Desafíos políticos y oportunidades de investigación en la era de los macrodatos

**[SP Kothari](https://www.sec.gov/biography/sp-kothari)*[Economista jefe y director de la División de Análisis Económico y de Riesgos](https://www.sec.gov/biography/sp-kothari)***

**Big Data y computación de alto rendimiento para la economía financiera, Oficina Nacional de Investigación Económica, Cambridge, MA**

**13 de julio de 2019**

Me complace tener la oportunidad de hablar en esta conferencia de la Oficina Nacional de Investigación Económica (NBER) sobre big data y computación de alto rendimiento. Antes de comenzar mis observaciones, debo mencionar que las opiniones que expreso hoy son mías y no reflejan necesariamente las opiniones de la Comisión o de su personal. [[1]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn1" \o ")

**1. Historia introductoria**

El término "big data" es nuevo, pero el fenómeno subyacente es todo menos nuevo y ciertamente no es exclusivo de la economía financiera. Considere, por ejemplo, el censo de los Estados Unidos, que se realiza cada diez años según lo exige la Constitución de los Estados Unidos. Es una tarea aparentemente simple contar personas y reportar información demográfica como el estado civil y el tamaño de la familia. Sin embargo, en 1870, la población estadounidense en rápida expansión obstaculizó la capacidad de la Oficina del Censo para tabular los resultados de manera efectiva. De hecho, el censo de 1880, que se contó a mano, tardó casi diez años en completarse. En otras palabras, el censo de 1880 involucró big data. Herman Hollerith vio la oportunidad y dejó la Oficina del Censo antes del censo de 1880 para desarrollar una máquina que pudiera contar y tabular los resultados. Su máquina fue probada en 1887, y la Oficina del Censo lo arrendó rápidamente para el censo de 1890. Su éxito en 1890 condujo a contratos con gobiernos extranjeros y empresas privadas. Las máquinas Hollerith se utilizaron en 1891 para los censos de Canadá, Noruega y Austria; las compañías ferroviarias los utilizaron para calcular la información de tarifas; etc. En otras palabras, las máquinas Hollerith resolvieron de manera eficiente muchos problemas importantes de big data del día.[[2]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn2" \o ")

Hoy, 150 años después, ¿dónde nos encontramos? Nos encontramos sobre montañas de datos que son inconcebiblemente más grandes. Según algunas estimaciones, el mundo genera más datos cada dos días que toda la humanidad generada desde los albores de los tiempos hasta el año 2003. [[3]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn3" \o ") ¿Cuántos datos se generan por o para la SEC? Una respuesta fácil es que el sistema de Recopilación, Análisis y Recuperación de Datos Electrónicos (o EDGAR) de la SEC recibe y procesa alrededor de 2 millones de presentaciones al año. Pero esas presentaciones son en sí mismas documentos complejos, muchos de los cuales contienen decenas de páginas, numerosos archivos adjuntos y muchos miles de piezas de información.

¿Qué es Big Data? Creo que es algo así como la vejez: cualquier persona mayor que yo es mayor, y cualquier conjunto de datos más grande de lo que mi sistema informático puede procesar es grande. ¿Qué significa "grande" para la SEC? La SEC procesa y mantiene varios grandes conjuntos de datos. Un ejemplo son los datos de la Autoridad de informes de precios de opciones o los datos de OPRA. Los datos de OPRA para un día son aproximadamente dos terabytes. [[4]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn4" \o ")

Los macrodatos a menudo se caracterizan por las llamadas "tres v", que son *volumen* , *velocidad* y *variedad* .

***El volumen*** es la cantidad de datos.

***La velocidad*** es la velocidad a la que se crean y almacenan los datos.

***La variedad*** es la heterogeneidad de los datos en términos de tipo y formato de datos.

A esta lista de tres, algunos agregarían una cuarta "v", *veracidad* .

***La veracidad*** es la calidad y precisión de los datos.

**2. Desafíos políticos**

Al igual que la Oficina del Censo hace 150 años, la SEC enfrenta hoy un problema de big data. Esto me lleva a la primera pregunta que quiero resaltar en esta charla: *¿Cuáles son los desafíos políticos que se derivan del big data en la SEC?*

Permítanme comenzar recordándoles que la misión de la SEC es (1) proteger a los inversionistas; (2) mantener un mercado justo, ordenado y eficiente; y (3) facilitar la formación de capital. Veo varios desafíos de política de big data a la luz de la triple misión de la SEC.

