# List of Suggested Reviewers or Reviewers Not To Include (optional)

## SUGGESTED REVIEWERS:

**Not Listed**

## REVIEWERS NOT TO INCLUDE:

**Not Listed**

---- List of Project Personnel and Partner Institutions ----

1. Oliver Kennedy; University at Buffalo; PI
2. Boris Glavic; Illinois Institute of Technology; PI
3. Juliana Freire; New York University; PI
4. Heiko Mueller; New York University; Postdoc
5. Dieter Gawlick; Oracle; Unpaid Collaborator
6. Zhen Hua Liu; Oracle; Unpaid Collaborator
7. Julia Lane; NYU Center for Urban Science and Progress; Unpaid Collaborator
8. Rebecca Rosen; NYU Center for Urban Science and Progress; Unpaid Collaborator
9. Ingrid Gould Ellen; NYU Furman Center; Unpaid Collaborator
10. Rodney Stiles; NYC Taxi and Limousine Commission; Unpaid Collaborator
11. Ronny Fehling; Airbus group; Unpaid Collaborator

=== PI Kennedy's Collaborators and Other Affiliations ==

A. Recent Collaborators

1. Sumit Agarwal; Gannett
2. Yanif Ahmad; Johns Hopkins University
3. Jerry Antony Ajay; University at Buffalo
4. Daniel Bellinger; Global Foundries
5. Geoffrey Challen; University at Buffalo
6. Varun Chandola; University at Buffalo
7. Sharath Chandrashekhara; University at Buffalo
8. Jan Chomicki; University at Buffalo
9. Patrick Coonan; University at Buffalo
10. Nick DiRienzo; University at Buffalo
11. Ronny Fehling; Oracle
12. Dieter Gawlick; Oracle
13. Boris Glavic; Illinois Inst. Tech.
14. Zhen Hua-Liu; Oracle
15. Kyungho Jeon; University at Buffalo
16. Steven Y. Ko; University at Buffalo
17. Christoph Koch; EPFL
18. GÃ¶khan Kul; University at Buffalo
19. Steve Lee; Microsoft Corp.
20. Charles Loboz; Microsoft Corp.
21. Duc Thanh Luong; University at Buffalo
22. Daniel Lupei; EPFL
23. Anudipa Maiti; University at Buffalo
24. Shikhar Mehra; University at Buffalo
25. Niccolo Meneghetti; University at Buffalo
26. Arindam Nandi; University at Buffalo
27. Anandatirtha Nandugudi; University at Buffalo
28. Suman Nath; Microsoft Research
29. Hung Ngo; University at Buffalo
30. Xuanlong Nguyen; University of Michigan
31. Milos Nicolic; EPFL
32. Andres NÃ¶tzli; Stanford
33. Amir Shaikhana; EPFL
34. Sriram Shantharam; University at Buffalo
35. Feng Shen; University at Buffalo
36. Jinghao Shi; University at Buffalo
37. Slawek Smyl; Microsoft Corp.
38. Guru Prasad Srinivasa; University at Buffalo
39. Ankur Upadhyay; FactSet
40. Shambhu Upadhyaya; University at Buffalo
41. Ting Xie; University at Buffalo
42. Ying Yang; University at Buffalo
43. Lukasz Ziarek; University at Buffalo

B. PhD and Postdoctoral Advisors

1. Christoph Koch; EPFL

C. Thesis Advisor and Postgraduate-Scholar Sponsor

> Graduated (1): Ankur Upadhyay (MS; Software Engineer; FactSet)

> Current at SUNY Buffalo (9): Arindam Nandi (MS), Poonam Kumari (PhD), Ting Xie (PhD), Ying Yang (PhD), Aaron Huber (PhD), John Logan

(PhD), Niccolo Meneghetti (PhD, Co-Advised), Razie Fathi (PhD, Co-Advised), Gokhan Kul (PhD Co-Advised)

   > Total Advisees:
       > Undergraduate Researchers: 3
       > Graduate Students: 16
       > Postdoctoral Researchers: 0

# Collaborators and Other Affiliations
# Dr. Juliana Freire

- *Collaborators and Co-Editors*: Luciano Barbosa (IBM Research), Philippe Bonnet (University of Copenhagen, Denmark), Daniel Fink (Cornell University), Nivan Ferreira (University of Arizona), Yolanda Gil (USC), Ingrid Gould-Ellen (NYU), Carlos Heuser (UFRGS, Brazil), Bill Howe (University of Washington), Steve Kelling (Cornell U.), David Koop (University of Massachusetts-Dartmouth), Ari Juels (Cornell Tech), Bertram Ludascher (UC Davis), Viviane Moreira (UFRGS, Brazil), Lauro Lins (AT&T Research), Luc Moreau (University of Southhampton), Valerio Pascucci (University of Utah), Jorge Poco (University of Washington), Dennis Shasha (NYU), Altigran Silva (UFAM, Brazil), Cláudio Silva (NYU), Torsten Suel (NYU), Lena Stromback (Linkoping University, Sweden), Andrew Terrel (Continuum Analytics), Matthias Troyer (ETH Zurich), Huy Vo (CUNY), Chris Wood (Cornell U.).

- *Graduate Advisor:* David S. Warren (SUNY at Stony Brook).

- *Thesis Advisor (18) and Postgraduate-Scholar Sponsor (3):* Aline Bessa, Masayo Otta, Fernando Chirigati, Tuan-Anh Hoang Vu, Kien Pham, Felipe Horta, Raoni Louren{cco, Lauro Lins, Viviane Moreira, Thanh Nguyen, Ronaldo Mello, Luciano Barbosa, Emanuele Santos, David Koop, Hoa Nguyen, Huong Nguyen, Ramesh Pinnamaneni, Sumit Tandon, Lorena Carlo, Eun Yong Kang, Fang Du, Lingzhi Zhang, Haojun Wang.

# Boris Glavic

**Collaborators & Other Affiliations**

Collaborators and Co-Editors

| | |
|---|---|
| Gustavo Alonso | ETH Zurich, Switzerland |
| Periklis Andritsos | University of Lausanne, Switzerland |
| Patricia R. Arocena | University of Toronto |
| Adriane P. Chapman | The MITRE Corporation |
| Radu Ciucanu | University of Oxfort, UK |
| Sarah Cohen-Boulakia | Université Paris-Sud, France |
| Kyumars Sheykh Esmaili | Bell Laboratories, Belgium |
| Ronny Fehling | Airbus |
| Peter M. Fischer | Albert-Ludwigs-Univ. Freiburg, Germany |
| Ian T. Foster | University of Chicago |
| Juliana Freire | New York University |
| Dieter Gawlick | Oracle |
| Laura M. Haas | IBM Research |
| Oliver Kennedy | SUNY Buffalo |
| Sven Köhler | Google |
| Vasudha Krishnaswamy | Oracle |
| Zhen Hua Liu | Oracle |
| Tanu Malik | University of Chicago |
| Marta Mattoso | Federal University of Rio de Janeiro, Brazil |
| Giansalvatore Mecca | Universita della Basilicata, Italy |
| Renee J. Miller | University of Toronto, Canada |
| Paolo Papotti | Arizona State University |
| Quan Pham | unkown |
| Venkatesh Radhakrishnan | Facebook |
| Ioan Raicu | IIT |
| Alexander Rasin | DePaul University |
| Donatello Santoro | Universita della Basilicata, Italy |
| Nesime Tatbul | Intel Labs and MIT CSAIL |

Graduate Advisors and Postdoctoral Sponsors

| | |
|---|---|
| Michael H. Böhlen | University of Zurich, Switzerland |
| Gustavo Alonso | ETH Zurich, Switzerland |
| Renee J. Miller | University of Toronto, Canada |

Thesis Advisor and Postgraduate-Scholar Sponsor

| | |
|---|---|
| Bahareh Arab | IIT |

| | |
|---|---|
| Jason Arnold | IIT |
| Xing Niu | IIT |
| Seokki Lee | IIT |
| Yuchen Tang | IIT |
| Beat Steiger | Inventage AG |
| Alessandro Vagliardo | BCP - Neplan |
| Oliver Wirz | nag informatik |
| Svetlana Gerster | Credit Suisse |
| Stephan Blatti | unknown |
| Sascha Nedkoff | unknown |
| Phillip Hochstrasser | unknown |
| Claude Humard | ELCA Informatik AG |
| Tobias Schlaginhaufen | Swiss Re |
| Zhen Wang | Amazon |
| Andrea Cornudella | unknown |

# Collaborators and Other Affiliations
# Dr. Heiko Müller

- *Collaborators and Co-Editors*: Peter Buneman (University of Edinburgh, U.K.), James Cheney (University of Edinburgh, U.K.), Mat Cook (Agriculture, CSIRO), Ritaban Dutta (Data61, CSIRO), Johann-Christoph Freytag (Humboldt University Berlin, Germany), Ulf Leser (Humboldt University Berlin, Germany), Felix Naumann (Hasso Plattner Institute, Germany), Yanfeng Shu (Data61, CSIRO).

# COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./DUE DATE
**NSF 16-530** **04/04/16**

☐ Special Exception to Deadline Date Policy

**FOR NSF USE ONLY**

**NSF PROPOSAL NUMBER**

**1640864**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)
**ACI - DATANET**

| DATE RECEIVED | NUMBER OF COPIES | DIVISION ASSIGNED | FUND CODE | DUNS# (Data Universal Numbering System) | FILE LOCATION |
|---|---|---|---|---|---|
| 04/04/2016 | 1 | 05090000 ACI | 7726 | 038633251 | 04/04/2016 4:54pm |

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)
**141368361**

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE
**SUNY at Buffalo**

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE
**402 Crofts Hall**
**Buffalo, NY 14260-7016**

AWARDEE ORGANIZATION CODE (IF KNOWN)
**0028373000**

NAME OF PRIMARY PLACE OF PERF
**SUNY at Buffalo**

ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE
**SUNY at Buffalo**
**338 Davis Hall**
**Buffalo ,NY ,142602500 ,US.**

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **CIF21 DIBBs: EI: Vizier, Streamlined Data Curation**

| REQUESTED AMOUNT | PROPOSED DURATION (1-60 MONTHS) | REQUESTED STARTING DATE | SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE |
|---|---|---|---|
| $ 2,725,699 | 36 months | 01/01/17 | |

THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW
☐ BEGINNING INVESTIGATOR (GPG I.G.2)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
☒ HISTORIC PLACES (GPG II.C.2.j)
☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date ________________
PHS Animal Welfare Assurance Number ________________
☒ FUNDING MECHANISM **Research - other than RAPID or EAGER**

☒ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number **FWA00008824**
Exemption Subsection **will apply** or IRB App. Date ________________
☒ INTERNATIONAL ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
**FI IN GM CH**
☒ COLLABORATIVE STATUS
**A collaborative proposal from one organization (GPG II.D.4.a)**

PI/PD DEPARTMENT
**Department of Computer Science**

PI/PD POSTAL ADDRESS
**338 Davis Hall**

**Buffalo, NY 142600000**
**United States**

PI/PD FAX NUMBER

| NAMES (TYPED) | High Degree | Yr of Degree | Telephone Number | Email Address |
|---|---|---|---|---|
| PI/PD NAME<br>Oliver A Kennedy | PhD | 2011 | 716-645-2634 | okennedy@buffalo.edu |
| CO-PI/PD<br>Juliana Freire | PhD | 1997 | 801-712-1363 | juliana.freire@nyu.edu |
| CO-PI/PD<br>Boris Glavic | PhD | 2010 | 312-567-3035 | bglavic@iit.edu |
| CO-PI/PD | | | | |
| CO-PI/PD | | | | |

# CERTIFICATION PAGE

### Certification for Authorized Organizational Representative (or Equivalent) or Individual Applicant

By electronically signing and submitting this proposal, the Authorized Organizational Representative (AOR) or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding conflict of interest (when applicable), drug-free workplace, debarment and suspension, lobbying activities (see below), nondiscrimination, flood hazard insurance (when applicable), responsible conduct of research, organizational support, Federal tax obligations, unpaid Federal tax liability, and criminal convictions as set forth in the NSF Proposal & Award Policies & Procedures Guide,Part I: the Grant Proposal Guide (GPG). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U.S. Code, Title 18, Section 1001).

### Certification Regarding Conflict of Interest

The AOR is required to complete certifications stating that the organization has implemented and is enforcing a written policy on conflicts of interest (COI), consistent with the provisions of AAG Chapter IV.A.; that, to the best of his/her knowledge, all financial disclosures required by the conflict of interest policy were made; and that conflicts of interest, if any, were, or prior to the organization’s expenditure of any funds under the award, will be, satisfactorily managed, reduced or eliminated in accordance with the organization’s conflict of interest policy. Conflicts that cannot be satisfactorily managed, reduced or eliminated and research that proceeds without the imposition of conditions or restrictions when a conflict of interest exists, must be disclosed to NSF via use of the Notifications and Requests Module in FastLane.

### Drug Free Work Place Certification

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent), is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

### Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency? Yes □ No ☒

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

### Certification Regarding Lobbying

This certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

### Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:
(1) No Federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any Federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, ‘‘Disclosure of Lobbying Activities,’’ in accordance with its instructions.
(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

### Certification Regarding Nondiscrimination

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

### Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:
(1) community in which that area is located participates in the national flood insurance program; and
(2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:
(1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
(2) for other NSF grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

### Certification Regarding Responsible Conduct of Research (RCR)
### (This certification is not applicable to proposals for conferences, symposia, and workshops.)

By electronically signing the Certification Pages, the Authorized Organizational Representative is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research.
The AOR shall require that the language of this certification be included in any award documents for all subawards at all tiers.

# CERTIFICATION PAGE - CONTINUED

## Certification Regarding Organizational Support

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that there is organizational support for the proposal as required by Section 526 of the America COMPETES Reauthorization Act of 2010. This support extends to the portion of the proposal developed to satisfy the Broader Impacts Review Criterion as well as the Intellectual Merit Review Criterion, and any additional review criteria specified in the solicitation. Organizational support will be made available, as described in the proposal, in order to address the broader impacts and intellectual merit activities to be undertaken.

## Certification Regarding Federal Tax Obligations

When the proposal exceeds $5,000,000, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal tax obligations.
By electronically signing the Certification pages, the Authorized Organizational Representative is certifying that, to the best of their knowledge and belief, the proposing organization:
(1) has filed all Federal tax returns required during the three years preceding this certification;
(2) has not been convicted of a criminal offense under the Internal Revenue Code of 1986; and
(3) has not, more than 90 days prior to this certification, been notified of any unpaid Federal tax assessment for which the liability remains unsatisfied, unless the assessment is the subject of an installment agreement or offer in compromise that has been approved by the Internal Revenue Service and is not in default, or the assessment is the subject of a non-frivolous administrative or judicial proceeding.

## Certification Regarding Unpaid Federal Tax Liability

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal Tax Liability:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has no unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have been exhausted or lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability.

## Certification Regarding Criminal Convictions

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Criminal Convictions:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has not been convicted of a felony criminal violation under any Federal law within the 24 months preceding the date on which the certification is signed.

## Certification Dual Use Research of Concern

By electronically signing the certification pages, the Authorized Organizational Representative is certifying that the organization will be or is in compliance with all aspects of the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.

| | | |
|---|---|---|
| AUTHORIZED ORGANIZATIONAL REPRESENTATIVE | SIGNATURE | DATE |
| NAME<br>Amy M Lagowski | Electronic Signature | Apr 4 2016 4:44PM |
| TELEPHONE NUMBER<br>716-645-4419 | EMAIL ADDRESS<br>Amy.Lagowski@buffalo.edu | FAX NUMBER<br>716-645-2760 |

# PROJECT SUMMARY

## Overview:

Data curation is a critical task in data science in which raw data is structured, validated, and repaired. Traditionally, curation has been carried out as a pre-processing task: after all data are selected for a study (or application), they are cleaned before an analysis can begin. This is problematic because while some cleaning constraints can be easily defined, others are only discovered as one analyzes the data. In this project, we propose to build Vizier, a system that unifies curation and data exploration through provenance. Vizier integrates and extends three existing systems that we have developed in previous work: Mimir, a system that supports probabilistic pay-as-you-go data curation operators; VisTrails, an NSF-supported open-source system designed for interactive data exploration; and GProM, a database middleware that efficiently supports fine-grained provenance. Vizier's interface combines elements of both notebooks and spreadsheets, allowing users to quickly apply precise data repairs that can later be generalized and deployed over a large scale dataset. Vizier also allows developers to defer resolving ambiguous data until they have had a chance to thoroughly explore the ambiguity and its implications. The system uses non-intrusive visual cues to explain how and why it arrived at specific results. Vizier guides the user in incrementally building and refining a curation workflow by providing recommendations on which cleaning operations to apply and how to tune them.

--- Project Participants ---

The proposed work will be supervised by Dr. Oliver Kennedy at the University at Buffalo, SUNY; Dr. Boris Glavic at the Illinois Institute of Technology; and Dr. Juliana Freire at New York University, with assistance from Dr. Heiko Mueller at New York University. As per his letter of collaboration, Dieter Gawlick and Zhen Hua-Liu of Oracle will provide guidance and feedback. As per their respective letters, we will receive assistance in designing and evaluating Vizier from Julia Lane and Rebecca Rosen of the NYU Center for Urban Science and Progress (CUSP), Ingrid Gould Ellen from NYU's Furman Center, Rodney Stiles from the NYC Taxi and Limousine Commission (TLC), and Ronny Fehling from the Airbus group.

--- Keywords ---

Provenance, Curation, Ambiguity, Urban Data, Sensor Data, Wrangling

--- Targetted NSF Directorates/Divisions ---

CISE/IIS, SBS/SBE

## Intellectual Merit :

The proposed work links research efforts on three systems: Mimir, VisTrails, and GProM. An expected outcome of the work will be a better understanding of provenance, particularly in the context of data curation. The project will also contribute to the state of the art in data curation, from the design of data curation languages, to interface design, and exploratory data processing. Besides advancing the state of the art in data curation, the proposed work will also improve research in other fields, notably, in social sciences. It will also contribute to an emerging field: urban science. Through the deployment of our tools within CUSP and the Furman Center, our work has the potential to impact a wide range of scientists and students, as well as the NYC agencies that are partnering with CUSP and Furman on various projects.

## Broader Impacts :

Data quality is a serious problem that impacts data scientists in all fields of study. As demonstrated by active interest from our collaborators at Oracle, Airbus, CUSP, TLC, and the Furman Center, this project will have significant positive impact on government, industry, and academic research. We have participants already eager to use the tools we will develop, so the impactfulness of this project hinges not on where it can be deployed, but rather on how soon can it be ready. Our work will positively impact government by improving data quality, which in turn will lead to better planning and policies and more efficient operations. The ability to publish provenance-rich, high-quality data will also positively impact governance and citizen engagement. In addition to improving the state of the art in data curation, the project will also result in support for four PhD students and one Postdoctoral researcher.

# TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

# CIF21 DIBBs: EI: Collaborative: Vizier, Streamlined Data Curation

4 April 2016

Oliver Kennedy, Boris Glavic, Juliana Freire

*NSF Divisions*: CISE/III, SBS/SES

*Keywords*: Provenance, Curation, Ambiguity, Urban Data, Sensor Data, Wrangling

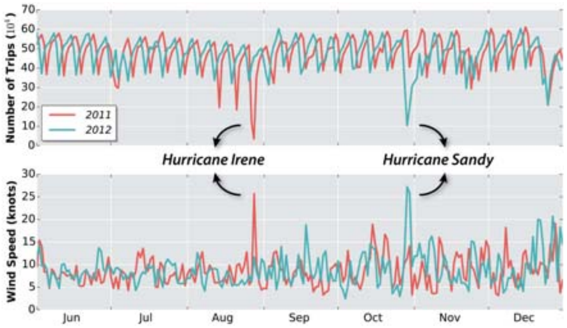


Figure 1: The plot on the top shows how the number of trips varies over 2011 and 2012. While the variation is similar for the two years, there are clear outliers, including large drops in August 2011 and in October 2012. These are not errors, but in fact correspond to hurricanes Sandy and Irene, as shown by the wind speed plot on the bottom.

**Curation as an Integral Component of Data Exploration** Data curation is a critical task in data science in which raw data is structured, validated, and repaired. Data validation and repair establish trust in analytical results, while appropriate structuring streamlines analytics. Unfortunately, even with advances in automated data cleaning tools, such as Oracle's Data Guide and Trifacta's Wrangler, curation is still a major bottleneck in data exploration. Traditionally, curation has been carried out as a pre-processing task: after all data are selected for a study (or application), they are cleaned and loaded into a database or data warehouse. This is problematic because while some cleaning constraints can be easily defined (e.g., checking for valid attribute ranges), others are only discovered as one analyzes the data. As one example, consider taxis in New York City [218]. Every day, there are over 500,000 taxi trips transporting about 600,000 people from Manhattan to different parts of the city [36]. Through the meters installed in each vehicle, the Taxi & Limousine Commission (TLC) captures detailed information about trips, including: GPS readings for pick-up and drop-off locations, pick-up and drop-off times, fare, and tip amount. These data have been used in several projects to understand different aspects of the city, from creating mobility models and analyzing the benefits and drawbacks of ride sharing, to detecting gentrification. In a recent study [89], we investigated quality issues in the taxi data. We found invalid values such as negative mile and fare values, as well as trips that started or ended in rivers or outside of the US. These are clearly errors in the data. Other issues are more nuanced. An example is a fare with a tip of US$938.02 (the maximum tip value for the 2010 dataset). While this could have be an error in the data acquisition or in the credit card information, it could also be the case that a wealthy passenger overtipped her taxi driver. Issues are often detected during analytics, as different slices of the data are aggregated. Figure 1 shows the number of daily taxi trips in New York City (NYC) during 2011 and 2012. Note the large drops in the number of trips in August 2011 and October 2012. Standard cleaning techniques are likely to classify these drastic reductions as outliers that represent corrupted or incorrect data. However, by integrating the taxi trips with wind speed data (bottom plot in Figure 1), we discover that the drops occur on days with abnormally high wind speeds, suggesting a causal relation: the effect of extreme weather on the number of taxi trips in NYC. Removing such outliers would hide an important phenomenon. Conversely, detecting it upfront requires identifying a non-obvious pattern in a very high-dimensional space [69].

Issues like these appear across all forms of analytics, making curation an integral component of data exploration. When erroneous features are identified, appropriate *cleaning operations should be applied on the fly*. Besides the need to refine a curation pipeline as the user gets more familiar with a dataset, different questions that arise during exploration may require different cleaning strategies. Thus, we need to move from the traditional cleaning function $DirtyData \rightarrow CleanData$, to a function that encapsulates the exploratory curation process: $DirtyData \times UserTask \rightarrow (CleanData, Explanation)$. The trial-and-error nature of the curation process poses several challenges. First, and foremost, the cleaned data must be accompanied by its provenance which *explains the transformations applied to the raw data*, as well as *ambiguities that arise while applying these transformations*. This is critical for subsequent analyses, as experts need to both assess the quality of the data and understand which assumptions they can rely on. If an operation is applied and later found to be incorrect (e.g., removing the outliers in Figure 1), it should be possible to *undo the operation and all of its direct and indirect effects*. It should also be possible to *modify an operation* (e.g., change the parameters of an outlier detection operation to not consider the $938.02 tip amount as an outlier) and the effects of this modification should be propagated to derived data. Furthermore, it is often necessary to *explore and compare alternative cleaning strategies and to consolidate their results*, especially when data curation is a collaborative effort. Currently, such corrections and refinements require an error-prone, time-consuming process to track different versions of scripts and queries used, and the data feeds they were applied to.

**Building Infrastructure for Data Curation** In this project, we propose to build Vizier, a system that unifies curation and data exploration through provenance. Vizier tracks the provenance of the exploratory process and all the computations performed [93, 98, 159, 224] as well as the provenance of the affected data items [215]. By connecting these two forms of provenance, Vizier is able to provide a detailed audit trail that explains which operations caused which changes to the data and to reproduce these steps on new data. Vizier will integrate and extend three production/near-production quality systems that we have developed in previous work: *Mimir* [143, 167, 231, 232], a system that supports probabilistic pay-as-you-go data curation operators; *VisTrails* [42, 43, 51, 99, 101, 128, 151, 200, 202, 209], an NSF-supported open-source system designed for interactive data exploration; and *GProM* [10, 11, 106, 107, 109, 116, 175], a database middleware that efficiently tracks fine-grained data provenance, even through update operations. A core challenge of implementing Vizier will be to integrate the data and provenance models of these systems and to develop a unified interface that synergystically combines the features of each.

To define the requirements for the system design, we have taken into consideration the curation needs of previous and ongoing projects led by the PIs in different domains—from urban to industrial data, as well as the results of a survey (see Section 1). Curation is a difficult task for experts in data management and computer science, and even more so for the growing number of *data enthusiasts*, with expertise in particular domains but that often lack training in computing. Furthermore, the process demands the iterative, trial-and-error application and tuning of data cleaning operations. To support these requirements, Vizier will provide a novel "sandbox" interface for exploration and curation that is easy to use, and whose unique features are enabled by the integration of fine- and coarse-grained provenance.

Vizier's interface is a hybrid of notebook-style interfaces and spreadsheets. Notebook interfaces, as exemplified in Jupyter (`http://jupyter.org`), interleave code, data, summaries, visualizations, and documentation into a single train of thought that allows readers to retrace an analyst's exploratory process. Conversely, spreadsheets allow analysts an unprecedented level of flexibility and freedom, while still supporting a powerful programming model. Thus, in contrast to classical notebook software where all data manipulations must be handled programmatically, Vizier allows data, summaries, and visualizations to be edited directly as in a spreadsheet *and* in a classical programming environment *simultaneously.*

Vizier uses provenance to aid the user in the exploration process by enabling her to navigate the evolution of her curation pipeline, to understand the impact of cleaning operations she devises, and to understand sources and effects of uncertainty or ambiguity in her data. Vizier also uses provenance to provide recommendations to the user on how to tune curation operations, which curation operations to apply next [152], and how to generalize her curation workflow developed over a small sample dataset to deploy it over a large scale dataset (e.g., to deploy the workflow on a Big Data platform).

