

# Tutorial: Introduction to Big Data

Marko Grobelnik

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

Jozef Stefan Institute, Slovenia

# Big-Data in numbers

# Big data—a growing torrent

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

**5 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

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

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



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**



320+ NEW **twitter** ACCOUNTS

100+ NEW **Linked in** ACCOUNTS

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!

1 associated content  
NEW ARTICLE IS PUBLISHED

20,000+ NEW POSTS ON **tumblr**

13,000+ **iPhone** APPLICATIONS DOWNLOADED



QUESTIONS ASKED ON THE INTERNET...

100+ 40+  
**Answers.com** **YAHOO! ANSWERS**



600+ NEW VIDEOS

25+ HOURS TOTAL DURATION

70+ DOMAINS REGISTERED

60+ NEW BLOGS

168 MILLION EMAILS ARE SENT

694,445 SEARCH QUERIES

1,700+ **Firefox** DOWNLOADS

695,000+ **facebook** STATUS UPDATES

50+ **WORDPRESS** DOWNLOADS

6,600+ NEW PICTURES ARE UPLOADED ON **flickr**



125+ **PLUGIN** DOWNLOADS

79,364 **WALL** POSTS

510,040 **COMMENTS**



1,500+ **BLOG** POSTS



**Google**

Google Search



# HOW PEOPLE - SPEND THEIR TIME - ONLINE



GLOBAL ONLINE POPULATION  
**2,095,006,005**

=



**30%**  
of World's  
Population.



GLOBAL TIME SPENT ONLINE / MONTH

**35 BILLION**

WHICH IS EQUIVALENT TO

**3,995,444**  
YEARS

## AVERAGE TIME SPENT BY :

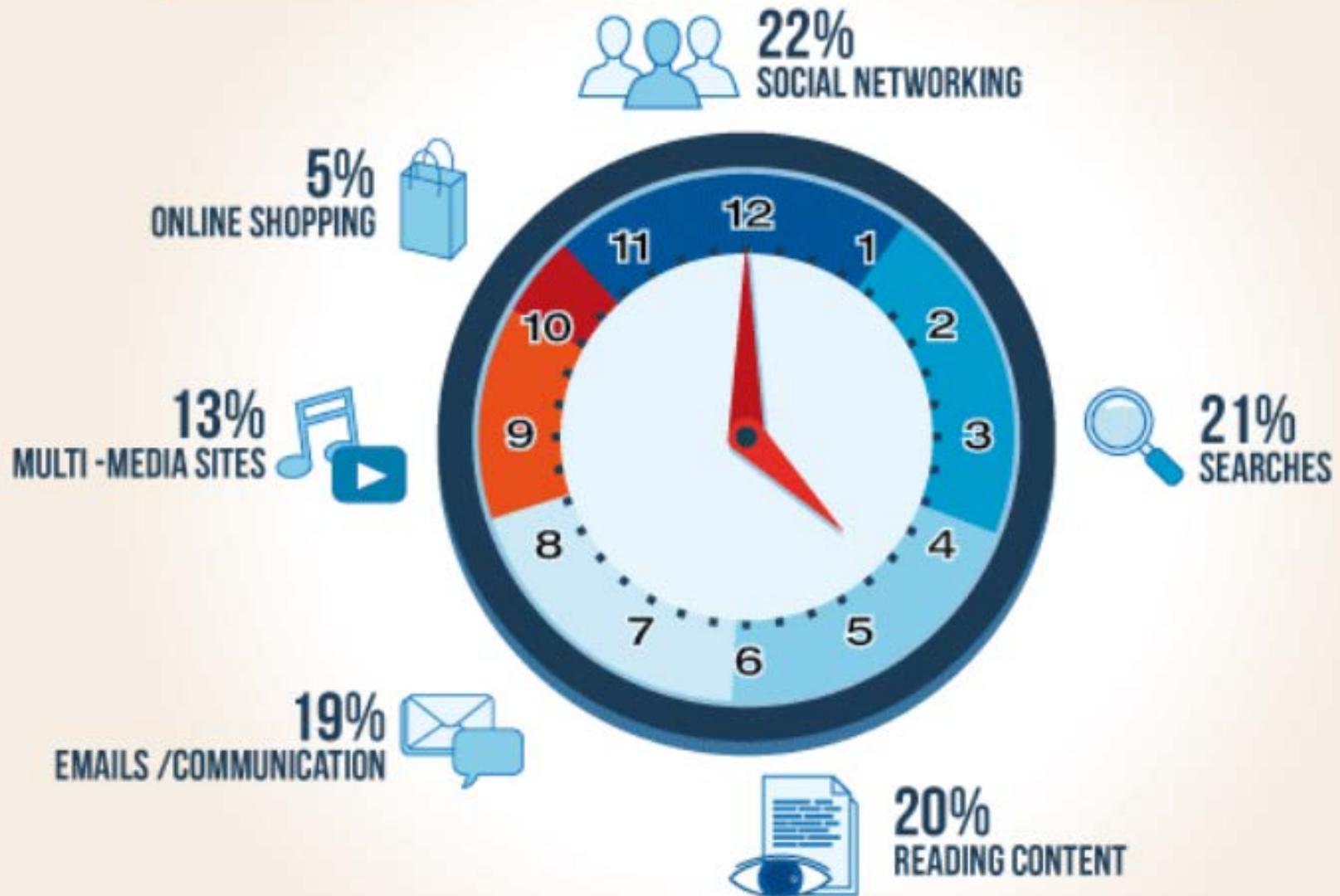
Global Internet user  
per month: **16 HOURS**



US Internet user  
per month: **32 HOURS**



# HOW PEOPLE SPEND THEIR TIME



# Big-Data Definitions

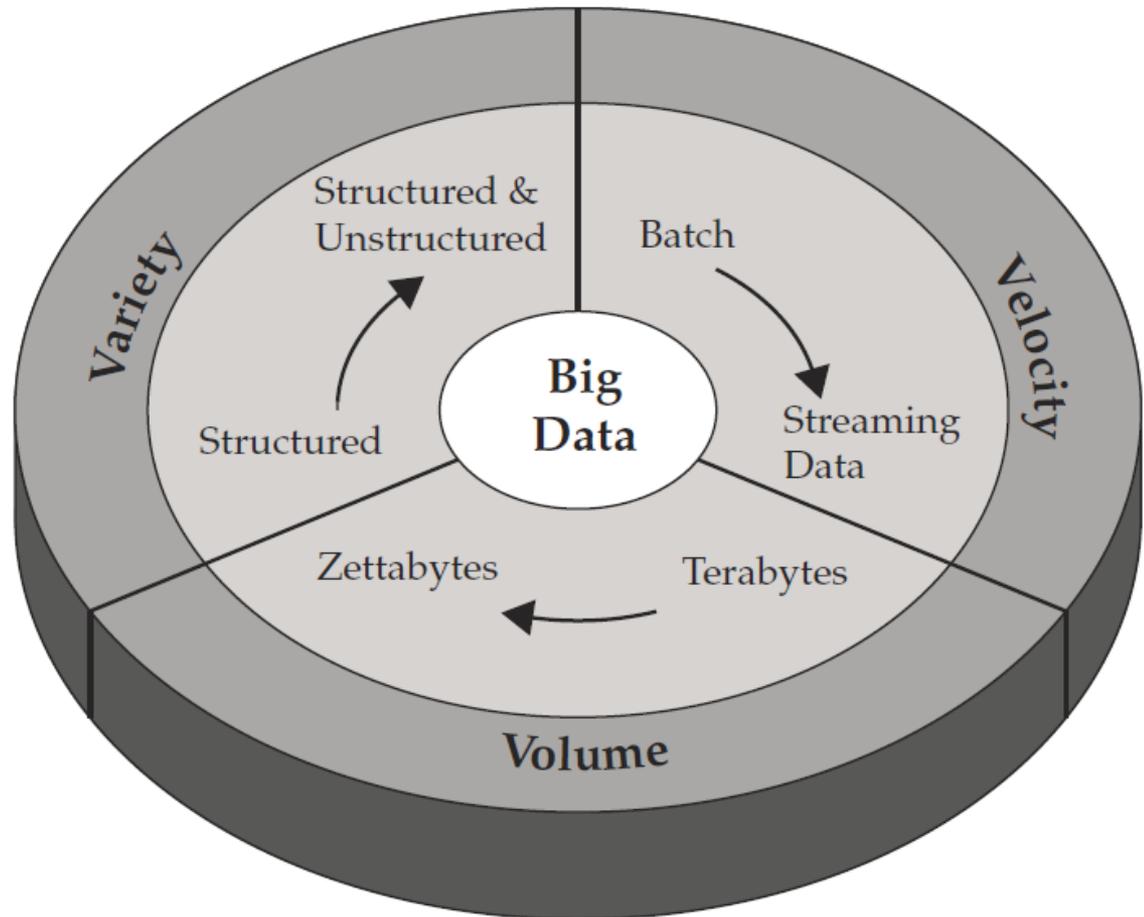
# ...so, what is Big-Data?

