TARC
TUNKU ABDUL RAHMAN
UNIVERSITY COLLEGE

BACS3013 Data Science

Chapter 1: Introduction to Data Science and Big Data Analytics

Content

- Big Data and data analytics
- Types of Data Scientists
- Types of analytics
- Analytics process model
- Related Software/Tools
- Data science applications

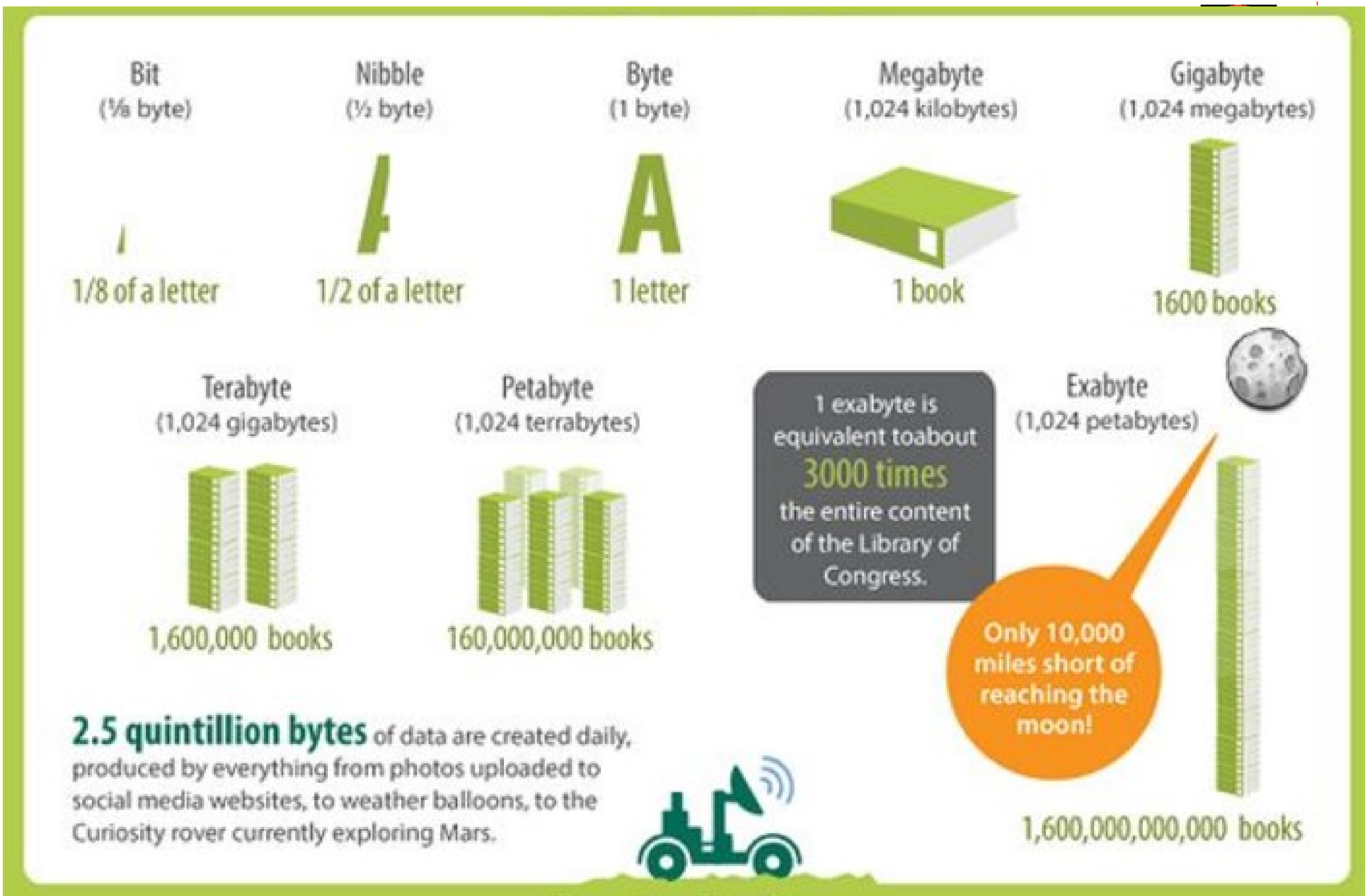
Big Data

How big should it be to be considered as Big Data?

