The Neural Hype and Comparisons Against Weak Baselines

Jimmy Lin
David R. Cheriton School of Computer Science, University of Waterloo

1 Introduction

Recently, the machine learning community paused in a moment of self-reflection. In a widely-discussed paper at ICLR 2018, Sculley et al. [13] wrote: "We observe that the rate of empirical advancement may not have been matched by consistent increase in the level of empirical rigor across the field as a whole." Their primary complaint is the development of a "research and publication culture that emphasizes *wins*" (emphasis in original), which typically means "demonstrating that a new method beats previous methods on a given task or benchmark". An apt description might be "leaderboard chasing"—and for many vision and NLP tasks, this isn't a metaphor. There are literally centralized leaderboards[1] that track incremental progress, down to the fifth decimal point, some persisting over years, accumulating dozens of entries.

Sculley et al. remind us that "the goal of science is not wins, but knowledge". The structure of the scientific enterprise today (pressure to publish, pace of progress, etc.) means that "winning" and "doing good science" are often not fully aligned. To wit, they cite a number of papers showing that recent advances in neural networks could very well be attributed to mundane issues like better hyperparameter optimization. Many results can't be reproduced, and some observed improvements might just be noise.

I'd like to suggest that similar self-examination is needed in our own community, especially with respect to the hype surrounding neural IR approaches. They are new, shiny, and have enchanted our youngest members—new students, many of whom find it hard to believe that anything even existed before neural networks. Yet it is unclear to me, at least for "classic" *ad hoc* retrieval problems without vast quantities of training data from behavior logs, whether neural techniques are actually more effective in absolute terms. As Sculley et al. suggest, (at least some) progress may be illusionary.

As in many other communities, a "win" is pretty much a prerequisite to getting a paper published at a top IR venue today. However, I am disappointed that it is not difficult to find neural ranking papers that demonstrate winning by showing statistically significant improvements over weak or inadequately-tuned baselines. These papers report comparisons to leaderboards selectively populated by entries of the authors' choosing, where better results are often ignored. To set up a contrast: for NLP tasks, even if there isn't a centralized leaderboard, it is customary for researchers to start with a previous paper that describes the state of the art, copy its results table, and then

[1] See, for example, `https://rajpurkar.github.io/SQuAD-explorer/`

add results from an improved model that claims some contribution. Thus, a *de facto* leaderboard is reproduced from paper to paper. The common behavior in IR where results are only measured against cherry-picked points of comparison would unlikely pass peer review in, for example, an ACL submission.

This observation isn't new.

About ten years ago, Armstrong et al. [4] published their now-famous "improvements that don't add up" paper illuminating the worrying state of affairs in empirical research on ranking models for *ad hoc* retrieval, wagging their fingers at the community. Writing in 2009, their conclusion was damning: "There is, in short, no evidence that ad-hoc retrieval technology has improved during the past decade or more." This finding was arrived at by a comprehensive longitudinal survey of research papers between 1998 and 2008 from major IR research venues that report results on a diverse range of TREC test collections. After careful analysis, the authors placed much of the blame on the "selection of weak baselines that can create an illusion of incremental improvement" and "insufficient comparison with previous results".

Ten years later, where are we now? I haven't replicated the careful meta-analysis of Armstrong et al. and thus cannot make generalizations about the literature as a whole, but it is still easy to find instances of comparisons to weak baselines. In some ways, things are actually worse. Armstrong et al. found it "surprising" that the lack of an upward trend in *ad hoc* retrieval effectiveness "appears to have gone largely unnoticed within the IR community." They wrote in their conclusion: "Perhaps most urgently of all, though, we should as a community take stock of the situation we find ourselves in." Ten years ago, the community had ignorance as a defense. Today, there is no excuse.[2] Had we collectively heeded the admonishments of Armstrong et al., the examples presented below should not exist in the literature (or should at least be difficult to find). To be absolutely clear, my arguments about weak baselines are existentially (not universally) quantified and this piece expresses my personal *opinion*.

Why do I specifically pick on neural IR? In a seminal essay, Ioannidis [7] suggested that "the hotter a scientific field… the less likely the research findings are to be true" and that "claimed research findings may often be simply accurate measures of the prevailing bias". Pfeiffer and Hoffmann [12] found empirical support for this inverse correlation between popularity and reliability in the scientific literature on protein interactions. While the issue of comparisons to weak baselines applies to *all* empirical studies of ranking models, the hype surrounding neural IR approaches makes them especially vulnerable to unwarranted claims.

