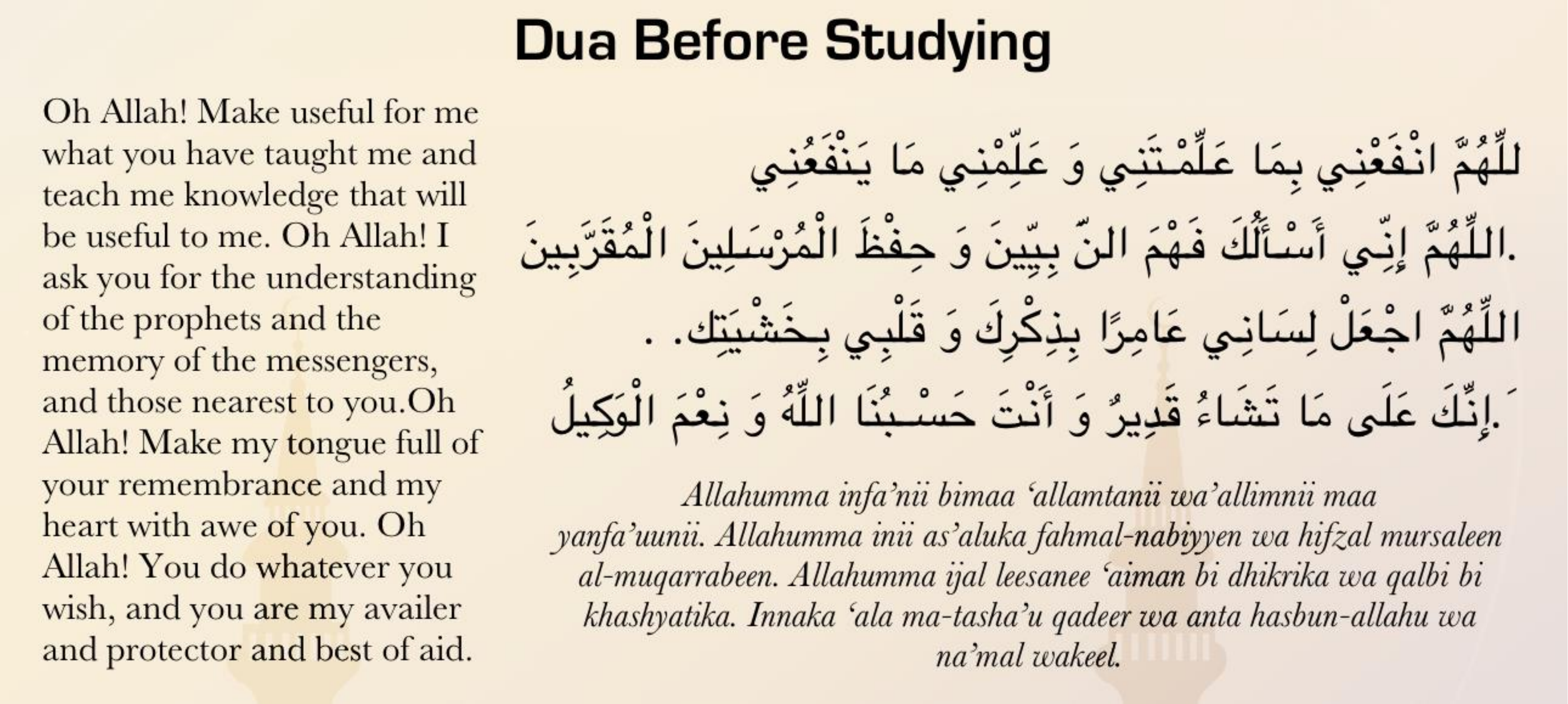
CSC 3303 BIG DATA ANALYTICS

DR. RAINI HASSAN

Department of Computer Science, KICT, IIUM
Office Room: C2-14 (Block C, Level 2)
Email: hrai@iium.edu.my
Office Telephone No: +603 6196 5655

Dua Before Studying

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ



Dua Before Studying

لِلّهُمَّ انْفَعْنِي بِمَا عَلَّمْتَنِي وَ عَلِّمْنِي مَا يَنْفَعُنِي

اللّهُمَّ إِنِّي أَسْأَلُكَ فَهْمَ النَّ بِيِّينَ وَ حِفْظَ الْمُرْسَلِينَ الْمُقَرَّبِينَ.

اللّهُمَّ اجْعَلْ لِسَانِي عَامِرًا بِذِكْرِكَ وَ قَلْبِي بِخَشْيَتِك. .

إِنَّكَ عَلَى مَا تَشَاءُ قَدِيرٌ وَ أَنْتَ حَسْبُنَا اللّهُ وَ نِعْمَ الْوَكِيلُ.َ

Allahumma infa'nii bimaa 'allamtanii wa'allimnii maa yanfa'uunii. Allahumma inii as'aluka fahmal-nabiyyen wa hifzal mursaleen al-muqarrabeen. Allahumma ijal leesanee 'aiman bi dhikrika wa qalbi bi khashyatika. Innaka 'ala ma-tasha'u qadeer wa anta hasbun-allahu wa na'mal wakeel.

Oh Allah! Make useful for me what you have taught me and teach me knowledge that will be useful to me. Oh Allah! I ask you for the understanding of the prophets and the memory of the messengers, and those nearest to you.Oh Allah! Make my tongue full of your remembrance and my heart with awe of you. Oh Allah! You do whatever you wish, and you are my availer and protector and best of aid.

Introduction to Big Data Analytics

| DATA COLLECTION, SAMPLING, AND PREPROCESSING, IN ANALYTICS IN A BIG DATA WORLD: THE ESSENTIAL GUIDE TO DATA SCIENCE AND ITS APPLICATIONS | BAESENS B. | 2012 | JOHN WILEY & SONS |

AND

| DATA SCIENCE AND BIG DATA ANALYTICS: DISCOVERING, ANALYZING, VISUALIZING AND PRESENTING DATA | EMC EDUCATION SERVICES | 2015 | INDIANA, USA: JOHN WILEY & SONS |

AND

| MY OWN ADDITIONAL OUTLINES AND CONTENTS, PREPARED FROM MULTIPLE RESOURCES |

Outlines

1. Big Data Overview
2. Basic Nomenclature
3. Examples of Big Data Analytics
4. State of the Practice in Analytics
5. Key Roles for the New Big Data Ecosystem
6. Analytics Process Model
7. Job Profiles Involved
8. Analytics
9. Analytical Model Requirements

1. Big Data Overview

Introduction [1]

- Data is created constantly, and at an ever-increasing rate.
- Mobile phones, social media, imaging technologies to determine a medical diagnosis…
  - all these and more create new data, and that must be stored somewhere for some purpose.
- Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.

Introduction [2]

- Merely keeping up with this huge influx of data is difficult.
- Substantially more challenging is analysing vast amounts of it…
  - especially when it does not conform to traditional notions of data structure,
  - to identify meaningful patterns and extract useful information.
- These challenges of the data deluge present the opportunity to transform…
  - business, government, science, and everyday life.

** Influx = an arrival or entry of large numbers of people or things; Deluge = a severe flood*
Source: Oxford Dictionary

Introduction [3]

- Several industries have led the way in developing their ability to gather and exploit data:

1. Credit card companies:

- Monitor every purchase their customers make and…
- can identify fraudulent purchases with a high degree of accuracy using rules…
- derived by processing billions of transactions.

Introduction [4]

- Several industries have led the way in developing their ability to gather and exploit data:

2. Mobile phone companies:

- Analyse subscribers' calling patterns to determine, for example:
  - whether a caller's frequent contacts are on a rival network.
- If that rival network is offering an attractive promotion that might cause the subscriber to defect…
- the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.

Introduction [5]

- Several industries have led the way in developing their ability to gather and exploit data:

3. Companies such as LinkedIn and Facebook:

- Data itself is their primary product.
- The valuations of these companies are heavily derived from the data they gather and host…
- which contains more and more intrinsic value as the data grows.

Introduction [6]

- Three attributes stand out as defining Big Data characteristics:

1. Huge volume of Data:

- Rather than thousands or millions of rows…
- Big Data can be billions of rows and millions of columns.

Introduction [7]

- Three attributes stand out as defining Big Data characteristics:

2. Complexity of data types and structures:

- Big Data reflects the variety of new data sources, formats, and structures…
- including digital traces being left on the web and other digital repositories for subsequent analysis.

Introduction [8]

- Three attributes stand out as defining Big Data characteristics:

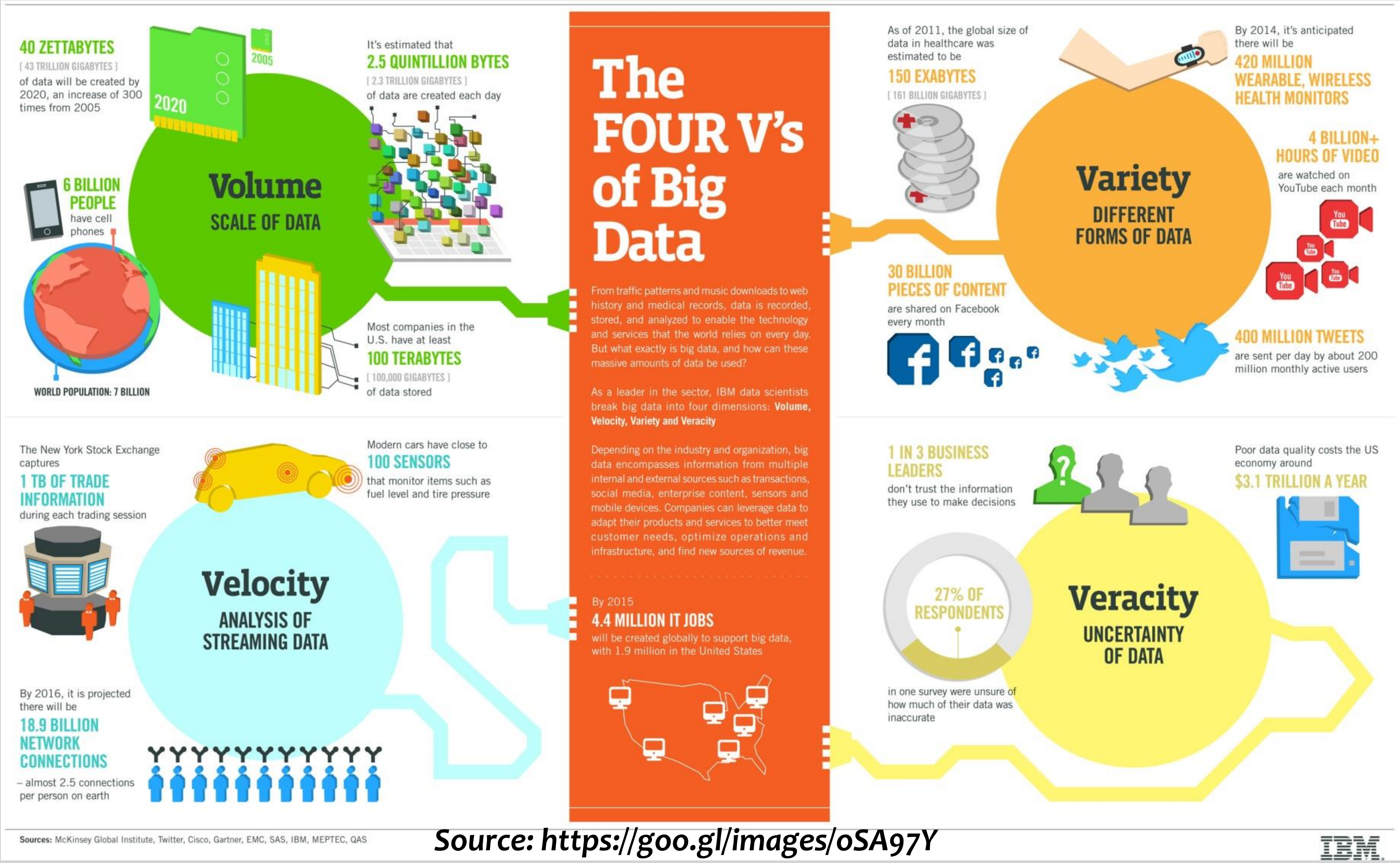
3. Speed of new data creation and growth:

- Big Data can describe high velocity data…
- with rapid data ingestion and near real time analysis.

Introduction [9]

- Although the volume of Big Data tends to attract the most attention…
  - generally the variety and velocity of the data provide a more apt definition of Big Data.
  - (Big Data is sometimes described as having 3 Vs: volume, variety, and velocity.)
- Due to its size or structure, Big Data cannot be efficiently analysed using only traditional databases or methods.

Introduction [10]



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

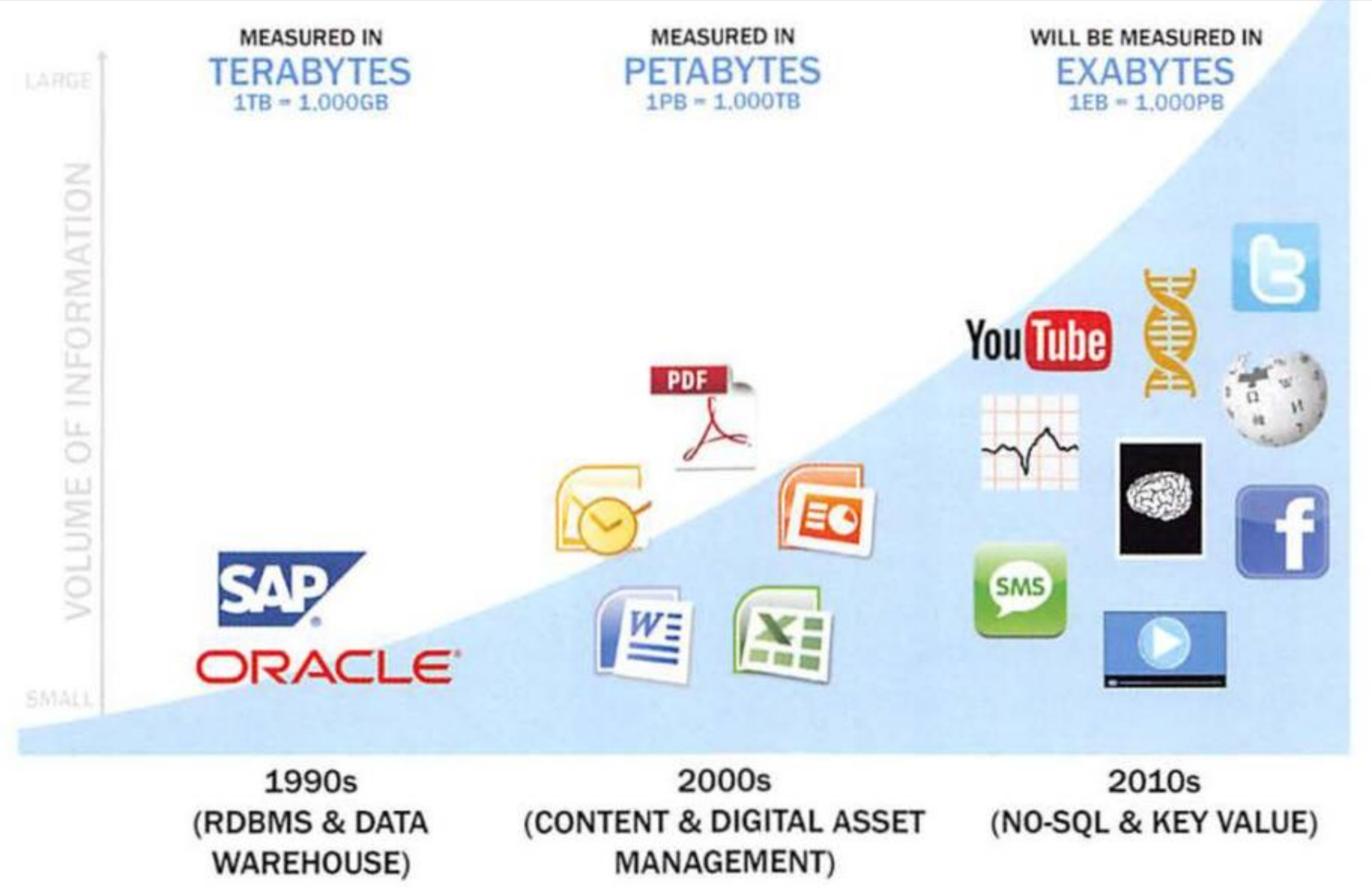
Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Source: https://goo.gl/images/oSA97Y

Introduction [11]



Introduction [12]

- Big Data problems require new tools and technologies to:
  - store,
  - manage, and
  - realize the business benefit.
- These new tools and technologies enable:
  - creation,
  - manipulation, and
  - management of large datasets and the storage environments that house them.

