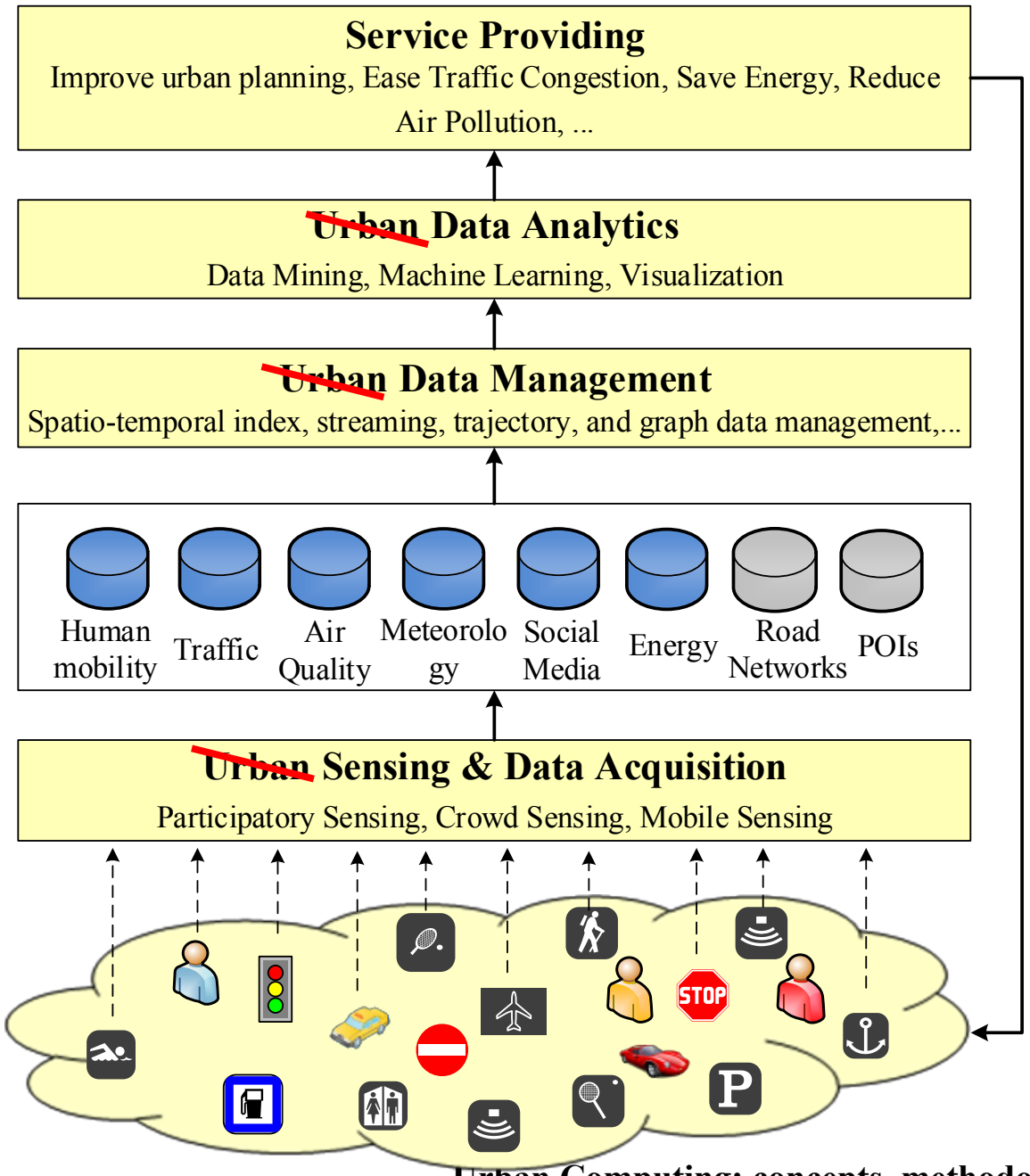
Welcome to

DS3010:
DS III: Computational Data Intelligence
Data acquisition and measurement

Prof. Yanhua Li

Time: 11:00am – 12:50pm Mon & Thur
Location: HL 114
D-term 2022

Data Pipeline



Urban Computing: concepts, methodologies, and applications.
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.

Data acquisition and measurement

via Sampling and Estimation

measurement distortions

"World Map" in 1459

- proved incomplete (Columbus et al. 1492)
- wrong proportions (Africa & Asia)

The Fra Mauro world map (1459)

