Intro to Big Data Analytics

CSC 410 Data Engineering
Spring 2020

Instructor: Dr. Md L Ali
E-mail: mdali@rider.edu

Contents

- Bid Data overview
- State of the practice in analytics
- Business Intelligence versus Data Science
- Key roles for the new Big Data ecosystem
- The Data Scientist
- Examples of Big Data analytics

Big Data Overview

- Industries that gather and exploit data
  - Credit card companies monitor purchase
    - Good at identifying fraudulent purchases
  - Mobile phone companies analyze calling patterns – e.g., even on rival networks
    - Look for customers might switch providers
  - For social networks data is primary product
    - Intrinsic value increases as data grows

Attributes Defining Big Data Characteristics

- Huge volume of data
  - Not just thousands/millions, but billions of items
- Complexity of data types and structures
  - Varity of sources, formats, structures
- Speed of new data creation and grow
  - High velocity, rapid ingestion, fast analysis

Big Data Characteristics

The characteristics of big data are often referred to as the three Vs:

- **Volume: How much data is there?** Big Data observes and tracks what happens from various sources which include business transactions, social media and information from machine-to-machine or sensor data. This creates large volumes of data.
- **Variety: How diverse are different types of data?** Data comes in all formats that may be structured, numeric in the traditional database or the unstructured text documents, video, audio, email, stock ticker data.
- **Velocity: At what speed is new data generated?** The data streams in high speed and must be dealt with timely. The processing of data that is, analysis of streamed data to produce near or real time results is also fast.

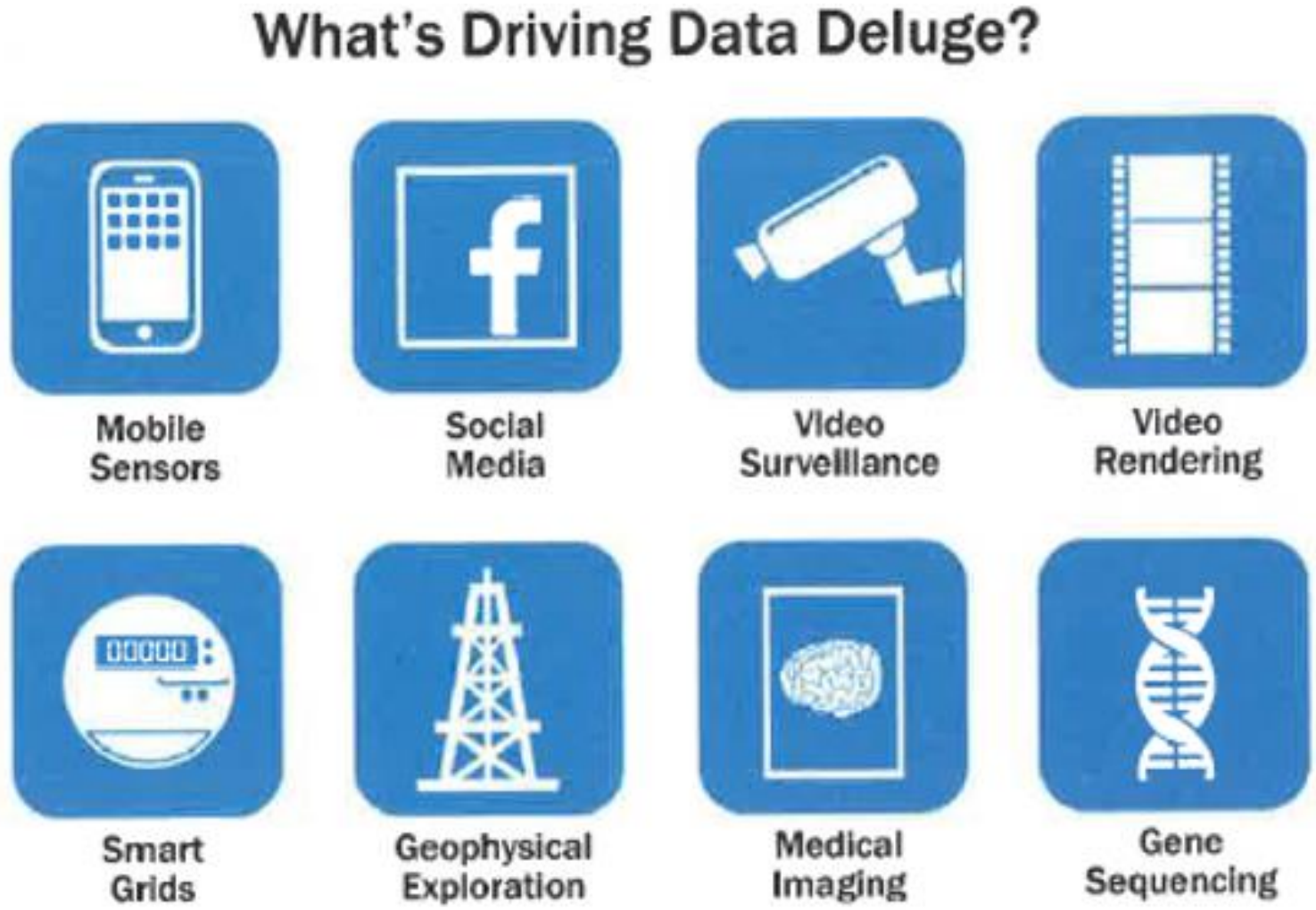
Big Data Analytics Importance

- **Cost Savings :** help in identifying more efficient ways of doing business.
- **Time Reductions :**helps businesses analyzing data immediately and make quick decisions based on the learnings.
- **New Product Development :** By knowing the trends of customer needs and satisfaction through analytics you can create products according to the wants of customers.
- **Understand the market conditions :** By analyzing big data you can get a better understanding of current market conditions.
- **Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company.

