CC5212-1

Procesamiento Masivo de Datos
Otoño 2018

Lecture 1: Introduction

Aidan Hogan
aidhog@gmail.com

The value of data

Soho, London, 1854

A COURT FOR KING CHOLERA.

Cholera: What we know now …

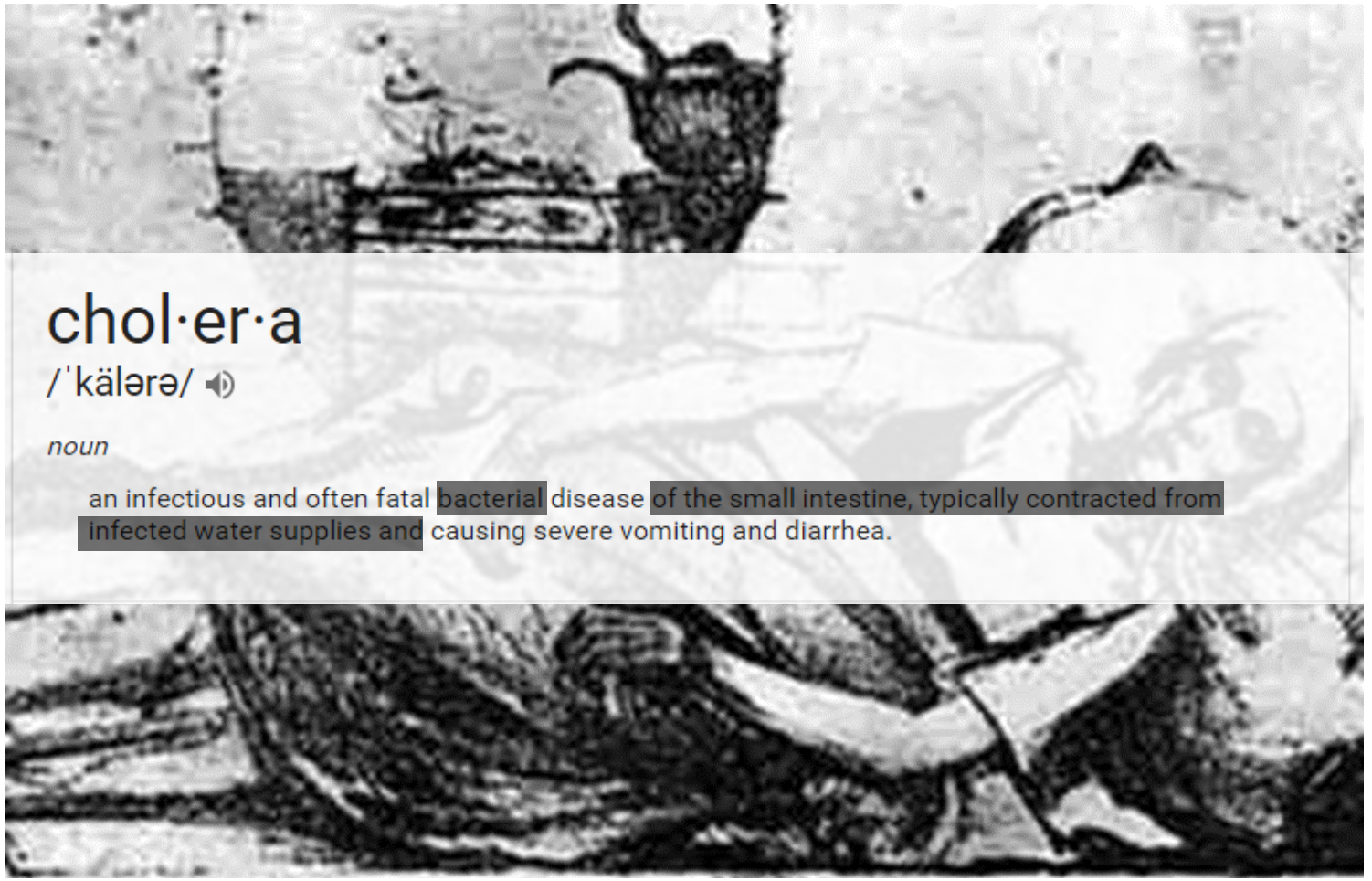
chol·er·a

/ˈkälərə/

noun

an infectious and often fatal bacterial disease of the small intestine, typically contracted from infected water supplies and causing severe vomiting and diarrhea.

Cholera: What we knew in 1854



chol·er·a

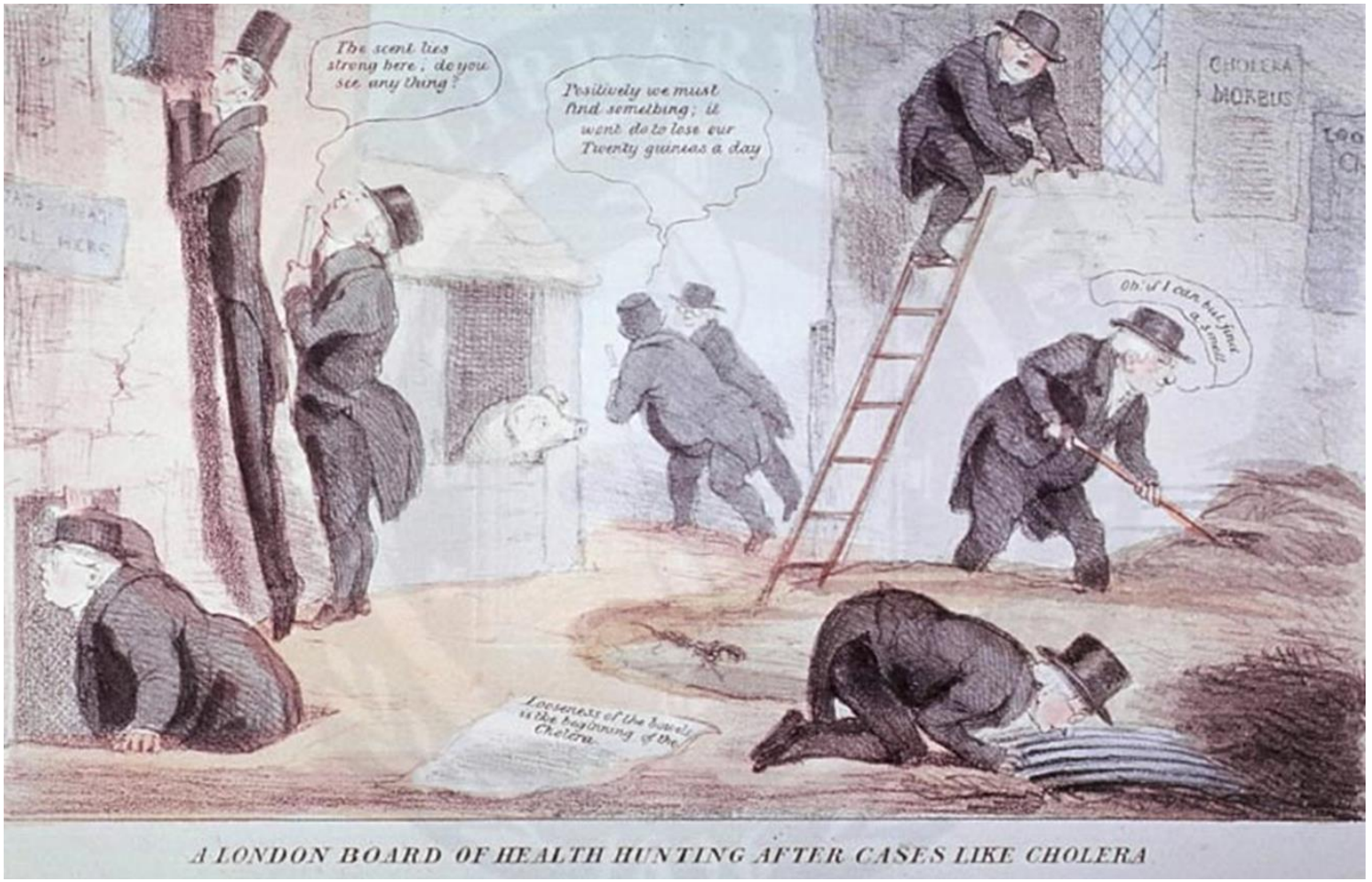
/ˈkälərə/

noun

an infectious and often fatal bacterial disease of the small intestine, typically contracted from infected water supplies and causing severe vomiting and diarrhea.

