فرهنگ و اندیشهٔ ریاضی
سال ۴۳، شمارهٔ ۱ (بهار و تابستان ۱۴۰۳)، صص ۲۷۳ تا ۳۲۷

پنجاه‌سال علم داده*

دیوید دونوهو
ترجمهٔ محمدقاسم وحیدی‌اصل

چکیده. بیش از ۵۰ سال پیش، جان توکی فراخوانی برای بازسازی آمار دانشگاهی داد. او در مقالهٔ «آیندهٔ تحلیل داده‌ها» به وجود علمی اشاره کرد که آن زمان به رسمیت شناخته نشده بود و موضوع مورد توجه آن یادگیری از داده‌ها یا «تحلیل داده‌ها» بود. پدیده‌ای مربوط به این اواخر که رو به گسترش نیز است، ظهور برنامه‌های «علم داده» در دانشگاه‌های بزرگ است. این مقاله در پی مرور برخی از اجزای «برههٔ علم دادهٔ» کنونی، ازجمله بر اظهارنظرهای اخیر دربارهٔ علم داده در رسانه‌های همگانی، است که آیا علم داده واقعاً با آمار تفاوت دارد یا خیر و اگر دارد، چگونه. برداشت فعلی از علم داده به‌مثابهٔ زیرمجموعه‌ای از رشته‌های آمار و یادگیری ماشین است که مقداری فنّاوری به منظور «ارتقای مقیاس» آن برای «داده‌های بزرگ» به آن افزوده شده است. انگیزهٔ انتخاب این زیرمجموعه، پیشرفت‌های تجاری است و نه اندیش‌ورزانه. چنین انتخابی محتملاً رویدادهای اندیش‌ورزانهٔ مهم ۵۰ سال آینده را نادیده خواهد گرفت. ازآنجاکه خودِ کلِّ علم به‌زودی بدل به داده می‌شود که می‌توان آن را کاوید، انقلاب در شرف وقوع در علم داده محدود به «ارتقای مقیاس» نبوده، بلکه در مقابل، در ظهور مطالعات علمی تحلیل داده‌ها در سطح کل علم است. دیدگاهی از علم داده را براساس فعالیت‌های افرادی که «از داده‌ها یاد می‌گیرند» ارائه می‌کنم و یک رشتهٔ دانشگاهی را که به بهبود این فعالیت به شیوه‌ای مبتنی‌بر شواهد اختصاص دارد توصیف می‌کنم. این رشتهٔ جدید وسعت‌بخشی علمی بهتری از آمار و یادگیری ماشین در مقایسه با ابتکارعمل‌های امروزی در علم داده است، و هم‌زمان توان آن را دارد که همان اهداف کوتاه‌مدت را برآورده کند.

۱ بُرههٔ علم دادهٔ امروزی

در سپتامبر ۲۰۱۵، زمانی که این یادداشت‌ها را آماده می‌کردم، دانشگاه میشیگان «ابتکارعمل علم

عبارات و کلمات کلیدی: علم داده، یادگیری ماشین، آمار، تحلیل داده‌ها، مدل‌بندی پیشگوگر
نوع مقاله: ترویجی؛ تاریخ دریافت: ۱۴۰۲/۳/۲۴؛ تاریخ پذیرش: ۱۴۰۲/۴/۶

*Donoho, D., 50 Years of Data Science, *J. Comput. Graph. Statist.*, **26** (2017), no. 4, 745-766.

doi: 10.30504/mct.2023.1398.1982

دادهٔ» (DSI) ۱۰۰ میلیون دلاری خود را اعلام و در نهایت ۳۵ عضو هیئت علمی جدید استخدام کرد. بیانیهٔ مطبوعاتی دانشگاه حاوی اظهارات جسورانه‌ای است:

> علم داده به چهارمین رویکرد در اکتشاف علمی، علاوه‌بر آزمایشگری، مدل‌بندی، و محاسبات، تبدیل شده است. (نقل از مارتا پولاک، رئیس دانشگاه)

وبگاه «ابتکارعمل علم داده» منظور از این را که علم داده چیست برای ما روشن می‌کند:

> این جفت‌شدگی اکتشاف علمی و عمل‌ورزی شامل جمع‌آوری، مدیریت، پردازش، تحلیل، دیداری‌سازی، و تفسیر حجم عظیمی از داده‌های ناهمگن مرتبط با آرایهٔ جورواجوری از کاربردهای علمی، واگردانی، و بین‌رشته‌ای است.

چنین اعلامی در خلأ صورت نمی‌گیرد. تعدادی ابتکارعمل شبیه ابتکارعمل علم داده اخیرا آغاز شده است، ازجمله الف) ابتکارعمل‌های در سطح دانشگاه در دانشگاه‌های نیویورک، کلمبیا، اِم‌آی‌تی ...، ب) برنامه‌های جدید کارشناسی ارشد در علم داده، به‌عنوان مثال، در دانشگاه‌های برکلی، نیویورک، استنفورد، کارنگی ملون، ایلینوی، .... هر هفته اعلامیه‌های جدیدی از چنین ابتکارعمل‌هایی دیده می‌شود.

۲ علم داده «علیه» آمار

بسیاری از مخاطبان من در گردهمایی سدهٔ توکی ــ جایی که این مطالب ابتدا در آنجا ارائه شد ــ آماردانان کاربردی بودند، و زندگی حرفه‌ای آن‌ها ازنظر خودشان، رشته‌ای طولانی از عمل‌ورزی‌ها در موضوعات فوق «جمع‌آوری، مدیریت، پردازش، تحلیل، دیداری‌سازی، و تفسیر حجم عظیمی از داده‌های ناهمگن مرتبط با آرایهٔ جورواجوری از . . . کاربردهاست.» درحقیقت، برخی از سخنرانی‌ها در سدهٔ توکی، روایت‌هایی نمونه‌گونه از «جمع‌آوری، مدیریت، پردازش، تحلیل، دیداری‌سازی، و تفسیر حجم عظیمی از داده‌های ناهمگن مرتبط با آرایهٔ جورواجوری از . . . کاربردها بوده است.»

پدیدهٔ DSI ممکن است برای آماردانان گیج‌کننده به نظر برسد. از نظر آماردانان، مدیران آن فعالیت‌هایی را به‌عنوان فعالیت‌های جدید می‌ستایند که شغل روزانهٔ آن‌ها در سرتاسر زندگی کاریشان بوده، و زمانی که آن‌ها در دورهٔ تحصیلات تکمیلی به تحصیل اشتغال داشتند، این فعالیت‌ها استاندارد تلقی می‌شد.

نکات زیر حرف‌های زیادی در مورد برنامهٔ DSI دانشگاه میشیگان برای چنین آماردانانی دارد:

- برنامهٔ DSI دانشگاه میشیگان در دانشگاهی با یک گروه آمار بزرگ و بسیار مشهور اتفاق می‌افتد.
- رهبران شناسایی‌شدهٔ این ابتکارعمل، از اعضای هیئت علمی گروه مهندسی برق و علوم کامپیوتر (اَل هیرو) و دانشکدهٔ پزشکی (برایان اِیتی) هستند.
- سمپوزیوم افتتاحیه، فقط یک سخنران از گروه آمار (سوزان مورفی) از بین بیش از ۲۰ سخنران داشت.

لاجرم، بسیاری از آماردانان دانشگاهی چنین برداشت خواهند کرد که آمار در اینجا به حاشیه رانده شده است. پیام ضمنی این ملاحظات آن است که آمار بخشی از آن ماجراهایی است که در علم داده روی می‌دهند، اما بخش خیلی بزرگی از آن نیست. درعین‌حال، بسیاری از توصیفات ملموس از کاری که DSI واقعاً انجام خواهد داد، به نظر آماردان‌ها، کار آماری یومیهٔ آن‌هاست. آمار ظاهراً کلمه‌ای است که آن‌ها جرأت به زبان آوردن آن را در ارتباط با چنین ابتکارعمل‌هایی ندارند!

با جستجو در اینترنت برای کسب اطلاعات بیشتر دربارهٔ اصطلاح درحال‌ظهور «علم داده»، با تعریف‌های زیر از «قواعد رفتاری» انجمن علم داده مواجه می‌شویم

> «دانشمند علم داده» به معنای فردیْ حرفه‌ای است که از روش‌های علمی برای آزادسازی و خلق معنا از داده‌های خام استفاده می‌کند.

برای آماردانان این گفته بسیار شبیه به همان کاری است که آماردانان کاربردی انجام می‌دهند: استفاده از روش‌شناسی برای استنباط از داده‌ها. و در ادامه می‌گویند:

> «آمار» به معنای عمل‌ورزی یا علم جمع‌آوری و تحلیل داده‌های عددی در اندازه‌های بزرگ است. از نظر آماردانان، این تعریف از آمار پیشاپیش دربرگیرندهٔ هر آن چیزی است که ممکن است شامل حال تعریف دانشمند داده بشود، اما تعریف آماردان به نظر محدودیت‌آمیز می‌رسد، زیرا بسیاری از کارهای آماری صراحتاً دربارهٔ استنباط‌هایی است که از نمونه‌های بسیار کوچک انجام می‌شود ــ این امر واقعاً برای صدها سال صادق بوده است. در واقعیت امر، آماردانان با داده‌ها دست‌وپنجه نرم می‌کنند ــ هر اندازه بزرگ یا هراندازه کوچک.

حرفهٔ آمار در برههٔ گیج‌کننده‌ای گرفتار آمده است: فعالیت‌هایی که در طول قرن‌ها آن را به خود مشغول کرده بود، اکنون در کانون توجه همگان قرار گرفته است، اما ادعا می‌شود که همان فعالیت‌ها

به‌طرز چشم‌نوازی جدیدند، و توسط افراد نوپا و غریبه انجام می‌شوند (هرچند در واقعیت امر توسط آن‌ها اختراع نشده است). سازمان‌های آمار حرفه‌ای گوناگون به این موضوع واکنش نشان می‌دهند:

- آیا کار ما علم داده نیست؟، سرمقالهٔ ماری داویدیان، رئیس انجمن آمار آمریکا در خبرنامهٔ انجمن آمار آمریکا، ژوئیهٔ ۲۰۱۳.
- یک مناقشهٔ بزرگ: آیا علم داده فقط یک «تغییر نام تجاری» آمار است؟، مارتین گودسون، از سازمان‌دهندگان نشست ۱۹ام ماه مه، ۲۰۱۵، دربارهٔ رابطهٔ آمار و علم داده، در پست‌های اینترنتی در تبلیغ آن رویداد.
- بگذارید ما مالک علم داده باشیم، خطابهٔ بین یو در مقام ریاست مؤسسهٔ آمار ریاضی (IMS)، چاپ‌شده در بولتن IMS، اکتبر ۲۰۱۴.

برای یافتن وبلاگ‌هایی که از این سردرگمی در مورد این وضعیت جدید بهره‌برداری می‌کنند، نیازی به این نیست که جای دوری برویم:

- حالا که قرن‌هاست آمار را در اختیار داریم، چه حاجتی به علم داده است؟ ایروینگ ولاداوسکی–برگر، وال استریت جورنال، گزارش ریاست هیئت مدیره، ۲ مه، ۲۰۱۲.
- علم داده همان آمار است.

> وقتی فیزیک‌دانان کارِ ریاضی انجام می‌دهند، آن‌ها نمی‌گویند که در حال انجام علم عددند. آن‌ها در حال انجام کار ریاضی هستند. اگر شما داده تحلیل می‌کنید، در حال انجام آمار هستید. می‌توانید آن را علم داده یا انفورماتیک یا تحلیل‌گری یا هر چیز دیگری بنامید، اما کار، هنوز کار آماری است. . . . شاید کاری را که برخی آماردانان انجام می‌دهند دوست نداشته باشید. ممکن است احساس کنید که آن‌ها وجه مشترکی با ارزش‌های شما ندارند. آن‌ها ممکن است باعث شرمساری شما باشند، اما این امر نباید موجب آن شود که اصطلاح «آمار» را کنار بگذاریم. (کارل برومن، دانشگاه ویسکانسین)

ازسوی‌دیگر، می‌توانیم آتش‌بیاران معرکه‌ای را مشاهده کنیم که (تقریباً) بی‌ربط بودن آمار را اعلام می‌کنند:

- علم داده بدون آمار ممکن و حتی مطلوب‌تر است. (وینسنت گرانویل، بلاگ مرکزی علم داده)

• آمار، کم‌اهمیت‌ترین جزء علم داده است. (اندرو گلمن، دانشگاه کلمبیا)

واضح است که دیدگاه‌های زیادی از علم داده و ارتباط آن با آمار وجود دارد. در بحث‌هایم با دیگران، با «میم»های مکرری مواجه شده‌ام. فعلاً به اصلی‌ترین آن‌ها می‌پردازم.

۱.۲ «میم» علم داده

بیانیهٔ مطبوعاتی‌ای را که ابتکارعمل علم دادهٔ دانشگاه میشیگان را اعلام کرد و ما این مقاله را با آن آغاز کردیم، در نظر بگیرید. رئیس دانشگاه میشیگان، مارک شلایسل، اصطلاح «داده‌های بزرگ» را به‌طور مکرر به کار می‌برد، اهمیت آن را در همهٔ زمینه‌ها می‌ستاید، و بر ضرورت علم داده برای مدیریت چنین داده‌هایی تاکید می‌کند. نمونه‌هایی از این گرایش تقریباً در همه‌جا حیّ و حاضرند.

می‌توانیم بلافاصله «داده‌های بزرگ» را به‌عنوان معیاری برای تمایز معنادار بین آمار و علم داده نفی کنیم.

تاریخ خودِ اصطلاح «آمار» در آغاز تلاش‌های نوین برای جمع‌آوری داده‌های سرشماری، یعنی داده‌های جامع دربارهٔ همهٔ سکنهٔ یک کشور، به‌عنوان مثال، فرانسه یا ایالات متحده، ابداع شد. داده‌های سرشماری تقریباً در مقیاس داده‌های بزرگ امروزی‌اند؛ اما بیش از ۲۰۰ سال است که در همین گوشه‌وکنار حضور دارند! یک آماردان، هولریت، اولین پیشرفت عمده در داده‌های بزرگ را به منصهٔ ظهور درآورد: کارت‌پانچ‌خوان که امکان جمع‌آوری کارآمد یک سرشماری جامع ایالات متحده را فراهم می‌کرد. این پیشرفت منجر به تشکیل شرکت IBM شد که در نهایت به نیروی محرکه‌ای بدل شد که محاسبات و داده‌ها را در مقیاس‌هایی بازهم بزرگ‌تر فراتر برد. مدت مدیدی است که آماردانان مشکلی با مجموعه‌داده‌های بزرگ نداشته‌اند، و برای چندین دهه کنفرانس‌هایی را برگزار کرده‌اند و متخصصان «مجموعه‌داده‌های بزرگ» را ــ هرچند که تعریف بزرگ همواره در حال گسترش یافتن بوده است ــ گرد هم آورده‌اند.

علم دهه‌هاست که پژوهشگران حوزهٔ آمار ریاضی، به دنبال درک علمی مجموعه‌داده‌های بزرگ بوده‌اند. آن‌ها بر این مطلب تمرکز داشته‌اند که وقتی پایگاه داده دارای تعداد زیادی از افراد یا تعداد زیادی از اندازه‌گیری‌ها یا هر دو باشد، چه اتفاقی می‌افتد. این تصور که جمع کثیری از آن‌ها به این کار نمی‌پردازند یا این کار مشغلهٔ فکری عمدهٔ آن‌ها نیست، کاملاً نادرست است. ازجملهٔ کشفیات اصلی آمار به عنوان یک رشتهٔ علمی، نمونه‌گیری و بسندگی بود که

امکان دست‌وپنجه نرم کردن با مجموعه‌داده‌های بسیار بزرگ را به‌طرزی بسیار کارآمد فراهم می‌کند. این ایده‌ها دقیقاً به این دلیل کشف شدند که آماردانان به مجموعه‌داده‌های بزرگ اهمیت می‌دهند.

چارچوب «علم داده = داده‌های بزرگ» به هیچ چیز عمیقاً ماهوی در مورد رشته‌های مربوط منتهی نمی‌شود.

۲.۲ «میم» مهارت‌ها

در مکالماتی که شاهد آن‌ها بوده‌ام، به نظر می‌رسد که حرفِ آخر متخصصان کامپیوتر، نکات محور بحث زیر بوده است:

الف) علم داده به داده‌های واقعاً بزرگ مربوط می‌شود که منابع محاسباتی سنتی نمی‌توانند آن‌ها را در خود جای دهند.

ب) مهارت‌آموختگان علم داده، مهارت‌های لازم برای مواجهه با چنین مجموعه‌داده‌های بزرگ را دارند.

این استدلال، مُهر تأکید بیشتری بر میم «داده‌های بزرگ» از طریق لایه‌بندی آن و قرار دادن «میم مهارت‌های داده‌های بزرگ» در لایهٔ اول می‌گذارد.

آن مهارت‌ها چیستند؟ در اوایل دههٔ ۲۰۱۰ بسیاری از افراد از تسلط بر هادوپ ــ گونه‌ای از نقشه‌نمایی کردن/ساده کردن برای استفاده با مجموعه‌داده‌هایی که در خوشه‌ای از کامپیوترها توزیع شده‌اند ــ یاد می‌کردند. به مرجع استاندارد هادوپ: راهنمای تامّ و تمام؛ ذخیره‌سازی و تحلیل در مقیاس اینترنت، ویراست چهارم نوشتهٔ تام وایت مراجعه کنید. آنجا به‌تفصیل یاد می‌گیریم که چگونه یک مجموعه‌دادهٔ انتزاعی را در تعداد زیادی پردازنده افرازبندی کنیم. سپس نحوهٔ محاسبهٔ ماکسیمم همهٔ اعداد واقع در ستون واحدی از این مجموعه‌دادهٔ را فرا می‌گیریم. این مطلب متضمن محاسبهٔ ماکسیمم روی پایگاه دادهٔ فرعی واقع در هر پردازنده و به دنبال آن ترکیب ماکسیمم هر پردازنده در راستای همهٔ پردازنده‌های متعدد به منظور به دست آوردن یک ماکسیمم کلی است. گرچه تابعک مورد محاسبه در این مثال، بی‌اندازه ساده است، نیاز به مهارت‌های بسیار کمی برای اجرای مثال در این مقیاس وجود دارد.

چیزی که در هیاهوی مربوط به چنین مهارت‌هایی به فراموشی سپرده می‌شود، این واقعیت خجل‌کننده است که روزی‌روزگاری شخص می‌توانست چنین کارهای محاسباتی و حتی جاه‌طلبانه‌تر

را بسیار راحت‌تر از آنچه در این زمینه‌سازی‌های پرزرق‌برق نوظهور مقدور است، انجام دهد! یک مجموعه‌داده را می‌توان بر روی یک پردازندهٔ واحد قرار داد، و ماکسیمم فراموضعی آرایهٔ «x» را می‌توان با قطعهٔ کد شش کاراکتری «$\max(x)$» ،مثلاً با متلب یا آر محاسبه کرد. زمینه‌چینی و استفاده از وظیفه‌های جاه‌طلبانه‌تر، مانند بهینه‌سازی یک تابع محدب در مقیاس بزرگ، کاری بود آسان. در آن دوره‌های کمتر توأم با سروصداهای تبلیغاتی، کمتر نیاز به مهارت‌هایی که امروزه تبلیغ می‌شوند، وجود داشت. درعوض، دانشمندان مهارت‌هایی را برای حل مسئله‌ای که واقعاً به آن علاقه‌مند بودند، با استفاده از ریاضیاتی برازنده و محیط‌های برنامه‌نویسی کمّی قدرتمند که حول آن ریاضیات مدل‌بندی شده بود، به وجود آوردند. آن محیط‌ها نتیجهٔ ۵۰ سال یا بیشتر پالایش مستمر بودند که پیوسته به سمت آرمان توانمندسازی برگرداندن بی‌درنگ تفکر انتزاعی روشن به نتایج محاسباتی، نزدیک و نزدیک‌تر می‌شدند.

مهارت‌های جدیدی که توجه رسانه‌ها را این‌همه به خود جلب می‌کنند، مهارت‌هایی برای حل بهتر مسئلهٔ واقعی استنباط از داده‌ها نبوده، بلکه مهارت‌های جواب‌گویی برای مواجهه با ابزارهای محاسبات خوشه‌ای در مقیاس بزرگ سازمانی هستند. مهارت‌های جدید با محدودیت‌های شدید جدید روی الگوریتم‌ها که توسط دنیای چند پردازنده/شبکه‌ای تحمیل می‌شود، رویارویی می‌کنند. در این دنیای به‌شدت محدودیت‌آمیز، گسترهٔ الگوریتم‌های به‌راحتی قابل‌ساخت، در مقایسه با مدل تک‌پردازنده به‌طور چشمگیری کاهش می‌یابد، بنابراین به‌ناگزیر میل به آن پیدا می‌شود که آن رویکردهای استنباطی که در زمان‌های قدیم ابتدایی یا حتی نامناسب تلقی می‌شد، در پیش گرفته شود. چنین رویارویی‌هایی، وقت و انرژی ما را هدر می‌دهند، قضاوت ما را در مورد آنچه مناسب است، دچار دگردیسی می‌کنند و ما را از دنبال کردن راهبردهای تحلیل داده‌ای که در غیر این صورت مشتاقانه در پیش می‌گرفتیم، باز می‌دارند.

بااین‌حال، هلهله‌چی‌های مقیاس‌بندی، از عمق ریه بانگ بر می‌آورند که به دلیل استفاده از داده‌های بیشتر، حقشان است که برایشان هورا بکشیم.

۳.۲ «میم» شغل‌ها

شوروشوق حول داده‌های بزرگ از موفقیت‌های قابل‌توجهی که در دههٔ گذشته توسط شرکت‌های صاحب‌نام تجاری فناوری اطلاعات جهانی مانند گوگل و آمازون به ثمر رسیده است، تغذیه می‌کنند و این‌ها موفقیت‌هایی هستند که امروزه ازطرف سرمایه‌گذاران و مدیران عامل به رسمیت شناخته

شده‌اند. در پی این امر، یک «کوهانهٔ» استخدامی در طول ۵ سال گذشته ایجاد شده، که در آن تقاضاهای شغلی زیادی برای مهندسانی هم با مهارت در پایگاه داده و هم آمار به وجود آمده است. در فرهنگ داده‌های بزرگ [۱] مایک بارلو وضعیت را از قول گارتنر چنین جمع‌بندی می‌کند که

> تا سال ۲۰۱۴، ۴٫۴ میلیون شغل در داده‌های بزرگ ایجاد و فقط یک سوم آن‌ها پر خواهد شد. پیشگویی گارتنر صحنه‌هایی از «یورش برای طلا» را برای افراد با استعداد داده‌های بزرگ، همراه با لشگرهایی از کارشناسان تحلیل کمّی دوآتشه که مدارک تحصیلی پیشرفتهٔ خود را با موقعیت‌های استخدامی پرسود تاخت می‌زنند، در ذهن تداعی می‌کنند.

