UNIVERSITÄT KLAGENFURT

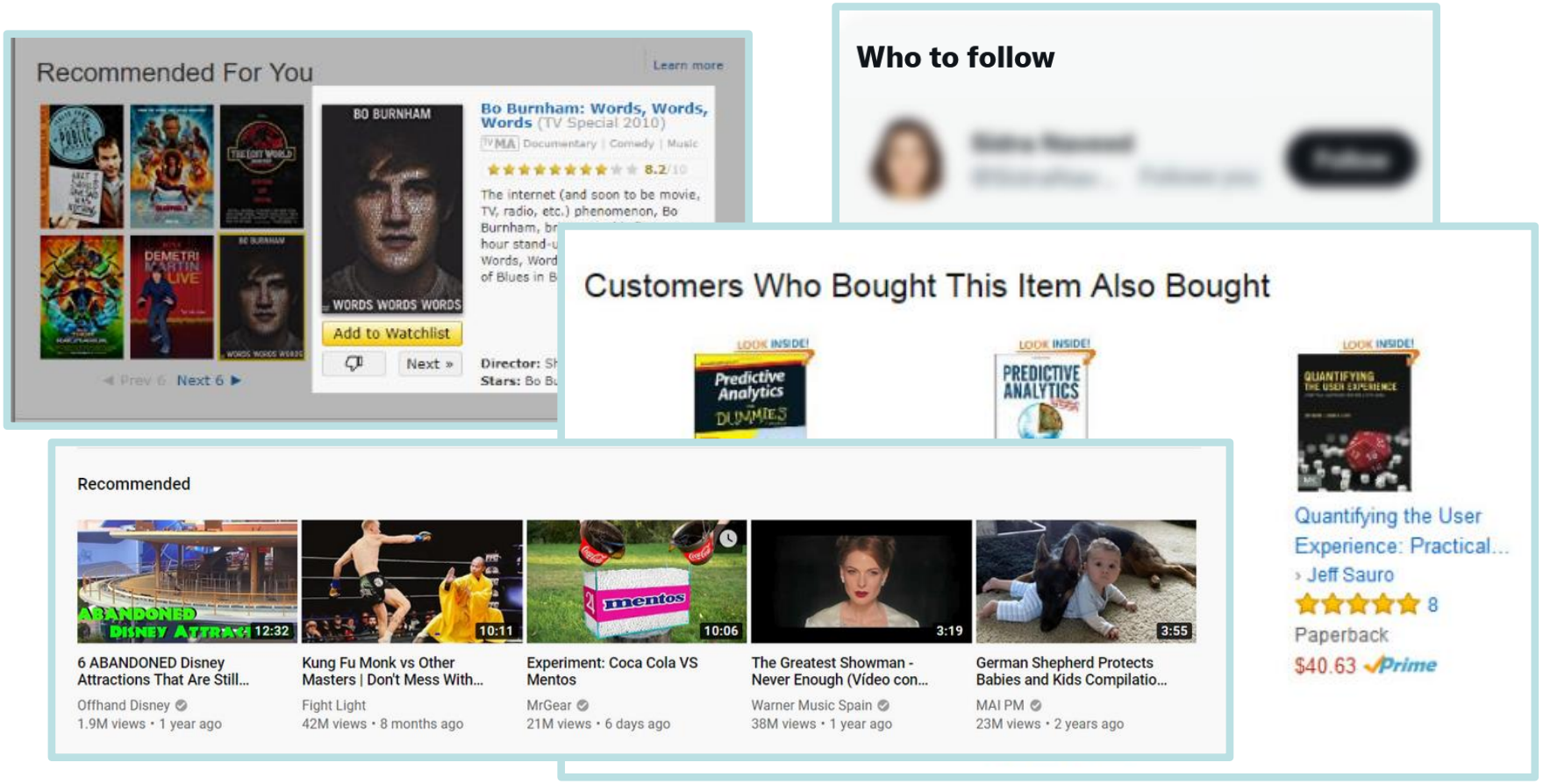
Recommender Systems, McNamara, and the Illusion of Progress

