کلان داده: مسائل و چالش های پیش روی آن

چکیده

کلان داده[1] به داده‌هایی با حجم زیاد، از اگزابایت[2] (10^{18}) و بیشتر از آن، گویند. این حجم از داده، از ظرفیت سیستم‌های پردازشی و سیستم‌های ذخیره آنلاین فعلی بیشتر است. داده[3]، اطلاعات[4] و دانش[5] در نرخی تولید و جمع آوری می‌شوند که خیلی سریع به حجم اگزابایت /سال می‌رسند. ایجاد و جمع آوری روز به روز سریعتر می‌شود و در طی چند سال به دامنه زتابایت[6]/سال می‌رسند. حجم[7]، تنها یک جنبه کلان داده است، صفات دیگر، تنوع[8]، سرعت[9]، ارزش[10] و پیچیدگی[11] هستند. ذخیره و انتقال داده مسائل مربوط به تکنولوژی هستند، که به نظر می‌رسد که در آینده نزدیک به مشکلات آن‌ها رسیدگی می‌شود، اما چالش‌های بلند مدتی را نشان می‌دهند که نیازمند پارادایم‌های جدید و پژوهش است. ما مسائل و چالش‌ها را با شروع یک برنامه پژوهشی همکارانه بر متدلوژی‌هایی برای طراحی و تحلیل کلان داده آغاز کردیم.

---

[1] Big Data
[2] exabyte
[3] Data
[4] information
[5] knowledge
[6] zettabyte
[7] Volume
[8] variety
[9] velocity
[10] value
[11] complexity

1. مقدمه

مفهوم کلان داده در علوم کامپیوتر از روزهای اولیه کامپیوتر شایع بوده است. "کلان داده" در اصل به معنی حجمی از داده است که نمی‌تواند (به صورت کارامدی) توسط ابزارها و متدهای پایگاه داده سنتی پردازش شود. هر بار که یک رسانه ذخیره سازی جدید اختراع می‌شود، مقدار داده قابل دسترسی بیش از حد می‌شود، چرا که این رسانه‌ها به سادگی قابل دسترسی هستند. تعریف اصلی بر داده ساختار یافته[12] متمرکز است، اما بیشتر پژوهشگران و متخصصان متوجه شده اند که بیشتر اطلاعات جهان به صورت اطلاعات ساختارنیافته و حجیم، و تا حدی در فرم متن و تصویر در دسترس هستند. انفجار داده ربطی به رسانه‌ها ذخیره سازی جدید ندارد.

ما "کلان داده" را به عنوان مقداری داده تعریف می‌کنیم که از نظر ذخیره، مدیریت و پردازش کارآمد فراتر از قابلیت تکنولوژی است. موارد گفته شده تنها توسط یک تحلیل قوی بر خود داده، بیان نیازهای پردازشی، و قابلیت‌های ابزارهای (سخت افزار، نرم افزار و متدهای) استفاده شده برای تحلیل آن، کشف شده است. با بوجود آمدن هر مشکل جدید، نتیجه چگونگی ادامه دادن ممکن است به این توصیه منجر شود که برای اجرای وظایف جدید باید ابزارهای جدیدی داشته باشیم.

تقریبا 5 سال پیش، ما تنها به فضای ذخیره حدود صدها گیگابایت برای کامپیوترهای خود فکر می‌کردیم. امروزه، ما به ده‌ها تا صدها ترابایت فکر می‌کنیم. بنابراین، کلان داده یک هدف رو به رشد است. به بیان دیگر، این مقدار داده فراتر از درک آنی ما است؛ به عبارت دیگر، برای ذخیره کردن آن، دسترسی آن، مدیریت آن، و پردازش آن، نیاز است که سخت کار کنیم.

نرخ رشد فعلی مقدار داده جمع آوری شده وحشتناک است. یک چالش اصلی پژوهشگران و متخصصین IT این است که این نرخ رشد خیلی سریع از توانایی ما برای (1) طراحی سیستم‌های مناسب برای مدیریت موثر داده و (2) تحلیل برای استخراج معانی متفاوت برای تصمیم گیری، سبقت می‌گیرد. در این مقاله مسائل مهم مرتبط با ذخیره، مدیریت

[12] structured data

و پردازش داده را بررسی می‌کنیم. تا انجا که می‌دانیم، ادبیات موضوعی پژوهشی خیلی به صورت موثری این مسائل را بررسی نکرده اند.

1.1 اهمیت کلان داده

در آگوست سال 2010، کاخ سفید، OMB، و OSTP اعلام کردند که در زمینه مراقبت‌های بهداشتی و امنیت ملی؛ کلان داده یک چالش ملی و اولویت دار است [1]. بنیاد ملی علوم، موسسات ملی بهداشت، سازمان زمین شناسی ایالت متحده، وزارت دفاع و انرژی، و آژانس پروژه‌های تحقیقات پیشرفته دفاعی یک طرح R&D مشترک را در مارس 2012 اعلام کردند که بیش از 200 میلیون دلار را بر توسعه تکنیک‌ها و ابزارهای جدید کلان داده سرمایه گذاری کردند. هدف این طرح پیشرفت "... درک ما از تکنولوژی‌های مورد نیاز برای دستکاری و کاووش گسترده مقادیر اطلاعات؛ استفاده از دانش برای دیگر زمینه‌های علمی" و همچنین بررسی اهداف ملی در عرصه‌های حفاظت از سلامت انرژی، آموزش و پرورش است" [14].

تاکید دولت بر این است که چگونه کلان داده "ارزشمند" –در طول دوره‌های و دامنه‌ها – ایجاد می‌شود. ارزش از توانایی تحلیل داده برای توسعه اطلاعات عملی بوجود می‌اید. بررسی ما بر ادبیات موضوعی فنی پنج روش عمومی را که کلان داده می‌تواند از ایجاد ارزش برای سازمان‌ها پشتیبانی کند را بیان می‌کند (جدول 1).

جدول 1: ارزش ایجاد شده از کلان داده

ایجاد شفافیت با در دسترس ساخت کلان داده برای تحلیل عملیاتی و تجاری (کیفیت، هزینه کمتر، کاهش زمان عرضه به بازار و غیره).
پشتیبانی از تحلیل آزمایشی در موقعیت‌هایی که افراد می‌توانند تصمیمات یا رویکردها را، مانند برنامه‌های بازاری خاص تست کنند.
ارزیابی، بر اساس اطلاعات مشتری، در تعریف بخش بندی بازار در یک سطح دقیق تر
پشتیبانی از تحلیل و تصمیمات بلادرنگ بر اساس تحلیل‌های پیچیده استفاده شده برای مجموعه‌های داده از مشتریان و حسگرهای تعبیه شده
تسهیل نوآوری به کمک کامپیوتر در محصولات بر اساس حسگرهای محصول تعبیه شده است که نشان دهنده واکنش مشتری هستند.

زمانی که دولت به دنبال این فرض بود که کاربران کلان داده موفق تر و بهره ورتر باشند و تاثیرات متمایزی در بسیاری از صنعت‌ها داشته باشند، نگرانی اصلی آن‌ها عدم وجود ابزارها و پرسنل آموزش دیده برای کار کردن متناسب با کلان داده بود. سایرین بیان کردند که تحلیل جنریک، تعاملات رسانه اجتماعی، پرونده‌های سلامتی، لاگ‌های تلفن؛ و پروندهای دولتی، خدمات و ابزار بهتری تولید نمی‌کنند، بلکه مجموعه جدیدی از تهاجم به حریم خصوص و بازاریابی تهاجمی و ناخواسته را ایجاد می‌کنند [3]، این نگرانی‌های متضاد دیدگاه رقابتی در مورد اینکه چگونه با کلان داده برخورد کنیم را تحریک می‌کند.

