II SEMESTER MCA

MCA 4251

DATA ANALYTICS

Poornima P Kundapur
Department of Data Science and Computer Applications
MIT, Manipal 576104


WELCOME BACK.
SECTION B

SYLLABUS for MCA 4251

Introduction: data science, need for analytics, steps in data analysis projects, **Data:** sources of data, data sets, data warehouses, data types, privacy and confidentiality, samples vs. population, **Data summarization and visualization**: tables and graphs, **Data Preprocessing:** cleaning, transformation, dimensionality reduction, **Data Analysis and Visualization:** descriptive, inferential statistics, uni-variate and multi-variate analysis, **Grouping:** Cluster Analysis: distance measures, partitioning, hierarchical, density based methods, Market Basket Analysis, Association Analysis, Market Basket Analysis, **Classifiers:** Bayesian, k-nearest neighbor, neural network, Support Vector Machine, Decision Trees, **Prediction**: Regression models, Evaluating Classification and Predictive performance, ensemble methods, Anomaly Detection, Forecasting models, **Applications in Data Analytics:** Case studies, Web Mining, Text Mining, Business Intelligence, Supply Chain Analytics, Time series, Spatial Data Analysis

DO WE HAVE A LAB?

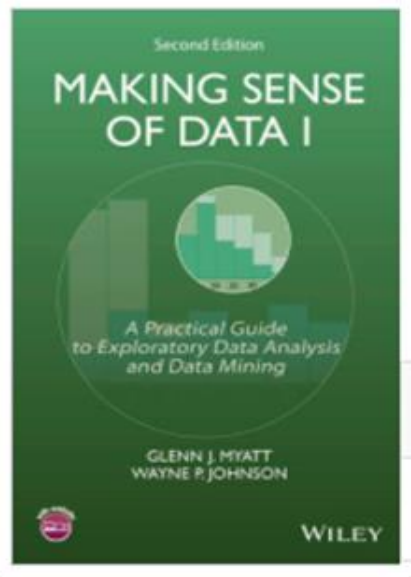
CLASS ROOM ENGAGEMENT.

- **Second Semester. Not new.**
- All laptop computers, mobile phones, tablet computers must be closed during all classroom hours.
- Computers distract the most people behind and around the user.
- ~~**Maintain social distance**~~ **and wear masks *(correctly)* at all times inside and outside class.**
- Make your own notes. Slides are not enough.
- Self study is paramount.
- **All homework** to be **completed** before class commences.
- **ASK QUESTIONS.**

UNLEARNING AND RELEARNING.

- Look at the notes/slides (IF ANY) before class.
- Attend all lectures (try IT, you will see the difference).
- Review the lecture during the evening.
- Rewrite and summarize the slides in your words
- Practice all examples and problems solved in class.
- Get a good night’s rest.

REFERENCES.

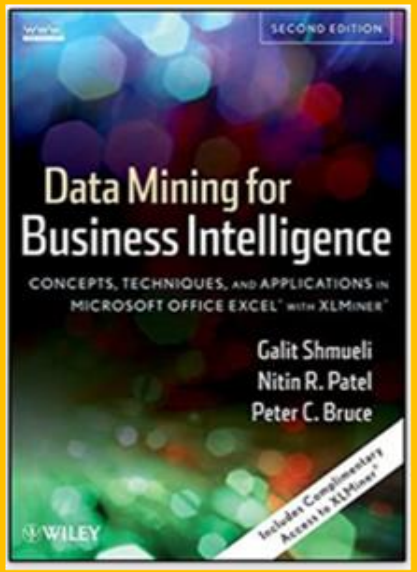


GLENN J. MYATT

Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John Wiley, November 2014

GLENN J. MYATT

Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications

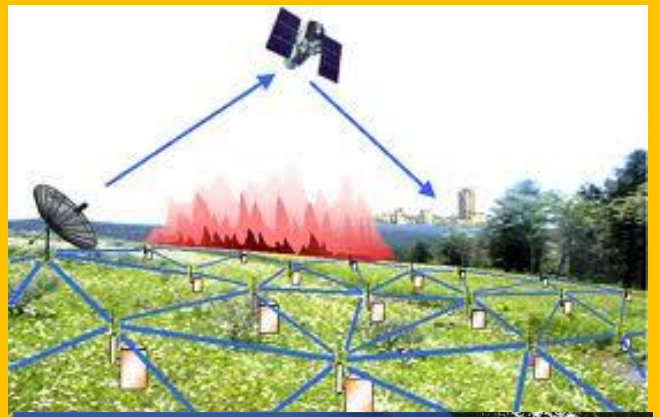


G. SHMUELI, N. R. PATEL, AND P.C. BRUCE

Data Mining for Business Intelligence, John Wiley and Sons, 2010

DATA ANALOGY.

Social media and networks
(All of us are generating data)

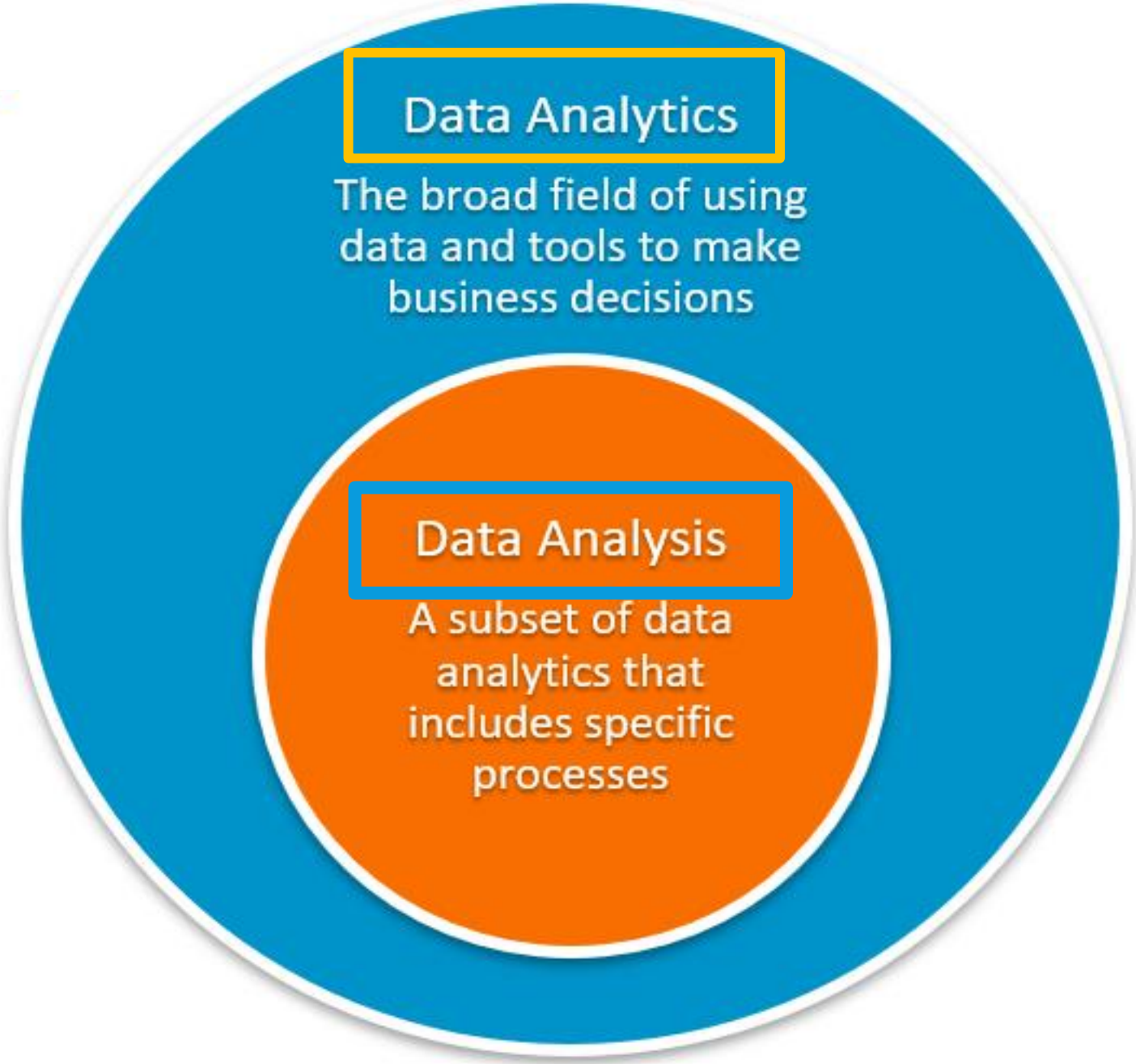


Scientific instruments
(Collecting all sorts of data)

Mobile devices
(Tracking all objects all the time)

