tvst

Editorial

Data Science in *Translational Vision Science and Technology*

What Is Data Science?

Data science involves the use of a variety of quantitative methods (e.g. mathematics, statistics, computer science) to extract useful information from structured and unstructured data.[1] Typically, data scientists undertake exploratory data analysis by deploying machine learning principles and algorithms to identify patterns in raw data with the purpose of understanding processes and predicting outcomes. These analytic approaches include predictive causal analytics, prescriptive analytics, and machine learning for pattern discovery and outcome prediction, and they require a large volume and variety of data (i.e. structured as well as unstructured data).

What Are Software Libraries and Why Are They Important?

Software libraries comprise suites of data and programming code used to develop software programs and applications. One might think of them as toolkits (code) created to solve a particular problem (e.g. visual field analysis).[2] Once situated in an open-access environment, these libraries can be used by others to solve the same problem when it recurs, and the code, if it is an open-source code, also can be improved by others for the benefit of all. In other words, these libraries are a repository for software reuse, thus eliminating the need for application developers to create software that has been developed elsewhere. Ideally, well-constructed software libraries become community standards that facilitate analysis of large data sets and communication among members of a particular scientific community.[3] Software libraries are essential for modern software development.

Why Are Open Data Sets Important?

Properly curated data sets are extremely valuable. Why? Because much work must be done to create a data set that can serve as a "laboratory" for hypothesis testing and outcome prediction. Raw data from an electronic health record, for example, must be categorized according to relevant attributes (e.g. height, weight, age, gender, blood pressure, visual acuity, intraocular pressure, cup/disc ratio, central foveal thickness, sensitivity to the size III test target within 10 degrees of fixation, etc.). Having been so structured, inconsistencies, such as missing values and incorrect data (e.g. a physiologically impossible temperature such as 4000 degrees Fahrenheit or an intraocular pressure of -100 mm Hg) must be identified and rectified ("data cleaning") as a prelude to data analysis. These curated data sets are valuable because they can be interrogated for a variety of purposes, not just those intended by the scientists that assembled and curated the data set.[4,5] The results of phase III randomized clinical trials comprise a valuable data set that, if available publicly, can accelerate hypothesis testing regarding disease pathogenesis (e.g. by analyzing the genetic background of enrolled patients)[6] or that might be used to identify subgroups of patients who are exceptionally resistant/responsive to therapy[7] or who are at high risk for severe complications from therapy (e.g. cerebrovascular accidents).[8,9] The availability of phase III trial data also affords an opportunity for independent investigators to reproduce the results reported by the original investigative team.

Unfortunately, investigators often cannot obtain access to high-quality datasets in their area of interest. In part, this obstacle may arise because the scientists who generate data sets may feel there is little incentive to share the data. Furthermore, when available, data sets may be annotated poorly or organized inconsistently. If different data sets for similar disease processes are structured consistently, then they may

tvst.arvojournals.org | ISSN: 2164-2591

be combined for subsequent analysis. We believe that vision science will benefit from harmonized methods for data representation. Properly annotated data sets can be used to develop and validate new artificial intelligence algorithms.

What Are Data Science Descriptor Articles and Why Should *TVST* Publish Them?

Data Science descriptor articles provide an introduction to and complete description of data sets or software libraries, as well as access to them via links to open access repositories. For readers, publishing these articles improves visibility and understanding of valuable underlying research resources. Publication in a peer-reviewed scientific journal provides an assurance that the featured articles meet standards established by the journal editors. Authors benefit by being credited with a citable publication in a reputable journal. Publication in an open access journal will enhance the availability and use of these data sets and software libraries.

For these reasons, *Translational Vision Science and Technology* (*TVST*) is establishing a **Data Science** section that will publish Data Science descriptor articles featuring external, scientifically valuable data sets and software libraries relevant to all aspects of vision science. The data sets can vary in nature and may include observational data sets (e.g. from data developed in laboratory experiments, such as genomewide association studies or from registration clinical trials) and computational data sets. These data sets and software libraries will be required to be open-access and open-source and are intended for reuse by the scientific community with the goal of accelerating scientific discovery. The data sets and libraries will incorporate documentation and narrative content with curated, structured descriptions (metadata) of the data set and/or code. The procedures used to develop the data sets will be described in detail in the publication. These descriptions should include machine readable metadata files and must provide information that: (1) will enable other investigators to interpret, reuse, and reanalyze the primary data set; (2) will enable other investigators to link to the data repository (e.g. figshare or Dryad, or other digital repositories) in which the data are stored (*TVST* will not host the data set); and (3) will enable investigators who have developed the data set to demonstrate to funding agencies that they have fulfilled data-sharing requirements. Additional details, such as: (1) specification of the data repositories approved by the journal, and (2) criteria used to determine whether proposed changes to an existing database or software library will be identified as a new version of the existing data set (e.g. version 1.1, 1.2, etc.) or constitute the basis for a new publication (e.g. version 2.0) will be provided as part of the instructions to authors. Hypothesis testing or extensive analysis of the data set should be provided as a separate publication (preferably in the same issue of *TVST*) and not in the publication featured in the **Data Science** section. Data sets and software papers can appear as standalone publications in the **Data Science** section of *TVST* (i.e. can be published without an accompanying manuscript that uses the data set or software).

