

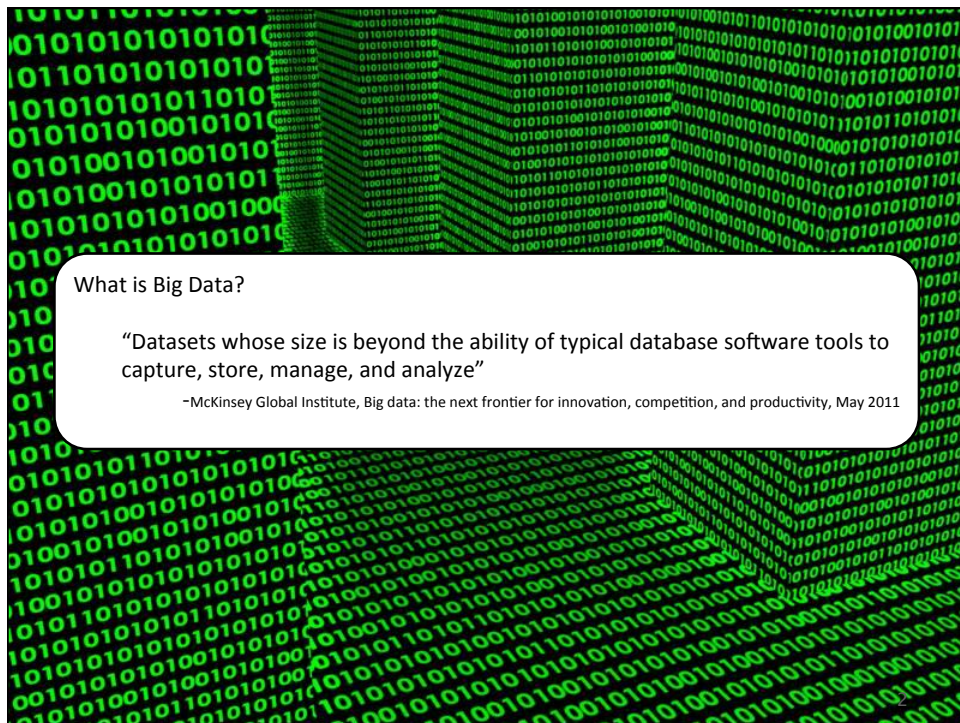
School of Communication
and Information

Making Sense of Big Data: Developing a Social Science Research Agenda

June 18, 2013
ICA Annual Conference
London, UK

Matthew Weber
Rutgers University
School of Communication & Information

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY



What is Big Data?

“Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”

-McKinsey Global Institute, Big data: the next frontier for innovation, competition, and productivity, May 2011

RUTGERS
School of Communication and Information

Opportunity: The Internet Archive contains the largest single record of the history of the World Wide Web from 1995 to the present—a wealth of untapped research data.

↓

Challenge #1: There is a significant lack of research-ready databases and tools available to the scholarly community

And who really needs 8.5PB of data?

3

The screenshot shows the Internet Archive homepage. At the top, it features the Internet Archive logo and navigation links for Web, Video, Texts, Audio, Projects, About, Account, TVNews, and OpenLibrary. A search bar is prominently displayed with the text "Hello KrisCarpenter" and an "Upload" button. Below the search bar, there are several content blocks:

- Announcements:** A notice about a 3-for-1 match for Internet Archive donations, stating that 10,000,000,000,000 bytes have been archived.
- Web:** A section for the Wayback Machine, showing a search input field and a "Take Me Back" button.
- Welcome to the Archive:** A brief introduction to the Internet Archive as a non-profit digital library.
- Video:** A section with 1,057,367 movies, featuring a "Curator's Choice" of "Grand Old Opry - 28/April/1956" with a 1-star rating.
- Live Music:** A section with 109,517 concerts, featuring a "Curator's Choice" of "Bob Weir Live at Thomas Wolfe Auditorium on..." with a 5-star rating.
- Audio:** A section with 1,447,101 recordings, featuring a "Curator's Choice" of "theta - labworks [quiz037]" with an average rating.
- Texts:** A section with 3,738,910 texts, featuring a "Curator's Choice" of "Notes historiques de biographies" with an average rating.

