

BIG DATA

DIPLOMADO DE DATOS 2021

Clase 1: Introducción

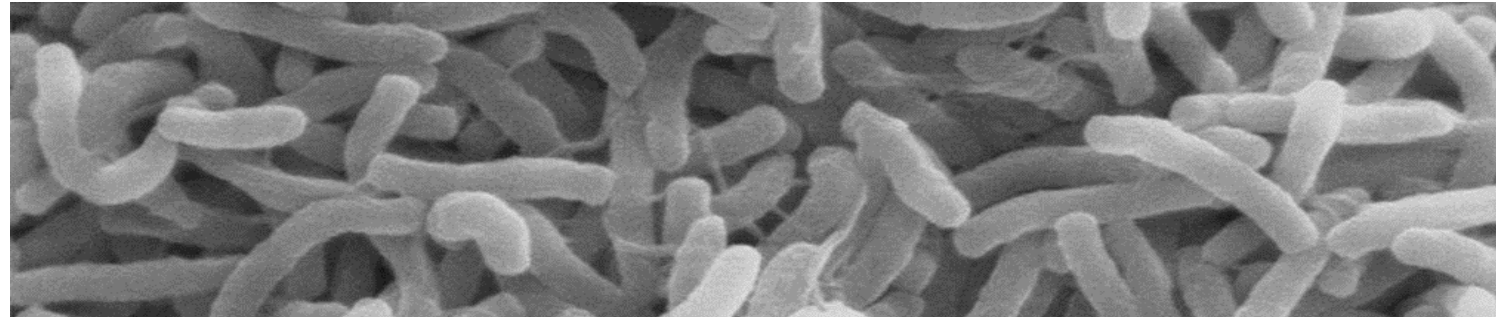
Aidan Hogan
aidhog@gmail.com

EL VALOR DE LOS DATOS

Soho, Londres, 1854

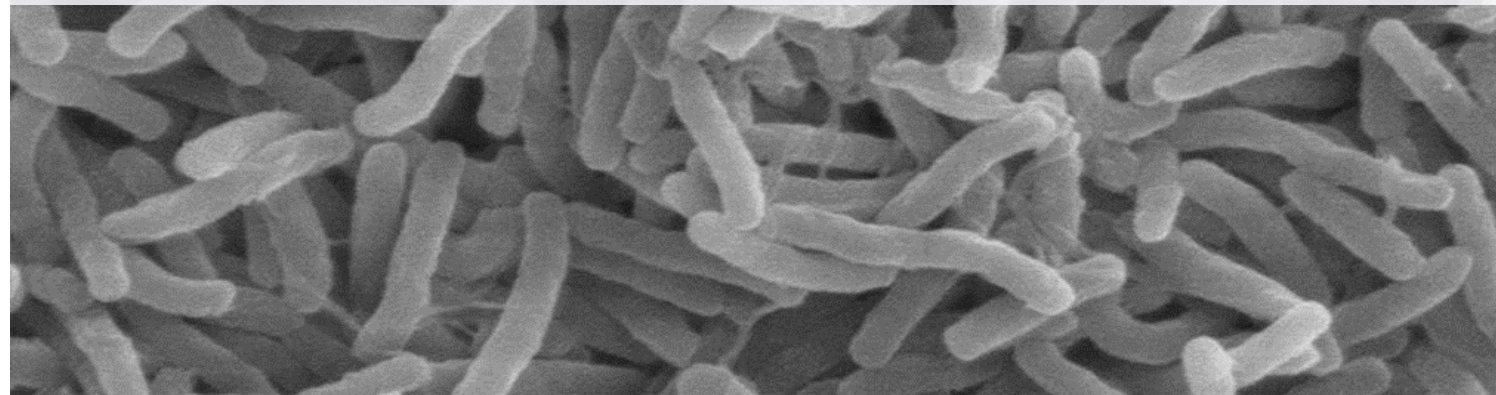


Cólera: Lo que sabemos hoy en día ...



cólera ²

1. m. **PAT.** Enfermedad infecciosa producida por una bacteria que se transmite a través de aguas contaminadas y que origina dolores abdominales, vómitos y diarreas que pueden causar la muerte.



Cólera: Lo que sabíamos en 1854



cólera ²

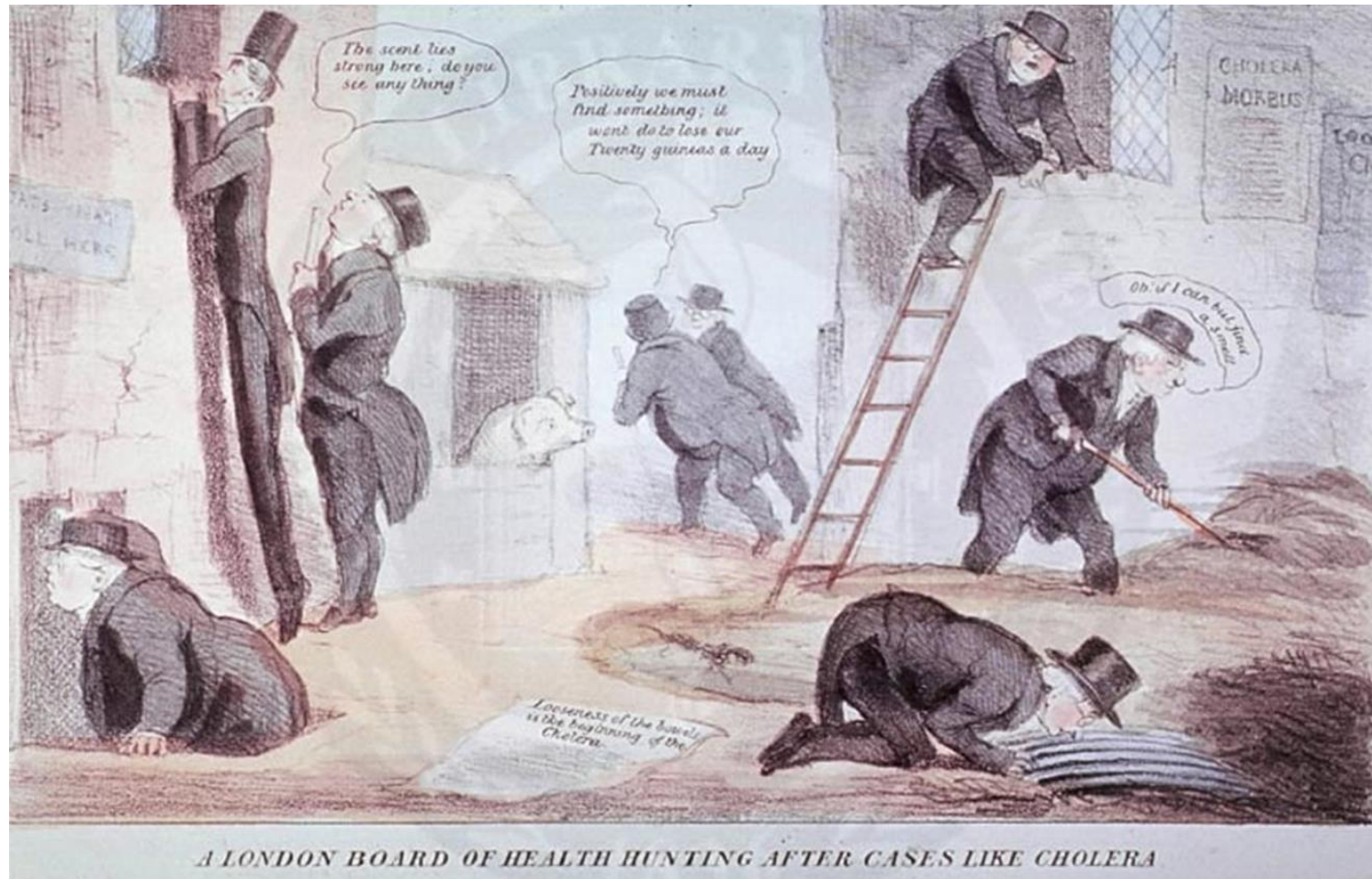
1. m. **PAT.** Enfermedad infecciosa producida por una bacteria que se transmite a través de aguas contaminadas y que origina dolores abdominales, vómitos y diarreas que pueden causar la muerte.



1854: La teoría del miasma de Galen



1854: La caza por el cólera invisible



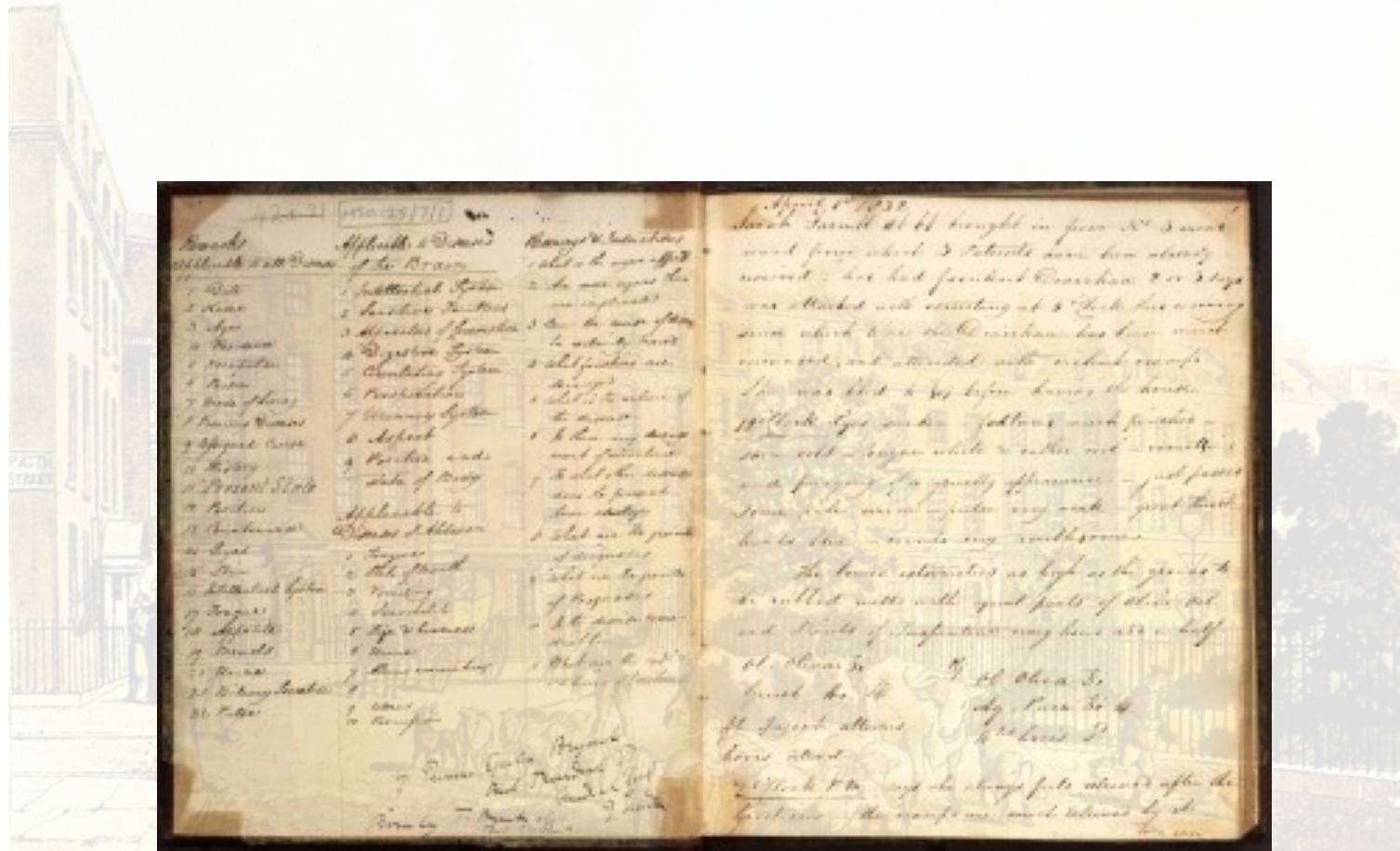
John Snow: 1813–1858



John Snow: 1813–1858



La encuesta de Soho



Recolección de datos ...

