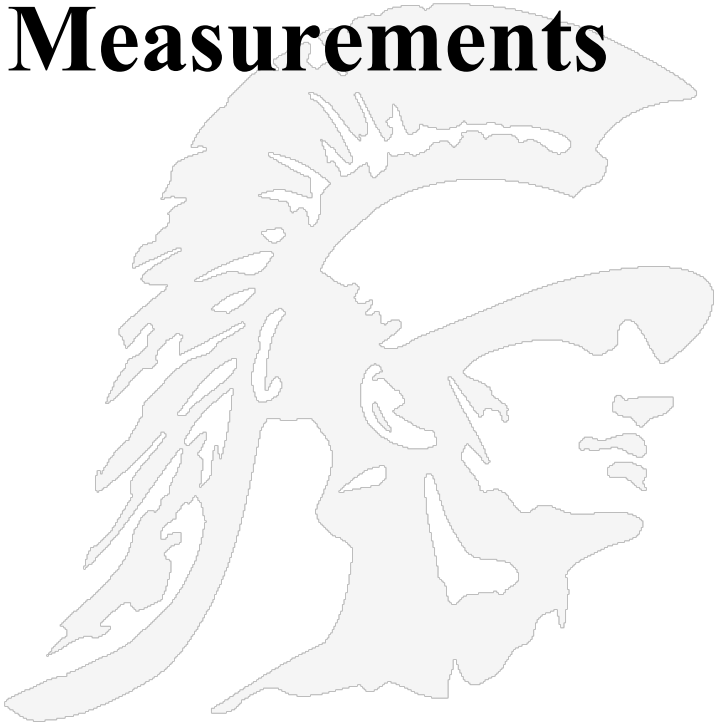
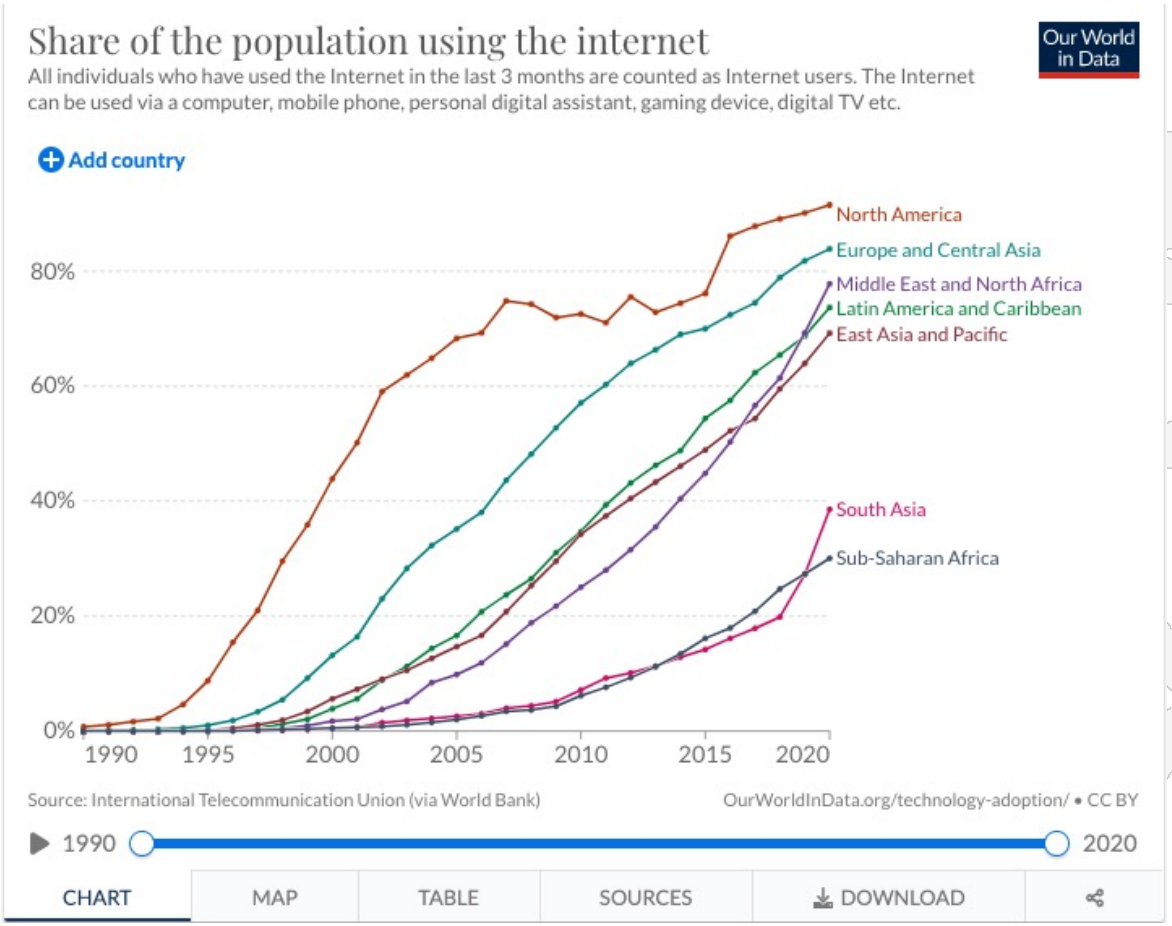
Web Trends and Measurements



Web Trends

- Web has changed dramatically over the last 30+ years
- If one is building a web search engine today it is important to understand the different dimensions of the web
  - Scale, complexity and growth are only a few important factors
- In today's lecture I try to quantify some of the trends to better understand where the web is going and why
- many of the early slides come from Mary Meeker (formerly of) Kleiner, Perkins, Caufield and Byers, a major venture capital firm, http://www.kpcb.com/

A total of **5.03 billion** people around the world use the internet today – equivalent to 63.1 percent of the world's total population

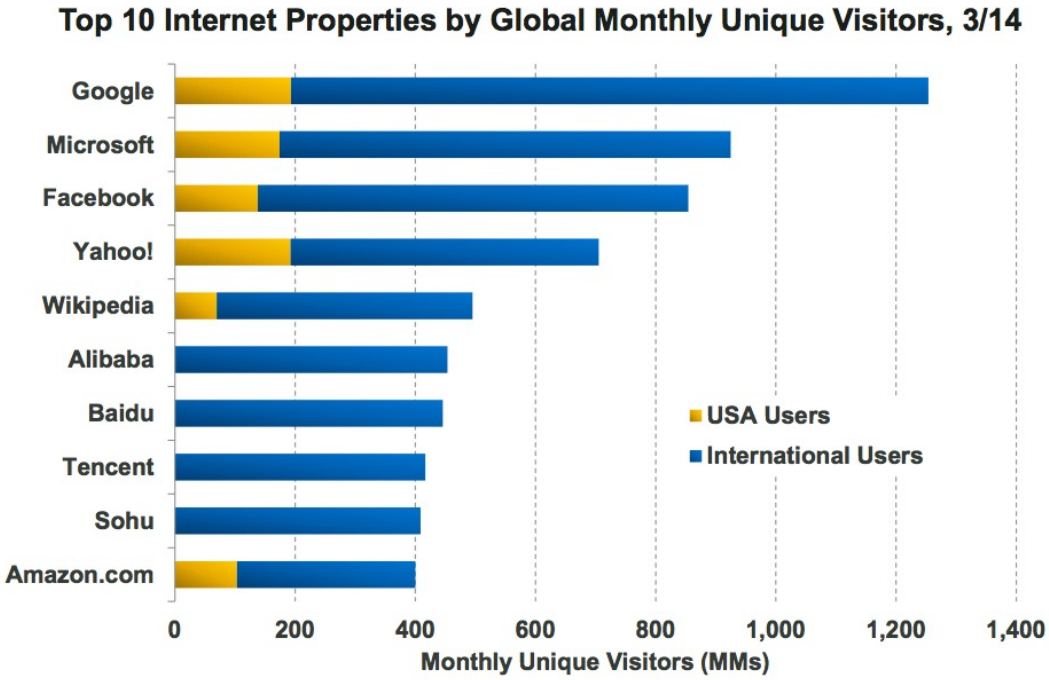


The US leads in the development of highly popular Internet websites;

Baidu is a Chinese search engine

Tencent is a Chinese holding company of Internet properties, among the most popular being, QQ, for chatting;
***Sohu**.com Inc. is a Chinese online media, search, gaming, community and mobile service group.*

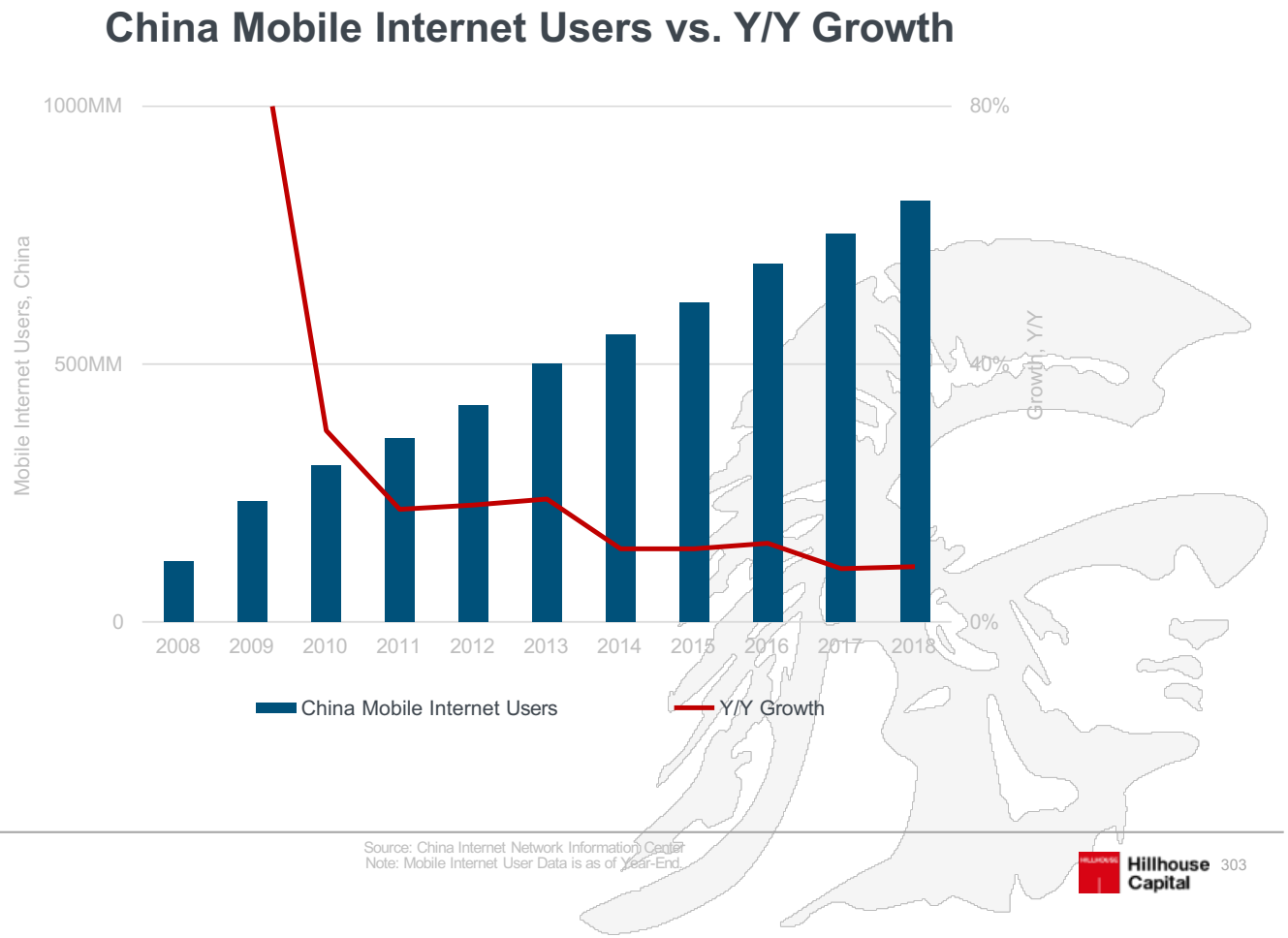
3/14 – 6 of Top 10 Global Internet Properties 'Made in USA'... >86% of Their Users Outside America...China Rising Fast



@KPCB Source: comScore, 3/14.

