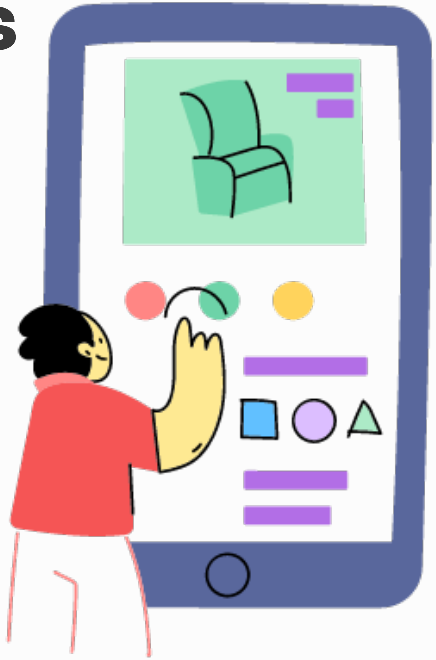
Enhancing Reproducibility in Recommender Systems

A Path Towards Scientific Integrity and Effective Implementation

ACM Europe School on Recommender Systems 2024

Saturday, October 12, 2024



Reproducibility-aware conferences

Reproducibility and Replicability Tracks

- ACM RecSys
- ACM SIGIR
- ACM UMAP
- ECIR
- ACM MultiMedia

Challenges and Competitions

- ACM RecSys (RecSys Challenge)
- ACM SigIR (eComm WS)
- WSDM
- KDD
- CIKM

Benchmarking

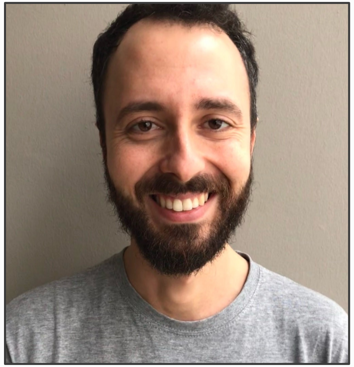
- NeurIPS
- ICIP
- CVPR

Conferences

- ACM REP

*non exaustive lists

Who's in front of you



Claudio

Antonio

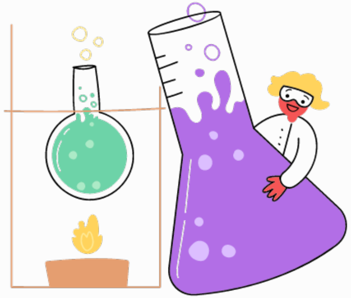
Assistant Professors at Politecnico di Bari (we received our PhDs here)

Born and raised right next to Bari

Working on recommender systems reproducibility, fairness, explanation, efficiency, and privacy

Let's talk about science

Am I a scientist?



What is science?

A mode of inquiry aiming to pose questions about the world, arriving at the answers and assessing their degree of certainty

- Describe the world
- Explain the world
- Predict what will happen
- Intervene in specific processes or systems

What is science?

How is the work of a scientist?

The scientists in the world follow a common approach

1. Introduce ideas and theories
2. Collect data
3. Analyze data and experiment
4. Communicate the results (e.g., through a scientific article)

The modern scientific method

The scientific method is an empirical method for acquiring knowledge about the world

1 **Observe something**

- Observe evidence systematically
- Document observations in an objective way

The modern scientific method

The scientific method is an empirical method for acquiring knowledge about the world

2 Develop a hypothesis

- Formulate a clear problem statement
- Identify the main question or investigation goal
- Pose a testable and measurable question

The modern scientific method

The scientific method is an empirical method for acquiring knowledge about the world

3 Collect data

- Gather relevant data systematically
- Use appropriate methods for data collection
- Organize data for analysis

The modern scientific method

The scientific method is an empirical method for acquiring knowledge about the world

4 Test with experiments

- Be aware of performing well-controlled experiments
- Control some parameters while manipulating others
- Collect result from the experiments

The modern scientific method

The scientific method is an empirical method for acquiring knowledge about the world

5 **Analyze results**

- Understand the meaning behind the results
- Establish cause-and-effect relationships
- Provide evidence to support or reject the hypothesis

The modern scientific method

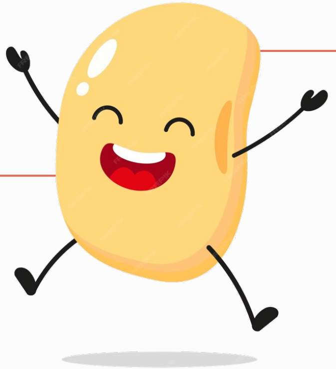
The scientific method is an empirical method for acquiring knowledge about the world

6 Report conclusions

- Share experiment outcomes through conferences or journal articles
- Contribute to the body of knowledge for future research
- Don’t forget to detail the project’s design, methods, and results

Let's develop a new idea

Story of a bean: ideas from deduction



Known rule

The beans in my bag are white
(i.e., if a bean is in my bag, then it is white)

Evidence

I have in my hand a bean from my bag

New theory The bean in my hand is white

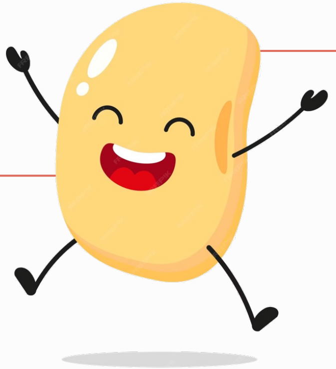
KB: $\forall x\ bag(x) \Rightarrow white(x)$
$bag(B)$

