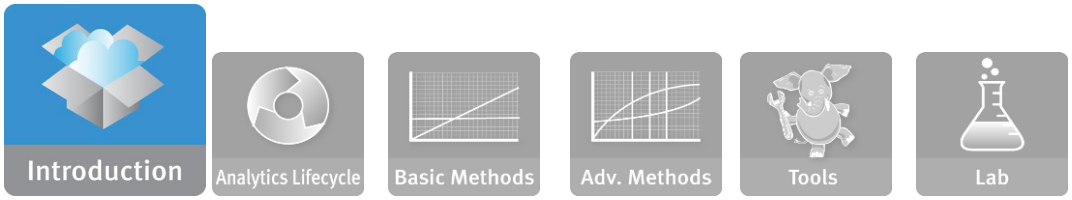
Module 2– Introduction to Big Data Analytics

EMC^2 PROVEN PROFESSIONAL

Module 2: Introduction to Big Data Analytics

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

- Define big data
- 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
- Cite at least three illustrative examples of big data opportunities



Module 2: Introduction to Big Data Analytics

Lesson 1: Big Data Overview

During this lesson the following topics are covered:

- Definition of big data
- Big data characteristics and considerations
- Unstructured data fueling big data analytics
- Analyst perspective on Data Repositories

Introduction to Big Data Analytics

Participate in this weeks discussion

What is *Big Data*?

What makes data, "*Big*" *Data*?

Big Data Defined

- ***"Big Data" is data whose 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, analytic sandboxes
  - New tools
  - New analytical methods
  - Integrating multiple skills into new role of data scientist

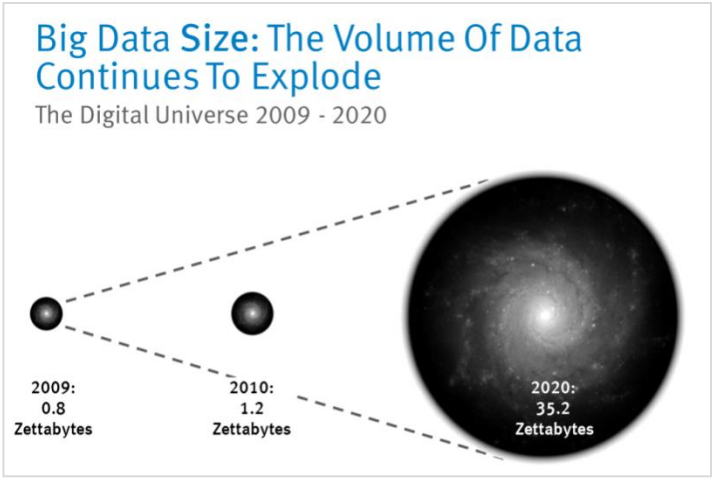
- Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities

Source: McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

Key Characteristics of Big Data

1. Data Volume

- 44x increase from 2010 to 2020
  (1.2zettabytes to 35.2zb)



2. Processing Complexity

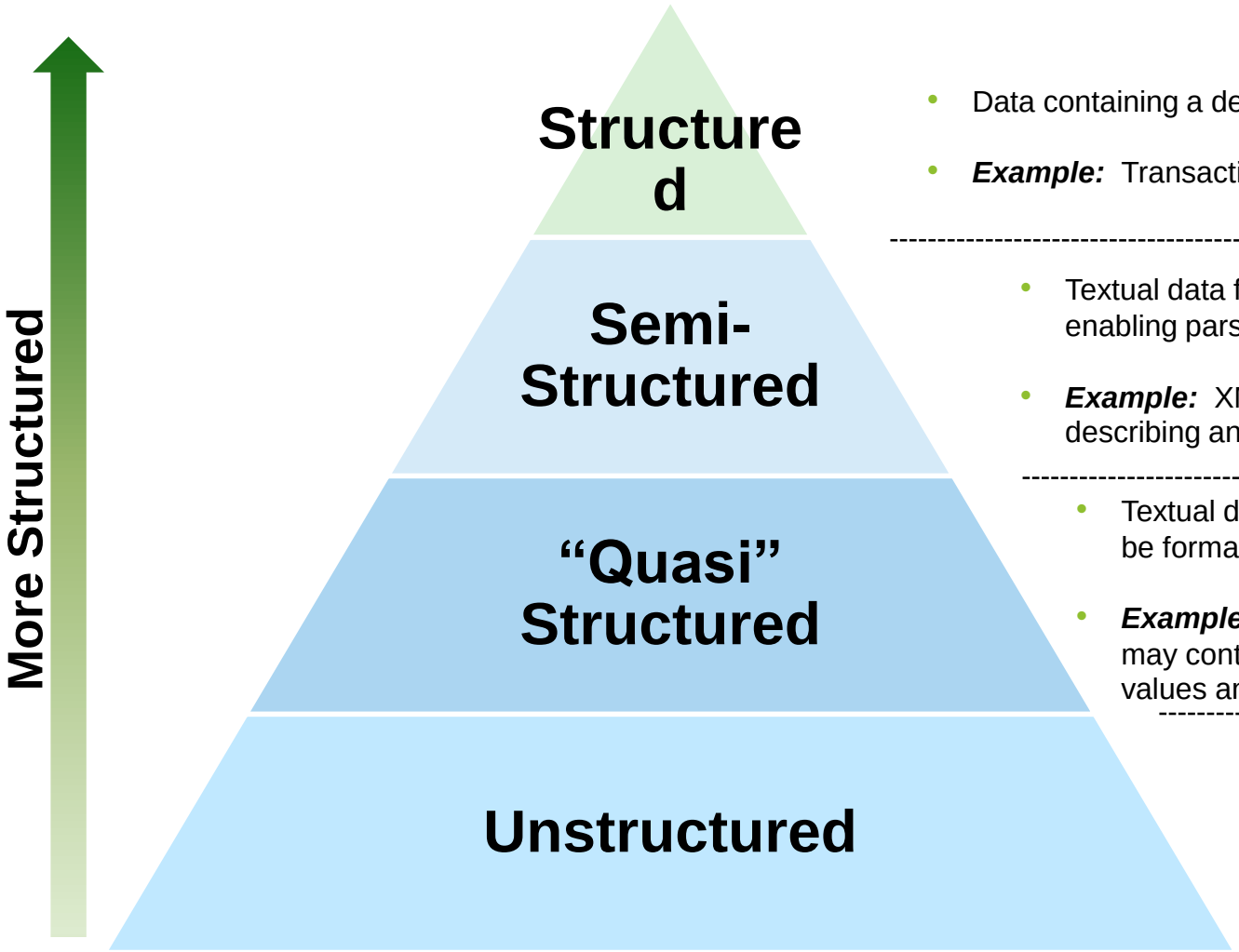
- Changing data structures
- Use cases warranting additional transformations and analytical techniques