- ▶ ‘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

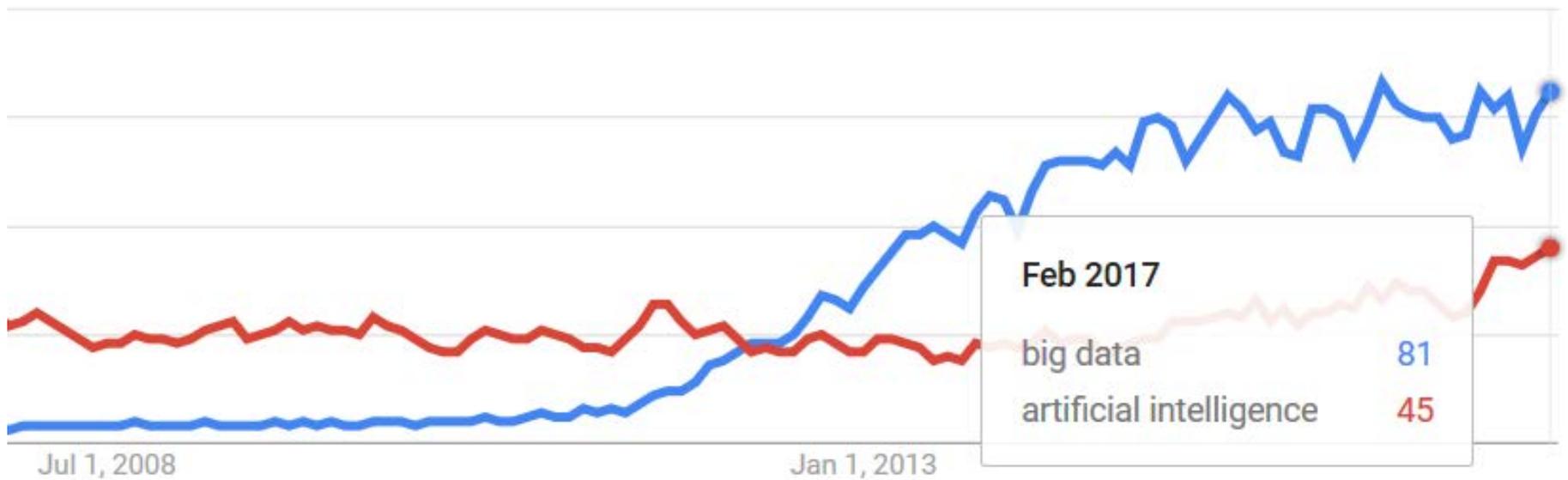
# The extended 3+n Vs of Big Data

- ▶ 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”)

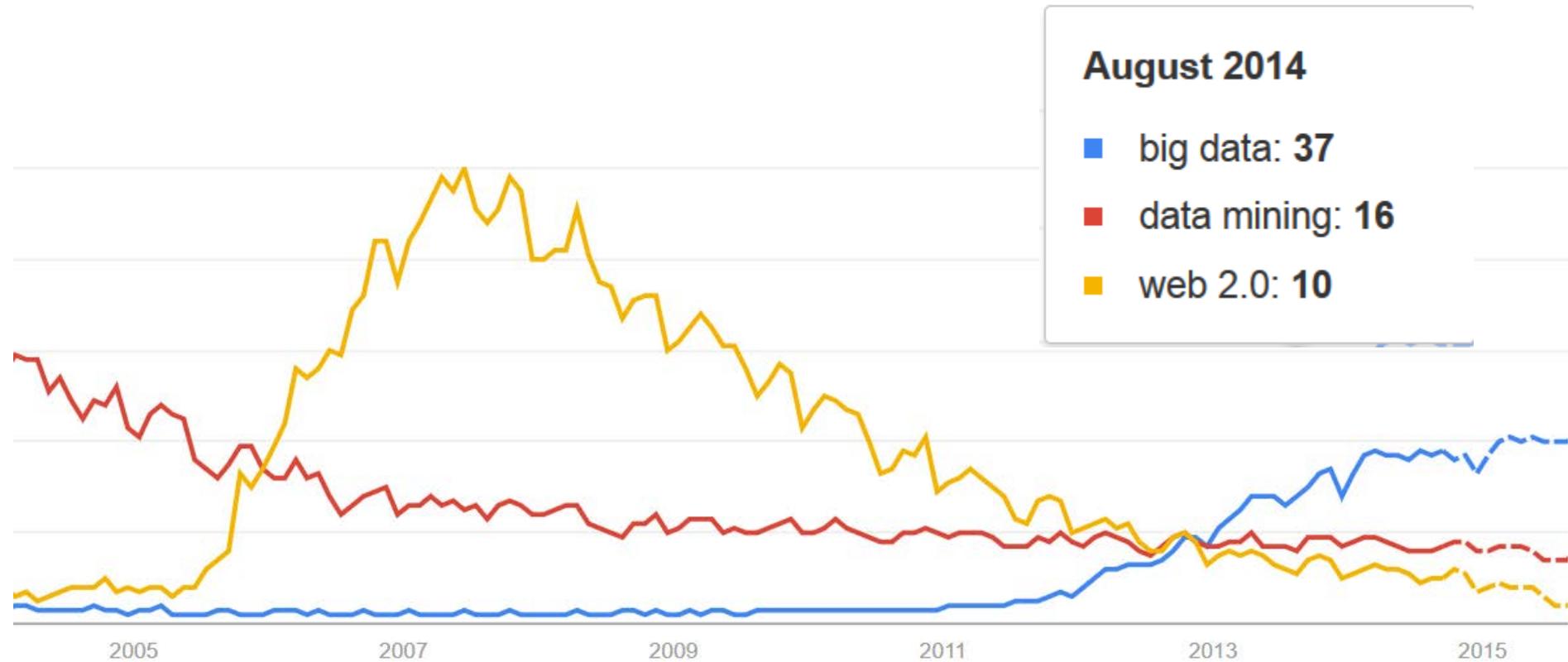
Comparing volume of “big data” and “artificial intelligence” queries



<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

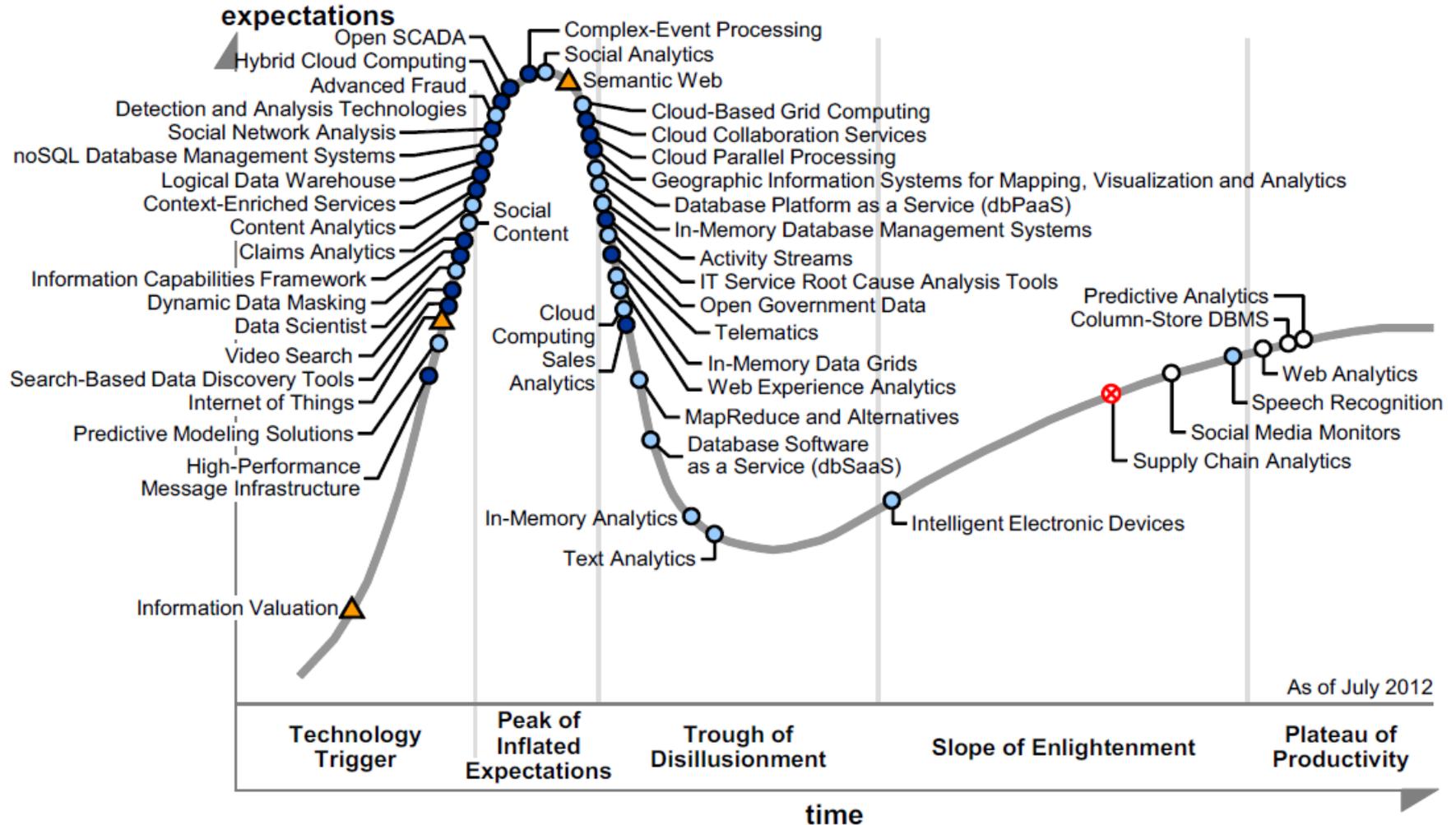


# Structure of Big Data:

## Gartner Hype Cycle for Big Data, 2012

(later Gartner stopped producing Big Data Hype Cycle)

Figure 1. Hype Cycle for Big Data, 2012



Plateau will be reached in:

