

---

THE GEORGE WASHINGTON UNIVERSITY

---

WASHINGTON, DC

# 8. The Cloud & Data

CSCI 2541 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

Wednesday: Exam review? Lab on session programming

Next Tuesday 3/8: Shopping cart due!

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

Office hours:

- Monday: Ethan 2-6pm and Deep (zoom) 2:30-3:30
- Tuesday: Alex 1-3pm and Cat 7-9pm
- Thursday: Jett 1-3pm
- Friday: Cat 1:30-3:30pm
- Saturday: Alex 12:30-2pm

Me: Tuesday/  
Thursday 7:30-8:30  
(zoom)

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

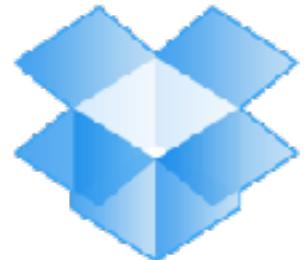
# Cloudy Buzz



Google™ Docs



iCloud



Dropbox flickr™

Fast!

XBOX  
LIVE.



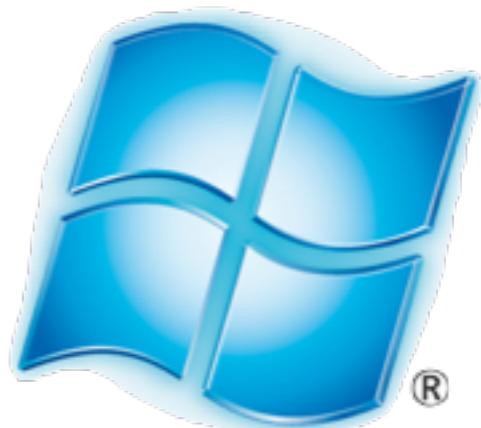
To the  
cloud!



amazon  
web services™

Free\*!

Powerful!



# What *is* a cloud?

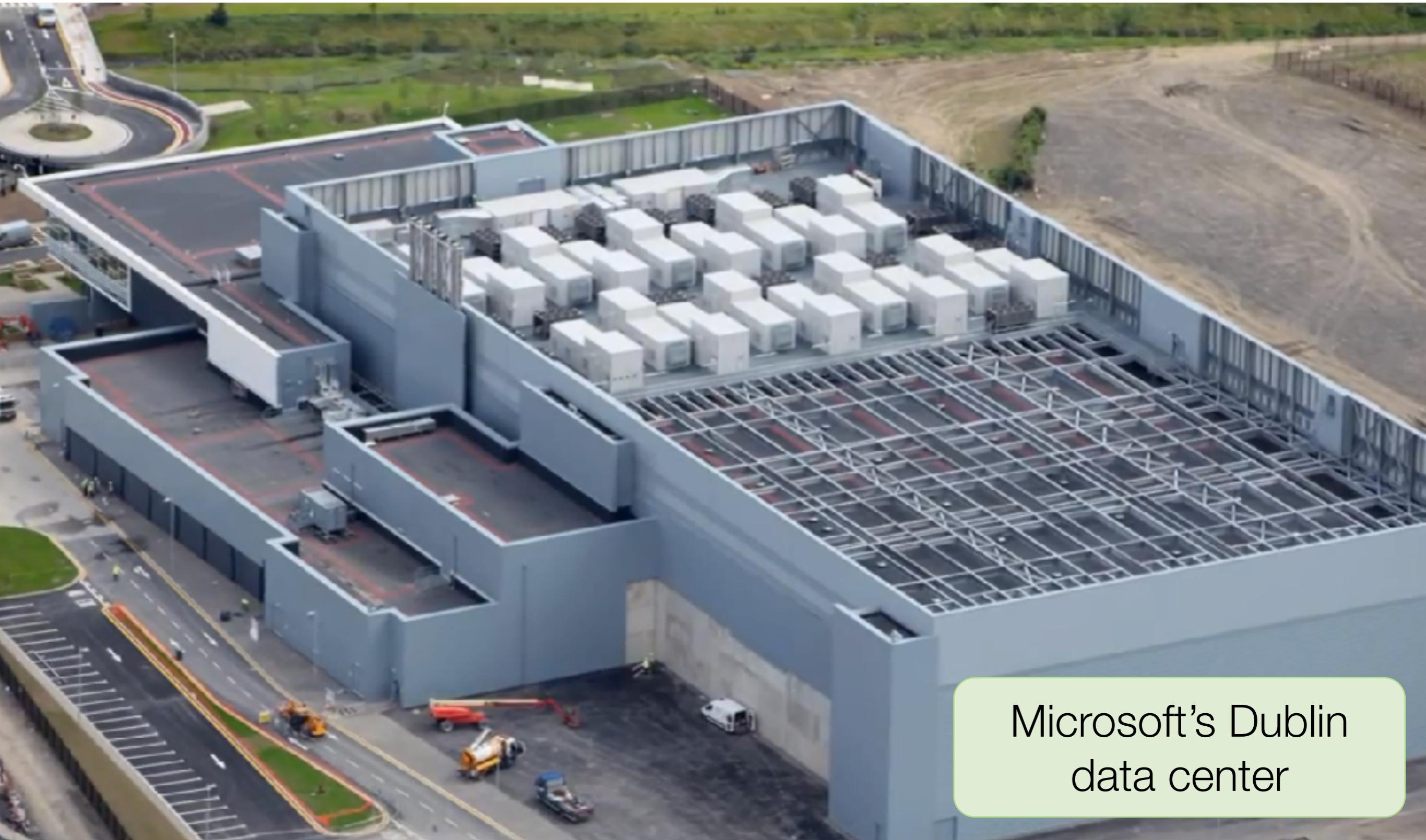
<spoiler alert>

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...



