Professional Statement: Tim Kraska

Data has gone from scarce to superabundant. Analyzing large amounts of data is no longer just a problem for a few research projects or giant Internet companies but concerns almost everyone. The massive increase in data volume brings huge new benefits and enables many new applications. At the same time, it also requires that we rethink the way we handle and analyze data. With the increasing number of people interested in data science, it is unrealistic to assume that all of them have both, deep domain knowledge and expertise in machine learning/statistical inference, data management, visualization and many other related fields. Moreover, with the end of Moore's law, the processing needs to keep up with the ever increasing amount of data has become a key challenge. Addressing these challenges requires throwing away old preconceptions about data management and analytics, and breaking down many of the traditional intra- and interdisciplinary boundaries in computer science.

My research aims to dramatically increase the efficiency of data-intensive systems and democratize data science by revisiting the way traditional data-intensive systems have been built, whether it is the interface or the algorithms and data structures the systems use. I strive to develop solutions that are both practical and principled, drawing on ideas from database systems, human-computer interaction, machine learning, networking, and statistics.

Over the past eight years, I have worked on two main research thrusts, SageDB at MIT and Northstar at Brown University. Both take an interdisciplinary approach to either significantly increase the efficiency of data processing systems (SageDB) or to make Data Science more accessible to a broader range of users (Northstar). These projects are summarized briefly below and discussed in greater detail in the next section.

- **Instance-optimized Systems (SageDB):** Database systems have a long history of carefully selecting efficient algorithms, e.g., a merge vs a hash-join, based on data statistics. Yet, existing databases remain general-purpose systems and are usually not purpose-built for a specific workload. As part of the SageDB project we are exploring how we can leverage machine learning (ML) and optimization methods to automatically optimize database components. For example, with learned indexes [1] we showed that we can use ML to create much more compact and faster indexes than traditional B-Trees when the data follows certain patterns. We later extended that work to multi-dimensional indexes [2, 3], DNA sequence search [4], and even sorting [5]. Further, with Neo [6] and BAO [7], which recently received the best paper award at SIGMOD, we took a learning-based approach to query optimization. Our results have driven the trend towards using machine learning to enhance systems [8, 9, 10], inspired significant follow-on work in and outside of the database community [11, 12, 13, 14, 15], and led to several companies aiming to integrate the ideas into their systems [16, 17].
- **Interactive Data Science (Northstar):** The interfaces people use to analyze data have not fundamentally changed over the past two decades. While these interfaces served generations of analysts well, they severely limit how teams can work together and still pose a significant barrier to entry for users without a computer science background. With Northstar, we developed an interactive data science platform from the ground up for teams and so-called Citizen Data Scientists, starting with the user experience rather than the system design. Interestingly, putting the user experience first required us to rethink the entire analytics stack and many of the involved algorithms. Key contributions of Northstar include a new visual data exploration paradigm [18, 19], new data structures/algorithms to better support interactive data exploration [20, 21, 22, 23, 24, 25, 26, 27], novel low-latency AutoML tools [28, 29, 30, 31], and techniques to protect the user from common mistakes [32, 33, 34, 35, 36]. Northstar is now commercialized by Einblick Analytics and used by several companies (e.g., BMW, Covestro) and government agencies (e.g., DARPA/JAIC and USDAA).

Thanks to my outstanding students, post-docs, and numerous collaborators from academia and industry, these projects led to 33 full research papers in top-tier conferences and journals (14xSIGMOD, 8xVLDB, 3xICDE, 3xCHI, 1xNeurIPS, 1xICLR, 1xKDD, 1xPoPP, 1xMobiSys), one best paper award, and two honorable mentions/best of citations since my start at MIT in 2018, and a total of 122 publications since my start as an assistant professor at Brown.

Instance-optimized Systems

Modern data-processing systems are designed to be general purpose, able to handle a wide variety of different schemas, data types, and workloads. This general-purpose nature results in systems that do not take advantage of a user's particular application and data, which often causes significant performance loss. With the SageDB project [37, 38], we are exploring a new way to engineer data-processing systems, where learning and program synthesis are used to specialize

components of the database for a given workload and data distribution. The goal is to build systems, that are able to support a wide range of workloads while providing the performance similar to what we would achieve by building and (hand-)tuning the entire system for a single application. As a first step towards this goal, we explored how different components, algorithms, and data structures of a database system could be optimized through machine learning.

In our SIGMOD 2018 paper, "The Case for Learned Index Structures" [1], we showed that traditional algorithms and data structures, particularly B-Trees, Hash-Maps, and Bloom-Filters, can be enhanced and sometimes entirely replaced by learned models with significant space and performance benefits, while providing the exact same semantic guarantees. To make this possible, we invented a new ML model, called RMI, with nano-second inference time and ways to combine probabilistic models with traditional data structures and algorithms. The result received substantial attention on (social) media (e.g., [8, 9, 10, 39, 40, 41, 42]) and led to significant follow-on work in both the database community and broader CS community [11, 38], as it provided a new perspective on how to optimize many fundamental algorithms and data structures in CS. At the same time, it also raised a lot of controversy because of its unintuitive result that machine learning can improve data structures with provable (sub-)linear complexity. Most notably, three individual blog posts by Prof. Neumann, Prof. Mitzenmacher, and a group at Stanford raised the question of whether other more traditional approaches can outperform Learned Range Indexes, Bloom-Filters, or Hash-Maps. Since the blogs were published, we joined forces with Prof. Neumann and his team to perform a joint evaluation, that showed that Learned Indexes are indeed much smaller while providing better or similar performance than traditional approaches [43, 44]. Subsequently, Ferragina et al. theoretically proved why learned indexes are more space efficient [45]. We and others showed that the advantage of a learned index on real-world systems is mainly due to its smaller size rather than its lower latency [16, 46]. We also joined forces with Prof. Mitzenmacher and recently published a new approach to Learned Bloom Filters called Partitioned Learned Bloom Filters (PLBF) [47]. PLBF can be regarded as a generalization with provable guarantees of Sandwiched Learned Bloom Filter (SBF) [48], an advancement by Mitzenmacher on our original work [1], and Adaptive Learned Bloom Filter (Ada-BF) [49], another follow-on work on [1] by Dai and Shrivastava. We also theoretically and experimentally evaluated when (RMI) models can be better hash functions for Hash-Maps [50] and developed new Learned Index Models with error guarantees [51, 52].

