Worst Practices for Designing Production Information Access Systems*

Fernando Diaz
Microsoft
fdiaz@microsoft.com

Abstract

Information access systems have become a core part of everyday life for a large variety of users. Behind these large systems are decades of academic information access research. Unfortunately, there is a significant gap between studying and implementing information access systems. In this presentation, I will concentrate on open problems in production systems that academia is better suited for addressing than industry.

1 Introduction

I want to begin by presenting the following device, invented in 1920 (Figure 1). The shoe-fitting fluoroscope provided shoe stores with the ability to customize shoes so that each customer received the ideally-sized shoe. At the time, it was considered a profound application of cutting edge technology to a consumer market. The device was commonplace in shoe stores. Customers used it regularly and loved the shoes. However, within a few years, it would be difficult to find a fluoroscope anywhere in the country. I will elaborate on the reason for this failure further along in this essay.

This essay deals with current practices employed in the development of online services that may have profound negative impact on our users and perhaps society in general. So, this presentation will be less of a celebration of industry in as much as it will be a critique of the state of the art. Hopefully, this will provide an opportunity for growth as a community.

I will be covering two problems in our current practices. There are many others.

2 Bias

Today, almost every component of an information access systems relies on machine learning. Those of us in industry build classifiers everyday, using off the shelf algorithms and sometimes even off the shelf models. This technology has permeated much of the information retrieval stack, including

*Based on an essay of the same title presented at the ECIR 2016 Industry Day.

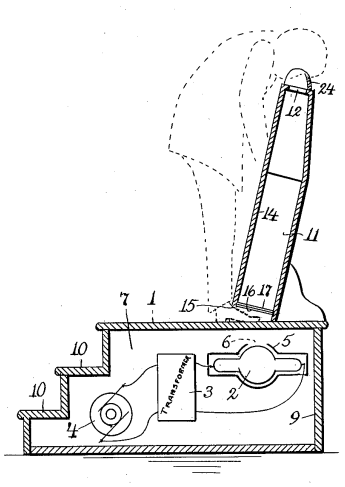

Figure 1: Figure from patent application for the shoe-fitting fluoroscope [14].

crawling, indexing, and ranking. Query classifiers decorate and modify user input with machine learned models. And, I am sure that, in a few years, information retrieval papers from the early 2000's will be seen as a series of papers applying machine learning to classic search subtasks.

At risk of alienating experts, I want to describe how these models are usually trained. I will focus on supervised classification, although other learning tasks follow. There are four parts to training a model to classify an object such as a document into a class such as relevance. First, we define a target or label we would like to predict. This may be 'a document's relevance to a query' or 'a query's classification'. Defining concepts like 'relevance' has been part of IR and continues to be debated, both in academia and industry. Second, we define attributes or features of the object that we believe are correlated with the label. In the context of document ranking, this includes features like BM25 or document quality. Third, we collect training data either from editorial assessment or behavioral signals. For example, Cranfield-style collections usually ask trained assessors to mark the relevance of documents to queries. In the final phase, we train a model that learns the statistical relationship between the features and relevance in the labeled data.

Many of us are familiar with this, or similar, strategies for problem-solving and it has been quite effective. However, I want to point out some specific ways this methodology can go very wrong. These examples are drawn from and inspired by Solon Barocas and Andrew D. Selbst's California Law Review article, "Big data's disparate impact", published earlier this year [3].

Consider first how we define our labels. We like to think that our definitions are constructed in some clean room, absent personal biases. But things like query classes presuppose that we can fit a spectrum of information needs into a single taxonomy for all users, regardless of background. In the best case, the system designer may make an effort to understand the complexity of possible intents, taking into account potential users' backgrounds. What is more likely to happen is that the system designer will adopt a taxonomy biased by their personal experience or a vague understanding of the users. As a result, a user's own nuanced behavior may be overlooked because they are a statistical outlier or belong to an underrepresented group.

Choosing which features to include in the model is equally problematic [5]. Even when using

overcomplete feature representations, this larger set is still hand-curated, potentially affected by biases of the system designer. Consider using a user's stated undergraduate institution in a reading level predictor. Such a model will likely find some correlation between the feature and reading level. But it will also be making decisions based on a feature that may result in estimating lower reading levels for individuals incapable of affording more expensive universities.