Registration Districts.	Registration Sub-Districts.	Population in 1851.	Estimated population supplied with water as under.			Deaths from cholera in 1854.		Calculated mortality in the population, supplied with water as under.			
			Southwark and Vauxhall Co.	Lambeth Co.	Both Companies together.	Total deaths.	Deaths per 10,000 living.	Southwark and Vauxhall Co. at 100 per 10,000.	Lambeth Co. at 57 per 10,000.	The two Companies.	Calculated deaths per 10,000 supplied by the two Companies.
St. Saviour, Southw.	1. Christchurch	10,022	2,915	13,234	16,149	113	71	46	36	82	57
	2. St. Saviour	10,709	10,337	898	17,235	378	192	261	2	263	153
St. Olave	1. St. Olave	8,015	8,745	0	8,745	161	201	140	0	140	160
	2. St. John, Horselydown	11,360	9,300	0	9,300	152	134	150	0	150	160
Bermondsey	1. St. James	18,899	23,173	603	23,866	362	192	370	2	372	156
	2. St. Mary Magdalen . .	13,934	17,258	0	17,258	247	177	276	0	276	160
	3. Leather Market	15,295	14,003	1,092	15,095	237	155	224	3	227	150
St. George, Southw.	1. Kent Road	18,126	12,630	3,997	16,627	177	98	202	11	213	134
	2. Borough Road	15,862	8,937	6,672	15,609	271	171	143	18	161	104
	3. London Road	17,836	2,872	11,497	14,369	95	53	46	31	79	55
Newington	1. Trinity	20,922	10,132	8,370	18,502	211	101	102	22	124	99
	2. St. Peter, Walworth . .	29,861	14,274	10,724	24,998	391	131	228	29	257	103
	3. St. Mary	14,033	2,983	5,484	8,467	92	66	48	15	63	74

CHOLERA AND THE WATER SUPPLY

Lo que los datos mostraron ...



the from cholera
in 1854.

Total deaths	Deaths per 10,000 living.
113	71
378	192
161	201
152	134
362	192
247	177
237	155
177	98
271	171
55	53
211	161
391	131
02	66



228	29	257	103
48	15	63	74

616 muertes, 8 días después ...



Lo que aprendimos ...



cólera ²

1. m. **PAT.** Enfermedad infecciosa producida por una bacteria que se transmite a través de aguas contaminadas y que origina dolores abdominales, vómitos y diarreas que pueden causar la muerte.



Cartel cólera ca. 1866 (aviso de hervir el agua)

CHOLERA
AND
WATER.

BOARD OF WORKS
FOR THE LIMEHOUSE DISTRICT,
Comprising Limehouse, Ratcliff, Shadwell,
and Wapping.

The INHABITANTS of the District within
which CHOLERA IS PREVAILING, are
earnestly advised

NOT TO DRINK ANY WATER
WHICH HAS NOT
PREVIOUSLY BEEN BOILED.

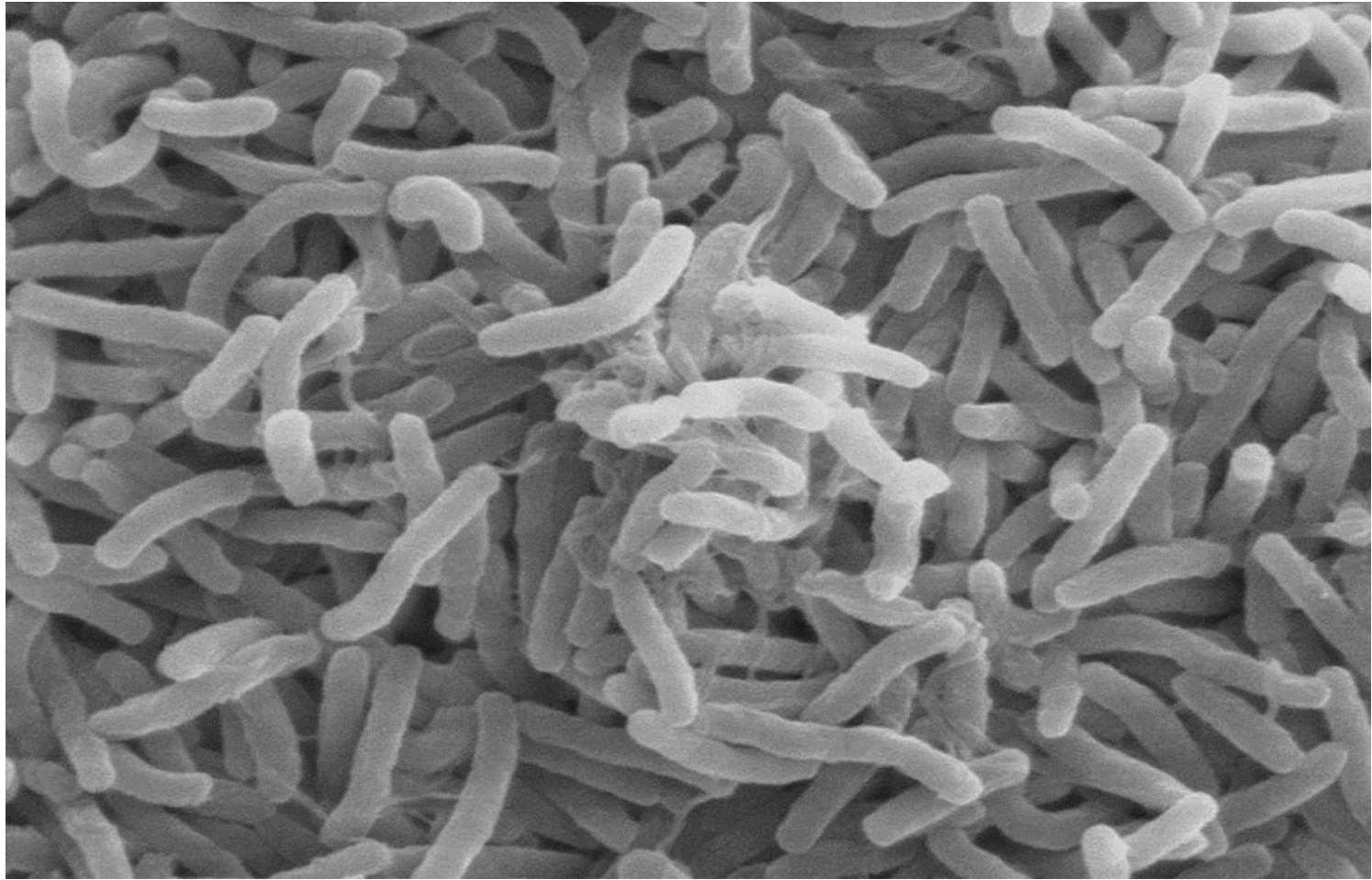
Fresh Water ought to be Boiled every
Morning for the day's use, and what
remains of it ought to be thrown away
at night. The Water ought not to stand
where any kind of dirt can get into it,
and great care ought to be given to see
that Water Butts and Cisterns are free
from dirt.

BY ORDER,
THOS. W. RATCLIFF,
CLERK OF THE BOARD.

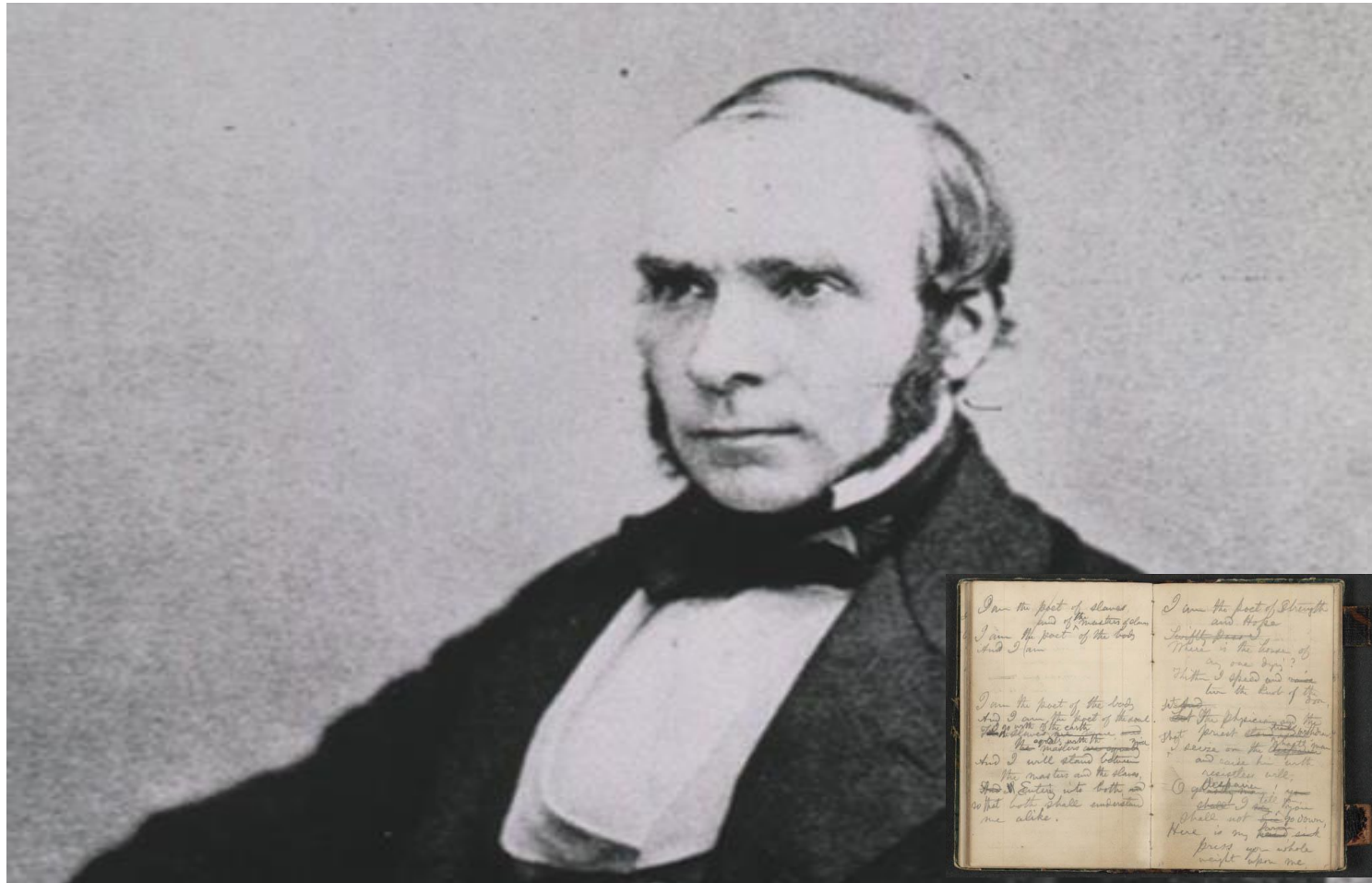
Head Office, White Horse Street,
St. James 1866.



