



-1- Introduction to Big Data Analytics



Lecture 1:

Introduction to Big Data Analytics

Upon completion of this module, you should be able to:

- Define big data: What is it? its Characteristics.
- Identify four business drivers for advanced analytics
- Distinguish the techniques for Business Intelligence from Data Science
- Describe the role of the Data Scientist within the new big data ecosystem

BI Vs DS

point / present - future
 Descriptive - Exploratory
 Decision - Strategic planning
 Dashboard, report - Statistical model predictive
 hypothesis
 - Structured & unstructured
 → Structured


Big Data

What is it?

What makes Data, "Big Data"

Big Data Defined

- ***“Big Data” is data where scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.***

- 
- Requires new data architectures
 - New tools
 - New analytical methods
 - Integrating multiple skills into new role of data scientist

- ***Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities***

Big Data Defined: Characteristics or V's

Big Data is sometimes described as having 3 characteristics or Vs: **V**olume, **V**ariety, and **V**elocity.

● Main characteristics of big data: (Named V's)

- ✓ **Huge volume of data (Volume):** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
- ✓ **Complexity of data types and structures (Variety):** Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis. With an increasing volume of unstructured data (*80-90% of the data in existence is unstructured*)
- ✓ **Speed of new data creation and growth (Velocity):** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Key Characteristics of Big Data

1. Data Volume

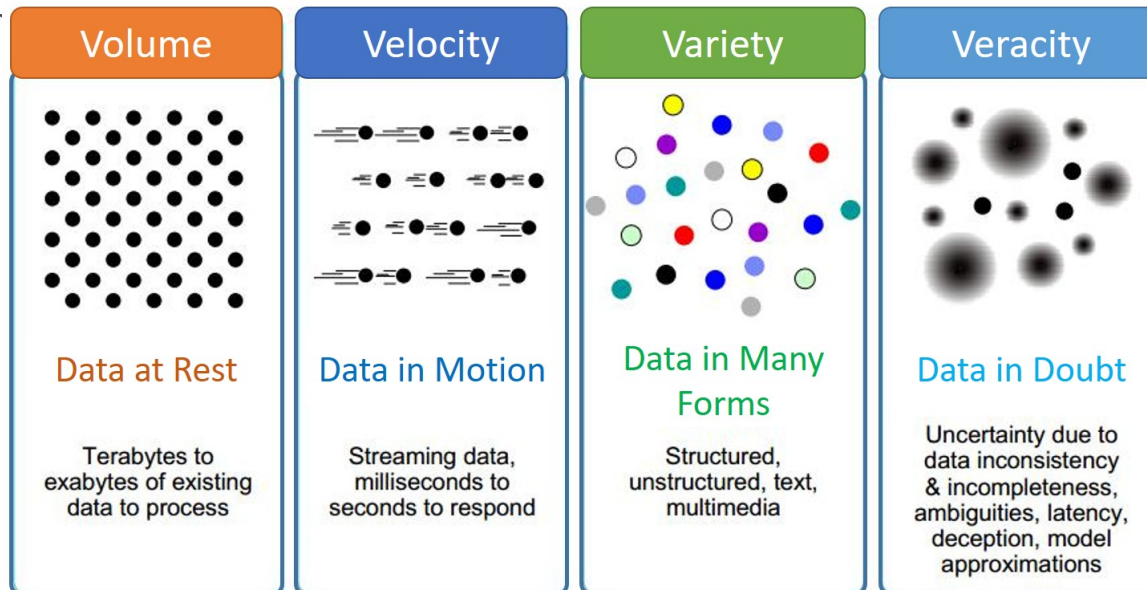
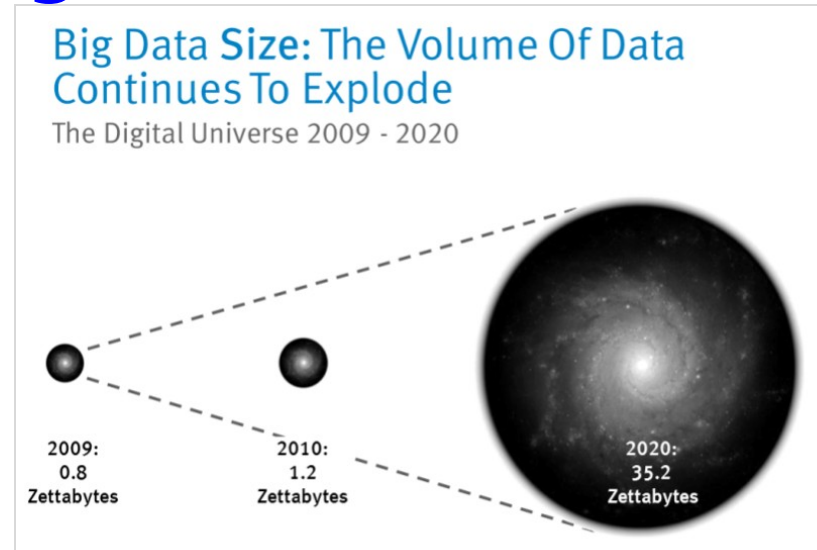
- 44x increase from 2009 to 2020
(0.8 zettabytes to 35.2zb)

