

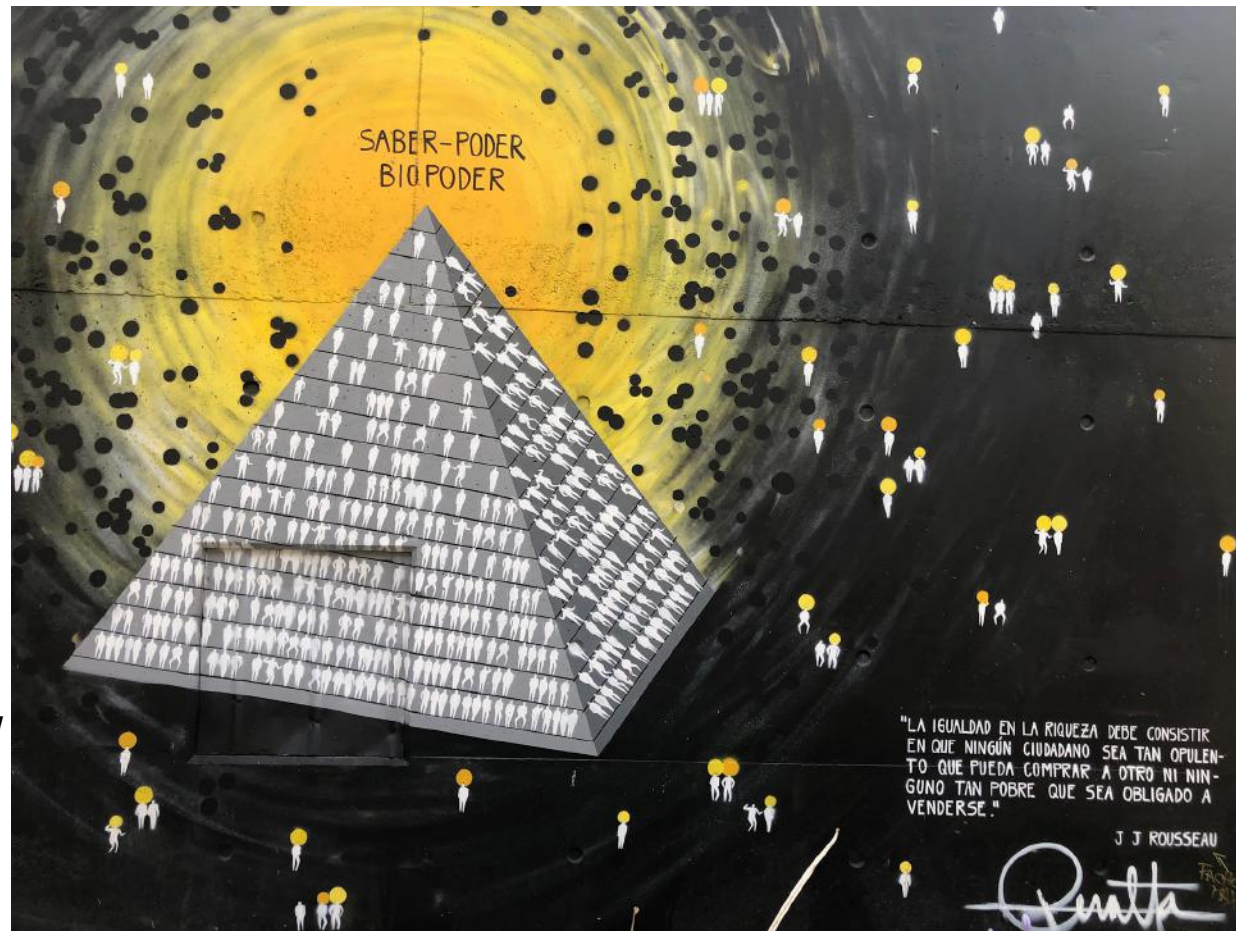
¿Representan los Datos de Medios Sociales a Chile?

¿Y es Big Data?

Ricardo Baeza-Yates

IMFD, DCC, Universidad de Chile
Northeastern University at Silicon Valley
Universitat Pompeu Fabra, Catalunya

@polarbeaRBY



[VIDEO] Fiscalía critica informe Big Data

Teletrece · El mes pasado



Tarud preocupado por imagen internacional ante informe Big Data: El Gobierno ha ridiculizado al país

BioBioChile · El mes pasado



[Video] Noticiario surcoreano toma para la chacota informe big data del Gobierno por apuntar a fanáticos del K-Pop

The Clinic · El mes pasado · Sátira



Gobierno se desmarca de polémico informe "big data": dice que llegó a la ANI

Diario Concepción · El mes pasado



Correlación no implica Causalidad

[VIDEO] Fiscalía critica informe Big Data

Teletrece · El mes pasado

#DebateLT | Reacciones al informe Big Data del Gobierno

LaTercera · El mes pasado



Tarud preocupado por imagen internacional ante informe Big Data: El Gobierno ha ridiculizado al país

BioBioChile · El mes pasado

Mon Laferte, K Pop y comediantes protagonizan informe de "Big Data" sobre disturbios en Chile

Fayerwayer · El mes pasado



Presencia no siempre es Influencia

[Video] Noticiario surcoreano toma para la chacota informe big data del Gobierno por apuntar a fanáticos del K-Pop

The Clinic · El mes pasado · Sátira

Las semejanzas con Chile del informe Big Data que analizó a los Chalecos Amarillos en Francia

Teletrece · El mes pasado



Gobierno se desmarca de polémico informe "big data" que dice que llegó a la ANI

Diario Concepción · El mes pasado

Twitter, el cuestionado big data y el fallido informe de inteligencia

pauta · El mes pasado



Las repercusiones del big data del gobierno: El hijo que nadie quiere reconocer en La Moneda

LaTercera · El mes pasado

Polémico informe de Big Data termina por sacar al jefe de la Secom del gobierno

Diario Financiero · Ayer



Karla Rubilar defiende informe del Big Data: "Hay mucha influencia extranjera que incentiva la violencia"

El Desconcierto · El mes pasado

**Presencia no siempre es influencia
ni el mismo mensaje**

Informe "Big Data" fue elaborado por la empresa española Alto Data Analytics

LaTercera · El mes pasado



El informe impresentable: menos BigData y más escucha ciudadana

El Desconcierto · El mes pasado

El big data del Gobierno: Los detalles del informe por el estallido social que entregó a la Fiscalía

LaTercera · El mes pasado



El peligro de juntar "big data", inteligencia y plataformas de redes sociales

El Desconcierto · El mes pasado

Big Data y K-Pop: la controversia por el informe que el gobierno entregó a Fiscalía

LaTercera · El mes pasado

DESPERT
DAT

¿Qué son Datos Masivos? (Big Data)

Los **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. (Wikipedia)

¿Cuán grandes? ¿Cuán complejos?

