

# Big Data

Source:

Tutorial: Introduction to Big Data

Marko Grobelnik, Blaz Fortuna, Dunja Mladenic

Jozef Stefan Institute, Slovenia

<http://ailab.ijs.si/~blazf/BigDataTutorial-GrobelnikFortunaMladenic-ISWC2013.pdf>

# Big-Data in numbers

# *Big data—a growing torrent*

**\$600** to buy a disk drive that can store all of the world's music

**\$5 million vs. \$400**

Price of the fastest supercomputer in 1975<sup>1</sup> and an iPhone 4 with equal performance

**5 billion** mobile phones in use in 2010

**30 billion** pieces of content shared on Facebook every month

**40%** projected growth in global data generated per year vs. **5%** growth in global IT spending

**235** terabytes data collected by the US Library of Congress by April 2011

**15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress





# TOP 10 MOST VISITED WEB PROPERTIES

Google™

Unique Visitors Per Month  
**153,441,000**

Time Spent Per Person  
Per Month in hh:mm:ss **1:47:42**

facebook

Unique Visitors Per Month  
**137,644,000**

Time Spent Per Person  
Per Month in hh:mm:ss **7:45:49**

	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
You Tube	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



# 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

# Big-Data Definitions

# ...so, what is Big-Data?

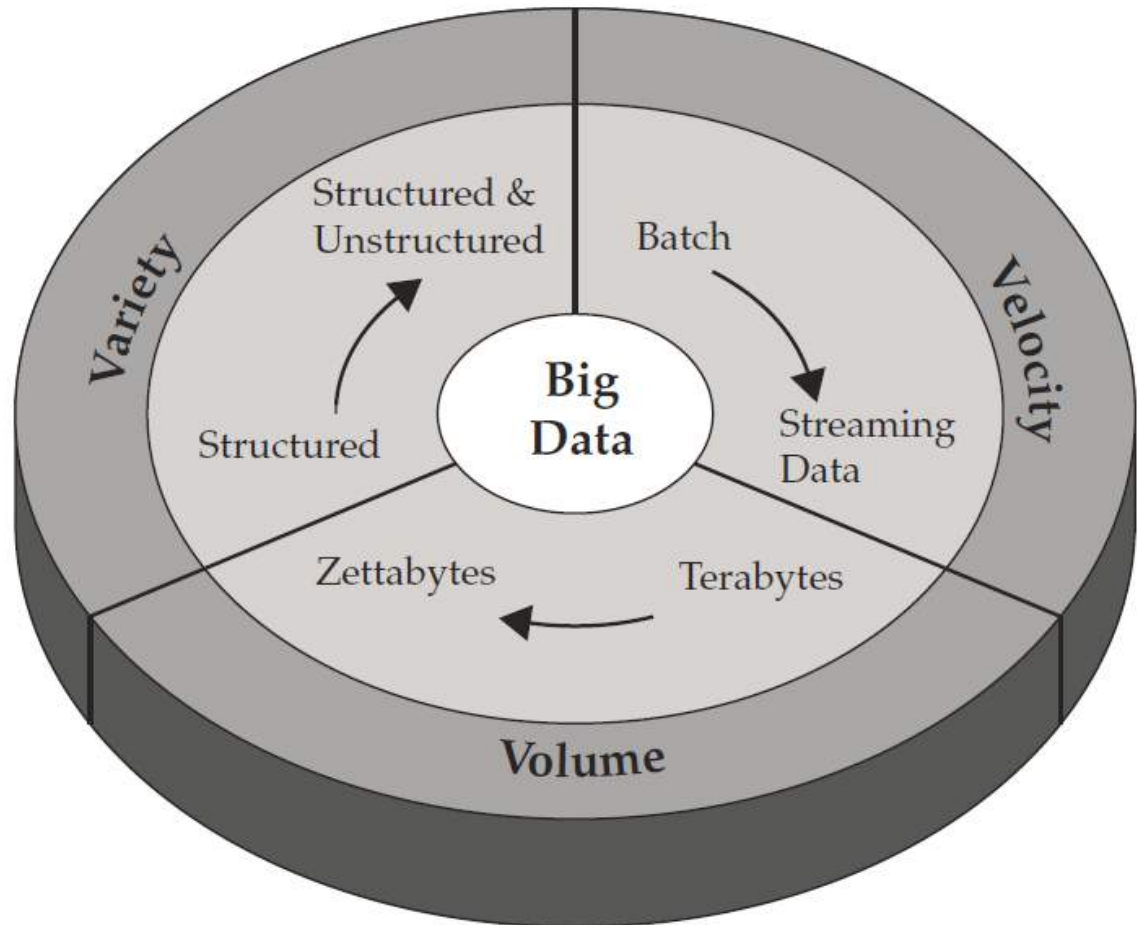
- ▶ 'Big-data' is similar to 'Small-data', but bigger
- ▶ ...but having data bigger it requires different approaches:
  - techniques, tools, architectures
- ▶ ...with an aim to solve new problems
  - ...or old problems in a better way.





# 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

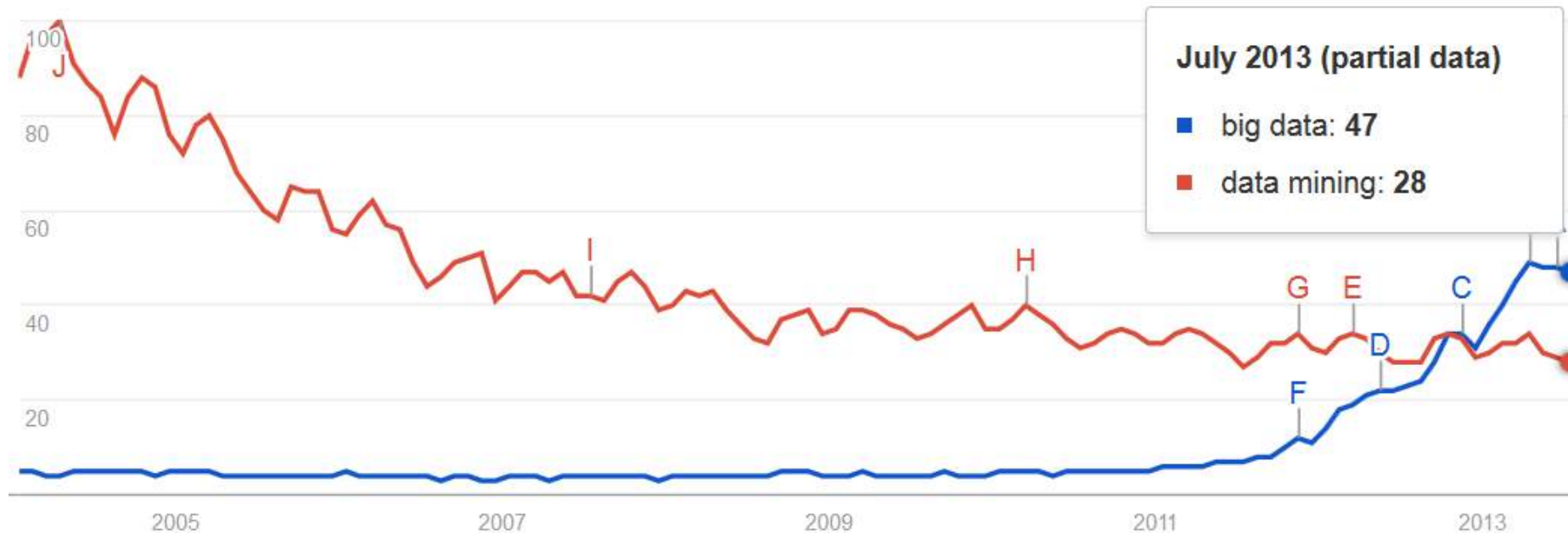
# The extended 3+n Vs of Big Data

- ▶ 1. **Volume** (lots of data = “Tonnabytes”)
- ▶ 2. **Variety** (complexity, curse of dimensionality)
- ▶ 3. **Velocity** (rate of data and information flow)
- ▶ 4. **Veracity** (need to keep data clean)
- ▶ 5. **Variability**
- ▶ 6. **Venue** (location)
- ▶ 7. **Vocabulary** (semantics)

# Motivation for Big-Data

# Big-Data popularity on the Web (through the eyes of “Google Trends”)

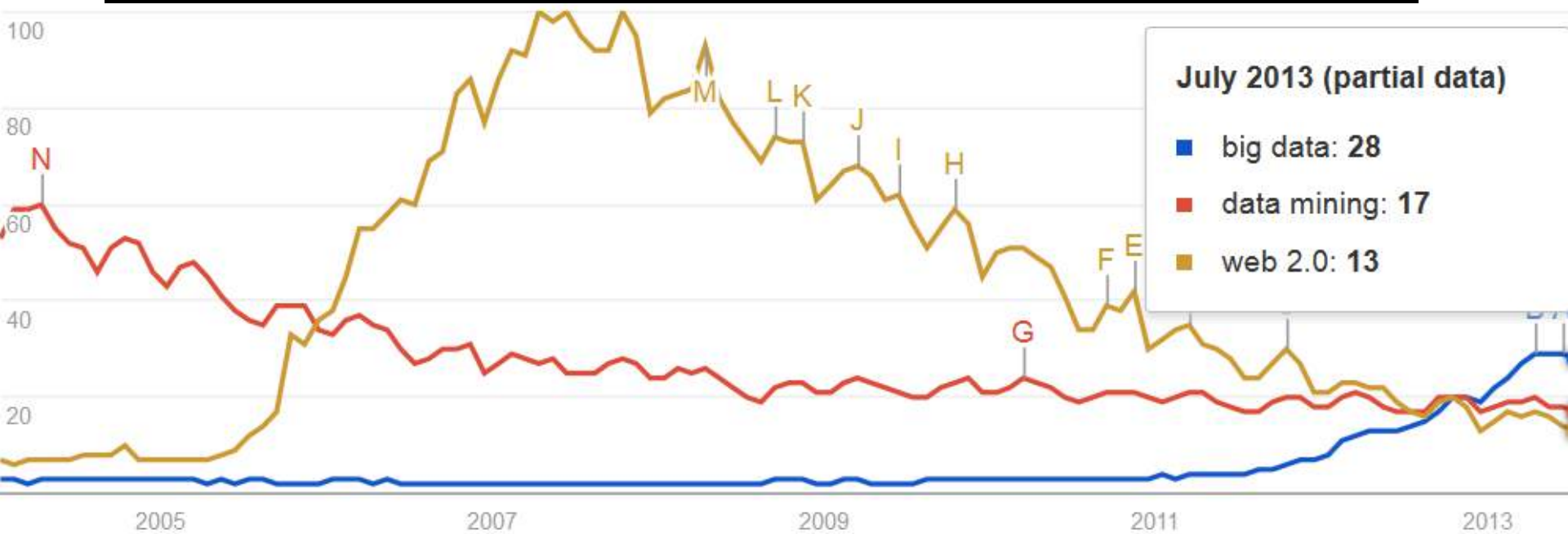
Comparing volume of “big data” and “data mining” queries





