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 3 Project Report

Note: Each response is 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]

Database research has focused on how to answer queries, but has not devoted much attention to discerning subtle qualities of the resulting claims, e.g., is a claim "cherry-picking"? We have been working on a framework that models claims based on structured data as parameterized queries. A key insight is *perturbation analysis*: we can learn a lot by perturbing a claim's parameters and seeing how its conclusion changes—basically by examining its QRS. This framework lets us formulate fact-checking tasks—reverse-engineering vague claims, and countering questionable claims—as computational problems. Based on this research, we have been developing a system called *iCheck*, which fact-checks claims about sports statistics, congressional voting records, publication records, etc.

In Year 3, we continued to investigate useful ways to present results from a complex QRS. Simply returning results with the highest strengths may not be useful, because many results can be similar. We want to pick results that are not only strong, but also diverse and able to provide good coverage for relevant portions of QRS. We have devised novel problem formulations and efficient algorithms for two scenarios. One scenario aims at selecting a small number of points to represent the high-strength regions of QRS. The second scenario is more specific: here QRS is the result of a multi-dimensional group-by aggregation query with categorical dimensional attributes, and the goal is to summarize the top result elements with a small number of clusters defined by multi-dimensional equality selection predicates.

In Year 3, we also discovered a deep connection between perturbation analysis and the so-called "iceberg" queries. Intuitively, both involve going through lots of parameter settings (or group-by attribute values in the case of iceberg queries), often in search for a few that stand out. Previous work on iceberg queries, however, did not deal with complex joins. With our observation, we developed a framework for combining a number of techniques such as a-priori and pruning, which had only been applied in separate, specific contexts. The framework applies to general SQL queries, exposing previously unexplored optimization opportunities. Encouraged by these results, we are now investigating the problem of approximately computing summaries of the QRS (an instance of which is estimating the result size of an iceberg query, or checking whether it is below a certain threshold), with the goal of handling general SQL queries.

We continued our investigation started in Year 2 on assessing a claim's "durability," the maximum possible time interval over which the claim remains true. We have also relaxed the notion of durability to allow a claim to be

temporarily false. Specifically, we have been working on efficient methods for finding objects in a time-series database that rank among the top *k* for a significant fraction of the time (i.e., higher than a given threshold) during a query interval.

Continuing our development of *ClaimBuster*, a primitive end-to-end fact-checking system is constructed and deployed in Year 3. The initial component of *ClaimBuster* which identifies important factual claims from text is now called the *ClaimBuster* Claim Spotter. We further developed several more components, including Claim Matcher, Claim Checker, and Fact-check Reporter. The claim matcher finds existing fact-checks that are closely-related or identical to the discovered claims. When a matching fact-check cannot be found, the claim checker queries external knowledge bases and the Web to vet the factual claims. The fact-check reporter compiles the evidence from the claim matcher and the claim checker, and presents fact-check reports to users through various channels, such as the project website, its Twitter account, a Slackbot, and a public API.

[Computational Lead-Finding]

Lead-finding involves looking for strong points in the QRS relevant to the context of interest. We continued to work on various problems in this area, such as ways to present results from a complex QRS, as described earlier, and implementing our results in *iCheck* and *FactWatcher*. In Year 3, we developed an instance of *iCheck* for the Duke University Athletic Department. It automatically finds interesting "factlets" about each Duke men's basketball game, which can be then shared on social media. The system is directly connected to the Athletic Department's game stats database and its factlets have become an integral part of the official stats website accessible to Duke fans.

Furthermore, we worked on extending our algorithms for finding prominent situational facts to deal with negative dimensions. During an ongoing event, values are accumulated in the opposite direction of users' preference on such dimensions. A prominent situational fact formed during the event may thus become less prominent. This makes discovering and maintaining such facts more challenging.

