

اشتراک گذاری داده‌ها و آینده علم

مهدی گلستانی نسب*

سمنان، دانشگاه سمنان، دانشکده علوم، گروه زیست شناسی

اشتراک‌گذاری داده‌ها برای چه کسانی مفید واقع خواهد شد؟ مشخصاً دانشمندان آینده، زیرا همان داده‌هایی که امروز به اشتراک گذاشته می‌شوند، پایه و اساس علم فردا را تشکیل خواهند داد. بر پایه شواهد موجود، مطالعات مبتنی بر تحلیل داده‌های قبلاً انتشار یافته می‌توانند به جای آنکه صرفاً تکرار مکررات مجموعه داده‌های قدیمی باشند، به همان اندازه پروژه‌های اولیه تاثیرگذار واقع شوند.

پست الکترونیکی: mgolestaninasab@semnan.ac.ir

تصویربرداری سیستم عصبی، مقالات منتشر شده بر پایه داده‌های به اشتراک‌گذاری شده در مقایسه با مقالات ارائه دهنده داده‌های اصلی از احتمال یکسانی برای انتشار در مجلات علمی با ضریب تأثیر بالا برخوردارند و همچنین به همان خوبی نیز مورد استناد واقع شده‌اند. هرچند استنادات یک مقاله و وجهه مجله علمی چاپ کننده آن معیارهای مستقیم کیفیت یا بدیع بودن یک برونداد علمی نیستند، با این حال، یافته‌های میلهام و همکاران می‌توانند قوت قلبی برای دانشمندان علوم اعصاب شناختی باشند که نگران هستند عدم گردآوری داده‌های اصلی ممکن است از تاثیرگذاری کار آنها بکاهد.

در واقع، اشتراک‌گذاری داده‌ها نه تنها مانعی در پیشبرد علوم نوین به شمار نمی‌رود، بلکه زمینه انجام پژوهش‌های جدیدی را نیز فراهم ساخته است. برای مثال، پژوهش با استفاده از مجموعه داده‌های پروژه Human Connectome Project (HCP) را در نظر بگیرید که به عنوان یکی از رویکردهای خلاقانه اشتراک‌گذاری داده‌ها در مطالعه میلهام و همکاران لحاظ شده است. در حال حاضر، HCP حاوی مجموعه گسترده‌ای از داده‌های fMRI، MRI ساختاری و داده‌های رفتاری ۱۲۰۰ داوطلب جوان بالغ و سالم است (<https://www.humanconnectome.org/study/hcp-young-adult>) و افزودن داده‌های مربوط به مغز کودکان، نوجوانان و بزرگسالان مسن‌تر به آن در دست انجام است. این داده‌ها برای تمامی پژوهشگران علاقه‌مند قابل دسترس هستند.

هر چند اشتراک‌گذاری داده‌ها در حوزه علوم اعصاب شناختی در بدو امر با مشکلاتی همراه بود (۴)، اما موفقیت HCP و بسیاری از مطالعات متأثر از آن نشان داد زمان

اشتراک‌گذاری داده‌ها در بسیاری از حوزه‌های پژوهشی از پیشینه بلند بالایی برخوردار است. هرچند اعمال فشار برای ترغیب کردن دانشمندان علوم اجتماعی و زیستی جهت اشتراک‌گذاری و عرضه عمومی یافته‌های خود اخیراً باب شده است (۱)، با این حال، در شماری از زمینه‌های دیگر، اشتراک‌گذاری داده‌ها از مدت‌ها پیش مرسوم بوده است. برای مثال، بیش از یک قرن است که بخش‌های زیادی از علوم اقتصاد و هواشناسی بر پایه داده‌های به اشتراک گذاشته شده که در دسترس عموم هستند، پیش برده شده است.

«این مسئله که عدم اشتراک‌گذاری داده‌ها سد راه دانشمندان آینده خواهد بود، در زمره مسائلی است که نیاز به بحث جدی دارد».

