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

Mobile!

Fast!

Powerful!

Free!*

What *is* a cloud?

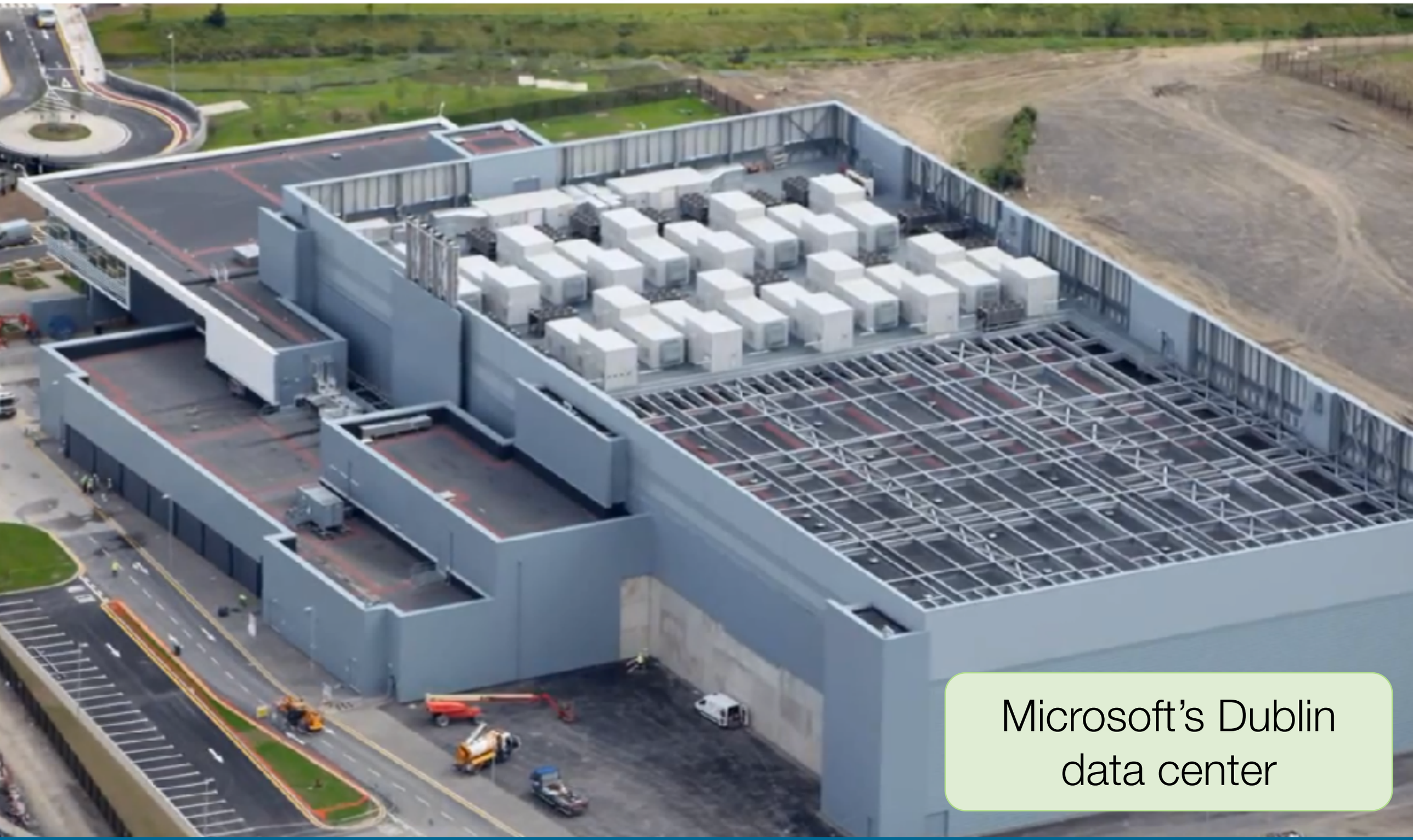
<spoiler alert>

It’s not in the sky

it’s not made of water droplets

</spoiler alert>

Some big buildings…



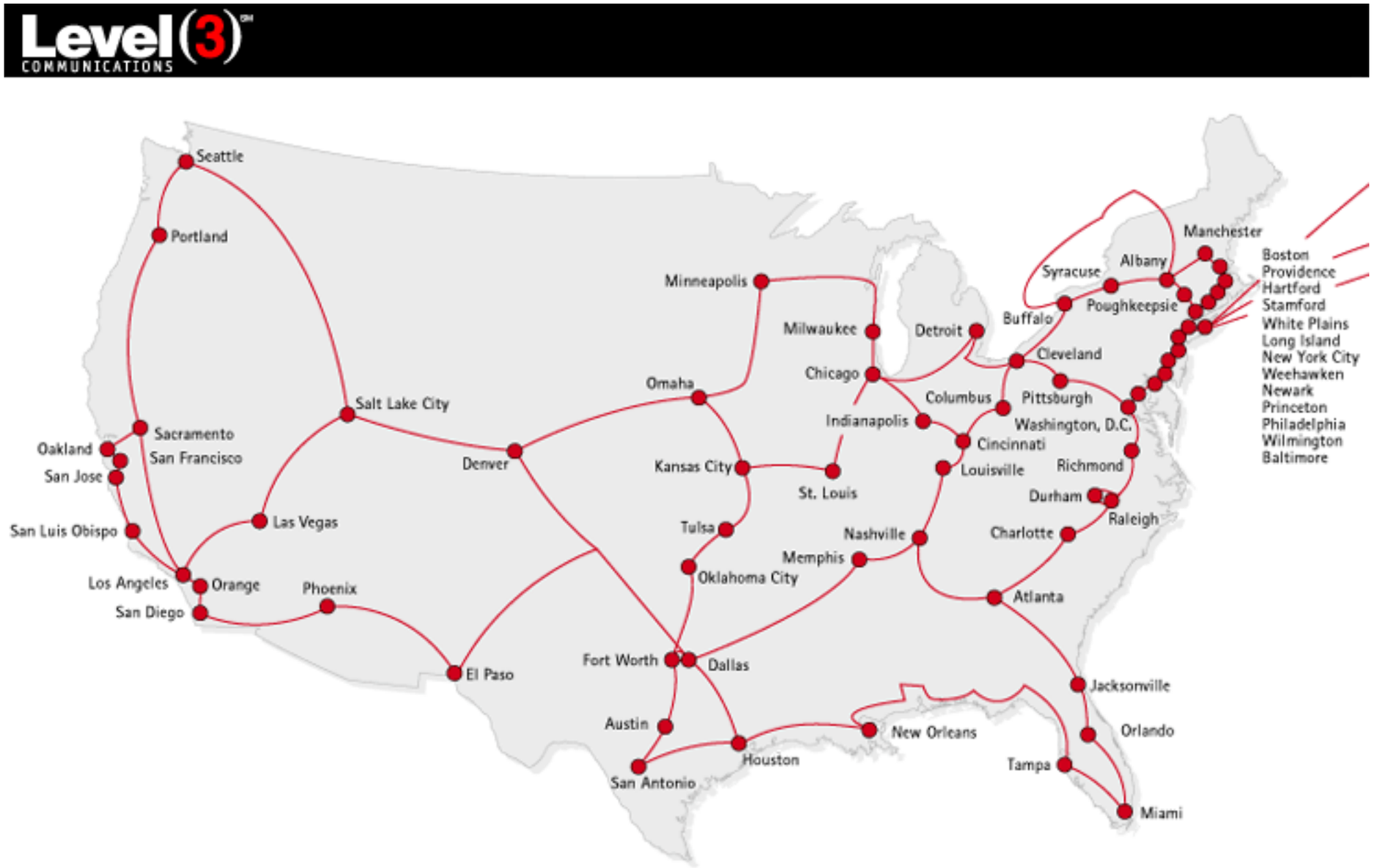
Microsoft's Dublin data center