Introduction [13]

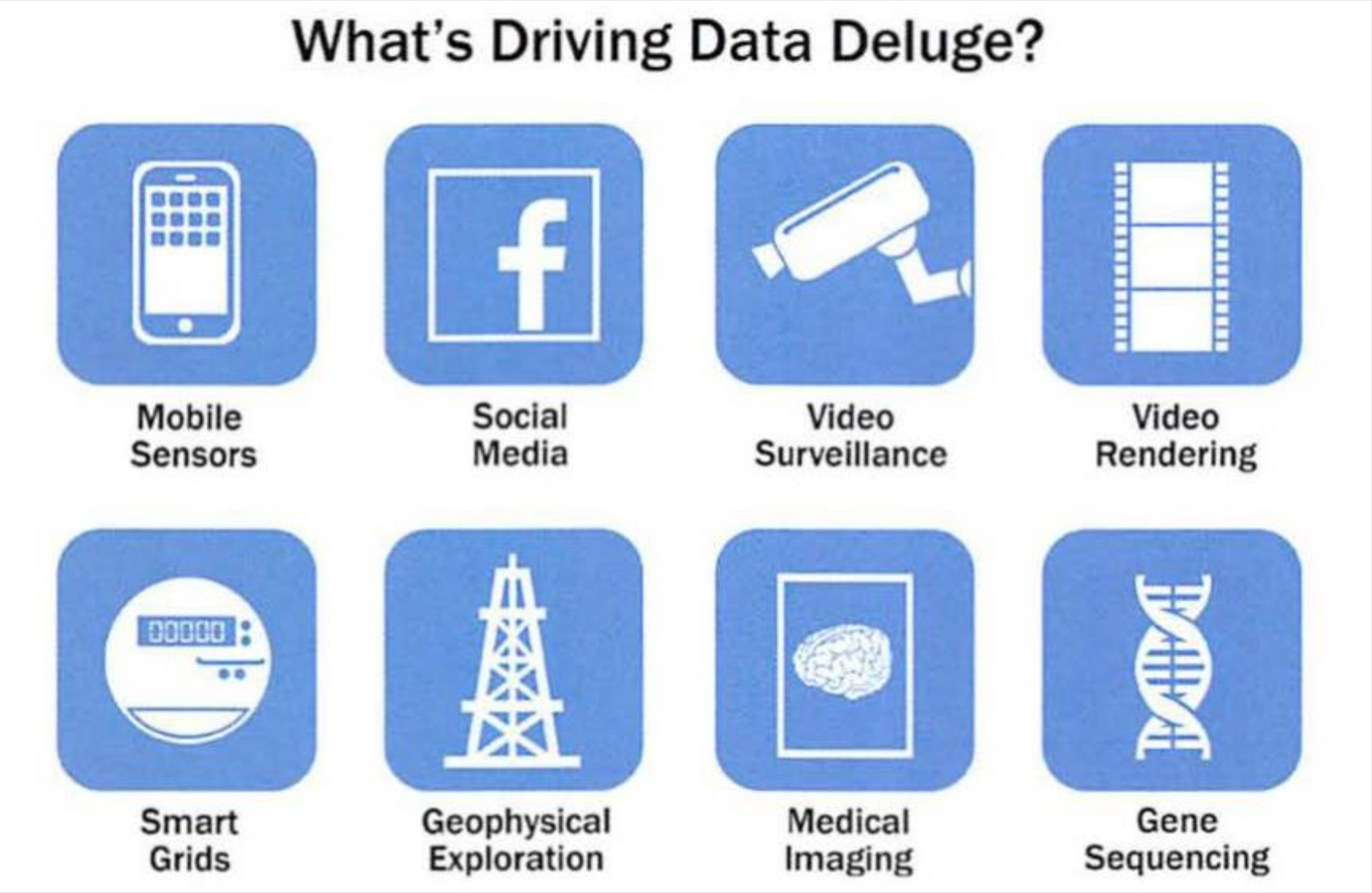
- Another definition of Big Data comes from the McKinsey Global report from 2011:

  "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 ne w sources of business value."

- McKinsey's definition of Big Data implies that organizations will need:
  - new data architectures and analytics and boxes,
  - new tools,
  - new analytical methods, and
  - an integration of multiple skills into the new role of the data scientist.

** Deluge = a severe flood*
Source: Oxford Dictionary

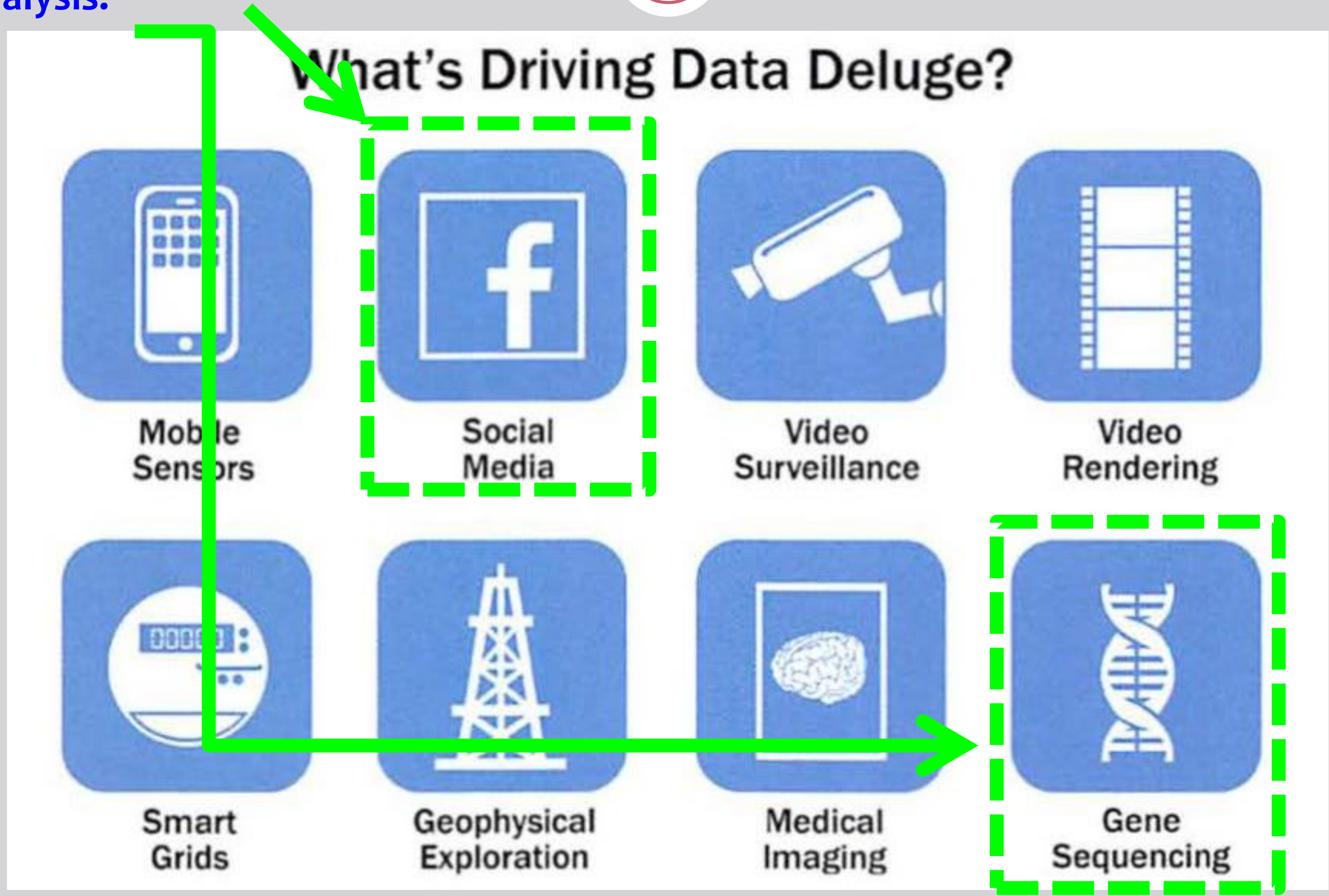
Introduction [14]



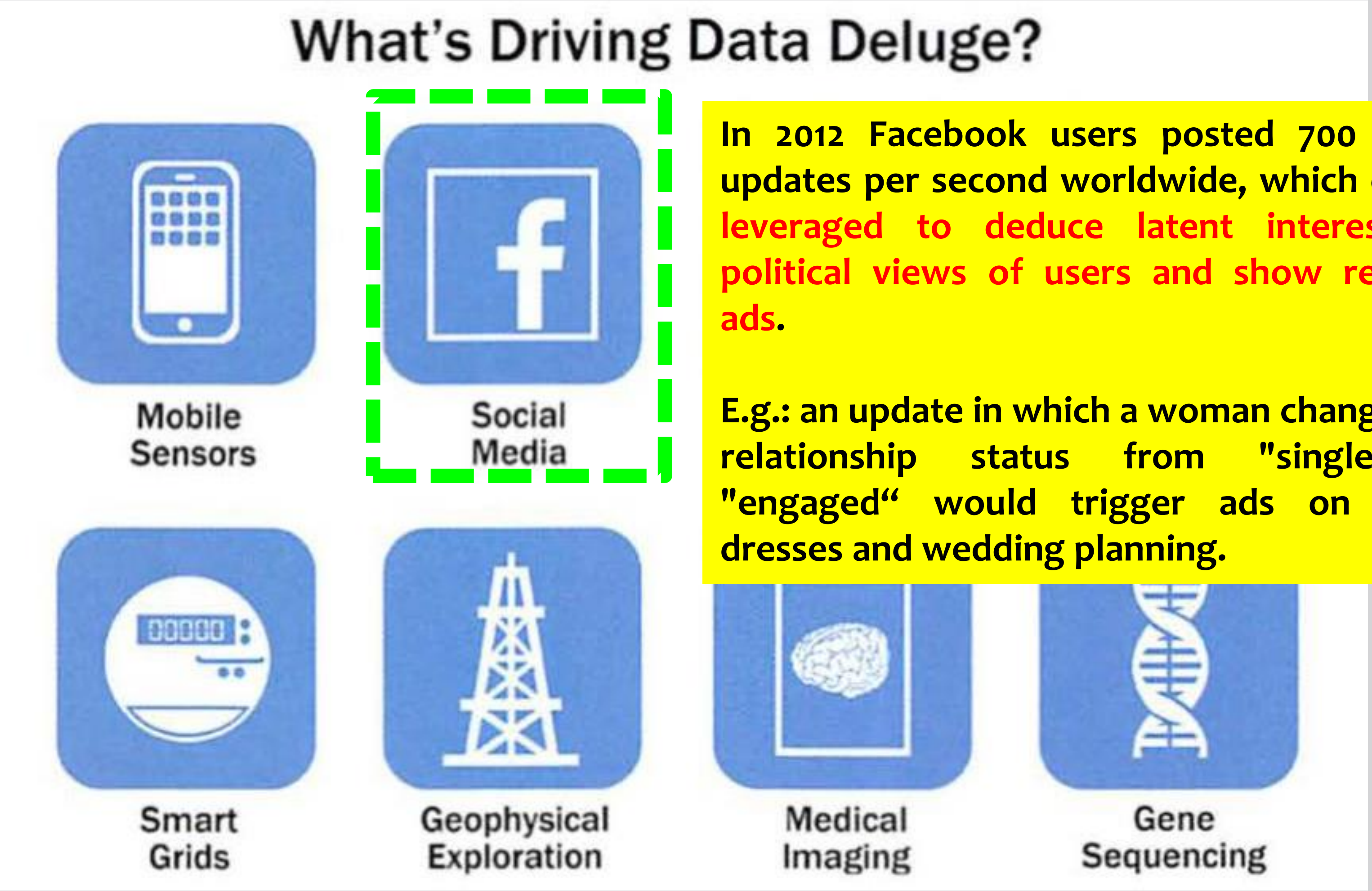
What's Driving Data Deluge?

Introduction [15]

The fastest-growing sources of Big Data and examples of untraditional sources of data being used for analysis.



Introduction [16]

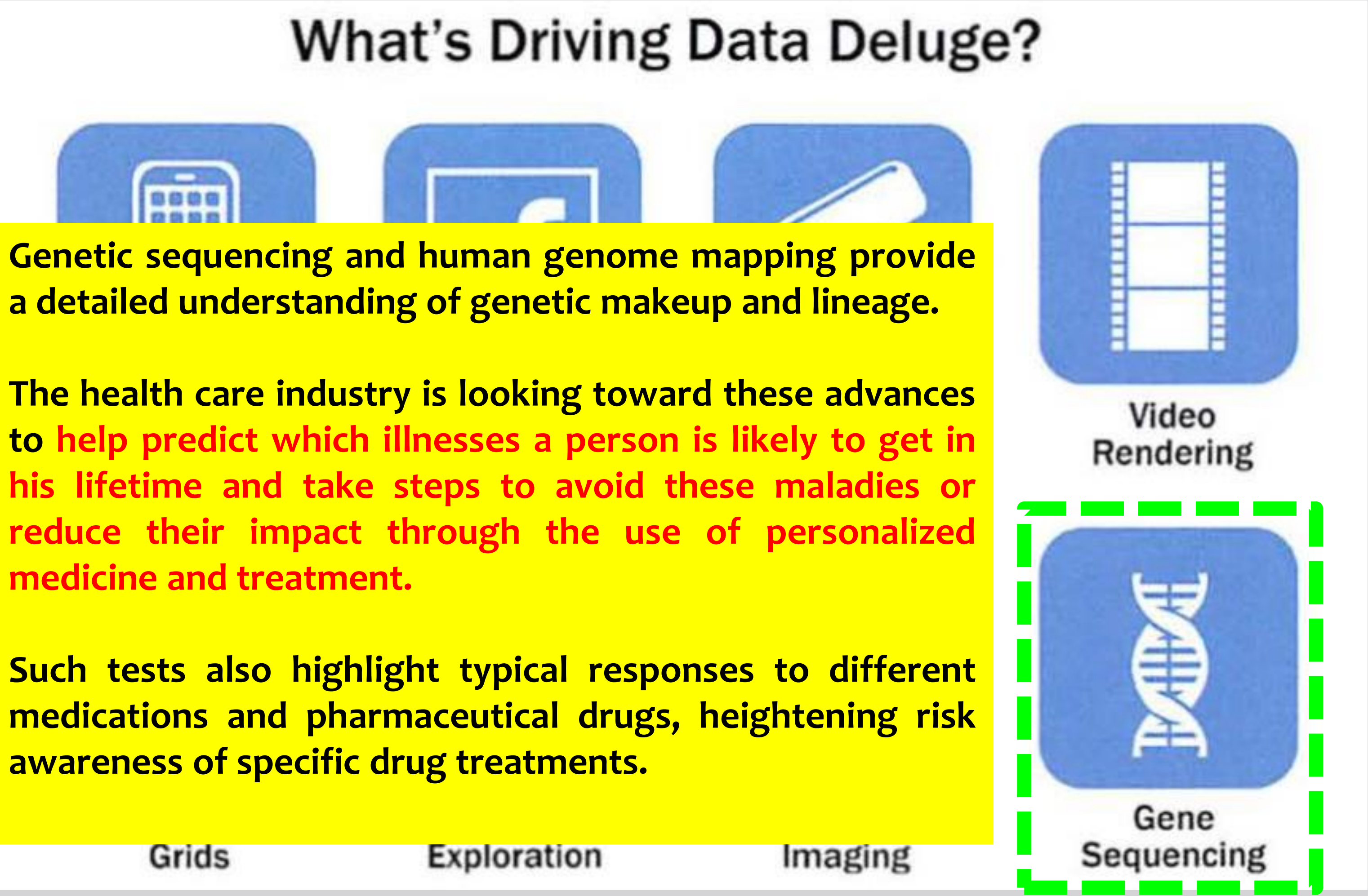


What's Driving Data Deluge?

In 2012 Facebook users posted 700 status updates per second worldwide, which can be leveraged to deduce latent interests or political views of users and show relevant ads.

E.g.: an update in which a woman changes her relationship status from "single" to "engaged“ would trigger ads on bridal dresses and wedding planning.

Introduction [17]



What's Driving Data Deluge?

Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage.

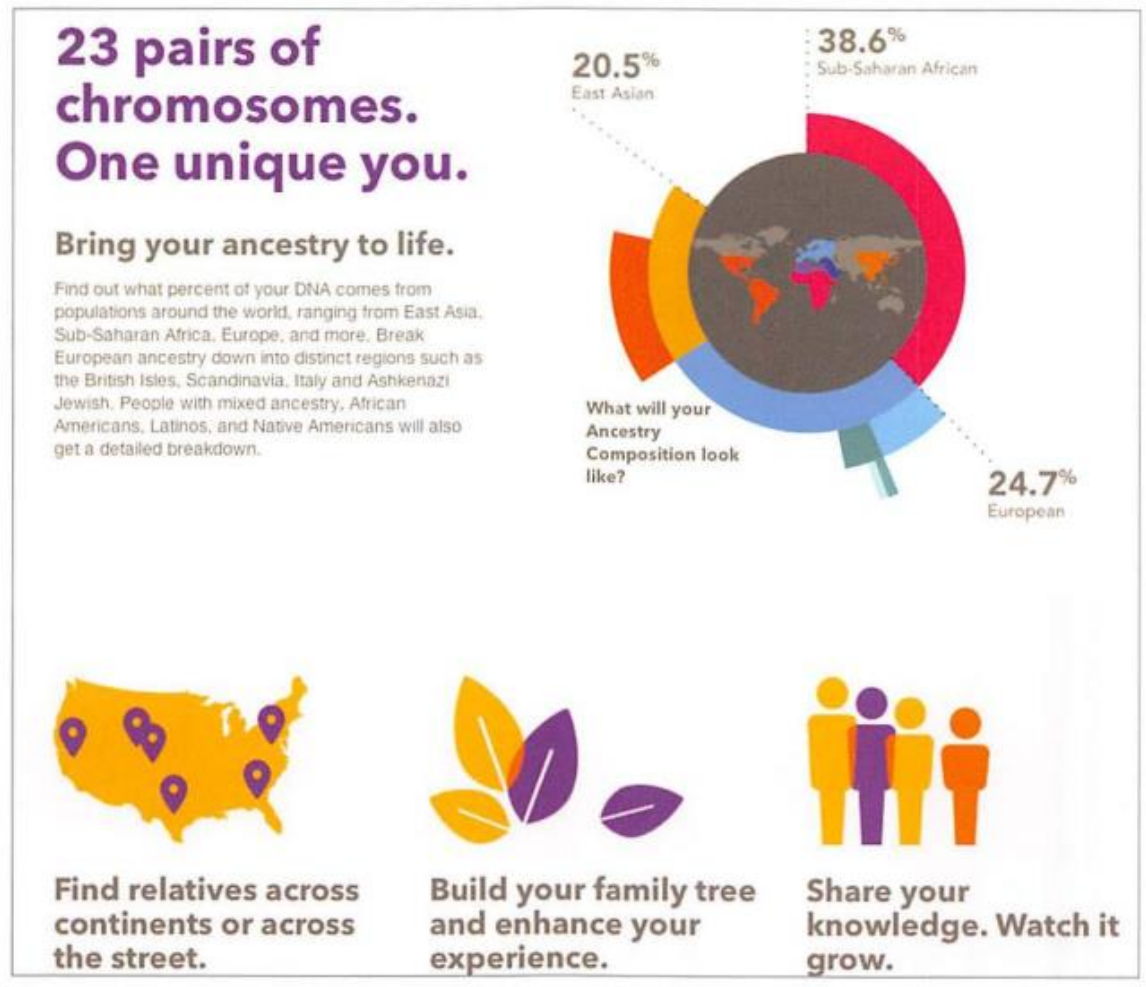
The health care industry is looking toward these advances to help predict which illnesses a person is likely to get in his lifetime and take steps to avoid these maladies or reduce their impact through the use of personalized medicine and treatment.

Such tests also highlight typical responses to different medications and pharmaceutical drugs, heightening risk awareness of specific drug treatments.

Introduction [18]

- While data has grown, the cost to perform this work has fallen dramatically.
- The cost to sequence one human genome has fallen from $100 million in 2001 to $10,000 in 2011, and the cost continues to drop.
- Now, websites such as 23andme (*next slide*) offer genotyping for less than $100.
- Although genotyping analyses only a fraction of a genome and does not provide as much granularity as genetic sequencing…
  - *it does point to the fact that data and complex analysis is becoming more prevalent and less expensive to deploy.*

Introduction [19]



23 pairs of chromosomes. One unique you.