The **Data Science** publications are subject to peer-review in order to validate: (1) the quality of the procedures used to generate the data set or software; (2) the completeness/appropriateness of the descriptors; (3) the functionality of the proposed software; and (4) the reuse value of the data set or software. These publications are citable and must be cited by investigators who use the data set or software, which will afford the team who created these valuable resources appropriate credit. As with all *TVST* publications, content in the **Data Science** section is indexed by PubMed, Scopus, MEDLINE, Google Scholar, and Clarivate.

Full release of the research data and/or software upon publication in *TVST* is mandatory. All content in the **Data Science** section of *TVST* is published under a Creative Commons Attribution 4.0 International License (CC BY) to enable maximum reuse of these open access materials. This license allows users to share and adapt the data set/software (including for commercial purposes) provided that publication in *TVST* is cited as specified by the authors. The **Data Science** publication costs are competitive. To promote this important initiative, we are starting with a charge of $300 per article. (Please note that publication costs for the manuscript that reports extensive analyses or hypothesis testing of the data set/library are the routine full *TVST* publication costs [currently, $1500 for ARVO members and $1850 for non-members].) Authors who cannot afford publication costs can apply to the ARVO Foundation for financial support. Although **Data Science** descriptor articles in *TVST* will be published under the CC BY license, *TVST* does not require that the authors use CC BY for the data or software, just that the authors must use some open-source license.

Analysis of large data sets is integral to the incorporation of artificial intelligence in science and medicine. The establishment of data sets such as the Intelligent Research in Sight (IRIS) registry and other health care data bases is a manifestation of this fact.[10]

Development of open-source software libraries catalyzes rapid evolution of the analysis of large data sets and facilitates communication among members of a particular scientific community. Placing scientifically valuable data sets and software libraries in the public domain is part of an overall trend in which information, including personal information, and computational tools are being configured in a digital format that radically facilitates access to said information and tools.[11,12] The **Data Science** section of *TVST* is thus intended to accelerate the pace of discovery in vision science.

Marco A. Zarbin[1], Aaron Y. Lee[2], Pearse A. Keane[3], and Michael F. Chiang[4]

[1] Rutgers – New Jersey Medical School, Newark, New Jersey, USA
[2] University of Washington, Seattle, WA, USA
[3] Moorfields Eye Hospital NHS Foundation Trust, Moorfields Eye Hospital NHS Foundation Trust, London, UK
[4] National Eye Institute, Bethesda, MD, USA

Correspondence: Marco A. Zarbin, Rutgers - New Jersey Medical School, 90 Bergen Street, 6th Floor, Newark, New Jersey 07101-1709, USA. e-mail: zarbin@earthlink.net

References

1. Dhar V. Data science and prediction. *Communications of the ACM.* 2013;56(12):64–73.
2. Marin-Franch I, Swanson WH. The visualFields package: a tool for analysis and visualization of visual fields. *J Vis.* 2013;13(4):10.
3. Sterling T, Anderson M, Brodowicz M. *High Performance Computing: Modern Systems and Practices.* Wonder Lake, IL: Morgan Kaufmann Publishers; 2018.
4. Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci.* 2018;59(7):2861–2868.
5. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng.* 2018;2(3):158–164.
6. Hoffman JD, van Grinsven MJ, Li C, et al. Genetic association analysis of Drusen progression. *Invest Ophthalmol Vis Sci.* 2016;57(4):2225–2231.
7. Chew EY, Klein ML, Clemons TE, et al. No clinically significant association between CFH and ARMS2 genotypes and response to nutritional supplements: AREDS report number 38. *Ophthalmology.* 2014;121(11):2173–2180.
8. Zarbin MA, Dunger-Baldauf C, Haskova Z, et al. Vascular safety of ranibizumab in patients with diabetic macular edema: a pooled analysis of patient-level data from randomized clinical trials. *JAMA Ophthalmol.* 2017;135(5):424–431.
9. Avery RL, Gordon GM. Systemic safety of prolonged monthly anti-vascular endothelial growth factor therapy for diabetic macular edema: a systematic review and meta-analysis. *JAMA Ophthalmol.* 2016;134(1):21–29.
10. Parke DW, II, Lum F, Rich WL. The IRIS(R) registry : purpose and perspectives. *Ophthalmologe.* 2017;114(Suppl 1):1–6.
11. Feng X, Yang L, Wang L, Vinel A. Internet of things. *Int J Communication Systems.* 2012;25:1101–1102.
12. Iansiti M, Lakhani K. Digital Ubiquity. How connections, sensors, and data are revolutionizing business. *Harvard Business Review.* 2014;11:90–99.

Amended July 26, 2021: Middle initials have been added to the authors' names.

translational vision science & technology