**Seguridad**

Permítanme comenzar con la seguridad, que es una de las principales preocupaciones de la SEC. El *volumen* , la *velocidad* y la *variedad.*de los macrodatos hacen que la seguridad sea particularmente desafiante por varias razones. Primero, los macrodatos son más difíciles de almacenar y mantener. Por ejemplo, es más difícil garantizar que solo las personas adecuadas en el momento adecuado tengan acceso solo a los datos correctos. En segundo lugar, los macrodatos son objetivos más importantes para los malos actores. Por ejemplo, los datos de tenencias de cartera de todos los asesores de inversión son más valiosos que los datos de tenencias de cartera de un asesor de inversiones, y los datos de tenencias de cartera semanales son más valiosos que los datos de tenencias de cartera anuales. Estos desafíos se vuelven más difíciles a medida que ciertos conjuntos de datos comienzan a incluir más información de identificación personal (PII) o identificadores que vinculan a los inversores e instituciones dentro y entre conjuntos de datos.

La SEC debe tener en cuenta los datos que recopila y su naturaleza sensible, y la SEC debe ser un usuario responsable y con principios de esos datos. Naturalmente, la recopilación de datos no es un fin en sí mismo: la SEC no debe estar en el negocio del almacenamiento de datos indefinido y mal definido. Por estas razones, la SEC continúa investigando si puede reducir los datos que recopila o reducir su sensibilidad. Un ejemplo de esto es el enfoque de la SEC para el Formulario N-PORT, que es un nuevo formulario para informar a la SEC sobre tenencias de carteras de fondos públicos y no públicos. La Comisión modificó recientemente los plazos de presentación de esta información con el fin de reducir el volumen de información sensible en poder de la SEC. Este simple cambio redujo el perfil de riesgo cibernético de la SEC sin afectar el momento o la cantidad de información que se pone a disposición del público.[[5]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn5" \o ")

**Tecnología**

Otro desafío político es la tecnología. Por ejemplo, las ganancias comerciales potenciales de tener sistemas informáticos y otras tecnologías que son incluso un poco más rápidas e inteligentes que la competencia son enormes. Por lo tanto, existe una carrera de armamentos tecnológicos entre empresas comerciales que se esfuerzan por obtener la mejor tecnología y el mejor personal. Los medios informan periódicamente sobre instituciones que están aumentando el uso de inteligencia artificial, aprendizaje automático y herramientas relacionadas. [[6]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn6" \o ") Sin embargo, puede haber costos fijos para el despliegue de estas tecnologías que excluyen a los inversores pequeños, fragmentados o con menos recursos.

En segundo lugar, existen diferencias culturales entre las organizaciones que afectan no solo la elección de qué tecnología implementar, sino también el momento de la implementación. Por ejemplo, los fondos de cobertura podrían adoptar nuevas tecnologías, como la computación en la nube, más rápidamente que los fondos de pensiones.

En tercer lugar, algunas tecnologías son intrínsecamente difíciles de monitorear para la SEC. Para mencionar solo un ejemplo, considere el comercio algorítmico artificialmente inteligente (comercio de algoritmos de IA), que se negocia a través del tiempo de formas no predecibles. Supongamos que un algoritmo de IA finalmente comienza a suplantar sin el conocimiento del creador del algoritmo. (La suplantación de identidad es una actividad prohibida que implica la creación y cancelación de una gran cantidad de operaciones en un intento de transmitir información falsa sobre la demanda del mercado). ¿Cómo debería responder la SEC a eso?

Y hablando de tecnología de rápido movimiento, ¿cómo desarrolla o atrae la SEC una fuerza laboral que no solo ve y comprende el estado actual de la técnica, sino que también puede visualizar y prepararse para el futuro? La SEC ha priorizado y respaldado el desarrollo de una fuerza laboral con habilidades y experiencia en big data. Durante los últimos 10 años, la plantilla de DERA ha crecido de poco más de 30 personas a casi 150 personas en la actualidad.

