CSCI446/946 Big Data Analytics

Week 1 Introduction to Big Data Analytics

School of Computing and Information Technology
University of Wollongong Australia

Introduction to Big Data Analytics

- Big Data Overview
- State of the practice in Analytics
- Key Roles for the New Big Data Ecosystem
- Examples of Big Data Analytics
  - See more details in Chapter 1 of *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, EMC Education Services (Editor)

Big Data Overview

- What’s your idea on Big Data?
- What’s driving data deluge?
  - Can you name a source of big data?

What’s Driving Data Deluge? Any more?

Mobile Sensors
Social Media
Video Surveillance
Video Rendering
Smart Grids
Geophysical Exploration
Medical Imaging
Gene Sequencing

Big Data Overview

- Keeping up with this high influx of data is difficult.
- Analysing vast amounts of data is more challenging, especially when the data does not conform to traditional structure.
- Can you name any real applications of Big Data Analytics you have been aware of?

Big Data Overview

- Four attributes define Big Data
  1. Volume of data.
  2. Variability and complexity of data types and data structures.
  3. Speed of new data creation and growth.
  4. Data quality, reliability (accuracy and truthfulness).
- 4V: Volume, Variety, Velocity and Veracity
- 5V: 4V + Value

Big Data Overview

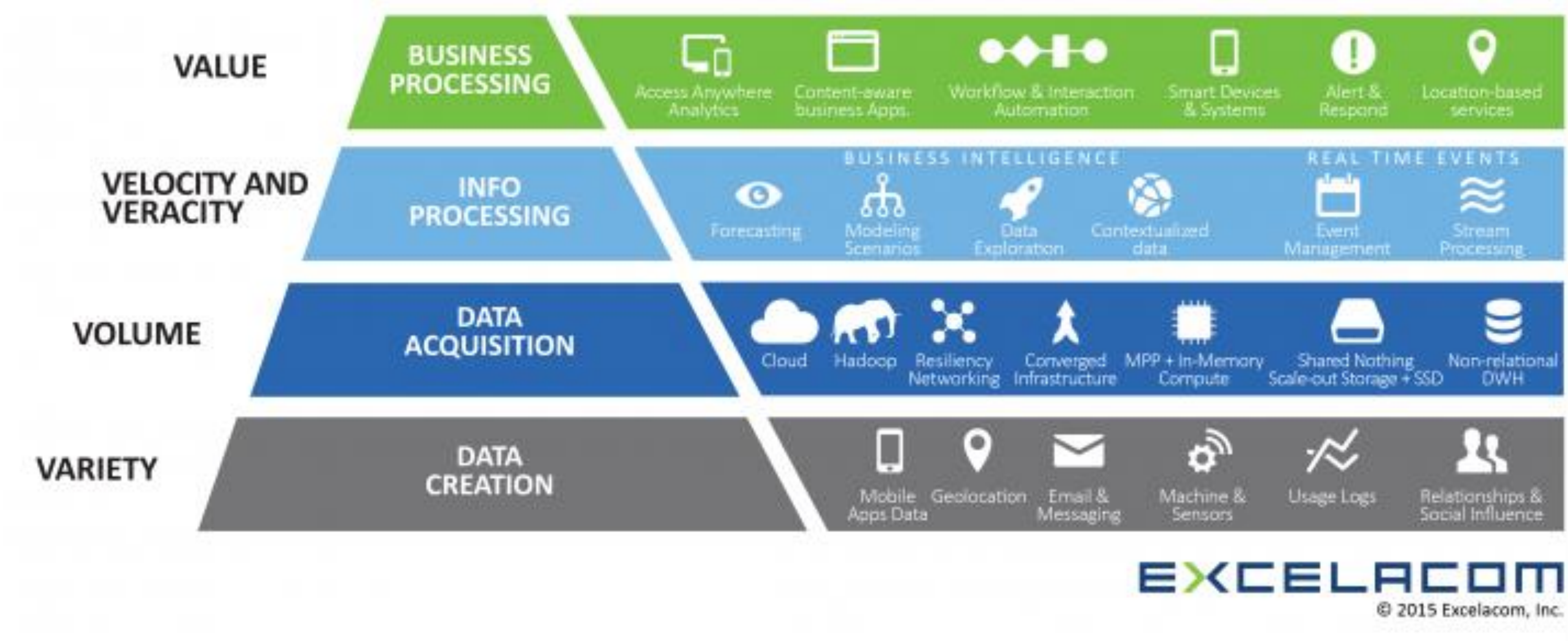
Characteristic differences of data in Data Mining and Big Data:

