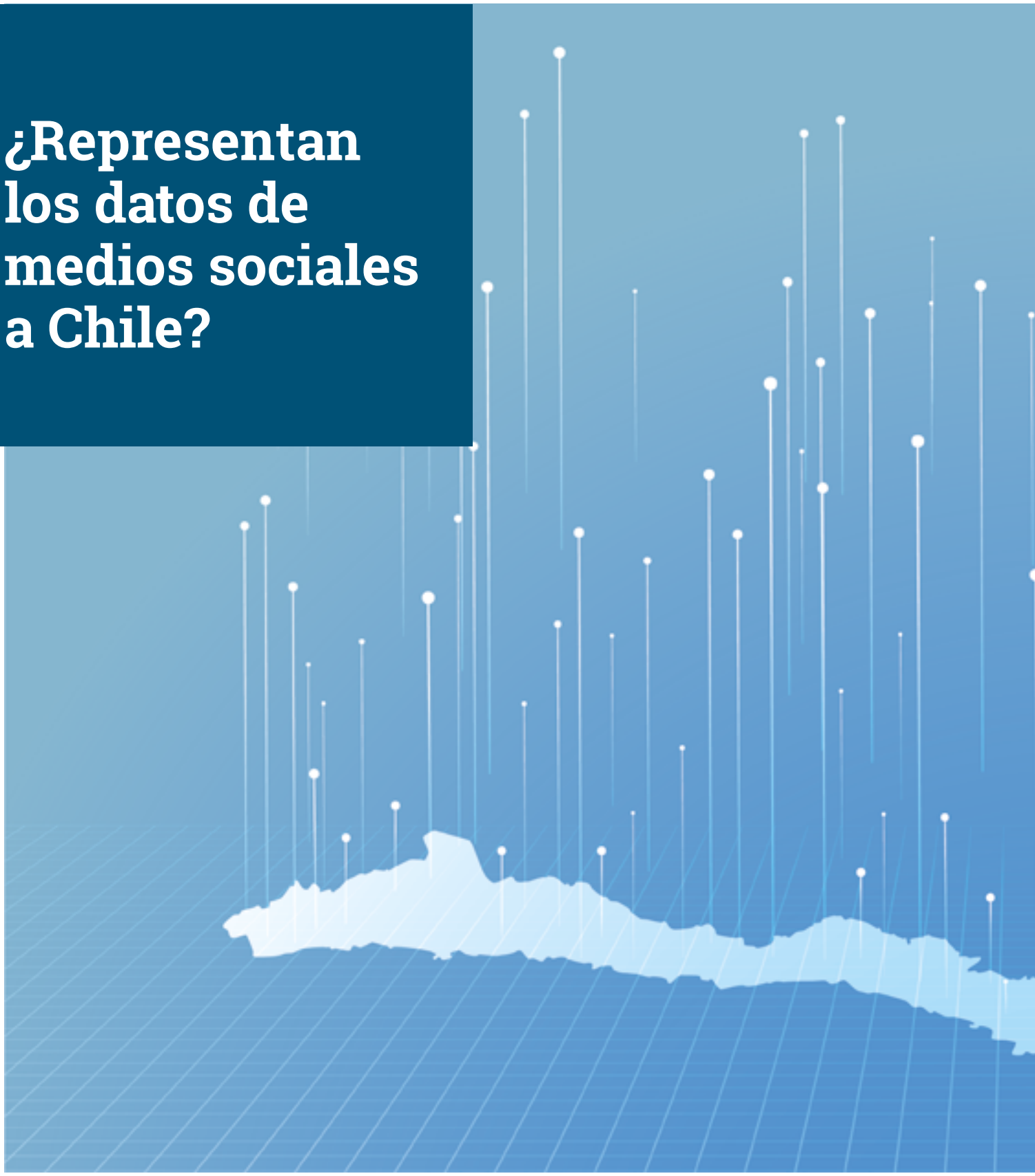
¿Representan los datos de medios sociales a Chile?



