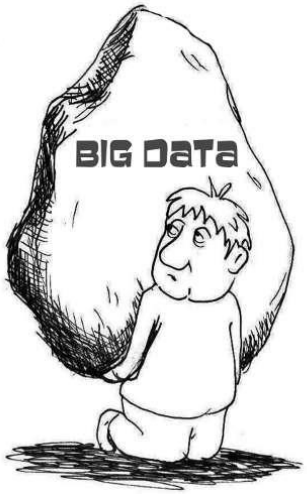


Quick Introduction to Big Data

Marko Grobelnik

Marko.Grobelnik@ijs.si

Jozef Stefan Institute, Slovenia

Brdo, Nov 10th 2015

Big Data Tutorial: http://www.slideshare.net/markogrobelnik/big-datatutorial-grobelnikfortunamladenicsydneyiswc2013

Big-Data Definitions

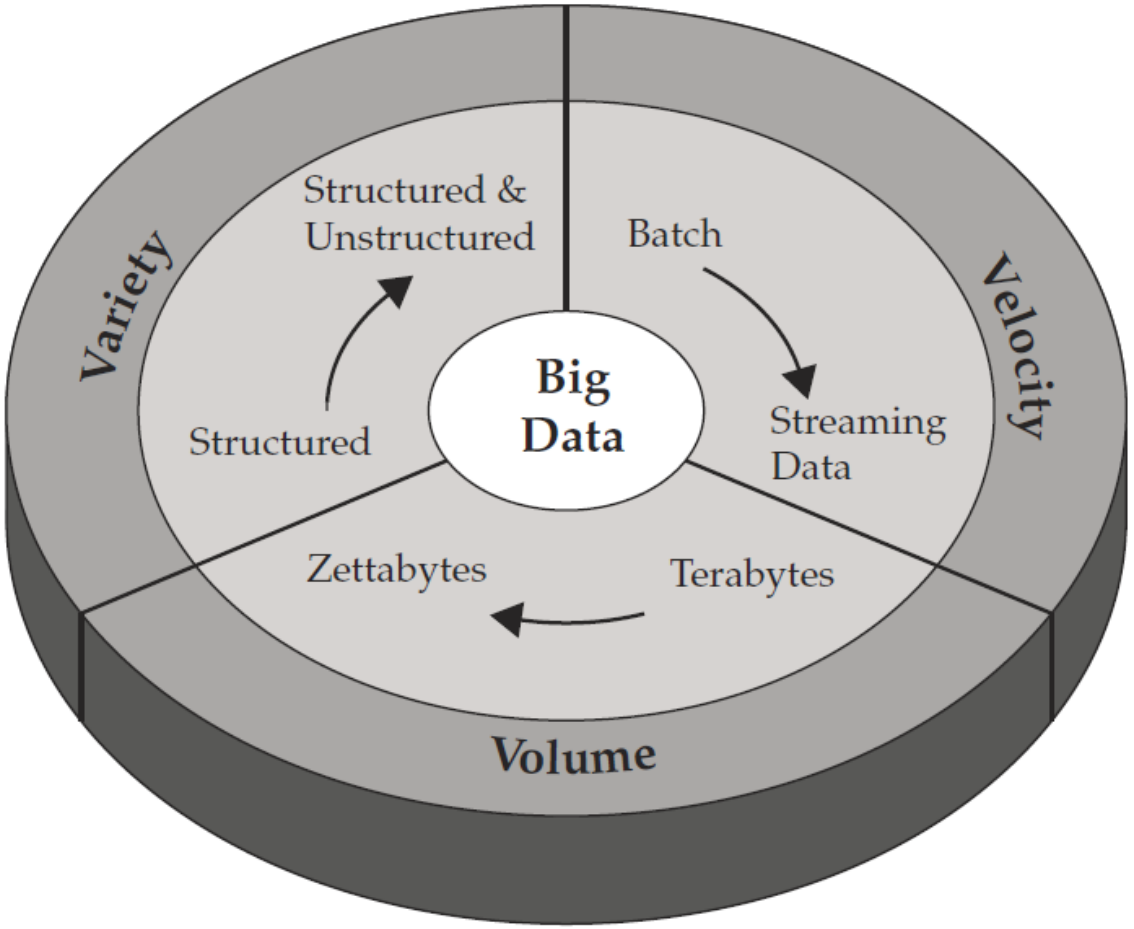
…so, what is Big-Data?

- 'Big-data' is similar to 'Small-data', but bigger
  - Recently getting popular expression "Midsize data"

- …but having data bigger it requires somewhat 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

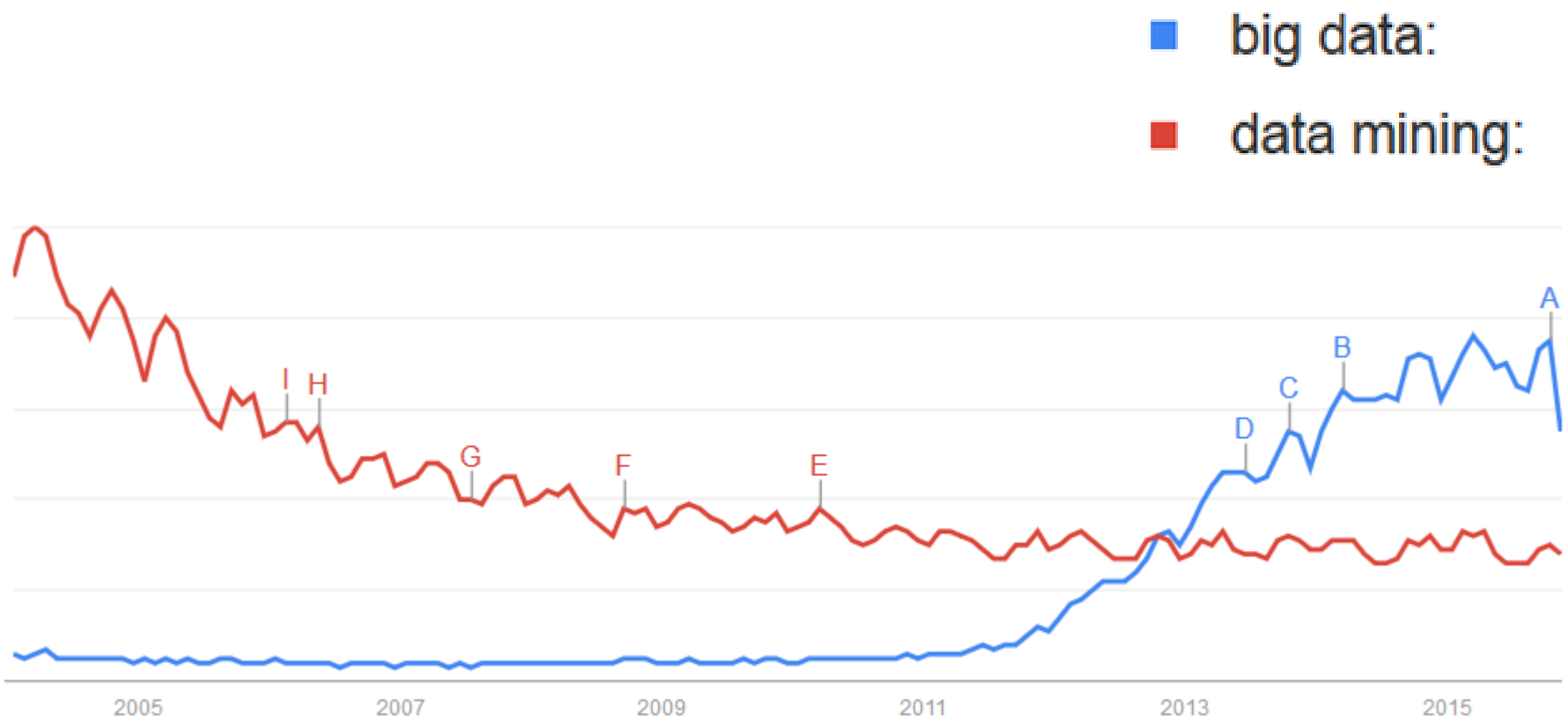
- 1. **Volume** (lots of data = "Tonnabytes")
- 2. **Variety** (complexity, curse of dimensionality)
- 3. **Velocity** (rate of data and information flow)