Bring your ancestry to life.

Find out what percent of your DNA comes from populations around the world, ranging from East Asia, Sub-Saharan Africa, Europe, and more. Break European ancestry down into distinct regions such as the British Isles, Scandinavia, Italy and Ashkenazi Jewish. People with mixed ancestry, African Americans, Latinos, and Native Americans will also get a detailed breakdown.

Find relatives across continents or across the street.

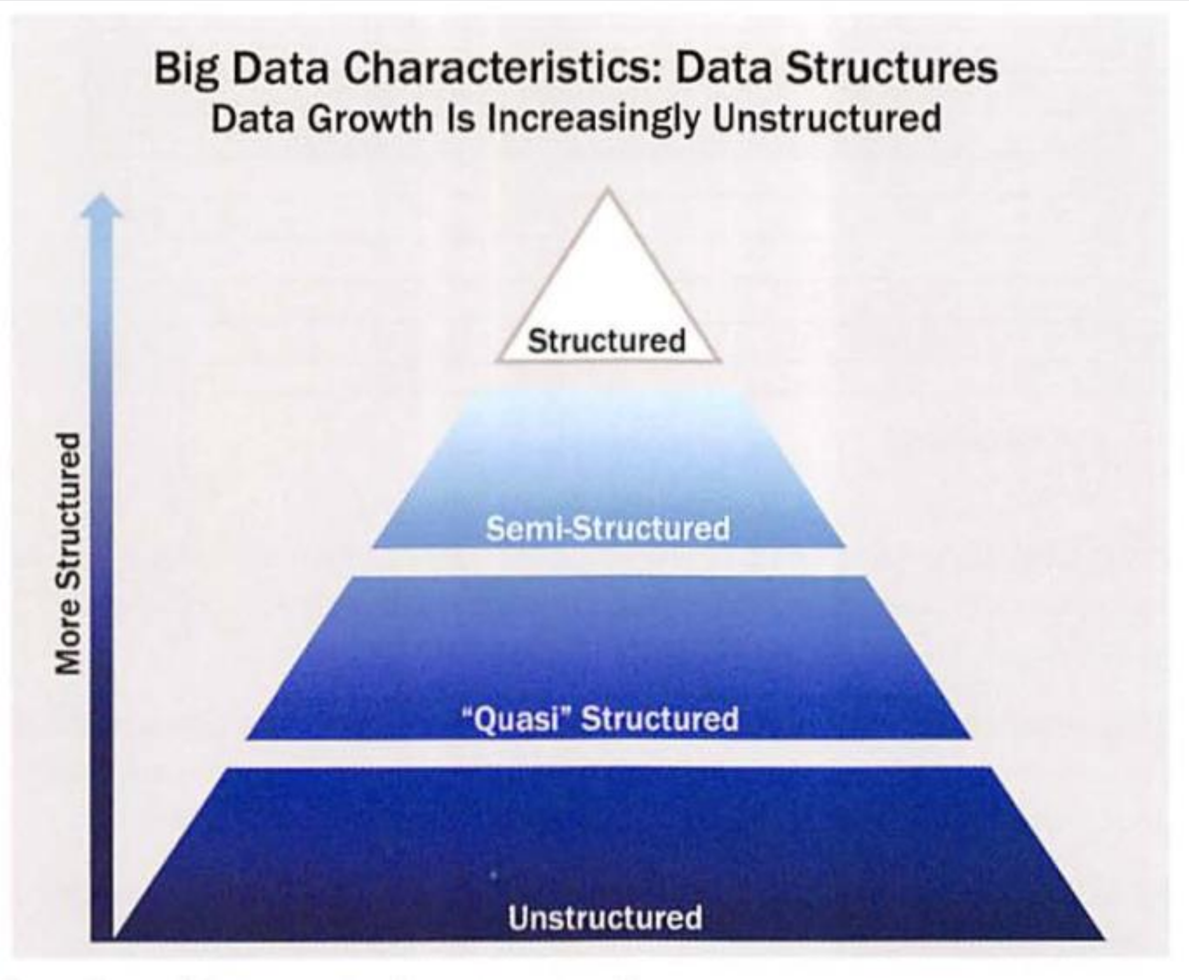
Build your family tree and enhance your experience.

Share your knowledge. Watch it grow.

Data Structures [1]

- Big data can come in multiple forms…
  - including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings.
- Contrary to much of the traditional data analysis performed by organizations…
  - most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyse.
- Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data.

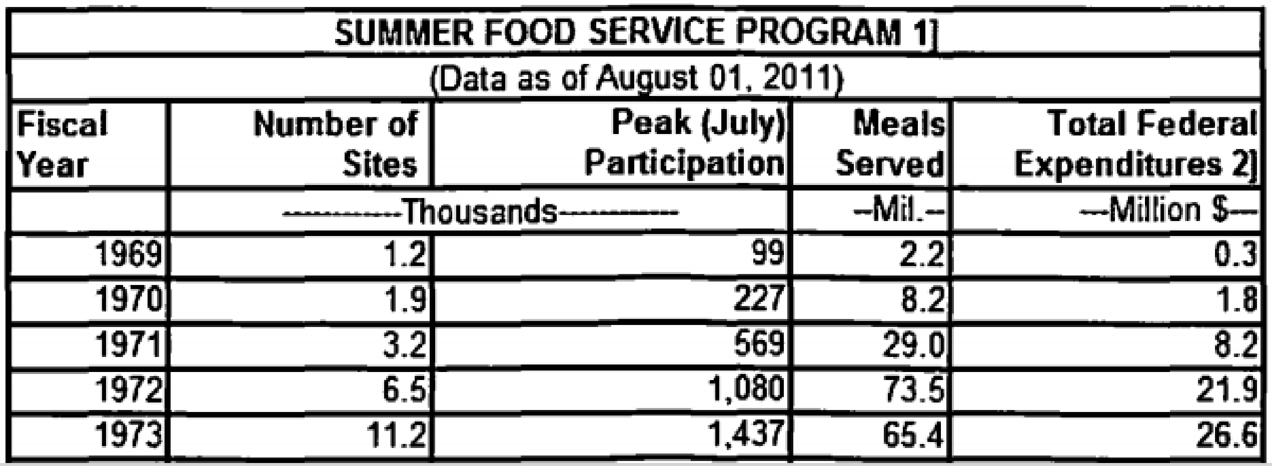
Data Structures [2]


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

Data Structures [3]

1. Structured Data:

- Data containing a defined:
  - data type,
  - format, and
  - structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets).

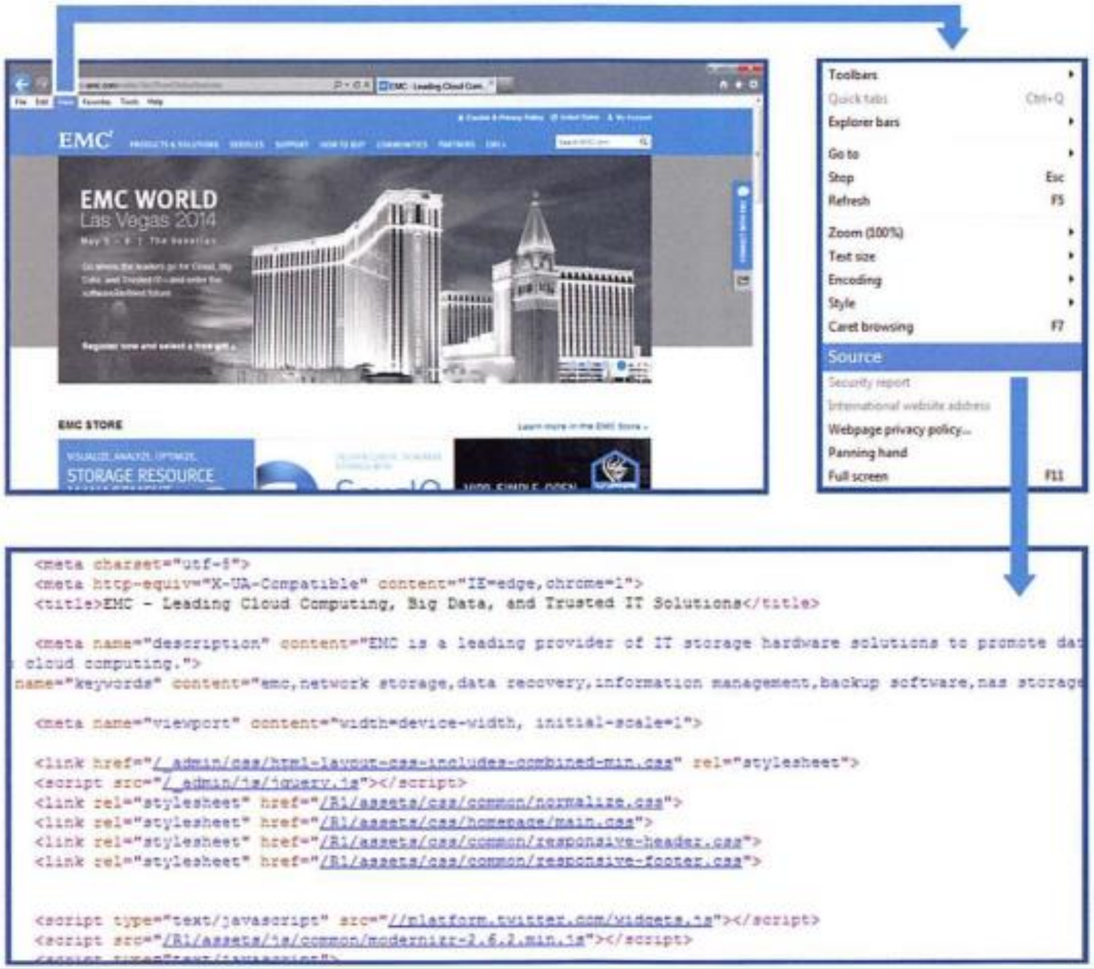


SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6

Data Structures [4]

2. Semi-Structured Data:

- Textual data files with a discernible pattern that enables parsing…
  - (such as Extensible Markup Language [XML] data files that are self describing and defined by an XML schema).

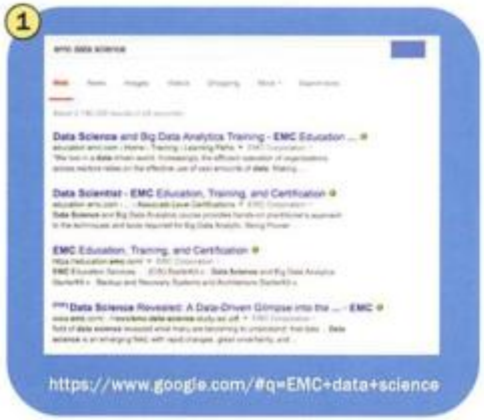


```
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

<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>
<script type="text/javascript">
```

Data Structures [5]

3. Quasi-Structured Data:

- Textual data with erratic data formats that can be formatted with effort, tools, and time…
  - (for instance, web clickstream data that may contain inconsistencies in data values and formats).



emc data science

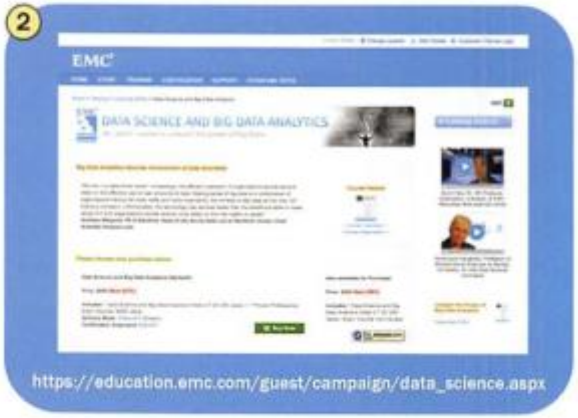
Data Science and Big Data Analytics Training - EMC Education ...

Data Scientist - EMC Education, Training, and Certification

EMC Education, Training, and Certification

Data Science Revealed: A Data-Driven Glimpse into the ... - EMC

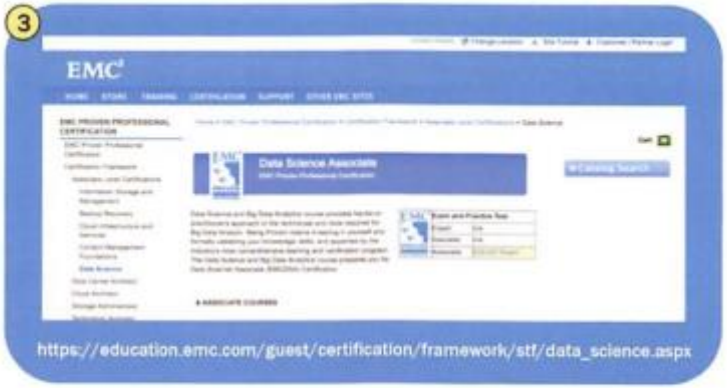
https://www.google.com/#q=EMC+data+science



EMC²

DATA SCIENCE AND BIG DATA ANALYTICS

https://education.emc.com/guest/campaign/data_science.aspx



EMC²

EMC PROVEN PROFESSIONAL CERTIFICATION

Data Science Associate

https://education.emc.com/guest/certification/framework/stf/data_science.aspx

Data Structures [6]

4. Unstructured Data:

- Data that has no inherent structure…
  - which may include text documents, PDFs, images, and video.

2. Basic Nomenclature

** Nomenclature* **= the devising or choosing of names for things**

Source: Oxford Dictionary

Basic Unit of Analysis [1]

Customers:

- Customers can be considered from various perspectives.
- Customer lifetime value (CLV) can be measured for either…
  - individual customers or
  - at the household level.
- Customers can also play different roles…
  - Payer vs the end user (*parents buying gifts to their children*)
  - Primary banking account vs secondary etc.
  - Main debtor vs guarantor

Basic Unit of Analysis [2]

Account behaviour:

- For example, consider a credit scoring exercise for which the aim is to predict whether the applicant will default on a particular mortgage loan account.
- The analysis can also be done at the transaction level. For example:
  - In insurance fraud detection, one usually performs the analysis at insurance claim level.
  - Also, in web analytics, the basic unit of analysis is usually a web visit or session.

Basic Unit of Analysis [3]

- In case of predictive analytics, the target variable needs to be appropriately defined.
- For examples:
  - when is a customer considered to be a churner (*customer turnover, i.e. no longer remain a customer*) or not,
  - a fraudster or not,
  - a responder or not, or
  - how should the CLV be appropriately defined?

3. Examples of Big Data Analytics

** Analytics = the systematic or computational analysis of data or statistics*

Source: Oxford Dictionary

Analytics is Everywhere [1]

1. Shopping habits:

- A specialized catalogue sent to the customer based on his/her characteristics and previous purchases behaviour.

2. Banking account:

- Checking account balance and credit payments during a particular period, together with other kinds of information available to the bank, to predict whether customers will default on loans during the next year.

Analytics is Everywhere [2]

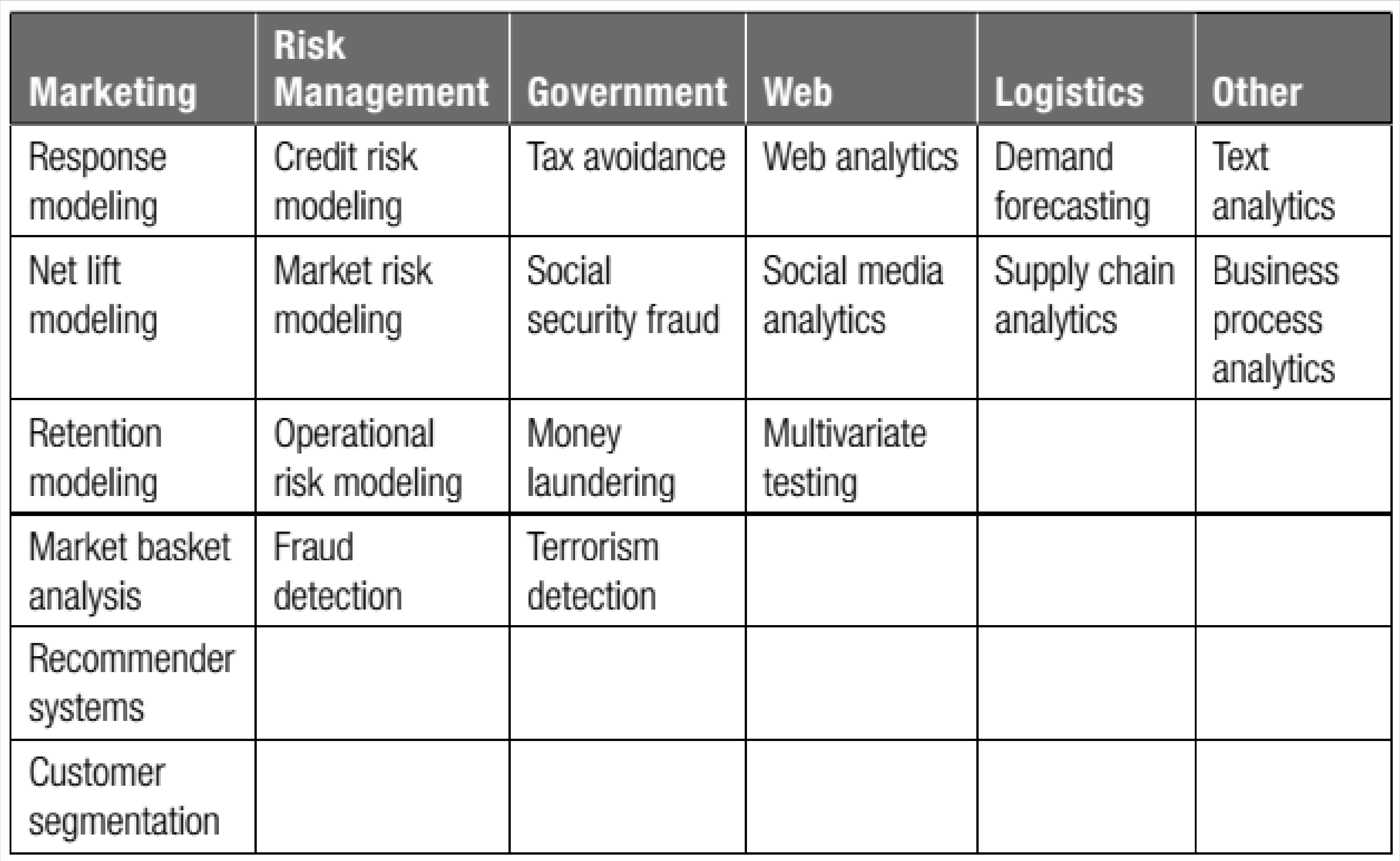
3. Telephone/mobile account:

- Analysing calling behaviour and account information to predict whether customers will churn in the next 3 months for example.