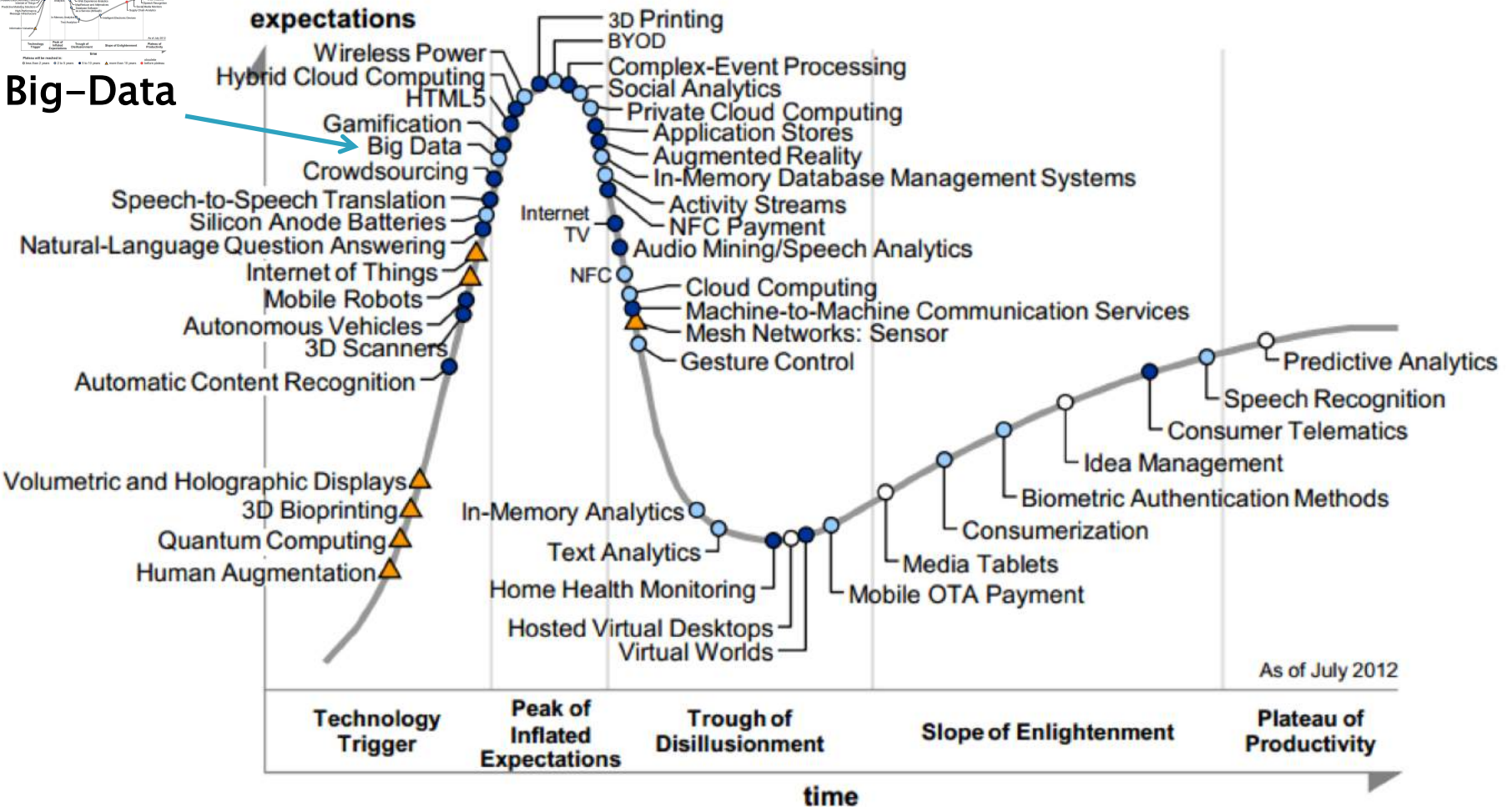
# ...but what can happen with “hypes”

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



# Emerging Technologies Hype Cycle 2012

Big-Data



Plateau will be reached in:

○ less than 2 years

○ 2 to 5 years

● 5 to 10 years

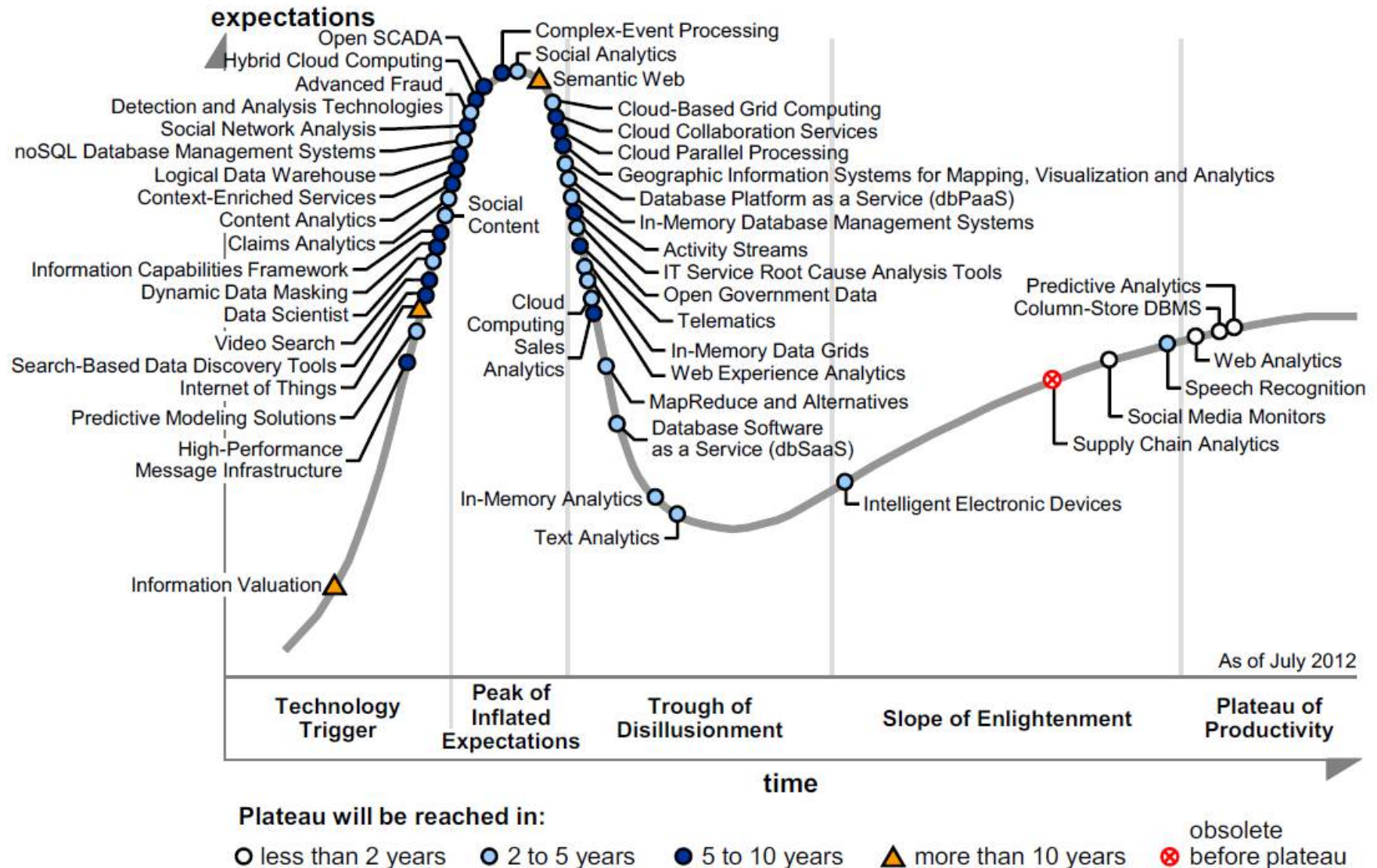
▲ more than 10 years

○ obsolete

⊗ before plateau

# Gartner Hype Cycle for Big Data, 2012

Figure 1. Hype Cycle for Big Data, 2012



# Why Big-Data?

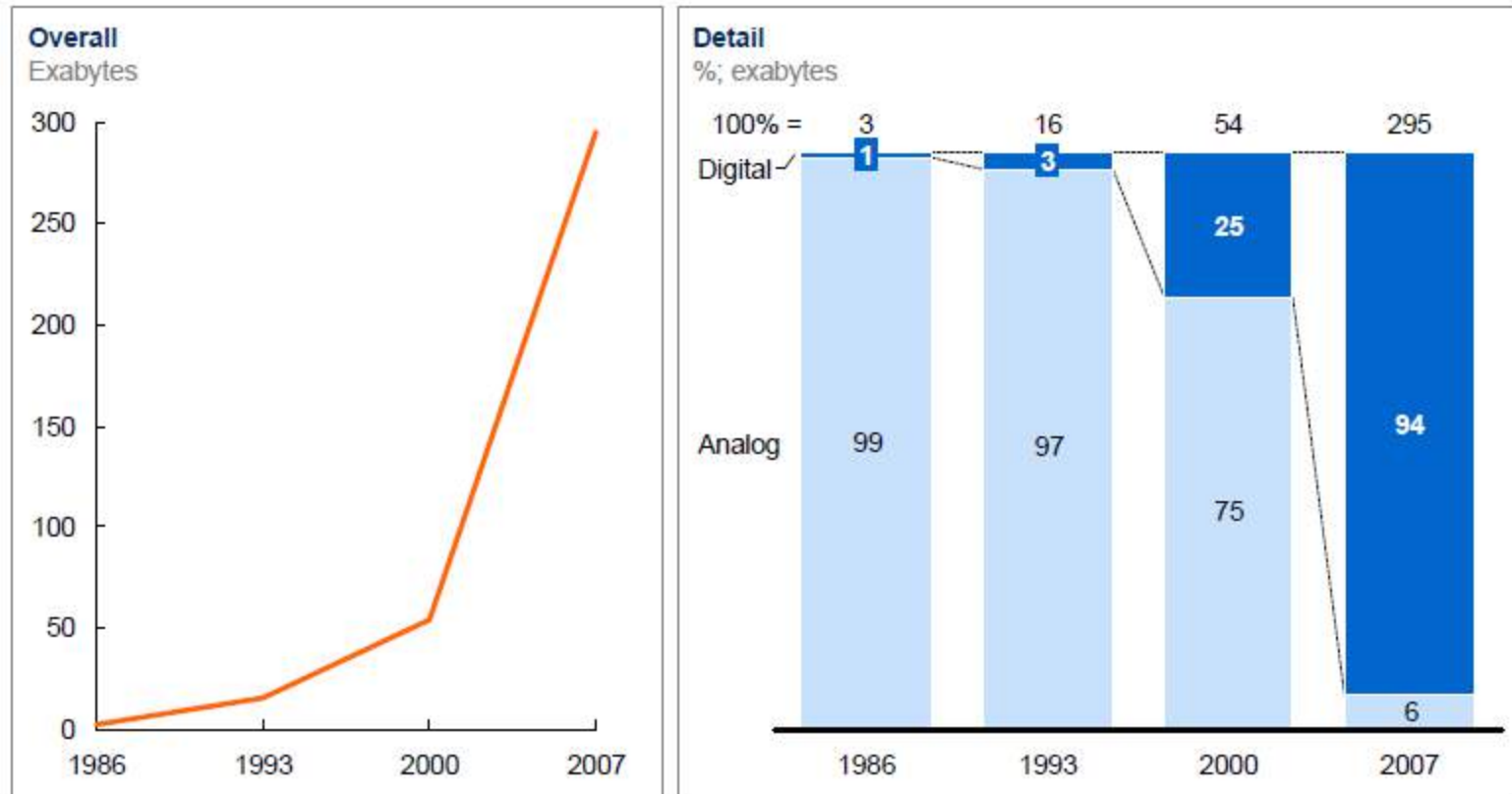
- ▶ Key enablers for the appearance and growth of “Big Data” are:
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data



