

Big-Data Tutorial

Marko Grobelnik

marko.grobelnik@ijs.si

Jozef Stefan Institute

Kalamaki, May 25th 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

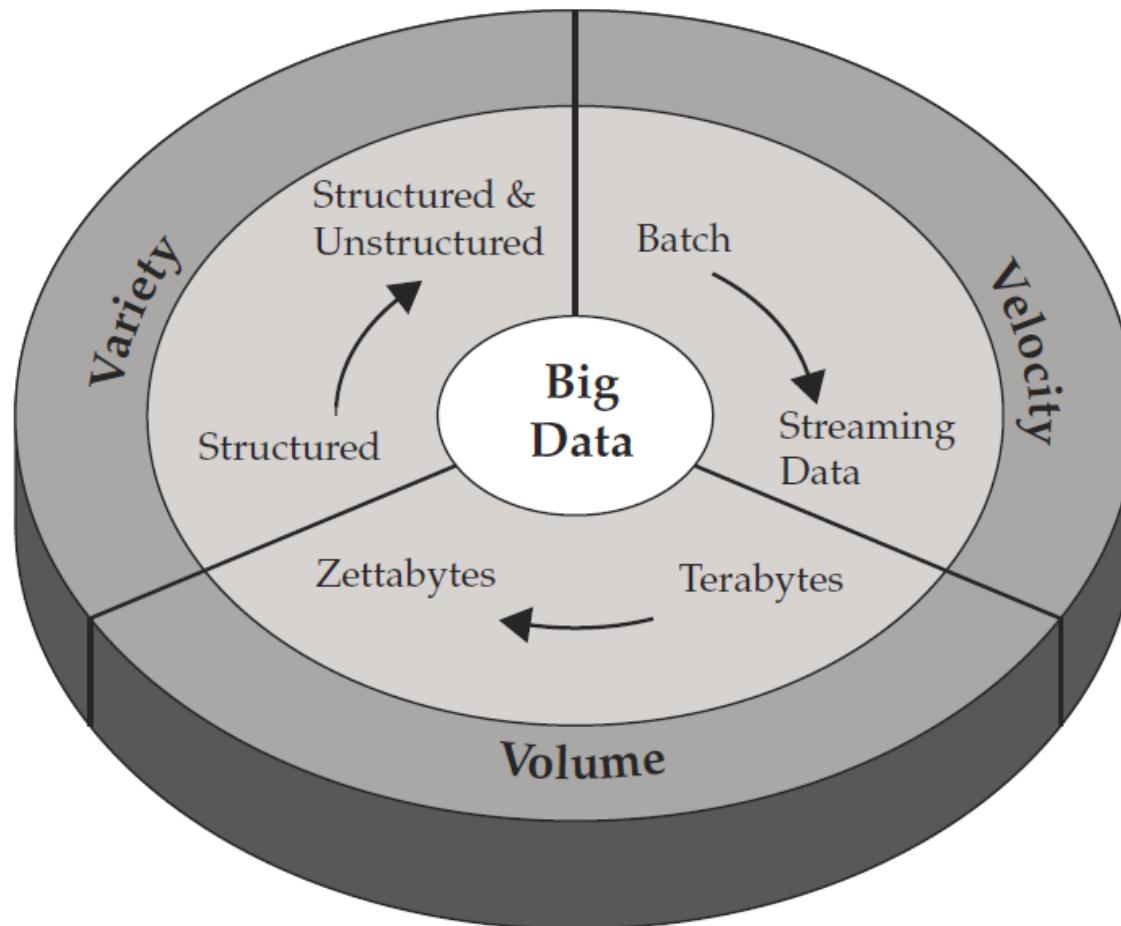
1.5 million

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

What is Big-Data?

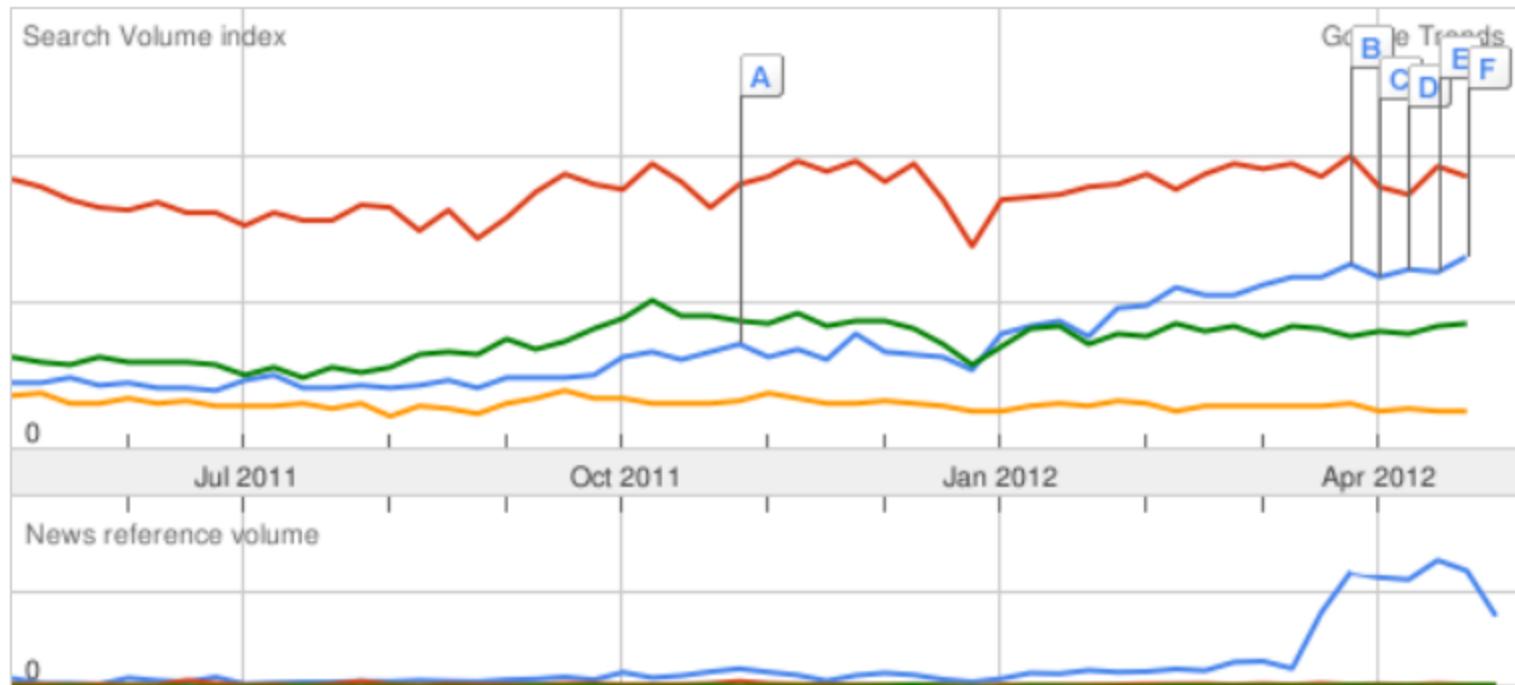
- ▶ ‘Big-data’ is similar to ‘Small-data’, but bigger
- ▶ ...but having data bigger consequently requires different approaches:
 - techniques, tools, architectures
- ▶ ...with an aim to solve new problems
 - ...and old problems in a better way.

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



Big-Data popularity on the Web

● big data ● data mining ● semantic web ● machine learning



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](#)

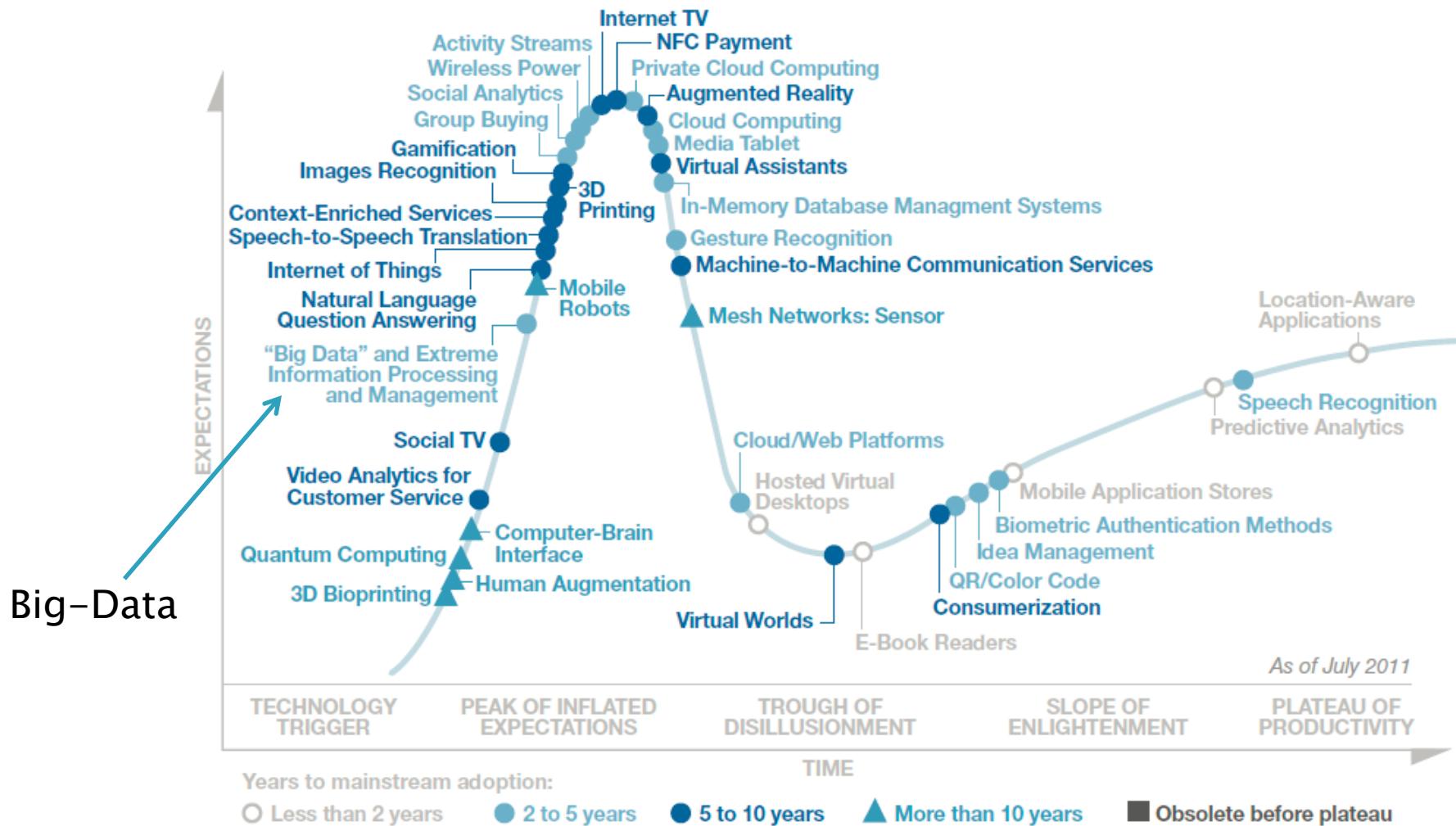
Houston Chronicle - Apr 17 2012

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



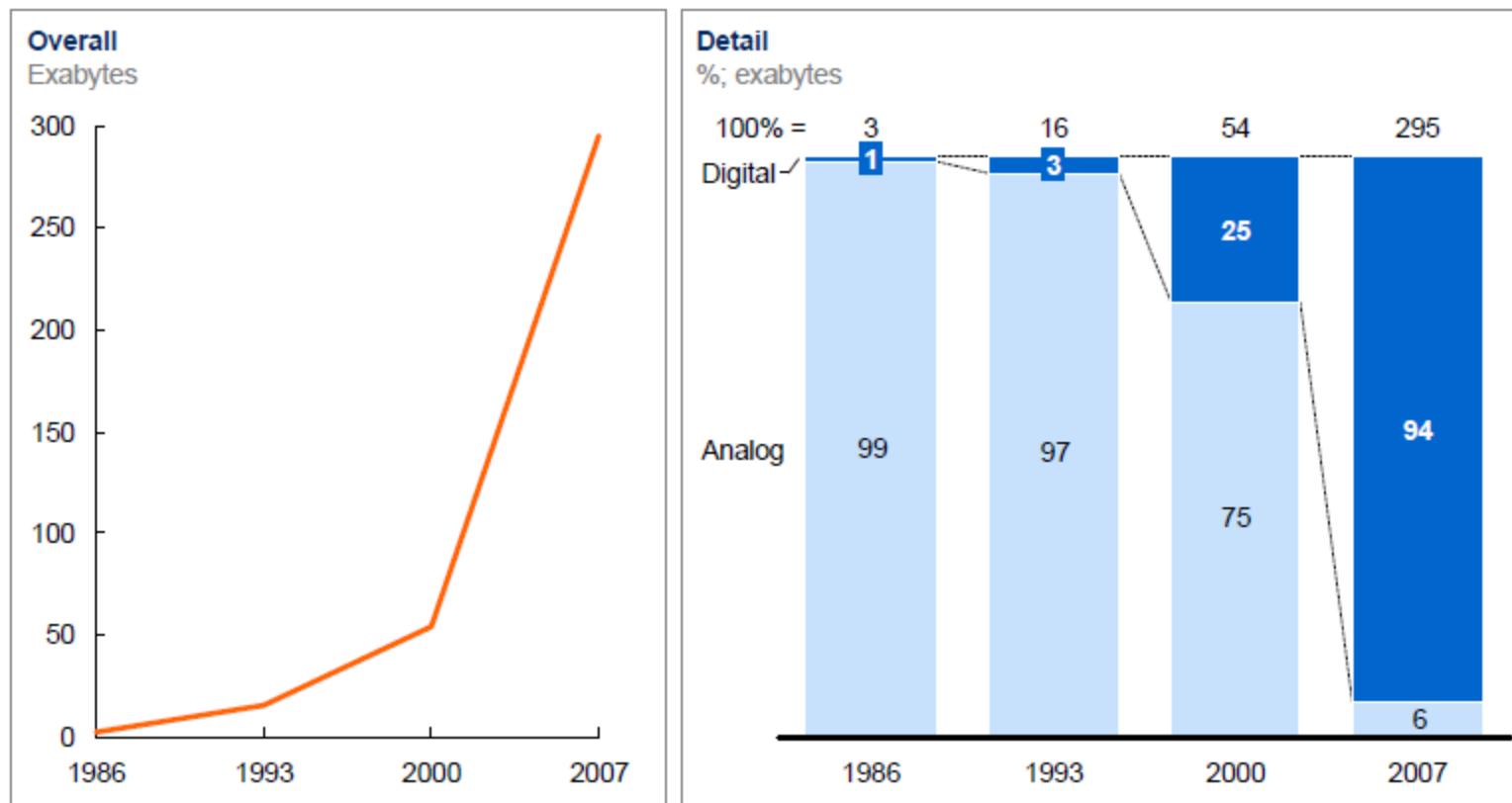
Why Big-Data?