Later, in the Flood project [2], we demonstrated how we can extend the idea of learned indexes to multi-dimensional indexes, which outperform alternative general-purpose data structures by orders of magnitude by automatically adjusting to the data distribution and workload. With Tsunami [3], we extended the Flood technique to create a first self-optimizing in-memory storage manager, which also takes advantage of correlation within the data. Together with Microsoft, we developed methods to automatically "instance-optimize" storage layouts for a given workload, considering not only single table accesses but also joins [53]. The Learned Index paper also inspired us to investigate other related problems. For example, in [5, 54] we created an ML-enhanced sorting algorithm, which outperforms the best sorting algorithms we were able to find by up to 30% in sorting throughput even when including the model training time.

In addition to developing fundamental ML-enhanced algorithms and data structures, my research group also worked on instance-optimizing higher-level components of a database management system [6, 7, 55, 56, 57]. For example, with NEO [6] we developed the first end-to-end learning-based query optimizer. Query optimization is a crucial component of any database system to achieving good performance. The key contributions of the paper are our model representation and our demonstration that a learned query optimizer can outperform traditional optimizers of open-source and commercial systems with much less development time. However, NEO can still take days to train and, similar to most deep learning approaches, NEO's decisions are hard for a human to understand. To address these shortcomings, we developed BAO [7], which learns to improve a traditional optimizer rather than replacing it entirely. In [7, 58, 17], we show how BAO can outperform the query optimizers of open-source and commercial row and column stores after a much shorter training time than NEO, and how it can be used to make automatic decisions as well as to function as an advisor for a database administrator.

Impact These projects are among the earliest successes in the area of learning-based database systems and ML-enhanced algorithms and data structures. Our techniques have spurred considerable interest from industry and follow-up research within and outside the database community (see also [38]). For example, Google integrated Learned Index structures into their BigTable system, improving the throughput by up to 2x [16], and research at the University of Wisconsin showed how Learned Indexes can improve "lookup performance by 1.23x-1.78x as compared to state-of-the-art production LSMs" [46]. To date, we have counted over 50 new variants of learned

indexes (see [11] for a list of papers) and Google lists our 2018 Learned Index paper as the most cited SIGMOD paper over the last 5 years [59]. The idea of Learned Index was also applied to other areas, including network package classification [60], DNA sequence search [4], longest prefix matches [61], and has been adopted for various modern hardware configurations such as multi-core environments [62], NVM [63], and RDMA [64]. Follow-on research on Learned Index Structures appeared in database (e.g., VLDB[65]), systems (e.g., OSDI [64]), machine learning (e.g., NeurIPS [49]), networking (e.g., SIGCOMM [60]), and theory (e.g., Theor. Comput. Sci. [66]) conferences. Researchers at Purdue even created an entire tutorial and taxonomy just on learned indexes [13, 14].

Similarly, NEO and BAO received a lot of attention from industry and academia. We collaborated with Microsoft to evaluate the benefits of BAO for Big Data workloads, showing up to 90% runtime latency savings for complex queries [17]. We are aware of at least two cloud database providers who are working on integrating BAO into their production systems. BAO also received the best paper award at SIGMOD'21 and our evaluation with Microsoft [17] a "Best Industrial Paper Honorable Mention."

Another indicator of the impact of our work is the fact that I have been invited to give nine keynotes on our ML for Systems work since 2019, including keynotes for O'Reilly AI and VLDB'21 [38], the largest database conference.

Interactive Data Science: Northstar

Northstar [18] is a visual interactive data science platform. It enables domain experts and data scientists to work together during a single meeting to visualize, transform, and analyze even the most complex data on the spot. In contrast to other systems, Northstar tries to always return a result in seconds, regardless of the operation complexity and data size, in order to foster interaction and collaboration. This new interactive mode of operation required us to rethink the full analytics stack, from the interface to the low level system internals.

Interface: Current analytics frameworks still focus on a text-based scripting interface, making it hard for domain experts and data scientist to collaborate in real time. While systems like Tableau make a step in the right direction, they still do not support the creation of sophisticated data pipelines. With Vizdom [19], Northstar's user interface for Interactive Data Science, we aimed to create a new interaction paradigm to enable true in-person and remote collaboration. Vizdom diverges from standard dashboards and workflow engines, and provides immediate feedback using progressive computation for any operation while also allowing users to compose complex analytic workflows. As we showed in a systematic user study [67], even a few seconds' delay can negatively impact the data exploration process, while progressive and approximate computation allows the user to stay immersed in the activity.

Backend: Surprisingly, today's analytics frameworks are ill-suited to support interactive visual frontends, even for simple operations over small data sets. Current frameworks like Spark are designed to process massive data sets distributed across huge clusters and scheduling a single job can already take longer than any reasonable interactivity latency threshold. With Tupleware [22], we explored the design of a new framework intended to provide interactive latencies for analytic database queries. Two key contributions of Tupleware are the close-to-zero execution overhead, and new query compilation optimization techniques [20, 21, 22, 23]. The latter fundamentally bridges the gap between query optimizers, which usually make high-level optimization decisions (e.g., picking a join algorithm), and compilers, which make low-level decisions (e.g., loop-unrolling). More recently, we extended the ideas of Tupleware to more efficiently compile Python code with error-handling as part of the Tuplex project [26, 27].

Even the fastest run-time can be too slow to guarantee interactive response times over large data sets. Approximate query processing (AQP) techniques can help in those situations but, as we showed in [68], require a significant amount of preprocessing time, which prevents ad hoc analysis. We therefore built Northstar's Interactive Data Exploration Accelerator [24, 69] and with it investigated new AQP techniques designed for visual interactive data science. The key contributions of this work are (1) new sample management techniques [24, 69], (2) VisTrees [70] to better support the access patterns created by visual front-ends, (3) a novel low-overhead query result re-use technique [25], and (4) a new AQP processing model and optimizations for interactive data exploration with visual interfaces [71].