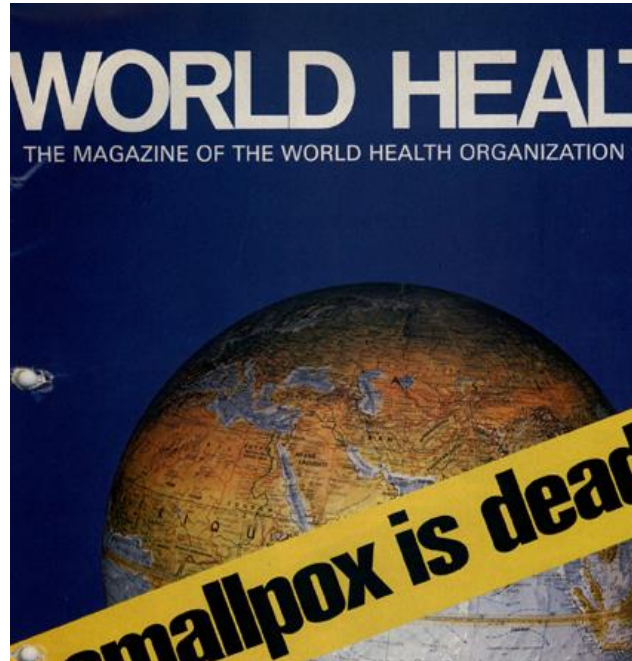
30 años antes (del descubrimiento) de *V. cholerae*



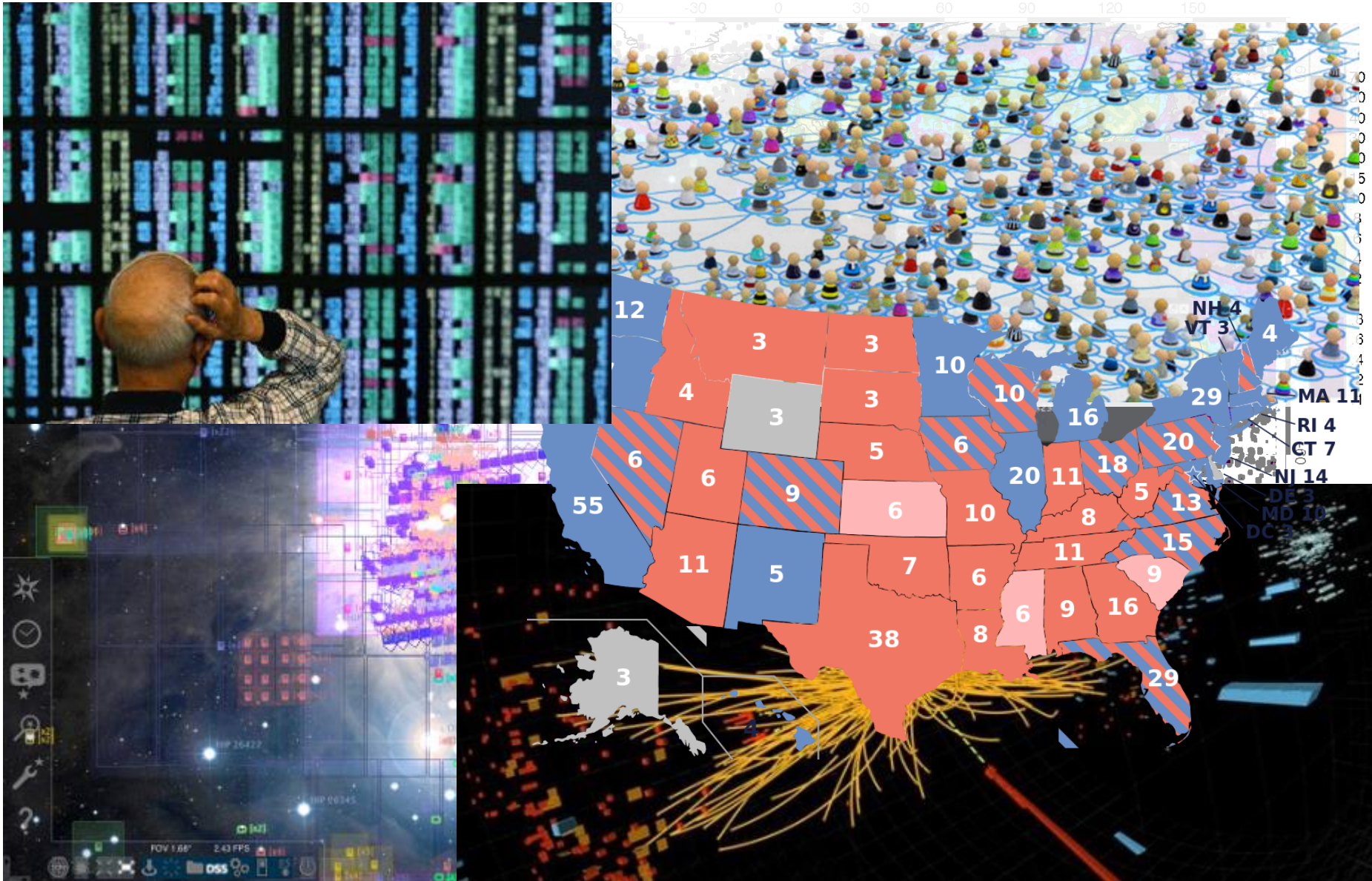
John Snow: El padre de la Epidemiología



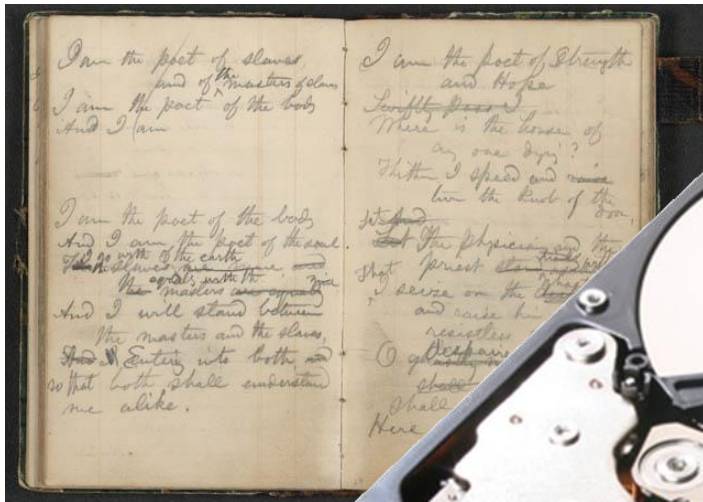
Historias de éxitos de la Epidemiología



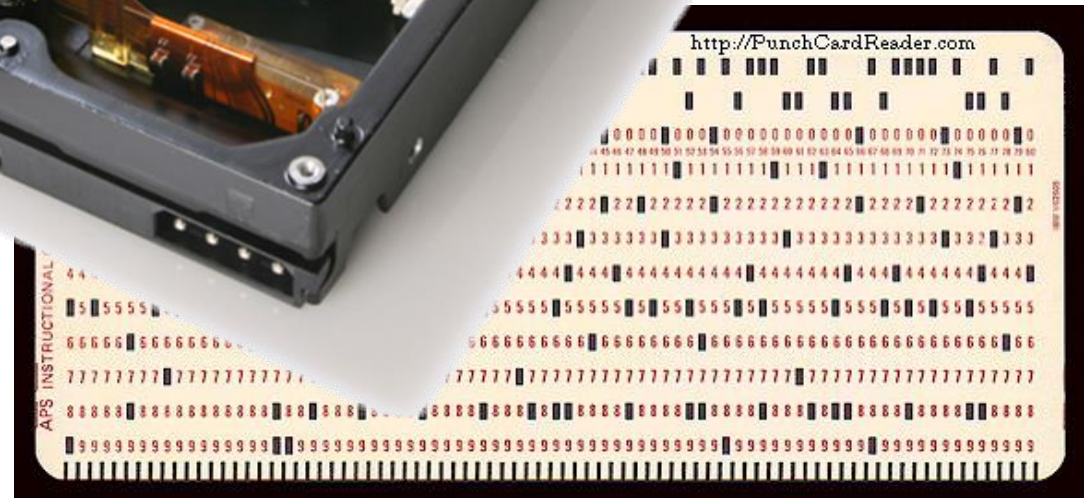
Valor de los datos: No sólo Epidemiología



Cuadernos no son suficientemente buenos



Download from
Download.com



EL CRECIMIENTO DE LOS DATOS

“Big Data”



English Wikipedia

≈ 51 GB de datos

(2015 dump)

(Texto; Datos actuales)

(XML; no comprimido)

WIKIPEDIA
The Free Encyclopedia

1 Wiki = 1 Wikipedia

“Big Data”

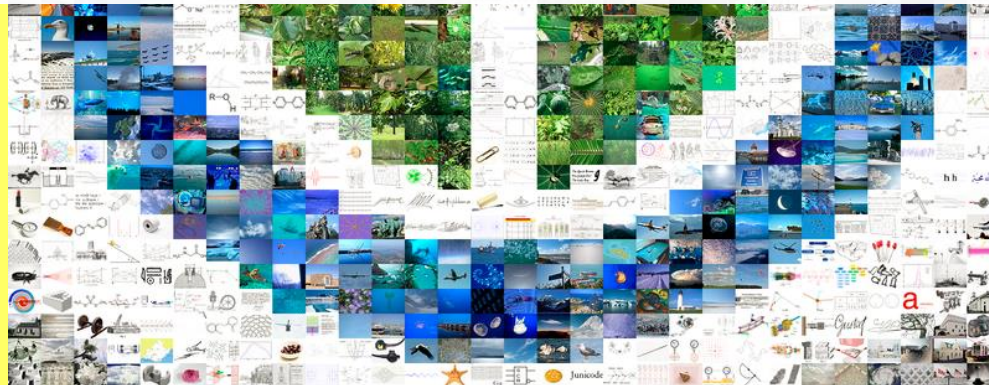


Wikimedia Commons

≈ 24 TB de datos

≈ 470.6 Wiki

(2014 dump)



“Big Data”

Twitter

≈ 8 TB / día

≈ 157 Wiki / día

(2013, generados)

The Twitter logo, a light blue silhouette of a bird in flight, is centered in the background of the slide.

twitter

“Big Data”



Large Synoptic Survey Telescope

≈ 15 TB / día (noche)

≈ 294 Wiki / día

(2020, generados)



“Big Data”

A large, semi-transparent watermark of the Facebook logo (a white lowercase 'f' on a blue circular background) is centered on the slide. The background of the slide is split into a yellow top half and a white bottom half.

Facebook

≈ 600 TB / día

≈ 11,764 Wiki / día

(2014, entrada, datos en Hive)

“Big Data”



Large Hadron Collider

≈ 1 PB / día

≈ 19,607 Wiki / día

(2017, datos filtrados)



“Big Data”

PRISM: Vigilancia de la NSA
≈ 29 PB / día
≈ 568,627 Wiki / día
(2013, procesados)



“Big Data”



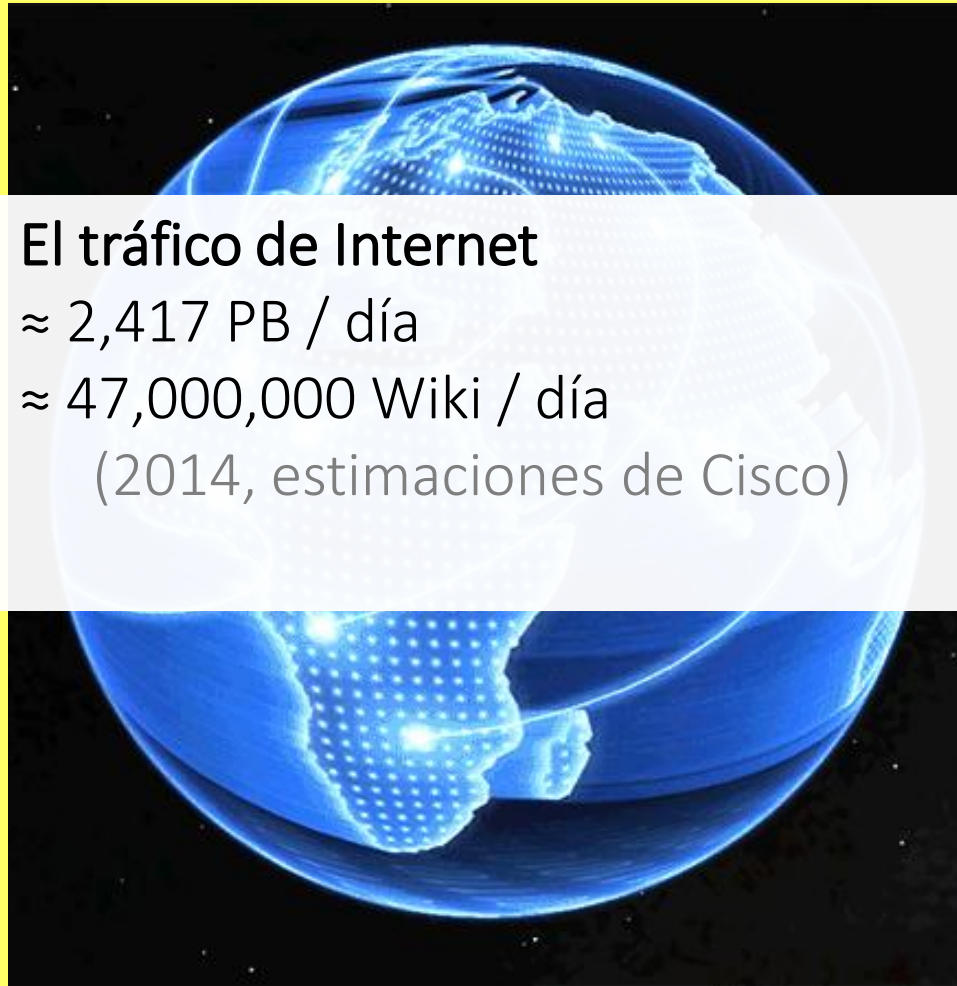
Google

≈ 100 PB / día

≈ 2,000,000 Wiki / día

(2014, procesados)

“Big Data”



El tráfico de Internet

≈ 2,417 PB / día

≈ 47,000,000 Wiki / día

(2014, estimaciones de Cisco)

Los datos: Un cuello de botella moderno?



