Web Trends and Measurements

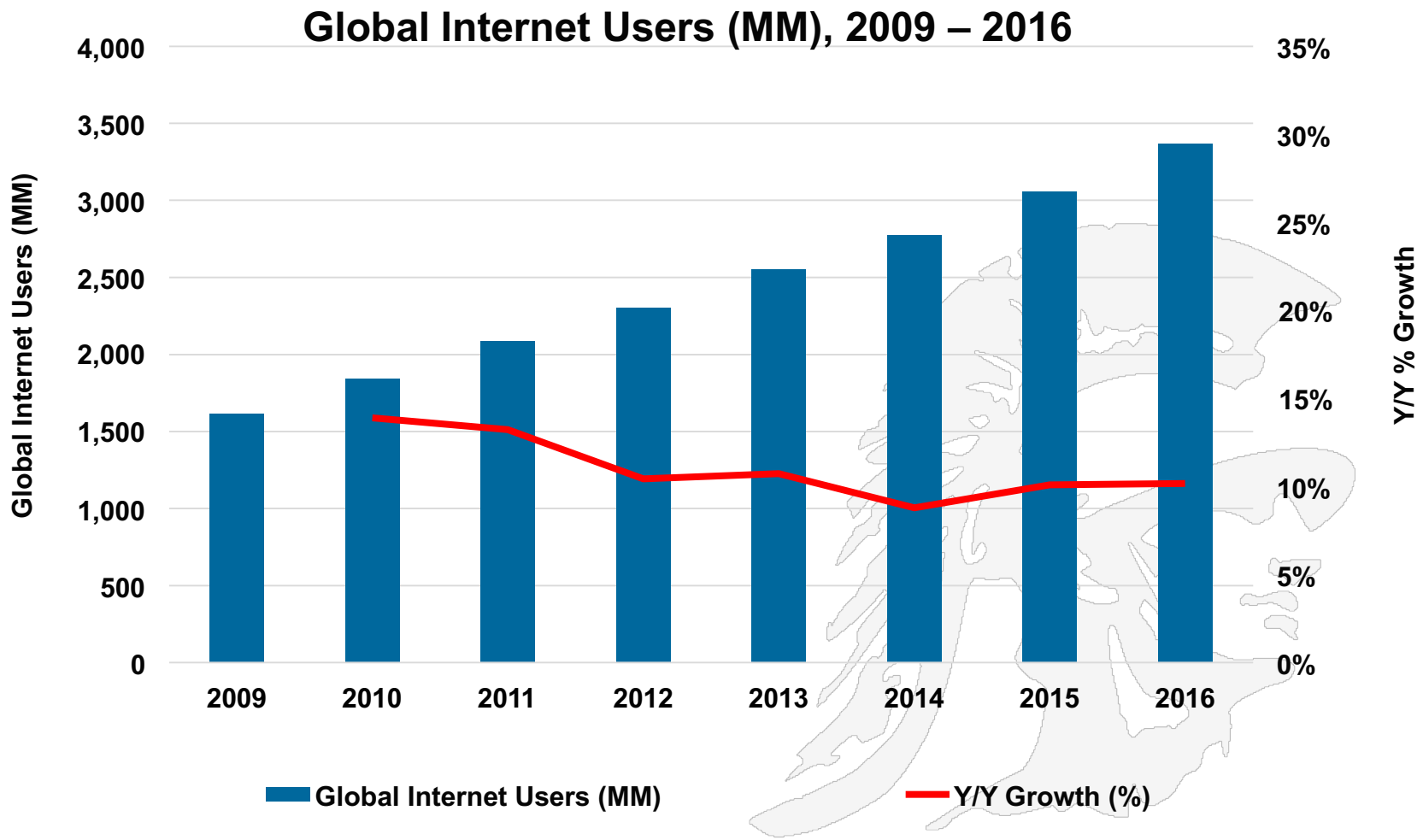
Copyright © Ellis Horowitz 1999-2019

Web Trends

- Web has changed dramatically over the last 20+ 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 slides come from Kleiner, Perkins, Caufield and Byers, a major venture capital firm, http://www.kpcb.com/

Copyright © Ellis Horowitz 1999-2019

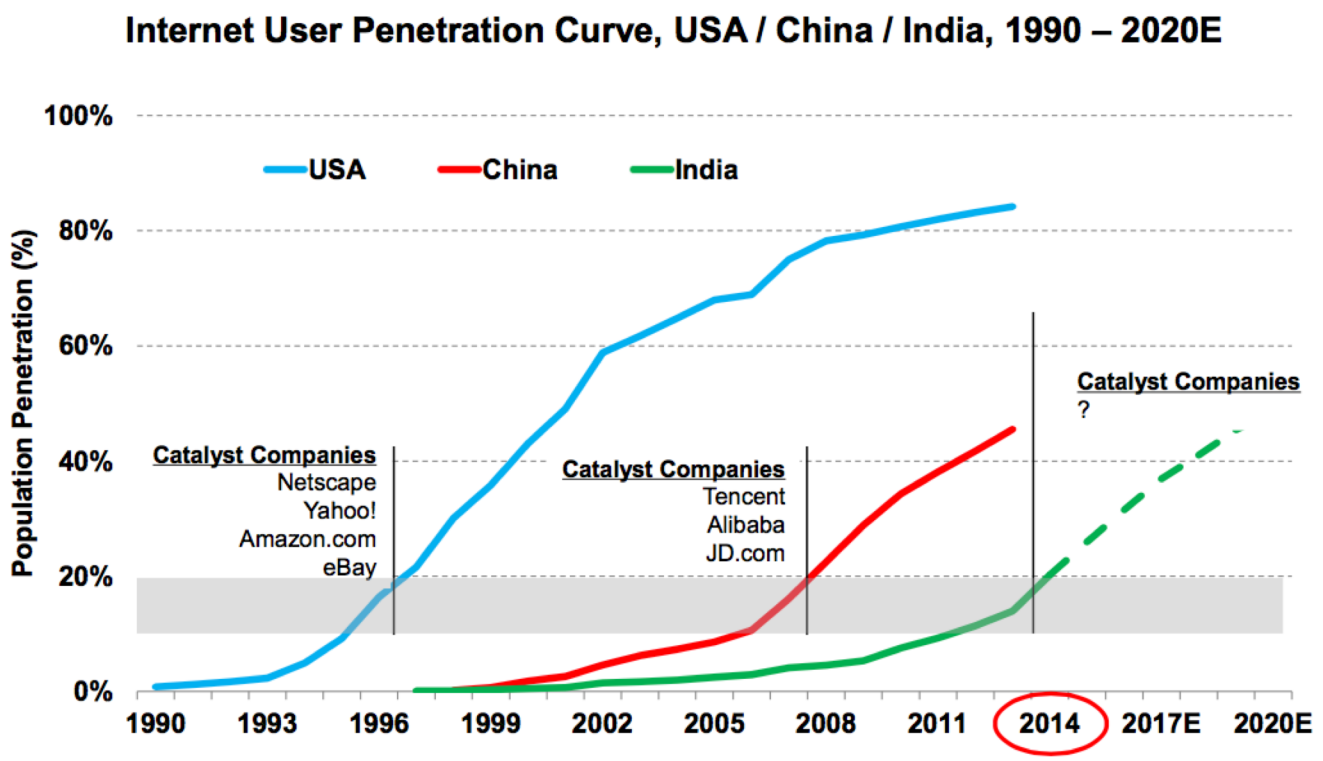
Global Internet Users = 3.4B @ 46% Penetration...



Source: United Nations / International Telecommunications Union, US Census Bureau. Internet user data is as of mid-year. Internet user data for: USA from Pew Research, China from CNNIC, Iran from Islamic Republic News Agency / InternetWorldStats / KPCB estimates, India from KPCB estimates based on IAMAI data, Indonesia from APJII.

KLEINER PERKINS
Copyright Ellis Horowitz, 2011-2019

India = Appears to Be @ Internet Penetration Growth Inflection



@KPCB Source: World Bank, Hillhouse Capital forecast for India beyond 2014.

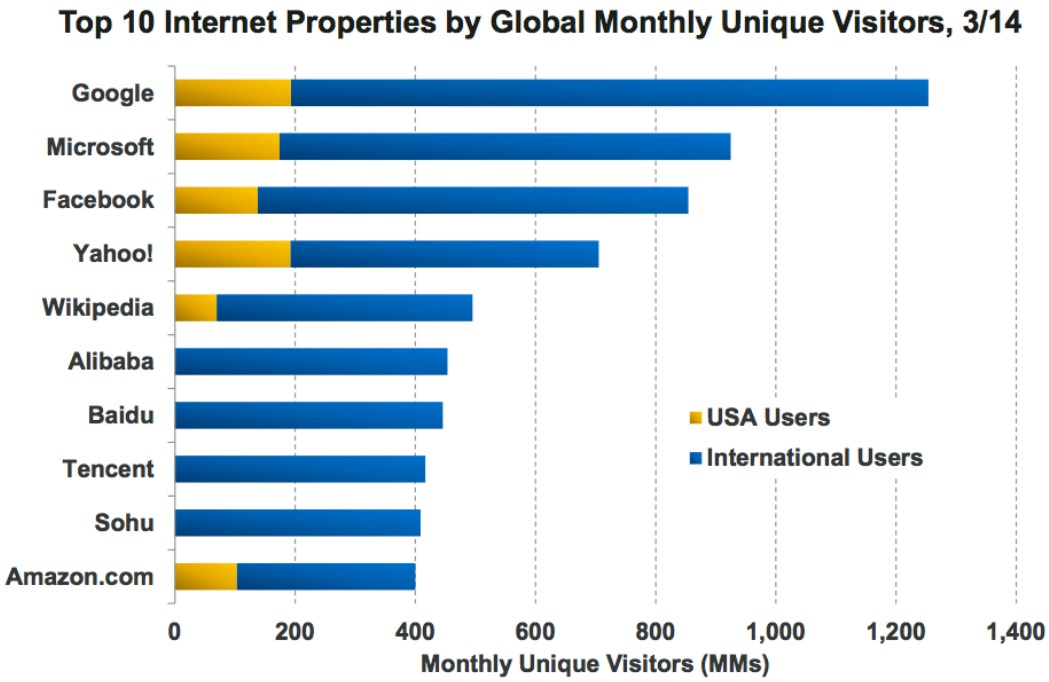
Hillhouse Capital
165

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

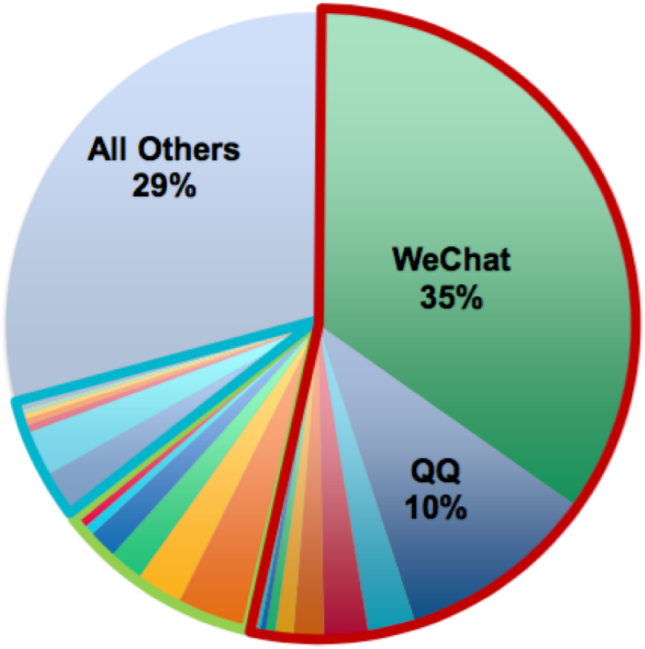


@KPCB Source: comScore, 3/14.