Data Mining	**Big Data**
• Large datasets* • Closed (fixed) datasets • Data from a known source • Data tends to be more reliable • Data type and structure is fixed.	• Large datasets* • Open ended data (data keeps coming) • Data come from a variety of sources. • Data quality tends to vary • Data type and structure can vary.

* The “size” property is relative to a domain or application.

Big Data Overview

- 5V: Volume, Variety, Velocity, Value and Veracity



From http://www.excelacom.com/resources/blog/the-5-vs-of-big-data-predictions-for-2016

Big Data Overview

- So, Big data analysis needs new tools and technologies

- *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*** *source of business value*

  - McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity, 2011

Big Data Overview

- This implies the need of
  - New data architectures
  - New analytic sandboxes
  - New tools
  - New analytical methods
  - An integration of multiple skills
  - New role of data scientist?

Big Data Overview

- Big Data aims at automating the processes as much as possible.
- The ultimate aim is to have tools that accept data and then produce valuable responses without user intervention.
  - Many challenges
  - Very active area of research.
  - We are still at the early beginnings.
  - Many unanswered questions.

Big Data Overview

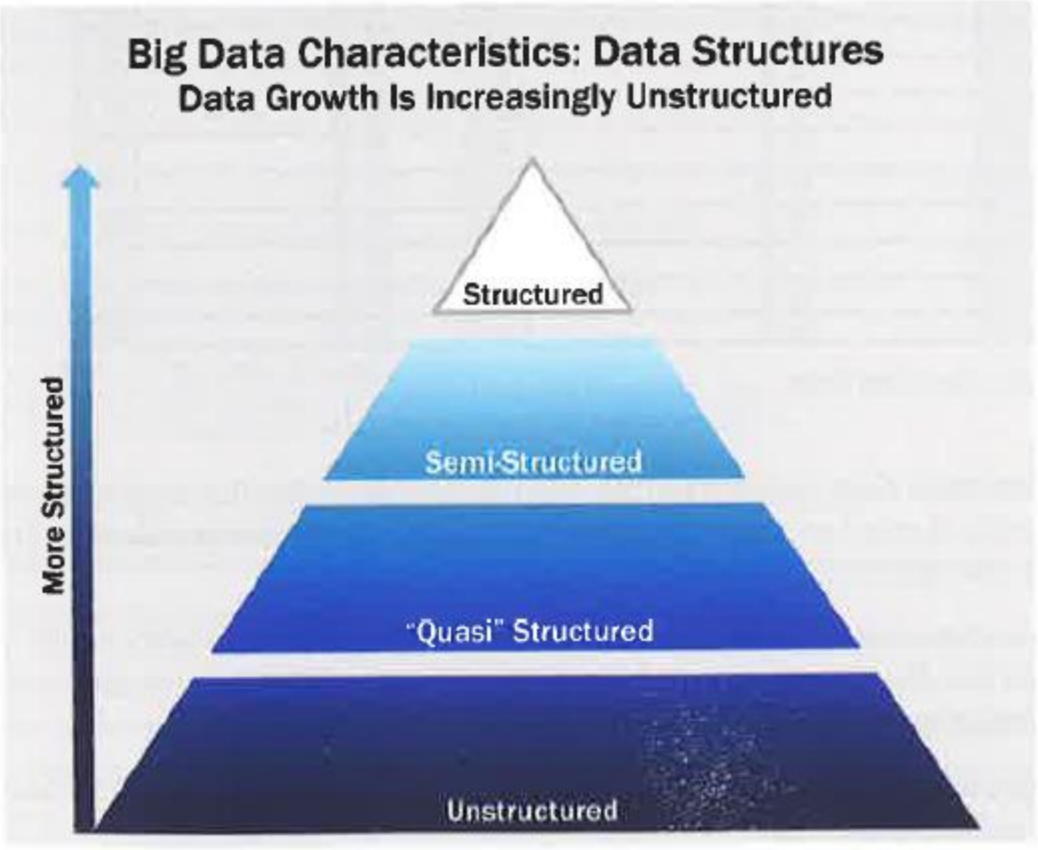
- It is believed that AI holds the key to success.
- Many machine learning algorithms in AI are:
  - Highly scalable methods
  - Relatively insensitive to variations in data quality
  - Enable the machine to solve a problem for us.
- Approach to Big Data is to enable AI methods to:
  - work on data streams
  - work with data from different sources
  - explain results/value