$$1 \text{ ZB} = 1000^7 \text{ bytes} = 10^{21} \text{ byte}$$

2. Speed or velocity of new data creation

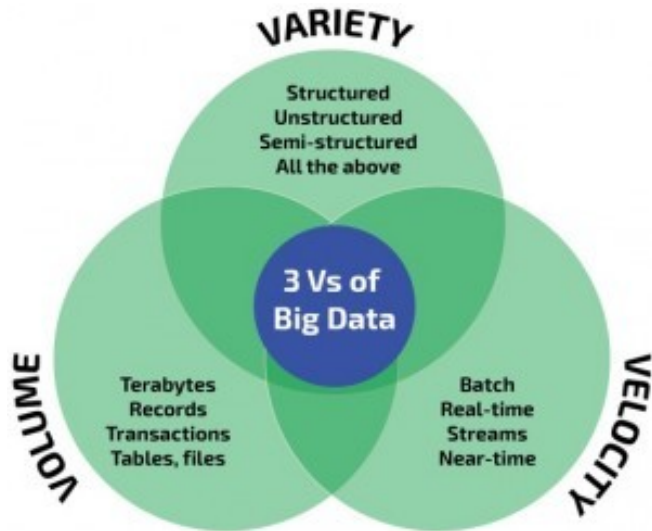
3. Data Structure

- Greater variety of data structures to mine and analyze



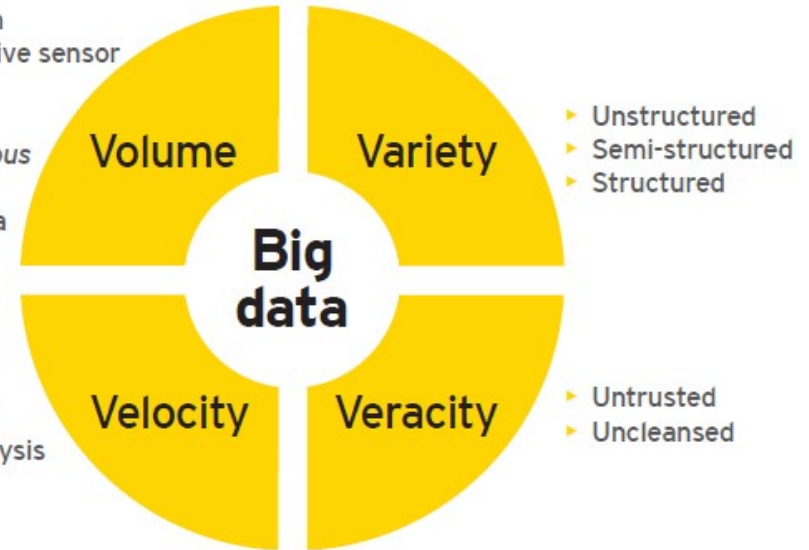
Characteristics of Big Data

Big Data Characteristics: V's

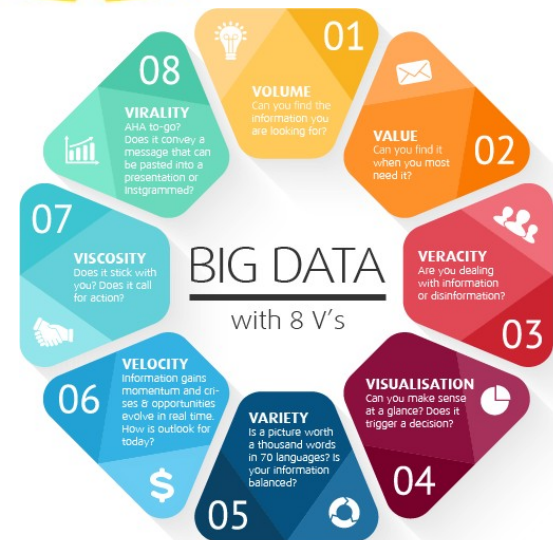


- ▶ Click stream
- ▶ Active/passive sensor
- ▶ Log
- ▶ Event
- ▶ Printed corpus
- ▶ Speech
- ▶ Social media
- ▶ Traditional