درحالی‌که به نظر بارلو هر مدرک کمّی پیشرفته در این محیط کافی خواهد بود، در ابتکارعمل‌های علم دادهٔ امروزی فی‌نفسه چنین وانمود می‌شد که مدارک آماری سنتی برای گیرآوردن شغل در این حوزه کافی نیستند ــ تأکید رسمی بر مهارت‌های محاسباتی و پایگاه داده باید بخشی از این آمیزه باشد.

ما واقعاً حقیقت ماجرا را نمی‌دانیم. جزوهٔ تحلیل تحلیل‌گرها: یک مطالعهٔ درون‌نگر از دانشمندان داده و کارِ آن‌ها [۲۳] خاطرنشان می‌کند که

> علی‌رغم جوش‌وخروشی که حول «علم داده»، «داده‌های بزرگ»، و «تحلیل‌گری» در جریان است، ابهام در این اصطلاحات منجر به ارتباط ضعیف بین دانشمندان داده و کسانی شده است که نیاز به کمک آن‌ها دارند.

وبلاگ یانیر سروسی بر این عقیده است که «موقعیت‌های شغلی واقعی کمی در علم داده برای افرادی که سابقهٔ کار ندارند، وجود دارد.»

> یک دانشمند دادهٔ موفق لازم است که توانایی آن را داشته باشد که با کاوش داده‌ها و به‌کارگیری تحلیل‌های آماری دقیق، با داده‌ها مانند دو جان در یک بدن باشد . . . اما دانشمندان خوب علم داده همچنین درک می‌کنند که برای به‌کارگیری مؤثر سیستم‌های تولید چه چیزی لازم است، و آمادگی آن را دارند که با نوشتن کدهایی که داده‌ها را پاک‌سازی یا در عملکرد اصلی سیستم مؤثرند، دست به کارِ گِل بزنند . . . به دست آوردن همهٔ این مهارت‌ها [ضمن کار] زمان می‌برد.

بارلو به‌طور ضمنی اشاره می‌کند که دانشمندان بالقوهٔ علم داده، ممکن است سال‌ها مواجه با نیاز به توسعهٔ مهارت‌های بیشتر پس‌از اخذ مدرک کارشناسی ارشد باشند، قبل‌از اینکه بتوانند چیز باارزشی بر داشته‌های سازمان محل کارِ خود بیفزایند. در یک سازمان دادهٔ بزرگ درحال فعالیت، زیرساختِ پردازش داده‌های تولید، دیگر استحکام یافته است. بعید به نظر می‌رسد که پایگاه‌های داده، نرم‌افزار، و گردش‌کار مدیریتی که در یک برنامهٔ کارشناسی ارشد علم داده تدریس می‌شود، همان باشد که توسط یک کارفرمای خاص به کار گرفته می‌شود. مصالحه‌ها و محدودیت‌های گوناگونی توسط سازمان‌های استخدام‌کننده انجام می‌شود و برای یک فرد تازه‌استخدام، تأثیر مثبت داشتن در این سازمان‌ها، یادگیری این مطلب است که چگونه با این محدودیت‌ها کنار بیاید و هنوز هم کاری انجام دهد.

مؤسسات مدرک‌دهندهٔ علم داده درواقع نمی‌دانند که چگونه ولع تقاضای ظاهری برای فارغ‌التحصیلان این رشته را برآورده کنند. همان‌طور که در زیر نشان می‌دهیم، نقش ویژهٔ [دارندهٔ] یک مدرک علم داده افزون بر یک مدرک آمار، دیدن آموزش اضافی فناوری اطلاعات است. بااین‌حال سازمان‌های استخدام‌کننده با مشکلاتی در استفاده از مهارت‌های خاص فنّاوری اطلاعات که در مقاطع تحصیلی آموزش داده می‌شود، مواجه هستند. درمقابل، تحلیل داده‌ها و آمار، مهارت‌هایی کاربردی هستند که قابلِ بردن از سازمانی به سازمان دیگر هستند.

۴.۲ چه چیزی در اینجا واقعی است؟

دیده‌ایم که رسانه‌های همگانی امروزی که در مورد علم داده صحبت می‌کنند، حتی تاب تحمل یک مداقهٔ بدوی را هم ندارند. این امر کاملاً قابل‌درک است: نویسندگان و مدیران دچار سراسیمگی شده‌اند. همه باور دارند که در امور انسانی با یک ناپیوستگی مرتبهٔ صفر روبه‌رو شده‌ایم.

اگر در سال ۲۰۱۰ یک کتاب راهنمای گردشگری را می‌خواندید، به شما گفته می‌شد که (مثلاً) تغییری در زندگی در روستاهای هند در طی هزاران سال رخ نداده است. اگر شما در سال ۲۰۱۵ به آن روستاها می‌رفتید، می‌دیدید که افراد بسیاری در آنجا حالا تلفن موبایل و بعضی‌ها گوشی‌های هوشمند دارند. این امر البته طلایه‌دار تغییری بنیادی محسوب می‌شود. چندان نخواهد گذشت که هشت میلیارد نفر به شبکه متصل، و بنابراین منابعی برای داده خواهند شد که مجموعهٔ گسترده‌ای از داده‌ها را در مورد فعالیت‌ها و ترجیحاتشان تولید خواهند کرد.

منتقل شدن به ارتباطات جهانی بسیار چشمگیر است؛ این کار درواقع، حجم عظیمی از داده‌های

تجاری تولید خواهد کرد. بهره‌برداری ار این داده‌ها مطمئناً یکی از دغدغه‌های اصلی دنیای تجارت در دهه‌های آینده است.

۵.۲ یک چارچوب بهتر

بااین‌حال، یک علم صرفاً به این دلیل به وجود نمی‌آید که سیل داده‌ها به همین زودی‌ها به یکباره سرورهای مخابراتی را پر خواهد کرد و یا به این دلیل که برخی از مدیران فکر می‌کنند که می‌توانند به روندهای استخدامی و تأمین منابع مالی دولتی متعاقب آن بو ببرند. خوشبختانه، شواهد قانع‌کننده‌ای به نفع خلق موجودیت مشخصی به نام «علم داده» وجود دارد که علمی واقعی خواهد بود: رویارویی با سؤال‌های اساسی دارای ماهیتی ماندگار و استفاده از تکنیک‌های به لحاظ علمی دقیق، برای مقابله با این سؤال‌ها.

حداقل ۵۰ سال است که آماردانان بصیر شالودهٔ لازم برای ساخت آن موجودیت بالقوه‌ای (would-be) را به‌صورت توسیعی از آمارسنتی دانشگاهی ریخته‌اند. این مفهوم بالقوه‌ای علم داده با آن علم داده که امروزه تبلیغ می‌شود، یکی نیست، اگرچه همپوشانی قابل‌توجهی بین آن‌ها وجود دارد. این مفهوم بالقوه‌ای به مجموعهٔ متفاوتی از گرایش‌های عاجل پاسخ می‌دهد ــ گرایش‌های فکری اما نه تجاری. مواجهه با این گرایش‌های فکری نیازمند بسیاری از همان مهارت‌هایی است که در رویارویی با گرایش‌های تجاری لازم است و به نظر می‌رسد که به همان اندازه احتمال دارد که با تقاضای آموزش دانشجویان آینده و تأمین مالی پژوهشی آینده مطابقت داشته باشد.

این مفهوم بالقوه‌ای، علم داده را به عنوان علم یادگیری از داده‌ها، با همهٔ تبعات آن در نظر می‌گیرد. این مفهوم با مهم‌ترین تحولات در علم که طی ۵۰ سال آینده به وجود خواهد آمد، مطابقت خواهد داشت، از این نظر که خود انتشارات علمی تبدیل به پیکره‌ای از داده‌ها می‌شود که می‌توانیم آن‌ها را تحلیل و مطالعه کنیم.

درک این مسائل فرصتی برای رؤسای دانشکده و دانشگاه فراهم می‌کند تا انرژی و شوروشوق در پسِ جنبش علم دادهٔ امروزی را مجدداً به سمت برنامه‌هایی ماندگار و عالی به منظور استانداردسازی یک رشتهٔ علمی جدید هدایت کنند.

من در این مقاله، بینش‌هایی را که در طول سال‌ها در مورد این رشتهٔ جدید بالقوه‌ای علم داده منتشر شده است، سروسازمان می‌بخشم و چارچوبی را برای درک پرسش‌های اساسی و رویه‌های آن ارائه می‌کنم. این چارچوب پیامدهایی، هم برای تدریس این مبحث و هم انجام تحقیقات علمی

دربارهٔ اینکه علم داده چگونه انجام می‌شود و چگونه می‌توان آن را بهبود بخشید، دارد.

۳ «آیندهٔ تحلیل داده‌ها»، ۱۹۶۲

این مقاله به‌عنوان یادداشتی برای عرضه در سدهٔ جان توکی تهیه شده بود. بیش از ۵۰ سال قبل، جان چنین از آینده خبر داده بود که برهه‌ای شبیه به علم دادهٔ امروزی فرا خواهد رسید. جان در مقالهٔ «آیندهٔ تحلیل داده‌ها» [۴۸] خوانندگان خود (آماردانان دانشگاهی) را در بندهای آغازین مقاله عمیقاً دچار بهت‌وحیرت کرد:

> مدت‌های مدیدی بر این تصور بودم که یک آماردانم و به استنتاج از جزء به کل علاقه دارم. اما در همان حینی که نظاره‌گر تحولات آمار ریاضی بودم، دلایلی برای حیرت و تردید یافتم. ... درکُل به این احساس رسیده‌ام که علاقهٔ اصلی من به تحلیل داده‌هاست، که به نظرم ازجملهٔ موارد دیگر، مشتمل است بر: شیوه‌هایی برای تحلیل داده‌ها، تکنیک‌هایی برای تفسیر نتایج چنین شیوه‌هایی، راه‌های برنامه‌ریزی برای جمع‌آوری داده‌ها به منظور آسان‌تر کردن ، بادقت‌تر کردن یا صحیح‌تر کردن تحلیل، و همهٔ ابزارآلات و نتایج آمار (ریاضی) که قابل‌اِعمال به تحلیل داده‌ها هستند.

مقالهٔ جان در سال ۱۹۶۲ در سالنامهٔ آمار ریاضی منتشر شد که محل اصلی انتشار پژوهش‌های آماری پیشرفتهٔ آن زمان بود. مقاله‌های دیگری که در آن زمان در این مجله منتشر می‌شدند، دقت ریاضی داشتند و به ارائهٔ تعریف‌ها، قضیه‌ها، و برهان‌ها می‌پرداختند. درمقابل، مقالهٔ جان نوعی اعتراف عمومی بود که توضیح می‌داد چرا از نظر او چنین تحقیقاتی بسیار باریک‌اندیش، و شاید بی‌فایده یا مضرند و گسترهٔ پژوهشی آمار نیاز به آن داشت که به طور چشمگیری پهناورتر شده و تغییر مسیر دهد.

پیتر هیوبر، که قرار بود کارهای علمی پیشگام او دربارهٔ برآورد استوار در همان مجله منتشر شود، اخیراً دربارهٔ «آیندهٔ تحلیل داده‌ها» چنین نظر داده است:

> نیم‌قرن پیش، توکی در مقاله‌ای با غایی‌ترین حد تأثیرگذاری، تعریف دوباره‌ای از مبحث ما به عمل آورد . . .[آن مقاله] اصطلاح «تحلیل داده» را به عنوان نامی برای آنچه آماردانان کاربردی انجام می‌دهند، معرفی کرد و تمایزی بین این اصطلاح با استنباط آماری صوری قائل شد. اما درواقع، به اذعان توکی، او «این اصطلاح را

از حوزهٔ زبانی آن فراتر برد» در آن حد که کلیت آمار را در بر بگیرد. (پیتر هیوبر، ۲۰۱۰)

بنابراین، چشم‌انداز توکی، آمار را در درون یک موجودیت بزرگ‌تر نشانید. ادعای اصلی توکی این بود که این موجودیت جدید، که او آن را «تحلیل داده» نامید یک علم جدید بود و نه شاخه‌ای از ریاضیات:

> دیدگاه‌های متفاوتی دربارهٔ اینکه عناصر سازندهٔ یک علم چیست، وجود دارد، اما اکثریت افراد، سه مؤلفه را برای آن، اساسی تشخیص می‌دهند، که عبارت‌اند از: (آ۱) محتوای فکری، (آ۲) سازمان، در شکلی قابل‌فهم، (آ۳) اتکا بر آزمونِ تجربه به‌عنوان استاندارد غایی معتبربودن.
>
> طبق این آزمون‌ها، ریاضیات علم نیست، زیرا استاندارد غایی برای معتبربودن آن، نوعی سازگاری منطقی و اثبات‌پذیریِ مطابق توافق است. از دیدگاه من، تحلیل داده‌ها، از این هر سه آزمون با موفقیت بیرون می‌آید و من آن را به عنوان یک علم تلقی می‌کنم، علمی که با یک مسئلهٔ همه‌جا حاضر تعریف می‌شود و نه یک موضوع ملموس. در این صورت لازم است که تحلیل داده‌ها و بخش‌هایی از آمار که به این قاعده التزام دارند، ویژگی‌های یک علم را به خود بگیرند و نه ویژگی‌های یک علم ریاضی را، . . .

این نکات باید جدی گرفته شوند.

توکی چهار نیروی محرکه را در این علم جدید شناسایی کرد:

> امروزه چهار عامل عمده بر تحلیل داده‌ها تأثیر گذاری دارند: ۱) نظریه‌های صوری آمار، ۲) پیشرفت‌های پرشتاب در کامپیوترها و دستگاه‌های نمایشگر، ۳) چالش، در بسیاری از حوزه‌ها، با پیکره‌هایی روزافزون از داده‌ها، ۴) تاکید بر کمّی‌سازی در طیف هرچه گسترده‌تری از رشته‌ها.

فهرست جان در سال ۱۹۶۲ به‌طرز شگفت‌آوری مدرن است و مشتمل‌بر همهٔ عواملی است که امروزه در بیانیه‌های مطبوعاتی‌ای که مُبلِّغِ ابتکارعمل‌های مربوط به علم دادهٔ امروزی هستند، به رخ کشیده می‌شود. آنچه در آن زمان تکان‌دهنده بود، مورد ۱) بود، که به‌طور ضمنی حاکی از این بود که نظریهٔ آمار تنها قسمتی (جزئی!) از این علم جدید است.

این علم جدید با علوم مستقر مقایسه می‌شود و در حد بیشتری نقش آمار را در محدودهٔ خود فرامی‌گیرد:

> . . . تحلیل داده‌ها حوزهٔ بسیار دشواری است. باید خود را با هر آنچه مردم قادرند و ضرورت دارد که با داده‌ها انجام دهند، سازگار کند. در این معنا که زیست‌شناسی پیچیده‌تر از فیزیک است، و علوم رفتاری پیچیده‌تر از هر دوی این‌هاست، این احتمال وجود دارد که مسائل کلی تحلیل داده‌ها پیچیده‌تر از هر سهٔ این‌ها باشد. انتظارِ اینکه دستورالعمل نزدیک به واقع و اثربخشی برای تحلیل داده‌ها با هر نوع ساختار بسیار رسمی، چه در حال حاضر و چه در آینده‌ای نزدیک داشته باشیم، توقع زیادی است. تحلیل داده‌ها می‌تواند از آمار صوری بسیار سود ببرد، اما فقط درصورتی که این ارتباط به اندازهٔ کافی سست نگه داشته شود.

بنابراین، تحلیل داده‌ها نه‌تنها یک رشتهٔ علمی است، بلکه به اندازهٔ هر رشتهٔ علمی عمده پیچیده است! و آمار نظری تنها می‌تواند نقشی جزئی در پیشرفت آن داشته باشد. عنوان مقالهٔ [۳۶] تکراری از این نکته است: «تحلیل داده‌ها، دربرگیرندهٔ آمار.»

۴ ۵۰ سال پس‌از «آیندهٔ تحلیل داده‌ها»

با آنکه توکی دعوت به [ایجاد] رشته‌ای بسیار گسترده‌تر از آمار شد، این کار ــ حتی به عنوان مجموعهٔ کل کار علمی یک فرد ــ یک‌شبه مقدور نبود.

هیوبر نوشت که «مقالهٔ توکی تأثیری فوری به جا نگذاشت. . . . چندین سال طول کشید تا اینکه اهمیت آن برایم جاافتاد.» (هیوبر، ۲۰۱۰). با ملاحظهٔ کار پیتر از نزدیک، می‌توان گفت که ۱۵ سال پس‌از مقالهٔ «آیندهٔ تحلیل داده‌ها»، او به‌وضوح با درس‌های آن کنار آمده است. درعین‌حال، شواهد کاملی از این تأثیر در مورد هیوبر حتی خیلی دیرتر اتفاق افتاده ــ کافی است به کتاب سال ۲۰۱۰ او، تحلیل داده‌ها: آنچه می‌توان از ۵۰ سال گذشته آموخت مراجعه شود که جمع‌بندی‌ای از نوشته‌های پیتر از دههٔ ۱۹۸۰ است و ۴۸ سال بعداز آن مقاله عرضه شده است.

۱.۴ اندرزها

در همان حال که هیوبر به‌وضوح به این انتخاب دست زد که به کاوش در چشم‌اندازهای دیدگاه توکی بپردازد، جامعهٔ دانشگاهی آمار در کلیّت خود چنین کاری نکرد. همکاران جان توکی در آزمایشگاه

بل، که در گروه‌های آمار دانشگاه‌ها مستقر نبودند، به‌راحتی دیدگاه جان را در مورد حوزه‌ای پهناورتر از آنچه آمار دانشگاهی می‌تواند ارائه دهد، پذیرفتند.

جان چمبرز، یکی از همکاران ابداع‌کنندهٔ زبان S برای آمار و تحلیل داده‌ها در آن هنگام که در آزمایشگاه بل بود، مقاله‌ای در سال ۱۹۹۳ با عنوان تحریک‌آمیز «آمار مهتر یا کهتر، انتخابی برای تحقیقات آینده» منتشر کرد [۷]. چکیدهٔ عاری از پرده‌پوشی او به شرح زیر بود:

> حرفهٔ آمار در تحقیقات آیندهٔ خود با انتخابی بین ادامهٔ تمرکز بر موضوعات سنتی ــ عمدتاً مبتنی‌بر تحلیل داده‌ها به پشتیبانی آمار ریاضی ــ و دیدگاهی گسترده‌تر ــ مبتنی‌بر یک مفهوم فراگیر یادگیری از داده‌ها ــ روبه‌روست. مسیر دوم، چالش‌هایی جدی و نیز فرصت‌هایی هیجان‌انگیز به ارمغان می‌آورد. مسیر اول، خطر روزافزون به حاشیه رانده شدن آمار را دارد.

این فراخوان برای اقدام از سوی آماردانی است که احساس می‌کند «قطار در حال راه افتادن از ایستگاه است.» مانند مقالهٔ توکی، در این مقاله چنین پیشنهاد می‌شود که ما می‌توانیم به دنبال تحقیقاتی باشیم که گستره‌ای بسیار وسیع‌تر از آن تحقیقات آماری را که امروزه انجام می‌دهیم، در بر بگیرد؛ چنین تحقیقاتی بر روی فرصت‌های فراهم‌شده توسط انواع جدید داده‌ها و انواع جدید نمایش‌ها متمرکز خواهد شد. چمبرز به‌صراحت اعلام کرد که این حوزهٔ وسعت‌یافته حتی از حوزهٔ تحلیل داده‌ها نیز بزرگ‌تر، و به‌طور مشخص نسبت به دیدگاه توکی در سال ۱۹۶۲، گسترده‌تر خواهد بود.

جف وو در مراسم معارفه به مناسبت انتخاب به عنوان استاد کارورِ آمارِ دانشگاه میشیگان، خطابه‌ای افتتاحیه‌ای با عنوان «آمار = علم داده؟» ایراد کرد که در آن از این ایده هواداری می‌کرد که آمار، علم داده و آماردانان، دانشمندان داده نامیده شوند. او با پیش‌اندیشی درس‌های ارشد دورهٔ مدرن کارشناسی ارشد علم داده، حتی به ایدهٔ یک مدرک کارشناسی ارشد جدید اشاره کرد که در آن حدود نیمی از درس‌ها خارج از گروه آمار گرفته می‌شدند. او کار آماری را سه‌گانه‌ای از جمع‌آوری داده‌ها، مدل‌بندی و تحلیل داده‌ها، و تصمیم‌سازی توصیف کرد. هیچ مقالهٔ مکتوب رسمی از این سخنرانی تهیه نشد، گرچه اسلایدهایی که ارائه کرد، در دسترس هستند.

کلیولند، هنگام کار در آزمایشگاه‌های بل، روش‌های آماری و نمایش‌های داده‌ای ارزشمندی به وجود آورد و به عنوان یکی از ویراستاران مجموعهٔ آثار توکی همکاری کرد. مقالهٔ او مربوط به سال

۲۰۰۱ [۱۱]، با عنوان علم داده: یک برنامهٔ عمل برای گسترش عرصه‌های فنی رشتهٔ آمار، در خطاب به گروه‌های آمار دانشگاهی بود و طرحی برای تغییر مسیر کار آن‌ها پیشنهاد کرد. چکیدهٔ او به شرح زیر است: تمرکز یک برنامهٔ عمل برای گسترش دادن عرصه‌های فنی آمار، بر روی تحلیلگر داده است. این برنامه، کار بر روی شش حوزهٔ فنی را برای یک گروه دانشگاهی در نظر می‌گیرد و از تخصیص منابع ویژه برای تحقیق و تدریس دروس در هر حوزه پشتیبانی می‌کند. ارزش کار فنی بر اساس میزانی که به تحلیلگر داده سود می‌رساند، چه به‌طور مستقیم و چه غیرمستقیم، مورد ارزیابی قرار می‌گیرد. این برنامه همچنین در مورد آزمایشگاه‌های تحقیقاتی دولتی و سازمان‌های تحقیقاتی شرکتی، قابل اجراست.

کلیولند در مقدمهٔ مقاله می‌نویسد:

> . . . [نتایج در] علم داده باید براساس میزانی که تحلیلگر را قادر می‌سازند تا از داده‌ها یاد بگیرد، ارزیابی شود . . . ابزارهایی که توسط تحلیلگر داده‌ها مورد استفاده قرار می‌گیرند، سودِ مستقیم دارند. نظریه‌هایی که به‌عنوان پایه‌ای برای ایجاد ابزارها به کار گرفته می‌شوند، سودِ غیرمستقیم دارند.

