REPRODUCIBILITY

A "datathon" model to support cross-disciplinary collaboration

Jerôme Aboab,[1]* Leo Anthony Celi,[1] Peter Charlton,[1] Mengling Feng,[1] Mohammad Ghassemi,[1] Dominic C. Marshall,[1†] Louis Mayaud,[1] Tristan Naumann,[1] Ned McCague,[1] Kenneth E. Paik,[1] Tom J. Pollard,[1] Matthieu Resche-Rigon,[1] Justin D. Salciccioli,[1] David J. Stone[2,3]

In recent years, there has been a growing focus on the unreliability of published biomedical and clinical research. To introduce effective new scientific contributors to the culture of health care, we propose a "datathon" or "hackathon" model in which participants with disparate, but potentially synergistic and complementary, knowledge and skills effectively combine to address questions faced by clinicians. The continuous peer review intrinsically provided by follow-up datathons, which take up prior uncompleted projects, might produce more reliable research, either by providing a different perspective on the study design and methodology or by replication of prior analyses.

The research problems of health care extend beyond clinical medicine and cannot be solved by physicians working in isolation. Clinicians are well aware of the uncertainties and information gaps that permeate the practice of medicine, which range from the trivial but annoying to the significant and seemingly intractable. However, busy clinicians typically do not have the time, energy, or training to address questions they encounter in day-to-day practice. Instead, scientists, physician-scientists, and engineers carry out biomedical research designed to decipher clinical problems, but these efforts often are isolated from critical clinical input. In addition, recent accusations against this research community include the production of an overwhelming number of published studies that are irreproducible, underpowered, poorly designed, and otherwise lack statistical rigor (*1*–*5*).

The typical clinician solves problems by applying basic medical-science concepts and diagnostic and treatment protocols mastered during medical training. In academic facilities, there are physicians who see patients but spend most of their time working on basic, translational, or clinical research. In fact, some physician-scientists—even those on medical school faculties—rarely have exposure to clinical medicine. The nonresearcher clinicians seem to be getting busier and busier with patient care, but their input is critically important to identify relevant medical questions and problems and to tackle them with a deep understanding of the clinical context. This must be accomplished in an era in which progressively limited resources are struggling to cope with increasingly expensive health care.

Over the past decade or so, data scientists and engineers have become increasingly drawn to and involved with health care (*5*). This interest has recently been accelerated by the now near-universal digitization of health care, which provides data scientists with a toehold and a progressively important role in daily clinical care as well as biomedical research. How can this group of diverse talents, interests, and schedules be combined for productive collaboration? We propose an adaptation of the computer industry's hackathon model—an event in which diverse professionals (entrepreneurs, software developers, designers, engineers) work together on software development over a short period of time. In the biomedical arena, a hackathon similarly would bring together participants with disparate but complementary knowledge and skills—for example, academic biomedical researchers, industry scientists, bioengineers, physicians, statisticians, and computational biologists—to address the myriad questions and unmet medical needs faced by clinicians.

It is difficult to establish a platform for the real-time, respectful, and effective exchange of ideas among specialists who are usually separated by time, space, methods, attitudes, and terminology (language). The hackathon provides just such a platform for initial conception and design of a study as well as subsequent analysis and publication (in the literal sense of making the results "public"). The hackathon also provides a real-time infrastructure for the kind of simultaneous translation of specialty terminologies and jargon that is necessary for effective communication and cooperation. Last, follow-up hackathons—events that pick up and advance prior uncompleted projects—offer a form of continuous peer review that might produce more reliable research either through replication of published results or by virtue of their differing perspectives on study conception and design.

HACKATHON FOR DATA ANALYSIS

Originally the brainchild of Silicon Valley, CA, hackathons have proven to be successful models for innovation in business settings and are typically organized as intense, short-duration, competitions in which teams generate innovative solutions (*5*). The hackathon model integrates collaboration, idea generation, and group learning by joining various stakeholders in a mutually supportive setting for a limited period of time.

For health data analysis, the goal of the hackathon is to assemble clinical experts, data scientists, statisticians, and those with domain-specific knowledge to create ideas and produce clinically relevant research that reduces or eliminates biases, relies on sound statistical rigor and adequate data samples, and aims to produce replicable results. For that purpose, we have coined the term "datathon" as a portmanteau of data + hackathon, accentuating the application of the hackathon model to data analytics. For example, a critical-care datathon is an event in which participants are brought together to form interdisciplinary teams and answer research questions in the field of critical (that is, intensive) care. In addition, using the term datathon avoids the potentially negative connotation of the term hackathon in the minds of some participants and observers.