4. Facebook:

- Social ads appearing there were based on analysing all information (posts, pictures, friends and their behaviour, etc.) available to Facebook.

Analytics is Everywhere [3]



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					

Example Analytics Applications

Analytics is Everywhere [4]

- The relevance, importance, and impact of analytics are now bigger than ever before and…
- given that more and more data are being collected and…
- that there is strategic value in knowing what is hidden in data, analytics will continue to grow.

Analytics is Everywhere [5]

Class Activity 1

- I am going to assign a number to each one of you (random).

1

2

3

4

5

6

7

8

9

10

Analytics is Everywhere [6]

Class Activity 1

- Based on the number assigned earlier, pick the place &:
  - discuss about the analytics that can be implemented.
  - Think thoroughly, logically and yet simple (*straight to the point*).

1. HOSPITAL

2. POST OFFICE

3. RESTAURANT

4. HOTEL

5. BANK

6. AIRPORT

7. THEME PARK

8. SUPERMARKET

9. MUSEUM

10. CINEMA

4. State of the Practice in Analytics

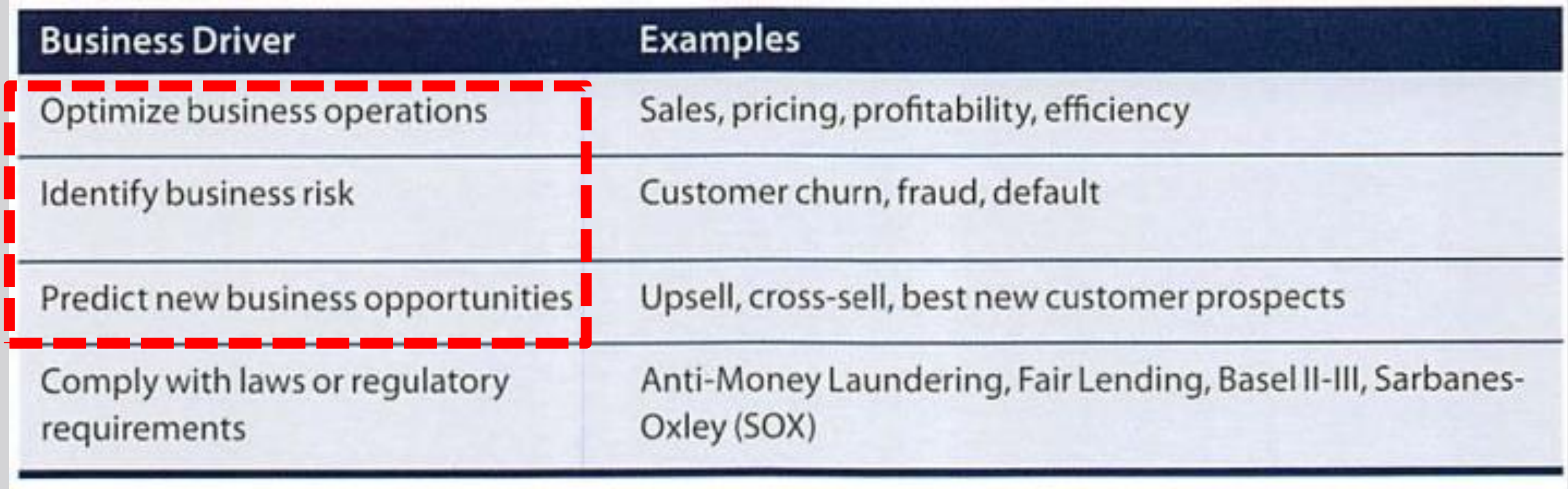
Business Drivers for Advanced Analytics [1]

- Current business problems provide many opportunities for organizations to become more analytical and data drive.
- There are four categories of common business problems that organizations…
  - contend with where they have an opportunity to leverage advanced analytics to create competitive advantage.

Business Drivers for Advanced Analytics [2]

- Rather than only performing standard reporting on these areas,…
  - organizations can apply advanced analytical techniques to optimize processes and
  - derive more value from these common tasks.

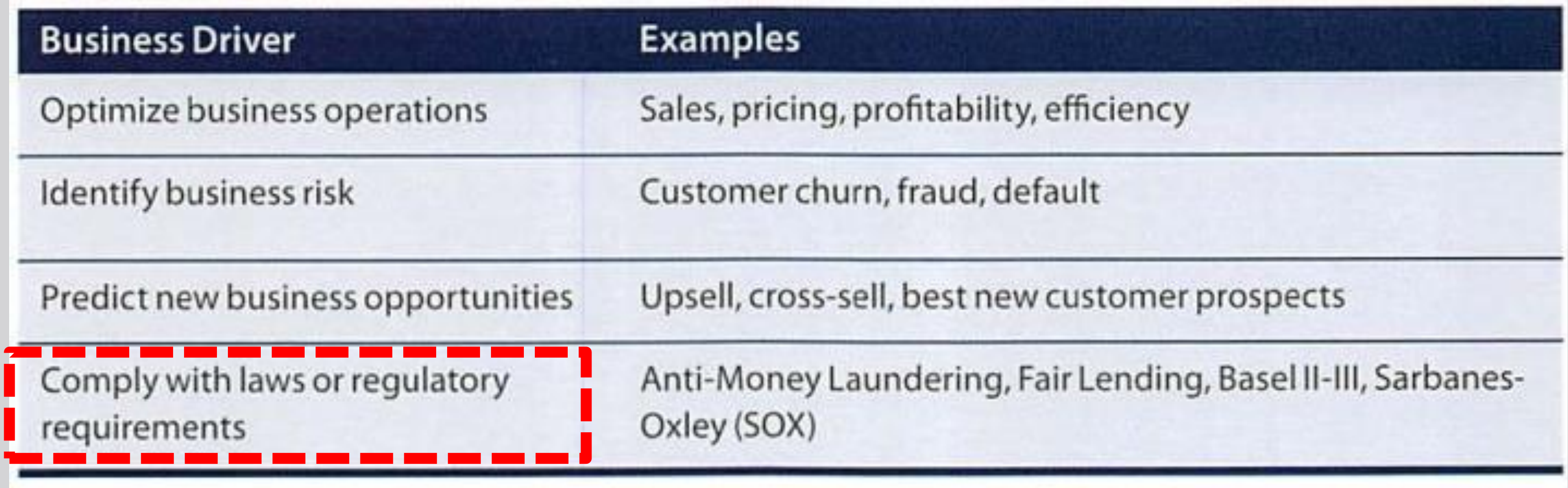
Business Drivers for Advanced Analytics [3]



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)

- The first 3 examples do not represent new problems.
- What is new is the opportunity to fuse advanced analytical techniques with Big Data…
  - to produce more impactful analyses for these traditional problem.

Business Drivers for Advanced Analytics [4]



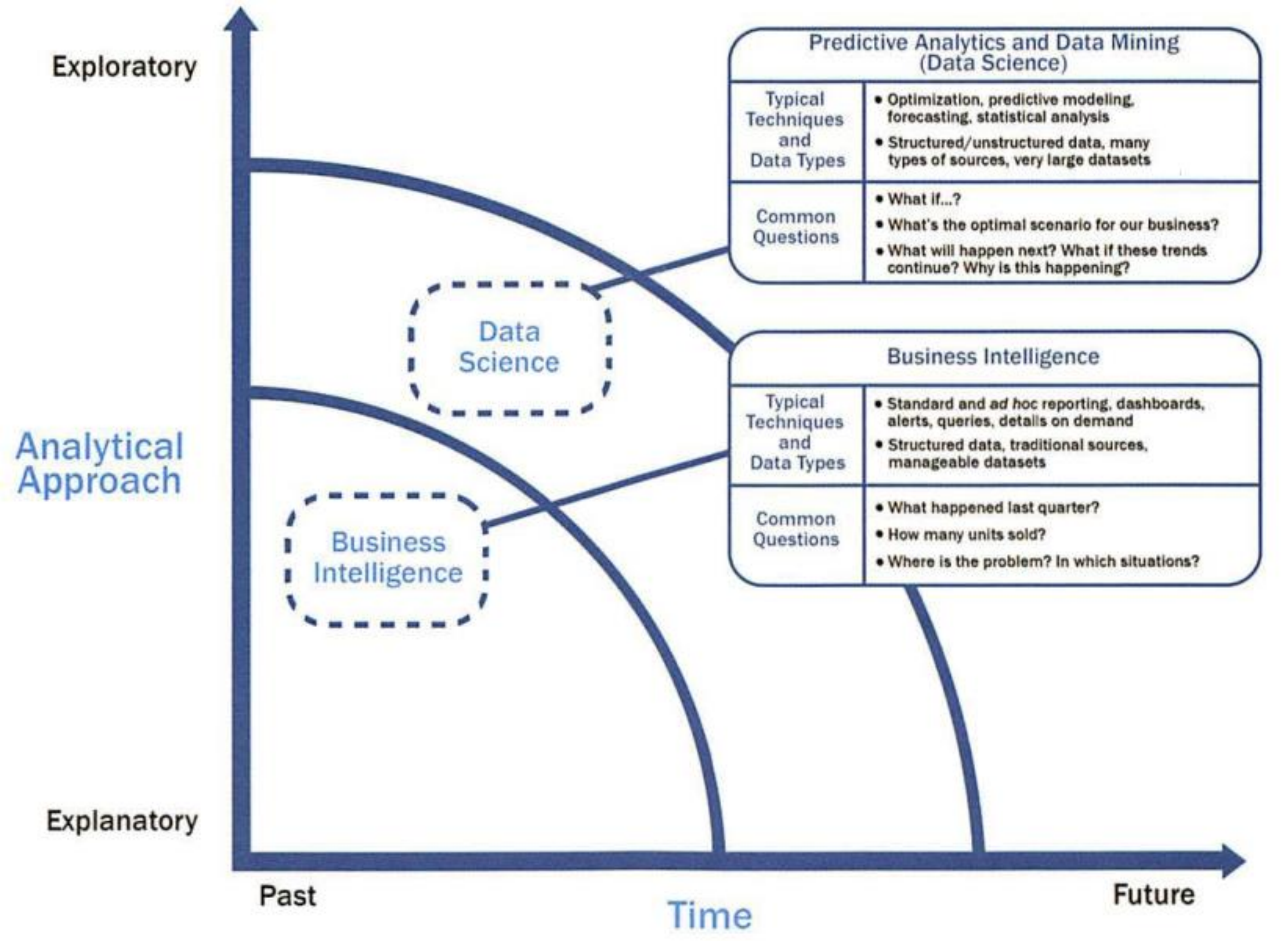
Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair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

- The last example portrays emerging regulatory requirements.
- Additional requirements of compliance and regulatory laws are added every year, which represent additional complexity and data requirements for organizations.
- Laws related to anti-money laundering (AML) and fraud prevention require advanced analytical techniques to comply with and manage properly.

BI versus Data Science [1]

- BI is referring to Business Intelligence.
- Although much is written generally about analytics, it important to distinguish between BI and Data Science.

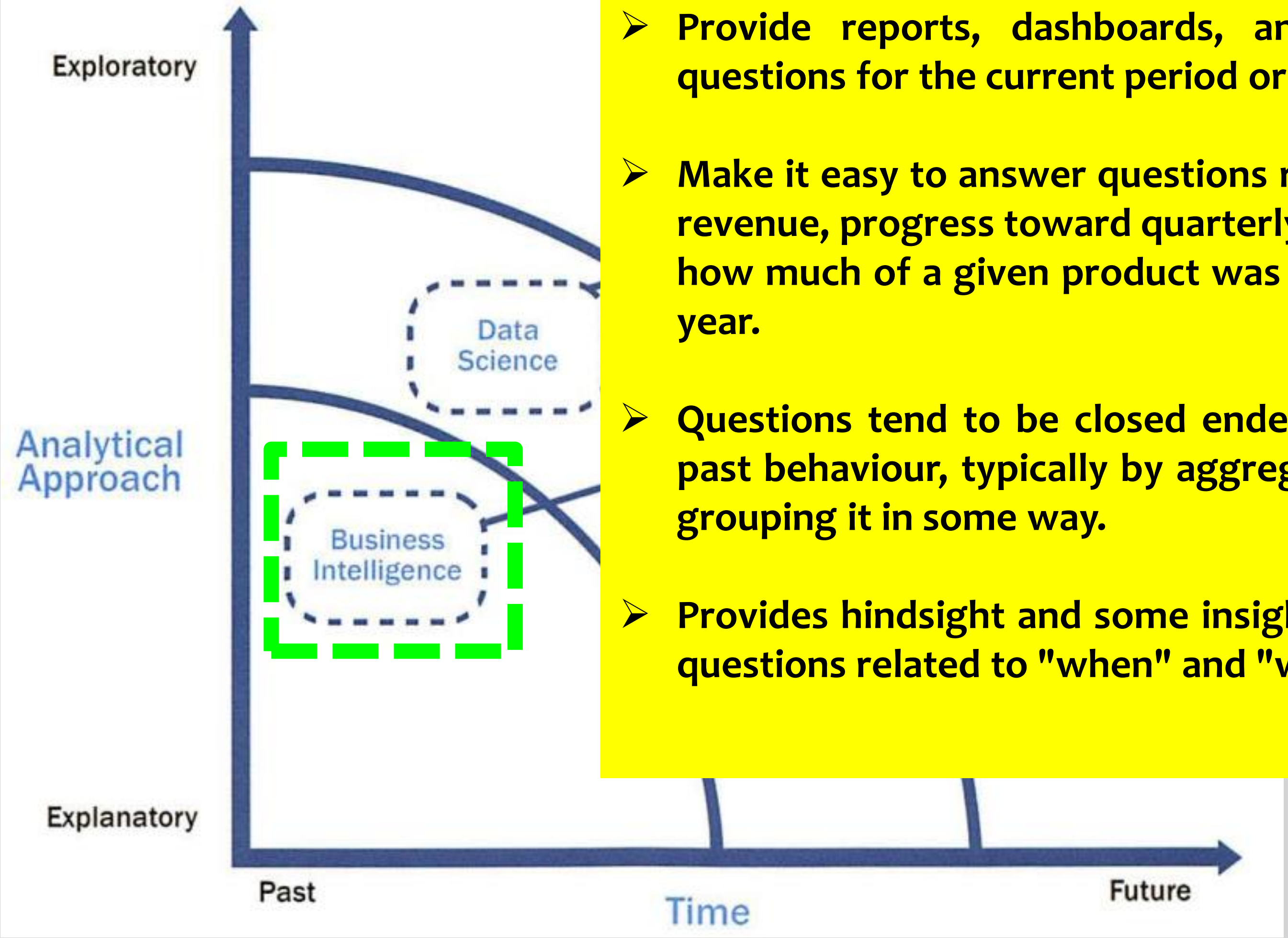
BI versus Data Science [2]



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?

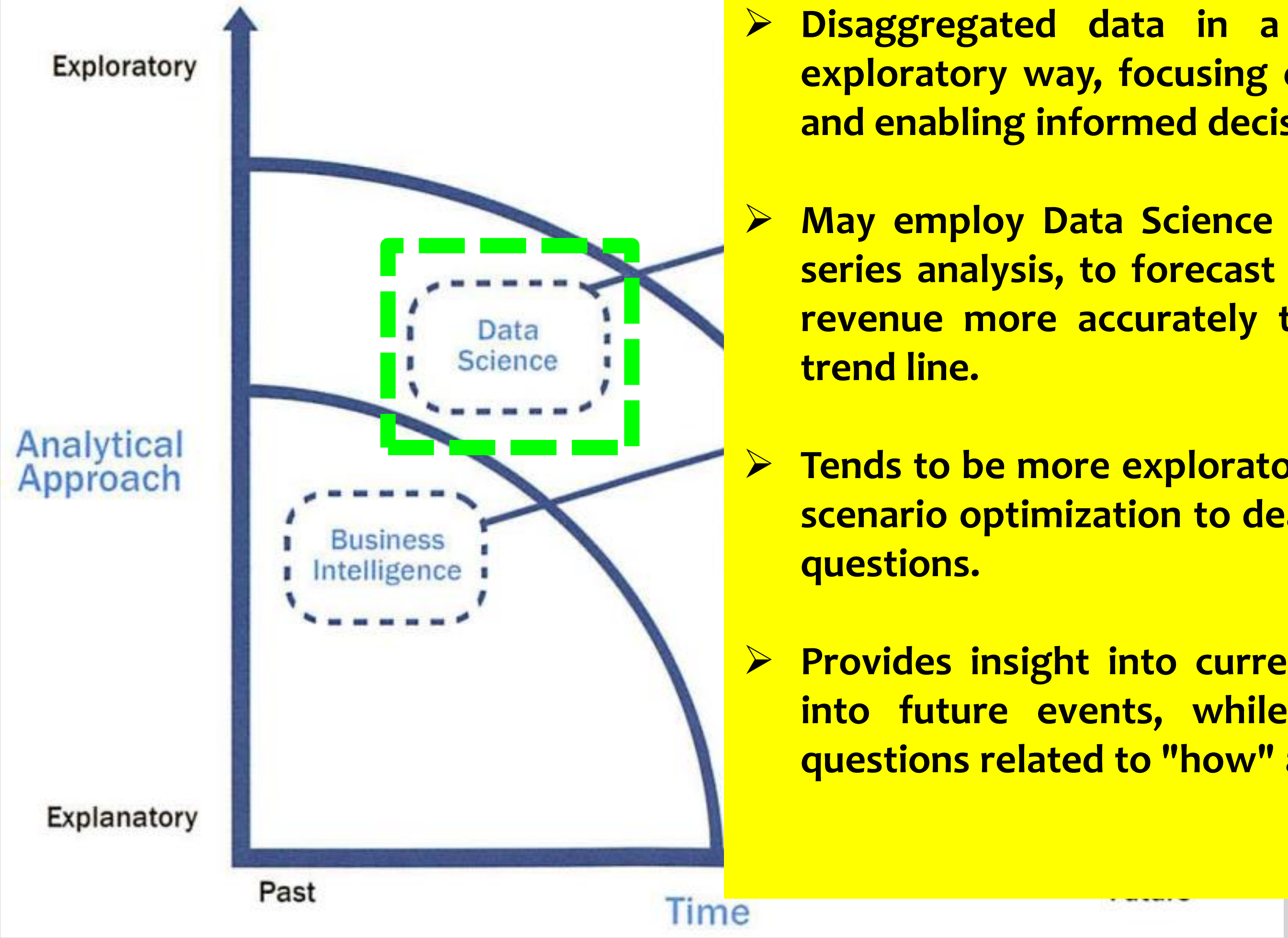
Business Intelligence	
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which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

BI versus Data Science [3]



- ➢ Provide reports, dashboards, and queries on business questions for the current period or in the past.
- ➢ Make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.
- ➢ Questions tend to be closed ended and explain current or past behaviour, typically by aggregating historical data and grouping it in some way.
- ➢ Provides hindsight and some insight and generally answers questions related to "when" and "where" events occurred.