کلیولند شش کانون فعالیت را مطرح و حتی اختصاص تلاش‌هایی را به آن‌ها پیشنهاد کرد: تحقیقات چندرشته‌ای (۲۵٪)؛ مدل‌ها و روش‌ها برای داده‌ها (۲۰٪)؛ محاسبات با داده‌ها (۱۵٪)؛ آموزش (۱۵٪)؛ ارزیابی ابزار (۵٪) نظریه (۲۰٪).

چندین گروه آمار در دانشگاه‌ها، که من به‌خوبی آن‌ها را می‌شناسم، در زمان انتشار مقالهٔ کلیولند، ۱۰٪ فعالیت خود را در ۲۰٪ مجاز کلیولند برای نظریه، اختصاص داده بودند. مقالهٔ کلیولند در سال ۲۰۱۴ بازنشر شد. من نمی‌توانم حتی یک گروه دانشگاهی را به تصور در بیاورم که امروزه ۱۵٪ از تلاش خود را صرف آموزش، یا ۱۵٪ را صرف محاسبات با داده‌ها کرده باشد. می‌توانم چندین گروه آمار از دانشگاه‌ها را به ذهن بیاورم که اساساً همهٔ فعالیت خود را در آخرین مقوله، یعنی نظریه، گنجانده‌اند. به‌طور خلاصه، آماردانان دانشگاهی بارهاوبارها در طول سال‌ها توسط جان توکی و برخی از همکارانش در آزمایشگاه‌های بل، و حتی توسط برخی از دانشگاهیان مانند پیتر هیوبر و جف وو، برای تغییر مسیر، به سمت تعریف بسیار گسترده‌تری از رشتهٔ خود اندرز داده شده‌اند. چنین توصیه‌هایی قبل از سال ۲۰۰۰ تأثیر ظاهری نسبتاً کمی داشته‌اند.

۲.۴ جان‌بخشی

یکی از موانعی که اولین اندرزها با آن روبه‌رو بودند این بود که مخاطبان بسیاری از اندرزدهندگان

نمی‌توانستند بفهمند که این قیل‌وقال برای چیست. ایجاد انگیزه برای ملموس‌تر ساختن و در معرض دید درآوردن فعالیتی که عنوان «تحلیل داده‌ها» داشت، در نهایت با کدنویسی میسر شد و نه با استفاده از کلمات.

در طول ۵۰ سال گذشته، بسیاری از آماردانان و تحلیلگران داده در ابداع و توسعهٔ محیط‌های محاسباتی برای تحلیل داده‌ها مشارکت کردند. چنین محیط‌هایی شامل بسته‌های آماری اولیه، مانند BMDP، SPSS، SAS، و مینی‌تب بودند، که همه ریشه در محاسبات با مین‌فریم در اواخر دههٔ ۱۹۶۰ داشتند، و در سال‌های آتی، بسته‌هایی مانند S، ISP، STATA ، و R، که ریشه در دوران مینی‌کامپیوترها/کامپیوترهای شخصی داشتند. این امر، تلاش عظیمی بود که توسط افراد بااستعداد بسیاری انجام شد ــ بسیار بیشتر از آنچه بتوان به‌طرز مناسبی از ابداع‌کنندگان آن‌ها به‌شایستگی یاد کرد.

برای تعیین میزان اهمیت این بسته‌ها، سعی کنید از نظارت‌گر N–نگارهای گوگل برای ترسیم فراوانی کلمات BMDP، SPSS، SAS، در کتاب‌های به زبان انگلیسی از۱۹۷۰ تا ۲۰۰۰ استفاده کنید؛ و برای مقایسه، فراوانی دونِگارهای تحلیل داده‌ها و «تحلیل آماری» را نیز رسم کنید. معلوم می‌شود که اصطلاحات SPSS، SAS در زبان انگلیسی در این دوره هر دو نسبت به «تحلیل داده‌ها» یا «تحلیل آماری» رایج‌تر ــ درواقع تقریباً دو برابر ــ رایج‌ترند.

جان چمبرز و همکارش ریک بِکِر در آزمایشگاه بل، محیط محاسباتی کمّی S را با شروع از اواسط دههٔ ۱۹۷۰ توسعه دادند؛ این محیط، زبانی برای توصیف محاسبات و بسیاری از ابزارهای آماری و دیداری‌سازی پایه‌ای را فراهم آورد. در دههٔ ۱۹۹۰، جنتلمن و ایهاکا سیستم مشابه R را به صورت یک پروژهٔ متن‌باز ابداع کردند که به‌سرعت گسترش یافت. امروزه R، محیط برنامه‌نویسی کمّی غالبِ مورد استفاده در آمار دانشگاهی، با دنبالگرهای برخط بسیار چشمگیری است.

محیط‌های برنامه‌نویسی کمّی، «اسکریپت‌ها» را اجرا می‌کنند که دقیقاً مراحل یک محاسبات را کدنویسی کرده، آن‌ها را در سطحی بسیار فراتر و انتزاعی‌تر از سطح زبان‌های کامپیوتری سنتی مانند C++ توصیف می‌کنند. چنین اسکریپت‌هایی امروزه اغلب گردش‌کار نامیده می‌شوند. زمانی که یک QEP در بخشی از جامعهٔ تحقیقاتی، همان‌طور که R در آمار دانشگاهی جنبهٔ غالب پیدا می‌کند گردش‌کار را می‌توان به‌طور گسترده در این جامعه به اشتراک گذاشت و به بازاجرای آن پرداخت، که این کار بر روی داده‌های اصلی (درصورتی که به اشتراک گذاشته شده باشد) یا به روی داده‌های جدید، انجام می‌شود. این، یک عامل تحول‌بخش است. آنچه قبلاً تا حدودی ابهام‌آلود بود ــ مثلاً توصیف

قسمت‌هایی از تحلیل داده‌ها به بیان نثر در یک مقالهٔ علمی ــ برعکس به چیزی ملموس و سودمند بدل می‌شود، زیرا می‌توان بلافاصله کد را دانلود و اجرا کرد. همچنین می‌توان به‌راحتی اسکریپت‌ها را دستکاری کرد تا ظرافت‌های داده‌های شخص را، برای مثال، با تغییر یک برآوردگر ماتریس کوواریانس استاندارد در اسکریپت اصلی به یک برآوردگر ماتریس کوواریانس استوار، منعکس کند. می‌توان بهبودهای عملکرد ناشی از ایجاد تغییرات در اسکریپت پایه را مستندسازی کرد. حالا منطقی است که بتوان از یک رویکرد علمی برای بهبود تحلیل داده‌ها، با اندازه‌گیری نحوهٔ عملکرد و به دنبال آن دستکاری اسکریپت، صحبت کرد. ادعای توکی مبنی‌بر اینکه مطالعهٔ تحلیل داده‌ها می‌تواند یک علم تلقی شود، حالا خودبه‌خود بدیهی می‌شود. هرکس ممکن است با فراخوان‌های چمبرز و کلیولند برای دست به کار شدن، موافق یا مخالف باشد؛ اما این امر ممکنی بود که تا سال ۲۰۰۱ ، همه با کلیولند، در این مطلب توافق داشته باشند که حالا این امکان به وجود آمده است که زمینه‌ای «با عنوان علم داده به منصهٔ ظهور درآمده باشد.»

۵ مقالهٔ «دو فرهنگ» برایمن

لئو برایمن، آماردان دانشگاه کالیفرنیا در برکلی، که پس از سال‌ها کار به‌عنوان مشاور آماری در طیف وسیعی از سازمان‌ها ازجمله سازمان حفاظت محیط زیست، دوباره وارد خدمات دانشگاهی شد، سررشتهٔ جدید و مهمی را طی مقاله‌اش در مجلهٔ علم آماری [۴] وارد بحث کرد. برایمن در این مقاله با عنوان «مدل‌بندی آماری: دو فرهنگ»، دو دیدگاه را توصیف کرد که از داده‌ها ارزش می‌آفرینند.

> آغاز آمار با داده‌هاست. داده‌ها را این‌گونه در نظر بیاورید که توسط یک جعبهٔ سیاه تولید می‌شوند که در آن بردار متغیرهای ورودی x (متغیرهای مستقل) از یک طرف وارد می‌شوند و متغیرهای پاسخ y از طرف دیگر بیرون می‌آیند. در داخل جعبهٔ سیاه، طبیعت در حال عمل است تا متغیرهای پیشگو را با متغیرهای پاسخ پیوند دهد.
>
> در تحلیل داده‌ها دو هدف وجود دارد:
>
> - پیشگویی. قادر بودن به پیشگویی اینکه پاسخ‌ها به متغیرهای ورودی آتی، چه خواهند بود؛
> - استنباط. برای [استنباط کردن] اینکه طبیعت چگونه متغیرهای پاسخ را با متغیرهای ورودی پیوند می‌دهد.

برایمن می‌گوید که کاربران داده‌ها بر مبنای پیروی اولیه‌شان از یکی از این اهداف، در یکی از این دو شاخهٔ فرهنگ قرار می‌گیرند.

فرهنگ «مدل‌بندی مولد» در صدد ایجاد مدل‌هایی تصادفی است که به داده‌ها برازش می‌یابند و سپس انجام استنباط‌هایی دربارهٔ سازوکار تولیدکنندهٔ داده‌ها براساس ساختار آن مدل‌هاست. چیزی که در دیدگاه آن‌ها به‌طور ضمنی مطرح است، این تصور است که یک مدل واقعی که داده‌ها را تولید می‌کند و در بیشتر موارد، یک واقعاً «بهترین» راه برای تحلیل داده‌ها وجود دارد. برایمن فکر می‌کرد که این فرهنگ ۹۸ درصد کل آماردانان دانشگاهی را دربرمی‌گیرد.

فرهنگ «مدل‌بندی پیشگوگر» پیشگویی را در اولویت قرار می‌دهد و با برآورد برایمن، ۲٪ از آماردانان دانشگاهی ــ ازجمله خود برایمن ــ و اما همچنین بسیاری از متخصصان کامپیوتر، و همان‌طور که بحث مقالهٔ او نشان می‌دهد، آماردانان مهم شاغل در صنعت را شامل می‌شود. مدل‌بندی پیشگوگر در مورد سازوکار زمینه‌ای تولیدکنندهٔ داده‌ها عملاً ساکت است، و الگوریتم‌های پیشگوگر متعددی را مجاز دانسته، ترجیح می‌دهد که فقط صحت پیشگویی‌های انجام‌شده توسط الگوریتم‌های مختلف روی مجموعه‌داده‌های مختلف را مورد بحث قرار دهد. برایمن رشتهٔ نسبتاً جدید یادگیری ماشین را، که اغلب در گروه‌های علوم کامپیوتر مستقر است، به عنوان مرکز فرهنگ مدل‌بندی پیشگوگر می‌شناسد.

بخشی از چکیدهٔ برایمن به شرح زیر است:

> جامعهٔ آماری تقریباً به‌طور انحصاری خود را متعهد به استفاده از مدل‌ها [ی مولد] کرده است. این تعهد منجر به نظریه‌های نامربوط و نتیجه‌گیری‌های سؤال‌برانگیز شده، و آماردان‌ها را از کار در زمینهٔ طیف وسیعی از مسائل جالبِ جاری باز داشته است. این مدل‌بندی [پیشگوگر]، چه در نظریه و چه در عمل‌ورزی، به‌سرعت در زمینه‌های بیرون از آمار توسعه یافته است. می‌توان از آن هم در مجموعهٔ داده‌های پیچیدهٔ بزرگ و هم به‌عنوان یک جایگزین دقیق‌تر و آگاهی‌بخش‌تر برای مدل‌بندی داده‌ها در مجموعه‌داده‌های کوچک‌تر استفاده کرد. اگر هدف ما به‌عنوان یک رشتهٔ علمی، استفاده از داده‌ها برای حل مسائل باشد، در این صورت باید از وابستگی انحصاری به مدل‌ها [ی مولد] فاصله گرفت . . ..

باز هم، رشتهٔ آمار فراخوانده شده است تا بر گستردگی خود بیفزاید.

در مباحثهٔ مقالهٔ برایمن، دو آماردان برجسته، سر دیوید کاکس از دانشگاه آکسفورد و بردلی افرون از دانشگاه استنفورد، هر دو به طرق مختلف بر این مؤکدسازی برایمن ایراد گرفتند.

- کاکس اظهار داشت که از نظر او، «موفقیت پیشگوگر مبنای اصلی برای انتخاب مدل نیست» و اینکه «روش‌های صوری انتخاب مدل که در آنها اهداف گسترده‌تر در نظر گرفته نمی‌شوند، سؤال‌برانگیزند . . . .»
- افرون بیان داشت که «پیشگویی، مطمئناً موضوع جالبی است اما مقالهٔ لئو هم در مورد نقش آن و هم در مورد عدم علاقه حرفهٔ ما به آن دچار اغراق شده است.»

در همین مباحثه، بروس هوُدلی ــ آماردانی از شرکت امتیازدهی اعتباری فیر، آیزک بااشتیاق در مورد اظهارات برایمن به بحث می‌پردازد:

> مقالهٔ پروفسور برایمن، واجد اهمیت زیادی برای آماردانان است. باید او و مجلهٔ علم آماری را تحسین کرد . . . نتیجه‌گیری‌های او اغلب با نحوهٔ به کار بسته شدن آمار در کسب‌وکار سازگاری دارد.

کسب‌وکار اصلی فیر، آیزک، پشتیبانی از میلیاردها تراکنش روزانهٔ کارت‌های اعتباری، با صدور بی‌درنگ (چیزی معادل) پیشگویی‌هایی است مبنی‌بر اینکه آیا تراکنش مورد درخواست، بازپرداخت خواهد شد یا نخواهد شد. فیر، آیزک نه‌تنها مدل‌های پیشگوگر ایجاد می‌کند، بلکه باید از آن‌ها در زمینهٔ کسب‌وکار اصلی خود استفاده کند و باید آن‌ها نسبت به توجیه دقت خود برای بانک‌ها، شرکت‌های کارت اعتباری و نهادهای نظارتی اقدام کنند. مربوط بودن فرهنگ پیشگوگر برایمن به کسب‌وکار آن‌ها آشکار و در ارتباط مستقیم است.

۶ چاشنی پنهان فرهنگ پیشگوگر

برایمن حق داشت که به آماردانان برای درک بهتر فرهنگ مدل‌بندی پیشگوگر اندرز دهد، اما مقالهٔ او «چاشنی پنهان» این فرهنگ را به‌وضوح آشکار نکرد.

۱.۶ چارچوب وظیفهٔ مشترک

به نظر من، روش‌شناسی پراهمیت اما دریافته‌نشدهٔ عامل محرک موفقیت مدل‌بندی پیشگوگر، همان چیزی است که زبان‌شناس محاسباتی، مارک لیبرمن [۳۲] آن را چارچوب وظیفهٔ مشترک (CTF) نامیده است. نمونه‌ای از چارچوب وظیفهٔ مشترک، شامل اجزای زیر است:

الف) یک مجموعه‌دادهٔ آموزشی در دسترس عموم، برای هر مشاهده، شامل فهرستی از (احتمالاً بسیاری از) اندازه‌گیری‌های ویژگی، و یک برچسب رده برای آن مشاهده باشد.

ب) مجموعه‌ای از رقبای ثبت‌نام‌کرده که وظیفهٔ مشترک آن‌ها استنباط یک قاعدهٔ پیشگویی رده از داده‌های آموزشی است.

پ) یک داور امتیازدهنده، که رقبای شرکت‌کننده می‌توانند قاعدهٔ پیشگویی خود را به او ارسال کنند. داور، قاعدهٔ پیشگویی را در برابر یک مجموعه‌دادهٔ آزمایشی، که به‌شدت از همه پنهان نگاه داشته می‌شود، اجرا می‌کند. داور به‌صورت عینی و خودکار، امتیاز (درستی پیشگویی) حاصل از قاعدهٔ ارائه‌شده توسط همهٔ رقبا را گزارش می‌کند.

همهٔ رقبا در وظیفهٔ مشترک آموزش یک قاعدهٔ پیشگویی که امتیاز خوبی دریافت خواهد کرد، تشریک مساعی می‌کنند و عنوان چارچوب وظیفهٔ مشترک به همین دلیل است.

یک مثال معروف مربوط به این اواخر، چالش نتفلیکس است که وظیفهٔ مشترک در آن، پیشگویی انتخاب‌های فیلم کاربران نتفلیکس بود. تیم برنده (شامل باب بل آماردان ATT) یک میلیون دلار برنده شد. مجموعه‌داده از داده‌های سوابق مشتریان اختصاصی نتفلیکس استفاده می‌کرد. بااین‌حال، نمونه‌های متعدد دیگری، اغلب (به‌طورضمنی) با پاداش‌های بسیار بیشتر در بین است.

۲.۶ تجربه با CTF

پیدایش الگوارهٔ چارچوب وظیفهٔ مشترک ارتباط جالبی با داستان ما دارد. در روایت مارک لیبرمن، شروع این موضوع با پیرس، یکی از همکاران توکی در آزمایشگاه بل بود. پیرس واژهٔ «ترانزیستور» را ابداع و بر توسعهٔ اولین ماهوارهٔ مخابراتی نظارت، و همراه با توکی در اوایل/اواسط دههٔ ۱۹۶۰ در کمیتهٔ مشورتی علم منصوب ریاست جمهوری خدمت کرده بود. در همان زمان که توکی در حال ارزیابی مشکلات در حال ظهور ناشی از استفادهٔ بیش از حد از آفت‌کش‌ها بود، از پیرس خواسته شده بود تا در مورد سرمایه‌گذاری گستردهٔ از مدت‌ها پیش در حال انجام در تحقیقات ترجمهٔ ماشینی، ارزیابی‌ای انجام دهد. به همان ترتیب که توکی آنچه را که عنوان تحقیقات آماری در دههٔ ۱۹۶۰در جریان بود، دوست نداشت، پیرس هم آنچه را که به عنوان تحقیقات ترجمهٔ ماشینی دههٔ ۱۹۶۰ در حال انجام بود، نمی‌پسندید.

حال نظرات مارک لیبرمن را به‌دقت بررسی می‌کنیم. پیرس با این نگاه که این حوزه سرشار از مستعد بودن به «اغواگری و فریب» است، موفق شد کل تلاش‌های تحقیقاتی ترجمهٔ ماشینی ایالات

متحده را فلج کند و آن را برای چندین دهه عملاً به صفر بازگرداند.

به‌عنوان نمونه‌هایی از اغواگری و فریب، پیرس به آن رویکردهای نظری در ترجمه انگشت گذاشت که منشأ آن‌ها، به‌عنوان‌مثال، از به‌اصطلاح نظریه‌های زبان چامسکی بود؛ درحالی‌که هیبت چنین نظریه‌هایی ظاهراً بسیاری از پژوهشگران زبان را در آن زمان اسیر خود کرده بود، پیرس آن پژوهشگران را فریب‌خوردهٔ اغواگری یک نظریه (بالقوه‌ای) می‌دید، و نه عملکرد واقعی در ترجمه.

تحقیقات در ترجمهٔ ماشینی سرانجام چندین دهه بعداز برزخ پیرسی دوباره سر بر آورد، اما تنها به این دلیل که راهی برای اجتناب از مستعد بودن به اتهامات پیرس مبنی‌بر اغواگری و فریب پیدا کرد. یک تیم تحقیقاتی در پردازش گفتار و زبان طبیعی در IBM، که شامل نابغه‌هایی واقعی مانند جان کوک و همچنین دانشمندان دادهٔ پیشانام‌گذاری امثال لالیت بال، پیتر براون، استفن، و وینسنت دلا پی‌یترا، و رابرت مرسر بود، شروع به پیشرفت قطعی به سمت ترجمهٔ ماشینی برپایهٔ استفاده از شکل اولیه‌ای از کاربرد چارچوب وظیفه مشترک کردند. یک منبع بسیار مهم برای این کار، داده‌ها بودند: آن‌ها یک نسخهٔ دیجیتالی از به‌اصطلاح هنساردهای کانادایی، مجموعه‌ای از اسناد دولتی که به دو زبان انگلیسی و فرانسوی ترجمه شده بودند، به دست آورده بودند. تا پیش از اواخر دههٔ ۱۹۸۰، دارپا متقاعد شد که CTF را به‌عنوان الگووارهٔ جدیدی برای تحقیقات ترجمهٔ ماشینی بپذیرد. با نیّتِ تولید داده‌های تحصیل‌شده و انجام داوری قرارداد بسته شد و دارپا تیم‌هایی از پژوهشگران را به چالش گرفت تا قواعدی برای رده‌بندی درست تحت CTF تولید کنند.

گونه‌های دیگری از CTF تاکنون توسط دارپا در مورد بسیاری از مسائل، توأم با موفقیت به کار برده شده است: ترجمهٔ ماشینی، شناسایی سخن‌گوینده، تشخیص اثر انگشت، بازیابی اطلاعات، OCR، شناسایی خودکار هدف، و غیره.

تجربهٔ کلی با CTF که توسط لیبرمن تلخیص شده، به شرح زیر است.

۱) کاهش نرخ خطا با درصد ثابتی در هر سال به یک مقدار مجانبی بسته به وظیفه و کیفیت داده‌ها.

۲) پیشرفت معمولاً از بهبودبخشی‌های کوچک بسیار ناشی می‌شود؛ یک تغییر یک‌درصدی می‌تواند دلیلی برای جشن گرفتن باشد.

۳) داده‌های تحت اشتراک‌گذاری، نقش مهمی ایفا می‌کنند ــ و به راه‌های غیرمنتظره‌ای مجددا مورد استفاده قرار می‌گیرند.

موفقیت نهایی بسیاری از فرایندهای خودکار که ما امروزه آن‌ها را امری بدیهی می‌انگاریم ــ

مترجمِ گوگل، شناسهٔ لمسی گوشی‌های هوشمند، تشخیص صدای گوشی‌های هوشمند ــ برگرفته از الگوارهٔ پژوهشی CTF، یا به‌طور مشخص‌تر اثر تجمعی آن پس‌از عمل کردن به آن برای چندین دهه در زمینه‌های خاص است. مهم‌ترین چیز برای داستان ما: آن زمینه‌هایی که یادگیری ماشین در آن‌ها موفقیت‌هایی کسب کرده است، اساساً آن زمینه‌هایی هستند که CFT به‌طور قاعده‌مند در آن‌ها به کار برده شده است.