3. Data Structure

- Greater variety of data structures to mine and analyze

Big Data Characteristics: Data Structures

Data Growth is Increasingly Unstructured



- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

---

- Textual data with erratic data formats, can be formatted with effort, tools, and time
- ***Example:*** Web clickstream data that may contain some inconsistencies in data values and formats

---

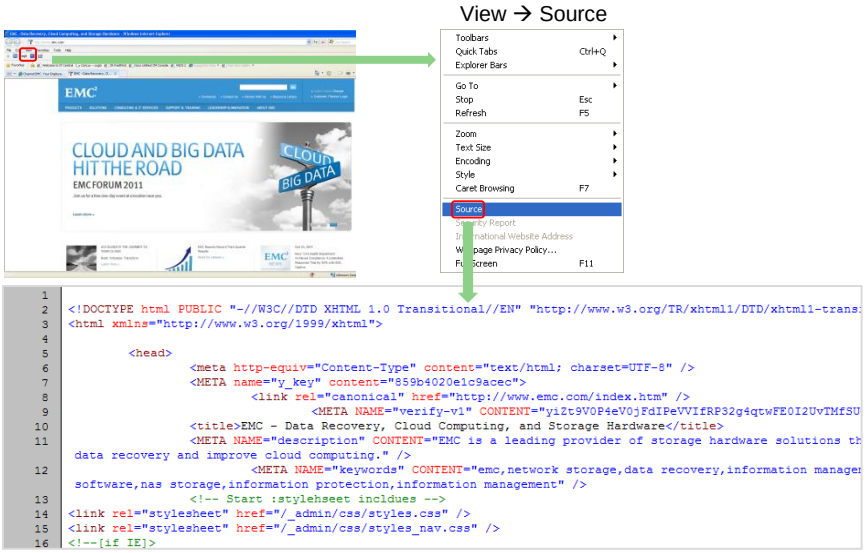
- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images and video

Four Main Types of Data Structures

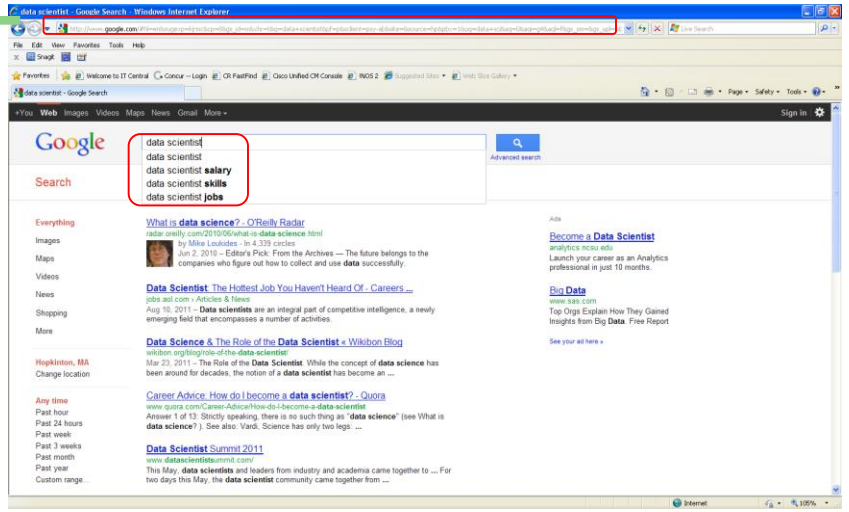
Structured Data

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



Quasi-Structured Data



http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=d566e0fbd09c8604&biw=1382&bih=651

Unstructured Data

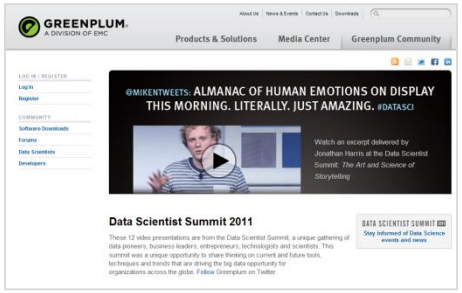
The Red Wheelbarrow, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.



