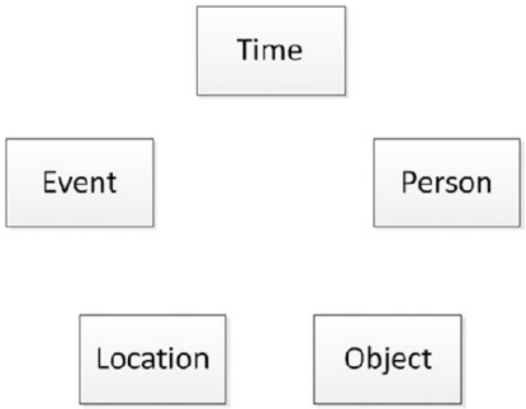
Data Science

Unit 4 Chapter 1: Process Superstep:

- **Introduction:**
- The Process superstep adapts the assess results of the retrieve versions of the data sources into a highly structured data vault that will form the basic data structure for the rest of the data science steps.

- **Five categories of data:**

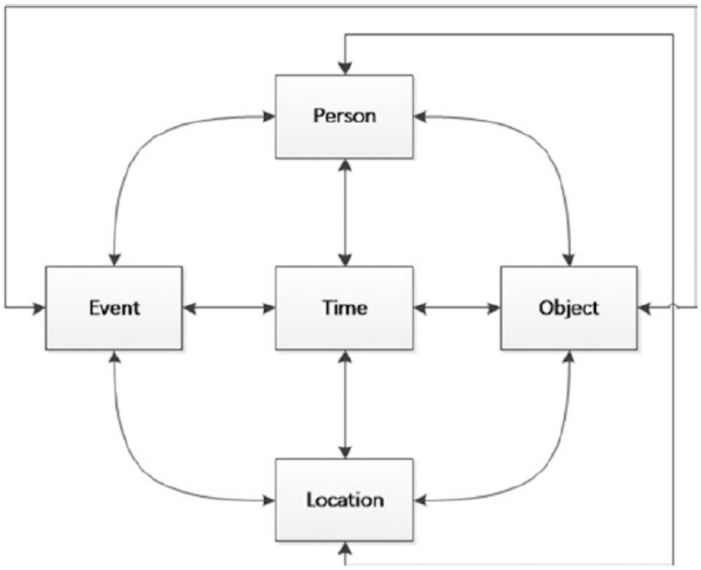


- **Data Vault:**
- The data structure is designed to be responsible for long-term historical storage of data from multiple operational systems.
- It supports chronological historical data tracking for full auditing and enables parallel loading of the structures.
- **Hubs:**
- Data vault hubs contain a set of unique business keys that normally do not change over time.
- Hubs hold a surrogate key for each hub data entry and metadata labeling the source of the business key.
- **Links:**
- Data vault links are associations between business keys.
- These links are essentially many-to-many joins, with additional metadata to enhance the particular link.
- **Satellites**
- Data vault satellites hold the chronological and descriptive characteristics for a specific section of business data.
- Satellites consist of characteristics and metadata linking them to their specific hub.
- Metadata labeling the origin of the association and characteristics, along with a time line with start and end dates for the characteristics, is put in safekeeping, for future use from the data section.
- Each satellite holds an entire chronological history of the data entities within the specific satellite.

- **Reference Satellites:**
- Reference satellites are referenced from satellites but under no circumstances bound with metadata for hub keys.
- They prevent redundant storage of reference characteristics that are used regularly by other satellites.
- Typical reference satellites are:
- Standard codes: These are codes such as ISO 3166 for country codes, ISO 4217 for currencies, and ISO 8601 for time zones.

- Fixed lists for specific characteristics: These can be standard lists that reduce other standard lists. For example, the list of countries your business has offices in may be a reduced fixed list from the ISO 3166 list.
- Conversion lookups: Look at Global Positioning System (GPS) transformations.

- **Time-Person-Object-Location-Event Data Vault**
- The data vault we use is based on the Time-Person-Object-Location-Event (T-P-O-L-E) design principle.



