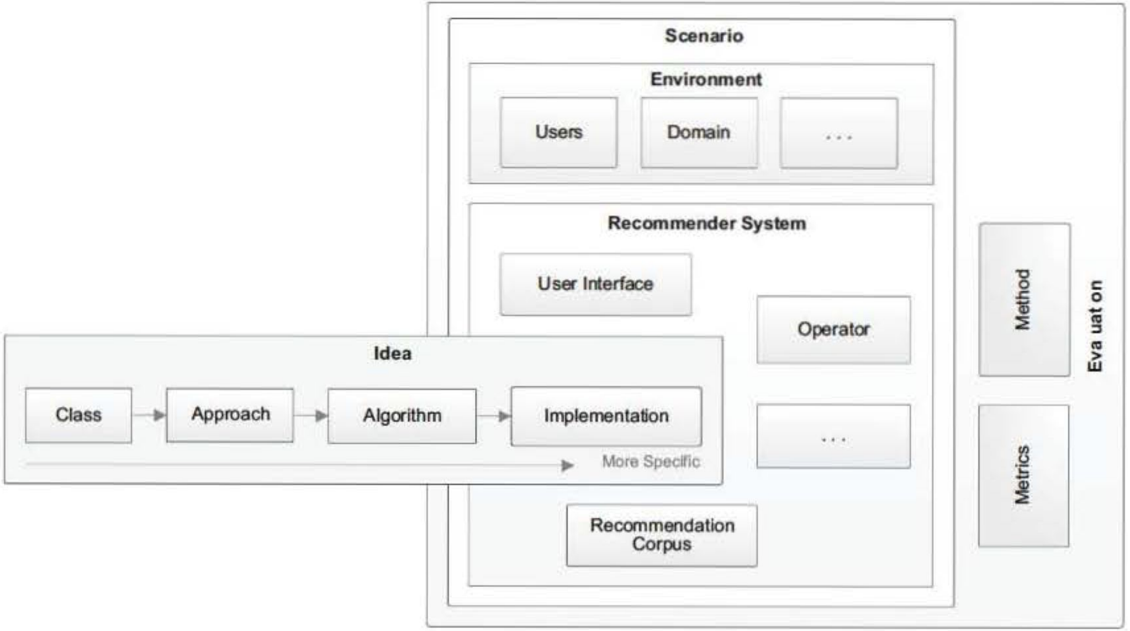
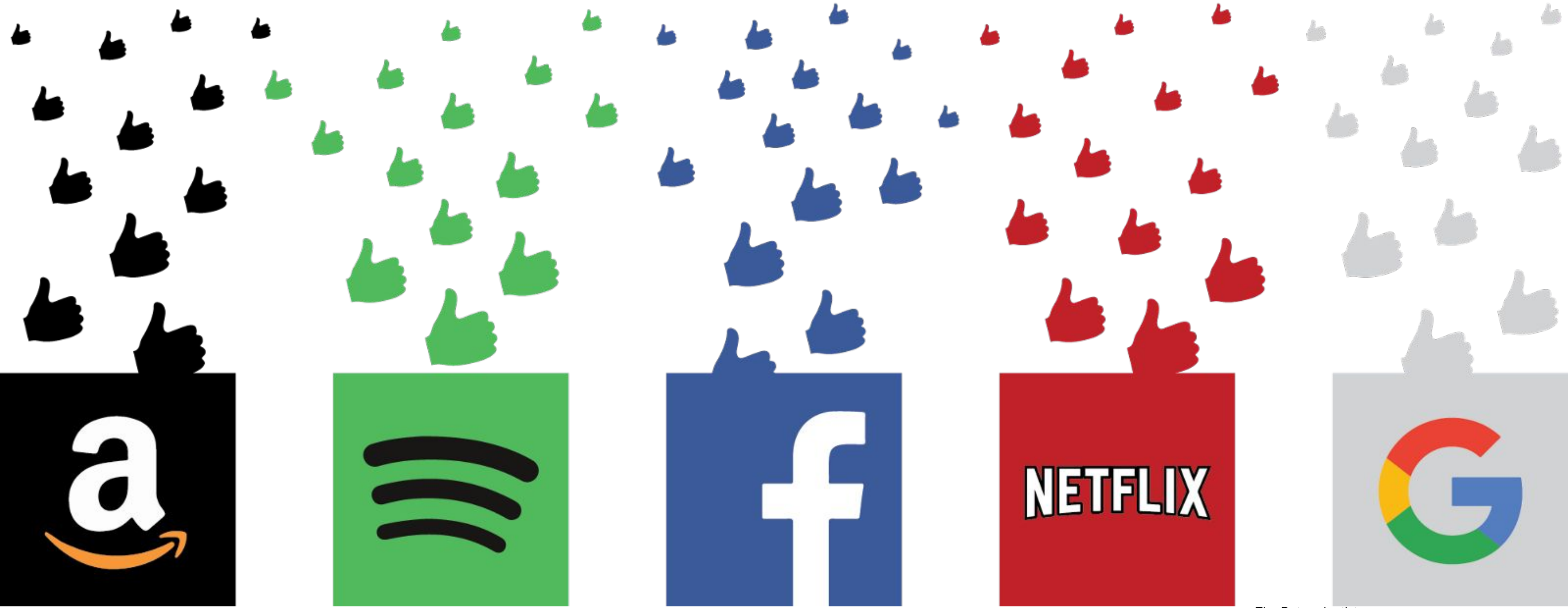


Fig. 4 Illustration of recommendation system terminology and concepts

Out in the Wild



The Data scientist

Out in the Wild (cont)

- Dating Apps (Hinge, OkCupid)
- Finance (eTraders)
- News and various “Feeds” (Twitter, StumbleUpon)
- Activities (Yelp, xbox, StubHub, etc)
- Food (from groceries like InstaCart to menus like GrubHub)

Welcome to the Recommended World

(old but classic example--->)


Forbes

Feb 16, 2012, 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Tech

Welcome to The Not-So Private Parts where technology & privacy collide

Follow

Forbes, Kashmir Hill

Why Care about Recommender Systems

- You can't escape it
    - 75% of Netflix watches
    - 70% of YouTube
    - 35% of Amazon sales

(according to McKinsley)t

Questions

- Does anyone here have a recommender system they use? Or a preference against competitors? (YouTube Music vs Spotify vs Apple Music vs Pandora, etc).

(according to McKinsley)t

We know they work, but what is “good?”

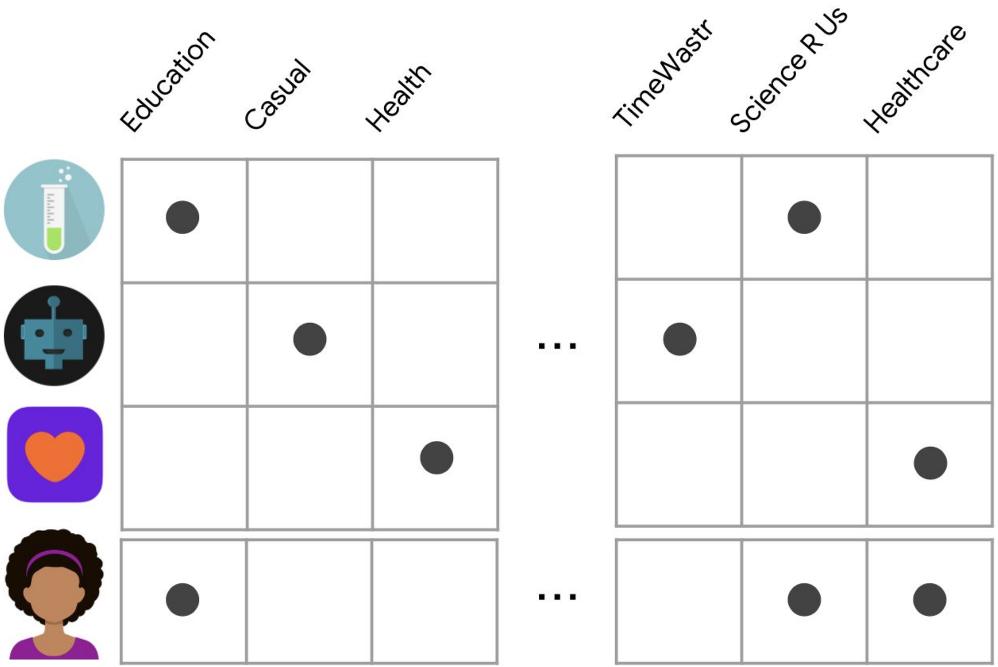
- User Studies
- Online Evaluation
- Offline Evaluation (p@N)
- Coverage
  - How much is seen by system?

Approaches to Recommendation

1. Stereotyping
    a. Exactly as it sounds. Once stereotypes made, easy and cheap to use.
2. **Content-based Filtering**
    a. User recs determined via given user's interactions with items
3. **Collaborative Filtering**
    a. User recs determined via other similar users
4. Co-Occurrence
    a. User recs determined via pairs e.g. Amazon's "Customers Who Bought This Item Also Bought…"
5. Graph-based
    a. User recs from graph metrics from domain / derived edges
6. Global Relevance
    a. One-fits-all / popularity contest
7. Hybrid
    a. Self explanatory

Content Based

- Most abundant for Academic Recommendations
- Based only on given user
- Domain knowledge required
- Struggles to diversify user



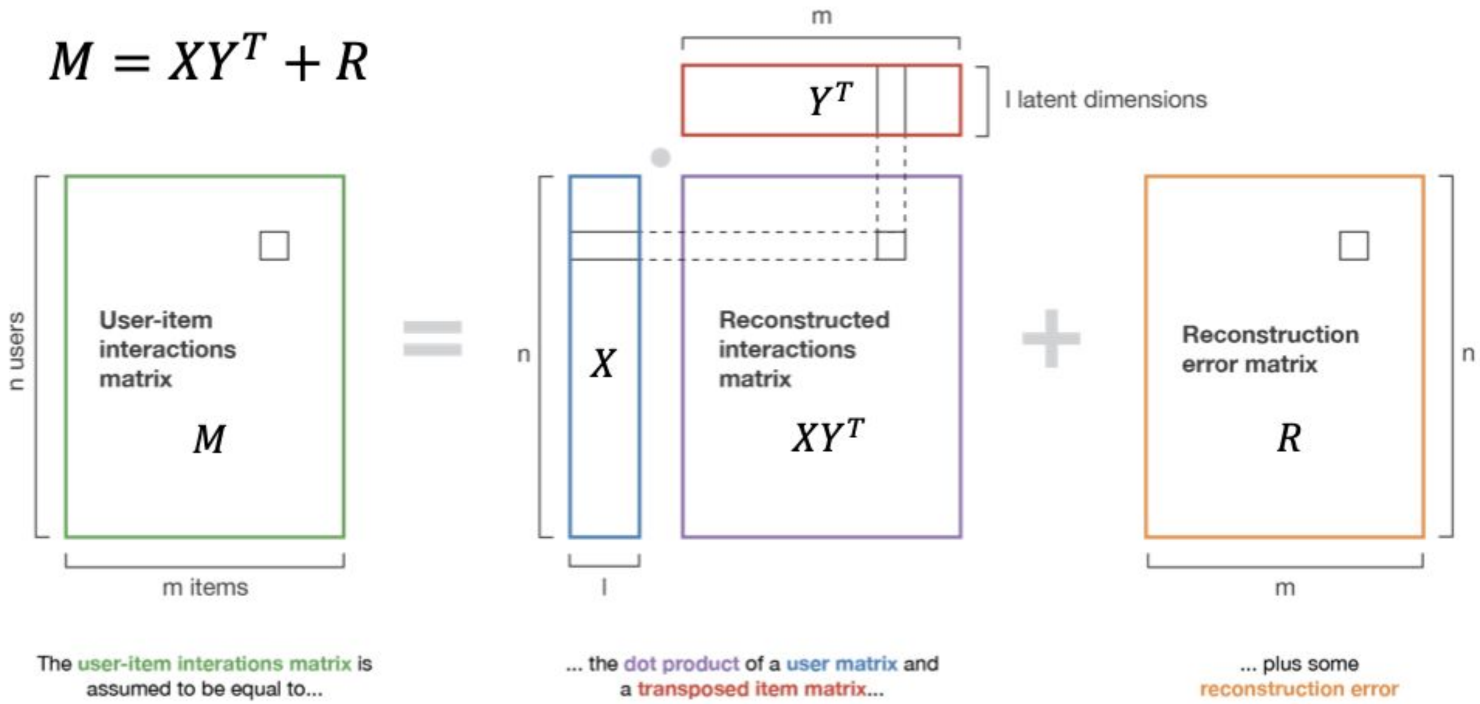
https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

Collaborative Filtering and the User-Item Matrix

- Amongst the most common and abundant approach
- Can be constructed explicitly or implicitly (who purchased what)
- Often sparse
- Common via Low Rank Approximation

