Big-Data Tutorial

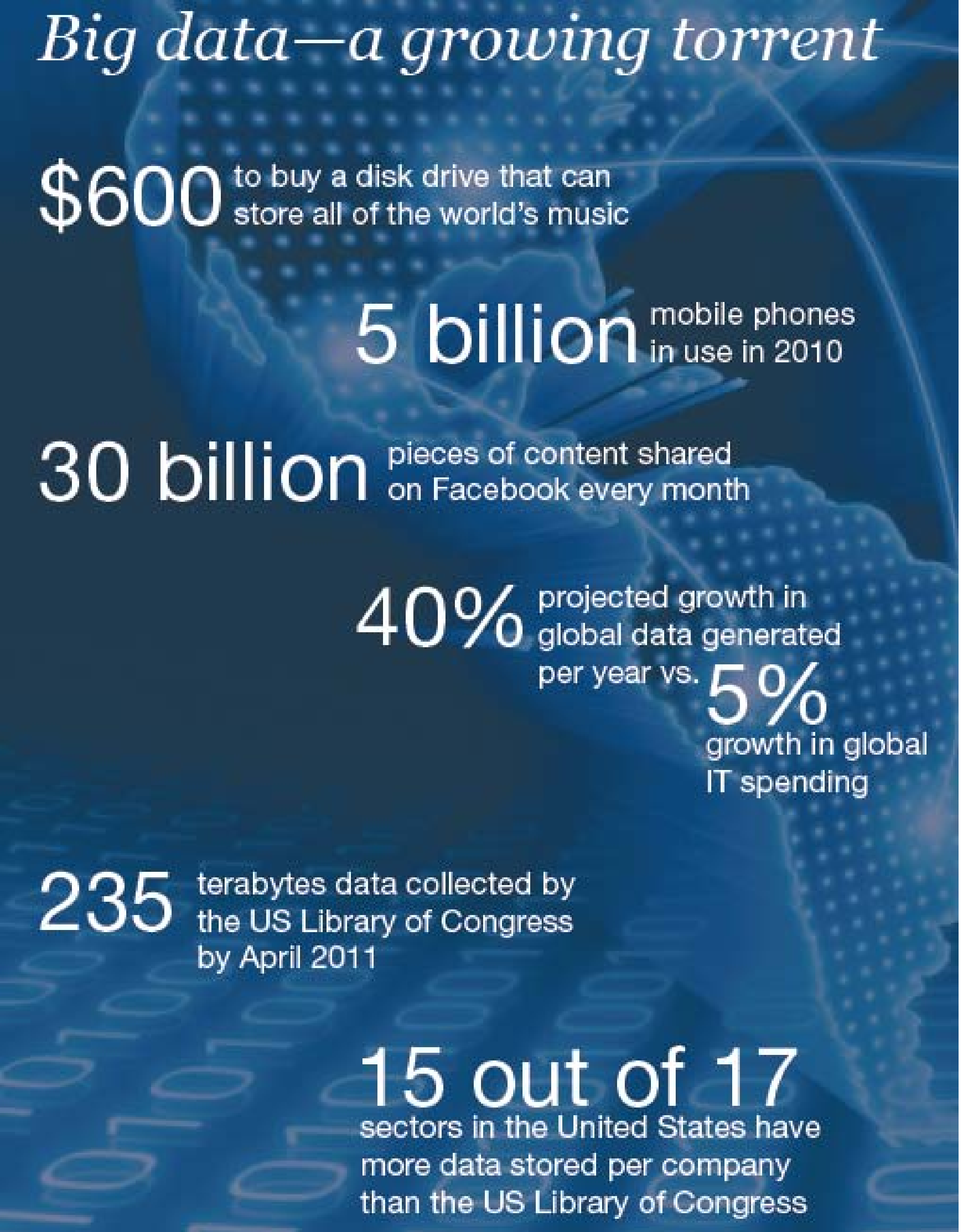
Marko Grobelnik
marko.grobelnik@ijs.si
Jozef Stefan Institute
Ljubljana, Slovenia

Stavanger, May 8th 2012

Outline

- Introduction
  - What is Big data?
  - Why Big-Data?
  - When Big-Data is really a problem?
- Techniques
- Tools
- Applications
- Literature

Big data—a growing torrent



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

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

Big data—capturing its value

$300 billion

potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion

potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion

potential annual consumer surplus from using personal location data globally

60%

potential increase in retailers' operating margins possible with big data

140,000–190,000

more deep analytical talent positions, and

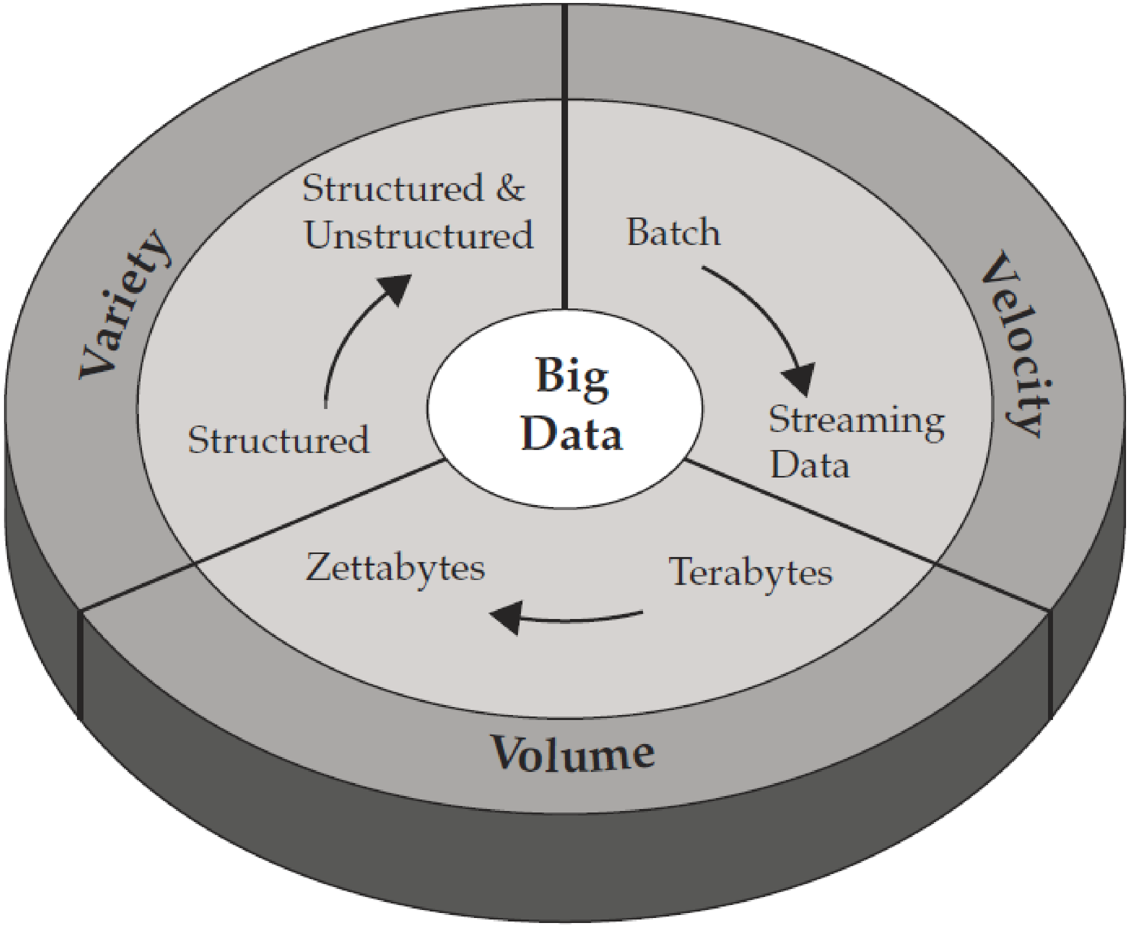
1.5 million

more data-savvy managers needed to take full advantage of big data in the United States

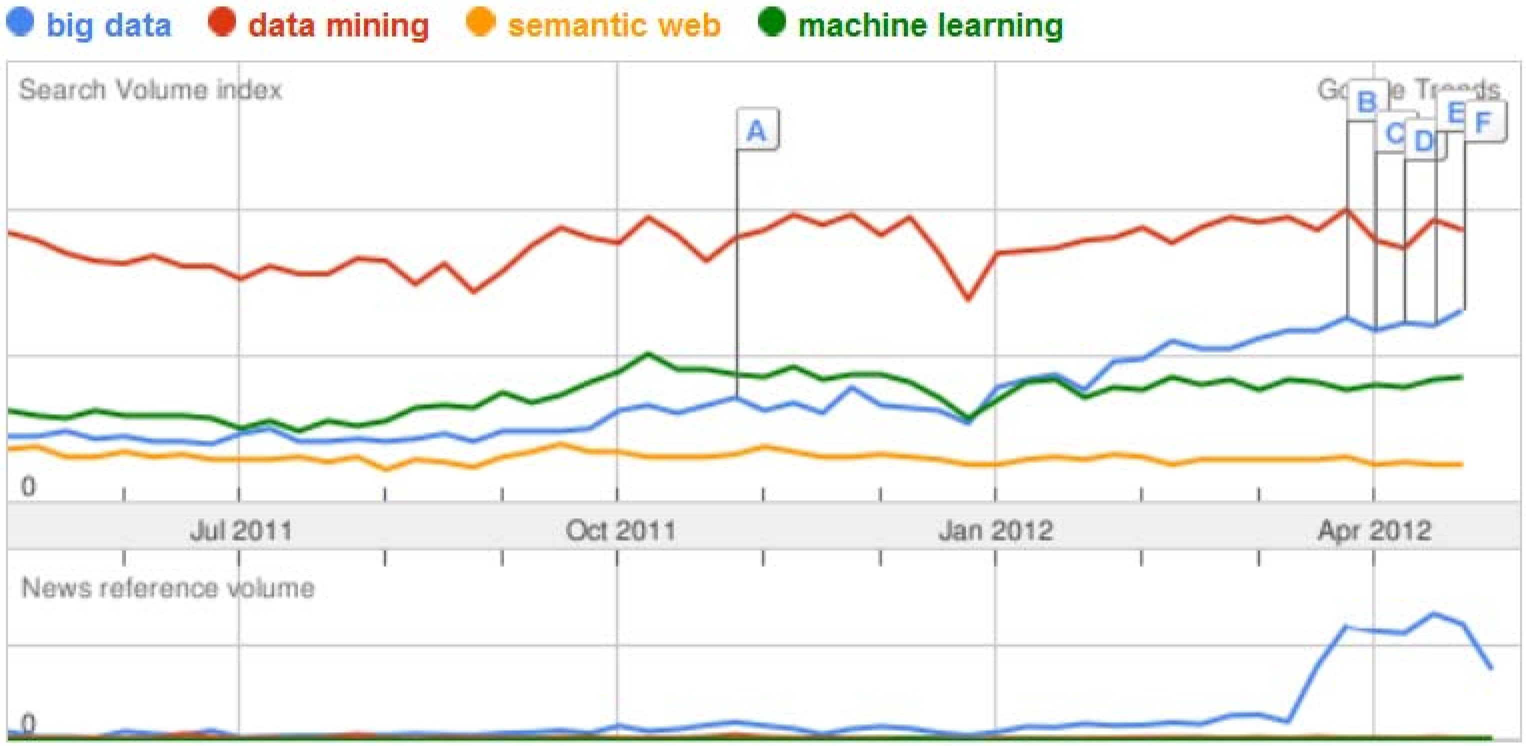
What is Big-Data?

- ‘Big-data’ is similar to ‘Small-data’, but bigger
- …but having data bigger consequently requires different approaches:
  - techniques, tools & architectures
- …to solve:
  - New problems…
  - …and old problems in a better way.

