Big Data and Open Data

Bebo White

SLAC National Accelerator Laboratory/
Stanford University

bebo@slac.stanford.edu

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

dekabytes

hectobytes

Hype

MacMania 17, Somewhere@Sea, June 2014

Big Data

Zb

Desktop

Gb

Internet

Pb

Hobbyist

Kb

The Data Deluge

- From the beginning of recorded time until 2003, mankind generated 5 exabytes of data
- In 2011, the same amount of data was generated every two days
- In 2013, the same amount of data was generated every 10 minutes

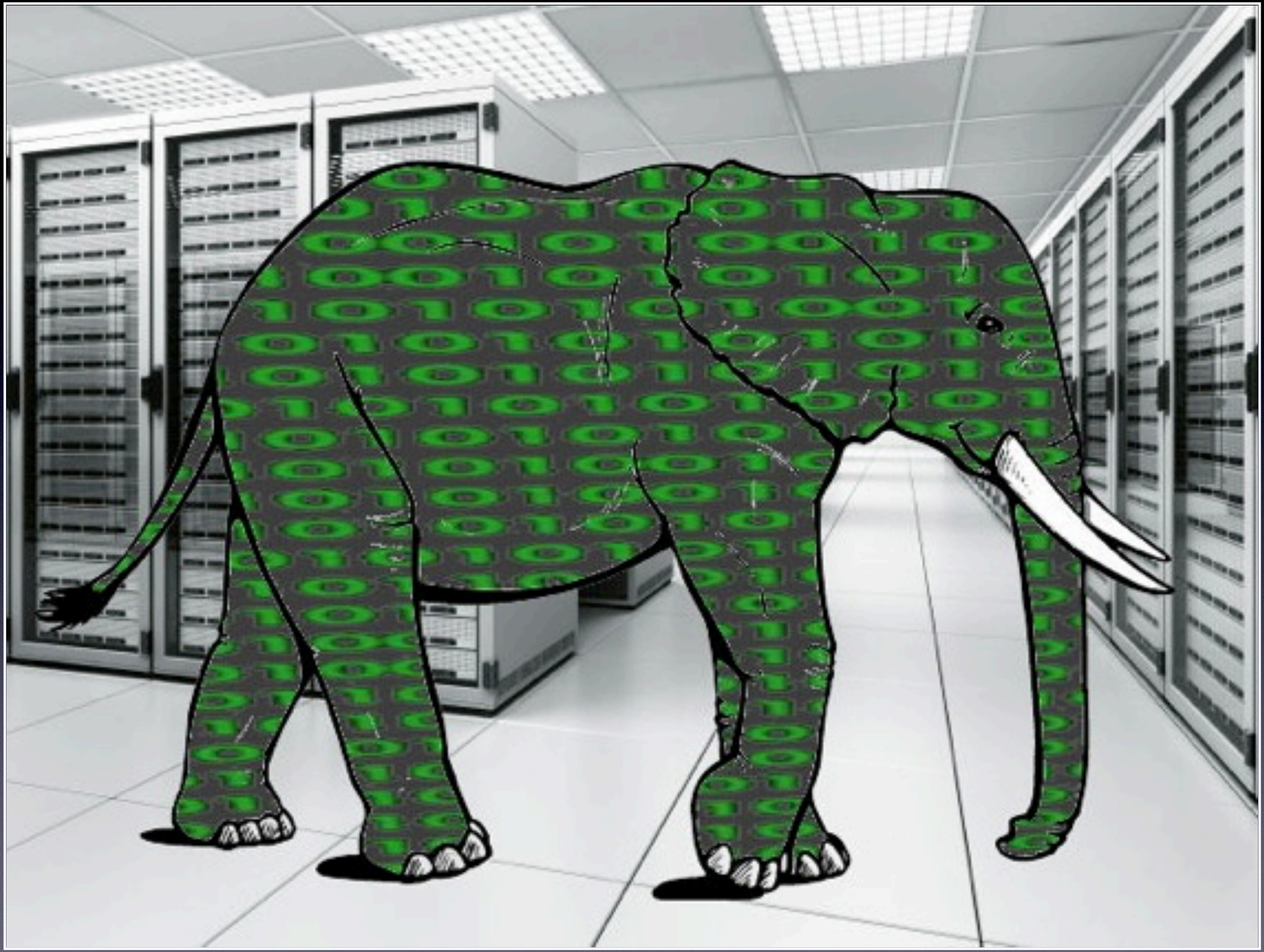

BIG
DATA
ANALYTICS
COMPLEX
SOCIAL
RELATIONAL
CITATION
LARGER
CONTINUES
SET
USE
ALSO
EVERY
TOOLS
DISK
TARGET
APPLIED
SHARED
SENSOR
DEFINITION
EXAMPLES
PRACTITIONERS
CAPTURE
INTERNET
BUSINESS
SETS
DISTRIBUTED
REMOVED
MANAGEMENT
TERABYTES
PETABYTES
INCLUDE
SYSTEMS
TOLERABLE
STORAGE
PARALLEL
SAN
MASSIVELY
GROW
SIZE
MAY
MOVING
WITHIN
ARCHIVES
MANAGE
TECHNOLOGIES
NEEDED
TIME
SOFTWARE
PERFORMANCE
RESEARCH
LOGS
COST
INFORMATION
STORE
BIOLOGICAL
PROCESSING
HUNDREDS
NETWORKS
DATABASES
RECORDS
SEARCH
GARTNER
DIFFICULTY

Big Data - *a Possible Definition*

- Refers to datasets whose size is beyond the ability of
  - Single storage devices
    - Typical database software tools to capture, store, manage, and analyze (McKinsey Global Institute)
- This definition is not defined in terms of data size (which will increase)
- It can vary by sector/usage