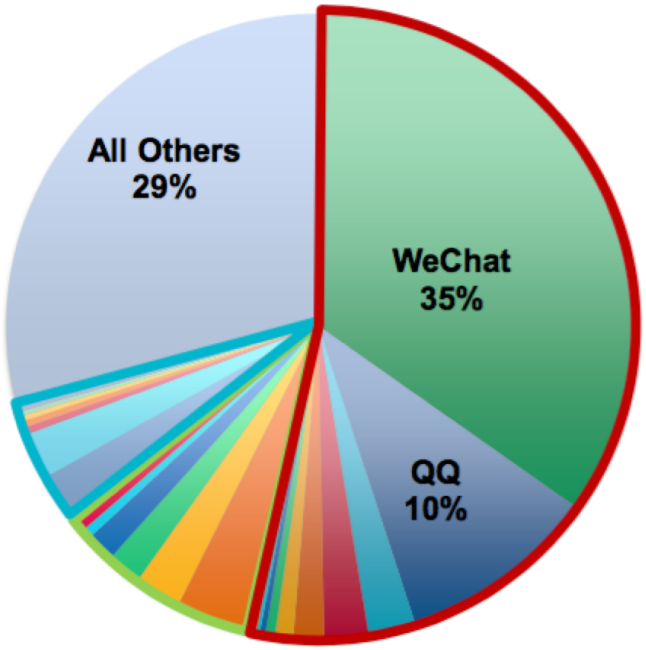
131

China Mobile Internet Users = 817MM.+9% vs. +8% Y/Y



China Mobile Internet Usage Leaders...
Tencent + Alibaba + Baidu = 71% of Mobile Time Spent

Share of Mobile Time Spent, April 2016
Daily Mobile Time Spent = ~200 Minutes per User, Average



Tencent
- WeChat
- QQ
- QQ Browser
- Tencent Video
- Tencent News
- Tencent Games
- QQ Music
- JD.com
- QQ Reading

Alibaba
- UCWeb Browser
- Taobao
- Weibo
- YouKu Video
- Momo
- Shuqi Novel
- AliPay
- AutoNavi

Baidu
- Mobile Baidu
- iQiyi / PPS Video
- Baidu Browser
- Baidu Tieba
- 91 Desktop
- Baidu Maps

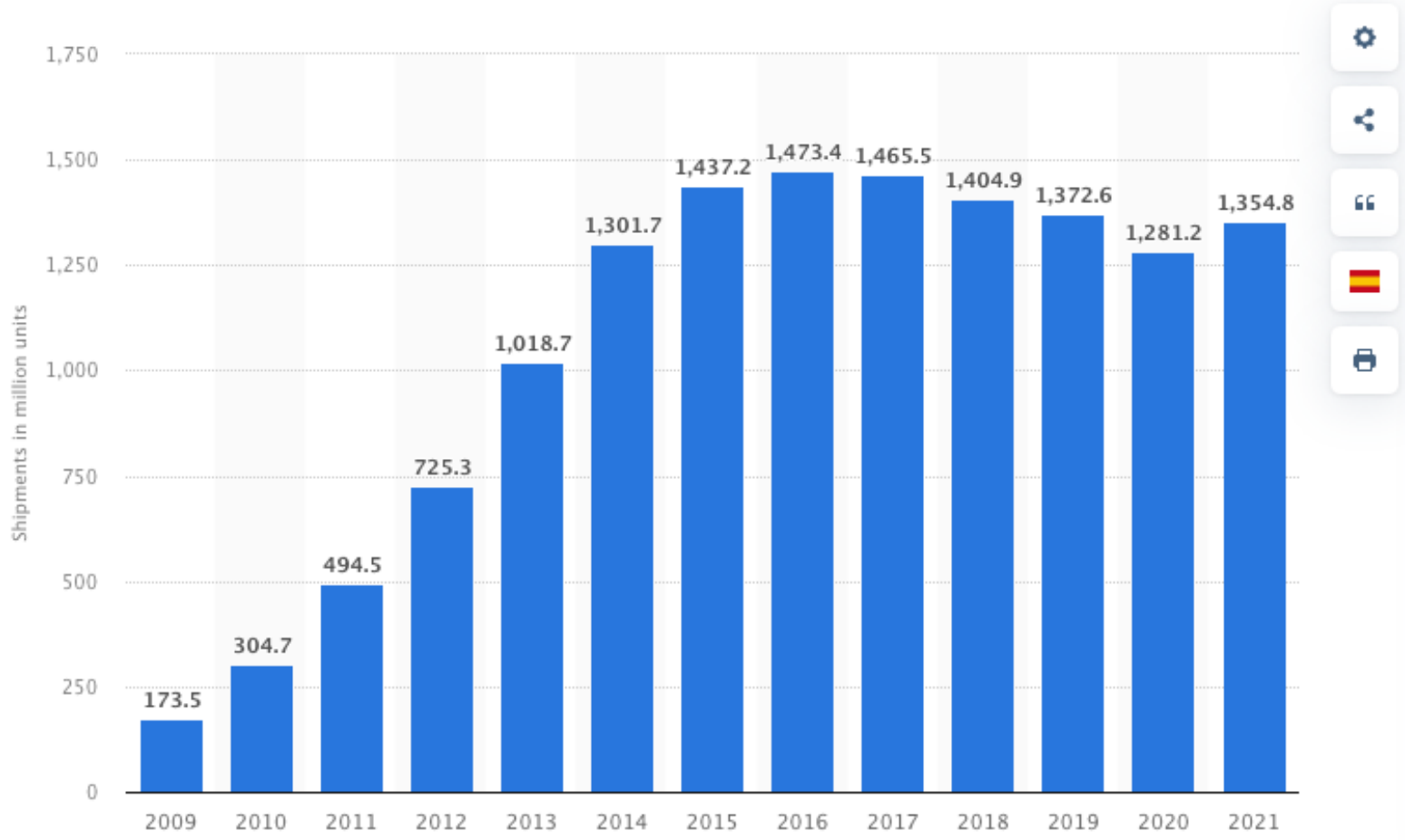
- All Other

@KPCB

Note: Grouping of apps include strategic investments made by Tencent, Alibaba and Baidu. Only apps in top 50 by time spent share are called out. Source: QuestMobile, Trustdata, and Hillhouse estimates.

Hillhouse Capital

Global smartphone shipments from 2009 to 2021

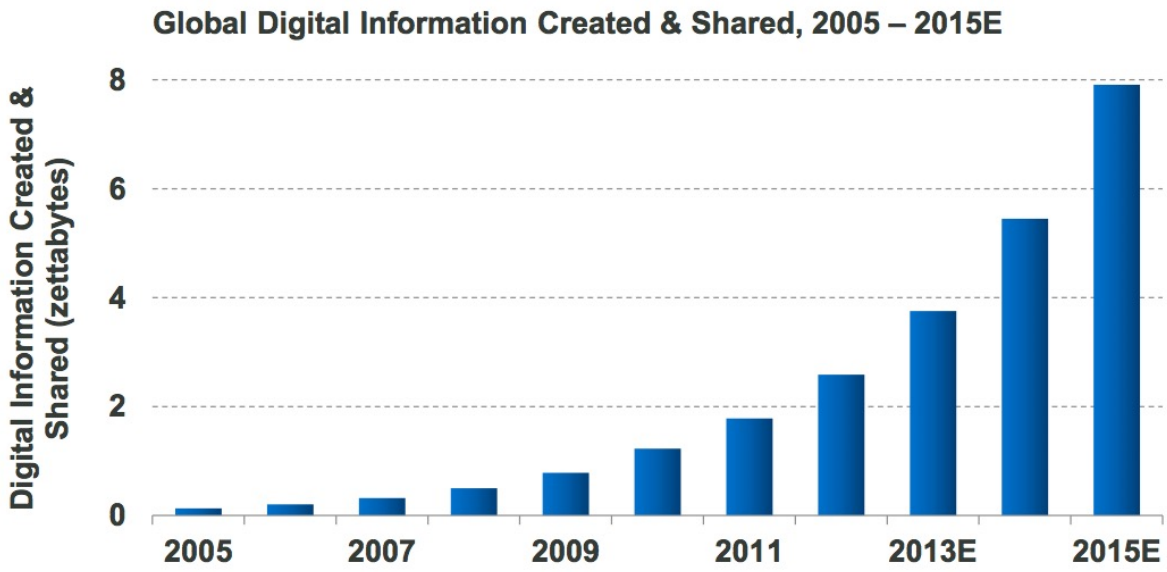


Details: Worldwide; IDC; 2009 to 2021

World's Content is Increasingly Findable + Shared + Tagged - Digital Info Created + Shared up 9x in Five Years

There has been exponential growth in online information;
1 Zettabyte = 1,024 Exabytes
1 Exabyte = 1,024 Petabytes
1 Petabyte = 1,024 Terabytes
1 Terabyte = 1,024 Gigabytes
or
1 Zettabyte = 1,000,000,000,000 gigabytes

Amount of global digital information created & shared – from documents to pictures to tweets – grew 9x in five years to nearly 2 zettabytes in 2011, per IDC.*



KPCB

Note: * 1 zettabyte = 1 trillion gigabytes. Source: IDC report "Extracting Value from Chaos" 6/11. 11

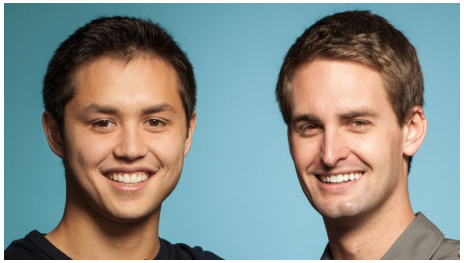
Photos Alone = 1.8B+ Uploaded & Shared Per Day...
Growth Remains Robust as New Real-Time Platforms Emerge

500 million photos are uploaded every day and that number is doubling every year

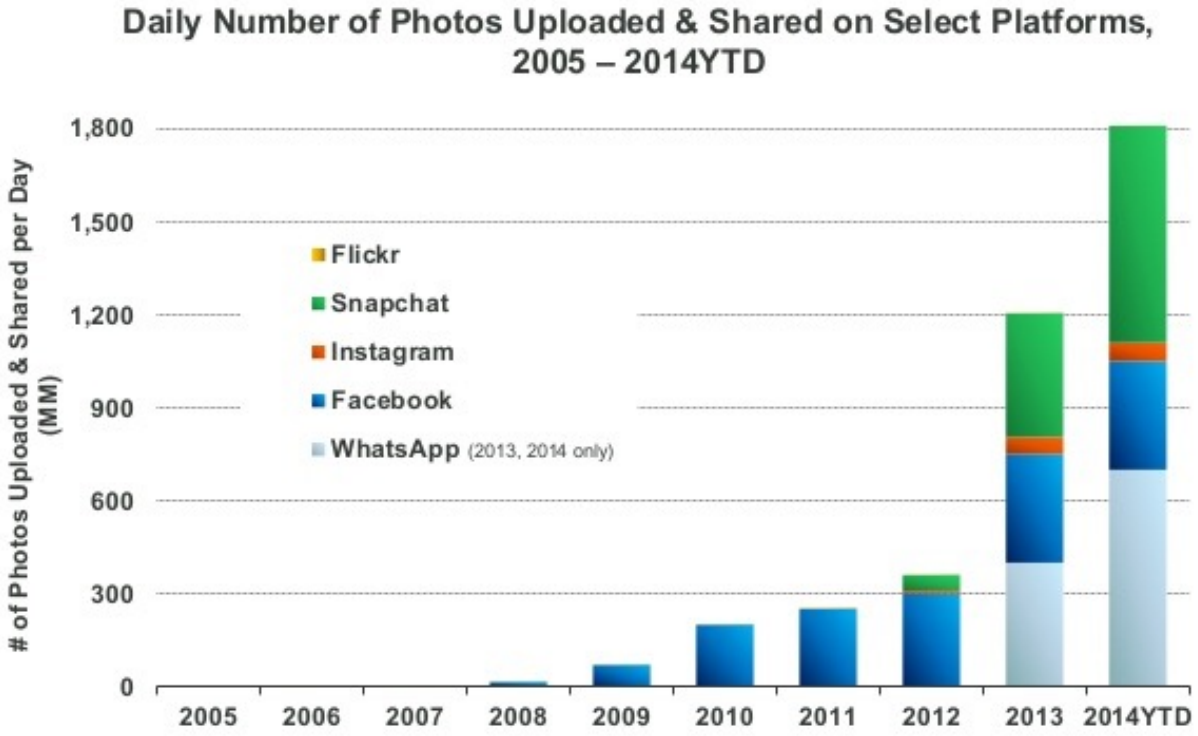
Yahoo has recently made a major upgrade to **Flickr**

