

# Structured Output Models of Recommendations, Activities, and Behavior

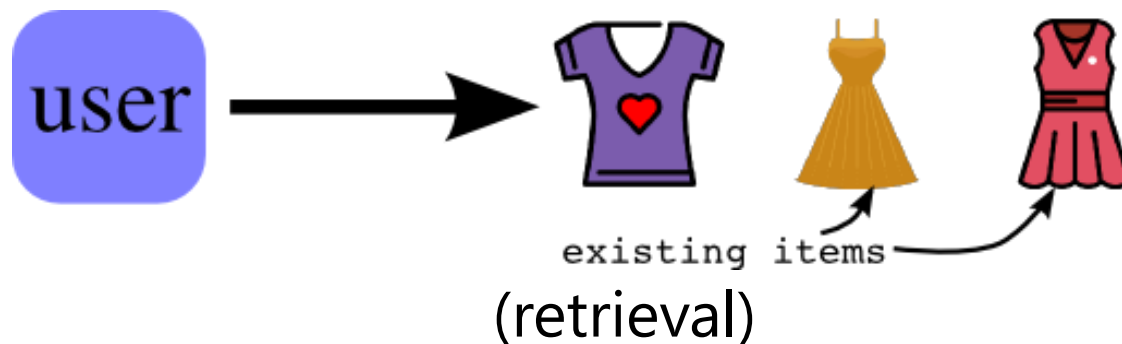
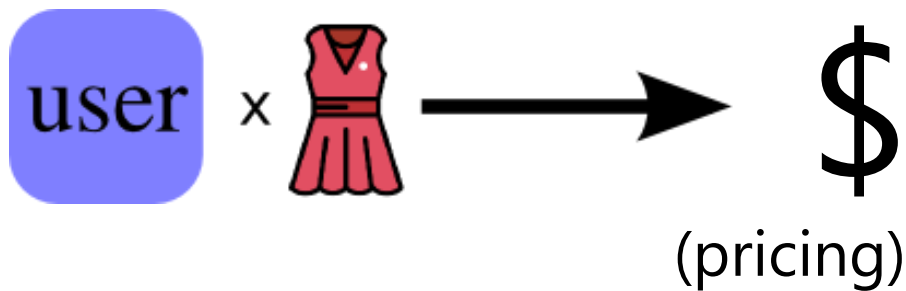
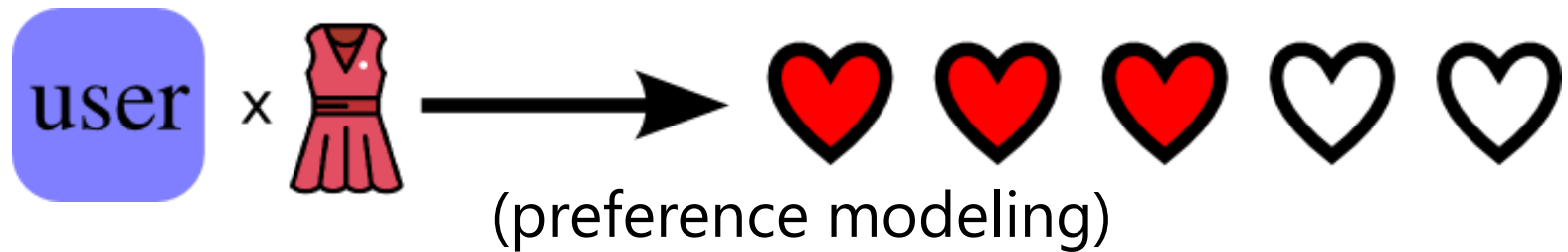
Feb 9, 2018

Julian McAuley

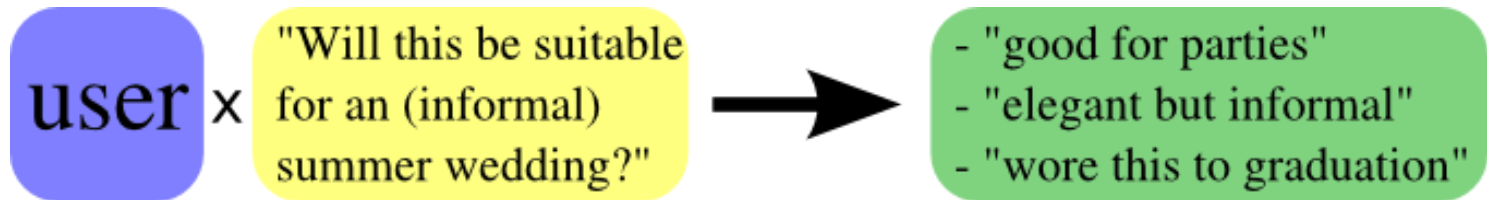
# Where are recommender systems used?



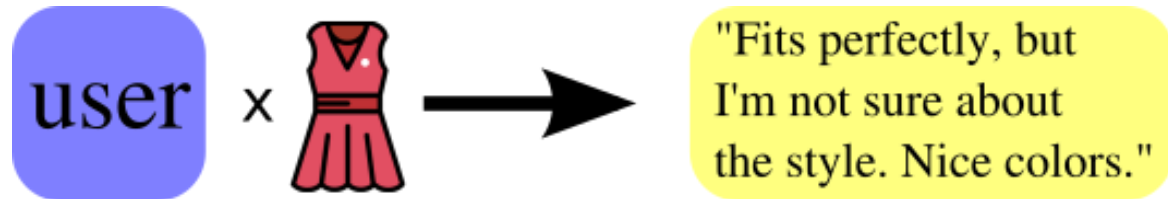
# What do recommender systems do?



