

«La ciencia es **un método**

para **encontrar la verdad,**

**no la verdad en sí»**

**SIOBHAN ROBERTS**

TRADUCIDO POR ANA MILUTINOVIC

03 JUNIO, 2021

**Cuando estalló la pandemia, el científico de datos especializado en finanzas Youyang Gu vio que los modelos de pronóstico del coronavirus eran muy inexactos y decidió usar sus conocimientos para crear uno mejor y más simple, que acabó convertido en referencia mundial.**

**E**l científico de datos Youyang Gu se considera realista y así lo afirma en su perfil de Twitter: «Presento opiniones sin sesgo. Realista».

Por eso, cuando la primavera pasada vio proyecciones muy diversas sobre el coronavirus (COVID-19), como un modelo que estimaba dos millones de muertes en Estados Unidos para el verano y otro que predijo 60.000, empezó a dudar de la calidad de los modelos informáticos. Así que decidió intentar crear un modelo propio.

El investigador recuerda: «Mi objetivo era generar el modelo más preciso posible. Sin los condicionantes (si esto) o (si aquello). Básicamente, no hay 'si'. Realmente no importa cuáles sean los escenarios. Solo quería exponerlo: (Este es el pronóstico más probable o realista de lo que va a suceder)». En una semana, logró construir un modelo de aprendizaje automático y lanzó el sitio web COVID-19 Projections. Gu ejecutaba el modelo todos los días (solo tardaba una hora en

su ordenador portátil) y publicaba las proyecciones sobre las muertes por COVID-19 para 50 estados de EE. UU., 34 condados y 71 países.

A finales de abril, atraía bastante atención: la web llegó a recibir millones de visitas a diario. El profesor de biología de la Universidad de Washington (EE. UU.) Carl Bergstrom se dio cuenta y comentó en Twitter que el modelo de Gu «hacía predicciones que parecían mejores que otras que veía». Y añadió: «Suelo ser algo escéptico en cuanto al aprendizaje automático. Pero en este caso, no piensen que por llamarse (aprendizaje automático) se trata de un engaño».

Cuando empezó la pandemia, Gu, de 27 años y graduado del MIT (EE. UU.) con un máster en Ingeniería Eléctrica e Informática (además de un título académico en Matemáticas) trabajaba en una *start-up* de análisis deportivo, pero pausó sus tareas por el parón de los deportes de las grandes ligas. Y luego, simplemente buscando en Google la palabra «epidemiología», comenzó su incursión en el modelado sobre la COVID-19.

El investigador recuerda: «No tenía experiencia en el modelado de enfermedades infecciosas». Pero sí tenía años de experiencia como científico de datos en finanzas, trabajando con modelos estadísticos que analizan los datos basados en ciertos supuestos estadísticos para hacer proyecciones sobre, digamos, cuál será el precio de una acción en el futuro.

«Resulta que muchos modelos de enfermedades infecciosas son básicamente modelos estadísticos», asegura. Y el objetivo de la precisión impulsado por los beneficios de la industria financiera le fue muy útil en el campo epidemiológico. «Si uno no puede hacer un modelo preciso en finanzas, se queda sin trabajo», resalta. En cambio, el objetivo en el mundo académico, al menos desde el punto de vista de Gu, no se centra tanto en crear modelos precisos, sino más bien en publicar artículos e informar sobre las políticas públicas. «Eso no quiere decir que no hagan modelos precisos, sino solo que no optimizan la precisión de forma específica», sostiene.

El modelo de Gu combina aprendizaje automático con un simulador clásico de enfermedades infecciosas llamado modelo SEIR (que incluye a las personas de la población que son susceptibles, expuestas, contagiosas, recuperadas o eliminadas por fallecimiento).

**el modelo de Gu combina aprendizaje automático con un simulador clásico de enfermedades infecciosas llamado modelo SEIR (que incluye a las personas de la población que son susceptibles, expuestas, contagiosas, recuperadas o eliminadas por fallecimiento).**



Youyang Gu, científico de datos

El componente SEIR utiliza como input un conjunto simulado de parámetros: el rango de mejor estimación para variables como el número básico de reproducción (la tasa a la que surgen nuevos casos en una población totalmente susceptible al inicio de un brote, antes de las intervenciones o la inmunidad), la tasa de contagio, la fecha del inicio del confinamiento, la fecha de la reapertura y el número de reproducción efectiva (la tasa a la que surgen nuevos casos después de algunas medidas de intervención). En cuanto los resultados u outputs, el simulador SEIR primero calcula las infecciones a lo largo del tiempo y luego las muertes (multiplicando los contagios por la tasa de letalidad por la infección).

