Primero

La palabra clave en "Data Science" no es Data, es Science

12 Dic 2013

Uno de mis colegas estaba en una conferencia donde vieron una presentación sobre el uso de datos para resolver un problema donde los datos no habían sido abundantes anteriormente. El orador afirmó que los datos eran "grandes datos" y una pregunta de la audiencia fue: "Bueno, eso no es realmente grandes datos es, es sólo X Gigabytes".

Si bien esa pregunta exacta provocaría gemidos de la mayoría de las personas que trabajan con datos, creo que destaca uno de los problemas clave con el pensamiento alrededor de la ciencia de los datos. La mayoría de las personas que se dedican a la ciencia de los datos se han centrado en la primera palabra: los datos. Ellos se preocupan por el volumen y la velocidad y cualquier otra palabra de moda describir los datos que es demasiado grande para que usted pueda analizar en Excel. Este bombo sobre el tamaño (relativo o absoluto) de los datos que se están recogiendo alimentado en la segunda categoría de bombo - hype sobre las herramientas. La gente lanzó alrededor de EC2, Hadoop, Pig, y tuvo enormes debates sobre Python versus R.

Pero la palabra clave en la ciencia de los datos no es "datos"; Es "ciencia". La ciencia de los datos sólo es útil cuando se utilizan los datos para responder a una pregunta. Esa es la parte de la ciencia de la ecuación. El problema con esta visión de la ciencia de datos es que es mucho más difícil que la visión que se centra en el tamaño de los datos o herramientas. Es mucho, mucho más fácil calcular el tamaño de un conjunto de datos y decir "Mis datos son más grandes que los tuyos" o decir, "puedo codificar en Hadoop, ¿no?" Que decir, "Tengo esta pregunta muy difícil , ¿Puedo responder con mis datos? ".

Algunas razones son más difíciles de enfocar en la ciencia que los datos / herramientas son:

La cita de John Tukey: "La combinación de algunos datos y un deseo dolorido pvVCBor una respuesta no asegura que una respuesta razonable pueda extraerse de un determinado cuerpo de datos". Usted puede tener 100 Gb y sólo 3 Kb son útiles para responder a la verdadera pregunta que le interesa.

Cuando empieza con la pregunta, a menudo descubre que necesita recopilar nuevos datos o diseñar un experimento para confirmar que está recibiendo la respuesta correcta.

Es fácil descubrir "estructura" o "redes" en un conjunto de datos. Siempre habrá correlaciones por mil razones si se recogen suficientes datos. Comprender si estas correlaciones son importantes para preguntas específicas e interesantes es mucho más difícil.

A menudo, la estructura que encontró en el primer paso se debe a un fenómeno (error de medición, artefactos, procesamiento de datos) que no responde a una pregunta interesante.

La cuestión es que el bombo en torno a la gran ciencia de datos / datos se apagará (ya lo es) si la ciencia de los datos se trata sólo de "datos" y no de "ciencia". El impacto a largo plazo de la ciencia de los datos se medirá por las preguntas científicas que podemos responder con los datos.

Segundo

Hace dieciocho meses, Li & Fung, una empresa que gestiona cadenas de suministro para los minoristas, vio 100 gigabytes de flujo de información a través de su red cada día. Ahora la cantidad se ha multiplicado por diez. Durante 2009, los aviones estadounidenses de aviones no tripulados sobrevolando Irak y Afganistán enviaron alrededor de 24 años de material de video. Los nuevos modelos desplegados este año producirán diez veces más flujos de datos que sus predecesores, y los de 2011 producirán 30 veces más.

Dondequiera que mires, la cantidad de información en el mundo está en alza. Según una estimación, la humanidad creó 150 exabytes (mil millones de gigabytes) de datos en 2005. Este año, creará 1.200 exabytes. Simplemente mantenerse al día con esta inundación, y almacenar los bits que podrían ser útiles, es bastante difícil. Analizarlo, detectar patrones y extraer información útil, es aún más difícil. Aún así, el diluvio de datos ya está empezando a transformar el negocio, el gobierno, la ciencia y la vida cotidiana (ver nuestro informe especial en este número). Tiene un gran potencial para el bien, siempre y cuando los consumidores, las empresas y los gobiernos tomen las decisiones correctas sobre cuándo restringir el flujo de datos y cuándo fomentarlo.