131

China Mobile Internet 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

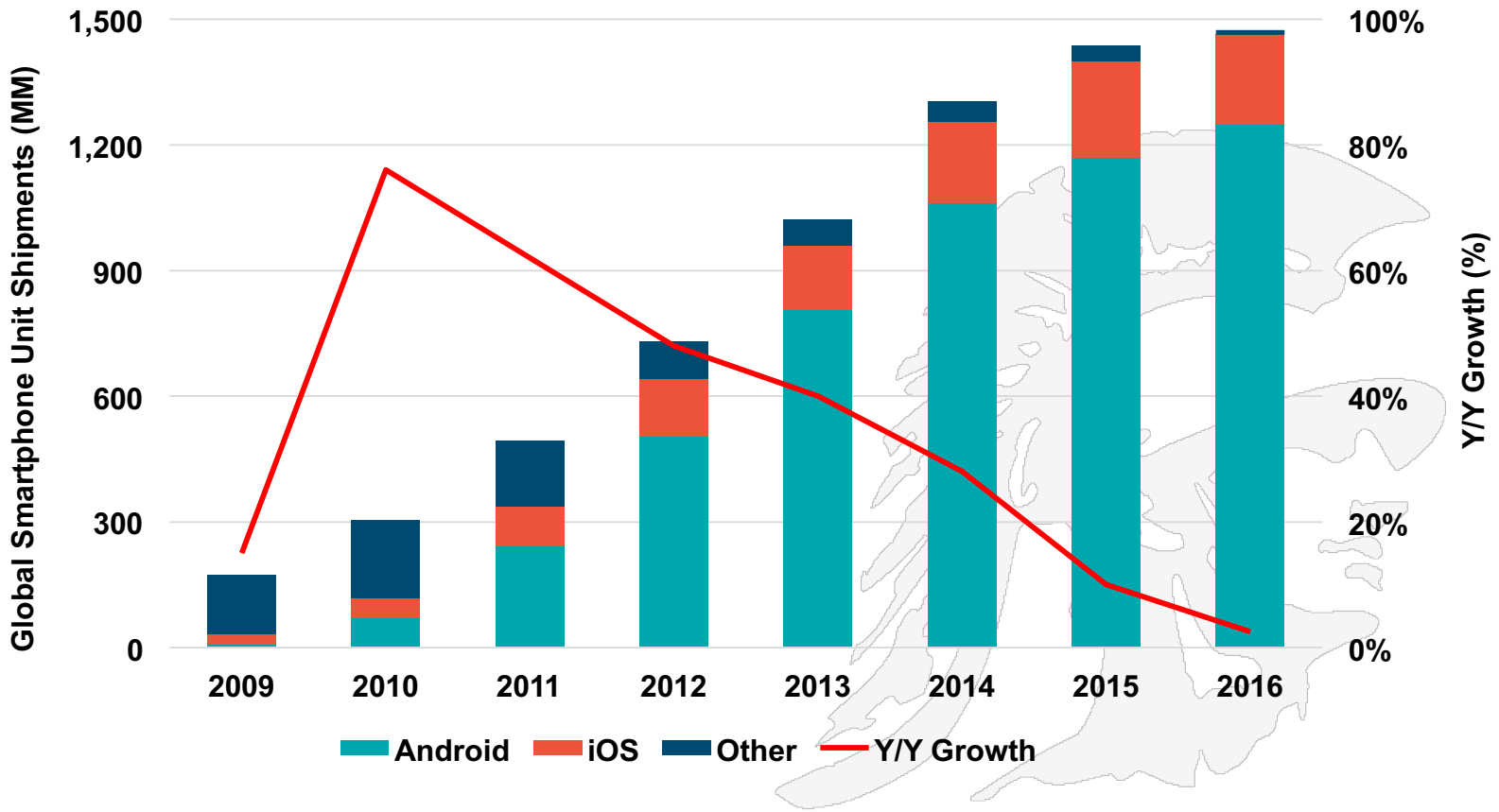
@KPCB

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 Unit Shipments = Continue to Slow...

Smartphone Unit Shipments by Operating System (MM), Global, 2009 – 2016



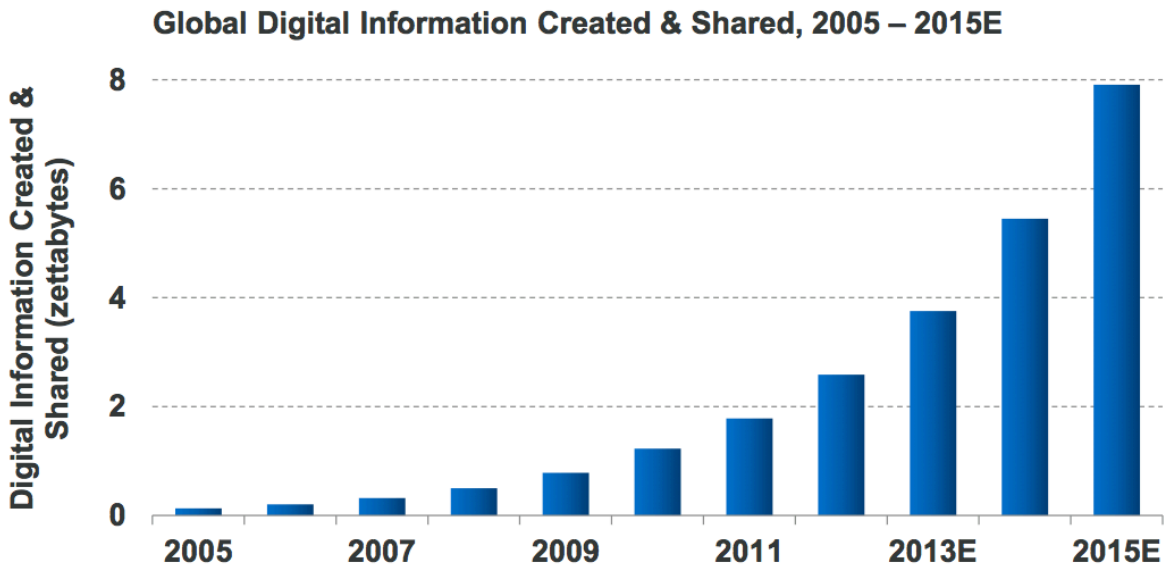
Source: Morgan Stanley Research (5/17)

Copyright Ellis Horowitz, 2011-2019

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.

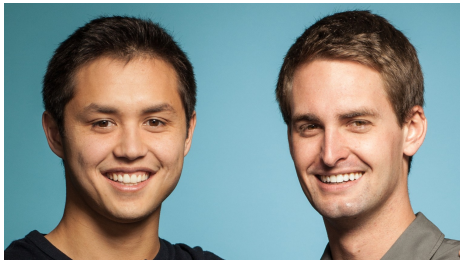
Photos Alone = 1.8B+ Uploaded & Shared Per Day...
Growth Remains Robust as New Real-Time Platforms Emerge

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

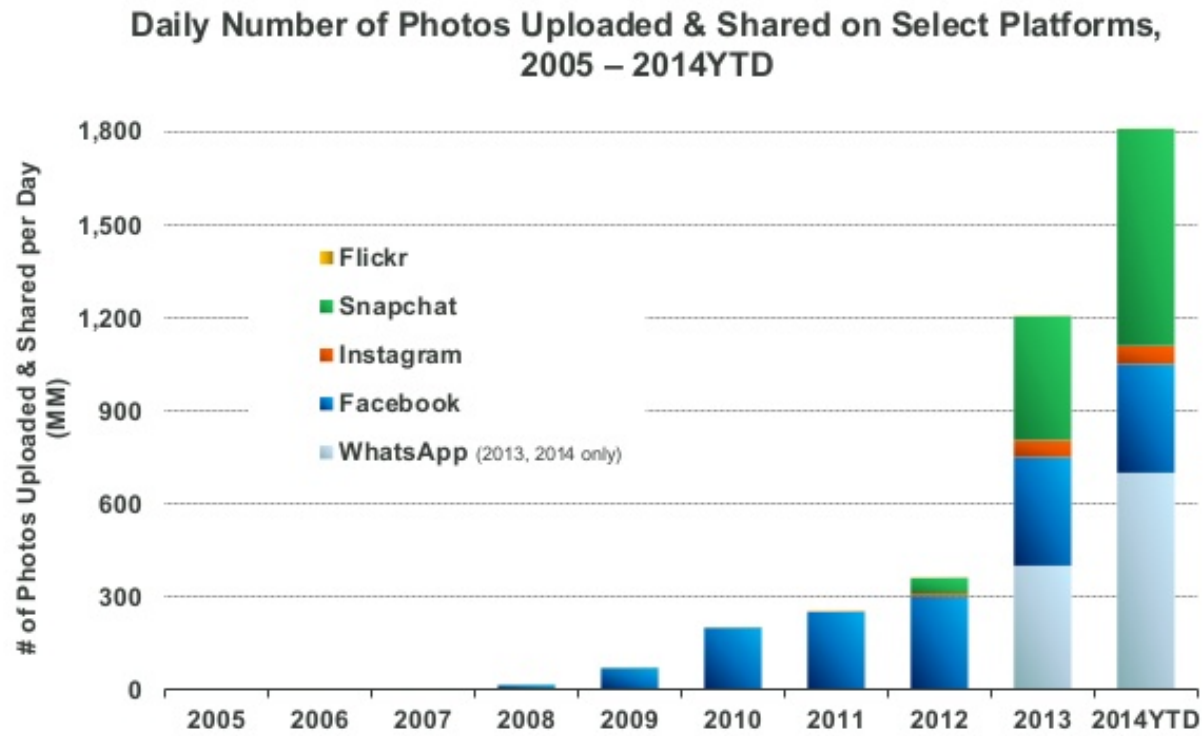
Yahoo has recently made a major upgrade to **Flickr**

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

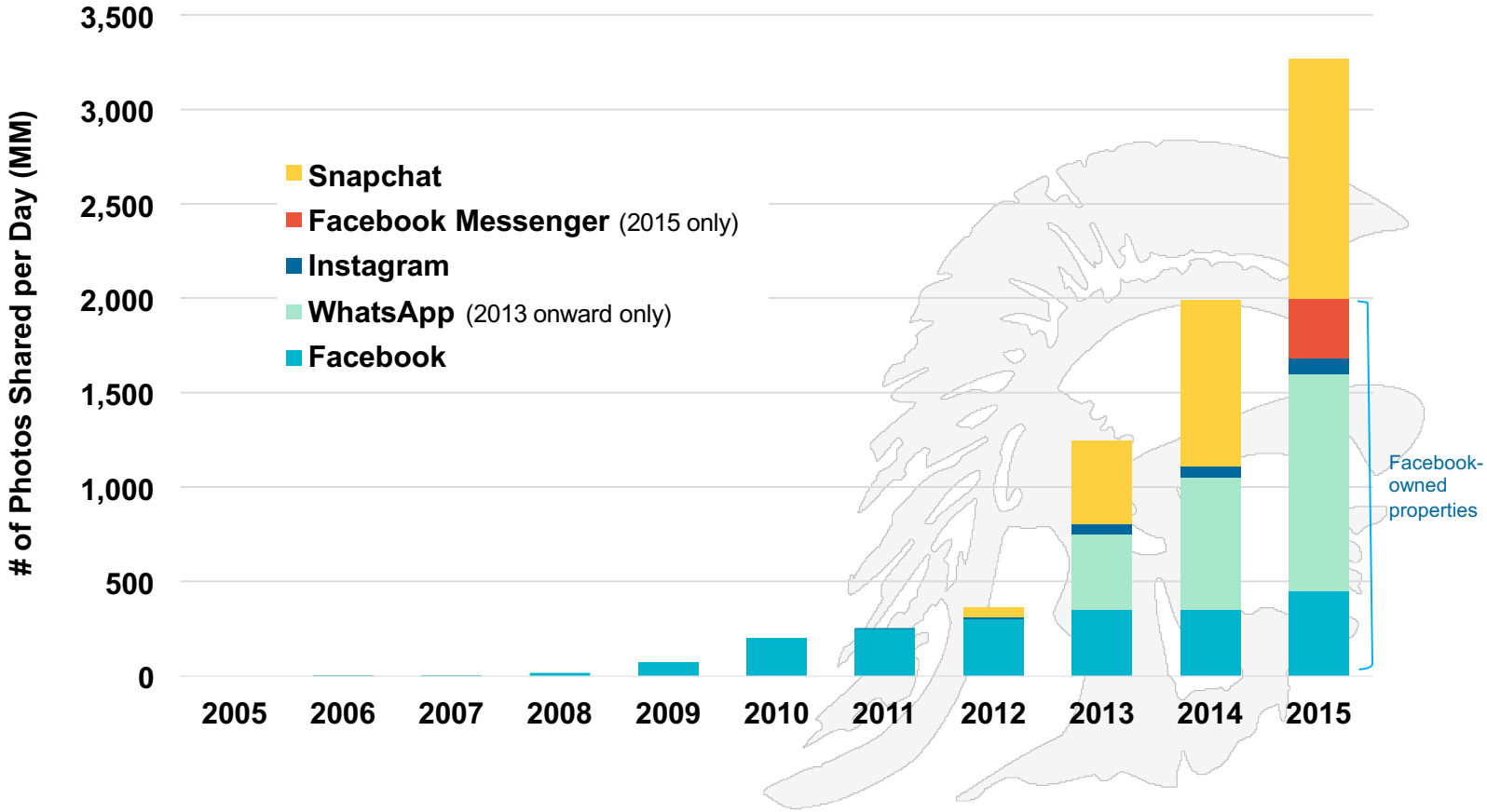


@KPCB Source: KPCB estimates based on publicly disclosed company data, 2014 YTD data per latest as of 5/14.

