13th ACM Conference on Recommender Systems

Jakub Macina
Deep Learning Vienna Meetup
September 24th 2019

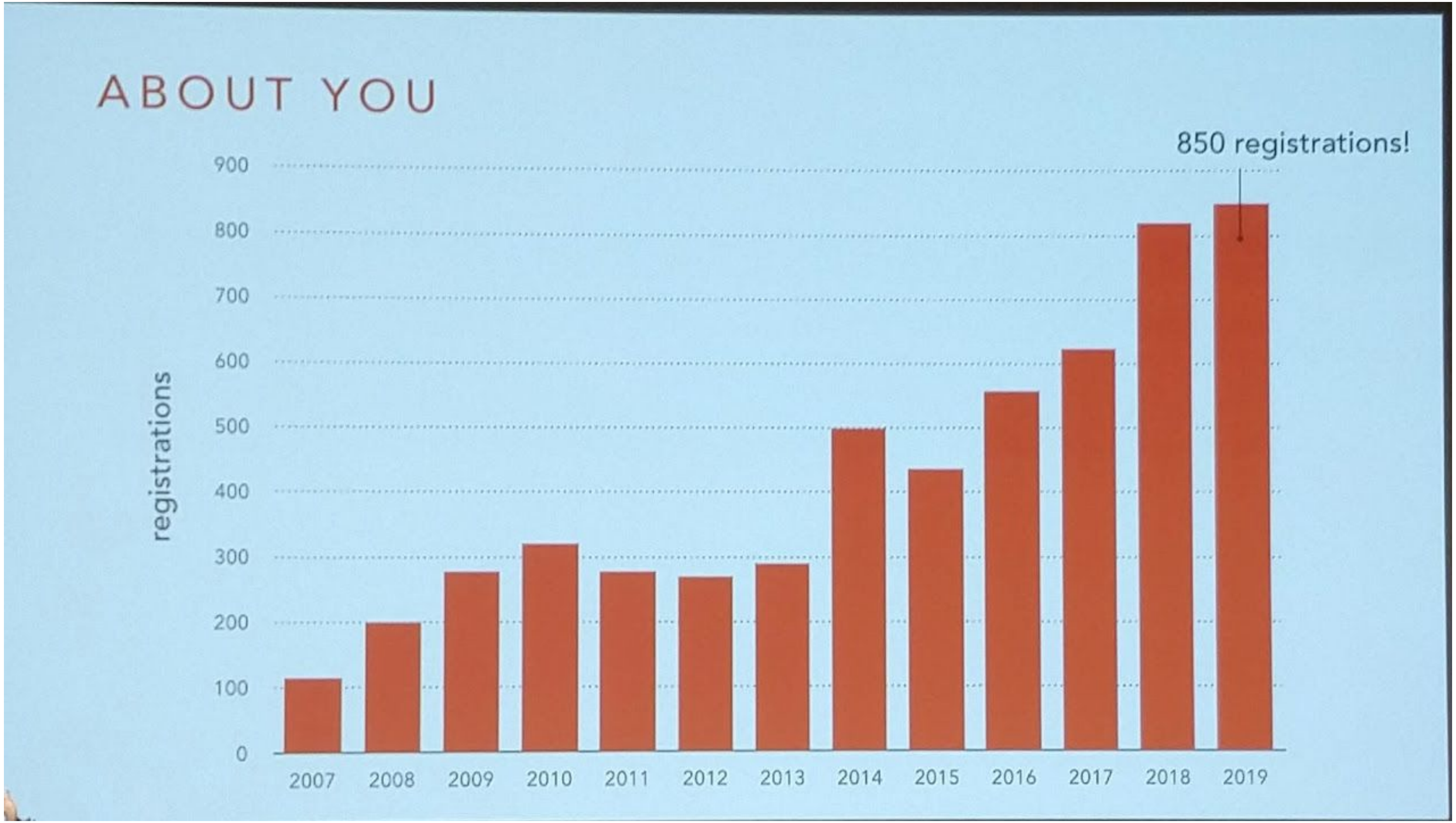
RECSYS 2019

EXPONEA

RecSys 2019

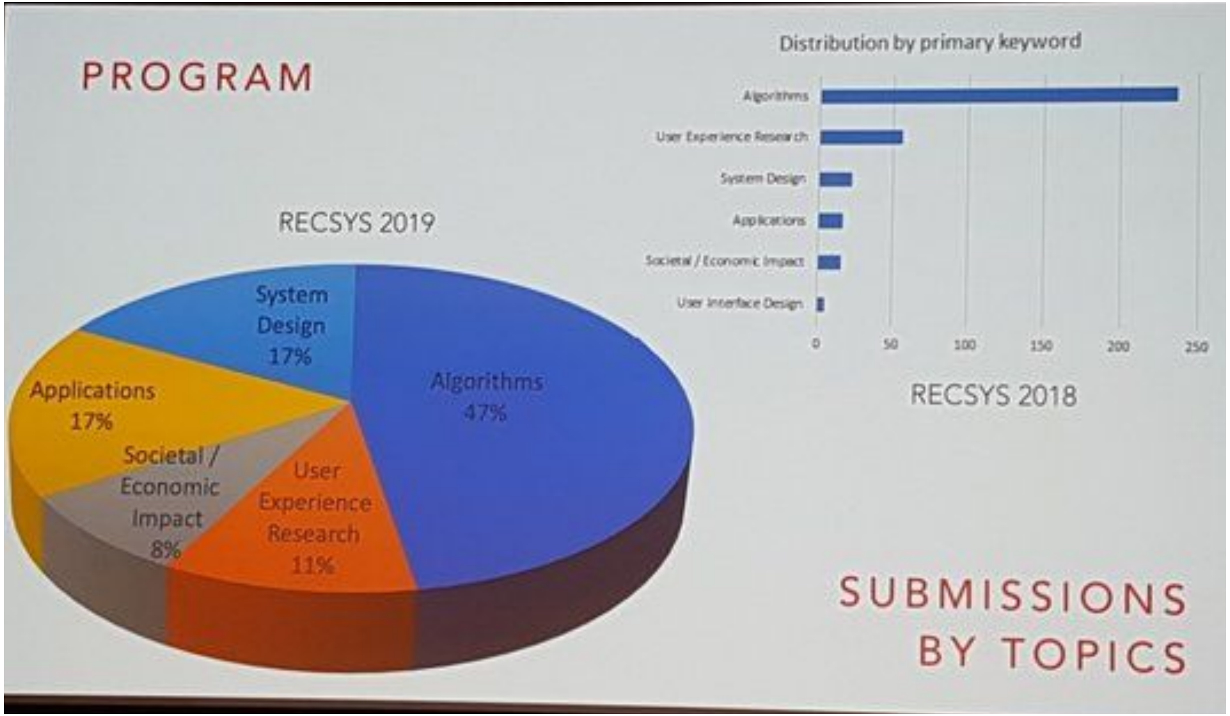
- premier international forum for the presentation of new research results, systems and techniques in the broad field of recommender systems
- Copenhagen, Denmark
- 909 participants
  - 73.4% from industry
- Acceptance rate
  - long papers - 19%
  - short papers - 24%

