# The First ACAIA Workshop: Approximate Computing for Affordable and Interactive Analytics

Yongjoo Park[1], Srikanth Kandula[3], Surajit Chaudhury[2], and Barzan Mozafari[4]

[1,4]University of Michigan, Ann Arbor [2,3]Microsoft Research

[1,4]{pyongjoo, mozafari}@umich.edu [2,3]{srikanth,surajitc}@microsoft.com

*Workshop Website:* `http://dbgroup.eecs.umich.edu/acaia/`
*San Jose, Nov 9, 2017*

## 1. GOALS

With regards to approximate computing, there continues to be a substantial gap between decades of academic research and the relatively modest adoption in industrial data platforms. Our purpose in organizing the first ACAIA (Approximate Computing for Affordable and Interactive Analytics) was to bridge this gap. By bringing together academic and industrial participants, roughly in a 50-50 ratio, we intended to facilitate a high-bandwidth conversation in both directions with the aim to understand practical use cases better, discuss adoption concerns for approximate analytics especially when users have different statistical capabilities and to determine the practical usefulness of recent research efforts in approximate computing.

We believe that the workshop has served these goals. We expand on this aspect in the rest of this document. The attendees included several large data platforms including some start-ups focused on this topic. We invited keynotes from leading researchers in the space. The agenda included demonstrations of recent advances and descriptions of use-cases from several industrial participants.

Our workshop had the following specific aims:

- Ensuring that attendees gain some basic familiarity with the promise and advantages of approximate computing
- Ensuring that attendees gain some basic familiarity with the limitations of approximate computing
- Ensuring that industry attendees learn about the state-of-the-art in academic research on approximate computing
- Ensuring that academic attendees learn about the practitioners' perspective on approximate computing
- Finding new use cases for approximate computing and approximate query processing
- Forming a diverse community of researchers and practitioners interested in approximate computing

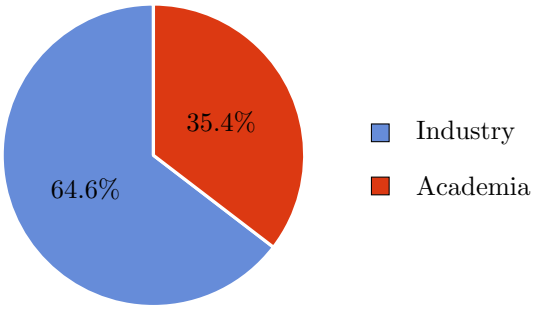



**Figure 1: Participation breakdown: out of the 65 registrations, 42 were from industry and 23 from academia.**

## 2. PARTICIPANTS

The organizing committee invited many individuals, companies, and research groups to attend. There was also a wide call for attendance where interested parties could apply and express how they would benefit from the workshop. The organizing committee reviewed all self-invitation applications and approved nearly all of them. At the end, 65 individuals registered to attend, including students (both undergraduate and graduate), professors, researchers, managers, developers, and practitioners from 39 companies and institutions. Out of the registered individuals, there were 50 participants plus 6 organizers, making a total of 56.

The following is the list of institutions that had at least one registrant: *Accenture, Amazon, Ampool, Inc., Apple Inc., BHGE, Dell / EMC, Dunnhumby, Facebook, Google, Hortonworks, Huawei, Lastline Inc., LendUp, LinkedIn, MemSQL, Microsoft, Netflix, Oracle, Palantir Technologies, PayPal Inc., Penn State University, Relational AI, Reliable Software, Salesforce, San Diego Supercomputer Center, Stanford University, Ohio State University, theScore Media Ventures, U.S. Census Bureau, Uber, UC Berkeley, UC Merced, UIUC, Univer-*

*sity of Michigan, University of Oklahoma, University of Southern California, University of Utah, University of Washington, UPC.*

## 3. WORKSHOP PROGRAM

The workshops program was as follows:

- 8:15 am: Continental Breakfast
- 9:00 am: Welcome and Overview
- 9:05 am: **Keynote 1.** Approximation and Interaction: A Progressive's View, Joseph M. Hellerstein (UC Berkeley)
- 10:00 am: **Tutorial.** AQP Motivations and Examples, Barzan Mozafari (University of Michigan, Ann Arbor)
- 10:30 am: Coffee Break
- 10:45 am: **Tutorial.** AQP Opportunities, limitations and key questions, Surajit Chaudhuri (Microsoft)
- 11:30 am: **Hands-on Session.** SQL Azure / Online AQP, Srikanth Kandula (Microsoft)
- 12:00 am: **Hands-on Session.** Verdict: Platform-independent Approximation, Yongjoo Park (University of Michigan, Ann Arbor)
- 12:30 am: Lunch and Breakout discussions (people in each group sit at the same table and discuss use cases over lunch)
- 2:00 pm: Break-out summary
- 2:50 pm **Keynote 2.** Relax! It's only analytics: Relaxing Consistency and Precision for High-Performance Analytics, Christopher Ré (Stanford)
- 3:45 pm: Coffee Break
- 4:00 pm: **Example Use-cases.** Brian Sullivan (Netflix), Robert Fink (Palantir), (Salesforce), (Google)
- 5:00 pm: Open MIC
- 5:30 pm: Closing Remarks