62

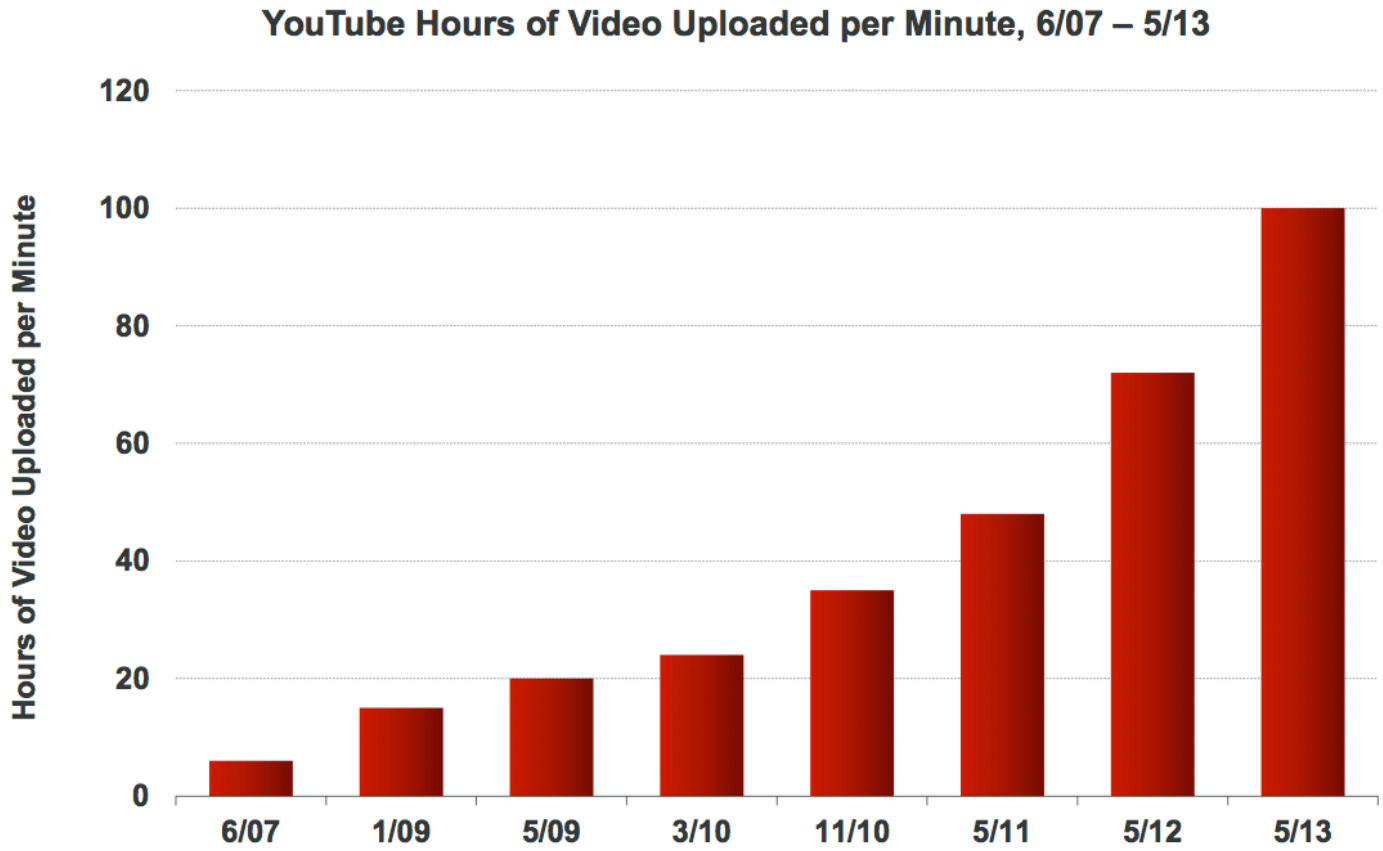
Image Growth Remains Strong

Daily Number of Photos Shared on Select Platforms, Global, 2005 – 2015



Copyright Ellis Horowitz, 2011-2019

Video = 100 Hours Per Minute Uploaded to YouTube, Up from ~Nada Six Years Ago



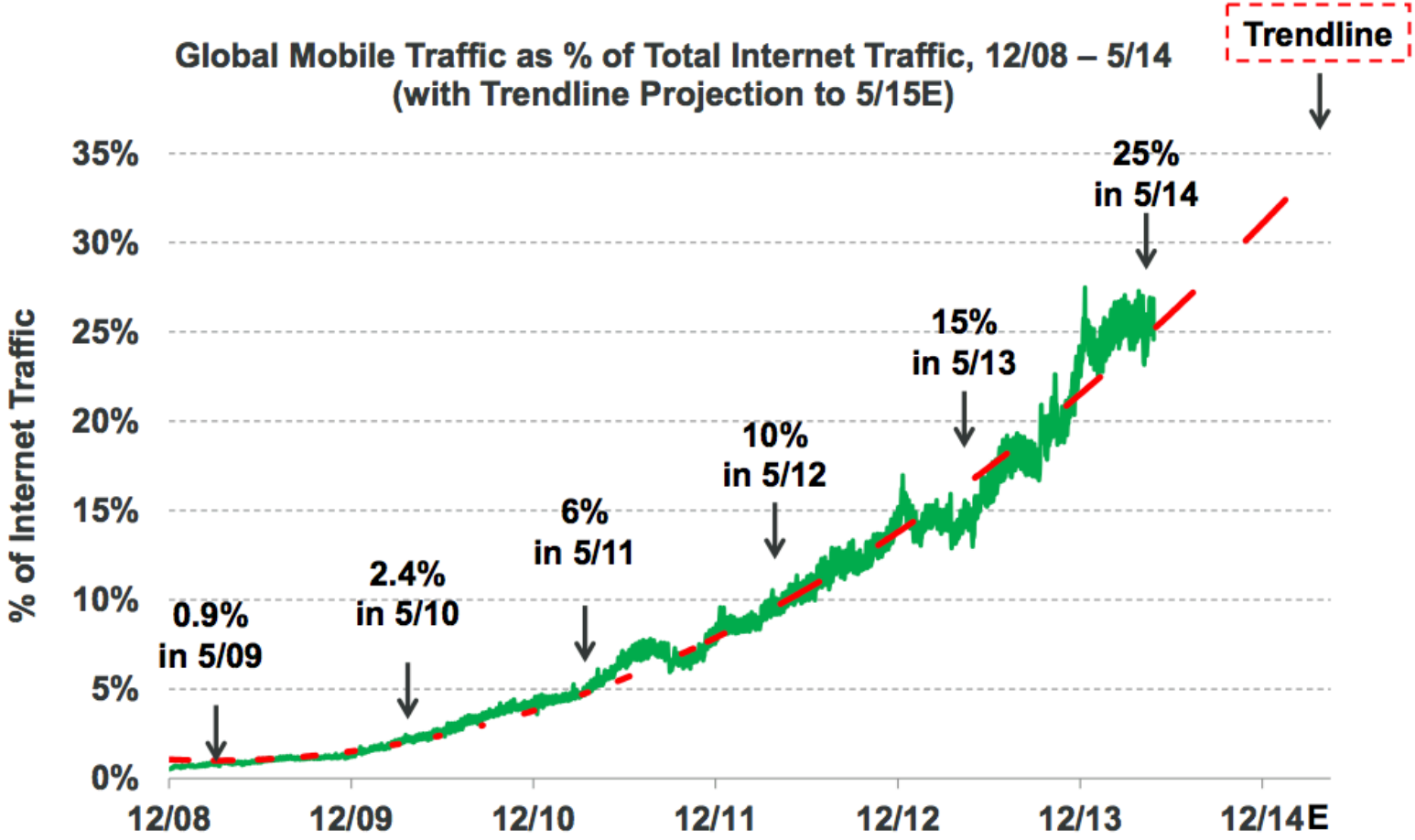
KPCB

As of 2018 300 hours of video are uploaded every minute

Source: YouTube.

17

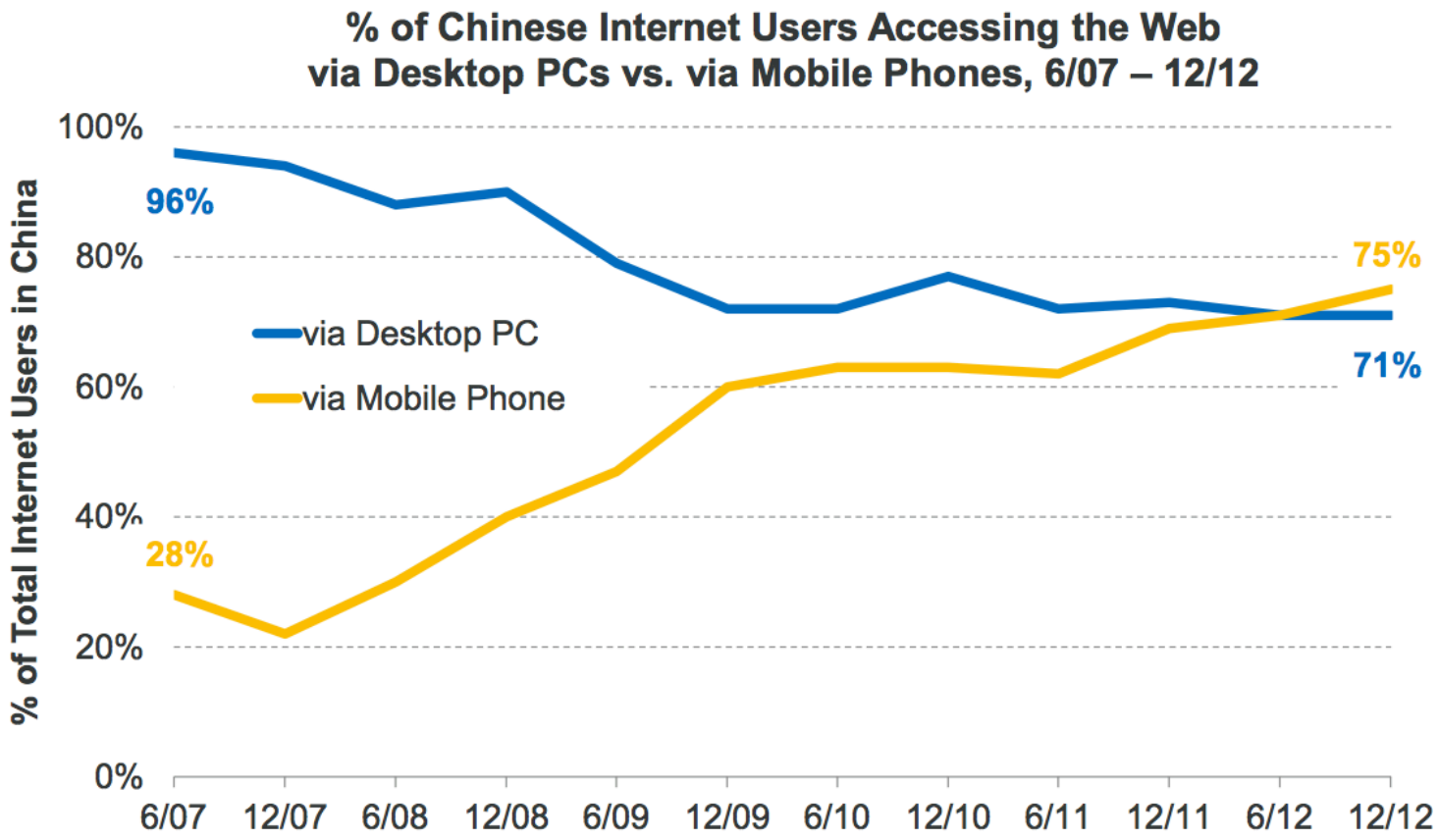
Mobile Traffic as % of Global Internet Traffic = Growing >1.5x per Year & Likely to Maintain Trajectory or Accelerate



@KPCB Source: StatCounter Global Stats, 5/14. Note that PC-based Internet data bolstered by streaming.