Dietmar Jannach, University of Klagenfurt, Austria
dietmar.jannach@aau.at

Presented NeuRec Workhsop @ ICML 2021
December 2021

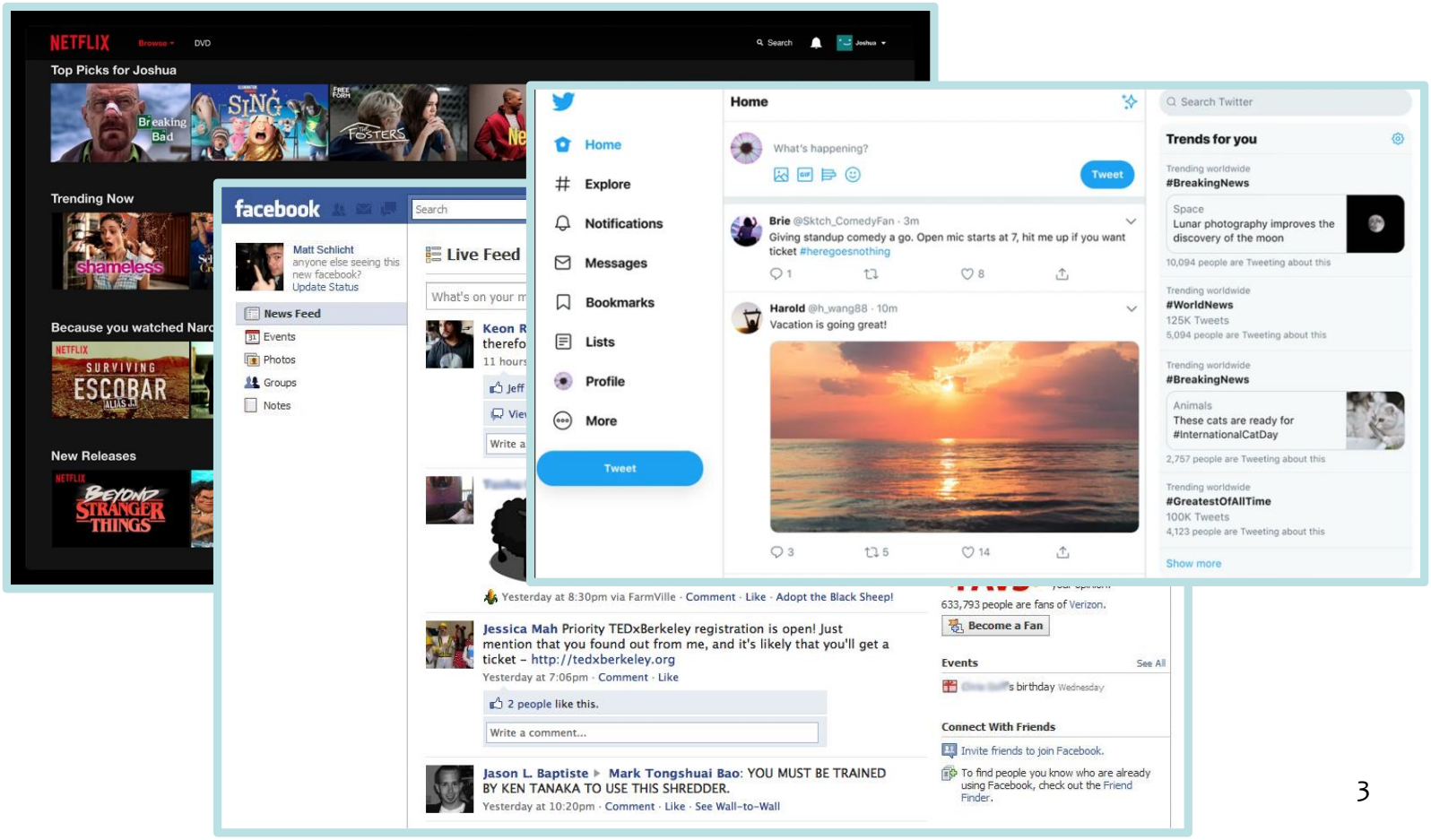
Recommender systems (RS)

- Automated recommendations everywhere



Recommender systems (RS)

- "Everything is a recommendation"



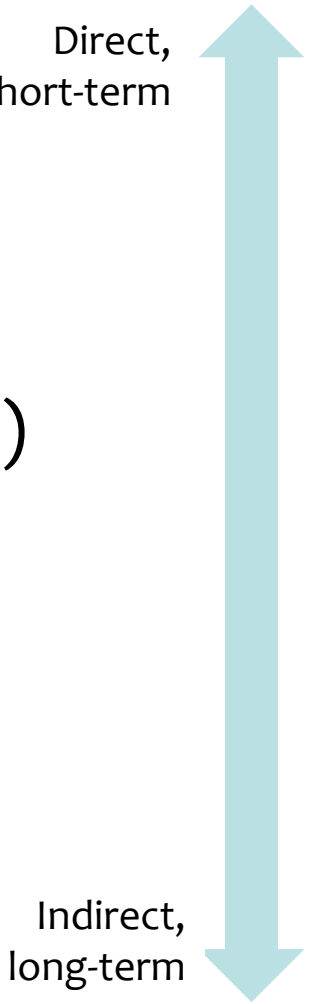
There's substantial business value[1]