○ less than 2 years

● 2 to 5 years

● 5 to 10 years

▲ more than 10 years

⊗ obsolete before plateau

# What about the future?

## Gartner Hype Cycle 2016

Figure 1. Hype Cycle for Emerging Technologies, 2016



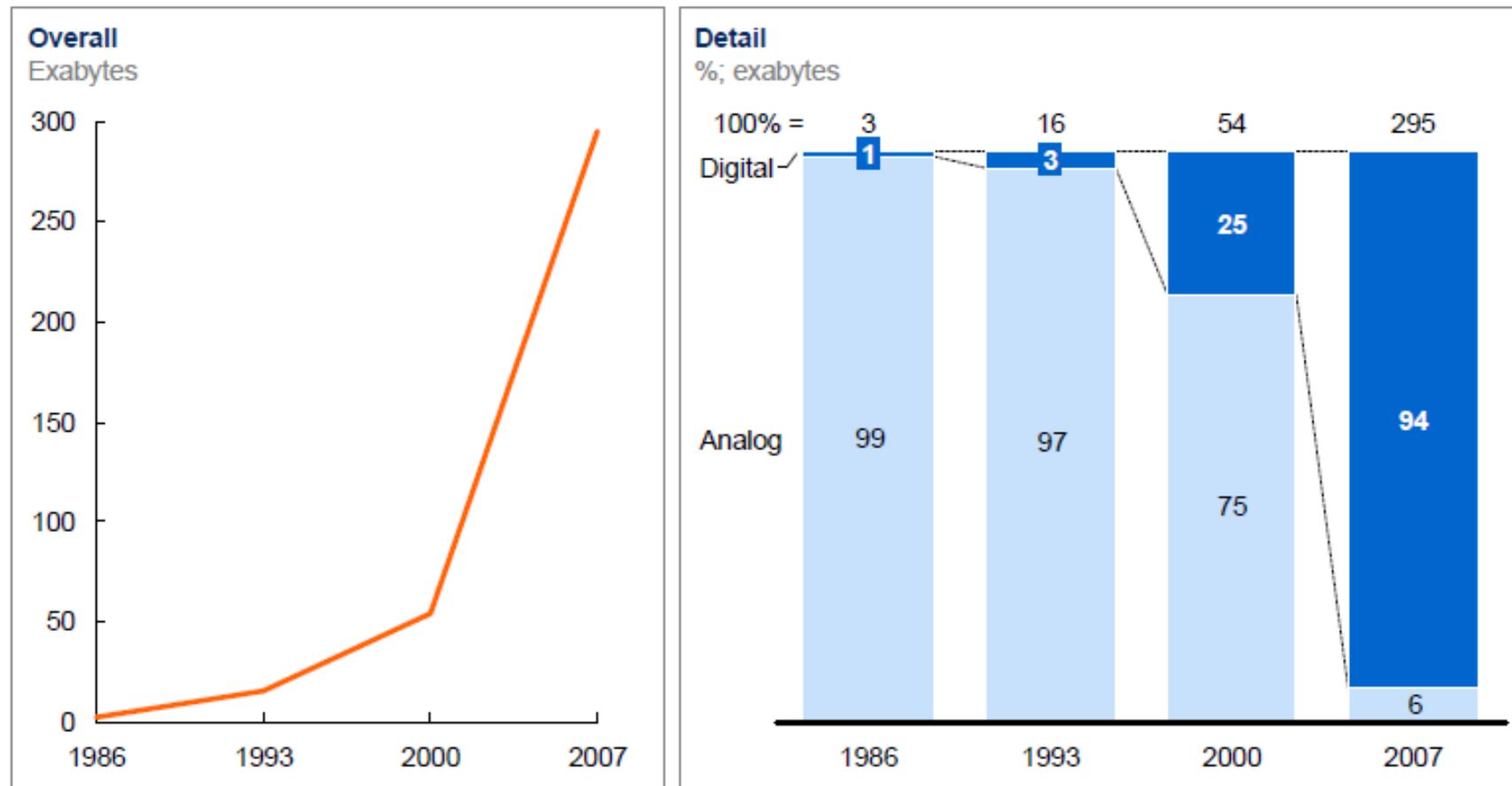
# Why Big-Data?

- ▶ 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



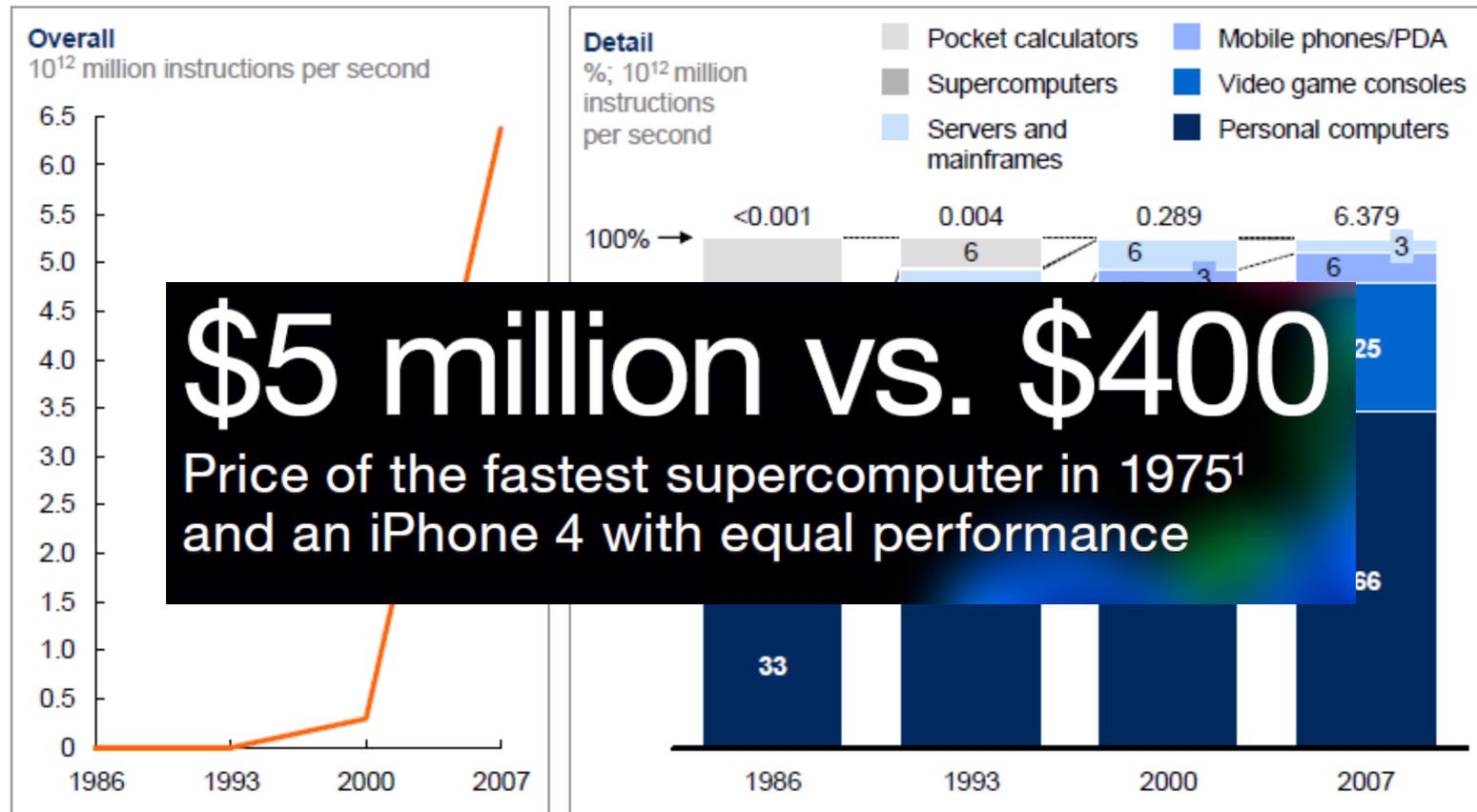
NOTE: Numbers may not sum due to rounding.

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

# Enabler: Computation capacity

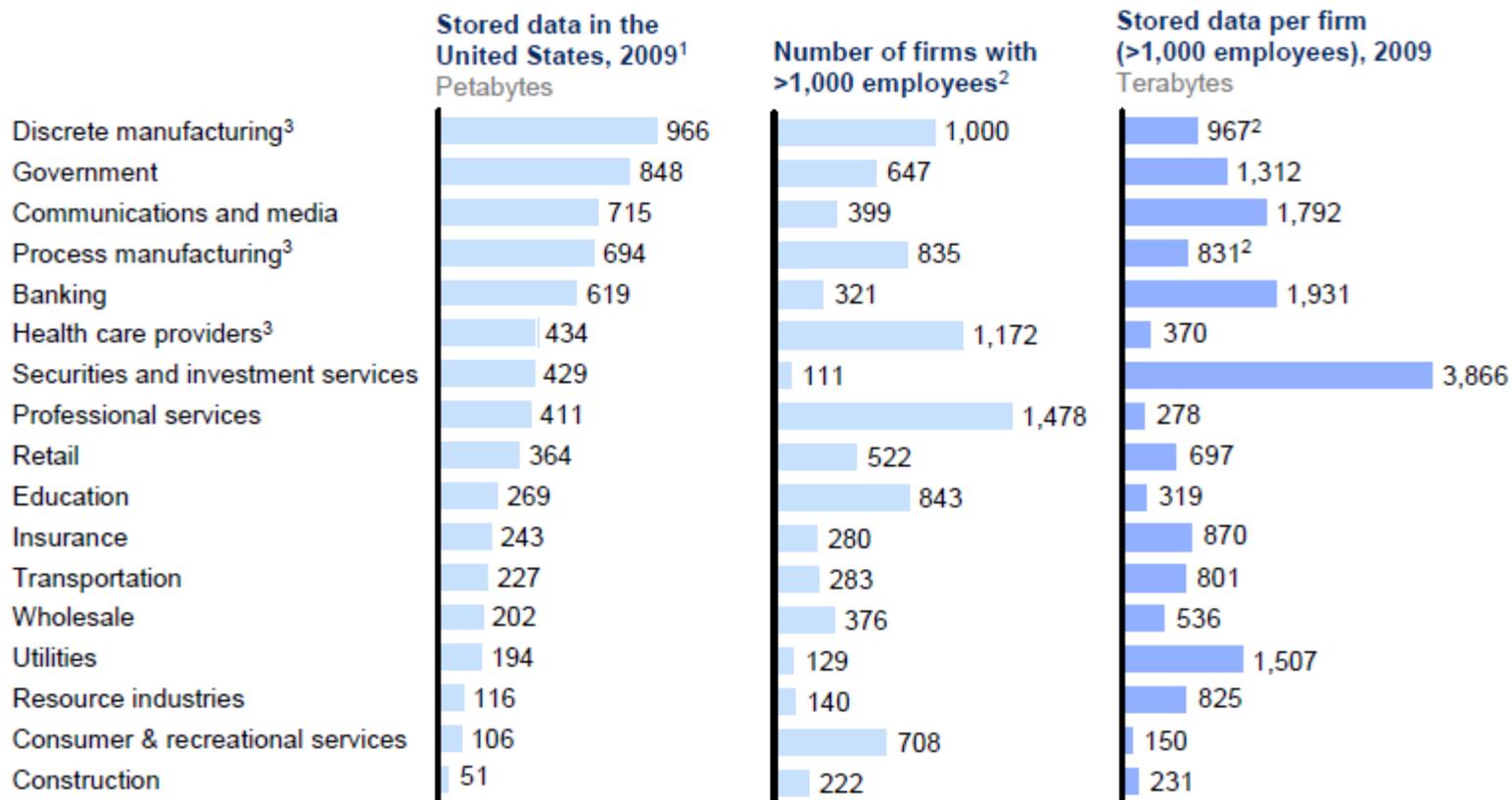
## Computation capacity has also risen sharply

