2018 IEEE 34th International Conference on Data Engineering

My Top Ten Fears about the DBMS Field

Michael Stonebraker
CSAIL, Massachussets Institute of Technology
32 Vassar Street, Cambridge, MA
stonebraker@csail.mit.edu

***Abstract*—In this paper, I present my top ten fears about the future of the DBMS field, with apologies to David Letterman. There are three ”big fears”, which I discuss first. Five additional fears are a result of the ”big three”. I then conclude with ”the big enchilada”, which is a pair of fears. In each case, I indicate what I think is the best way to deal with the current situation.**

I. Big Fears About Our Field

A. Major Fear #1: ”The Hollow Middle”

I did a quick survey of the percentage of SIGMOD papers that deal with our field as it was defined in 1977 (storage structures, query processing, security, integrity, query languages, data transformation, data integration, and database design). Today, I would call this the core of our field. Here is the result:

TABLE I
Percentage of papers belonging to core DBMS research as defined in 1977

1977	100%	21/21
1987	93%	40/43
1998	68%	30/44
2008	52%	42/80
2017	47%	42/90

Note that I only include research papers, and not papers from the industrial or demo tracks, when these tracks came into existence. Notice that the core is being ”hollowed out”, as our researchers drift into working on what would have been called applications forty years ago. In my opinion, the reason for this shift is that the historical uses of DBMS technology (OLTP, business data warehouses) are fairly mature. As a result, researchers have largely moved on to other challenges.

However, there is little or no commonality across the various applications areas. What is important in Natural Language Processing (NLP) is totally different from what is important in machine learning (ML) or complex analytics. The net effect is that we have essentially ”multi-furcated” into subgroups that don’t communicate with each other. This is reminiscent of the 1982 bifurcation that occurred when SIGMOD and PODS split.

There is little that binds these various subgroups together other than they must have something to do with data processing (but not necessarily storage). Put differently, our field is defined by what around 200 researchers are working on; which may have little to do with the management of data.

My fear is the obvious one. If there is nothing that binds our community together other than 200 researchers who will look favorably on each others papers, then we will inevitably collapse. In my opinion, the best solution is to recognize this fact, and decompose the major DBMS conferences (SIGMOD, VLDB, ICDE) into multiple (say 5 or 6) separate tracks with independent program committees. (collocated or not). In other words, multi-furcate along the lines of the SIGMOD/PODS division many years ago. If this does not happen, then I strongly suspect that the ”systems folks” will declare a divorce and start their own conference. Other subgroups may follow.

You might ask ”Where is there room for cross cultural research?” I have not seen people from other fields at SIGMOD conferences in quite a while. Equally, program committees do not have such multi-culturalism. The obvious answer is to have more cross cultural conferences. In olden times, we used to have such things, but they have fallen out of favor in recent years.

B. Major Fear #2: We Have Been Abandoned by our Customer

Forty years ago, there was a cadre of ”industry types” who came to our conferences. They were early users (evangelists) of DBMS technology and came from financial services, insurance, petroleum exploration, etc. As such, they provided a handy reality check on whether any given idea had relevance in the real world. In effect, they provided a view of our ”customer”. Hence, our mission was to provide better DBMS support for the broad customer base of DBMS technology, represented by the evangelists.

Over the years, these evangelists have largely disappeared from our conferences. As such, there is no customer facing influence for our field. Instead, there are representatives of the large internet vendors, who have their own priorities, and represent the largest 0.01% of DBMS users, whom I call ”the whales”. Hence, they hardly represent the real world. In effect, our customer has vanished and been replaced by either the whales or by a vacuum.

This loss of our customer has resulted in a collection of bad effects, especially the content of fear #5. In my opinion we desperately need to reconnect with the ”real world”. This could be done by giving free registration to real world users, organizing panels of real users, inviting short problem commentaries from real users, etc.

2375-026X/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDE.2018.00012

C. Major Fear #3: Diarrhea of the Papers

When I graduated from Michigan in 1971 with a PhD, my resume consisted of zero papers. Five years later, I was granted tenure at Berkeley with a resume of half a dozen papers. Others from my age cohort (e.g. Dave Dewitt) report similar numbers. Today, to get a decent job with a fresh PhD, one needs around 10 papers; to get tenure more like 40 is the goal. It is becoming common to take a two-year Postdoc to build up one's resume before hitting the academic job market. Another common tactic these days is to accept an academic job and then delay starting the tenure clock by taking a Postdoc for a year. This was done recently, for example, by Peter Bailis and Leilani Battle. The objective in both situations is to get a head start on the paper deluge required for tenure.

