III: Medium: Collaborative Research: From Answering Questions to Questioning Answers (and Questions)---Perturbation Analysis of Database Queries

Jun Yang (PI), Bill Adair, Pankaj K. Agarwal: Duke University (NSF-IIS-1408846)
James T. Hamilton: Stanford University (NSF-IIS-1408915)
Chengkai Li: University of Texas at Arlington (NSF-IIS-1408928)

Year 1 Project Report

Note: Responses are limited to 8000 characters max.

Accomplishments

What are the major goals of the project?

In the age of data ubiquity, decision making is increasingly driven by data. Oftentimes, database queries are used to identify issues, debate strategies, make choices, and explain decisions. How these database queries are formulated can significantly influence the decision making process. A poor choice of query parameters---be it intentionally or accidentally---may give a biased view of the underlying data, and lead to decisions that are wrong, misguided, or "brittle" when reality deviates from assumptions. Database research has in the past focused on how to answer queries, but has not devoted much attention to how queries impact decision making, or how to formulate "good" queries from the outset. This project aims to fill this void. The key insight is perturbation analysis of data queries---i.e., studying how perturbations of the query form and parameters affect the query result. For example, slight query perturbations leading to very different results help identify potential pitfalls in decision making. In general, perturbation analysis of database queries reveals how queries affect the robustness and objectivity of decisions, and helps decision makers identify "good" queries that will influence their decisions.

This project plans to carry out a systematic study of perturbation analysis of database queries. On the modeling front, the project proposes query response surface (QRS) over the parametric space as a framework for perturbation analysis. Intuitive notions of query "goodness" (for the purpose of supporting decisions), such as fairness and robustness, can be formulated as statistical,

geometric, and topological properties of the QRS. The framework also allows practical problems to be formulated in terms of the QRS. For example, a brittle decision can be illustrated by identifying its pitfalls, which can be cast as an optimization problem of searching the QRS for slight perturbations with large result deviations; the problem of finding "good" queries that will influence a decision can be cast as that of finding points with desired properties in the relevant region of the QRS. On the algorithmic front, fundamental research problems arise in coping with the complexity of QRS and the vast space of perturbations. While there has been much study on perturbations of data, considering perturbations of queries poses novel challenges and compounds existing ones. The project will develop both efficient representations of QRS and fast algorithms for exploring and analyzing the QRS, using scalable techniques for indexing, optimization, and incremental evaluation that rely on sampling, approximation, and geometric insights. On the systems and applications front, this project plans to deliver the core features of perturbation analysis as a web service with a public API, and address the design and scalability challenges. The project will produce a general-purpose website for applying perturbation analysis of database queries, as well as websites customized for several domains of public interest. The websites will include a facet-driven interface and features that help collaboration and dissemination.

In today's data-driven society, there is increasing demand for the proposed research in many application domains such as public policy, urban planning, business intelligence, and health care. This project will significantly expand the functionality of database systems, making them easier to use (and harder to misuse) for a new generation of data-driven decision makers, especially those outside the traditional "data-heavy" disciplines such as computer science and statistics. This project will develop courses, seminars, and workshops targeting this much broader population of data-driven decision makers, to help train them in data and quantitative analysis, and in interpreting results critically.

What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?

Major Activities

[Computational Fact-Checking]

Our news is saturated with claims of “facts” made from data. Database research has in the past focused on how to answer queries, but has not devoted much

attention to discerning more subtle qualities of the resulting claims, e.g., is a claim "cherry-picking"? We have been working on developing a framework that models claims based on structured data as parameterized queries. A key insight is perturbation analysis: we can learn a lot about a claim by perturbing its parameters and seeing how its conclusion changes—basically by examining its QRS. This framework lets us formulate practical fact-checking tasks—reverse-engineering (often intentionally) vague claims, and countering questionable claims—as computational problems. Along with the modeling framework, we develop an algorithmic framework that enables efficient instantiations of "meta" algorithms by supplying appropriate algorithmic building blocks. These results have been published in PVLDB 2014 and a journal version is under review. A system called *iCheck* was demonstrated in SIGMOD 2014 and the 2014 Computational+Journalism Symposium.

We note that the same framework for checking claims can also be used to automatically look for interesting claims (of a give template) in a dataset. We will elaborate under [Computational Lead-Finding] below.