- ▶ Speed of generation
- ▶ Rate of analysis

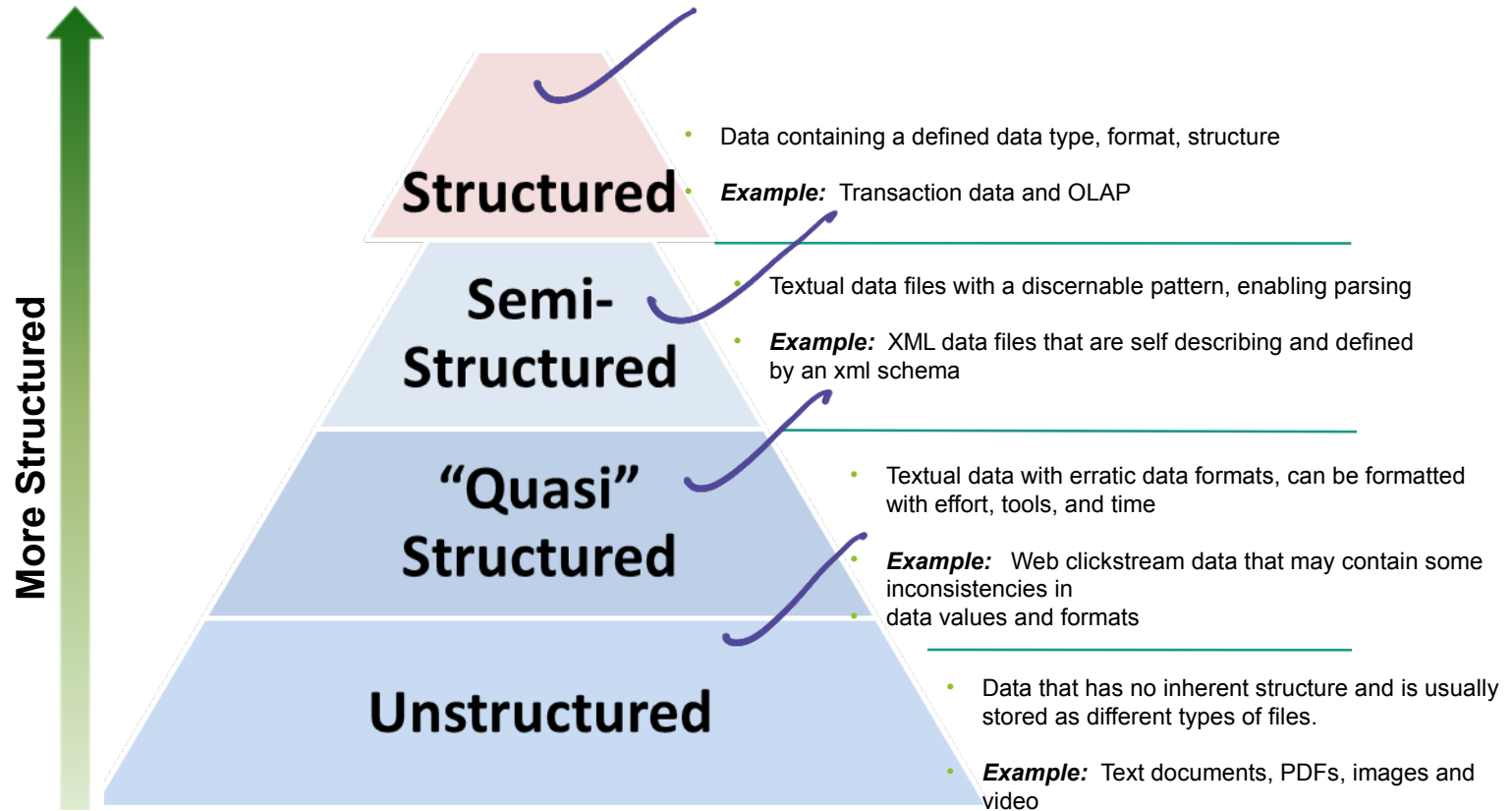


VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.



Big Data Characteristics: Data Structures

Data Growth is Increasingly Unstructured



Four Main Types of Data Structures

Complex & Varied of Data Structures

Structured Data

SUMMER FOOD SERVICE PROGRAM 1)				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2)
	-----Thousands-----	---Mil.---		---Million \$---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3)	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1

Semi-Structured Data



View →
Source

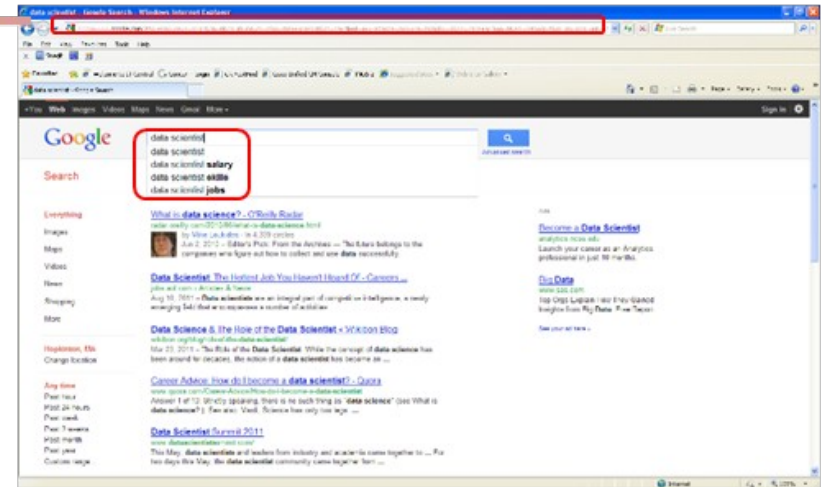


```

1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-trans:
2 <html xmlns="http://www.w3.org/1999/xhtml">
3
4 <head>
5 <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
6 <META name="y_key" content="859b4020c1c9a0cc">
7 <link rel="canonical" href="http://www.emc.com/index.htm" />
8 <META NAME="verify-v1" CONTENT="y1t9V0P4eV0j9dIeV1eRfP32g4qswF0120VtM50
9 <title>EMC - Data Recovery, Cloud Computing, and Storage Hardware</title>
10 <META NAME="description" CONTENT="EMC is a leading provider of storage hardware solutions th
11 data recovery and improve cloud computing." />
12 <META NAME="keywords" CONTENT="emc,network storage,data recovery,information manager
13 software,sas storage,information protection,information management" />
14 <!-- Start stylesheet includes -->
15 <link rel="stylesheet" href="/_admin/css/styles.css" />
16 <link rel="stylesheet" href="/_admin/css/styles_nav.css" />
17 </-->

```

Quasi-Structured Data



http://www.google.com/#hl=en&sugexp=kjmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&scient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aql=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or,r_gc_r_pw,cf.osb&fp=d566e0fdb09c8604&biw=1382&bih=651

Unstructured Data

The Data Wheelbarrow, by William Carlos Williams





From Data Analytics to Big Data Analytics

Business Drivers for Analytics

Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

Driver		Examples
1	Desire to optimize business operations	Sales, pricing, profitability, efficiency
2	Desire to identify business risk	Customer churn, fraud,
3	Predict new business opportunities	Attempt to make a more profitable sale, best new customer prospects

Analytics

- Decision makers may choose to make decisions based on past experiences or rules of thumb, but unless data is considered, it would not be an analytical decision-making process.