2 Empirical Support

To support the assertions made in the introduction, I discuss two recent neural IR papers that in my opinion report comparisons to weak baselines. For reference, I present the effectiveness of a number of well-known (and in some cases, decades-old) ranking models after careful tuning on the test collection from the TREC 2004 Robust Track: documents from TREC Disks 4 & 5 (minus Congressional Records) with topics 301–450 and 601–700 (henceforth, Robust04).

[2] And it's not a matter of not knowing about the work: As of October 2018, the paper has been downloaded around 700 times according to the ACM Digital Library and has been cited over 130 times according to Google Scholar. Entire workshops have been devoted to issues raised in the paper [2].

This test collection was selected primarily for two reasons: First, it is generally acknowledged as being a high-quality resource, with a sufficient number of topics to draw generalizable conclusions and reasonably complete relevance judgments (compared to, say, ClueWeb collections). Second, I wish to focus on models that do not depend on having access to large amounts of behavioral data—in other words, the "classic" ad hoc retrieval task (more discussion on this later).

The ranking models below were considered, with the following parameters explored via (exhaustive) grid search:

- BM25, k_1 from $[0.1, 4.0]$ with a step size of 0.1 and b from $[0.1, 1.0]$ with a step size of 0.05.
- Query likelihood with Dirichlet smoothing (QL), the smoothing parameter μ from $[0, 10000]$ with a step size of 250.
- RM3 variant of relevance models [1]: the number of feedback documents N and the number of feedback terms M from $[1, 50]$ with a step size of 1, the weight of the original query λ from $[0, 1]$ with a step size of 0.1.

All experiments were conducted with Anserini [16, 17], an open-source information retrieval toolkit built on Lucene.[3]

How do results from Anserini compare to the latest papers on neural ranking models? I discuss two papers published within the last two months that conducted evaluations on Robust04, referring to these papers simply as Paper 1 and Paper 2 (purposely omitting references). It is not my intention to personally attack or to single out any particular group of researchers for criticism, but to merely document instances of what I believe to be a prevalent problem.

Paper 1 appears in a top IR research venue. Table 1 copies results from that paper on Robust04, where the names of the models have been blinded. Neural$_3$ and its variant Neural$_3{}'$ are the contributions of the authors, while Neural$_1$ and Neural$_2$ are other points of comparisons. Paper 1 performs two-fold cross validation, and our experiments replicate the same conditions, using the same folds (which we obtained from the authors). Anserini results are shown in the bottom rows of the table. We see that our QL results are fairly close to the QL results reported in Paper 1, which indicates that the authors did a fine job tuning the QL baseline in their paper. However, it appears that BM25 provides a slightly better bag-of-words baseline. Therefore, for the query expansion model RM3, we decided to use BM25 as the base ranking model for initial retrieval as well as for the expanded query. This differs from Paper 1, which uses QL as the base retrieval model. The query expansion model was tuned in two ways, what we call "independent" and "joint". In the first method, we fixed the underlying BM25 parameters and tuned only the RM3 parameters. In the second method, we performed end-to-end optimization and tuned all the parameters at once; for efficiency reasons, we started with parameters from the "independent" condition and explored a more restricted range around those.

We observe that a well-tuned query expansion model (RM3) achieves effectiveness on par with Neural$_3{}'$ in terms of AP (although P20 is slightly lower). In other words, the improvements that the authors of Paper 1 report mostly disappear when compared to a well-tuned baseline. Without access to the original runs, it is not possible to conduct significance testing. However,

[3] Anserini is available at `anserini.io`. All experiments were conducted on a post-v0.2.0 release, commit id `2c8cd7a550`.

Condition	AP	P20
QL	0.2499	0.3556
QL + RM3	0.2865	0.3773
Neural_1	0.2815	0.3752
Neural_2	0.2801	0.3764
Neural_3	0.2856	0.3766
$\text{Neural}_3{}'$	0.2971	0.3948
Anserini: QL	0.2496	0.3543
Anserini: BM25	0.2526	0.3604
Anserini: BM25 + RM3 (independent)	0.2954	0.3885
Anserini: BM25 + RM3 (joint)	0.2973	0.3871

Table 1: Comparison of Anserini results to Paper 1.