# What *could* recommender systems do?



## 1. Question answering

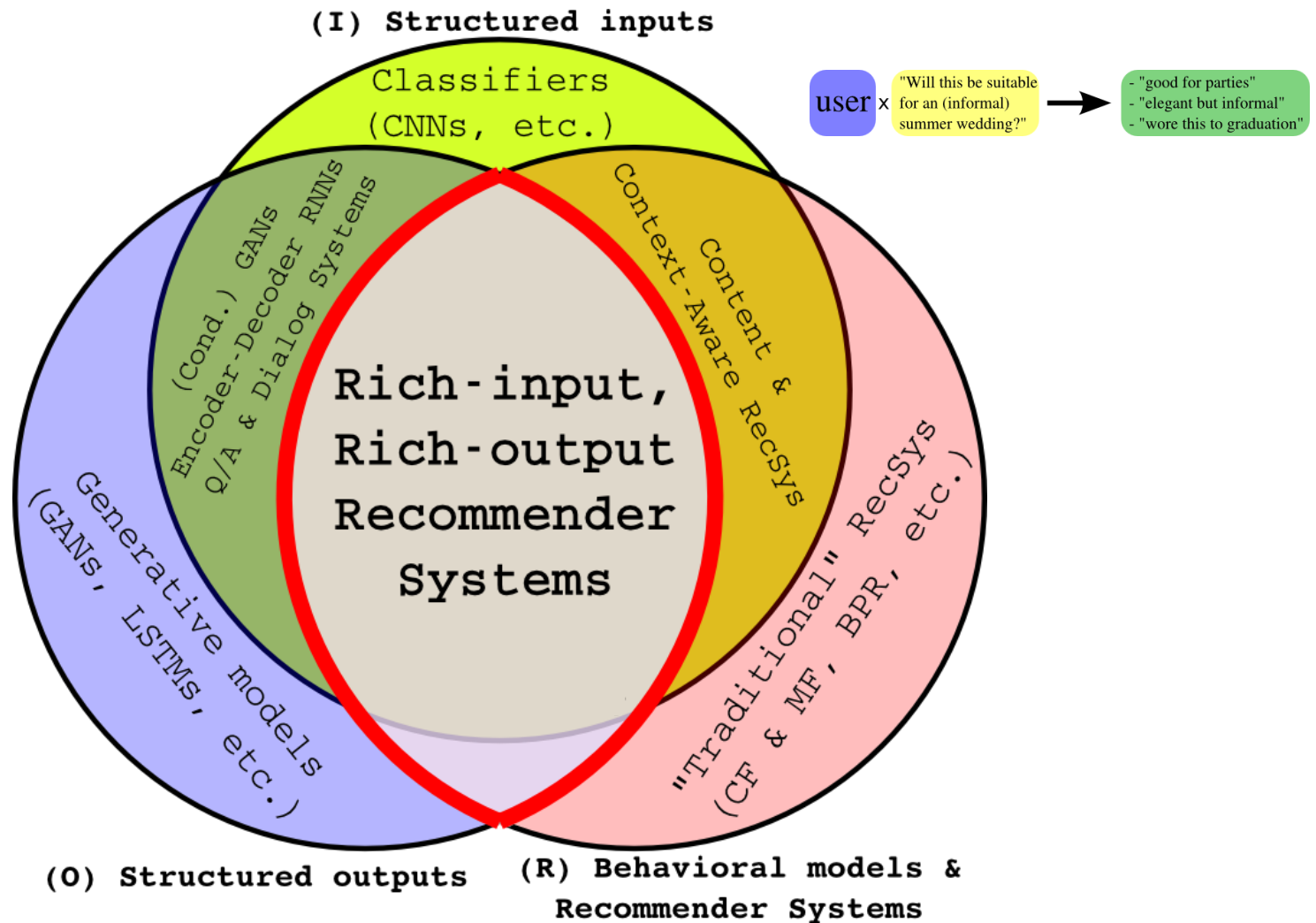


## 2. Estimating reactions



## 3. Generating content

# Recommender systems + structured output / generative modeling




# Rich-input, rich-output recommender systems

1. How can we extend **Q/A** systems to deal with issues of **personalization and subjectivity**?
2. How can we extend **generative text models** to **estimate nuanced reactions**?
3. How can we extend **Generative Adversarial Nets** to generate **personalized content**?

# Goals of my lab's research

**Machine Learning:** new methodology



**Goal 1:** Extending structured output models to account for **variance across users**

**Goal 2:** Building recommender systems with **rich, structured outputs**



**Recommender Systems:**  
New applications

# Data




**~100M** reviews, **~10M** items, **~20M** users  
**1.4M** questions and answers

**Beer**advocate

**~3M** reviews, **~60k** items, **~30k** users





1. Answering  
personalized and  
subjective  
questions

# Answering product-related queries



**Q:** "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

Suppose we want to answer the question above.  
Should we:

- 1) Wade through (hundreds of!) existing reviews looking for an answer → time consuming
- 2) Ask the community via a Q/A system? → have to wait
- 3) Can we answer the question **automatically?**

# Answering product-related queries



**Q:** "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

## Challenging!

- The question itself is complex (not a simple query)
- Answer (probably?) won't be in a knowledge base
- Answer is subjective (how loud is "loud enough"?)

# Answering product-related queries



**Q:** "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

So, let's use **reviews** to find possible answers:

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

# Answering product-related queries



**Q:** "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

## Still challenging!

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

- Text is only tangentially related to the question
- Text is linguistically quite different from the question
- Combination of positive, negative, and lukewarm answers to resolve

# Answering product-related queries



**Q:** "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

So, let's aggregate the results of many reviews

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

"If you are looking for a water resistant blue tooth speaker you will be very pleased with this product."

Yes

"However if you are looking for something to throw a small party this just doesn't have the sound output."

No

**=Yes**



# Challenges

- 1.** Question, answers, and reviews are linguistically heterogeneous
- 2.** Questions may not be answerable from the knowledge base, or may be subjective
- 3.** Many questions are non-binary

# Linguistic heterogeneity

## Question, answers, and reviews are linguistically heterogeneous

How might we estimate whether a review is “relevant” to a particular question?

1. Cosine similarity?  (won't pick out important words)
2. Tf-idf (e.g. BM25 or similar)?  (won't handle synonyms)
3. **Bilinear models**

$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} W \mathbf{x}_{\text{review}}^T$$



# Linguistic heterogeneity

$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} W \mathbf{x}_{\text{review}}^T$$



$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} (AB^T + \Delta) \mathbf{x}_{\text{review}}^T$$

- $A$  and  $B$  embed the text to account for synonym use,  $\Delta$  accounts for (weighted) word-to-word similarity
  - But how do we learn the parameters?

# Parameter fitting

- We have a high-dimensional model whose parameters describe how relevant each review is to a given question
  - But, we have no **training data** that tells us what is relevant and what isn't
- But we *do* have training data in the form of **answered questions!**

**Idea:** A **relevant** review is one that helps us to **predict** the correct answer to a question

# Parameter fitting

$$p(\text{answer is yes} \mid \text{question } q) = \sum_{r \in \text{reviews}} \underbrace{p(r \text{ is relevant} \mid q)}_{\substack{\propto \mathbf{x}_{\text{question}}(AB^T + \Delta)\mathbf{x}_{\text{review}}^T \\ \text{"relevance"}}} \underbrace{p(\text{yes} \mid r, q)}_{\substack{\propto \mathbf{x}_{\text{question}}(A'B'^T + \Delta')\mathbf{x}_{\text{review}}^T \\ \text{"prediction"}}$$

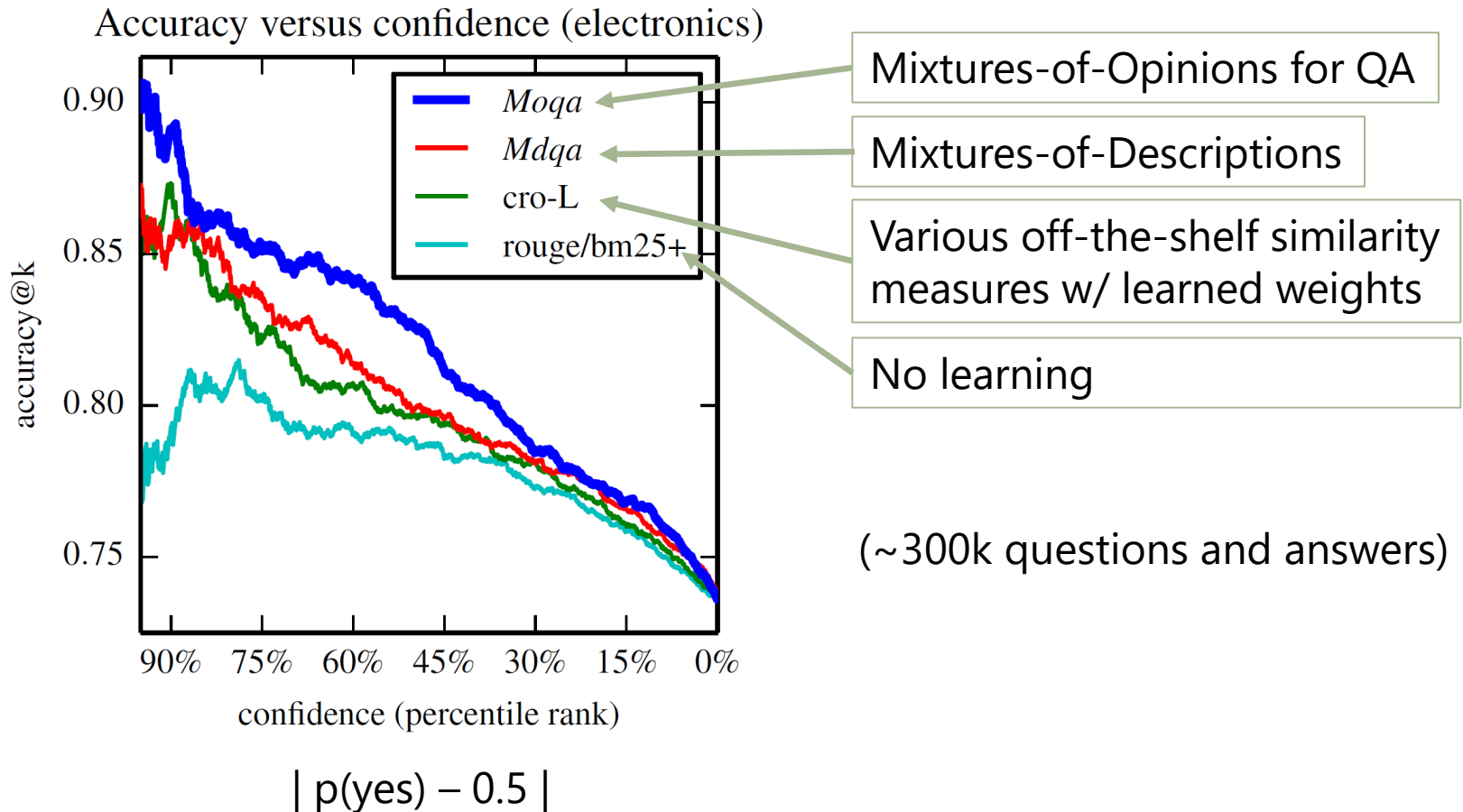
"mixture of experts"

Fit by maximum-likelihood:

**Extracting yes/no questions:**  
"Summarization of yes/no questions using a feature function model" (He & Dai, '11)

$$\ell(\text{corpus}) = \sum_{q \in q_{\text{yes}}} \log p(\text{yes} \mid q) + \sum_{q \in q_{\text{no}}} \log(1 - p(\text{yes} \mid q))$$

# Evaluation – binary questions




# Evaluation – user study

## mturk interface:

**Instructions**

Consider a customer's query about the following product:

Think King Mighty Buggy Hook for Stroller, Wheelchair, Rollator, Walker, 2 Pack



" Since the hooks attach with velcro, do they slide or do they stay in place? "

Which of the following sentences is **most relevant** to the above question?