## Undersea Cables

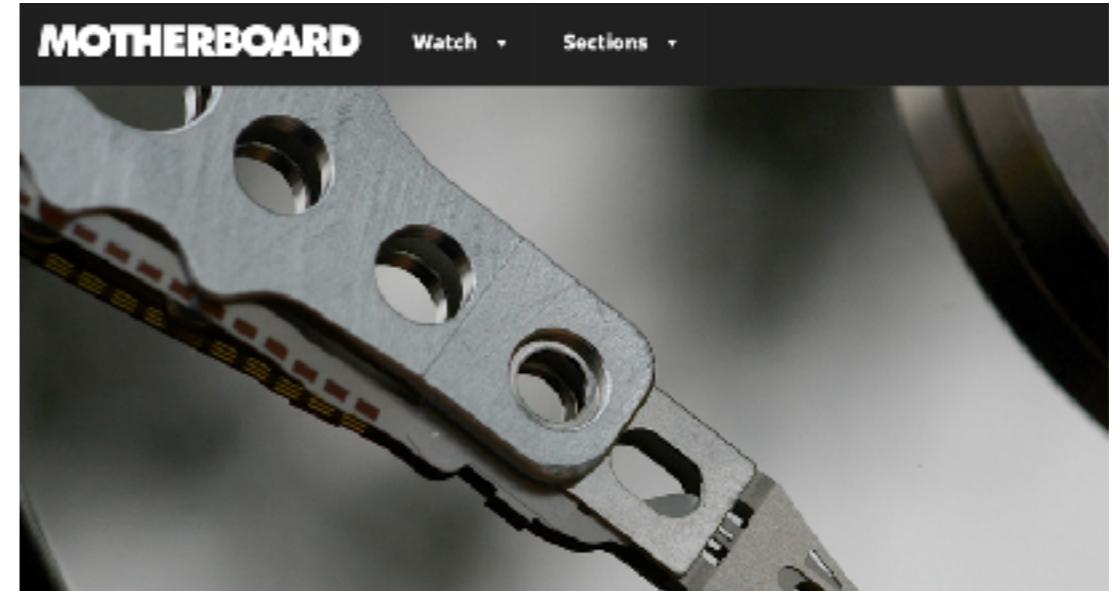
- Connect all continents except Antarctica
- First deployed in 1850s