…and computers…

Giant warehouses

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

…interconnected…



...around the world...

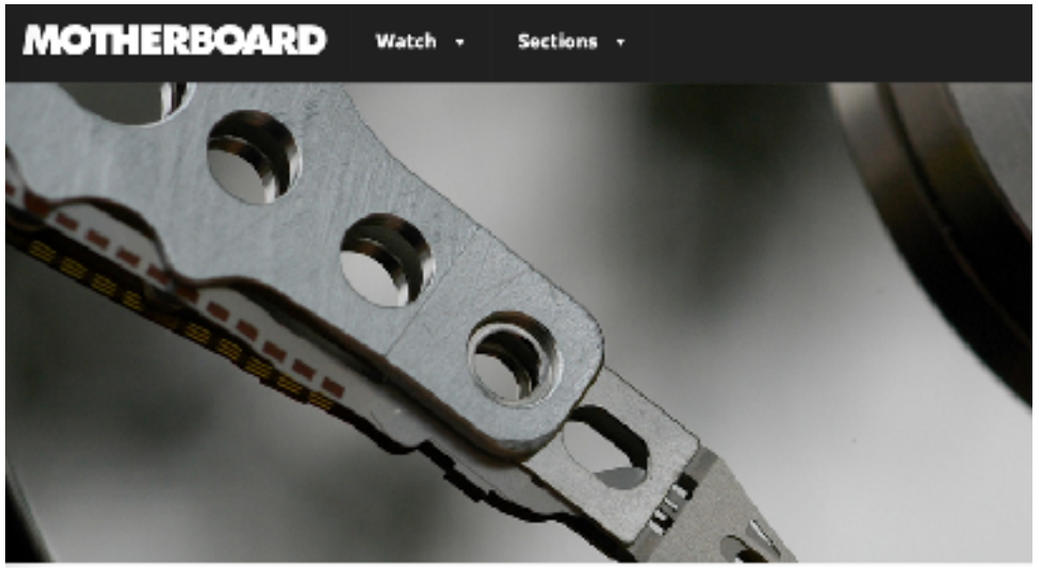
http://submarine-cable-map-2013.telegeography.com/

Undersea Cables

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

…that break a lot.

