CSCI446/946 Big Data Analytics

Week 1 – Introduction to Big Data Analytics

School of Computing and Information Technology
University of Wollongong Australia
Spring 2024

Prof Wanqing Li
Email: wanqing@uow.edu.au
Office: 3-101

Content

- Big Data Overview
- State of the practice in Analytics
- Key Roles for and in the 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 is your idea about Big Data?

Big Data Overview

- What's driving data deluge?
  - Would you know some sources of big data?

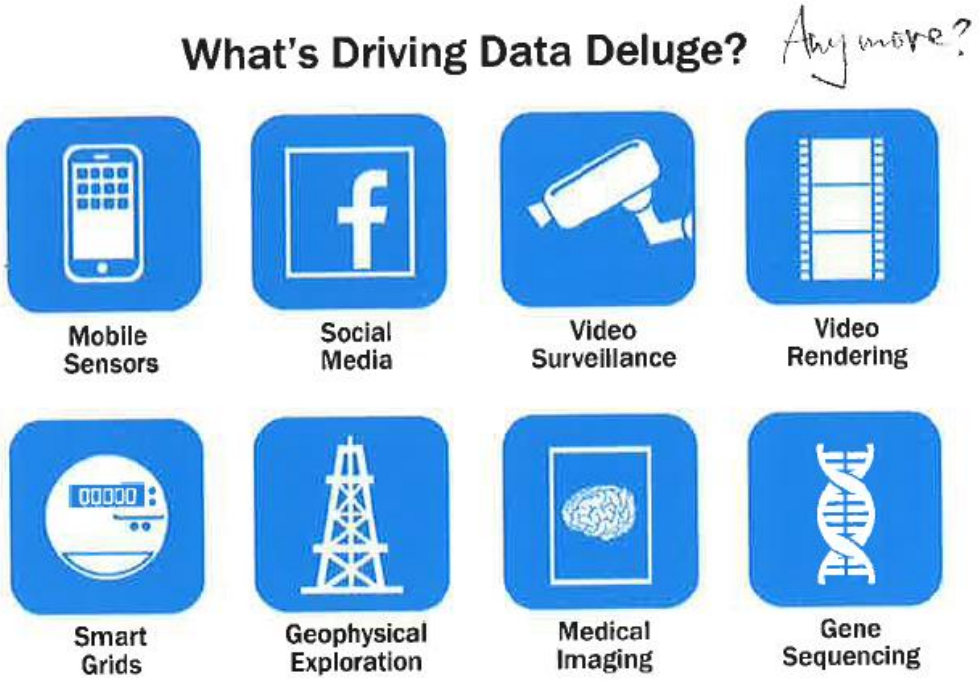


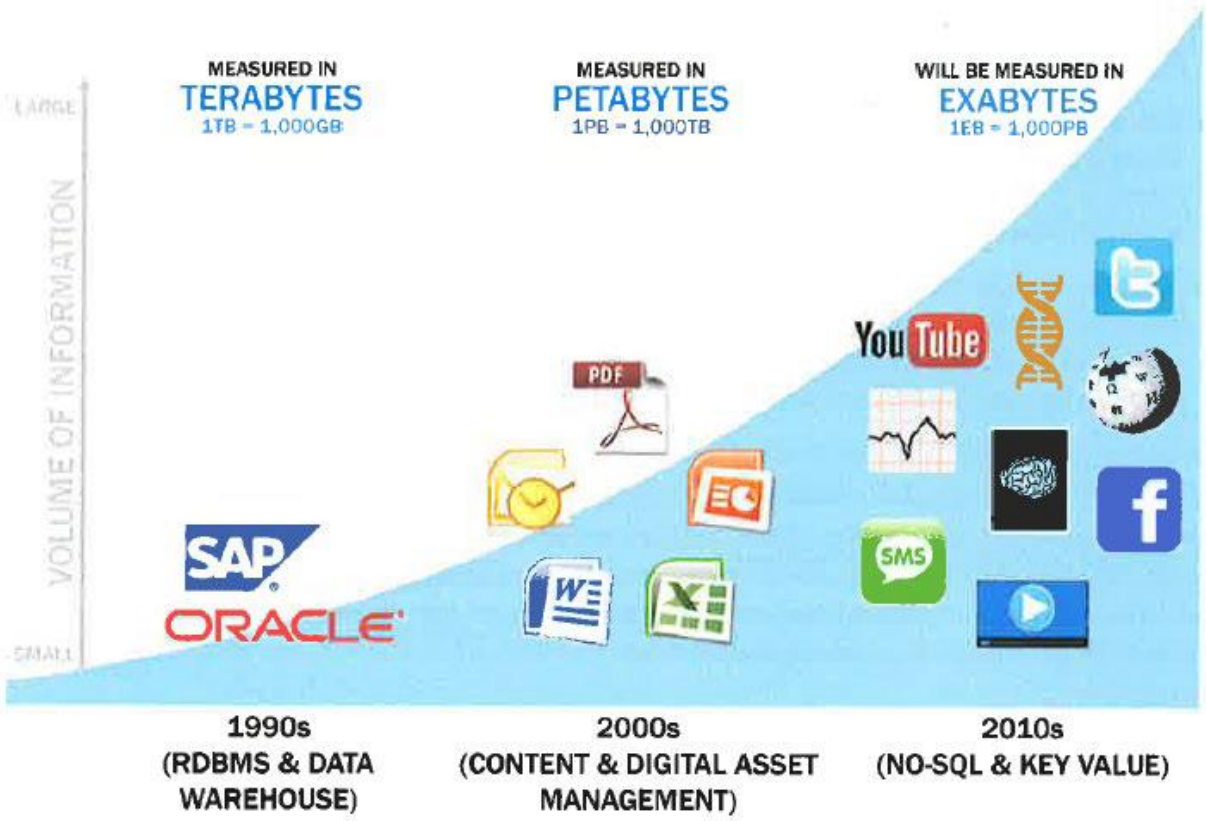
Figure source: Book - Data Science and Big Data Analytics Chapter1 Figure 1.1