Smart Assistants: A visual interactive data exploration system alone does not make data science readily accessible for domain experts without a deep technical background in CS, statistics/ML, etc. Rather, we need to provide a lot of (automatic) assistants to help the domain experts to achieve the task. We therefore developed a series of techniques to help users to (1) build models [28, 29, 30, 31, 72], (2) find interesting and significant insights [73, 74, 75, 76, 77], (3) deal with data quality issues [78, 79, 80, 81, 82, 83, 84, 85], (4) create labels [86], and

(5) prevent common mistakes [32, 33, 34, 35, 36]. The following highlights some of these results.

ML Assistant: One key promise of Northstar is to help users to quickly arrive at an initial solution, often a model, in a collaborative setting. Hence, we designed the first interactive AutoML tool, called Alpine Meadow [31], which returns a first high-quality model in just a few seconds, which it then continuously refines in the background. Alpine Meadow improves upon our previous AutoML work MLbase [29] and TUPAQ [30] by combining transfer learning with a novel sampling and pruning technique in order to achieve interactive model training times. Since 2017, Alpine Meadow has dominated the DARPA D3M AutoML competition (see [31] for more details) before Northstar was spun out as a company. More recently, we expanded Alpine Meadow to automatically find and join related datasets to increase the prediction quality of a model [72], which received a "Best of VLDB" invite from TODS.

Data Quality Assistant: Working with our industry partners, we found (unsurprisingly) that data scientists commonly integrate disparate data sources into one data set, and that the integrated data is almost never complete or 100% clean. As a first step to address this problem, we aimed to estimate the impact of the missing data (a.k.a. unknown unknowns) on aggregate query results. While sounding impossible, our previous work on "Crowdsourced enumeration queries" [87], which received the ICDE best paper award and was listed in ACM's Computing Review "Best of Computing," laid the foundation for it by proposing techniques to estimate the completeness of a set. Building on these results, we developed a novel bucket estimator, capable of estimating not only the completeness of a data set, but also the potential impact of the missing data under loose conditions. The result of this work was published at SIGMOD [80] and received a TODS' Best of SIGMOD 2015 invite [83]. Together with Brown's Center for Evidence Synthesis in Health we then explored how these techniques can lower the cost to create systematic reviews (published in a medical journal [88]). Similarly, we created techniques to estimate the number of remaining data errors within a data set [79] and how sampling [78, 82] or crowd-sourcing [89, 85] can be used to more (cost-)efficiently clean a data set. In addition we also developed assistants to recommend visualizations [90, 76], type detection [75], and insight recommendation [73, 74], and worked on techniques to automatically protect users from the multiple hypothesis pitfall [32, 33, 34, 76] and the Simpson paradox [36].

Impact: Northstar was featured several times in the media (e.g., Techcrunch [91] and Science [92]) and we received several awards for the work including the best demo award at VLDB 2015, a TODS' Best of SIGMOD 2015 and, a Best of VLDB 2020 invite for papers out of the Northstar project. Our research results influenced many follow-up works, and the cumulative effect was that I received the VLDB Early Career Contribution award for "advancing systems research on interactive data analytics" and was invited to give a keynote at SIGMOD in 2017 on the Northstar system [93]. After multiple demo deployments in industry, Northstar is currently being commercialized by Einblick Analytics, Inc., as a venture-backed MIT/Brown spin-off, and used by large enterprises (e.g., BMW and Covestro) and government agencies (e.g., DARPA/JAIC, USDAA, the Ethiopian government/Gates Foundation).

Other projects

NAM-DB: Traditional distributed database systems are designed assuming that the network is the bottleneck and thus must be avoided as much as possible. However, we observed that with the next generation of networks this assumption is no longer true [94]. We therefore suggested a new database design, called NAM-DB [94], which is able to take full advantage of high-bandwidth networks. We further showed that, contrary to common wisdom, distributed transactions can be made scalable [95], and developed new partitioning [96] and replication protocols [97] to take full advantage of next generation networks. This work was done in close collaboration with Oracle and has wide reaching implications for distributed (DB) system design. We received an "ACM SIGMOD Research Highlight Award" for our partition paper [96] and Erfan Zamanian received the "Honorable Mention for the Jim Gray Dissertation Award" for his work on NAM-DB.

Crowd sourcing: Before starting as an Assistant Professor, I worked on CrowdDB [98, 99], a hybrid human-machine database system that automatically uses crowdsourcing to integrate human input for processing queries that neither database systems nor search engines can adequately answer (e.g., certain cleaning or data collection tasks). CrowdDB won the best demo award at VLDB'11 [99] and I gave the first crowdsourcing tutorial [100] for DB researchers at VLDB in 2011. Today, most database conferences feature a crowd-sourcing track.

I further worked on protocols for geo-replication as part of the MDCC project [101, 102], and was the first to explore a cloud database architecture that separates compute and storage [103, 104, 105], which now is the industry standard for most cloud database systems.

Teaching and Service

In 2014, I created the first "Data Science" course at Brown University to train students in the interdisciplinary field of data science. I designed the course to fill a gap in the course offerings at Brown as none of the existing courses taught how the different disciplines involved in data science (visualizations, statistics/machine learning, databases, etc.) actually worked together. One challenge with the course was its increasing class size: 45 in 2014, 120 in 2015, and over 200 initially registered students in 2016 (the class was capped by Brown). To keep the class still interactive, I started to use iClickers, introduced reversed classrooms, and did conference-style poster session for the final project presentations.