۳.۶ چاشنی پنهان

اغراق نیست اگر بگوییم که ترکیبی از یک فرهنگ مدل‌بندی پیشگوگر همراه با CTF، «چاشنی پنهان» یادگیری ماشین است. هم‌افزایی به حداقل رساندن خطای پیشگویی به کمک CTF، ارزش چندانی ندارد. این ترکیب مستقیماً به تمرکز کامل بر بهینه‌سازی عملکرد تجربی منجر می‌شود که، همان‌گونه که مارک لیبرمن خاطر نشان کرده است، به تعداد کثیری از محققان اجازه می‌دهد در هر چالش وظیفهٔ مشترک مشخص رقابت کنند و اجازه داوری کارآمد و غیراحساسی برندگان چالش را می‌دهد. این کار همچنین بلافاصله منجر به کاربردهایی در یک کاربرد دنیای واقعی می‌شود. در فرایند برنده شدن در یک رقابت، یک قاعدهٔ پیشگویی لزوماً به آزمون گذاشته می‌شود، و بنابراین اساساً آمادهٔ به کار گرفته شدن در عمل فوری است

بسیاری از «افراد بیرونی» از ماهیت الگوارهٔ CTF و نقش اصلی آن در بسیاری از موفقیت‌های یادگیری ماشین آگاه نیستند. افراد بیرونی، ممکن است نام چالش نتفلیکس را شنیده باشند، بدون اینکه نقش CTF را در آن چالش درک کرده باشند. آن‌ها ممکن است متوجه شده باشند که «یادگیری عمیق» به یک موضوع سرِزبان‌ها در رسانه‌های فنّاوری‌های پیشرفته تبدیل شده است، بدون اینکه بدانند این همهمه به‌خاطر موفقیت‌های هواداران یادگیری عمیق در رقابت‌های سازگار با CTF چندگانه است.

در میان افراد بیرونی، بسیاری از آماردانان دانشگاهی در ظاهر با خط فکری غالب دانشگاهی وجود دارند که به نظر می‌رسد درک کمی از قدرت CTF در ایجاد پیشرفت، در انواع حوزه‌های فنّاوری داشته باشند. من به یاد ندارم که CTF در یک سخنرانی کنفرانسی در یک کنفرانس آمار حرفه‌ای یا سمینار دانشگاهی در یک دانشگاه پژوهشی بزرگ، خودی نشان داده باشد.

نویسنده معتقد است که چارچوب وظیفه مشترک، تنها ایده‌ای از یادگیری ماشین و علم داده است که بیشتر از هر چیز از بذل توجه در آموزش آماری امروزی برخوردار نشده است.

۴.۶ مهارت‌های مورد نیاز

چارچوب وظیفهٔ مشترک، خواسته‌های متعددی را بر کارکنان یک رشته تحمیل می‌کند:

- کارکنان باید مدل‌های پیشگوگری ارائه دهند که بتوان آن‌ها را با روش امتیازدهی CTF مورد بحث، ارزیابی کرد. بنابراین آن‌ها باید خود را تسلیم نظم و انضباط تحمیل‌شده توسط ابداع‌کنندگان CTF کنند.
- کارکنان حتی ممکن است نیاز به اجرای یک CTF مشتری‌مدار برای مسئلهٔ خود داشته باشند؛ بنابراین آن‌ها باید هم یک حوزهٔ فناوری اطلاعات برای ارزیابی قواعد امتیازدهی را ایجاد کنند و هم باید مجموعه‌داده‌ای را به دست آورند که بتواند پایهٔ منبع داده‌های به‌اشتراک‌گذارده در قلب CTF را تشکیل دهد.

به‌طور خلاصه، مهارت‌های فناوری اطلاعات در قلب الزامات مورد نیاز برای کار در مدل‌بندی پیشگوگر قرار دارند. این مهارت‌ها مشابهِ مهارت‌های آزمایشگاهی است که یک دانشمندِ آزمایشگاه محلول‌ها برای انجام آزمایش به آن نیاز دارد، بدون نیاز به ریاضیات.

استفاده از CTFها تقریباً در همان زمان به مرحلهٔ عمل رسید که جنبش نرم‌افزار متن‌باز و در تعاقب آن، ورود محیط‌های برنامه‌نویسی کمّی که بر جوامع تحقیقاتی خاص حکم‌فرما بودند، آغاز شد. تسلط QEP این امکان را برای محققان فراهم آورد که اسکریپت‌ها، به‌ویژه آن اسکریپت‌هایی را که یک مدل پیشگوگر پایه‌ای یا یک گردش‌کار امتیازدهی پایه‌ای را پیاده‌سازی می‌کنند، به‌سهولت در بین جوامع خود به اشتراک بگذارند. بنابراین مهارت‌های لازم برای کار در یک CTF، بسیار خاص و بسیار قابل‌آموزش شد ــ آیا می‌توانیم مجموعه‌ای از اسکریپت‌ها را دانلود کرده و به شکلی سازنده در آن دست ببریم؟

۷ آموزش علم دادهٔ مورداجماع امروزی

نگاه کردن به آنچه در برنامه‌های علم دادهٔ امروزی، که در برخی از دانشگاه‌هایی که اخیراً این رشته را تاسیس کرده‌اند، آموزش داده می‌شود، می‌تواند روشنگر باشد. بگذارید وبگاه جذاب و آگاهی‌بخش کارشناسی ارشد علم دادهٔ شعبهٔ برکلی دانشگاه کالیفرنیا را در نظر بگیریم. با مرور برنامهٔ درسی در وبگاه مربوط[1] به پنج درس پایهٔ زیر برمی‌خوریم: طرح تحقیق و کاربرد برای داده‌ها و تحلیل؛ کاوش و تحلیل داده‌ها؛ ذخیره‌سازی و بازیابی داده‌ها؛ یادگیری ماشین کاربردی؛ دیداری‌سازی و ارتباطات.

1. https://datascience.berkeley.edu/academics/curriculu

به نظر آشکار می‌رسد که تنها «ذخیره‌سازی و بازیابی داده‌ها» است که در گروه‌های آمار سنتی تدریس نمی‌شود؛ و بررسی دقیق واژه‌ها نشان می‌دهد که مبحثی با کمترین میزان سنتی بودن در میان مباحث دیگر، یعنی «یادگیری ماشین کاربردی»، از نظر آماردانی که موضوعات واقعی پوشش داده شده را از ذهن می‌گذراند، بسیار شبیه به آن چیزی است که شاید یک گروه آمار ارائه می‌دهد یا باید ارائه دهد ــ بااین‌حال، استفاده از «یادگیری ماشین» در عنوان درس، هشدار پنهانی است مبنی بر اینکه کفهٔ رویکرد ممکن است به جای استنباط، به‌شدت به سمت مدل‌بندی پیشگوگر سنگین شده باشد.

یادگیری ماشین یک زمینهٔ به‌سرعت در حال رشد در محل تلاقی علوم کامپیوتر و آمار و موضوعش یافتن الگوها در داده‌هاست. این مبحث، عامل پیشرفت‌های شگرف در فنّاوری، از توصیه‌های مختصِ هر شخص محصولات گرفته تا تشخیص گفتار در تلفن‌های همراه است. این درس، آشنایی گسترده‌ای با ایده‌های کلیدی در یادگیری ماشین فراهم می‌آورد. تأکید، به‌جای نتایج نظری، بر شهود و نمونه‌های عملی خواهد بود؛ گرچه تجربیاتی با احتمال، آمار، و جبر خطی مهم خواهد بود.

انتخاب مباحث ممکن است تنها ایده‌ای جزئی از آنچه در این درس می‌گذرد، ارائه دهد. تحت عنوان «ابزارها»، آرایه‌ای از هسته‌های فناوری اطلاعات را پیدا می‌کنیم. کتابخانه‌های پایتون برای جبر خطی، رسم نمودار، یادگیری ماشین: numpy، matplotlib، sk-learn/گیت‌هاب برای ارسال کد پروژه.

به‌طور خلاصه، شرکت‌کنندگان در این درس در حال تولید و ارسال کد هستند. توسعهٔ کد برای آموزش آمار هنوز به‌طور کامل جزو آداب و اصول در نظر گرفته نشده است و در بسیاری از درس‌های آمار با استفاده از کدهای R یا دیگر محیط‌های برنامه‌نویسی کمّی انجام می‌شود که استفاده از آن برای تحلیل داده‌ها برای دانشجویان بسیار «آسان‌تر» است، به این دلیل که کل کار تحلیل داده‌های مدرن پیشتر عملاً در آن برنامه‌نویسی شده است. بااین‌حال، این شهرت در مورد آر به وجود آمده است که نسبت به پایتون، کمتر می‌تواند خود را با مقیاس‌های مسئله‌های با اندازه‌های بزرگ وفق دهد. این مطلب بدان معناست که این‌گونه تصور می‌شود که فردی که کار خود را با پایتون انجام می‌دهد، شاید سخت‌تر کار کرده و پشتکار و تمرکز بیشتری نسبت به کسی که همان کار را با R انجام می‌دهد، از خود نشان داده است.

چنین تصوراتی، زمانی که درس‌های پیشرفته را در نظر می‌گیریم، همچنان به قوت خود باقی

می‌مانند. آزمایش‌ها و استنباط علّی؛ رگرسیون کاربردی و تحلیل سری‌های زمانی؛ ملاحظات حقوقی، سیاست‌گذاری، و اخلاقی برای دانشمندان داده؛ یادگیری ماشین در مقیاس لازم؛ ارتقای مقیاس! داده‌های واقعاً بزرگ.

دو درس اول مانند درس‌های آماری متداول به نظر می‌رسند که هر گروه آمار در هر دانشگاه پژوهش‌محور قادر به تدریس آن است. سومی کمتر آشنا به نظر می‌آید اما با درس‌های «ملاحظات حقوقی، سیاست‌گذاری، و اخلاقی برای دانشمندان داده» که برای مدت‌های طولانی در دانشگاه‌های پژوهش‌محور دایر بوده‌اند، هم‌پوشانی دارد.

دو درس آخر، چالش ارتقای مقیاس فرایندها و رویّه‌ها برای داده‌های واقعاً بزرگ را مورد توجه قرار می‌دهند. این‌ها درس‌هایی هستند که معمولاً در گروه‌های آمار سنتی ارائه نمی‌شوند.

اعضای هیئت علمی برنامهٔ علم دادهٔ شعبهٔ برکلی دانشگاه کالیفرنیا چه کسانی هستند؟ ظاهراً آماردانان دانشگاهی مجرّب سنتی نیستند. در بخش وبگاه «دربارهٔ هیئت علمی MIDS» ، در روز جمعه ۱۱ سپتامبر ۲۰۱۵، توانستم فقط شرح‌حال‌های کوتاهی از اعضای هیئت علمی مرتبط با درس‌های عمدتاً غیرآماری (مانند «ارتقای مقیاس! داده‌های واقعاً بزرگ» یا «یادگیری ماشین در مقیاس لازم») را پیدا کنم. برای تقریباً ۵۰ درصد درس‌هایی که موضوعات آماری سنتی را پوشش می‌دادند، شرح‌حال‌های کمتری در دسترس بود، و از آن‌ها چنین برداشت می‌شد که اشاره به مسیرهای شغلی‌ای متفاوت از دکترای آمار سنتی ــ دکترای جامعه‌شناسی یا دکترای علوم اطلاعات ــ دارند. خود برنامه تحت مدیریت پردیس اطلاعات قرار دارد.

توکی در «آیندهٔ علم داده» استدلال کرد که آموزش آمار به عنوان شاخه‌ای از ریاضیات، موجب عقب ماندن تحلیل داده‌ها می‌شود. او مهارت‌آموزی تحت نظر تحلیلگران دادهٔ واقعی و بنابراین داده‌های واقعی را به عنوان راه‌حل می‌دید:

> همهٔ علوم مقدار زیادی هنر در ساختار خود دارند. گذشته از حقایق آموزشی و ساختارهای خوش‌بنیان، همهٔ علوم باید به شاگردان خود یاد بدهند که چگونه در مورد هرچیز به روال آن علمِ خاص فکر کنند، و باورها و عمل‌ورزی‌های کنونی آن چه‌ها هستند. تحلیل داده‌ها نیزباید چنین کند. وظیفهٔ این مبحث، به‌طرز غیرقابل‌اجتنابی، دشوارتر از بسیاری از علوم خواهد بود. فیزیک‌دان‌ها معمولاً به مدتی طولانی و متمرکز، تحت نظر کسانی که قبلاً در این زمینه استاد بوده‌اند، قرار داشته‌اند. تحلیلگران داده حتی اگر آماردانان حرفه‌ای هم باشند، در طول دورهٔ

آموزشی خود به‌مراتب کمتر در معرض مواجههٔ حرفه‌ای با تحلیلگران داده بوده‌اند. امروزه سه دلیل برای این امر وجود دارد، و در بهترین حالت می‌توان به آرامی آن را تغییر داد:

(ج۱) آمار عموماً به‌صورت بخشی از ریاضیات تدریس می‌شود.
(ج۲) در یادگیری فی‌نفسهٔ آمار، توجه محدودی به تحلیل داده‌ها شده است.
(ج۳) تعداد سال‌های ارتباط صمیمانه و جدی با افراد حرفه‌ای برای دارندگان مدرک دکترای آمار بسیار کمتر از این امر برای دارندگان مدرک دکترای فیزیک یا ریاضیات است.

بنابراین تحلیل داده‌ها و آمار پایبند به آن، با مشکلِ به‌طرزی غیرعادی دشوارِ انتقال اصول اساسی آن روبه‌روست، که چیزی است که احتمالاً نمی‌توان آن را در این حوزه و نیز در اغلب زمینه‌ها با گفتمان غیرمستقیم و کار در کنار هم برآورده کرد.

برنامهٔ کارشناسی ارشد علم دادهٔ برکلی یک درس سنگ‌بنایی دارد که دربرگیرندهٔ یک پروژهٔ تحلیل داده‌ها با یک مجموعه‌دادهٔ بزرگ است. در بخشی از ریزمواد این درس گفته می‌شود که

پروژهٔ نهایی . . . تجربه‌ای در فرمول‌بندی و به اجرا در آوردن یک دورهٔ کاری پایدار، منسجم، و پراهمیت فراهم می‌آورد که منجر به یک پروژهٔ ملموس تحلیل علم داده با داده‌های دنیای واقعی می‌شود . . . این درس به صورت یک پروژهٔ گروهی/تیمی (۳-۴ دانشجو) کامل می‌شود، و تمرکز هر پروژه بر روی داده‌های ثانویهٔ دردسترس عموم و از قبل در معرض وجود، خواهد بود.

به نظر می‌رسد که این پروژه برخی از فرصت‌های «مهارت‌آموزی» را که جان توکی به دلیل کارش در حین کسب مدرک شیمی دانشگاهی با آن آشنا و چیزی بود که او آن را برای تحلیل داده‌ها مهم می‌دانست، ارائه می‌کند.

توکی بر این نکته پافشاری می‌کرد که دقت ریاضی، ارزش بسیار محدودی در آموزش تحلیل داده دارد. این دیدگاه در نقل‌قول اخیر دربارهٔ «آیندهٔ علم داده» کاملاً اشکار بود. توکی در جای دیگری در آن مقاله چنین گفته است:

آموزش تحلیل داده آسان نیست و زمان مقرر برای این کار، همیشه بسیار کمتر از حد کافی است. اما پیرو این دیدگاه که «پرهیز از روش کتاب آشپزی و رشد ادراک تنها

> با روش ریاضی، با تأکید بر اثبات، مقدور می‌شود»، این دشواری‌ها بیشتر افزایش یافته‌اند. مشکل کتاب آشپزی‌گونه، تنها مختص تحلیل داده نیست. اما راه‌حل تمرکز بر ریاضیات و اثبات، مختص به آن است.

از نظر توکی، تحلیل داده مانند سایر علوم بود و نه مانند ریاضیات، از این لحاظ که دانشی در میان بود که نیازمند واگویی بود و نه قضیه‌هایی که لازم بود اثبات شوند. توکی دوباره، با کشیده شدن به پیشینهٔ شیمی خود، خاطرنشان کرد که

> رشتهٔ بیوشیمی امروزه حاوی دانش بسیار مفصل‌تری در مقایسه با حوزهٔ تحلیل داده‌هاست. مسئلهٔ کلی تدریس، در آنجا دشوارتر است. بااین‌حال، کتاب‌های درسی تلاش می‌کنند تا حقایق را تاآنجاکه ممکن است، همراه با جزئیات واگویی کنند.

او همچنین پیشنهاد کرد که آزمایشگاه‌های تجربی، راهی برای یادگیری آمار به دانشجو ارائه می‌کنند:

> این حقایق کمی پیچیده‌اند و تدریس آن‌ها ممکن است بی‌نهایت آسان از کار در نیاید، اما هر کلاس می‌تواند تقریباً هریک از آن‌ها را با انجام نمونه‌گیری تجربی خودش بررسی کند.

می‌توان گمانه‌زنی کرد که ازنظر جان توکی، شاید دست کشیدن دانشجویان از درس‌های آمار و رفتن به سراغ درس‌های علم دادهٔ هم‌ارز، احتمالاً چیز بدی نبوده است.

لئو برایمن در مقالهٔ «مدل‌بندی آماری: دو فرهنگ» خود، استدلال کرد که تدریس مدل‌سازی تصادفی و استنباط تا حدِ حذفِ مدل‌بندی پیشگوگر، بر توانایی آمار در تهاجم به جالب‌ترین مسائلی که او آن‌ها را در افق می‌دید، آسیب می‌رساند. مسائلی که او در آن زمان به آن‌ها اشاره کرد، امروزه کاربردهای داغ علم داده را تشکیل می‌دهند. بنابراین، ممکن است که برایمن از برنامه‌های آموزشی که توازن را بین استنباط و پیشگویی برعکس می‌کند، استقبال کرده باشد، یعنی برنامه‌هایی ازقبیل کارشناسی ارشد علم دادهٔ دانشگاه کالیفرنیا در برکلی.

اگرچه قهرمانان من، توکی، چمبرز، کلیولند، و برایمن حتماً متوجه ویژگی‌های مثبتی در این برنامه‌ها می‌شوند، گفتن اینکه آیا آن‌ها جهت‌گیری بلندمدت این برنامه‌ها را تأیید می‌کنند یا خیر ــ یا اینکه اصلاً جهت‌گیری بلندمدتی وجود دارد که دربارهٔ آن بتوان اظهارنظر کرد ــ دشوار است. به این تعریف تحقیرآمیز توجه کنید:

دانشمند داده: کسی که در آمار بهتر از هر مهندس نرم‌افزار است و در مهندسی نرم‌افزار بهتر از هر آماردان است.

این تعریف ریشه در واقعیت دارد. برنامه‌های درسی کارشناسی ارشد علم داده، ریشه در انواع مصالحه‌ها دارند: حذف برخی مطالب از برنامهٔ کارشناسی ارشد آمار برای ایجاد فضا برای آموزش پایگاه دادهٔ بزرگ؛ یا، به همان روال، حذف برخی از مطالب از برنامهٔ کارشناسی ارشد پایگاه داده در علوم کامپیوتر و گنجاندن مقداری آمار و یادگیری ماشین. چنین مصالحه‌ای به مدیران کمک می‌کند تا به‌سرعت یک برنامهٔ درسی منجر به مدرک راه بیندازند، بدون اینکه هیچ رهنمودی در مورد جهت‌گیری بلندمدت برنامه و در مورد تحقیقاتی که هیئت علمی مربوط دنبال خواهند کرد، ارائه کنند. قهرمانان من امکان داشت که چه رهنمودهای بلندمدتی ارائه کنند؟

۸ گسترهٔ کامل علم داده

جان چمبرز و بیل کلیولند هریک یک حوزهٔ بالقوه‌ای را در ذهن می‌پروراندند که به‌طور قابل‌ملاحظه‌ای بزرگ‌تر از برنامهٔ کارشناسی ارشد علم دادهٔ مورداجماعی است که ما دربارهٔ آن بحث می‌کردیم، اما برنامه‌ای که درعین‌حال به‌لحاظ فکری پربارتر و ماندگارتر می‌بود.

چشم‌انداز بزرگ‌تر، فردِ شاغل در این حرفه را در جایگاهی می‌بیند که در تکاپوی استخراج اطلاعات از داده‌هاست ــ دقیقاً آن‌گونه که قبلاً درتعریف‌های علم داده دیدیم. این حوزهٔ بزرگ‌تر به تک‌تک گام‌هایی که فرد شاغل این حرفه باید بردارد، اهمیت می‌دهد؛ گام‌هایی که مشتمل‌اند بر آشنایی پیدا کردن با داده‌ها تا رسیدن به نتایجی مبتنی‌بر آن، و گسترش دادن آن حتی به‌مرور مداوم شواهد موجود دربارهٔ بهترین شیوه‌های عمل‌ورزیِ خودِ کلِ آن رشته، توسط آن فرد حرفه‌ای.

در پیروی از چمبرز، اجازه دهید که گردایهٔ فعالیت‌هایی را تاکنون ذکر شده‌اند، به علم دادهٔ کهتر (LDS) و رشتهٔ بزرگ‌تر بالقوه‌ای را علم دادهٔ مهتر (GDS) بنامیم. چمبرز و کلیولند هریک موضوع بسط‌یافتهٔ خود را به بخش‌ها/مباحث/ زیرشاخه‌های خاص فعالیت تفکیک کردند. من ادغام کردن، برچسب‌گذاری مجدد، و تعمیم دو جزئی را که آن‌ها پیشنهاد کردند، سودمند می‌بینم. در این بخش، این رده‌بندی علم دادهٔ مهتر آن‌ها ارائه می‌شود و سپس مورد بحث قرار می‌گیرد.

۱.۸ شش قسمت

فعالیت‌های GDS به شش قسمت رده‌بندی می‌شود:(۱) جمع‌آوری ، آماده‌سازی، و کاوش داده‌ها، (۲) نمایش و تبدیل داده‌ها، (۳) محاسبات با داده‌ها، (۴) مدل‌بندی داده‌ها، (۵) دیداری‌سازی و

ارائهٔ داده‌ها، (۶) علم دربارهٔ علم داده.

اکنون اجازه دهید به جزئیات هر بخش بپردازیم.