- **Amazon in 2006**: About 35 % of revenue attributed to cross-selling, e.g., through "customers who bought ...."
- **YouTube in 2010:** About 60 % of the clicks on the home screen are on the recommendations
- **Netflix in 2012:** 75 % of what people watch is from some sort of recommendation

[1]Jannach, D. and Jugovac, M.: *"Measuring the Business Value of Recommender Systems"*. ACM Transactions on Management Information Systems, Vol. 10(4). 2019

Business value

- Direct profitability
  - Sales, revenue, profit
  - Effects on sales distribution
- Click-through-rate (CTR)
- Adoption of recommendation (“Long-CTR”)
- Engagement with service
  - Time spent on the platform, discovery effects
  - Number of interactions per session or user
- Retention rates
- …



Business value

- Key Performance Indicators
  - General ones
    - Revenue, Clicks/Page Impressions, etc.
  - Service-specific, business-model specific ones
    - eBay: "purchase-through-rate", "bid-through-rate"
    - LinkedIn: Contact with employer made
    - Paper recommendation: "link-through", "cite-through"
    - E-Commerce marketplace: "click-outs"
    - Online dating: "open communications", "positive contacts per user"

Exciting times

- Huge economic value of recommendations
- Wide adoption in industry
- Modern technology used in organizations
  - Variety of machine learning techniques in place
  - Deep Learning adopted in industry
    - YouTube, Netflix, Facebook, Twitter …
  - Shortage of experts (like you)
- General expectations in AI / machine learning

Booming academic research

- The Beginnings
  - Tapestry news filtering (1992)
  - GroupLens (1994)
    - Matrix completion, nearest neighbors
- Success reports in e-commerce (early 2000s)
- Netflix Prize (2006-2009)
  - Matrix Factorization
- AI/ML Boom (- present)
  - Deep Learning

Booming academic research

- Countless papers published each year
  - KDD, TheWebConf, IJCAI, SIGIR, NeurIPS …
    - 117 papers alone in AAAI and IJCAI in 2018/2019[1]
  - ACM TOIS, IEEE TKDE, UMUAI, …
- ACM Conference on Recommender Systems
  - Constantly growing since 2007
- ACM Transactions on Recommender Systems
  - Established in 2021

Association for Computing Machinery

[1]Jannach, D. and Bauer, C.: *"Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research"*. AI Magazine, Vol. 41(4). 2020, pp. 79-95

Continued progress

teractions. Extensive experiments on a real-world dataset show that DAM outperforms the state-of-the-art solution, verifying the effectiveness of our attention design and multi-

level and individual-level. Extensive experiments

erent domains

ificant perfor-

state-of-the-art

ables us to directly learn a ranking function. Extensive online and offline experiments deployed on a commercial platform demonstrate that our models significantly increase diversity while preserving accuracy compared to the state-of-the-art sequential recommendation model, and consequently our models improve user satisfaction.

tion
two
pro
tain
sup
ommendatio

In addition, experiments on two real-world datasets demonstrate the superiority of our approach against other state-of-the-art approaches in terms of ranking accuracy and efficiency.

But wait …

ables us to directly learn a ranking function. Extensive online and offline experiments deployed on a commercial platform demonstrate that our models significantly increase diversity while preserving accuracy compared to the state-of-the-art sequential recommendation model, and consequently our models improve user satisfaction.

But wait …

- Higher prediction accuracy **consequently** leads to higher user satisfaction?

Notorious B.I.G.: Mo money mo problems

Measurements, revisited

- Real-world deployment:
  - Sales, Revenue, CTR, Engagement, Retention rate, Sales distributions, Conversion Rate, Time on site, …
- Academic papers, predominantly:
  - RMSE, MAE, NDCG, Precision/Recall, …

Use of **proxy measures / offline metrics**:

- **Academia**: Because we mostly don't have access to real systems
- **Industry:** Because A/B tests are expensive/risky

An important question

Computer Science > Information Retrieval

[Submitted on 7 Nov 2020]

Do Offline Metrics Predict Online Performance in Recommender Systems?

Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, Michael I. Jordan

Recommender systems operate in an inherently dynamical setting. Past recommendations influence future behavior, including which data points are observed and how user preferences change. However, experimenting in production systems with real user dynamics is often infeasible, and existing simulation-based approaches have limited scale. As a result, many state-of-the-art algorithms are designed to solve supervised learning problems, and progress is judged only by offline metrics. In this work we investigate the extent to which offline metrics predict online performance by evaluating eleven recommenders across six controlled simulated environments. We observe that

A plausible assumption

- Ranking performance **is** important
  - More relevant items appear high up in the list ➲
  - Making it easier for consumers to find them ➲
    - Leading to more consumption/sales
    - Leading to more satisfaction
      - e.g., because something useful was found
- Nothing wrong with this

Consider this case

- Imagine the user liked this

- We recommend

A good recommendation?