Finally, provenance enables the use of ambiguity-aware data transformations, including automated zero-configuration curation operators called lenses [167, 232]. Unlike classical clean-before-use approaches to data curation, Vizier allows data wranglers to defer resolving ambiguities. Ambiguities persist through transformations and appear as annotations (e.g., asterisks) on data in Vizier that indicate the cause and quantify the effects of the ambiguity. By deferring ambiguity resolution until after the user has had a chance to explore, the analyst can better decide how to resolve the ambiguity, or even whether she needs to resolve the ambiguity at all. Lenses also form the basis for Vizier's extensibility, as they allow data cleaning heuristics or tools to be just "plugged" in, even if the tools have overlapping use cases or conflicting effects.

The interface features of Vizier are enabled by tracking *provenance.* Edits, whether to code or data, are all *transparently* recorded and associated with the resulting data. Guesses made by the system on behalf of the user are recorded similarly. Provenance persists through computations and modifications, providing an audit trail and allowing the system to explain the reasoning behind results it shows to the analyst. Provenance allows potentially suspicious or ambiguous results to be flagged as such, provides context for results, and by using well-established techniques for probabilistic and uncertain data management [214], also quantifies the impact of the ambiguity on the result. An important technical challenge we will tackle in this regard is how to integrate the workflow (evolution) provenance of VisTrails with the fine-grained data provenance of GProM and the ambiguity tracking of Mimir.

Interactive exploration also requires interactive latencies, even for large datasets. We will build Vizier to provide interactive response times through two techniques. First, through its facilities for tracking provenance, Vizier can enact a form of program slicing, a more general form of incremental view maintenance that allows data representations produced by Vizier to be updated rapidly in response to small changes to input data and parameters of curation workflows. Second, similar to Trifacta's Wrangler, Vizier allows analysts to extract and use samples of large datasets for preliminary development efforts. Unlike Wrangler, however, Vizier can measure the quality or representativeness of a sample set with respect to a given summary, representation, or view of the data. Furthermore, Vizier guides the user in tweaking a curation workflow developed over a sample dataset to ensure that it generalizes to the complete data.

**Community Building.** Users in a broad range of application domains from urban sciences to business intelligence, have committed to collaborate with the PIs on this proposal (see Section 1 and letters of collaboration). This diversity of collaborators will ensure that Vizier is relevant to real-world needs and help to establish a healthy user base. The latter is particularly important to sustain development beyond the initial funding period. Our strategy for building a healthy community around Vizier includes demonstrations, workshops, publicly released code, and ensuring proper documentation for users and developers, and development of project governance. Our sustainability plan is detailed in Section 4.

**The Team.** The PIs bring cross-cutting expertise from different areas of research on data management, provenance, and visualization. Systems developed by each of the PIs: Mimir, GProM and VisTrails, will serve as the core building blocks of our proposed system Vizier. Additionally, PIs Kennedy and Glavic already have a record of collaboration that has resulted in joint papers (under submission) related to data curation, uncertainty, and provenance. Both PIs have ongoing projects in this domain that are sponsored by Oracle (see letter of collaboration from Gawlick and Hua-Liu). PI Kennedy's expertise covers optimization and incremental computation [3,138,145], data structures [144], uncertain data management [140–143,167, 168,232], online aggregation [137,141], and mobile systems [49,139]. He is a member of UB's Center for Multisource Information Fusion and National Center for Geographic Information Awareness, and has collaborations [49,139] based on UB's NSF-funded PhoneLab smartphone experimental platform. PI Glavic's research is focused on database systems with a strong track record in provenance [9,11,105–117,174,175, 182,183] and data integration [12–15,109,110]. He has designed and implemented several provenance-aware systems including Perm [106–108,116], GProM [9,11,174,175], Vagabond [109,110], Ariadne [111,112], and LDV [114,182,183]. PI Freire's research has focused on big-data analysis and visualization, large-scale information integration, provenance management, and computational reproducibility. She has a track record of successful interdisciplinary collaborations and her work has had impact in a number of domains beyond computer science, from physics and biology to climate research and social sciences [18,29,83,125,127,194, 195,199]. She has co-authored multiple open-source systems [1,34,189,216,221,224] (see `https://github.com/ViDA-NYU`), including VisTrails which is used by high-visibility scientific applications in diverse fields. As the Executive Director of the Moore-Sloan Data Science Environment at NYU, a faculty member at the NYU Center for Data Science and at the NYU Center for Urban Science and Progress, she currently leads several projects where data curation is a key challenge.

# 1 Applications and Requirement Gathering

The need for data curation arises in all applications of data science. To ensure that our efforts will see use in practice, the PIs will deploy Vizier through established collaborative efforts with data scientists in academia and industry. These collaborations will serve as a platform to evaluate Vizier, as well as a source of feedback, helping us to identify and address critical pain points in data curation workflows. Additionally, we have conducted an informal survey, reaching out to several affiliated data science communities and potential consumers of Vizier for feedback and desiderata. In this section, we outline our two primary collaborative efforts, as well as the high-level feedback garnered from our informal survey.

## 1.1 Curating Urban Data

The NYU team is working on several projects that involve curation, analysis and integration of urban data [28, 33, 47, 70, 71, 84, 102, 125, 180, 181, 184, 216]. PI Freire is a faculty member at the NYU Center for Urban Science and Progress (CUSP). CUSP is set up as a unique public-private research center that uses New York City as its laboratory and classroom to help cities around the world become more productive, livable, equitable, and resilient. Research and development at CUSP focuses on the collection, integration, and analysis of data to improve urban systems. The social sciences play an integral role in CUSP activities– people are the customers and operators of urban systems, so that understanding them and their behavior is essential to the CUSP mission. Conversely, CUSPs large, multi-modal data sets and technology have the potential to revolutionize the social sciences. CUSP has established a data facility (CDF) to support the empirical study of cities. Freire was one of the original designers of CDF and part of the vision of this proposal was motivated by the needs of CDF users, notably: different users (and projects) need to combine and clean datasets in different ways, depending on their research questions, which change over time as the process evolves and new hypotheses are formulated; users come from widely different backgrounds, and include social scientists, physicists, computer scientists, civil engineers, policy makers, and students. As Professor Lane and Dr. Rosen state in their letter of collaboration, the CDF needs a data curation infrastructure such as the one we propose to build. Vizier will bring many benefits to CDF, including: users will be to able collaboratively curate data; curated data will include detailed provenance, allowing them to be re-used; and curation pipelines will also be shared within the facility, enabling users to benefit from the collective wisdom of their peers by re-using and building upon these pipelines.

Freire also has an ongoing collaboration with the NYU Furman Center for Real Estate and Policy [125]. The Furman Center uses urban data to explore the effect that land use regulations, real estate development, and other public and private place-based investments have on the quality, affordability, and character of neighborhoods and on individual well-being (see e.g., [76, 77, 126, 205]). They take an interdisciplinary approach, applying policy and legal analyses, as well as econometric techniques to study critical current issues. Over the years, they have collected a broad array of data on demographics, neighborhood conditions, infrastructure, housing stock and other aspects of New York City's neighborhoods and real estate market [103]. They also produce a series of data products that help inform community groups, developers, policymakers and investors about trends, challenges and opportunities in particular neighborhoods [31, 32, 45, 46]. Given that their research has direct impact on policies and the data they release is widely used, for them, data quality is of utmost importance (see letter from Professor Gould Ellen).

Freire's group has multiple ongoing collaborations with NYC agencies. She is a member of the NYC Taxi & Limousine Commission (TLC) Data/Technology Advisory Committee. She has collaborated with the TLC on different projects, from the study of privacy and quality issues in data they release [89] to the deployment of TaxiVis, an open-source tool for the analysis of spatio-temporal data [84, 102, 216]. As with many other agencies that use data to improve their operations and policies, and that publicly release these data [179], the TLC faces tremendous challenges around data quality (see letter by Rodney Styles, TLC).

Our collaborators will provide us unique data and feedback that we will use in the design of Vizier, and we will collaborate with them on the deployment of the system at CUSP, Furman Center, and TLC.

## 1.2 Industrial Strength Data Lakes

Our second effort is part of a large, ongoing collaborative project between PIs Kennedy and Glavic and Oracle's Dieter Gawlick and Zhen Hua-Liu, as well as with Ronny Fehling, Head of Data Driven Technologies and Advanced Analytics at Airbus (see attached letters). One specific part of this effort concerns a large scale project at Airbus to merge data sources from different sectors of its business into a single, company-wide data lake. In addition to the issues of data quality that arise in our collaboration with CUSP, an operation of this scale presents several additional challenges:

**Cross-Domain Data Mapping.** Datasets from different sectors have distinct, non-overlapping schemas, as well as distinct attribute domains recorded at different granularities. There are also more subtle conflicts. For example, one assumption often made when using a dataset is temporal stability [143]. An anecdotal example encountered during past collaborative efforts involved an auto-generated report that computed revenue totals for sales groups at a large company. Revenue totals would fluctuate unexpectedly over time, even decreasing for some groups. The cause was mismatched temporal assumptions between an append-only sales history dataset and a HR database linking salespeople to their groups. As salespeople moved between groups, the HR database was updated and past sales would be re-attributed to their new group.

**Data Discovery.** Given the number of distinct data sources and data sets available at a company of this scale, simply finding datasets applicable to a specific problem is a challenge. Forcing participants in the data lake to properly create and curate metadata for their datasets is infeasible, necessitating alternative approaches to data discovery.

**Heterogeneity.** The data lake draws on a heterogeneous mix of storage layers and data sources including Hadoop, client APIs, external tables, and live data feeds, a fact that makes other data quality challenges more difficult. Cross domain data mapping is more difficult, as schemas and attribute domains may not be immediately accessible. Data discovery also becomes more difficult, as complete datasets are not available and associated metadata must either be tracked independently or through purpose-built adaptors.

**Versioning.** As analysts progressively refine and extend a dataset, multiple revisions of the same data become available. Using an earlier revision may require an analyst to duplicate effort, while a later revision may be modified in unexpected ways that are inconsistent with the analyst's current goals.

### 1.3 Community Feedback

As a preliminary effort at reaching a broader audience, we have conducted an informal survey to gauge preliminary interest in Vizier, as well as to solicit feedback from different communities about the data curation challenges that they face. The survey was distributed directly to many of the PI's colleagues and collaborators at potential deployment sites like UB's National Center for Geographic Information and Analysis. Over the course of a one-week informal survey period, we received 8 detailed responses from parties expressing interest in improved tools for data cleaning and a further 3 additional off-the-record responses expressing frustration with the state of the art in data curation. The preliminary responses indicate that *spreadsheets are the tool of choice for data curation* (used by 6 out of the 8 respondents). The two respondents who do not use spreadsheets, regularly work with data at sizes to which spreadsheets do not scale. When asked informally, one of these respondents stated that a spreadsheet-style interface used to design scalable curation workflows over samples would be useful to them. When asked about the biggest challenges in their most recent analytics task, the two most common answers were: *the time taken to perform computations* and *data quality.* For the first case, standard database techniques (indexing, incremental maintenance, set-at-a-time) are often sufficient, but require too much effort to deploy, configure, and load data into. A concern with missing or garbage data was also common, and several respondents in particular noted issues of data quality when working with different schema versions simultaneously.

Throughout our project, we will continue to seek feedback from experts in different domains. This will not only ensure the proposed infrastructure will be widely available, but we also hope this will help create and sustain a strong user community for Vizier.

## 2 The Vizier System

The goal of Vizier is to **unify the tasks of data exploration and data curation** under a single interface (illustrated in Figure 3) that combines characteristics of a notebook and a spreadsheet. Our aim is to provide a tool that allows users to quickly and easily tabulate, visualize, clean, and summarize large datasets. As the user explores her data, Vizier provides a variety of tools for curation, from the ability to manually edit individual records, to batch operators that **automate common tasks** like schema merging, outlier detection, missing value imputation, entity resolution, constraint repair, and others. While the user is exploring and curating the data, Vizier records her curation activities to **incrementally build a workflow** [5, 51, 64, 65, 101, 105, 128, 202], as illustrated in Figure 2a. During this process, the system **recommends curation steps and tuning parameters** to the user based on successful past curation efforts on datasets with similar characteristics [200]. For example, the system may detect that the user is constructing a typical address cleaning workflow and suggest curation steps that are commonly used to clean addresses such as using an external zip-city lookup table. For large datasets, a user would typically first curate and explore a small sample and only **deploy** her curation workflow over the full dataset once she

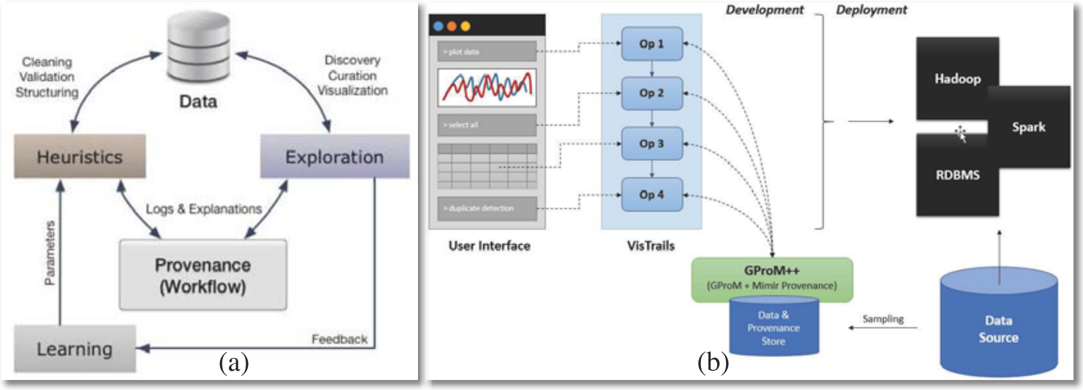
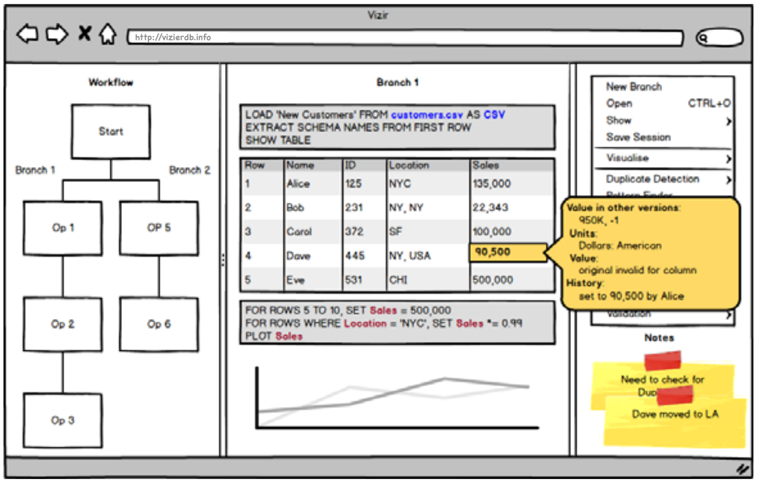


Figure 2: The Vizier System. (a) Vizier backs freeform data exploration with heuristic data curation operators and a feedback and provenance-driven learning engine that continually improves its curation heuristics. (b) Workflows are designed through Vizier's hybrid notebook-spreadsheet interface on small or sampled dataset processed locally by an engine backed by GProM and Mimir. When workflows are ready, they can also be deployed to the cloud, to a Hadoop or Spark cluster, or to a relational database system.

is sufficiently certain that the workflow resolves all data quality issues. This type of deployment at-scale is supported in Vizier, as illustrated in Figure 2b. One challenge in this setting is that the curation steps applied over the sample dataset need to be generalized to correctly apply to the full dataset. For example, the user may manually remove taxi trips with unrealistically large fare values from the sample dataset. However, the full dataset will contain many additional such trips that would not be removed when the workflow is deployed, unless the deletion of a fixed set of trips is *generalized* to a deletion of trips based on their fare value. Vizier aids the user in the deployment by providing recommendations on how to **generalize** her operations based on data characteristics and provenance. For example, the system may detect that all deleted trips have an unusually high fare in common and offer to delete all records with similarly high fares in one operation. Combining workflow provenance, data provenance, and uncertainty management, Vizier can offer non-intrusive visualizations that **fully explain** how data was produced and how ambiguities in the input data or curation steps have affected it. For example, clicking on a value in Vizier's interface will bring up an explanation view showing the value's provenance, ambiguities in its derivation, and how it compares to related values (e.g., those in the same column).

**Interface.** Vizier's interface (illustrated in Figure 3) combines elements of both notebooks and spreadsheets. Notebook interfaces like Jupyter use an analogy of pages in a notebook that consist of a block of code, as well as an output for the block like a table, visualization, or documentation text. Blocks are part of a continuous program, allowing a user to quickly probe intermediate state or to safely insert hypothetical, exploratory modifications by adding or disabling pages. Spreadsheets give users an infinite 2-dimensional grid of cells that can hold either constant values or computed values derived from other cells. Instead of classical programmatic specification of bulk, set-at-a-time operations, spreadsheets use the metaphor of copying code and relative, positional data dependencies to "map" operations over collections defined by contiguous regions of data. Thus, the ability to change any value anywhere in the execution process, and simple integrated visualizations combine to make spreadsheets a very viable tool for data curation and exploration. The simplicity of spreadsheets has encouraged many database-driven efforts to resolve the impedance mismatch between positional and set-at-a-time query semantics [132,160], make spreadsheets more structured [16, 17] or make databases more spreadsheet-like [131]. Vizier builds on these efforts, creating a hybrid notebook-spreadsheet interface by making a notebook's output dynamic. Vizier's users can edit tables and visualizations directly, and have those edits reflected in the notebook's code block through database-style table updates. As a result, the user's edits, however they are applied, are recorded in the notebook as a part of the workflow (see Figure 2b). Although we will not reproduce the full spreadsheet interface entirely, our goal is to replicate as many of the flexible data and schema manipulation features of spreadsheets as possible. Vizier allows users to overwrite arbitrary values, embed formulas into table cells, cut/copy/paste cells, add (or delete) columns or rows, and sort or filter the data. In addition to low-level modifications, the user can also apply higher-level curation operations, including ones for which exact configuration parameters may not be known ahead of time (we return to this later). Unique to Vizier is its ability to present useful recommendations to the user based on provenance, to expose ambiguity through visual hints in the interface, and to explain why and how values where derived.

**Incrementally building curation workflows.** The workflows that are *incrementally* constructed based on the user's operations serve several purposes. By leveragering VisTrails [42, 99, 202] as a system to manage these workflows and their evolutionary construction through user actions [43], they allow seamless repeatability [9, 10]. This enables users to easily identify and revert erroneous changes and helps Vizier to learn from the user's activities. To support a user in efficiently revising a workflow, we can use provenance to propagate changes to data and workflow steps throughout dependent workflow stages. Techniques such as program slicing [2,153,230] and incremental view maintenance [3,35,38,50,118,135,138,145] can help to improve the performance of this process. It can be further optimized in the case where update histories are available through a technique called reenactment which is readily supported in GProM [9,10].



```
LOAD 'New Customers' FROM customers.csv AS CSV
EXTRACT SCHEMA NAMES FROM FIRST ROW
SHOW TABLE
```

| Row | Name | ID | Location | Sales |
|---|---|---|---|---|
| 1 | Alice | 125 | NYC | 135,000 |
| 2 | Bob | 231 | NY, NY | 22,343 |
| 3 | Carol | 372 | SF | 100,000 |
| 4 | Dave | 445 | NY, USA | 90,500 |
| 5 | Eve | 531 | CHI | 500,000 |

```
FOR ROWS 5 TO 10, SET Sales = 500,000
FOR ROWS WHERE Location = 'NYC', SET Sales *= 0.99
PLOT Sales
```

Figure 3: An example of Vizier's UI

Furthermore, a workflow generated for one dataset can be easily adapted for use with different, but similar data sets, a task for which research by PI Freire [200] provides a preliminary approach. Inputs can take the form of CSV files, relational database tables, JSON files, unstructured data, or external data sources like web APIs or HTML pages. Workflows are part of the extensive provenance tracking of Vizier [51,65], helping the user and her collaborators to explain the system's outputs and audit their curation and exploration activities. Unique to Vizier is that workflow steps are not considered as black boxes, but correspond to programs in a declarative language described below. Vizier also supports ambiguity-tolerant operators called lenses and tracks fine-grained provenance at the data level.

**Exploration, testing, and deployment.** As mentioned earlier, a user would typically want to design a workflow over a small sample dataset before testing or deploying it at scale. One of the compelling benefit of workflows is the ability to adapt and re-use them across different contexts. A data curation process *designed through an exploration-friendly spreadsheet* interface can then be adapted to other datasets, replicated for recurring data curation tasks like monthly reports, or run at scale on a cluster. Like Wrangler [121, 134], Vizier allows users to easily develop data curation workflows on a small sample of the data before scaling to tera- or peta-byte scale data on a Hadoop cluster or RDBMS. In contrast to Wrangler which forces users to generalize edits upfront, Vizier instead helps users to generalize a workflow that includes manual curation steps into a workflow suitable for the complete dataset.

**Curation DSL with Provenance.** Vizier's flexibility is derived, in part, from a new domain-specific language for exploratory curation that we will develop as part of the proposed work. The objective of this DSL is to *integrate closely with the interface* by ensuring a 1-1 mapping between operations in the language and edits to Vizier's tabular outputs. In addition to facilitating the hybrid notebook/spreadsheet environment, the Vizier DSL serves as an intermediary between the workflows and a variety of back-end execution environments. In addition to executing locally, we will develop modules to compile the Vizier DSL down to a variety of languages for deployment on external computing platforms, such as SQL for Relational DBs or Map/Reduce or Spark programs for cloud computing clusters. To maximize compatibility with GProM and Mimir, Vizier's DSL will start as an instance of classical extended-relational algebra [123]. For a more user-friendly imperative flavor and to make it easier to mirror manual edits from the spreadsheet onto the target program, we will add operations from SQL's DML that edit specific fields, batches of fields, and insert or delete rows; operations from SQL's DDL that add, remove, and hide columns or manipulate constraints; and automated data curation operations based on Mimir's lenses, as discussed below. In spite of the imperative flavor of the language, these operators effectively modify a table-valued query "object" modelling the whole sequence of operations [9, 10], and can be thought of as operators in a relational monad [39] that can be reduced to a single query.

**Taming uncertainty and explainability.** Lenses [167,232], part of the Mimir system discussed below, are data curation operators that use heuristics to minimize or eliminate configuration. In lieu of configuration, a lens is allowed to produce multiple ambiguous *possible outputs* alongside a single *best guess output*. Possible outputs remain associated with the output of queries or transformations over lenses, allowing potentially ambiguous results to be identified and explained and allowing the uncertainty stemming from this ambiguity to be quantified. For example, depending on the user's goals outliers like the $938 tip in the TLC dataset might be kept as-is, elided from the dataset, or replaced with a reasonable educated guess. In the latter case, there are also numerous heuristics that can approximate a user's best guess, including sequential interpolation along each of several candidate dimensions [59,67,136], a classifier trained on the adjacent attributes [228], a classifier that treats the data as the output of a Markov process [157,158], and numerous others. A user could be asked to select from among these options upfront, but she may not have sufficient information about the data to decide which is relevant (e.g., is the tip a data error or a legitimate outlier). Furthermore, depending on the user's goals, the choice may not even be relevant (e.g., she is generating a table of average tips by year). Workflows also provide context for data presented through the notebook interface. As illustrated in Figure 3, users can obtain statistics about output values through the Vizier UI, including the record's dependencies and formula, variations in the record's value over time, the system's certainty in the result (discussed further below), and other features like the set of of edits with the greatest impact on the value.

**Automation and Recommendation.** In keeping with best practices for user interface design [177], a goal of Vizier's interface is retaining the user's sense of control over the system's behavior. Accordingly, we adopt two general strategies for streamlining the user's interaction with the system that we refer to as Automation and Recommendation. Automation allows users to accomplish complex high-level tasks without concern for the low-level details of the task. Effective use of automation requires both the user's consent and awareness, as well as effective communication in the case of ambiguities. Automation in Vizier occurs through lenses. Lens behaviors are precisely defined in terms of desired effects. By requesting that a lens be applied, a user indicates both consent and an awareness of what the lens is trying to accomplish. As discussed above, ambiguities are communicated through explanations over results, minimizing upfront effort from the user, but keeping them aware of potential repercussions of using automated tools. By integrating these techniques into Vizier we enable users to combine these techniques with each other and with simpler manual curation operations (e.g., manually deleting dirty rows). Even more important, through their integration with Vizier, these operations can benefit from the uncertainty, provenance, recommendation, and deployment features of Vizier. Recommendations, instead help users to quickly reach relevant curation tools and to discover tools that they may not be aware of. Vizier uses a repository of collected provenance to provide suggestions to the user based on the current workflow and data characteristics. This is similar to Wrangler [134] which trains a Markov model on sequences of transformations to suggest transformations commonly performed together in sequence. However, our recommendations are much more advanced in that they are not just based on the current structure of the workflow, but also on the fine-grained data and workflow provenance. Four types of recommendations will be supported: (1) Recommendations on how to tune the parameters of a cleaning operation in the workflow. For example, the sensitivity of an outlier detection step could be tuned based on successful values for past outlier detection methods over data with similar characteristics. Similarly, if a past workflow containing a particular data curation operation has a similar structure as the current workflow this can also be an indication that similar tuning parameter values should be applied; (2) Recommendations on which data curation steps to apply next. These recommendations will be based both on the characteristics of the workflow as well as the data. (3) Recommendations on how to generalize specific curation steps, e.g., updating values based on a condition instead of updating a fixed set of rows. (4) Finally, Vizier offers facilities for automated data discovery, both through simple keyword search, and offering suggestions based on datasets frequently used together.

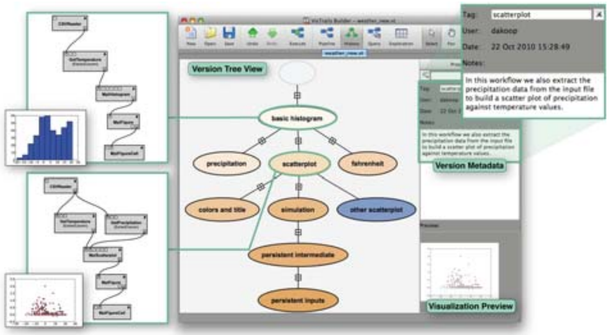
## 2.1 Comparison to Existing Tools

Spreadsheets are an extremely common tool for data curation. Vizier borrows the structural and visual programming elements of spreadsheets, allowing users to freely interact with data and gracefully supporting corner cases. Although Vizier will not provide full freedom of spreadsheets, curation efforts on spreadsheets can not be easily generalized into repeatable, shareable workflows. Scripting languages like Python are another common tool for curation work, but suffer from several limitations that Vizier addresses. First, the link between output and code is unidirectional: Edits to the output are overwritten the next time the code runs, making it harder to apply one-off changes or to explore hypothetical what-if scenarios. Moreover, once a curation script is developed through interactive design, it must still be manually adapted for parallel execution via map/reduce or a tool like apache spark [212].