Algunas industrias han liderado el camino en su capacidad para recopilar y explotar datos. Las compañías de tarjetas de crédito supervisan cada compra y pueden identificar fraudulentas con un alto grado de precisión, utilizando reglas derivadas de crujir a través de miles de millones de transacciones. Tarjetas de crédito robadas son más propensos a ser utilizados para comprar licor duro que el vino, por ejemplo, porque es más fácil de cercar. Las firmas de seguros también son buenas en combinar pistas para detectar demandas sospechosas: las reclamaciones fraudulentas tienen más probabilidades de hacerse un lunes que un martes, ya que los asegurados que acaban con accidentes tienden a reunir a sus amigos como falsos testigos durante el fin de semana. Mediante la combinación de muchas de estas reglas, es posible determinar qué cartas son más probables haber sido robadas, y cuáles afirmaciones son poco fiables.

Los operadores de telefonía móvil, por su parte, analizan los patrones de llamadas de los suscriptores para determinar, por ejemplo, si la mayoría de sus contactos frecuentes se encuentran en una red rival. Si esa red rival está ofreciendo una promoción atractiva que podría hacer que el suscriptor de defecto, él o ella puede entonces ser ofrecido un incentivo para quedarse. Las industrias más viejas crujen datos con tanto entusiasmo como las nuevas en estos días. Los minoristas, tanto offline como en línea, son dueños de la minería de datos (o "business intelligence", como se conoce actualmente). Al analizar los "datos de la cesta", los supermercados pueden adaptar promociones a las preferencias de los clientes particulares. La industria petrolera utiliza supercomputadores para arrastrar datos sísmicos antes de perforar pozos. Y los astrónomos son tan propensos a señalar una herramienta de consulta de software en una encuesta de cielo digital como punto de un telescopio en las estrellas.

Hay mucho más por recorrer. A pesar de los años de esfuerzo, las bases de datos de las agencias de inteligencia y agencias de inteligencia no están vinculadas en gran medida. En la atención de la salud, la digitalización de los registros haría mucho más fácil detectar y vigilar las tendencias de la salud y evaluar la eficacia de los diferentes tratamientos. Pero los esfuerzos a gran escala para computarizar los registros sanitarios tienden a tener problemas burocráticos, técnicos y éticos. La publicidad en línea ya está mucho más dirigida que el tipo offline, pero hay margen para una personalización aún mayor. Los anunciantes entonces estarían dispuestos a pagar más, lo que a su vez significaría que los consumidores dispuestos a optar por tales cosas podrían ofrecer una gama más rica y más amplia de servicios gratuitos en línea. Y los gobiernos se acercan tardíamente a la idea de poner en el dominio público más información -como cifras de delitos, mapas, detalles de contratos gubernamentales o estadísticas sobre el desempeño de los servicios públicos-. La gente puede reutilizar esta información de maneras novedosas para construir negocios y hacer que los funcionarios electos tomen en cuenta. Las empresas que capten estas nuevas oportunidades, o provean las herramientas para que otros lo hagan, prosperarán. Business Intelligence es una de las partes de más rápido crecimiento de la industria del software.

Ahora las malas noticias

Pero el diluvio de datos también plantea riesgos. Los ejemplos abundan de bases de datos robadas: los discos llenos de datos de seguridad social desaparecen, las laptops cargadas con registros de impuestos se dejan en los taxis, los números de las tarjetas de crédito son robados a los minoristas en línea. El resultado son violaciones de privacidad, robo de identidad y fraude. Las infracciones de privacidad también son posibles incluso sin este juego sucio: son testigos de los problemas periódicos cuando Facebook o Google cambian inesperadamente la configuración de privacidad en sus redes sociales en línea, haciendo que los miembros revelen información personal involuntariamente. Una amenaza más siniestra proviene de la Gran Hermandad de varios tipos, particularmente cuando los gobiernos obligan a las empresas a entregar información personal sobre sus clientes. En lugar de poseer y controlar sus propios datos personales, las personas encuentran muy a menudo que han perdido el control de la misma.

