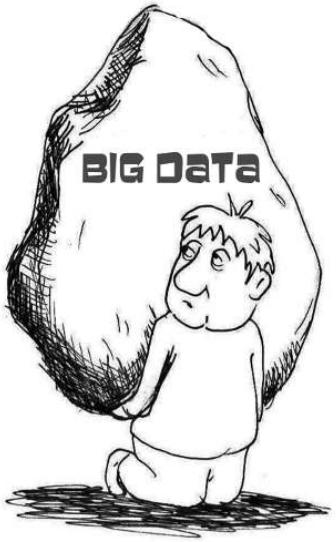


Tutorial: Introduction to Big Data

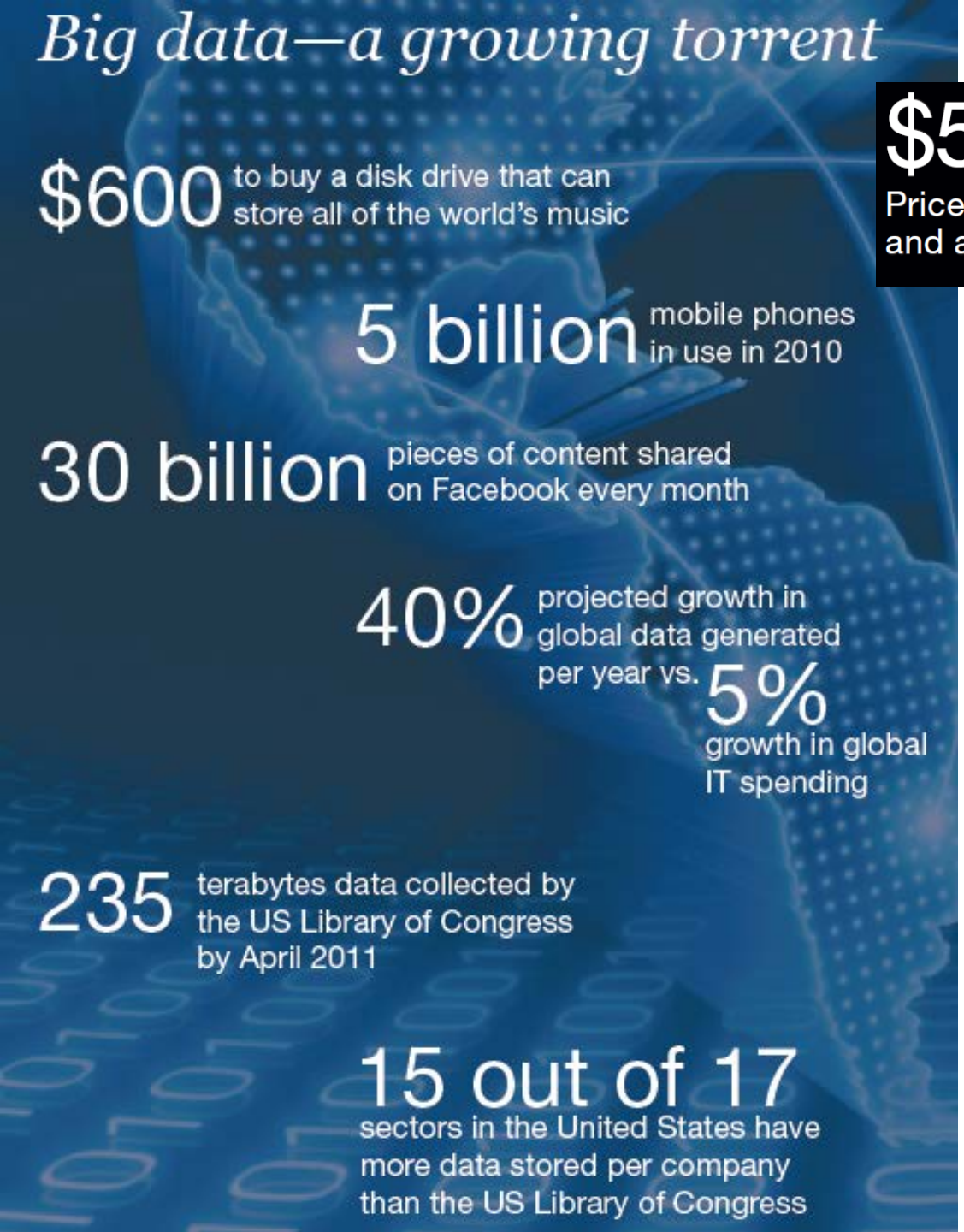
Marko Grobelnik

Luka, Bradesko, Blaz Fortuna, Blaz Fortuna, Gregor Leban

Jozef Stefan Institute, Slovenia

Ljubljana, Feb 22 2017

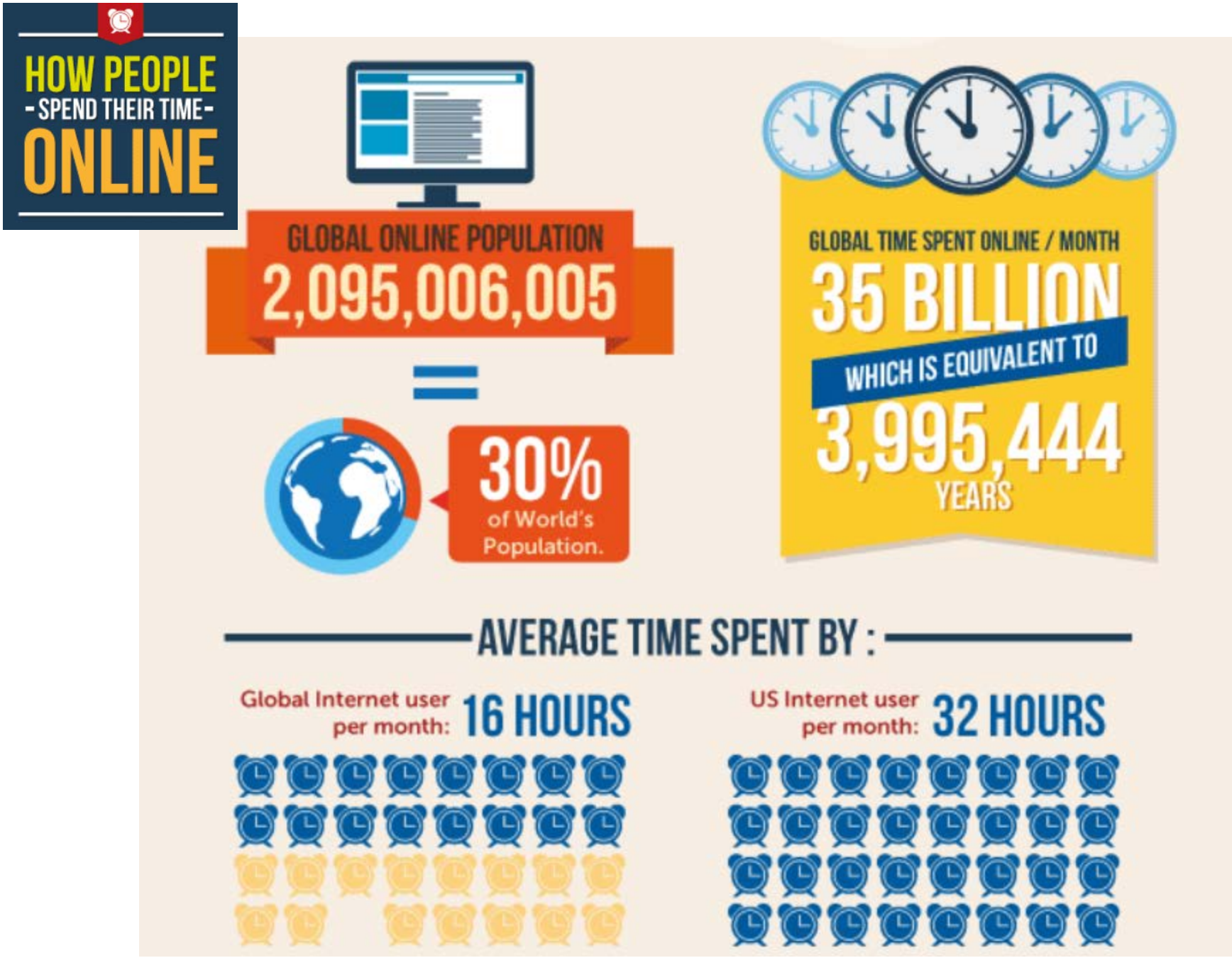
Big-Data in numbers



$5 million vs. $400

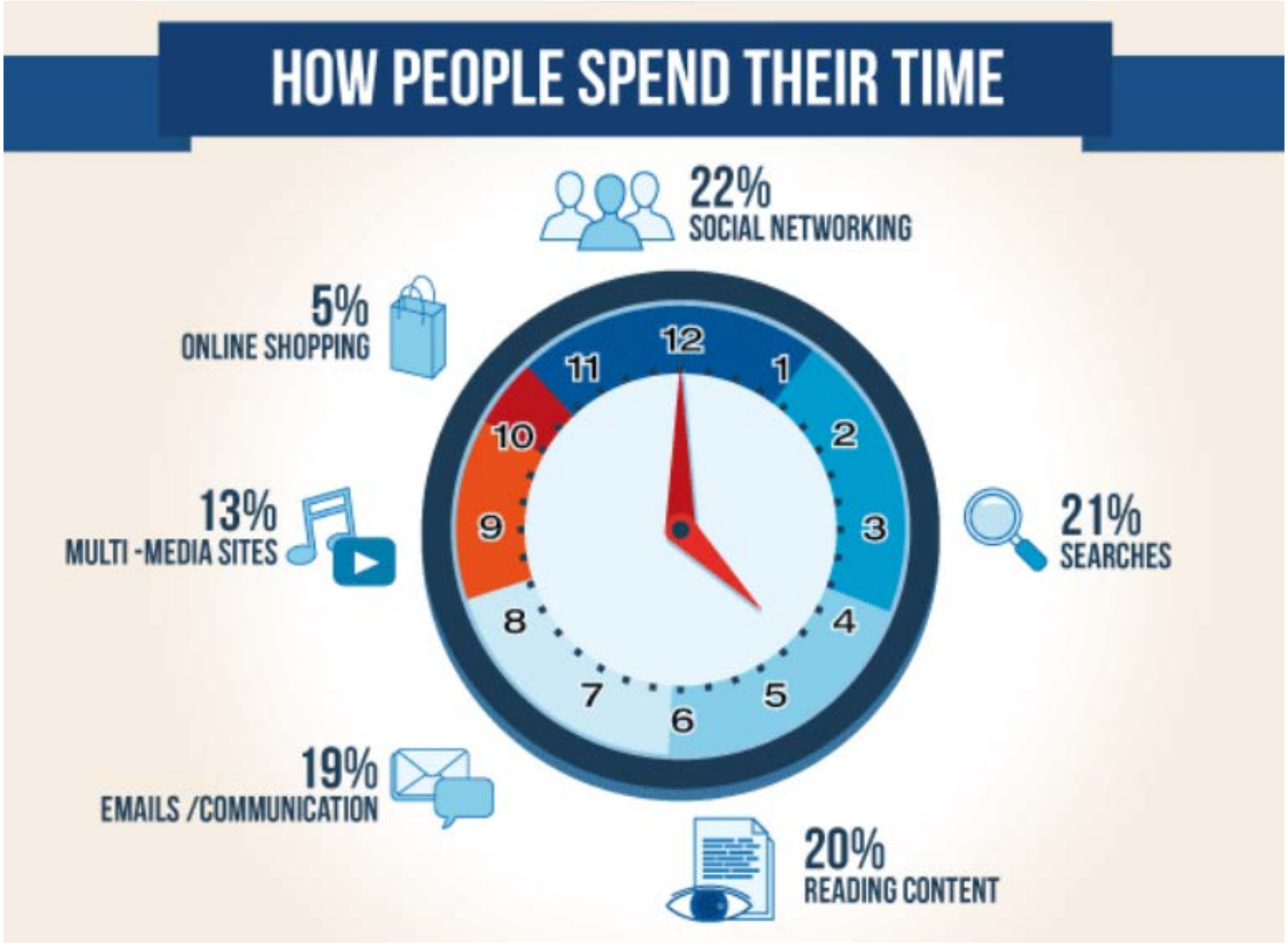
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance


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
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
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/

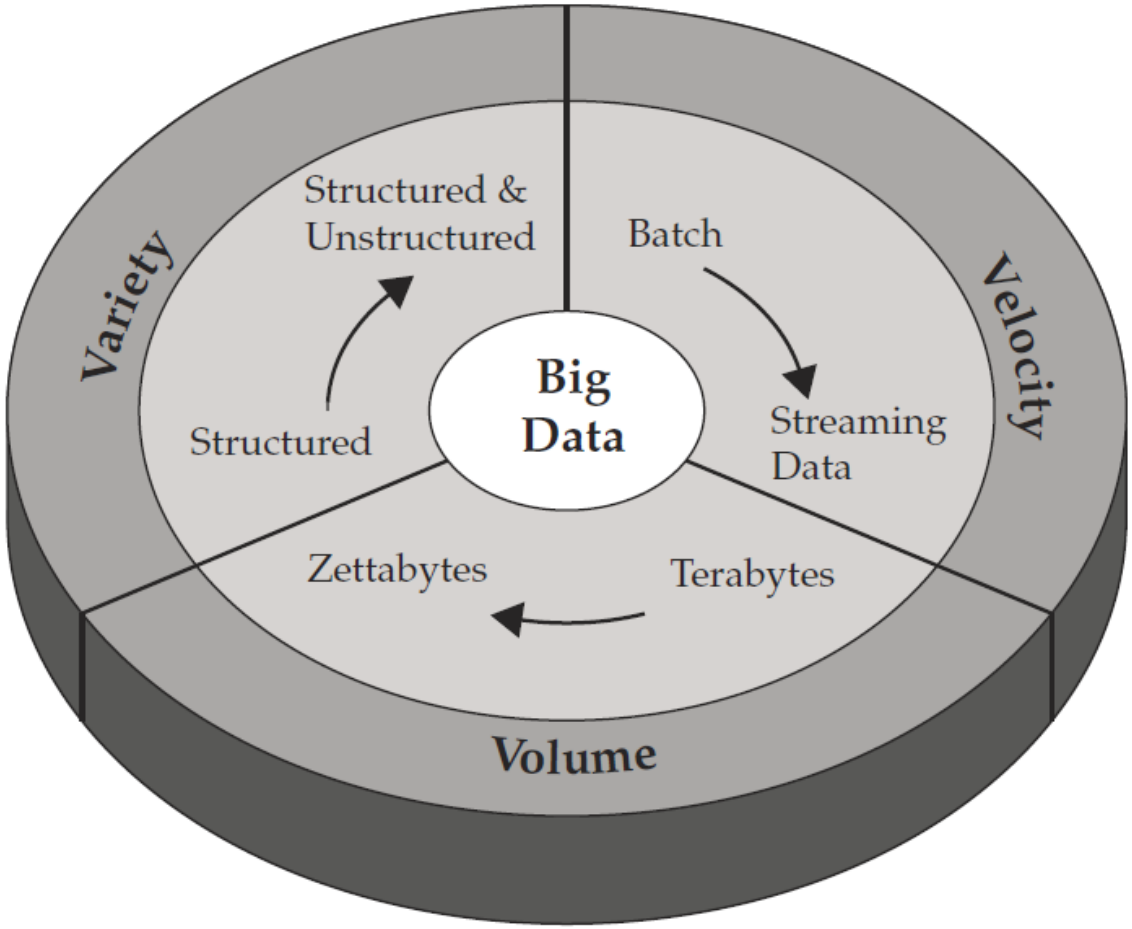
Big-Data Definitions

…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

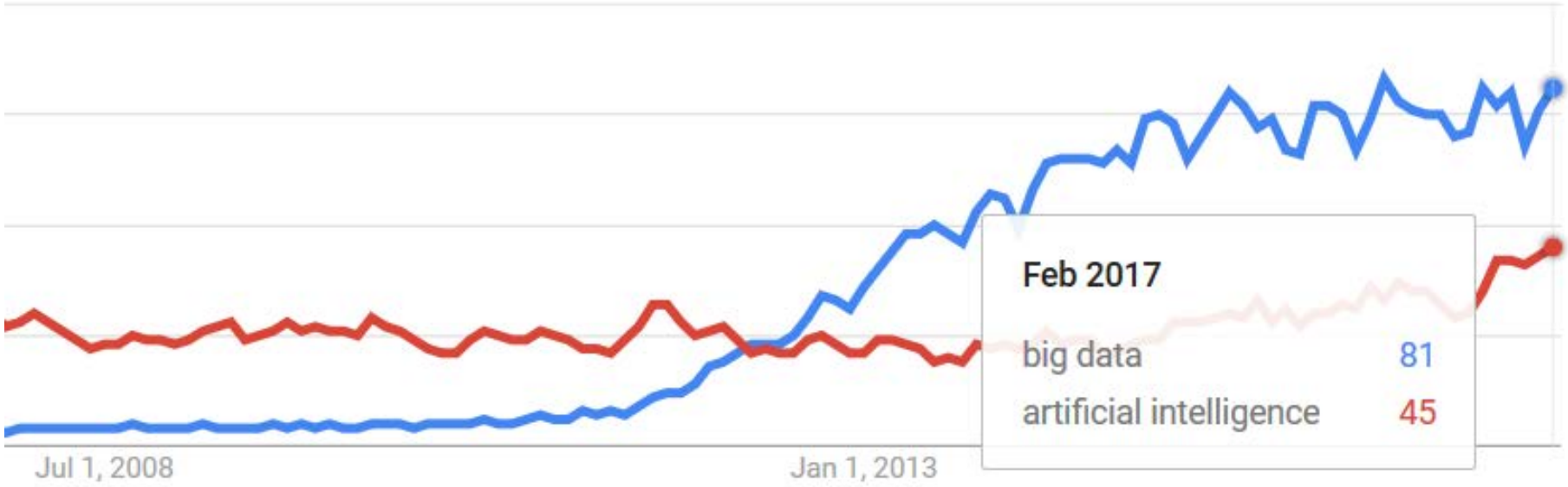
- 1. **Volume** (lots of data = "Tonnabytes")
- 2. **Variety** (complexity, curse of dimensionality)
- 3. **Velocity** (rate of data and information flow)

- 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")

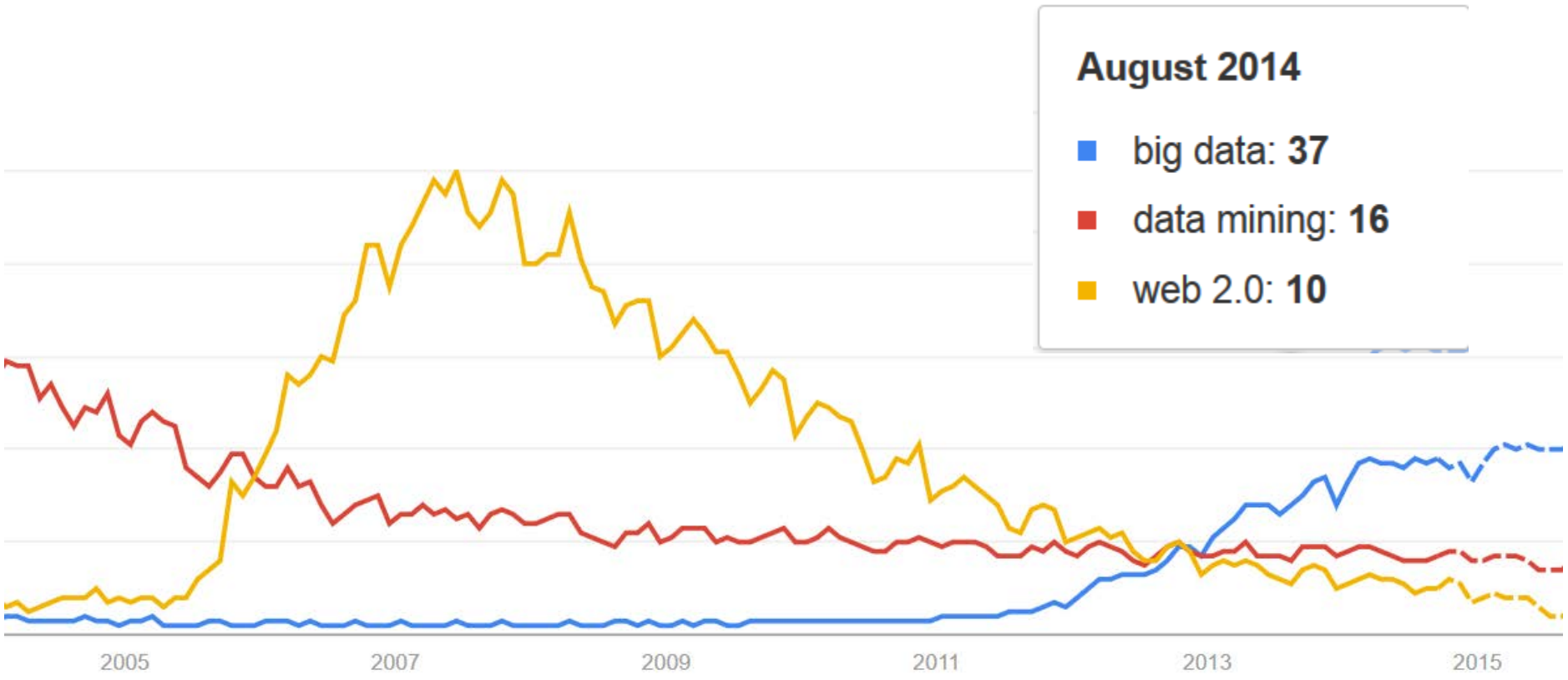
Comparing volume of "big data" and "artificial intelligence" queries



https://trends.google.com/trends/explore?date=all&q=big%20data,artificial%20intelligence

…but what can happen with “hypes”