Sources of Big Data Deluge

- Mobile sensors – GPS, accelerometer, etc.
- Social media – 700 Facebook updates/sec in 2012
- Video surveillance – street cameras, stores, etc.
- Video rendering – processing video for display
- Smart grids – gather and act on information
- Geophysical exploration – oil, gas, etc.
- Medical imaging – reveals internal body structures
- Gene sequencing – more prevalent, less expensive, healthcare would like to predict personal illnesses

Sources of Big Data Deluge



Example: Genotyping from 23andme.com

Fig. Examples of what can be learned through genotyping, from 23andme.com

Data Structures: Characteristics of Big Data

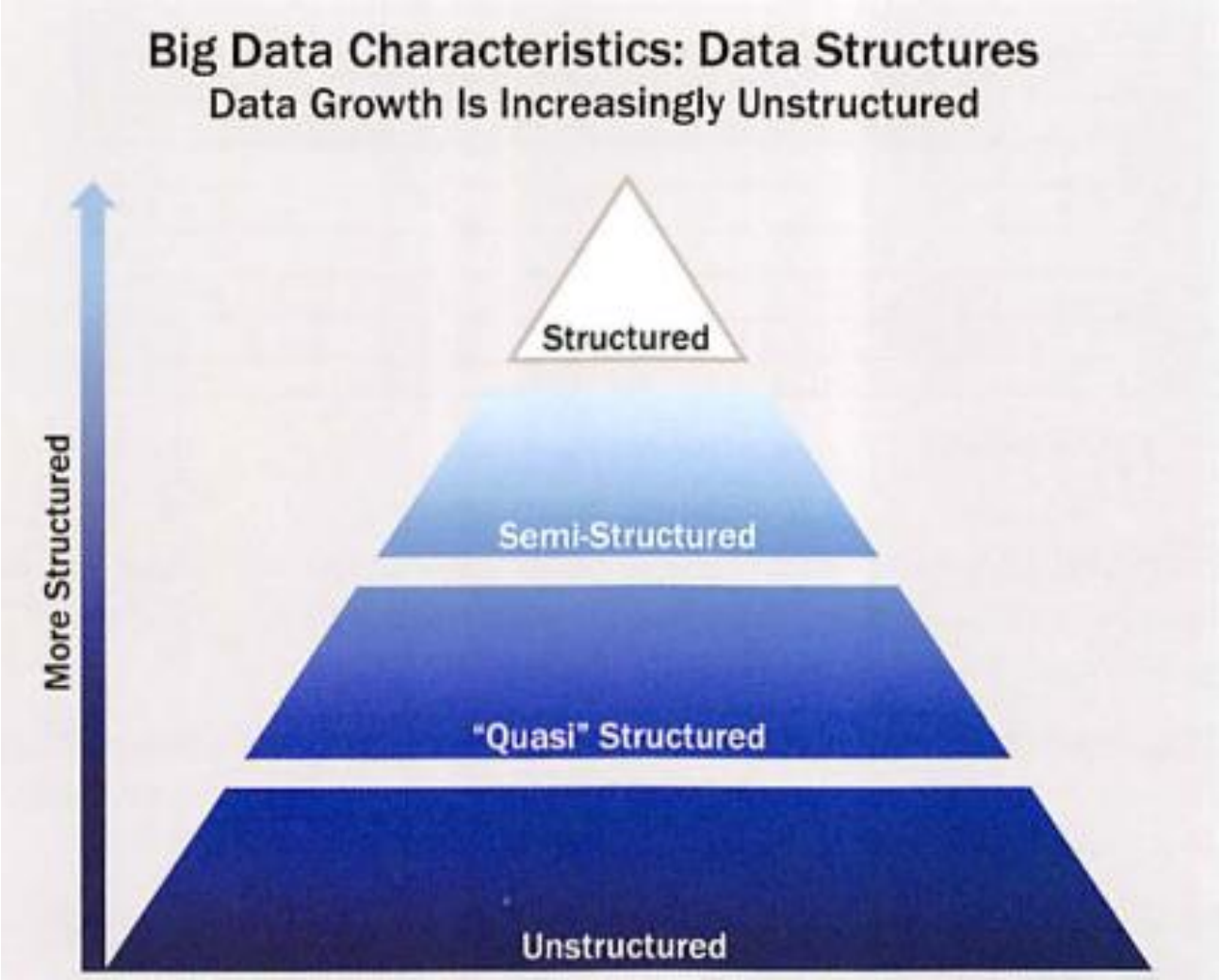


Fig. Big Data growth is increasingly unstructured

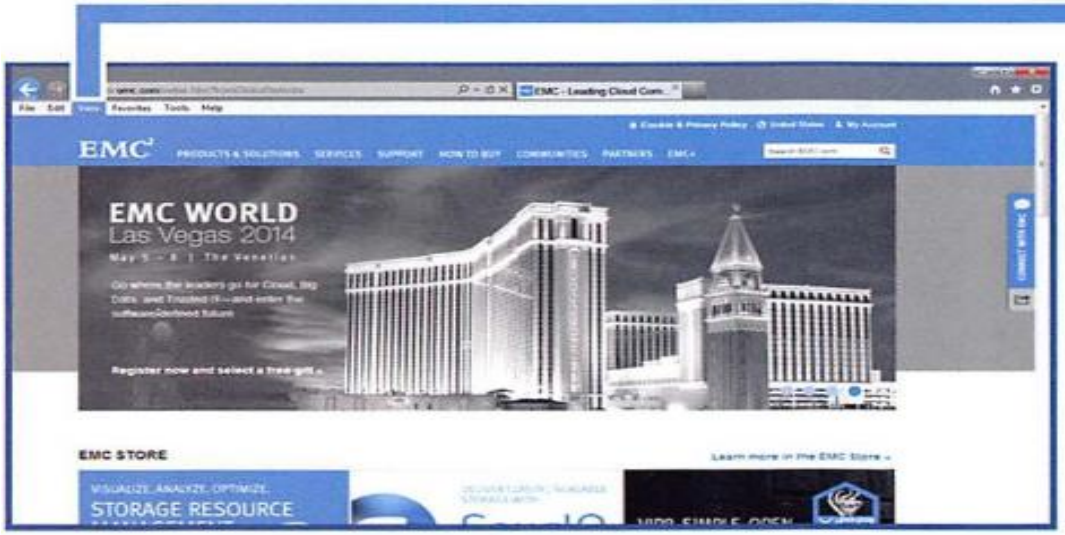
Data Structures: Characteristics of Big Data