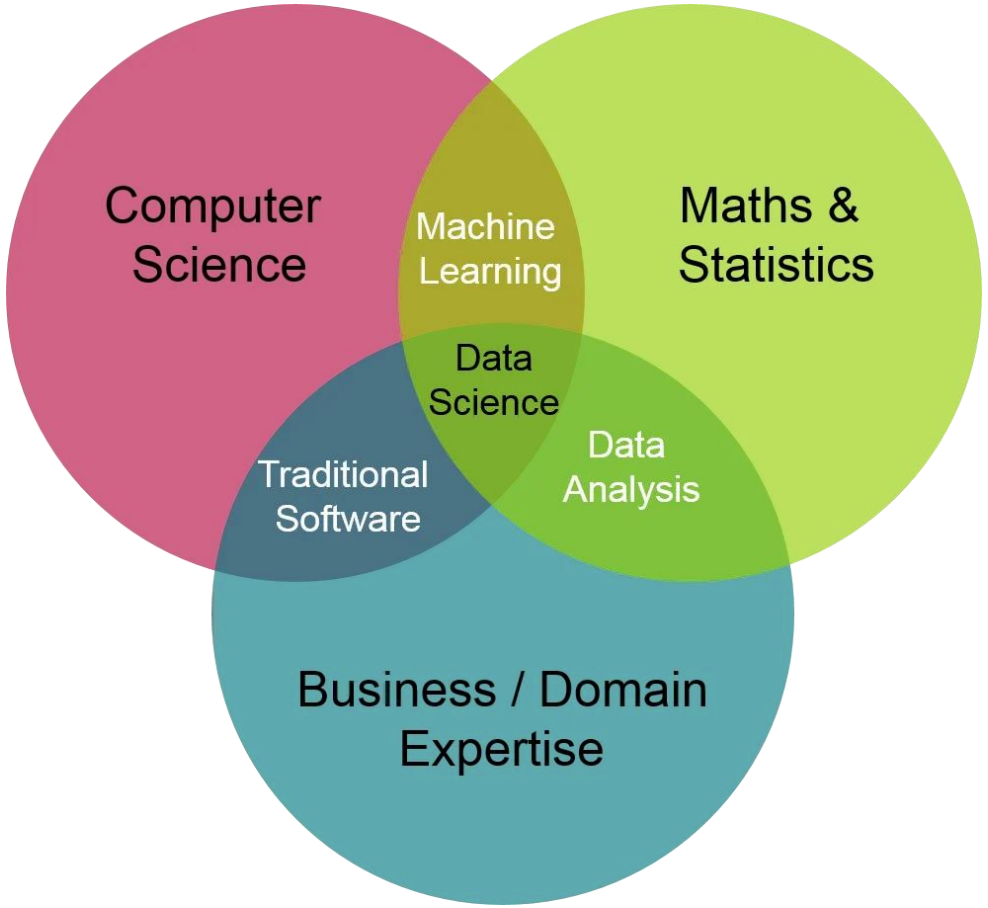
Sensor technology and networks
(Measuring all kinds of data)

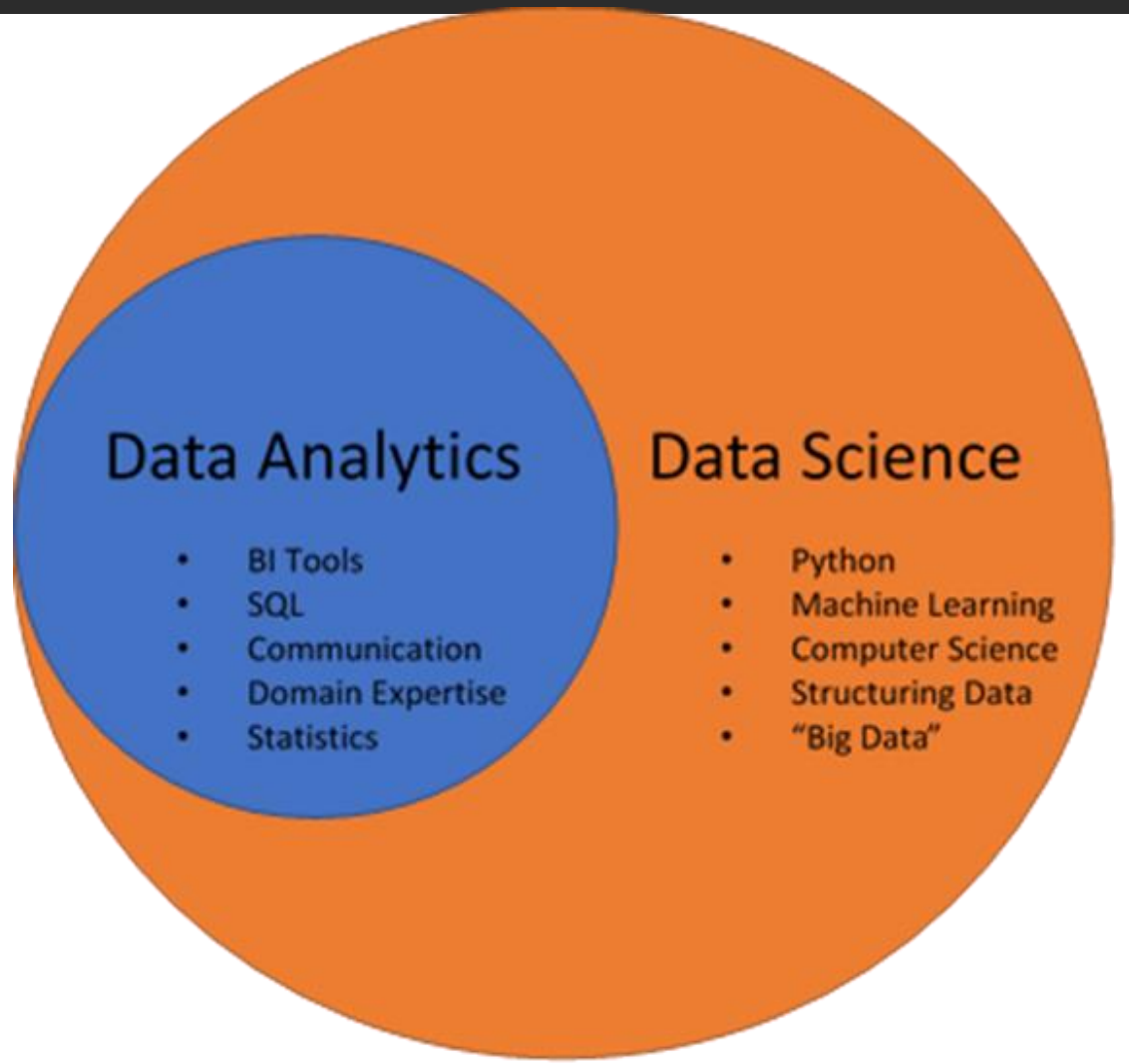

BI
Backward Looking (Reactive)
Primarily answers predefined questions like
What happened, Why did it happen, and What is the current trend?
METHODS
• Reporting
• Automated Monitoring
• Ad Hoc Query
• OLAP
• Dashboards and Scorecards
DA
Forward Looking (Proactive)
Used to derive new insights by asking questions like
What will happen What should we do, and What will be the future trend?
METHODS
• Predictive Modelling
• Data Mining
• Statistical/Quantitative Modelling
• Simulation and Optimization


Data Analytics
The broad field of using data and tools to make business decisions
Data Analysis
A subset of data analytics that includes specific processes

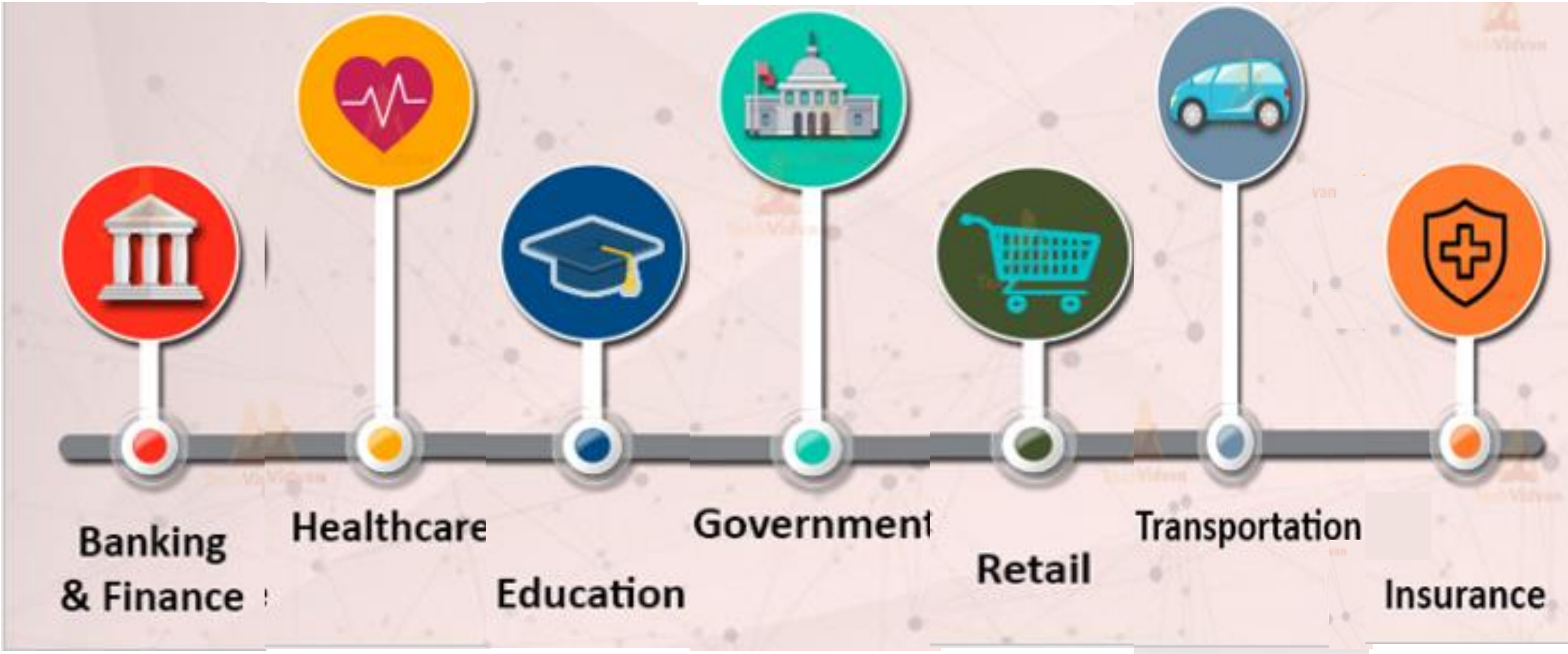
DATA ANALYTICS vs DATA SCIENCE

DATA ANALYTICS VS DATA SCIENCE.