Lightning causes Amazon outages (2009 and 2011)



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

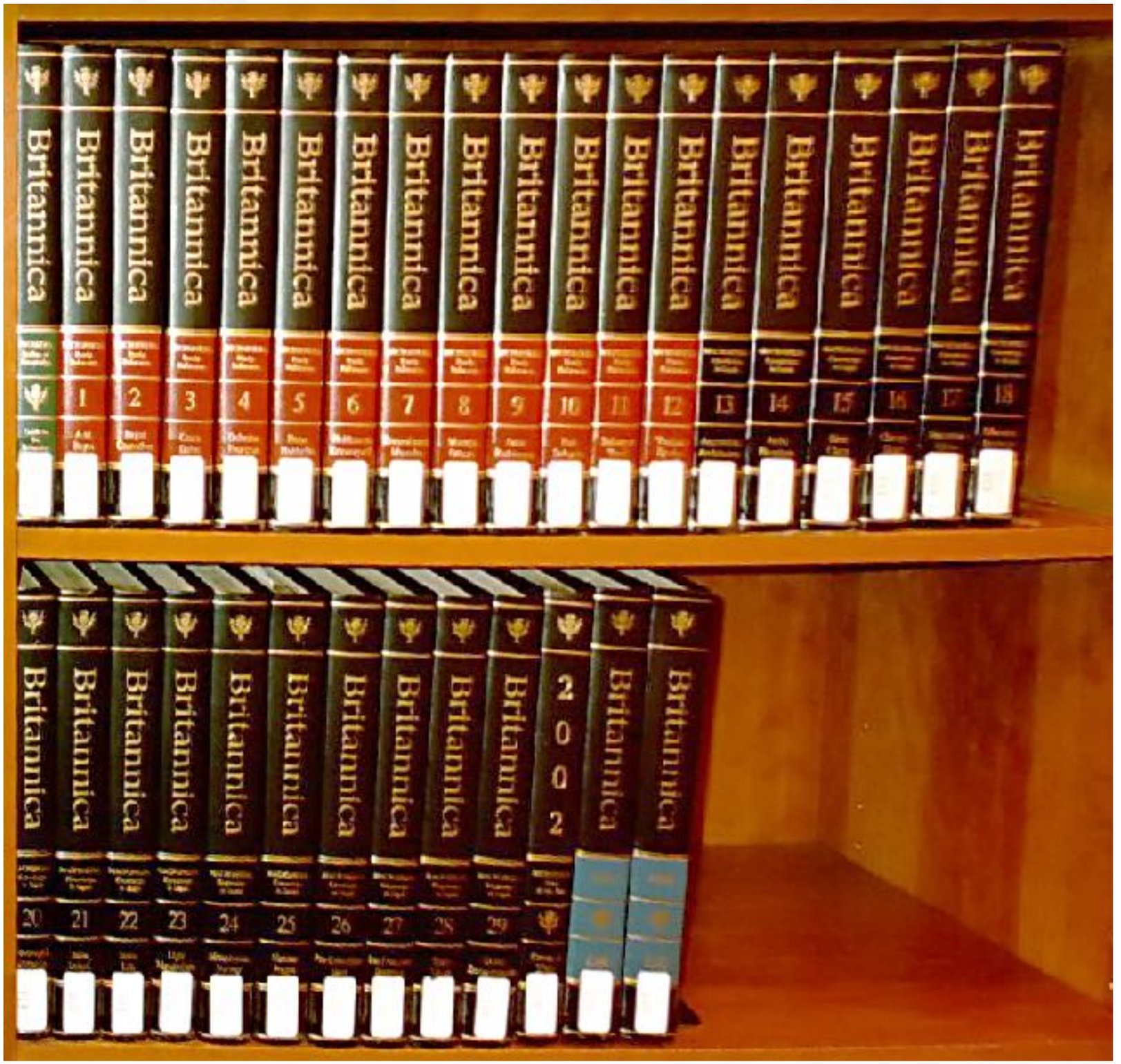
- ~10,000 servers and 200 million users in 2013

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?

