¿Qué Hace que Alguien sea un Científico de Datos?

¿Qué Hace que Alguien sea un Científico de Datos?

Ahora que sabe lo que hay en el libro, es hora de poner algunas definiciones. A pesar de su uso ubicuo, el consenso evade las nociones de big data y data science. La pregunta, “¿quién es un científico de datos?” Está muy viva y es cuestionada por individuos, algunos de los cuales simplemente están interesados en proteger su disciplina o territorio académico. En esta sección, intento abordar estas controversias y explicar por qué una definición estrechamente interpretada de big data o data science dará como resultado la exclusión de cientos de miles de personas que recientemente se han volcado al campo emergente.

"Todo el mundo ama a un científico de datos", escribió Simon Rogers (2012) en The Guardian. El sr. Rogers también rastreó el nuevo amor por el cálculo de números a una cita de Hal Varian de Google, quien declaró que "el trabajo sexy en los próximos diez años serán los estadísticos".

Mientras que Hal Varian nombró a los estadísticos “sexy”, se cree ampliamente que lo que realmente quiso decir eran científicos de datos. Esto plantea varias preguntas importantes:

• ¿Qué es la ciencia de datos?

• ¿Cómo difiere de las estadísticas?

• ¿Qué hace que alguien sea un científico de datos?

En tiempos de big data, una pregunta tan simple como, “¿qué es la ciencia de datos?” puede dar lugar a muchas respuestas. En algunos casos, la diversidad de opiniones sobre estas respuestas raya en la hostilidad.

Defino al científico de datos como alguien que encuentra soluciones a los problemas mediante el análisis de datos grandes o pequeños utilizando las herramientas apropiadas y luego cuenta historias para comunicar sus hallazgos a las partes interesadas relevantes. No uso el tamaño de los datos como una cláusula restrictiva. Un dato por debajo de cierto umbral arbitrario no lo convierte en un científico de datos menos. Mi definición de científico de datos tampoco se limita a herramientas analíticas particulares, como el aprendizaje automático. Mientras uno tenga una mente curiosa, fluidez en el análisis y la capacidad de comunicar los hallazgos, considero a la persona un científico de datos.

Defino ciencia de datos como algo que hacen los científicos de datos. Hace años, como estudiante de ingeniería en la Universidad de Toronto, me quedé atrapado con la pregunta: ¿Qué es la ingeniería? Escribí mi tesis de maestría sobre el pronóstico de los precios de la vivienda y mi tesis doctoral sobre el pronóstico de las elecciones de los constructores de viviendas relacionadas con lo que construyen, cuándo construyen y dónde construyen nuevas viviendas. En el departamento de ingeniería civil, otros trabajaban en el diseño de edificios, puentes, túneles y se preocupaban por la estabilidad de las pendientes. Mi trabajo, y el de mi supervisor, no era su ingeniería tradicional de variedades de jardín. Obviamente, otros me preguntaron repetidamente si mi investigación era realmente ingeniería.

Cuando compartí estas preocupaciones con mi supervisor doctoral, el profesor Eric Miller, se rió. El Dr. Miller pasó toda una vida investigando el uso del suelo urbano y el transporte, y anteriormente obtuvo un doctorado del MIT. "La ingeniería es lo que hacen los ingenieros", respondió. Durante los siguientes 17 años, me di cuenta de la sabiduría en su declaración. Primero se convierte en ingeniero al obtener un título y luego registrarse en el organismo profesional local que regula la profesión de ingeniería. Ahora eres ingeniero. Puedes cavar túneles; escribir códigos de software; componentes de diseño de un iPhone o un jet supersónico. Eres un ingeniero. Y cuando lidera la respuesta global a la crisis financiera en su papel de economista jefe del Fondo Monetario Internacional (FMI), como lo hizo el Dr. Raghuram Rajan, usted es ingeniero.

El profesor Raghuram Rajan realizó su primer título en ingeniería eléctrica del Instituto Indio de Tecnología. Se dedicó a la economía en estudios de posgrado, más tarde se convirtió en profesor en una prestigiosa universidad y finalmente aterrizó en el FMI. Actualmente se desempeña como el 23er Gobernador del Banco de la Reserva de la India. ¿Podría alguien argumentar que su destreza intelectual se basa solo en su formación como economista y que los fundamentos que aprendió como estudiante de ingeniería no desempeñaron ningún papel en el desarrollo de sus habilidades para resolver problemas?

El profesor Rajan es ingeniero. También lo son Xi Jinping, presidente de la República Popular de China, y primer ministro griego que está obligando al mundo a repensar los fundamentos de la economía global. Es posible que no estén diseñando nuevos circuitos, equipos de destilación o puentes, pero están ayudando a construir mejores sociedades y economías y no puede haber una mejor definición de ingeniería e ingenieros, es decir, individuos dedicados a construir mejores economías y sociedades.

