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 2 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 publication records, baseball statistics, and congressional voting records.

In Year 2, we continued to improve our models and algorithms. One line of work focuses on selecting a few 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. Previous work on result diversity requires user-specified parameters that are unable to account for variations within a complex QRS, and the results do not necessarily represent the overall QRS. To overcome these shortcomings, we have devised a novel problem formulation and efficient algorithms.

Another investigation we started in Year 2 is how to assess a claim's "durability," the maximum possible time interval over which the claim remains true. The longer the durability, the better the claim. We have been studying how to compute durability and find durable claims efficiently for different claim types, in hope of developing a general algorithmic framework for durability.

Furthermore, we have begun working towards the vision of an end-to-end fact-checking system, called *ClaimBuster*, which automatically detects textual claims and checks them. We are putting together state-of-the-art technologies in question generation, question answering, and knowledge bases. A preliminary version of this system will help us understand the limitations of current methods; insights gained in these areas will prepare us for breakthroughs toward automated fact-checking.

[Computational Lead-Finding]

Lead-finding involves looking for strong points in the QRS relevant to the context of interest. We worked on a number of problems in this area, including discovering prominent streaks from time series, finding one-of-few claims from multidimensional data, incrementally computing prominent situational facts,

and mining frequent episodes online. We implemented and demonstrated these algorithms in *iCheck* and *FactWatcher*.

Another aspect of lead-finding addresses how to find the claims to check in the first place. We applied machine learning and NLP for automatic discovery of factual claims worthy of checking from political debates and social media. We built a website to collect labeled examples from all US presidential debates. We also developed a system that, as part of *ClaimBuster*, applied the trained model to detect and highlight sentences worthy of checking from text. In Year 2, we focused on extending the system to more sources, including closed captions of live debates, interviews and speeches, transcripts of US presidential debates and Australian Parliamentary debates ("Hansard"), as well as Twitter tweets. We also collected more labeled training data and enhanced the website with visualization and social features.

[System Support]

We developed a system called *Perada*, for parallel perturbation analysis of database queries. *Perada* simplifies the development of general, ad hoc perturbation analysis by providing a flexible API to support a variety of optimizations such as grouping, memoization, and pruning; by automatically optimizing performance through run-time observation, learning, and adaptation; and by hiding the complexity of concurrency and failures from its developers. We demonstrate *Perada*'s efficacy and efficiency with real workloads in computational journalism. Although *Perada* has greatly simplified implementation and tuning, it is still not fully declarative. To further improve usability, we are working on automatic derivation of memoization and pruning opportunities from declarative specifications.

[Visualization Support]

Visualization is a powerful aid to perturbation analysis. In Year 1, in collaboration with Google, we worked on 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 (for the overall distribution) and a scatterplot (for the outliers). The challenge is that computing the points to visualize from raw data may be slow. We devised sampling-based algorithms with as little loss in fidelity as possible given a computation budget.

In Year 2, we continued to improve *iCheck*, extending it with interactive data visualization and exploration. Developing interactive visualization is hard. While a database offers a declarative way to specify views of data for visualization, a naive implementation has poor performance and is unsuitable for interactive use. So far, we have hand-optimized the division of work between the database server and the visualization client, which required substantial effort and expertise. We are starting to investigate methods for automatically optimizing declarative data visualization code, so that it can deliver good performance with much less developer effort. Exploiting limits of human perception and predictability of interactions, we plan to apply techniques such as precomputation, group and incremental processing, and approximation to speed up visualization queries, decrease server load, and reduce network communication.

[Coping with Uncertainty]

Besides perturbing parameter settings of a query, the dataset itself can be perturbed too—this interesting perspective is motived by data uncertainty, which is common and costly to resolve. We are interested in understanding how data uncertainty affects query answer (and hence claim quality). The first problem we studied is range-max—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 a query hyper-rectangle. We are also studying the range-sum problem under uncertainty—what is 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.

To help fact-checking involving uncertain data, we are also working on "targeted data cleaning," which selects data items to clean under a budget with the goal of finding strong counterarguments to a given claim. We are given a space of possible counterarguments and a distribution of possible outcomes for 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 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. We model

the above problems as budgeted stochastic optimization problems and find techniques to guarantee good approximate 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 various visualizations that are perceptually accurate at interactive speeds from large datasets.