α: $white(B)$

Story of a bean: ideas from induction

Facts

I have in my hand
some white beans



Knowledge

All these beans come
from my bag

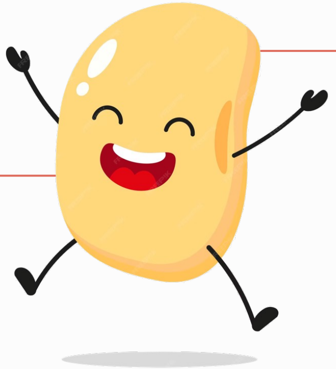
General rule

The beans in my bag are white
(i.e., if a bean is in my bag, then it is white)

Story of a bean: ideas from abduction

Known rule

The beans in my bag are white
(i.e., if a bean is in my bag, then it is white)



Evidence

I have in my hand a white bean (let's call it Fagiolino)

Supposed explanation Fagiolino comes from my bag

Story of a bean: ideas from abduction

Let’s rewrite the last story in a more formal way

KB: $\forall x \; bag(x) \Rightarrow white(x)$

α: white(Fagiolino)

With deduction we cannot conclude that the bean comes from my bag
But abduction can help us explain why the bean in my hand is white!

Let’s define some potential hypotheses

Fagiolino fell from the sky	Fagiolino comes from my pocket	Fagiolino comes from my bag

Story of a bean: ideas from abduction

Now, add each hypothesis to the knowledge base and check whether it can be valid

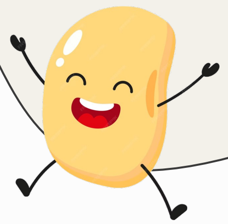
(h_1) Fagiolino fell from the sky: sky(Fagiolino)

$KB \cup h_1$: $[\forall x \; bag(x) \Rightarrow white(x)] \cup sky(Fagiolino)$

α: white(Fagiolino)

With this hypothesis, we cannot deduce α from $KB \cup h_1$

Story of a bean: ideas from abductic



Now, add each hypothesis to the knowledge base and check whether it can be valid

(h_2) Fagiolino comes from my pocket: pocket(Fagiolino)

KB $\cup$ h_2: [$\forall x$ bag(x) $\Rightarrow$ white(x)] $\cup$ pocket(Fagiolino)

α: white(Fagiolino)

With this hypothesis, we cannot deduce α from KB $\cup$ h_2

Story of a bean: ideas from abduction

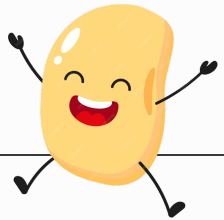
Now, add each hypothesis to the knowledge base and check whether it can be valid

(h_3) Fagiolino comes from my bag: bag(Fagiolino)

KB ∪ h_3: [∀x bag(x) ⇒ white(x)] ∪ bag(Fagiolino)

α: white(Fagiolino)

With this hypothesis, we can deduce the bean is ______ white!
We have found a potentially valid hypothesis



Story of a bean: ideas from abduction

Now, add each hypothesis to the knowledge base and check whether it can be valid

(h_3) Fagiolino comes from my bag: bag(Fagiolino)

KB ∪ h_3: [∀x bag(x) ⇒ white(x)] ∪ bag(Fagiolino)

α: white(Fagiolino)

And… what if I have more than one potentially valid hypothesis?

Select the most simple and elegant

(see the Occam's razor)

Story of a bean: ideas from abduction

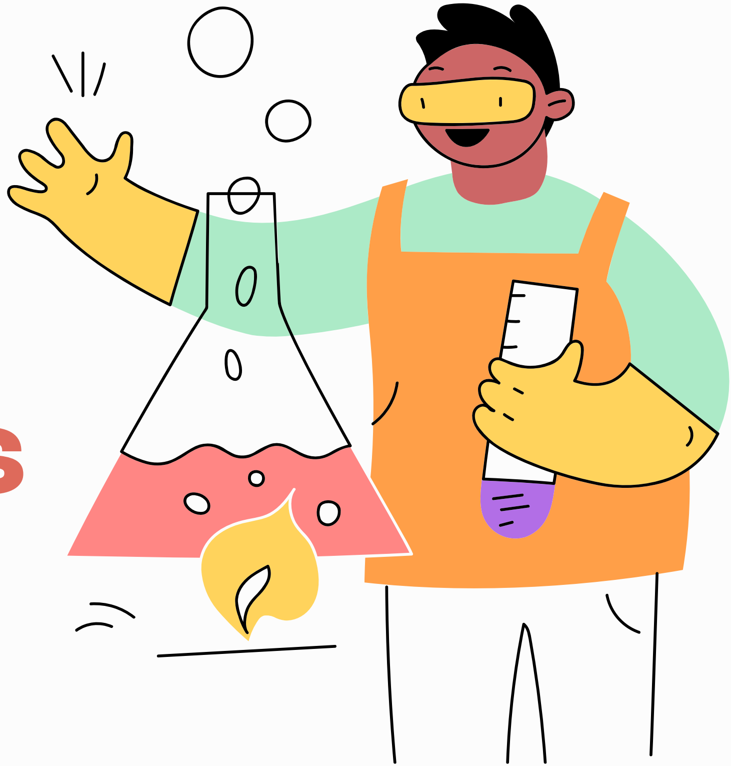
OK, but is this enough?