RecSys 2019 sold out



RecSys 2019 keywords

RecSys 2019 topics

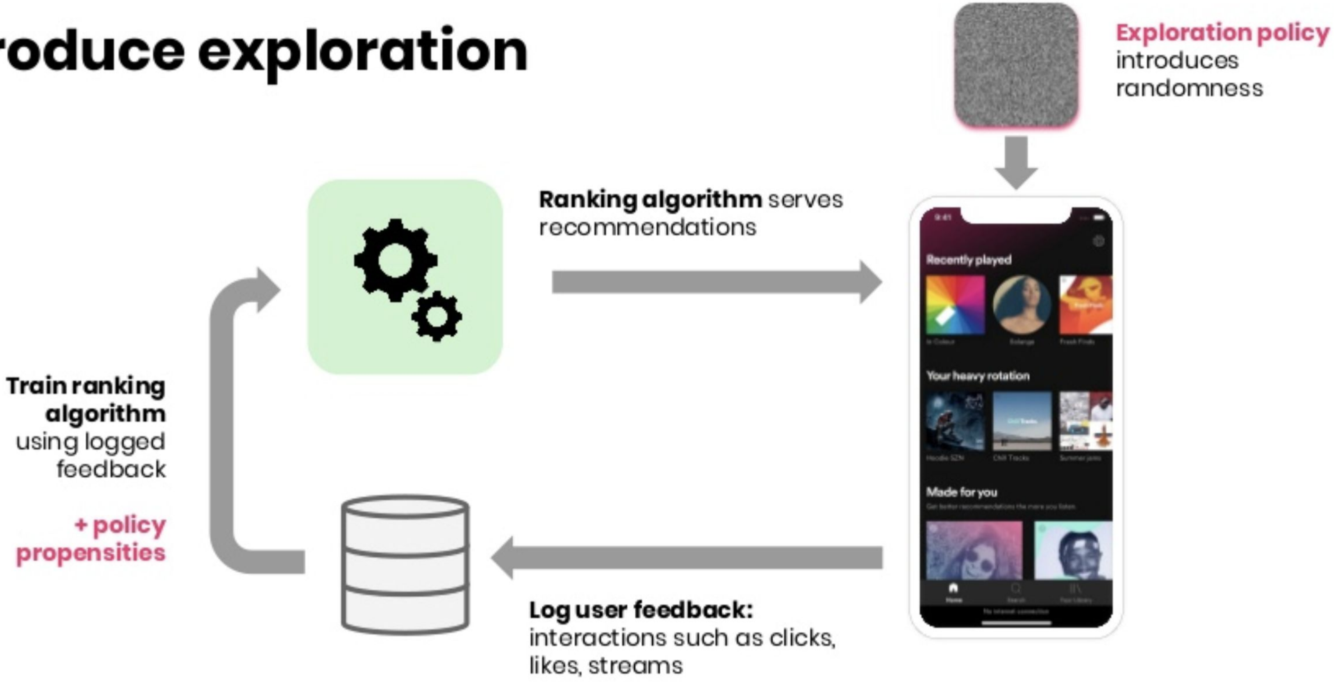


RecSys 2019 recent research challenges

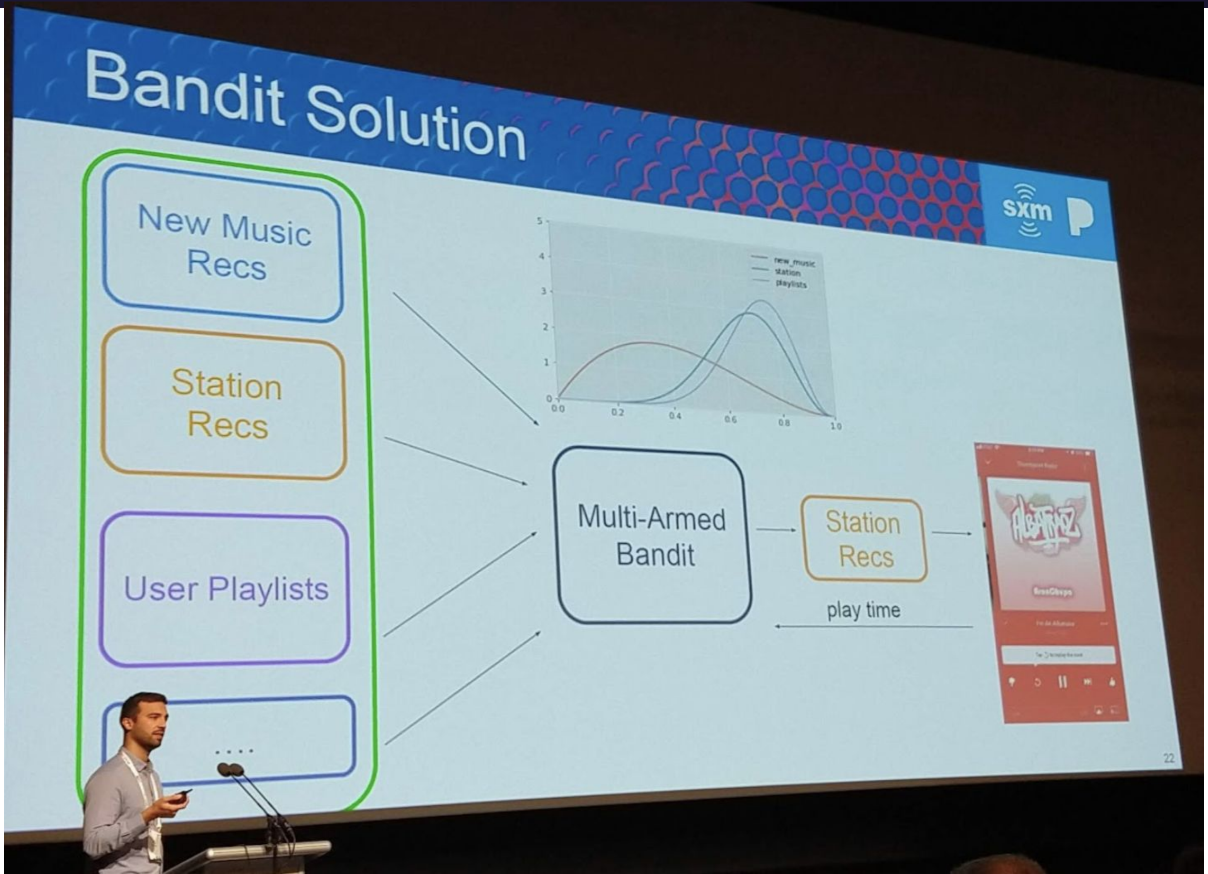
- Multi-armed bandits and Reinforcement Learning
- Offline evaluation and causality
- Fairness and responsible recommendation