## 4. DISCUSSION SUMMARY

Industrial participants mostly brought up the chicken-and-egg problem: without approximation support being available in a data platform, it is anybody's guess as to what barriers may stall user adoption barriers but, on the other hand, without a clear case for value it is somewhat unreasonable to expect data platforms to invest the substantial engineering resources needed to make approximation support available.

Other concerns that were discussed included an open discussion on user penchant to accept an approximate answer: (1) scenarios where errors are universally accepted as being vanishingly small are likely to be quickly adopted (e.g., as with the case of using the hyperloglog sketch to estimate COUNT DISTINCT, where prior published works express confidence that errors will be small in many scenarios) and (2) scenarios where user's can reason about and be comfortable with the error model offered by the approximation method.

Academic participants brought up particular use-cases such as data-cleaning and distributed model training where approximation can play a key role (e.g., keynotes from Joe and Chris). There was also substantial reflection regarding the key technical advances that may warrant re-examining the space of approximate analytics (e.g., Florin Rusu questioning).

### 4.1 Group Discussions

Here, we share a summary of the group-level discussions shared at the workshop. The participants were divided up into three groups, each consisting of a diverse group of individuals from both academia and industry.

**Group A** Participants in this group pointed out that one of the biggest challenges for approximate query processing is adoption by users. One way to address this would be help users easily migrate their existing queries into an approximate query processing interface without having to make any changes to the queries themselves, but perhaps only making references to a new abstraction. This group also mentioned several relevant projects from the University of Southern California, such as location sampling for geographic data and time series sampling for healthcare data.

Also, notably, developers from Google mentioned a related technique offered by Google's PowerDrill [13]. Although PowerDrill currently lacks the ability to control the level of approximation, many internal users at Google rely on this feature for dashboards purposes.

**Group B** This group presented several possible use-cases of approximate query processing. First, for domains where immediate answers are crucial (e.g., healthcare), approximation techniques could be beneficial. Also, one group member was involved in an online education platform; one of his main tasks was predicting if students would drop a course in the middle. This application could be considered as a traditional classification task but one where approximate answers are still tolerable.[1]

This group also pointed out several challenges facing approximate query processing. First, approximate query processing typically involves more parameter configurations (e.g., accuracy level), which makes writing queries more complex. Furthermore, for anomaly detection, sampling could be less useful since a random sample may miss those "outlier" data points. In addition, there could be higher chances of a malicious user injecting some skewed information into the sample. This would in turn affect the decision making. Finally, an acceptable accuracy can be very subjective and domain-specific, e.g., 0.1% error may be permissible to one application, while another application might require no

[1]This and similar use case, later motivated the BlinkML project [31].

more than 0.001% error.

**Group C** This group shared some of their own experience in applying approximate query processing for cardinality estimation, a crucial operation in database systems' query planning and optimization. This group also pointed out approximate query processing can offer low-latency data exploration even to people who don't have access to high computing power.

This group pointed out that one of the biggest challenges in adopting approximate query processing is understanding the underlying semantics by non-expert users. One participant posed an interesting question that, it would be great if the system could almost detect the user's intention and automatically identify when they need approximate query processing.

This group included developers from Oracle, who shared their own experience regarding approximate query processing: they added sketch-based techniques (e.g., approximate counting, approximate percentile, and approximate top-N) and observed that with these techniques, query latencies go down from 50 minutes to 6 minutes.

## 4.2 Company-specific Usecases

Several companies provided a more detailed usecase, which we provide a summary of below.

**Netflix Case Study** An invited speaker from Netflix, Brian Sullivan, shared their use of approximate query processing in quickly estimating the median play delays for various platforms (e.g., Comcast, mobile devices). Unlike simple aggregations (e.g., sum and average), one of the difficulties in estimating median (and percentiles in general) is that a nice mathematical property—called associativity—no longer holders for median, implying that we cannot simply compute the median of Group 1 + Group 2 even if we know their respective medians, i.e., the median of Group 1 and the median of Group 2.