Instagram was in 2010 purchased by Facebook for $1 billion

Snapchat is a photo messaging application developed by two Stanford students ($9B valuation);



bobby Murphy - Evan Spiegel



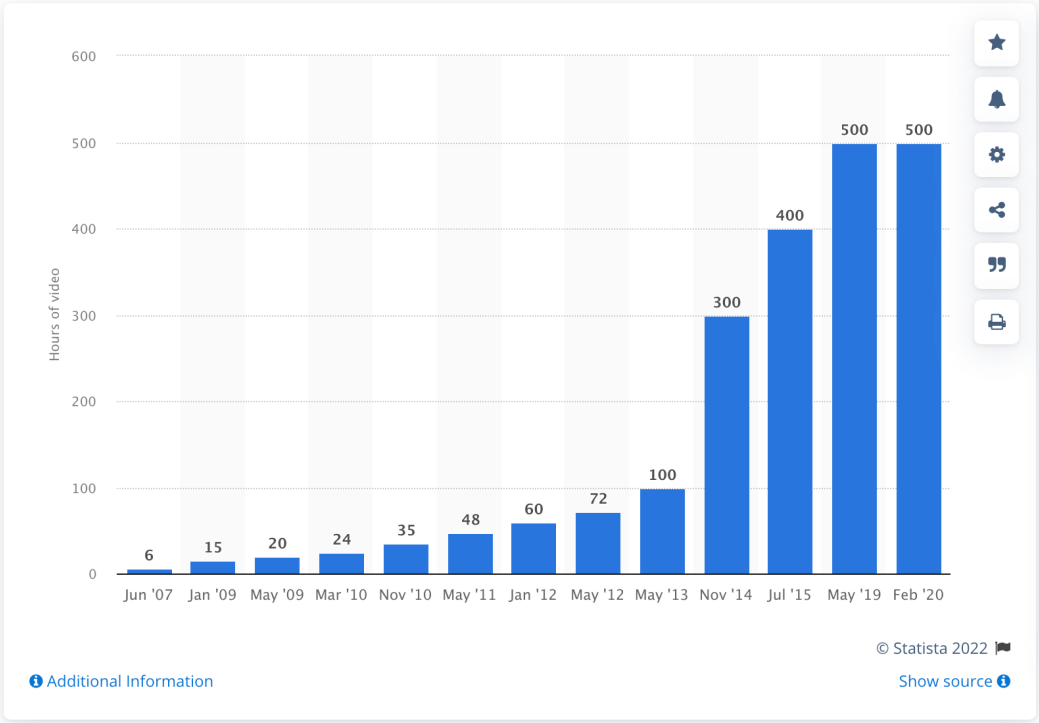
@KPCB Source: KPCB estimates based on publicly disclosed company data, 2014 YTD data per latest as of 5/14.

62

Content Uploaded to YouTube

Internet › Online Video & Entertainment

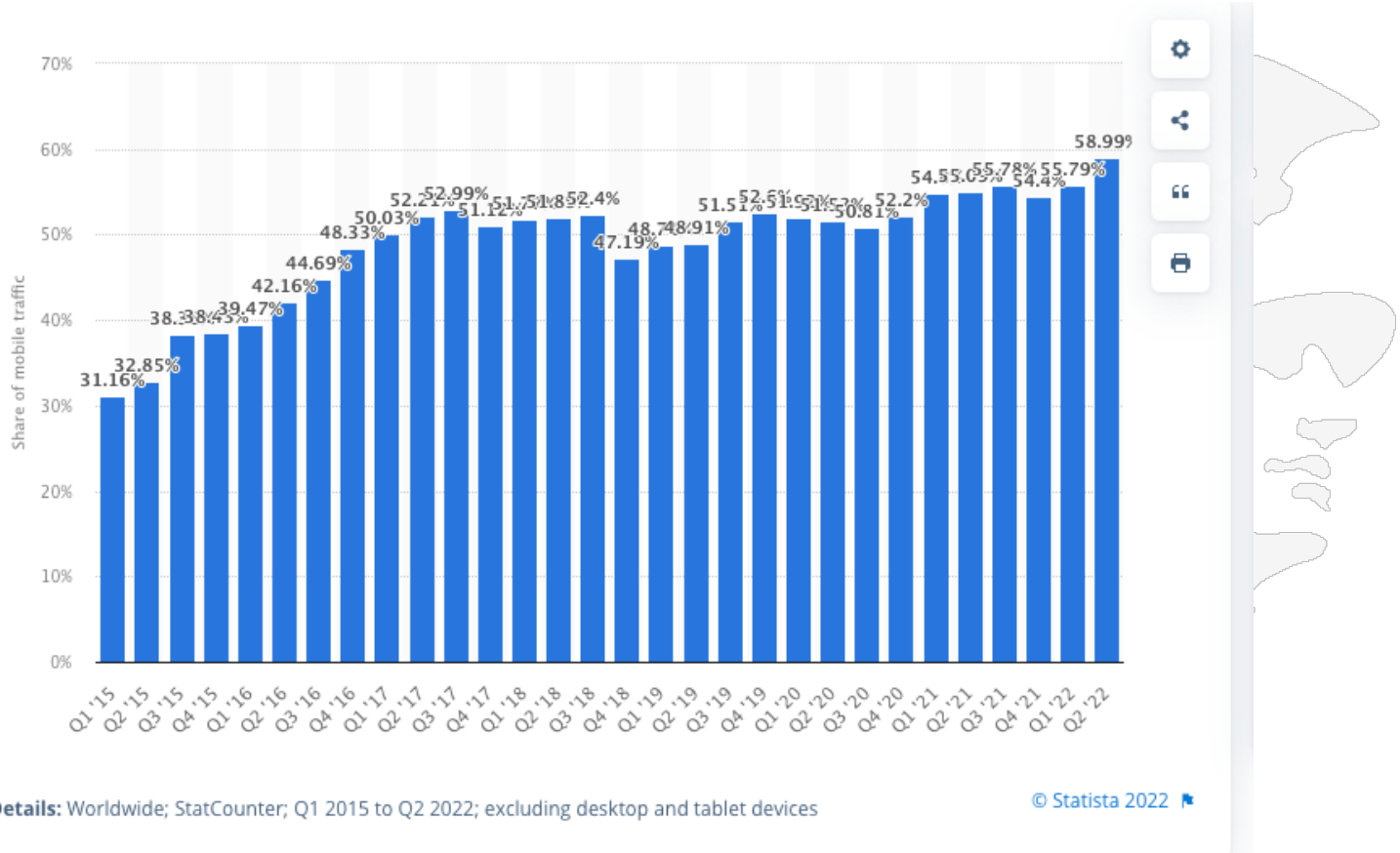
Hours of video uploaded to YouTube every minute as of February 2020



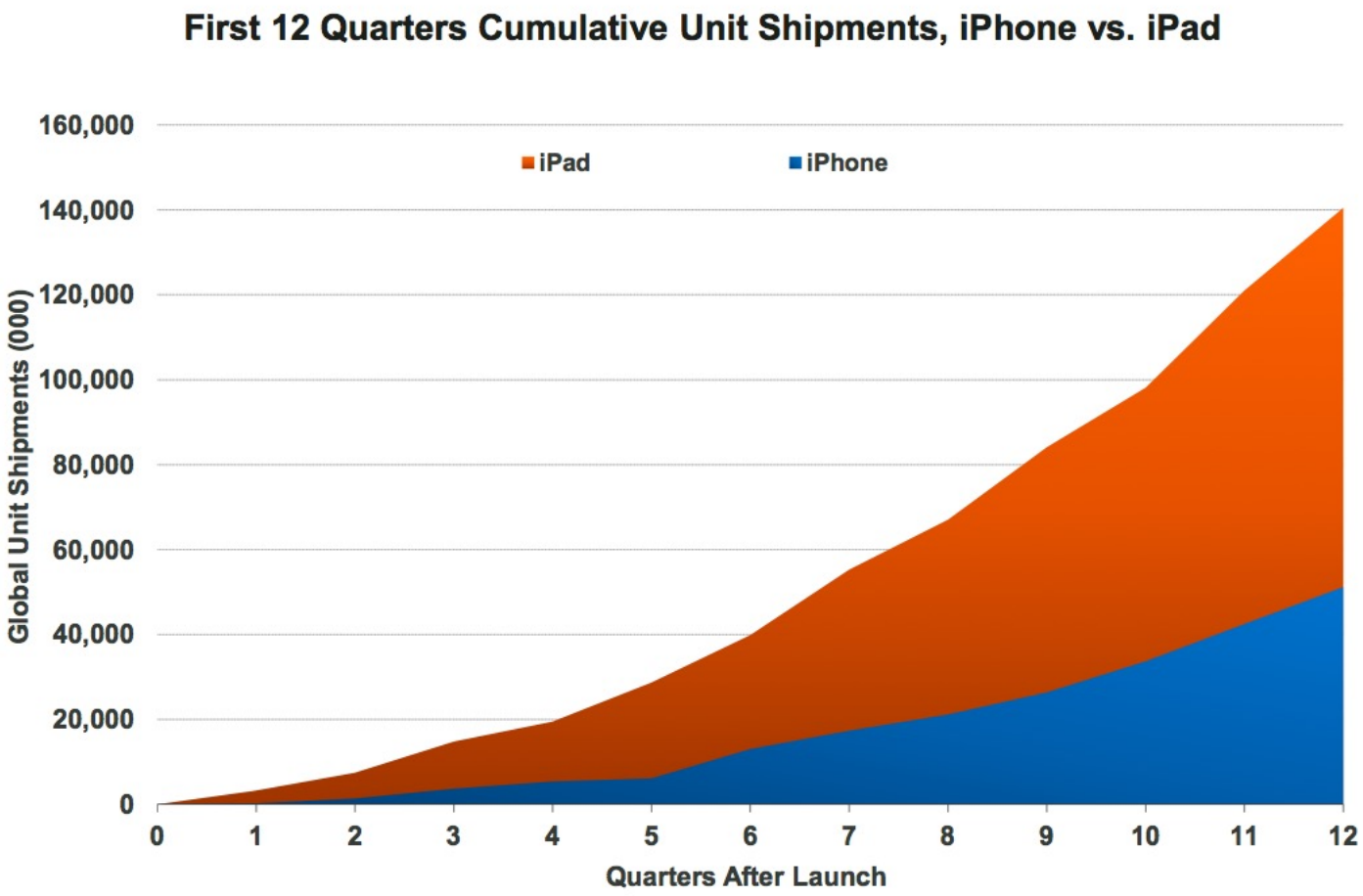
- As of February 2020, more than 500 hours of video were uploaded to YouTube every minute.
- This equates to approximately 30,000 hours of newly uploaded content per hour.
- The number of video content hours uploaded every 60 seconds grew by around 40 percent between 2014 and 2020.

Mobile Accounts for 50% of Web Traffic

In the second quarter of 2022, mobile devices (excluding tablets) generated 58.99 percent of global website traffic, consistently hovering around the 50 percent mark since the beginning of 2017 before permanently surpassing it in 2020.



Tablet Growth =
More Rapid than Smartphones, iPad = ~3x iPhone Growth



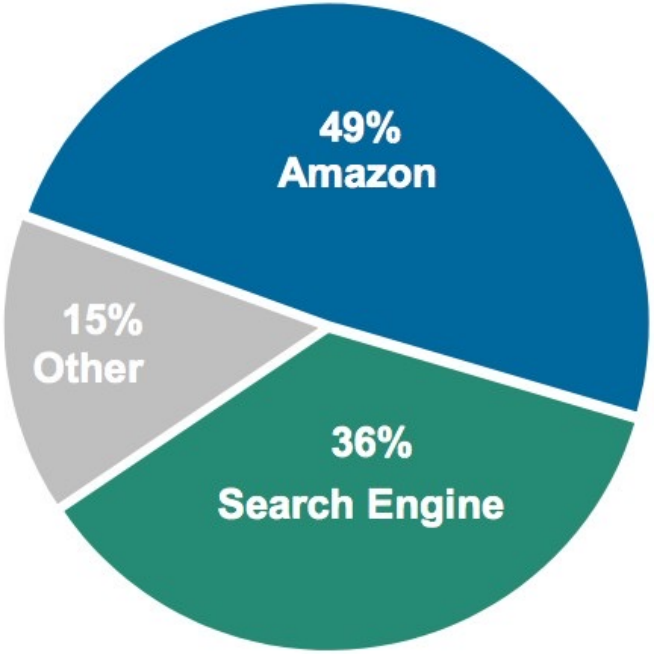
KPCB

Source: Apple, as of CQ1:13 (12 quarters post iPad launch).
Launch Dates: iPhone (6/29/07), iPad (4/3/10).

44

Product Finding =
Often Starts @ Search (Amazon + Google...)

Where Do You Begin Your Product Search?