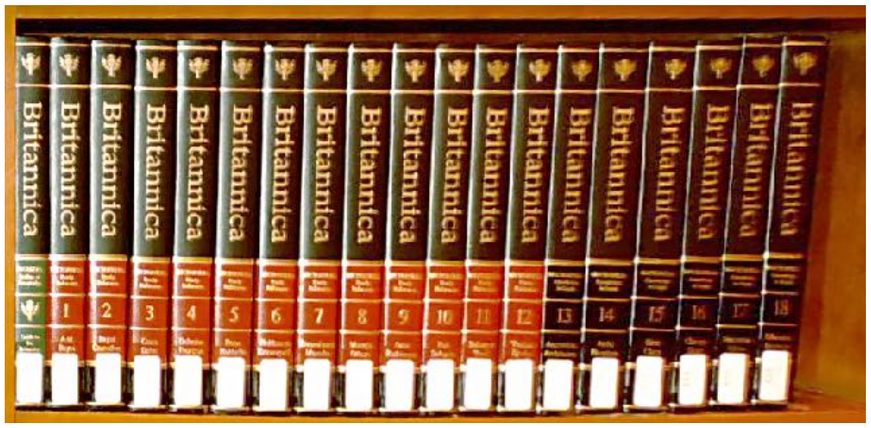
What is this???



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 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**

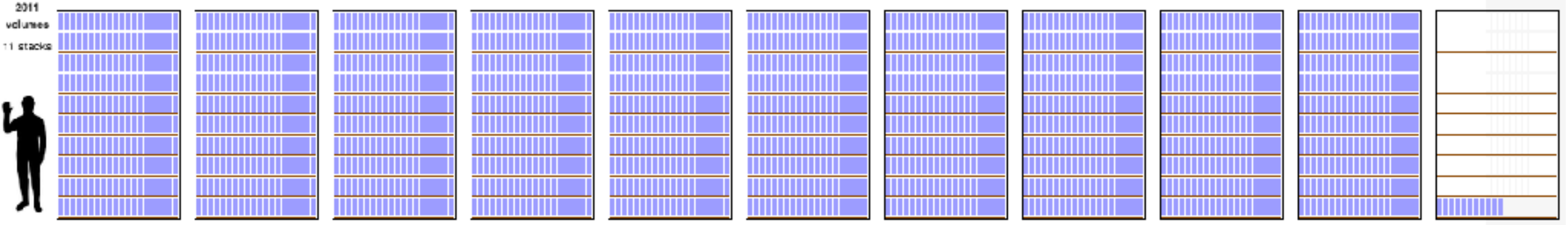
200 photos vs 1.4 million photos

Encyclopedias

Wikipedia... in print

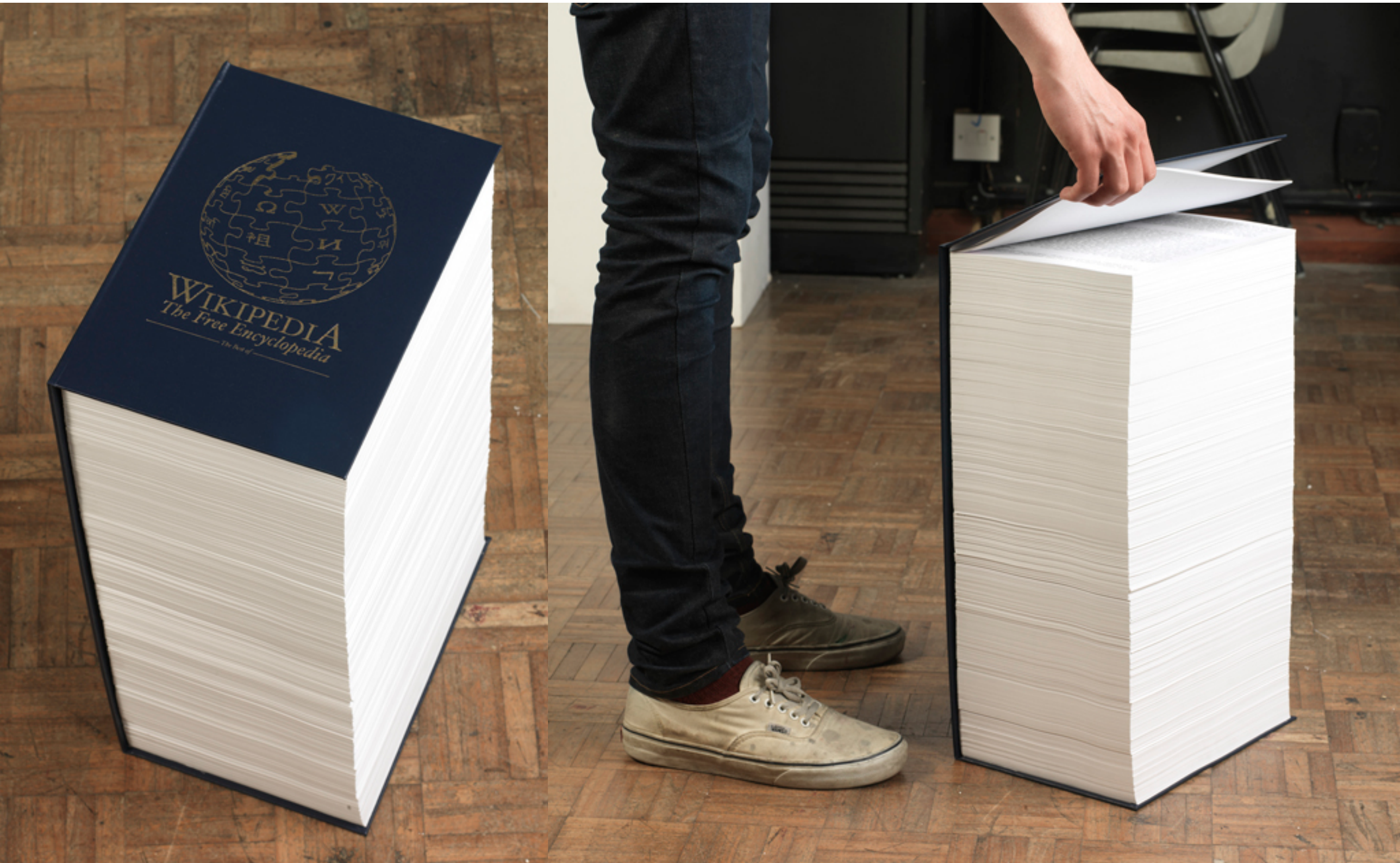
- ~~1,763 volumes~~
- (no, this does not exist)