یک مثال از زمینه پزشکی تشریح می‌کند که چگونه و چرا کلان داده و تحلیل‌های جدید ممکن است سودمند باشند. Fox [6] تشریح کرد که چگونه داده‌های فعلی در پرونده‌های پزشکی بیماران و وضعیت سلامتی فعلی برای برنامه ریزی و مشارکت بیماران در برنامه‌های مدیریت سلامت و بیماری استفاده می‌شود. Fox ادعا کرد که دکترها (و شرکت‌های بیمه) باید به جای بیماری، بیمار را درک کنند. برای انجام این کار، آن‌ها باید داده‌ای را- "داده‌های رفتاری و اجتماعی مهمی که بر انتخاب بیمار برای مشارکت، سطح مشارکت، و تناسب داده‌های عمومی با داده‌های رفتاری و اطلاعات سلامتی تاثیر داشته اند را- فراتر از موارد مربوط به شرایط پزشکی بیمار" جمع آوری و تحلیل کند. بنابراین،

برنامه‌ها ممکن است تعیین کنند که چگونه هدف بهتر، افراد را در برنامه‌ها با تحریک مدل‌های پیش بینی کننده که می‌توانند دکترها و مدیرانی که به دنبال تاثیر مثبت رفتار بیمار بر بیماری مزمن هستند را حفظ کنند.

1.2 مشخصه‌های کلان داده

یک دیدگاه، حمایت شده توسط Doug Laney بیانگر این است که کلان داده سه بعد دارد: حجم، تنوع، سرعت. بنابراین، IDC بیان کرد که "فناوری‌های کلان داده یک نسل جدید از تکنولوژی و معماری طراحی شده برای استخراج اقتصادی ارزشمند از حجم بسیار بزرگی از طیف وسیعی از داده، با فعال سازی اتخاذ، کشف و یا تحلیل سرعت بالا را تشریح می‌کنند" [8]. دو مشخصه دیگر نیز به نظر مرتبط می‌رسند: ارزش و پیچیدگی. این دو مشخصه را در جدول 2 خلاصه کردیم.

1.3 کلان داده- کجاست؟

اگر چه ممکن است متوجه آن نباشیم، اما کلان داده ما را احاطه کرده است(جدول 3). بخشی از مسئله این است که، به جز در شرایط غیر معمول، بیشتر ما با حجم زیادی از داده در زندگی روزانه خود سر و کار نداریم. با توجه به عدم وجود همچین تجربه‌ای، اغلب در درک فرصت‌ها و چالش‌های بوجود آمده توسط کلان داده با شکست روبرو می‌شویم. به همین دلیل، در بررسی این شرایط بوجود آمده با مشکلاتی مواجه هستیم.

1.4 مسائل

بیان می‌کنیم که سه مسئله اصلی وجود دارد که نیاز است که در سر و کار داشتن با کلان داده بررسی شوند: مسائل ذخیره سازی، مسائل مدیریتی، و مسائل پردازشی. هر یک از این موارد مجموعه بزرگی از مسائل پژوهشی فنی را به نوبه خود نشان می‌دهد.

جدول 2: مشخصه‌های کلان داده

حجم داده(Data Volume): حجم داده مقدار داده در دسترس برای یک سازمان را در نظر می‌گیرد، چرا که ضرورتا سازمان همیشه به همه داده‌هایی که در اختیار دارد دسترسی ندارد. با افزایش حجم داده، ارزش رکوردهای متفاوت داده متناسب با سن، نوع، دسترسی، و مقدار و فاکتورهای دیگر کاهش می‌یابند.
سرعت داده (Data Velocity): سرعت داده سرعت ایجاد داده، جریان یافتن آن، تجمع را اندازه می‌گیرد. تجارت الکترونیک خیلی سریع سرعت و غنی سازی داده استفاده شده برای تراکنش‌های تجاری متفاوت (وب سایت‌های کلیکی برای مثال) را افزایش می‌دهد. مدیریت سرعت داده خیلی بیشتر از مسئله پهنای باند است؛ این یک مسئله ingest است.
تنوع داده (Data Variety): تنوع داده یک معیار غنی بودن نمایش داده – متنی، تصویر، ویدیو، صدا و غیره – است. از یک چشم انداز تحلیلی، احتمالا این بزرگترین مانع برای استفاده موثر از حجم زیاد داده است. فرم‌های داده ناسازگار، ساختمان داده‌های غیر هم تراز، و داده‌های معنایی ناسازگار چالش‌های قابل توجهی را نشان می‌دهند که به پراکندگی تحلیلی منجر می‌شود.
ارزش داده (Data Value): ارزش داده سودمندی داده در تصمیم گیری را نشان می‌دهد. اشاره می‌کند که "هدف رایانش بینش است، نه اعداد". علوم داده در فهمیدن داده اکتشافی مفید است؛ اما "علوم تحلیلی" از قدرت پیش بینی کننده کلان داده تشکیل شده است.
پیچیدگی(Complexity): پیچیدگی درجه‌ای از همبستگی (بسیار بزرگ) و وابستگی متقابل در ساختارهای کلان داده مانند یک تغییر کوچک (یا ترکیبی از تغییرات کوچک) در یک یا چند عنصر را اندازه می‌گیرد که می‌تواند تغییر بسیار بزرگ و بسیار کوچکی را حاصل کند که در سراسر سیستم به صورت پی در پی بر رفتار تاثیر می‌گذارد، یا به هیچ وجه عوض نمی‌شود.

1.4.1 مسائل مربوط به ذخیره سازی و انتقال

هر بار که رسانه ذخیره سازی جدیدی ابداع می‌شود مقدار داده رو به سمت انفجار می‌رود. آنچه که در مورد انفجار اخیر– تا حدی در رسانه‌های اجتماعی- متفاوت بود این است که هیچ رسانه ذخیره جدیدی در آن دخیل نبود. علاوه

بر این، داده توسط هر کسی و هر چیزی(مانند دستگاه ها)- نه فقط، توسط افراد حرفه‌ای مانند دانشمندان، روزنامه نگاران و نویسندگان- ایجاد می‌شود.

جدول 3: برخی از مثال‌های کلان داده

تشریح	مجموعه داده/دامنه
15-13 پتابایت در سال 2010	برخورددهنده هادرونی بزرگ/ فیزیک ذرات (CERN)
667 اگزابایت در سال 2013	ارتباطات اینترنتی (Cisco)
12- تی بایت از توییت‌هایی روزانه و در حال رشد. میانگین ریتوییت ها برابر 144 به ازای هر توییت است.	رسانه اجتماعی
Zbyte 1.7 (2011) -> Zbyte 7.9 در سال 2015	جهان دیجیتال انسانی
~ 110 TBytes به ازای هر خزنده دامنه که باید آرشیو شود.	وب سایت کتابخانه انگلستان
RFID، متر برقی هوشمند،4.6 میلیارد تلفن دوربینی GPS/w	سایر موارد

محدودیت‌های تکنولوژی دیسک فعلی حدود 5 ترابایت در هر دیسک است. لذا، 1اگزابایت نیازمند 25000 دیسک است. حتی اگر یک اگزابایت داده بتواند در یک سیستم کامپیوتری واحد پردازش شود، قادر به پیوست به تعداد دیسک‌های مورد نیاز نیست. دسترسی به آن داده باعث قطعی شبکه‌های ارتباطی فعلی می‌شود. فرض کنید که یک شبکه 1 گیگابتی در ثانیه دارای نرخ انتقال پایدار موثر 80٪ باشد، پهنای باند پایدار حدود 100 مگابایت است. بنابراین، انتقال یک اگزابایت حدود 2800 ساعت طول می‌کشد، اگر فرض کنیم که انتقال بتواند به صورت پایداری حفظ شود. ممکن است انتقال داده از یک نقطه تجمعی یا ذخیره سازی به نقطه پردازشی دیگر، به دلیل توانایی پردازشی، بیشتر طول بکشد!