BI versus Data Science [4]



- ➢ **Disaggregated data in a more forward-looking, exploratory way, focusing on analysing the present and enabling informed decisions about the future.**
- ➢ **May employ Data Science techniques such as time series analysis, to forecast future product sales and revenue more accurately than extending a simple trend line.**
- ➢ **Tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions.**
- ➢ **Provides insight into current activity and foresight into future events, while generally focusing on questions related to "how" and "why" events occur.**

BI versus Data Science [5]

- BI problems tend to require highly structured data organized in rows and columns for accurate reporting.
- Data Science projects tend to use many types of data sources, including large or unconventional datasets.
- Depending on an organization's goals, it may choose to embark on a…
- BI project if it is doing reporting, creating dashboards, or performing simple visualizations, or…
- it may choose Data Science projects if it needs to do a more sophisticated analysis with disaggregated or varied datasets.

5. Key Roles for the New Big Data Ecosystem

Emerging Big Data Ecosystem and a New Approach to Analytics [1]

- Organizations and data collectors are realizing that…
  - the data they can gather from individuals contains intrinsic value and,
  - as a result, a new economy is emerging.
- As this new digital economy continues to evolve…
  - the market sees the introduction of data vendors and data cleaners that use crowdsourcing (such as Mechanical Turk and GalaxyZoo)…
  - to test the outcomes of machine learning techniques.

Emerging Big Data Ecosystem and a New Approach to Analytics [2]

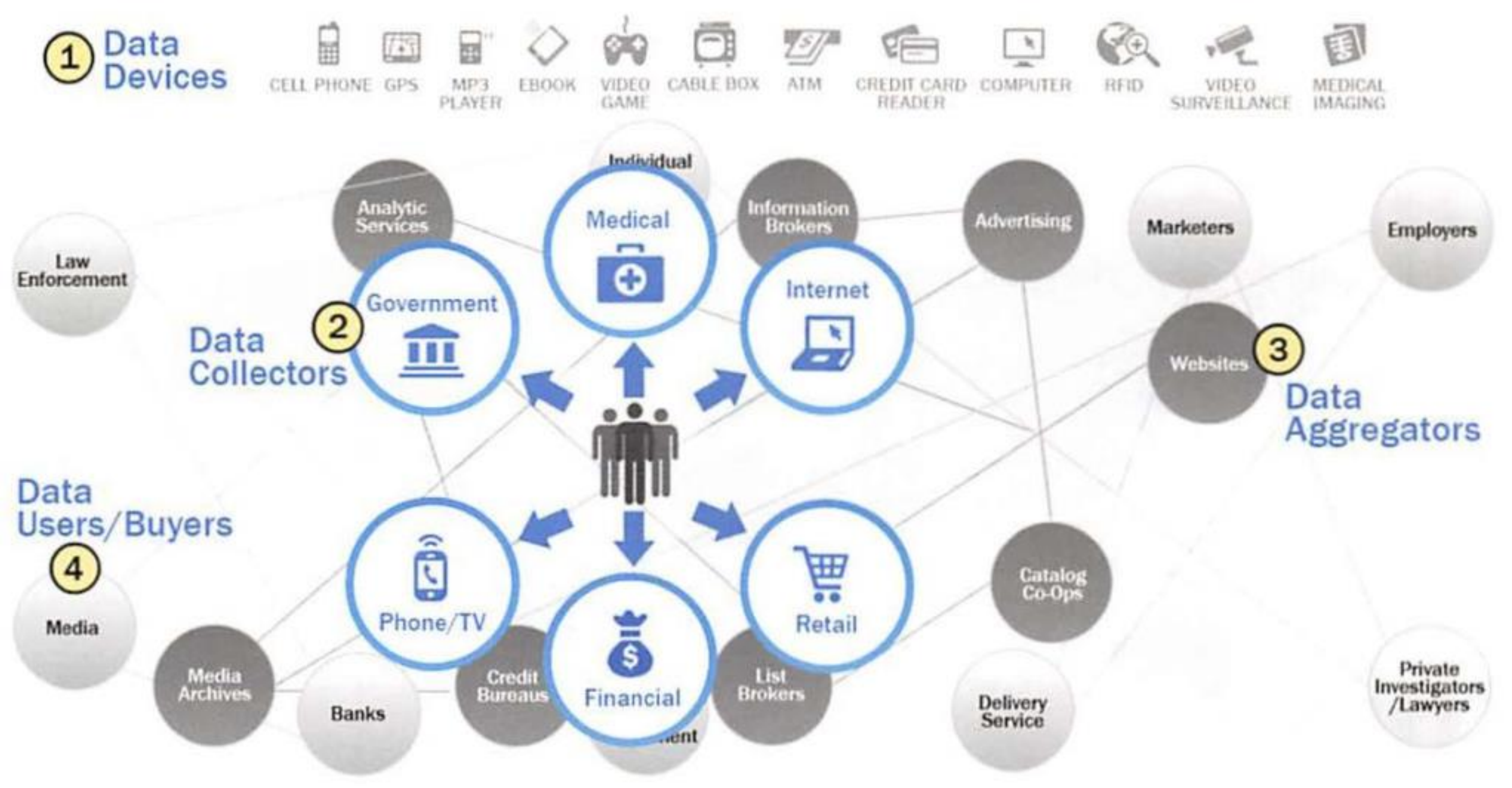
- Other vendors offer added value by…
  - repackaging open source tools in a simpler way and
  - bringing the tools to market.
- Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

Emerging Big Data Ecosystem and a New Approach to Analytics [3]

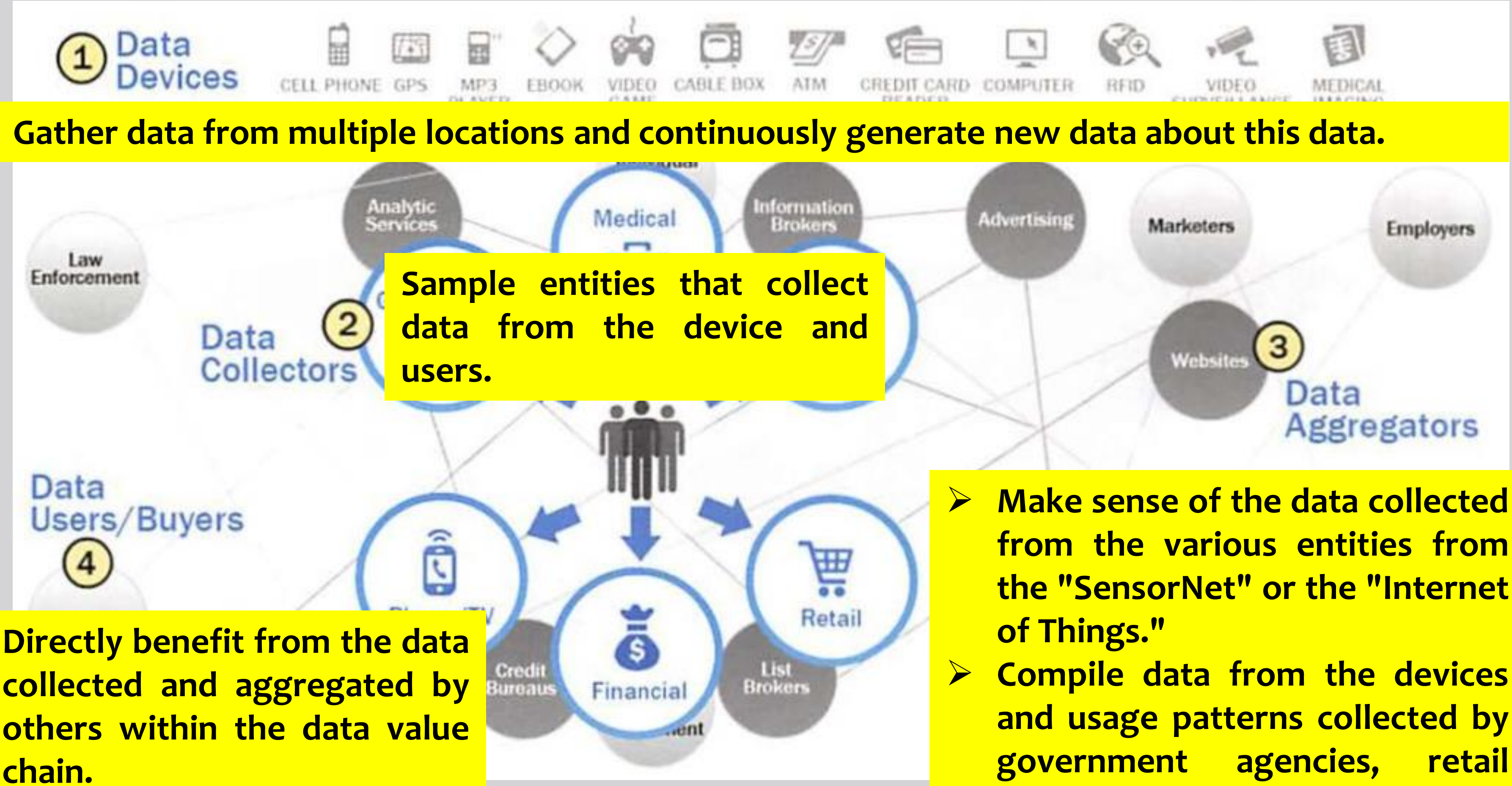
- As the new ecosystem takes shape, there are four main groups of players within this interconnected web:

1. Data devices
2. Data collectors
3. Data aggregators
4. Data users and buyers

Emerging Big Data Ecosystem and a New Approach to Analytics [4]



Emerging Big Data Ecosystem and a New Approach to Analytics [5]



Gather data from multiple locations and continuously generate new data about this data.

Sample entities that collect data from the device and users.

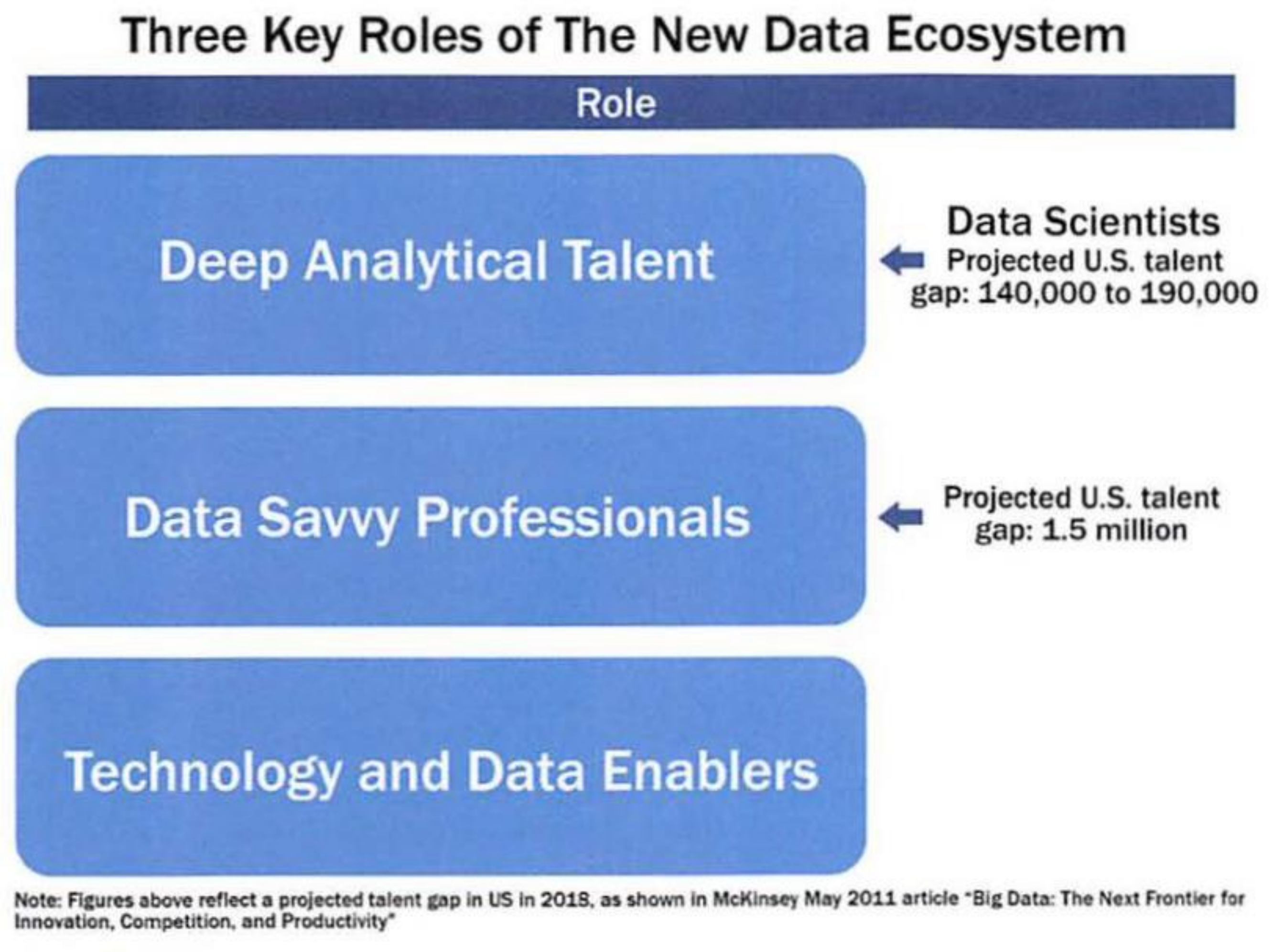
- Make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things."
- Compile data from the devices and usage patterns collected by government agencies, retail stores, etc.

Directly benefit from the data collected and aggregated by others within the data value chain.

Key Roles for the New Big Data Ecosystem [1]

- The Big Data ecosystem demands three categories of roles:
  1. Deep Analytical Talent
  2. Data Savvy Professional
  3. Technology and Data Enablers

Key Roles for the New Big Data Ecosystem [2]



Three Key Roles of The New Data Ecosystem

Role

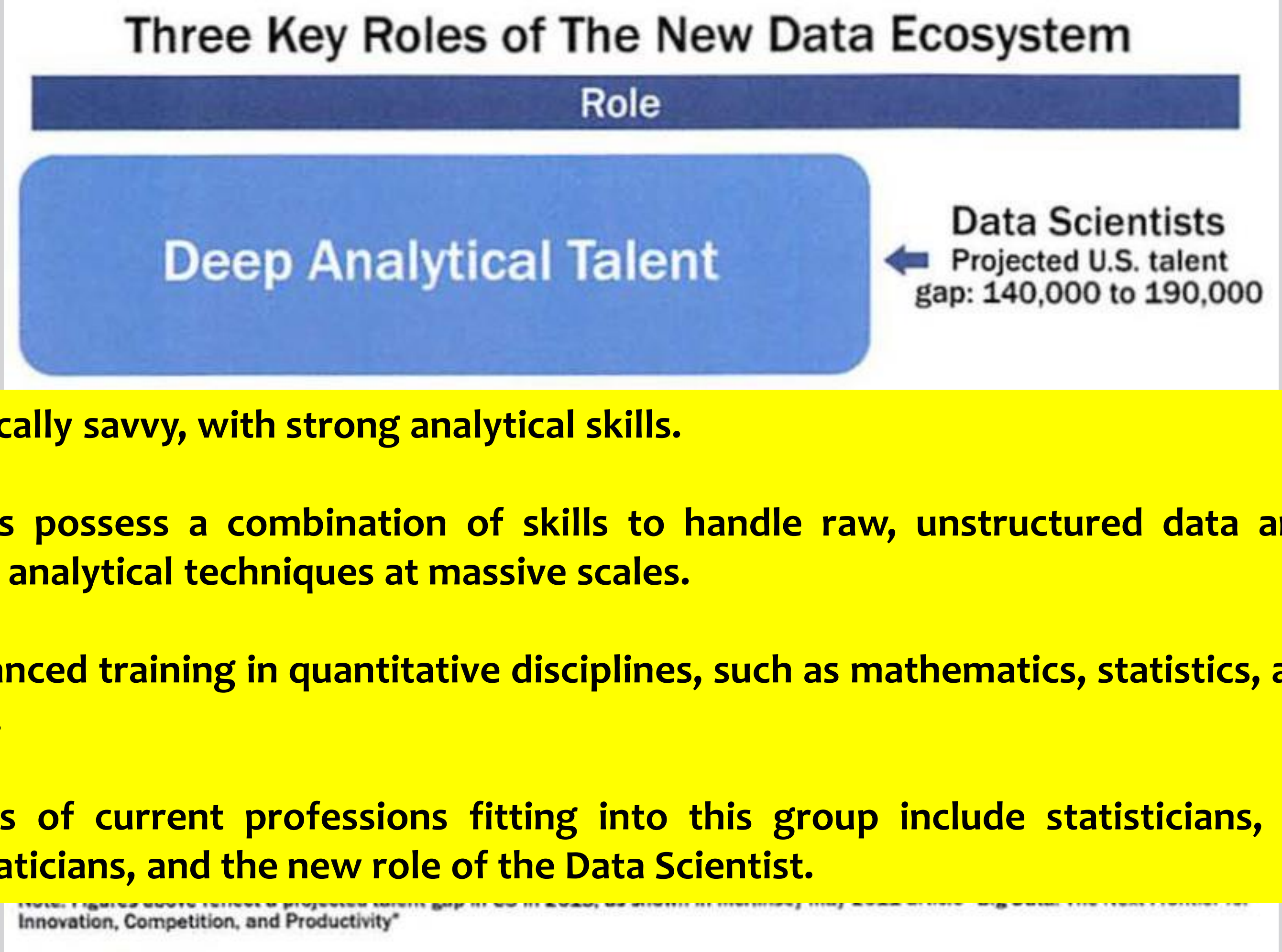
Deep Analytical Talent ← Data Scientists
Projected U.S. talent gap: 140,000 to 190,000

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 Big Data Ecosystem [3]