- 4. **Veracity** (verifying inference-based models from comprehensive data collections)
- 5. **Venue** (location)
- 6. **Vocabulary** (semantics)
- 7., 8., 9. …: V…, V…, V…

Motivation for Big-Data

Big-Data popularity on the Web
(through the eyes of "Google Trends")

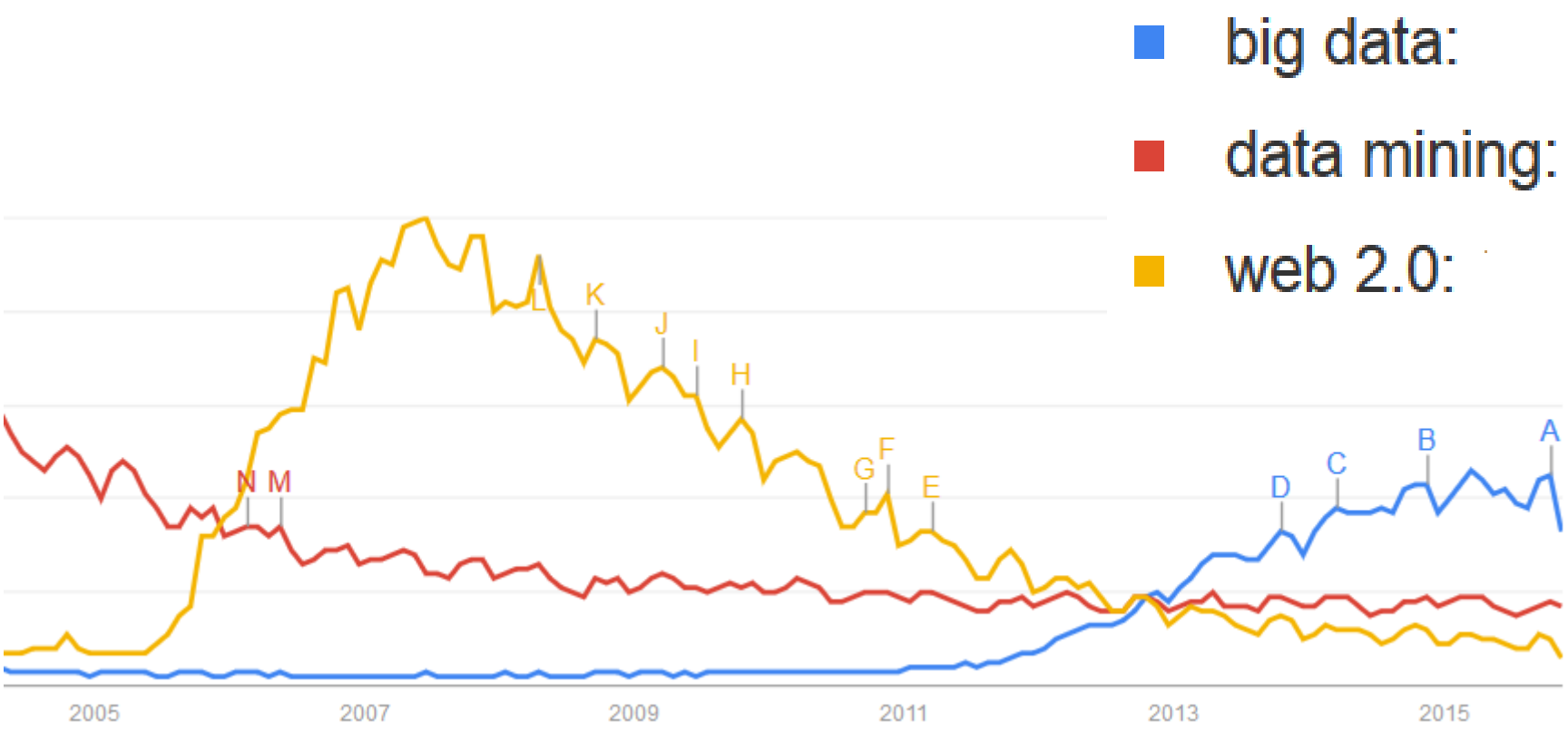
Comparing volume of "big data" and "data mining" queries