La capa de aprendizaje automático de Gu genera miles de combinaciones diferentes para esos conjuntos de datos para tratar de encontrar los parámetros reales para cada región geográfica. Aprende qué parámetros generan las proyecciones más precisas de fallecimientos al comparar las predicciones de SEIR con

los datos reales sobre las muertes diarias de la Universidad Johns Hopkins (EE. UU.). «Descubre qué conjuntos de parámetros generan el número de muertos que se asemeja más a los datos reales observados, mirando hacia atrás. Y luego usa esos parámetros para pronosticar y hacer proyecciones sobre las futuras muertes», explica su creador.

Sus pronósticos resultaron extraordinariamente precisos. Por ejemplo, el 3 de mayo, Gu apareció en CNN Tonight y compartió las proyecciones de su modelo que estimaban que Estados Unidos alcanzaría 70.000 muertes el 5 de mayo, 80.000 el 11 de mayo, 90.000 el 18 de mayo y 100.000 el 27 de mayo. El 28 de mayo, Gu tuiteó, «COVID19-projections.com acertó exactamente las 4 fechas». Con algo de redondeo, sí que resultó cierto.

El modelo no fue perfecto, por supuesto, pero impresionó al bioestadístico e investigador de enfermedades infecciosas de la Universidad de Massachusetts, en Amherst (EE. UU.) Nicholas Reich, cuyo laboratorio, en colaboración con los Centros para el Control y la Prevención de Enfermedades de EE. UU., añade resultados de aproximadamente 100 equipos internacionales de modelado. Entre todos los modelos agregados, Reich observó que el modelo de Gu estaba «siempre entre los mejores».

El 6 de octubre, Gu publicó su último pronóstico sobre fallecimientos, justo antes de la ola de otoño. El modelo proyectaba que habría 231.000 muertes en Estados Unidos para el 1 de noviembre. El total registrado a esa fecha fue: 230.995.

Dejó de usar su primer modelo a principios de octubre porque entonces ya había muchos equipos haciendo buenos pronósticos del número de fallecidos. En cambio, se dedicó a modelar el contagio real frente al registrado. Y luego en diciembre comenzó el seguimiento de la vacunación y el escurridizo «camino hacia la inmunidad de grupo», lo que a principios de 2021 se empezó a denominar «camino a la normalidad». Mientras que la inmunidad colectiva o de grupo se consigue cuando una parte suficiente de una población es inmune al virus, lo que reduce la propagación, Gu define la normalidad como «el levantamiento de todas las restricciones relacionadas con la COVID-19 para la mayoría de los estados de EE. UU.».

Y explica: «Quedó claro que no alcanzaríamos la inmunidad de grupo en 2021, al menos no en todo el país. Y creo que resulta importante, especialmente si se quiere infundir confianza, que hagamos caminos sensatos para cuando podamos volver a la normalidad. No deberíamos fijarlo a un objetivo poco realista como alcanzar la inmunidad colectiva. Sigo siendo cautelosamente optimista de que mi primer pronóstico de febrero, sobre la vuelta a la normalidad en el verano, será válido».

A principios de marzo, dejó de realizar proyecciones por completo; pensó que ya había hecho todo lo que podía para contribuir. Recuerda: «Quería dar un paso atrás y dejar que los demás modeladores y expertos hicieran su trabajo. No quiero liar ese campo».

Todavía sigue atento a los datos, investigando y analizando las variantes, la vacunación y la cuarta ola. Gu cuenta: «Si veo algo especialmente preocupante o problemático de lo que creo que la gente no está hablando, definitivamente lo publicaré». Pero por el momento se dedica a otros proyectos, como la plataforma de análisis de cotizaciones bursátiles «YOLO Stocks». Su principal trabajo sobre la pandemia es como miembro del grupo asesor técnico de la Organización Mundial de la Salud sobre la evaluación de la mortalidad por COVID-19, donde comparte sus conocimientos como consultor externo. Y confiesa: «Definitivamente he aprendido mucho. El año pasado fue muy revelador».