Data Repositories, An Analyst Perspective

Data Islands
"Spreadmarts"

Isolated data marts

- Spreadsheets and low-volume DB's for recordkeeping
- Analyst dependent 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 & DBAs for data access and schema changes
- Analysts must spend significant time to get extracts from multiple sources

Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- *"Analyst-owned" rather than "DBA owned"*

Module 2: Introduction to Big Data Analytics

Lesson 1: Summary

During this lesson the following topics were covered:

- Definition of big data
- Big data characteristics and considerations
- Unstructured data fueling big data analytics
- Analyst perspective on Data Repositories

Module 2: Introduction to Big Data Analytics

Lesson 2: State of the Practice in Analytics

During this lesson the following topics are covered:

- Business drivers for analytics
- Current analytical architecture
- Business intelligence vs. data science
- Drivers of big data and new big data ecosystem

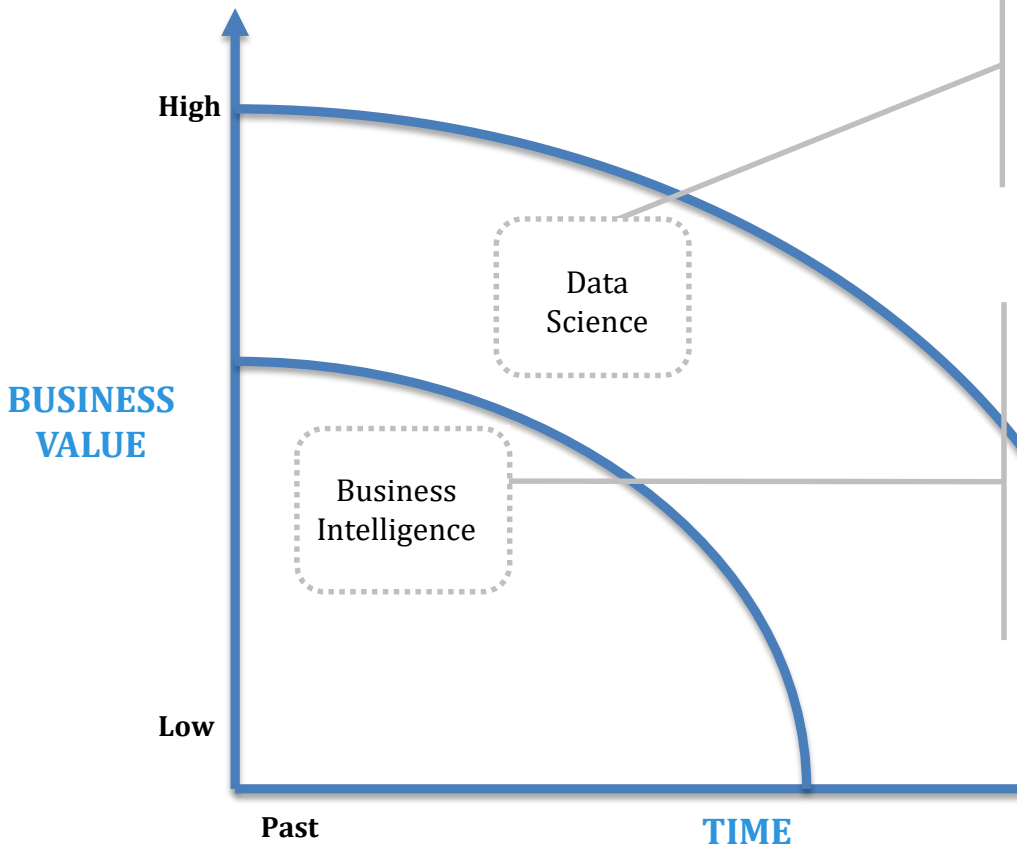
Business Drivers for Analytics

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, default
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II