[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 the QRS-based framework that formulates 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 optimization problems. We designed an architecture with "pay-as-you-go" support for efficient fact-checking. New claims can be supported by baseline algorithms with little work upfront. With better knowledge of the structure of the parameter space and/or query template, users can supply building blocks that can achieve higher efficiency when plugged into our generic meta-algorithms. These results were published in PVLDB 2014 and a journal version is under review. The system was demonstrated in SIGMOD 2014 and the 2014 Computational+Journalism Symposium.

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. In theory, *k*-means-like algorithms give optimal solutions for our utility function if run with many initial states. Because our utility function is submodular, 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. We are currently preparing this work for submission.

We have built a preliminary version of *ClaimBuster* by integrating a question generator and a question answering (QA) system, specifically WolframAlpha. The question generator converts a natural language sentence into a question,

which is fed to the QA system. We compare the answer from the QA system with the answer extracted from the original sentence, to see if the claim checks out. We prepared four different datasets: claims in Twitter tweets, fact-checks from PolitiFact.com, sports records from ESPN's *Elias Says*, and Wikipedia sentences. The preliminary *ClaimBuster* exposed several limitations. While the question generator produces reasonable questions for a large percentage of factual claims, it misses many complex claims that are common; the QA system answers almost none of the questions, due to lack of data and poor understanding of the questions. We are gaining insights on how to improve it. We are also conducting a comprehensive survey of real-world claim templates, which will help us develop template-specific and category-specific enhancements to question generation and QA, as well as techniques for checking claims using structured queries over databases and knowledge graphs.

[Computational Lead-Finding]

To find leads automatically from data, we developed efficient algorithms for many claim templates, including discovering prominent streaks, finding one-of-few claims, incrementally computing prominent situational facts, and mining frequent episodes online. Implementations of these algorithms were demonstrated at SIGMOD 2014, VLDB 2014, and the 2014 Computation+Journalism Symposium. The VLDB 2014 *FactWatcher* demo won the best demo award.

To find claims worth checking, we have accomplished about 80% of our data collection objective. We extracted a total of 20,788 sentences during 30 debates in the past 11 presidential elections, and we have recruited 350+ coders to participate in labeling. We evaluated several supervised learning methods. Thousands of sentence features (e.g., length, words, sentiments, part-of-speech tags) were extracted and important one were selected. Preliminary 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. These results were presented at CIKM 2015 and the 2015 Computation+Journalism Symposium.

[System Support]

We have finished building *Perada*, a system aimed at reducing development and execution costs of perturbation analysis of SQL queries. Currently, it relies on

users to provide a small number of optimization hooks, but 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. Experiments reveal that *Perada* produces substantial improvement in execution time over naive approaches, while keeping development complexity low. This work was submitted for publication in PVLDB and a revision is under review.

[Visualization Support]

For efficient approximate 2-d visualization, we have developed a two-stage sampling-based algorithm. In the first stage, query is performed on a sample of all objects' data. Based on the results, 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. Experiments show that our approach can generate high-quality visualizations much faster and using far fewer data accesses than evaluation on the full dataset. The results appeared in PVLDB 2015.

We also added many interactive visualization elements to *iCheck*. This system has become our testbed for investigating automatic optimization techniques for interactive data visualizations.

[Coping with Uncertainty]

We developed the first subquadratic space/sublinear query time index for finding the expected maximum and the most likely maximum value in any dimension. Under some natural assumptions, we significantly improved the performance of the index for the most likely maximum problem: if values are assigned to points randomly, we give a near-linear size index that returns the most likely maximum value in polylog(n) expected time. We also developed an approximation algorithm for computing the expected maximum value inside a query range. Using the prophet-inequality from stochastic optimization, we proposed a near linear index that finds a 1/2-approximation of the expected maximum value in polylog(n) time in any dimension. Furthermore, we proved the first nontrivial lower bound for the most likely maximum problem in 2 or higher dimensions, doing a reduction from the set intersection problem. Finally,

we generalized the above indexes to other models of uncertainty, like the location or value uncertainty model. The results appeared in PODS 2016.