159

China –
Mobile Internet Access Surpassed PC, Q2:12

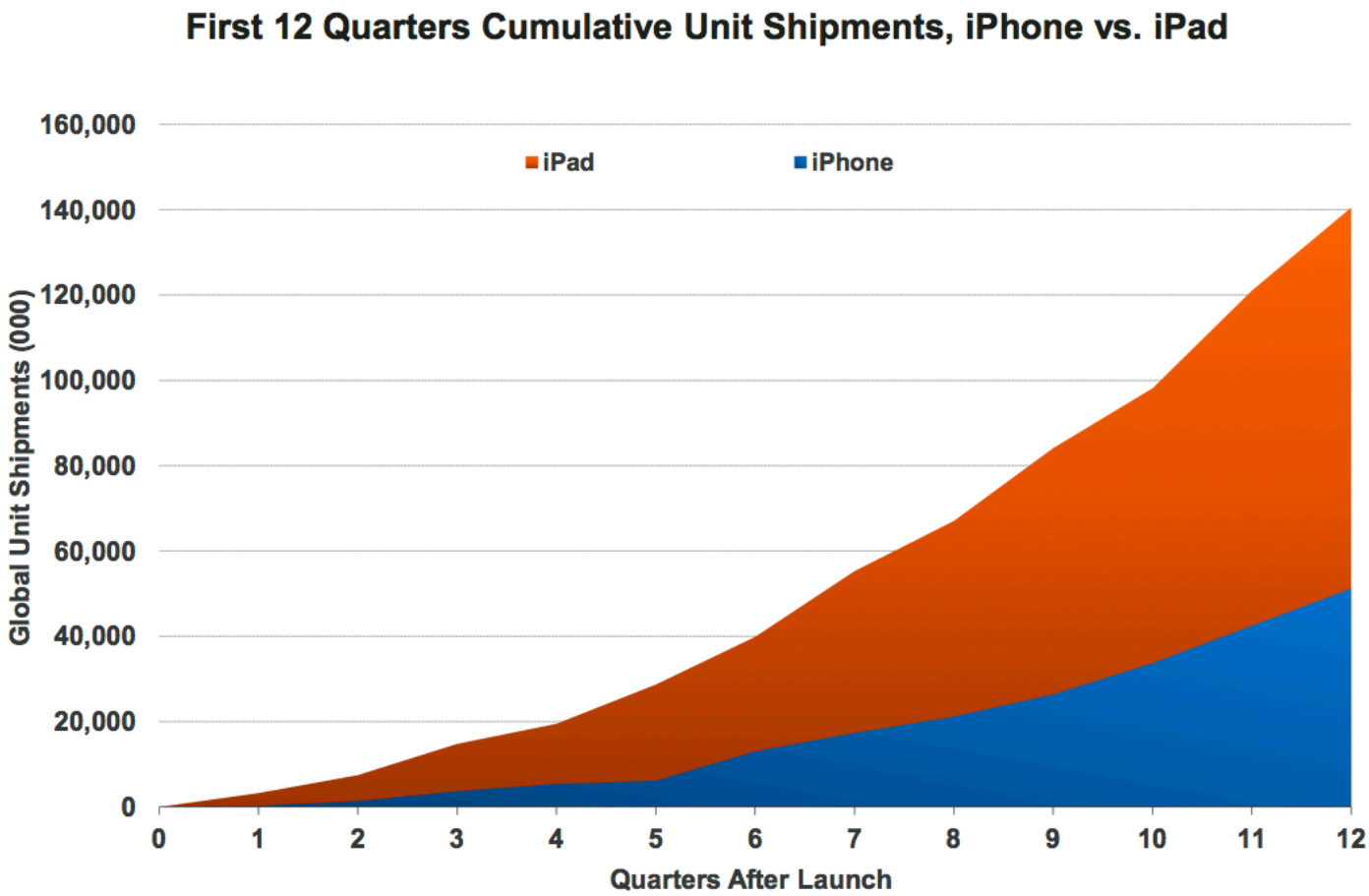


KPCB

Source: CNNIC, 1/13.

33

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



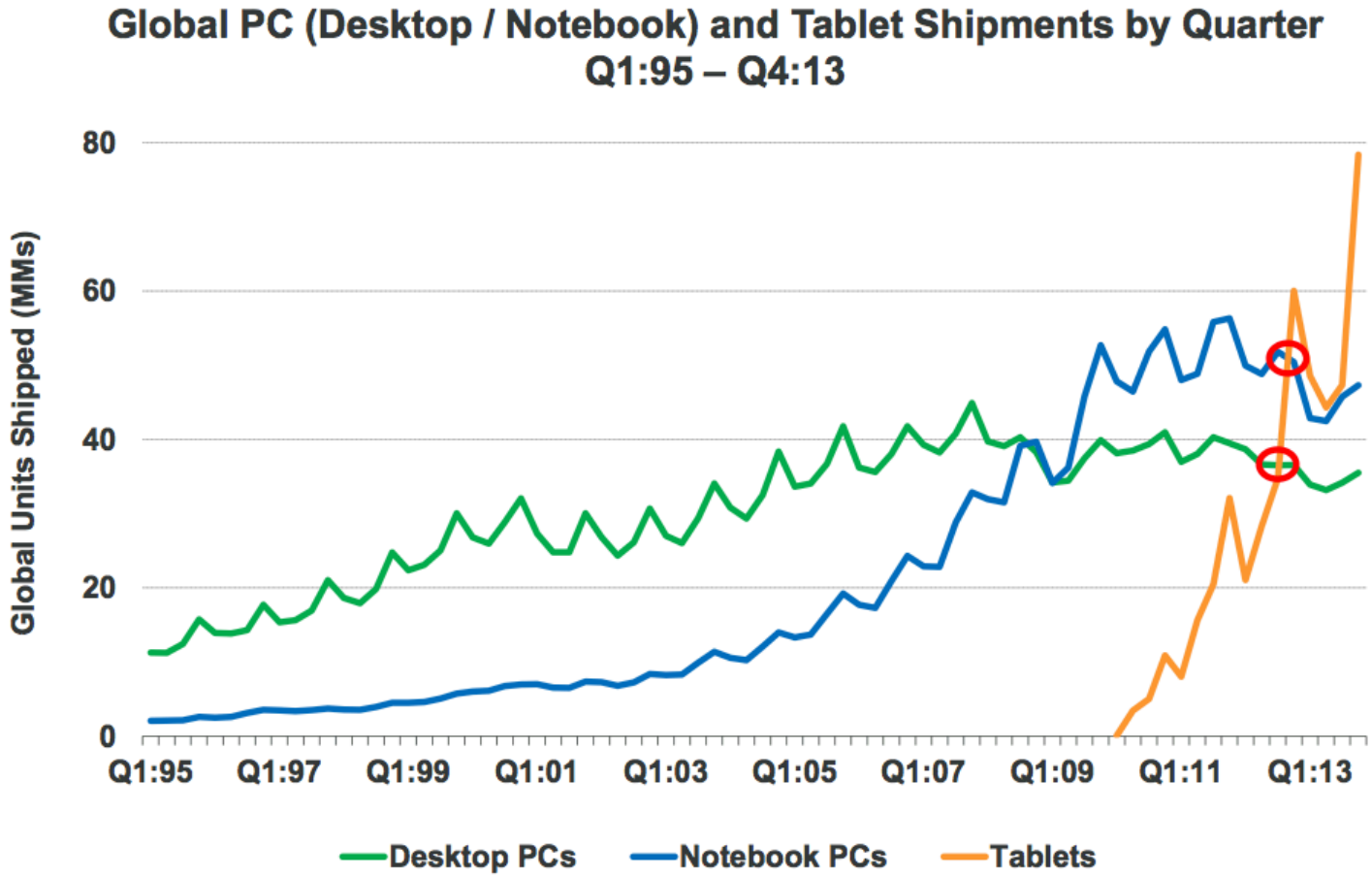
KPCB

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

44

Copyright © Ellis Horowitz 2019

Tablet Units = Growing Faster Than PCs Ever Did... +52%, 2013



@KPCB Source: Morgan Stanley Research. Note: Notebook PCs include Netbooks.

7

Copyright © Ellis Horowitz 2019

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

Mainframe Computing 1960s	*Mini* Computing 1970s	*Personal* Computing 1980s	*Desktop Internet* Computing 1990s	*Mobile Internet* Computing 2000s	*Wearable / Everywhere* Computing 2014+

Others?