- "Big data" typically refers to data on the scale of terabytes and petabytes .
- The tools of data science are as appropriate for gigabyte as they are for petabyte scale datasets.

https://datascience.berkeley.edu/about/what-is-data-science/

The Size of Data



Bit (⅛ byte)

1/8 of a letter

Nibble (½ byte)

1/2 of a letter

Byte (1 byte)

1 letter

Megabyte (1,024 kilobytes)

1 book

Gigabyte (1,024 megabytes)

1600 books

Terabyte (1,024 gigabytes)

1,600,000 books

Petabyte (1,024 terrabytes)

160,000,000 books

1 exabyte is equivalent toabout 3000 times the entire content of the Library of Congress.

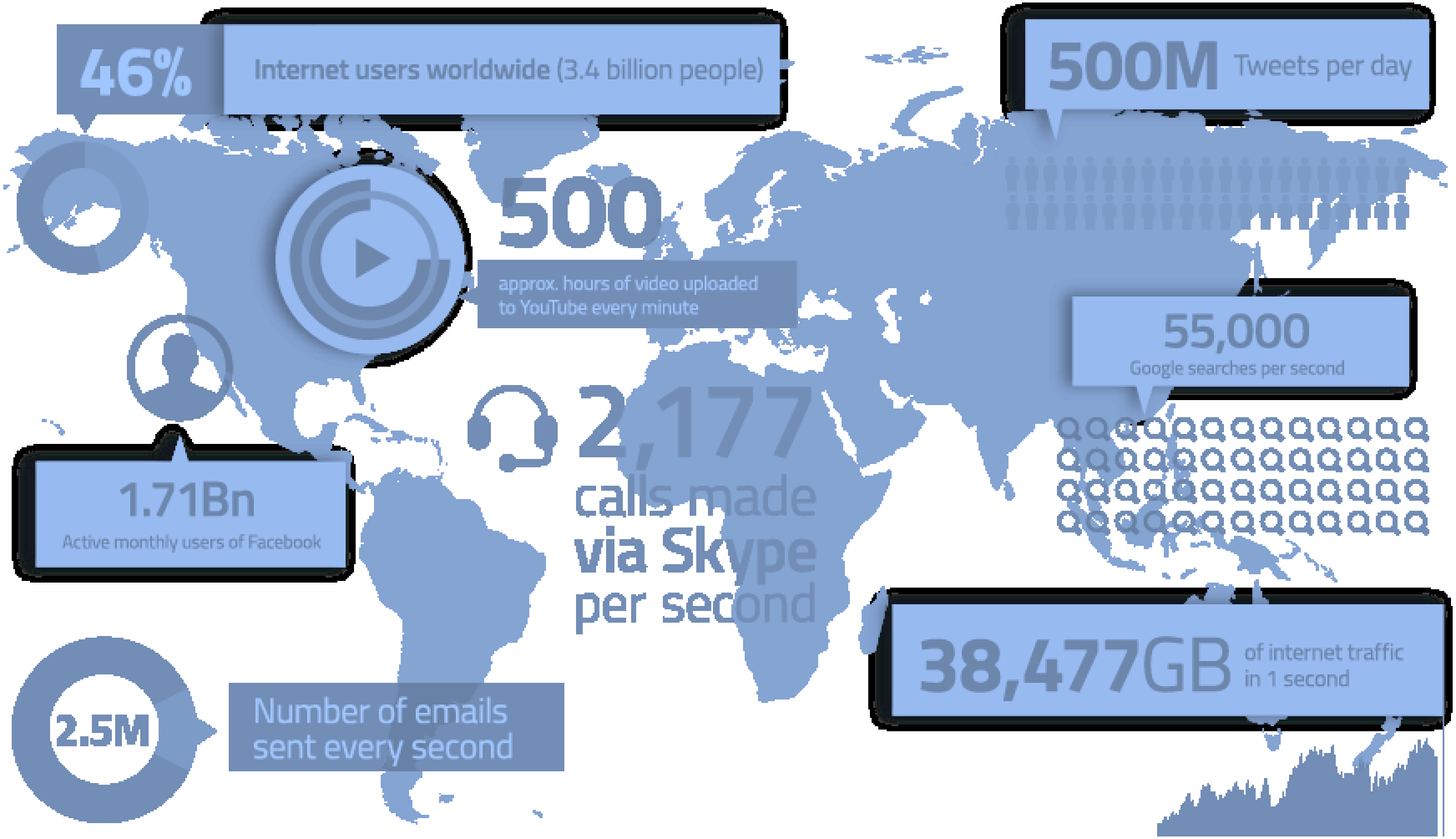
Exabyte (1,024 petabytes)