Big Data Overview

- Data Structures:
  1. Structured data
     - Can you name some examples?
  2. Non-structured data (80-90% of data growth)
     - Semi-structured (XML data file)
     - Quasi-structured (Web clickstream data)
     - Unstructured (text documents, images, videos)

Big Data Overview

- Data Structures


Big Data Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

Big Data Overview

- Analyst Perspective on Data Repositories
  - Data accuracy and availability
  - Flexibility and agility of analysis
- Types of data repositories
  - Spreadsheets and data marts
  - Data Warehouses
  - Analytics Sandbox (workspaces)
  - Cloud
- Approach shall fit with the desired goals

State of the Practice in Analytics

- Business drivers for Advanced Analytics
  - Optimise business operations
  - Identify business risk
  - Predict new business opportunities
  - Comply with laws or regulatory requirements
- Leverage advanced analytics to create competitive advantage
- Advanced analytical techniques + Big Data

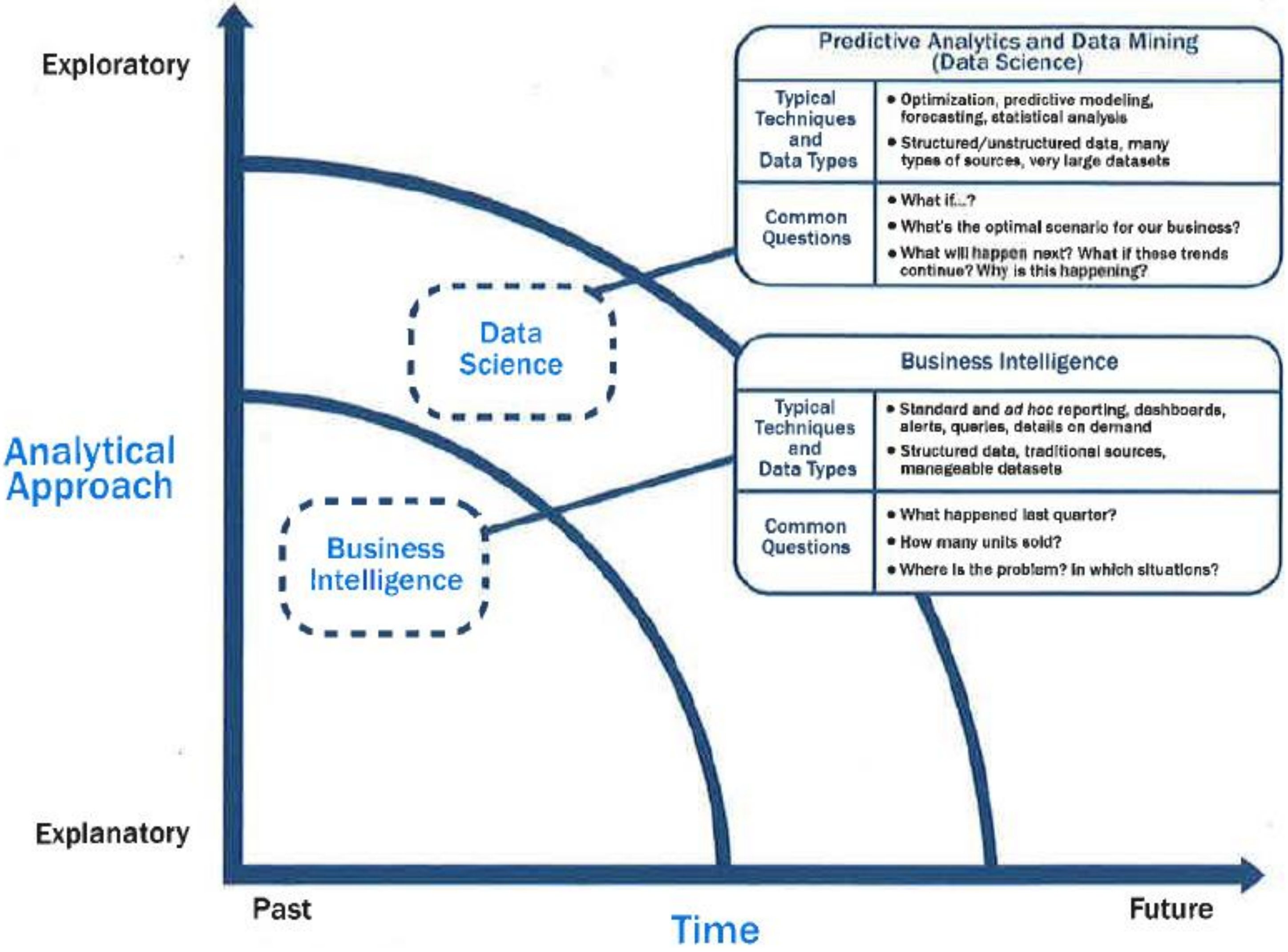
  → More impactful analyses

https://au.linkedin.com/jobs/view/head-of-ai-machine-learning-at-woolworths-supermarkets-746700893

State of the Practice in Analytics

- Business Intelligence vs. Data Science
  - Scope of time,
  - Analytical Approach,
  - Data type,
  - …
- Both analyse data (reflecting the past) to help with making decisions (reflecting the future).
  - What & How have we done in the past?
  - What & How can we do in the future?
- But they differ in scope…