¿Qué son Datos Masivos? (Big Data)

Volumen: la cantidad de datos generados y guardados

Variedad: el tipo y naturaleza heterogénea de los datos

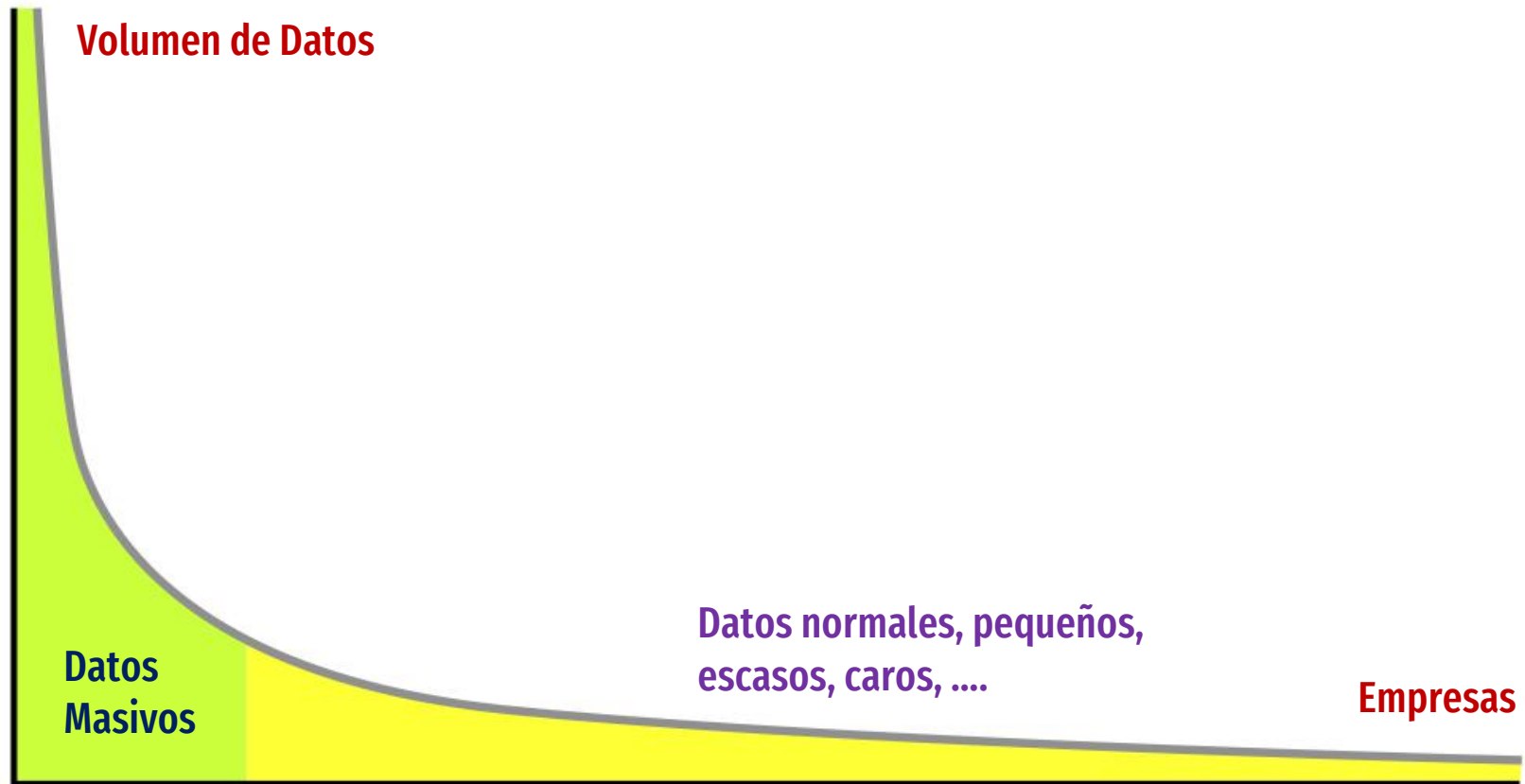
Velocidad: tasa a la cual se generan y procesan los datos

Veracidad: calidad de los datos capturados

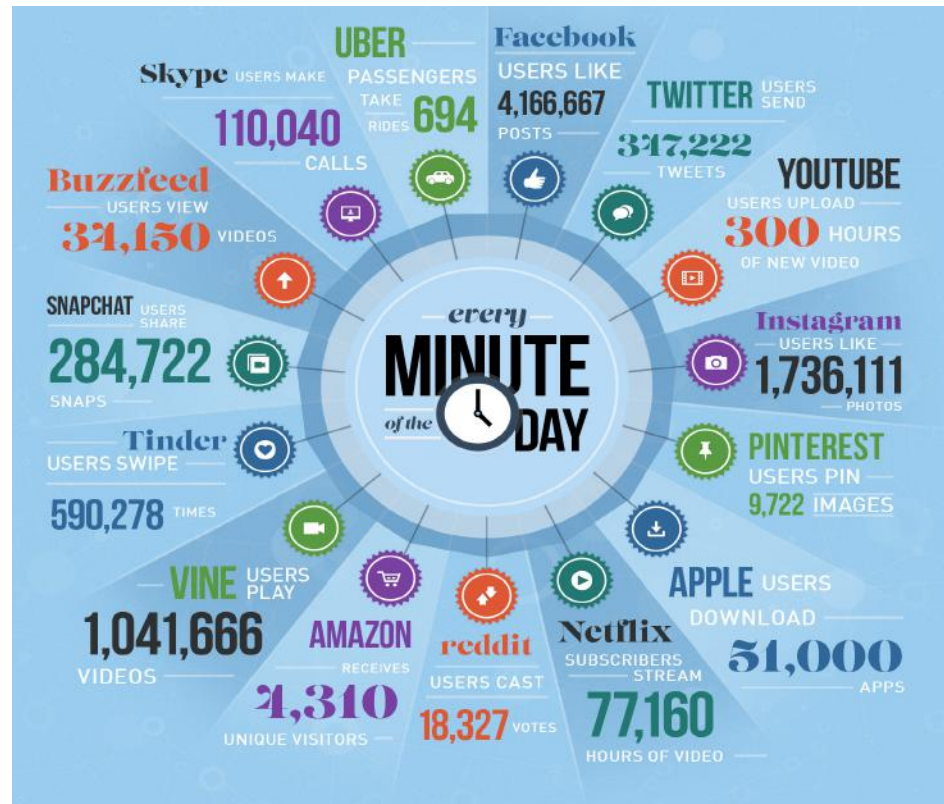
Valor: los datos generados deben ser útiles y accionables

Característica	Problema de los datos	Problema de procesamiento
Volumen	Escala, Redundancia	Escalabilidad
Variedad	Heterogeneidad, Complejidad	Adaptabilidad, Extensibilidad
Veracidad	Exhaustividad, Sesgo, Escasez, Ruido, Spam	Fiabilidad, Confianza
Velocidad	Tiempo real	En línea
Valor	Utilidad, Privacidad	Depende del negocio

¿Quiénes tienen Big Data?



