Understanding and Managing Data Science and Big Data

Erik Perjons

DSV, Stockholm University

Understanding Data Science

- A way to understand what data science is by comparing it to business intelligence

Business Intelligence – a definition

Business intelligence (BI) is an umbrella term that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help users make better decisions.

Wixom and Watson, 2010

Data Science – a definition

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

Wikipedia

Business Intelligence vs Data Science

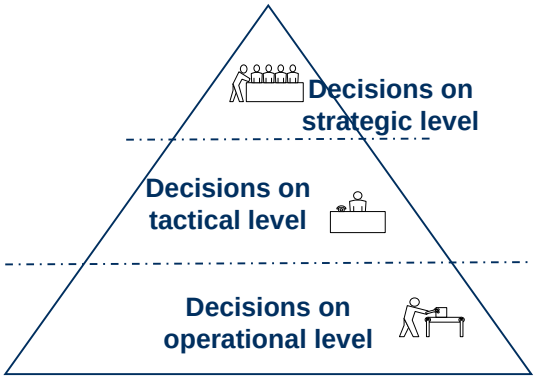
Business intelligence (BI) is an umbrella term that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help users make better decisions.

Wixom and Watson, 2010

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

Wikipedia

Business intelligence – an overview



Data sources

- Business systems

BI technologies

- Dashboard systems
- Analytical decision support tools
- Expert systems
- Data warehouse systems
- OLAP tools
- Data mining tools
- Reporting tools

BI related methods

- Performance management
- Six sigma
- Balanced scorecard
- Lean
- Activity based costing
- Data and process mining
- Decision and risk analysis

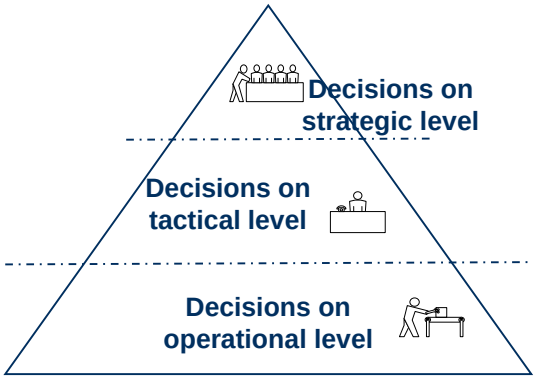
System integration tools

- ETL tools

Visualization

- Visualization techniques

Data science – an overview



Data sources

- Business systems

Additional data sources

- Social media
- Internet of Things
- Web
- Video/Camera
- Machines/Devices
- Maps
- Satellites

Data science technologies

- Hadoop
- MapReduce YARN/SPARK
- SAS/R

Data science related methods

- Data mining
- Machine learning
- Language technology
- Process mining
- Web mining

- CRISP-DM
- Knowledge Discovery in Databases (KDD)

Visualization

- Visualization techniques

Data science

Data science (Wikipedia) - is an **inter-disciplinary field** that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

Data science (Lewis, 2004) - is about **finding new variables** and metrics **that are better predictors** of performance

Data science vs. Big data

Big data (Wikipedia) - is a **field** that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Big data (Gartner) - is high-volume, and high-velocity or high-variety **information assets** that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Big data (Schmarzo, 2014) - is a **key enabler** of a new discipline called **data science** that seeks to leverage new sources of structured and unstructured data, coupled with predictive and prescriptive analytics, to uncover new variables and metrics that are better predictors of performance

… vs. Big Data analytics

Big data analytics (https://searchbusinessanalytics.techtarget.com/) - is the often complex **process of examining big data** to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Data science, Big data, v.s Big data analytics

Data science (Wikipedia) - is an **inter-disciplinary field** that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

Big data (Gartner) - is high-volume, and high-velocity or high-variety **information assets** that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Big data analytics (https://searchbusinessanalytics.techtarget.com/) - is the often complex **process of examining big data to uncover information --** such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions**.**

Business intelligence vs. Data science

The differences between BI and Data science is discussed in a chapter by Schmarzo (2016):

- The questions are different
- The analyst characteristics are different
- The analytic approaches are different
- The data models are different
- The views on business are different

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The questions are different

Business Intelligence

- ***Focus on descriptive analytics:*** ”What happend?” type of questions: How many units of products X did we sell in Jan 2017

Data Science

- ***Focus on predictive analytics:*** ”What is likely to happend? type of questions: How many units of products X will we sell in Jan 2018?
- ***Focus on prescriptive analytics:*** ”What should we do?” type of question: How many components A, B, C should I order to support the sales of product X?