State of the Practice in Analytics

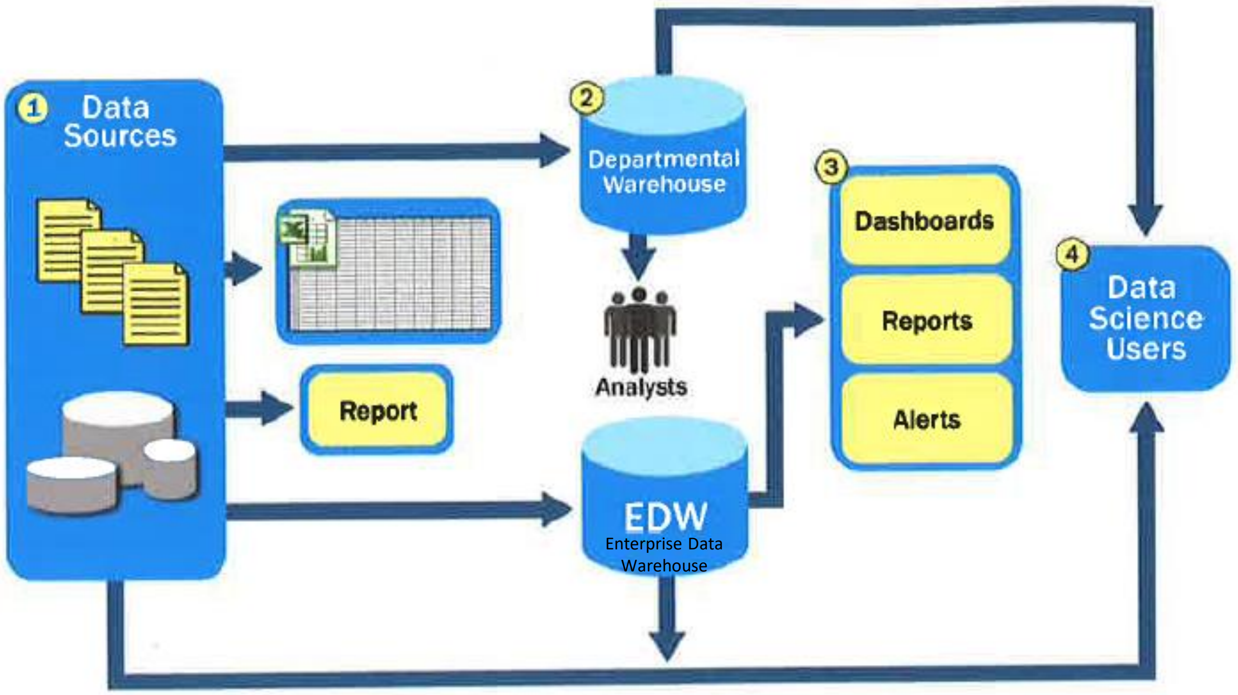


Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our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

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