Si, las Redes Sociales Generan Big Data



DOMO, 2015

Y en Chile También

Población: 19M

Internet: 15M de usuarios (79%)

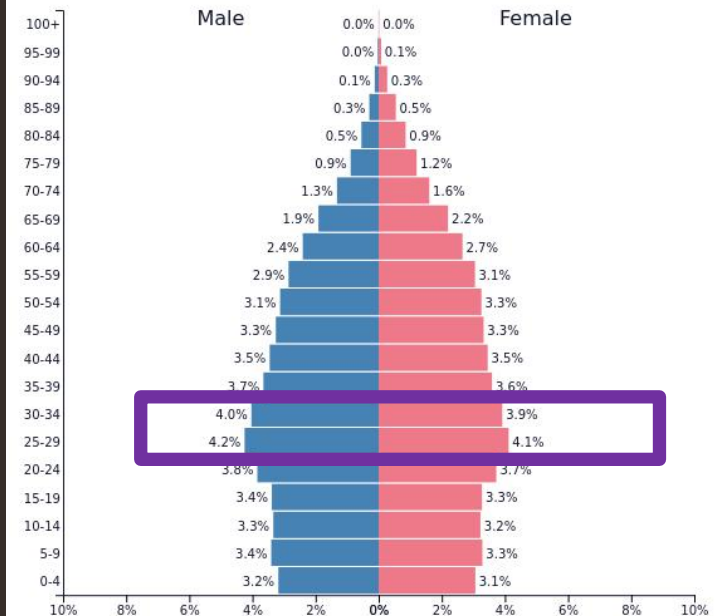
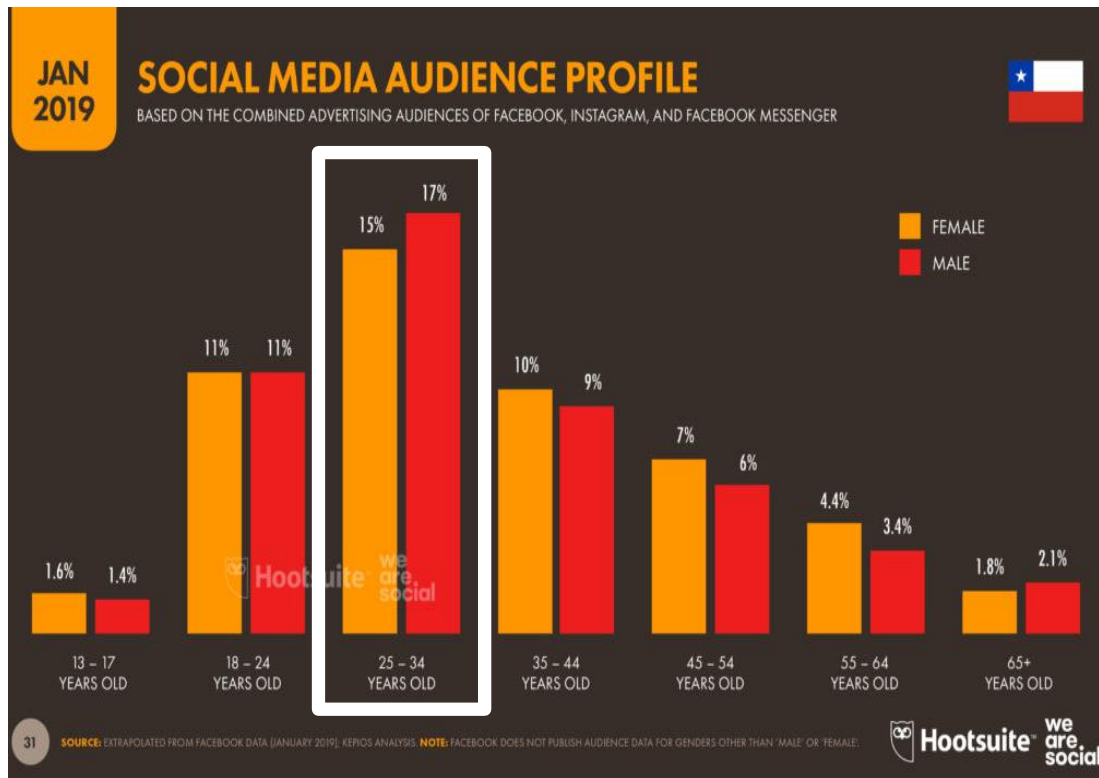
en Celulares: 14M (74%)

Red Social	¿Datos Públicos?	Usuarios Chilenos*
Whatsapp (implícita)	Ninguno	14M
Facebook	Pocos	13M
Instagram	Algunos	7,3M
YouTube	Muchos	?
Twitter	Todos	1,5M

* Audiencia publicitaria

Usuarios reales de Twitter: a lo más 3,5M (18%)