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



Big-Data popularity on the Web



A [Spectra Logic Delivers ExaScale Storage for 'Big Data'; Announces Series of Products and Advancements and Unveils World's Highest Capacity Storage System](#)
MarketWatch - Nov 1 2011

B [Webcast: Obama Goes Big on Big Data](#)
Wired News - Mar 27 2012

C [Cisco Joins Forces with EMC to Advance IT Skills in Cloud, Big Data and Data Center Technologies](#)
Justmeans - Apr 3 2012

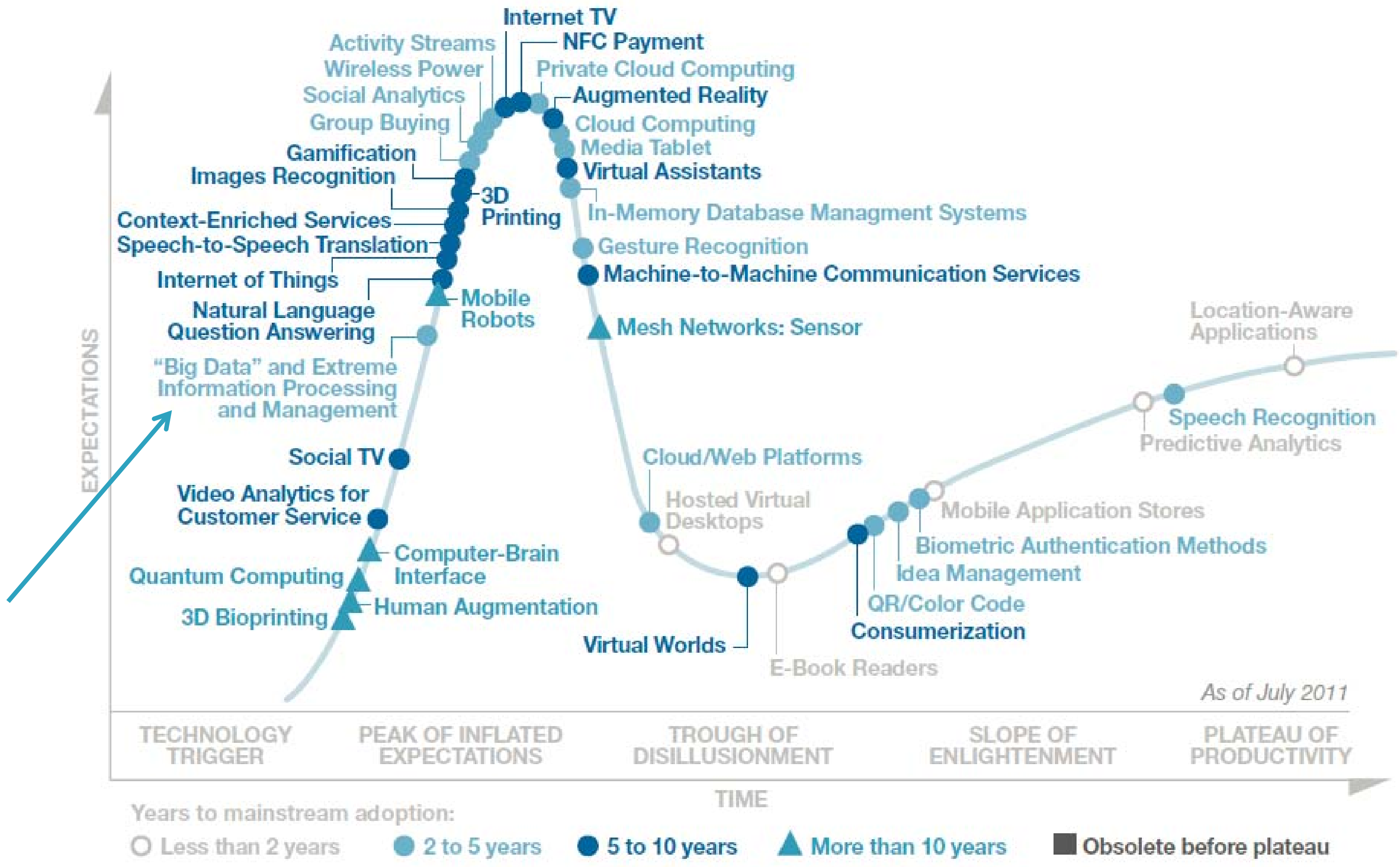
D [Ferranti Unveils its MECOMS™ “Big Data” Strategy for Utility Meter Data Management and Real Time Billing](#)
Victoria Times Colonist - Apr 10 2012

E [Deconstructing Big Data - BuildZoom Launches an Article Series that Reveals the Hype and Substance Behind Big Data](#)
Houston Chronicle - Apr 17 2012

F [Harvard Releases Big Data for Books](#)
New York Times - Apr 24 2012

Big-Data in Gartner Hype-Cycle 2011

Hype Cycle for Emerging Technologies, 2011



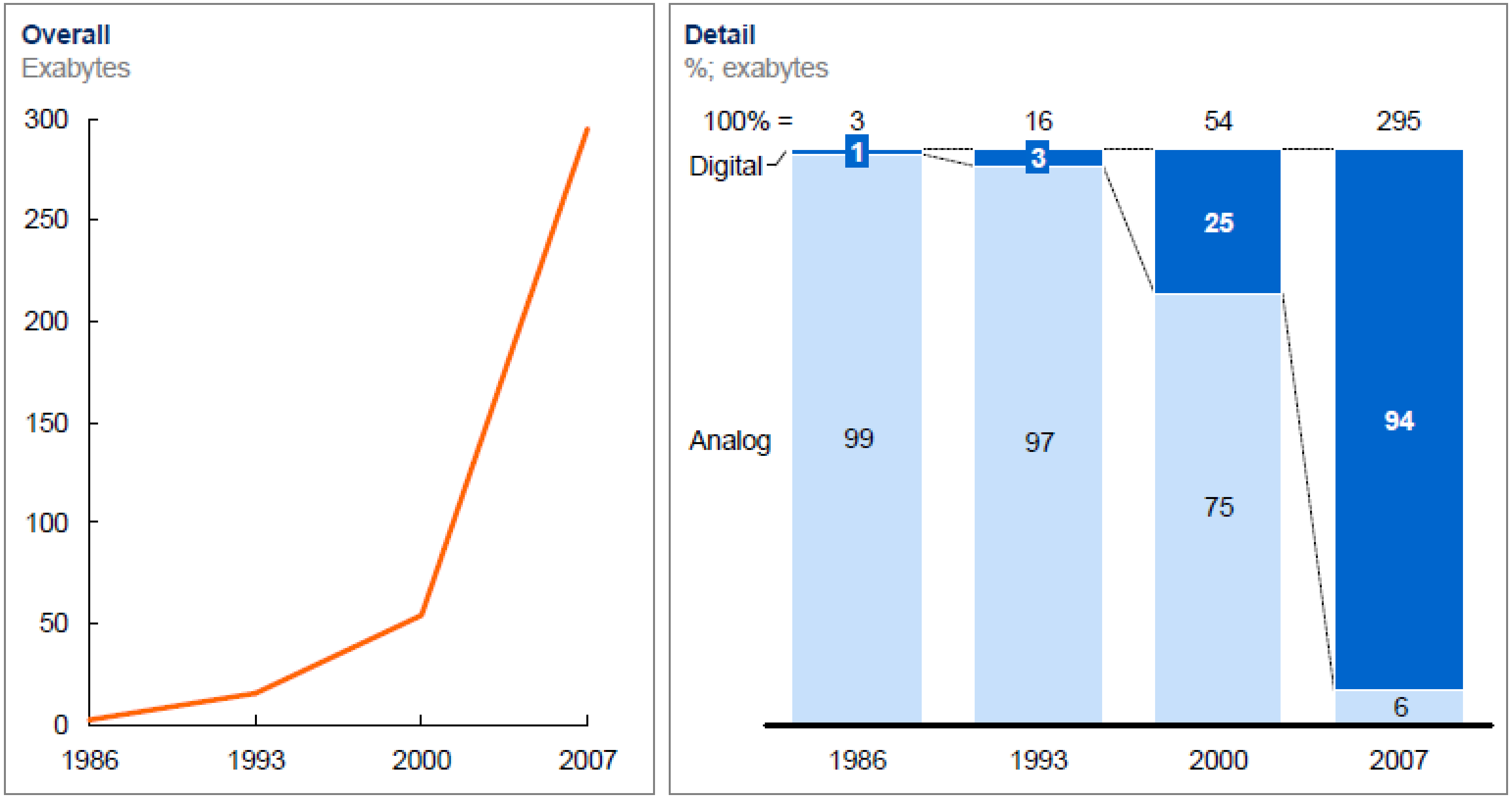
Why Big-Data?

- Key enablers for the 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



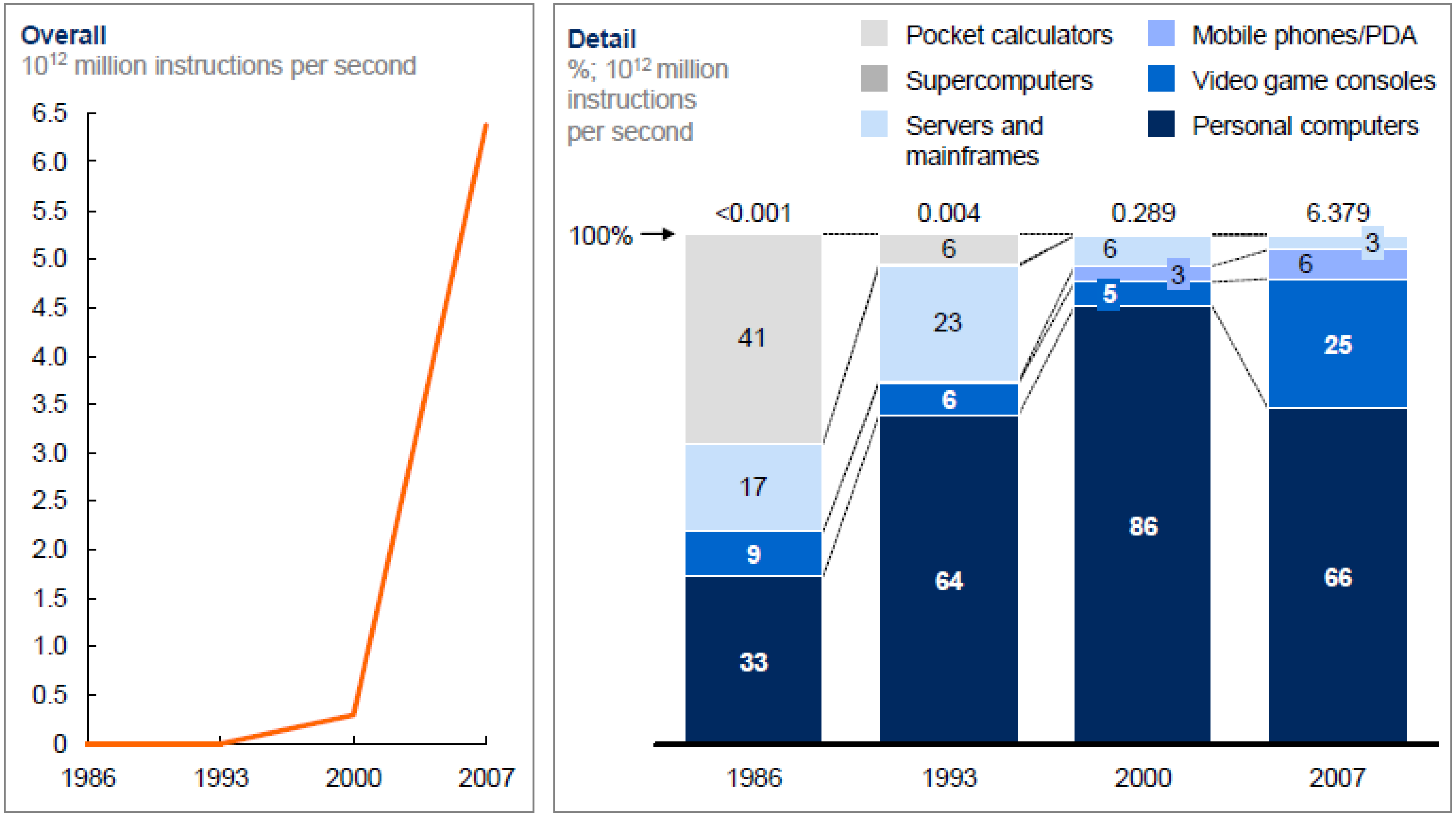
NOTE: Numbers may not sum due to rounding.