با این حال، نگرانی در خصوص اشتراک‌گذاری داده‌ها کماکان در جامعه علمی به ویژه در برخی رشته‌های خاص رواج دارد. در رابطه با مسائلی که باعث می‌شوند برخی پژوهشگران تمایلی برای اشتراک‌گذاری داده‌های خود نداشته باشند، بحث‌های زیادی انجام شده است (۲)، اما پژوهشگرانی که در نظر دارند داده‌های اشتراک گذاشته شده را اساس پژوهش خود قرار دهند نیز نگرانی‌هایی دارند: «من که در نظر دارم یک اثر کاملاً تاثیرگذار منتشر کنم، آیا نیاز به گردآوری داده‌های جدید ندارم؟» اما آیا این گردآوری داده‌های اولیه و اصلی است که یک مطالعه را به اثری تازه و بدیع تبدیل می‌کند؟

مزایای اشتراک‌گذاری داده‌ها فراتر از آن هستند که بتوان آنها را برشمرد. اما مطالعه مایکل پ. میلهام (Milham M.) و همکاران (۳) شواهدی مستقیم ارائه می‌دهد که در حوزه

که فراتر از مقیاس‌های زمانی مقرر انجام پژوهش یا چرخه‌های تأمین مالی عمل می‌کنند. گستره اشتراک‌گذاری داده‌ها در این زمینه به مرور زمان تحول پیدا کرده است (۱۰) اما آثار انتشار یافته تأثیرگذار هم اکنون بیش از هر زمان دیگری بر پایه پایگاه‌های داده تحت پشتیبانی شبکه‌های بزرگ پژوهشگران ارائه می‌شوند (۱۱).

این مثال‌ها یک مزیت بارز اشتراک‌گذاری داده‌ها را نشان می‌دهند، بدین ترتیب که با در اختیار قرار دادن مجموعه داده‌های بزرگ که گردآوری آنها هزینه قابل توجهی را می‌طلبد، پژوهشگران انفرادی قادر خواهند بود محدودیت‌ها و مشکلات مرتبط با کمبود منابع مالی خود را کنار بگذارند. بدین ترتیب، اشتراک‌گذاری داده‌ها در نهایت افق‌های پژوهشی پیش‌بینی نشده را پیش روی آنها باز خواهد کرد. این تنها در رابطه با راهکارهای اشتراک‌گذاری داده‌های بزرگ-مقیاس صدق نمی‌کند: حتی مجموعه داده‌های نسبتاً کوچک نیز در صورت اشتراک‌گذاری می‌توانند همانند داده‌های بزرگ موثر واقع شده و به شیوه‌هایی دور از انتظار در پیشبرد اکتشافات علمی آینده ایفای نقش کنند. برای مثال، در پزشکی، متاتحلیل در سطح بیماران در خصوص تعداد زیادی از کارآزمایی‌های بالینی گذشته به یافته‌های جدید متعددی منتهی شده است که بسیار فراتر از هدف اولیه پژوهش‌های تولید کننده این داده‌ها بوده است (بری مثال، منبع ۱۲ را ببینید).

بنابراین، اشتراک‌گذاری داده‌ها نه تنها روشی برای بهبود قابلیت بازتولید و اعتمادپذیری علم امروز است (۱۳)، بلکه می‌تواند زمینه‌ساز دستیابی به علم جدید برای فردا نیز باشد. با توجه به اینکه امروزه قادر نیستیم پیش‌بینی کنیم یک مجموعه داده معین در آینده تا چند اندازه ارزشمند خواهد بود، این مسئله جای بحث جدی دارد که عدم اشتراک‌گذاری داده‌ها به منزله سنگ گذاشتن پیش پای دانشمندان آینده قلمداد می‌شود. در واقع، می‌توان زمانی را متصور شد که اشتراک‌گذاری داده‌ها دیگر یک رویکرد نوآورانه محسوب نخواهد شد، بلکه به یک بخش معمول و اساسی فرآیند علمی مبدل می‌شود، تقریباً شبیه آنچه در فرآیند مرور-همتا نظاره‌گر هستیم.

روآوردن به این امر مهم فرا رسیده است. بدون اشتراک‌گذاری داده‌ها، اسکن کردن ۱۲۰۰ نفر برای یک گروه پژوهشی واحد کاملاً دور از ذهن به نظر می‌رسد. اسکن‌های MRI پرهزینه هستند و مطالعات تصویربرداری سیستم عصبی که از داده‌های اصلی استفاده می‌کنند، عموماً ۲۰ تا ۵۰ شرکت کننده را در بر می‌گیرند. هرچند این حجم نمونه برای انجام آن دسته مطالعاتی که یک دهه قبل پیشرو قلمداد می‌شدند کفایت می‌کرد، اما امروزه روش‌های پیشرفته‌تر به داده‌های بسیار بیشتری نیاز دارند.