For the targeted data-cleaning problem, we have developed efficient algorithms for determining data items to clean in order to maximize the probability of finding a counterargument. We map this problem to a variant of the stochastic knapsack problem and adapt some of the techniques thereof. We also consider the problem of finding data items to clean to minimize uncertainty on various measures of claim quality. For instance, to minimize the uncertainty in ***fairness***, we map our problem to a stochastic knapsack problem and give a pseudo-polynomial algorithm. To minimize the uncertainty in ***uniqueness*** and ***robustness***, we formulate the problem as minimizing a submodular non-decreasing function under a linear constraint. Using known results and taking advantage of the specific properties of our functions, we give a polynomial-time approximation algorithm with theoretical guarantees. We also have run experiments comparing our proposed algorithms with other simpler greedy techniques. We are preparing the results for publication.

[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. We are currently preparing *iCheck* on congressional voting records for a public release.

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

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 worthy checking from text, as part of *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 contributions from more than 350 coders. We have applied *ClaimBuster* live on every Democratic and Republican primary debate of the 2016 election. A paper on *ClaimBuster* was published in CIKM 2015 and a demo was presented at the 2015 Computation+Journalism Symposium.

[System Support]

The first iteration of *Perada* is functional. Our experiments show that it produces a substantial improvement in execution time over naive approaches, while keeping development complexity low. A paper describing *Perada* is currently under review for publication in PVLDB.

[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.

We have extended *iCheck* for congressional voting records with interactive visualization features, and we plan to release the system in the summer of 2016.

[Coping with Uncertainty]

We have developed a suite of algorithms for range-max query over uncertain data under different models and assumptions. Our results have been published in PODS 2016.

We are currently preparing for submission our results on cleaning uncertain data for the purpose of fact-checking.

[Economics of Investigative Journalism]

Hamilton finished his book on the economics of investigative reporting, *Democracy's Detectives: The Economics of Investigative Journalism*, which is due out in September from 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.

[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 7 undergraduate students in Computer Science. At UT Arlington, the project provided training for 4 PhD, 4 MS students, and 2 undergraduate students in Computer Science. 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 2014, and the Tech & Check conference in 2016, where journalists, computer scientists, and researchers from public policy participated.

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, the IEEE Computer Society Fort Worth section meeting, the Arlington Technology Association monthly meeting, Nanjing University, and Shandong University. 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. Project team members from Duke, Stanford, and UT participated in the Computation+Journalism Symposium in 2014 and 2015, which attracted both journalists and computer scientists. The team presented papers describing the project and gave demos of *iCheck*, *FactWatcher*, and *ClaimBuster*. 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?" On Dec. 3, 2015, Adair and Li participated in the American Press Institute's second Thought Leader Summit, "Fact-Checking the 2016 Elections: What's New and What's Next." On Feburary 25, 2016, Adair gave a presentation at the annual staff meeting of PolitiFact, a fact-checking organization, titled "The Quest to Automate Fact-Checking." At the Global Fact-Checking Summit (annual meeting of the world's fact-checkers, with about 100 attendees) in Buenos Aires, Argentina, June 9-10, 2016, Adair made presentation on automating fact-checking and described iCheck and ClaimBuster in detail; in the brainstorming sessions at the summit, participants recommended that our systems to be included in a more complete ecosystem. At the Investigative Reporters and Editors Conference, New Orleans, June 17, 2016, Adair participated in a panel called "Robots that report: How fact-checkers worldwide are experimenting with automation," and made a presentation about ClaimBuster and iCheck. In June

2016, Hamilton gave a talk on a panel at Columbia University's FOIA@50 Conference, entitled "What Difference Does FOIA Make? Some Noteworthy New Data" based on his forthcoming book.

At Duke, Adair organized the Tech & Check conference, March 31-April 1, 2016 (https://reporterslab.org/tech-check-new-ideas-automate-fact-checking/). The conference was co-hosted with Poynter's International Fact-Checking Network. Attendees came from around the world and included journalists, computer science faculty and students, as well as technologists from Google and IBM.

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. On May 13, 2016, Hamilton planned and co-hosted a day-long Workshop on Media Innovation and an evening panel discussion on Digital News Strategies (see http://www.125yearsofjournalism.org/ and http://news.stanford.edu/2016/05/18/innovation-amplifies-old-school-news-sense-unprecedented-journalistic-impact-stanford-alumni-panel-says/). Hamilton is co-hosting the 2016 Computation+Journalism Symposium, to be held at Stanford, Sept. 30 – Oct. 1, 2016.