- Highly accurate ☑
- But useful?
  - Probably already seen
  - Monotone and obvious, satisfying?
  - Maybe not leading to additional engagement
  - Maybe a missed sales opportunity

Maybe still useful as a reminder or recommendation shortcut[1]

[1]Lerche, L., Jannach, D. and Ludewig, M.: *"On the Value of Reminders within E-Commerce Recommendations"*. In: Proceedings UMAP 2016

Purpose and usefulness

- What is a good (or: useful) recommendation?
- Recommender systems can serve a variety of purposes[1] for different stakeholders[2]
  - Consumer, e.g.:
    - Reduction of choice overload, discovery, understanding the space of options, entertainment, …
  - Provider, e.g.:
    - Sales, revenue, customer retention, …

[1]Jannach, D. and Adomavicius, G.: "Recommendations with a Purpose". In: Proceedings RecSys 2016
[2]Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J. and Pizzato, L.: "Multistakeholder Recommendation: Survey and Research Directions". UMUAI, Vol. 30. 2020, pp. 127–158

Purpose and usefulness

- Recommender systems can serve a variety of purposes for different stakeholders
  - Moreover: usefulness may depend on a variety of factors like user context, revenue models, etc.
- But what do we do?
  - Abstract from domain/application specifics ☑
  - Use ranking accuracy as **only** and **universal** proxy for everything
    - (Yes, there is research on diversity, serendipity etc., but suffering from similar problems)

The McNamara Fallacy

- Robert McNamara
  - US Secretary of Defense ('61 – '68)
- Vietnam War:
  - Assessing progress (also) with the enemy body count
- The Fallacy …
  - "… involves making a decision based solely on quantitative observations (or metrics) and ignoring all others." (Wikipedia)

In recommender systems research

- We focus on what we can easily measure
  - Prediction accuracy in offline experiments
- But do our assumptions actually hold?
  - Will (sometimes tiny) increases in accuracy on historical data lead to better recommender systems?
  - Where is the evidence?

Some studies

A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems

Joeran Beel[1] and Stefan Langer[1]

[1] Otto-von-Guericke Unive

Abstract. The evaluation of rec
ommender systems in practice. H
tle attention in the recommend
search-paper recommender syste
of different evaluation methods,
the context of research-paper re
filtering approaches in the resear

Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study

PAOLO CREMONESI and FRANCA GARZOTTO, Politecnico di Milano
ROBERTO TURRIN, M

User Perception of Differences in Recommender Algorithms

strand[1,2], F. Maxwell Harper[2], Martijn C. Willemsen[3], and Joseph A. Konstan[2]

Computer
ence
e University
s, TX, USA
txstate.edu

[2]GroupLens Research
Dept. of Comp. Sci. & Eng.
University of Minnesota
Minneapolis, MN, USA
{harper,konstan}@cs.umn.edu

[3]Human-Technology Interaction Group
School of Innovation Sciences
Eindhoven University of Technology
Eindhoven, The Netherlands
M.C.Willemsen@tue.nl

user evaluation of recommender systems
rful new tools for understanding what makes
ive and useful. We apply these methods to

algorithms with comparable accuracy [21, 12]. To map out some of those differences in the movie recommendation domain, we present a user study aimed at understanding the subjective differences users perceive between different collaborative filtering algorithms and how

Offline and Online Evaluation of News Recommender Systems at swissinfo.ch

Florent Garcin
Artificial Intelligence Lab
Ecole Polytechnique Fédérale de Lausanne
Switzerland
firstname.lastname@epfl.ch

Boi Faltings
Artificial Intelligence Lab
Ecole Polytechnique Fédérale de Lausanne
Switzerland
firstname.lastname@epfl.ch

Olivier Donatsch
SWI swissinfo.ch
Swiss Broadcasting Corp.
Switzerland

Ayar Alazzawi
SWI swissinfo.ch
Swiss Broadcasting Corp.
Switzerland

Christophe Bruttin
SWI swissinfo.ch
Swiss Broadcasting Corp.
Switzerland

Amr Huber
SWI swissinfo.ch
Swiss Broadcasting Corp.
Switzerland

ABSTRACT

We report on the live evaluation of various news recommender systems conducted on the website *swissinfo.ch*. We demonstrate that there is a major difference between offline

on Context Trees (CT) [7,
systems adapts its model t
erences. The model evolves
it is always up to date and
in real-time.