Put differently, there has been an order of magnitude escalation in desired paper production. Add to this fact that there are (say) an order of magnitude more DBMS researchers today than 40 years ago, and we get paper output rising by two orders of magnitude. There is no possible way to cope with this deluge. There are several implications of this escalation.

First, the only way that I would ever read a paper is if somebody else said it was very good or if it was written by one of a handful of researchers that I routinely follow. As a result, we are becoming a ”word of mouth” distribution system. That makes it nearly impossible for somebody from the hinterlands to get well known. In other words, you either work with somebody from the ”in crowd” or you are in ”Outer Siberia”. This makes for an un-level tenure-track playing field.

In addition, everybody divides their ideas into Least Publishable Units (LPUs) to generate the kind of volume that a tenure case requires. Generally, this means there is no seminal paper on a particular idea, just a bunch of LPUs. This makes it difficult to follow the important ideas in the field. It also ups the number of papers researchers must read, which makes us all grumpy.

Lastly, few graduate students are willing to undertake significant implementation projects. If you have to write 10 papers in (say) 4 productive years, that is a paper every 5 months. You cannot afford the time for significant implementations. This results in graduate students being focused on ”quickies”, and tilts paper production toward theory papers. More on this later.

So how did this paper explosion occur? In my opinion it is driven by a collection of Universities mostly in the Far East whose Deans and Department Chairmen are too lazy to actually evaluate the contribution of a particular researcher. Instead they just count papers as a surrogate. This also appears to exist at some second and third rate US and European Universities.

My fear is that this phenomenon will just get worse off into the future. Hence, we should actively put a stop to it, and here is a simple idea. It would be fairly straightforward to get the Department Chairmen of (say) the fifteen best US Universities to adopt the following principle:

Any DBMS applicant for an Assistant Professor position would be required to submit a resume with at most three papers on it. Anybody coming up for tenure, could submit a resume with at most 10 papers. If an applicant submitted a longer resume, it would be sent back to the applicant for pruning. Within a few years, this would radically change publication patterns. Who knows, it might even spread to other disciplines in Computer Science.

II. Other Fears Which Result From the ”Big Three”

What follows is five more fears, which follow directly as consequences of the three already discussed.

A. Fear #4: Reviewing is Getting Very Random

In general, the quality of reviewing stinks. In my experience, about half of the comments from reviewers are way off the mark. Moreover, the variance of reviews is very high. Of course, the problem is that a program committee has about 200 members, so it is a hit-or-miss affair. The biases and predispositions of the various members just increase the variance. Given the ”hollow middle”, the chances of getting three reviewers who are experts in the application domain of any given paper is low, thereby augmenting the variance. Add to this the paper deluge and you get very high volume and low reviewing quality.

So what happens? The average paper is rewritten and resubmitted multiple times. Eventually, it generally gets accepted somewhere. After all, researchers have to publish or perish!

In ancient times, the program chairman read all the papers and exerted at least some uniformity on the reviewing process. Moreover, there were face-to-face program committee meetings where differences were hashed out in front of a collection of peers. This is long gone overrun by the size (some 800 papers) of the reviewing problem. In my opinion, the olden times strategy produced far better results. The obvious helpful step would be to decompose the major DBMS conferences into subconferences as noted in fear #1. Such subconferences would have (say) 75 papers. This would allow ”old school” reviewing and program committees. This subdivision could be adopted easily by putting the current ”area chairman” concept on steroids. These subconferences could be co-located or not; there are pros and cons to both possibilities.

Another solution would be to dramatically change the way paper selection is done. For example, we could simply accept all papers, and make reviews public, along with reviewers' scores. A researcher could then put on his resume his paper and the composite score he received. Papers would get exposure at a conference (long slot, short slot, poster) based on the scores. However, the best solution of all would be to solve fear #3.

If the status quo persists, variance will just increase, resulting in more and more overhead for poorer and poorer results.

B. Fear #5: Research Taste Has Disappeared

Because our community has become disconnected from the real world (Fear #2) and pays attention (at best) to a small collection of internet vendors, we act like lemmings when the next ”silver marketing bullet” from some large internet vendor occurs. Our community has embraced and then rejected (when it became apparent that the idea was terrible) OLEDB, MapReduce, the Semantic WEB, Object Data Bases, XML, and data lakes, just to name a few.

We are very uncritical of systems written by the large internet vendors that solve application specific problems which have been written, until recently, by a development team with little background in DBMSs. As such, they have tended to reinvent the wheel. I am especially amused by Google's adoption and then rejection of MapReduce and eventual consistency.