No.

Abuctive reasoning helps us in generating
new hypotheses that must be validated

We won’t be sure about them until we are not able
to somehow prove their validity

Let’s do some experiments



Experimenting with a new RS

Observation

My recommender considering only the last interaction of a user isn’t working well 😢

Hypothesis

The low performance is due to the limited user representation and a new model considering a longer user history would perform better

Experiments

The new model considering a longer history shows its effectiveness over the previous model

Experimenting with a new RS

Do the experiments prove the new model is «the best»?

Do the experiments prove the new model works in any scenarios?

The scientific method never proves something with absolute certainty

Instead, the scientific method provides a structured process for testing, evaluating, and validating hypothesis and solutions through evidence

Making hypotheses and experiments reliable

How to make hypotheses (models, ideas, ...) more and more reliable?

Allow others to verify our findings.

Other people should be able to:

- check the **validity and generalizability** of our results,
- or **contradict** our evidence (according to Popper, the progress does not consist in the accumulation of certainties, but in the progressive elimination of errors)

Making hypotheses and experiments reliable

But, why people want to check my findings?

Sometimes there could be **mistakes and they just want to check**

Sometimes they want to explore the limits of the findings and relationships you discovered to make other inquiry

Sometimes… a **young researcher** may be pressed to publish papers to improve their CVs
This pressure may lead to **overstate the importance of the results** and **increase the risk of bias** in data collection, analysis, and reporting

Making hypotheses and experiments reliable

Remember that nature is not capricious
and follows rules that are consistent overtime and across different contexts

So… redoing an experiment, people should observe no difference between the original and the reproduction

Yes, this is what we call

reproducibility

Four questions about reliability of hypotheses and experiments

1 Are the data and the analyses laid with **sufficient transparency and clarity** that the results can be checked?

Reproducible research is research that is capable of being checked because the data, code, and methods of analysis are available to other researchers

Four questions about reliability of hypotheses and experiments

2 If checked, do the data and analysis offered in support of the result in fact **support that result**?

Research is reproducible if another researcher uses the available data and code and obtains the same results

Four questions about reliability of hypotheses and experiments

3 If the data and analysis are shown to support the original result, can the result reported be found again in the specific study context?

To answer this question, a researcher must redo the study, following the original methods as closely as possible and collecting new data, aimed at the same or a similar scientific question as the original research

This is no more reproducibility, but what we call

replicability

Four questions about reliability of hypotheses and experiments

4 Can the result reported or the inference drawn be found again in a broader set of study contexts?

A researcher could take a variety of paths: choose a new condition of analysis, conduct the same study in a new context, or conduct a new study aimed at the same or similar research question

And this is the notion of

generalizability

Let’s tidy things up

Reproducibility

Obtaining **consistent results** using the **same** input data, computational steps, methods, code, and conditions of analysis;
a.k.a. transparency and “**computational reproducibility**”

Replicability

Obtaining **consistent results** across studies aimed at answering the same scientific question, each of which has obtained its own data

Generalizability

Exploring a similar scientific question but in **other contexts or populations** that differ from the original one and finding consistent results

Note that historically, various disciplines have used different nomenclatures, sometimes even reversing the terms "reproducibility" and "replicability" (ACM itself previously used this reverse terminology)

ACM EIGREP

The mission of EIGREP is to foster a broad and inclusive intellectual community around the issues of reproducibility of computational research

Reproducibility is a **cornerstone of the scientific method** and central to research integrity

ACM REP

Let’s rep recommenders

Reproducibility vs. Replicability

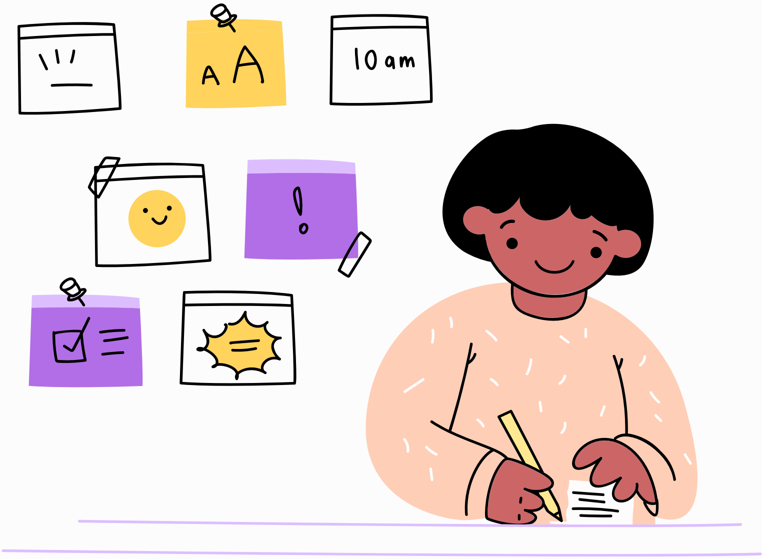
Reproducibility

- Avoids changes
- Allows others to inspect and validate the experiment
- Expected from any well-controlled experiment, it is crucial for trasparent and accountable research

Replicability

- Requires changes
- Validates the experiment's core ideas, ensuring results aren't due to ad-hoc design choices
- Essential for corroborating findings and advancing inquiry

To-do list for a reproducible work



Provide a detailed description of:

- Dataset collection
- Data splitting
- Implementation details of the recommendation algorithms
- Parameters
- Candidate Item Filtering
- Evaluation
- Statistical Testing

To-do list for a replicable work

TL;DR: Try to change your environment:

- Dataset
- Parameters
- …

Are your findings still confirmed?