Three Key Roles of The New Data Ecosystem

Role

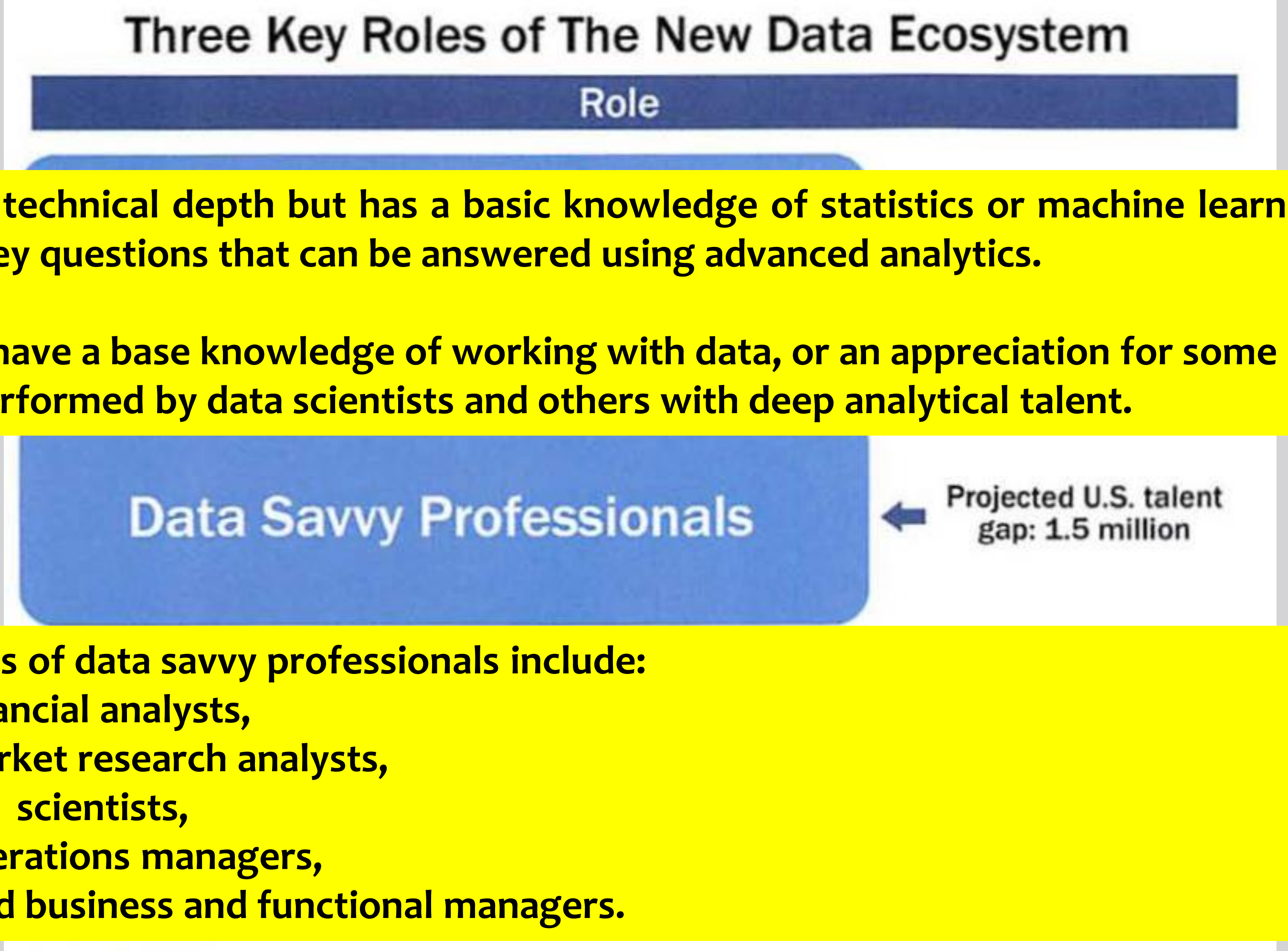
Deep Analytical Talent

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

- ➢ **Is technically savvy, with strong analytical skills.**
- ➢ **Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales.**
- ➢ **Has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.**
- ➢ **Examples of current professions fitting into this group include statisticians, economists, mathematicians, and the new role of the Data Scientist.**

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 Big Data Ecosystem [4]



Three Key Roles of The New Data Ecosystem

Role

Data Savvy Professionals

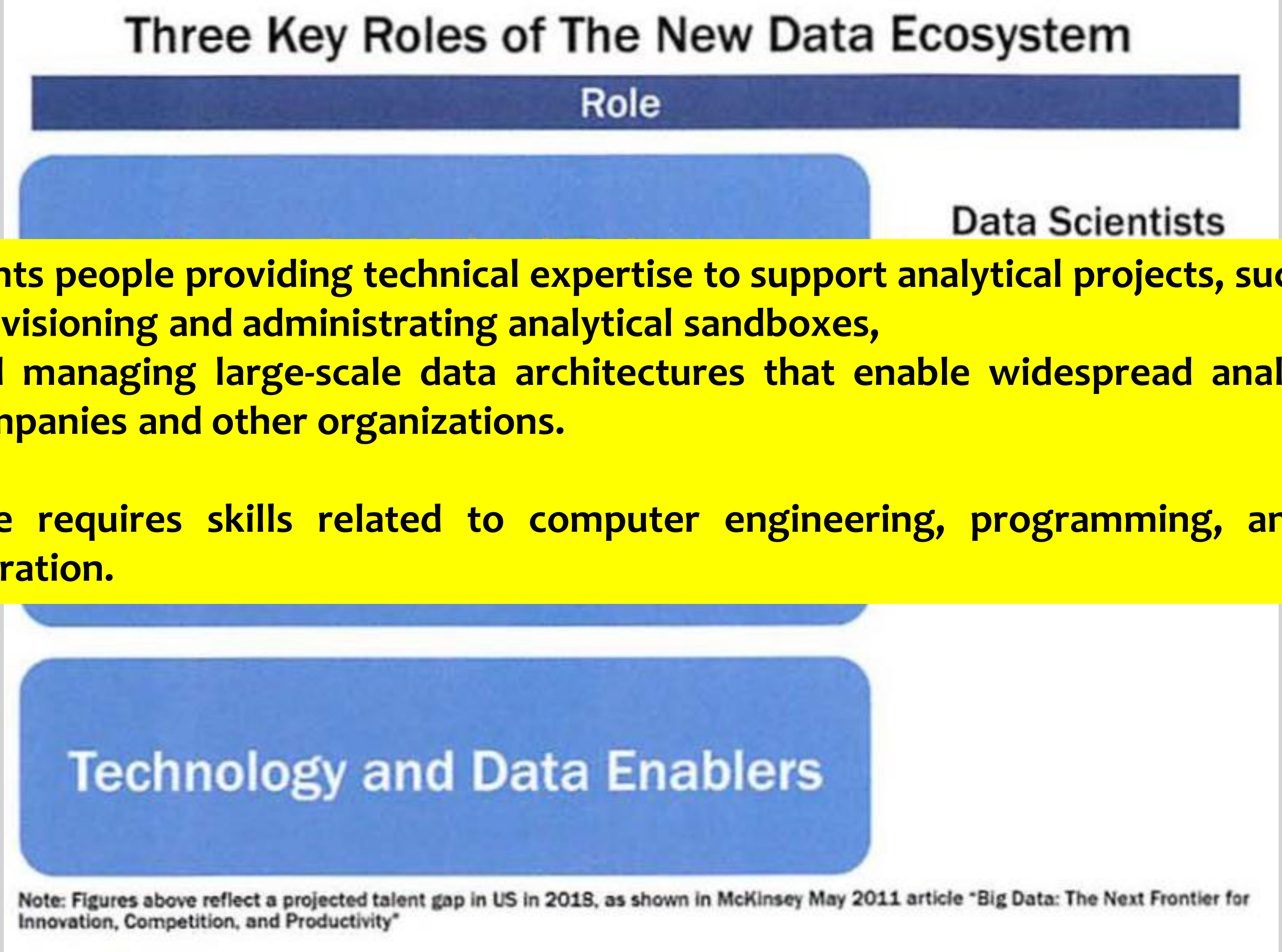
Projected U.S. talent gap: 1.5 million

- ➢ **Has less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics.**

- ➢ **Tend to have a base knowledge of working with data, or an appreciation for some of the work being performed by data scientists and others with deep analytical talent.**

- ➢ **Examples of data savvy professionals include:**
  - ➢ **financial analysts,**
  - ➢ **market research analysts,**
  - ➢ **life scientists,**
  - ➢ **operations managers,**
  - ➢ **and business and functional managers.**

Key Roles for the New Big Data Ecosystem [5]



Three Key Roles of The New Data Ecosystem

Role

Data Scientists

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"

- ➢ Represents people providing technical expertise to support analytical projects, such as
  - ➢ provisioning and administrating analytical sandboxes,
  - ➢ and managing large-scale data architectures that enable widespread analytics within companies and other organizations.

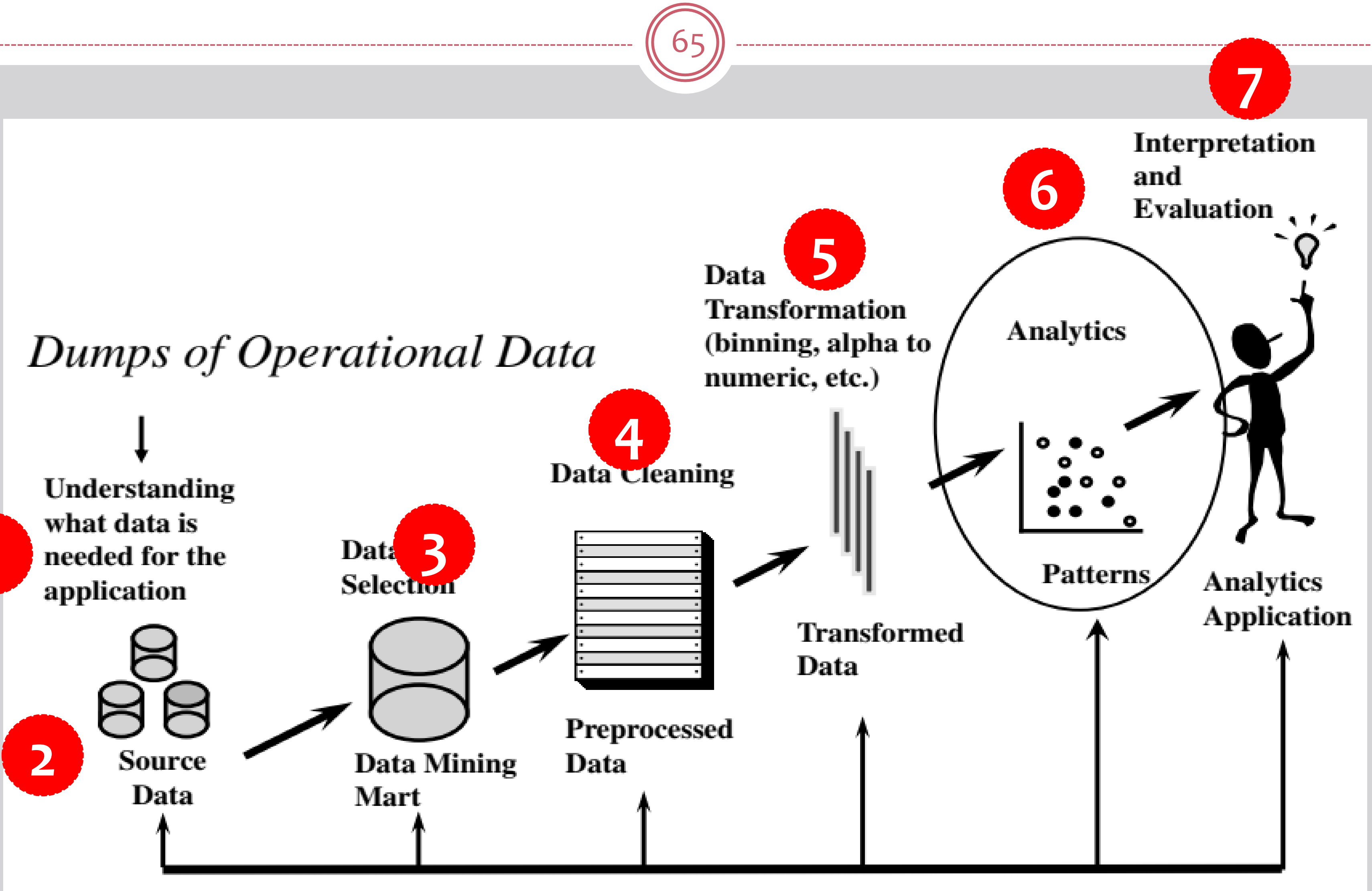
- ➢ This role requires skills related to computer engineering, programming, and database administration.

Key Roles for the New Big Data Ecosystem [6]

- In specific to Data Scientist, there are 3 recurring set of activities to 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.

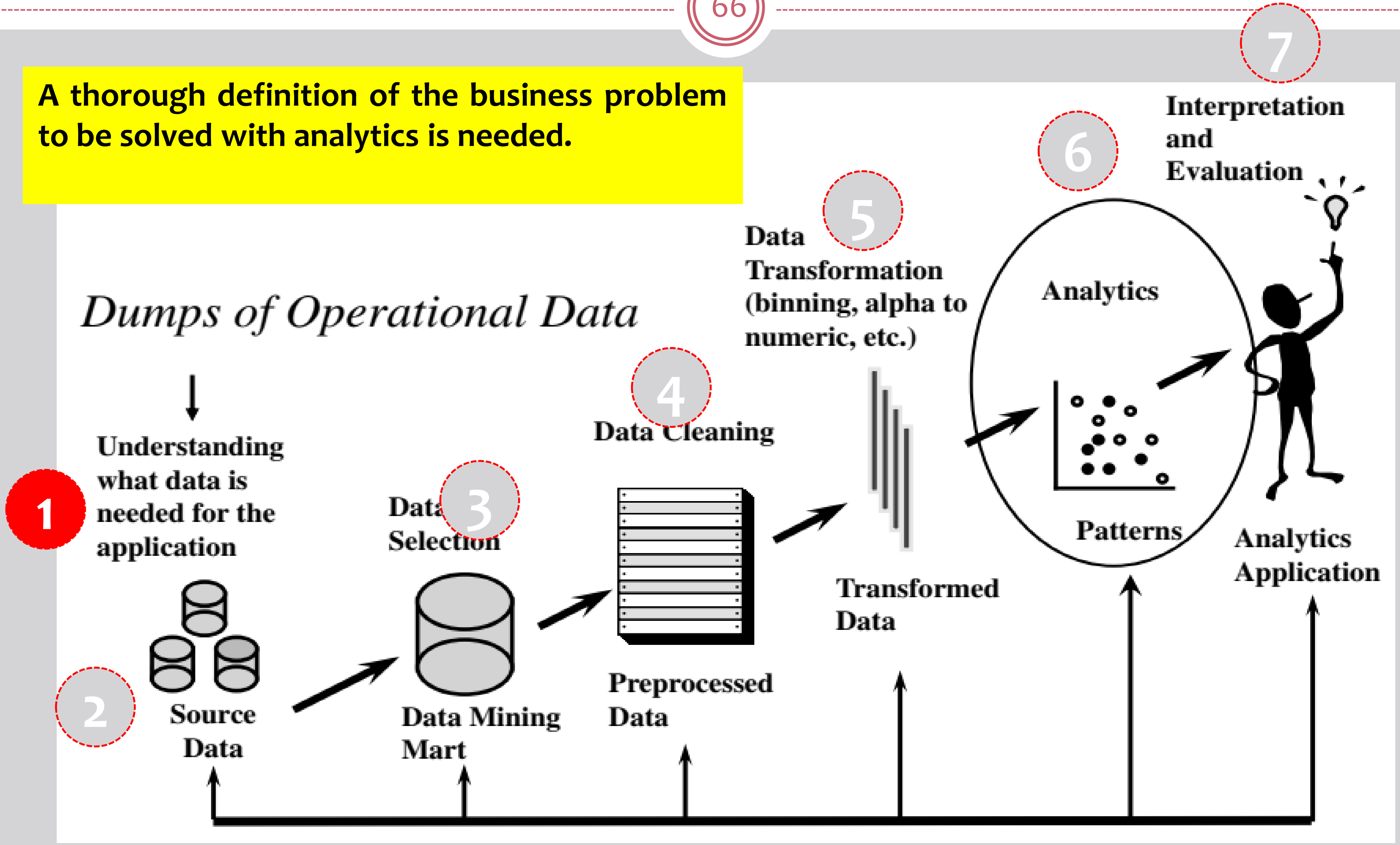
6. Analytics Process Model

Analytics Process Model [1]


Dumps of Operational Data
1
Understanding what data is needed for the application
2
Source Data
3
Data Selection
Data Mining Mart
4
Data Cleaning
Preprocessed Data
5
Data Transformation (binning, alpha to numeric, etc.)
Transformed Data
6
Analytics
Patterns
7
Interpretation and Evaluation
Analytics Application

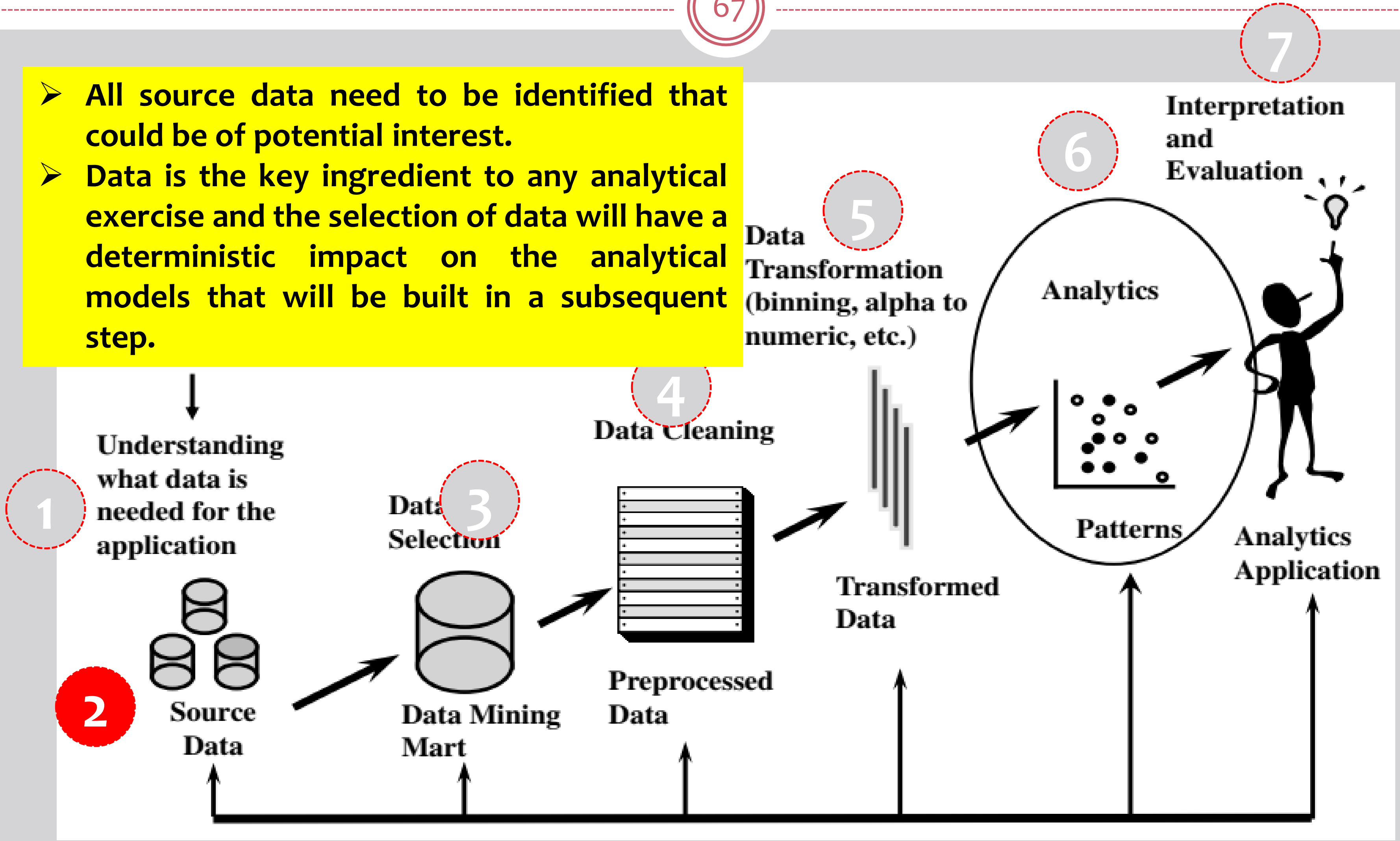
Analytics Process Model [2]

A thorough definition of the business problem to be solved with analytics is needed.