Analytical Approaches for Meeting Business Drivers

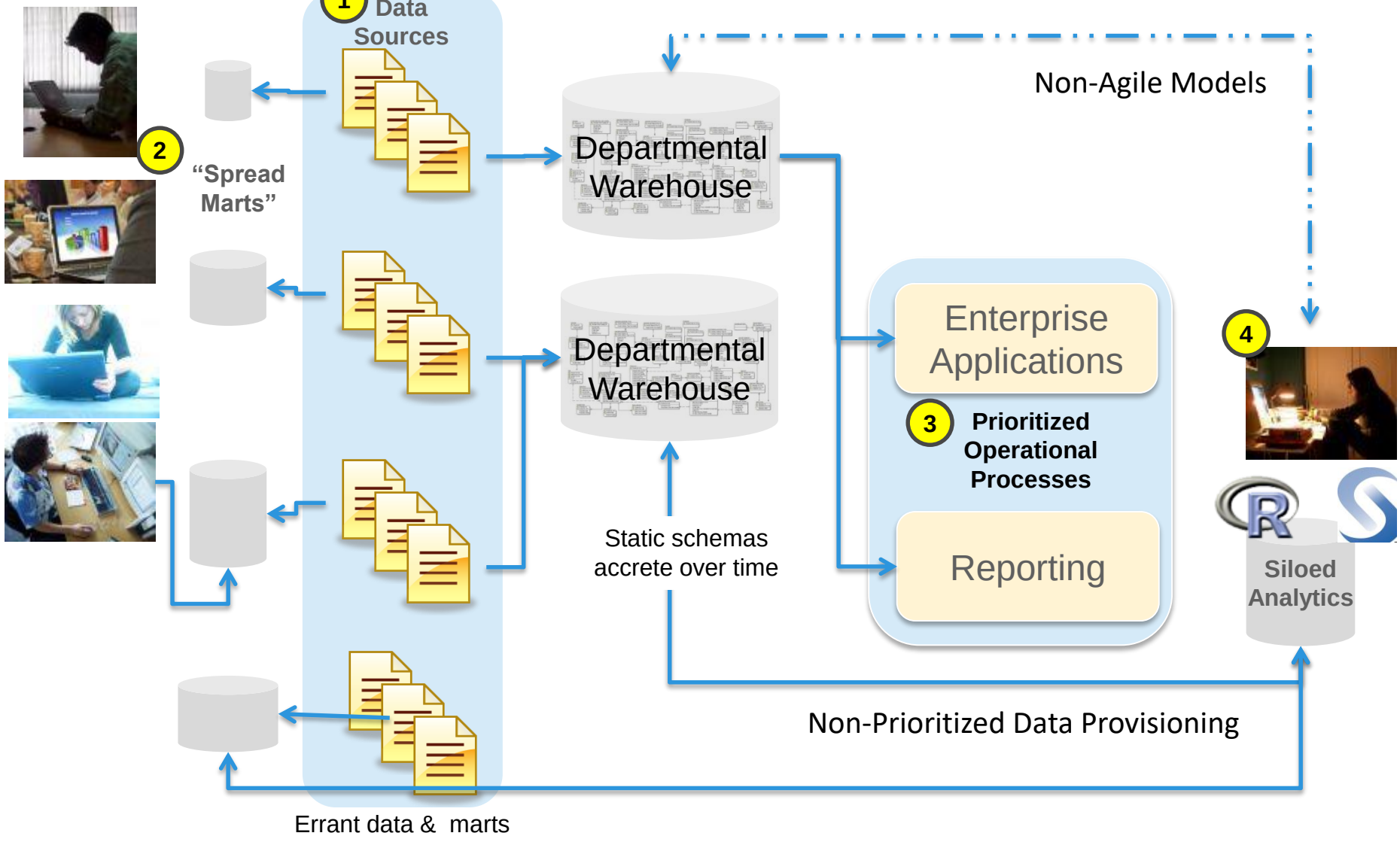
Business Intelligence vs. Data Science



Predictive Analytics & Data Mining (Data Science)	
Typical Techniques & Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if…..? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques & Data Types	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

A Typical Analytical Architecture



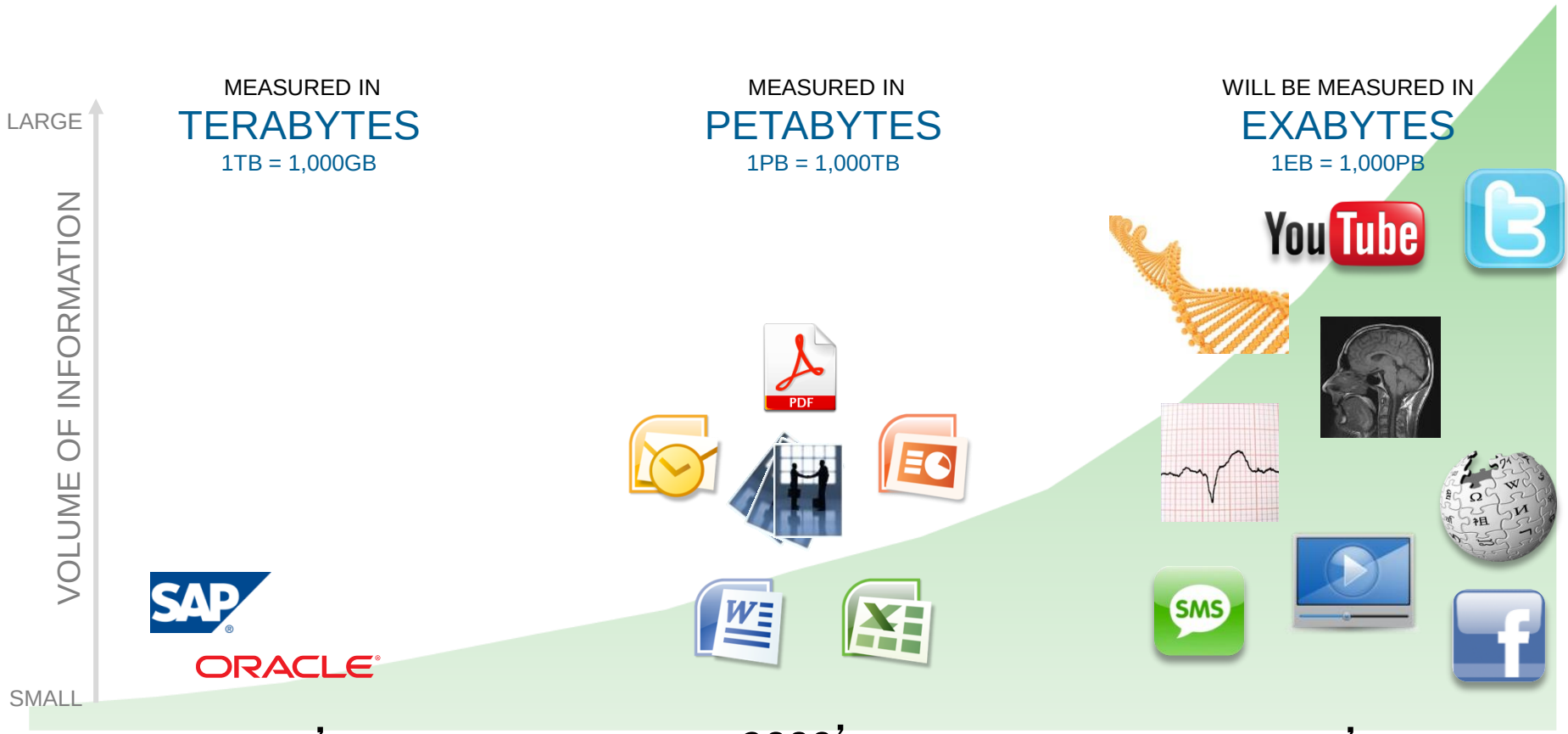
Implications of Typical Architecture for Data Science