Las 'V's de "Big Data"



“BIG DATA” NECESITA

“GESTIÓN DE DATOS (MASIVOS)” ...

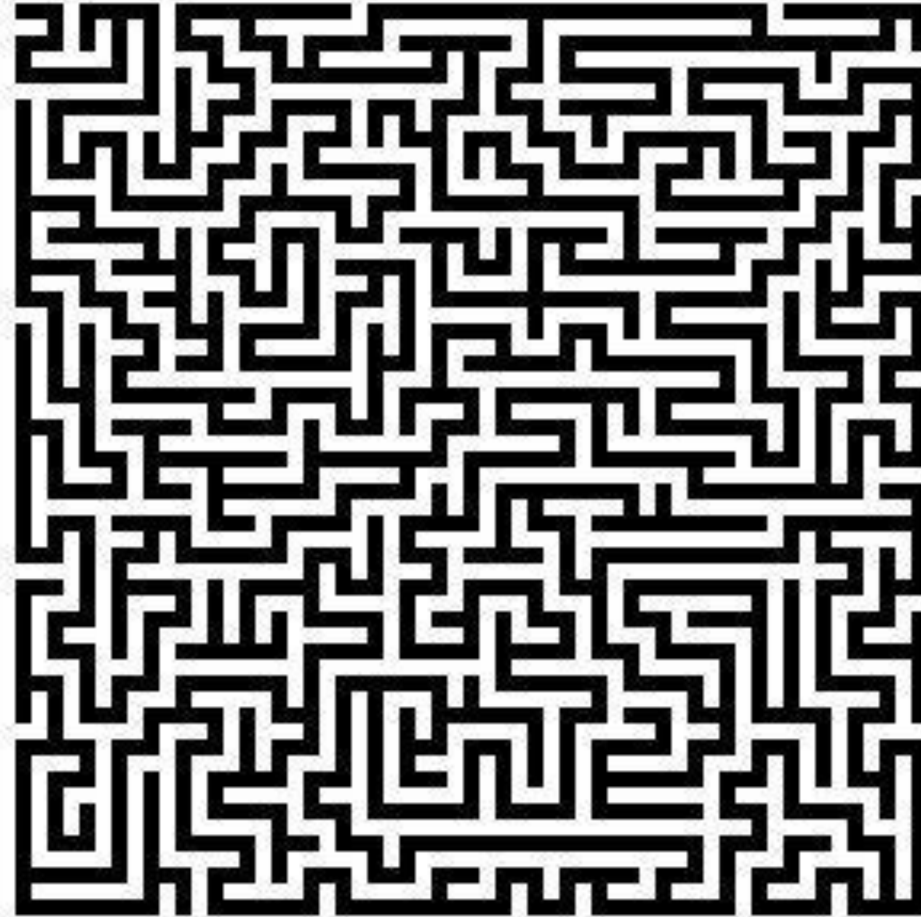
Cada aplicación es distinta ...

- **Datos** pueden ser
 - (semi-)estructurados
 - (Relational DBs, JSON, XML, CSV)
 - sin estructura
 - (documentos de texto, tweets, comentarios)
 - y cualquier cosa entre medio!

Cada aplicación es distinta ...

- **Procesamiento** puede involucrar
 - Gestión de Datos Estructurados
 - ([indexación](#), [consultas](#), [joins](#), [agregación](#))
 - Procesamiento de Lenguaje Natural
 - ([búsqueda de texto](#), [clasificación de texto](#), [análisis de sentimiento](#), [relevancia y similitud](#), etc.)
 - Minería de Datos y Aprendizaje
 - ([regresión](#), [reconocimiento de patrones](#), [clasificación](#), [detección de eventos](#), etc.)
 - Y cualquier cosa entre medio.

¿Por dónde deberíamos empezar?



GESTIÓN DE DATOS (MASIVOS)

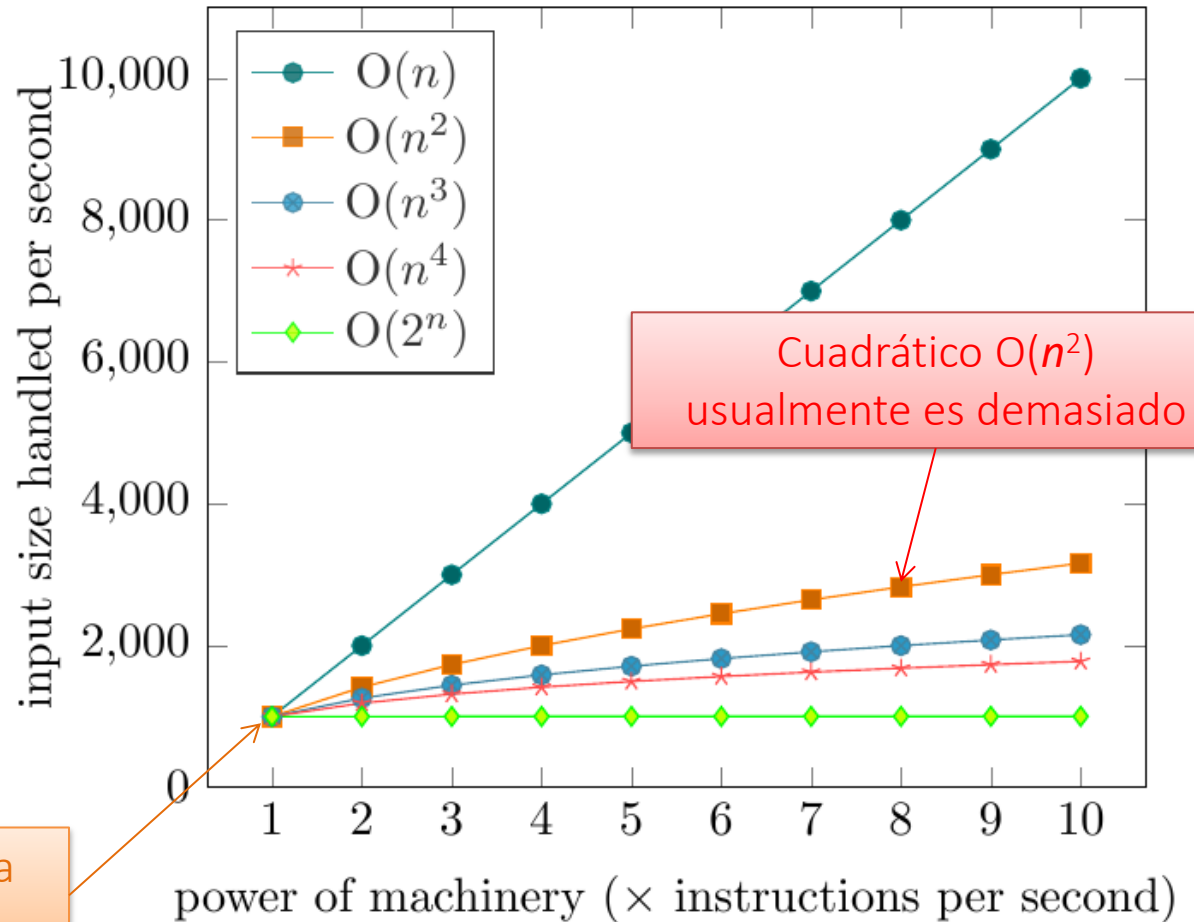
La escala es un factor importante ...

Tengo un algoritmo. 

Tengo una máquina que puede procesar 1.000 entradas por hora.

Si compro una máquina que es n veces más potente, ¿cuántas entradas puedo procesar?

¡Depende del algoritmo! 



Nota: No la misma máquina!

La escala es un factor importante ...

- ¿Una máquina que es n veces más potente?
- ¿ n máquinas que son igualmente potentes entre ellas?

VS.



¿Cuál es mejor?

¡Depende de la aplicación!



La escala es un factor importante ...

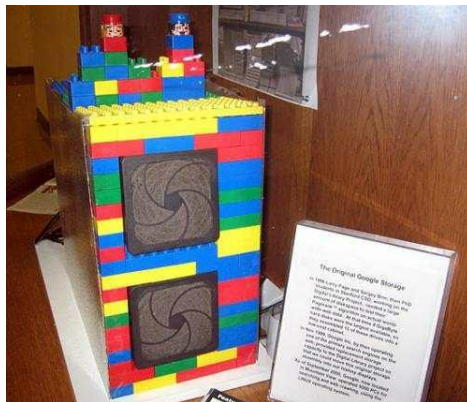
- Intensivo en los datos (nuestro foco!)
 - Algoritmos baratos / Grandes entradas
 - p.ej., Google, Facebook, Twitter
- Intensivo en computo (no es nuestro foco!)
 - Algoritmos más caros / Entradas más pequeñas
 - p.ej., simulaciones de clima, ajedrez, etc.
- No es blanco y negro

"GESTIÓN DE DATOS (MASIVOS)" NECESITA
"COMPUTACIÓN DISTRIBUIDA"

Computación distribuida

- Necesita más de una máquina
- Google ca. 1998:

GOOGLE

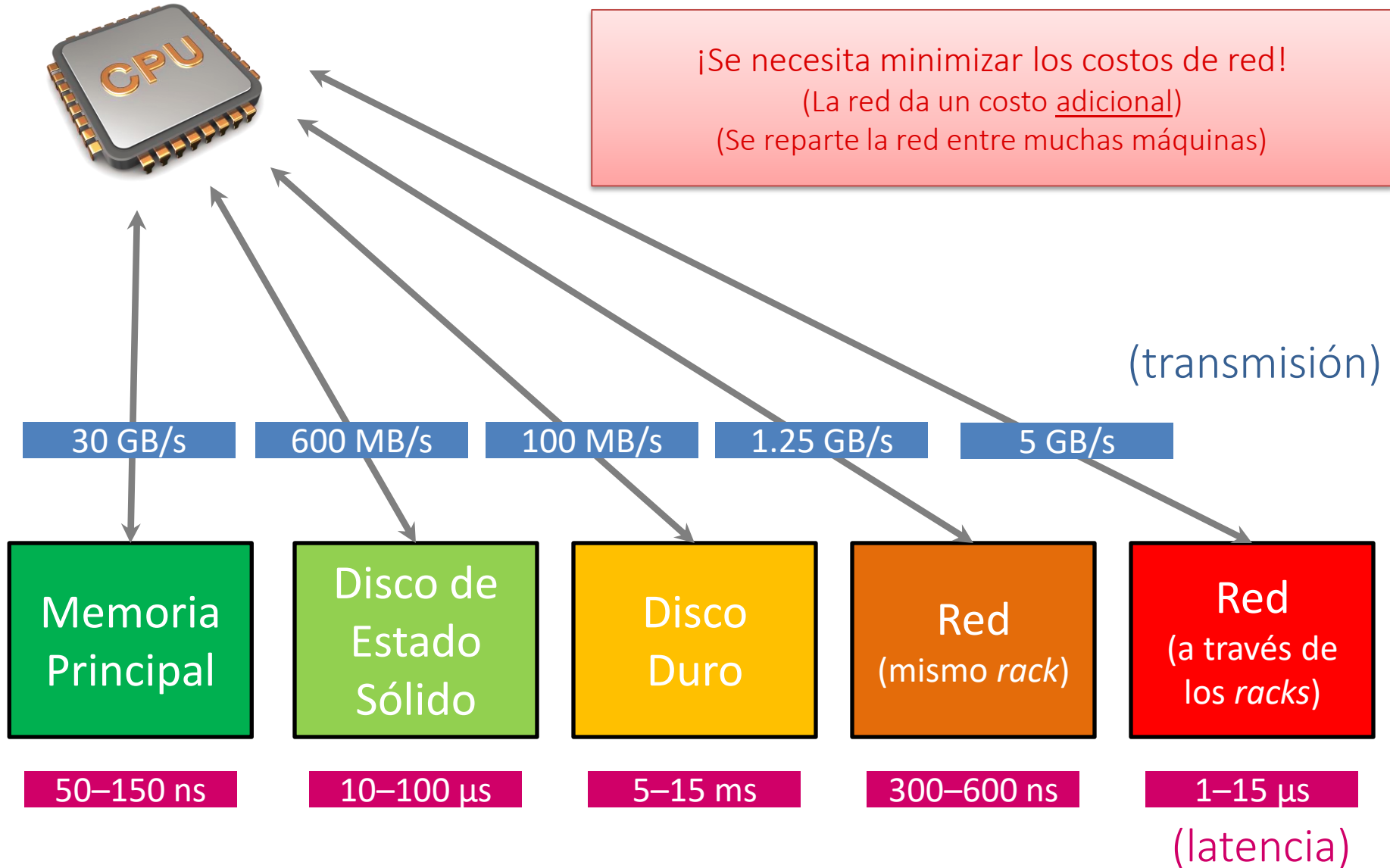


Computación distribuida

- Necesita más de una máquina
- Google ca. 2018:




Costos de transporte de los datos (estimaciones)




Colocación de los datos

- Hay que pensar cuidadosamente dónde poner qué datos

Tengo cuatro máquinas para  correr mi página web. Tengo 10 millones de usuarios.

Cada usuario tiene un perfil personal, fotos, amigos y juegos.

¿Cómo debería dividir los datos en las máquinas?

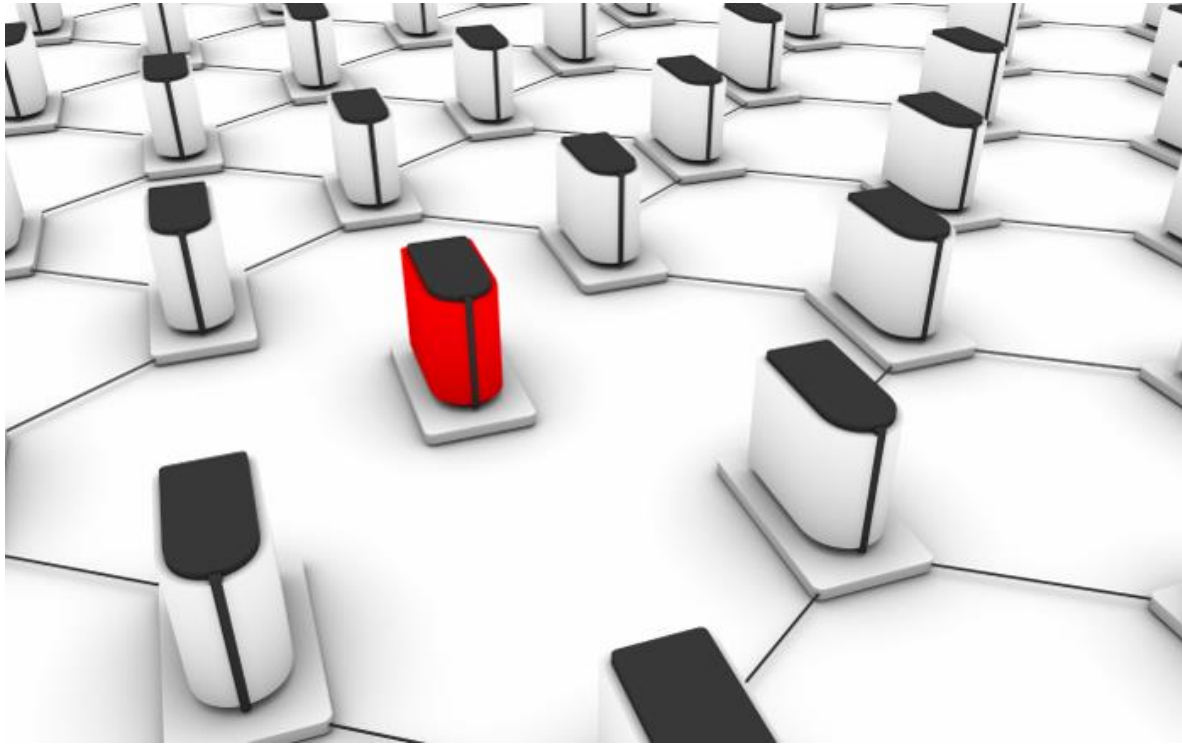
¡Depende de la aplicación! 

(Pero buenos principios de diseño aplican universalmente.)




Fallas de red/nodo

- Si tenemos miles de máquinas, ¡hay que pensar en las fallas!




Colocación de los datos

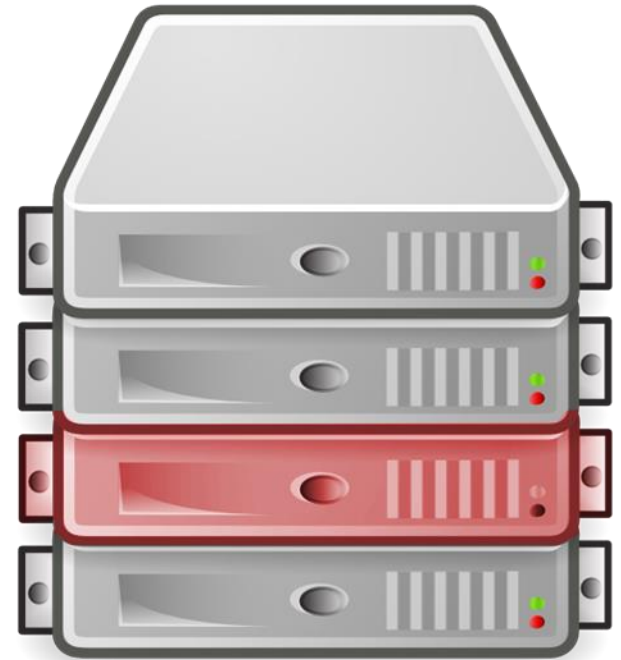
- Hay que pensar (**¡aún más!**) cuidadosamente dónde poner qué datos

Tengo cuatro máquinas para correr mi página web. Tengo 10 millones de usuarios. 

Cada usuario tiene un perfil personal, fotos, amigos y juegos.

¿Cómo debería dividir los datos en las máquinas?

¡Depende de la aplicación! 
(**de nuevo**)
(Pero buenos principios de diseño aplican universalmente.)

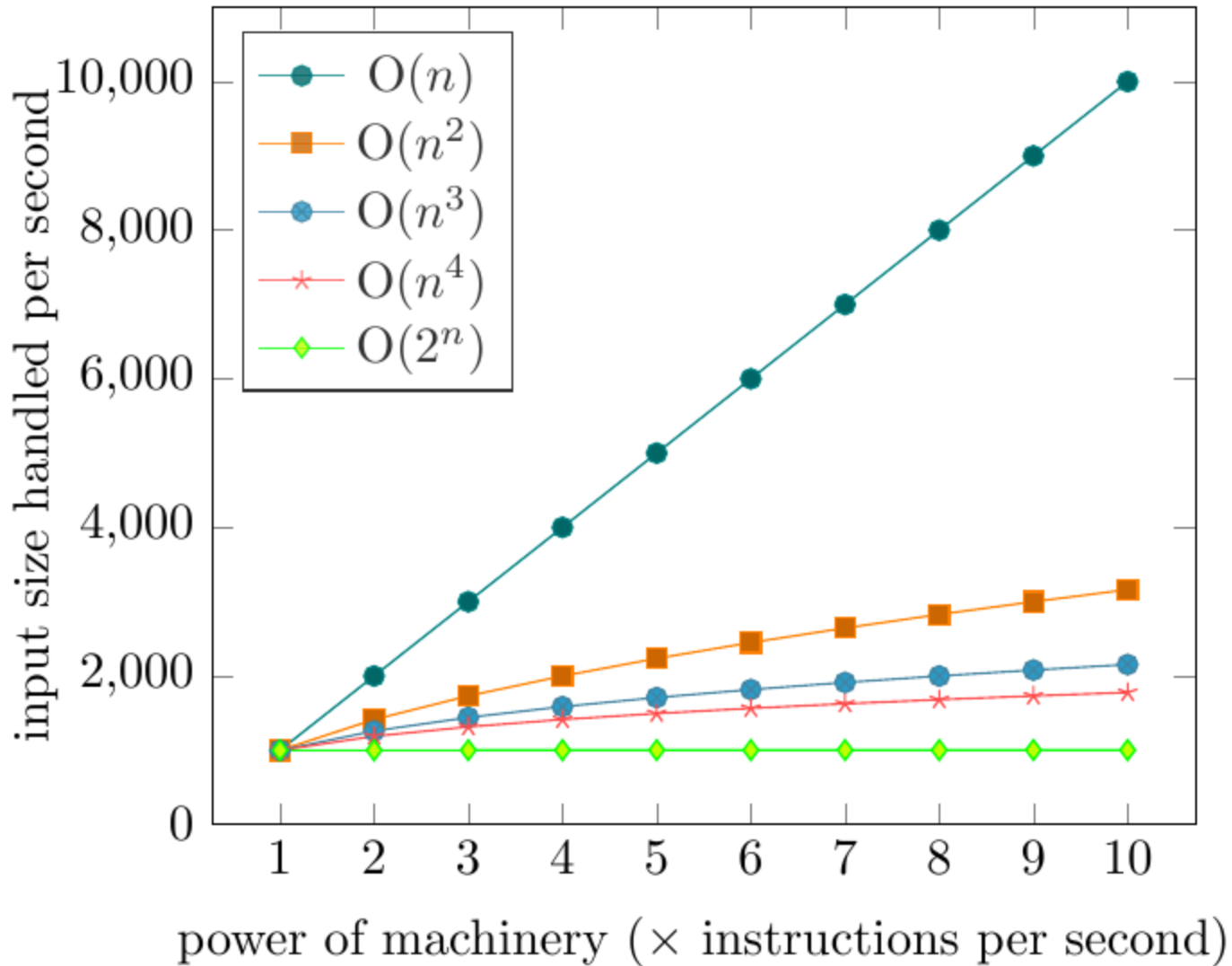


Computación distribuida humana



"COMPUTACIÓN DISTRIBUIDA"
LIMITACIONES Y DESAFÍOS ...

¡Distribución no es siempre aplicable!



