Introduction to Data Science

Lecture 02

Data Quality

Data Quality is an essential characteristic that determines the reliability of data for making decisions.

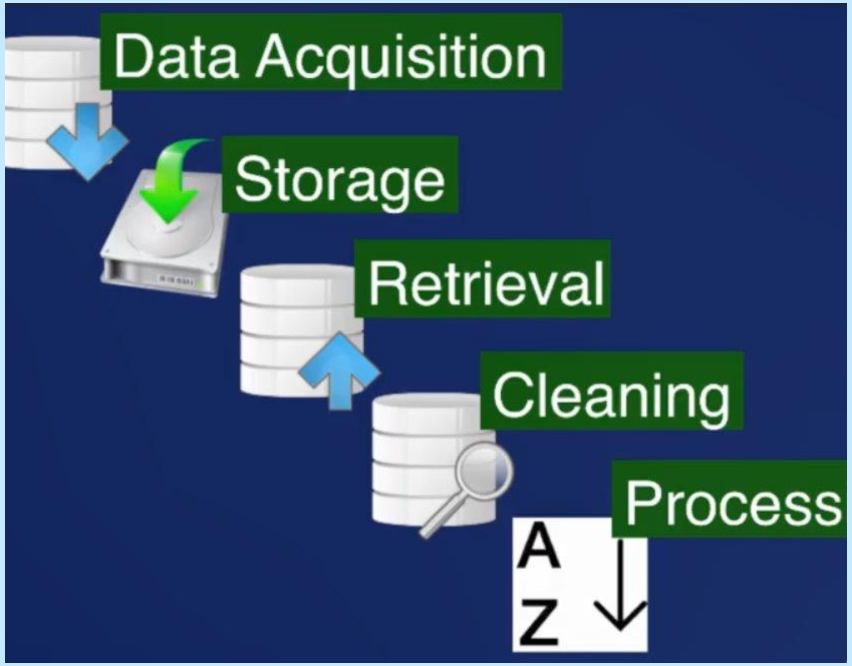
http://bigdata.black/training/tutorials/what-is-data-quality/

Data Quality

- Timeliness: Does data arrive on time? Does it meet a refreshing schedule? Does it meet the requirements for the time interval from collection to processing to analysis?
- Readability: Is the content and format easy to understand?
- Authorization: Does using the data require certain rights or permissions and what limitations are there?
- Structure: Do you have the technology to transform unstructured data into structured data?
- Credibility: What is your confidence in this data?

Challenges

- Confirmation of unstructured data is often time consuming and costly.
- The costs and time of the process of acquiring, storing, cleaning, retrieving, and processing unstructured data can add up to quite and investment before we can start reaping value from this process



Challenges

Use this information to personalize their communications with their costumers, which in turns leads to better met consumer expectations and happier customers

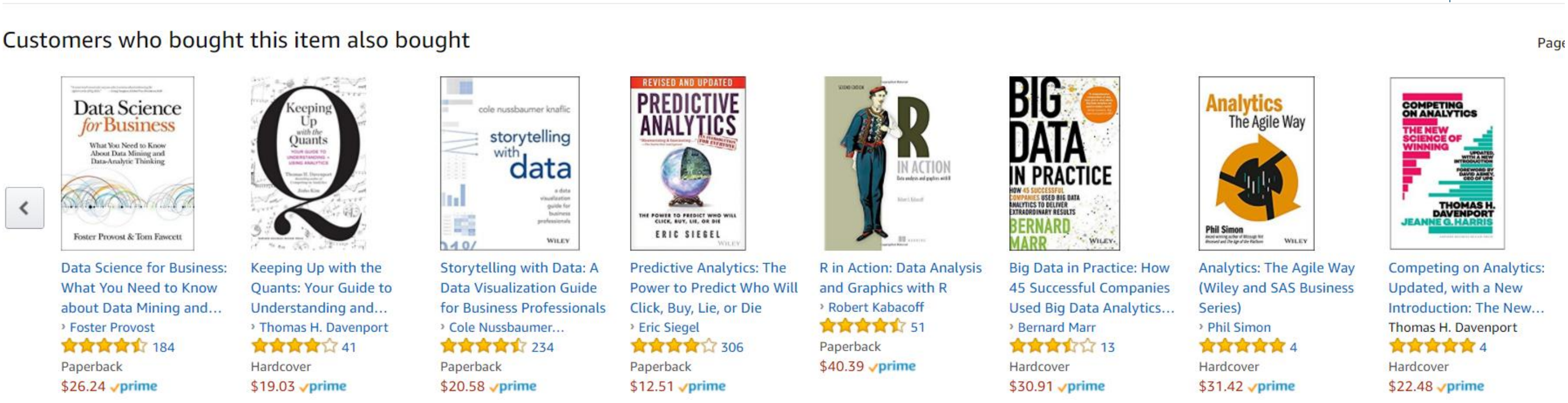
Data enabled personalized marketing. - accessible data through social media sites, like Twitter or Facebook