Data Analytics

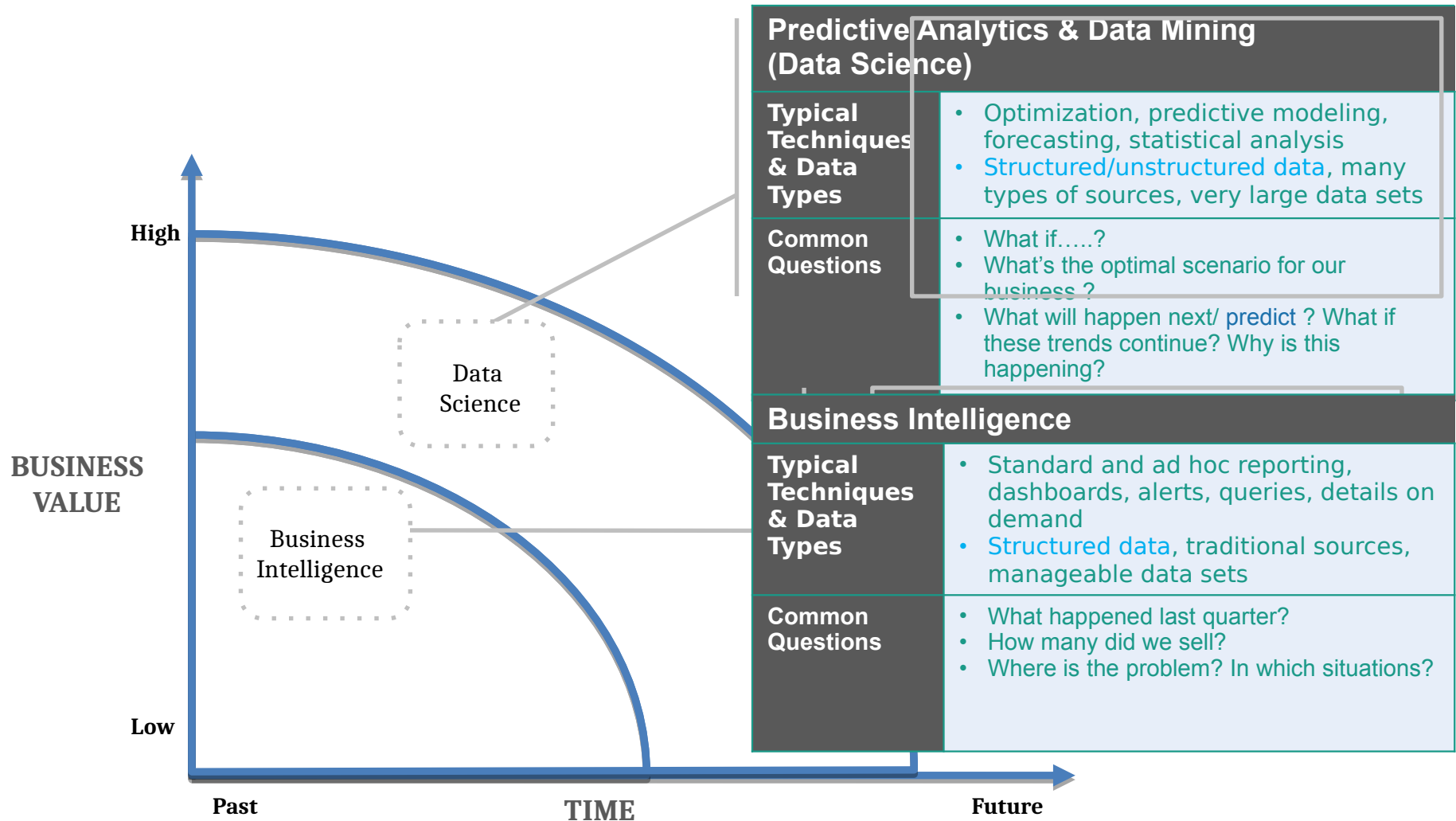
- Suppose your street ice cream vendor stopped servicing your street but still serving the next street. Then one day you asked him/her why?
- He / She tells you that your street customers continually bargain and hence he loses a lot of money and time, but on the street next to yours he has some great customers for whom he provides excellent service in short time.
- Your ice cream vendor TESTED servicing your street and within one month he/she DECIDED to stop servicing your street, and even if you ask him/her, he/she will not show up.
- Because he/she analyzed the figures of his/her expenses against the income in your street and realized that he/she is losing money and time. The vendors used some kind of data analytics and decided not to servicing your street.
- Can you tell us a real story in which you took a decision based on some kind of data analytics?

