Business Intelligence

Lecture 1

Or Peretz

About Me

Current

- (in progress) Ph.D in Information/Data Science
- Lecturer / Researcher – Probabilistic models, Massive data methodologies, Music technology and collaborative platforms

Past

- BS.c in Computer Science (learning environments and algorithms)
- MS.c in Computer Science (collaborative platforms and science education)
- Full stack software engineer

Logistics:

1. 3 weekly hours
2. Assignments (30%), Project (70%)
3. Communication through Moodle / Email
4. Weekly reception hour
5. Email - or.izhakp@afeka.ac.il

What You Are Getting

1. Describe the concepts and components of business Intelligence.
2. Use of BI for supporting decision making in an organization.
3. Understand and use the technologies and tools that make up BI:
    1. Data warehousing
    2. Data reporting
    3. Analytics
4. Build, manage and maintenance of data-driven project.

What You Are Giving

1. Four assignments (groups of 2-4 members)

   BI developers / analysts always work as a team!

2. Final project

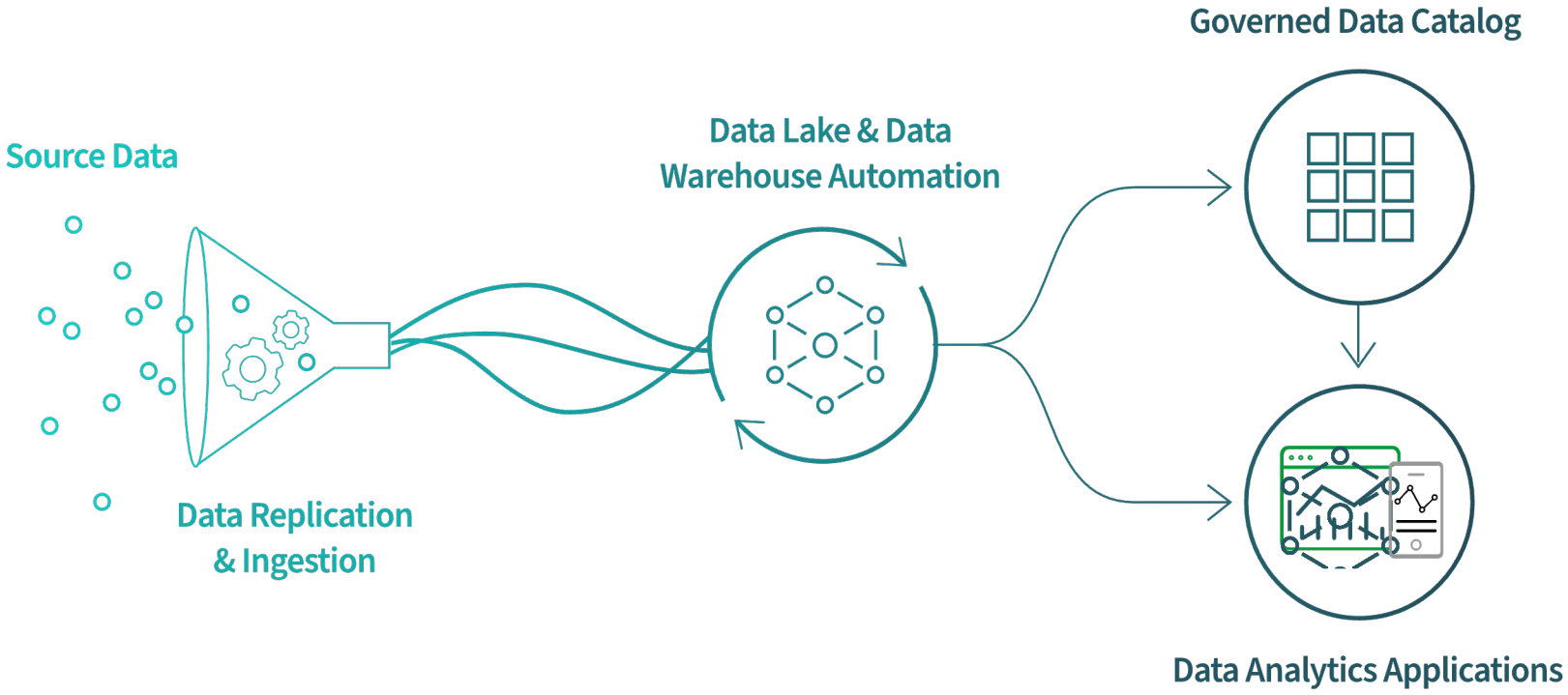
3. Class participation

4. **Smiles and fun ☺**

Ready?

Introduction

Business Intelligence (BI) is a set of ideas, methodologies, processes, architectures, and technologies **that change raw data into significant and useful data for business purpose**.



Introduction

Benefits:

- Can handle large amounts of data to help identify and evolve new opportunities.
- Making use of these new opportunities and applying a productive scheme.
- Provide a comparable market benefit and long-term stability.

Use the technology as a tool!

Example

Suppose we have company data of 3-6 months. Assume that we have candles product. We have three kinds of candles in this class say Candle A, Candle B, Candle C.

On studying the data we come to know that sale of candle C was at peak out of these three classes, and the sale was maximum between the time intervals of 9 am to 11 am.

Example

Now, let’s apply Business Intelligence for this analysis**.**

What the organization can do?

- Get other material that can be used in church and place them nearby those candles.
- Enhance the sales
- Enhance the income
- May help customers to find other relative products

Data

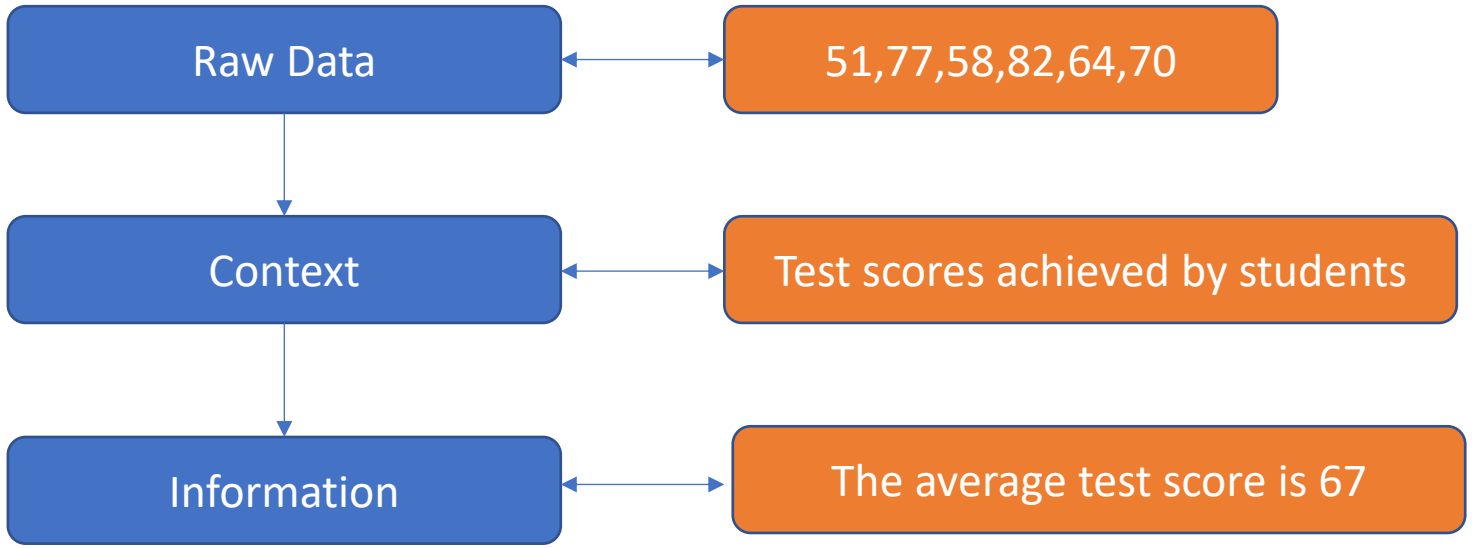
Data - raw facts that describe the characteristics of an event or object

StudentID	Name	Credit	GPA
42271	Ross	77	2.96
65655	Rachel	43	2.34
44311	Monica	65	3.14
72313	Chandler	56	3.81
63223	Joey	44	2.98

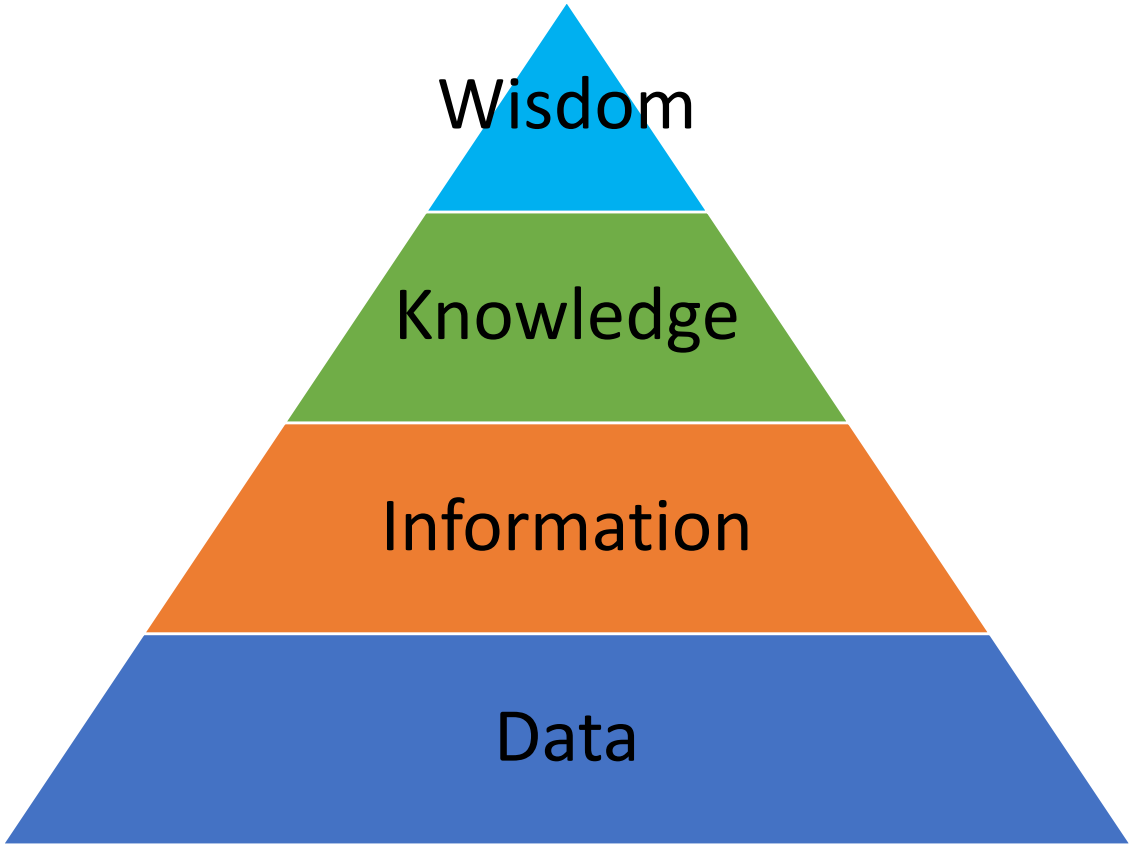
Data

Information – data converted into a meaningful insights

Knowledge – the skills and experience coupled with information and intelligence, that creates a person's intellectual resources.