Recently, several new tools for data cleaning or "wrangling" have emerged from academia and industry and are gaining traction in the data science community. NADEEF [61, 73, 74, 82] uses a rule-based system to detect and repair violated constraints. The set-at-a-time interaction model of a rule-based system works well when a user knows what properties the data should satisfy upfront, but does not permit easy discovery of such properties as in Vizier. SampleClean [122, 154, 229] uses sample-based cleaning to make unbiased estimates for aggregate results over messy or dirty data; This technique is orthogonal to Vizier's workflow generalization, and could conceivably be eventually incorporated into Vizier. Habitat [130] is an extensible framework for data curation, provenance, and discovery, into which specialized modules and interfaces can be deployed; Vizier could eventually be adapted to use Habitat as a deployment target. The Data Tamer project [120, 213], commercialized as Tamr, focuses more on data integration issues like schema matching and entity resolution. These are both important tasks in data curation, and appear as operators in Vizier.

The most similar production system is the Wrangler [121, 134, 185] project, commercialized as Trifacta [220]. Here too, the goal is to develop a repeatable curation workflow. We borrow Trifacta's sample-based development model, its ability to apply global edits directly on its output, as well as the idea of predictive suggestions. However, Vizier distinguishes itself in four ways. First, Trifacta forces users into the clean-before-use model and is thus optimized for developing only generalized set-at-a-time cleaning programs. Second, Vizier's notebook metaphor allows users to explore the data simultaneously from multiple perspectives or with different hypothetical modifications. Third, Vizier tracks provenance through the workflow, making it easier to debug values and sanity check results. Finally, provenance also permits Vizier to *safely* automate common tasks, even in the presence of ambiguity; if they are later found to be incorrect, the operations can be undone.

## 2.2 VisTrails



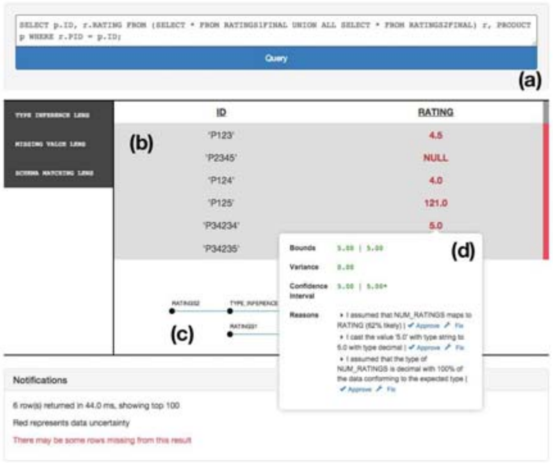
Tag: scatterplot
User: dakoop
Date: 22 Oct 2010 15:28:49
Notes:
In this workflow we also extract the precipitation data from the input file to build a scatter plot of precipitation against temperature values.

Figure 4: In VisTrails, provenance of exploration is represented as a version tree. Each node represents a workflow and an edge between two nodes encodes the set of changes applied to the parent to derive the child.

The open-source VisTrails system was designed to support exploratory and creative computational tasks [41,86,100,133,204,210,224,225], including data exploration, visualization, mining, machine learning, and simulations. The system allows users to create complex workflows (computational processes) that encompass important steps of data science, from data gathering and manipulation to complex analyses and visualizations. A new concept we introduced with VisTrails is the *provenance of workflow evolution* [40, 98]. In contrast to previous workflow and visualization systems which maintain provenance only for derived data products, VisTrails treats workflows as first-class data items and maintains their provenance (see *version tree* in Figure 4). Workflow evolution provenance supports reflective reasoning, allowing users to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward [176]. It also works as a *version control system for computational pipelines*, thus naturally supporting collaboration [78]. Users can easily *navigate through the space of workflows* created for a given investigation, visually *compare workflows and their results*, and *explore parameter spaces* [98]. Users are encouraged to *re-use knowledge by exploring and leveraging provenance information* through specific, easy-to-use components. These include a query-by-example interface, a mechanism for refining workflows by analogy [203], and a recommendation system that aids users in the design of workflows [152]. Pipelines and their provenance can be shared through Web-based interfaces, allowing *others to validate and reproduce computational experiments* [161, 192, 197]. The system has an active community of developers and users and has been adopted in several scientific projects, both nationally and internationally, across different domains. It has enabled and supported new research in environmental science [18,48,58,127,129], psychiatry [8], astronomy [219], cosmology [7], high-energy physics [68], molecular modeling [124], quantum physics [29, 85], earth observation [60, 222] and habitat modeling [165]. Besides being a stand-alone system, VisTrails has been used as a key component of domain-specific tools including DoE's UV-CDAT [191,195,221]; USGS's SAHM [165,190]; and NASA's DV3D [227]. VisTrails was featured as an NSF Discovery [178].

## 2.3 Mimir

Mimir [167, 231, 232] is a system that extends existing relational databases with support for so-called on-demand, or pay-as-you-go data curation through lenses, already introduced above in the description of Vizier. Mimir's support comes through a form of an annotated cursor that identifies ambiguous values and rows who's presence in the result set is ambiguous. Furthermore, the annotated cursor can explain ambiguity, both through English statements like *"I replaced* `TIP` *with* `NULL` *on row 5239865 because you asked me to remove outliers,"* and by quantifying the effect of ambiguity on results, as in *"the* `AVERAGE(TIP)` *is* $\$10 \pm 2$ *with 95% confidence"*. To convey the output of this annotated cursor to users, Mimir's front-end interface (illustrated in Figure 5) provides a standard SQL interface (a). Ambiguous outputs are marked (b), and in addition to showing their lineage (c), Mimir produces explanations (d) upon request.



```
SELECT p.ID, r.RATING FROM (SELECT * FROM RATINGS1FINAL UNION ALL SELECT * FROM RATINGS2FINAL) r, PRODUCT
p WHERE r.PID = p.ID;
```

| ID | RATING |
|---|---|
| 'P123' | 4.5 |
| 'P2345' | NULL |
| 'P124' | 4.0 |
| 'P125' | 121.0 |
| 'P34234' | 5.0 |
| 'P34235' | |

Bounds 5.00 | 5.00

Variance 0.00

Confidence Interval 5.00 | 5.00+

Reasons

- I assumed that NUM_RATINGS maps to RATING (62% likely) | ✓ Approve Fix
- I cast the value '5.0' with type string to 5.0 with type decimal | ✓ Approve Fix
- I assumed that the type of NUM_RATINGS is decimal with 100% of the data conforming to the expected type | ✓ Approve Fix

Notifications

6 row(s) returned in 44.0 ms, showing top 100

Red represents data uncertainty

There may be some rows missing from this result

Figure 5: Mimir's User Interface shows and explains uncertain or ambiguous data

## 2.4 GProM

GProM (**G**eneric **Pro**venance **M**iddleware) [10, 11, 113, 174, 175] is a database-independent middleware for tracking and querying fine-grained provenance of database queries, updates, and transactions (see Figure 6). GProM supports multiple database backends and can be extended to new backends through plug-ins. The user interacts with the system through one of the system's declarative frontend languages (currently dialects of SQL and Datalog with constructs for requesting provenance). Importantly, GProM tracks data provenance in a non-intrusive manner, without requiring any changes to applications or the database backend. Using **reenactment** [10], a declarative replay technique which simulates the effects of past updates or transactions using queries with time travel, the provenance of an update or transaction is computed retroactively through replaying the operation instrumented to capture provenance. Time travel is supported by most commercial systems and can be implemented using, e.g., triggers in systems

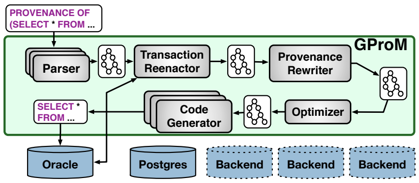


Figure 6: GProM system overview

that do not support it natively (see, e.g., [211]). GProM will be used to supply fine-grained provenance for curation operations be they queries or updates. Vizier uses this functionality to generate explanations for curated data and the curation process as well as to extract training data for the recommendations based on provenance. Furthermore, using reenactment it is possible to efficiently propagate changes to data and operations through a workflow during the exploratory phase of curation workflow construction.

# 3 Management Plan

Our implementation strategy for Vizier is to gradually integrate these three systems into a front-end interface that combines elements of VisTrails and Mimir, and a back-end component combining elements of Mimir and GProM. Our approach is iterative: many of our deliverables are useful stand-alone data curation tools in their own right that become progressively more powerful as they are extended and combined.

**Task 1: Bulletproof GProM and Mimir.** Both GProM and Mimir are stable enough for general use, but have primarily been developed as proofs of concept. The first task will be to thoroughly stress-test both systems to identify and repair any bugs, and to evaluate whether any critical core functionality necessary for Vizier is missing from either system.
**Deliverable**: *Production quality releases of GProM and Mimir*

**Task 2: Unify GProM and Mimir.** Our first integration target will be the back-end for Vizier: a data processing system with fine-grained provenance support, able to link outputs to edits applied both by users and by automated curation tasks. The back-end component will combine GProM's reenactment facilities with support for Mimir's lenses and explanations of ambiguity.
**Deliverable**: *A provenance and ambiguity-aware data processing system supporting queries and updates*

**Task 3: Add a notebook Interface to VisTrails.** VisTrails will serve as a central dispatcher for Vizier, linking the front-end interface to the fine-grained provenance and data-processing capabilities of the backend developed in Task 2. In this task, we will extend VisTrails with a UI shell based on Jupyter notebook, which will eventually serve as a front-end for Vizier. For this task, like VisTrails, the notebook shell will remain agnostic to the implementation of workflow steps.
**Deliverable**: *A notebook-style UI for VisTrails*

**Task 4: The Vizier DSL.** In parallel with the previous tasks, we will begin a research effort to design a DSL for our hybrid notebook/spreadsheet environment. As noted, our starting point will be relational algebra with additional operations based on SQL's DDL and DML, Mimir's lenses, as well as operators from existing data curation systems like Wrangler or NADEEF. Our goal is to create a language with a 1-1 mapping between edits applied to notebook's tabular output and the program generating that output.
**Deliverable**: *A language specification for Vizier's DSL and compliant parser*

**Task 5: A Prototype of Vizier's Hybrid Notebook and Spreadsheet UI.** As the DSL and the front- and back-end components mature, we will begin integrating them together using the DSL as glue. The first part of the integration process will be to create the hybrid notebook/spreadsheet UI proposed earlier by enabling spreadsheet-style editing for the notebook's tabular outputs. We then also need to create the UI elements necessary to mirror edits between the spreadsheet and code elements. Time permitting, we will explore visual cues to help users to follow the flow of code or data dependencies.
**Deliverable**: *A prototype front-end for Vizier*

**Task 6: Implement the DSL in the Vizier Back-End.** Simultaneously, we will add DSL support to the back-end. Both GProM and Mimir use an intermediate representation similar to relational algebra with updates. We anticipate that it will be possible to implement the DSL with minimal extension to this basic model (e.g., GProM already supports multiple front-end languages and, thus, it is reasonable to assume that adding support for the DSL would be possible with reasonable effort). Our initial plan will be to support local evaluation of the DSL on common formats including CSV and JSON, first to permit low-latency development, and second to create a prototype as quickly as possible. We will return to add support for partial or full deployment to external data management systems later (Task 8).
**Deliverable**: *A back-end with support for the Vizier DSL.*

**Task 7: Link the Back- and Front-end components.** The DSL developed in Task 4 serves as a common interface between the front- and back-end components of Vizier. Tasks 5 and 6 prepare these components for integration, establishing a DSL-aware front-end, and a back-end capable of evaluating DSL programs. We anticipate this task to require only minor refinements, but extensive stress-testing.
**Deliverable**: *A beta version of Vizier*

**Task 8: External Datasource Connectors.** Our next goal will be to enable access to data hosted in distributed resources such as relational databases, HDFS clusters, NoSQL data stores, generic REST APIs and other internet resources. A connector will require three components: First a way to access the complete data, second a way to sample from the available data, and finally an optional DSL translator to allow Vizier to deploy workflows to it.
**Deliverable**: *Connectors between Vizier and external data sources*

**Task 9: Extending Mimir's Library with Automated Data Curation Operations.** With the intent of getting users up and running as quickly as possible, our next goal will be to extend Mimir's existing library of Lenses with new data curation operations as well as operations that evaluate data quality and unearth interesting features of the data. Our initial goal will be to develop three modules: One for identifying potential outliers in a dataset [119], one for detecting correlations or dependencies between attributes and/or successive rows, and one for suggesting visual representations of a dataset [151]. We will revise this list based on feedback from our user community collaborators.
**Deliverable**: *Extended library of automated data curation operations and quality evaluation modules*

**Task 10: Provenance-based Suggestions.** Vizier will track the user's activities as a way to explain outputs to the user. The system will leverage the resulting repositories of both workflow provenance and fine-grained data provenance to offer suggestions to users [152,200], e.g., by identifying common relationships between user edits [134] or common properties of data elements that the user appears to be trying to fix.
**Deliverable**: *Automatic provenance-based recommendation module*

**Task 11: Revise and Refine.** We anticipate active community involvement in the development of Vizier. However, once we have assembled a beta version of the complete system, we expect to receive a substantially higher volume and quality of feedback from our collaborators and users. Our timeline includes an explicit period in which we expect to focus entirely on revising and refining the system.
**Deliverable**: *A full release of Vizier*

**Task 12: Version Tree Operations.** The change-based provenance adopted by VisTrails enables a series of operations that streamline exploration, including the ability to visually compare workflows [98] and to modify workflows by analogy [203]. We will extend VisTrails to support additional operations, including the ability to merge/reconcile curation workflows and propagate updates to multiple workflows.
**Deliverable**: *A comprehensive set of version tree manipulations required by exploratory curation*

## 3.1 Development Timeline

We have prepared a 3-year development timeline. In addition to the 3 PIs and a senior research engineer responsible for the project, we have budgeted for 4 students, 1 post-doctoral associate, and 2 developers for the full 3 years. The post-doc and developer hosted at NYU will be responsible for the implementation of Tasks 3-5, 7, 9-12. One developer hosted at UB will be implement the components in Tasks 1-2, 4, 6-8, 11. Students, one hosted at NYU and UB each, and two hosted at IIT, and the post-doc will be responsible for research & development components of the proposal. The post-doc will also serve as a bridge to our collaborators and work on outreach, giving talks and tutorials about the system. A yearly demonstration of technical capabilities, as requested by the DIBBS solicitation, will be based on the deliverables described for each task above, grouped into yearly units of work as described below.

**Year 1.** Preliminaries: Bulletproofing and unifying GProM and Mimir (Tasks 1-2) and preparing the user interface (Task 3). While these tasks are underway, all three sites will collaborate on a research effort to design a notebook/spreadsheet DSL (Task 4). Towards the end of the year, we expect both developers to begin prototyping efforts for their respective parts of Vizier (Tasks 5-6).

**Year 2.** Continued efforts to prototype the Front- and Back-end components of Vizier: We expect to see early efforts to integrate both elements (Task 7) begin early-to-mid year 2, and a beta version of the system available late in year 2. Research efforts in year 2 will include a preliminary exploration of provenance-aware learning and heuristic quality assessment (Tasks 9 and 10), and performance-tuning, for example through techniques like reenactment and incremental maintenance (Task 11).

**Year 3.** Year 3 will serve as a buffer for time-overruns and as a period of refinement: improving compatibility (Task 8), making Vizier smarter (Tasks 9 and 10), and incorporating community feedback (Task 11) and adding versioning (Task 12).

### 3.2 Evaluation Strategy

As our goal is to simplify data curation, we will evaluate Vizier and its components primarily through community feedback and expert studies facilitated by our collaborators. We will also conduct user studies to evaluate the effectiveness of specific features or interface elements on an as-needed basis, pending IRB approval for each case. As performance is also a significant user concern, we will evaluate Vizier's back-end components using data from collaborators including Airbus and the NYC Taxi and Limousine Commission, standard benchmarks like the MayBMS probabilistic data benchmark [217], openly accessible datasets, and benchmarking tools for cleaning and integration [12–15].

With the exception of Tasks 4–6, each sub-goal results in a complete software artifact that can be evaluated in isolation. Tasks 1–3 result in stand-alone components with clear, self-contained goals. Tasks 7–12 result in versions of Vizier with progressively more features that can be evaluated through expert studies and deployment at our collaborator's sites. In each case, the task has clear, measurable deliverables. The interface design of Task 5 will be evaluated and refined primarily through expert studies, and IRB approval permitting, user studies as well. The back-end design of Task 6 will be evaluated in terms of its performance and compliance with the DSL's specifications. The DSL created in Task 4 will be evaluated and refined based on its suitability for Tasks 5 and 6.

### 3.3 Risk Mitigation

The primary elements of Vizier are already well-established systems and we do not anticipate any significant issues with the feasibility of the proposed system. One possible risk is time overruns due to unforeseen difficulty or unexpected events. To mitigate this risk we have budgeted a part of year 3 as a buffer. If absolutely necessary, we can also scale back the goals of Tasks 8-10 to ensure the delivery of a stable system by the end of the grant period. Another risk is that of a developer leaving. We will mitigate this potential risk to institutional knowledge by ensuring that developers actively document their efforts, holding regular code reviews, and ensuring that the student researchers are actively engaged in the design process with the developers. A third risk is the possibility that by the time a prototype of Vizier is ready, our collaborators will no longer be able to provide us with their expected contributions. To mitigate this risk, we have engaged a large, diverse group of collaborators and will continuously seek out new partnerships over the course of the grant. Finally, the risk of running out of development resources is addressed in our sustainability plan.

## 4 Sustainability Plan

To ensure the continuation of the project beyond the three year grant period, we will build on our existing efforts to establish a strong community of users and developers with a vested interest in Vizier. To this end, we have already engaged a diverse set of expert leaders from communities in industry, academia, and government, all interested in actively participating in the design of Vizier and helping us to evaluate and refine the results. Our preliminary set of collaborators includes representatives from NYU's Center of Urban Science and Progress (CUSP), Furman Center, The Airbus group, the NYC Taxi and Limousine Commission (TLC), and Oracle (see letters of collaboration). Oracle, in particular, has already been supporting PIs Kennedy and Glavic's research through unrestricted gifts for the past 2 and 3 years, respectively. Their collaborators at Oracle, Dieter Gawlick and Zhen Hua-Liu, actively provide feedback and guidance for the Mimir and GProM projects. We have also begun efforts to reach out to additional communities (e.g., the informal survey in Section 1.3). We will expand on these efforts as additional components of Vizier are completed through further surveys, workshops, webinars, and demonstration videos. We have also budgeted for a publicly-available demonstration release to be hosted on a cloud-computing platform.

Our primary approach to building community buy-in will be to engage prospective users in the design and development of Vizier. The first step in this process will be to release Vizier under a permissive license (e.g., Apache), and provide access via a public code repository such as GitHub. However, simply releasing code is insufficient to build an engaged community. Ensuring high-quality user and developer documentation, as well as well-structured, well-commented, and readable code will help newcomers to make use of Vizier, and eventually to extend it as well. A stable community also needs communication. We will ensure the availability of vectors for both asynchronous communication (e.g., wikis or bug/feature request trackers) and synchronous communication (IRC, Slack, or similar) with project staff. We note that the use of such tools for communication is already necessary to support collaboration between the three project sites and that making these resources publicly accessible presents a negligible overhead. Finally, to provide the community with effective leadership, we will establish a common set of guidelines for code review and community governance during our first annual on-site meeting, and will review these guidelines during subsequent annual meetings.

# 5 Collaboration Plan

The PIs will closely collaborate on the project and synchronize their work progress using bi-weekly virtual meetings. These meetings will be attended by the PIs, the developers funded through this project, involved Ph.D. students and PostDocs, and collaborators from the target communities (see letters of collaboration). The purpose of these meetings is to (1) plan short-term and long-term implementation efforts, (2) elicit feedback from community leaders on the current version of the Vizier system, (3) strategic planning of development direction, and (4) coordination among all involved project sites. All three PIs have extensive experience working with users from the sciences and industry and have been involved in collaborations that span multiple disciplines. The PIs and project staff will also meet in person regularly at conferences (funds for conference travel are included in our respective budgets), and at a yearly on-site meeting. We have budgeted travel funds to allow PIs, developers, students, PostDocs, and key collaborators from target communities to gather at a single (rotating) project site for a workshop event once per year. These yearly meetings will revolve around a series of lightning talks from project staff to convey detailed information about each group's activities, areas of expertise, and project requirements. The lightning talks will be followed by a strategic planning session, a review of project and community governance policies, a breakout session for free-form discussions among project participants, and a wrap-up discussion to discuss outcomes arrived at during the break-out sessions. The PIs are using web-based collaboration software called GitLab that provides document sharing, a collaborative wiki, a bug tracker, and other tools for group management, as well as other cloud collaboration tools like Dropbox and Skype. If funded, project staff will also deploy and use tools for community-management, including group chat software (e.g., Slack, IRC, or similar), forums (e.g., PHPBB or Disqus), and blogging software (e.g., Wordpress). Finally, when Vizier is ready for a preliminary release, we will begin hosting it on a public open-source code repository such as GitHub.

# 6 Broader Impacts

The overarching goal of this proposal, to simplify data curation and exploration, has the potential to be transformative and positively impact the state of data science in many different domains. Our initial set of collaborators (see attached letters) includes representatives from government (TLC), academia (NYU's CUSP and Furman Center), and industry (Oracle and Airbus), all heavily invested in the issues of data quality that we are poised to address. Our informal survey turned up numerous colleagues and collaborators from across multiple fields of research and sectors of industry, all of whom grapple with data quality on a regular basis. Thus, the impact of the project can be immediate. Besides advancing the state of the art in data curation, the proposed work will improve research in other fields, notably, in social sciences. It will also contribute to an emerging field: urban science. Through the deployment of our tools within CUSP and the Furman Center, our work has the potential to impact a wide range of scientists and students, as well as the NYC agencies that are partnering with CUSP and Furman on various projects. If successful, our work will positively impact government by improving data quality, which in turn will lead to better planning and policies and more efficient operations. The ability to publish provenance-rich, high-quality data will also positively impact governance and citizen engagement.

**Integration of Research and Education.** We see education in a broader context. In our interactions with domain scientists, it has become clear that many of them are not up-to-date of recent developments around data curation techniques and tools. Thus, we believe that we not only need to educate our own students, but also inform the scientific community at large of the benefits and technologies related to curation. We are well positioned to reach "across our disciplines", given our on-going multi-disciplinary collaborations and the make up of our team. Four PhD students and one Postdoctoral researcher will participate in this project. The grant will also support the PIs in their existing educational outreach efforts: PI Kennedy is working with contacts at local high schools, gained through his membership in the WNY chapter of the ACM Computer Science Teacher's Association, to develop an open-data after-school program at high schools near UB. PI Glavic is actively working on integrating data cleaning and integration into IIT's CS graduate programs including the development of a new course on data integration and provenance. Vizier will be used as a tool in future installments of this course.

**Minority and Undergraduate Involvement.** We are committed to recruiting and mentoring minorities. All the PIs have been involved in different efforts to foster minority involvement. NYU Tandon School of Engineering ranked #1 in the nation by US News and World Report in Racial Diversity, and #3 in Economic Diversity among private universities in 2009. Typically, around half of our incoming domestic CS Ph.D. students are women or underrepresented minorities. It is well known that attracting minorities to STEM fields has been a great challenge. We believe that our research topic can contribute to attracting more minorities to computer science. The M.S. programs at NYU CUSP and at NYU CDS (which PI Freire directs) have close to 50% female enrollments. If funded, the PIs will also apply for REUs to facilitate involvement from undergraduate researchers at their respective universities.

**Technology Transfer and Software Tools.** Our team has an unassailable record of translating research results into practice across the sciences, including urban science, as well as to government, industry, and the general public. The software we will develop will be released as open source. Value-added open data derived by our methods will also will be made available under creative commons licenses where permitted.

# 7 Results from Prior NSF Support

**Dr. Oliver Kennedy.** Dr. Oliver Kennedy has been tenure-track faculty since Fall of 2012 and has been a PI on one NSF award. *Intellectual Merit:* Since receiving his first award 1.5 years ago, Dr. Kennedy's funding has resulted in 2 workshop publications [155,156] and one further conference paper under submission. *Broader Impacts:* NSF: CNS-1409551 (2014-2018; $976k) has funded the development of Ettu, a tool for summarizing query logs to enable insider threat analysis, as well as the collection and anonymization of a one-week trace of *all* SQL query activity at a major US bank. The project actively supports three PhD students, and an REU supplement is actively supporting one undergraduate and supported one undergraduate student who graduated in Winter of 2015. One graduate student has completed his thesis under Dr. Kennedy's supervision. Dr Kennedy is currently advising 5 PhD students (including 2 female students), one MS student, and one BS student, and co-advising 3 PhD students (including one female student).

**Dr. Juliana Freire.** Since joining academia in late 2002, Dr. Freire has been PI, co-PI, or senior investigator on fourteen NSF awards. *Intellectual Merit:* Dr. Freire's NSF funding has resulted in over 87 publications [4, 6, 18–27, 29, 30, 37, 40, 41, 44, 52–56, 63, 66, 71, 72, 75, 78–81, 83, 84, 86–88, 90, 91, 93, 94, 96–98, 100, 104, 127, 146–150, 152, 152, 159, 161–164, 166, 169–172, 172, 173, 186, 187, 193–199, 201, 203, 204, 206–208, 210, 223, 226], including a IEEE Visualization 2007 best paper award and an Eurographics Educational Program best paper award. *Broader Impact:* NSF IIS-0513692 (2005-2008; $499k)) [95] and NSF CNS-1405927 (2014-16; $530k) [92] have funded the development of VisTrails [86,224], an open-source data analysis and visualization tool that provides a comprehensive provenance infrastructure. VisTrails has been adopted in several scientific projects, both nationally and internationally, including in large NSF-funded projects [57, 58, 62, 188]; and it has had impact in different scientific domains [7,8,18,48,58,68,124,127,129,219]. Thirteen graduate students have completed their degrees under Dr. Freire's supervision—these include five female and three Hispanic students. She has also supervised post-doctoral assistants and several undergraduate students. She currently advises 7 Ph.D. students, 5 of which are from under-represented minorities.