To address this challenge, Netflix has started to look at sketching-based methods, such as *t-digest*. t-digest offers approximate percentiles (and median, as a special case). The nice property of this sketching-based method is that it satisfies associativity, enables Netflix to precompute medians in a space-efficient manner.

To build this new analytics platform, Netflix has combined various open-source platforms. The main reason behind this choice is the flexibility in extending existing systems. They first build sketches using big data computing platform, such as Apache Spark, and ship those constructed sketches into ElasticSearch and Druid. For query processing, they use both ElasticSearch and Druid, depending on the nature of the usecase at hand.

**Salesforce Case Study** There was also an invited speaker from Salesforce, Vijay Devadhar, who shared their use of approximate query processing. Their usecases were quite unique: they use approximation for managing resources and systems. In particular, Salesforce has a multi-tenant computing environment, where it is crucial to collect tenant-level statistics on a weekly basis. They then conduct A/B testing for query planning in order to adapt their caching and traffic sharding strategies.

# 5. OPEN CHALLENGES

Several open challenges were brought up during the workshop; while not per se novel, we mention a few of these challenges to emphasize their value. End-users want a simple usage model: an approximation system that takes as input some user-specified error metric (say confidence intervals on each aggregate) and computes such an answer in the most performant way. It remains an open problem to answer this question for complex queries (e.g., TPC-DS queries); although some research since the workshop has expanded the space of queries for which this use-case is possible [30].

## 5.1 Vendor Resistance

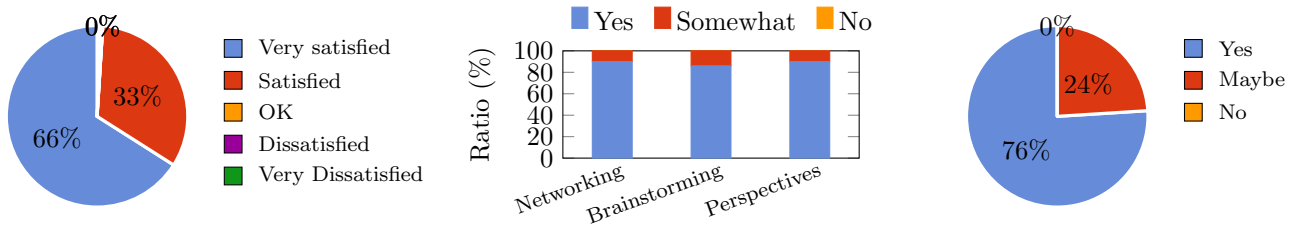
Since its inception [35], there has been much work on approximate query processing (AQP) in academic research [7, 6, 2, 3, 24, 34, 4, 10, 16, 18, 8, 20, 33, 9, 23, 36, 8, 11, 25, 5, 28, 29, 27, 32, 14, 17]. Despite this long history, however, AQP has had little success in terms of industrial adoption [21]. One of the key obstacles is the reluctance of major vendors in adding new technologies to their legacy codebases. For example, it took nearly a decade before column stores were adopted by mainstream vendors for analytical workloads. However, this reluctance is even more severe for approximate query processing. Unlike columnar formats which only affected internal implementation but left the user interface intact (i.e., the familiar SQL semantics), AQP requires modifications to both: incorporate approximation features means changes to both database internals as well as output semantics.

A common set of modifications that need be made to database internals include:

1. **Error Estimation:** BlinkDB [4], G-OLA [37], Quickr [18]
2. **Query Evaluation:** Online Aggregation [37, 12, 26, 15], Join Synopses [2, 1, 3], Dynamic Sample Selection [7, 6]
3. **New Relational Operations:** Analytical Bootstrap [38], Wander Join [19]

## 5.2 Incompatibility with BI Tools

Now, with AQP, users must cope with the uncertainty in their query output, such as errors and probability of existence. However, existing BI tools are *not* designed




**(a) Overall satisfaction** **(b) Useful aspects** **(c) Will you attend again?**

**Figure 2: Summary of the post-workshop survey.**

to support these types of uncertainty; that is, existing BI tools do not translate these different notions of errors into the form that can easily be understood by end users (e.g., confidence intervals, error bars, jiggle animations, etc.).

### 5.3 Offline Provisioning vs. Generality

Offline AQP solutions build samples *a priori*, often based on past workload or based on common columnsets [22]. As such, offline AQP engines offer one or two orders of magnitude speedup. However, this is contingent upon having built the *right* set of samples and structures in advance. Online sampling approaches do not this limitations. However, they only offer modest speedups compared to offline engines.