این مهم که اشتراک‌گذاری داده‌ها نوع مطالعاتی که پژوهشگران قادر به انجام آنها هستند را تغییر داده است، صرفاً محدود به علوم اعصاب نمی‌شود. در علم ژنتیک، ژنومیکس و زیست شناسی ساختاری، مجموعه داده‌های اشتراک‌گذاری شده بزرگ رایج هستند (برای مثال، منبع ۵ را ببینید) و پژوهشگران از مجموعه داده‌های منتشر شده قبلی، برای دستیابی به یافته‌های جدید در این حوزه‌ها استفاده کرده یا حتی مجدداً استفاده کرده‌اند (۶).

در علوم فیزیک نیز اشتراک‌گذاری داده‌ها به سرعت رواج پیدا کرد. برای مثال، در نجوم و فیزیک نجوم داده‌های تلسکوپی اساساً در دسترس عموم قرار دارند (۷)؛ بدون این گونه اشتراک‌گذاری، اغلب گروه‌های پژوهشی قادر به ارائه دستاوردهای اکتشافی پیشرو نخواهد بود زیرا از منابع مالی مورد نیاز برای ساخت این قبیل تلسکوپ‌های بزرگ ضروری برای پژوهش‌های نجومی نوین برخوردار نیستند. اشتراک‌گذاری داده‌ها در علم نجوم به اندازه‌ای گسترش یافته است که از طریق برنامه SETI@home تحت پشتیبانی دانشگاه برکلی به رایانه‌های شخصی نیز راه پیدا کرده است و زمینه مشارکت علم شهروندی (citizen science) در تحلیل داده‌ها را نیز فراهم می‌آورد (۸).

رشته اکولوژی به واسطه اشتراک‌گذاری داده‌ها تحت شبکه پژوهش‌های اکولوژی بلندمدت (Long-Term Ecological Research) یا LTER ایالات متحده آمریکا، گام‌های بلندی در این خصوص برداشته است (۹). این شبکه که حاوی مجموعه‌ای از مشاهدات بلندمدت در میان اکوسیستم‌های مختلف است، برای اکولوژیست‌ها فرصتی را فراهم ساخته است تا آن دسته از الگوهای مهم تأثیرگذار را کشف نمایند

نتوانسته‌اند علی‌رغم بهره‌گیری از داده‌های اشتراک‌گذاری شده به اسرار مغز انسان پی ببرند، با توجه به اینکه حدود ۸۶ میلیارد نورون در مغز هر شخص وجود دارد (۱۴)، شناخت همه آنها همکاری نزدیک بین دانشمندان را می‌طلبد.

هرچند SETI@home هنوز هیچ هوش بیگانه‌ای را کشف نکرده است، اما نشان داده است که میلیاردها ستاره در کهکشان ما وجود دارد: به راستی چگونه می‌توانستیم این همه ستاره را مطالعه کنیم، اگر به جای کار مشترک جداگانه کار می‌کردیم؟ دانشمندان علوم اعصاب تاکنون

منابع

1. Gewin, V. Data sharing: an open mind on open data. *Nature* 529, 117–119 (2016).
2. Tenopir, C. et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE* 10, e0134826 (2015).
3. Milham, M. P. et al. Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.* 9 (2018). <https://doi.org/10.1038/s41467-018-04976-1>.
4. Van Horn, J. D. & Gazzaniga, M. S. Why share data? Lessons learned from the fMRIDC. *Neuroimage* 82, 677–682 (2013).
5. Genome Aggregation Database (gnomAD). <http://gnomad.broadinstitute.org/>.
6. Bonàs-Guarch, S. et al. Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* 9, 321 (2018).
7. How big data advances physics. Marc Chahin June 27, 2017 (blog post).
8. SETI@home. <https://setiathome.berkeley.edu/>.
9. Long-Term Ecological Research Network (LTER). <https://lternet.edu/>.
10. Michener, W. K. Ecological data sharing. *Ecol. Inform.* 29, 33–44 (2015).
11. The Earth Microbiome Project. <http://www.earthmicrobiome.org/>.
12. Fournier, J. C. et al. Antidepressant drug effects and depression severity: a patient-level metaanalysis. *JAMA* 303, 47–53 (2010).
13. On data availability, reproducibility and reuse. *Nat. Cell Biol.* 19, 259 (2017).
14. Azevedo, F. A. et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513, 532–541 (2009).