Biases can similarly be introduced in the data labeling phase. When data is editorially labeled, not only do we have potential biases in the label definition, but now we compound this with biases in an assessor's interpretation of the guidelines. Imagine the potential issues that might arise if we ask assessors for judgments about whether a user with specific demographics would find specific documents relevant. When labels are gathered from behavioral data such as real search traffic, a similar problem arises because it is often an engineer or system designer indicating which signals are correlated with the target. A good case here arises when we use cursor movement for inferring relevance, as has been advocated by many researchers, myself included [8]. This may seem innocuous until we think about populations of users who systematically underutilize their mouse, such as those using screen readers.

This is related to biases which may arise when deciding what data to collect for training our model. Indeed, the drawing of training data from internet users in general excludes the 60% of the world population without internet access [11]. When we consider that, even amongst individuals with access, internet usage may be higher for certain socioeconomic sectors and may result in overemphasis of those users' data.

Putting this all together, we are often left with a very distorted perspective on our target, features, and data. Unfortunately, our models to date are not built to compensate for these biases. As a result, any biases in the data are incorporated into the model with little opportunity for detection.

Now, I've been describing this at a high level. So, let me assure you that these biases are having real effects. As a sanity check, I would like to confirm that even human assessors are capable of biased decisions in online systems. Facebook's Real Name policy is intended to remove fake accounts, usually by flagging names as potentially fictitious and then conducting manual review [15]. Unfortunately, even though the review is done by a human, certain demographic groups were disproportionately denied access [2]. These included Native Americans and members of the LGBT community. Moritz Hardt, in his 2014 article, "How Big Data is Unfair", describes how an *automated* system for fake name detection could suffer from similar disparate impact [10]. Similar bias problems have impacted face tracking software, image labeling, and speech recognition.

I suspect that the effects of many biases are more subtle and personalized, making them difficult to detect by an individual user. At the same time, this personalization makes investigation by researchers in academia, government, or the media extremely difficult. As a result, there is a risk that our information access systems are polluted with biases that are impacting individuals to various degrees.

3 Experimentation

Measurement and experimentation are part of most online services today. Experimentation drives everything from search engine design to news feed organization to page layout.

Let me begin by describing what online experimentation involves. I will focus on classic A/B testing but other methods such as interleaving follow similarly. There are five parts to an online experiment, which follow a similar pattern to machine learned systems described earlier. The first part is to define an objective. This may be user satisfaction, user task completion, corporate revenue, or some other abstract concept. The second part is to define a metric that measures how well a system is doing at achieving its objective. Those of us here will understand metrics like mean average precision and expected reciprocal rank, which are appropriate for Cranfield style experiments. These are meant to capture a user's satisfaction with a ranking. In online experiments, metrics are based on logged user behavior signals. We may, for example, measure user satisfaction by observing user clicks or session-based behavior. The next phase is to consider a set of treatments. These may include subtle changes to certain parameter weights or dramatic layout changes. Having defined a metric and treatments, we conduct the experiment, with each treatment applied to a different partition of users. Often the partitions are random samples from the population of users but we might try more clever sampling methods as well. After the experiment concludes, the metrics are computed, aggregated, and then compared across the treatments. Insights from these experiments are then documented and used for future decision-making.

Measurement and experimentation are part of the DNA of information retrieval so, while the specifics of online experimentation may be obscure to many, the process should be familiar to all of us here.

In as much as human decision-making can introduce bias, the concerns from the previous section apply to experimentation as well. Problematic areas include defining an objective, metric, treatments, partitions, and aggregation. However, conducting online experiments introduces new concerns.

First, experiments, left unreviewed, run the risk of causing harm to our users. Consider an innocuous experiment to test whether users are tolerant to nonrelevant content. If a user is in the middle of an urgent task, such as treating a critical medical condition or evacuating in response to a disaster, being presented inferior results can have real impact on the user's quality of life. Although this example may seem a bit contrived, people do turn to our online services during health or crisis episodes [19, 18]. And there are likely other subpopulations of users who would also be impacted but similar experiments. Measuring risk is difficult because conditions like health crises may be statistically overwhelmed in the data or considered outliers. But this also results from the incomplete representation that systems have of users; a system may not be able to tell that a user is in a vulnerable state.