La mejor manera de hacer frente a estos inconvenientes del diluvio de datos es, paradójicamente, hacer más datos disponibles de manera correcta, exigiendo mayor transparencia en varias áreas. En primer lugar, se debe dar a los usuarios mayor acceso y control sobre la información almacenada sobre ellos, incluyendo a quién se comparte. Google permite a los usuarios ver qué información contiene sobre ellos, y les permite eliminar sus historiales de búsqueda o modificar la orientación de la publicidad, por ejemplo. En segundo lugar, se debería exigir a las organizaciones que revelen detalles de las infracciones de seguridad, como ya ocurre en algunas partes del mundo, para alentar a los jefes a tomar más en serio la seguridad de la información. En tercer lugar, las organizaciones deben someterse a una auditoría anual de seguridad, con el grado resultante hecho público (aunque los detalles de cualquier problema expuesto no sería). Esto alentaría a las empresas a mantener actualizadas sus medidas de seguridad.

Los incentivos de mercado entrarán en juego ya que las organizaciones que manejan bien los datos son favorecidas sobre las que no. Una mayor transparencia en estas tres áreas mejoraría la seguridad y daría a las personas más control sobre sus datos sin la necesidad de una regulación intrincada que pudiera sofocar la innovación. Después de todo, el proceso de aprender a lidiar con el diluvio de datos, y trabajar la mejor manera de aprovecharlo, apenas ha comenzado.

Tercero

Los grandes datos se convertirán en una base clave de la competencia, apoyando las nuevas oleadas de crecimiento de la productividad, la innovación y el excedente del consumidor, siempre y cuando las políticas correctas y los facilitadores estén en su lugar.

La cantidad de datos en nuestro mundo ha estado explotando y el análisis de grandes conjuntos de datos -los llamados grandes datos- se convertirá en una base clave de la competencia, apoyando nuevas oleadas de crecimiento de la productividad, innovación y excedente del consumidor, según la investigación de MGI y Oficina de Tecnología Empresarial de McKinsey. Los líderes de todos los sectores tendrán que lidiar con las implicaciones de los grandes datos, no sólo con unos pocos administradores orientados a los datos. El creciente volumen y detalle de la información capturada por las empresas, el auge de los multimedia, las redes sociales y la Internet de las Cosas impulsarán un crecimiento exponencial de los datos en el futuro previsible.

Talento analítico profundo: ¿Dónde están ahora?

Las investigaciones realizadas por MGI y la Oficina de Tecnología Empresarial de McKinsey examinan el estado de los datos digitales y documentan el valor significativo que potencialmente puede desbloquearse.

MGI estudió grandes datos en cinco ámbitos: la asistencia sanitaria en los Estados Unidos, el sector público en Europa, el comercio minorista en los Estados Unidos y los datos de fabricación y datos personales en todo el mundo. Los datos grandes pueden generar valor en cada uno. Por ejemplo, un minorista que utilice grandes datos al máximo podría aumentar su margen operativo en más del 60 por ciento. Aprovechar grandes datos en el sector público también tiene un enorme potencial. Si la atención sanitaria estadounidense utilizara los datos grandes de forma creativa y eficaz para impulsar la eficiencia y la calidad, el sector podría crear más de 300.000 millones de dólares en valor cada año. Dos tercios de esto estaría en la forma de reducir los gastos de salud de los Estados Unidos en un 8 por ciento. En las economías desarrolladas de Europa, los administradores gubernamentales podrían ahorrar más de 100.000 millones de euros (149.000 millones de dólares) en mejoras de la eficiencia operativa solo utilizando grandes datos, sin utilizar grandes datos para reducir el fraude y los errores y aumentar la recaudación de impuestos. Y los usuarios de servicios habilitados por datos de ubicación personal podrían capturar 600 mil millones de dólares en excedentes de consumo. La investigación ofrece siete ideas clave.