State of the Practice in Analytics

- Current Analytical Architecture



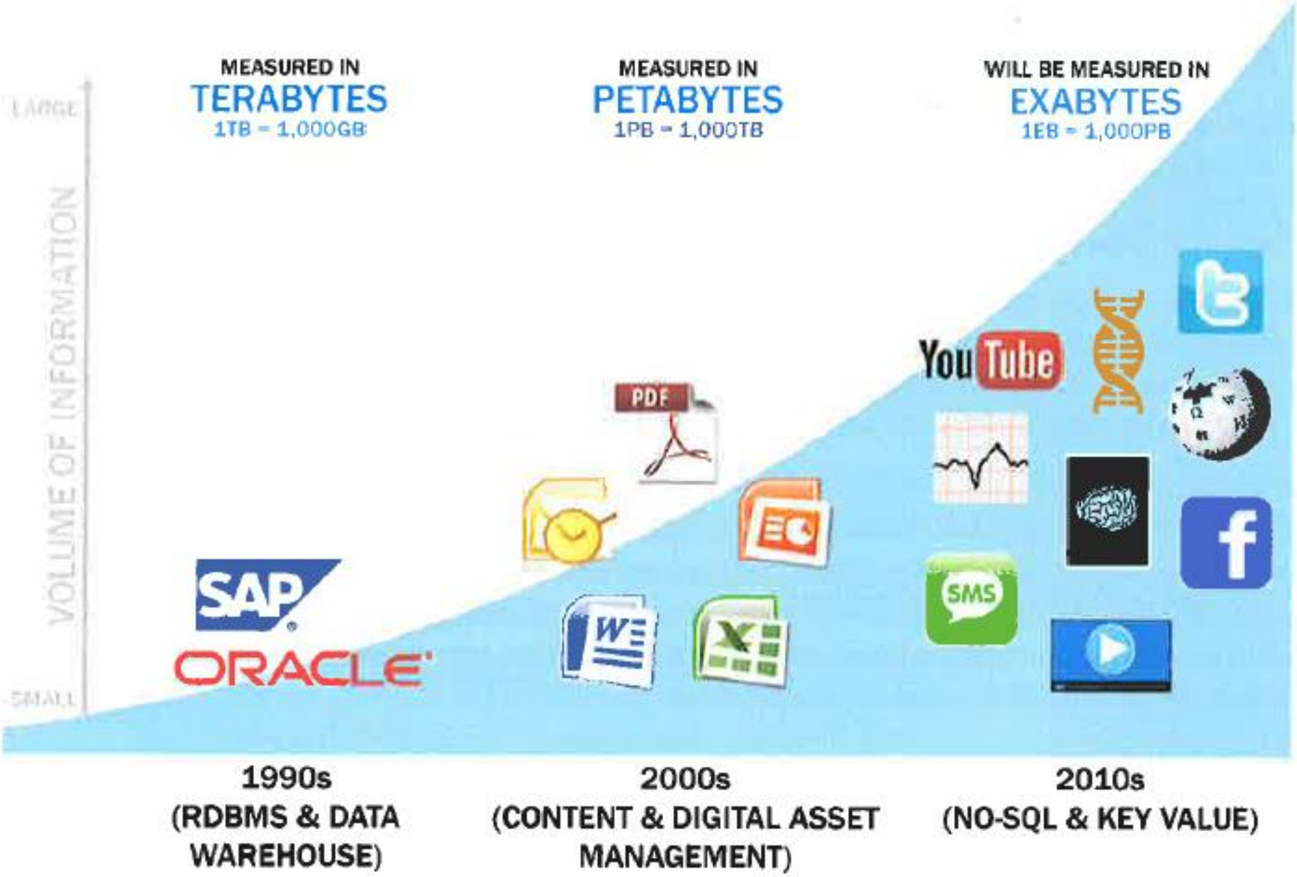
  - Traditional data architectures inhibit data exploration and more sophisticated analysis