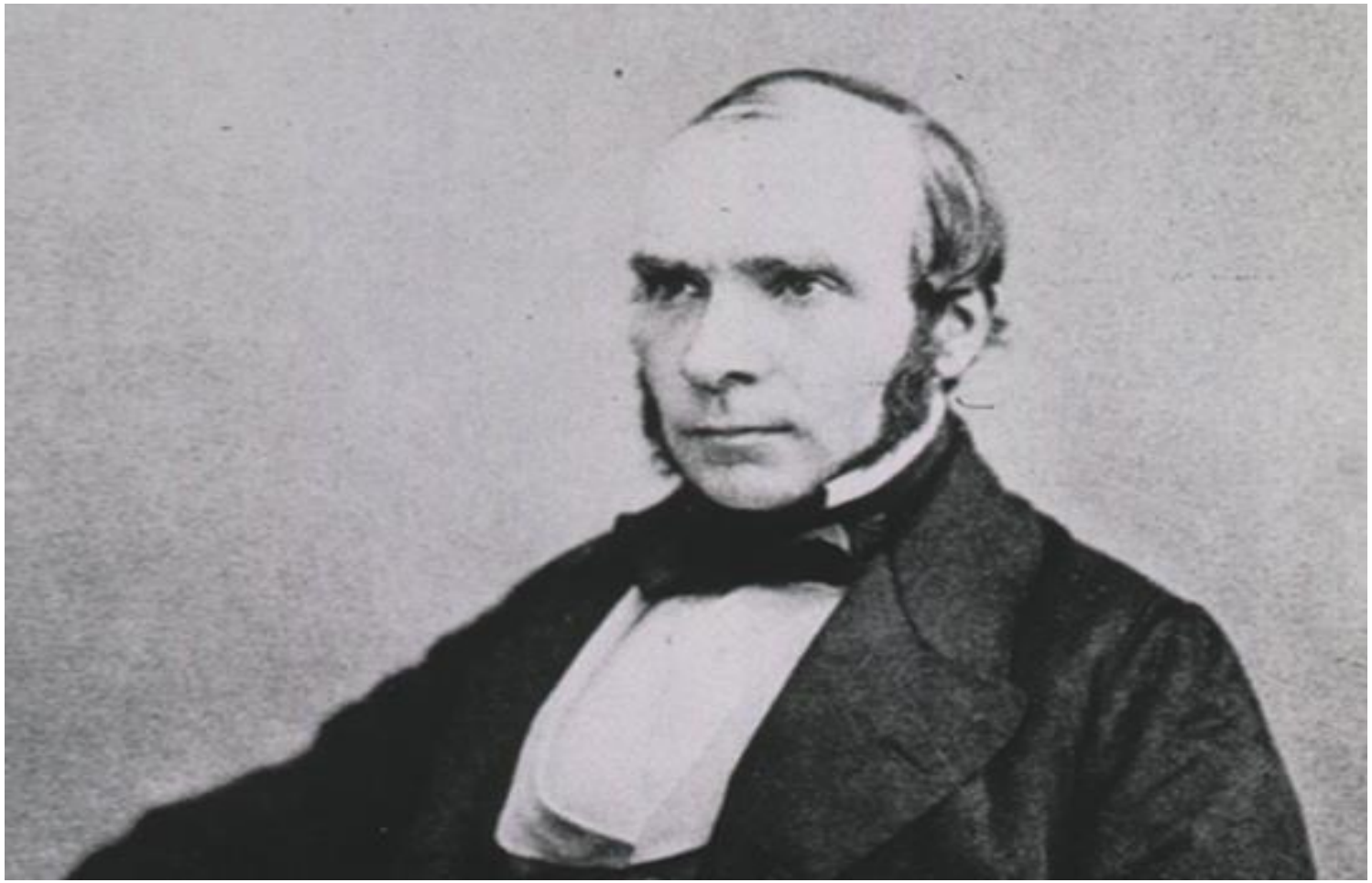
1854: Galen's miasma theory of cholera

1854: The hunt for the invisible cholera



A LONDON BOARD OF HEALTH HUNTING AFTER CASES LIKE CHOLERA

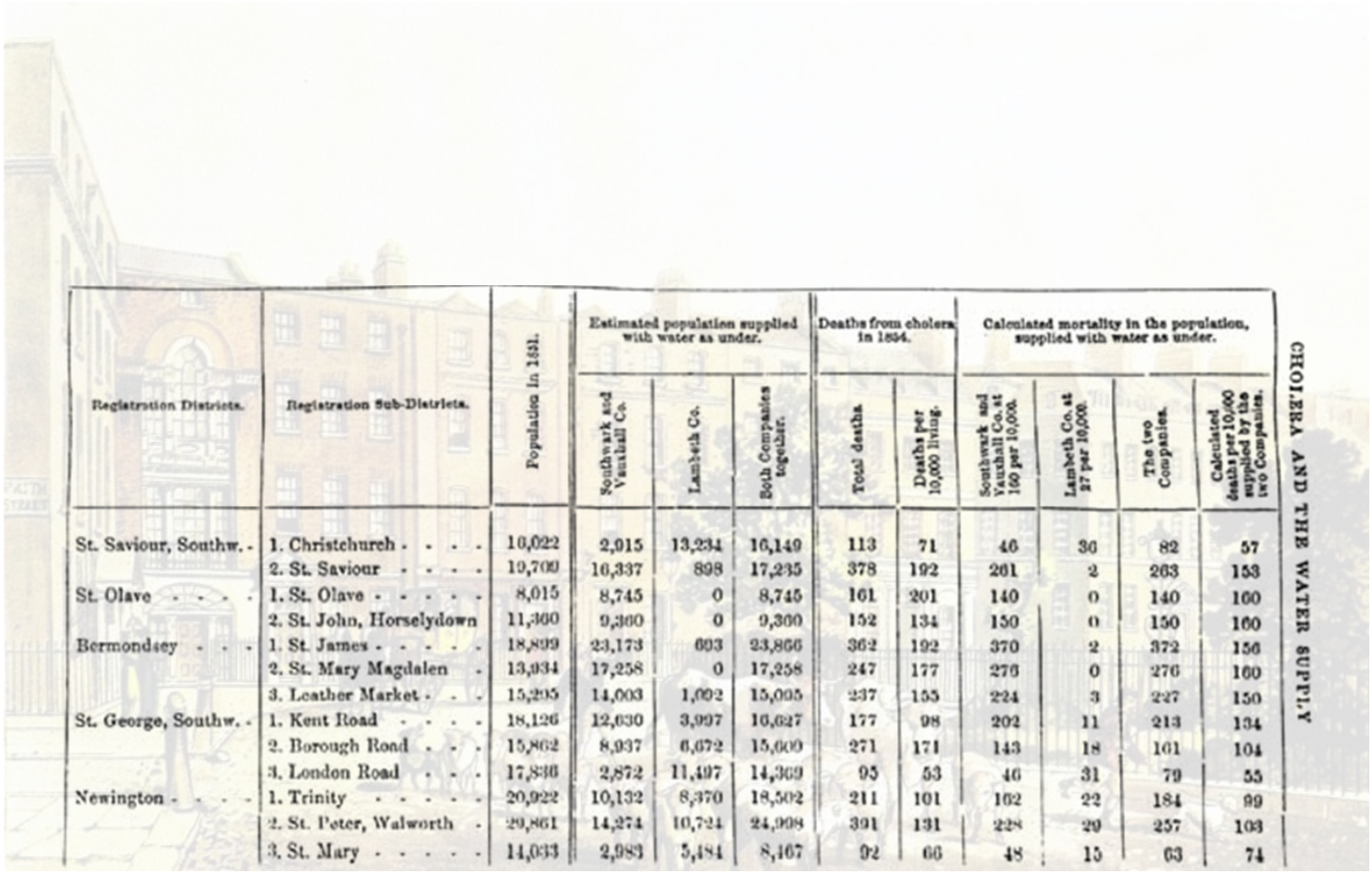
John Snow: 1813–1858



John Snow: 1813–1858

The Survey of Soho

The Survey of Soho

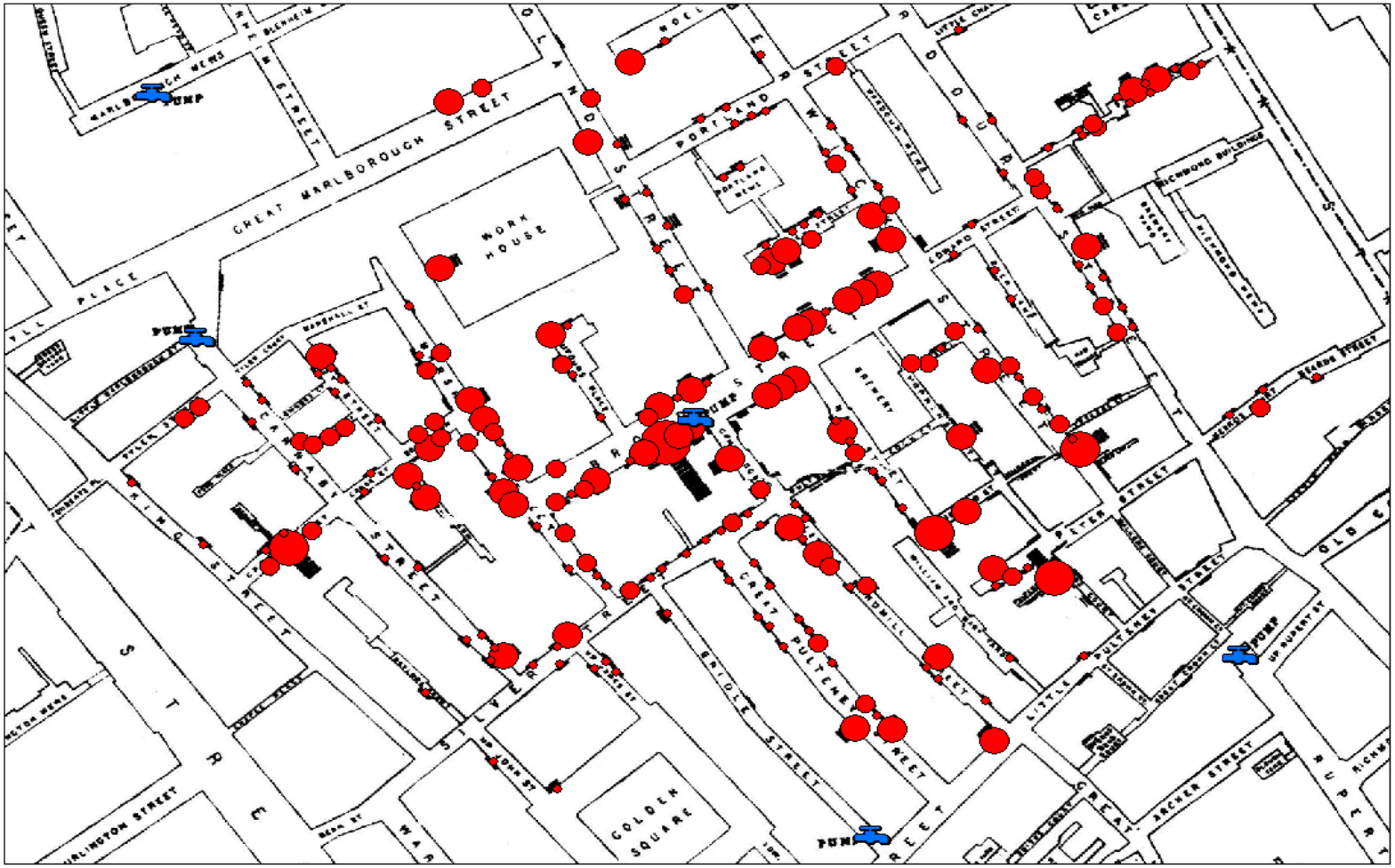


Registration Districts.	Registration Sub-Districts.	Population in 1851.	Estimated population supplied with water as under.			Deaths from cholera in 1854.		Calculated mortality in the population, supplied with water as under.			
			Southwark and Vauxhall Co.	Lambeth Co.	Both Companies together.	Total deaths	Deaths per 10,000 living.	Southwark and Vauxhall Co. at 150 per 10,000.	Lambeth Co. at 37 per 10,000.	The two Companies.	Calculated deaths per 10,000 supplied by the two Companies.
St. Saviour, Southw. -	1. Christchurch - - - -	16,022	2,915	13,234	16,149	113	71	46	36	82	57
	2. St. Saviour - - - -	19,709	16,337	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	9,360	0	9,360	152	134	150	0	150	160
Bermondsey - - -	1. St. James - - - - -	18,899	23,173	693	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	162	22	184	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

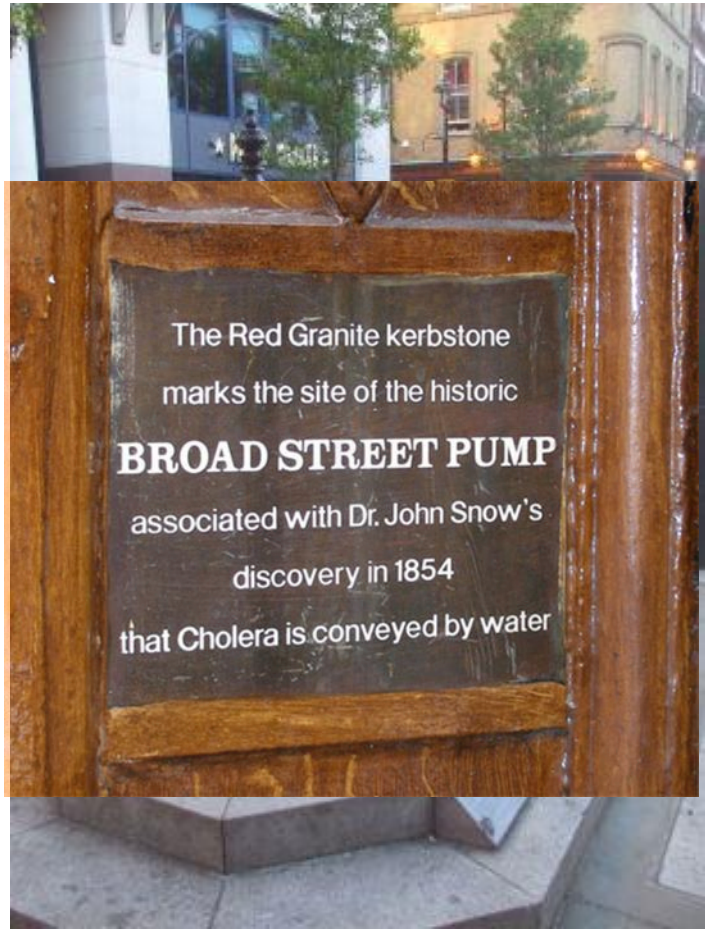
CHOLERA AND THE WATER SUPPLY.

What the data showed …

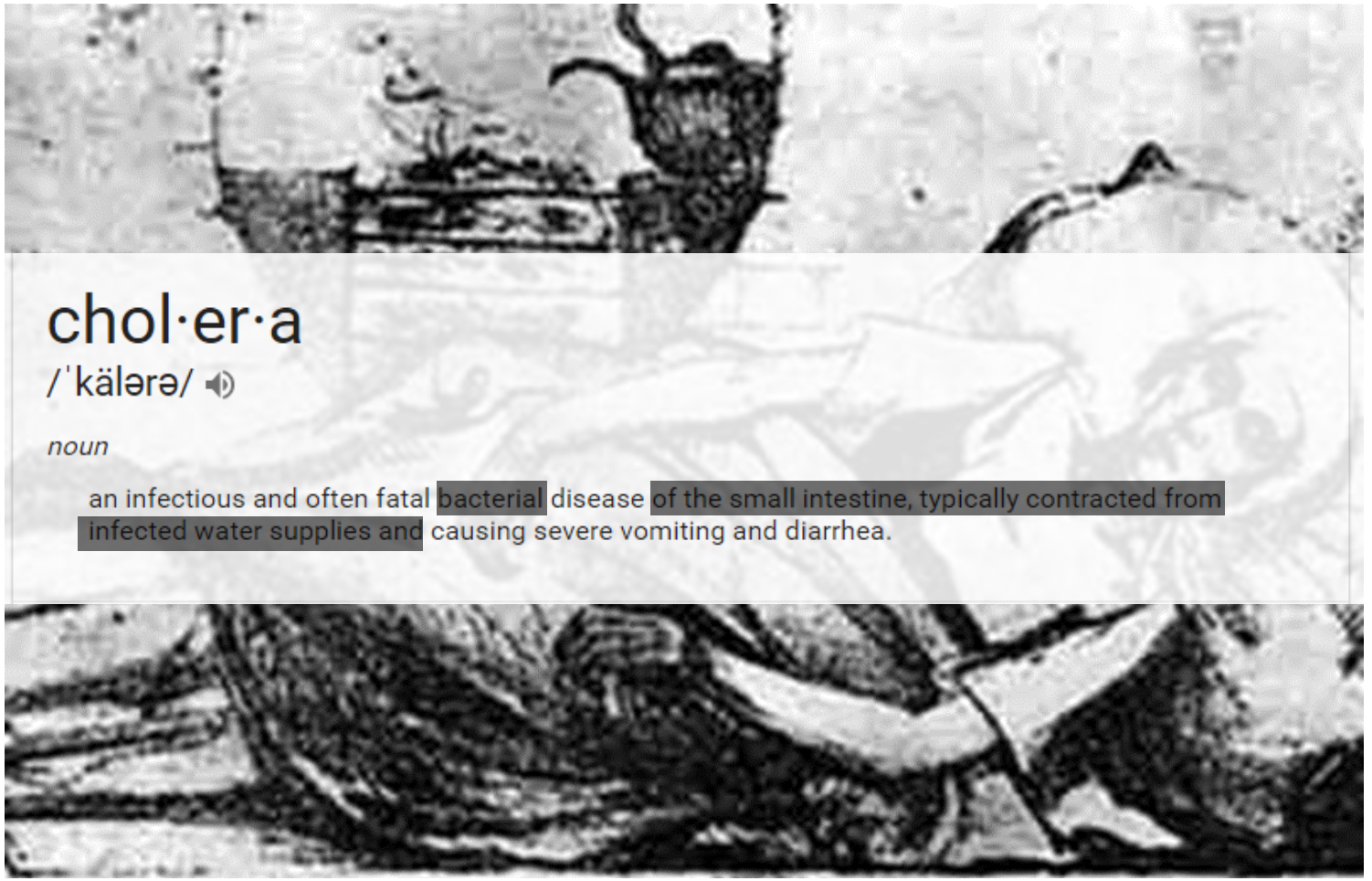
What the data showed …



616 deaths, 8 days later …



Cholera: What we knew in 1855



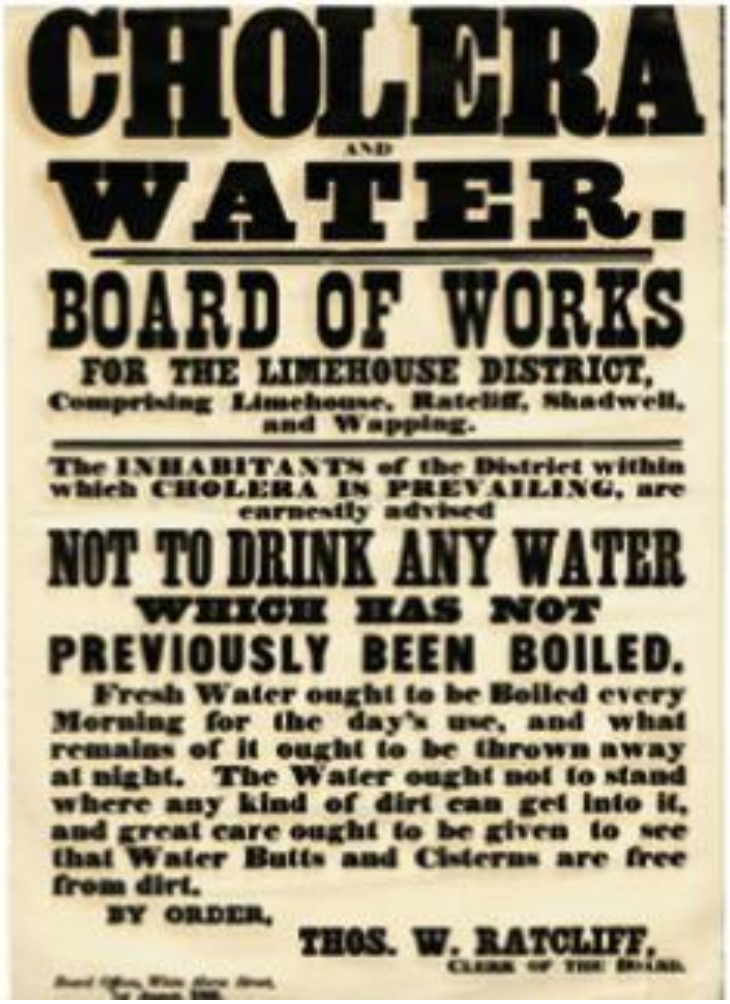
chol·er·a

/ˈkälərə/

noun

an infectious and often fatal bacterial disease of the small intestine, typically contracted from infected water supplies and causing severe vomiting and diarrhea.