**Comunicación**

Otro desafío de la política de big data es la comunicación porque la SEC tiene diversas partes interesadas. La SEC se centra en los inversores de "Main Street", es decir, inversores minoristas individuales que normalmente invierten a través de sus planes tipo 401 (k). Pero nuestros accionistas también incluyen fondos de pensiones, emisores de bonos municipales, firmas de corretaje, fondos de cobertura y el Congreso. Los problemas que rodean a los macrodatos son complejos y cada vez más requieren una formación especializada para comprenderlos. Por lo tanto, es un desafío comunicar las partes esenciales de estos mercados a cada grupo de partes interesadas. De hecho, una talla NO sirve para todos.

Mientras hablo de comunicación, me gustaría mencionar un detalle importante sobre la historia de Herman Hollerith. Una idea clave del problema de los datos del censo fue la comprensión de que la *variedad* de los datos podría reducirse drásticamente al requerir que los datos se transcriban en lo que ahora llamaríamos tarjetas perforadas. Con todos los datos en una forma estandarizada, fue relativamente fácil construir una máquina que pudiera tabular la información. Este principio sigue siendo válido en la actualidad. Por ejemplo, la SEC ha requerido que los contribuyentes etiqueten algunos datos utilizando métodos como XML, FIX, FpML, XBRL y, más recientemente, Inline XBRL. Reduciendo drásticamente la *variedad*de los datos, el etiquetado hace que un documento electrónico pase de ser legible por humanos a uno que también sea legible por máquina. Un desafío permanente de la SEC es encontrar formas rentables de reducir la *variedad* de datos financieros sin pérdida de información sustancial.

Una característica adicional del etiquetado de datos son los efectos de red. Es bien sabido que los datos de los 10-K etiquetados se pueden vincular a los datos de otros formularios y otras empresas. Quizás sea menos apreciado que los datos en los documentos etiquetados puedan estar vinculados a través de fronteras regulatorias e incluso fronteras nacionales siempre que la comunidad reguladora requiera un marcado de datos similar. Para la SEC, un beneficio clave de la coherencia entre reguladores en los datos etiquetados es la capacidad de comprender mejor la naturaleza de los riesgos en los mercados financieros. Los mercados de hoy no se detienen en las fronteras nacionales, por lo que mirar solo los datos intranacionales proporciona solo una imagen parcial del riesgo del sistema.

**3. Oportunidades de investigación**

La segunda pregunta clave de mi charla es sobre *las oportunidades de investigación en la era del big data.*

Veo muchas oportunidades de investigación para los economistas financieros de DERA, para los académicos, para la industria y para cualquiera que valore los datos financieros. En general, veo oportunidades basadas en grandes bases de datos que están disponibles ahora o que podrían estar disponibles en un futuro cercano. También veo aún más oportunidades basadas en los cambios que se están realizando en las fuentes de datos existentes. Además de una gran cantidad de preguntas académicas, el big data continuará ayudando a la SEC y a otros reguladores del mercado a identificar y cerrar a los malos actores.

Además de la base de datos OPRA que ya he mencionado, me gustaría destacar una base de datos adicional. Posteriormente, destacaré otras dos áreas que abrirán puertas a nuevas oportunidades de investigación.

**La pista de auditoría consolidada**

El 11 de julio de 2012, la Comisión votó para adoptar la Regla 613 bajo el Reglamento NMS. Este fue un marcador de milla significativo a lo largo del camino para crear e implementar la pista de auditoría consolidada (CAT). Cuando lo completen las organizaciones autorreguladoras (SRO), el CAT proporcionará una base de datos única y completa que permitirá a los reguladores realizar un seguimiento más eficiente y exhaustivo de toda la actividad comercial de acciones y opciones en los mercados estadounidenses. Esto transformará las funciones de vigilancia del mercado y ejecución de los reguladores. Por ejemplo, los reguladores podrán rastrear la actividad de un solo individuo que opera en múltiples mercados a través de múltiples corredores de bolsa. El CAT no estará disponible para académicos o la industria con fines de investigación, marketing u otros fines. [[7]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn7" \o ")

**Lenguajes estructurados estandarizados**

Las tres v de big data son *volumen* , *velocidad* y *variedad* . Es difícil imaginar que los conjuntos de datos financieros futuros tendrán menos *volumen* o menor *velocidad* que en la actualidad. Entonces, quizás la mejor manera de hacer que los conjuntos de datos futuros sean más manejables es imitar la solución del censo de Herman Hollerith atacando la *variedad* .