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

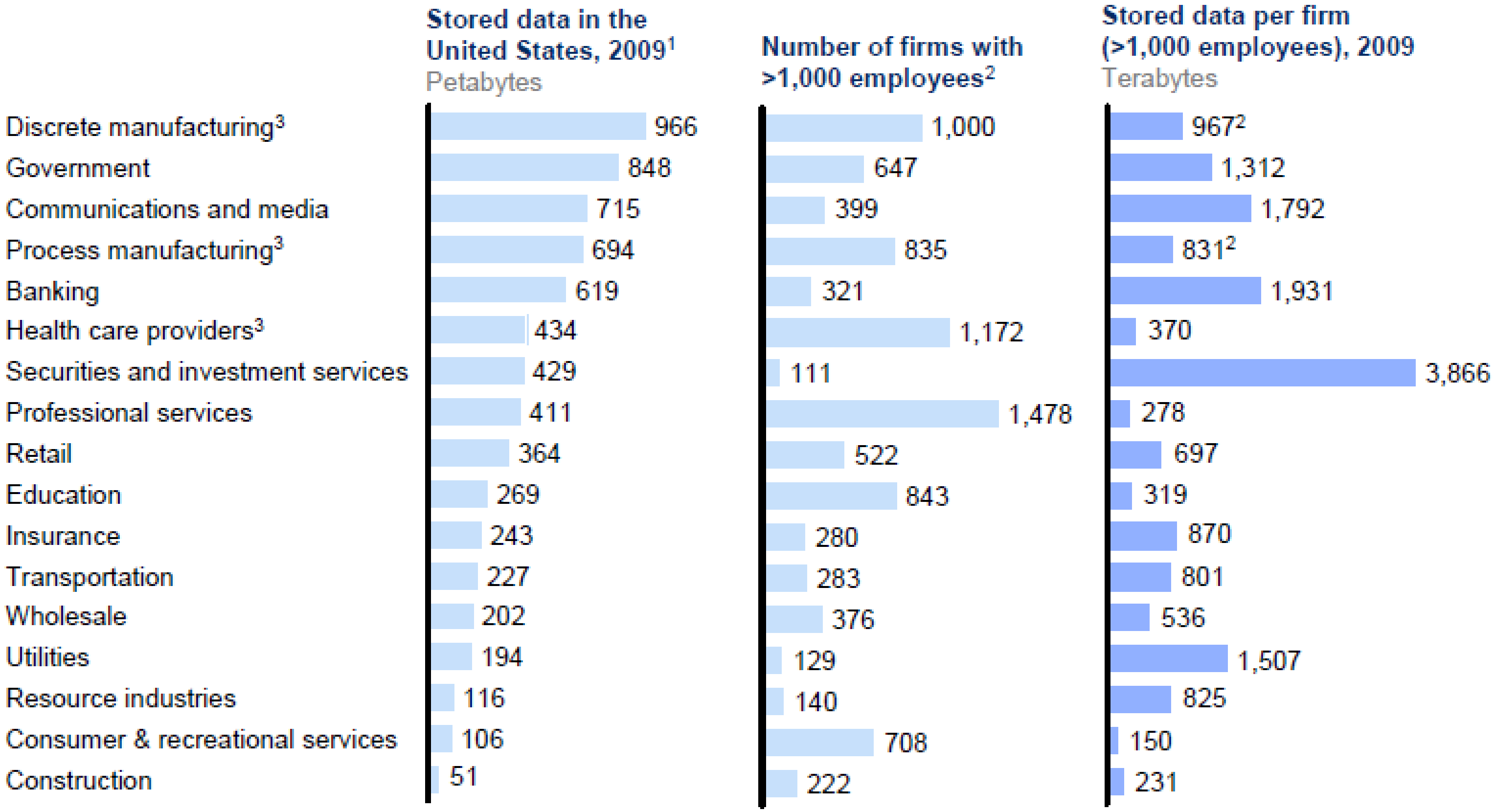


NOTE: Numbers may not sum due to rounding.

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.

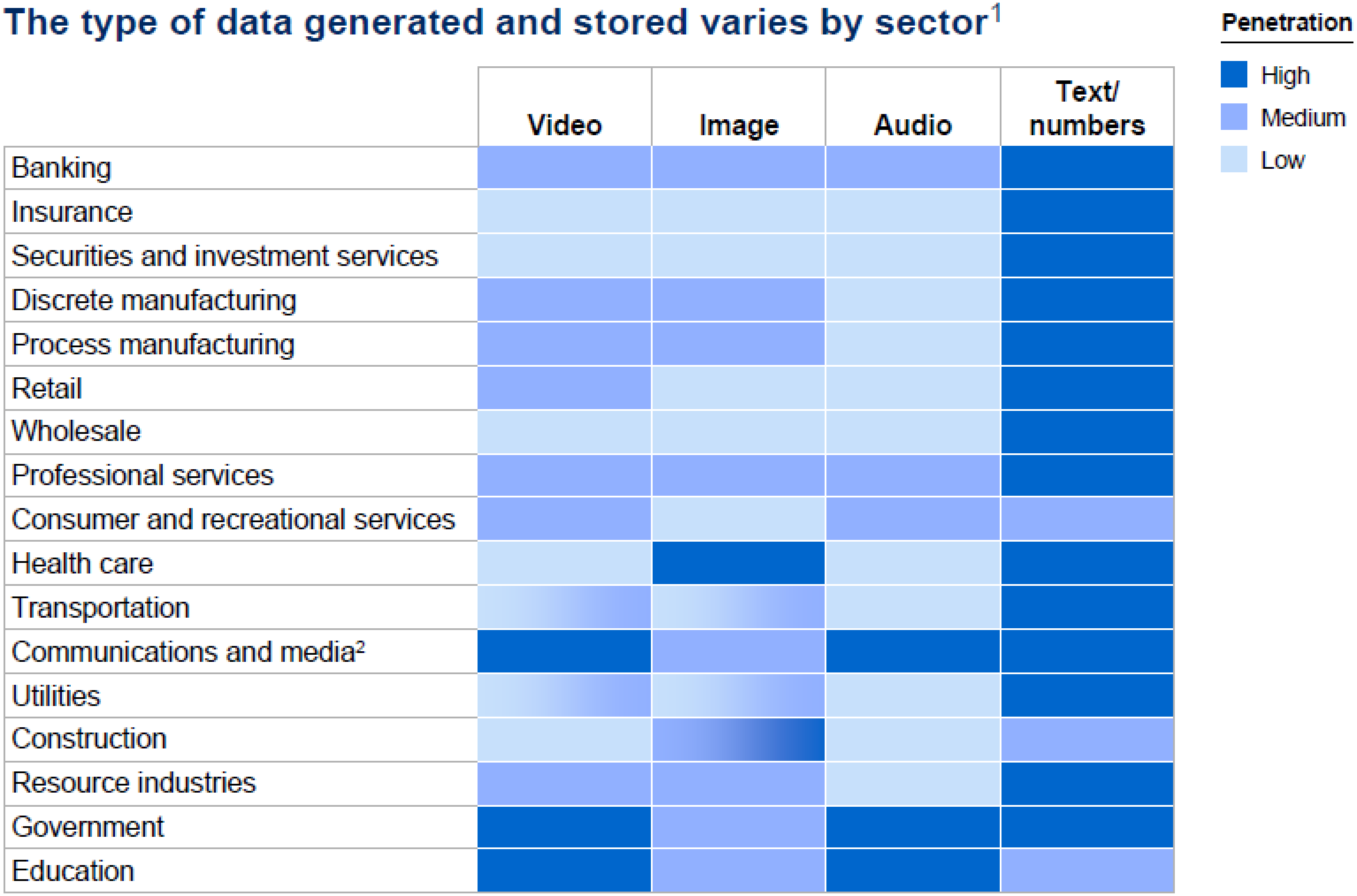
[2] Firm data split into sectors, when needed, using employment

[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	■	▨	■	▨

Penetration

■ High

▨ Medium

□ Low

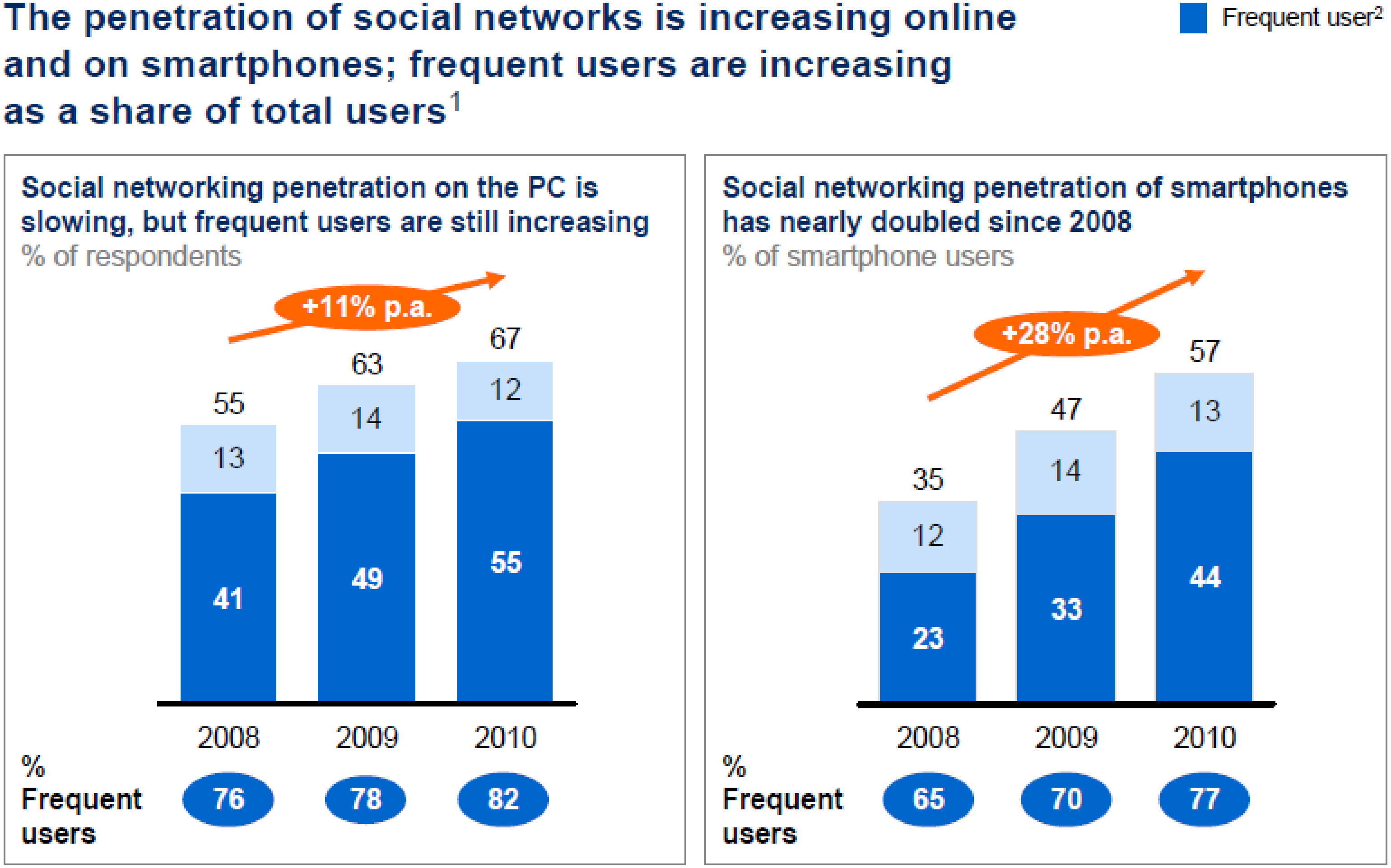
[1] We compiled this heat map using units of data (in files or minutes of video) rather than bytes.