Cholera boil notice ca. 1866



CHOLERA
AND
WATER.

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

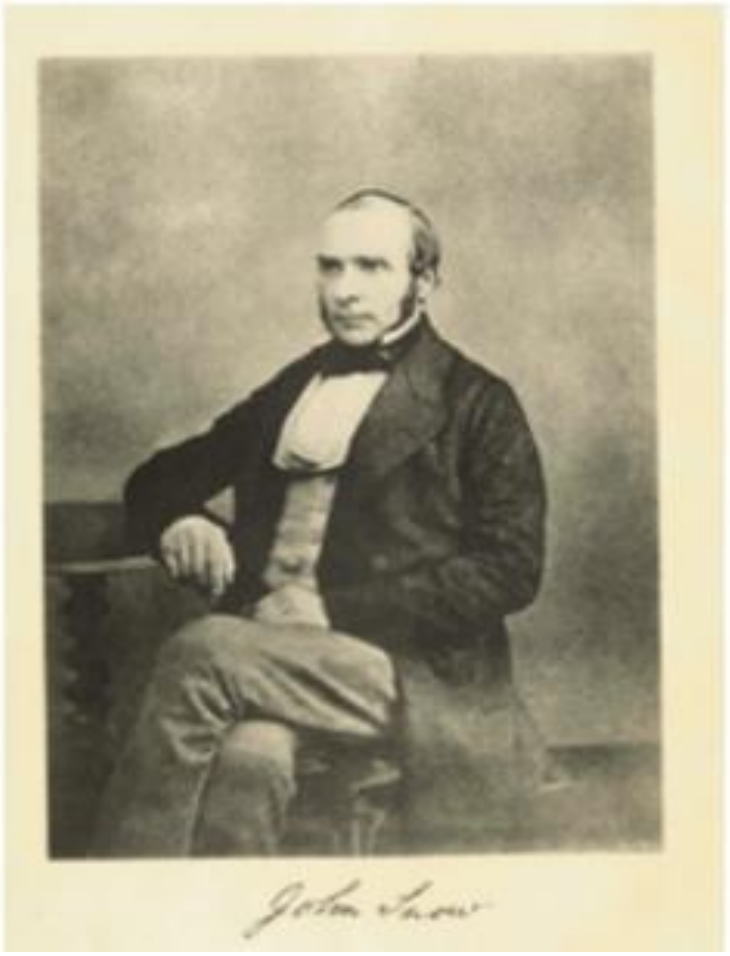
The INHABITANTS of the District within which CHOLERA IS PREVAILING, are earnestly advised

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

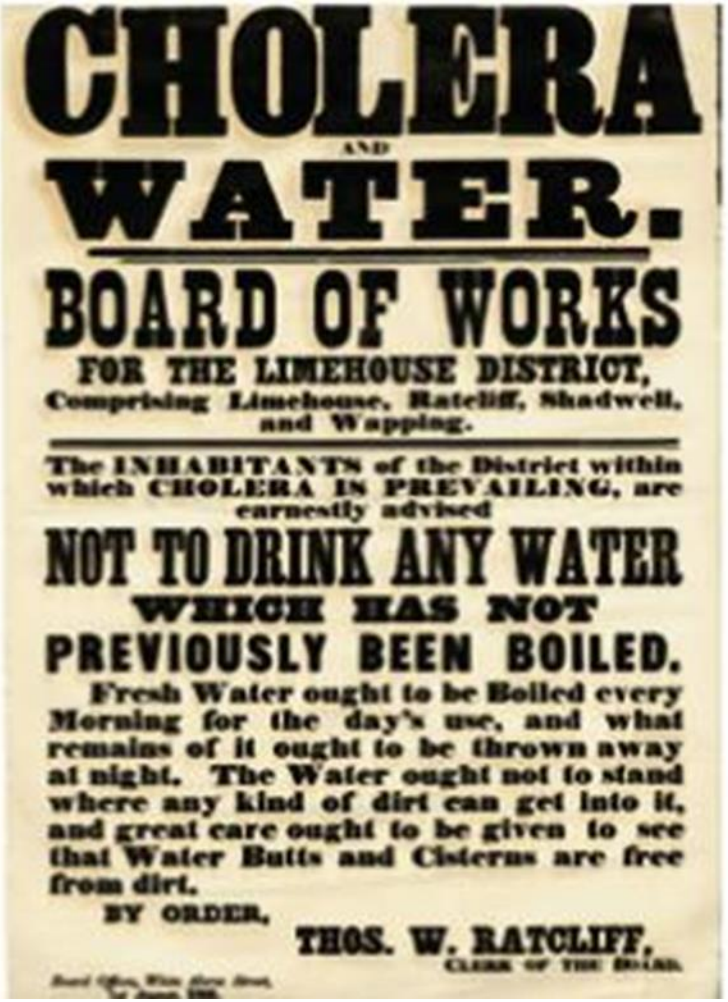
Fresh Water ought to be Boiled every Morning for the day's use, and what remains of it ought to be thrown away at night. The Water ought not to stand where any kind of dirt can get into it, and great care ought to be given to see that Water Butts and Cisterns are free from dirt.

BY ORDER,

THOS. W. RATCLIFF,
CLERK OF THE BOARD.



Cholera boil notice ca. 1866



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

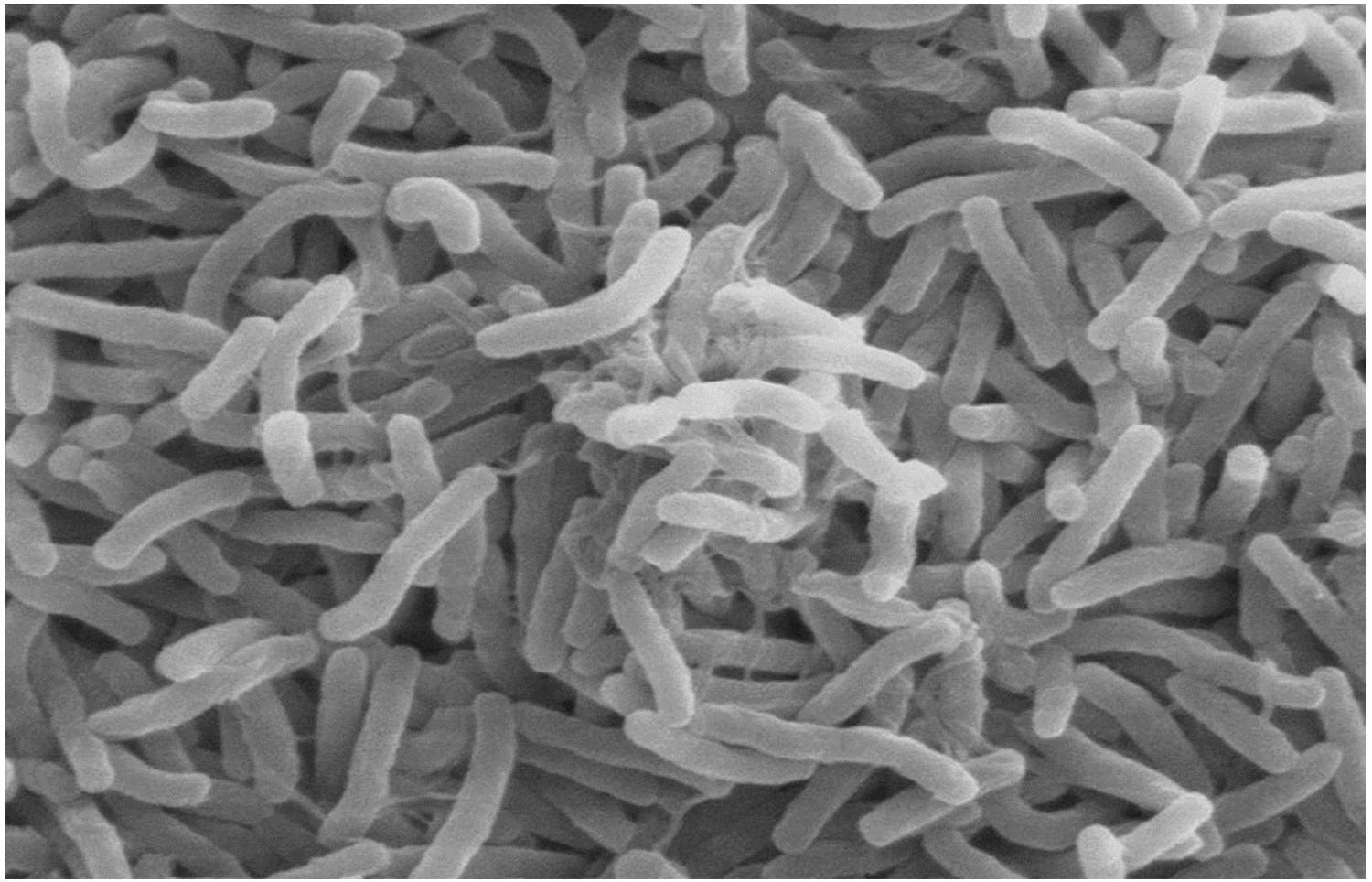
NOT TO DRINK ANY WATER
WHICH HAS NOT
PREVIOUSLY BEEN BOILED.

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.

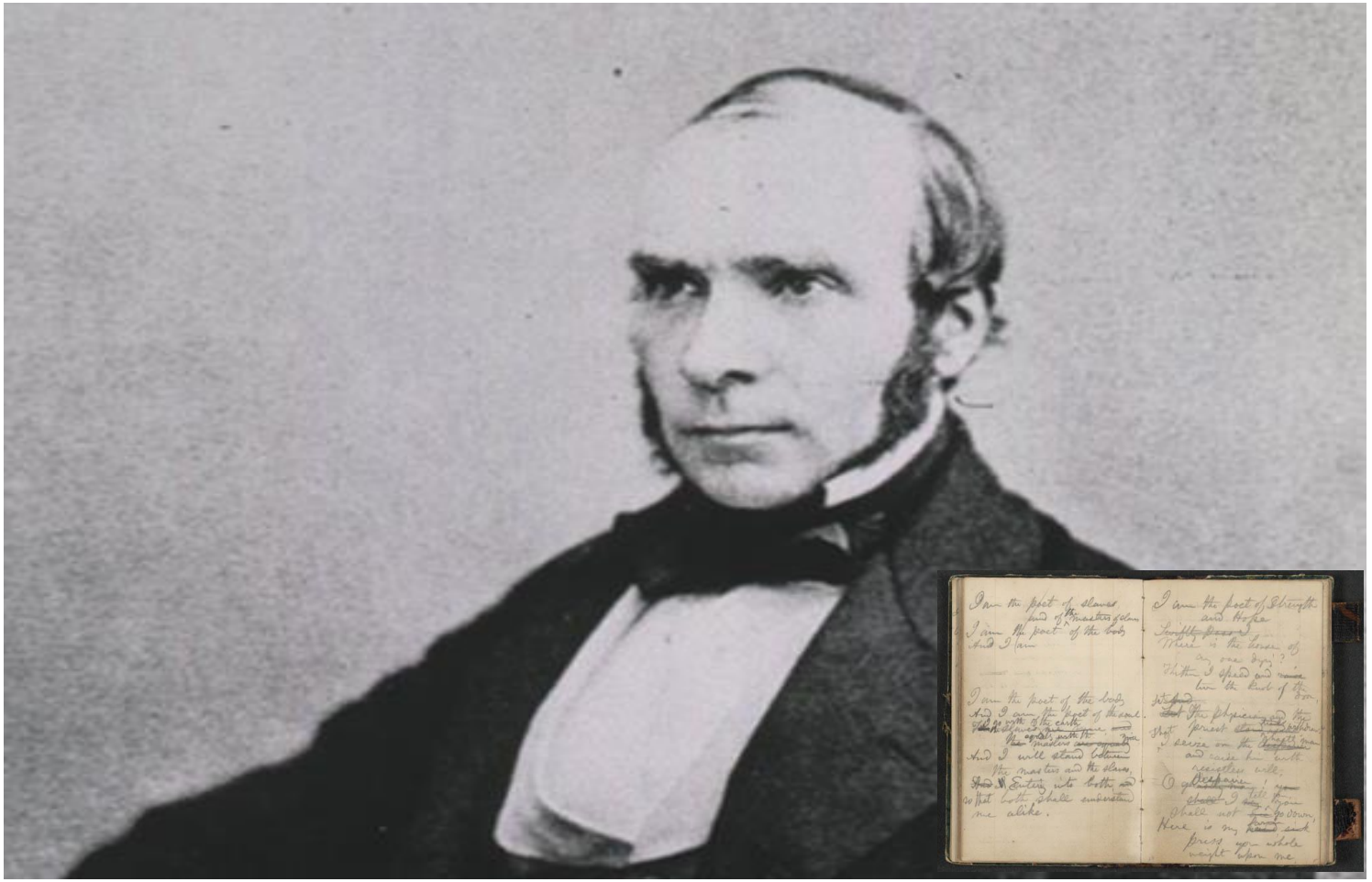
BY ORDER,

THOS. W. RATCLIFF,
CLERK OF THE BOARD.