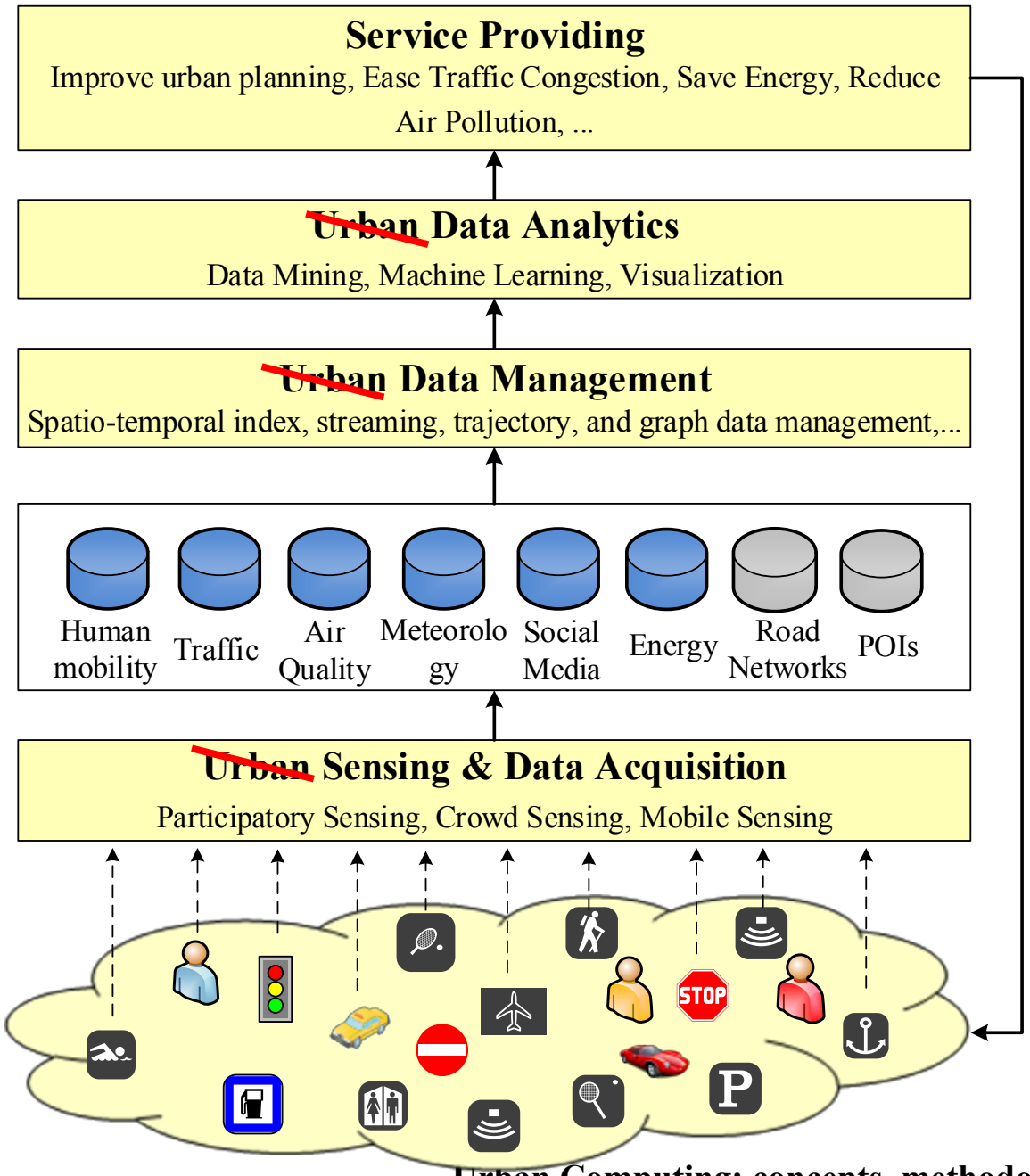
source: Wikipedia

outline

- Why sampling?

- Sampling methods

Data Pipeline



Urban Computing: concepts, methodologies, and applications.
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.

Motivation

- Measurement studies aid understanding existing systems and user behaviors.

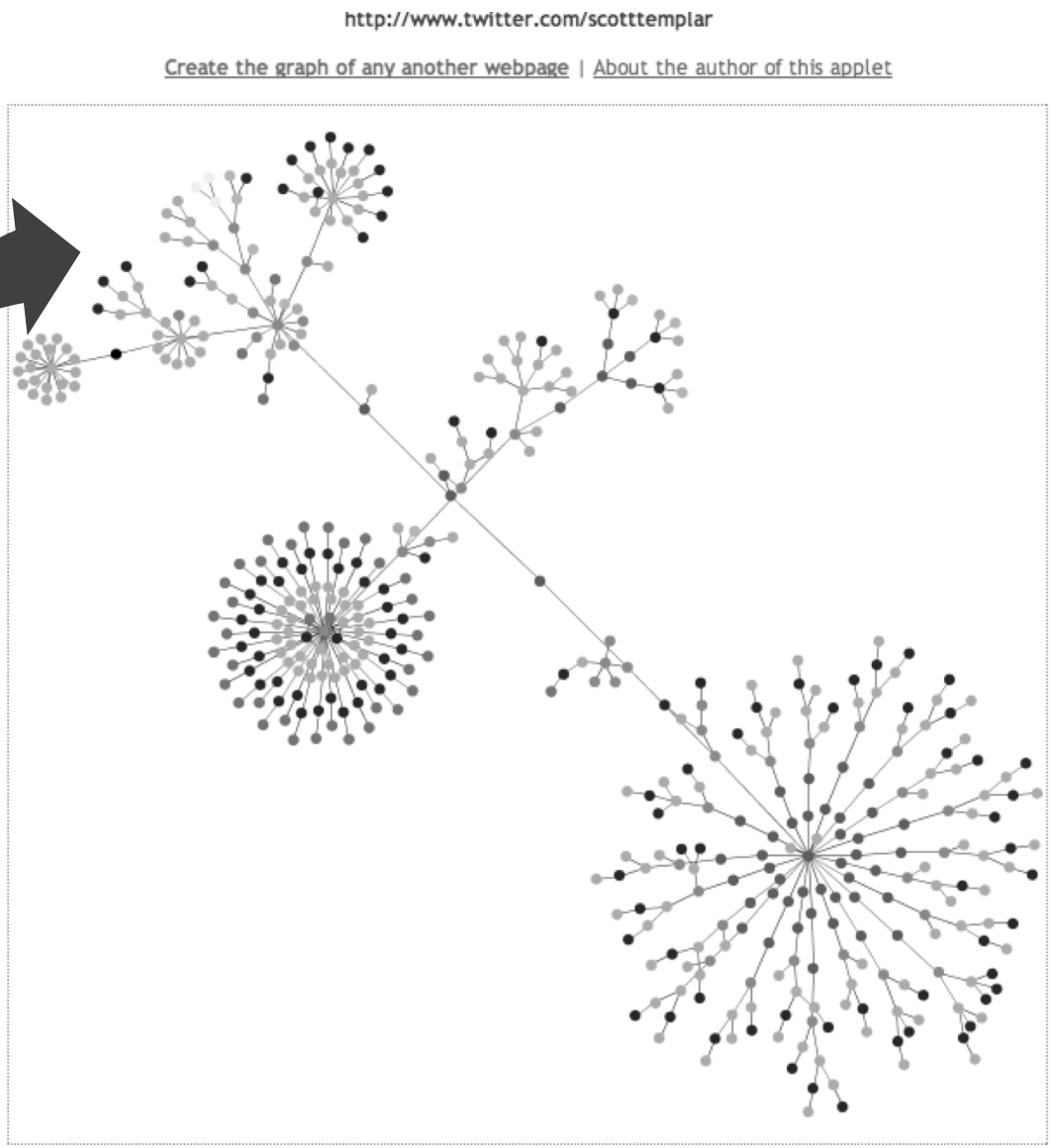
- Capturing an accurate global "snapshot" is often infeasible.