We continue to improve our models and algorithms. The QRS may be a complex surface. For the problem of finding interesting claims or counterarguments, it is not always clear what are the most interesting points on the QRS to return—the user may be overwhelmed by a large number of results, some of which can be quite similar. Therefore, we are working on a new problem formulation that allows us to pick points that are not only strong themselves but together also provide good coverage for all relevant portions of QRS. While previous work on top-*k* and search also addresses result diversity, the results do not necessarily represent the overall QRS. Thus, we have developed new algorithms for our problem formulation. The results are currently being prepared for submission.

[Computational Lead-Finding]

Lead-finding from data can be modeled as a computational task over the QRS—in essence, we look for strong points in the QRS that are relevant to the context of interest. When the claim template is given, we can develop efficient algorithms tailored for the given claim template. We have worked on a number problem in this area, including discovering prominent streaks from time series (KDD 2011, TKDD 2014), finding one-of-few claims from multidimensional data (KDD 2012), incrementally computing prominent situational facts (ICDE 2014), and mining frequent episodes online (ICDE 2015). Implementations of these lead-finding algorithms have been demonstrated at SIGMOD 2014, VLDB 2014, and the 2014

Computation+Journalism Symposium. The VLDB 2014 *FactWatcher* demo won the best demo award.

Another aspect of lead-finding addresses how to find the claims to check in the first place. We are working on applying machine learning and NLP methods for automatic discovery of factual claims worthy of checking from political debates. We have built a website to collect labeled examples for sentences from all US presidential debates. We have also developed a demo that applies the trained model to detect and highlight sentences worthy of checking from passages of text. A paper describing these results has been accepted to CIKM 2015.

[System Support]

We are developing a system that simplifies scalable perturbation analysis. Algorithms tailored towards particular query templates are efficient but require efforts to develop; algorithms that assume nothing about the query function are generic but less efficient. The new system we are building will support general SQL queries, which represent a practical alternative between the two extremes. It will feature automatic parallel execution of a query template with different parameter settings, significantly reducing the development and execution times of perturbation analysis in time-critical applications such as journalism. Besides parallelization, we have identified several effective optimizations, such as memoization and pruning. In the first iteration of our system, we will rely on user to annotate memoization points and supply pruning functions, but in the future we plan to study how these tasks can be automated. The first iteration of our system prototype is already running, and we are working on a paper describing our results on this problem.

[Visualization Support]

Visualization can be a powerful and convenient way to aid intuitive perturbation analysis. In collaboration with Google, we have worked on efficiently visualizing a large set of 2-d points that are result of evaluating a query over data on different entities (such as basketball players, authors, politicians). The point set can be visualized as a combination of a heat map (to capture the overall distribution) and a scatterplot (to capture the outliers). The computational challenge is that computing the points to visualize from raw data may be expensive and slow. To improve efficiency and interactivity, we have devised sampling-based algorithms that generate visualizations with as little loss in

fidelity as possible given a computation budget. The results will appear in PVLDB 2015.

[Coping with Uncertainty]

Besides perturbing parameter settings for a query, the dataset itself can be perturbed too—this interesting perspective is motived by data uncertainty, which is commonplace in reality and very costly to resolve. We are interested in understanding how data uncertainty affects query answer (and hence claim quality) as well. Lots of problems are still open in this area. The first problem we studied is the range-max problem—which, given a set of weighted points in d dimensions, finds the maximum weight point in a query hyper-rectangle. We consider various models of uncertainty—e.g., the weight or location of a point is specified by a distribution, or a point may exist with some probability. We study the problem of computing the expected or most-likely maximum weight inside an axis-parallel orthogonal query range. We investigate both exact and approximation algorithms.

To help fact-checking using data containing uncertainty, we are also working on the problem of "targeted data cleaning," which selects data items to clean under a budget with the goal of finding strong counterarguments to a given claim. One problem formulation assumes that we are given a space of possible counterarguments and a distribution of possible outcomes for the action of cleaning each data item; we need to find a set of items to clean to maximize the expected strength of the best counterargument over all possible cleaning outcomes. Another possible formulation assumes that the counterarguments are from a distribution as well; in this case, we might want to find items that maximize the probability that a randomly drawn counterargument is good enough. Our approach is to model the above problems as budgeted stochastic optimization problems and find techniques to guarantee good approximation solutions.

[Education, Dissemination, and Broader Impact Activities]

Because of space constraints, we describe other activities for education, dissemination and broader impacts under other sections of this report.