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The questions are different

Busines

- ***Focus*** How many units c

Data Sc

- ***Focus*** estions: How many

- ***Focus*** on: How many components A, B, C should I order to support the sales of product X?

The major **difference between predictive and prescriptive** is that the former forecasts potential future outcomes, while the latter helps you draw up specific recommendations (https://www.businessnewsdaily.com/)

To answer the predictive and prescriptive questions, the data scientist build analytic models in order to quantify cause and effect relationships

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The questions are different

Busines

- ***Focus*** **How many units c**

Data Sc

- ***Focus*** **estions: How many**
- ***Focus*** **on: How many components A, B, C should I order to support the sales of product X?**

To answer the predictive and prescriptive questions, the data scientist build analytic models in order to quantify cause and effect relationships

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analysts' characteristics are different

The attitude and work approach among BI analysts and data scientists differs:

AREA	BI ANALYST	DATA SCIENTIST
Focus	Trends, KPIs	Pattern, Correlations, Models
Process	Static	Exploratory, experimentation, visual, agile
Data sources	Pre-planned, added slowly	On the fly, as needed
Transform data	Carefully planned	On demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analysis	Descriptive	Predictive, prescriptive

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analytic approaches are different 1(2)

Stockholms universitet

BI analytic approach

Step 1: Pre-build a data model (Schema on load). This is done by gather requirements from business users, by asking ”What do you want to know?”

Step 2: Define the report/queries. This is done by using a BI tool, in which attributes/dimensions/facts, etc, are selected for creating SQL queries

Step 3: Generate SQL queries. This is done automatically by using the BI tool

Step 4: Create the report/dashbord widget. The BI tool is used to apply SQL queries against the data in the DW and visualize the result graphically

Data science analytic approach

Step 1: Define hypothesis

Step 2: Gather and test data from multiple data sources using an analytic sandbox

Step 3: Build data model when knowing what data sources to use (Schema on query)

Step 4: Use visualization tool to identify intresting correlations and outliers to test

Step 5: Build analytic models using different analytic techniques/algorithms

Step 6: Evaluate the models goodness of fit

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analytic approaches are different 1(2)

Stockholms universitet

BI analytic approach

Step 1: Pre-build a data model (Schema on load). This is done by gather requirements from business users, by asking ”What do you want to know?”

Step 2: Define the report/queries. This is done by using a BI tool, in which attributes/dimensions/facts, etc, are selected for creating SQL queries

Step 3: Generate SQL queries. This is done automatically by using the BI tool

Step 4: Create the report/dashbord widget. The BI tool is used to apply SQL queries against the data in the DW and visualize the result graphically

Data science analytic approach

Step 1: Define hypothesis

Step 2: Gather and test data from multiple data sources using an analytic sandbox

Step 3: Build data model when knowing what data sources to use (Schema on query)

Step 4: Use visualization tool to identify intresting correlations and outliers to test

Step 5: Build analytic models using different analytic techniques/algorithms

Step 6: Evaluate the models goodness of fit

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analytic approaches are different 1(2)

Stockholms universitet

BI analytic approach

Step 1: Pre-
done by gat
asking "Wh

Step 2: Def
a BI tool, i
etc, are se

Step 3: Ge
automatic

Step 4: Cre
is used to apply SQL queries against the data in the DW and visualize the result graphically

What is the benefit with schema on the load?

What would happend if you want to add new data into the data warehouse/BI environment?