- *How can we collect representative samples?*

Motivation

Sample data
to estimate the
statistics, i.e., size,
degree distribution,
etc.

sample of social networks

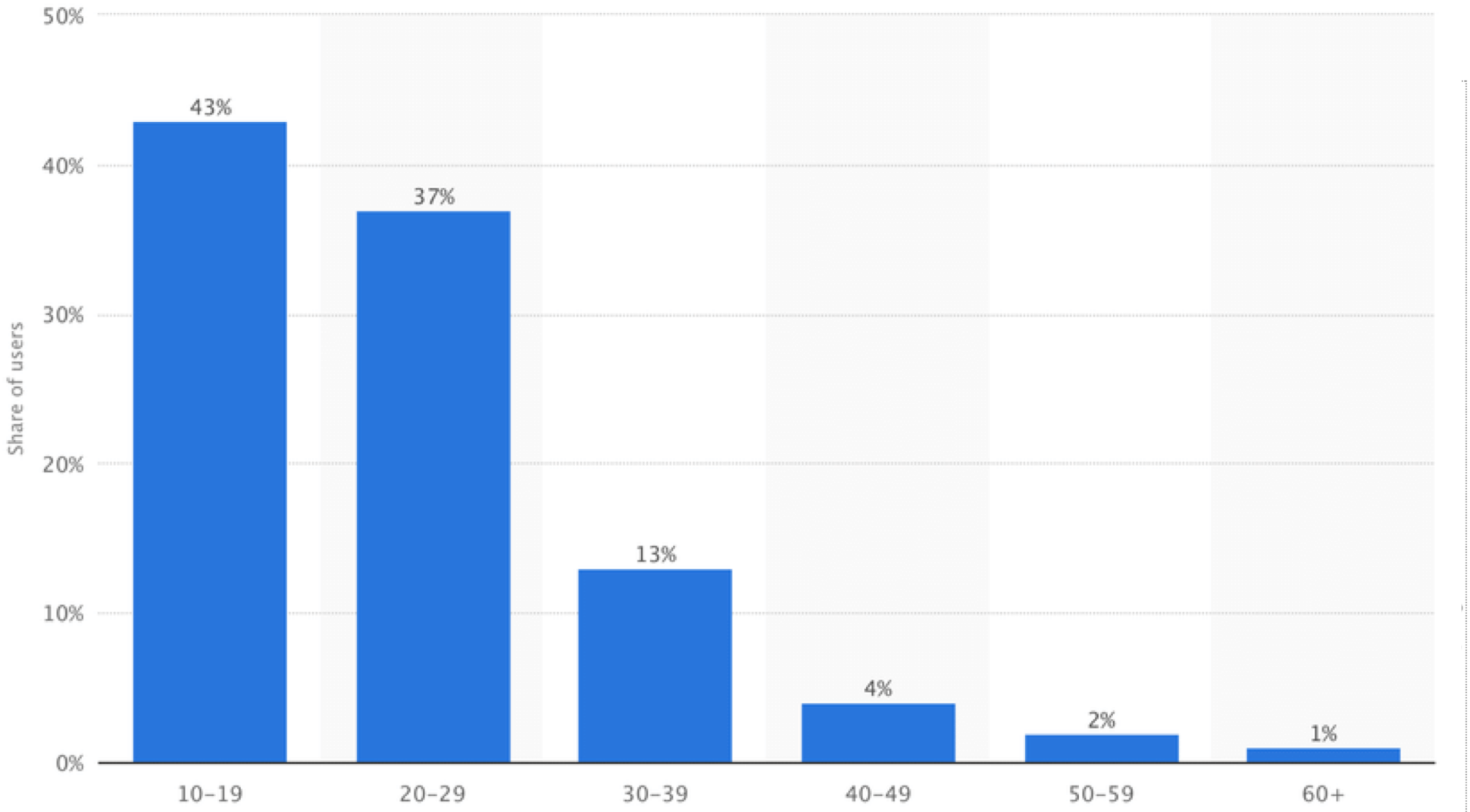


- Capturing an accurate global "snapshot" is often infeasible.