This brings us to another problem. Even if our abstract objective has the best interest of the user in mind, our metric is only an approximation of that objective. In fact, it probably is quite a crude reflection of the user's satisfaction. As a result, an experiment may need to be very aggressive in order to observe a positive or, more importantly, negative outcome. An anecdote that helps me explain this concern came to me last year. My cat, Millie, woke up one morning in terrible pain. She was meowing constantly and limping. I immediately took her to the veteranarian. Before examining her, the vet informed me that, "Since Millie cannot talk to me, I need to poke her until she tries to bite me. Then I'll know exactly where the problem is." He then proceded to poke her until she, as predicted, tried to bite him. He diagnosed her with a strained back paw and, after a few days of treatment, she was back to normal. So, even though the vet

had Millie's best interest in mind, he required a very invasive experiment to test his hypothesis because he did not have a better signal from the cat. In many ways, experiments in online systems suffer from the same impoverished signals and may need to perform invasive tests to provoke a recognizable response.

Almost all experiments run in online services argue that users provide the best data when unaware that an experiment is being conducted, the so-called Hawthorne effect. As a result, we can imagine the adversarial role of an experimenter intentionally trying to obscure their experiment in order to preserve 'ideal experimental conditions'. Indeed, the public reaction to the Facebook emotional contagion experiment provides another justification for companies hiding experimentation. Even though we, as computer scientists, know it happens constantly, experimentation has become subliminal.

Even if users were aware of experimentation in general, there are issues of informed consent. Users often provide consent in a very long terms of service provision with a vague description of experimentation. Specifics of particular experiments are rarely exposed before, during, or after an experiment. We sacrifice the user's autonomy for experimental flexibility.

All of these issues become even more problematic when we consider automated experimentation systems [16, 7]. Increasingly, bandit style algorithms–or their sequential decision making generalizations–are being adopted to streamline optimization and accelerate the experimental loop. This results in a process that can leave even system designers uninformed about what experiments are being conducted and on who.

4 Remedies

A colleague pointed out to me that pessimism does not help progress. So, I would like to suggest a few possible remedies to the issues. These are no doubt holes in these proposals. And I invite critique.

4.1 Bias

As we described earlier, there are several sources of bias that can be introduced into a production information access system.

Auditing is an approach that should be familiar to information retrieval researchers, especially those involved in measurement and evaluation. Specifically, I want to advocate research into auditing the various stages of data definition, acquisition, and model training. The information retrieval community is *excellent* at detecting and quantifying relevance, user satisfaction, or task completion. Applying this perspective to bias involves asking questions about whether we can detect bias in the target definition, selected features, assessor judgments, sampled instances, or trained models.

Fortunately, there is precedent for this sort of work emerging. From the systems community, Roxana Geambasu has developed methods for externally auditing production systems by using synthetic users to conduct reverse-A/B testing [12]. This technique essentially constructs an artificial population of users with demographic diversity. By controlling the composition of the population, we can ask causal questions about protected attributes–such as race or gender–and system behavior. Similar work has been developed by Amit Datta and colleagues at Carnegie

Melon University [6]. While these techniques focus on detection of bias of a production system, we need better diagnostic methods for determining where bias is entering the system.

Auditing allows us to monitor system biases for manual treatment by, for example adjusting label definitions, feature sets, or sampling. In order to scale this approach, we need to develop learning models that automatically balance fairness and effectiveness. Again, there is some existing work in the machine learning community attempting to accomplish this. For example, Feldman and colleagues have studied ways of repairing the training data to avoid unfair models [9].

4.2 Experimentation

Turning to experimentation, we, fortunately, have a great deal of work in the area from disciplines such as medicine and the social sciences. Despite this, information access systems introduce new problems which may require further dialog. In preparation for the WSDM 2016 workshop on online experimentation, Solon Barocas and I developed the following motivating themes for thinking about online experimentation.

First, experiments in corporate context are different from academic experiments because they often have business interests in mind. Should the ethical questions asked of experiments conducted for scholarly knowledge be different from those asked of experiments conducted for business interest? On the one hand, because we are often not interested in generalizable knowledge, we do not have the same responsibly as academic research. On the other hand, academic style review processes are motivated by a respect for the user. It is not clear to me that, as a discipline, the criteria should be different.

Second, those of you in academia are familiar with review processes for human subjects experiments. Given that we believe there should be *some* review of experiments, how we implement such a process in a company expecting to run hundreds of new experiments each day? Clearly there needs to be some research into scaling review processes.