Only 10,000 miles short of reaching the moon!

1,600,000,000,000 books

2.5 quintillion bytes of data are created daily, produced by everything from photos uploaded to social media websites, to weather balloons, to the Curiosity rover currently exploring Mars.

Big Data Examples



46% Internet users worldwide (3.4 billion people)

500M Tweets per day

500 approx. hours of video uploaded to YouTube every minute

55,000 Google searches per second

1.71Bn Active monthly users of Facebook

2,177 calls made via Skype per second

38,477GB of internet traffic in 1 second

2.5M Number of emails sent every second

WHAT MAKES BIG DATA SO BIG?

90%
of the world's data has been created in the past two years!

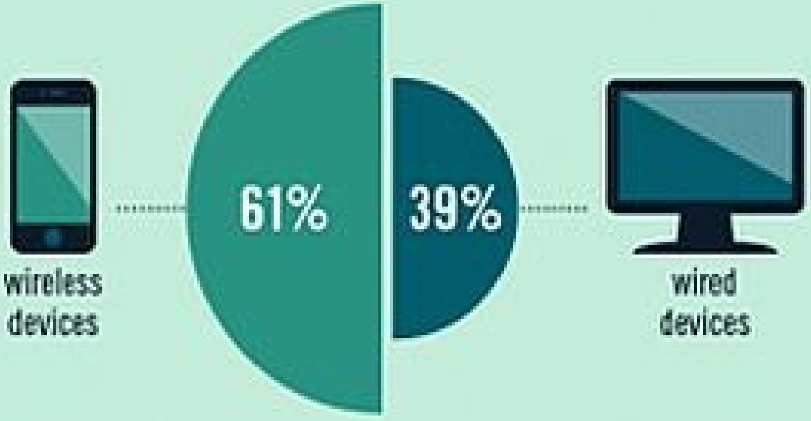
6 BILLION mobile subscriptions worldwide = **87%** of the world's population