دو راه حل به چشم می‌خورد. ابتدا، پردازش داده "در محل" و انتقال اطلاعات نتیجه، به عبارت دیگر، " وارد کردن کد به داده" در برابر متد سنتی "وارد کردن داده به کد"و دوم، انجام تریاژ بر روی داده و تنها انتقال داده‌هایی که برای تحلیل پایین دستی مهم هستند. در هر مورد، یکپارچگی و اصل فراداده باید در زمان انتقال داده حقیقی رعایت شود.

1.4.2 مسائل مدیریتی

مدیریت، شاید، سخت ترین مشکل در بررسی کلان داده باشد. این مسئله از یک دهه پیش در طرح‌های علوم الکترونیکی UK که در آن داده به صورت جغرافیایی توزیع شده است و توسط نهادهای متعدد "مدیریت شده" و به نهادهای متعددی تعلق دارد، دیده شده است. حل مسئله دسترسی، فرا داده، استفاده، به روزرسانی، مدیریت، و ارجاع (در کل) یک مسئله بزرگ ثابت شده است.

بر خلاف جمع آوری داده با متدهای دستی، که در آن پروتکل‌های دقیق اغلب به منظور تضمین دقت و اعتبار دنبال می‌شوند، جمع آوری مجموعه داده دیجیتال خیلی راحت تر انجام می‌شود. غنی سازی نمایش داده دیجیتال استفاده از یک روش قراردادی برای جمع آوری داده را ممنوع کرده است. صلاحیت سنجی داده اغلب بیشتر بر داده‌های از دست رفته یا پرتی به جای تلاش برای اعتبارسنجی هر آیتم تمرکز می‌کند. داده اغلب جزئی است مانند clickstream یا داده‌های سنجشی. با توجه به حجم، اعتبارسنجی هر آیتم داده غیر عملی است: رویکردهای جدیدی برای سنجش، اعتبار و صلاحیت سنجی داده نیاز هستند.

منبع داده‌ها متغیر است- از نظر زمانی و فضایی، فرمت و متد جمع آوری. افراد در اشتراک گذری داده‌های دیجیتال در رسانه‌هایی که دسترسی به ان‌ها راحت است، سهم دارند: مستندات، تصاویر، ضبط صدا و تصویر، مدل‌ها، رفتارهای نرم افزاری، طراحی واسط کاربر- با و بدون ارائه داده کافی که توضیح دهد که چه چیزی، چه زمانی، کجا، چه کسی، چرا و چگونه جمع آوری شده اند و منبع آن‌ها چیست. هنوز هم این داده‌ها به آسانی برای بررسی و تحلیل در دسترس هستند.

هر چه جلوتر می‌رویم، حفظ اصل اطلاعات و داده به یک مسئله مهم تبدیل می‌شود. JASON [10] اشاره کرد که "هیچ روش پذیرفته شده کلی برای ذخیره داده خام،... داده کاهش یافته، و.... کد و انتخاب پارامتر وجود ندارد که داده را تولید کند". علاوه بر این، آن‌ها نوشتند: "ما از هر راه حل مقاوم، متن باز، مستقل از پلت فرم برای این مسئله آگاه نیستیم". تا انجا که می‌دانیم، این قضیه امروز هم صدق می‌کند. برای خلاصه، هیچ راه حل مدیریت کلان داده مناسبی وجود ندارد. این مسائل یک شکاف مهم را در ادبیات پژوهشی بر کلان داده نشان می‌دهند، و نیاز است که این شکاف پر شود.

1.4.3 مسائل پردازش

فرض کنید که یک اگزابایت داده نیاز است که در تمامیت خود پردازش شود. برای سادگی، فرض کنید داده به بلوک‌های 8 کلمه‌ای تقسیم شود، لذا 1 اگزابایت برابر 1K petabytes است. فرض کنید که یک پردازنده 100 دستور را در یک بلوک 5 گیگاهرتزی پردازش می‌کند، و زمان مورد نیاز برای پردازش انتها به انتها برابر 20 نانوثانیه است. برای پردازش 1K petabytes نیازمند مجموع زمان پردازش انتها به انتهای تقریبا 635 سال هستیم. بنابراین، پردازش موثر اگزابایت داده نیازمند پردازش موازی گسترده و الگوریتم‌های تحلیلی جدید برای ارائه اطلاعات به موقع و عملی است.

2. چالش‌های طراحی پویا

چالش‌های بسیاری وجود دارد که نیازمند پژوهش بلند مدت بر کار با کلان داده هستند. Stonebreaker و Hong [18] استدلال کردند که طراحی برای سیستم‌ها و مولفه‌هایی که با کلان داده کار می‌کنند هنوز هم نیازمند درک نیازهای کاربران و تکنولوژی‌هایی هستند که می‌توانند برای حل مسئله بررسی شده استفاده شوند- برای مثال، همه داده‌های کلان و پیش نیازهای آن‌ها یکی نیست. در این نمونه، از آنجایی که این نوع داده به تازگی ایجاد شده است (پیش بینی و جمع اوری شده)، و نه به درستی شناخته شده و نه درک شده است، طراحان نیاز دارند که واسط ها،

گرافیک‌ها و آیکون ها، سازمان کاربردی و مدل‌های مفهومی، استعاره‌ها و قابلیت‌ها را در نظر بگیرند. چرا که کاربران نهایی اغلب طراحان سیستم هستند، این یک چالش طراحی اضافی را ارائه می‌دهد.

با هر افزایش در مقیاس و توسعه تحلیل جدید چالش‌های ناشناخته‌ای بروز می‌کنند. برخی از این چالش‌ها با ابزارها و تکنیک‌های در دست قابل تنظیم هستند. باور داریم که با رفتن به افق فراتر از مجموعه داده‌هایی با اندازه zettabyte با چالش‌هایی مواجه می‌شویم.

2.1 فرآیندهای ورودی و خروجی داده

یک مسئله اصلی که در طراحی کلان داده باید در نظر گرفته شود فرآیند خروجی است. Jacobs [9] این مسئله را بسیار مختصر توضیح داده است – ".... دریافت اطلاعات خیلی ساده تر از خروج اطلاعات است". کار او نشان می‌دهد که ورود و ذخیره داده می‌تواند با فرآیندهایی که در حال حاضر برای پایگاه داده‌های رابطه ای[۱۳] استفاده می‌شوند مدیریت شود. اما، ابزارهای طراحی شده برای پردازش تراکنش، می‌توانند مقدار کم تا زیاد داده را اضافه کنند، به روزرسانی کنند، جستجو و بازیابی کنند، که برای حجم زیاد داده قابل انجام نیست و نمی‌تواند در عرض چند ثانیه تا چند دقیقه انجام شود.