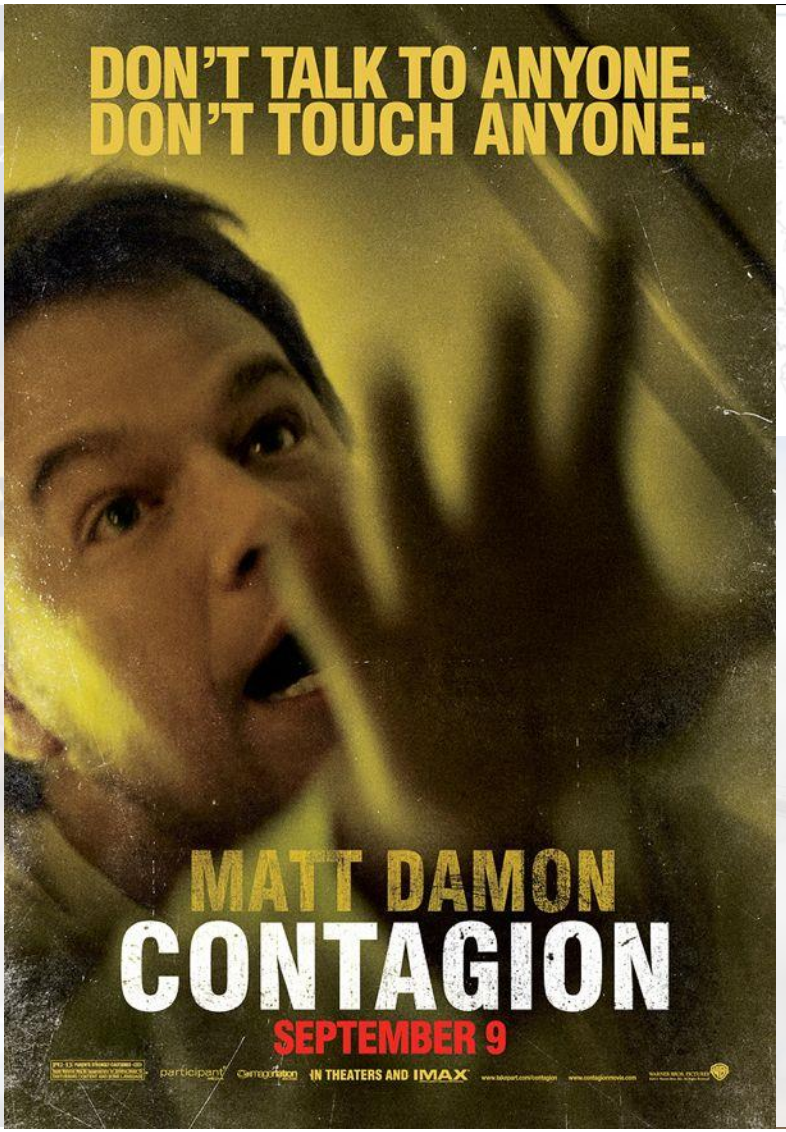
Thirty years before discovery of *V. cholerae*

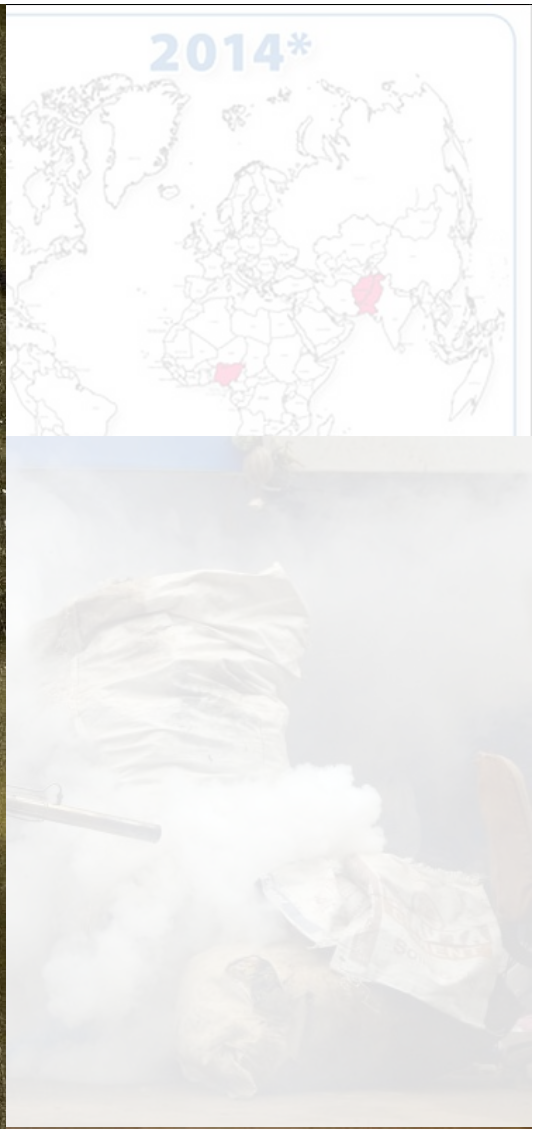


John Snow: Father of Epidemiology



Epidemiology's Success Stories



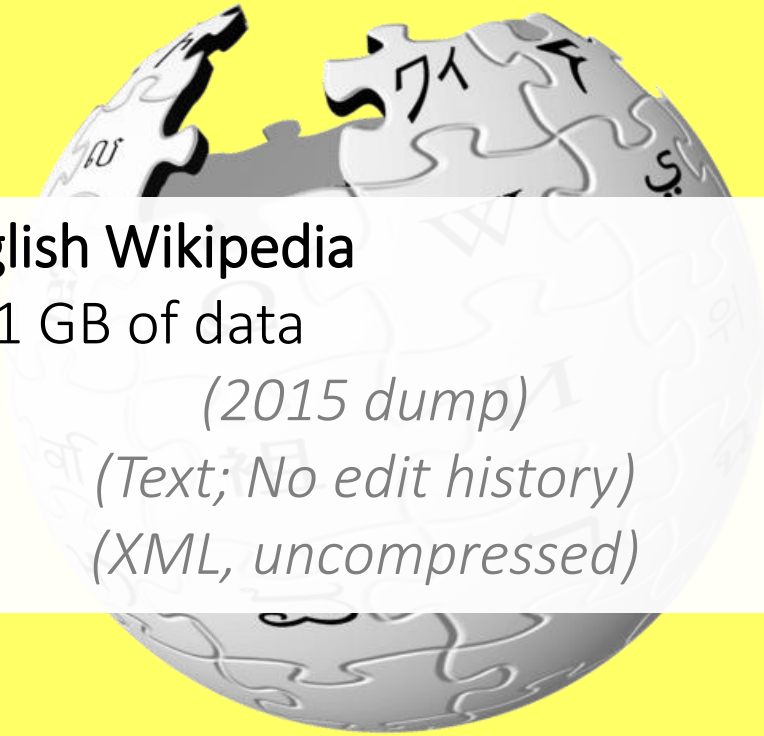


Value of data: Not just epidemiology

(Paper) Notebooks no longer good enough

The Growth of Data

“Big Data”



English Wikipedia

≈ 51 GB of data

(2015 dump)

(Text; No edit history)

(XML, uncompressed)

WIKIPEDIA

The Free Encyclopedia

1 Wiki = 1 Wikipedia

“Big Data”

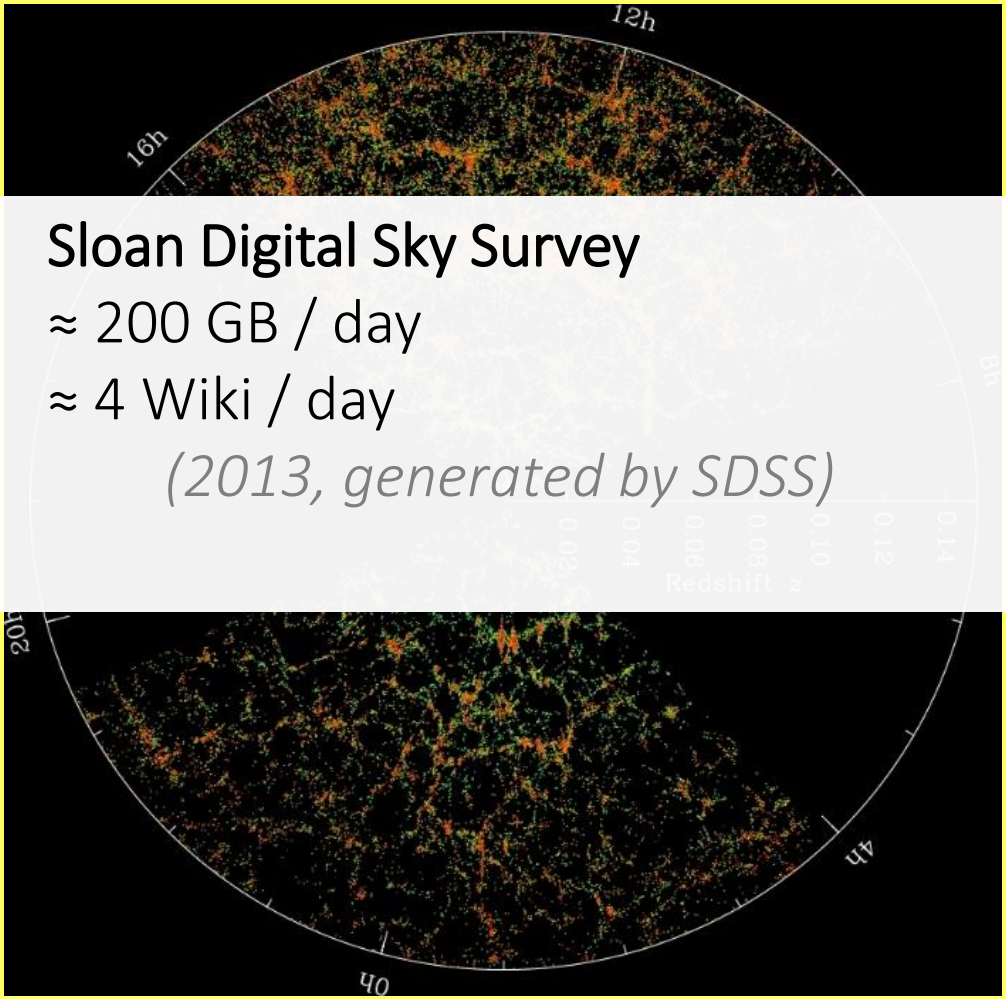
Wikimedia Commons

≈ 24 TB of data

≈ 470.6 Wiki

(2014 dump)

“Big Data”



Sloan Digital Sky Survey

≈ 200 GB / day

≈ 4 Wiki / day

(2013, generated by SDSS)

“Big Data”

Twitter

≈ 8 TB / day

≈ 157 Wiki / day

(2013, generated)

“Big Data”

Large Hadron Collider

≈ 68 TB / day

≈ 1,370 Wiki / day

(2012, collision data generated)

“Big Data”

Facebook

≈ 600 TB / day

≈ 11,764 Wiki / day

(2014, incoming Hive data)

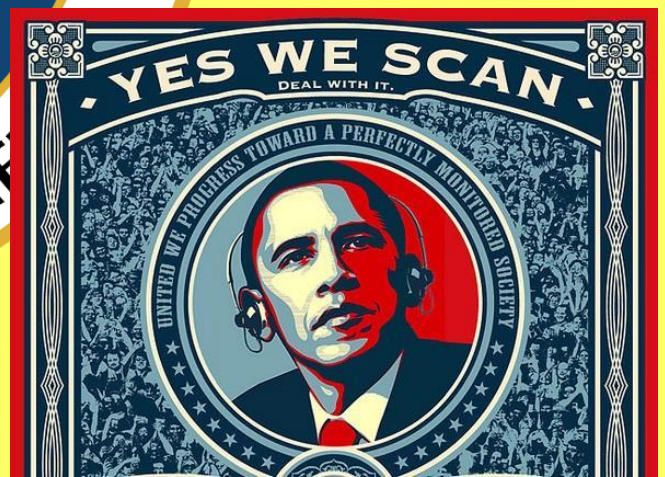
“Big Data”

NSA Surveillance

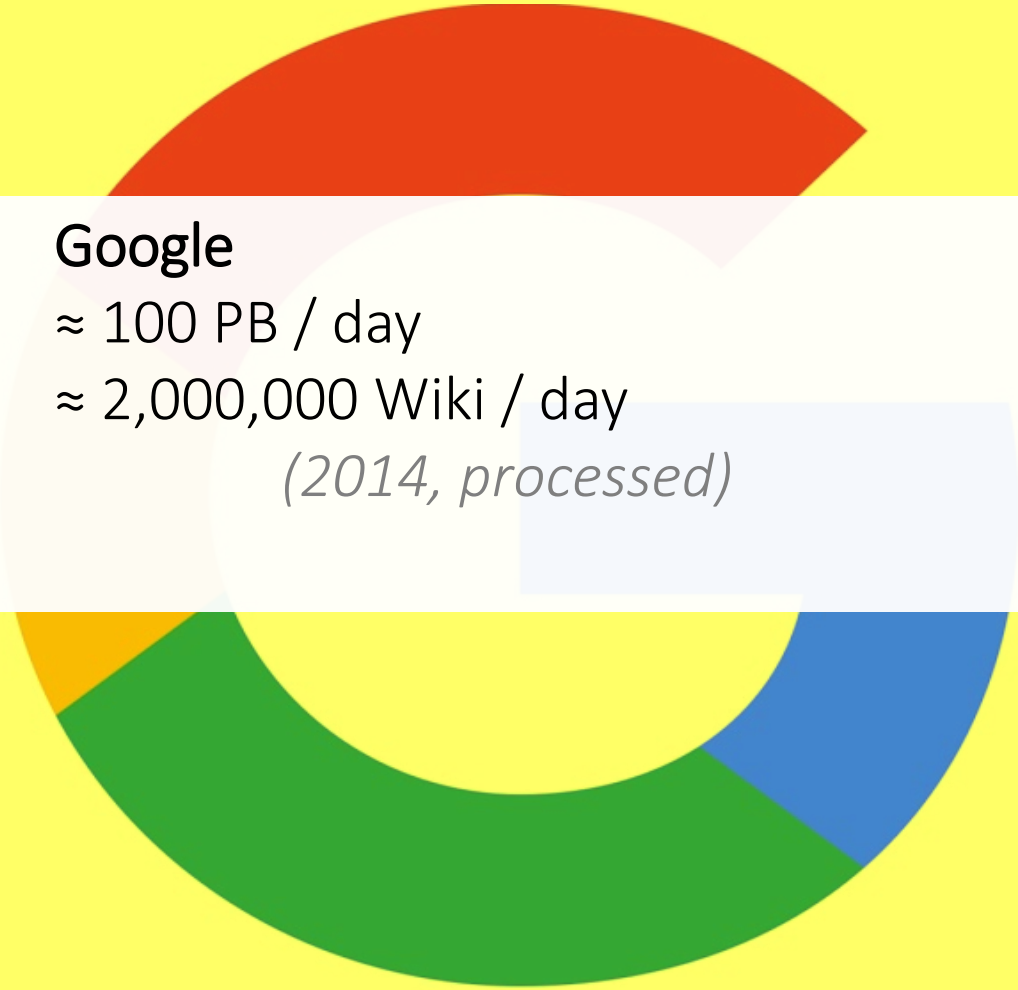
≈ 29 PB / day

≈ 568,627 Wiki / day

(2013, processed)