Matrix Factorization

$$M = XY^T + R$$

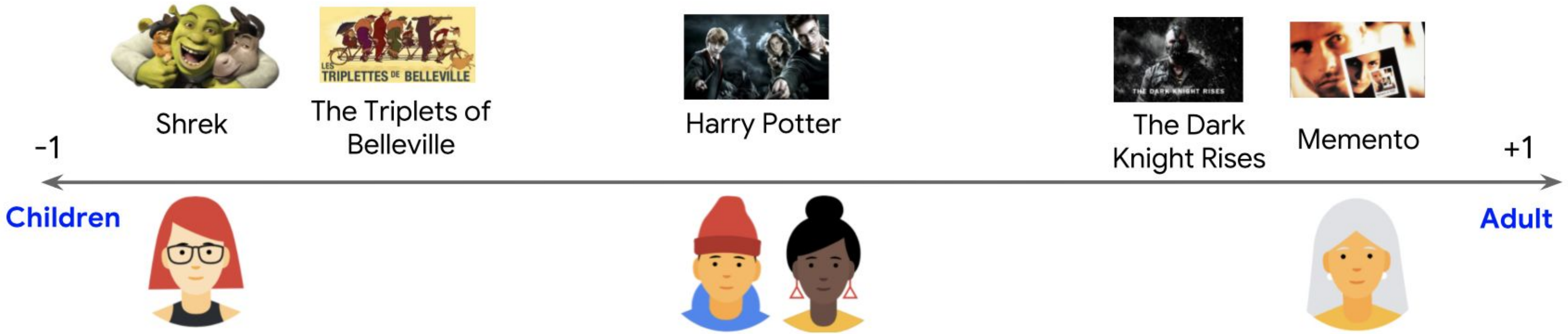


Jason Yanglu

Movie Example

Movie	Rating	Description
The Dark Knight Rises	PG-13	Batman endeavors to save Gotham City from nuclear annihilation in this sequel to The Dark Knight, set in the DC Comics universe.
Harry Potter and the Sorcerer's Stone	PG	A orphaned boy discovers he is a wizard and enrolls in Hogwarts School of Witchcraft and Wizardry, where he wages his first battle against the evil Lord Voldemort.
Shrek	PG	A lovable ogre and his donkey sidekick set off on a mission to rescue Princess Fiona, who is emprisoned in her castle by a dragon.
The Triplets of Belleville	PG-13	When professional cycler Champion is kidnapped during the Tour de France, his grandmother and overweight dog journey overseas to rescue him, with the help of a trio of elderly jazz singers.
Memento	R	An amnesiac desperately seeks to solve his wife's murder by tattooing clues onto his body.

https://developers.google.com/machine-learning/recommendation/overview/candidate-generation



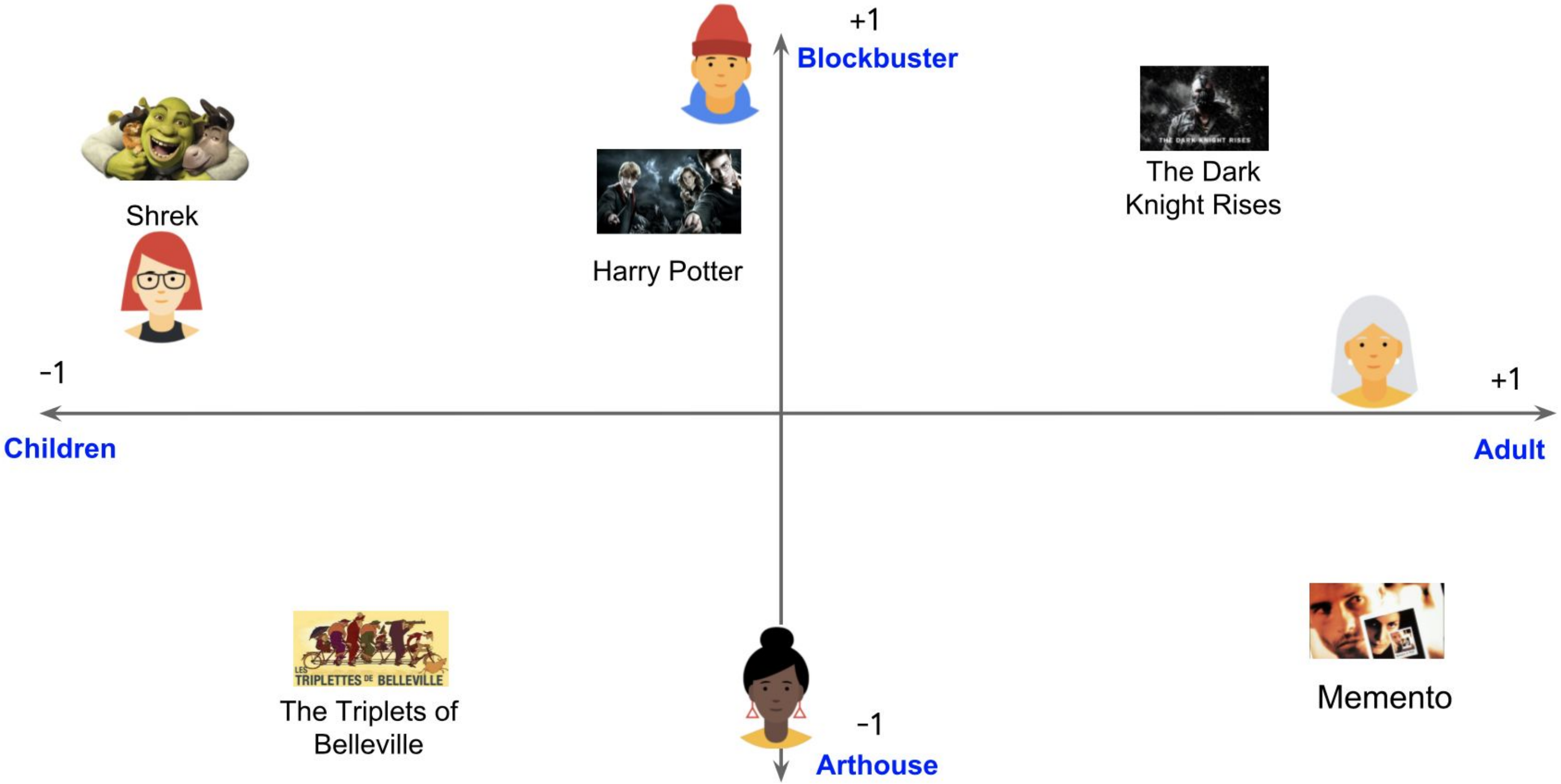
https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

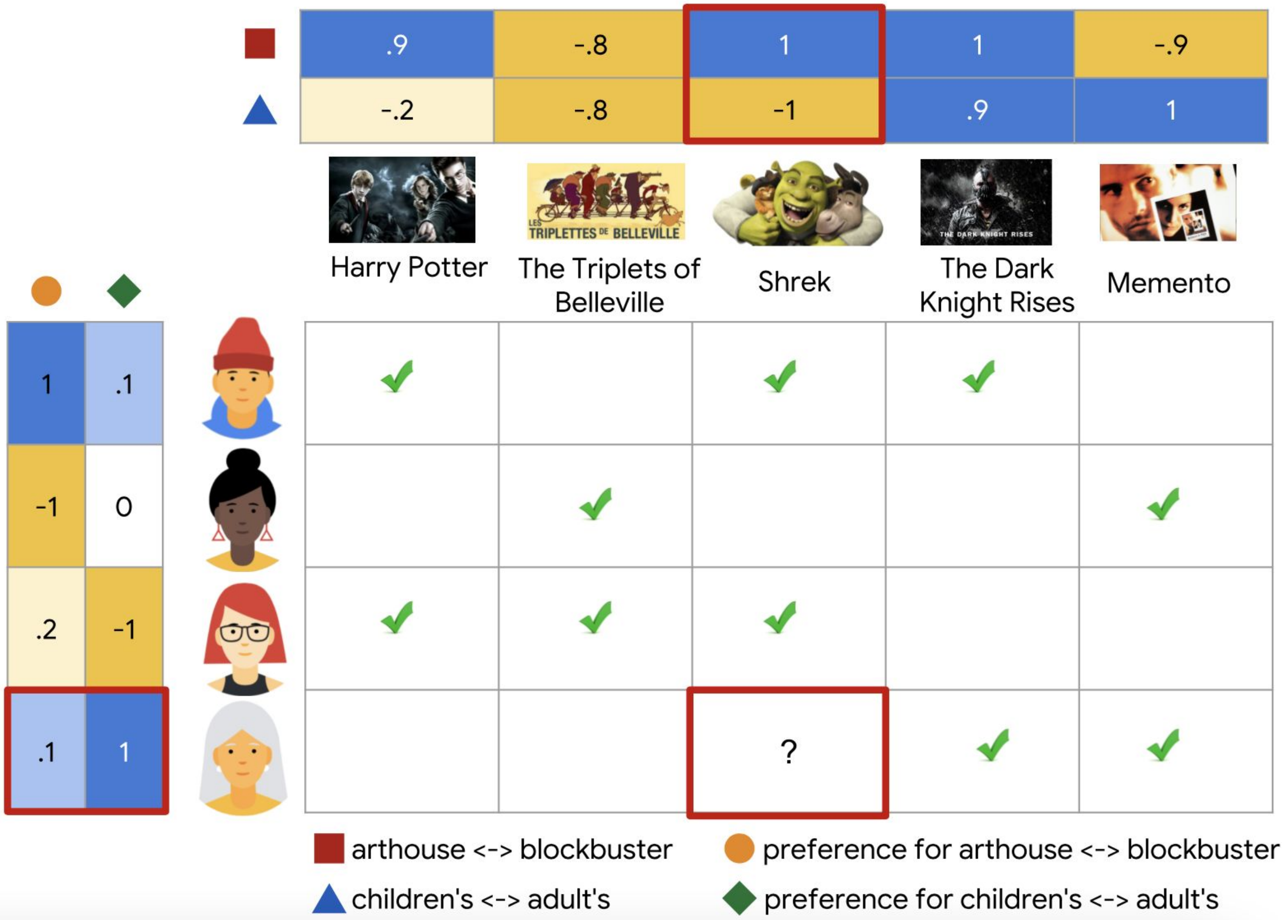
◆	▲	Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
		-.2	-.8	-1	.9	1
.1		✓		✓	✓	
0			✓			✓
-1		✓	✓	✓		
1					✓	✓

▲ children's <-> adult's

◆ preference for children's <-> adult's



https://developers.google.com/machine-learning/recommendation/overview/candidate-generation



https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

Matrix Factorization

	Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
User 1	✔		✔	✔	
User 2		✔			✔
User 3	✔	✔	✔		
User 4				✔	✔

≈

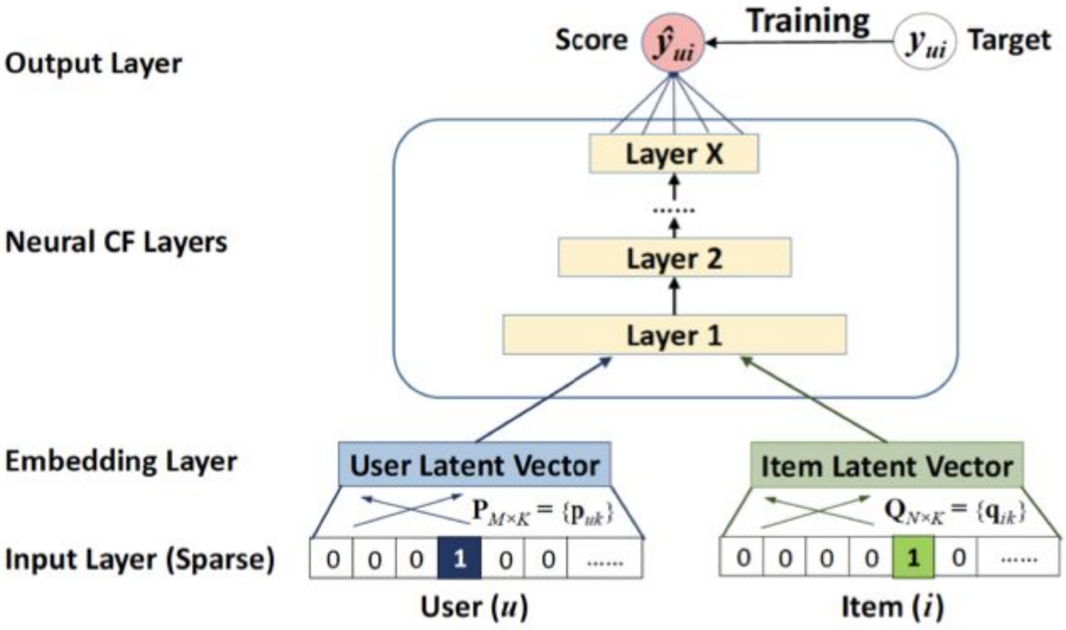
.9	-1	1	1	-.9
-.2	-.8	-1	.9	1

1	.1
-1	0
.2	-1
.1	1

.88	-1.08	0.9	1.09	-0.8
-0.9	1.0	-1.0	-1.0	0.9
0.38	0.6	1.2	-0.7	-1.18
-0.11	-0.9	-0.9	1.0	0.91

https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

Neural Matrix Factorization



https://slidetodoc.com/recommender-systems-twenty-years-of-research-lior-rokach/

Questions

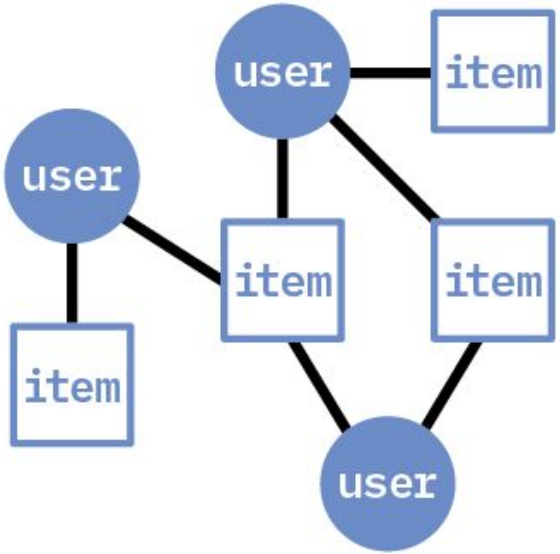
- How to attenuate RecSys Focus?
  - Don’t always need “similar” items (e.g. two types of chocolate chips, instead flour and eggs to bake cookies)
- How to ensure Diversity?
  - “Bubble” feed (cough cough YouTube)
- Who matters more?
  - Operator or User?