Data to Wisdom



Applied	I better stop the car!
Context	The traffic light has turned red
Meaning	Traffic lights on the main street
Raw	Red, 192, 235, V2

Business Analytics, BI, Big Data, Data Mining - What's the difference?

1. **Business Analytics**

   Tools to explore past data to gain insight into future business decisions.

2. **BI**

   Tools and techniques to turn data into meaningful information.

3. **Big Data**

   data sets that are so large or complex that traditional data processing applications are inadequate.

4. **Data Mining**

   Tools for discovering patterns in large data sets.

Businesses Need Support for Decision Making

1. Uncertain economics
2. Rapidly changing environments
3. Global competition
4. Demanding customers
5. Taking advantage of information acquired by companies is a Critical Success Factor.

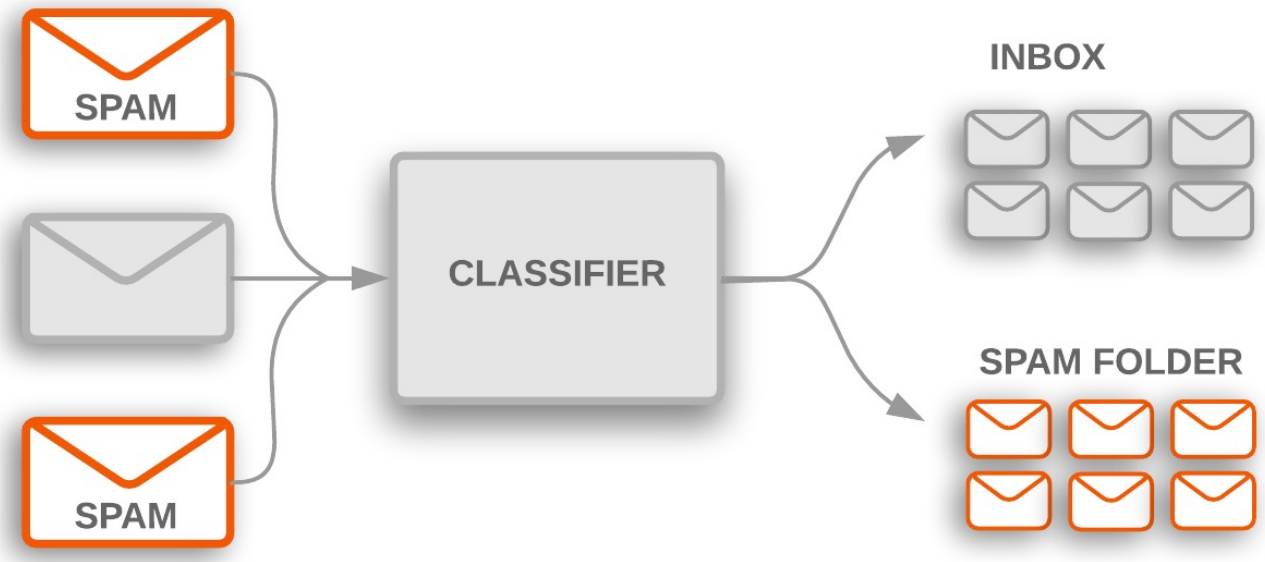
Data Mining

"*Data mining* is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large *data* sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems." - Wikipedia

1. Examining large databases to produce new information.
2. Uses statistical methods and artificial intelligence to analyze data
3. Finds hidden features of the data that were not yet known.

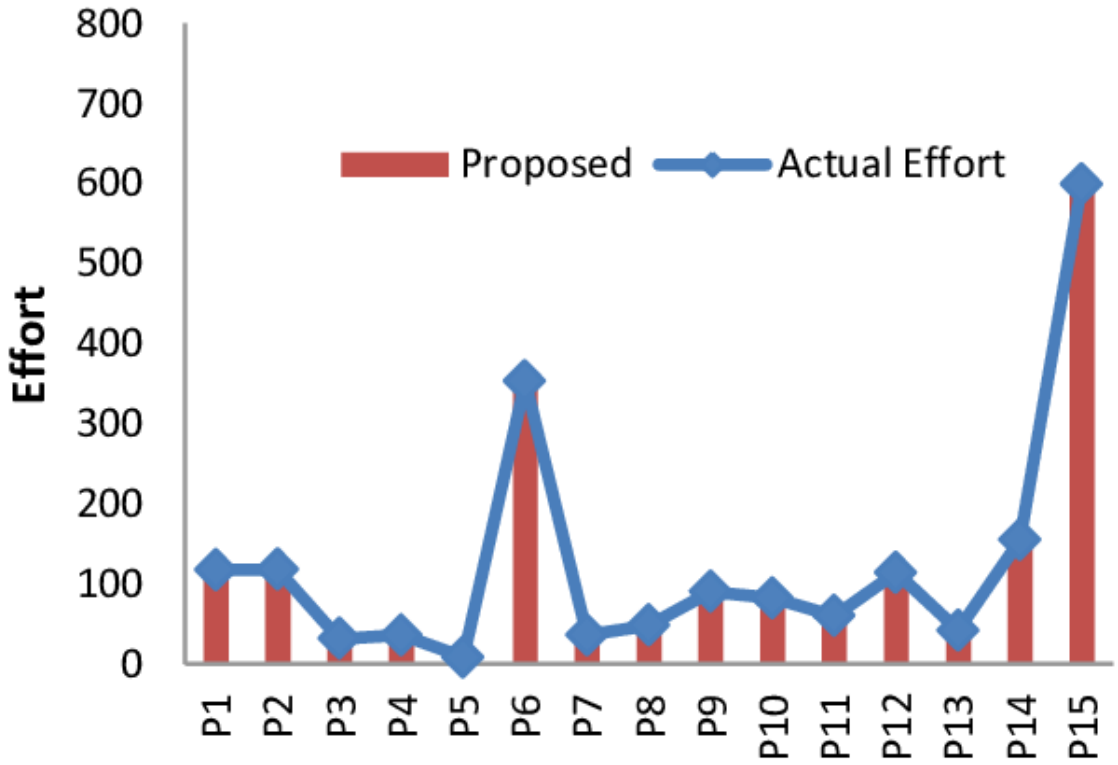
Tasks of Data Mining in Business

Classification – Categorizing data into actionable groups.



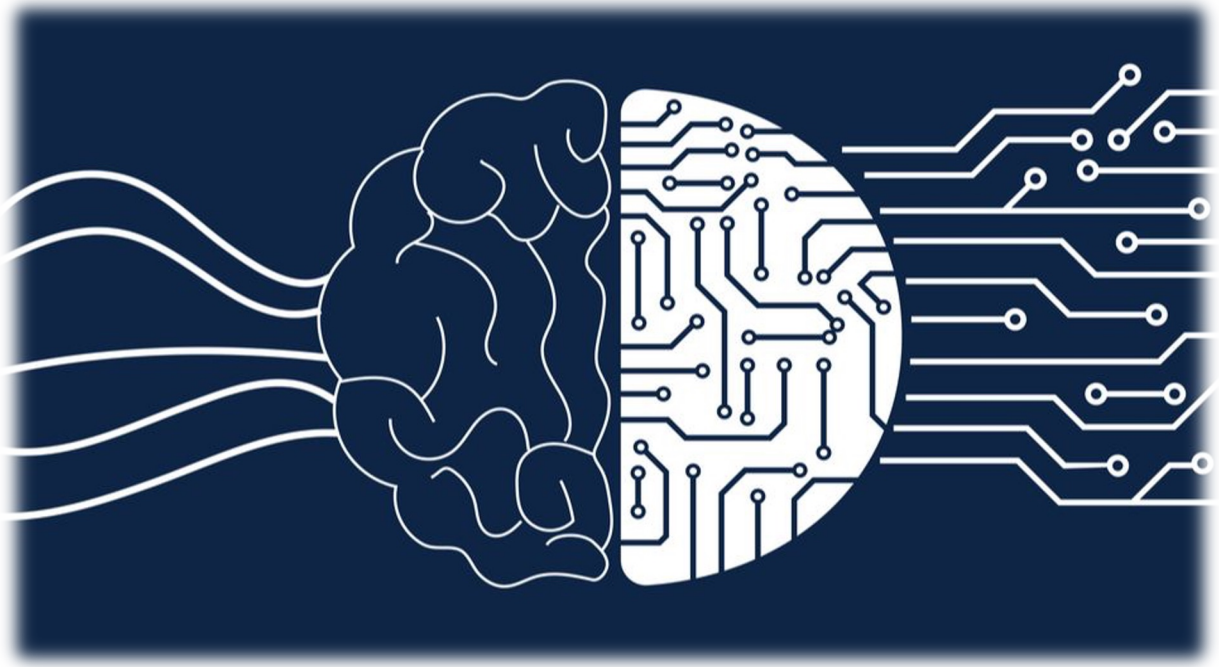
Tasks of Data Mining in Business

Estimation – Response rates, probabilities of responses.