RICARDO BAEZA-YATES

Director tecnológico de NTENT, empresa de tecnología de búsqueda semántica basada en California, director de los programas de postgrado en ciencia de datos de la Northeastern University, sede Silicon Valley e investigador senior del Instituto Milenio Fundamentos de los Datos. Profesor del Departamento de Ciencias de la Computación de la Universidad de Chile. PhD en ciencia de la computación por la Universidad de Waterloo, Canadá. Sus áreas de investigación son búsqueda en la Web, minería de datos y algoritmos. Es ACM Fellow e IEEE Fellow.

http://www.baeza.cl/

Desde el estallido social del 18 de octubre de 2019, los medios sociales en Internet (*social media* en inglés) y particularmente Twitter, han sido usados para tratar de entender las motivaciones y opiniones de los chilenos, culminando con el escándalo del famoso informe de big data que recibió el Gobierno. Motivado por estos hechos, en este artículo analizamos los sesgos de los datos de los medios sociales y si es posible mitigarlos para que realmente representen la opinión del país. También de paso explicamos qué son los datos masivos (big data en inglés) y mostramos cómo incluso datos correctos pueden ser manipulados para difundir información falsa. Una presentación preliminar basada en estas ideas se puede encontrar en [2].

Datos masivos y medios sociales

Wikipedia define datos masivos o macrodatos de la siguiente manera: *"Macrodatos es un término que hace referencia a conjuntos de datos tan grandes y complejos como para que hagan falta aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente"*. Esta definición es bastante abstracta pues no nos dice a partir de qué tamaño son masivos ni cuán complejos deben ser. Por esto se usan las siguientes características para precisar esta definición:

- ***Volumen***: la cantidad de datos generados y guardados.
- ***Variedad***: el tipo y naturaleza heterogénea de los datos.

Cualidad	Problemas de datos	Problemas de computación	Problemas humanos
Volumen	escala, **redundancia**	escalabilidad	sobrecarga de información
Variedad	heterogeneidad, complejidad	adaptabilidad, extensibilidad	complejidad
Veracidad	exhaustividad, **sesgo, escasez, ruido, spam**	fiabilidad, confianza	**sesgo, escasez, ruido, spam**
Velocidad	tiempo real (instantáneo)	en línea (menos de unos segundos)	sobrecarga de información
Valor	utilidad, **privacidad**	depende del objetivo	**privacidad, ética y legalidad**

Figura 1. Características distintivas del big data y problemas asociados a cada una de ellas.

Medio social	¿Datos públicos?	Usuarios chilenos	Penetración aproximada
Facebook	pocos	13,0M*	71%
WhatsApp	ninguno	12,4M	68%
YouTube	muchos	11,7M	64%
FB Messenger	ninguno	7,3M	40%
Instagram	bastantes	7,0M*	38%
LinkedIn	muchos	4,8M	26%
Twitter	**todos**	1,5M*	8%
Snapchat	pocos	1,1M*	6%

Figura 2. Penetración de los medios sociales más populares en Chile.

- ***Velocidad***: tasa a la cual se generan y procesan los datos.
- ***Veracidad***: calidad de los datos obtenidos.
- ***Valor***: los datos obtenidos deben ser útiles y accionables.

[Además de los sesgos de género, edad y demografía] los datos de Twitter tienen otros sesgos, empezando con que representan a menos del 20% de la población.

Notar que aunque tengamos datos masivos, éstos no necesariamente son big data si ellos no son veraces o no tienen relación (valor) con el análisis que queremos realizar. En cada una de estas dimensiones, hay problemas inherentes a los datos, al procesamiento informático de los mismos y a las personas que los usan, como mostramos en la Figura 1, donde hemos destacado con negrita los más importantes.