- Structured – Data containing a defined data type, format, structure (Transactional data, OLAP cubes, RDBMS, CVS files, spreadsheets)
- Semi-structured – Textual data files with a discernable patterns that enable parsing – e.g., XML data
- Quasi-structured - Textual data with erratic data formats that can be formatted with effort, tools, and time– e.g., web clickstream data
- Unstructured - Data with no inherent structure – text docs, PDF's, images, video

Example of 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
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Example of Semi-Structured Data




```
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

<meta name="description" content="EMC is a leading provider of IT storage hardware solutions to promote dat
cloud computing.">
name="keywords" content="emc,network storage,data recovery,information management,backup software,nas storage

<meta name="viewport" content="width=device-width, initial-scale=1">

<link href="/_admin/css/html-layout-css-includes-combined-min.css" rel="stylesheet">
<script src="/_admin/js/jquery.js"></script>
<link rel="stylesheet" href="/R1/assets/css/common/normalize.css">
<link rel="stylesheet" href="/R1/assets/css/homepage/main.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-header.css">
<link rel="stylesheet" href="/R1/assets/css/common/responsive-footer.css">


<script type="text/javascript" src="//platform.twitter.com/widgets.js"></script>
<script src="/R1/assets/js/common/modernizr-2.6.2.min.js"></script>
```

Example of Quasi-Structured Data

visiting 3 websites adds 3 URLs to user's log files

Visiting these three websites add three URLs to the log files monitoring the user's computer or network use. Together, this comprises a clickstream that can be parsed and mined by data scientist to discover usage patterns and uncover relationships among clicks and areas of interest on a website or group of site .

Example of Unstructured Data

Video about Antarctica Expedition

Types of Data Repositories from an Analyst Perspective

Table 1.1 Types of Data Repositories, from an Analyst Perspective

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

State of the Practice in Analytics

- Business Intelligence (BI) versus Data Science
- Current Analytical Architecture
- Drivers of Big Data
- Emerging Big Data Ecosystem and a New Approach to Analytics

