THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

8. The Cloud & Data

CSCI 2541W Database Systems & Team Projects

Wood - 2022

Upcoming

Last week: Exam

- Some students have not taken it, please do not discuss

Today:

- Large scale data and web applications
- Teamwork

Wednesday:

- Project introduction!
- Lab on session programming + Wordle

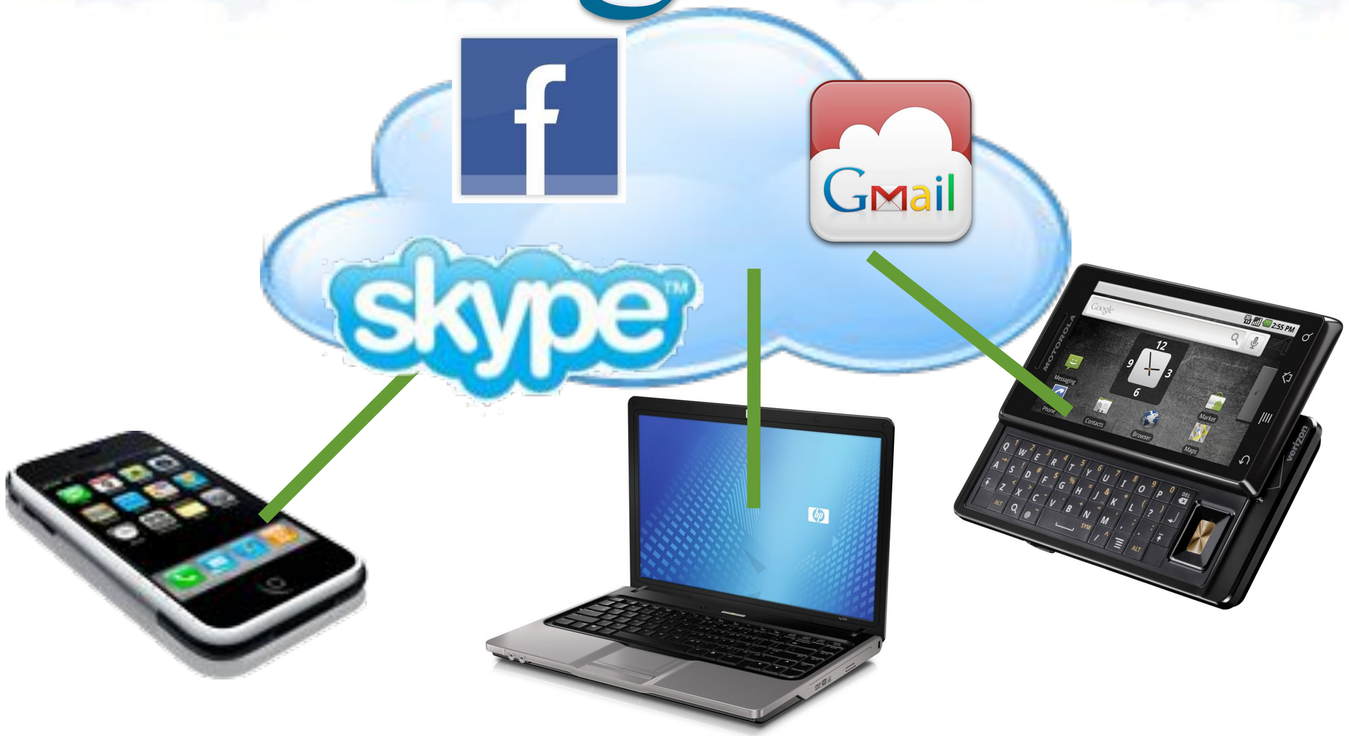
Next Tuesday 3/12: Shopping cart due!

- If you aren't at least halfway done, you are behind!

What is the oldest piece of software you remember using?

Software Changed

Then



Now

Where and how we run programs has changed

- Network connected
- Mobile
- Multi-media content
- Shared by lots of users

Cloudy Buzz

Mobile
!

Google Docs

iCloud

Dropbox

flickr

Fast!

Powerful!

amazon web services

XBOX LIVE

Gmail

Free!*

What *is* a cloud?

<spoiler alert>
It’s not in the sky
it’s not made of water droplets
</spoiler alert>

Some big buildings…

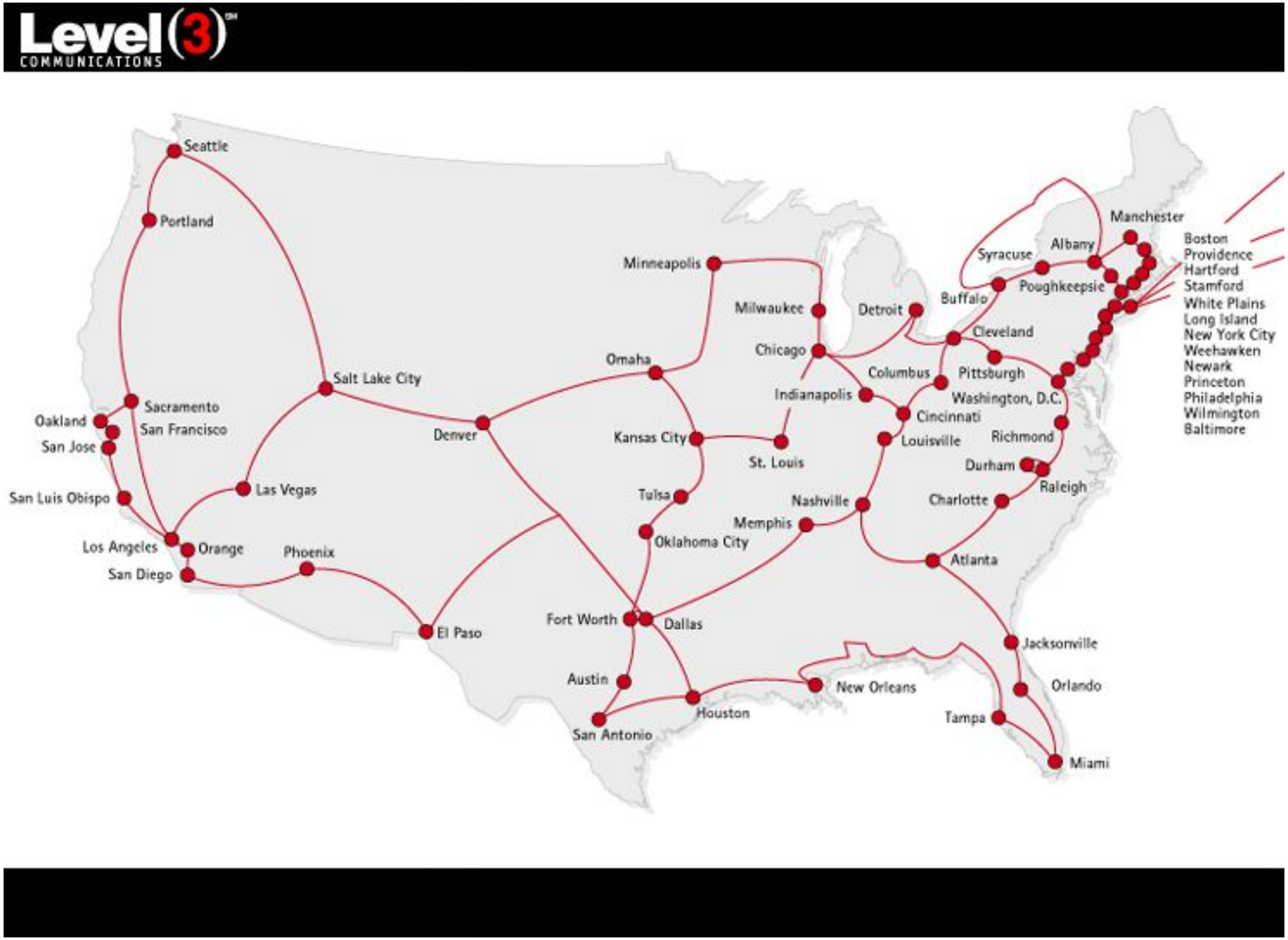
Microsoft’s Dublin data center

...and computers...