State of the Practice in Analytics

- Traditional data architectures have several additional implications for data scientists
  - Predictive analytics and data mining activities are last in the line for data (i.e., low priority)
  - Limited to perform in-memory analytics, restricting the size of the datasets they can use
  - Projects remain isolated and ad hoc, rather than centrally managed. Exist as nonstandard initiatives
- One solution: analytic sandboxes

State of the Practice in Analytics

- Drivers of Big Data



State of the Practice in Analytics

- Emerging Big Data Ecosystem & a New Approach to Analytics
  - Data → intrinsic value → a new economy
  - Data vendors, data cleaners
  - Repackaging and simplifying open source tools
  - Data is the king!

https://www.linkedin.com/pulse/data-king-one-who-masters-rule-world-so-you-ready-vinay-rao-mle-

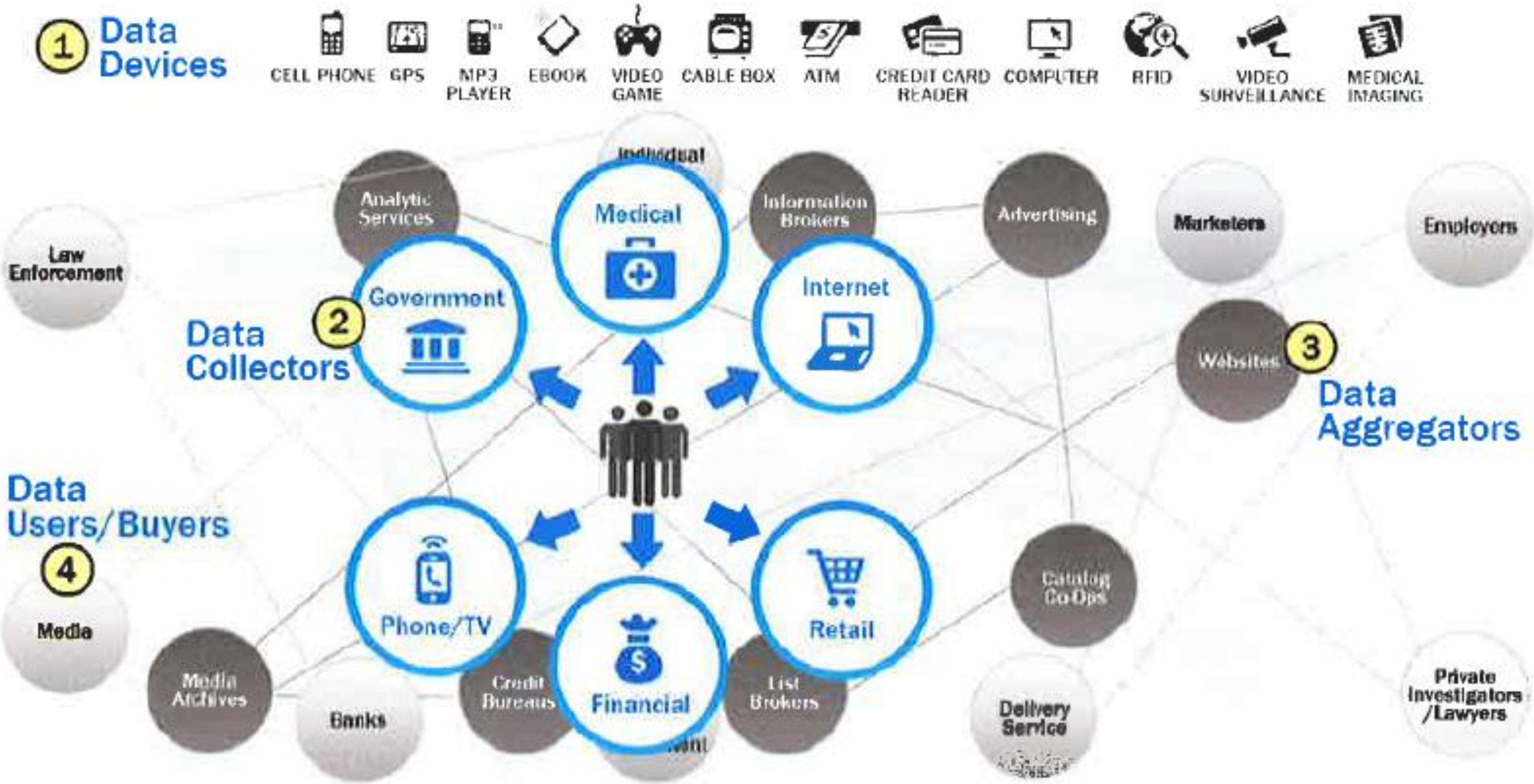
State of the Practice in Analytics

- How big is Big Data?
  - Is there a size requirement on the data?
  - Is there a threshold value on the minimum size of the amount of data?
- Answer depends on the domain.
- Example: Youtube vs. climate modelling.
  - Both create a continuous stream of data.
  - The rate by which data is created differs significantly.
- Big Data does not necessarily imply that TB of data need to be processed at a given time.
  - We may only need to process a few KB in some domains.