More recently, I designed 6.S080 "Software Systems for Data Science" at MIT. Similar to the "Data Science" course at Brown, the course offers an integrated perspective on how data cleaning, data integration, scalable systems, fundamental statistics, machine learning, and scalable visualization have to come together to create data products. I helped craft the "Data Science Education" report as part of a workshop organized by the NSF in Washington DC, and participated in several Data Science Education panels [106, 107]. As a member of Brown's Data Science Initiative, I significantly contributed to the creation of Brown's Master of Data Science curriculum and taught a new variant of my Data Science course as part of the master's program in fall 2017. At MIT I taught 6.814/6.830 "Database Systems" as an instructor and 6.033 "Computer Systems Engineering" as a recitation instructor. As part of 6.033, I helped to develop new active learning techniques, such as the database Jeopardy game, that is now used by all recitation instructors. For the fall 2021, together with Mohammad Alizadeh and in collaboration with Microsoft, I am planning to offer a new course on "Machine Learning for Systems."

However, my real passion is advising students and helping them to become professors, researchers, entrepreneurs, or leaders in industry. I am proud that my first five students (one co-advised with Andy Van Dam at Brown) all successfully graduated with a strong record. For example, Erfan Zamanian received an honorable mention for the Jim Gray dissertation award [108]. All my current students form a very strong team and I am proud of their academic achievements, which have received external recognition in the form of various fellowships (e.g., Andrew Crotty received the Google PhD Fellowship, and Jialin Ding and Leonhard Spiegelberg the Facebook PhD Fellowship). Without them, it would be impossible for our ML for Systems research agenda to have the academic success, visibility, and impact it is currently enjoying.

Service

After starting at MIT, I co-founded the MIT Data Systems and AI Lab (DSAIL), an industry sponsored lab that I now lead together with Sam Madden. The lab consists of eight faculty members and is focused on exploring how to enhance systems using ML and build systems for ML. As part of the lab, I organize yearly retreats with close to 100 participants and helped establish several successful research collaborations, which lead to numerous joint papers.

I was the main representative of Rhode Island and member of the steering committee at the NSF Northeast Big Data Hub before transitioning to the advisory board. I regularly serve on the PhD admission committee, have served on over 30 program committees and was a program chair at nine conference tracks/workshops, including co-chairing one of the largest database conferences, SIGMOD, in 2016 and 2019.

Community Engagement

I also strive to improve the database community and help junior researchers to find their footing. For example, I helped to implement the Toronto Paper Matching System (TPMS) at SIGMOD'19. This was one of many improvement Amol Deshpande, Anastasia Ailamaki, and I made to improve the review quality for SIGMOD'19 (see [109] for an overview). Furthermore, I regularly participate at Young Researchers symposiums/workshops (e.g., I gave the keynote at the VLDB'21 and VLDB'16 PhD workshop and the SIGMOD'15 New Researcher Symposium) and served as a mentor and judge at HackMIT in 2020 and 2021 to help junior CS students. I continuously advise undergraduate researchers as part of MIT's UROP program. Several undergraduates have been involved in papers and demos and my SuperUROP Yaateh received the "2020 SuperUROP award" for his work. To date, I wrote recommendation letters for over 75 researchers for PhD programs, tenure track positions, and promotions cases. More recently, I organized LADSIOS@VLDB (`https://www.ladsios.org/`), a new workshop, that is intended as a tutorial on recent work on learned algorithms, data structures, and instance-optimized systems to help young researchers to enter the field of ML for Systems. All 14 talks at the workshop will be given by young researchers, providing them with a unique platform to increase the visibility of their work, network, and form new collaborations.

References

[1] T. Kraska, A. Beutel, E. H. Chi, J. Dean, and N. Polyzotis, "The case for learned index structures," in *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018* (G. Das, C. M. Jermaine, and P. A. Bernstein, eds.), pp. 489–504, ACM, 2018.

[2] V. Nathan, J. Ding, M. Alizadeh, and T. Kraska, "Learning multi-dimensional indexes," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020* (D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, eds.), pp. 985–1000, ACM, 2020.

[3] J. Ding, V. Nathan, M. Alizadeh, and T. Kraska, "Tsunami: A learned multi-dimensional index for correlated data and skewed workloads," *Proc. VLDB Endow.*, vol. 14, no. 2, pp. 74–86, 2020.

[4] D. Ho, J. Ding, S. Misra, N. Tatbul, V. Nathan, V. Md, and T. Kraska, "LISA: towards learned DNA sequence search," *CoRR*, vol. abs/1910.04728, 2019.

[5] A. Kristo, K. Vaidya, U. Çetintemel, S. Misra, and T. Kraska, "The case for a learned sorting algorithm," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020* (D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, eds.), pp. 1001–1016, ACM, 2020.

[6] R. C. Marcus, P. Negi, H. Mao, C. Zhang, M. Alizadeh, T. Kraska, O. Papaemmanouil, and N. Tatbul, "Neo: A learned query optimizer," *Proc. VLDB Endow.*, vol. 12, no. 11, pp. 1705–1718, 2019.

[7] R. Marcus, P. Negi, H. Mao, N. Tatbul, M. Alizadeh, and T. Kraska, "Bao: Making learned query optimization practical," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021* (G. Li, Z. Li, S. Idreos, and D. Srivastava, eds.), pp. 1275–1288, ACM, 2021.

[8] O. Peckham, "MIT Is Developing a Tool for Machine Learning-Powered Data Retrieval." `https://www.datanami.com/2020/08/14/mit-is-developing-a-tool-for-machine-learning-powered-data-retrieval/`, 2020. [Online; accessed 7-July-2021].

[9] W. Knight, "Your next computer could improve with age." `https://www.technologyreview.com/2018/03/12/144740/your-next-computer-could-improve-with-age/`, 2018. [Online; accessed 7-July-2021].

[10] A. KATTE, "Indexing By Learning: A Revolutionary Idea That Can Shake The Core Of Computer Science." `https://analyticsindiamag.com/indexing-by-learning-a-revolutionary-idea-than-can-shake-the-core-of-computer-science/`, 2018. [Online; accessed 7-July-2021].

[11] DSAIL, "Ml for systems papers." `http://dsg.csail.mit.edu/mlforsystems/papers/`, 2021.