En base a esta definición, no es difícil concluir que la mayoría de las organizaciones (empresas, instituciones, etc.) no tienen big data y nunca lo tendrán pues no cumplen con alguna de las tres primeras V's (ver Figura 1). De hecho, personalmente opino que el verdadero problema hoy son los datos normales, pequeños o *small data,* pues así más organizaciones podrían tener la posibilidad de aprovechar los avances en aprendizaje automático y minería de datos [1, 7].

Por otro lado, ciertamente los datos de medios sociales cumplen con las tres primeras V's y si son bien usados, también son veraces y tienen valor. Esto sigue siendo válido si nos circunscribimos a Chile. Sin embargo, cuando el análisis es realizado con una muestra de datos en un segmento de tiempo predeterminado, deja de ser big data, pues ya no posee velocidad.

La Figura 2 muestra los ocho medios sociales más usados en Chile a comienzos de 2019 [4], donde algunos son redes sociales implícitas (es decir, suponemos que si una persona conoce la identidad de otra persona, están conectados), destacadas en negrita. También mostramos la penetración de cada una de ellas considerando una población de 18,3 millones que da una penetración de Internet del 82% (15 millones de personas conectadas) que baja al 77% si consideramos usuarios activos en medios sociales (14 millones) y al 71% si solo consideramos teléfonos celulares (13 millones) [4]. Las cifras marcadas con un "*" se han calculado usando la audiencia para publicidad en Internet que representa los usuarios activos mensuales y no el número total de usuarios, a excepción de LinkedIn donde se consideran todos los usuarios chilenos registrados. También indicamos si los datos son públicos o no, destacando que Twitter es el único medio completamente público. Finalmente, recalcamos que cuatro de los cinco primeros medios sociales pertenecen a Facebook, a excepción de YouTube que pertenece a Google.

Todos estos números anteriores son aproximados por diversas razones. Primero, solo cada medio sabe exactamente el número de usuarios registrados y activos por mes en cada país. Segundo, estos usuarios incluyen organizaciones y *bots* (usuario que es un agente de software) que no son personas, además de individuos que pueden tener más de un perfil.

Sesgos demográficos en medios sociales

Si consideramos que hay 14 millones de usuarios activos en medios sociales (77%), estos usuarios también tienen sesgos demográficos. Si comparamos los porcentajes de hombres y mujeres por rangos de edad considerando la estimación de usuarios activos de Hootsuite [4] (incluyendo un ajuste de 0,7% por una suma incorrecta que da sobre el 100%) y las estimaciones a partir del

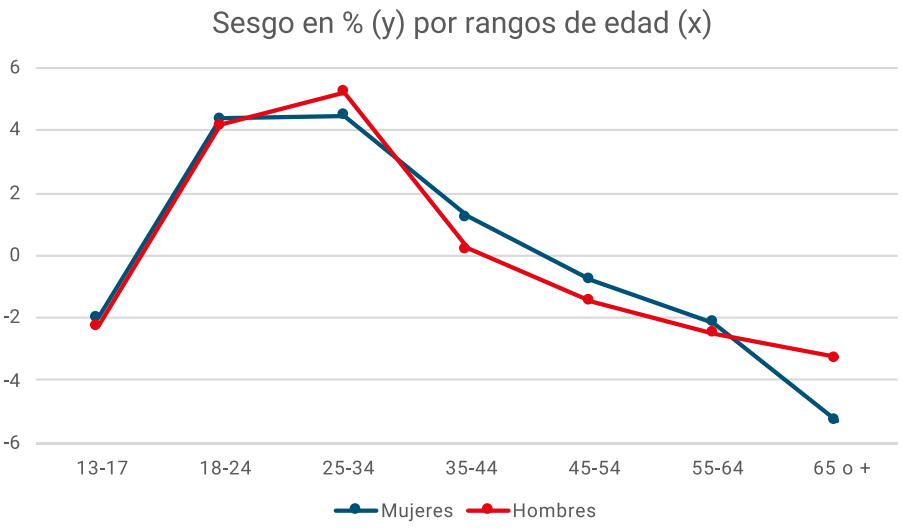


Figura 3. Sesgo de participación entre hombres y mujeres en una muestra de datos de Twitter.