1. Los datos han llegado a todas las industrias y funciones empresariales y ahora son un importante factor de producción, junto con el trabajo y el capital. Estimamos que en 2009 casi todos los sectores de la economía estadounidense tenían por lo menos un promedio de 200 terabytes de datos almacenados (el doble del tamaño del almacén de datos Wal-Mart en 1999) por empresa con más de 1.000 empleados.
2. 2. Hay cinco maneras amplias en que el uso de datos grandes puede crear valor. Primero, los datos grandes pueden desbloquear un valor significativo al hacer que la información sea transparente y utilizable a una frecuencia mucho mayor. En segundo lugar, a medida que las organizaciones crean y almacenan más datos transaccionales en forma digital, pueden recopilar información de rendimiento más precisa y detallada sobre todo, desde inventarios de productos hasta días enfermos y, por tanto, exponer la variabilidad y aumentar el rendimiento. Las principales empresas están utilizando la recopilación y el análisis de datos para llevar a cabo experimentos controlados para tomar mejores decisiones de gestión; Otros están utilizando los datos para la previsión básica de baja frecuencia a alta frecuencia de predicción inmediata para ajustar sus palancas de negocio justo a tiempo. En tercer lugar, los grandes datos permiten una segmentación cada vez más estrecha de los clientes y, por lo tanto, productos o servicios mucho más precisos. En cuarto lugar, una analítica sofisticada puede mejorar sustancialmente la toma de decisiones. Por último, se pueden utilizar grandes datos para mejorar el desarrollo de la próxima generación de productos y servicios. Por ejemplo, los fabricantes están utilizando los datos obtenidos de los sensores incorporados en los productos para crear innovadoras ofertas de servicio post-venta, tales como el mantenimiento proactivo (medidas preventivas que tienen lugar antes de que se produzca un fallo o incluso se nota).
3. 3. El uso de grandes datos se convertirá en una base clave para la competencia y el crecimiento de las empresas individuales. Desde el punto de vista de la competitividad y la captura potencial de valor, todas las empresas necesitan tomar en serio los datos importantes. En la mayoría de las industrias, los competidores establecidos y los nuevos participantes aprovecharán las estrategias basadas en datos para innovar, competir y capturar el valor de la información profunda y en tiempo real. De hecho, hemos encontrado ejemplos tempranos de tal uso de datos en cada sector que examinamos.
4. 4. El uso de grandes datos servirá de base a nuevas oleadas de crecimiento de la productividad y excedentes de los consumidores. Por ejemplo, estimamos que un minorista que utilice grandes datos al máximo tiene el potencial de aumentar su margen de operación en más del 60 por ciento. Los grandes datos ofrecen considerables beneficios tanto para los consumidores como para las empresas y organizaciones. Por ejemplo, los servicios habilitados por datos de ubicación personal pueden permitir a los consumidores capturar $ 600 mil millones en excedente económico.
5. 5. Si bien el uso de grandes datos importará a través de sectores, algunos sectores están dispuestos a mayores ganancias. Se comparó la productividad histórica de los sectores en los Estados Unidos con el potencial de estos sectores para capturar el valor de los grandes datos (utilizando un índice que combina varias métricas cuantitativas) y encontró que las oportunidades y desafíos varían de un sector a otro. Los sectores informático y de los productos electrónicos y de la información, así como las finanzas y los seguros, y el gobierno están preparados para ganar sustancialmente con el uso de grandes datos.
6. 6. Habrá una escasez de talento necesario para que las organizaciones aprovechen los grandes datos. Para el año 2018, sólo los Estados Unidos podrían enfrentar una escasez de 140,000 a 190,000 personas con profundas habilidades analíticas así como 1,5 millones de gerentes y analistas con el conocimiento para usar el análisis de datos grandes para tomar decisiones efectivas.
7. 7. Habrá que abordar varias cuestiones para captar todo el potencial de los grandes datos. Las políticas relacionadas con la privacidad, la seguridad, la propiedad intelectual e incluso la responsabilidad deberán abordarse en un gran mundo de datos. Las organizaciones no sólo tienen que poner el talento y la tecnología adecuados sino también estructurar los flujos de trabajo e incentivos para optimizar el uso de los grandes datos. El acceso a los datos es fundamental: las empresas tendrán cada vez más que integrar la información de múltiples fuentes de datos, a menudo de terceros, y los incentivos deben estar en su lugar para permitir esto.