http://www.google.com/trends/explore#q=big%20data%2C%20data%20mining

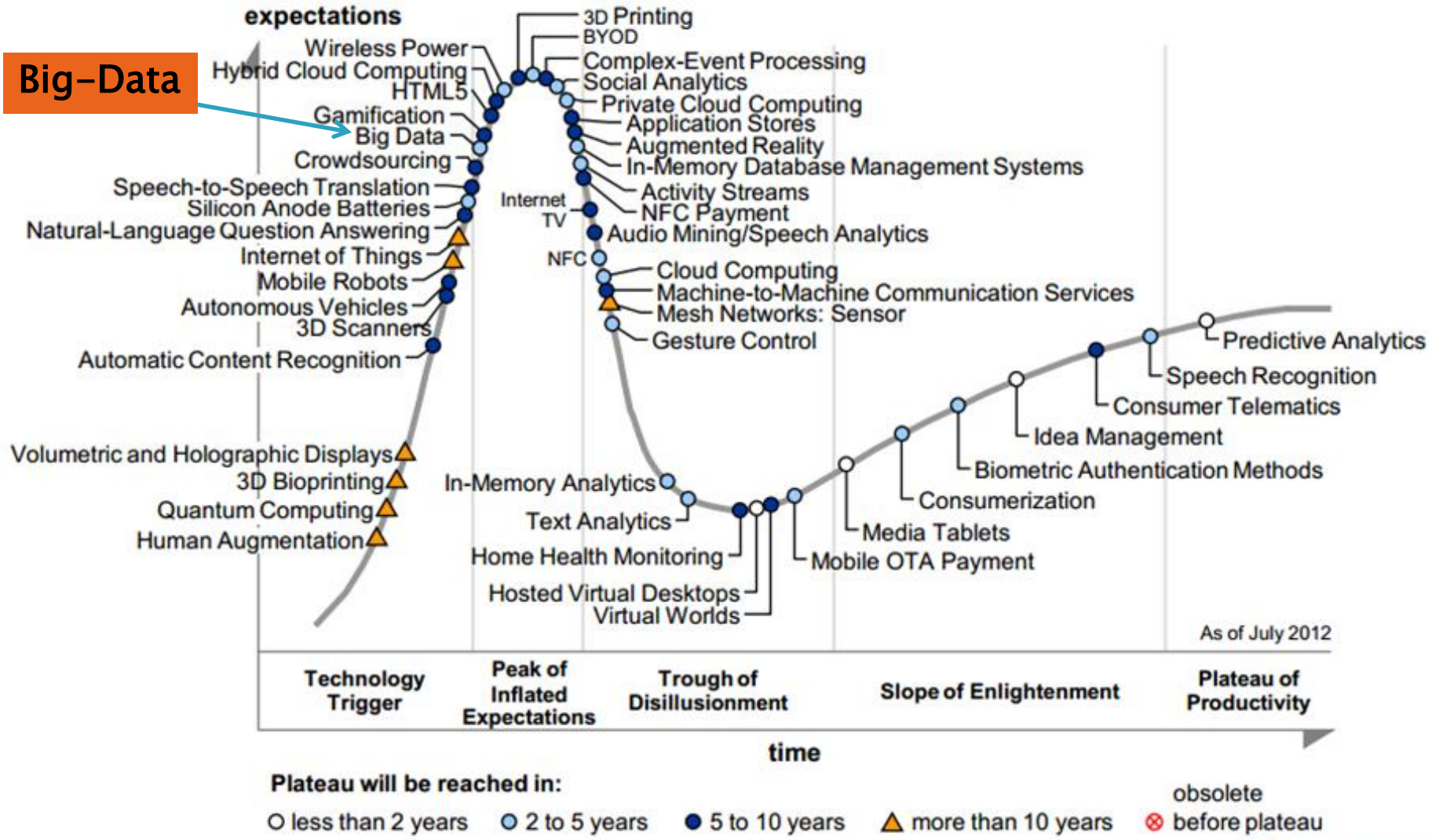
…but what can happen to “hypes”

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



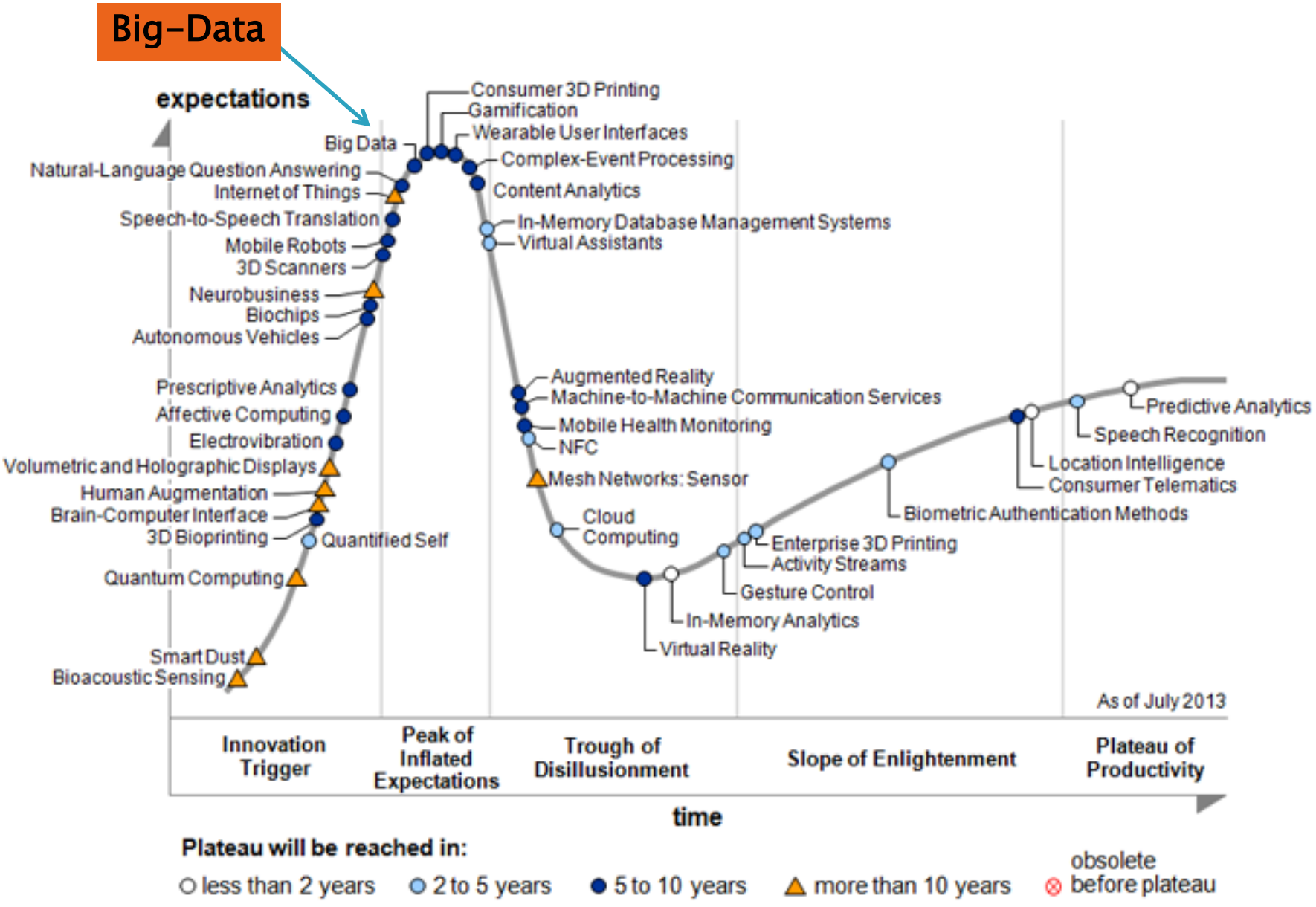
http://www.google.com/trends/explore#q=big%20data%2C%20data%20mining%2C%20web%202.0