I am skeptical that the proposed models "significantly outperform competitive baselines", as the authors claim. Note that our RM3 results are in the same ballpark as the 0.306 AP reported by Benham et al. [5] on Robust04; that figure is with five-fold cross-validation and under slightly different (easier) experimental settings, so it is not directly comparable to the numbers in Table 1. Nevertheless, there are other examples of effective and well-tuned baselines in the literature.

Paper 2 appears in a top research venue, although not in information retrieval. Table 2 copies results from that paper, where the names of the neural models have also been blinded: "A", "B", and "M" are elements of the paper's contributions, building on existing models, Neural_x and Neural_y. The Anserini results shown at the bottom of the table are from five-fold cross validation using the same folds as the authors' work. Because of the different folds, Anserini results from Table 1 and Table 2 are slightly different, but the models and parameter tuning methods ("independent" and "joint") are the same (although it appears that joint optimization overfits in this case). In Paper 2, the authors only measured their contributions against two variants of BM25: an off-the-shelf implementation and the same implementation with light feature engineering. Both in my opinion are weak baselines and inadequately-tuned, since their reported BM25 results are quite a bit worse than Anserini's. The paper claims that "these IR baselines" (referring to BM25 and BM25 + Features) "are very strong", which is not a characterization I would agree with. None of the neural models in Paper 2 come close to the effectiveness of the RM3 implementation in Anserini.

To be completely fair, the primary experiments in Paper 2 focused on another collection with different characteristics, and the authors present results on Robust04 primarily to provide context. While the goal of their work is to improve upon existing neural models, which they do indeed demonstrate, it would have been desirable to present results from other mature models (such as RM3) to give the reader a better sense of how the neural models stack up in absolute terms.

My arguments thus far have only focused on comparisons to weak baselines, but even well-tuned baselines should be viewed as a minimum threshold. In my opinion, newly-proposed ranking models should be compared against the state of the art, such as the 0.3277 AP on Robust04 reported by Dalton et al. [6] for their entity query feature expansion (EQFE) model. Or the 0.3331 AP of run `pircRB04t3` at Robust04 (the highest score that year), or any of the three teams

Condition	AP	P20
BM25	0.238	0.354
BM25 + Features	0.250	0.367
Neural$_x$	0.258	0.372
Neural$_y$	0.256	0.370
Neural$_x$ + Neural$_y$	0.259	0.373
A + Neural$_y$	0.263	0.380
A + Neural$_y$ + M	0.265	0.380
B + Neural$_y$	0.270	0.383
B + Neural$_y$ + M	0.272	0.386
Anserini: QL	0.2481	0.3517
Anserini: BM25	0.2528	0.3598
Anserini: BM25 + RM3 (independent)	0.2991	0.3901
Anserini: BM25 + RM3 (joint)	0.2956	0.3931

Table 2: Comparison of Anserini results to Paper 2.

that reported AP over 0.3 [15]. When presenting results on a test collection, one shouldn't simply ignore previous work that reports higher effectiveness. As in other communities, top entries on the leaderboard for a particular task should be included, at the very least, to provide context.

Having discussed these two papers, I wish to address a few possible objections and discuss one additional question.

But you're not taking advantage of behavioral data! Correct, and this is intentional. I have specifically focused on neural ranking models that do not require behavioral data (e.g., query and click logs). Yes, it is certainly believable that search engine companies can pump vast quantities of behavioral data into training neural models to achieve impressive gains, but even conceding this, several issues remain:

First, these results are neither repeatable, replicable, nor reproducible[4] by academics. For competing search engine companies, such results are neither repeatable nor replicable, but possibly reproducible. However, given commercial interests and IP issues (e.g., patents) it is highly unlikely that reproducibility efforts (either positive or negative) will be publicized. To be harsh, neural ranking models that depend on large amounts of behavioral data are unscientific.