You will not escape User-Item Graph Factorization

(when talking about Recommender Systems)

1. Harry Potter and the Chamber of Secrets
   —because you liked Sorcerer's Stone
2. Lord of the Rings: Return of the King
   —because you liked Fellowship of the Ring
3. Game of Thrones
   —because you liked Fellowship of the Ring

Content-based



Collaborative Filtering

https://session-based-recommenders.fastforwardlabs.com

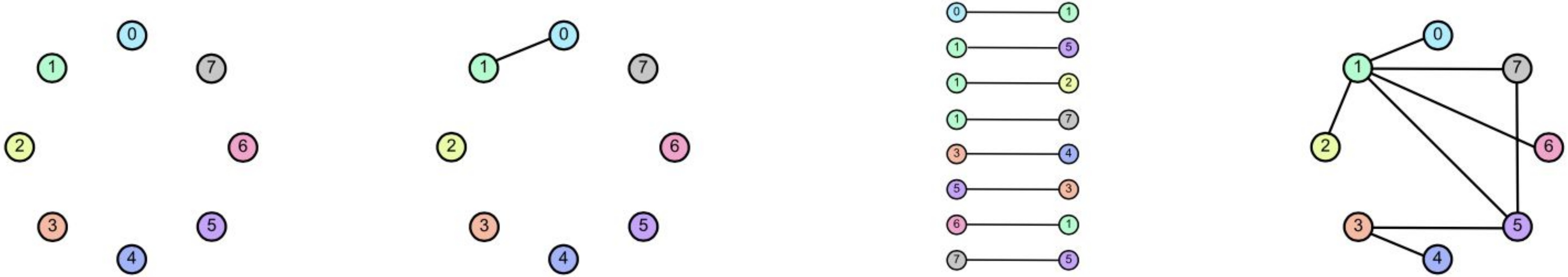
Augmenting the User-Item Graph with Textual Similarity Models

Required Paper 1

Graph Theory Crash Course

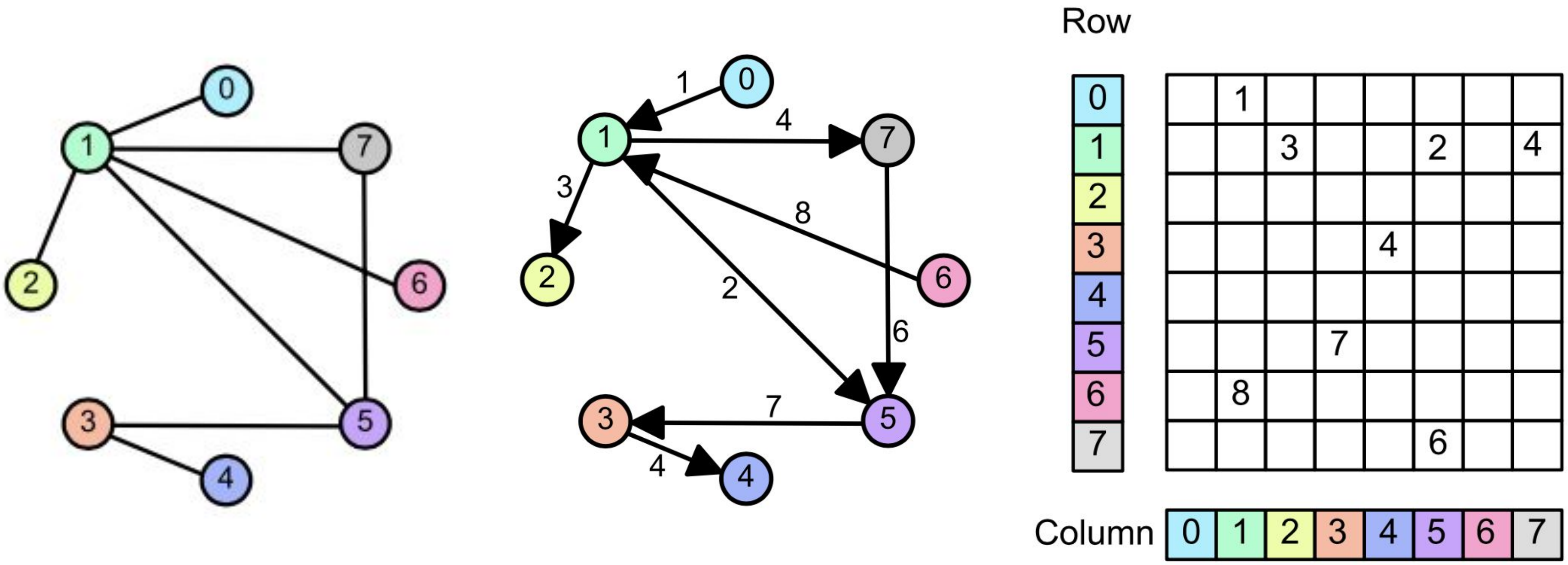
Graph to Matrix

- Review Adjacency Matrices as a Graph Structure

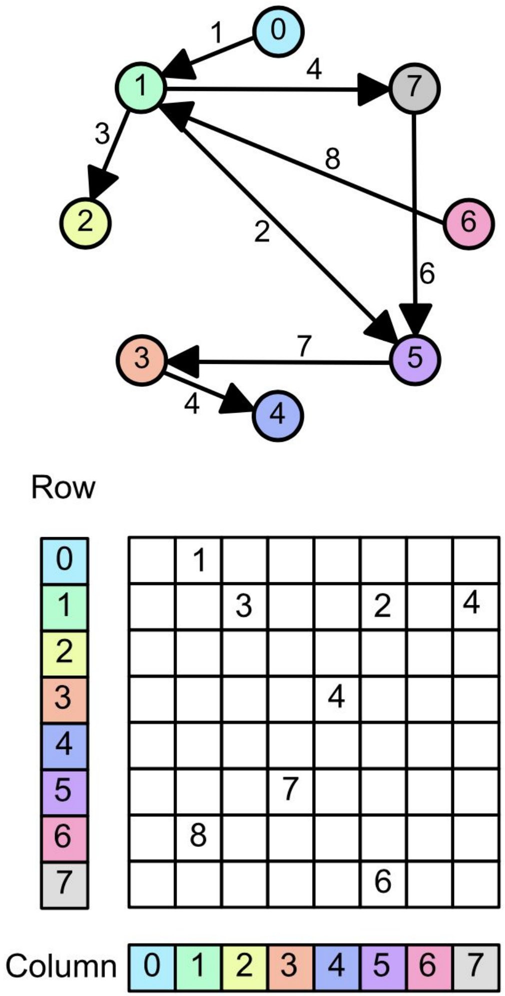


(a) A vertex set $|\mathbf{V}|$ (b) An edge $0 \bullet\!\!-\!\!\bullet 1$. (c) An edge set $|\mathbf{E}|$ (d) A graph $\mathscr{G}^U$

Matrix to Graph

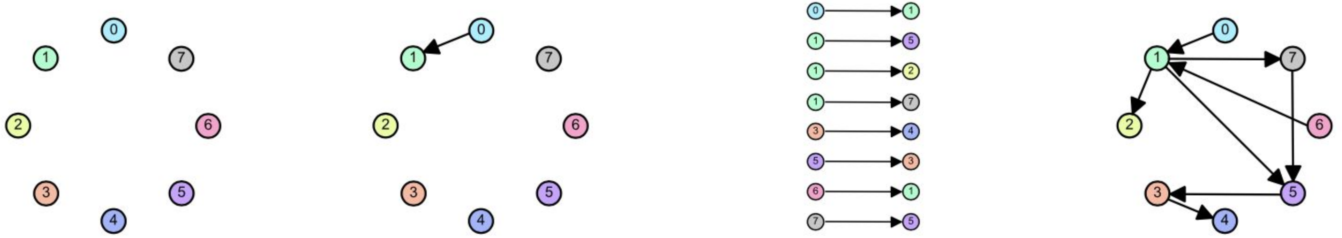


GRAND, Sumner Magruder



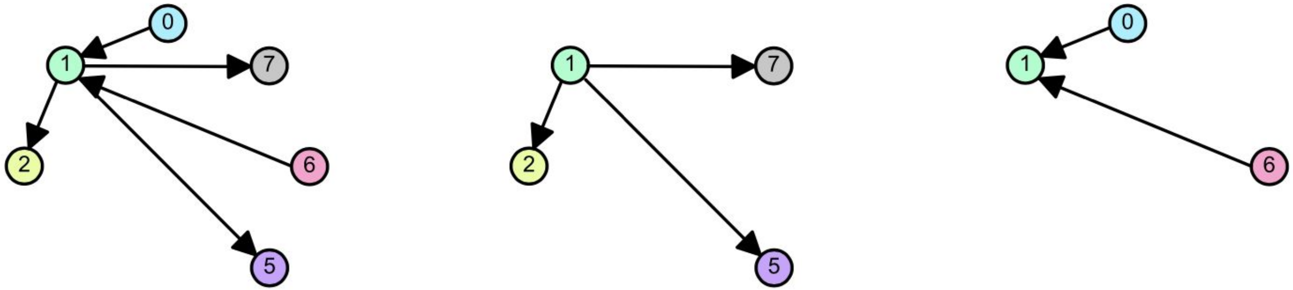
Some methods use directed graphs, but most use undirected graphs.

https://developers.google.com/machine-learning/recommendation/overview/candidate-generation



(a) A vertex set $|\mathbf{V}|$ (b) An arc $0 \bullet\!\!\rightarrow 1$. (c) An arc set $|\mathbf{A}|$ (d) A graph $\mathscr{G}^D$

Figure 3.2: Construction of a directed graph



(a) Neighborhood of v_1 (b) Successors of v_1. (c) Predecessors of v_1

GRAND, Sumner Magruder

Successors the successors of vertex v, $\mathrm{N}^+(v)$ is the set of vertices **v** $\subset$ **V** such that there exists at least one arc $a = v \hookrightarrow u \; \forall \; u \in$ **v** (fig. 3.3b).

Predecessors the predecessors of vertex v, $\mathrm{N}^-(v)$ is the set of vertices **v** $\subset$ **V** such that there exists at least one arc $a = u \hookrightarrow v \; \forall \; u \in$ **v** (fig. 3.3c).

Path a path p from v_i to $v)j$ of length k is a sequence of unique arcs $a_1, a_2, \ldots \; a_k \in$ **A** which starts at v_i, ends at v_j and contains k arcs of the same direction.

Neighborhood the neighborhood of vertex v, $\mathrm{N}(v)$, is a set of vertices **v** $\subset$ **V** such that u is adjacent with $v \; \forall \; u \in$ **v** on the corresponding *undirected* graph (fig. 3.3a). Likewise, one can also define the neighborhood as follows: $\mathrm{N}(v) = \mathrm{N}^+(v) \cup \mathrm{N}^-(v)$

Out-degree the out degree, δ^+ of vertex v is the defined as $\delta^+(v) = |\mathrm{N}^+(v)|$

In-degree the in degree, δ^- of vertex v is the defined as $\delta^-(v) = |\mathrm{N}^-(v)|$

Degree the degree δ of vertex v is the defined as $\delta(v) = \delta^-(v) + \delta^+(v)$

Graphs are two sets:
V(ertices) and
E(dges)

Ollivier-Ricci Curvature & Hyperbolic geometric Spaces

Why care about Hyperbolic Geometry?