Specific Objectives

[Computational Fact-Checking] Our objectives are to develop a feasible computational approach toward fact-checking; demonstrate the practicality and generality of the framework; and develop efficient algorithms for various claim templates. While our goals do not include handling all forms of claims or replacing human fact-checkers, we wish to use the computational approach to significantly reduce manual efforts for a large class of common claims, thereby making human fact-checkers more effective and allowing them to focus on more challenging tasks.

[Computational Lead-Finding] For finding interesting claims from data, our objective is to devise effective and efficient algorithms for a variety of interesting claim templates, and understand what generic optimizations are possible for general, black-box functions. For finding claims to check, our objectives are to collect manually labeled training examples from past US presidential debates, and to develop a supervised learning algorithm with reasonable accuracy such that it can significantly narrow down the list of candidates that need to be further considered by human experts.

[System Support] Our objective is to develop a system that allows easy specification and fast execution of perturbation analysis for general SQL query templates, such that it can reduce the development and execution times of fact-checking and lead-finding tasks to the point where their applications in journalism become feasible.

[Visualization Support] Our objective is to develop fast algorithms to produce, at interactive speeds from large datasets, approximate heat maps and scatterplots that are visually accurate.

[Coping with Uncertainty] Our objectives are to develop a deeper understanding of how data uncertainty (or perturbation) affects query results, to develop solutions to problems in querying uncertain data and cleaning uncertain data for the purpose of fact-checking.

[Education, Dissemination, and Broader Impact] Our objectives are to develop course materials, give presentations and lectures, participate in and/or organize panels, workshops, and conferences both in and out of computer science to help 1) attract the computer science community to this line of research in the public interest; and 2) engage the journalism community and the public with applications of this research.

Significant Results

[Computational Fact-Checking]

We have developed QRS-based framework that allows us to formulate various fact-checking tasks as computational ones. We show that claim qualities such as uniqueness, fairness, and robustness can be defined as properties of the QRS; the tasks of finding counter-arguments and reverse-engineering vague claims can be defined as bi-criteria optimization problems, where the criteria are claim strength and sensibility (relative to the claim context). We designed an extensible architecture with "pay-as-you-go" support for efficient fact-checking. New claims can be supported with little work upfront, by baseline algorithms that exhaustively examine all possible perturbations. With knowledge of the structure of the parameter space and/or query template, advanced users can supply low-level algorithmic building blocks that can be plugged into our generic meta-algorithms to achieve higher efficiency. We evaluated the generality, effectiveness, and efficiency of our approach on multiple real claim templates and datasets.

For the problem of finding *k* strong representatives of QRS, *k*-means-like algorithms, in theory, gives optimal solutions for our utility function if run with many initial states. We show that because our utility function is a submodular set function, a more efficient greedy algorithm can provide constant-factor approximation. We further improve the greedy algorithm's quality with post-processing, and improve its efficiency using geometric properties of the utility function.

[Computational Lead-Finding]

To find leads automatically from data, we developed efficient algorithms for a series of claim templates, including discovering prominent streaks from time series (e.g., "*the Nikkei 225 closed below 10000 for the 12th consecutive week, the longest such streak since June 2009*"), finding one-of-few claims from multidimensional data (e.g., "*only 10 players in NBA history had more points, more rebounds, and more assists per game than Sam Lacey in their career*"), incrementally computing prominent situational facts (e.g., "*Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992*"), and mining frequent episodes online.

To find claims worth checking, we have accomplished about 50% of our data collection objective. We extracted every sentence spoken during 30 debates in the past 11 presidential elections, a total of 20,788 sentences. So far we have recruited 223 coders to participate in labeling data. We evaluated several supervised learning models using the labeled data. Thousands of sentence features (e.g., length, words, sentiments, part-of-speech tags) were extracted and important features were selected. Preliminary experiment results show that our system is accurate 85% of the time when it declares an important factual claim, and 65% of truly important factual claims are deemed by our system as important. This degree of accuracy is comparable to top-quality human coders.