Through such data, the companies are able to see their purchase history, what they searched for, what they watched, where they have been, and what they're interested in through their likes and shares.

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

https://commons.wikimedia.org/wiki/File:Big_data_cartoon_t_gregorius.jpg

Amazon’s Recommendation Engine



Customers who bought this item also bought

Page

Data Science for Business: What You Need to Know about Data Mining and...
› Foster Provost
★★★★☆ 184
Paperback
$26.24 ✓prime

Keeping Up with the Quants: Your Guide to Understanding and...
› Thomas H. Davenport
★★★★☆ 41
Hardcover
$19.03 ✓prime

Storytelling with Data: A Data Visualization Guide for Business Professionals
› Cole Nussbaumer...
★★★★☆ 234
Paperback
$20.58 ✓prime

Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die
› Eric Siegel
★★★★☆ 306
Paperback
$12.51 ✓prime

R in Action: Data Analysis and Graphics with R
› Robert Kabacoff
★★★★☆ 51
Paperback
$40.39 ✓prime

Big Data in Practice: How 45 Successful Companies Used Big Data Analytics...
› Bernard Marr
★★★☆☆ 13
Hardcover
$30.91 ✓prime

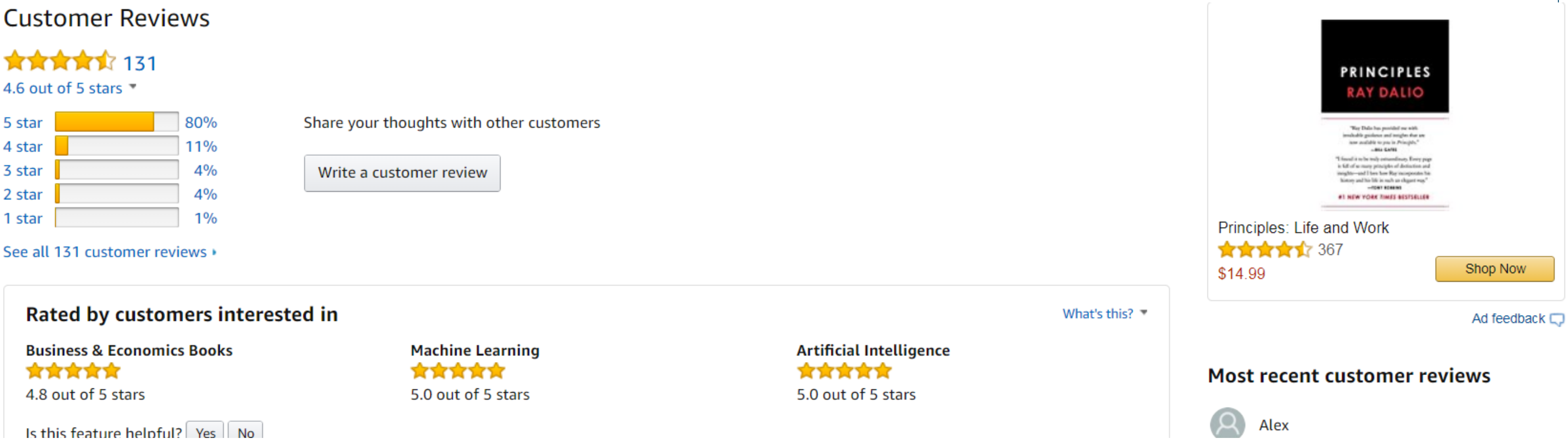
Analytics: The Agile Way (Wiley and SAS Business Series)
› Phil Simon
★★★★★ 4
Hardcover
$31.42 ✓prime