1.01 BILLION Facebook users worldwide = **604 MILLION** users log-in monthly from mobile devices

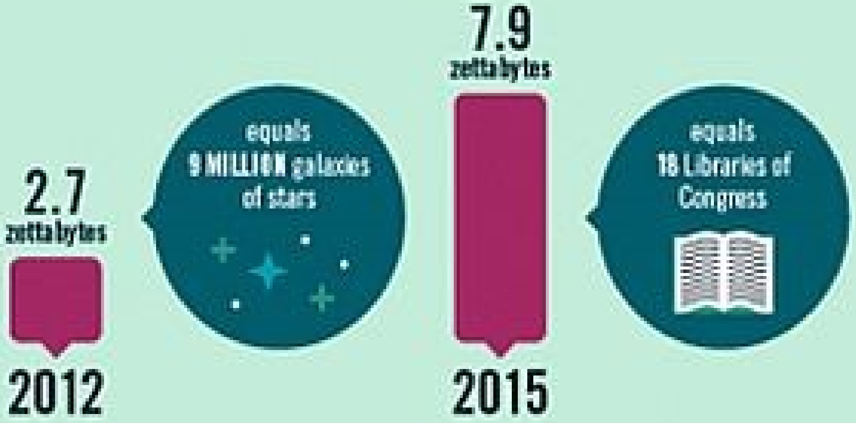
400 MILLION Tweets per day = **84 MILLION** users access Twitter via mobile

And Big Data will only get bigger as traffic from smartphones and tablets outpaces traditional devices.

Percentage of Web Traffic by 2016:



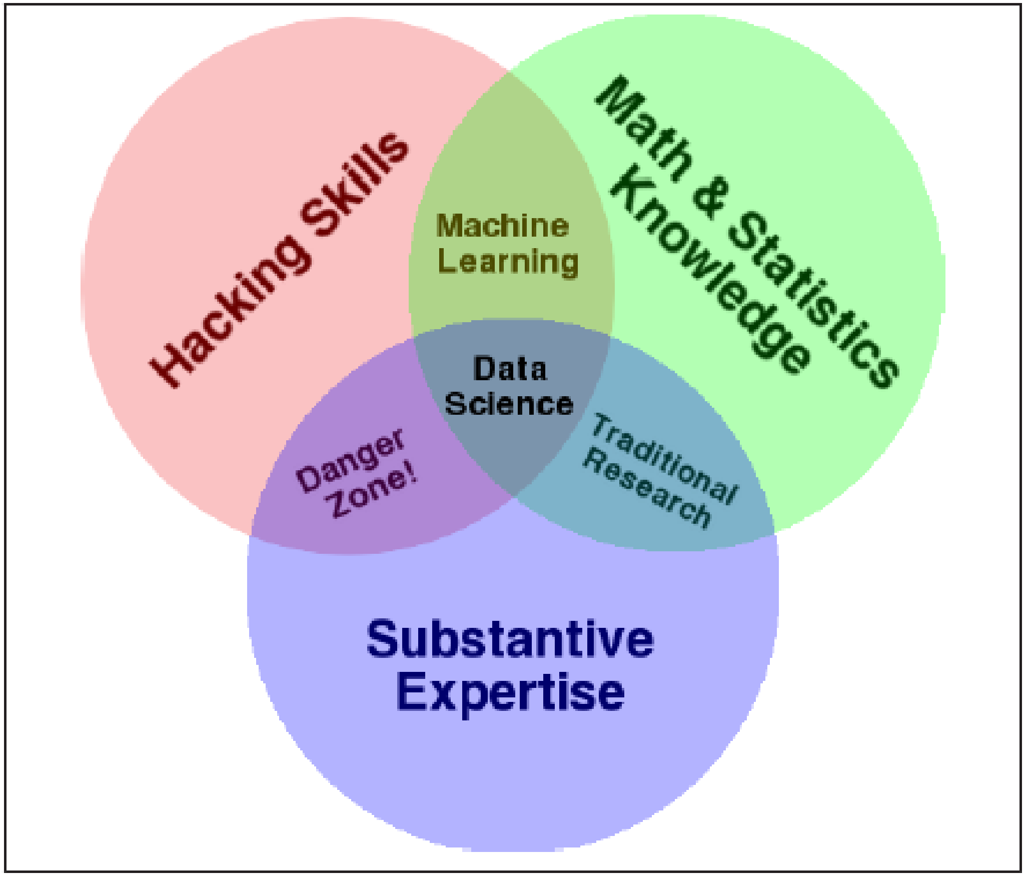
Volume of Digital Content:



Big Data and Analytics

- Gartner projects that by 2015, 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage. About 4.4 million jobs will be created around big data (Baesens et al, 2003).
- A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and "data scientist" talent required to exploit big data.