In another line of work, we have studied the problem of summarizing the input data for answering queries efficiently. In many instances, there is no single best answer, and often a very large number of incomparable objects satisfy the user's query. A regret minimizing set Q is a small-size representation of a much larger

database P so that user queries executed on Q return answers whose scores are not much worse than those on the entire dataset. A k-regret minimizing set has the property that the regret ratio between the score of the top-1 item in Q and the score of the top-k item in P is minimized. The problem is challenging because the goal is to find a representative set Q whose regret ratio is small with respect to all queries.

Another aspect of lead-finding addresses how to find the claims to check in the first place. The Claim Spotter in *ClaimBuster* uses supervised learning to find factual claims worthy of checking from political debates and social media. In Year 3 we investigated how the size of training data impacts the performance of *ClaimBuster*'s scoring model. We also investigated the impact of using named entities as features in the scoring model. Furthermore, we experimented with a few deep learning approaches for building the scoring model.

In a related effort (not supported by this grant), we also began to develop tools that will take transcripts of cable news programs and state legislative debates and submit them to *ClaimBuster* where they can be scored. The highest rated claims, which are the most value, will then be emailed to fact-checkers.

[System Support]

We have completed the *Perada* system for parallel perturbation analysis of database queries. In Year 3, we primarily worked on overcoming a limitation of *Perada*, namely its reliance on developers to specify memoization and pruning opportunities. This work led to our investigation of automatic optimization of iceberg queries over complex joins described earlier. We have implemented the optimization techniques in PostgreSQL, in order to validate support for general SQL queries and the ease of adoption by existing database systems. If implemented in *Perada*, these techniques would further simplify the development of general perturbation analysis in a cluster computing setting.

As discussed earlier, we have begun working on approximate computation of the summaries of the QRS. Following our approach to supporting iceberg queries, we plan to implement our techniques in PostgreSQL as well.

[Coping with Uncertainty]

In Year 3, we continued to work on "targeted" data cleaning for fact-checking, which selects data items to clean under a budget with the goal of checking a

claim. Intuitively, one should prioritize efforts towards those parts of data that "matter the most" to fact-checking. Besides formalizing the optimization problem and developing efficient algorithms with good approximation guarantees, we considered the interesting question of how the choice of the optimization objective (i.e., what it means to "matter the most") affects fact-checking. One reasonable objective is to minimize the uncertainty in a numeric measure of the claim's quality—intuitively, the goal here is to assess the claim. Another possibility is to maximize the probability that we can find a counterargument to the given claim—intuitively, the goal here purely is to counter the claim. In general, these two goals may not align, which implies that we need better data cleaning guidelines for fact-checkers to avoid potential biases introduced by the desire to counter claims.

In some scenarios, data cleaning does not lead to completely certain data; instead, it results in data with uncertainty captured by probability distributions (e.g., if one replaces erroneous and missing values using multiple imputation). In Year 3, we continued to work on the problem of lead-finding and query-answering over uncertain data. A specific instance of the problem arose from finding factlets about player performance from Duke men's basketball stats database: we were interested in the most likely skyline (records not dominated by others) in the presence of some uncertain records.

Besides perturbing parameter settings of a query, the dataset itself can be perturbed too—this interesting perspective is motivated by data uncertainty, which is common and costly to resolve. We continued our study on how data uncertainty affects query answer (and hence claim quality). We focused on the range-sum problem under uncertainty—calculating the probability that the sum of weights inside a query range is at least W, or more generally, compute the distribution of this sum approximately.

[Support for Interactive Analysis]

In Years 1 and 2, our work on enhancing fact-checking and lead-finding with interactive analysis focused primarily on visualization. In Year 1, we worked on generating 2-d visualizations combining heat maps and scatterplots, using efficient sampling-based techniques. In Year 2, we implemented interactive data visualization for *iCheck* on congressional voting records, and started exploring automatic optimization techniques for database-backed visualizations.