Desarrollo distribuido es difícil

- Sistemas Distribuidos pueden ser complejos
- Con múltiples máquinas hay que ocuparse de:
 - Datos en diferentes localizaciones
 - Logs y mensajes en diferentes lugares
 - La eficiencia de la red
 - ¡Hay que manejar fallas!
 - ¡Hay que balancear carga!
- ¡Tareas toman mucho tiempo!
 - Bugs pueden no ser evidentes por horas
 - Muchos datos = muchos contra ejemplos

Frameworks/abstracciones pueden ayudar

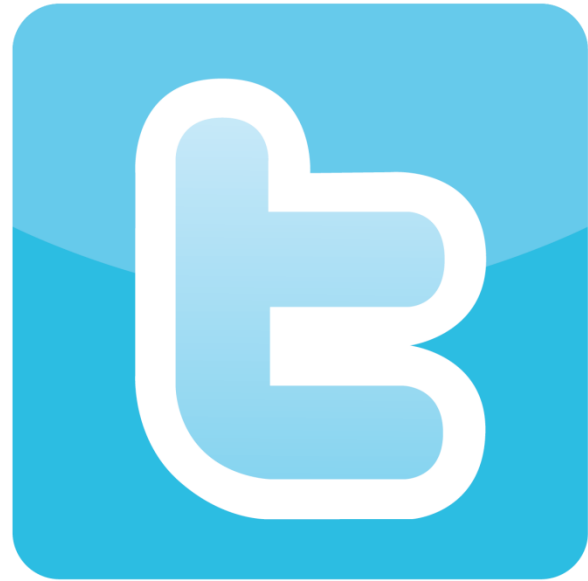
Para procesamiento distribuido
(p.ej.)



Frameworks/abstracciones pueden ayudar

Para almacenamiento distribuido
(p.ej.)





¿CÓMO FUNCIONA(BA) TWITTER?

Basado en las dispositivas del 2013, del Arquitecto Principal de Twitter: Raffi Krikorian



Big Data en Twitter

- 150 millones de usuarios activos
- 400 millones de tweets por día
 - 4.600 tweets por segundo
 - max: 143.199 tweets por segundo
- 300 mil consultas/s por timelines de usuarios
- 6 mil consultas/s por búsqueda personalizada

¿Qué debería ser la
prioridad al optimizar?



Twitter Timeline

The image shows a screenshot of the Twitter homepage. At the top, there is a navigation bar with icons for Home, Moments, Notifications (with a '2' badge), and Messages (with a '1' badge). A search bar labeled 'Search Twitter' and a 'Tweet' button are on the right. The main content area is titled 'What's happening?' and shows a list of tweets. The first tweet is from 'jon hendren @fart' about Doritos skywriting. The second is from 'demi adejuyigbe @electrolemon' about Beyoncé's performance. Below these are 'Top Tweets you might enjoy' from 'Fred Delicious' and 'jomny sun'. A promotional tweet from 'Twitter Small Biz' is at the bottom. On the left, there is a profile for 'leon @leyawn' and a 'United States Trends' sidebar. On the right, there is a 'Who to follow' sidebar and a footer with copyright information.

Home Moments Notifications Messages Search Twitter Tweet

What's happening?

Now viewing Top Tweets. [Switch to Most Recent Tweets?](#)

leon @leyawn
TWEETS 11.3K FOLLOWING 714 FOLLOWERS 43K

United States Trends · [Change](#)
#AskAlexa
Promoted by Amazon Echo
#TheSuperBowl
Howard Dean
#ILoveYouZayn
#SelenaGomezLive
#ItsChineseNewYear
#sundaymotivation
Free Beer
Alexa Says
Katie Holmes
Every Vote Counts

jon hendren @fart · 7h
the doritos corporation is skywriting over town for the super bowl. fools i already told you we are a #FunyunsFamily
17 retweets 117 likes

demi adejuyigbe @electrolemon · 1h
my mom just called me to say she doesn't think the halftime show was beyonce's best but she likes that coldplay guy. the polls are closed
36 retweets 383 likes

Here are some Top Tweets you might enjoy.
[Refresh](#) · [View all](#)

Fred Delicious @Fred_Delicious · 9h
Accidentally glued myself to the ceiling again
152 retweets 266 likes

jomny sun @jonnyusun · 1h
great super bowl evreybody see u next year
239 retweets 593 likes

Twitter Small Biz @TwitterSmallBiz · Jan 27
Start promoting your business on Twitter with a budget that works for you.

Who to follow · [Refresh](#) · [View all](#)

Daniel S. Johnson @linern...
[Follow](#)

Bill Maher @billmaher
[Follow](#)

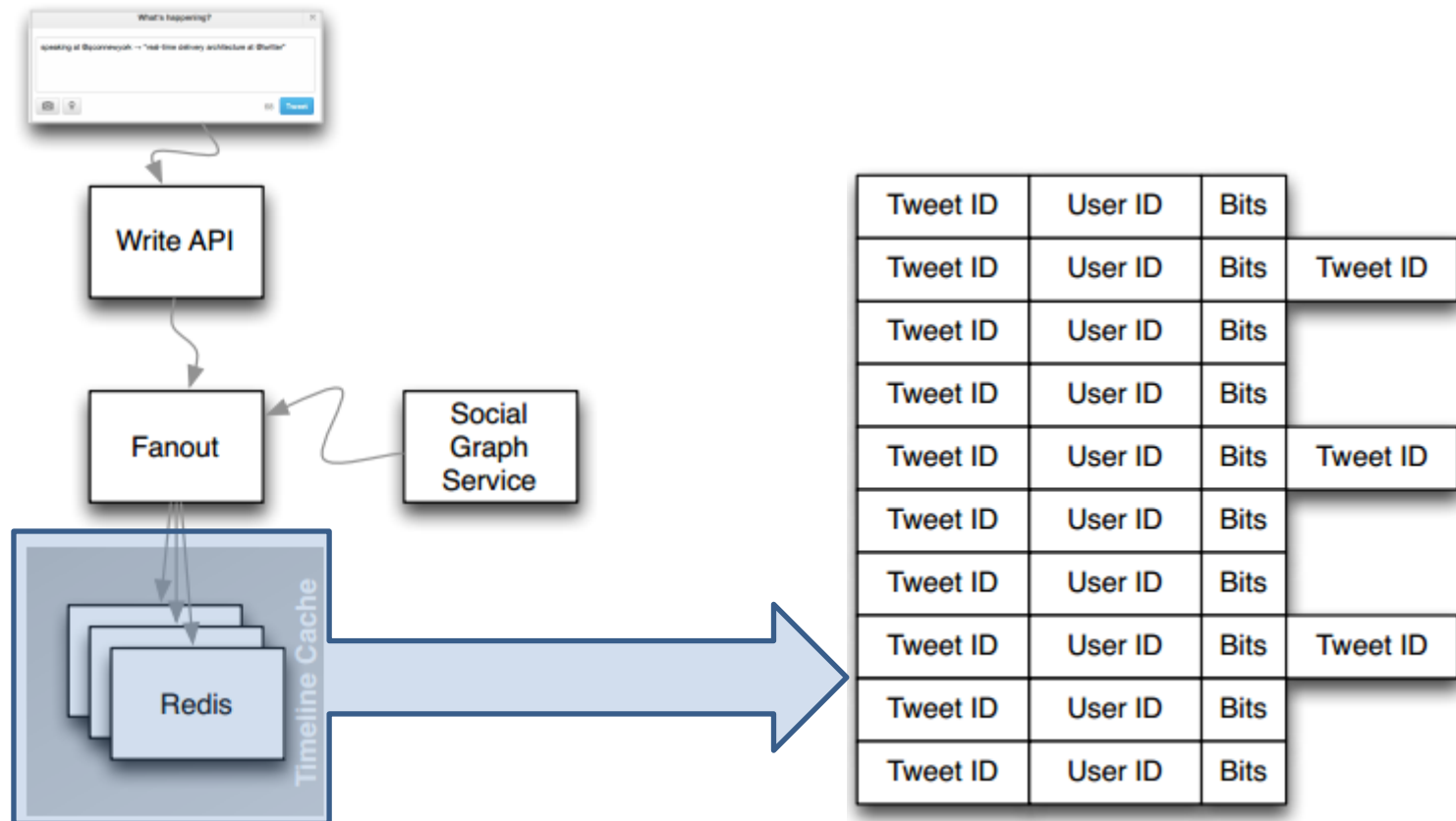
Edward Snowden @Sno...
[Follow](#)

[Find friends](#)

© 2016 Twitter [About](#) [Help](#) [Terms](#) [Privacy](#) [Cookies](#) [Ads info](#) [Brand](#) [Blog](#) [Status](#) [Apps](#) [Jobs](#) [Advertise](#) [Businesses](#) [Media](#) [Developers](#)

Implementando timelines: Escritura

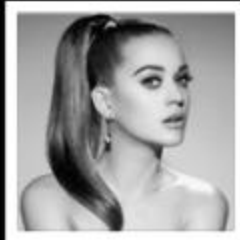
- 4.600 tweets por segundo (en promedio)



Nodos con alto grado



@ladygaga ✓
31 million followers



@katyperry ✓
28 million followers



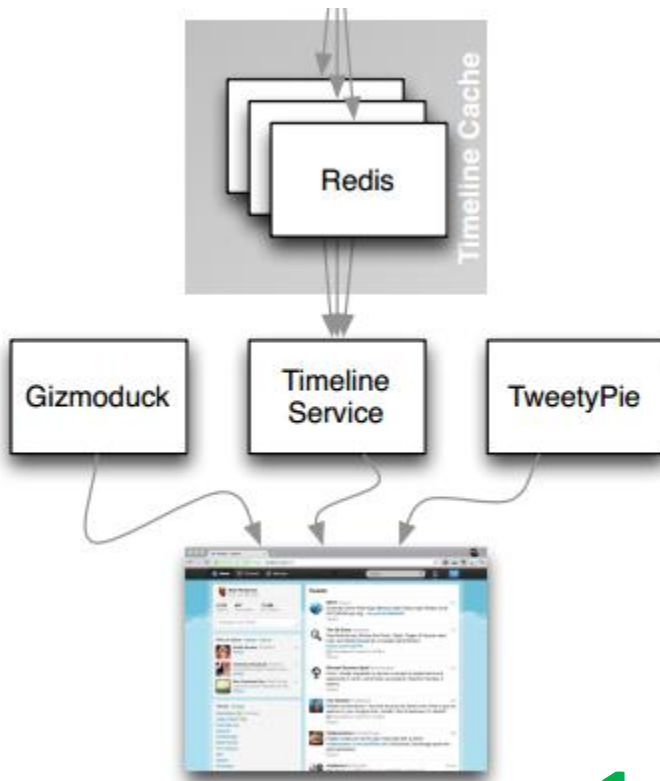
@justinbieber ✓
28 million followers



@barackobama ✓
23 million followers

Implementando timelines: Lectura

- 300.000 consultas por segundo (en promedio)



Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	

1ms @p50
4ms @p99

Búsqueda de texto



Home

About

See what's happening right now

terremoto

terremoto

#terremotos

terremoto chile

terremoto en irak



Terremoto Magazine @terremoto_mx



TERREMOTO en vivo @terremotolive



Pablo Ampuero @DocTerremoto



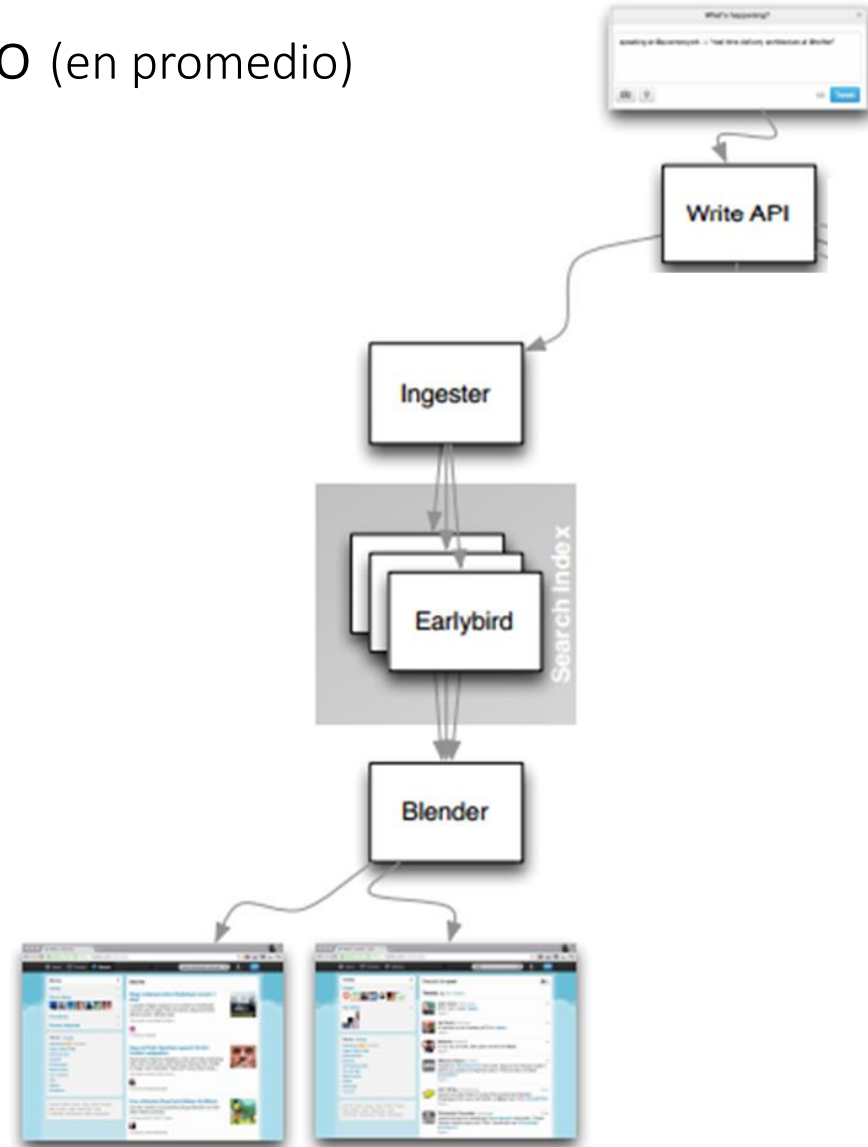
Terremoto Coffee @TerremotoCoffee



Terremoto Tequila @Terremoto

Implementando búsqueda de texto

- 6.000 consultas por segundo (en promedio)



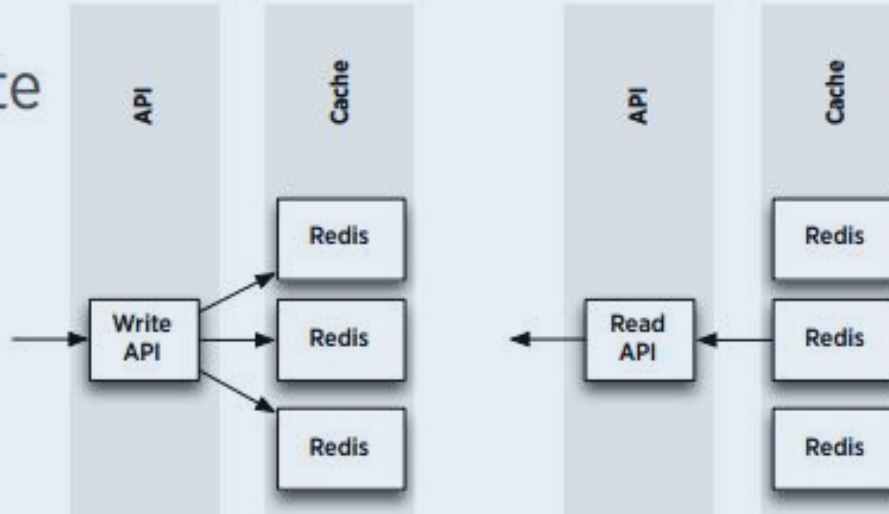
Timeline vs. Búsqueda

4.600 peticiones/s

300.000 peticiones/s

→ $O(n)$ write

→ $O(1)$ read

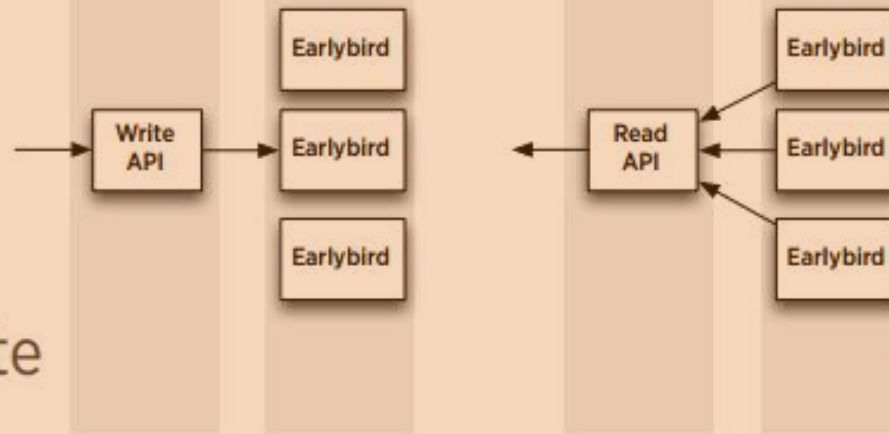


→ $O(1)$ write

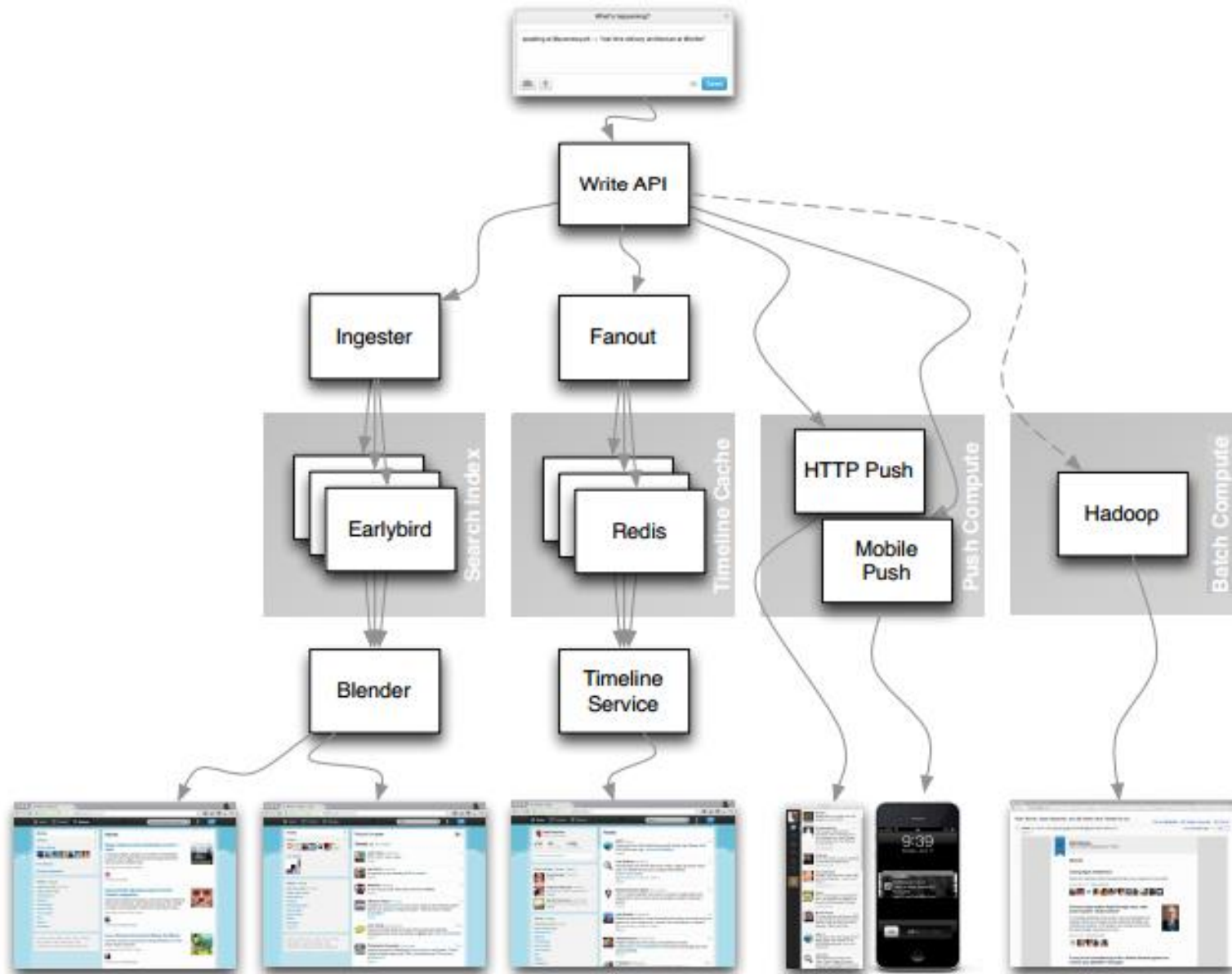
→ $O(n)$ read

4.600 peticiones/s

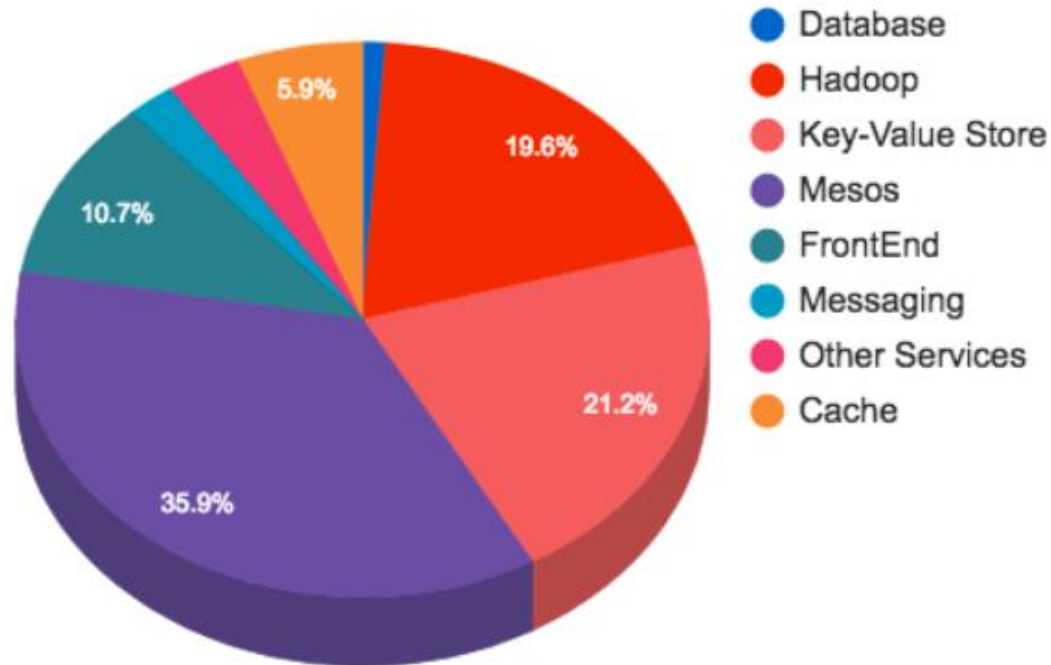
6.000 peticiones/s



Twitter: Arquitectura Completa



Twitter en ~~2021~~ 2017?



https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale.html

"GESTIÓN DE DATOS"

ACERCA DEL CURSO

¿Qué es el curso/No es?