Business Intelligence Versus Data Science

- There are two types of data analytics:
 - **Descriptive:** its purpose is to summarize/describe what happened,
 - **Predictive:** its purpose is to forecast or predict what might happen in the future
- **Business Intelligence Technology (BI)** is very useful for **descriptive analytics**
 - Useful for closed-ended and explanation of current or past behavior, typically by aggregating historical data and grouping it in some way.
 - Provides hindsight and some insight and generally answers questions related to “when” and “where” events occurred through reports, dashboards, and queries on business questions for the current period or in the past.
- **Data Science (DS)** combines statistics, mathematics and computing concepts, methodologies and tools to undertake **predictive analytics** on big data
 - to see patterns,
 - to discover relationships, and
 - to make sense of stunningly varied images and information.
- Data Science developed **for undertaking Analytics on Big Data** as Physics, Chemistry and Biology sciences were developed to study physical environment, chemical elements, and living things.

Business Intelligence Versus Data Science

Analytical Approaches for Meeting Business Drivers



Example of BI Queries

- Find the courses whose grades increased by 5% compared to the two last year,
- Identify the schools from which the best students came from in the last three years,
- Find out three of the most frequent reasons for which students left university compared to the last two years,
- Find cities whose purchases grew by more than 20% during the specified 3-month period, versus the same 3-month period last year,
- Find the shares in sales for the same period a year ago and then calculate the change in share between the two years.

Example of DS Queries

- How much increase in students enrolment for next year,
- How many faculty should we recruit in the next three years,
- How to reduce to 4% the rate of students leaving the university in the next two years,
- How to increase students enrollment in a particular course/program?
- What happens if we change the lecture starting time to 7am?
- What happens if we reduce car parking slots in the next two years?
- How to attract customers to buy our products?

Big Data Analytics

- Today, businesses of all sizes use data analytics to answer complex queries and take decisions.
- If the ice cream vendor can answer why he stopped serving your street, How many Big businesses with thousands/millions of customers today could answer questions like:
 - Who their MOST PROFITABLE CUSTOMERS are?
 - Do they know who their MOST COST GENERATING customers are?
 - How should they target their efforts to ACQUIRE the MOST PROFITABLE customers?
- These questions are very difficult to answer when data is growing exponentially in today's internet, social networks, sensors etc. to become what is

Considerations for Big Data Analytics

Criteria for Big Data Projects

Criteria for Big Data Projects

1. Speed of decision making
2. Throughput
3. Analysis flexibility

Video on the Use of Big Data Analytics

- **How Netflix uses Data Analytics to Launch a new TV Series?**





The Data in Big Data?

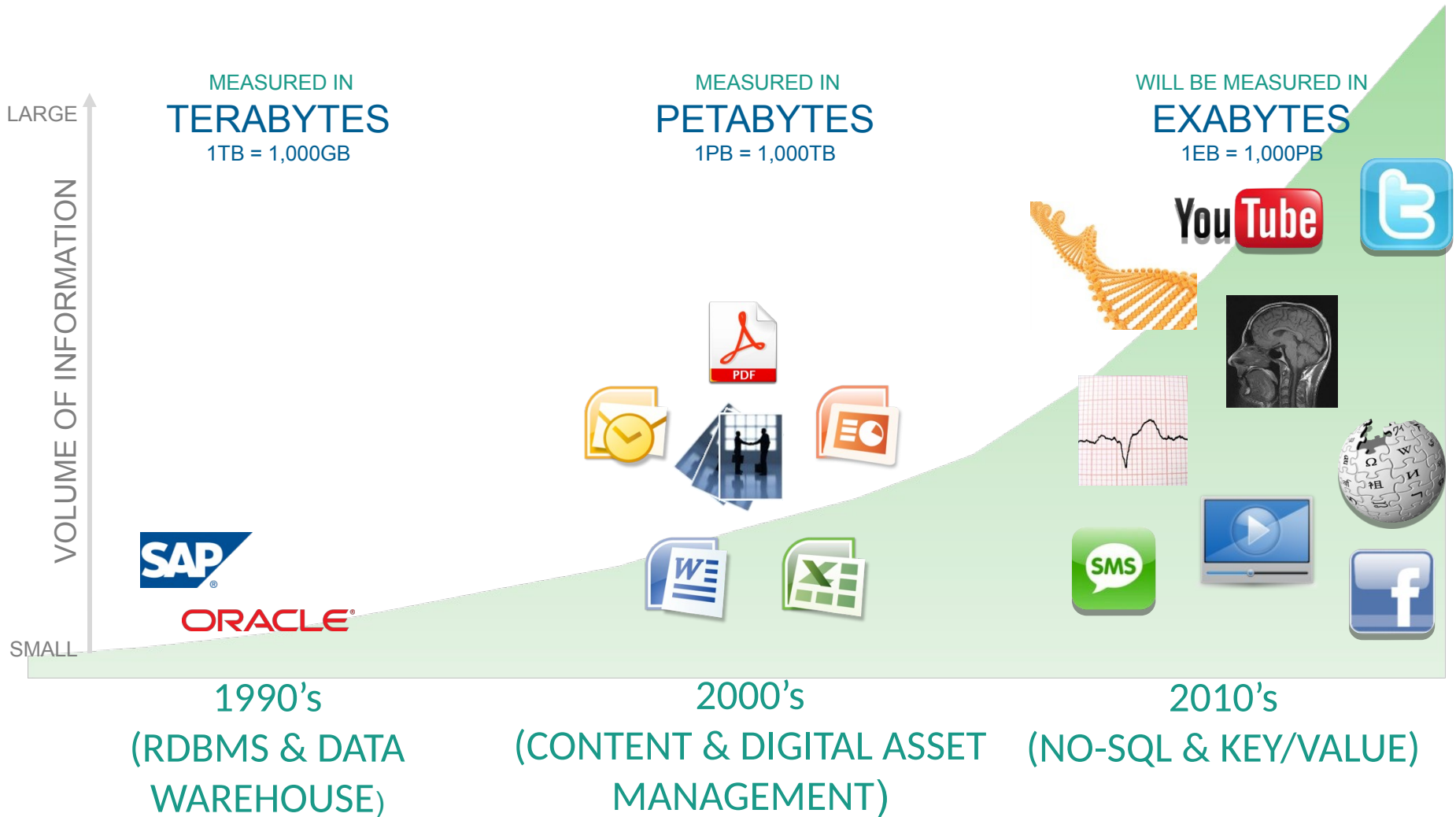
The Big Data?