- High-value data is hard to reach and leverage
- Predictive analytics & data mining activities are last in line for data
  - Queued after prioritized operational processes
- Data is moving in batches from EDW to local analytical tools
  - In-memory analytics (such as R, SAS, SPSS, Excel)
  - Sampling can skew model accuracy
- Isolated, *ad hoc* analytic projects, rather than centrally-managed harnessing of analytics
  - Non-standardized initiatives
  - Frequently, not aligned with corporate business goals

Slow "time-to-insight" & reduced business impact

Opportunities for a New Approach to Analytics

New Applications Driving Data Volume



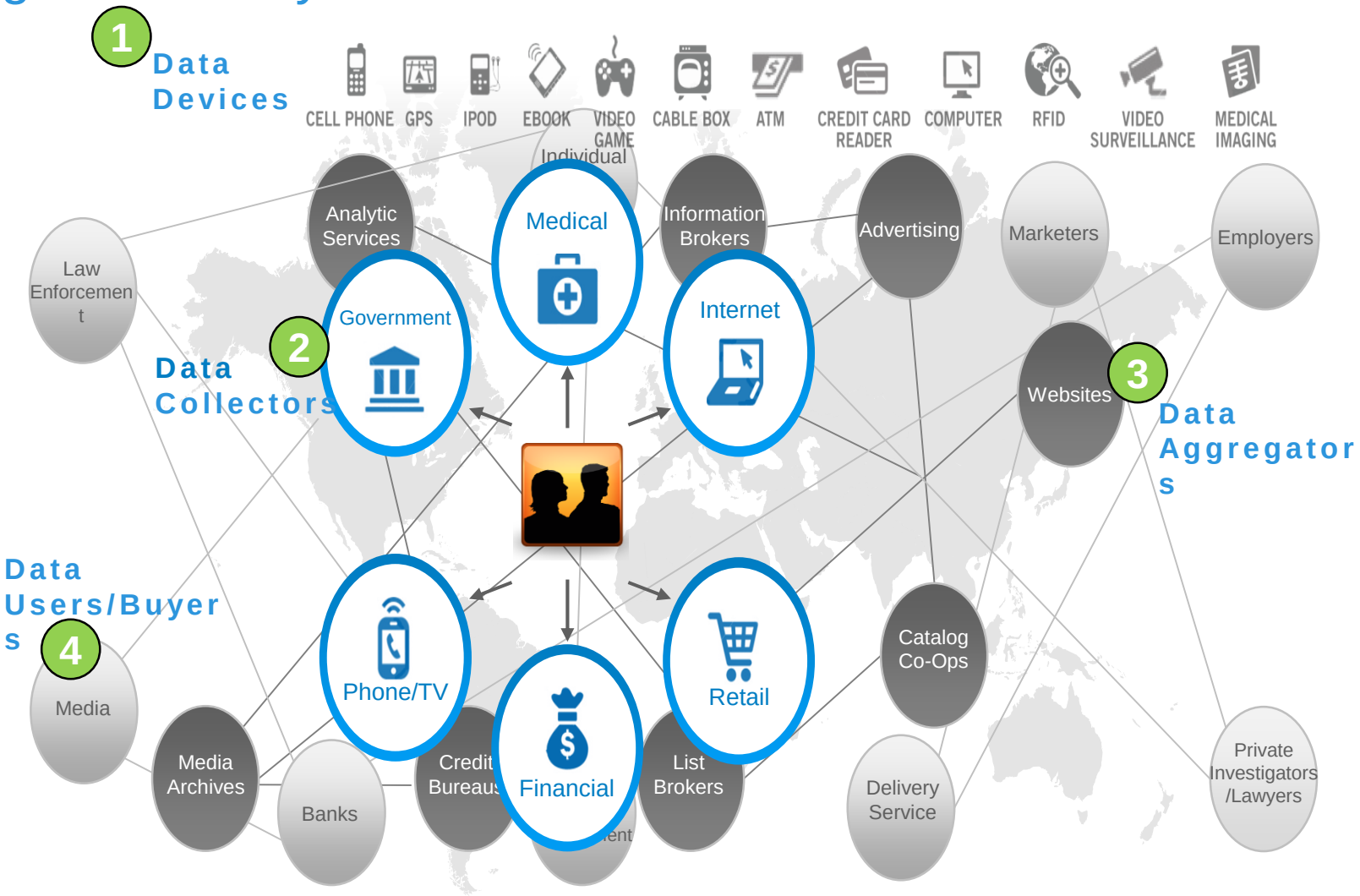
1990's (RDBMS & DATA WAREHOUSE)

2000's (CONTENT & DIGITAL ASSET MANAGEMENT)

2010's (NO-SQL & KEY/VALUE)