- **Time Section:**
- The time section contains the complete data structure for all data entities related to recording the time at which everything occurred.
- **Time Hub**
- The time hub consists of the following fields:
- CREATE TABLE [Hub-Time] (
- IDNumber VARCHAR (100) PRIMARY KEY,
- IDTimeNumber Integer,
- ZoneBaseKey VARCHAR (100),
- DateTimeKey VARCHAR (100),
- DateTimeValue DATETIME
-);

- **Time Links:**
- The time links link the time hub to the other hubs
- The following links are supported.
- **1) Time-Person Link:**
- This connects date-time values within the person hub to the time hub.
- Dates such as birthdays, marriage anniversaries, and the date of reading this book can be recorded as separate links in the data vault.
- The normal format is BirthdayOn, MarriedOn, or ReadBookOn. The format is simply a pair of keys between the time and person hubs.
- **2) Time-Object Link**
- This connects date-time values within the object hub to the time hub.
- Dates such as those on which you bought a car, sold a car, and read this book can be recorded as separate links in the data vault.
- The normal format is BoughtCarOn, SoldCarOn, or ReadBookOn. The format is simply a pair of keys between the time and object hubs.

- **3) Time-Location Link:**
- This connects date-time values in the location hub to the time hub.
- Dates such as moved to post code SW1, moved from post code SW1, and read book at post code SW1 can be recorded as separate links in the data vault.
- The normal format is MovedToPostCode, MovedFromPostCode, or ReadBookAtPostCode. The format is simply a pair of keys between the time and location hubs.
- **4)Time-Event Link:**
- This connects date-time values in the event hub with the time hub.
- Dates such as those on which you have moved house and changed vehicles can be recorded as separate links in the data vault.
- The normal format is MoveHouse or ChangeVehicle. The format is simply a pair of keys between the time and event hubs.

- **Time Satellites:**
- Time satellites are the part of the vault that stores the following fields.
- CREATE TABLE [Satellite-Time-<Time Zone>] (
- IDZoneNumber VARCHAR (100) PRIMARY KEY,
- IDTimeNumber INTEGER,
- ZoneBaseKey VARCHAR (100),
- DateTimeKey VARCHAR (100),
- UTCDateTimeValue DATETIME,
- Zone VARCHAR (100),
- DateTimeValue DATETIME
-);

Person Section:
- The person section contains the complete data structure for all data entities related to recording the person involved.
- **Person Hub:**
- The person hub consists of a series of fields that supports a "real" person. The person hub consists of the following fields:
- CREATE TABLE [Hub-Person] (
- IDPersonNumber INTEGER,
- FirstName VARCHAR (200),
- SecondName VARCHAR (200),
- LastName VARCHAR (200),
- Gender VARCHAR (20),
- TimeZone VARCHAR (100),
- BirthDateKey VARCHAR (100),
- BirthDate DATETIME
-);

- **Person Links:**
- This links the person hub to the other hubs
- **1) Person-Time Link:**
- This link joins the person to the time hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Person-Time] (
- IDPersonNumber INTEGER,

- IDTimeNumber INTEGER,
- ValidDate DATETIME
-);
- **2) Person-Object Link:**
- This link joins the person to the object hub to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Person-Object] (
- IDPersonNumber INTEGER,
- IDObjectNumber INTEGER,
- ValidDate DATETIME
-);
- **3)Person-Location Link:**
- This link joins the person to the location hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Person-Time] (
- IDPersonNumber INTEGER,
- IDLocationNumber INTEGER,
- ValidDate DATETIME
-);
- **4)Person-Event Link:**
- This link joins the person to the event hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Person-Time] (
- IDPersonNumber INTEGER,
- IDEventNumber INTEGER,
- ValidDate DATETIME
-);

- **Person Satellites:**
- The person satellites are the part of the vault that stores the temporal attributes and descriptive attributes of the data. The satellite is of the following format:
- CREATE TABLE [Satellite-Person-Gender] (
- PersonSatelliteID VARCHAR (100),
- IDPersonNumber INTEGER,
- FirstName VARCHAR (200),
- SecondName VARCHAR (200),
- LastName VARCHAR (200),
- BirthDateKey VARCHAR (20),
- Gender VARCHAR (10),
-);

Object Section:

- The object section contains the complete data structure for all data entities related to recording the object involved.

- **Object Hub:**
- The object hub consists of a series of fields that supports a "real" object. The object hub consists of the following fields:
- CREATE TABLE [Hub-Object-Species] (
- IDObjectNumber INTEGER,
- ObjectBaseKey VARCHAR (100),
- ObjectNumber VARCHAR (100),
- ObjectValue VARCHAR (200),
-);

- **Object Links:**
- These link the object hub to the other hubs
- **1) Object-Time Link:**
- This link joins the object to the time hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Object-Time] (
- IDObjectNumber INTEGER,
- IDTimeNumber INTEGER,
- ValidDate DATETIME
-);
- **2)Object-Person Link:**
- This link joins the object to the person hub to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Object-Person] (
- IDObjectNumber INTEGER,
- IDPersonNumber INTEGER,
- ValidDate DATETIME
-);
- **3)Object-Location Link:**
- This link joins the object to the location hub, to describe the relationships between the two hubs. The link consists of the following fields:
- CREATE TABLE [Link-Object-Location] (
- IDObjectNumber INTEGER,
- IDLocationNumber INTEGER,
- ValidDate DATETIME
-);

- **Object-Event Link:**
- This link joins the object to the event hub to describe the relationships between the two hubs.