Global installed computation to handle information



# Enabler: Data availability

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



1 Storage data by sector derived from IDC.

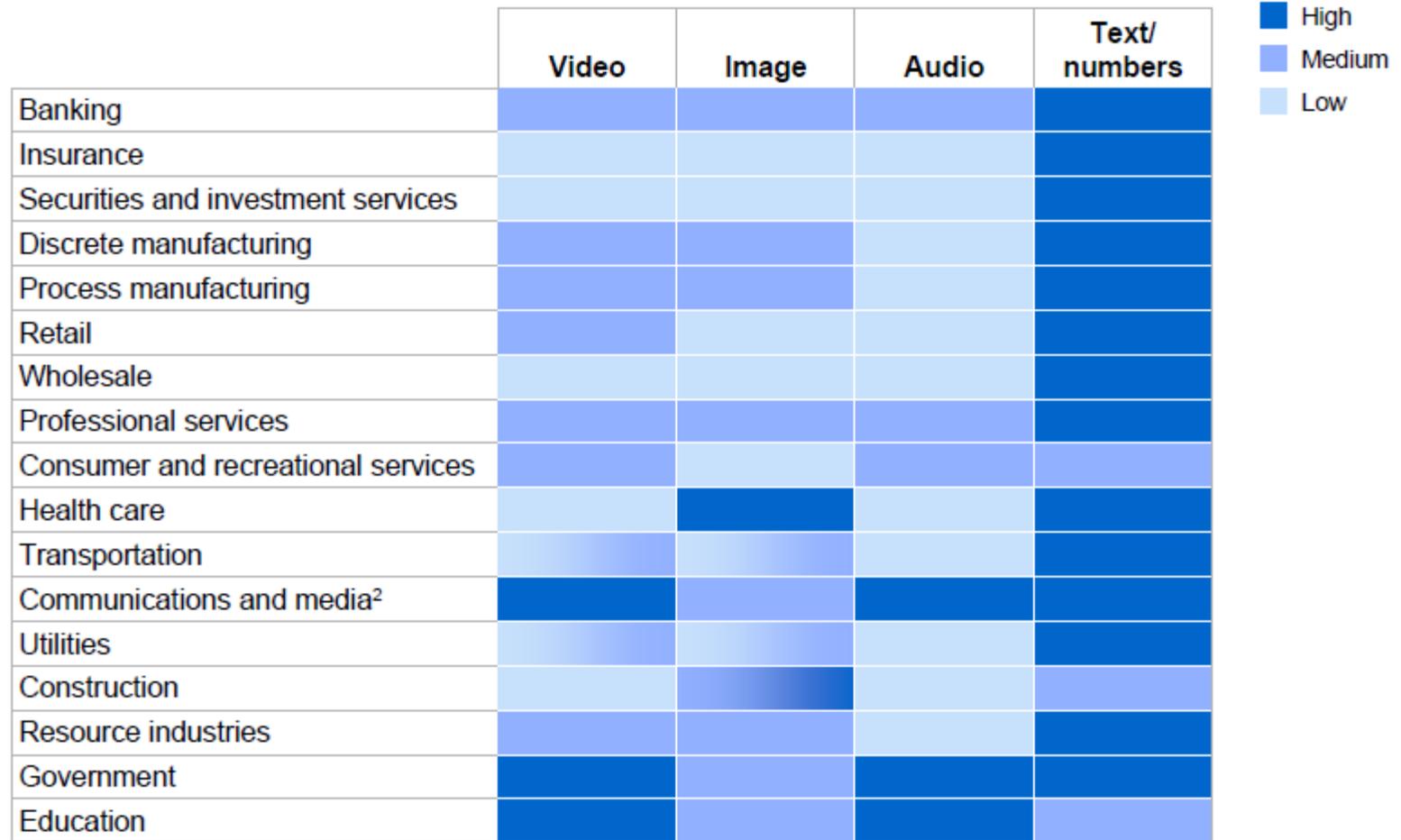
2 Firm data split into sectors, when needed, using employment

3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

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

# Type of available data

The type of data generated and stored varies by sector<sup>1</sup>



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

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

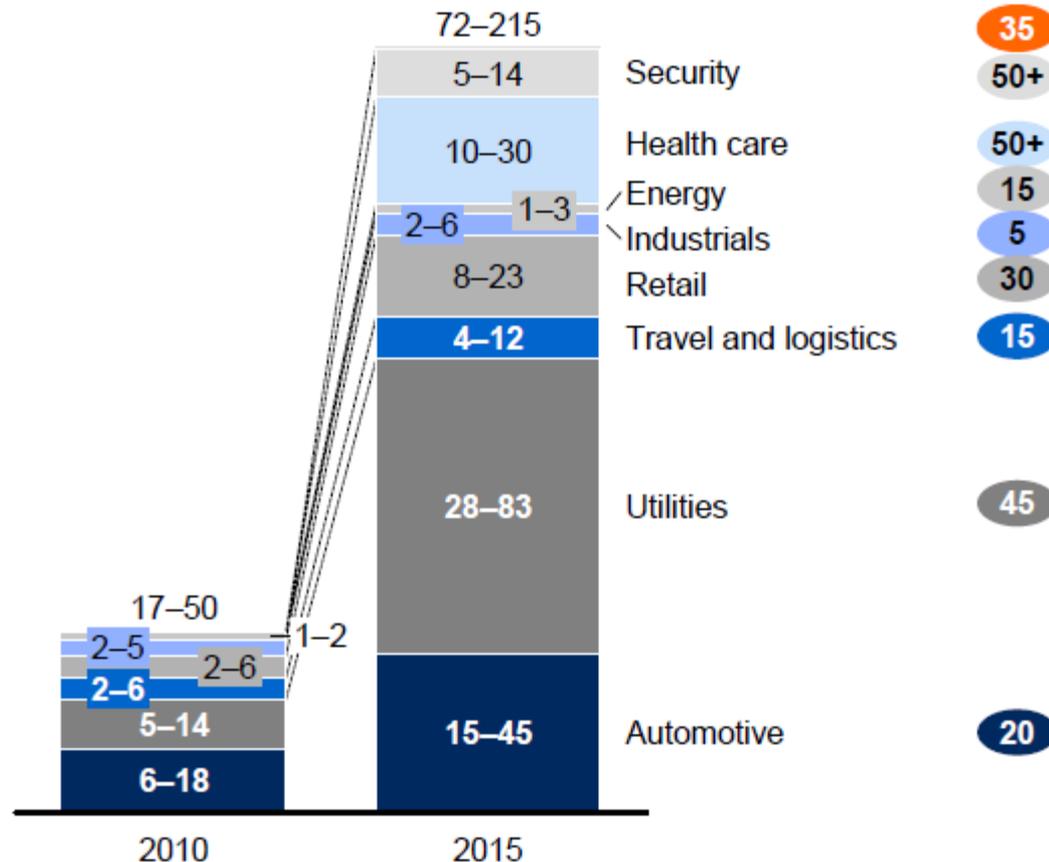
SOURCE: McKinsey Global Institute analysis