# ...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

# 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

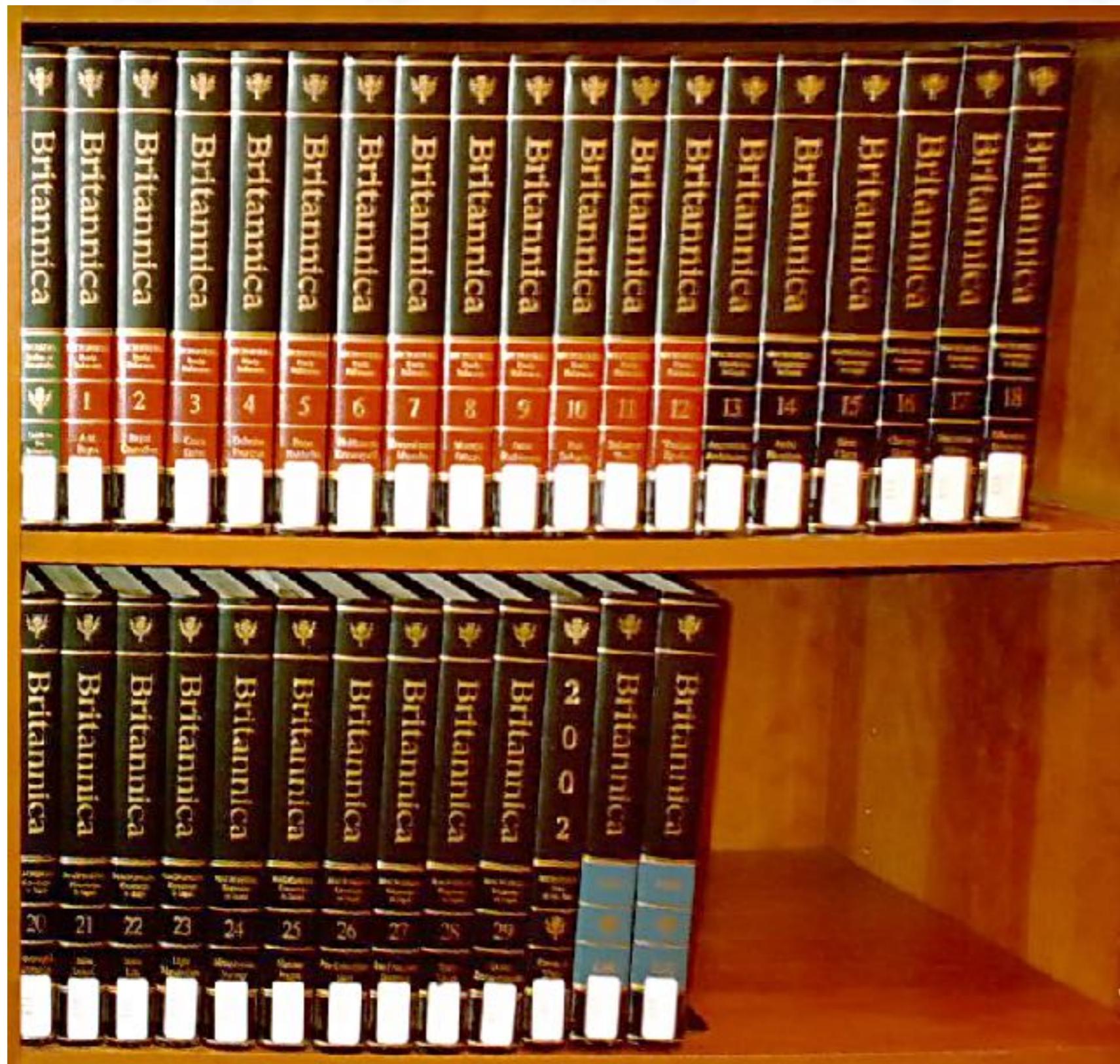


Shared computing infrastructure that developers, companies, and students can easily get access to

- ~1.4 million servers in 2014

Why do we need all of  
this physical  
infrastructure?

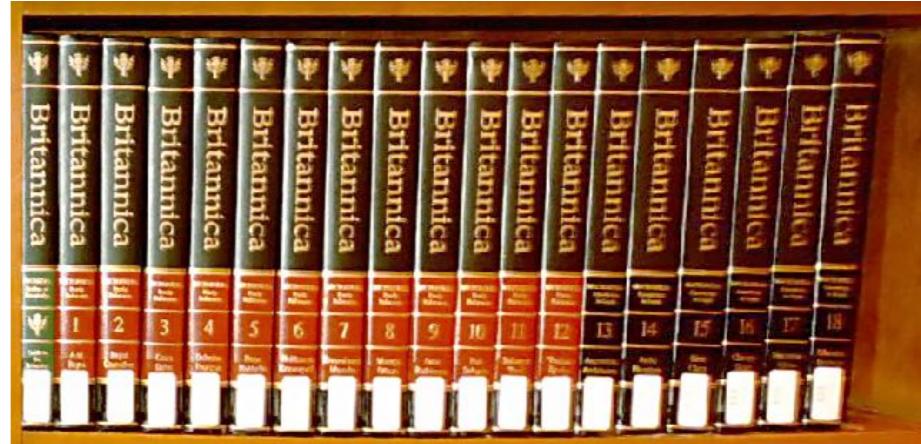
# What is this???



# 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	$1024 \times 1024 =$ $\sim 1,000,000$
Giga	Billion	$1024 \times 1024 \times 1024 =$ $\sim 1,000,000,000$
Tera	Trillion	$1024 \times 1024 \times 1024 \times 1024 =$ $\sim 1,000,000,000,000$