Estos Usuarios tienen Sesgos Demográficos de Género y Edad

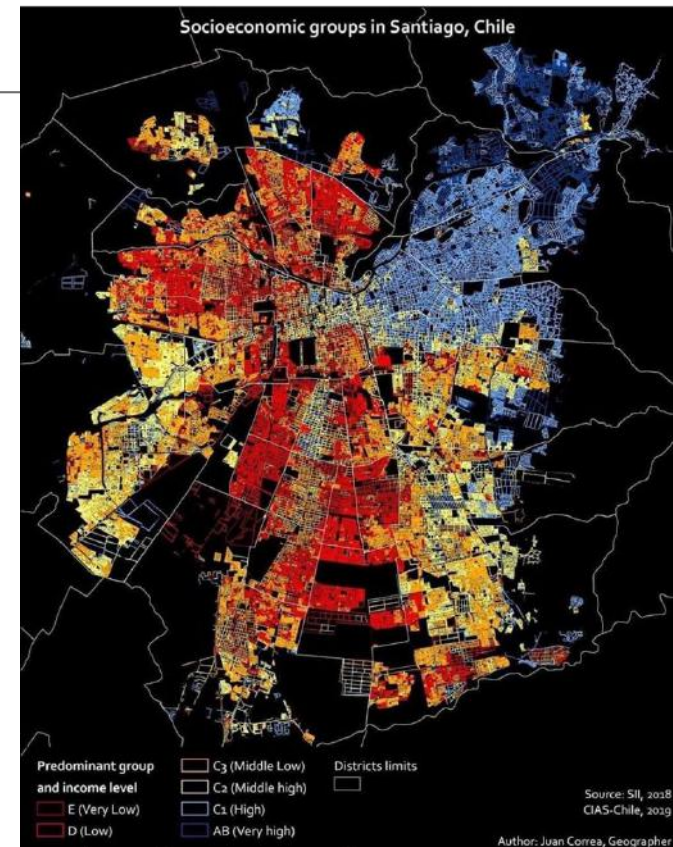
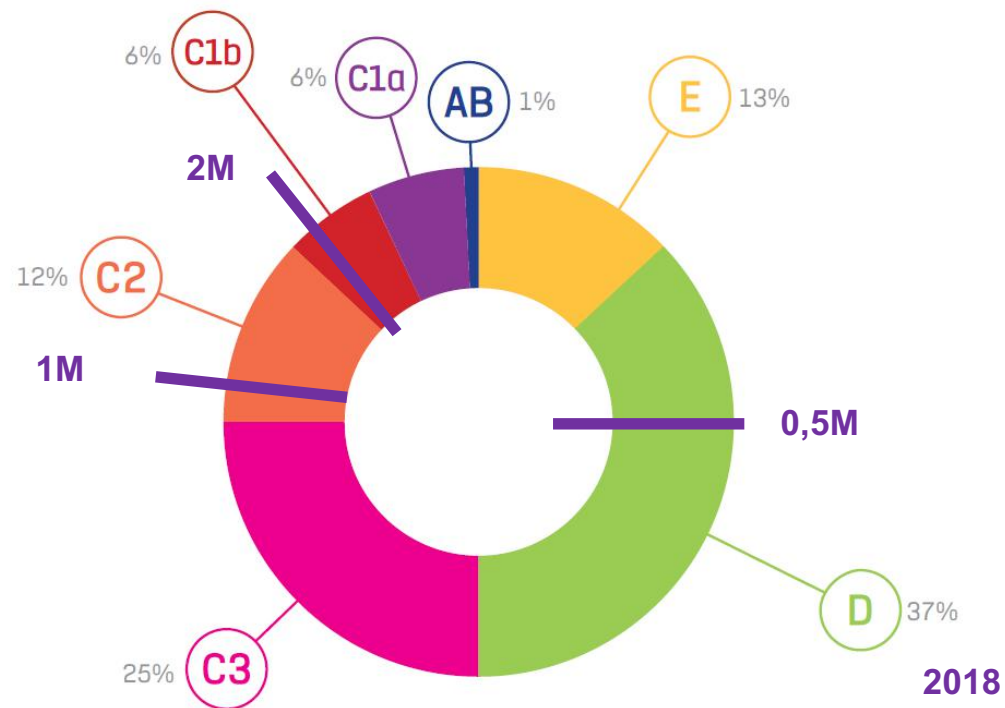


PopulationPyramid.net

Chile - 2019
Population: **18,952,035**

Y Geográficos, Económicos, Educacionales, Tecnológicos, etc.

% de hogares de cada GSE, NIVEL PAÍS.



Otros Sesgos de Muestreo

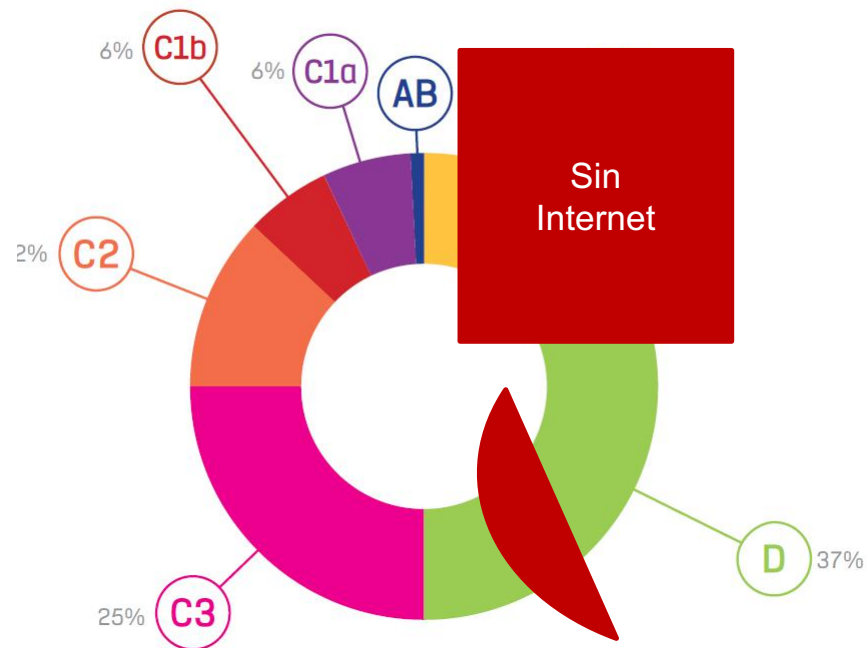
- * Los datos que entrega la API de Twitter ya son una muestra
 - * Que tweets seleccionamos: #hashtags, palabras claves, etc.
- completitud de opiniones de usuarios menos expertos

Y una muestra de big data NO es big data

Por lo tanto Twitter es una muestra sesgada



de hogares de cada GSE, NIVEL PAÍS.



¿Podemos Mitigar estos Sesgos?

- **Si, parcialmente**
- **Podemos segmentar la muestra y darle a cada segmento el peso correcto que debería tener**
 - Segmentación demográfica: género, edad, etc. y usar el último censo
- **Más segmentos, menos sesgo pero más error de muestreo**
- **La segmentación debe decidirse SIN mirar los datos (*p-hacking*)**

Análisis Longitudinal de Discusión de **Ley de Aborto** en Chile y Argentina

Eduardo Graells-Garrido
Ricardo Baeza-Yates
Mounia Lalmas

rakelilla79@twitter



**Contexto: legislación de Aborto en Chile y Argentina,
influencia entre países entre 2015-2018.**

Leyes discutidas en Twitter extensivamente.

Pero:

¿Qué tan representativa es la discusión?

¿Qué podemos aprender de ella?