Cuarto

***For Today’s Graduate, Just One Word: Statistics***

En Harvard, Carrie Grimes se especializó en antropología y arqueología y se aventuró a lugares como Honduras, donde estudió los patrones de asentamientos mayas mediante la cartografía donde se encontraron artefactos. Pero ella se sintió atraída por lo que ella llama "todas las cosas de computación y matemáticas" que formaba parte del trabajo.

"La gente piensa en la arqueología de campo como Indiana Jones, pero gran parte de lo que realmente haces es el análisis de datos", dijo.

Ahora la Sra. Grimes hace otro tipo de excavación. Ella trabaja en Google, donde utiliza el análisis estadístico de montones de datos para llegar a formas de mejorar su motor de búsqueda.

La Sra. Grimes es una estadística de la Internet-edad, una de muchas que están cambiando la imagen de la profesión como un lugar para los nerds número dronish. Se están encontrando cada vez más en la demanda - e incluso fresco.

Continuar leyendo la historia principal

"Sigo diciendo que el trabajo sexy en los próximos 10 años serán estadísticos", dijo Hal Varian, economista jefe de Google. Y no estoy bromeando.

La creciente estatura de los estadísticos, que pueden ganar 125.000 dólares en las mejores empresas en su primer año después de obtener un doctorado, es un subproducto de la reciente explosión de datos digitales. En campo tras campo, la informática y la Web están creando nuevos reinos de datos para explorar - señales de sensores, cintas de vigilancia, charlas de redes sociales, registros públicos y más. Y el aumento de los datos digitales sólo promete acelerar, quintuplicando para 2012, según una proyección de IDC, una firma de investigación.

Sin embargo, los datos son meramente la materia prima del conocimiento. "Estamos entrando rápidamente en un mundo donde todo puede ser monitoreado y medido", dijo Erik Brynjolfsson, economista y director del Centro de Negocios Digitales del Instituto de Tecnología de Massachusetts. "Pero el gran problema va a ser la capacidad de los seres humanos para usar, analizar y dar sentido a los datos".

La nueva generación de estadísticos aborda ese problema. Ellos usan computadoras de gran alcance y modelos matemáticos sofisticados para buscar patrones y percepciones significativas en vastos troves de datos. Las aplicaciones son tan diversas como mejorar la búsqueda en Internet y la publicidad en línea, eliminar la información de secuenciación de genes para la investigación del cáncer y analizar los datos de sensores y localización para optimizar el manejo de envíos de alimentos.

Incluso el recién terminado concurso Netflix, que ofreció $ 1 millón a cualquiera que pudiera mejorar significativamente el sistema de recomendación de películas de la compañía, fue una batalla librada con las armas de las estadísticas modernas.

Aunque en el primer plano, los estadísticos son sólo una pequeña parte de un ejército de expertos que utilizan técnicas estadísticas modernas para el análisis de datos. La informática y las habilidades numéricas, dicen los expertos, importan mucho más que los grados. Así que los nuevos detectives de datos vienen de antecedentes como la economía, la informática y las matemáticas.

Sin duda son bienvenidos en la Casa Blanca en estos días. Peter R. Orszag, director de la Oficina de Gestión y Presupuesto, declaró en un discurso en mayo que "datos robustos e imparciales son el primer paso para abordar nuestras necesidades económicas a largo plazo y nuestras principales prioridades políticas". Más tarde ese mismo día, el Sr. Orszag confesó en una entrada en el blog que su charla sobre la importancia de las estadísticas era un tema "cercano a mi corazón (que ciertamente era un poco raro").

I.B.M., viendo una oportunidad en servicios de caza de datos, creó un grupo de Business Analytics y Servicios de Optimización en abril. La unidad aprovechará la experiencia de los más de 200 matemáticos, estadísticos y otros analistas de datos en sus laboratorios de investigación, pero ese número no es suficiente. I.B.M. planea reciclar o contratar a 4.000 analistas más en toda la compañía.