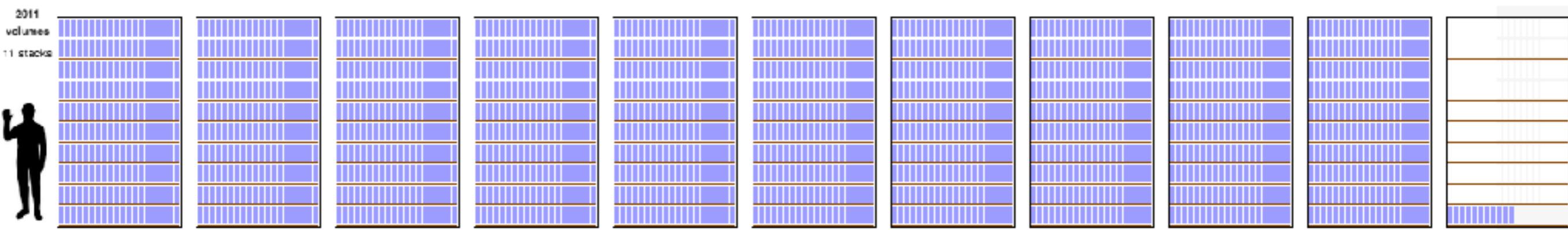
200 photos vs 1.4 million photos

# Encyclopedias

## Wikipedia... in print

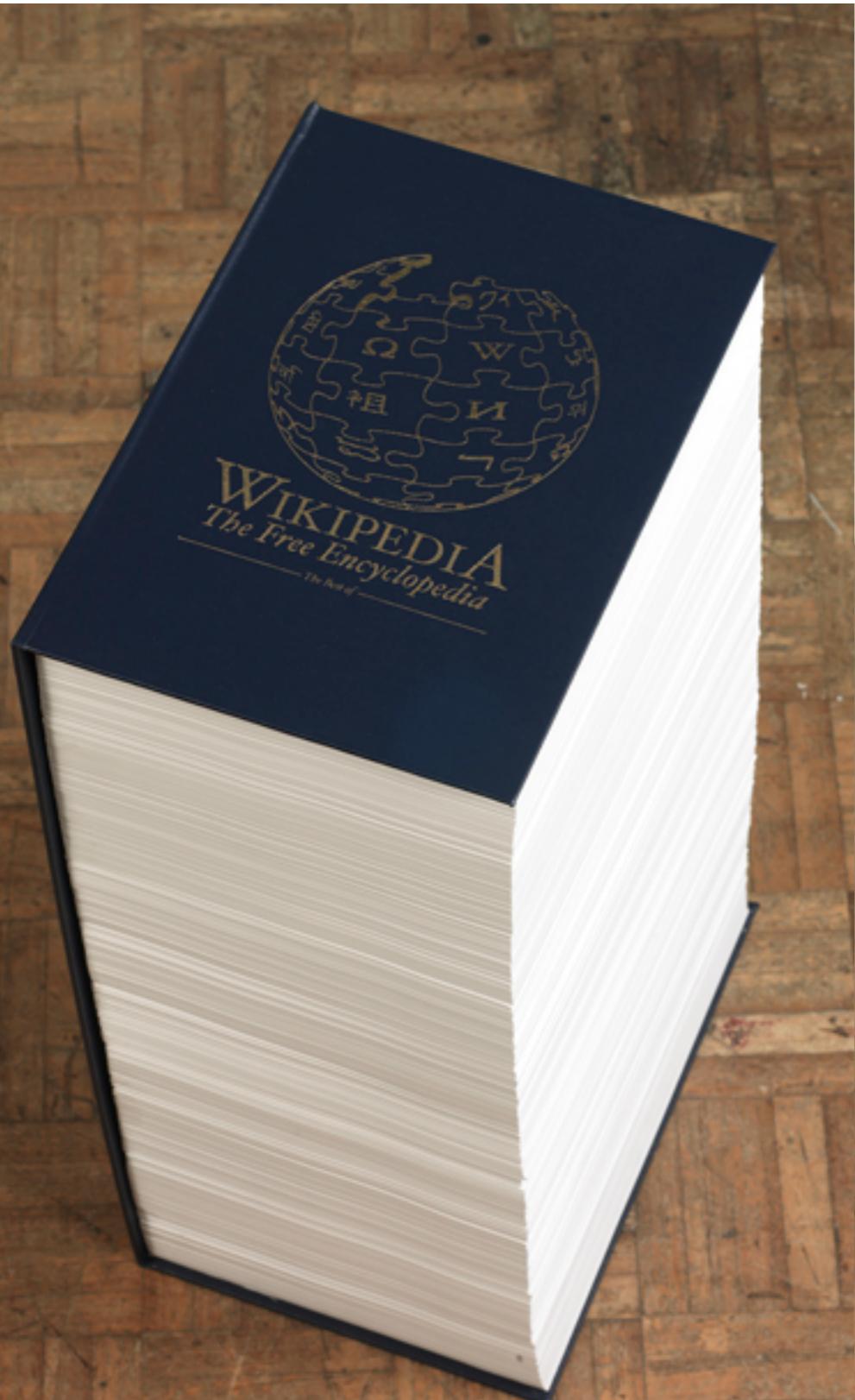
- ~~1,763 volumes~~
- (no, this does not exist)

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



[http://en.wikipedia.org/wiki/Wikipedia:Size\\_in\\_volumes](http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes)

# 0.01% of Wikipedia



# It exists! (sort of)

863	864	865	866	867	868	869	870	871	872	873
ARS TO ART	ART TO ART									

# Own it!

Just \$80\*!!!