[System Support]

We have finished our first iteration of the system aimed at reducing development and execution costs of scalable perturbation analysis of SQL queries. Currently, it relies on users to provide a small number of optimization hooks, but it hides from users low-level implementation details and optimization knobs, including how to parallelize computation on Spark, how to use distributed cache and replicated SQL stores for memoization and pruning, and how to balance parallelism and sequentiality (for pruning). The system monitors and learns from execution traces to dynamically adjust the optimization knobs. Our experiments reveal that our system produces a substantial improvement in execution time over naive approaches, while keeping development complexity at a minimum. For many datasets, simple pruning functions can effectively trim the number of parameter settings requiring evaluation by up to 90%.

[Visualization Support]

For efficient approximate production of 2-d visualization, we have developed a two-stage sampling-based algorithm aimed at reducing the amount of data and computation required. In the first stage, query evaluation is performed on a sample of all objects' data. Base on the results on samples, a small number of objects are selected as potential outliers. In the second stage, evaluation is performed on the full data associated with these objects to obtain a scatterplot. For the remaining objects, a small random subset is chosen to generate a heat map. We use an adaptation of the *Earth Mover's Distance* to measure the quality of the approximate heat map, and standard *precision*, *recall*, and *F-score* to measure the quality of the approximate scatterplot. Experiments on real datasets show that our approach is able to generate high-quality visualizations much faster and using far fewer data accesses than evaluation on the full dataset.

[Coping with Uncertainty]

We have developed an efficient exact algorithm to answer range expected max queries that works for each model and dimension. The same algorithm also generalizes to find the most-likely maximum point in a query hyper-rectangle. Besides the exact algorithm, we have results for approximation solutions for the expected case. Using prophet-inequalities from stochastic optimization we propose a 1/2-approximation algorithm for the expected maximum weight in poly-logarithmic time that works in any dimension and defined model. Assuming bounded probabilities or bounded spread of weights we also give arbitrary constant approximation factor algorithms. Under some more natural conditions we can significantly improve the complexity of the most-likely maximum weight. If weights are assigned to points randomly, we give an exact algorithm with poly-logarithmic expected time and near linear space. Finally, we prove a lower bound for the expected maximum weight in orthogonal queries for points in dimension greater than 1, using a reduction from the integer set intersection problem.

Work on cleaning uncertain data for the purpose of fact-checking is still ongoing.

[Education, Dissemination, and Broader Impact]

We describe the results of activities for education, dissemination and broader impacts under other sections of this report.

Key Outcomes or Other Achievements:

[Computational Fact-Checking]

Our results were published in PVLDB 2014 and a journal version is under review. A system called *iCheck*, which fact-checks claims about publication records, baseball statistics, and congressional voting records, was demonstrated in SIGMOD 2014 and the 2014 Computational+Journalism Symposium.

[Computational Lead-Finding]

Our results on finding leads from data were published in a series of papers in KDD 2011, KDD 2012, ICDE 2014, and TKDD 2014. Implementations of the lead algorithms were demonstrated at SIGMOD 2014, VLDB 2014, and the 2014

Computation+Journalism Symposium. The VLDB 2014 *FactWatcher* demo won the best demo award.

Our system for finding claims worthy checking from text, called *ClaimBuster*, is up and running with accuracy that rivals top-quality human coders. Our data collection website has been operating for a time now, with more than two hundred coders participating. We are now readying *ClaimBuster* for the 2016 elections. A paper describing these results has been accepted to CIKM 2015.

[System Support]

The first iteration of our system prototype is functional. Our experiments show that it produces a substantial improvement in execution time over naive approaches, while keeping development complexity at a minimum. We are working on a paper describing our results.

[Visualization Support]

We have developed an efficient sampling-based method of generating approximate heat map and scatterplot visualizations for 2-d result sets. The method is fast enough to support interactive data exploration, and accurate enough for human perception. Our results are published in PVLDB 2015.

[Coping with Uncertainty]

We have developed a suite of algorithms for range-max query over uncertain data under different models and assumptions. We are currently preparing the paper for submission.

[Education, Dissemination, and Broader Impact]

We describe the outcomes and achievements of our activities for education, dissemination and broader impacts under other sections of this report.

What opportunities for training and professional development has the project provided?

At Duke, the project provided training for 3 PhD students and 4 undergraduate students in Computer Science. At UT Arlington, the project provided training for 1 PhD and 1 MS students. Aspects of this project have been integrated into

undergraduate computer science courses that Yang and Li are teaching at Duke and UT Arlington. The project has also served as a student recruiting tools for the computer science departments at these institutions.

The PIs at Duke organized a seminar series on Data+Journalism in which journalists, computer scientists, and researchers from public policy were invited.