- Continuous analogue to discrete tree-like structures!
- Embedding norm is corollary of hierarchy depth
- Distance between embeddings is affinity of respective items
- Have constant negative curvature

Hyperbolic Geometry (cont)

- This paper users n-dimensional Poincare-Ball

$$\mathbb{H}^n = \{x \in \mathbb{R}^n : ||x|| < 1\}$$

For two points $x, y \in \mathbb{H}^n$ the distance in this space is defined as:

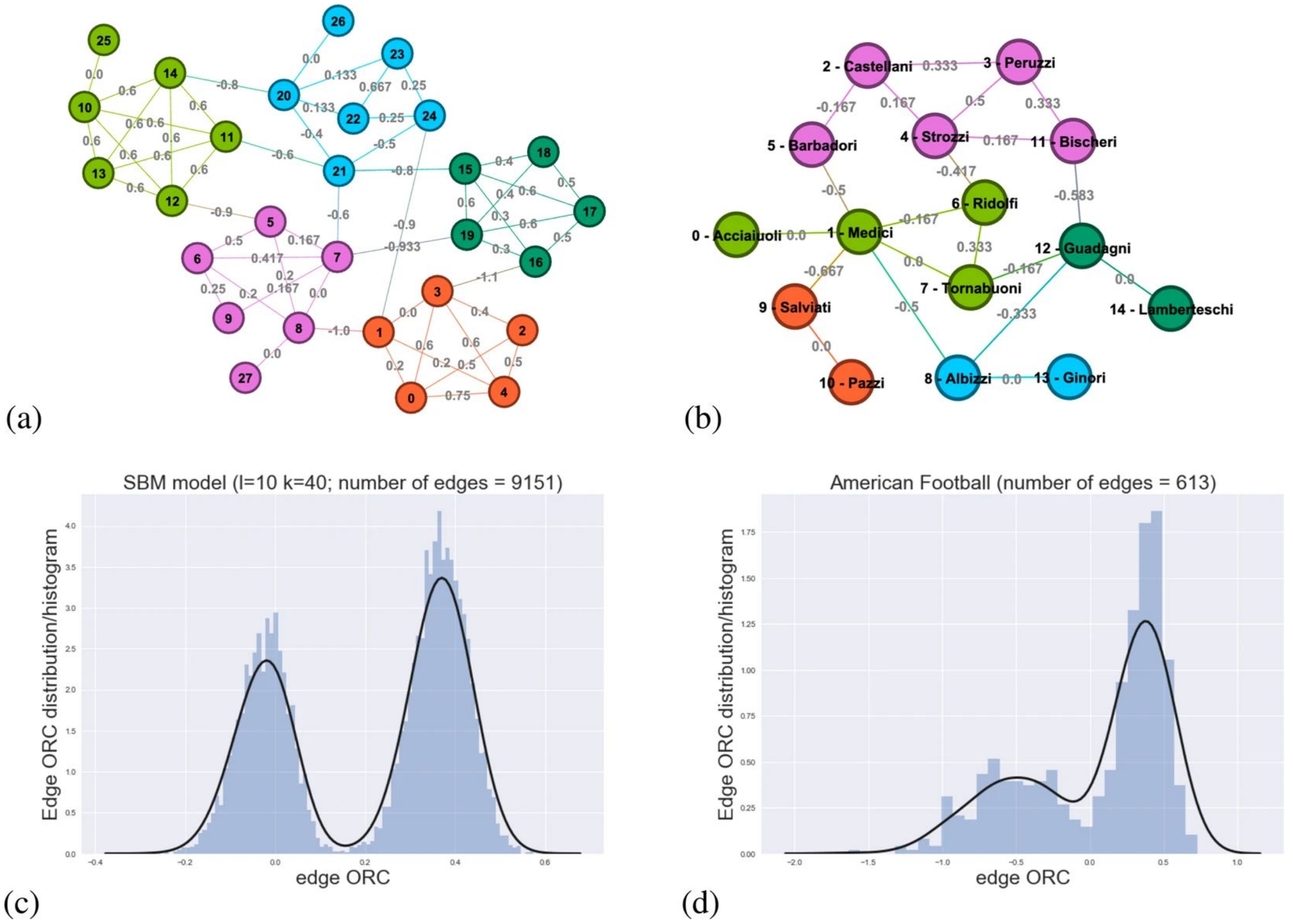
$$d_{\mathbb{H}}(x, y) = \cosh^{-1}\left(1 + 2\frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}\right)$$

https://arxiv.org/pdf/2109.09358v1.pdf

Ollivier- Ricci Curvature

- The Ollivier-Ricci curvature (ORC) captures two fundamental properties of the structure of complex networks
  - ORC of an edge encodes shortest path characteristics.
  - ORC provides information about the frequency of triangles, characterized by the clustering coefficient, within a neighborhood of two adjacent vertices
- Graphs do not have manifold structure, ORC is local approximate

Example



https://www.nature.com/articles/s41598-019-46079-x

Computation

Neighbor: $x \sim y$, $\exists\, e(x,y)$

vertices

is there an edge

Neighborhood:

$$\Gamma_x = \{ v \ \forall v \in G(V,E) : x \sim y \}$$

vertex

vertex set

Edge set

"All vertices touching 'x'"

Degree: $K_x = |\Gamma_x|$

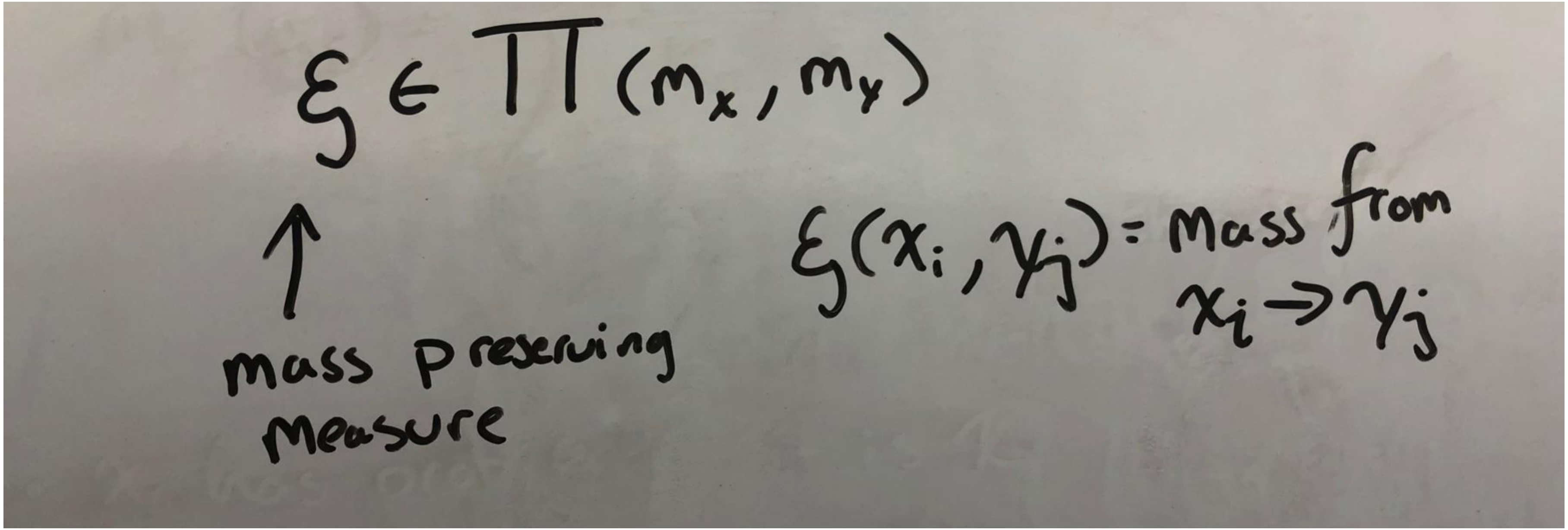
"number of neighbors"

Measure

$$m_x(x_i) = \begin{cases} \alpha, & x_i = x \\ \frac{1-\alpha}{k_x}, & x_i \in \Gamma_x \\ 0, & \text{otherwise} \end{cases}$$

- x_i has prob. α if it is x_i
- neighbors of x have uniform prob. of "what's left" $(1-\alpha)/k_x$

Transportation Plan



Sum over all x_i is m_y and sum over all y_j is m_x

L1 Wasserstein Distance satisfies this and is minimum average traveling distance that can be achieved by any transport plan

L1 Wasserstein Distance

$$W_1(m_x, m_y) = \inf_{\xi} \sum_{x_i \in V} \sum_{y_j \in V} d(x_i, y_j)\,\xi(x_i, y_j)$$

Scaled by d(x_i, y_j), the distance / weight of the edge.

Finally Ollivier-Ricci Curvature

If x is a neighbor of y then the curvature of its edge

$$\underset{\text{(on edge)}}{\mathrm{ORC}} :\rightarrow (x,y) = 1 - \underbrace{\frac{W_1(m_x, m_y)}{d(x,y)}}_{\frac{\text{effort to move } x \rightarrow y}{\text{length of which to move}}}$$

Vertex is ave OCR

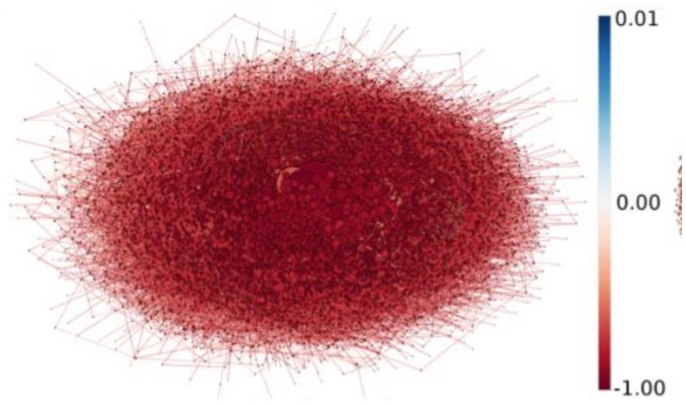
ORC is a discrete estimate of graph curvature

I thought this was about Recommender Systems?

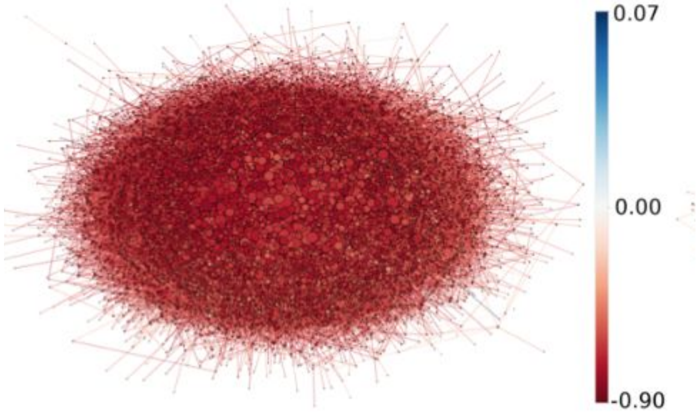
Interaction Graphs have Negative Curvature

From Amazon's Standard Dataset

MusIn Graph



VGame Graph



https://arxiv.org/pdf/2109.09358v1.pdf

Augmenting the Graph

- Use USE (Universal Sentence Encoder) to embed item descriptions, reviews, etc
- If high similarity between two items, add an edge

https://arxiv.org/pdf/1803.11175.pdf

Augmenting the Graph

More edges decrease their importance but nonetheless remain in hyperbolic space

López et al.

