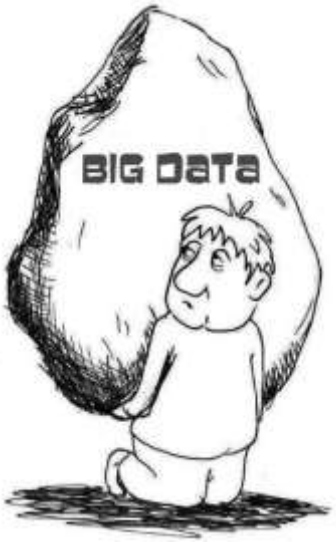


Tutorial: Introduction to Big Data

Marko Grobelnik, Blaz Fortuna, Dunja Mladenic
Jozef Stefan Institute, Slovenia

Sydney, Oct 22nd 2013

Outline

- Big-Data in numbers

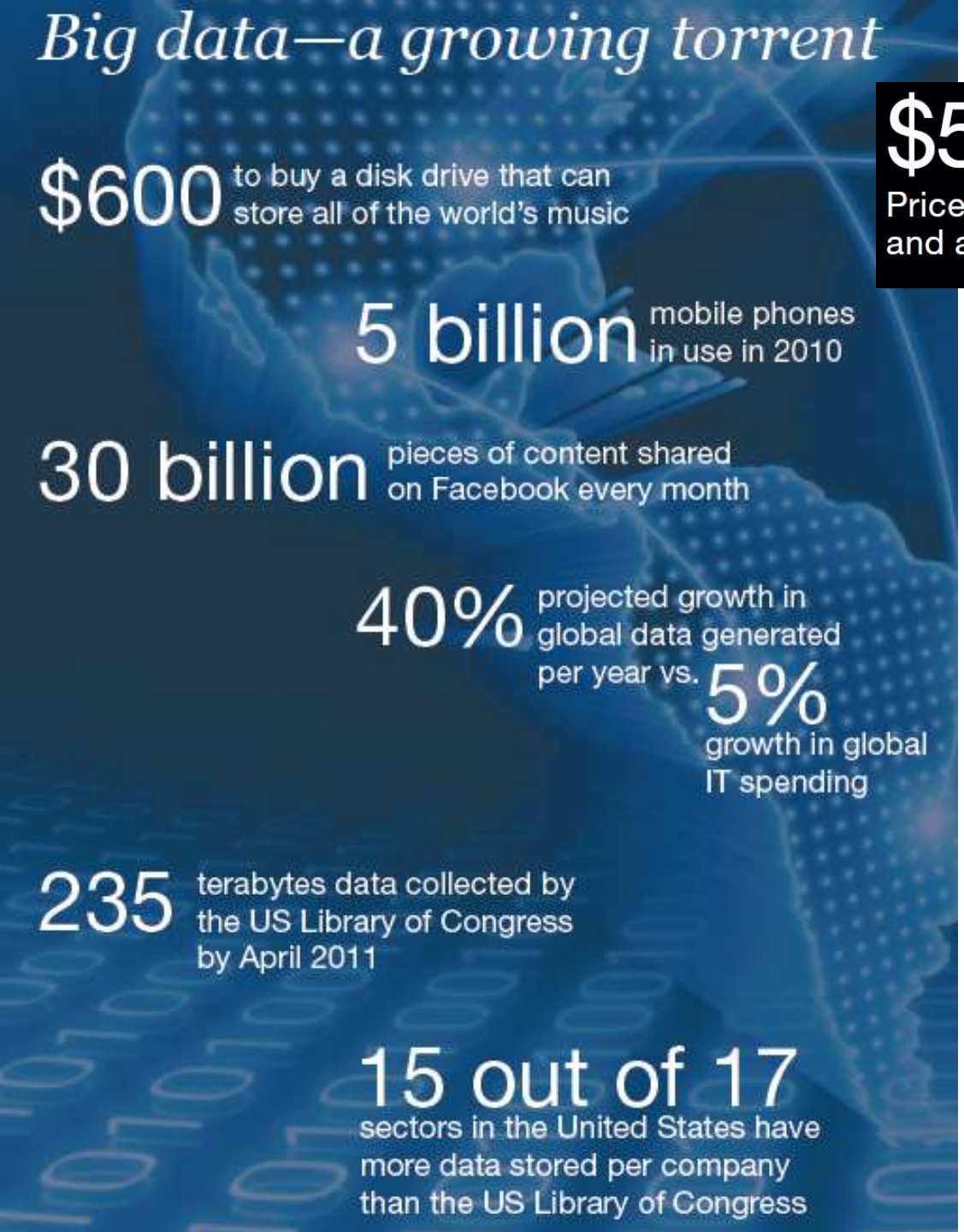
- Big-Data Definitions
- Motivation
- State of Market

- Techniques
- Tools
- Data Science

- Applications
  - Recommendation, Social networks, Media Monitoring

- Concluding remarks

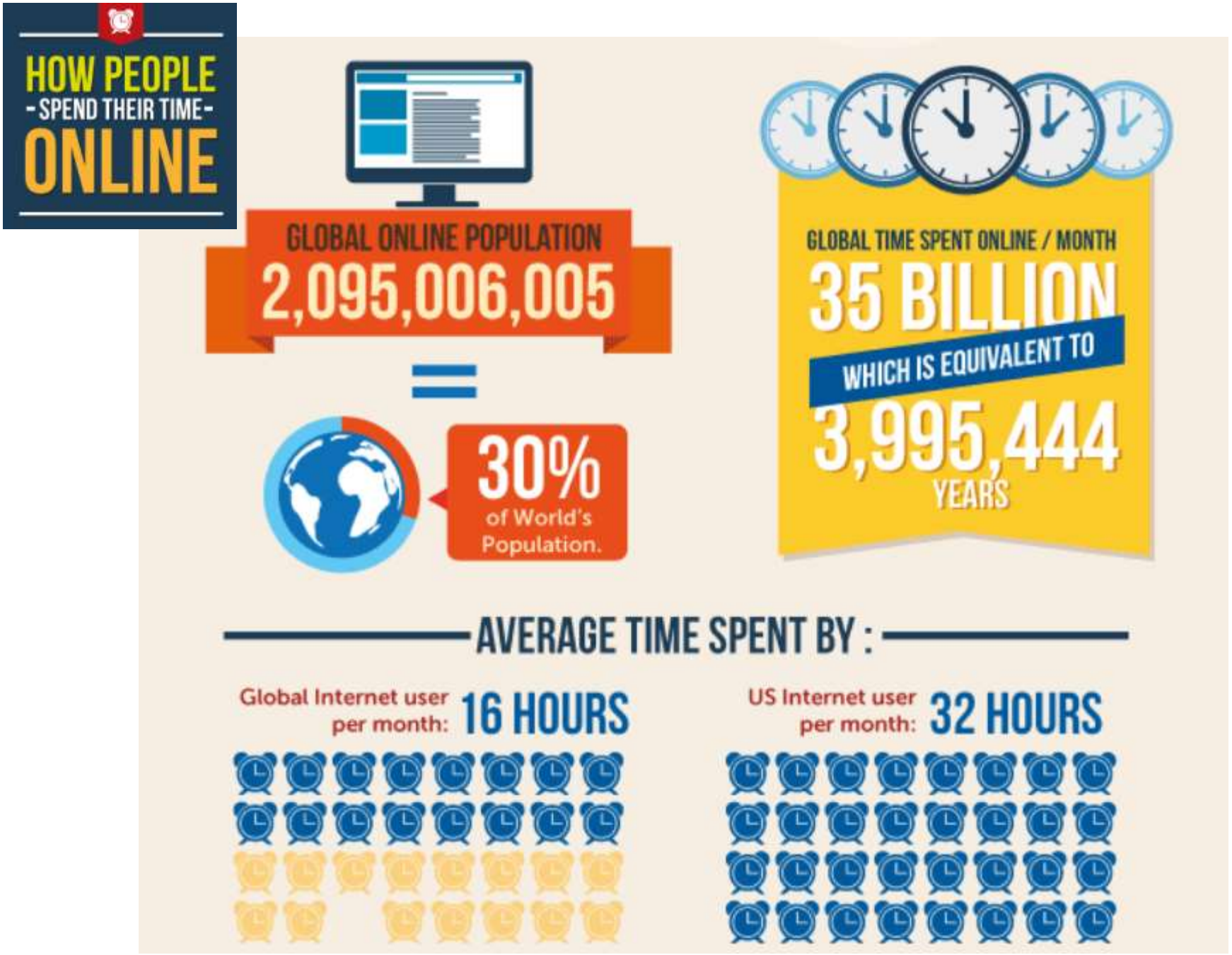
Big-Data in numbers



$5 million vs. $400

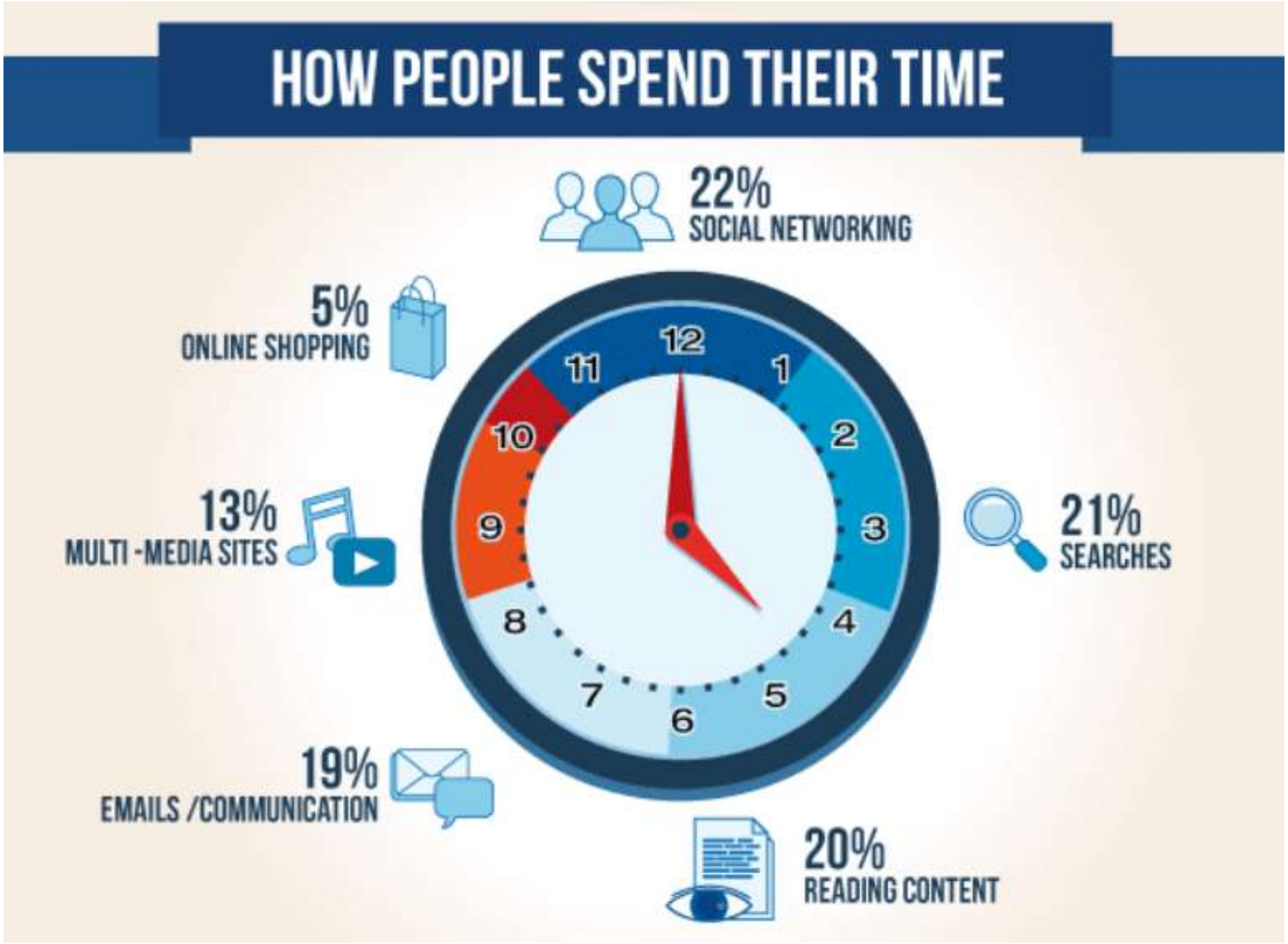
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WordPress DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
Google
Google Search
GO-Globe.com



http://www.go-gulf.com/blog/online-time/

HOW PEOPLE SPEND THEIR TIME



http://www.go-gulf.com/blog/online-time/

TOP 10 MOST VISITED WEB PROPERTIES

	Unique Visitors Per Month	Time Spent Per Person Per Month in hh:mm:ss
YAHOO!	130,121,000	2:12:08
msn bing	115,890,000	1:43:45
YouTube	106,692,000	1:41:27
Microsoft	83,691,000	0:45:05
Aol.	74,633,000	2:52:52
	62,097,000	0:18:03
	61,608,000	1:06:15
Ask	60,552,000	0:12:27

http://www.go-gulf.com/blog/online-time/

INTERESTING FACTS

More than **56%** of Social Networking Users have used Social Networking Sites for spying on their partners.

Brazilians have the highest online friends averaging **481** friends per user, whereas Japanese have the least average of only **29** friends.

Chinese users spend the maximum time of more than **5** hours a week, in shopping online.

More than **1 Billion** Search Queries per day on Google.

4 Billion views per day on Video Sharing Website YouTube. Video content of more than **60 hours** gets uploaded every minute onto YouTube.

More than **250 Million** Tweets per day.

More than **800 Million** updates on Facebook per day

http://www.go-gulf.com/blog/online-time/

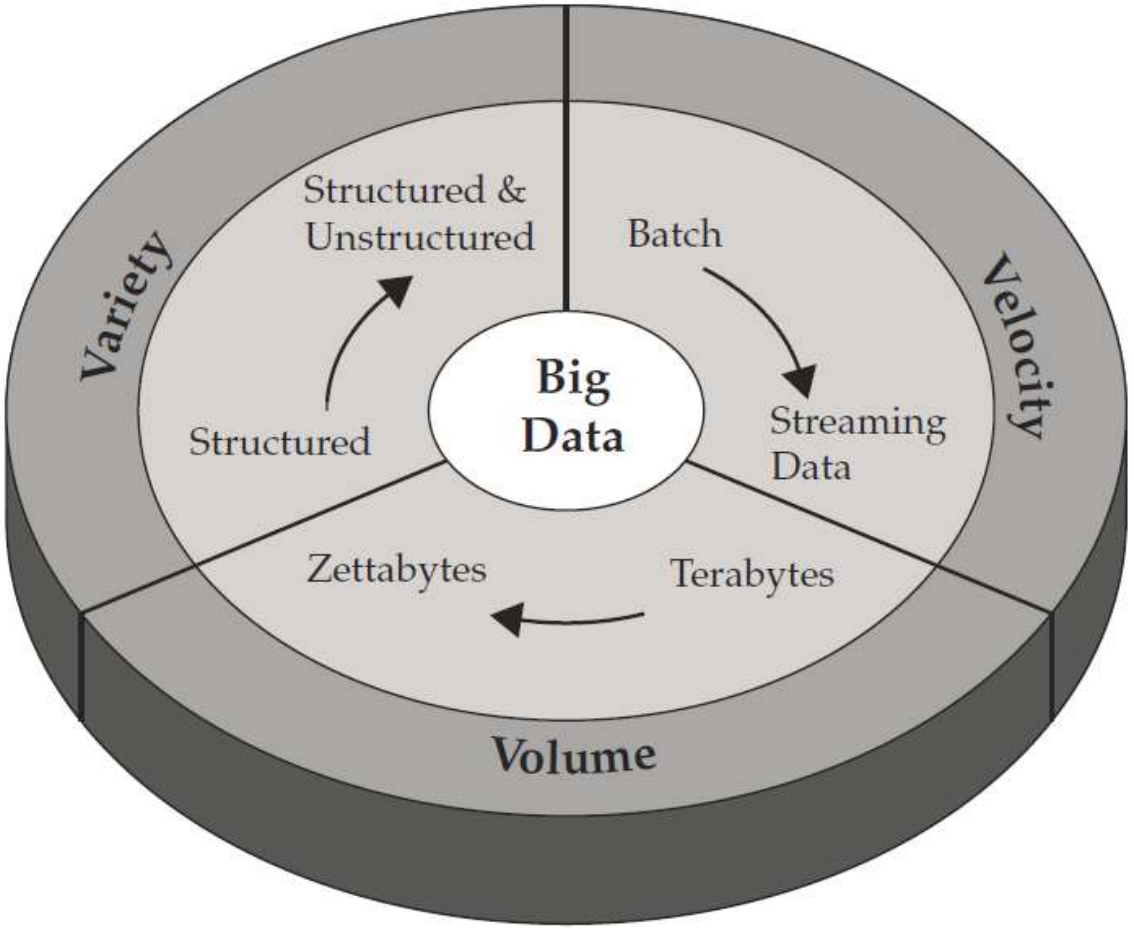
Big-Data Definitions

…so, what is Big-Data?

- ‘Big-data’ is similar to ‘Small-data’, but bigger
- …but having data bigger it requires different approaches:
  - techniques, tools, architectures
- …with an aim to solve new problems
  - …or old problems in a better way.

Characterization of Big Data: volume, velocity, variety (V3)

- **Volume** – challenging to load and process (how to index, retrieve)
- **Variety** – different data types and degree of structure (how to query semi-structured data)
- **Velocity** – real-time processing influenced by rate of data arrival



From "Understanding Big Data" by IBM

The extended 3+n Vs of Big Data

- 1. **Volume** (lots of data = "Tonnabytes")
- 2. **Variety** (complexity, curse of dimensionality)
- 3. **Velocity** (rate of data and information flow)

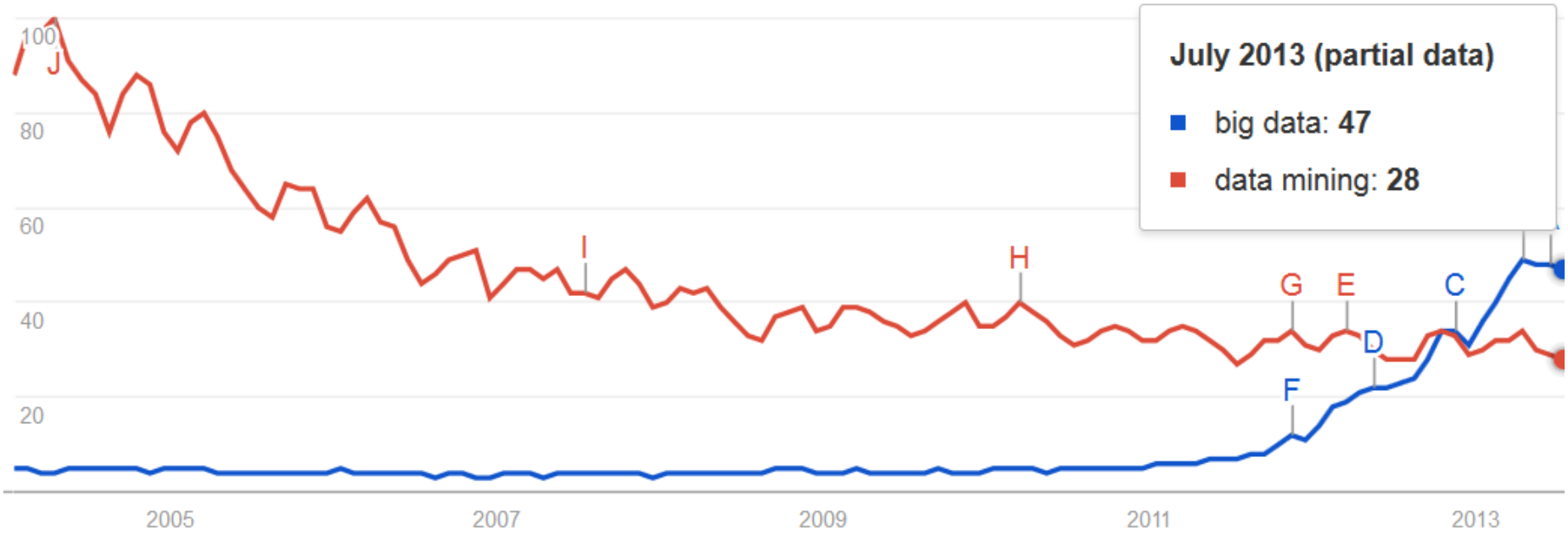
- 4. **Veracity** (verifying inference-based models from comprehensive data collections)
- 5. **Variability**
- 6. **Venue** (location)
- 7. **Vocabulary** (semantics)

Motivation for Big-Data

Big-Data popularity on the Web

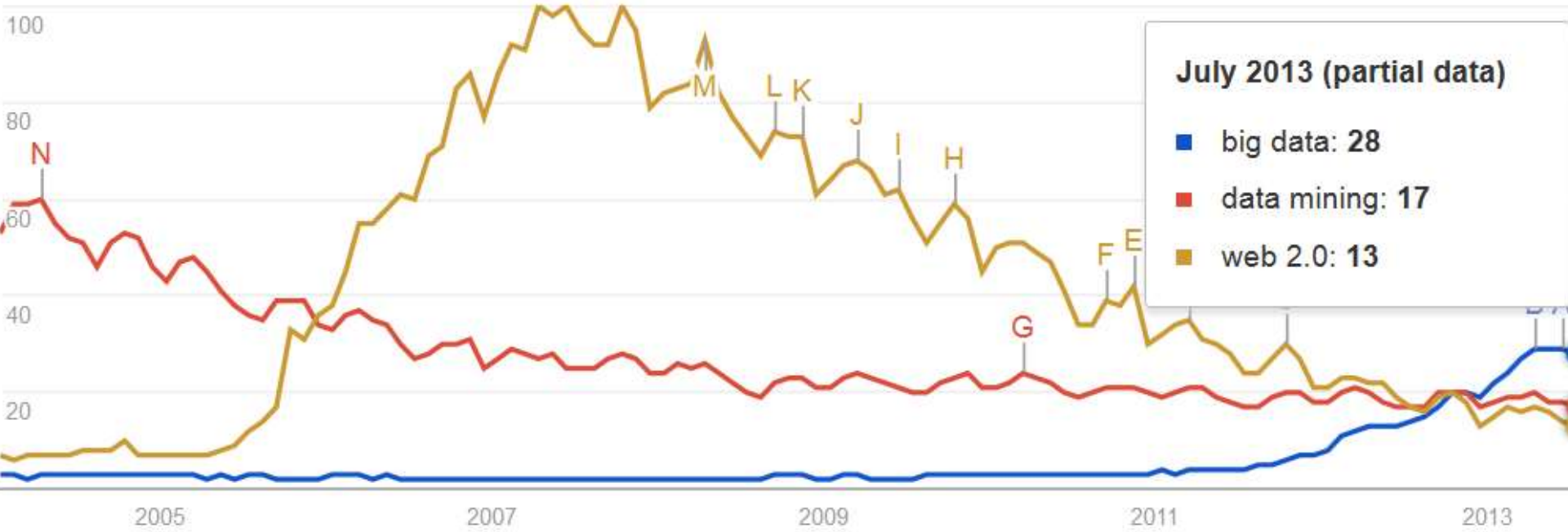
(through the eyes of "Google Trends")