\*per volume

7,473 volumes  
each with 700  
pages

Print on demand

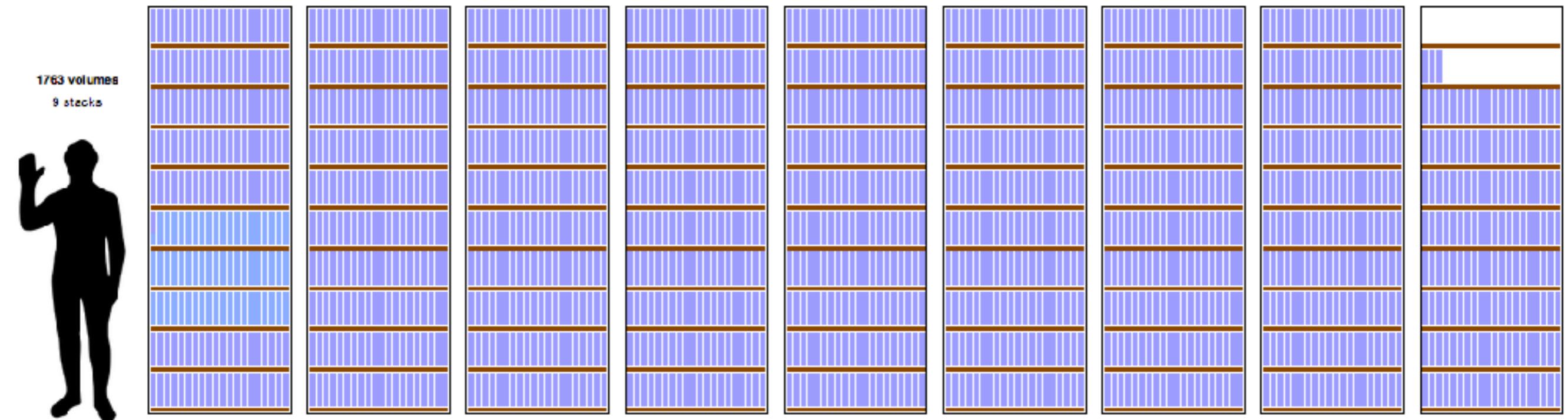
<https://printwikipedia.com>

The screenshot shows a web browser window with a title bar "Print Wikipedia". The address bar displays the URL "https://printwikipedia.com". The main content area features a grid of 12 cards, each representing a different Wikipedia volume. The volumes are arranged in three rows of four. Each card includes the word "Wikipedia" and "Volume" followed by a number. Below the volume number, there is a brief summary of the contents.

Volume 285	Volume 286	Volume 287	Volume 288
2007–08 Atlanta Thrash... – 2007–08 FA Cup	2007–08 FA Cup Qualifying Rounds – 2007–08 Louisville Cardinal...	2007–08 Luge World Cup – 2007–08 Scottish Junior Cup	2007–08 Scottish League Cup – 2007–08 WHL season
2007–08 Wichita State Shock... – 2008 ANZ Championship season	2008 ANZ Championship Trans... – 2008 Boise State Broncos football team	2008 Boleslaw Chrobry Tournament – 2008 Chinese milk scandal Official tes...	2008 Chinese motorcycle Grand Prix – 2008 Euroleague Final Four
2008 European Allro... – 2008 Fresno State Bulldog...	2008 Fresno State Bulld... – 2008 IIHF World Ranking – 2008 in rugby league Championshi...	2008 IIHF World Ranking – 2008 in rugby league	2008 in rugby union – 2008 Liga Indonesia Premie...