Tasks of Data Mining in Business

Prediction – Predicting customer behavior.

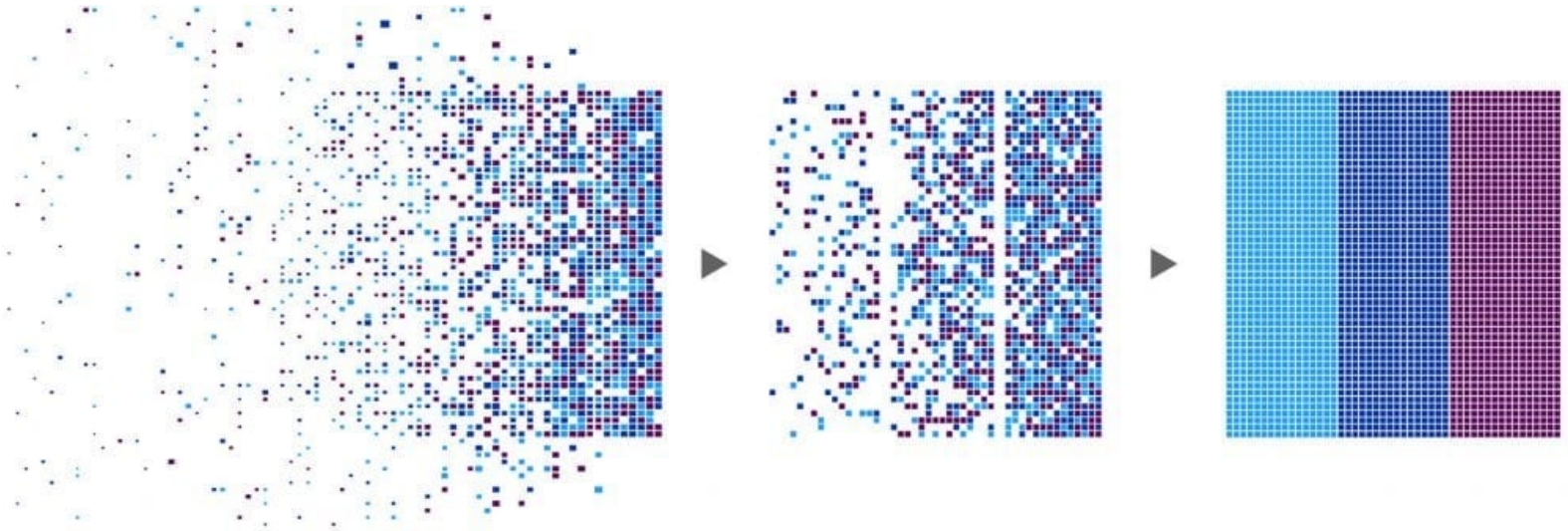


Tasks of Data Mining in Business

Affinity Grouping – What items or services are customers likely to purchase together?

Tasks of Data Mining in Business

Description – Finding interesting patterns.



Data Mining Techniques

1. **Market Basket Analysis**

   Finding patterns or sequences in the way that people purchase products and services.

2. **Cluster Analysis**

   Grouping data into like clusters based on specific attributes

3. **Principal Components Analysis (PCA)**

4. **Decision Trees and Rule Induction**

5. **Logistic / Linear Regressions**

Business Intelligence

1. Collecting and refining information from many sources (internal and external)
2. Analyzing and presenting the information in useful ways (dashboards, visualizations) So that people can make **better decisions**
3. Tools and techniques to turn data into meaningful information.

   Process: Methods used by the organization to turn data into knowledge.

   Product: Information that allows businesses to make decisions.

BI Initiatives

- 75% of BI projects fail because of **poor communication** and **not understanding** what to ask. (Goodwin, 2010)
- 65% of BI projects fail because of technology and **lack of infrastructure** (Lapu, 2007)
- 70% of senior executives report that analytics will be important for competitive advantage. Only 2% feel that they've achieved competitive advantage.

Data Warehouse

A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources.

The concept of the data warehouse is a lone scheme that is the repository of all of the organization’s data (or simply data) in a pattern that can be competently analyzed.

Data Warehouse - Challenges

1. Data should be acquired from a variety of incompatible systems.
2. The identical piece of data might reside in the databases of distinct systems in distinct types.
3. A specific data item might not only be represented in distinct formats, which value is the correct one?
4. Data is continually altering. How often should the Data warehouse be revised to contemplate a sensibly current view?

Data Warehouse - Main Challenge

The amount of data is massive. How is it analyzed and presented easily so that it is useful?

1. **Extract, Transform, and Load** (ETL) utilities for the moving of data from the diverse data sources to the common data warehouse.
2. **Data-mining** pushes for complex predetermined analysis of the data retained in the data warehouse.
3. **Reporting** tools to provide management employees with the outcomes of the analysis in very simple to absorb formats.

Data Warehouse - Strategies

Data mart

1. The most common approach
2. Begins with a single mart and architected marts are added over time for more subject areas
3. Relatively inexpensive and easy to implement
4. Can be used as a proof of concept (POC) for data warehousing
5. Can postpone difficult decisions and activities
6. Requires an overall integration plan

Data Warehouse - Strategies

Enterprise-wide

1. A comprehensive warehouse is built initially
2. An initial dependent data mart is built using a subset of the data in the warehouse
3. Additional data marts are built using subsets of the data in the warehouse
4. Expensive, time consuming, and prone to failure
5. When successful, it results in an integrated, scalable warehouse

ETL – Extract, Transform, Load

Extracts data from one or more data-sources

Transforms it and cleanses it to be optimized for reporting and analysis

Loads it into a data store or data warehouse

There are many different models of ETL tools in today's BI market, from complex, specialized products to light, web-based solutions that work easily with multiple data sources.

Reasons for “Dirty” Data

- Dummy values
- Absence of data
- Multipurpose fields
- Cryptic data
- Contradicting data
- Violation of business rules
- Reused primary keys
- Non-unique identifiers (keys)
- Data integration problems

Data Cleaning

Parsing - locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

- Parsing the first, middle, and last name; street number and street name.

Correcting -corrects parsed individual data components using sophisticated data algorithms and secondary data sources.

- Replacing a vanity address and adding a zip code.

Data Cleaning

Standardizing - standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.

- Adding a pre name
- Replacing a nickname

Data Cleaning

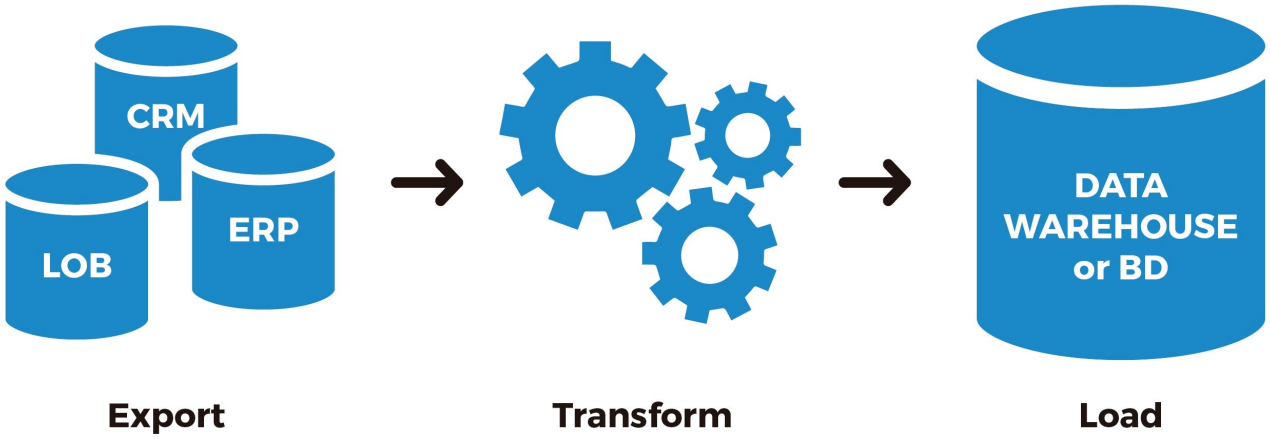
Matching - searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

- Identifying similar names and addresses.

Consolidating - analyzing and identifying relationships between matched records and consolidating/merging them into ONE representation.

Data Transformation and Loading

- Transforms the data with business rules and standards that have been established
  - format changes, deduplication, splitting up fields, replacement of codes, and aggregates
- Data are **physically** moved to the data warehouse