## 6. POST-WORKSHOP FEEDBACK

After the workshop, an online survey was sent to all participants. In total, 21 individuals participated in the survey. Below is the summary of the main questions on the post-workshop survey:

1. Please rate your overall satisfaction with the overall format of the workshop: (Very satisfied: 66%, Satisfied: 33%, OK: 0%, Dissatisfied: 0%, Very Dissatisfied: 0%)
2. Did you find the workshop useful in enabling the following activities?
   - Networking (Yes: 90%, Somewhat: 10%, No: 0%)
   - Brainstorming Ideas (Yes: 86%, Somewhat: 14%, No: 0%)
   - Hearing different perspectives (Yes: 90%, Somewhat: 10%, No: 0%)
3. How useful was this workshop to you?
   - I am now much more familiar with approximate analytics and query processing (Yes: 100%, Somehow: 0%, No: 0%)
   - I will consider using approximate query processing at my own team/company (Yes: 76%, Somehow: 24%, No: 0%)
4. If ACAIA were to be offered next year,
   - Would you attend again? (Yes: 76%, Maybe: 24%, No: 0%)
   - Would you or your company be interested in getting involved in the organization of ACAIA? (Yes: 38%, Maybe: 24%, No: 38%)

## 7. CONCLUSION AND NEXT STEPS

As a first-time experience, ACAIA was extremely encouraging, both quantitatively (participation numbers) and qualitatively (discussions and post-workshop survey). In addition to raising awareness and taking the initial steps towards building a home community for approximate query processing, the various forms of feedback and the post-workshop survey have been extremely encouraging.

From a logistical perspective, the most challenging aspect of the ACAIA workshop was reaching the target audience and inviting interested parties across the industry. The community is more easily identified and reached, whereas identifying companies, teams and individuals from engineering teams who might have an interest in approximate technology proved to be the most time-consuming aspect of the workshop organization.

One particular idea that came out of the discussions at ACAIA was to bring several major vendors together to form an AQP consortium. The purpose would be to create a unified body to discuss and oversee AQP standards, language extensions, and even joint educational efforts. Another suggestion was to also to have the ACAIA workshop be overseen by this consortium.

Another joint decision after the workshop was to conduct future ACAIA workshops in conjunction with a major database conference, such as VLDB or ACM SIGMOD. Such an arrangement would significantly simplify the logistics, lower the organizational burden, and reduce the monetary costs. The downsides of this arrangement, however, would be two-fold. First, these conferences are overwhelmingly attended by academics, and may defeat the purpose of ACAIA which is to bring academics and practitioners. Second, the location of the first ACAIA was intentionally chosen to be in Silicon Valley (San Jose, CA), in an attempt to maximize industry participation. If held jointly with VLDB or SIGMOD, we will not have any control over the physical location of the workshop.

Creating a call for papers for the subsequent ACAIA workshops would be another means of increased participation (perhaps, more so from academia). A small program committee would need to be formed to peer-review the publications.

Finally, the unanimous conclusions seemed to be the dire need for a more transparent discussion of AQP usecases. Numerous companies seem to be resorting to various forms of approximation (for both analytical and operational purposes), but are less open to sharing their usecases, requirements and experiences with the rest of the community.

## 8. ACKNOWLEDGEMENTS

The ACAIA workshop was funded by National Science Foundation, through award number 1748047, under a grant entitled *Bridging the Gap between Academia and Industry: Workshop on Approximate Computing.* This workshop would not have been possible without the generous support and guidance of Nan Zhang, as well as the administrative support from the department of computer science and engineer at the University of Michigan, Ann Arbor. The organizers are also grateful to many individuals who participated in the workshop, contributed to lively discussions, and provided valuable feedback and suggestions.

## References

S. Acharya, P. B. Gibbons, and V. Poosala. Aqua: A fast decision support system using approximate query answers. In *VLDB*, 1999.

S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy. Join synopses for approximate query answering. In *SIGMOD*, 1999.

S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy. The Aqua Approximate Query Answering System. In *SIGMOD*, 1999.

S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. BlinkDB: queries with bounded errors and bounded response times on very large data. In *EuroSys*, 2013.