What is Data Science?



Drew Conway's Venn diagram of data science (2010)

Data Science and Data Analytics

Analytics is a term that is often used interchangeably with data science, data mining, knowledge discovery, and others.

A few definitions of Data Science

- Data science, or data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.
- A concept to unify statistics, data analysis and their related methods in order to "understand and analyze actual phenomena" with data.
- It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the sub-domains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

What Data Scientists do?

- A data scientist is a hybrid statistician, software engineer and social scientist*.
- Expert in computer science, statistics, communication and visualization, and to have extensive domain expertise.
- A scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem.

How many of you could achieve the above alone?

**Social scientist deals with human or user behavior.*

The Roles in a Data Analytics Team

- A team of data scientists may involves roles as follows:

Database or data warehouse administrator (DBA)

- Aware of data available within the firm, the storage details, and the data definitions. Crucial in feeding the analytical modeling exercise with its data.

Business Expert

- Has extensive business experience and business common sense. Helps to steer the analytical modeling exercise and interpret its key findings.

The Roles in a Data Analytics Team

- A team of data scientists may involves roles as follows:

Legal expert

- Given that not all data can be used in an analytical model because of privacy, discrimination, etc. The regulation of such protection is vary depending on geographical region. Legal expert tells what data can be used, when, and what regulation applies in what location.

Data scientist/ data miner/ data analyst

- Do the actual analytics.
- Possess a thorough understanding of all techniques involved and know how to implement them using the appropriate software.
- Have good communication and presentation skills to report the analytical findings to other parties.

The Roles in a Data Analytics Team

- A team of data scientists may involves roles as follows:

Software Tool Vendors

- provide tools to automate specific steps of the analytical modeling process (e.g., data preprocessing).
- Provide software that covers the entire analytical modeling process.
- Provide analytics-based solutions for specific application areas, such as risk management, marketing analytics and campaign management, and so on.