Comparing volume of "big data" and "data mining" queries



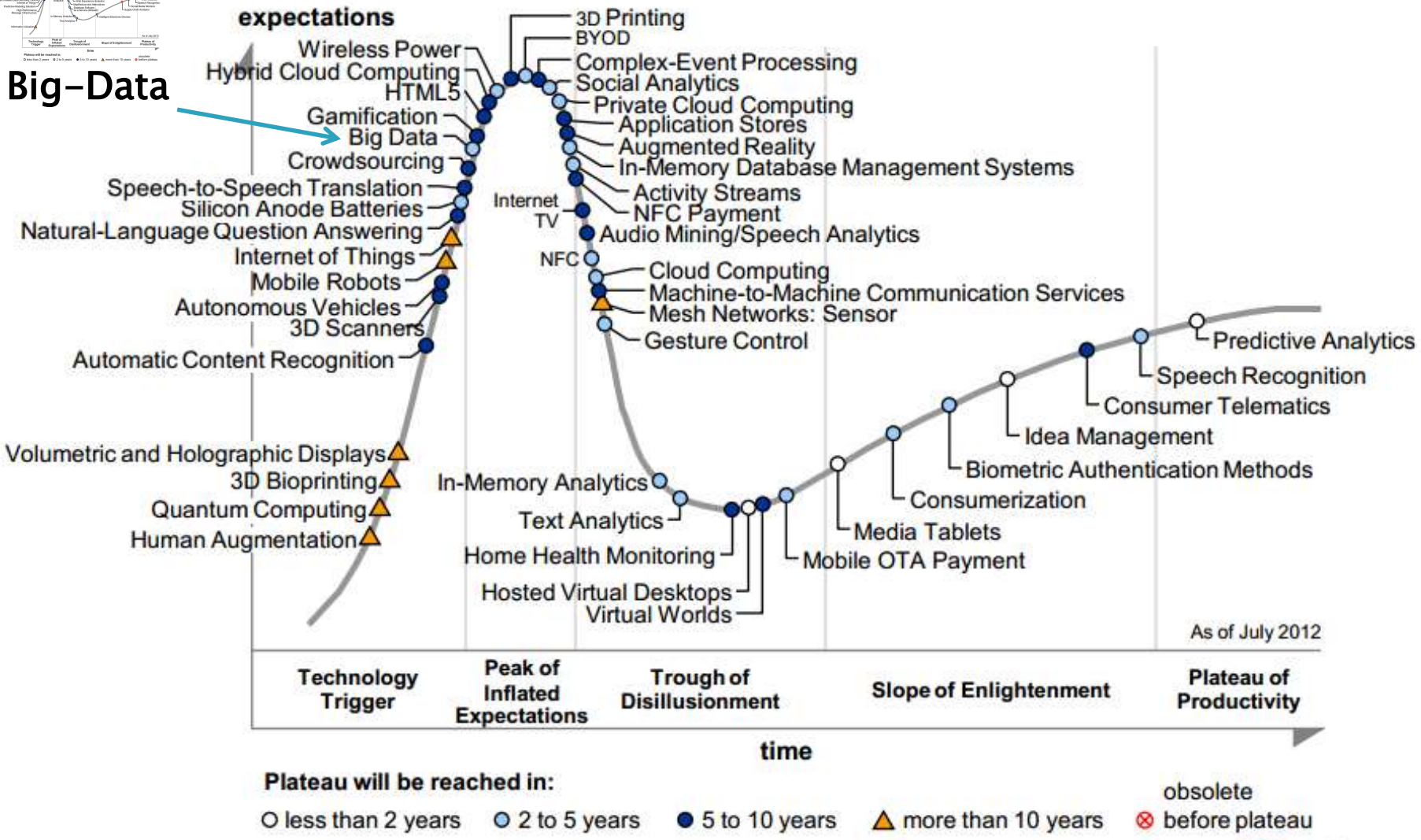
…but what can happen with “hypes”

…adding “web 2.0” to “big data” and “data mining” queries volume



Emerging Technologies Hype Cycle 2012

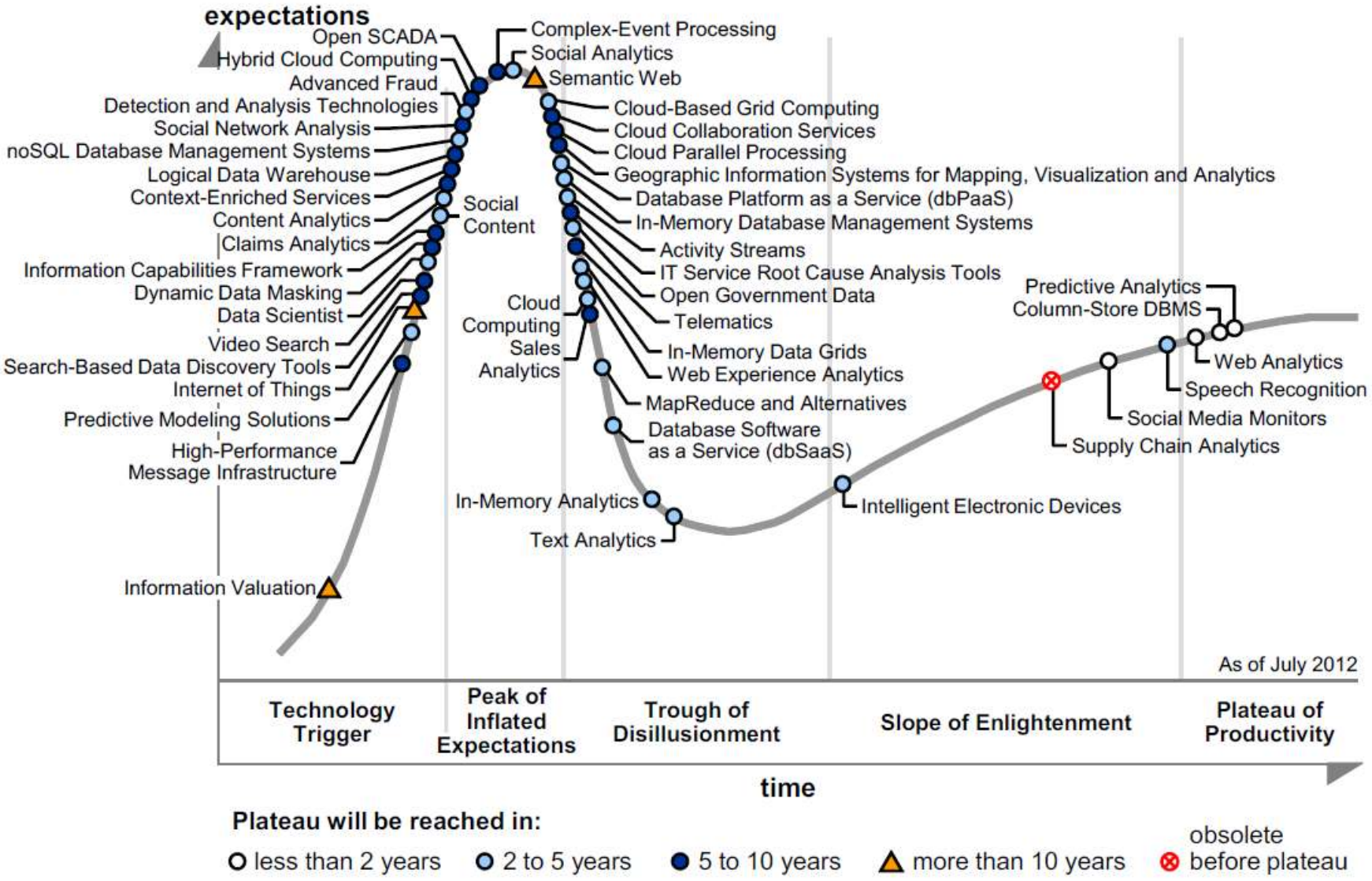
Big-Data



Gartner

Gartner Hype Cycle for Big Data, 2012

Figure 1. Hype Cycle for Big Data, 2012



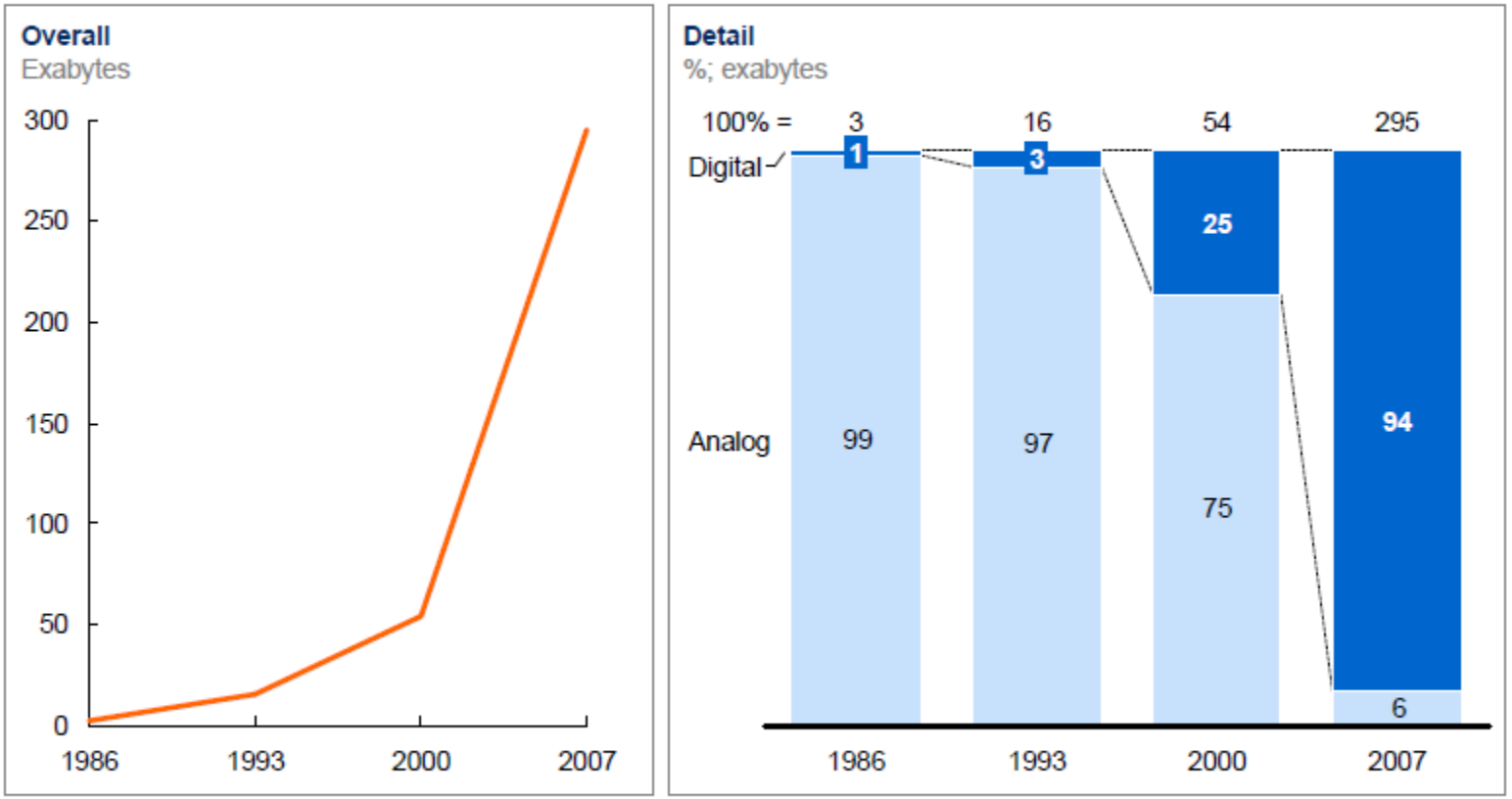
Why Big-Data?

- Key enablers for the appearance and growth of “Big Data” are:
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data

Enabler: Data storage

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



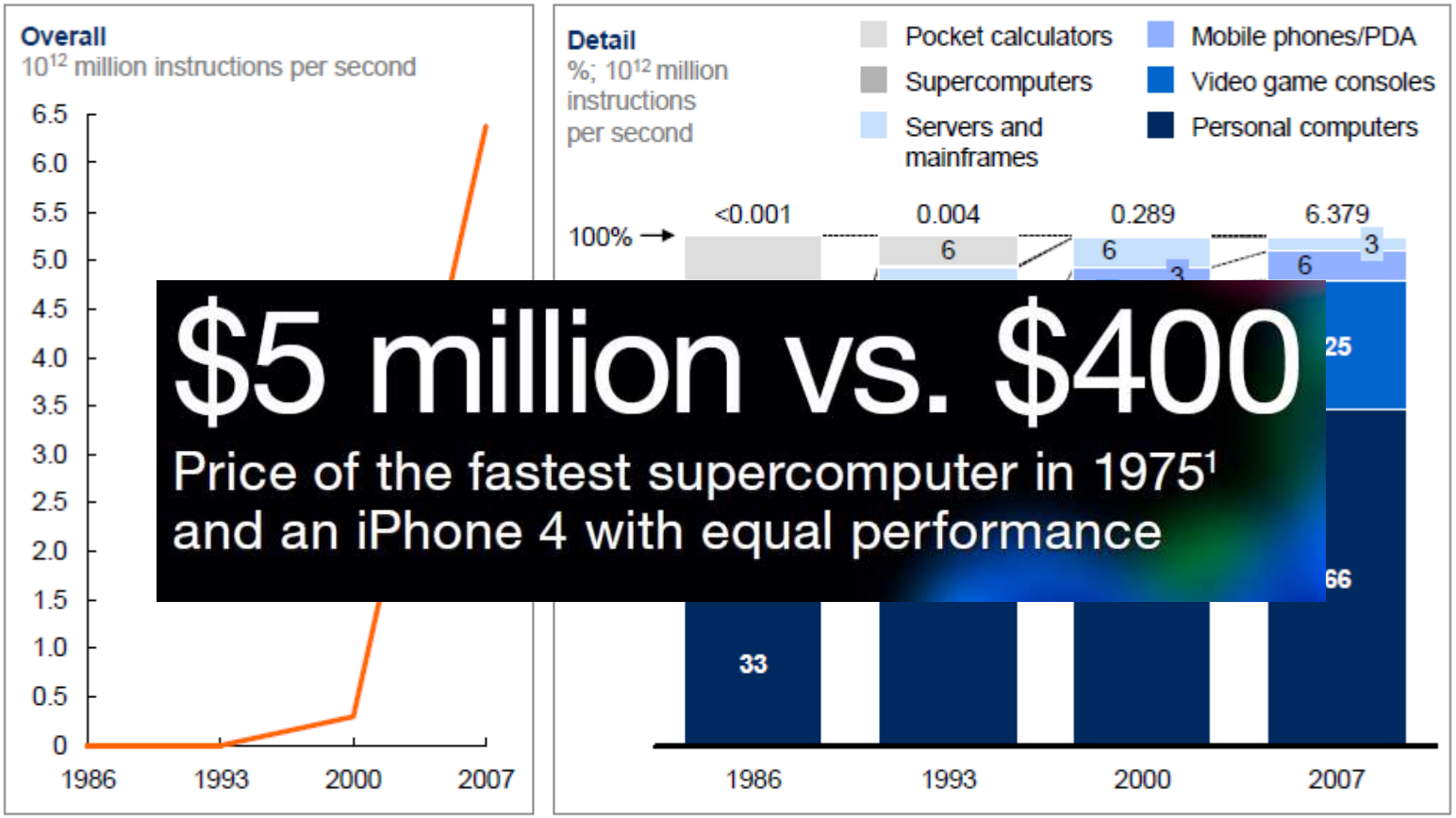
NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Computation capacity

Computation capacity has also risen sharply

Global installed computation to handle information



$5 million vs. $400

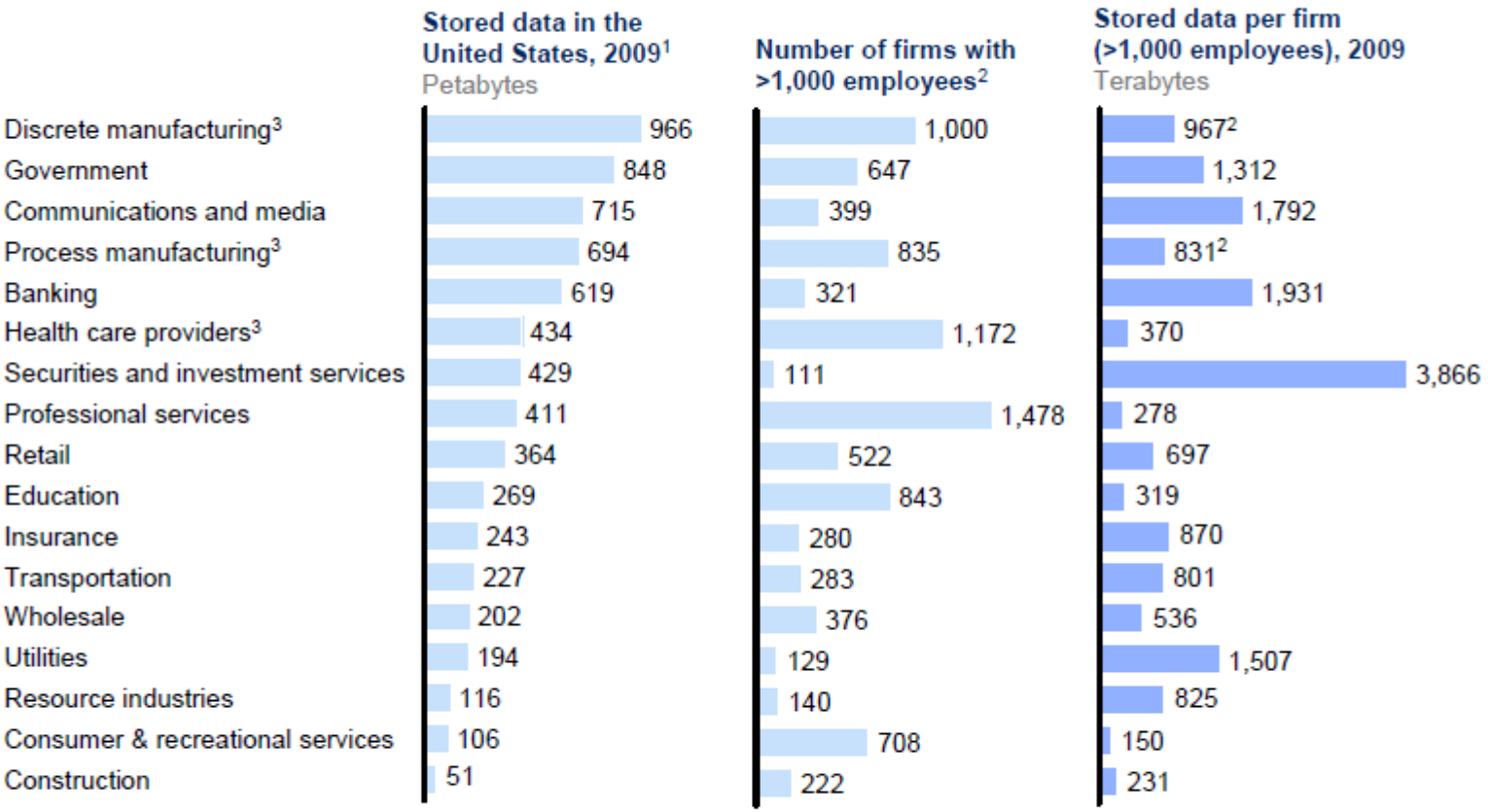
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Data availability

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

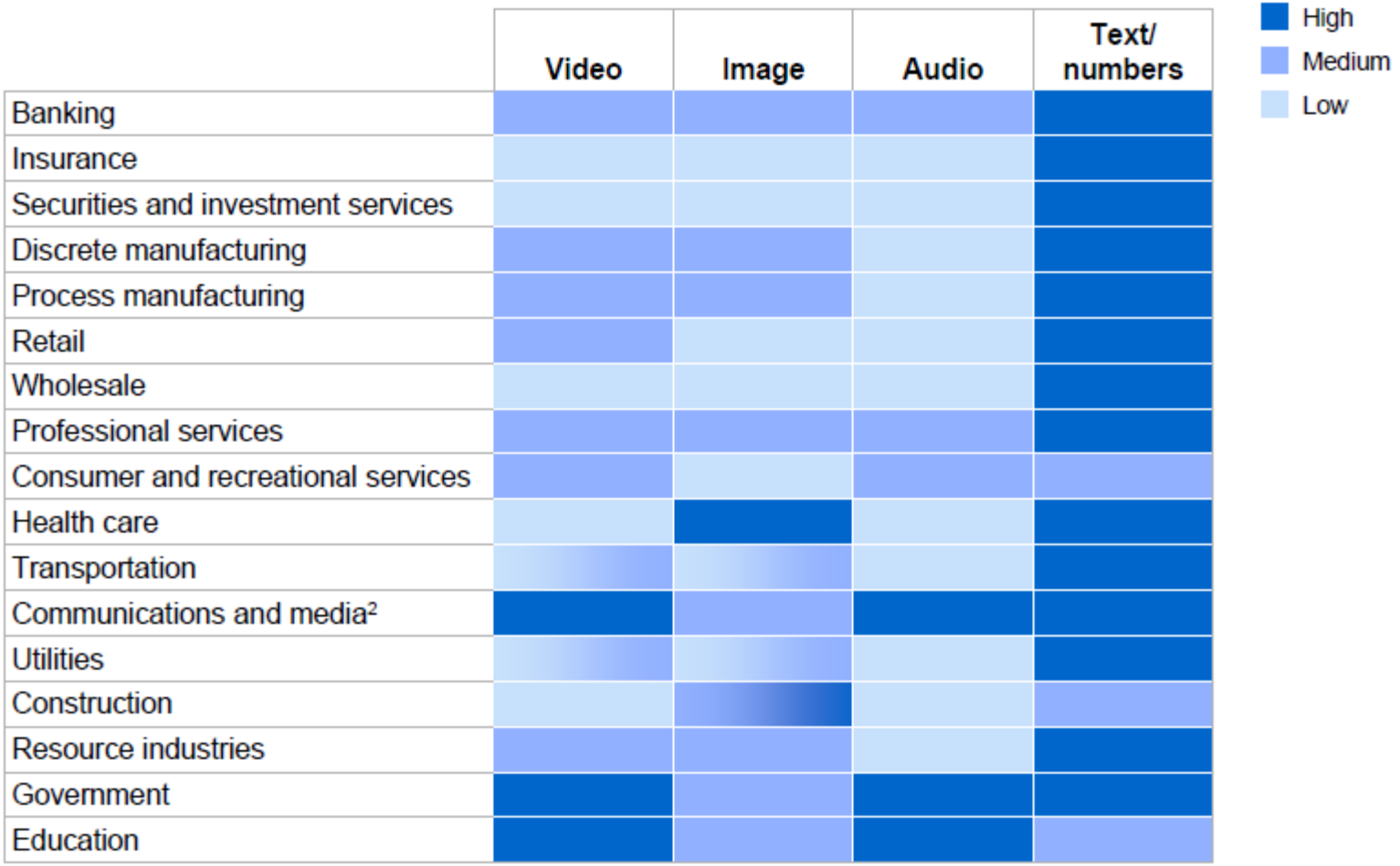


1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

Type of available data

The type of data generated and stored varies by sector[1]



	Video	Image	Audio	Text/ numbers
Banking				
Insurance				
Securities and investment services				
Discrete manufacturing				
Process manufacturing				
Retail				
Wholesale				
Professional services				
Consumer and recreational services				
Health care				
Transportation				
Communications and media[2]				
Utilities				
Construction				
Resource industries				
Government				
Education				

1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

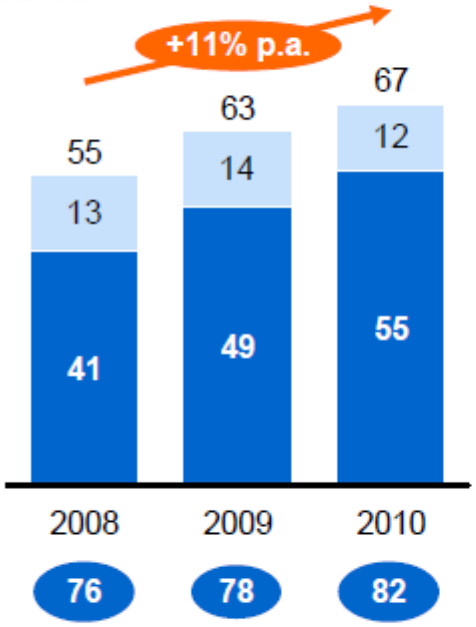
Data available from social networks and mobile devices

