

# Рекомендательные системы глазами исследователя

Нейросетевые рекомендательные  
системы 2025/26, семинар 3

Кирилл Хрыльченко

Старший преподаватель,  
ФКН ВШЭ

Яндекс

# Что мы хотим?

- Хотим делать значимый ресерч, который будет активно использоваться другими людьми
- Что мы для этого можем делать:
  - Выбирать важные задачи, находить общие, фундаментальные идеи
  - Поддерживать реалистичность экспериментального сетапа
  - Работать над презентацией своей научной работы
  - Прокачивать visibility
  - Завоёывать доверие других людей

# Порочные практики оффлайн-экспериментов

По мотивам следующих докладов с RecSys Summer School 2025:

- Best Practices for Offline Evaluation, Lien Michiels
- Fairness in Recommender Systems, Michael D. Ekstrand

Начнём: если просуммировать приrostы из всех статей, которые клеймят новую SOTA, получим 200+% accuracy. Почему?

# Оффлайн-оценка качества

	Anelli et al. [3]	Rendle et al. [58]
EASE	0.336	0.449
SLIM	0.335	0.447
iALS	0.306	0.453
NeuMF	0.277	0.477

Table 1. NDCG@10 for the ML1M dataset for all algorithms that were evaluated in both Anelli et al. [3] and Rendle et al. [58]. Results taken directly from the respective papers.

# Оффлайн-оценка качества

Процесс оффлайн-оценки качества:

- Есть датасет с историческими данными
- В датасете присутствует пользовательский фидбек
- Датасет делится на **train, validation, test**
- С помощью **метрик** замеряется качество различных алгоритмов

На RecSys'23 оффлайн-оценку качества использовали 87% статей, и во многих случаях – неправильно.

# Аспекты оффлайн-оценки качества

1. Stakeholders & objectives
2. Research questions, гипотезы
3. Датасеты
4. Алгоритмы
5. Метрики

# Stakeholders & objectives

- Что и для кого хотим улучшить?
- Для кого: для пользователей, авторов, платформы
- Что хотим:
  - Онлайн метрики качества
  - Скорость
  - Потребление ресурсов
  - Справедливость (fairness)
- Важны конечные цели, для которых делается ресерч

# RQ и гипотезы

- Хорошо задизайненный эксперимент отвечает на **ресерч вопросы (RQ)** и проверяет гипотезы
- За гипотезами (и RQ) должна стоять какая-то интуиция, теория, логика, философия, и её тоже необходимо объяснять
- **Научный метод:** формулируем гипотезы, ставим эксперименты, делаем выводы, формулируем новые гипотезы, etc

# Что нам важно про данные

Вопросы про датасет:

- Доступен ли он публично?
- Когда он был собран?
- За какой период?
- Было ли какое-то сэмплирование / фильтрации?
- Что за рексистема была на момент генерации данных?
- Кто пользователи? Что за взаимодействия?
- Какие статистики у датасета (количество пользователей, айтемов, взаимодействий)

# MovieLens

MovieLens – очень популярный датасет с рейтингами фильмов:

- Отражает фидбек только тех пользователей, которые пользуются самой платформой
- Специфика платформы – нельзя ничего посмотреть, можно только ставить рейтинги
- Соцдем – мужчины 20+ лет. Почему?
- Средний рейтинг фильмов – 3.6; почему?
- А ещё: внутри одного дня взаимодействия пользователя отсортированы по ID фильма

# Датасеты

- Используйте **публичные датасеты**, чтобы ваши результаты можно было воспроизводить
- Используйте **несколько датасетов**, чтобы можно было делать обобщения
- Используйте **свежие и большие датасеты** вместо устаревших и небольших
- Если вы делаете предобработку данных (e.g., core-фильтрацию) – объясните причину и демонстрируйте как это повлияло на данные
  - Проговаривайте обработку целиком (e.g., фильтрация unseen items / users)
- **Правильный сплит в рекомендациях – temporal split**

# Датасеты

TABLE 2  
Statistics of the six selected datasets.

Dataset		ML-1M	Yelp	LastFM	Epinions	Book-X	AMZe
origin	#User	6,038	1,326,101	1,892	22,164	105,283	4,201,696
	#Item	3,533	174,567	17,632	296,277	340,556	476,002
	#Record	575,281	5,261,669	92,834	922,267	1,149,780	7,824,482
	Density	2.697e-2	2.273e-5	2.783e-3	1.404e-4	3.207e-5	3.912e-6
5-filter	#User	6,034	227,109	1,874	21,995	22,072	253,994
	#Item	3,125	123,985	2,828	31,678	43,748	145,199
	#Record	574,376	3,419,587	71,411	550,117	623,405	2,109,869
	Density	3.046e-2	1.214e-4	1.348e-2	7.895e-4	6.456e-4	5.721e-5
10-filter	#User	5,950	96,168	1,867	21,111	12,720	63,161
	#Item	2,811	80,351	1,530	14,030	18,318	85,930
	#Record	571,549	2,458,153	62,984	434,162	443,196	949,416
	Density	3.412e-2	3.181e-4	2.205e-2	1.466e-3	1.902e-3	1.749e-4
Timestamp		✓	✓	✗	✓	✗	✓