En otro signo del creciente interés en el campo, unas 6.400 personas asisten a la conferencia anual de la profesión de estadísticas en Washington esta semana, frente a 5.400 en los últimos años, según la Asociación Estadounidense de Estadística. Los asistentes, hombres y mujeres, jóvenes y encanecidos, se parecían mucho a cualquier otra multitud de turistas en la capital de la nación. Pero sus intercambios extenuantes estaban llenos de conversaciones sobre asignación al azar, parámetros, regresiones y clusters de datos. El aumento de los datos es elevar una profesión que tradicionalmente abordó el trabajo menos visible y menos lucrativo, como calcular la esperanza de vida de las compañías de seguros.

La Sra. Grimes, de 32 años, obtuvo su doctorado en estadísticas de Stanford en 2003 y se unió a Google más tarde ese año. Ahora es una de las muchas estadísticas en un grupo de 250 analistas de datos. Ella utiliza el modelado estadístico para ayudar a mejorar la tecnología de búsqueda de la compañía.

Por ejemplo, la Sra. Grimes trabajó en un algoritmo para afinar el software de rastreo de Google, que recorre la Web para actualizar constantemente su índice de búsqueda. El modelo aumentó las posibilidades de que el rastreador explorara páginas web actualizadas con frecuencia y hiciera menos viajes a los más estáticos.

El objetivo, explicó la Sra. Grimes, es hacer pequeñas ganancias en la eficiencia del uso de computadoras y redes. "Incluso una mejora de un por ciento o dos puede ser enorme, cuando haces cosas sobre los millones y miles de millones de veces que hacemos cosas en Google", dijo.

Es el tamaño de los conjuntos de datos en la Web que abre nuevos mundos de descubrimiento. Tradicionalmente, las ciencias sociales rastreaban el comportamiento de las personas entrevistándolas o examinándolas. "Pero la Web proporciona este increíble recurso para observar cómo millones de personas interactúan", dijo Jon Kleinberg, un científico de la computación y un investigador de redes sociales en Cornell.

Por ejemplo, en la investigación que acaba de publicarse, Kleinberg y dos colegas siguieron el flujo de ideas a través del ciberespacio. Ellos rastrearon 1,6 millones de sitios de noticias y blogs durante la campaña presidencial de 2008, utilizando algoritmos que buscaban frases asociadas con temas de noticias como "lápiz de labios en un cerdo".

Los investigadores de Cornell encontraron que, en general, los medios de comunicación tradicionales conducen y los blogs siguen, por lo general en 2,5 horas. Pero un puñado de blogs fueron más rápidos a citas que más tarde ganó gran atención.

El rico hilo de datos web, advierten los expertos, tiene sus peligros. Su volumen escarpado puede fácilmente superar modelos estadísticos. Los estadísticos también advierten que las fuertes correlaciones de datos no necesariamente demuestran un vínculo de causa y efecto.

Por ejemplo, a fines de la década de 1940, antes de que hubiera una vacuna contra la poliomielitis, expertos en salud pública en América señalaron que los casos de poliomielitis aumentaron a medida que se consumía helado y refrescos, según David Alan Grier, historiador y estadístico de George Washington Universidad. La eliminación de tales golosinas fue incluso recomendada como parte de una dieta anti-polio. Resultó que los brotes de poliomielitis eran más comunes en los meses calurosos del verano, cuando la gente naturalmente comía más helado, mostrando sólo una asociación, dijo Grier.

Si la explosión de los datos amplía los problemas de estadística de muchos años, también abre nuevas fronteras.

"La clave es permitir que las computadoras hagan lo que son buenos, lo que está arrastrando estos enormes conjuntos de datos para algo que es matemáticamente extraño", dijo Daniel Gruhl, investigador de I.B.M, cuyo trabajo reciente incluye la extracción de datos médicos para mejorar el tratamiento. "Y eso facilita que los seres humanos hagan lo que son buenos - explican esas anomalías".