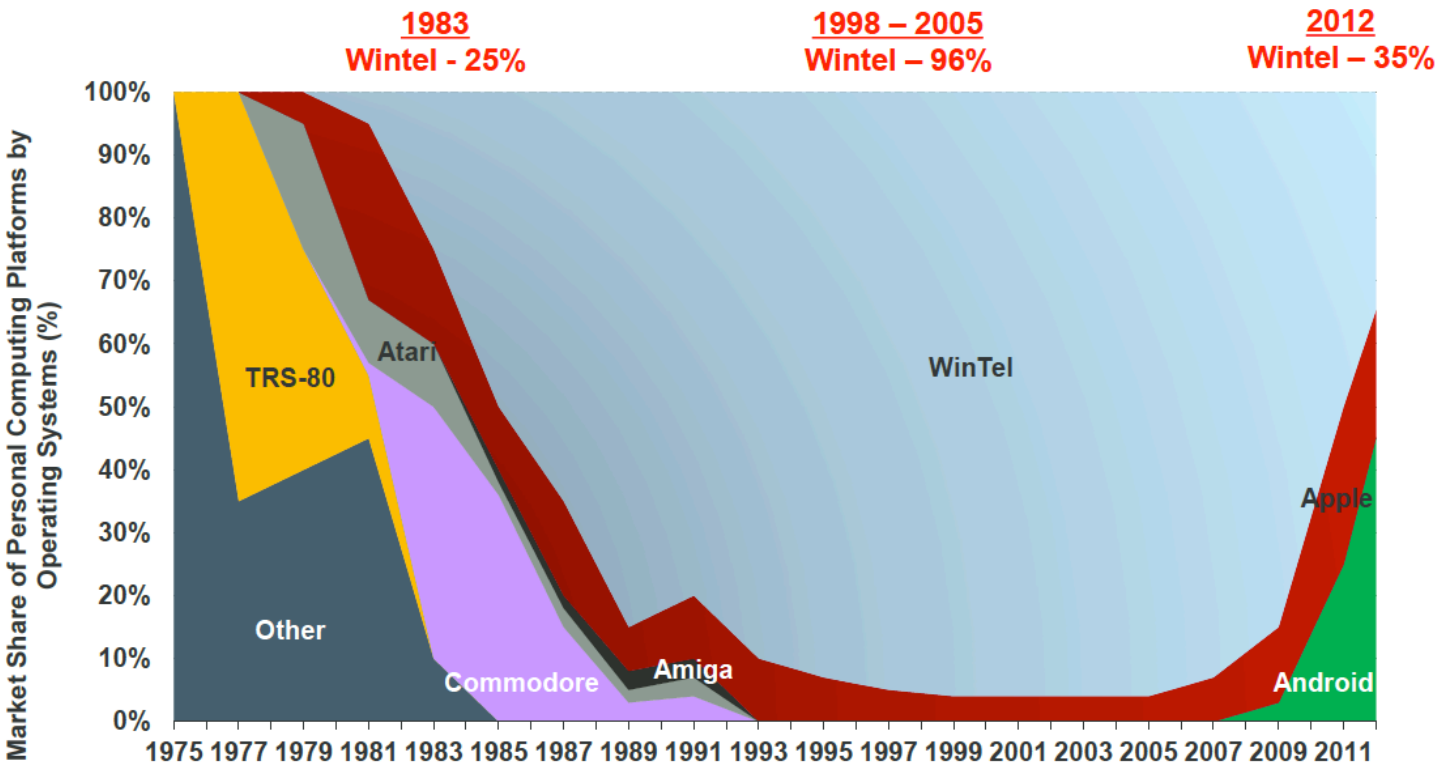
KPCB

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

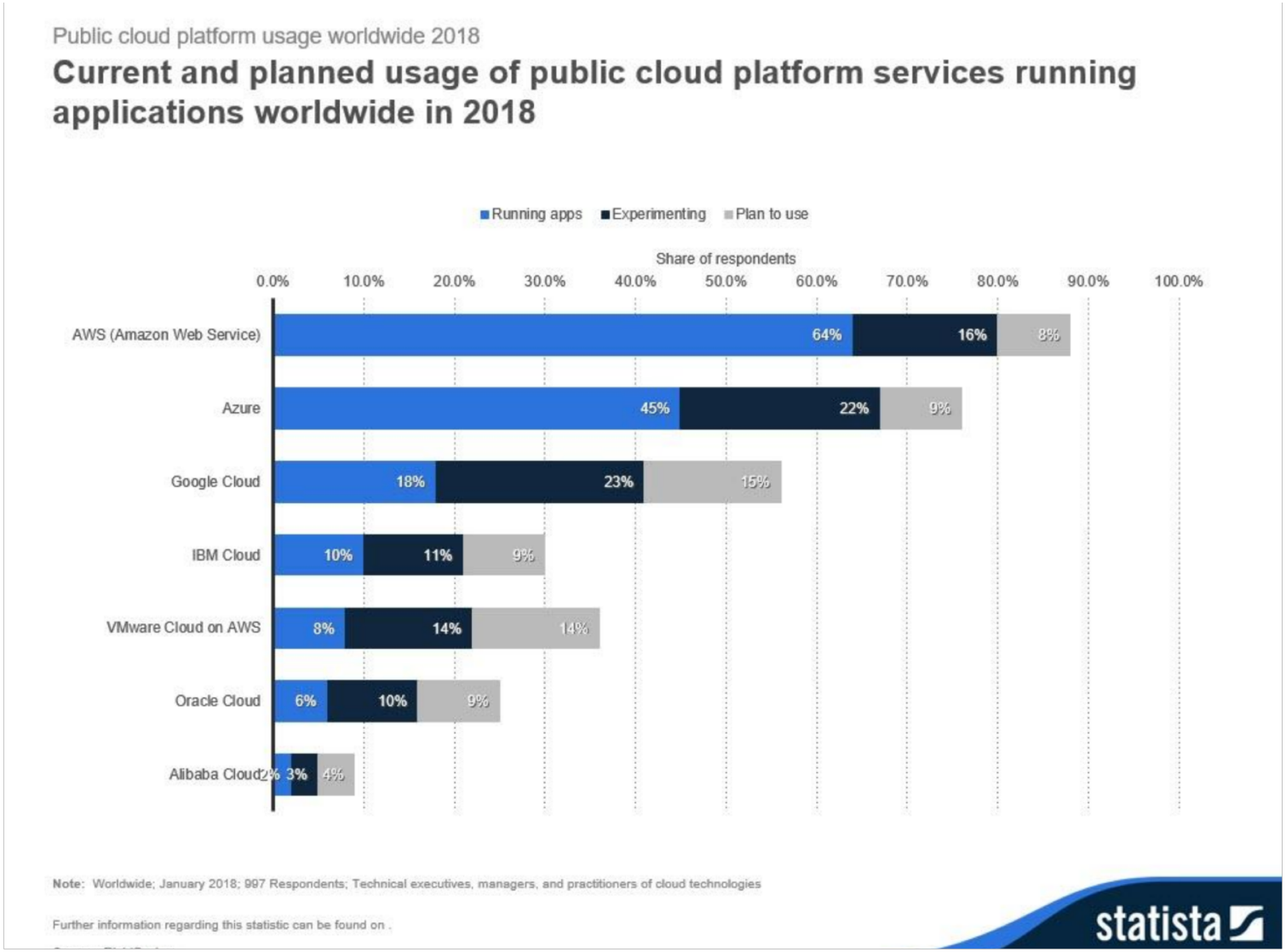


KPCB

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

109

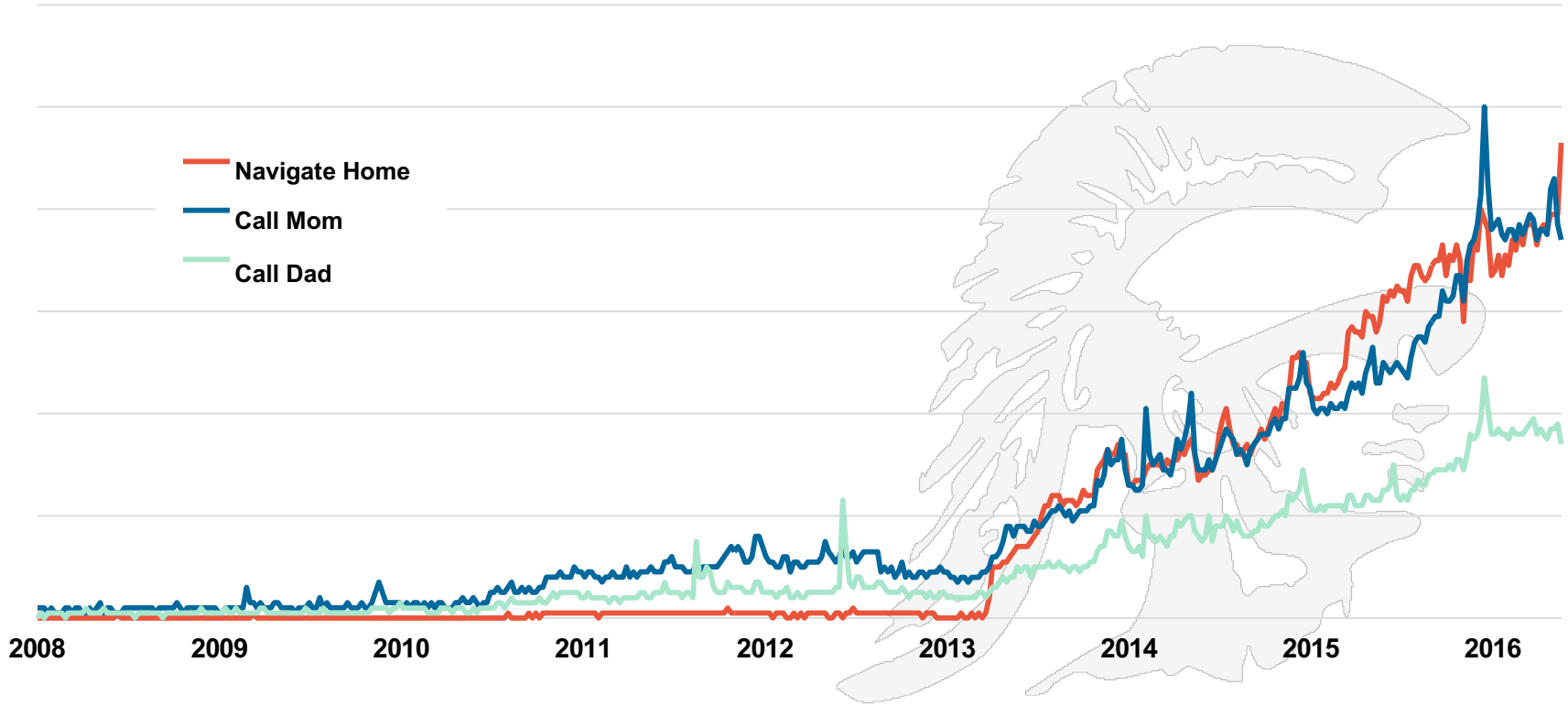
Major Cloud Providers



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 – 2016



Copyright Ellis Horowitz, 2011-2019

Summary of Recent Trends in Web/Internet Development

- Growth in number of users connected
- Growth in Smartphone use, particularly iOS and Android
- Growth in digital data, especially photos and video
- Growth in Social Media
- Growth in Internet use from Mobile over desktop/laptop
- Growth in tablet usage over desktops/laptops
- Decreased dominance of Microsoft Windows

Copyright © Ellis Horowitz 1999-2019

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)

Copyright Ellis Horowitz, 2011-2019

A Summary of Some Web Facts

- **How many websites?** *~1.7 billion (see Netcraft survey)*
- **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**
- **How many pages of different content types: HTML, PDF, Word, Excel, PPT, etc?**
  - *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*
  - *Google processes 24 petabytes per day, http://en.wikipedia.org/wiki/Petabyte*
  - *The Internet Archive has 10 petabytes as of 10/2012, http://en.wikipedia.org/wiki/Internet_Archive*
  - *1 petabyte storage costs under $1,000*
- **What are the languages in which the documents are written?**
  - *As of 04/2013, about 55% is in English; other popular languages include: Russian, German, Chinese, see http://en.wikipedia.org/wiki/Languages_used_on_the_Internet*
- **How often are web pages changing**
  - *Studies done in 1999, 2002, 2004 are now old and out-of-date*
- **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*

Copyright Ellis Horowitz, 2011-2019

Number of Websites

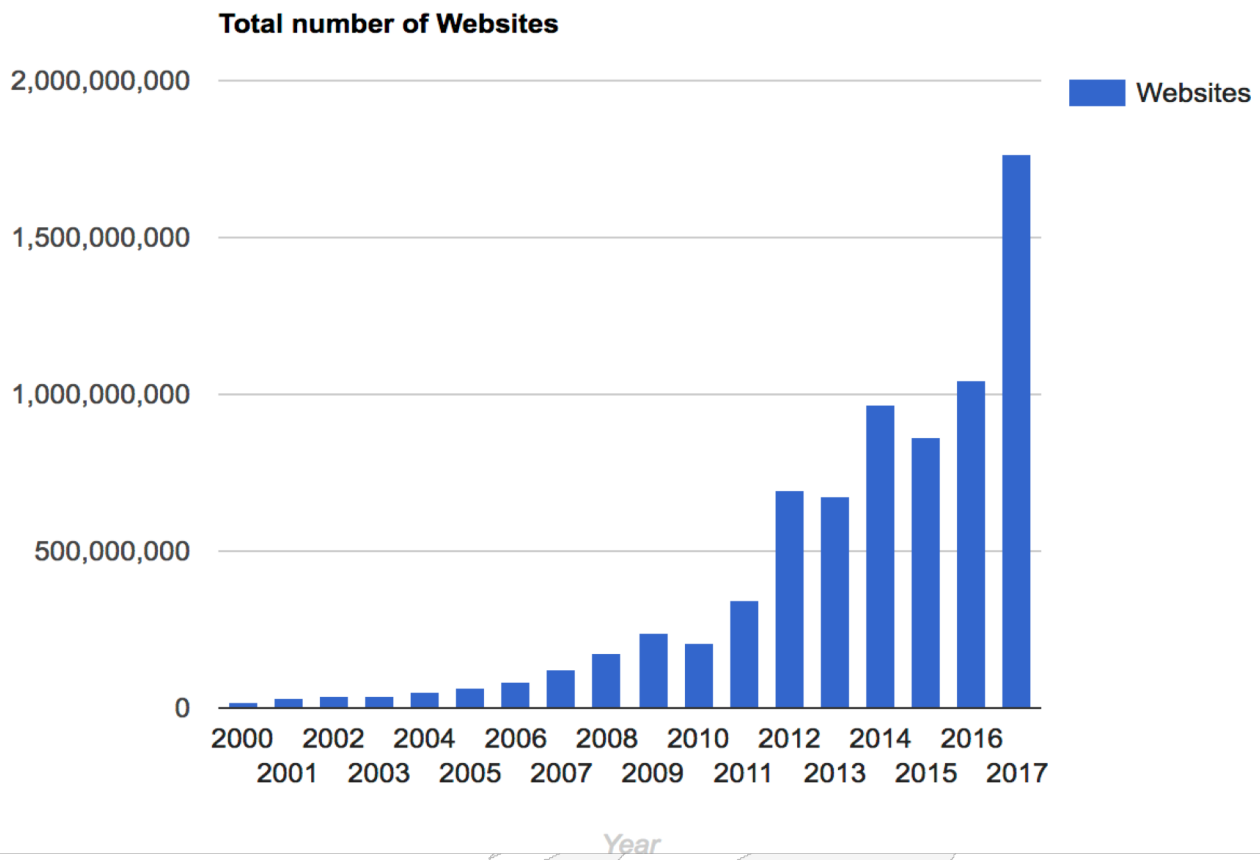
Aug 2018:
1,943,986,530 sites

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/

Copyright © Ellis Horowitz 1999-2017

Web Page Content Distribution

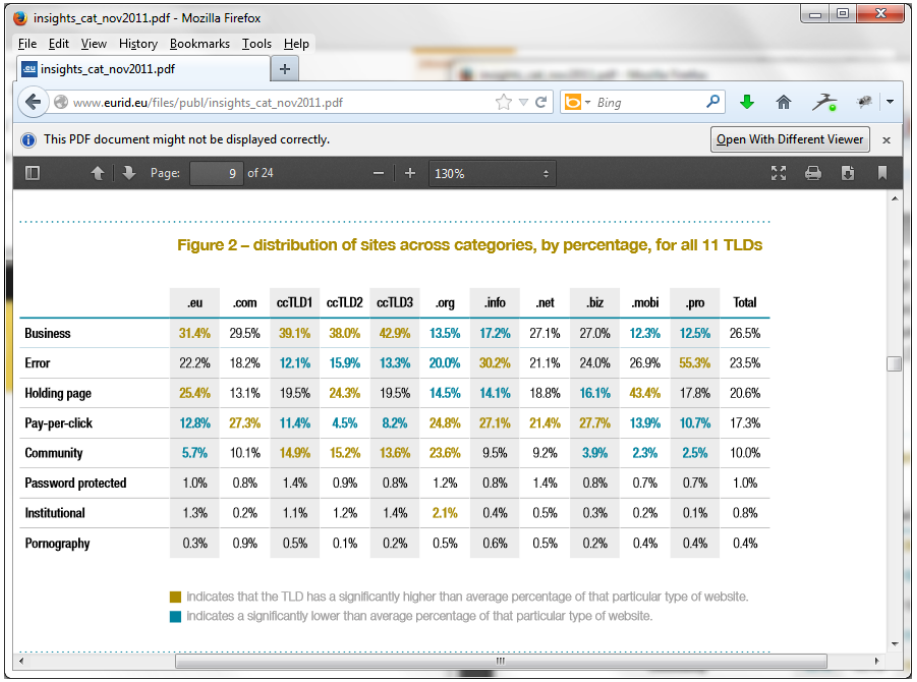
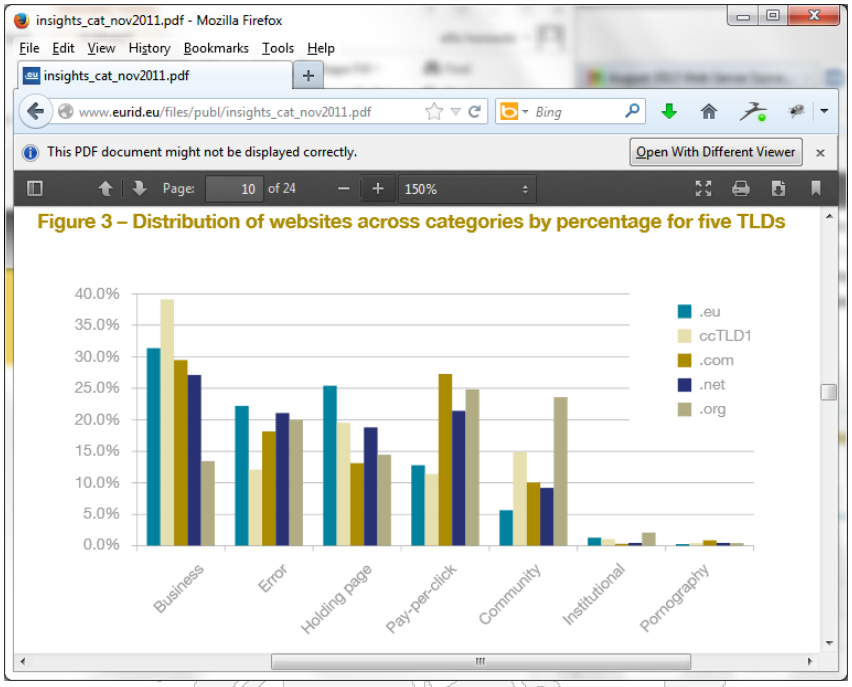


