

El gripazo de Google muestra las flaquezas del ‘big data’

- [Javier Salas](#)
- 27/03/2014

Una herramienta de Google pasó de predecir con exactitud los casos de gripe en cada región de EEUU gracias a las consultas del buscador a fallar en todos sus pronósticos. Un artículo en ‘Science’ señala este caso como una parábola de la arrogancia asociada al campo del ‘big data’



Una trabajadora en el cerebro de Google, su Data Center de Oregón. / [Google](#)

El *big data* prometía soluciones, respuestas. La formidable capacidad de procesar ingentes cantidades de datos permite descubrir información desconocida, perspectivas sorprendentes, conclusiones sólidas. En 2009 se produjo un gran acontecimiento en este sentido: la revista *Nature* publicó un estudio que mostraba cómo las consultas en el motor de búsqueda de Google se habían traducido en [una predicción casi exacta de la incidencia de la gripe en cada región de EEUU](#). El *big data* relucía en las portadas, una herramienta que ayudaría a conservar la salud de la gente mejor que los epidemiólogos.

Sin embargo, sus últimas previsiones han duplicado el dato real ofrecido por las autoridades sanitarias. En un lustro, Google pasó de clavar el pronóstico a fallar como escopeta de feria.

¿Qué ha ocurrido en estos años que ha llevado al fracaso de [Google Flu Trends](#) (tendencias de la gripe)? Hay quien directamente lo atribuye a la arrogancia del *big data* y la ausencia de espíritu científico de Google, pero no conviene perder de vista todos los matices si queremos sacar conclusiones útiles. ¿Cuáles son las flaquezas del *big data* que muestra este caso y hasta qué punto son representativas de lo que ocurre con otros *oráculos* basados en grandes cantidades de información social, [como ocurre con la capacidad predictiva de Twitter](#)?

Los expertos señalan varios puntos clave, como la opacidad de las empresas que cuentan con mayores bases de datos, los cambios del algoritmo, cierta desconexión al confiar ciegamente en la cantidad de datos olvidando el método científico, las dificultades para predecir comportamientos sociales, la facilidad para manipular desde el exterior la cosecha de información y que hoy somos más exigentes con unos resultados que hace años parecían formidables.

Hacer predicciones sobre el comportamiento humano no es fácil: “Un electrón siempre va a tener la misma carga, pero las personas cambian”

Repasemos [lo ocurrido con Flu Trends](#) para visualizar su talón de Aquiles. La gigantesca multinacional que (casi) todo lo sabe, podía cabalgar sobre millones de búsquedas relacionadas con la gripe —“síntomas gripe”, “virus gripe”, etcétera— para decirnos casi al instante si habría colas en las urgencias en determinado punto del país. Mientras, los sistemas predictivos de los Centros para el Control y la Prevención de Enfermedades de EEUU (CDC), necesitaban entre una o dos semanas para recoger la información necesaria.

El artículo en [Nature](#) que consagró a Flu Trends lo firmaron conjuntamente Google y los CDC ([PDF](#)). *The New York Times* [hablaba](#) de “un matrimonio fructífero” entre la medicina y el comportamiento de las masas. “Desde un punto de vista tecnológico, es solo el comienzo”, dijo el presidente de Google, Eric Schmidt. Desde entonces, el sistema ha sobreestimado constantemente las visitas a centros médicos relacionadas con la gripe, siendo especialmente inexacto en los picos de la temporada de gripe, cuando más útiles son las previsiones. En la temporada 2012-2013, predijo el doble de visitas al médico que las registradas por los CDC y en la 2011-2012 se sobreestimó en más de un 50%: [no atina desde agosto de 2011](#).

La arrogancia del ‘big data’

Las primeras veces que falló, Google lo quiso atribuir a una maniobra malintencionada: no es difícil alterar el resultado, tan solo hay que fingir que se busca información sobre esta enfermedad. Además, se decía desde la empresa, el interés que despertó Flu Trends en los medios provocó que la herramienta recibiera una importante atención por parte del público, lo que terminó [por falsear el resultado](#).

Criticaron “la suposición implícita de que pueden sustituir a la recopilación de datos y el análisis tradicional”

Varios expertos en *big data* [publicaron recientemente en Science un análisis de los fallos de Google Flu Trends](#), que consideraban una parábola de lo que está ocurriendo en su campo. "La arrogancia del *big data*", arrancan, "es la suposición implícita de que pueden sustituir, en lugar de complementar, a la recopilación de datos y el análisis tradicional". "La cantidad de datos no significa que uno puede ignorar los problemas fundamentales de la medición", añaden estos autores, liderados por [David Lazer](#), de la Universidad de Harvard (EEUU).

Google tiene [más datos que nadie](#) y, en muchas ocasiones, el debate sobre el valor del *big data* se centra únicamente en la cantidad. "Es verdad que se da un poco de arrogancia del tipo 'mi base de datos es más grande que la tuya'. Muchas veces se queda en segundo plano la importancia del proceso de transformación del dato en valor", reconoce [Esteban Moro](#), investigador de la Universidad Carlos III de Madrid.