"I originally purchased the Mommy Hooks for our stroller and loved the durability of the metal and being able to put large amounts of stuff on them, but i ended up hating how big and clunky they are especially when folding the stroller and they are not stationary, always sliding around." ☐

"With the hooks attached at the highest part of the main handle, bags that are hung from the hooks press against both the bassinet and the footrest of the backwards-facing seat, but not in such a way that the hooks are unusable." ☐

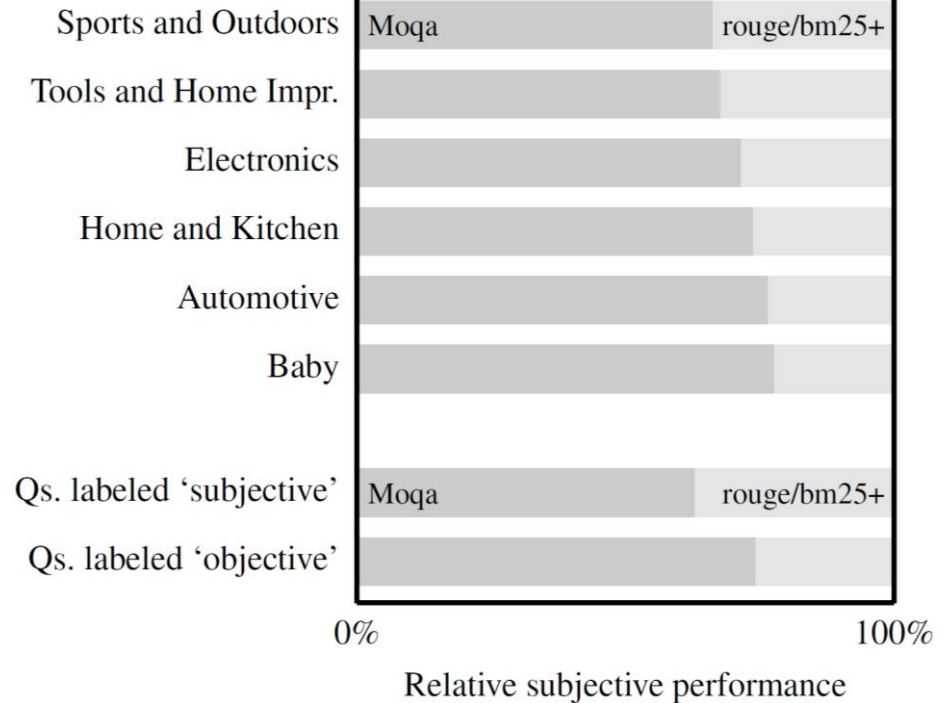
"The hooks stay in place even with multiple bags hanging on it." ☐

Would you say that this question is **subjective**?

Yes ☐

No ☐

## Mechanical turk study



# Evaluation – binary examples

**Product:** Schwinn Searcher Bike ([amazon.com/dp/B007CKH61C](https://www.amazon.com/dp/B007CKH61C))

**Question:** "Is this bike a medium? My daughter is 5'8"."

**Ranked opinions:** "The seat was just a tad tall for my girl so we actually sawed a bit off of the seat pole so that it would sit a little lower." (yes, .698); "The seat height and handlebars are easily adjustable." (yes, .771); "This is a great bike for a tall person." (yes, .711)

**Response:** Yes (.722)

**Actual answer:** My wife is 5'5" and the seat is set pretty low, I think a female 5'8" would fit well with the seat raised



**Product:** Davis & Sanford EXPLORERV ([amazon.com/dp/B000V7AF8E](https://www.amazon.com/dp/B000V7AF8E))

**Question:** "Is this tripod better than the AmazonBasics 60-Inch Lightweight Tripod with Bag one?"

**Ranked opinions:** "However, if you are looking for a steady tripod, this product is not the product that you are looking for" (no, .295); "If you need a tripod for a camera or camcorder and are on a tight budget, this is the one for you." (yes, .901); "This would probably work as a door stop at a gas station, but for any camera or spotting scope work I'd rather just lean over the hood of my pickup." (no, .463)

**Response:** Yes (.863)

**Actual answer:** The 10 year warranty makes it much better and yes they do honor the warranty. I was sent a replacement when my failed.

## Follow-up work

- **ICDM 2016** (with M. Wan)
- Adds “**personalization**” terms to the model to capture quirks of the questioner and answerer
- Considers the *distribution* of answers to each question
  - Generalization to open-ended questions
- Considers various product metadata



## 2. Generative models of reactions





# Richer recommenders

have:

$$f(u, i) : U \times I \rightarrow \{1, 2, 3, 4, 5\}$$

want:

$$f(u, i) : U \times I \rightarrow$$

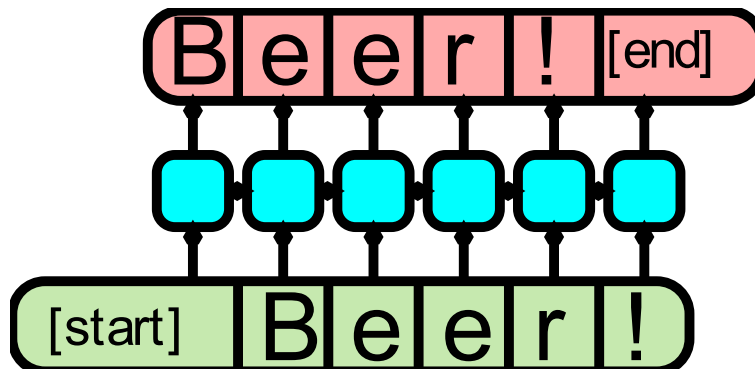
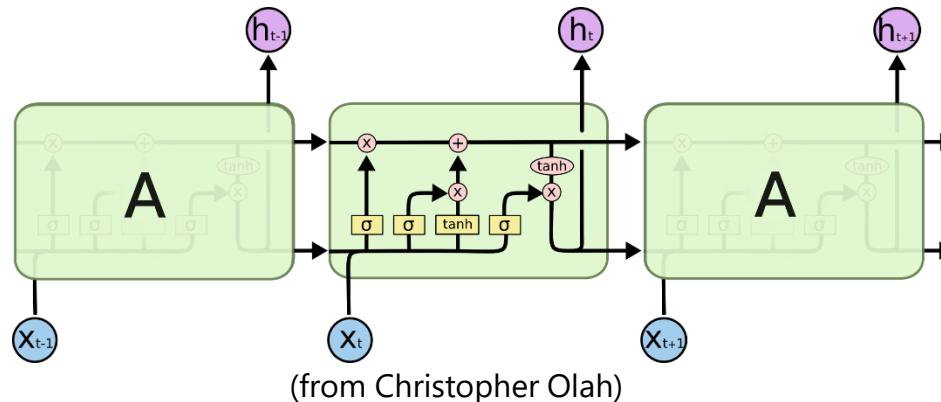
Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from 'Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to 'Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