# Data available from “Internet of Things”

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

Estimated number of connected nodes  
Million

Compound annual  
growth rate 2010–15, %

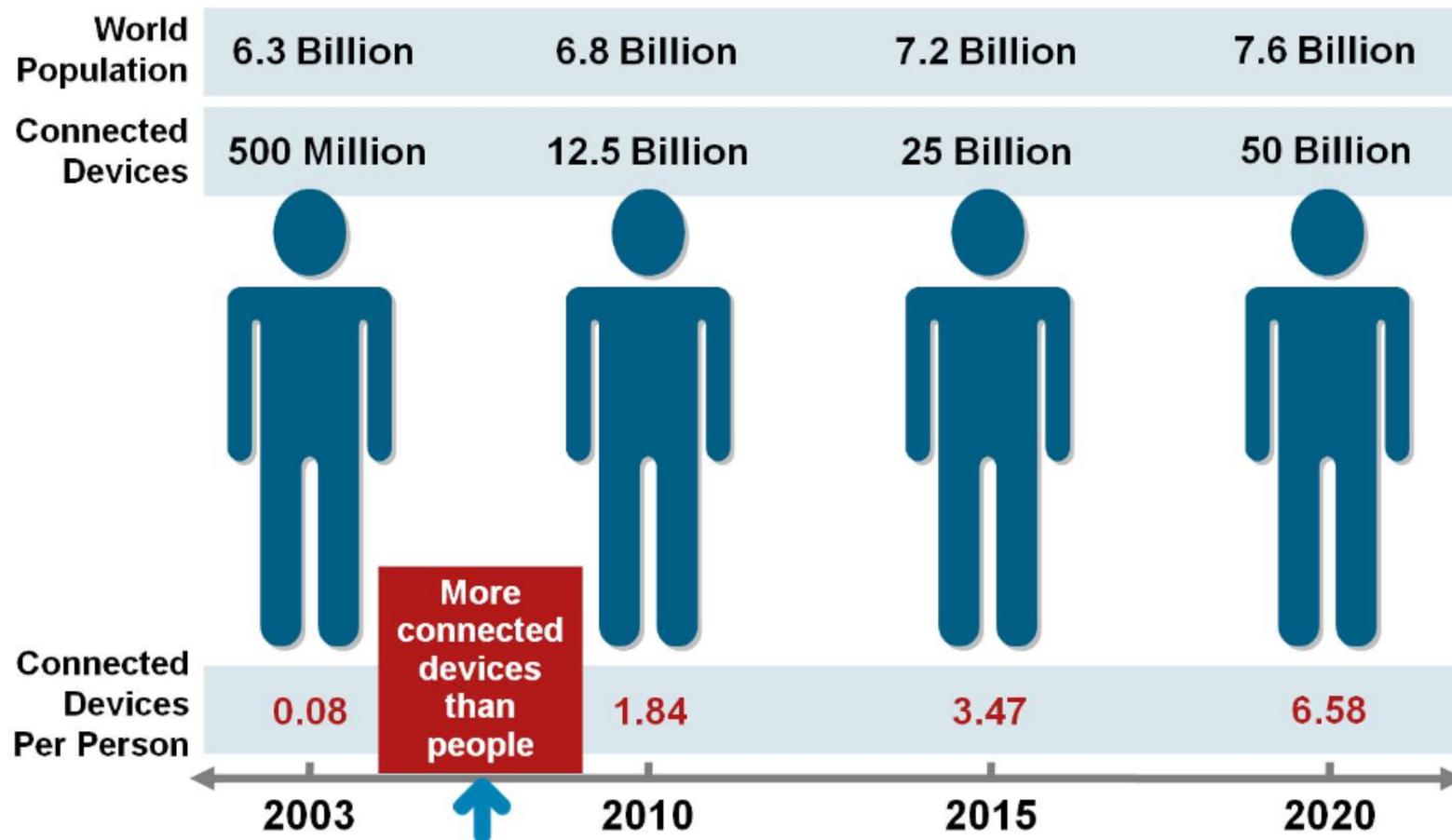


NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

# Birth & Growth of “Internet of Things”

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

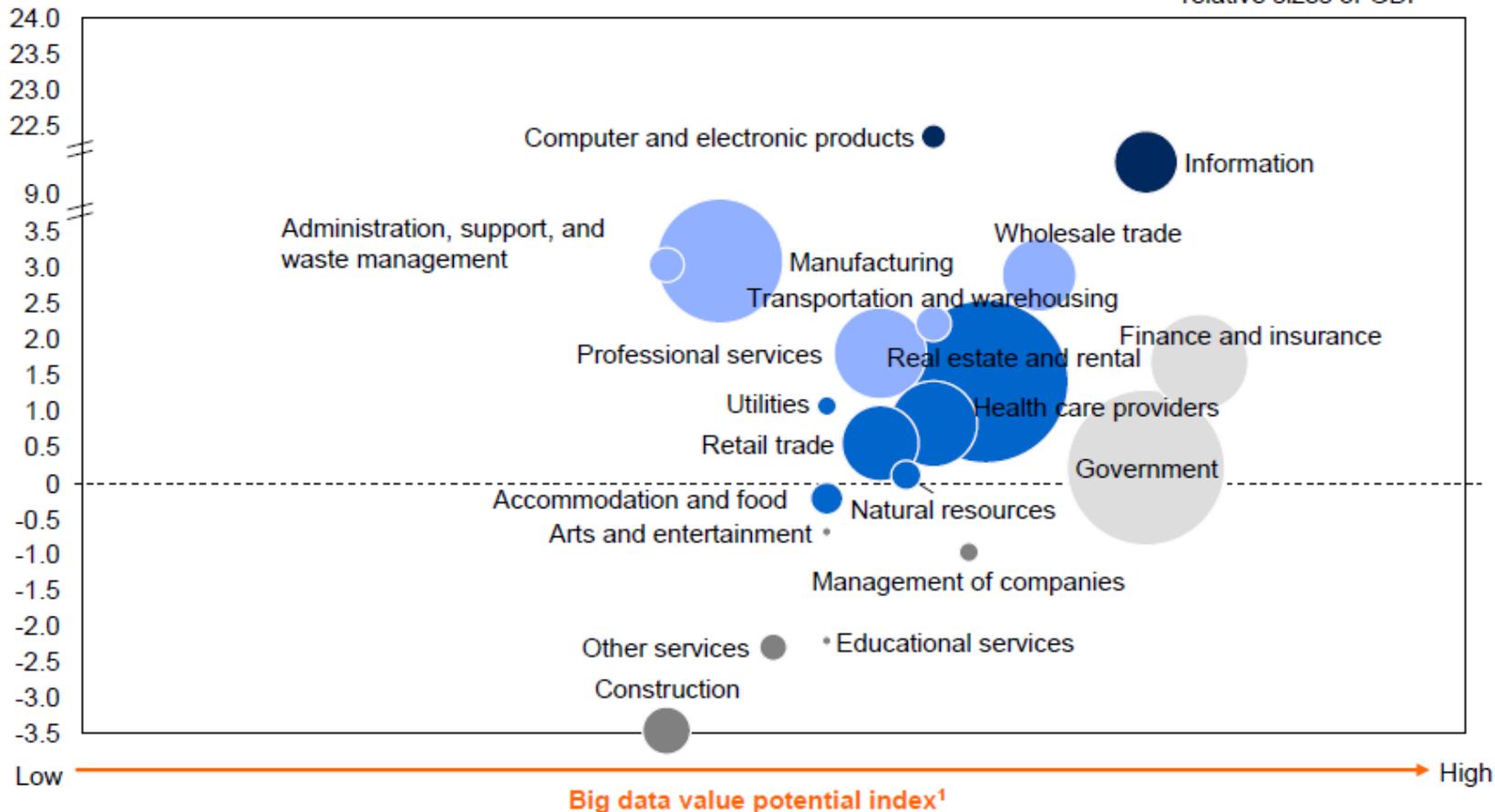


# Gains from Big-Data per sector

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.

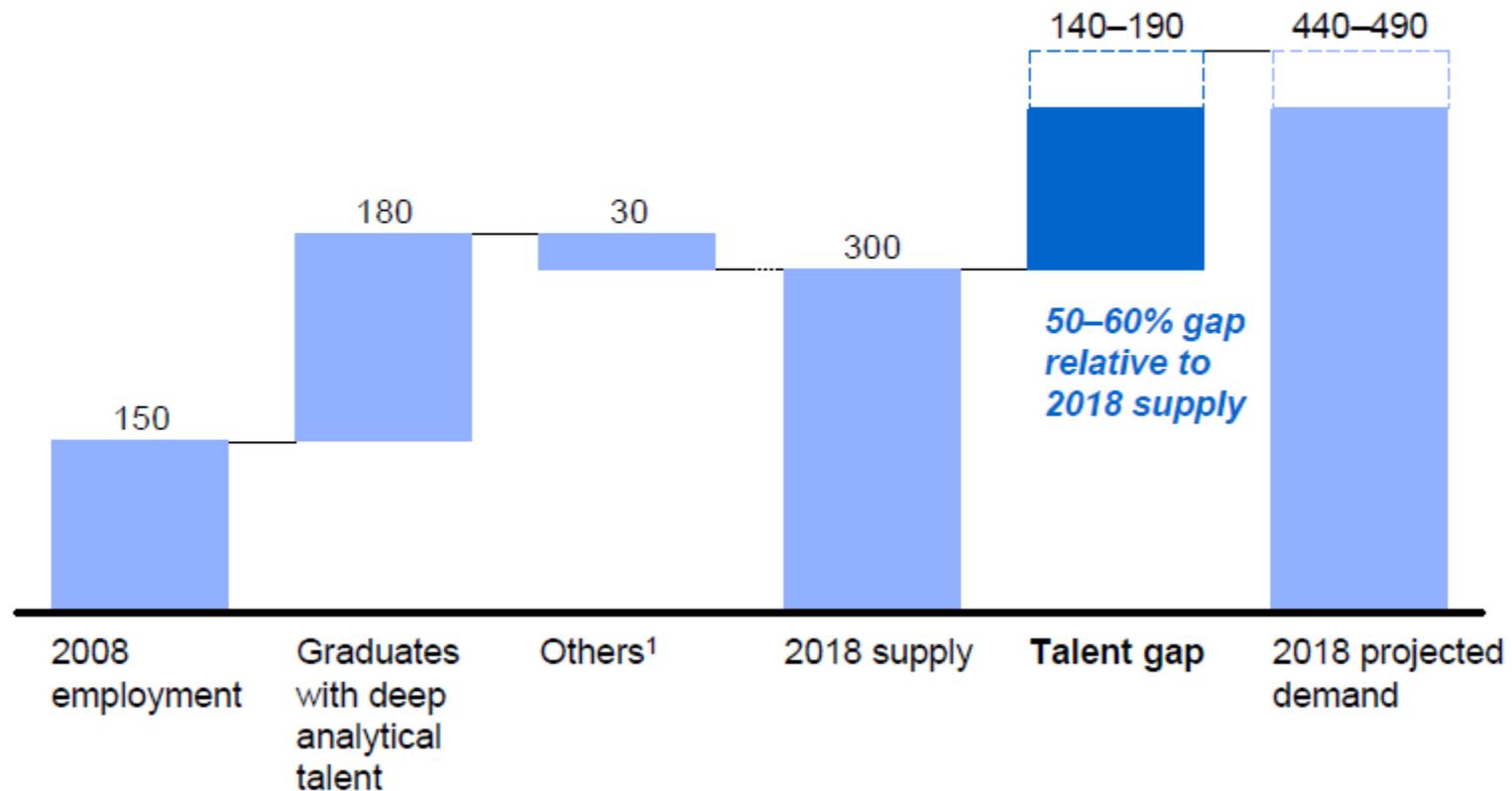
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

# 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



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

# Big Data Market

# Big Data Landscape 2016 (Version 3.0)

## Infrastructure

**On-Premise**  
 cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

**Hadoop in the Cloud**  
 amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubale

**Spark**  
 databricks, GridGain, TACHYON NEXUS

**Cluster Services**  
 amazon, kubernetes, docker, HIPCC SYSTEMS, MESOSPHERE, CoreOS, pepperdata, StackIQ

## Analytics

**Analyst Platforms**  
 Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

**Analytics Platforms**  
 Microsoft, GUAVUS, Datameer, Bottlenose, interlana

**Data Science Platforms**  
 context relevant, DataRobot, CONTINUUM ANALYTICS, Alpine, MODE, plotly, ARIMO, dataiku, dominio, sense, yhat, ALGORITHMIA

**Visualization**  
 tableau, Google Cloud Platform, Qlik, looker, Roambi, BISSENSE, FOCMDATA, datarama, CHARTIO

## Applications

**Sales & Marketing**  
 RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, bsense, QUANTIFIND, ACTIONIQ, fuse|machines, EN G A G I O

**Customer Service**  
 MEDALLIA, ATTENTIFY, CLARABRIDGE, CLICKFOX, STELLASERVICE, NGDATA, Preact, DigitalGenius, appurfi, Wiseio

**Human Capital**  
 gild, Connectifier, textic, entelo, hiQ

**Legal**  
 RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

**NoSQL Databases**  
 amazon, dynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, <EROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

**NewsSQL Databases**  
 SAP, Clustrix, Pivotal, paradigm4, nuODB, memsql, splice MACHINE, MariaDB, VOLTDDB, citusdata, deepdb, Trafaldrion, Cockroach LABS