Figure 2 – distribution of sites across categories, by percentage, for all 11 TLDs

	.eu	.com	ccTLD1	ccTLD2	ccTLD3	.org	.info	.net	.biz	.mobi	.pro	Total
Business	31.4%	29.5%	39.1%	38.0%	42.9%	13.5%	17.2%	27.1%	27.0%	12.3%	12.5%	26.5%
Error	22.2%	18.2%	12.1%	15.9%	13.3%	20.0%	30.2%	21.1%	24.0%	26.9%	55.3%	23.5%
Holding page	25.4%	13.1%	19.5%	24.3%	19.5%	14.5%	14.1%	18.8%	16.1%	43.4%	17.8%	20.6%
Pay-per-click	12.8%	27.3%	11.4%	4.5%	8.2%	24.8%	27.1%	21.4%	27.7%	13.9%	10.7%	17.3%
Community	5.7%	10.1%	14.9%	15.2%	13.6%	23.6%	9.5%	9.2%	3.9%	2.3%	2.5%	10.0%
Password protected	1.0%	0.8%	1.4%	0.9%	0.8%	1.2%	0.8%	1.4%	0.8%	0.7%	0.7%	1.0%
Institutional	1.3%	0.2%	1.1%	1.2%	1.4%	2.1%	0.4%	0.5%	0.3%	0.2%	0.1%	0.8%
Pornography	0.3%	0.9%	0.5%	0.1%	0.2%	0.5%	0.6%	0.5%	0.2%	0.4%	0.4%	0.4%

indicates that the TLD has a significantly higher than average percentage of that particular type of website.
indicates a significantly lower than average percentage of that particular type of website.



Business is largest category in 6 out of 11 TLDs;
Least popular with Business category are .org, .mobi, .pro

.org maintains a community profile;
.eu has higher than average holding pages
as do ccTLDs

A study done in 2011 using 8 Categories of content include: 1. Business website; 2. Community or Personal website; 3. Pay-per-Click website (mainly advertising); 4. Holding page (corporate, under construction, held by ISP); 5. Government website; 6. Password protected website; 7. Error (not properly configured); 8. Pornography;

See http://www.eurid.eu/files/publ/insights_cat_nov2011.pdf

Copyright Ellis Horowitz, 2011-2019

Distribution Across TLDs

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

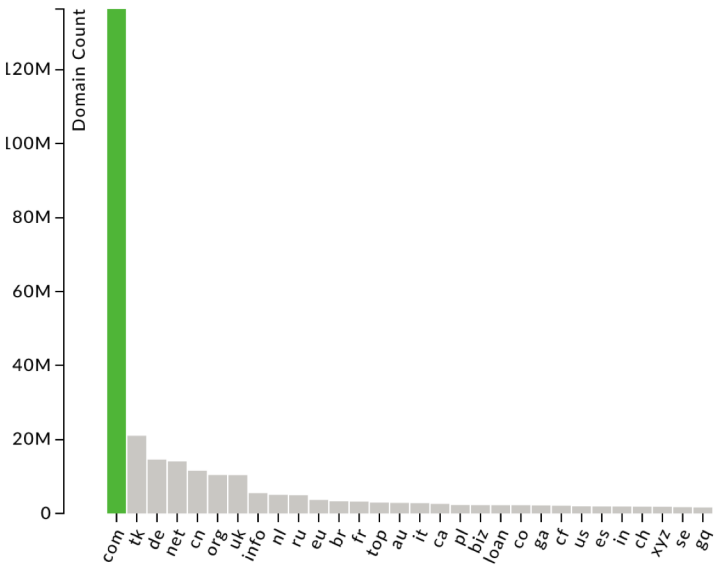
what is .tk and why is it so large? (Tokelau)

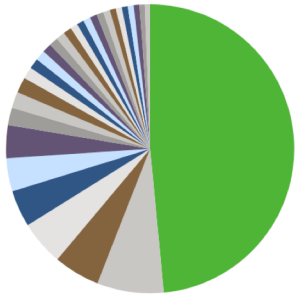
Domain Count Statistics for TLDs

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

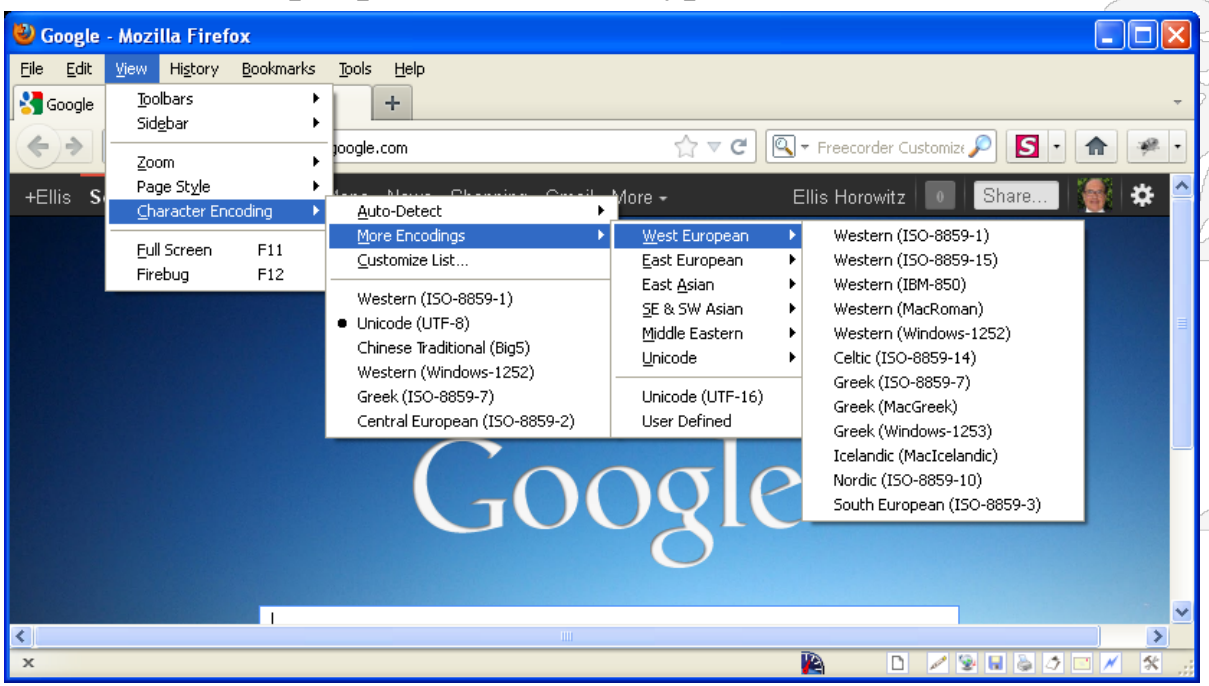
Display: Top25 - Default - None





Web Page Language Diversity

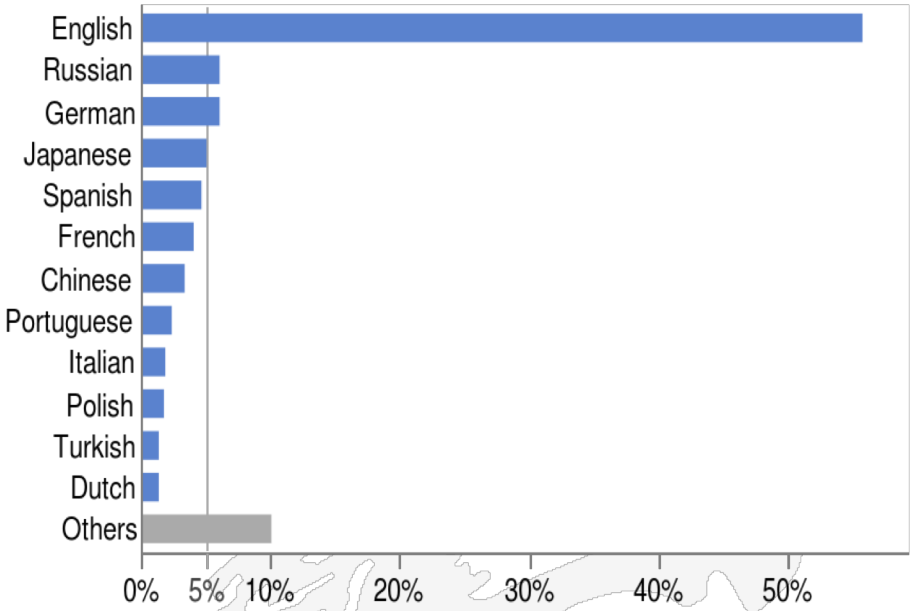
- **The Web contains pages in many different languages**
- Characters in a language are encoded such that each character is paired with a number
- Unicode and its parallel standard, the ISO/IEC 10646 Universal Character Set, together constitute a modern, unified character encoding.
- Most modern web 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%**

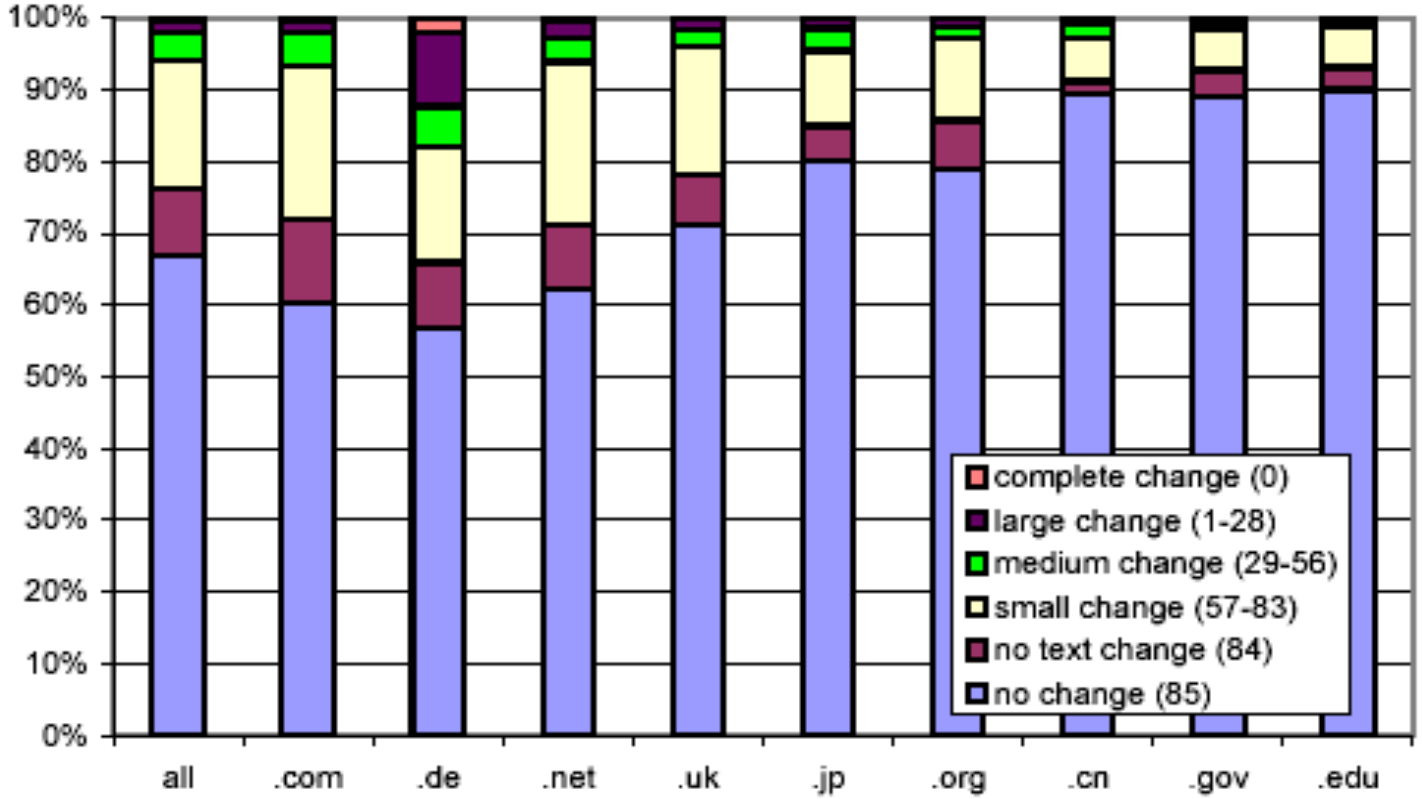


Content languages for websites as of 12 March 2014[2]

Copyright Ellis Horowitz, 2011-2019

Static Pages: Rate of Change

- Fetterly et al. study several views of data, 150 million pages over 11 weekly crawls
  - Divided into 6 groups by extent of change
- Changes in *.com are more frequent than .gov or .edu



Copyright Horowitz 1999-2017

Rate of Change

Cho and Garcia-Molina did a study of 720K pages from 270 popular sites sampled daily

Mathematically, what does this seem to be?