- "Richer" recommendations, but can also be "reversed", and used for search

# Generative models of text

(a) Standard generative RNN

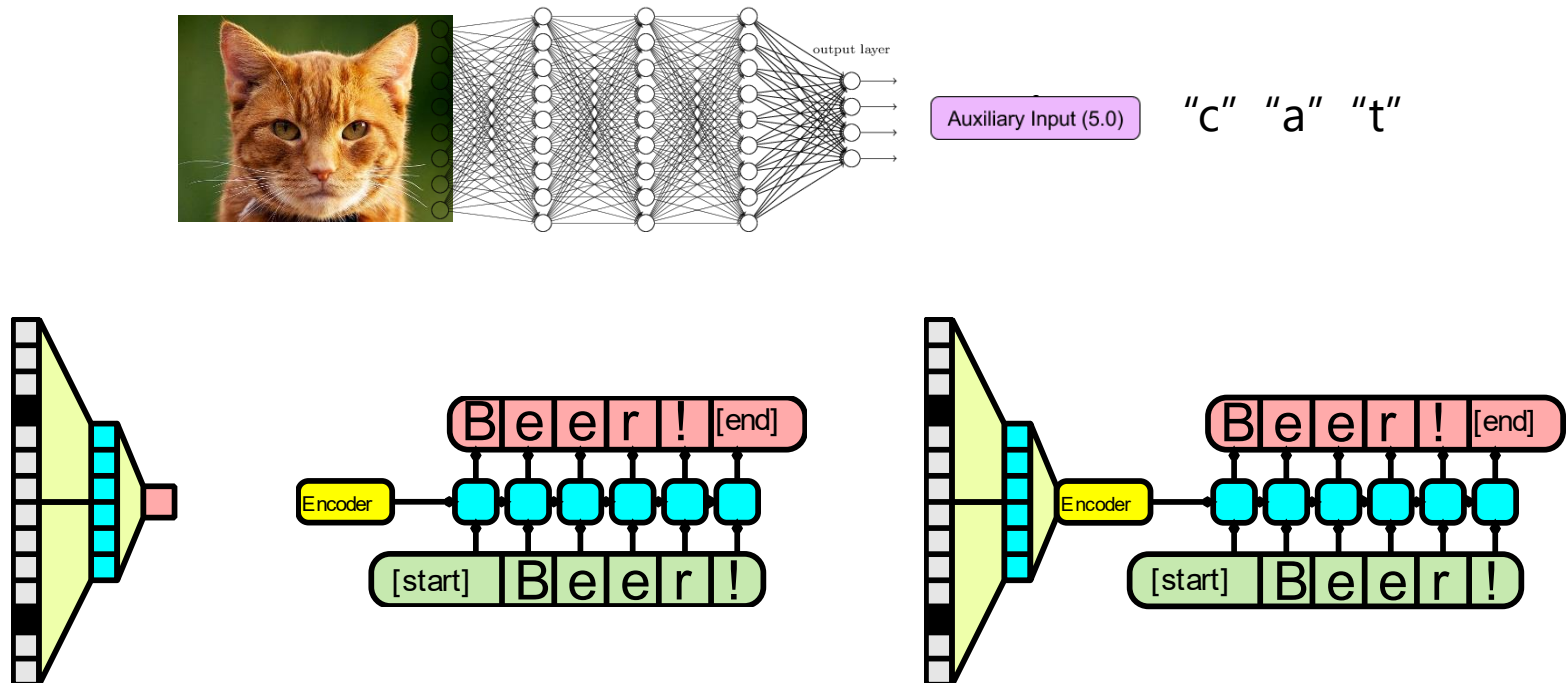


- train on ~200k reviews
- generate new reviews following the language model
- generates "plausible" reviews, **but isn't personalized**

(see e.g. "Learning to generate reviews and discovering sentiment", Radford et al. 2017)

# Need a model of users / items

## (b) Encoder-decoder RNN

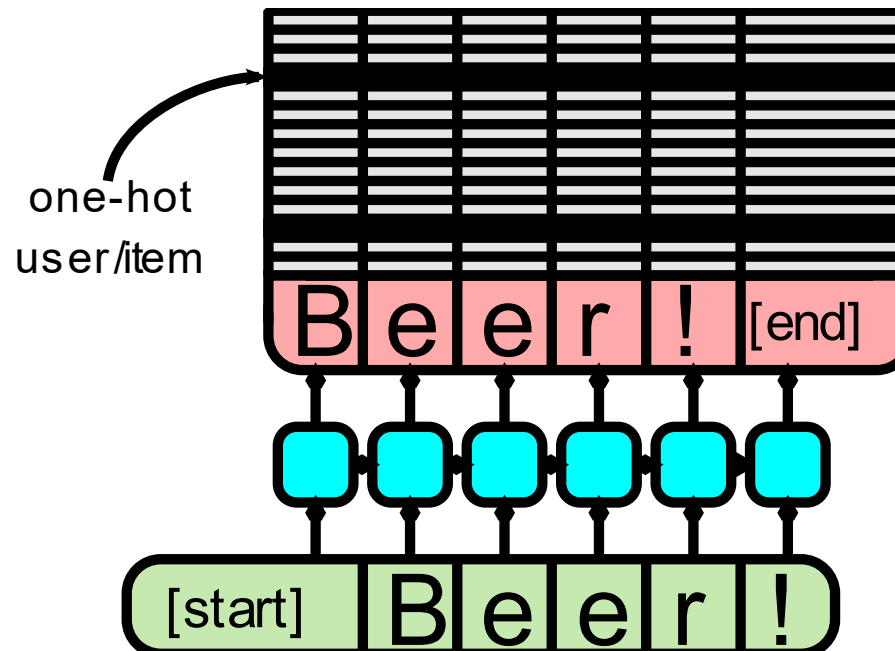


- Is personalized, but struggles with **long sequences**

(see e.g. "Neural rating regression with abstractive tips generation", Li et al. 2017)

# Need a model of users / items

## (c) "Generative Concatenative" RNN



(see e.g. "Generative Concatenative Networks", Lipton et al. 2017)

# Generating reviews

Poured from 12oz bottle into half-liter Pilsner Urquell branded pilsner glass. **Appearance:** Pours a cloudy golden-orange color with a small, quickly dissipating white head that leaves a bit of lace behind. **Smell:** Smells HEAVILY of citrus. By heavily, I mean that this smells like kitchen cleaner with added wheat. **Taste:** Tastes heavily of citrus- lemon, lime, and orange with a hint of wheat at the end. Mouthfeel: Thin, with a bit too much carbonation. Refreshing. **Drinkability:** If I wanted lemonade, then I would have bought that.

Actual review

Poured from a 12oz bottle into a 16oz Samuel Adams Perfect Pint glass. **Appearance:** Very pale golden color with a thin, white head that leaves little lacing. **Smell:** Very mild and inoffensive aromas of citrus. **Taste:** Starts with the same tastes of the citrus and fruit flavors of orange and lemon and the orange taste is all there. There is a little bit of wheat that is pretty weak, but it is sort of harsh (in a good way) and ends with a slightly bitter aftertaste. Mouthfeel: Light body with a little alcohol burn. Finish is slightly dry with some lingering spice. **Drinkability:** A decent beer, but not great. I don't think I would rate this anytime soon as it says that there are other Belgian beers out there, but this is a good choice for a warm day when it's always available in the North Coast Brewing Company party.