Marcha en defensa de los derechos a abortar en Argentina, 2018.
Source: flickr(CC licensed).

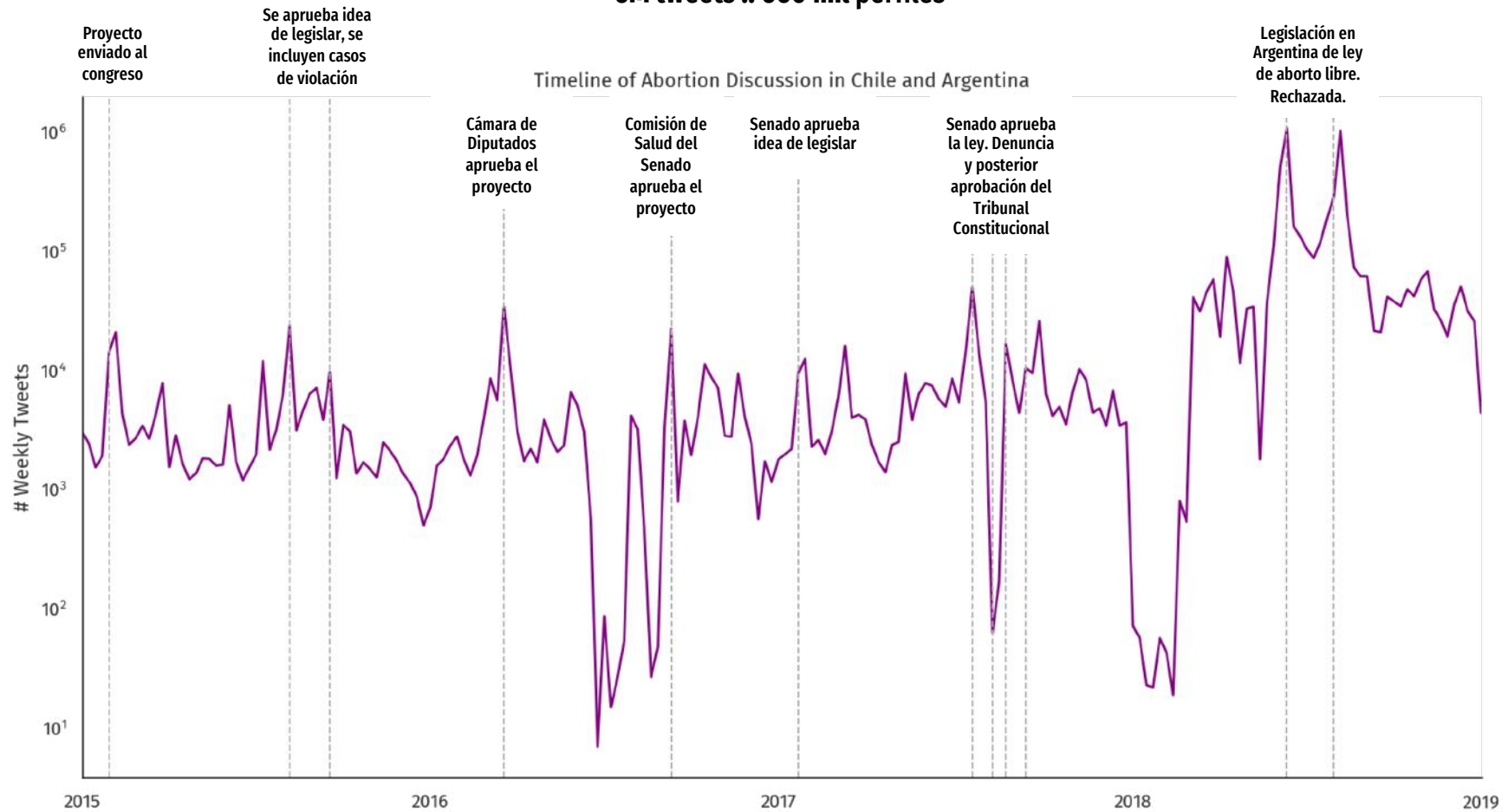


Marcha en defensa de los derechos a abortar en marzo de 2015.
Source: MovilH (CC licensed).

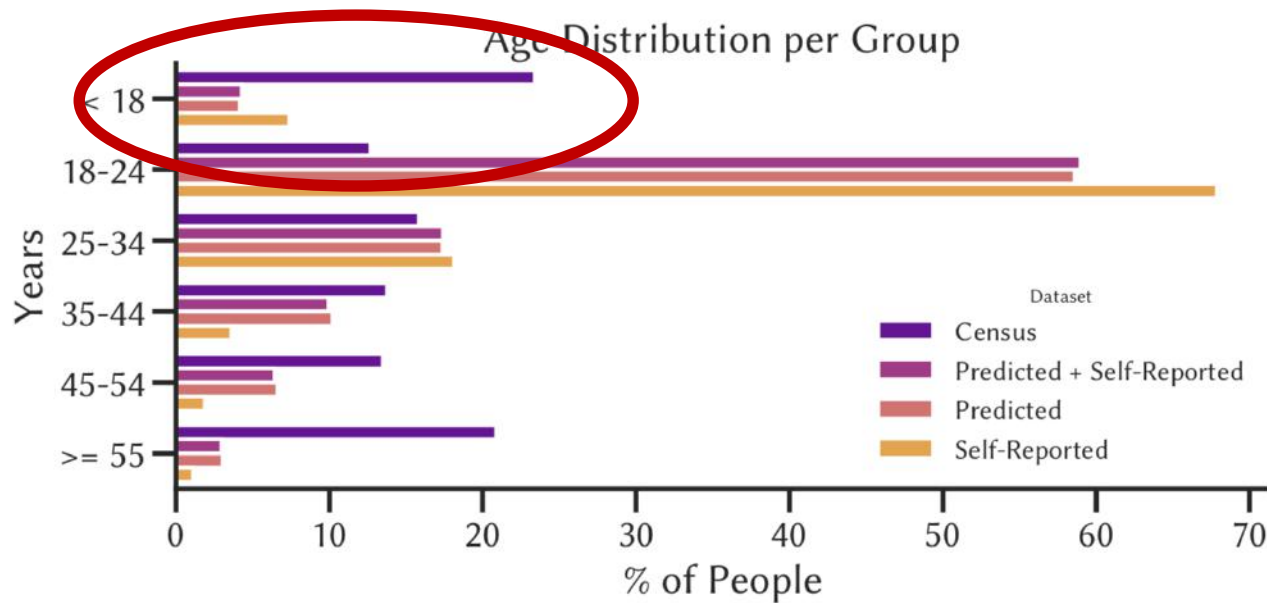


Manifestación pública en oposición a los derechos a abortar en marzo de 2016.
Source: [Radio BioBio](#).