State of the Practice in Analytics

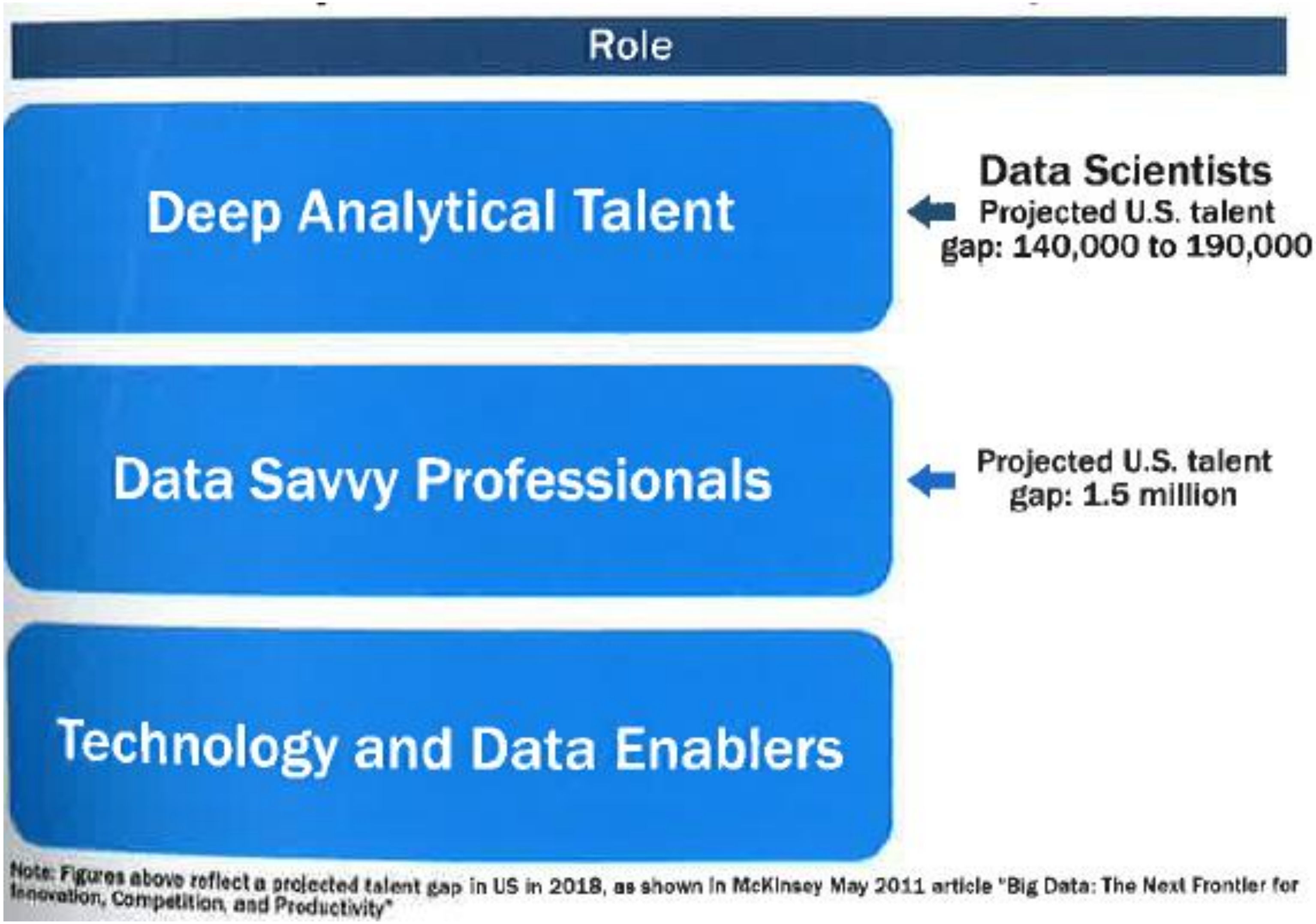
- Four main groups of players here
  - Data devices
    - Video game, Smartphone, Retail shopping card
  - Data collectors
    - Service providers, shopping cart with RFID chips
  - Data aggregators
    - Compile, transform and package data to sell
  - Data users and buyers
    - Retail banks, common people
- Each with commercial interests.

State of the Practice in Analytics



- So, Big Data problems and projects require new approach to succeed

Key roles for the New Ecosystem



Role

Deep Analytical Talent

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

Data Savvy Professionals

Projected U.S. talent
gap: 1.5 million

Technology and Data Enablers

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"

Key roles for the New Ecosystem

- Data Analytical Talent (Data Scientist)
  - Advanced training in mathematics, statistics, and machine learning
  - Newest role, least understood
- Data Savvy Professionals
  - Less technical depth but can define key questions
- Technology and Data Enablers
  - Support data analytical projects
- These three groups must work together

Key roles for the New Ecosystem

- What do data scientists do?
  - Reframe business challenges to analytical challenges
  - Design, implement, and deploy statistical models and data mining techniques on Big Data
    - This is mainly what people think about them
  - Develop insights that lead to actionable recommendations to derive new business value

Examples of Big Data Analytics

- Some examples
  - US retailer Target

    - Infer Marriage, Divorce, and Pregnancy
    - Manage its inventory correspondingly
  - IT Infrastructure
    - Apache Hadoop
    - Process vast amount of information in parallel.
  - Social media
    - Leverage social interactions to derive new insights.

Summary

- Big Data comes from myriad of sources.
- Big Data addresses business needs and solves complex problems.
- Companies and organisations move toward Data Science.
- Require new architectures, new ways of working, new skill sets, new roles, etc.
- A growing talent gap.

Questions for you

- What are the four (or five) characteristics of Big Data?
- What is an analytic sandbox, and why is it important?
- Explain the difference between BI and Data Science.
- Describe the challenges of the current analytical architecture for data scientists.
- What are the key skills and characteristics of a data scientist?
- How much data is involved in big Data?

Q&A

Images Courtesy of Google Image