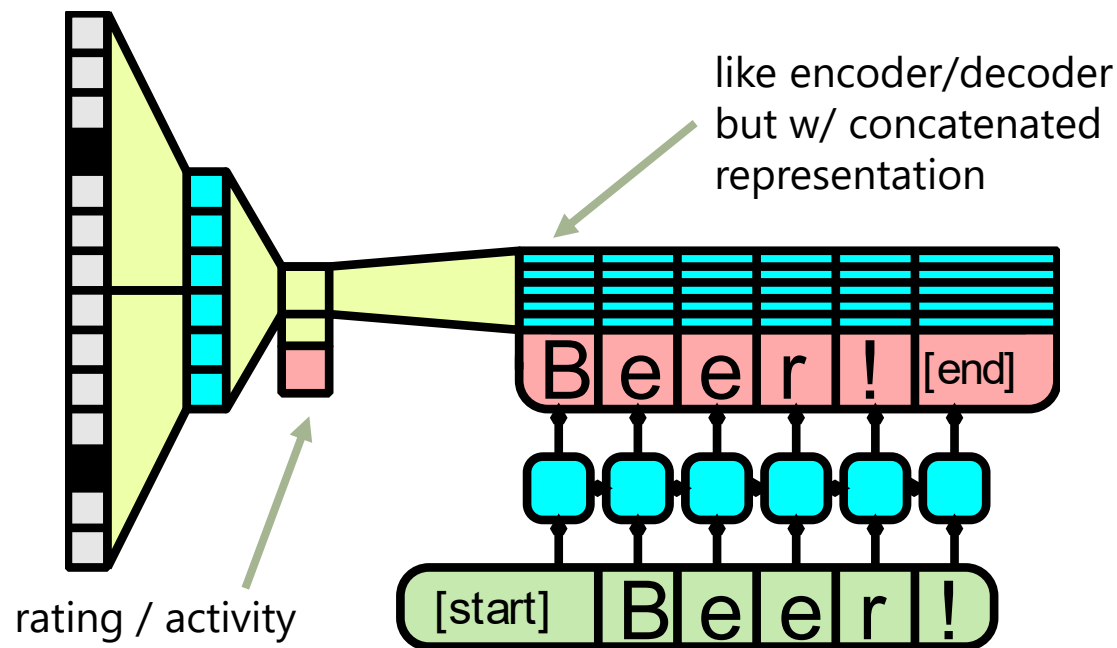
Synthetically generated review

Yes but...

- Requires on the order of ~1 week of training to handle ~200k reviews
- Requires ~100 reviews per user/item to learn a reasonable representation
- Still not particularly useful as a “recommender system”

# Low-rank concatenative networks

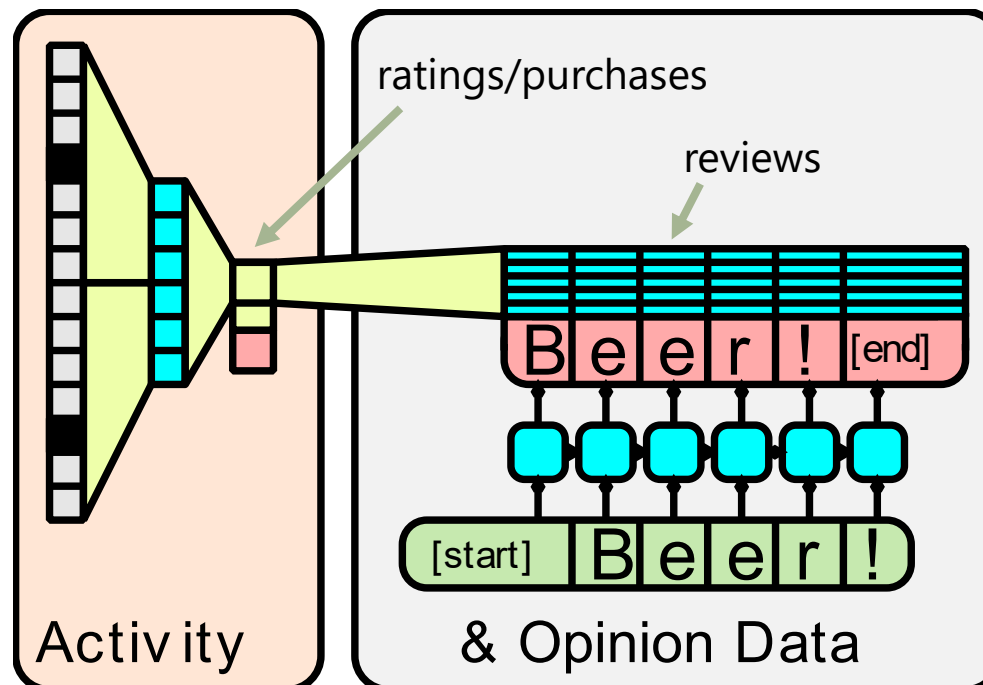
## (d) Low-rank Generative Concatenative RNN



- Facilitates much more efficient training
- Simultaneously predicts preferences and generates reviews

# Semi-supervised Low-rank concatenative networks

## (e) Semi-supervised Generative Concatenative RNN



- Can now train on millions of ratings + a limited number of reviews
- Can predict reviews even for users who have written none!



# Generating reviews

12 oz. bottle, excited to see a new Victory product around, **A: Pours a dark brown, much darker than I thought it would be, rich creamy head, with light lace.** **S:** Dark cedar/pine nose with some dark bread/pumpernickel. **T:** This ale certainly has **a lot of malt**, bordering on Barleywine. **Molasses, sweet maple** with a clear **bitter** melon/white grapefruit hop flavour. Not a lot of complexity in the hops here for me. Booze is noticable. **M:** Full-bodied, creamy, resinous, nicely done. **D:** A good beer, it isn't exactly what I was expecting. **In the end above average**, though I found it monotonous at times, hence the 3. **A sipper for sure.**

Actual review

**A: Pours a very dark brown** with a **nice finger of tan head that produces a small bubble and leaves decent lacing** on the glass. **S:** Smells like a nut brown ale. It has **a slight sweetness** and a bit of a woody note and a little cocoa. The nose is **rather malty** with some chocolate and coffee. The taste is strong but not overwhelmingly sweet. The sweetness is overpowering, but not overwhelming and is a pretty strong **bitter** finish. **M:** Medium bodied with a slightly thin feel. **D: A good tasting beer. Not bad.**

Synthetically generated review

# Recommending products

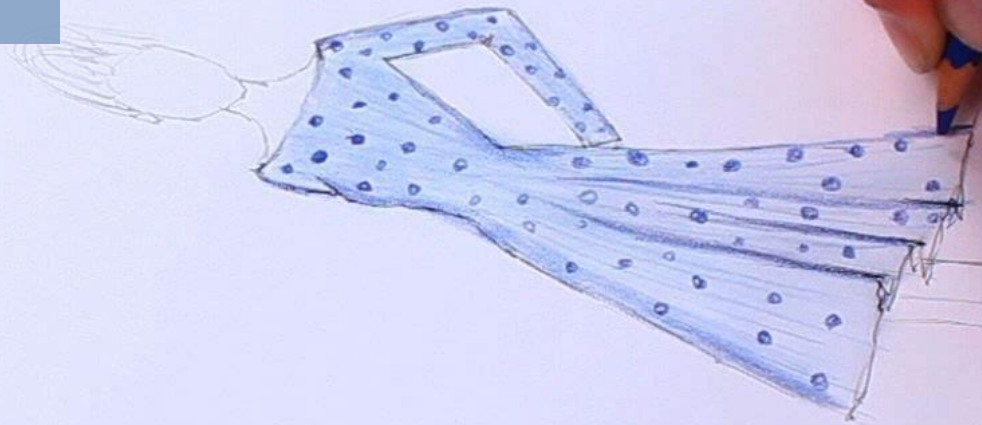
- We can see (modest) improvements in perplexity over non-personalized language models

Dataset	char-LSTM	CF-GCN
BeerAdvocate	2.370	2.329
Amazon Electronics	3.033	2.959
Yelp	2.916	2.809