GDS1: جمع‌آوری، آماده‌سازی، و کاوش داده‌ها برخی می‌گویند که ۸۰٪ از تلاش‌های اختصاص‌یافته به علم داده صَرفِ غرقه شدن در داده‌های درهم‌وبرهم خودِ شخص و یکی‌شدن با آن است تا اصول پایه‌ای هر آنچه را که در آن‌ها هست یاد بگیرد، به‌طوری که بتوان داده‌ها را برای بهره‌برداری بیشتر آماده کرد. ما سه فعالیت زیر را مشخص کرده‌ایم:

- جمع‌آوری. این جزء شامل طرح آزمایش سنتی به همان صورتی است که برای بیش از یک قرن توسط آماردانان عمل‌ورزی می‌شود، اما انواع گوناگونی از روش‌های مدرن جمع‌آوری داده‌ها و منابع داده را نیز شامل می‌شود. به‌عنوان مثال، نظاره‌گر N–نگار گوگل می‌تواند کل پیکرهٔ نوشتگان بین سال‌های ۱۵۰۰–۲۰۰۸ را کمّی‌سازی کند، گوگل تِرِندز می‌تواند جستجوهای وب موردعلاقهٔ زمان‌های اخیرِ کلِ جمعیت و حتی مربوط به مکان‌ها را کمّی‌سازی کند، انسان‌ها در سال یک تریلیون عکس می‌گیرند که بسیاری از آن ها در رسانه‌های اجتماعی ارسال می‌شوند[1]، میلیاردها گفتار در رسانه‌های اجتماعی ارسال می‌شود[2]، فنّاوری‌های جدید داده‌سازی مانند نسل بعدی توالی در زیست‌شناسی محاسباتی، ثابت‌سازی‌های مکانی جی‌پی‌اِس، داده‌های اسکن‌کنندهٔ سوپرمارکت‌ها را داریم. مهارت‌های نسل بعدی می‌تواند شامل اسکرپینگ وب، پاب‌مِد[3]، پردازش تصویر، و دست‌بَری[4] در توییتر، فیس‌بوک، و رِدیت[5] باشد.
- آماده‌سازی. بسیاری از مجموعه‌داده‌ها حاوی ناهنجاری‌ها و چیزهای ساختگی هستند. هر پروژهٔ داده‌رهنمون نیاز به شناسایی و توجه به چنین مسائلی با ذهن باز دارد. پاسخ‌ها از فرمت کردن مجدد و کدگذاری مجددِ خودِ مقدارها، تا پیش‌پردازش بلندپروازانه‌تر، از قبیل گروه‌بندی، هموارسازی و زیرمجموعه‌سازی در تغییرند. امروزه غالباً، با آب‌وتاب از پاک‌سازی داده‌ها و کلنجار با داده‌ها صحبت می‌شود.
- کاوش. از زمان ابداع اصطلاح «تحلیل کاوشگرانهٔ داده‌ها» (EDA) توسط جان توکی، همهٔ ما قبول داریم که هر دانشمند داده زمان و تلاشی جدی برای کاویدن داده‌ها به منظور بررسی سلامت پایه‌ای‌ترین ویژگی‌ها و آشکارسازی ویژگی‌های غیرمنتظره اختصاص می‌دهد.

1. https://arxiv.org/abs/1706.01869 2. https://arxiv.org/abs/1704.05579 3. http://jamanetwork.com/journals/jama/fullarticle/2503172 4. munging 5. Reddit

کارهایی تجسسی از این نوع، بینش‌های مهمی را به هر تلاش داده‌رهنمون اضافه می‌کند.

GDS2: نمایش و تبدیل داده‌ها. یک دانشمند داده طی دورهٔ کاری خود با منابع دادهٔ مختلف بسیاری کار می‌کند. این‌ها معمولاً طیف وسیعی از فرمت‌ها را به خود می‌گیرند که اغلب وابسته به موردهای فردی هستند و دانشمند داده باید به‌راحتی خود را با همهٔ آن‌ها وفق دهد. محدودیت‌های سخت‌افزاری و نرم‌افزاری فعلی، بخشی از این تنوع است؛ به این دلیل که دسترسی و پردازش ممکن است مستلزم تجهیز کردن دقیق منابع توزیع‌شده باشد.

دانشمندان داده اغلب به این نتیجه می‌رسند که یک گام اساسی در کار آن‌ها به اجرا درآوردن یک تبدیل مناسب است که ساختاربندی مجددی به داده‌های در دست اولیه بدهد و آن را به شکلی جدید و خودنمایان‌تر درآورد.

دانشمندان داده، مهارت‌هایی در دو حوزهٔ خاص کسب می‌کنند:

- پایگاه‌های دادهٔ مدرن. گسترهٔ امروزی نمایش داده‌ها مشتمل بر همه‌چیز از فایل‌های متنی خانگی و صفحه‌گسترده‌ها تا پایگاه‌های دادهٔ SQL و noSQL، پایگاه‌های دادهٔ توزیع‌شده، و جریان‌های دادهٔ پخش زنده است. لازم است که دانشمندان داده ساختارها، تبدیل‌ها، و الگوریتم‌های دخیل در استفاده از همهٔ این نمایش‌ها را بدانند.
- بازنمایی‌های ریاضی. این‌ها ساختارهای ریاضی جالب و سودمندی برای نمایش داده‌های از انواع خاص، ازجمله صوتی، تصویری، حسگر، و داده‌های شبکه‌ای هستند. به‌عنوان مثال، برای دریافت ویژگی‌ها با داده‌های صوتی، فرد اغلب تبدیل سپستروم یا تبدیل فوریه انجام می‌دهد؛ برای داده‌های تصویر و حسگر، تبدیل موجک یا هر تبدیل چندمقیاسی دیگر را انجام می‌دهد (به‌عنوان مثال، پیرامیدها در یادگیری عمیق). دانشمندان داده، توانِ کار راحت با چنین ابزارهایی را در خود می‌پرورانند و در مورد به کار گماردن آن‌ها به بلوغ قضاوتی می‌رسند.

GDS3: محاسبات با داده‌ها. هر دانشمند داده باید چندین زبان بداند و برای تحلیل داده‌ها و پردازش داده‌ها از آن‌ها استفاده کند. این‌ها ممکن است شامل زبان‌های عامه‌پسند مانند R و پایتون، اما همچنین زبان‌های ویژه‌ای برای تبدیل و دستکاری متن و برای مدیریت خطوط پیچیده ارتباطی محاسباتی باشند. درگیر شدن در پروژه‌های بلندپروازانه با استفاده از هماهنگی نیم‌دوجین زبان تعجبی بر نخواهد انگیخت. علاوه‌بر دانش اولیهٔ زبان‌ها، دانشمندان داده برای استفادهٔ کارآمد

آن زبان‌ها، باید خود را از بابت اصطلاحات جدید، به‌روز نگاه دارند و لازم است که مسائل عمیق‌تر مرتبط با کارایی محاسباتی را درک کنند. محاسبات خوشه‌ای و ابری و توانایی اجرای تعداد زیادی کار در چنین خوشه‌هایی، به یک عنصر فوق‌العاده قدرتمند در چشم‌انداز مدرن محاسباتی بدل شده است. برای استفاده از این فرصت، دانشمندان داده گردش‌کارهایی را ایجاد می‌کنند که کار را طوری سازماندهی می‌کند که به تکلیف‌های متعددی تقسیم شود تا به‌صورت پی‌درپی یا به‌نحوی دیگر روی ماشین‌های زیادی اجرا شوند. دانشمندان داده همچنین گردش‌کارهایی را ایجاد می‌کنند که مراحل یک تحلیل دادهٔ فردی یا پروژهٔ تحقیق را مستندسازی می‌کند. سرانجام، دانشمندان داده، بسته‌هایی را ایجاد می‌کنند که برای اجزای متداول گردش‌کارِ استفاده‌شده، چکیده درست می‌کنند و آن‌ها را برای استفاده در پروژه‌های آینده در دسترس قرار می‌دهند.

GDS4: دیداری‌سازی و نمایش داده‌ها. دیداری‌سازی داده‌ها در یک انتها با نمودارهای بسیار سادهٔ EDA ــ بافت‌نگاشت‌ها، نمودارهای پراکنش، نمودارهای سری‌های زمانی ــ هم‌پوشانی دارد، اما در عمل‌ورزی مدرن می‌توان آن را تا موارد بسیار بهتری فراتر برد. دانشمندان داده اغلب برای تزئین نمودارهای ساده با رنگ‌ها یا نمادهای اضافی به منظور وارد کردن یک عامل مهم جدید، وقت بسیار زیادی صرف می‌کنند و اغلب درک خود از مجموعه‌داده را با خلق یک نمودار جدید متبلور می‌کنند که آن را کدگذاری می‌کند. دانشمندان داده‌ها همچنین داشبوردهایی را برای پایش خطوط پردازش داده‌ها ایجاد می‌کنند که به داده‌های جریانی یا توزیع‌شده با گستردگی زیاد، دسترسی دارند. سرانجام، آن‌ها دیداری‌سازی‌هایی را ایجاد می‌کنند که نتیجه‌های مربوط به یک کاربست مدل‌بندی یا چالش CTF را ارائه کنند.

GDS5: مدل‌بندی داده‌ها. هر دانشمند داده در عمل از ابزارها و دیدگاه‌های هر دو فرهنگ مدل‌بندی لئو برایمن استفاده می‌کند.

- مدل‌بندی مولد، که در آن فرد، یک مدل تصادفی را پیشنهاد می‌کند که ممکن است داده‌ها را تولید کرده باشد، و روش‌هایی را استخراج می‌کند تا ویژگی‌های سازوکار مولدی زمینه‌ای را استنباط کند. این کار به‌طور تقریبی با آمار سنتی دانشگاهی و شاخه‌های آن انطباق دارد.
- مدل‌بندی پیشگوگر، که در آن شخص به ساختن روش‌هایی اقدام می‌کند که در یک جهان مفروض داده‌ها ــ یعنی یک مجموعه‌دادهٔ واقعی بسیار خاص ــ به‌خوبی پیشگویی انجام می‌دهد. این کار به‌طور تقریبی با یادگیری ماشین مدرن و شاخه‌های صنعتی آن مطابقت دارد.

GDS6: علم دربارهٔ علم داده. توکی این مطلب را مطرح کرد که یک «علمِ تحلیل داده» در عالَم وجود هست و باید به‌عنوان یکی از پیچیده‌ترینِ علم‌ها به رسمیت شناخته شود. او از مطالعهٔ آنچه که تحلیلگران داده «فارغ از هر قیدوبند» درعمل انجام می‌دهند، جانبداری و به ما یادآوری کرد که کارآیی واقعی یک ابزار با احتمال به‌کارگماری ضرب در احتمال نتایج مؤثر پس‌از به‌کارگماری، مرتبط است.

دانشمندان داده، علم دربارهٔ علم داده می‌ورزند، هر زمان که گردش‌کارهای معمول تحلیل/پردازش را، برای مثال، در موقع استفاده از داده‌ها دربارهٔ فراوانی رخدادهای آن‌ها در یک حوزهٔ دانشگاهی یا کسب‌وکار، مشخص می‌کنند؛ هر زمان که مؤثر بودن گردش‌کارهای استاندارد را ازنظر زمان انسانی، منبع محاسباتی، اعتبار تحلیل، یا هر سنجهٔ عملکرد دیگر اندازه‌گیری می‌کنند، و هر زمان که پدیده‌هایی نوظهوری در تحلیل داده‌ها، به‌عنوان‌مثال، الگوهای جدید ناشی از گردش‌کارهای تحلیل داده‌ها، یا مصنوعات ناخوشایند در نتایج تحلیل منتشرشده را برملا می‌کنند.

گسترهٔ موردنظر در اینجا، همچنین دربردارندهٔ کار بنیادی برای ممکن‌سازی چنین علمی در آینده ــ مانند رمزگذاری مستندسازی تحلیل‌ها و نتیجه‌های فردی در یک فرمت دیجیتال استاندارد ــ برای خوشه‌چینی در آینده و فراتحلیل است.

در همان حال که تحلیل داده و مدل‌بندی پیشگوگر به یک پیشهٔ گسترنده‌تر جهانی تبدیل می‌شود، «علم در بارهٔ علم داده» به لحاظ اهمیت، به‌طور چشمگیری رشد خواهد یافت.

۲.۸ بحث

این شش رسته از فعالیت‌ها، زمانی که از ظرفیت کامل آن‌ها استفاده شود، زمینه‌ای ازجدّوجهد را پوشش می‌دهند که بسیار گسترده‌تر از آن چیزی است که درتلاش‌های فعلی دانشگاهی تدریس یا مطالعه می‌شود. برنامهٔ ۲۰۰۱ کلیولند برای علم داده شامل چندین رسته است که می‌توان آن‌ها را روی (زیرمجموعه‌هایی) از آنچه در اینجا پیشنهاد شده، نگاشت، به‌عنوان مثال:

- رسته‌های «نظریه» و «مدل‌های تصادفی و روش‌های آماری» را می‌توان به GDS یا روی زیرمجموعهٔ «مدل‌های مولد» «GDS5: مدل‌بندی داده‌ها»، یا روی خود «GDS5: مدل‌بندی داده‌ها» تصویر کرد.
- رسته‌بندی «محاسبات با داده»ی او به زیرمجموعه‌ای از رستهٔ GDS با همان نام نگاشته می‌شود؛ رستهٔ GDS گسترش یافته است تا پیشرفت‌هایی از قبیل هادوپ را که در ۲۰۰۱

هنوز قابل‌مشاهده نبودند، پوشش دهد.

- رستهٔ «ارزیابی ابزار» کلیولند را می‌توان بر روی زیرمجموعه‌ای از «GDS6: علم دربارهٔعلم داده» نگاشت.

درواقع، یک رستهٔ واحد ــ GDS5: مدل‌بندی داده‌ها ــ بر نمایش علم داده در گروه‌های دانشگاهی امروزی، چه در گروه‌های آمار و چه در گروه‌های ریاضی از طریق آموزش و تحقیق آمار سنتی یا در گروه‌های علوم کامپیوتر از طریق یادگیری ماشین، دست برتر را دارد.

این تقطیع، بازتاب‌دهندهٔ نکات مختلفی است که پیشتر در صدد بیان آن‌ها بوده‌ایم:

- موضوع انشعاب‌گری که متخصصان کامپیوتر از آن برای جداسازی "علم داده» از «آمار» استفاده می‌کنند، در اینجا با برهم‌افزودن هم «GDS3: محاسبات با داده‌ها» و هم «GDS2: نمایش داده‌ها» به عنوان بخش‌های اصلی در کنار «GDS5: مدل‌بندی داده‌ها»، مورد توجه قرار گرفته است.
- بر تنش بین یادگیری ماشین و آمار دانشگاهی در رده‌بندی بالا سرپوش گذاشته شده است. بخش اعظم آن به آنچه دانشمندان داده به‌صورت روزانه انجام می‌دهند، ربطی ندارد. همان‌طور که در بالا گفتم، دانشمندان داده باید هم از مدل‌بندی مولد و هم پیشگوگر استفاده کنند.
- هیاهوی مربوط به پایگاه‌های دادهٔ توزیع‌شده، نقشه‌نمابی کردن/ساده کردن و هادوپ در رده‌بندی بالا آشکار نیست. چنین ابزارهایی با «GDS2: نمایش داده‌ها» و «GDS3: محاسبات با داده‌ها» ربط دارند، اما با اینکه امروزه بسیار به آن‌ها استناد می‌شود، آن‌ها صرفاً توانمندسازهای امروزی برای برخی فعالیت‌های بزرگ‌تر هستند. چنین فعالیت‌هایی به‌طور دائمی در بین خواهند بود، درحالی‌که نقش توانمندسازهایی مانند هادوپ رفته‌رفته از بین خواهد رفت.
- برنامه‌های کارشناسی ارشد فعلی در علم داده، تنها جزئی از قلمروی را که در اینجا به تصویر کشیده شد، پوشش می‌دهد. فارغ‌التحصیلان چنین برنامه‌هایی، به‌قدر کافی در معرض کاوش داده‌ها، پاک‌سازی داده‌ها، کلنجار با داده‌ها، تبدیل داده‌ها، علم دربارهٔ علم داده، و دیگر موضوعات ذکر شده در GDS قرار نگرفته‌اند.

سایر ویژگی‌های این سیاهه در زیر ظاهر می‌شود.

۳.۸ آموزش GDS

شناخت کامل گسترهٔ GDS نیاز به پرداختن به هریک از شش شاخهٔ آن دارد. این امر مستلزم تغییرات اساسی در آموزش است.

جنبهٔ رسمی دادن و آموزش دادن «GDS5: مدل‌بندی داده‌ها»، بخش آسانِ علم داده است؛ ما این کار را طی نسل‌ها در درس‌های آمار و برای یک دهه یا بیشتر در درس‌های یادگیری ماشین انجام داده‌ایم؛ و این الگو در برنامه‌های کارشناسی ارشد علم داده که در دوروبر ما ارائه می‌شوند و بخش عمدهٔ کار موظف تدریس ما صرف آن می‌شود، ادامه دارد. بااین‌حال، این «مطالب ساده» تنها جزئی از تلاش مورد نیاز برای استفادهٔ سازنده از داده‌ها را پوشش می‌دهد.

«GDS1: جمع‌آوری، آماده‌سازی، و کاوش داده‌ها»، با حساب زمانی که عمل‌ورزان صرف آن می‌کنند، مهم‌تر از «GDS5: مدلبندی داده‌ها» است، اما تلاش‌های کمی برای جنبهٔ رسمی دادن به کاوش و پاک‌سازی داده‌ها صورت گرفته و موضوعاتی از این دست هنوز هم در تدریس مورد غفلت قرار می‌گیرند. به دانشجویانی که فقط به تحلیل داده‌های ازپیش آماده می‌پردازند، فرصتی برای یادگیری این مهارت‌های ضروری داده نمی‌شود.

تدریس تنها، چگونه می‌تواند، کفاف چنین کاری را بدهد؟ من پیشنهاد می‌کنم که خواننده دو کتاب زیر را با دقت (باهم) مطالعه کند.

- کتاب [۴۷] به تحلیل مجموعه‌ای از پایگاه‌های داده می‌پردازد که تمام جنبه‌های بازی آمریکایی لیگ برتر بیس‌بال، ازجمله هر بازی انجام‌شده در دهه‌های اخیر و هر بازیکنی را که تا به حال در چنین بازی‌هایی شرکت داشته است، شامل می‌شود. این اثرِ به‌طرزِ شگفت‌انگیزی جامع، فهرستی تقریباً از هر سؤالی را که ممکن است در مورد عملکرد کمّی راهبردهای مختلف بیس‌بال داشته باشیم، در نظر می‌گیرد و به‌دقت توضیح می‌دهد که چگونه می‌توان به چنین سؤالاتی با استفاده از یک چنین پایگاه داده‌ای، معمولاً به کمک یک آزمون آماری دونمونه‌ای (یا در قالب اصطلاحات بازاریابی اینترنتی، یک آزمون A/B) پاسخ داد.
- تحلیل داده‌های بیس‌بال با R [۳۴] نحوهٔ دسترسی به خزانهٔ چشمگیر داده‌های بیس‌بال دردسترس را با استفاده از اینترنت و نحوهٔ استفاده از آر به آن داده‌ها را برای تحلیل توأم با بینش نشان داده‌اند.

دانشجویی که قادر باشد نشان دهد که چگونه می‌توان به‌صورت نظام‌مند از ابزارها و روش‌های

آموزش داده شده در کتاب دوم استفاده کرده، به برخی از سؤالات جالب کتاب اول پاسخ دهد، به اعتقاد من، خبرگی واقعی را در قسمت «GDS1: جمع‌آوری، آماده‌سازی و کاوش» در خود پرورانده است. پروژه‌های مشابهی را می‌توان برای سایر قسمت‌های علم دادهٔ «جدید» به وجود آورد. در «GDS3: محاسبات با داده‌ها»، می‌توان به دانشجویان آموزش داد که شخصاً به خلق بسته‌های جدید آر و گردش‌کارهای تحلیل داده‌ای جدید به صورت عملی مبادرت کنند.

بن باومر و همکاران تجربیات خود در [۲۵] را مرور می‌کنند و باومر به مرور نحوهٔ تدریس نخستین و دومین درس در علم داده/آمار که با این رویکرد سازگارند، می‌پردازد.

خواننده، این دغدغه را خواهد داشت که گسترهٔ پروسعت GDS، بسیار وسیع‌تر از آن چیزی است که به آموزش آن عادت کرده‌ایم. توکی، با اشاره به اینکه کتاب‌های درسی بیوشیمی به نظر مطالب بسیار بیشتری را در مقایسه با کتاب‌های درسی آمار در بر دارند، چنین ایرادهایی را پیش‌بینی کرده بود؛ به گمان او به محض اینکه این رشته خود را مقید به آموزش جاه‌طلبانه‌تر بکند، می‌تواند به‌سرعت راه بیفتد

۴.۸ تحقیق در GDS

به محض اینکه قالب GDS را در ذهن مجسم کنیم، می‌توانیم متوجه شویم که امروزه انواع و اقسام «تحقیق GDS» جالب ــ و بسیار تاثیرگذار ــ وجود دارند. بسیاری از آن‌ها، هنوز «خانه»‌ای طبیعی ندارند، اما GDS چارچوبی برای سازماندهی و دسترس‌پذیری بیشتر آن فراهم می‌کند. چند مثال برای دادن انگیزه به خوانندگان می‌آوریم.

۱.۴.۸ محیط‌های برنامه‌نویسی کمّی: R

موضوع کلی «محاسبات با داده‌ها» ممکن است در وهلهٔ اول چنین جلوه کند که گویی قابل کش آوردن برای پوشش دادن مقادیر زیادی از علوم کامپیوتر بابِ روز دانشگاهی است و از آن این‌گونه برداشت شود که شاید هیچ تفاوت واقعی بین علم داده و علوم کامپیوتر وجود ندارد. برعکس، «محاسبات با داده‌ها» هسته‌ای متمایز و هویتی جدای از علوم کامپیوتر دانشگاهی دارد. آزمون تورنسل این است که آیا تمرکز کار بر نیاز به تحلیل داده است یا غیر آن.

قبلاً استدلال کردیم که سیستم آر، عمل‌ورزی تحلیلگران داده‌ها را با ایجاد یک زبان استاندارد که تحلیلگران مختلف همه می‌توانند از آن برای برقراری ارتباط و به‌اشتراک‌گذاری الگوریتم‌ها و گردش‌کارها استفاده کنند، دچار تغییر کرده است. تصور بِکِر و چَمبرز (با S) و بعدها ایهاکا،

جنتلمن، و اعضای تیم هستهٔ R از کار خود، تحقیق در مورد این بود که چگونه می‌توان به بهترین وجه، محاسبات با داده‌های آماری را سازماندهی کرد. من نیز این را در ردهٔ تحقیق قرار می‌دهم و مقولهٔ «GDS3: محاسبات با داده‌ها» را در خطاب به آن می‌بینیم. لطفاً توجه داشته باشید که این تلاش اساساً چه اندازه بلندپروازانه و چقدر تأثیرگذار بوده است. در بررسی اخیر بسیاری از ارائه‌های برخط در مورد ابتکارعمل‌های علم داده، من از دیدن شدت اتکا بر R، حتی توسط مدرسان علم داده که ادعا می‌کنند اصلاً کار آماری نمی‌کنند، دچار حیرت شدم.