# Big Data in Perspective

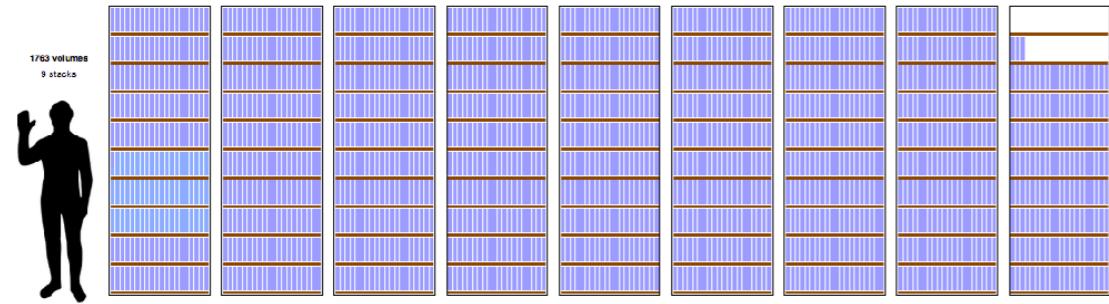
**Wikipedia** - 5TB of text



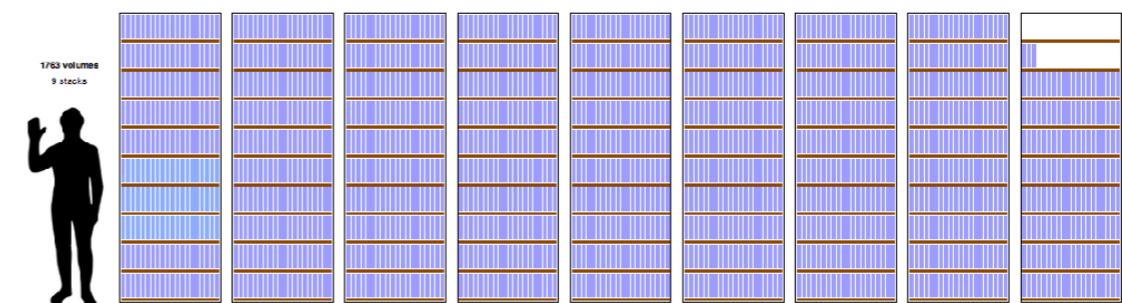
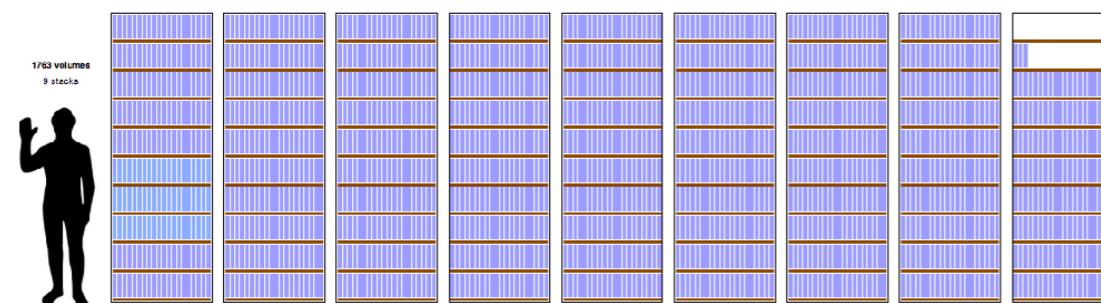
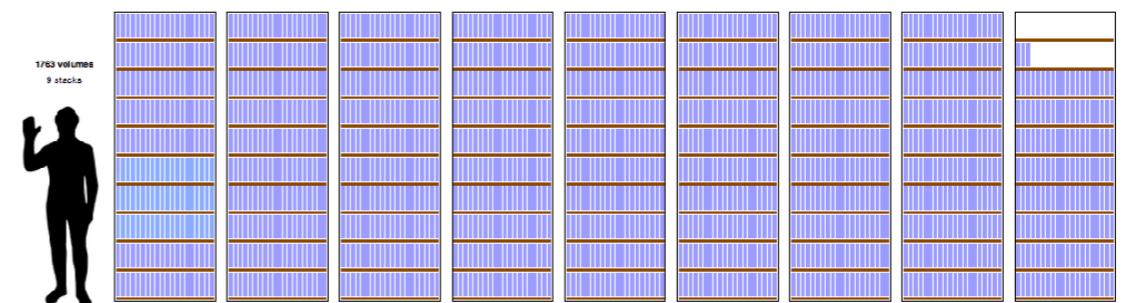
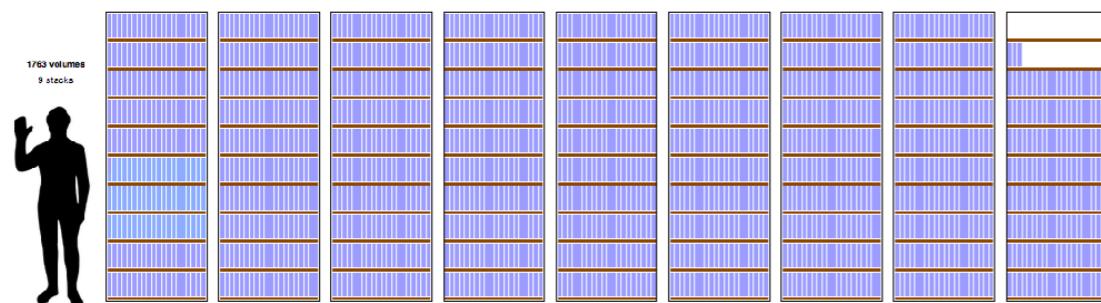
**Facebook** - ???