- ▶ 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



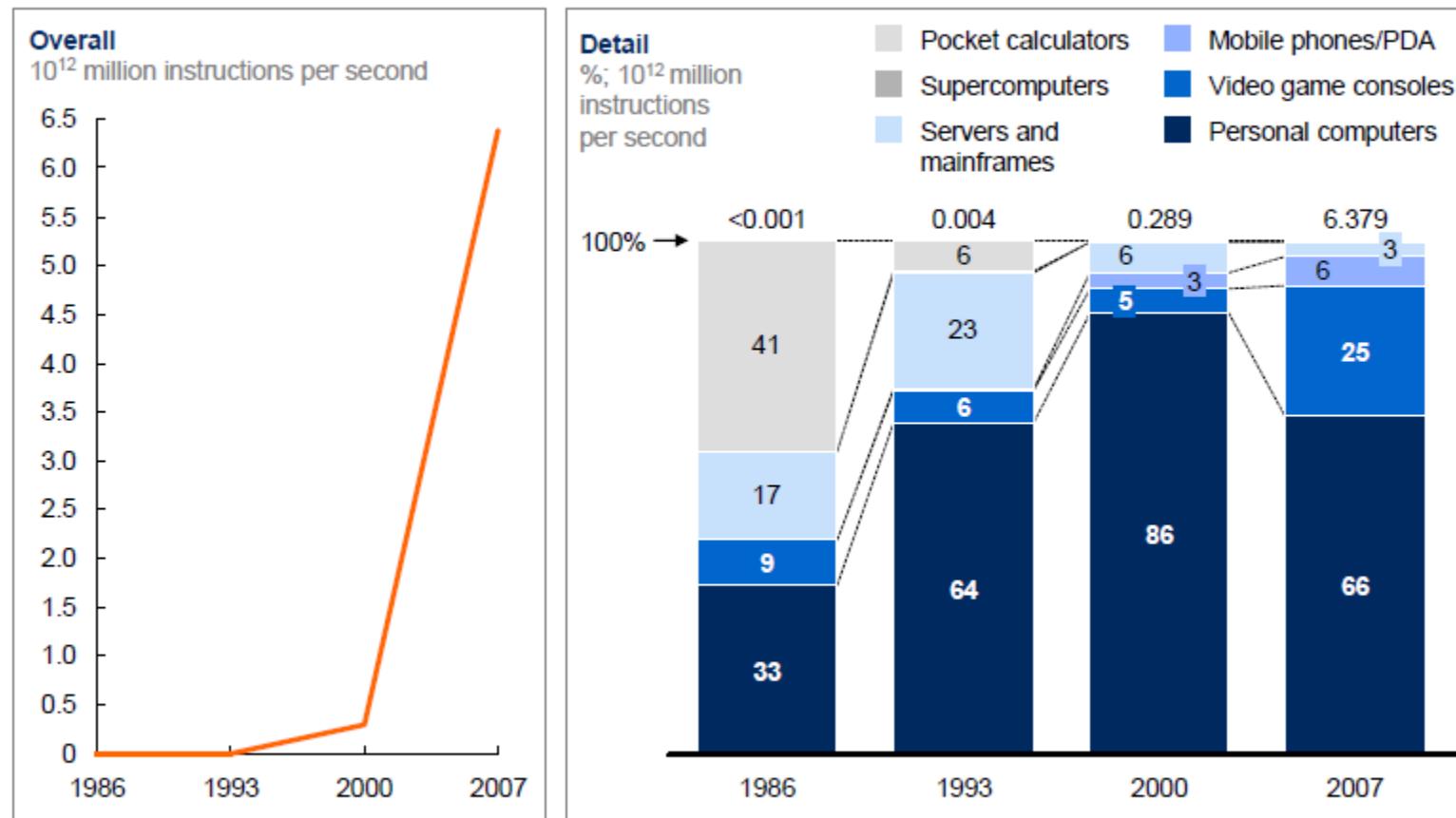
NOTE: Numbers may not sum due to rounding.

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

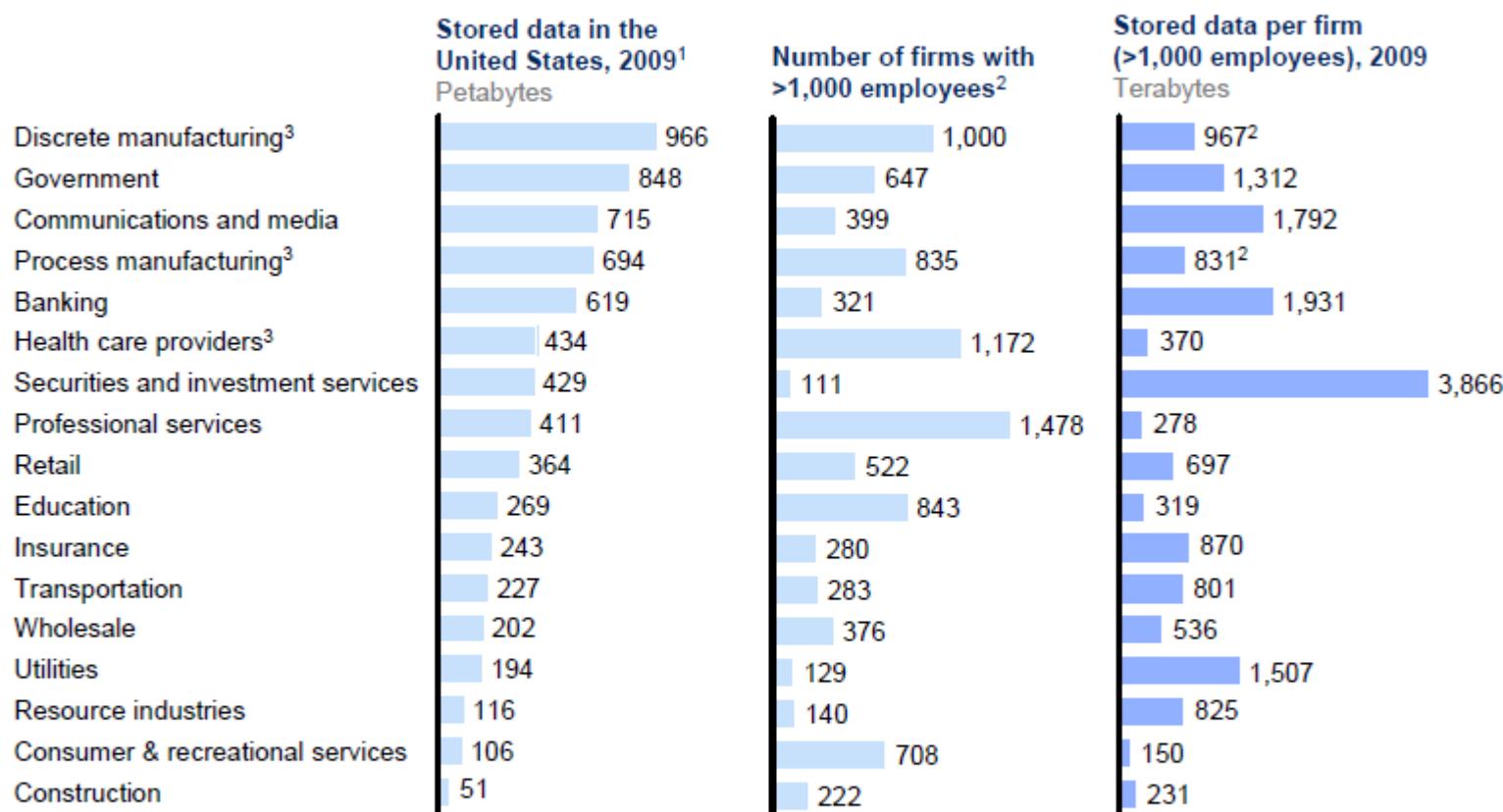


NOTE: Numbers may not sum due to rounding.

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



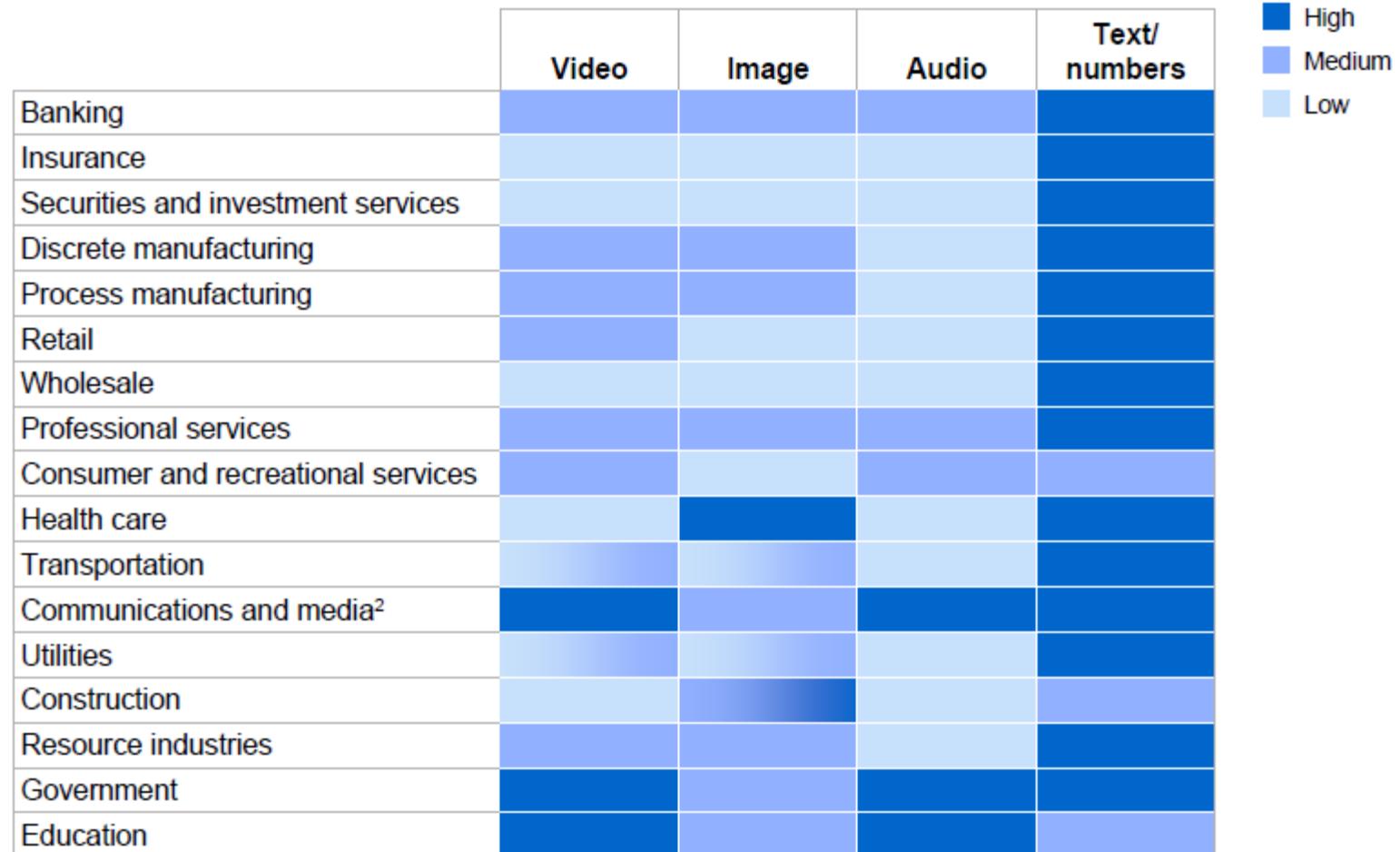
1 Storage data by sector derived from IDC.

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.

Type of available data

The type of data generated and stored varies by sector¹



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

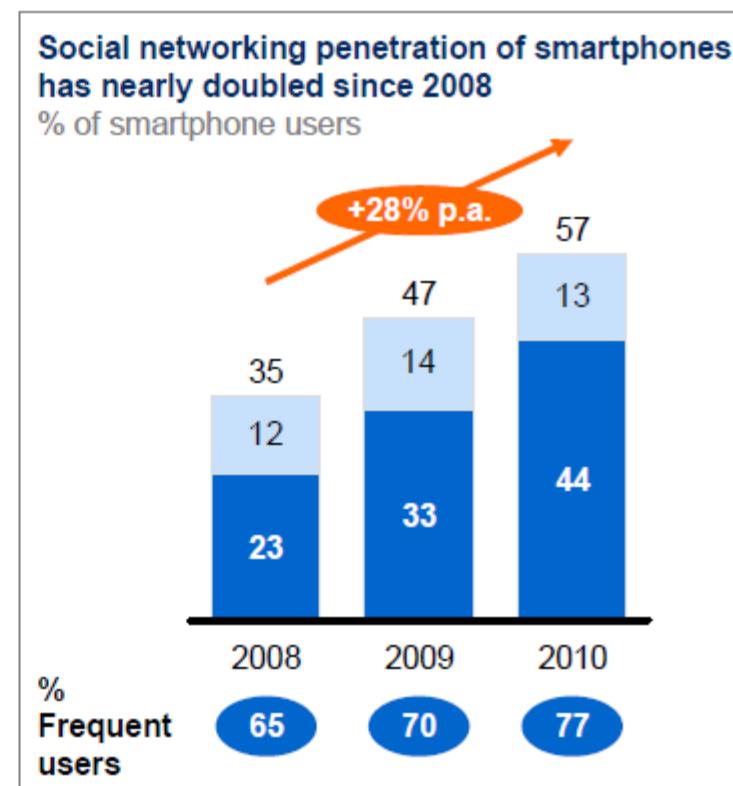
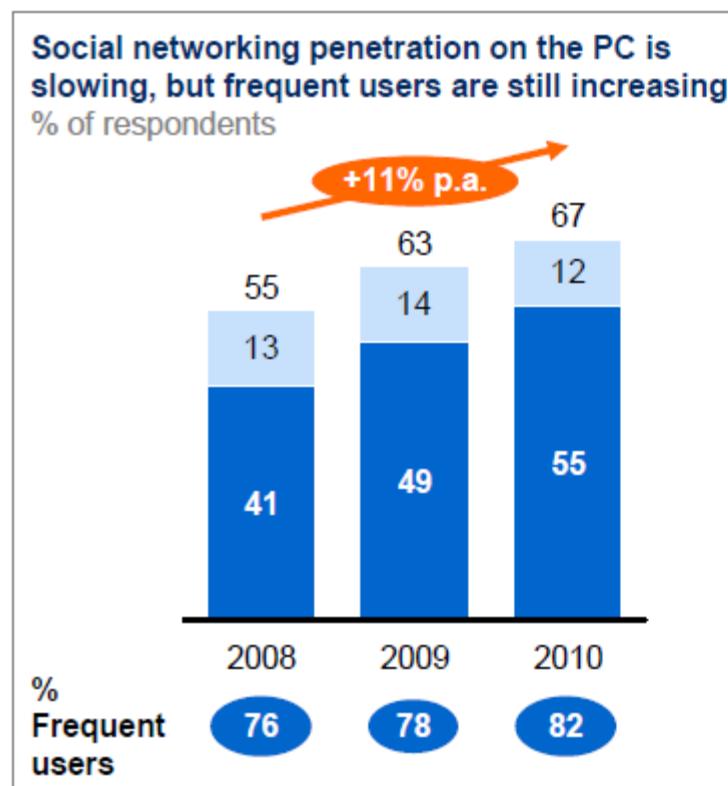
2 Video and audio are high in some subsectors.