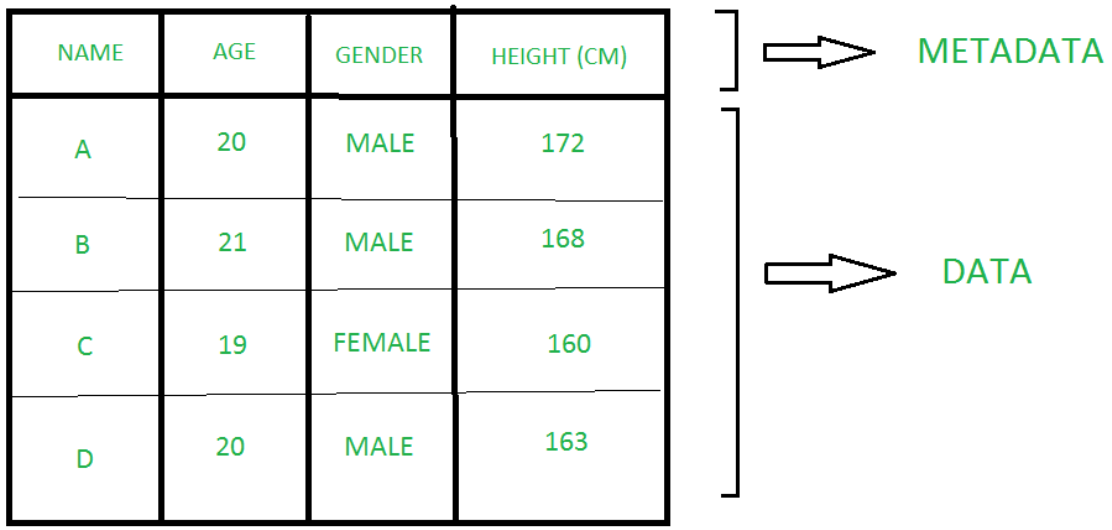


META Data

1. Data about data, includes data sources and targets, database, table and column names.

2. Needed by both information technology and users

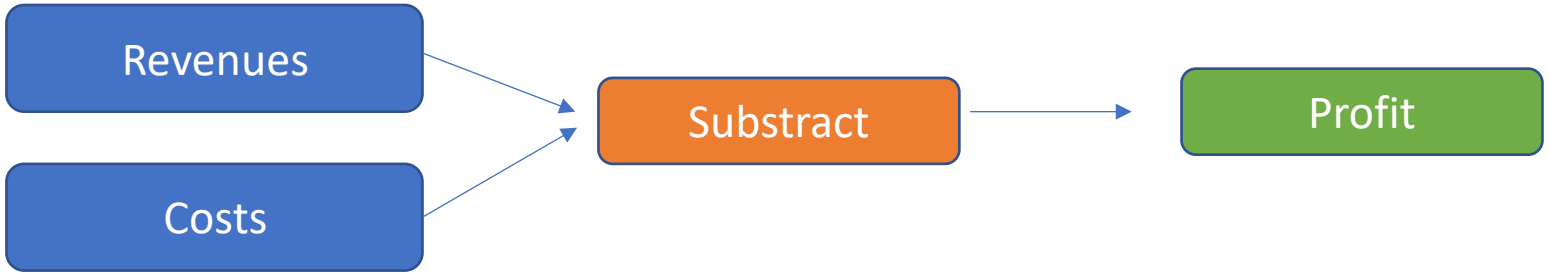
   - Users need to know entity/attribute definitions
   - Reports/query tools
   - Report distribution information



Mathematical Models

- Mathematics can be used to represent real-world situations.
- The instruments used to represent real are called **mathematical models**.
- They help us understand how the real-world works.

A simple mathematical model representing profit calculation is illustrated as



Use of Mathematical Models in Business

Decision Making

- Multiple participants with conflicting views
- Use input variables and a set of conditions to arrive at a decision
- Decide whether to invest in a project or not

Use of Mathematical Models in Business

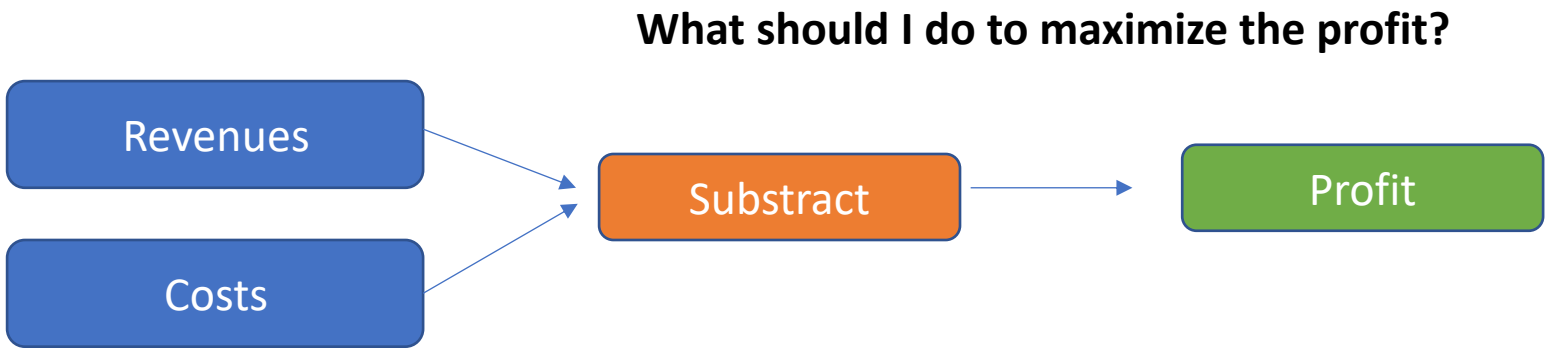
Making predictions

- Requirement of predicting certain factors, such as revenue, growth rate, costs.
- Used in case of new product launch, change in strategy, investment needs, etc.
- Predictive mathematical models are used that analyze historical data and use probability distribution as input for predicting the future values.
- Regression analysis

Use of Mathematical Models in Business

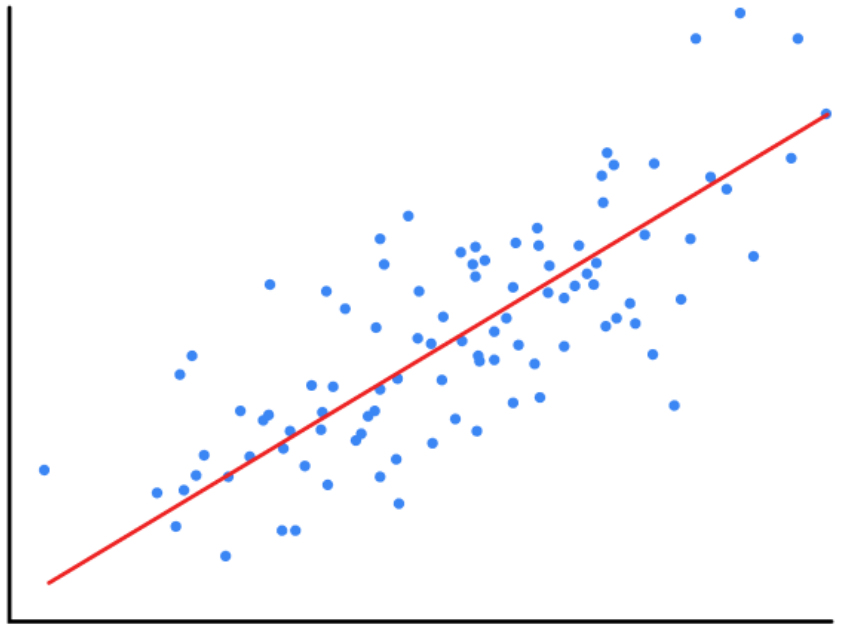
Optimizing

- These models often maximize or minimize a quantity by making changes in another variable or a set of variables.