[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]



[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.

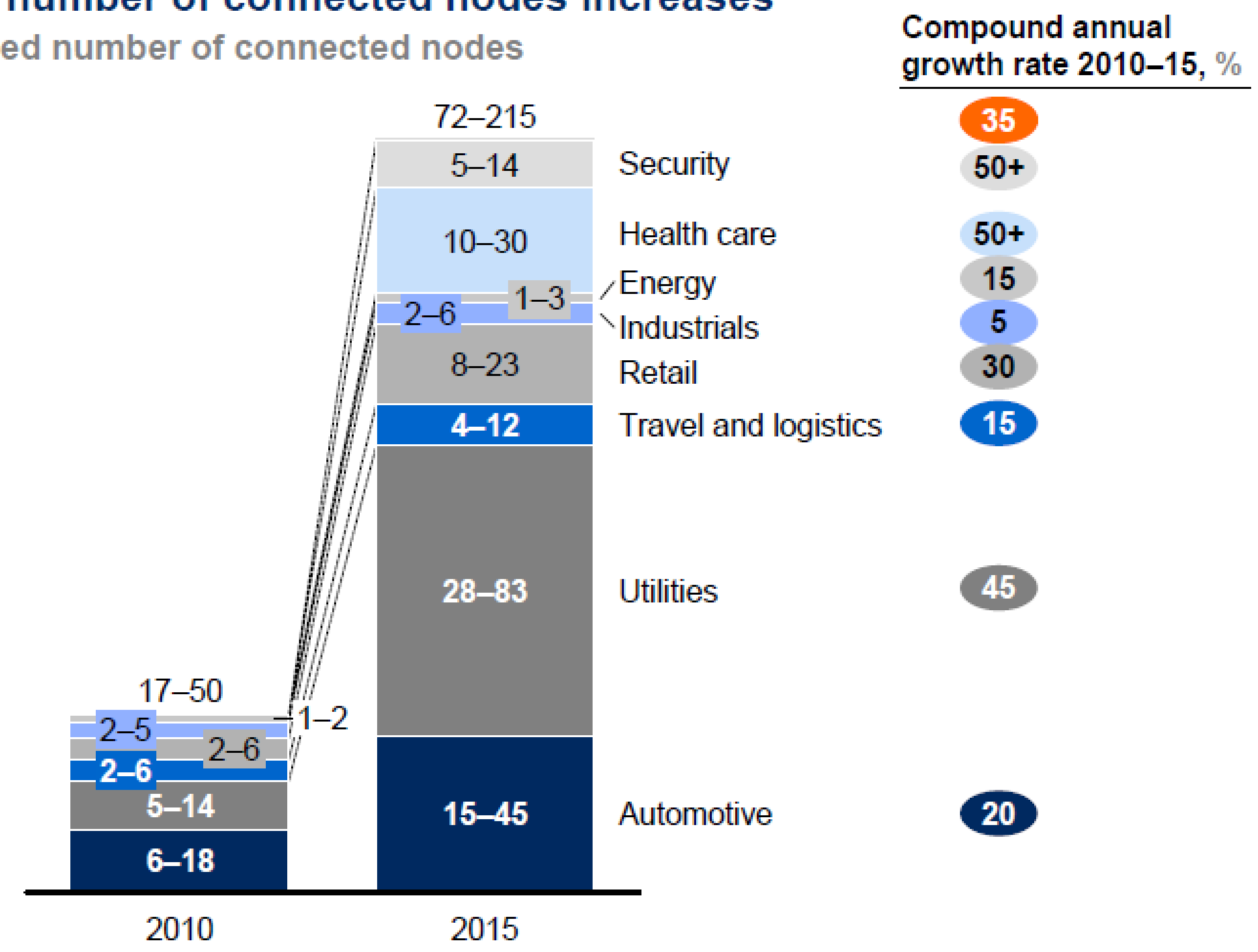
SOURCE: McKinsey iConsumer Survey

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

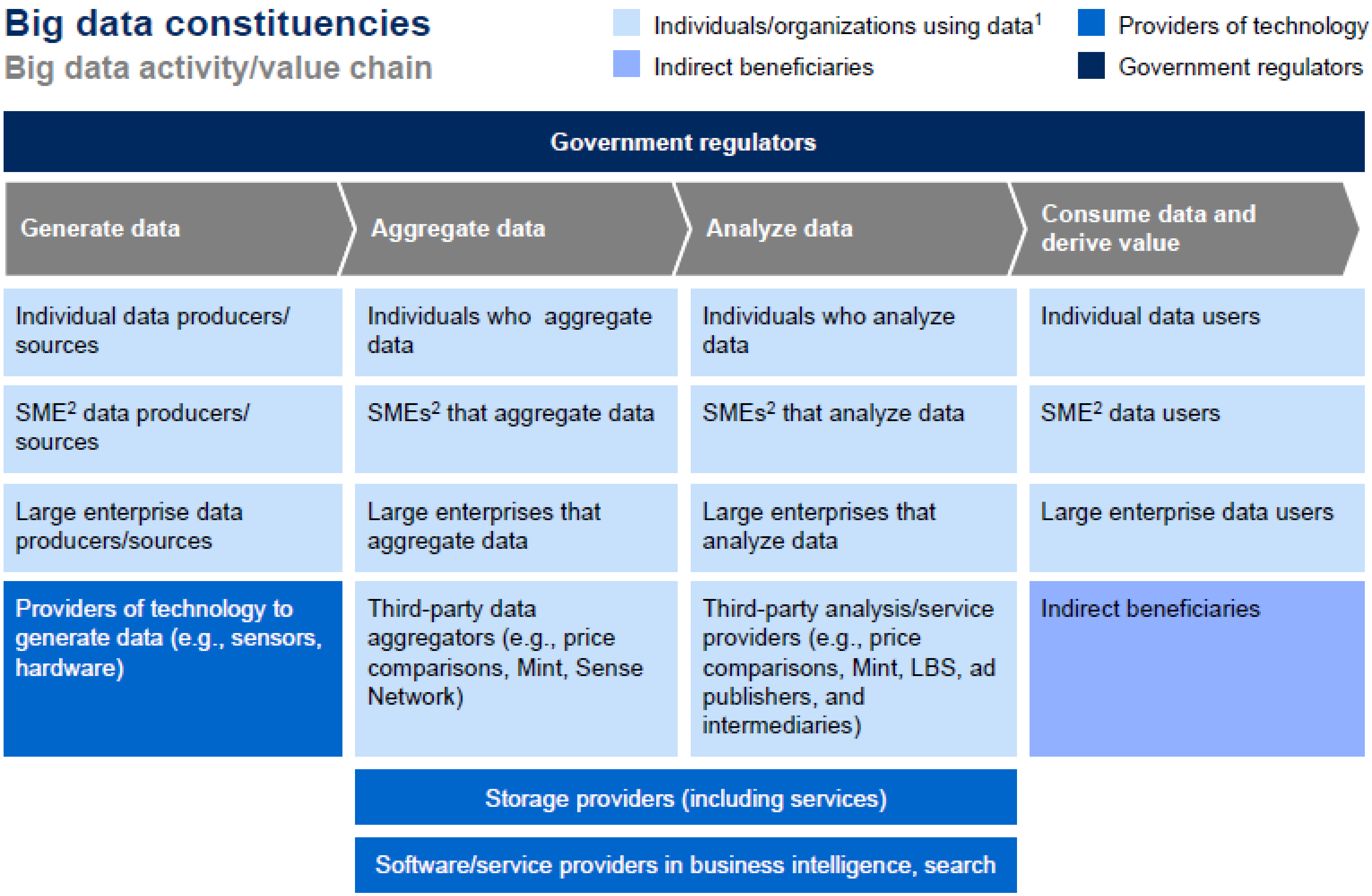
Million



NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

Big-data value chain



Big data constituencies

Big data activity/value chain

- ▢ Individuals/organizations using data[1]
- ▨ Indirect beneficiaries
- ■ Providers of technology
- ■ Government regulators

Government regulators

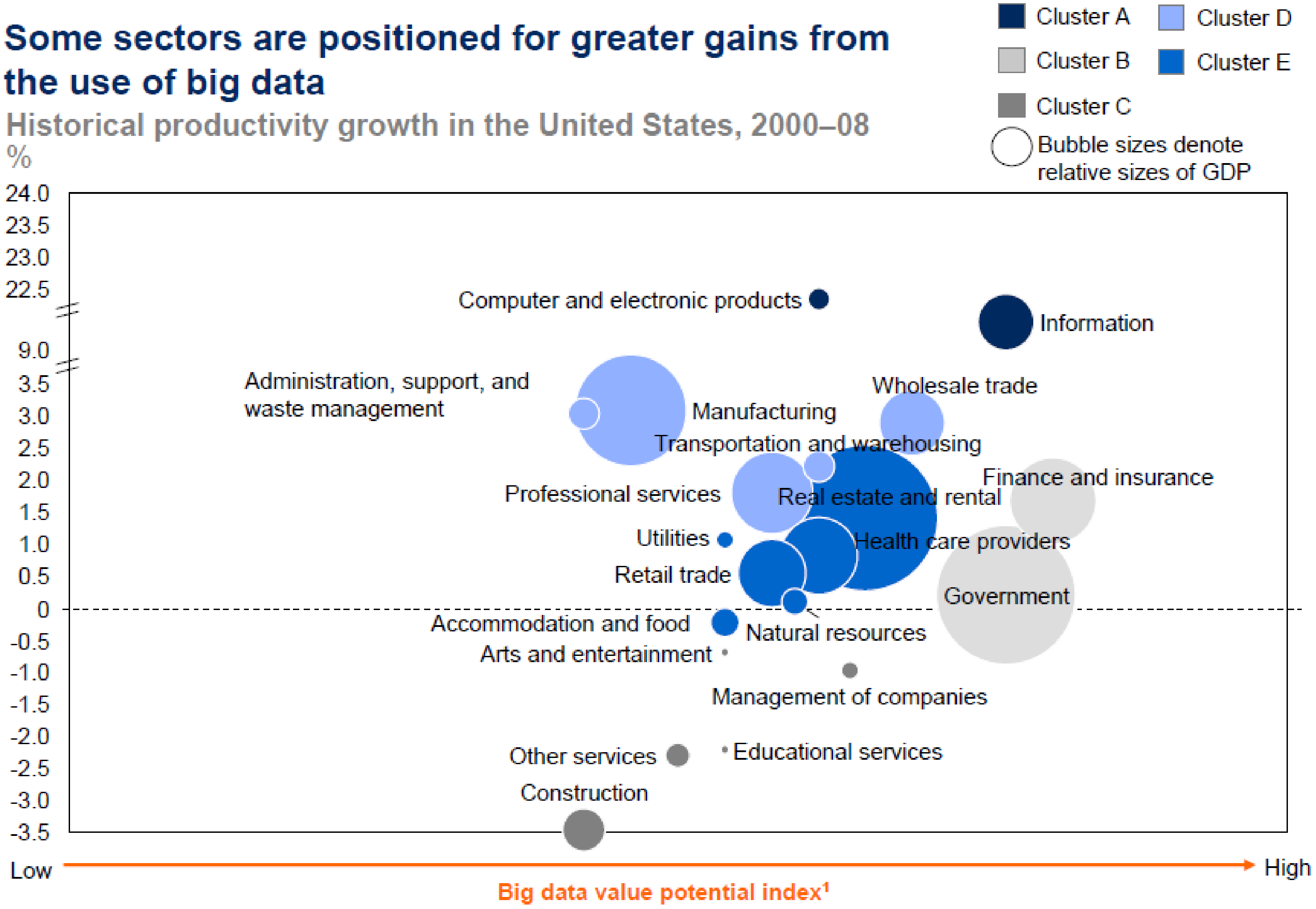
Generate data	Aggregate data	Analyze data	Consume data and derive value
Individual data producers/ sources	Individuals who aggregate data	Individuals who analyze data	Individual data users
SME[2] data producers/ sources	SMEs[2] that aggregate data	SMEs[2] that analyze data	SME[2] data users
Large enterprise data producers/sources	Large enterprises that aggregate data	Large enterprises that analyze data	Large enterprise data users
Providers of technology to generate data (e.g., sensors, hardware)	Third-party data aggregators (e.g., price comparisons, Mint, Sense Network)	Third-party analysis/service providers (e.g., price comparisons, Mint, LBS, ad publishers, and intermediaries)	Indirect beneficiaries
	Storage providers (including services)		
	Software/service providers in business intelligence, search		

[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



Some sectors are positioned for greater gains from the use of big data

Historical productivity growth in the United States, 2000–08

[1] See appendix for detailed definitions and metrics used for value potential index.