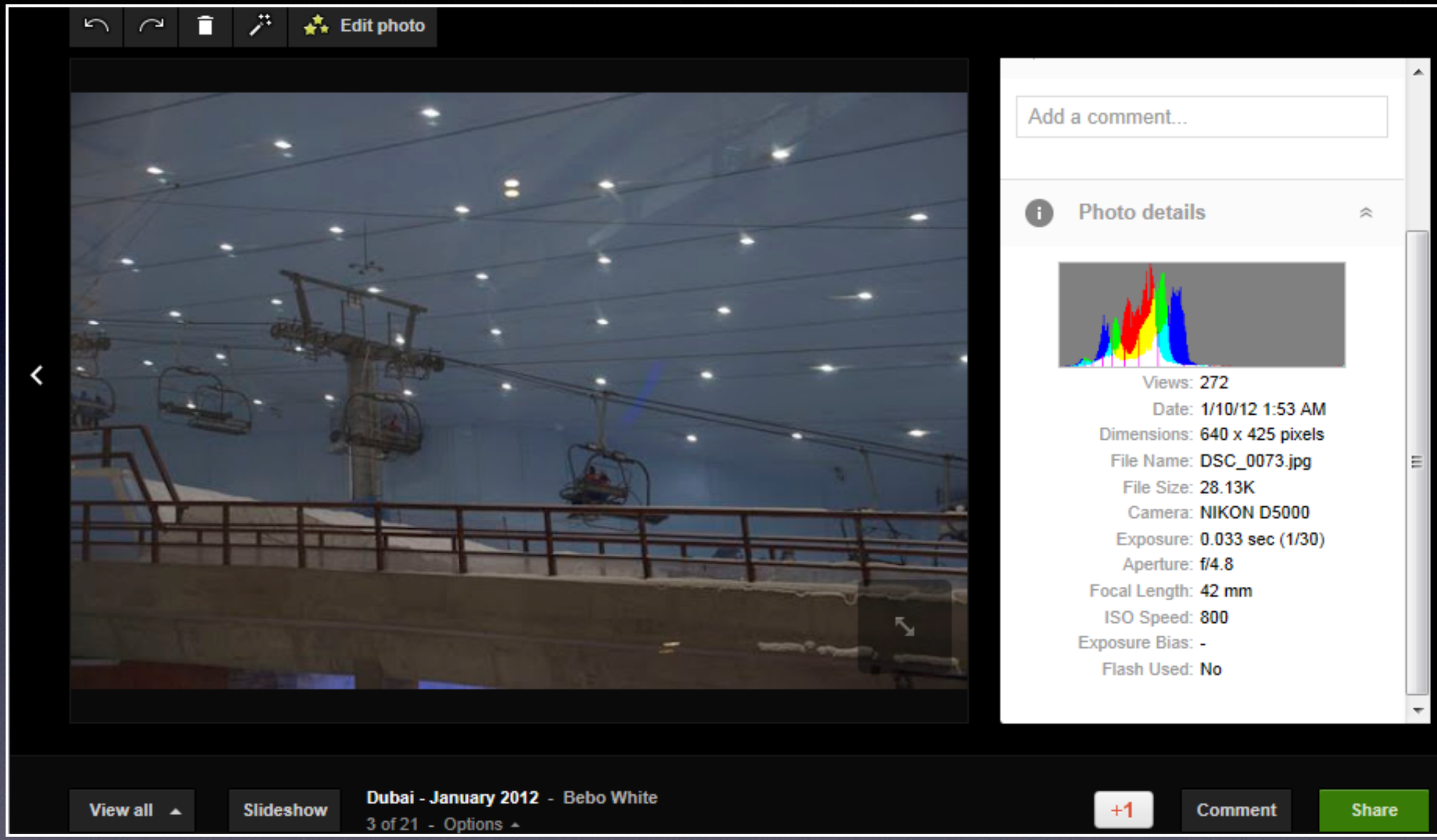
So no cool Big Data apps for Mac or iOS - yet

MacMania 17, Somewhere@Sea, June 2014



Where is Big Data Coming From?

- EVERYWHERE!
- Any communication over a network involves transfer of data that is meaningful to someone
- Every e-mail, every tweet, every transaction, every social media interaction, etc.,etc.
- Sensors - "The Internet of Things"


Edit photo
Add a comment...
Photo details
Views: 272
Date: 1/10/12 1:53 AM
Dimensions: 640 x 425 pixels
File Name: DSC_0073.jpg
File Size: 28.13K
Camera: NIKON D5000
Exposure: 0.033 sec (1/30)
Aperture: f/4.8
Focal Length: 42 mm
ISO Speed: 800
Exposure Bias: -
Flash Used: No
View all
Slideshow
Dubai - January 2012 - Bebo White
3 of 21 - Options
+1
Comment
Share

Google™
processes 20 PB a day (2008)
crawls 20B web pages a day (2012)

JPMorganChase
150 PB on 50k+ servers
running 15k apps (6/2011)

ebay
>10 PB data, 75B DB
calls per day (6/2012)

INTERNET ARCHIVE
Wayback Machine: 240B web
pages archived, 5 PB (1/2013)

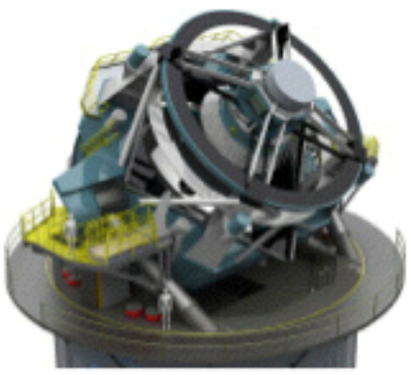
>100 PB of user data +
500 TB/day (8/2012)

facebook

LHC: ~15 PB a year

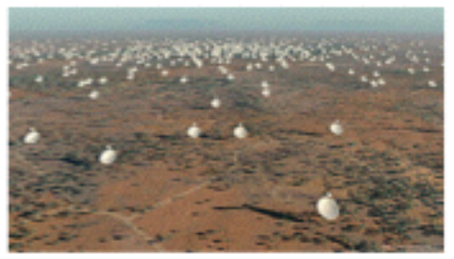
CERN

amazon web services™
S3: 449B objects, peak 290k
request/second (7/2011)
1T objects (6/2012)



LSST: 6-10 PB a year
(~2015)

SKA: 0.3 – 1.5 EB
per year (~2020)



How much data?