CNN.com WITH FREE VIDEO Member Services International Edition | Netscape MAKE CNN.com YOUR HOME PAGE

SEARCH THE WEB CNN.com SEARCH Powered by Yahoo! search

Updated: 4:41 a.m. EDT (08:41 GMT), August 27, 2005

Louisiana and Mississippi declare emergencies



People along the Gulf Coast got some very unwelcome news just before the weekend — Hurricane Katrina appears headed their way and may be a destructive monster by the time it arrives. Governors of Louisiana and Mississippi declared states of emergency Friday. Damages from Katrina are estimated at between \$600 million and \$2 billion.

FULL STORY

- Watch: [Forecast](#) | [Death and destruction](#)
- Watch: [Helicopter Rescue](#) | [Citizen Journalist](#)
- Special Report: [Katrina's predicted path](#)
- Gallery: [Storm damage](#) | [Your pictures](#)
- [Helicopter rescue for stranded family](#)

MORE NEWS Most Popular

- [Brothers held again in Natalee Holloway case](#) | Watch
- [Al Qaeda in Iraq issues virulent manifesto](#)
- [SI.com: Rolling Stones concerts trash Red Sox's field](#)
- [Ford gives sneak peek at sleek new Jaguar](#)
- [CNN/Money: Two nabbed in computer worm case](#) | Watch
- [Panel cites teens for delaying morning-after pill decision](#)
- [Vote to save bases in 2 states, lawsuit in third](#) | Watch
- [Slim Jim the firefighter rescues toddler](#) | Watch
- [Oprah 'furious' at snubbing allegations](#)

WATCH FREE VIDEO Browse/Search

[Pregnant teen problem](#) (3:30)  [Baby panda](#) (:40) 

Now In The News: Your quick news update

Featured Video More Video Picks

BUSINESS at CNNMoney ON CNN TV Schedule

Markets: As of close Aug 26

DOW	▼ -53.34	10,397.29
NAS	▼ -13.60	2,120.77
S&P	▼ -7.29	1,205.10

© BigCharts

Enter Symbol: GET

Tune In: **8 p.m. ET**

Dead Wrong: Inside an Intelligence Meltdown
How did U.S. intelligence get it so wrong? CNN investigates how faulty intelligence led to the Iraq war and its affect on U.S. credibility.

CNN.com WITH FREE VIDEO Member Services International Edition | Netscape MAKE CNN.com YOUR HOME PAGE

SEARCH THE WEB CNN.com SEARCH Powered by Yahoo! search

Updated: 2:22 a.m. EDT (06:22 GMT), August 29, 2005

'Big Easy' braces for the 'big one'



Residents of New Orleans and the surrounding area jammed highways Sunday, evacuating from the path of Hurricane Katrina. A direct hit from the huge storm could be "potentially catastrophic" for New Orleans with officials warning of power outages and widespread flooding. But more than 10,000 people opted to remain in the city, taking shelter at the Louisiana Superdome.



FULL STORY

- Watch: [FEMA prepares](#) | [Hurricane insurance](#)
- Watch: [Citizen journalist video](#) | [Seeking shelter](#)
- Special Report: [File citizen journalist stories](#)
- Interactive: [Projected path](#) | [Progress](#) | [Gallery](#)

MORE NEWS Most Popular

- [Katrina may be 'our tsunami'](#) | Watch
- [CNN/Money: Katrina slashes Gulf oil output](#) | Watch
- [Suicide bomber injures 21 in Israel](#) | Watch
- [Committee signs Iraq's draft constitution](#) | Watch
- [Sharpton, Martin Sheen visit antiwar camp](#) | Watch
- [Report: More journalists killed in Iraq than Vietnam](#)
- [Police hunt man who shot rap mogul](#)
- [Big night for Green Day at MTV awards](#) | Gallery
- ['Virgin' stays on top at box office](#)

WATCH FREE VIDEO Browse/Search

[Worst case scenario](#) (4:31)  [What about the pets?](#) (1:39) 