**Dr. Boris Glavic.** does not have NSF support yet. He is currently advising three Ph.D. students (one female), three MS students, and is co-advising one Ph.D. student.

# 7 References

[1] ACHE. `https://github.com/ViDA-NYU/ache`.

[2] Hiralal Agrawal and Joseph R. Horgan. Dynamic program slicing. *SIGPLAN Not.*, 25(6):246–256, June 1990.

[3] Yanif Ahmad, **Oliver Kennedy**, Christoph Koch, and Milos Nikolic. DBToaster: Higher-order delta processing for dynamic, frequently fresh views. *PVLDB*, 2012.

[4] Sihem Amer-Yahia, Fang Du, and **Juliana Freire**. A Comprehensive Solution to the XML-to-Relational Mapping Problem. In *Proceedings of ACM WIDM*, pages 31–38, 2004.

[5] Y. Amsterdamer, S.B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting Lipstick on Pig: Enabling Database-style Workflow Provenance. *Proceedings of the VLDB Endowment*, 5(4):346–357, 2011.

[6] Erik Anderson, Steven P. Callahan, David A. Koop, Emanuele Santos, Carlos E. Scheidegger, Huy T. Vo, **Juliana Freire**, and Cláudio T. Silva. *VisTrails: Using Provenance to Streamline Data Exploration. In *Poster Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS)*, page 8, 2007.

[7] Erik W. Anderson, James P. Ahrens, Katrin Heitmann, Salman Habib, and Cláudio T. Silva. Provenance in comparative analysis: A study in cosmology. *Computing in Science and Engineering*, 10(3):30–37, 2008.

[8] Erik W. Anderson, Gil A. Preston, and Cláudio T. Silva. Towards development of a circuit based treatment for impaired memory: A multidisciplinary approach. In *IEEE EMBS Neural Engineering*, 2007.

[9] Bahareh Arab, Dieter Gawlick, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Reenacting transactions to compute their provenance. Technical Report IIT/CS-DB-2014-02, Illinois Institute of Technology, 2014.

[10] Bahareh Arab, Dieter Gawlick, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Formal foundations of reenactment and transaction provenance. Technical Report IIT/CS-DB-2016-01, Illinois Institute of Technology, 2016.

[11] Bahareh Arab, Dieter Gawlick, Venkatesh Radhakrishnan, Hao Guo, and **Boris Glavic**. A generic provenance middleware for database queries, updates, and transactions. In *Proceedings of the 6th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2014.

[12] Patricia C. Arocena, Radu Ciucanu, **Boris Glavic**, and Renée J. Miller. Gain Control over your Integration Evaluations. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 8(12):1960 – 1971, 2015.

[13] Patricia C. Arocena, **Boris Glavic**, Radu Ciucanu, and Renée J. Miller. The iBench Integration Metadata Generator. *Proceedings of the VLDB Endowment (PVLDB)*, 9(3):108–119, 2015.

[14] Patricia C. Arocena, **Boris Glavic**, Giansalvatore Mecca, Renée J. Miller, Paolo Papotti, and Donatello Santoro. Messing Up with Bart: Error Generation for Evaluating Data-Cleaning Algorithms. *Proceedings of the VLDB Endowment (PVLDB)*, 9(2):36–47, 2015.

[15] Patricia C. Arocena, **Boris Glavic**, and Renée J. Miller. Value invention for data exchange. In *Proceedings of the 39th International Conference on Management of Data (SIGMOD)*, pages 157–168, 2013.

[16] Eirik Bakke and Edward Benson. The schema-independent database ui a proposed holy grail and some suggestions. 2011.

[17] Eirik Bakke, David Karger, and Rob Miller. A spreadsheet-based user interface for managing plural relationships in structured data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2541–2550, New York, NY, USA, 2011. ACM. `http://doi.acm.org/10.1145/1978942.1979313`

[18] António Baptista, Bill Howe, **Juliana Freire**, David Maier, and Cláudio T. Silva. *Scientific Exploration in the Era of Ocean Observatories. *Computing in Science and Engineering*, 10(3):53–58, 2008.

[19] António Baptista, Todd Leen, Y. Zhang, A. Chawla, David Maier, Wu-Chi Feng, Wu-Chang Feng, Jon Walpole, Cláudio Silva, and **Juliana Freire**. Environmental observation and forecasting systems: Vision, challenges and successes of a prototype. In *Conference on Systems Science and Information Technology for Environmental Applications (ISEIS 2003)*, 2003.

[20] Denilson Barbosa, **Juliana Freire**, and Alberto Mendelzon. Information preservation in xml-to-relational mappings. In *Proceedings of XML Database Symposium (XSym)*, pages 66–81, 2004.

[21] Luciano Barbosa and **Juliana Freire**. Siphoning hidden-web data through keyword-based interfaces. In *Proceedings of the Brazilian Symposium on Databases (SBBD)*, pages 309–321, 2004.

[22] Luciano Barbosa and **Juliana Freire**. Automatically constructing collections of online databases (poster). In *Proceedings of CIKM*, pages 796–797, 2006.

[23] Luciano Barbosa and **Juliana Freire**. *Siphoning Hidden-Web Data through Keyword-Based Interfaces. *JIDM*, 1(1):133–144, 2010.

[24] Luciano Barbosa and **Juliana Freire**. *Siphoning Hidden-Web Data through Keyword-Based Interfaces: Retrospective. *JIDM*, 1(1):145–146, 2010.

[25] Luciano Barbosa and **Juliana Freire**. *Using Latent-Structure to Detect Objects on the Web. In *Proceedings of WebDB*, 2010.

[26] Luciano Barbosa, **Juliana Freire**, and Altigran Soares da Silva. Organizing hidden-web databases by clustering visible web documents. In *IEEE International Conference on Data Engineering (ICDE)*, pages 326–335, 2007.

[27] Luciano Barbosa, Hoa Nguyen, Thanh Nguyen, Ramesh Pinnamaneni, and **Juliana Freire**. *Creating and exploring web form repositories. In *SIGMOD Conference*, pages 1175–1178, 2010.

[28] Luciano Barbosa, Kien Pham, Cláudio Silva, Marcos Vieira, and **Juliana Freire**. *Structured Open Urban Data: Understanding the Landscape. *Big Data*, 2(3), 2014.

[29] B. Bauer et al. The alps project release 2.0: open source software for strongly correlated systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(05):P05001, 2011.

[30] Louis Bavoil, Steve Callahan, Patricia Crossno, **Juliana Freire**, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *VisTrails: Enabling Interactive Multiple-View Visualizations. In *Proceedings of IEEE Visualization*, pages 135–142, 2005.

[31] Vicki Been, Sam Dastrup, Ingrid Gould Ellen, Ben Gross, Andrew Hayashi, Susan Latham, Meghan Lewit, Josiah Madar, Vincent Reina, Mary Weselcouch, and Michael Williams. State of new york city's housing and neighborhoods, 2011.

[32] Vicki Been, Sam Dastrup, Ingrid Gould Ellen, Ben Gross, Andrew Hayashi, Susan Latham, Meghan Lewit, Josiah Madar, Vincent Reina, Mary Weselcouch, and Michael Williams. State of new york city's housing and neighborhoods, 2012.

[33] Aline Bessa, Fernando de Mesentier Silva, Rodrigo Frassetto Nogueira, Enrico Bertini, and **Juliana Freire**. *RioBusData: Outlier Detection in Bus Routes of Rio de Janeiro. In *IEEE Symposium on Visualization in Data Science*, 2015.

[34] The BirdVis System. `http://www.birdvis.org`.

[35] Jose A. Blakeley, Per-Ake Larson, and Frank Wm Tompa. Efficiently updating materialized views. *SIGMOD Rec.*, 15(2):61–71, June 1986.

[36] Michael R. Bloomberg and David Yassky. 2014 taxicab fact book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[37] Philippe Bonnet, Stefan Manegold, Matias Bjørling, Wei Cao, Javier Gonzalez, Joel Granados, Nancy Hall, Stratos Idreos, Milena Ivanova, Ryan Johnson, David Koop, Tim Kraska, René Müller, Dan Olteanu, Paolo Papotti, Christine Reilly, Dimitris Tsirogiannis, Cong Yu, **Juliana Freire**, and Dennis Shasha. Repeatability and workability evaluation of sigmod 2011. *SIGMOD Record*, 40(2):45–48, 2011.

[38] O. Peter Buneman and Eric K. Clemons. Efficiently monitoring relational databases. *ACM Trans. Database Syst.*, 4(3):368–382, September 1979.

[39] Peter Buneman, Shamim Naqvi, Val Tannen, and Limsson Wong. Principles of programming with complex objects and collection types. *Theoretical Computer Science*, 149(1):3 – 48, 1995. Fourth International Conference on Database Theory (ICDT '92).

[40] Steve Callahan, **Juliana Freire**, Emanuele Santos, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *Managing the Evolution of Dataflows with VisTrails *(Extended Abstract).* In *IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow)*, 2006.

[41] Steve Callahan, **Juliana Freire**, Emanuele Santos, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *VisTrails: Visualization meets Data Management. In *Proceedings of ACM SIGMOD*, pages 745–747, 2006.

[42] Steven Callahan, **Juliana Freire**, Emanuele Santos, Carlos Eduardo Scheidegger, Claudio T. Silva, and Huy Vo. VisTrails: Visualization meets Data Management. In *SIGMOD '06: Proceedings of the 32th SIGMOD International Conference on Management of Data (demonstration)*, pages 745–747, 2006.

[43] Steven P Callahan, **Juliana Freire**, Emanuele Santos, Carlos E Scheidegger, Claudio T Silva, and Huy T Vo. Managing the evolution of dataflows with vistrails. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 71–71. IEEE, 2006.

[44] Steven P. Callahan, **Juliana Freire**, Carlos Eduardo Scheidegger, Cláudio T. Silva, and Huy T. Vo. *Towards Provenance-Enabling ParaView. In *IPAW*, pages 120–127, 2008.

[45] Sean Capperis, Jorge De la Roca, Ingrid Gould Ellen, , Brian Karfunkel, Yiwen (Xavier) Kuai, Shannon Moriarty, Justin Steil, Eric Stern Michael Suher, Max Weselcouch, Mark Willis, and Jessica Yager. State of new york city's housing and neighborhoods, 2014.

[46] Sean Capperis, Jorge De la Roca, Kevin Findlan, Ingrid Gould Ellen, Josiah Madar, Shannon Moriarti, Justin Steil, Mary Weselcouch, and Mark Williams. State of new york city's housing and neighborhoods, 2013.

[47] Daniel Castellani Ribeiro, Huy T. Vo, **Juliana Freire**, and Cláudio T. Silva. *An Urban Data Profiler. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1389–1394. ACM, 2015.

[48] Cdat newsletter: Cdat v5.0 - highlights. http://www-pcmdi.llnl.gov/software-portal/Newsletter/Vol3/news.html, June 2007.

[49] Geoffrey Challen, Jerry Antony Ajay, Nick DiRienzo, **Oliver Kennedy**, Anudipa Maiti, Anandatirtha Nandugudi, Sriram Shantharam, Jinghao Shi, Guru Prasad Srinivasa, and Lukasz Ziarek. maybe we should enable more uncertain mobile app programming. In *HotMobile*, pages 105–110, 2015.
`http://doi.acm.org/10.1145/2699343.2699361`

[50] Surajit Chaudhuri, Ravi Krishnamurthy, Spyros Potamianos, and Kyuseok Shim. Optimizing queries with materialized views. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 190–200, Washington, DC, USA, 1995. IEEE Computer Society.
`http://dl.acm.org/citation.cfm?id=645480.655434`

[51] Fernando Chirigati and **Juliana Freire**. Towards integrating workflow and database provenance. In *Provenance and Annotation of Data and Processes*, pages 11–23. Springer, 2012.

[52] Fernando Seabra Chirigati and **Juliana Freire**. Towards integrating workflow and database provenance. In *IPAW*, pages 11–23, 2012.

[53] Fernando Seabra Chirigati, **Juliana Freire**, David Koop, and Cláudio T. Silva. *VisTrails provenance traces for benchmarking. In *EDBT/ICDT Workshops*, pages 323–324, 2013.

[54] Fernando Seabra Chirigati, Dennis Shasha, and **Juliana Freire**. Packing experiments for sharing and publication. In *SIGMOD Conference*, pages 977–980, 2013.

[55] Fernando Seabra Chirigati, Dennis Shasha, and **Juliana Freire**. *ReproZip: Using Provenance to Support Computational Reproducibility. In *USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2013.

[56] Fernando Seabra Chirigati, Matthias Troyer, Dennis Shasha, and **Juliana Freire**. *A Computational Reproducibility Benchmark. *IEEE Data Eng. Bull.*, 36(4):54–59, 2013.

[57] CLEO Experiment.

[58] NSF Center for Coastal Margin Observation and Prediction (CMOP).

[59] Daniel Crankshaw, Peter Bailis, Joseph E. Gonzalez, Haoyuan Li, Zhao Zhang, Michael J. Franklin, Ali Ghodsi, and Michael I. Jordan. The missing piece in complex analytics: Low latency, scalable model management and serving with velox. 09 2014.

[60] Council for Scientific and Industrial Research (CSIR) in South Africa.

[61] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. Nadeef: A commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 541–552, New York, NY, USA, 2013. ACM.
`http://doi.acm.org/10.1145/2463676.2465327`

[62] The Data Observation Network for Earth (DataONE).

[63] Susan B. Davidson, Sarah Cohen Boulakia, Anat Eyal, Bertram Ludäscher, Timothy M. McPhillips, Shawn Bowers, Manish Kumar Anand, and **Juliana Freire**. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007.

[64] Susan B. Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludäscher, Timothy McPhillips, Shawn Bowers, and **Juliana Freire**. Provenance in Scientific Workflow Systems. *IEEE Data Engineering Bulletin*, 32(4):44–50, 2007.

[65] Susan B Davidson and **Juliana Freire**. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350. ACM, 2008.

[66] Susan B. Davidson and **Juliana Freire**. *Provenance and scientific workflows: challenges and opportunities. In *SIGMOD*, pages 1345–1350, 2008.

[67] Amol Deshpande and Samuel Madden. Mauvedb: Supporting model-based user views in database systems. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 73–84, New York, NY, USA, 2006. ACM.
`http://doi.acm.org/10.1145/1142473.1142483`

[68] Andrew Dolgert, Lawrence Gibbons, Christopher D. Jones, Valentin Kuznetsov, Mirek Riedewald, Daniel Riley, Gregory J. Sharp, and Peter Wittich. Provenance in high-energy physics workflows. *Computing in Science and Engineering*, 10(3):22–29, 2008.

[69] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C.T. Silva. Using topological analysis to support event-guided exploration in urban data. *IEEE TVCG*, 20(12):2634–2643, 2014.

[70] H. Doraiswamy, H. Vo, C.T. Silva, and J. Freire. *A GPU-Based Index to Support Interactive Spatio-Temporal Queries over Historical Data. In *ICDE*, 2016.

[71] Harish Doraiswamy, Nivan Ferreira, Theodoros Damoulas, **Juliana Freire**, and Cláudio T. Silva. *Using Topological Analysis to Support Event-Guided Exploration in Urban Data. *IEEE Trans. Vis. Comput. Graph.TVCG*, 20(12):2634–2643, 2014.

[72] Fang Du, Sihem Amer-Yahia, and **Juliana Freire**. Shrex: Managing xml documents in relational databases. In *Proceedings of VLDB*, pages 1297–1300, 2004.

[73] Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Arnulfo Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF: A generalized data cleaning system*, volume 6, pages 1218–1221. 12 edition, 8 2013.

[74] Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Si Yin. Nadeef: A generalized data cleaning system. *Proc. VLDB Endow.*, 6(12):1218–1221, August 2013.

[75] Eric Eide, Tim Stack, Leigh Stoller, **Juliana Freire**, and Jay Lepreau. *Integrated scientific workflow management for the Emulab network testbed. In *Proceedings of USENIX*, pages 363–368, 2006.

[76] I.G. Ellen, J. Lacoe, and C.A. Sharygin. Do foreclosures cause crime? *Journal of Urban Economics*, 74:59–70, 2013.

[77] Ingrid Gould Ellen and Johanna Ruth Lacoe. Do foreclosures cause crime? Technical report, New York University, 2013.

[78] Tommy Ellkvist, David Koop, Erik W. Anderson, **Juliana Freire**, and Cláudio T. Silva. Using provenance to support real-time collaborative design of workflows. In *IPAW*, pages 266–279, 2008.

[79] Tommy Ellkvist, David Koop, **Juliana Freire**, Cláudio Silva, and Lena Strömbck. *Using Mediation to Achieve Provenance Interoperability (Extended Abstract). In *IEEE International Conference on eScience*, pages 398–399, 2008.

[80] Tommy Ellkvist, Lena Strömbäck, Lauro Didier Lins, and **Juliana Freire**. A first study on strategies for generating workflow snippets. In *Proceedings of the ACM SIGMOD Intenational Workshop on Keyword Search on Structured Data (KEYS)*, pages 15–20, 2009.

[81] Tommy Ellqvist, David Koop, **Juliana Freire**, Claudio Silva, and Lena Stromback. Using mediation to achieve provenance interoperability. *IEEE Congress on Services*, pages 291–298, 2009.

[82] Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. Nadeef/er: Generic and interactive entity resolution. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1071–1074, New York, NY, USA, 2014. ACM.
`http://doi.acm.org/10.1145/2588555.2594511`

[83] Nivan Ferreira, Lauro Lins, Daniel Fink, Steve Kelling, Chris Wood, **Juliana Freire**, and Cláudio Silva. *BirdVis: Visualizing and Understanding Bird Populations. *IEEE Transactions on Visualization and Computer Graphics*, 17:2374–2383, 2011.

[84] Nivan Ferreira, Jorge Poco, Huy T. Vo, **Juliana Freire**, and Claudio T. Silva. *Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[85] M. H. Freedman, J. Gukelberger, M. B. Hastings, S. Trebst, M. Troyer, and Z. Wang. Galois conjugates of topological phases. *Phys. Rev. B*, 85:045414, Jan 2012.

[86] J. Freire, D. Koop, E. Santos, C. Scheidegger, C. Silva, and H. T. Vo. **The Architecture of Open Source Applications*, chapter VisTrails. Lulu.com, 2011.

[87] **Juliana Freire**. Provenance management: Challenges and opportunities. In *Datenbanksysteme in Business, Technologie und Web (BTW)*, page 4, 2009.

[88] **Juliana Freire** and Michael Benedikt. Managing xml data: An abridged overview. *IEEE Computing in Science & Engineering*, 6(4):12–19, 2004.

[89] **Juliana Freire**, Aline Bessa, Fernando Seabra Chirigati, Huy Vo, and Kai Zhao. Exploring what not to clean in urban data:
a study using new york city taxi trips. *IEEE Data Eng. Bull.*, 2016.

[90] **Juliana Freire**, Philippe Bonnet, and Dennis Shasha. *Exploring the Coming Repositories of Reproducible Experiments: Challenges and Opportunities. *PVLDB*, 4(12):1494–1497, 2011.

[91] **Juliana Freire**, Philippe Bonnet, and Dennis Shasha. *Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *SIGMOD Conference*, pages 593–596, 2012.

[92] **Juliana Freire** and David Koop. Ci-en: Enhancing and supporting a community-based data analysis, visualization, and provenance platform, 2014.

[93] **Juliana Freire**, David Koop, Emanuele Santos, and Cláudio T. Silva. *Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

[94] **Juliana Freire**, M. Ramanath, and L. Zhang. A flexible infrastructure for gathering XML statistics and estimating query cardinality. In *IEEE International Conference on Data Engineering (ICDE)*, 2004.

[95] **Juliana Freire** and Cláudio Silva. Managing complex visualizations, July 2005.

[96] **Juliana Freire** and Cláudio Silva. *Simplifying the Design of Workflows for Large-Scale Data Exploration and Visualization. In *Proceedings of the Microsoft eScience Workshop*, 2008.

[97] **Juliana Freire** and Claudio Silva. *Towards Enabling Social Analysis of Scientific Data. In *ACM CHI Social Data Analysis Workshop*, 2008.

[98] **Juliana Freire**, Cláudio Silva, Steve Callahan, Emanuele Santos, Carlos Scheidegger, and Huy Vo. *Managing Rapidly-Evolving Scientific Workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18. Springer Verlag, 2006.

[99] **Juliana Freire** and Cláudio T. Silva. Making computations and publications reproducible with vistrails. *Computing in Science and Engineering*, 14(4):18–25, 2012.

[100] **Juliana Freire** and Cláudio T. Silva. *Making Computations and Publications Reproducible with VisTrails. *Computing in Science and Engineering*, 14(4):18–25, 2012.

[101] **Juliana Freire**, Cláudio T Silva, Steven P Callahan, Emanuele Santos, Carlos E Scheidegger, and Huy T Vo. Managing rapidly-evolving scientific workflows. In *Provenance and Annotation of Data*, pages 10–18. Springer, 2006.

[102] **Juliana Freire**, Cláudio T. Silva, Huy T. Vo, Harish Doraiswamy, Nivan Ferreira, and Jorge Poco. *Riding from Urban Data to Insight Using New York City Taxis. *IEEE Data Eng. Bull.*, 37(4):43–55, 2014.

[103] Furman center: Data services. `http://furmancenter.org/data`.

[104] Robert B. Gilbert, Fulvio Tonon, **Juliana Freire**, Cláudio Silva, and David R. Maidment. Visualizing uncertainty with uncertainty multiples. In *GeoCongress 2006: Geotechnical Engineering in the Information Technology Age*, pages 1–6. ASCE, 2006.

[105] **Boris Glavic**. Big data provenance: Challenges and implications for benchmarking. In *2nd Workshop on Big Data Benchmarking (WBDB)*, pages 72–80, 2012.

[106] **Boris Glavic** and Gustavo Alonso. Perm: Processing Provenance and Data on the same Data Model through Query Rewriting. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE)*, pages 174–185, 2009.

[107] **Boris Glavic** and Gustavo Alonso. Provenance for Nested Subqueries. In *Proceedings of the 12th International Conference on Extending Database Technology (EDBT)*, pages 982–993, 2009.

[108] **Boris Glavic** and Gustavo Alonso. The Perm Provenance Management System in Action. In *Proceedings of the 35th ACM SIGMOD International Conference on Management of Data (SIGMOD) (Demonstration Track)*, pages 1055–1058, 2009.

[109] **Boris Glavic**, Gustavo Alonso, Renée J. Miller, and Laura M. Haas. TRAMP: Understanding the Behavior of Schema Mappings through Provenance. *Proceedings of the Very Large Data Bases Endowment (PVLDB)*, 3(1):1314–1325, 2010.

[110] **Boris Glavic**, Jiang Du, Renée J. Miller, Gustavo Alonso, and Laura M. Haas. Debugging Data Exchange with Vagabond. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 4(12):1383–1386, 2011.

[111] **Boris Glavic**, Kyumars Sheykh Esmaili, Peter M. Fischer, and Nesime Tatbul. Ariadne: Managing fine-grained provenance on data streams. In *Proceedings of the 7th ACM International Conference on Distributed Event-Based Systems (DEBS)*, pages 291–320, 2013.

[112] **Boris Glavic**, Kyumars Sheykh Esmaili, Peter M. Fischer, and Nesime Tatbul. Efficient stream provenance via operator instrumentation. *Transactions on Internet Technology (TOIT)*, 13(1):7:1–7:26, 2014.

[113] **Boris Glavic**, Sven Köhler, Sean Riddle, and Bertram Ludäscher. *Towards Constraint-based Explanations for Answers and Non-Answers. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2015.

[114] **Boris Glavic**, Tanu Malik, and Quan Pham. *Making Database Applications Shareable. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP) (Poster)*, 2015.

[115] **Boris Glavic** and Renée J. Miller. Reexamining Some Holy Grails of Data Provenance. In *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[116] **Boris Glavic**, Renée J Miller, and Gustavo Alonso. Using sql for efficient generation and querying of provenance information. In *In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman*, pages 291–320. Springer, 2013.

[117] **Boris Glavic**, Javed Siddique, Periklis Andritsos, and Renée J. Miller. Provenance for data mining. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2013.

[118] Timothy Griffin and Leonid Libkin. Incremental maintenance of views with duplicates. *SIGMOD Rec.*, 24(2):328–339, May 1995.

[119] K.C. Gross, R.M. Singer, S.W. Wegerich, J.P. Herzog, R. VanAlstine, and F. Bockhorst. *Application of a model-based fault detection system to nuclear plant signals*. May 1997.

[120] M. Gubanov, M. Stonebraker, and D. Bruckner. Text and structured data fusion in data tamer at scale. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 1258–1261, March 2014.

[121] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 65–74, New York, NY, USA, 2011. ACM.
http://doi.acm.org/10.1145/2047196.2047205

[122] Daniel Haas, Sanjay Krishnan, Jiannan Wang, Michael J. Franklin, and Eugene Wu. Wisteria: Nurturing scalable data cleaning infrastructure. *Proc. VLDB Endow.*, 8(12):2004–2007, August 2015.

[123] Garcia-Molina Hector, Jeffrey D Ullman, and Jennifer Widom. *Database systems: The complete book*. Prentice-Hall, 2002.

[124] Randy Heiland, Maciek Swat, Benjamin Zaitlen, James Glazier, and Andrew Lumsdale. Workflows for parameter studies of multi-cell modeling (hpc). In *Proceedings of the ACM High Performance Computing Symposium*, 2010.

[125] Tuan-Anh Hoang-Vu, Vicki Been, Ingrid Gould Ellen, Max Weselcouch, and **Juliana Freire**. Towards understanding real-estate ownership in new york city: Opportunities and challenges. In *Proceedings of the Workshop on Data Science for Macro-Modeling*, 2014.

[126] Keren Mertens Horn, Ingrid Gould Ellen, and Amy Ellen Schwartz. Do housing choice voucher holders live near good schools? *Journal of Housing Economics*, 24(0):109–121, 2014.