The Cloud/IOT/WOT is/ will be a very "noisy" place

- An unbelievable of objects (theoretically more than 10e38) will be able to talk to us and to each other (orders of magnitude more than now)
- We will be interested in hearing what *some* of them have to say
- How can we manage these conversations?
- Traditional interfaces break down

Why big data?

Science
Engineering
Commerce

Source: Wikipedia (Everest)

Science

Emergence of the 4th Paradigm

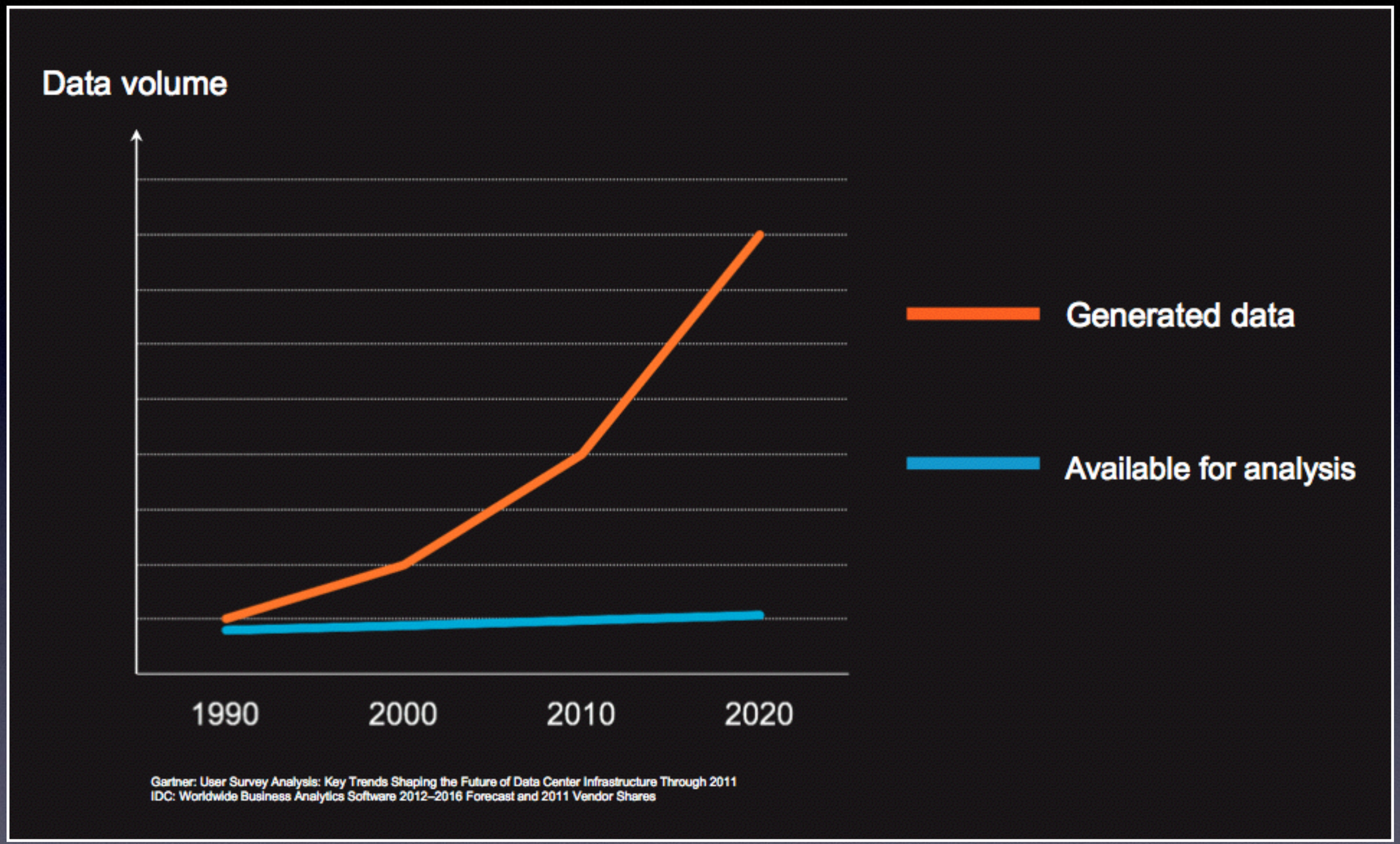
Data-intensive e-Science

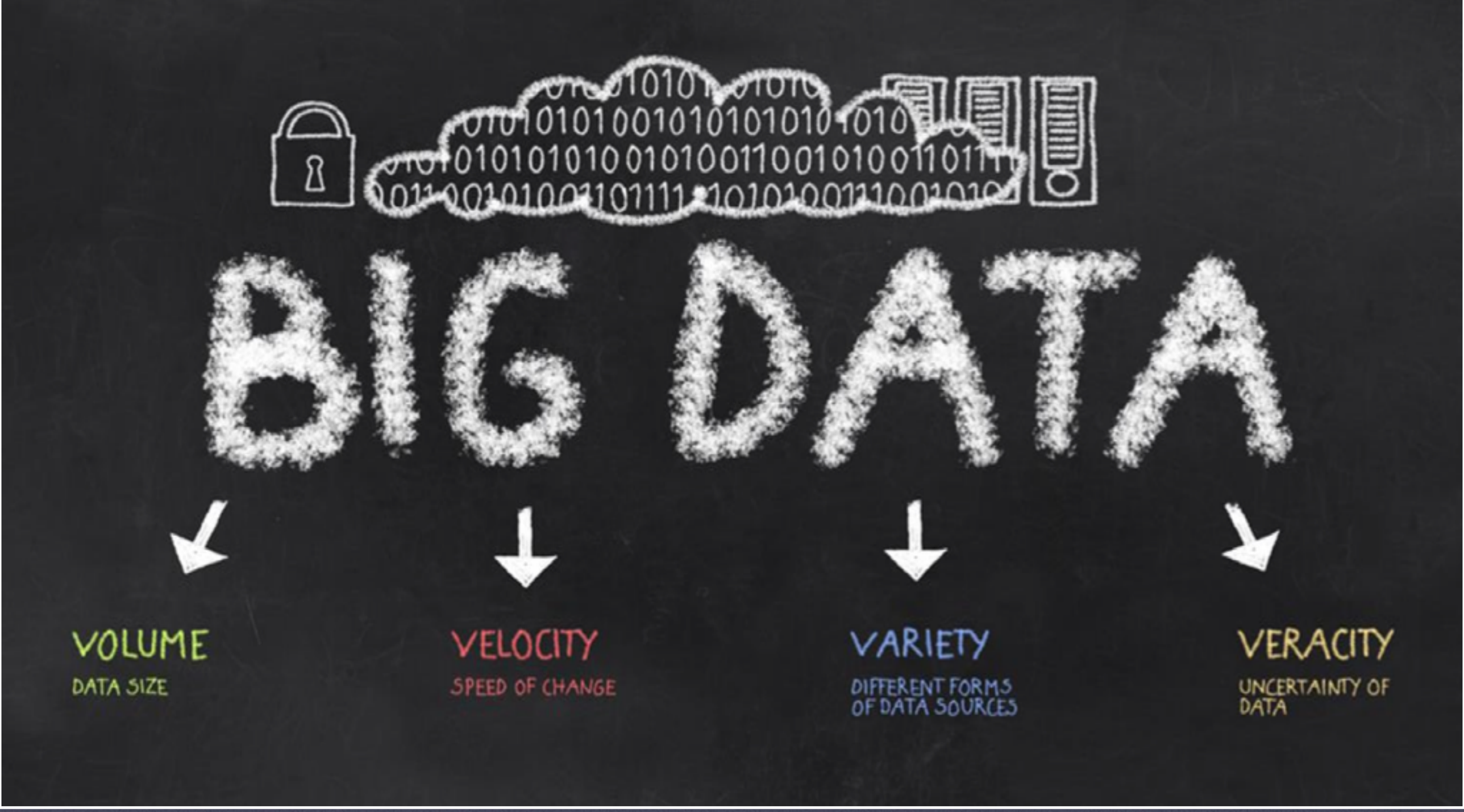
Know thy customers

Data → Insights → Competitive advantages

Commerce

Source: Wikiedia (Shinjuku, Tokyo)


Data volume
Generated data
Available for analysis
1990
2000
2010
2020
Gartner: User Survey Analysis: Key Trends Shaping the Future of Data Center Infrastructure Through 2011
IDC: Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares



"High-volume, -velocity, and -variety information assets that demand cost-effective innovative forms of information processing for enhanced insight and decision making"