- Very useful Data is created constantly, and at an ever-increasing rate through the internet and different electronic devices:
 - Mobile phones, social media e.g. Facebook, twitter, etc.
 - Imaging technologies to determine a medical diagnosis,
 - Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.
- This huge data streams of records, documents, messages, images and videos



Opportunities for a New Approach to Analytics

New Applications/Tools Driving Data Volume

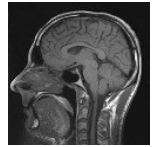


Opportunities for a New Approach to Analytics

New Applications/Tools Driving Data Volume

These data come from multiple sources, including:

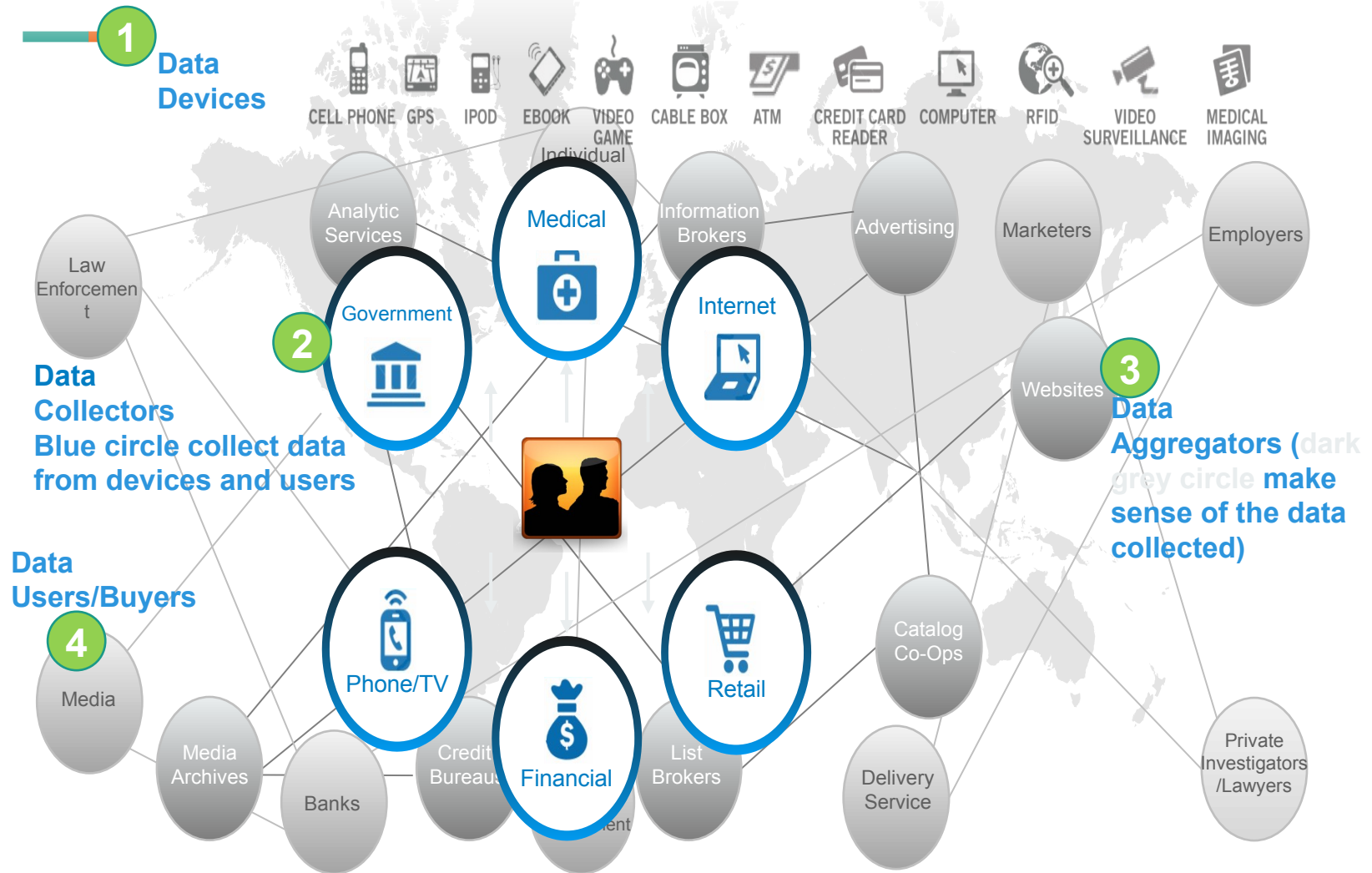
- **Medical Information**, such as genomic sequencing and MRIs
- Increased use of broadband on the Web – including the 2 billion photos each month that Facebook users currently upload as well as the innumerable videos uploaded to YouTube and other multimedia sites
- **Video surveillance (airport)**
- Increased global use of **mobile devices** – the torrent of texting is not likely to cease
- **Smart devices** – sensor-based collection of information from smart electric grids, **smart buildings** and many other public and industry infrastructure
- Non-traditional IT devices – including the use of RFID readers, GPS navigation systems