چگونگی دسترسی به مقادیر زیاد داده‌های ساختارنیافته یا نیمه ساختاریافته، و چگونگی استفاده از طرح‌های ابزار مجهول هنوز ناشناخته اند. شفاف است که مسئله ممکن است که نه با مدل سازی بعدی و نه با پردازش تحلیلی آنلاین[۱۴] (OLAP)، که ممکن است از لحاظ عملیاتی آهسته یا محدود باشند، که داده را به سادگی از حافظه نمی‌خوانند، حل نشود. ملاحظات فنی که باید در طراحی اعمال شود شامل نسبت سرعت خواندن دیسک‌های پی در پی متناسب با سرعت دسترسی تصادفی به حافظه است. تکنولوژی حاضر نشان می‌دهد که دسترسی تصادفی به حافظه 150000 برابر آهسته تر از دسترسی ترتیبی است. جدول‌های پیوندی، یک پیش نیاز مفروض حجم بزرگ داده‌های

[13] relational databases
[14] online analytical processing

پراکنده اما تا حدی مرتبط است، شاید مشاهدات در طول زمان، هزینه‌های عملکردی زیادی داشته باشند (Jacobs, 2009).

2.2 کیفیت در برابر کمیت

یک چالش بوجود آمده برای کاربران کلان داده "کمیت در برابر کیفیت" است. از آنجایی که کاربران به داده‌های بیشتری(کمیت) دسترسی دارند، اغلب چیز بیشتری می‌خواهند. برای برخی از کاربران، جمع آوری داده به یک اعتیاد تبدیل می‌شود. شاید، به دلیل اینکه آن‌ها باور دارند که با داده کافی، قادر هستند که به صورت مناسبی توضیح دهند که به دنبال چه پدیده‌ای هستند.

برعکس، یک کاربر کلان داده ممکن است بر کیفیت تمرکز کند که به معنی این است که همه داده‌ها را در دسترس ندارد، اما مقدار زیادی (خیلی زیادی) داده با کیفیت بالا دارد که می‌توانند برای استنتاج نتایج دقیق و با ارزش استفاده شوند (جدول 3).

روش دیگر نگاه کردن به این موضوع این است که، کاربر به چه سطح دقتی نیاز دارد. برای مثال، تحلیل روند ممکن است نیازمند دقتی که سیستم‌های DB سنتی باید داشته باشند؛ نباشد، اما نیازمند پردازش حجیم در یک محیط کلان داده باشد. این مسئله خود را در چالش "سرعت در برابر مقیاس" بحث شده در زیر آشکار می‌کند.

جدول 4: برخی از چالش‌های کمیت و کیفیت

چگونه تصمیم می‌گیریم که کدام داده بی ربط است و مرتبط ترین داده را انتخاب می‌کنیم؟
چگونه تضمین می‌کنیم که همه داده‌ها از یک نوع مشخص قابل اعتماد و دقیق هستند؟ یا، شاید تقریبا دقیق باشد؟
چقدر داده برای ایجاد یک تخمین یا پیش بینی احتمال خاص و دقت یک رویداد مشخص نیاز است؟
چگونه "ارزش" داده در تصمیم گیری را ارزیابی می‌کنیم؟ لزوما بهتر است؟

2.3 رشد کلان داده در برابر گسترش داده‌ها

بیشتر سازمان‌ها انتظار دارند که داده‌های آن در طول عمر آن‌ها با افزایش خدمات آن ها، کسب و کار آن‌ها، شرکا و ارباب رجوعان آن‌ها، پروژه‌ها و تاسیسات آن‌ها، و کارمندان آن‌ها، افزایش یابد. چند کسب و کار به صورت کافی گسترش داده را در نظر گرفتند، چرا که زمانی رخ می‌دهد که غنی سازی رکوردهای داده رشد می‌کند، و زمانی که در طول زمان با اطلاعات اضافی همراه با تکامل تکنیک‌های جدید، فرآیندها و تقاضا برای اطلاعات تکامل می‌یابد. بیشتر این داده‌ها متغیر زمانی هستند- آیتم‌های داده مشابه می‌تواند بیشتر و بیشتر با مقادیر متفاوت بر اساس مهر زمانی جمع آوری شوند. بسیاری از این داده‌ها برای تحلیل گذشته نگر - به خصوص موارد استفاده شده در تحلیل‌های تخمینی و پیش بینی کننده نیاز هستند.

2.4 سرعت در برابر مقیاس

به عنوان حجمی از داده‌های رو به رشد؛ "کلان[15]" ممکن است استعاره‌ای از مقیاس انبار داده‌ها نسبت به مقدار داده‌ای باشد که می‌تواند در یک بازه مشخص، در 24 ساعت برای مثال، پردازش شود. کسب بینش در مورد مسائل تحلیل شده اغلب مهم تر از پردازش همه داده‌ها است. نسبت زمان به اطلاعات در زمانی که یک نفر فرآیندهای بلادرنگ (یا

[15] big

تقریبا بلادرنگ) را در نظر می‌گیرد و داده تقریبا پیوسته را تولید می‌کند، مانند شناسه فرکانس رادیویی[16] (RFID-برای خواندن بی سیم داده‌های الکترونیک استفاده می‌شود، مانند تگ‌های EZPass) و دیگر انواع حسگرها مهم است. یک سازمان باید تعیین کند که چقدر داده در تنظیم بازه پردازش کافی است چرا که این معماری سیستم پردازشی، مشخصه موتورهای محاسباتی، اجرا و ساختار الگوریتم را تحریک می‌کند.

گفته می‌شود، چالش اصلی دیگر انتشار داده است. میان افزارهای ارتباطی مانند گلوگاهی دیده می‌شون. در حالی که سرعت سخت افزارهای ارتباطی با تکنولوژی‌های جدید در حال افزایش است؛ سرعت انتقال پیام به آرامی رو به کاهش است. محاسبات در مقابل دوگانگی ارتباطات توسط سیستم‌های ذخیره کلان داده مانند HDFS یا مجموعه داده‌ها با اندازه Exabyte کاملا برطرف نشده است.

2.5 داده‌های ساختاریافته در مقابل داده‌های ساختارنیافته

تفسیر بین داده ساختار یافته با داده به خوبی تعریف شده (اغلب در جدول ها) به عنوان داده ذخیره شده در پایگاه داده رابطه‌ای تعریف می‌شود، و داده ساختارنیافته (مستثنی از متن، گرافیک، چند رسانه ای) که مناسب تحلیل هستند می‌توانند عملکرد پردازش انتها به انتها را مانع شوند. ظهور پایگاه داده‌های غیر رابطه ای، توزیع شده، و تحلیلگرا مانند NoSQL, MongoDB, SciDB و DBهای داده‌های پیوندی انعطاف پذیر پویایی را در نمایش و سازمان دهی اطلاعات فراهم می‌کند.

برخلاف یک مجموعه داده، یک منبع داده هیچ آغاز و هیچ انتهایی ندارد. جمع اوری داده آغاز می‌شود و همچنان تا زمانی که داده کافی وجود دارد یا دوره انتظار به سر رسد یا پول تمام شود یا هر دو، ادامه می‌یابد. جریان‌های داده با سرعت، فراوانی، حجم و پیچیدگی متغیر، تغییر می‌کنند. جریان داده ممکن است به طور پویایی در دو جهت تغییر کند: (1) تغییر فرمت داده؛ تغییرات لازم در نحوه تجزیه و تحلیل داده، یا (2) تغییر خود داده نیازمند تحلیل‌های متفاوت برای پردازش آن است. یک فاکتور پیچیده یک فرض ضمنی است مبنی بر اینکه جریان‌های داده به خوبی

[16] radio frequency identifiers

رفتار می‌کنند و داده به ترتیب بیشتر یا کمتر می‌رسد. در واقعیت، جریان‌های داده خیلی هم خوب رفتار نمی‌کنند و اغلب داده‌های ترکیبی و اختلالاتی، داده‌های بی ربط به داده مورد نظر تجربه می‌شوند. نیاز به یک تفکر مجدد بر پرازش جریان داده، برای تاکید بر تحلیل پیوسته جریان‌های داده توزیع شده و گسسته خود داریم.