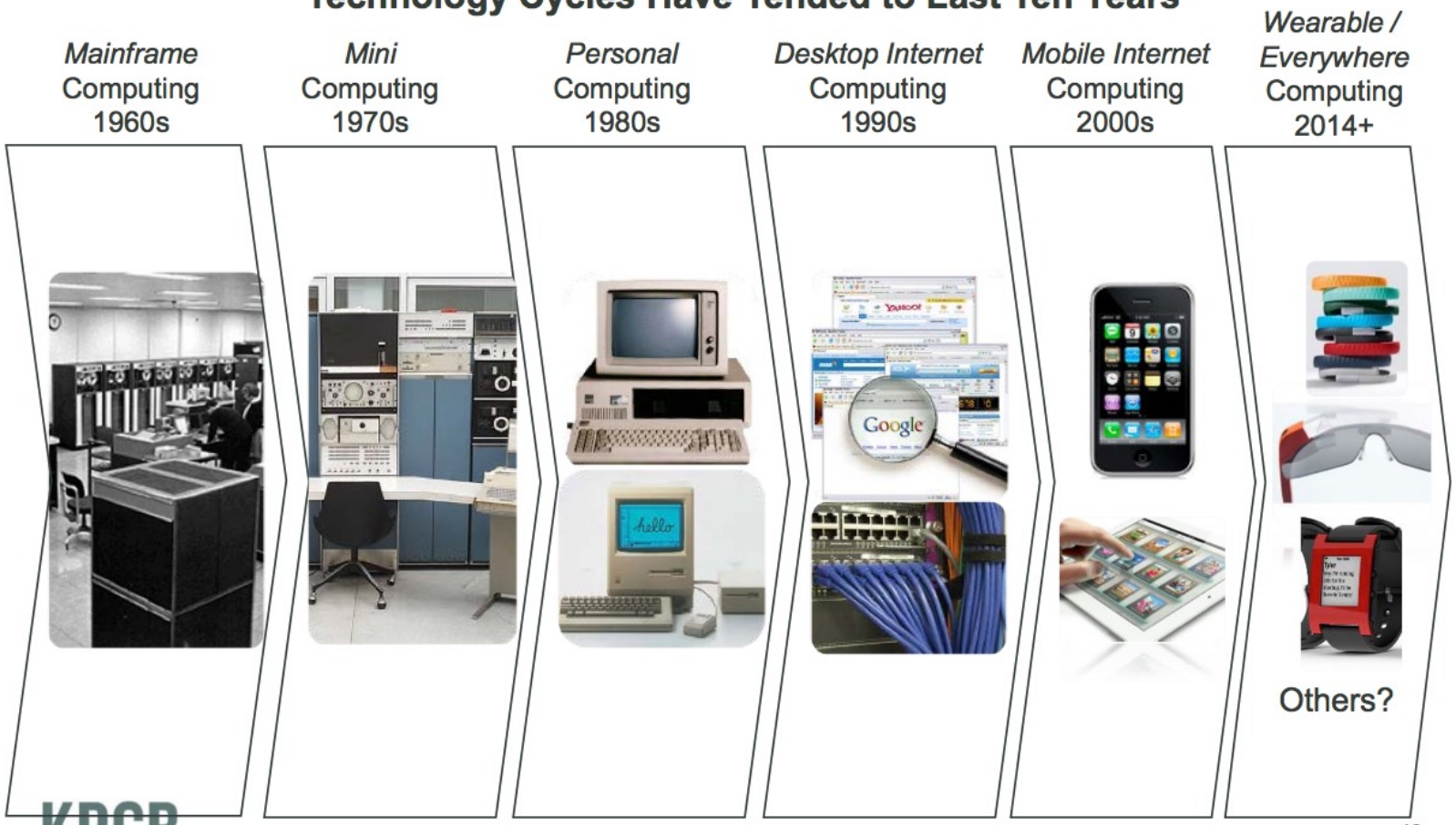


KLEINER PERKINS
2018
INTERNET TRENDS

Source: Survata (9/17). Note: n = 2,000 USA customers.

Technology Cycles – Still Early Cycle on Smartphones + Tablets, Now Wearables Coming on Strong, Faster than Typical 10-Year Cycle

Technology Cycles Have Tended to Last Ten Years



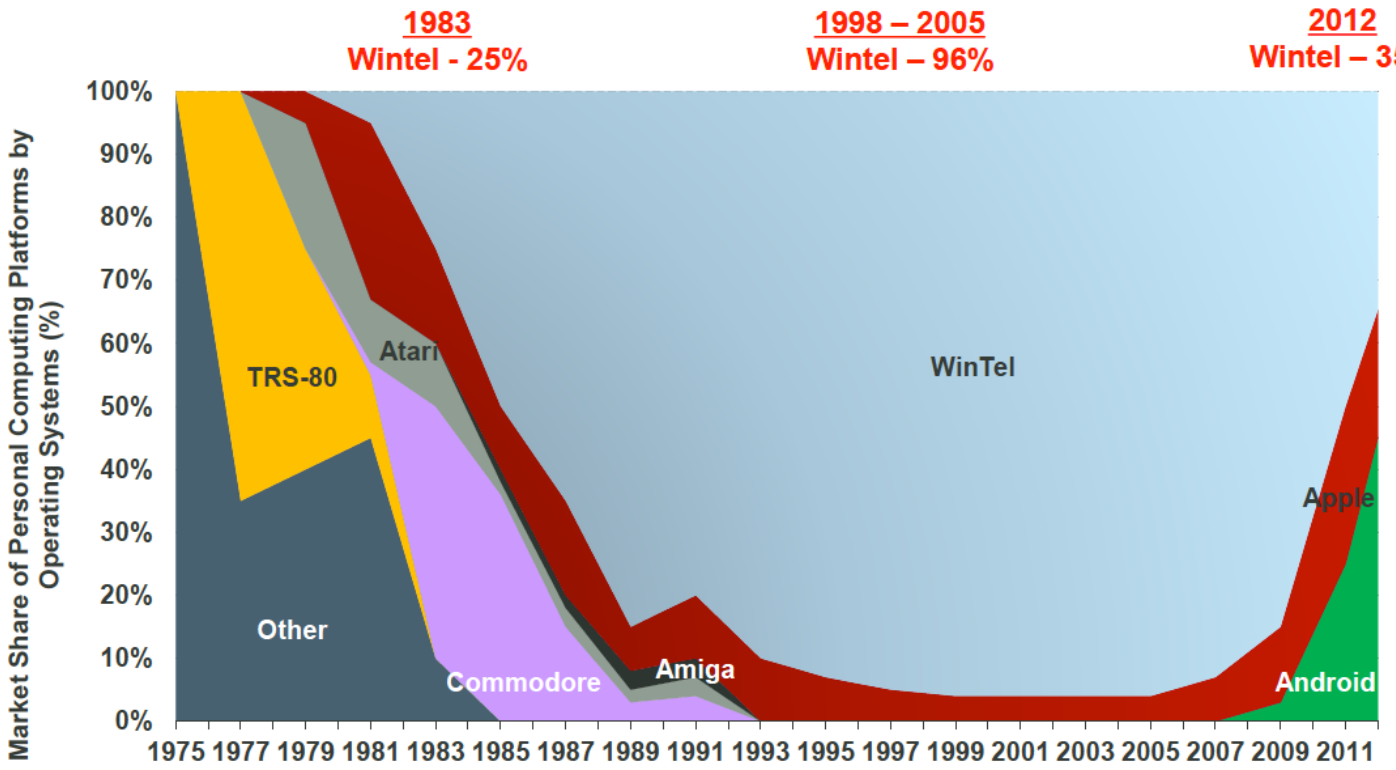
KPCB

49

Image Source: Computersciencelab.com, Wikipedia, IBM, Apple, Google, NTT docomo, Google, Jawbone, Pebble.

Re-Imagination of Computing Operating Systems - iOS + Android = 60% Share vs. 35% for Windows

Global Market Share of Personal Computing Platforms by Operating System Shipments, 1975 – 2012



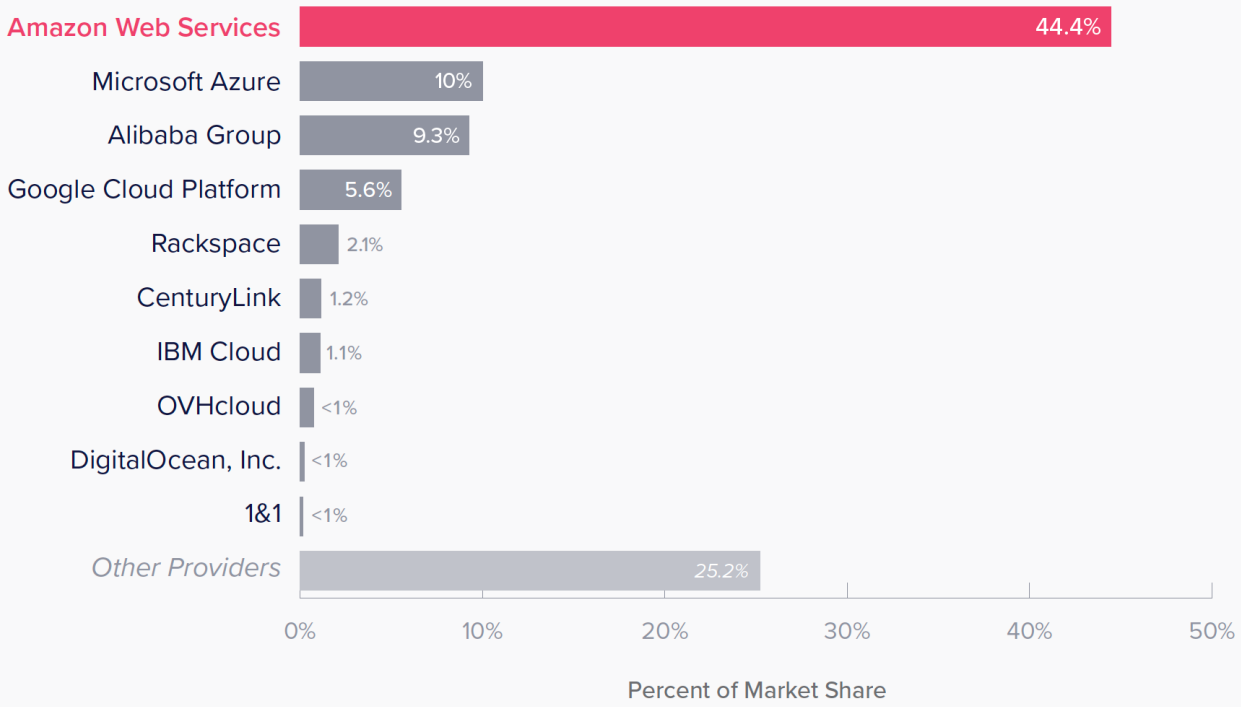
KPCB

Source: Asymco.com (as of 2011), Public Filings, Morgan Stanley Research, Gartner for 2012 data. 109

Major Cloud Providers

Market Share Of Leading Cloud Hosting Providers

Top 10 Providers by Total 2020 Market Share



Source: Intricately data, April 2021

...While The Cloud Rises

Amazon Web Services (AWS) Leading Cloud Charge...

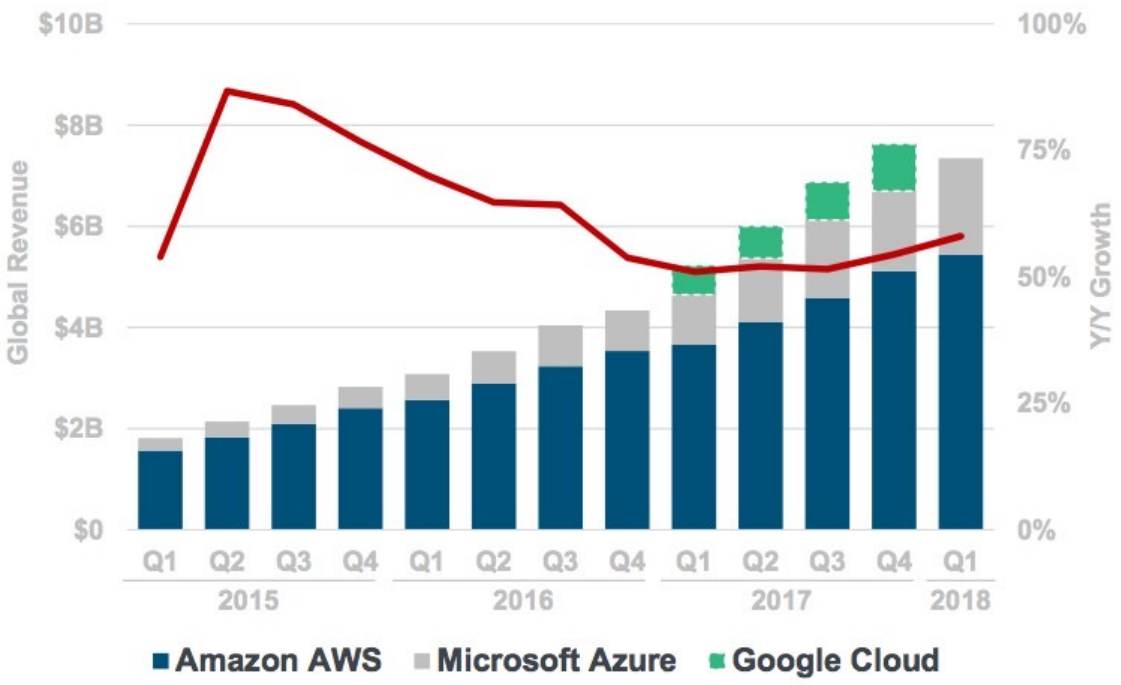
@KPCB

*Note: S3 is AWS' storage product and used as proxy for AWS scale / growth .
Source: Company data.