[12] S. Idreos and T. Kraska, "From auto-tuning one size fits all to self-designed and learned data-intensive systems," in *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019* (P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, eds.), pp. 2054–2059, ACM, 2019.

[13] A. A. Mamun, H. Wu, and W. G. Aref, "A tutorial on learned multidimensional indexes." `https://www.cs.purdue.edu/homes/aref/learned-indexes-tutorial.html`, 2020.

[14] A. Al-Mamun, H. Wu, and W. G. Aref, "A tutorial on learned multi-dimensional indexes," in *SIGSPATIAL '20: 28th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 3-6, 2020* (C. Lu, F. Wang, G. Trajcevski, Y. Huang, S. D. Newsam, and L. Xiong, eds.), pp. 1–4, ACM, 2020.

[15] P. Ferragina and G. Vinciguerra, "Learned data structures," in *Recent Trends in Learning From Data* (L. Oneto, N. Navarin, A. Sperduti, and D. Anguita, eds.), pp. 5–41, Springer International Publishing, 2020.

[16] H. Abu-Libdeh, D. Altinbüken, A. Beutel, E. H. Chi, L. Doshi, T. Kraska, X. Li, A. Ly, and C. Olston, "Learned indexes for a google-scale disk-based database," in *Proceedings of the Workshop on ML for Systems at NeurIPS*, 2020. `http://mlforsystems.org/assets/papers/neurips2020/learned_abu-libdeh_2020.pdf`.

[17] P. Negi, M. Interlandi, R. Marcus, M. Alizadeh, T. Kraska, M. Friedman, and A. Jindal, "Steering query optimizers: A practical take on big data workloads," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021* (G. Li, Z. Li, S. Idreos, and D. Srivastava, eds.), pp. 2557–2569, ACM, 2021.

[18] T. Kraska, "Northstar: An interactive data science system," *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 2150–2164, 2018.

[19] A. Crotty, A. Galakatos, E. Zgraggen, C. Binnig, and T. Kraska, "Vizdom: Interactive analytics through pen and touch," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 2024–2027, 2015.

[20] A. Crotty, A. Galakatos, K. Dursun, T. Kraska, C. Binnig, U. Çetintemel, and S. Zdonik, "An architecture for compiling udf-centric workflows," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1466–1477, 2015.

[21] A. Crotty, A. Galakatos, and T. Kraska, "Tupleware: Distributed machine learning on small clusters," *IEEE Data Eng. Bull.*, vol. 37, no. 3, pp. 63–76, 2014.

[22] A. Crotty, A. Galakatos, K. Dursun, T. Kraska, U. Çetintemel, and S. B. Zdonik, "Tupleware: "big" data, big analytics, small clusters," in *Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, www.cidrdb.org, 2015.

[23] A. Crotty, A. Galakatos, and T. Kraska, "Getting swole: Generating access-aware code with predicate pullups," in *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pp. 1273–1284, IEEE, 2020.

[24] A. Crotty, A. Galakatos, E. Zgraggen, C. Binnig, and T. Kraska, "The case for interactive data exploration accelerators (ideas)," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2016, San Francisco, CA, USA, June 26 - July 01, 2016* (C. Binnig, A. D. Fekete, and A. Nandi, eds.), p. 11, ACM, 2016.

[25] K. Dursun, C. Binnig, U. Çetintemel, and T. Kraska, "Revisiting reuse in main memory database systems," in *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017* (S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, eds.), pp. 1275–1289, ACM, 2017.

[26] L. F. Spiegelberg, R. Yesantharao, M. Schwarzkopf, and T. Kraska, "Tuplex: Data science in python at native code speed," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021* (G. Li, Z. Li, S. Idreos, and D. Srivastava, eds.), pp. 1718–1731, ACM, 2021.

[27] L. F. Spiegelberg and T. Kraska, "Tuplex: Robust, efficient analytics when python rules," *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1958–1961, 2019.

[28] E. R. Sparks, A. Talwalkar, V. Smith, J. Kottalam, X. Pan, J. E. Gonzalez, M. J. Franklin, M. I. Jordan, and T. Kraska, "MLI: an API for distributed machine learning," in *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013* (H. Xiong, G. Karypis, B. M. Thuraisingham, D. J. Cook, and X. Wu, eds.), pp. 1187–1192, IEEE Computer Society, 2013.

[29] T. Kraska, A. Talwalkar, J. C. Duchi, R. Griffith, M. J. Franklin, and M. I. Jordan, "Mlbase: A distributed machine-learning system," in *Sixth Biennial Conference on Innovative Data Systems Research, CIDR 2013, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*, www.cidrdb.org, 2013.

[30] E. R. Sparks, A. Talwalkar, D. Haas, M. J. Franklin, M. I. Jordan, and T. Kraska, "Automating model search for large scale machine learning," in *Proceedings of the Sixth ACM Symposium on Cloud Computing, SoCC 2015, Kohala Coast, Hawaii, USA, August 27-29, 2015* (S. Ghandeharizadeh, S. Barahmand, M. Balazinska, and M. J. Freedman, eds.), pp. 368–380, ACM, 2015.

[31] Z. Shang, E. Zgraggen, B. Buratti, F. Kossmann, P. Eichmann, Y. Chung, C. Binnig, E. Upfal, and T. Kraska, "Democratizing data science through interactive curation of ML pipelines," in *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019* (P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, eds.), pp. 1171–1188, ACM, 2019.

[32] C. Binnig, L. D. Stefani, T. Kraska, E. Upfal, E. Zgraggen, and Z. Zhao, "Toward sustainable insights, or why polygamy is bad for you," in *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*, www.cidrdb.org, 2017.

[33] Z. Zhao, L. D. Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska, "Controlling false discoveries during interactive data exploration," in *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017* (S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, eds.), pp. 527–540, ACM, 2017.

[34] Z. Zhao, E. Zgraggen, L. D. Stefani, C. Binnig, E. Upfal, and T. Kraska, "Safe visual data exploration," in *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017* (S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, eds.), pp. 1671–1674, ACM, 2017.