The penetration of social networks is increasing online and on smartphones; frequent users are increasing as a share of total users[1]

Frequent user[2]

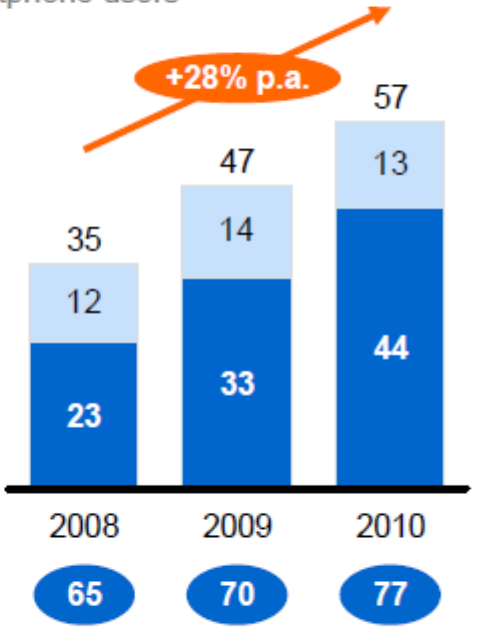
Social networking penetration on the PC is slowing, but frequent users are still increasing

% of respondents



Social networking penetration of smartphones has nearly doubled since 2008

% of smartphone users

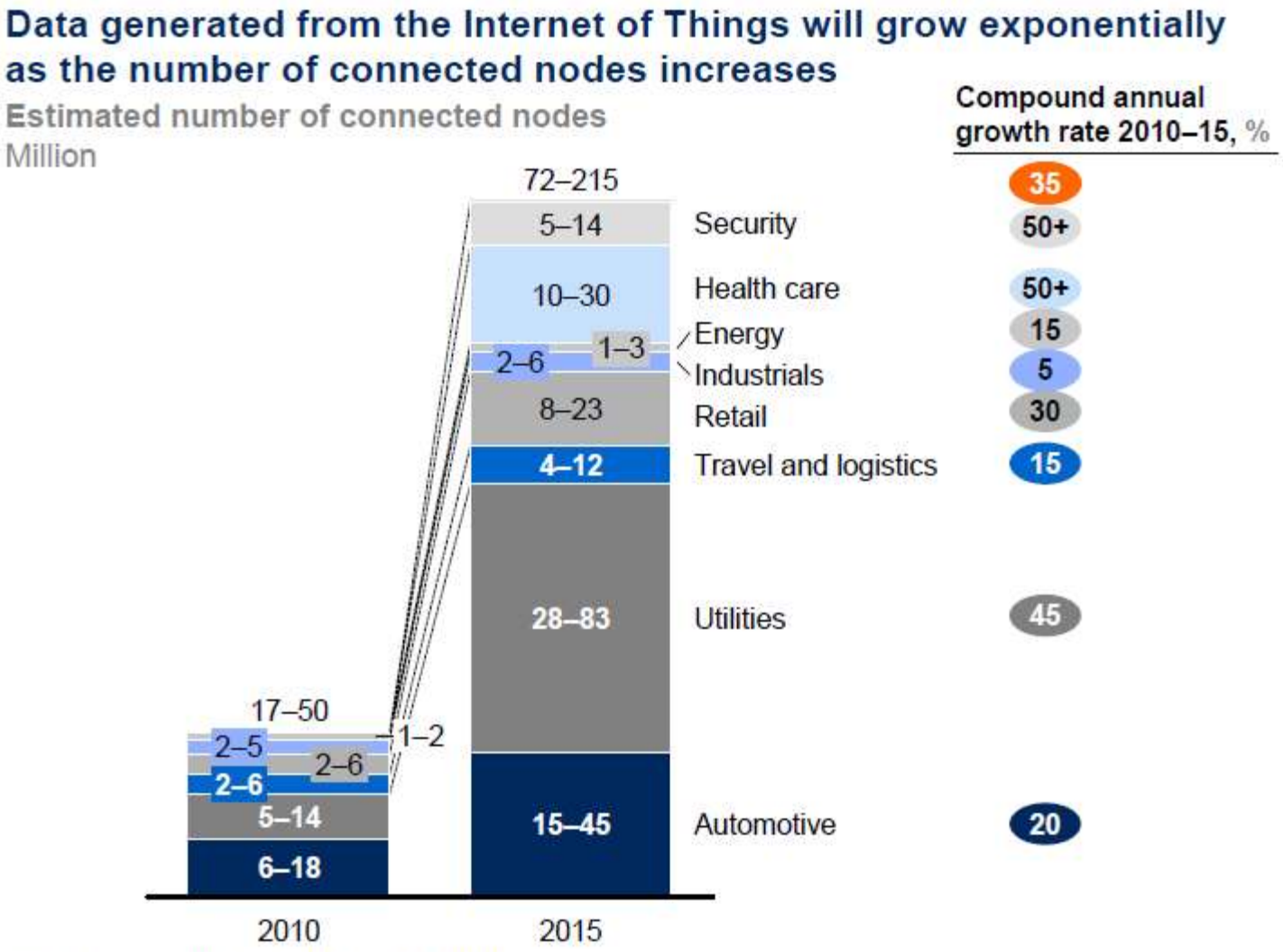


1 Based on penetration of users who browse social network sites. For consistency, we exclude Twitter-specific questions (added to survey in 2009) and location-based mobile social networks (e.g., Foursquare, added to survey in 2010).

2 Frequent users defined as those that use social networking at least once a week.

SOURCE: McKinsey iConsumer Survey

Data available from "Internet of Things"

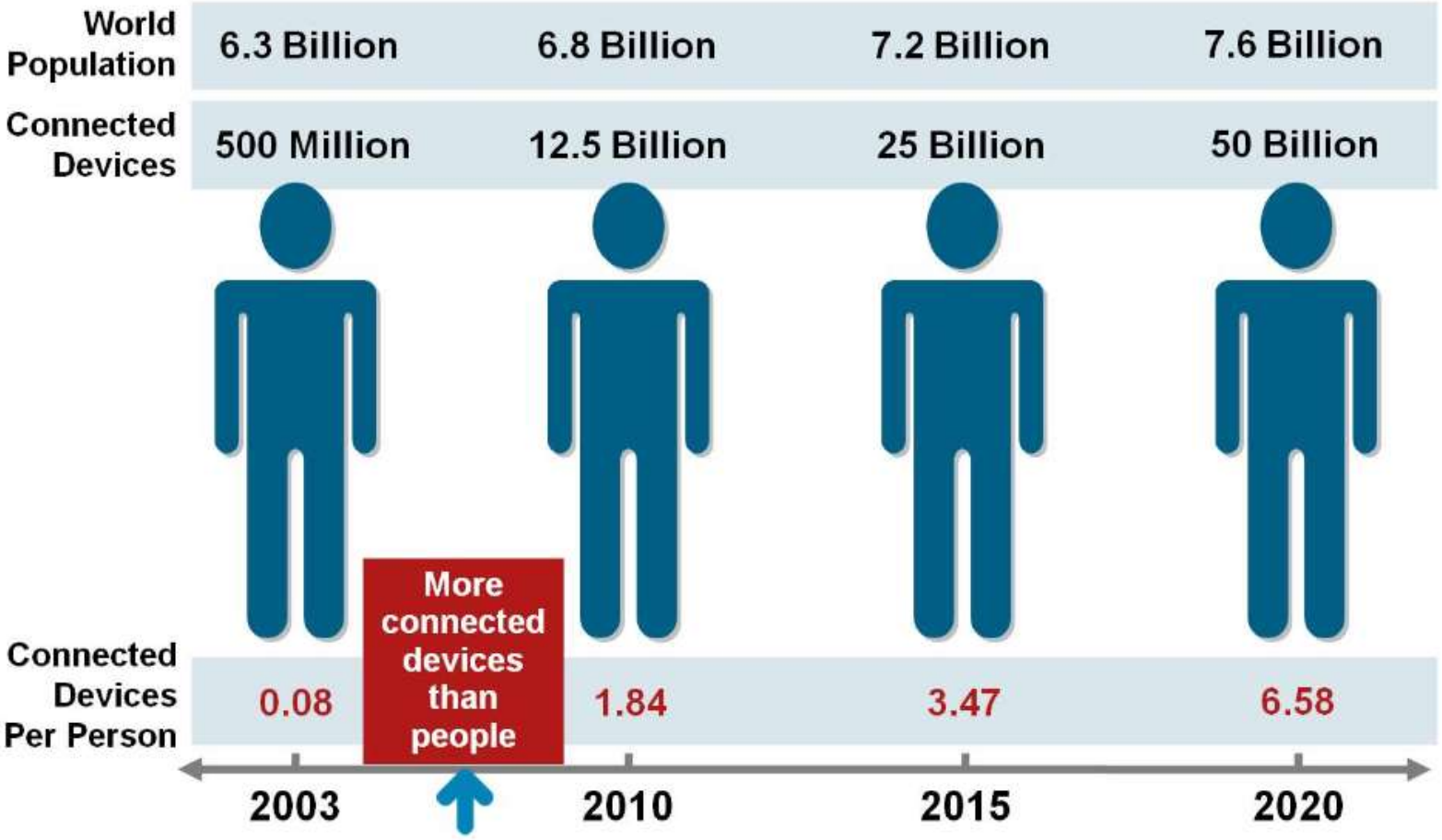


NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

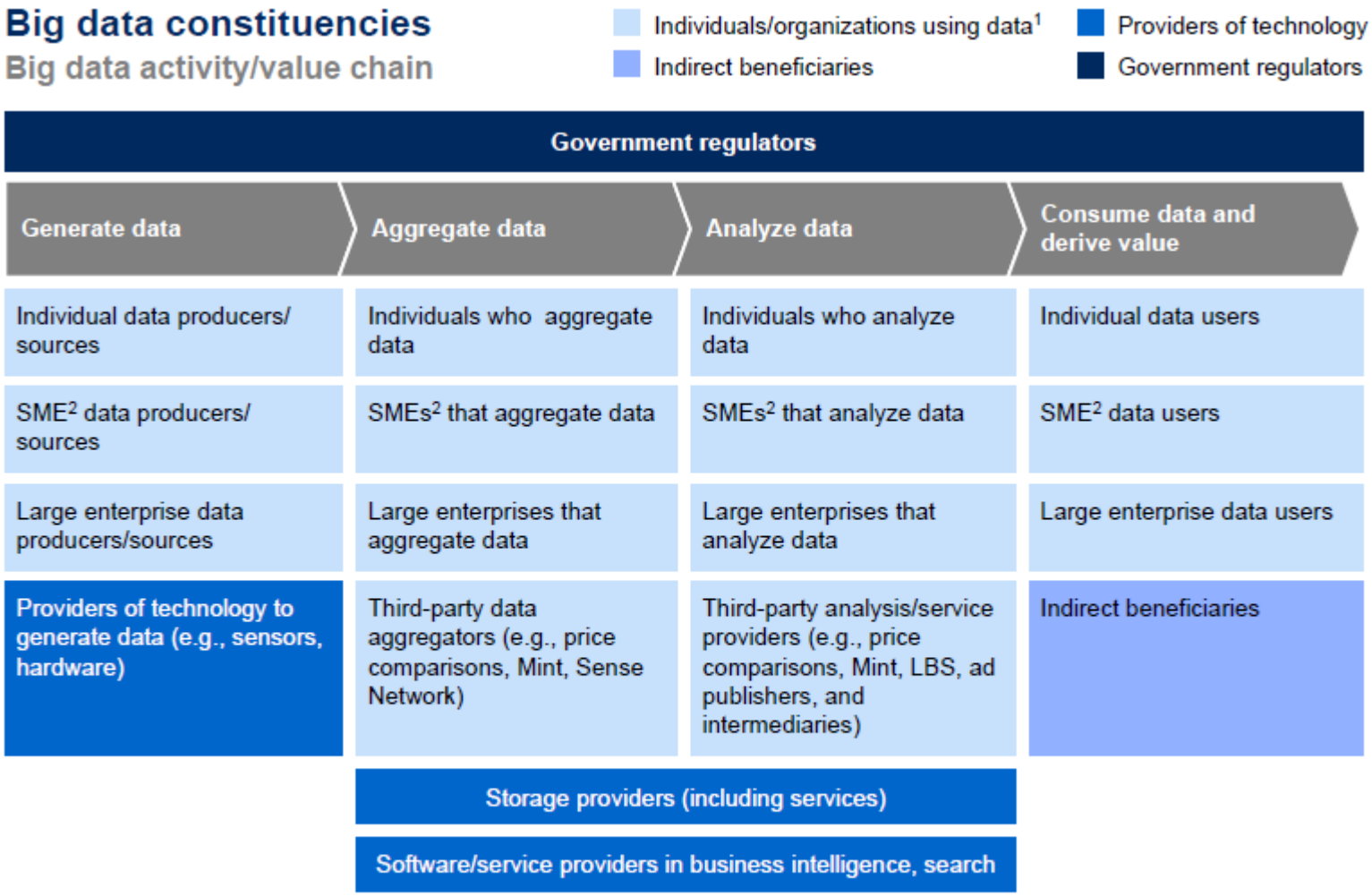
Birth & Growth of "Internet of Things"

Figure 1. The Internet of Things Was "Born" Between 2008 and 2009



Source: Cisco IBSG, April 2011

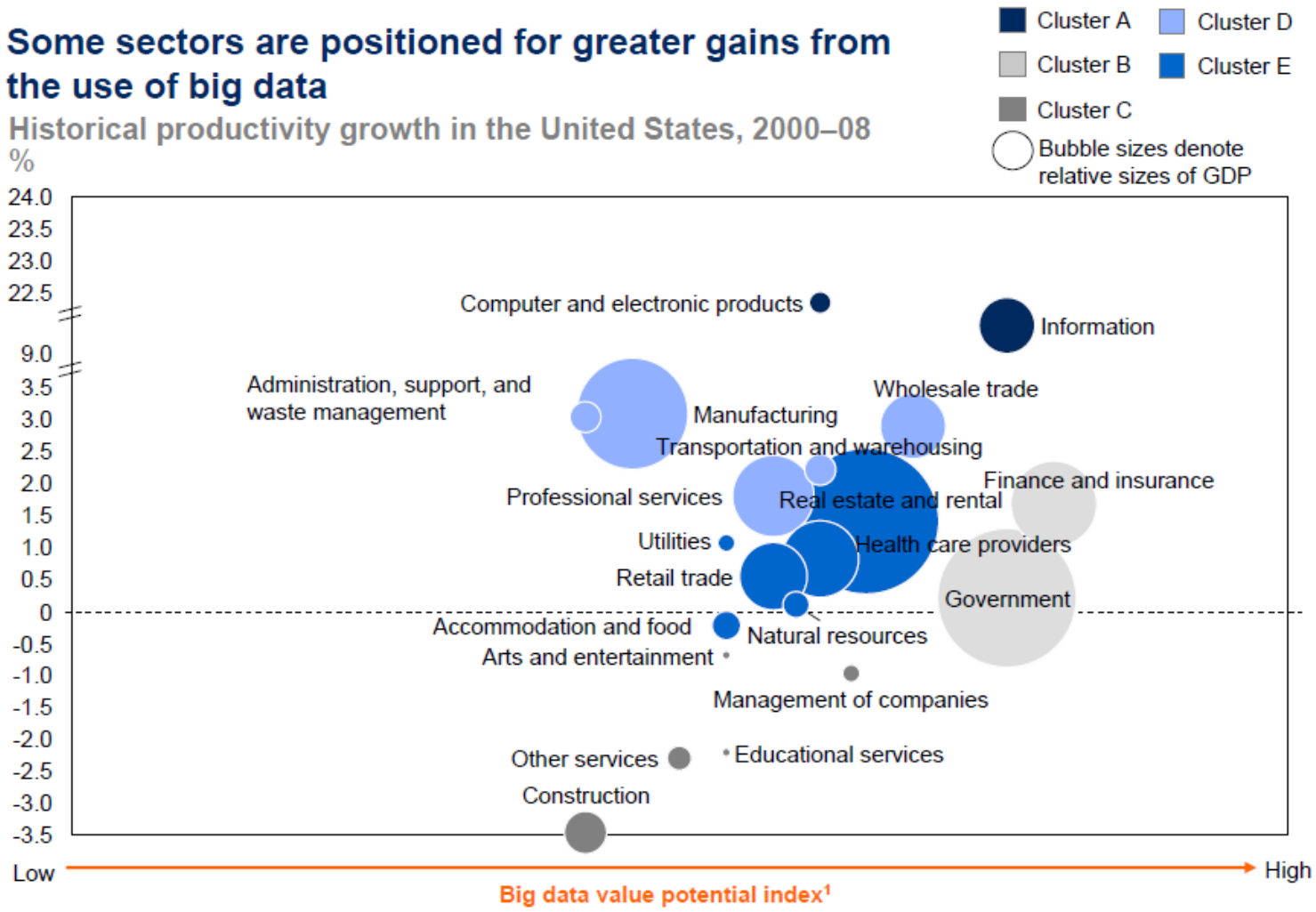
Big-data value chain



1 Individuals/organizations generating, aggregating, analyzing, or consuming data.
2 Small and medium-sized enterprises.

SOURCE: McKinsey Global Institute analysis

Gains from Big-Data per sector

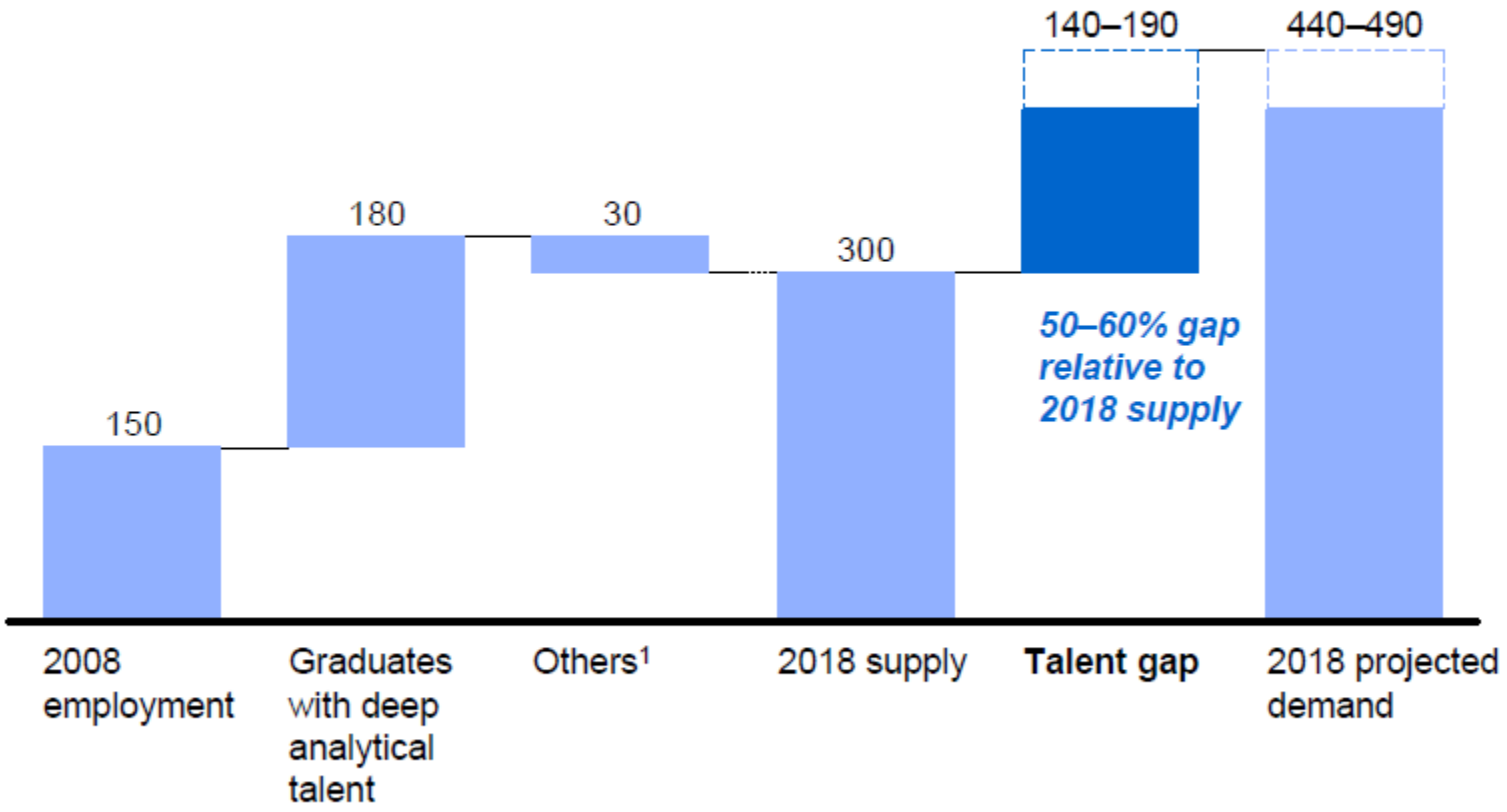


1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Predicted lack of talent for Big-Data related technologies

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Big Data Market

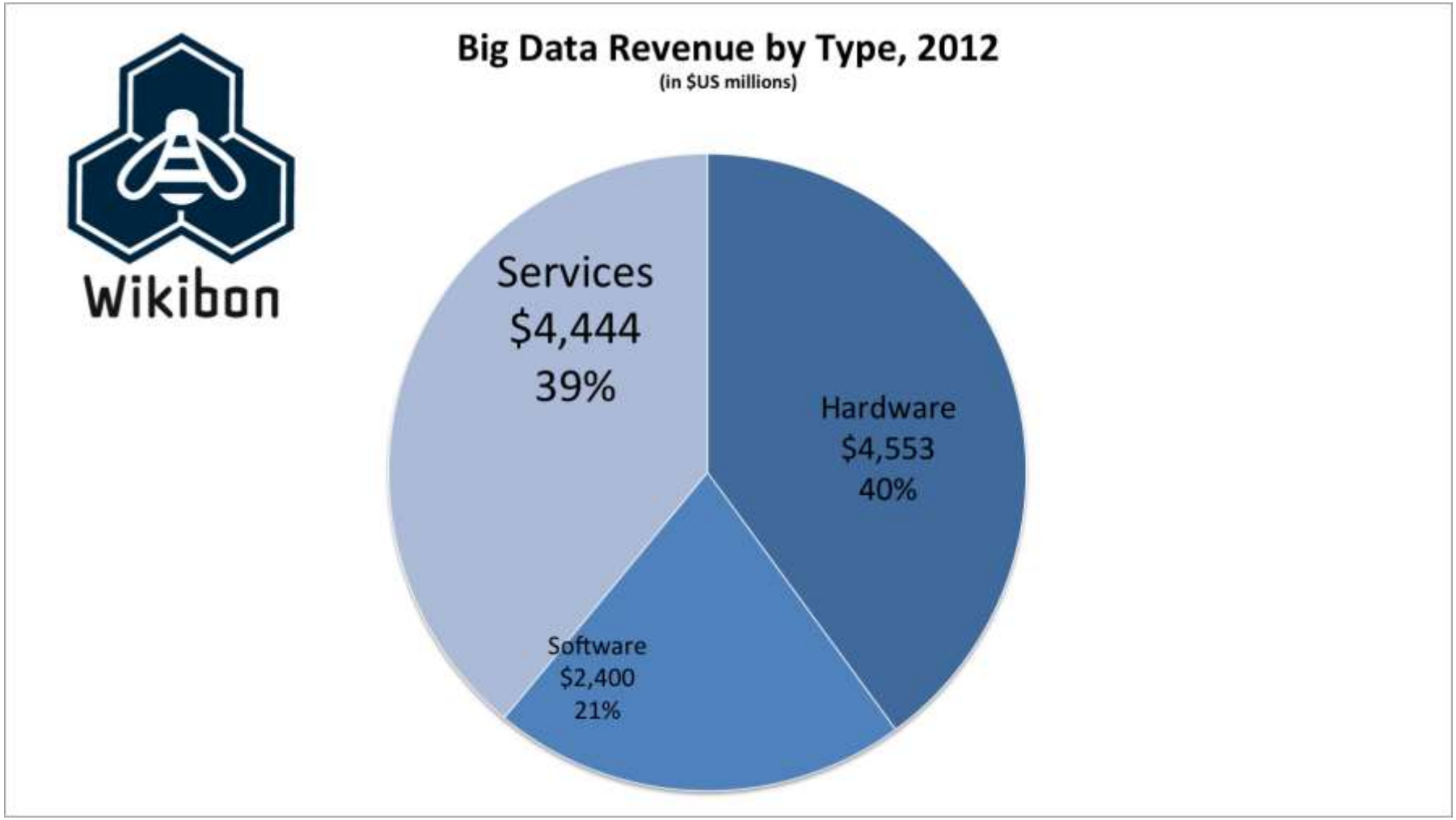
2012 Worldwide Big Data Revenue by Vendor ($US millions)

Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	$1,352	$103,930	1%	22%	33%	44%
HP	$664	$119,895	1%	34%	29%	38%
Teradata	$435	$2,665	16%	31%	28%	41%
Dell	$425	$59,878	1%	83%	0%	17%
Oracle	$415	$39,463	1%	25%	34%	41%
SAP	$368	$21,707	2%	0%	67%	33%
EMC	$336	$23,570	1%	24%	36%	39%
Cisco Systems	$214	$47,983	0%	80%	0%	20%
Microsoft	$196	$$71,474	0%	0%	67%	33%
Accenture	$194	$29,770	1%	0%	0%	100%
Fusion-io	$190	$439	43%	71%	0%	29%
PwC	$189	$31,500	1%	0%	0%	100%
SAS Institute	$187	$2,954	6%	0%	59%	41%

Source: WikiBon report on "**Big Data Vendor Revenue and Market Forecast 2012-2017**", 2013

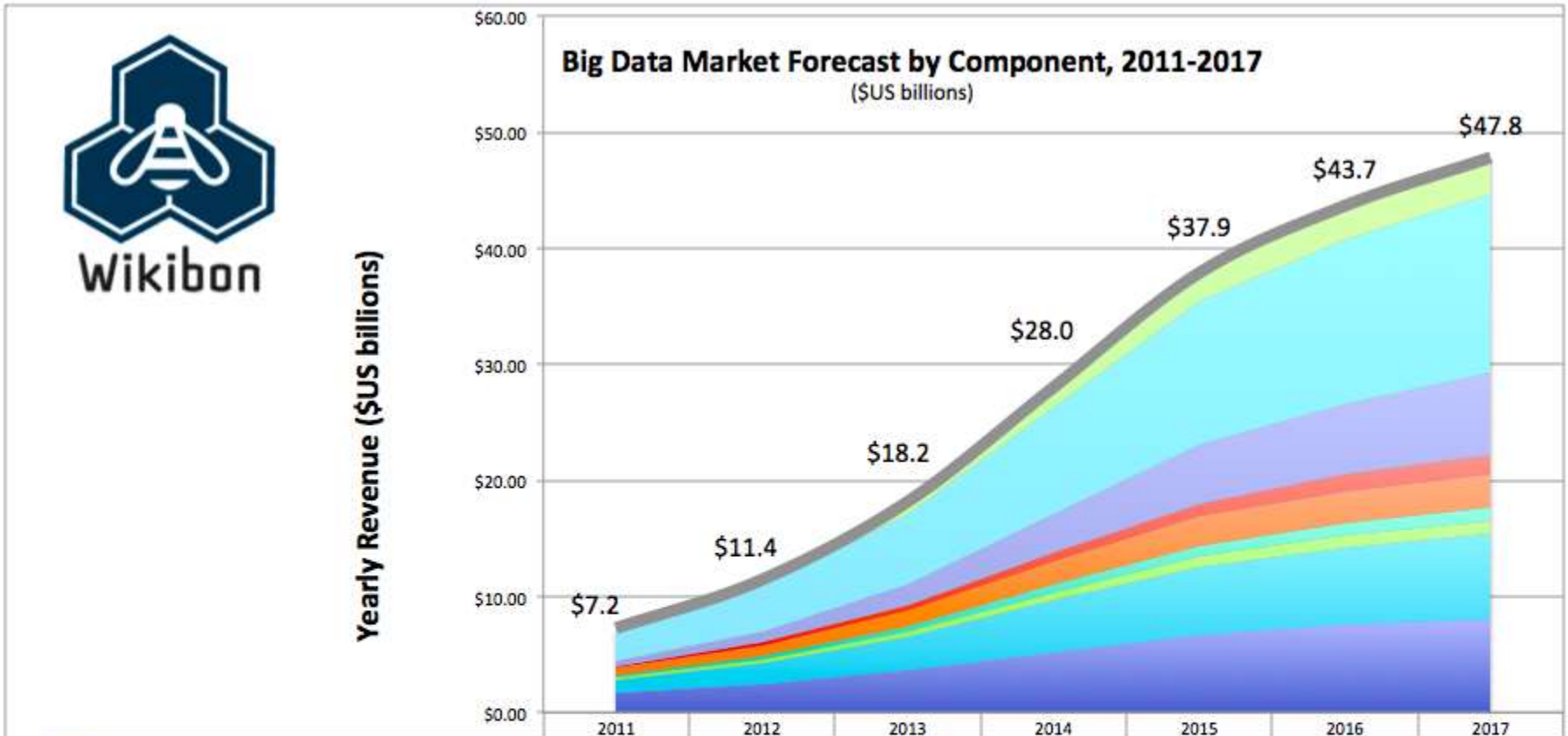
Big Data Revenue by Type, 2012

(http://wikibon.org/w/images/f/f9/Segment_-_BDMSVR2012.png)



Big Data Market Forecast (2011–2017)

(http://wikibon.org/w/images/b/bb/Forecast-BDMSVR2012.png)



	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.35	$0.61	$1.05	$1.74	$2.47	$2.91	$3.24
Big Data Professional Services Revenue	$2.45	$3.87	$6.10	$9.29	$12.37	$14.14	$15.38
Big Data Application (Analytic and Transactional) Software	$0.49	$0.94	$1.80	$3.29	$5.02	$6.15	$7.00
Big Data NoSQL Database Software	$0.10	$0.19	$0.39	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Software	$0.72	$1.02	$1.45	$1.99	$2.47	$2.73	$2.90
Big Data Infrastructure Software	$0.16	$0.26	$0.43	$0.70	$0.96	$1.12	$1.24
Big Data Networking Revenue	$0.18	$0.28	$0.44	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$1.16	$1.83	$2.89	$4.40	$5.86	$6.70	$7.28
Big Data Compute Revenue	$1.64	$2.45	$3.64	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$7.2	$11.4	$18.2	$28.0	$37.9	$43.7	$47.8

Techniques

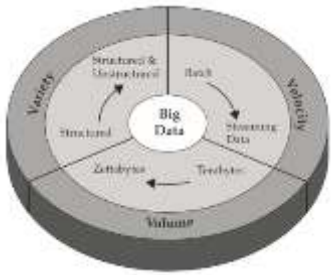
When Big-Data is really a hard problem?

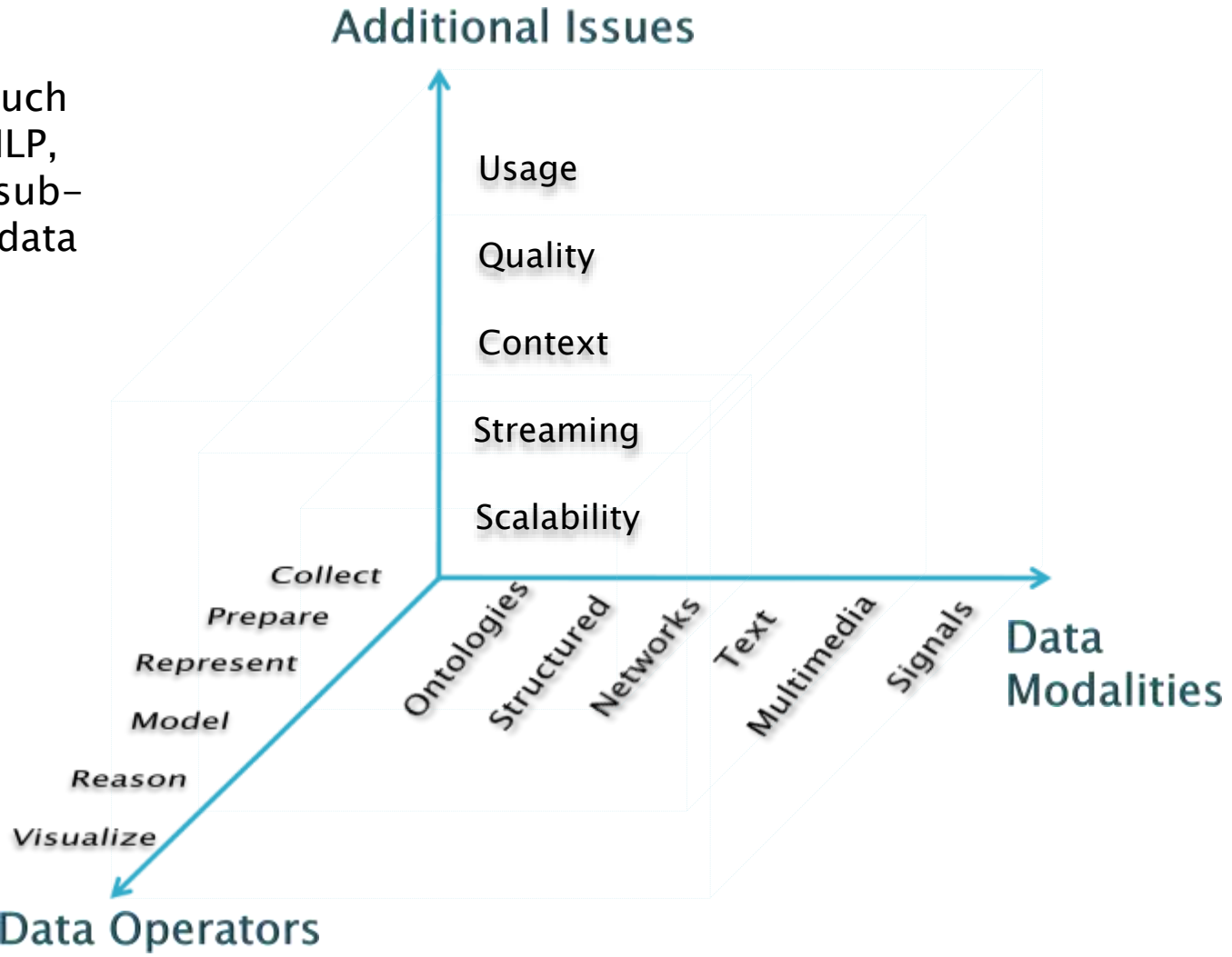
- …when the operations on data are complex:
  - …e.g. simple counting is not a complex problem
  - Modeling and reasoning with data of different kinds can get extremely complex

- Good news about big-data:
  - Often, because of vast amount of data, modeling techniques can get simpler (e.g. smart counting can replace complex model-based analytics)…
  - …as long as we deal with the scale

What matters when dealing with data?

- Research areas (such as IR, KDD, ML, NLP, SemWeb, …) are sub-cubes within the data cube





Meaningfulness of Analytic Answers (1/2)

- A risk with "Big-Data mining" is that an analyst can "discover" patterns that are meaningless
- Statisticians call it Bonferroni's principle:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Meaningfulness of Analytic Answers (2/2)

Example:

- We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day
  - 10^9 people being tracked.
  - 1000 days.
  - Each person stays in a hotel 1% of the time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels).
  - If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?
- Expected number of "suspicious" pairs of people:
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

Example taken from: Rajaraman, Ullman: Mining of Massive Datasets

What are specific operators used in Big-Data applications

- **Smart sampling** of data
  - …reducing the original data while not losing the statistical properties of data
- **Finding similar items**
  - …efficient multidimensional indexing
- **Incremental updating** of the models
  - (vs. building models from scratch)
  - …crucial for streaming data
- **Distributed linear algebra**
  - …dealing with large sparse matrices

Analytical operators on Big-Data

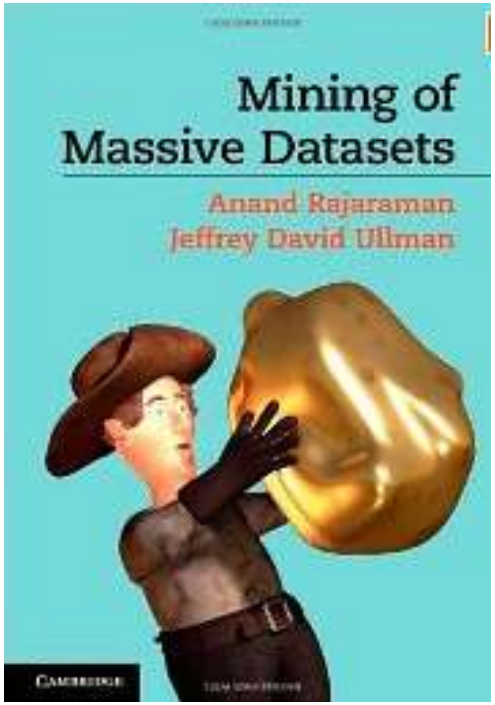
- On the top of the previous ops we perform usual data mining/machine learning/statistics operators:
  - **Supervised** learning (classification, regression, …)
  - **Non-supervised** learning (clustering, different types of decompositions, …)
  - …
- …we are just more careful which algorithms we choose
  - typically linear or sub-linear versions of the algorithms

…guide to Big-Data algorithms

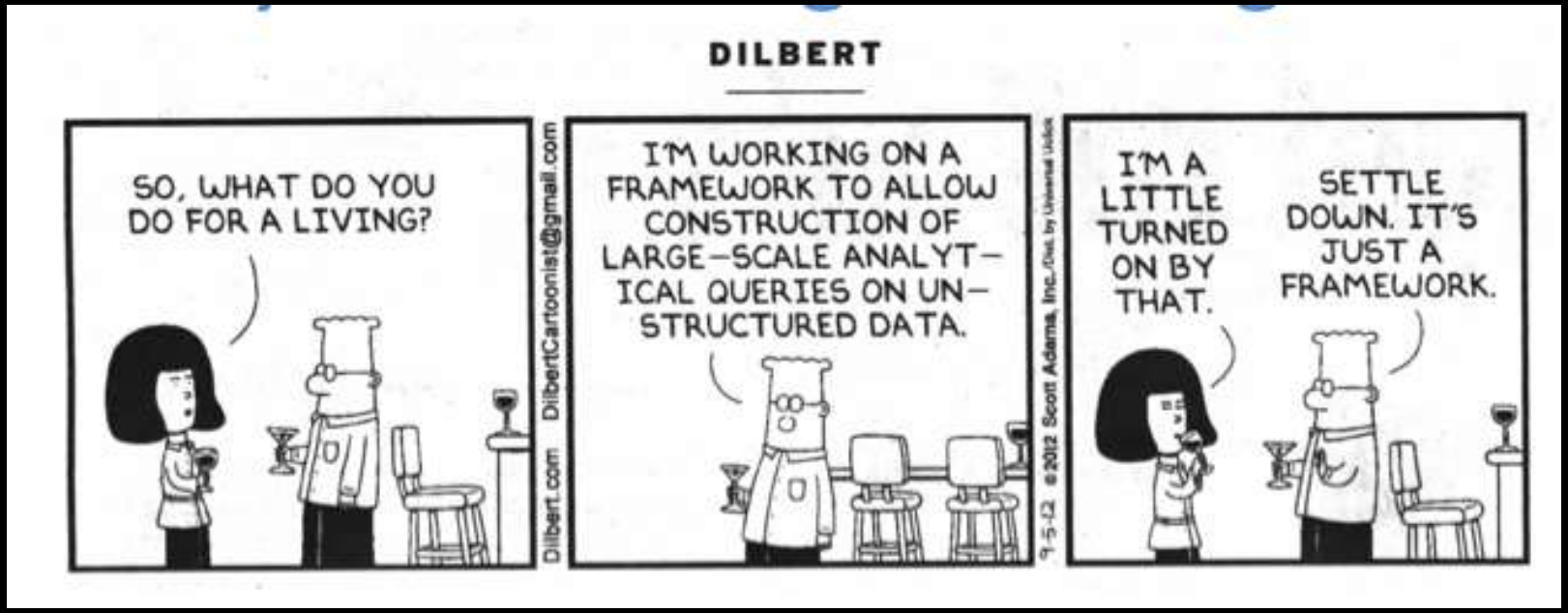
- An excellent overview of the algorithms covering the above issues is the book "**Rajaraman, Leskovec, Ullman: Mining of Massive Datasets**"

- Downloadable from:

http://infolab.stanford.edu/~ullman/mmds.html



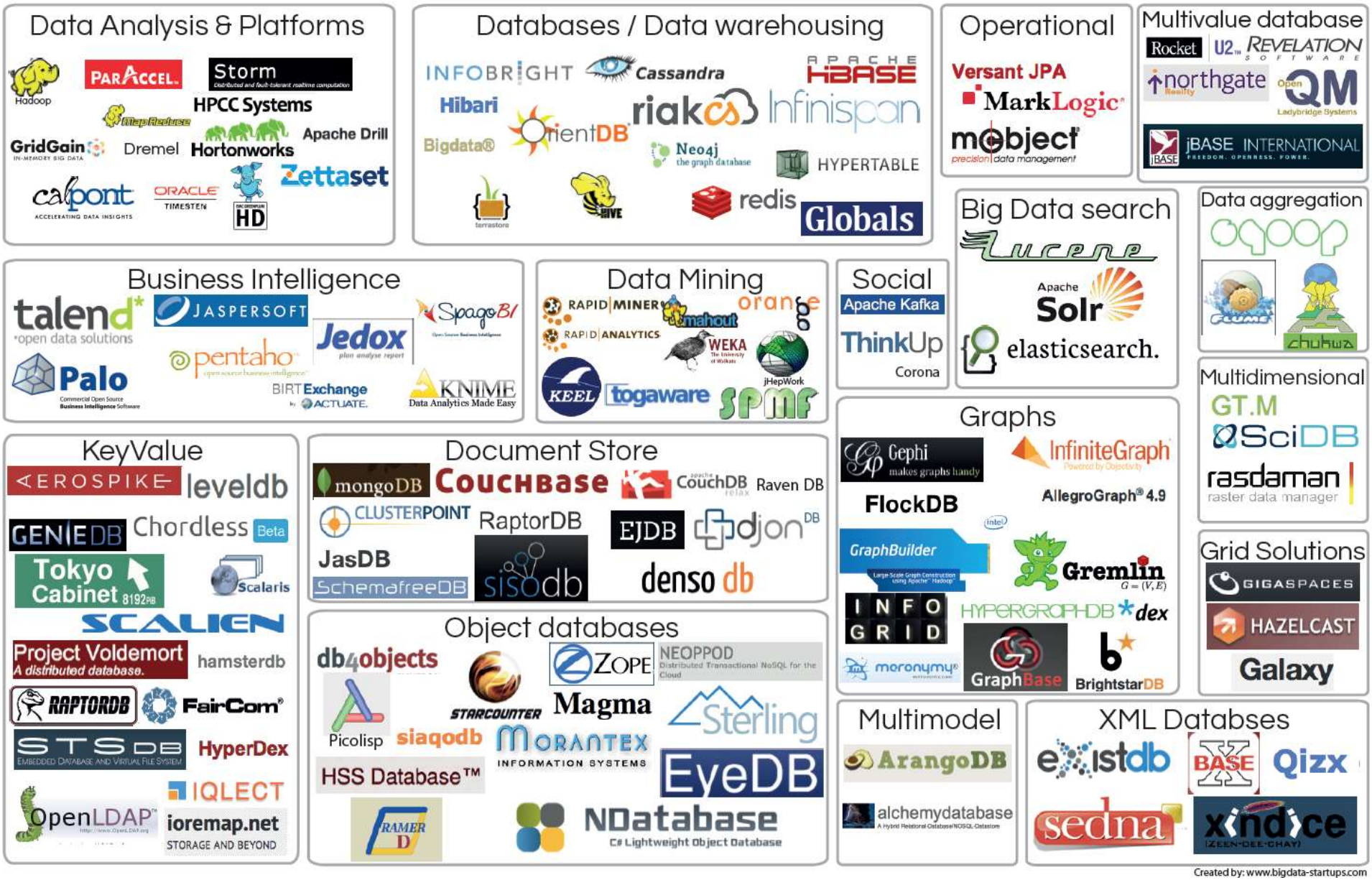
Tools


DILBERT
SO, WHAT DO YOU DO FOR A LIVING?
I'M WORKING ON A FRAMEWORK TO ALLOW CONSTRUCTION OF LARGE-SCALE ANALYT-ICAL QUERIES ON UN-STRUCTURED DATA.
I'M A LITTLE TURNED ON BY THAT.
SETTLE DOWN. IT'S JUST A FRAMEWORK.
Dilbert.com DilbertCartoonist@gmail.com
9-5-12 ©2012 Scott Adams, Inc./Dist. by Universal Uclick

Types of tools typically used in Big-Data scenarios

- Where processing is **hosted**?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is **stored**?
  - Distributed Storage (e.g. Amazon S3)
- What is the **programming model**?
  - Distributed Processing (e.g. MapReduce)
- How data is **stored & indexed**?
  - High-performance schema-free databases (e.g. MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

Plethora of "Big Data" related tools



http://www.bigdata-startups.com/open-source-tools/

Distributed infrastructure

- Computing and storage are typically hosted transparently on cloud infrastructures
  - …providing scale, flexibility and high fail-safety

- Distributed Servers
  - Amazon-EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Distributed Storage
  - Amazon-S3, Hadoop Distributed File System

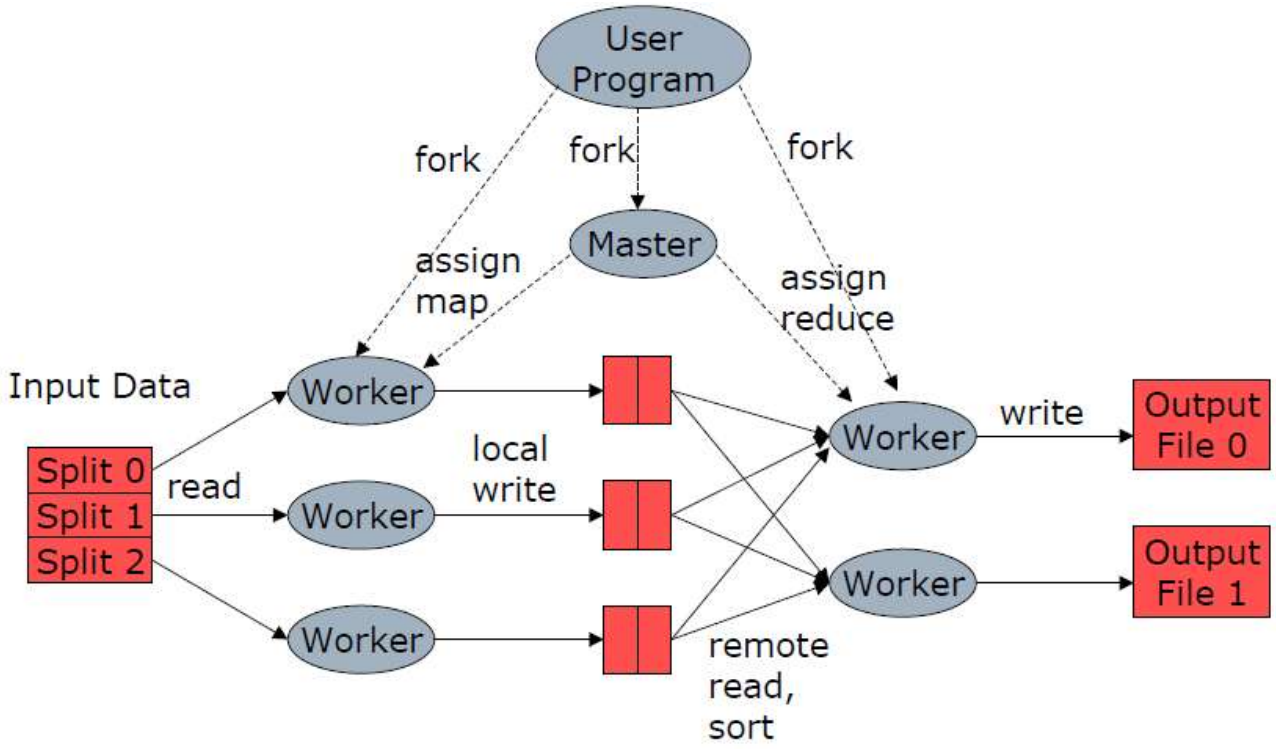
Distributed processing

- Distributed processing of Big-Data requires non-standard programming models
  - …beyond single machines or traditional parallel programming models (like MPI)
  - …the aim is to simplify complex programming tasks