Second, one of the purported advantages of neural ranking models—and the continuous representations that underlie them—is the ability to capture semantics (i.e., these might be more precisely called semantic matching models). If this were indeed the case, then there should be no need for behavioral data to improve *ad hoc* retrieval effectiveness. If we look at the natural language processing community, for example, we see that neural networks have yielded demonstrable gains in a wide variety of tasks *without* requiring large amounts of behavioral data. If

[4] I use these terms precisely as ACM defines them at `https://www.acm.org/publications/policies/artifact-review-badging`: repeatable ("a researcher can reliably repeat her own computation"), replicable ("an independent group can obtain the same result using the author's own artifacts"), and reproducible ("an independent group can obtain the same result using artifacts which they develop completely independently").

the neural hype is justified, then why haven't we seen similarly convincing gains in IR? A more accurate summary might be that for problems where there is a plethora of behavioral data from a multitude of users, neural networks can yield impressive gains in effectiveness. I would have no objections to more carefully-scoped claims along these lines.

Neural approaches open up new opportunities for exploration. This is fundamentally a diversity argument, where the claim is that existing models represent local maxima, and thus we should explore new parts of the landscape. This is a valid argument: an analogy might be the introduction of language modeling techniques in the 1990s. In the beginning, it was unclear whether language modeling was actually better than what came before in terms of effectiveness (e.g., BM25), but the approach undeniably opened up the field to thinking about the search problem in new and interesting ways. Similar arguments might be made about the axiomatic approach to IR—questionable metric improvements, but new insights in terms of a formal framework for ranking. Such an argument recognizes that "winning" isn't the only game, which is positive for the community—but it's not consonant with the neural hype. I would have no complaints if neural IR papers conceded that their proposed models fall short in terms of state-of-the-art effectiveness, but claimed contributions in delivering insight.

Neural ranking models can always take advantage of strong baselines. Most neural ranking models are deployed as rerankers over initial sets of candidate results retrieved using a simple model like BM25. It would be natural to argue that neural models can also take advantage of strong baselines. However, Armstrong et al. explicitly dispelled this "additivity of improvements" argument. Techniques that yield improvements over weak baselines often do not show similar improvements over strong baselines. In other words, we cannot treat ranking models as a bunch of individual innovations, throw them all together, and expect the improvements to be additive. The follow-up work of Kharazmi et al. [8] validated these findings, showing that in some cases, a technique that improves a weak baseline actually *harms* a strong baseline. Thus, while it may indeed be true that some neural ranking models can exploit strong baselines, such claims should not be taken without empirical support.

A related point is that many neural ranking models only "work" after they incorporate a feature like BM25 explicitly, the simplest approach being interpolation between the BM25 retrieval score and the score from the neural network. Often, after normalization, most of the weight is placed on the BM25 score. In other words, the neural ranking model is "doing a lot" (for example, millions of parameters and operations in feedforward inference) for not a lot of gain.

But this is only one test collection. Sure, but Robust04 is one of the most widely used newswire collections available, among the largest (if not the largest) in terms of available high-quality relevance judgments. If neural ranking models "can't make it work" here, I'm skeptical that more collections will change the general contours of my argument. Absolutely, more experiments need to be conducted to support proper generalizations.

Why are these baselines better? An interesting question is why implementations in this paper appear to be more effective than other implementations of the *same model*? There are several explanations, the first of which is that model descriptions are woefully under-specified, a phenomenon previously noted [9]. For example, Mühleisen et al. [10] examined four different systems that all purport to implement BM25, but their rankings differ in effectiveness substantially. Similarly, Trotman et al. [14] cataloged at least half a dozen variants of what researchers have referred to generically as "BM25" or "query likelihood with Dirichlet smoothing". Ultimately, experiments compare *implementations*, not *mathematical models*. Dozens, if not hundreds, of seemingly small decisions (tokenization, stemming, stopwords, pruning of the vocabulary space, edge cases, etc.) are all consequential, especially in aggregate. In most cases, there is no documentation of these small individual design choices other than the code itself. Thus, although the *name* of the model may be the same, what in fact is being compared may be very different.

However, I believe that the biggest difference is the amount of effort devoted to parameter tuning. I suspect that, in general, researchers are not as rigorous about baseline tuning as they should be. In fact, a carefully-tuned baseline only reduces the "amount of winning" of the proposed contribution. In short, there is likely at least some confirmation bias at play.[5]

3 Constructive Suggestions

I'd like to conclude with some constructive suggestions:

Pick the best implementation. Despite our fondness for beautiful mathematical models, information retrieval is at its core an empirical discipline. Implementations matter, more so than models, and thus it makes sense to pick the best one.