Spotify: Multi-armed bandits for homepage recommendation

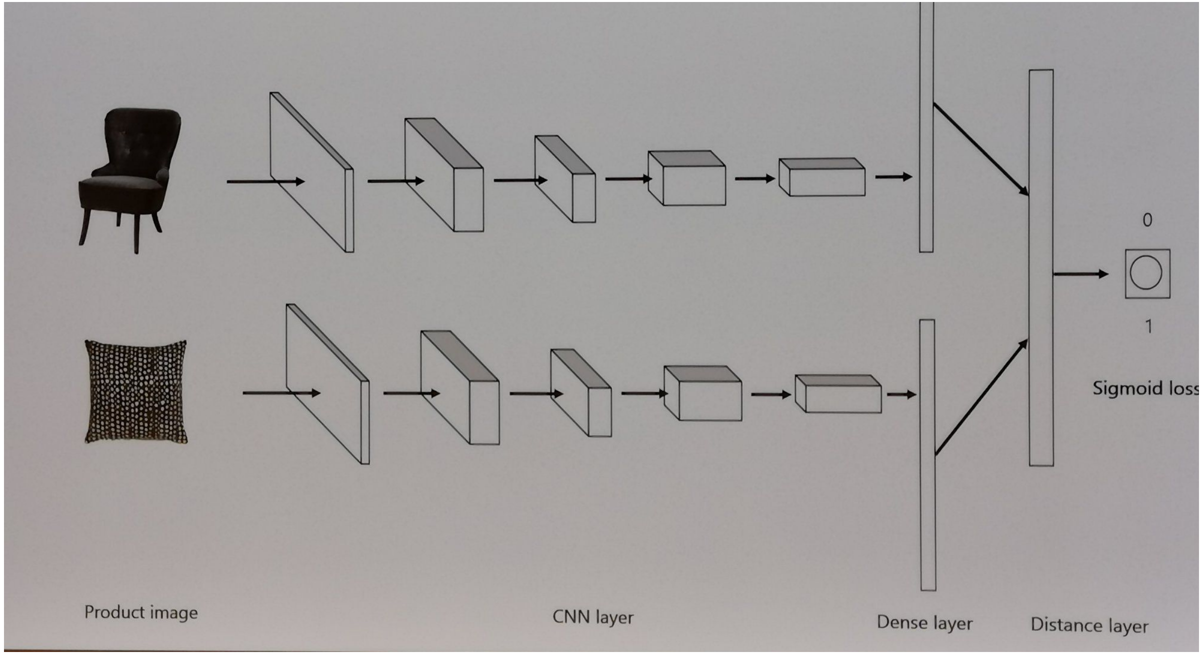
Introduce exploration



Pandora: Multi-armed bandits for homepage recommendations



Ikea: Siamese neural networks for add-to-cart recommendations



0 : Products with different color, material, series and never bought together

1 : Put together by designers + transactional data

Best paper award: Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

- top-n recommendations task (trained on interactions without additional metadata)
- DL approaches did not consistently outperform a well-tuned non-neural linear ranking method
- reproducibility is low
  - code, data

Table 1: Reproducible works on deep learning algorithms for top-n recommendation per conference series from 2015 to 2018.