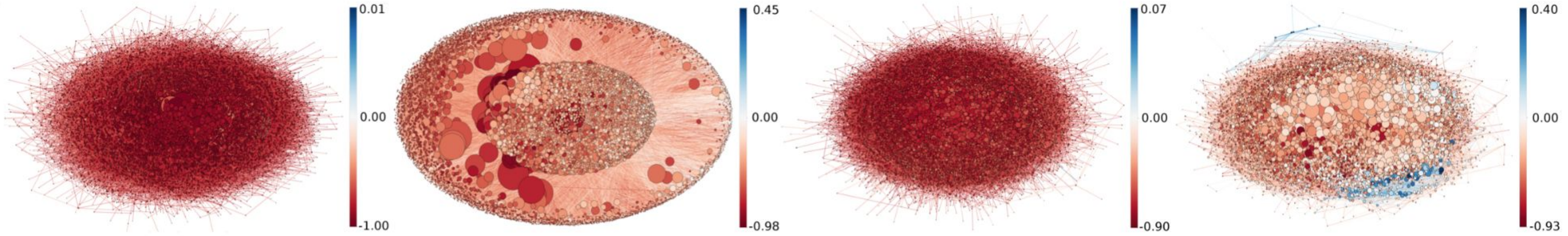


Figure 1: Visualization of Ollivier-Ricci curvature for different graphs. Left to right: a) "MusIns" user-item, b) "MusIns" augmented graph, c) "VGames" user-item, d) "VGames" augmented graph. Node size represents node degree. The vast majority of nodes and edges in red depict negative curvature, in correspondence with the negative curvature of hyperbolic space.

https://arxiv.org/pdf/2109.09358v1.pdf

Augmented Graphs all have improved Hit Ratio

		Musical Instruments					Video Games					Arts, Crafts & Sewing				
		User-item		Augmented			User-item		Augmented			User-item		Augmented		
Model	$\mathcal{S}$	DCG	HR	DCG	HR	Δ%	DCG	HR	DCG	HR	Δ%	DCG	HR	DCG	HR	Δ%
BPR	$\mathbb{R}$	.266	43.03	.308	49.00	13.9	.365	57.67	.397	59.42	3.0	.320	48.80	.390	56.79	16.4
HyperML	$\mathbb{H}$	.295	48.72	.264	47.26	-3.0	.329	53.41	.340	56.31	5.4	.318	50.57	.322	53.13	5.1
CML++	$\mathbb{R}$	.282	48.82	.345	55.87	14.4	.323	55.41	.384	62.31	12.4	.291	50.47	.358	59.60	18.1
NeuMF	$\mathbb{R}$	.320	49.88	.347	55.05	10.4	.419	62.84	.448	66.93	6.5	.363	53.35	.420	61.46	15.2
TransE	$\mathbb{R}$	.284	44.86	.346	54.13	20.7	.387	58.47	.455	67.82	16.0	.330	48.49	.421	60.61	25.0
TransH	$\mathbb{R}$	.284	45.05	.342	54.07	20.0	.390	58.73	.440	67.51	14.9	.329	48.22	.422	61.63	27.8
DistMul	$\mathbb{R}$	.251	40.58	.310	49.75	22.6	.346	54.85	.370	57.54	4.9	.290	44.85	.321	48.61	8.4
RotatE	$\mathbb{C}$	.251	40.22	.332	53.31	32.6	.345	54.49	.405	63.80	17.1	.300	46.37	.389	58.37	25.9
RotRefEu	$\mathbb{R}$	.317	51.63	.366	57.82	12.0	.408	63.12	.447	67.31	6.6	.371	56.43	**.433**	63.77	13.0
RotRefHy	$\mathbb{H}$	.317	51.63	**.368**	57.86	12.1	.400	62.78	.450	68.64	9.3	.363	55.76	.421	**64.09**	14.9
MuRE	$\mathbb{R}$	**.337**	52.81	.365	**57.99**	9.8	.405	63.11	.462	68.80	9.0	**.372**	56.72	.422	63.95	12.8
MuRP	$\mathbb{H}$	.332	**52.84**	**.368**	57.86	9.5	**.424**	**64.92**	**.467**	**70.49**	8.6	.366	**56.83**	.426	64.06	12.7

Table 4: Results for "User-item" and "Augmented" graph setups, for models operating in Euclidean ($\mathbb{R}$), hyperbolic ($\mathbb{H}$) and complex ($\mathbb{C}$) space. Δ% shows the hit rate improvement over the user-item graph when data is augmented. All HR improvements are statistically significant (one-tailed Mann-Whitney U test, $p \leq 0.05$).

https://arxiv.org/pdf/2109.09358v1.pdf

Augmented via USE

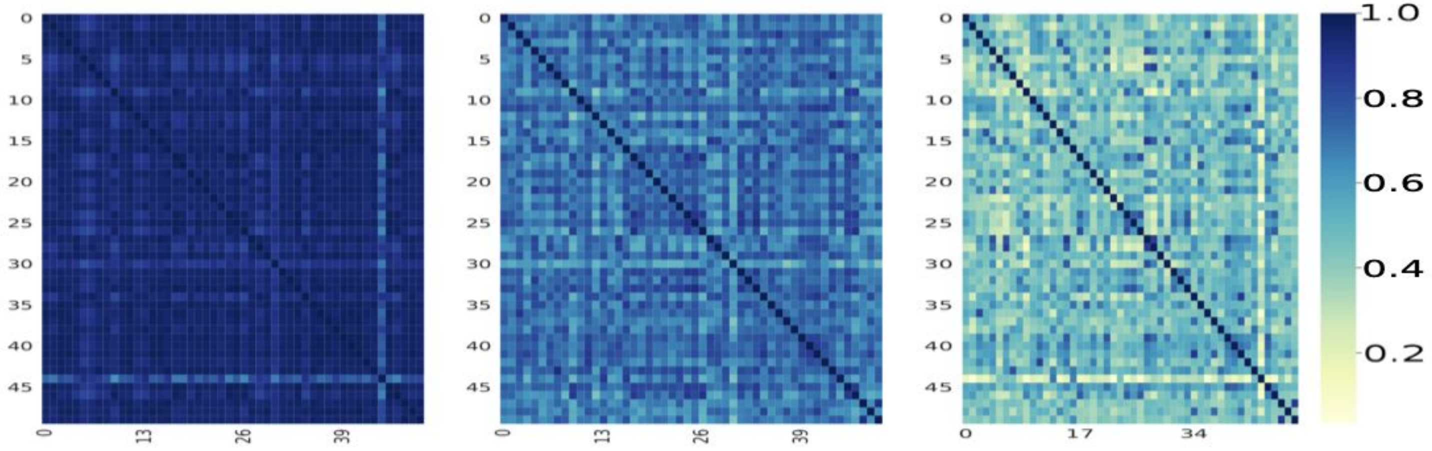


Figure 2: Cosine similarity of BERT (left), S-BERT (center) and USE (right) review embeddings for 50 items.

https://arxiv.org/pdf/2109.09358v1.pdf

Questions

- Common approaches to improve recommendations appear to rely on augmenting the recommend item space (often by reviews), are there other approaches?
- How does shifting geometric spaces inform analysis?

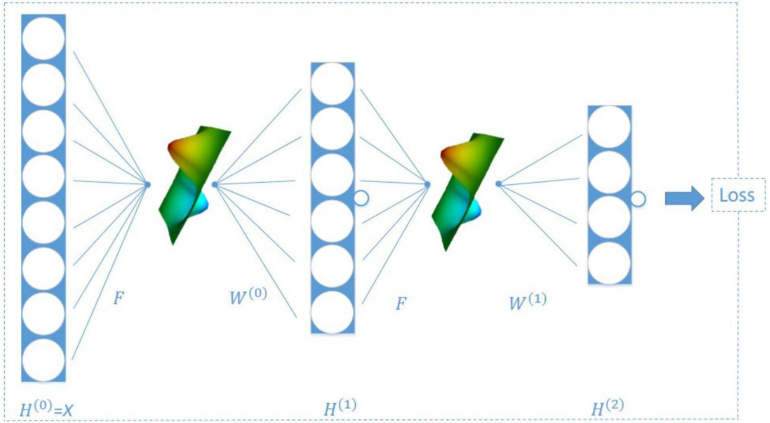


Figure 5. Ricci curvature-based graph convolutional neural network (RCGCN) architecture.

Table 2. Summary of results in the form of classification accuracy (percent).

Method	Cora	Citeseer	Cornell	Texas	Wisconsin
DeepWalk [4]	67.2	43.2	58.38	57.84	50.39
Planetoid [25]	75.7	64.7	45.41	63.24	63.14
GCN [15]	81.5	70.3	62.43	60.81	51.37
GAT [16]	**82.4**	72.6	68.92	62.16	38.43
RCGCN	82.0	**73.2**	**79.19**	**81.89**	**82.94**

https://www.mdpi.com/1099-4300/23/3/292

User-Item Graphs have negative curvature

Augmenting User-Item Graphs improve models

Generate Natural Language Explanations for Recommendation

Additional Paper 1

Very long and dense paper

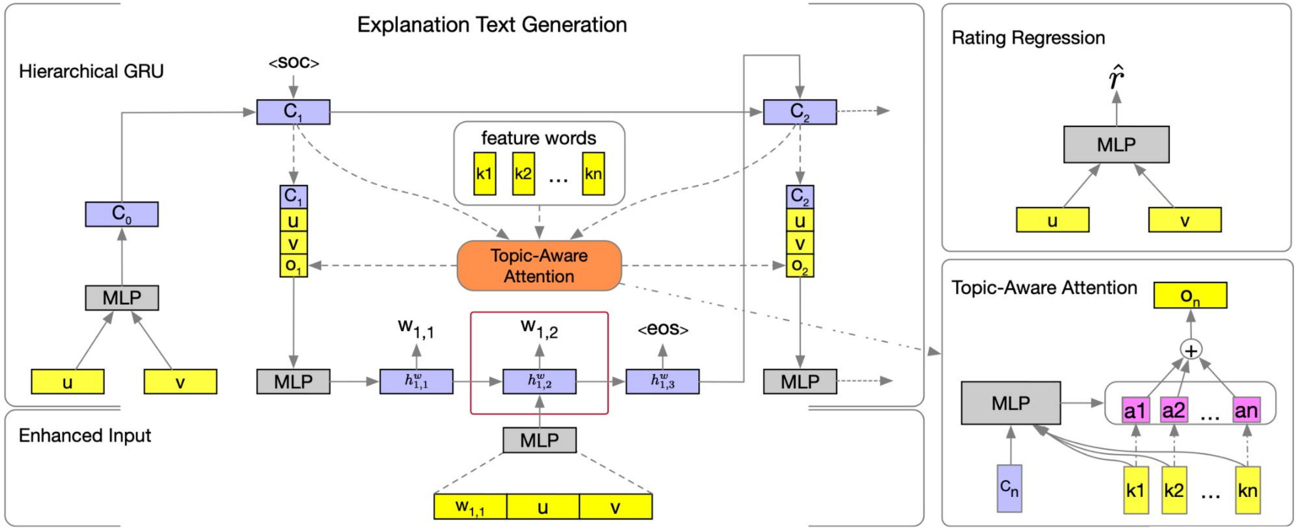
- Things we have seen before:
  - Explainability
  - Multitask Learning
  - Custom Attention
- Why Care?
  - Item recommendations are a very powerful feature of RecSys (companies even pay people to write reviews)
  - Knowing what is a “good” review can be used to filter your data

Description	Explanation Sentences
good explanation on Beauty	The **bottle** is very **light** and the **smell** is very strong.
good explanation on Electronics	The **price** is great. The **sound quality** is great
cover feature words but not fluent	The **scent** is a good **product**. I have to use this **product**. I have used to use the **hair**.
fluent but wrong description	the **price** is a great. The **sound** is great

https://arxiv.org/pdf/2101.03392.pdf

Interesting notes

- Improve latent representation of users and items via rating regression
- Denoise reviews by selecting sentences therein with highest number of keywords



https://arxiv.org/pdf/2101.03392.pdf

Avenues of RecSys research exist outside of User-Item graphs

Advances and Challenges in Conversational Recommender Systems: A Survey

Additional Paper 2

Why Care?

- We've seen GPT-3
- We've seen chat bots
- Can these bots do anything useful?
  - Yes & you've probably encountered several

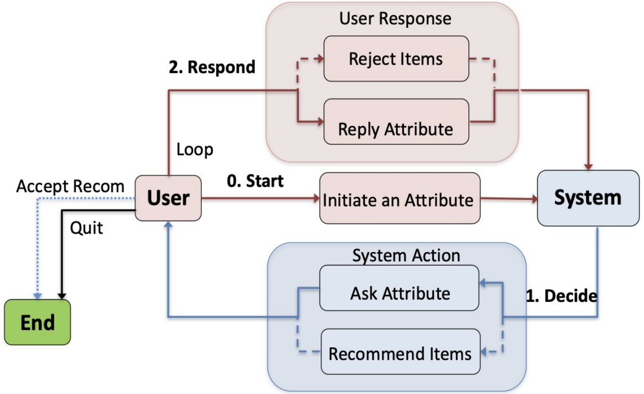


Figure 5: The estimation-action-reflection workflow. Credits:

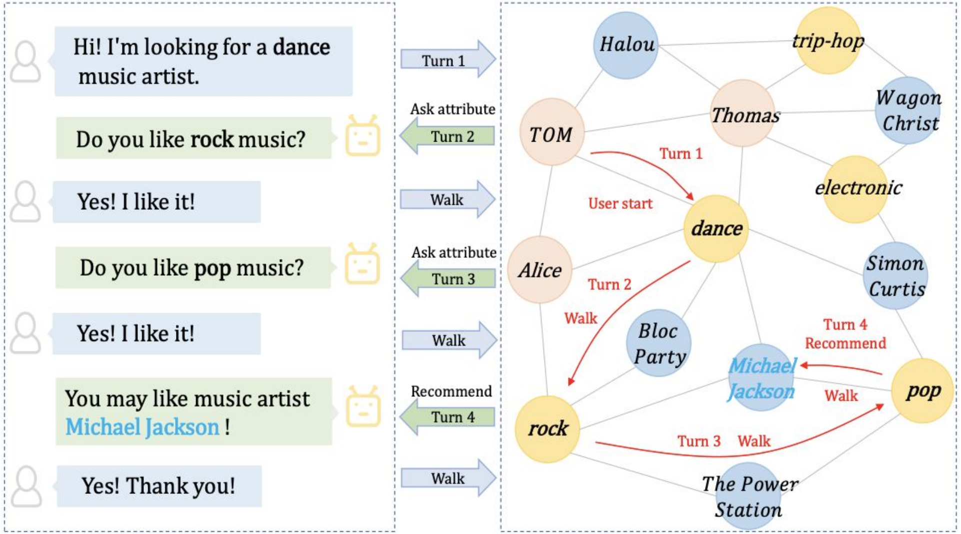


Figure 4: An illustration of interactive path reasoning in the conversational path reasoning (CPR) model. Credits: Lei et al.