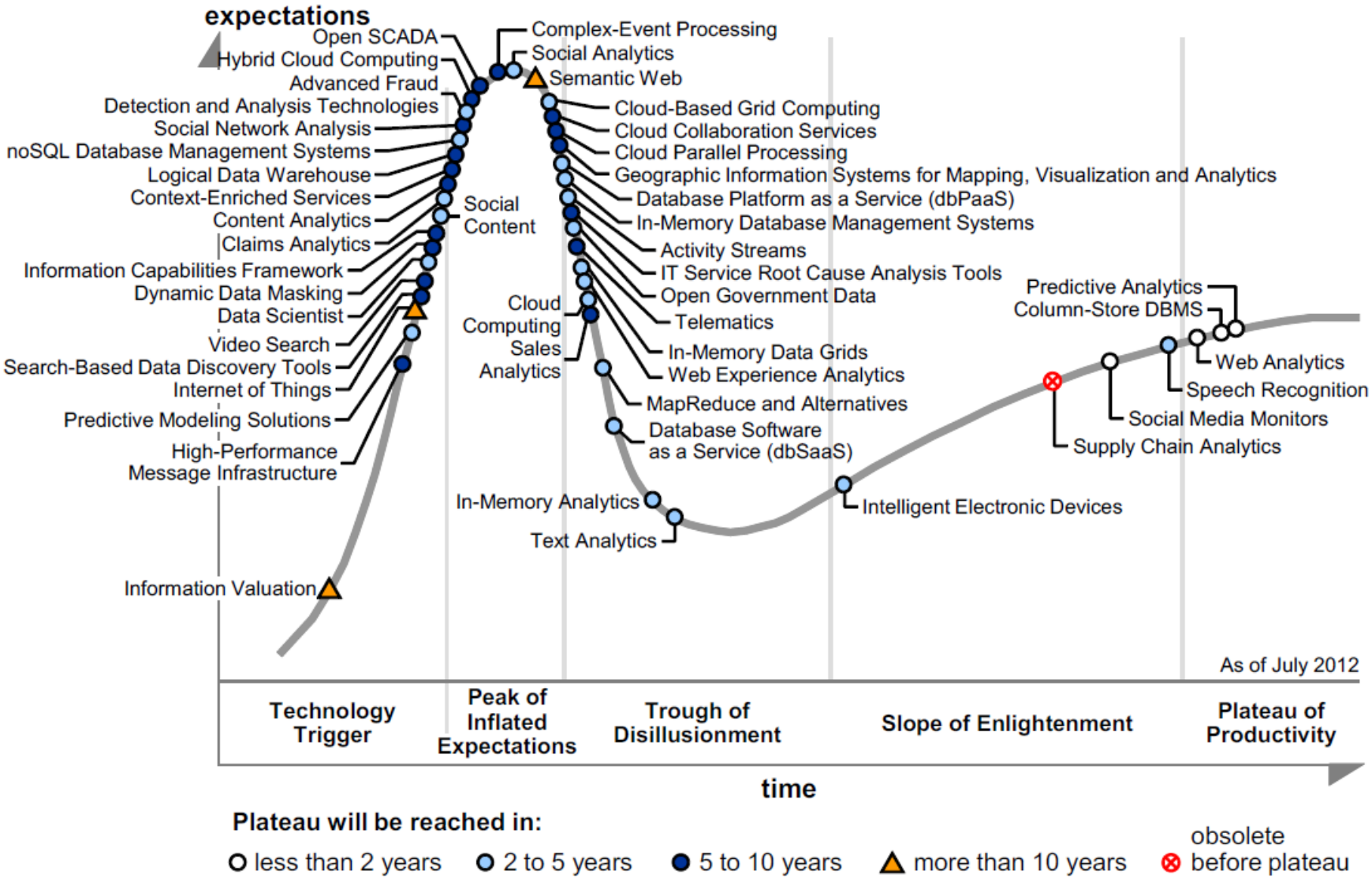
…adding “web 2.0” to “big data” and “data mining” queries volume



Structure of Big Data:
Gartner Hype Cycle for Big Data, 2012

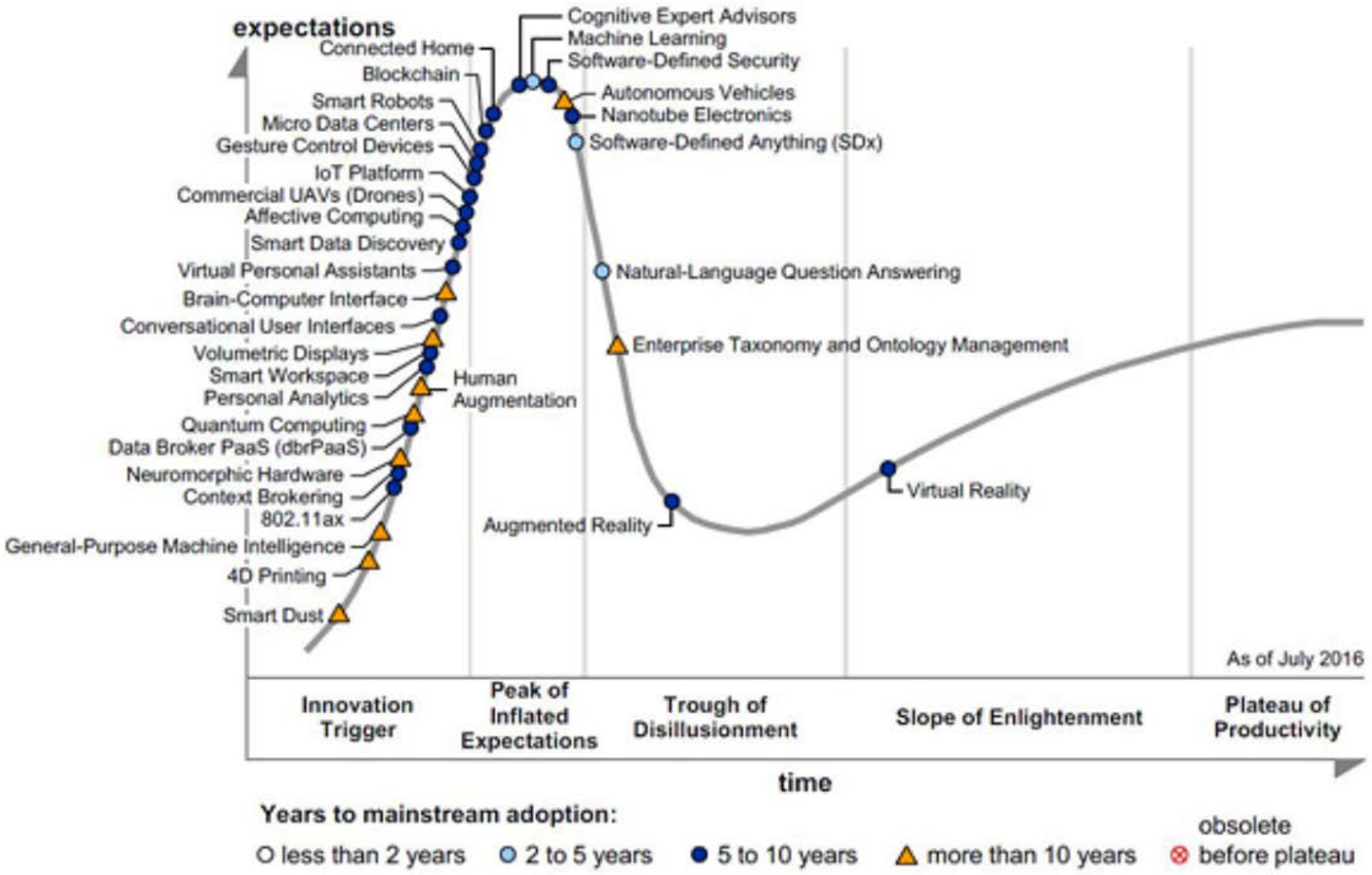
(later Gartner stopped producing Big Data Hype Cycle)

Figure 1. Hype Cycle for Big Data, 2012



What about the future?
Gartner Hype Cycle 2016

Figure 1. Hype Cycle for Emerging Technologies, 2016



Source: Gartner (July 2016)

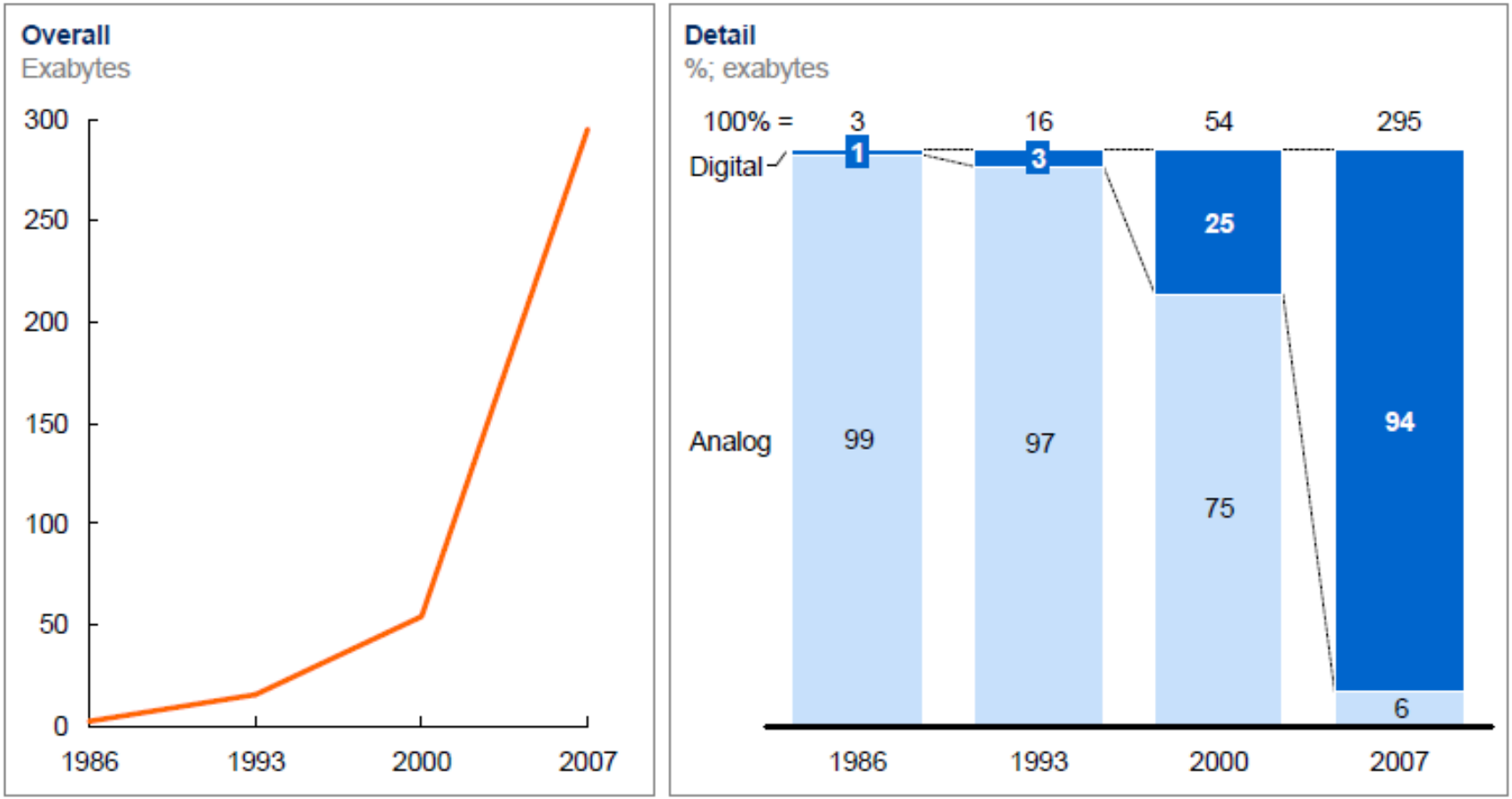
Why Big-Data?