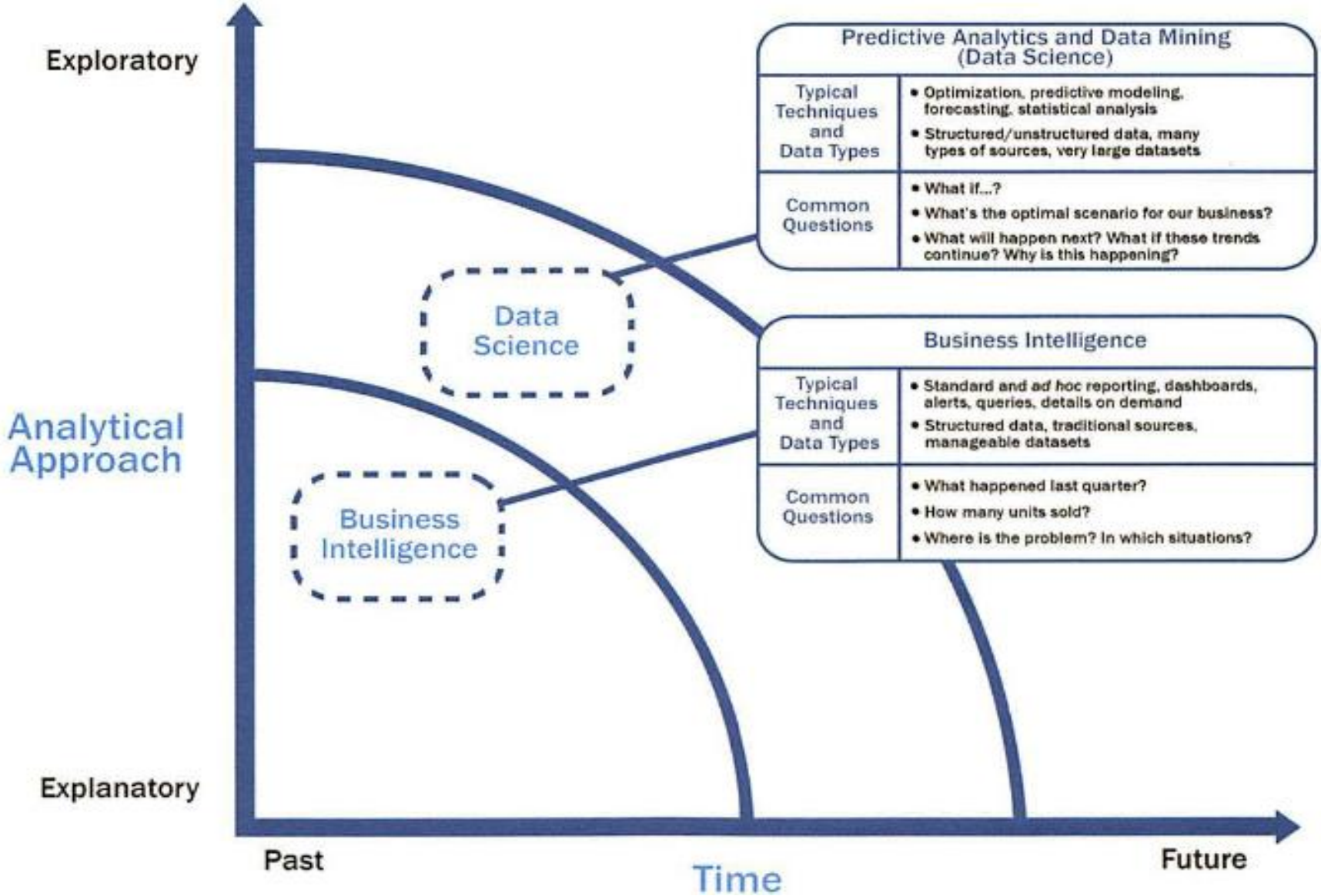
Business Drivers for Advanced Analytics

Current business problems provide many opportunities for organizations to become more analytical and data driven, as shown in Table 1.2.

Table 1.2 Business Drivers for Advanced Analytics

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

Business Intelligence (BI) versus Data Science



Current Analytical Architecture

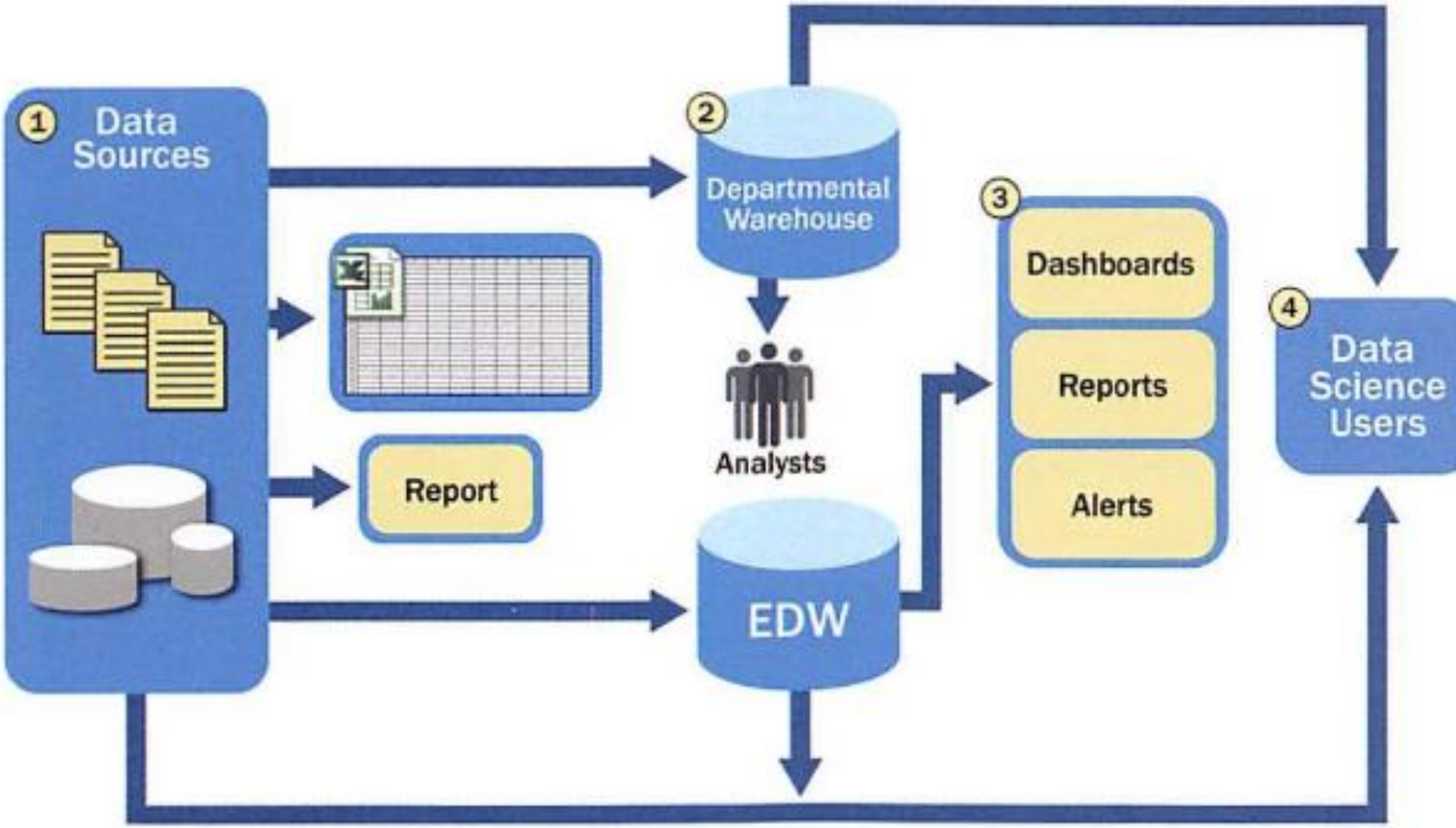


Fig. Typical Analytic Architecture

Current Analytical Architecture

The figure in the previous slide shows a typical data architecture and several of the challenges it presents to data scientists and others trying to do advanced analytics.

Data sources must be well understood, structured, and normalized

EDW – Enterprise Data Warehouse

From the EDW data is read by applications

Data scientists get data for downstream analytics processing

Current Analytical Architecture - Problem

High-value data is hard to reach and leverage, and predictive analytics and data mining activities are last in line for data.

Data scientists are limited to performing in-memory analytics, which will restrict the size of the datasets . So Analyst works on sampling, which can skew model accuracy.

Data Science projects will remain isolated rather than centrally managed. The implication of this is that the organization can never tie together the power of advanced analytics.