Although this approach might be more difficult for some study designs, the analysis of health data is especially suited to benefit from this model, because the data have already been collected and are readily available to researchers in a machine-readable format. Whereas the process flow in translational medicine is conventionally from the research bench to the bedside, the flow in the datathon model is from the point of care to research, or more specifically, to the database and the analytical team. Approaching health care in this manner

[1]Massachusetts Institute of Technology (MIT) Critical Data, MIT, Cambridge, MA 02139, USA. [2]Departments of Anesthesiology and Neurosurgery, University of Virginia School of Medicine, Charlottesville, VA 22908, USA. [3]Center for Wireless Health, University of Virginia School of Engineering and Applied Science, Charlottesville, VA 22904, USA.
*All authors contributed equally to this manuscript.
†Corresponding author. E-mail: dominic.marshall12@imperial.ac.uk

is a form of reverse engineering, because it examines the data record of the system inputs and elements that produce the observed outcomes. As databases become larger and more detailed, they will come to represent more and more accurate depictions of a kind of virtual clinical reality providing even richer resources for research.

Health care data admittedly are subject to intrinsic and extrinsic problems of accuracy. Therefore, analytical studies based on the secondary use of health care data inherently suffer this particular fragility. The data can be wrong for a large variety of reasons, including human and machine misentry, missing data, and faulty assessments, among others. However, all scientific research suffers from some degree of data unreliability—those who have worked in wet labs understand that not all reported results reflect repeatable experimental perfection. One further advantage of a team approach is that faulty data potentially can be recognized from a variety of viewpoint axes by the various expert types required to do so. For example, there may be technical problems whose recognition requires a data scientist or software engineer, clinical issues requiring a clinician, and other issues best recognized by domain experts.

AN EXEMPLAR DATABASE

The MIMIC (Medical Information Mart for Intensive Care) database contains patient-level data for intensive care unit (ICU) admissions at Beth Israel Deaconess Medical Center from 2001 to 2012 (*6*). Developed and maintained by the Laboratory of Computational Physiology at MIT, the publically accessible database contains anonymized data from more than 60,000 ICU admissions, with data stored across multiple tables and thousands of fields.

In order to give all teams in the datathon simultaneous and continuous access to the MIMIC database, the data were stored in the SAP HANA in-memory relational database management system and hosted on cloud servers. Database instances were deployed dynamically according to the data query demands of participants to allow for more efficient queries.

Note that the current MIMIC database serves as a model for the clinical databases of the future, which will grow more complete and, therefore, more useful as electronic health records become nearly universally implemented and an open, shared data philosophy becomes more widespread. Already, discussions are ongoing to build an international consortium that leverages data from ICUs in the United States, the United Kingdom, France, Belgium, Brazil, Japan, Australia, and New Zealand. Other specialties are following suit: in April 2015, the American Heart Association convened a 1.5-day forum to discuss critical issues in the acquisition, analysis, and sharing of data in the field of cardiovascular and stroke science (*7*).

CRITICAL CARE DATATHON

Our group hosted International Critical Care datathons in September 2014 and September 2015. The 2014 datathon spanned three days and three countries, with teams located in Cambridge, MA; London; and Paris. More than 200 participants across all locations gathered to conduct secondary data analysis using the MIMIC database. The organizing committee made a particular effort to prepare participants in advance of the event. Video lectures and online tutorials were available on the event website to educate participants about the MIMIC database, including guidelines to define clinical questions and tips to extract variables.

The introductory Friday evening session began with an overview of expectations for the weekend; a list of dos and don'ts were highlighted to make the weekend as productive as possible. Participants were encouraged to collaborate, fail fast, and iterate. This session was followed by problem-based pitches, in which participants shared clinical problems and knowledge gaps that could be addressed using the MIMIC database. After problem pitching, participants divided into specifically diverse teams during a stand-up dinner. Teams were required to have at a minimum one clinician, one data engineer, and one data scientist. We also encouraged participants to promote the event on Twitter (with labels #criticaldata and #criticaldata2014). The session ended with an overview of the MIMIC database and a hands-on tutorial, which helped minimize the technical hurdles related to software installation, cloud server connection, and programming syntax.