74

...Computing Big Bangs Volume Effects = Cloud Revenue Re-Accelerating +58% vs. +54% Q/Q

Cloud Service Revenue – Amazon + Microsoft + Google



KLEINER PERKINS
2018
INTERNET TRENDS

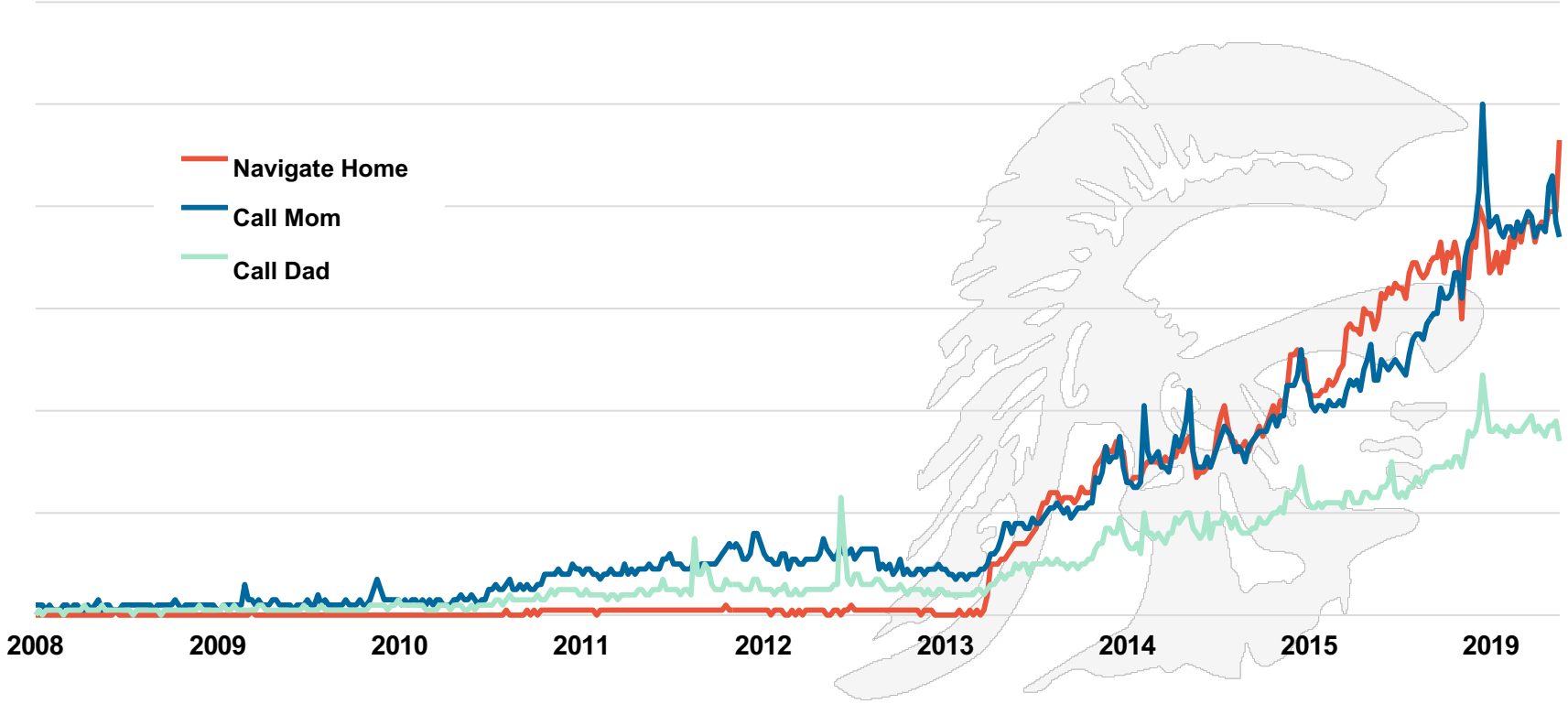
Source: Amazon AWS = Company filings, Microsoft Azure = Keith Weiss @ Morgan Stanley (4/18), Google Cloud = Brian Nowak @ Morgan Stanley (5/18). Note: Google Cloud revenue excluded in Y/Y growth rate calculation due to limited quarterly estimates.

185

Google Voice Search Queries

Google Trends imply queries associated with voice-related commands have risen >35x since 2008 after launch of iPhone & Google Voice Search

Google Trends, Worldwide, 2008 – 2019

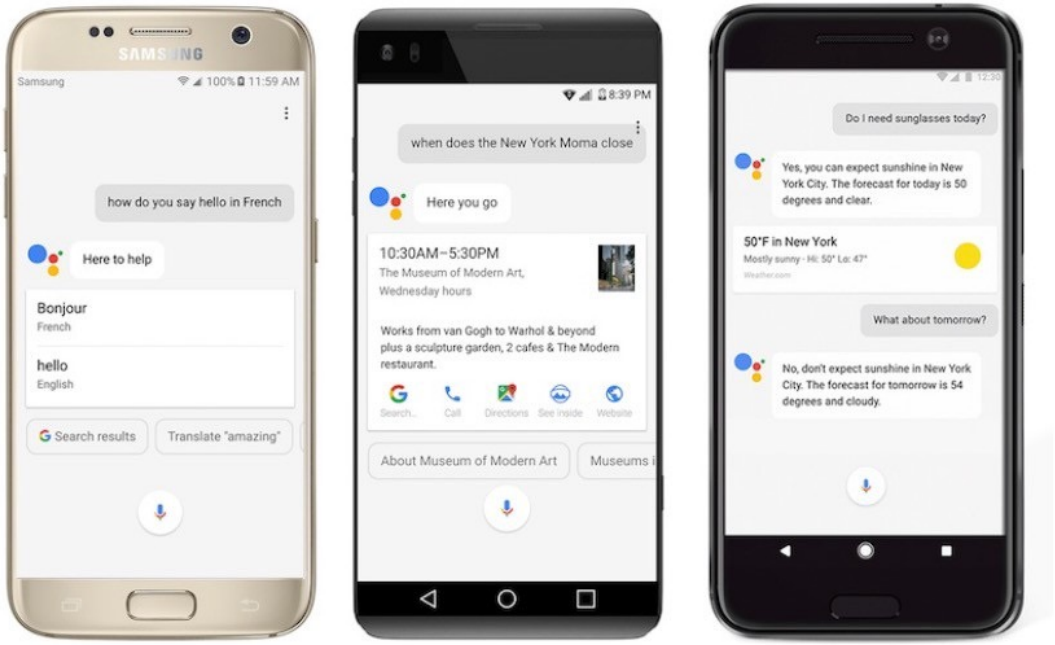


Voice-Based *Mobile* Platform *Front-Ends* = Voice Can Replace Typing

Google Assistant

Nearly 70% of Requests are Natural / Conversational Language, 5/17

20% of Mobile Queries Made via Voice, 5/16

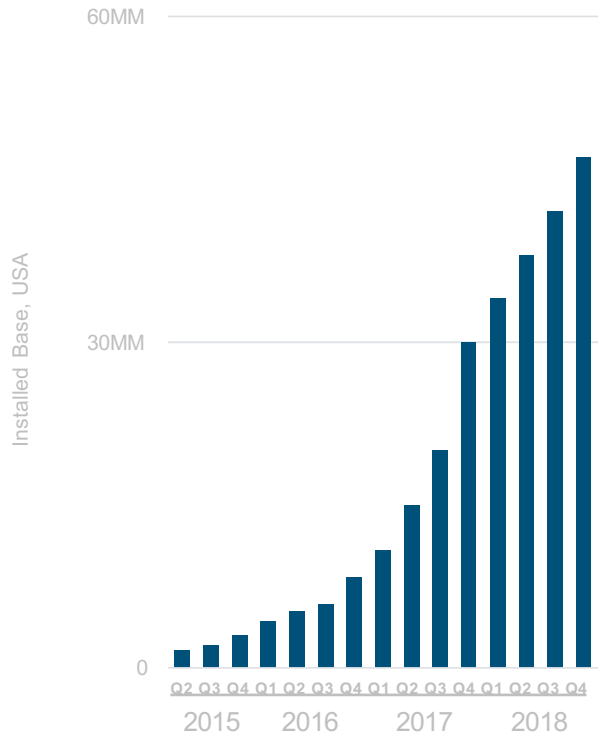


KLEINER PERKINS

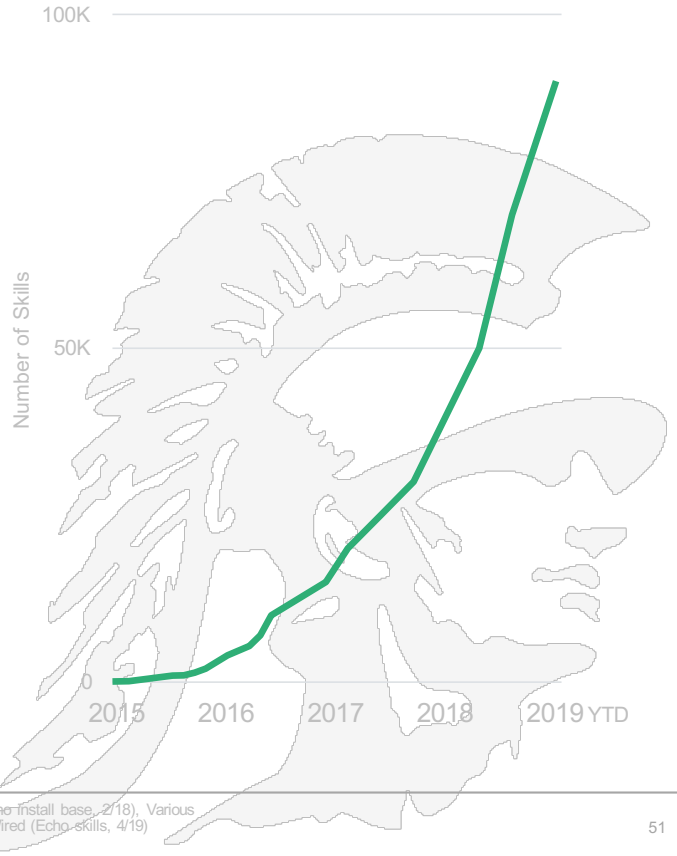
Source: Google I/O (5/16), Image: Macrumors (2/17)

.Voice =
47MM Amazon Echo Base + ~2x in One Year

Amazon Echo Installed Base



Amazon Echo Skills



BOND
Internet Trends
2019

Source: Consumer Intelligence Research Partners LLC (Echo install base, 2/18), Various media outlets including Geekwire, TechCrunch, & Wired (Echo skills, 4/19)

51

21

Summary of Recent Trends in Web/Internet Development

- Growth in number of users connected
- Growth in Smartphone use
- Growth in digital data, especially photos and video
- Growth in Social Media as an advertising platform
- Transition from desktop/laptop use to mobile
- Growth in tablet usage over desktops/laptops
- Decreased dominance of Microsoft Windows
- Move away from server farms to cloud computing
- Growth in voice communication with devices

Measuring the Web

- The World Wide Web (the Web, the publicly accessible web) is so dynamic it is hard to describe it and have the description be valid for very long
- In this lecture we look at what is known,
  - Measuring the Web by number of web sites
  - Measuring the Web by the Languages of Web Pages
  - Measuring the Web by Rate of Change of Pages
  - Measuring the Web by Document Content Type
  - Measuring the Web by linkage
  - Measuring the Web as a Graph
  - Measuring the Web by Content
  - (using the best statistics we can find)