Brevemente, diría que la ciencia de datos es lo que hacen los científicos de datos.

Otros tienen definiciones muy diferentes. En septiembre de 2015, un co-panelista en una reunión organizada por BigDataUniversity.com en Toronto limitó la ciencia de datos al aprendizaje automático. Ahí tiene. Si no está utilizando las cajas negras que componen el aprendizaje automático, según algunos expertos en el campo, no es un científico de datos. Incluso si descubriera la cura de una enfermedad que amenaza la vida de millones, los colegas que protegen su territorio lo excluirán del club de ciencia de datos.

El Dr. Vincent Granville (2014), autor de ciencia de datos, ofrece ciertos umbrales que cumplir para ser un científico de datos. En las páginas 8 y 9 en Developing Analytic Talent, el Dr. Granville describe al nuevo profesor de ciencias de datos como un instructor no titular en una universidad no tradicional, que publica resultados de investigación en blogs en línea, no pierde el tiempo escribiendo subvenciones, trabaja desde casa, y gana más dinero que los profesores tradicionales. Baste decir que la próspera comunidad académica de científicos de datos podría estar en desacuerdo con el Dr. Granville.

El Dr. Granville utiliza restricciones en el tamaño y los métodos de datos para definir qué es la ciencia de datos. Define a un científico de datos como alguien que puede "procesar fácilmente un conjunto de datos de 50 millones de filas en un par de horas: y que desconfía de los modelos (estadísticos). Distingue la ciencia de datos de las estadísticas. Sin embargo, enumera álgebra, cálculo y capacitación. en probabilidad y estadística como antecedentes necesarios para comprender la ciencia de datos" (página 4).

Algunos creen que big data se trata simplemente de cruzar un cierto umbral en el tamaño de los datos o el número de observaciones, o se trata del uso de una herramienta en particular, como Hadoop. Dichos umbrales arbitrarios en el tamaño de los datos son problemáticos porque con la innovación, incluso las computadoras normales y el software comercial han comenzado a manipular conjuntos de datos muy grandes. Stata, un software comúnmente utilizado por científicos de datos y estadísticos, anunció que ahora se pueden procesar entre 2 y 24,4 mil millones de filas utilizando sus soluciones de escritorio. Si Hadoop es la contraseña para el club de big data, la capacidad de Stata para procesar 24.4 mil millones de filas, bajo ciertas limitaciones, acaba de colarse en esa fiesta de big data.

Es importante darse cuenta de que alguien que intenta establecer umbrales arbitrarios para excluir a otros es probable que encuentre inconsistencias. El objetivo debería ser definir la ciencia de los datos en un contexto más exclusivo, independiente de la disciplina y la plataforma, sin tamaño, donde la resolución de problemas centrada en los datos y la capacidad de tejer narrativas fuertes sean el centro de atención.

Dada la controversia, preferiría consultar a otros para ver cómo describen a un científico de datos. ¿Por qué no volvemos a consultar al Científico en Jefe de Datos de los Estados Unidos? Recordemos que el Dr. Patil le dijo al periódico Guardian en 2012 que un "científico de datos es esa combinación única de habilidades que pueden desbloquear los conocimientos de los datos y contar una historia fantástica a través de los datos". Lo que es admirable de la definición del Dr. Patil es que incluye a personas de diversos antecedentes académicos y capacitación, y no restringe la definición de un científico de datos a una herramienta en particular ni la somete a un cierto umbral mínimo arbitrario de tamaño de datos.

El otro ingrediente clave para un científico de datos exitoso es un rasgo de comportamiento: la curiosidad. Un científico de datos tiene que ser uno con una mente muy curiosa, dispuesta a dedicar mucho tiempo y esfuerzo para explorar sus corazonadas. En periodismo, los editores lo llaman tener nariz para las noticias. No todos los periodistas saben dónde están las noticias. Solo aquellos que tienen la nariz para las noticias entienden la historia. La curiosidad es igualmente importante para los científicos de datos como lo es para los periodistas.

Rachel Schutt es la Científica de Datos en Jefe de News Corp., imparte un curso de ciencias de datos en la Universidad de Columbia. También es autora de un excelente libro, Doing Data Science. En una entrevista con el New York Times, la Dra. Schutt definió a un científico de datos como alguien que es en parte científico de la computación, en parte ingeniero de software y en parte estadístico (Miller, 2013). Pero esa es la definición de un científico de datos promedio. "Los mejores", sostuvo, "tienden a ser personas realmente curiosas, pensadores que hacen buenas preguntas y están bien lidiando con situaciones no estructuradas y tratando de encontrar estructura en ellas".