Conference	Rep. ratio	Reproducible
KDD	3/4 (75%)	[17], [23], [48]
RecSys	1/7 (14%)	[53]
SIGIR	1/3 (30%)	[10]
WWW	2/4 (50%)	[14], [24]
Total	7/18 (39%)	

Non-reproducible: KDD: [43], RecSys: [41], [6], [38], [44], [21], [45], SIGIR: [32], [7], WWW: [42], [11]

Best paper award: Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

- reproducibility is low
    - evaluation and hyperparameter optimization is not shared
- baselines
    - wrong choice
    - lack of optimization
- No standard dataset - 20+ public datasets
- Hunt for better accuracy
    - Wrong choice of metrics

Table 6: Experimental results for NCF.

	Pinterest			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1663	0.1065	0.2744	0.1412
UserKNN	0.7001	**0.5033**	0.8610	**0.5557**
ItemKNN	**0.7100**	**0.5092**	**0.8744**	**0.5629**
$P^3\alpha$	0.7008	**0.5018**	0.8667	**0.5559**
$RP^3\beta$	**0.7105**	**0.5116**	**0.8740**	**0.5650**
NeuMF	0.7024	0.4983	0.8719	0.5536
	Movielens 1M			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.3043	0.2062	0.4531	0.2542
UserKNN	0.4916	0.3328	0.6705	0.3908
ItemKNN	0.4829	0.3328	0.6596	0.3900
$P^3\alpha$	0.4811	0.3331	0.6464	0.3867
$RP^3\beta$	0.4922	0.3409	0.6715	0.3991
NeuMF	0.5486	0.3840	0.7120	0.4369
SLIM	**0.5589**	**0.3961**	**0.7161**	**0.4470**

Best paper award: Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

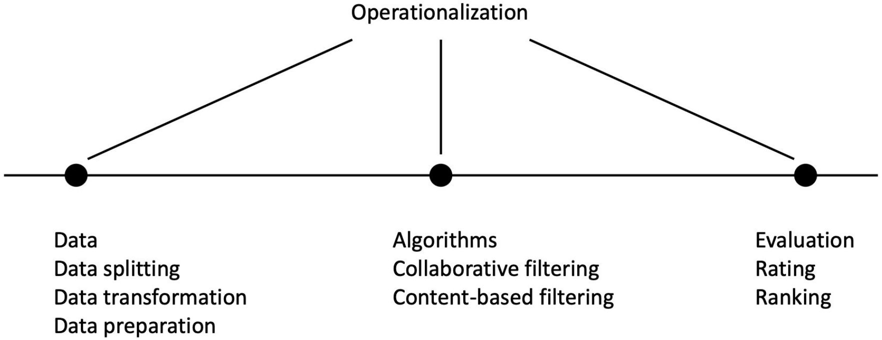
- DL approaches did not consistently outperform a well-tuned non-neural linear ranking method
- Reviewers + community should push for reproducibility
- https://github.com/MaurizioFD/RecSys2019_DeepLearning_Evaluation
- Continue research
    - 253 days of Amazon AWS

DL algorithms outperforming our baselines

Algorithm	CF + CBF	CF + CBF + NP	CF + CBF + NP + SLIM
MCRec	-	-	-
SpectralCF	-	-	-
CMN	4/12 - 30%	-	-
NeuMF	6/12 - 50%	6/12 - 50%	-
CDL	9/24 - 37%	9/24 - 37%	9/24 - 37%
CVAE	9/24 - 37%	9/24 - 37%	9/24 - 37%
Mult-VAE	12/12 - 100%	12/12 - 100%	10/12 - 83%