- **Object Satellites:**
- Object satellites are the part of the vault that stores and provisions the detailed characteristics of objects.
- The typical object satellite has the following data fields:
- CREATE TABLE [Satellite-Object-Make-Model] (
- IDObjectNumber INTEGER,
- ObjectSatelliteID VARCHAR (200),

- ObjectType VARCHAR (200),
- ObjectKey VARCHAR (200),
- ObjectUUID VARCHAR (200),
- Make VARCHAR (200),
- Model VARCHAR (200)
-);

Location Section:

- The location section contains the complete data structure for all data entities related to recording the location involved.
- **Location Hub:**
- The location hub consists of a series of fields that supports a GPS location.
- The location hub consists of the following fields:
- CREATE TABLE [Hub-Location] (
- IDLocationNumber INTEGER,
- ObjectBaseKey VARCHAR (200),
- LocationNumber INTEGER,
- LocationName VARCHAR (200),
- Longitude DECIMAL (9, 6),
- Latitude DECIMAL (9, 6)
-);

- **Location Links:**
- The location links join the location hub to the other hubs
- **1)Location-Time Link:**
- The link joins the location to the time hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Location-Time] (
- IDLocationNumber INTEGER,
- IDTimeNumber INTEGER,
- ValidDate DATETIME
-);
- These links support business actions such as ArrivedAtShopAtDateTime or ShopOpensAtTime.
- **2)Location-Person Link:**
- This link joins the location to the person hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Location-Person] (
- IDLocationNumber INTEGER,
- IDPersonNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as ManagerAtShop or SecurityAtShop.
- **3)Location-Object Link:**
- This link joins the location to the object hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Location-Object] (

- IDLocationNumber INTEGER,
- IDObjectNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as ShopDeliveryVan or RackAtShop.
- **4)Location-Event Link:**
- This link joins the location to the event hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Location-Event] (
- IDLocationNumber INTEGER,
- IDEventNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as ShopOpened or PostCodeDeliveryStarted.

- **Location Satellites:**
- The location satellites are the part of the vault that stores and provisions the detailed characteristics of where entities are located. The typical location satellite has the following data fields:
- CREATE TABLE [Satellite-Location-PostCode] (
- IDLocationNumber INTEGER,
- LocationSatelliteID VARCHAR (200),
- LocationType VARCHAR (200),
- LocationKey VARCHAR (200),
- LocationUUID VARCHAR (200),
- CountryCode VARCHAR (20),
- PostCode VARCHAR (200)
-);

- **Event Section:**
- The event section contains the complete data structure for all data entities related to recording the event that occurred.
- **Event Hub:**
- The event hub consists of a series of fields that supports events that happens in the real world.
- The event hub consists of the following fields:
- CREATE TABLE [Hub-Event] (
- IDEventNumber INTEGER,
- EventType VARCHAR (200),
- EventDescription VARCHAR (200)
-);

Event Links:
- Event links join the event hub to the other hubs
- **1) Event-Time Link:**
- This link joins the event to the time hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Event-Time] (