All of Saturday and part of Sunday were spent "hacking." During the event, participants were encouraged to share code (for example, SQL, Python, R, SAS, and STATA) used for data extraction and statistical analysis in a GitHub repository. Use of the GitHub repository served multiple purposes: (i) facilitated code review among members of the same team; (ii) facilitated interteam code review to cross-check methods for similarly themed projects (*8*, *9*); and (iii) allowed code to be reused for future related studies, while permitting customizations to be made based on modifications in a study design. On Sunday afternoon, preliminary findings were presented by each team and focused on the difficulties encountered and insights gained during the course of the weekend. However, the three simultaneous events unfolded with subtle differences and have provided insight for improvement in future datathons.

The 2014 Cambridge event saw more than 100 registrants with a wide range of self-identified professions, including biomedical/computer engineers (n = 37), physicians (n = 33), data scientists (n = 33), entrepreneurs (n = 22), biostatisticians (n = 11), patients (n = 4), nurses (n = 3), and pharmacists (n = 2). The presence of four patients was singular to the Cambridge hackathon and demonstrated a valuable opportunity for researchers to include a population usually relegated to study participation as subjects. In addition, the Cambridge event included some experienced participants who had previously attended other hackathons, including our inaugural, single-site MIT Critical Care datathon in January 2014, from which we gained valuable experience in the planning and execution of an event of this kind.

In London, more than 40 participants formed seven interdisciplinary teams. Research topics included examination of the relationship between lactic acidosis and mortality, the association between supranormal oxygen levels in the blood and survival after subarachnoid hemorrhage, and visualization of outcomes among obese patients in the ICU. The event also had a number of invited speakers who discussed opportunities arising from the increasing availability of data in health care.

The Paris event was structured to ensure that the relationships formed at the datathon would transition into long-term collaborations. To accomplish this, five teams were limited to three participants each, and each team included one intensivist, one biostatistician, and one data engineer. Instead of problem-pitching, the clinical questions were selected from previously submitted ideas. Last, in the spirit of continuous collaboration and peer review, the Paris-based organizing committee scheduled regular follow-up meetings and progress presentations with the participating teams.

The MIT Critical Care datathon has served as a model for and ushered in similar events in other specialties. The following year, in June 2015, the Cross Neurodegenerative Diseases datathon was held in Boston, MA (*10*), and

focused on developing approaches to understand similarities and differences across a variety of neurodegenerative diseases. About 30 scientists from leading institutions around the world participated in five teams, which had access to public datasets of interest to research teams studying Parkinson's and other neuromuscular diseases.

In September 2015, the follow-up second international MIT Critical Care datathon was held, bringing back participants from the previous year's event as well as new ones. A software platform (Fig. 1) was provided that integrated with Jupyter Notebook and included a custom Python package to facilitate data processing and analysis (*11*). The platform connected directly to the database, allowed documentation and information sharing among the team members as regards the research question and study design, and, most importantly, facilitated archiving and sharing of queries and codes among the teams for cohort selection, variable extraction, and data visualization and analysis. For example, teams assisted each other in identifying vasopressor and mechanical ventilation use, their doses or settings, and durations of use. Last, we encouraged the teams to publish their project notebooks at the MIMIC website once they publish their manuscripts, in order to share their patient cohorts, queries, and codes used for analysis.

Whereas the objective of the first International Critical Care datathon was to draw data scientists and frontline clinicians to a research community around MIMIC, the second international event focused on attracting participants who are committed to publishing their findings. The first datathon produced one publication (*12*), and another is currently under review. The second datathon is now poised to deliver its goal: Six out of the 10 teams submitted abstracts of their projects for presentation at the 2016 American Thoracic Society meeting in San Francisco, CA. All six projects were accepted.