[127] Bill Howe, Peter Lawson, Renee Bellinger, Erik Anderson, Emanuele Santos, **Juliana Freire**, Carlos Scheidegger, António Baptista, and Cláudio Silva. *End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. In *IEEE International Conference on eScience*, pages 127–134, 2008.

[128] Bill Howe, Peter Lawson, Renee Bellinger, Erik W. Anderson, Emanuele Santos, **Juliana Freire**, Carlos Eduardo Scheidegger, Antonio Baptista, and Claudio T. Silva. End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. In *eScience '08: Proceedings of the 4th IEEE International Conference on eScience*, pages 127–134, 2008.

[129] Bill Howe, Claudio Silva, and **Juliana Freire**. A science cloud on your desktop: Vistrails + gridfields, 2009.

[130] Zachary G Ives, Zhepeng Yan, Nan Zheng, Brian Litt, and Joost B Wagenaar. Looking at everything in context.

[131] H. V. Jagadish, Li Qian, and Arnab Nandi. Organic databases. *IJCSE*, 11(3):270–283, 2015.

[132] HV Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu. Making database systems usable. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 13–24, 2007.

[133] Emanuele Santos Juliana Freire and Cláudio Silva. Provenance-enabled data exploration and visualization with vistrails. In *SciDAC*, volume 125, 2009.

[134] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3363–3372, New York, NY, USA, 2011. ACM.
http://doi.acm.org/10.1145/1978942.1979444

[135] Manos Karpathiotakis, Ioannis Alagiannis, Thomas Heinis, Miguel Branco, and Anastasia Ailamaki. Just-In-Time Data Virtualization: Lightweight Data Management with ViDa. In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2015.

[136] Yannis Katsis, Yoav Freund, and Yannis Papakonstantinou. Combining databases and signal processing in plato. In *CIDR*, 2015.

[137] O. Kennedy, C. Koch, and A. Demers. Dynamic approaches to in-network aggregation. In *ICDE*, pages 1331–1334, 2009.

[138] **Oliver Kennedy**, Yanif Ahmad, and Christoph Koch. DBToaster: Agile views for a dynamic data management system. In *CIDR*, pages 284–295, 2011.

[139] **Oliver Kennedy**, Jerry Ajay, Geoffrey Challen, and Lukasz Ziarek. Pocket Data: The need for TPC-MOBILE. In *TPC Technology Conference on Performance Evaluation & Benchmarking*, 2015.

[140] **Oliver Kennedy** and Christoph Koch. Pip: A database system for great and small expectations. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 157–168. IEEE, 2010.

[141] **Oliver Kennedy**, Steve Lee, Charles Loboz, Slawek Smyl, and Suman Nath. Fuzzy prophet: Parameter exploration in uncertain enterprise scenarios. In *SIGMOD*, pages 1303–1306, 2011.
`http://doi.acm.org/10.1145/1989323.1989482`

[142] **Oliver Kennedy** and Suman Nath. Jigsaw: Efficient optimization over uncertain enterprise data. In *SIGMOD*, pages 829–840, 2011.
`http://doi.acm.org/10.1145/1989323.1989410`

[143] **Oliver Kennedy**, Ying Yang, Jan Chomicki, Ronny Fehling, Zhen Hua Liu, and Dieter Gawlick. *Enabling Real-Time Business Intelligence: International Workshops, BIRTE 2013, Riva del Garda, Italy, August 26, 2013, and BIRTE 2014, Hangzhou, China, September 1, 2014, Revised Selected Papers*, chapter Detecting the Temporal Context of Queries, pages 97–113. Springer Berlin Heidelberg, 2015.

[144] **Oliver Kennedy** and Lukasz Ziarek. Just-in-time data structures. In *CIDR*, 2015.

[145] Christoph Koch, Yanif Ahmad, **Oliver Andrzej Kennedy**, Milos Nikolic, Andres Nötzli, Daniel Lupei, and Amir Shaikhha. DBToaster: Higher-order delta processing for dynamic, frequently fresh views. *VLDBJ*, 2013.

[146] D. Koop, J. Freire, and C.T. Silva. Visual summaries for graph collections. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 57–64, 2013.

[147] David Koop and **Juliana Freire**. *Reorganizing Workflow Evolution Provenance. In *USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2014.

[148] David Koop, Emanuele Santos, Bela Bauer, Matthias Troyer, **Juliana Freire**, and Cláudio T. Silva. Bridging workflow and data provenance using strong links. In *SSDBM*, pages 397–415, 2010.

[149] David Koop, Emanuele Santos, Phillip Mates, Huy T. Vo, Philippe Bonnet, Bela Bauer, Brigitte Surer, Matthias Troyer, Dean N. Williams, Joel E. Tohline, **Juliana Freire**, and Cludio T. Silva. *A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers. *Procedia Computer Science*, 4:648–657, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011.

[150] David Koop, Carlos Scheidegger, **Juliana Freire**, and Cláudio T. Silva. *The Provenance of Workflow Upgrades. In *IPAW*, pages 2–16, 2010.

[151] David Koop, Carlos E Scheidegger, Steven P Callahan, **Juliana Freire**, and Cláudio T Silva. Viscomplete: Automating suggestions for visualization pipelines. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1691–1698, 2008.

[152] David Koop, Carlos Eduardo Scheidegger, Steven P. Callahan, **Juliana Freire**, and Cláudio T. Silva. *VisComplete: Automating Suggestions for Visualization Pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1691–1698, 2008.

[153] Bogdan Korel and Janusz Laski. Dynamic program slicing. *Information Processing Letters*, 29(3):155 – 163, 1988.

[154] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning while learning convex loss models. 01 2016.

[155] Gokhan Kul, Duc Thanh Luong, Ting Xie, Patrick Coonan, Varun Chandola, **Oliver Kennedy**, and Shambhu Upadhaya. * Ettu: Analyzing query intents in corporate databases. In *ERMIS*, 2016.

[156] Gokhan Kul and Shambhu Upadhyaya. * A preliminary cyber ontology for insider threats in the financial sector. In *Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats*, MIST '15, pages 75–78, New York, NY, USA, 2015. ACM.
`http://doi.acm.org/10.1145/2808783.2808793`

[157] Julia Maureen Letchner. *Lahar: warehousing markovian streams.* PhD thesis, University of Washington, 2010.

[158] Julie Letchner, Christopher Ré, Magdalena Balazinska, and Matthai Philipose. Lahar demonstration: warehousing markovian streams. *Proceedings of the VLDB Endowment*, 2(2):1610–1613, 2009.

[159] Lauro Lins, David Koop, Erik W. Anderson, Steven P. Callahan, Emanuele Santos, Carlos Eduardo Scheidegger, **Juliana Freire**, and Cláudio T. Silva. Examining statistics of workflow evolution provenance: A first study. In *SSDBM*, pages 573–579, 2008.

[160] Bin Liu and HV Jagadish. A spreadsheet algebra for a direct data manipulation query interface. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 417–428. IEEE, 2009.

[161] Phillip Mates, Emanuele Santos, **Juliana Freire**, and Cláudio T. Silva. *CrowdLabs: Social Analysis and Visualization for the Sciences. In *SSDBM*, pages 555–564, 2011.

[162] Sergio L. S. Mergen, **Juliana Freire**, and Carlos A. Heuser. Querying structured information sources on the web. In *iiWAS*, pages 470–476, 2008.

[163] Luc Moreau, **Juliana Freire**, Joe Futrelle, Robert McGrath, Jim Myers, and Patrick Paulson. The open provenance model (v1.00), December 2007.

[164] Luc Moreau, **Juliana Freire**, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick Paulson. The open provenance model: An overview. In *IPAW*, pages 323–326, 2008.

[165] J. Morisette, C. Jarnevich, T. Holcombe, C. Talbert, D. Ignizio, M. Talbert, C. T. Silva, D. Koop, A. Swanson, and N. Young. VisTrails SAHM: Visualization and workflow management for ecological niche modeling. *Ecography*, 2012. To appear.

[166] Leonardo Murta, Vanessa Braganholo, Fernando Seabra Chirigati, David Koop, and **Juliana Freire**. noworkflow: Capturing and analyzing provenance of scripts. In *IPAW*, 2014.

[167] Arindam Nandi, Ying Yang, **Oliver Kennedy**, **Boris Glavic**, Ronny Fehling, Zhen Hua Liu, and Dieter Gawlick. Mimir: Bringing ctables into practice. *CoRR*, abs/1601.00073, 2016.

[168] Suman K Nath, Seung Ho Lee, Slawomir Smyl, Charles Z Loboz, and **Oliver Andrzej Kennedy**. Efficient optimization over uncertain data, December 20 2012.

[169] Hoa Nguyen, Eun Yong Kang, and **Juliana Freire**. Automatically extracting form labels. In *ICDE*, pages 1498–1500, 2008.

[170] Hoa Nguyen, Thanh Nguyen, and **Juliana Freire**. Learning to extract form labels. *PVLDB*, 1(1):684–694, 2008.

[171] Huong Nguyen, Thanh Nguyen, Hoa Nguyen, and **Juliana Freire**. *Querying Wikipedia Documents and Relationships. In *Proceedings of WebDB*, 2010.

[172] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and **Juliana Freire**. *Multilingual Schema Matching for Wikipedia Infoboxes. *PVLDB*, 2012. Conditionally accepted.

[173] Thanh Hoang Nguyen, Hoa Nguyen, and **Juliana Freire**. Prusm: a prudent schema matching approach for web forms. In *CIKM*, pages 1385–1388, 2010.

[174] Xing Niu, Raghav Kapoor, Dieter Gawlick, Zhen Hua Liu, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Interoperability for provenance-aware databases using prov and json. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2015.

[175] Xing Niu, Raghav Kapoor, and **Boris Glavic**. Heuristic and cost-based optimization for provenance computation. In *TaPP*, 2015.

[176] Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine.* Addison Wesley, 1994.

[177] Donald A Norman. *The design of everyday things: Revised and expanded edition.* Basic books, 2013.

[178] Nsf discovery: A new vision for scientific visualizations. http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=114322 March 2009.

[179] NYC OpenData. `https://nycopendata.socrata.com`.

[180] Masayo Ota, Huy T. Vo, Cláudio T. Silva, and **Juliana Freire**. *A scalable approach for data-driven taxi ride-sharing simulation. In *IEEE International Conference on Big Data*, pages 888–897, 2015.

[181] Cesar Palomo, Zhan Guo, Cláudio T. Silva, and **Juliana Freire**. *Visually Exploring Transportation Schedules. *IEEE Trans. Vis. Comput. Graph.*, 22(1):170–179, 2016.

[182] Quan Pham, Tanu Malik, **Boris Glavic**, and Ian Foster. *LDV: Light-weight Database Virtualization. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE)*, pages 1179–1190, 2015.

[183] Quan Pham, Richard Whaling, **Boris Glavic**, and Tanu Malik. *Sharing and Reproducing Database Applications. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 8(12):1988 – 1999, 2015.

[184] Jorge Poco, Harish Doraiswamy, Huy Vo, João LD Comba, **Juliana Freire**, Cláudio Silva, et al. *Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions. *Computer Graphics Forum*, 34(3):161–170, 2015.

[185] Vijayshankar Raman and Joseph M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 381–390, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
`http://dl.acm.org/citation.cfm?id=645927.672045`

[186] Maya Ramanath, **Juliana Freire**, Jayant Haritsa, and Prasan Roy. Searching for efficient xml-to-relational mappings. In *Proceedings of XML Database Symposium (XSym)*, pages 19–36, 2003.

[187] Maya Ramanath, Lingzhi Zhang, **Juliana Freire**, and Jayant Haritsa. Imax: Incremental maintenance of schema-based xml statistics. In *IEEE International Conference on Data Engineering (ICDE)*, pages 273–284, 2005.

[188] Remote data analysis and visualization (rdav), 2009.

[189] ReproZip. https://github.com/ViDA-NYU/reprozip.

[190] Software for Assisted Habitat Modeling Package for VisTrails (SAHM: VisTrails).

[191] E. Santos, J. Poco, Yaxing Wei, Shishi Liu, B. Cook, D.N. Williams, and C.T. Silva. Uv-cdat: Analyzing climate datasets from a user's perspective. *Computing in Science and Engineering*, 15(1):94–103, 2013.

[192] Emanuele Santos. *Simplifying the Creation and Deployment of Collaborative Data Analysis and Visualization Tools*. PhD thesis, University of Utah, 2010.

[193] Emanuele Santos, **Juliana Freire**, and Claudio Silva. Information sharing in science 2.0: Challenges and opportunities. In *ACM CHI Workshop on The Changing Face of Digital Science: New Practices in Scientific Collaborations*, 2009.

[194] Emanuele Santos, **Juliana Freire**, Claudio Silva, Ayla Khan, Julien Tierny, Brad Grimm, Lauro Lins, Valerio Pascucci, Scott A. Klasky", Roselyne D. Barreto, and Norbert Podhorszki. Enabling advanced visualization tools in a simulation monitoring system. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 358–365. IEEE, December 2009.

[195] Emanuele Santos, David Koop, Thomas Maxwell, Charles Doutriaux, Tommy Ellqvist, Gerald Potter, **Juliana Freire**, Dean Williams, and Claudio Silva. Designing a provenance-based climate data analysis application. In *IPAW*, pages 214–219, 2012.

[196] Emanuele Santos, David Koop, Huy T. Vo, Erik W. Anderson, **Juliana Freire**, and Cláudio T. Silva. *Using Workflow Medleys to Streamline Exploratory Tasks. In *SSDBM*, pages 292–301, 2009.

[197] Emanuele Santos, Lauro Lins, James Ahrens, **Juliana Freire**, and Cláudio T. Silva. *VisMashup: Streamlining the Creation of Custom Visualization Applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539–1546, 2009.

[198] Emanuele Santos, Lauro Lins, James P. Ahrens, **Juliana Freire**, and Cláudio T. Silva. *A First Study on Clustering Collections of Workflow Graphs. In *IPAW*, pages 160–173, 2008.

[199] Emanuele Santos, Phillip Mates, Erik Anderson, Brad Grimm, **Juliana Freire**, and Claudio Silva. Towards supporting collaborative data analysis and visualization in a coastal margin observatories. ACM CSCW Workshop on The Changing Dynamics of Scientific Collaborations, 2010.

[200] Carlos Scheidegger, Huy Vo, David Koop, **Juliana Freire**, and Claudio Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization & Computer Graphics*, 13(6):1560–1567, 2007.

[201] Carlos Eduardo Scheidegger, David Koop, Emanuele Santos, Huy T. Vo, Steven P. Callahan, **Juliana Freire**, and Cláudio T. Silva. *Tackling the Provenance Challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008.

[202] Carlos Eduardo Scheidegger, Huy Vo, David Koop, **Juliana Freire**, and Claudio T. Silva. Querying and Re-using Workflows with VisTrails. In *SIGMOD '08: Proceedings of the 34th SIGMOD International Conference on Management of Data*, pages 1251–1254. ACM, 2008.

[203] Carlos Eduardo Scheidegger, Huy T. Vo, David Koop, **Juliana Freire**, and Cláudio T. Silva. *Querying and Creating Visualizations by Analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007.

[204] Carlos Eduardo Scheidegger, Huy T. Vo, David Koop, **Juliana Freire**, and Cláudio T. Silva. *Querying and re-using workflows with VisTrails. In *SIGMOD*, pages 1251–1254, 2008.

[205] Amy Ellen Schwartz, Ingrid Gould Ellen, Ioan Voicu, and Michael H. Schill. The external effects of place-based subsidized housing. *Regional Science and Urban Economics*, 36(6):679 – 707, 2006.

[206] Cláudio Silva, **Juliana Freire**, and Steven P. Callahan. *Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science & Engineering*, 9(5):82–89, 2007.

[207] Cláudio T. Silva, Erik Anderson, Emanuele Santos, and **Juliana Freire**. *Using VisTrails and Provenance for Teaching Scientific Visualization. In *Proceedings of the Eurographics Education Program*, 2010.

[208] Cláudio T. Silva and **Juliana Freire**. *Software Infrastructure for exploratory visualization and data analysis: past, present, and future. *Journal of Physics: Conference Series*, 25:012100 (15pp), 2008. SciDAC 2008 Conference.

[209] Claudio T. Silva, **Juliana Freire**, and Steven Callahan. Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science and Engineering*, 9(5):82–89, 2007.

[210] Cludio T. Silva, Erik Anderson, Emanuele Santos, and **Juliana Freire**. *Using VisTrails and Provenance for Teaching Scientific Visualization. *Computer Graphics Forum*, 30(1):75–84, 2011.

[211] Richard T Snodgrass, Jim Gray, and Jim Melton. *Developing time-oriented database applications in SQL*, volume 42. Morgan Kaufmann Publishers San Francisco, 2000.

[212] Apache Spark. Apache spark–lightning-fast cluster computing. 2014.

[213] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cherniack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.

[214] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic databases, synthesis lectures on data management*. Morgan & Claypool, 2011.

[215] Wang Chiew Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, 30(4):3–12, 2007.

[216] TaxiVis. `https://github.com/ViDA-NYU/TaxiVis`.

[217] The MayBMS project. Pdbench. http://pdbench.sourceforge.net.

[218] TLC Trip Record Data. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`, 2015.

[219] Joel E. Tohline, Jinghya Ge, Wesley Even, and Erik Anderson. A customized python module for cfd flow analysis within vistrails. *Computing in Science and Engineering*, 11(3):68–73, 2009.

[220] Trifacta. Trifacta wrangler. https://www.trifacta.com.

[221] Ultrascale Visualization - Climate Data Analysis Tools (UV-CDAT).

[222] TL Van Zyl, G McFerren, and A Vahed. Earth observation scientific workflows in a distributed computing environment. Technical Report 7727, CSIR, 2011.

[223] Karane Vieira, André Luiz Costa Carvalho, Klessius Berlt, Edleno S. Moura, Altigran S. Silva, and **Juliana Freire**. On finding templates on web collections. *World Wide Web*, 12(2):171–211, 2009.

[224] VisTrails. http://www.vistrails.org.

[225] VisTrails Users Guide.

[226] H. T. Vo, J. Bronson, B. Summa, J. Comba, J. Freire, B. Howe, V. Pascucci, , and C. Silva. *Parallel Visualization on Large Clusters using MapReduce. In *Proceedings of IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV)*, pages 81–88, 2011.

[227] vtDV3D VisTrails Package.

[228] Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakis, and Joseph M. Hellerstein. Bayesstore: Managing large, uncertain data repositories with probabilistic graphical models. *Proc. VLDB Endow.*, 1(1):340–351, August 2008.

[229] Jiannan Wang, Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Tim Kraska, and Tova Milo. A sample-and-clean framework for fast and accurate query processing on dirty data. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 469–480, New York, NY, USA, 2014. ACM.
`http://doi.acm.org/10.1145/2588555.2610505`

[230] Mark Weiser. Program slicing. In *Proceedings of the 5th International Conference on Software Engineering*, ICSE '81, pages 439–449, Piscataway, NJ, USA, 1981. IEEE Press.
`http://dl.acm.org/citation.cfm?id=800078.802557`

[231] Ying Yang. On-demand query result cleaning. In *VLDB PhD Workshop*, 2014.

[232] Ying Yang, Niccolo Meneghetti, Ronny Fehling, Zhen Hua Liu, and **Oliver Kennedy**. Lenses: an on-demand approach to etl. *Proceedings of the VLDB Endowment*, 8(12):1578–1589, 2015.

# Biographical Sketch — Oliver Kennedy

## Education (Chronological)

- New York University; New York City, NY, USA *(Major: Comp. Sci.)* — *B.S.* 2005
- Stevens Institute of Technology; Hoboken, NJ, USA *(Major: Comp. Eng.)* — *B.E.* 2005
- Cornell University; Ithaca, NY, USA *(Major: Comp. Sci.)* — *M.S.* 2009
- Cornell University; Ithaca, NY, USA *(Major Area: Comp. Sci.)* — *Ph.D.* 2011
  Dissertation Title: "*Watch out for... What?: Monitoring and Uncertainty in Scientific Computing*"
  Advisor: Christoph Koch

## Professional Appointments (Reverse Chronological)

- *Assistant Professor,* Comp. Sci. and Eng., SUNY at Buffalo — Aug 2012–present
- *Postdoctoral Fellow,* Faculté IC, EPFL Switzerland — June 2011–July 2012

## Five Products Closely Related to Proposed Project

1. "Lenses: An On-Demand Approach to ETL" (paper; Yang, Meneghetti, Fehling, Hua-Liu, Kennedy) - *VLDB* 2015
2. "Detecting the Temporal Context of Queries" (paper; Kennedy, Yang, Chomicki, Fehling, Hua-Liu, Gawlick) - *BIRTE* 2014
3. "Jigsaw: Efficient optimization over uncertain enterprise data" (paper; Kennedy, Nath) - *SIGMOD* 2011
4. "PigOut: Making Multiple Hadoop Clusters Work Together" (paper; Jeon, Chandrasekhara Shen, Mehra, Kennedy, Ko) - *IEEE BigData* 2014
5. "Just in Time Datastructures" (paper; Kennedy, Ziarek) - *CIDR* 2015

## Five Other Significant Products

1. "PIP: A Database System for Great and Small Expectations" (paper; Kennedy, Koch) - *ICDE* 2010
2. "DBToaster: Higher-order Delta Processing for Dynamic, Frequently Fresh Views" (paper; Koch, Ahmad, Kennedy, Nicolic, Nötzli, Lupei, Shakhana) - *VLDBJ* 2014
3. "Pocket Data: The Need for TPC-MOBILE" (paper; Kennedy, Ajay, Challen, Ziarek) - *TPC-TC* 2015
4. "`maybe` We Should Enable More Uncertain Mobile App Programming" (paper; Challen, Ajay, DiRienzo, Kennedy, Maiti, Nandugudi, Prasad, Shantharam, Shi, Ziarek) - *HotMobile* 2015
5. "Ettu: Analyzing Query Intents in Corporate Databases" (paper; Kul, Luong, Xie, Coonan, Chandola, Kennedy, Upadhyaya) - *ERMIS* 2016

## Selected Professional and Synergistic Activities

- Professional Memberships: ACM (SIGMOD, CSTA), IEEE
- Teacher Development Workshops at WNY-CSTA (Fa2012), CSTA CS4HS (2013, 2014); with Lukasz Ziarek and Sarbani Banerjee
- PC/Panel Member: SIGMOD (2015-2017), VLDB (2017) NSF (2014-2015), SIGMOD Reproducability (2015-2016), HILDA (2016), ProvenanceWeek (2016)
- CSTA WNY Chapter Secretary (2014-2015)
- Volunteer: Science is Elementary-Buffalo, Liberty Partnerships

# Biographical Sketch
# Dr. Juliana Freire

Department of Computer Science and Engineering
New York University
6 Metrotech Place, Brooklyn, NY 11201

Phone: (718) 260-4128
*juliana.freire@nyu.edu*

## Professional Preparation

| | | |
|---|---|---|
| Federal University of Ceará (Brazil) | Computer Science | B.S., 1991 |
| State University of New York at Stony Brook | Computer Science | M.S., 1992 |
| State University of New York at Stony Brook | Computer Science | Ph.D., 1997 |

## Appointments

| | | |
|---|---|---|
| NYU Moore Sloan Data Science Environment | Executive Director | 7/2015– |
| NYU Center for Data Science | Professor | 7/2014– |
| NYU Center for Data Science | Director of Graduate Studies | 7/2014– |
| NYU Center for Urban Science and Progress | Faculty Member | 2/2014– |
| New York University | Professor | 1/2014– |
| Polythechnic Institute of New York University | Professor | 7/2011– |
| Courant Institute, New York University | Affiliated Professor | 12/2011– |
| University of Utah | Associate Professor | 7/2008–6/2011 |
| University of Utah | Assistant Professor | 7/2005–6/2008 |
| OGI School of Science & Engineering at OHSU | Assistant Professor | 9/2002–6/2006 |
| Bell Labs Research, Lucent | Member of Technical Staff | 12/1997–9/2002 |

## Products

### Five most relevant products

1. *Collecting and Analyzing Provenance on Interactive Notebooks: When IPython Meets No Workflow.* J. Pimentel, V. Braganholo, L. Murta, and J. Freire. Proceedings of USENIX TAPP, 2015. Open-source system available at `https://github.com/gems-uff/noworkflow`.

2. *An Urban Data Profiler.* D. Ribeiro, H. Vo, C. Silva, and J. Freire. Proceedings of WWW, pp. 1389-1394, 2015.

3. *Using Topological Analysis to Support Event-Guided Exploration in Urban Data.* H. Doraiswami, N. Ferreira, T. Damoulas, J. Freire, and C. Silva. IEEE TVCG, 20(12), pp. 2634-2643, 2014.

4. *Structured Open Urban Data: Understanding the Landscape.* L. Barbosa, K. Pham, C. Silva, M. Vieira, and J. Freire. Big Data Journal, 2(3): 144-154, 2014.

5. *Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips.* N. Ferreira, J. Poco, H. Vo, J. Freire, and C. Silva. IEEE TVCG 19(12): 2149-58, 2013. Open-source system available at `https://github.com/ViDA-NYU/TaxiVis`.

### Five other relevant products

1. *Towards Understanding Real-Estate Ownership in New York City: Opportunities and Challenges.* T. Hoang-Vu, V. Been, I. G. Ellen, M. Weselcouch and J. Freire. Proceedings of the Workshop on Data Science for Macro-Modeling, 2014.