Current Analytical Architecture - Solution

- One solution to this problem is to introduce **analytic sandboxes** to enable data scientists to perform advanced analytics in a controlled and sanctioned way. Meanwhile, the current data Warehousing solutions continues offering reporting and BI services to support management and mission-critical operations.

Drivers of Big Data

Data Evolution & Rise of Big Data Sources

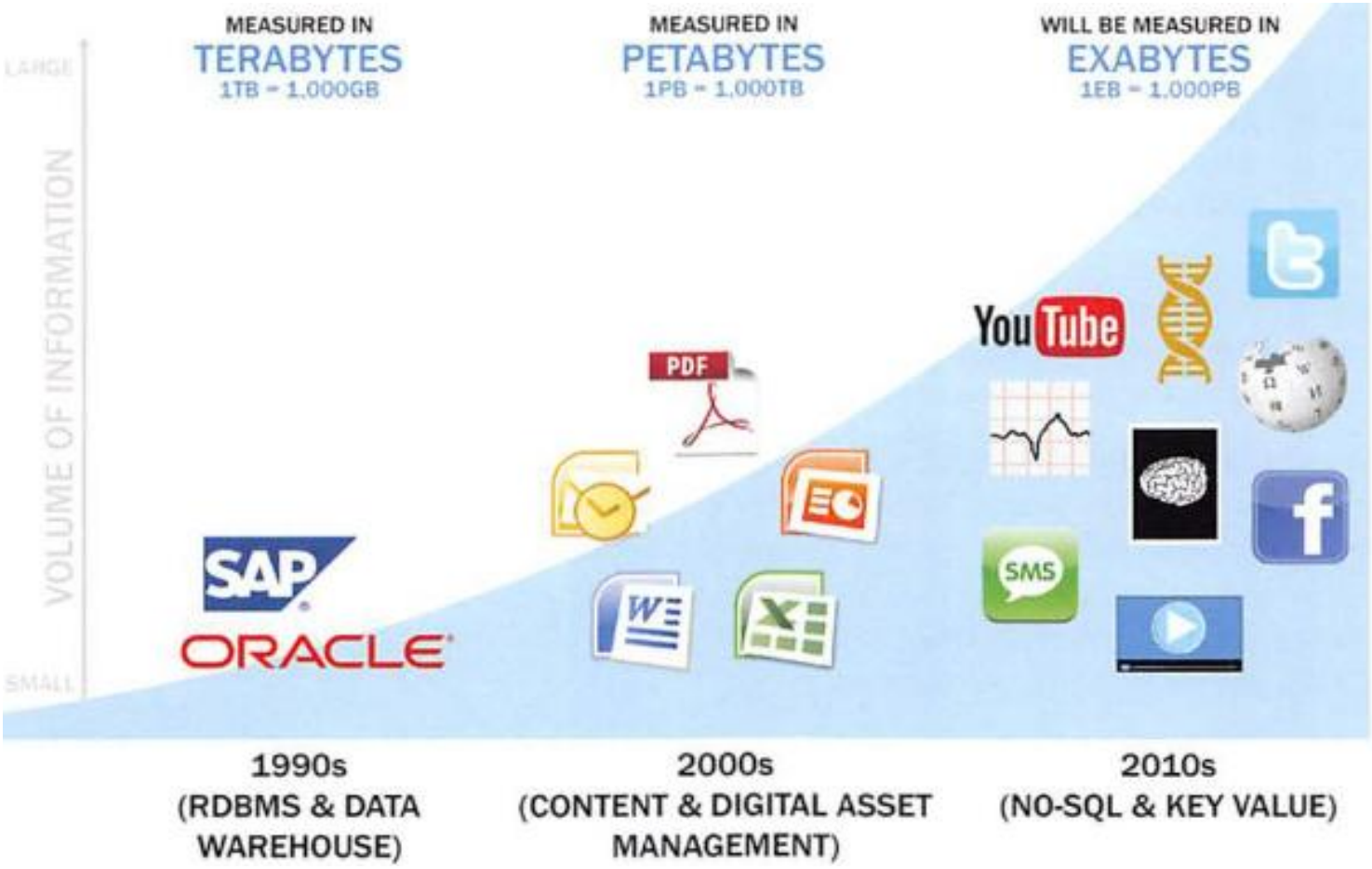


Figure shows a summary perspective on sources of Big Data generated by new applications and the scale and growth rate of the data.

Drivers of Big Data

Data Evolution & Rise of Big Data Sources

The data now comes from multiple sources, such as these:

- Medical information, such as diagnostic imaging
- Photos and video footage uploaded to the World Wide Web
- Video surveillance, such as the thousands of video cameras across a city
- Mobile devices, which provide geospatial location data of the users
- metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart
- Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