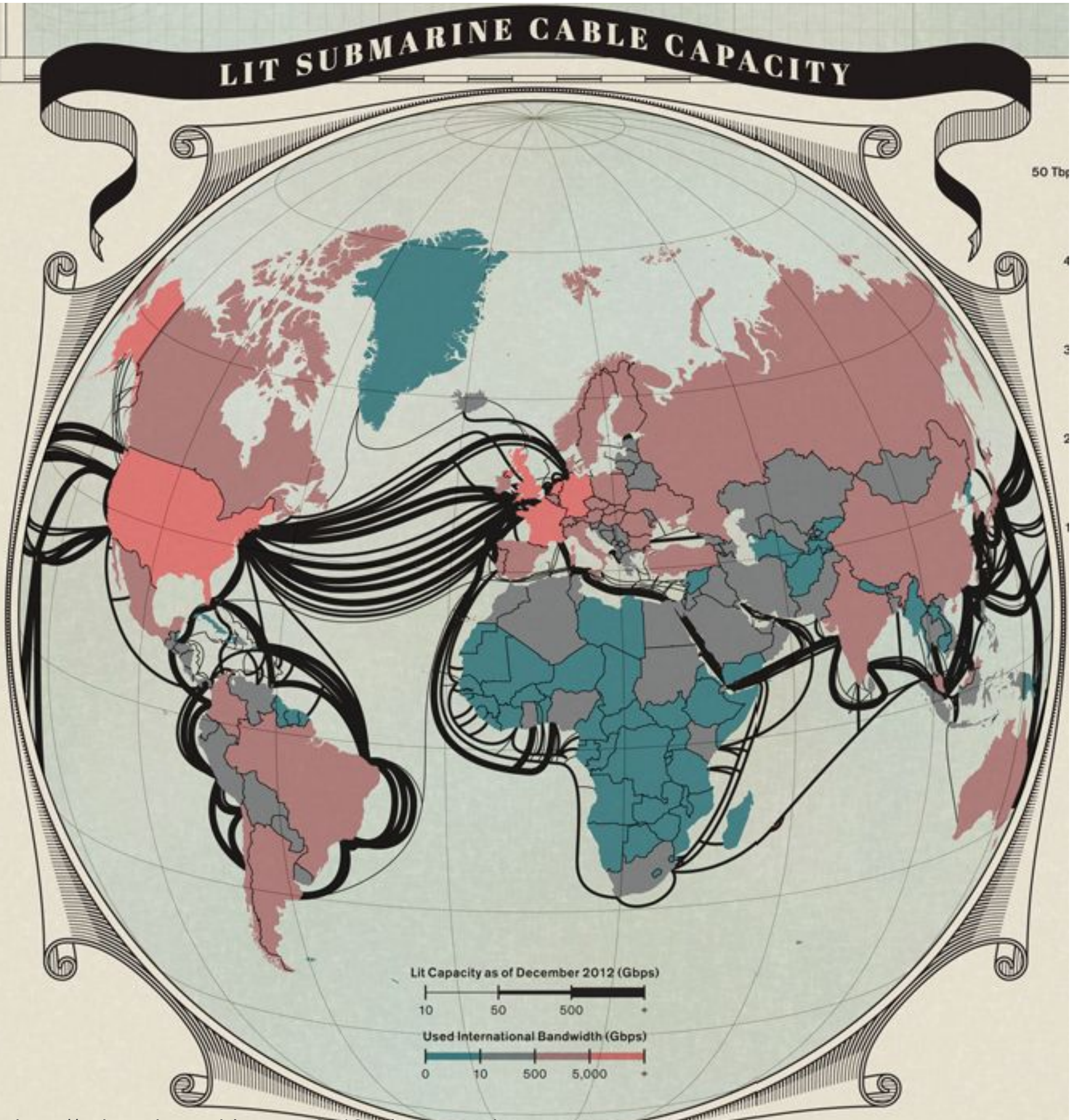
Giant warehouses

- The size of 10 football fields
- 10s of thousands of servers
- Petabytes of storage

…interconnected…



...around the world...


LIT SUBMARINE CABLE CAPACITY
50 Tbp
Lit Capacity as of December 2012 (Gbps)
10
50
500
+
Used International Bandwidth (Gbps)
0
10
500
5,000
+

http://submarine-cable-map-2013.telegeography.com/

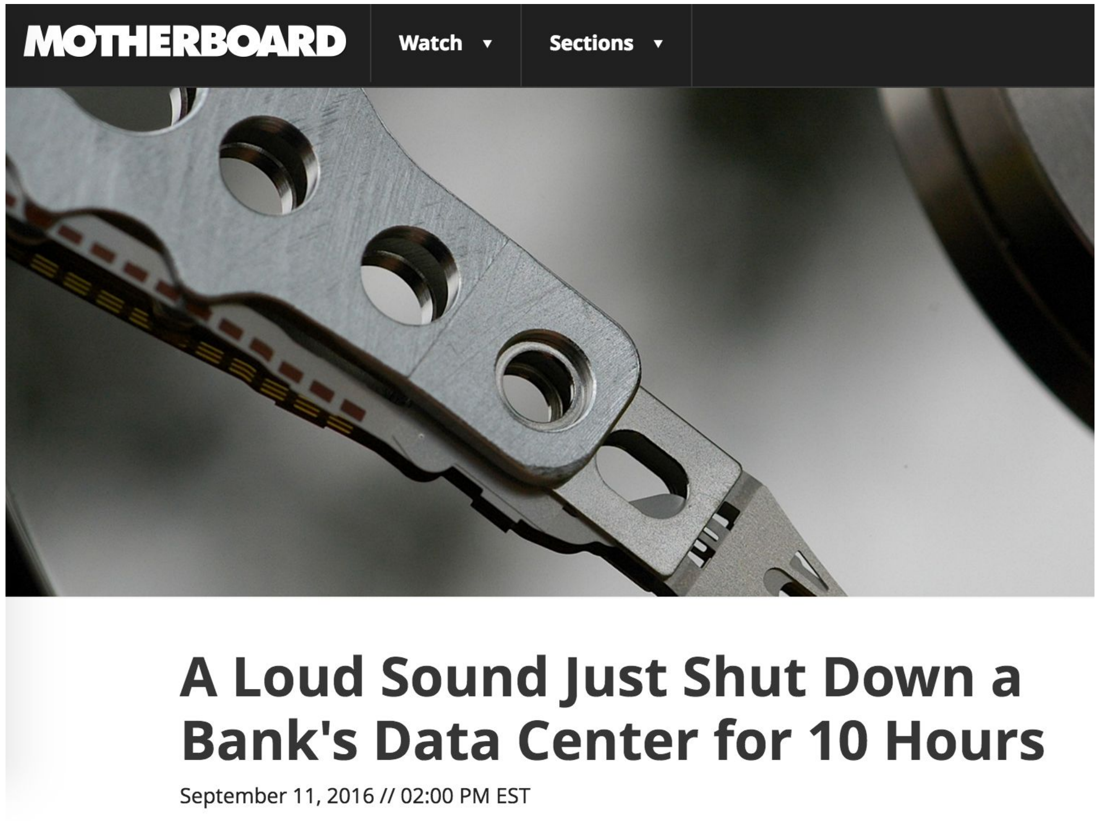
Undersea Cables

- Connect all continents except Antarctica
- First deployed in 1850s

http://www.cyprusupdates.com

...that break a lot.

Lightning causes Amazon outages (2009 and 2011)



A Loud Sound Just Shut Down a Bank's Data Center for 10 Hours

September 11, 2016 // 02:00 PM EST

Comcast down after hunter shoots cable (2008)

Anchor hits underwater Internet cable (Feb 2012)

Or if you're really unlucky...

vs

Cloud Defined

cloud: */kloud/* noun

A **large** collection of computers, accessible over a **network**, running many different types of software as a **shared** service

Must be:

efficient, scalable, secure, reliable, *elastic*

Cloud Examples

Shared, worldwide infrastructure to host email services for many users and organizations

- ~900,000 servers in 2014

Shared storage service

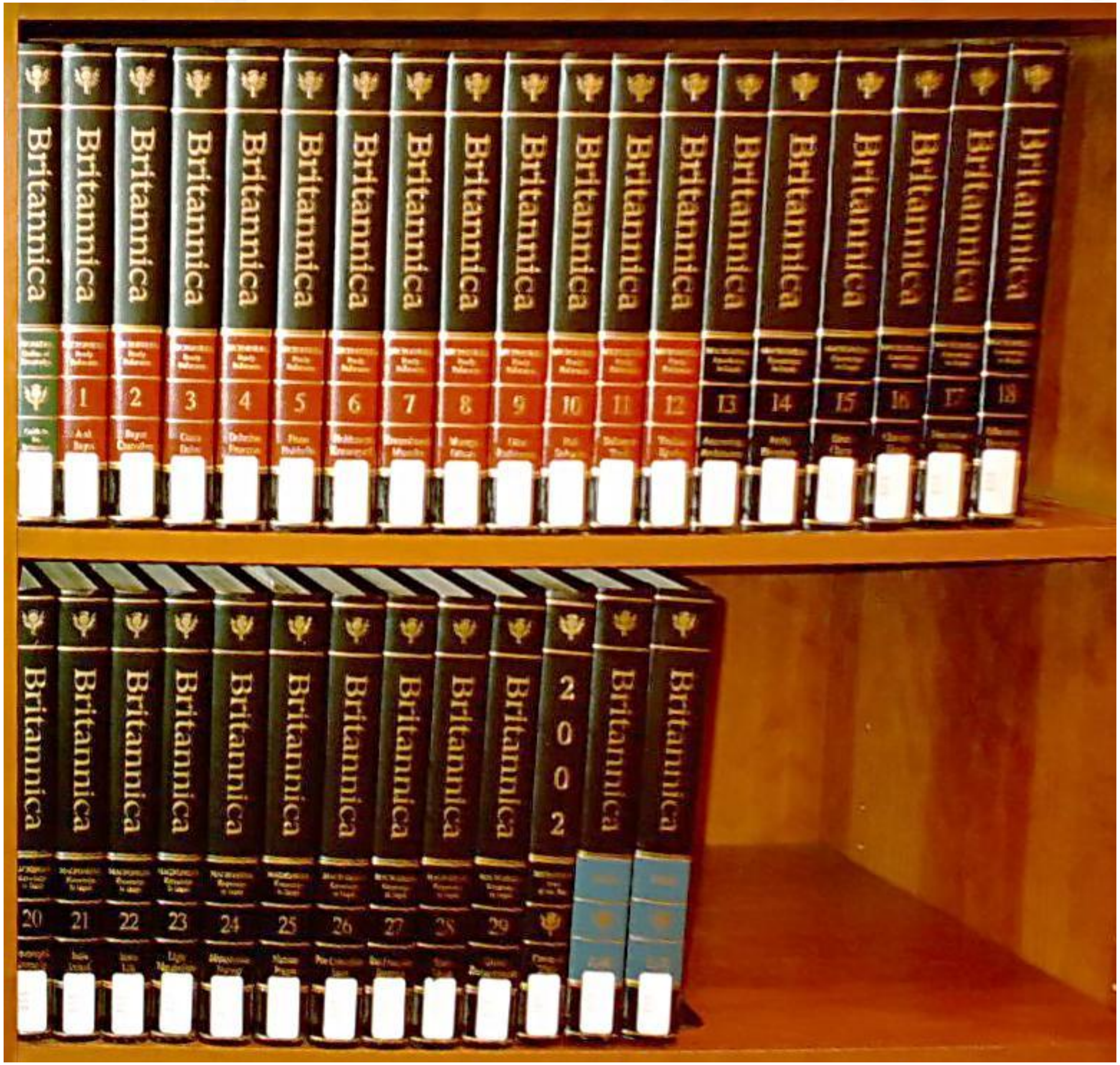
- ~10,000 servers and 200 million users in 2013