Number of Websites

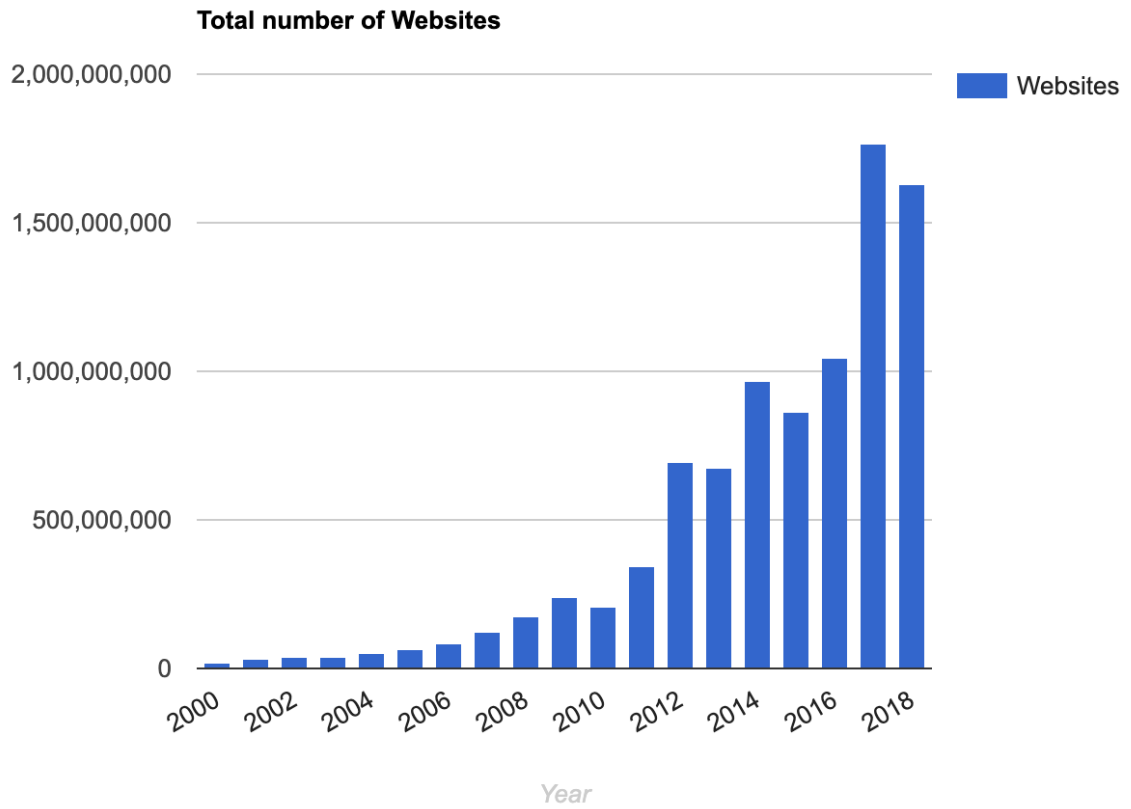
Jan. 2020:
~1.7 Billion sites

nginx web server had the largest growth;

Over 50% of websites Are hosted either by Apache or nginx;

But Microsoft web servers still power 43.2% of all sites

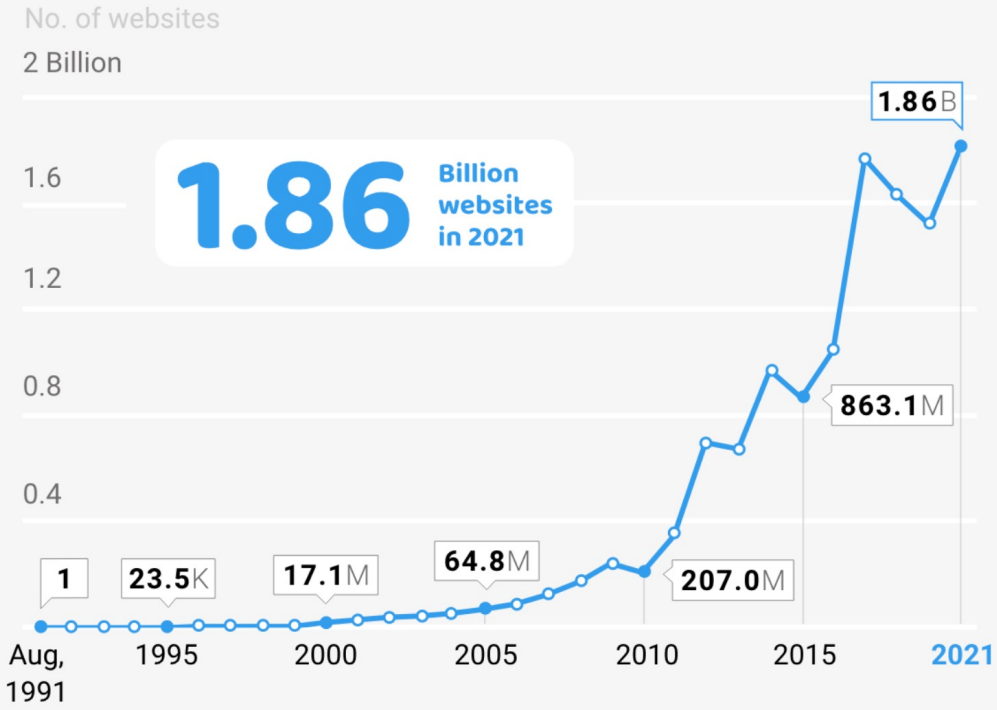
Around 75% of websites are not active, but parked



http://www.internetlivestats.com/total-number-of-websites/

Number of websites in the world

The global number of websites has more than doubled from 2015 to 2021. Websites growth rate from 1991 to 2021



Distribution Across TLDs

136 million in .com, 21million in .tk, 14 million in .de, etc

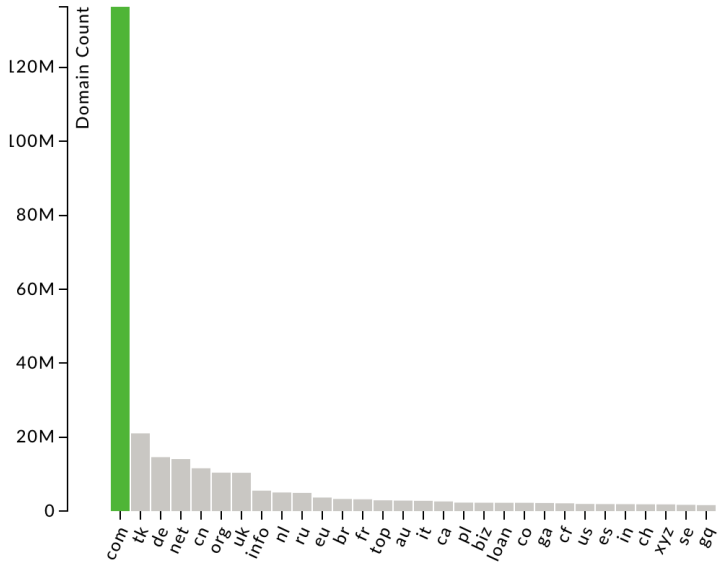
what is .tk and why is it so large? (Tokelau)

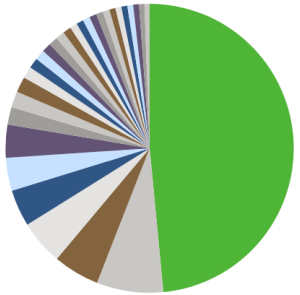
Domain Count Statistics for TLDs

This page displays the count of all Domains in each TLD. For Registry's publishing a domain count, "Our Count" should closely match their published record. For registry's that don't provide a zone file or publish an up-to-date record, Our Count represents all domains we know about, which is usually more accurate.

TLD	Our Count
.com	136,346,743
.tk	21,014,704
.de	14,586,920
.net	14,061,971
.cn	11,579,296
.org	10,398,501
.uk	10,361,314
.info	5,528,696
.nl	5,042,167
.ru	4,928,746
.eu	3,673,059
.br	3,282,608
.fr	3,195,248
.top	2,921,081
.au	2,845,979
.it	2,786,780
.ca	2,611,854
.pl	2,302,199
.biz	2,269,454
.loan	2,253,686

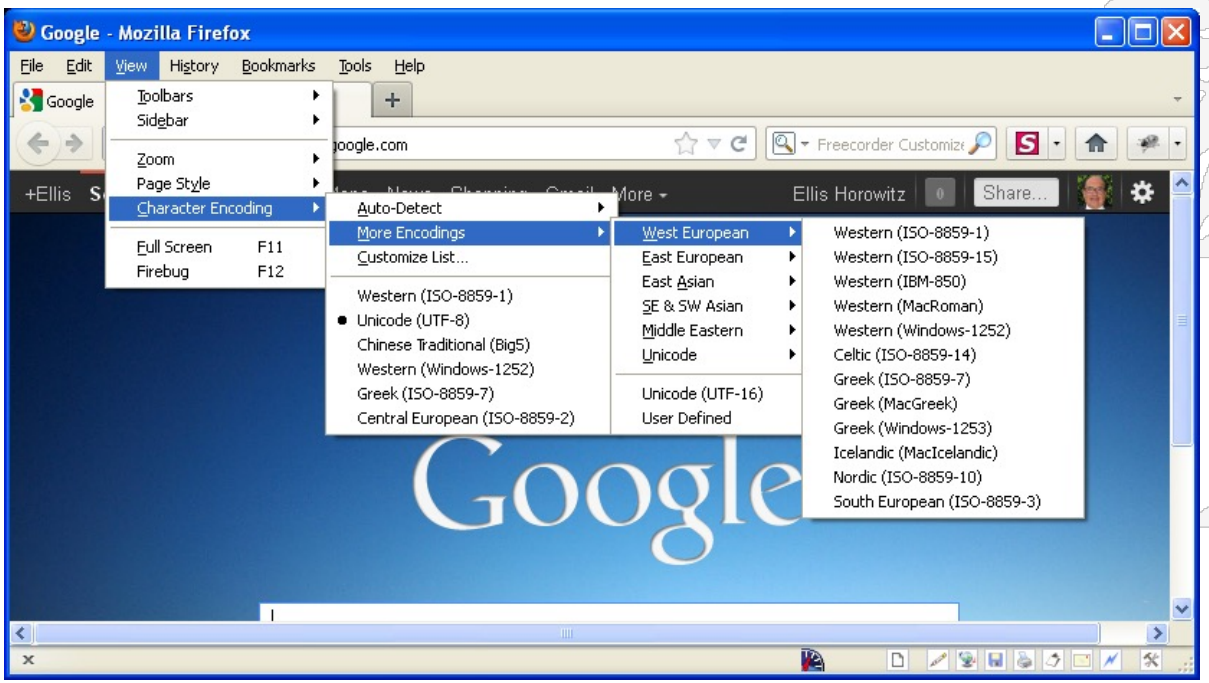
Display: Top25 - Default - None





Web Page Language Diversity

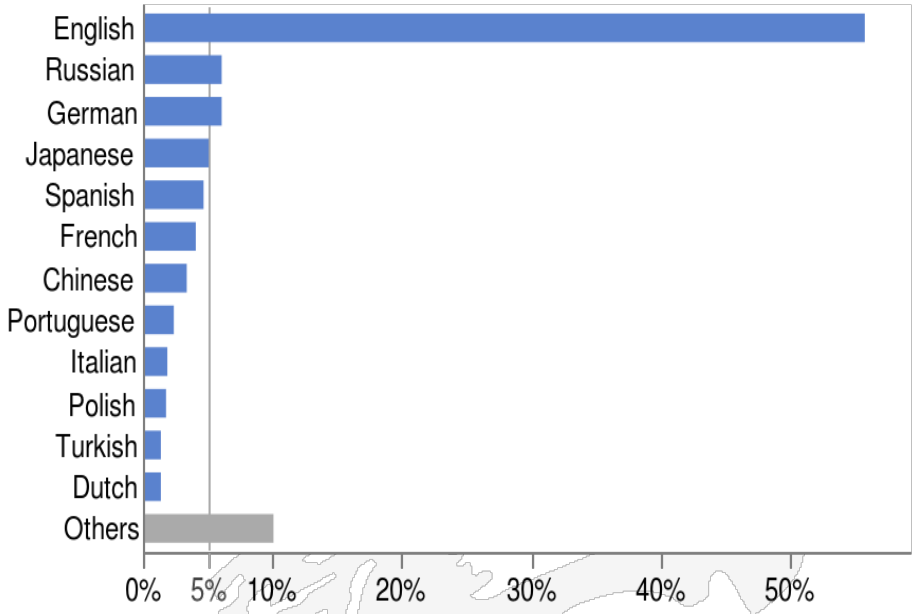
- **The Web contains pages in many different languages**
- Characters in a language are encoded such that each character is paired with a number
- Unicode and its parallel standard, the ISO/IEC 10646 Universal Character Set, together constitute a modern, unified character encoding.
- Most modern web browsers feature automatic character encoding detection. In Firefox, for example, see the View/Character Encoding submenu, shown below
- In HTML one can specify the character encoding using
- **<meta http-equiv="Content-Type content="text/html" charset=utf-8>**



- If charset is missing ISO-8859-1 is taken as the default unless there is a browser setting;
- Websites in non-western languages typically use UTF-8

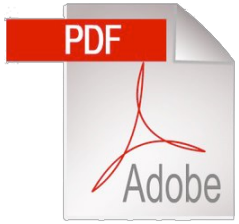
Measuring Language Diversity

- **It is estimated that about 40,000 different languages have been created by human beings**
- **Only between 6,000-9,000 are still in use**
- **Study done by the United Nations**
  - http://unesdoc.unesco.org/images/0018/001870/187016e.pdf
  - The methodology was to examine the pages on a search engine and attempt to identify the primary language in which the page is written
  - **Conclusions**
    - **From 1996 – 2008 English was predominant, occupying roughly 80% of web pages**
    - **At the same time the number of Internet users who had English as their primary language dropped from 80% to 40%**