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 users¹

Frequent user²



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.

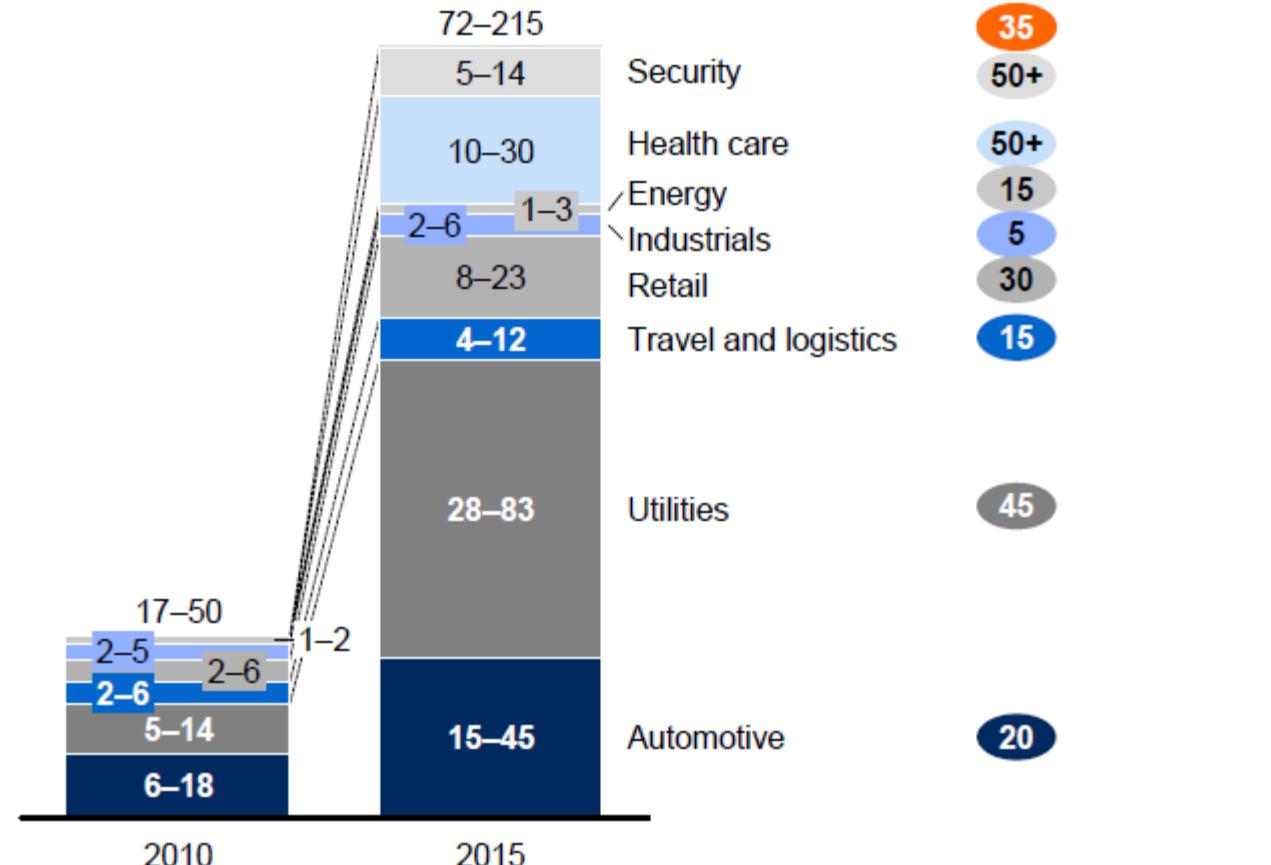
Data available from “Internet of Things”

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

Big-data value chain

Big data constituencies

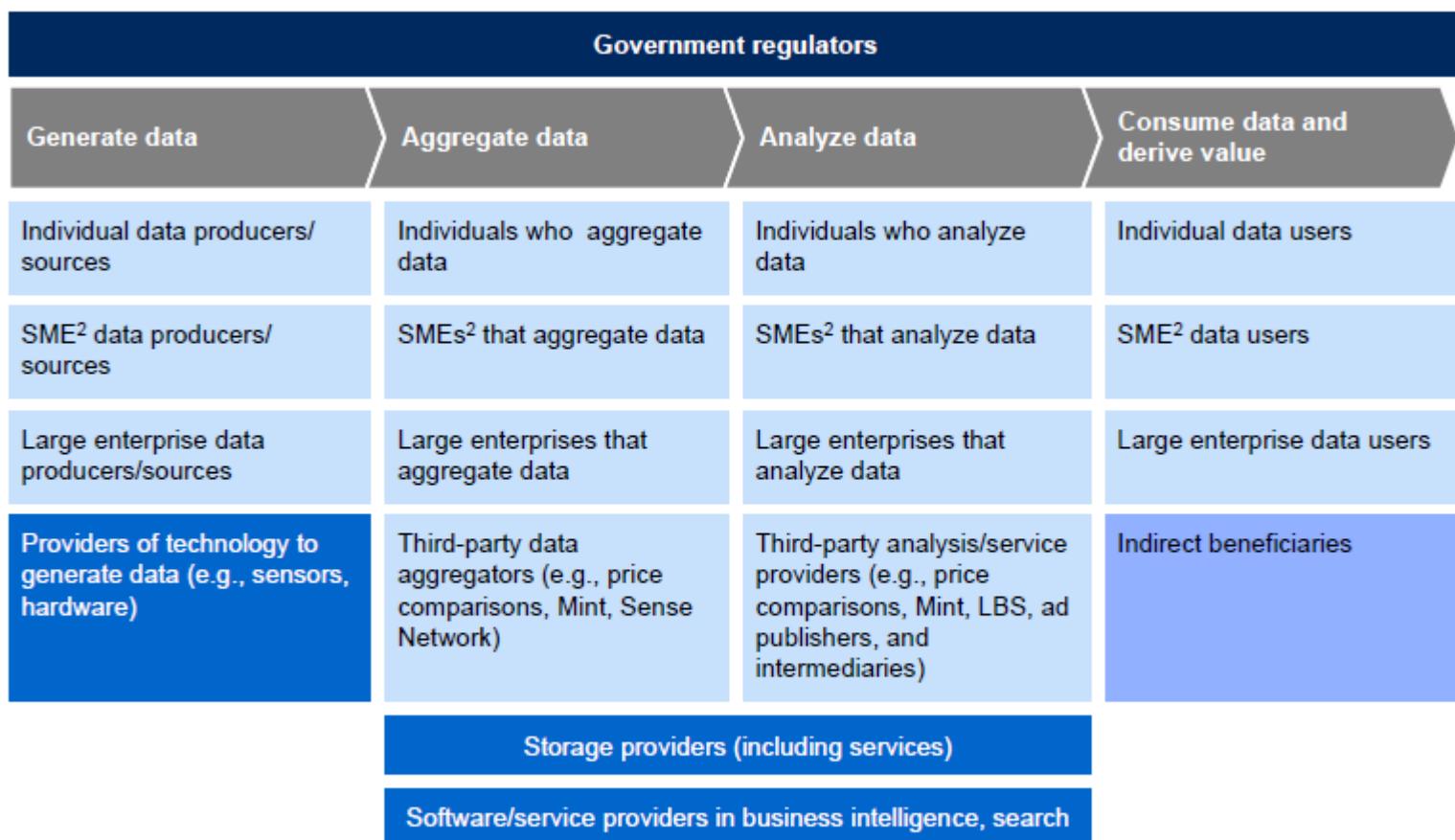
Big data activity/value chain

Individuals/organizations using data¹

Indirect beneficiaries

Providers of technology

Government regulators



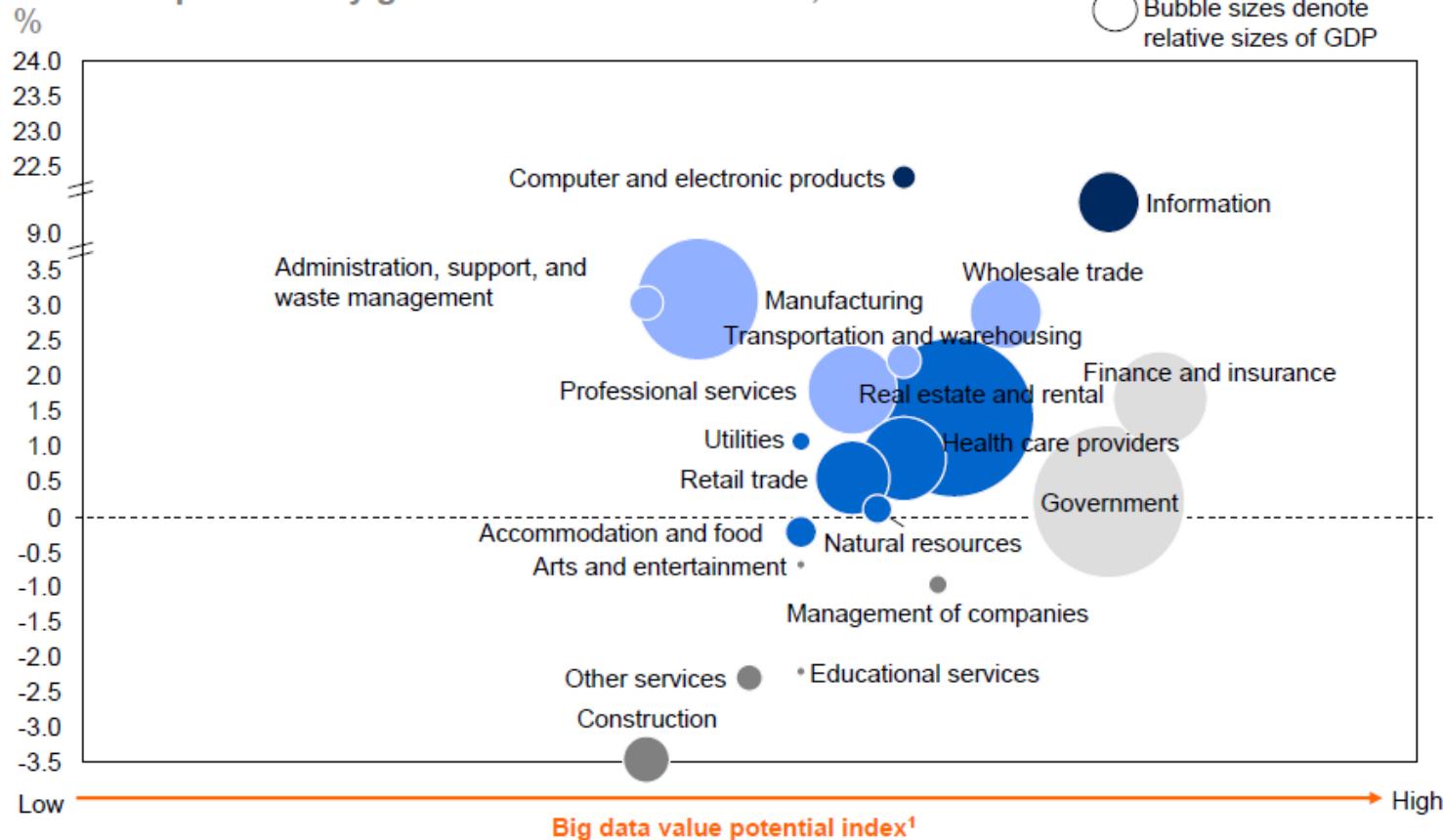
¹ Individuals/organizations generating, aggregating, analyzing, or consuming data.

² Small and medium-sized enterprises.

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.

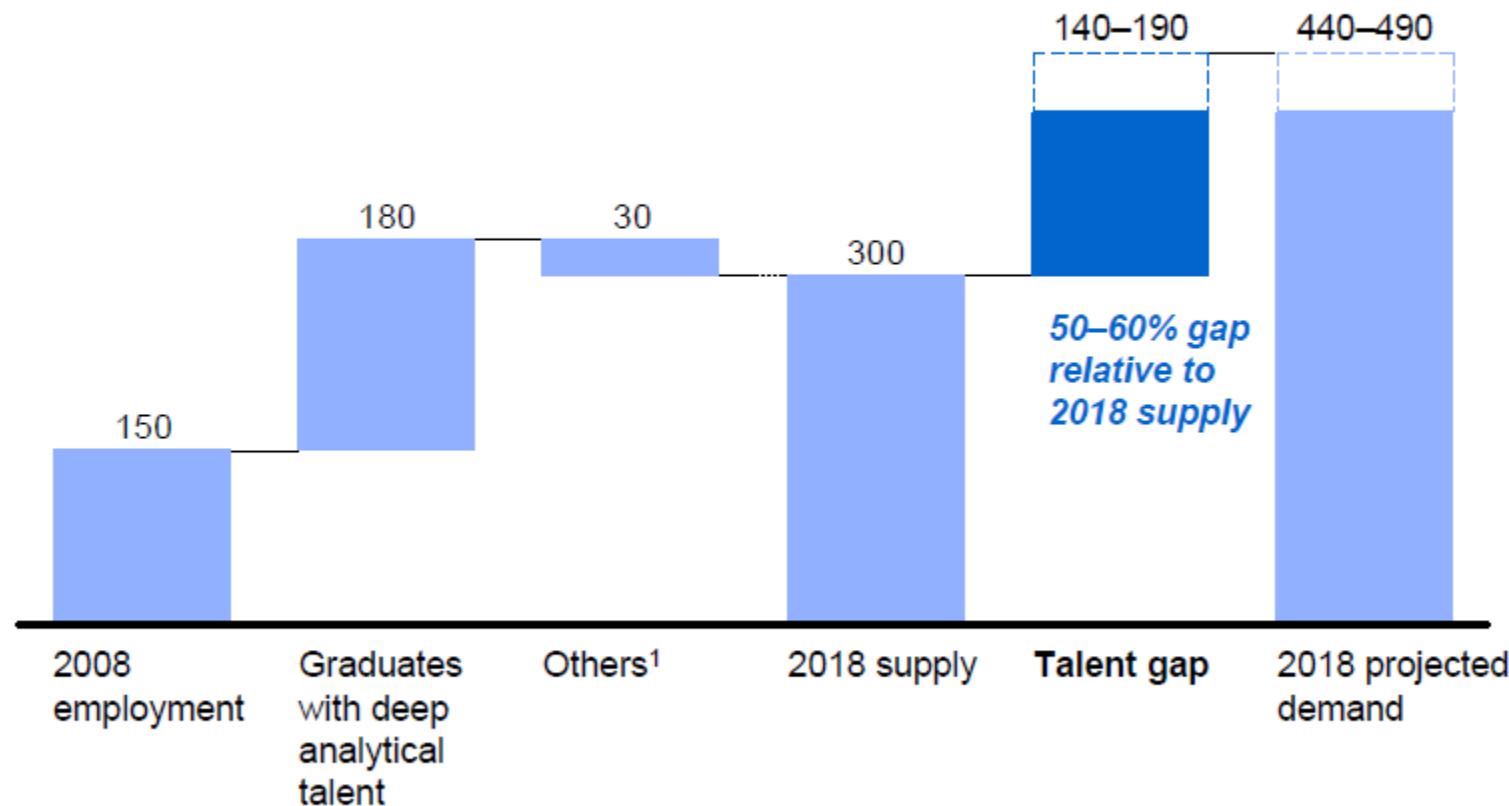
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