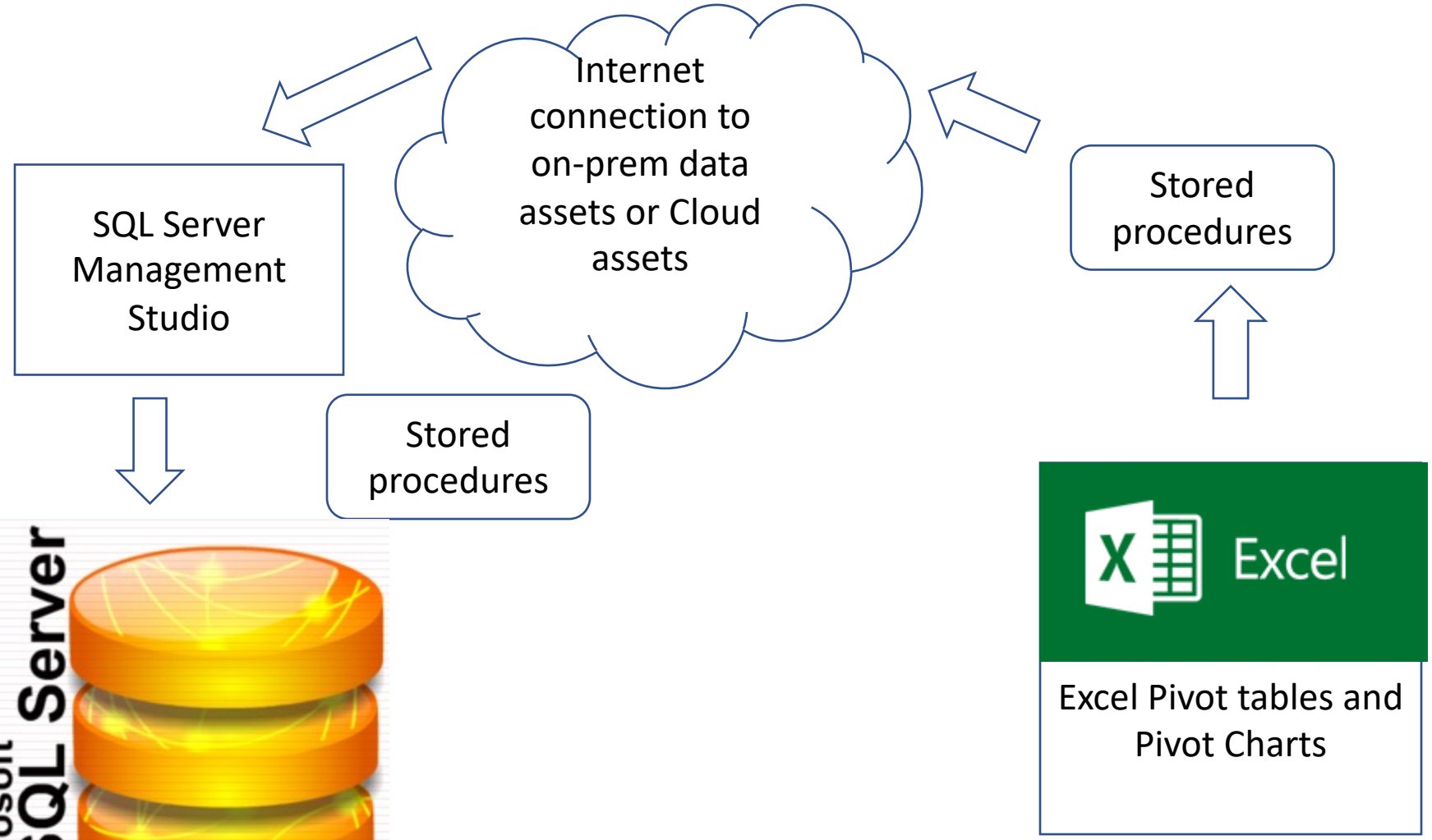


Transforming Data into Decisions

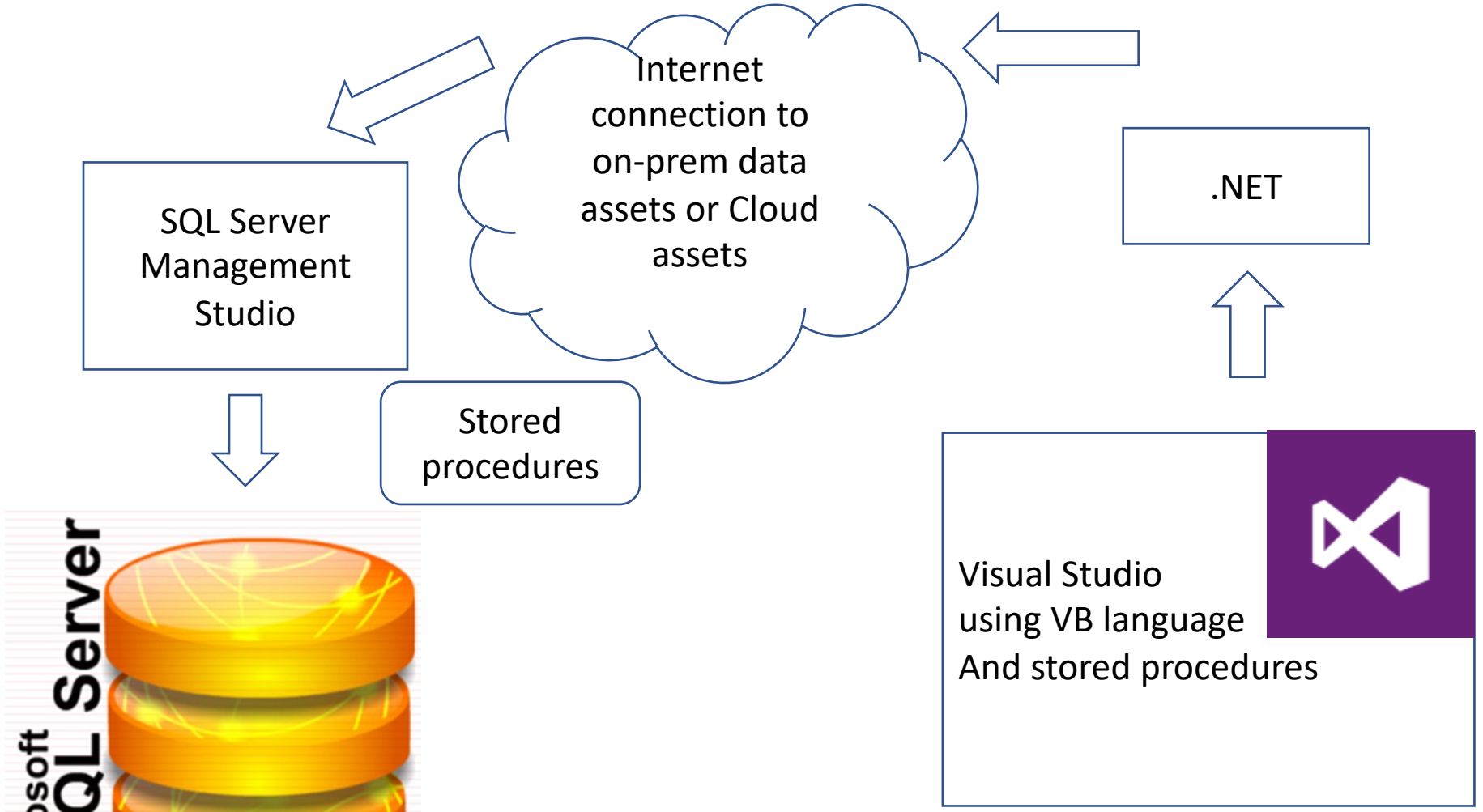
1. Define objectives and information needs
2. Collect data
3. Analyze data
4. Present information
5. Make **data-driven** decisions