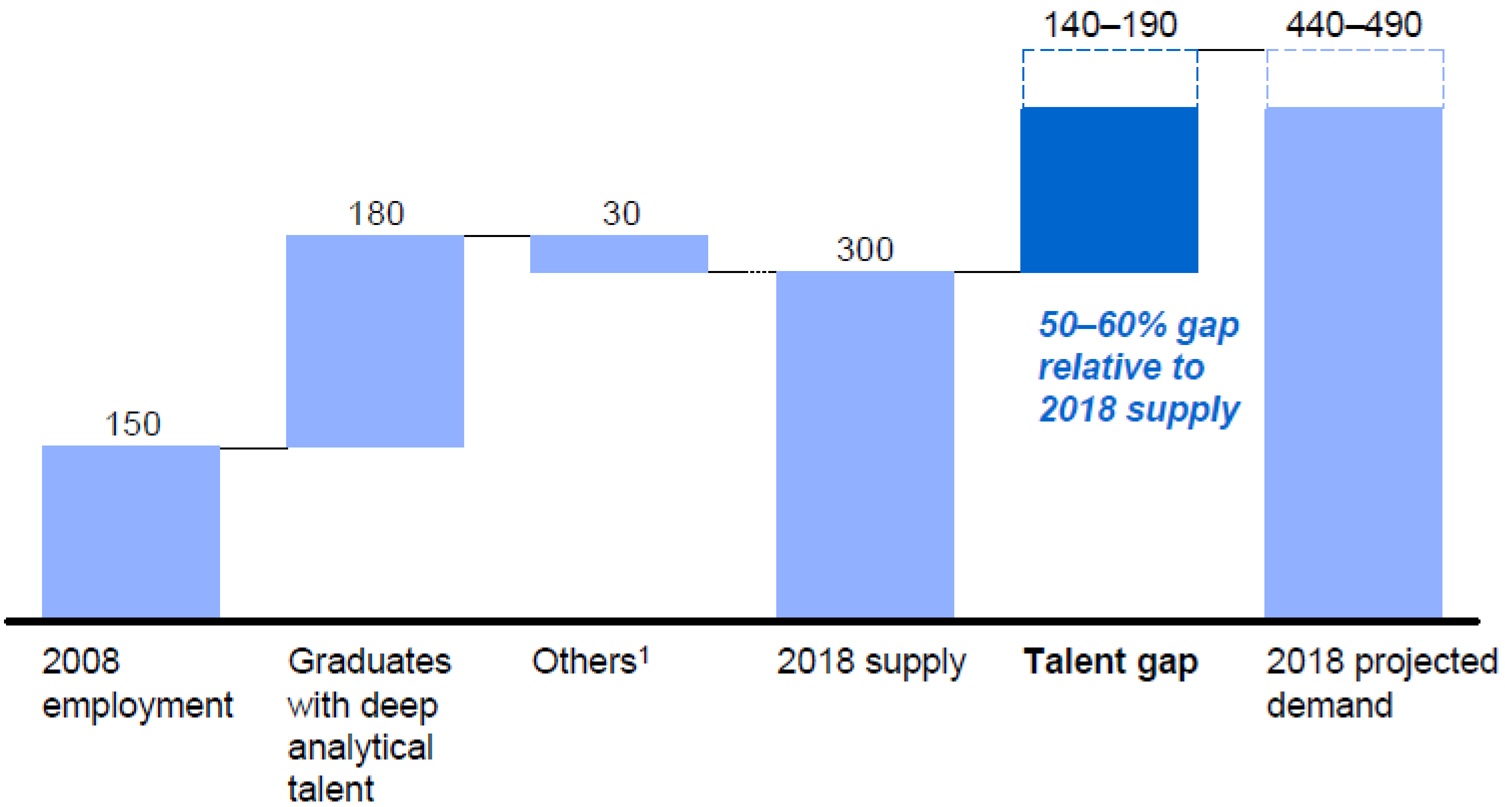
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

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

Tools

Tools typically used in Big-Data scenarios

- NoSQL
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Techniques

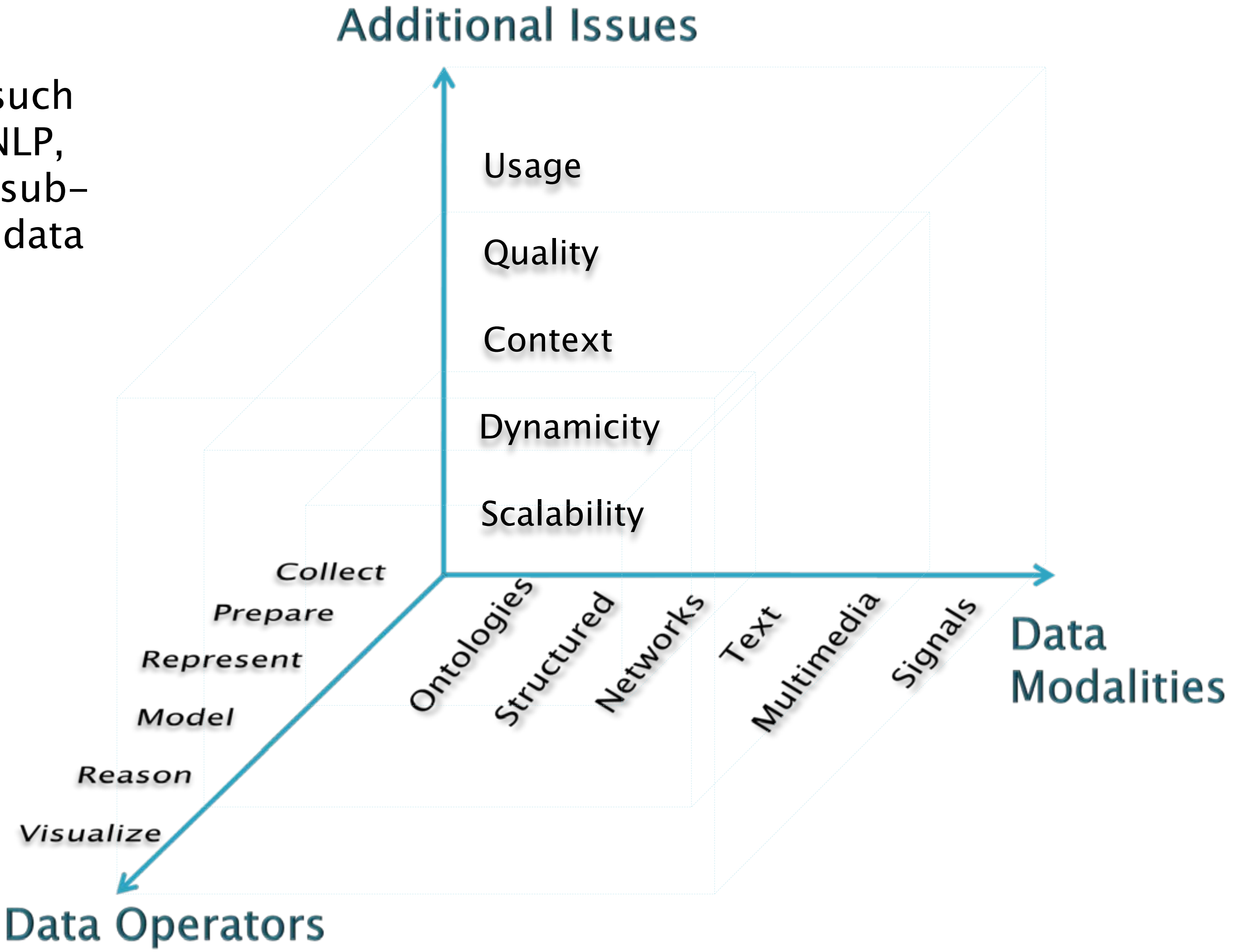
When Big-Data is really a hard problem?

- …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



Additional Issues

Usage

Quality

Context

Dynamicity

Scalability

Data Modalities

Ontologies

Structured

Networks

Text

Multimedia

Signals

Data Operators

Collect

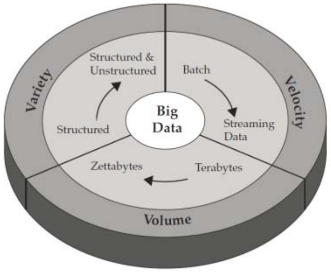
Prepare

Represent

Model

Reason

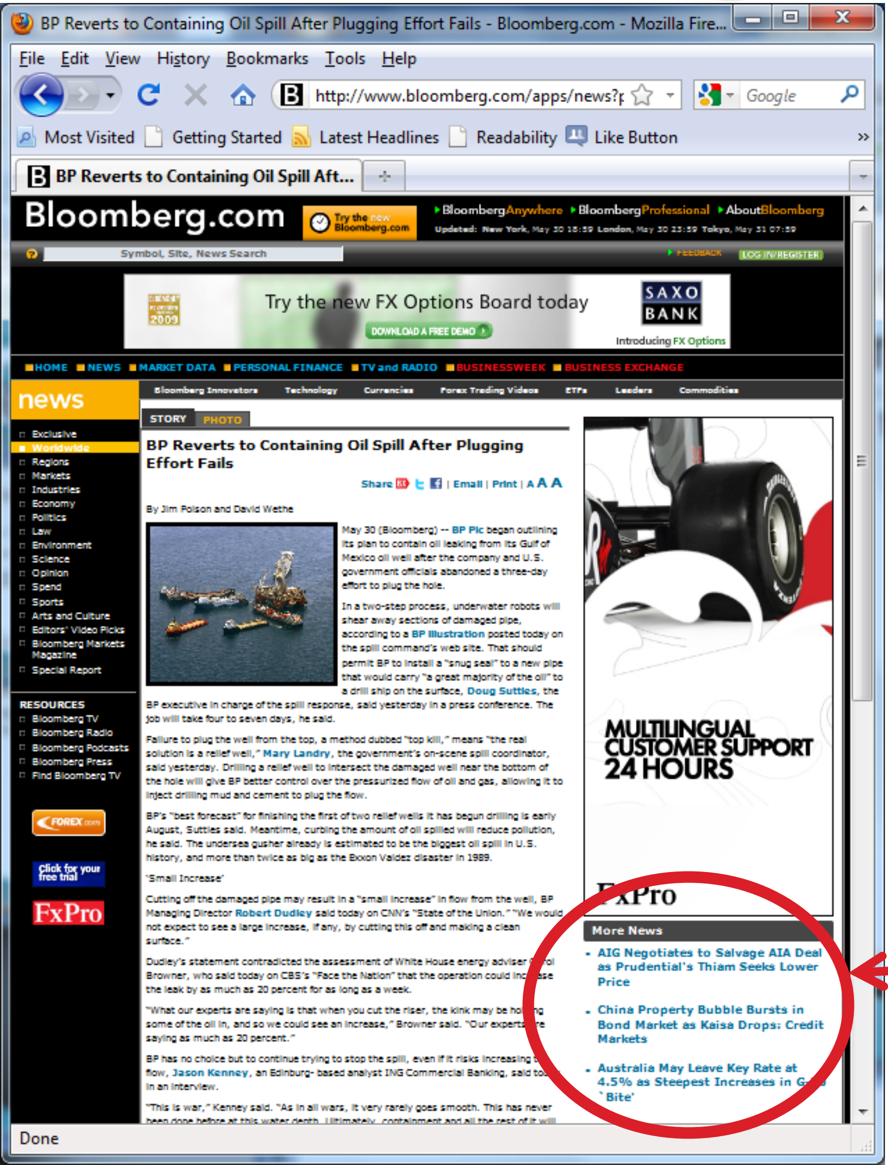
Visualize



Applications

Recommendation

…an example: recommendation @Bloomberg.com



BP Reverts to Containing Oil Spill After Plugging Effort Fails

By Jim Polson and David Wethe

May 30 (Bloomberg) -- BP Plc began outlining its plan to contain oil leaking from its Gulf of Mexico oil well after the company and U.S. government officials abandoned a three-day effort to plug the hole.