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

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 (e.g. R, OWLIM)

Distributed infrastructure

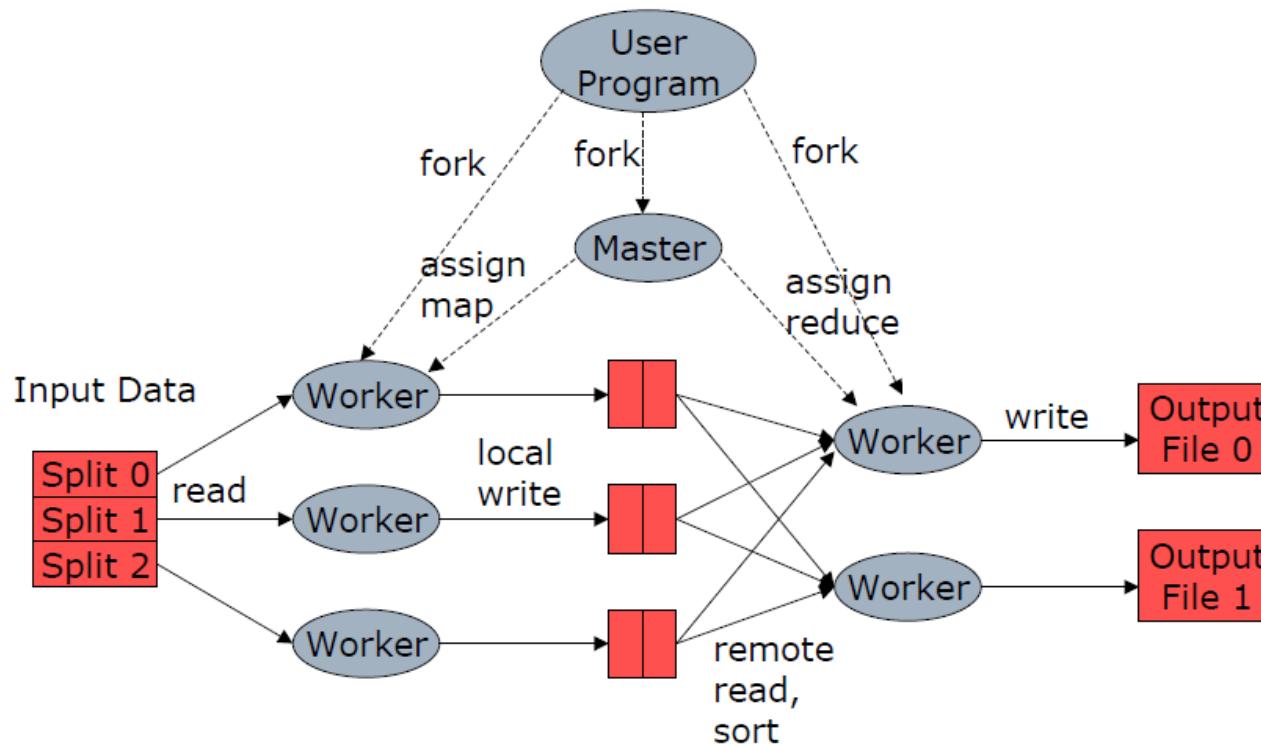
- ▶ Computing and storage are typically hosted transparently on cloud infrastructures
 - ...providing scale, flexibility and high fail-safety
- ▶ Distributed Servers
 - Amazon-EC2, Google App Engine, Elastic, Beanstalk, Heroku
- ▶ Distributed Storage
 - Amazon-S3, Hadoop Distributed File System

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
- ▶ Implementations of **MapReduce**
 - Hadoop (<http://hadoop.apache.org/>), Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum

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)



High-performance schema-free databases

- ▶ NoSQL class of databases have in common:
 - To support large amounts of data
 - Have mostly non-SQL interface
 - Operate on distributed infrastructures (e.g. Hadoop)
 - Are based on key-value pairs (no predefined schema)
 - ...are flexible and fast
- ▶ Implementations
 - MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper...

```
Spike:~ petewarden$ mongo
MongoDB shell version: 1.0.1
url: test
connecting to: test
type "help" for help
> db.users.save({name:"Pete Warden", eyes:"Blue"});
> db.users.find({name:"Pete Warden"});
{"_id" : ObjectId("4e48683fc6092f1f77ffac16") , "name" : "Pete Warden" , "eyes" : "Blue"}
> █
```

Techniques

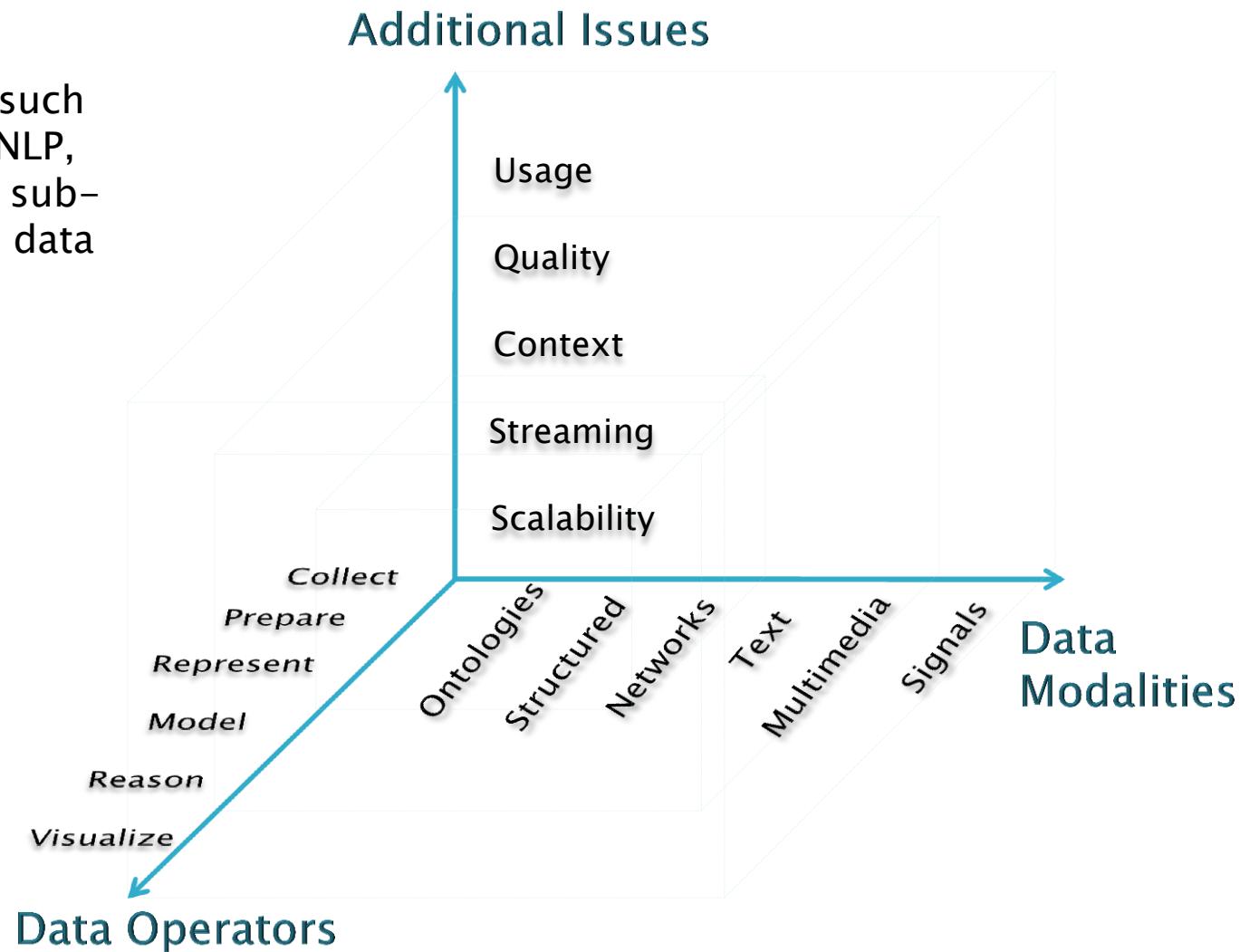
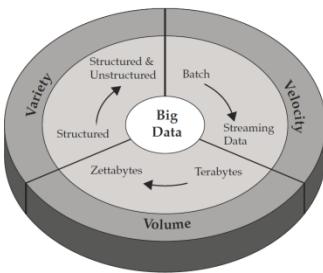
When Big-Data is really a hard problem?

- ▶ ...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

What are “atypical” operators on Big-Data

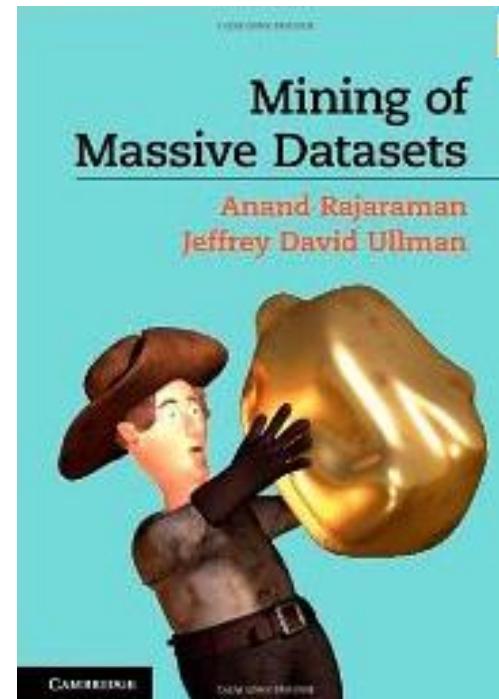
- ▶ **Smart sampling** of data
 - ...reducing the original data while not losing the statistical properties of data
- ▶ **Finding similar items**
 - ...efficient multidimensional indexing
- ▶ **Incremental updating** of the models
 - (vs. building models from scratch)
 - ...crucial for streaming data
- ▶ **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)

...guide to Big-Data algorithmics

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



Applications

Application: Recommendation

BP Reverts to Containing Oil Spill After Plugging Effort Fails - Bloomberg.com - Mozilla Fire...

File Edit View History Bookmarks Tools Help

Most Visited Getting Started Latest Headlines Readability Like Button

B Bloomberg.com Try the app Bloomberg.com BloombergAnywhere BloombergProfessional AboutBloomberg Updated: New York, May 30 16:59 London, May 30 23:59 Tokyo, May 31 07:59

BP Reverts to Containing Oil Spill Aft... +

Bloomberg.com Try the app Bloomberg.com BloombergAnywhere BloombergProfessional AboutBloomberg Updated: New York, May 30 16:59 London, May 30 23:59 Tokyo, May 31 07:59

Symbol, Site, News Search

HOME NEWS MARKET DATA PERSONAL FINANCE TV and RADIO BUSINESSWEEK BUSINESS EXCHANGE

news STORY PHOTO

BP Reverts to Containing Oil Spill After Plugging Effort Fails

By Jim Polson and David Wethe Share | Email | Print | AAA

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 undersurface 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 its flow, Jason Kenney, an Edinburgh-based analyst at ING Commercial Banking, said 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

MULTILINGUAL CUSTOMER SUPPORT 24 HOURS

ExPro

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-Bite

► Good recommendations can make a big difference when keeping a user on a web site

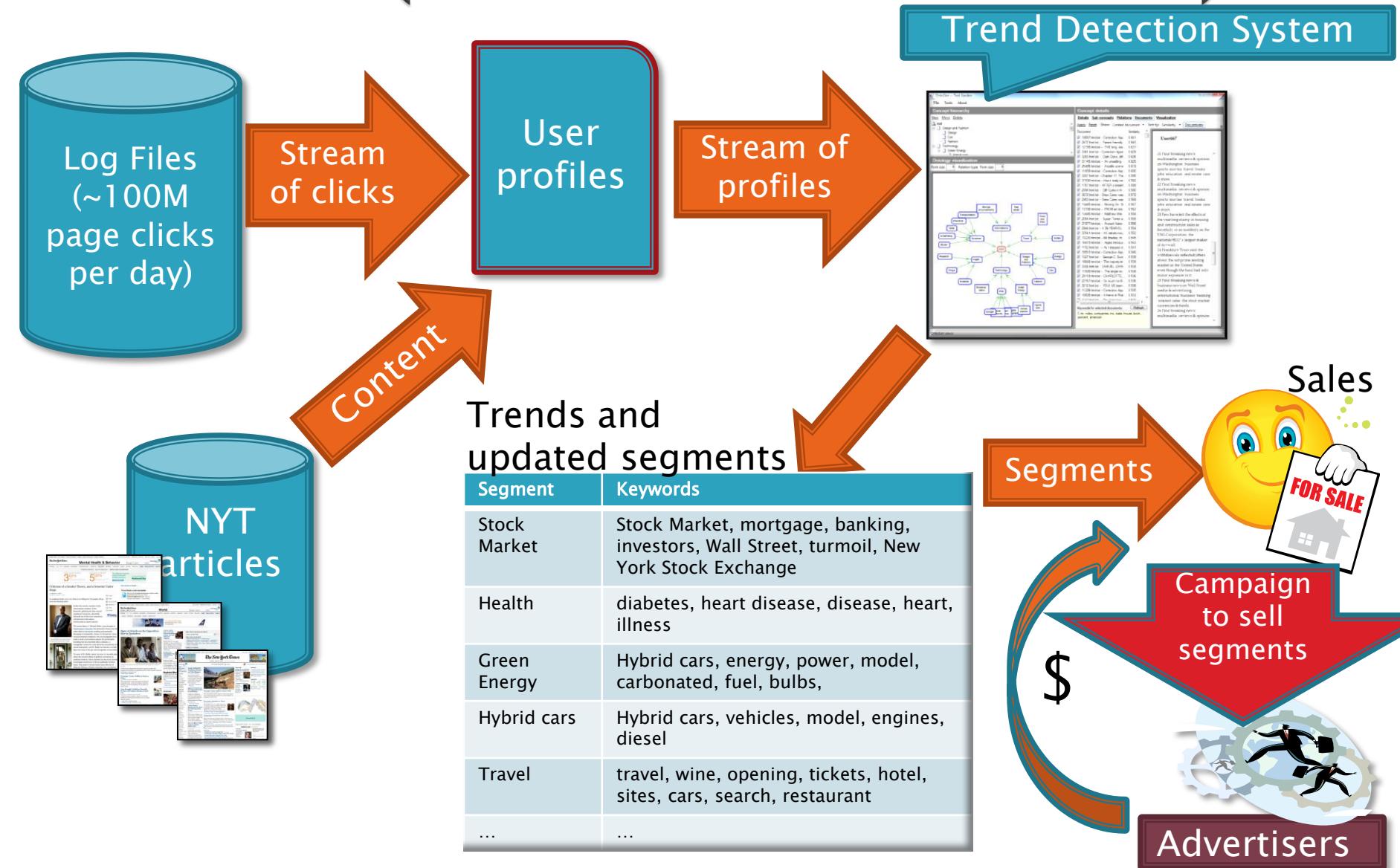
- ...the key is how rich the context model a system is using to select information for a user
- Bad recommendations <1% users, good ones >5% users click
- 200clicks/sec