“Big Data”



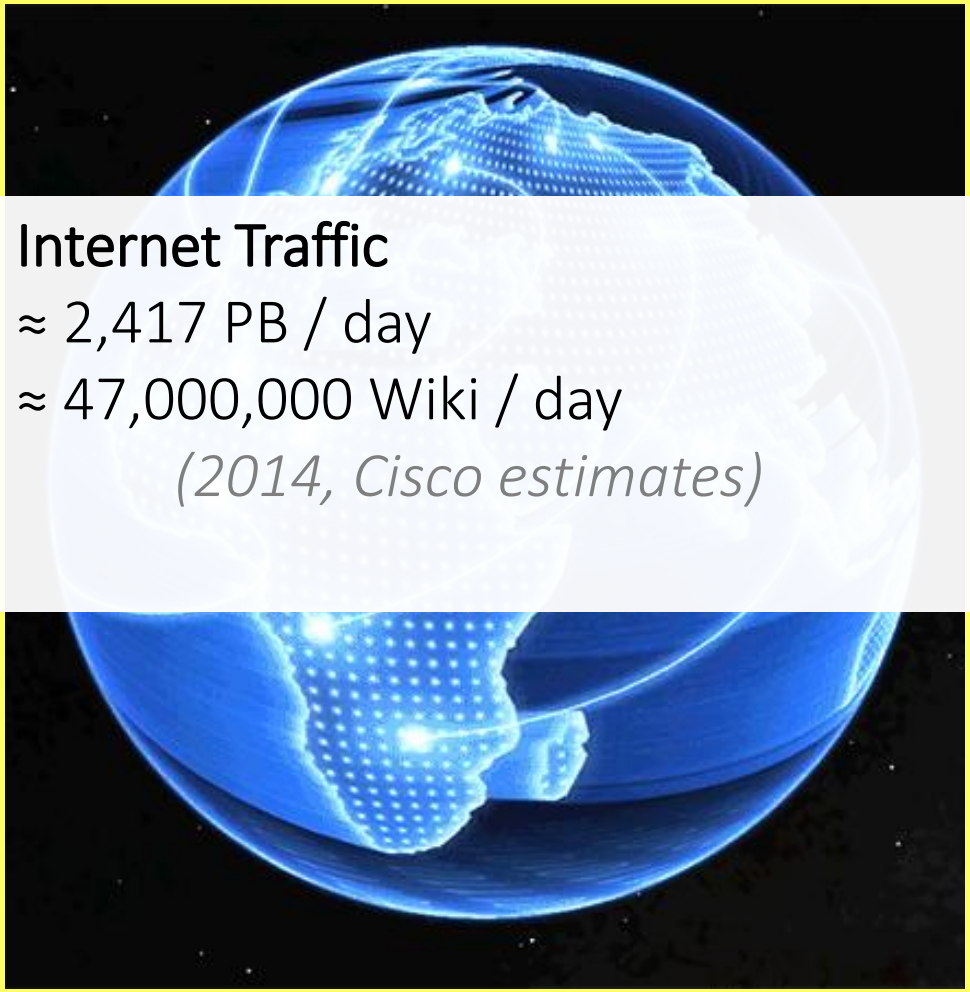
Google

≈ 100 PB / day

≈ 2,000,000 Wiki / day

(2014, processed)

“Big Data”



Internet Traffic

≈ 2,417 PB / day

≈ 47,000,000 Wiki / day

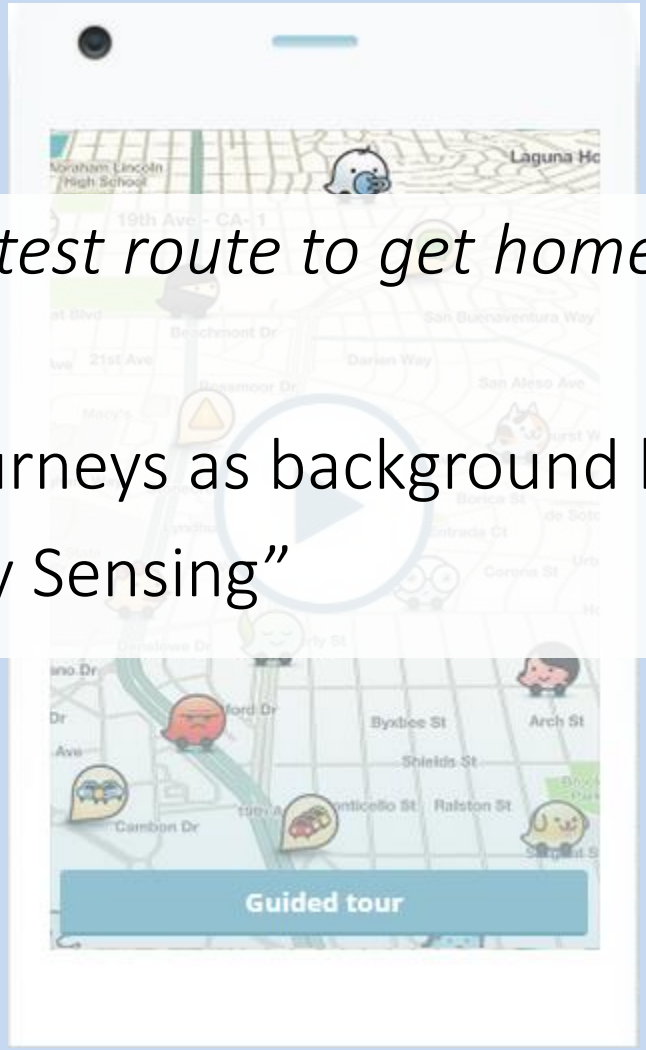
(2014, Cisco estimates)

Data: A Modern-day Bottleneck?

The ‘V’s of “Big Data”

“Big Data” in action …

Getting Home (Waze)



"What's the fastest route to get home right now?"

- Processes journeys as background knowledge
- "Participatory Sensing"

Predicting Pre-crime (PredPol)

PredPol®
Predict Crime in Real Time

"What areas of the city are most need of police patrol at 13:55 on Mondays?"

- PredPol system used by Santa Cruz (US) police patrols
- Predictions based on 8 years of historical crime data

Getting Elected President (Narwhal)

"Who are the undecided voters and how can I convince them to vote for me?"

- User profiles built and integrated from online sources
- Targeted messages sent to voters based on profile

Winning Jeopardy (IBM Watson)

"Can a computer beat human experts at Jeopardy?"

- Indexed 200 million pages of content
- An ensemble of 100 processing techniques

“Big Data” needs “Massive Data Processing” …

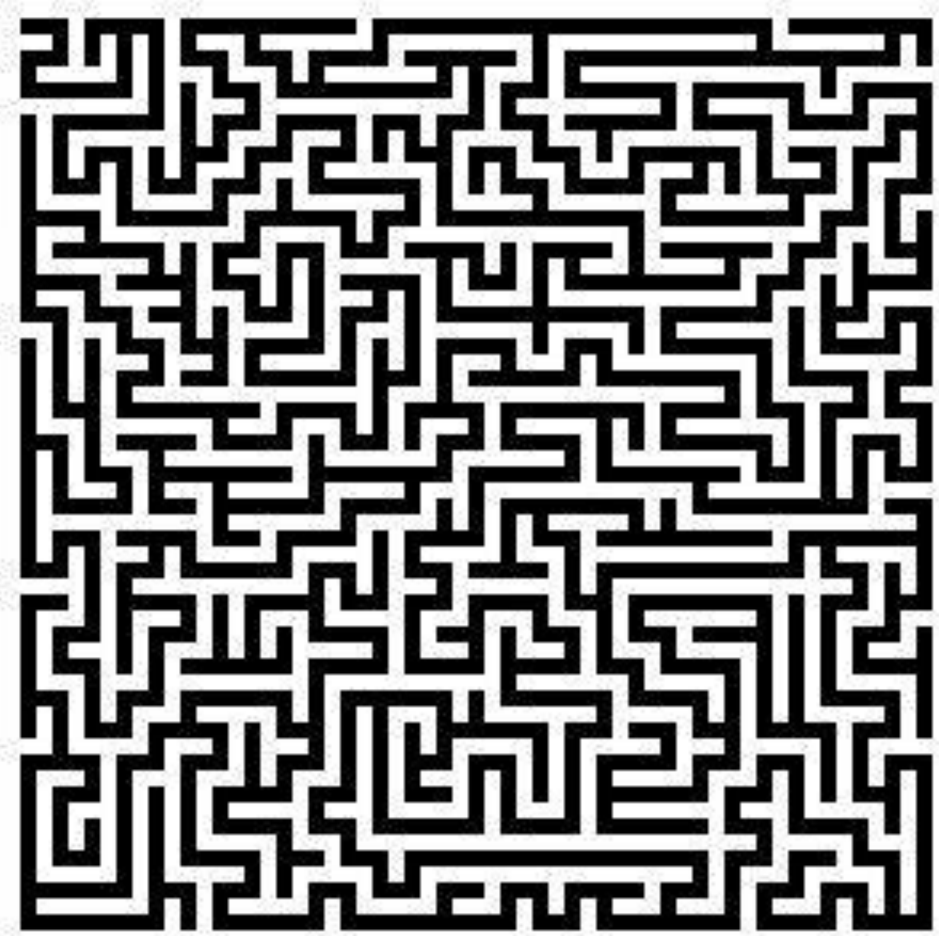
Every Application is Different …

- **Data** can be
  - (Semi-)Structured data
    - (Relational DBs, JSON, XML, CSV, HTML form data)
  - Unstructured data
    - (text document, comments, tweets)
  - And everything in-between!

Every Application is Different …

- **Processing** can involve:
  - Database Management/Analytics
    - (indexing, querying, joins, aggregation)
  - Natural Language Processing
    - (keyword search, topic extraction, entity recognition, machine translation, sentiment analysis, etc.)
  - Data Mining and Statistics
    - (pattern recognition, classification, event detection, recommendations, etc.)
  - Or something else / A mix

So where to start?



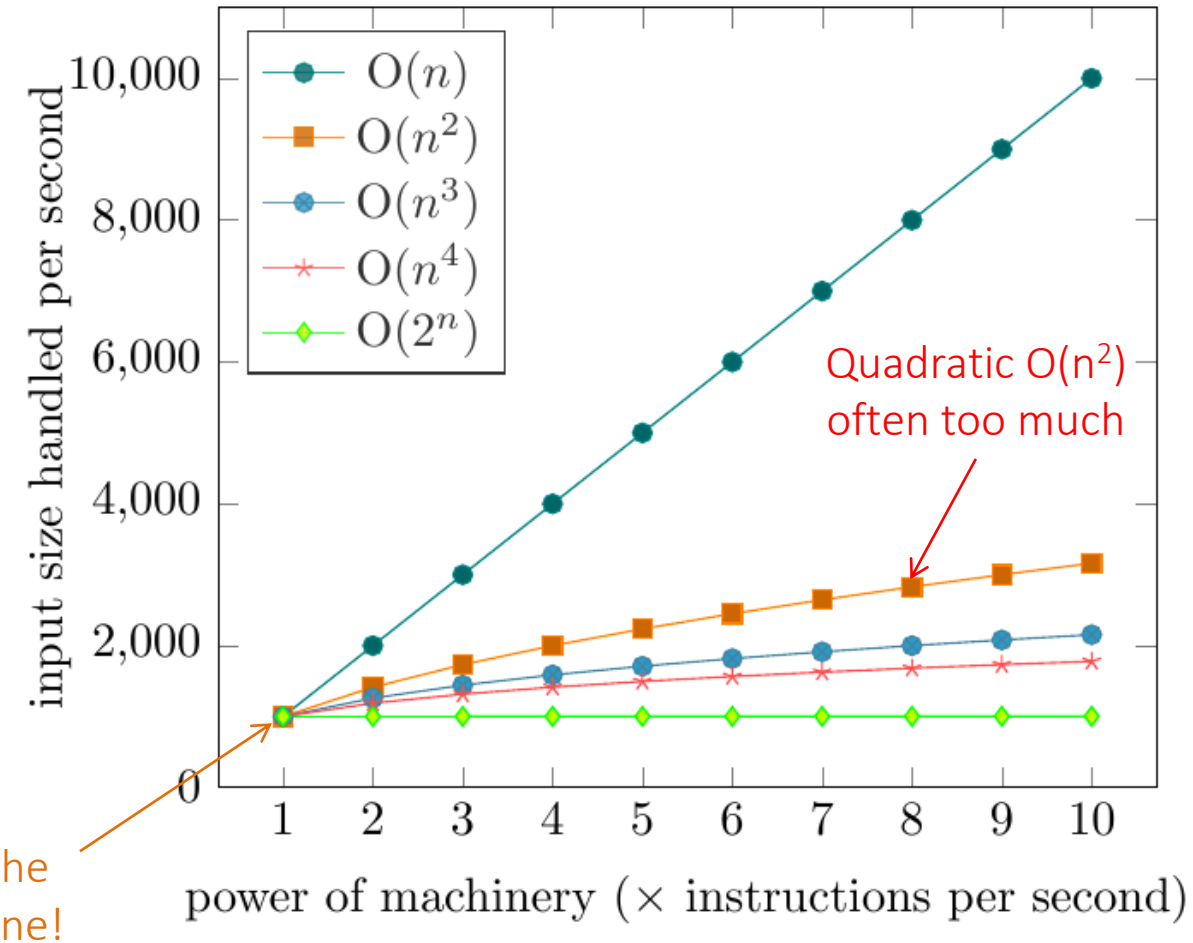
Scale is a Common Factor …

I have an algorithm.

I have a machine that can process 1,000 input items in an hour.

If I buy a machine that is *n* times as powerful, how many input items can I process in an hour?