Microsoft recommenders open source library

- https://github.com/microsoft/recommenders



Algo	MAP	nDCG@k	Precision@k	Recall@k	RMSE	MAE	R^2	Explained Variance
ALS	0.004732	0.044239	0.048462	0.017796	0.965038	0.753001	0.255647	0.251648
SVD	0.012873	0.095930	0.091198	0.032783	0.938681	0.742690	0.291967	0.291971
SAR	0.113028	0.388321	0.333828	0.183179	N/A	N/A	N/A	N/A
NCF	0.107720	0.396118	0.347296	0.180775	N/A	N/A	N/A	N/A
FastAI	0.025503	0.147866	0.130329	0.053824	0.943084	0.744337	0.285308	0.287671

Recommending What Video to Watch Next: A Multitask Ranking System

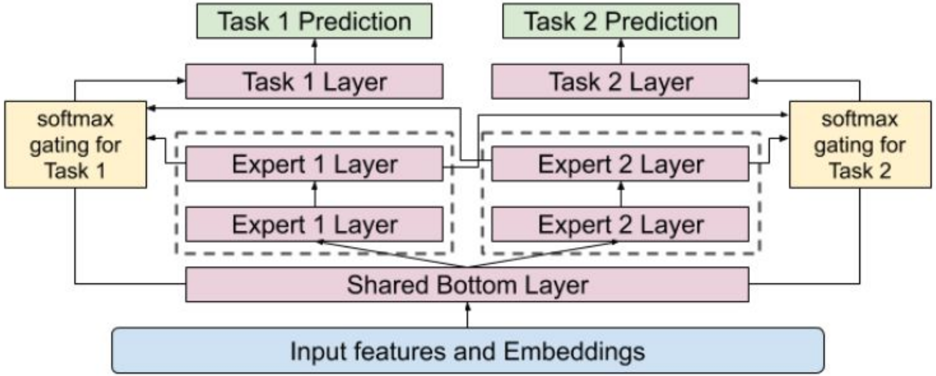
- Google AI
- ranking part - given a video which a user is currently watching, recommend the next video that the user might watch and enjoy
- multiple competing ranking objectives - watching time vs. sharing
    - soft-parameter sharing techniques - Multi-gate Mixture-of-Experts (KDD18)
- implicit selection biases in user feedback - positioning bias
    - shallow tower - Wide & Deep framework

Model Architecture	Number of Multiplications	Engagement Metric	Satisfaction Metric
Shared-Bottom	3.7M	/	/
Shared-Bottom	6.1M	+0.1%	+ 1.89%
MMoE (4 experts)	3.7M	+0.20%	+ 1.22%
MMoE (8 Experts)	6.1M	+0.45%	+ 3.07%

Table 1: YouTube live experiment results for MMoE.

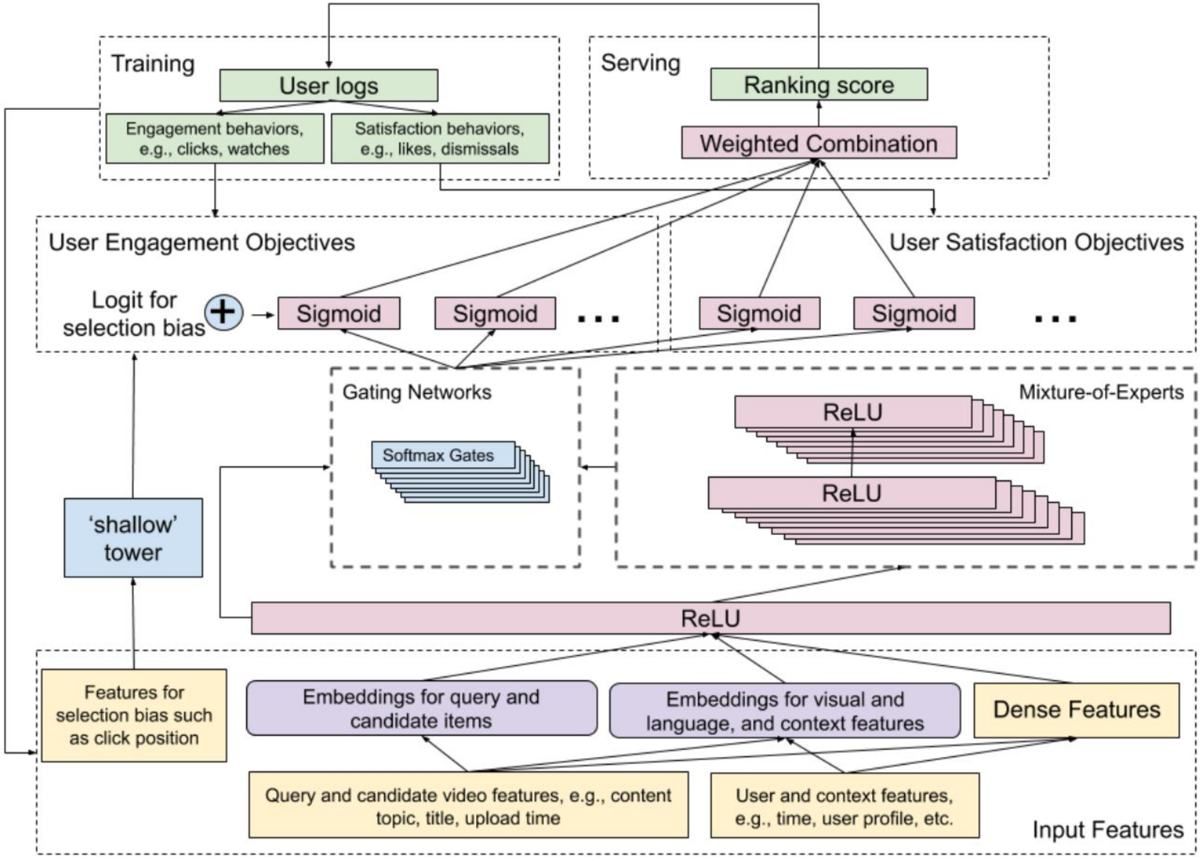
Recommending What Video to Watch Next: A Multitask Ranking System