Shared computing infrastructure that developers, companies, and students can easily get access to

- ~1.4 million servers in 2014

Why do we need all of this physical infrastructure?

What is this???

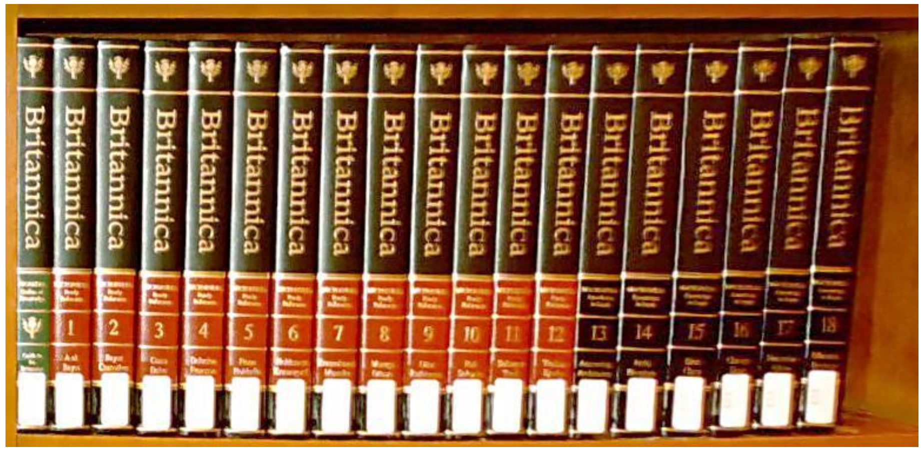


Britannica
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
21
22
23
24
25
26
27
28
29
2002

Encyclopedias

Encyclopedia Britannica

- 40,000+ articles
- 32 hard bound volumes (32,640 pages)



Microsoft Encarta

- 60,000+ articles
- 1 CD-ROM (**700 MB**)

Wikipedia

- 6,383,000 articles (in English)
- More than **5 TB** of text (about 7,500 CDs)

Mega whats?

700MB vs 5TB

Mega	Million	1024 x 1024 = ~1,000,000
Giga	Billion	1024 x 1024 X 1024 = ~1,000,000,000
Tera	**Trillion**	1024 x 1024 x 1024 x 1024= **~1, 000,000,000,000**

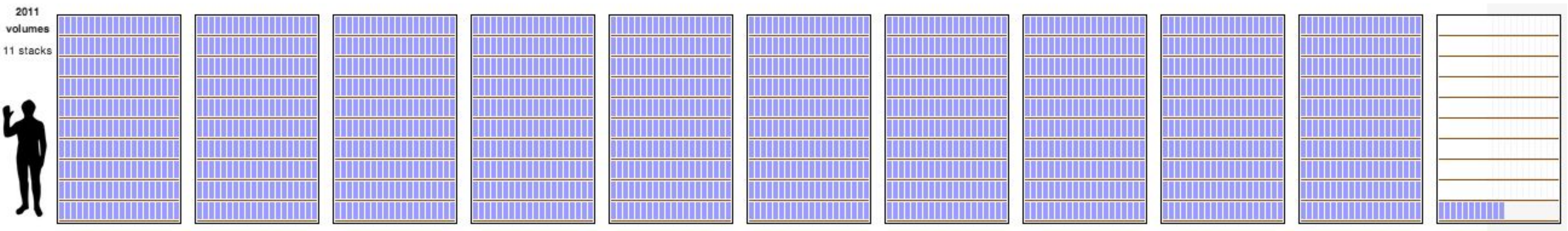
200 photos vs 1.4 million photos

Encyclopedias

Wikipedia... in print

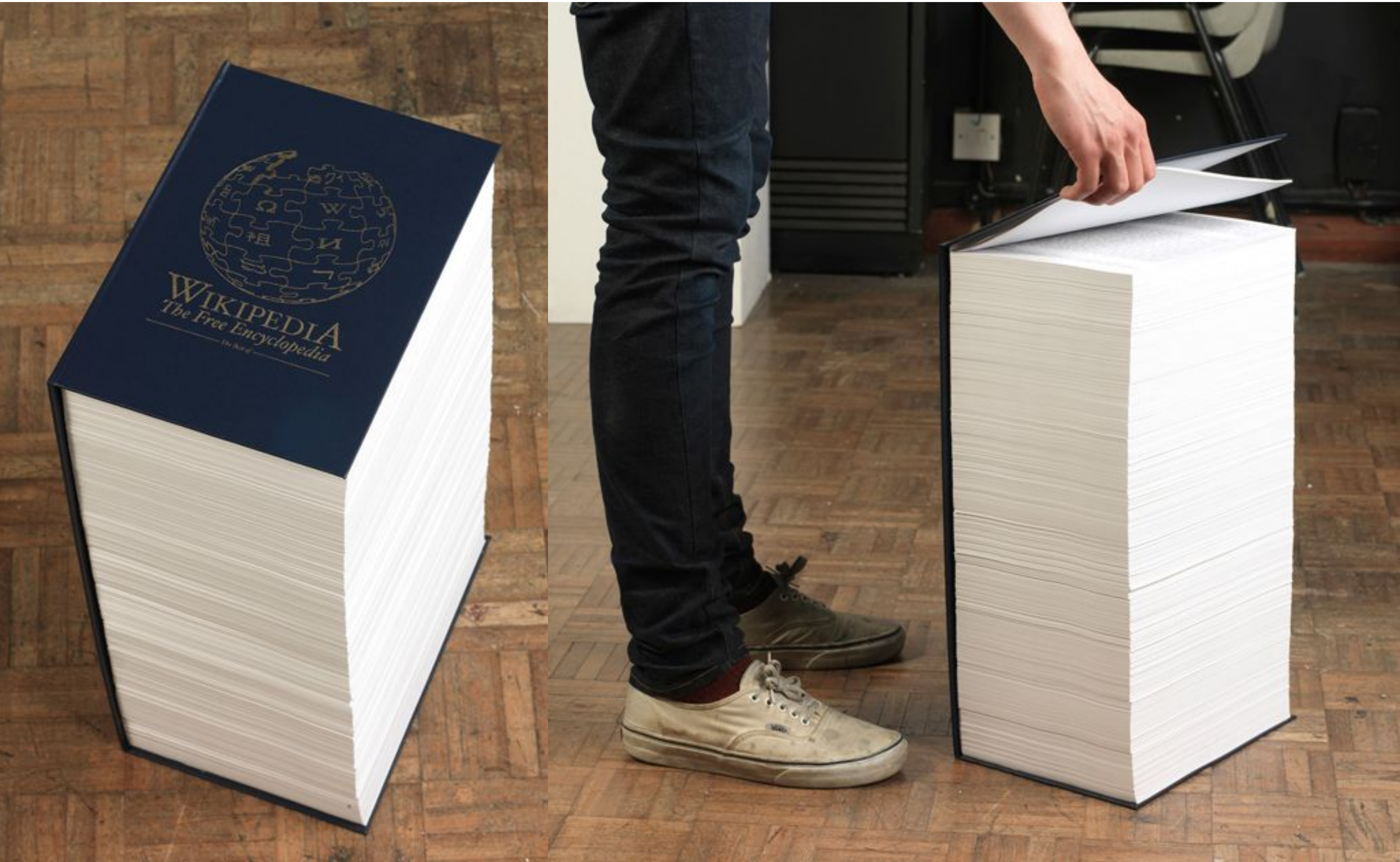
- ~~1,763 volumes~~
- (no, this does not exist)

Now grown to **3,024** volumes and >30TB of data!