Gartner: Emerging Technologies Hype Cycle 2012

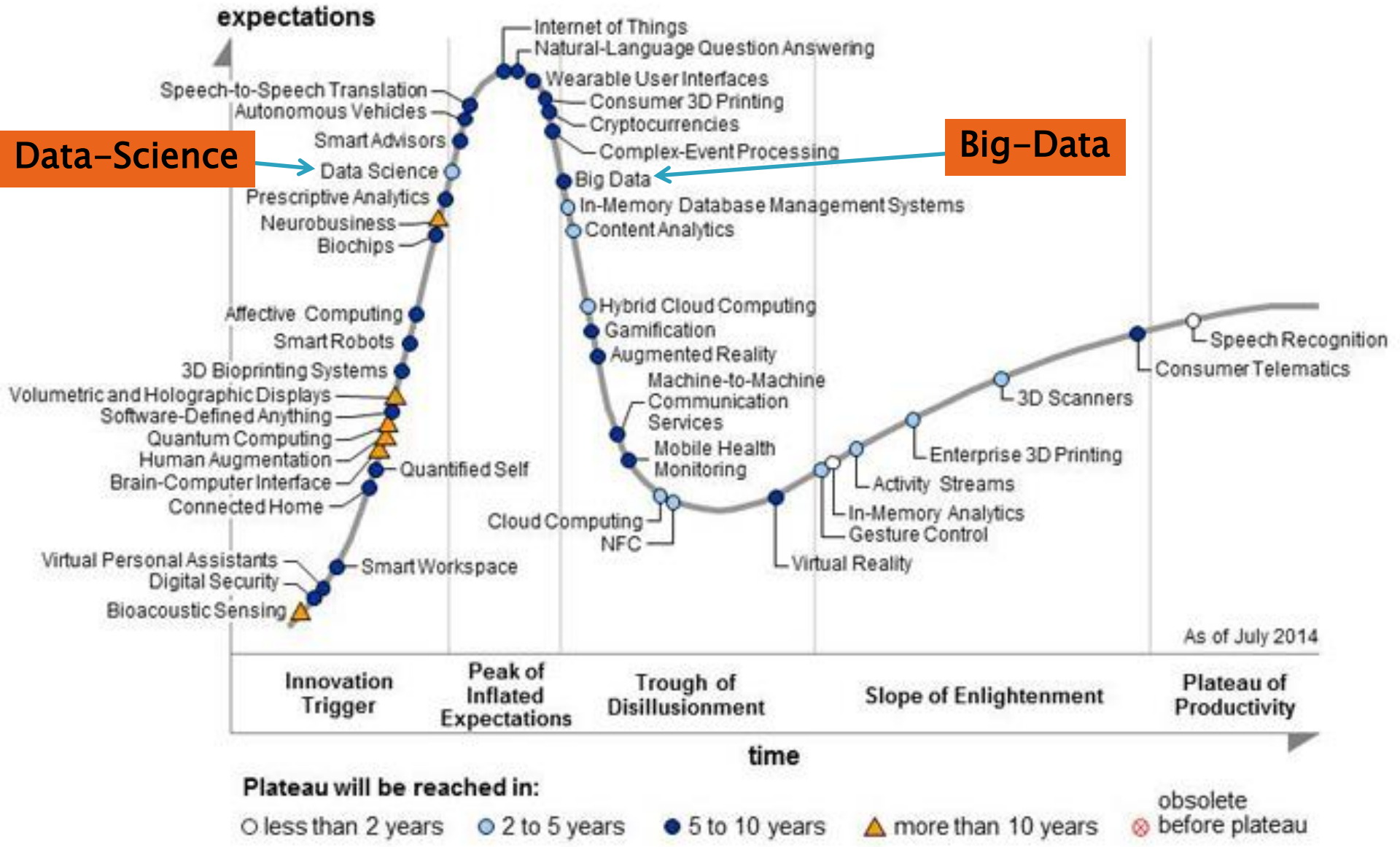


Gartner

Gartner: Emerging Technologies Hype Cycle 2013

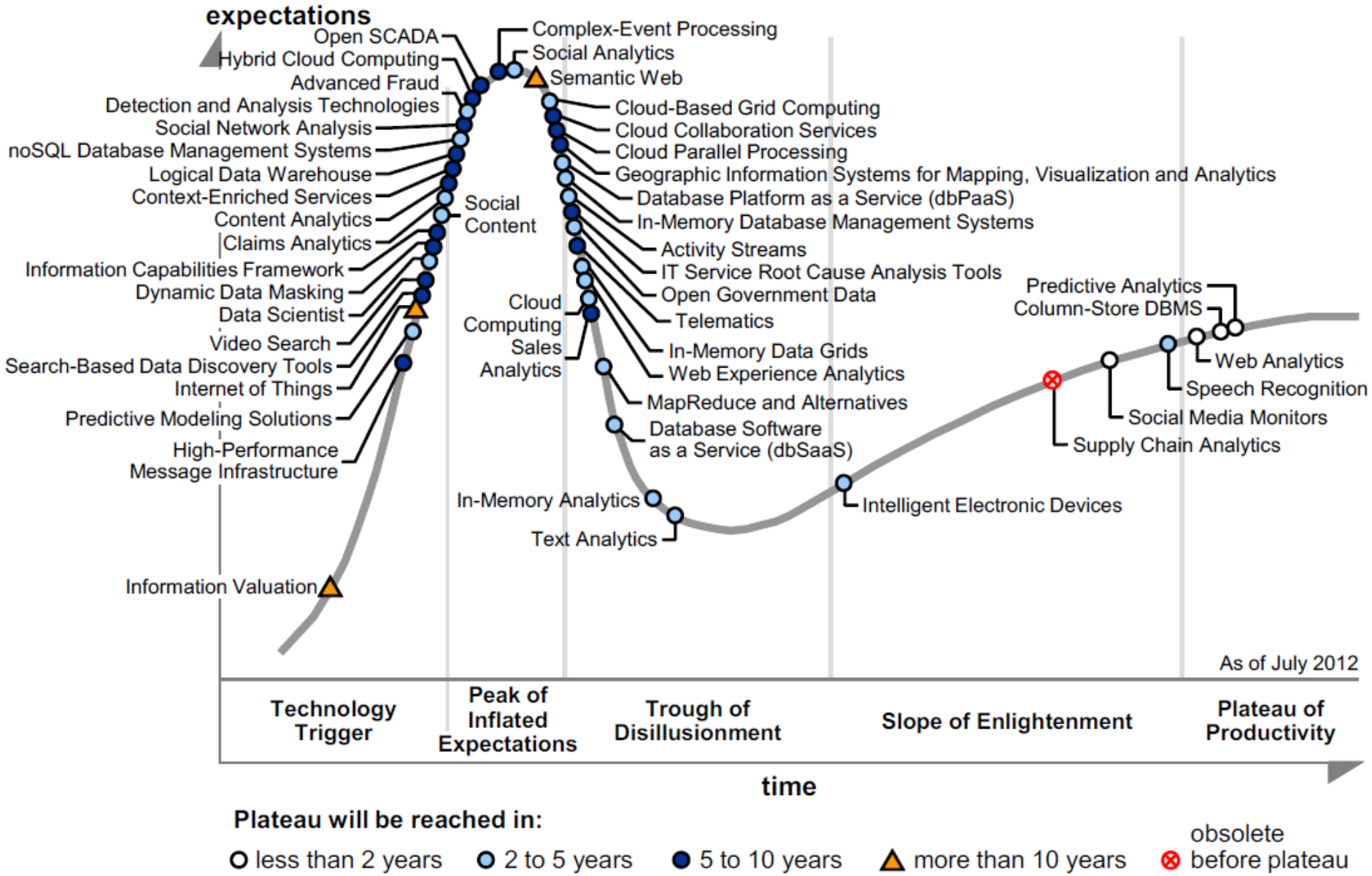


Gartner: Emerging Technologies Hype Cycle 2014