- The most popular programming model is **MapReduce** approach
  - …suitable for commodity hardware to reduce costs

MapReduce

- The key idea of the MapReduce approach:
  - A target problem needs to be parallelizable
  - First, the problem gets split into a set of smaller problems (Map step)
  - Next, smaller problems are solved in a parallel way
  - Finally, a set of solutions to the smaller problems get synthesized into a solution of the original problem (Reduce step)



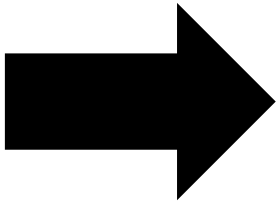
MapReduce example: Counting words in documents

Google Maps charts new territory into businesses

Google selling new tools for businesses to build their own maps

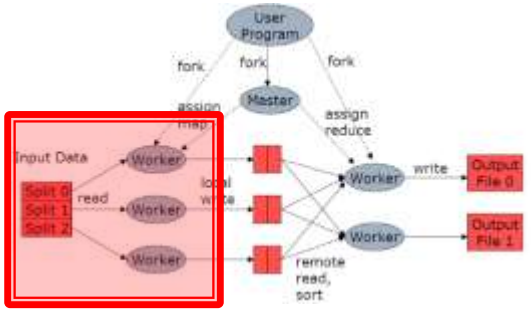
Google promises consumer experience for businesses with Maps Engine Pro

Google is trying to get its Maps service used by more businesses



Google	4
Maps	4
Businesses	4
New	1
Charts	1
Territory	1
Tools	1
…	

MapReduce example: Map task



Google Maps charts new territory into businesses

Google selling new tools for businesses to build their own maps

Map 1

Businesses	2
Charts	1
Maps	2
Territory	1
…	

Google promises consumer experience for businesses with Maps Engine Pro

Google is trying to get its Maps service used by more businesses

Map 2

Businesses	2
Engine	1
Maps	2
Service	1
…	

MapReduce example: Group and aggregate

- Split according to the hash of a key
- In our case: key = word, hash = first character

Task 1

Businesses	2
Charts	1
Maps	2
Territory	1
…	

Task 2

Businesses	2
Engine	1
Maps	2
Service	1
…	

Reduce 1

Businesses	2
Charts	1
…	

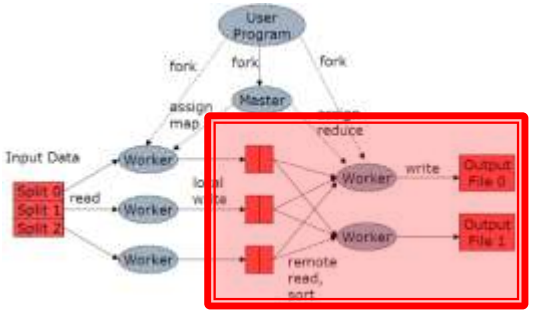
Businesses	2
Engine	1
…	

Reduce 2

Maps	2
Territory	1
…	

Maps	2
Service	1
…	

MapReduce example: Reduce task



Businesses	2
Charts	1
…	

Businesses	2
Engine	1
…	

Reduce 1

Businesses	4
Charts	1
Engine	1
…	

Maps	2
Territory	1
…	

Maps	2
Service	1
…	

Reduce 2

Maps	4
Territory	1
Service	1
…	

MapReduce example: Combine

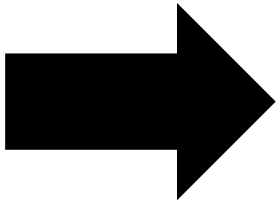
- We concatenate the outputs into final result

Reduce 1

Businesses	4
Charts	1
Engine	1
…	

Reduce 2

Maps	4
Territory	1
Service	1
…	



Businesses	4
Charts	1
Engine	1
…	
Maps	4
Territory	1
Service	1
…	

MapReduce Tools

- Apache Hadoop [http://hadoop.apache.org/]
  - Open-source MapReduce implementation
- Tools using Hadoop:
  - **Hive**: data warehouse infrastructure that provides data summarization and ad hoc querying (HiveQL)
  - **Pig**: high-level data-flow language and execution framework for parallel computation (Pig Latin)
  - **Mahout**: Scalable machine learning and data mining library
  - **Flume**: Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data
  - Many more: Cascading, Cascalog, mrjob, MapR, Azkaban, Oozie, …

History repeats

Hype on Databases from nineties == Hadoop from now

http://iggyfernandez.wordpress.com/2013/01/21/dilbert-likes-hadoop-clusters/

NoSQL Databases

- "[…] need to solve a problem that relational databases are a bad fit for", Eric Evans

- Motives:
  - **Avoidance of Unneeded Complexity** – many use-case require only subset of functionality from RDBMSs (e.g ACID properties)
  - **High Throughput** – some NoSQL databases offer significantly higher throughput then RDBMSs
  - **Horizontal Scalability, Running on commodity hardware**
  - **Avoidance of Expensive Object-Relational Mapping** – most NoSQL store simple data structures
  - **Compromising Reliability for Better Performance**

Based on "NoSQL Databases", Christof Strauch http://www.christof-strauch.de/nosqldbs.pdf

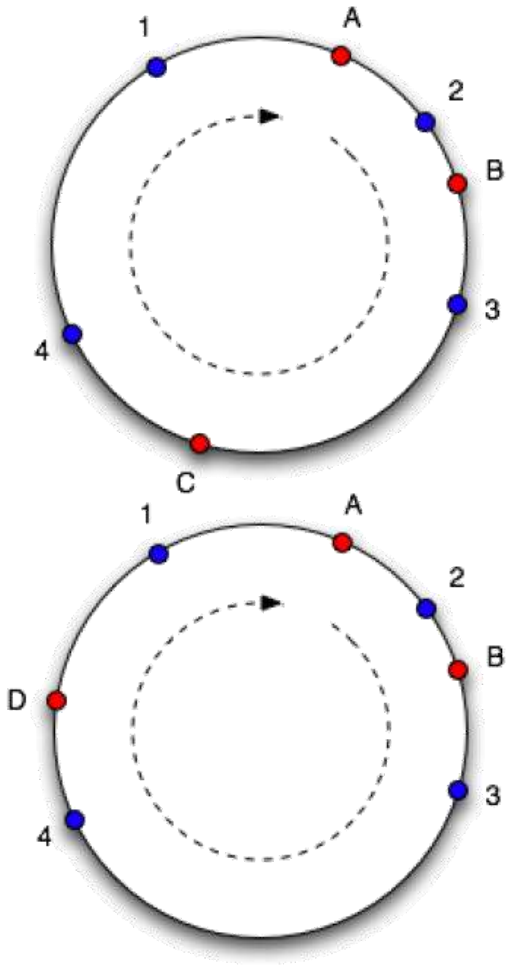
Basic Concepts – Consistency

- BASE approach
  - Availability, graceful degradation, performance
  - Stands for "**B**asically **a**vailable, **s**oft state, **e**ventual consistency"

- Continuum of tradeoffs:
  - **Strict** – All reads must return data from latest completed writes
  - **Eventual** – System eventually return the last written value
  - **Read Your Own Writes** – see your updates immediately
  - **Session** – RYOW only within same session
  - **Monotonic** – only more recent data in future requests

Basic Concepts – Partitioning

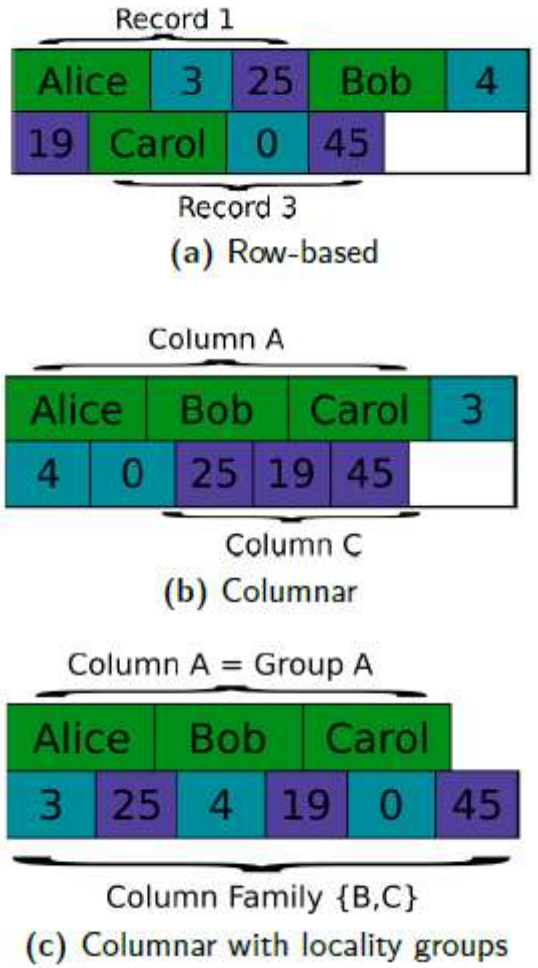
- Consistent hashing
  - Use same function for hashing objects and nodes
  - Assign objects to nearest nodes on the circle
  - Reassign object when nodes added or removed
  - Replicate nodes to *r* nearest nodes



White, Tom: *Consistent Hashing*. November 2007. – Blog post of 2007-11-27. http://weblogs.java.net/blog/tomwhite/archive/2007/11/consistent_hash.html

Other Basic Concepts

- Storage Layout
  - Row-based
  - Columnar
  - Columnar with Locality Groups
- Query Models
  - Lookup in key-value stores
- Distributed Data Processing via MapReduce



Lipcon, Todd: *Design Patterns for Distributed Non-Relational Databases*. June 2009. – Presentation of 2009-06-11. http://www.slideshare.net/guestdfd1ec/design-patterns-for-distributed-nonrelationaldatabases

Key-Value Stores

- Map or dictionary allowing to add and retrieve values per keys
- Favor scalability over consistency
  - Run on clusters of commodity hardware
  - Component failure is “standard mode of operation”

- Examples:
  - Amazon Dynamo
  - Project Voldemort (developed by LinkedIn)
  - Redis
  - Memcached (not persistent)

Document Databases

- Combine several *key-value pairs* into *documents*
- Documents represented as JSON

```
" Title " : " CouchDB ",
" Last editor " : "172.5.123.91" ,
" Last modified ": "9/23/2010" ,
" Categories ": [" Database ", " NoSQL ", " Document Database "],
" Body ": " CouchDB is a ..." ,
" Reviewed ": false
```

- Examples:
  - Apache CouchDB
  - MongoDB

Column-Oriented

- Using columnar storage layout with locality groups (column families)

- Examples:
  - Google Bigtable
  - Hypertable, HBase
    - open source implementation of Google Bigtable
  - Cassandra
    - combination of Google Bigtable and Amazon Dynamo
    - Designed for high write throughput

Open Source Big Data Tools

Infrastructure:

- Kafka [http://kafka.apache.org/]
  - A high-throughput distributed messaging system
- Hadoop [http://hadoop.apache.org/]
  - Open-source map-reduce implementation
- Storm [http://storm-project.net/]
  - Real-time distributed computation system
- Cassandra [http://cassandra.apache.org/]
  - Hybrid between Key-Value and Row-Oriented DB
  - Distributed, decentralized, no single point of failure
  - Optimized for fast writes

Open Source Big Data Tools Machine Learning

- Mahout
  - Machine learning library working on top of Hadoop
  - http://mahout.apache.org/
- MOA
  - Mining data streams with concept drift
  - Integrated with Weka
  - http://moa.cms.waikato.ac.nz/

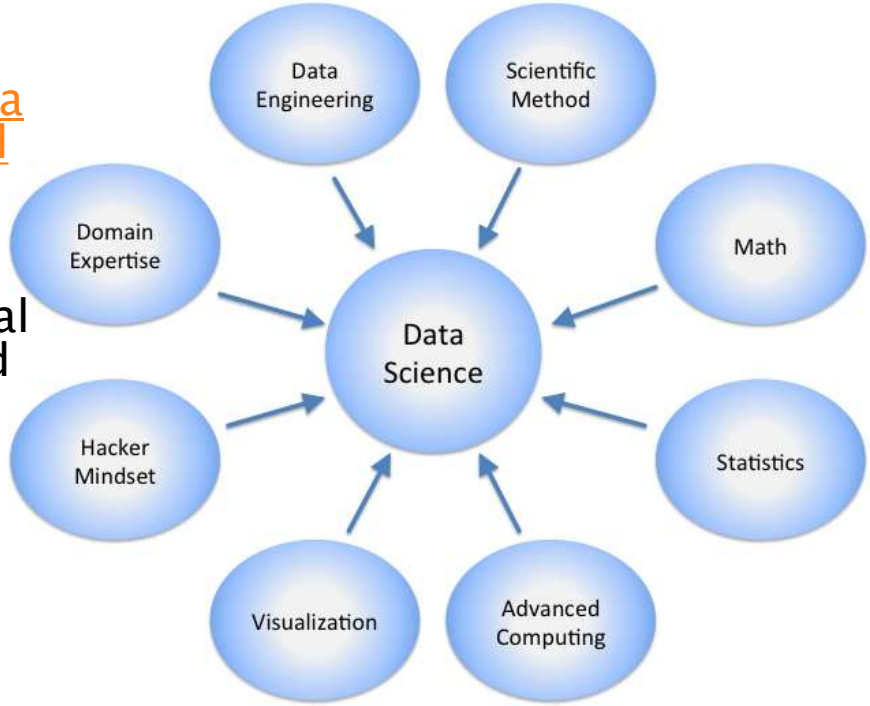
Mahout currently has:

- Collaborative Filtering
- User and Item based recommenders
- K-Means, Fuzzy K-Means clustering
- Mean Shift clustering
- Dirichlet process clustering
- Latent Dirichlet Allocation
- Singular value decomposition
- Parallel Frequent Pattern mining
- Complementary Naive Bayes classifier
- Random forest decision tree based classifier

Data Science

Defining Data Science

- Interdisciplinary field using techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.



- Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common.
- Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners.

http://en.wikipedia.org/wiki/Data_science

Statistics vs. Data Science

	Statistician	Data Scientist
Image	Baseball (Cricket)	HBR Sexiest Job of 21st Century
Mode	Reactive	Consultative
Works	Solo	In a team
Inputs	Data File, Hypothesis	A Business Problem
Data	Pre-prepared, clean	Distributed, messy, unstructured
Data Size	Kilobytes	Gigabytes
Tools	SAS, Mainframe	R, Python, awk, Hadoop, Linux, …
Nouns	Tables	Data Visualizations
Focus	Inference (why)	Prediction (what)
Output	Report	Data App / Data Product
Latency	Weeks	Seconds
Stars	G.E.P Box Trevor Hastie	Hilary Mason Nate Silver

http://blog.revolutionanalytics.com/data-science/

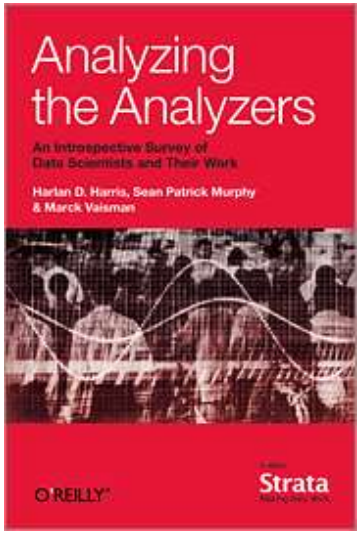
Business Intelligence vs. BI

	Business Intelligence	Data Science
Perspective	Looking backwards	Looking forwards
Actions	Slice and Dice	Interact
Expertise	Business User	Data Scientist
Data	Warehoused, Siloed	Distributed, real-time
Scope	Unlimited	Specific business question
Questions	What happened?	What will happen? What if?
Output	Table	Answer
Applicability	Historic, possible confounding factors	Future, correcting for influences
Tools	SAP, Cognos, Microstrategy, SAS	Revolution R Enterprise QlikView, Tableau, Jaspersoft
Hot or not?	*So* 1997	Transformational

http://blog.revolutionanalytics.com/data-science/

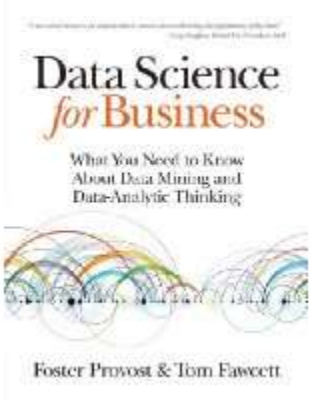
Relevant reading

Analyzing the Analyzers
An Introspective Survey of Data Scientists and Their Work
By Harlan Harris, Sean Murphy, Marck Vaisman
Publisher: O'Reilly Media
Released: June 2013



An Introduction to Data
Jeffrey Stanton, Syracuse University School of Information Studies
Downloadable from http://jsresearch.net/wiki/projects/teachdatascience
Released: February 2013

Data Science for Business: What you need to know about data mining and data-analytic thinking
by Foster Provost and Tom Fawcett
Released: Aug 16, 2013



Applications

- Recommendation
  Social Network Analytics
  Media Monitoring

Application: Recommendation

Data

- User visit logs
  - Track each visit using embedded JavaScript
- Content
  - The content and metadata of visited pages
- Demographics
  - Metadata about (registered) users

Visit log Example

User ID cookie: 1234567890

IP: 95.87.154.251 (Ljubljana, Slovenia)

Requested URL: http://www.bloomberg.com/news/2012-07-19/americans-hold-dimmest-view-on-economic-outlook-since-january.html

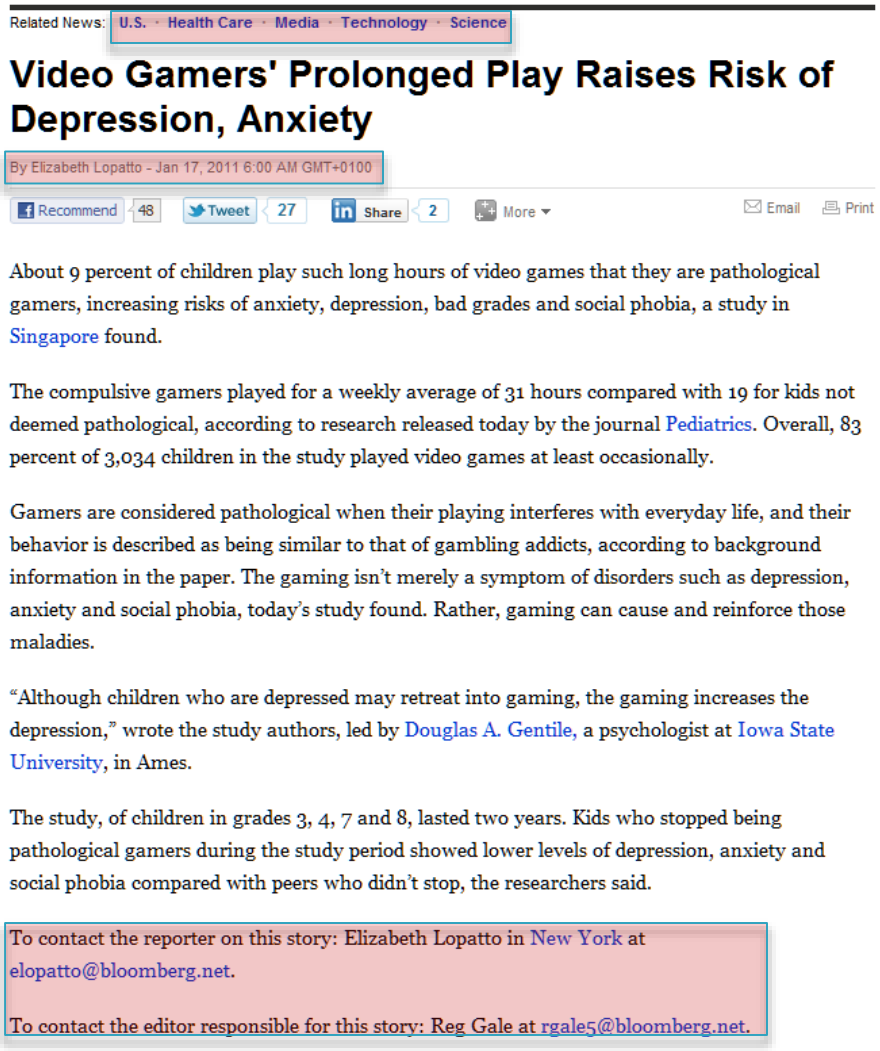
Referring URL: http://www.bloomberg.com/

Date and time: 2009-08-25 08:12:34

Device: Chrome, Windows, PC

Content example (1)

- **News-source:**
  - www.bloomberg.com
- **Article URL:**
  - http://www.bloomberg.com/news/2011-01-17/video-gamers-prolonged-play-raises-risk-of-depression-anxiety-phobias.html
- **Author:**
  - Elizabeth Lopatto
- **Produced at:**
  - New York
- **Editor:**
  - Reg Gale
- **Publish Date:**
  - Jan 17, 2011 6:00 AM
- **Topics:**
  - U.S., Health Care, Media, Technology, Science



Related News: U.S. · Health Care · Media · Technology · Science

Video Gamers' Prolonged Play Raises Risk of Depression, Anxiety

By Elizabeth Lopatto - Jan 17, 2011 6:00 AM GMT+0100

Recommend 48 Tweet 27 Share 2 More Email Print

About 9 percent of children play such long hours of video games that they are pathological gamers, increasing risks of anxiety, depression, bad grades and social phobia, a study in Singapore found.

The compulsive gamers played for a weekly average of 31 hours compared with 19 for kids not deemed pathological, according to research released today by the journal Pediatrics. Overall, 83 percent of 3,034 children in the study played video games at least occasionally.

Gamers are considered pathological when their playing interferes with everyday life, and their behavior is described as being similar to that of gambling addicts, according to background information in the paper. The gaming isn't merely a symptom of disorders such as depression, anxiety and social phobia, today's study found. Rather, gaming can cause and reinforce those maladies.

"Although children who are depressed may retreat into gaming, the gaming increases the depression," wrote the study authors, led by Douglas A. Gentile, a psychologist at Iowa State University, in Ames.

The study, of children in grades 3, 4, 7 and 8, lasted two years. Kids who stopped being pathological gamers during the study period showed lower levels of depression, anxiety and social phobia compared with peers who didn't stop, the researchers said.

To contact the reporter on this story: Elizabeth Lopatto in New York at elopatto@bloomberg.net.

To contact the editor responsible for this story: Reg Gale at rgale5@bloomberg.net.