The Roles in a Data Analytics Team

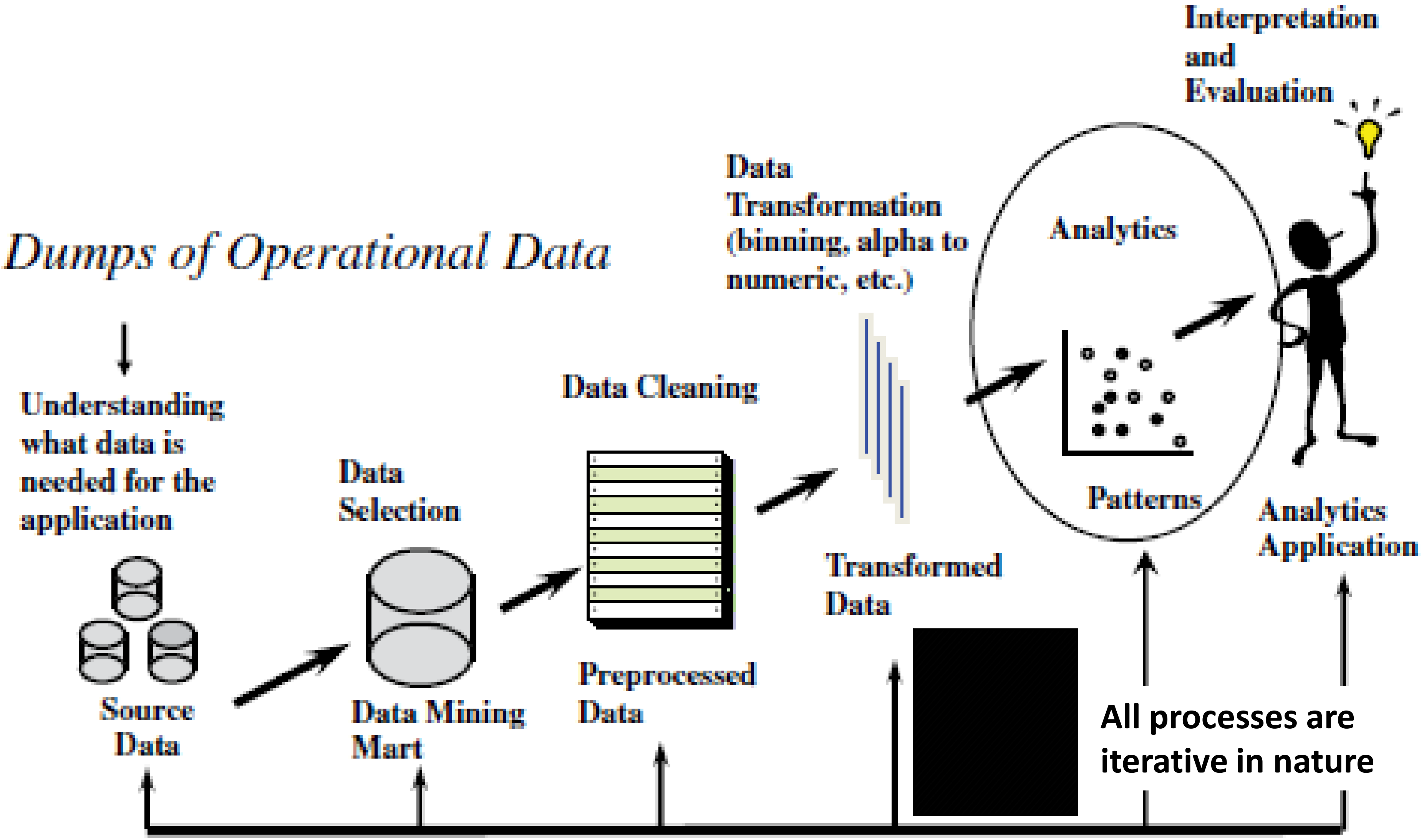
- A team of data scientists may involves roles as follows:

Chief Data Scientist

- setting the data strategy of the company,
- setting everything up from the engineering and infrastructure for collecting data and logging, to privacy
- manage a team of engineers, scientists, and analysts
- communicate with leaders across the company, including the CEO, CTO, and product leaders.
- concerned with patenting innovative solutions and setting research goals.

Types of Analytics

The Analytics Process Model



Dumps of Operational Data

Understanding what data is needed for the application

Source Data

Data Selection

Data Mining Mart

Data Cleaning

Preprocessed Data

Data Transformation (binning, alpha to numeric, etc.)

Transformed Data

[redacted]

Analytics

Patterns

Interpretation and Evaluation

Analytics Application

All processes are iterative in nature

(Baesens, 2014)

Analytics Process Model

1. A thorough definition of business problem to be solved with analytics.
2. All source data need to be identified that could be of potential interest.
3. All data are gathered in a staging area, e.g. a data mart or data warehouse.
4. Basic exploratory analysis can be considered, e.g. online analytical processing (OLAP) for multidimensional data analysis.
5. Data cleaning to get rid of all inconsistencies, e.g. missing values, outliers and duplicate data.