# Enabler: Data storage

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

Global installed, optimally compressed, storage



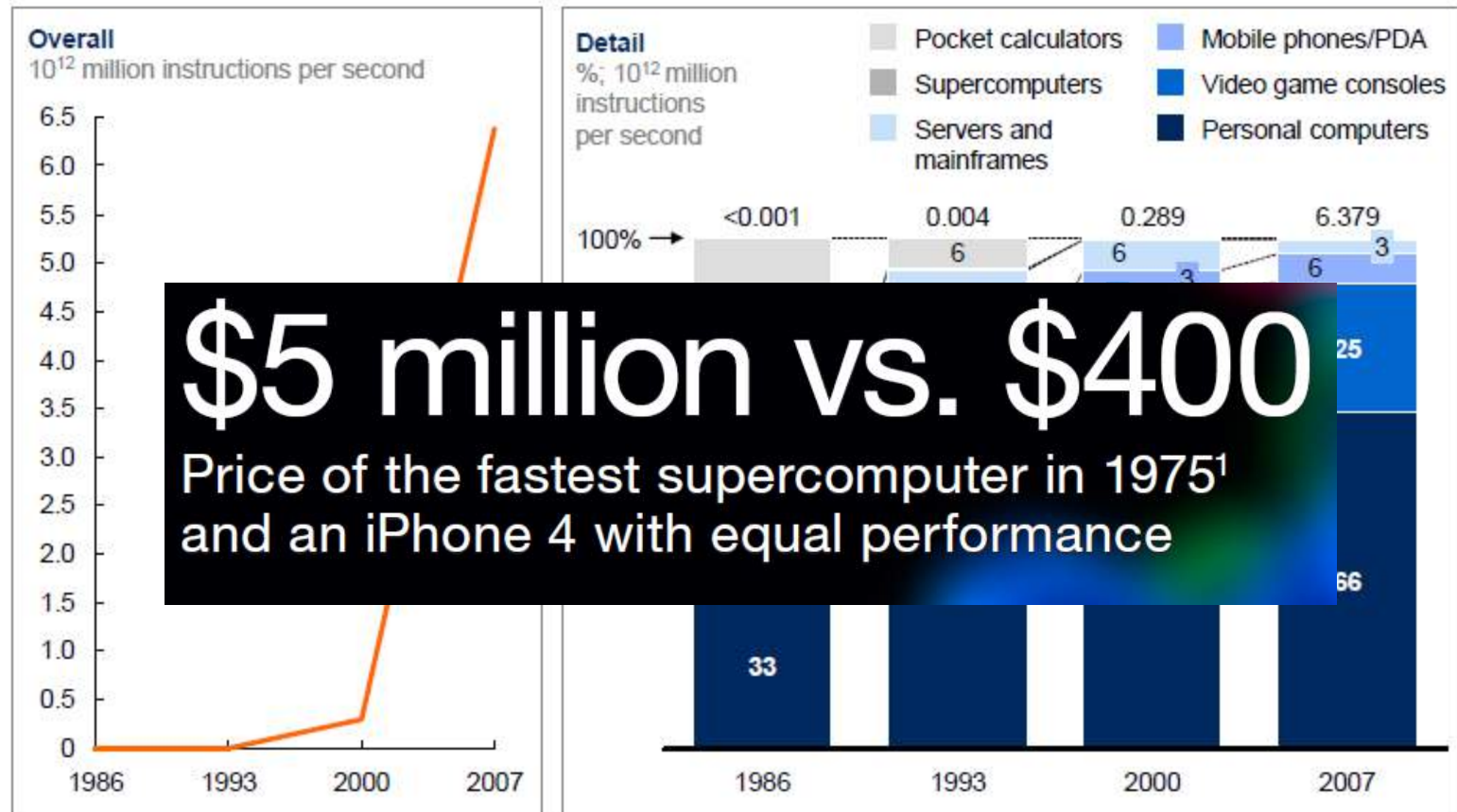
NOTE: Numbers may not sum due to rounding.

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

# Enabler: Computation capacity

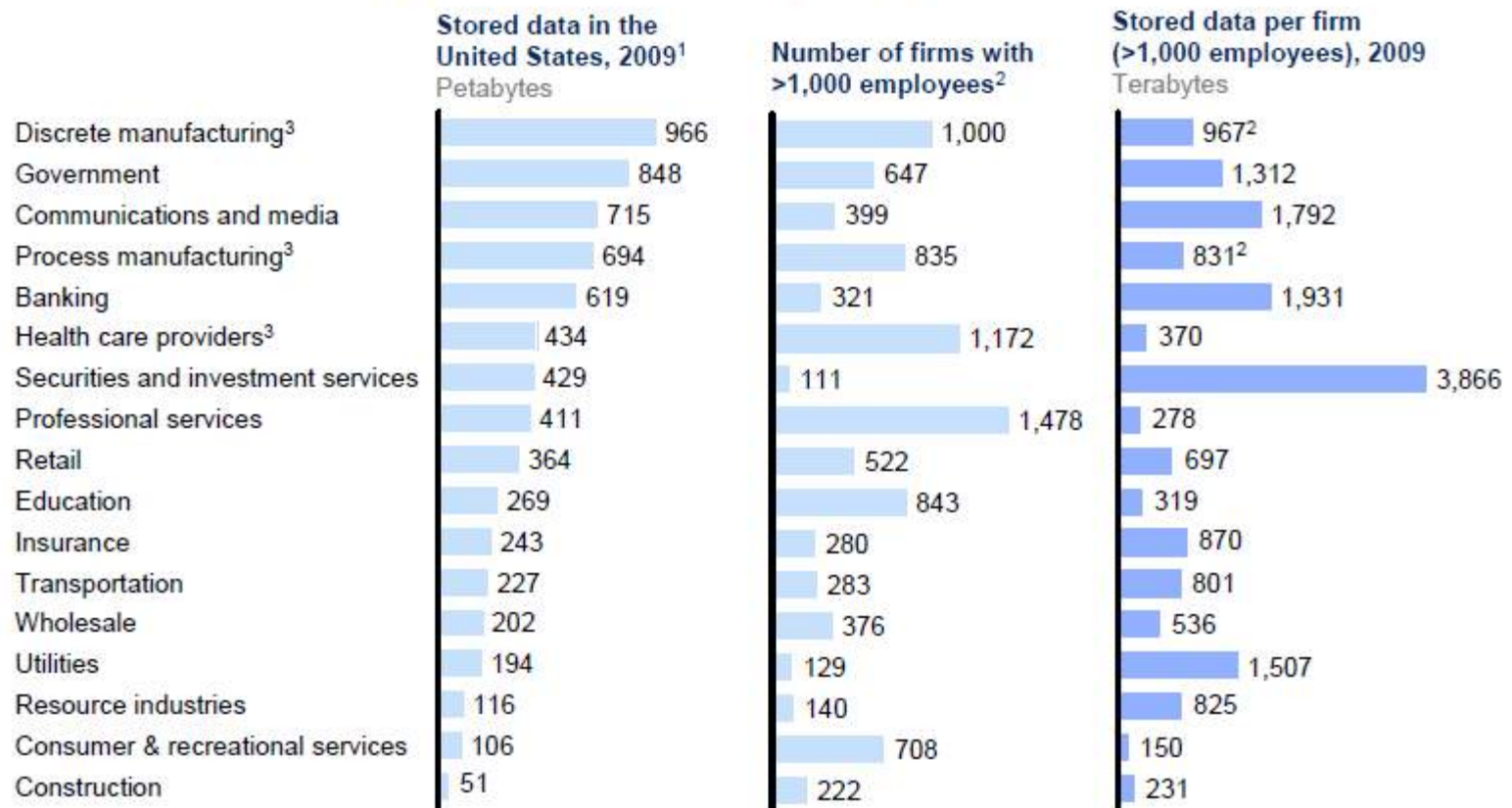
## Computation capacity has also risen sharply

Global installed computation to handle information



# 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.

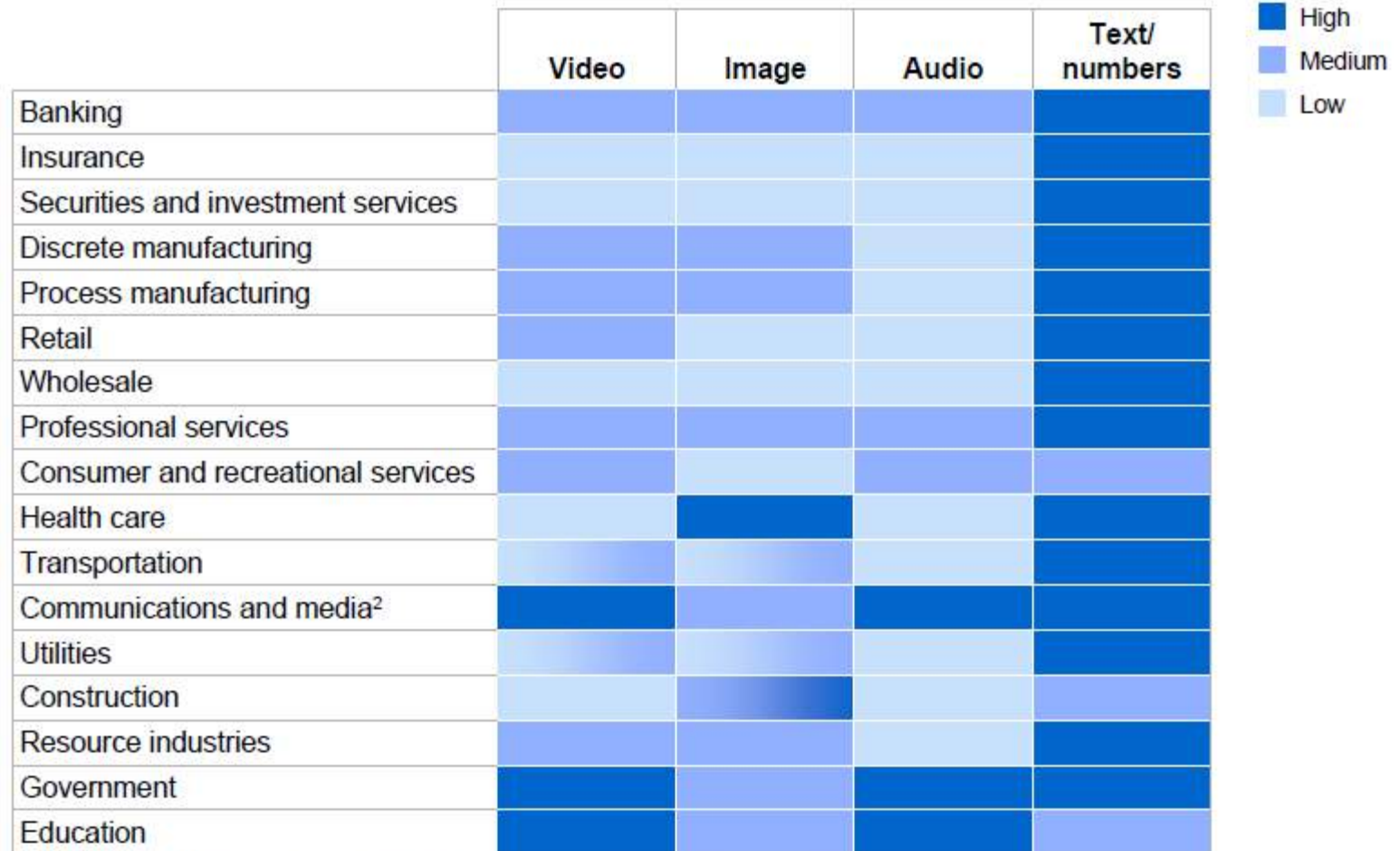
2 Firm data split into sectors, when needed, using employment.