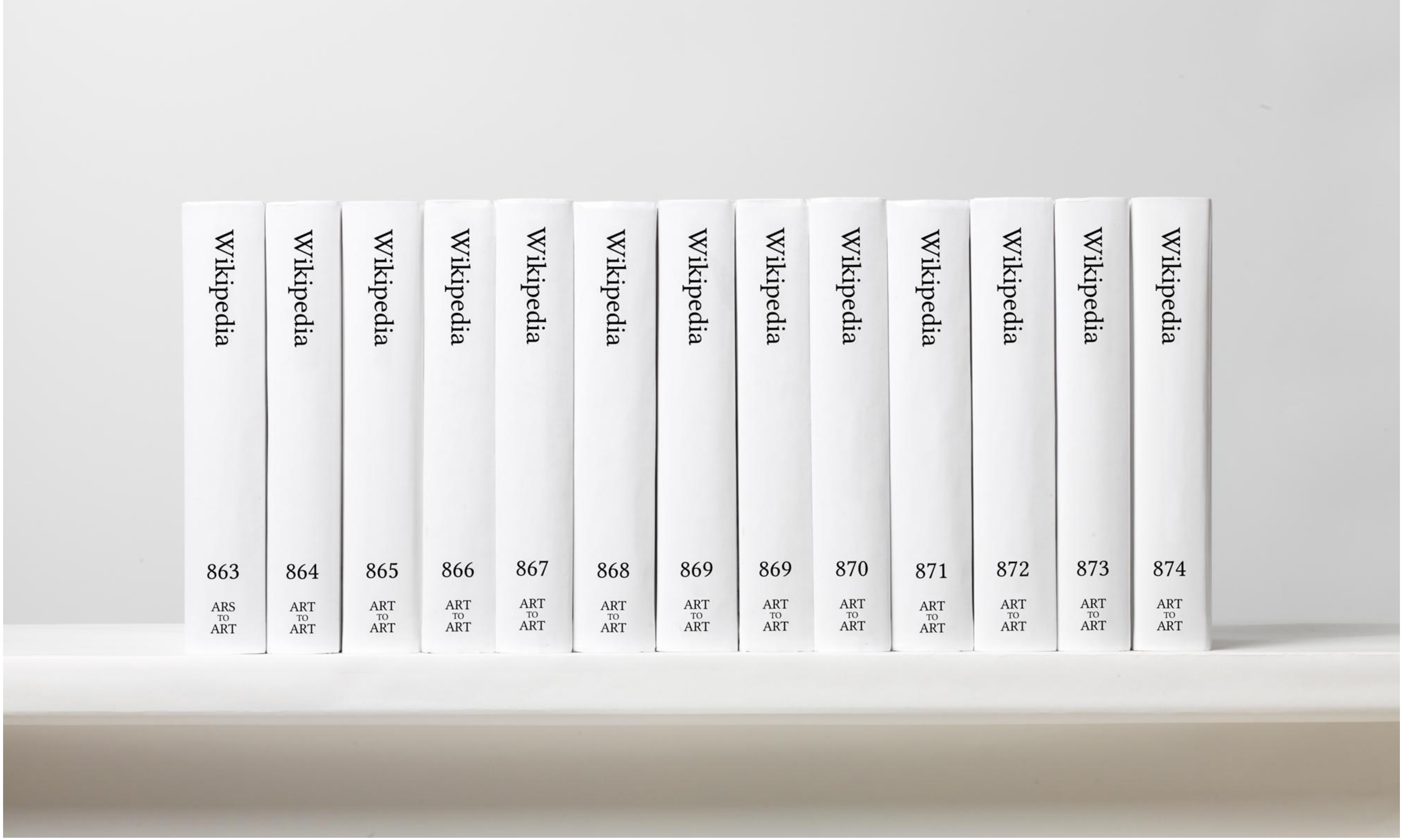


http://en.wikipedia.org/wiki/Wikipedia:Size_in_volume
s

0.01% of Wikipedia



It exists! (sort of)



Own it!

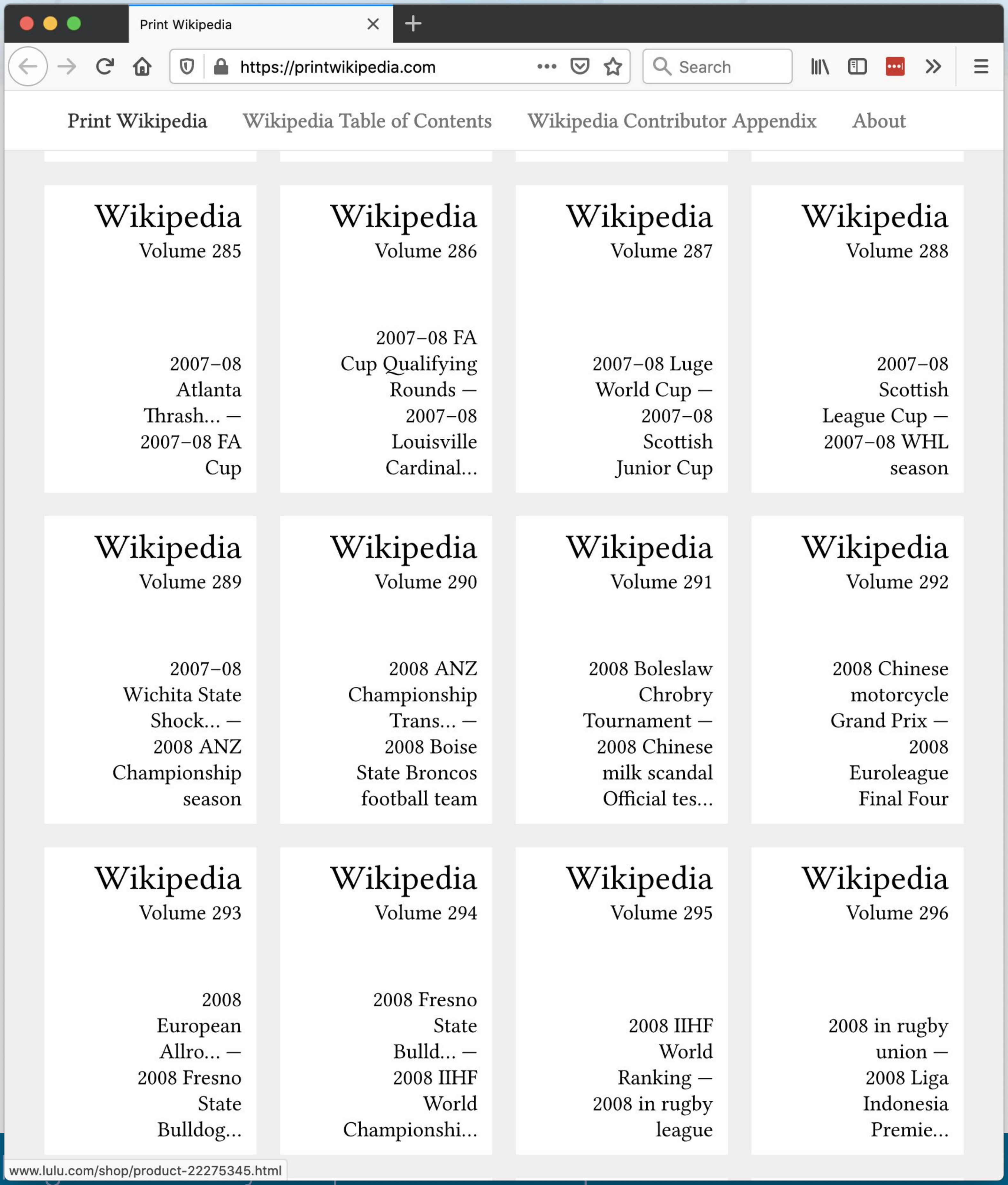
Just $80*!!!

*per volume

7,473 volumes each with 700 pages

Print on demand

https://printwikipedia.com



Print Wikipedia   Wikipedia Table of Contents   Wikipedia Contributor Appendix   About

Wikipedia
Volume 285
2007–08 Atlanta Thrash... — 2007–08 FA Cup

Wikipedia
Volume 286
2007–08 FA Cup Qualifying Rounds — 2007–08 Louisville Cardinal...

Wikipedia
Volume 287
2007–08 Luge World Cup — 2007–08 Scottish Junior Cup

Wikipedia
Volume 288
2007–08 Scottish League Cup — 2007–08 WHL season

Wikipedia
Volume 289
2007–08 Wichita State Shock... — 2008 ANZ Championship season

Wikipedia
Volume 290
2008 ANZ Championship Trans... — 2008 Boise State Broncos football team

Wikipedia
Volume 291
2008 Boleslaw Chrobry Tournament — 2008 Chinese milk scandal Official tes...

Wikipedia
Volume 292
2008 Chinese motorcycle Grand Prix — 2008 Euroleague Final Four

Wikipedia
Volume 293
2008 European Allro... — 2008 Fresno State Bulldog...

Wikipedia
Volume 294
2008 Fresno State Bulld... — 2008 IIHF World Championshi...