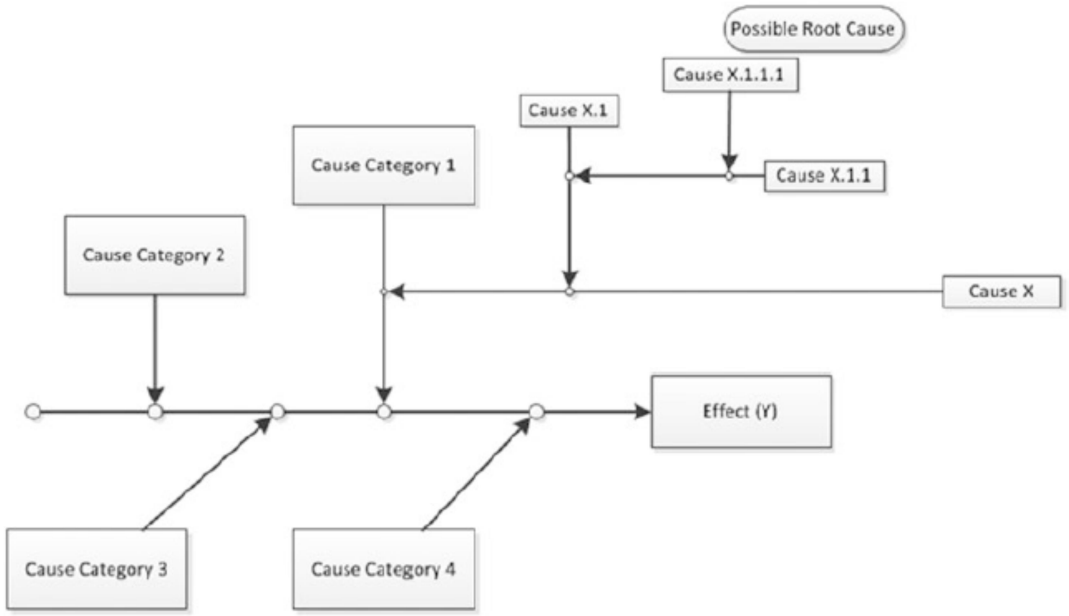
- IDEventNumber INTEGER,
- IDTimeNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as DeliveryDueAt or DeliveredAt.
- **2)Event-Person Link:**
- This link joins the event to the person hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Event-Person] (
- IDEventNumber INTEGER,
- IDPersonNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as ManagerAppointAs or StaffMemberJoins.
- **3) Event-Object Link:**
- This link joins the event to the object hub, to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Event-Object] (
- IDEventNumber INTEGER,
- IDObjectNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as VehicleBuy, VehicleSell, or ItemInStock.
- **4)Event-Location Link:**
- The link joins the event to the location hub to describe the relationships between the two hubs.
- The link consists of the following fields:
- CREATE TABLE [Link-Event-Location] (
- IDEventNumber INTEGER,
- IDTimeNumber INTEGER,
- ValidDate DATETIME
-);
- These links support such business actions as DeliveredAtPostCode or PickupFromGPS.

- **Event Satellites:**
- The event satellites are the part of the vault that stores the details related to all the events that occur within the systems you will analyze with your data science.

- **Data Science Process:**
- **Roots of Data Science:**
- Data science is at its core about curiosity and inquisitiveness.
- This core is rooted in the 5 Whys.
- The 5 Whys is a technique used in the analysis phase of data science.
- **Benefits of the 5 Whys:**
- The 5 Whys assist the data scientist to identify the root cause of a problem and determine the relationship between different root causes of the same problem.
- It is one of the simplest investigative tools—easy to complete without intense statistical analysis.

- **When Are the 5 Whys Most Useful?**
- The 5 Whys are most useful for finding solutions to problems that involve human factors or interactions that generate multilayered data problems.
- In day-to-day business life, they can be used in real-world businesses to find the root causes of issues.
- **How to Complete the 5 Whys:**
- Write down the specific problem. This will help you to formalize the problem and describe it completely.
- It also helps the data science team to focus on the same problem.
- Ask why the problem occurred and write the answer below the problem.
- If the answer you provided doesn't identify the root cause of the problem that you wrote down first, ask why again, and write down that answer.
- Loop back to the preceding step until you and your customer are in agreement that the problem's root cause is identified.
- Again, this may require fewer or more than the 5 Whys.

- **Fishbone Diagrams:**
- The diagram is drawn up as you complete the 5 Whys process, as you will discover that there are normally many causes for why specific facts have been recorded.