Depends on what the algorithm is!!



Scale is a Common Factor …

- One machine that's *n* times as powerful?

vs.

- *n* machines that are equally as powerful?

Scale is a Common Factor …

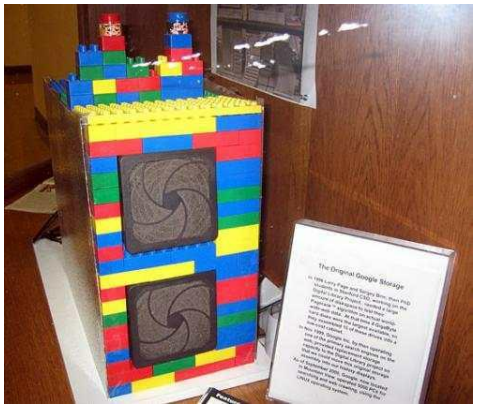
- Data-intensive (our focus!)
  - Inexpensive algorithms / Large inputs
  - e.g., Google, Facebook, Twitter
- Compute-intensive (not our focus!)
  - More expensive algorithms / Smaller inputs
  - e.g., climate simulations, chess games, combinatorials
- No black and white!

“Massive Data Processing” needs “Distributed Computing” …

Distributed Computing

- Need more than one machine!

- Google ca. 1998:



Distributed Computing

- Need more than one machine!

- Google ca. 2014:

Data Transport Costs

- Need to divide tasks over many machines
  - Machines need to communicate

    … but not too much!
  - Data transport costs (*simplified*):

Need to minimise network costs!

Data Placement

- Need to think carefully about where to put what data!

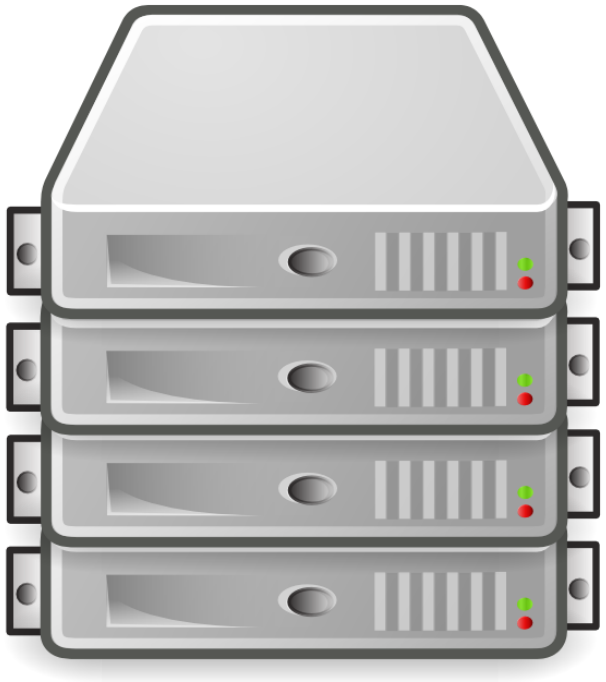
I have four machines to run a website. I have 10 million users.

Each user has personal profile data, photos, friends and games.

How should I split the data up over the machines?

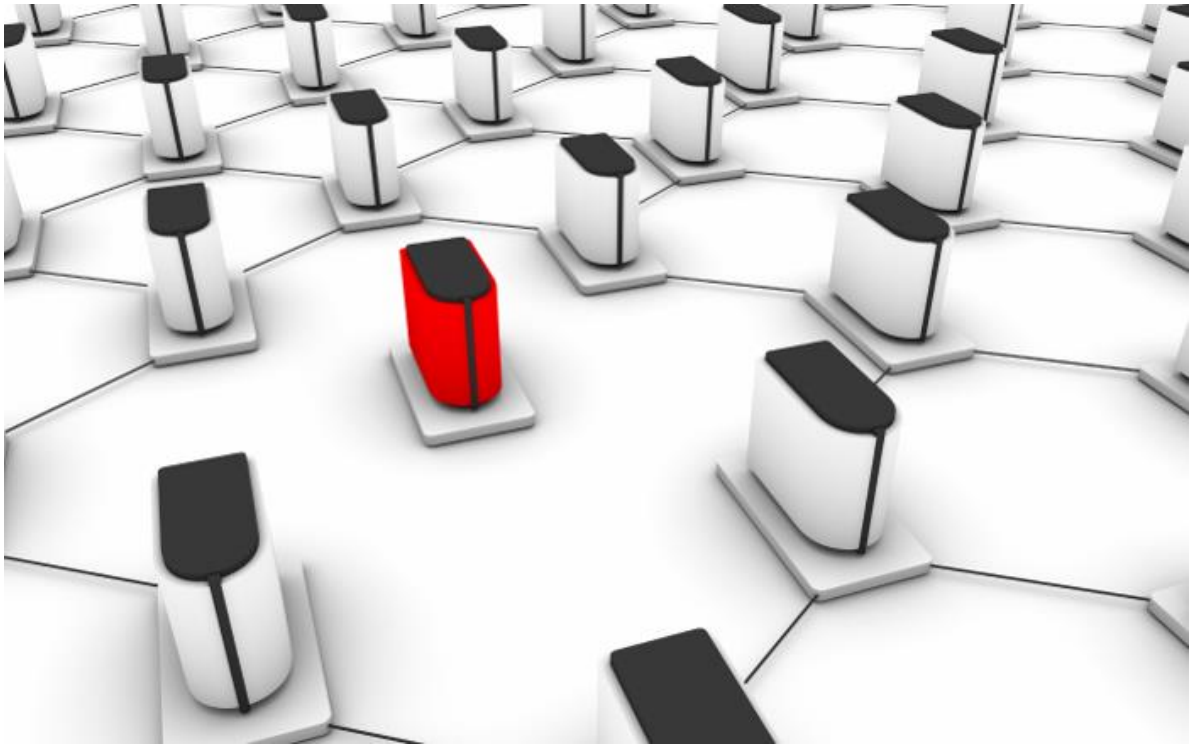
Depends on the application!

But some general principles and design choices apply.



Network/Node Failures

- Need to think about failures!



Network/Node Failures

- Need to think (even more!) carefully about where to put what data!

I have four machines to run a website. I have 10 million users.

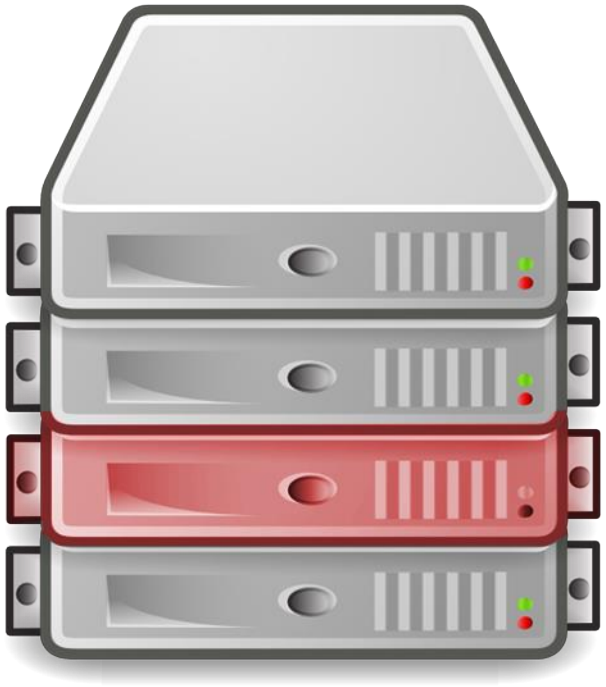
Each user has personal profile data, photos, friends and games.

How should I split the data up over the machines?

(Again)

Depends on the application!

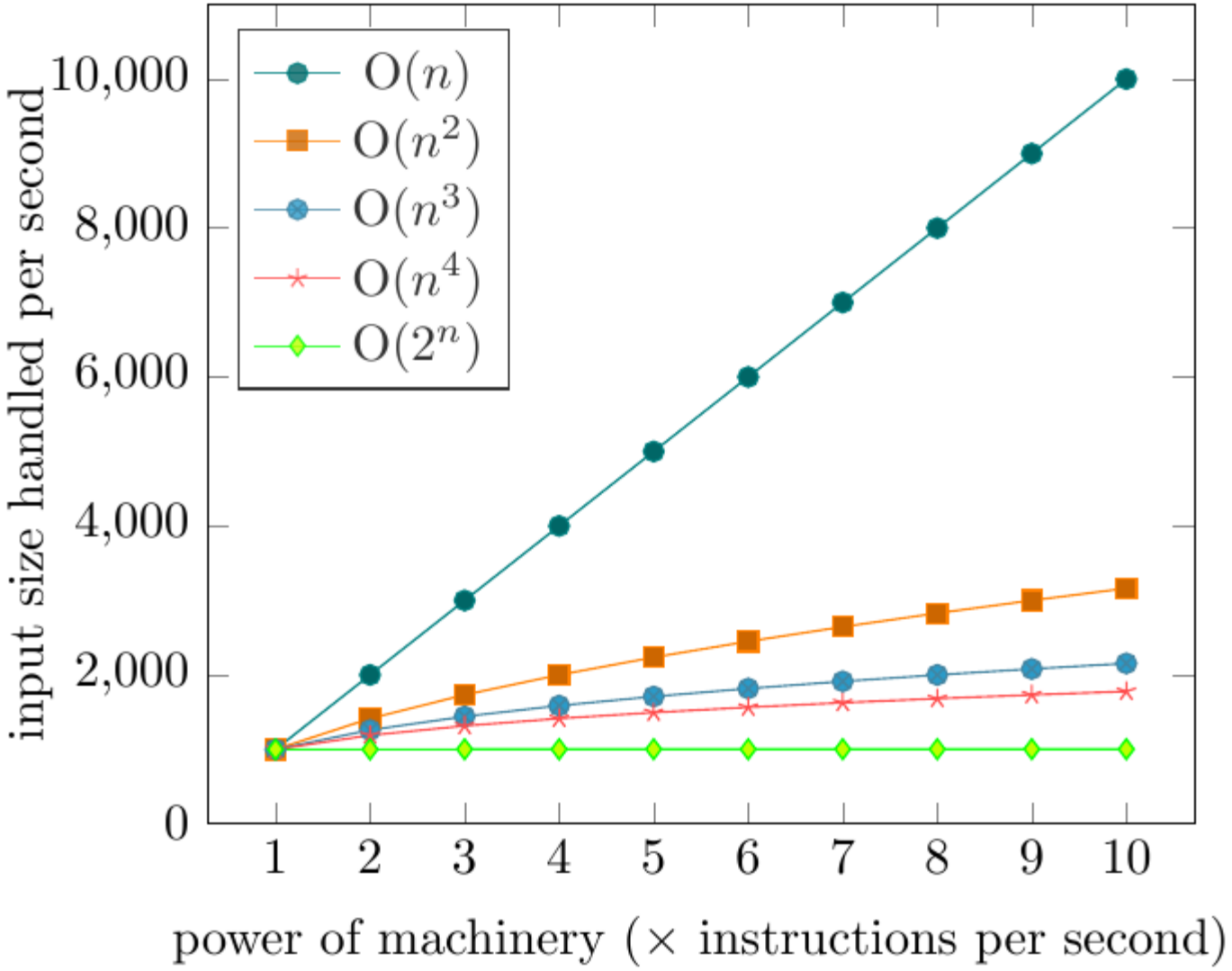
But some general principles and design choices apply.



Human Distributed Computation

“Distributed Computing” Limits & Challenges …

Distribution Not Always Applicable!



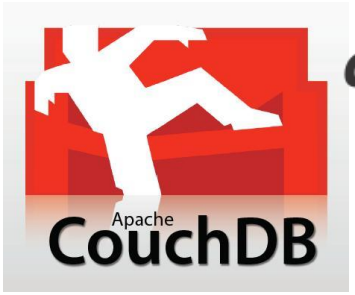
Distributed Development Difficult

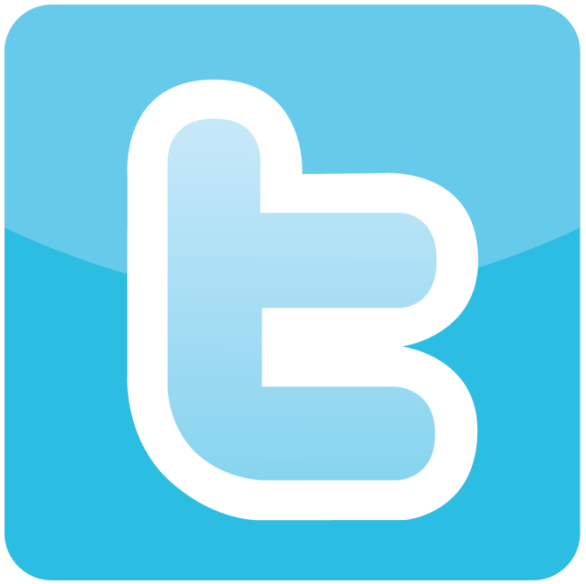
- Distributed systems can be complex
- Multiple machines; need to take care of
  - Data in different locations
  - Logs and messages in different places
  - Different users with different priorities
  - Different network capabilities
  - Need to balance load!
  - Need to handle failures!
- Tasks may take a long time!
  - Bugs may not become apparent for hours
  - Lots of data = lots of counter-examples

Frameworks/Abstractions can Help

- For Distrib. Processing

- For Distrib. Storage





HOW DOES TWITTER WORK?