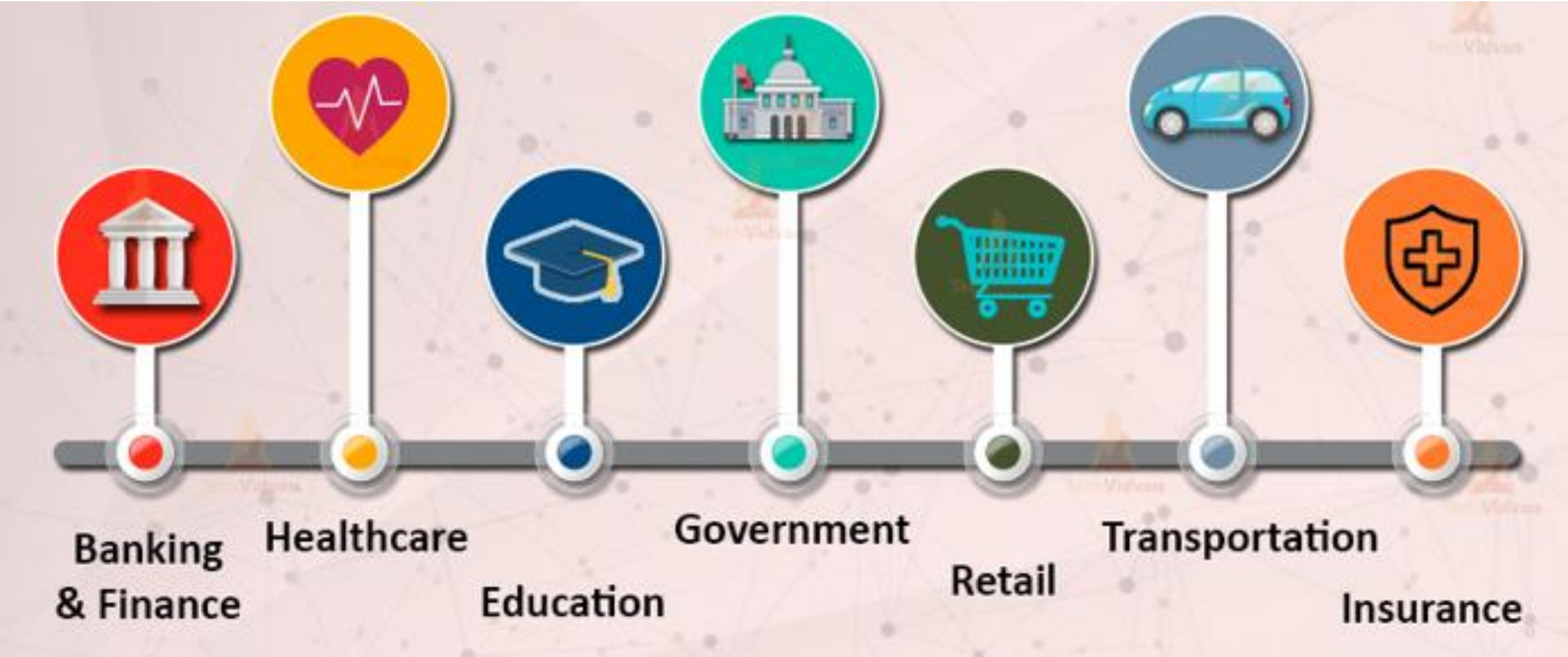




THE OVERVIEW.

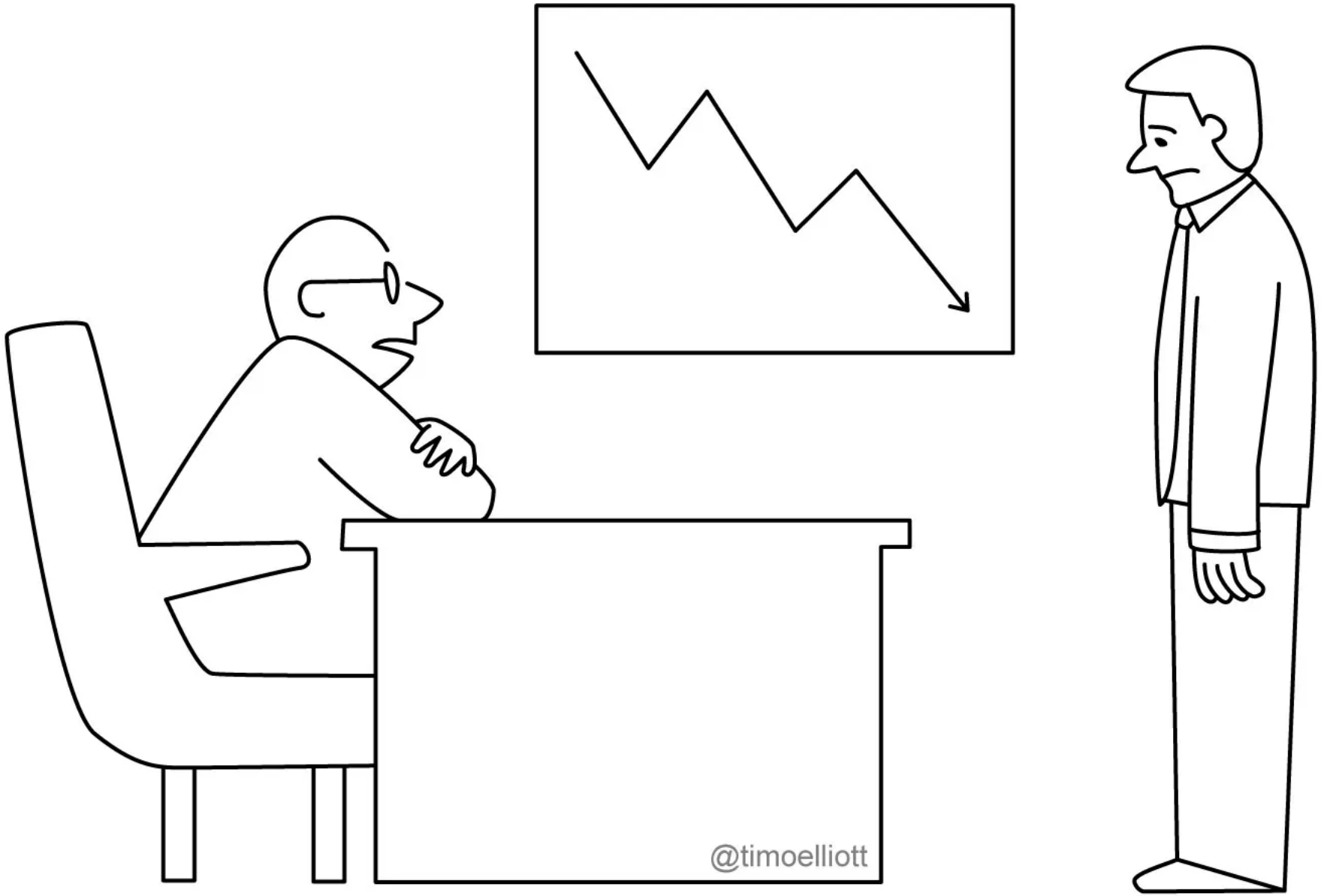


THE OVERVIEW.