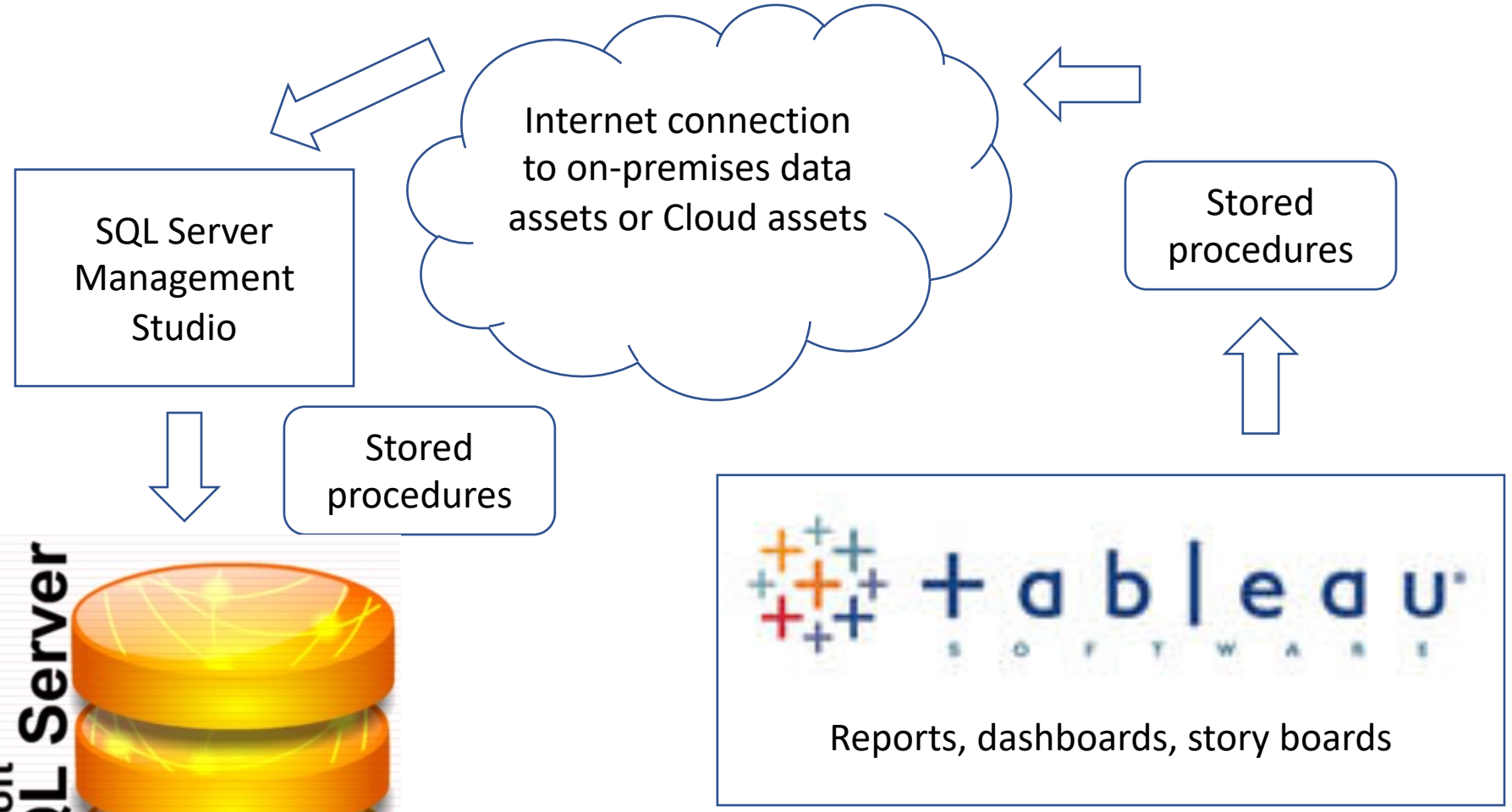
Connection to Data



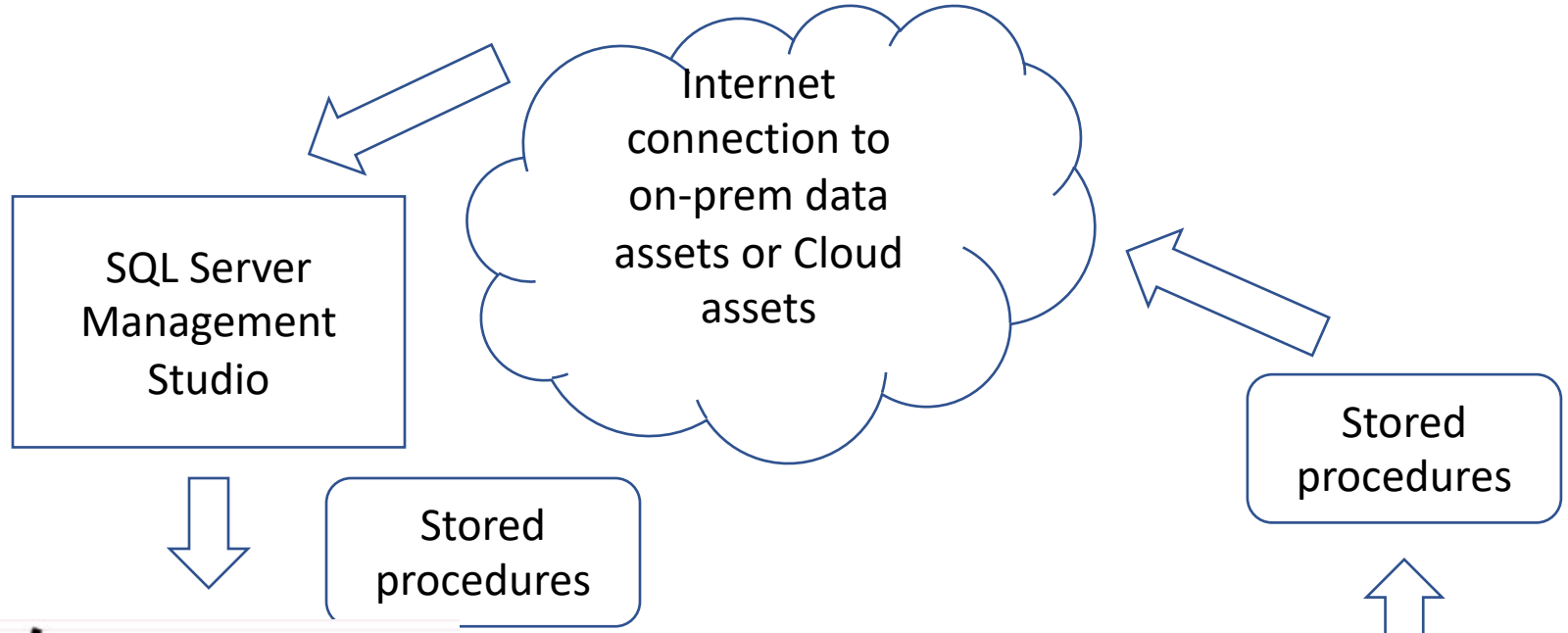
Connection to Data



Connection to Data



Connection to Data

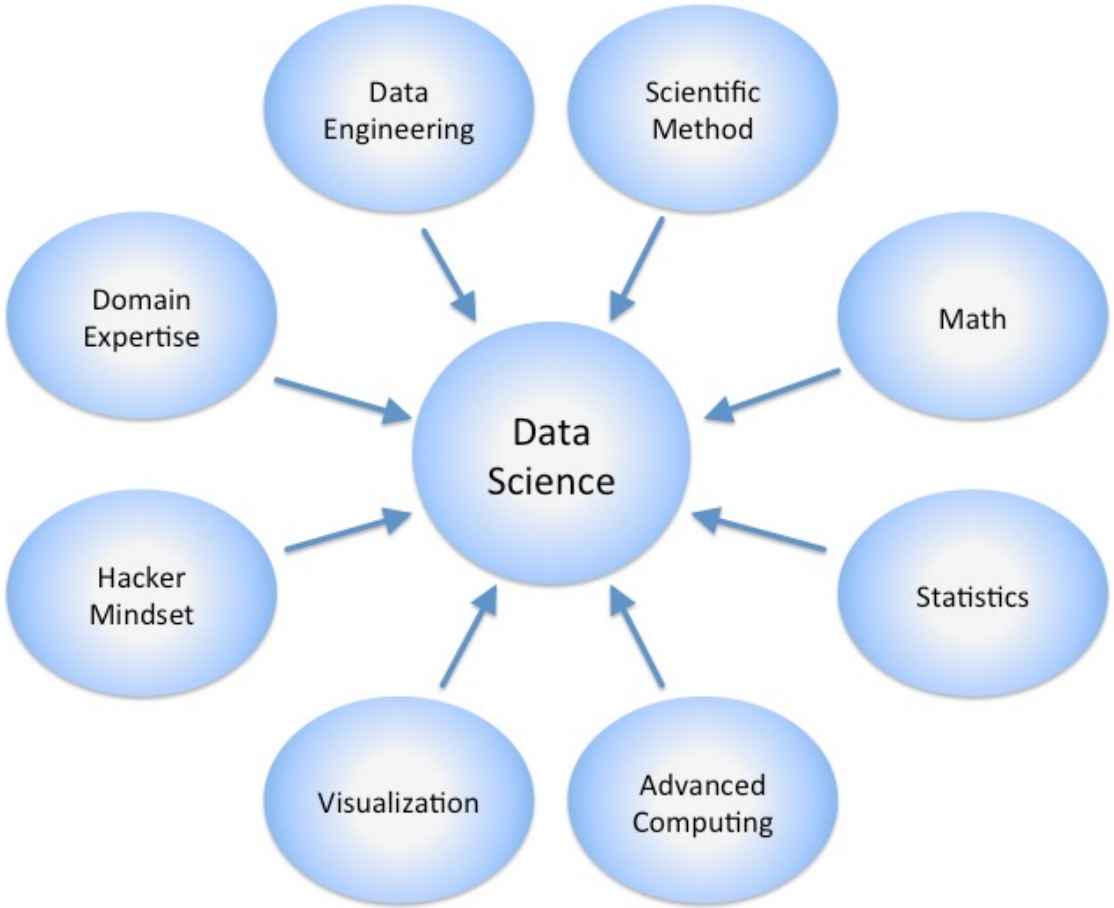


Microsoft® SQL Server®
Reporting Services

Why so many IT approaches?

- The more technologies you are familiar with the more you understand the overall BI Architecture picture and can lead a project
- A mastery over technical tools fuels your ability to be a competent analyst
- **The more tools you have under your belt, the more likely you will NOT get an entry-level BI or analyst position (rather rise in stature)**

Will I become a Data Scientist?



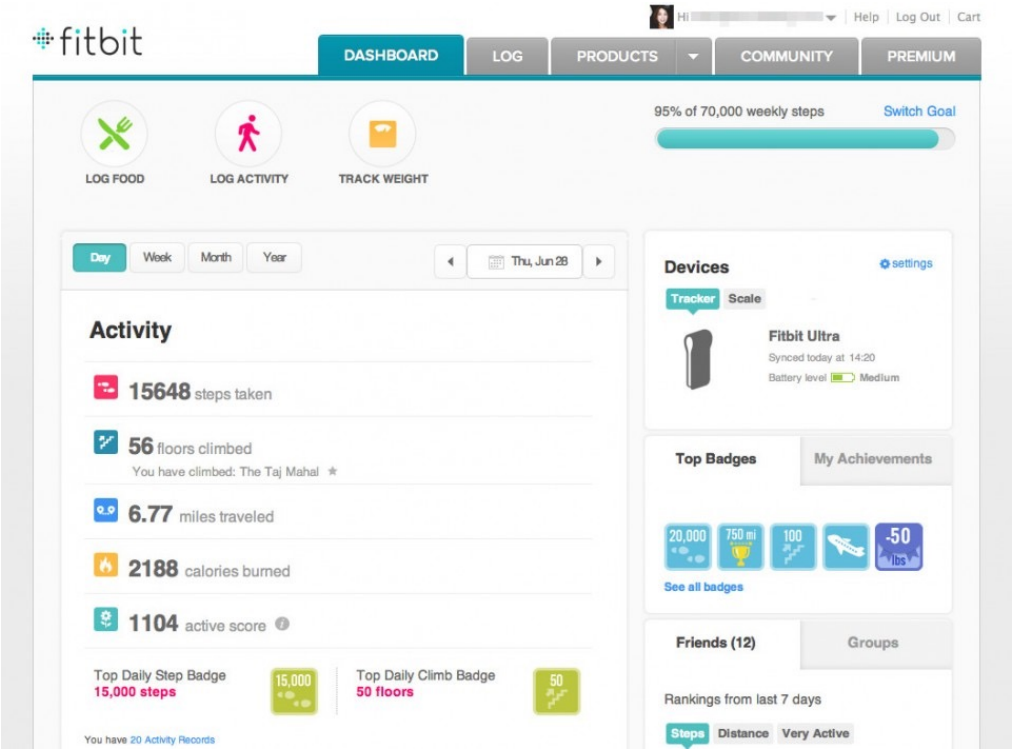
Wikipedia.com

Reporting Tools

The knowledge created by a data-mining engine is not very useful unless it is presented easily and clearly to those who need it. There are many formats for reporting information and knowledge results. One of the common techniques for displaying information is the **digital dashboard.**

Reporting Tools

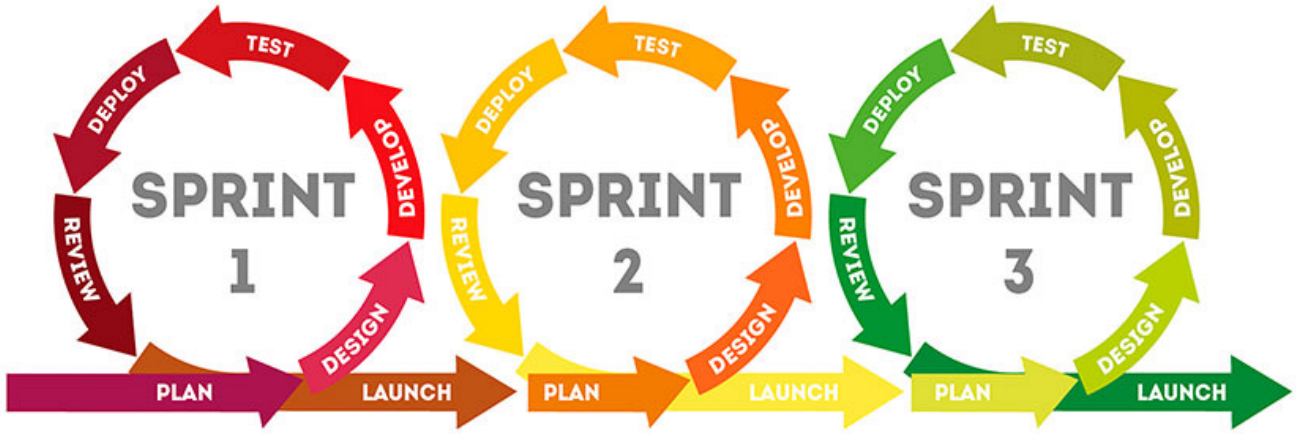
The business / customers should find the answers to their questions in the dashboard



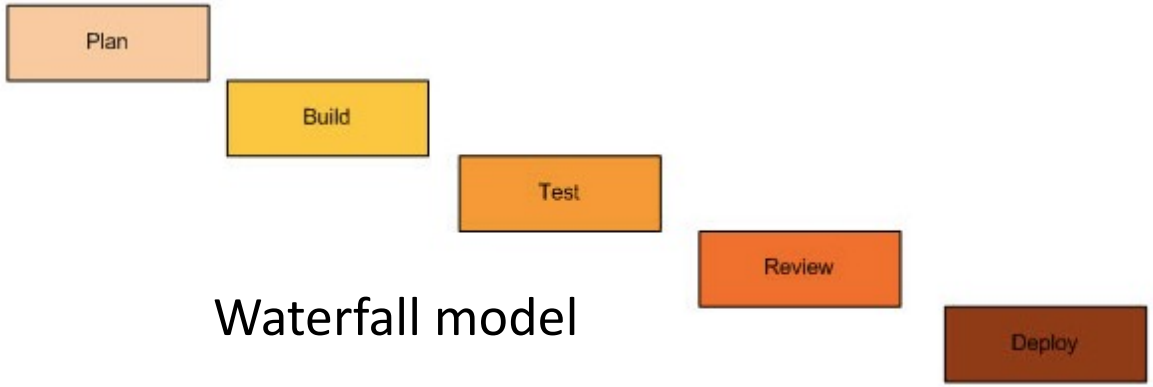
Support Agile Methodology

The Agile methodology is a way to manage a project by breaking it up into several phases.