Gartner: Hype Cycle for Big Data

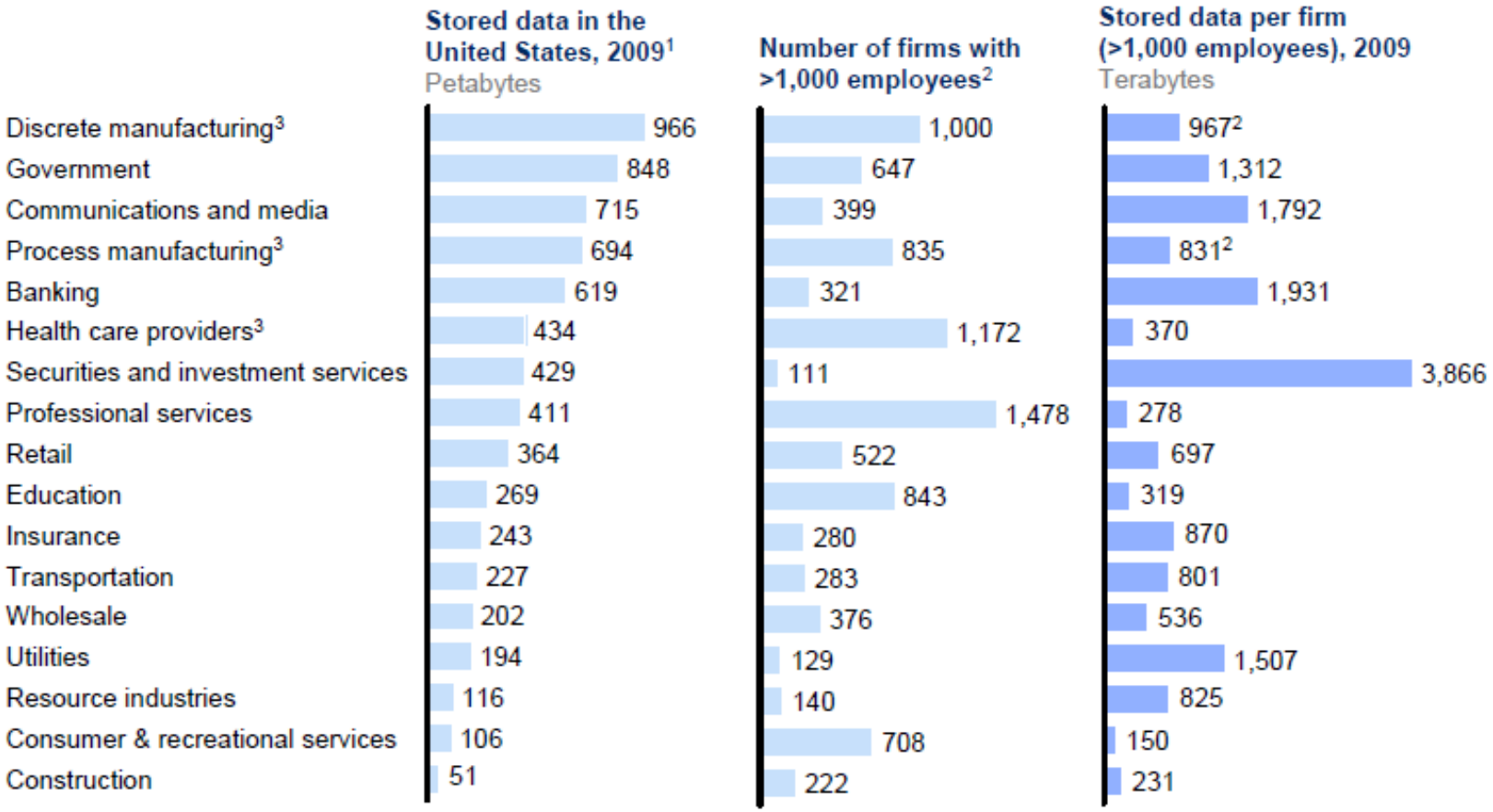
Figure 1. Hype Cycle for Big Data, 2012



Big Data Market

Enabler: Data availability

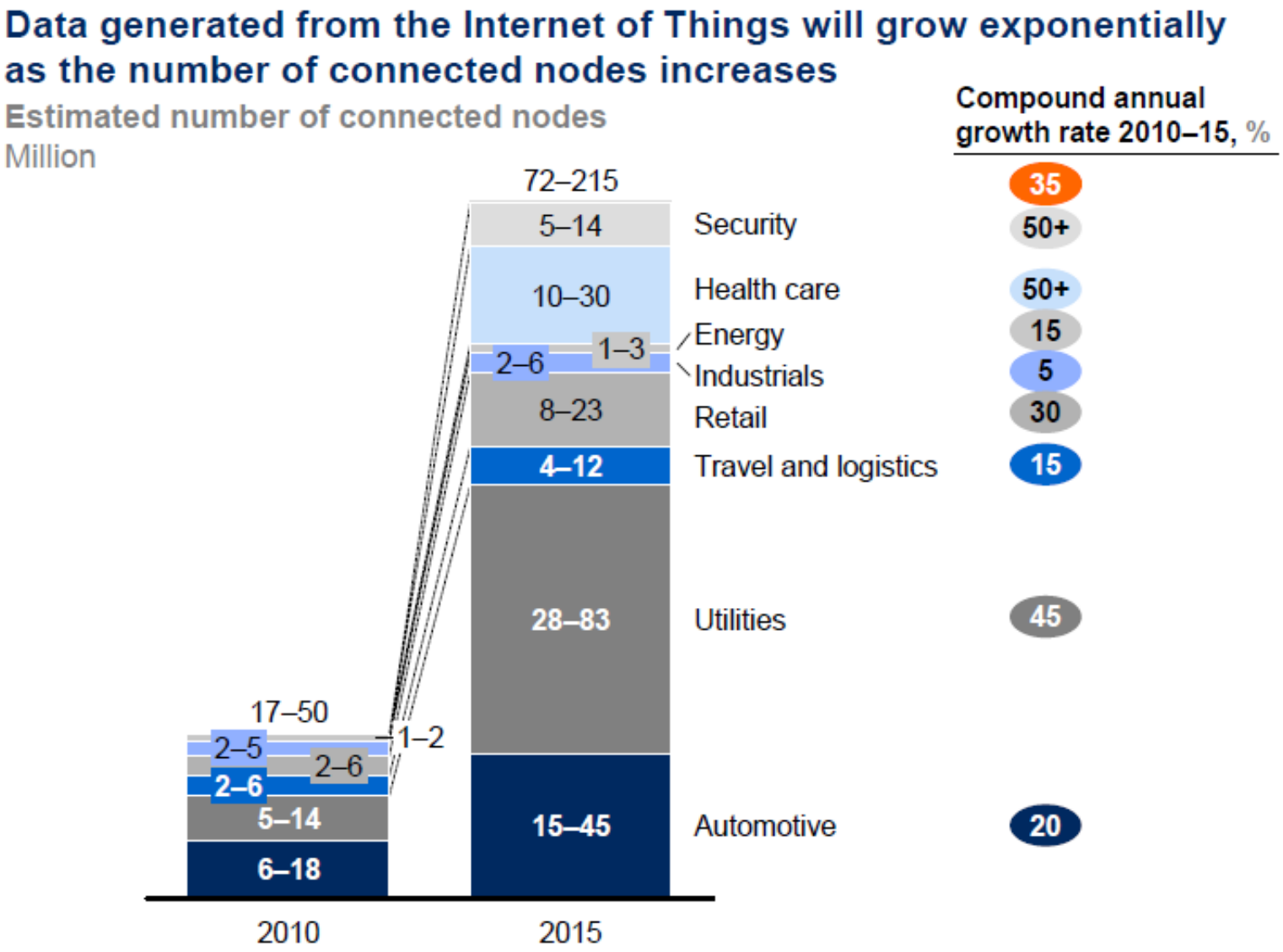
Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



