Provenance is the chronological history of creation, ownership, chain of custody, and location of an object. In its original and still most frequently used sense, provenance is used to authenticate and trace the legitimate ownership of a work of art; it confers, creates, or adds value to the work itself. But provenance can be constructed, identified, or traced for any object, including data. Data provenance confers or adds value—as trustworthiness—to data, but data provenance also can be used to reproduce computational analyses and validate scientific conclusions. In short, whereas the existence of provenance establishes value of artwork, the use of provenance establishes value of data.

For more than a decade, our group (Box 1; Figure 1) has guided the development of a set of tools (Figure 2) that uses data provenance to enhance trustworthiness and reproducibility of data, the associated analytical processes (software) that created them, and the publications and conclusions derived from them.

But the provenance of our End-to-End Provenance project spans a much longer period of time. The roots of our End-to-End Provenance project extend back more than two decades in time to the LASER (laboratory for advanced software engineering research) group at the University of Massachusetts at Amherst, led by Professors Leon Osterweil and Lori Clarke. Two members of our current team, Emery Boose and Aaron Ellison, worked with LASER on a project aimed at establishing a process-definition formalism that could be used to describe scientific workflows. In those early days, we were interested in collecting provenance to be able to evaluate the correctness of the workflows that were carried out; the LASER group, including Barbara Lerner (then a research assistant professor at the University of Massachusetts), developed Little-JIL, a graphical language with rigorously defined operational semantics in which one could program

---

Box 1. The Current Main Characters of the End-to-End Provenance Project

**The Visionary:** Margo Seltzer is Canada 150 Research Chair in computer systems and the Cheriton Family chair in computer science at the University of British Columbia. She studies systems sensu lato—systems for capturing and using data provenance, file systems, databases, transaction processing systems, storage and analysis of graph-structured data, new architectures for parallelizing execution, and systems for discrete optimization.

**The Developers and Maintainers:** Emery Boose and Barbara Lerner have been our system designers and developers from the get-go. Emery is information manager and a senior scientist at the Harvard Forest. His research interests include data provenance, ecoinformatics, hurricane modeling, meteorology, and hydrology. Barbara is a professor of computer science at Mount Holyoke College. She develops software that data analysts can use to help understand their scripts and is passionate about increasing participation of women in computing.

**The Translator:** Aaron Ellison is the senior research fellow in ecology at Harvard University and a senior ecologist and the deputy director of the Harvard Forest. His overlapping interests in ecological processes, publishing and open science, and cultural and technical challenges for collecting provenance and archiving data have positioned him as the person who brings reality into software engineering and translates software engineering concepts back to domain scientists.

Undergraduates who have worked on the project are listed in Table 1.
coordination among processes, document their execution sequence, and re-execute them.

Over time we shifted our focus from Little-JIL to R, a language widely used by scientists for data analysis and statistics. In a fortuitous coincidence, Barbara Lerner and Margo Seltzer served together on a grant-review panel for the National Science Foundation’s Computer and Information Science and Engineering directorate. In discussing research over a break, they learned of their shared interest in provenance, and particularly in developing provenance support for languages that scientists actively used. They shared a vision to bring provenance tools to domain scientists instead of trying to convince them to change how they worked. Bringing in Margo (then a professor of computer science at Harvard) and her group broadened our overall focus beyond evaluating workflows to include system-level processes, provenance storage, and end-to-end solutions.

The tools that we have developed, informed by this broadened perspective, transparently capture and use data provenance from workflows and analytical pipelines in multiple languages (R, Python) and more generically (Camflow) (Figure 2). These tools, and provenance-based tools developed by other groups (see summary in Lerner et al.), improve transparency, trustworthiness, and reproducibility of data analysis and associated results and facilitate debugging and improve understanding of why different runs of a seemingly identical script can yield different outcomes. More recent applications of provenance include its use in system security, including visualization and explanation of software faults, intrusion detection, and compliance with regulations involving protection of personal data. A thorough review of these topics will be
part of a future paper in Patterns, and all packages discussed are available from https://end-to-end-provenance.github.io/.

The End-to-End group is, of course, much larger than the PIs and senior scientists. In fact, much of the behind-the-scenes work has been done by teams of undergraduates participating in the Harvard Forest Summer Undergraduate Research Program and a handful of graduate students (Xueyuan Han, Jingmei Hu, Jackson Okuhn, Narun Raman) and postdocs (Matthew Lau [now with the Chinese Academy of Sciences] and Thomas Pasquier [now at the University of Bristol]). Although most of the undergraduates have gone on to careers in data science and software engineering in the private sector, Morgan Vigil is now an assistant professor in computer science at Northern Arizona University, and Joe Wonsil is now a PhD student with Margo Seltzer at the University of British Columbia (Table 1). As all the participants of the End-to-End Provenance project grow their careers, spread