In my opinion, our community needs to become more assertive at pointing out flawed ideas and ”reinventions of the wheel”. Otherwise, the traditional mantra ”people who do not understand history will be condemned to repeat it” will continue to be true. In addition, we need to reconnect with the ”real world” as noted in fear #2.

C. Fear #6: We are Polishing a Round Ball

With real customers absent (Fear #2) and faced with a required paper production (Fear #3), our field has drifted into solving artificial problems, and especially into making 10% improvements on previous work (so called Least Publishable Units (LPUs). The number of papers at major DBMS conferences that seem completely irrelevant to anything real seems to be increasing over time. Of course, the argument is that it is impossible to decide whether a paper will have impact at some later point in time. However, the number of papers that make a 10% improvement over previous work seems very large. A complex algorithm that makes a 10% improvement over an existing simpler one is just not worth doing. Authors of such papers are just exhibiting poor engineering taste. I have generally felt that we were polishing a round ball for about the last decade. I would posit the following question: ”What was the last paper that made a dramatic contribution to our field?” If you said a paper written in the last 10 years, I would like to know what it is.

A possible strategy would be to require every Assistant Professor to spend a year in industry pre tenure. Nothing generates a reality check better than some time spent in the real world. Of course, implementing this tactic would require a solution to Fear #3. In the current paper climate, it is foolhardy to spend a year not grinding out papers.

D. Fear #7: Irrelevant Theory is Taking Over

Given that our customer has vanished and given the required paper production, the obvious strategy is to grind out ”quickies”. The obvious way to optimize quickies is to include a lot of theory, whether relevant to the problem at hand or not. This has two major benefits. First, it makes for quicker papers, and therefore more volume. Second, it is difficult to get a major conference paper accepted without theorems, lemma and proofs. Hence, this optimizes acceptance probability.

This focus on theory, relevant or not, effectively guarantees that no big ideas will ever be presented at our conference. It also guarantees that no ideas will ever be accepted until they have been polished to be pretty round. My personal experience is that experimental papers are difficult to get by major conference reviewers, mostly because they have no theory. Once we move to polishing a round ball, then the quality of papers is not measured by the quality of the ideas, but by the quality of the theory. To put a moniker on this effect, I call this ”excessive formalism”, which is lemmas and proofs which do nothing to enhance a paper except to give it theoretical standing. Such ”irrelevant theory” essentially guarantees that conference papers will diverge from reality. Effectively, we are moving to papers whose justification has little to nothing to do with solving a real world problem. Because of this attitude, our community has moved from serving a customer (some real world person with a problem) to serving ourselves with interesting math. Of course, this is tied to fear #3; getting tenure is optimized by ”quickies”. I have nothing against theoretical papers, just a problem with irrelevant theory.

Of course, our researchers will assert that it is too difficult to get access to real world problems. In effect, the community has been rendered sterile by the refusal of real enterprises to partner with us in a deep way. The likes of Google, Amazon, Microsoft, et. al. also refuse to share data with our community. In addition, my efforts to obtain data on software faults from a large investment bank were stymied because the bank did not want to make their DBMS vendor look bad, given the frequency of their crashes. I have also been refused access to software evolution code by several large organizations, who apparently have decided that their coding techniques or their code or both were proprietary.

As a result, we deal primarily with artificial benchmarks (such as YCSB) or benchmarks far outside the mainstream (such as Wikipedia). I am particularly reminded of a thread of research that artificially introduced faults into a data set and then proved that the algorithms being presented could find the faults they injected. In my opinion, this proves absolutely nothing about the real world.

Until (and unless) the community finds a way to solve fear #2 and to engage real enterprises in order to get real data on real problems, then we will live in the current theory warp. The wall between real enterprises and the research community will have to come down!

E. Fear #8: We are Ignoring the Hardest Problems

A big problem facing most enterprises is the integration of disparate data sources (data silos). Every large enterprise divides into semi-independent business units, so business agility is enabled. However, this creates independently constructed ”data silos”. It is clearly recognized that silo integration is hugely valuable, for cross selling, social networking, single view of a customer, etc. This is the problem which is the achilles heel of data management, and there is ample evidence of this fact. Data scientists routinely say that they spend at least 80% of their time on data integration, leaving at most 20% for the tasks for which they were hired. Many enterprises report data integration (data curation) is their most difficult problem.

So what is our community doing? There was some work on data integration in the 1980's as well as work on federated data bases over the last 30 years. However, federating data sets is of no value unless they can be cleaned, transformed and deduplicated. In my opinion, insufficient effort has been directed at this problem or at data cleaning, which is equally difficult.