**BI Platforms**  
 Power BI, amazon, DOMO, Wave Analytics, GoodData, platforma, atscale, ACADIA, BISSENSE

**Statistical Computing**  
 SAS, SPSS, MATLAB

**Log Analytics**  
 splunk, sumologic, kibana, CLOUD PHYSICS, loggly

**Social Analytics**  
 Hootsuite, NETBASE, DATASIFT, track, bitly, synthesio, simplereach

**Ad Optimization**  
 AppNexus, MediaMath, criticoL, OpenX, rocketfuel, Integral, theTradeDesk, Adgorithms, dstillery, Livelihood, TAFAD, DataXu, Appier, MOAT

**Security**  
 CYCLANCE, CounterTack, cyberreason, AREA 1 SECURITY, ThreatMetrix, Recorded Future, SentinelOne, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SCINIFYD

**Vertical AI Applications**  
 facebook, Clara, KASIST@, lumiatia

**Graph Databases**  
 neo4j, ORIENT DB, InfiniteGraph

**MPP Databases**  
 TERADATA, NETEZZA, Acton, Kognitio, SAS SQL, dremio

**Cloud EDW**  
 amazon, Microsoft Azure, Pivotal, snowflake, WATERLUNE, Infoworks

**Data Transformation**  
 alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

**Data Integration**  
 informatica, MuleSoft, snaplogic, BedrockData, xplenty

**Real-Time**  
 amazon, METAMARKETS, Striim, confluent, DATATORRENT, dataArtisans

**Machine Learning**  
 Azure Machine Learning, H2O, amazon, SKYTRIP, rapidminer, DATAROBOT, deepinsight, VISENZE, PredictionIO, glowfish

**Speech & NLP**  
 NarrativeScience, NUANCE, semantic machines, ARRIA, api.ai, corticoL, maluba, MindMeld, IDIBON, vscope

**Horizontal AI**  
 IBM Watson, Cortana, sentient, viv, nervana, vicarious, nora, Numenta, Discretas Labs, clarifai, META-MIND

**Publisher Tools**  
 Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

**Govt / Regulation**  
 Socrata, OPENGOV, FN FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

**Finance**  
 affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tdemark, Payoff, INSIKT, uora, Dataminr, Lendio, KENSHO, AIDYIA, ISENTIUM, Quantopian, sentient technologies

**Management / Monitoring**  
 New Relic, APPDYNAMICS, actifio, Numerify, splunk, DATADOG, FROCANO, DRIVEN, Anodot

**Security**  
 TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrl, BlueTalon

**Storage**  
 amazon, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

**App Dev**  
 apigee, GASK, KitemIO, Typesafe, DRIVEN

**Crowd-sourcing**  
 amazon, mechanicalturk, CrowdFlower, WorkFusion

**Search**  
 hp, ORACLE, ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

**Data Services**  
 OPERA, MU SIGMA, EXL, DATA SCIENCE, kaggle, datascope, DataKind

**For Business Analysts**  
 OrigamiLogic, ClearStory, CIRRO, import io

**Web / Mobile / Commerce**  
 Google Analytics, mixpanel, R.J. Metrics, BLUECORE, AMPLITUDE, granify, sumall, Airtable, retention custora

**Education / Learning**  
 KNEWTON, Clever, Declara, PANORAMA, knowre

**Life Sciences**  
 23andMe, COUNSYL, PATHWAY GENOMICS, RECOMBINE, FLATIRON, KYRUUS, HealthTap, zymogen, ZEPHYR HEALTH, ovia, METABIOTA, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

**Industries**  
 OPOWER, eHarmony, RetailNext, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, Seeq, FarmLogs, SwiftKey, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

## Cross-Infrastructure/Analytics

amazon, Google, Microsoft, IBM, SAP, sas, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

## Open Source

**Framework**  
 hadoop, HADOOP HOPS, HADOOP IMPERIAL, YARN, MESOS, TEZ, Flink, CDAP

**Query / Data Flow**  
 SLAMDATA, ADVANTAGE DRILL, Google Cloud Dataflow, cassandra, CouchDB, riak, OPENTSDB, nifi

**Data Access**  
 accumulo, mongoDB, kafka, nifi

**Coordination**  
 talend, Apache Zookeeper, Apache Ambari

**Real-Time**  
 STORM, Spark, APEX, Flink, TACHYON, druid

**Stat Tools**  
 ScalaLab, Numpy, SciPy

**Machine Learning**  
 mllib, Apache SINGA, MADlib, Aerosolve, Caffe, CNTK, TensorFlow, VELES, WEKA, FeatureFu, DIMSUM, jupyter, DL4J

**Search**  
 elasticsearch, Solr, Lucene

**Security**  
 Apache Ranger, Zeppelin

## Data Sources & APIs

**Health**  
 Apple, JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

**IOT**  
 UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

**Financial & Economic Data**  
 Bloomberg, DOW JONES, THOMSON REUTERS, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, StockTwits, estimote, PLAID

**Air / Space / Sea**  
 PLANET LABS, spire, WINDWARD, CRUISE, SKY CATCH, Airware, DroneDeploy

**Location / People / Entities**  
 axiomatic, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, Crism Hexagon, CARTODB, factual, PlaceIQ, CIRCLATE, placemeter, BASIS, Sense

**Other**  
 qualtrics, panjiva, DATA.GOV

**Incubators & Schools**  
 GA, PLURALSIGHT, DataCamp, DataElite, The Data Incubator, METIS

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

Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	\$1,352	\$103,930	1%	22%	33%	44%
HP	\$664	\$119,895	1%	34%	29%	38%
Teradata	\$435	\$2,665	16%	31%	28%	41%
Dell	\$425	\$59,878	1%	83%	0%	17%
Oracle	\$415	\$39,463	1%	25%	34%	41%
SAP	\$368	\$21,707	2%	0%	67%	33%
EMC	\$336	\$23,570	1%	24%	36%	39%
Cisco Systems	\$214	\$47,983	0%	80%	0%	20%
Microsoft	\$196	\$71,474	0%	0%	67%	33%
Accenture	\$194	\$29,770	1%	0%	0%	100%
Fusion-io	\$190	\$439	43%	71%	0%	29%
PwC	\$189	\$31,500	1%	0%	0%	100%
SAS Institute	\$187	\$2,954	6%	0%	59%	41%

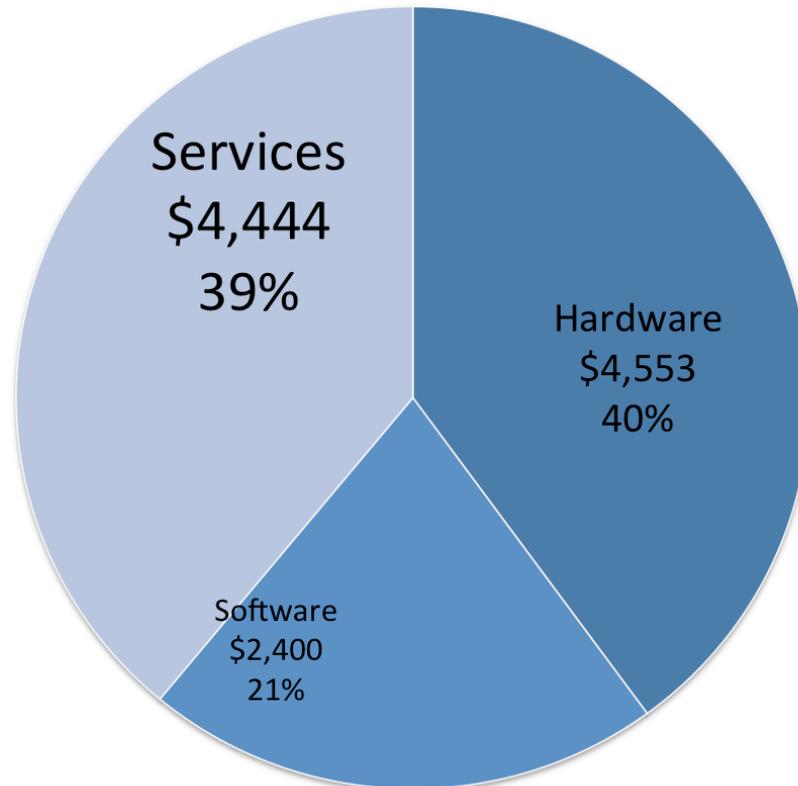
Source: WikiBon report on “Big Data Vendor Revenue and Market Forecast 2012–2017”, 2013