In Year 3, we put on hold work on automatic optimization of database-backed visualizations, in order to focus on another problem that arose in interactively exploring multi-dimensional group-by aggregation query results using summary clusters, as discussed earlier. Supporting interactivity introduces several challenges. First, computation of summary clusters requires setting several parameters, but users typically do not know the appropriate settings in advance. Therefore, our system would preview the space of possibilities upfront and suggest interesting settings for users to explore. Also, small changes to parameter settings (through a user-interface element such as a slider) should be handled at interactive speed. Finally, after changing parameter settings, we would like to present the differences in clustering results visually to users. Choosing the best visualization involves a non-trivial problem of picking the optimal layout of visual elements to highlight strong relationships and minimize clutter.

[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 objective is to investigate approaches that can improve *ClaimBuster*'s accuracy in spotting check-worthy claims.

[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.

[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 or lead-finding.

[Support for Interactive Analysis] Our objective is to develop fast algorithms to support various visualizations and explorations of large datasets at interactive speeds.

[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]

In earlier years of this project, we developed the QRS-based framework that formulates fact-checking tasks as computational ones, which laid the foundation for much of our subsequent work. The results were published in PVLDB 2014 and an extended journal version was published in TODS 2017. The system was demonstrated in SIGMOD 2014 and the 2014 Computational+Journalism Symposium. In Year 3, we built *iCheck* for U.S. congressional voting records (http://icheckuclaim.org). It was demonstrated at the 2016 Computational+Journalism Symposium and released to the public in September 2016. The website analyzes the voting records from January 2009 to September 2016, and lets users compare how legislators vote with party majorities and the President, and more importantly, explore how the comparison stacks up under different contexts—over time, among groups of peers, and for "key votes" identified by lobbying/political organizations.

For the problem of selecting strong, representative results to return from a QRS, we have developed a novel, clustering-inspired problem formulation, which allows users to specify soft preferences that better accommodate varying local

features of the QRS. We gave an efficient greedy algorithm providing constant-factor approximation. This result was published in PVLDB 2016 and will be presented in VLDB 2017. For the scenario of summarizing multi-dimensional aggregate results, we have developed a problem formulation that is also clustering-based and takes both coverage and diversity into consideration, but restricts clusters to be defined by equality selection predicates. We obtained complexity results for the optimization problem and proposed efficient algorithms at interactive speeds (more on these results later under [Support for Interactive Analysis]). We are currently revising a paper and a demonstration proposal for submission.

Our observation of the deep connection between perturbation analysis and iceberg queries led to the development of an automatic optimization framework for iceberg queries over complex joins, allowing us to combine techniques such as a-priori and pruning in novel ways. Implementation in PostgreSQL demonstrated the feasibility for existing systems to incorporate our techniques, as well as significant performance gains over baseline approaches. This result was published in SIGMOD 2017.

On the problem of finding "durable" objects from a time-series database (i.e., those that rank among the top k for at least a given fraction of the time during a query interval), we have developed both exact and approximation algorithms. Experiments on real-life data show that our approximation algorithm is efficient in both space and time—it works by intelligently and selectively precomputing a data structure within a storage budget to minimize run-time error. We are currently preparing a manuscript for submission.

The *ClaimBuster* system is hosted at http://idir.uta.edu/ClaimBuster and its features are being constantly expanded. The claim monitor interfaces various data sources (social media, broadcasted TV programs, and websites) with *ClaimBuster*. The claim spotter identifies check-worthy factual claims in verbose text from the data sources. The claim matcher finds existing fact-checks that are closely-related or identical to the discovered claims. In this way, we fully leverage well-researched fact-checks from professional fact-checkers. This is particularly useful, because oftentimes the same false claims are repeated. When a matching fact-check cannot be found, the claim checker queries external knowledge bases and the Web to vet the factual claims. The fact-check reporter compiles the evidence from the claim matcher and the claim checker, and presents fact- check reports to users through various channels, such as the project website, its Twitter account, a Slackbot, and a public API. The Slackbot has been

developed for users to supply their own text (i.e., directly as input or through a shared Dropbox folder) and receive the claim spotter score and fact-check report for that piece of text. The Slackbot has been published in the public Slack App directory and can also be installed by clicking the "ClaimBuster Slackbot" button on the project's website. We also made available a public *ClaimBuster* API to allow developers create their own fact-checking applications.