## LECCIÓN 1: CENTRARSE EN LOS ASPECTOS FUNDAMENTALES

Gu explica: «Desde la perspectiva de la ciencia de datos, mis modelos han demostrado la importancia de la simplicidad, que a menudo se subestima». Su modelo de pronóstico de fallecimientos era simple no solo en su diseño (el componente SEIR con una capa de aprendizaje automático), sino también en su enfoque muy reducido y «de abajo hacia arriba» con respecto a los datos introducidos. Afirma que de abajo hacia arriba significa «comenzar desde lo básico y añadir complejidad según sea necesario», y añade: «Mi modelo solo usa los fallecimientos pasados para predecir las futuras muertes. No utiliza ninguna otra fuente real de datos».

## «Desde la perspectiva de la ciencia de datos, mis modelos han demostrado la importancia de la simplicidad, que a menudo se subestima».

Notó que otros modelos se basaban en una variedad ecléctica de datos sobre los casos, las hospitalizaciones, los test, la movilidad, el uso de mascarillas, las patologías, la distribución por edad, la demografía, la estacionalidad de la neumonía, la tasa anual de mortalidad por neumonía, la densidad de la población, la contaminación del aire, la altitud, los datos sobre el tabaquismo, los contactos auto-registrados, el tráfico de pasajeros de las aerolíneas, los puntos de atención, los termómetros inteligentes, las publicaciones en Facebook, las búsquedas en Google, y más.

Y cuenta: «Existe la creencia de que, si se introducen más datos al modelo, o si es más sofisticado, funcionará mejor. Pero en situaciones

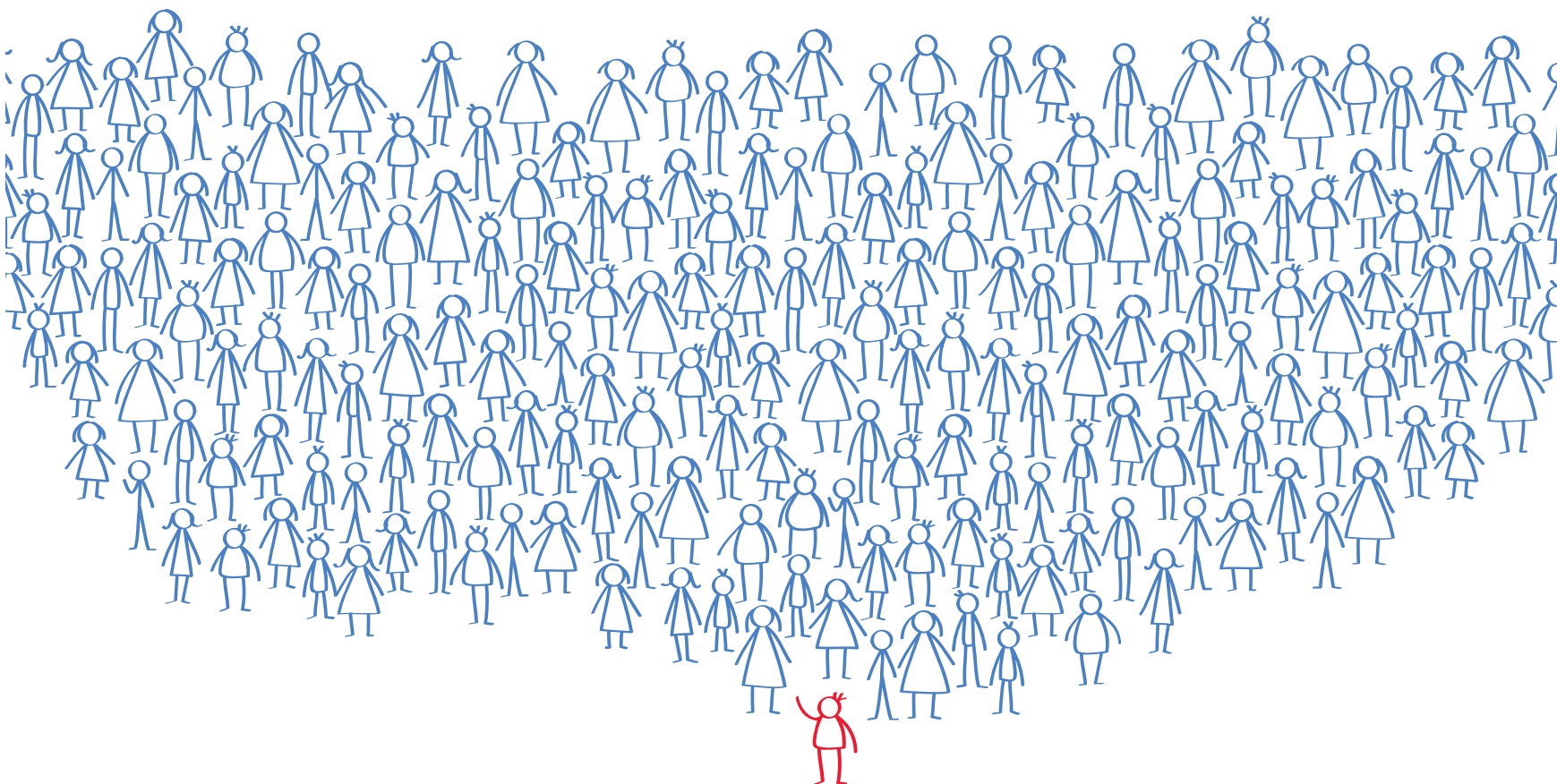
reales como en una pandemia, donde los datos son tan ruidosos, hay que mantener las cosas lo más simples posible. Desde el principio decidí que las muertes ocurridas son el mejor predictor de los fallecimientos futuros. Es muy simple: *input, output*. Introducir más fuentes de datos solo complica extraer la señal del ruido».