Deluge: *noun*

a great flood of water; a drenching rain; anything that overwhelms like a flood

Big Data Overview

- Drivers of Big Data



Big Data Overview

- When is data “Big”?
  - 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. air temperature 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.
  - The term “Big” in Big Data is often misrepresented in the media.

Big Data Overview

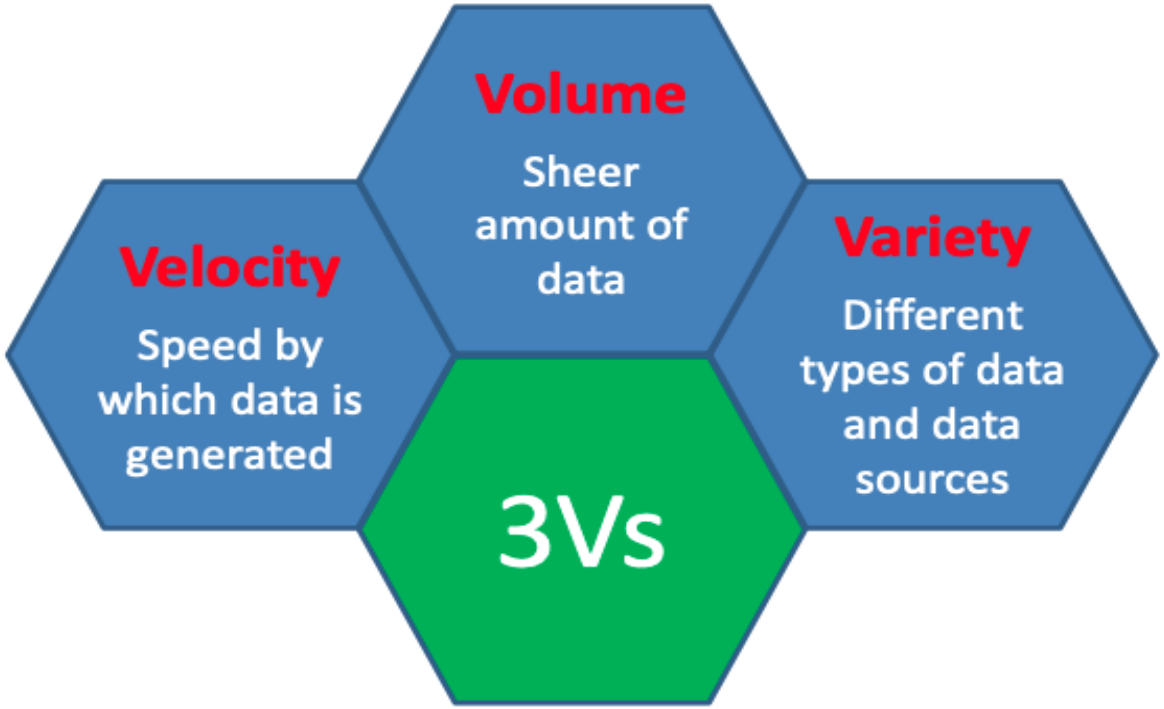
- Keeping up with this increasingly high influx of data is difficult.
- Analysing amounts of data in real time is more challenging, especially when the data does not conform to traditional structure.
- Can you name any real applications of Big Data Analytics you have been aware of?

Social media case

- Fast
- Large-scale
- Different data

- more?