Contextual personalized recommendations generated in ~20ms

The context of each click on the web site used for recommendation

- ▶ 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
- ▶ GeolP Country
- ▶ GeolP State
- ▶ GeolP 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

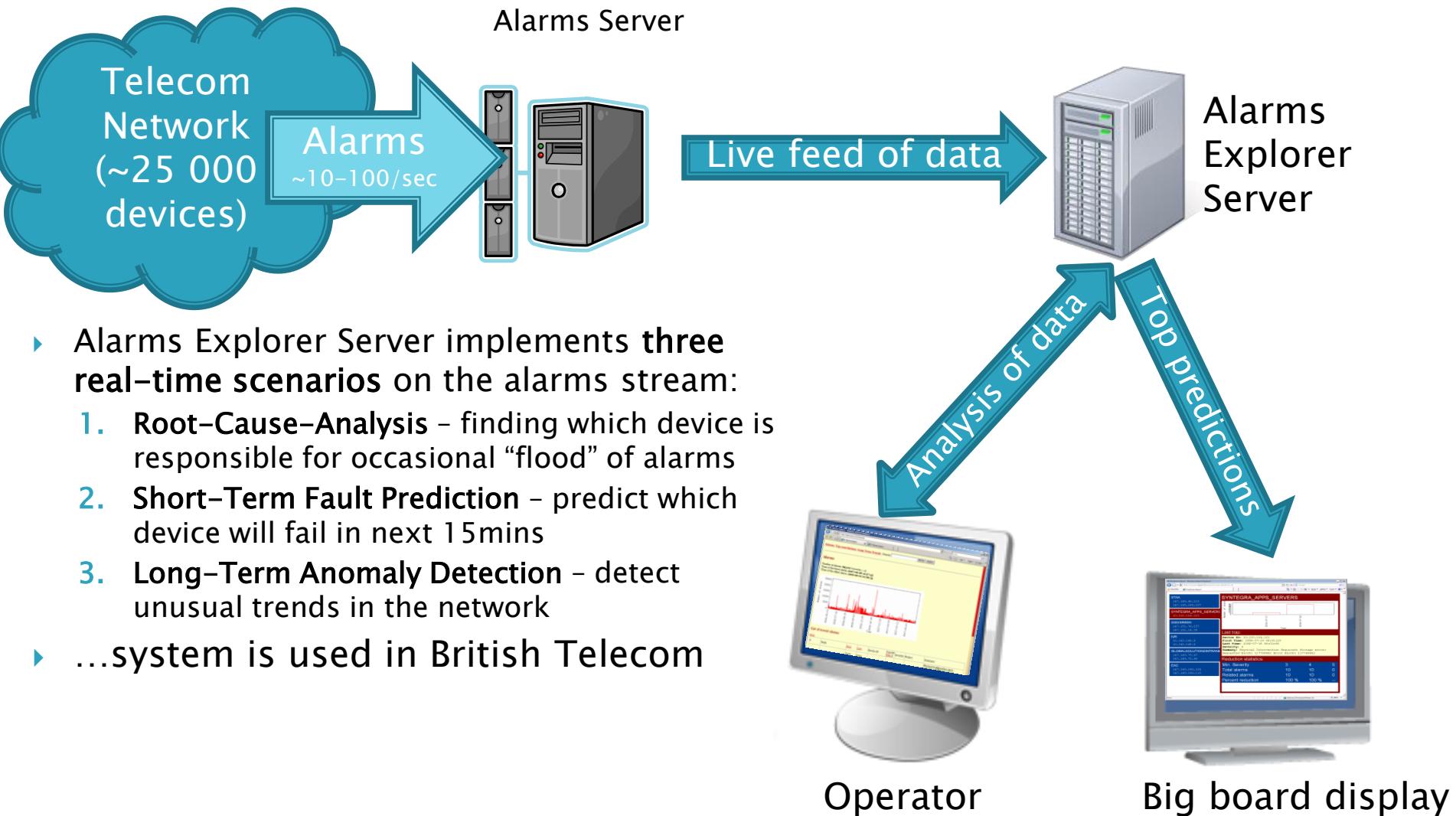
Application: Online Advertising for NYTimes (microtrends detection)



Scale of one day of NYTimes data

- ▶ 50Gb of uncompressed log files
- ▶ 50–100M clicks
- ▶ 4–6M unique users
- ▶ 7000 unique pages with more than 100 hits

Application: Telecommunication Network Monitoring



Application: Monitoring global main stream news

- ▶ The aim of the project is to collect and analyze most of the main-stream media across the world
 - ...from 35,000 publishers (180K RSS feeds) crawled in real time (few ~10articles per second)
 - ...each article document gets extracted, cleaned, semantically annotated, structure extracted
- ▶ Challenges are in terms of complexity of processing and querying of the extracted data
 - ...and matching textual information across the languages (cross-lingual technologies)

Firefox ▾

http://newsfeed.ijs.si/visual_demo/ +

newsfeed.ijs.si/visual_demo/ Google

Most Visited Real-time News Reco... Real-Time Insights Fin... Research Participant P... Selerity • Home RapiData LLC - Welco... Event-Driven Architect... GSN Bookmarks

Real-time newsfeed demo

Since this page was opened: 471 articles received, 232 skipped for legibility.

#47665170 @ 2012-05-07 18:38:00 (UTC) by flwoutdoors.com
Sullivan grinds it out on Fort Gibson Lake

#47669656 @ 2012-05-08 05:11:40 (UTC) by tt.bernerzeitung.ch
Ab 23. Mai wird im Bernaqua wieder gebadet

#47666591 @ 2012-05-08 00:32:00 (UTC) by hoy.co.jp
Eco menu

#47668122 @ 2012-05-08 11:22:46 (UTC) by wesh.com
Lawmer: Bomb Plot Shows New Level Of Sophistication

#47657268 @ 2012-05-08 11:41:37 (UTC)* by morgenpost.de
Berliner Arzt zur Behandlung von Timoschenko in Charkow

#47669984 @ 2012-05-07 18:00:00 (UTC) by frankston-leader.wherelive.com.au
Evidence points to an arresting day in Frankston

#47665623 @ 2012-05-07 10:00:00 (UTC) by ads.pheedo.com
Presented By:

#47669009 @ 2012-05-07 22:00:00 (UTC) by gundem.milliyet.com.tr
Suriye'de kaybolan Türk gazetecilerden haber var

#47667995 @ 2012-05-08 01:42:38 (UTC) by localnews8.com
IRS Forms Show Charity's Money Isn't Going To Disabled Vets

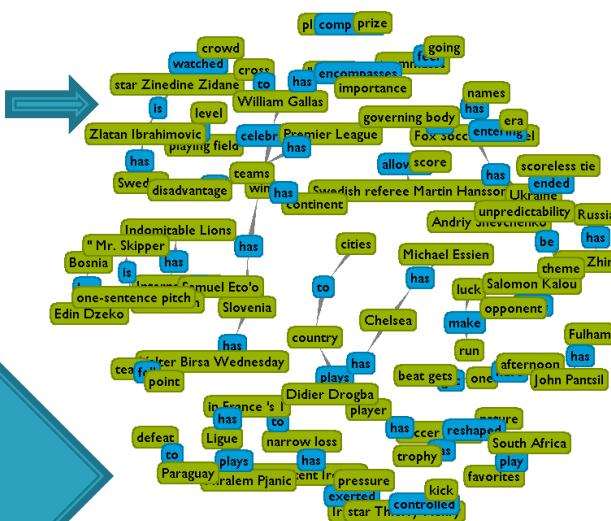
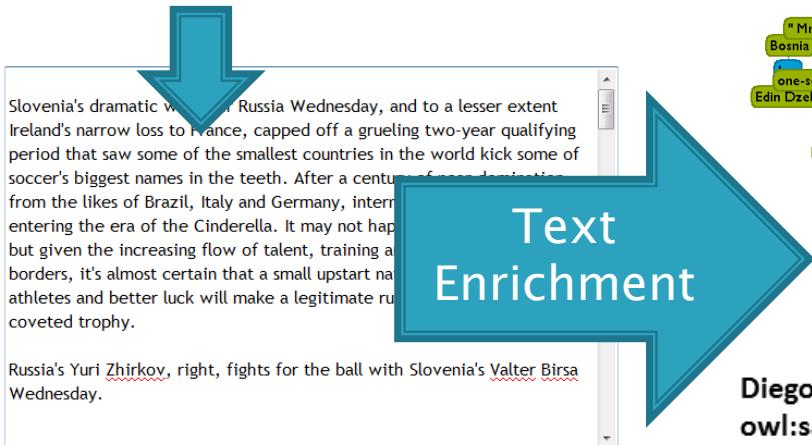
#47669161 @ 2012-05-07 01:37:42 (UTC) by valldaily.com
Ask Waste Watchers: How to recycle batteries

#47667226 @ 2012-05-08 01:07:00 (UTC) by freep.com
'Lost' star Matthew Fox arrested for DUI in

Map Satellite

Google Imagery ©2012 NASA, TerraMetrics - Terms of Use

Semantic text enrichment (DBpedia, OpenCyc, ...) with Enrycher (<http://enrycher.ijcns.si/>)



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](#)
- [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/ArgentinianInternationalFootballers>
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>

Application: Text visualization

- ▶ The aim is to use analytic techniques to visualize documents in different ways:
 - Topic view
 - Social view
 - Temporal view