MacMania 17, Somewhere@Sea, June 2014

- Volume?
  - ~ data volume worldwide in 2013 = 3.5 ZB (including 400 billion feature length HD movies)
- Velocity?
  - Every 60 sec. on Facebook - 510K posted comments; 293K status updates; 136K uploaded photos
  - 30 billion shares
  - 20 million apps installed

- Variety?
  - Any type of data both meaningful and meaningless
- Veracity?
  - How is trust established?
  - What does “like” really mean?

Usage

Plus:
ISPs
Utilities
Academic institutions
Everyone

What/How Does Target Know About Pregnant Women?

Challenges of Harnessing Big Data

- Datamining huge datasets
- Shortages of Big Data experts
- Privacy, legal, and social issues
- Strategies for acquiring Big Data - a new form of currency

Big Data Analytics and Data Science

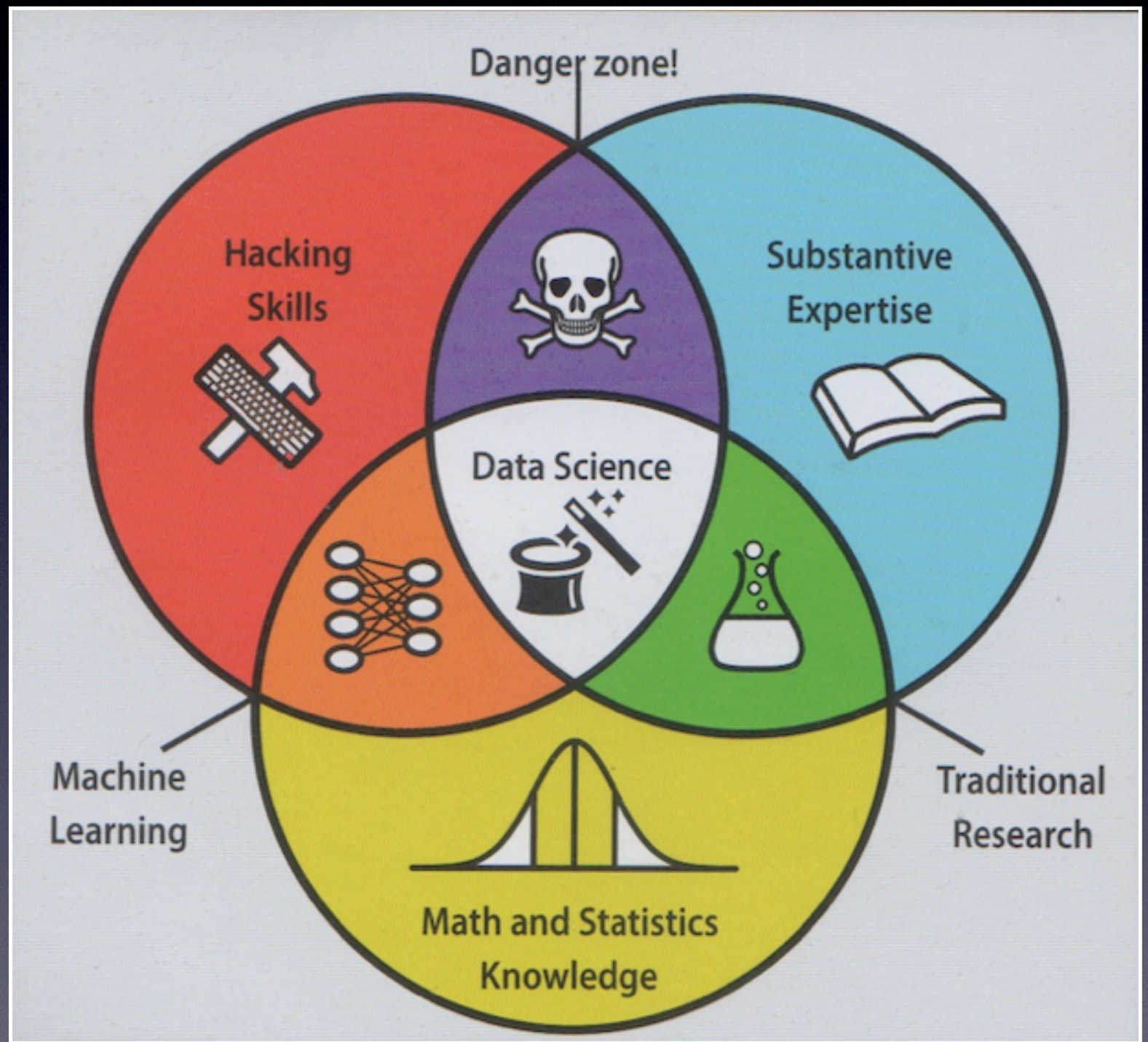
"… [data] analytics is the process of obtaining an optimal or realistic decision based on existing data."