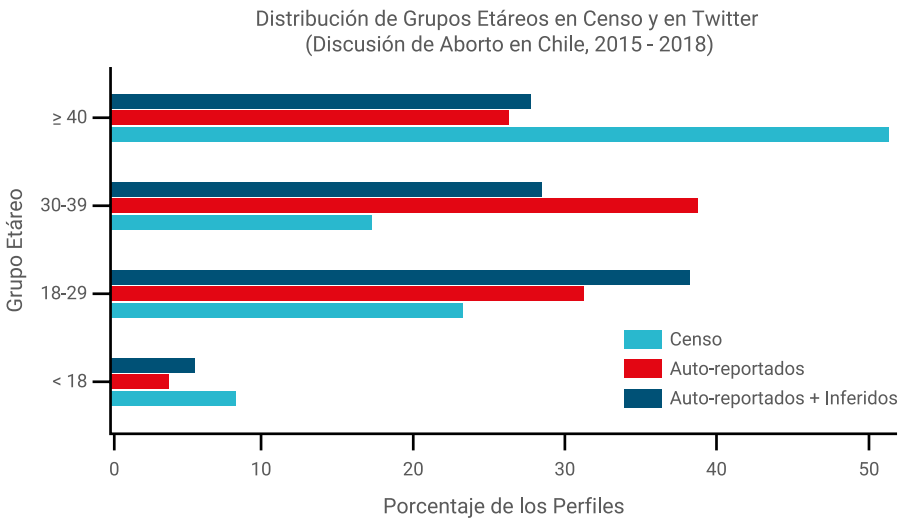


Figura 4. Perfil de los distintos grupos etáreos en Twitter (auto-reportados y autoreportados + inferidos) y en la población general chilena (censo de 2017).[1]

último censo [8], obtenemos las diferencias para personas de al menos 13 años (que excluye al 8.6% de la población) en la Figura 3.

Como podemos ver, entre 18 y 44 años (y particularmente entre 18 y 34) tenemos una sobrerrepresentación mientras que en el resto es lo opuesto. También los hombres están más representados que las mujeres en general (un 9% más en el rango de 25 a 34 años y 17% más para mayores de 64 años), pero en el rango de 35 a 64 años, son las mujeres las que están sobrerrepresentadas en mayor medida.

Considerando que la mayoría de las redes sociales son principalmente privadas, usar los datos que son públicos sesga cualquier muestra a personajes públicos o extrovertidos y en general solo temas de interés público, entre otros. Por esta razón, la mayoría de los análisis de medios sociales usa Twitter, donde potencialmente todos los datos son públicos.

Sesgos de datos en Twitter

Por supuesto los datos de Twitter tienen otros sesgos, empezando con que representan a menos del 20% de la población, ya que el número de usuarios registrados chilenos es menor a 3,5 millones si consideramos los seguidores de los medios de comunicación más populares, lo que es una cota superior de los usuarios registrados. Por otro lado, solo podemos conseguir datos de usuarios activos y, por lo tanto, considera a menos del 10% de los chilenos.

Respecto a los sesgos demográficos ya mencionados, el sesgo de género en Twitter es más pronunciado que en el promedio de todas las redes sociales, pues se estima que solo el 29% de los usuarios son mujeres [4]. Este sesgo de género es el primero que hay que mitigar, seguido de los sesgos de edad.

Otro sesgo de Twitter es que los datos de la API pública (interfaz para solicitar datos) son ya una muestra y no sabemos si Twitter los selecciona aleatoriamente o hace algún tipo de filtrado por temas (por ejemplo, eliminando contenido adulto o de incitación al odio). Además para seleccionar contenido muchas veces se usan palabras claves (*e.g.*, "estallido social") o términos temáticos (*hashtags*, e.g., "**#chiledesperto**") y por supuesto esto deja fuera a todos los usuarios que no usan estas palabras claves o términos temáticos, normalmente por pereza o ignorancia, lo que es común en personas que no se preocupan o no saben usar bien la tecnología.