[Computational Lead-Finding]

To find leads automatically from data, we developed efficient algorithms for many claim templates and demonstrated their applications at SIGMOD 2014, VLDB 2014, and the 2014 Computation+Journalism Symposium. The VLDB 2014 *FactWatcher* demo won the best demo award. In Year 3, we have developed a system for finding "factlets" about each Duke men's basketball game. Our system runs after each game as soon as official stats become ready, and is able to produce a variety of interesting factlets within minutes. We have also run this system retroactively for Duke games in recent years. As of May 2017, these factlets have become a part of the official Duke men's basketball stats website, and are available for fans to explore and share on social media.

We investigated the performance of *ClaimBuster*'s scoring model under various dataset sizes (4,000, 8,000, ..., 20,000 sentences). We observed that the performance of SVM remained stable when dataset size was increased whereas the performance of NBC got better. This can be explained by how SVM and NBC work. SVM may have already discovered the decision boundary of the problem space with 4,000 − 8,000 training instances. Hence, more training instances afterwards did not change the boundary much. On the other hand, NBC kept updating the conditional probabilities when more training data became available. We also investigated the impact of using named entities as features in the scoring model. The results using various methods suggest that using named entities does not improve the model's accuracy. Furthermore, we also experimented with a few deep learning approaches for building the scoring model. No clear performance improvement was observed. We will continue to deepen our understanding and analyses.

The approach to selecting representative results to return from a QRS, discussed earlier, can be applied to find leads automatically from a dataset. In another line of work, we show that the k-regret minimization problem, described earlier, is NP-Complete for all dimensions $d \geq 3$. On the positive side, we developed two approximation algorithms for computing a small size k-regret minimizing set (k-

RMS), both with provable guarantees. The first one is guaranteed to compute a k-RMS of a small size. The other one computes an k-RMS whose size is within log-factor of the optimal k-RMS. We perform extensive experimental evaluation of our algorithms, using both real world and synthetic data, and compare their performance against the previously best known methods for this problem. The results show that our algorithms are significantly faster and scalable to much larger sets than the previous algorithms.

[System Support]

We have finished building *Perada*, a system aimed at reducing development and execution costs of perturbation analysis of SQL queries. It hides low-level 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 execution and dynamically adjusts the optimization knobs. This work has been published in PVLDB 2016 and will be presented in VLDB 2017. In Year 3, we also worked on overcoming a limitation of *Perada*, namely its reliance on developers to specify optimization opportunities. This work led to our investigation of automatic optimization of iceberg queries over complex joins, which was published in SIGMOD 2017 as described earlier.

[Coping with Uncertainty]

For the targeted data-cleaning problem, we have developed efficient algorithms for determining what data items to clean in order to reduce the uncertainty in some numeric measure of the given claim's quality (such as fairness, uniqueness, and robustness), or to maximize the probability of finding a counterargument to the given claim. For linear queries, we show how to solve the optimization problem under different objectives, drawing techniques from stochastic knapsack and submodular function optimization problems to obtain pseudo-polynomial or polynomial-time approximation algorithms with theoretical guarantees. We also empirically compare these algorithms with simpler greedy algorithms. We are currently preparing a manuscript for submission.

For the problem of lead-finding and query-answering over uncertain data, we investigated the problem of most likely skyline on probabilistic data. Under the unipoint model of uncertainty, we provide a polynomial-time algorithm for the case of 2-d points, and conjecture that the problem becomes NP-hard in higher dimensions. Under the multipoint model, we show that the problem is NP-hard

even in 2-d. We also develop Monte Carlo approximation algorithms for both models. We plan to further improve our results and prepare a manuscript for submission.

We developed the first data structure for answering range-sum queries under uncertainty. Specifically, given a set of weighted "uncertain" points, each with a probability of existence, we build an index that given a query rectangle and a weight W, returns the probability of the sum of the weights of points inside the query rectangle being at least W. Computing such a probability exactly is intractable, so we focus on estimating this quantity within a given error.