Content Example (2)

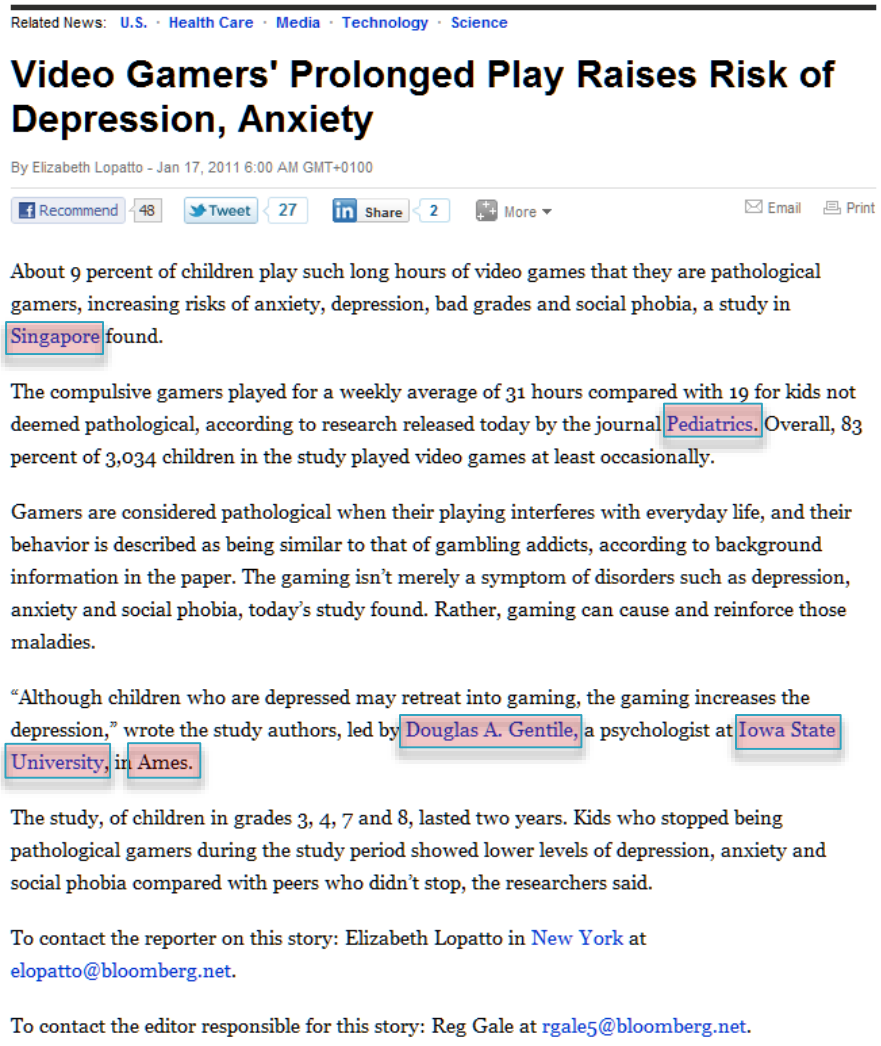
Topics (e.g. DMoz):
- Health/Mental Health/…/Depression
- Health/Mental Health/Disorders/Mood
- Games/Game Studies

Keywords (e.g. DMoz):
- Health, Mental Health, Disorders, Mood, Games, Video Games, Depression, Recreation, Browser Based, Game Studies, Anxiety, Women, Society, Recreation and Sports

Locations:
- Singapore (sws.geonames.org/1880252/)
- Ames (sws.geonames.org/3037869/)

People:
- Duglas A. Gentile

Organizations:
- Iowa State University (dbpediapa.org/resource/Iowa_State_University)
- Pediatrics (journal)



Related News: U.S. · Health Care · Media · Technology · Science

Video Gamers' Prolonged Play Raises Risk of Depression, Anxiety

By Elizabeth Lopatto - Jan 17, 2011 6:00 AM GMT+0100

Recommend 48 Tweet 27 Share 2 More Email Print

About 9 percent of children play such long hours of video games that they are pathological gamers, increasing risks of anxiety, depression, bad grades and social phobia, a study in Singapore found.

The compulsive gamers played for a weekly average of 31 hours compared with 19 for kids not deemed pathological, according to research released today by the journal Pediatrics. Overall, 83 percent of 3,034 children in the study played video games at least occasionally.

Gamers are considered pathological when their playing interferes with everyday life, and their behavior is described as being similar to that of gambling addicts, according to background information in the paper. The gaming isn't merely a symptom of disorders such as depression, anxiety and social phobia, today's study found. Rather, gaming can cause and reinforce those maladies.

"Although children who are depressed may retreat into gaming, the gaming increases the depression," wrote the study authors, led by Douglas A. Gentile, a psychologist at Iowa State University in Ames.

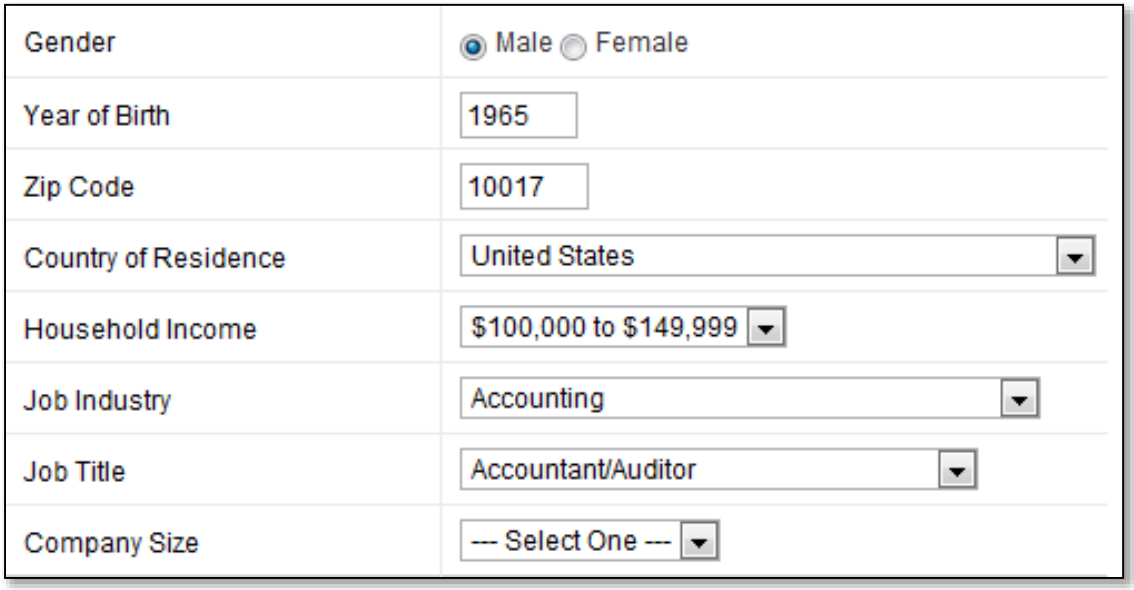
The study, of children in grades 3, 4, 7 and 8, lasted two years. Kids who stopped being pathological gamers during the study period showed lower levels of depression, anxiety and social phobia compared with peers who didn't stop, the researchers said.

To contact the reporter on this story: Elizabeth Lopatto in New York at elopatto@bloomberg.net.

To contact the editor responsible for this story: Reg Gale at rgale5@bloomberg.net.

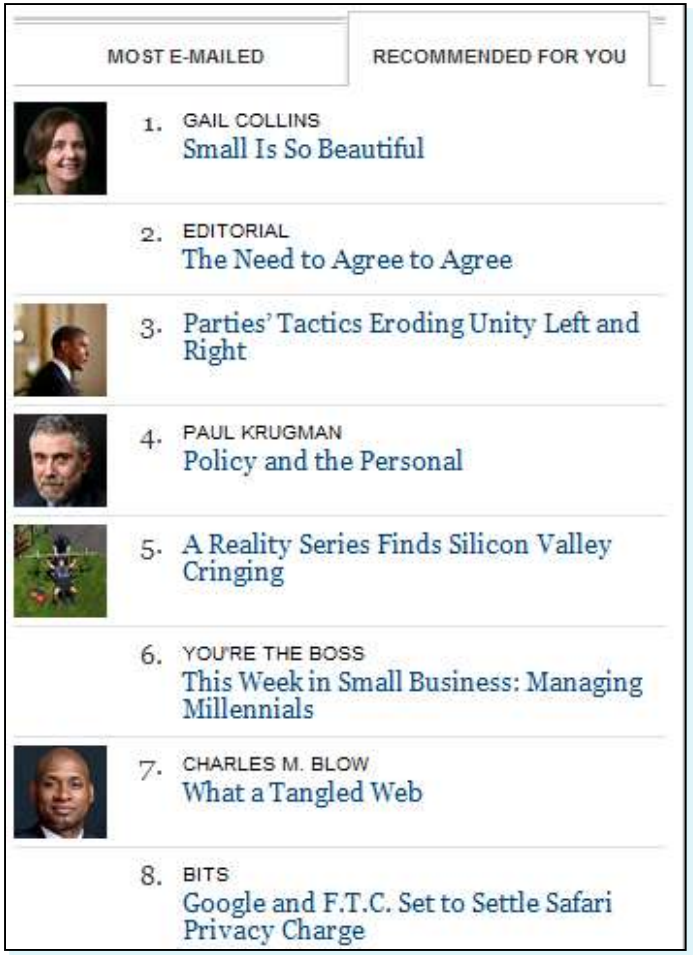
Demographics Example

- Provided only for registered users
  - Only some % of unique users typically register
- Each registered users described with:
  - Gender
  - Year of birth
  - Household income
- Noisy



News recommendation

- List of articles based on
  - Current article
  - User's history
  - Other Visits
- In general, a combination of **text stream** (news articles) with **click stream** (website access logs)
- The key is a rich context model used to describe user



Why news recommendation?

- “Increase in engagement”
  - Good recommendations can make a difference when keeping a user on a web site
  - Measured in number of articles read in a session
- “User experience”
  - Users return to the site
  - Harder to measure and attribute to recommendation module

- Predominant success metric is the attention span of a user expressed in terms of time spent on site and number of page views.

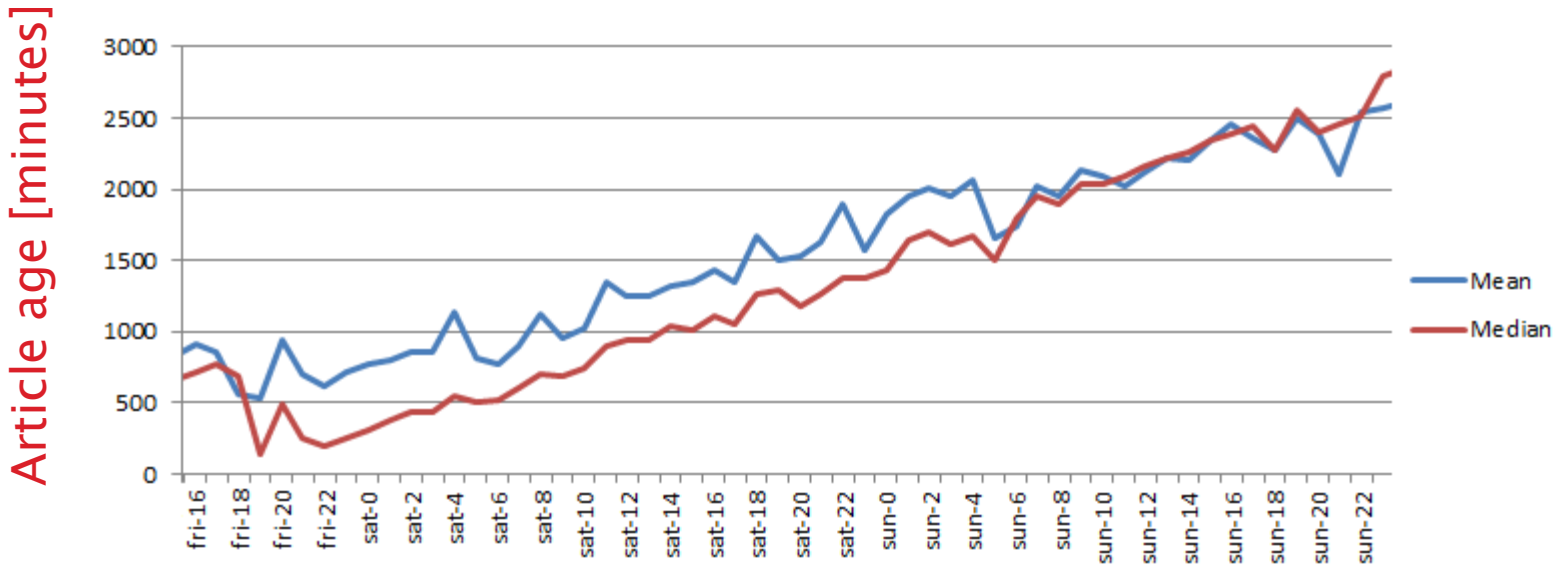
Why is it hard?

- Cold start
  - Recent news articles have little usage history
  - More sever for articles that did not hit homepage or section front, but are still relevant for particular user segment

- Recommendation model must be able to generalize well to new articles.

Example: Bloomberg.com

- Access logs analysis shows, that half of the articles read are less then ~8 hours old
- Weekends are exception



User profile

- History
  - Time
  - Article
- Current request:
  - Location
  - Requested page
  - Referring page
  - Local Time

History:

- Top 2% Not Job Creators or Millionaires in Tax Debate
- Looming Copper Surplus Contracting as Mining Fails: Commodities
- Mayer Becomes Highest-Profile Pregnant Woman Hired as CEO
- Zuckerberg's Loan Gives New Meaning to the 1%
- Brees and Saints Agree on New Contract 3 Days Prior to Deadline
- JPMorgan's Drew Forfeits 2 Years' Pay as Managers Ousted
- JPMorgan's $4.4 Billion CIO Loss Drives Profit Down 9%
- JPMorgan's London Whale Could Use New Nickname

User and current visit:

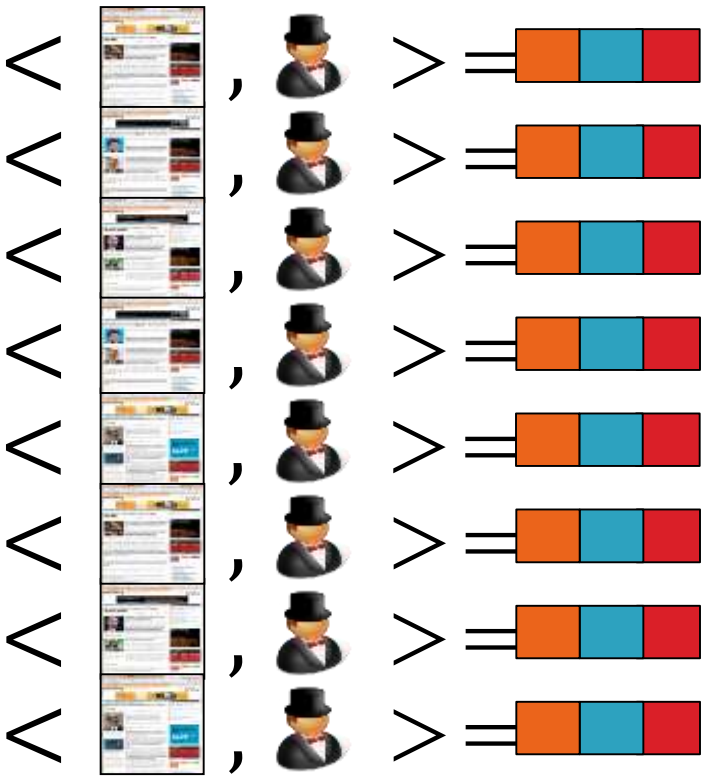
User: 20208908727046187 **Country:** US [NC] **City:** High Point **#Visits:** 8

Refer: http://www.bloomberg.com/

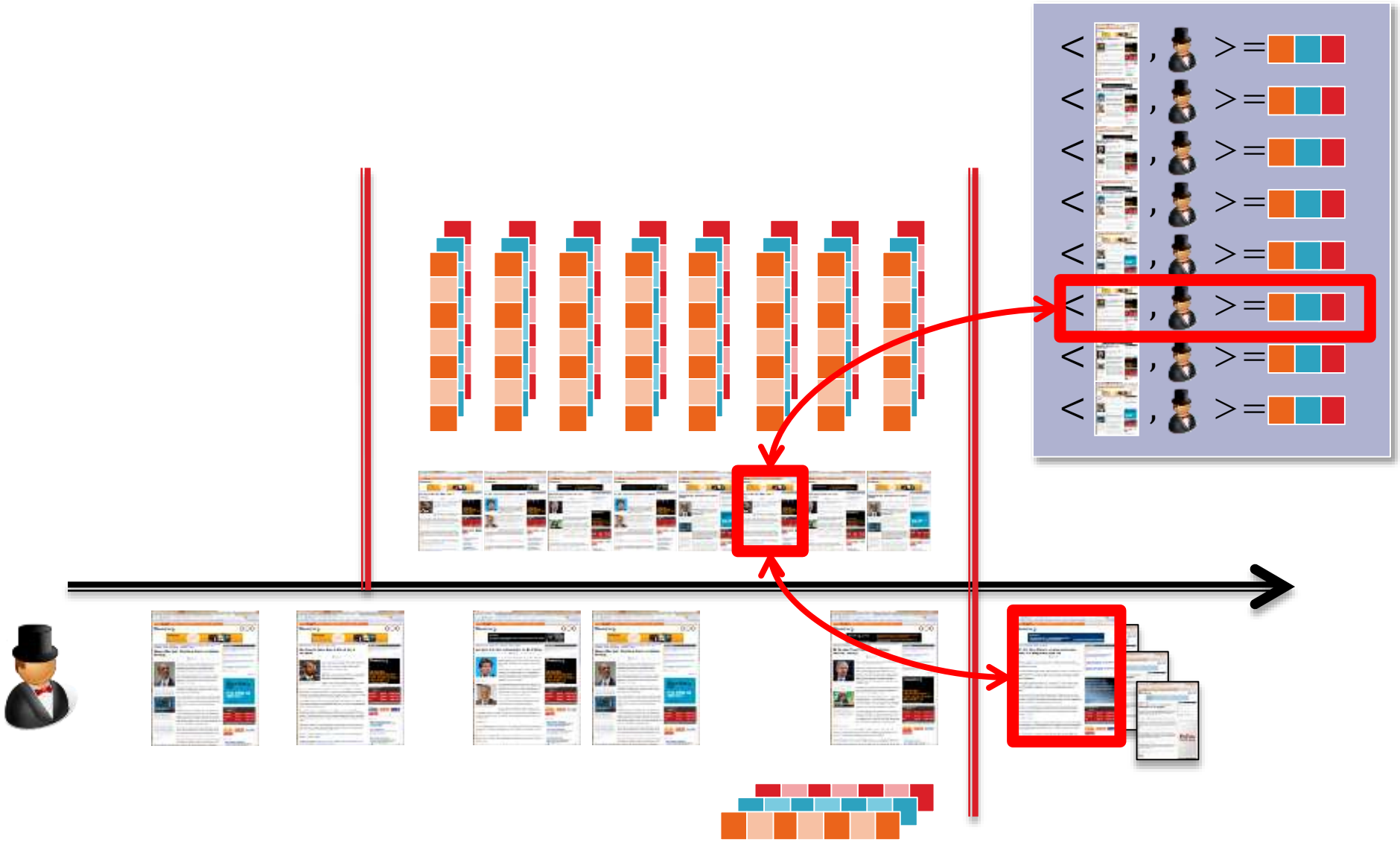
Request: http://www.bloomberg.com/news/2012-07-20/top-2-not-job-creators-or-millionaires-in-tax-debate.html

Features

- Each article from the time window is described with the following features:
  - Popularity (user independent)
  - Content
  - Meta-data
  - Co-visits
  - Users
- Features computed by comparing article's and user's feature vectors
- Features computed on-the-fly when preparing recommendations



Algorithm



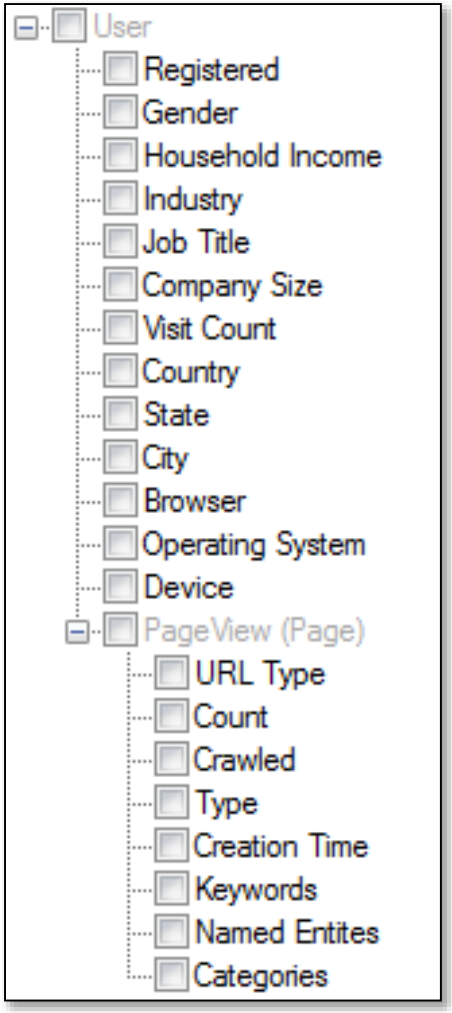
Evaluation

- Measure how many times one of top 4 recommended article was actually read

1	2-10	11-50	51-
21%	24%	32%	37%

User Modeling

- Feature space
  - Extracted from subset of fields
  - Using vector space model
  - Vector elements for each field are normalized
- Training set
  - One visit = one vector
  - One user = a centroid of all his/her visits
  - Users from the segment form positive class
  - Sample of other users form negative class
- Classification algorithm
  - Support Vector Machine
  - Good for dealing with high dimensional data
  - Linear kernel
  - Stochastic gradient descent
    - Good for sampling



Experimental setting

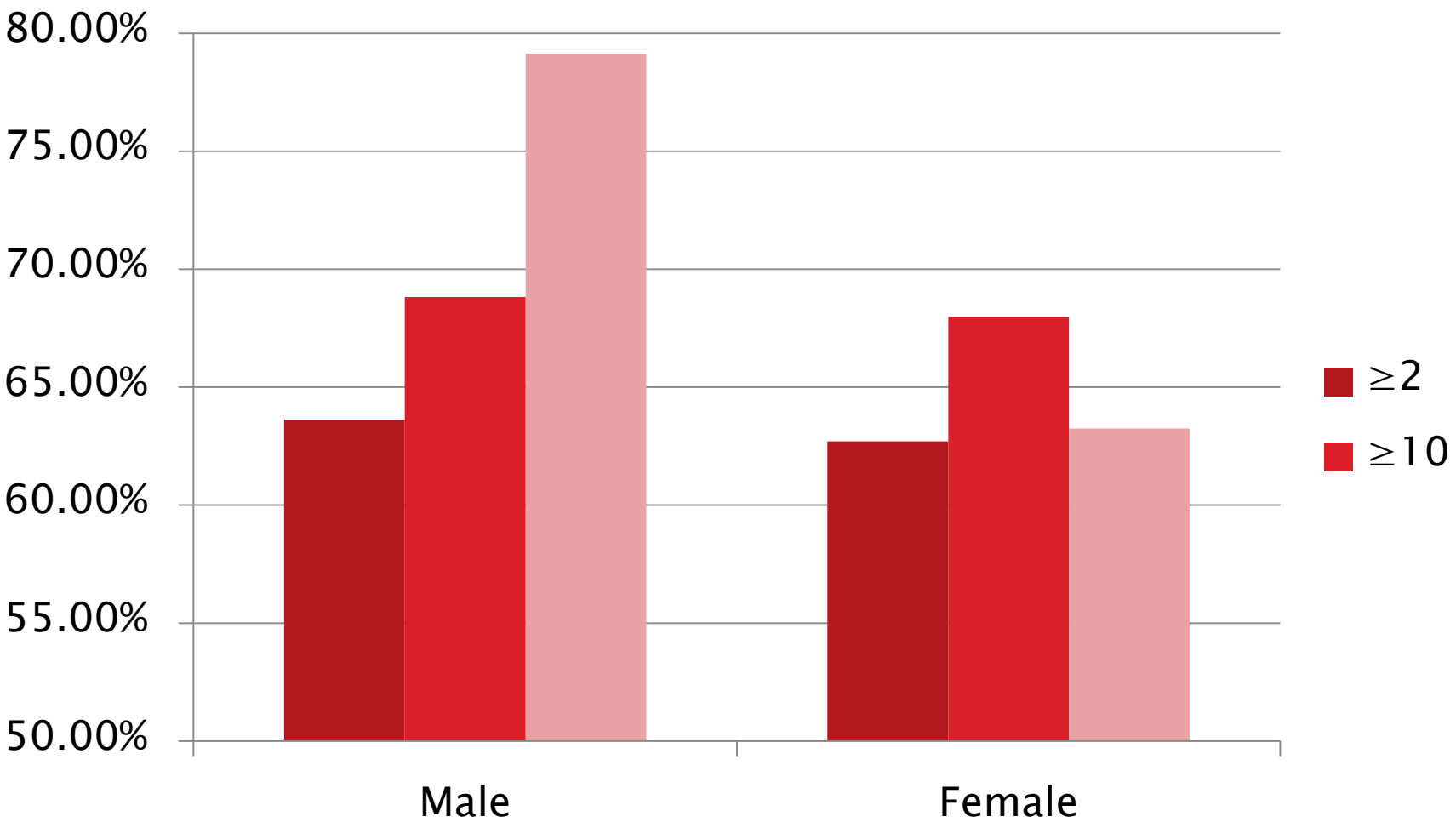
- Real-world dataset from a major news publishing website
  - 5 million daily users, 1 million registered
- Tested prediction of three demographic dimensions:
  - Gender, Age, Income
- Three user groups based on the number of visits:
  - ≥2, ≥10, ≥50
- Evaluation:
  - Break Even Point (BEP)
  - 10-fold cross validation