User-Centric Evaluation of Session-Based Recommendations for an Automated Radio Station

Malte Ludewig
TU Dortmund, Germany
malte.ludewig@tu-dortmund.de

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

What does the literature tell us?

- The choice of algorithm matters in practice
  - But often different families of approaches compared
- Offline/Online comparisons with user studies
  - Accuracy-based ranking of algorithms in most cases **not aligned** with user-related quality factors
  - Often only small differences in offline rankings
- Industry/Netflix:
  - *".. we do not find [offline experiments] to be as highly predictive of A/B test outcomes as we would like."*[1]

[1]Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 6, 4

McNamara Fallacy in RS

- We are obsessed with accuracy improvements
  - Even though algorithms with the same accuracy might recommend very different things[1]
  - Even though we do not know if small improvements matter in practice
  - Even though we do not know if offline accuracy is a good proxy for any business KPI or user-perceived qualities

[1]Jannach, D., Lerche, L., Kamehkhosh, I. and Jugovac, M.: "What recommenders recommend: an analysis of recommendation biases and possible countermeasures". User Modeling and User-Adapted Interaction, Vol. 25(5).

The question

- Why aren't we looking at relevant or interesting) questions?
  - Is it easier to publish such research?
    - Algorithms paper needs no explicit research question
    - Paper needs no underlying theory (e.g., about human behavior), even though recommending is an HCI problem
    - No need to involve users in the research
    - No need to craft an experimental design and defend it against reviewers
    - No need for complex statistical analyses
    - No need to discuss research limitations

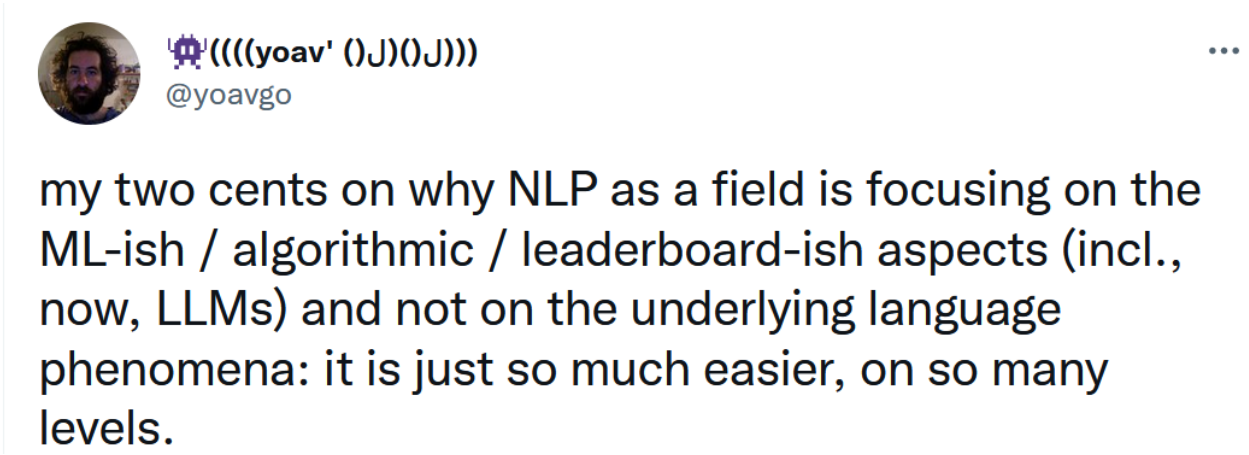
The question

- Why aren't we looking at relevant or interesting) questions?
  - Is it easier to publish such research?
    - Getting a paper through at a good (ML) conference or journal remains difficult, of course
    - A lot of competition
    - Mixed review quality
    - Beefing it up with math might help[1]

[1]Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. Queue, 17(1), February 2019.

A general "applied ML" problem?