[35] E. Zgraggen, Z. Zhao, R. C. Zeleznik, and T. Kraska, "Investigating the effect of the multiple comparisons problem in visual analysis," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018* (R. L. Mandryk, M. Hancock, M. Perry, and A. L. Cox, eds.), p. 479, ACM, 2018.

[36] Y. Guo, C. Binnig, and T. Kraska, "What you see is not what you get!: Detecting simpson's paradoxes during data exploration," in *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2017, Chicago, IL, USA, May 14, 2017* (C. Binnig, J. M. Hellerstein, and A. G. Parameswaran, eds.), pp. 2:1–2:5, ACM, 2017.

[37] T. Kraska, M. Alizadeh, A. Beutel, E. H. Chi, A. Kristo, G. Leclerc, S. Madden, H. Mao, and V. Nathan, "Sagedb: A learned database system," in *9th Biennial Conference on Innovative Data Systems Research, CIDR 2019, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings*, www.cidrdb.org, 2019.

[38] T. Kraska, "Towards instance-optimized data systems," *Proc. VLDB Endow.*, vol. 14, no. 12, pp. 3222–3232, 2021.

[39] C. Mannig, "Machine learning just ate algorithms in one large bite." `https://twitter.com/chrmanning/status/940230539126046720`, 2017.

[40] K. Borne, "Wow! this could have huge benefits ¿¿ the case for learned index structures." `https://twitter.com/KirkDBorne/status/941162217809969152`, 2017.

[41] R. Rodger, "Learned indexes: a new idea for efficient data access." `https://2019.berlinbuzzwords.de/18/session/learned-indexes-new-idea-efficient-data-access.html`, 2019.

[42] A. COLYER, "The case for learned index structures – part I." `https://blog.acolyer.org/2018/01/08/the-case-for-learned-index-structures-part-i/`, 2018.

[43] R. Marcus, A. Kipf, A. van Renen, M. Stoian, S. Misra, A. Kemper, T. Neumann, and T. Kraska, "Benchmarking learned indexes," *Proc. VLDB Endow.*, vol. 14, no. 1, pp. 1–13, 2020.

[44] DSAIL, "(Learned Index Leaderboard." `https://learnedsystems.github.io/SOSDLeaderboard/leaderboard/`, 2021. [Online; accessed 7-July-2021].

[45] P. Ferragina, F. Lillo, and G. Vinciguerra, "Why are learned indexes so effective?," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 3123–3132, PMLR, 13–18 Jul 2020.

[46] Y. Dai, Y. Xu, A. Ganesan, R. Alagappan, B. Kroth, A. Arpaci-Dusseau, and R. Arpaci-Dusseau, "From wisckey to bourbon: A learned index for log-structured merge trees," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pp. 155–171, USENIX Association, Nov. 2020.

[47] K. Vaidya, E. Knorr, M. Mitzenmacher, and T. Kraska, "Partitioned learned bloom filters," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[48] M. Mitzenmacher, "A model for learned bloom filters and optimizing by sandwiching," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 462–471, 2018.

[49] Z. Dai and A. Shrivastava, "Adaptive learned bloom filter (ada-bf): Efficient utilization of the classifier with application to real-time information filtering on the web," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[50] I. Sabek, K. Vaidya, D. Horn, A. Kipf, and T. Kraska, "When are learned models better than hash functions?," in *International Workshop on Applied AI for Database Systems and Applications (AIDB@VLDB)*, AIDB '21, to appear.

[51] A. Kipf, R. Marcus, A. van Renen, M. Stoian, A. Kemper, T. Kraska, and T. Neumann, "Radixspline: a single-pass learned index," in *Proceedings of the Third International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, aiDM@SIGMOD 2020, Portland, Oregon, USA, June 19, 2020* (R. Bordawekar, O. Shmueli, N. Tatbul, and T. K. Ho, eds.), pp. 5:1–5:5, ACM, 2020.

[52] A. Galakatos, M. Markovitch, C. Binnig, R. Fonseca, and T. Kraska, "Fiting-tree: A data-aware index structure," in *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019* (P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, eds.), pp. 1189–1206, ACM, 2019.

[53] J. Ding, U. F. Minhas, B. Chandramouli, C. Wang, Y. Li, Y. Li, D. Kossmann, J. Gehrke, and T. Kraska, "Instance-optimized data layouts for cloud analytics workloads," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pp. 418–431, ACM, 2021.

[54] A. Kristo, K. Vaidya, and T. Kraska, "Defeating duplicates: A re-design of the learnedsort algorithm," *CoRR*, vol. abs/2107.03290, 2021.

[55] H. Mao, P. Negi, A. Narayan, H. Wang, J. Yang, H. Wang, R. Marcus, R. Addanki, M. K. Shirkoohi, S. He, V. Nathan, F. Cangialosi, S. B. Venkatakrishnan, W. Weng, S. Han, T. Kraska, and M. Alizadeh, "Park: An open platform for learning-augmented computer systems," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, eds.), pp. 2490–2502, 2019.

[56] P. Negi, R. Marcus, H. Mao, N. Tatbul, T. Kraska, and M. Alizadeh, "Cost-guided cardinality estimation: Focus where it matters," in *36th IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2020, Dallas, TX, USA, April 20-24, 2020*, pp. 154–157, IEEE, 2020.

[57] P. Negi, R. Marcus, A. Kipf, H. Mao, N. Tatbul, T. Kraska, and M. Alizadeh, "Flow-loss: Learning cardinality estimates that matter," *Proc. VLDB Endow.*, to appear.

[58] R. Marcus, "More Bao Results: Learned Distributed Query Optimization on Vertica, Redshift, and Azure Synapse." `https://learnedsystems.mit.edu/bao-distributed/`, 2021.

[59] Google, "ACM SIGMOD International Conference on Management of Data - Top Publication over the last 5 years." `https://scholar.google.com/citations?hl=en&vq=eng_databasesinformationsystems&view_op=list_hcore&venue=u1CjH9_75_cJ.2021`, 2021.