Now In The News: Your quick news update

Featured Video More Video Picks

BUSINESS at CNNMoney ON CNN TV Schedule

Markets: As of close Aug 26

DOW	▼ -53.34	10,397.29
NAS	▼ -13.60	2,120.77
S&P	▼ -7.29	1,205.10

© BigCharts

CNN TV American Morning
Was Katrina underestimated? We'll look at damage and tell you how much steam the storm has left. CNN, your Hurricane

RUTGERS School of Communication and Information
 WITH FREE VIDEO
 CNN.com International Edition Netscape
 Member Services MAKE CNN.com YOUR HOME PAGE
 SEARCH THE WEB CNN.com SEARCH Powered by YAHOO! search

Updated: 1:38 p.m. EDT (17:38 GMT), August 31, 2005

'NIGHTMARE'



Rescuers, residents struggle

- Katrina death toll estimated to be at least 120
- Superdome refugees moving to Astrodome
- Breached New Orleans levee an "engineering nightmare"
- Mississippi governor: More damage than Camille

FULL STORY

- Watch: [Levee breached](#) | [Dome must be evacuated](#)
- Watch: [For Slidell, very bad bump in road](#)
- Storm Roundup: [Katrina's effects at a glance](#)
- Special Report: [Trail of destruction](#) | [Gallery](#) | [Levee system](#)
- Citizen Journalist: [Your e-mails](#) | [Gallery](#) | [Share your story](#)
- Hotlines: [Where to get and give help](#)

David Kelfer, leads his sister and son through flooded streets in New Orleans today.

SERVICES

- E-mail Newsletters
- Your E-mail Alerts
- RSS [HOME](#)
- CNNtoGO
- Contact Us

MORE NEWS [Most Popular](#)

- [Superdome refugees moving to Texas](#) | [Watch](#)
- [Feds fan out for storm recovery efforts](#) | [Watch](#)
- [Wife makes desperate trip with husband's body](#)
- Watch: [Wife searches for missing husband](#)
- ["It's like being in a Third World country"](#) | [Watch](#)
- [Mayor blasts failure to patch levees](#) | [Watch](#)
- [Bush may visit hurricane sites](#) | [Watch](#)

WATCH FREE VIDEO [Browse/Search](#)

- [Family faces uncertain future](#) (2:23)
- [A survivor's story](#) (3:25)

Now in **The News (1:20 p.m. ET)**: Your quick news update

[Featured Video](#) [More Video Picks](#)

RUTGERS School of Communication and Information

Challenges in Tool Development

- Challenging to develop longitudinal data sets on an ad-hoc basis
 - Longitudinal studies are reduced to annual data summaries (Payne & Thelwall, 2007)
- Existing social science methods for handling data and conducting analysis have not proven adequate for large-scale data (Lazer et. Al. 2009)
- IA is structured implicitly, as patterns of associations and interactions are inferred from the nature of content with the sites themselves rather than by formal structures (Arms et. al. 2006)

8

RUTGERS School of Communication and Information

Opportunity: Scholars from a wide range of fields have the potential to benefit from the advancement of research ready tools to access archival Internet data

Challenge #2: Expertise in large-scale data analysis resides across multiple disciplines

RUTGERS School of Communication and Information

HistoryTracker Tool

20th Century Collection @ RU
Curated Data Sets

Version 2.0

PIG Scripts in Hadoop Environment

RU High-Speed Computing Cluster

Link Lists & Other Data

- 01:09 Action A1
- 01:09 Action A2
- 01:09 Action A3
- 01:09 Action A4
- 01:09 Action A5
- 01:09 Action A6

10

HistoryTracker: The Data

WATs

- A compressed metadata format
 - Essential metadata for many types of analysis
- Avoids barriers to data exchange

- WATs
 - Metadata format
 - Date archived
 - Content type
 - Outlinks
 - Anchor text for links
 - Size of record
 - Pointer to actual content

11



Anderson (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. *Wired*

12

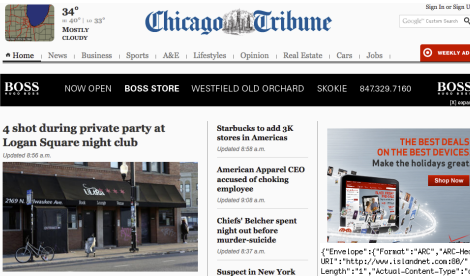
RUTGERS School of Communication and Information

Challenge #3: Data are dirty.