Opportunities for a New Approach to Analytics

Big Data Ecosystem



Considerations for Big Data Analytics

Criteria for Big Data Projects

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

New Analytic Architecture

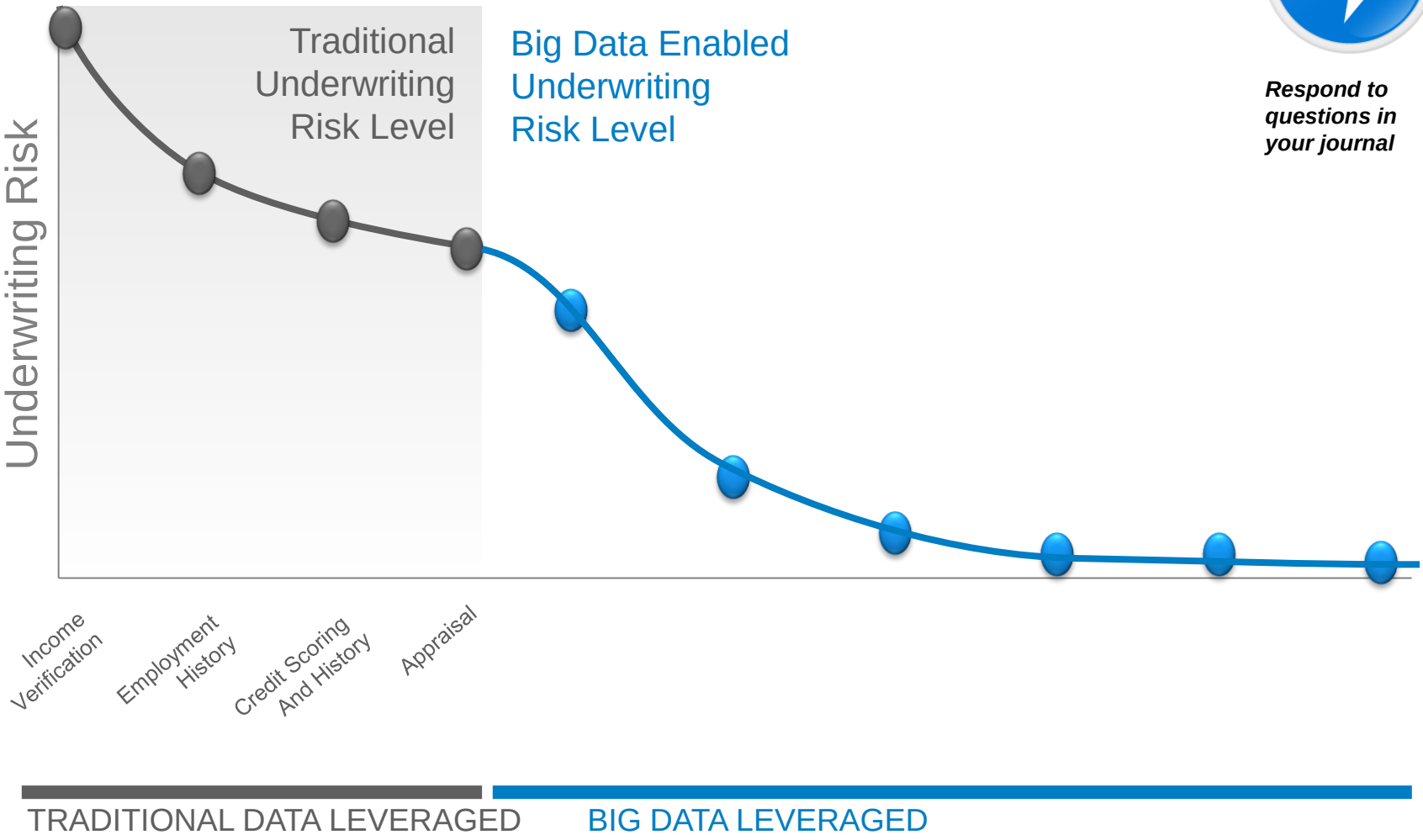
Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- “Analyst-owned” rather than “DBA owned”

State of the Practice in Analytics: Mini-Case Study

Big Data Enabled Loan Processing at Yoyodyne



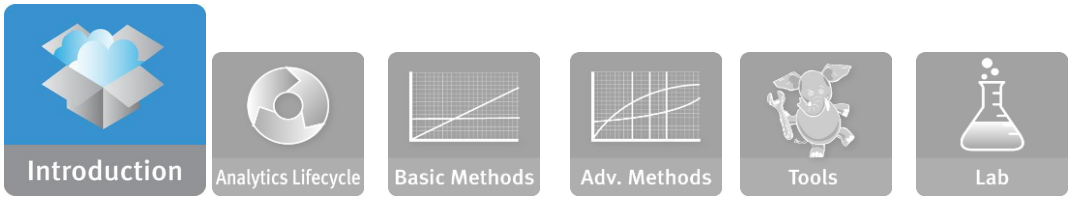
Respond to questions in your journal

Module 2: Introduction to Big Data Analytics

Lesson 2: Summary

During this lesson the following topics were covered:

- Business drivers for analytics
- Current analytical architecture
- Business intelligence vs. data science
- Drivers of big data and new big data ecosystem



Module 2: Introduction to Big Data Analytics

Lesson 3: The Data Scientist

During this lesson the following topics are covered:

- Key Roles of the New Big Data Ecosystem
- Profile of a Data Scientist

Skills Needed In the New Data Ecosystem

Respond to questions in your journal

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

Three Key Roles of the New Data Ecosystem

Data Scientists

Projected U.S. talent gap: 140,000 to 190,000

Analysts & Data Savvy Managers

Projected U.S. talent gap: 1.5 million

Role	Role Description
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
Technology & Data Enablers	People providing technical expertise to support analytical projects. Skills sets including computer programming and database administration