https://arxiv.org/pdf/2101.09459.pdf

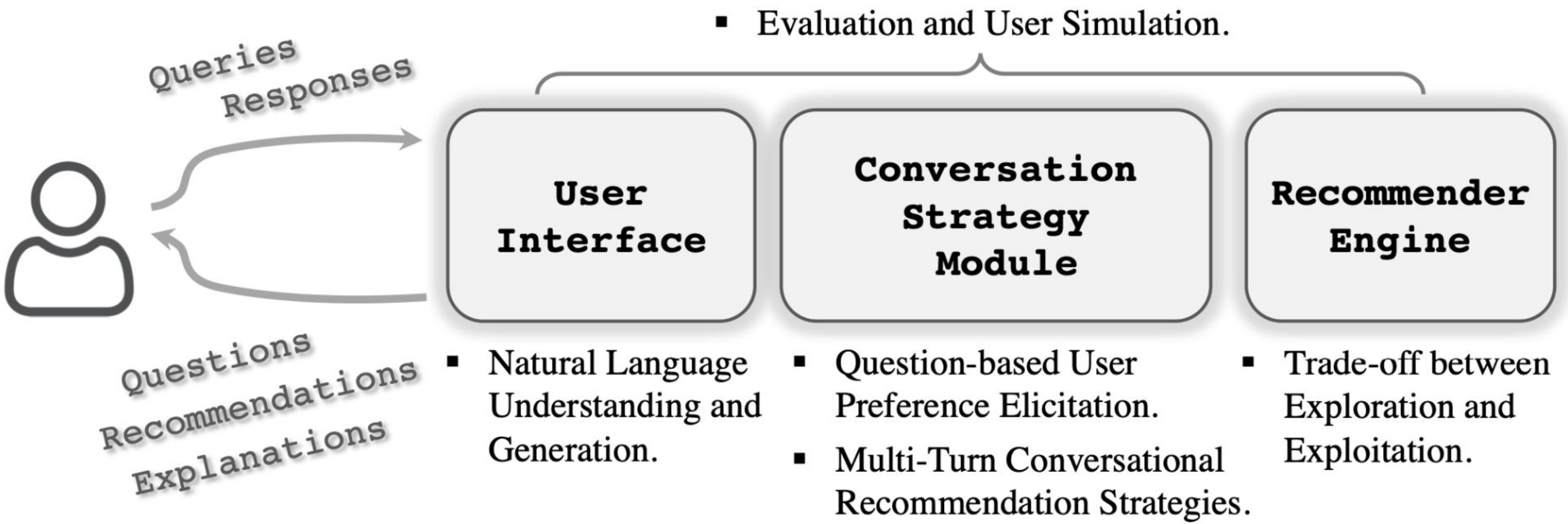


Figure 3: Illustration of the general framework of CRSs and our identified five primary challenges on the three main components.

https://arxiv.org/pdf/2101.09459.pdf

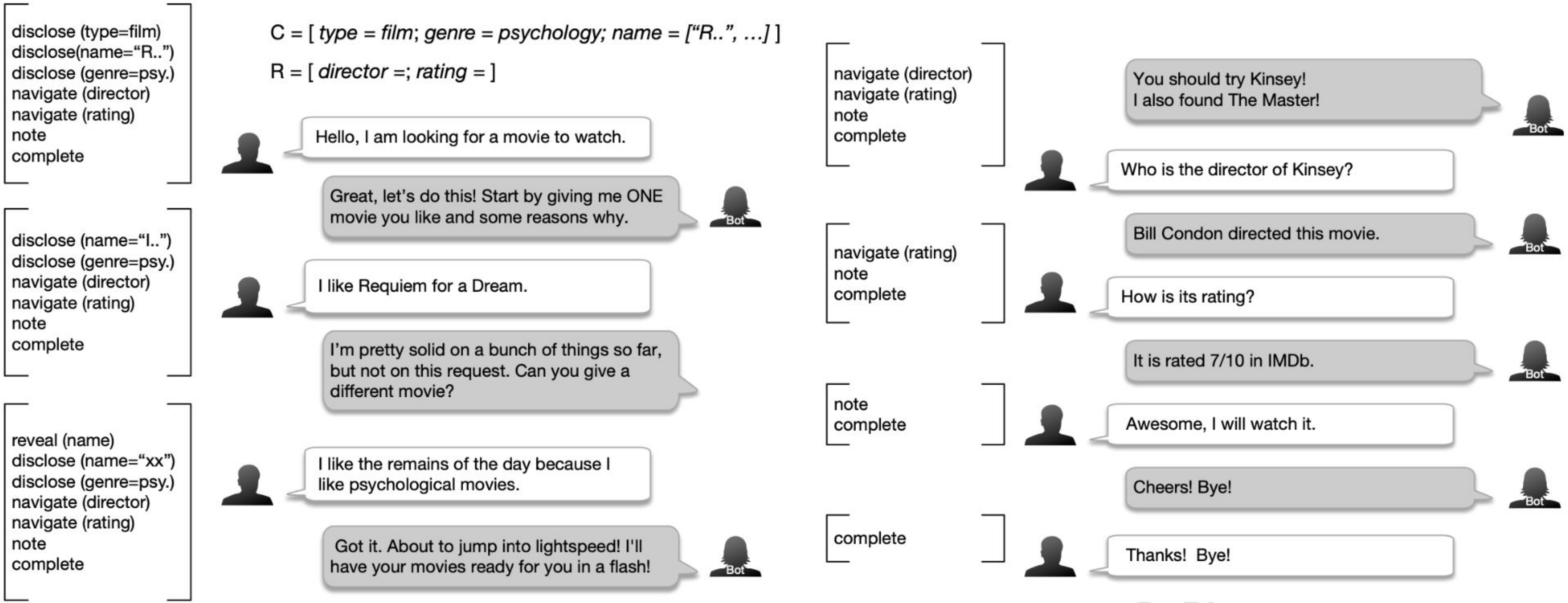


Figure 8: Example dialogue with agenda sequence and state transition. The agenda is shown in square brackets. The third agenda is a result of a push operation, all other agendas updates are pull operations. Credits: Zhang and Balog [246].

https://arxiv.org/pdf/2101.09459.pdf

Primary Challenges in CRSs	Contributions of Existing Studies
Question-based User Preference Elicitation	Asking about items
	Asking about attributes
Multi-turn Conversational Strategies	Explicit strategies
	Leading diverse topics
Language Understanding and Generation	End-to-end dialogue systems
Exploration and Exploitation Trade-offs	Leveraging multi-armed bandits
Evaluation and User Simulation	Evaluation
	User simulation

https://arxiv.org/pdf/2101.09459.pdf

Conversational RecSys are a hat-on-a-hat

Closing Question 1

Non-graph based recommendation approaches?

I thought this was about Natural Language Processing?

Session (Sequential) Models

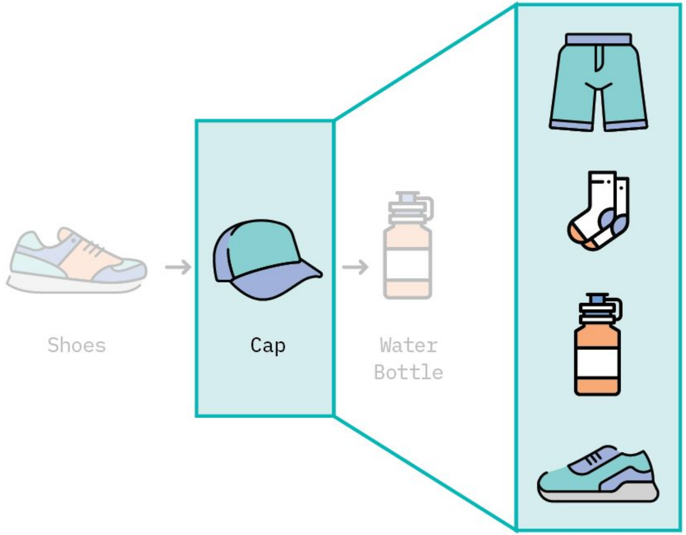
- Rather than look at the graph as a whole look at interactions (views, clicks, purchases) in order

Note: Session models are a type of sequential model. They can also include a “bag” of user-item interactions not scoped to a single user
https://medium.com/nvidia-merlin/transformers4rec-4523cc7d8fa8

Word2Vec

Just like the n-gram Word2Vec model, try to learn items that surround the query item.

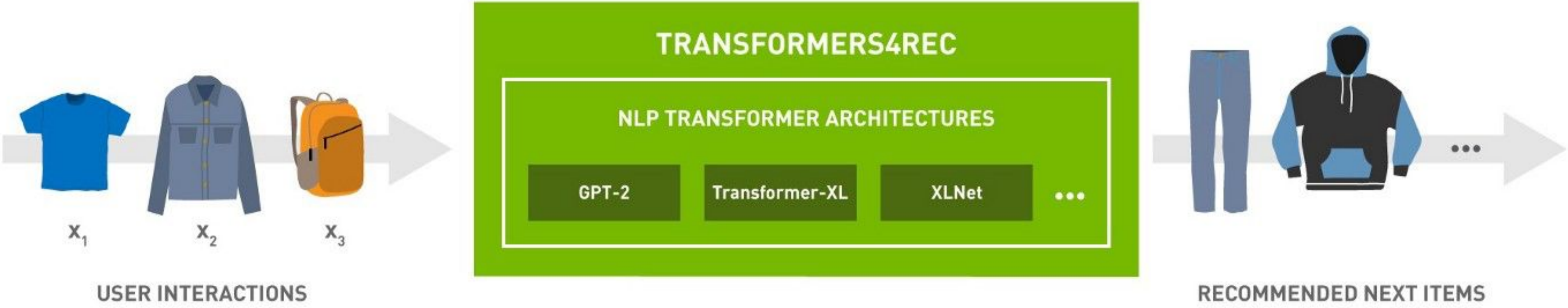
- Scales poorly



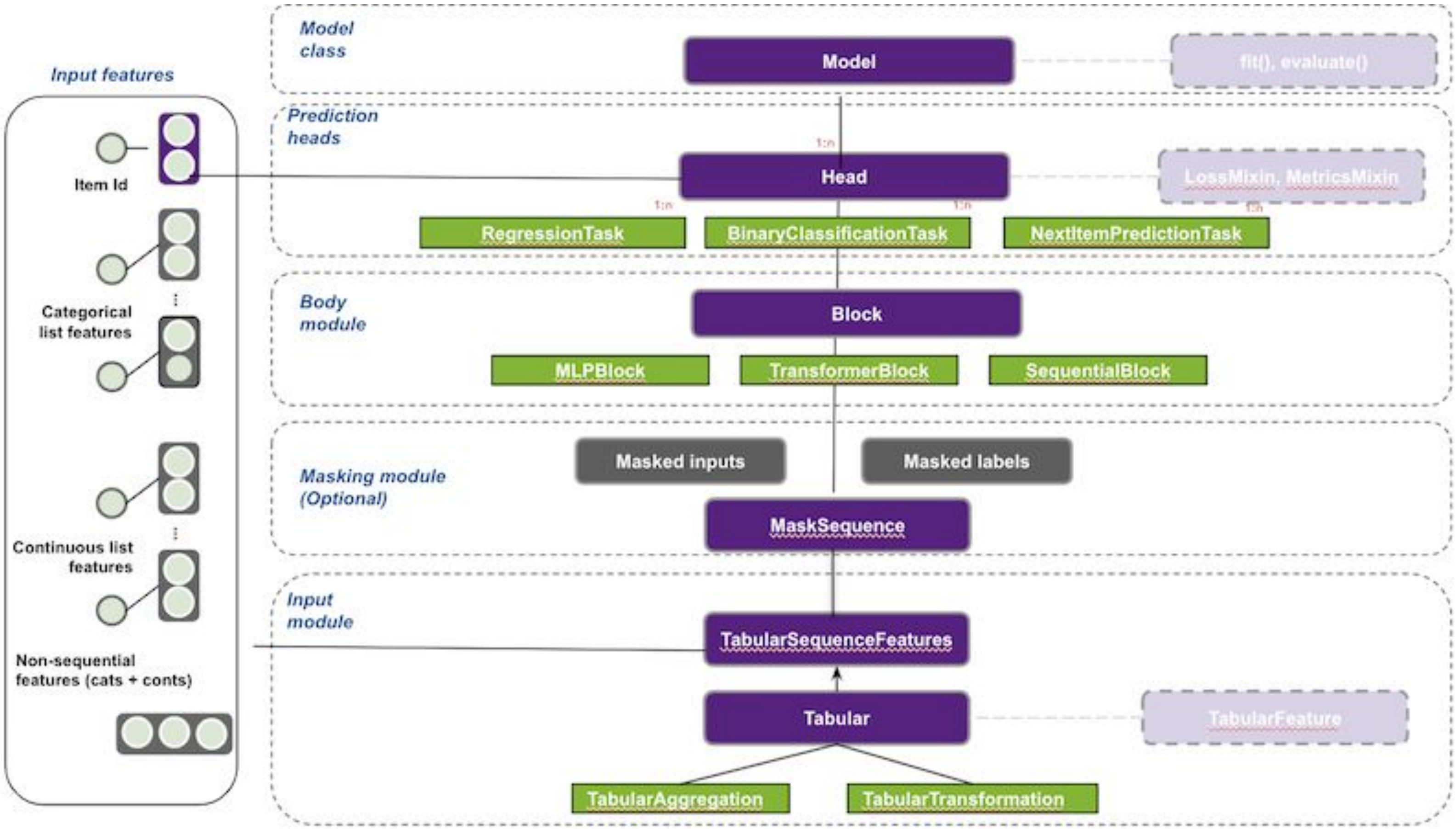
https://session-based-recommenders.fastforwardlabs.com