Properties of Big Data

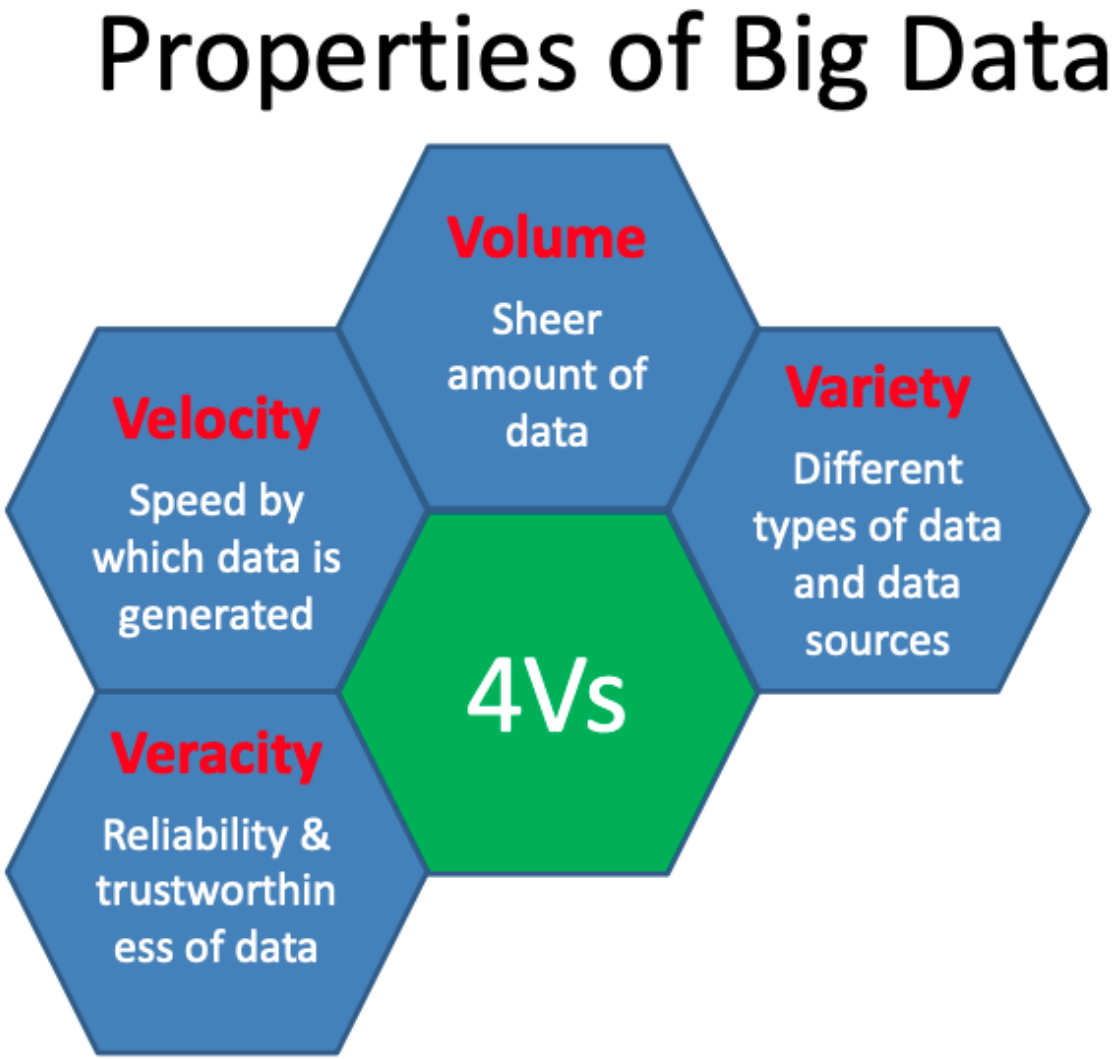


In 2001 MetaGroup (now Gartner) associated three key properties of emerging data. The term "Big Data" was eventually coined in 2005 (by Roger Mougalas).

Social media case

- Fast
- Large-scale
- Different data
- Real data

- more?