(Wikipedia)

"[data analytics is]…the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions."

(*Competing on Analytics: The New Science of Winning*; Thomas Davenport and Jeanne Harris, Harvard Business Press, 2007)

Data Science Skill Set



DATA SCIENCE WORKFLOW

1 Data acquisition and cleanup

Many **Python** libraries and specialized tools like **OpenRefine** and **Wrangler** aim to lower costs of data cleanup, which can claim up to 80% of development time.

2 Storage and management

Novel tools such as **NoSQL** and **MapReduce** are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.

Analysis often involves revisiting raw data

3 Analysis

Data scientists who use open-source tools such as statistical packages in **R** and **Python** report higher salaries than those who use commercial software.

4 Visualization

Flexible visualization tools such as **D3.js** and **Processing** extract insight from data and easily integrate with existing frameworks.

5 Communication

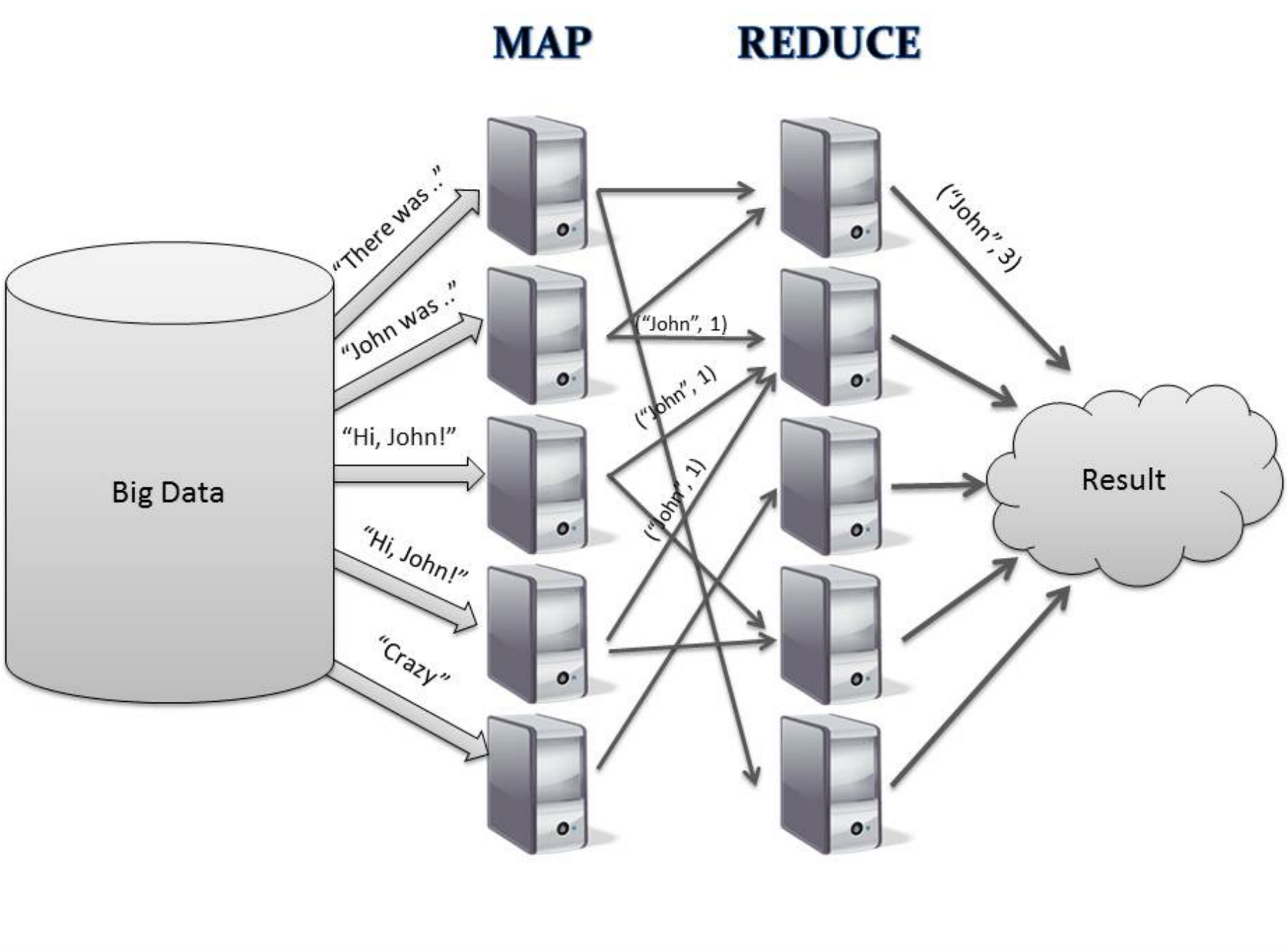
Collaborative services such as **GitHub** and **Bitbucket** simplify sharing code and distributing results, which in turn increases reproducibility.