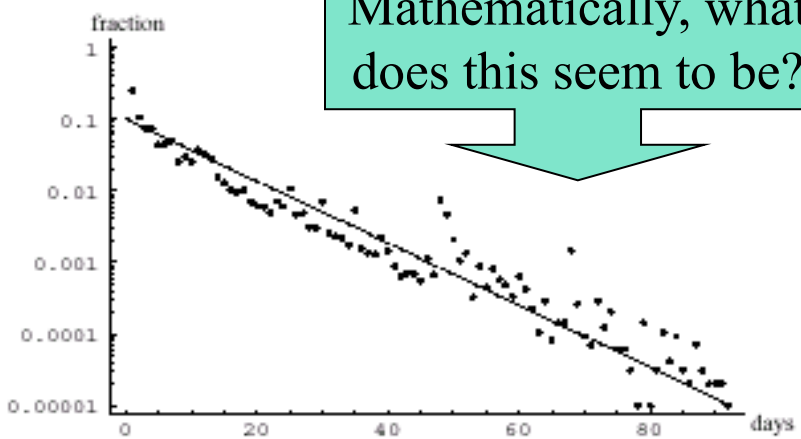


Figure 11: Change intervals for pages with the average change interval of 10 days

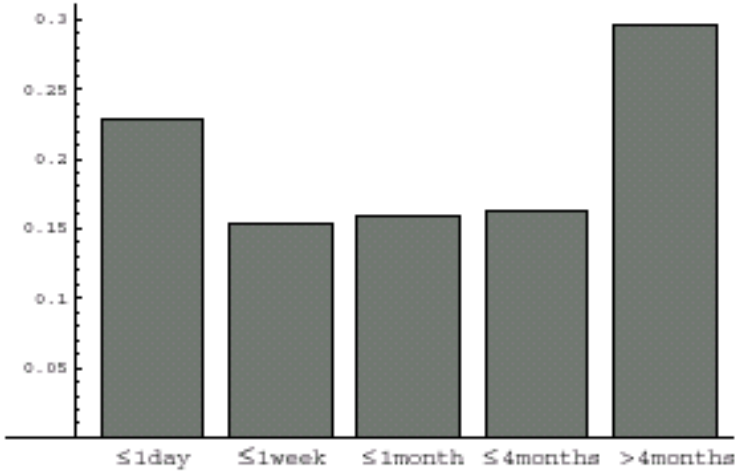


Figure 12: Percentage of pages with given average interval of change

Complexity of Data Types

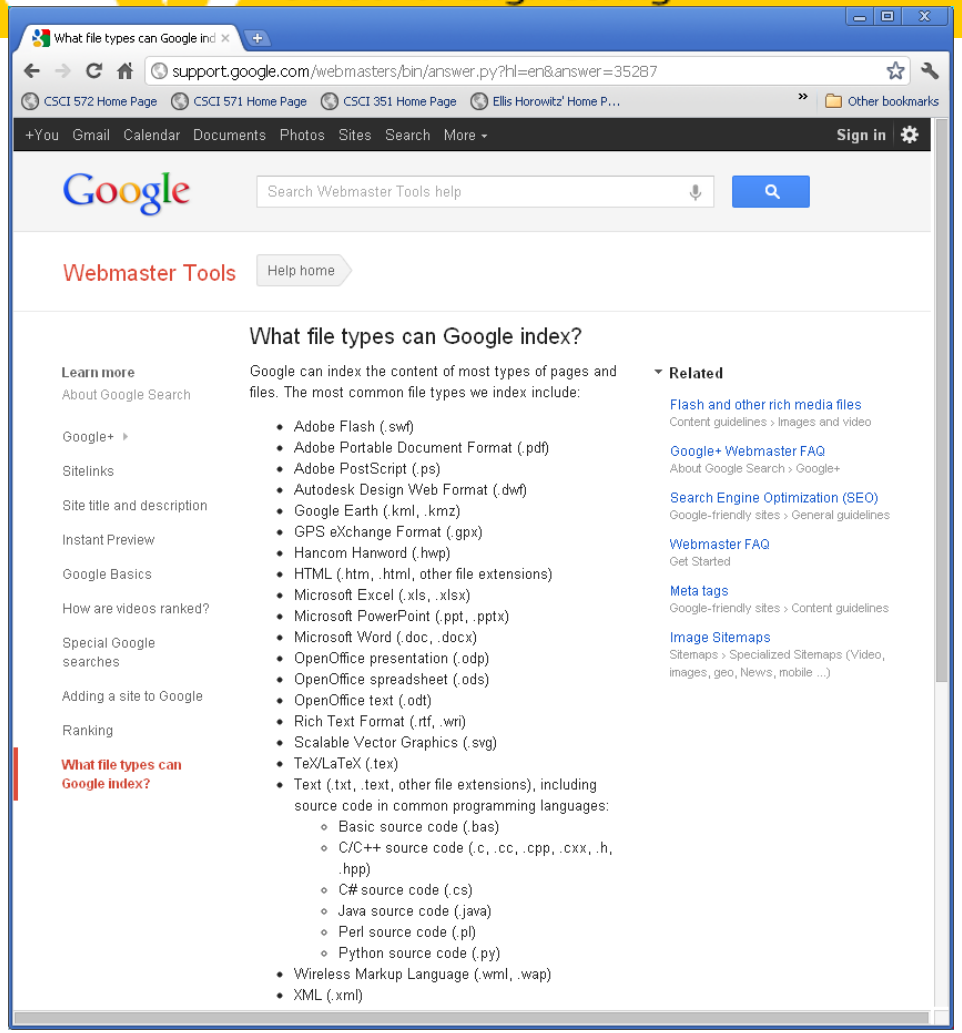
Proliferation of Content Types Available

- **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

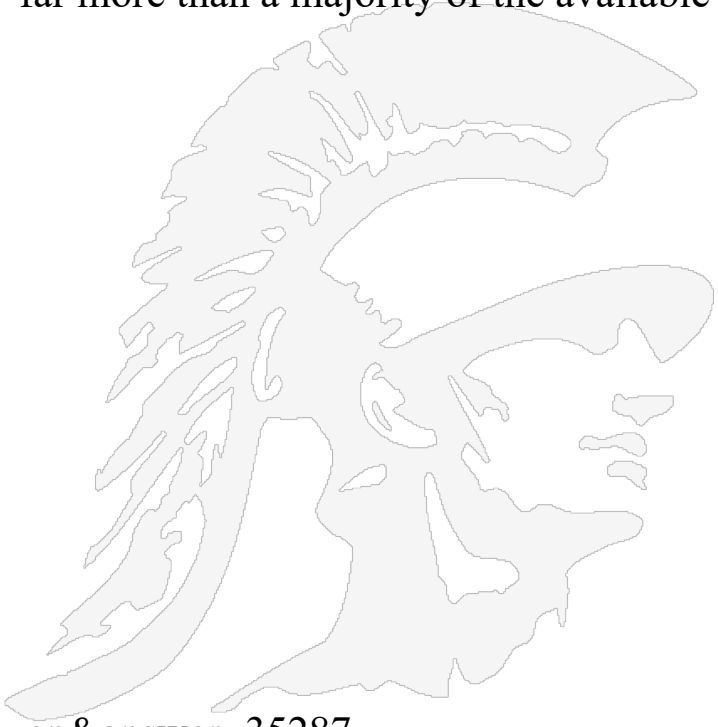
*http://filext.com/ (see if you can name the top 20 file extensions)

Copyright Ellis Horowitz, 2011-2019

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

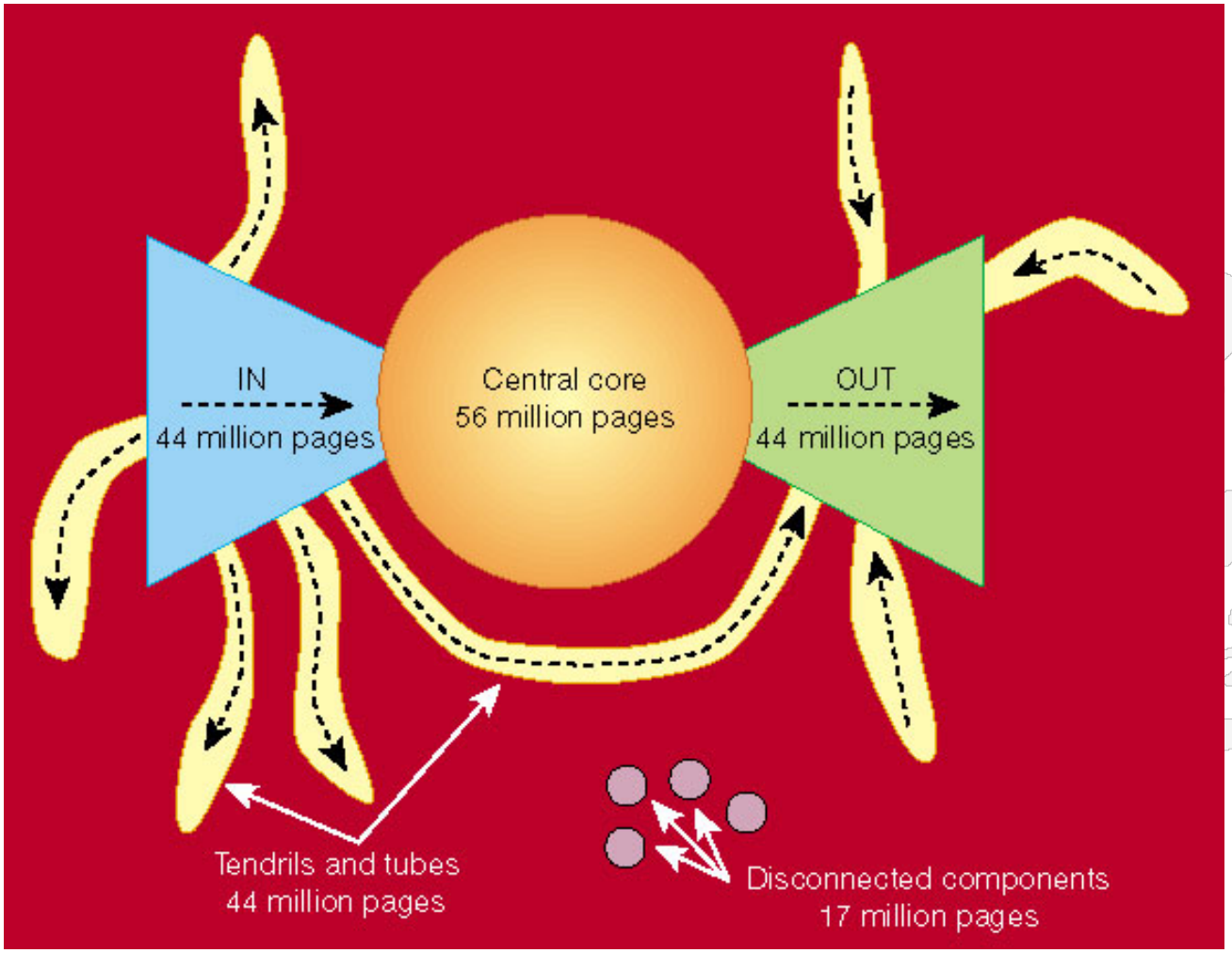
Copyright Ellis Horowitz, 2011-2019

Web Characteristics When Viewed as a Graph

- **Significant duplication**
  - Syntactic – 30%-40% (near) duplicates
    - [Brod97, Syntactic Clustering of the Web, Broder, Glassman, Manasse, Zweig, http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-TN-1997-015.pdf],
    - [Shiv99b, Finding Replicated Web Collections, Cho, Shivakumar, Garcia-Molina]
- **High linkage**
  - More than 8 links/page in the average
- **Complex graph topology**
  - Not a small world; bow-tie structure Graph Structure in the Web [Brod00]
- **Spam**
  - 100s of millions of pages

Copyright Ellis Horowitz, 2011-2019

Graph Structure in the Web (an old study)