LESSONS LEARNED

Biomedical research is often undertaken by investigators operating independently and working in isolation (*13*). For experiments such as double-blind randomized controlled trials, independence and isolation are critical measures to reduce bias. However, the overemphasis on isolation and independence in all types of research has come at the cost of validity and reproducibility. This problem is only exacerbated by the academic reward

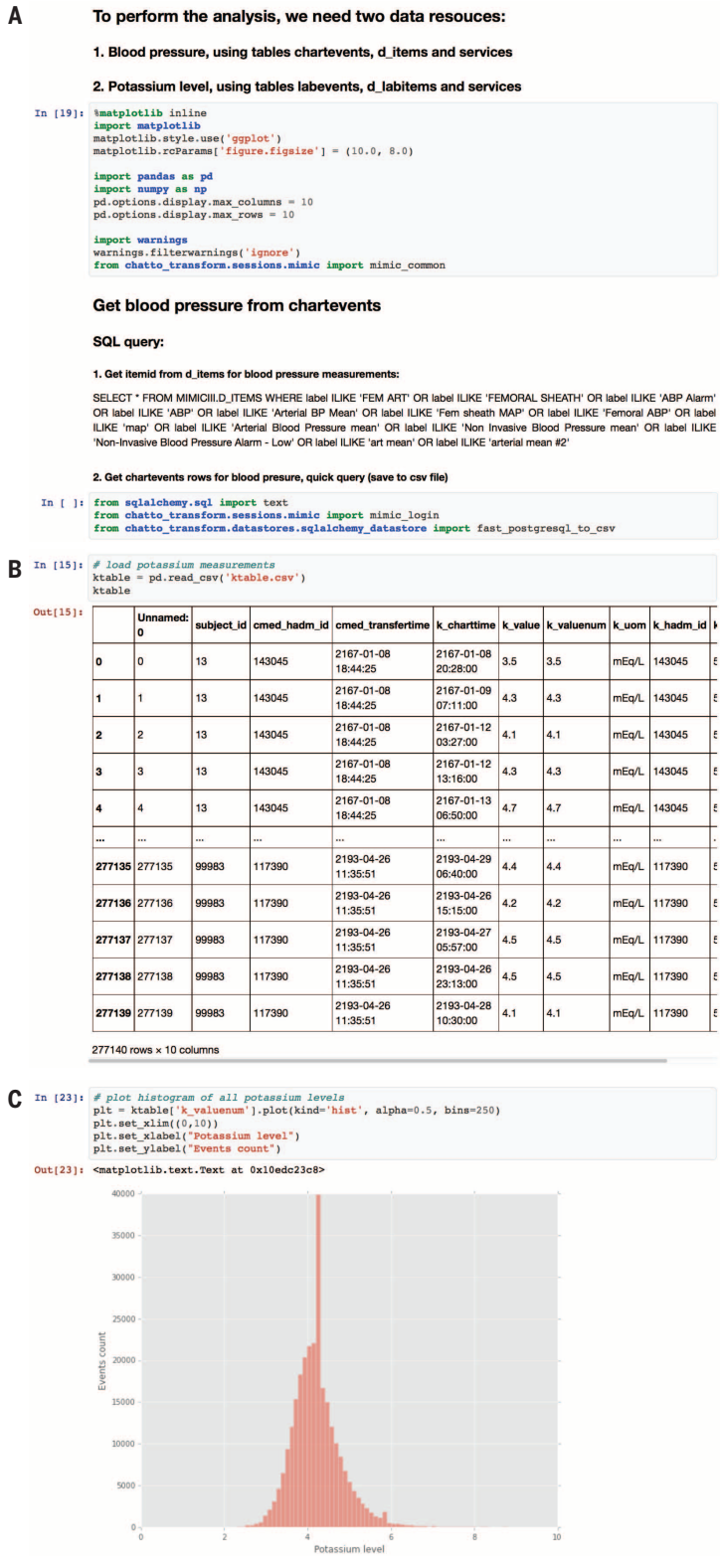

Fig. 1. Sample snapshots. Shown are screen shots of the Jupyter Notebooks used in the Second International Critical Care datathon to facilitate sharing of queries and codes among the teams. (**A**) A custom Python package was developed to assist with transforming and analyzing data. (**B**) Notebooks simplified data exploration. (**C**) Data visualization.

system, which is largely tailored toward the promotion of individual researchers or multiple researchers working in the same area of basic science, rather than work by interdisciplinary teams. Further, the current academic system effectively hinders potentially beneficial idea generation and collaboration across disparate entities, because data sharing restraints keep valuable data within host institutions. In contrast, our datathons aim to accelerate the discovery of evidence-based knowledge, increase collaboration via an open, cross teams–based approach during the design phase of research, and create a system of iterative peer review that might improve the reliability of research.