Quinto

Para algunas personas R es sólo la letra decimoctava del alfabeto. Para otros, es la calificación de las películas raciales, una medida del aislamiento de un ático o lo que dicen los piratas en las películas.

R es también el nombre de un popular lenguaje de programación utilizado por un número creciente de analistas de datos dentro de las corporaciones y el mundo académico. Se está convirtiendo en su lingua franca en parte porque la minería de datos ha entrado en una edad de oro, ya sea que se utiliza para fijar los precios de anuncios, encontrar nuevos medicamentos más rápidamente o afinar los modelos financieros. Empresas tan diversas como Google, Pfizer, Merck, Bank of America, InterContinental Hotels Group y Shell lo utilizan.

Pero R también ha encontrado rápidamente un seguimiento porque los estadísticos, los ingenieros y los científicos sin conocimientos de programación informática lo encuentran fácil de usar.

"R es realmente importante hasta el punto de que es difícil de sobrevalorar", dijo Daryl Pregibon, un científico investigador de Google, que utiliza el software ampliamente. "Permite a los estadísticos realizar análisis complejos y complicados sin conocer la sangre y las entrañas de los sistemas informáticos".

También es gratis. R es un programa de código abierto, y su popularidad refleja un cambio en el tipo de software utilizado en las corporaciones. El software de código abierto es gratuito para que cualquiera pueda usarlo y modificarlo. I.B.M., Hewlett-Packard y Dell hacen miles de millones de dólares al año vendiendo servidores que ejecutan el sistema operativo Linux de código abierto, que compite con Windows de Microsoft. La mayoría de los sitios web se muestran utilizando una aplicación de código abierto llamada Apache, y las empresas cada vez más dependen de la base de datos de código abierto de MySQL para almacenar su información crítica. Muchas personas ven los resultados finales de toda esta tecnología a través del navegador web de Firefox, también software de código abierto.

R es similar a otros lenguajes de programación, como C, Java y Perl, ya que ayuda a las personas a realizar una amplia variedad de tareas de computación dándoles acceso a varios comandos. Para los estadísticos, sin embargo, R es particularmente útil porque contiene una serie de mecanismos integrados para organizar datos, ejecutar cálculos sobre la información y crear representaciones gráficas de conjuntos de datos.

Algunas personas familiarizadas con R lo describen como una versión sobrecargada del software de hoja de cálculo de Microsoft Excel que puede ayudar a iluminar las tendencias de datos más claramente de lo que es posible mediante la introducción de información en filas y columnas.

Lo que hace R tan útil -y ayuda a explicar su rápida aceptación- es que los estadísticos, ingenieros y científicos pueden mejorar el código del software o escribir variaciones para tareas específicas. Los paquetes escritos para R añaden algoritmos avanzados, gráficos de colores y texturas y técnicas de minería para profundizar en bases de datos.

Cerca de 1.600 paquetes diferentes residen en uno de los muchos sitios web dedicados a R, y el número de paquetes ha crecido exponencialmente. Un paquete, llamado BiodiversityR, ofrece una interfaz gráfica para facilitar los cálculos de tendencias ambientales.

Otro paquete, llamado Emu, analiza patrones de habla, mientras que GenABEL se utiliza para estudiar el genoma humano.

La comunidad de servicios financieros ha demostrado una afinidad particular por R; Existen docenas de paquetes para análisis de derivados solamente.

"La gran belleza de R es que se puede modificar para hacer todo tipo de cosas", dijo Hal Varian, economista jefe de Google. "Y tienes un montón de cosas preenvasadas que ya están disponibles, por lo que estás de pie sobre los hombros de gigantes."

R apareció por primera vez en 1996, cuando los profesores de estadística Ross Ihaka y Robert Gentleman de la Universidad de Auckland en Nueva Zelanda publicaron el código como un paquete de software libre.

Según ellos, la noción de idear algo como R surgió durante una conversación en el pasillo. Ambos querían tecnología mejor adaptada para sus estudiantes de estadística, que necesitaban analizar datos y producir modelos gráficos de la información. La mayor parte del software comparable había sido diseñado por científicos de la computación y resultó difícil de usar.