Third, informed consent is fundamental to preserving our users' sense of autonomy. And we are doing a disservice to our users by hiding consent behind a vague terms of service. There needs to be research here, perhaps from an interface perspective but also perhaps from a model interpretability perspective, focusing on how to communicate potentially complex experiments to our users. In this way, some of the information asymmetry can be resolved. At the same time, though, we want to preserve some of the benefits of existing experimentation methodologies (i.e. naturalistic user interaction). So perhaps we need to think about when informed consent is required for online experiments.

Fourth, from a measurement perspective, we need to have a better understanding about the groups or individuals adversely affects by current experiments or the accumulation of experiments. In this way, we can better diagnose and address systemic over-experimentation.

Finally, our algorithms for automated experiments–that is, bandit or reinforcement learning algorithms–are developing quickly. We need to think deeply about how to quantify risks to users and incorporate them into models. I'm not optimistic since part of what makes ethics important is disagreement and dialog amongst humans. As a result, we may have to think about hybrid systems that combine automated experimentation with human review. Again, issues of scale arise.

4.3 Training

There is a long term solution that recurs throughout discussions with individuals inside or out of our community: learning how to think about and discuss somewhat foreign concepts likes ethics.

First, we can begin by engaging our peers outside of computer science on these issues. Many have thought deeply about these issues in general, in related domains, or in precisely our domain. To this end, there have been several workshops attempting to bring together computer scientists and non-computer scientists. The Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) focuses on many of the issues discussed earlier in the context of discrimination. The Workshop on the Ethics of Online Experimentation, held at WSDM this year, went deep into issues of online experimentation. One thing that has been disappointing so far is the underattendance by computer scientists. A dialog requires two parties and I strongly encourage individuals here to attend.

Second, my colleague Hanna Wallach has written [17],

> ...if technology companies and government organizations...are going to take issues like bias, fairness, and inclusion seriously, they need to hire social scientists–the people with the best training in thinking about important societal issues. Moreover, it is important that this hiring is done not just in a token, "hire one social scientist for every hundred computer scientists" kind of way, but in a serious, "creating interdisciplinary teams" kind of way.

This is one area where researchers in academia can *really* help those of us in industry. You are sitting in the same organization as the exact people we need to address the issues I have raised. There are many days I *wish* I had the opportunity to walk across campus to an entire department of communications, media studies, anthropology, or sociology to discuss questions of fairness or experimentation. Fortunately, within the information retrieval community, we have fantastic peers from the library and information science community, many of whom have studied these topics.

Finally, how should the training of current and future computer scientists and researchers be changed, if at all? Very few of us–especially those with computer science degrees–received any sort of ethics training. Part of this has to do with the fact that we were developing systems relatively distant from humans, be they operating systems kernels, compilers, or index architectures. We learned abstraction and applied it liberally. However, we are building systems that are very close to users these days. And perhaps abstraction is inappropriate when designing for humans. That data point is more than a 100-dimensional vector, it is a human with human rights. If there is one thing that I have learned from interacting with researchers outside of the computer science community, it is to attempt to have empathy for users. That is not an easy exercise when trained in abstraction. But, it is an important process to try to incorporate into existing practices.

5 Conclusion

In closing, I want to return to the fluoroscope and why it disappeared from the retail landscape.

A first clue as to why can be found by reading promotional text for the device. Take for example, the following [1],

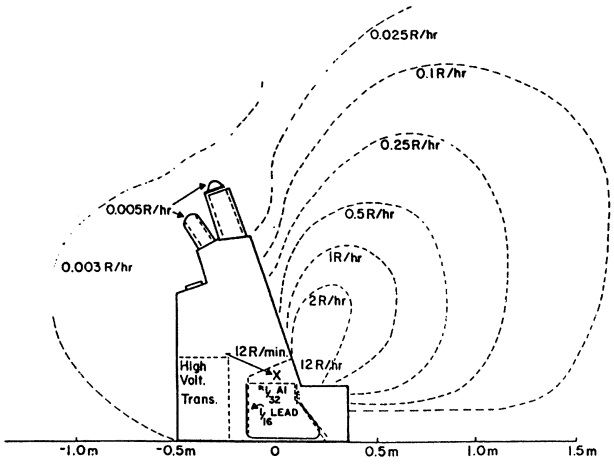

Figure 2: Radiation exposure from the shoe-fitting fluoroscope [4].