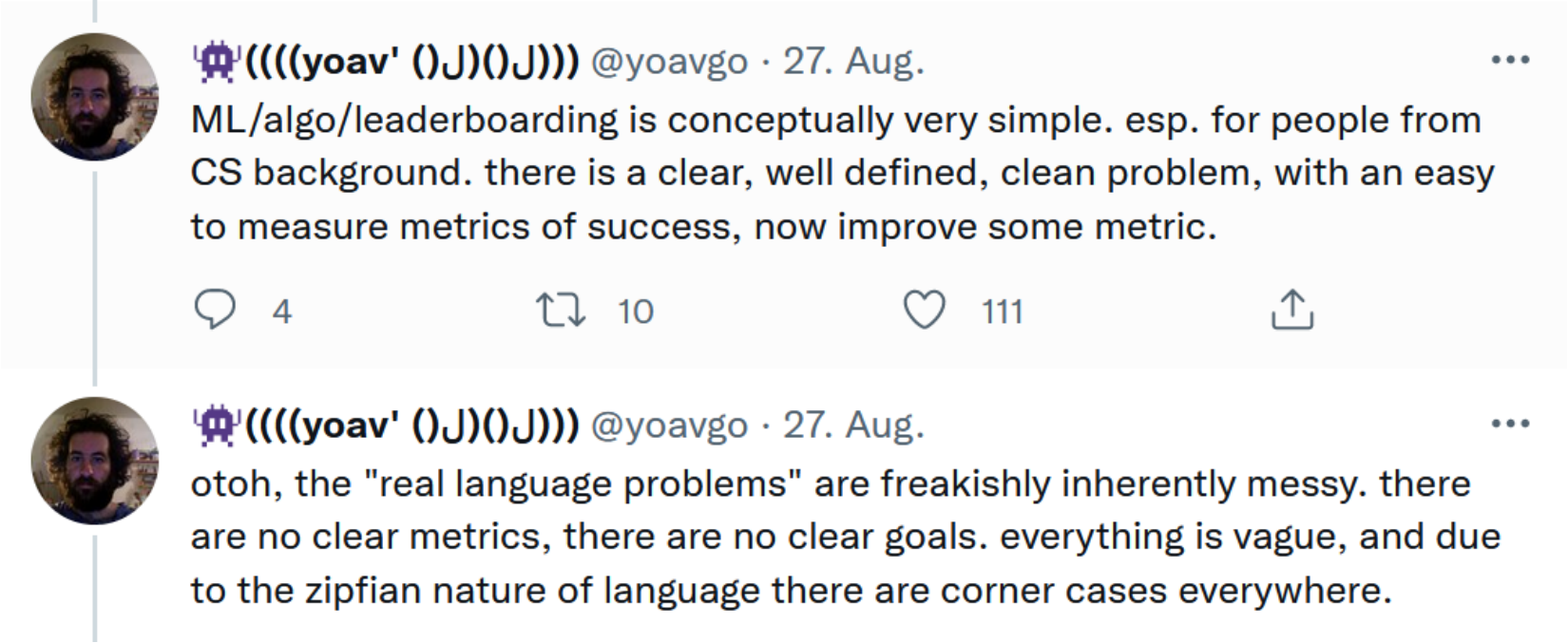
- Known hyper-focus on accuracy in applied ML[1]
- Critical voices: NLP[2]



[1]Kiri L. Wagstaff. Machine learning that matters. In Proceedings of the Twenty-Ninth International Conference on Machine Learning, ICML '12, pages 529–534, 2012.
[2] https://twitter.com/yoavgo/status/1431284873151528960

A general “applied ML” problem?

- Critical voices: NLP



(((yoav' ()J)()J))) @yoavgo · 27. Aug.

ML/algo/leaderboarding is conceptually very simple. esp. for people from CS background. there is a clear, well defined, clean problem, with an easy to measure metrics of success, now improve some metric.

4 10 111

(((yoav' ()J)()J))) @yoavgo · 27. Aug.

otoh, the "real language problems" are freakishly inherently messy. there are no clear metrics, there are no clear goals. everything is vague, and due to the zipfian nature of language there are corner cases everywhere.

A general “applied ML” problem?

- Critical voices

Buckman's Homepage

Please Commit More Blatant Academic Fraud

Posted on May 29, 2021

[1] https://jacobbuckman.com/2021-05-29-please-commit-more-blatant-academic-fraud/

A general "applied ML" problem?

- Critical voices
  - Day-to-day fraud
    - "Trying that shiny new algorithm out on a couple dozen seeds, and then only reporting the best few. "
    - "Running a big hyperparameter sweep on your proposed approach but using the defaults for the baseline."
    - "Cherry-picking examples where your model looks good, or cherry-picking whole datasets to test on, where you've confirmed your model's advantage"
  - As a result
    - "… even at top conferences, the median published paper contains no truth or insight."