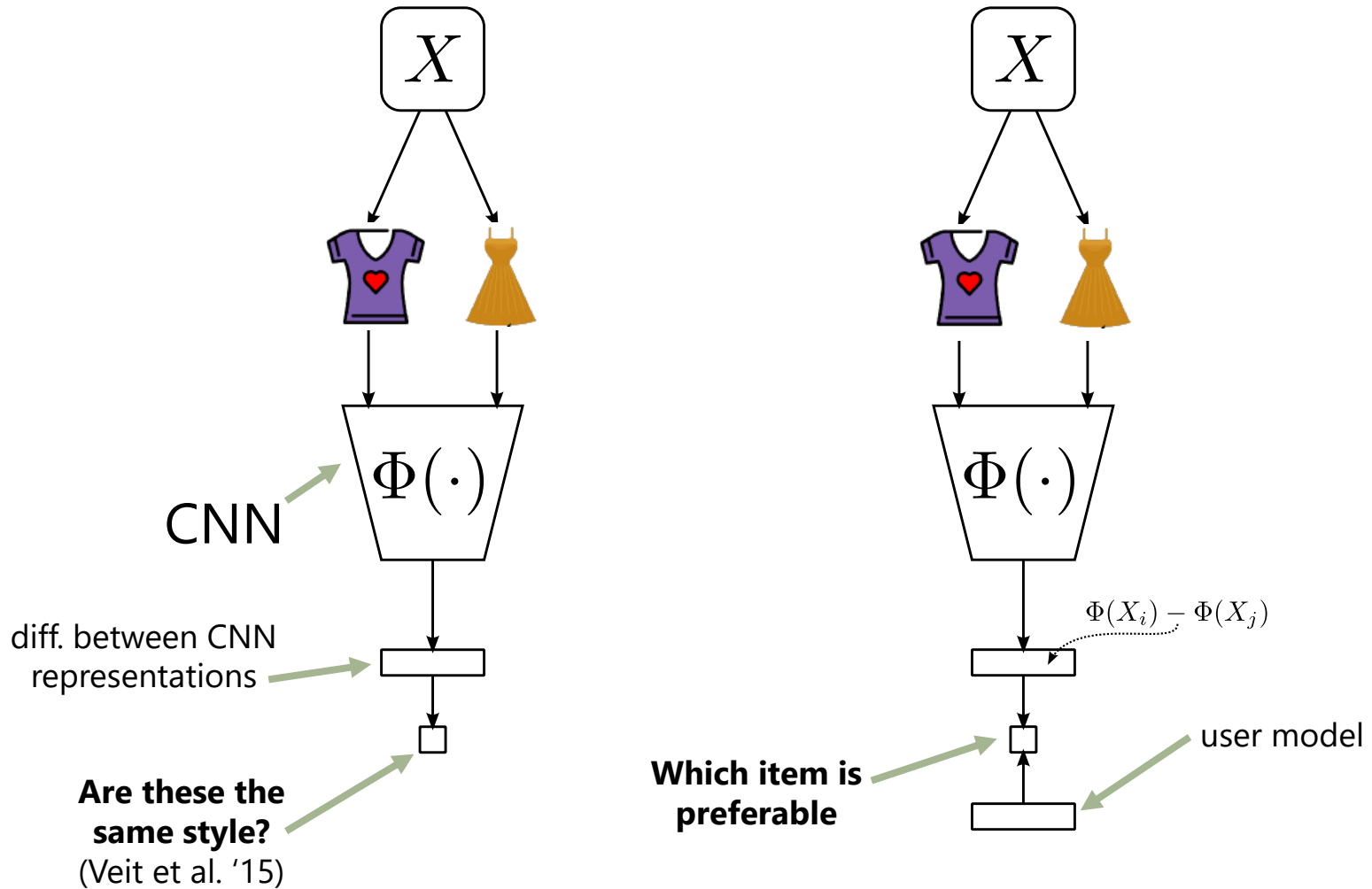
- Having a better language model can also lead to better recommendations (in terms of AUC):

Dataset	BPR	GMF	CF-GCN
<u>BeerAdvocate</u>	0.826	0.847	<b>0.861</b>
<u>Amazon Electronics</u>	0.690	0.746	<b>0.779</b>
<u>Yelp</u>	0.899	0.895	<b>0.902</b>

### 3. Generative models of content

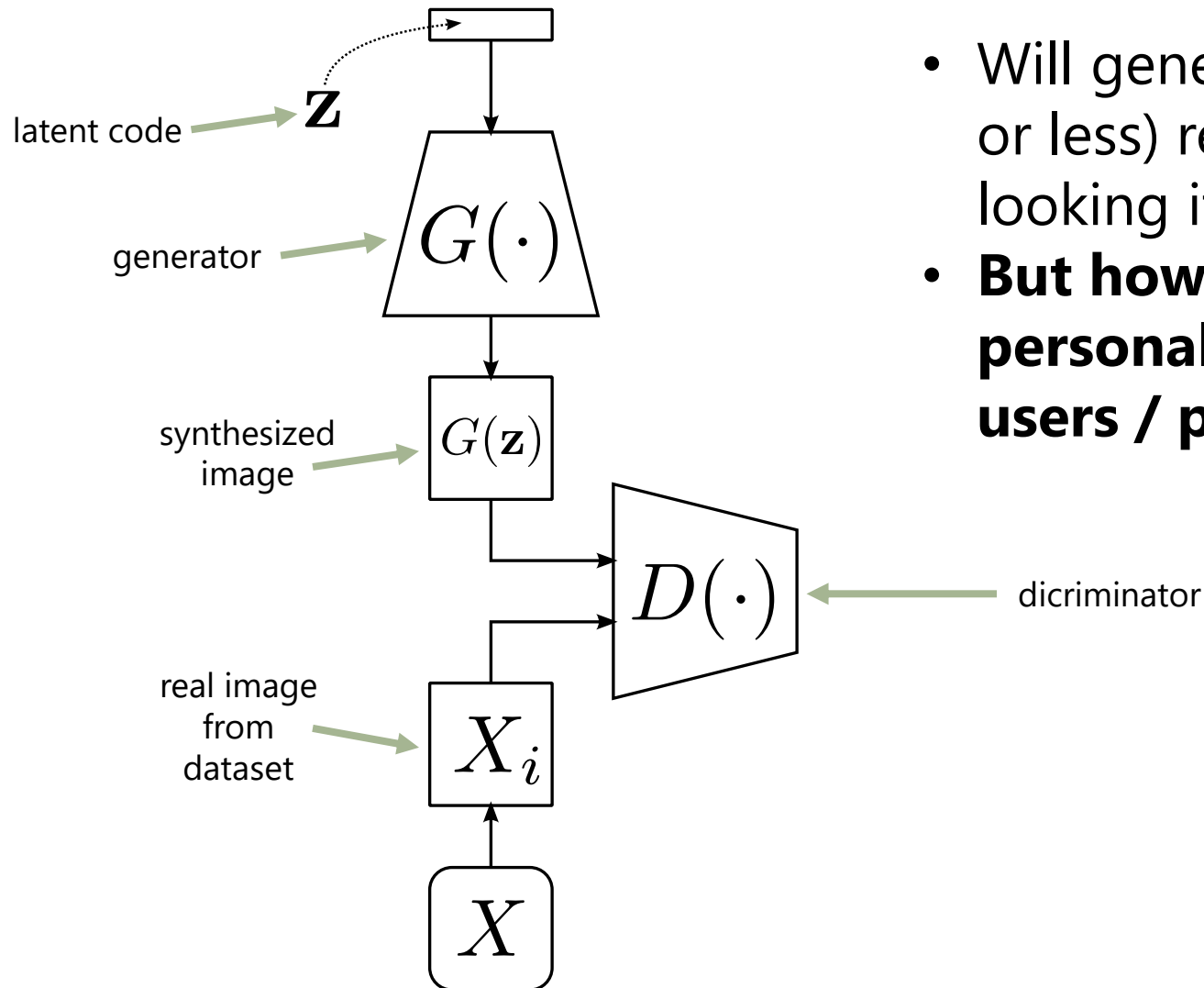


# Image models for recommendation



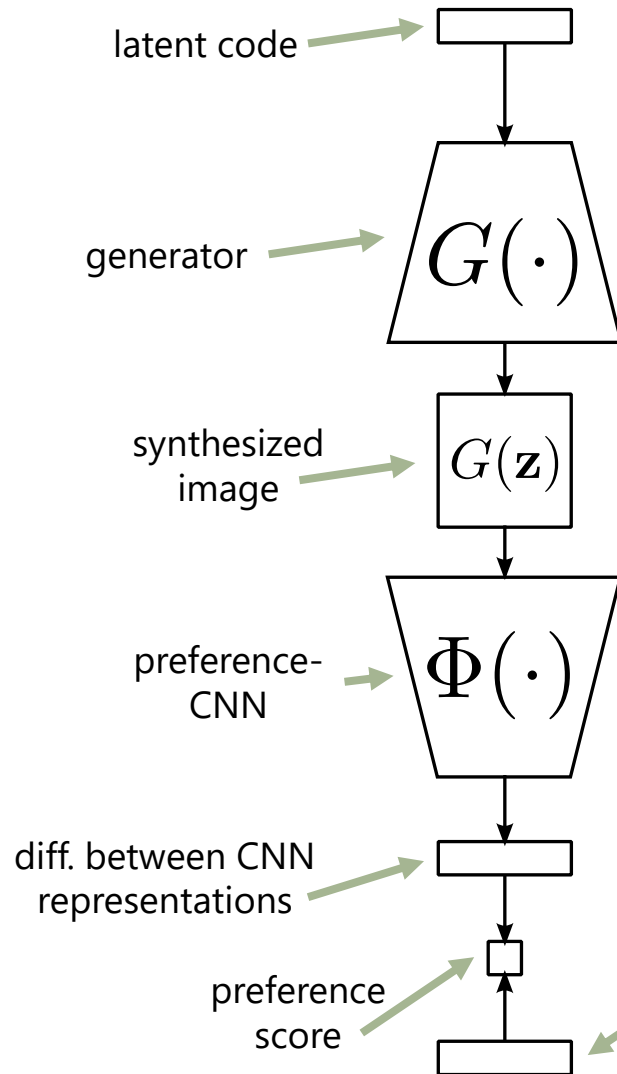
(see e.g. "Siamese Nets", Hadsell et al. 2006)