Data science analytic approach

ultiple
box for
nesis

wing what
ery)

ntify
to test

different analytic techniques/algorithms

Step 6: Evaluate the models goodness of fit

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analytic approaches are different 2(2)

Schema on load

– a schema must be built prior to loading data into the data warehouse

Schema on query/read

– a schema is defined based on data that support the hypothesis

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The data models are different

Business Intelligence

- Schema on load
- Often star join schemas – multiples tables, many (comparable slow) joins

Data Science

- Schema on query/read
- Often flattened tables – few (flattened) tables with a lot of data, few joins

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The views on business are different

Business Intelligence

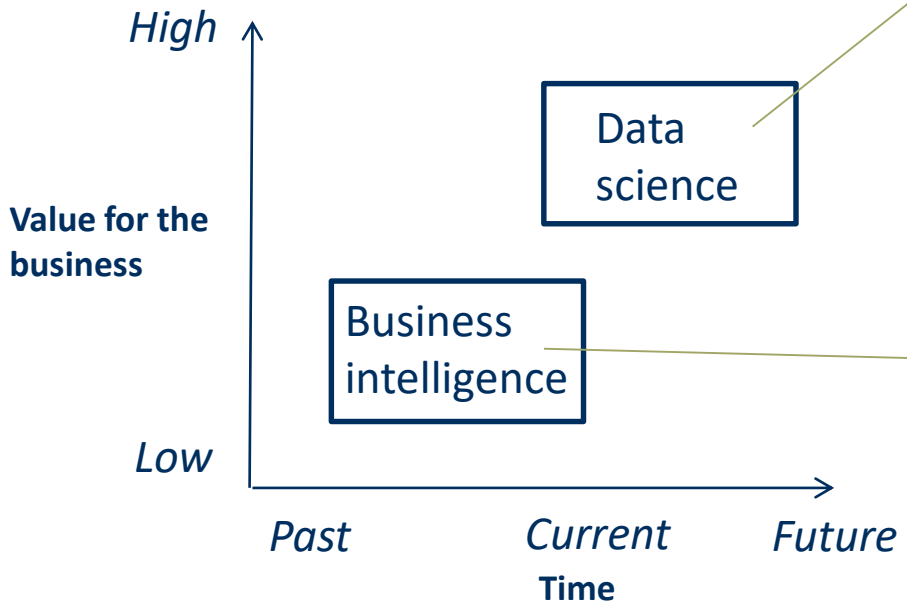
- Aggregated data on business entities, such as customers, products

Data Science

- Build **analytic profiles** on each business entity. Example of business entities are customers, partners/suppliers, devices, machines
- For exampel, analytic profiles for customers could be used for managing customer rentension rate

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Business intelligence vs. Data science



- Predictive analytics
- Prescriptive analytics
- On the fly use of large amount of data
- Exploratory, experimental process
- Focus on pattern, correlations
- Data model – schema on query/read
- Data quality – good enough, propabilities

- Desciptive analytics
- Pre-planned use of data (via ETL)
- Static and pre-planned process
- Focus on trends, use of KPIs
- Data model – schema on load
- Data quality – high quality, single source of truth

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Big Data: The Management Revolution

- McAfee & Brynjolfsson (2012) claim that **big data will have a large impact on the organisations' performance**

- **Why? Well, big data enable managers to measure more about their business then they did before, and therefore know radically more about it. This knowledge will improve decision making and performance**

- **Is that true? How do we know that big data and data science will improve performance?**

McAfee&Brynjolfsson (2012) Big Data: The Management Revolution

How do we know that big data will improve performance?

- An **investigation was carried** out by MIT, McKinsey and a researcher from Wharton
- Methods: **Interviews with managers. Analysis of performance data (financial and operational results, 2005-2009). 330 North American companies**
- **Result:** The more **companies characterized themselves as data-driven, the better they performed** on objective measures of financial and operational results
- **Result:** Organizations – that are in the top third of their industry – and that are **using evidence based decision making** - were, **on average, 5 % more productive and 6 % more profitable**

McAfee&Brynjolfsson (2012) Big Data: The Management Revolution

Management Challenges

The management challenges that need to be addressed:

- First of all, decision need to be made based on data, evidence (and not using the HIPPO style)
- Understand how to combine domain expertise (who understand what questions to ask) with data science expertise
- Information and decision rights need to be put in the same location – ”people who understand a problem needs to be brought together with right data”
- The skills of the data scientist are crucial
- Privacy concerns need to be addressed
-

McAfee&Brynjolfsson (2012) Big Data: The Management Revolution