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
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 and an iPhone 4 with equal performance
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

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
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

Computation capacity has also risen sharply
Global installed computation to handle information

$5 million vs. $400
Price of the fastest supercomputer in 1975 and an iPhone 4 with equal performance

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

<table>
<thead>
<tr>
<th>Sectors</th>
<th>Stored data in the United States, 2009</th>
<th>Number of firms with &gt;1,000 employees</th>
<th>Stored data per firm (&gt;1,000 employees), 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete manufacturing(^3)</td>
<td>966</td>
<td>1,000</td>
<td>967(^2)</td>
</tr>
<tr>
<td>Government</td>
<td>848</td>
<td>647</td>
<td>1,312</td>
</tr>
<tr>
<td>Communications and media</td>
<td>715</td>
<td>399</td>
<td>1,792</td>
</tr>
<tr>
<td>Process manufacturing(^3)</td>
<td>694</td>
<td>835</td>
<td>831(^2)</td>
</tr>
<tr>
<td>Banking</td>
<td>619</td>
<td>321</td>
<td>1,931</td>
</tr>
<tr>
<td>Health care providers(^3)</td>
<td>434</td>
<td>1,172</td>
<td>370</td>
</tr>
<tr>
<td>Securities and investment services</td>
<td>429</td>
<td>111</td>
<td>3,866</td>
</tr>
<tr>
<td>Professional services</td>
<td>411</td>
<td>522</td>
<td></td>
</tr>
<tr>
<td>Retail</td>
<td>364</td>
<td>843</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>269</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td>243</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>227</td>
<td>236</td>
<td></td>
</tr>
<tr>
<td>Wholesale</td>
<td>202</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Utilities</td>
<td>194</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>Resource industries</td>
<td>116</td>
<td>708</td>
<td></td>
</tr>
<tr>
<td>Consumer &amp; recreational services</td>
<td>106</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>Construction</td>
<td>51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>Video</th>
<th>Image</th>
<th>Audio</th>
<th>Text/numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Insurance</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Securities and investment services</td>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Discrete manufacturing</td>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Process manufacturing</td>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Retail</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Wholesale</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Professional services</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Consumer and recreational services</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Health care</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Transportation</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Communications and media</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Utilities</td>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Construction</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Resource industries</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Government</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Education</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Penetration:
- **High**
- **Medium**
- **Low**

1. We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2. Video and audio are high in some subsectors.

SOURCE: McKinsey Global Institute analysis
Data available from “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):
- 2010: 5-14
- 2015: 72-215

<table>
<thead>
<tr>
<th>Sector</th>
<th>2010-2015 Growth Rate (2010-15, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>50+</td>
</tr>
<tr>
<td>Health care</td>
<td>50+</td>
</tr>
<tr>
<td>Energy</td>
<td>15</td>
</tr>
<tr>
<td>Industrials</td>
<td>5</td>
</tr>
<tr>
<td>Retail</td>
<td>30</td>
</tr>
<tr>
<td>Travel and logistics</td>
<td>15</td>
</tr>
<tr>
<td>Utilities</td>
<td>45</td>
</tr>
<tr>
<td>Automotive</td>
<td>20</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>World Population</th>
<th>Connected Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>6.3 Billion</td>
<td>500 Million</td>
</tr>
<tr>
<td>2010</td>
<td>6.8 Billion</td>
<td>12.5 Billion</td>
</tr>
<tr>
<td>2015</td>
<td>7.2 Billion</td>
<td>25 Billion</td>
</tr>
<tr>
<td>2020</td>
<td>7.6 Billion</td>
<td>50 Billion</td>
</tr>
</tbody>
</table>

Source: Cisco IBSG, April 2011
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

- Computer and electronic products
- Information
- Professional services
- Health care providers
- Wholesale trade
- Manufacturing
- Transportation and warehousing
- Retail trade
- Real estate and rental
- Accommodation and food
- Natural resources
- Utilities
- Other services
- Construction