Category	Size
Male	250,000
Female	250,000

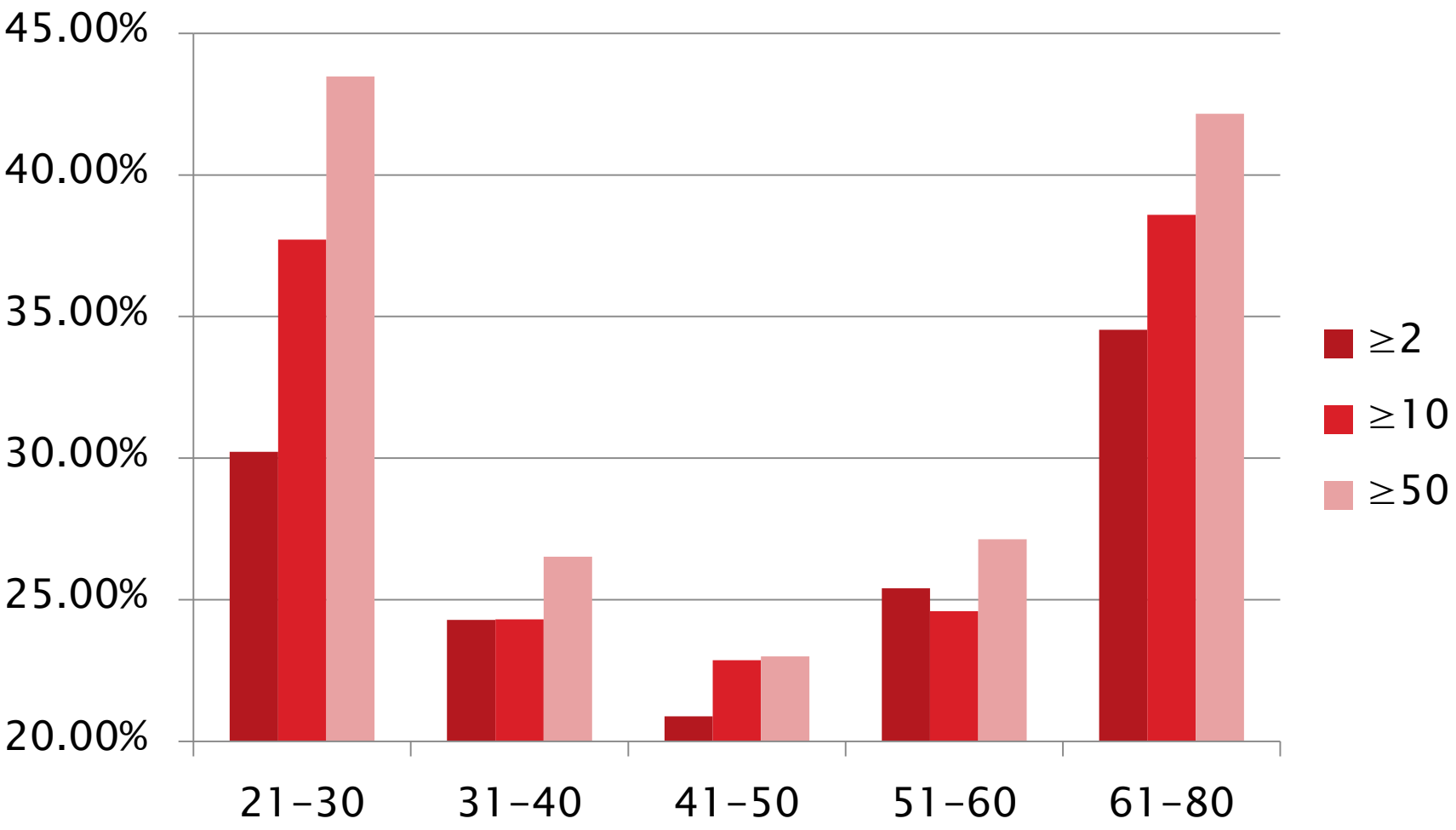
Category	Size
21-30	100,000
31-40	100,000
41-50	100,000
51-60	100,000
61-80	100,000

Category	Size
0-24k	50,000
25k-49k	50,000
50k-74k	50,000
75k-99k	50,000
100k-149k	50,000
150k-254k	50,000

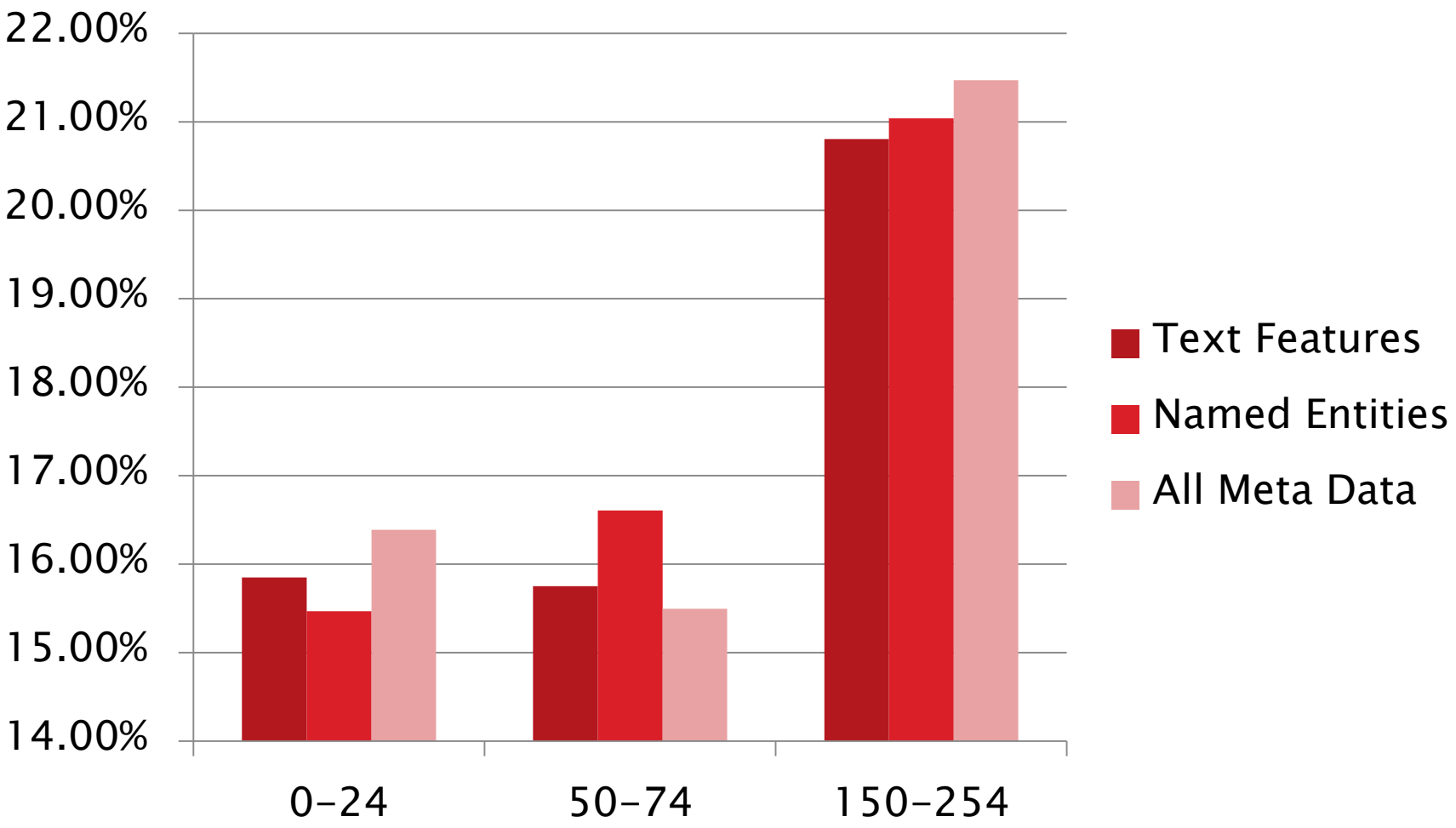
Gender



Age



Income (≥10 visits)



Application: Social-network Analysis

Application: Analysis of MSN-Messenger Social-network

- Observe social and communication phenomena at a *planetary* scale
- **Largest social network analyzed till 2010**

Research questions:

- How does communication change with user demographics (age, sex, language, country)?
- How does geography affect communication?
- What is the structure of the communication **network**?

"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Data statistics: Total activity

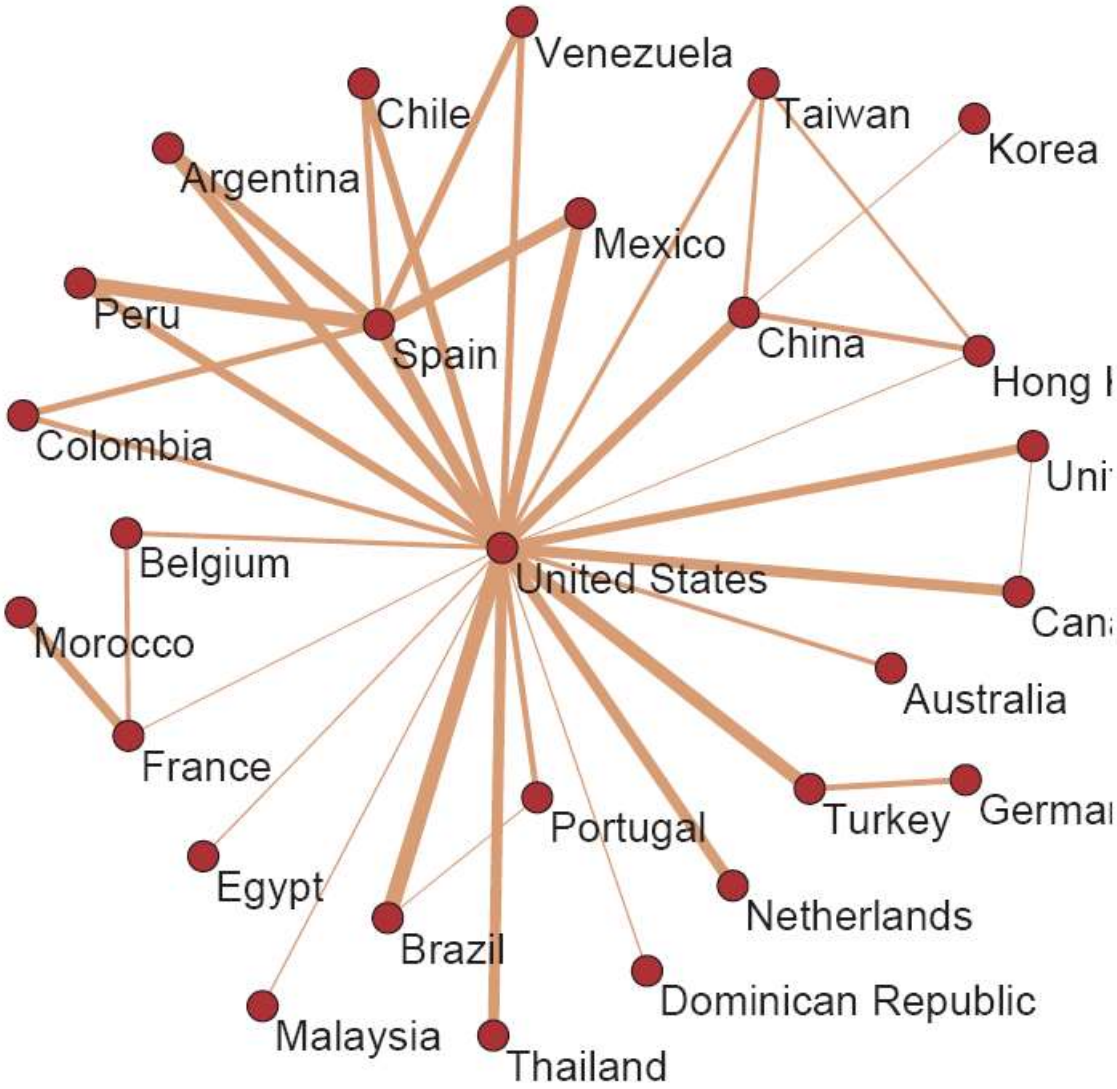
- We collected the data for **June 2006**
- Log size:

  150Gb/day (compressed)
- Total: 1 month of communication data:

  4.5Tb of compressed data
- **Activity over June 2006 (30 days)**
  - 245 million users logged in
  - 180 million users engaged in conversations
  - 17,5 million new accounts activated
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

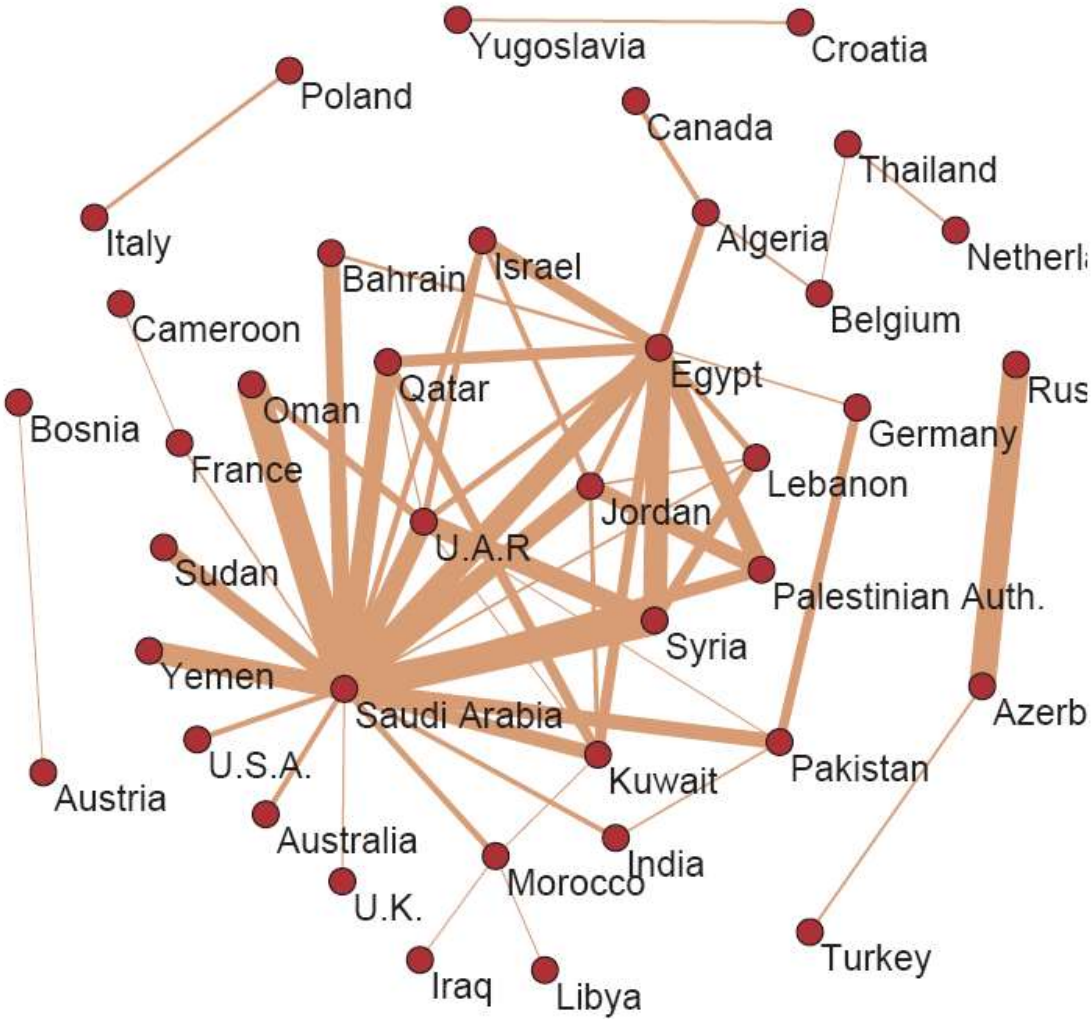
"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Who talks to whom: Number of conversations



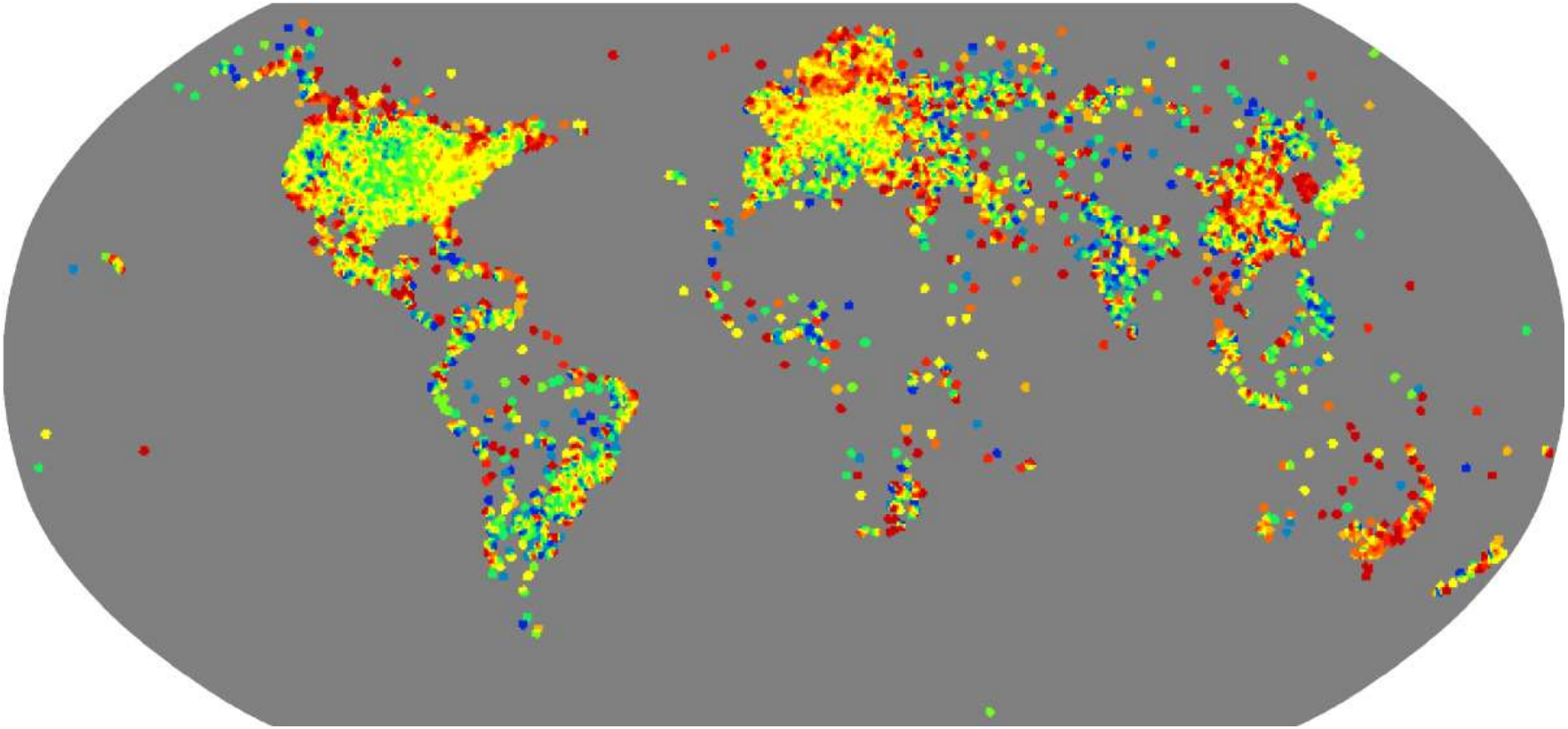
"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Who talks to whom: Conversation duration



"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Geography and communication