The Big Data trend is generating an enormous amount of information that requires **advanced analytics** and new market players to take advantage of it.

Opportunities for a New Approach to Analytics

New Applications for Big Data Ecosystem



Data form Devices... 1

- Data devices and the “Sensornet” gather data from multiple locations and continuously generate new data about this data.
- For each gigabyte of new data created, an additional petabyte of data is created about that data.
- Consider someone playing an online video game through a PC, game console, or smartphone.
 - In this case, the video game provider captures data about the skill and levels attained by the player fine-tune the difficulty of the game,
 - suggest other related games that would most likely interest the user, and
 - offer additional equipment and enhancements for the character based on the user’s age, gender, and interests

Data From Devices...2


- Smartphones provide another rich source of data.
 - In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location.
 - This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars,
 - The relative traffic congestion on busy roads.
 - GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.
- Retail shopping loyalty cards record not just the amount an individual spends,
 - but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often,
 - the combinations of products purchased together.



Data Scientist Profile: Skills Needed In the New Data Ecosystem

- What new **skill sets** do you need to take advantage of the big data?
- Do most large organizations have people with these **skill sets**?
- If so, **who are they**?

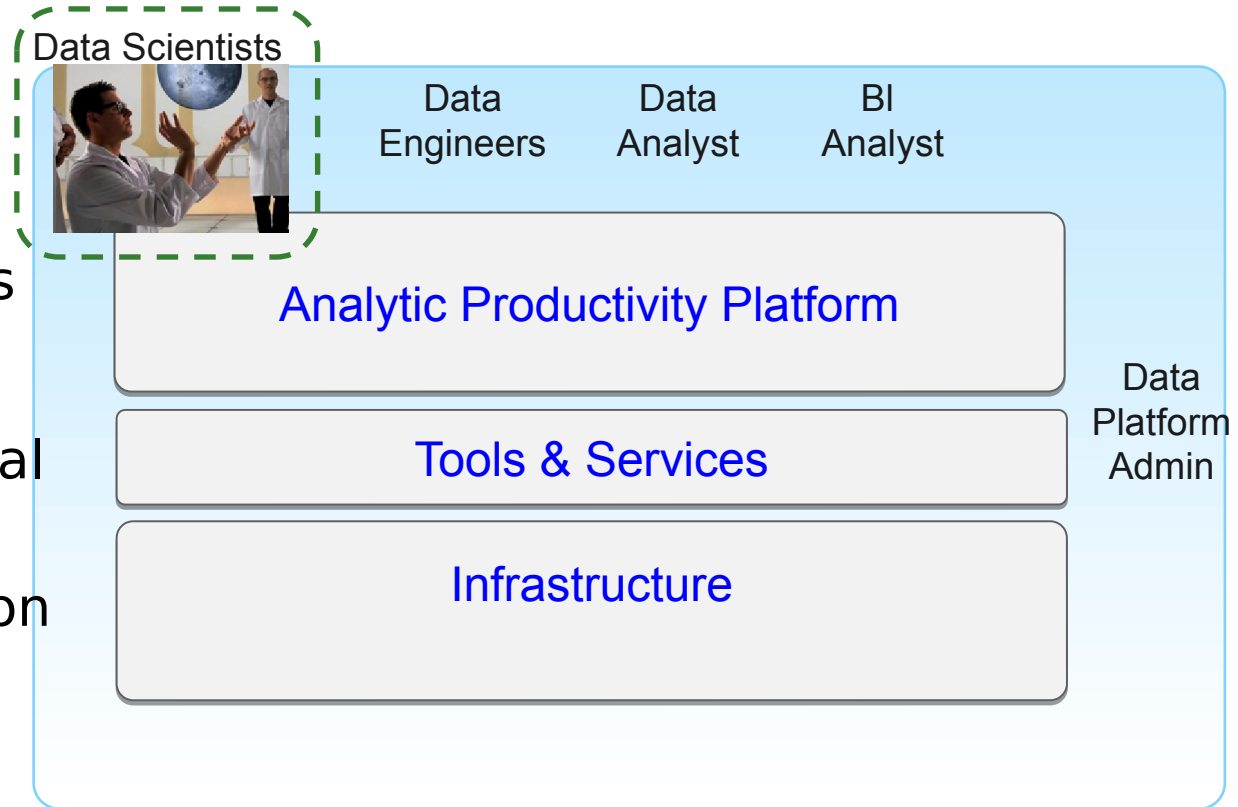
Three Key Roles of the New Data Ecosystem