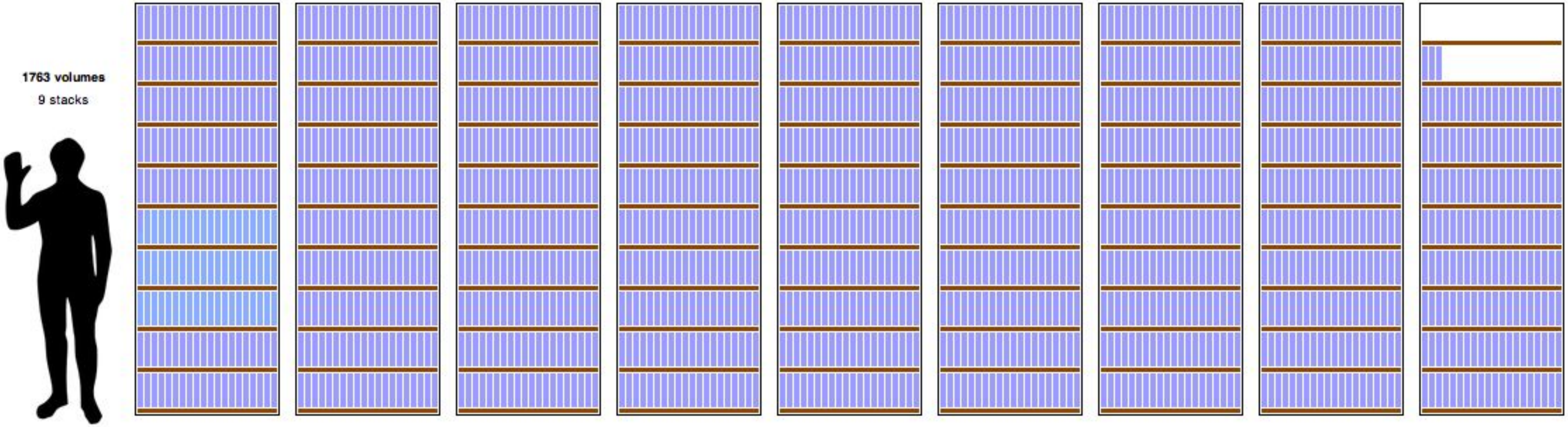
Wikipedia
Volume 295
2008 IIHF World Ranking — 2008 in rugby league

Wikipedia
Volume 296
2008 in rugby union — 2008 Liga Indonesia Premie...

www.lulu.com/shop/product-22275345.html

Big Data in Perspective

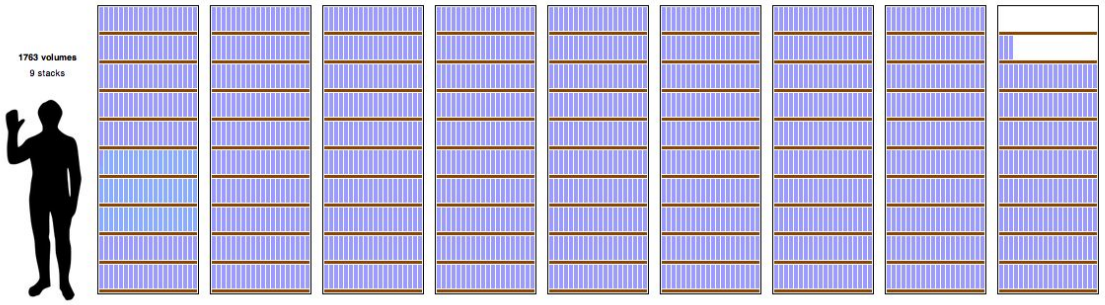
Wikipedia - 5TB of text



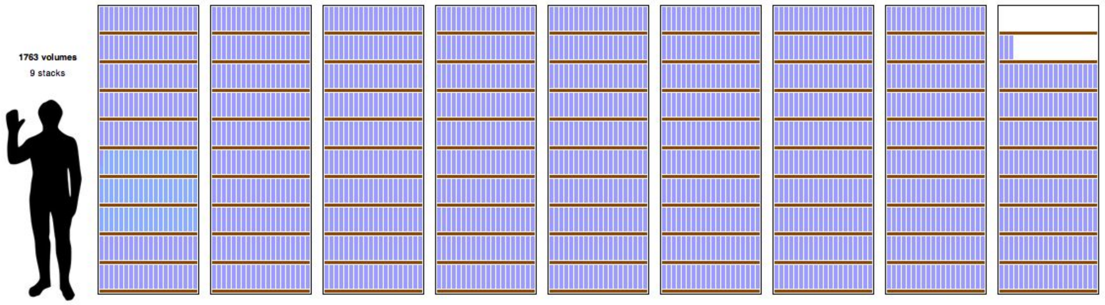
Facebook - ???