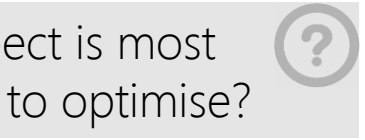
Based on 2013 slides by Twitter lead architect: Raffi Krikorian

"Twitter Timelines at Scale"

Big Data at Twitter

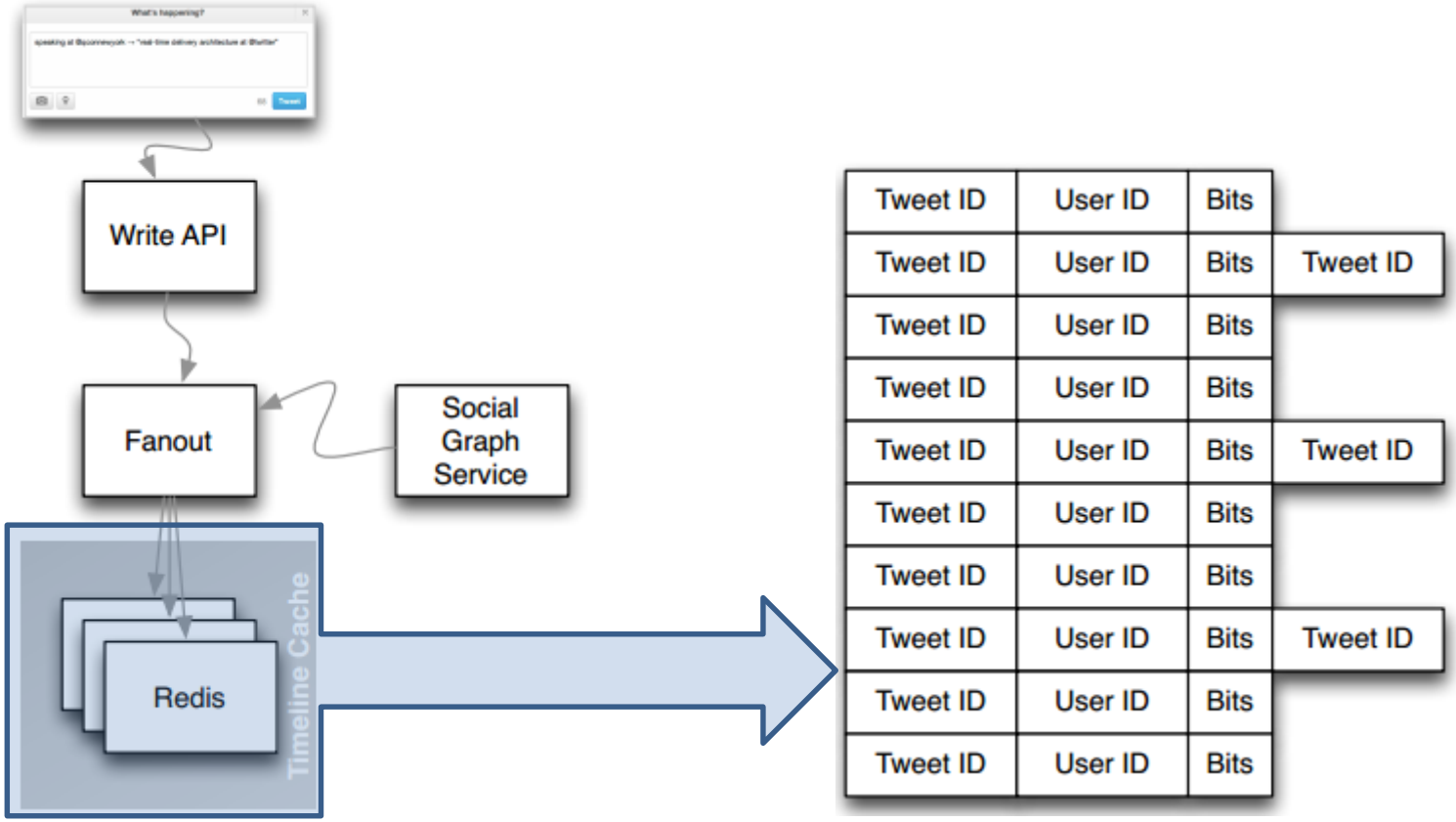
- 150 million active worldwide users
- 400 million tweets per day
  - mean: 4,600 tweets/second
  - max: 150,000 tweets/second
- 300,000 queries/second for user timelines
- 6,000 queries/second for custom search

Which aspect is most important to optimise?

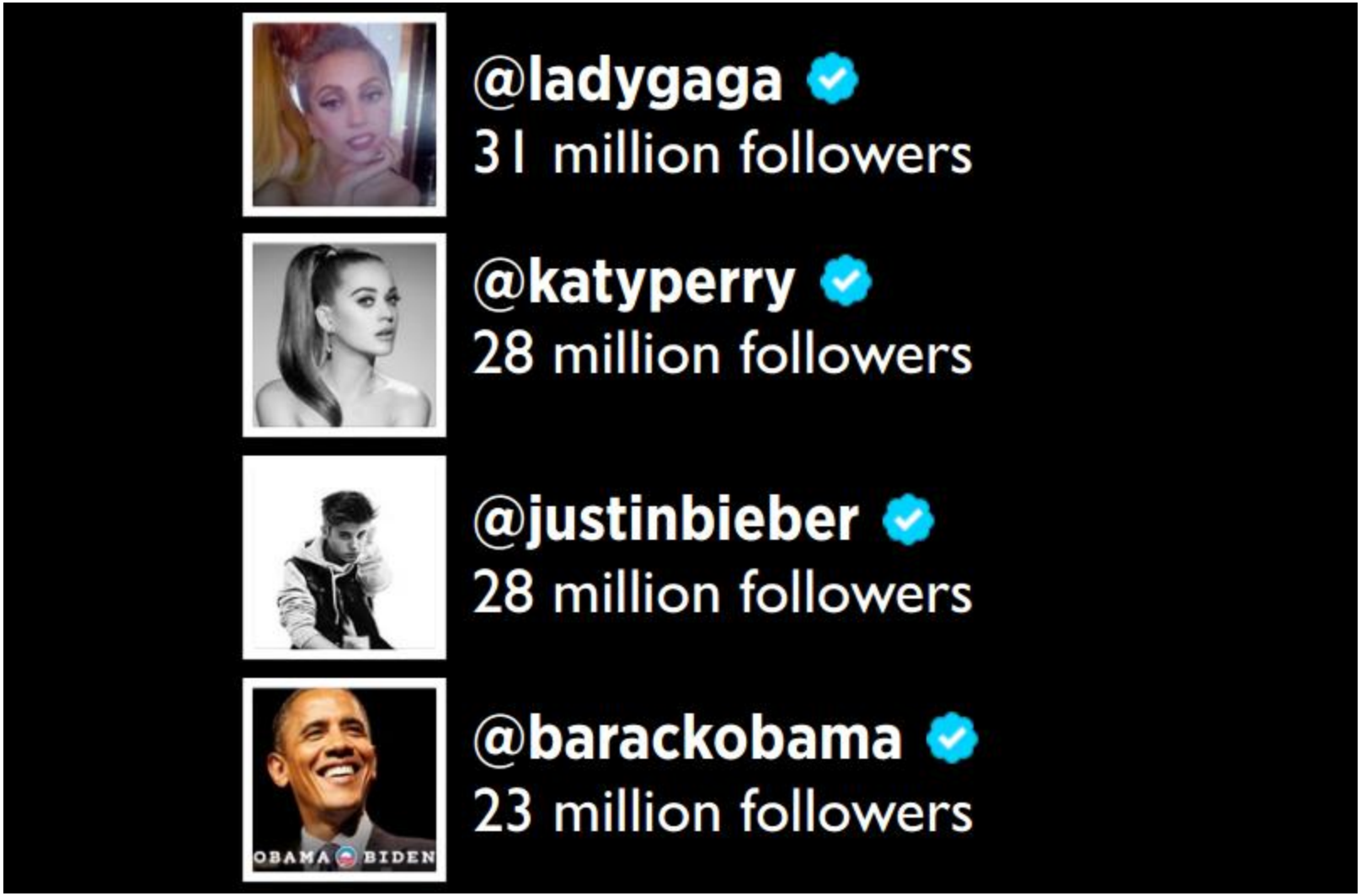


Supporting timelines: write

- mean: 4,000 tweets/second

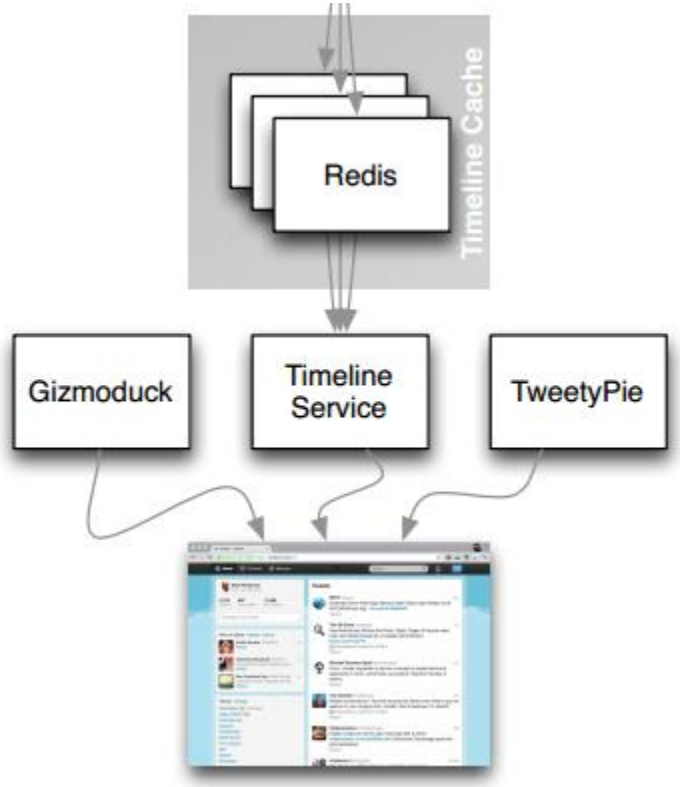


High-fanout


@ladygaga
31 million followers
@katyperry
28 million followers
@justinbieber
28 million followers
@barackobama
23 million followers
OBAMA BIDEN

Supporting timelines: read

- 300,000 queries/second

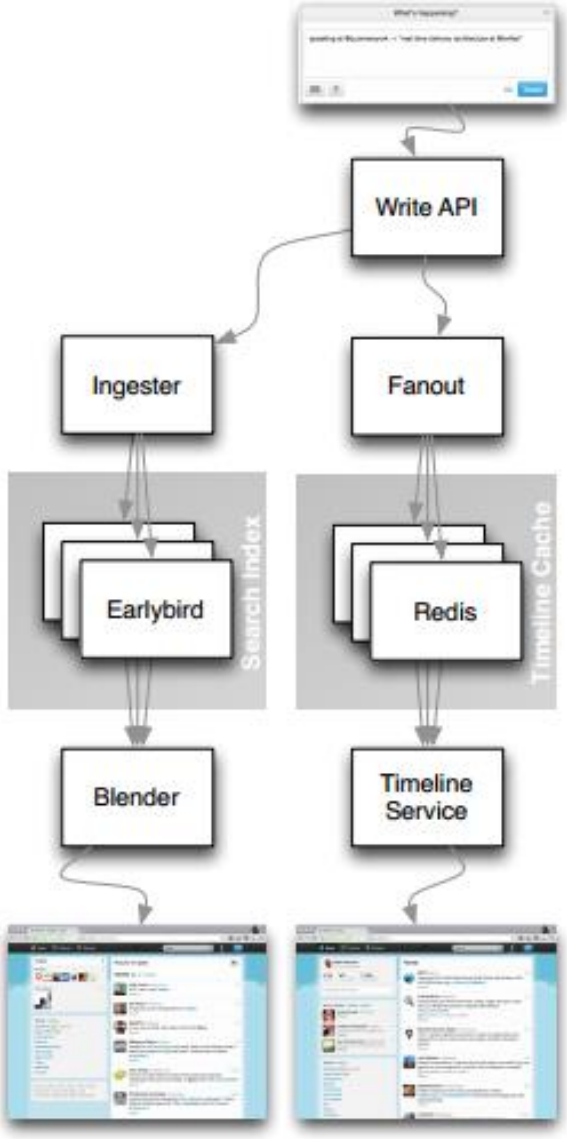


Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	Tweet ID
Tweet ID	User ID	Bits	
Tweet ID	User ID	Bits	