2.6 مالکیت داده

مالکیت داده یک چالش مهم را، به خصوص در عرصه رسانه اجتماعی، نشان می‌دهد، در حالی که مقدار petabytes از داده‌های رسانه اجتماعی در سرورهای Facebook, MySpace, and Twitte، در حقیقت به آن‌ها تعلق ندارد (اگر چه ممکن است ادعای آن را داشته باشند). اساسا،"مالکان" صفحات یا حساب‌ها باور دارند که ان‌ها مالک داده هستند. این دوگانگی باید در دادگاه بررسی شود. Kaisler، Money and Cohen [12] این مسئله را با توجه به رایانش ابری و همچنین دیگر جنبه‌های قانونی بررسی کردند.

با مالکیت حداقل مسولیت برای اطمینان از صحت بوجود می‌اید. این ممکن است برای افراد زیاد مهم نباشد، اما تقریبا سازمان‌های عمومی و کسب و کارها به آن نیاز دارند. به هر حال، اجرای همچین فرضیه‌ای (با سیاست کمتر) به شدت مشکل است. از انجا که هیچ تامین کننده رسانه اجتماعی منبعی برای بررسی هر آیتم داده در سرورهای خود نیست، توافقات ساده کاربر کافی نیست.

با پیشرفت بسیاری از سایت‌های رسانه اجتماعی، روندی در تحلیل کلان داده به سمت ترکیب داده‌های اول شخص، معتبر، با داده‌های خارجی عمومی و سوم شخص بوجود امد، که تا حدی توسط هر متدلوژی رسمی اعتبارسنجی و تایید نمی‌شدند. علاوه بر داده‌های تایید نشده: وفاداری مجموعه داده‌ها به خطر می‌افتد؛ ممکن است نهادهای غیرمرتبط را معرفی کند؛ و ممکن است به پیوندهای بیشماری در میان نهادها منجر شود. در نتیجه، دقت نتایج از موارد استنتاج شده از پردازش این داده‌های ترکیبی به صورت گسترده تغییر می‌کنند.

جدول 5: برخی از چالش‌های مالکیت کلان داده

چه زمانی اعتبار داده (در دسترس) منقضی می‌شود؟
اگر اعتبار داده منقضی شود، آیا داده باید از وب سایت‌های عمومی یا مجموعه‌های داده حذف شود؟ کجا و چگونه داده‌های منقضی را آرشیو می‌کنیم؟ آیا باید آن‌ها را آرشیو کنیم؟
داده‌های منقضی شده در کجا و چگونه آرشیو می‌شوند؟ آیا باید آن‌ها را آرشیو کنیم؟
چه کسی مسئول وفاداری و دقت داده است؟ آیا کاربر باید از آن مطلع شود ؟

2.7 تطبیق و امنیت

در دامنه‌های اصلی، مانند رسانه اجتماعی و اطلاعات مراقبت‌های سلامت و بهداشت، هر چه داده بیشتری در مورد افراد جمع آوری شود، ترس این وجود دارد که آن سازمان اصلی اطلاعات بیشتری در مورد افراد داشته باشد. برای مثال، داده‌های جمع آوری شده در سیستم‌های ثبت الکترونیکی سلامت مطابق با مقررات HIPAA/HITECH تاکنون نگرانی‌هایی را در مورد نقض حریم خصوصی بروز داده اند. الگوریتم‌های در حال توسعه که داده شخصی را در میان یک مجموعه داده به اندازه کافی بزرگ برای تضمین حفظ حریم خصوصی تصادفی سازی می‌کنند یک مسئله پژوهشی کلیدی هستند.

شاید بزرگترین تهدید برای امنیت شخصی جمع آوری غیرقانونی داده توسط بسیاری از شرکت‌های رسانه‌های اجتماعی باشد. این داده‌ها یک نگرانی امنیتی شدید را، به خصوص در زمانی که افراد بسیاری تمایل دارند اطلاعات خود را ارائه دهند نشان می‌دهد. سوالاتی از دقت، انتشار، انقضا و دسترسی فراوان وجود دارند. برای مثال، ایالت مایلند اولین ایالتی است که بنا به قانون کارمندان را از پرسش رمزعبور فیس بوک و دیگر رسانه‌های اجتماعی در طول مصاحبه‌های شغلی و بعد از آن، منع کرد.

شرکت بین المللی داده (IDC) عبارت "سایه دیجیتال[17]" را برای انعکاس مقداری از داده جمع آوری شده، سازمان یافته و شاید تحلیل شده مربوط به یک فرد، برای تشکیل یک "تصویر" تجمعی از فرد، استفاده کرد. اطلاعاتی که در مورد شما وجود دارد بسیار بیشتر از اطلاعاتی است که شما در مورد خود ایجاد و یا منتشر می‌کنید. یک مسئله کلی این است که چقدر از این اطلاعات- اصلی یا بدست آمده – می‌خواهیم که خصوصی باقی بمانند؟

واضح است که، برخی از کلان داده‌ها باید با توجه به قوانین و مقررات امنیت و حفظ حریم شخصی امن بمانند. IDC پنج سطح افزایش امنیت را بیان کرده است [8]: حریم شخصی، سازگاری، نگهداری، محرمانگی و مستندسازی. پژوهش بیشتری برای تعریف شفاف این سطوح امنیت و نگاشت آن‌ها در برابر قوانین فعلی و تحلیل‌های جاری نیاز است. برای مثال، در فیس بوک، می‌توان صفحه را به "دوستان" خود محدود کرد. اما، اگر فیس بوک یک تحلیل را بر روی پایگاه داده خود برای استخراج همه پیوندهای دوستانه در یک گراف در حال بسط اجرا کند، در چه سطح امنیتی این تحلیل باید صورت گیرد؟ به عبارت دیگر، اگر فرد این دوستان را در سطوح امنیتی مشخصی، نشانه گذاری کرده باشند، چقدر از دوستان وی باید در تحلیل شرکت داشته باشند؟

2.8 ارزش "برخی از داده ها" در برابر "همه داده ها"

توجه داشته باشید که همه داده‌ها به صورت برابر ایجاد می‌شوند؛ برخی داده‌ها نسبت به داده‌های دیگر- از لحاظ زمانی، مفهومی و غیره، ارزشمندتر هستند. قبلا، محدودیت‌های فضای ذخیره نیازمند فیلتر کردن داده و تصمیم گیری بر این بود که چه داده‌هایی باید حفظ شوند. در گذشته، از آنچه که می‌خواستیم استفاده می‌کردیم و انچه که نمی‌خواستیم را دور می‌انداختیم.

[17] digital shadow

جدول 6: برخی چالش‌های سازگاری کلان داده

چه قواعد و مقرراتی باید با توجه به ترکیب داده از منابع متعدد در مورد افراد در یک مخزن واحد وجود داشته باشد؟
آیا قوانین سازگاری (مانند HIPAA) بر کل انبار داده اعمال می‌شوند یا فقط بخشی که حاوی داده‌های مرتبط هستند؟
چه قوانین و مقرراتی باید برای ممانعت از جمع آوری و ذخیره داده در مورد افراد – به صورت متمرکز یا توزیع شده- باید وجود داشته باشد؟
آیا یک تجمع داده باید در یک سطح بالاتر از عناصر تشکیل دهنده آن، امنیت داشته باشد؟
با توجه به گروه بندی امنیت IDC، چه درصدی از داده باید در هر گروه قرار گیرد؟ چه مکانیزم‌هایی به داده اجازه میدهند که بین گروه‌ها حرکت کند؟

مفهوم " محاسبات کمی و کیفی" بیانگر این است که ما برای تبدیل داده‌های نهان، متون ساختارنیافته، تصویر یا اطلاعات صوتی و تصویر به شاخص‌های عددی برای اینکه از لحاظ محاسباتی قابل ردیابی باشند، نیازمند مکانیزم‌هایی هستیم. با کلان داده و قابلیت‌های تحلیلی افزایش یافته ما، روند به سمت حفظ هر چیزی با این فرض که اهمیت تحلیلی در طول زمان افزایش می‌یابد؛ می‌رویم. به هر حال، در هر نقطه زمانی مقدار داده‌ای که نیاز است که به تصمیات خاصی تحلیل شوند تنها یک کسر بسیار کوچک از همه داده‌های در دسترس را در یک منبع داده نشان می‌دهند و بیشتر داده‌ها بدون تحلیل رها می‌شوند.