Theory is not dead.

13

RUTGERS School of Communication and Information



4 shot during private party at Logan Square night club
Updated 8:56 a.m.

Starbucks to add 3K stores in Americas
Updated 8:58 a.m.

American Apparel CEO accused of choking employee
Updated 8:58 a.m.

Chief's Belcher spent night out before murder-suicide
Updated 8:57 a.m.

Suspect in New York

THE BEST DEALS ON THE BEST DEVICES. Make the holidays great.
Shop Now

```

{"Envelope":{"Format":"ABC","ARC-Header-Metadata":{"Date":"1996112108034","Content-Length":"4267","Content-Type":"text/html","Target-URI":"http://www.islandnet.com/80/","IPAddress":"207.162.137.20"},"ARC-Header-Length":"74","Payload-Metadata":{"Trailing-Stop-Length":"1","Actual-Content-Type":"application/javascript; charset=utf-8"},"HTTP-Response-Metadata":{"Headers":{"Content-Type":"text/html","Date":"Thu, 21 Nov 1996 19:11:58 GMT","Server":"Apache/1.0.5"},"Headers-Length":"383","Entity-Length":"4264","Entity-Trailing-Stop-Byte":"0","Response-Message":{"Status":"200","Reason":"HTTP/1.0","Reason-Phrase":"OK"},"HTML-Metadata":{"Links":[{"Path":"/BOSS/W/ background","url":"/%26o.gif"},"{"Path":"/IM6/src","url":"/logo.gif"},"{"Path":"/IM6/src","url":"/welcome.gif"},"{"Path":"/IM6/src","url":"/spoon.gif"},"{"Path":"/IM6/src","url":"/island.gif"},"{"Path":"/News & Announcements","path":"/A/ href","url":"/news/press/island.news"},"{"Path":"/IM6/src","url":"/island.gif"},"{"Path":"/About Island Net","path":"/A/ href","url":"/about.html"},"{"Path":"/IM6/src","url":"/island.gif"},"{"Path":"/Contacting Island Net","path":"/A/ href","url":"/contact.html"},"{"Path":"/IM6/src","url":"/island.gif"},"{"Path":"/Services and Pricing","path":"/A/ href","url":"/service.html"},"{"Path":"/IM6/ src","url":"/island.gif"},"{"Path":"/Joining Island Net","path":"/A/ href","url":"/joining.html"},"{"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Policies & FAQs","path":"/A/ href","url":"/policies.html"},"{"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Web Gadgets (GIGs)","path":"/A/ href","url":"/gadgets.html"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Net Web Links","path":"/A/ href","url":"/net.html"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Your account status","path":"/A/ href","url":"/status.html"}," {"Path":"/IM6/ src","url":"/island.gif"}," {"Path":"/Web Server Status","path":"/A/ href","url":"/webstatus.html"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Purchase Time","path":"/A/ href","url":"/purchase.html"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/IM6/src","url":"/A/ href","url":"/http://www.islandnet.com/members/guestbook.html"}," {"Path":"/IM6/src","url":"/island.gif"}," {"Path":"/Other Island Net Servers","path":"/A/ href","url":"/servers.html"}," {"Path":"/FORM/action","method":"post","url":"/cgi-bin/reflect.cgi"}," {"Path":"/support@islandnet.com","path":"/A/ href","url":"/mailto:support@islandnet.com"}," {"Path":"/ISLAND_NEWS","path":"/A/ href","url":"/news// news.islandnet.com/island.news"}," {"Path":"/Island Net A/E Solutions Group Inc.,"path":"/A/ href","url":"/http://www.islandnet.com/"}," {"Header":{"Meta":{"Content":"Island Net is Vancouver Island's Premier Internet Provider","name":"description"},"Title":"Welcome to ISLAND NET"},"Title-Original":"http://www.islandnet.com/","Block-Original":"http://www.islandnet.com/","Actual-Content- Length":"4267"},"Container":{"Container-Type":"text/html","Footer-Length":"1441","Header-Length":"10","Infolated- DRC":"-119286275","Infolated-Length":"4342","Offset":"5581","Filename":"967d799c3ed4c39ed245d81d6c1662_8c-800001.src.org"}
}
ARC-Header-Metadata
ARC-Type: text/html
ARC-Target-URI: http://www.kiss.com/80/
ARC-Date: 1996-11-21 19:08:02
ARC-Record-ID: csm:au:at:66600-5903-4576-2840-2596c20e7959
ARC-Header-Tag: csm:src:967d799c3ed4c39ed245d81d6c1662_8c-800001.src.org:4942
Content-Type: application/json
Content-Length: 2565
  