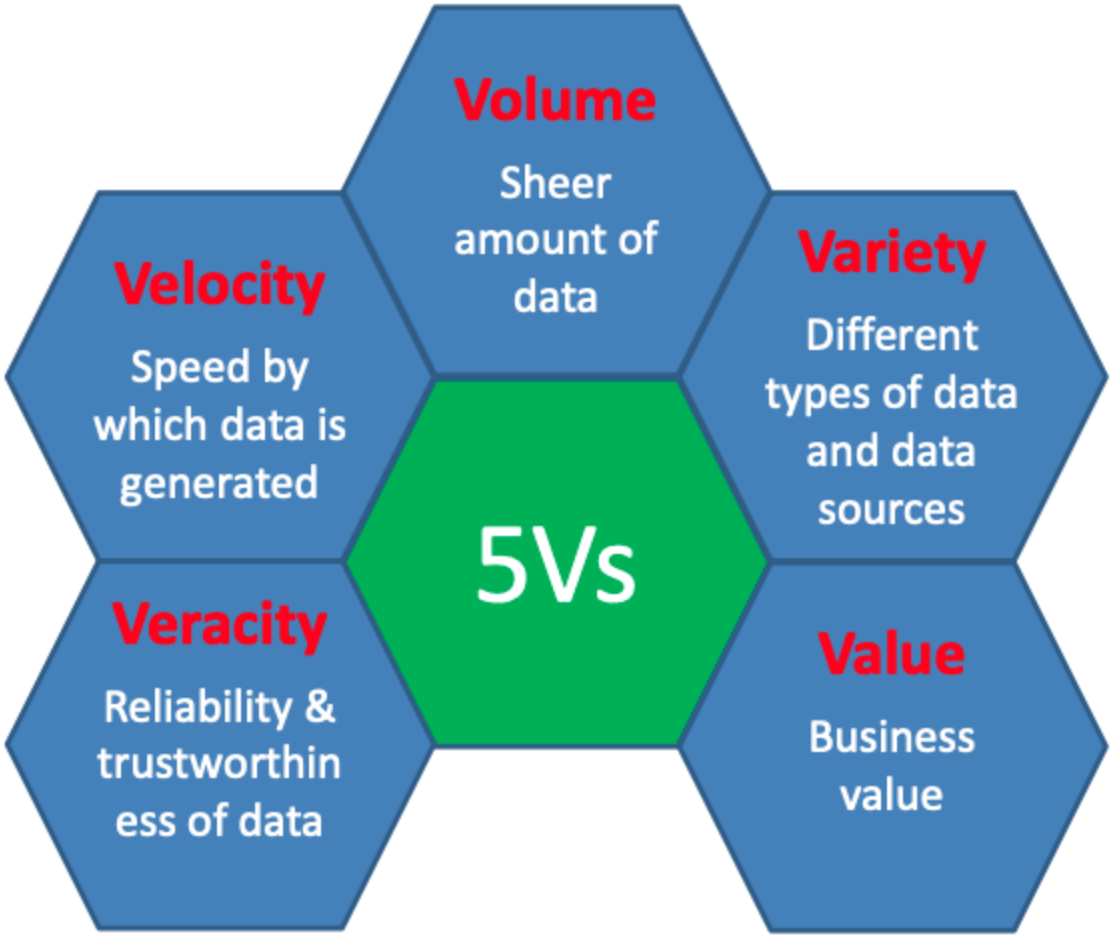
IBM expanded this by a fourth property: Veracity

Social media case

- Fast
- Large-scale
- Different data
- Real data
- Valuable

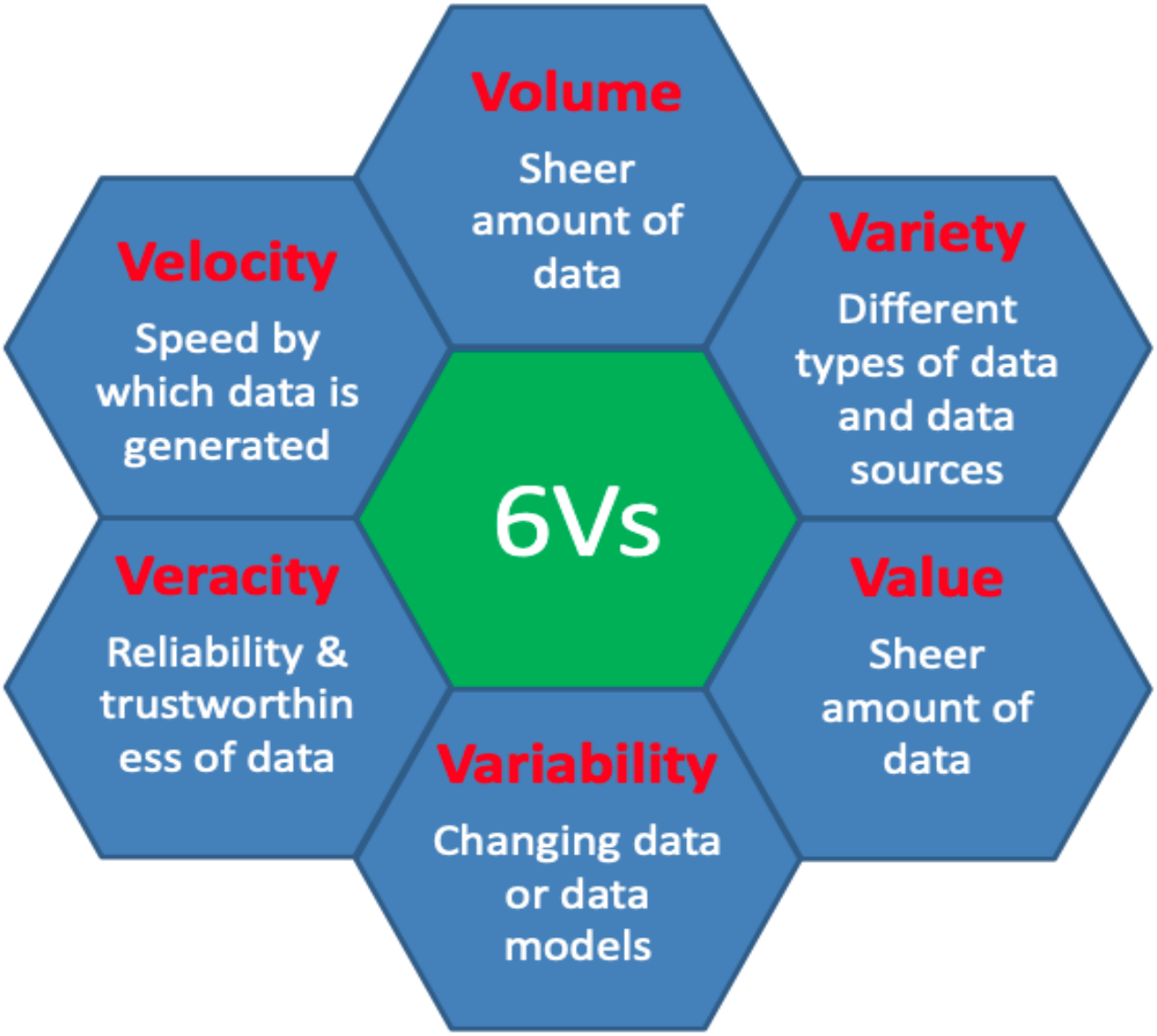
- more?