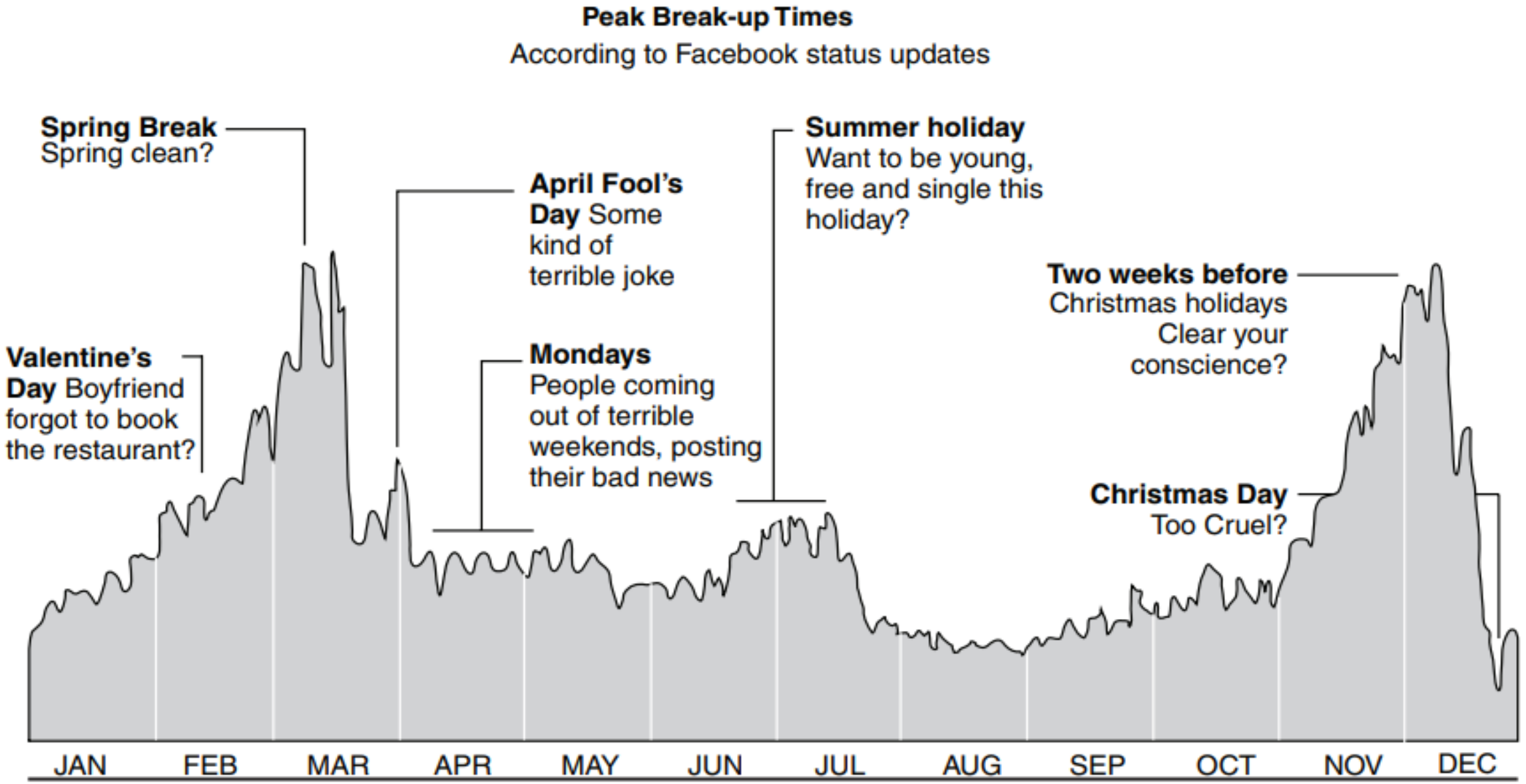


NEED FOR MAKING SENSE OF DATA.

- Unprecedented amount of data is being generated
  - Data rich but information poor situation
  - Data repositories becomes Data tombs
  - Decision making was based on intuition rather than information
  - Expert systems rely on domain experts to manually input system knowledge to knowledge bases
- Leads to information overload
- Need for making sense of the data
  - Analysis of data includes
    - Summarizing and interpret the data
    - how to identify nontrivial facts, patterns, and relationships in the data
    - how to make predictions from the data.



"It would appear, Hopkins, that your gut feel was only indigestion"


Peak Break-up Times
According to Facebook status updates
Spring Break
Spring clean?
April Fool's Day Some kind of terrible joke
Summer holiday
Want to be young, free and single this holiday?
Two weeks before
Christmas holidays
Clear your conscience?
Valentine's Day Boyfriend forgot to book the restaurant?
Mondays
People coming out of terrible weekends, posting their bad news
Christmas Day
Too Cruel?
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
David McCandless & Lee Byron
InformationIsBeautiful.net / LeeByron.com
Source: Facebook Lexicon 2008

WHAT IS DATA ANALYTICS

Analytics is the **systematic** computational **analysis of data or statistics**. It is used for the **discovery**, **interpretation**, and **communication** of **meaningful patterns** in data. It also entails applying data patterns towards effective **decision-making**. It can be valuable in areas rich with recorded information; analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. **Wikipedia**

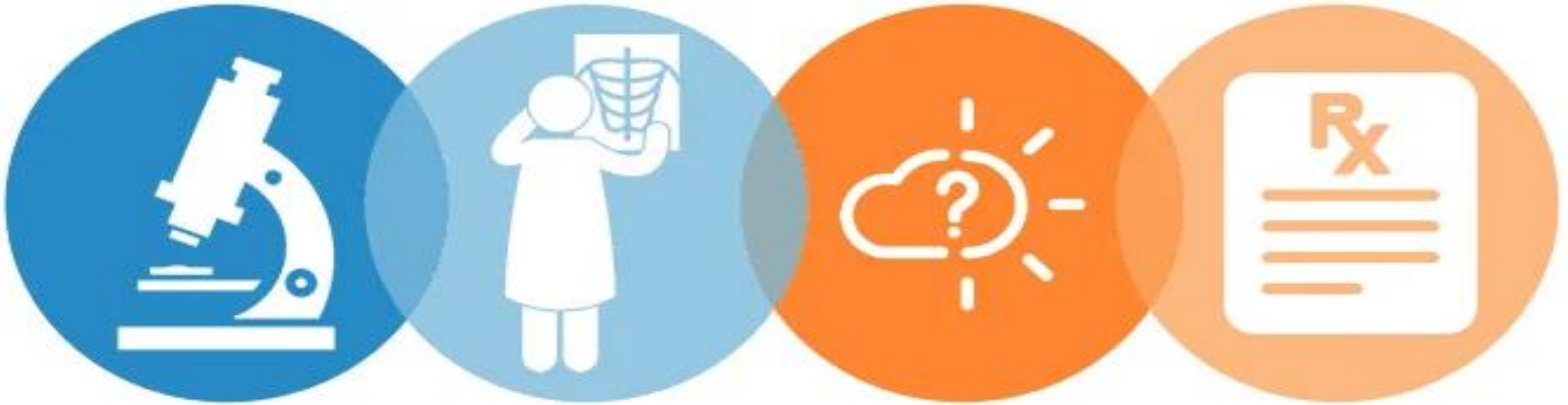
Data analytics enables organizations to analyze all their data (**real-time, historical, unstructured, structured, qualitative**) to identify patterns and **generate insights** to inform and, in some cases**, automate decisions**, connecting intelligence and action.