A. Arasu and G. S. Manku. Approximate counts and quantiles over sliding windows. In *PODS*, 2004.

B. Babcock, S. Chaudhuri, and G. Das. Dynamic sample selection for approximate query processing. In *VLDB*, 2003.

S. Chaudhuri, G. Das, and V. Narasayya. Optimized stratified sampling for approximate query processing. *TODS*, 2007.

J. Considine, F. Li, G. Kollios, and J. Byers. Approximate aggregation techniques for sensor databases. In *ICDE*, 2004.

W. Fan, F. Geerts, Y. Cao, T. Deng, and P. Lu. Querying big data by accessing small data. In *PODS*, 2015.

V. Ganti, M.-L. Lee, and R. Ramakrishnan. Icicles: Self-tuning samples for approximate query answering. In *VLDB*, 2000.

W. Gatterbauer and D. Suciu. Approximate lifted inference with probabilistic databases. *PVLDB*, 2015.

P. J. Haas and J. M. Hellerstein. Ripple Joins for Online Aggregation. In *SIGMOD*, pages 287–298, 1999.

A. Hall, O. Bachmann, R. Büssow, S. Gănceanu, and M. Nunkesser. Processing a trillion cells per mouse click. *PVLDB*, 2012.

W. He, Y. Park, I. Hanafi, J. Yatvitskiy, and B. Mozafari. Demonstration of VerdictDB, the platform-independent AQP system. In *SIGMOD*, 2018.

J. M. Hellerstein, P. J. Haas, and H. J. Wang. Online aggregation. In *SIGMOD*, 1997.

K. Hose, D. Klan, and K.-U. Sattler. Distributed data summaries for approximate query processing in pdms. In *IDEAS*, 2006.

N. Kamat, P. Jayachandran, K. Tunga, and A. Nandi. Distributed and interactive cube exploration. In *ICDE*, 2014.

S. Kandula, A. Shanbhag, A. Vitorovic, M. Olma, R. Grandl, S. Chaudhuri, and B. Ding. Quickr: Lazily approximating complex adhoc queries in bigdata clusters. In *SIGMOD*, 2016.

F. Li, B. Wu, K. Yi, and Z. Zhao. Wander join: Online aggregation via random walks. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, 2016.

A. Meliou, C. Guestrin, and J. M. Hellerstein. Approximating sensor network queries using in-network summaries. In *IPSN*, 2009.

B. Mozafari. Approximate query engines: Commercial challenges and research opportunities. In *SIGMOD*, 2017.

B. Mozafari and N. Niu. A handbook for building an approximate query engine. *IEEE Data Eng. Bull.*, 2015.

B. Mozafari and C. Zaniolo. Optimal load shedding with aggregates and mining queries. In *ICDE*, 2010.

C. Olston, E. Bortnikov, K. Elmeleegy, F. Junqueira, and B. Reed. Interactive Analysis of Web-Scale Data. In *CIDR*, 2009.

D. Olteanu, J. Huang, and C. Koch. Approximate confidence computation in probabilistic databases. In *ICDE*, 2010.

N. Pansare, V. R. Borkar, C. Jermaine, and T. Condie. Online aggregation for large mapreduce jobs. *PVLDB*, 4, 2011.

Y. Park. Active database learning. In *CIDR*, 2017.

Y. Park, M. Cafarella, and B. Mozafari. Neighbor-sensitive hashing. *PVLDB*, 2015.

Y. Park, M. Cafarella, and B. Mozafari. Visualization-aware sampling for very large databases. *ICDE*, 2016.

Y. Park, B. Mozafari, J. Sorenson, and J. Wang. Verdictdb: universalizing approximate query processing. In *SIGMOD*, 2018.

Y. Park, J. Qing, X. Shen, and B. Mozafari. Blinkml: Efficient maximum likelihood estimation with probabilistic guarantees. In *SIGMOD*, 2019.

Y. Park, A. S. Tajik, M. Cafarella, and B. Mozafari. Database Learning: Towards a database that becomes smarter every time. In *SIGMOD*, 2017.

N. Potti and J. M. Patel. Daq: a new paradigm for approximate query processing. *PVLDB*, 2015.

L. Sidirourgos, M. L. Kersten, and P. A. Boncz. SciBORQ: Scientific data management with Bounds On Runtime and Quality. In *CIDR*, 2011.

S. Vrbsky, K. Smith, and J. Liu. An object-oriented semantic data model to support approximate query processing. In *Proceedings of IFIP TC2 Working Conference on Object-Oriented Database Semantics*, 1990.

F. Xu, C. Jermaine, and A. Dobra. Confidence bounds for sampling-based group by estimates. *TODS*, 2008.

K. Zeng, S. Agarwal, A. Dave, M. Armbrust, and I. Stoica. G-OLA: Generalized on-line aggregation for interactive analysis on big data. In *SIGMOD*, 2015.

K. Zeng, S. Gao, B. Mozafari, and C. Zaniolo. The analytical bootstrap: a new method for fast error estimation in approximate query processing. In *SIGMOD*, 2014.