7
Interpretation and Evaluation
6
5
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Dumps of Operational Data
4
Data Cleaning
Understanding what data is needed for the application
1
Data Selection
3
Patterns
Analytics Application
Transformed Data
Preprocessed Data
2
Source Data
Data Mining Mart

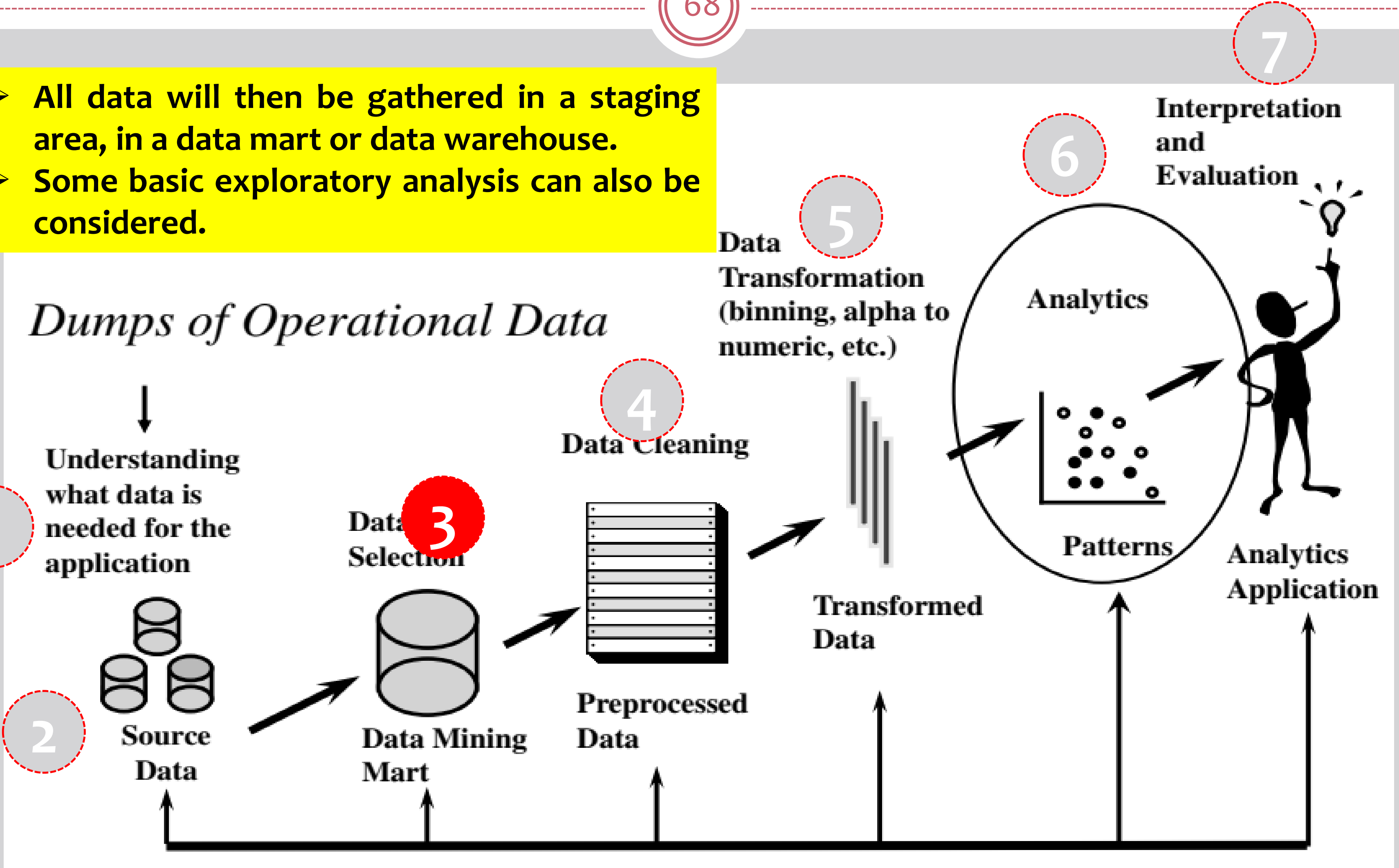
Analytics Process Model [3]

- All source data need to be identified that could be of potential interest.
- Data is the key ingredient to any analytical exercise and the selection of data will have a deterministic impact on the analytical models that will be built in a subsequent step.


7
Interpretation and Evaluation
6
5
Data Transformation (binning, alpha to numeric, etc.)
Analytics
4
Data Cleaning
Understanding what data is needed for the application
1
Data Selection
3
Patterns
Analytics Application
Transformed Data
2
Source Data
Data Mining Mart
Preprocessed Data

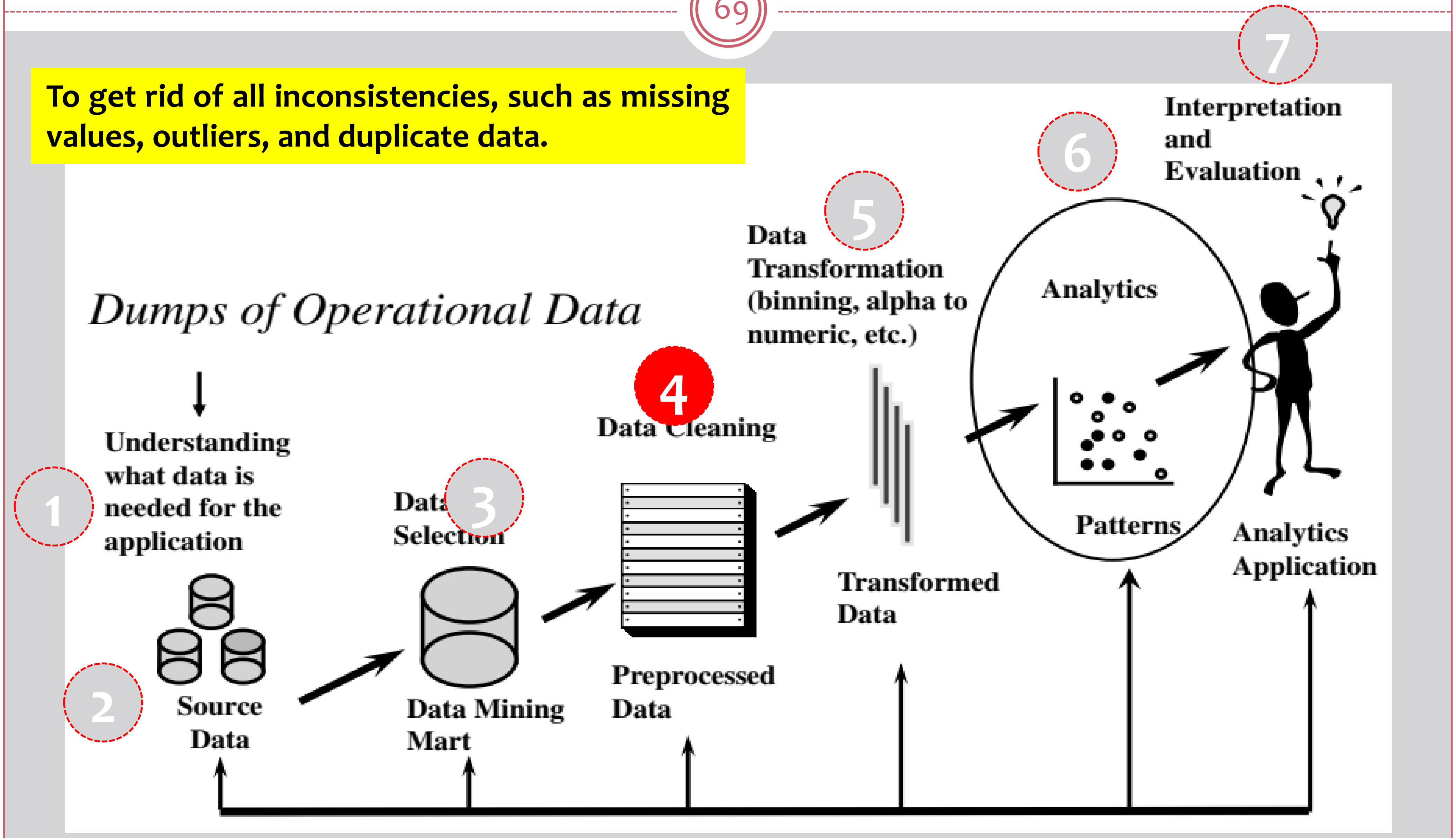
Analytics Process Model [4]

- All data will then be gathered in a staging area, in a data mart or data warehouse.
- Some basic exploratory analysis can also be considered.


7
Interpretation and Evaluation
6
5
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Dumps of Operational Data
4
Data Cleaning
Understanding what data is needed for the application
1
Data Selection
3
Patterns
Analytics Application
Transformed Data
Preprocessed Data
2
Source Data
Data Mining Mart

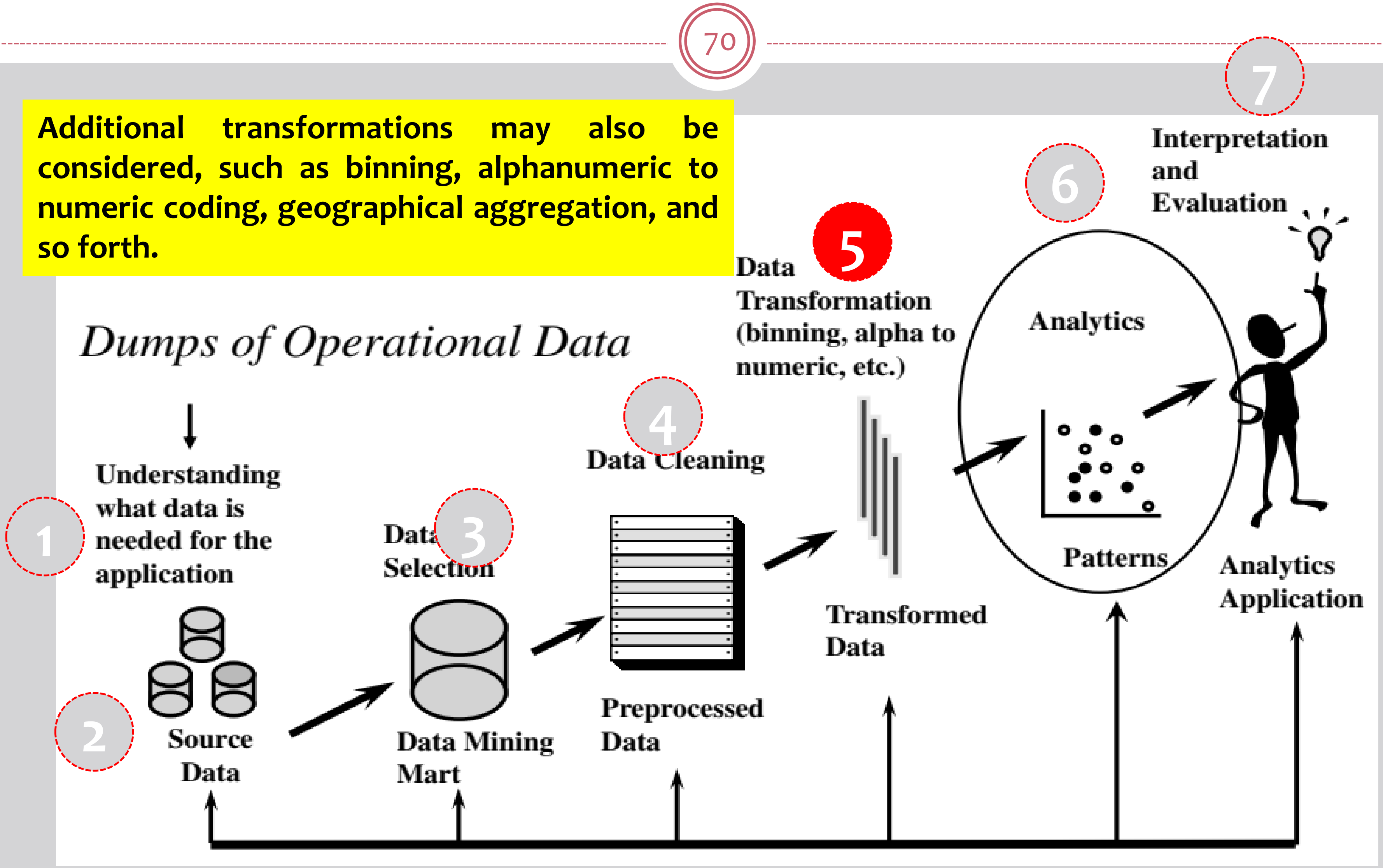
Analytics Process Model [5]

To get rid of all inconsistencies, such as missing values, outliers, and duplicate data.


7
Interpretation and Evaluation
6
5
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Dumps of Operational Data
4
Data Cleaning
Understanding what data is needed for the application
1
Data Selection
3
Patterns
Analytics Application
Transformed Data
2
Source Data
Data Mining Mart
Preprocessed Data

Analytics Process Model [6]

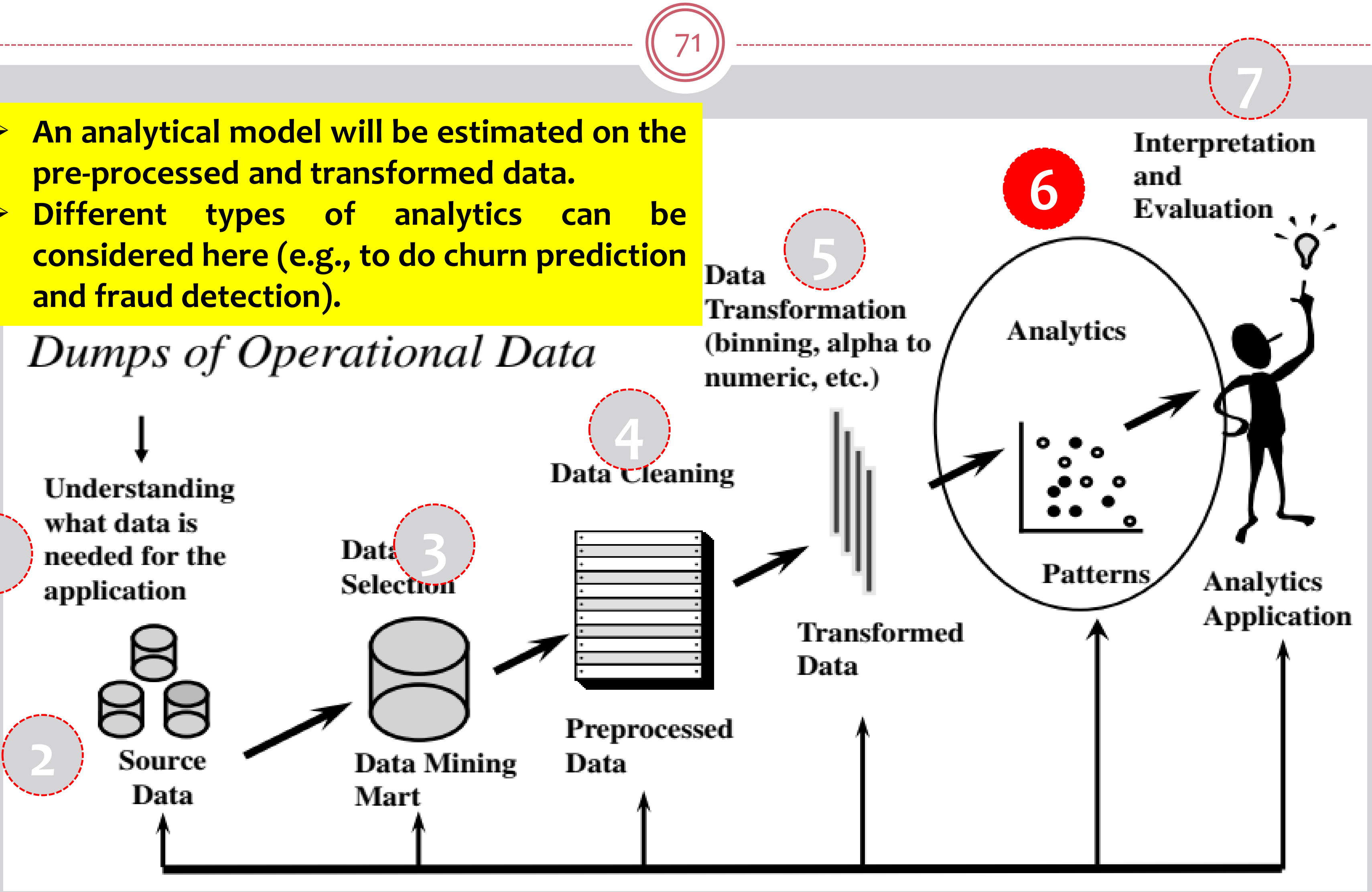
Additional transformations may also be considered, such as binning, alphanumeric to numeric coding, geographical aggregation, and so forth.