To-do list for a replicable work

Example

- We have created a **new graph recommender system**
- Our **hypothesis** is that the new recommender system works better than the state-of-the-art graph recommenders
- Let's create an **experimental environment** and test our algorithm:
    1. Choose a dataset
    2. Preprocess it to remove cold users and items according to a threshold
    3. Select a candidate items protocol
    4. And…

To-do list for a replicable work

Example (cont’d)

	nDCG@10
Our model	0.21

WOW! This is a very good performance 🤩

Not at all. 😏

What about the other recommender systems?

Ok. Let’s read other papers and pick their results

To-do list for a replicable work

Example (cont'd)

	nDCG@10
Our model	0.21
Other graph model	0.10
Item kNN	0.06
Matrix Factorization	0.05

WOW! We are still the best! 🥳

No, these results won't be replicable 😞

Remember that the experimental environment (thus, the evaluation protocol) dramatically impacts the observed results

To-do list for a replicable work

Digression on the impact of the experimental setup

Let's have a look at two works experimenting with MovieLens 100K

	Algorithm				
Metric	*k*-Item	*k*-User	PureSVD	*Pop-item*	IMM
P@5	0.00135	0.006	0.067	0.227	0.267
NDCG@5	0.0036	0.0091	0.0566	0.216	0.245
MAP	0.013	0.041	0.061	0.119	0.156

Gorla et al, 2013

	Baseline(Test)
MAP	0.447
MRR	0.889
NDCG@10	0.720
NDCG@5	0.570
NDCG@3	0.447

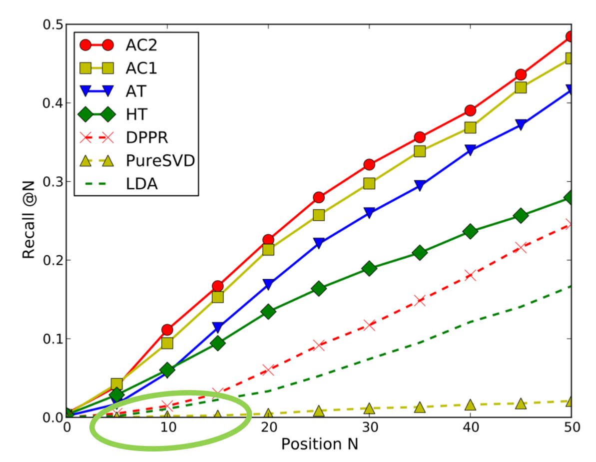
Jambor & Wang, 2010

MAP and nDCG seem ten times different!

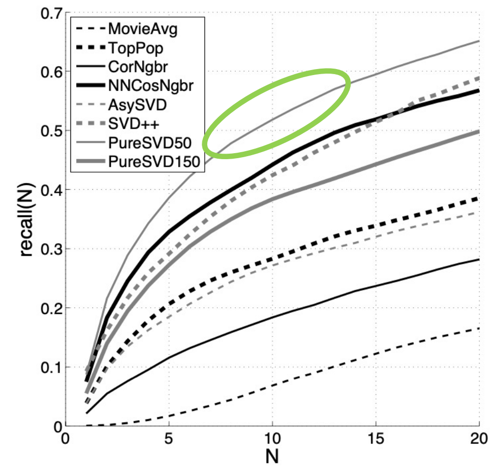
To-do list for a replicable work

Digression on the impact of the experimental setup

Both these works experiment with MovieLens 1M but report recall values that differ by one order of magnitude



Yin et al, 2012



Cremonesi et al, 2010

To-do list for a replicable work

Digression on the impact of the experimental setup

Remember that a lot of factors influence the results of an evaluation pipeline

- Splitting methods
- The selection of the items candidate to ranking (test ratings, test items, training items, all items, …)
- The use of different implementations of the same metric (e.g., normalizations, compensations, treatment of equal scores, …)
- Ability to predict for all items or users
- …

To-do list for a replicable work

Example (cont'd)

Ok, you got me! We have to **replicate** the other baselines in our environment

	nDCG@10
Our model	0.21
Other graph model	0.20
Item kNN	0.17
Matrix Factorization	0.16

The final (replicable?) finding: our graph recommender system improves the state of the art of graph recommender systems

To-do list for a replicable work

Example (cont'd)

	nDCG@10
Our model	0.21
Other graph model	0.10
Item kNN	0.06
Matrix Factorization	0.05

Still not sure about the replicability

Who is the «other graph model»?
Is it recent enough? Is it competitive enough?

Often, improved scores surpass **outdated baselines** and don't trend upwards over time, as baselines are **rarely recent or competitive** and **fail to reflect new discoveries**

To-do list for a replicable work

Example (cont’d)

	nDCG@10
Our model	0.21
Other graph model	0.10
Item kNN	0.06
Matrix Factorization	0.05

Still not sure about the replicability

Who are the other two baselines?
How our scientific findings relate to the two non-graph baselines?
Are they useful to confirm our hypothesis?