MapReduce

Source: Google

Typical Big Data Problem

- Iterate over a large number of records
- Extract something of interest from each **(MAP)**
- Shuffle and sort immediate results
- Aggregate immediate results **(REDUCE)**
- Generate final output


MAP
REDUCE
"There was .."
"John was .."
"Hi, John!"
"Hi, John!"
"Crazy"
Big Data
("John", 1)
("John", 1)
("John", 1)
("John", 3)
Result

MapReduce Can Refer to…

- The programming model
- The execution framework (aka “runtime”
- The specific implementation

MapReduce Implementations

- Google has a proprietary implementation in C++
  - Bindings in Java, Python
- Hadoop is an open-source implementation in Java
  - Development led by Yahoo!, now an Apache project
  - Used in production at Yahoo!, Facebook, Twitter, LinkedIn, Netflix, etc.
  - The *de facto* Big Data processing platform
  - Lots of custom research implementations

Example - “Sentiment Analysis”

- Goal - gauging mood on social network data
- Not a traditional survey or focus group
  - Social sites operate 24/7
  - Timeliness - not subject to time lags
- Useful to marketers, IT, customers, etc.

Difficult Comment Analysis (1/2)

- False negatives - “crying” & “crap” (negative) vs. “crying with joy” & “holy crap!” (positive)
- Relative sentiment - “I bought a Honda Accord” - great for Honda, bad for Toyota
- Compound sentiment - “I love the phone but hate the network”
- Conditional sentiment - “If someone doesn’t call me back, I’m never doing business with them again!”

Difficult Comment Analysis Problems (2/2)

- Scoring sentiment - “I like it” vs. “I really like it” vs. “I love it”
- Sentiment modifiers - “I bought an iPhone today :-)” “Gotta love the telephone company ;-<“
- International/cultural sentiments
  - Japanese - unique emoticons for crying - (;_;)
  - Italians - effusive, grandiose
  - British - drier, less effusive

Linked Data

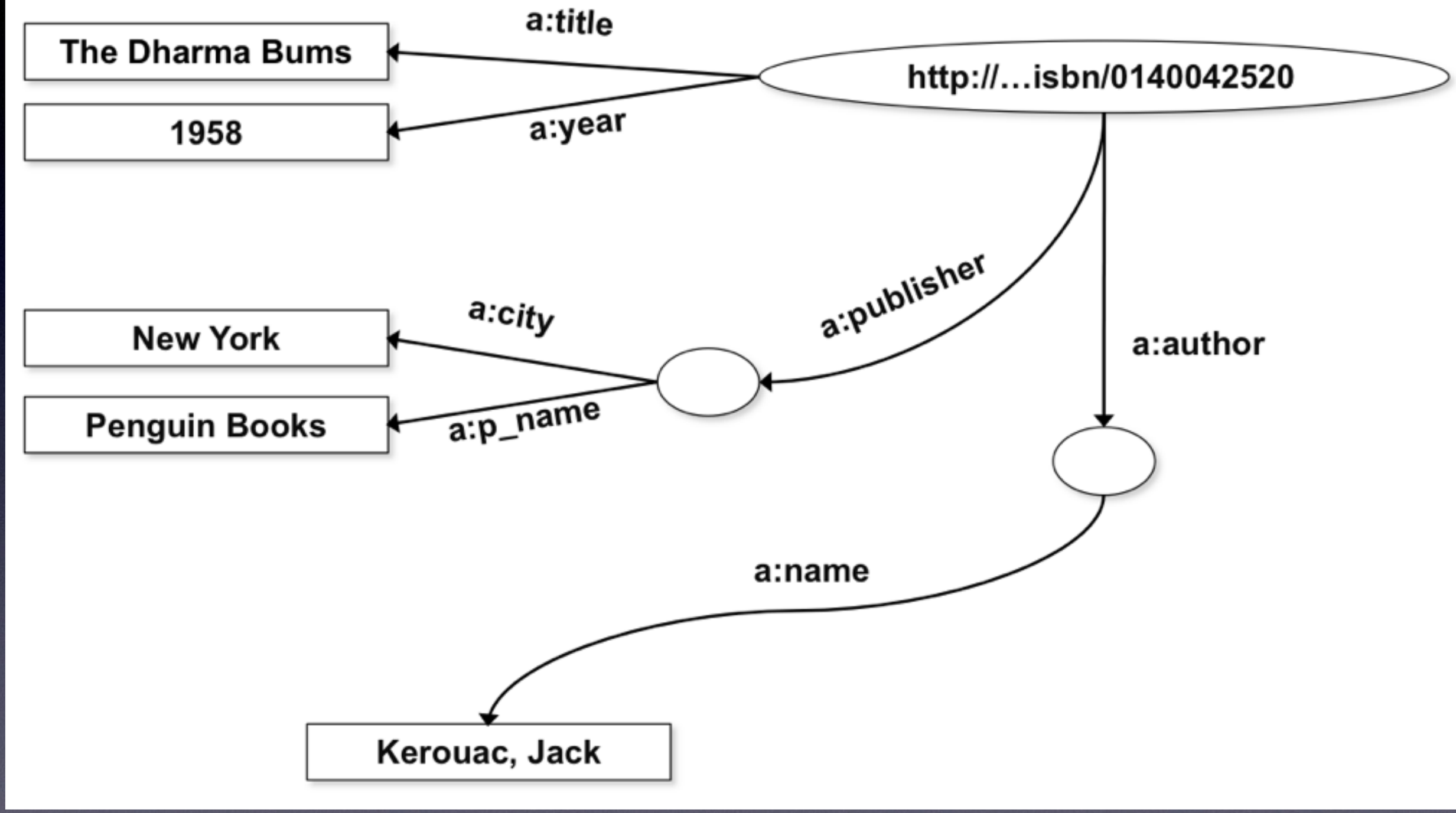
- Provides access to the semantics of data items
- Based upon Semantic Web technologies and ontologies
- Designed for machines first and humans later
- Degree of structure in descriptions of things is high

- Big Data tends to be unstructured data (e.g., lists, e-mails, tweets, etc.)