6M tweets :: 660 mil perfiles



Segmentación en Chile: Género y edad



ACM Web Science 2019, Boston, Junio 2019



Twitter Profiles per 1,000 inhabitants

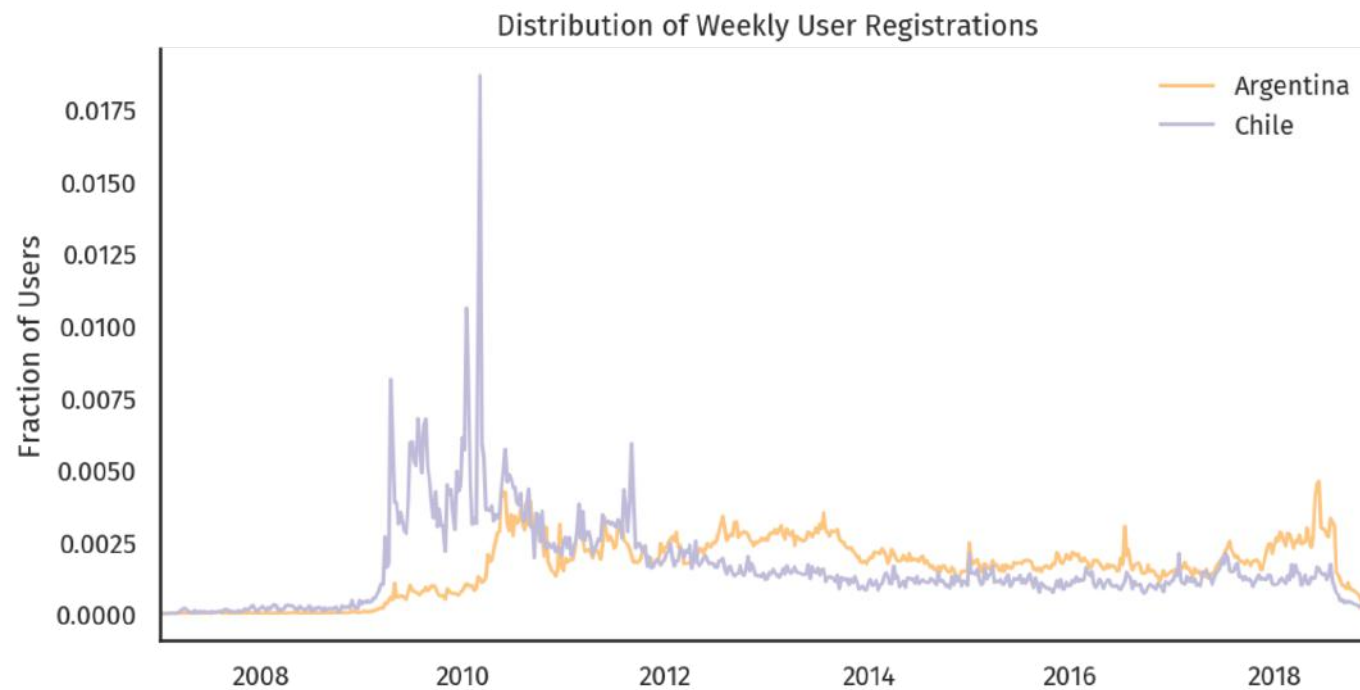
	total_gender		total_age
	female	male	
< 18	1.2	0.23	0.73
18-29	1.9	1.6	1.8
30-39	1.6	2	1.8
>= 40	0.51	0.65	0.58
total_gender	6.1	6.6	6.3



Twitter Profiles per 1,000 inhabitants

	total_gender		total_age
	female	male	
< 18	35	5.1	20
18-29	13	3.8	8.3
30-39	0.66	0.51	0.58
>= 40	0.56	0.7	0.63
total_gender	25	11	18

Argentina tiene más perfiles, y mayor representación de mujeres menores a 18 años.
Los atributos de perfiles fueron clasificados utilizando XGBoost.



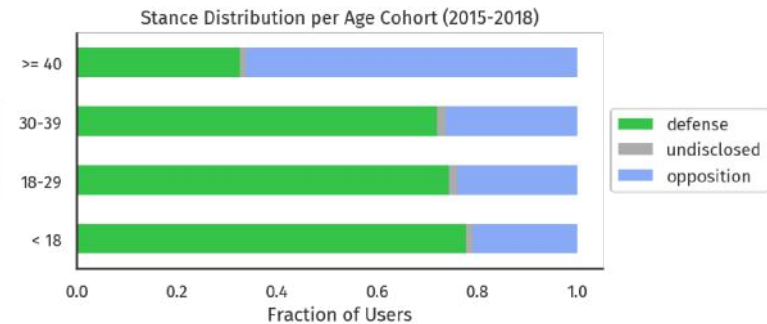
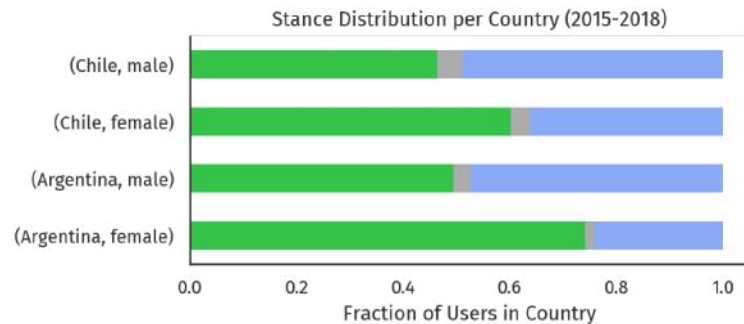
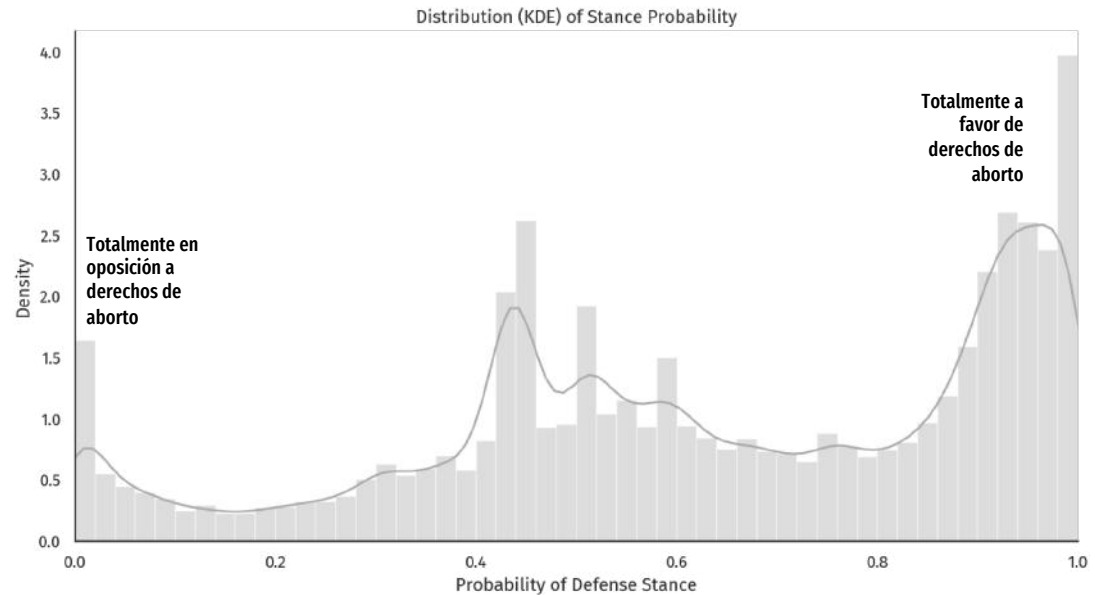
Cuentas que participaron en la discusión son en su mayoría consolidadas en la red. Solamente en el estallido social se registró más gente que en el terremoto del 27 de Febrero de 2010.