2. *Making Computations and Publications Reproducible with VisTrails.* J. Freire and Clúdio Silva. Computing in Science and Engineering 14(4): 18-25, 2012. Open-source system available at `http://vistrails.org`.
3. *VisComplete: Automating Suggestions for Visualization Pipelines*, D. Koop,C. Scheidegger, S. Callahan, H. Vo, J. Freire, and C. Silva. IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1691-1698, 2008.
4. *Provenance for Computational Tasks: A Survey*, J. Freire, D. Koop, E. Santos, C. Silva. In IEEE Computing in Science & Engineering, 10(3), pp. 11-21, 2008.
5. *Querying and Creating Visualizations by Analogy*, C. Scheidegger, H. Vo, D. Koop, J. Freire, C. Silva. IEEE Transactions on Visualization and Computer Graphics, 13(6) pp. 1560-1567. *Best paper award at IEEE Visualization 2007.*

### Synergistic Activities

- **Awards:** ACM Fellow, 2014; Google Faculty Award, 2013; NSF CAREER, 2008; IBM Faculty Award, 2008 and 2014; Best paper award at IEEE Visualization 2007; Best paper award at Eurographics Educator Program, 2010; Best poster award at SBBD2009.
- **Conference organization:** *Demo Program Chair*: ACM SIGMOD, 2015. *Group Leader*: ACM SIGMOD, 2012 and 2017. *Program Chair*: Experiment and Analysis Track of the International Conference on Very Large Databases (VLDB), 2012. *Program Chair*: Workshop on the Theory and Practice of Provenance (TAPP), June 2011. *Program Chair*, World Wide Web Conference (WWW), April 2010; *Program Chair*, International Provenance and Annotation Workshop (IPAW), June 2008; *Program Chair*, HPDC Provenance Challenge Workshop, June 2007; *Program Vice-chair*, International World Wide Web Conference (WWW), 2005, 2008;
- **Program Committee Member:** (over 70 events, including) International World Wide Web Conference (WWW), 2000–2003, 2005–2008; ACM SIGMOD International Conference on Management of Data (SIGMOD), 2004, 2007, 2012, 2014; International Conference on Very Large Databases (VLDB), 2003, 2007, 2008, 2012; IEEE International Conference on Data Engineering (ICDE), 2001, 2002, 2005, 2007, 2008;
- **Boards:** VLDB Journal, IEEE Data Engineering Bulletin, IEEE Transactions on Knowledge and Data Engineering, Journal of Information and Data Management, VLDB Endowment, NYC Taxi & Limousine Commission (TLC) Data/Technology Advisory Committee, SOCIAM Project (UK).
- **Minority Involvement**: Chair of the Diversity and Outreach committee, School of Computing, University of Utah (2007-2009); Member of the Academic Alliance for the National Center for Women in Technology (NCWIT).

# Biographical Sketch

**Boris Glavic**, Assistant Professor

Illinois Institute of Technology, Department of Computer Science, Stuart Building, 10 West 31st Street, Chicago, IL 60616. Phone: 312-567-5205. Email: `bglavic@cs.iit.edu`.

## a. Professional Preparation

| | | | | |
|---|---|---|---|---|
| **RWTH Aachen** | Aachen, Germany | Computer Science | **Diploma** | (2005) |
| **University of Zurich** | Zurich, Switzerland | Computer Science | **Ph. D.** | (2010) |
| **University of Toronto** | Toronto, Canada | Computer Science | **PostDoc** | (2010 - 2012) |

## b. Appointments

| | | |
|---|---|---|
| **Adjunct Assistant Professor (Status only)** | University of Toronto, Toronto, Canada | (2016-present) |
| **Assistant Professor** | Department of Computer Science, Illinois Institute of Technology, Chicago, USA | (2012-present) |
| **Post-Doctoral Fellow** | Department of Computer Science, University of Toronto, Toronto, Canada | (2010-2012) |
| **Research Assistant** | Department of Computer Science, University of Zurich, Zurich, Switzerland | (2005-2010) |

## c. Products

### (i) Products Related to Proposed Project

[1] B. Arab, D. Gawlick, V. Radhakrishnan, H. Guo, **B. Glavic**. *A Generic Provenance Middleware for Queries, Updates, and Transactions.* In **TaPP**, 2014.

[2] **Perm**: *Open-source provenance-aware database.*
Available at `https://sourceforge.net/projects/permdbms/`

[3] P. C. Arocena, **B. Glavic**, G. Mecca, and R. J. Miller, P. Papotti, D. Santoro. *Messing Up with BART: Error Generation For Evaluating Data-Cleaning Algorithms.* In **PVLDB**, vol. 9, no. 2, pp. 36–47, 2015.

[4] **B. Glavic** and G. Alonso. *Perm: Processing Provenance and Data on the same Data Model through Query Rewriting.* In **ICDE**, pp. 174–185, 2009.

[5] P. C. Arocena, **B. Glavic**, and R. J. Miller. *Value Invention for Data Exchange.* In **SIGMOD**, pp. 157-168, 2013

### (ii) Other Products

[1] P. C. Arocena, **B. Glavic**, R. Ciucanu, and R. J. Miller. *The iBench Integration Metadata Generator.* In **PVLDB**, vol. 9, no. 3, pp. 108–119, 2015.

[2] Q. Pham, T. Malik, **B. Glavic**, and I. Foster. *LDV: Light-weight Database Virtualization,* **ICDE**, pp. 1179-1190, 2015

[3] **B. Glavic**, G. Alonso, and R. J. Miller. *Using SQL for Efficient Generation and Querying of Provenance Information.* In In search of elegance in the theory and practice of computation, LNCS, pp. 291-320, 2013.

[4] **B. Glavic**, G. Alonso, R. J. Miller, and L. M. Haas. *TRAMP: Understanding the Behavior of Schema Mappings through Provenance.* In **PVLDB**, vol. 3, no. 1, pp. 1314–1325, 2010.

[5] **B. Glavic** and G. Alonso. *Provenance for Nested Subqueries.* In **EDBT**, pp. 982–993, 2009.

## d. Synergistic Activities

### (i) Program Committee Member and Chair

| | |
|---|---|
| 2017 | SIGMOD, VLDB |
| 2016 | IPAW (**PC Chair**), SIGMOD, ICDE (demo) |
| 2015 | SIGMOD, SIGMOD (demo), EDBT, ICDE, DATA, TaPP, WebDB, WBDB |
| 2014 | SIGMOD (demo), TaPP, DATA |
| 2013 | WBDB, TaPP, SSDBM |
| 2012 | AWM |
| 2011 | SIGMOD |

### (ii) Journal Reviewer

| | |
|---|---|
| 2015 | The VLDB Journal, Information Systems, Artificial Intelligence Review, JBI |
| 2014 | The VLDB Journal, ACM JDIQ |
| 2013 | ACM TOIT, IEEE TKDE |
| 2012 | ACM TODS |
| 2011 | ACM TODS, IEEE TKDE |

### (iii) Teaching

Hot topics in database systems: Data Provenance, Computer science for economists II (part 1) - introduction to databases, Advanced Database Organization, Database Organization, Data Integration, Warehousing, and Provenance, Doctoral Seminar.

### (iv) NSF Panel Reviews, other Grant Reviews, Event Reviews

| | |
|---|---|
| 2016 | Israel Science Foundation (ISF) - Grant Review |
| 2015 | Fields Institute, Toronto - Event Proposal Review |
| 2014 | Marsden Fund Grant Review (The Royal Society of New Zealand) |
| 2013 | NSF CISE-IIS panel |

### (v) University

Member of the graduate admission committee, Editor of the department newsletter, Member of the undergraduate committee.

# Biographical Sketch
# Dr. Heiko Müller

Center for Data Science
New York University
726 Broadway, New York, NY 10003

Phone: (929) 392-7621
*heiko.mueller@nyu.edu*

## Professional Preparation

| | | |
|---|---|---|
| Technical University of Berlin (Germany) | Computer Science | M.S., 2000 |
| Humboldt University Berlin (Germany) | Computer Science | Ph.D., 2006 |

## Appointments

| | | |
|---|---|---|
| NYU Center for Data Science | Research Engineer | 11/2015– |
| Data61, CSIRO (Australia) | Team Leader | 7/2012–11/2015 |
| ICT Centre, CSIRO (Australia) | Research Scientist | 10/2010–6/2012 |
| School of Informatics, University of Edinburgh (U.K.) | Research Fellow | 11/2006–9/2010 |
| Max-Planck-Institute for Molecular Genetics (Germany) | Software Engineer | 3/2006–10/2006 |
| Kelman GmbH (Germany) | Software Enginieer | 6/2000–12/2000 |

## Products

### Five most relevant products

1. *A Use Case in Semantic Modelling and Ranking for the Sensor Web.* L. Cabral, M. Compton, H. Müller. International Semantic Web Conference (IWSC), 2014.
2. *The Database Wiki Project: A General-Purpose Platform for Data Curation and Collaboration.* P. Buneman, J. Cheney, S. Lindley, H. Mller. SIGMOD Record, Vol. 40, No. 3, September 2011. Open-source system available at `https://github.com/jamescheney/database-wiki`
3. *Detecting Inconsistencies in Distributed Data.* W. Fan, F. Geerts, S. Ma, H. Müller IEEE International Conference on Data Engineering (ICDE), 2010.
4. *Sorting Hierarchical Data in External Memory for Archiving.* I. Koltsidas, H. Müller, S. Viglas Proceedings of the VLDB Endowment, Volume 1, Issue 1, 2008.
5. *XArch: Archiving Scientific and Reference Data.* H. Müller, P. Buneman, I. Koltsidas ACM International Conference on Management of Data (SIGMOD), Demo Track, 2008. Open-source system available at `http://xarch.sourceforge.net`

### Five other relevant products

1. *Link My Data: Community-based Curation of Environmental Sensor Data.* H. Müller, C. Peters, P. Taylor, A. Terhorst Intl. Symposium on Spatial and Temporal Databases (SSTD), Demo Track, 2013.
2. *Discovering conditional inclusion dependencies.* J. Bauckmann, Z. Abedjan, U. Leser, H. Müller, F. Naumann ACM Conf. on Information and Knowledge Management (CIKM), 2012.
3. *Improving data quality by source analysis.* H. Müller, J.-C. Freytag, U. Leser. ACM J. Data and Information Quality, Vol. 2, Issue 4, March 2012.
4. *Towards Content-Aware SPARQL Query Caching for Semantic Web Applications.* Y. Shu, M. Compton, H. Müller, K. Taylor Web Information Systems Engineering (WISE), 2013.
5. *Describing Differences between Databases.* H. Müller, J.-C. Freytag, U. Leser ACM Conf. on Information and Knowledge Management (CIKM), 2006.

## Synergistic Activities

- **Awards:** Asia Pacific ICT (APICTA) Award for Sustainability and Green IT, 2012 (The South Esk Hydrological Sensor Web: Next Generation Catchment Management); CSIRO ICT Centre Teamwork Award, 2012; Australian iAward for Green IT and Sustainability, 2011 (The South Esk Hydrological Sensor Web: Next Generation Catchment Management).
- **Program Committee Member:** GI-Fachtagung Datenbanksysteme für Business, Technologie und Web, 2009; International Colloquium on Data Provenance and Data Management for eScience, 2011;

# SUMMARY PROPOSAL BUDGET

YEAR 1

| | | | |
|---|---|---|---|
| | **FOR NSF USE ONLY** | | |
| ORGANIZATION<br>**SUNY at Buffalo** | PROPOSAL NO. | DURATION (months) | |
| | | Proposed | Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>**Oliver Kennedy** | AWARD NO. | | |

| | | | | | |
|---|---|---|---|---|---|
| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months | | | Funds Requested By proposer | Funds granted by NSF (if different) |
| | CAL | ACAD | SUMR | | |
| 1. **Oliver Kennedy - PI** | 0.00 | 0.00 | 1.00 | **10,757** | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( **0** ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | **0** | |
| 7. ( **1** ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | **10,757** | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( **0** ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | **0** | |
| 2. ( **1** ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | **100,000** | |
| 3. ( **1** ) GRADUATE STUDENTS | | | | **24,000** | |
| 4. ( **0** ) UNDERGRADUATE STUDENTS | | | | **0** | |
| 5. ( **0** ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | **0** | |
| 6. ( **0** ) OTHER | | | | **0** | |
| TOTAL SALARIES AND WAGES (A + B) | | | | **134,757** | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | **51,913** | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | **186,670** | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)<br>**See budget justification** **$** **0** | | | | | |
| TOTAL EQUIPMENT | | | | **0** | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | **5,000** | |
| 2. FOREIGN | | | | **4,000** | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ **0** | | | | | |
| 2. TRAVEL **0** | | | | | |
| 3. SUBSISTENCE **0** | | | | | |
| 4. OTHER **0** | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( **0** ) TOTAL PARTICIPANT COSTS | | | | **0** | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | **4,000** | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | **0** | |
| 3. CONSULTANT SERVICES | | | | **0** | |
| 4. COMPUTER SERVICES | | | | **3,900** | |
| 5. SUBAWARDS | | | | **537,534** | |
| 6. OTHER | | | | **19,782** | |
| TOTAL OTHER DIRECT COSTS | | | | **565,216** | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | **760,886** | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>**MTDC (Rate: 59.5000, Base: 253570)** | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | **150,874** | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | **911,760** | |
| K. SMALL BUSINESS FEE | | | | **0** | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | **911,760** | |

| | |
|---|---|
| M. COST SHARING PROPOSED LEVEL $ **0** | AGREED LEVEL IF DIFFERENT $ |

| | | | |
|---|---|---|---|
| PI/PD NAME<br>**Oliver Kennedy** | **FOR NSF USE ONLY** | | |
| | INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>**Amy Lagowski** | Date Checked | Date Of Rate Sheet | Initials - ORG |

1 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

YEAR 2

| | | | |
|---|---|---|---|
| ORGANIZATION<br>**SUNY at Buffalo** | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>**Oliver Kennedy** | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months CAL | ACAD | SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. **Oliver Kennedy - PI** | 0.00 | 0.00 | 1.00 | **10,972** | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( **0** ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | **0** | |
| 7. ( **1** ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | **10,972** | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( **0** ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | **0** | |
| 2. ( **1** ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | **102,000** | |
| 3. ( **1** ) GRADUATE STUDENTS | | | | **24,480** | |
| 4. ( **0** ) UNDERGRADUATE STUDENTS | | | | **0** | |
| 5. ( **0** ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | **0** | |
| 6. ( **0** ) OTHER | | | | **0** | |
| TOTAL SALARIES AND WAGES (A + B) | | | | **137,452** | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | **54,216** | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | **191,668** | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)<br>**$ 0** | | | | | |
| TOTAL EQUIPMENT | | | | **0** | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | **5,000** | |
| 2. FOREIGN | | | | **4,000** | |
| F. PARTICIPANT SUPPORT COSTS<br>1. STIPENDS $ **0**<br>2. TRAVEL **0**<br>3. SUBSISTENCE **0**<br>4. OTHER **0** | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( **0** ) TOTAL PARTICIPANT COSTS | | | | **0** | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | **2,000** | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | **0** | |
| 3. CONSULTANT SERVICES | | | | **0** | |
| 4. COMPUTER SERVICES | | | | **3,900** | |
| 5. SUBAWARDS | | | | **544,663** | |
| 6. OTHER | | | | **21,564** | |
| TOTAL OTHER DIRECT COSTS | | | | **572,127** | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | **772,795** | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>**MTDC (Rate: 59.5000, Base: 206568)** | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | **122,908** | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | **895,703** | |
| K. SMALL BUSINESS FEE | | | | **0** | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | **895,703** | |

| | | | |
|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ **0** | AGREED LEVEL IF DIFFERENT $ | | |
| PI/PD NAME<br>**Oliver Kennedy** | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>**Amy Lagowski** | Date Checked | Date Of Rate Sheet | Initials - ORG |

2 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

YEAR 3

| | | | |
|---|---|---|---|
| ORGANIZATION<br>SUNY at Buffalo | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Oliver Kennedy | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months CAL | ACAD | SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. Oliver Kennedy - PI | 0.00 | 0.00 | 1.00 | 11,191 | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 1 ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | 11,191 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 0 ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | 0 | |
| 2. ( 1 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | 104,040 | |
| 3. ( 1 ) GRADUATE STUDENTS | | | | 24,970 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 140,201 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 55,301 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 195,502 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)<br>$ 0 | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 5,000 | |
| 2. FOREIGN | | | | 4,000 | |
| F. PARTICIPANT SUPPORT COSTS<br>1. STIPENDS $ 0<br>2. TRAVEL 0<br>3. SUBSISTENCE 0<br>4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 2,000 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 3,900 | |
| 5. SUBAWARDS | | | | 559,137 | |
| 6. OTHER | | | | 23,508 | |
| TOTAL OTHER DIRECT COSTS | | | | 588,545 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 793,047 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>MTDC (Rate: 59.5000, Base: 210402) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 125,189 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 918,236 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 918,236 | |

| | | | |
|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ 0 | AGREED LEVEL IF DIFFERENT $ | | |
| PI/PD NAME<br>Oliver Kennedy | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | Date Checked | Date Of Rate Sheet | Initials - ORG |

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

Cumulative

| | | | |
|---|---|---|---|
| ORGANIZATION<br>**SUNY at Buffalo** | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>**Oliver Kennedy** | AWARD NO. | | |

| | | | | | |
|---|---|---|---|---|---|
| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months<br>CAL | <br>ACAD | <br>SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
| 1. **Oliver Kennedy - PI** | 0.00 | 0.00 | 3.00 | **32,920** | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | **0** | |
| 7. ( **1** ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 3.00 | **32,920** | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( **0** ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | **0** | |
| 2. ( **3** ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 36.00 | 0.00 | 0.00 | **306,040** | |
| 3. ( **3** ) GRADUATE STUDENTS | | | | **73,450** | |
| 4. ( **0** ) UNDERGRADUATE STUDENTS | | | | **0** | |
| 5. ( **0** ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | **0** | |
| 6. ( **0** ) OTHER | | | | **0** | |
| TOTAL SALARIES AND WAGES (A + B) | | | | **412,410** | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | **161,430** | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | **573,840** | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)<br>**$ 0** | | | | | |
| TOTAL EQUIPMENT | | | | **0** | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | **15,000** | |
| 2. FOREIGN | | | | **12,000** | |
| F. PARTICIPANT SUPPORT COSTS<br>1. STIPENDS $ **0**<br>2. TRAVEL **0**<br>3. SUBSISTENCE **0**<br>4. OTHER **0** | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( **0** ) TOTAL PARTICIPANT COSTS | | | | **0** | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | **8,000** | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | **0** | |
| 3. CONSULTANT SERVICES | | | | **0** | |
| 4. COMPUTER SERVICES | | | | **11,700** | |
| 5. SUBAWARDS | | | | **1,641,334** | |
| 6. OTHER | | | | **64,854** | |
| TOTAL OTHER DIRECT COSTS | | | | **1,725,888** | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | **2,326,728** | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | **398,971** | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | **2,725,699** | |
| K. SMALL BUSINESS FEE | | | | **0** | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | **2,725,699** | |

| | | | |
|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ **0** | AGREED LEVEL IF DIFFERENT $ | | |
| PI/PD NAME<br>**Oliver Kennedy** | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>**Amy Lagowski** | Date Checked | Date Of Rate Sheet | Initials - ORG |

C ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# Budget Justification

## Senior Personnel

PIs Kennedy is budgeted one month of summer salary and will apply his expertise and experience in the areas of databases, incremental computation, and probabilistic data curation. PI Kennedy will also take responsibility for (1) managing the project developer hosted by the University at Buffalo in their responsibilities as described below, (2) coordinating student-driven research efforts as described below, (3) coordinating his team's interactions with teams at other participating institutions, and (4) helping to drive the project's community-building efforts.

## Other Personnel

Funding is requested for one computer science graduate student assistant for three years. The two-semester and summer salary for the student is $24,000, with a standard 2% raise per year ($24,480 in Year 2 and $24,970 in Year 3). The graduate student assistant will be expected to (1) participate in the conceptual development of Vizier, (2) drive the project's research efforts, in particular those pertaining to probabilistic data curation, (3) document their findings through workshop, conference, and journal papers, blog posts, and other media as appropriate.

Funding is also requested for one developer for three years. The full-year salary for the developer is $100,000 with a standard 2% raise per year ($102,000 in Year 2 and $104,040 in Year 3). The developer will be responsible for the project's back-end implementation efforts; in particular focusing on aspects of Vizier related to the Mimir and GProM systems. For the first year, the developer will be tasked with bulletproofing, refining, and integrating both systems. In subsequent years, the developer's focus will remain on Vizier's compute and provenance management infrastructure.

## Fringe Benefits and Indirect Costs

Fringe benefit rates are based on the applicable federally negotiated rates published at
`http://www.research.buffalo.edu/sps/about/rates.cfm`

## Equipment

N/A

## Travel

Travel may include trips to NSF meetings, conferences and workshops, and any PI meetings. Major conferences such as SIGMOD, VLDB, and ICDE, typically last 4-5 days, and are located both domestically and internationally. Workshops are often affiliated with major conferences, and attendees frequently attend both. We have budgeted for both the student and developer to attend one domestic conference, and for the student to attend one foreign conference.

**Domestic Conferences** As an example of a domestic conference, we use SIGMOD 2016 being held in San Fransisco, CA. We anticipate a lodging cost of $99 per night and a $59 perdiem. The subtotal for 2 attendees over 5 nights is $790. We expect airfare of $630 and average conference registration fees of $600 per person for a total domestic travel cost of about $4000 per year.
**Foreign Conferences** As an example of a foreign conference, we use ICDE 2016 being held in Helsinki, Finland. We anticipate a lodging cost of $200 per person, and a $260 perdiem. The subtotal for 1 attendee over 5 nights is $2,300 per person. We expect airfare of $1000 and average conference registration fees of $700 per person for a total international travel cost of $4,000.
**Other Domestic Travel** We have budgeted an additional $1000 per year for travel to NSF PI meetings, a yearly on-site meeting/workshop, and community outreach efforts.

## Other Direct Costs

### Computer Services

The negotiated rate with the Department of Computer Science and Engineering for computer services is $156 per month of effort from faculty, students, and staff, or $4212 for 12 months of student effort, 12 months of developer effort, and 1 month of faculty effort.

### Materials and Supplies

$2,000 is requested in year 1 for Materials and Supplies to purchase desktop computers for the graduate research student, developer, and faculty working on this project. The computers will be used for code development, experimental evaluation, paper writing and typesetting and other efforts related to this project. An additional $2,000 is requested in all years to pay for resources on a cloud compute platform such as Amazon's Web Services, Microsoft Azure, or similar. Cloud resources will be used for experimental validation, performance testing, enabling collaboration between group participants, and for hosting demonstration versions of Vizier.

### Other

Tuition is budgeted at the standard University at Buffalo rates for a senior Graduate Research Assistant at 9 credit hours per semester. The anticipated out-of-state student tuition for one student is $19,782 in Year 1, increasing to $21,564 in Year 2 and $23,508 in Year 3.

### Indirect Costs

Indirect cost rates are based on the applicable federally negotiated rates published at `http://www.research.buffalo.edu/sps/about/rates.cfm`.

### Budget Overview

With the standard fringe rates, student tuition, and university overhead, the requested budget is $1,054,615.

# SUMMARY PROPOSAL BUDGET

YEAR 1

| | | | |
|---|---|---|---|
| ORGANIZATION<br>**Illinois Institute of Technology** | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>**Boris Glavic** | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months<br>CAL | <br>ACAD | <br>SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. **Boris Glavic - PI** | 0.00 | 0.00 | 1.00 | **10,751** | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( **0** ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | **0** | |
| 7. ( **1** ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | **10,751** | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( **0** ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | **0** | |
| 2. ( **0** ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 0.00 | 0.00 | 0.00 | **0** | |
| 3. ( **2** ) GRADUATE STUDENTS | | | | **48,000** | |
| 4. ( **0** ) UNDERGRADUATE STUDENTS | | | | **0** | |
| 5. ( **0** ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | **0** | |
| 6. ( **0** ) OTHER | | | | **0** | |
| TOTAL SALARIES AND WAGES (A + B) | | | | **58,751** | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | **806** | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | **59,557** | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | **0** | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | **3,000** | |
| 2. FOREIGN | | | | **3,000** | |
| F. PARTICIPANT SUPPORT COSTS<br>1. STIPENDS $ **0**<br>2. TRAVEL **0**<br>3. SUBSISTENCE **0**<br>4. OTHER **0** | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( **0** ) TOTAL PARTICIPANT COSTS | | | | **0** | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | **3,000** | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | **0** | |
| 3. CONSULTANT SERVICES | | | | **0** | |
| 4. COMPUTER SERVICES | | | | **0** | |
| 5. SUBAWARDS | | | | **0** | |
| 6. OTHER | | | | **25,200** | |
| TOTAL OTHER DIRECT COSTS | | | | **28,200** | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | **93,757** | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>**MTDC (Rate: 53.0000, Base: 68557)** | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | **36,335** | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | **130,092** | |
| K. SMALL BUSINESS FEE | | | | **0** | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | **130,092** | |

| | | | |
|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ **0** | AGREED LEVEL IF DIFFERENT $ | | |
| PI/PD NAME<br>**Boris Glavic** | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>**Amy Lagowski** | Date Checked | Date Of Rate Sheet | Initials - ORG |

1 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

YEAR 2

| | | | |
|---|---|---|---|
| ORGANIZATION<br>Illinois Institute of Technology | FOR NSF USE ONLY | | |
| | PROPOSAL NO. | DURATION (months) | |
| | | Proposed | Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Boris Glavic | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months CAL | ACAD | SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. Boris Glavic - PI | 0.00 | 0.00 | 1.00 | 11,181 | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 1 ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | 11,181 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 0 ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | 0 | |
| 2. ( 0 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 0.00 | 0.00 | 0.00 | 0 | |
| 3. ( 2 ) GRADUATE STUDENTS | | | | 49,920 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 61,101 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 839 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 61,940 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 3,000 | |
| 2. FOREIGN | | | | 3,000 | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ 0 | | | | | |
| 2. TRAVEL 0 | | | | | |
| 3. SUBSISTENCE 0 | | | | | |
| 4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 3,120 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 26,208 | |
| TOTAL OTHER DIRECT COSTS | | | | 29,328 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 97,268 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>MTDC (Rate: 53.0000, Base: 71060) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 37,662 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 134,930 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 134,930 | |
| M. COST SHARING PROPOSED LEVEL $ 0 | AGREED LEVEL IF DIFFERENT $ | | | | |

| | | | |
|---|---|---|---|
| PI/PD NAME<br>Boris Glavic | FOR NSF USE ONLY<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | Date Checked | Date Of Rate Sheet | Initials - ORG |

*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

# SUMMARY PROPOSAL BUDGET

YEAR 3

| FOR NSF USE ONLY | | |
|---|---|---|
| PROPOSAL NO. | DURATION (months) | |
| | Proposed | Granted |
| AWARD NO. | | |

ORGANIZATION
Illinois Institute of Technology

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR
Boris Glavic

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months | | | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| | CAL | ACAD | SUMR | | |
| 1. Boris Glavic - PI | 0.00 | 0.00 | 1.00 | 11,628 | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 1 ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 1.00 | 11,628 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 0 ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | 0 | |
| 2. ( 0 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 0.00 | 0.00 | 0.00 | 0 | |
| 3. ( 2 ) GRADUATE STUDENTS | | | | 51,916 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 63,544 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 872 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 64,416 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 3,000 | |
| 2. FOREIGN | | | | 3,000 | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ 0 | | | | | |
| 2. TRAVEL 0 | | | | | |
| 3. SUBSISTENCE 0 | | | | | |
| 4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 3,245 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 27,256 | |
| TOTAL OTHER DIRECT COSTS | | | | 30,501 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 100,917 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>MTDC (Rate: 53.0000, Base: 73661) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 39,040 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 139,957 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 139,957 | |

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

PI/PD NAME
Boris Glavic

ORG. REP. NAME*
Amy Lagowski

| FOR NSF USE ONLY | | |
|---|---|---|
| INDIRECT COST RATE VERIFICATION | | |
| Date Checked | Date Of Rate Sheet | Initials - ORG |
| | | |

*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

# SUMMARY PROPOSAL BUDGET

Cumulative

| | | | |
|---|---|---|---|
| | FOR NSF USE ONLY | | |
| ORGANIZATION<br>Illinois Institute of Technology | PROPOSAL NO. | DURATION (months) | |
| | | Proposed | Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Boris Glavic | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months | | | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| | CAL | ACAD | SUMR | | |
| 1. Boris Glavic - PI | 0.00 | 0.00 | 3.00 | 33,560 | |
| 2. | | | | | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 1 ) TOTAL SENIOR PERSONNEL (1 - 6) | 0.00 | 0.00 | 3.00 | 33,560 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 0 ) POST DOCTORAL SCHOLARS | 0.00 | 0.00 | 0.00 | 0 | |
| 2. ( 0 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 0.00 | 0.00 | 0.00 | 0 | |
| 3. ( 6 ) GRADUATE STUDENTS | | | | 149,836 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 183,396 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 2,517 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 185,913 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 9,000 | |
| 2. FOREIGN | | | | 9,000 | |
| F. PARTICIPANT SUPPORT COSTS<br>1. STIPENDS $ 0<br>2. TRAVEL 0<br>3. SUBSISTENCE 0<br>4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 9,365 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 78,664 | |
| TOTAL OTHER DIRECT COSTS | | | | 88,029 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 291,942 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 113,037 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 404,979 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 404,979 | |

| | | | | |
|---|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ 0 | AGREED LEVEL IF DIFFERENT $ | | | |
| PI/PD NAME<br>Boris Glavic | | FOR NSF USE ONLY | | |
| | | INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | | Date Checked | Date Of Rate Sheet | Initials - ORG |

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

## Budget Justification

### Salary - Senior Personnel

We request  1.00 month summer salary for the PI for the duration of the project. The PI together will be responsible for managing the project as a whole and supervising all involved students.