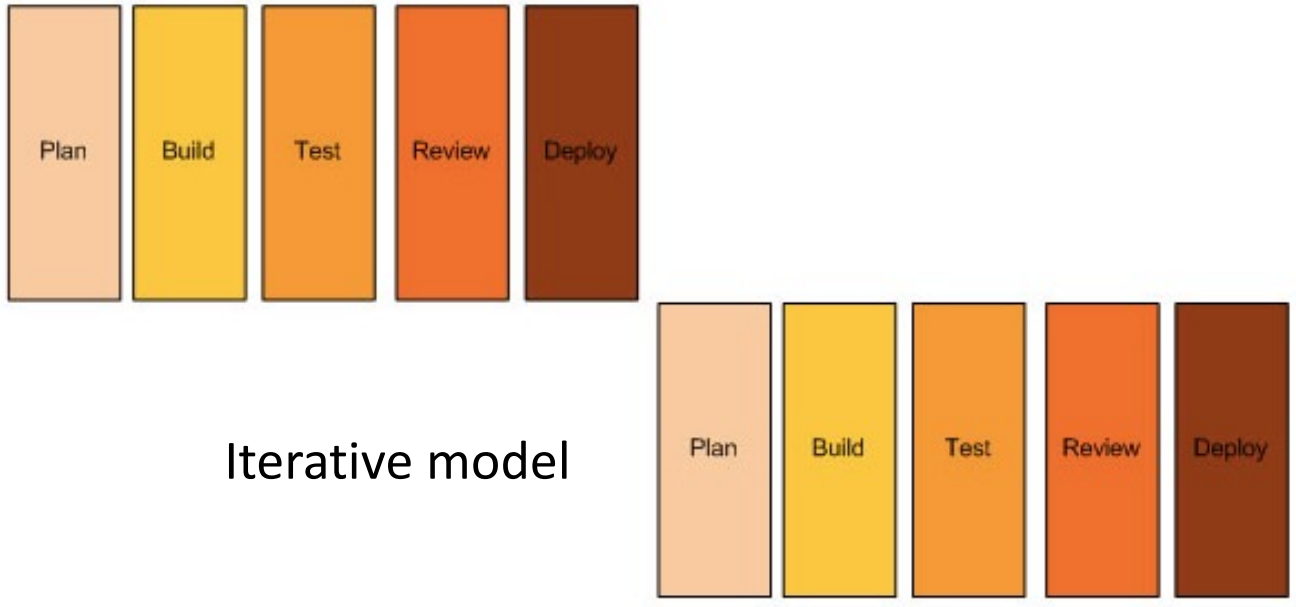
- It involves constant collaboration and continuous improvement at every stage.
- Teams cycle through a process of planning, executing, and evaluating.
- Continuous collaboration is **vital**.



The Basics: From Traditional to Agile



Waterfall model



Iterative model

Agile Principles

- **Our highest priority is to satisfy the customer.**
- Welcome changing requirements, even late in development.
- Deliver working software frequently, from a couple of weeks to a couple of months (with a preference to the shorter timescale).
- The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.

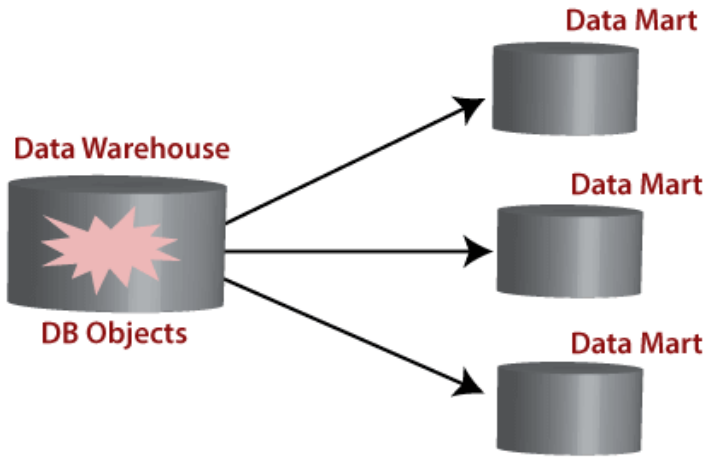
Agile Principles

- Business people and developers must work together daily throughout the project.
- **At each stage/sprint, you are holding a deliverable project!**
- Working software is the primary measure of progress.
- The sponsors, developers, and users should be able to maintain a constant working.
- The best architectures, requirements, and designs emerge from self-organizing teams.

Data Mart and Data Sources

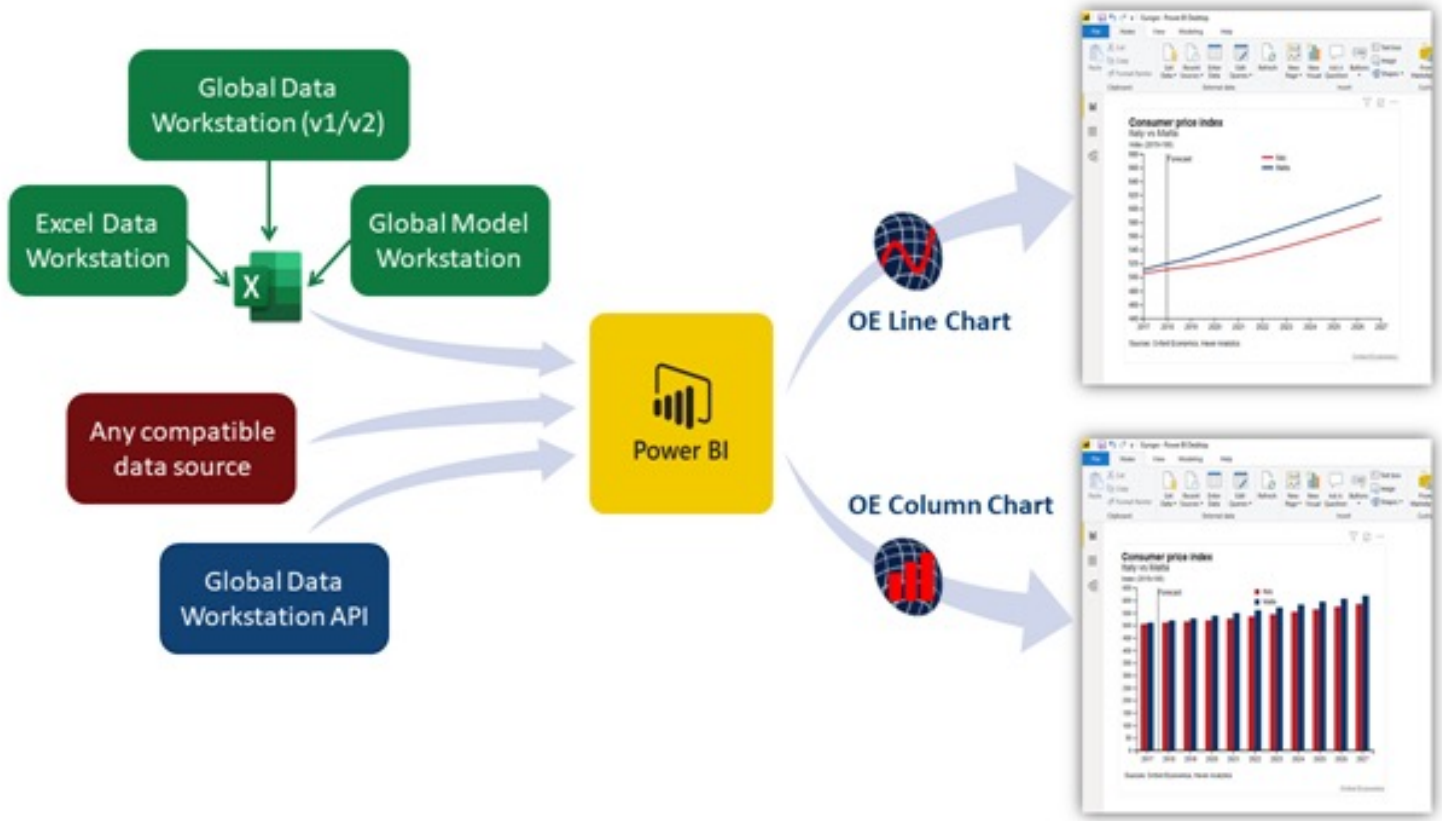
A data mart is a structure / access pattern specific to data warehouse environments.

Used to retrieve client-facing data.