# Датасеты

## A Critical Study on Data Leakage in Recommender System Offline Evaluation

YITONG JI, Nanyang Technological University, Singapore  
AIXIN SUN, Nanyang Technological University, Singapore  
JIE ZHANG, Nanyang Technological University, Singapore  
CHENLIANG LI, Wuhan University, China

We show the temporal dynamics of users and items through their average active time in four major datasets and highlight that users' last interactions may occur at any time point, and items may be released at any time point along the global timeline. Due to the nature of collaborative filtering, if the train/test data split does not observe the global timeline and all training instances are fed to the recommender as a whole, then the model could learn from data instances that are not available at the time point of the test instance. Through carefully designed experiments, we show that models with data leakage do recommend future items which are not available to the system at the time point of a test instance. We also show that more future data leads to a more different

# Датасеты

Dataset	SASRec performance degradation						
	HitRate@10			NDCG@10			
	Before shuffle	After shuffle	Relative change	Before shuffle	After shuffle	Relative change	
<b>Beauty</b>	0.042	0.026	-39%	0.019	0.011	-43%	
<b>Diginetica</b>	0.333	0.286	-14%	0.161	0.149	-7%	
<b>OTTO</b>	0.205	0.143	-30%	0.120	0.086	-28%	
<b>RetailRocket</b>	0.326	0.315	-4%	0.195	0.190	-2%	
<b>MegaMarket</b>	0.192	0.101	-47%	0.111	0.062	-45%	
<b>Sports</b>	0.032	0.023	-28%	0.016	0.011	-32%	
<b>Yoochoose</b>	0.396	0.308	-22%	0.228	0.167	-27%	
<b>Games</b>	0.052	0.035	-33%	0.025	0.015	-38%	
<b>Steam</b>	0.110	0.099	-10%	0.053	0.047	-12%	
<b>ML-20m</b>	0.075	0.031	-59%	0.036	0.014	-61%	
<b>30Music</b>	0.198	0.020	-90%	0.136	0.010	-92%	
<b>Zvuk</b>	0.216	0.069	-68%	0.112	0.034	-70%	
<b>Foursquare</b>	0.353	0.328	-7%	0.224	0.213	-5%	
<b>Gowalla</b>	0.301	0.277	-8%	0.186	0.170	-8%	
<b>Yelp</b>	0.044	0.043	-2%	0.021	0.022	+5%	

- [1. Does It Look Sequential? An Analysis of Datasets for Evaluation of Sequential Recommendation](#)
- [2. An Analysis of Sequential Patterns in Datasets for Evaluation of Sequential Recommendations](#)

# Алгоритмы

- Не копируйте результаты бейзлайнов из других статей, всегда запускайте их на своём сетапе
- Используйте сильные бейзлайны (EASE, SLIM, MF, Item-to-Item)
- Нужно тюнить все алгоритмы, а не только свои

# Алгоритмы

## A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research

MAURIZIO FERRARI DACREMA, SIMONE BOGLIO, and PAOLO CREMONESI, Politecnico

di Milano, Italy

DIETMAR JANNACH, University of Klagenfurt, Austria

- *Lack of proper tuning of baselines:* This is probably the most striking observation of our analysis and is not specifically tied to deep learning approaches [57] or to recommendation problems [39, 40]. Researchers apparently invest significant efforts in optimizing their own new method but do not pay the same attention to their baselines. Sometimes, authors simply pick the hyperparameter settings reported to be optimal from a previous paper, even though those may refer to a different dataset or experimental procedure. Probably, this behavior might be the result of a *confirmation bias*, i.e., the tendency to search for results that affirm rather than refute prior research hypotheses.

# Алгоритмы

## **Everyone's a Winner! On Hyperparameter Tuning of Recommendation Models**

FAISAL SHEHZAD and DIETMAR JANNACH, University of Klagenfurt, Austria

learning rate that was used for all models and datasets. Furthermore, in this case, the embedding size was kept constant across all compared models “for fair comparison”. In reality, however, embeddings sizes are hyperparameters to tune, and fixing them to one specific value (without much justification) may actually lead to an unfair comparison. In the

# Алгоритмы

**Recommendation Performance.** We perform an extensive analysis of our proposed TIGER on the sequential recommendation task and compare against the baselines above. The results for all baselines, except P5, are taken from the publicly accessible results<sup>3</sup> made available by Zhou *et al.* [44]. For P5, we use the source code made available by the authors. However, for a fair comparison, we updated the data pre-processing method to be consistent with the other baselines and our method. We provide further details related to our changes in Appendix D.