Transformers4Rec

- By Nvidia



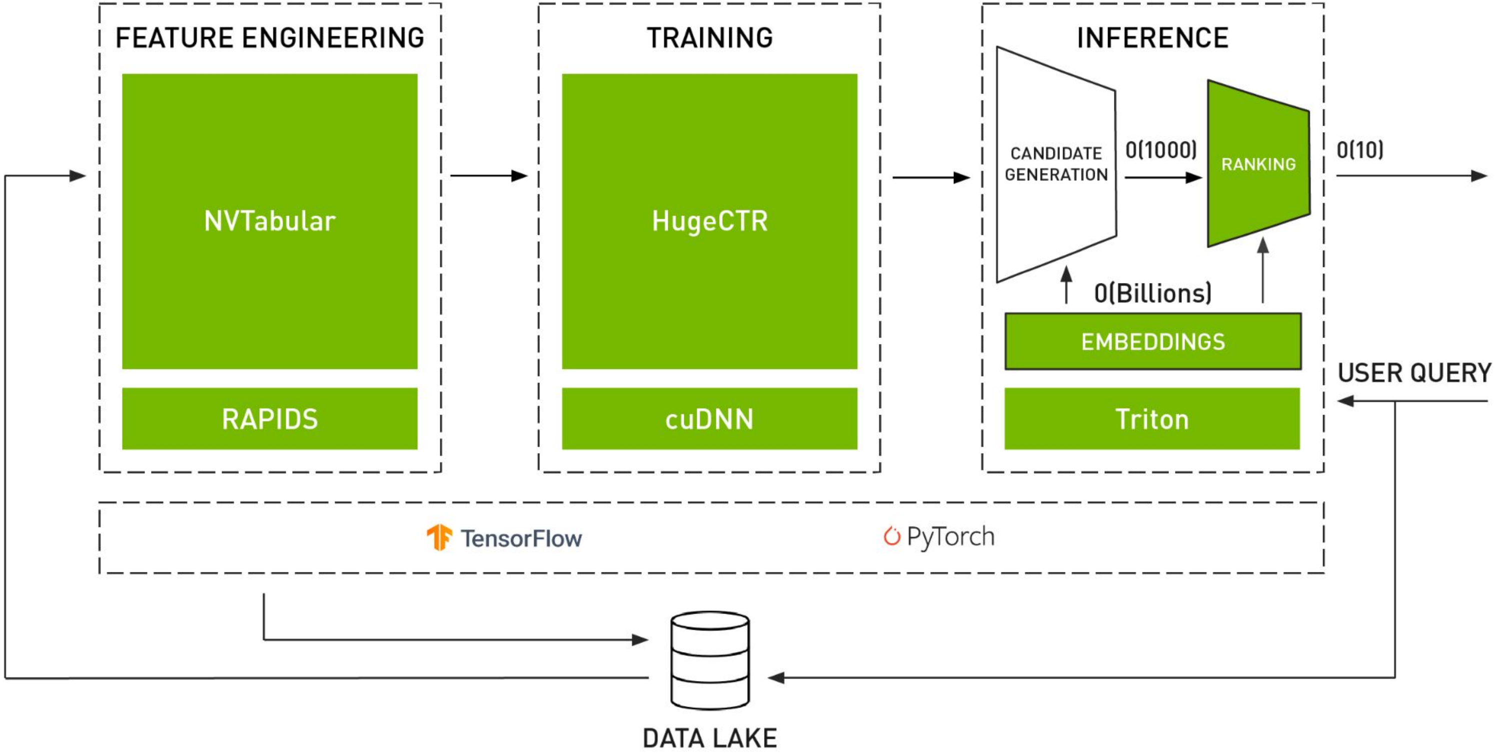
https://medium.com/nvidia-merlin/transformers4rec-4523cc7d8fa8



https://medium.com/nvidia-merlin/transformers4rec-4523cc7d8fa8

NVIDIA MERLIN COMPONENTS

INTEROPERABILITY WITH OPEN SOURCE



https://developer.nvidia.com/nvidia-merlin?ncid=so-medi-636648#cid=dl19_so-medi_en-us

Session-based Recommender Systems

- Solve cold start problem
- More suitable to time
  - What is more important short / long term user intent?
- RecSys already large as is

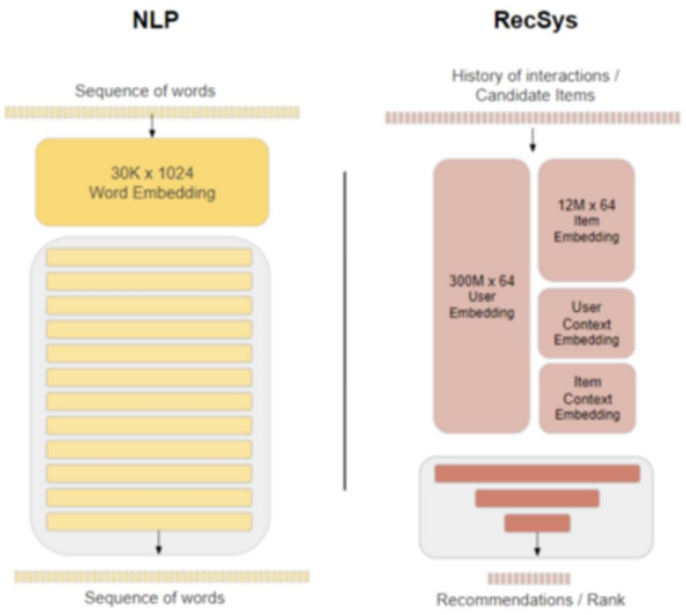


Figure 15: Model and embedding sizes between NLP and recommender systems vary significantly.

https://developer.nvidia.com/blog/how-to-build-a-winning-recommendation-system-part-2-deep-learning-for-recommender-systems/

Cons Review

Content Based:

- Hand engineer features
- Struggles with coverage

Collaborative Filtering:

- Cold Start
- Struggles with side information

Session Based:

- More susceptible to noise
- Scale of intent difficult

Architecture Overview

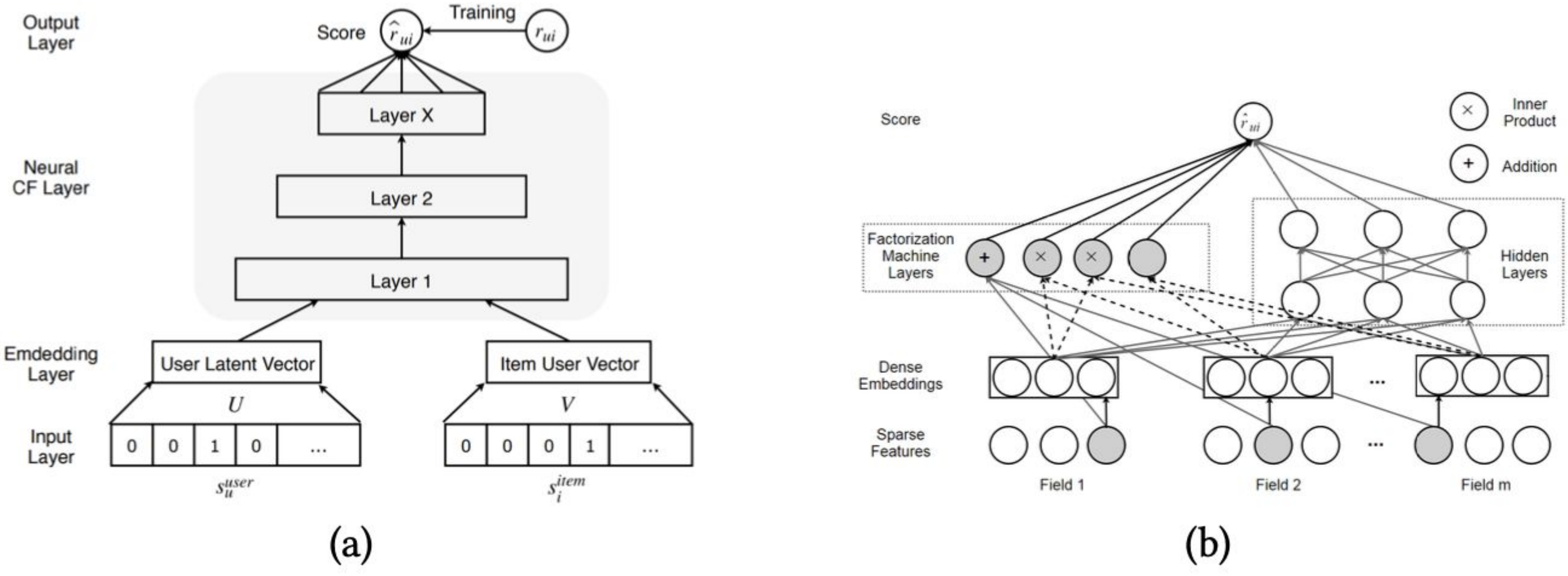


Fig. 2. Illustration of: (a) Neural Collaborative Filtering; (b) Deep Factorization Machine.

https://arxiv.org/abs/1707.07435

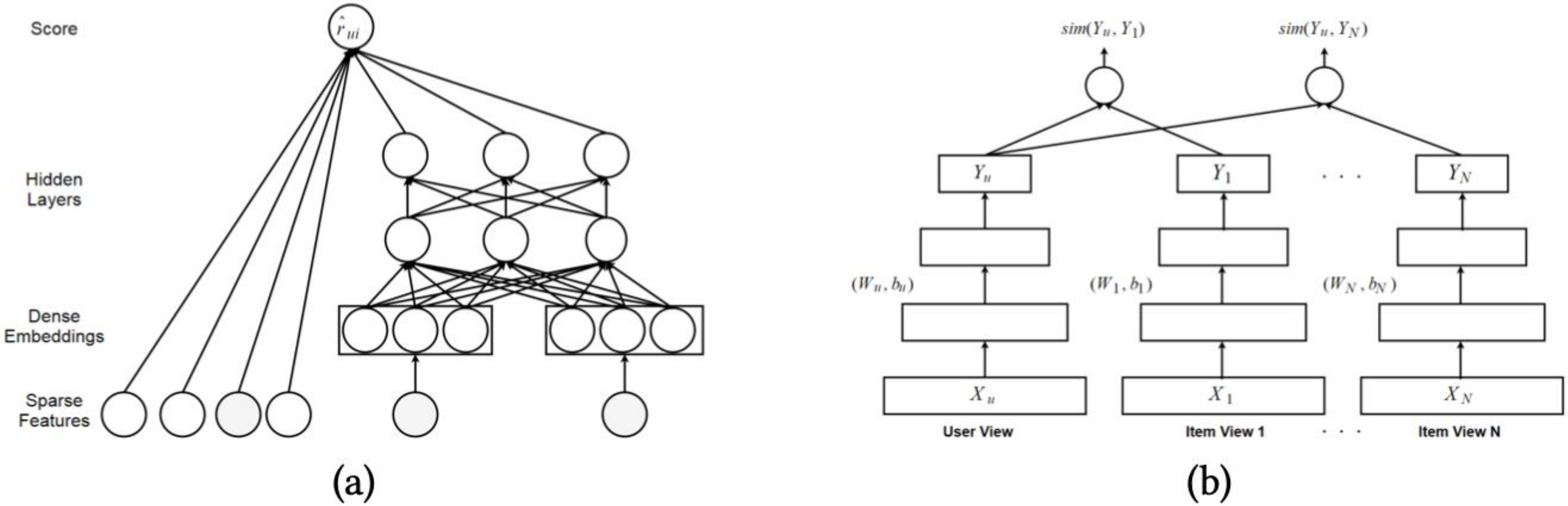


Fig. 3. Illustration of: (a) Wide & Deep Learning; (b) Multi-View Deep Neural Network.