How can we claim to have the research mandate of management of data, if we are ignoring the most important management problem? We have become a community that looks for problems with a clean theoretical foundation that beget mathematical solutions, not one that tries to solve important real world problems. Obviously, this attitude will drive us toward long-term irrelevance.

Of course, this is an obvious result of the necessity of publishing mountains of papers. I.e. don't work on anything hard, whose outcome is not guaranteed to be a paper. It is equally depressing that getting tenure does not stop this paper grind, because your students still need to churn out the required number of papers to get a job. I would advise everybody to take a sabbatical year in industry and delve into data quality or data integration issues. Of course, this is a hollow suggestion, given the current publication requirements on faculty.

Data integration is not the only incredibly important issue facing our users. Evolution of schemas as business conditions change (data base design) is horribly broken, and real users don't follow our traditional wisdom. It is also widely reported that new DBMSs still require some $20M in capital to get to production readiness. For a mature discipline this is appalling. Database applications still require a user to understand way too much about DBMS internals to effectively perform optimization.

In other words, there is no shortage of very important stuff to work on. However, it often does not make for good theory or quickies and often requires massaging a lot of ugly data that is hard to come by. As a community, we need to reset our priorities!

However, these fears pall in comparison to my final pair of fears.

III. The Big Enchilada

Right now most DBMS research occurs in Academia. The attractiveness of universities depends on two major things:

- Money
- Quality of life

The next two fears discuss each in turn. Of course, my perspective is a largely American one.

A. Fear #9: We are Being Starved

The success rate for NSF proposals is down to about 7%. In effect, it is easier to get into an Ivy League school than to receive an NSF grant. At a time when the number of mouths to feed is increasing, grant support is decreasing. As a result, US researchers are being starved by the US Government. The likely effect is the best people will depart for greener pastures, whether abroad or in industry.

In this climate, industry is not picking up the slack. To a first approximation, US enterprises are not funding research, either internally or externally. This presents a double whammy: Government support is declining, and US industry is not picking up the slack. This presents many US universities with a cruel dilemma: Wither away or exist on life support complements of foreign governments and foreign enterprises.

In this climate of tough funding, it is imperative that we make a strong case for our field. In other words, we need to develop a crisp coherent story to tell to the world. Also, we must redouble our efforts to partner with industry. Absent a big crisis, I do not expect the government to pick up the slack.

B. Fear #10: The Quality of Life in US Universities is Declining

Our field exists primarily in a university system, being overrun with students who want to study Computer Science. It is crystal clear that CS is rapidly becoming central to most academic disciplines, and students are voting with their feet. For example, at MIT, 40% of the undergraduates are majoring in Computer Science; if you add in Mechanical Engineering (robotics), the number increases to 50%. There is no indication that this trend is going to slow down anytime soon, if ever.

As a result, the obvious question to ask is "How should Computer Science be organized in universities to achieve maximum student benefit?" Inevitably faculty talk about a School of Computing. However, few universities are currently organized that way. At MIT, for example, CS is the major portion of the EECS department, but there are at least four other units on campus which focus on some portion of CS. In my opinion, universities would be best served by "putting all their wood behind a single arrowhead", to borrow a quote from Scott McNeely, former CEO of Sun Microsystems. In addition, Computer Science needs a LOT more resources (faculty, TAs, graders, etc.) going forward.

My biggest fear is that the historical legacy of organizing computing at many universities will result in bad solutions going forward. In other words, the organization of Computer Science will get caught up in university politics and will not optimize for the needs of the students.

The net result is likely to be a difficult money environment in a world being overrun with students. The obvious solution is to depart for an industry post, where the pay is much higher, the quality of life is good and the customer (the enterprise you work for) has clear DBMS needs. It will be a sad day if the best researchers depart for greener pastures. The long term consequences are particularly dire.

Of course, we have little control over government funding. However, universities with bold administrations can do a lot about quality of life issues. Sadly, I know few such administrators.

IV. Summary

I look out at our field with a hollow middle and increasingly working on applications with little commonality to each other. Restructuring our publication system seems desperately needed. In addition, there is increasing pressure to be risk-adverse and theoretical, so as to grind out the required number of publications. This is an environment of incrementalism, not one that will enable breakthrough research. In my opinion, we are headed in a bad direction.

Most of my fears are rectifiable, given enlightened leadership by the elders of our community. This paper is a plea for action, and quickly! In my opinion, the ”5 year assessment of our field” which is scheduled for the fall of this year should focus primarily on the fears in this paper.