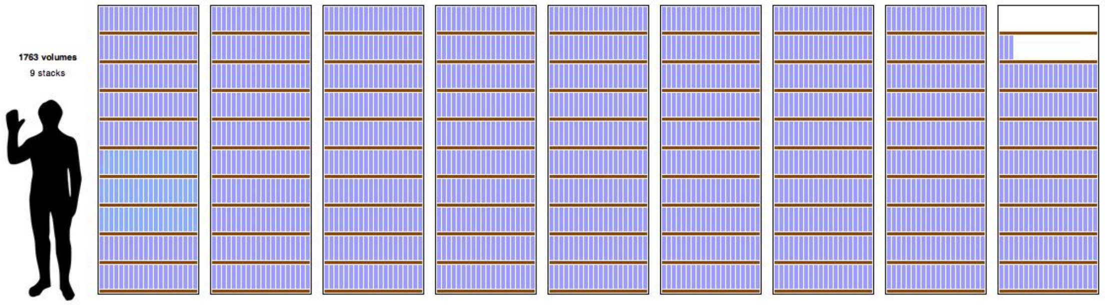
Big Data in Perspective

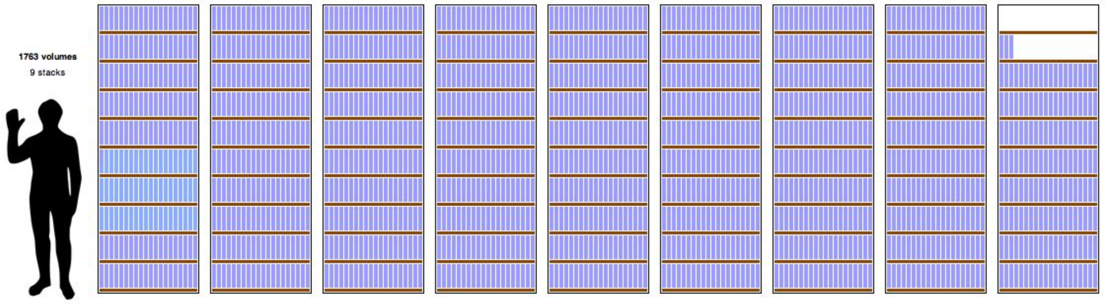
Wikipedia - 5TB of text

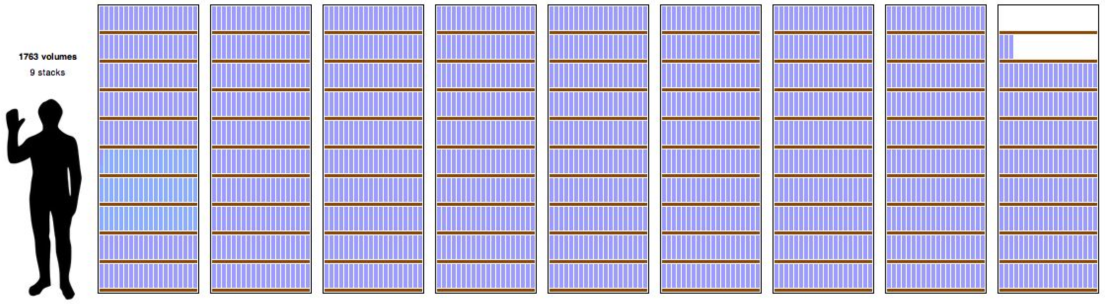


Facebook - 20TB of photos added *each week*

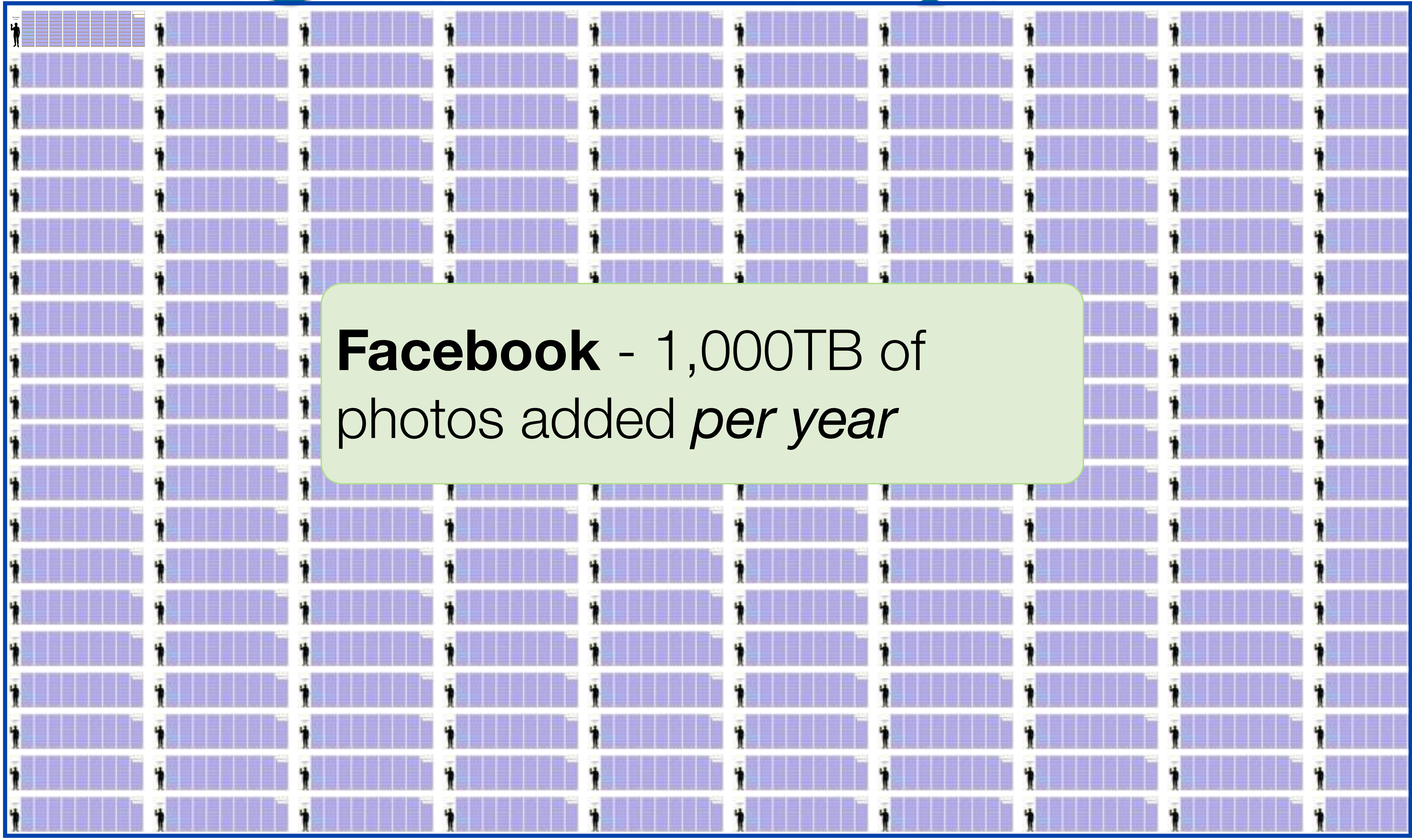








Big Data in Perspective



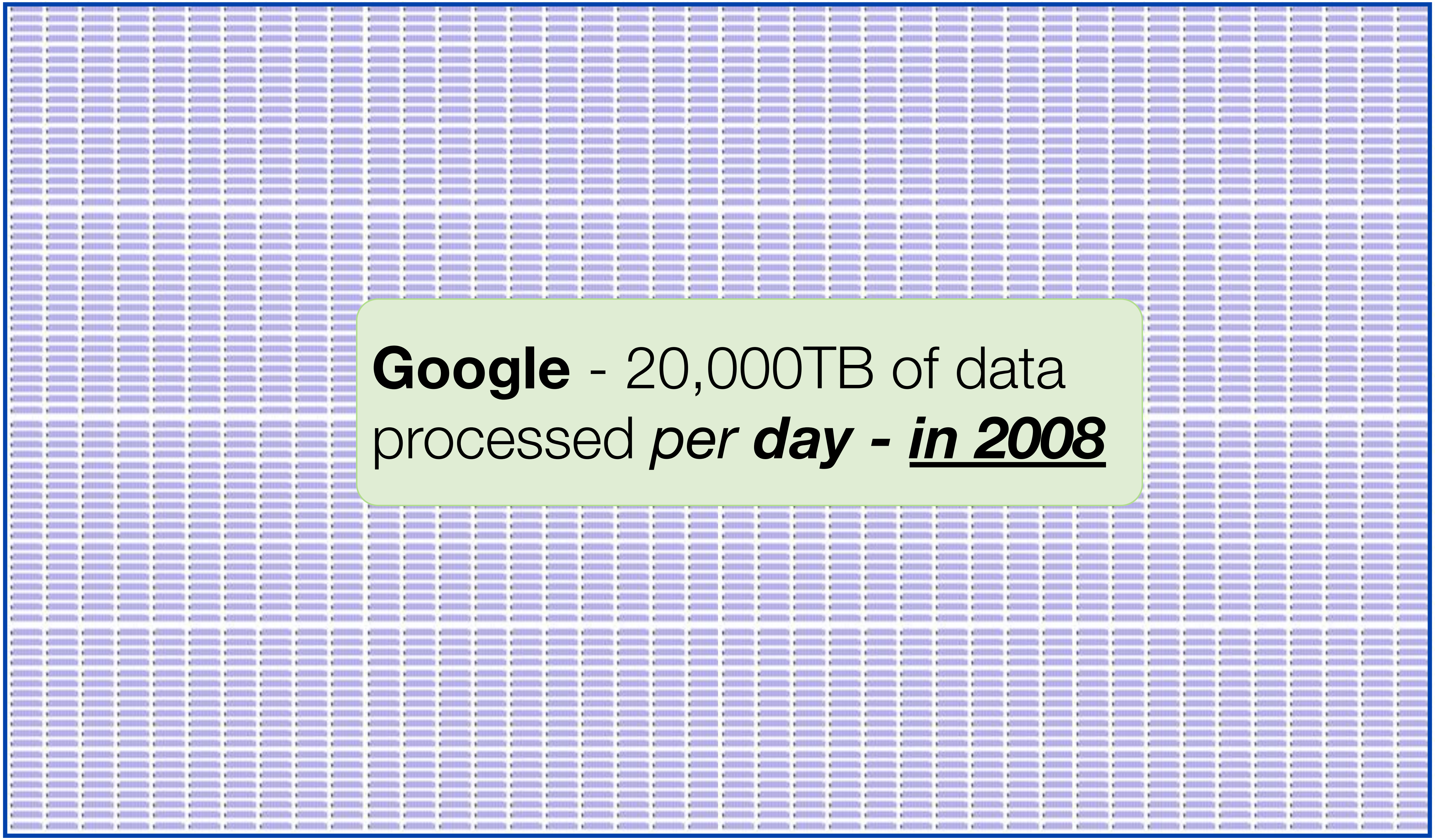
Facebook - 1,000TB of photos added *per year*

Big Data in Perspective

Facebook - 1,000TB of photos added *per year*

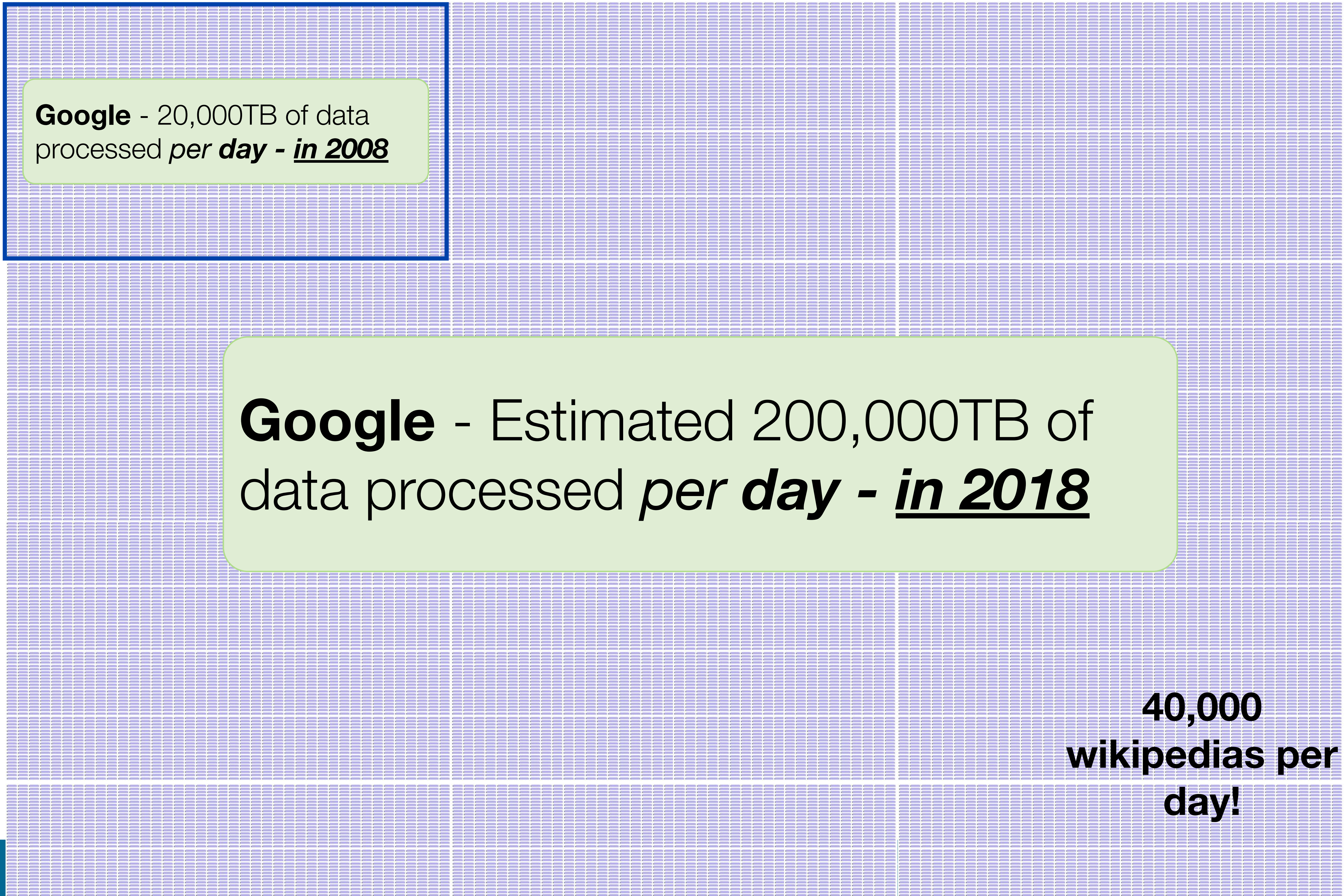
Google - 20,000TB of data processed *per **day***