- *How can we collect representative samples?*

Age distribution of Twitter users worldwide as of October 2013

This statistic gives information on the age distribution of Twitter users worldwide. As of October 2013, 43 percent of global active Twitter users were between 10 and 19 years old.

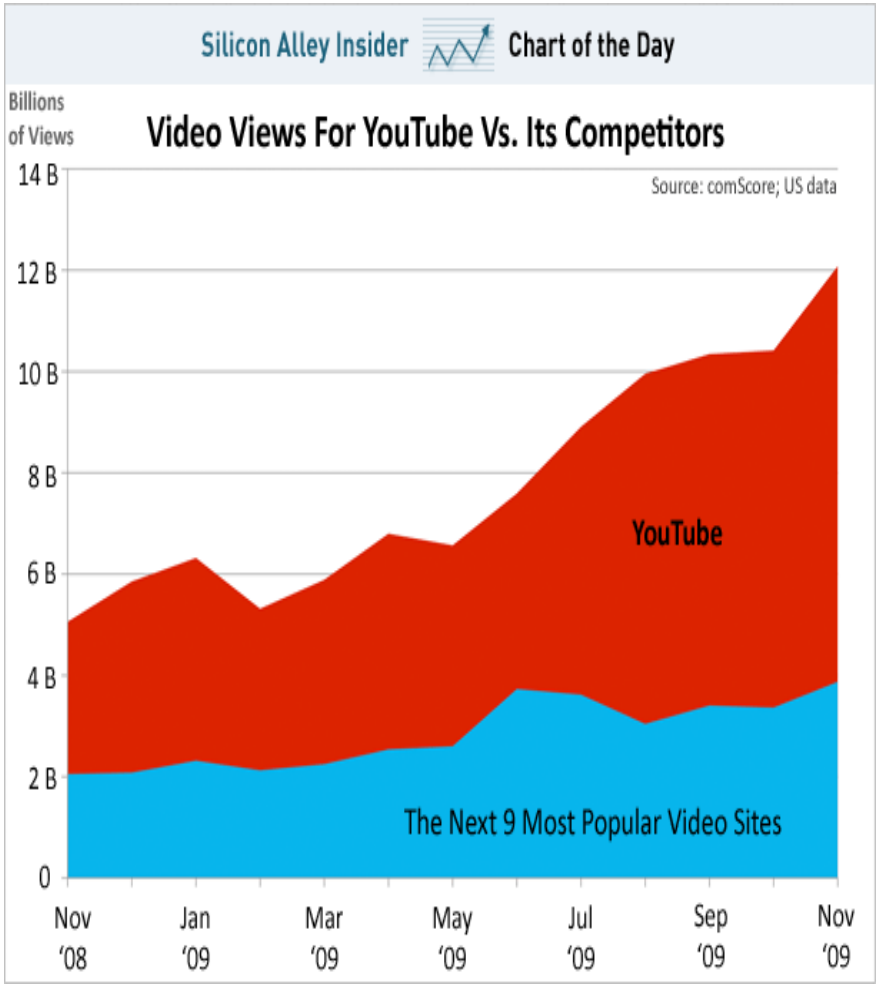


Counting YouTube Videos via Sampling

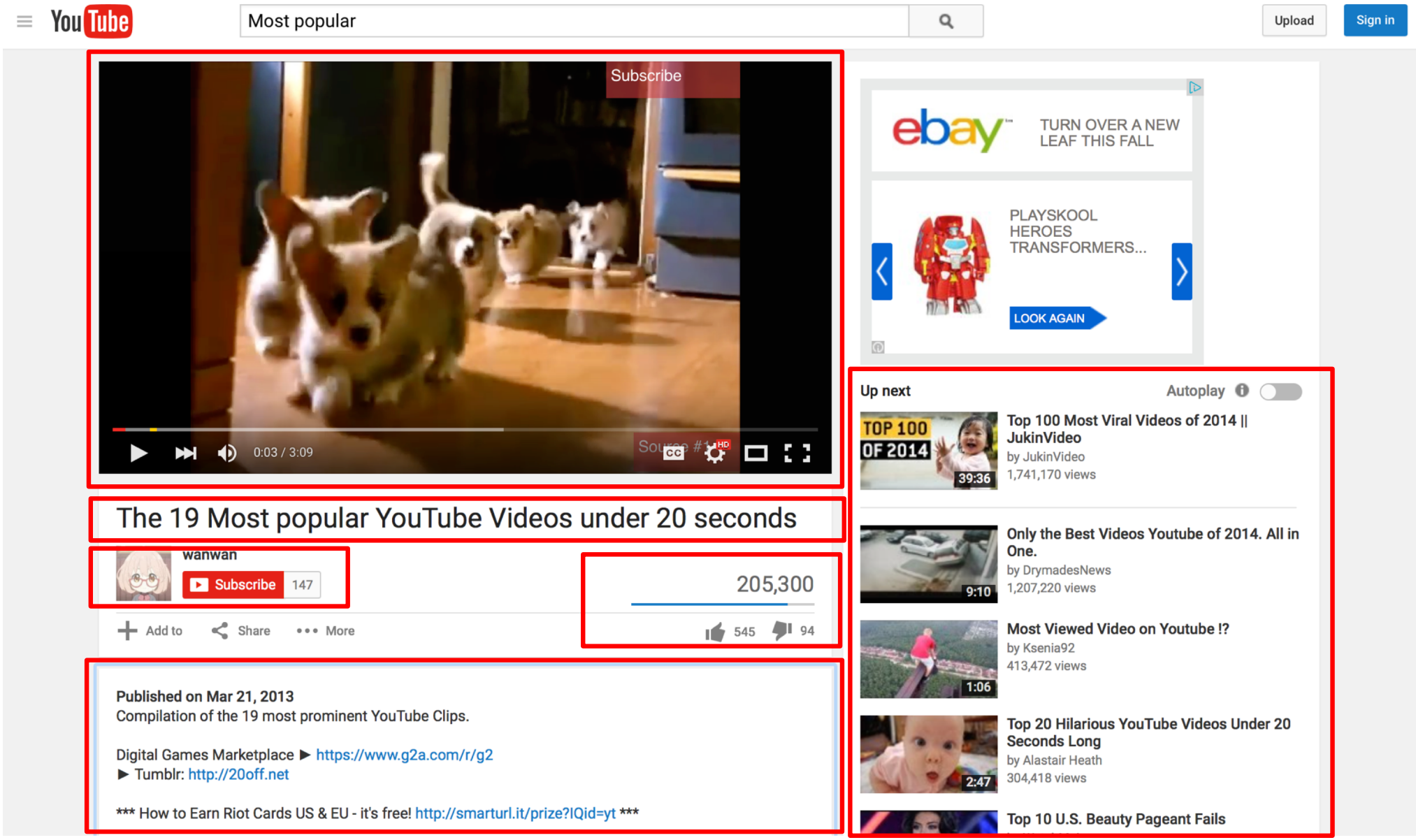
Why YouTube?

World's largest (mostly user-generated) global video delivery service

- More than 13 million hours of video were uploaded during 2010 and 35 hours of video are uploaded every minute.
- More videos are uploaded to YouTube in 60 days than the 3 major US networks created in 60 years
- 70% of YouTube traffic comes from outside the US
- YouTube reached over 700 billion playbacks in 2010
- YouTube mobile gets over 100 million views a day



YouTube Video



Comments from other YouTube users

Socio-technical Aspects of YouTube: Counting Videos & Views

Why Counting YouTube Videos and Views::

- YouTube traffic contributes to a significant portion of inter-domain network traffic
- Knowing the total number of videos and view counts per day can shed light on
  - the total amount of storage
  - as well as the system capacity needed to store and deliver YouTube videos

Challenges:

- These statistics are *not* made available publicly by YouTube
- Even for YouTube, it is costly to get an exact answer.

Challenges for Counting Videos & Views

- Video id space is extremely large, of the order $O(64^{11})$
  - brute-force survey of the entire YouTube video population will be too costly
  - direct application of (uniform) random sampling to the video id space will be ineffective
- Existing methods for *collecting* YouTube videos following the “related videos” links produce a biased sample

Goal

- Develop a sampling model of an *unbiased* estimator for the total number of YouTube videos

- Apply the sampling method to
  - Estimate the total number of videos and analyze its dynamics