Properties of Big Data



Over time the list of properties grew…

Properties of Big Data



Properties of Big Data

- The increase in the list of properties make the Big Data ecosystem more complex and more challenging to solve.
  - The Big Data ecosystem is not fixed; it changes over time.

- Note: Gaining value from data is the main objective of Big Data analytics as opposed to the other Vs which are a property of the data in Big Data analytics

10Vs


Volume
Size of Data
Validity
Data quality, Governace, Moster Data Management on Massive
Velocity
The Speed at which Data is Generated
Variability
Dynamic, Evolving Behavior in Data Source
Variety
Different type of Data
BigData
Venue
Distributed Heterogeneous Data from Multiple Platforms
Veracity
Data Accuracy
Vocabulary
Data Models, Semantics that describes data Structure
Value
Useful Data
Vagueness
Confusion over Meaning of BigData and Tools used

Properties of Big Data

- The increase in the list of properties

Why ? Changing needs from Big Data.

What's impact ?

make the Big Data ecosystem* more complex and more challenging to solve.

What's benefit ?

Gaining value from data is the main objective of Big Data analytics.

* Big data ecosystem is **the comprehension of massive functional components with various enabling tools**.

Big Data Overview

- Data Mining and Big Data is related but not the same.
- Characteristic differences of data used for Data Mining and for Big Data:

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

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

Big Data Overview

- There are many well established tools for Data Mining.
- Big Data analytics 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 extract 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 appears to imply the need of:
  - New data architectures
  - New data management tools.
  - New analytic sandboxes.
  - New data processing tools.
  - New analytical methods.
  - An integration of multiple skills.
  - New expertise?
  - New role of data scientist?
  - …

Big Data Overview

- Big Data Analytics aims at:
  - Extracting value from data
  - 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 stages.
  - Many unanswered questions.

Approaches to Big Data Analytics

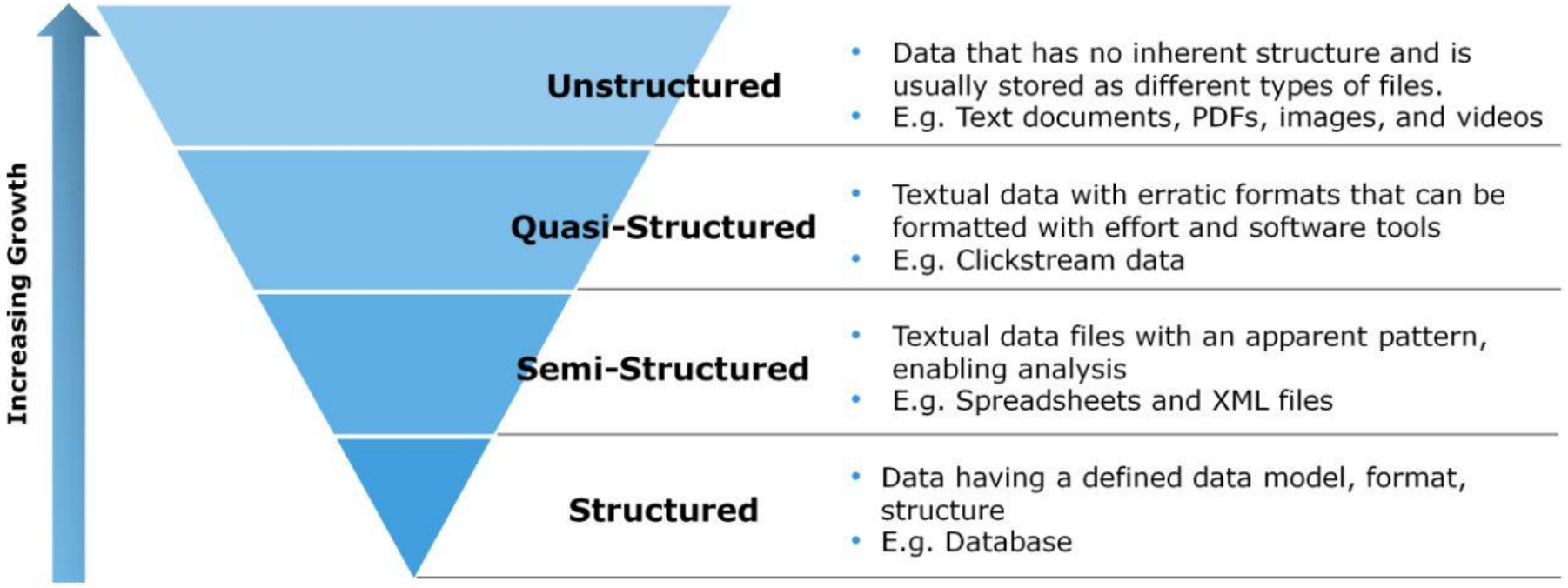
- It is believed that AI holds the key to success.
- Many machine learning algorithms in AI are:
  - Highly scalable methods
  - Relatively insensitive to variations in data quality
  - Enable the machine to solve a problem for us.
- Approach to Big Data is to enable AI methods to:
  - work on data streams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