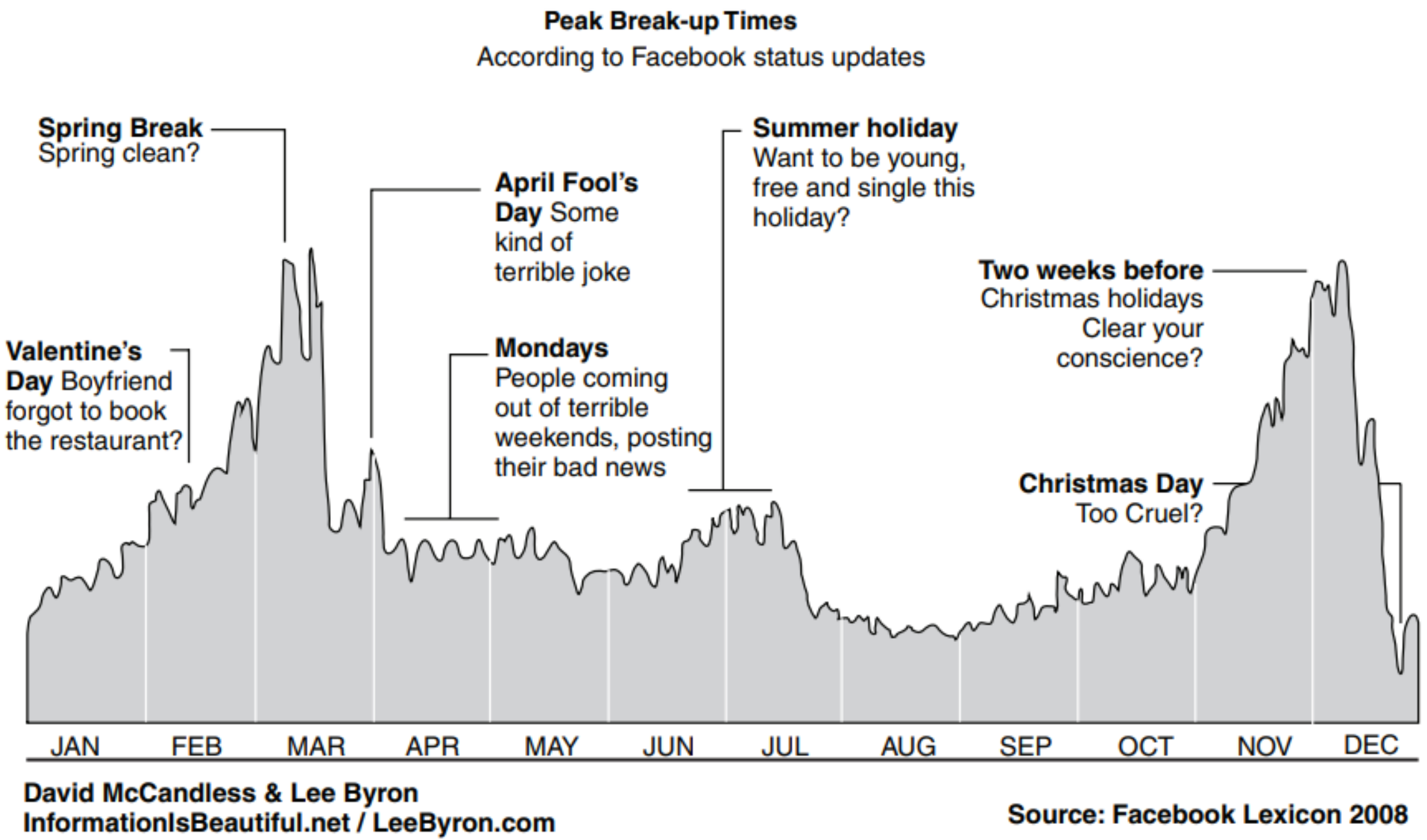
WHAT IS DATA ANALYTICS

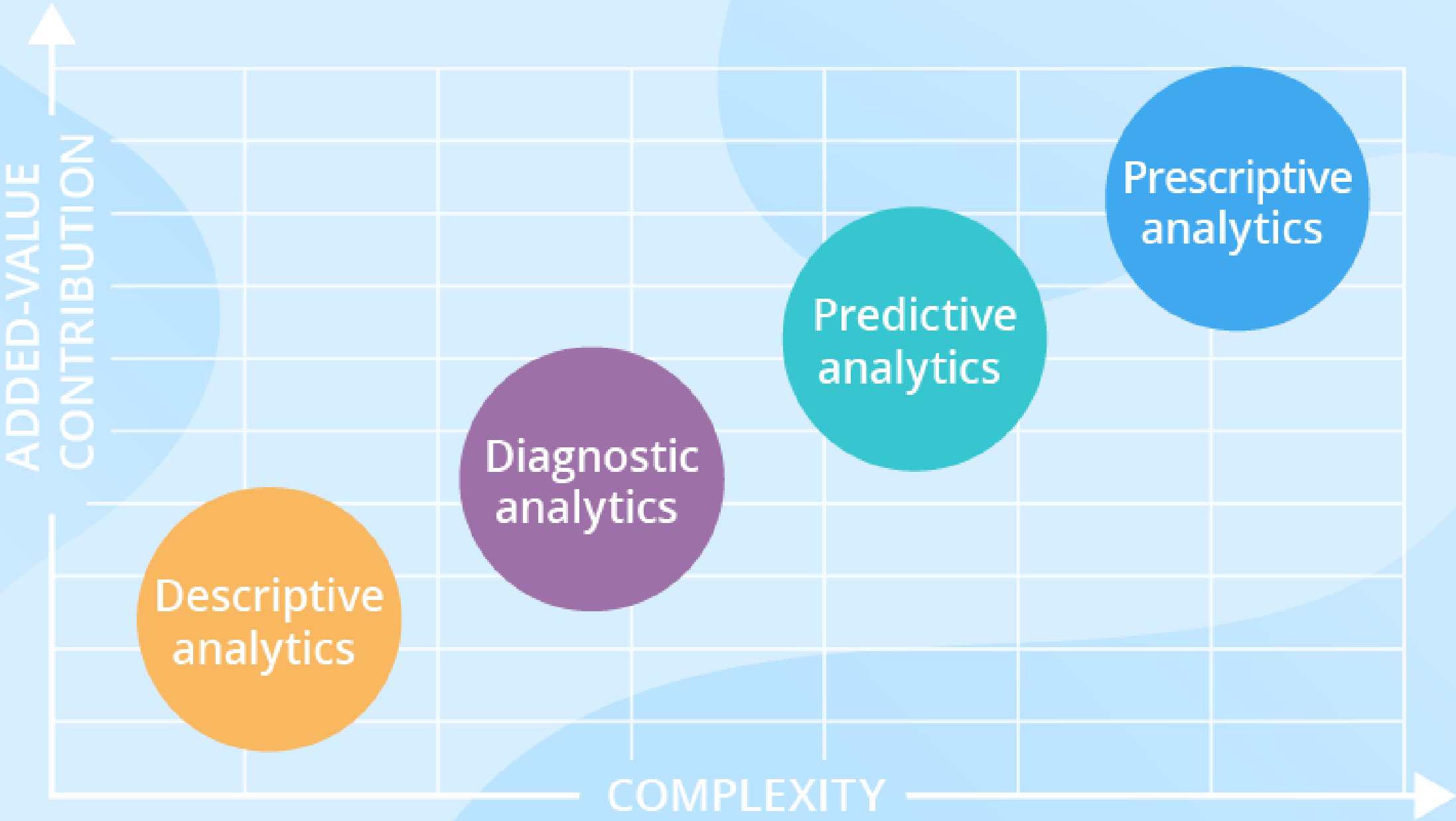
Data analytics is the process of analyzing raw data in order to draw out meaningful, actionable insights.

Data Analytics refers to the technologies, skills and practices for interactive and investigative analysis which are **used to drawing out new, useful insights to improve business planning and boost future performance**, through algorithms and mechanical processes.

TYPES OF DATA ANALYTICS.




Peak Break-up Times
According to Facebook status updates
Spring Break
Spring clean?
April Fool's Day Some kind of terrible joke
Summer holiday
Want to be young, free and single this holiday?
Two weeks before
Christmas holidays
Clear your conscience?
Valentine's Day Boyfriend forgot to book the restaurant?
Mondays
People coming out of terrible weekends, posting their bad news
Christmas Day
Too Cruel?
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
David McCandless & Lee Byron
InformationIsBeautiful.net / LeeByron.com
Source: Facebook Lexicon 2008


ADDED-VALUE CONTRIBUTION
Descriptive analytics
Diagnostic analytics
Predictive analytics
Prescriptive analytics
COMPLEXITY

DATA ANALYTICS.

- Is the science of examining raw data with the purpose of drawing conclusions about that information.
- To boost **data driven decision making**
- Consists of
  - **Exploratory data analysis (EDA)**
  - **Confirmatory data analysis (CDA)**
  - **Qualitative data analysis (QDA)**

EXPLORATORY DATA ANALYSIS (EDA)

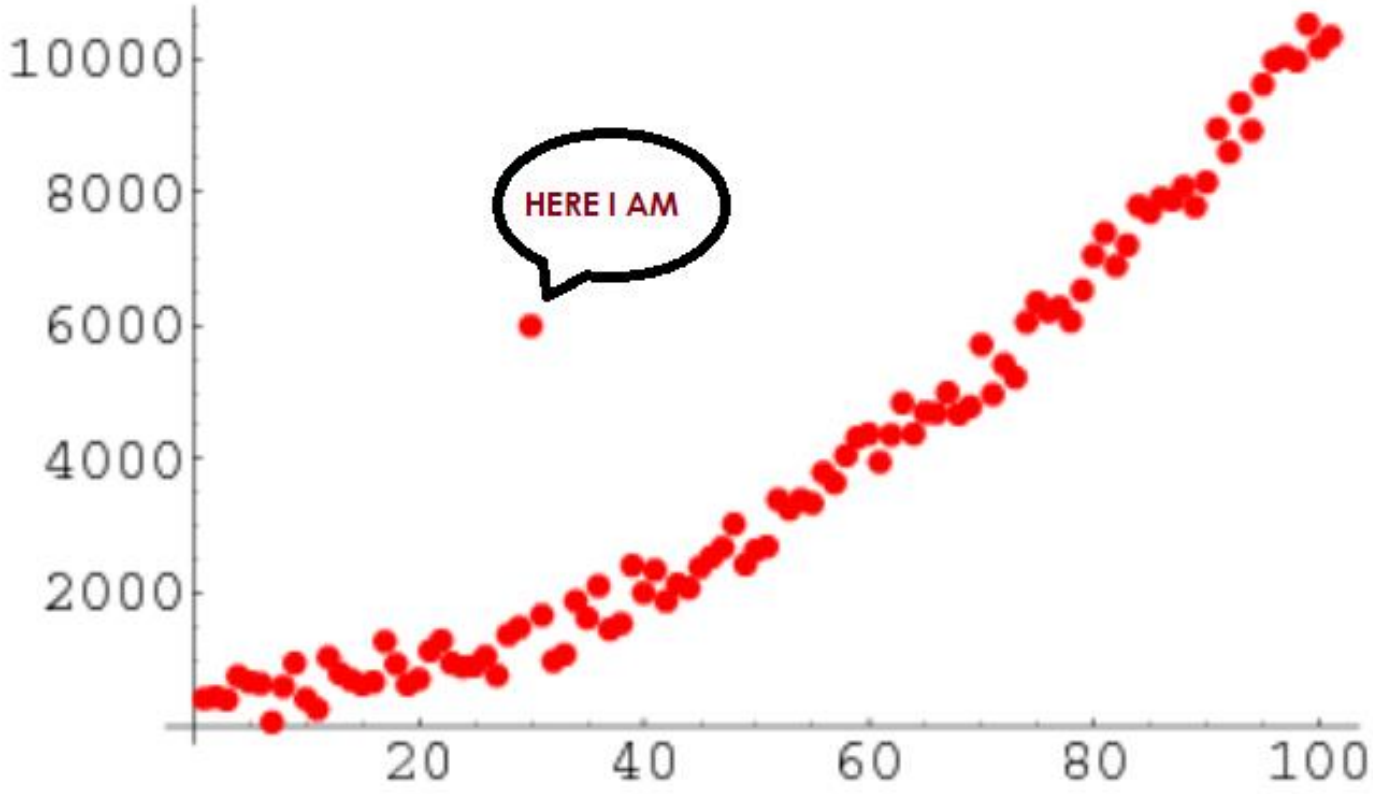
Exploratory Data Analysis is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) in order:

1. To maximize insight into a data set;
2. Uncover underlying structure;
3. Extract important variables;
4. Detect **outliers /** anomalies;
5. Test underlying assumptions;
6. Develop parsimonious models; and
7. Determine optimal factor settings.

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

OUTLIERS.