Big Data in Perspective



Google - 20,000TB of data processed *per* ***day*** ***- <u>in 2008</u>***

Big Data in Perspective



Google - 20,000TB of data processed *per **day** - **<u>in 2008</u>***

Google - Estimated 200,000TB of data processed *per **day** - **<u>in 2018</u>***

40,000 wikipedias per day!

How can google process *so much* information *so quickly*?

Processing Data Quickly

1. 3 + 6 = ?

Buy a **faster** computer

Buy **another** computer

Processing Data in PARALLEL

1.

2

3

4

5

6

7

8

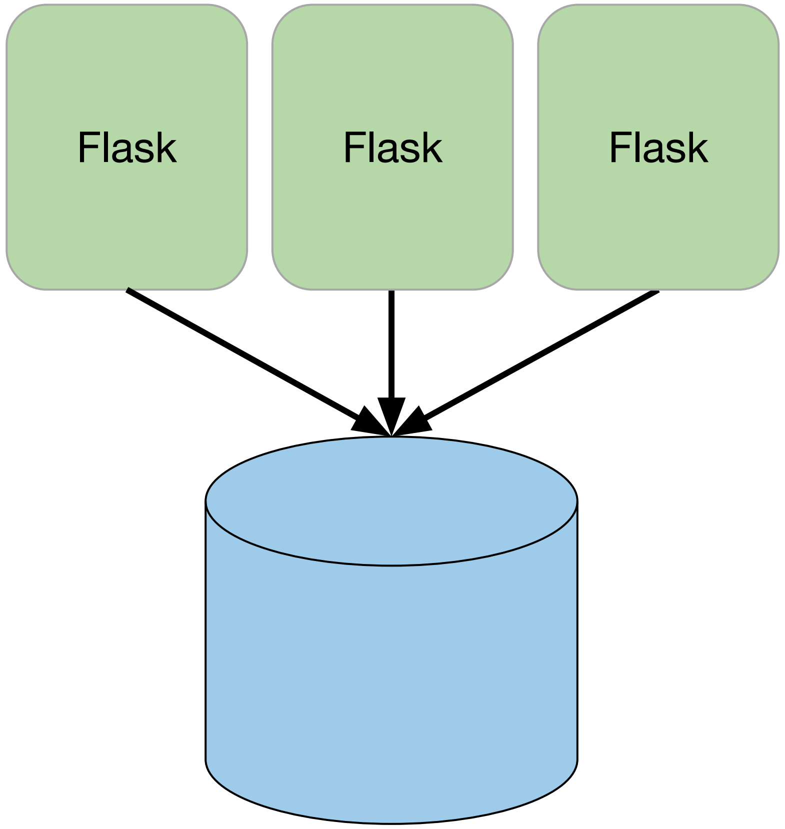
9

10

11

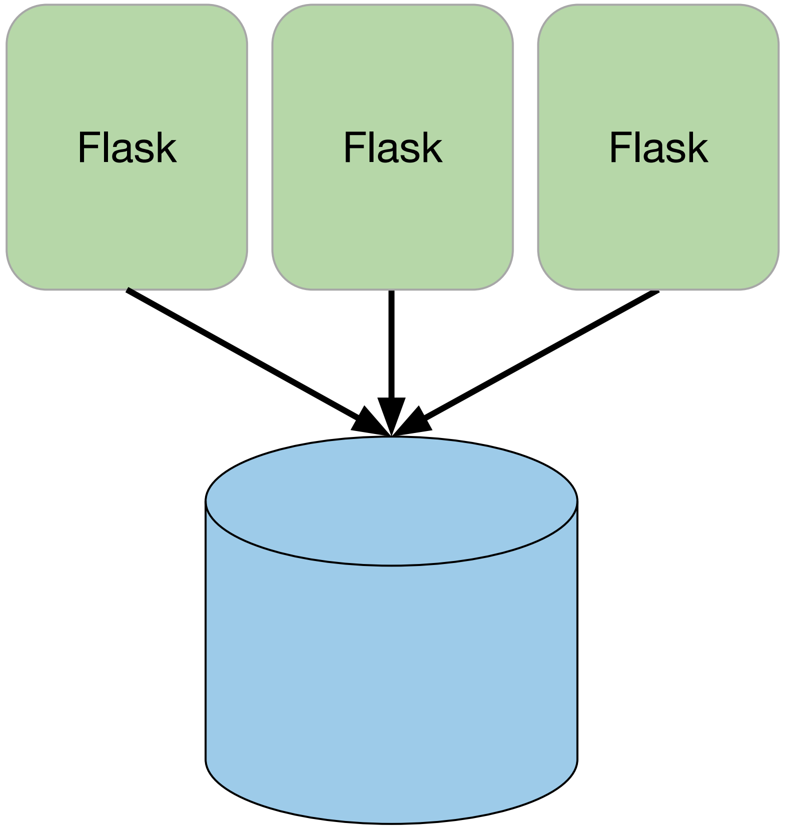
12.

Processing Data via DB

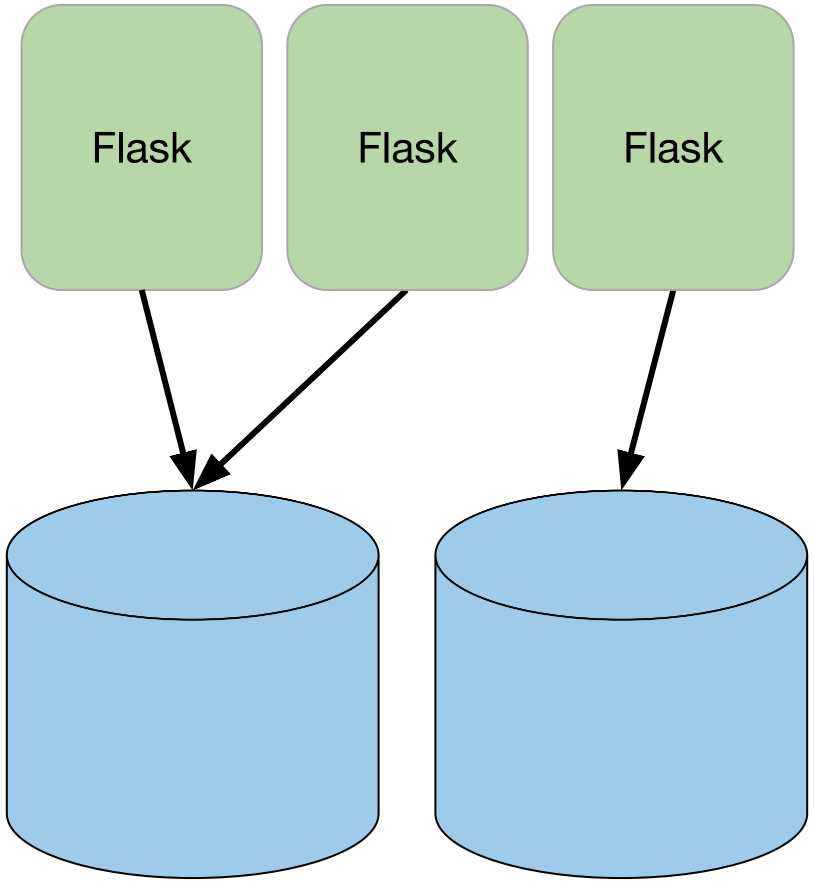


Problems?