Al carecer de un profundo entrenamiento en ciencias de la computación, los profesores consideraban que sus esfuerzos de codificación eran más un juego académico que cualquier otra cosa. Sin embargo, a partir de 1991, trabajaron a tiempo completo. "Estuvimos prácticamente inseparables durante cinco o seis años", dijo el señor Gentleman. "Una persona haría el mecanografiar y una persona haría el pensamiento."

Algunos estadísticos que tomaron una mirada temprana en el software lo consideraron áspero alrededor de los bordes. Pero a pesar de sus deficiencias, R inmediatamente ganó un seguimiento con personas que vieron las posibilidades en la personalización del software libre.

John M. Chambers, un ex investigador de Bell Labs que ahora es profesor consultor de estadísticas en la Universidad de Stanford, fue uno de los primeros campeones. En Bell Labs, el Sr. Chambers había ayudado a desarrollar S, otro proyecto de software de estadísticas, que tenía por objeto dar a los investigadores de todas las franjas una herramienta de análisis de datos accesible. Sin embargo, no fue un proyecto de código abierto.

El software no generó interés general y, en última instancia, los derechos de S terminaron en manos de Tibco Software. Ahora R está superando lo que Mr. Chambers había imaginado posible con S.

"La diversidad y la emoción en torno a lo que todas estas personas están haciendo es grande", dijo Chambers.

Si bien es difícil calcular exactamente cuántas personas utilizan R, los más familiarizados con el software estiman que cerca de 250.000 personas trabajan con él regularmente. La popularidad de R en las universidades podría amenazar SAS Institute, la empresa privada de software de negocios que se especializa en software de análisis de datos. SAS, con más de $ 2 mil millones en ingresos anuales, ha sido la herramienta preferida de académicos y gerentes corporativos.

"R realmente se ha convertido en el segundo idioma para las personas que salen de la escuela de postgrado ahora, y hay una cantidad increíble de código que está siendo escrito para él", dijo Max Kuhn, director asociado de estadísticas no clínicas de Pfizer. "Usted puede mirar en los tableros del mensaje de SAS y ver allí es una disminución proporcional en tráfico."

SAS dice que ha notado la creciente popularidad de R en las universidades, a pesar de los descuentos educativos en su propio software, pero descarta la tecnología como de interés para un grupo limitado de personas que trabajan en tareas muy difíciles.

"Creo que se dirige a un nicho de mercado para los analistas de datos de gama alta que quieren código libre y fácilmente disponible", dijo Anne H. Milley, directora de marketing de productos tecnológicos de SAS y agregó: "Tenemos clientes que construyen motores para aviones. Estoy feliz de que no están usando freeware cuando me meto en un jet ".

Pero mientras SAS rechaza el atractivo corporativo de R, compañías como Google y Pfizer dicen que usan el software para casi cualquier cosa que puedan. Google, por ejemplo, aprovecha R para ayudar a entender las tendencias en los precios de los anuncios y para iluminar los patrones de los datos de búsqueda que recoge. Pfizer ha creado paquetes personalizados para que R permita a sus científicos manipular sus propios datos durante estudios de drogas no clínicas en lugar de enviar la información a un estadístico.

Los co-creadores de R expresan satisfacción de que estas empresas se beneficien de los frutos de su trabajo y el de cientos de voluntarios.

El Sr. Ihaka continúa enseñando estadísticas en la Universidad de Auckland y quiere crear software más avanzado. El Sr. Gentleman está aplicando el software basado en R, llamado Bioconductor, en el trabajo que está haciendo en biología computacional en el Fred Hutchinson Cancer Research Center en Seattle.

"R es una demostración real del poder de la colaboración, y no creo que puedas construir algo como esto de otra manera", dijo Ihaka. "Podríamos haber elegido ser comerciales, y habríamos vendido cinco copias del software".

<https://github.com/GuilleTlapale/Programacion_Actuarial_III/blob/master/Historial%20Academico.Rmd>

08f6ee21f30e2a085393718f94621313a2773b7c