An ***outlier*** is an observation that lies an abnormal distance from other values in a random sample from a population



CONFIRMATORY DATA ANALYSIS (CDA)

CDA is the process used to evaluate evidence by challenging their assumptions about the data

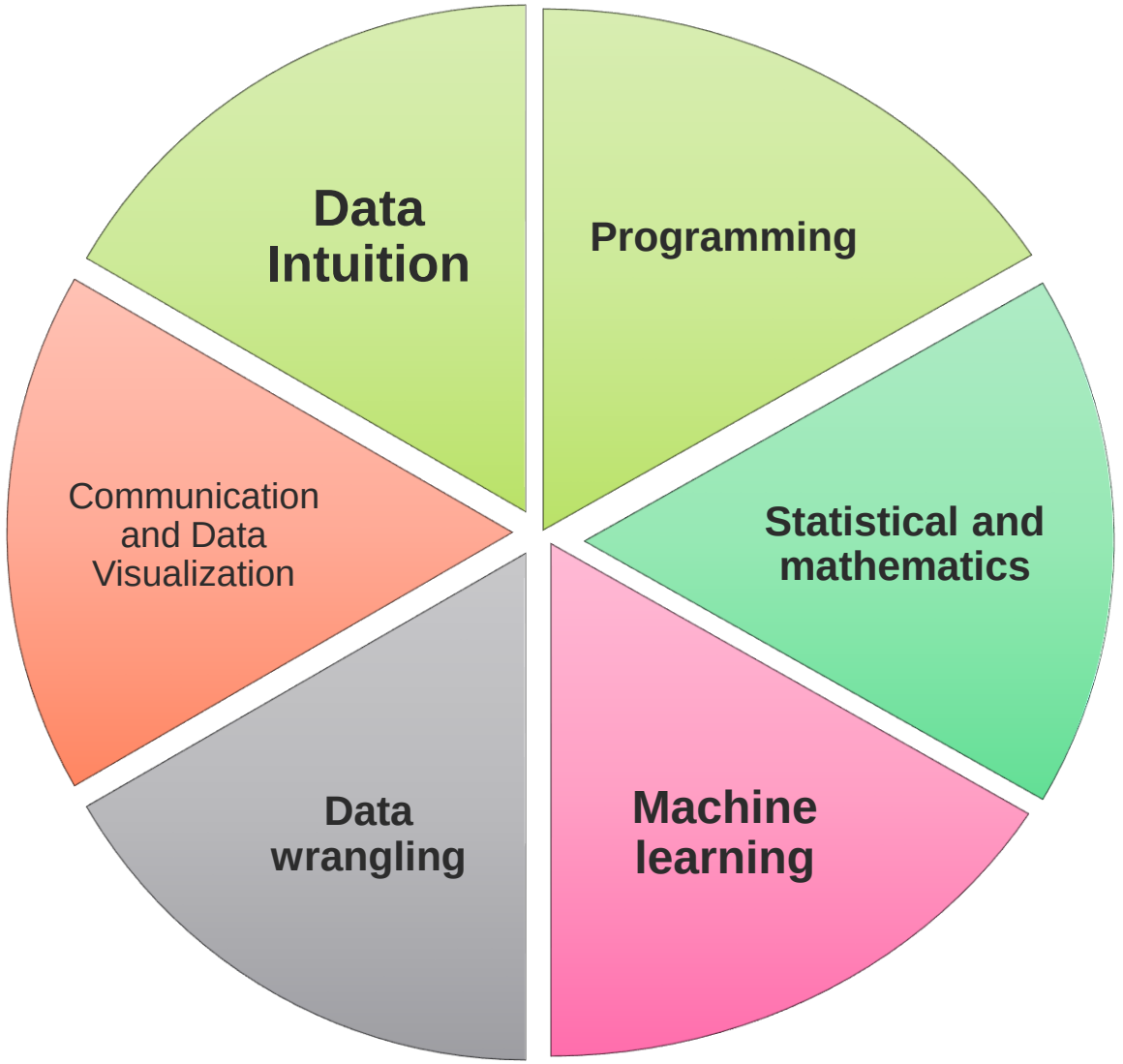
CDA involves processes like
testing hypotheses,
producing estimates,
regression analysis
(estimating the relationship between variables)
and,
variance analysis
(evaluating the difference between the planned and actual outcome).

QUALITATIVE DATA ANALYSIS (QDA)

Qualitative data refers to non-numeric information such as interview transcripts, notes, video and audio recordings, images and text documents.

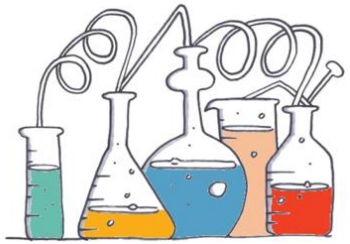
- Content analysis.
- Narrative analysis.
- Discourse analysis..
- Framework analysis.
- Grounded theory.

SKILLS FOR A DATA ANALYST.



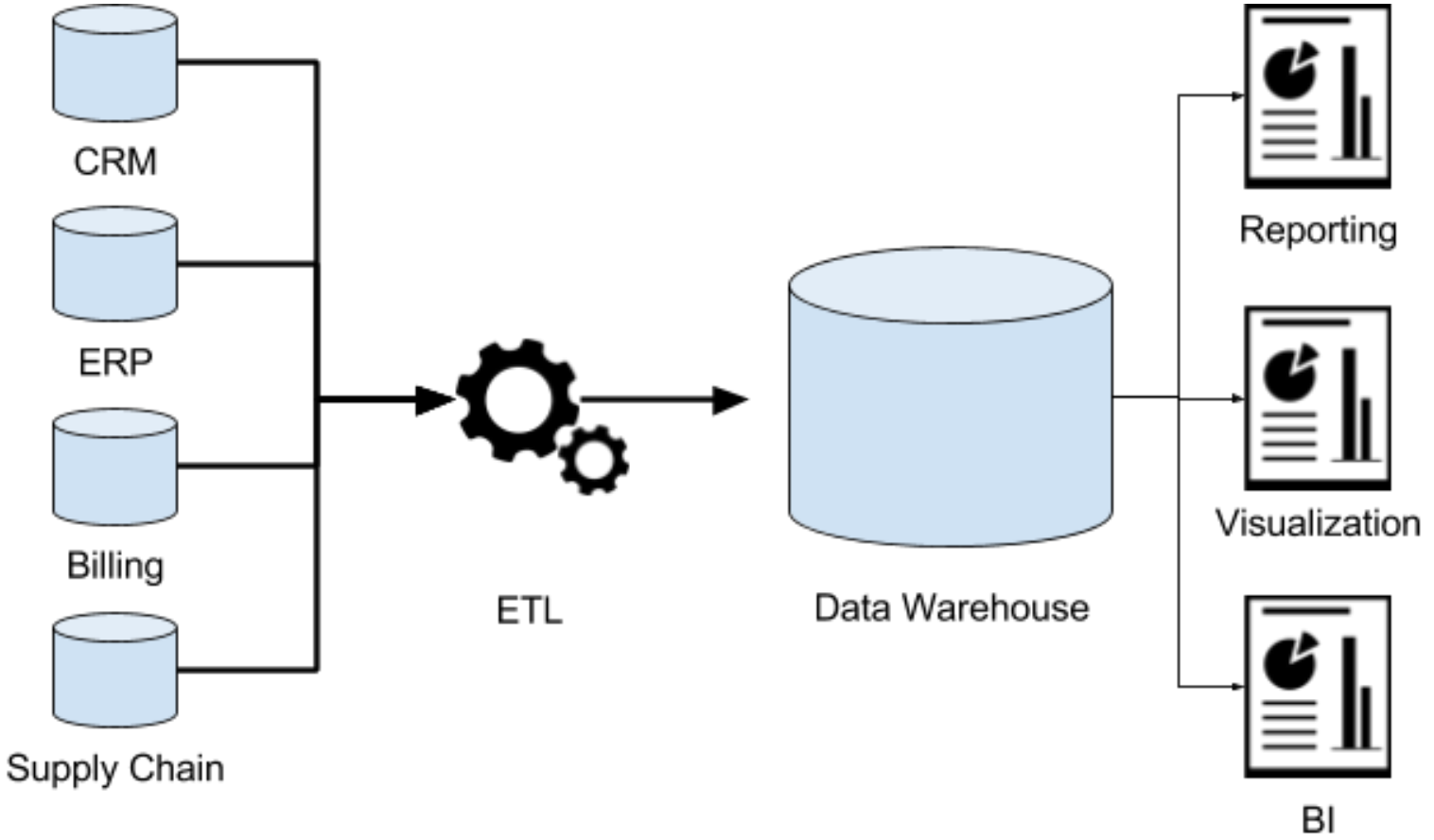
SOURCES OF DATA.

Surveys & Polls



Experiments

Operational DBs



PROCESS FOR MAKING SENSE OF DATA.

The problem to be solved and the projected deliverables should be clearly defined and planned, and an appropriate team should be assembled to perform the analysis.

Prior to starting a data analysis or data mining project, the data should be **collected, characterized, cleaned, transformed**, and partitioned into an appropriate form for further processing.

Based on the information from steps 1 and 2, **appropriate data analysis and data mining techniques** should be selected. These methods often need to be optimized to obtain the best results.

The results from step 3 should be **communicated and/or deployed** to obtain the projected benefits identified at the start of the project.

Summarization
process in which the data is reduced for interpretation without sacrificing any important information. Summaries can be developed for the data as a whole or any portion of the data. **For example**, a retail company that collected data on its transactions could develop summaries of the total sales transactions.