۲.۴.۸ کلنجار با داده‌ها: داده‌های پاکیزه

هادلی ویکام، به‌عنوان مؤلف بسته‌های متعدد که در همه‌جا در بین کاربران R محبوبیت دارند، یکی از سهم‌آوران شناخته‌شده به جهان محاسبات آماری است؛ این بسته‌ها مشتمل‌اند بر `ggplot2`، `reshape2`، `plyr`، `tidyr`، `dplyr`، [۵۳، ۵۲، ۵۴]. این بسته‌ها به تجرید برخی از مسائل رایج در داده‌ها پرداخته، به تهاجم به آن‌ها در زیرحوزهٔ علم دادهٔ «GDS2: نمایش و تبدیل داده‌ها» و نیز زیرحوزهٔ «GDS2: دیداری‌سازی و نمایش» اقدام می‌کنند، و ابزارهای ویکام برای عدّهٔ بسیاری به عنوان جزئی جدانشدنی پذیرفته شده‌اند. ویکام در [۵۴] مفهوم داده‌های پاکیزه را مورد بحث قرار داد. ویکام با مورد توجه قرار دادن این برآورد متداول (چیزی که من هم در بالا به آن توجه داشتم) که ۸۰٪ کار تحلیل داده، صرف فرایند پاک‌سازی و آماده‌سازی داده‌ها می‌شود، یک روش منظم تفکر در مورد فرمت‌های داده‌ای «درهم‌وبرهم» مدون می‌کند و مجموعه‌ای از ابزارها را در آر وارد می‌کند که آن‌ها را به یک فرمت داده‌ای جهانی «پاکیزه» برگردان می‌کند. او چندین فرمت داده درهم‌وبرهم را که معمولاً در تحلیل داده‌ها با آن‌ها روبه‌رو می‌شویم، شناسایی می‌کند و نحوهٔ تبدیل هر یک از این فرمت‌ها را به یک فرمت پاکیزه، با استفاده از ابزارهای `melt` و `cast` خود نشان می‌دهد. به محض اینکه داده‌ها «ذوب» شدند، می‌توان با استفاده از ابزارهای کتابخانهٔ `plyr` به‌آسانی روی آن‌ها عملیات انجام داد و سپس داده‌های خروجی حاصل را می‌توان برای استفادهٔ بیشتر به یک شکل نهایی «قالب‌ریزی» کرد.

کتابخانهٔ `plyr` برخی فرایندهای تکراری را که در تحلیل داده‌ها بسیار متداول هستند، به شکل «فلان و بهمان تابع را در مورد هر عنصر/ستون/سطر/برش» از یک آرایه «اِعمال کن»، تجرید می‌کند. ایدهٔ کلی این مطلب ریشه در زبان برنامه‌نویسی APL 360 سال ۱۹۶۰ کِنِت آیویسون و عملگر فروکاست که در آنجا رسمیت یافت، دارد [۳۰]؛ خوانندگان جوان‌تر حتماً استفاده از ایده‌های برگرفته

در ارتباط با نقشه‌نمابی کردن/ساده کردن و هادوپ را، که مواد سازندهٔ اِعمال تابع‌ها بر روی چندین پردازندهٔ موازی را اضافه کرد، دیده‌اند. `plyr` هنوزهم یک تجرید بسیار پرثمر برای کاربران آر ارائه می‌کند، و به‌ویژه به کاربران آر مطالب قابل‌توجهی در مورد پتانسیل روش خاص آر برای اعمال تابع‌ها به عنوان بستارهایی در درون محیط‌ها آموزش می‌دهد.

ویکام نه‌تنها یک بستهٔ آر را توسعه داده که ابزارهای داده‌های پاکیزه را در دسترس می‌گذارد؛ مقاله‌ای هم نوشته است که به کاربران آر در مورد پتانسیل این طرز عمل آموزش می‌دهد. این تلاش ممکن است تأثیر بیشتری بر روی عمل‌ورزی امروزی تحلیل داده‌ها در مقایسه با بسیاری از مقاله‌های آمار نظریِ بسیار باارزش، داشته باشد.

۳.۴.۸ ارائهٔ تحقیق: Knitr

به عنوان یک رویداد چشمگیر سوم، ما به کار ییهوئی شی روی بستهٔ `knitr` در آر اشاره می‌کنیم. این بسته به تحلیلگران داده که بخواهند اسناد منبع بنویسند، کمک می‌کند که کدهای آر را با متن ترکیب کنند و سپس آن اسناد را با اجرای کد آر برگردان کرده، نتایج را از محاسبات زنده استخراج و آن‌ها را به‌صورت فایل PDF، صفحهٔ وب HTML، یا هر محصول خروجی دیگر درج کنند.

درواقع، کلیت گردش‌کار یک تحلیل داده با تفسیر نتایج، ذخیره‌کردن حجم عظیمی از خروجی‌های محاسباتی متحرک مستعد خطا با برش و چسباندن دستی و مکان آن‌ها در سند، درهم‌تنیده شده است.

ازآنجاکه تحلیل داده نوعاً دربرگیرندهٔ ارائهٔ نتیجه‌گیری‌ها می‌شود، شکی نیست که فعالیت‌های علم داده، در معنای گسترده‌تر GDS، شامل تهیه گزارش‌ها و ارائه‌هاست. تحقیقاتی که آن گزارش‌ها و ارائه‌ها را به‌شکلی بنیادی بهبود می‌بخشند، مطمئناً کمک‌حال GDS خواهد بود. در این حالت، می‌توانیم آن را به عنوان بخشی از «GDS3: محاسبات با داده‌ها» تلقی کنیم، زیرا به‌این‌ترتیب، گردش‌کار یک تحلیل را فراچنگ می‌آوریم. همان‌طور که بعداً نشان می‌دهیم، این کار همچنین توانایی تحقیقاتی مهم در «GDS6: علم دربارهٔ علم داده» را امکان‌پذیر می‌کند.

۵.۸ بحث

می‌توان بر تعداد مثال‌های بالا افزود و تحقیقات در GDS را حتی ملموس‌تر کرد. دو مورد خیلی خلاصه:

- برای زیرشاخهٔ «GDS4:دیداری‌سازی و نمایش داده‌ها»، می‌توان به چند کار پژوهشی سرمشق‌گونه

اشاره کرد: کار بیل کلیولند روی گرافیک‌های آماری [۹، ۱۱] همراه با [۵۶] و کتاب‌های دستور زبان گرافیک [۵۲].

- برای زیرشاخهٔ «GDS1: کاوش و نمایش داده‌ها»، البته تحقیقات دست‌اول مربوط به مدت‌ها قبل جان توکی دربارهٔ EDA [۴۹]؛ کار متأخر کوک و سوین دربارهٔ گرافیک پویا [۱۴] را داریم.

نکات اصلی ما در مورد تمام تحقیقات یادشده:

(الف) تحقیق سنتی به معنای آمار ریاضی یا حتی یادگیری ماشین نیست؛

(ب) برای دانشمندان داده بسیار تأثیرگذار از کار درآمده است.

(پ) تحقیقات بسیار بیشتری از این دست را می‌توان و باید انجام داد.

بدون رده‌بندی‌ای از جنس GDS، دانستن اینکه کجا باید «همه چیز را قرار داد» یا اینکه آیا یک برنامهٔ علم دادهٔ معین به‌اندازهٔ کافی برای دانشوران/محققان در گسترهٔ کامل این مبحث تجهیز شده یا خیر، کاری دشوار خواهد بود.

۹ علم دربارهٔ علم داده

مجموعه گسترده‌ای از فعالیت‌های فنی، لزوماً یک علم نیست؛ این می‌تواند صرفاً یک پیشه مانند آشپزی یا یک زمینهٔ فنی مانند مهندسی زمین‌فنّاوری باشد. برای اینکه حق استفاده از واژهٔ «علم» را داشته باشیم، باید رویکردی دایماً در حال تحول و مبتنی‌بر شواهد داشته باشیم. «GDS6: علم دربارهٔ علم داده» اصل را بر این رویکرد قرار می‌دهد؛ ما به‌طور خلاصه برخی از کارهایی را که نشان می‌دهد ما واقعاً می‌توانیم تحلیل داده‌های مبتنی‌بر شواهد در دست داشته باشیم، مرور می‌کنیم. در هر مورد، به نقش اساسی مهارت‌های فنّاوری اطلاعات، میزانی که کار «شبیه علم داده به نظر می‌رسد» و سابقهٔ حرفه‌ای محققان شاغل در این کار اشاره می‌کنیم.

۱.۹ فراتحلیل در سطح کلِ علم

توکی در «آیندهٔ علم داده» پیشنهاد کرد که آماردانان باید این موضوع را مطالعه کنند که مردم امروزه داده‌ها را چگونه تحلیل می‌کنند.

با جنبهٔ رسمی دادن به مفهوم مقایسه‌های چندگانه [۵۰] توکی این ایده را وارد کار کرد که تمامیت پیکرهٔ نتیجه‌گیری‌های تحلیل را می‌توان به‌صورت آماری ارزیابی کرد.

ترکیب چنین ایده‌هایی به فاصلهٔ اندکی به فراتحلیل منجر می‌شود که در آن ما همهٔ تحلیل‌های داده‌ای منتشرشده دربارهٔ یک مبحث معین را مطالعه می‌کنیم. توکی در سال ۱۹۵۳، در مقدمهٔ مقالهٔ خود [۵۰]، یک مثال در مقیاس بسیار کوچک را با شش مقایسهٔ مختلفِ تحتِ مطالعه در نظر گرفت. امروزه سالانه بیش از یک میلیون مقالهٔ علمی، فقط در تحقیقات پزشکی بالینی منتشر می‌شود، و مطالعات تکراری زیادی از مداخله‌ای واحد در دست است. تحلیل داده‌های فراوانی برای فرامطالعه در اختیار داریم!

طی ۱۰ سال گذشته، گسترهٔ چنان فراتحلیل‌هایی به‌طرز چشمگیری پیشرفت کرده است، تصور ما حالا از کل نوشتگان علمی، پیکره‌ای از متن است که باید محصول‌برداری، پردازش و «اسکریپینگ» داده شود تا از داده‌های عددی تعبیه شده در آن جدا شود. این داده‌ها برای یافتن سرنخ‌هایی دربارهٔ فرامسائل به همان روشی که در تمامی علم مورد تحلیل قرار می‌گیرند، تحلیل می‌شوند. می‌توانم به [۲۷، ۲۸، ۲۹] و [۳۸] ؛ و برای آماردانان مقالهٔ «برآوردی از نرخ کشف نادرست در سطح سرتاسر علم . . . » از جاگر و لیک [۳۱] همراه با همهٔ بحث‌های ارائه شده در دنبالهٔ مقاله، اشاره کنم.

به‌طور خاص، فراتحلیلگران دریافته‌اند که بخشی نومیدکننده از نتیجه‌گیری‌ها در نوشتگان علمی بی‌تردید نادرست است (یعنی بسیار بیشتر از ٪۵)، اینکه در بیشتر اندازه‌های اثر منتشرشده اغراق شده است، اینکه بسیاری از نتایج تکرارپذیر نیستند، و الخ.

دولت ما هر سال ده‌ها میلیارد دلار برای تولید بیش از یک میلیون مقالهٔ علمی هزینه می‌کند. بنابراین دانستن اینکه آیا علم به‌صورتی که عملاً به آن پرداخته می‌شود، موفقیت‌آمیز است یا خیر، یا حتی دانستن اینکه چگونه علم را به عنوان یک کل می‌شود بهبود بخشید، دنیایی اهمیت دارد.

قسمت اعظم این نوع تحقیقات، در جامعهٔ آمار کاربردی در معنای موسع آن، به‌عنوان مثال، در دانشکده‌های آموزش، پزشکی، بهداشت عمومی و غیره، اتفاق افتاده است. بخش بزرگی از دستاوردهایی که تا همین اواخر خیره‌کننده بوده، به «پردازش متن» بستگی دارد، یعنی، اسکریپینگ داده‌ها از چکیده‌های گذاشته شده در پایگاه‌های دادهٔ برخط، یا برداشته‌شده از فایل‌های PDF و غیره. طی این فرایند است که «داده‌های بزرگ» را به وجود می‌آوریم، به‌عنوان مثال، یوئانیدیس و همکاران اخیراً همهٔ پی‌مقدارهای موجود در همهٔ چکیده‌های پاب‌مد را جمع‌آوری کردند [۸]. مشارکت‌کنندگان در این زمینه در علم داده کار می‌کنند و هدف آن‌ها پاسخ به سؤالات اساسی در مورد روش علمی به صورتی است که امروزه در حال انجام است.

۲.۹ تحلیل مطالعهٔ متقابل

ازآنجاکه تحقیقات پزشکی، بسیار گسترده و مخاطرات در آن بسیار بالاست، اغلب مطالعات متعددی از یک مداخلهٔ بالینی پایه‌ای واحد، هریک توسط تیم مشخصی با شیوهٔ خاص آن تیم صورت می‌گیرد. تیم‌های مختلف پیشگویی‌های متفاوتی از نتایج مربوط به بیمار و ادعاهای مختلفی از عملکرد پیشگوهای خود ارائه می‌کنند. کدام‌یک از پیشگوها و اصلاً آیا یکی از آن‌ها واقعاً عمل می‌کند؟

جیووانی پارمیگیانی در دانشکدهٔ بهداشت عمومی هاروارد توضیحاتی به من دربارهٔ یک کار انجام‌شدهٔ اعتبارسنجی مطالعهٔ متقابل ارائه کرد [۳] که در آن او و همکاران، گروهی از مطالعاتی را که روش‌هایی را برای پیشگویی بقای سرطان تخمدان از روی اندازه‌گیری‌های بیان ژن توسعه می‌دهند، در نظر گرفتند. آن‌ها از ۲۳ مطالعهٔ سرطان تخمدان با داده‌های در دسترس عموم، یک مجموعه‌دادهٔ ترکیبی دست‌چین دربرگیرندهٔ داده‌های بیان ژن و بقا، شامل ۱۰ مجموعه‌داده جمعاً با ۱۲۵۱ بیمار به وجود آوردند. از ۱۰۱ مقالهٔ منتخب در نوشتگان، آن‌ها ۱۴ مدل پیش‌سنجی مختلف را برای پیشگویی نتیجهٔ بیمار مشخص کردند. این‌ها فرمول‌هایی برای پیشگویی بقا از روی بیان ژن مشاهده‌شده بودند؛ فرمول‌ها توسط تحلیلگران اصلی مربوط، با مجموعه‌داده‌های مطالعهٔ فردی برازش داده شده، و در برخی موارد در برابر مجموعه‌داده‌های جدید جمع‌آوری شده توسط مطالعات دیگر، اعتبارسنجی شده بودند.

پارمیگیانی و همکاران شیوهٔ اعتبارسنجی مطالعهٔ متقابل زیر را در نظر گرفتند: هریک از ۱۴ مدل را به یکی از ۱۰ مجموعه‌داده برازش دهید، و سپس آن را بر روی هریک از بقیهٔ مجموعه‌داده‌ها اعتبارسنجی کنید، همنوایی مخاطرهٔ پیشگویی‌شده را با حکم موت واقعی اندازه‌گیری کنید، و به‌این‌ترتیب یک ماتریس ۱۴ در ۱۰ تولید کنید که امکان مطالعهٔ مدل‌های فردی را در راستای مجموعه‌داده‌ها، و نیز امکان مطالعهٔ مجموعه‌داده‌های فردی را در راستای مدل‌ها، می‌دهد.

نتیجه‌گیری‌های مطالعهٔ متقابل شگفت‌انگیزی به دست آمد. در وهلهٔ اول مشخص شد که مدل یکی از تیم‌ها به‌وضوح بهتر از همه است گرچه در انتشار اولیه، میانی‌ترین عملکرد اعتبارسنجی را گزارش کرده بود. دوم اینکه، پیشگویی یک مجموعه‌داده، آشکارا به‌مراتب «دشوارتر» از بقیه، در معنای نرخ بدرده‌بندی گزارش‌شدهٔ اولیه بود، اما دقیقاً همین مجموعه‌داده بود که درمجموع بهترین مدل را به دست داد.

جدول ۱. مجموعه‌داده‌های OMOP ارقام عددی تعداد افراد یا اشیا را نشان می‌دهند. به‌عنوان‌مثال، M ۴۶٫۵ در گوشهٔ سمت چپ بالا به معنای ۴۶٫۵ میلیون نفر است؛ درحالی‌که M ۱۱۰ در گوشهٔ سمت چپ پایین به معنای ۱۱۰ میلیون روش است.

شیوه	وضعیت	داروها	دورهٔ زمانی	منبع	اندازهٔ جامعه	سرنام
CCAE	۴۶/۵ M	Private	۲۰۰۳–۲۰۰۹	۱/۳۰ B	۱/۲۶ B	۱/۹۸ B
MDCD	۲۰/۸	Medicaid	۲۰۰۲–۲۰۰۷	۳۶۰ M	۵۵۲ M	۵۵۸ M
MDCR	۴/۶ M	Medicare	۲۰۰۳–۲۰۰۹	۴۰۱ M	۴۰۵ M	۴۷۸ M
MSLR	۱/۲ M	lab	۲۰۰۳–۲۰۰۷	۳۸ M	۵۰ M	۶۹ M
GE	۱۱/۲ M	EHR	۱۹۹۶–۲۰۰۸	۱۸۲ M	۶۶ M	۱۱۰ M

این فرامطالعه نشان می‌دهد که هم با دسترسی به همهٔ داده‌های قبلی از گروهی از مطالعه‌ها و هم با آزمایش کردن همهٔ رویکردهای مدل‌بندی قبلی روی همهٔ مجموعه‌داده‌ها، هم می‌توان نتیجه‌ای بهتر و هم درک کامل‌تری از مشکلات و کاستی‌های تحلیل‌های داده‌ای واقعی به دست آورد.

تلاش انجام شده در اجرای این مطالعه نفس‌گیر بوده است. نویسندگان به کندوکاو در جزئیات بیش از ۱۰۰ مقالهٔ علمی پرداختند و نحوهٔ پاک‌سازی داده‌ها و برازش داده‌ها را در هر مورد به‌طور کامل دریافتند. به همهٔ داده‌های زمینه‌ای دسترسی به عمل آمد و در یک فرمت واحد جدیدِ به‌دقت ساخته‌وپرداخته بازپردازش شدند و همهٔ مراحل برازش داده‌ها به‌صورت الگوریتمی بازسازی شد، به‌طوری که امکان اینکه آن‌ها را در مورد مجموعه‌داده‌های دیگر به کار برد، فراهم شد. بازهم، فنّاوری اطلاعات نقش کلیدی ایفا می‌کند؛ بخش اعظم برنامه‌نویسی برای این پروژه در آر انجام شد. پارمیگیانی و همکاران آمارزیستی‌دانانی هستند که به‌شدت در توسعهٔ بسته‌های R دخالت دارند.

۳.۹ تحلیل گردش‌کار متقابل

یکی از مؤلفه‌های پنهان مهم تغییرپذیری در علم، گردش‌کار تحلیل است. مطالعه‌های مختلف از مداخله‌ای یکسان، ممکن است از گردش‌کارهای متفاوتی پیروی کند که امکان دارد باعث آن شود که مطالعه‌ها به نتیجه‌گیری‌های مختلف بینجامند. [۶] گردش‌کار تحلیل را در ۲۴۱ مطالعهٔ fMRI بررسی کرد. او متوجه شد که تقریباً به تعداد مطالعات، گردش‌کار انحصاری وجود دارد! به‌عبارت دیگر محققان، تقریباً برای هر مطالعهٔ fMRI یک گردش‌کار جدید ایجاد می‌کنند.

دیوید مادیگان و همکاران [۴۱، ۳۳] اثر انعطاف‌پذیریِ تحلیل روی اندازه‌های اثر را در مطالعات مشاهده‌ای مورد مطالعه قرار دادند؛ همکاری آن‌ها از این پس OMOP نامیده خواهد

شد. نویسندگانOMOP به عنوان انگیزه، خاطرنشان می‌کنند که در نوشتگان تحقیقات بالینی، مطالعاتی با مجموعه‌داده‌های یکسان، و مداخله و نتیجه‌ای یکسان، اما با گردش‌کار تحلیل متفاوت وجود دارند و نتایج منتشرشده در مورد مخاطرهٔ مداخله برعکس شده‌اند. مادیگان مثال صریح قرار گرفتن در معرض پیوگلیتازون و سرطان مثانه را ارائه می‌کند که در آن، مقاله‌های منتشرشده در BJMP و BMJ در پایگاه دادهٔ زمینه‌ای یکسان، به نتیجه‌گیریهایی در نقطهٔ مقابل هم رسیدند!

نویسندگان OMOP، پنج مجموعه‌دادهٔ مشاهداتی بزرگ به دست آوردند که مجموعاً بیش از ۲۰۰ میلیون سال ــ بیمار را پوشش می‌داد (جدول ۱ را ببینید).

گروه OMOP، چهار نتیجهٔ مختلف را با کدهای «آسیب حاد کلیه»، «آسیب حاد کبد»، «میوکارد حاد»، «خون‌ریزی دستگاه گوارش» مورد بررسی قرار دادند. آن‌ها طیف گسترده‌ای از مداخلات ممکن را برای هر سنجش پیامد در نظر گرفتند، از این قبیل که، آیا بیمارانی که داروی X مصرف می‌کردند بعداً دچار پیامد Y شدند یا خیر. در زیر، «آسیب حاد کبد» به نشانهٔ پیوند «قرارگرفتن در معرض X و آسیب حاد کبد» است.

برای هر نتیجهٔ هدف، محققان مجموعه‌ای از شاهدهای مثبت و منفی شناخته‌شده را مشخص کردند، مداخله‌های X که برای آن‌ها گزاره‌های صدق زمینه مانند «قرارگرفتن در معرض X با آسیب حاد کبد پیوند دارد» معلوم در نظر گرفته شد. با استفاده از چنین شاهدهایی آن‌ها توانستند یک قابلیت شیوهٔ استنباطی را برای تشخیص صحیح پیوندها، با استفاده از اندازهٔ مساحت زیر خم عملیاتی AUC، کمّی‌سازی کنند.

OMOP هفت شیوهٔ مختلف را برای استنباط از مطالعات مشاهده‌ای، با برچسب‌های «CC»، «CM»، «DP»، «ICTPD»، «LGPS»، «O»، و «SCCS» در نظر گرفت. به‌عنوان‌مثال، «CC» به نشانهٔ مطالعات مورد ــ شاهد است، درحالی‌که SCCS به نشانهٔ سری موارد خودشاهد است. در هر حالت، روش استنباط را می‌توان کاملاً خودکار کرد.

OMOP در مطالعه خود، برای هر پایگاه داده، برای هر نتیجهٔ ممکن، هر یک از هفت نوع روش مطالعهٔ مشاهداتی (CC، . . .، SCCS) را در نظر گرفت.