2.9 پردازش توزیع شده و داده توزیع شده

جذابیت تکرار سخت افزار و قابلیت ارتقای سیستم همانطور که توسط رایانش ابری همراه با سیستم‌های برنامه نویسی موازی MapReduce و واسط تبادل پیام[۱۸] (MPI) نشان داده شده راه حلی برای این چالش‌ها با استفاده از یک

[18] Message Passing Interface

رویکرد توزیع شده است. حتی با این رویکرد، تضعیف قابل توجه عملکرد می‌تواند به دلیل نیاز به ارتباط بین گره‌ها رخ دهد.

جدول 7: برخی از چالش‌های ارزش داده

برای دامنه یک مسئله مشخص، حداقل حجم داده مورد نیاز برای تشریح، تخمین، پیش بینی و تحلیل تحویزی و مدلسازی تصمیم با یک دقت خاص چیست؟
برای یک سرعت داده مشخص، چگونه حجم داده خود را برای تضمین دقت مستمر و پشتیبانی از پردازش (تقریبا) بلادرنگ به روزرسانی کنیم؟
برای یک دامنه مسئله مشخص، چه علم تحلیلی برای داده‌های غیر عددی ایجاد می‌شود؟
"اگر همه چیز را بدانیم؟" – کار بعدی که انجام می‌دهیم چیست؟

یک سوال پژوهشی باز این است که، کدام مشکل کلان داده "MapReducible" است؟ الگوریتم‌های توزیع شده خاص، ضرورتا بر اساس پارادایم‌های MPI یا MapReduce نیستند، ممکن است نیازمند وظایف تکمیلی برای حداقل سازی ارتباطات مورد نیاز بین گره‌ها باشند. سرانجام؛ اگر پردازش توزیع شده به عنوان یک گزینه دیده شود، قابلیت اعتماد کلی سیستم نیاز است که برای تضمین اینکه هیچ گره ساده‌ای یا ارتباط مورد نیازی با شکست مواجه نمی‌شود افزایش یابد. هردو سیستم‌های Google و Hadoop's MapReduce مراحل آغازینی را برای تضمین تحمل خطا برای سخت افزار، نرم افزار سیستم، و تا حدی، برای الگوریتم‌هایی در نرم افزاری کاربردی کاربر، اتخاذ می‌کنند.

3. پردازش کلان داده: چالش‌های تحلیلی

پردازش کلان داده یک چالش اصلی است، شاید چیزی بیشتر از مسائل مربوط به مدیریت و ذخیره سازی. انواع بسیاری از تحلیل‌ها وجود دارد: توصیفی، تخمینی، پیش بینانه، و تجویزی، که به انواع گوناگون مدل‌های تصمیم و بهینه سازی منجر می‌شود. برخی از تحلیل‌های کسب و کار مشترک در شکل 1 نشان داده شده اند. Kaisler [11] دیگر تجزیه

تحلیل تجزیه شده به 16 گروه بر اساس انواع مسائلی که باید بررسی شود را نشان داد، که شامل مدل‌های اقتصادی، نظریه بازی، نظریه کنترل، محاسبات تکاملی، و مدل‌های شبیه سازی است. مورد نرمال جدید تحلیل چابک، پیشرفته، پیش بینی کننده است که به راحتی با مجموعه داده‌ها و جریان‌های در حال تغییر منطبق می‌شود و اطلاعات و دانشی را برای بهبود خدمات و عملیات در سرار دولت، صنعت و دانشگاه حاصل می‌کند.

3.1 مقیاس بندی

یک مسئله مهم ایت است که آیا یک فرآیند تحلیلی با افزایش مجموعه داده با مرتبه بزرگی، بزرگ می‌شود. هر الگوریتم یک "زانو" دارد- نقطه‌ای که در آن عملکرد الگوریتم به صورت خطی با افزایش منابع محاسباتی افزایش می‌یابد و از نقطه اوج، پیک و... شروع به کاهش می‌کند. حل این مسئله نیازمند یک الگوریتم جدید برای مسئله، یا بازنویسی الگوریتم فعلی برای "تفسیر" زانو بیشتر از مقیاس است. یک سوال پژوهشی باز این است که آیا برای هر الگوریتم مشخص، محدودیتی در قابلیت مقیاس پذیری آن وجود دارد. این محدودیت‌ها برای الگوریتم‌های خاص با اجرای خاص بر ماشین‌های خاص در مقیاس‌های خاص شناخته شده هستند. راه حل‌های محاسباتی عمومی، به خصوص موارد استفاده کننده از داده‌های ساختارنیافته، هنوز شناخته شده نیستند. جدول 8 مثال‌هایی از رویکردهای تحلیلی ارائه می‌دهد که ممکن است در مقیاس خطی نباشند. به راحتی، پردازش کلان داده می‌تواند به یکی از سه روش نشان داده شده در جدول 9 مشخص شود.

شکل 1: مثال‌هایی از انواع تحلیل‌ها

جدول 8: مثال‌هایی از تحلیل‌هایی که ممکن است به مقیاس Zettabytes نرسند

تکنیک‌های یادگیری ماشین
تحلیل ویدیو/تصویر/متن ساختارنیافته
بصری سازی
رایانش ابری
داده کاوی
الگوریتم‌های مش و گراف
الگوریتم‌های مشترک در مجموعه‌های داده‌های ساختار یافته

شکل 2: سوزن در انبار کاه

جدول 9: پردازش کلان داده

هدف کشف و استخراج یک بخش مهم اطلاعات است که کاربر را قادر می‌کند بر شرایطی غلبه کند.	**1.** یافتن سوزن در انبار کاه
هیچ راه حل نهایی وجود ندارد، اما ترکیبی از راه حل‌ها سته به اینکه چگونه مسئله ارائه می‌شود وجود دارد. هدف انتخاب بهترین، و نه ضرورتا بهینه ترین، راه حل ممکن است.	2. تعمیم کاه به طلا
هدف همگرایی پاسخ همزمان با همگرایی سوال است- یک استراتژی outer-in، اما احتمال دارد که پاسخ‌های بسیاری را حاصل کند.	3. ترکیبی از تکنیک‌ها

3.2 یافتن سوزن در انبار کاه

این چالش بر یافتن داده‌های کلیدی تمرکز می‌کند که اهرمی برای تصمیم گیری در یک فضای مسئله است. یک مسئله سوزن در انبار کاه یکی از مواردی است که در ان پاسخ درست بسیار به سختی مشخص می‌شود، اما زمانی که بدانید که سوزن در کجا است، خیلی راحت پیدا می‌شود [5]. فرض کنید که ان مسئله را به عنوان یافتن یک پاسخ درست در یک مخزن از 1000000 پاسخ اشتباه مشخص می‌کنیم. اگر فرآیند تصمیم 0.1٪ مواقع اشتباه باشد، سپس از 100 پاسخ پیشنهادی پردازش شده؛ 101 پاسخ وجود دارد، که تنها یکی از آن‌ها "درست" است. همانطور که Felten اشاره کرد، هر عرصه پژوهشی بسته به این رویکرد از این مسئله رنج می‌برد، به خصوص اگر بر تحلیل آماری متکی باشد. تنها دو روش برای مقابله با این مسئله وجود دارد: کاهش اندازه پشته علف، یا بهبود رویه تصمیم گیری، جستجو و تحلیل.