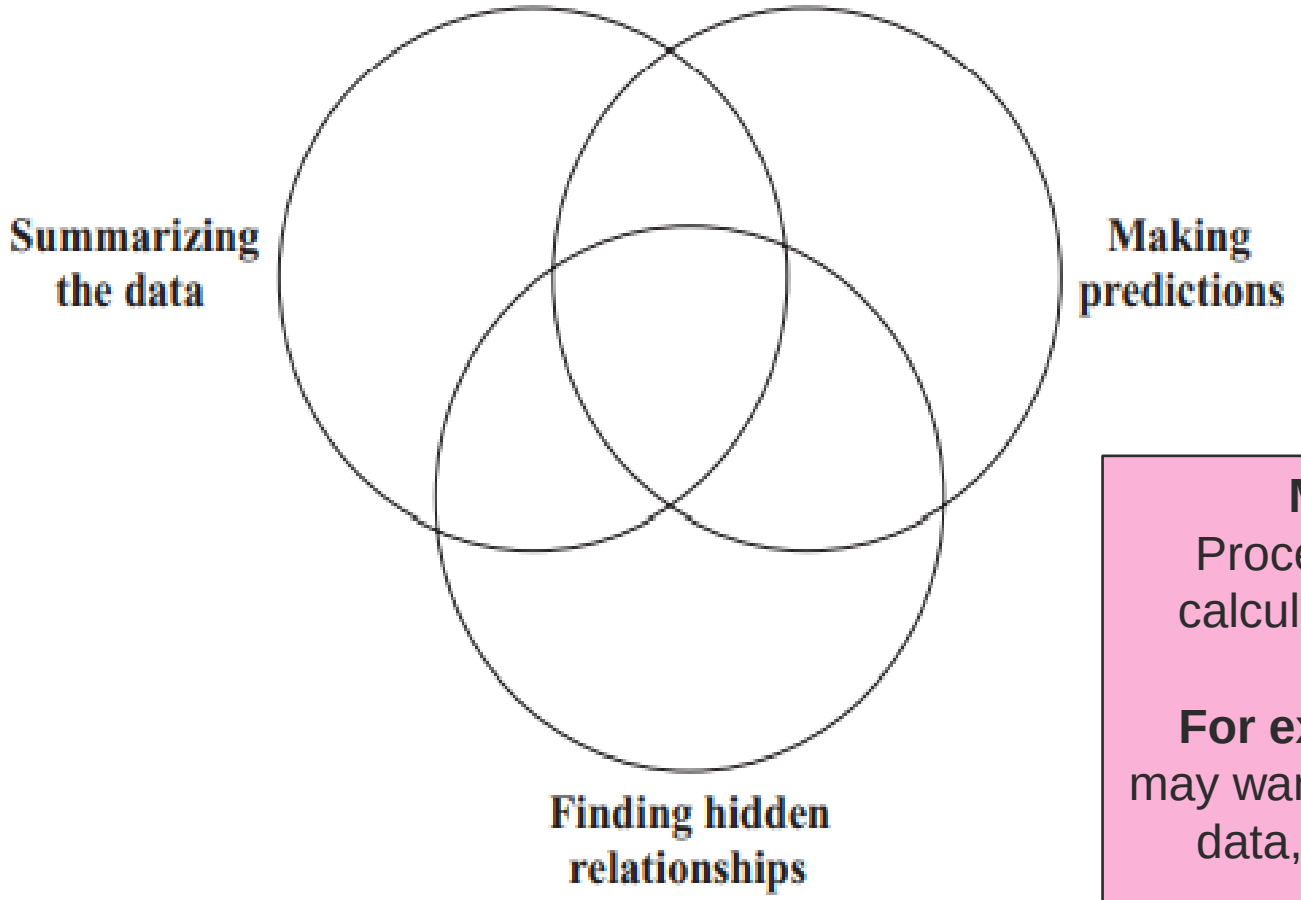


Figure 1.1. Data analysis tasks

Making Predictions
Process where an estimate is calculated for something that is unknown.
For example, a retail company may want to predict, using historical data, the sort of products that specific consumers may be interested in.

FINDING hidden relationships
identification of important facts, relationships, anomalies or trends in the data, which are not obvious from a summary alone. **For example**, a retail company may want tounderstand customer profiles lead to the purchase of certain products.

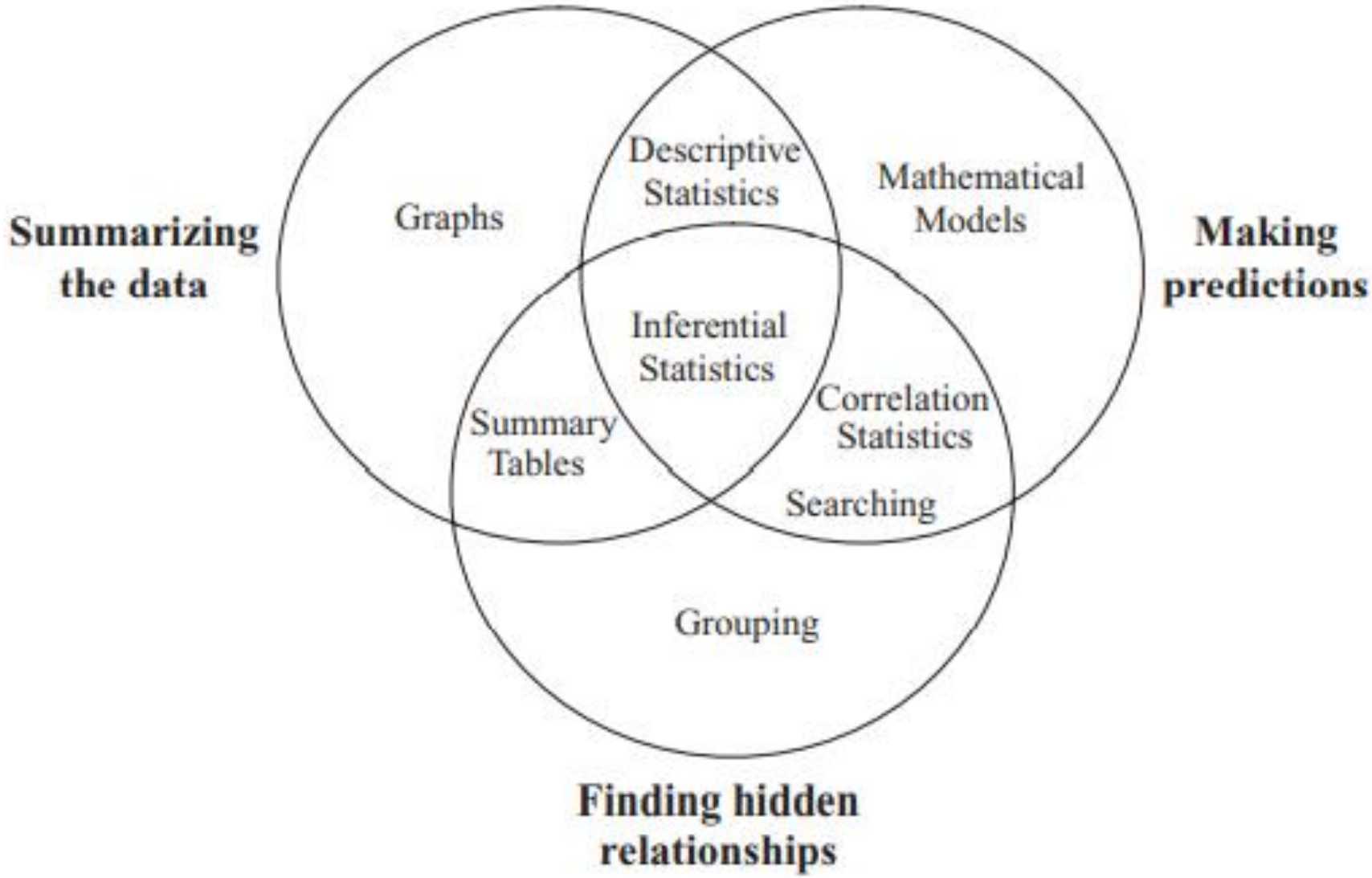


Figure 1.2. Data analysis tasks and methods

PROCESS FOR MAKING SENSE OF DATA.

1. Problem definition and planning

Problem definition and planning

- Identify the problem or need to be addressed
- List the project's deliverables
- Generate success factors
- Understand each resource and other limitations
- Put together an appropriate team
- Create a plan
- Perform a costs/benefits analysis

1. Problem definition and planning

- **Objectives**
  - to define the business or scientific problem to be solved
  - Helps to create a focused plan to execute
  - Success criteria for the project should be defined and measurable
  - Collection of suitable information must be available
  - **For example**: Make recommendations to improve sales on the web site by a specific amount. Sub Objectives are:
    1. Identify categories of web site users (on the basis of demographic information) that are more likely to purchase from the web site.
    2. Categorize users of the web site on the basis of usage information.
    3. Determine if there are any relationships between buying patterns and web site usage patterns.
- **Deliverables**
  - Will the solution be a report, a computer program to be used for making predictions, a new workflow or a set of business rules
  - When developing predictive models, it is useful to understand any required level of accuracy
  - It is also important to understand the consequences of answering questions incorrectly.
  - In many situations, the time to create a model can have an impact on the success of the project

1. Problem definition and planning

Roles and responsibilities

- **Project leader**: responsible for putting together a plan and ensuring the plan is executed.
- **Subject matter experts and/or business analysts**: have specific knowledge of the subject matter or business problems including
  (1) how the data was collected,
  (2) what the data values mean,
  (3) the level of accuracy of the data,
  (4) how to interpret the results of the analysis,
  (5) the business issues being addressed by the project.
- **Data analysis/data mining expert**: familiar with statistics, data analysis methods and data mining approaches as well as issues of data preparation.
- **IT expert**: expertise in pulling data sets together (e.g., accessing databases, joining tables, pivoting tables, etc.) as well as knowledge of software and hardware issues important for the implementation and deployment steps.
- **Consumer:** will use the information derived from the data in making decisions, either as a one-off analysis or on a routine basis.