At UT Arlington, Li organized a fact-checking workshop in June 2015. The attendees were graduate and undergraduate students. The workshop helped collect more labeled data for *ClaimBuster*, and served an educational purpose. The workshop began with a training session explaining the task of identifying claims worth checking. Participants also discussed more intricate cases. Through the workshop, the students obtained better understanding of fact-checking in general and learned to appreciate the subtlety of discerning important factual claims from other statements.

The data collection effort at UT Arlington also involved many students in the Department of Communication. Instructors of their courses collaborated with the research team and offered students extra credits for participating in data collection. The experience of learning about fact-checking and contributing labeled data was valuable to them as some of them may work in related fields after graduation.

How have the results been disseminated to communities of interest?

In the computer science community, our research results have been disseminated through papers and demonstrations listed earlier (and under [Products]), as well as presentations at other institutions and venues. During 2014-15, Yang gave invited talks about this project at MIT and Tsinghua University; Li gave invited talks in PyData Dallas 2015 and IEEE Computer Society Fort Worth section meeting. Yang co-organized the *First International Workshop on Bringing the Value of "Big Data" to Users* (https://sites.google.com/site/data4u2014/), on Sep. 1, 2014, in conjunction with VLDB 2014. The workshop helped promote fresh, user-centric looks at the big data challenges, such as the use of big data in applications for public interest like this project.

The project team has also been keen on disseminating the results beyond the computer science community. At Stanford, Hamilton organized two conferences with themes related to the project. *Data Driven: Coding and Writing Transportation's Future* (https://comm.stanford.edu/data-driven-conference-drives-discussions-on-transportation-data/), on Feb. 13, 2015, is a conference on how data from cars, including sensor data, can help predict problems with

institutions; the conference had panel with journalists discussing how to use such data to monitor problems with local and state governments. The 49th *Annual McClatchy Symposium* (https://comm.stanford.edu/mcclatchy/), on Apr. 16, 2015, was titled *Corruption: Who Plays? Who Pays?* Journalists from *LA Times*, *NY Times*, and Center for Investigative Reporting spoke with Stanford social scientists about how to spot patterns in campaign finance data that might be indicative to problems. Both conferences relate to the project in that they involve thinking about how to use databases to spot interesting, policy-related anomalies. One product of the first conference is a public website (http://www.datadrivenstanford.org/) with a big collection of transportation datasets that provide interesting materials for fact-checking and lead-finding.

Many project team members from Duke, Stanford, and UT participated in the 2014 Computation+Journalism Symposium, which attracted both journalists and computer scientists. The team presented a paper describing the project and gave demos of *iCheck* and *FactWatcher*. The team wrote a short article for *American Journalism Review* on computational fact-checking. Adair and Yang participated in the American Press Institute's Thought Leader Summit, "Truth in Politics 2014: A Status Report on Fact-Checking Journalism," on Dec. 10, 2014. Adair moderated the panel "An Insider Perspective," and Yang was a panelist for "Can Technology Change Fact-Checking?"

Adair hosted visits by Adam Long of *Automated Insights* (a Durham-based company that does computational narratives) on Mar. 3, 2015, and Julian Rademeyer of *Africa Check* (a fact-checking organization in South Africa) on Apr. 2, 2015. The team gave presentations and demos of the project to the visitors. A number of team members led by Yang paid a reciprocal visit to *Automated Insights* on May 1, 2015. Yang delivered a lecture on computational journalism at the Workshop on Journalism and Public Policy for media professionals visiting Duke from Nanjing, China in December 2014. At UT Arlington, Li hosted a departmental colloquium on March 27, 2015, in which Jon McClure and Daniel Lathrop from the *Dallas Morning News* presented their data journalism projects. The team also made presentations and demos of the project to the visitors. On April 29, 2015, Li and his students visited the *Dallas Morning News* and gave a talk about the project. All these activities provided opportunities to directly engage media professionals.

For the *ClaimBuster* system currently under active development at UT Arlington, websites for labeled data collection (http://idir-server2.uta.edu/classifyfact_survey/) and claim-finding demo

(http://idir-server2.uta.edu/classifyfact_survey/claimbuster/) are both available to the general public.

What do you plan to do during the next reporting period to accomplish the goals?