Figure 2. Our End-to-End Provenance Tools

Tools include packages for the R software system that take advantage of a lightweight provenance collection tool (rdtLite) that collects provenance during a console session or as an R script executes. Taking advantage of the prov.json-encoded provenance and internal parsing and graphing functions (provParseR, provGraphR), provSummarizeR provides a high-level summary of the computing environment, loaded libraries, sourced scripts, and I/O; provExplainR helps users identify differences in results derived from multiple executions of a script; and provDebugR supports “time-traveling debugging” of a script without the need to set breakpoints or insert print statements and rerun the script. containR is a provenance-based virtual machine for reproduction and re-execution of R scripts. ProvBuild provides provenance-based debugging tools and builds off the noWorkflow project. CamFlow (Cambridge information flow architecture) is a Linux security module designed to capture data provenance for the purpose of system auditing. It is being leveraged by Unicorn, an anomaly-based detector of advanced persistent threats (APTs) that are otherwise difficult to detect because of their “low-and-slow” attack patterns and frequent use of zero-day exploits.*

Table 1. Undergraduate Students Who Have Worked on the End-to-End Provenance Project in the Last Decade and Their Postgraduate Trajectories

| REU year | Student | Institution | Graduated | Present position |
|----------|---------|-------------|-----------|-----------------|
| 2009     | Cory Teshera-Sterne | Mt. Holyoke  | 2010  | Data Coordinator, NeighborWorks Home Partners |
| 2010     | Morgan Vigil | Westmont  | 2011  | Asst. Prof. of CS, Northern Arizona U. |
| 2011     | Andy Kaldunski | Ripon  | 2012  | Deceased |
| 2012     | Miruna Oprescu | Harvard  | 2015  | Sr Data & Applied Scientist, Microsoft Research |
| 2013     | Yujia Zhou | Dickinson  | 2013  | PhD Analyst, L.E.K. Consulting |
| 2013     | Shay Adams | Mt. Holyoke  | 2014  | Application Developer, U Wisconsin |
| 2014     | Luis Perez | Harvard  | 2016  | Research Engineer, DeepMind |
| 2015     | Marios Dardas | Holy Cross  | 2016  | Data Analyst, McKinsey & Co. |
| 2015     | Alex Liu | Amherst  | 2019  | Flow Volatility Trading Analyst, Barclays Investment Bank |
| 2017     | Connor Gregoric-Trevor | Grinnell  | –  | – |
| 2018     | Dia Poulos | Mt. Holyoke  | –  | – |
| 2016     | Moe Pwint Phyu | Mt. Holyoke  | 2018  | Software Development Engineer, Workday |
| 2017     | Jen Johnson | Middlebury  | –  | – |
| 2018     | Orenna Brand | Columbia  | Enrolled  | – |
| 2019     | Joe Wonsil | Carthage  | 2019  | PhD student, U. British Columbia |
| 2019     | Khanh Ngo | Mt. Holyoke  | Enrolled  | – |
| 2019     | Erick Odunuiyi | Kansas  | Enrolled  | – |

*4-year liberal arts college
throughout the world, and continue to
develop more useful tools, awareness
of their provenance will ensure that
the people behind the tools are
valued, too.

ACKNOWLEDGMENTS

Our work on end-to-end provenance has been
supported by grants from the US National Science
Foundation (DEB-1237491, DBI-1459519, and
SSI-1450277), a Charles Bullard Fellowship to
B.S.L. at Harvard University, and a faculty fellow-
ship to B.S.L. from Mount Holyoke College. This
paper is a contribution from the Harvard Forest
Long-Term Ecological Research (LTER) program,
supported since 1990 by the US National Science
Foundation.

REFERENCES

1. Becker, R.A., and Chambers, J.M. (1986).
Auditing of Data Analyses. SIAM J. Sci. Statist.
Comput. 9, 78–86.

2. Boose, E.R., and Lerner, B.A. (2017).
Replication of data analyses: provenance in R.
In Stepping in the Same River Twice:
Replication in Biological Research, A. Shavit
and A.M. Ellison, eds. (Yale University Press),
pp. 195–212.

3. Lerner, B., Boose, E., and Perez, L. (2018).
Using introspection to collect provenance in R.
Informatics 5, 12.

4. Pasquier, T., Lau, M.K., Trisovic, A., Boone,
E.R., Couturier, B., Crosas, M., Ellison, A.M.,
Gibson, V., Jones, C.R., and Seltzer, M.
(2017). If these data could talk. Sci. Data 4,
170114.

5. Cass, A.G., Lerner, B.S., McCall, E.K.,
Osterweil, L.J., Sutton, S.M.J., and Wise, A.
(2000). Little-JIL/Juliette: a process definition
language and interpreter. In Proceedings of
the 2000 International Conference on Software
Engineering (IEEE), pp. 754–757.

6. R Development Core Team (2020). R: A lan-
guage and environment for statistical
computing (R Foundation for Statistical
Computing).

7. Pasquier, T., Eyers, D., and Seltzer, M. (2019).
From Here to Provtopia,. In Heterogeneous
Data Management, Polystores, and Analytics
for Healthcare, V. Gadepally, et al., eds.,
pp. 54–67.

8. Han, X., Pasquier, T., Bates, A., Mickens, J., and
Seltzer, M. (2020). UNICORN: Runtime
Provenance-Based Detector for Advanced
Persistent Threats. arXiv, 2001.01525 https://
arxiv.org/abs/2001.01525.