In a two-step process, underwater robots will shear away sections of damaged pipe, according to a BP illustration posted today on the spill command's web site. That should permit BP to install a "snug seal" to a new pipe that would carry "a great majority of the oil" to a drill ship on the surface, Doug Suttles, the BP executive in charge of the spill response, said yesterday in a press conference. The job will take four to seven days, he said.

Failure to plug the well from the top, a method dubbed "top kill," means "the real solution is a relief well," Mary Landry, the government's on-scene spill coordinator, said yesterday. Drilling a relief well to intersect the damaged well near the bottom of the hole will give BP better control over the pressurized flow of oil and gas, allowing it to inject drilling mud and cement to plug the flow.

BP's "best forecast" for finishing the first of two relief wells it has begun drilling is early August, Suttles said. Meantime, curbing the amount of oil spilled will reduce pollution, he said. The undersea gusher already is estimated to be the biggest oil spill in U.S. history, and more than twice as big as the Exxon Valdez disaster in 1989.

'Small Increase'

Cutting off the damaged pipe may result in a "small increase" in flow from the well, BP Managing Director Robert Dudley said today on CNN's "State of the Union." "We would not expect to see a large increase, if any, by cutting this off and making a clean surface."

Dudley's statement contradicted the assessment of White House energy adviser Carol Browner, who said today on CBS's "Face the Nation" that the operation could increase the leak by as much as 20 percent for as long as a week.

"What our experts are saying is that when you cut the riser, the kink may be holding some of the oil in, and so we could see an increase," Browner said. "Our experts are saying as much as 20 percent."

BP has no choice but to continue trying to stop the spill, even if it risks increasing the flow, Jason Kenney, an Edinburg- based analyst ING Commercial Banking, said today in an interview.

"This is war," Kenney said. "As in all wars, it very rarely goes smooth. This has never been done before at this water depth. Ultimately, containment and all the rest of it will

More News

- AIG Negotiates to Salvage AIA Deal as Prudential's Thiam Seeks Lower Price
- China Property Bubble Bursts in Bond Market as Kaisa Drops: Credit Markets
- Australia May Leave Key Rate at 4.5% as Steepest Increases in G-10 `Bite'

- Good recommendations can make a big difference when keeping a user on a web site
  - …the key is how rich context model a system is using to select information for a user
  - Bad recommendations <1% users, good ones >5% users click

Contextual personalized recommendations generated in ~20ms

Each click on the web site is enriched and indexed using:

- Domain
- Sub-domain
- Page URL
- URL sub-directories

- Page Meta Tags
- Page Title
- Page Content
- Named Entities

- Has Query
- Referrer Query

- Referring Domain
- Referring URL
- Outgoing URL

- GeoIP Country
- GeoIP State
- GeoIP City

- Absolute Date
- Day of the Week
- Day period
- Hour of the day
- User Agent

- Zip Code
- State
- Income
- Age
- Gender
- Country
- Job Title
- Job Industry

Application: Online Advertising for NYTimes (microtrends detection)

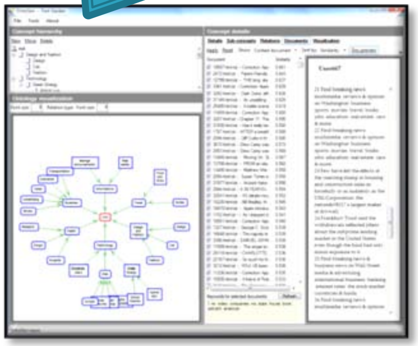
Log Files (~100M page clicks per day)

Stream of clicks

User profiles

Stream of profiles

Trend Detection System



NYT articles

Content

Trends and updated segments

Segment	Keywords
Stock Market	Stock Market, mortgage, banking, investors, Wall Street, turmoil, New York Stock Exchange
Health	diabetes, heart disease, disease, heart, illness
Green Energy	Hybrid cars, energy, power, model, carbonated, fuel, bulbs,
Hybrid cars	Hybrid cars, vehicles, model, engines, diesel
Travel	travel, wine, opening, tickets, hotel, sites, cars, search, restaurant
…	…

Segments

Sales

Campaign to sell segments

$

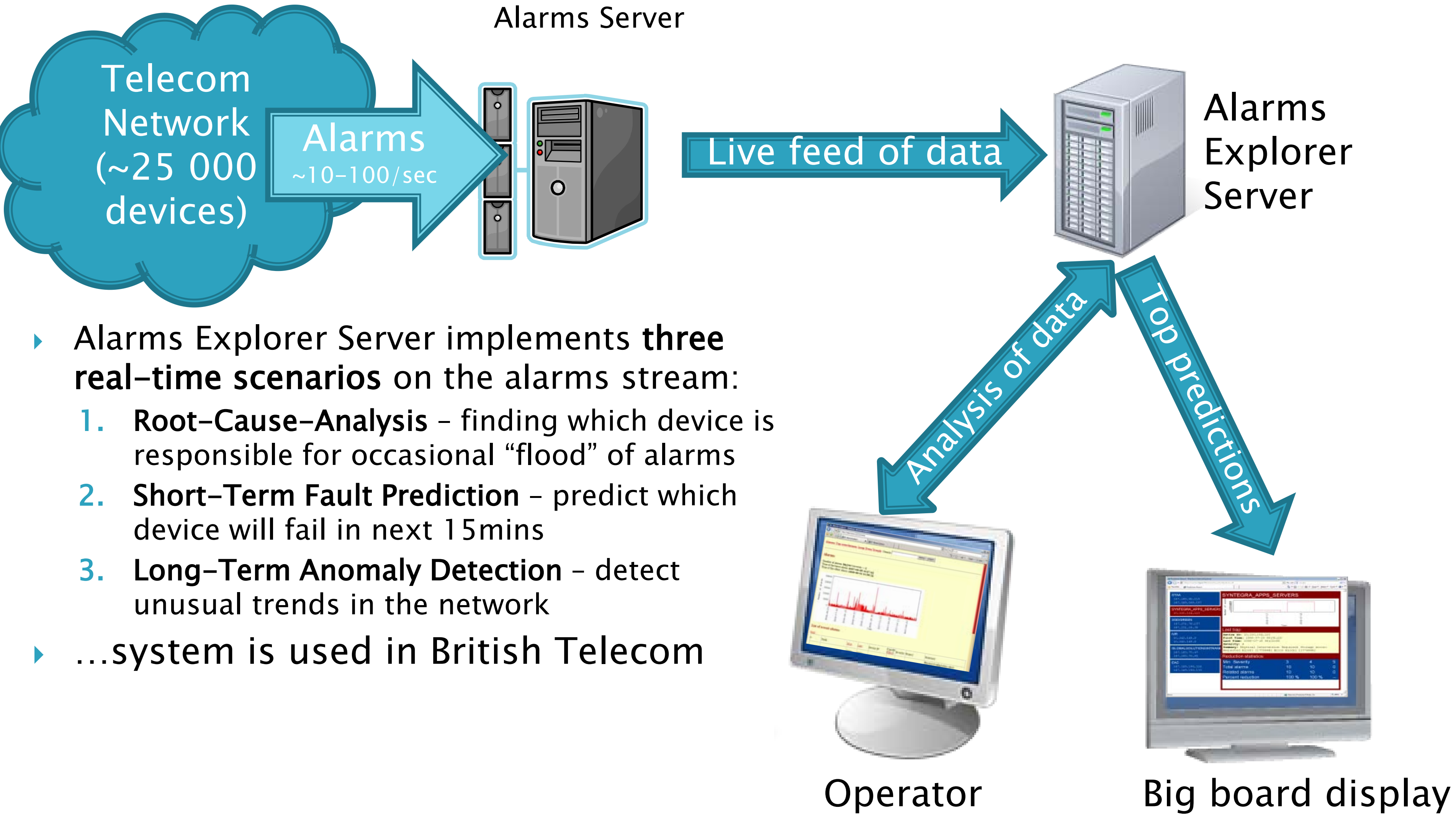
Advertisers

Figures for one day of NYTimes

- 50Gb of uncompressed log files
- 10Gb of compressed log files
- 0.5Gb of processed log files
- 50-100M clicks
- 4-6M unique users
- 7000 unique pages with more then 100 hits
- Index size 2Gb
- Pre-processing & indexing time
  - ~10min on workstation (4 cores & 32Gb)
  - ~1hour on EC2 (2 cores & 16Gb)

Root-cause analysis

Applications: Telecommunication Network Monitoring



- Alarms Explorer Server implements **three real-time scenarios** on the alarms stream:
  1. **Root-Cause-Analysis** – finding which device is responsible for occasional “flood” of alarms
  2. **Short-Term Fault Prediction** – predict which device will fail in next 15mins
  3. **Long-Term Anomaly Detection** – detect unusual trends in the network
- …system is used in British Telecom

Analysis of MSN-Messenger Social-network

- Presented in “Planetary-Scale Views on a Large Instant-Messaging Network” by Jure Leskovec and Eric Horvitz WWW2008

Instant Messenger – Phenomena at a planetary scale

- Observe social and communication phenomena at a *planetary* scale
- Largest social network analyzed to date

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**?