### LECCIÓN 2: MINIMIZAR LAS SUPOSICIONES

Gu considera que tuvo una ventaja al abordar el problema con un papel en blanco: «Mi objetivo era simplemente seguir los datos de la COVID-19 para aprender sobre la COVID-19. Es uno de los principales beneficios de la perspectiva de alguien ajeno». Pero al no ser epidemiólogo,

también tenía que asegurarse de no realizar suposiciones incorrectas o inexactas.

Gu explica: «Mi función consiste diseñar un modelo capaz de descubrir las suposiciones por mí. Cuando aparecen nuevos datos que van en contra de nuestras ideas, a veces tendemos a pasarlos por alto, y eso puede tener repercusiones en el futuro. Desde luego que me vi siendo víctima de eso, y sé que muchas otras personas también lo han sido. Por lo tanto, es muy importante ser consciente del posible sesgo que tenemos y reconocerlo, y poder ajustar nuestras ideas previas, si los nuevos datos las contradicen, especialmente en un entorno de rápido movimiento como el que hemos visto con la COVID-19.»



**LECCIÓN 3: PONER A PRUEBA LAS HIPÓTESIS**

Gu sostiene: «Lo que he notado en los últimos meses es que cualquiera puede hacer afirmaciones o manipular los datos para que se ajusten a la narrativa de lo que quieren creer». Esto resalta la importancia de hacer hipótesis comprobables.

Y añade: «Para mí, esa es la base de mis proyecciones y pronósticos. Tengo un conjunto de suposiciones, y si son ciertas, entonces esto es lo que predecimos que sucederá en el futuro. Y si las suposiciones acaban siendo incorrectas, entonces, por supuesto, tenemos que admitir que las suposiciones que hicimos no son ciertas y ajustarlas en consecuencia. Si no se formulan

hipótesis comprobables, no hay forma de demostrar si en estamos en lo cierto o no».

**LECCIÓN 4: APRENDER DE LOS ERRORES**

Gu también admite: «No todas las proyecciones que hice fueron correctas». En mayo de 2020, proyectó 180.000 muertes en EE. UU. para agosto. «Muchas más de las que se produjeron», recuerda. Su hipótesis comprobable resultó incorrecta, y afirma: «Eso me obligó a ajustar mis suposiciones».