- Intensivo en datos | No intensivo en computo
- Tareas distribuidas | No crear redes
- Hardware no especializado | No supercomputadores
- Métodos generales | No algoritmos específicos
- Métodos prácticos | Con poco teoría

Estructura del curso

- Primera mitad de la sesión: Clase
 - Segunda mitad de la sesión: Práctica
-
1. Introducción Java: Conteo local de palabras
 2. GFS & MapReduce HDFS & Hadoop: Conteo de palabras
 3. Pig Pig: Contando IMDb co-actores
 4. Spark Spark: Analizando series de televisión
 5. Crawling & Índices Invertidos D. Elasticsearch: Búsqueda sobre Wikipedia
 6. PageRank & Grafos Giraph: PageRank sobre Wikipedia
 7. NoSQL I Cassandra: Consultas e indexación
 8. NoSQL II MongoDB: Consultas sobre series de televisión

Nota final: 100% prácticas (8 en total, 12,5% cada una)

Descargar datos



<http://aidanhogan.com/teaching/data/wiki/es/es-wiki-abstracts.txt.gz>



Preguntas?

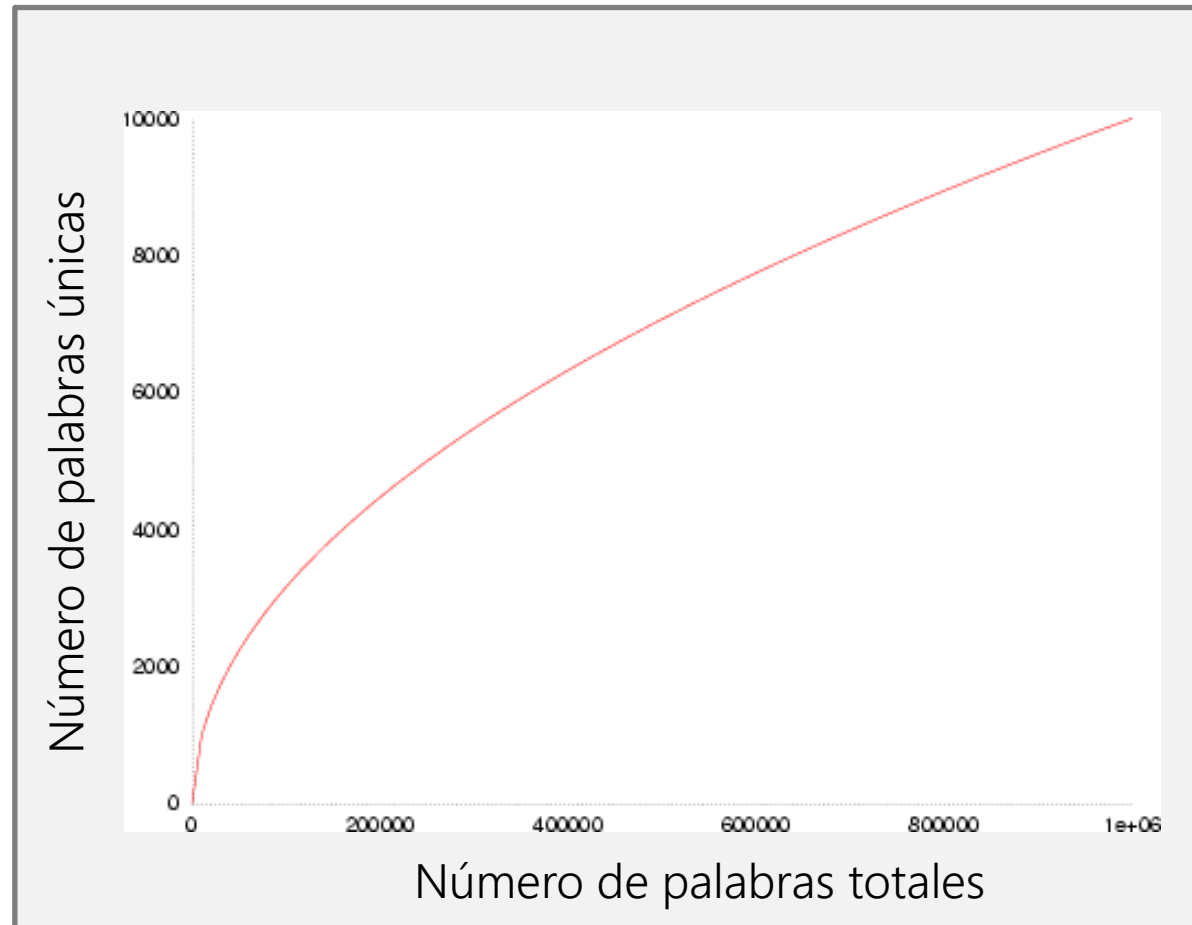
LAB

¿Por qué funcionó?

Procesamos muchos datos. ¿Por qué funcionó en memoria?

No hay tantas palabras únicas ...

- Ley de Heap:

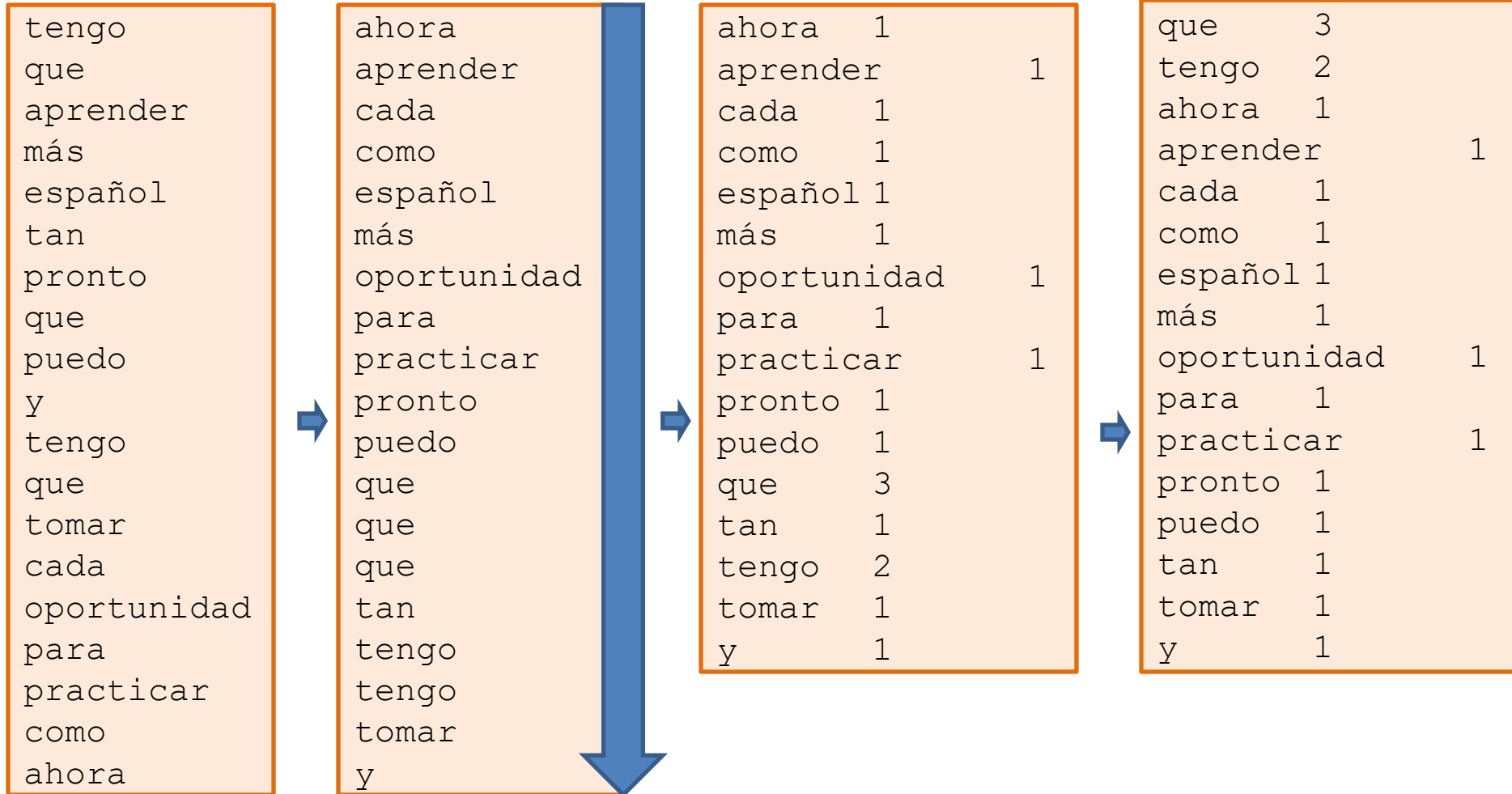


¿Y si no funciona?

Y ¿si no funciona en memoria?



Ordenar los datos



¿Cómo podemos usar el disco para ordenar los datos?

Ordenamiento Externo 1: Lotes

- Ordenar los datos en lotes

Entrada (disco)
(Tamaño: n)

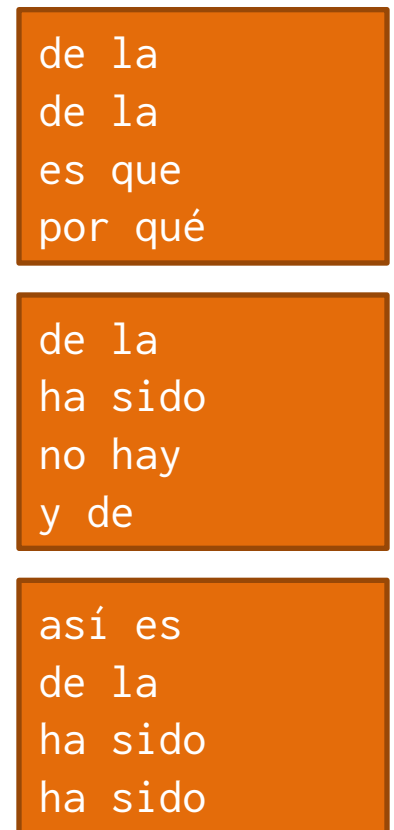
```
es que
de la
por qué
de la
ha sido
no hay
de la
y de
ha sido
de la
así es
ha sido
```

Ordenar (en memoria)
(Lote: b)



```
así es
de la
ha sido
ha sido
```

Salida intermedia (disco)
($\lceil n/b \rceil$ lotes)



```
de la
de la
es que
por qué

de la
ha sido
no hay
y de

así es
de la
ha sido
ha sido
```

Ordenamiento Externo 2: Combinar (*Merge*)

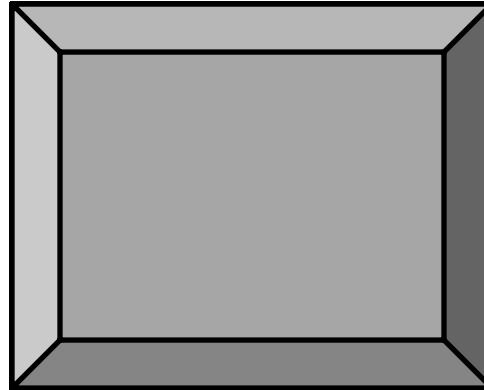
Salida intermedia (disco)
($\lceil n/b \rceil$ lotes)

[1]
de la
de la
es que
por qué

[2]
de la
ha sido
no hay
y de

[3]
así es
de la
ha sido
ha sido

Ordenar (en memoria)
(Espacio: $\lceil n/b \rceil$)



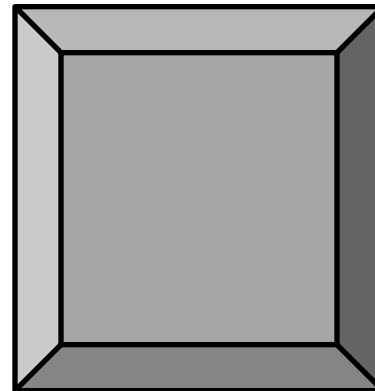
Salida final (disco)
(Tamaño: n)

así es
de la
de la
de la
de la
es que
ha sido
ha sido
ha sido
no hay
por qué
y de

Contar

así es
de la
de la
de la
de la
es que
ha sido
ha sido
ha sido
no hay
por qué
y de

Podríamos ordenar de nuevo, esta vez por frecuencia, usando el mismo método



así es,	1
de la,	4
es que,	1
ha sido,	3
no hay,	1
por qué,	1
y de,	1



¿Escalar más?