The importance of addressing communication failure and a lack of shared responsibility among collaborators was evident in an event involving a Duke laboratory that jolted the research world (*14*). Flawed gene-expression tests developed in the laboratory were used in three clinical trials to determine which chemotherapy treatment patients with lung or breast cancer would receive. Throughout this process, the responsibilities of the co-investigators on the research team and lines of accountability were apparently unclear. The U.S. Institute of Medicine, in response, published a manuscript in 2012 entitled *Evolution of Translational Omics: Lessons Learned and the Path Forward* (*15*). The report emphasized that given the complexity and multidisciplinary nature of omics research, there should be heightened attention in promoting a culture of teamwork across disciplines, in addition to scientific integrity and transparency.

Last, a potential advantage of upfront and ongoing collaborative interdisciplinary review is better recognition of which ideas are clearly worth investigation, which are interesting but not particularly promising, and which are extremely unlikely to be successfully pursued. At times, the latter are just the kinds of ideas that are new and innovative and should therefore be pursued. But at other times, wild ideas might simply lead to dead ends and wasted efforts. Team collaboration and discussion, in addition to the group's development of ideas in the first place, should help in determining whether such ideas are worthy of the effort involved to investigate them. With the potential for more efficiency in research, there might also be the opportunity for pilot efforts, in which unlikely but interesting ideas are investigated at a smaller scale by a select smaller team broken off from the main group for this purpose. Overall, the goal is to pursue important and interesting ideas that seem likely to pay off but also not to neglect unusual ideas that might ultimately be correct.

DOING DATATHONS RIGHT

With the increasing complexity of research in a world where the low-hanging fruit has been picked and the issues are often at the scale of "big data," research has already to some extent become a team-based process. We propose that multi-expertise viewpoints inserted openly into the process would aid in the conception, development, processing, and publication of research that is more reliable while hopefully remaining as interesting, innovative, and important as that produced by the current system. The datathon approach offers potential solutions to research problems such as poor upfront study design with inadequate statistics and an insufficient number of data samples. It might also provide the benefits of more and continuous transparency, as the input of many and more objective "eyes" is applied to the entire enterprise; this could lead to more objective and valid analyses and contents being provided to those in the next stage of peer review.

Although our datathons have not yet involved basic biomedical scientists, participation in these kinds of events might provide this group with ideas and opportunities. The early datathons have already given data-science experts a critical toehold in the space and allowed them to leverage their skills while exploring their evolving interests in the area. The datathons have also given a variety of clinicians normally excluded from the research process the opportunity to lend their practical experiential insights and have opened up the research arena to entirely new and valuable groups such as entrepreneurs and patients. Future datathons are likely to build and expand on these early innovations in collaborative research.

If half of reported research is indeed unreliable (*3*), those who read the literature are not only wasting half their time, they are actually poisoning half their time with falsehoods. There is an unacceptable time delay in the feedback loop between reported research results and the subsequent discovery that the results are not valid. In fact, the results might never even be carefully scrutinized and remain as unknown falsehoods in academic-literature limbo. It is important to deliver a research output that clinicians and patients can rely on to the extent that everything we read (including this paper) should be fundamentally reliable, while still approached with critical skepticism.

We are not so naïve as to believe that this kind of paradigm change will be easy or unchallenged. But the current level of research unreliability is unacceptable and is only likely to increase with the deluge of digital health data unless the fundamental fragilities of the research system are addressed. We also understand that, while the proposed datathon model of collaborative work might yield better science and better use of interdisciplinary personnel and resources, it is unlikely to have real impact until the academic reward system changes. The current research enterprise is an open control loop in which the output is not continuously evaluated for its impact on the plant. Academic pressures, sometimes compounded by influential funding by nondisinterested commercial parties, tend to move the research process relentlessly forward with a speed that at times compromises due diligence. A datathon model with its interdisciplinary team approach has the potential to raise and answer important new questions while potentially reducing the wasteful unreliability of the current system.