[60] A. Rashelbach, O. Rottenstreich, and M. Silberstein, "A computational approach to packet classification," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '20, (New York, NY, USA), p. 542–556, Association for Computing Machinery, 2020.

[61] S. Higuchi, J. Takemasa, Y. Koizumi, A. Tagami, and T. Hasegawa, "Feasibility of longest prefix matching using learned index structures," *SIGMETRICS Perform. Eval. Rev.*, vol. 48, p. 45–48, May 2021.

[62] C. Tang, Y. Wang, Z. Dong, G. Hu, Z. Wang, M. Wang, and H. Chen, "Xindex: A scalable learned index for multicore data storage," in *Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '20, (New York, NY, USA), p. 308–320, Association for Computing Machinery, 2020.

[63] B. Lu, J. Ding, E. Lo, U. F. Minhas, and T. Wang, "APEX: A high-performance learned index on persistent memory," *CoRR*, vol. abs/2105.00683, 2021.

[64] X. Wei, R. Chen, and H. Chen, "Fast rdma-based ordered key-value store using remote learned cache," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pp. 117–135, USENIX Association, Nov. 2020.

[65] J. Wu, Y. Zhang, S. Chen, Y. Chen, J. Wang, and C. Xing, "Updatable learned index with precise positions," *Proc. VLDB Endow.*, vol. 14, no. 8, pp. 1276–1288, 2021.

[66] P. Ferragina, F. Lillo, and G. Vinciguerra, "On the performance of learned data structures," *Theor. Comput. Sci.*, vol. 871, pp. 107–120, 2021.

[67] E. Zgraggen, A. Galakatos, A. Crotty, J. Fekete, and T. Kraska, "How progressive visualizations affect exploratory analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 8, pp. 1977–1987, 2017.

[68] P. Eichmann, E. Zgraggen, C. Binnig, and T. Kraska, "Idebench: A benchmark for interactive data exploration," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020* (D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, eds.), pp. 1555–1569, ACM, 2020.

[69] Z. Shang, E. Zgraggen, B. Buratti, P. Eichmann, N. Karimeddiny, C. Meyer, W. Runnels, and T. Kraska, "Davos: A system for interactive data-driven decision making," *Proc. VLDB Endow.*, to appear.

[70] M. El-Hindi, Z. Zhao, C. Binnig, and T. Kraska, "Vistrees: fast indexes for interactive data exploration," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2016, San Francisco, CA, USA, June 26 - July 01, 2016* (C. Binnig, A. D. Fekete, and A. Nandi, eds.), p. 5, ACM, 2016.

[71] A. Galakatos, A. Crotty, E. Zgraggen, C. Binnig, and T. Kraska, "Revisiting reuse for approximate query processing," *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1142–1153, 2017.

[72] N. Chepurko, R. Marcus, E. Zgraggen, R. C. Fernandez, T. Kraska, and D. Karger, "ARDA: automatic relational data augmentation for machine learning," *Proc. VLDB Endow.*, vol. 13, no. 9, pp. 1373–1387, 2020.

[73] Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, "Slice finder: Automated data slicing for model validation," in *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pp. 1550–1553, IEEE, 2019.

[74] Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, "Automated data slicing for model validation: A big data - AI integration approach," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 12, pp. 2284–2296, 2020.

[75] M. Hulsebos, K. Z. Hu, M. A. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, Ç. Demiralp, and C. A. Hidalgo, "Sherlock: A deep learning approach to semantic data type detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019* (A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, eds.), pp. 1500–1508, ACM, 2019.

[76] L. D. Stefani, L. F. Spiegelberg, E. Upfal, and T. Kraska, "Vizcertify: A framework for secure visual data exploration," in *2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019, Washington, DC, USA, October 5-8, 2019* (L. Singh, R. D. D. Veaux, G. Karypis, F. Bonchi, and J. Hill, eds.), pp. 241–251, IEEE, 2019.

[77] K. Z. Hu, S. N. S. Gaikwad, M. Hulsebos, M. A. Bakker, E. Zgraggen, C. A. Hidalgo, T. Kraska, G. Li, A. Satyanarayan, and Ç. Demiralp, "Viznet: Towards A large-scale visualization learning and benchmarking repository," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019* (S. A. Brewster, G. Fitzpatrick, A. L. Cox, and V. Kostakos, eds.), p. 662, ACM, 2019.

[78] S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, T. Kraska, T. Milo, and E. Wu, "Sampleclean: Fast and reliable analytics on dirty data," *IEEE Data Eng. Bull.*, vol. 38, no. 3, pp. 59–75, 2015.

[79] Y. Chung, S. Krishnan, and T. Kraska, "A data quality metric (DQM): how to estimate the number of undetected errors in data sets," *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1094–1105, 2017.

[80] Y. Chung, M. L. Mortensen, C. Binnig, and T. Kraska, "Estimating the impact of unknown unknowns on aggregate query results," in *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016* (F. Özcan, G. Koutrika, and S. Madden, eds.), pp. 861–876, ACM, 2016.

[81] Y. Chung, S. Servan-Schreiber, E. Zgraggen, and T. Kraska, "Towards quantifying uncertainty in data analysis & exploration," *IEEE Data Eng. Bull.*, vol. 41, no. 3, pp. 15–27, 2018.

[82] J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo, "A sample-and-clean framework for fast and accurate query processing on dirty data," in *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014* (C. E. Dyreson, F. Li, and M. T. Özsu, eds.), pp. 469–480, ACM, 2014.

[83] Y. Chung, M. L. Mortensen, C. Binnig, and T. Kraska, "Estimating the impact of unknown unknowns on aggregate query results," *ACM Trans. Database Syst.*, vol. 43, no. 1, pp. 3:1–3:37, 2018.

[84] S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, and T. Kraska, "Privateclean: Data cleaning and differential privacy," in *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016* (F. Özcan, G. Koutrika, and S. Madden, eds.), pp. 937–951, ACM, 2016.

[85] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1483–1494, 2012.