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<sup>1</sup>



<sup>1</sup> We compiled this heat map using units of data (in files or minutes of video) rather than bytes.

<sup>2</sup> Video and audio are high in some subsectors.

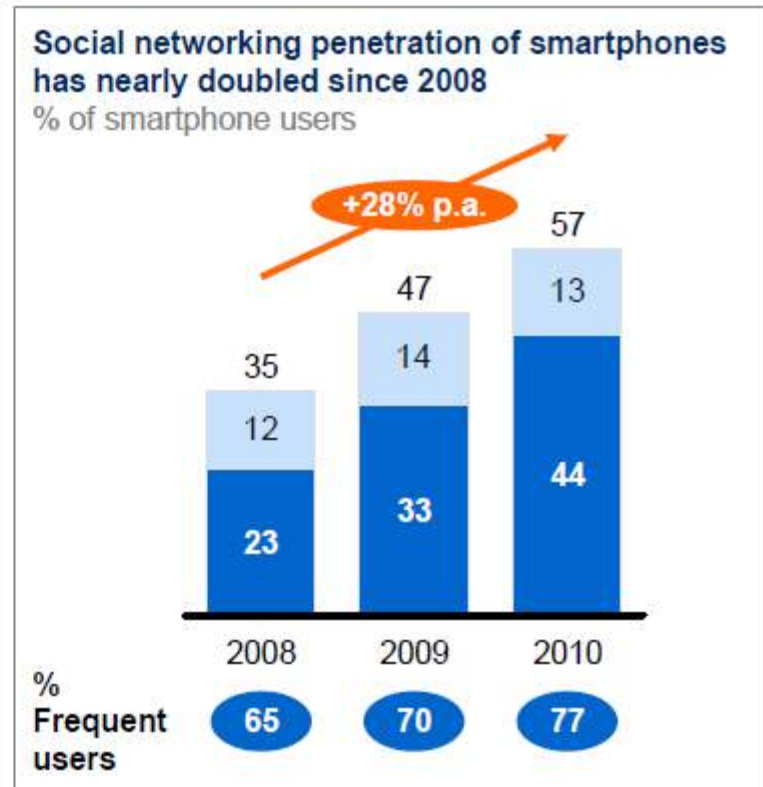
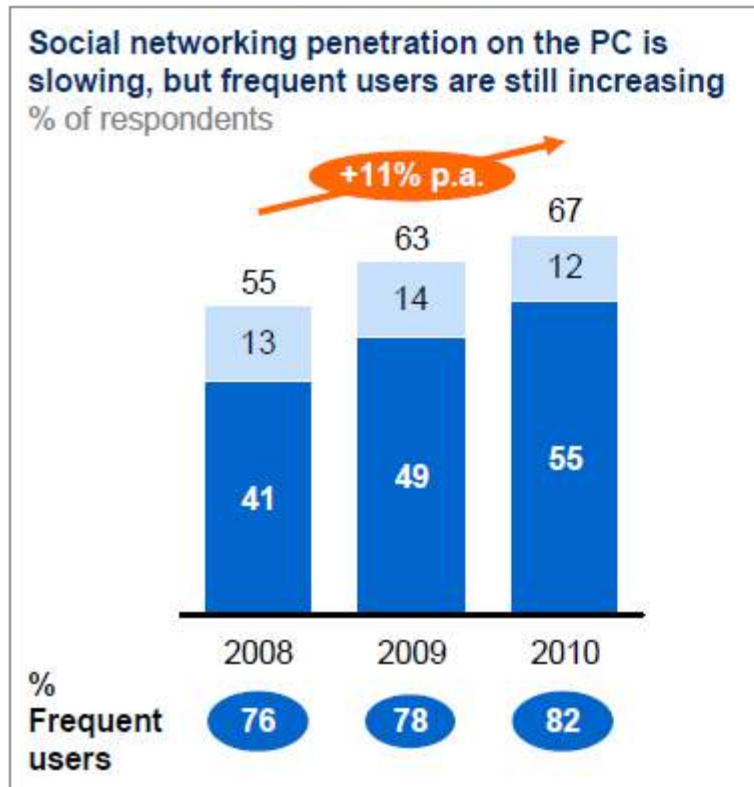
SOURCE: McKinsey Global Institute analysis



# 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<sup>1</sup>

■ Frequent user<sup>2</sup>



- 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

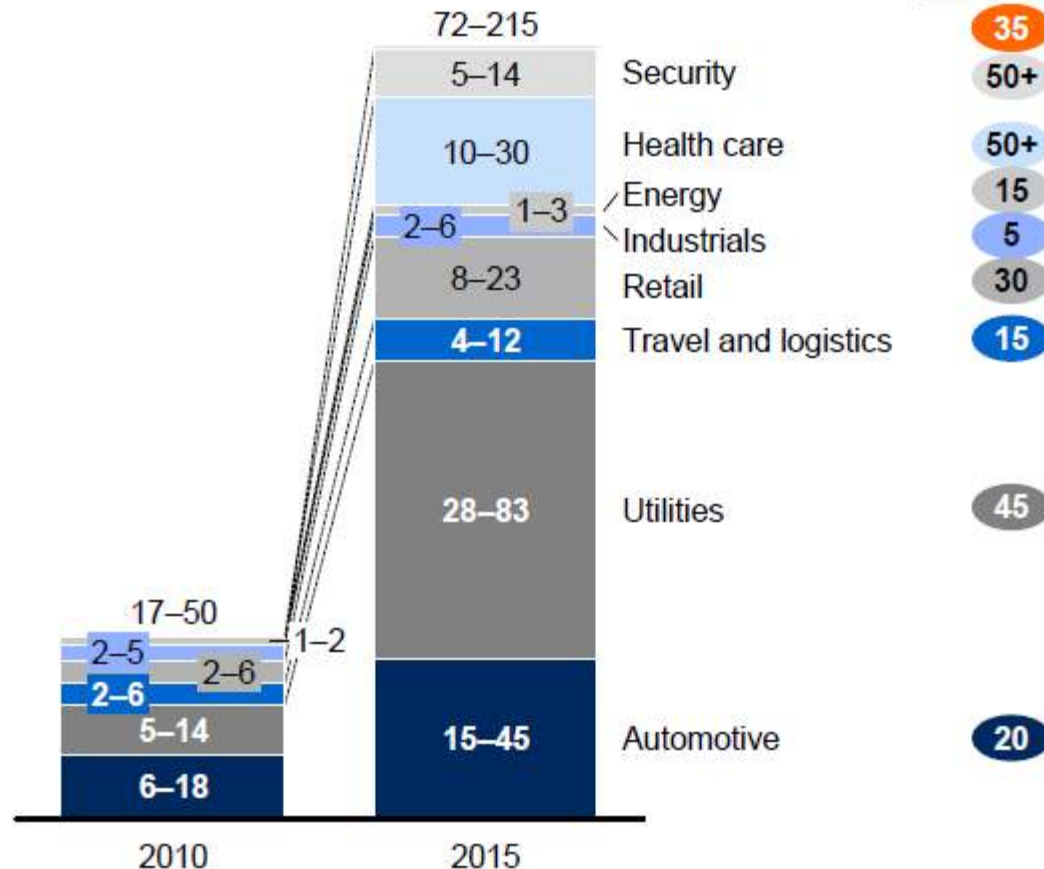
# Data available from “Internet of Things”

**Data generated from the Internet of Things will grow exponentially as the number of connected nodes increases**

Estimated number of connected nodes

Million

Compound annual growth rate 2010–15, %

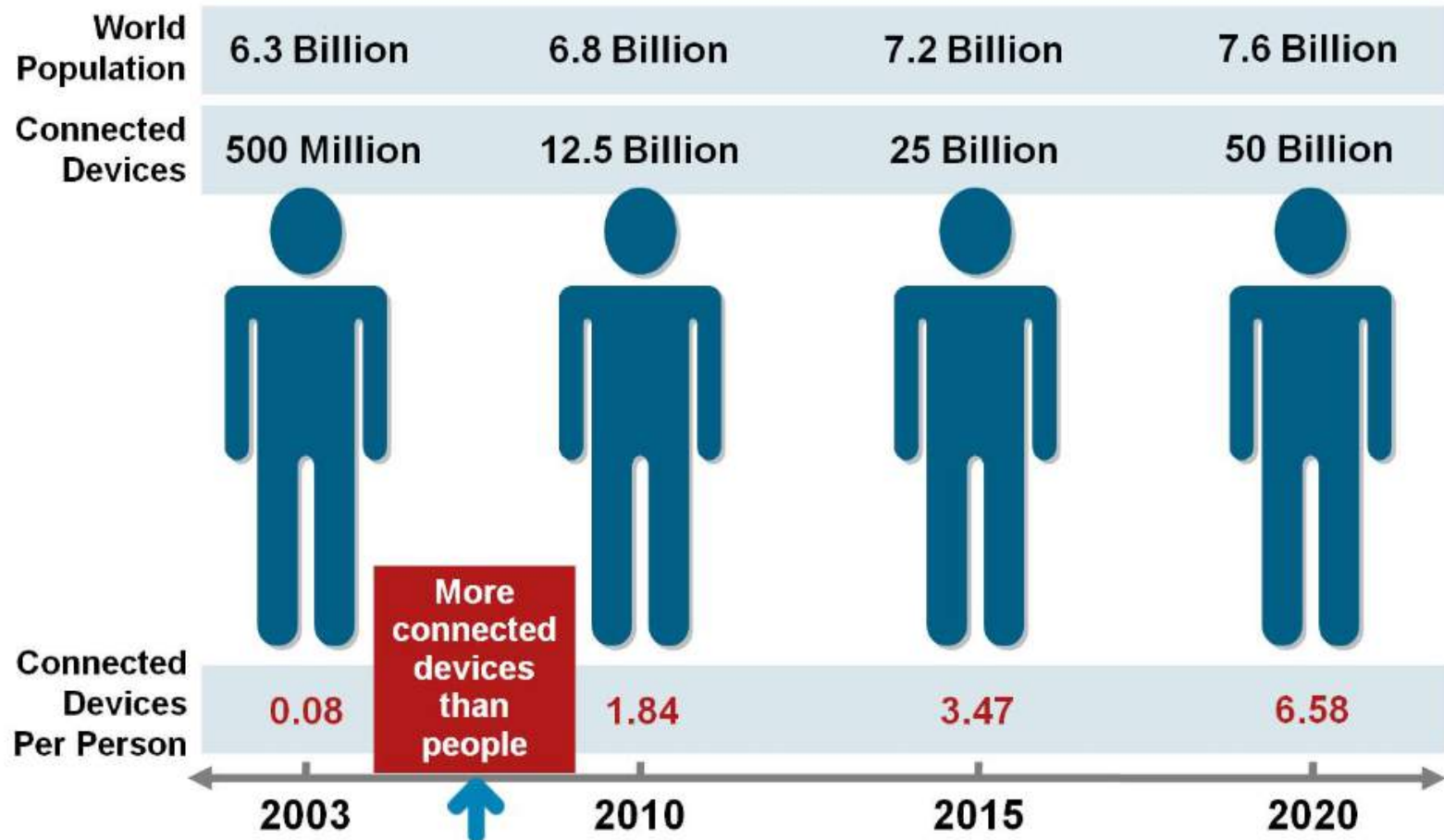


NOTE: Numbers may not sum due to rounding.