Desde mediados de la década de 1990, la mayoría de los documentos de la SEC se han presentado a EDGAR. Aunque las presentaciones son electrónicas y pueden ser leídas fácilmente por un ser humano en cualquier computadora, no son legibles por máquina porque son esencialmente papel electrónico no estructurado. El contenido no solo cambió entre los declarantes y a lo largo del tiempo, sino también el formato: texto sin formato, html, pdf y otros. Las iniciativas posteriores de la SEC han facilitado que las personas, las máquinas y los reguladores lean y comprendan las divulgaciones en EDGAR.

Se alcanzó un hito importante el 7 de mayo de 2003 cuando la SEC adoptó su requisito inicial de presentar los formularios 3, 4 y 5 utilizando el Lenguaje de marcado extensible (o XML). [[8]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn8" \o ") Creo que la estructuración de estos formularios en XML redujo los costos de acceso y los costos analíticos, haciendo que esta información sea más valiosa para el mercado. Desde 2003, ahora se envían muchos más formularios en XML, FIX, FpML, XBRL y, más recientemente, Inline XBRL.

Estructurar las divulgaciones para que sean legibles por máquina facilita el acceso y análisis más rápidos que pueden mejorar la toma de decisiones de los inversores y reducir la capacidad de los declarantes para ocultar el fraude. La información estructurada también puede ayudar a automatizar las presentaciones reglamentarias y el procesamiento de información comercial. En particular, al etiquetar los elementos de divulgación numéricos y narrativos de los estados financieros y los resúmenes de riesgo / rendimiento en XBRL, esos elementos de divulgación se estandarizan y pueden ser procesados ​​inmediatamente por software para análisis. Esta estandarización permite la agregación, la comparación y los análisis estadísticos a gran escala que son menos costosos y más oportunos para los usuarios de datos que si la información se presentara en un formato no estructurado. [[9]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn9" \o ") Los datos estructurados probablemente impulsarán la investigación futura en finanzas corporativas y macroeconomía.

**Identidad estandarizada: el LEI**

Otro problema común de big data es la conexión precisa y oportuna de conjuntos de big data dispares para análisis. Este problema se ve agravado por la amplia gama de identificadores utilizados por las agencias federales: el IRS tiene el Número de identificación del empleador (EIN); la Reserva Federal tiene el identificador de descuento de supervisión de estadísticas de investigación (ID de RSSD); FINRA tiene el Depositario de Registro Central (CRD); y la SEC tiene la clave de índice central (CIK). Un informe reciente identificó 36 agencias federales que utilizan hasta 50 sistemas distintos de identificación de entidades incompatibles. En mi opinión, estas diferencias elevan costos y cargas tanto para las agencias federales como para sus entidades reguladas.

El Identificador Global de Entidades Jurídicas (LEI) es un código alfanumérico de 20 caracteres que proporciona un único identificador internacional único que permite la identificación precisa de las entidades jurídicas. Como tal, ofrece un único conector internacional para conjuntos de macrodatos dispares y al mismo tiempo reduce la carga regulatoria actual asociada con el sistema de identificación único de cada agencia. El LEI incluye datos de "nivel 1" que sirven como tarjetas de presentación corporativas. (Responde a las preguntas "¿quién es quién?") El LEI también incluye datos de "nivel 2" que muestran las relaciones entre diferentes entidades. (Responde a la pregunta "¿quién es el propietario de quién?") El LEI sirve como una piedra de Rosetta para identificar de forma clara y única a las empresas y entidades que participan en los mercados financieros mundiales.