3.3 تعمیم کاه به طلا

این چالش بر پردازش یک مجموعه بزرگ از نقاط داده گسسته به درون داده‌های به شدت ارزشمند تمرکز می‌کند. شکل 3 را در زیر در نظر بگیرید- یک بصری سازی داده Kenneth از دوستان فیس بوک Freeman در دسامبر 2011 است[7]. این زیر مجموعه کوچکی از صدهای میلیون یا افرادی که از فیس بوک استفاده می‌کنند را نشان می‌دهد. با افزایش حذف تعداد یال‌ها از گره‌های "اصلی"، پیچیدگی کلی مش به صورت غیرخطی افزایش می‌یابد. یافتن زیرگراف‌ها درون گراف‌ها با مجموعه‌های خاص ویژگی ممکن است یک مسئله به لحاظ محاسباتی قابل ردیابی با استفاده از الگوریتم‌های تحیلی و پیمایش گراف استاندارد باشد.

یک رویکرد برای حل مسائل نمایندگی متن نیمه ساختار یافته پراکنده و تبدیل آن به داده‌های پیوندی با استفاده از فرمت سه گانه چارچوب توصیف منابع[19] (RDF) است. انفجار متن نتیجه شده اغلب با توجه به استفاده از تگ‌های RDF برای شناسایی مولفه‌های یک ساختار RDF، دارای مرتبه 10:1 هستند. این مسئله را از پردازش متون نیمه

[19] Resource Description Framework

ساختار یافته به یافتن روابطی بر مش‌های بسیار بزرگ، جهان واقعی، تا حدی متصل، تفسیر می‌کند. استخراج ویژگی‌های ساختاری مش برای شناسایی الگوها و آنومالی‌ها مهم است. استنتاج در زیرساختارهای مش به "گناه جمعی" شباهت دارد، برای مثال، اگر یک فرد معتاد به مصرف مواد باشد، احتمال دارد که دوستان او نیز اینگونه باشند. باورهای منتشر شده در مش در انفجار داده بیشتر نتیجه می‌دهد [13].

چالش دیگر ماهیت متغیر زمان گراف‌های بسیار بزرگ است. گراف Freeman یک اسنپ شات است. تعیین تغییر بین دو اسنپ شات- به صورت ایستا یا به صورت پی در پی در فاصله معین- یک از نظر محاسباتی یک مسئله انفجاری است. این نوع مسئله غالبا رخ می‌دهد، اما در زمان اعمال در تحلیل‌های تقریبا بلادرنگ مانند نظارت بر حمله به شبکه، نیازمند الگویتم‌های جدید و محاسبات فشرده تر است.

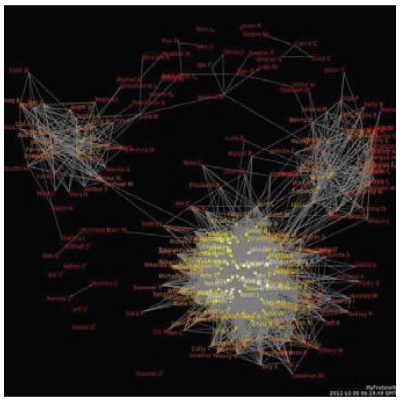
برای ما، واضح است که: استخراج طلا برابر داده کاوی نیست! الگوریتم‌های متفاوت با اعتماد بیشتر بر استدلال (یادگیری ماشین- سمبلیک نه اماری)، علوم اجتماعی محاسباتی، و تحلیل مبتنی بر دامنه برای دیدن "تصویر بزرگ" به منظور تفسیر و استخراج الگوهای عملی-رفتار، معانی و قطعات هوشی برای تصمیم گیری آگاهانه ضروری هستند.

3.4 یک ترکیب از تکنیک‌ها

با توجه به یک مجموعه داده ناهمگن بزرگ، یک چالش اصلی تصمیم گیری بر این است که چه داده‌ای باید تحلیل شود و چگونه باید تحلیل شود. برخلاف دو بخش قبلی، مجموعه داده‌های ترکیبی ترکیب شده از بسیاری از دیگر مجموعه داده‌ها نسبت به پاسخ‌های بلافصل، موارد جالب تری در خود دارند. برای تحلیل این داده‌ها نیازمند ادغام و پذیریش تکنیک‌های تحیلی متعدد برای ‘‘دیدن گوشه‌ها’’ هستیم، به عبارت دیگر، متوجه می‌شویم که دانش جدید احتمال دارد در یک روش غیرخطی بروز کند. واضح نیست که متدهای تحلیل آماری، همانطور که Ayres [2] استدلال کرد، کل پاسخ باشند.

Nassim Taleb [19] پتانسیلی را برای اثرات غیرمستقیم و پیش بینی نشده بروز کرده از رخدادهایی که منشا آن‌ها در خارج از قلمروی انتظارات معمول هستند را بررسی کرده اند، چرا که هیچ چیز در گذشته نمی‌تواند به صورت

متقاعد کننده‌ای به امکان پذیری آنها اشاره کند. همچین رخدادی اغلب یک تاثیر نهایی دارد – یک "شوک" برای سیستم که می‌تواند رفتار جدید را بوجود آورد. به دلیل اینکه همچین انتظاری نداریم، نمی‌توانیم آن را پیش بینی کنیم. بنابراین، تنها برای توضیح اینکه چه اتفاقی افتاده است تلاش می‌کنیم. با داده بیشتر، احتمال شناسایی همچین رویدادهایی بروز می‌کند و ما را بر آن می‌دارد که مجدد ابزارهای تحیلی پیش بینی کننده و تخمینی خود را ارزیابی کنیم.



شکل 3 : دوستان فیس بوکی Freeman

3.5 جهان را بشناسید

با فراوانی بیش از حد داده‌های در دسترس ما، یک سوال پژوهشی اصلی این است که : آیا می‌توانیم سیستم جهانی را، به منظور تمایل به مدل سازی و پیش بینی آب و هوا، مدل کنیم؟ برای مثال، آیا می‌توانیم پایداری اقتصادی یا جهانی را در یک بازه زمانی مشخص پیش بینی کنیم؟ اساس این چالش‌ها سوال‌هایی در مورد مدلسازی علوم طبیعی، تعاملات اجتماعی فرهنگی در مقیاس‌های گوناگون، درک نحوه عملکرد جوامع و علل ناآرامی جهانی؛ و درک اینکه چگونه دانشمندان انسانی منابع و جریان منابع در اطراف جهان را تولید و مصرف می‌کنند هستند. این سوال‌ها برای تعیین کسب و کار جهانی و در حال گذر به سمتی که منابع و سرمایه‌ای به زیرساخت‌های تخصیص داده می‌شود مهم هستند.

ایجاد مدل در یک کامپیوتر یک علم استاندارد است. اما، ایجاد مدل‌های تشکیل دهدنده جهان (یا حتی، محدود به دامنه) هنوز امکان پذیر نیست. با این وجود، چالش آغاز این مدل‌ها است که به ما اجازه می‌دهند که سیستم‌ها را در هردو دامنه و گرانولیته لازم برای پاسخ به سوالات اصلی علت و معلول درک کنیم.