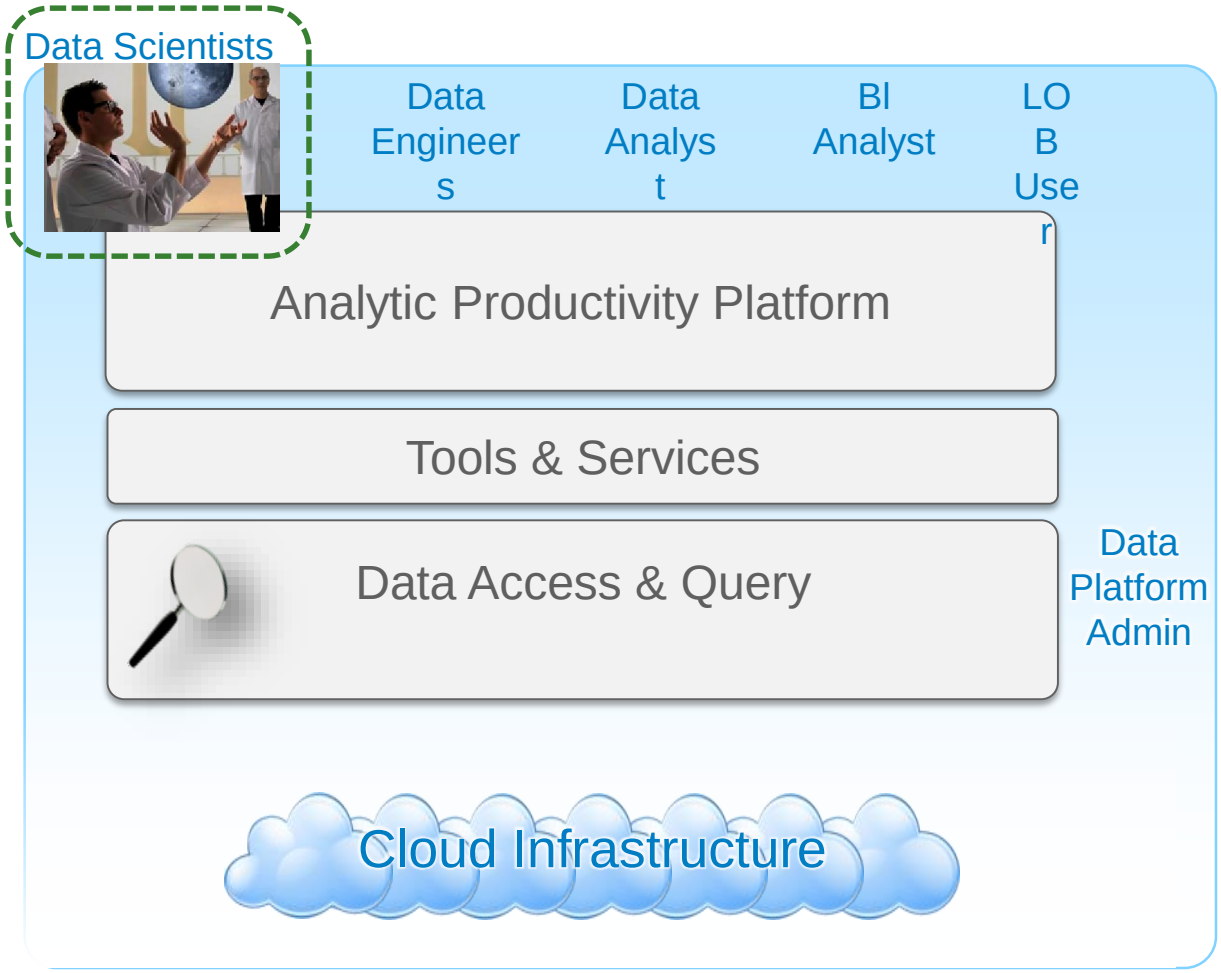
Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