Big data value potential index

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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Employment</th>
<th>Graduates with deep analytical talent</th>
<th>Others ¹</th>
<th>2018 supply</th>
<th>Talent gap</th>
<th>2018 projected demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>150</td>
<td>180</td>
<td>30</td>
<td>300</td>
<td>140–190</td>
<td>440–490</td>
</tr>
</tbody>
</table>

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

Big Data Revenue by Type, 2012

(in $US millions)

Hardware
$4,553
40%

Software
$2,400
21%

Services
$4,444
39%

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

Big Data Market Forecast by Component, 2011-2017
($US billions)

Yearly Revenue ($US billions)

<table>
<thead>
<tr>
<th>Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data XaaS Revenue</td>
<td>$0.35</td>
<td>$0.61</td>
<td>$1.05</td>
<td>$1.74</td>
<td>$2.47</td>
<td>$2.91</td>
<td>$3.24</td>
</tr>
<tr>
<td>Big Data Application (Analytic and Transactional) Software</td>
<td>$0.49</td>
<td>$0.94</td>
<td>$1.80</td>
<td>$3.29</td>
<td>$5.02</td>
<td>$6.15</td>
<td>$7.00</td>
</tr>
<tr>
<td>Big Data NoSQL Database Software</td>
<td>$0.10</td>
<td>$0.19</td>
<td>$0.39</td>
<td>$0.73</td>
<td>$1.14</td>
<td>$1.41</td>
<td>$1.62</td>
</tr>
<tr>
<td>Big Data SQL Database Software</td>
<td>$0.72</td>
<td>$1.02</td>
<td>$1.45</td>
<td>$1.99</td>
<td>$2.47</td>
<td>$2.73</td>
<td>$2.90</td>
</tr>
<tr>
<td>Big Data Infrastructure Software</td>
<td>$0.16</td>
<td>$0.26</td>
<td>$0.43</td>
<td>$0.70</td>
<td>$0.96</td>
<td>$1.22</td>
<td>$1.24</td>
</tr>
<tr>
<td>Big Data Networking Revenue</td>
<td>$0.18</td>
<td>$0.28</td>
<td>$0.44</td>
<td>$0.67</td>
<td>$0.89</td>
<td>$1.02</td>
<td>$1.11</td>
</tr>
<tr>
<td>Big Data Storage Revenue</td>
<td>$1.16</td>
<td>$1.83</td>
<td>$2.89</td>
<td>$4.40</td>
<td>$5.86</td>
<td>$6.70</td>
<td>$7.28</td>
</tr>
<tr>
<td>Big Data Compute Revenue</td>
<td>$1.64</td>
<td>$2.45</td>
<td>$3.64</td>
<td>$5.23</td>
<td>$6.70</td>
<td>$7.50</td>
<td>$8.06</td>
</tr>
<tr>
<td>Total Big Data Revenue</td>
<td>$7.2</td>
<td>$11.4</td>
<td>$18.2</td>
<td>$28.0</td>
<td>$37.9</td>
<td>$43.7</td>
<td>$47.8</td>
</tr>
</tbody>
</table>
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.
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
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
SO, WHAT DO YOU DO FOR A LIVING?

I'M WORKING ON A FRAMEWORK TO ALLOW CONSTRUCTION OF LARGE-SCALE ANALYTICAL QUERIES ON UNSTRUCTURED DATA.

I'M A LITTLE TURNED ON BY THAT.

SETTLE DOWN. IT'S JUST A FRAMEWORK.
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 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**

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
Life as an Analyst

WALLY, DID YOU FINISH THE ANALYSIS FOR TOMORROW?

NO.

I'M WAITING UNTIL THE LAST MINUTE SO YOU WON'T HAVE TIME TO ASK FOR UNNECESSARY CHANGES.

I'M A STEP AHEAD OF HIM — THE ANALYSIS ITSELF IS UNNECESSARY.
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

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

## Business Intelligence vs. BI

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

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)