- objectives
  - 1) engagement objectives - e.g. user clicks
  - 2) satisfaction objectives - e.g. liking a video on YouTube



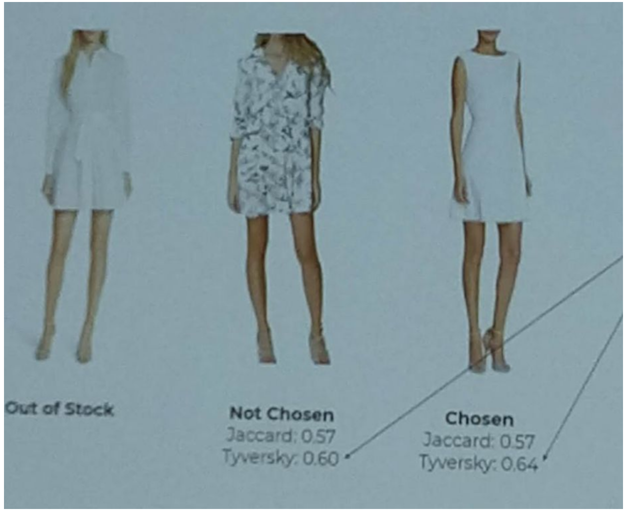
(b) Multi-gate Mixture-of-Expert Model with one shared bottom layer and separate hidden layers for two tasks.

Recommending What Video to Watch Next: A Multitask Ranking System



Users in the Loop: A Psychologically-Informed Approach to Similar Item Retrieval

- User study on Amazon Mechanical Turk
- Psychologically-informed similarity function (i.e., Tversky contrast model) outperforms a psychologically-naive similarity function (i.e., Jaccard similarity)
- Demonstrated that users' behavior violates properties and assumptions of commonly used mathematical similarity metrics such as symmetry (i.e., $Sim(a, b) = Sim(b, a)$)
- Results:
  - Some features of fashion items are more important than others like *dress length and sleeve length*
  - Users' similarity judgments are asymmetric
  - Common features are more influential than distinctive features



$$S_J(a, b) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$S_T(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{1}$$

RecSys 2019 conclusions

- Lot of open research questions
- Other topics:
  - Fashion recommendations
    - Size recommendations
    - Complementary recommendations
  - Negative sampling
  - Multi-stakeholder recommendation
  - Calibration and Bias Disparity
  - GDPR
- Finally no ratings predictions this year
- Everybody is hiring
- Oral presentation quality is low and time was usually just for 1 question

RECSYS 2020

JAKUB MACINA
Machine Learning Engineer

jakub.macina@exponea.com
+421 910 926 484

www.exponea.com

EXPONEA