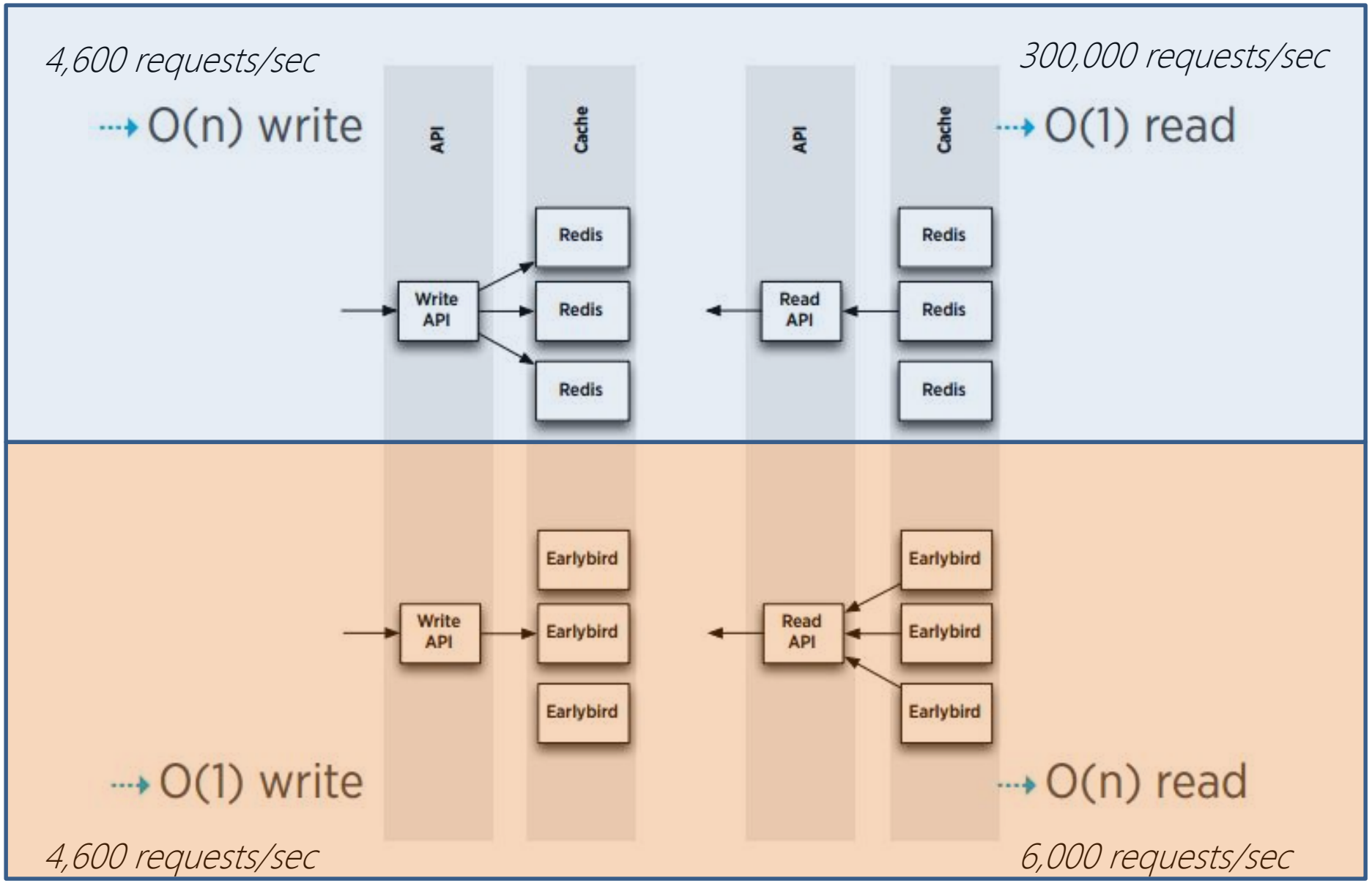
1ms @p50
4ms @p99

Supporting text search

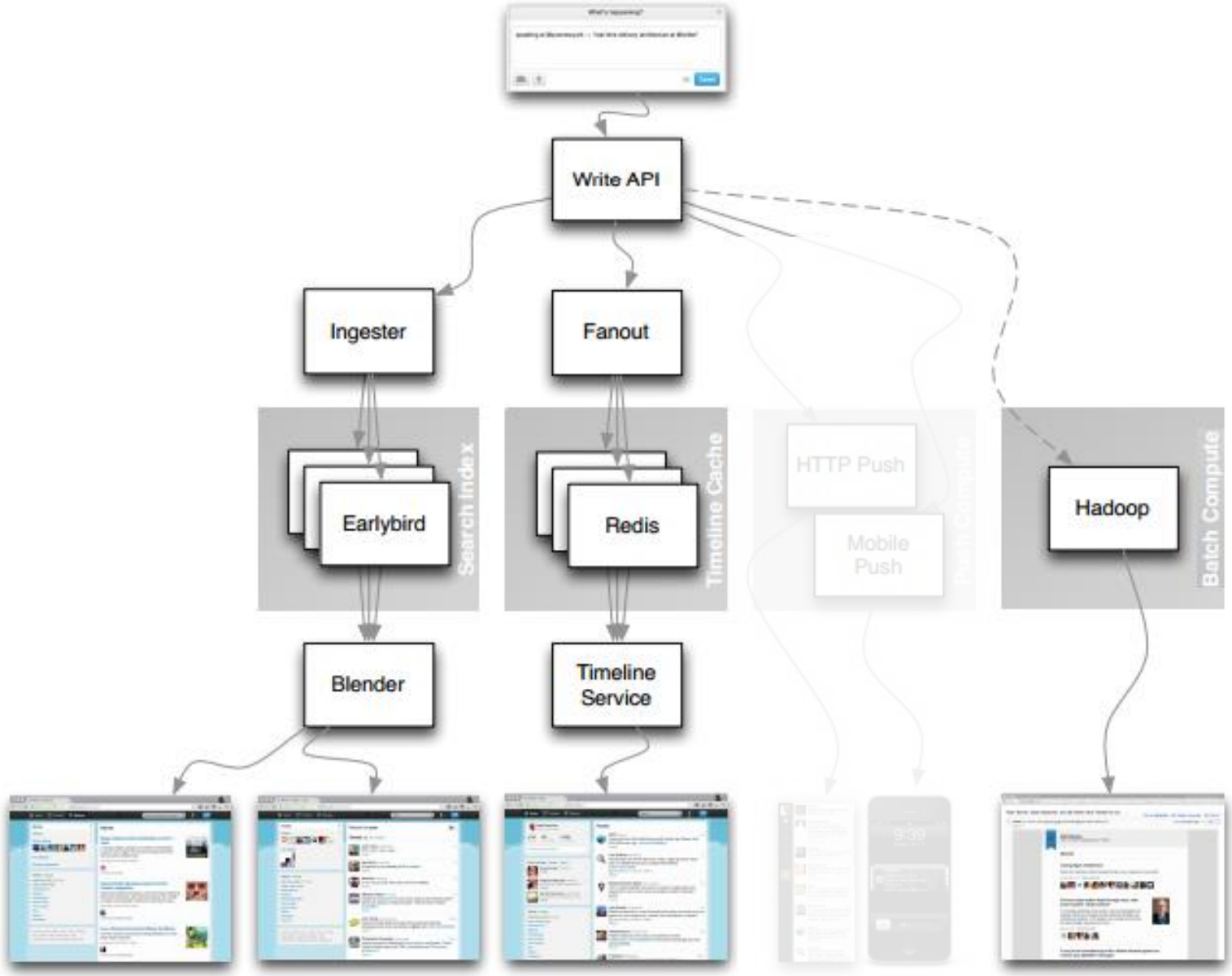
- Information retrieval
  - Earlybird: Lucene clone
  - Write once
  - Query many



Timeline vs. Search



Twitter: Full Architecture



“Procesamiento Masivo de Datos”
About the course …

What the Course Is/Is Not

- Data-intensive not compute-intensive
- Distributed tasks not networking
- Commodity hardware not supercomputers
- General methods not specific algorithms
- Practical methods with a little theory

What the Course Is

- Principles of Distributed Computing [1 week]
- Distributed Processing Frameworks [4 weeks]
- Information Retrieval [3 weeks]
- Principles of Distributed Databases [3 weeks]
- Projects [1–2 weeks]

Course Structure

- ~1.5 hours of lectures per week [Monday]
- 1.5 hours of labs per week [Wednesday]
  - To be turned in by next Monday evening
  - Mostly Java
- Auxiliar Session [Friday]
  - Mostly for controls
  - We will announce in the forum if there's a session

http://aidanhogan.com/teaching/cc5212-1-2018/

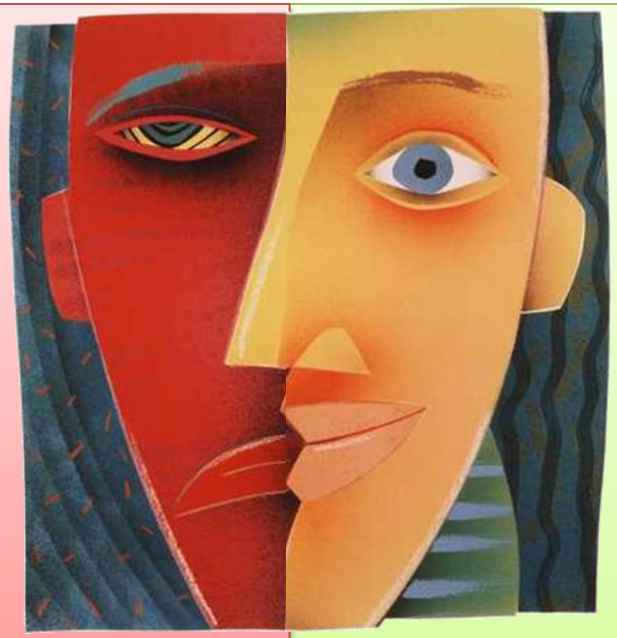
Course Marking

- 55% for Weekly Labs (~5% a lab!)
- 15% for Class Project
- 30% for 2x Controls

Assignments each week

Controls

Working in groups



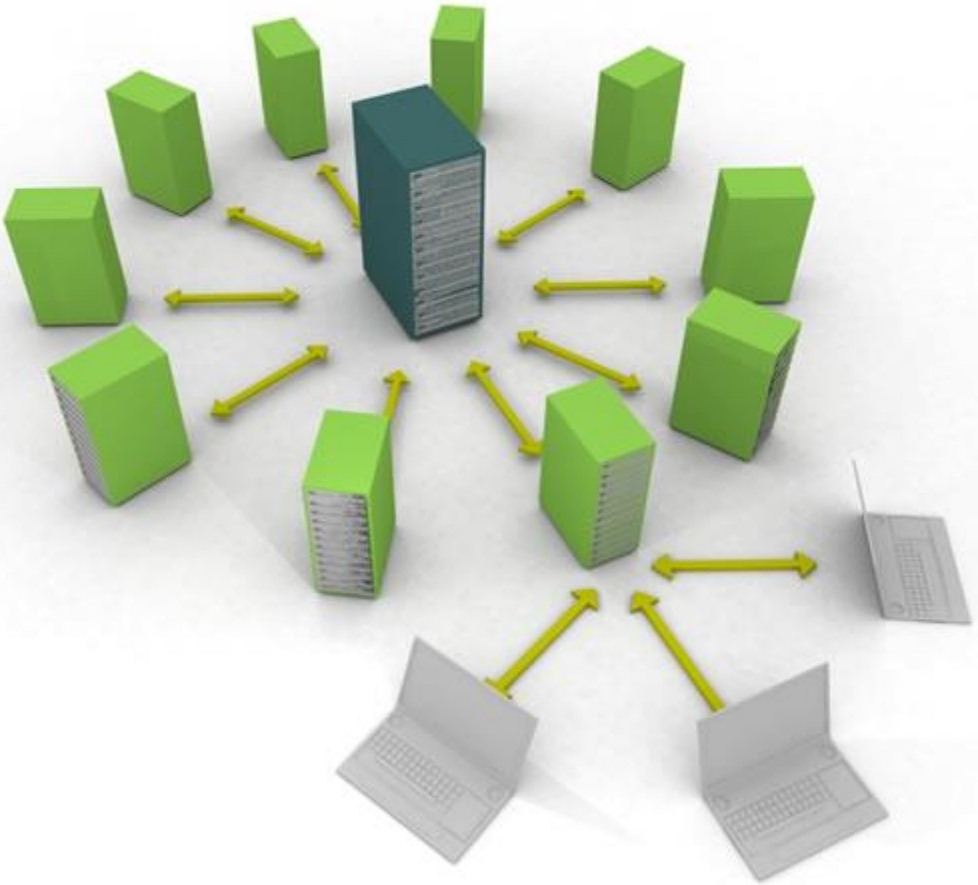
Only need to pass overall!

No final exam!

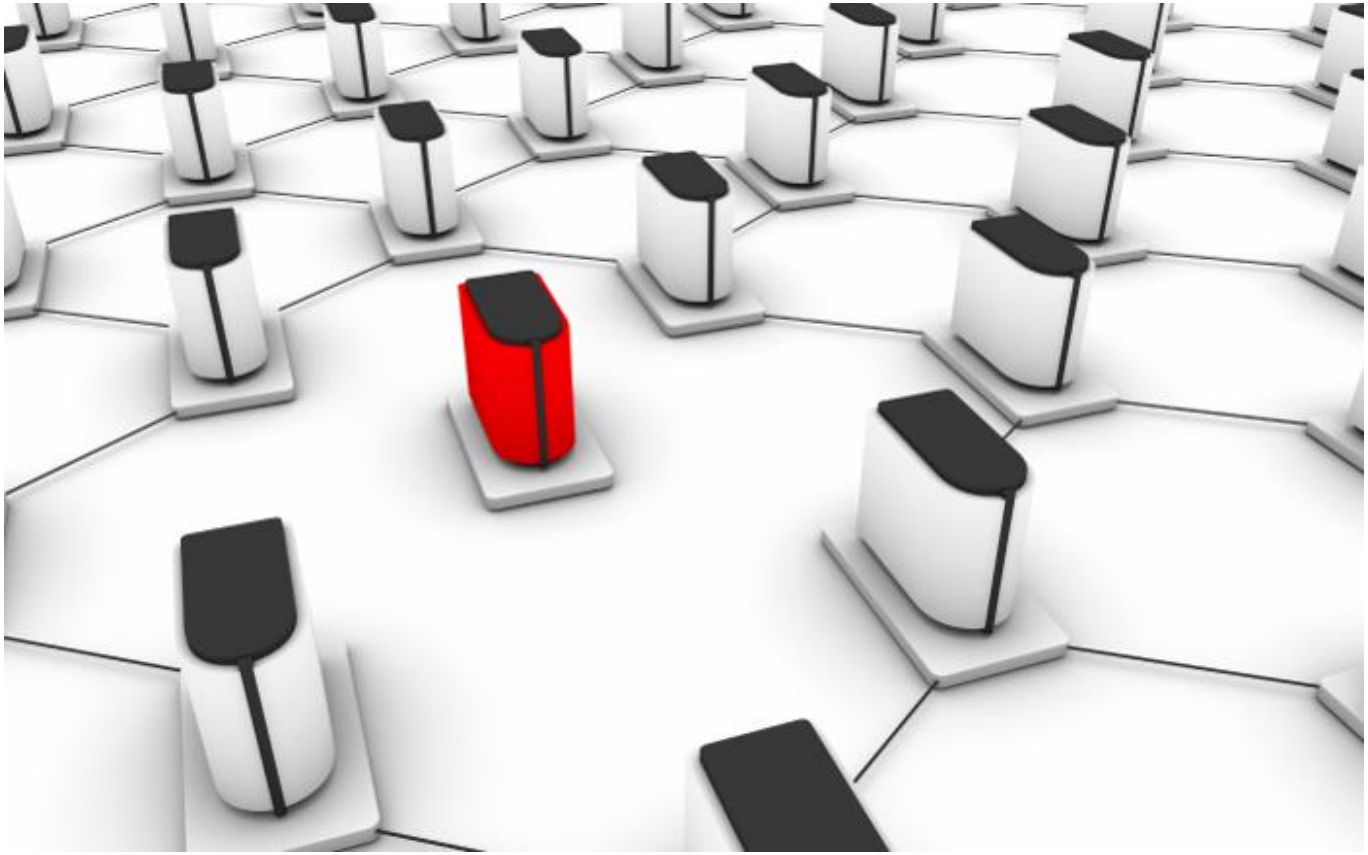
Working in groups!

Outcomes!

Outcomes!



Outcomes!

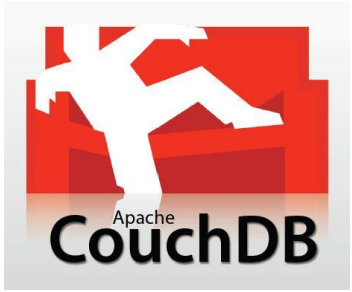


Outcomes!

hadoop

HBASE

cassandra


Apache
CouchDB

Apache
Solr

APACHE
GIRAPH

mongoDB

accumulo

Outcomes!

Questions?