- Data sources gather and integrate the data stored
- May include unstructured documents (such as data received from external providers)

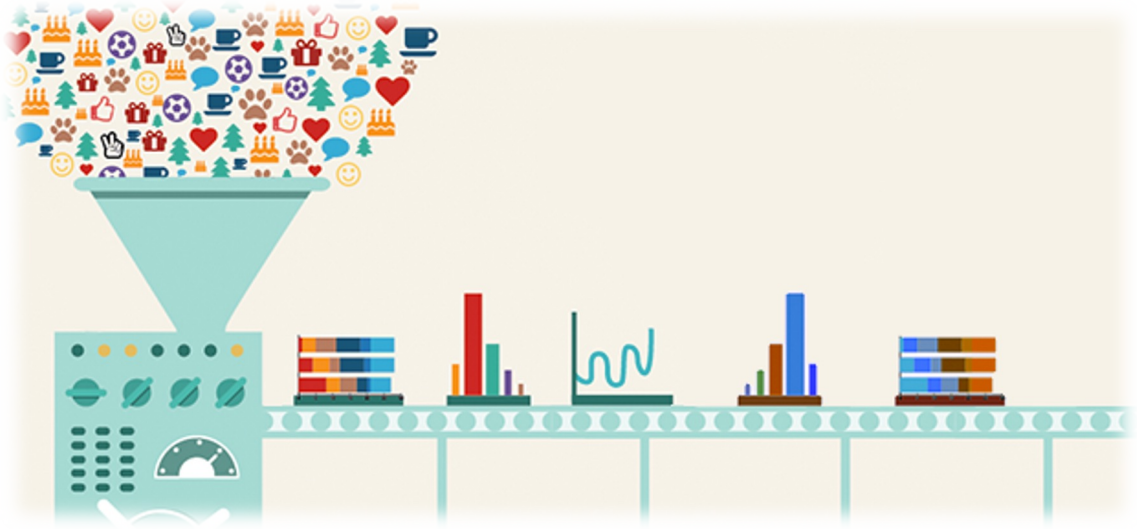
Data Mart and Data Sources



Data Driven Project

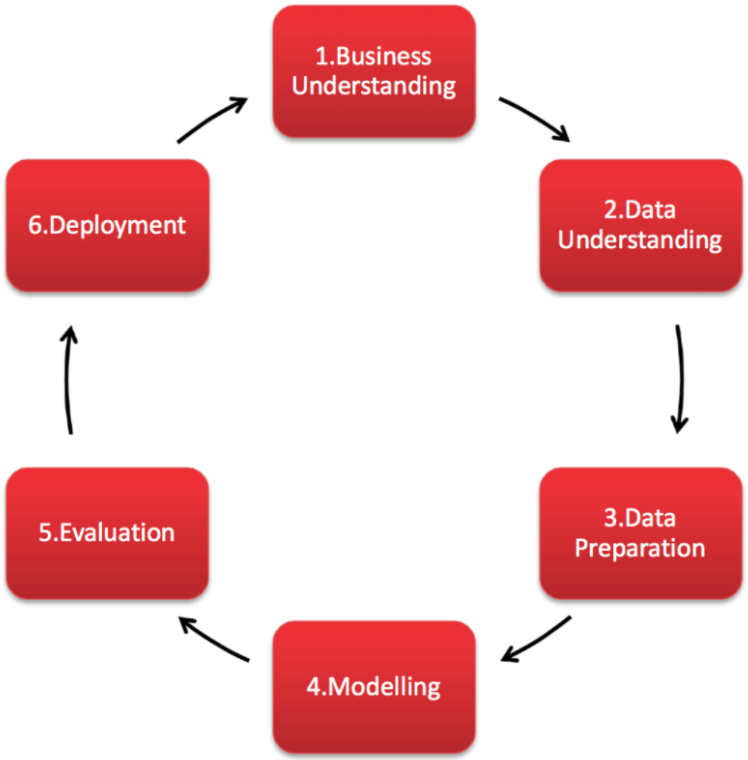
When a company employs a “data-driven” approach, it means it makes **strategic decisions** based on data analysis and interpretation.

A **data-driven approach** enables companies to examine and organize their data with the goal of better serving their customers and consumers.



CRISP-DM

Cross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining, known as CRISP-DM, is a model that describes common approaches used by data mining experts. It is the most widely-used analytics model. CRISP breaks the process of data mining into six major phases.



POC

Proof Of Concept

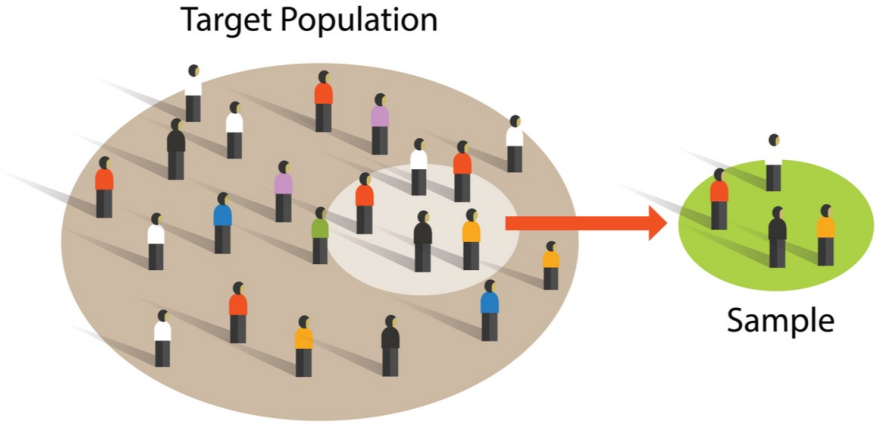
Proof of concept (**POC**), also known as **proof of principle**, is a realization of a certain method or idea in order to demonstrate its feasibility or a demonstration in principle with the aim of verifying that some concept or theory has practical potential.

- A proof of concept is usually small and may or may not be complete.
- Test feasibility of business concepts and proposals to solve business problems and accelerate business innovation goals

Tips for Successful POC

A POC that doesn’t use your own data, doesn’t prove anything

- Limit the scope of data sources involved
- Trimming the data down
- Using sample data sets instead of your own data



Tips for Successful POC

Do not get distracted by pretty visuals

With visualization software components a dime a dozen, a vendor can easily ‘fake’ these pretty graphics.

Address future and present requirements

it doesn't matter who you are or how much experience you have – it is almost impossible to know in advance what your future requirements will be.

BI requirements tend to be highly dynamic because businesses change all the time and business users are continually refining and adjusting their requirements.

Tips for Successful POC

Consult your own IT professionals, even if they are not directly involved

In many organizations business analytics solutions are already set up that are highly reliant on IT. Business professionals that are not being able to extract relevant data quickly and independently, will look for BI solutions that cut IT completely out of the loop.

A POC should not require you to make a financial investment

Demand one solid report or dashboard running over your own data before you agree to any financial commitment.

Mockup

Mockup

Mockup is a scale or full-size model of a design or device, used for teaching, demonstration, design evaluation, promotion, and other purposes.

A mockup may be a *prototype* if it provides at least part of the functionality of a system and enables testing of a design.

Mock-ups are used by designers mainly to acquire feedback from users.

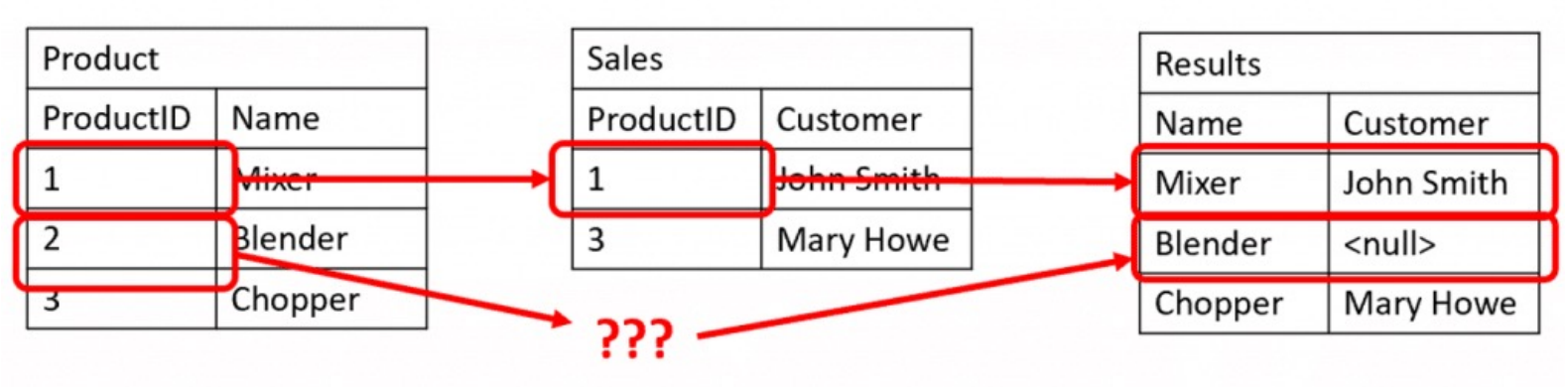
Optimizing Queries in SQL

Or Peretz

Merge Multiple Sources

- Merge different data sources by key

- Can be applied by JOIN or SubQueries



JOIN vs SubQueries

Use both Joins and subqueries to query data from different tables.

Though they may even share the same query plan, are many differences between them.

Knowing the differences and when to use either a join or subquery to search data from one or more tables is key to mastering SQL.

Subquery

A common use for a subquery may be to calculate a summary value for use in a query. Notice how the subqueries are queries unto themselves.

```
SELECT ProductID, Name, ListPrice, (SELECT AVG(ListPrice)
                                    FROM Production.Product) AS AvgListPrice
FROM Production.Product
WHERE ListPrice > (SELECT AVG(ListPrice)
                   FROM Production.Product)
```

JOIN

Combine rows from one or more tables based on a match condition.

```
SELECT Product.Name, ProductModel.Name AS ModelName
FROM Production.Product JOIN Production.ProductModel
ON Product.ProductModelID = ProductModel.ProductModelID
```

Types: left, right, full, inner

JOIN vs SubQueries

- The subquery returns a single result, which then filters the records.
- The join is an integral part of the select statement.
- Join can not stand on its own as a subquery can.
- Joins are used (mostly) in the FROM clause
- Subquries can appear anywhere

Advantages of JOIN

- Executes faster than subquery (running time). The retrieval time of the query using joins almost always will be faster than that of a subquery.
- Can maximize the calculations on the database.

  For example, instead of multiple queries using one join query.
- Have many types of joins.

Disadvantages of JOIN

- Not easy to read as subqueries.
- More joins in a query means the database server has to do more work
- Can be confusing which join is the appropriate type of join to use to yield the correct desired result set.
- Joins cannot be avoided when retrieving data from a normalized database

Advantages of SubQuery

- Divide the complex query into isolated parts so that a complex query can be broken down into a series of logical steps.
- Easy to understand and maintenance the code.
- Subqueries allow you to use the results of another query in the outer query.
- In some cases, subqueries can replace complex joins and unions.

Disadvantages of SubQuery

- The optimizer is more mature for MYSQL for joins than for subqueries.
- A subquery can be executed more efficiently if you rewrite it as join.
- Can not modify a table and select from the same table within a subquery in the same SQL statement.