Competing on Analytics: Updated, with a New Introduction: The New...
Thomas H. Davenport
★★★★★ 4
Hardcover
$22.48 ✓prime

One area we are all familiar with are the recommendation engines. These engines leverage user patterns and product features to predict best match product for enriching the user experience. If you ever shopped on Amazon, you know you get recommendations based on your purchase

https://www.amazon.com/s?k=data+science+books&ref=nb_sb_noss_1

Sentiment Analysis



Customer Reviews

131

4.6 out of 5 stars

5 star	80%
4 star	11%
3 star	4%
2 star	4%
1 star	1%

See all 131 customer reviews

Share your thoughts with other customers

Write a customer review

Rated by customers interested in

What's this?

Business & Economics Books	Machine Learning	Artificial Intelligence
4.8 out of 5 stars	5.0 out of 5 stars	5.0 out of 5 stars

Is this feature helpful? Yes No

PRINCIPLES RAY DALIO

Principles: Life and Work

367

$14.99

Shop Now

Ad feedback

Most recent customer reviews

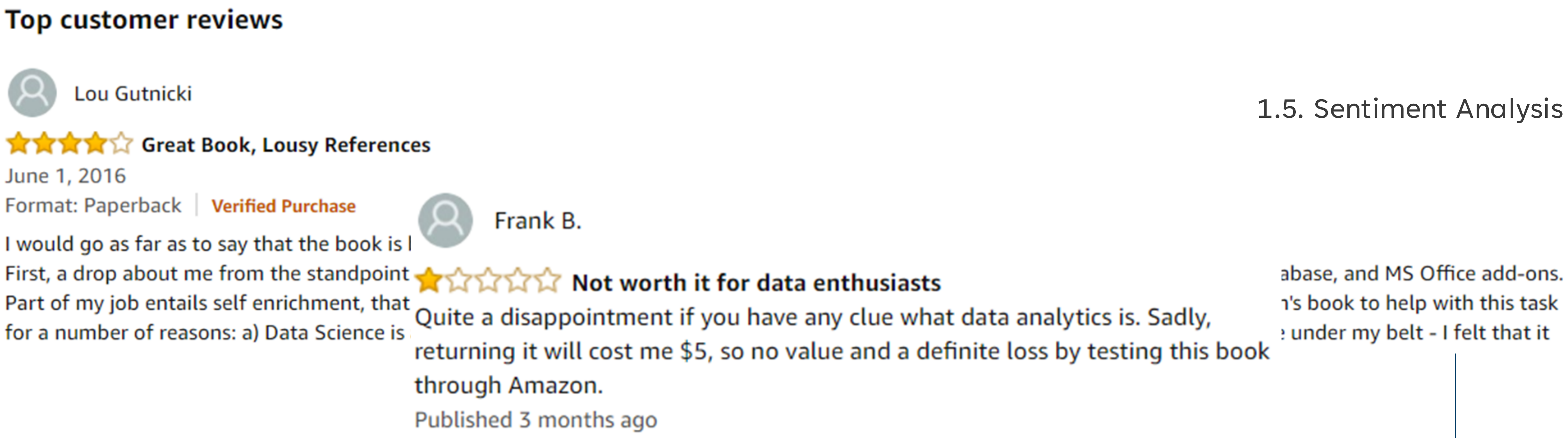
Alex

Another technique that companies use is sentiment analysis, or in simple terms, analysis of the feelings around events and products. Remember the book (Data Science) I purchased on Amazon.com? I not only can read the reviews before purchasing them, I can also write a product review once I receive my plates.

This way, other customers can be informed. But more importantly, Amazon can keep a watch on the product reviews and trends for a particular product. In this case, book on Data Science. For example, they can judge if a product review is positive or negative.

https://www.amazon.com/s?k=data+science+books&ref=nb_sb_noss_1

1.5. Sentiment Analysis



Top customer reviews

Lou Gutnicki

★★★★☆ **Great Book, Lousy References**

June 1, 2016

Format: Paperback | Verified Purchase

I would go as far as to say that the book is l
First, a drop about me from the standpoint ... abase, and MS Office add-ons.
Part of my job entails self enrichment, that ... 's book to help with this task
for a number of reasons: a) Data Science is ... under my belt - I felt that it

Frank B.