The notion of "best", however, is hard to define and comprises many characteristics that are no doubt desirable. However, I argue that Lucene is the Pareto-optimum solution. Experiments here and elsewhere [16, 17] have established that Lucene provides effective rankings; additional studies have shown that Lucene is also reasonably efficient [9, 17]. Beyond the academic ivory tower, Lucene and its derivatives (Solr and Elasticsearch) have become the *de facto* platform for building real-world search applications outside a handful of web search engine companies that deploy custom infrastructure. Lucene powers search at organizations as diverse as Twitter, Bloomberg, Reddit, Wikimedia, and Target. This production experience means that the Lucene codebase is battle-tested and industrial-strength. From the perspective of software engineering practices, quality of artifacts, and diversity of capabilities, there isn't much contest between the legions of salaried developers and volunteers who contribute to Lucene vs. the meager resources of an academic research group in building and maintaining their own search engine.

Thus, I recommend that the research community rally around Lucene and adopt it as the foundation on which to build. The Anserini project represents an effort to support academic IR research with Lucene by implementing missing capabilities: for example, Anserini provides all the necessary components to replicate standard TREC experiments out of the box. One simply needs to copy and paste a few commands from a readme to obtain reasonable (untuned) results. Scripts for parameter tuning, like those used to generate the results reported here, are also included. For a neural ranking model deployed as a reranker over an initial set of candidate documents,

[5] To be absolutely clear, I am not suggesting that anyone engaged in untoward behavior—simply that fallible human beings are susceptible to cognitive biases.

Anserini can provide the strong baselines necessary to demonstrate genuine progress. As a bonus, Lucene-based implementations provide a smoother transition path to production deployment, to help bridge the chasm between research and practice in information retrieval.

Build an execution-centric leaderboard. Although "leaderboard chasing" no doubt has its downsides, I believe it remains an important component of a healthy research ecosystem that also values knowledge and insights. However, we can and should be much more rigorous in the comparisons made in papers, and one way to encourage researchers to deploy appropriate baselines is to make clear what those baselines should be. To complement their paper, Armstrong et al. developed EvaluateIR [3], an online tool for comparing IR systems, whereby researchers could upload runs to a central repository—in essence, a leaderboard! Using the tool, researchers, reviewers, and readers can "easily establish whether published results demonstrate a genuine advance in effectiveness" via comparisons to other submitted runs. Unfortunately, EvaluateIR never gained traction, and a number of similar efforts following it have also floundered. Nevertheless, the underlying ideas remain solid, and I believe that if the community builds a better leaderboard, "wins" can be more confidently validated and this will, I argue, actually move the field away from leaderboard chasing. How? Allow me to explain.

I offer two additional suggestions on how a renewed push building on EvaluateIR might be successful. The first is a technical improvement, while the second is a policy prescription. In my opinion, code execution is a vital element missing from the original effort, since it focused on archiving *run files* and did not include a mechanism to capture code that generated those results. The reproducibility effort I organized in 2015 [9] explicitly gathered scripts necessary to replicate runs on a standard test collection, but focused mostly on bag-of-words baselines and lacked the underlying analytical features of EvaluateIR. Both are needed, and modern tools like Docker make such efforts practical. In other words, what the community needs is a leaderboard in which execution is a first-class citizen. One could easily imagine that submissions to the leaderboard take the form of Docker images that the evaluation infrastructure then executes on a blind held-out test set. With such an execution-centric leaderboard (coupled with suggestions below), our community can combat weak baselines as well as repeatability and replicability in one swoop.

Bootstrap cultural changes with top-down edicts. Upton Sinclair, a prolific American author, once quipped that "it is difficult to get a man to understand something, when his salary depends upon his not understanding it". This quote, I believe, explains the underlying psychology of inaction. In this analogy, our "salary" represents the primary metric by which research productivity is measured: publication output. Our community is faced with a collective action problem—unless everyone also does it, holding oneself to a higher standard (i.e., "doing good science") is actively harmful for an individual researcher or group, since it means a higher bar for publication, and hence diminished output for the same amount of effort. Self-interested actors are behaving in a rational manner within the current incentive structures by continuing to compare against weak baselines. Thus, until we can address this issue, no amount of ranting or technical infrastructure will make a meaningful difference. In other words, little progress will be made until there's a cultural change.