# Big Data in Perspective

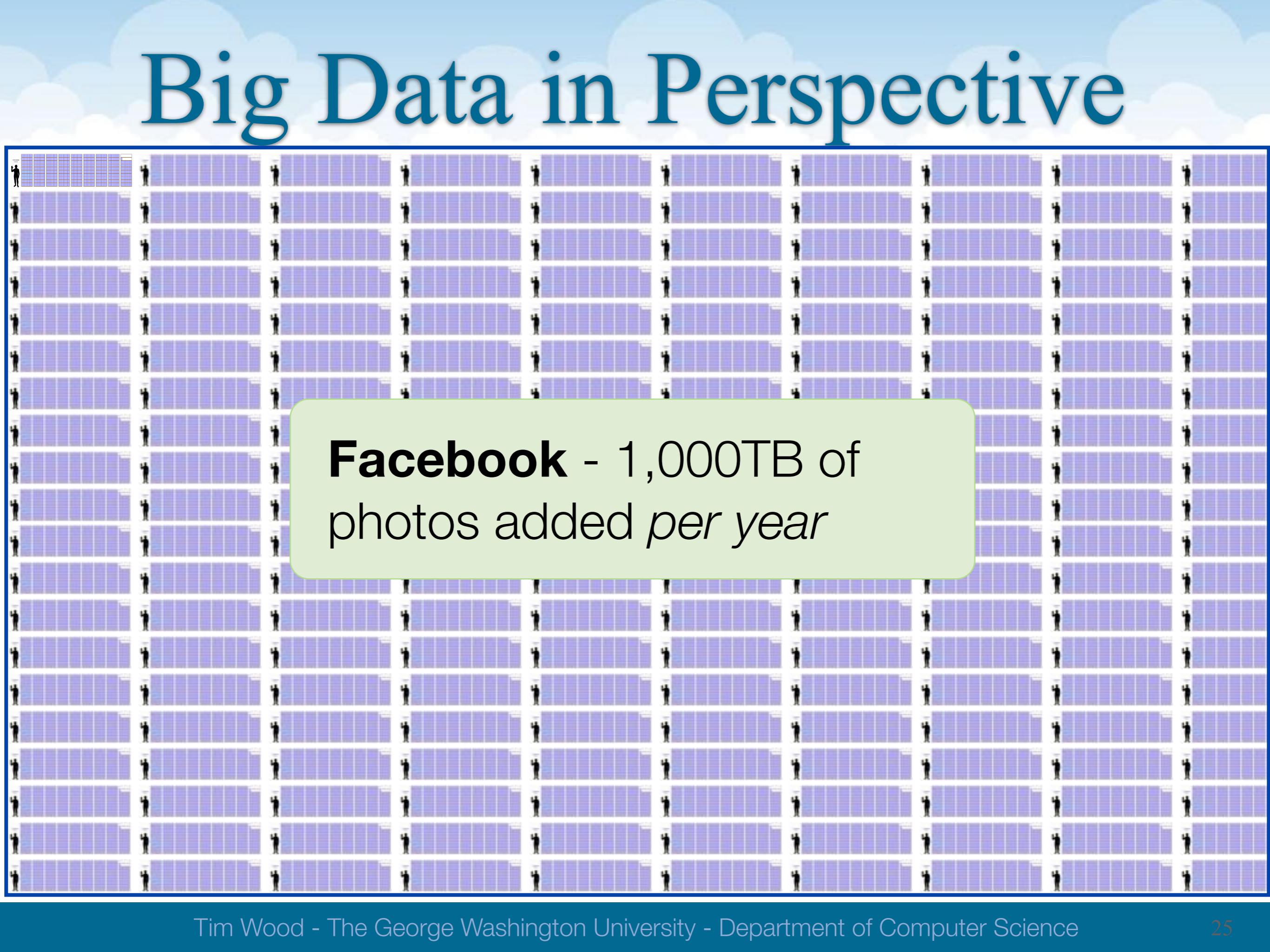
**Wikipedia** - 5TB of text



**Facebook** - 20TB of photos added each week

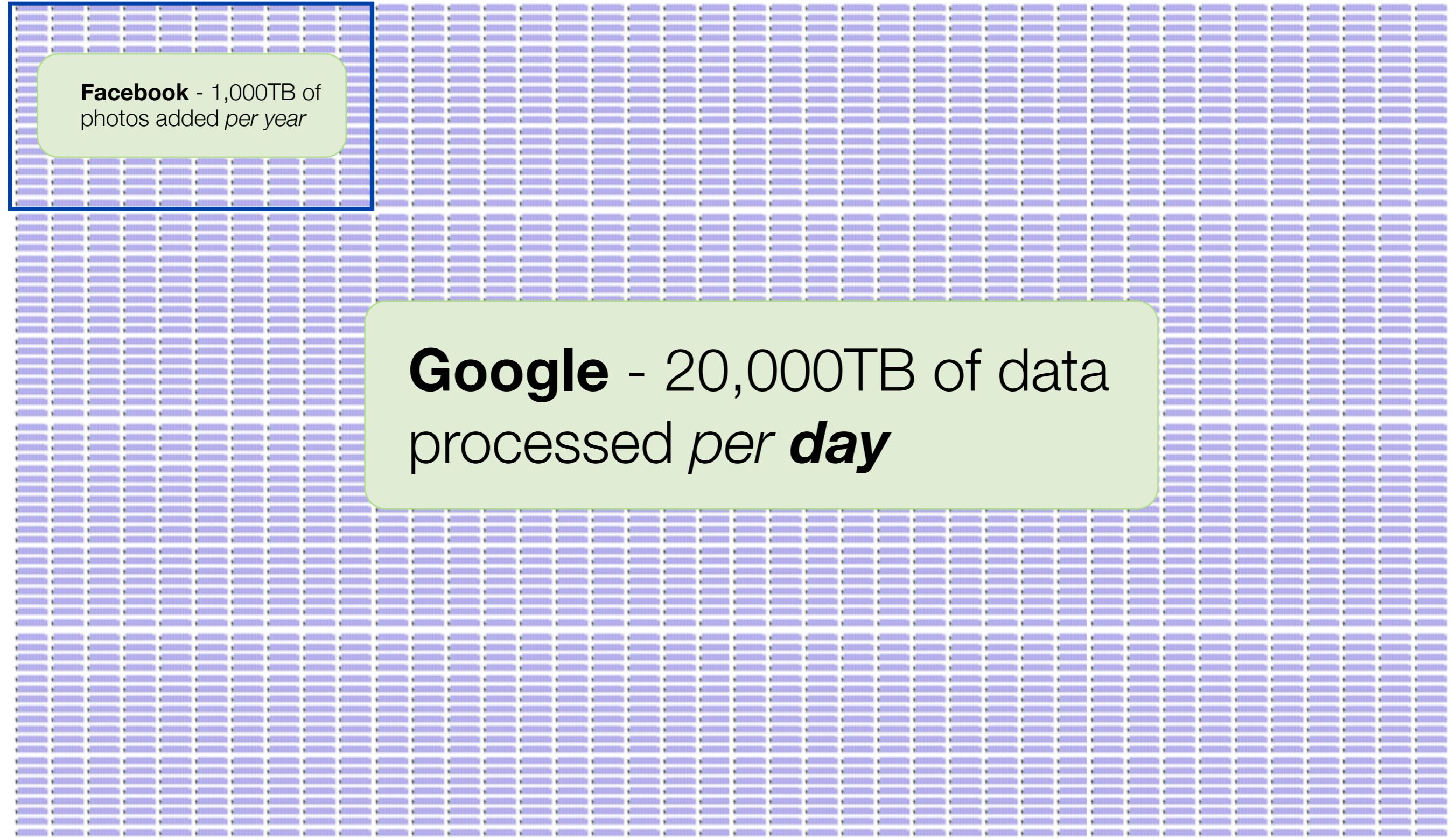


# Big Data in Perspective



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

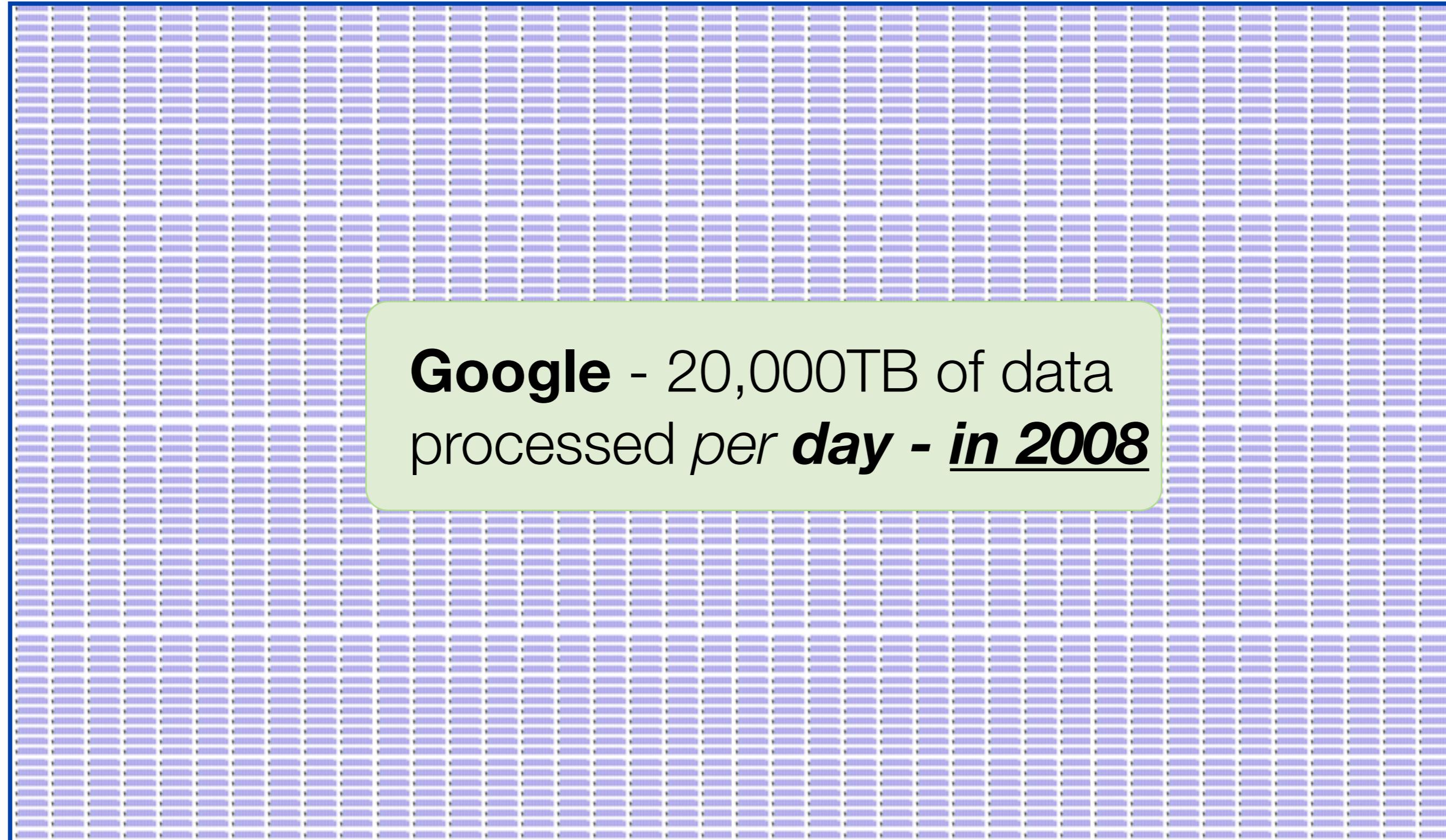
# Big Data in Perspective



**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** - **in 2008**

# Big Data in Perspective

**Google** - 20,000TB of data  
processed *per day - in 2008*

**Google** - Estimated 200,000TB of  
data processed *per day - in 2018*

40,000  
**wikipedias per  
day!**

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

# Processing Data Quickly

$$1. \quad 3 + 6 = ?$$

Buy a **faster** computer

Buy **another** computer

# Processing Data in PARALLEL

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.
- 8.
- 9.
- 10.
- 11.
- 12.

# Let's try it at scale

I have lots of questions I need answered... help me out! Slightly more complicated

$$10 = 3 + A$$

16 questions per page, 10 pages... how long should it take?



What problems did we hit?

How could we optimize the process?

What things can't we prevent?