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

Data available from “Internet of Things”

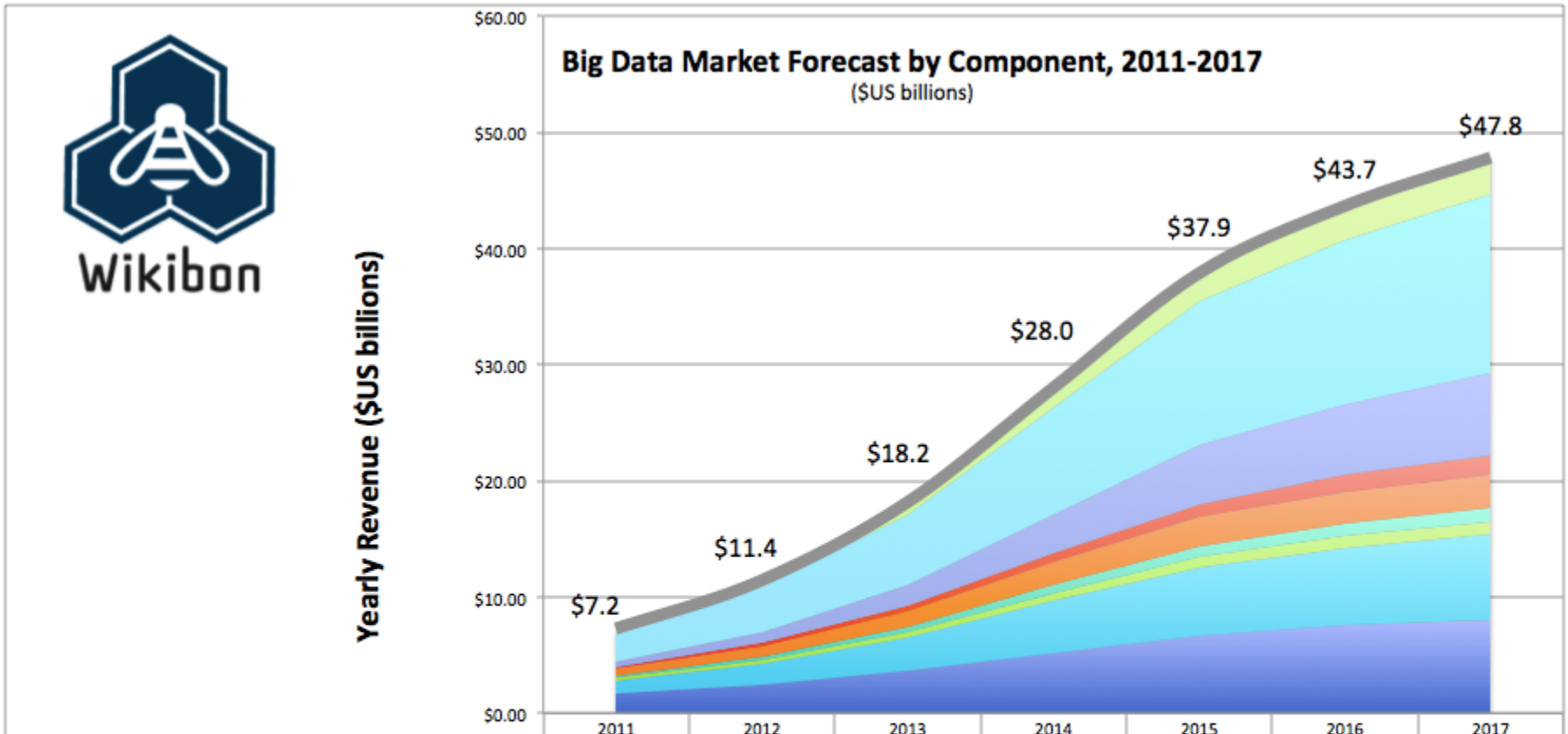


NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

Big Data Market Forecast (2011–2017)