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. On March 10, 2016, Adair visited UT Arlington and gave a talk on fact-checking in the Department of Communication, hosted by Mark Tremayne, the team's collaborator and Assistant Professor in Broadcast Communication at UT Arlington. Adair, Li, Tremayne and their students exchanged ideas related to the project. In April 2016, Yang paid an extended visit to Cong Yu, the team's long-time collaborator, at Google NYC. During this trip he visited Reg Chua at Thompson Reuters, Sarah Cohen at the New York Times, and Mark Hansen at Columbia University and Brown Institute for Media Innovation. All these activities provided opportunities to 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 claim durability assessment. Another big push will be improving the generality and usability of *Perada*, which supports a wider range of ad hoc perturbation analysis.

We will continue the development of *ClaimBuster* in several directions. We will conduct a comprehensive survey of the categories and templates of real-world factual claims from various domains (such as political claims, sports, finance, common knowledge, and Twitter data in general). We will investigate category-specific and template-specific approaches to question generation, question answering, and structured query generation, for automated fact-checking. We will continue to collect labeled data for the problem of finding check-worthy

claims, improve our machine learning algorithms, and enhance the *ClaimBuster* website with visualization and social features.

For uncertainty support, while we continue to deepen our understanding of the effect of data uncertainty on query results, we will complete our work on 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. Next, we will also study the impact of the uncertainty in query parameters on fact checking and lead finding.

Finally, we expect to make initial progress on how to support fast, interactive data visualization 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.

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:

- "Is that a fact? Checking politicians' statements just got a whole lot easier." *The Guardian*. April 18, 2016. http://www.theguardian.com/commentisfree/2016/apr/19/is-that-a-fact-checking-politiciansstatements-just-got-a-whole-lot-easier/
- "The Future of Political Fact-Checking / New fact-checking projects use advanced technologies to automate and accelerate the process." *Nieman Reports*, Harvard. March 23, 2016. http://niemanreports.org/articles/the-future-of-political-fact-checking/
- "Automated Fact-checking: The Holy Grail of Political Communication." *Nordic APIs*. February 25, 2016. http://nordicapis.com/automated-fact-checking-the-holy-grail-of-political-communication/
- "New Software Developed at UTA Tracks Candidates Statements." *NBC 5 News*. January 7, 2016. http://www.nbcdfw.com/news/politics/New-Software-Developed-at-UTA-Tracks-Candidates-Statements-364565851.html
- "In search of fact checking's 'Holy Grail': News outlets might not get there alone." *Medium.com*. October 30, 2015. https://medium.com/1st-draft/automated-fact-checking-and-verification-will-only-happen-if-organizations-other-than-newsrooms-can-84cf40689924/
- "The 'Holy Grail' of computational fact checking – and what we can do in the meantime." *Poynter*. October 21, 2015. http://www.poynter.org/2015/the-holy-grail-of-computational-fact-checking-and-what-we-can-do-in-the-meantime/379687/
- "Sifting balderdash from truth gets a boost from computers." *Austin American-Statesman*. August 8, 2015. http://www.mystatesman.com/news/news/meet-the-robots-that-fact-check/nnCxy/
- "Getting It Right: Fact-Checking in the Digital Age." *American Journalism Review*. April 21, 2015. http://ajr.org/2015/04/21/fact-checking-tools/

What is the impact on the development of human resources?

So far, the project has supported close training and development of 5 PhD students and 7 undergraduate students at Duke, as well as 4 PhD students, 4 MS students, and 2 undergraduate students at UT Arlington. Yu Wu graduated his PhD from Duke in the summer of 2015. Naeemul Hassan graduated with his PhD from UT Arlington in the summer of 2016.

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, we are in the process of making *iCheck* for congressional voting records available to the public. 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. Jun Yang visited Google NYC during May 21-30, 2016. You Wu, a 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.

Chengkai Li organized a team that participated in the NSF I-Corps program from October 2015 to December 2015. The team members are Naeemul Hassan and Gensheng Zhang (Entrepreneurial Leads; both are UT Arlington PHD students), Chengkai Li (Principal Investigator) and Harold Strong (Mentor; Co-founder of Inovatx Consulting Group). In the program, the team conducted market research and explored commercialization opportunities for *ClaimBuster* through participating in workshops organized by the I-Corps training staff, interviewing potential customers, and attending industrial conferences and meetings.

A number of media and fact-checking organizations have expressed 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.