To-do list for a replicable work

Digression on Top-N Recommendation Algorithms

Algorithm	Top@10					
	nDCG	MAP	MRR	Pre	Rec	F1
EASE^R	**0.336**	0.335	**0.583**	0.274	**0.194**	0.190
SLIM	0.335	**0.337**	0.580	**0.275**	0.189	0.188
MF2020	0.329	0.327	0.563	0.272	0.190	**0.192**
UserKNN	0.315	0.314	0.554	0.256	0.183	0.179
RP$^3\beta$	0.315	0.313	0.556	0.256	0.184	0.179
iALS	0.306	0.304	0.542	0.252	0.179	0.176
MultiVAE	0.294	0.284	0.514	0.243	0.183	0.175
ItemKNN	0.292	0.293	0.518	0.242	0.163	0.163
NeuMF	0.277	0.275	0.494	0.232	0.157	0.158
BPRMF	0.275	0.271	0.502	0.226	0.166	0.161
MostPop	0.159	0.159	0.317	0.137	0.084	0.086
Random	0.008	0.007	0.020	0.007	0.004	0.004

Accuracy Results for MovieLens-1M. The tables are sorted by nDCG in descending order.

The paper *Top-N Recommendation Algorithms: A Quest for the State-of-the-Art* shows consistent performance by linear models, nearest-neighbor methods, and traditional matrix factorization on modest-sized, commonly-used datasets

Each algorithm is “competitive” in a different way w.r.t. the objective as measured by different metrics

To-do list for a replicable work

Example (cont'd)

	nDCG@10
Our model	0.21
Other graph model	0.10
Item kNN	0.06
Matrix Factorization	0.05

Still not sure about the replicability

Is the metric properly chosen for the task?
E.g., if our aim is to recommend just one item, in what helps nDCG?

To-do list for a replicable work

Example (cont'd)

	nDCG@10	
Our model	0.21	
Other graph model	0.10	
Item kNN	0.06	
Matrix Factorization	0.05	

Still not sure about the replicability

Are we including all the metrics needed to analyze and justify our findings?

To-do list for a replicable work

Example (cont'd)

	nDCG@10
Our model	0.21
Other graph model	0.10
Item kNN	0.06
Matrix Factorization	0.05

Still not sure about the replicability

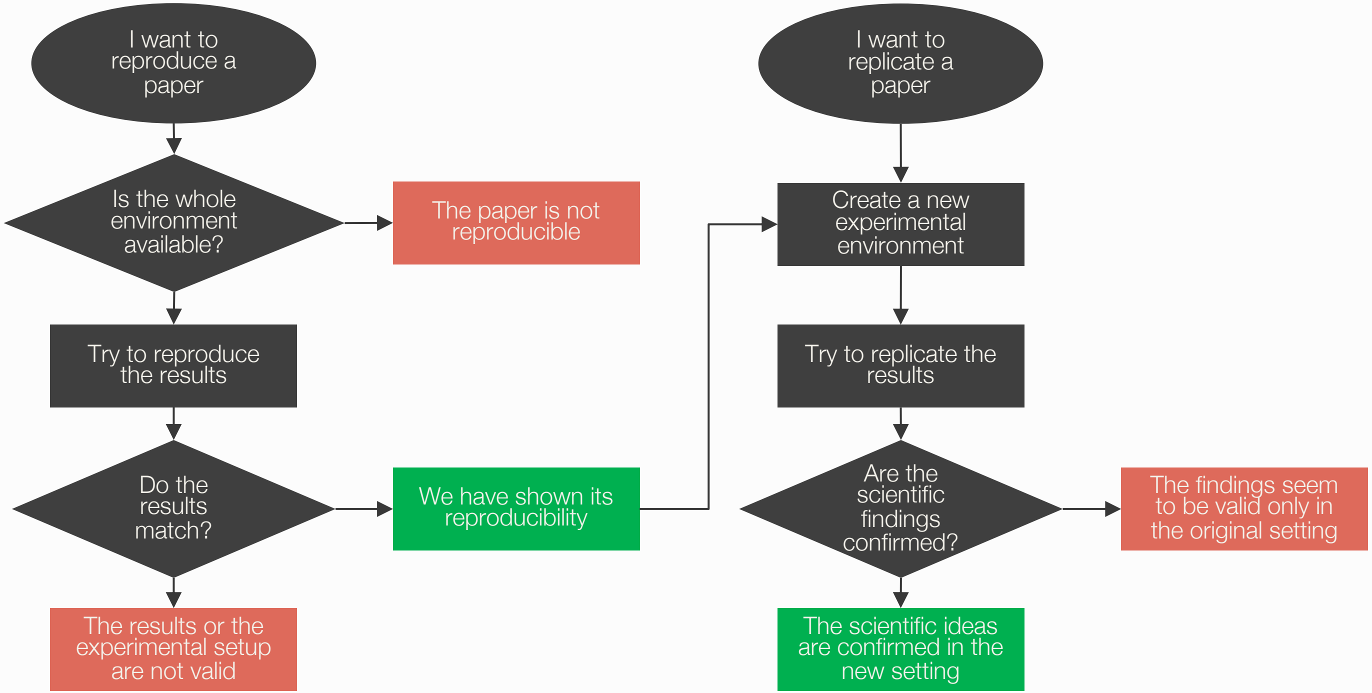
What about other datasets?
Is this (are these) dataset(s) enough to prove our findings?