# Could they also be used for design?



- Will generate (more or less) realistic looking items
- **But how can it be personalized to users / populations?**

# Simple GAN architecture



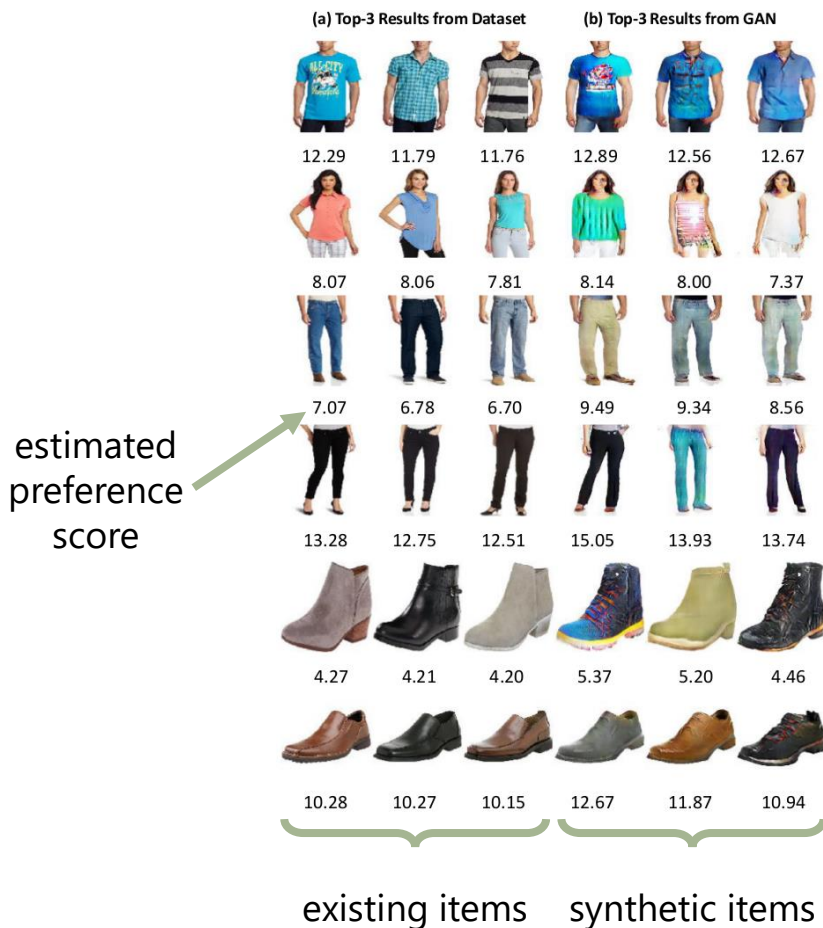
- Generated items are now personalized to each individual user

$$\arg \max_{e \in G(\cdot)} \underbrace{\theta_u^T \Phi(e)}_{\text{preference score}} - \underbrace{\eta [D(e) - 1]^2}_{\text{'plausibility'}}$$