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

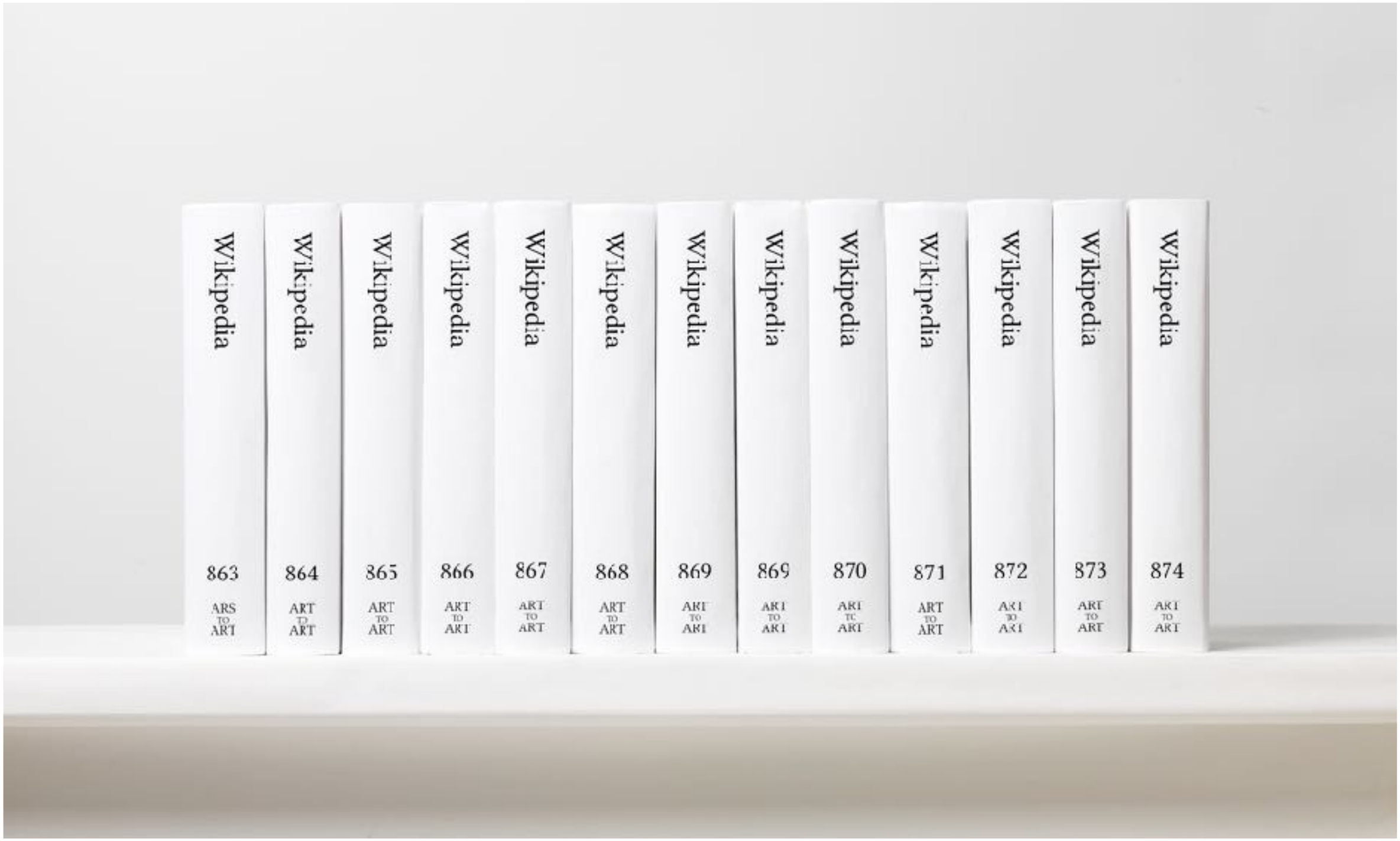


http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes

0.01% of Wikipedia



It exists! (sort of)



Own it!

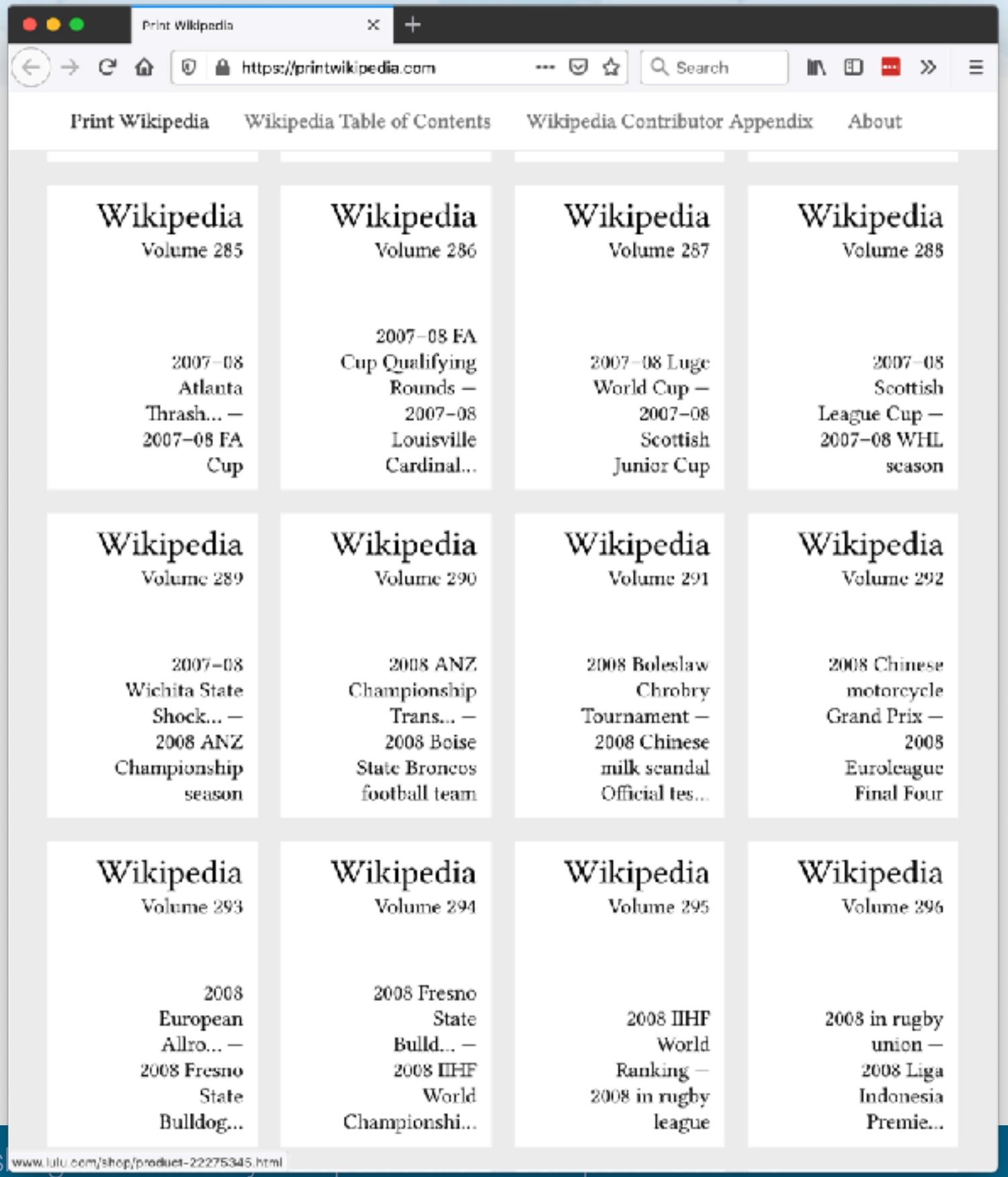
Just $80*!!!