Posiciones

A pesar del sesgo poblacional, los resultados son coherentes con la encuesta CEP.

Por ejemplo:

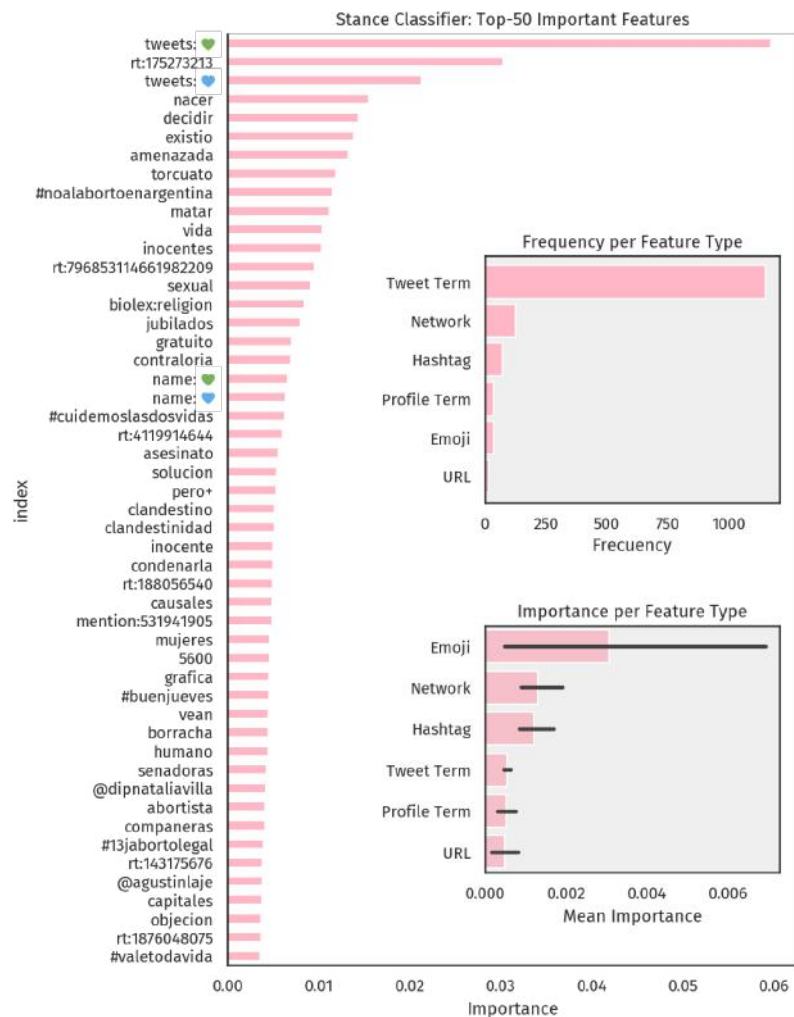
- Probabilidad de estar en oposición a los derechos de abortar es 0.06 mayor en hombres.
- Probabilidad de estar en oposición aumenta con la edad.



Comparación con Encuesta CEP

	Twitter	Weighted	CEP Survey
Intercept (female, 18–24)	1.34***	1.35***	1.30***
male	-0.29***	-0.29***	-0.23
< 18	-0.25***	-0.25***	—
25–34	-0.18***	-0.18***	-0.06
35–44	-0.27***	-0.27***	-0.31
45–54	-0.38***	-0.38***	-0.50
≥ 55	-0.48***	-0.47***	-0.58

*** $p < 0.001\%$



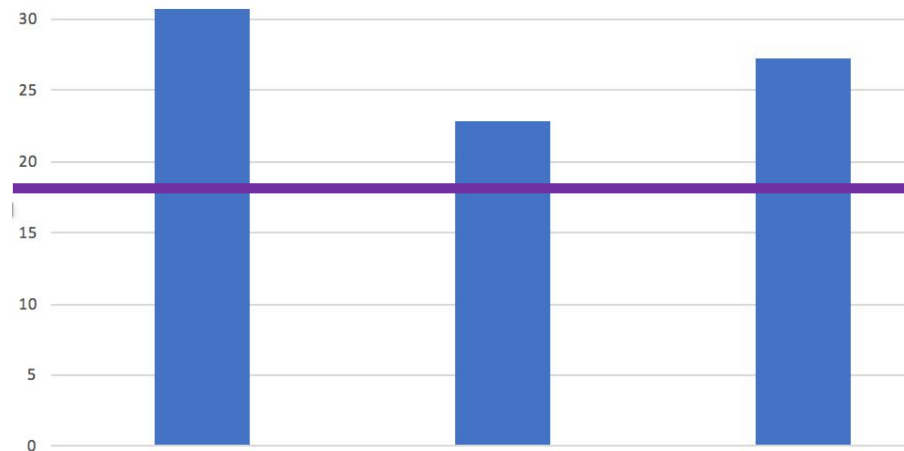
¿Qué permite identificar la posición de las personas?

El lenguaje que utilizan (*matar, decidir*) y las hashtags que difunden.

¡Pero también los emoji! Tanto en los tweets como en los nombres.