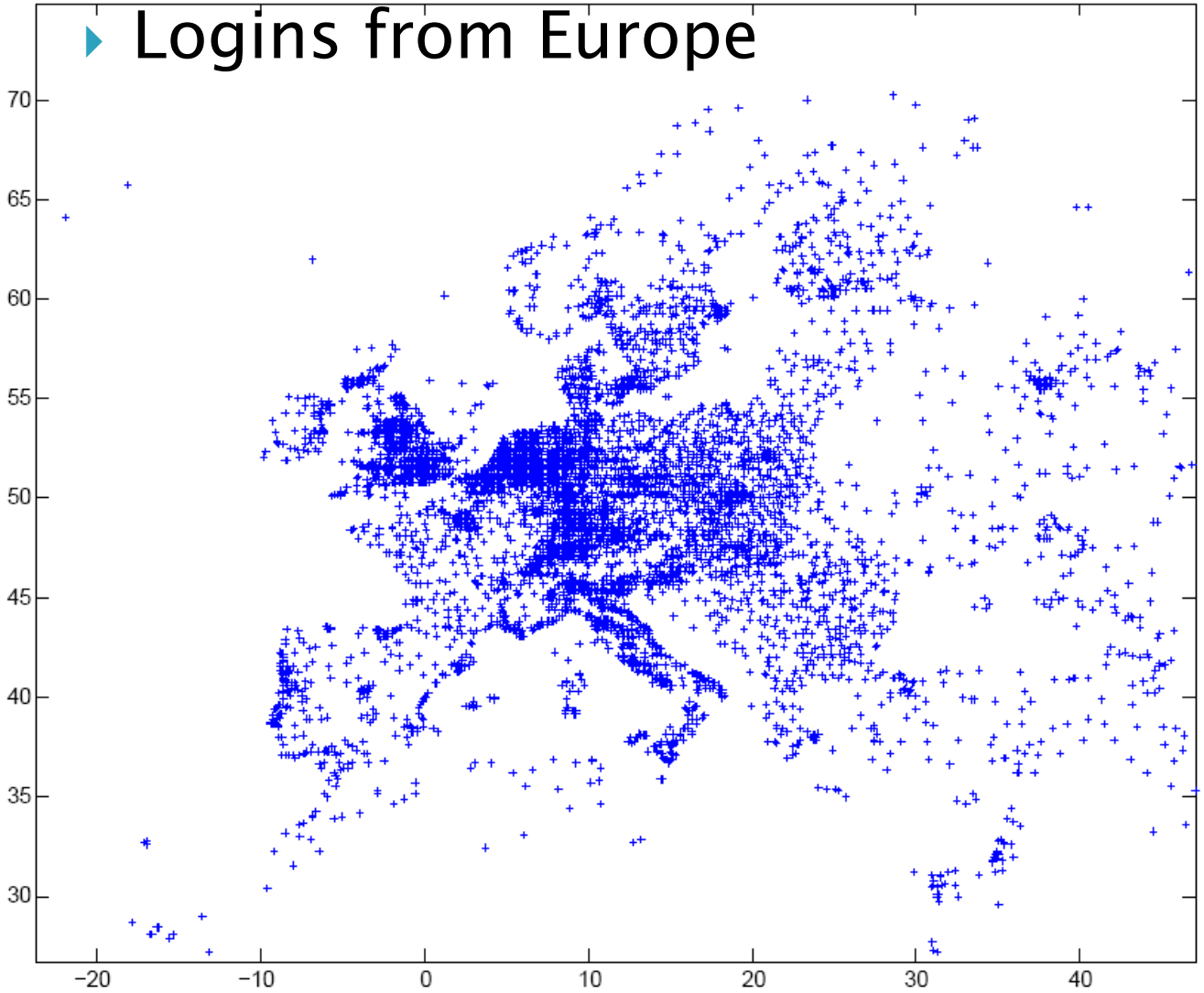


- Count the number of users logging in from particular location on the earth

"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

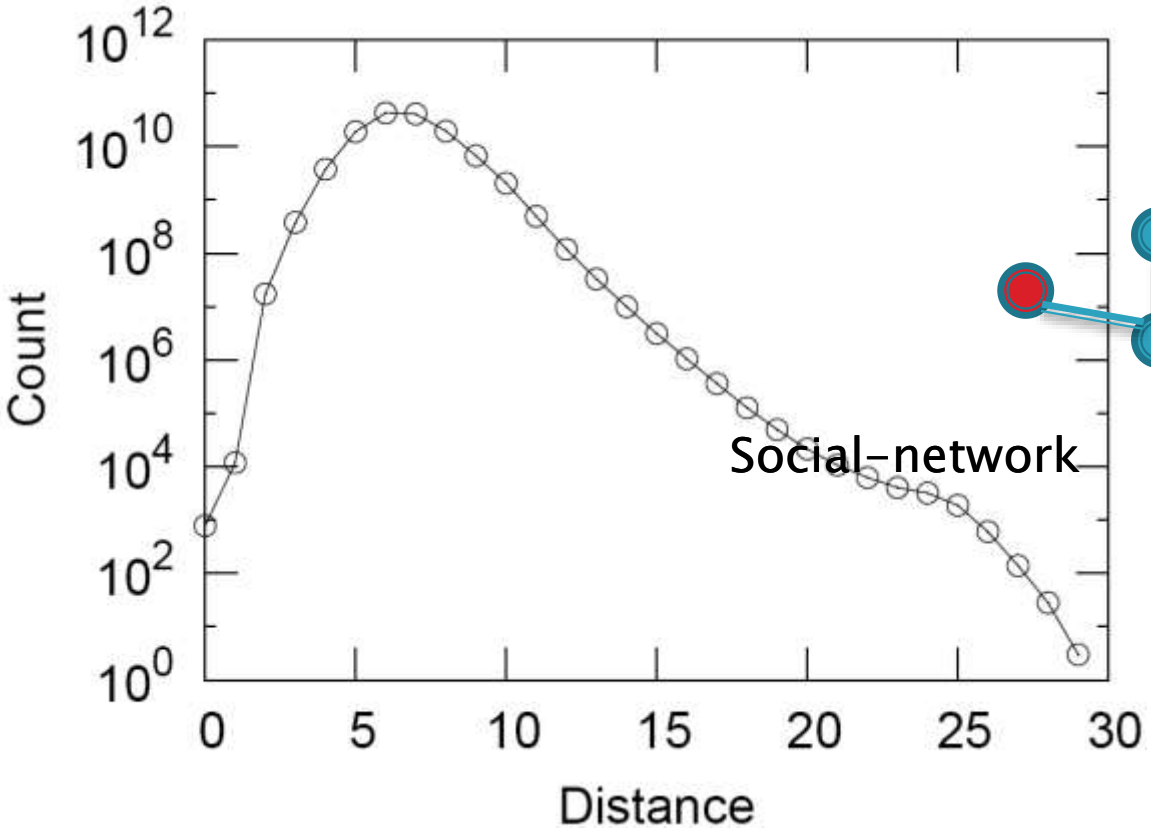
How is Europe talking

- Logins from Europe



"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Network: Small-world



Hops	Nodes
1	10
2	78
3	396
4	8648
5	3299252
6	28395849
7	79059497
8	52995778
9	10321008
10	1955007
11	518410
12	149945
13	44616
14	13740
15	4476
16	1542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

- 6 degrees of separation [Milgram '60s]
- Average distance between two random users is 6.6
- 90% of nodes can be reached in < 8 hops

"Planetary-Scale Views on a Large Instant-Messaging Network" Leskovec & Horvitz WWW2008

Application: Global Media Monitoring

Application: Monitoring global media

- The aim of the project is to collect and analyze global main-stream and social media
  - …documents are crawled from 100 thousands of sources
  - …each crawled document gets cleaned, linguistically and semantically enriched
  - …we connect documents across languages (cross-lingual technology)
  - …we identify and connect events

SEVENTH FRAMEWORK PROGRAMME

http://render-project.eu/ http://www.xlike.org/

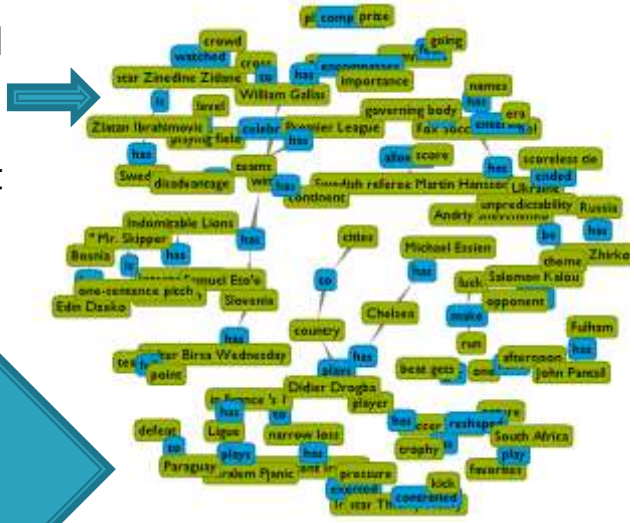
Collecting global media in near-real-time (http://newsfeed.ijs.si)

- The NewsFeed.ijs.si system collects
  - 40.000 main-stream news sources
  - 250.000 blog sources
  - Twitter stream
- …resulting in ~500.000 documents + #N of twits per day
- Each document gets cleaned, linguistically and semantically annotated

Semantic text enrichment (DBpedia, OpenCyc, …) with Enrycher (http://enrycher.ijs.si/)

Plain text

Extracted graph of triples from text



Text Enrichment

Slovenia's dramatic w… Russia Wednesday, and to a lesser extent Ireland's narrow loss to F…ance, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a centu… from the likes of Brazil, Italy and Germany, interr… entering the era of the Cinderella. It may not hap… but given the increasing flow of talent, training a… borders, it's almost certain that a small upstart na… athletes and better luck will make a legitimate ru… coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Valter Birsa Wednesday.

entities

- Brazil
- Italy
- Germany
- Cinderella
- Paris
- John O'Shea
- Manchester United
- Robbie Keane
- Shay Given
- Greece
- Portugal
- Bosnia-Herzegovina
- Cristiano Ronaldo
- Uruguay

keywords

Sports, Soccer, CONCACAF, Competitions, United States, Sports and Hobbies, Kids and Teens, World Cup, Women,

categories

- Top/Kids_and_Teens /Sports_and_Hobbies /Sports/Soccer
- Top/Sports/Soccer /Competitions
- Top/Sports/Soccer /Competitions/World_Cup
- Top/Sports/Soccer /CONCACAF

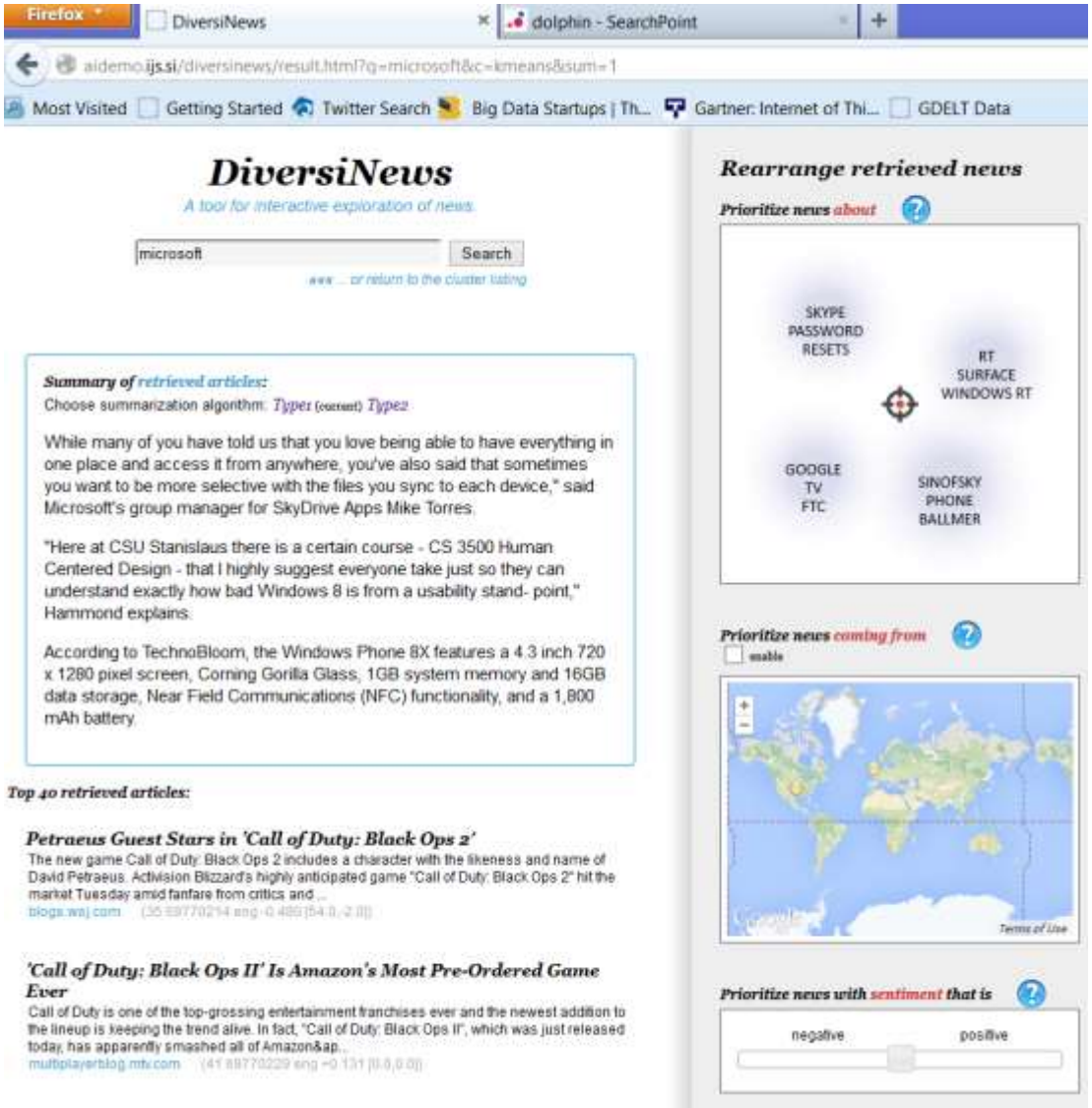
Diego Maradona Semantics:
owl:sameAs: http://dbpedia.org/resource/Diego_Maradona
owl:sameAs: http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA
rdf:type: http://dbpedia.org/class/yago/ArgentinaInternationalFootballers
rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers

Robbie Keane Semantics:
owl:sameAs: http://dbpedia.org/resource/Robbie_Keane
rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly
rdf:type: http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers

"Enrycher" is available as as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization

DiversiNews – exploring news diversity (http://aidemo.ijs.si/diversinews/)

- Reporting has bias – same information is being reported in different ways
- DiversiNews system allows exploring news diversity along:
  - Topicality
  - Geography
  - Sentiment

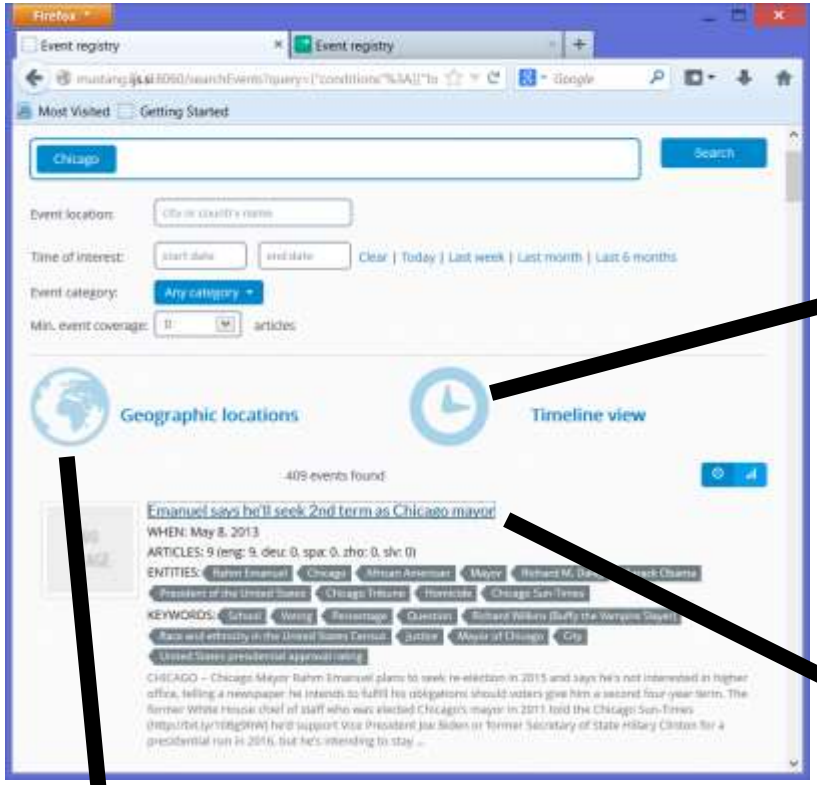


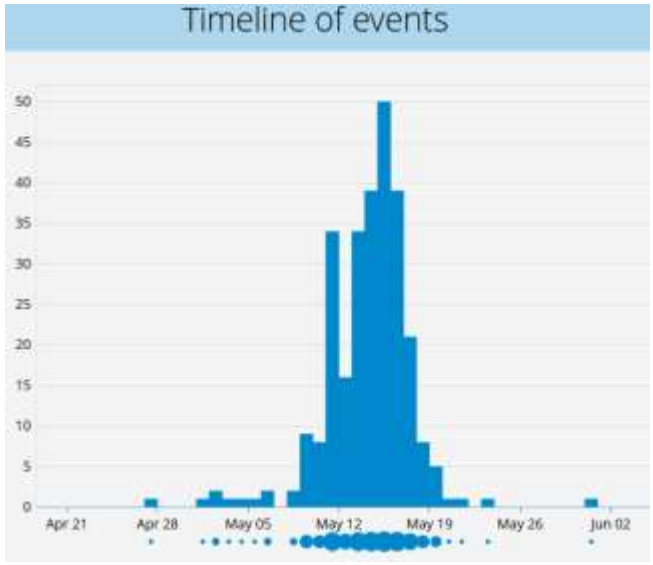
“Event Registry” system for event identification and tracking

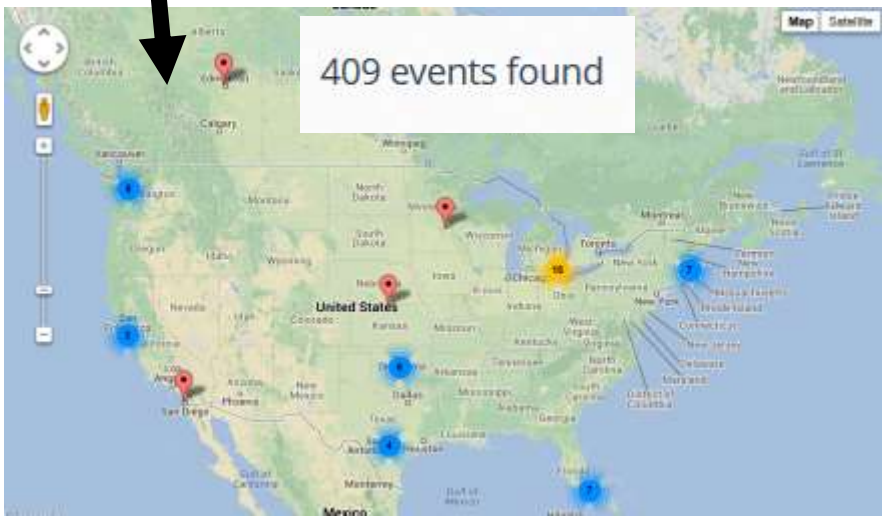
- Having stream of news & social media, the task is to structure documents into events
- “Event Registry” system allows for:
  - Identification of events from documents
  - Connecting documents across many languages
  - Tracking events and constructing story-lines
  - Describing events in a (semi)structured way
  - UI for exploration through Search & Visualization
  - Export into RDF (Storyline ontology)

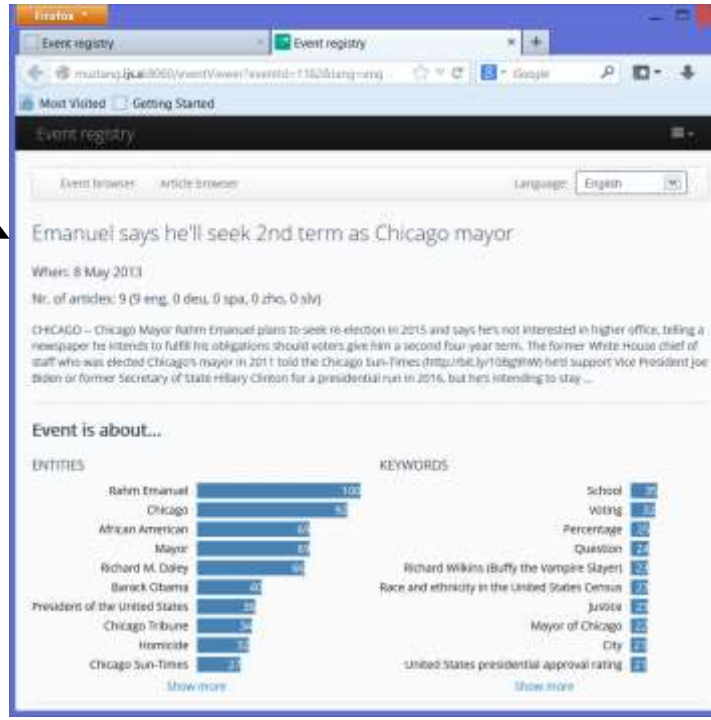
- Prototype operating at
  - http://mustang.ijs.si:8060/searchEvents

“Event Registry” example on “Chicago” related events



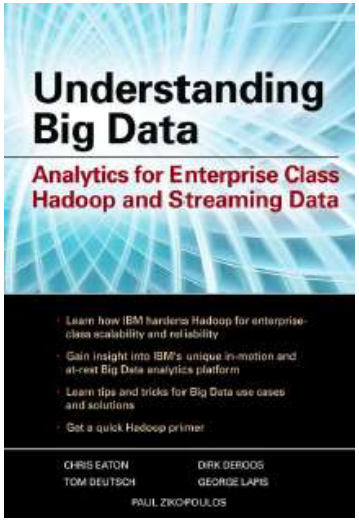


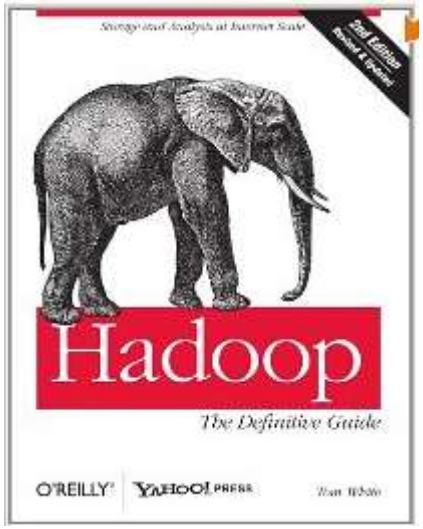


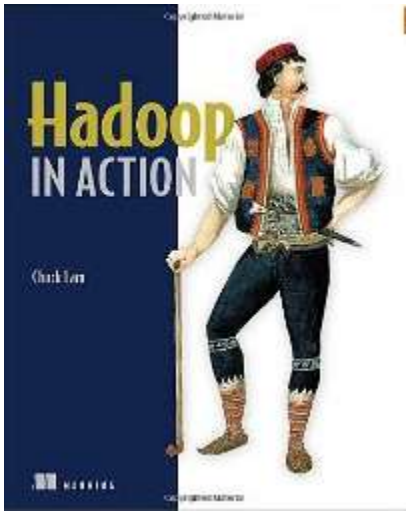


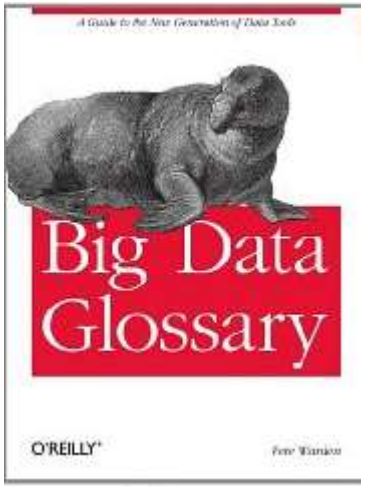
Final thoughts

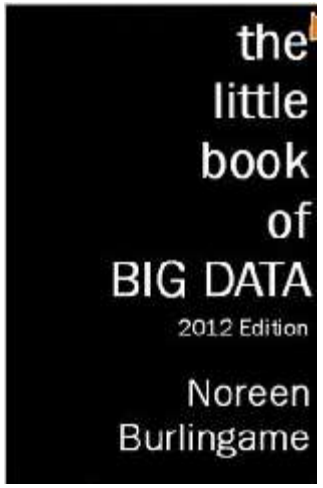
Literature on Big-Data

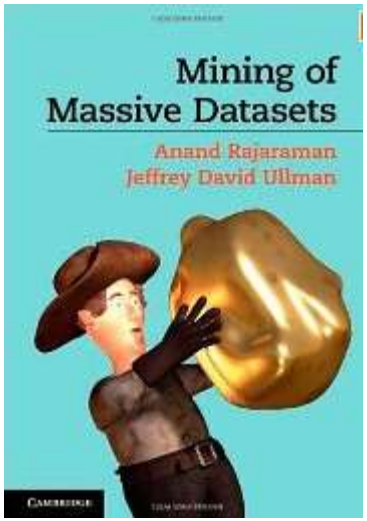


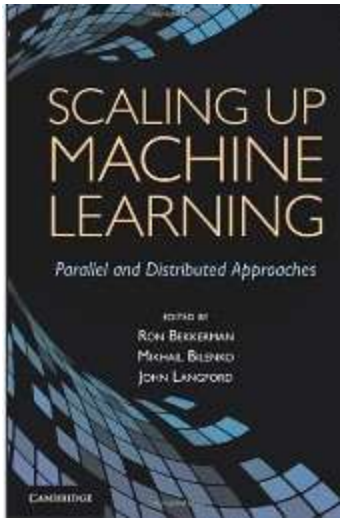












…to conclude

- Big-Data is everywhere, we are just not used to deal with it
- The “Big-Data” hype is very recent
  - …growth seems to be going up
  - …evident lack of experts to build Big-Data apps
- Can we do “Big-Data” without big investment?
  - …yes – many open source tools, computing machinery is cheap (to buy or to rent)
  - …the key is knowledge on how to deal with data
  - …data is either free (e.g. Wikipedia) or to buy (e.g. twitter)