  - Estimate the views counts and study its properties

Sampling Techniques to Count Population

- **German Tank Problem**
- Panther tanks, 1943.
- World War II
- Estimate # German Tanks (*N*)
- the problem of estimating

the maximum of a discrete uniform distribution from Sampling without replacement

- *m* : the max series number
- *k* : total number of tanks observed
- Estimator: ???

Sampling Techniques to Count Population

- **German Tank Problem**
- Panther tanks, 1943.
- World War II
- Estimate # German Tanks (*N*)
- the problem of estimating

the maximum of a discrete uniform distribution from Sampling without replacement

- *m* : the max series number
- *k* : total number of tanks observed
- **Estimator:** $\hat{N} = m(1 + k^{-1}) - 1$
- the sample maximum plus the average gap between observations in the sample.
- **Minimum-variance unbiased estimator**
- https://en.wikipedia.org/wiki/German_tank_problem#Frequentist_analysis

Sampling Techniques to Count Population

- **Mark and recapture**
- a method commonly used in ecology to estimate an animal population's size *N*.

- Step 1: A portion of the population *K* is captured, marked, and released.

- Step 2: Later, another portion *n* is captured and the number of marked individuals within the sample is counted *k*.

- Estimation: ???

Sampling Techniques to Count Population

- **Mark and recapture**

- *N* = Number of animals in the population
- *K* = Number of animals marked on the first visit
- *n* = Number of animals captured on the second visit
- *k* = Number of recaptured animals that were marked

- Estimator: ???

Sampling Techniques to Count Population

- **Mark and recapture**

- *N* = Number of animals in the population
- *K* = Number of animals marked on the first visit
- *n* = Number of animals captured on the second visit
- *k* = Number of recaptured animals that were marked

- Assumption: Each animal has an equal probability *p* being captured
- Thus, $p = \frac{k}{K} = \frac{n}{N}$

- The estimator is obtained, as $\hat{N} = \frac{Kn}{k}$.