Mitigando sesgos demográficos

Una forma de mitigar los sesgos demográficos es segmentar la muestra y sopesar cada segmento para que represente la población real. Esto es lo que hicimos en [3] para contrastar los cambios de opinión en Twitter con respecto al proceso de legislación sobre la ley que despenaliza el aborto en Chile y compararlos con los de la población en general. Para ello usamos los usuarios que indicaban su género y edad para entrenar modelos basados en aprendizaje automático para predecir el género y rango de edad para el resto de los usuarios. Luego sopesamos su opinión (a favor o contra el aborto) y la comparamos con la encuesta CEP de este tema en 2017, encontrando un error de solo 3% para las mujeres y de 7% para los hombres.

En la Figura 4 mostramos el proceso de mitigación para la edad, donde los datos de entrenamiento (auto-reportados) han sido usados para predecir el resto de la muestra, comparando el resultado con los datos del censo de 2017 [6].

1 | Agradecemos a Eduardo Graells por la realización de este gráfico.

Aquí vemos que los menores de 18 y los mayores de 39 están subrepresentados y, por lo tanto, se necesita multiplicar por un factor mayor a 1, mientras que entre 18 y 39 están sobrerrepresentados y necesitamos multiplicar por un factor menor que 1. En el caso de género encontramos que el 55,7% de la muestra eran hombres y 43,3% mujeres (mayor al mencionado anteriormente, seguramente por el tema en discusión, el aborto). Considerando que en el censo el 48,6% eran hombres y 51,4% mujeres, tenemos que multiplicar por 0,87 la opinión de los hombres y por 1,19 la de las mujeres, para que sean representativas.

Incluso los datos correctos pueden ser manipulados

En 1907, Mark Twain en su autobiografía, menciona que "hay tres tipos de mentiras: mentiras blancas, mentiras malditas y estadísticas", atribuyendo este texto erróneamente al primer ministro británico Disraeli. Esta frase describe el poder persuasivo de los números, el cual puede usarse para todo tipo de fines [5]. En los años sesenta, durante una charla en la Universidad de Virginia, el premio Nobel de economía de 1991, Ronald Coase, dijo que "si torturamos los datos el tiempo suficiente, ellos confesarán lo que queramos". Esto puede ser hecho de una manera burda, de una manera sutil o incluso sin