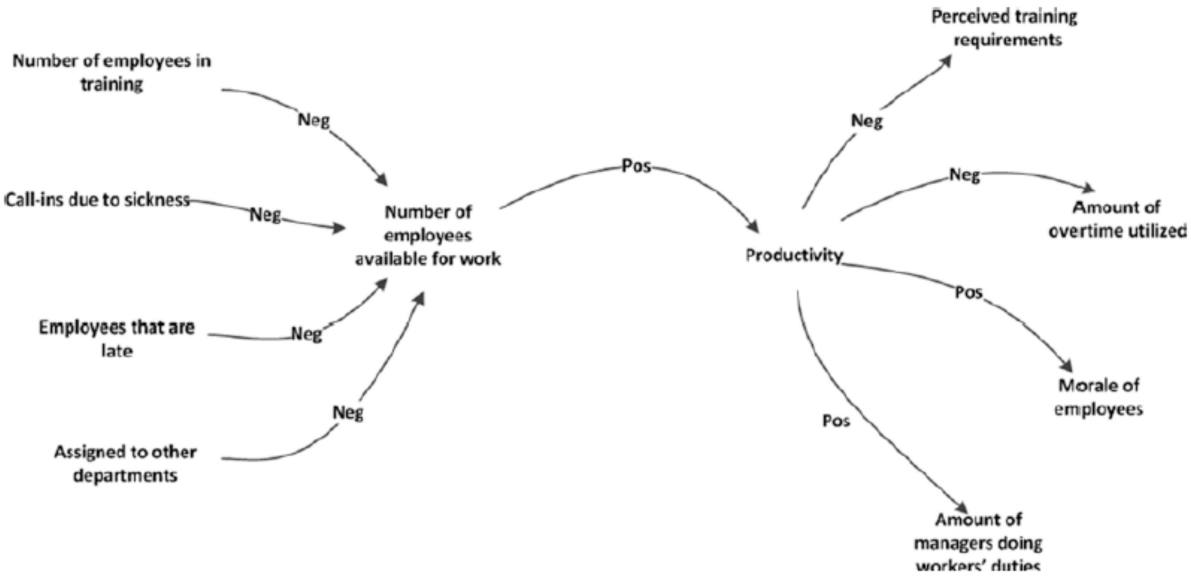
- The ten cans are the effect (Y), but the four root causes of the purchase are
- 1) I was hungry, so I bought ten tins. I did not like the brand of curry that I bought 10 cans of the previous week.
- 2) My neighbor needed five cans, as she was no longer able to walk, and she requested the brand that I purchased.
- 3) I fed two cans to the dog, because I feel dog food is not nutritious, but I was not prepared to buy a more expensive brand of canned beef curry for the dog.
- 4) I put three cans in the charity bin outside the local school.

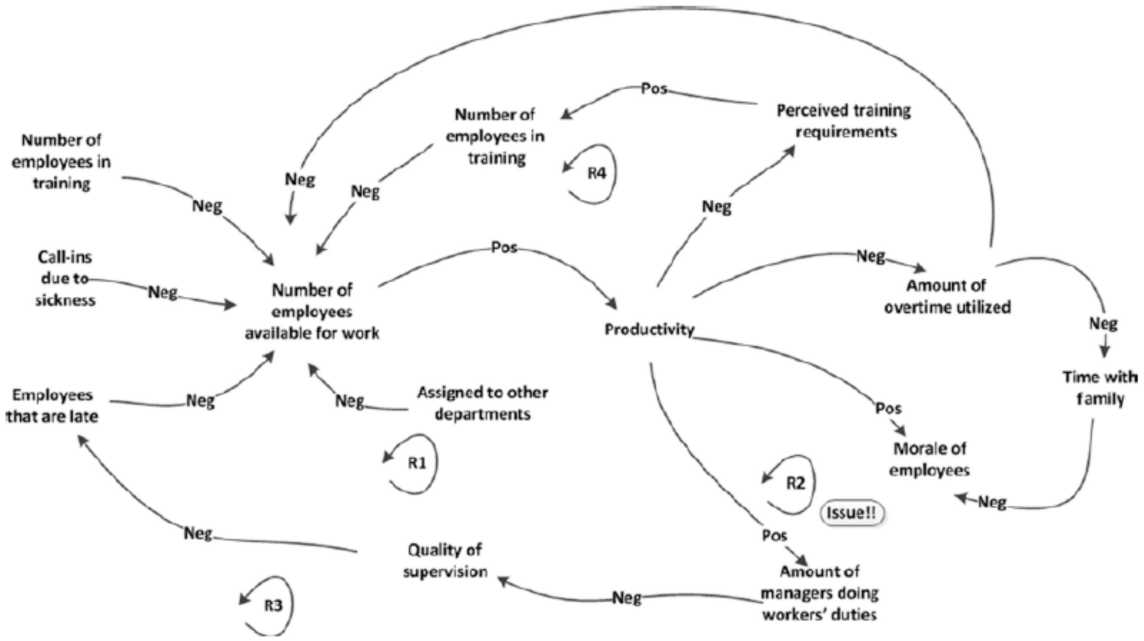
- **5 Whys Example:**
- Problem Statement: Customers are unhappy because they are being shipped products that don't meet their specifications.
- **1. Why are customers being shipped bad products?**
- • Because manufacturing built the products to a specification that is different from what the customer and the salesperson agreed to.