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



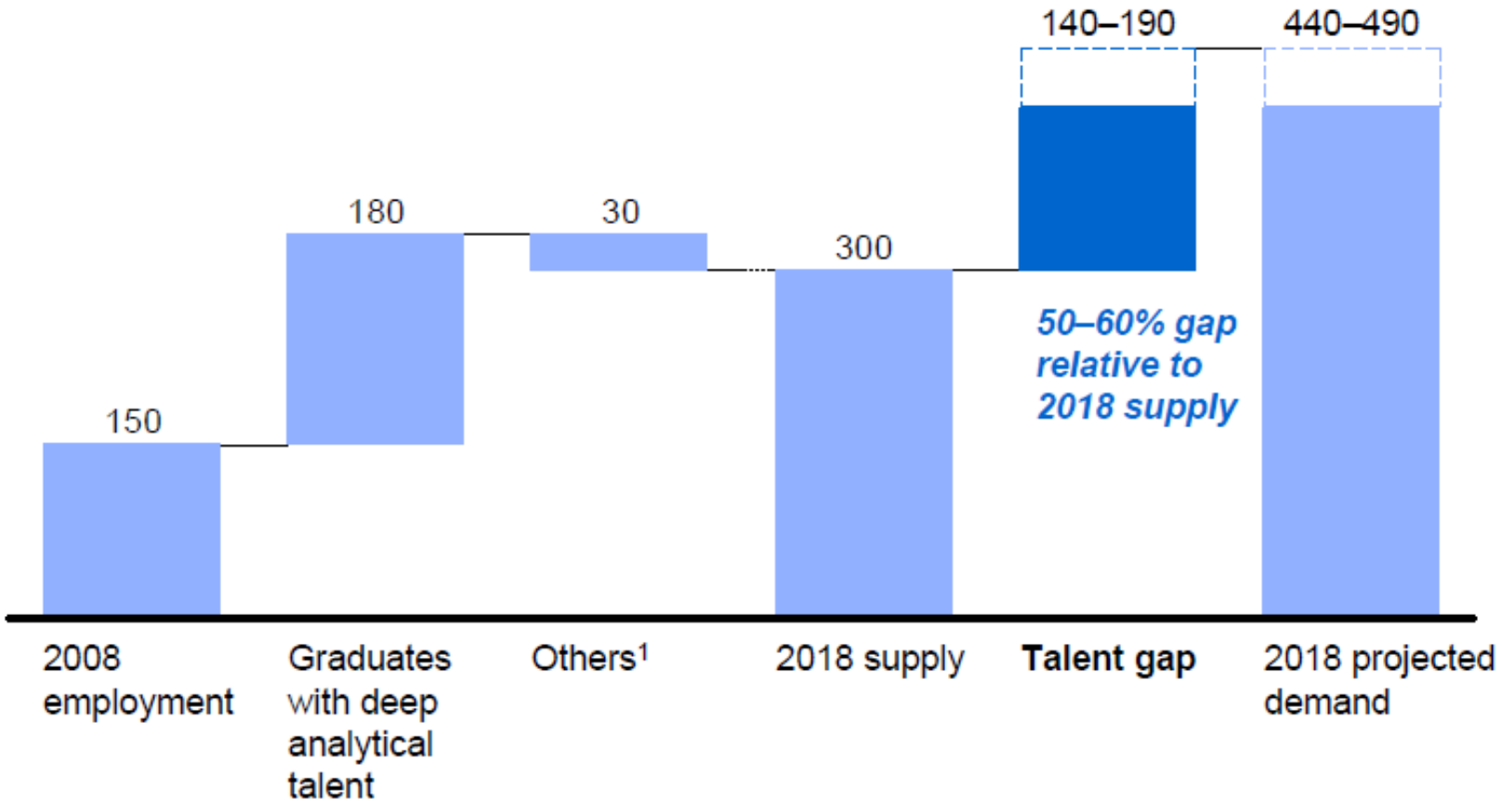
	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.35	$0.61	$1.05	$1.74	$2.47	$2.91	$3.24
Big Data Professional Services Revenue	$2.45	$3.87	$6.10	$9.29	$12.37	$14.14	$15.38
Big Data Application (Analytic and Transactional) Software	$0.49	$0.94	$1.80	$3.29	$5.02	$6.15	$7.00
Big Data NoSQL Database 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

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

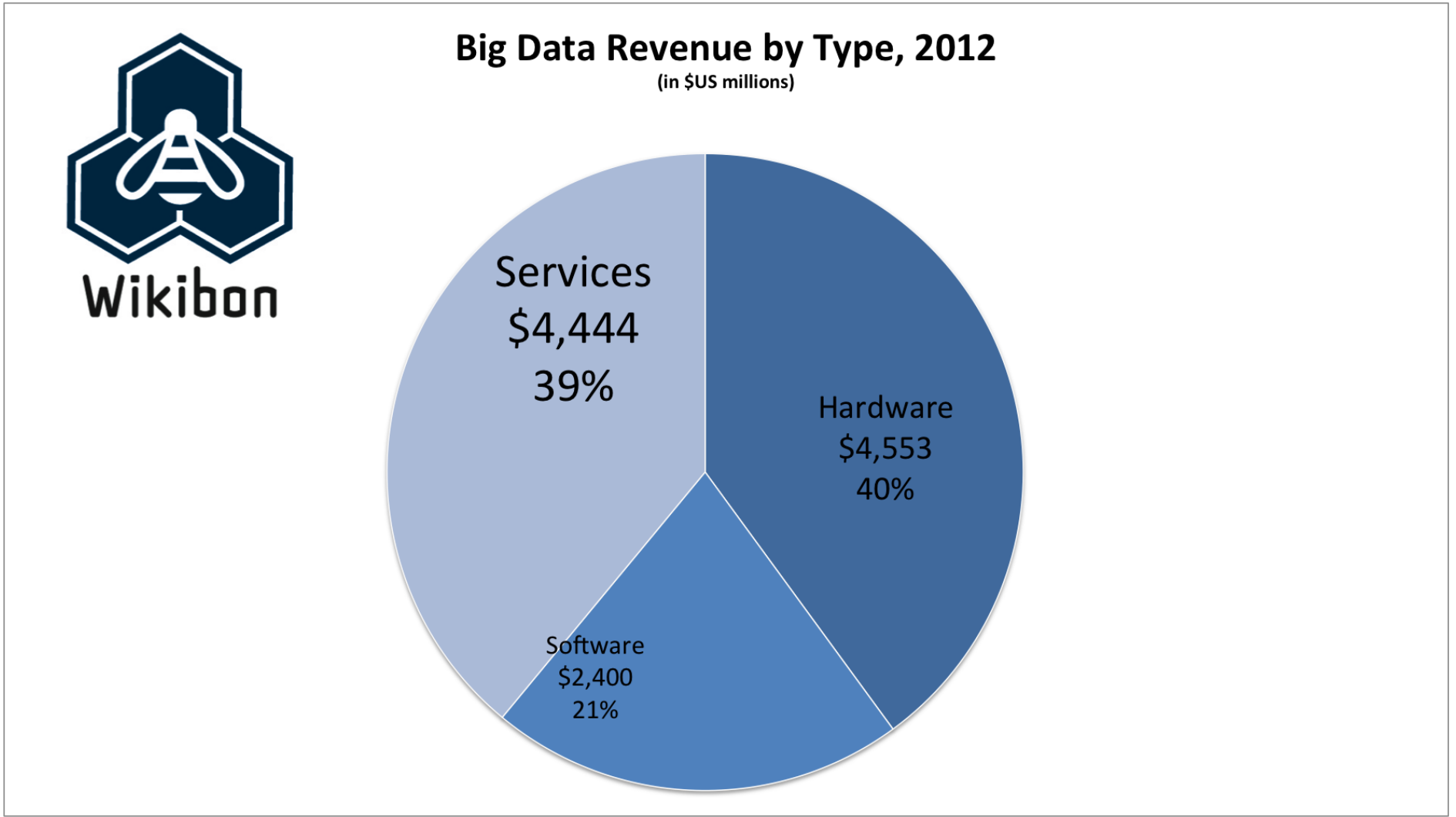
2012 Worldwide Big Data Revenue by Vendor ($US millions)