[Support for Interactive Exploration]

We have worked on efficient generation of 2-d visualizations combining heat maps and scatterplots (published in PVLDB 2015), and interactive data visualization for *iCheck* on congressional voting records (available to the public at http://icheckuclaim.org/ since September 2016). In Year 3, we focused on techniques for interactively exploring multi-dimensional group-by aggregation query results using summary clusters. To help users choose parameters for the clustering problem and to meet the stringent real-time performance requirement, we have developed efficient precomputation and incremental computation techniques for solving the clustering problem. To help users visualize how clusters change in response to parameter settings, we find the best visualization layout by solving an optimization problem; we have explored how different types of layouts affect the optimization complexity and hence their suitability for interactive exploration. We are currently preparing a manuscript for submission.

[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]

A paper describing our vision of automated fact-checking appeared in 2015 Computation+Journalism Symposium.

Our QRS-based framework was published in PVLDB 2014 and an extended version was published in TODS 2017. A system called *iCheck*, which fact-checks claims about sports statistics, congressional voting records, publication records, etc., was demonstrated in SIGMOD 2014 and the 2014 Computational+Journalism Symposium. An instance of *iCheck* for congressional voting records (http://icheckuclaim.org/) was made public in September 2016, and demonstrated at the 2016 Computational+Journalism Symposium.

Our results on selecting strong, representative results to return from a QRS were published in PVLDB 2016 and will be presented in VLDB 2017. Our results on automatically optimizing iceberg queries over complex joins were published in SIGMOD 2017.

We submitted two conference papers on the *ClaimBuster* system. Both papers have been accepted and will be presented in August. The KDD 2017 paper describes the system architecture, design of the components, and the evaluation. The VLDB 2017 demonstration paper explains the user-facing features of the system and describes a demonstration scenario.

[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-finding 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 about Duke men's basketball stats has been up since May 2017 and supplying interesting factlets about each game for the official Duke men's basketball stats website.

[System Support]

We have finished building *Perada*, and the work has been published in PVLDB 2016 and will be presented in VLDB 2017. We also have developed techniques to support automatic identification of optimization opportunities (in the context of optimizing iceberg queries over complex joins, published in SIGMOD 2017); these were implemented and evaluated in PostgreSQL, although they should be applicable to *Perada* as well.

[Coping with Uncertainty]

Our conference paper on range-max queries was invited to a special issue of Journal of Computer Science and Systems (JCSS). We are currently preparing submissions based on our results on targeted data for fact-checking, and on computing the most likely skyline over uncertain data.

[Support for Interactive Exploration]

Our results on efficiently generating approximate heat map and scatterplot visualizations were published in PVLDB 2015. We have implemented various visualizations for *iCheck* on congressional voting records, and they have been available to the public at http://icheckuclaim.org/ since September 2016. Our results on interactive exploration of multi-dimensional group-by aggregation query results are currently being prepared for submission.

[Economics of Investigative Journalism]

Hamilton finished his book on the economics of investigative reporting, *Democracy's Detectives: The Economics of Investigative Journalism*, which was published in September 2016 by Harvard University Press (http://www.hup.harvard.edu/catalog.php?isbn=9780674545502). The last chapter, "Accountability and Algorithms," relates to how computational techniques can help with investigative journalism tasks such as fact-checking and lead-finding, and is described at http://cjlab.stanford.edu/democracys-detectives-jay-hamilton/.

[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 has provided training for 6 PhD students and 10 undergraduate students in Computer Science. It has also trained a volunteer researcher at Duke as well as two students from local high schools. In the the Duke journalism program, two students have been involved in developing tools for fact-checkers or consumer. At UT Arlington, 3 PhD students, 4 MS students, and 2 undergraduate students have participated in the project, including three in

traditionally underrepresented categories (one female and two Hispanic students). Aspects of this project have been integrated into various computer science courses that Yang, Agarwal, 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 2014, and the Tech & Check conference in 2016 and 2017, where journalists, computer scientists, and researchers from public policy participated.