To-do list for a replicable work

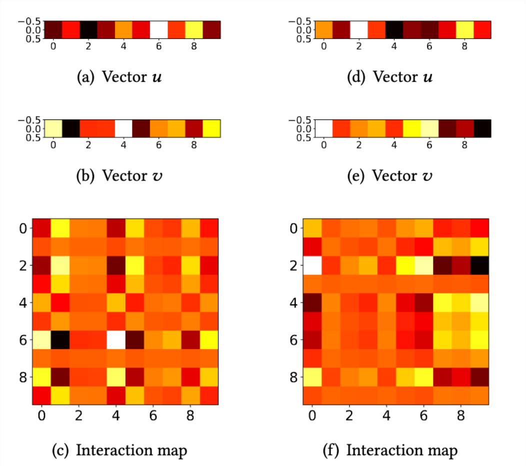
Take-home message

- Many possible mistakes can hinder the replicability of our work
- Carefully check your experimental environment to ensure your hypotheses are as strongly validated as possible within your context
- Test your findings with other experimental setups
- Make your paper (at least) reproducible to promote transparency, facilitate verification, and simplify future replicability efforts

How to reproduce or replicate a work



Convolutions over User-Item Embedding Maps?



Effects of permutating the columns of vectors u and v on their resulting outer product (the interaction map)

The paper *Critically Examining the Claimed Value of Convolutions over User-Item Embedding Maps for Recommender Systems* poses questions about CNN advantages

CNNs leverage the position of each “pixel” to discover “semantic” patterns
Does it make sense in user-item matrices?

CNN-based models cannot offer the claimed advantages (think about permutations of rows)

Convolutions over User-Item Embedding Maps?

They used the original code, data, data splits, as well as hyperparameters that were provided by the authors

	@5		@10		@20	
	HR	NDCG	HR	NDCG	HR	NDCG
TopPopular	0.0817	0.0538	0.1200	0.0661	0.1751	0.0799
UserKNN CF	**0.2068**	**0.1355**	**0.3126**	**0.1695**	0.4401	**0.2017**
ItemKNN CF	**0.2521**	**0.1686**	**0.3669**	**0.2056**	**0.4974**	**0.2385**
$P^3\alpha$	**0.2146**	**0.1395**	**0.3211**	**0.1737**	0.4442	**0.2049**
$RP^3\beta$	**0.2202**	**0.1431**	**0.3323**	**0.1793**	**0.4667**	**0.2132**
SLIM	**0.2330**	**0.1535**	**0.3475**	**0.1904**	**0.4799**	**0.2238**
PureSVD	**0.2011**	**0.1307**	0.3002	**0.1626**	0.4238	0.1938
iALS	**0.2048**	**0.1348**	**0.3080**	**0.1680**	0.4319	**0.1993**
ConvNCF	0.1947	0.1250	0.3059	0.1608	0.4446	0.1957

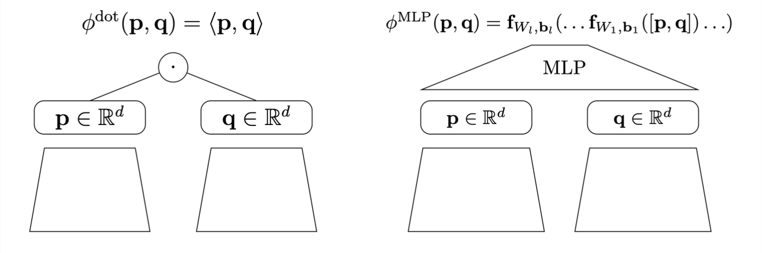
	@ 1		@ 5		@ 10	
	HR	NDCG	HR	NDCG	HR	NDCG
TopPopular	0.1593	0.1593	0.4217	0.2936	0.5813	0.3451
UserKNN CF	0.3540	0.3540	0.6884	0.5324	0.8060	0.5704
ItemKNN CF	0.3305	0.3305	0.6682	0.5080	0.7940	0.5488
$P^3\alpha$	0.3316	0.3316	0.6543	0.5031	0.7687	0.5402
$RP^3\beta$	0.3464	0.3464	0.6743	0.5198	0.7959	0.5591
SLIM	**0.3906**	**0.3906**	**0.7116**	**0.5625**	**0.8315**	**0.6014**
PureSVD	**0.3735**	**0.3735**	**0.7088**	**0.5522**	0.8132	**0.5861**
iALS	**0.3816**	**0.3816**	**0.7121**	**0.5581**	0.8200	**0.5933**
CoupledCF	0.3522	0.3522	0.7018	0.5374	0.8247	0.5775