- 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

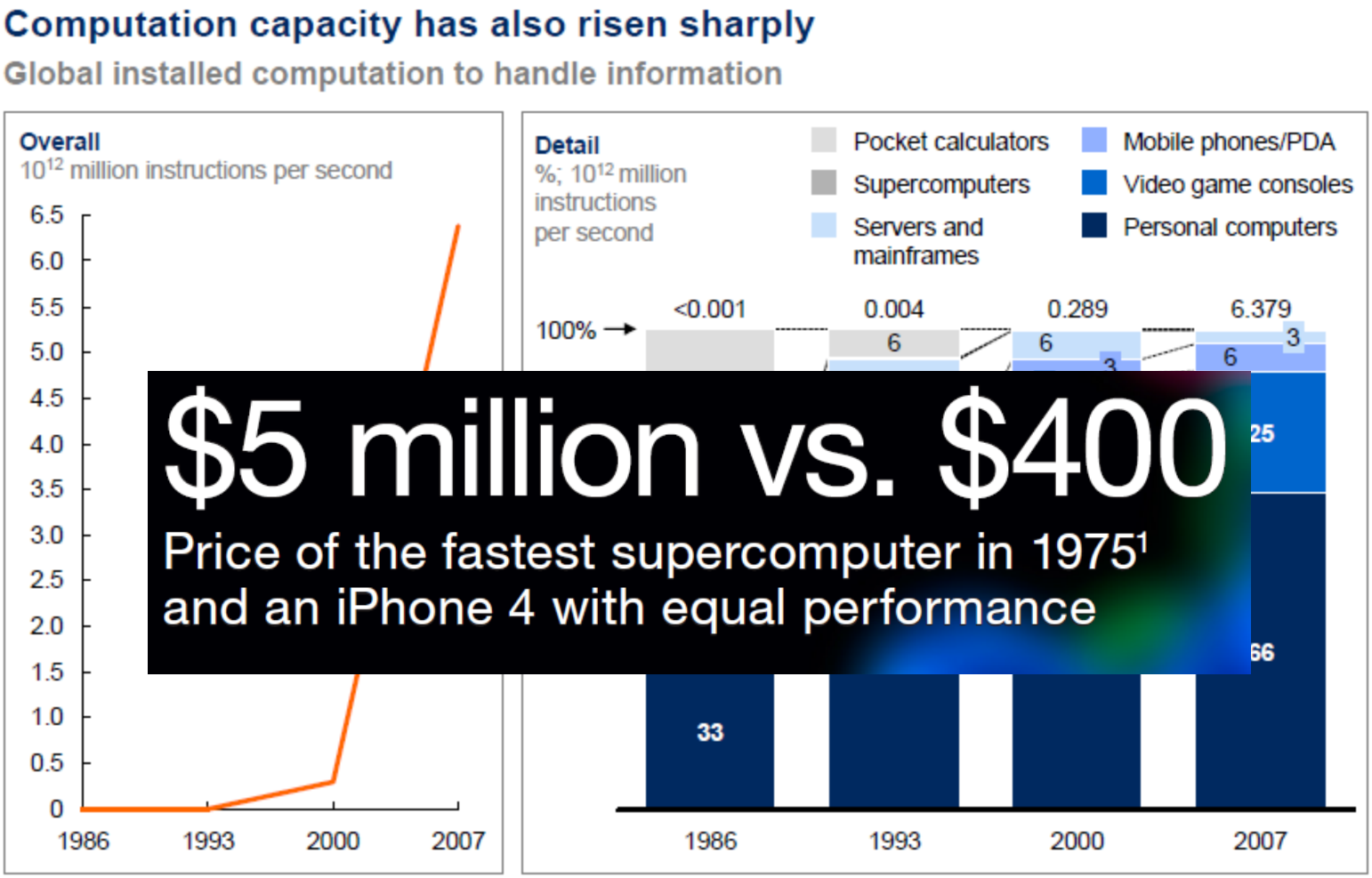
Global installed, optimally compressed, storage



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

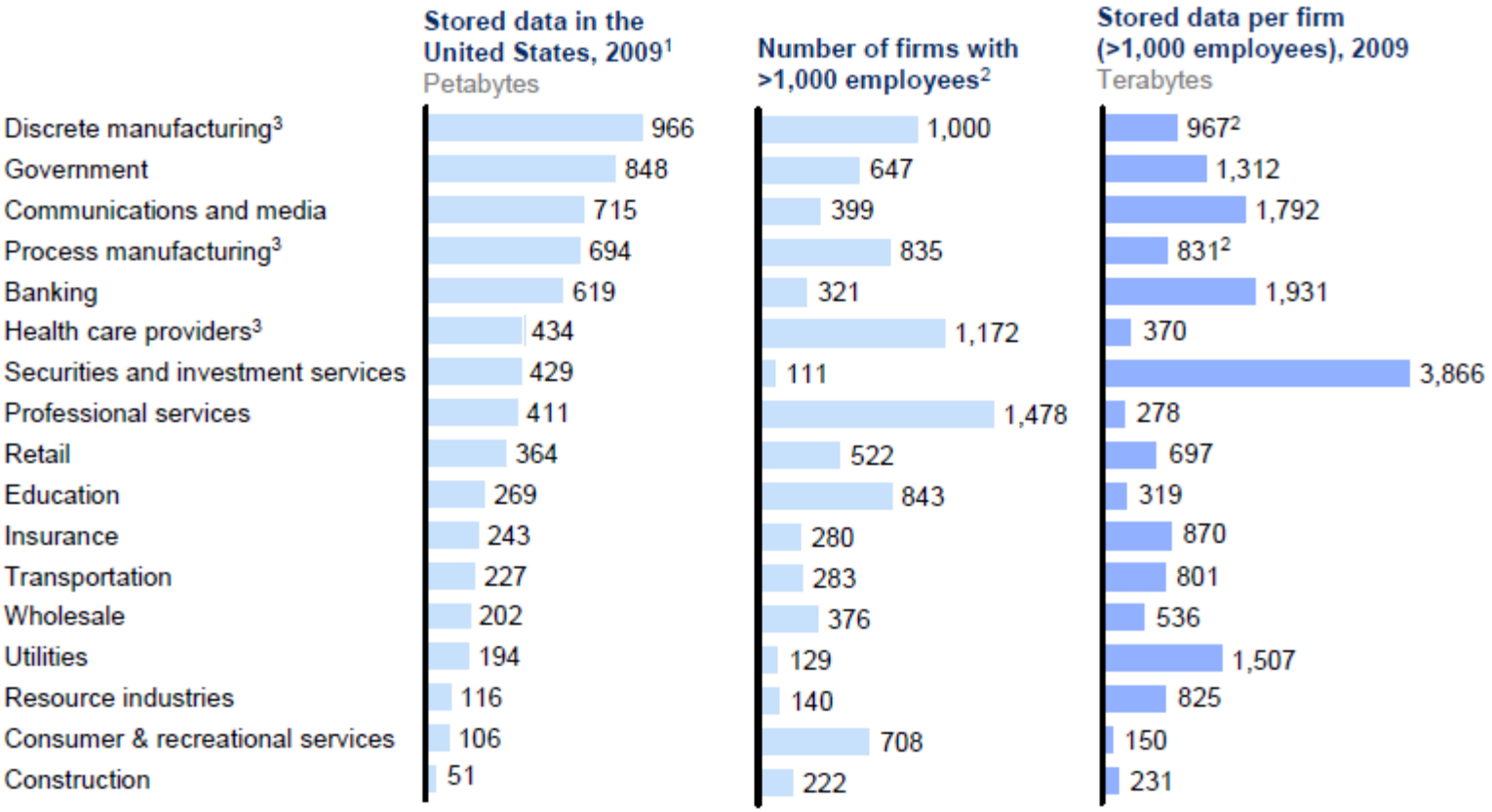


NOTE: Numbers may not sum due to rounding.