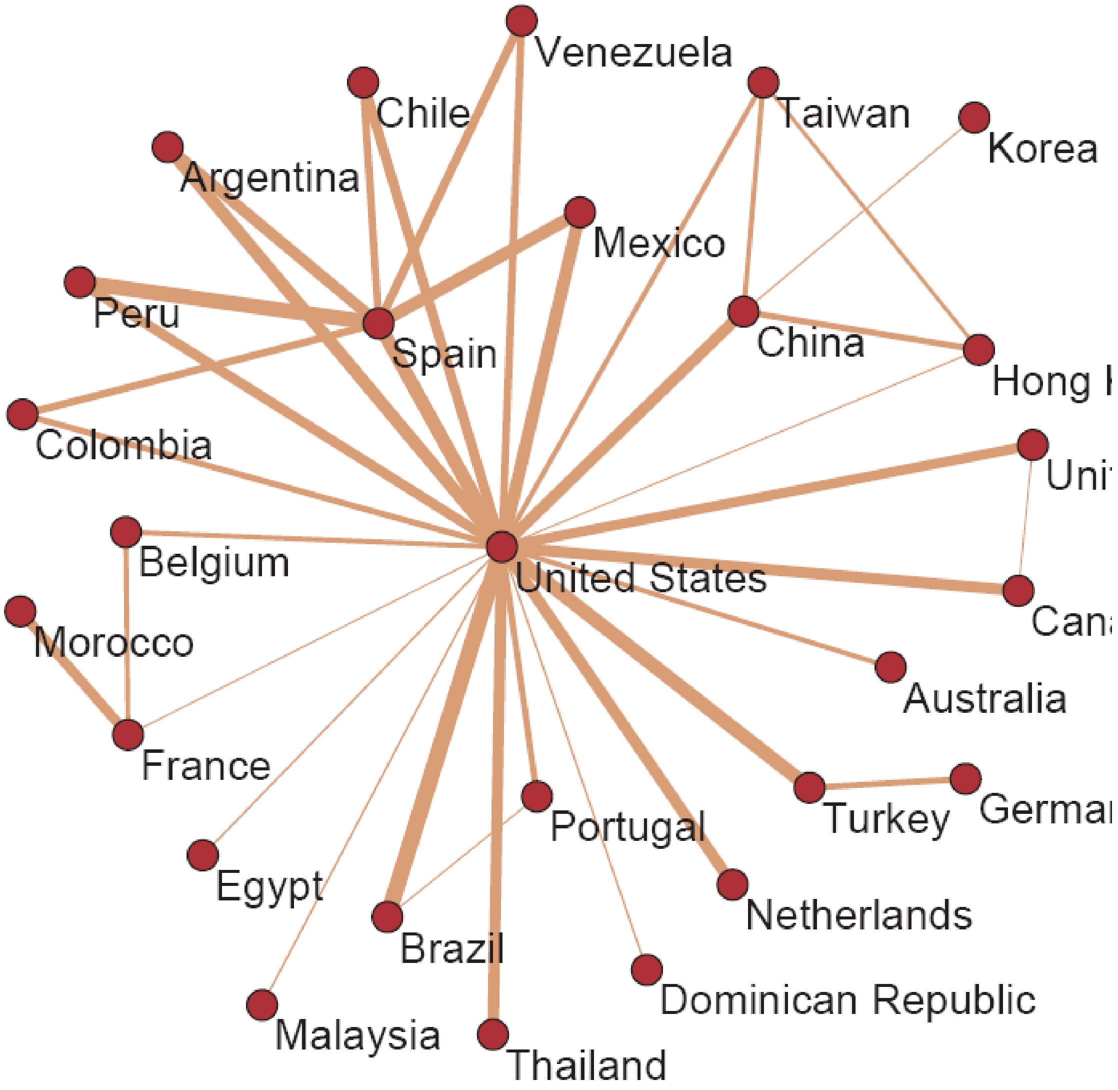
Data statistics: Total activity

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

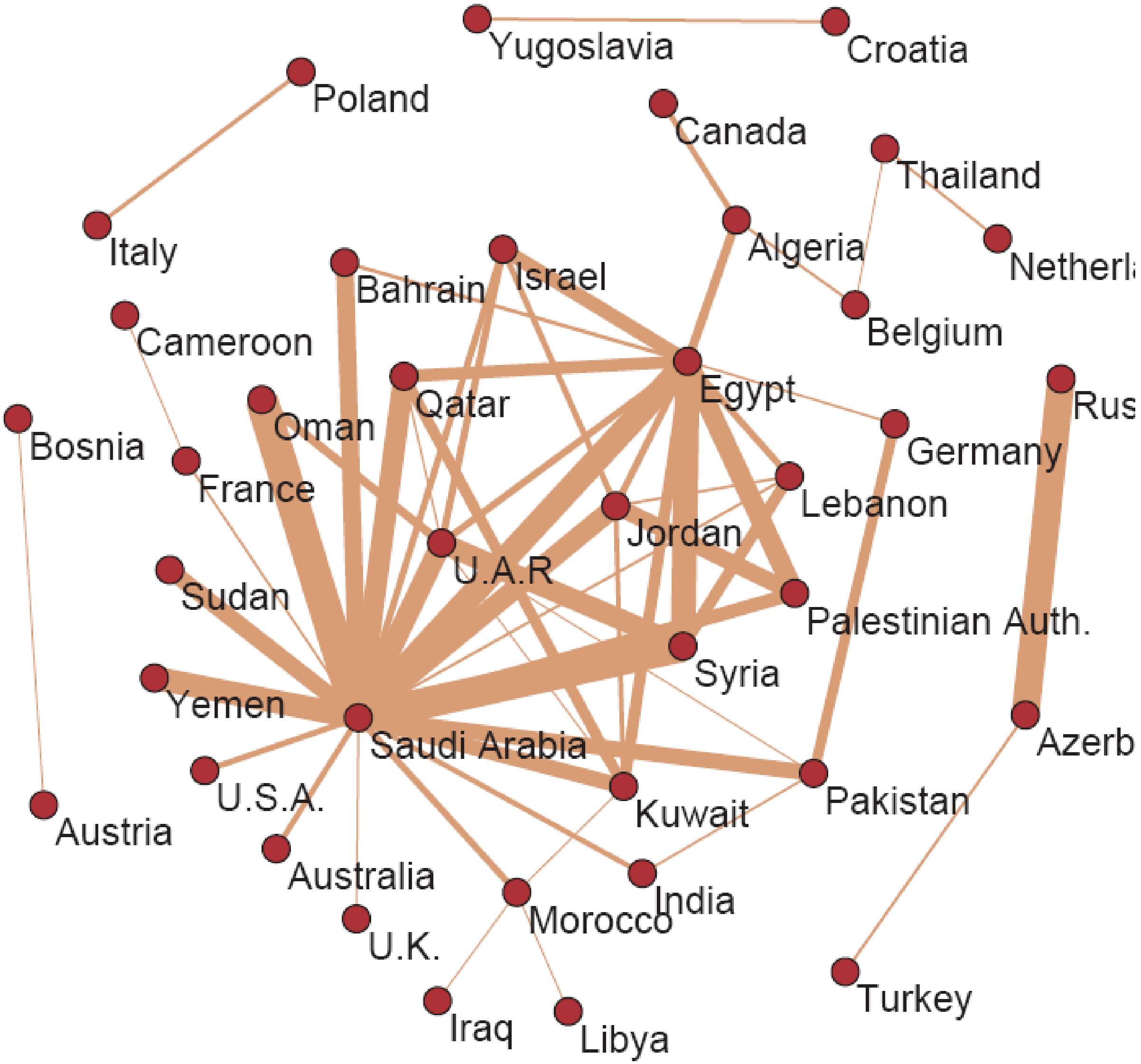
  150Gb/day (compressed)
- Total: 1 month of communication data:

  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

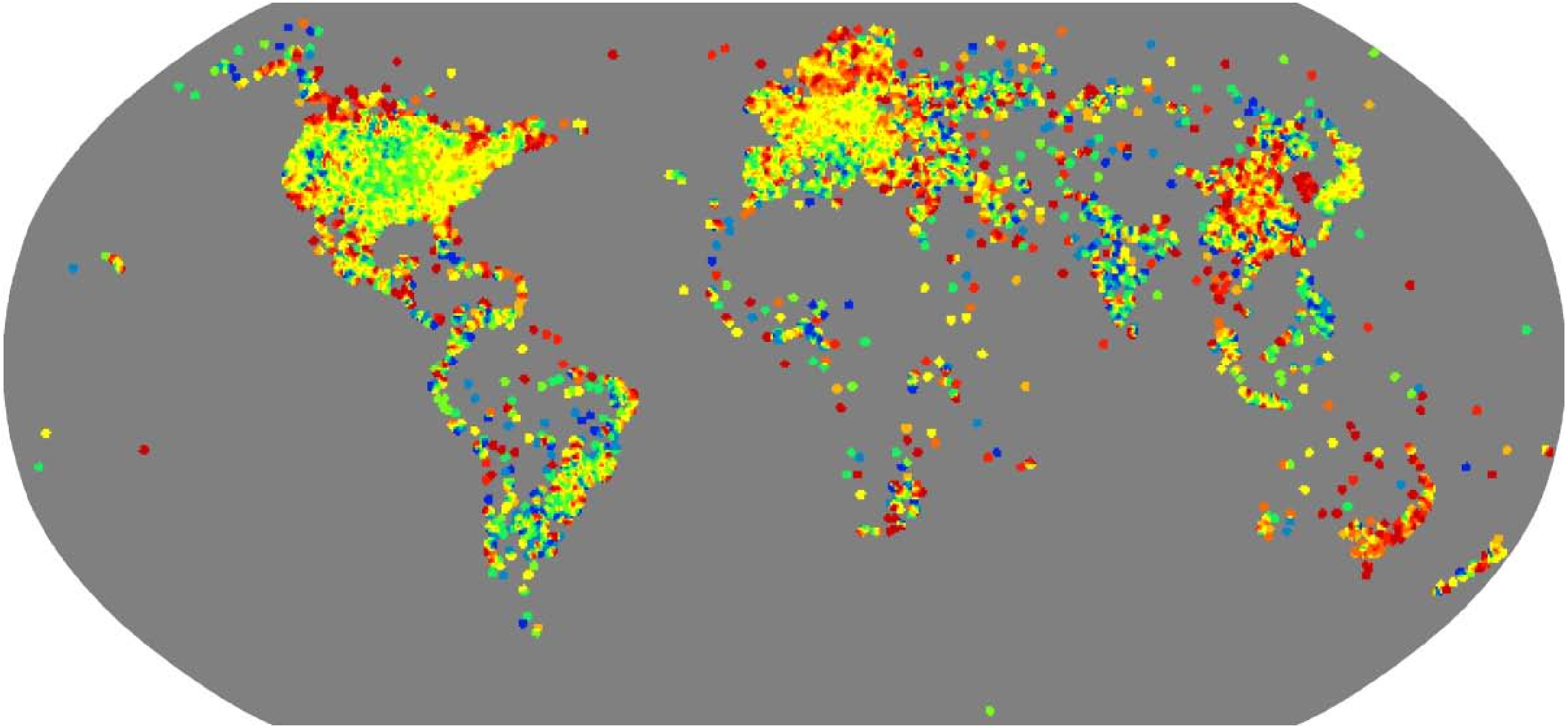
Who talks to whom: Number of conversations