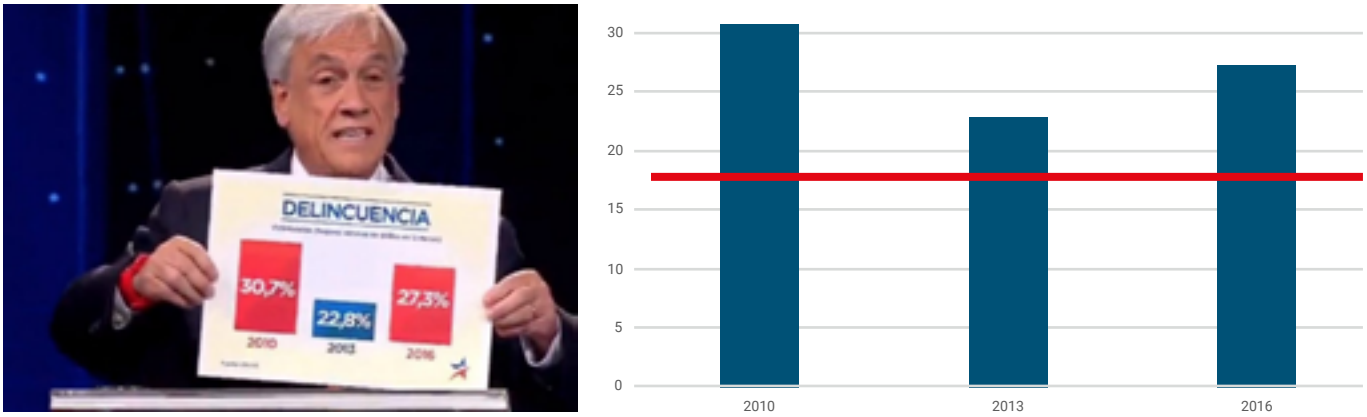


Figura 5. Izquierda: gráfico comparativo sobre delincuencia mostrado por Sebastián Piñera en el debate presidencial de 2017. Derecha: gráfico en su escala "real".

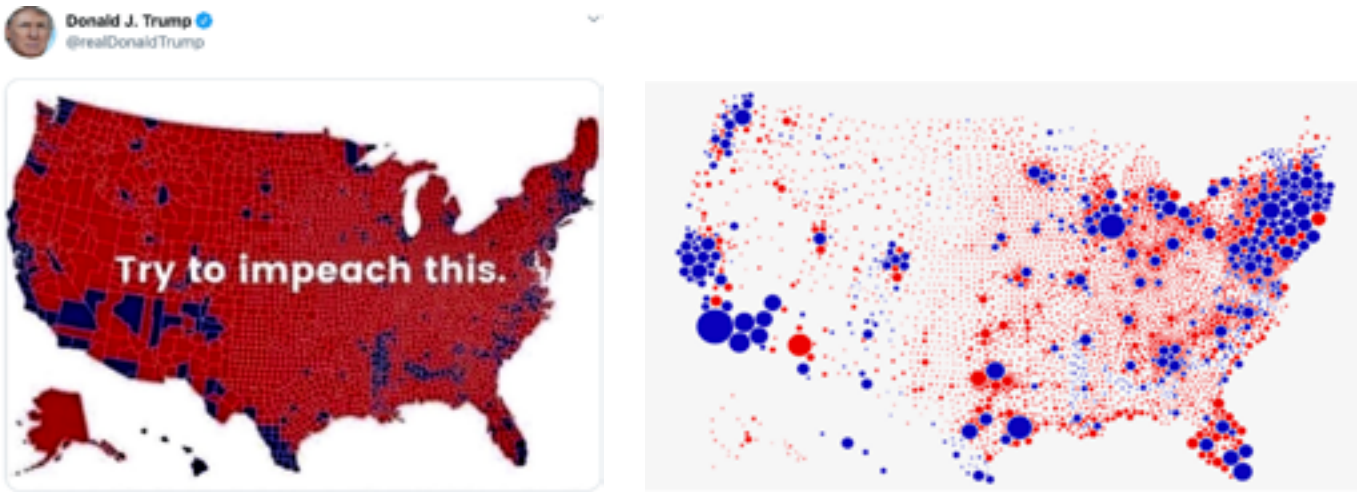


Figura 6. Izquierda: Mapa publicado por Donald Trump mostrando su apoyo en Estados Unidos. Derecha: Mapa ponderado por la densidad poblacional.

"Si torturamos los datos el tiempo suficiente, ellos confesarán lo que queramos". Ronald Coase, premio Nobel de economía.

intención. Veamos algunos ejemplos, enfatizando que datos no es lo mismo que información.

Durante un debate presidencial en 2017, Sebastián Piñera mostró el gráfico a la izquierda de la Figura 5 sobre cifras de delincuencia, donde el número durante su primer mandato parece ser la mitad de los números durante los gobiernos de Michelle Bachelet. Por supuesto esto no es efectivo pues 22,8% no es la mitad de 27,3%. Para que se visualice esto de esa manera, se ha elegido comenzar las barras desde el 18% en vez del 0, como lo muestra el gráfico de la derecha.