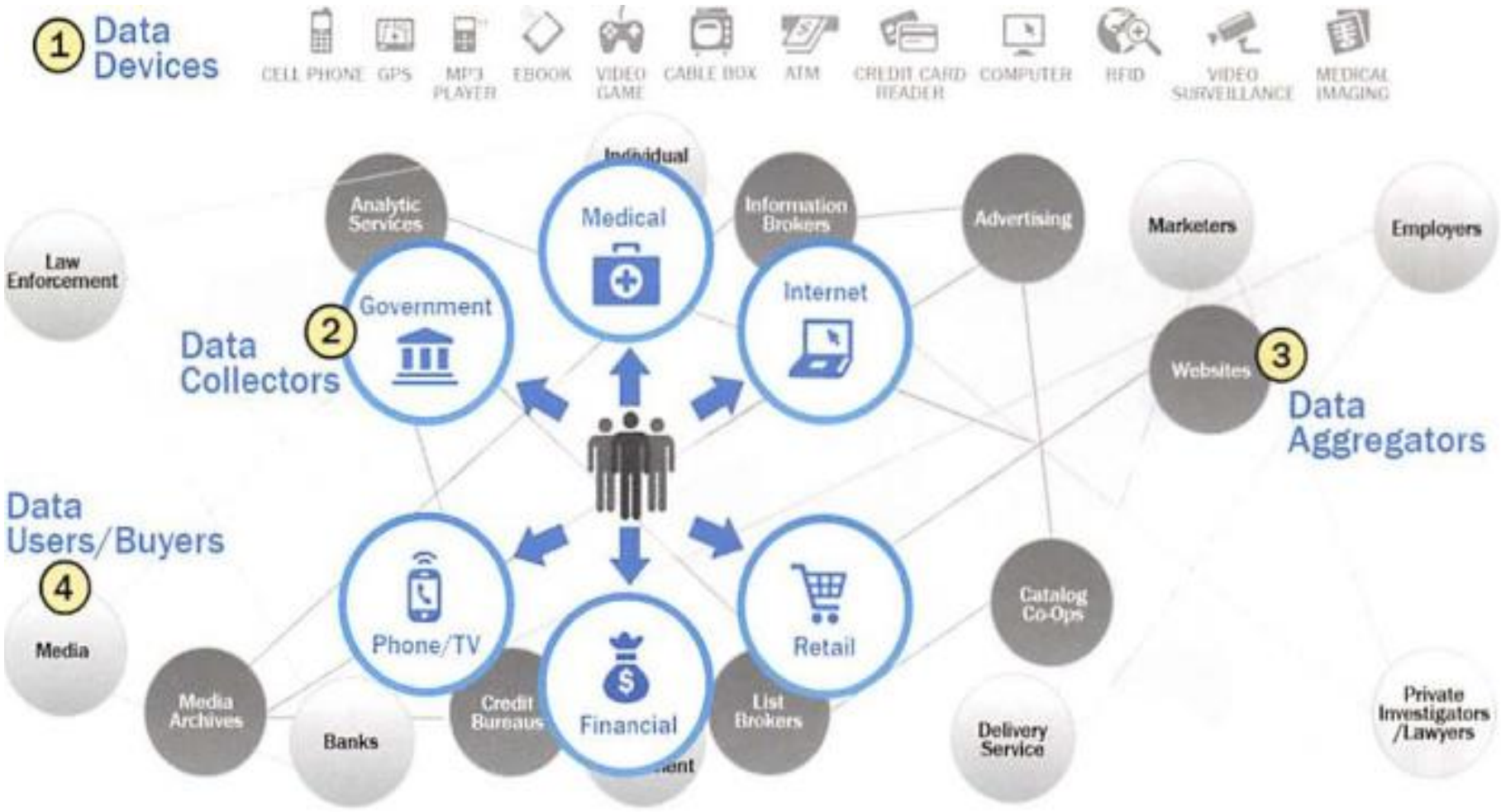
Emerging Big Data Ecosystem and a New Approach to Analytics

- As the new ecosystem takes shape, there are four main groups of players within this interconnected web
  - Data devices
    - Games, smartphones, computers, etc.
  - Data collectors
    - Phone and TV companies, Internet, Gov't, etc.
  - Data aggregators – make sense of data
    - Websites, credit bureaus, media archives, etc.
  - Data users and buyers
    - Banks, law enforcement, marketers, employers, etc.

Emerging Big Data Ecosystem and a New Approach to Analytics

- Data devices
  - 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.
  - For example, playing an online video game, Smartphones data, Retail shopping loyalty cards data
- Data collectors
  - Include sample entities that collect data from the device and users.
  - For example, Retail stores tracking the path a customer
- Data aggregators – make sense of data
  - They transform and package the data as products to sell to list brokers for specific ad campaigns.
- Data users and buyers
  - These groups directly benefit from the data collected and aggregated by others within the data value chain.
  - For Example, People want to determine public sentiments toward a candidate by analyzing related blogs and online comments

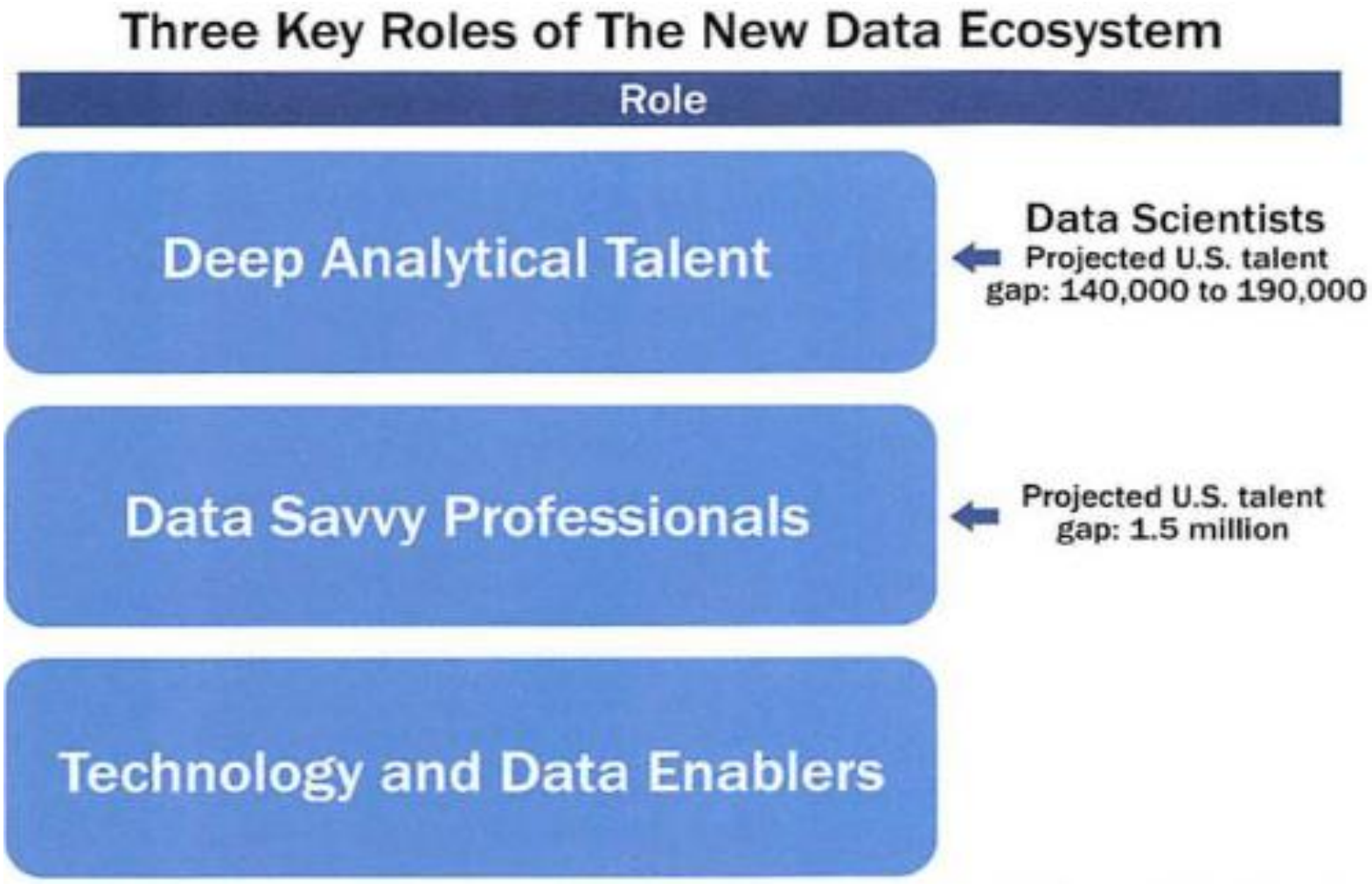
Emerging Big Data Ecosystem and a New Approach to Analytics