Roles Needed for Analytical Projects

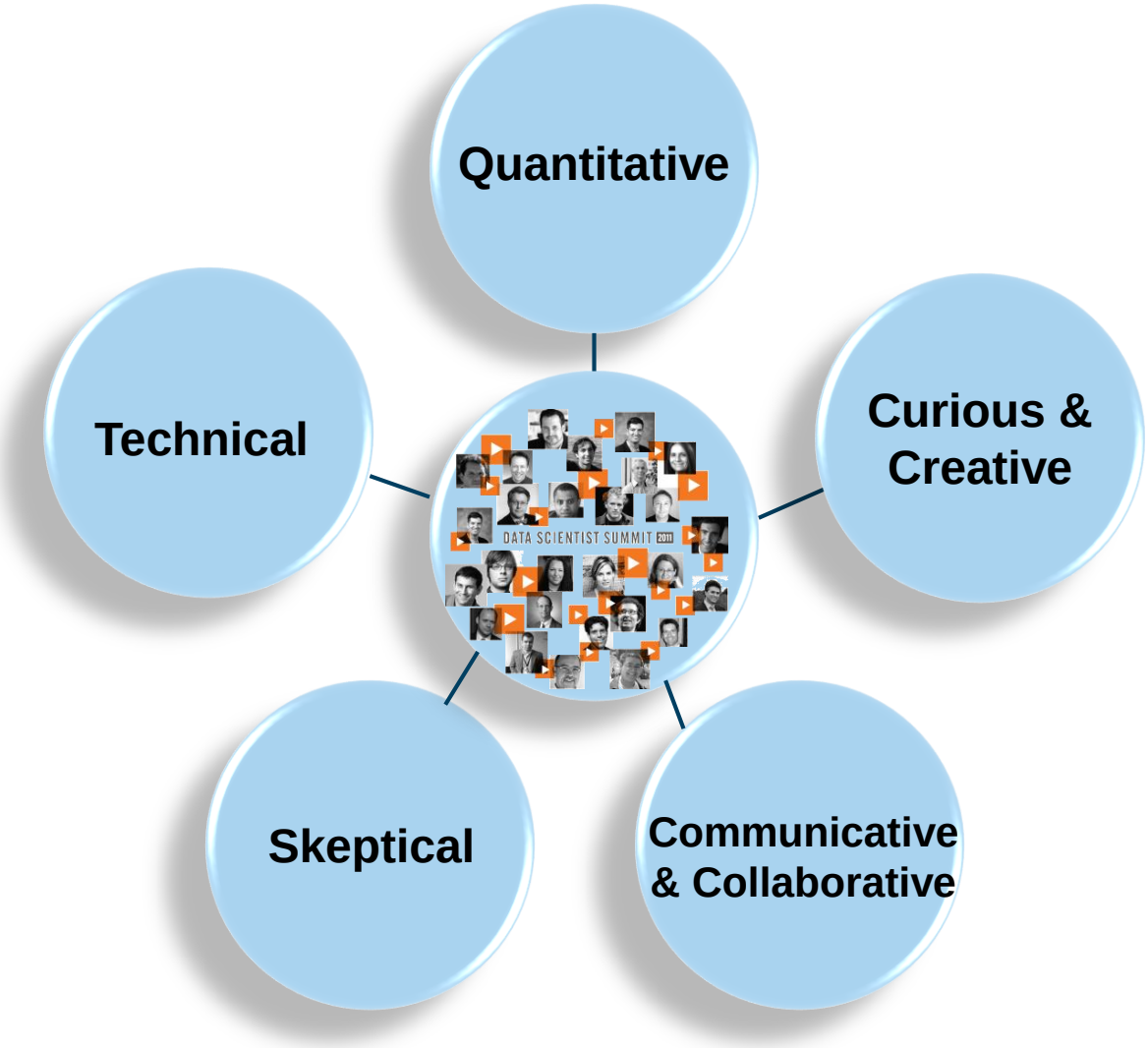
Data Scientist *Key Activities*

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



Module 2: Introduction to Big Data Analytics

Lesson 3: Summary

During this lesson the following topics were covered:

- Key Roles of the New Big Data Ecosystem
- Profile of a Data Scientist

Module 2: Introduction to Big Data Analytics

Lesson 4: Big Data Analytics in Industry Verticals

During this lesson we cover the following representative examples:

- Health Care
- Public Services
- Life Sciences
- IT Infrastructure
- Online Services

Big Data Analytics: Industry Examples

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

1 Big Data Analytics: *Healthcare*

Situation

- Poor police response and problems with medical care, triggered by shooting of a Rutgers student
- The event drove local doctor to map crime data and examine local health care

Use of Big Data

- Dr. Jeffrey Brenner generated his own crime maps from medical billing records of 3 hospitals

Key Outcomes

- City hospitals & ER's provided expensive care, low quality care
- Reduced hospital costs by 56% by realizing that 80% of city's medical costs came from 13% of its residents, mainly low-income or elderly
- Now offers preventative care over the phone or through home visits

2 Big Data Analytics: *Public Services*

Situation

- Threat of global pandemics has increased exponentially
- Pandemics spreads at faster rates, more resistant to antibiotics

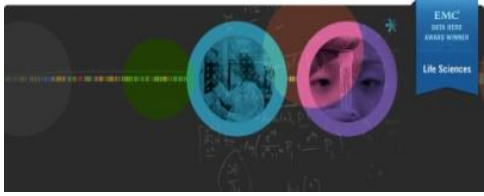
Use of Big Data

- 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 predications 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

3 Big Data Analytics: *Life Sciences*



Situation

- Broad Institute (MIT & Harvard) mapping the Human Genome

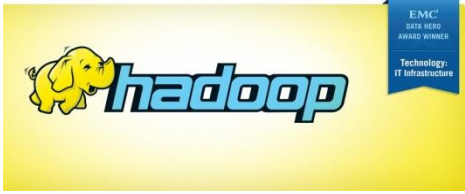
Use of Big Data

- In 13 yrs, mapped 3 billion genetic base pairs; 8 petabytes
- Developed 30+ software packages, now shared publicly, along with the genomic data

Key Outcomes

- Using genetic mappings to identify cellular mutations causing cancer and other serious diseases
- Innovating how genomic research informs new pharmaceutical drugs

4 Big Data Analytics: *IT Infrastructure*



Situation

- Explosion of unstructured data required new technology to analyze quickly, and efficiently

Use of Big Data

- Doug Cutting created Hadoop to divide large processing tasks into smaller tasks across many computers
- Analyzes social media data generated by hundreds of thousands of users

Key Outcomes

- New York Times used Hadoop to transform its entire public archive, from 1851 to 1922, into 11 million PDF files in 24 hrs
- Applications range from social media, sentiment analysis, wartime chatter, natural language processing

5 Big Data Analytics: *Online Services*

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

Module 2: Introduction to Big Data Analytics

Lesson 4: Summary

During this lesson the following representative examples were covered:

- Health Care
- Public Services
- Life Sciences
- IT Infrastructure
- Online Services

Check Your Knowledge

Take quiz in Blackboard

Module 2: Summary

Key points covered in this module:

- Big data was defined
- Four business drivers for advanced analytics were identified
- The techniques for Business Intelligence were distinguished from those of Data Science
- The role of the Data Scientist within the new big data ecosystem was described
- Multiple illustrative examples of big data opportunities were cited