Terabytes de datos

"Esto es algo que sabe cualquier persona que se dedique a aplicar el método científico, el dato en sí no tiene ningún valor", insiste Moro. "Antes, los datos cabían en una hoja. Pero ahora, por tener terabytes de datos, no tienes mejor información". ¿Cuál ha sido el problema de Google con la transformación de los datos? Lazer apuesta en *Science* por los distintos (cientos, miles) de cambios que el propio Google introduce en el algoritmo de su buscador.

"Por tener terabytes de datos no tienes mejor información", señala Esteban Moro

"La forma en que se presentan los resultados también cambia la forma de buscarlos", resume Moro, en referencia a novedades como el autocompletado del buscador, que dificulta conocer con exactitud lo que iba a buscar el usuario. Quizá el internauta iba a teclear "gripe en ancianos", preocupado por su abuelo, pero al terminar de escribir la primera palabra el buscador le sugirió otras opciones que alteraron su intención.

"Lo lógico es que lo hubieran refinado", reflexiona [Daniel Gayo](#), investigador de la Universidad de Oviedo. "En este campo, los cambios más triviales pueden afectar tremendamente: la forma de ordenar los resultados, la interfaz, el algoritmo... Es posible que no haya colaboración entre equipos, que dentro de la propia empresa no haya comunicación entre los departamentos responsables de los algoritmos del buscador y de la confección de Flu Trends", sugiere Gayo.

El ser humano es imprevisible

Además, Gayo recuerda que "el *big data* es un subproducto, hay que vigilar cómo se origina", en alusión a esas búsquedas: los internautas no siempre se van a comportar como *deberían*, cumpliendo como simples sensores, "sino por antojos, modas y disfrutando de su albedrío". Moro coincide en que los comportamientos sociales son mucho más difíciles de calibrar: en su día a día, este investigador se sirve de los datos para la predicción del fraude o las bajas en los servicios telefónicos y "se trata de un comportamiento que cambia permanentemente, de un día para otro". Lo que servía en 2009 para vaticinar estornudos, en 2013 puede estar desfasadísimo.

Los internautas no siempre se van a comportar como 'deberían', como simples sensores

Un gran obstáculo que identifica Moro es el de saltar de la correlación a la predicción: “Aunque encontremos correlaciones significativas, pueden no ser suficientemente buenas para aventurarse a realizar una predicción”. Y en el momento en que las encontremos, apunta Gayo, los interesados harán lo posible para manipularlo: “Es lo que ocurre en política y las predicciones electorales en Twitter, [todos van a querer cambiar el resultado a su favor](#)”.

Otro de los problemas científicos que plantea la experiencia de Flu Trends es el de la replicabilidad de los resultados y la falta de transparencia. ¿Quién va a comprobar sus números y predicciones si ni siquiera sabemos cuáles son las búsquedas que se tienen en cuenta para hacerlas? Google es una empresa y sus datos atañen a la privacidad de millones de personas. En *Science* critican que los materiales que aportan desde Flu Trends no cumplen con los estándares científicos emergentes y que, no obstante, hay otros que sí se podrían liberar para ayudar a la ciencia.

Falta de transparencia

“Te tienes que creer que funciona, dar por buenos los datos que proporcionan las empresas. Sin una mínima transparencia, no se puede hacer ciencia a partir de su trabajo, que sería lo deseable”, señala el investigador de la Universidad de Oviedo. Gayo considera que, en este caso, los autores de *Science* se han podido pasar de frenada elevando a categoría este fallo: más allá de EEUU, Flu Trends sí funciona para numerosos países y los picos coinciden con la realidad, aunque no en tamaño. “Hace unos años sorprendió gratamente el descubrimiento, pero ahora somos más exigentes con los resultados”, resume.

Es decir, tampoco conviene ser reduccionistas y que la mala experiencia de Flu Trends nos lleve a decir que “el *big data* no merece la pena”, [como señala el experto Kaiser Fung en la revista *Harvard Business Review*](#). Lazer y los coautores de su texto creen que falta coordinación entre el *big data* y la confección de datos de toda la vida. Gayo opina que, tal y como se plantea en *Science*, el caso no es generalizable como un inconveniente común para la ciencia de los datos. Y Moro, que admite que “muchísima gente se dejó llevar por el entusiasmo”, sugiere que finalmente el factor humano es más determinante de lo que pensábamos: “Un electrón siempre va a tener la misma carga, pero las personas cambian”.

REFERENCIA

['The Parable of Google Flu: Traps in Big Data Analysis' DOI: 10.1126/science.1248506](#)

MÁS INFO

- » [Detecting influenza epidemics using search engine query data](#) (Nature)
- » [Google Flu Trends gets it wrong three years running](#) (New Scientist)