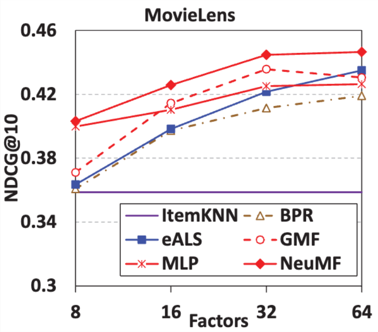
	@5		@10		@20	
	HR	NDCG	HR	NDCG	HR	NDCG
TopPopular	0.0016	0.0009	0.0023	0.0011	0.0033	0.0014
UserKNN CF	**0.5964**	**0.4527**	**0.6715**	**0.4773**	**0.7032**	**0.4855**
ItemKNN CF	**0.5975**	**0.4425**	**0.6776**	**0.4689**	**0.7070**	**0.4764**
$P^3\alpha$	**0.6327**	**0.4929**	**0.6744**	**0.5066**	**0.7014**	**0.5135**
$RP^3\beta$	**0.5896**	**0.4458**	**0.6756**	**0.4739**	**0.7071**	**0.4821**
SLIM	**0.6674**	**0.5169**	**0.6972**	**0.5267**	**0.7102**	**0.5300**
PureSVD	**0.4026**	**0.3117**	**0.4891**	**0.3397**	**0.5652**	**0.3590**
iALS	**0.6110**	**0.4811**	**0.6735**	**0.5017**	**0.7033**	**0.5093**
CFM	0.2241	0.1485	0.3338	0.1839	0.4661	0.2173

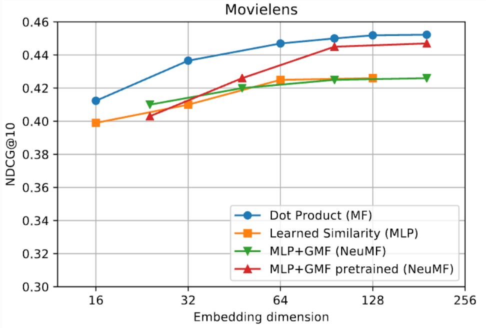
Experimental results for ConvNCF, CoupledCF, and CFM for Yelp, MovieLens1M, and Last.fm respectively

Neural Collaborative Filtering vs. Matrix Factorization Revisited

Rendle et al. show that a well-tuned simple dot product outperforms MLPs (NeuMF) in both effectiveness and efficiency for estimating the similarity between a user and an item



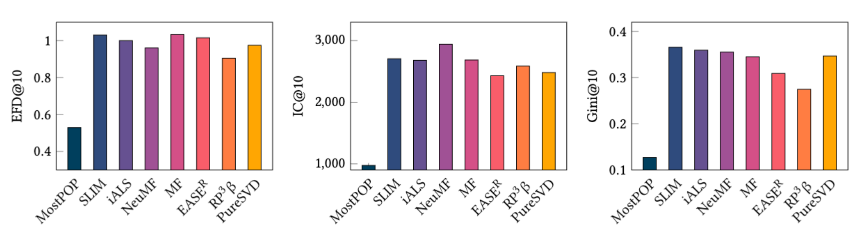




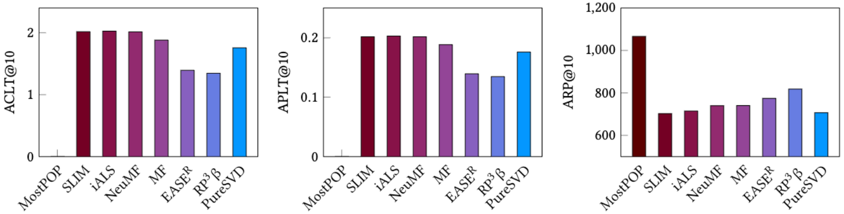
Performance of NDCG@10 w.r.t. the number of predictive factors on MovieLens1M. Comparison of the results of the two papers.

Reenvisioning Collaborative Filtering vs Matrix Factorization

Anelli et al. **reproduce and replicate** experiments from *Neural Collaborative Filtering vs Matrix Factorization* and **extend the original findings** confirming that MF provides better accuracy, especially on long-tail items, but NeuMF offers better coverage and diversification



Diversity comparison of NeuMF and MF with various baselines (higher is better)



Analysis of Bias for NeuMF, MF and various baselines considering a cutoff @10

A Troubling Analysis of Recommender Systems Research

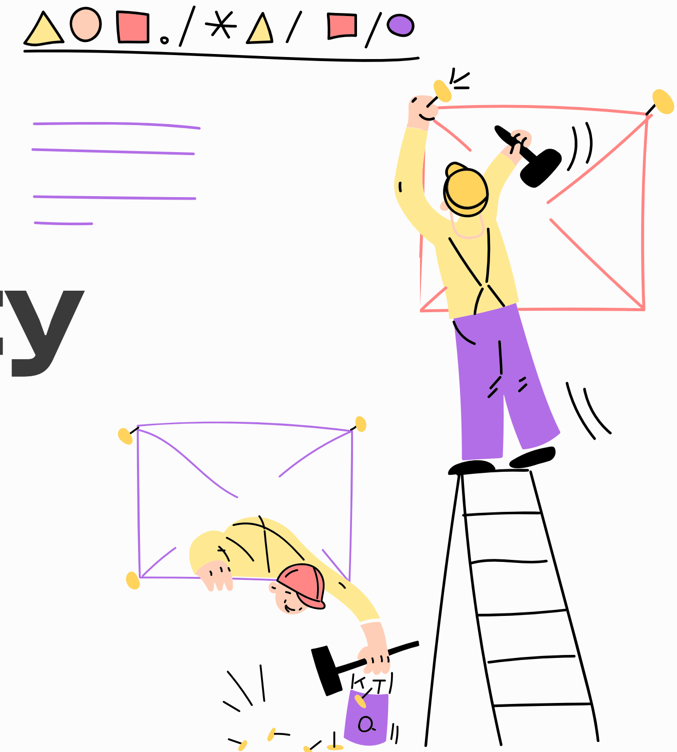
In *A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research*, Ferrari Da Crema et al. survey papers published between 2015 and 2018 in top-conferences They identify 26 relevant papers and, among these, only 12 were considered having a reproducible experimental setup (evidencing a reproducibility crisis)

In a lot of cases, they also evidence a lack of replicability

Authors report papers showing only favorable results, thus inflating the risk of presenting only "virtual" progress

They confirm a propagation of weak baselines: relying on methods like NeuMF as state-of-the-art can mislead research, as they may not outperform simpler techniques.