★☆☆☆☆ **Not worth it for data enthusiasts**

Quite a disappointment if you have any clue what data analytics is. Sadly, returning it will cost me $5, so no value and a definite loss by testing this book through Amazon.

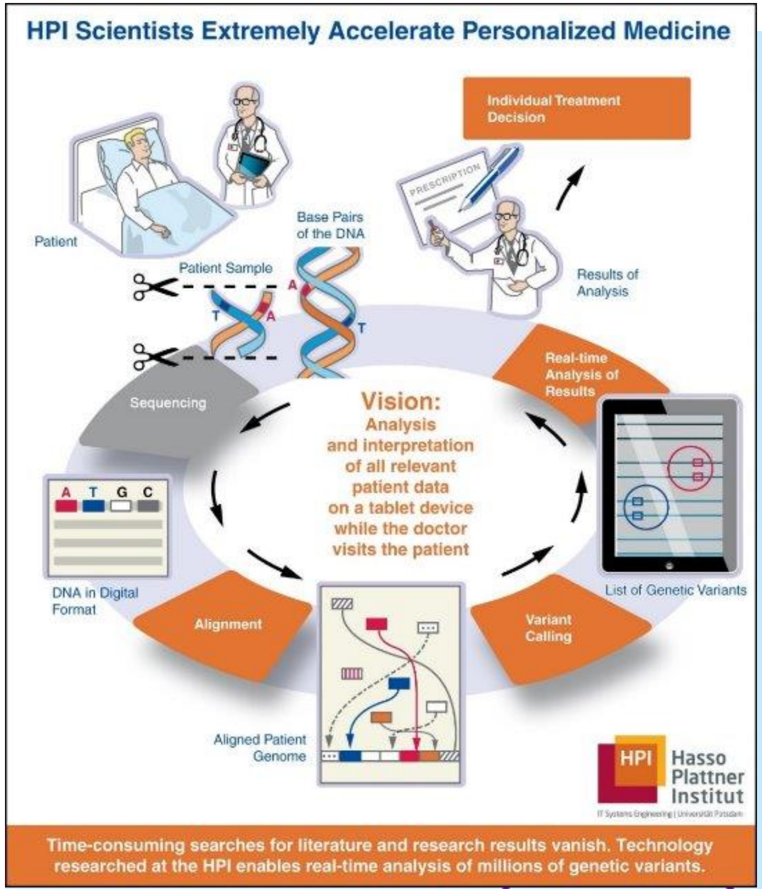
Published 3 months ago

1.5. Sentiment Analysis

Since these reviews are written in English using a technique called natural language processing, and other analytical methods, Amazon can analyze the general opinion of a person or public about such a product.

This is why sentiment analysis often gets referred to as opinion mining. News channels are filled with Twitter feed analysis every time an event of importance occurs, such as elections.

Personalized Medical Treatment



HPI Scientists Extremely Accelerate Personalized Medicine

Individual Treatment Decision
Patient
Base Pairs of the DNA
Patient Sample
Results of Analysis
Real-time Analysis of Results
Sequencing
Vision: Analysis and interpretation of all relevant patient data on a tablet device while the doctor visits the patient
A T G C
DNA in Digital Format
List of Genetic Variants
Alignment
Variant Calling
Aligned Patient Genome
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Time-consuming searches for literature and research results vanish. Technology researched at the HPI enables real-time analysis of millions of genetic variants.

The Human Genome Project is an initiative that aims to map all the 3 billion units of human genome to find the genetic roots of diseases and thereby cure for them

- Satellite (Image Data)
- Research Center/Fire Station (sensor data)
- Social Media (from the people)

Wild fire

https://www.nbcnews.com/news/us-news/worst-california-fires-may-be-over-n899906

Rare Event

The key observation is that, even if the data is random, then with enough data, you will still observe a large number of rare events

- **Rare events** are events that occur with low frequency, and the term is often used in particular reference to infrequent or hypothetical events that have potentially significant impact.