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

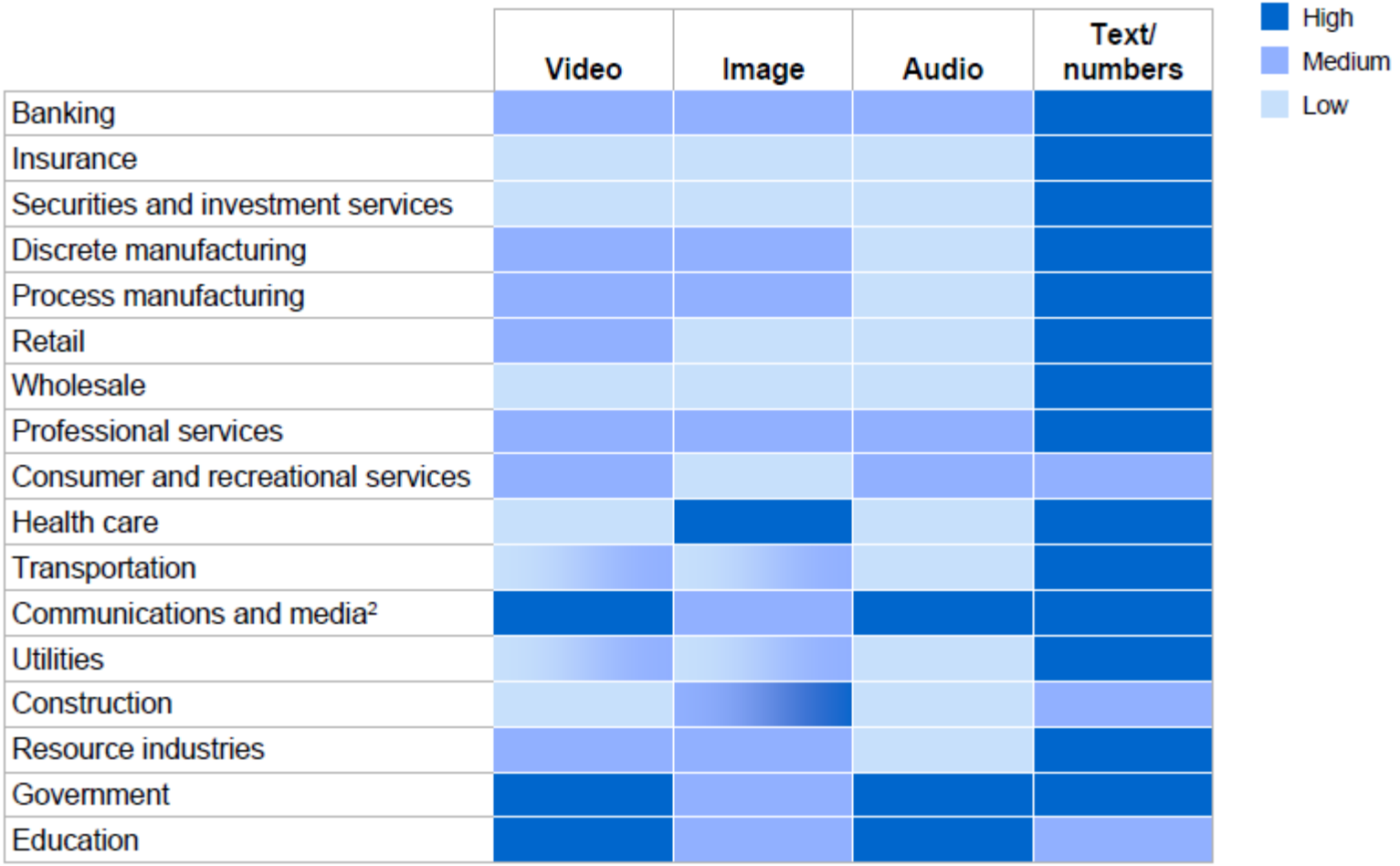


1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

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

Type of available data

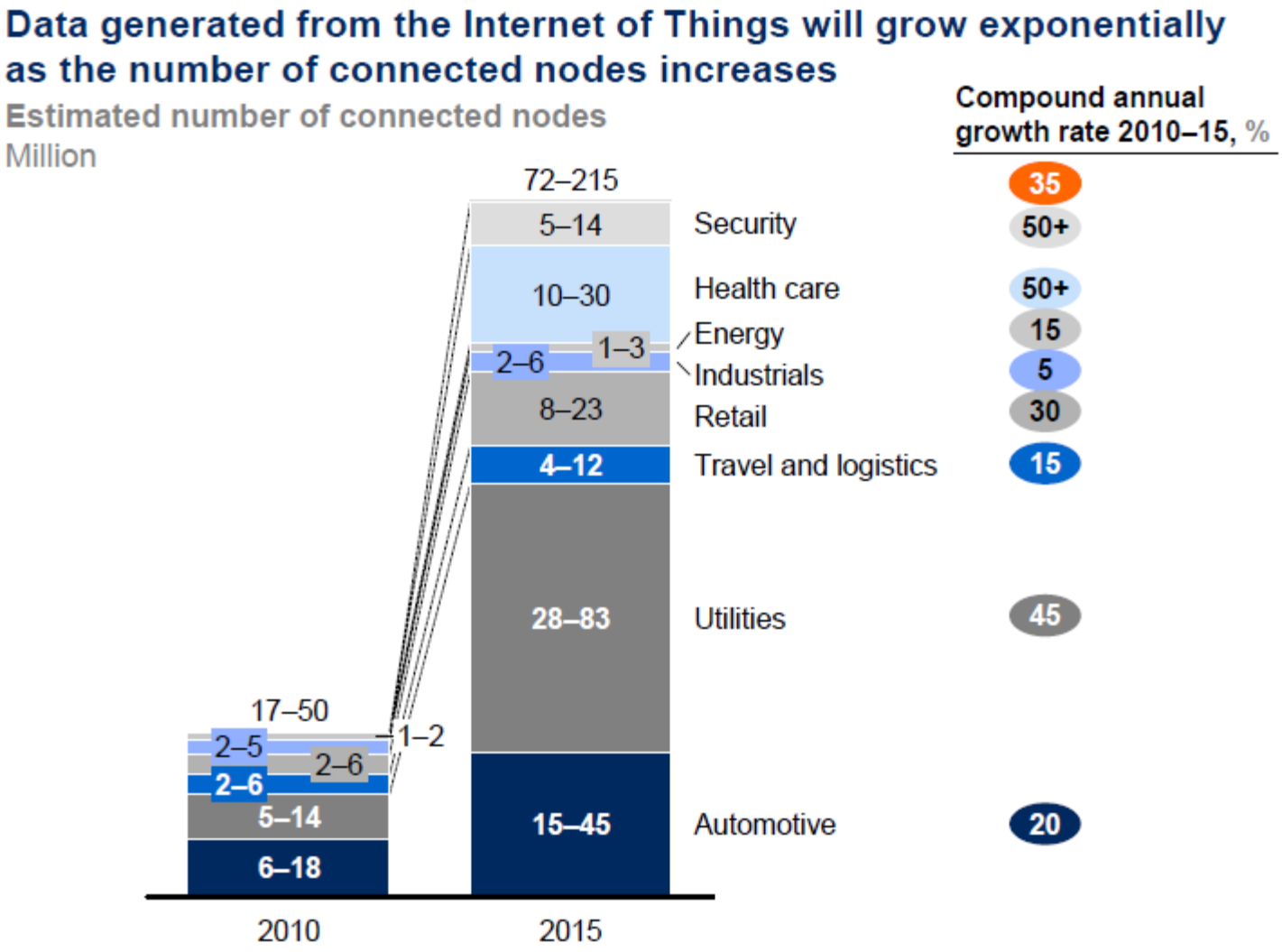
The type of data generated and stored varies by sector[1]



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

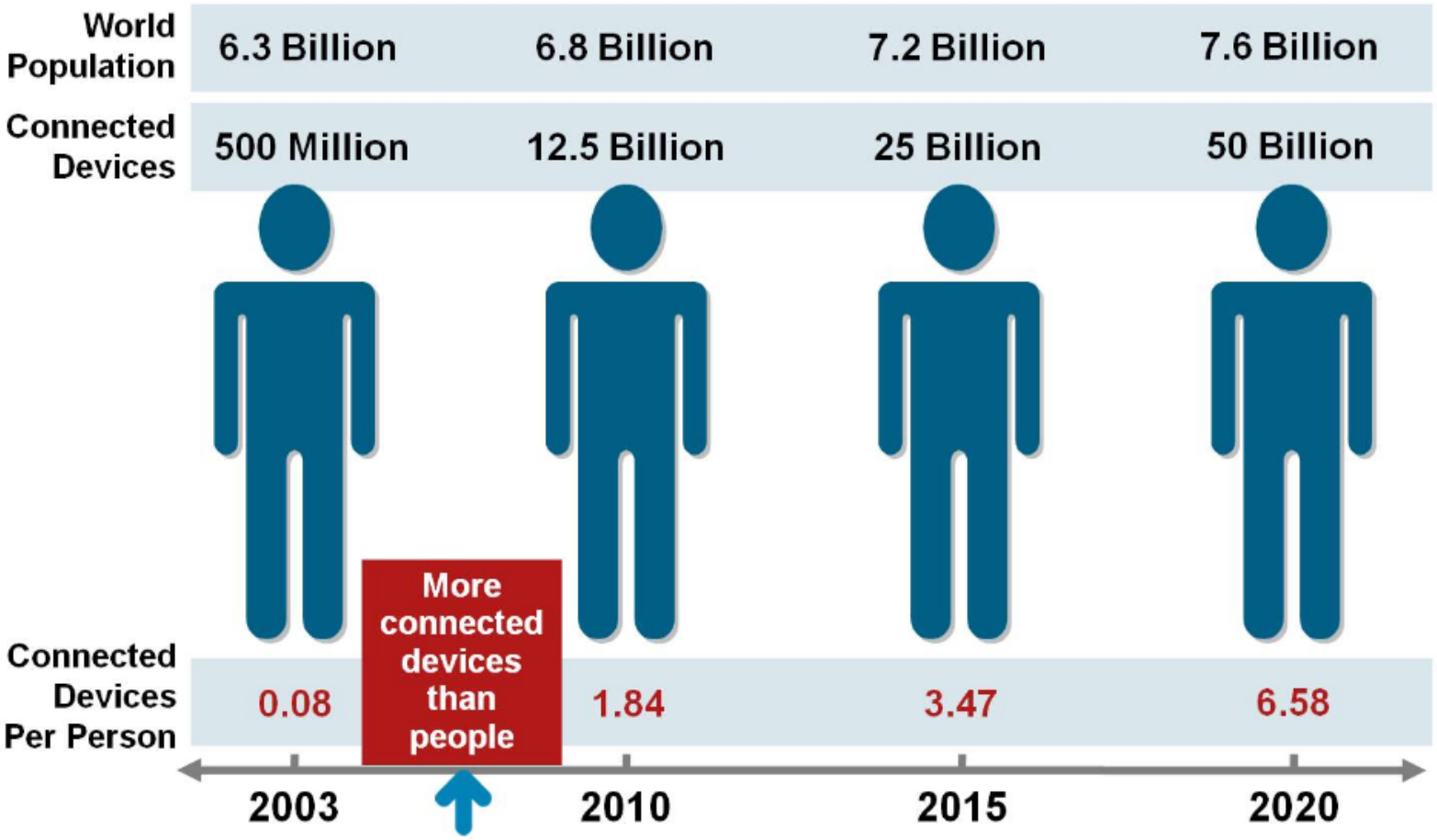


NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

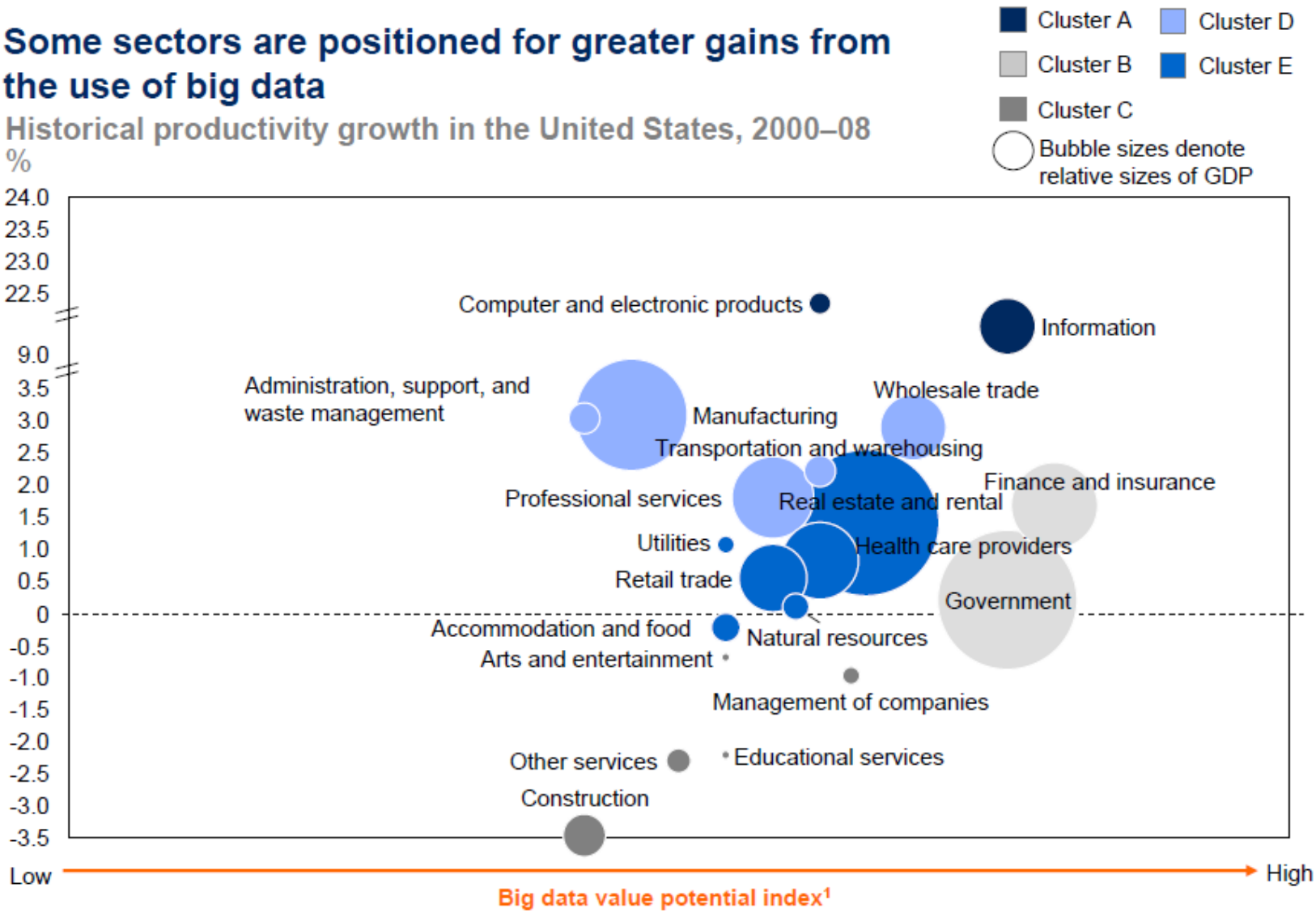
Birth & Growth of "Internet of Things"