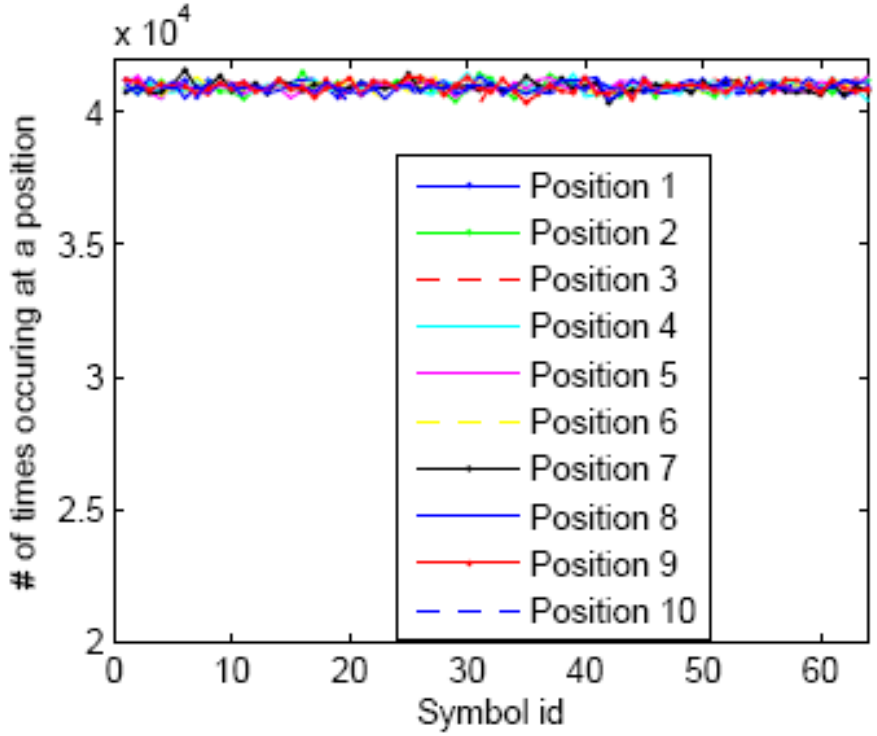
YouTube Video ID Space

Each YouTube id consists of 11 characters

The first 10 characters of a valid id contain any of the characters in S = {0–9,_,–,A–Z, a–z}

The last (11-th) character only comes from T = {0, 4, 8, A,E, I,M, Q,U, Y, c, g, k, o, s,w}

A YouTube video id is randomly generated from the id space $\mathcal{S}$



Prefix Search in YouTube

https://www.youtube.com/watch?v=kSbJAg5nPWk

https://www.youtube.com/watch?v=FAgabbBVL6Y

- Key unique property of YouTube search API we accidentally stumble on
  - When searching using a keyword string of the format ”watch?v=xy-...z”
  - YouTube returns a list of videos whose id’s begin with “xy-”, if they exist.
- The above property is well validated by three real datasets

Certain return limits apply, e.g., maximum # of videos returned is 200.

Prefix Sampling

- Total number of YouTube video IDs: 16×64^{10}
- Total number of prefixes if prefix length is 4: 64^4=16 million
- Each prefix has the same number of video IDs, e.g., for prefix length of 4,

  there are $16 \times 64^{(10-4)}$=1 trillion video ids.

can we use German Tank and Mark-recapture method to estimate the YouTube video population size, and why?

Prefix Sampling

- Total number of YouTube video IDs: 16×64^{10}
- Total number of prefixes if prefix length is 4: 64^4=16 million
- Each prefix has the same number of video IDs, e.g., for prefix length of 4,

  there are $16 \times 64^{(10-4)}$=1 trillion video ids.

- X_i is the number of valid videos in the i-th sampled prefix
- Total valid YouTube video estimated using X_i is

Prefix Sampling

- Total number of YouTube video IDs: 16×64^{10}
- Total number of prefixes if prefix length is 4: 64^4=16 million
- Each prefix has the same number of video IDs, e.g., for prefix length of 4,