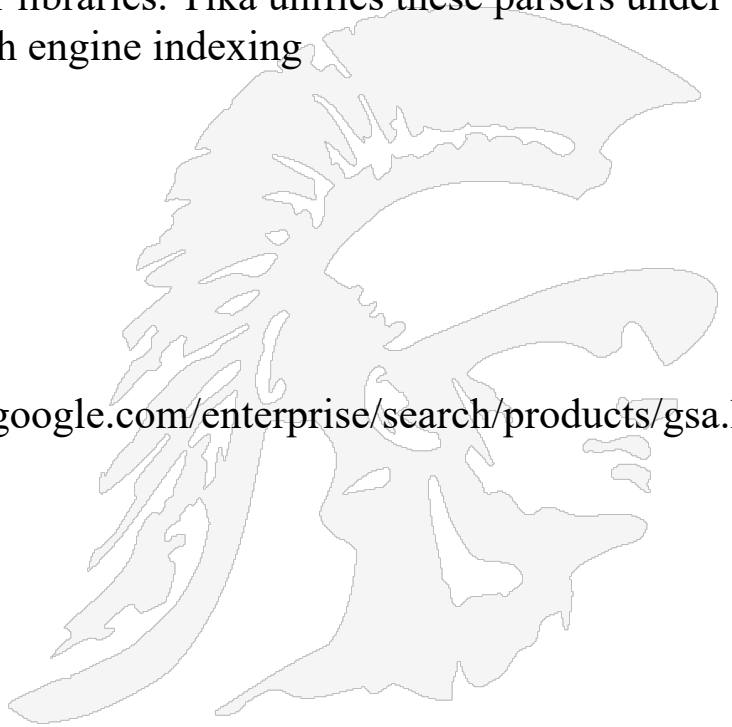
Content languages for websites as of 12 March 2014[2]

Complexity of Data Types



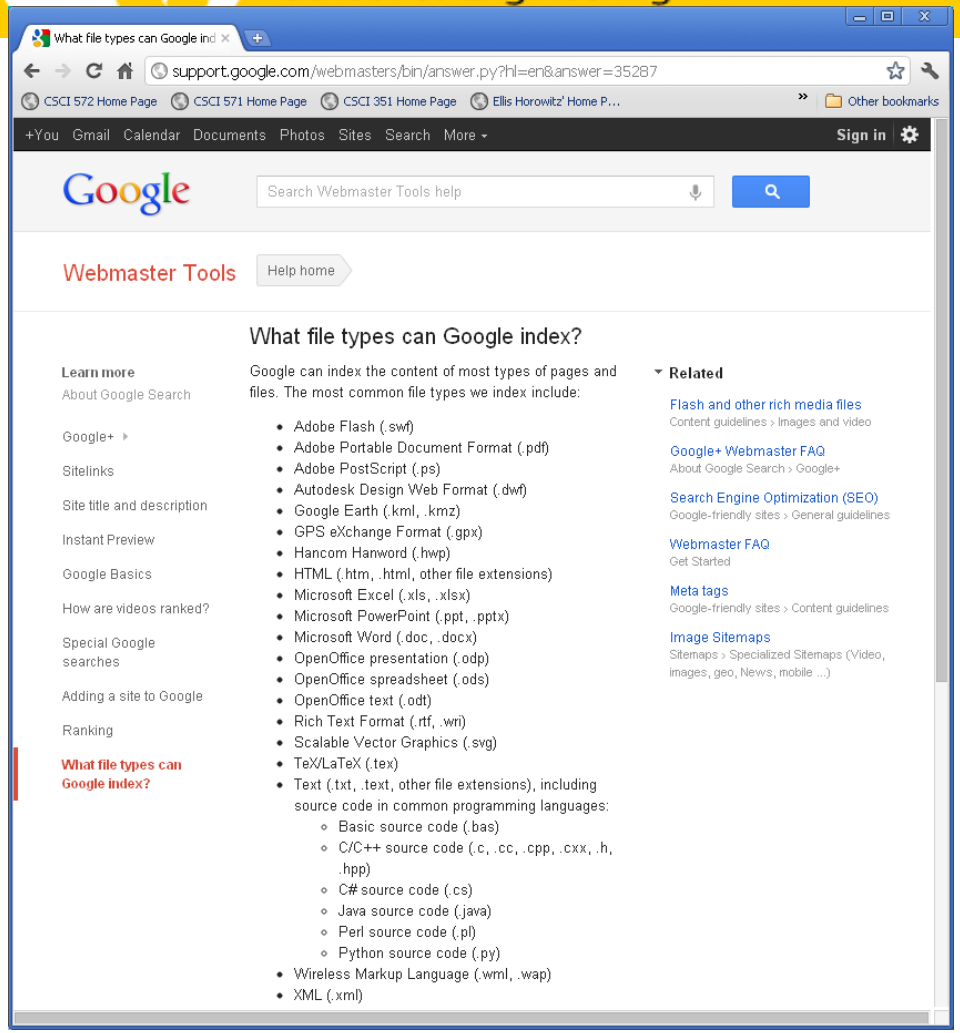
Proliferation of Content Types Available

- **By some accounts, there are 16,000 to 51,000 content types***
- **What to do with content types?**
  - **Parse them**
    - **How?** The Apache Tika™ toolkit detects and extracts metadata and text content from various documents using existing parser libraries. Tika unifies these parsers under a single interface. Tika is useful for search engine indexing
    - Extract their text and structure
  - **Index their metadata**
    - Use an indexing technology like
    - Lucene, http://lucene.apache.org/
    - Solr, http://lucene.apache.org/solr/
    - Google Search Appliance (http://www.google.com/enterprise/search/products/gsa.html)
  - **Identify what language they belong to**
    - N-grams

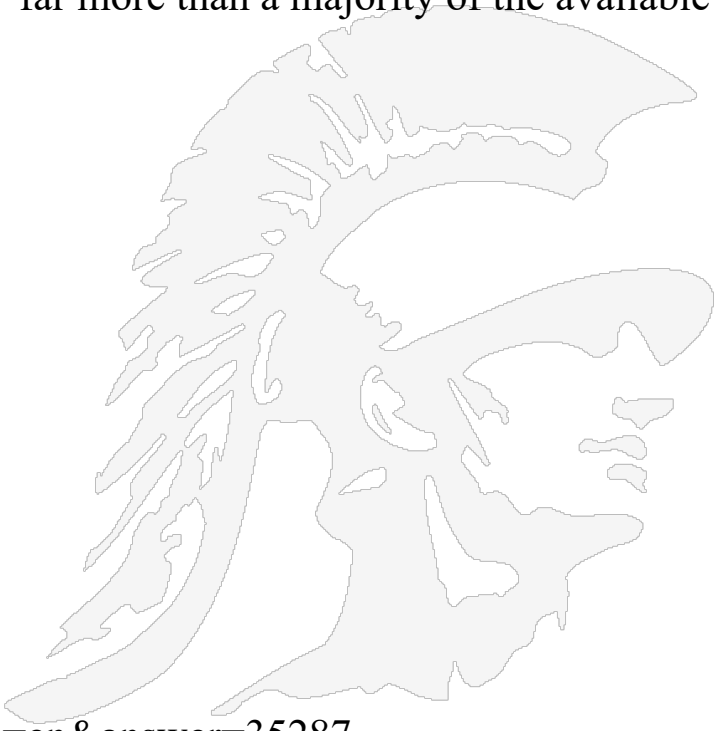


*http://filext.com/ (see if you can name the top 20 file extensions)

Content Types Indexed by Google



Considering the fact that there are thousands of file types of content stored on the web, Google actually indexes only a small number, less than 3 dozen, but they may well constitute far more than a majority of the available content



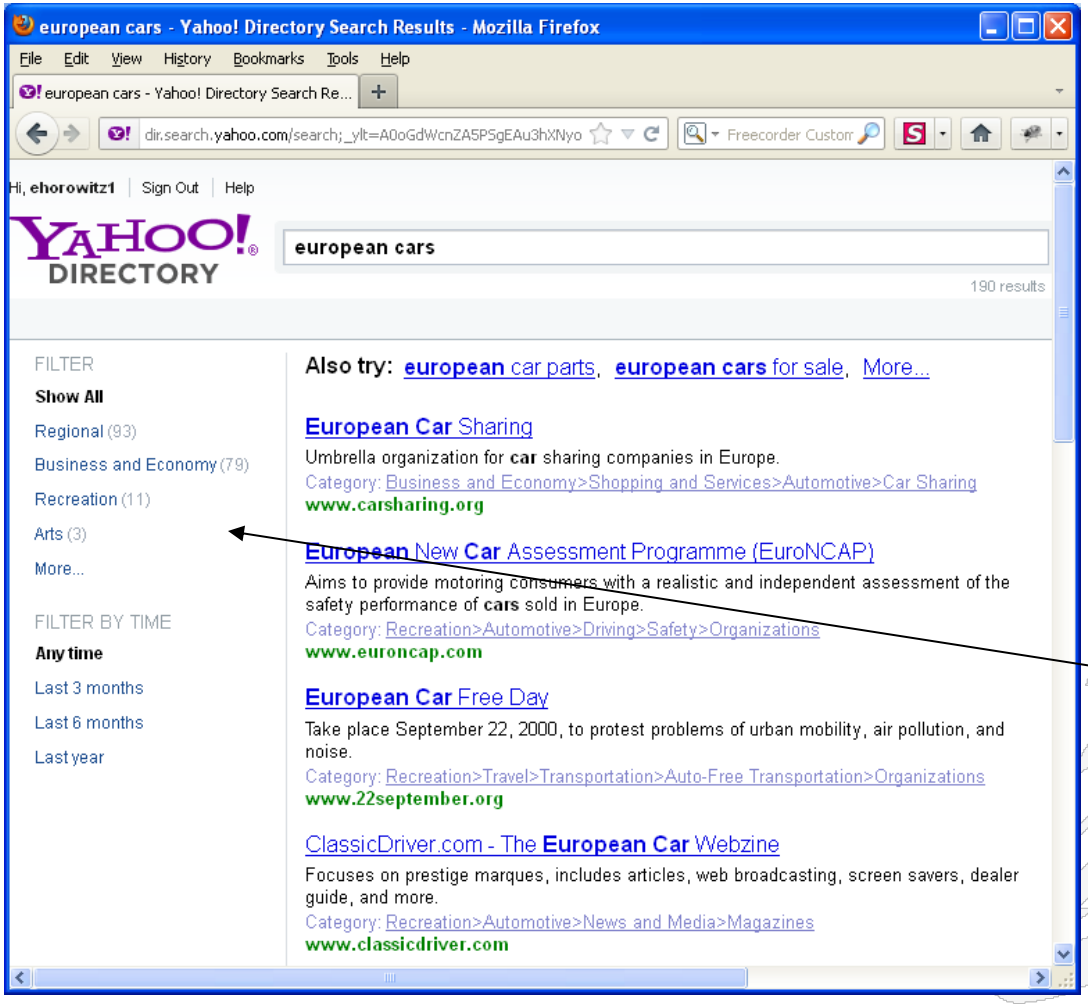
http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35287

A Summary of Some Web Facts

- **How many websites?** *~1.86 billion*
- **How are they distributed across TLDs or across countries?**
  - *112 million out of 148 million belong to .com or about 72%*
- **How many web pages are there?** *30 trillion unique URLs from Google found in 2012,*