```

14

HistoryTracker: The Data

WARC/1.0
 WARC-Type: metadata
 WARC-Target-URI: http://state.tn.us/robots.txt
 WARC-Date: 2009-03-12T22:31:30Z
 WARC-Record-ID: <urn:uuid:6fe4e186-97d6-4e39-8c68-932137c281e1>
 WARC-Refers-To: <urn:uuid:129978d2-04cc-4601-8b67-047aeae49fc2>
 Content-Type: application/json
 Content-Length: 1361

```
{
  "Envelope": {
    "Format": "WARC",
    "WARC-Header-Length": 343,
    "Block-Digest": "sha1:R2HAVAYDRXLBXOPOMIZX7IGNSKQJBRZY",
    "Actual-Content-Length": 306,
    "WARC-Header-Metadata": {
      "WARC-Type": "response",
      "WARC-Date": "2009-03-12T22:31:30Z",
      "Content-Length": 306,
      "WARC-Record-ID": "",
      "WARC-IP-Address": "170.143.36.24",
      "WARC-Payload-Digest": "sha1:Q4FXJXFW7O2MF52UIFEEA5BLPKDUTAFU",
      "WARC-Target-URI": "http://state.tn.us/robots.txt",
      "Content-Type": "application/http; msgtype=response",
      "Payload-Metadata": {
        "Trailing-Slop-Length": 4,
        "Actual-Content-Type": "application/http; msgtype=response",
        "HTTP-Response-Metadata": {
          "Headers": {
            "ETag": "\"c51f7-16-34ce5610\"",
            "Date": "Thu, 12 Mar 2009 22:31:30 GMT",
            "Content-Length": 22,
            "Last-Modified": "Tue, 27 Jan 1998 21:48:00 GMT",
            "Content-Type": "text/plain",
            "Connection": "close",
            "Accept-Ranges": "bytes",
            "Server": "Oracle-Application-Server-10g/9.0.4.0.0 Oracle-HTTP-Server",
            "Headers-Length": 284,
            "Entity-Length": 22,
            "Entity-Trailing-Slop-Bytes": 0,
            "Response-Message": {
              "Status": "200 OK",
              "Version": "HTTP/1.1",
              "Reason": "OK",
              "Entity-Digest": "sha1:Q4FXJXFW7O2MF52UIFEEA5BLPKDUTAFU"
            }
          }
        }
      }
    }
  },
  "Container": {
    "Compressed": true,
    "Gzip-Metadata": {
      "Footer-Length": 8,
      "Deflate-Length": 457,
      "Header-Length": 10,
      "Inflated-CRC": "125238600",
      "Inflated-Length": 653,
      "Offset": 677,
      "Filename": "TENN-000001.warc.gz"
    }
  }
}
```

15

Previous Research Sets a Path

- Use of the IA to analyze the frequency of updates based on content analysis (Brock, 2005; Veronin, 2002)
- Measures of age and frequency of updating have been validated against third-party data (Murphy, Hashim and O'Connor, 2005)
- Longitudinal study of e-commerce Web sites to examine development of content (Chu, Leung, Van Hui and Cheung 2007)
- Evolution of organizational hyperlinks in relation to fluctuations in resources (Weber, 2012)