اثرات بلابای طبیعی (مانند سونامی که بر تصمیمات ژاپن در زمینه انرژی هسته‌ای تاثیر گذاشت)، شکست در سیستم اقتصادی (بحران بانک و مسکن امریکا)، رکود / بحران و بهار عربی باعث تکامل عمده‌ای در دولت ها، یا نوآوری تکنولوژیکال می‌شود (مانند افزایش رسانه اجتماعی). سوال کلیدی برای بسیاری از تصمیم گیرندگان- تجاری، آکادمیک، دولتی- این است که " منظور از این موارد چیست؟" و سپس "احتمال دارد که بعدا چه اتفاقی رخ دهد؟". این‌ها همه "مسائل بدخیم[۲۰]" هستند که توسط Ritchey [16] تعریف شده است. یک Ritchey یکی از مواردی است که دارای پیش نیازهای ناقض، متناقض و اغلب در حال تغییر است [17]. به دلیل وابستگی متقابل پیچیده عناصر آن، اغلب تشخیص اینکه حتی یک راه حل جزئی بدست می‌اید سخت است. علاوه بر این، در زمان تلاش برای حل یک مسئله بدخیم، راه حل جزئی اغلب مسائل پیچیده تری را نشان می‌دهد یا ایجاد می‌کند. سیستم‌های اصلی ظهور می‌کنند، سیستم‌های تطبیقی به معنی این است که سیستم به صورت چشمگیری رفتار و توانایی تطبیق با شرایط جدید خود را تغییر می‌دهد. مدلسازی این نوع سیستم‌ها باید به صورت پیوسته‌ای به منظور پشتیبانی از آگاهی بخشی از شرایط عرصه گسترده تصمیم گیرنده تکامل یابد.

4. نتیجه گیری و کار آینده

کلان داده مرز علوم اجتماعی و کسب و کار "جدید" است. مقدار اطلاعات و دانشی که می‌تواند از جهان دیجیتال استخراج شود همچنان با آشنا شدن کاربران با روش‌های جدید ارسال پیام و پردازش داده، رو به گسترش است. علاوه بر این، واضح است که " داده بیشتر تنها به معنی داده بیشتر نیست"، بلکه "داده بیشتر به معنی چیزهای متفاوت تری است".

[20] wicked problems

"کلان داده" تنها آغاز مسئله است. تکامل و جایگزینی تکنولوژی تضمین می‌کند که در عرض چند سال داده بیشتری در یک سال در دسترس خواهد بود، داده‌ای بیشتر از آن چه از آغاز بشریت جمع اوری شده است. اگر فیس بوک و توییتر حدود 50 گیگابایت داده را در هر روز تولید کنند، جمع آوری کنند، و این حجم هر سال سه برابر شود، در طی چند سال (شاید 3-5) براستی با چالش " کلان داده واقعا کلان داده می‌شود" مواجه می‌شویم.

ما- به عنوان یک جامعه جهانی- از جوامع داده محور به سمت جوامع دانش محور تکامل می‌یابیم. دانش ما به صورت گسترده توزیع شده است و به همان اندازه قابل دسترسی است. یک برنامه بررسی این مسئله تعامل معنایی جامعه فمنیسم (SICoP) است که از مدل در حال تکامل پشتیبانی می‌کند: دولت شهروندی- سیستم‌هایی که می‌دانند، تحلیل پیشرفته- سیستم‌هایی که یاد می‌گیرند؛ و عملیات هوشمند – سیستم‌هایی که استدلال می‌کنند. این سیستم‌ها نیازمند کلان داده هستند. داده در یک یا چند محل ذخیره نمی‌شود؛ تنها در یک یا چند فرمت نیستند؛ تنها یک یا چند تحلیل بر آن‌ها اعمال نمی‌شود؛ و در میان عناصر داده متفاوت تنها یک یا چند پیوند برقرار نیست. بنابراین، این یک مثال از برخی از مسائلی است که در این مقاله بررسی می‌کنیم. حل این مسائل و چالش‌های بررسی شده در این مقاله نیازمند تلاش‌های پژوهشی متمرکز است- بر چیزی که انتظار داریم که در چند سال آینده تکامل یابد.

این مقاله یک تلاش پژوهشی همکارانه را برای آغاز بررسی مسائل و چالش‌های کلان داده آغاز می‌کند. برخی از مسائل اصلی را در فضای ذخیره، مدیریت و پردازش کلان داده شناسایی کردیم. برخی از چالش‌های اصلی- آینده – را شناسایی کردیم که باور داریم که باید در دهه بعد بررسی شوند و چارچوبی را برای minitrack کلان داده ما در جلسات HICSS آینده ایجاد می‌کند. پژوهش آینده ما بر توسعه یک درک کامل از مسائل مرتبط با کلان داده متمرکز است، و این فاکتورها ممکن است در نیاز به تحلیل کلان داده و طراحی متدلوژی سهم داشته باشند. ما بررسی راه حل‌های برخی از مسائلی که در این مقاله با تلاش‌های پژوهشی همکارانه بروز کرده است را آغاز می‌کنیم.

5. References

[1] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[2] Ayres, I. 2007. *Supercrunchers*, Bantam Books, New York, NY

[3] Boyd, D. and K. Craford. 2011. "Six Provocations for Big Data", Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society"

[4] The Economist. 2010. "Data, Data Everywhere", (online edition, February 28)

http://www.economist.com/node/15557443

[5] Felten, E. 2010. "Needle in a Haystack Problems",

https://freedom-to-tinker.com/blog/felten/needle-haystack-problems/

[6] Fox, B. 2011. "Leveraging Big Data for Big Impact", Health Management Technology, http://www.healthmgttech.com/

[7] Freeman, K. 2011. http://en.wikipedia.org/wiki/File:Kencf0618FacebookNetwork.jpg

[8] Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC

[9] Jacobs, A. 2009. "Pathologies of Big Data", *Communications of the ACM*, 52(8):36-44

[10] JASON. 2008. "Data Analysis Challenges", The Mitre Corporation, McLean, VA, JSR-08-142

[11] Kaisler, S. 2012. "Advanced Analytics", CATALYST Technical Report, i_SW Corporation, Arlington, VA

[12] Kaisler, S., W. Money, and S. J. Cohen. 2012. "A Decision Framework for Cloud Computing", *45th Hawaii International Conference on System Sciences*, Grand Wailea, Maui, HI, Jan 4-7, 2012

[13] Kang, U. 2012. "Mining Tera-scale Graphs with MapReduce: Theory, Engineering, and Discoveries", PhD. Thesis, Computer Science, Carnegie-Mellon University, Pittsburgh, PA

[14] Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012

[15] Popp, R., S. Kaisler, et al. 2006. "Assessing Nation-State Fragility and Instability", *IEEE Aerospace Conference*, 2006, Big Sky, MT

[16] Ritchey, T. 2005. "Wicked Problems: Structuring Social Messes with Morphological Analysis", Swedish Morphological Society, http://www.swemorph.com/wp.html

[17] Rittel, H. and M. Webber. *1973.* "Dilemmas in a General theory of Planning*", in Policy Sciences*, Vol. 4, Elsevier Scientific, Amsterdam, the Netherlands, pp. 155-169

[18] Stonebraker, M. and J. Hong. 2012. "Researchers' Big Data Crisis; Understanding Design and Functionality", Communications of the ACM, 55(2):10-11

[19] Taleb, N. 2010. *The Black Swan: The Impact of the Highly Improbable*, Random House, New York, NY