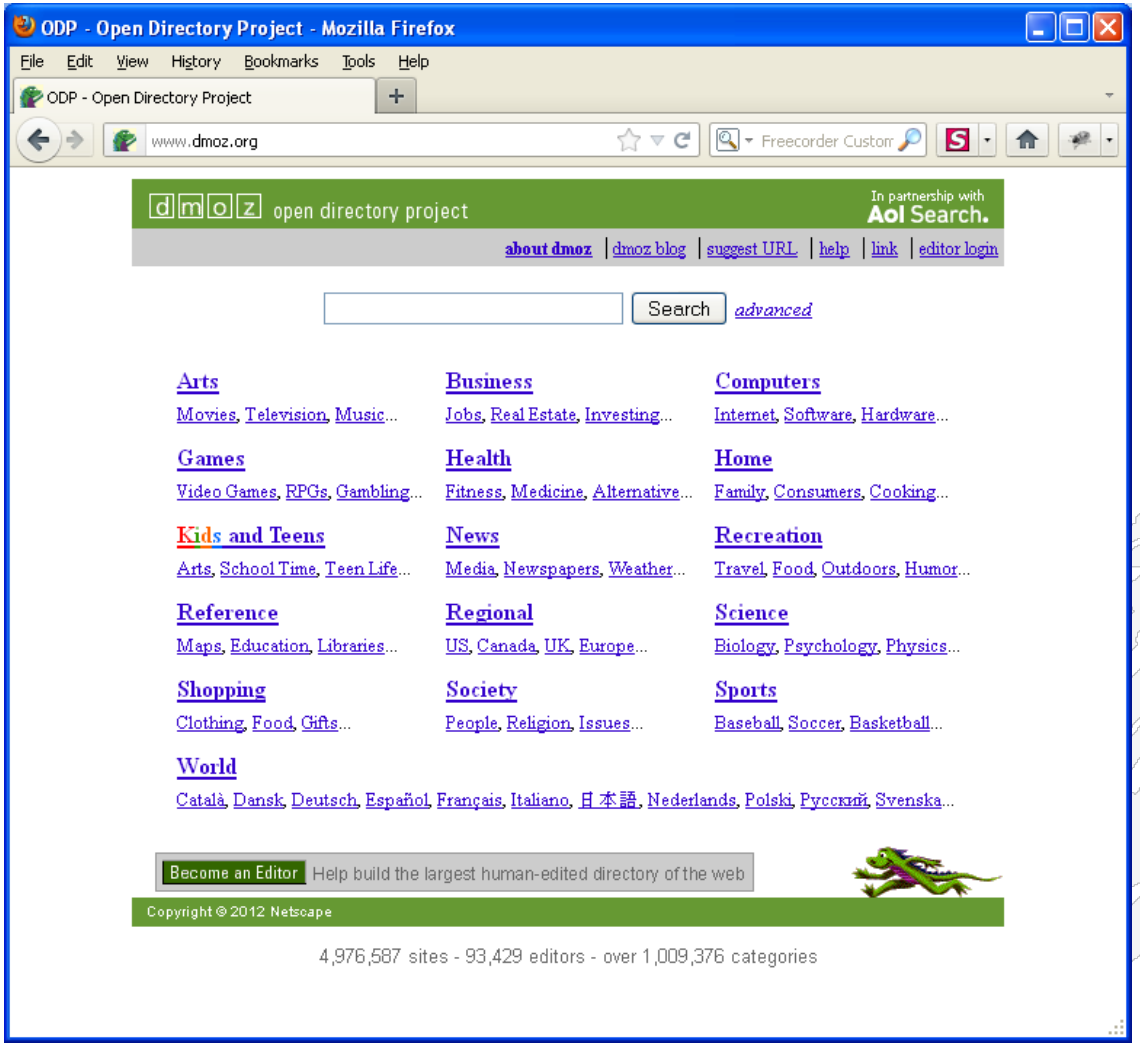
  see **http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html**
- **Which content types hold the most information: HTML, PDF, Word, Excel, PPT, others?**
  - *There are thousands of different content types*
- **How much storage is required to hold a single snapshot of the Web?**
  - *1 trillion web pages at 100K bytes per page requires 100 petabytes*
    - *1 petabyte storage costs under $1,000, so $100,000 of equipment will work*
  - *Google processes 24 petabytes per day, http://en.wikipedia.org/wiki/Petabyte*
  - *The Internet Archive has more than 10 petabytes, http://en.wikipedia.org/wiki/Internet_Archive*
- **What are the languages in which the documents are written?**
  - *According to the Internet Archive, about 55% is in English; other popular languages include: French, German, Spanish and Chinese, also see http://en.wikipedia.org/wiki/Languages_used_on_the_Internet*
- **General properties of the Web graph**
  - *In-degree and out-degree distribution follows a power law*
- **Categories of Content: pornography, spam, mirrors**
  - *Presumably there is a lot of the above, but little concrete data on how much*

Manual Hierarchical Web Taxonomies



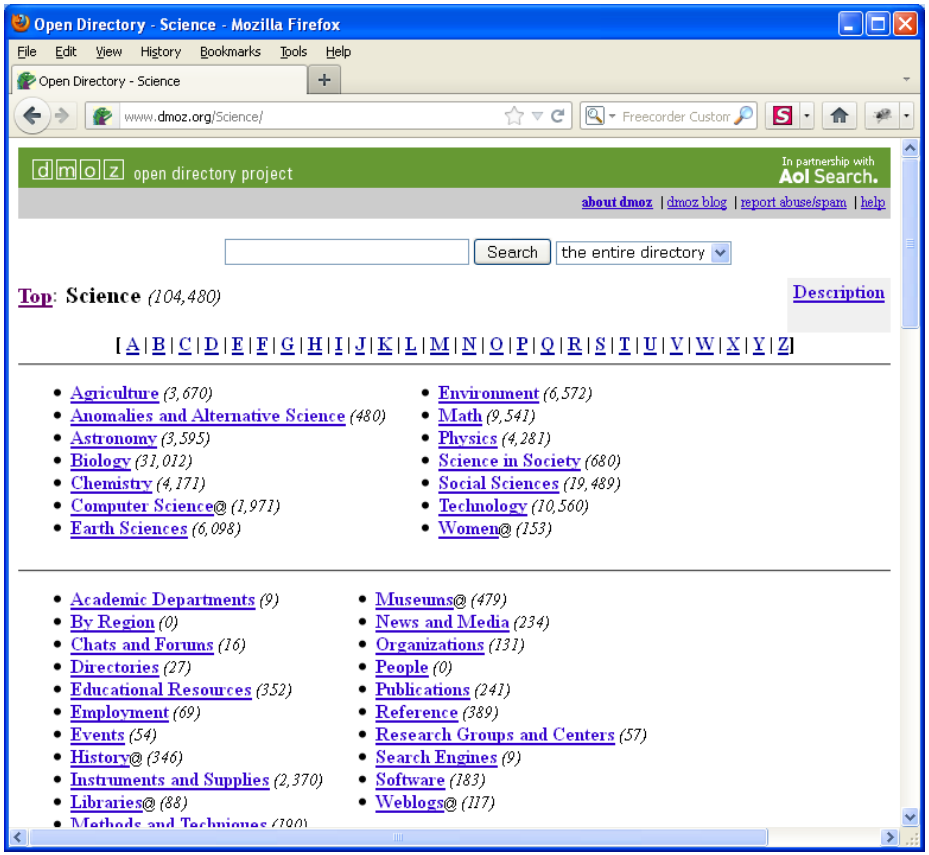
- **Yahoo originally used human editors to assemble a large hierarchically structured directory of web pages.**
  - **http://www.yahoo.com/**
  - **Yahoo still retains the hierarchy as seen to the left; under european cars we see categories: regional with 93 matches, business & economy with 79 matches, etc**

Open Directory Project

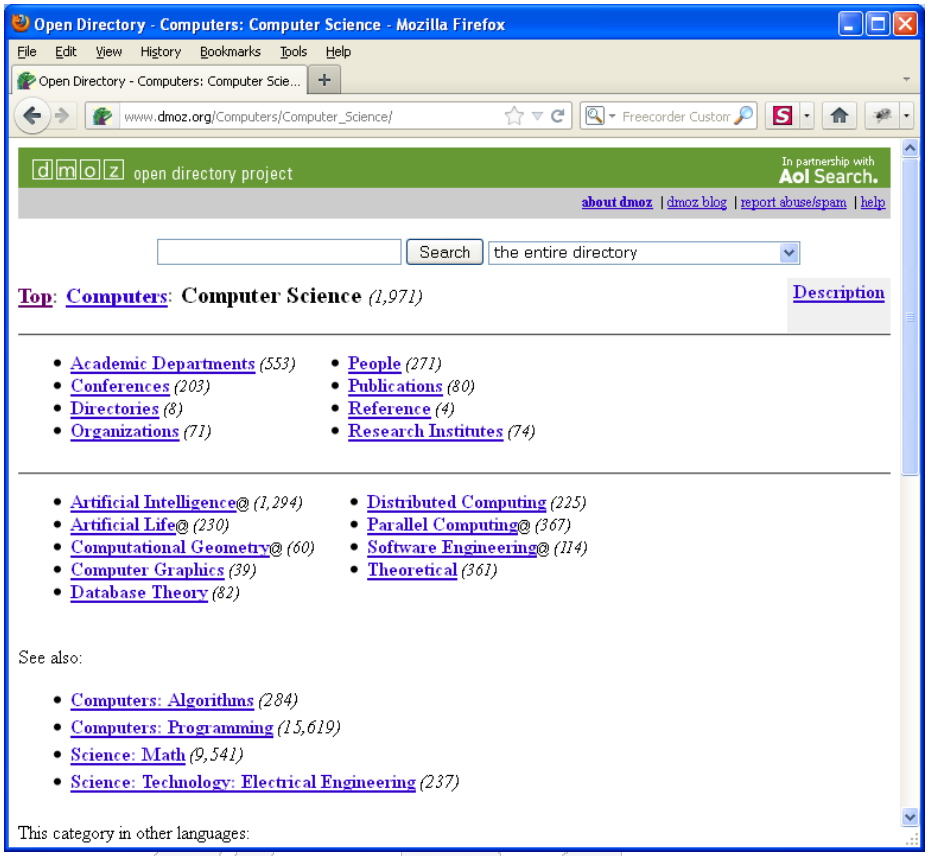


- **Open Directory Project, known as DMoZ, is an effort to organize the web according to an ontology;**
- **An approach similar to Yahoo's;**
- **Based on the distributed labor of volunteer editors ("net-citizens provide the collective brain").**
- **Used by most other search engines.**
- **Started by Netscape.**
  - **http://www.dmoz.org/**
- **Distributes its data using RDF format**
- **DMOZ shut down in 2016**

Drilling Down By Category



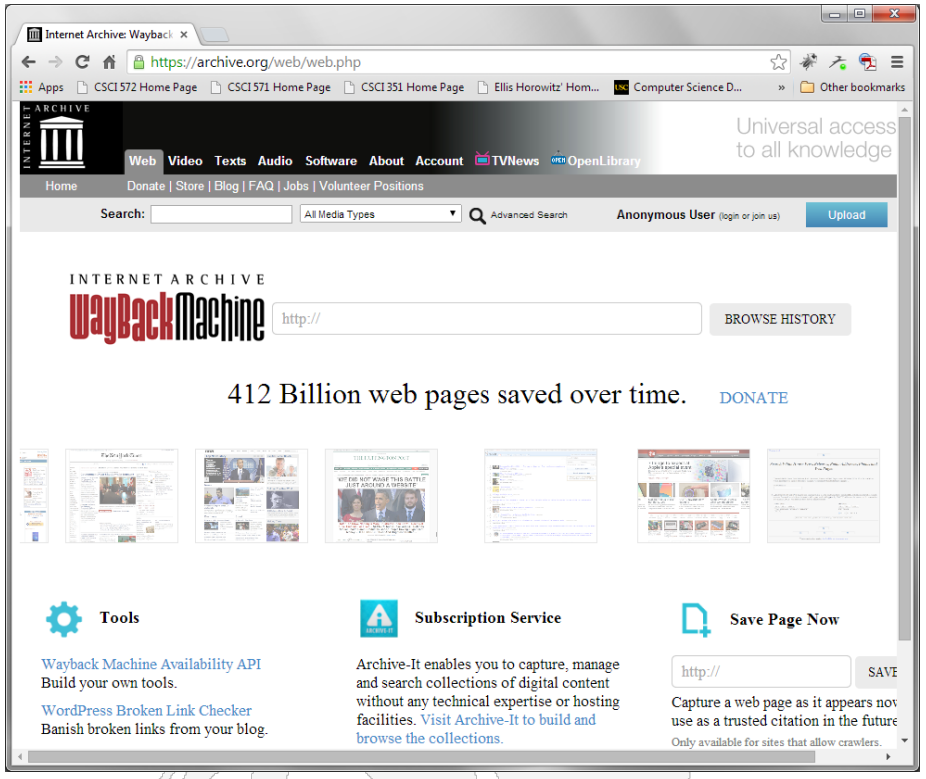
Selecting Category "Science"



Selecting Category "Computer Science"

Internet Archive

- The Internet Archive has been taking a snapshot of the World Wide Web every two months since 1997 – has used **Apache Nutch**
- The results are made available through the Wayback Machine,
- Its database is approximately 4.5 petabytes
- The founder is Brewster Kahle
- For the past 13 years, the Internet Archive has been growing rapidly, most recently by about 100TB of data per month.
- Their crawler surveys the web every two months The algorithm first performs a broad crawl that starts with a few "seed sites," such as Yahoo's directory. After snapping a shot of the home page, it then moves to any referable pages within the site until there are no more pages to capture. If there are any links on those pages, the algorithm automatically opens them and archives that content as well.



Surface Web


SURFACE WEB
Google
Bing
Wikipedia
DEEP WEB
Academic Information
Medical Records
Legal Documents
Scientific Reports
Subscription Information
Contains 90% of the information on the Internet, but is not accessible by Surface Web crawlers.
Social Media
Multilingual Databases
Financial Records
Government Resources
Competitor Websites
Organization-specific Repositories
(DARK WEB)
A part of the Deep Web accessible only through certain browsers such as Tor designed to ensure anonymity. Deep Web Technologies has zero involvement with the Dark Web.
Illegal Information
TOR-Encrypted sites
Political Protests
Drug Trafficking sites
Private Communications