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)



Types of tools typically used in Big-Data scenarios

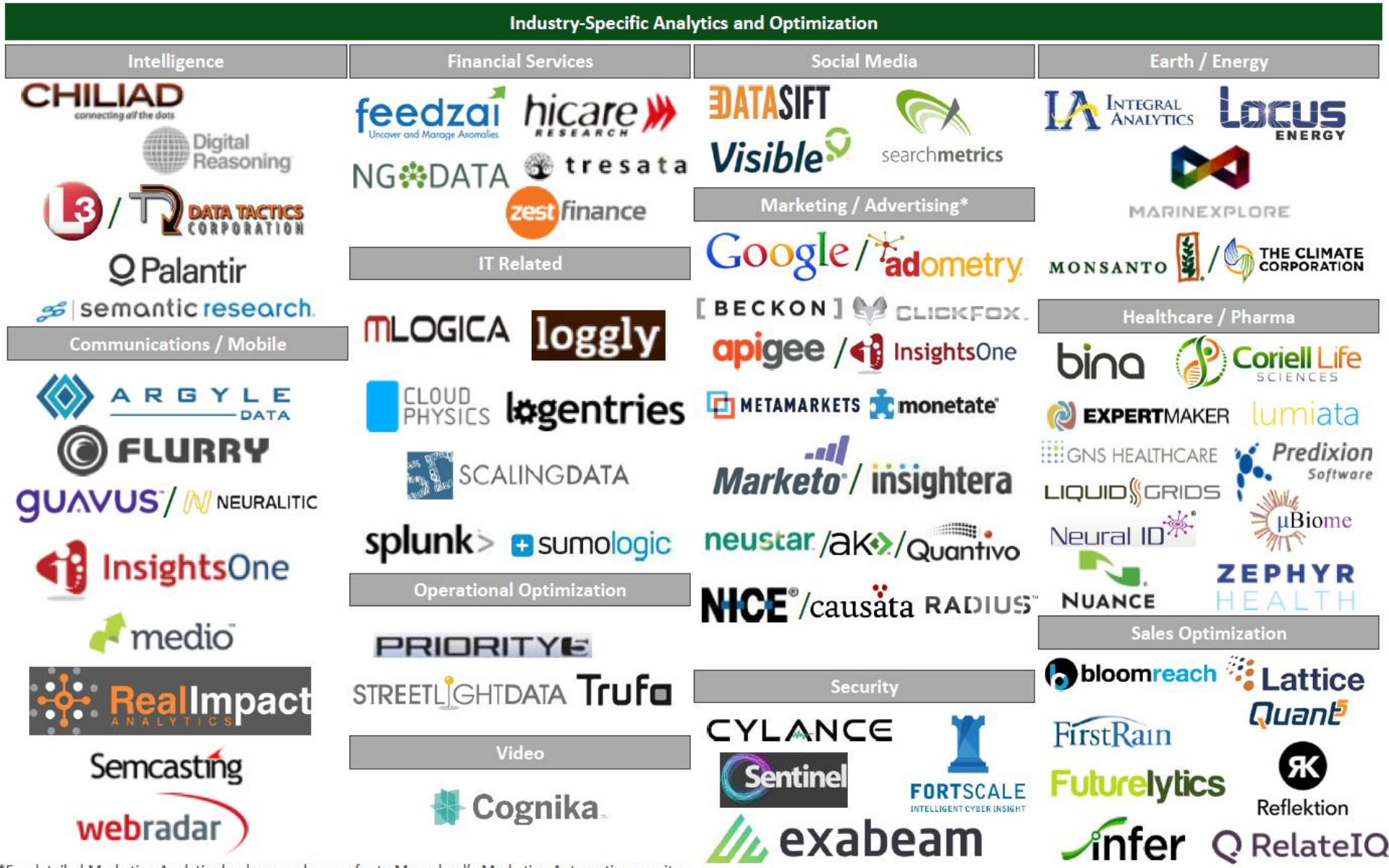
- Where processing is **hosted**?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is **stored**?
  - Distributed Storage (e.g. Amazon S3)
- What is the **programming model**?
  - Distributed Processing (e.g. MapReduce)
- How data is **stored & indexed**?
  - High-performance schema-free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

Landscape of Big Data tools (1/2)

(Mooreland Monitor)

Landscape of Big Data tools (2/2)

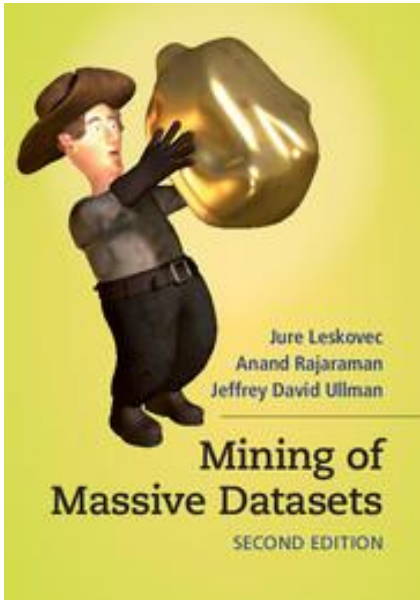
(Mooreland Monitor)



*For detailed Marketing Analytics landscape, please refer to Mooreland's Marketing Automation monitor

Guide to Big-Data algorithms

- An excellent overview of the "Big Data" algorithms is the book "**Leskovec, Rajaraman, Ullman: Mining of Massive Datasets**"
  - Downloadable from: http://www.mmds.org/
  - Associated MOOC (from Oct 2014): https://www.coursera.org/course/mmds



…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)