Rare Event

Think about a very basic calculation first. Assume that you toss a coin 10 times. And consider the events that all of these coin tosses are heads

The probability for this is 2 to the power of minus 10,which is 1 over 1,024, which is quite small. So if you only repeat this experiment a few times- let's say two or three times - then the probability that one of these experiments shows all 10 heads is very small

Rare Event

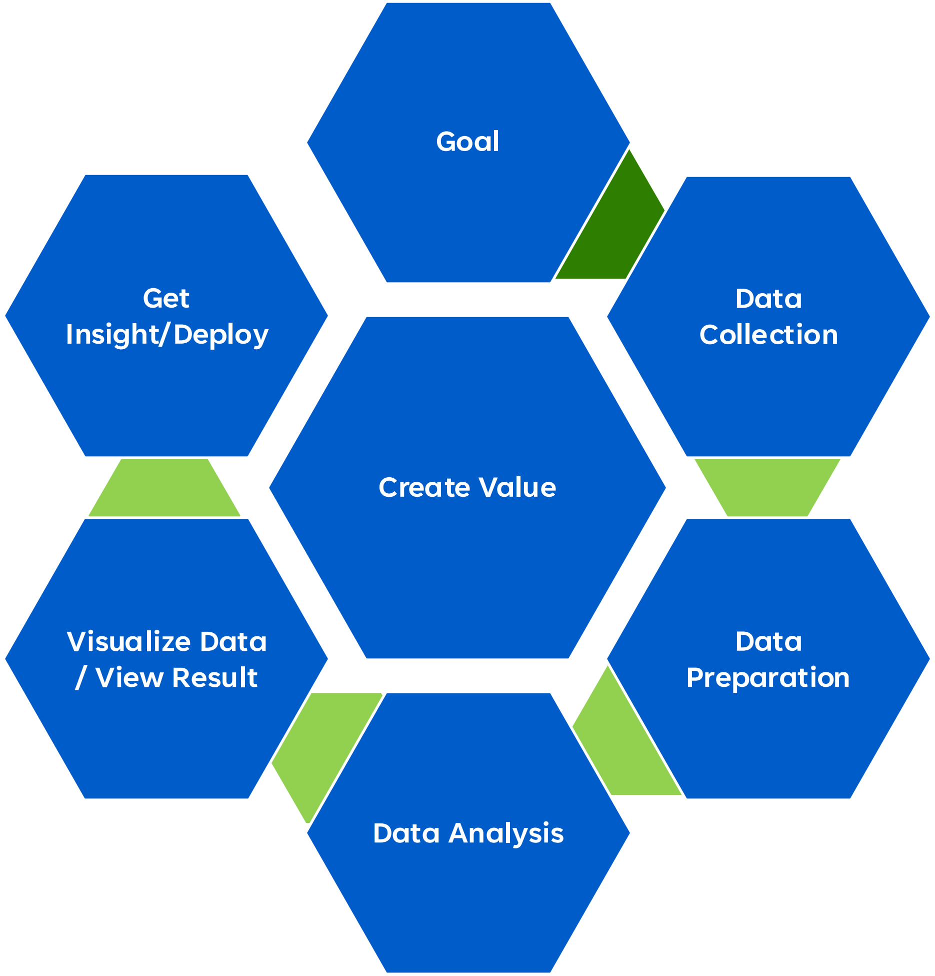
- However, assume that you repeat this experiment 10,000 times. Then the expected number of times you see 10 heads is 10,000 over 1,024, which is approximately 10.
- If you repeat this experiment a billion times, you would expect to see 10 heads approximately 1 million times. This phenomenon often occurs in big data

1.8. Rare Event

Example of Rare events

- Face recognition system
- Earthquake – Tsunami etc

Overview of Data Science Processing



Basic Steps in Data Science Processing

- Define the Goal
- Data Collection
- Data Preparation (Exploring and Cleaning Data)
- Data Analysis
- Visualize Data
- Get Insight (Predictive)
- Deploy ML (Iterate)

Define the Goal

- Problem Definition would be the hardest part – as in most cases we don’t know what we exactly need.