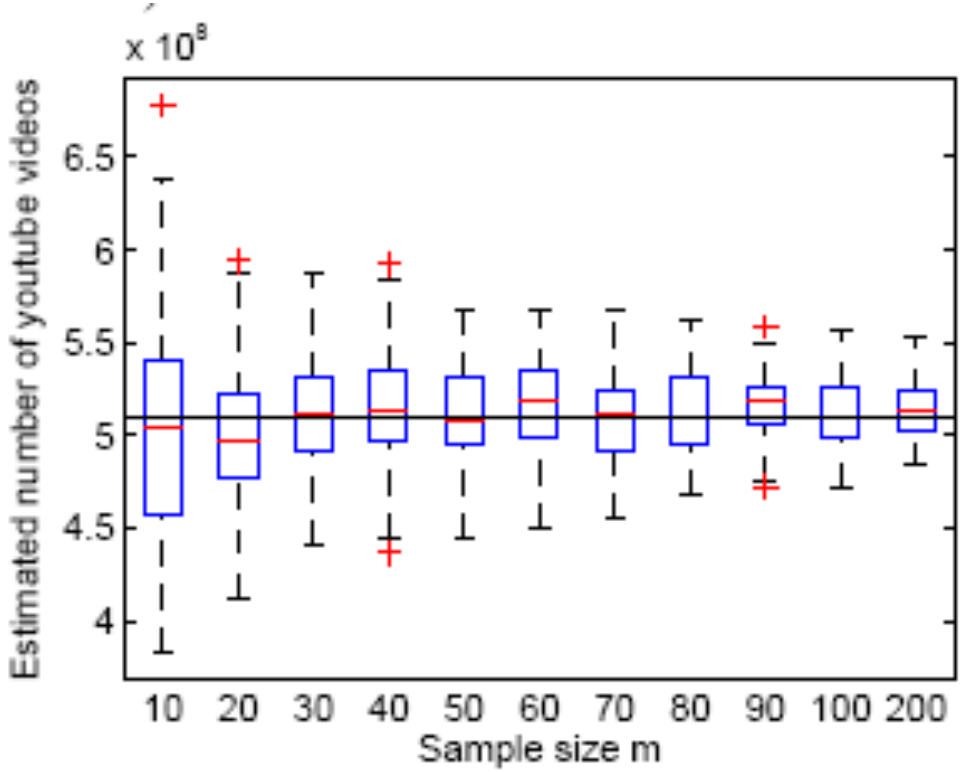
  there are $16 \times 64^{(10-4)}$=1 trillion video ids.

- X_i is the number of valid videos in the i-th sampled prefix
- Total valid YouTube video estimated using X_i is $X_i * 64^4$

Unbiased Estimator for the Total Number of Videos

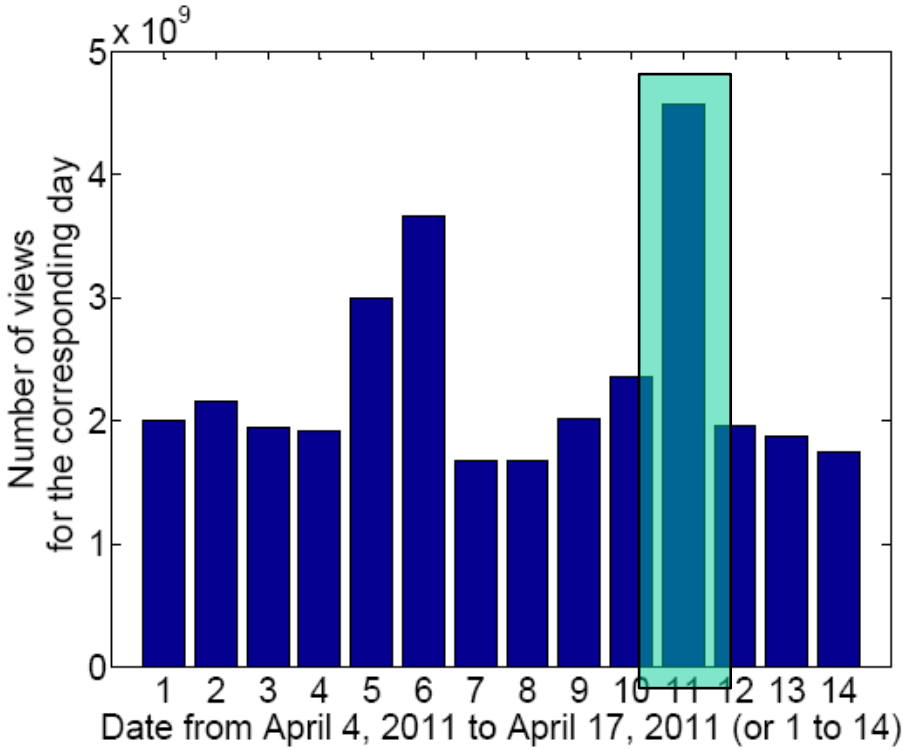
- Given m samples X_i by querying randomly generated prefixes of the same length e.g., 4,
- we have the unbiased estimator of total number of videos $\widehat{N} = \frac{1}{m}\sum_{i=1}^{m}(64^4\ X_i)$

Estimated number of YouTube videos by 05/12/2011



- The estimated result becomes more stable with more samples
- Around half a billion videos by May 2011