# Big Data Revenue by Type, 2012

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



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

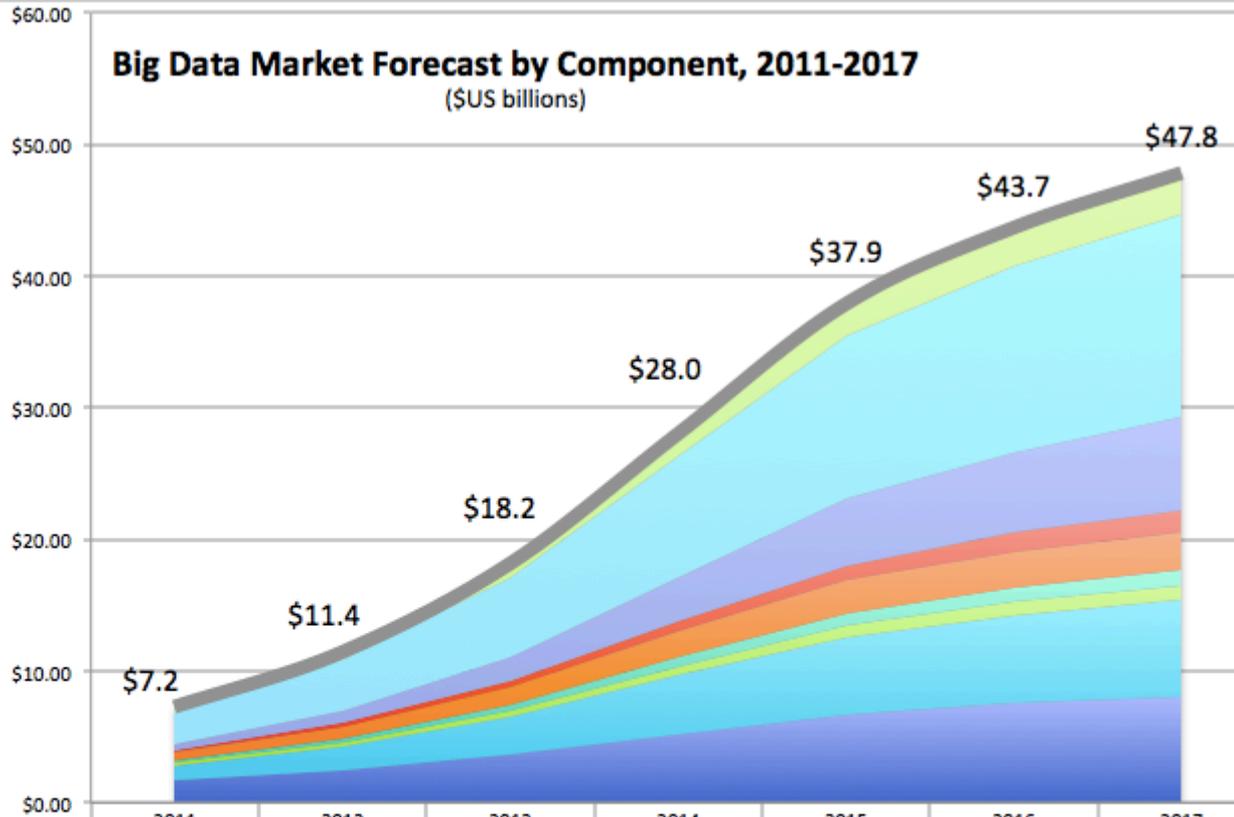


# Big Data Market Forecast (2011–2017)

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



Yearly Revenue (\$US billions)



	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

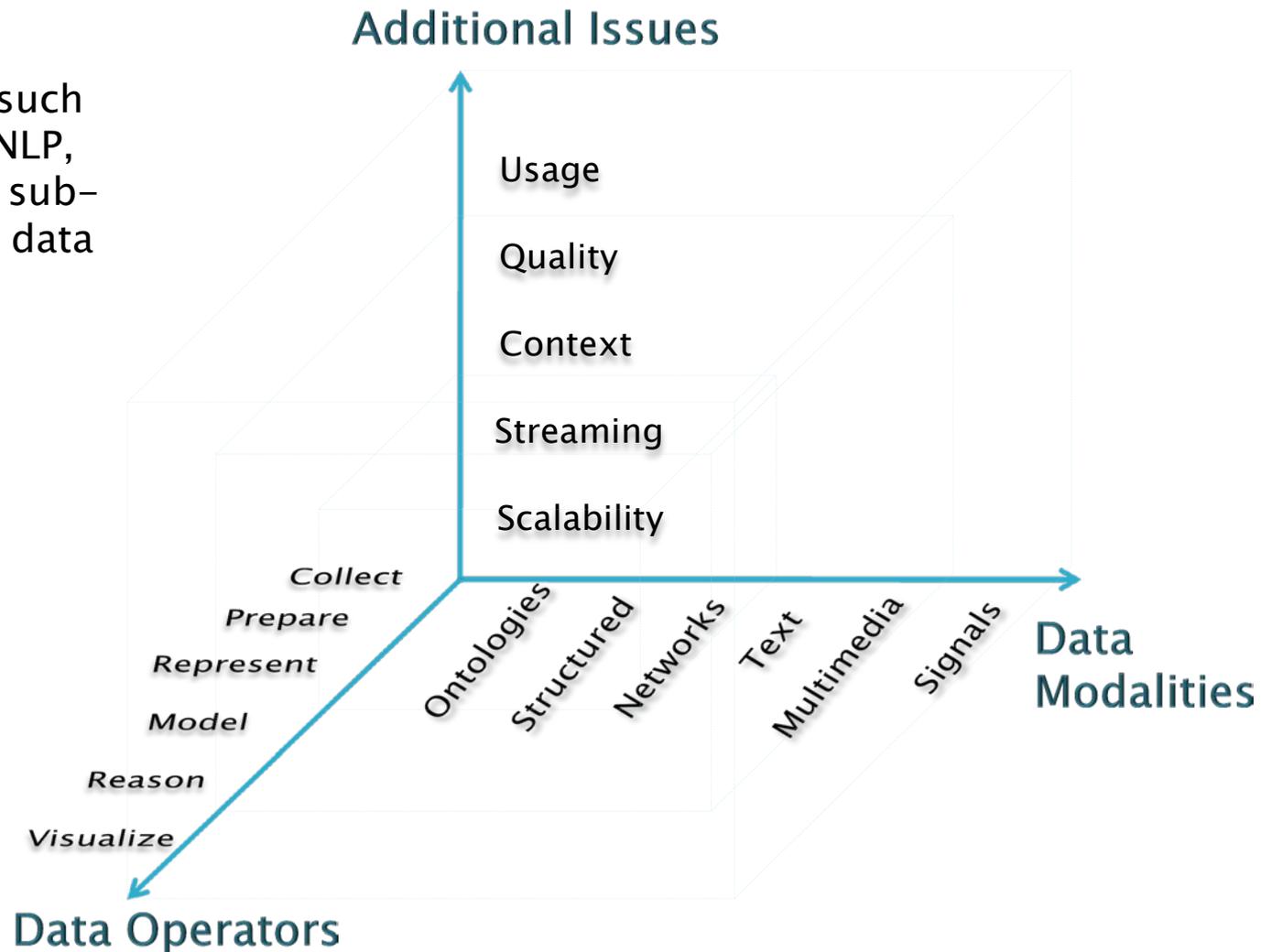
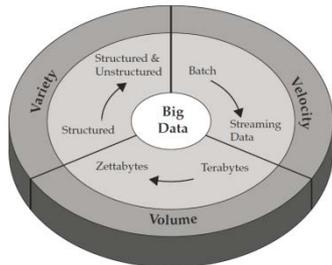
# Techniques

# 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

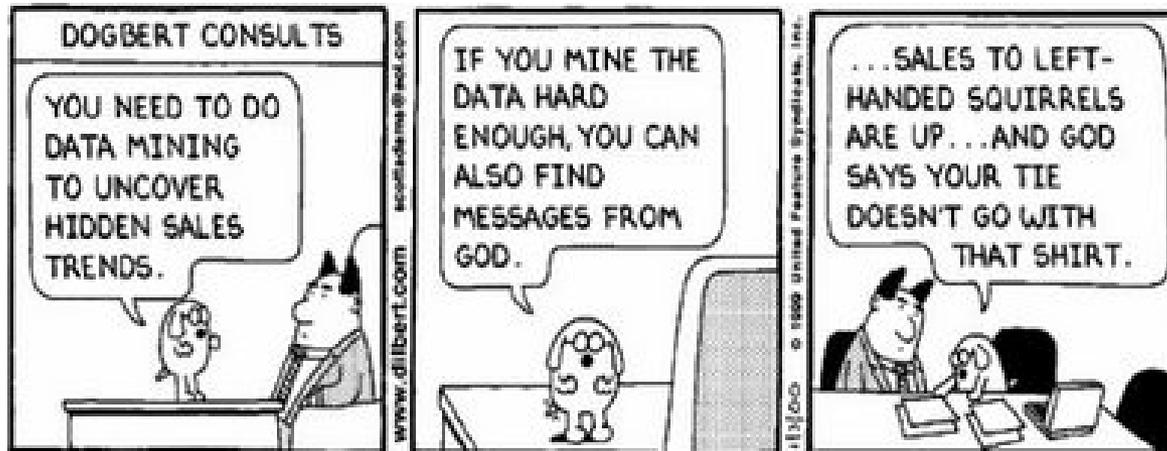
# What matters when dealing with data?

- ▶ 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

# What are specific operators used in Big-Data applications

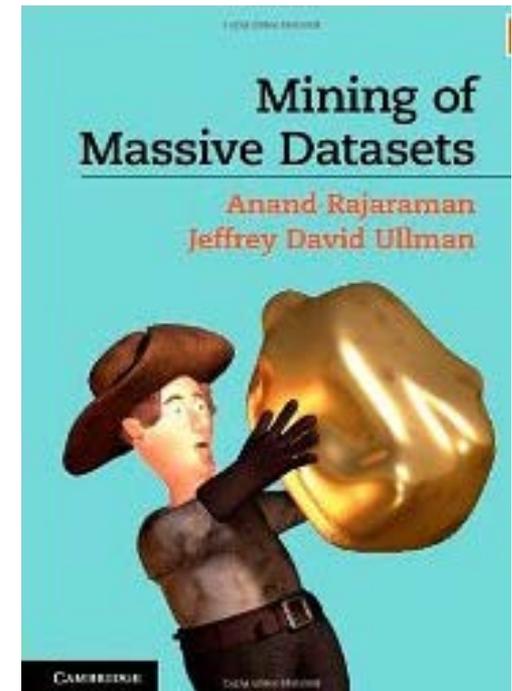
- ▶ **Smart sampling of data**
  - ...reducing the original data while not losing the statistical properties of data
- ▶ **Finding similar items**
  - ...efficient multidimensional indexing
- ▶ **Incremental updating of the models**
  - (vs. building models from scratch)
  - ...crucial for streaming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

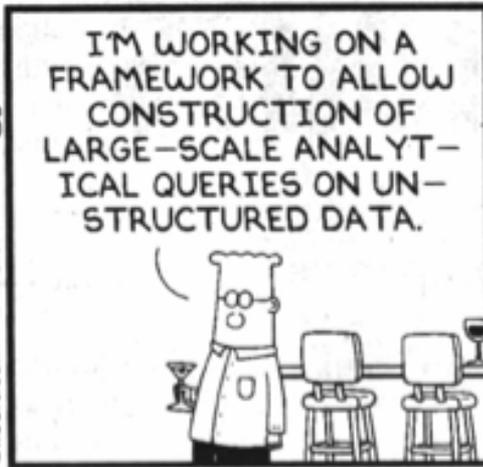


# Tools

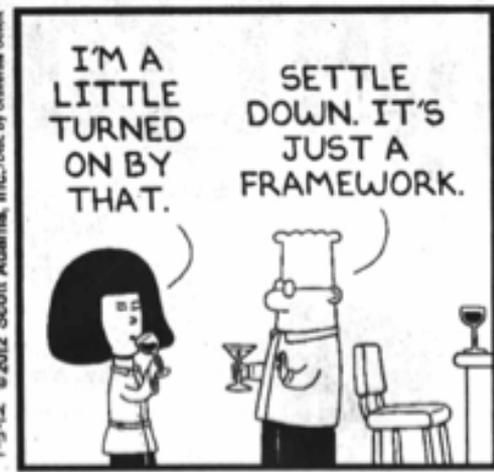
## DILBERT



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



# NoSQL Databases



- ▶ “[...] need to solve a problem that relational databases are a bad fit for”, Eric Evans
- ▶ Motives:
  - **Avoidance of Unneeded Complexity** – many use-case require only subset of functionality from RDBMSs (e.g ACID properties)
  - **High Throughput** – some NoSQL databases offer significantly higher throughput than RDBMSs
  - **Horizontal Scalability, Running on commodity hardware**
  - **Avoidance of Expensive Object-Relational Mapping** – most NoSQL store simple data structures
  - **Compromising Reliability for Better Performance**