Data Collection

- **Data collection** is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes

http://ixd.prattsi.org/2016/04/1606/

Data Collection

(is the process of gathering and measuring information)

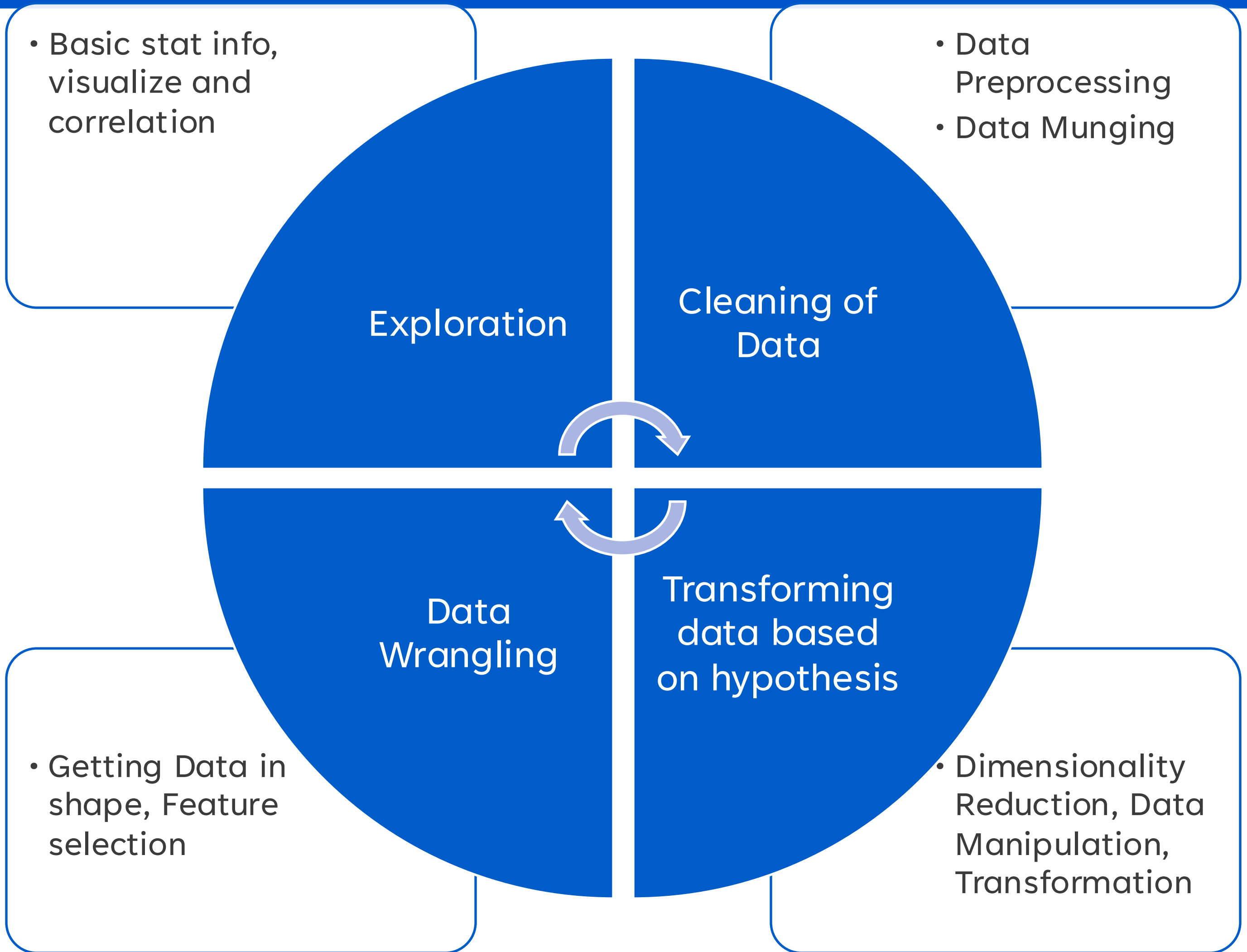
How Data are collected

- Survey
- Experiment
- Observation
- Sampling

From Where :

- Social Media
- Sensors
- Satellite
- Internet
- Telecommunication
- Database
- Open Data