It gets even worse

- But we sometimes don't even make true progress on these problematic measures
- We reproduced more than 2 dozen algorithms
  - Traditional matrix completion setups
  - Session-based recommendation
  - Session-aware recommendation

Ferrari Dacrema, M., Boglio, S., Cremonesi, P. and Jannach, D.: "A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research". ACM Transactions on Information Systems, Vol. 39(2). 2021
Ferrari Dacrema, M., Cremonesi, P. and Jannach, D.: "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches". In: Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019). Copenhagen, 2019
Ludewig, M., Latifi, S., Mauro, N. and Jannach, D.: "Empirical Analysis of Session-Based Recommendation Algorithms". User Modeling and User-Adapted Interaction, Vol. 31(1). 2021, pp. 149–181
Latifi, S., Mauro, N. and Jannach, D.: "Session-aware Recommendation: A Surprising Quest for the State-of-the-art". Information Sciences, Vol. 573. 2021, pp. 291-315

It gets even worse

- We reproduced more than 2 dozen algorithms
  - Using the original code
  - Using the original datasets
  - Using the original evaluation
- We found
  - Only one single case, where the proposed neural method was not outperformed by existing techniques (e.g., nearest neighbors)
    - In one case even recommending popular items was better
  - Severe methodological issues
  - Low reproducibility in general

What to do?

- The situation is difficult
  - Established research paths
  - Academic incentive system will not change
- Yoshua Bengio:
  - Time to rethink the publication process in machine learning (Blog post, 2020)
  - *"[…] our current system incentivizes incremental work and creates a lot of pressure and stress on grad students (and researchers in general) to submit as many papers as possible at each deadline."*
  - (Bengio's name appeared on > 80 papers in 2020)

What to do?

- Such a change in publication culture will take time
- Too late for current students
- No incentives or signs for such a change

The absolute minimum

- Make sure that our work is **reproducible**, share:
  - Code
    - Own model, baseline models, hyperparameter optimization procedure, evaluation, data pre-processing
  - Data
    - Original data, pre-processed data, data splits for evaluation
  - Documentation & "Infrastructure"
    - Readme, installation instructions
    - Scripts to run everything end to end
    - Maybe: Docker image

The absolute minimum

- Make sure that our work **progresses** the field
  - Systematically tune all baselines properly
  - Don't copy results from other papers without reproducing
  - Do not only include algorithms of one type (e.g., the latest deep learning models)
    - Sanity check: popular item recommendations, nearest neighbors
  - Justify the choice of
    - Datasets
    - Evaluation protocol, metrics and cut-offs based on the claims you make

What would be better?

- Investigate more relevant, but **challenging** questions, there are indeed many
  - (This is an advice for professors)
- For example:
  - Recommendation largely is a problem **of human-computer interaction** (HCI)
    - How to explain recommendations, how to persuade users, how to help them make better decisions?
    - How to increase their trust?
    - How to interact with them in a **conversation**?
    - How to visually present things?

The HCI Perspective

- Often more important than algorithms
  - A/B at swissinfo.ch
    - Around 30% increase in CTR with algorithm
    - **But:** "When the recommendations are at the top-right corner of the page, the CTR is more than double (2.25) the one at the bottom."
  - Also, recently:

The 'Unreasonable' Effectiveness of Graphical User Interfaces for Recommender Systems

Joeran Beel, University of Siegen, Germany, joeran.beel@uni-siegen.de
Haley Dixon, Trinity College Dublin, Ireland, dixonh@tcd.ie

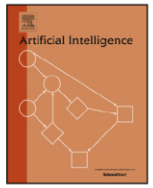
Garcin et al.: Offline and online evaluation of news recommender systems at swissinfo.ch", RecSys '14