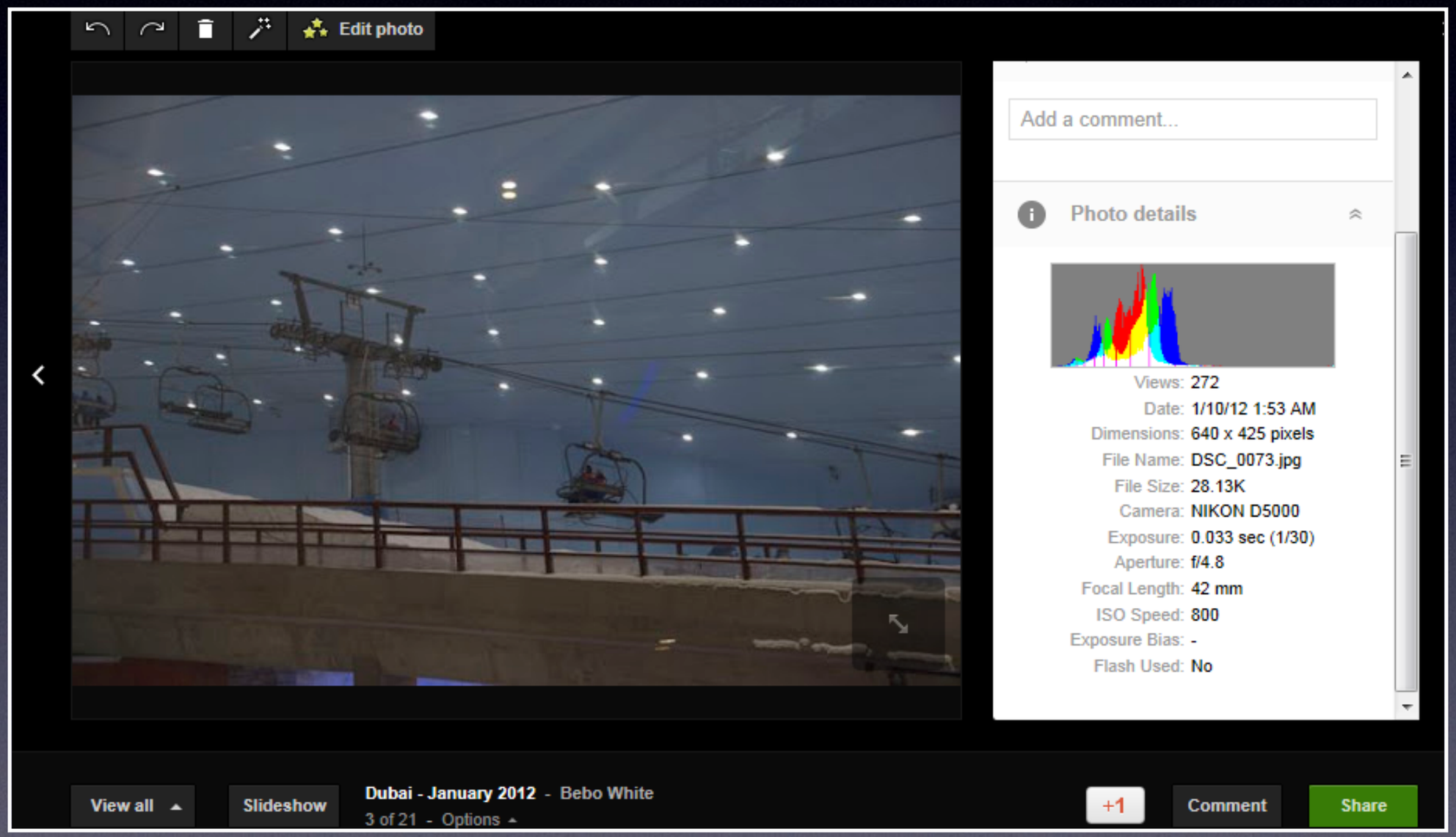
- Therefore it tends to be “thin” rather than “thick”

- “Thin” means very little (if any) context - Suppose I send some e-mail stating “My favorite book is The Dharma Bums”

- What can be added to this data to change it from “thin” to “thick?”


The Dharma Bums
1958
New York
Penguin Books
Kerouac, Jack
a:title
a:year
a:city
a:p_name
a:publisher
a:author
a:name
http://…isbn/0140042520

Linked Data is similar to Metadata but provides Context



Linked Data Pros

- Far more "parseable" and "machine processable" than raw unstructured data
- Enhances data descriptions for complex analyses
- Can contribute to the VERACITY of our data
- Wide variety of discipline/data ontologies available

Linked Data Cons

- Much harder to do than adding keyword metadata
- Building efficient processing applications and parsers
- Implementing effective linked data stores

Linked Open Data

- LOD refers to data stores of Linked Data that are published (made available online and accessed via URLs) and free to use
- Open data means it must be available to all without copyright or ownership
- There is an increasing trend towards "opening" government data (US and UK, San Francisco and more) and scientific results
- Provides unprecedented ability to build "mashup" applications

PUBLISHING

US science to be open to all

Government mandates that taxpayer-funded research be freely available within 12 months.

BY RICHARD VAN NOORDEN

The rumours have been buzzing around Capitol Hill since before last year's election, and last week, supporters of open-access publication in the United States got most of what they wanted. The White House declared that government-funded research would be made free for all to read, rather than kept behind paywalls. However, those hoping that the government would require papers to be free from the time of publication were disappointed.

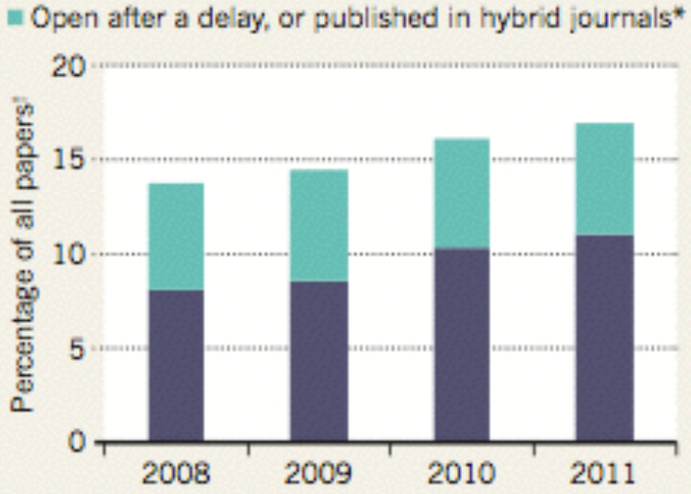
In a 22 February memo, John Holdren, director of the White House's Office of Science and Technology Policy (OSTP), gave federal agencies until 22 August to produce plans for making the data and papers from the research they fund more accessible to the public. The move, he says, would "accelerate scientific breakthroughs and innovation" and boost economic growth. Agencies should aim to make research papers free by 12 months after publication — a concession to publishers, who say that a year's delay is needed to maintain their revenue from subscriptions.