Data Preparation



Exploration

- Basic stat info, visualize and correlation

Cleaning of Data

- Data Preprocessing
- Data Munging

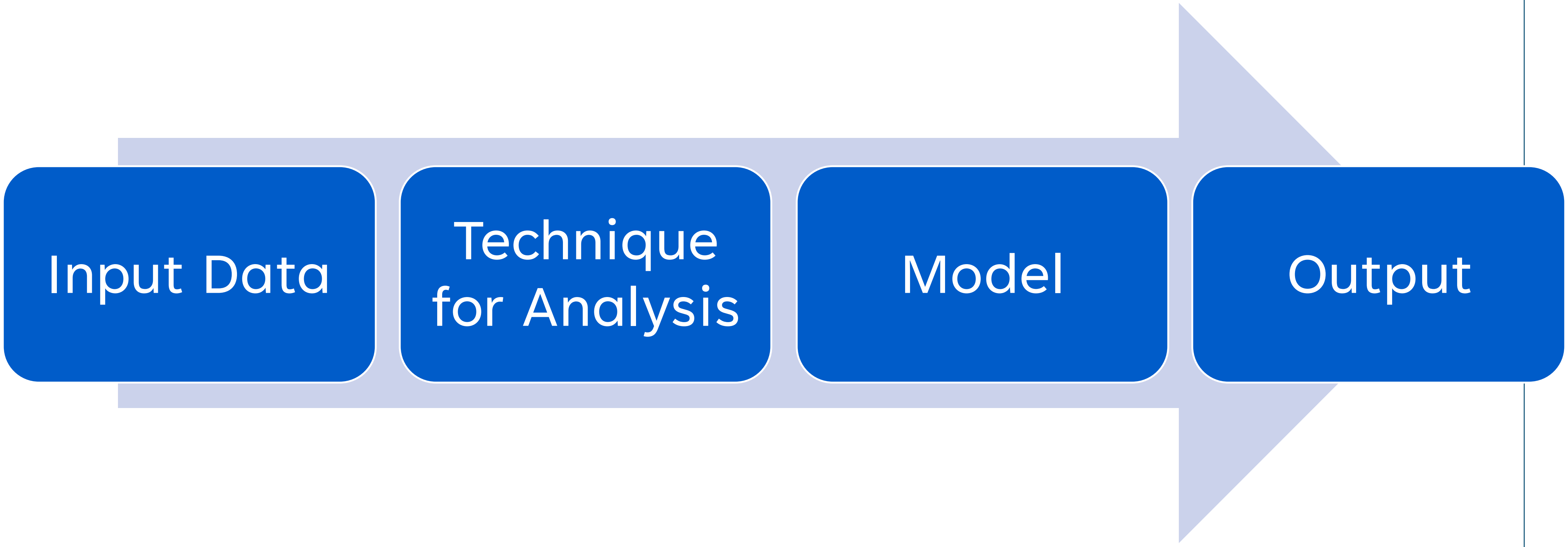
Data Wrangling

- Getting Data in shape, Feature selection

Transforming data based on hypothesis

- Dimensionality Reduction, Data Manipulation, Transformation

Data Analysis

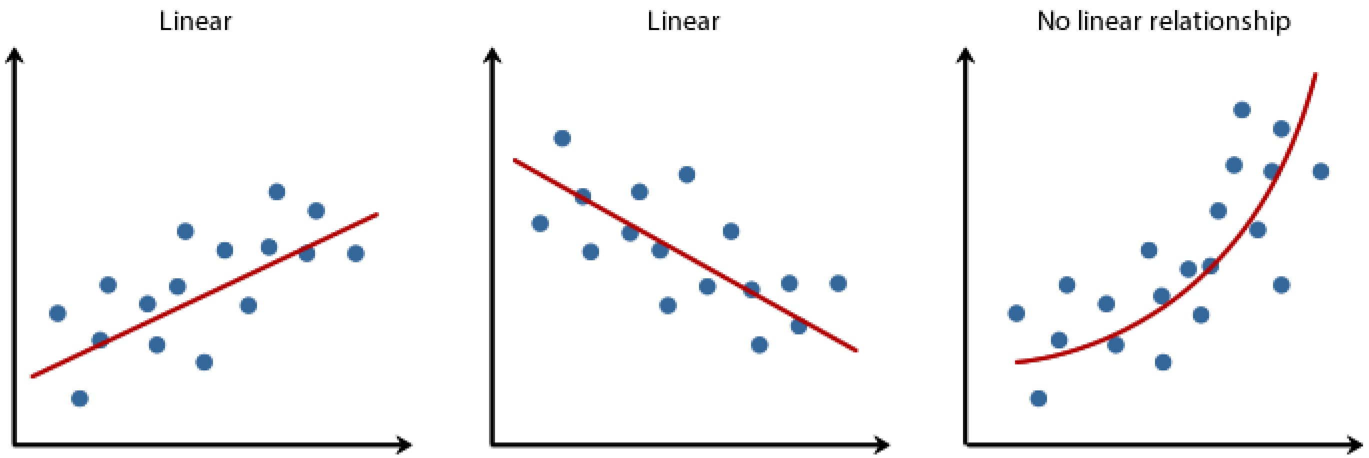

Input Data
Technique for Analysis
Model
Output

Analysis Techniques

- Regression
- Classification