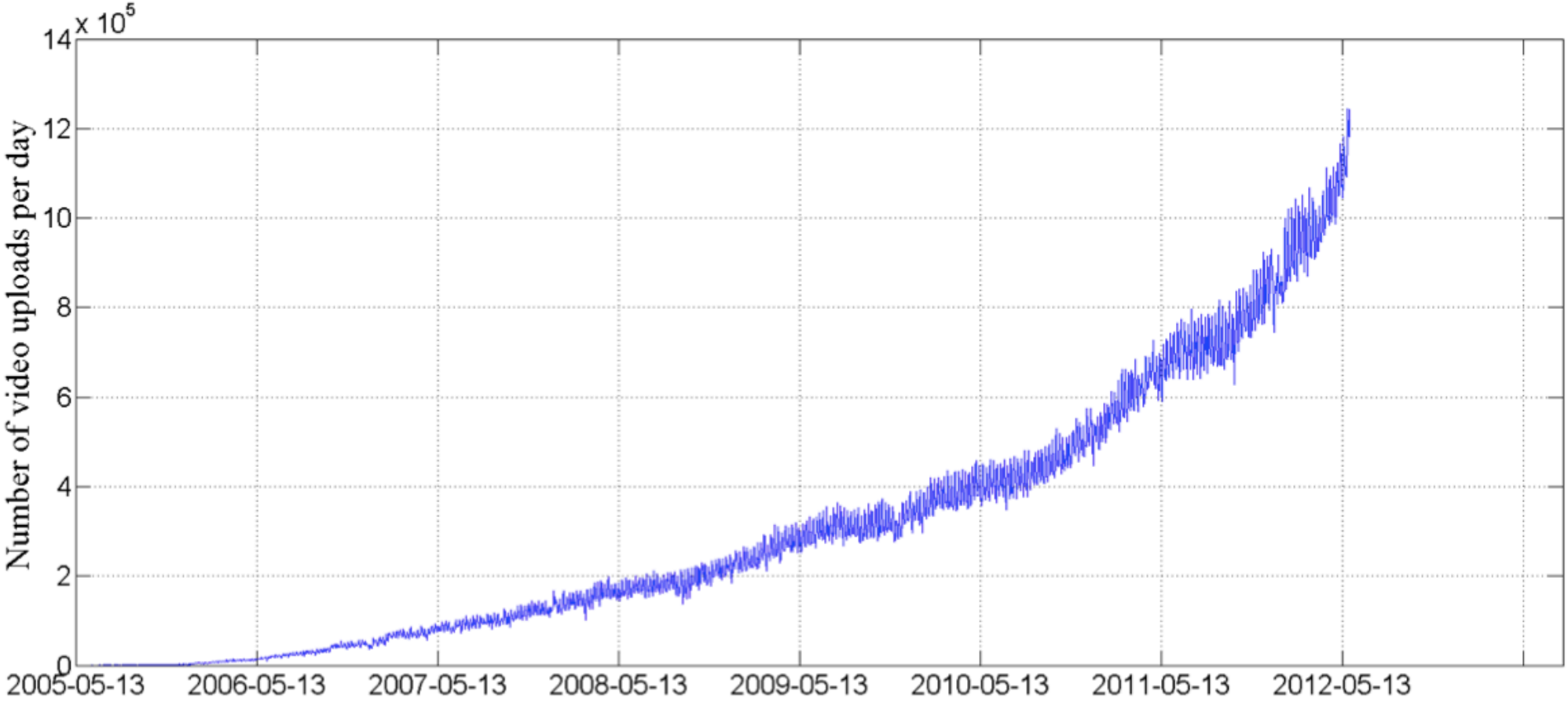
Number of Views for a two week period



On average it is 2.3 billion per day

For some day it can be as large as over 4.6 billions or over twice of the average, e.g., April 11, 2011

Daily YouTube video uploads



Slow in the first two years but increase more and more quickly in the following years;

Sampled Data

- Q00I-y9iePw|Tech|2008-08-19T02:52:52.000Z|23|blessingsolarenergy
- q00i--f2s4s|Entertainment|2008-10-12T18:29:22.000Z|602|corester69
- q00j-Zrs730|Music|2009-08-04T08:27:38.000Z|323|jeppeli123
- q00j-9vwAEA|Games|2009-08-15T19:36:50.000Z|64|GMLEGENDAZTEK
- Q00J-XhwEqA|People|2009-04-23T22:56:54.000Z|72|sjohnsgeo
- Q00j-9h8g0k|Games|2010-10-14T11:44:13.000Z|29|bebelulu91
- q00k-mgp9ak|Music|2008-02-12T16:51:02.000Z|169|grizzly9587
- Q00K-TZ53IY|People|2009-02-17T23:58:46.000Z|535|83diogosampaio
- q00K-VR6xT0|Comedy|2011-02-13T18:04:26.000Z|71|WhatsUpTay
- Q00L-OsxpfM|Comedy|2008-04-11T00:46:39.000Z|94|feergi
- Q00m-hFq_0Y|Music|2010-01-02T02:15:10.000Z|212|BakhtiyarHajiyev
- q00m-44nU7o|Sports|2007-07-23T21:17:16.000Z|27|smashingSurfer
- Q00m-Qha_nE|People|2009-11-29T03:54:40.000Z|29|swaggaqueens
- Q00N-LAzRgI|Entertainment|2010-12-12T03:03:20.000Z|321|BNMASS

Example result in JSON from YouTube


```
{
    "kind": "youtube#searchListResponse",
    "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/PaiEDiVxOyCWelLPuuwa9LKz3Gk\"",
    "nextPageToken": "CAUQAA",
    "regionCode": "KE",
    "pageInfo": { "totalResults": 4249, "resultsPerPage": 3 },
    "items": [
            {"kind": "youtube#searchResult",
              "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/QpOIr3QKlV5EUlzfFcVvDiJT0hw\"",
              "id": {
                        "kind": "youtube#channel",
                        "channelId": "UCJowOS1R0FnhipXVqEnYU1A"
                    }
            },
            { "kind": "youtube#searchResult",
              "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/AWutzVOt_5p1iLVifyBdfoSTf9E\"",
              "id":
                        { "kind": "youtube#video",
                          "videoId": "Eqa2nAAhHN0" } },
           { "kind": "youtube#searchResult",
             "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/2dIR9BTfr7QphpBuY3hPU-h5u-4\"",
             "id": { "kind": "youtube#video",
                       "videoId": "IirngItQuVs" } }
              ]
}
```