### Salary - Other Personnel

Two Ph.D. students will be supported for 3 years of the project working on discovery and repeatability components.

### Fringe Benefits

Fringe benefit rate is 18.6% for academic year salary and 7.5% for the summer month salary. Rate for staff is 19.7%.  Rate for students is 0%.

### Equipment

None requested.

### Materials and Supplies

We need to purchase devices so we can conduct experiments, for developing and debugging the system, and evaluate its scalability to ensure that it is efficient enough for the user communities that will use it. Typically we conduct experiments and debug on laptop or desktop machines. Additionally, some of the funds will be used to pay for cloud resources such as AWS to conduct experiments. We budget $3000 for each year for such purchases and other miscellaneous office expenses for the PI's lab.

### Travel

The PI requests travel support for him and his graduate students to attent top-level database conference and provenance workshops such as SIGMOD, VLDB, ICDE, EDBT, TaPP, and IPAW. These conference are held in both the USA and foreign countries. Participation in such forums is essential for disseminating the results and to keep informed of developments in the field. Additionally, these funds will also be used to attend meetings among the PIs and between the PIs and involved communities that will use Vizier.

Furthermore, traveling the PI expects to travel to NSF and to travel to give presentations to potential user communities in addition to the communities that have already committed to use the tool.

### Other Direct Costs

Tuition support for two Ph.D. students is requested at 9 credit hours per semester for all 3 years of the project with a total of $25,200 in the first year, $26,208 in the second year, and $27,256 in the third year.

### Indirect Costs

The indirect cost rate is 53% for the entire duration of the project.

An inflationary rate of 1.04% is used for all categories for all years of the project.

# SUMMARY PROPOSAL BUDGET

YEAR 1

| | | | |
|---|---|---|---|
| ORGANIZATION<br>New York University | FOR NSF USE ONLY<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Juliana Freire | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months<br>CAL | <br>ACAD | <br>SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. Juliana Freire - PI | 0.00 | 0.00 | 0.50 | 11,278 | |
| 2. Heiko Mueller - Co-I | 2.00 | 0.00 | 0.00 | 20,000 | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 2 ) TOTAL SENIOR PERSONNEL (1 - 6) | 2.00 | 0.00 | 0.50 | 31,278 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 1 ) POST DOCTORAL SCHOLARS | 6.00 | 0.00 | 0.00 | 37,500 | |
| 2. ( 1 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | 100,000 | |
| 3. ( 1 ) GRADUATE STUDENTS | | | | 34,800 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 203,578 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 49,790 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 253,368 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 10,000 | |
| 2. FOREIGN | | | | 8,000 | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ 0 | | | | | |
| 2. TRAVEL 0 | | | | | |
| 3. SUBSISTENCE 0 | | | | | |
| 4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 10,000 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 17,748 | |
| TOTAL OTHER DIRECT COSTS | | | | 27,748 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 299,116 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>MTDC (Rate: 38.5000, Base: 281367) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 108,326 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 407,442 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 407,442 | |

| | |
|---|---|
| M. COST SHARING PROPOSED LEVEL $ 0 | AGREED LEVEL IF DIFFERENT $ |

| | | | |
|---|---|---|---|
| PI/PD NAME<br>Juliana Freire | FOR NSF USE ONLY<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | Date Checked | Date Of Rate Sheet | Initials - ORG |

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

YEAR 2

| | | | |
|---|---|---|---|
| ORGANIZATION<br>New York University | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Juliana Freire | AWARD NO. | | |

| | NSF Funded Person-months | | | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | CAL | ACAD | SUMR | | |
| 1. Juliana Freire - PI | 0.00 | 0.00 | 0.50 | 11,560 | |
| 2. Heiko Mueller - Co-I | 2.00 | 0.00 | 0.00 | 20,500 | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 2 ) TOTAL SENIOR PERSONNEL (1 - 6) | 2.00 | 0.00 | 0.50 | 32,060 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 1 ) POST DOCTORAL SCHOLARS | 6.00 | 0.00 | 0.00 | 38,438 | |
| 2. ( 1 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | 102,500 | |
| 3. ( 1 ) GRADUATE STUDENTS | | | | 35,670 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 208,668 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 51,033 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 259,701 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 10,000 | |
| 2. FOREIGN | | | | 8,000 | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ 0 | | | | | |
| 2. TRAVEL 0 | | | | | |
| 3. SUBSISTENCE 0 | | | | | |
| 4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 5,000 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 18,192 | |
| TOTAL OTHER DIRECT COSTS | | | | 23,192 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 300,893 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>MTDC (Rate: 38.5000, Base: 282701) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 108,840 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 409,733 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 409,733 | |

| | | | |
|---|---|---|---|
| M. COST SHARING PROPOSED LEVEL $ 0 | AGREED LEVEL IF DIFFERENT $ | | |
| PI/PD NAME<br>Juliana Freire | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | Date Checked | Date Of Rate Sheet | Initials - ORG |

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

YEAR 3

| | | | |
|---|---|---|---|
| ORGANIZATION<br>**New York University** | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed | <br>Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>**Juliana Freire** | AWARD NO. | | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months<br>CAL | <br>ACAD | <br>SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. **Juliana Freire - PI** | 0.00 | 0.00 | 0.50 | **11,849** | |
| 2. **Heiko Mueller - Co-I** | 2.00 | 0.00 | 0.00 | **21,013** | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( **0** ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | **0** | |
| 7. ( **2** ) TOTAL SENIOR PERSONNEL (1 - 6) | 2.00 | 0.00 | 0.50 | **32,862** | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( **1** ) POST DOCTORAL SCHOLARS | 6.00 | 0.00 | 0.00 | **39,398** | |
| 2. ( **1** ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 12.00 | 0.00 | 0.00 | **105,063** | |
| 3. ( **1** ) GRADUATE STUDENTS | | | | **36,562** | |
| 4. ( **0** ) UNDERGRADUATE STUDENTS | | | | **0** | |
| 5. ( **0** ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | **0** | |
| 6. ( **0** ) OTHER | | | | **0** | |
| TOTAL SALARIES AND WAGES (A + B) | | | | **213,885** | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | **52,309** | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | **266,194** | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | **0** | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | **10,000** | |
| 2. FOREIGN | | | | **8,000** | |

F. PARTICIPANT SUPPORT COSTS

| | |
|---|---|
| 1. STIPENDS | $ **0** |
| 2. TRAVEL | **0** |
| 3. SUBSISTENCE | **0** |
| 4. OTHER | **0** |

| | | |
|---|---|---|
| TOTAL NUMBER OF PARTICIPANTS ( **0** ) TOTAL PARTICIPANT COSTS | **0** | |
| G. OTHER DIRECT COSTS | | |
| 1. MATERIALS AND SUPPLIES | **5,000** | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | **0** | |
| 3. CONSULTANT SERVICES | **0** | |
| 4. COMPUTER SERVICES | **0** | |
| 5. SUBAWARDS | **0** | |
| 6. OTHER | **18,646** | |
| TOTAL OTHER DIRECT COSTS | **23,646** | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | **307,840** | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)<br>**MTDC (Rate: 38.5000, Base: 289194)** | | |
| TOTAL INDIRECT COSTS (F&A) | **111,340** | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | **419,180** | |
| K. SMALL BUSINESS FEE | **0** | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | **419,180** | |

| | |
|---|---|
| M. COST SHARING PROPOSED LEVEL $ **0** | AGREED LEVEL IF DIFFERENT $ |

| | | | |
|---|---|---|---|
| PI/PD NAME<br>**Juliana Freire** | **FOR NSF USE ONLY**<br>INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>**Amy Lagowski** | Date Checked | Date Of Rate Sheet | Initials - ORG |

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

# SUMMARY PROPOSAL BUDGET

Cumulative

| | | |
|---|---|---|
| ORGANIZATION<br>New York University | **FOR NSF USE ONLY**<br>PROPOSAL NO. | DURATION (months)<br>Proposed / Granted |
| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR<br>Juliana Freire | AWARD NO. | |

| A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets) | NSF Funded Person-months CAL | ACAD | SUMR | Funds Requested By proposer | Funds granted by NSF (if different) |
|---|---|---|---|---|---|
| 1. Juliana Freire - PI | 0.00 | 0.00 | 1.50 | 34,687 | |
| 2. Heiko Mueller - Co-I | 6.00 | 0.00 | 0.00 | 61,513 | |
| 3. | | | | | |
| 4. | | | | | |
| 5. | | | | | |
| 6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE) | 0.00 | 0.00 | 0.00 | 0 | |
| 7. ( 2 ) TOTAL SENIOR PERSONNEL (1 - 6) | 6.00 | 0.00 | 1.50 | 96,200 | |
| B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS) | | | | | |
| 1. ( 3 ) POST DOCTORAL SCHOLARS | 18.00 | 0.00 | 0.00 | 115,336 | |
| 2. ( 3 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.) | 36.00 | 0.00 | 0.00 | 307,563 | |
| 3. ( 3 ) GRADUATE STUDENTS | | | | 107,032 | |
| 4. ( 0 ) UNDERGRADUATE STUDENTS | | | | 0 | |
| 5. ( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY) | | | | 0 | |
| 6. ( 0 ) OTHER | | | | 0 | |
| TOTAL SALARIES AND WAGES (A + B) | | | | 626,131 | |
| C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS) | | | | 153,132 | |
| TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C) | | | | 779,263 | |
| D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) | | | | | |
| TOTAL EQUIPMENT | | | | 0 | |
| E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS) | | | | 30,000 | |
| 2. FOREIGN | | | | 24,000 | |
| F. PARTICIPANT SUPPORT COSTS | | | | | |
| 1. STIPENDS $ 0 | | | | | |
| 2. TRAVEL 0 | | | | | |
| 3. SUBSISTENCE 0 | | | | | |
| 4. OTHER 0 | | | | | |
| TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS | | | | 0 | |
| G. OTHER DIRECT COSTS | | | | | |
| 1. MATERIALS AND SUPPLIES | | | | 20,000 | |
| 2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION | | | | 0 | |
| 3. CONSULTANT SERVICES | | | | 0 | |
| 4. COMPUTER SERVICES | | | | 0 | |
| 5. SUBAWARDS | | | | 0 | |
| 6. OTHER | | | | 54,586 | |
| TOTAL OTHER DIRECT COSTS | | | | 74,586 | |
| H. TOTAL DIRECT COSTS (A THROUGH G) | | | | 907,849 | |
| I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) | | | | | |
| TOTAL INDIRECT COSTS (F&A) | | | | 328,506 | |
| J. TOTAL DIRECT AND INDIRECT COSTS (H + I) | | | | 1,236,355 | |
| K. SMALL BUSINESS FEE | | | | 0 | |
| L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K) | | | | 1,236,355 | |

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

| | | | |
|---|---|---|---|
| PI/PD NAME<br>Juliana Freire | **FOR NSF USE ONLY** | | |
| | INDIRECT COST RATE VERIFICATION | | |
| ORG. REP. NAME*<br>Amy Lagowski | Date Checked | Date Of Rate Sheet | Initials - ORG |

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

# BUDGET JUSTIFICATION
## New York University Tandon School of Engineering
## Department of Computer Science & Engineering

### A. Senior Personnel:

PI: Juliana Freire, will serve as Principal Investigator on this research project and will devote 0.50 summer month during each year of the project. Dr. Freire will direct the project, having responsibility for the project outcomes.

Co-I: Heiko Mueller is an expert in data management, provenance and data cleaning. We are requesting 2.0 month of support per year for Dr. Mueller to perform the work outlined in the proposal.

Developer: We are requesting 12.0 month of support per year for a developer to perform the work outlined in the proposal.

Postdoc: We are requesting 6.0 month of support per year for a postdoc to perform the work outlined in the proposal.

### B. Other Personnel:

Graduate Research Assistant (GRA): We are requesting 12.0 months of support per year for a graduate student. The student will work on the tasks outlined in the proposal.

**C. Fringe Benefits:** The negotiated and approved fringe benefit rate for faculty & full time personnel is 29.5%. Salaries have been increased at a rate of 2.5% per year to reflect cost of living escalation for faculty, staff, and the GRA.

**D. Equipment**:

### E. Travel:

- **Domestic:** $10,000 is requested in each year of the project to allow the PI and/or other project personnel to travel for collaborative purposes and to attend conferences.
- **Foreign:** $8,000 is requested in each year of the project to allow the PI and/or other project personnel to travel for collaborative purposes and to attend conferences.

### G. Other Direct Costs:

**1. Materials & Supplies:** $10,000 is requested in year 1, and $5,000 is requested in years 2 and 3 of the project for the purchase of necessary materials and supplies. In year 1, materials include laptops and/or desktops for the project members.

**Other**: Graduate Students receive tuition remission in lieu of fringe benefits. Tuition Remission is calculated at 51% of salary for each student supported on the project.

**I. Indirect Costs:** F&A is at 38.5% of MTDC, per NYU-Tandon's negotiated and approved rate agreement with ONR, dated 6/23/2014.

## Current and Pending Support

**(See GPG Section II.D.8 for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Oliver Kennedy

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
CIF21 DIBBs: EI: Vizier, Streamlined Data Curation

Source of Support: NSF: ACI: DIBBS

Total Award Amount: $ 2725699 Total Award Period Covered: 01/2017 -- 12/2019

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
DD: Success in Partnership: Computing for all of STEM

Source of Support: NSF: DRL: STEM+C

Total Award Amount: $ 2500000 Total Award Period Covered: 08/2016 -- 07/2019

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
Intuitive Data Interpretation

Source of Support: Oracle University Relations

Total Award Amount: $ 89187 Total Award Period Covered: 05/2016 -- 05/2017

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
CI-P: Supporting Pocket Scale Data Management Research

Source of Support: NSF: CISE: IIS: CRI

Total Award Amount: $ 100000 Total Award Period Covered: 08/2016 -- 08/2017

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 0.5

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
III: Small: Just in Time Datastructures

Source of Support: NSF: CISE: IIS: III

Total Award Amount: $ 494274 Total Award Period Covered: 05/2016 -- 04/2019

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Current and Pending Support

**(See GPG Section II.D.8 for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Oliver Kennedy

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
Expressing Uncertainty Using the maybe System

Source of Support: Google Research Awards

Total Award Amount: $ 38656 Total Award Period Covered: 08/2015 -- 07/2016

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 0

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
Intuitive Data Interpretation

Source of Support: Oracle University Relations

Total Award Amount: $ 90455 Total Award Period Covered: 03/2015 -- 03/2016

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:
TWC: Medium: Collaborative: Data is Social: Exploiting Data Relationships to Detect Insider Attacks

Source of Support: NSF: CISE: SaTC

Total Award Amount: $ 959999.0 Total Award Period Covered: 09/2014 -- 08/2017

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal: 0 Acad: 0 Sumr: 1

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Current and Pending Support

**(See GPG Section II.C.2.h for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Juliana Freire

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: CI-EN: Enhancing and Supporting a Community-Based Data Analysis, Visualization

Source of Support: NSF

Total Award Amount: $ 499,962 Total Award Period Covered: 09/01/14 - 08/31/16

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.50

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: MRI: Acquisition of an Infrastructure for Prototyping Next-Generation Algorithms for Large-Scale Visualization, Data Processing and Analysis

Source of Support: NSF

Total Award Amount: $ 799,999 Total Award Period Covered: 09/01/12 - 08/31/16

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.00

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: Towards Locating and Exploring Hard-to-Find Information on the Web

Source of Support: DARPA

Total Award Amount: $ 3,600,000 Total Award Period Covered: 10/01/14 - 09/30/17

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 2.50 Sumr: 0.00

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: Post-doctoral Fellowship for Juliana Freire's Lab, NYU Center for Data Science

Source of Support: Vivaki, Inc.

Total Award Amount: $ 225,000 Total Award Period Covered: 03/31/14 - 08/31/20

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.00

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: Data Science Environments Program

Source of Support: Alfred P. Sloan Foundation

Total Award Amount: $ 2,600,000 Total Award Period Covered: 11/01/13 - 01/01/17

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Summ: 0.50

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Current and Pending Support

**(See GPG Section II.C.2.h for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Juliana Freire

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support
Project/Proposal Title: Data Science Environments Program

Source of Support: Gordon & Betty Moore Foundation
Total Award Amount: $ 10,000,000 Total Award Period Covered: 10/18/13 - 12/31/18
Location of Project: NYU
Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.50

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support
Project/Proposal Title: BIGDATA:F: Understanding Spatio-Temporal Urban Data through their Many-to-Many Relationships

Source of Support: NSF
Total Award Amount: $ 1,350,166 Total Award Period Covered: 09/01/16 - 08/31/19
Location of Project: NYU
Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support
Project/Proposal Title: CIF21 DIBBs: EI: Vizier, Streamlined Data Curation

Source of Support: NSF
Total Award Amount: $ 2,725,699 Total Award Period Covered: 01/01/17 - 12/31/19
Location of Project: NYU
Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.50

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Current and Pending Support

**(See GPG Section II.C.2.h for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Boris Glavic

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: NeTS: Medium: Ensuring Privacy, Accountability, and Traceability in Data Publishing, Sharing, and Trading Networks

Source of Support: NSF

Total Award Amount: $ 1,199,760 Total Award Period Covered: 08/01/16 - 07/31/20

Location of Project: Illinois Institute of Technology

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: III: Small: Scalable Exploration and Comprehensible Summarization of Why & Why-Not Provenance

Source of Support: NSF

Total Award Amount: $ 499,979 Total Award Period Covered: 06/01/16 - 05/31/19

Location of Project: Illinois Institute of Technology

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: III: Small: Collaborative Research: Parrot: Enabling Sharing and Reproducibility for Computational Experiments

Source of Support: NSF

Total Award Amount: $ 241,478 Total Award Period Covered: 06/01/16 - 05/31/19

Location of Project: Illinois Institute of Technology

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: CIF21 DIBBs: EI: Vizier, Streamlined Data Curation

Source of Support: NSF

Total Award Amount: $ 2,725,699 Total Award Period Covered: 01/01/17 - 12/31/19

Location of Project: University at Buffalo, Buffalo, NY

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Current and Pending Support

**(See GPG Section II.C.2.h for guidance on information to include on this form.)**

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Heiko Mueller

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: □ Current ☒ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title: CIF21 DIBBs: EI: Vizier, Streamlined Data Curation

Source of Support: NSF: ACI: DIBBS

Total Award Amount: $ 2,725,699 Total Award Period Covered: 01/01/17 - 01/12/19

Location of Project: NYU

Person-Months Per Year Committed to the Project. Cal:2.00 Acad: 0.00 Sumr: 0.00

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: □ Current □ Pending □ Submission Planned in Near Future □ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

USE ADDITIONAL SHEETS AS NECESSARY

# Facilities Statement

## University at Buffalo, SUNY

### Department of Computer Science and Engineering

General research facilities at UB's Department of Computer Science and Engineering include more than 200 i386 servers, x86_64 servers, SPARC-servers, PCs, and thin-client workstations. The systems include three Dell Optiplexes, twenty-two Dell PowerEdges, two Dell Precision, one Sun 5220, one Sun Blade 1000, two Sun Fire 280R, two Sun Fire V20Zs, more than 100 Sun Ray thin client terminals, one Sun Ultra-60s, and a Dell PowerEdge based virtual machine infrastructure. These systems are attached to a NetApp FAS2050A installed with 13 TB of disk space. Nine Dell Optiplex systems are configured as Linux and Windows student workstations. Laser printing resources are readily available. Program-specific research facilities include over 50 dedicated research systems. The Computer-Assisted Diagnostics Interventions (CADI) Lab houses a 21-member compute cluster connected to a data storage device with 10 TB of installed capacity. The CyberInfrastructure (CI) Lab houses a 14-member Open Science Grid (OSG) compute cluster connected to a 15 TB (installed) data storage device. The Bioinformatics, Database, Data Mining, and Multimedia Group houses dedicated Oracle, MySQL, and compute servers. The Vision and Perceptual Machines Lab (VPML) include two Apple Xserves. Lab space and offices provided in Davis Hall houses workstations and specialized equipment for researchers. The departmental network-attached storage device provides 13 TB of installed disk space for general and specific research projects. The university provides office space for PI Kennedy and his students. PI Kennedy has improved these spaces using funding from his start-up package.

### The ODIn Lab @ UB

PI Kennedy's **O**nline **D**ata **In**teractions Lab at the University at Buffalo (`http://odin.cse.buffalo.edu`) maintains additional resources specifically for internal use, including multiple x86 workstations, laptops and low-power development boards (Raspberry Pis and Intel Galileos) for general lab member use, a 16-node Hadoop cluster shared with 3 other labs, a 4-node Raspberry Pi embedded database benchmarking micro-cluster, as well as three servers: (1) A 32-core AMD Opteron being used as an application server, hosting the lab's project management system, teaching support applications, and trial deployments of lab-developed software; (2) A 16-core Intel Xeon database server testbed, and (3) A 12-core Intel Xeon, 128GB RAM testbed server. Lab workstations and laptops are configured with OS X, Ubuntu Linux, or Windows. Servers are configured with Redhat Enterprise Linux.

## Illinois Institute of Technology

### DBGroup Lab

PI Glavic is the head of the database group (DBGroup) at IIT (`http://www.cs.iit.edu/~dbgroup/`). The university provides office space for PI Glavic and his students. PI Glavic has improved these office spaces using funding from his start-up package.

### Equipment

PI Glavic bought two servers from the funds provided through an unrestricted gift from the Oracle Corporation. Both machines have 2 x 3.3Ghz AMD Opteron CPUs, 128GB RAM, and 4 x 1TB hard drives configured as RAID 5. These machines are administered by the research data center at IIT. PI Glacic's laboratory has 5 desktop machines for student use.

### Computer Science Department

The computer science department provides several machines that are used in courses taught by PI Glavic which are related to this proposal including a course on data integration and provenance (`http://www.cs.iit.edu/~cs520/`). These machines are used to run DBMS instances (e.g., Oracle, Postgres) that are used by students to practice database operations (e.g., SQL, programmatic access to a database, run data cleaning operations).

### IIT

IIT was listed on the National Register of Historic Places in 2005. The proposed research activities will not make any physical changes to IIT's campus and buildings.

## New York University

### Tandon School of Engineering

The Tandon School of Engineering (Tandon) at New York University is located in the Metrotech Center in Brooklyn, which is shared by a number of large modern buildings occupied by major corporations in the financial services industry and New York City agencies.

The School of Engineering has recently expanded its campus with 120,000 sq ft of newly leased space in 2 MetroTech Center and 15 MetroTech Center, which is part of both NYU's NYU 2031: NYU in NYC city-wide plan for academic space development and Tandon's capital plan, called the i2e (invention, innovation, and entrepreneurship) Campus Transformation, which is focused on attracting top-level faculty and high-achieving students and firmly establishing Tandon as a world-class center of applied science, technology and engineering.

The Computer Science and Engineering (CSE) department has 29 faculty members, and excellent computing facilities, including a number of teaching and research labs. CSE has just expanded into 2 MetroTech Center. The 10th floor space features faculty offices, computational labs, and workstations for post-docs and student researchers, among other facilities. The ninth floor has also been recently renovated to create large capacity classrooms and research space, including the Center for Advanced Technology in Telecommunications, and the Wireless Internet Center for Advanced Technology (WICAT) laboratory — the largest National Science Foundation-funded industry/academic cooperative research center.

VIDA: The Visualization and Data Analysis (VIDA) center is a relatively new addition to the School of Engineering. Silva is a faculty member in VIDA. VIDA is located on the 10th floor of 2 Metrotech Center and it houses students, staff, postdocs, and faculty in VIDA. The existing infrastructure in VIDA includes a number of dedicated servers, including machines that drive the birdvis.org, vistrails.org and crowdlabs.org domains. In order to enable collaborative large-scale software development, VIDA maintains machines that serve svn, git, trac, and wiki functionality.