Dumps of Operational Data
1
Understanding what data is needed for the application
2
Source Data
3
Data Selection
Data Mining Mart
4
Data Cleaning
Preprocessed Data
5
Data Transformation (binning, alpha to numeric, etc.)
Transformed Data
6
Analytics
Patterns
7
Interpretation and Evaluation
Analytics Application

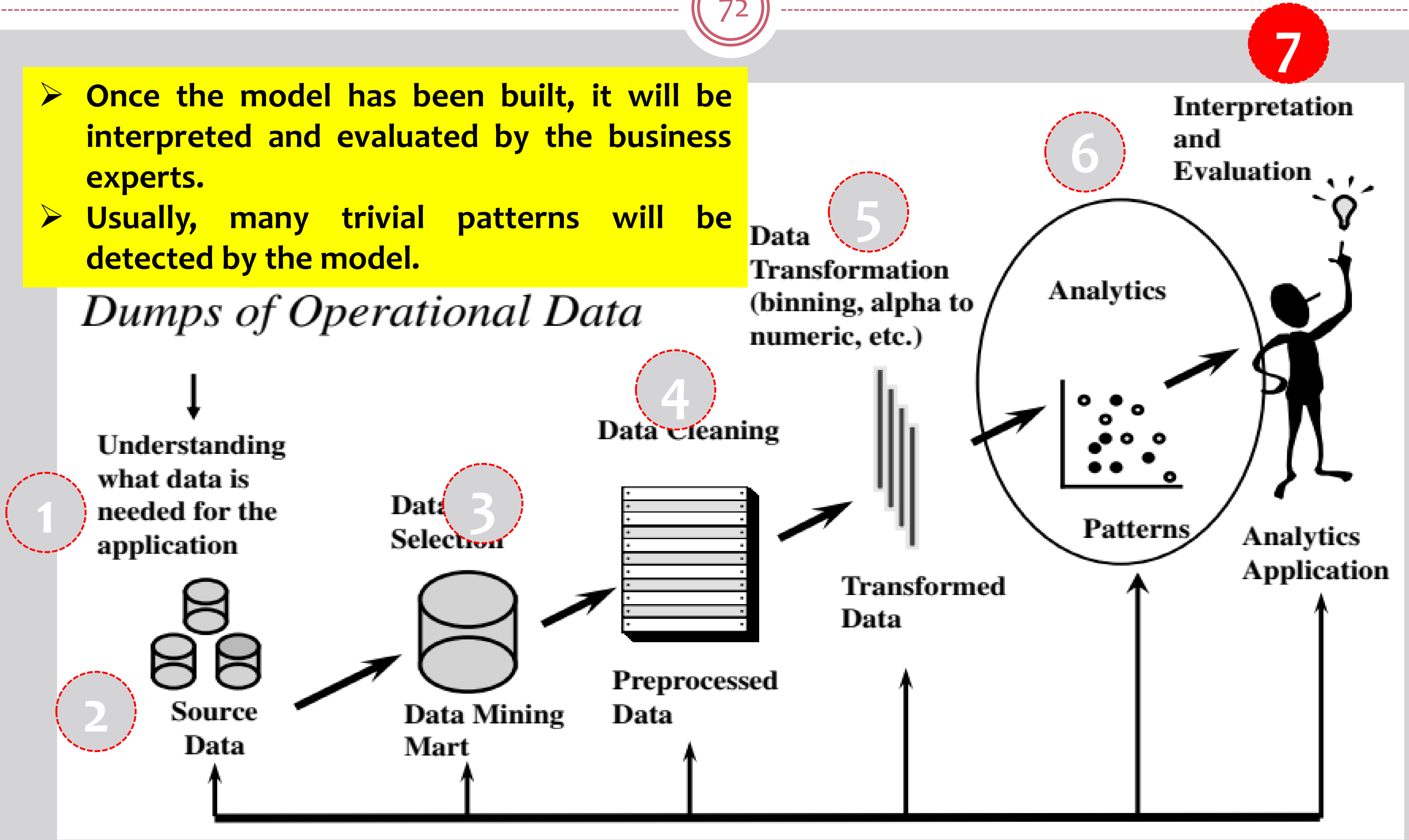
Analytics Process Model [7]

- An analytical model will be estimated on the pre-processed and transformed data.
- Different types of analytics can be considered here (e.g., to do churn prediction and fraud detection).


Dumps of Operational Data
1
Understanding what data is needed for the application
2
Source Data
3
Data Selection
Data Mining Mart
4
Data Cleaning
Preprocessed Data
5
Data Transformation (binning, alpha to numeric, etc.)
Transformed Data
6
Analytics
Patterns
7
Interpretation and Evaluation
Analytics Application

Analytics Process Model [8]

- Once the model has been built, it will be interpreted and evaluated by the business experts.
- Usually, many trivial patterns will be detected by the model.


7
Interpretation and Evaluation
6
5
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Dumps of Operational Data
4
Data Cleaning
Understanding what data is needed for the application
1
Data Selection
3
Patterns
Analytics Application
Transformed Data
Preprocessed Data
2
Source Data
Data Mining Mart

Analytics Process Model [9]

- This process model is iterative in nature.
- For example, during the analytics step, the need for additional data may be identified…
  - which may necessitate additional cleaning, transformation, and so forth.
- The most time consuming step is…
- the data selection and pre-processing step;
- this usually takes around 80% of the total efforts needed to build an analytical model.

7. Job Profiles Involved

Job Profiles [1]

- Analytics is essentially a multidisciplinary exercise in which many different job profiles need to collaborate together.

1. Database or Data Warehouse Administrator (DBA):

- Aware of all the data available within the firm, the storage details, and the data definitions.
- Hence, the DBA plays a crucial role in feeding the analytical modeling exercise with its key ingredient, which is data.
- Because analytics is an iterative exercise, the DBA may continue to play an important role as the modeling exercise proceeds.

Job Profiles [2]

2. Business Expert:

- For example, be a credit portfolio manager, fraud detection expert, brand manager, or e-commerce manager.
- This person has extensive business experience and business common sense, which is very valuable.
- It is precisely this knowledge that will help to steer the analytical modeling exercise and interpret its key findings.

Job Profiles [3]

3. Legal Expert:

- This job profile is becoming more and more important given that not all data can be used in an analytical model because of privacy, discrimination, and so forth.
- For examples:
  - In credit risk modeling, one can typically not discriminate good and bad customers based upon gender, national origin, or religion.
  - In web analytics, information is typically gathered by means of cookies, which are filesles that are stored on the user's browsing computer. However, when gathering information using cookies, users should be appropriately informed.

Job Profiles [4]

4. Data Scientist, Data Miner or Data Analyst:

- Responsible for doing the actual analytics.
- Possesses a thorough understanding of all techniques involved…
- and know how to implement them using the appropriate software.
- Should also have good communication and presentation skills to report the analytical findings back to the other parties involved.

Job Profiles [5]

5. Software Tool Vendors:

- Some vendors only provide tools to automate specific steps of the analytical modeling process (e.g., data pre-processing).
- Others sell software that covers the entire analytical modeling process.
- Some vendors also provide analytics-based solutions for specific application areas, such as:
  - risk management,
  - marketing analytics and
  - campaign management, and so on.

8. Analytics

Analytics [1]

- Analytics is a term that is often used interchangeably with data science, data mining, knowledge discovery, and others.
- The distinction between all those is not clear cut.
- All of these terms essentially refer to extracting useful business patterns or mathematical decision models from a pre-processed data set.

Analytics [2]

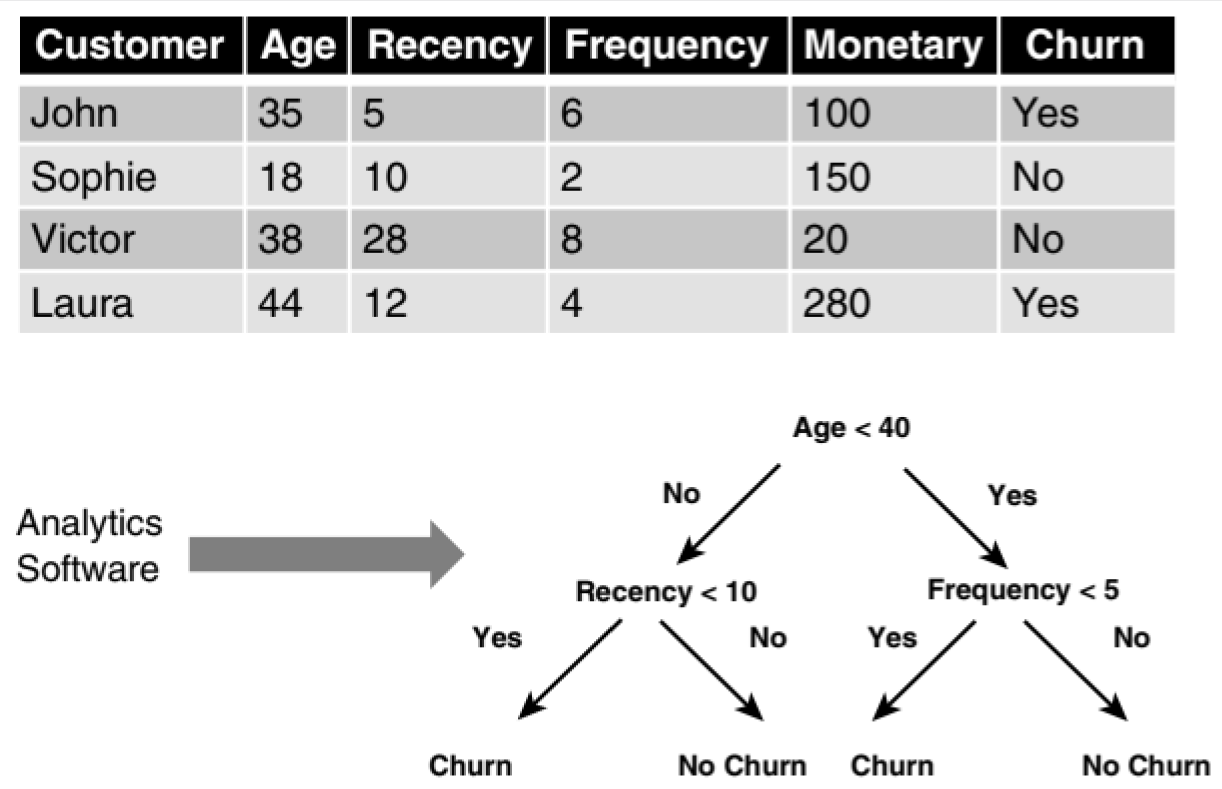
- Different underlying techniques can be used for this purpose…
- stemming from a variety of different disciplines, such as:
  - Statistics (e.g., linear and logistic regression)
  - Machine learning (e.g., decision trees)
  - Biology (e.g., neural networks, genetic algorithms, swarm intelligence)
  - Kernel methods (e.g., support vector machines)

Analytics [3]

- Basically, a distinction can be made between predictive and descriptive analytics.
- A target variable is typically available, which can either be categorical (e.g., churn or not, fraud or not) or continuous (e.g., customer lifetime value, loss given default).
- No such categorical target variable for descriptive analytics.

Analytics [4]

- Common examples for predictive analysis:
  - Association rules
  - Sequence rules
  - Clustering



Customer	Age	Recency	Frequency	Monetary	Churn
John	35	5	6	100	Yes
Sophie	18	10	2	150	No
Victor	38	28	8	20	No
Laura	44	12	4	280	Yes

An example of a decision tree in a classification predictive analytics setting for predicting churn.

9. Analytical Model Requirements

Analytical Model Requirements [1]

- A good analytical model should satisfy several requirements, depending on the application area.
- So that the analytical models developed are in the most optimal way.
- There are 7 common requirements.

Analytical Model Requirements [2]

1. Business relevance:

- The analytical model should actually solve the business problem for which it was developed.
- It makes no sense to have a working analytical model that got side-tracked from the original problem statement.
- In order to achieve business relevance, it is of key importance that the business problem to be solved is…
  - appropriately defined,
  - qualified, and a
  - greed upon by all parties involved at the outset of the analysis.

Analytical Model Requirements [3]

2. Statistical performance:

- The model should have statistical significance and predictive power.
- How this can be measured will depend upon the type of analytics considered.
- For examples:
  - In a classification setting (churn, fraud), the model should have good discrimination power.
  - In a clustering setting, the clusters should be as homogenous as possible.

Analytical Model Requirements [4]

3. Interpretability:

- Refers to understanding the patterns that the analytical model captures.
- This aspect has a certain degree of subjectivism, since interpretability may depend on the business user's knowledge.
- In many settings, however, it is considered to be a key requirement.
- For example, in credit risk modeling or medical diagnosis…
  - interpretable models are absolutely needed to get good insight into the underlying data patterns.

Analytical Model Requirements [5]

4. Justifiability:

- Refers to the degree to which a model corresponds to prior business knowledge and intuition.
- For example, a model stating that a higher debt ratio results in more creditworthy clients may be interpretable…
  - but is not justifiable because it contradicts basic financial intuition.

Analytical Model Requirements [6]

- Note that both interpretability and justifiability often need to be balanced against statistical performance.
- Often one will observe that high performing analytical models are incomprehensible and black box in nature.
- A popular example of this is neural networks…
  - which are universal approximators and are high performing,
  - but offer no insight into the underlying patterns in the data.
- On the contrary…
  - linear regression models are very transparent and comprehensible,
  - but offer only limited modeling power.

Analytical Model Requirements [7]

5. Operational efficient:

- Refers to the efforts needed to:
  - collect the data,
  - pre-process it,
  - evaluate the model, and
  - feed its outputs to the business application (e.g., campaign management, capital calculation).
- Especially in a real-time online scoring environment (e.g., fraud detection) this may be a crucial characteristic.
- Operational efficiency also entails the efforts needed to monitor and back-test the model, and re-estimate it when necessary.

Analytical Model Requirements [8]

6. Economic cost:

- Refers to the inclusion of the costs to:
  - gather and pre-process the data,
  - the costs to analyze the data, and
  - the costs to put the resulting analytical models into production.
- In addition, the software costs and human and computing resources should be taken into account here.
- It is important to do a thorough cost–benefit analysis at the start of the project.

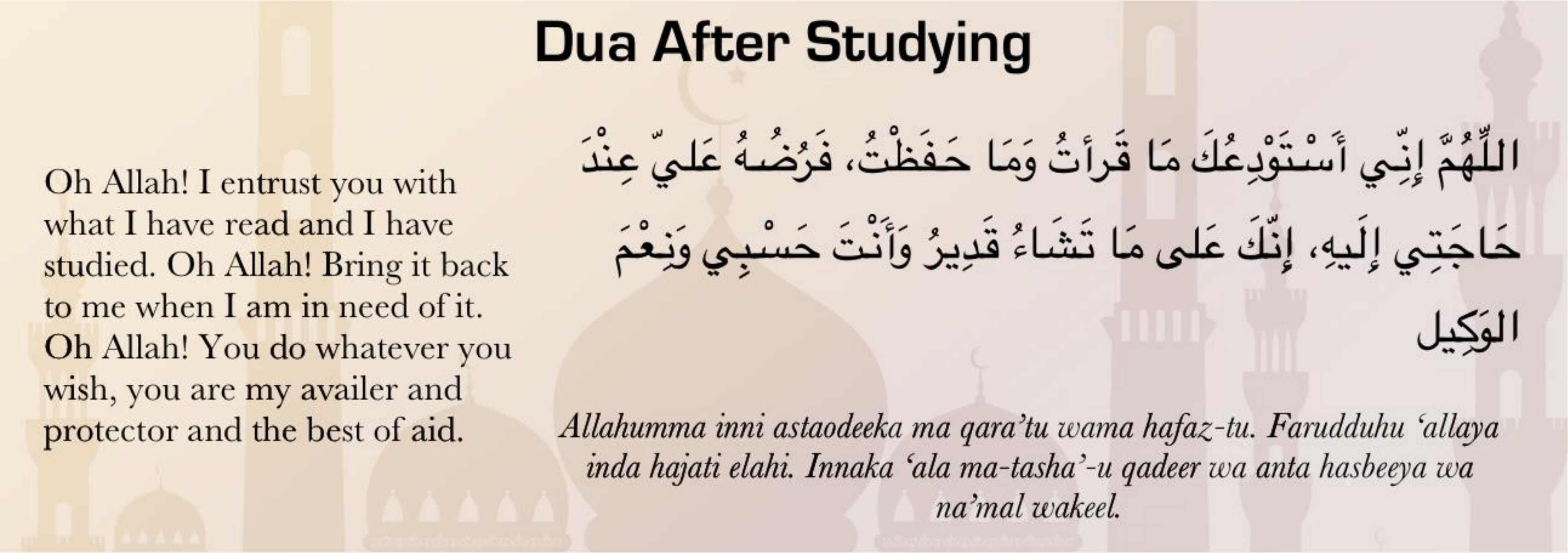
Analytical Model Requirements [9]

7. International regulation and legislation:

- The analytical models should also comply with both local and international regulation and legislation.
- For example:
  - In a credit risk setting, the Basel II and Basel III Capital Accords have been introduced to appropriately identify the types of data that can or cannot be used to build credit risk models.
  - In an insurance setting, the Solvency II Accord plays a similar role.
- Additionally, in the context of privacy, many new regulatory developments are taking place at various levels.
- A popular example here concerns the use of cookies in a web analytics context.

Dua After Studying

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ



Dua After Studying

Oh Allah! I entrust you with what I have read and I have studied. Oh Allah! Bring it back to me when I am in need of it. Oh Allah! You do whatever you wish, you are my availer and protector and the best of aid.

اللِّهُمَّ إِنّي أَسْتَوْدِعُكَ مَا قَرأتُ وَمَا حَفَظْتُ، فَرُضُهُ عَليّ عِنْدَ
حَاجَتِي إِلَيهِ، إِنَّكَ عَلى مَا تَشَاءُ قَدِيرُ وَأَنْتَ حَسْبِي وَنِعْمَ
الوَكِيل

Allahumma inni astaodeeka ma qara'tu wama hafaz-tu. Farudduhu 'allaya inda hajati elahi. Innaka 'ala ma-tasha'-u qadeer wa anta hasbeeya wa na'mal wakeel.