Key Roles for the New Big Data Ecosystem

1. Deep analytical talent
   - Advanced training in quantitative disciplines – e.g., math, statistics, machine learning
2. Data savvy professionals
   - Savvy but less technical than group 1
3. Technology and data enablers
   - Support people – e.g., DB admins, programmers, etc.
   - This group represents people providing technical expertise to support analytical projects,
   - such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures

Three Key Roles of the New Big Data Ecosystem



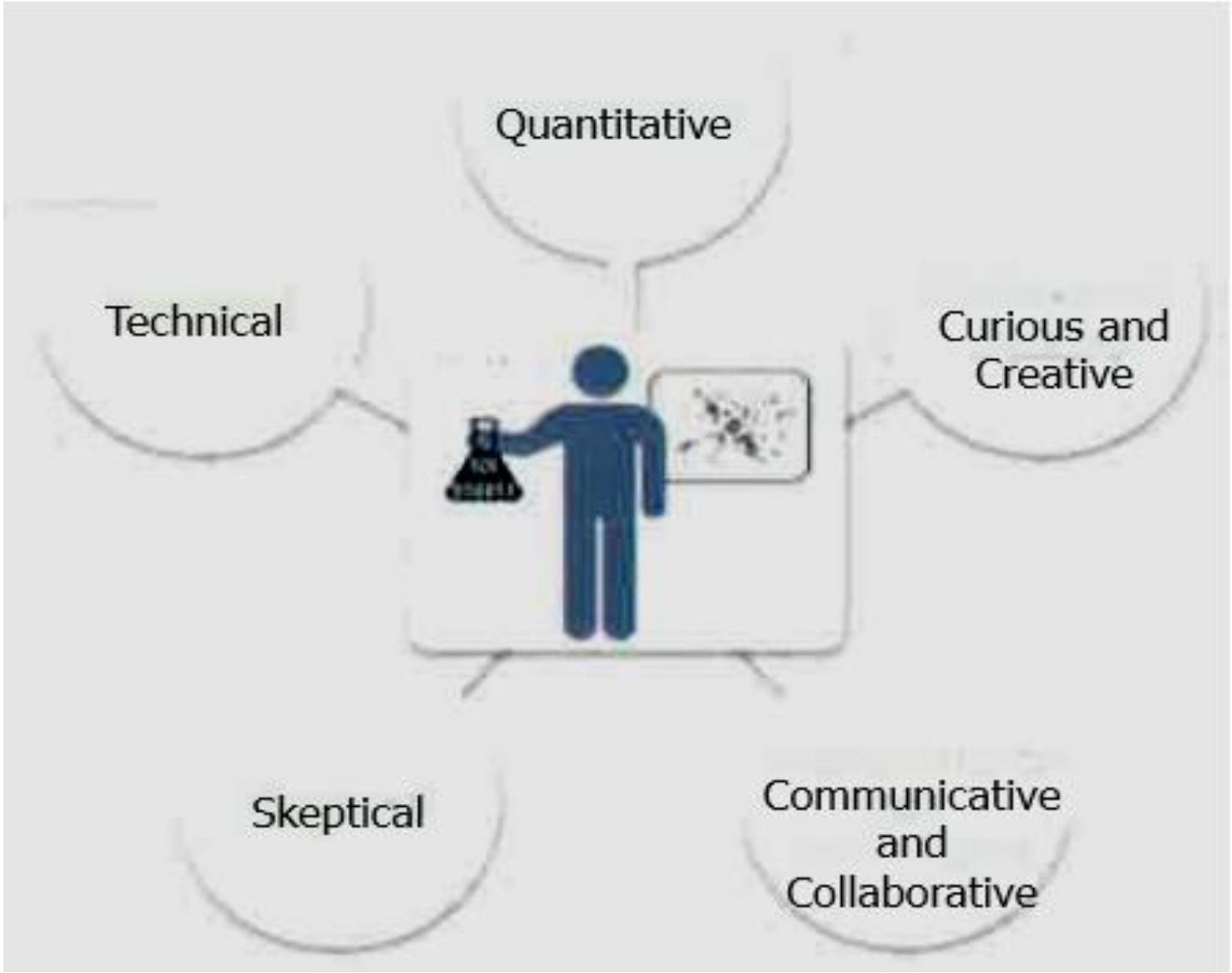
Three Recurring Data Scientist Activities