1. Problem definition and planning

Current Situation

Define the constraints on the project
The sources and locations of the data to be identified.
Privacy or legal issues to be listed.

Timeline

Preliminary implementation plan should be put together.
Time to be set aside for iteration of activities as the solution is optimized

Costs and benefits

budget based on the plan could be used, alongside the business success criteria, to understanding the cost/benefits for the project

Steps	Tasks	Parties involved	Deliverables	Timeline	Budget
				Apr May Jun Jul Aug Sep Oct Nov	
Preparation	Kickoff meeting	All			$500
	Create data tables	Lee	Data tables		$4,000
	Prepare data for analysis	Pam, Tony	Prepared data for analysis		$10,000
	Meeting to review data	All	Plan for implementation		$500
Implementation	Find hidden relationships	Pam, Tony	Key facts and trends		$5,000
	Build and optimize model	Pam, Tony	Model with 70% accuracy		$3,000
	Create a report	All	Report		$1,000
	Meet to review	All	Plan for next steps		$500
Deployment	Double blind test	Pam, Lee	Business impact projections		$5,000
	Production rollout	Pam, Lee	Ongoing review		$5,000
	Assess project	All	Project assessment report		$1,000
					$35,500

Figure 2.1. Project timeline

PROCESS FOR MAKING SENSE OF DATA.

2. Data preparation

Data Preparation

- Access and combine data tables
- Summarize data
- Look for errors
- Transform data
- Segment data

Understanding the data and getting it ready for analysis is the most time-consuming step in the process, since the data is usually integrated from many sources, with different representations and formats.

PROCESS FOR MAKING SENSE OF DATA.

3. Analysis

Analysis

- Summarizing data
- Exploring relationships between attributes
- Grouping the data
- Identifying non-trivial facts, patterns, and trends
- Building regression models
- Building classification models

- **Summarizing the data**
- Methods include
  - Summary tables
  - Graphs
  - Descriptive statistics
  - Inferential statistics
  - Correlation statistics
- **Finding hidden relationships**
- Methods include
  - Data mining
  - Market Basket Analysis
  - Clustering
  - Making prediction

PROCESS FOR MAKING SENSE OF DATA.

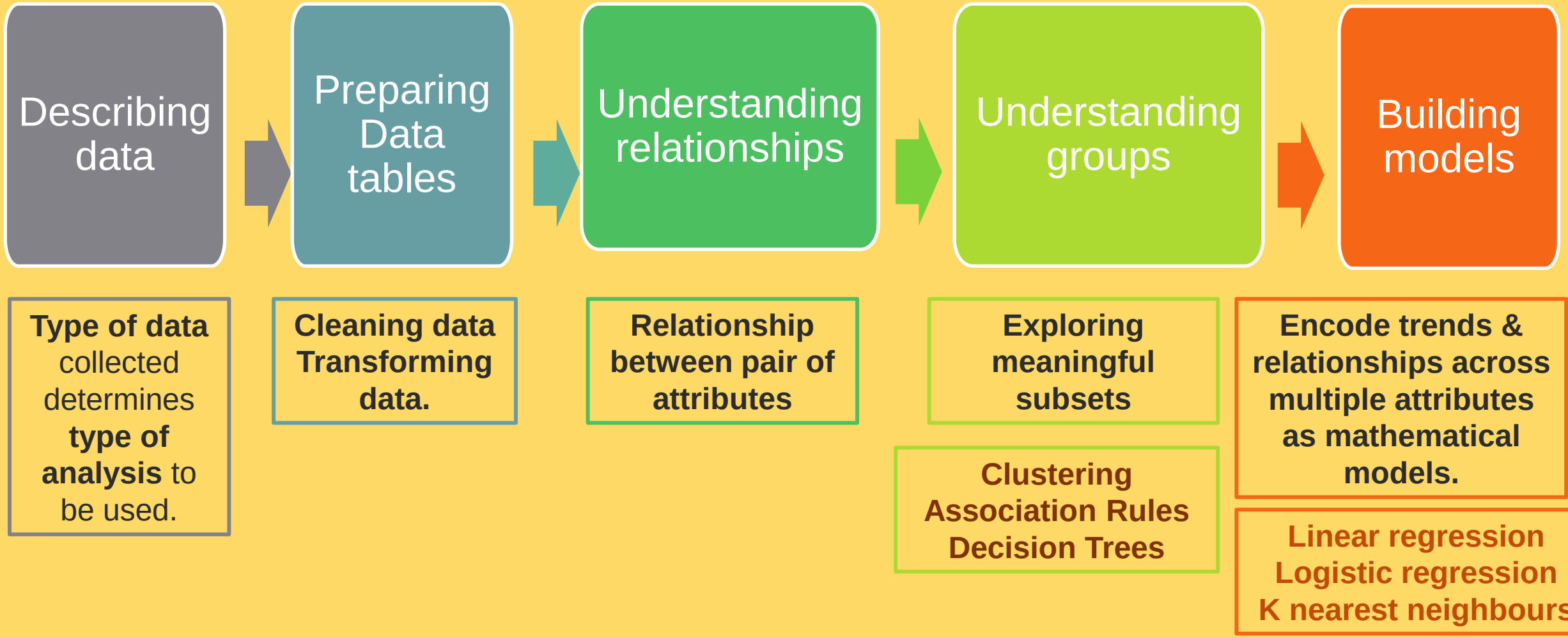
4. Deployment

Deployment

- Generate report
- Deploy standalone or integrated decision-support tool
- Measure business impact

- Analysis is translated into a benefit to the business
- Plan and execute deployment based on the definition in step 1
- Deliverables could be
  - Could be static report of the analysis to management or to the customer
  - Could be a predictive model deployed as standalone or integrated with other software such as spreadsheets or web pages.
  - Measure and monitor performance
- Review the project

GOING FORWARD.



Now for the actual stuff.

Let's

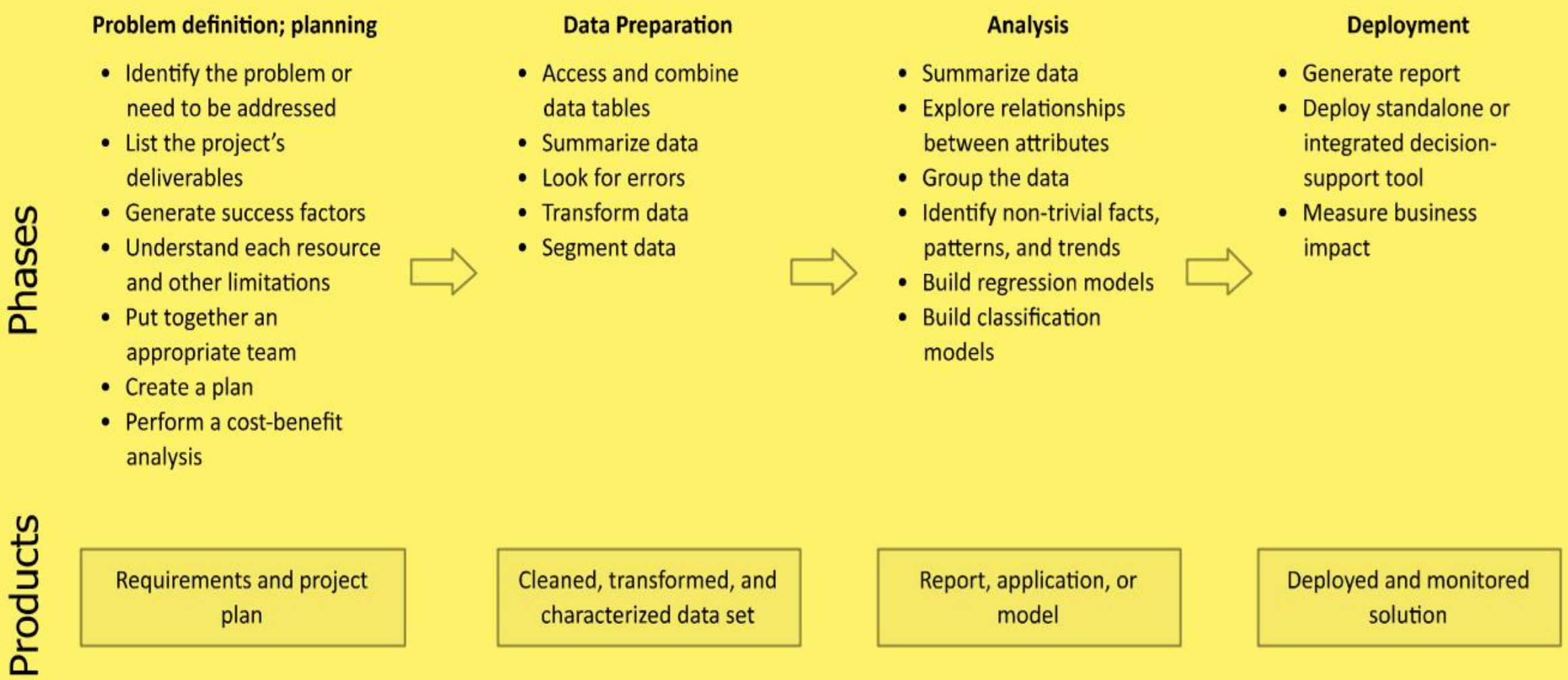


FIGURE 1.6 Summary of steps to consider in developing a data analysis or data mining project.

THANK YOU.