> Scientific shoe-fitting at its best. On Dr. Scholl's Fluoroscopic Shoe X-ray you can see the position of the bones in your feet right through the shoe. In addition to this checkup, other methods of scientific shoe fitting will be employed here during this special demonstration.
>
> Dr. Scholl's Shoe-Fitting Experts from the Chicago factory will be in our store. Monday, February 15th. They bring with them the complete line of Dr. Scholl's Shoes (622 fittings) ... every size, width, style–for every type of foot. X-ray fitting–as well as other Dr. Scholl shoe fitting devices. Now you can obtain the shoe that will give you perfect satisfaction–and if you have foot troubles you will be shown how to obtain relief, quickly and inexpensively. Be sure to attend this great display and demonstration ... first of its kind in this city.
>
> George S. Merchant
>
> Winter Garden, Florida

In 1950, thirty years after the invention of the fluoroscope, Leon Lewis and Paul Caplan published an article in the journal California Medicine which measured the intensity of radiation in 40 sample machines [13]. The authors found that, after one minute of usage, the median exposure to the foot was more than one hundred times the weekly accepted exposure to radiation. Even when not directly placing one's foot in the machine, there was a large amount of stray radiation detected as far as six feet away (Figure 2).

The effect of this exposure had impact on several populations. An excerpt from the installation directions suggests the type of customer directly affected by using a fluoroscope,

> Before putting the tube in the X-ray Machine, place the machine in the most desirable location... We would suggest that you center the machine in the store so that it will be equally accessible from any point. Of course, it should face the ladies' and children's departments by virtue of the heavier sales in these departments.

In addition to customers, salespeople were often also exposed to radiation and developed dermatitis.

And while governments began regulating fluoroscopes, a federal ban was never implemented in the US. However, the last fluoroscope was seen in 1970.

When I think about science, I think about this story. Not only because it highlights the fact that science can be wrong. Nor because it has real implications for real people. But rather because it reflects the amount of trust given to scientists and engineers by our users. It is absolutely in our best interest to question ourselves, our research, our methodologies, our products not only from a scientific perspective but from perspectives outside of computer science. This is a rigor that keeps preserves the honesty of our field and the wellbeing of our users.

6 Acknowledgements

I want to thank Omar Alonso and Pavel Serdyukov for the invitation to present at the ECIR 2016 Industry Day. I would also like to thank several researchers who have helped me develop these thoughts and, in most cases, have written more eloquently on similar subjects. These individuals include Solon Barocas, Sarah Bird, danah boyd, Kate Crawford, Sorelle Friedler, Mary Gray, and Hanna Wallach. I also would like to thank Yubin Kim and Alexandra Olteanu for feedback.

References

[1] Shoe-fitting fluoroscope (ca. 1930-1940). `https://www.orau.org/ptp/collection/shoefittingfluor/shoe.htm`, 2010.

[2] Facebook real-name policy controversy. `https://en.wikipedia.org/wiki/Facebook_real-name_policy_controversy`, 2016.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[4] S. C. Bushong and W. D. West. Radiation exposure from a shoe-fitting fluoroscope. *Health physics*, 18(5):575–577, 1970.

[5] Kate Crawford. The hidden biases in big data. `https://hbr.org/2013/04/the-hidden-biases-in-big-data`, April 2013.

[6] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.

[7] Fernando Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.

[8] Fernando Diaz, Ryen W. White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM conference on Information and knowledge management (CIKM 2013)*, 2013.

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[10] Moritz Hardt. How big data is unfair. `https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de`, 2014.

[11] ICT Data and Statistics Division. Itu ict facts and figures – the world in 2015, May 2015.

[12] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 554–566, New York, NY, USA, 2015. ACM.

[13] Leon Lewis and Paul E. Caplan. The shoe-fitting fluoroscope as a radiation hazard. *California Medicine*, 72(1):26–30, January 1950.

[14] Jacob J. Lowe. Method and means for visually determining the fit of footwear, January 1927.

[15] Justin Osofsky and Todd Gage. Community support fyi: Improving the names process on facebook. `http://newsroom.fb.com/news/2015/12/community-support-fyi-improving-the-names-process-on-facebook/`, December 2015.

[16] Filip Radlinski and Thorsten Joachims. Active exploration for learning rankings from clickthrough data. In *KDD*, pages 570–579, 2007.

[17] Hanna Wallach. Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. `https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d#.u7inp4wwx`, December 2014.

[18] Ryen W. White and Eric Horvitz. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *JAMIA*, 21(1):49–55, 2014.

[19] Elad Yom-Tov and Fernando Diaz. Out of sight, not out of mind: on the effect of social and physical detachment on information need. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 385–394, New York, NY, USA, 2011. ACM.