Recientemente, la Comisión publicó reglas que exigen el uso de LEI cuando se asocia con transacciones de swap basadas en valores. El LEI es ahora un componente de los informes obligatorios de transacciones de swaps en EE. UU., Europa y Canadá. Europa ha exigido el uso futuro de LEI ampliamente, incluso en las actividades de pago y liquidación, así como en la financiación estructurada. [[10]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn10" \o ")

Creo que aún no se han obtenido todos los beneficios del LEI. Como algunas empresas pueden tener cientos o miles de subsidiarias o afiliadas que operan en todo el mundo, se avecinan más beneficios a medida que el LEI se utilice de manera más amplia e integral. El LEI permite una mayor transparencia con respecto a las jerarquías y el mapeo de relaciones. Esto apoyará mejores análisis de los riesgos a medida que se agregan y potencialmente se vuelven sistémicos. [[11]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftn11" \o ")

**4. Conclusión**

Realmente espero con ansias la sólida discusión de hoy *sobre los desafíos* de las *políticas y las oportunidades de investigación en la era de los macrodatos* . Está ayudando a expandir el futuro de las finanzas de formas importantes que seguramente tendrán efectos positivos en los mercados, los inversores y las empresas. Gracias.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[[1]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref1" \o ") La Comisión de Bolsa y Valores se exime de responsabilidad por cualquier publicación o declaración privada de cualquier empleado o comisionado de la SEC. Este discurso expresa las opiniones del autor y no refleja necesariamente las de la Comisión, los Comisionados u otros miembros del personal.

[[2]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref2" \o ") Consulte el sitio web de la oficina del censo, incluido <https://www.census.gov/history/www/through_the_decades/overview/1890.html> y [https://www.census.gov/history/www/innovations/technology/ tabulation\_and\_processing.html](https://www.census.gov/history/www/innovations/technology/tabulation_and_processing.html) y <https://www.census.gov/history/www/census_then_now/notable_alumni/herman_hollerith.html>

[[3]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref3" \o ") International Data Corp (IDC).

[[4]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref4" \o ") Véase Scott W. Bauguess, The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective (21 de junio de 2017), disponible en: [https://www.sec.gov/news/speech / bauguess-big-data-ai](https://www.sec.gov/news/speech/bauguess-big-data-ai)

[[5]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref5" \o ") Véase el presidente Jay Clayton, discurso de apertura en la Conferencia Regional del Atlántico Medio (4 de junio de 2019), disponible en: [https://www.sec.gov/news/speech/clayton-keynote-mid-atlantic-regional -conferencia-2019](https://www.sec.gov/news/speech/clayton-keynote-mid-atlantic-regional-conference-2019)

[[6]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref6" \o ") Ver, por ejemplo, <https://www.bloomberg.com/news/articles/2019-07-02/jpmorgan-to-start-ai-hedge-fund-strategy-in-technology-arms-race> y <https://www.forbes.com/sites/bernardmarr/2019/02/15/the-revolutionary-way-of-using-artificial-intelligence-in-hedge-funds-the-case-of-aidyia/#56037dfe57ca>

[[7]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref7" \o ") Véase el presidente Jay Clayton, discurso de apertura en la Conferencia Regional del Atlántico Medio (4 de junio de 2019), disponible en: [https://www.sec.gov/news/speech/clayton-keynote-mid-atlantic-regional -conferencia-2019](https://www.sec.gov/news/speech/clayton-keynote-mid-atlantic-regional-conference-2019)

[[8]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref8" \o ")<https://www.sec.gov/rules/final/33-8230.htm>

[[9]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref9" \o ") Véase el Comisionado Michael S. Piwowar, Comentarios en la Cumbre de datos de RegTech 2018 — Old Fields, New Corn: Innovation in Technology and Law, disponible en: [https://www.sec.gov/news/speech/piwowar-old -campos-nuevo-maíz-innovación-tecnología-ley](https://www.sec.gov/news/speech/piwowar-old-fields-new-corn-innovation-technology-law)

[[10]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref10" \o ") Véase la comisionada Kara M. Stein, Quality Data and the Power of Prevention: Remarks at Meet the Market, Norteamérica, disponible en: [https://www.sec.gov/news/speech/quality-data-and- el-poder-de-prevención.html](https://www.sec.gov/news/speech/quality-data-and-the-power-of-prevention.html)

[[11]](https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-data" \l "_ftnref11" \o ")*Id.*