https://arxiv.org/abs/1707.07435

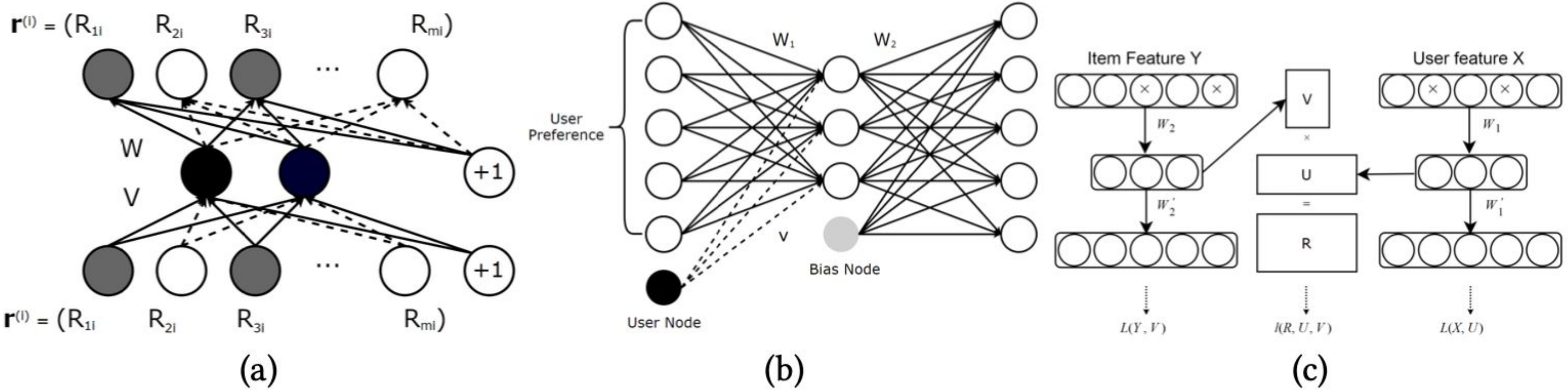


Fig. 4. Illustration of: (a) Item based AutoRec; (b) Collaborative denoising autoencoder; (c) Deep collaborative filtering framework.

https://arxiv.org/abs/1707.07435

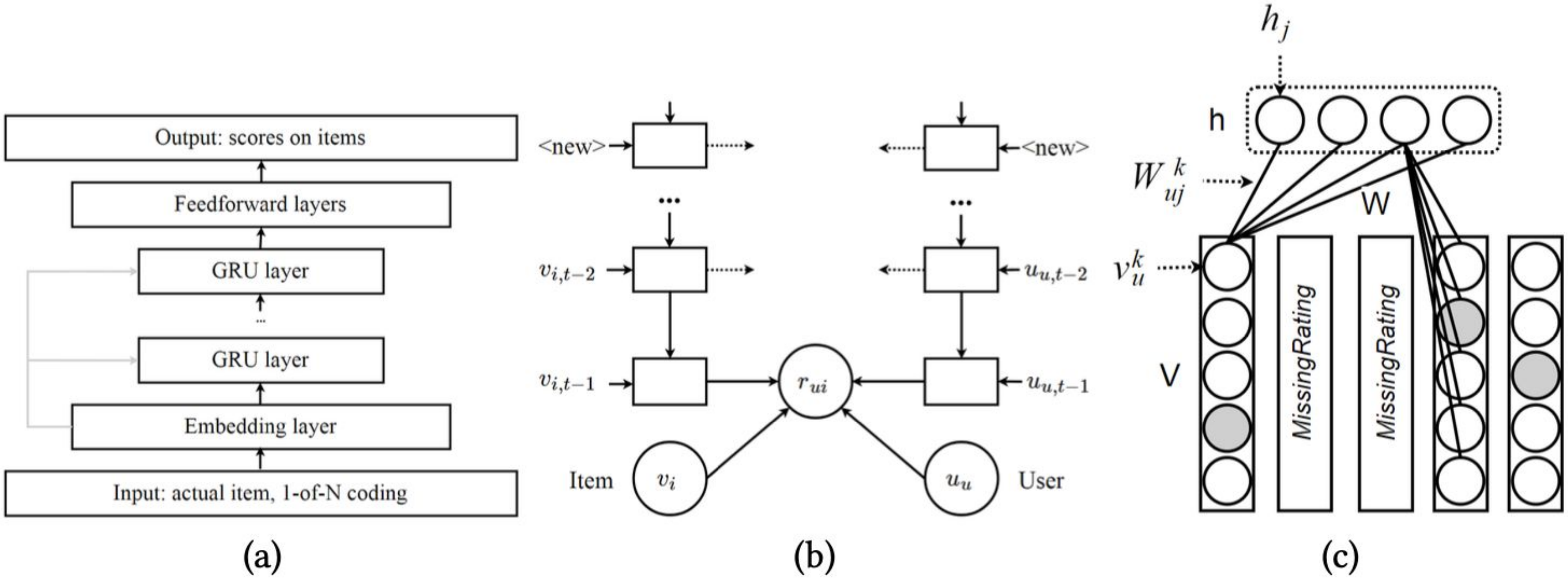


Fig. 6. Illustration of: (a) Session-based recommendation with RNN; (b) Recurrent recommender network; (c) Restricted Boltzmann Machine based Collaborative Filtering.

https://arxiv.org/abs/1707.07435

Closing Question 2

Is the ubiquity of RecSys a good thing?

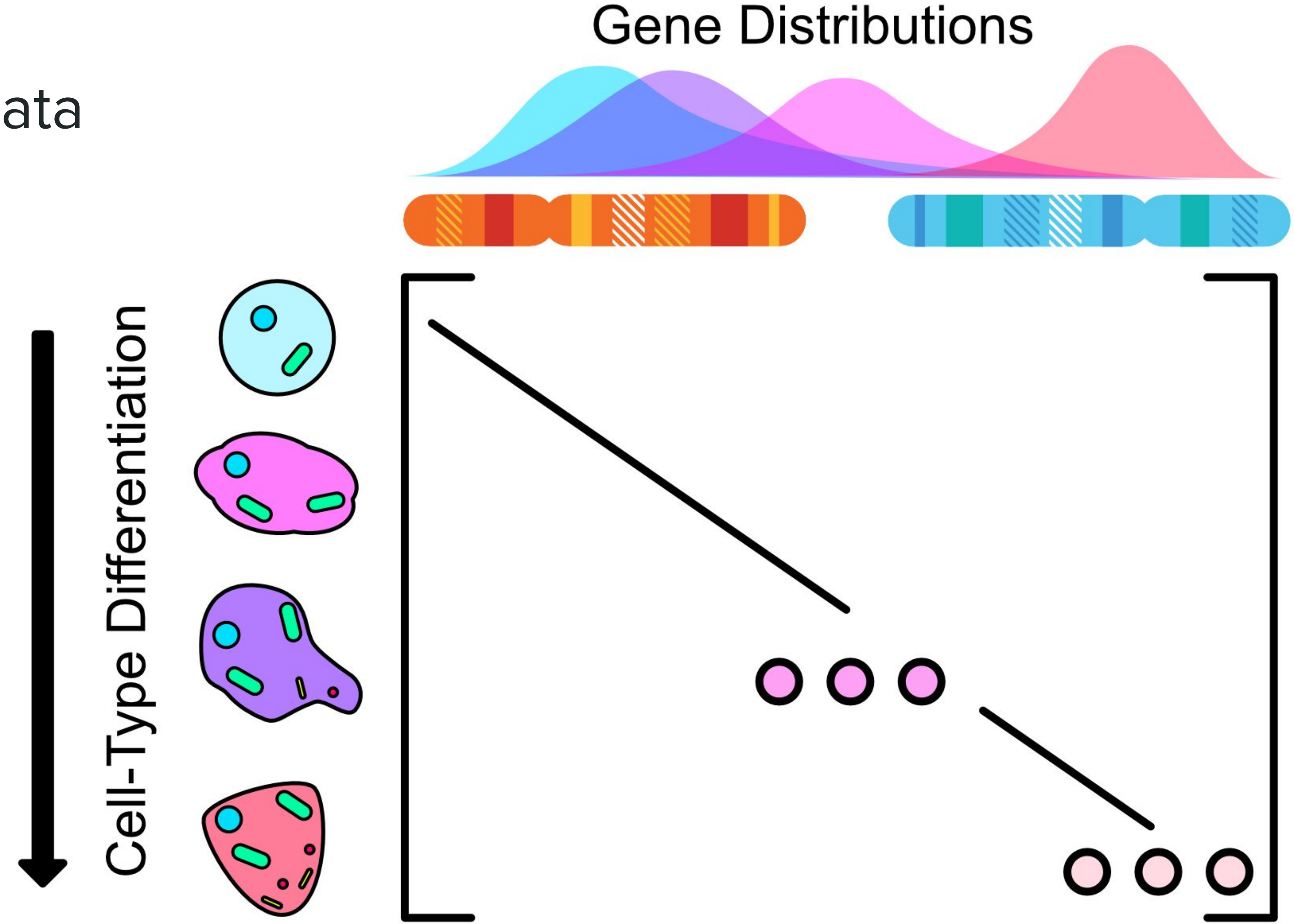
Closing Question 3

What would you like a RecSys to do which at the moment it can not?

Closing Question 4

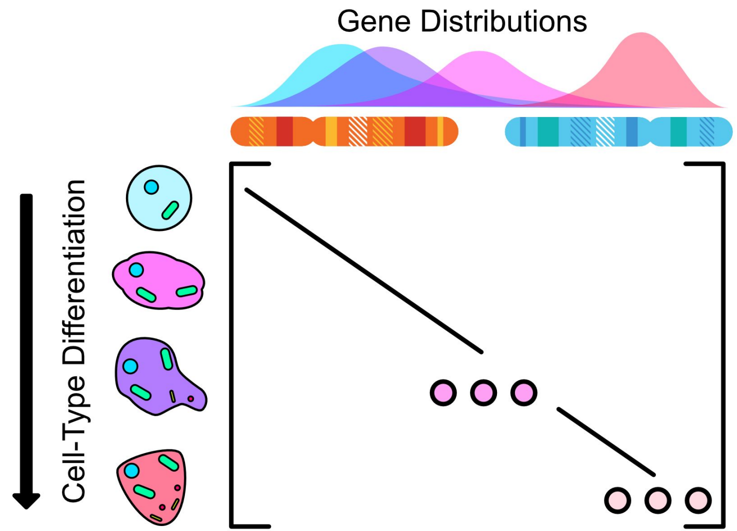
Can you find a way to repurpose advances in RecSys to other fields?

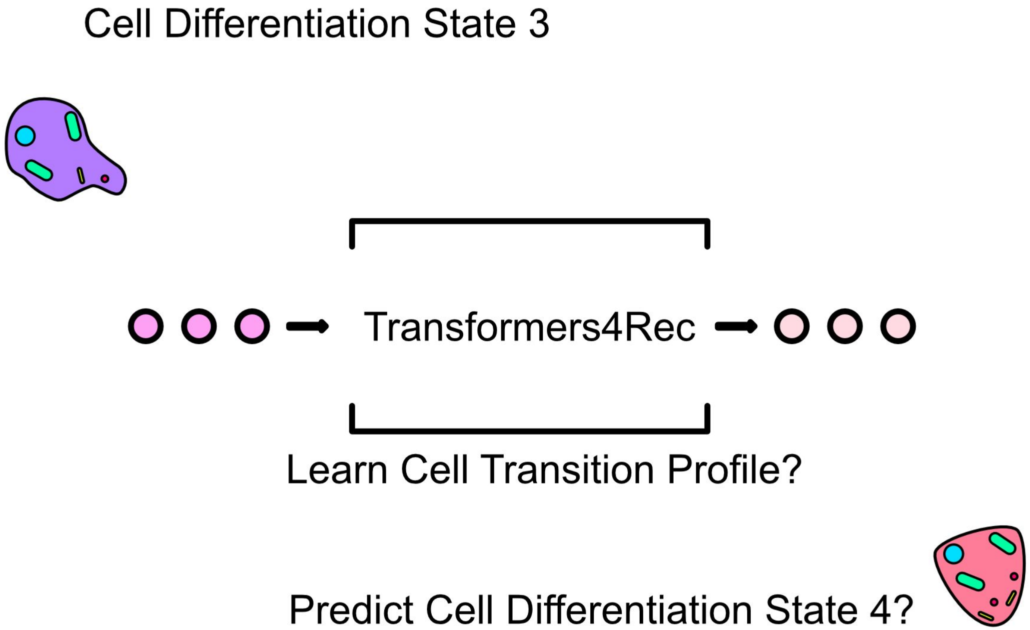
Single Cell Data



Sumner Magruder

Single Cell Data





Sumner Magruder

Other Neat Stuff

Papers / Architectures

DASO (Deep Adversarial Social Recommendation) https://www.ijcai.org/proceedings/2019/0187.pdf

DEERS - Recommendation with Negative Feedback via Pairwise Deep Reinforcement Learning https://arxiv.org/pdf/1802.06501.pdf

Transformers4Rec https://resources.nvidia.com/en-us-merlin/transformers4rec-bringing-the-gap?lx=97GH0Q

Graph Neural Networks in Recommender Systems: A Survey https://arxiv.org/pdf/2011.02260.pdf

Ricci Curvature-Based Semi-Supervised Learning on an Attributed Network https://www.mdpi.com/1099-4300/23/3/292