Structures of Big Data

- Structured data
  - Can you name some examples?

- Non-structured data (80-90% of data growth)
  - Semi-structured (XML data file,…)
  - Quasi-structured (Web clickstream data,…)
  - Unstructured (text documents, images, videos, audio,…)

Structures of Big Data



Big Data Analytics may take all data structures

Detailed definitions & a similar Figure in Book - Data Science and Big Data Analytics Chapter1

Data Repositories

- Analytic Perspective on Data Repositories
  - Data completeness, structure, and accessibility
  - Flexibility and agility of analysis
  - Types of data repositories
    - Spreadsheets
    - Data Warehouses (DW), Enterprise DW, and data marts.
    - Analytics Sandbox(workspaces)
    - Cloud
    - …
- Repository shall be compatible with the desired goals

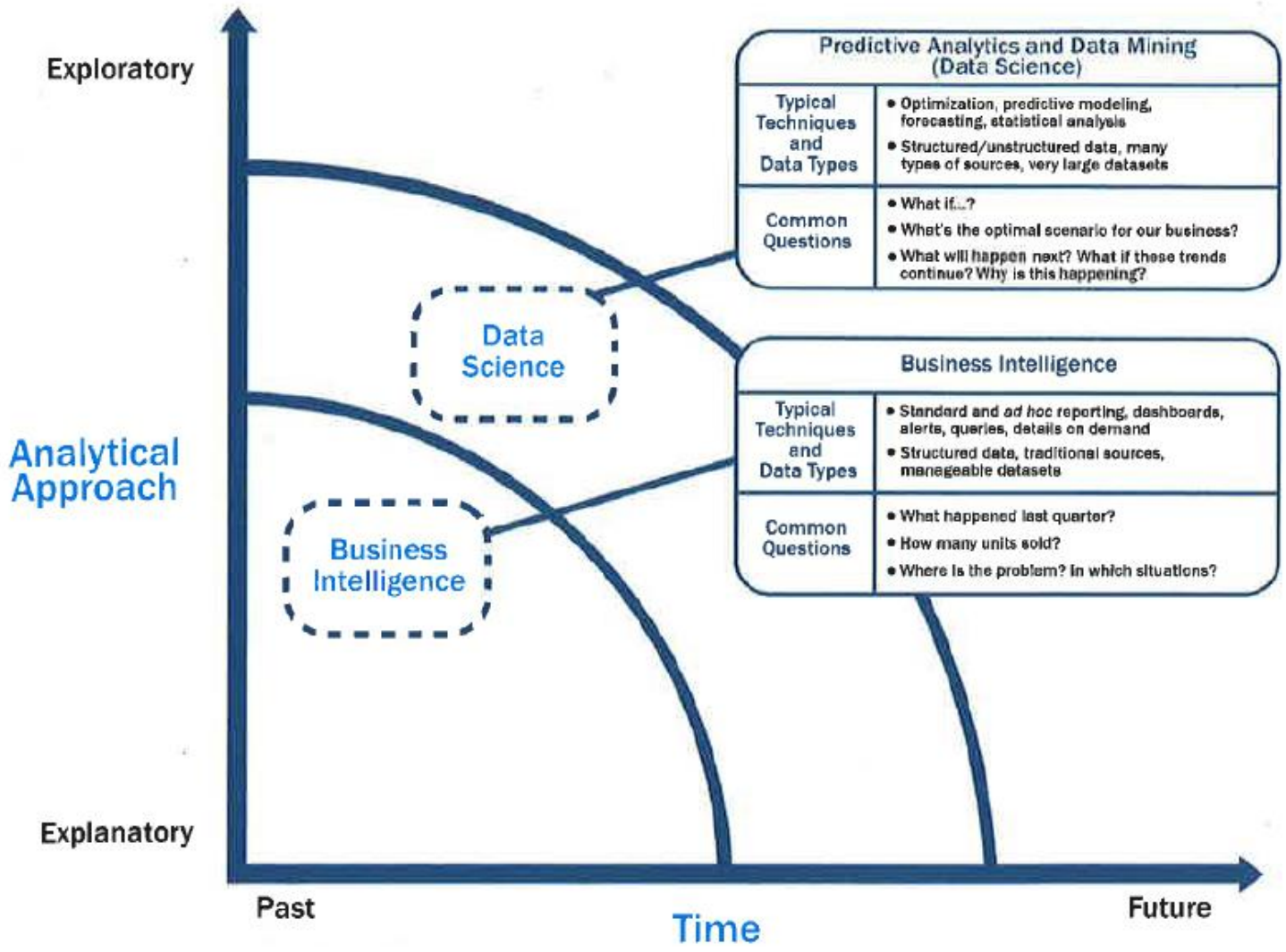