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



Source: Cisco IBSG, April 2011

Gains from Big-Data per sector

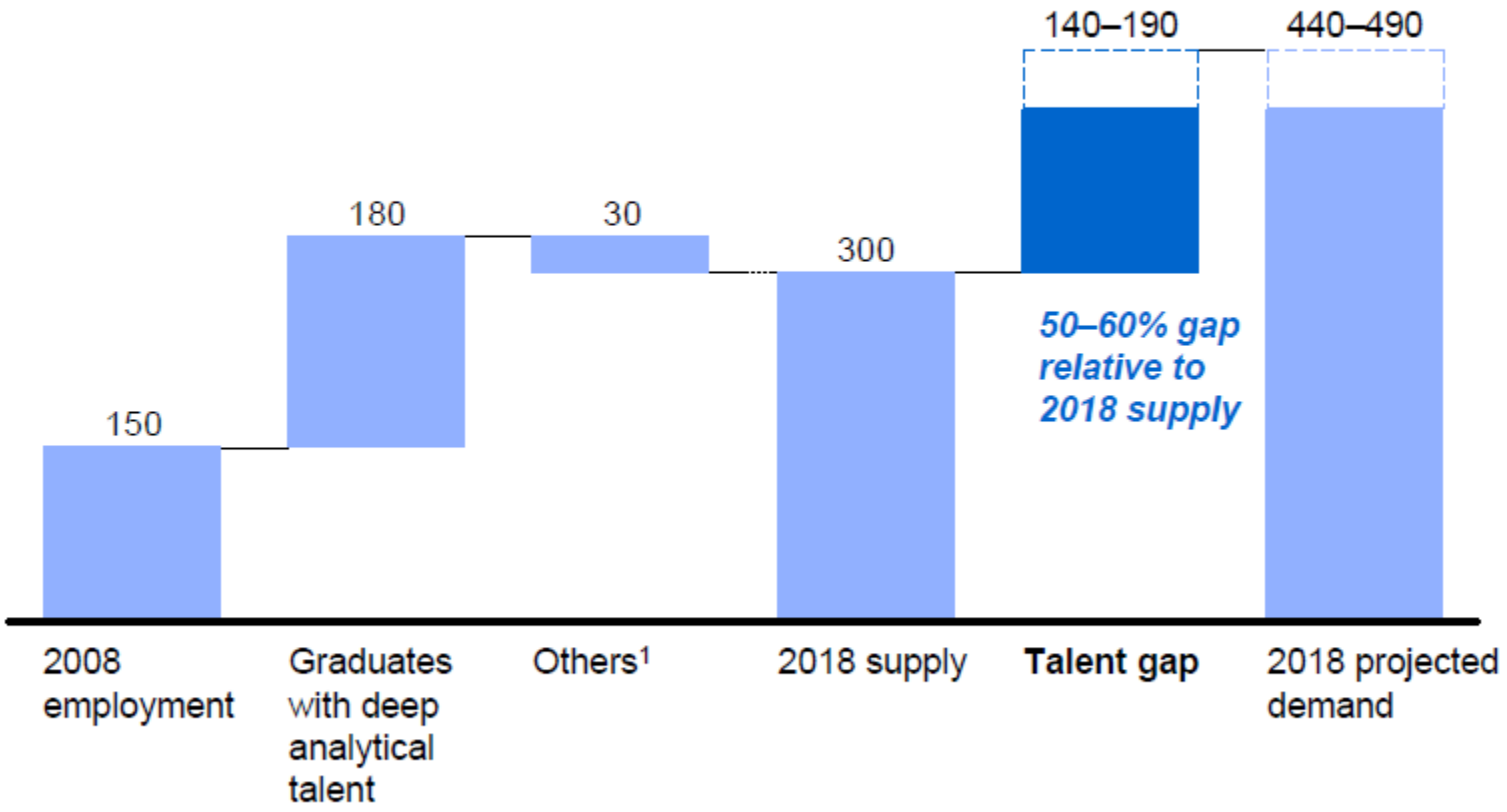


1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

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

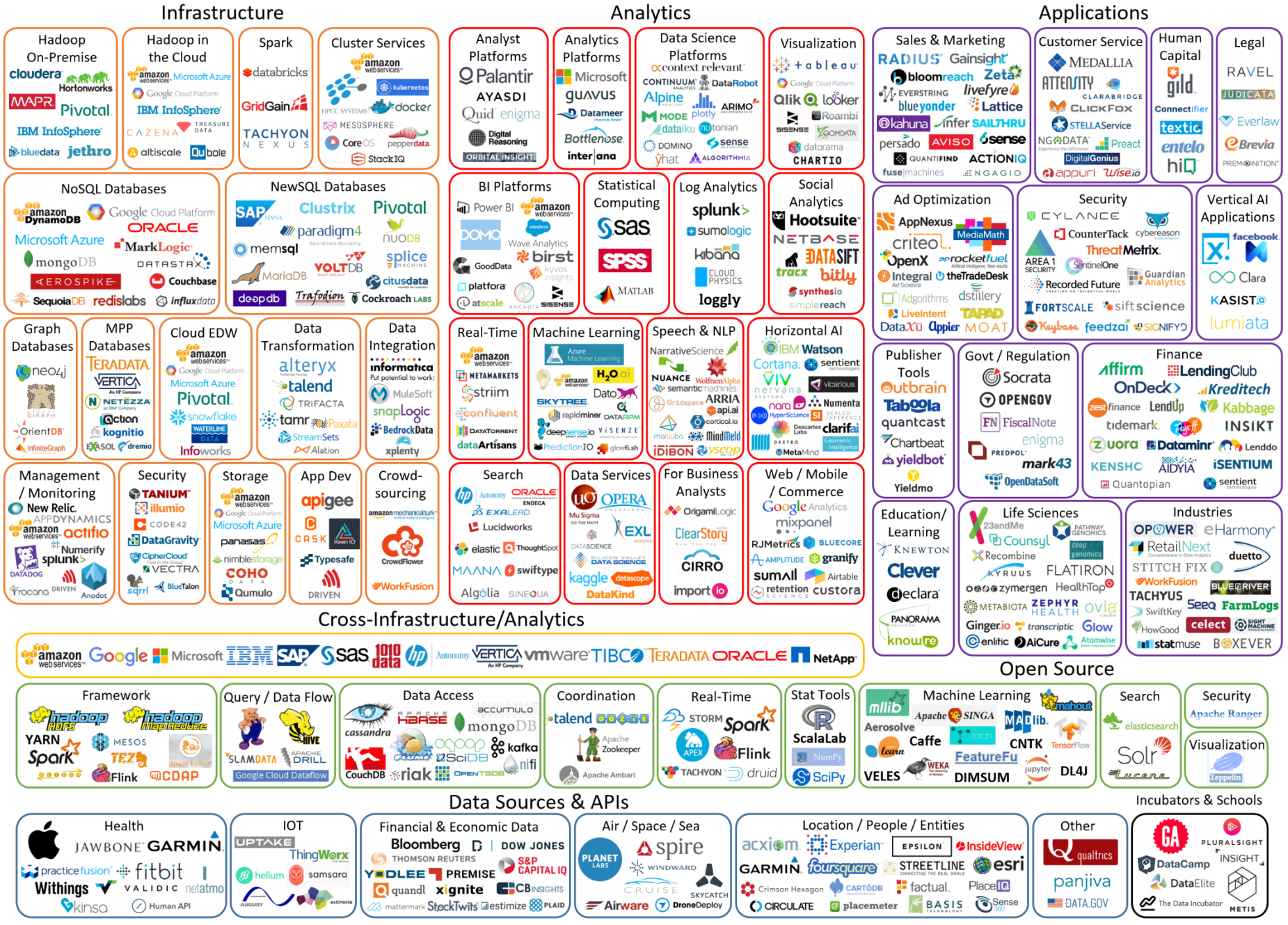


1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

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

Big Data Market

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

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

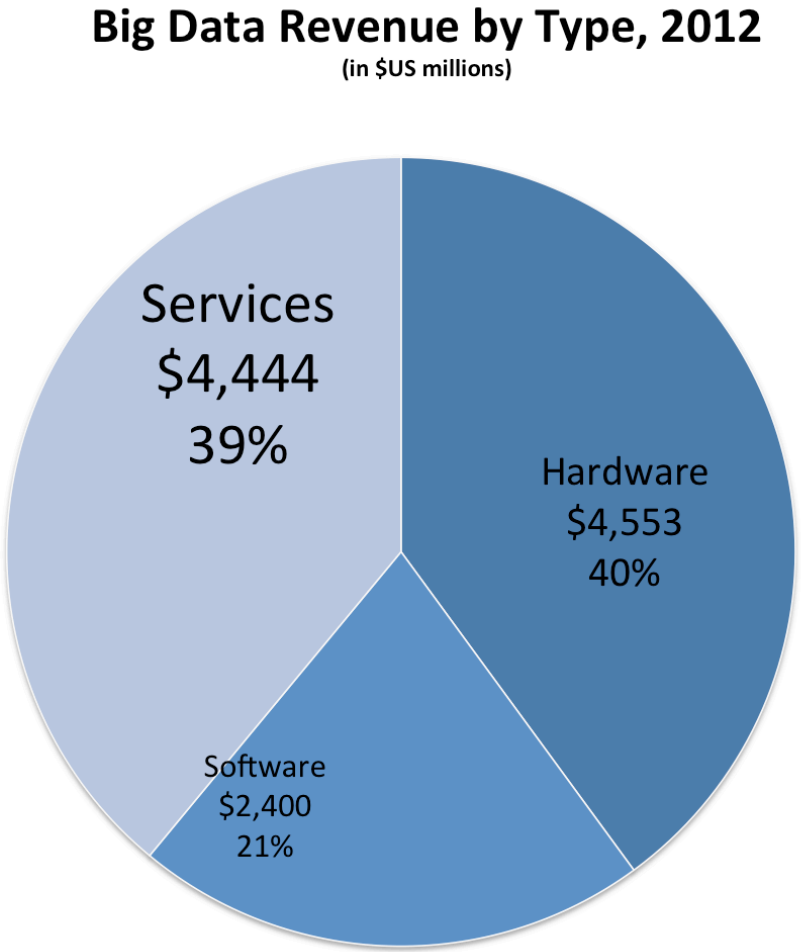
Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	$1,352	$103,930	1%	22%	33%	44%
HP	$664	$119,895	1%	34%	29%	38%
Teradata	$435	$2,665	16%	31%	28%	41%
Dell	$425	$59,878	1%	83%	0%	17%
Oracle	$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