Copyright © Ellis Horowitz
1999-2018

http://www9.org/w9cdrom/160/160.html

Distance Measurement Results

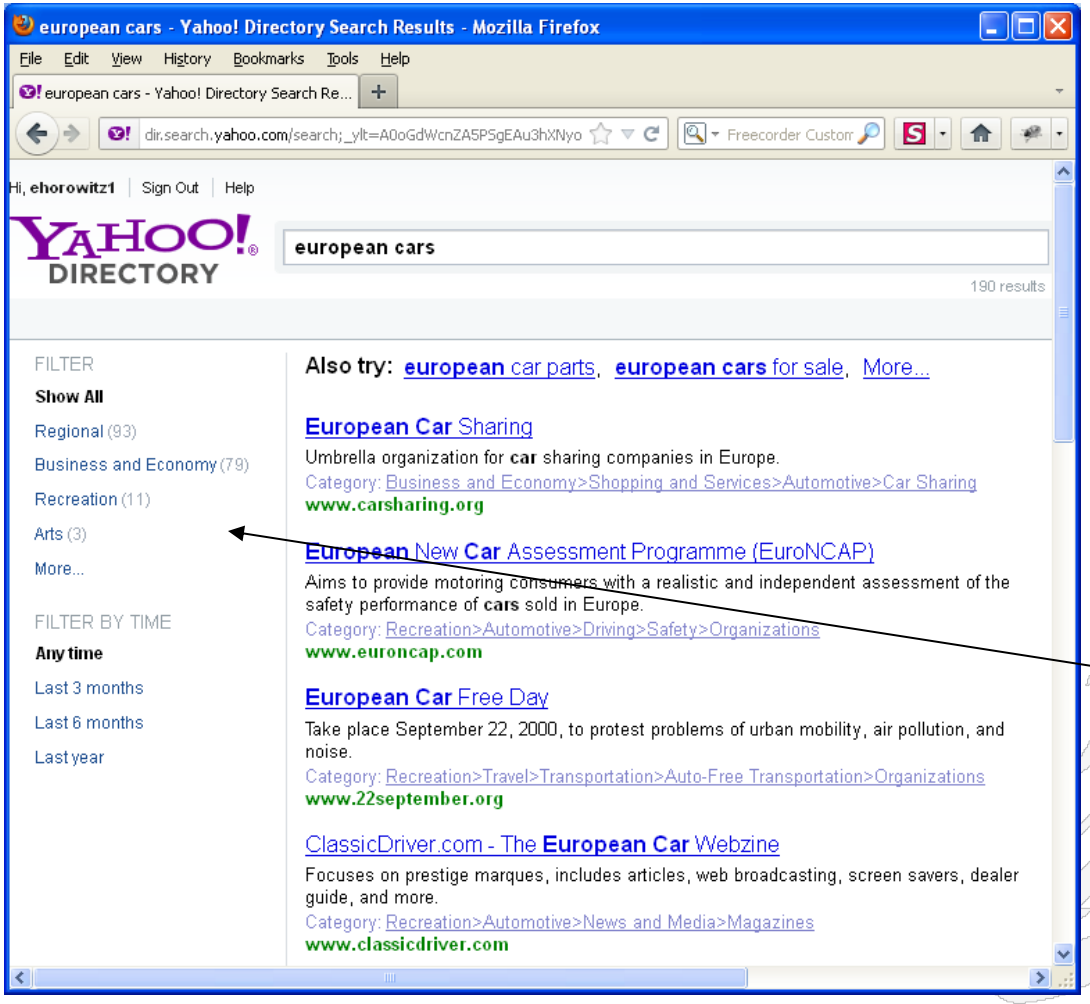
- **For random pages *p1*,*p2*:**

  Pr[*p1* reachable from *p2*] ~ 1/4
- **Maximum directed distance between 2 SCC nodes: >28**
- **Maximum directed distance between 2 nodes, given there is a path: > 900**
- **Average directed distance between 2 SCC nodes: ~16**
- **Average undirected distance: ~7**

- **These properties are likely no longer true, but no one has attempted a similar study**

Copyright Ellis Horowitz, 2011-2019

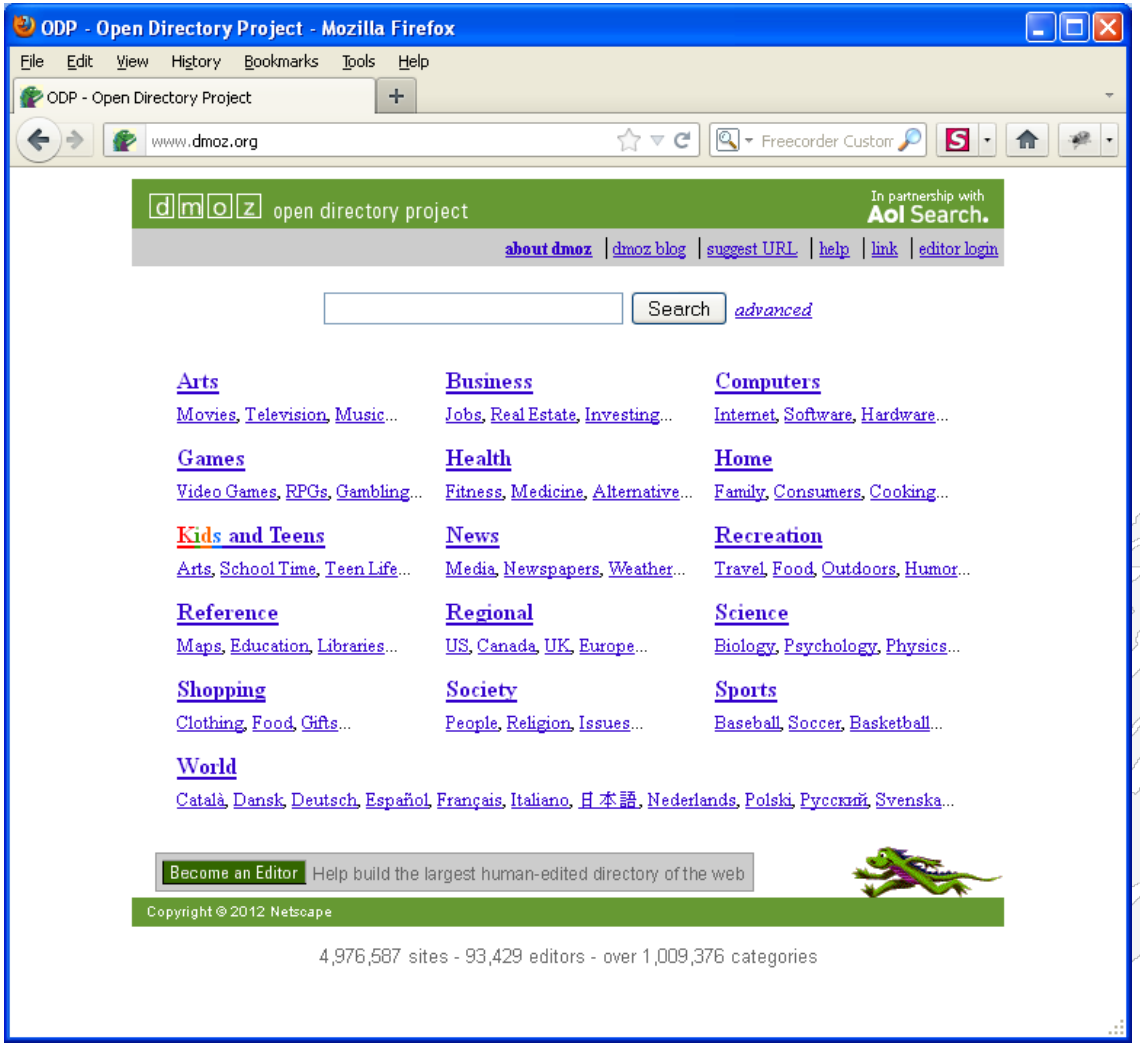
Manual Hierarchical Web Taxonomies



- **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**

Copyright Ellis Horowitz, 2011-2019

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

Copyright Ellis Horowitz, 2011-2019

Drilling Down By Category

Open Directory - Science - Mozilla Firefox

www.dmoz.org/Science/

dmoz open directory project — In partnership with AOL Search.

about dmoz | dmoz blog | report abuse/spam | help

Search — the entire directory

Top: **Science** *(104,480)* — Description

[A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z]

- Agriculture *(3,670)*
- Anomalies and Alternative Science *(480)*
- Astronomy *(3,595)*
- Biology *(31,012)*
- Chemistry *(4,171)*
- Computer Science@ *(1,971)*
- Earth Sciences *(6,098)*
- Environment *(6,572)*
- Math *(9,541)*
- Physics *(4,281)*
- Science in Society *(680)*
- Social Sciences *(19,489)*
- Technology *(10,560)*
- Women@ *(153)*

- Academic Departments *(9)*
- By Region *(0)*
- Chats and Forums *(16)*
- Directories *(27)*
- Educational Resources *(352)*
- Employment *(69)*
- Events *(54)*
- History@ *(346)*
- Instruments and Supplies *(2,370)*
- Libraries@ *(88)*
- Methods and Techniques *(190)*
- Museums@ *(479)*
- News and Media *(234)*
- Organizations *(131)*
- People *(0)*
- Publications *(241)*
- Reference *(389)*
- Research Groups and Centers *(57)*
- Search Engines *(9)*
- Software *(183)*
- Weblogs@ *(117)*

Selecting Category "Science"

Open Directory - Computers: Computer Science - Mozilla Firefox

www.dmoz.org/Computers/Computer_Science/

dmoz open directory project — In partnership with AOL Search.

about dmoz | dmoz blog | report abuse/spam | help

Search — the entire directory

Top: **Computers**: **Computer Science** *(1,971)* — Description

- Academic Departments *(553)*
- Conferences *(203)*
- Directories *(8)*
- Organizations *(71)*
- People *(271)*
- Publications *(80)*
- Reference *(4)*
- Research Institutes *(74)*

- Artificial Intelligence@ *(1,294)*
- Artificial Life@ *(230)*
- Computational Geometry@ *(60)*
- Computer Graphics *(39)*
- Database Theory *(82)*
- Distributed Computing *(225)*
- Parallel Computing@ *(367)*
- Software Engineering@ *(114)*
- Theoretical *(361)*

See also:

- Computers: Algorithms *(284)*
- Computers: Programming *(15,619)*
- Science: Math *(9,541)*
- Science: Technology: Electrical Engineering *(237)*

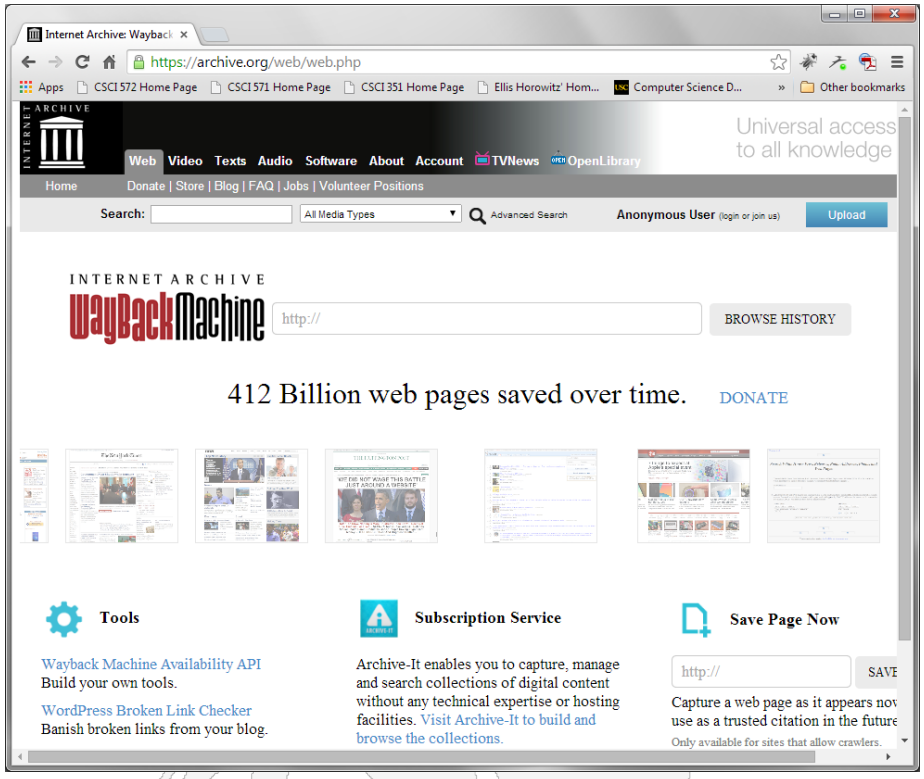
This category in other languages:

Selecting Category "Computer Science"

Copyright © Ellis Horowitz 1999-2018

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.



Copyright Ellis Horowitz, 2011-2019

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
Drug Trafficking sites
TOR-Encrypted sites
Political Protests
Private Communications