Who talks to whom: Conversation duration



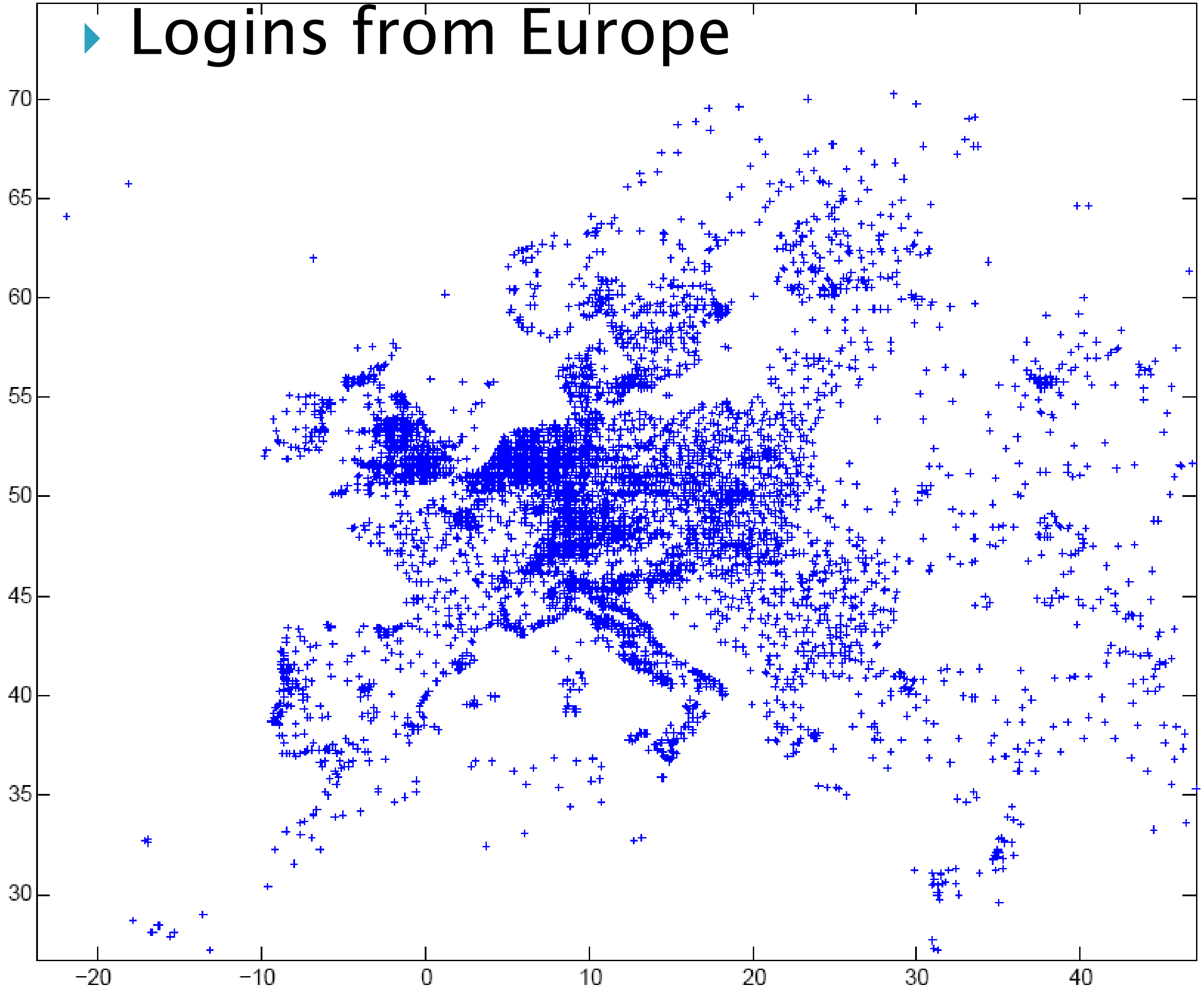
Geography and communication



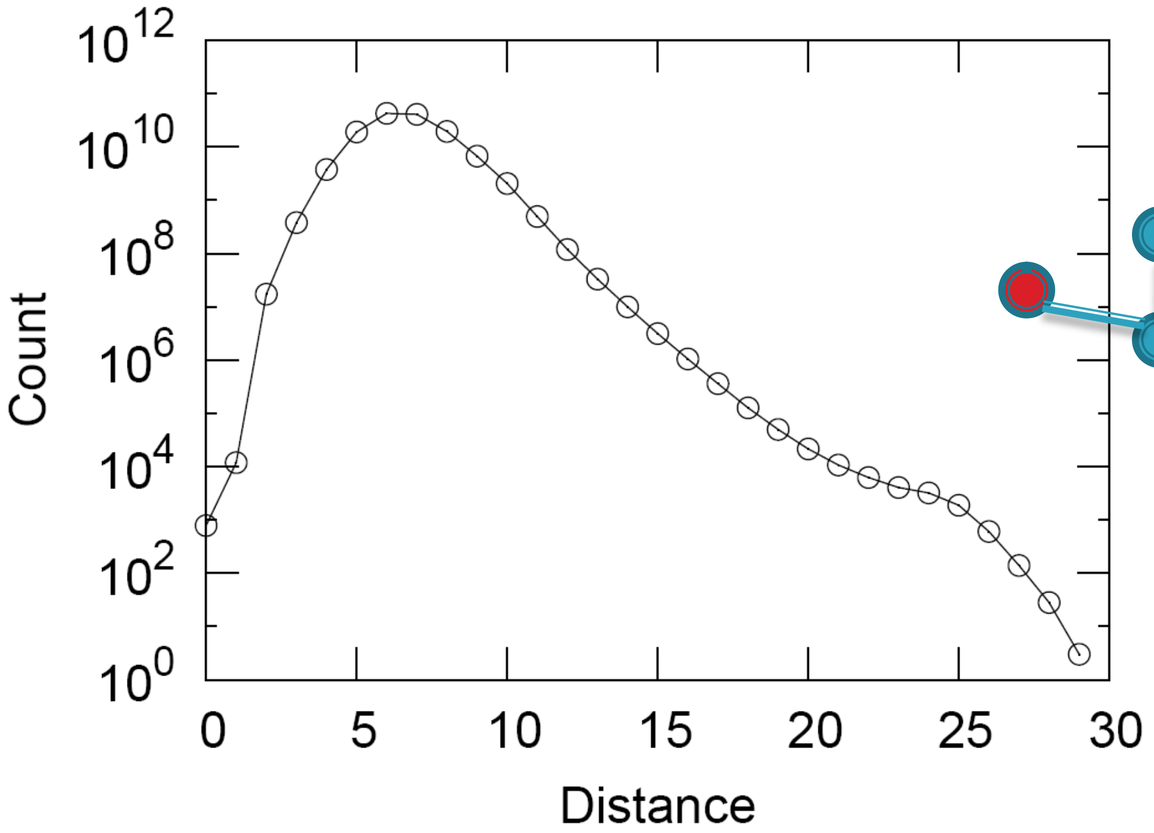
- Count the number of users logging in from particular location on the earth

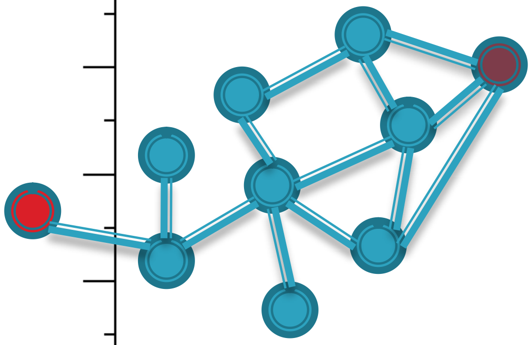
How is Europe talking

- Logins from Europe



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

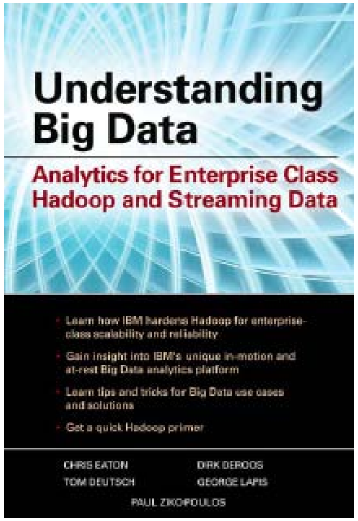
Web-of-Things

Literature on Big-Data

McKinsey & Company

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

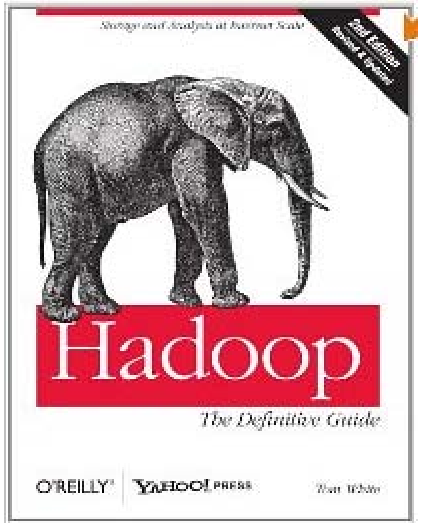


Understanding Big Data

Analytics for Enterprise Class Hadoop and Streaming Data

- Learn how IBM hardens Hadoop for enterprise-class scalability and reliability
- Gain insight into IBM's unique in-motion and at-rest Big Data analytics platform
- Learn tips and tricks for Big Data use cases and solutions
- Get a quick Hadoop primer

CHRIS EATON DIRK DEROOS
TOM DEUTSCH GEORGE LAPIS
PAUL ZIKOPOULOS



Storage and Analysis at Internet Scale

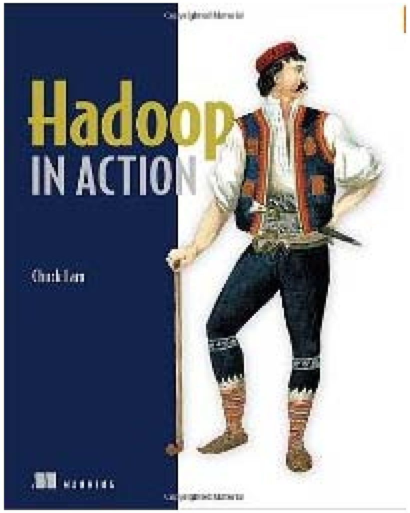
2nd Edition Revised & Updated

Hadoop

The Definitive Guide

O'REILLY® YAHOO! PRESS

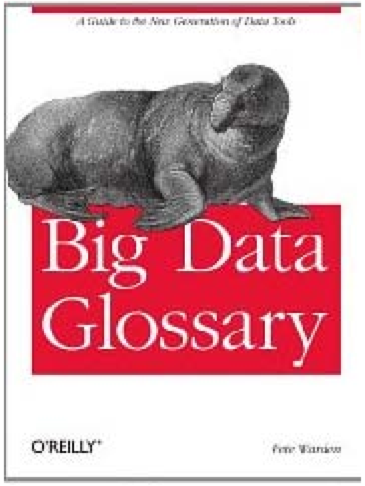
Tom White



Hadoop IN ACTION

Chuck Lam

MANNING

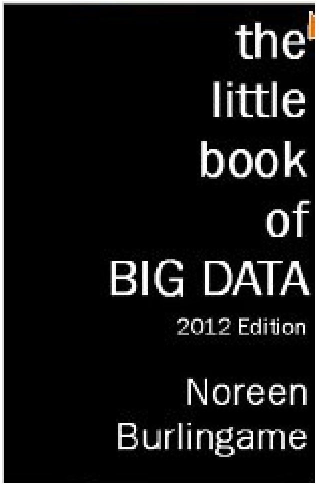


A Guide to the New Generation of Data Tools

Big Data Glossary

O'REILLY®

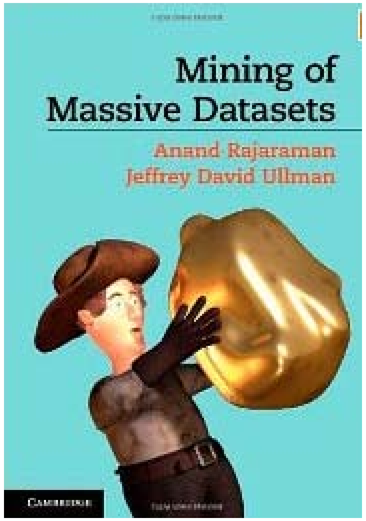
Pete Warden



the little book of BIG DATA

2012 Edition

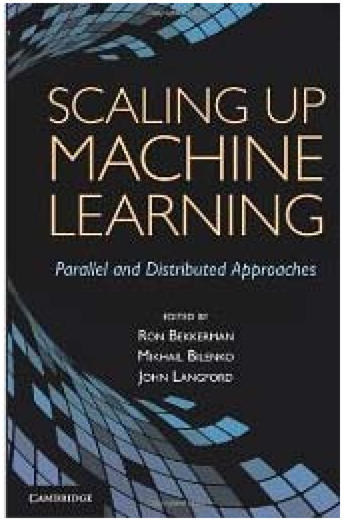
Noreen Burlingame



Mining of Massive Datasets

Anand Rajaraman
Jeffrey David Ullman

CAMBRIDGE



SCALING UP MACHINE LEARNING

Parallel and Distributed Approaches

EDITED BY
RON BEKKERMAN
MIKHAIL BILENKO
JOHN LANGFORD

CAMBRIDGE