Topic landscape of the query “Clinton” from Reuters news 1996–1997

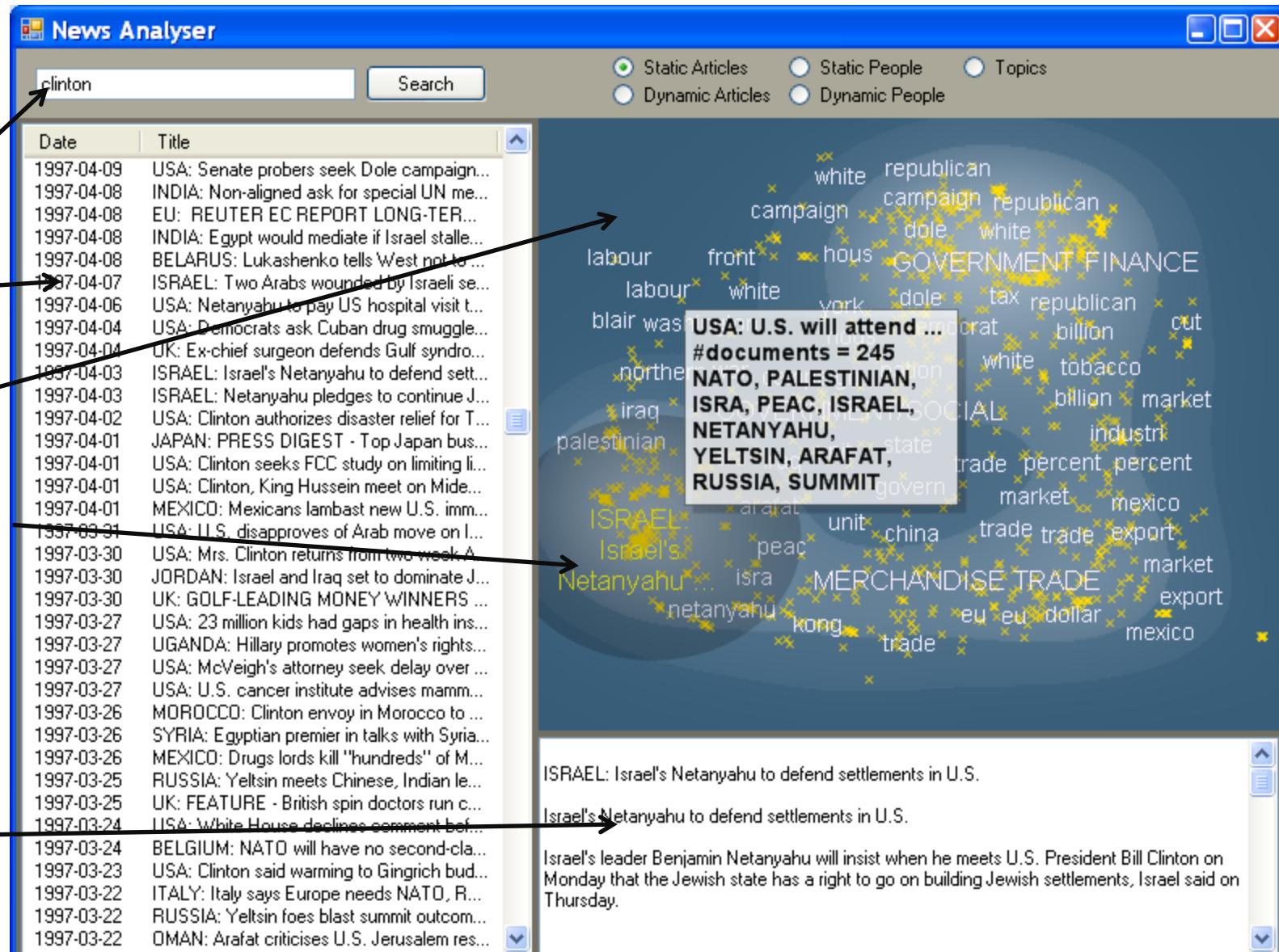
Query

Search Results

Topic Map

Selected group of news

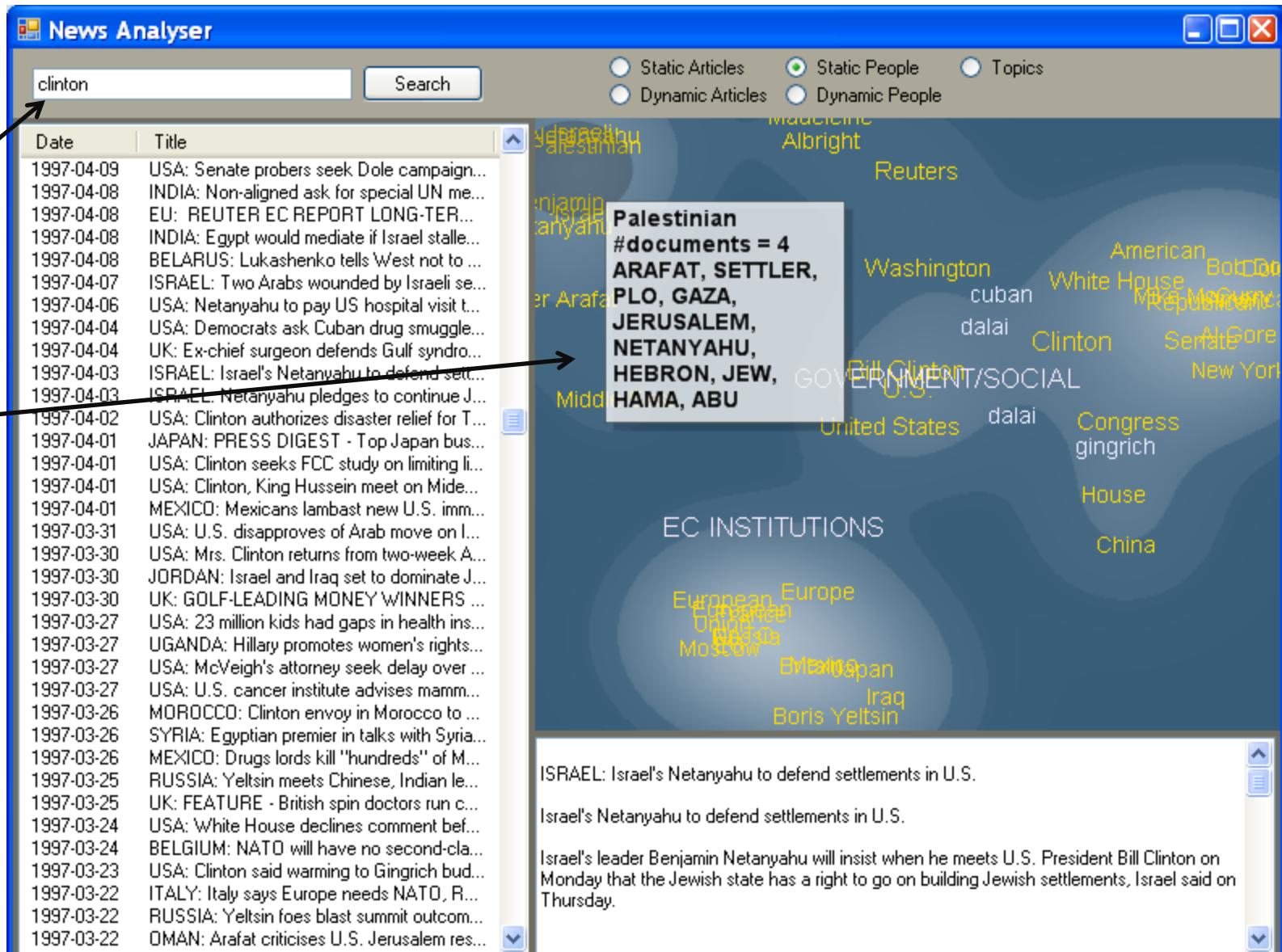
Selected story



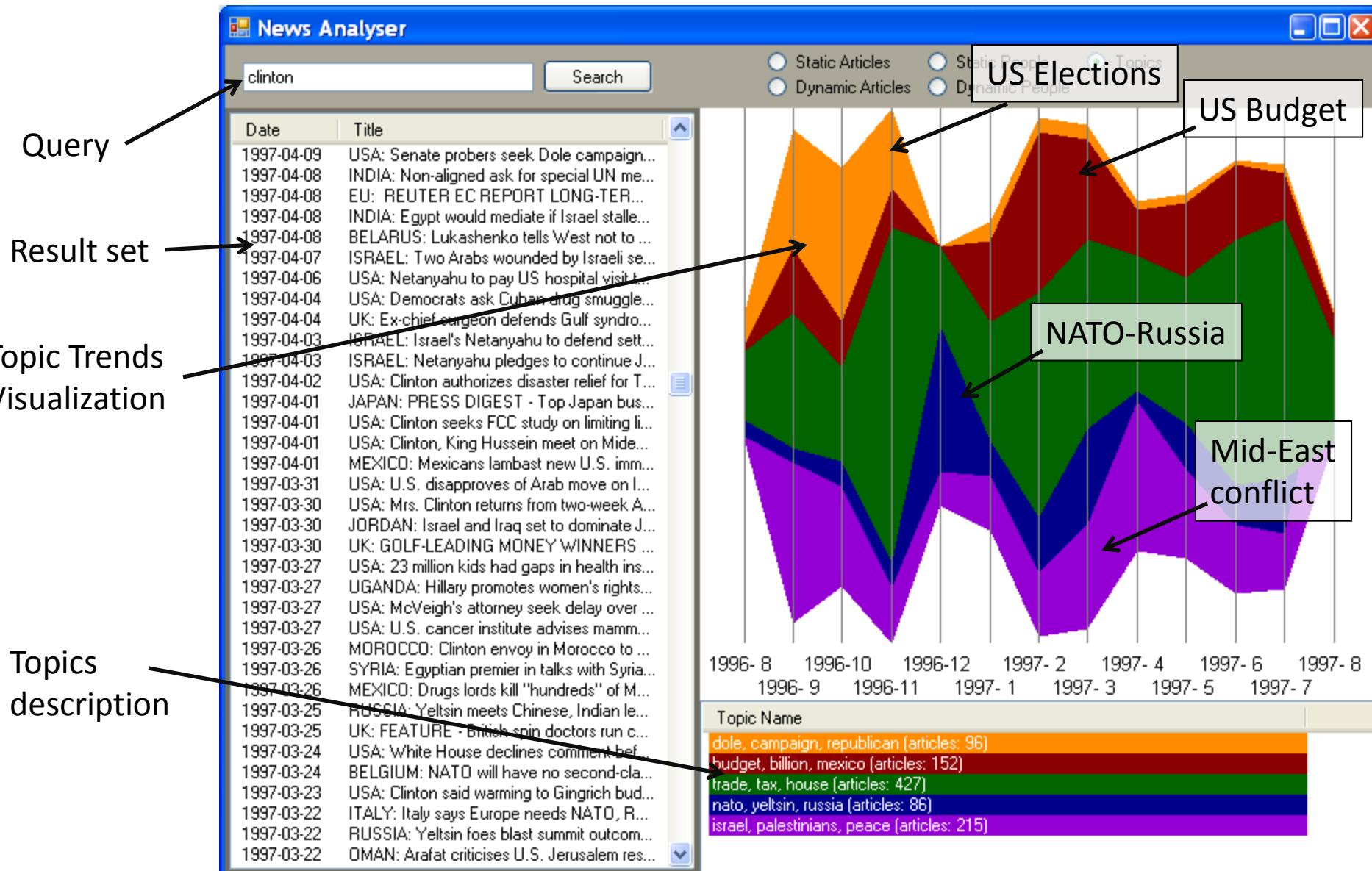
Visualization of social relationships between “Clinton” and other entities

Query

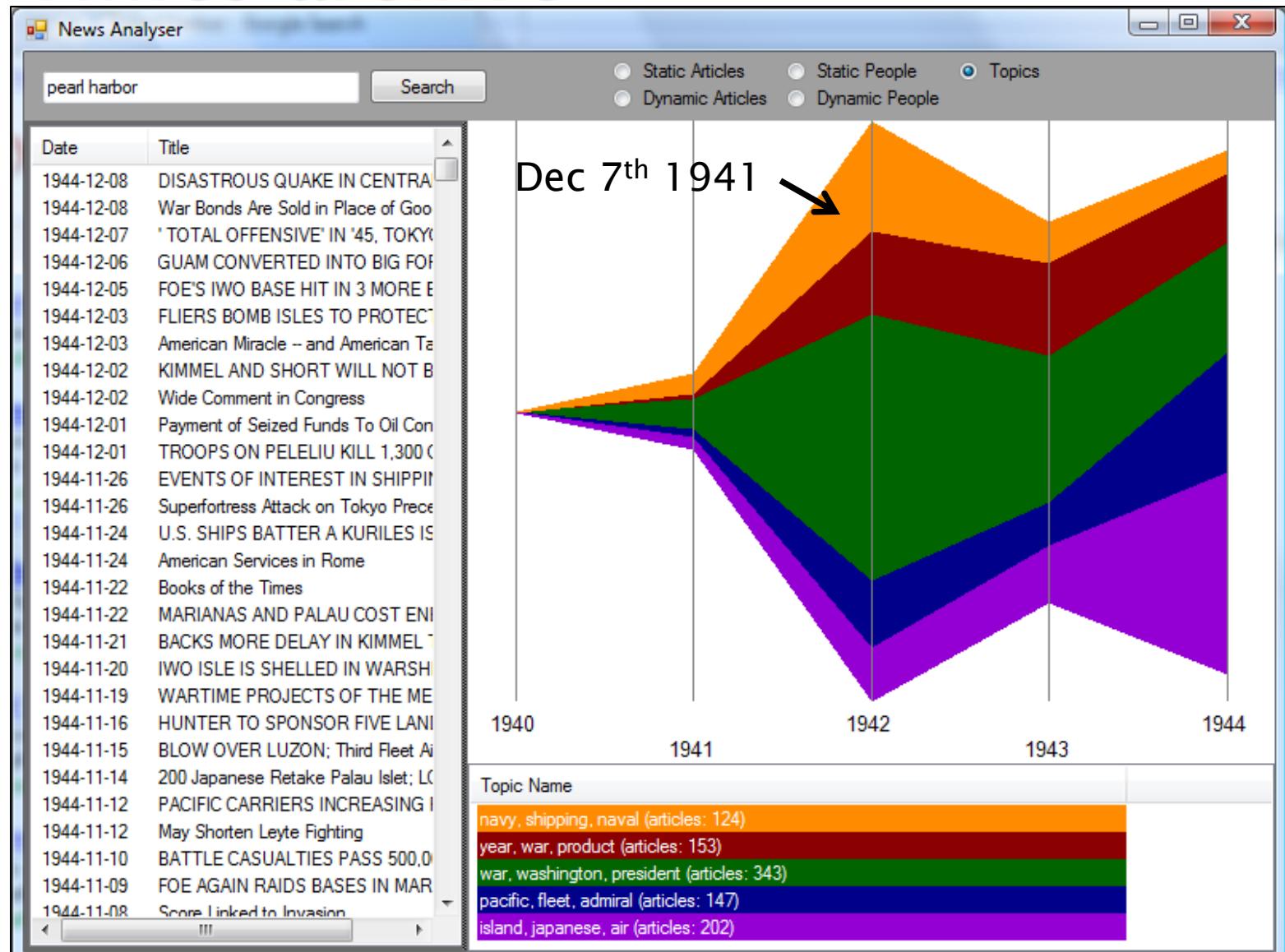
Named entities in relation



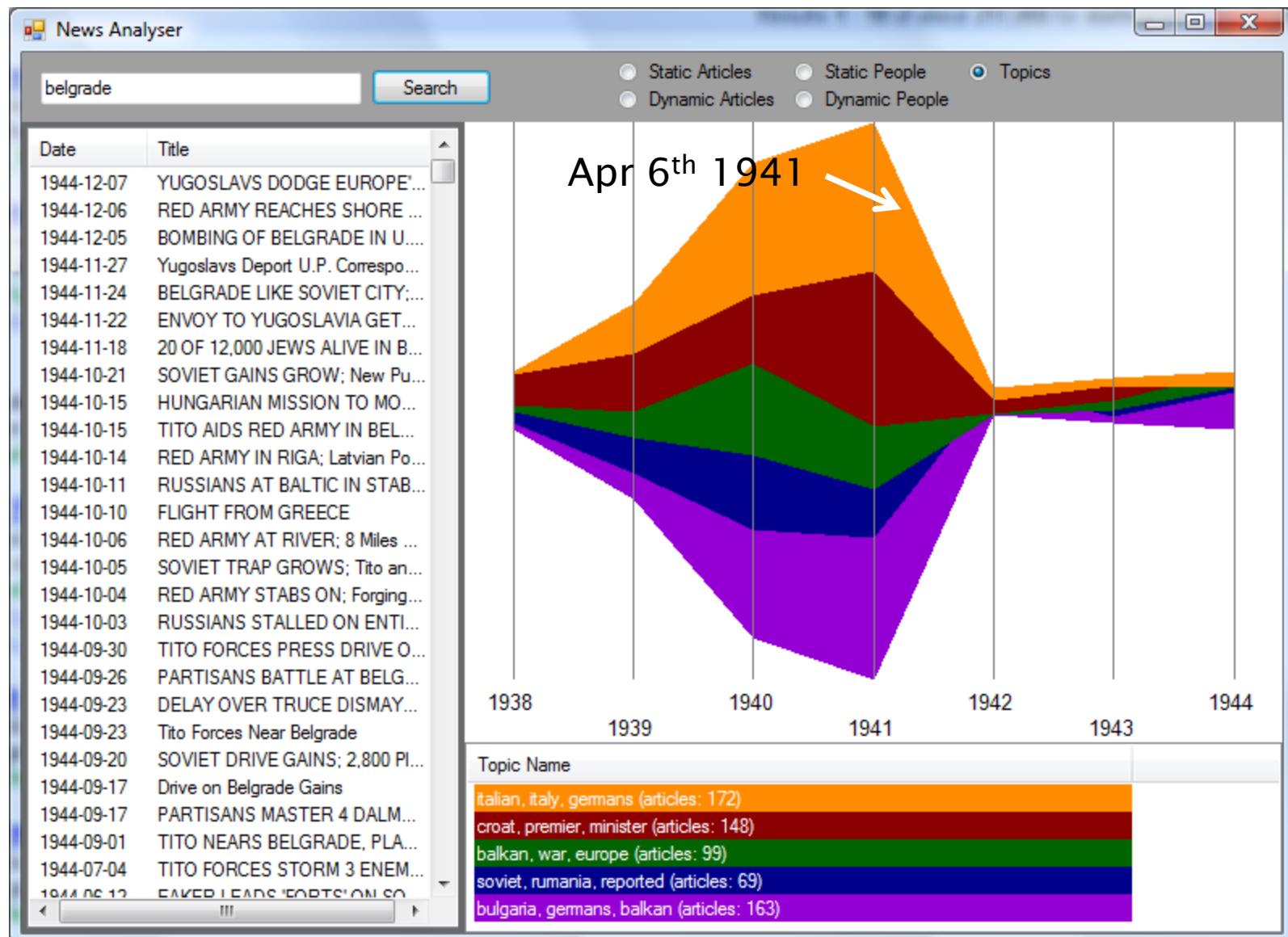
Topic Trends Tracking of the documents including “Clinton”