Normally, top-down efforts at affecting cultural change are doomed to failure, but I believe there are tenable solutions in this case. The PC chairs of future IR conferences can simply "outlaw" the use of weak baselines. This can be instituted in a variety of ways: In the call for papers, it can simply be declared that “bag of words” models do not qualify as “competitive baselines”. In the review form, a question that explicitly asks about the quality of the experimental comparisons (if applicable) can be added. In the meta-reviewing form, senior PC members can be asked to verify the assessment of the reviewers. At the PC meeting, the PC chairs can orchestrate cross-paper analyses to ensure fair treatment.

As long as these procedural changes are transparent and applied consistently, this collective action dilemma can be solved by edict until the cultural changes become ingrained. Such actions are completely within the purview of PC chairs and require no new “powers” to be granted to them. Our community already has well-developed policies against double submission and plagiarism, so a check weeding out weak baselines can simply be added to the “standard operating procedure”. PC chairs are given substantial discretion to affect change, and indeed they periodically exercise such powers. Increasing page lengths, introducing entirely new categories of papers, and restructuring reviewer pools are all changes made to SIGIR in recent memory. Should future PC chairs wish to take up this challenge, I’ve sketched out a possible roadmap.

Would these changes make it harder to publish? Perhaps, but shouldn’t we prioritize good science? Instead of artificially manufacturing wins, we might begin to admit that genuine progress is difficult. Instead of wins becoming a necessary condition for publication, perhaps we’ll come to recognize knowledge and insight as valuable currency as well. In other words, I believe that building a better leaderboard and instituting rigorous process checks will make “wins” more difficult to achieve, and thus encourage the community to demonstrate contributions in other ways.

Back to basics and the self-driving search engine. Nearly a decade ago, Armstrong et al. concluded their paper as follows:

> Indeed, as a concrete challenge, perhaps it is time for us to take on what should be an attainable goal - let us build a public system that matches the BM25 run in the 1994 TREC-3 experiment, and then add to it the fruits of the past fifteen years’ research, to form a new baseline against which future effectiveness improvements can be properly measured.

Over the past few years, the Anserini team has been exactly trying to do this.

To complement this effort, I propose a new research program called “back to basics” (which meshes well with the resurgence of interest in “classic” *ad hoc* retrieval represented by the TREC “Common Core” Track). My hypothesis is that the community is already equipped with the techniques to make substantial improvements to *ad hoc* retrieval, we just need to learn how to better deploy the tools at our disposal. What does “better deployment” mean? Let’s just consider the simple case of parameter tuning, the focus of much of this piece.

Table 3 reports results on Robust04 under several hypothetical scenarios. The rows marked “(optimal, overall)” report effectiveness under optimal parameter settings across all topics—that is, if an oracle revealed the best parameter setting. The rows marked “(optimal, per topic)” report effectiveness under optimal *per topic* parameter settings—that is, if an oracle revealed the best parameter setting *for each topic*. These results show that large gains are possible if we simply twiddled the knobs correctly. Comparing the “(optimal, overall)” results with figures in Table 1

Condition	AP	P20
Anserini: QL (optimal, overall)	0.2496	0.3572
Anserini: QL (optimal, per topic)	0.2703	0.4008
Anserini: BM25 (optimal, overall)	0.2532	0.3614
Anserini: BM25 (optimal, per topic)	0.2921	0.4426
Anserini: BM25 + RM3 (optimal, overall)	0.3020	0.4012
Anserini: BM25 + RM3 (optimal, per topic)	0.4402	0.6054

Table 3: Anserini effectiveness under various upper-bound conditions.

and Table 2, we see that cross validation gets us pretty close to optimal parameter settings across topics. However, properly adjusting parameters per topic could further increase BM25 effectiveness by quite a bit—in fact, not too far from the best results reported in Paper 1. With RM3, optimal per-topic tuning yields impressive gains as well. Of course, these are oracle upper bounds; it remains to be seen how much of these gains can be actually realized.

Am I seriously proposing an entire research program around parameter tuning? Yes, and the initial reaction might be... how boring! To make this research sexy, a rebranding effort is needed. First, I propose to rename all parameters hyperparameters, thus enabling smart-sounding statements like "we apply meta-learning techniques to optimize the hyperparameters b and k_1 for BM25". Second, I propose to call this effort "building a self-driving search engine". Pavlo et al. [11] shamelessly stole this slogan from self-driving vehicles and applied it to databases. The catchphrase has become so successful that even Oracle got on board, introducing "the world's first self-driving database".[6] The "self-driving" notion builds on the decades-old idea that a database should be able to tune its own performance by monitoring query execution. Hey, the phrase has already been ripped off once—let's do it again for search engines.