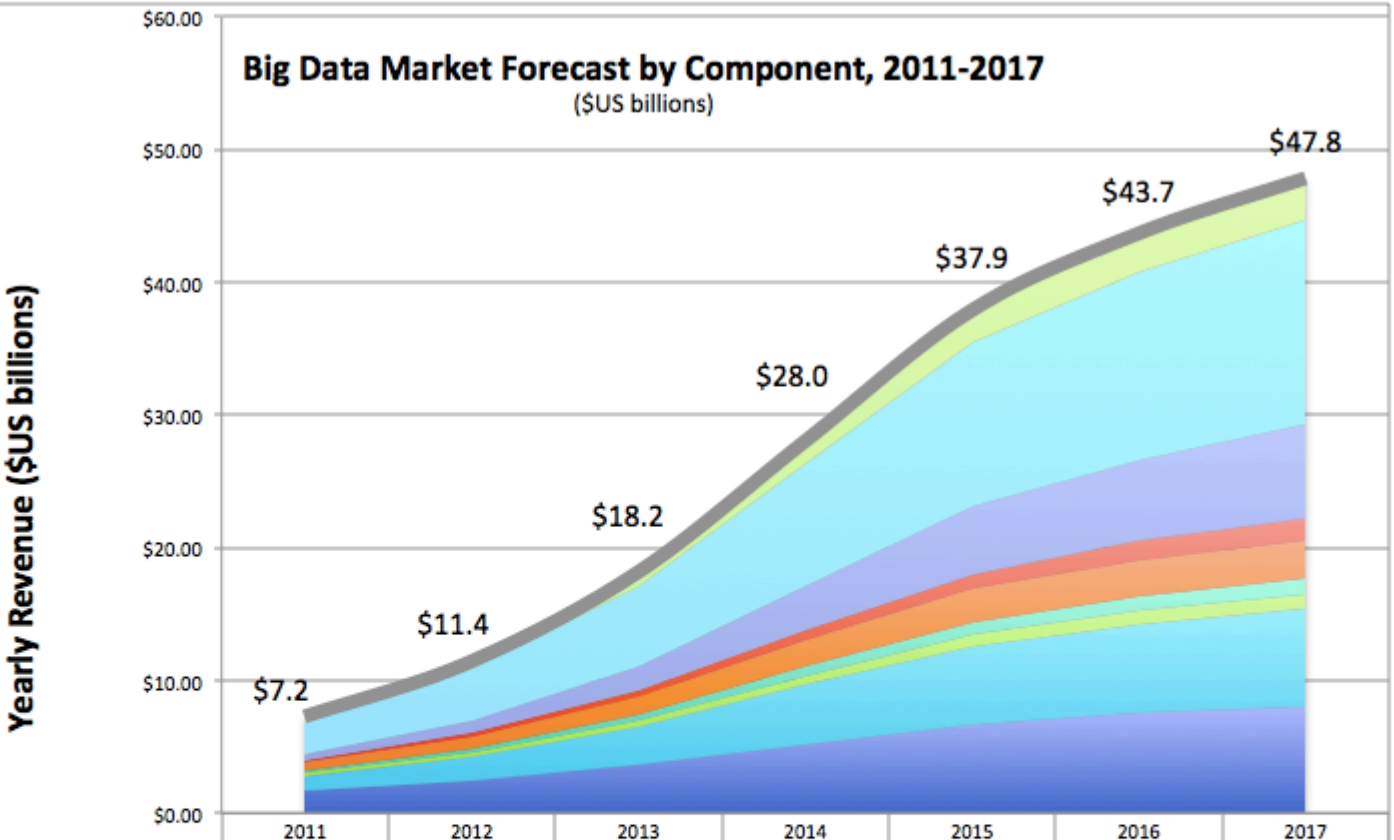
(http://wikibon.org/w/images/f/f9/Segment_-_BDMSVR2012.png)

Wikibon



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

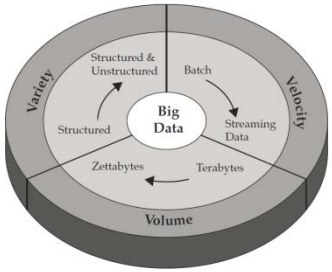
When Big-Data is really a hard problem?

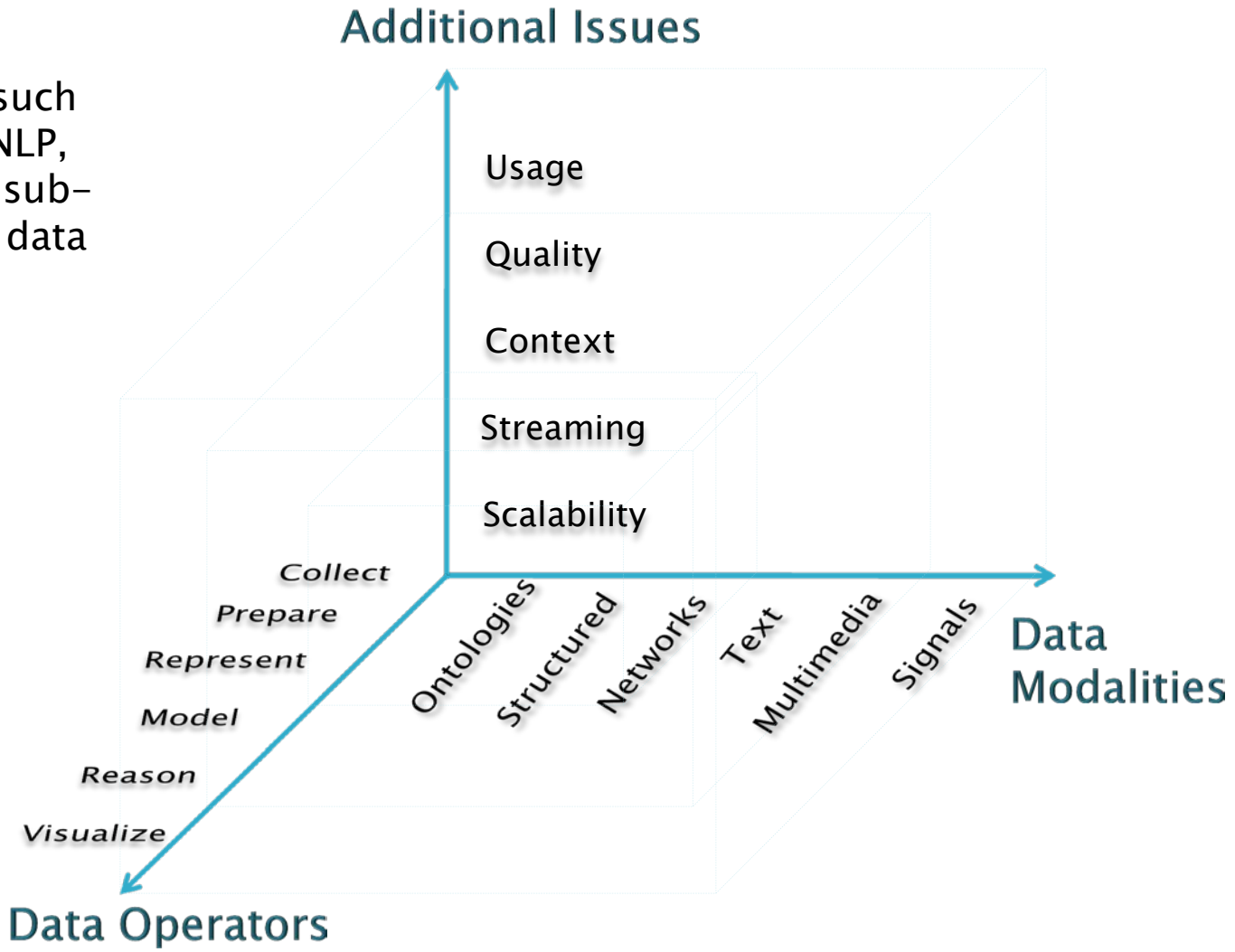
- …when the operations on data are complex:
  - …e.g. simple counting is not a complex problem
  - Modeling and reasoning with data of different kinds can get extremely complex

- 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?

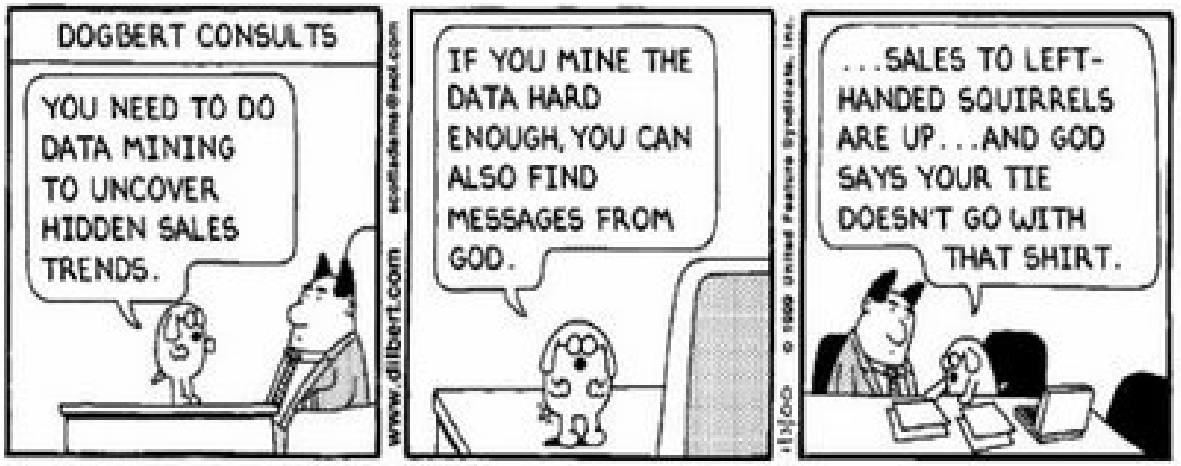
- Research areas (such as IR, KDD, ML, NLP, SemWeb, …) are sub-cubes within the data cube





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

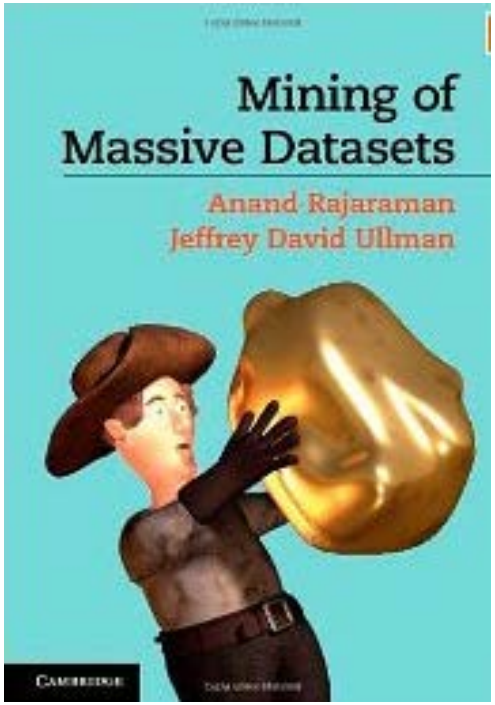
- **Smart sampling** of data
  - …reducing the original data while not losing the statistical properties of data
- **Finding similar items**
  - …efficient multidimensional indexing
- **Incremental updating** of the models
  - (vs. building models from scratch)
  - …crucial for streaming data
- **Distributed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