- **2. Why did manufacturing build the products to a different specification than that of sales?**
- • Because the salesperson accelerates work on the shop floor by calling the head of manufacturing directly to begin work.
- An error occurred when the specifications were being communicated or written down.
- **3. Why does the salesperson call the head of manufacturing directly to start work instead of following the procedure established by the company?**
- • Because the "start work" form requires the sales director's approval before work can begin and slows the manufacturing process (or stops it when the director is out of the office).
- **4. Why does the form contain an approval for the sales director?**
- • Because the sales director must be continually updated on sales for discussions with the CEO, as my retailer customer was a topten key account.
- In this case, only four whys were required to determine that a non-value-added signature authority helped to cause a process breakdown in the quality assurance for a key account.

- **Monte Carlo Simulation:**
- This technique performs analysis by building models of possible results, by substituting a range of values—a probability distribution—for parameters that have inherent uncertainty.
- It then calculates results over and over, each time using a different set of random values from the probability functions.
- Depending on the number of uncertainties and the ranges specified for them, a Monte Carlo simulation can involve thousands or tens of thousands of recalculations before it is complete.
- Monte Carlo simulation produces distributions of possible outcome values.
- As a data scientist, this gives you an indication of how your model will react under real-life situations.
- It also gives the data scientist a tool to check complex systems, wherein the input parameters are high-volume or complex.

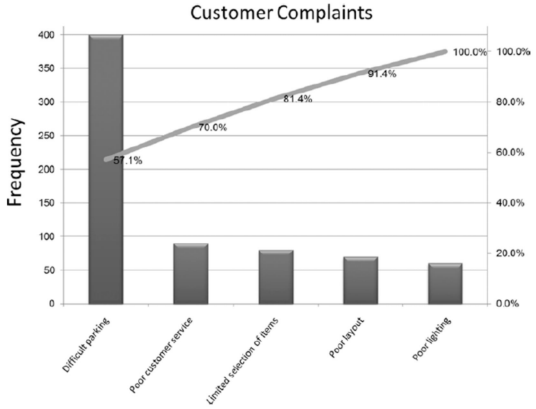
- **Causal Loop Diagrams:**
- A causal loop diagram (CLD) is a causal diagram that aids in visualizing how a number of variables in a system are interrelated and drive cause-and-effect processes.
- The diagram consists of a set of nodes and edges.
- Nodes represent the variables, and edges are the links that represent a connection or a relation between the two variables.
- Example: The challenge is to keep the "Number of Employees Available to Work and Productivity" as high as possible.



- 

- ***Figure 9-12.*** *Monte Carlo result*
- The result was "Managers need to manage not work." The R2—percentage of manage doing employees' duties—was the biggest cause and impact driver in the system.

- **Pareto Chart:**
- Used to perform a rapid processing plan for the data science.
- Pareto charts can be constructed by segmenting the range of the data into groups (also called segments, bins, or categories).

- 

- Questions the Pareto chart answers:
- • What are the largest issues facing our team or my customer’s business?
- • What 20% of sources are causing 80% of the problems (80/20 Rule)?
- • Where should we focus our efforts to achieve the greatest improvements?

- **Forecasting:**
- Forecasting is the ability to project a possible future, by looking at historical data.
- The data vault enables these types of investigations, owing to the complete history it collects as it processes the source’s systems data.
- You will perform many forecasting projects during your career as a data scientist and supply answers to such questions as the following:
- • What should we buy?
- • What should we sell?
- • Where will our next business come from?
- People want to know what you calculate to determine what is about to happen.

- **Data Science:**
- You must understand that data science works best when you follow approved algorithms and techniques.
- **data science that works follows these basic steps:**
- 1. Start with a question. Make sure you have fully addressed the 5 Whys.
- 2. Follow a good pattern to formulate a model.
- Formulate a model, guess a prototype for the data, and start a virtual simulation of the real-world parameters.
- Mix some mathematics and statistics into the solution, and you have the start of a data science model.
- 3. Gather observations and use them to generate a hypothesis.
- Start the investigation by collecting the required observations, as per your model.
- Process your model against the observations and prove your hypothesis to be true or false.
- 4. Use real-world evidence to judge the hypothesis.
- Relate the findings back to the real world and, through storytelling, convert the results into real-life business advice and insights.
- 5. Collaborate early and often with customers and with subject matter experts along the way.
- You also must communicate early and often with your relevant experts to ensure that you take them with you along the journey of discovery.
- Businesspeople want to be part of solutions to their problems. Your responsibility is to supply good scientific results to support the business.