Analytics Process Model

6. Additional transformation can be considered, e.g. binning, alphanumeric to numeric coding, geographical aggregation, etc.
7. The analytics step involves an analytical model of the preprocessed and transformed data. Different types of analytics can be considered (e.g. fraud detection, customer segmentation, market basket analysis, etc.)
8. Once the model is built, then interpret and evaluate it by business experts.
9. Once the analytical model is validated and approved, then it can be put into production as an analytics application (e.g. decision support system)

Example Applications

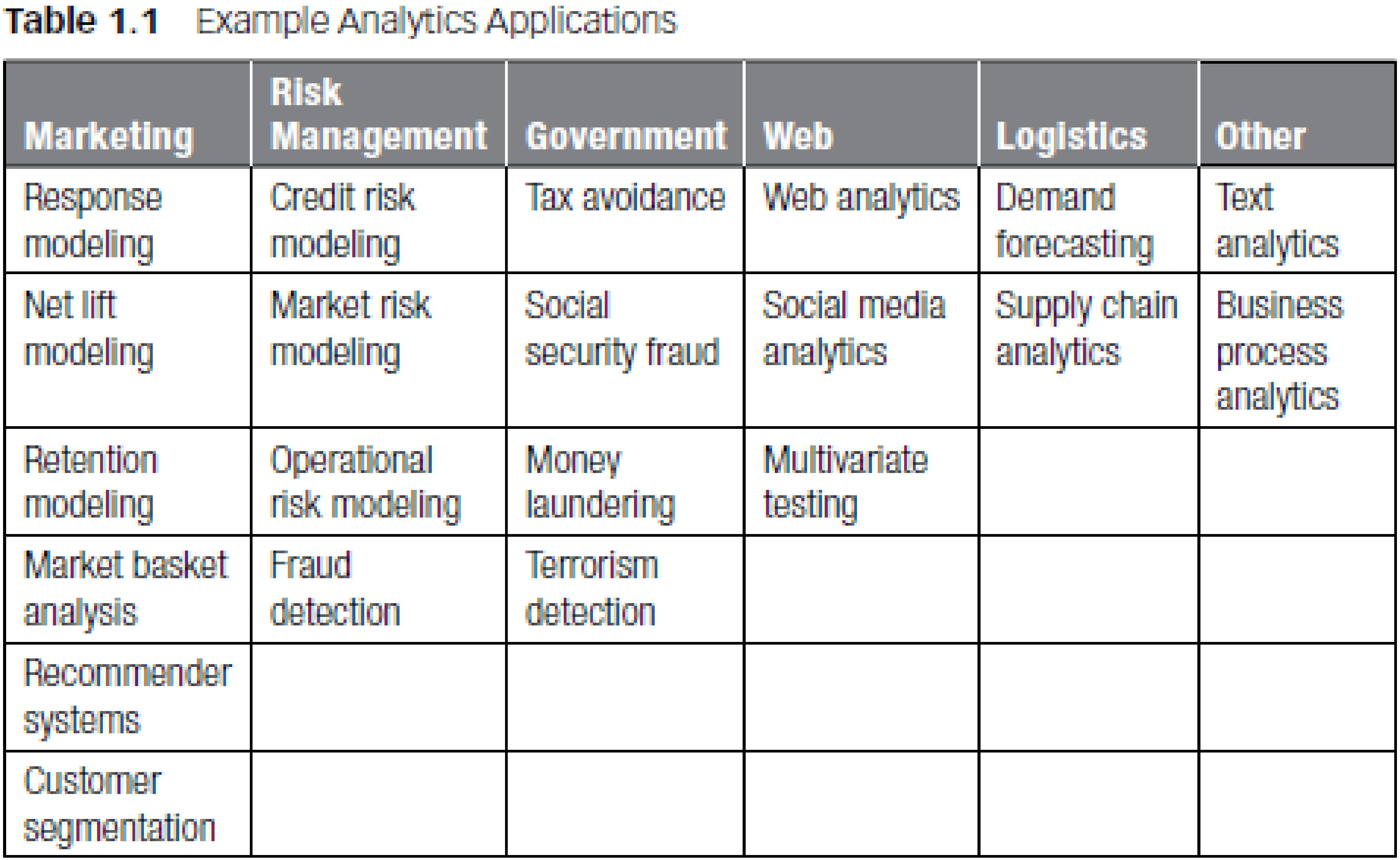


Table 1.1 Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

(Baesens, 2014)