Let's make reproducibility easier



Towards a easier reproducibility

We have seen how replicability is strictly related with good hypotheses and evaluation methodologies properly chosen to confirm the hypotheses

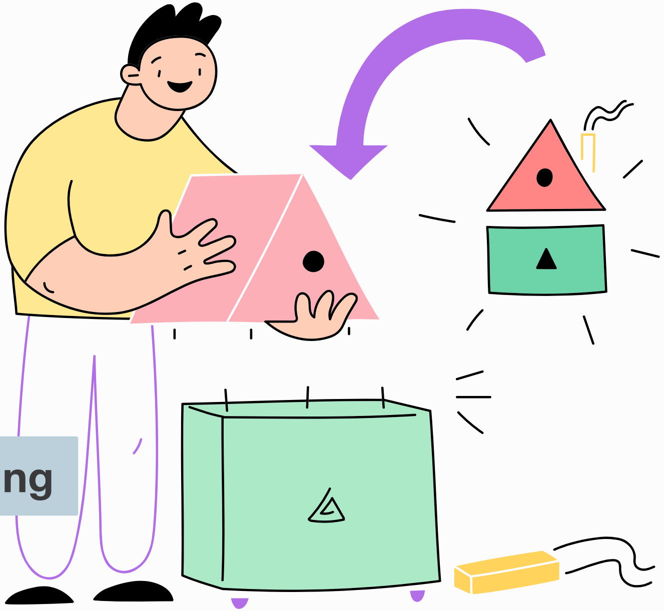
Reproducibility, instead, is related to rigourously provide code, data, and artifacts that lead to the same experimental results

But how hard can be guarantee (at least) reproducibility without any errors? Remember that in our works we should «reimplement» the baselines, so that this «reimplementation» and the experimental settings are in turn reproducible

Towards a easier reproducibility

Reproducibility would be for sure easier with…

1. Common practices for artifact sharing
2. Shared baseline implementations
3. Shared metrics implementations
4. Common practices for data preprocessing



Yes, this is what we call

reproducibility framework

Reproducibility frameworks

- **~2010**: Lenskit (Java), MyMediaLite, RankSys
- **2019**: Recommenders
- **2020**: Cornac, DaisyRec, Lenskit (Python), ReChorus
- **2021**: Elliot, FuxiCTR, RecBole
- **2022**: ClayRS, RecList, RecPack
- **2024**: Mab2Rec, ReChorus 2.0

The RecSys CfP suggests using one of the frameworks above for the submitted papers and sharing the used *experimental environment*

Reproducibility frameworks

Data-pipeline	Item selection	Models
Metrics	Tuning	Statistical tests
Configuration	APIs and UIs	Results (CSV/LaTeX)

References

- Anelli, Vito Walter, et al. "Top-n recommendation algorithms: A quest for the state-of-the-art." Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. 2022.
- Anelli, Vito Walter, et al. "Reenvisioning the comparison between neural collaborative filtering and matrix factorization." Proceedings of the 15th ACM Conference on Recommender Systems. 2021.
- Rendle, Steffen, et al. "Neural collaborative filtering vs. matrix factorization revisited." Proceedings of the 14th ACM Conference on Recommender Systems. 2020.
- Ferrari Dacrema, Maurizio, et al. "A troubling analysis of reproducibility and progress in recommender systems research." ACM Transactions on Information Systems (TOIS) 39.2 (2021): 1-49.
- Ferrari Dacrema, Maurizio, et al. "Critically examining the claimed value of convolutions over user-item embedding maps for recommender systems." Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.
- Gorla, Jagadeesh, et al. "Probabilistic group recommendation via information matching." Proceedings of the 22nd international conference on World Wide Web. 2013
- Jambor, Tamas, and Jun Wang. "Goal-driven collaborative filtering–a directional error based approach." European Conference on Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
- Cremonesi, Paolo, Yehuda Koren, and Roberto Turrin. "Performance of recommender algorithms on top-n recommendation tasks." Proceedings of the fourth ACM conference on Recommender systems. 2010.
- Yin, Hongzhi, et al. "Challenging the long tail recommendation." arXiv preprint arXiv:1205.6700 (2012).
- Armstrong, Timothy G., et al. "Improvements that don't add up: ad-hoc retrieval results since 1998." Proceedings of the 18th ACM conference on Information and knowledge management. 2009
- Cockburn, Andy, et al. "Threats of a replication crisis in empirical computer science." Communications of the ACM 63.8 (2020): 70-79.
- Popper, K. All Life Is Problem Solving. Routledge, 1999.
- Burch, Robert. "Charles sanders peirce." (2001).

Thanks!

Do you have any questions?

 claudio.pomo@poliba.it

 antonio.ferrara@poliba.it

Let's dive right into the hands-on session!