Another instance of this "back to basics" research program is exemplified by the work of Benham et al. [5], who noted that multiple query variants combined with rank fusion techniques can yield quite impressive results: 0.325 average precision on a variant condition of Robust04. Like parameter tuning, these techniques do not require new ranking models, just better "deployment". In truth, all the machinery for building the self-driving search engine already exists. Threads in the literature on performance prediction, query expansion, rank fusion, and risk minimization provide productive starting points. The funny thing is, neural networks can still be exploited to build the self-driving search engine,[7] thereby rechanneling the raw energies of the neural hype into an alternative endeavor.

To conclude, if the original Armstrong et al. paper was the seminal "the emperor has no clothes" moment, the last decade has shown the emperor's lack of shame, since he's still wandering around naked today. The ultimate purpose of this opinion piece isn't simply to embarrass the emperor, but to affect real positive change. I've called out the fact that IR papers *still* bash weak baselines and proposed a number of constructive and practical suggestions. Let this opportunity for reflection serve as a "call to arms" to the community to do better science. Hopefully, in another ten years, there won't be a need to write something similar again.

[6] `https://www.oracle.com/database/autonomous-database/feature.html`

[7] For example, Pavlo et al. [11] use RNNs for workload forecasting in their self-driving database.

4 Acknowledgments

I'd like to acknowledge Peilin Yang and Kuang Lu for code contributions to Anserini that enabled the experiments reported in this paper. Thanks to Tim Armstrong, Charlie Clarke, Shane Culpepper, Hui Fang, Claudia Hauff, Alistair Moffat, Arjen de Vries, Chenyan Xiong, and Justin Zobel for helpful comments and suggestions on previous drafts. The authors of Paper 1 and Paper 2 were provided an earlier draft of this piece; comments from the lead author of both papers have been incorporated in the final draft.

References

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, M. D. Smucker, T. Strohman, H. Turtle, and C. Wade. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, Maryland, 2004.

[2] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. Report on the SIGIR 2015 workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2):107–116, 2015.

[3] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. EvaluatIR: An online tool for evaluating and comparing IR systems. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 833–833, 2009.

[4] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 601–610, 2009.

[5] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. Towards efficient and effective query variant generation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, Bertinoro, Italy, 2018.

[6] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 365–374, New York, NY, USA, 2014. ACM.

[7] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[8] S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. Examining additivity and weak baselines. *ACM Transactions on Information Systems*, 34(4):Article 23, 2016.

[9] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. Toward reproducible baselines: The open-source IR reproducibility

challenge. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*, pages 408–420, Padua, Italy, 2016.

[10] H. Mühleisen, T. Samar, J. Lin, and A. de Vries. Old dogs are great at new tricks: Column stores for ir prototyping. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 863–866, 2014.

[11] A. Pavlo, G. Angulo, J. Arulraj, H. Lin, J. Lin, L. Ma, P. Menon, T. C. Mowry, M. Perron, I. Quah, S. Santurkar, A. Tomasic, S. Toor, D. V. Aken, Z. Wang, Y. Wu, R. Xian, and T. Zhang. Self-driving database management systems. In *Proceedings of the 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017)*, Chaminade, California, 2017.

[12] T. Pfeiffer and R. Hoffmann. Large-scale assessment of the effect of popularity on the reliability of research. *PLoS ONE*, 4(6):e5996, 2009.

[13] D. Sculley, J. Snoek, A. Rahimi, and A. Wiltschko. Winner's curse? On pace, progress, and empirical rigor. In *Proceedings of the 6th International Conference on Learning Representations, Workshop Track (ICLR 2018)*, 2018.

[14] A. Trotman, A. Puurula, and B. Burgess. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 58:58–58:65, 2014.

[15] E. M. Voorhees. Overview of the TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, Maryland, 2004.

[16] P. Yang, H. Fang, and J. Lin. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1253–1256, 2017.

[17] P. Yang, H. Fang, and J. Lin. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16, 2018.