¡El corazón verde (o celeste) es equivalente al pañuelo verde (o celeste) en las manifestaciones en el mundo físico!

Incluso cuando lo hacemos bien, siempre podemos manipular la interpretación de los resultados

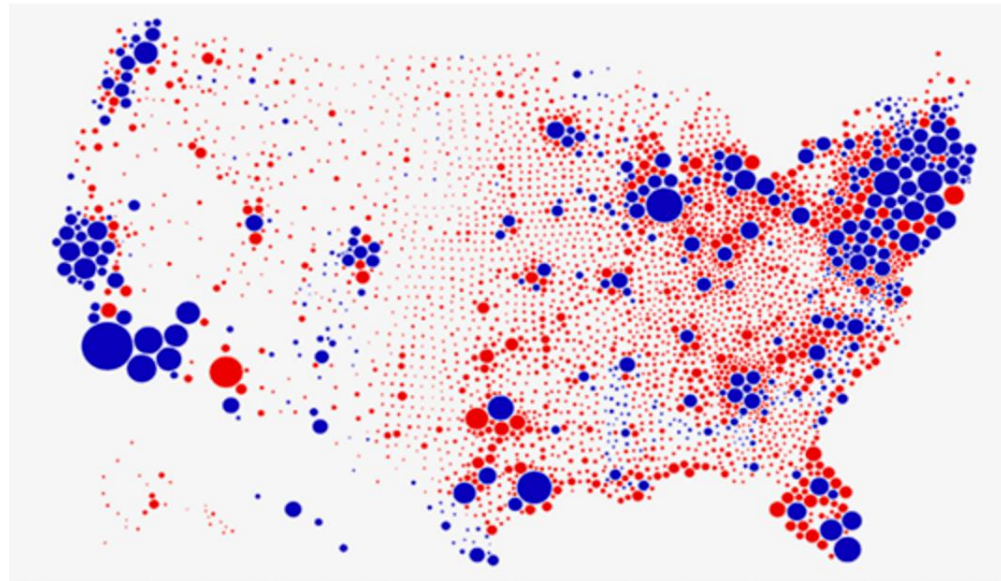


Y podemos manipular mejor

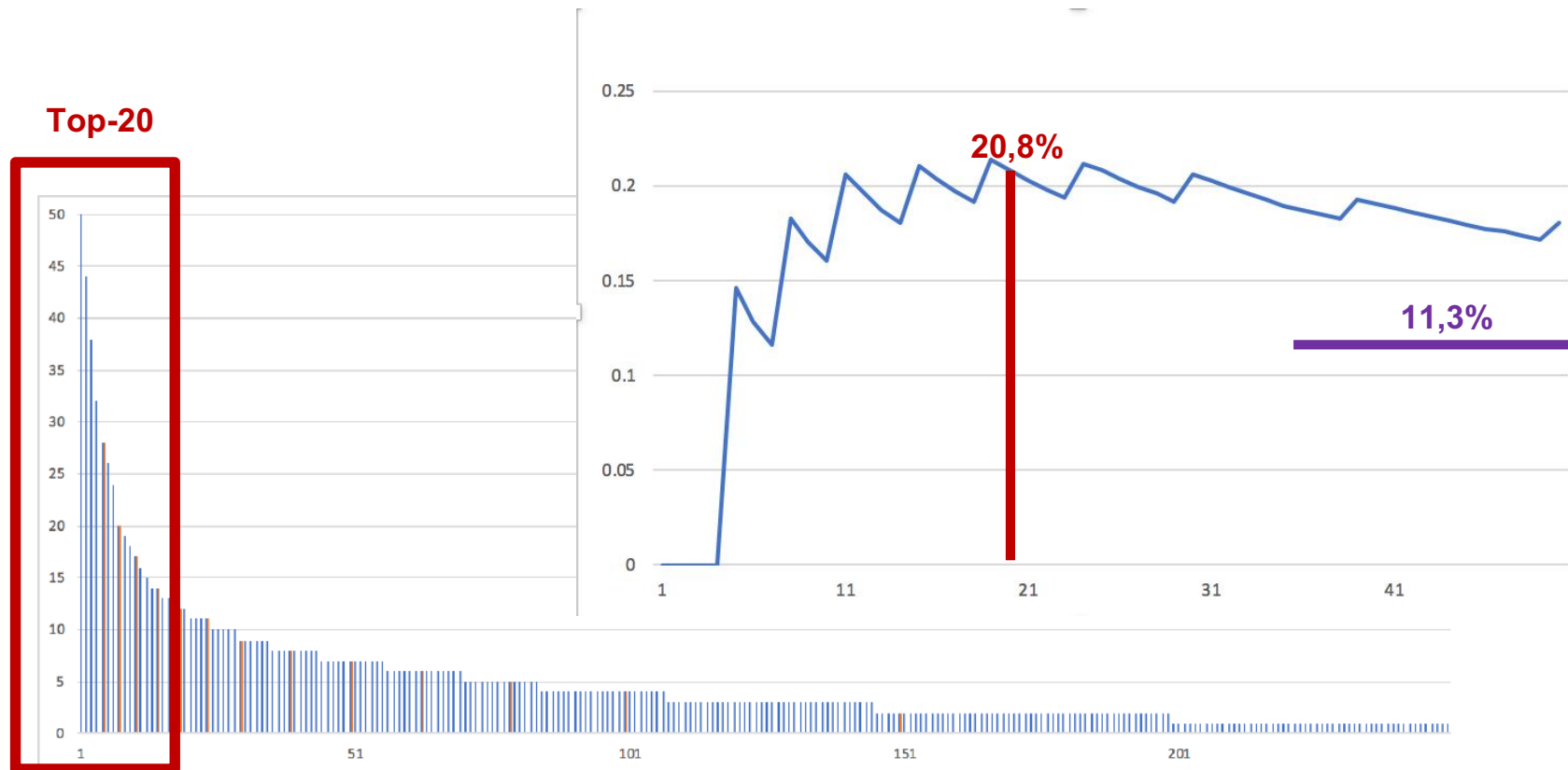


4:05 AM · Oct 1, 2019 · [Twitter for iPhone](#)

61.7K Retweets 227.4K Likes



Ejemplo más complejo: influencia en Twitter



"Torture the data enough and it will confess to anything" (1960s)
(Tortura bastante los datos y confesarán lo que quieras)

Ronald Coase, Nobel Prize Laureate in Economics (1991)

**"Las personas más fáciles de manipular son
las que creen que no pueden ser manipuladas"**

Yuval Noah Harari, 2019

[@polarbeaRBY](#)

www.ciencia.cl

www.baeza.cl