…guide to Big-Data algorithms

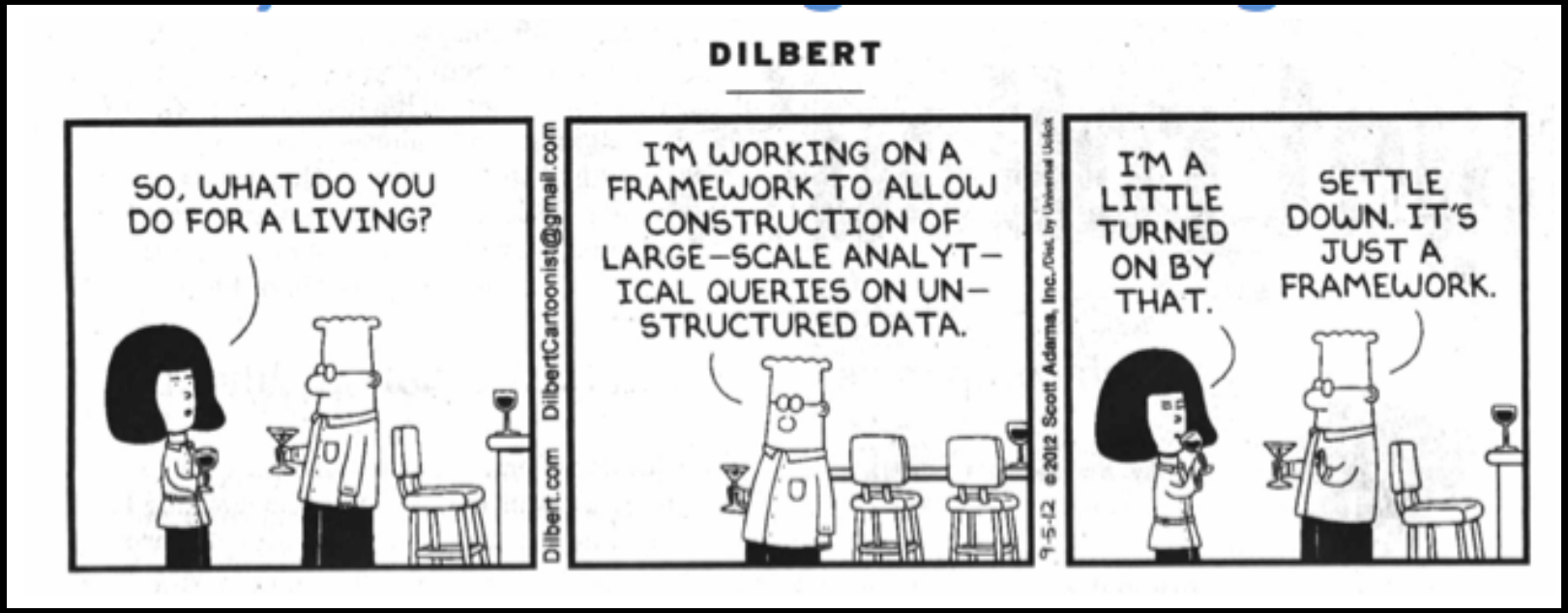
- An excellent overview of the algorithms covering the above issues is the book "**Rajaraman, Leskovec, Ullman: Mining of Massive Datasets**"

- 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

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

Open Source Big Data Tools

Infrastructure:

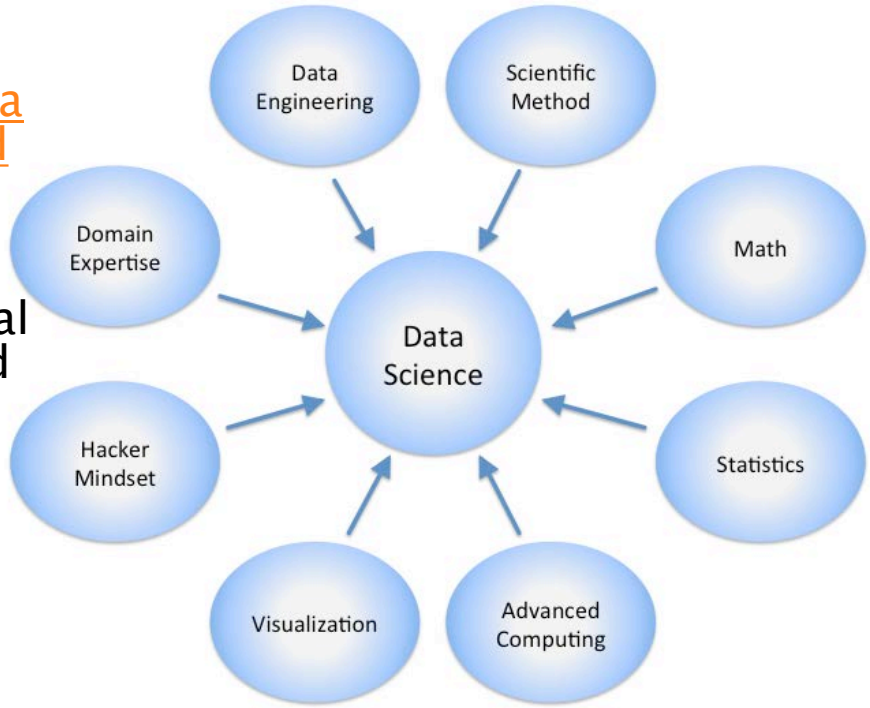
- Kafka [http://kafka.apache.org/]
  - A high-throughput distributed messaging system
- Hadoop [http://hadoop.apache.org/]
  - Open-source map-reduce implementation
- Storm [http://storm-project.net/]
  - Real-time distributed computation system
- Cassandra [http://cassandra.apache.org/]
  - Hybrid between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

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/

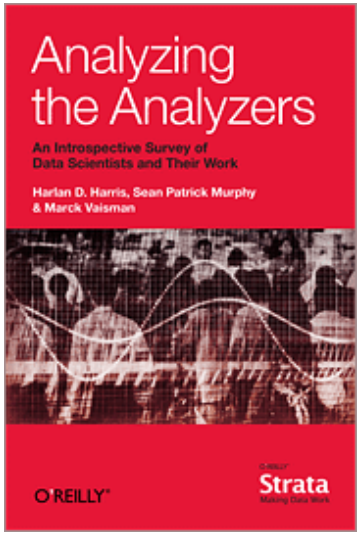
Business Intelligence vs. BI

	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

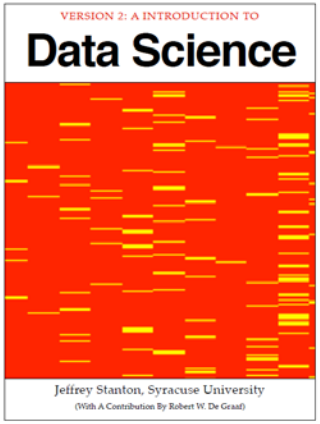
http://blog.revolutionanalytics.com/data-science/

Relevant reading

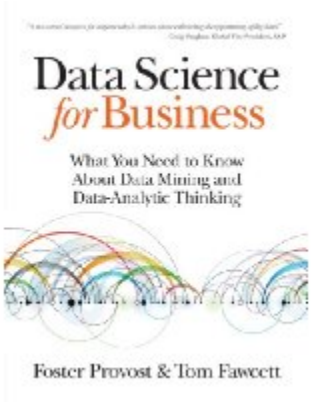
Analyzing the Analyzers
An Introspective Survey of Data Scientists and Their Work
By Harlan Harris, Sean Murphy, Marck Vaisman
Publisher: O'Reilly Media
Released: June 2013



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

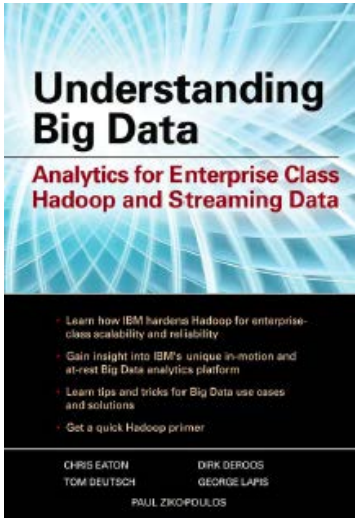


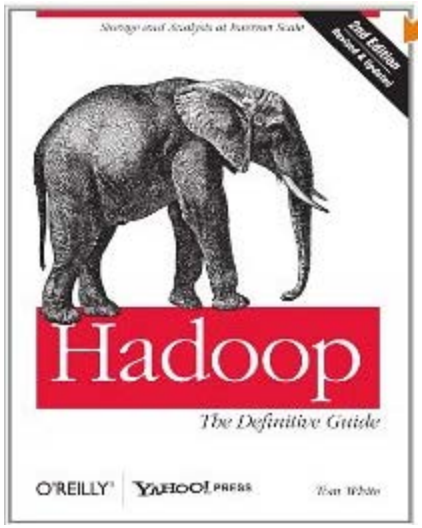
Data Science for Business: What you need to know about data mining and data-analytic thinking
by Foster Provost and Tom Fawcett
Released: Aug 16, 2013

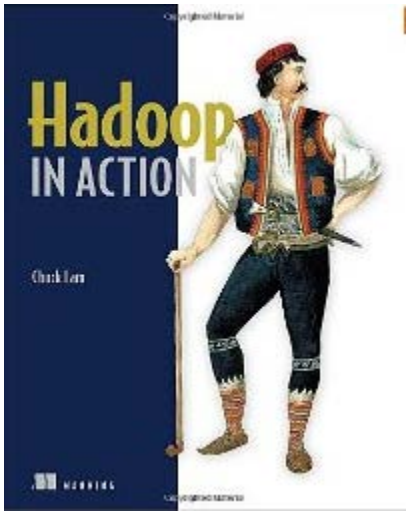


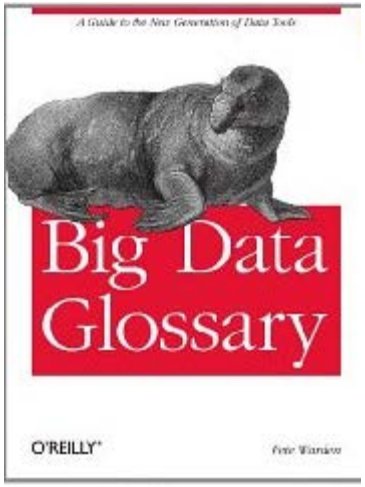
Final thoughts

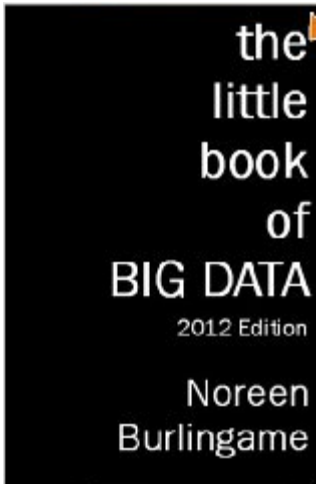
Literature on Big-Data

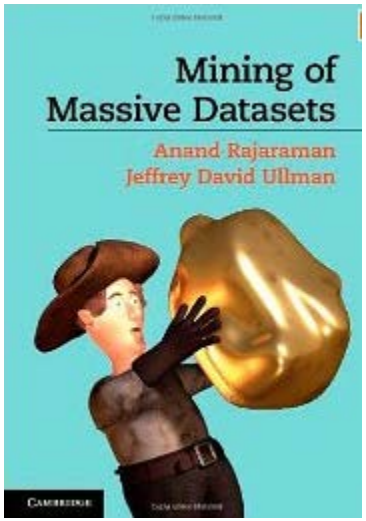


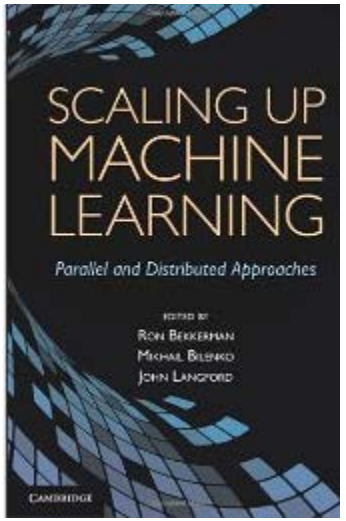












…to conclude

- 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)