The policy applies to an estimated 19 federal agencies, which each spend more than US$100 million on research and development. It would roughly double the number of articles made publicly available each year to about 180,000, according to the Scholarly Publishing and Academic Resources Coalition, an open-access advocacy group in Washington DC, which called the memo a "landmark". Until now, only the US National Institutes of Health (NIH) has required its research to be publicly available after 12 months.

The latest move is a response to the 2011 reauthorization of the 2007 America COMPETES Act, which included billions of dollars for science, and also charged the OSTP with improving public access to research (see 'Into the open'). Another spur came in May 2012, when thousands petitioned the White House to require free access to journal articles arising from US taxpayer-funded research. Agencies such as the National Science Foundation and the Department of Energy have been laying the groundwork with publishers for the

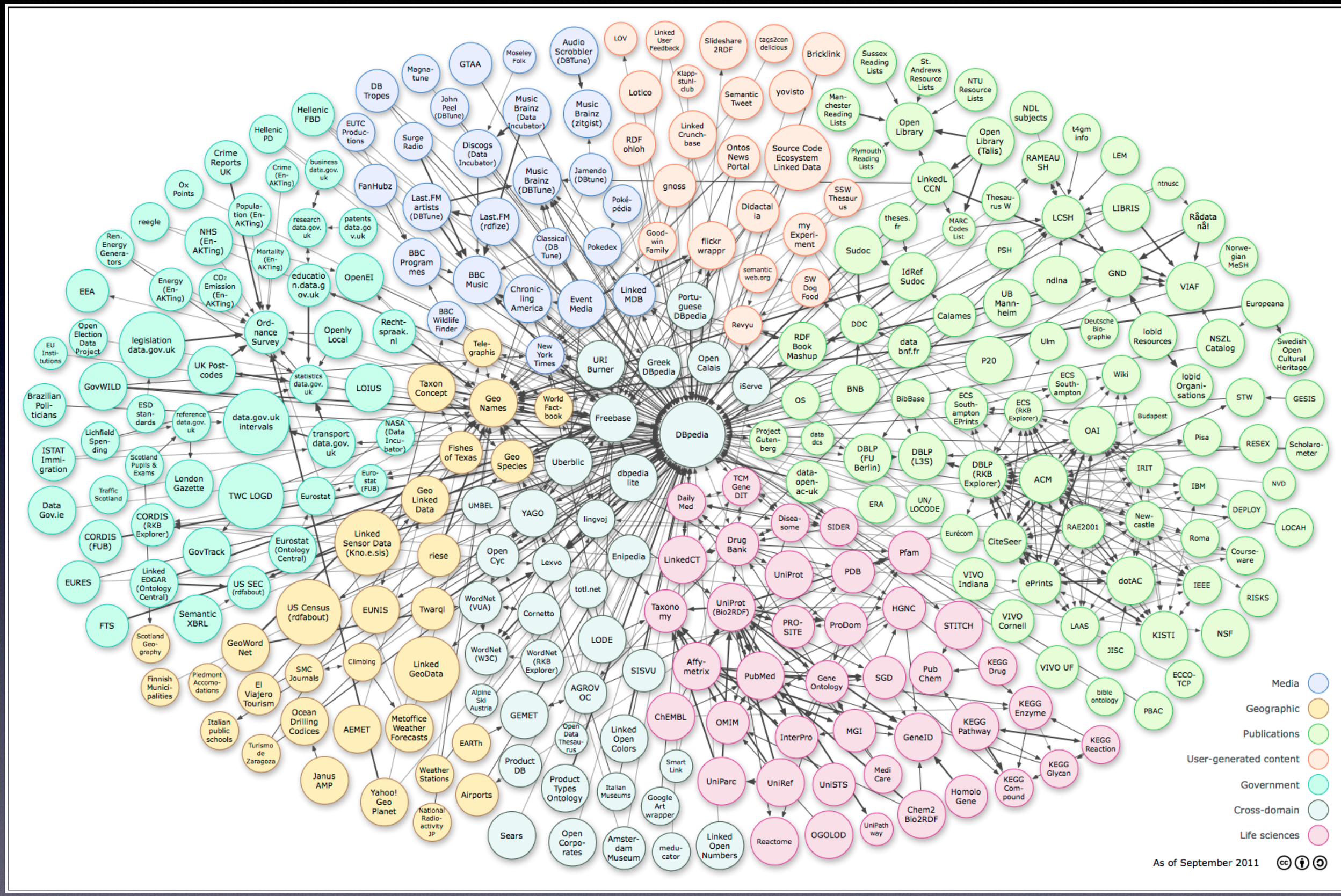
INTO THE OPEN

Publishers are making an increasing proportion of papers free to the public on their websites.



*Hybrid journals: subscription journals that publish some articles open access. †All papers indexed in Elsevier's Scopus database.

SOURCE: M. LAAKSO & B.-C. BJÖRK *BMC MED.* **10**, 124 (2012)


DBpedia
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

But we are here to discuss Mac and iOS

- I believe that there is great future potential for powerful and creative Mac and iOS apps that leverage
  - Big Data analytics results
  - Linked Data and LOD data stores
- Great possibilities in E-Commerce, Education, Productivity, Social Interaction, etc., etc.
- The people in this room can help define, drive and evangelize these concepts!

Thank You!
Questions? Comments?

bebo@slac.stanford.edu

Want copies of these slides?