There are three recurring sets of activities that data scientists perform:

1. Reframe business challenges as analytics challenges
2. Design, implement, and deploy statistical models and data mining techniques on Big Data
3. Develop insights that lead to actionable recommendations

Profile of Data Scientist

Five Main Sets of Skills



Profile of Data Scientist
Five Main Sets of Skills

- Quantitative skill – e.g., math, statistics
- Technical aptitude – e.g., software engineering, programming
- Skeptical mindset and critical thinking – ability to examine work critically
- Curious and creative – passionate about data and finding creative solutions
- Communicative and collaborative – can articulate ideas, can work with others

Examples of Big Data Analytics

Three examples of Big Data Analytics in different areas:

- Retailer Target: Target's statisticians determined that the retailer made a great deal of money from three main life event situations.
  - Marriage, when people tend to buy many new products
  - Divorce, when people buy new products and change their spending habits
  - Pregnancy, when people have many new things to buy and have an urgency to buy them
- Apache Hadoop
  - Open source Big Data infrastructure innovation
  - MapReduce paradigm, ideal for many projects
- Social Media Company LinkedIn
  - Social network for working professionals
  - Can graph a user's professional network
  - 250 million users in 2014

Data Visualization of User's Social Network Using InMaps

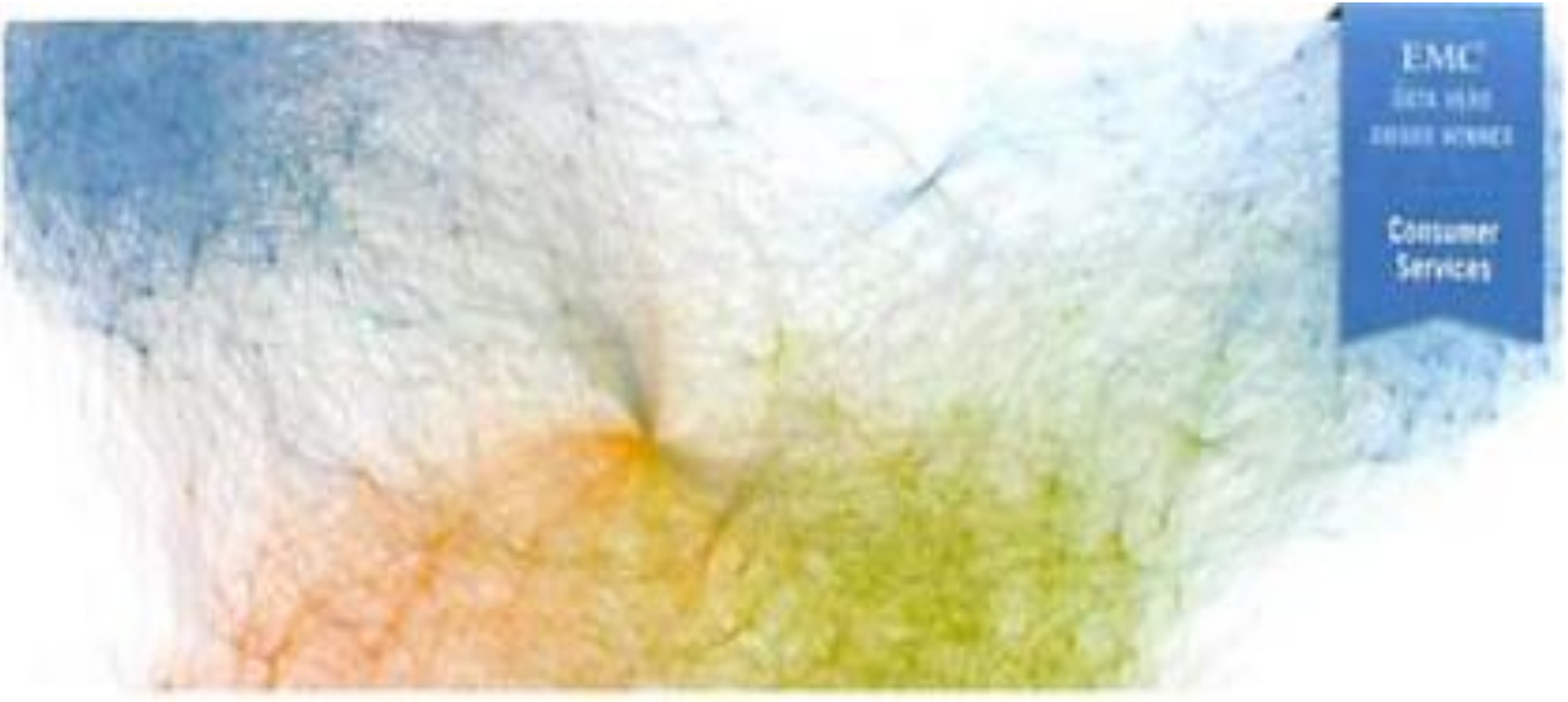


Figure shows an example of an InMap visualization that enables a LinkedIn user to get a broader view of the interconnectedness of his contacts and understand how he knows most of them.

Summary

- Big Data comes from myriad sources
  - Social media, sensors, IoT, video surveillance, and sources only recently considered
- Companies are finding creative and novel ways to use Big Data
- Exploiting Big Data opportunities requires
  - New data architectures
  - New machine learning algorithms, ways of working
  - People with new skill sets