How have the results been disseminated to communities of interest?

For dissemination efforts in Years 1 and 2 of the project, please refer to our earlier reports. This report focuses on Year 3. 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 2016-2017, Yang and Walenz (a senior PhD student working on the project) gave about the project at the Duke Visualization Forum, the seminar series for summer undergraduate researchers at Duke, and a meeting of the Duke Information Technology Advisory Council. In May 2017 Li gave a talk about *ClaimBuster* at Google, hosted by Cong Yu. Agarwal has given a number of invited talks that covered work done in this project.

The project team has also been keen on disseminating the results beyond the computer science community. Project team members from Duke, Stanford, and UT participated in the Computation+Journalism Symposium in 2014, 2015, and 2016, which attracted both journalists and computer scientists. The team presented papers describing the project and gave demos of *iCheck*, *FactWatcher*, and *ClaimBuster*.

At Duke, following the success of the first Tech & Check conference in March 2016, Adair organized a second conference in January 2017 with the co-PIs of this project as well as representatives from Google, the New York Times and the Internet Archive.

At Stanford, Hamilton co-hosted the 2016 Computation+Journalism Symposium, Sept. 30 – Oct. 1, 2016 (see http://journalism.stanford.edu/cj2016/).

Adair made presentations about the project at Global Fact 3, the annual meeting of the world's fact-checkers in Buenos Aires in July 2016; in a presentation at Virginia Wesleyan College, "The Truth-O-Meter, Pants on Fire and Fact-Checking the 2016 Campaign," in Norfolk, Va., in October 2016; in a speech to

Duke Science & Society in Durham, N.C., in October, 2016; and in a panel presentation called "Robots that report: How fact-checkers worldwide are experimenting with automation" at the conference of the Investigative Reporters and Editors in New Orleans, La. in June 2016.

Adair organized and moderated an SXSW (South by Southwest) 2017 panel on "How Bots Are Automating Fact-Checking." The panel took place at March 13, 2017 in Austin, TX. Li participated in the event as a panelist.

At Duke, Yang participated in the Social Innovation and Entrepreneurship Faculty Fellowship program in 2016-17, for help with broadening the impact of the project.

Several of the systems developed by this project are now available to the public. *iCheck* for congressional voting records (http://icheckuclaim.org/) was released in September 2016. In October 2016, Yang and Walenz gave a Science Cafe presentation to the general public at the North Carolina Museum of Natural Sciences. An instance of *iCheck* has also been deployed on Duke men's basketball stats website since May 2017. The *ClaimBuster* website (http://idir.uta.edu/claimbuster) is under continuous improvement and expansion. We recently have released a few new features on the website, including a public API, a Slackbot, a factual claim taxonomy, and a primitive end-to-end fact-checking system.

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 continue our progress on assessing claim durability and automatically optimizing general perturbation analysis. We also expect to make new progress on approximate computation of QRS summaries. We will continue to improve our two public-facing *iCheck* systems (for congressional voting records and for Duke men's basketball stats), which will serve as testbeds for our ideas and techniques.

We will continue the development of *ClaimBuster* in several directions. We will further polish the factual claim taxonomy and work on automatic classification of claims based on the taxonomy. We will also work on improving the accuracy of the claim spotter's scoring model by using neural networks.

Along the line of research related to data uncertainty, we continue to deepen our understanding of the effect of data uncertainty on fact-checking and querying in general. We expect to wrap up our work on targeted data cleaning for fact-checking, and on finding leads (such as the mostly likely skyline) from uncertain data.

We also expect to make more progress on how to support fast, interactive data exploration for lead-finding and fact-checking, where interaction can be seen as a form of perturbation. We will continue to extend and enhance our systems with interactive data visualization components.

Finally, as a conclusion to the project, we plan to collaborate on a retrospective on our experience of this project, as well as a survey and tutorial on fact-checking (primarily from the perspective of data management research).

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.