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

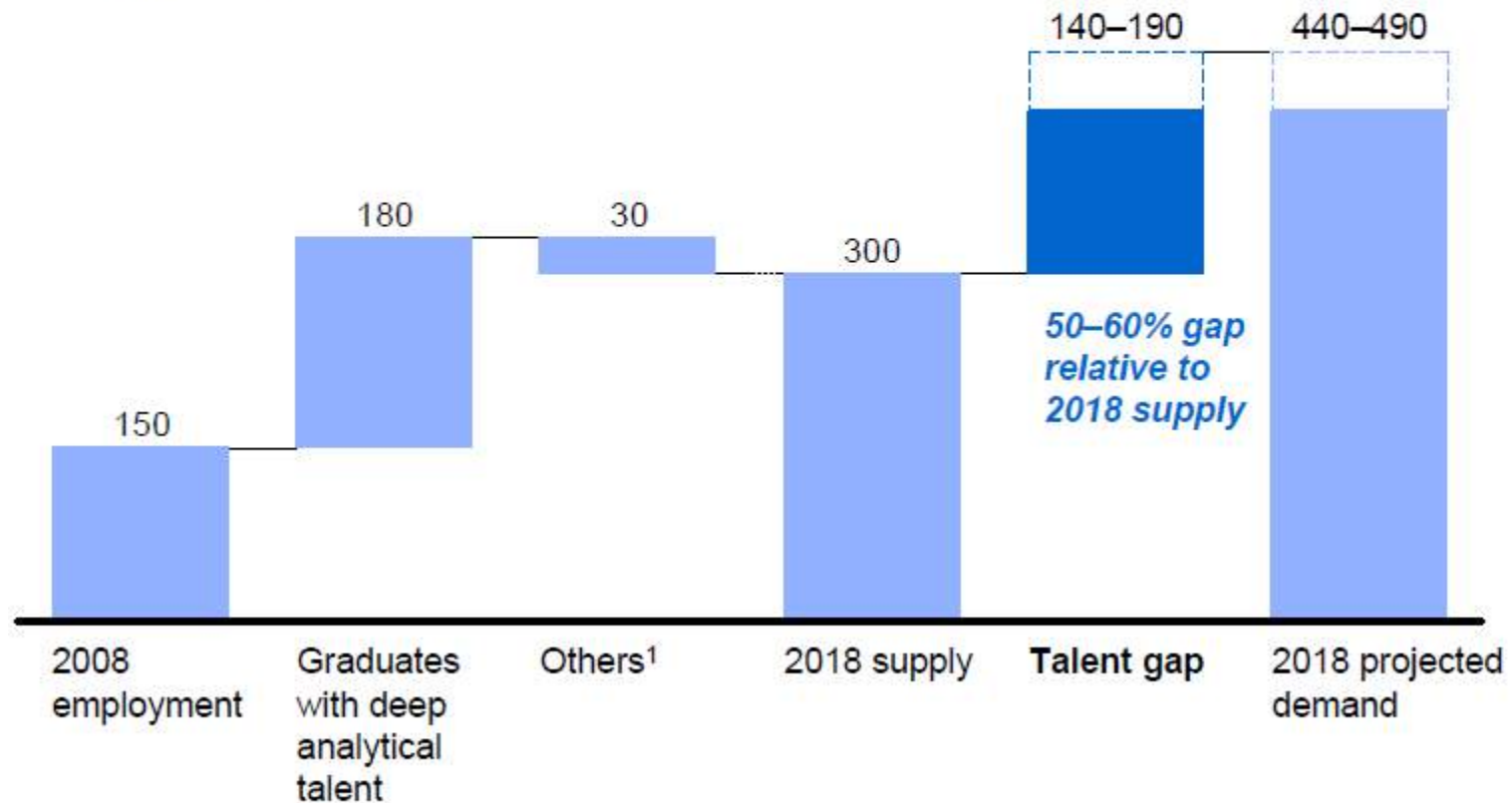


# Predicted lack of talent for Big-Data related technologies

**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



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



# 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%

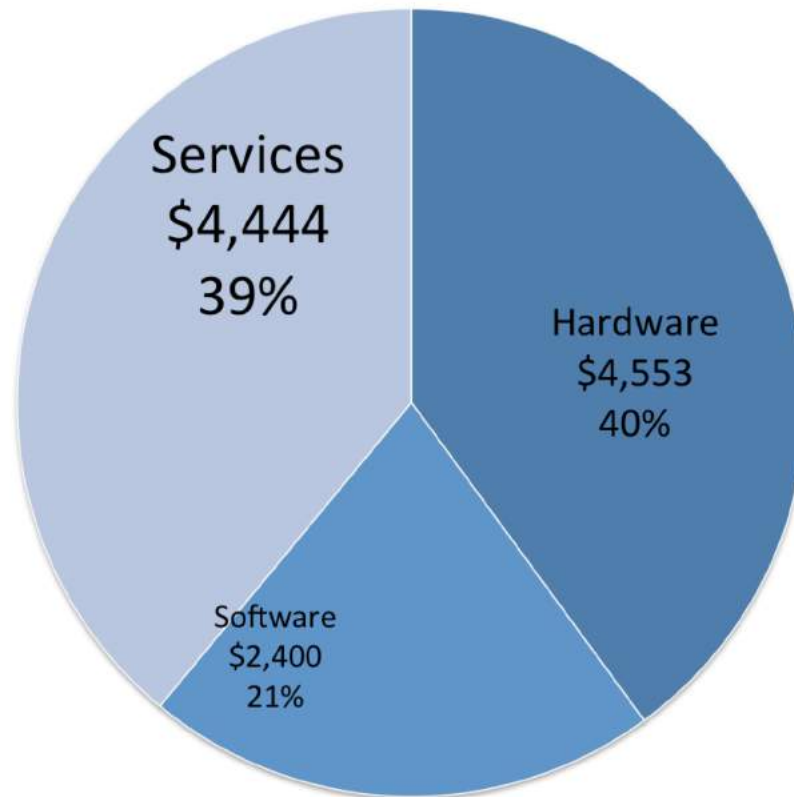
Source: WikiBon report on “Big Data Vendor Revenue and Market Forecast 2012–2017”, 2013

# Big Data Revenue by Type, 2012

([http://wikibon.org/w/images/f/f9/Segment\\_-\\_BDMSVR2012.png](http://wikibon.org/w/images/f/f9/Segment_-_BDMSVR2012.png))



**Big Data Revenue by Type, 2012**  
(in \$US millions)

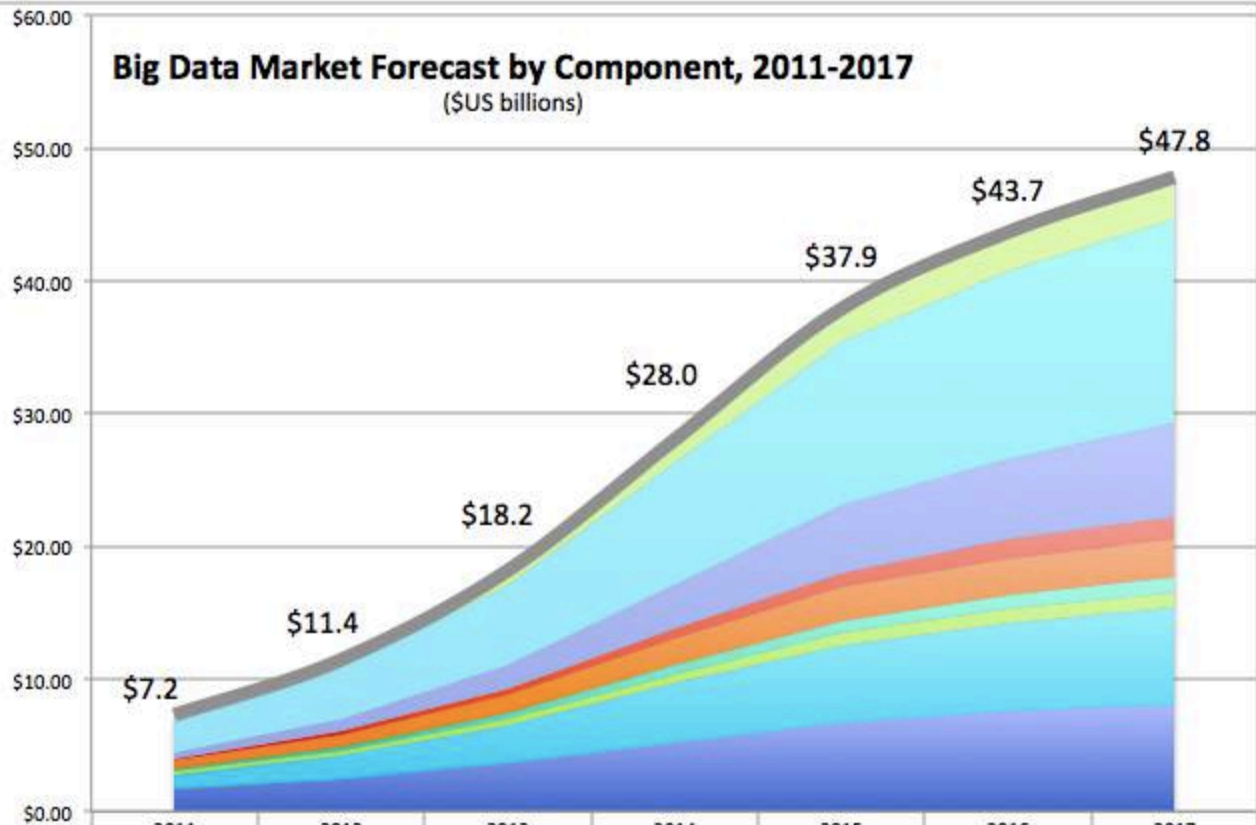


# Big Data Market Forecast (2011–2017)

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



Yearly Revenue (\$US billions)