The group also has extensive visualization and data analysis facilities, including a high-resolution display wall, and a stereo video wall. Each graduate student, post-doctoral associate, software developer, and research scientists in the group has access to an individual high-end workstation.

Tandon has been awarded an equipment award, CNS-1229185, "MRI: Acquisition of an Infrastructure for Prototyping Next-Generation Algorithms for Large-Scale Visualization, Data Processing and Analysis" which includes a computing cluster and large disk storage that will be instrumental for this proposal. That proposed equipment is being installed at NYU's state-of-the-art machine room in a secure location in downtown Manhattan. The School of Engineering has excellent networking and power infrastructure, and it is connected to the Internet by high-speed links. This infrastructure will be instrumental in this project, allowing us to experiment with the new functionality to execute analyses and event detection using GPUs and in parallel and cloud-based infrastructure.

The Visualization and Data Analysis (VIDA) center has extensive computing infrastructure for this project. The existing infrastructure in VIDA includes a number of dedicated servers, including an enterprise level SAN system with 288TB storage capacity, 19 Hadoop nodes, each with 64 cores (AMD Opteron 2.3GHz), 256 GB of RAM and 12TB of disk and Infiniband interconnects, for a total of another 228 TB of disk. We also have specialized GPU nodes (12 cores 256GB RAM, 8 TB of disk and 3 GPUs (Geforce Titan, 6GB of VRAM each)).

As one of the 14 schools of New York University, Tandon has access to other substantial computational resources, including NYU's High Performance Computing Service.

Program URL: `http://engineering.nyu.edu`

## NYU Center for Urban Science and Progress

New York University's Center for Urban Science and Progress is a multi-sector, interdisciplinary research collaboration of partners from academia, industry, and U.S. national labs and the City of New York with the mission to make cities more productive, livable, equitable, and resilient. Our strategy and vision are data-driven and leverage advances in areas of computing such as optimization, modeling, simulation, prediction and inference; large-scale data management, visualization, and analytics; advanced sensing techniques; social computing; infrastructure design, control and management; and intelligent systems and decision-making.

Data Warehouse. CUSP's Data Warehouse is being developed as a significant resource for research and will both be leveraged to support this project and benefit from its outcomes. Envisioned as a user facility, the Data Warehouse provides: infrastructure for cataloging and storing data; a production environment that supports analysis of large urban data sets (e.g., cluster and cloud computing facilities, relational and NoSQL databases); and protocols and infrastructure for data access that ensure compliance with privacy and security constraints for individual datasets. A research and management group, including a data curator and privacy expert, support the facility and conduct related research in techniques methods and tools. Included in this effort is collaboration with CUSP researchers to: assist researchers in the design of data strategy to enable the reproducibility of results and data preservation after the completion of projects; and develop technology and solutions for processing urban data sets, such as cleaning and integration of disparate datasets, visualization, and analysis, including mining and application of machine learning techniques.

CUSP leverages a diverse range of existing, emerging, and new data flows, including data generated by and/or collected by the City of New York; sensor data, both from the City and from networks developed by CUSP for research; and novel data streams collected by CUSP researchers.

The Data Warehouse currently houses 6TB of data, with an additional 400GB expected in the next few months. Data from the City of New York is acquired in part through a unique collaboration between CUSP and the City that underlies CUSP's mission. CUSP specifically collaborates with more than 15 City agencies and divisions, including: Department of Transportation, Department of Buildings, Department of Sanitation, Department of Citywide Administrative Services, Department of Design and Construction, Department of City Planning, Department of Health and Mental Hygiene, Department of Environmental Protection, Department of Information Technology and Telecommunications, Department of Parks and Recreation, City Police Department, City Fire Department.

Computing. CUSP's Brooklyn space houses a cluster with 20 HADOOP nodes, each with 24TB of disk, 256GB of RAM, and 64 AMD cores; 1 management node; 2 (high memory) server nodes, each with 1GB of RAM and 32 Intel cores; 2 (utility) serve nodes, each with 32GB of RAM and 8 Intel cores; 10GB Ethernet networking for the cluster; and an IBM General Parallel File System supported by a two server configuration and approximately 1PB raw storage. Cloud computing resources are provided by CUSP's industry partners.

Office. CUSP currently occupies a 27,000 square foot full floor of 1 MetroTech Center located in Downtown Brooklyn, adjacent to NYU Polytechnic School of Engineering, with classroom space next door in NYU's Media and Games Network (MAGNET) at 2 MetroTech Center. As of fall 2014, our faculty, students, and staff will also have access to an additional 30,000 square feet of space in a neighboring building before moving into a total of 150,000 square feet of space in 370 Jay Street, Brooklyn (projected move-in, 2017 based on completion of renovation).

Program URL: `http://cusp.nyu.edu`

# Data Management Plan

1. **Types of data and other materials produced during the course of the project.**
   There are several main types of materials that the investigators will produce during the course of this project:

   (a) A public binary and source-code release of the Vizier data curation tool, including user and developer documentation;

   (b) A library of example curation notebooks and datasets;

   (c) minutes and proceedings of the proposed workshop;

   (d) scientific papers describing the tools and datasets generated;

   (e) educational materials, including lecture notes, problem sets, and demonstrations.

2. **Standards to be used for data format and content.**
   The standard format the dissemination of research results for the computer science community is to provide access to the research papers in a Portable Document Format (i.e., .pdf) in several online repositories such as the IEEE and Association for Computing Machinery (`http://ieeexplore.ieee.org/Xplore` and `http://dl.acm.org/`) digital libraries. It is also standard practice to post preprint copies on a personal webpage (depending on the copyright restrictions of peer-reviewed journals) and through open-access repositories such as `www.arXiv.org`. Source code is also stored on personal webpages or at public repositories such as `www.github.com`. The proposed budget includes an allocation for a demonstration deployment of the proposed system on a cloud-hosting platform such as Amazon EC2 (`http://aws.amazon.com`).

   Where practical, example data will be stored in standard formats such as CSV and JSON. Example curation notebooks will be made available in a self-documenting format like JSON, YAML or Markdown. Documentation will be stored in plaintext, HTML, Markdown, and/or PDF.

   Educational materials are stored in a variety of formats from PDF files of lecture notes and problem sets, to PowerPoint, HTML, or PDF slides will be posted on the personal webpages of the course instructors. Interactive course materials built on top of Vizier will be made publicly accessible through the demonstration deployment.

   System implementations, log parsers, and other programs produced as part of this project will be in standard programming languages such as Scala, Java, JavaScript, Ruby, Python, and/or C.

3. **Methods and policies for providing access and enabling sharing.**
   All public data will be stored either in publicly accessible databases or websites hosted by one or more participating departments, or made accessible via the demonstration deployment of Vizier.

4. **Provisions for re-use, re-distribution, and the production of derivatives.**
   We will use Creative Commons licenses `http://creativecommons.org/licenses/` for the re-use, re-distribution, and the production of derivatives of the project data, while respecting the limitations on copyrighted material from published journals. We will use Apache

licenses `http://www.apache.org/licenses/` or similarly permissive licenses for the re-use, re-distribution, and the production of derivatives of all project source code, while respecting the licenses of any library dependencies.

5. **Methods for archiving and preserving access to data and materials.**
   All data and materials will be stored on back up systems indefinitely.

   The investigators will mainly rely upon the facilities of their respective participating departments to maintain web access to the course materials, sources codes, and papers.

## Mentoring Plan for Postdoctoral Researcher

NYU is committed to advancing and developing the careers of young researchers. In addition to engaging postdoctoral fellow(s) in active research, we will employ the following mentoring plan.

**Orientation** will include in-depth conversations between the supervisor (Freire) and the Postdoctoral Researcher (PR). Mutual expectations will be discussed and agreed upon in advance. Orientation topics will include (a) the amount of independence a PR requires, (b) interaction with coworkers, (c) productivity including the importance of scientific publications, and (d) documentation of research methodologies and experimental details so that the work can be continued by other researchers in the future.

**Career Counseling** will be directed at providing the Postdoctoral Researcher with the skills, knowledge, and experience needed to excel in his/her chosen career path. In addition to guidance provided by the respective supervisors, the PR will be encouraged to discuss career options with researchers and supervisors at the NYU Tandon and at CUSP.

**Experience with Preparation of Grant Proposals** will be gained by direct involvement of the PR in proposals prepared by the supervisors. The PR will have opportunities to learn best practices in proposal preparation including identification of key research questions, definition of objectives, description of approach and rationale, and construction of a work plan, timeline, and budget.

**Publications and Presentations** are expected to result from the work supported by the grant. These will be prepared under the direction of the respective supervisors and in collaboration with other researchers participating in this project at the other institutions, as appropriate. The PR will receive guidance and training in the preparation of manuscripts for scientific journals and presentations at conferences.

**Teaching and Mentoring Skills** will be developed in the context of regular project meetings, during which graduate students and postdoctoral researchers will describe their work to colleagues within the group and assist each other with solutions to challenging research problems, often resulting in cross fertilization of ideas. The PR will be encouraged to develop independent projects with graduate students.

**Instruction in Professional Practices** will be provided on a regular basis in the context of the research work and will include fundamentals of the scientific method and other standards of professional practice. In addition, the PR will be encouraged to affiliate with one or more professional societies in his/her chosen field.

**Technology Transfer activities** will include training on the patent process, and will include regular patent search activities. The collaborative team will be pro-active in the process of identifying contributions of commercial value, and will teach the PR how to protect the generated intellectual property. The PR will be given an opportunity to become familiar with the university-industry relationship including applicable confidentiality requirements and preparation of invention disclosure applications.

**University Programs** NYU's Office of Postdoctoral Affairs (OPA) aims to provide a community for NYU postdocs. The OPA also offers Workshops for Emerging Scientists, addressing topics such as research misconduct, conflict of interest, mentor/trainee responsibilities, collaboration in science, publication practices and responsible authorship, ethical considerations in research with human and animal subjects, ownership, acquisition, storage and sharing of research data, and survival skills for a career in research. These workshops are augmented with ongoing programs provided through the Science Alliance Program of the New York Academy of Sciences, a consortium of universities and research institutions committed to advancing the next generation of scientists. Along with a wide range of career mentoring activities, the Science Alliance also offers conferences, workshops, discussion groups, eBriefings, and encounters with industry representatives and other potential employers.

**Success of the Mentoring Plan** will be assessed by monitoring the personal progress of the PR through tracking of the PR's progress toward his/her career goals after finishing the postdoctoral program.

CENTER FOR URBAN
SCIENCE+PROGRESS

1 MetroTech Center, 19FL
Brooklyn, New York 11201
tel: 646.997.0500
fax: 646.997.0560
web: cusp.nyu.edu

April 1, 2016

Dear Juliana,

I strongly support your NSF proposal entitled "CIF21 DIBBs: PD: Collaborative: Streamlining and Understanding Curation with Vizier" and I look forward to collaborating with you in this endeavor.

As you know, we manage the Data Facility at the NYU Center for Urban Science and Progress (CUSP). CUSP leverages a diverse range of existing, emerging, and new data flows, including data generated by and/or collected by the City of New York; sensor data, both from the City and from networks developed by CUSP for research; and novel data streams collected by CUSP researchers. The Data Facility has been established to support the empirical study of cities in conjunction with New York based researchers, agencies, and citizens. The CUSP Data Catalog is online at datahub.cusp.nyu.edu/catalog and currently hosts over 65TB of data.

Data from the City of New York is acquired in part through a unique collaboration between CUSP and the City that underlies CUSP's mission. CUSP specifically collaborates with more than 15 City agencies and divisions, including: Department of Transportation, Department of Buildings, Department of Sanitation, Department of Citywide Administrative Services, Department of Design and Construction, Department of City Planning, Department of Health and Mental Hygiene, Department of Environmental Protection, Department of Information Technology and Telecommunications, Department of Parks and Recreation, City Police Department, City Fire Department. Under this agreement, CUSP works with City agencies to: (1) identify significant real world problems affecting the delivery of municipal services and critical challenges to the urban environment and economy, and (2) develop and implement research approaches to these problems and challenges. The goal is to understand and improve urban systems, the urban quality of life and city planning.

In addition to data from the City of New York, CUSP is working with the City and other community and industry partners to develop and deploy sensor networks and other novel data generation practices for research to support our mission to make cities more productive, livable, equitable, and resilient.

Data quality is crucial for the CUSP Data Facility (CDF) and it is also a major challenge. As one of the original designers of the CDF, you are aware that given the large number of heterogeneous data sets we store, and the broad range of projects that use subsets of these data, it is not possible for us to adopt the traditional warehouse approach -- extract data, transform, and load (ETL). As you describe in your proposal, even for a given data set, different projects (and research questions) require different curation pipelines, thus there is no one-size-fits-all solution we can adopt. This problem is compounded due to the fact that many of the CDF users have little or no programming expertise.

We have a clear need for a tool such as the one you propose. Besides streamlining the curation process and enabling domain experts to curate the data they need as they explore the data, we see a great opportunity for such a tool to enable collaboration and re-use. A shared repository of curation pipelines will enable our users to benefit from the collective wisdom: they will be able to search for pipelines, re-use them, and learn by example how to build new pipelines. Another key feature that your tool provides is detailed provenance capture for both the data and the curation process. The various projects using the CDF produce a large volume of derived data that are also stored in the facility. To re-use and properly analyze these derived data, users need their provenance to understand the curation process applied to the data, and to assess whether that is compatible with their analysis needs. Besides, since much of the data we keep come from feeds that change over time, we need to also carefully track which curation pipelines were applied to which feeds.

The Urban Data Profiler system you designed and developed has already been deployed at the CDF. The system has been successfully used to derive metadata for many of our data feeds, as well as to identify data quality issues. We see the work you propose on Vizier a natural next step. We look forward to working with you on this project and we hope to deploy the system in the CUSP Data Facility.

Julia Lane
Professor of Practice, NYU Center for Urban Science and Progress
Professor, NYU Wagner School of Public Policy

CUSP
CENTER FOR URBAN
SCIENCE+PROGRESS

1 MetroTech Center, 19FL
Brooklyn, New York 11201
tel: 646.997.0500
fax: 646.997.0560
web: cusp.nyu.edu

Rebecca Rosen, PhD
Associate Director, Data Resources & Data Strategy
Center for Urban Science + Progress, NYU

NYC

Taxi & Limousine
Commission

**Meera Joshi**
Commissioner

**Rodney Stiles**
Executive Director of
Policy and Analytics
rodney.stiles@tlc.nyc.gov

**33 Beaver Street**
**22nd Floor**
**New York, NY 10004**

**+1 212 676 1183** tel
**+1 212 676 1101** fax

National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

March 25, 2016

**Re: DIBBs Solicitation, NSF 16-530**

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's transportation for hire industries, including yellow and green taxicabs, liveries and black cars. With around 90,000 licensed vehicles, we are continually gathering and analyzing data to improve transit service for all New Yorkers. As part of our commitment to become a data-informed city agency, we undertook an initiative in 2015 to publish trip records collected from licensees or about licensee operations online. As a result, academics, researchers, and the public are now able to study our data and gain insight into our city's movements. We welcome these activities, as they provide another means of determining whether our regulated industries are as efficient and accessible as possible and provide service where and to whom it is needed. We understand, and plan to continue to use, the transformational potential of data-driven policymaking, and plan on publicizing more data as time goes on.

New York University has approached TLC with an opportunity to collaborate on the Vizier system NYU is currently building, with the grant opportunity offered through NSF 16-530. We see great potential in the Vizier system as a form of data exploration. This framework streamlines the data curation process by distinguishing between outlying data features worthy of analysis versus data errors that are missing or require cleaning. This system could greatly improve the TLC's ability to analyze data that we provide publically, and as such aligns with our mission to provide more and better data.

Our data is extremely important to the work that we do, and we support this proposal for collaboration. We look forward to collaborating with NYU in several ways: 1) offering a variety of data for all yellow and green taxis and for-hire vehicles; 2) providing feedback on the techniques developed to automate the data cleaning process; and 3) testing the system that will be built.

Please feel free to contact us with any questions regarding our planned support of this project.

Sincerely,

Rodney Stiles

Rodney Stiles
Executive Director of Policy and Analytics

3/31/2016

**Dieter Gawlick**
Architect
Oracle Corporation
Oracle ST
500 Oracle Parkway
Redwood City, CA 94065

Dear NSF Panel:

In my role as Architect at Oracle, I wish to express my wish to collaborate with the PIs on the proposal entitled *CIF21 DIBBs: PD: Streamlining and Understanding Curation with Vizier*. Maintaining high quality data is an important issue for Oracle, as many of Oracle's clients must regularly manage unstructured, incomplete, or otherwise uncertain data. Current data curation products do not expose uncertainty information to the user and do not support the exploratory, incremental construction of a data curation workflow. Once curated data is digested into a database, it is typically hard to determine whether this data is trustworthy, because of lack of provenance information about the curation steps that were applied to the data. Our clients need a platform that tracks provenance, exposes uncertainty, and provides full accountability given the frequent changes applied to a curation workflow under development. Ideally, such a platform should guide a user through the steps of creating a curation pipeline, providing recommendations along the way on what curation steps to apply and how to tune them.

The proposed research is novel, because it is the first solution that 1) supports incremental development of data curation workflows with full explainability (using provenance to explain the existence of every piece of data, where is coming from, how it got transformed and why, and any uncertainty associated with it); 2) leverages past curation efforts through recommendations based on provenance; and 3) naturally supports the exploratory nature of curation through versioning of workflows and unlimited undo of operations. Vizier is transformative, because by involving several communities with dire curation needs, the project will have significant impact on real world curation practice. Oracle is interested in integrating the library of data curation steps developed as part of Vizier and its provenance tracking and recommendation facilities into our own data curation products.

I have been collaborating with the PIs Kennedy and Glavic on the Mimir and GProM projects respectively. Note that these projects are two out of the three major building blocks of the Vizier system that will be build in this proposal. Oracle has been supporting both PIs' research efforts for the past years through unrestricted gifts totaling roughly $500,000. I have been in contact with PI Freire for several years and am well aware of her seminal work on VisTrails, the third major component of the proposed system.

If the proposal is selected for funding by the NSF, it is my intent to collaborate and commit resources as detailed in the Project Description and described in the following. Any resource allocation has of course to go through Oracle's approval process as was the case for the current collaborations. I expect to spend time to advice the PIs by participating in the regular meetings that are part of the projects collaboration plan. My decade-long experience in working on the technical and research aspects of implementing database systems makes me an ideal candidate to advise on architectural, conceptual, and implementation issues. Furthermore, I can provide advice on experimental evaluations of Vizier that use our products. Last but not least, as a veteran in the database industry I have a thorough understanding of our customer's needs and am very well connected. I plan to leverage my connections and understanding to broaden the user base of the project.

If you have any questions, please contact me at dieter.gawlick@oracle.com or (650) 506 8706.

Yours sincerely,

D. Gawlick

Dieter Gawlick, Architect at Oracle

NYU | WAGNER

**INGRID ELLEN**
*Paulette Goddard Professor of Urban Policy and Planning*

**Robert F. Wagner Graduate School of Public Service**
New York University
295 Lafayette Street, 2nd Floor
New York, NY 10012

**P:** 212 998 7533
**F:** 212 995 4162

ingrid.ellen@nyu.edu

April 1, 2016

Dear Juliana,

I am writing to support your NSF DIBBs proposal entitled " Streamlining and Understanding Curation with Vizier" and to confirm my interest in collaborating with you on this project.

I am a Professor at the NYU Wagner School and the Faculty Director of the NYU Furman Center, a research center devoted to the study of housing, land use, and neighborhoods. We are committed to producing research that influences urban policy. Some of the Furman Center's early work provided some of the first hard evidence of the benefits that subsidized housing investments can provide to communities. More recently, our analysis of the data on the subprime debacle and the foreclosure crisis has helped quantify the costs of the crisis to children, neighbors, and communities. Our assessment of crime data and property-level foreclosure data led to the finding that properties on the way to foreclosure (rather than those that have already been foreclosed) invite crime.

As you know, a common theme in our projects is the need to analyze multiple data sets. Over the years, at the Furman Center we have accumulated a broad array of data on demographics, neighborhood conditions, infrastructure, housing stock and other aspects of New York City's neighborhoods and real estate market. Due to the nature of our work, and the potential impact it has on policy making, data quality is crucial for us. Moreover, we make much of these data available online as part of the Furman Center Data Services (http://furmancenter.org/data), and every year, we publish a report on the "State of New York City's Housing & Neighborhoods", a compendium of data and analysis about New York City's housing, land use, demographics, and quality of life indicators for each borough and the city's 59 community districts.

For each analysis we perform, a lot of effort is devoted to curating the data sets we use. As you articulate in your proposal, often, it is during these analyses that we discover serious data quality issues. Currently, we perform data cleaning in an ad-hoc fashion for each project we carry out. Besides being a time-consuming process, we face several challenges regarding (1) re-use---as we obtain new data sets or data feeds, it is hard to assemble new cleaning pipelines; (2) lack of systematic provenance capture---as we iteratively clean our data, it is hard to precisely document the cleaning steps applied to the data and to keep track of the different versions that are derived. As a result, curation is a major bottleneck in our research.

The tool you propose to develop has great potential for practical impact in the work we do. By combining data curation and exploration, and capturing detailed provenance, it addresses many of the challenges we have to deal with on a daily basis. I welcome the opportunity to contribute to the design of the tool and I hope to deploy it at the Furman Center.

I look forward to continuing to collaborate with you and your group.

Sincerely,

Ingrid G Ellen

Ingrid Gould Ellen

4/2/2016

**Ronny Fehling**
Vice President, Head of
Data Driven Technologies
and Advanced Analytics
*Airbus Group*

Dear NSF Panel:

In my role as Vice President, Head of Data Driven Technologies and Advanced Analytics at Airbus, I wish to express my wish to collaborate with the PIs on the proposal entitled *CIF21 DIBBs: EI: Vizier, Streamlined Data Curation*. Airbus has an ongoing need to analyze large, noisy heterogeneous data. Sensors deployed on aircraft, satellites, or in test environments, simulation results, GPS data, manufacturing, repair logs, and other records are regularly analyzed to develop platforms that are safer, cheaper, cleaner, and more efficient. I am presently coordinating several data integration and data curation efforts at Airbus, including an effort to build a company-wide data lake. There are several key challenges in this effort: (1) Enabling efficient discovery of datasets relevant to an analyst's goals, (2) Integrating datasets from different sectors of Airbus that may have different schemas, data domains, or underlying assumptions, (3) Ensuring that proprietary and sensitive data remains secure, even through data derivatives, (4) Harvesting data and lineage information from opaque, static APIs, (5) Operating over a heterogeneous environment of Hadoop, client APIs, many databases, external tables, and dynamic data sources, (6) Helping analysts quickly visualize and contextualize datasets.

I have been collaborating with PI Kennedy on the Mimir project, one of the three major building blocks of the Vizier system, and believe that Vizier is well positioned to help us to address many of these challenges. If the proposal is selected for funding by the NSF, it is my intent to collaborate and commit resources as detailed in the Project Description and described in the following: (1) I expect to spend time to advise the PIs by participating in the regular meetings that are part of the project's collaboration plan, (2) I will provide the PIs with examples of datasets that exemplify data quality challenges faced by Airbus, (3) I will facilitate discussions between the PIs and prospective users at Airbus for the purpose of providing feedback about Vizier prototypes and the project's goals, (4) Once the project reaches a mature state, I will help to facilitate a preliminary deployment.

If you have any questions, please contact me at `ronny.fehling@airbus.com`.

Yours sincerely,

Ronny Fehling

4/4/2016

**Zhen Hua Liu**
Architect
Oracle Corporation
Oracle ST
500 Oracle Parkway
Redwood City, CA 94065

Dear NSF Panel:

In my role as Architect at Oracle, I wish to express my wish to collaborate with the PIs on the proposal entitled *CIF21 DIBBs: PD: Streamlining and Understanding Curation with Vizier*. Maintaining high quality data is an important issue for Oracle, as many of Oracle's clients must regularly manage unstructured, incomplete, or otherwise uncertain data. Current data curation products do not expose uncertainty information to the user and do not support the exploratory, incremental construction of a data curation workflow. Once curated data is digested into a database, it is typically hard to determine whether this data is trustworthy, because of lack of provenance information about the curation steps that were applied to the data. Our clients need a platform that tracks provenance, exposes uncertainty, and provides full accountability given the frequent changes applied to a curation workflow under development. Ideally, such a platform should guide a user through the steps of creating a curation pipeline, providing recommendations along the way on what curation steps to apply and how to tune them.

The proposed research is novel, because it is the first solution that 1) supports incremental development of data curation workflows with full explainability (using provenance to explain the existence of every piece of data, where is coming from, how it got transformed and why, and any uncertainty associated with it); 2) leverages past curation efforts through recommendations based on provenance; and 3) naturally supports the exploratory nature of curation through versioning of workflows and unlimited undo of operations. Vizier is transformative, because by involving several communities with dire curation needs, the project will have significant impact on real world curation practice. Oracle is interested in integrating the library of data curation steps developed as part of Vizier and its provenance tracking and recommendation facilities into our own data curation products.

I have been collaborating with the PIs Kennedy and Glavic on the Mimir and GProM projects respectively. Note that these projects are two out of the three major building blocks of the Vizier system that will be build in this proposal. Oracle has been supporting both PIs' research efforts for the past years through unrestricted gifts totaling roughly $500,000. I have been in contact with PI Freire for several years and am well aware of her seminal work on VisTrails, the third major component of the proposed system.

If the proposal is selected for funding by the NSF, it is my intent to collaborate and commit resources as detailed in the Project Description and described in the following. Any resource allocation has of course to go through Oracle's approval process as was the case for the current collaborations. I expect to spend time to advice the PIs by participating in the regular meetings that are part of the projects collaboration plan. My decade-long experience in working on the technical and research aspects of implementing database systems makes me an ideal candidate to advise on architectural, conceptual, and implementation issues. Furthermore, I can provide advice on experimental evaluations of Vizier that use our products. Last but not least, as a veteran in the database industry I have a thorough understanding of our customer's needs and am very well connected. I plan to leverage my connections and understanding to broaden the user base of the project.

If you have any questions, please contact me at `zhen.liu@oracle.com`.

Yours sincerely,

Zhen Hua-Liu, Architect at Oracle

4/4/2016