		Role	Role Description
Data Scientists <i>Projected U.S. talent gap: 140,000 to 190,000</i>		Deep Analytical Talent	People with advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.
		Data Savvy Professionals	People with a basic knowledge of statistics and/or machine learning, who can define key questions that can be answered using advanced analytics
Analysts & Data Savvy Managers <i>Projected U.S. talent gap: 1.5 million</i>		Technology & Data Enablers	People providing technical expertise to support analytical projects. Skills sets including computer programming and database administration

Note: Figures above reflect a projected talent gap in US **in 2018**, as shown in McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

Data Scientist Key Activities

- Reframe business challenges as analytics challenges
- Design, implement and deploy statistical models and data mining techniques on big data
- Create insights that lead to actionable recommendations



Profile of a Data Scientist

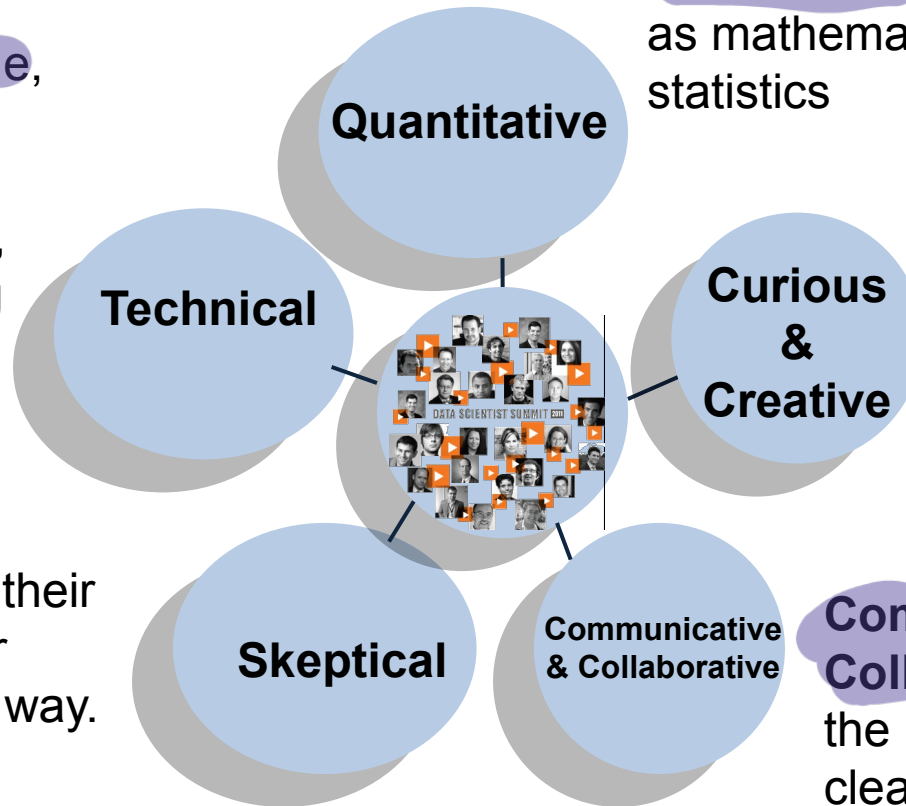
Technical aptitude, such as software engineering, machine learning, and programming skills.

Quantitative skills, such as mathematics or statistics

Curious & Creative, must be passionate about data and finding creative ways to solve problems and portray information

Skeptical examine their work critically rather than in a one-sided way.

Communicative & Collaborative: articulate the business value in a clear way, and work collaboratively with project sponsors and key stakeholders.





Big Data Analytics Case Examples

Big Data Analytics: Industry Examples

- 1 Health Care
 - Reducing Cost of Care
- 2 Public Services
 - Preventing Pandemics
- 3 IT Infrastructure
 - Unstructured Data Analysis
- 4 Online Services
 - Social Media for Professionals



Big Data Analytics: Health/*Public* Services



Situation

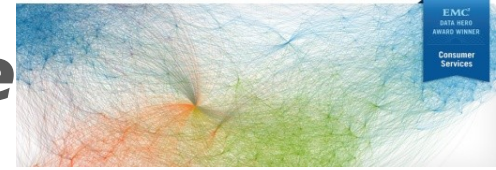
- Threat of global pandemics has increased exponentially
- Pandemics spreads at faster rates

Use of Big Data

- Created a network of viral listening posts
- Combines data from viral discovery in the field, research in disease hotspots, and social media trends
- Using Big Data to make accurate predictions on spread of new pandemics

Key Outcomes

- Identified a fifth form of human malaria, including its origin
- Identified why efforts failed to control swine flu
- Proposing more proactive approaches to preventing outbreaks



Situation

- Opportunity to create social media space for professionals

Use of Big Data

- Collects and analyzes data from over 100 million users
- Adding 1 million new users per week

Key Outcomes

- LinkedIn Skills, Job Recommendations, Recruiting
- Established a diverse data scientist group, as founder believes this is the start of Big Data revolution



Check Your Knowledge

Check Your Knowledge: (From the Textbook)

1. What are the most important characteristics of Big Data, **(slide 5-7)** and what are the main considerations in processing Big Data? **slide 19 / TextBook page 11**
2. Explain the differences between BI and Data Science. **slide 15/TextBook page 12 & 13**
3. Describe the challenges of the current analytical architecture for data scientists.
4. What are the key skill sets and behavioral characteristics of a data scientist? **slide 33**