State of the Practice in Analytics

- Business drives for Advanced Analytics
  - Optimise business operations
  - Identify business position and risk
  - Predict new business opportunities
  - Comply with laws or regulatory requirements
  - Provide advanced decision support.
- Leverage advanced analytics to create competitive advantage
- Advanced analytical techniques + Big Data
  - More impactful analyses

State of the Practice in Analytics

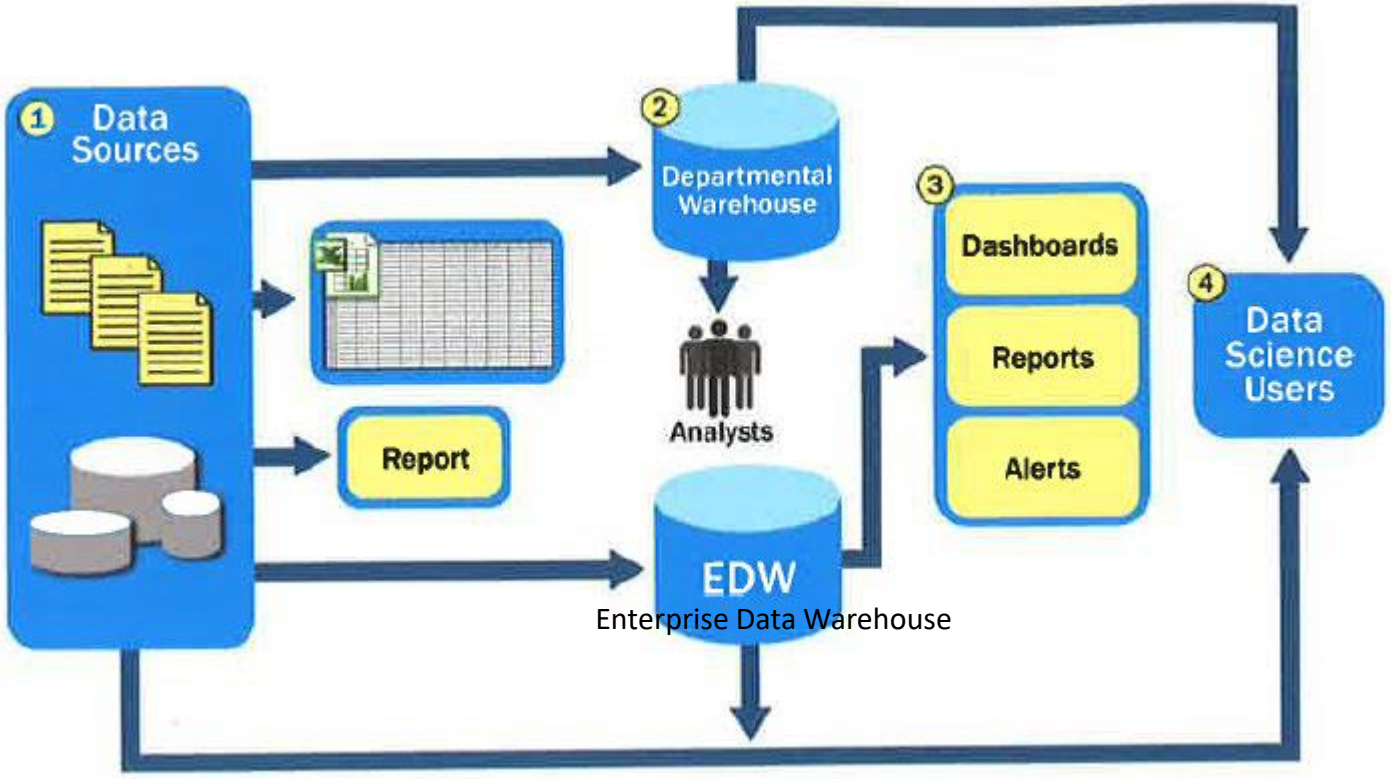
- Business Intelligence vs. Data Science
  - Both analyse data (reflecting the past) to help with making decisions (reflecting the future).
    - What & How have we done in the past? (descriptive)
    - What is the current situation and what led to it? (descriptive)
    - What & How can we do in the future? (predictive)
  - But they differ in scope…