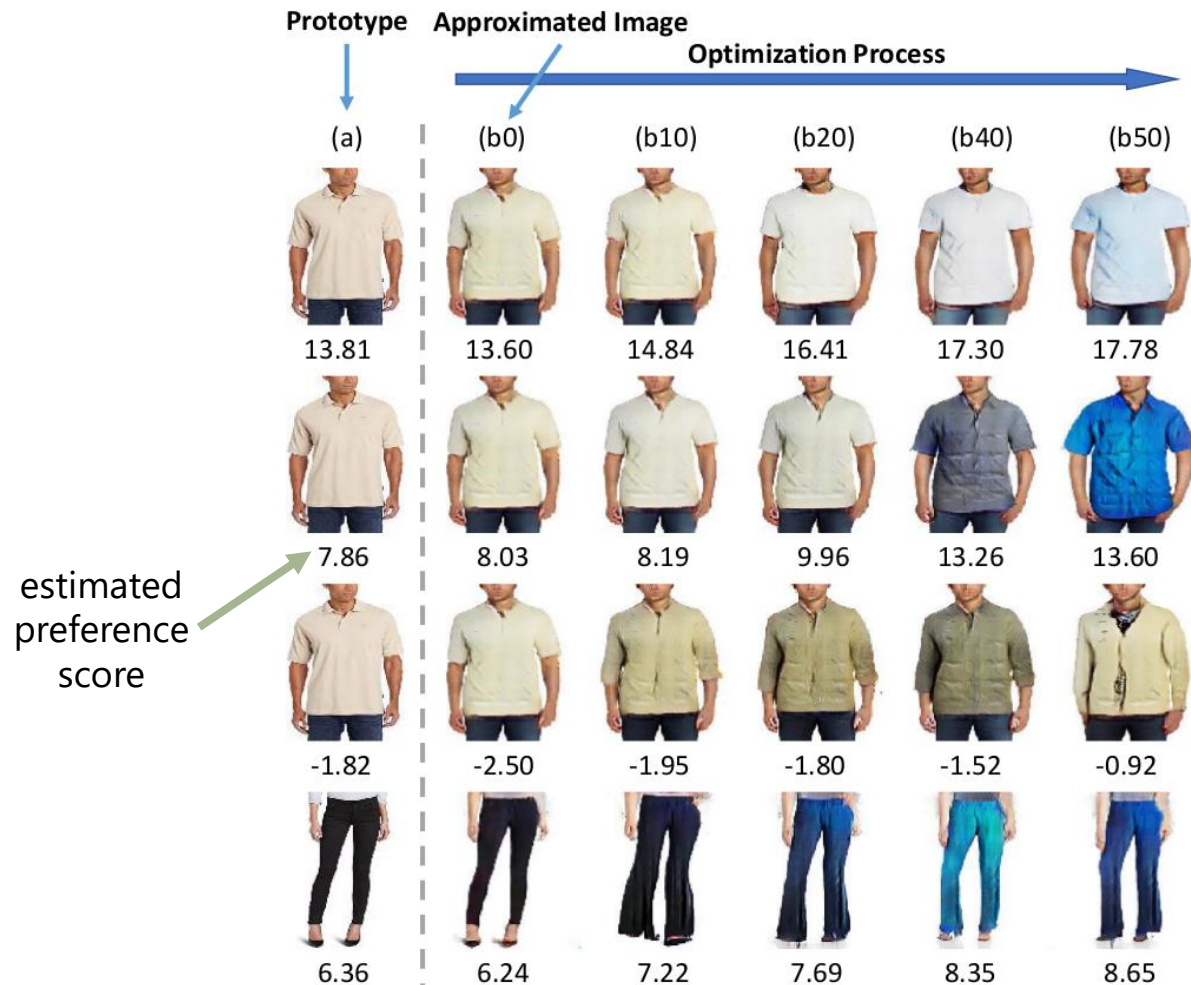
# GAN-generated outfits

- Sample new items matching users' preferences



# Optimization of existing content

- Optimize existing items to better match user





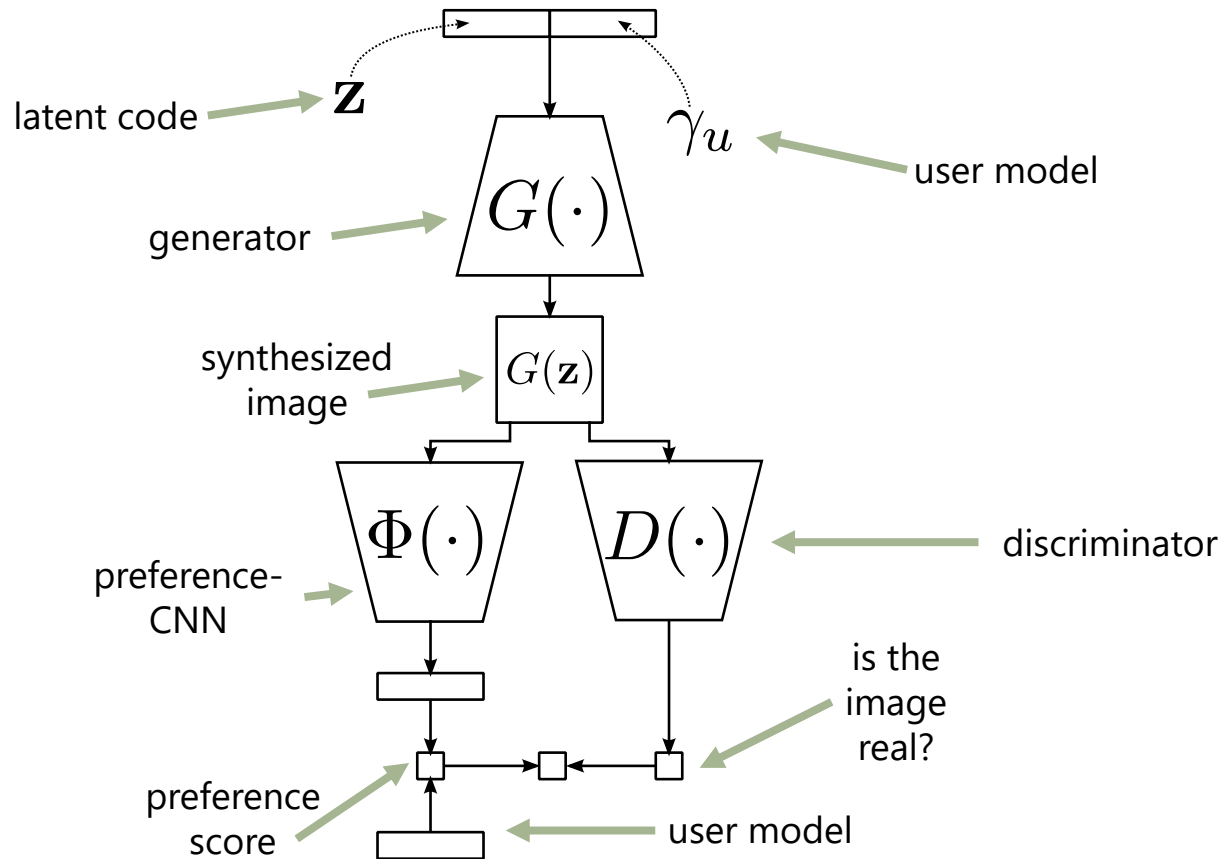
# Recommendation

- Using a “deep” model leads to better results in terms of traditional recommendation objectives (e.g. AUC)

Dataset	BPR (no visual features)	VBPR (pretrained CNN)	'Deep' VBPR
Amazon fashion	0.628	0.748	<b>0.796</b>
Tradesy.com	0.586	0.750	<b>0.786</b>

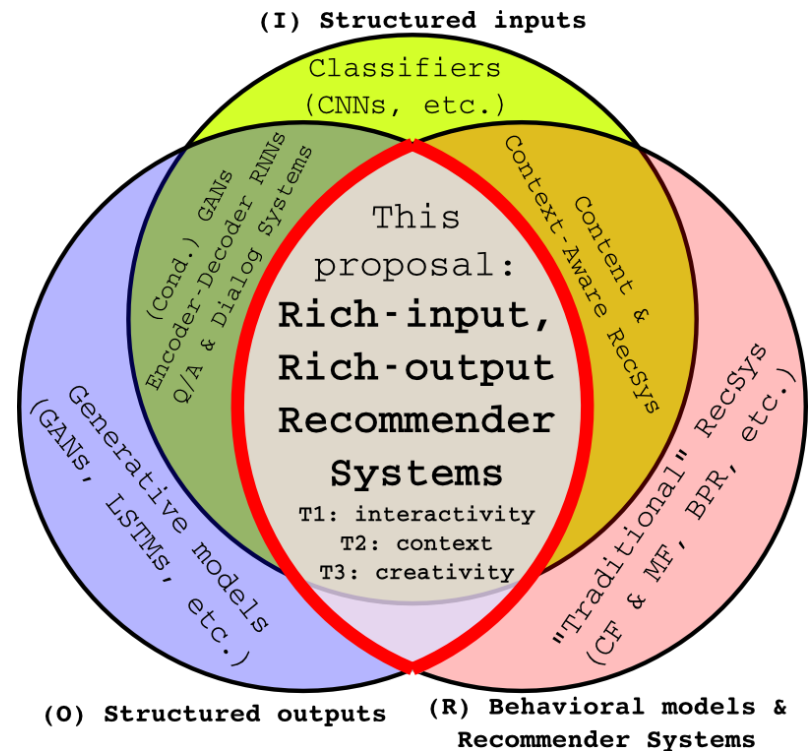
# Future work: joint training schemes

- In the future, we hope to investigate joint training schemes that simultaneously learn to generate and personalize



# Summary

- New class of models and applications at the intersection of structured input/output modeling and "traditional" recommender systems
- As well as generating rich output types, such systems can also improve performance at traditional recommendation objectives



# Summary

- Besides recommender systems, such models can also be applied to data like medical dialogues (top) and heartrate data (bottom)
- In both cases, we need to generate complex, structured outputs, while also accounting for variance between users

## Neurotology intake dialogue:

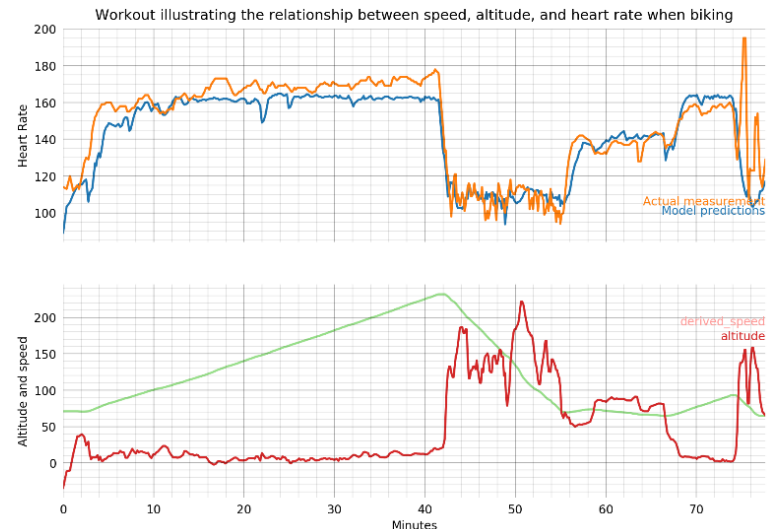
Q: Where in your head do you hear the click?

A: I am not sure what area the clicking comes from but I hear it in my head almost as if it is behind my throat.

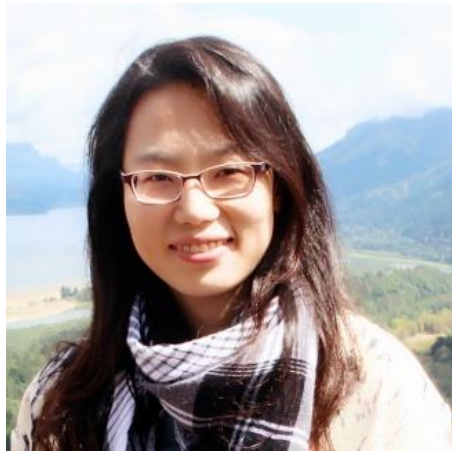
Q: Does the clicking sound only happen when your head is forward or does it happen other times?

A: I seem to hear it when I am walking in a quiet environment and my head is down. It seems when I stand straight up, I don't hear it as much.

## Exercise heartrate data:



# Thanks!



- Mengting Wan (personalized Q/A)
- Zachary Lipton, Jianmo Ni (generative models of text)
- Wang-Cheng Kang (generative image models)

code and data on: <http://cseweb.ucsd.edu/~jmcauley/>