REFERENCES AND NOTES

1. F. Godlee, Research misconduct is widespread and harms patients. *BMJ* **344**, e14 (2012).
2. A.-W. Chan, F. Song, A. Vickers, T. Jefferson, K. Dickersin, P. C. Gøtzsche, H. M. Krumholz, D. Ghersi, H. B. van der Worp, Increasing value and reducing waste: Addressing inaccessible research. *Lancet* **383**, 257–266 (2014).
3. J. P. Ioannidis, How to make more published research true. *PLOS Med.* **11**, e1001747 (2014).
4. M. R. Macleod, S. Michie, I. Roberts, U. Dirnagl, I. Chalmers, J. P. Ioannidis, R. Al-Shahi Salman, A.-W. Chan, P. Glasziou, Biomedical research: Increasing value, reducing waste. *Lancet* **383**, 101–104 (2014).
5. E. T. Moseley, D. J. Hsu, D. J. Stone, L. A. Celi, Beyond open big data: Addressing unreliable research. *J. Med. Internet Res.* **16**, e259 (2014).
6. L. A. Celi, R. G. Mark, D. J. Stone, R. A. Montgomery, "Big data" in the intensive care unit. Closing the data loop. *Am. J. Respir. Crit. Care Med.* **187**, 1157–1160 (2013).
7. E. M. Antman, E. J. Benjamin, R. A. Harrington, S. R. Houser, E. D. Peterson, M. A. Bauman, N. Brown, V. Bufalino, R. M. Califf, M. A. Creager, A. Daugherty, D. L. Demets, B. P. Dennis, S. Ebadollahi, M. Jessup, M. S. Lauer, B. Lo, C. A. MacRae, M. V. McConnell, A. T. McCray, M. M. Mello, E. Mueller, J. W. Newburger, S. Okun, M. Packer, A. Philippakis, P. Ping, P. Prasoon, V. L. Roger, S. Singer, R. Temple, M. B. Turner, K. Vigilante, J. Warner, P. Wayte; American Heart Association Data Sharing Summit Attendees, Acquisition, analysis, and sharing of data in 2015 and beyond: A survey of the landscape: A conference report from the American Heart Association data summit 2015. *J. Am. Heart Assoc.* **4**, e002810 (2015).
8. D. C. Ince, L. Hatton, J. Graham-Cumming, The case for open computer programs. *Nature* **482**, 485–488 (2012).
9. G. Wilson, D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, P. Wilson, Best

practices for scientific computing. *PLOS Biol.* **12**, e1001745 (2014).
10. C. I. Barash, K. Elliston, R. Potenzone, tranSMART Foundation Datathon 1.0: The cross neurodegenerative diseases challenge. *Applied Transl. Genomics* **6**, 42–44 (2015).
11. F. Pérez, B. E. Granger, IPython: A system for interactive scientific computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
12. J. D. Salciccioli, D. C. Marshall, M. A. Pimentel, M. D. Santos, T. Pollard, L. A. Celi, J. Shalhoub, The association between the neutrophil-to-lymphocyte ratio and mortality in critical illness: An observational cohort study. *Crit. Care* **19**, 13 (2015).
13. J. P. Ioannidis, S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz, R. Tibshirani, Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166–175 (2014).
14. *The Economist* (2011) An array of errors. 10th September 2011.
15. C. M. Micheel, S. J. Nass, G. S. Omenn, *Evolution of Translational Omics: Lessons Learned and the Path Forward* (National Academies Press, Washington, DC, 2012).

Acknowledgments: The MIMIC database is funded by the U. S. National Institutes of Health grant R01 EB001659. The MIT Critical Care datathons were funded by the MIT International Science and Technology Initiatives. The London datathon was supported by The IET: The Institution of Engineering and Technology. **Competing interests:** The authors declare that they have no competing interests.

10.1126/scitranslmed.aad9072

Citation: J. Aboab, L. A. Celi, P. Charlton, M. Feng, M. Ghassemi, D. C. Marshall, L. Mayaud, T. Naumann, N. McCague, K. E. Paik, T. J. Pollard, M. Resche-Rigon, J. D. Salciccioli, D. J. Stone, A "datathon" model to support cross-disciplinary collaboration. *Sci. Transl. Med.* **8**, 333ps8 (2016).

A ''datathon'' model to support cross-disciplinary collaboration

Jerôme Aboab, Leo Anthony Celi, Peter Charlton, Mengling Feng, Mohammad Ghassemi, Dominic C. Marshall, Louis Mayaud, Tristan Naumann, Ned McCague, Kenneth E. Paik, Tom J. Pollard, Matthieu Resche-Rigon, Justin D. Salciccioli and David J. Stone

(April 6, 2016)
Science Translational Medicine **8** (333), 333ps8. [doi: 10.1126/scitranslmed.aad9072]

Editor's Summary

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.