State of the Practice in Analytics



State of the Practice in Analytics

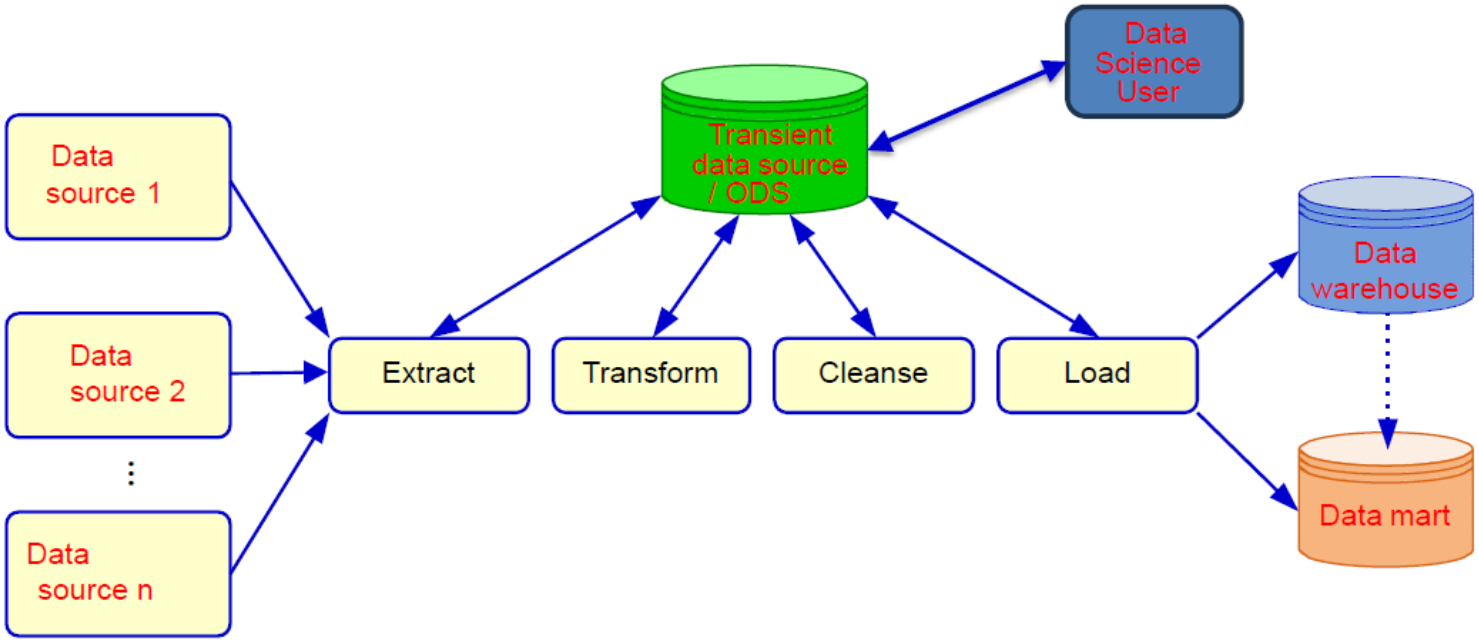
- Typical Analytical Architecture



This data architectures inhibit rapid data access, exploration and more sophisticated analysis.

State of the Practice in Analytics

- Processing high-velocity data needs faster access to data i.e. via the use of a transient data store



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.
  - Analytics takes place in a DW production system.
- One solution: analytic sandboxes

State of the Practice in Analytics

- Emerging Big Data Ecosystem & a New Approach to Analytics
  - Data -> intrinsic value -> a new economy
    - "Data is the new oil"
  - New professions: Data vendors, data cleaners,…
  - New opportunities for software developers:
    - 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

- 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.

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 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, Pregnancy, …
    - Manages its inventory correspondingly
  - IT Infrastructure
    - Apache Hadoop
    - Process vast amount of information in parallel.
  - Social media
    - Leverage social interactions to derive new insights.
- Free economy?
  - Facebook, WhatsApp, Beidu, retail memberships, … can be used free of charge.
    - Are we in a wonderful world where businesses provide services to end users at no cost, make no profit, and pay for all expenses?

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, or six) 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 key skills and roles of a data scientist?
- How much data is involved in Big Data?

Q&A

Images Courtesy of Google Image