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



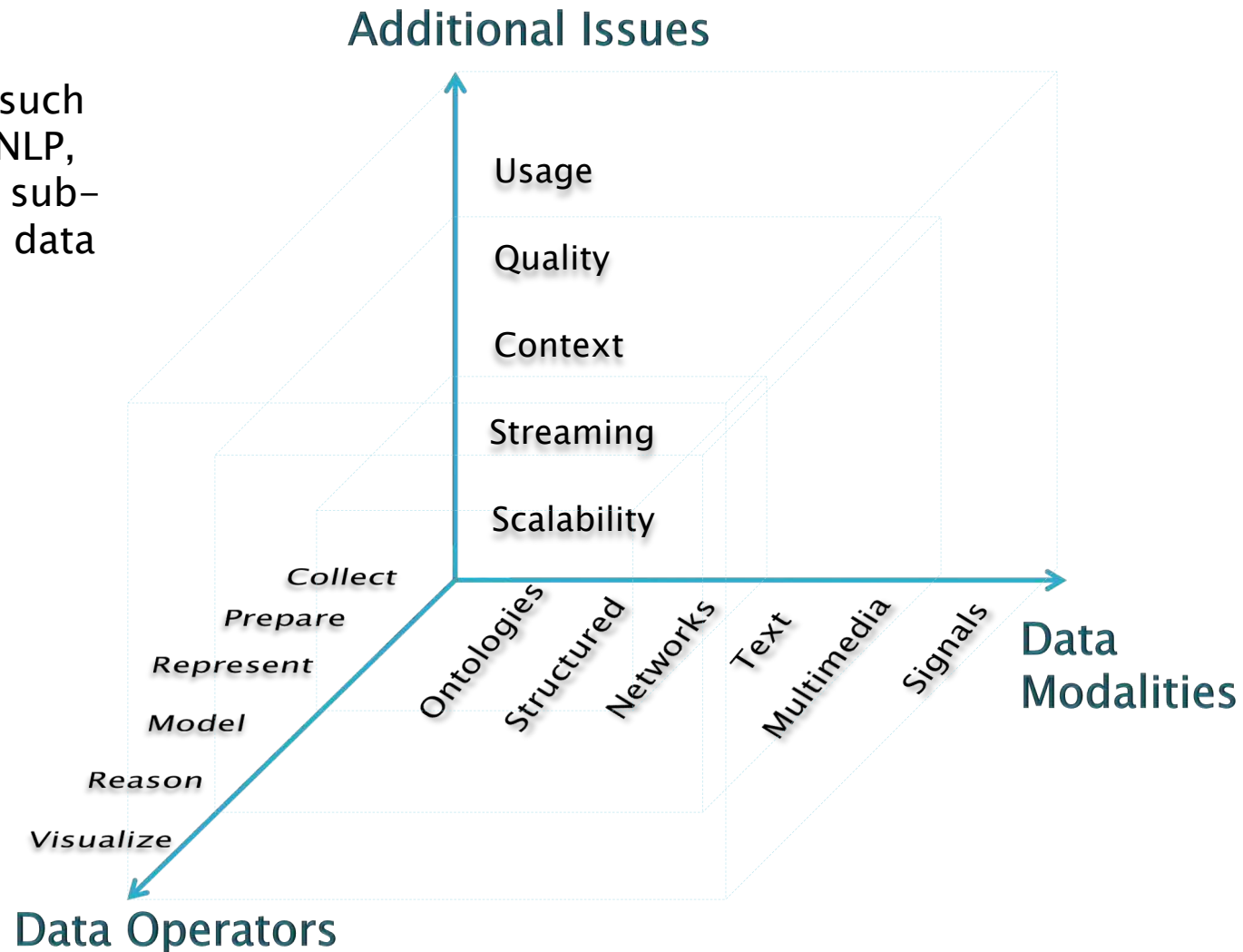
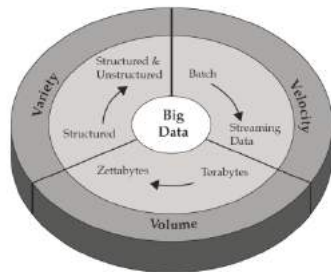
# Techniques

# 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

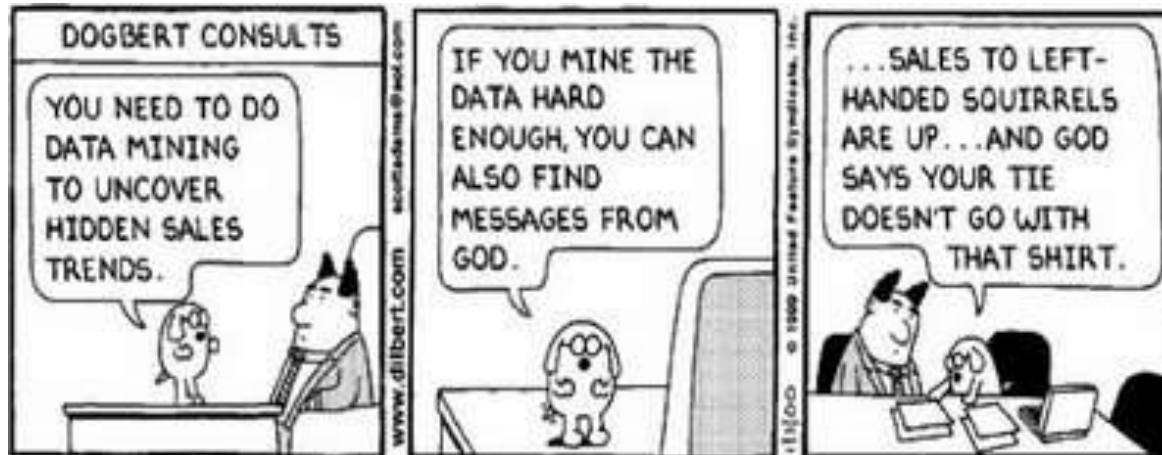
# What matters when dealing with data?

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



# Meaningfulness of Analytic Answers

- ▶ A risk with “Big-Data mining” is that an analyst can “discover” patterns that are meaningless
- ▶ Statisticians call it **Bonferroni's principle**:
  - Roughly, as the amount of data grows, you may find events that are a statistical artifact and not a true instance of what you are looking for





# Meaningfulness of Analytic Answers

- ▶ Suppose you have a certain amount of data, and you look for events of a certain type within that data.
- ▶ You can expect events of this type to occur, even if the data is completely random, and the number of occurrences of these events will grow as the size of the data grows.
- ▶ These occurrences are “bogus,” in the sense that they have no cause other than that random data will always have some number of unusual features that look significant but aren’t.

# Meaningfulness of Analytic Answers

- ▶ Calculate the expected number of occurrences of the events you are looking for, on the assumption that data is random.
- ▶ If this number is significantly larger than the number of real instances you hope to find, then you must expect almost anything you find to be bogus, i.e., a statistical artifact rather than evidence of what you are looking for.

# Meaningfulness of Analytic Answers

Example:

- ▶ We want to find terrorists: (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - $10^9$  people being tracked.
  - Each person stays in a hotel 1% of the time (1 day out of 100)
  - Hotels hold 100 people (so  $10^9 * 10^{-2} * 10^{-2} = 10^5$  hotels).
  - 1000 days.
  - If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?

# Meaningfulness of Analytic Answers

- ▶ Suppose, however, that there really are no evil-doers.
- ▶ That is, everyone behaves at random, deciding with probability 0.01 to visit a hotel on any given day, and if so, choosing one of the  $10^5$  hotels at random.
- ▶ Would we find any pairs of people who appear to be evil-doers?



# Meaningfulness of Analytic Answers

- ▶ The probability of any two people both deciding to visit a hotel on any given day is .0001.
- ▶ The chance that they will visit the same hotel is this probability divided by  $10^5$
- ▶ Thus, the chance that they will visit the same hotel on one given day is  $10^{-9}$
- ▶ The chance that they will visit the same hotel on two different given days is the square of this number,  $10^{-18}$

# Meaningfulness of Analytic Answers

- ▶ Now, we must consider how many events will indicate evil-doing. An “event” in this sense is a pair of people and a pair of days, such that the two people were at the same hotel on each of the two days.
- ▶ Note that for large  $n$ ,  $\binom{n}{2}$  is about  $n^2/2$ .
- ▶ The number of pairs of people is  $\binom{10^9}{2} = 5 \times 10^{17}$
- ▶ The number of pairs of days is  $\binom{1000}{2} = 5 \times 10^5$
- ▶ The expected number of events that look like evil-doing is  $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250,000$

# Meaningfulness of Analytic Answers

- ▶ That is, there will be a quarter of a million pairs of people who look like evildoers, even though they are not.
- ▶ Now, suppose there really are 10 pairs of evil-doers out there.
- ▶ The police will need to investigate a quarter of a million other pairs in order to find the real evil-doers.
- ▶ In addition to the intrusion on the lives of half a million innocent people, the work involved is sufficiently great that this approach to finding evil-doers is probably not feasible.

# 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

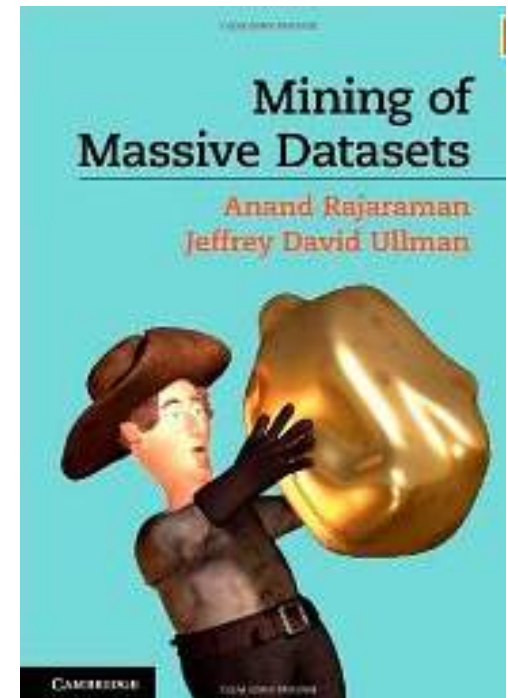


# ...guide to Big-Data algorithms

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

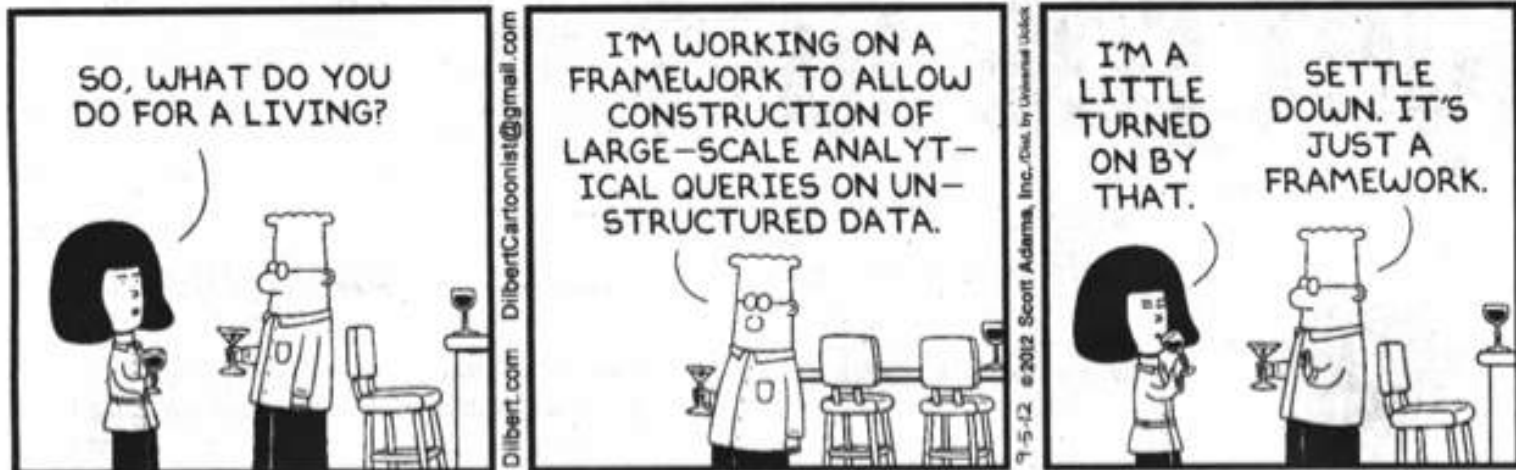
- ▶ Downloadable from:

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



# Tools

## DILBERT



# 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

## Data Analysis & Platforms



## Databases / Data warehousing



## Operational



## Multivalued database



## Big Data search



## Data aggregation



## Business Intelligence



## Data Mining



## Social



## Graphs



## Multidimensional



## KeyValue



## Document Store



## Object databases



## Multimodel



## XML Databases



Created by: www.bigdata-startups.com

<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, 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

# 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 than RDBMSs
  - **Horizontal Scalability, Running on commodity hardware**
  - **Avoidance of Expensive Object-Relational Mapping** – most NoSQL store simple data structures
  - **Compromising Reliability for Better Performance**

# 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

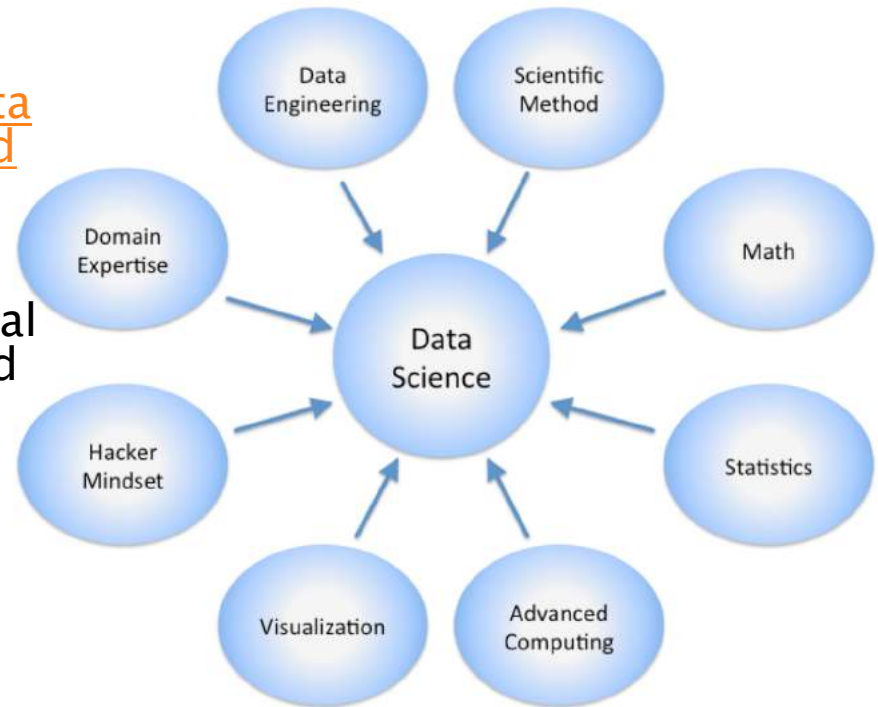
# Data Science

## Life as an Analyst



# 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.





# Applications

- ▶ Recommendation
- ▶ Social Network Analytics

# Application: Recommendation of News Articles

# 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](http://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

f Recommend

48

Tweet

27

in 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](mailto:elopatto@bloomberg.net).

To contact the editor responsible for this story: Reg Gale at [rgale5@bloomberg.net](mailto: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/](http://sws.geonames.org/1880252/))
- Ames ([sws.geonames.org/3037869/](http://sws.geonames.org/3037869/))

## People:

- Douglas A. Gentile

## Organizations:

- Iowa State University ([dbpediapa.org/resource/iowa\\_state\\_university](http://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

[f Recommend](#) 48 [Tweet](#) 27 [in 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](mailto:elopatto@bloomberg.net).

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

Gender	<input checked="" type="radio"/> Male <input type="radio"/> Female
Year of Birth	<input type="text" value="1965"/>
Zip Code	<input type="text" value="10017"/>
Country of Residence	<input type="text" value="United States"/> ▼
Household Income	<input type="text" value="\$100,000 to \$149,999"/> ▼
Job Industry	<input type="text" value="Accounting"/> ▼
Job Title	<input type="text" value="Accountant/Auditor"/> ▼
Company Size	<input type="text" value="--- Select One ---"/> ▼

# 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

MOST E-MAILED		RECOMMENDED FOR YOU	
	1.	GAIL COLLINS	<a href="#">Small Is So Beautiful</a>
	2.	EDITORIAL	<a href="#">The Need to Agree to Agree</a>
	3.	<a href="#">Parties' Tactics Eroding Unity Left and Right</a>	
	4.	PAUL KRUGMAN	<a href="#">Policy and the Personal</a>
	5.	<a href="#">A Reality Series Finds Silicon Valley Cringing</a>	
	6.	YOU'RE THE BOSS	<a href="#">This Week in Small Business: Managing Millennials</a>
	7.	CHARLES M. BLOW	<a href="#">What a Tangled Web</a>
	8.	BITS	<a href="#">Google and F.T.C. Set to Settle Safari Privacy Charge</a>

# 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 severe 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.

# 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

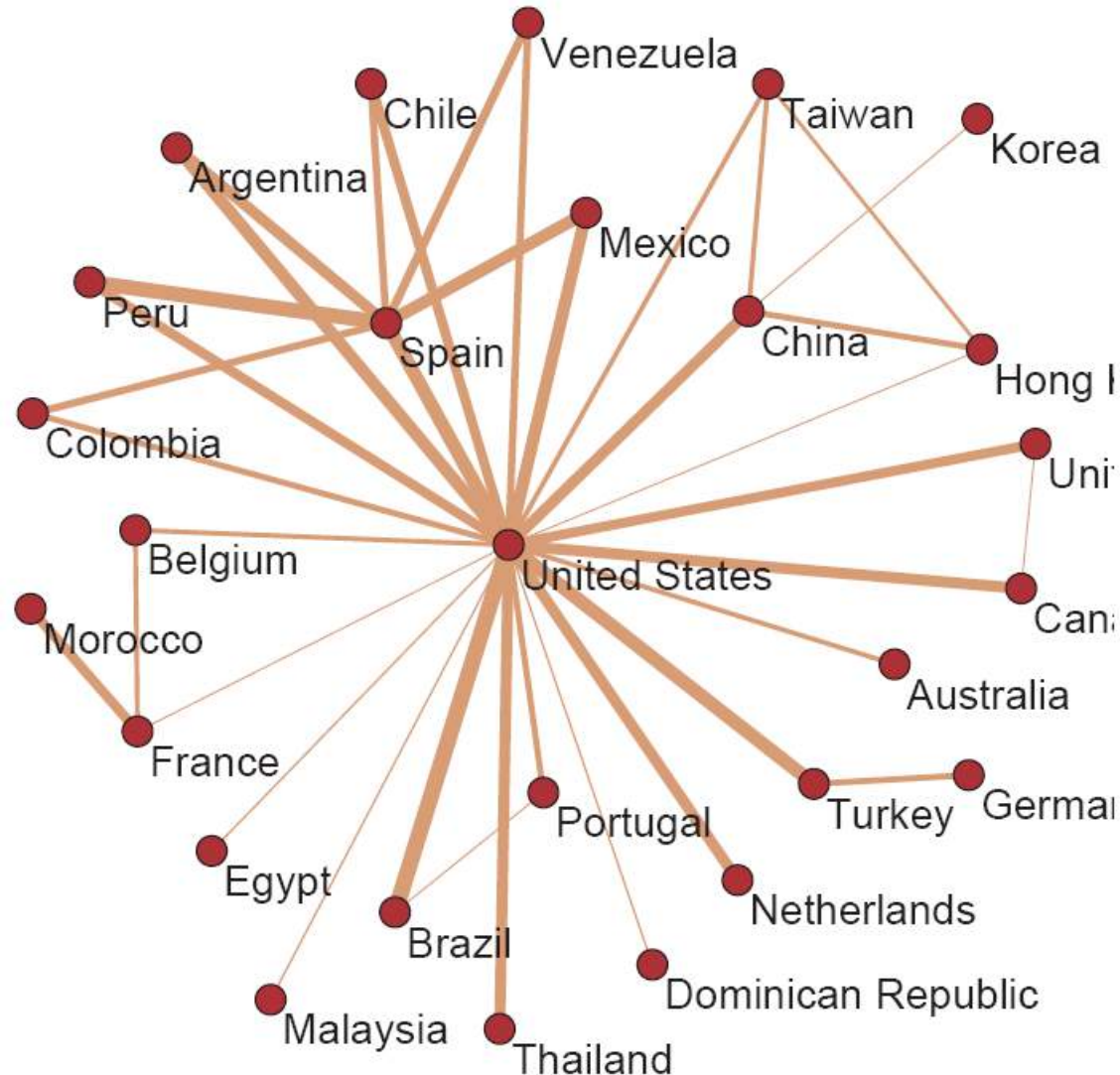
## Research questions:

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

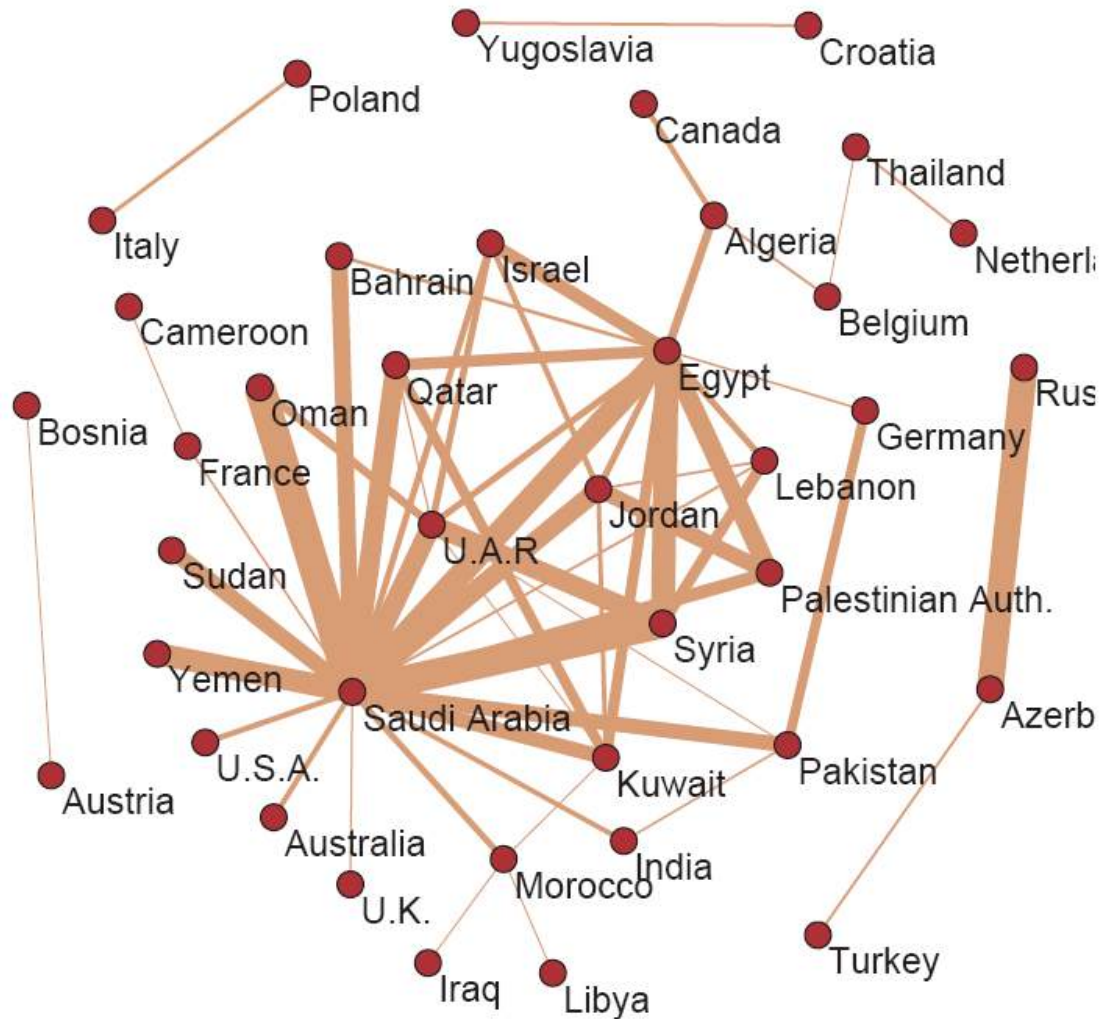
# Data statistics: Total activity

- ▶ Data collected 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

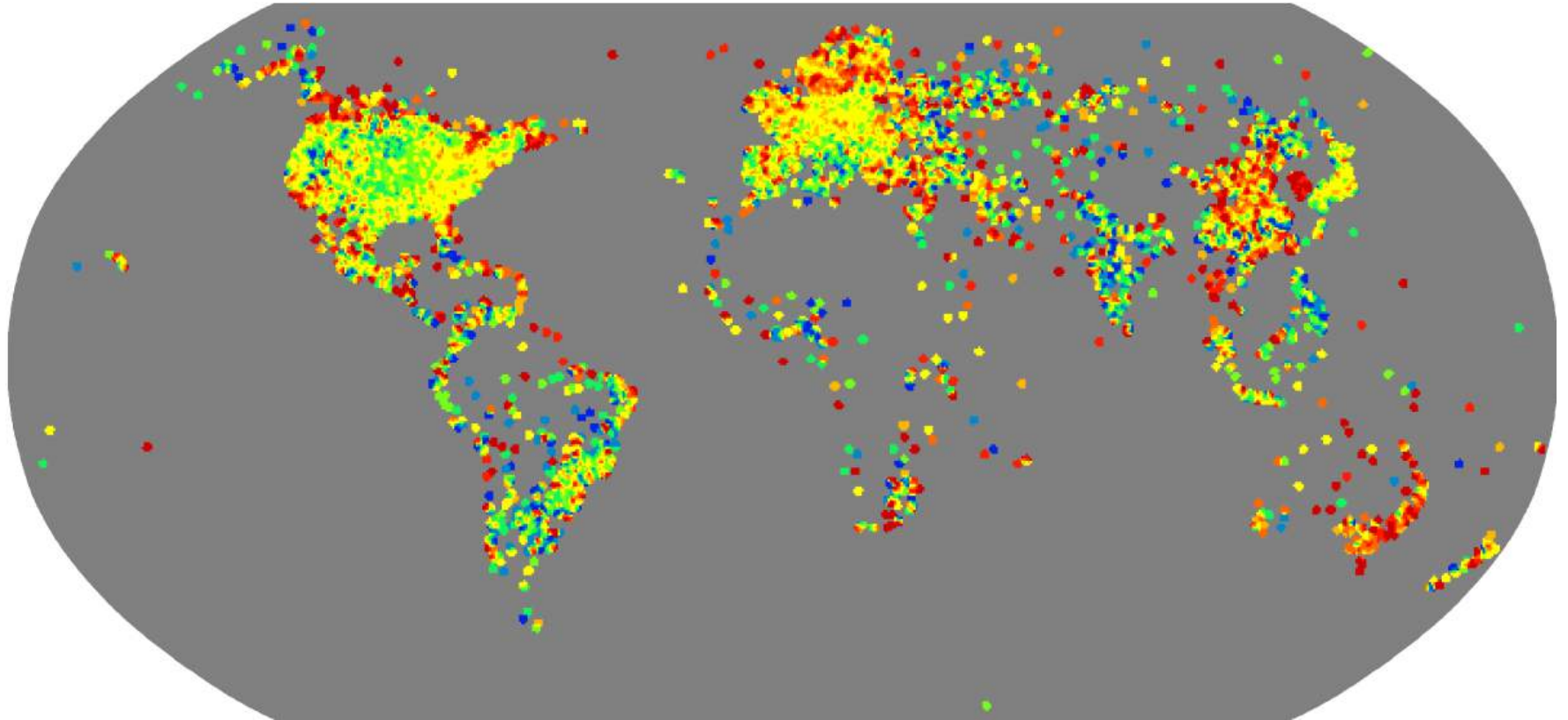
# Who talks to whom: Number of conversations



# Who talks to whom: Conversation duration

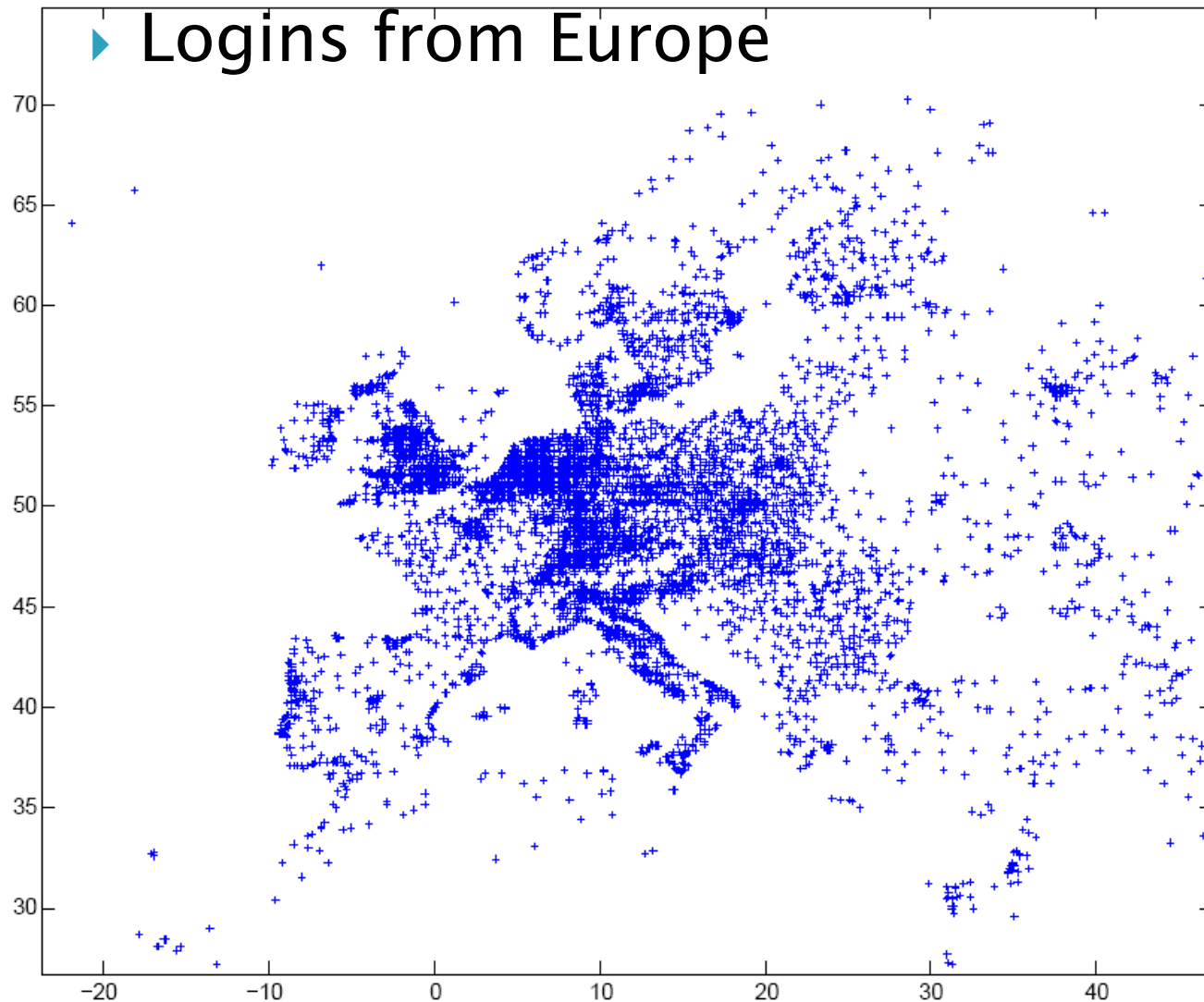


# Geography and communication



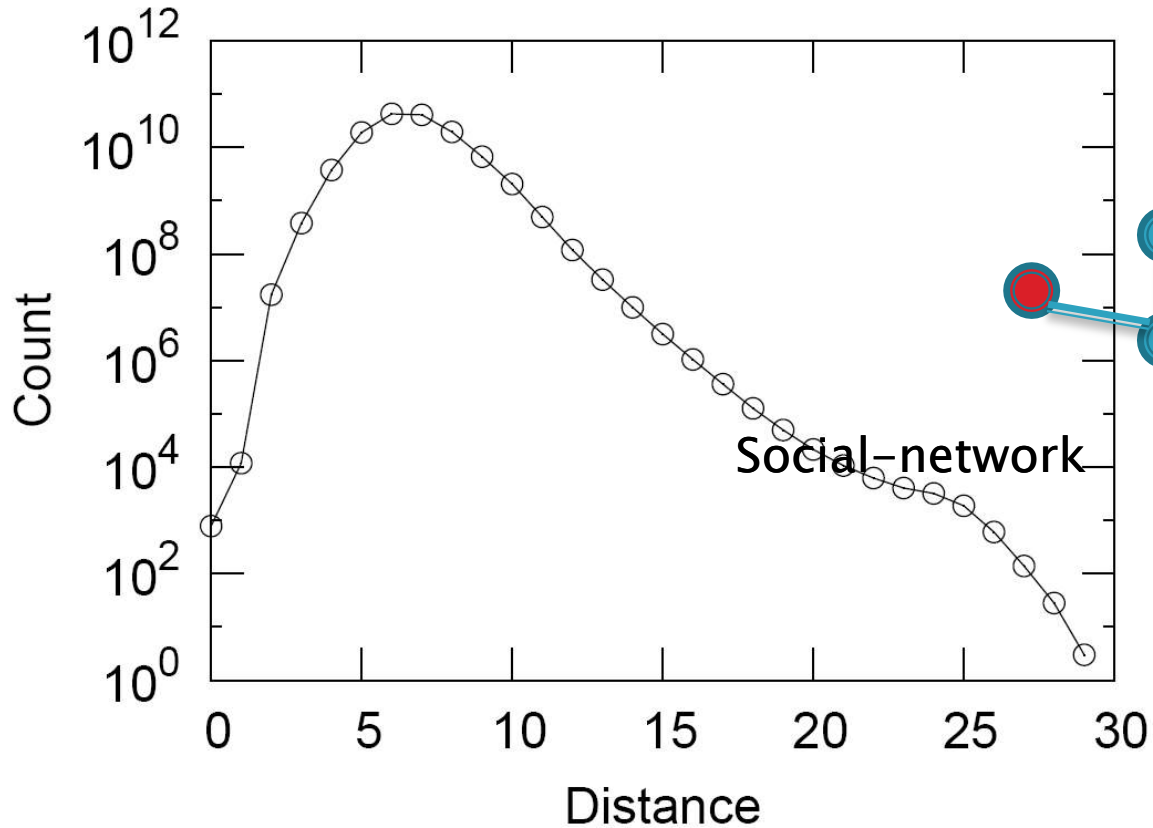
- ▶ Count the number of users logging in from particular location on the earth

# How is Europe talking





# Network: Small-world



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

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

# ...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)