# Open Source Big Data Tools

## 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

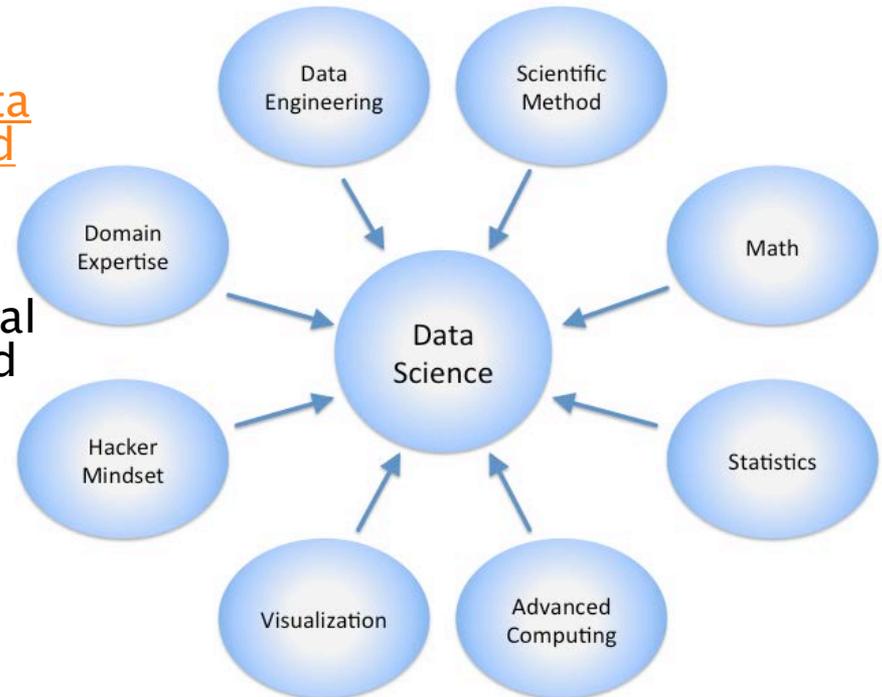
# Data Science

## Life as an Analyst



# Defining Data Science

- ▶ Interdisciplinary field using techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.
- ▶ Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common.
- ▶ Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners.



# Statistics vs. Data Science

	Statistician	Data Scientist
<b>Image</b>	Baseball (Cricket)	HBR Sexiest Job of 21 <sup>st</sup> Century
<b>Mode</b>	Reactive	Consultative
<b>Works</b>	Solo	In a team
<b>Inputs</b>	Data File, Hypothesis	A Business Problem
<b>Data</b>	Pre-prepared, clean	Distributed, messy, unstructured
<b>Data Size</b>	Kilobytes	Gigabytes
<b>Tools</b>	SAS, Mainframe	R, Python, awk, Hadoop, Linux, ...
<b>Nouns</b>	Tables	Data Visualizations
<b>Focus</b>	Inference (why)	Prediction (what)
<b>Output</b>	Report	Data App / Data Product
<b>Latency</b>	Weeks	Seconds
<b>Stars</b>	G.E.P Box Trevor Hastie	Hilary Mason Nate Silver

# Business Intelligence vs. BI

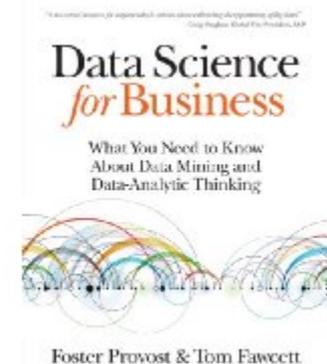
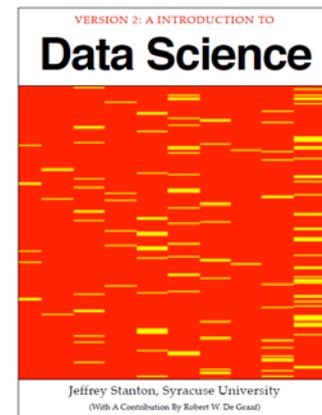
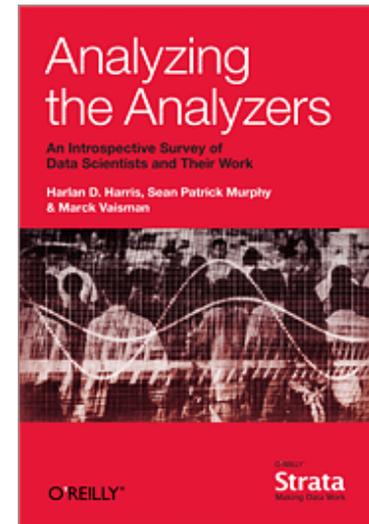
	<b>Business Intelligence</b>	<b>Data Science</b>
<b>Perspective</b>	Looking backwards	Looking forwards
<b>Actions</b>	Slice and Dice	Interact
<b>Expertise</b>	Business User	Data Scientist
<b>Data</b>	Warehoused, Siloed	Distributed, real-time
<b>Scope</b>	Unlimited	Specific business question
<b>Questions</b>	What happened?	What will happen? What if?
<b>Output</b>	Table	Answer
<b>Applicability</b>	Historic, possible confounding factors	Future, correcting for influences
<b>Tools</b>	SAP, Cognos, Microstrategy, SAS	Revolution R Enterprise QlikView, Tableau, Jaspersoft
<b>Hot or not?</b>	So 1997	Transformational

# 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

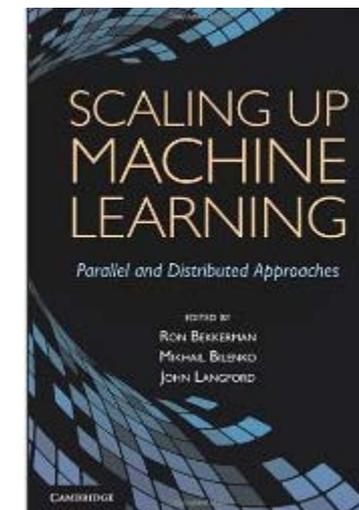
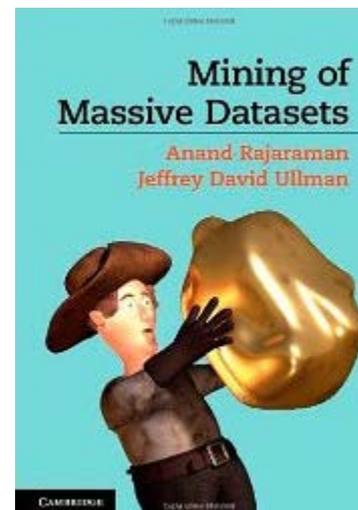
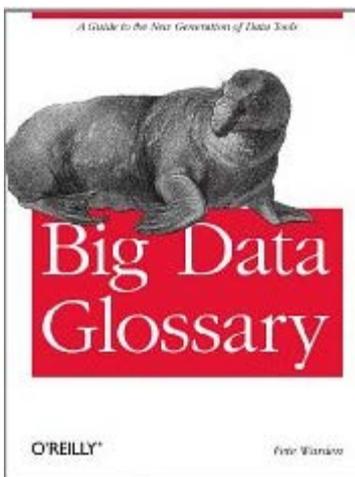
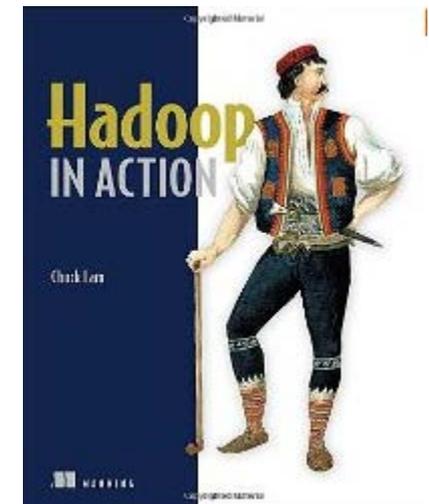
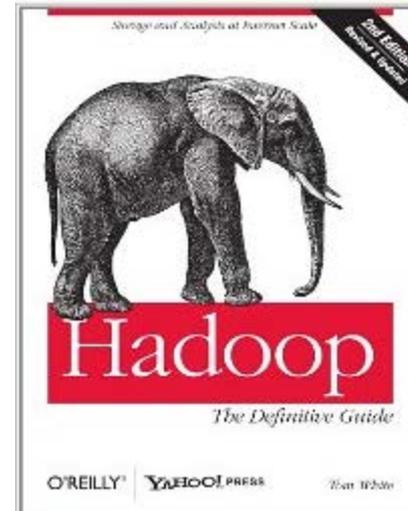
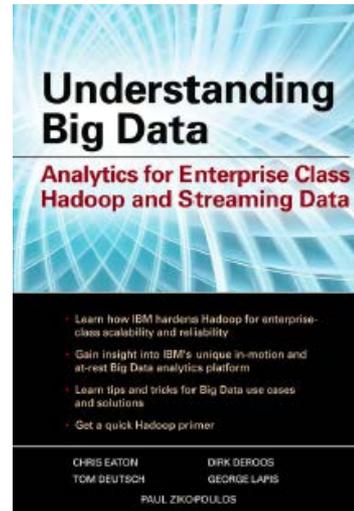
[An Introduction to Data](#)  
Jeffrey Stanton, Syracuse University School of Information Studies  
Downloadable from <http://jsresearch.net/wiki/projects/teachdatascience>  
Released: February 2013

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)