For computational fact-checking and lead-finding, we plan to continue enriching the modeling power of our framework for a wider range of claim types, and improving efficiency and accuracy of our algorithms. We expect to make good progress on intelligent result selection and system support for perturbation analysis. We will continue to collect labeled data for the problem of finding check-worthy claims, and improve our machine learning algorithms. We plan to extend our simple demos into full-fledged fact-checking websites intended for public consumption. We will make our tools available to fact-checkers to help them find checkable claims during the 2016 campaign. We have also obtained a Duke basketball dataset that can be used on public websites with fewer restrictions.

Motivated by our experience with labeled data collection, we also plan to investigate how to generalize the data collection methodology to other types of tasks. Particularly, we will study approaches for accurately estimating the work quality of crowdsourcers and incentive mechanisms for improving their work quality.

For uncertainty support, while we continue to deepen our understanding of the effect of data uncertainty on query results, we plan to devote more efforts to the problem of "targeted" data cleaning, which selects data items to clean under a budget with the goal of verifying a claim or finding strong counterarguments.

Impacts

What is the impact on the development of the principal discipline(s) of the project?

Traditional database research has focused on answering queries, but has not devoted much attention to discerning the quality of the resulting claims, or to formulating good queries from the outset. This project fills this void, by advancing the understanding of what makes for a high-quality claim based on data, and how to find queries that lead there. This project is helping to lay the foundation for perturbation analysis of database queries, by tackling multiple aspects of the problem, from algorithmic to system-building challenges, from

visualization to connection with data uncertainty, and from cluster computing to crowdsourcing.

What is the impact on other disciplines?

The project has identified compelling applications of perturbation analysis, namely fact-checking and lead-finding using structured data for journalism. The project has built working systems to demonstrate that much of work traditionally done by journalists by hand can in fact be formulated as computational tasks and hence automated, leaving journalists with more time for more important tasks, and enabling more timely and comprehensive news coverage, potentially in areas that have been traditionally underserved.

What is the impact on the development of human resources?

So far, the project has supported close training and development of 4 PhD students, 1 MS student, and 4 undergraduate students at Duke and UT Arlington. Yu Wu plans to graduate with his PhD from Duke in the summer of 2015.

We have revamped the undergraduate computer science curriculum at Duke and UT Arlington to incorporate elements of this research. At UT Arlington, our research has also been introduced through courses at the Department of Communication, exposing students outside computer science to computational, data-driven approaches to problems in news and media.

What is the impact on physical resources that form infrastructure?

Nothing to report.

What is the impact on institutional resources that form infrastructure?

Nothing to report.

What is the impact on information resources that form infrastructure?

We have built a number of websites that enrich the public information infrastructure. At Stanford, we are making available a big collection of transportation datasets that provide interesting materials for fact-checking and lead-finding. At UT Arlington, *ClaimBuster* serves as an important resource to help fact-checkers prioritize their work; *ClaimBuster*'s labeled data collection website serves an educational resource for the public, and will also provide good training data for researchers working on related problems. At Duke and UT Arlington, we are also in the process of making our other fact-checking and lead-finding demos available to the general public.

What is the impact on technology transfer?

The project team has collaborated closely with Cong Yu at Google. You Wu, a PhD student working on this project, interned at Google for two summers, and will be joining Cong Yu's team at Google in fall 2015. Through these connections, we are hopeful that some results from this project will be incorporated into Google's commercial offerings.

A number of media and fact-checking organizations have expressed in our research, and we are actively engaging them with visits and presentations. For details, please see the section of this report on [How have the results been disseminated to communities of interest?].

What is the impact on society beyond science and technology?

This project benefits many domains where decisions are increasingly driven by data, e.g., public policy, business intelligence, homeland security, and health care. The impact of this research extends beyond fact-checking and lead-finding, because it advances fundamental understanding of how query results respond to perturbations in query parameters and/or data, a core database problem with applications ranging from optimization of marketing strategies to impact evaluation of public policies.

A focus application of this project is public interest journalism, as resources are severely strained and innovation is pressingly needed in this area. The decline of traditional media in recent years has led to dwindling support for public interest reporting, which is vitally important in holding governments, corporations, and powerful individuals accountable to society. Meanwhile, with the current movement of "democratizing data," data-driven fact-checking and lead-finding are growing in importance. Taking advantage of data availability, this project helps reduce cost, increase effectiveness, and broaden participation for journalism, by putting practical tools in the hands of journalists and citizens alike.