En octubre de 2019, el presidente de Estados Unidos, Donald Trump, publicó un tweet con el mapa a la izquierda de la Figura 6, que visualmente hace parecer que la mayoría de la población de ese país lo apoya. Sin embargo, esto supone que la densidad de población es uniforme, lo que no es cierto en ningún país. Si convertimos este mismo mapa a esferas que indican el tamaño de la población, como el que mostramos al lado derecho, vemos que la ilusión de mayoría se desvanece.

Finalmente, incluso cuando todo parece estar correcto, podemos engañarnos. Consideren el gráfico de la Figura 7 que muestra la influencia de 250 usuarios de mayor a menor (por ejemplo, el número de seguidores en Twitter). Ésta es una típica distribución de ley de potencias donde hemos incluido un 5% de usuarios extranjeros que están destacados en rojo. Escojamos ahora calcular la influencia extranjera en los 20 usuarios más populares (noten que normalmente se escogen múltiplos de 10, un sesgo antropomórfico, que es algo completamente arbitrario). Si ahora calculamos la suma de los seguidores de usuarios extranjeros comparados con los usuarios chilenos en los 20 primeros, obtenemos una influencia del 20,8%. ¿Preocupante, no?

¿Dónde está el engaño? Bien, este porcentaje en realidad depende de cuántos usuarios populares escogemos. De hecho, si hubiéramos elegido 19, habríamos encontrado el máximo posible, 21,4%. El gráfico de la Figura 8 calcula el porcentaje de influencia extranjera dependiendo del número de usuarios populares que escogemos. Si usamos los 250, llegaríamos al valor correcto que es el mínimo, 11,3%. Por supuesto, otra falacia de este análisis es suponer que seguir a alguien significa ser influenciado por esa persona, pocas veces ocurre esto e incluso muchas de las personas a las que seguimos opinan exactamente lo contrario a nosotros, pero los seguimos porque nos interesa conocer la opinión contraria. Volveremos a esto antes de terminar.

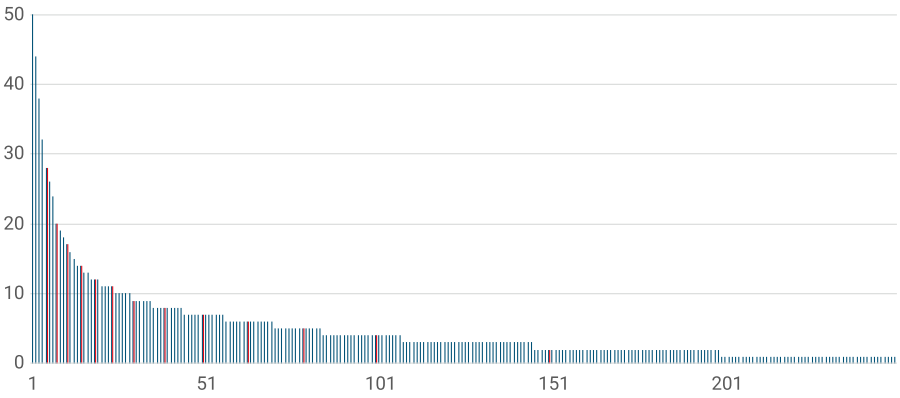


Figura 7. Influencia (medida en cantidad de seguidores) de 250 usuarios ficticios. Las barras en rojo representan usuarios extranjeros.