Deciding What Matters

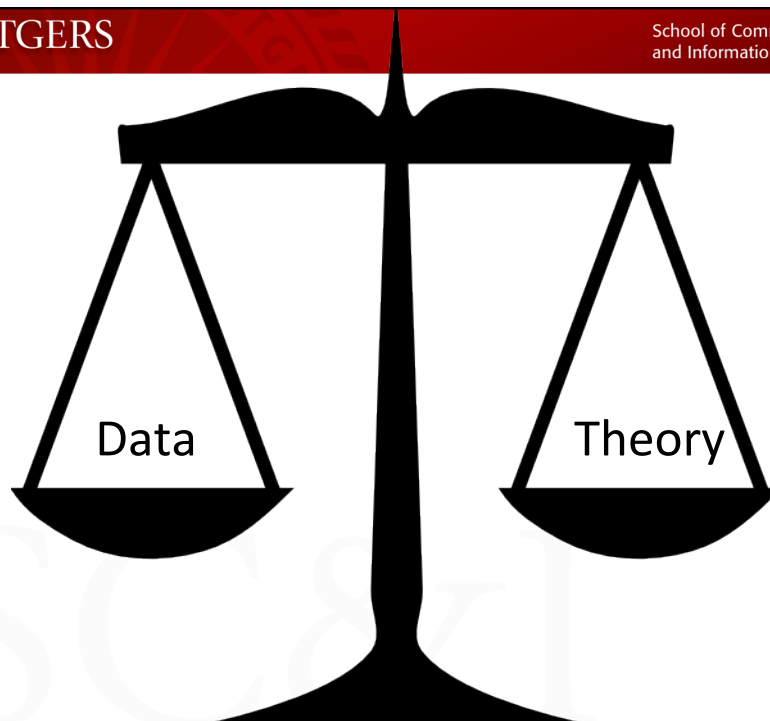
- For a crawl of the 2012 extraction of the Internet:
 - Map output records 1,127,187,232,984 Links

- When does a link matter?

ELEMENT	COUNT
– 	3,212
– 	3,195

- But how do you define what matters?
 - Subcomponents?
 - Text? (anchor text!)

17



18

RUTGERS School of Communication and Information

Research Subset: Political Activity

- Database on 109th, 110th, 111th, 112th congresses
 - .gov domain & congressional pages
 - Full snapshots of the full .gov domain taken at the end of each congressional session
- Research Potential
 - Web promises potential to study the growth of political activity in online environments (Adamic & Glance, 2005; Bruns, 2007; Chang & Park, 2012)

SC&I

19

RUTGERS School of Communication and Information

Research Subset: Media Organizations

- Database on US Media organizations
 - Major media outlets, blogs, opinion sites, etc.
- Synopsis
 - Dataset extracted from a seed list of 4,891 news media organizations
 - 2008 - 2012
 - 1.7 million URL captures from 2008 – 2012
- Research Potential
 - Previous studies of news media organizations (Greer & Mensing, 2006; Weber, 2012; Weber & Monge, 2011) set forth methodology for examining media organizations via online representations

SC&I

20

RUTGERS School of Communication and Information

Research Subset: Social Movements

- Database on social movements
 - Occupy Wall Street
 - Focus on supporting existing avenues of research
- Synopsis
 - 2010 – 2012
 - 528 seed URLs
- Research Potential
 - Previous research on NGOs in the online environment (Bach & Stark, 2004; Shumate, 2003, 2012; Shumate, Fulk, & Monge, 2005)

21

RUTGERS School of Communication and Information

Research Subset: Disasters

- Database on social movements
 - Superstorm Sandy
 - Hurricane Katrina
- Synopsis
 - 2003 – 2012
 - 672 seed URLs
- Research Potential
 - Online networks as enabling organizational resilience (Chewning, Lai and Doerfel, 2012; Perry, Taylor and Doerfel, 2003)

22

RUTGERS School of Communication and Information

- Want to get involved?
 - Email me! matthew.weber@rutgers.edu
 - ResearchHub launching in August – go to netsci.rutgers.edu
 - Workshop: Spring 2014 @ Harvard Institute for Quant. Social Science
- Collaborators
 - Kris Carpenter, Internet Archive
 - David Lazer, Northeastern University
 - Peter Monge, University of Southern California



USC ANNENBERG 
SCHOOL FOR COMMUNICATION

Research support by NSF Award #1244727

23

RUTGERS School of Communication and Information

Research Example

- Political Networks

Subset	Year of Archive	Recorded Links	Reciprocal
109 th	2006	54K	10K
110 th	2008	82K	18K
111 th	2010	253K	34K
112 th	2012	346K	43K

24