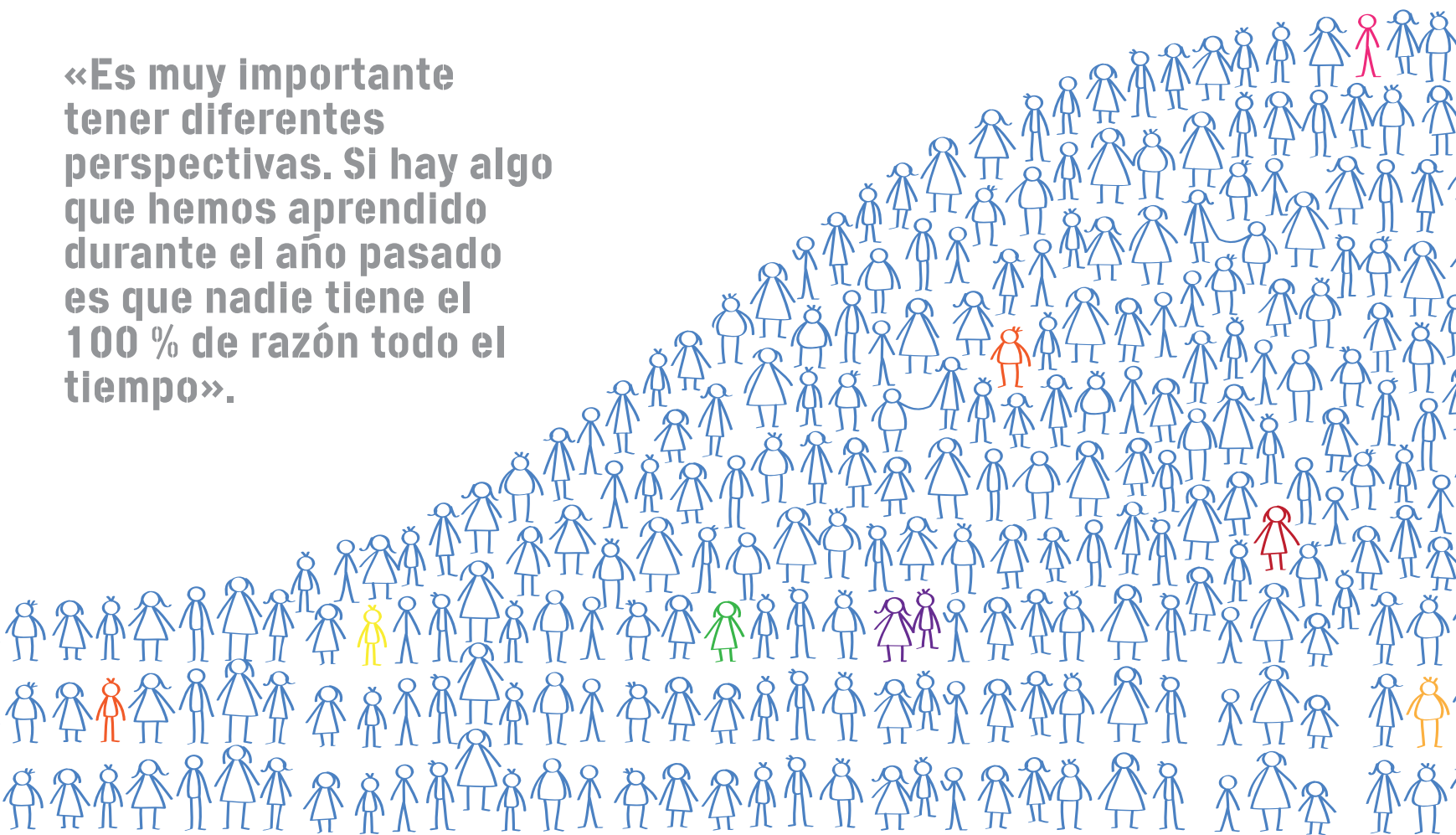
En ese momento, Gu usaba una tasa fija de mortalidad por infección de aproximadamente