گزارش OMOP نتیجه می‌گیرد که این سه به‌اصطلاح روش خودشاهد روی‌هم‌رفته بهتر از سایر روش‌ها عمل می‌کنند، و SCCS به‌طور خاص روی هم رفته خوب است. بنابراین مطالعهٔ آن‌ها مطالب قابل‌توجهی دربارهٔ اثربخشی شیوه‌های مختلف استنباط برملا می‌کند و ایده‌ای دربارهٔ اینکه استنباط بهبودیافته چه شکلی است و تا چه اندازه ممکن است دقیق باشد، به دست می‌دهد.

این کار، نشان‌دهندهٔ تلاش گسترده‌ای از سوی OMOP است: ساخته‌وپرداخته کردن ماهرانهٔ داده‌ها، برنامه‌ریزی الگوریتم‌های استنباط به‌شکلی یکپارچه، و اجرای آن‌ها در راستای رشته‌ای از وضعیت‌های زمینه‌ای. دست‌وپنجه نرم کردن با داده‌های بزرگ، بخشی اساسی از این پروژه بود؛ اما انگیزهٔ هدایتگر آن، درک این مطلب بود که نوشتگان علمی حاوی منبعی از تغییرات ــ تغییرات روش‌شناختی ــ است که تأثیر آن روی استنباط آینده در این زمینه ممکن است درک و موجب محدود شدن یا حتی کاسته شدن آن شود. مشارکت‌کنندگان در این طرح آماردان و متخصص آمارزیستی بودند.

۴.۹ خلاصه

به نظر می‌رسد که کاستی‌های قابل‌توجهی در اعتبار علمی نوشتگان وجود داشته باشد [۲۸، ۴۶، ۴۰]. قرن گذشته شاهد توسعهٔ گردایهٔ بزرگی از روش‌شناسی آمار، و جدّوجهدی گسترده در استفاده از این روش‌شناسی برای تقویت انتشارات علمی بوده است. جامعهٔ بسیار بزرگی از استفاده‌کنندگان متخصص و نه چندان متخصص این روش‌شناسی وجود دارد. ما اطلاع زیادی در مورد نحوهٔ استفاده از پیکرهٔ این روش‌شناسی نداریم و چیز زیادی نیز از کیفیت نتایج به‌دست‌آمده نمی‌دانیم.

دانشمندان داده نباید کورکورانه به تولید حجم انبوهی از این روش‌شناسی، بدون توجه به نتایجی که در عمل به دست می‌آیند، بپردازند. مطالعاتی که ما با عنوان « GDS6: علم دربارهٔ علم داده» رده‌بندی کرده‌ایم، به ما کمک می‌کنند تا بفهمیم که تحلیل داده‌ها به‌گونه‌ای که در عمل انجام می‌شود، چگونه بر «کل علم» تأثیر می‌گذارد.

مهارت‌های فناوری اطلاعات مطمئناً در تحقیقاتی که هم اینک مورد بحث قرار داده‌ایم، حرف اول را می‌زند. بااین‌حال، صندلی راننده به‌طورکامل در اختیار درک علمی و بینش آماری است.

۱۰ علم داده طی ۵۰ سال آینده

علم داده در سال ۲۰۶۵ در کجا خواهد بود؟ شواهدی که تاکنون ارائه شده، حاوی سرنخ‌های مهمی است که آن‌ها راحالا گرد می‌آوریم.

۱.۱۰ علم مجدداً بر همه چیز مسلط می‌شود

علی‌الاصول، هدف از انتشار علمی توانمندسازیِ بازتولیدپذیریِ یافته‌هایِ پژوهشی است. برای قرن‌ها، نتایج محاسباتی و تحلیل‌های داده‌ها در انتشارات علمی مورد استناد قرارگرفته‌اند، اما برای

خوانندگان از پیچیدگی کامل تحلیل داده‌هایی که توصیف شده‌اند، عموماً فقط اشاره‌ای رفته است. به‌تدریج که عظمت محاسبات جاه‌طلبانه‌تر شده، شکاف بین آگاهی خوانندگان از آنچه نویسندگان انجام داده‌اند، بسیار زیاد شده است. بیست سال قبل، جان باکهایت و من درس‌هایی را که با جان کلائربوت در استنفورد آموخته بودیم، چنین خلاصه کردیم:

> مقاله‌ای در مورد علم محاسباتی در یک نشریهٔ علمی فی‌نفسه کاری دانشورانه نبوده، بلکه صرفاً تبلیغی از دانشوری است. دانشوری واقعی، توسعهٔ محیط کامل نرم‌افزار و مجموعهٔ کامل دستورالعمل‌هایی است که اعداد و ارقام را تولید کرده‌اند.

برای نیل به هدفِ اصلیِ انتشار علمی، باید کد و داده‌های زمینه‌ای به اشتراک گذارده شوند. علاوه‌براین، مزایایی نیز در این کار برای نویسندگان وجود دارد. کار کردن از ابتدا مطابق با یک برنامه برای به‌اشتراک‌گذاری کد و داده‌ها منجر به کار با کیفیت بالاتر می‌شود و این امر را تضمین می‌کند که نویسندگان بتوانند به آثار قبلی خود و نویسندگان همکارشان، دانشجویان و محققان دوره‌های پسادکترا دسترسی داشته باشند [۱۷]. در گذر سالیان، چنین عمل‌ورزی‌هایی بهتر درک شده‌اند [۴۲، ۴۳] و رشد یافته‌اند [۱۹، ۴۴] اگرچه هنوز از جهان‌شمولی فاصلهٔ بسیار دارند. فارغ از مقایسه، مقدار پژوهش‌های اساساً غیرقابل‌بازتولید، به‌مراتب بسیار بیشتر از قبل است [۴۳].

محاسبات قابل‌بازتولید امروزه سرانجام، توسط بسیاری از پیشروان علمی به‌عنوان یک لازمهٔ اصلی برای معتبربودن انتشار علمی به رسمیت شناخته می‌شود. پیام سالانهٔ رالف سیسرون، رئیس آکادمی ملی علوم ایالات متحده، مربوط به سال ۲۰۱۵ تأکیدی بر این موضوع است؛ درحالی‌که مؤسسات تأمین‌کنندهٔ منابع مالی [۱۳] و چندین مجلهٔ کلیدی [۳۹، ۲۴، ۳۵] به تدوین رشته‌ای از ابتکارعمل‌های مربوط به قابلیت‌بازتولید پرداخته‌اند.

برای کار قابل‌بازتولید در محیط محاسباتی امروزی، شخص یک گردش‌کار خودکار ایجاد می‌کند که همهٔ محاسبات و تمامی تحلیل‌ها در یک پروژه را تولید می‌کند. به عنوان فرعی بر آن، شخص سپس می‌تواند به‌راحتی و به‌طورطبیعی به‌طورمداوم به تظریف و بهبودبخشی کار قبلی اقدام کند.

نتایج محاسباتی باید در انتشارات نهایی ادغام شوند. روش‌های سنتی ــ اجرای کارها به‌صورت تعاملی با دست، فرمت‌بندی مجدد داده‌ها با دست، جستجو برای یافتن نتایج محاسباتی، و کپی کردن و چسباندن در اسناد ــ امروزه به نشانهٔ مسئولیت‌ناپذیری تلقی می‌شود. اخیراً چندین چارچوبِ جالبِ که اسکریپت‌نویسی محاسباتی تعبیه‌شده را با سندنگاری ترکیب می‌کند، توسعه داده شده است. با کار کردن در درون نظمی که چنین سیستم‌هایی تحمیل می‌کنند، مستندسازی کلِ محاسبات، کار

بسیار آسانی شده و منجر به یک نتیجهٔ خاص در مقاله‌ای خاص می‌شود. کار ییهوئی شی با بستهٔ knitr ــ که پیش از این ذکر شد ــ یکی از چنین نمونه‌هایی است.

قابل‌بازتولیدبودن آزمایش‌های محاسباتی برای علم دادهٔ صنعتی به همان اندازه اهمیت دارد که برای انتشار علمی مهم است. این کار، یک رویکرد نظم‌یافته را برای مطرح کردن و ارزیابی به منظور اصلاحات بالقوهٔ سیستم و انتقال آسان اصلاحات اعتبارسنجیده را برای استفاده در تولید امکان‌پذیر می‌کند.

محاسبات قابل‌بازتولید در رده‌بندی ما، هم برازندهٔ « GDS4: نمایش داده‌ها» و هم « GDS6: علم دربارهٔ علم داده» است. به‌ویژه، آموزش دادن دانشجویان برای آنکه کار قابل‌بازتولید انجام دهند، ارزیابی آسان‌تر و ژرفترِ کار آن‌ها را امکان‌پذیر می‌کند؛ واداشتن آن‌ها به بازتولید بخش‌هایی از تحلیل‌های دیگران، به آن‌ها امکان می‌دهد که مهارت‌هایی از قبیل تحلیل کاوشگرانهٔ داده‌ها را، که معمولاً در عمل انجام می‌شود اما هنوز به‌طور نظام‌مند تدریس نمی‌شود، بیاموزند؛ و تربیت آن‌ها برای انجامِ کارِ قابل‌بازتولید، کارِ پس‌از فارغ‌التحصیلی آن‌ها را قابل‌اعتمادتر خواهد کرد.

مؤسسات تامین‌کنندهٔ منابع مالی علوم، مدت‌هاست که در سیاست‌های تأمین مالی خود یک شرط مفهومی را گنجانده‌اند که پژوهشگران، کد و داده‌ها را در دسترس دیگران قرار بدهند. بااین‌حال، این بند در هیچ زمانی اجرایی نمی‌شده و همیشه این بهانه وجود داشت که هیچ راه استانداردی برای به‌اشتراک‌گذاری کد و داده‌ها وجود ندارد. امروزه تلاش‌های ایجادگرانهٔ مداومی برای توسعهٔ ابزارهای استانداردی که قابلیت بازتولید را امکان‌پذیر می‌کند، در جریان است [۱۹، ۴۴، ۴۵] بعضی از آن‌ها بخشی از پروژه‌های معروف بنیادهای مور و سیمونز هستند. می‌توانیم بااطمینان پیشگویی کنیم که قابل‌بازتولیدبودن در سال‌های آینده به‌طور گسترده‌ای مورد عمل‌ورزی قرارخواهد گرفت.

۲.۱۰ علم به‌منزلهٔ داده

ضمیمه‌ای مفهومی به یک انتشار علمی، مقدار بسیار زیادی از اطلاعات عددی ــ برای مثال، پی‌مقدارهای گزارش‌شده در متن آن است [۸]. چنین اطلاعاتی باید به منزلهٔ داده مورد مطالعه قرار گیرند. امروزه به دست آوردن چنان داده‌هایی مشکل‌آفرین است؛ این کار ممکن است متضمن خواندن فردبه‌فرد مقاله‌ها و بیرون کشیدن دستی و ترکیب کردن آن‌ها، یا اسکریپینگ وب و پاک‌سازی داده‌ها باشد. هر دو راهبرد، مستعد خطا و زمان‌بر هستند.

با پذیرش گستردهٔ علم باز طی ۵۰ سال آینده، افق جدیدی نمایان می‌شود. نتایج محاسباتی

فردی گزارش‌شده در یک مقاله و کد و داده‌های زمینه‌ای این نتایج، به صورت جهان‌شمول قابل استناد و ازلحاظ برنامه‌نویسی قابل بازیابی خواهد بود. متان گاویش [۲۱، ۲۰] و من چند مقاله نوشتیم که راهی برای گشودن آن دنیای جدید پیشنهاد می‌کرد و سپس به کندوکاو در آیندهٔ علم در چنین جهانی می‌پرداخت.

آن مقاله‌ها مفهوم یک نتیجهٔ محاسباتی قابل‌تأیید (VCR)، یک نتیجهٔ محاسباتی، و فراداده در مورد نتیجه، تغییرناپذیری وابسته به یک URL، و بنابراین به لحاظ برنامه‌نویسی دائماً قابل‌استناد و قابل‌بازیابی‌بودن را تعریف کردند. گاویش با ترکیب کردن محاسبات ابری و ذخیره‌سازی ابری، چارچوب‌هایی سِروِری توسعه داد که مفهوم VCR را پیاده‌سازی کرده، هر نتیجهٔ کلیدی را به‌طور دائم در سرور ثبت می‌کرد و URL استناددهنده را بازمی‌گرداند. او همچنین کتابخانه‌های سمت کاربر (به‌عنوان مثال، برای متلب) ارائه کرد که امکان ایجاد VCRها را می‌داد و لینک مربوط را بازمی‌گرداند، و نیز دسترسی برنامه‌ای به داده‌هایی را که توسط لینک ارجاع‌دهی می‌شد، بازمی‌گرداند. در سمت ایجاد سند، او بسته‌های کلان را ارائه کرد که چنین لینک‌هایی را در اسناد TeX منتشرشده جاسازی می‌کرد. درنتیجه، می‌شد به‌سهولت اسنادی را نوشت که در آن‌ها هر نتیجهٔ عددی محاسبه‌شده برای یک مقاله، قابل‌استناد عمومی و قابل‌وارسی بود و این کار نه‌تنها درموردِ مقدارهای عددی مقدور بود، بلکه اسکریپت محاسباتی زمینه‌ای، هم قابل‌مشاهده و هم قابل‌مطالعه بود.

در دنیایی که در آن هر نتیجهٔ عددی در یک اثر انتشار یافتهٔ علمی، به همراه الگوریتم زمینه‌ای که آن را تولید کرده، قابل‌استناد و قابل‌بازیابی باشد، اجرای رویکردهای فعلی برای فراتحلیل بسیار آسان‌تر است. شخص به‌راحتی می‌تواند همهٔ پی‌مقدارها را، به روشی جهانی و کاملاً قابل‌تأیید، از یک مقالهٔ سازگار با VCR استخراج کند، یا همهٔ نقاط داده‌ای در یک گراف در داخل آن را استخراج کند. در این دنیای آینده، عمل‌ورزی فراتحلیل از نوعی که ما در مورد آن در بخش ۱.۹ صحبت کردیم، البته گسترش بیشتری خواهد یافت. اما فرصت‌های علمی جدید بسیاری نیز به وجود می‌آید. به دو مثال اشاره می‌کنیم:

- به‌اشتراک‌گذاری شاهدها در مطالعهٔ متقابل. در این دنیای جدید، می‌توان داده‌های شاهدی را از مطالعات قبلی استخراج کرد .[۵۱] فرصت‌های جدید مشتمل‌اند بر: (الف) در اختیار داشتن مجموعه‌های شاهدِ به‌طرزی انبوه بزرگ‌تر در مطالعات آینده، (ب) کمّی‌سازی تأثیر گروه‌های شاهد خاص و تفاوت‌های آن‌ها بر نتیجه‌گیری‌های مطالعهٔ فردی، و (پ) ممارست‌های واسنجی «دنیای واقعی» گسترده که در آن هردو گروه درواقع گروه‌های شاهدند.

- مقایسه‌های مطالعهٔ متقابل. مقایسه‌های مطالعهٔ متقابل بخش‌های ۲.۹ و ۳.۹ نیازمند تلاش‌های گسترده‌ای برای بازسازی تحلیل‌ها در مطالعات قبلی توسط نویسندگان دیگر به‌صورت دستی، و سپس ساخته‌وپرداخته کردن داده‌های آن‌ها به‌صورت دستی بودند. زمانی که مطالعه‌ها به‌لحاظ محاسباتی قابل بازتولید و امکان به اشتراک‌گذاری کد و داده‌ها فراهم باشد، طبیعی خواهد بود که الگوریتم مقالهٔ (الف) را روی داده‌های مقالهٔ (ب) اجرا کرد و درنتیجه دریافت که گردش‌کارهای مختلف و مجموعه‌داده‌های مختلف چگونه موجب تغییرات در نتیجه‌گیری‌ها می‌شوند. انتظار می‌رود که این کار به روند غالب در تحقیقات الگوریتمی تبدیل شود.

امکانات بیشتر در [۲۰] مورد بحث قرار گرفته است.

۳.۱۰ تحلیل علمی داده‌ها با آزمون تجربی آن‌ها

در همان حال که علم به‌طور فزاینده‌ای از نظر داده‌ها و الگوریتم‌ها قابل‌کاوش می‌شود، رویکردهای به اشتراک‌گذاری داده‌ها و به اشتراک‌گذاری گردش‌کارها در مطالعه‌های متقابل که در بخش‌های ۲.۹ و ۳.۹ در بالا مورد بحث قرار گرفت، به‌طور گسترده‌ای اشاعه پیدا خواهند کرد. در ۵۰ سال آینده، داده‌های فراوانی برای اندازه‌گیری عملکرد الگوریتم‌ها در راستای کلیتی از مجموعهٔ وضعیت‌ها در دسترس خواهند بود. این امر، روش‌شناسی آماری را دگرگون خواهد کرد. به‌جای به دست آوردن شیوه‌های بهینه تحت مفروضات ایده‌آل در مدل‌های ریاضی، ما عملکرد را به‌دقت با روش‌های تجربی، براساس کل نوشتگان علمی یا زیرمجموعه‌های مرتبط با آن، اندازه‌گیری خواهیم کرد.

بر بسیاری از داوری‌های فعلی دربارهٔ اینکه کدام الگوریتم‌ها برای کدام اهداف خوب هستند، قلم بطلان کشیده خواهد شد. ما به سه مرجع در مورد موضوع اصلی رده‌بندی، با کمی جزئیات، استناد می‌کنیم.

۱.۳.۱۰ مقالهٔ هند و همکاران

هند در [۲۲] وضعیت تحقیقات رده‌بندی‌گر را در سال ۲۰۰۶ خلاصه کرد. او می‌نویسد:

> به‌این‌ترتیب به نظر می‌رسد که اوضاع تا به امروز در یک وضعیت پیشرفت بسیار اساسی نظری است که منجر به تحولات عمیق نظری و قدرت پیشگویی افزایش‌یافته در کاربردهای عملی می‌شود. درحالی‌که همهٔ این موارد درست‌اند، احتجاج مقالهٔ حاضر این است که تأثیر عملی تحولات دچار تورم شده است؛ اینکه گرچه پیشرفتهایی

حاصل شده، اما ممکن است آن‌گونه که مطرح می‌شود، در حد عالی نباشد.

> لُب‌کلام بحث [در این مقاله] آن است که بهبودهایی که به پیشرفت‌های پیشگامانه‌تر و متأخرتر نسبت داده می‌شود، کوچک هستند و اینکه جنبه‌های مسائل عملی واقعی، اغلب چنین تفاوت‌های کوچکی را بی‌ربط یا حتی غیرواقعی می‌سازند، به‌طوری که دستاوردهای گزارش‌شده در زمینه‌های نظری، یا دربارهٔ مقایسه‌های تجربی حاصل از مجموعه‌داده‌های شبیه‌سازی‌شده یا حتی واقعی، به مزیت‌های واقعی در عمل تبدیل نمی‌شوند. به‌عبارت‌دیگر، پیشرفت به‌مراتب بسیار کمتر از آن چیزی است که به نظر می‌رسد.

هند چگونه از چنین ادعای جسورانه‌ای طرفداری کرد؟ از جنبهٔ تجربی، او از «نمونه‌ای به‌تصادف انتخاب‌شده از ۱۰ مجموعه‌داده» از نوشتگان استفاده کرد و نرخ رده‌بندی تجربی را در نظر گرفت. او نشان داد که تحلیل تمایزی خطی، که ریشه در کار فیشر (۱۹۳۶) دارد، به کسری قابل‌توجه (۹۰٪ یا بیشتر) از بهبود قابل‌دستیابی، بالاتر از یک خط‌مبنای حاصل از حدس تصادفی، دست می‌یابد. روش‌های با عملکرد بهتر بسیار بغرنج‌تر و پیچیده‌تر بودند ــ اما عملکرد افزایشی بالاتر از LDA نسبتاً کوچک بود.

نکتهٔ نظری هند دقیقاً با نکته‌ای که توکی در «آینده علم داده» در مورد بهینگی نظری متذکر شده بود، همریخت بود: بهینه‌سازی تحت یک مدل نظریِ باریک‌بین، به بهبودبخشی‌های عملکرد به هنگام عمل‌ورزی منجر نمی‌شود.

۲.۳.۱۰ مقالهٔ دونوهو و جین

برای ملموس ساختن کامل نکتهٔ هند، کار روی رده‌بندی در ابعاد بالا توسط من و جیاشون جین را در نظر بگیرید [۱۶].

فرض کنید که داده‌های X_{ij} متشکل از $1 \le i \le n$ مشاهده روی p متغیر و برچسب‌های دودویی $Y_i \in \{+1, -1\}$ را در دست داریم. ما در جستجوی یک رده‌بندکنندهٔ $T(X)$ هستیم که با یک بردار ویژگی بدون برچسب ارائه شده و برچسب Y را پیشگویی می‌کند. فرض می‌کنیم که ویژگی‌های پرتعدادی وجود دارند، یعنی p در مقایسه با n به‌مراتب بزرگ‌تر است.

یک روش بسیار ناجذاب را در نظر بگیرید: یک رده‌بندی‌گر خطی، که صرفاً ویژگی‌های انتخاب‌شده را با وزن‌های ۱+ یا ۱- ترکیب می‌کند. این روش ویژگی‌هایی را انتخاب می‌کند

که در آن‌ها قدرمطلق نمرهٔ tی تک‌متغیره از یک مقدار آستانه‌ای تجاوز می‌کند و به عنوان علامت ضریب ویژگی، صرفاً از علامت نمرهٔ tی آن ویژگی استفاده می‌کند. آستانه، با نقادی محتوی تعیین می‌شود. در مقالهٔ منتشرشده آن را HC-clip نامیدند؛ این قاعده، به‌شدت ساده است، حتی بسیار ساده‌تر از تحلیل تمایزی خطی فیشر، از این لحاظ که از ماتریس کوواریانس استفاده نمی‌کند و حتی نیازی به استفاده ازضرایب با اندازه‌های مختلف ندارد. تنها نکتهٔ ظریف در آن، استفاده از نقادی محتوا برای انتخاب آستانه است. از سایر جهات، HC-clip، بازگشتی به یک محیط قبل از ۱۹۳۶ است، یعنی قبل از اینکه فیشر (۱۹۳۶) نشان دهد که حتماً «باید» از ماتریس کوواریانس در رده‌بندی استفاده کرد

دتلینگ (۲۰۰۴) چارچوبی را برای مقایسهٔ رده‌بندی‌گرها ایجاد کرد که در یادگیری ماشین براساس رشته‌ای استاندارد از مجموعه‌داده‌ها متداول بودند (در حالت ۲–رده‌ای، این مجموعه‌داده‌ها به ترتیب ALL، لوکیمیا و پروستات نامیده می‌شوند). او این مجموعه‌داده‌ها را در مورد طیفی از تکنیک‌های رده‌بندی استاندارد که در بین جامعهٔ یادگیری آماری متداول هستند (درخت‌های تصمیم تقویت‌شده، جنگل‌های تصادفی SVM، KNN، PAM ، و DLDA) به کار برد. آن روش‌های یادگیری ماشین که دتلینگ مورد مقایسه قرار داد، اغلب «جذاب» و در حال حاضر با تعداد بالایی از استنادات و طرفدارانی پرجوش‌وخروش هستند.