Our work received good coverage by the press (see project reports for earlier reports on coverage before July 2016):

- ”Post-truth v tech: could machines help us call out politicians’ and journalists’ lies?” NewStatesman. August 17, 2016. http://www.newstatesman.com/2016/08/post-truth-v-tech-could-machines-help-us-call-out-politicians-and-journalists-lies
- “Fail and move on: Lessons from automated fact-checking experiments.” Poynter. September 7, 2016. http://www.poynter.org/2016/fail-and-move-on-lessons-from-automated-fact-checking-experiments/429232/
- “Fact-checking Senate campaign ads just got easier.” *Duke Today*. September 29, 2016. https://today.duke.edu/2016/09/icheck
- “‘It makes them more honest’: New site provides data for fact-checkers to hold politicians accountable.” The Chronicle. October 12, 2016. http://www.dukechronicle.com/article/2016/10/it-makes-them-more-honest-new-site-provides-data-for-fact-checkers-to-hold-politicians-accountable
- “Fake news and fact-checking: Trump is demonstrating how to outsmart an AI.” The Guardian. January 31, 2017. https://www.theguardian.com/science/2017/jan/31/fake-news-and-fact-checking-trump-is-demonstrating-how-to-outsmart-an-ai-artificial-intelligence
- “How do you solve fake news problem in the post-truth era?” New Atlas. February 7, 2017. http://newatlas.com/how-to-fix-fake-news-problem/47800/
- “A Heroic AI Will Let You Spy on Your Lawmakers’ Every Word.” Wired. February 7, 2017. https://www.wired.com/2017/02/soon-bots-will-spying-state-lawmakers-every-move/
- “Digital Democracy: government transparency website keeps track of state legislators.” Plaid Zebra. March 6, 2017. http://www.theplaidzebra.com/government-transparency-website-keeps-track-state-legislators/
- “SXSW 2017: Embattled Journalism in Focus.” MediaShift. March 21, 2017. http://mediashift.org/2017/03/sxsw-2017-embattled-journalism-in-focus/
- “UTA-led interdisciplinary team to construct computer program to identify fake news.” University of Texas at Arlington. June 12, 2017. http://www.uta.edu/news/releases/2017/06/bots-fake-news-grant.php
- “How the world gets its facts straight.” Policy Options, Institute for Research on Public Policy. July 17, 2017. http://policyoptions.irpp.org/magazines/july-2017/how-the-world-gets-its-facts-straight/
- “Professors from UT-Arlington, UT-Dallas join forces to fight fake news.” The Dallas Morning News. July 20, 2017.

https://www.dallasnews.com/news/higher-education/2017/07/20/professors-ut-arlington-ut-dallas-join-forces-fight-fake-news

What is the impact on the development of human resources?

So far, the project has supported close training and development of 6 PhD students and 10 undergraduate students at Duke, as well as 3 PhD students, 4 MS students, and 2 undergraduate students at UT Arlington. You Wu graduated his PhD from Duke in the summer of 2015 and joined Google Research. Naeemul Hassan graduated with his PhD from UT Arlington in the summer of 2016 and joined University of Mississippi as assistant professor.

We have revamped the undergraduate computer science curriculum at Duke and UT Arlington to incorporate elements of this research. At Duke, Yang taught a new graduate course on data cleaning in Spring 2017, which covered latest research on data cleaning and had projects investigating the interaction between data cleaning and lead-finding.

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, we have made *iCheck* for congressional voting records available to the public as of September 2016, and *iCheck* for Duke men's basketball stats in May 2017.

What is the impact on technology transfer?

The project team has collaborated closely with Cong Yu at Google. Jun Yang and Chengkai Li both visited Google NYC (during May 21-30, 2016, and June 12-23, 2017, respectively). You Wu, a Duke PhD student working on this project, interned at Google for two summers, and has joined Cong Yu's team at Google

since fall 2015. Through these connections, we are hopeful that some results from this project will be incorporated into Google/Jigsaw's commercial offerings.

A number of media and fact-checking organizations have expressed interests in our research, and we are actively engaging them with visits, presentations, and collaborations. 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.