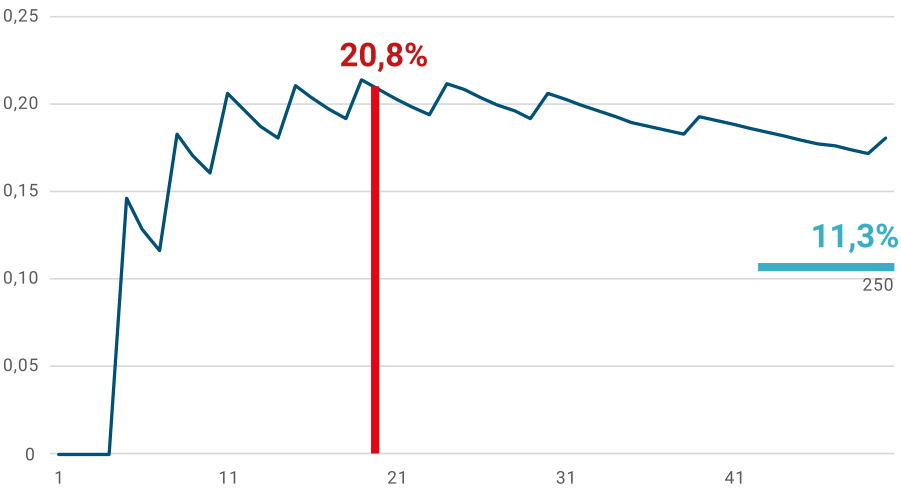


Figura 8. Fracción extranjera (eje vertical) dependiendo del número de usuarios influyentes seleccionados (eje horizontal).

Para recordar

Volviendo al famoso informe de big data, este suceso muestra también otras suposiciones que mucha gente hace sin ninguna justificación, claramente retratadas por las noticias en la Figura 9. Primero, asociación (o correlación) no implica causalidad. Por ejemplo, no es extraño que manifestantes jóvenes gusten del K-Pop, pues es un gusto típico para su edad. De allí al hecho de que eso implique una influencia coreana hay mucho trecho. Segundo, que personalidades de la música aparezcan, no significa que sean importantes, ya que presencia no siempre implica influencia. Finalmente, presencia tampoco implica tendencia, ya que podemos seguir a muchas personas de tendencias opuestas y, por lo tanto, suponer que toda posible influencia tiene el mismo mensaje es una generalización simplista y errónea.

En marzo de 2019, durante un evento en Stanford en el que participé, el famoso historiador israelí Yuval Harari dijo: "Las personas más fáciles de manipular son las que creen que no pueden ser manipuladas". Así que la próxima vez que su sesgo de confirmación le haga creer que lo que le están comunicando es cierto, recuerde esta frase e intente vencer sus sesgos cognitivos. ■

Figura 9. Selección de noticias relacionadas con el informe del big data.

Nota: Después de escribir este artículo, Alto Analytics publicó lo que parece ser un resumen del informe de big data para Chile y Colombia, el que contiene muchos errores conceptuales y analíticos, incluyendo los sesgos demográficos y el de la Figura 8. Mi análisis de este estudio se puede encontrar en Los Asombrosos Errores del Análisis de Redes Sociales Chilenas de Alto Analytics, Medium, febrero 2020, https://medium.com/@rbaeza_yates/los-asombrosos-errores-del-2b0225c2e622.

REFERENCIAS

[1] R. Baeza-Yates. BIG, small or Right Data: Which is the proper focus? KDnuggets. 2018. https://www.kdnuggets.com/2018/10/big-small-right-data.html

[2] R. Baeza-Yates. ¿Representan los medios sociales a Chile? XI Encuentro Sociedad y Tecnologías de la Información. 2020. https://www.elperiodista.cl/encuentro-big-data-y-social-media/

[3] E. Graells-Garrido, R. Baeza-Yates, M. Lalmas. How Representative is an Abortion Debate on Twitter? ACM Web Science, Boston. 2019. https://dl.acm.org/doi/10.1145/3292522.3326057

[4] Hootsuite & We Are Social. Digital 2019 in Chile. 2019. https://www.slideshare.net/-DataReportal/digital-2019-chile-january-2019-v01

[5] D. Huff. How to Lie with Statistics. W. W. Norton & Company. 1993.

[6] Instituto Nacional de Estadísticas. Censo 2017. 2018. http://www.censo2017.cl/microdatos/

[7] M. Lindstrom. Small Data: The Tiny Clues that Uncover Huge Trends. St Martin's Press, New York, EE.UU. 2016.

[8] Population Pyramid. Pirámide demográfica de Chile. 2019. https://www.populationpyramid.net/-chile/2019/