[86] S. Goldberg, D. Z. Wang, and T. Kraska, "CASTLE: crowd-assisted system for text labeling and extraction," in *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, November 7-9, 2013, Palm Springs, CA, USA* (B. Hartman and E. Horvitz, eds.), AAAI, 2013.

[87] B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar, "Crowdsourced enumeration queries," in *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013* (C. S. Jensen, C. M. Jermaine, and X. Zhou, eds.), pp. 673–684, IEEE Computer Society, 2013.

[88] M. L. Mortensen, G. P. Adam, T. A. Trikalinos, T. Kraska, and B. C. Wallace, "An exploration of crowdsourcing citation screening for systematic reviews," *Res Synth Methods*, vol. 8, no. 3, p. 366–386, 2017.

[89] J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng, "Leveraging transitive relations for crowdsourced joins," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013* (K. A. Ross, D. Srivastava, and D. Papadias, eds.), pp. 229–240, ACM, 2013.

[90] K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo, "Vizml: A machine learning approach to visualization recommendation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019* (S. A. Brewster, G. Fitzpatrick, A. L. Cox, and V. Kostakos, eds.), p. 128, ACM, 2019.

[91] D. Etherington, "MIT's new interactive machine learning prediction tool could give everyone AI superpowers." `https://techcrunch.com/2019/06/27/mits-new-interactive-machine-learning-prediction-tool-could-give-everyone-ai-superpowers/`, 2019. [Online; accessed 7-July-2021].

[92] M. Hutson, "No coding required: Companies make it easier than ever for scientists to use artificial intelligence." https://www.sciencemag.org/news/2019/07/no-coding-required-companies-make-it-easier-ever-scientists-use-artificial-intelligence, 2019. [Online; accessed 7-July-2021].

[93] T. Kraska, "Approximate query processing for interactive data science," in *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017* (S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, and D. Suciu, eds.), p. 525, ACM, 2017.

[94] C. Binnig, A. Crotty, A. Galakatos, T. Kraska, and E. Zamanian, "The end of slow networks: It's time for a redesign," *Proc. VLDB Endow.*, vol. 9, no. 7, pp. 528–539, 2016.

[95] E. Zamanian, C. Binnig, T. Kraska, and T. Harris, "The end of a myth: Distributed transaction can scale," *Proc. VLDB Endow.*, vol. 10, no. 6, pp. 685–696, 2017.

[96] E. Zamanian, J. Shun, C. Binnig, and T. Kraska, "Chiller: Contention-centric transaction execution and data partitioning for modern networks," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020* (D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, eds.), pp. 511–526, ACM, 2020.

[97] E. Zamanian, X. Yu, M. Stonebraker, and T. Kraska, "Rethinking database high availability with RDMA networks," *Proc. VLDB Endow.*, vol. 12, no. 11, pp. 1637–1650, 2019.

[98] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "Crowddb: answering queries with crowdsourcing," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011* (T. K. Sellis, R. J. Miller, A. Kementsietsidis, and Y. Velegrakis, eds.), pp. 61–72, ACM, 2011.

[99] A. Feng, M. J. Franklin, D. Kossmann, T. Kraska, S. Madden, S. Ramesh, A. Wang, and R. Xin, "Crowddb: Query processing with the VLDB crowd," *Proc. VLDB Endow.*, vol. 4, no. 12, pp. 1387–1390, 2011.

[100] A. Doan, M. J. Franklin, D. Kossmann, and T. Kraska, "Crowdsourcing applications and platforms: A data management perspective," *Proc. VLDB Endow.*, vol. 4, no. 12, pp. 1508–1509, 2011.

[101] T. Kraska, G. Pang, M. J. Franklin, S. Madden, and A. D. Fekete, "MDCC: multi-data center consistency," in *Eighth Eurosys Conference 2013, EuroSys '13, Prague, Czech Republic, April 14-17, 2013* (Z. Hanzálek, H. Härtig, M. Castro, and M. F. Kaashoek, eds.), pp. 113–126, ACM, 2013.

[102] G. Pang, T. Kraska, M. J. Franklin, and A. D. Fekete, "PLANET: making progress with commit processing in unpredictable environments," in *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014* (C. E. Dyreson, F. Li, and M. T. Özsu, eds.), pp. 3–14, ACM, 2014.

[103] T. Kraska, M. Hentschel, G. Alonso, and D. Kossmann, "Consistency rationing in the cloud: Pay only when it matters," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 253–264, 2009.

[104] M. Brantner, D. Florescu, D. A. Graf, D. Kossmann, and T. Kraska, "Building a database on S3," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008* (J. T. Wang, ed.), pp. 251–264, ACM, 2008.

[105] C. Binnig, D. Kossmann, T. Kraska, and S. Loesing, "How is the weather tomorrow?: towards a benchmark for the cloud," in *Proceedings of the 2nd International Workshop on Testing Database Systems, DBTest 2009, Providence, Rhode Island, USA, June 29, 2009* (B. Dageville and C. Binnig, eds.), ACM, 2009.

[106] B. Howe, M. J. Franklin, L. M. Haas, T. Kraska, and J. D. Ullman, "Data science education: We're missing the boat, again," in *33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017*, pp. 1473–1474, IEEE Computer Society, 2017.

[107] B. Howe, M. J. Franklin, J. Freire, J. Frew, T. Kraska, and R. Ramakrishnan, "Should we all be teaching "intro to data science" instead of "intro to databases"?," in *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014* (C. E. Dyreson, F. Li, and M. T. Özsu, eds.), pp. 917–918, ACM, 2014.

[108] E. Zamanian, "Scalable distributed transaction processing on modern rdma-enabled networks," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021* (G. Li, Z. Li, S. Idreos, and D. Srivastava, eds.), p. 8, ACM, 2021.

[109] A. Ailamaki, P. Chrysogelos, A. Deshpande, and T. Kraska, "The SIGMOD 2019 research track reviewing system," *SIGMOD Rec.*, vol. 48, no. 2, pp. 47–54, 2019.