ما مطالعهٔ دتلینگ را، با اضافه کردن قاعدهٔ بسیار سادهٔ برش خود در ترکیب، گسترش دادیم. ما تأسف (یعنی نسبت خطای بدرده‌بندی در یک مجموعه‌دادهٔ معین به بهترین خطای بدرده‌بندی در بین تمام روش‌های موجود روی آن مجموعه‌دادهٔ خاص) را در نظر گرفتیم. عملکردِ پیشنهاد سادهٔ ما روی این مجموعه‌داده به همان خوبی روش‌های دیگر بود؛ حتی دارای بهترین تأسف بدترین حالت است. به‌عبارت‌دیگر، هر یک از تکنیک‌های جذاب‌تر، تأسف بیشینهٔ بدترین را تجربه می‌کند. تقویت، جنگل‌های تصادفی، و امثال آن‌ها؛ به‌طرزی چشمگیر، پیچیده‌ترند و به‌طرزی متناظر، جذابیتی بالاتر بین جامعهٔ یادگیری ماشین دارند. اما در قبال یک سری از محک‌های ازقبل موجود که بین جامعهٔ یادگیری ماشین توسعه یافته، روش‌های پرجاذبه، عملکردی بهتر از ساده‌ترین روش‌ها ــ برش ویژگی با انتخاب دقیق ویژگی‌ها ــ ندارند.

در مقایسه با کار هند، کارِ ما از یک مجموعه‌دادهٔ ازقبل موجود استفاده می‌کرد که ممکن است کمتر در معرض سوگیری گزینش قرار داشته باشد، از این لحاظ که قبلاً در کارزارهای رده‌بندی‌گرهای چندگانه، توسط متخصصان یادگیری ماشین مورد استفاده بوده است.

۳.۳.۱۰ مقالهٔ ژائو و همکاران

در یک پروژهٔ بسیار جالب [۵۷]، پارمیگیانی و همکاران آنچه را که رده‌بندی‌گر Más-o-Menos نامیده‌اند، مورد بحث قرار دادند، که یک رده‌بندی‌گر خطی بود که در آن ویژگی‌ها می‌توانند فقط ضرایب ± 1 داشته باشند؛ این رده‌بندی‌گر بسیار شبیه روش HC-clip است که در بالا مورد بحث قرار گرفت، و درواقع یکی از گونه‌های آن‌ها فقط شامل آن ویژگی‌هایی بود که توسط HC-clip انتخاب شده بودند ــ یعنی روش بخش قبل. ما دوباره به زمینه‌چینی پیش از فیشر ۱۹۳۶، که می‌گوید از ماتریس کوواریانس استفاده کنید، برگشته‌ایم.

ژائو و همکاران در مطالعهٔ خود، Más-o-Menos را با رده‌بندی‌گرهای «پیچیده» براساس تاوانیدن (به‌عنوان‌مثال، لاسو، ستیغی) مقایسه کردند.

مهم‌تر از همه، این نویسندگان گام بنیادی مقایسهٔ عملکرد را روی یک جهان از مجموعه‌داده‌های مورد استفاده در تحقیقات پزشکی بالینی منتشرشده برداشتند. به‌طور خاص، آن‌ها مجموعه‌ای از مجموعه‌داده‌ها را از نوشتگان مربوط به درمان مثانه، پستان، و سرطان تخمدان، ساخته‌وپرداخته کردند و عملکرد پیشگویی هر روش رده‌بندی را بر روی این جهان ارزیابی کردند.

> ما . . . در یک تجزیه و تحلیل گستردهٔ مطالعات بیان ژن واقعی سرطان نشان دادیم که [Más-o-Menos] واقعاً می‌تواند به عملکرد تمایزی خوبی در زمینه‌چینی‌های واقعی، حتی در مقایسه با رگرسیون لاسو و ستیغی دست یابد. نتایج ما توجیهی برای پشتیبانی از استفادهٔ گستردهٔ از آن در عمل ارائه می‌کند. ما امیدواریم که کار ما از تأکید بر تلاش‌های مدل‌بندی پیشگویی در حال انجام در ژنومیکس در توسعهٔ مدل‌های پیچیده به مسائل مهم‌تر طراحی مطالعه، تفسیر مدل، و اعتبارسنجی مستقل تغییر یابد.

نکتهٔ ضمنی، بازهم تلاش‌های نابه‌جای اختصاص‌یافته به روش‌های به‌ظاهر پرزرق‌وبرق در مقایسه با سایر مسائل مهم‌تر است. آن‌ها چنین ادامه می‌دهند:

> یکی از دلایلی که چرا Más-o-Menos با روش‌های پیچیده‌تر مانند رگرسیون تاوانیده قابل‌مقایسه است، شاید این باشد که ما اغلب از یک مدل پیشگویی آموزش‌دیده بر روی مجموعه‌ای از بیماران به منظور ایجاد تمایز بین زیرگروه‌ها در یک نمونهٔ مستقل استفاده می‌کنیم که معمولًا از یک جامعهٔ اندکی متفاوت جمع‌آوری شده و در

یک آزمایشگاه متفاوت پردازش شده است. این تغییرات مطالعهٔ متقابل، به کمک تحلیل‌های نظری استاندارد فراچنگ نمی‌آیند، بنابراین روش‌های به لحاظ نظری بهینه، ممکن است در کاربردهای واقعی عملکرد خوبی نداشته باشند.

در مقایسه با مقاله‌های بحث شده در زیربخش‌های قبلی [۲۲، ۱۶]، این کار، با کاوش در نوشتگان علمی، مستقیماً با عمل‌ورزان رده‌بندی در یک زمینهٔ خاص صحبت می‌کند و رهنمودهایی مبتنی‌بر شواهد ارائه می‌دهد در مورد هر آن چیزی که برای مطالعات انجام‌شده تا به امروز در آن زمینه درست بوده است، درصورتی‌که معلوم می‌شد که همهٔ افراد از تکنیک توصیه‌شده استفاده کرده‌اند.

۴.۱۰ علم داده در سال ۲۰۶۵

در آینده، روش‌شناسی علمی به‌صورت تجربی مورد اعتبارسنجی قرار خواهد گرفت. به اشتراک‌گذاری کد و به اشتراک‌گذاری داده‌ها، اجازهٔ آن را خواهد داد که مجموعه‌داده‌ها و گردش‌کارهای تحلیل از مطالعات در سطح کل علم استخراج شوند. این‌ها در پیکره‌های مجموعه‌داده‌ها و گردش‌کارها ساخته‌وپرداخته خواهند شد. عملکرد روش‌های آماری و یادگیری ماشین، مآلاً متکی‌بر رویکردهای مطالعهٔ متقابل و گردش‌کارهای متقابل خواهد بود که در بخش‌های ۲.۹ و ۳.۹ بحث کردیم. رویکردهای ناظر بر کمّی‌سازی عملکرد، دوباره به دلیل به اشتراک‌گذاری کد و داده‌ها، به‌شکل استاندارد درخواهند آمد. چارچوب‌های وظیفهٔ مشترک جدید بسیاری به عرصه در خواهند آمد؛ بااین‌حال، این موارد جدید همیشه دقت پیشگویی را برای متریک عملکرد خود نخواهند داشت. عملکرد ممکن است متضمن معتبر بودن نتیجه‌گیری‌های حاصل یا خطای نوع II و IIای تجربی نیز باشد. پژوهش به یک سطح فرا [متا]، انتقال خواهد یافت که سؤال در آن به این صورت درمی‌آید: «اگر ما از چنین و چنان روش در سطح کلِّ علم استفاده کنیم، نتیجهٔ در سطح کلِّ علم فراموضعی، چقدر بهبود خواهد یافت؟» درصورتی‌که اندازه‌گیری با استفاده از یک پیکرهٔ پذیرفته‌شده که نمایندهٔ خود علم است، انجام شده باشد.

در سال ۲۰۶۵، استنتاج و برهان ریاضی، بر نتیجه‌گیری‌های مشتق از تجربه‌گرایی وضعیت موجود علم، سبقت نخواهد گرفت. با بازگویی نکتهٔ بیل کلیولند، نظریه‌ای که روش‌شناسی جدیدی را برای استفاده در تحلیل داده‌ها یا یادگیری ماشین تولید می‌کند، بر اساس سودِ قابل‌کمّی‌سازی آن در مسائلی که بیشتر با آن‌ها مواجه می‌شویم، ارزشمند تلقی خواهد شد، درصورتی‌که معیارِ آن، آزمون تجربی باشد

۱۱ نتیجه‌گیری

هر مفهوم پیشنهادی در علم داده، متضمن مقداری بزرگ‌سازی آمار دانشگاهی و یادگیری ماشین است. گونهٔ «GDS» که به‌طور خاص در این مقاله مورد بحث قرار گرفت، از بینش‌های مربوط به تحلیل داده و مدل‌بندی ناشی می‌شود که سابقهٔ آن به دهه‌های گذشته بازمی‌گردد. در گونهٔ مورد بحث، انگیزهٔ اصلی برای گسترش دادن آن به علم داده، جنبهٔ اندیش‌ورزی آن است. در آینده، ممکن است تقاضاهای زیادی در حوزهٔ صنعت برای مهارت‌های پرورانده‌شده توسط GDS به وجود بیاید؛ بااین‌حال، سؤال‌های اصلی که راه‌برندهٔ این رشتهٔ علمی هستند، صنعتی نیستند.

GDS این موضوع را مطرح می‌کند که علم داده، علم یادگیری از داده‌هاست؛ این علم، روش‌های متضمن در تحلیل و پردازش داده‌ها را مطالعه می‌کند و فنّاوری لازم برای بهبود روش‌ها را به شیوه‌ای مبتنی‌بر شواهد مطرح می‌کند. گستره و تأثیر این علم، درهمان‌حال که داده‌های علمی و داده‌های دربارهٔ خود علم به شکلی همه‌جا موجود در دسترس قرار می‌گیرند، در دهه‌های آینده به‌شدت گسترش خواهد یافت. جامعه در حال حاضر سالانه ده‌ها میلیارد دلار برای تحقیقات علمی هزینه می‌کند و بخش اعظم این تحقیقات در دانشگاه‌ها انجام می‌شود. کارِ GDS به‌طور ذاتی برای درک و بهبود اعتبار نتایج حاصل از تحقیقات دانشگاهی است و می‌تواند نقشی کلیدی در تمام پردیس‌های دانشگاهی که در آن تحلیل و مدل‌بندی داده‌ها از فعالیت‌های عمده محسوب می‌شود، ایفا کند.

مؤخره

«نسخهٔ ۱٫۰۰» این مقاله، تاریخ ۱۸ سپتامبر ۲۰۱۵ را خورد. از زمان انتشار آن من ده‌ها ایمیل از خوانندگان همراه با نظرات آن‌ها دریافت کرده‌ام. چهار مجموعه از این نظرها بسیار ارزشمند بوده‌اند، و من آن‌ها را به همراه پاسخ خود در اینجا مرور خواهم کرد.

علم داده در حکمِ برندسازی

جف وو، استاد مهندسی صنایع و سیستم‌ها در جورجیا تک، ضمن نوشتهٔ خود، خاطرنشان کرد که او از دههٔ ۱۹۹۰ در حال استفاده از اصطلاح «علم داده» بوده است. قبلاً در بخش ۱.۴، به سخنرانی افتتاحیهٔ کارور او در دانشگاه میشیگان اشاره کرده‌ایم. وو در آن سخنرانی پیشنهاد کرد که آمار به تغییر برند خود بپردازد.

قبلاً متذکر شدیم که انجمن سلطنتی آمار میزبانی «مناظره»ای را در ماه مه سال ۲۰۱۵ ــ

ویدیوئی از آن به‌صورت برخط منتشر شده است ــ به عهده داشت و در آن پرسیده شده بود که آیا علم داده در واقع امر، صرفاً یک چنین برندسازی مجدد آن، یا چیزی بزرگ‌تر است. پیشنهاد علم دادهٔ وو جلوتر از زمان خود بود.

من در اینجا استدلال کرده‌ام که علم داده یک تغییر نام تجاری صِرف یا تغییر عنوان آمار نیست. علم دادهٔ مورد اجماع امروزی، آمار را به عنوان یک زیرمجموعه در خود فرامی‌گیرد.

یادداشت ویراستار بخش‌هایی از مقالهٔ اصلی که به سپاسگزاری از افراد مربوط می‌شد و پانویس‌های مؤلف و مترجم به دلیل محدودیت در تعداد صفحه‌های مجله در ترجمه حذف شده‌اند.

مراجع

[1] Barlow, M., *The Culture of BigData, Sebastopol*, O'Reilly Media, Inc., OCA, 2013.

[2] Baumer, B., A Data science course for undergraduates: Thinking with data, *The American Statistician*, **69** (2015), 334-342.

[3] Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., Trippa, L., Cross-Study validation for the assessment of prediction algorithms, *Bioinformatics*, **30** (2014), i105-i112.

[4] Breiman, L., Statistical modeling: The two cultures, *Statistical Science*, **16** (2001), 199-231.

[5] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., Munafò, M. R., Power failure: Why small sample size undermines the reliability of neuroscience, *Nature Reviews Neuroscience*, **14** (2013), 365-376.

[6] Carp, J., *The Secret Lives of Experiments: Methods Reporting in the fMRI Literature*, Neuroimage, **63**, 289–300, (2012). [759]

[7] Chambers, J. M., Greater or lesser statistics: A choice for future research, *Statistics and Computing*, **3** (1993), 182-184.

[8] Chavalarias, D., Wallach, J., Li, A., Ioannidis, J. A., Evolution of reporting p values in the biomedical literature, 1990–2015, *Journal of the American Medical Association*, **315** (2016), 1141-1148.

[9] Cleveland, W. S., *The eElements of Graphing Data, Monterey*, Wadsworth Advanced Books and Software, CA. 1985.

[10] Cleveland, W. S., *Visualizing Data*, Hobart Press, Summit, NJ, 1993.

[11] Summit, NJ: Data science: An action plan for expanding the technical areas of the field of statistics, *Su International Statistical Review*, **69** (2001), 21-26.

[12] Coale, A. J., Stephan, F. F., The case of the indians and the teen-age widows, *Journal of the American Statistical Association*, **57** (1962), 338-347.

[13] Collins, F., and Tabak, L. A., Policy: NIH plans to enhance reproducibility, *Nature*, **505** (2014), 612-613.

[14] Cook, D., Swayne, D. F., *Interactive and Dynamic Graphics for Data Analysis:With R and Gobi*, Springer Science & Business Media, New York, 2007.

[15] Dettling, M., Bag boosting for tumor classification with gene expression data, *Bioinformatics*, **20** (2004), 3583-3593.

[16] Donoho, D., Jin, J., Higher criticism thresholding: Optimal feature selection when useful features are rare and weak, *Proceedings of the National Academy of Sciences*, **105** (2008), 14790-14795.

[17] Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., Stodden, V., Reproducible research in computational harmonic analysis, *Computing in Science and Engineering*, **11** (2009), 8-18.
[18] Fisher, R. A., The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7** (1936), 179-188.
[19] Freire, J., Bonnet, P., Shasha, D., Computational reproducibility: State-of-the-art, challenges, and database research opportunities, in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12*, ACM, 2012, 593-596.
[20] Gavish, M., Three dream applications of verifiable computational results, *Computing in Science & Engineering*, **14** (2012), 26-31.
[21] Gavish, M., Donoho, D., A universal identifier for computational results, *Procedia Computer Science*, **4** (2011), 637-647.
[22] Hand, D. J., Classifier technology and the illusion of progress, *Statistical Science*, **21** (2006), 1-14.
[23] Harris, H., Murphy, S.,Vaisman, M., *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, Sebastopol, O'Reilly Media, Inc, CA, 2013.
[24] Heroux, M. A., *Editorial: ACM TOMS replicated computational results initiative*, *ACM Transactions on Mathematical Software*, **13** (2015), 41, 1-13.
[25] Horton, N. J., Baumer, B. S., Wickham, H., Taking a chance in the classroom: Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics, *CHANCE*, **28** (2015), 40-50.
[26] Hotelling, H., The teaching of statistics, *Ann. of Math. Statistics*, **11** (1940), 457-470.
[27] Ioannidis, J. P. A., Contradicted and initially stronger effects in highly cited clinical research, *Journal of the American Medical Association*, **294** (2005), 218–228.
[28] Ioannidis, J. P. A., Non-replication and inconsistency in the genome-wide association setting, *Human Heredity*, **64** (2007), 203–213.
[29] Ioannidis, J. P. A., Why most discovered true associations are inflated, *Epidemiology*, **19** (2008), 640-648.
[30] Iverson, K. E., A personal view of APL, *IBM Systems Journal*, **30** (1991), 582-593.
[31] Jager, L. R., Leek, J. T., An estimate of the science-wise false discovery rate and application to the top medical literature, *Biostatistics*, **15** (2014), 1-12.
[32] Liberman, M., Fred Jelinek, *Computational Linguistics*, **36** (2010), 595-599.
[33] Madigan, D., Stang, P. E., Berlin, J. A., Schuemie, M., Overhage, J. M., Suchard, M. A., Dumouchel, B., Hartzema, A. G., Ryan, P. B., A systematic statistical approach to evaluating evidence from observational studies, *Annual Review of Statistics and Its Application*, **1** (2014), 11–39.
[34] Marchi, M., Albert, J., *Analyzing Baseball Data with R*, CRC Press, Boca Raton, FL, 2013.
[35] McNutt, M., Reproducibility, *Science*, **343** (2014), 229.
[36] Mosteller, F., Tukey, J. W., Data analysis, including statistics, in *Handbook of Social Psychology*, G. Lindzey, E.Aronson, eds. Addison-Wesley, Reading, MA, 1968, 80-203.
[37] Open Science Collaboration et al., Estimating the reproducibility of psychological science, *Science*, **349** (2015), aac4716.
[38] Pan, Z., Trikalinos, T. A., Kavvoura, F. K., Lau, J., Ioannidis, J. P. A., Local literature bias in genetic epidemiology: An downloaded by empirical evaluation of the chinese literature, *PLoS Medicine*, **2** (2005), 1309.
[39] Peng, R. D., Reproducible research and biostatistics, *Biostatistics*, **10** (2009), 405–408.
[40] Prinz, F., Schlange, T., Asadullah, K., Believe it or not: How much can we rely on published data on potential drug targets?, *Nature Reviews Drug Discovery*, **10** (2011), 712-712.
[41] Ryan, P. B., Madigan, D., Stang, P. E., Overhage, J. M., Racoosin, J. A., Hartzema, A. G., Empirical assessment of methods for risk identification in healthcare data: Results fromthe experiments of the observational medical outcomes partnership, *Statistics in Medicine*, **31** (2012), 4401-4415.

[42] Stodden, V., Reproducible research: Tools and strategies for scientific computing, *Computing in Science and Engineering*, **14** (2012), 11-25.
[43] Stodden, V., Guo, P., Ma, Z., Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals, *PLoS ONE*, **8** (2013), e67111.
[44] Stodden, V., Leisch, F., Peng, R. D., eds., *Implementing Reproducible Research*, Chapman & Hall/CRC, Boca RatonFL, 2014.
[45] Stodden, V., Miguez, S., Best practices for computational science: Software infrastructure and environments for reproducible and extensible research, *Journal of Open Research Software*,**1** (2014), e21.
[46] Sullivan, P. F., Spurious genetic associations, *Biological Psychiatry*, **61** (2007), 1121-1126.
[47] Tango, T. M., Lichtman, M. G., Dolphin, A. E., *The Book: Playing the Percentages in Baseball*, Potomac Books, Inc, Lincoln, NE, 2007.
[48] Tukey, J. W., The future of data analysis, The Annals of Mathematical *Statistics*, **33** (1962), 1-67.
[49] Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
[50] Tukey, J. W., *The Collected Works of JohnW. Tukey*, Multiple Comparisons, vol. 1, H. I. Braun, eds., Wadsworth&Brooks/Cole, Pacific Grove, CA, 1994.
[51] Wandell, B. A., Rokem, A., Perry, L. M., Schaefer, G., Dougherty, R. F., Quantitative biology – Quantitative methods, Bibliographic code (2015), available at `arXiv: 150206900W`.
[52] Wickham, H., Reshaping data with the reshape package, *Journal of Statistical Software*, **21** (2007), 1-20.
[53] Wickham, H., ggplot2, *Wiley Interdisciplinary Reviews: Computational Statistics*, **3** (2011), 180-185.
[54] Wickham, H., The split-apply-combine strategy for data analysis, *Journal of Statistical Software*, **40** (2011), 1-29.
[55] Wickham, H., Tidy data, *Journal of Statistical Software*, **59** (2014), 1-23.
[56] Wilkinson, L., *The Grammar of Graphics*, Springer Science & Business Media, New York, 2006.
[57] Zhao, S. D., Parmigiani, G., Huttenhower, C., Waldron, L., Más-o-Menos: A simple sign averaging method for discrimination in genomic data analysis, *Bioinformatics*, **30** (2014), 3062-3069.

محمدقاسم وحیدی‌اصل: دانشگاه شهید بهشتی، دانشکدهٔ علوم ریاضی
رایانامه: `m-vahidi@sbu.ac.ir`

Mathematical Culture and Thought
Volume 43, Number 1, Spring/Summer 2024, Pages 273–327
doi: 10.30504/mct.2023.1398.1982

50 Years of Data Science*

D. Donoho
Translated by M. Q. Vahidi-Asl[1]

[1] Shahid Beheshti University, Iran

Abstract. More than 50 years ago, J. Tukey called for a reformation of academic statistics. In "The future of data analysis," he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or "data analysis." A recent and growing phenomenon has been the emergence of "data science" programs atmajor universities. This article reviews some ingredients of the current "data science moment," including recent commentary about data science in the popular media, and about how/whether data science is really different from statistics. The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for "scaling up" to "big data." Because all of science itself will soon become data that can be mined, the imminent revolution in data science is not about mere "scaling up," but instead the emergence of scientific studies of data analysis science-wide. I present a vision of data science based on the activities of people who are "learning from data," and I describe an academic field dedicated to improving that activity in an evidence-basedmanner. This new field is a better academic enlargement of statistics and machine learning than today's data science initiatives, while being able to accommodate the same short-term goals.

Keywords: cross-study analysis, data analysis, data science, meta analysis, predictive modeling, quantitative programming environments, statistics
Article history: Recieved 14 June 2023; Accepted 27 June 2023
Article type: translation

* Donoho, D., 50 Years of Data Science, *J. Comput. Graph. Statist.*, **26** (2017), no.4, 745-766.

1. m-vahidi@sbu.ac.ir