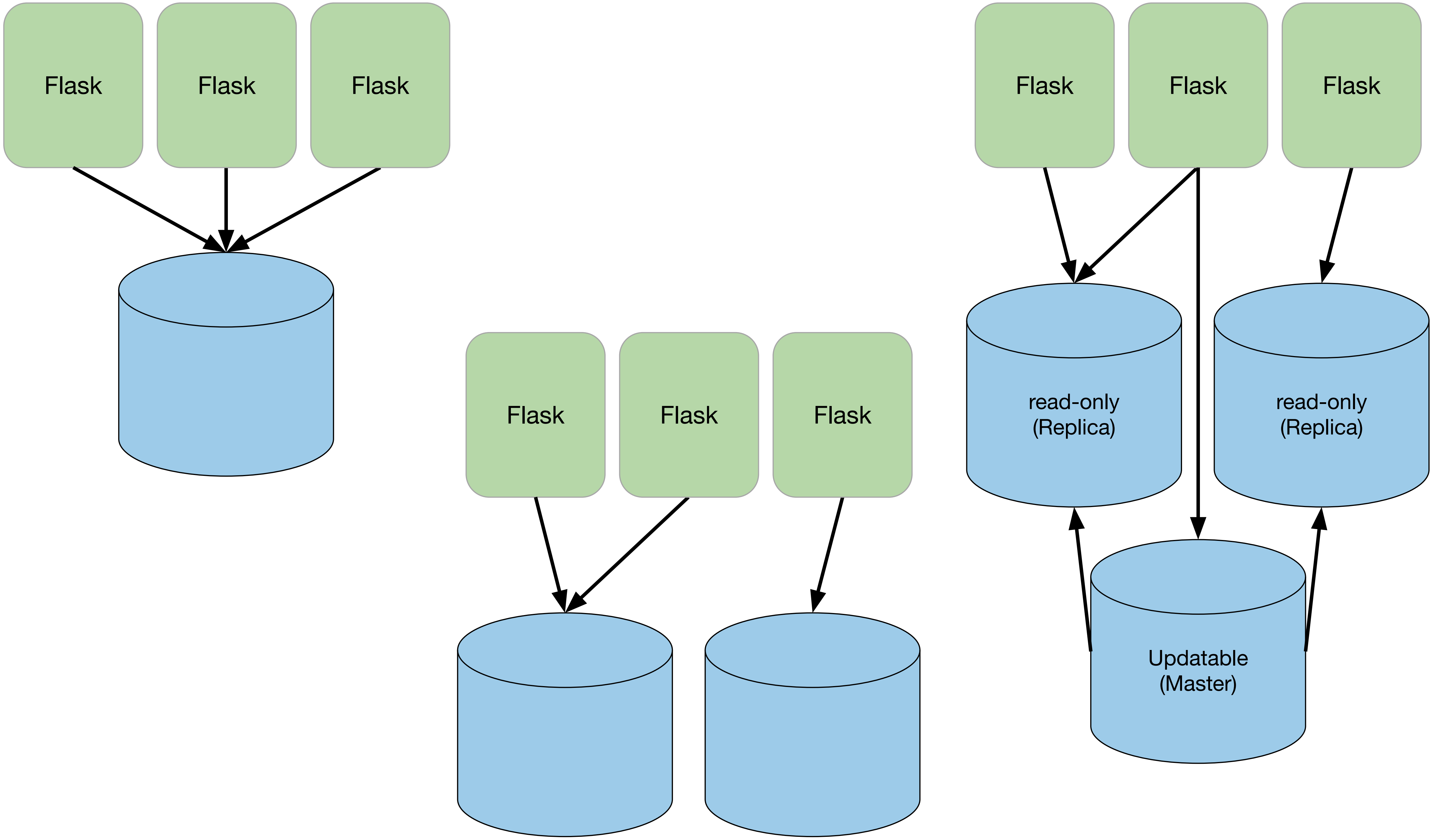
Processing Data via DB



Problems?



Processing Data via DB



Teamwork

Extensive set of rules you must follow

- Communication – early and often
- Respect – responsibility and follow-through
- Planning – everyone on the same page
- Allocating work – clear division of labor/plan
- No procrastination – makes everything else impossible
- Early problem notification – let me help you
- Flexibility – different working styles/schedules
- Professionalism – co-worker not college kid, SEH not dorm
- No type-casts – everyone does interesting work
-

Teamwork

Extensive set of rules you must follow, and **common traps**

- Radio silence
- “do the documentation” “...project management”
- “I’ll just do it all”
- Altering other’s code without their approval
- Their code, your commit
- Falling behind, but not helping a back-up plan
  - “I’ll have it done tomorrow…”
- Gossip
- Asserting organization – need consensus on, e.g. when to meet

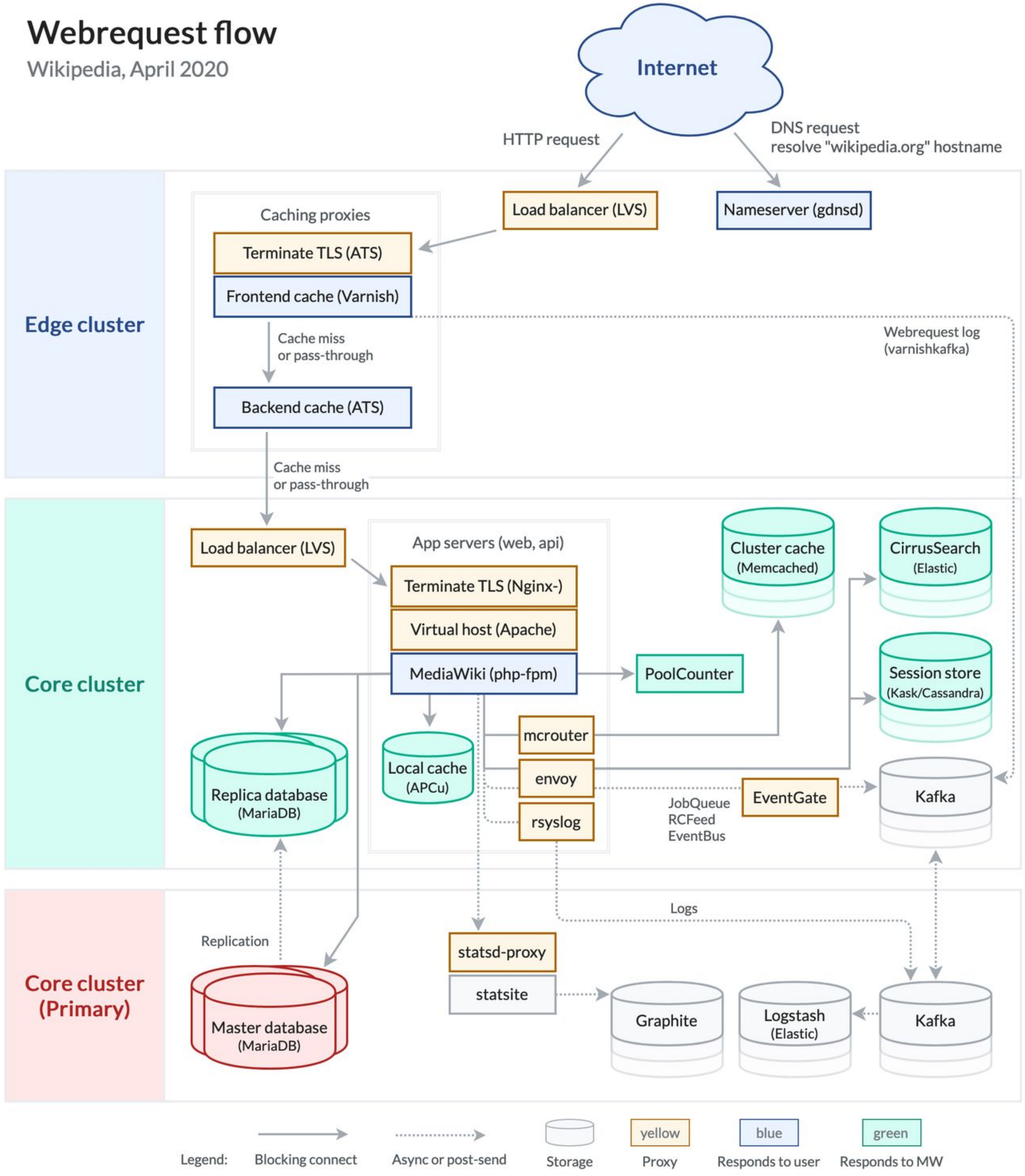
`

What problems did we hit?

How could we optimize the process?

What things can't we prevent?

How does even Wikipedia work?



Webrequest flow
Wikipedia, April 2020

Internet
HTTP request
DNS request
resolve "wikipedia.org" hostname

Edge cluster
Load balancer (LVS)
Nameserver (gdnsd)
Caching proxies
Terminate TLS (ATS)
Frontend cache (Varnish)
Cache miss
or pass-through
Backend cache (ATS)
Webrequest log
(varnishkafka)
Cache miss
or pass-through

Core cluster
Load balancer (LVS)
App servers (web, api)
Terminate TLS (Nginx-)
Virtual host (Apache)
MediaWiki (php-fpm)
PoolCounter
Cluster cache
(Memcached)
CirrusSearch
(Elastic)
Session store
(Kask/Cassandra)
mcrouter
envoy
rsyslog
Local cache
(APCu)
Replica database
(MariaDB)
JobQueue
RCFeed
EventBus
EventGate
Kafka

Core cluster
(Primary)
Replication
Logs
statsd-proxy
statsite
Master database
(MariaDB)
Graphite
Logstash
(Elastic)
Kafka

Legend: Blocking connect | Async or post-send | Storage | yellow Proxy | blue Responds to user | green Responds to MW

https://meta.wikimedia.org/wiki/Wikimedia_servers
https://grafana.wikimedia.org/