*per volume

7,473 volumes each with 700 pages

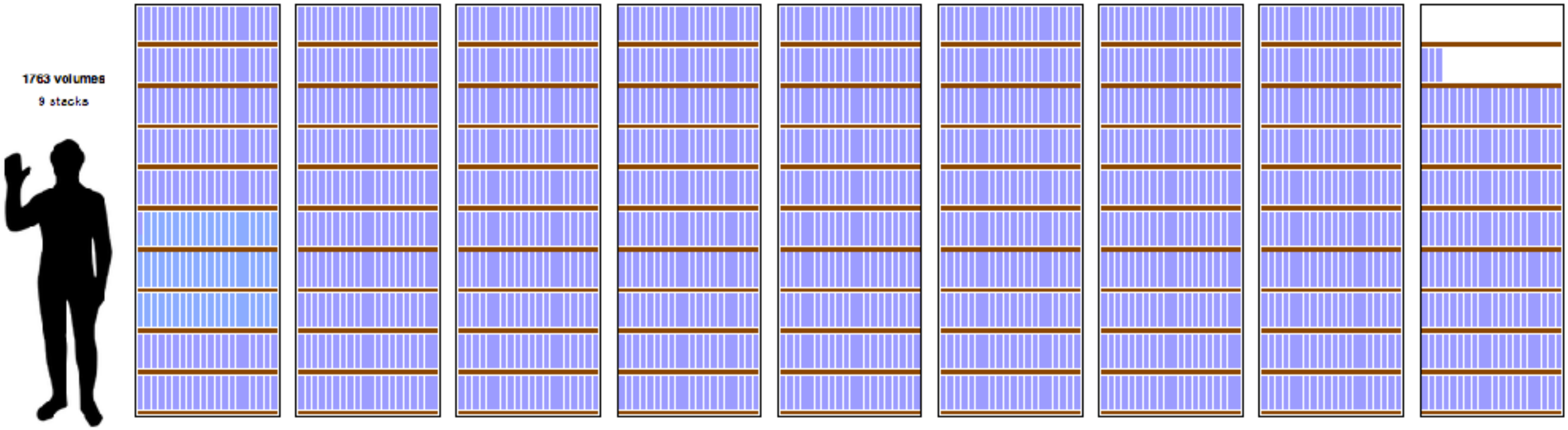
Print on demand

https://printwikipedia.com



Big Data in Perspective

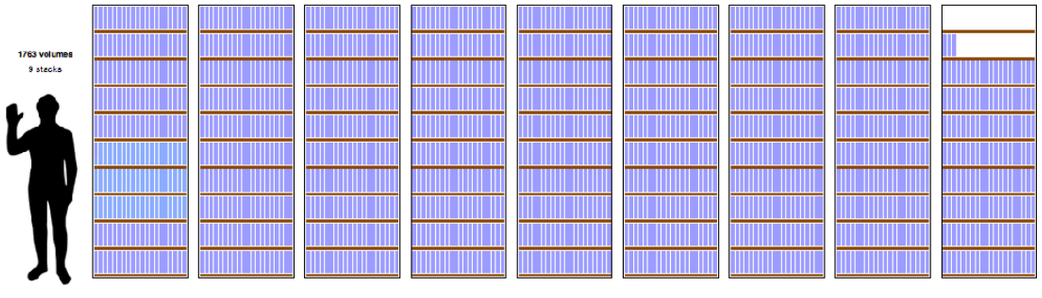
Wikipedia - 5TB of text



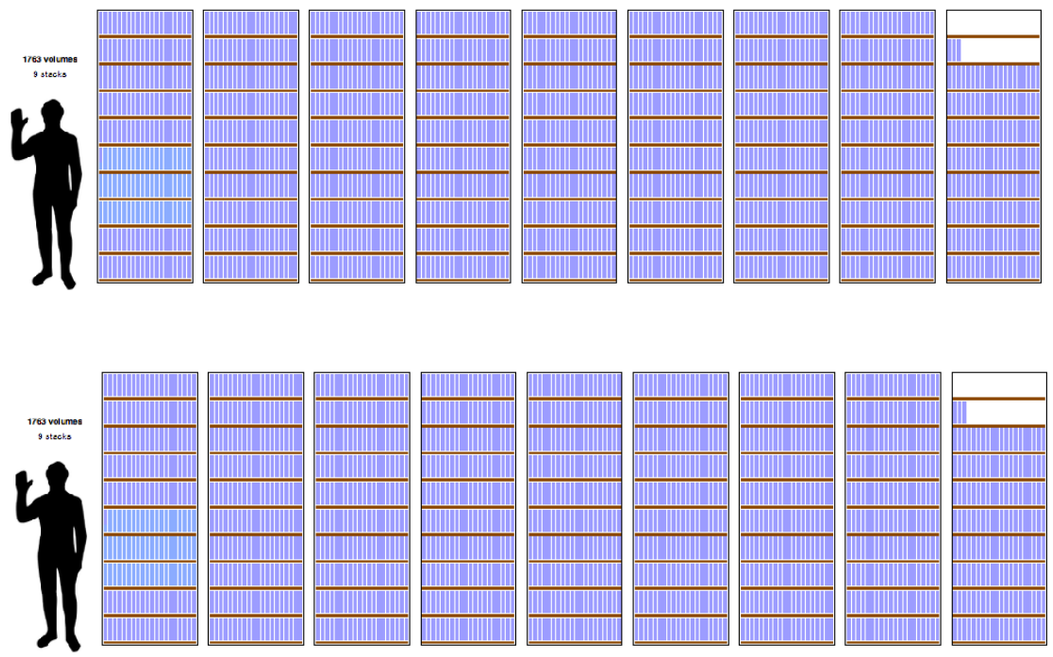
Facebook - ???