Social Sciences Perspective - Explanations

Contents lists available at ScienceDirect

Artificial Intelligence

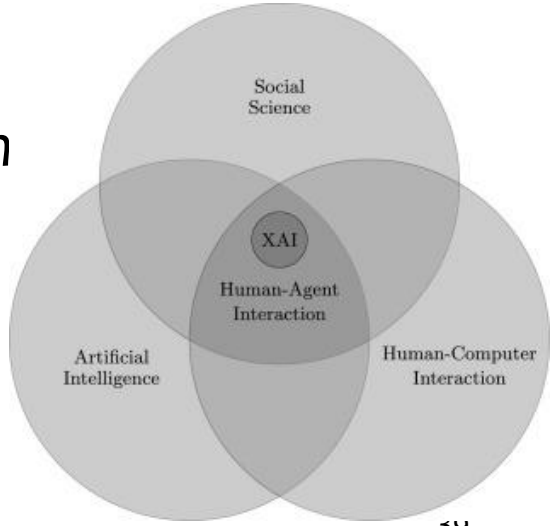
www.elsevier.com/locate/artint



Explanation in artificial intelligence: Insights from the social sciences

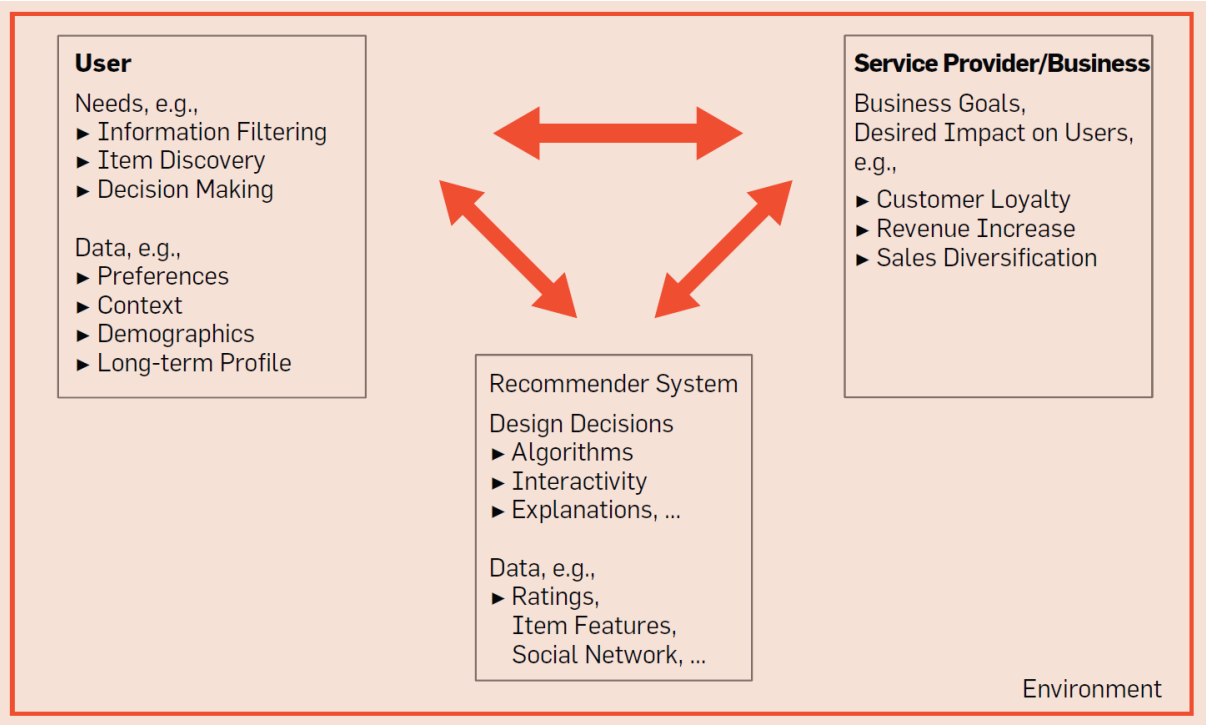
Tim Miller

"[…] explainable AI is not just 'more AI'. Ultimately, it is a human–agent interaction problem."



The Information Systems Perspective – Impact on Users and Providers

- Business value, multi-stakeholder perspectives



Jannach, D., Resnick, P., Tuzhilin, A. and Zanker, M.: "Recommender Systems - Beyond Matrix Completion". Communications of the ACM, Vol. 59(11). Association for Computing Machinery (ACM), 2016, pp. 94-102

What about offline + algorithms?

- Offline analyses remain relevant as well
- But maybe a shift would be helpful
  - Understand the problem first
    - What is a good recommendation in a given application context?
      - Diversity/popularity may be good in one application (or a specific user situation) but not in another
      - More analytics-based approaches required first
    - Choose evaluation designs with goal and purpose in mind
    - Validate the proxies
  - Only then try to generalize

Jannach, D. and Adomavicius, G.: "Recommendations with a Purpose". In: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016). Boston, Massachusetts, USA, 2016, pp. 7-10

What about algorithms?

- Evaluate with users!
- Exploring Simulation-based approaches
  - Offline A/B tests, counterfactual evaluation
    - Reduce the offline/online gap
    - "Ground-truth" often remains simulated, various assumptions
  - Agent-based Modeling and Simulation
    - Model system to observe emergent behavior
      - e.g. effects on provider strategy on long-term profitability
    - Typically: various simplifications, mainly to detect trends

Towards a more comprehensive research approach

- Avoid to be a "one-trick pony"
- Let's grow our methodological repertoire
  - Algorithms-based research
  - Studies with users (randomized controlled trials)
  - Qualitative research methods
    - Surveys, focus groups, towards a better understanding of the problem first
- Let's take a more scientific approach
  - What are our research hypotheses? What are our underlying theories? Are our experiments suited to support our hypothesis? How can our claims be falsified?

Thank you for your attention

dietmar.jannach@aau.at

https://tinyurl.com/rs-mc-namara

Announcement

Association for Computing Machinery

ACM Transactions on Recommender Systems

- A new journal entirely devoted to recommender systems
- Submissions open in January 2022
  - Technical papers, surveys, industry reports and case studies, opinion pieces