# Алгоритмы

In this work, we report the results of a systematic analysis of algorithmic proposals for top-n recommendation tasks. Specifically, we considered 18 algorithms that were presented at top-level research conferences in the last years. Only 7 of them could be reproduced with reasonable effort. For these methods, it however turned out that 6 of them can often be outperformed with comparably simple heuristic methods, e.g., based on nearest-neighbor or graph-based techniques. The remaining one clearly outperformed the baselines but did not consistently outperform a well-tuned non-

# Алгоритмы

Numerical evaluations with comparisons to baselines play a central role when judging research in recommender systems. In this paper, we show that running baselines properly is difficult. We demonstrate this issue on two extensively studied datasets. First, we show that results for baselines that have been used in numerous publications over the past five years for the MovieLens 10M benchmark are suboptimal. With a careful setup of a vanilla matrix factorization baseline, we are not only able to improve upon the reported results for this baseline but even outperform the reported results of any newly proposed method. Secondly, we recap the tremendous effort that was required by the community to obtain high quality results for simple methods on the Netflix Prize. Our results indicate that empirical findings in research papers are questionable unless they were obtained on standardized benchmarks where baselines have been tuned extensively by the research community.

# Алгоритмы

Five of the six third-party versions implement the same *architecture* (RQ1) as the original. However, “GRU4Rec” of the Microsoft Recommenders collection only has two things in common with original: its name and that it uses GRU layers. The list of differences starts with the embedding layer, where this reimplementation utilizes extra information (item categories) besides the item ID. While changing the representation of the events within the session does not result in a new algorithm, it already undermines the goal of reproducing the original work. The list continues with the training of the GRU layer, where this algorithm requires sequences of equal length and processes multiple time steps in one forward pass, ignoring one of the distinctive features of the original. Scoring is also different and deeply flawed as well. In the reimplementation, a feedforward

# Метрики

- **Beyond accuracy:**
  - Делайте замеры coverage, training / inference time
  - Используйте разумные метрики для ваших stakeholders & objectives
- **Не сэмплируйте негативы для замеров, используйте весь каталог или большой индекс (похожий на реальный продовыи)**
  - Если пришлось сэмплировать – фиксируйте сиды и предоставляйте результаты сэмплирования
- **Объясняйте свой выбор метрик**
  - Почему в этом RQ нужна эта метрика, что конкретно она проверяет?

## A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research

MAURIZIO FERRARI DACREMA, SIMONE BOGLIO, and PAOLO CREMONESI, Politecnico di Milano, Italy

DIETMAR JANNACH, University of Klagenfurt, Austria

### 3.3 Early Stopping Approach

Many machine learning models are trained for a *number of epochs* in which the model's performance varies, first increasing and then stabilizing, while usually exhibiting some degree of variance. The number of epochs therefore represents another important parameter to be determined. However, it is worth noting that in the articles we have analyzed neither the number of epochs nor the stopping criteria are usually mentioned. The procedure in which this parameter was chosen in the original articles is therefore not clear. Looking at the code shared by the authors we could observe that, in some cases, the number of epochs was inappropriately selected via an evaluation done on the test data, therefore causing information leakage from the test data. In other cases, the reported metric values were inappropriately taken from different epochs.

# Как завоевать доверие?

- **Не делайте ошибок, будьте внимательны**
  - Одна ошибка – и ты ошибся :)
  - В подсчете метрик, в обучении и применении алгоритмов, в математике
- **Обеспечивайте воспроизводимость**
  - Предоставляйте код и все остальные артефакты, нужные для получения ваших результатов
- **Аргументируйте свои выборы**
  - Датасетов, алгоритмов, метрик
- **Формируйте фундаментальные вопросы, а не тезисы вида "Мы сделали новый SOTA"**
- **Делайте тщательный обзор литературы (do your research)**

# Как обеспечить реалистичность?

- **Обеспечивайте стат. значимость**
  - Делайте повторные запуски экспериментов и усредняйте результаты
- **Используйте похожие на прод экспериментальные сетапы**
  - Большие датасеты для обучений и замеров, большие индексы
  - Почему замеряем Recall@1000, а не Recall@10? А nDCG?
- **Не игнорируйте тяжелые хвост и проблему холодного старта**
  - Оставляйте коротких и новых пользователей
  - Оставляйте тяжелый хвост и новые айтемы
- **Замеряйте distribution drift**
  - Тестируйте дообучение вашего подхода на новых данных, эмулируйте стриминговый сценарий

# An Opinionated Guide to ML Research

- Honing Your Taste
  - Read a lot of papers
  - Work in a research group
  - Seek advice from experienced researchers
  - Spend time reflecting on what research is useful
- Idea-Driven vs Goal-Driven Research
  - **Idea-Driven:** As you read a paper, you have an idea how to do it better
  - **Goal-Driven:** Develop a vision of some new capabilities you'd like to achieve
  - You need your own perspective
- Aim High, and Climb Incrementally Towards High Goals