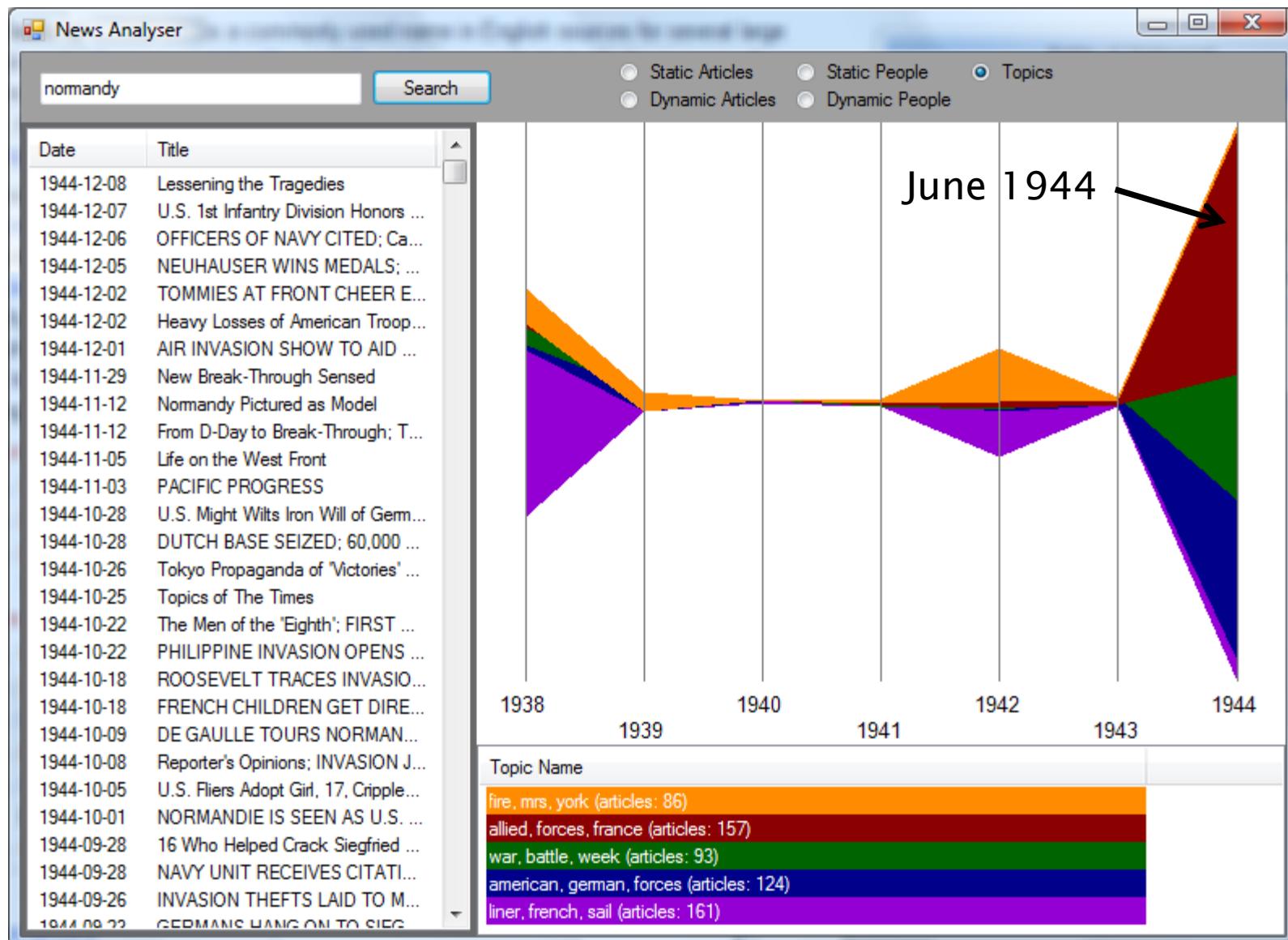
WW2 query “Pearl Harbor” into NYTimes archive



WW2 query “Belgrade” into NYTimes archive



WW2 query “Normandy” into NYTimes archive



Application: Context sensitive search ranking: <http://searchpoint.ijs.si>

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point

The screenshot shows a Windows Internet Explorer window with the title 'jaguar - SearchPoint - Windows Internet Explorer'. The address bar contains 'http://searchpoint.ijs.si/Result.aspx'. The search query 'jaguar' is entered in the search bar. Below the search bar are three buttons: 'Search via topics', 'Search via query to ontology', and 'Search via hits to ontology'. The main content area displays a list of search results:

- (9) [Jaguar](http://www.abf90.dial.pipex.com/jaguar.htm)
General information and facts from Big Cats Online.
<http://www.abf90.dial.pipex.com/jaguar.htm>
- (59) [Jaguar, Jaguar Profile, Facts, Information, Photos, Pictures ...](http://animals.nationalgeographic.com/animals/mammals/jaguar.html)
Get jaguar profile, facts, information, photos, pictures, sounds, habitats, reports, news, and more from National Geographic.
<http://animals.nationalgeographic.com/animals/mammals/jaguar.html>
- (0) [Jaguar - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Jaguar)
The jaguar (*Panthera onca*) is a New World mammal of the Felidae family and one of four "big cats" in the *Panthera* genus, along with the tiger, ...
<http://en.wikipedia.org/wiki/Jaguar>
- (11) [Jaguar](http://www.bigcatrescue.org/jaguar.htm)
Jaguar Facts, Jaguar Photos and Jaguars in the news at the world's largest big cat rescue and sanctuary.
<http://www.bigcatrescue.org/jaguar.htm>
- (1) [Jaguar](http://www.jaguar.com/)
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets.
<http://www.jaguar.com/>
- (32) [Jaguar](http://www.abf90.dial.pipex.com/bco/jaguar.htm)
Contains extensive information about the Jaguar. Information includes habitat, body size, and life span.
<http://www.abf90.dial.pipex.com/bco/jaguar.htm>
- (2) [Jaguar UK - Jaguar Cars](http://www.jaguar.co.uk/)
Jaguar & Ownership. Highlights. Gallery. Models & Pricing. Design Your XK. TEST DRIVE. Brochure. Dealer. eNewsletter ...
<http://www.jaguar.co.uk/>
- (17) [Jaguar Enthusiasts' Club](http://www.jec.org.uk/)
World's largest audited membership. UK-based, JEC's site has extensive resources available for the enthusiast, including information about their Sections, ...
<http://www.jec.org.uk/>
- (20) [San Diego Zoo's Animal Bytes: Jaguar](http://www.sandiegozoo.org/animalbytes/jaguar.html)
Get fun and interesting jaguar facts in an easy-to-read style from the San Diego Zoo's Animal Bytes website.
<http://www.sandiegozoo.org/animalbytes/jaguar.html>

A conceptual map is overlaid on the right side of the results. It features a central node labeled 'Top' with various connections to other nodes: 'Parts and Accessories', 'Vehicles', 'Mammalia', 'Sports', 'NFL', 'Enthusiasts', 'Recreation', 'Society', 'Games', 'Console Platforms', 'Aviation', 'Aircraft', 'Enthusiasts', 'Sports', 'NFL', 'Shopping', 'France', and 'Top'. The word 'Top' is highlighted in red.

Application: Analysis of MSN–Messenger Social–network

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

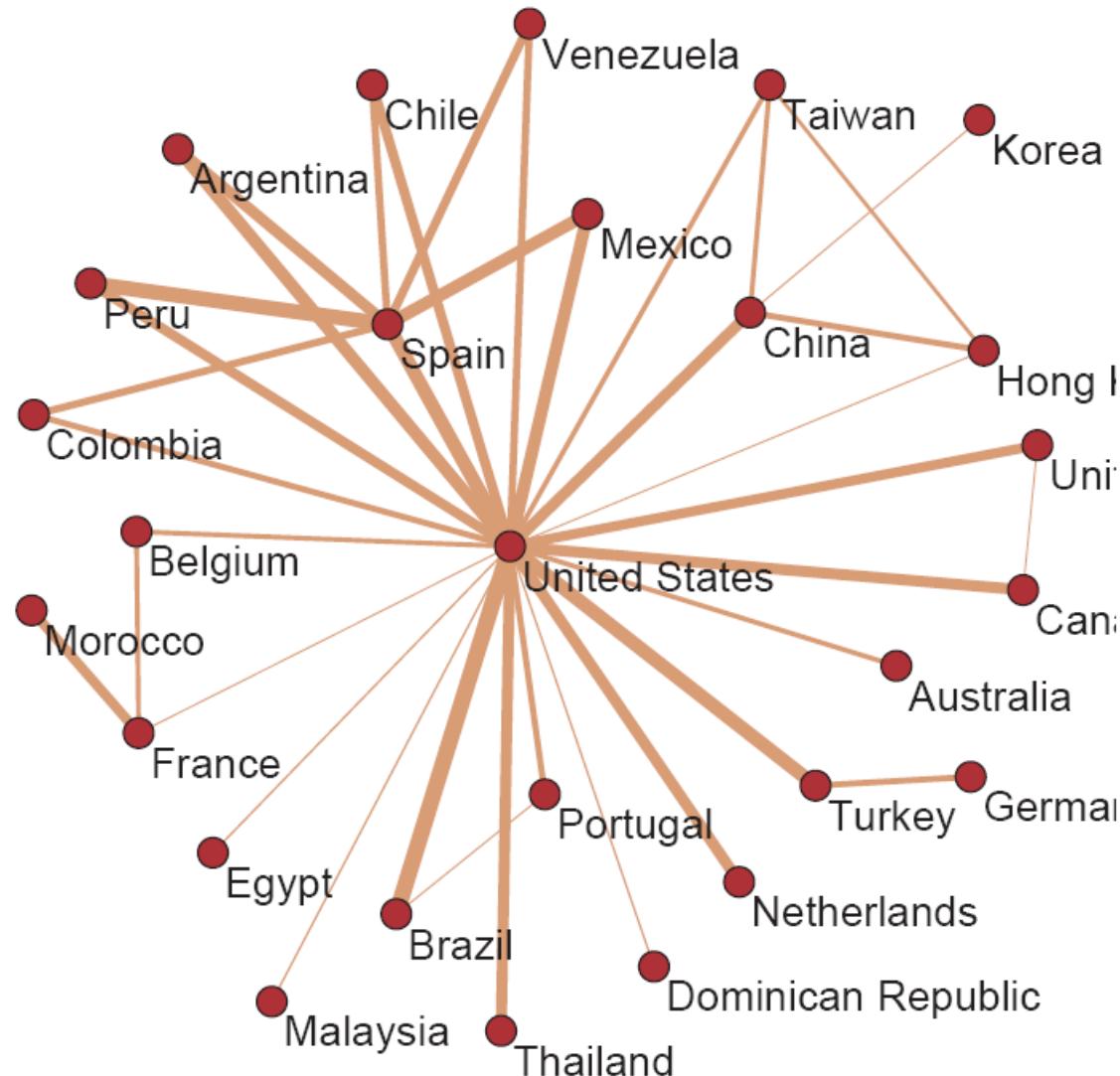
Research questions:

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

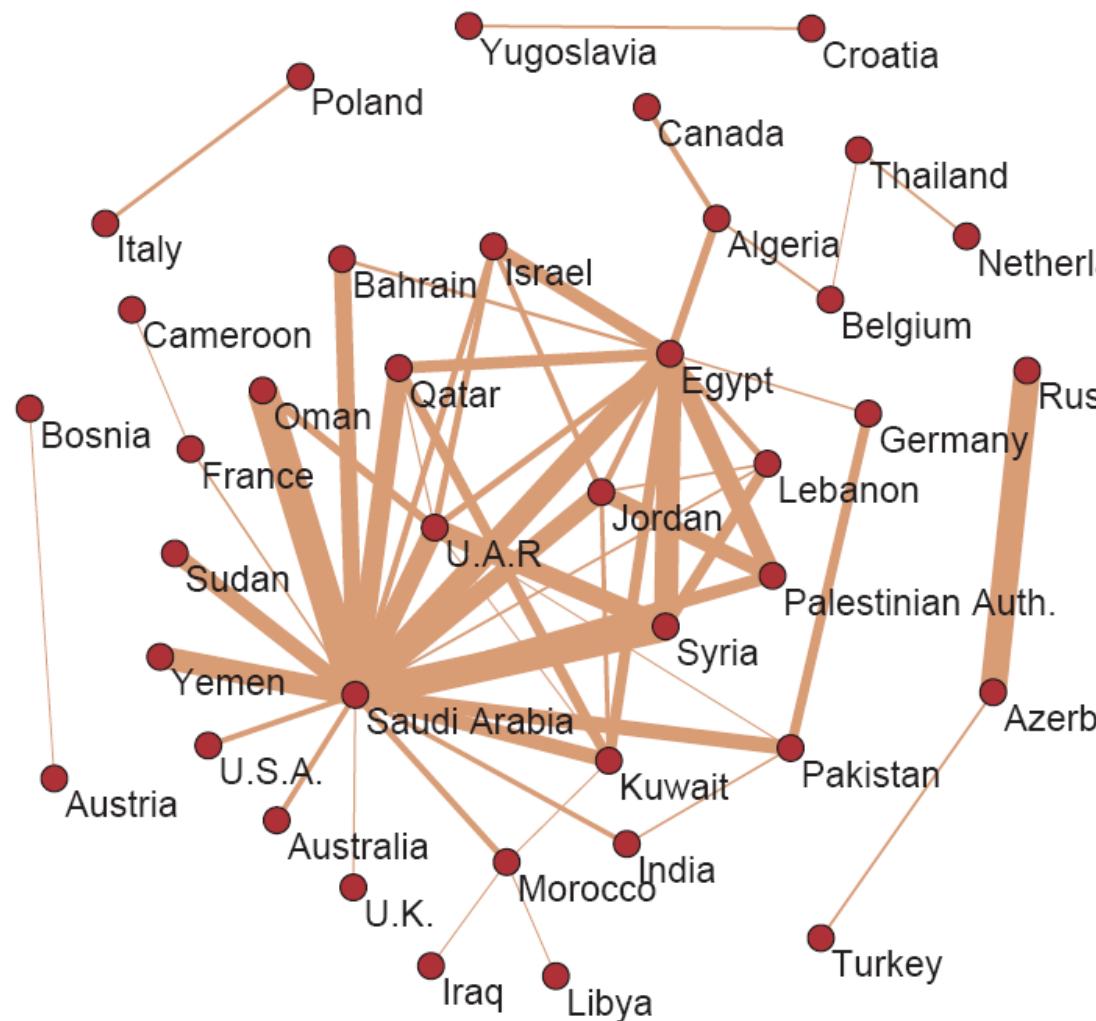
Data statistics: Total activity

- ▶ 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

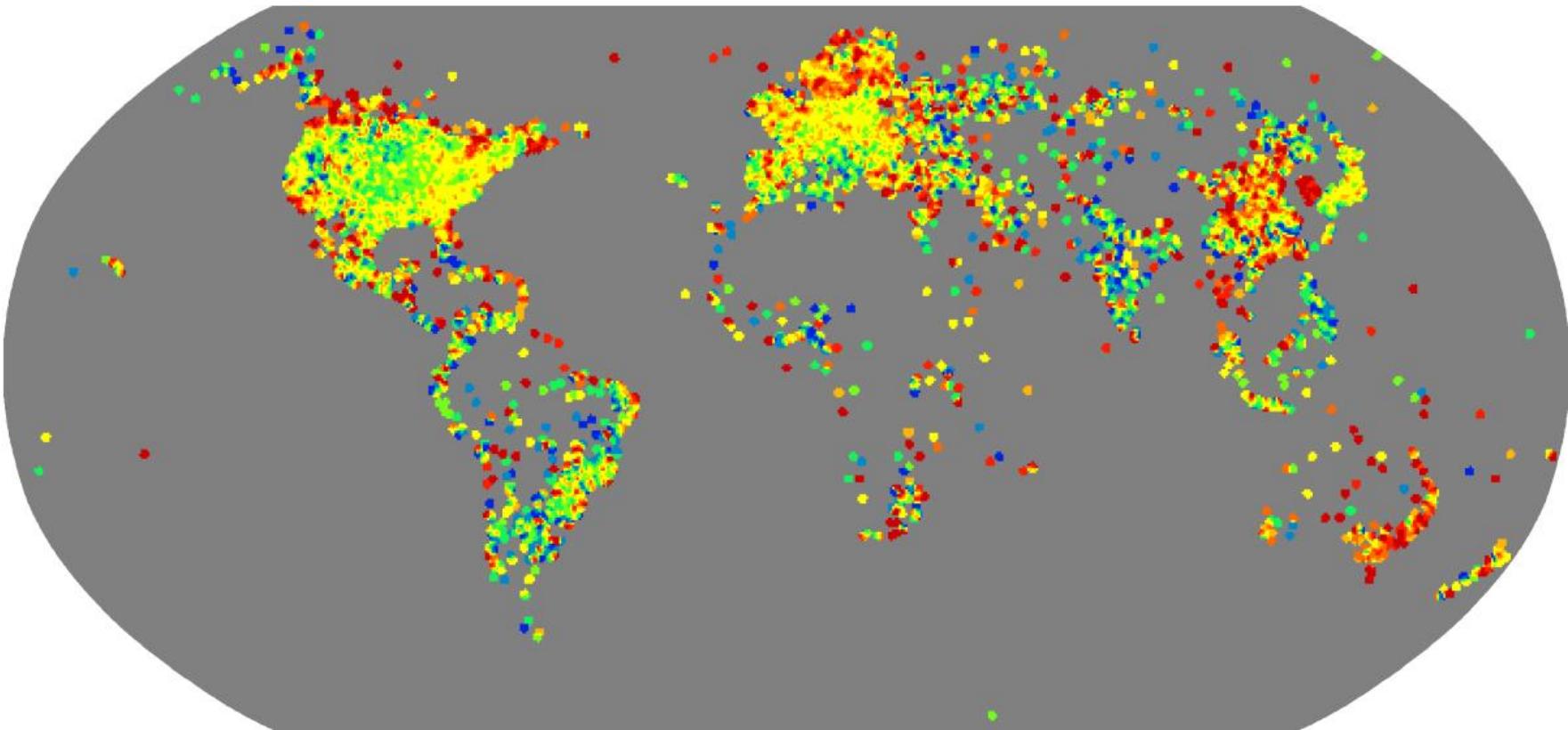
Who talks to whom: Number of conversations



Who talks to whom: Conversation duration



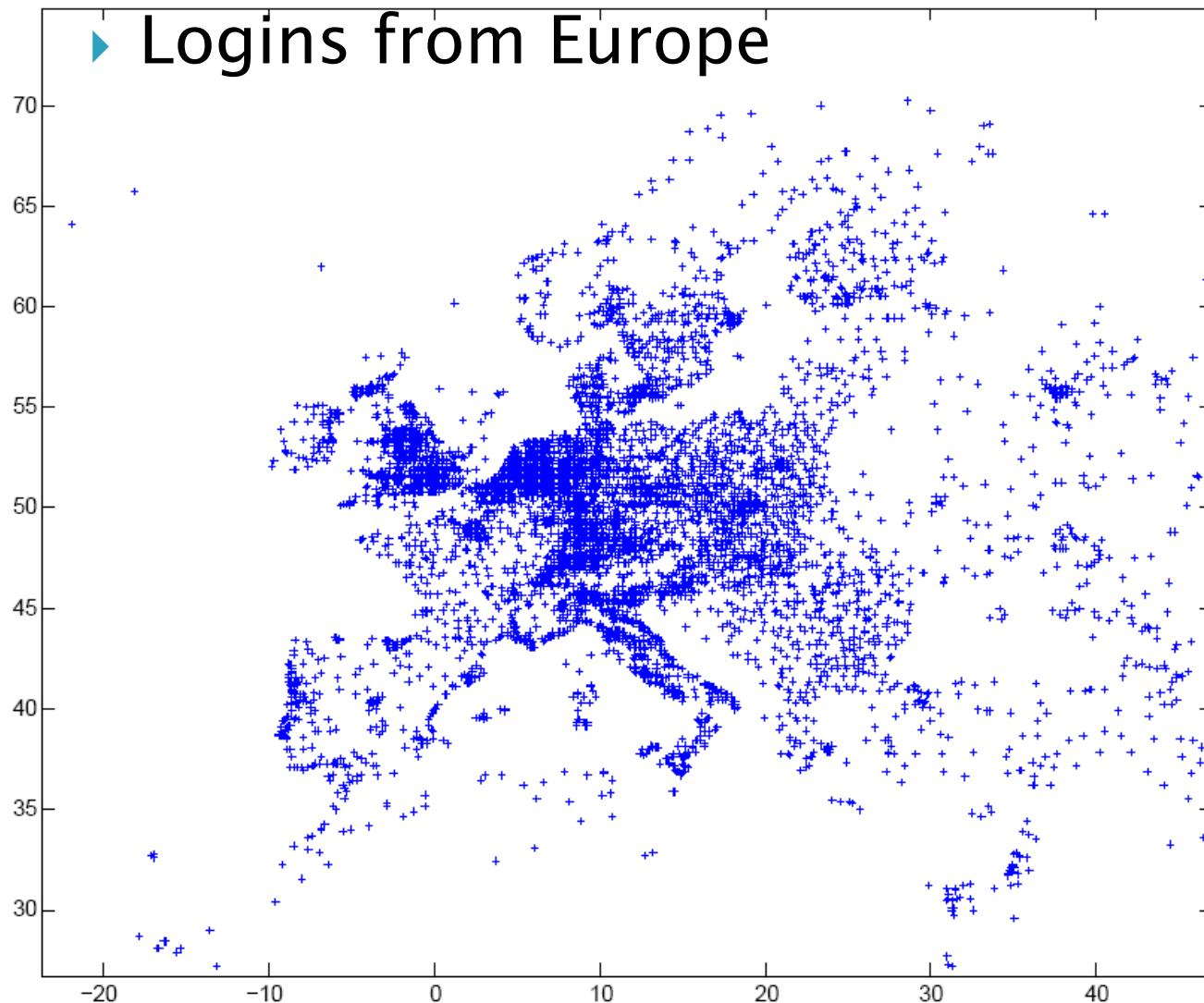
Geography and communication



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

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

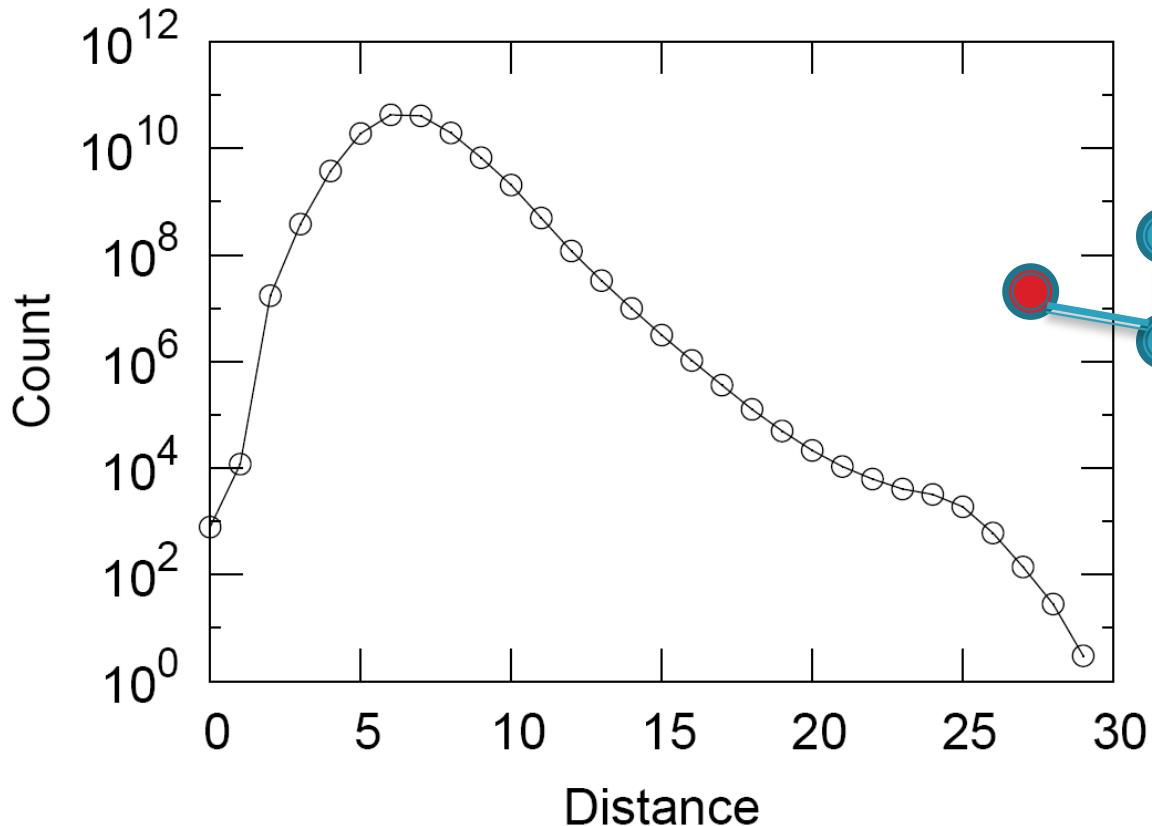
How is Europe talking



Hops Nodes

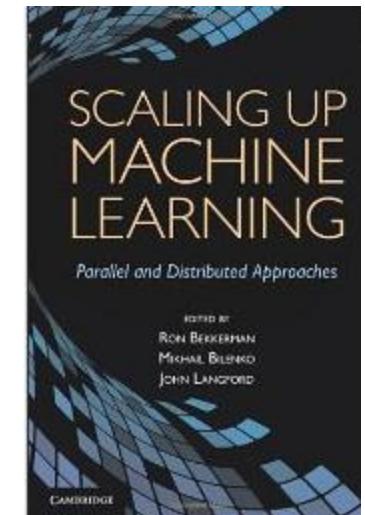
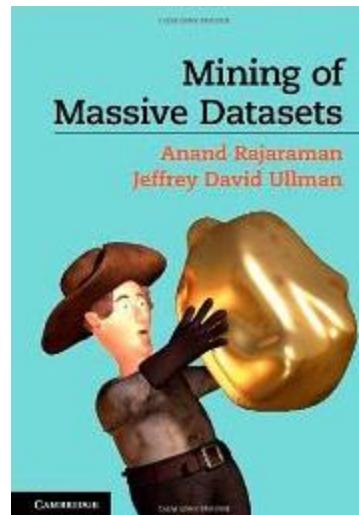
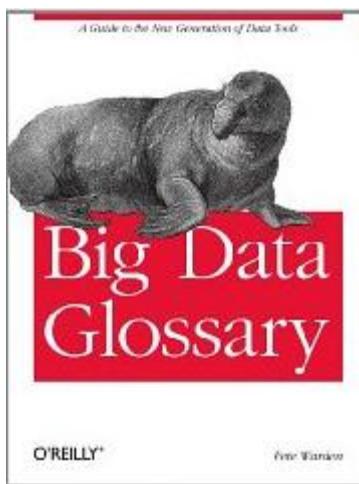
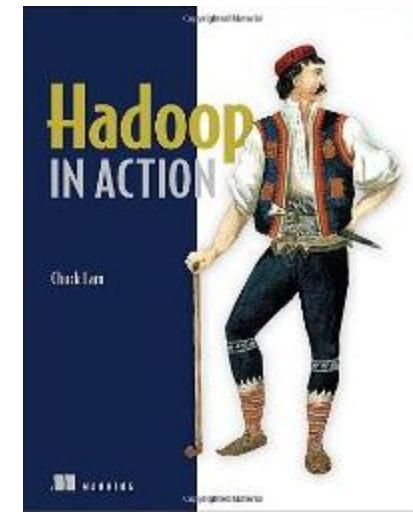
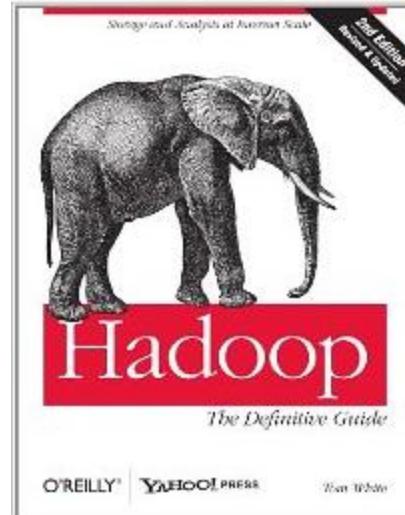
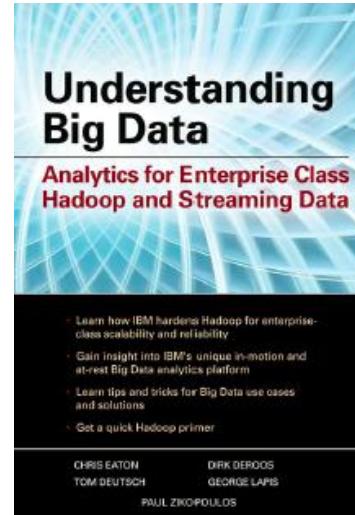
1	10
2	78
3	396
4	8648
5	3299252
6	28395849
7	79059497
8	52995778
9	10321008
10	1955007
11	518410
12	149945
13	44616
14	13740
15	4476
16	1542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Network: Small-world



- ▶ 6 degrees of separation [Milgram '60s]
- ▶ Average distance between two random users is 6.6²¹
- ▶ 90% of nodes can be reached in < 8 hops

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)