Big Data in Perspective

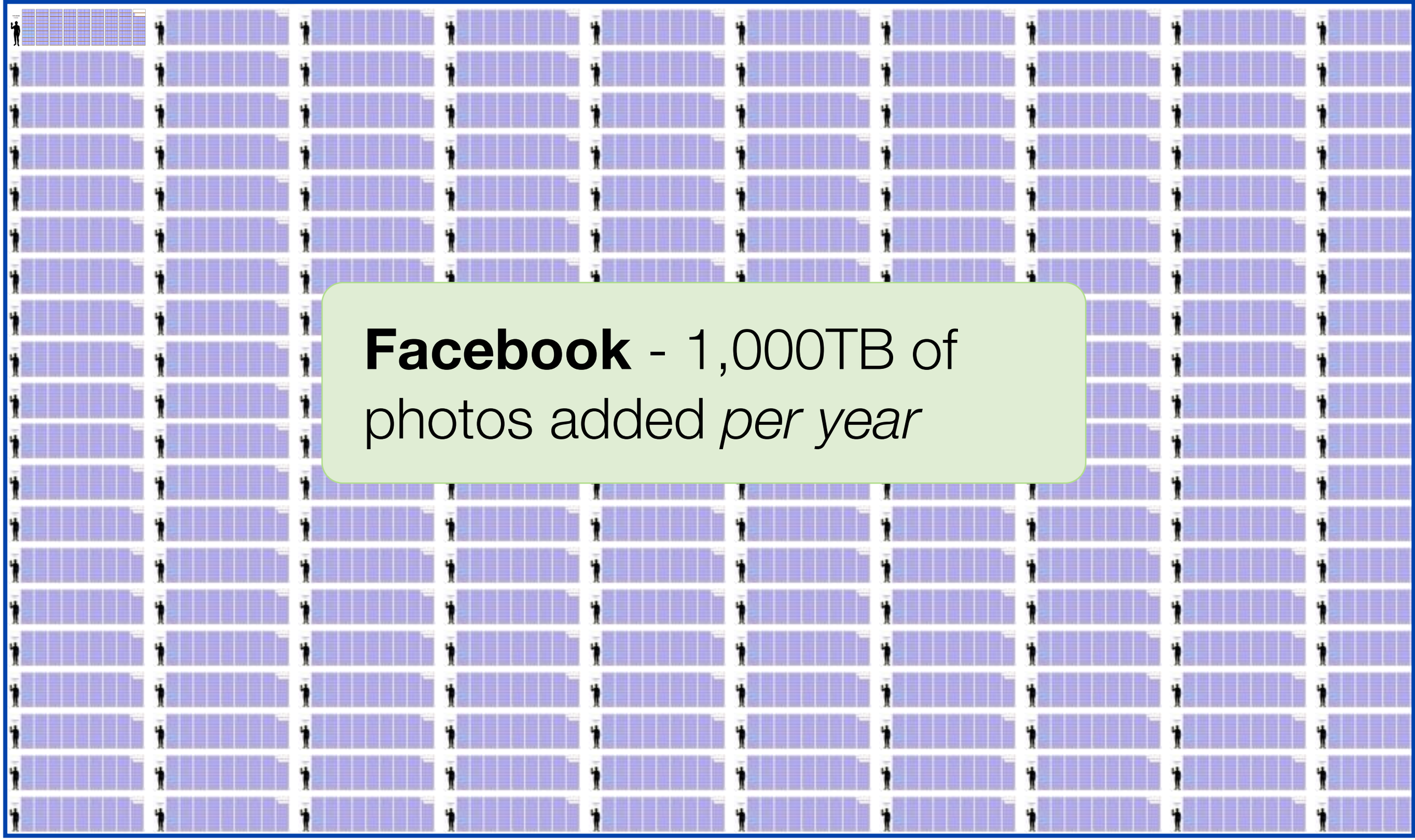
Wikipedia - 5TB of text



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 - <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.

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?