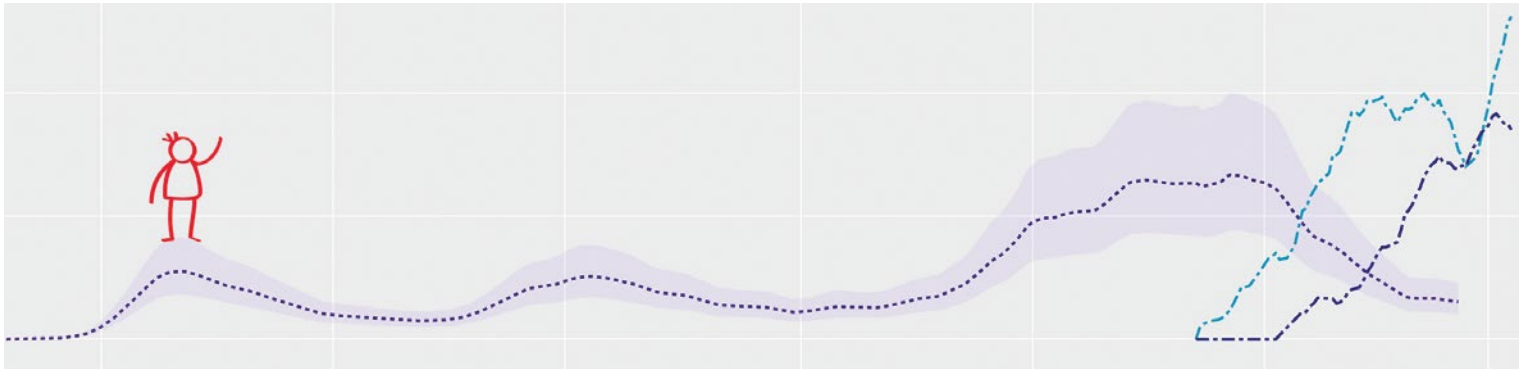
el 1 % como constante en el simulador SEIR. Cuando en el verano la tasa de mortalidad por infección se redujo a aproximadamente el 0,4 % (y luego a cerca de 0,7 %), sus proyecciones volvieron a un nivel más realista.

**LECCIÓN 5: INVOLUCRAR A LOS CRÍTICOS**

El investigador confiesa: «No todo el mundo estará de acuerdo con mis ideas, y lo agradezco». Tras haber usado Twitter para publicar sus proyecciones y análisis, cuenta: «Intento responder a la gente tanto como puedo, defender mi posición y debatir. Eso me obliga a pensar cuáles

«Es muy importante tener diferentes perspectivas. Si hay algo que hemos aprendido durante el año pasado es que nadie tiene el 100 % de razón todo el tiempo».





**es importante cuestionar los resultados, pero de una forma saludable. Se trata de una línea muy fina. Porque mucha gente simplemente rechaza rotundamente la ciencia, y esa tampoco es la forma más correcta de hacerlo.**

son mis suposiciones y por qué creo que son correctas».

Y añade: «Es algo que se remonta al sesgo de confirmación. Si no puedo defender adecuadamente mi posición, entonces ¿acaso es realmente correcto, y debería hacer estas afirmaciones? Relacionarme con otras personas me ayuda a comprender cómo pensar sobre estos problemas. Cuando otros presentan evidencias que contradicen mis posiciones, tengo que ser capaz de reconocer cuándo puedo estar equivocado en algunas de mis suposiciones. Y eso me ha ayudado enormemente a mejorar mi modelo».

**LECCIÓN 6: EJERCER UN ESCÉPTICISMO SALUDABLE**

Por último, Gu afirma: «Ahora soy mucho más escéptico con la ciencia, y eso no es nada malo. Creo que es importante cuestionar siempre los resultados, pero de una forma saludable. Se trata de una línea muy fina. Porque mucha gente simplemente rechaza rotundamente la

ciencia, y esa tampoco es la forma más correcta de hacerlo. Pero creo que también es importante no confiar ciegamente en la ciencia. Los científicos no son perfectos».

En su opinión, si algo no parece correcto, lo apropiado es hacer preguntas y encontrar las explicaciones. Y concluye: «Es muy importante tener diferentes perspectivas. Si hay algo que hemos aprendido durante el año pasado es que nadie tiene el 100 % de razón todo el tiempo. No puedo hablar por todos los científicos, pero mi trabajo es superar todo el ruido y llegar a la verdad. No estoy diciendo que haya sido perfecto. Durante el último año me he equivocado muchas veces. Pero creo que todos podemos aprender a acercarnos a la ciencia como un método para encontrar la verdad, en vez de creer que la ciencia en sí es la propia verdad». </>

La autora es editora senior de *MIT Technology Review*. Es periodista científica, biógrafa e historiadora de las matemáticas.

El artículo original «La ciencia es un método para encontrar la verdad, no la verdad en sí.» pertenece a la edición digital de *MIT Technology Review*.

Los contenidos bajo el sello *MIT Technology Review* están protegidos enteramente por copyright. Ningún material puede ser reimpresso parcial o totalmente sin autorización.

Si quisiera syndicar el contenido de la revista *MIT Technology Review*, por favor contáctenos.

E-mail: [redaccion@technologyreview.com](mailto:redaccion@technologyreview.com)

Tel: +34 911 284 864