- Clustering
- Association Analysis
- Graph analytics

Regression

- **Regression analysis** is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

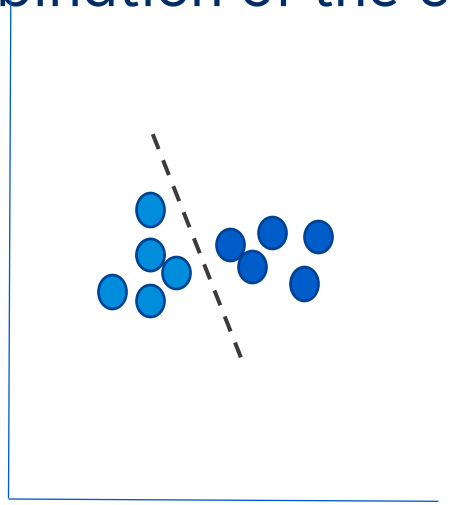


Copyright 2014. Laerd Statistics.

https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

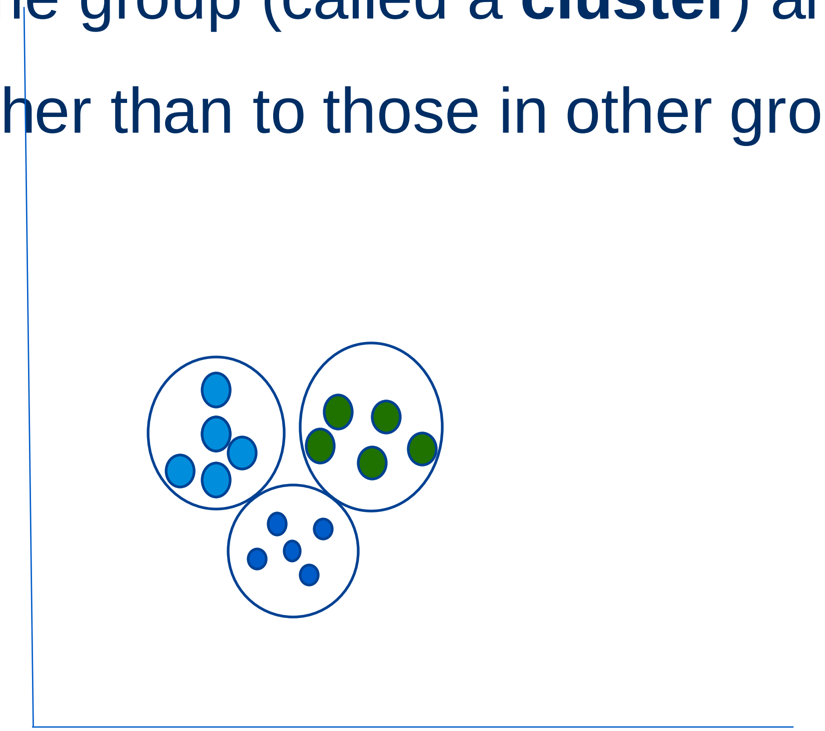
Classification

- Example : A **linear classifier** achieves this by making a classification decision based on the value of a linear combination of the characteristics



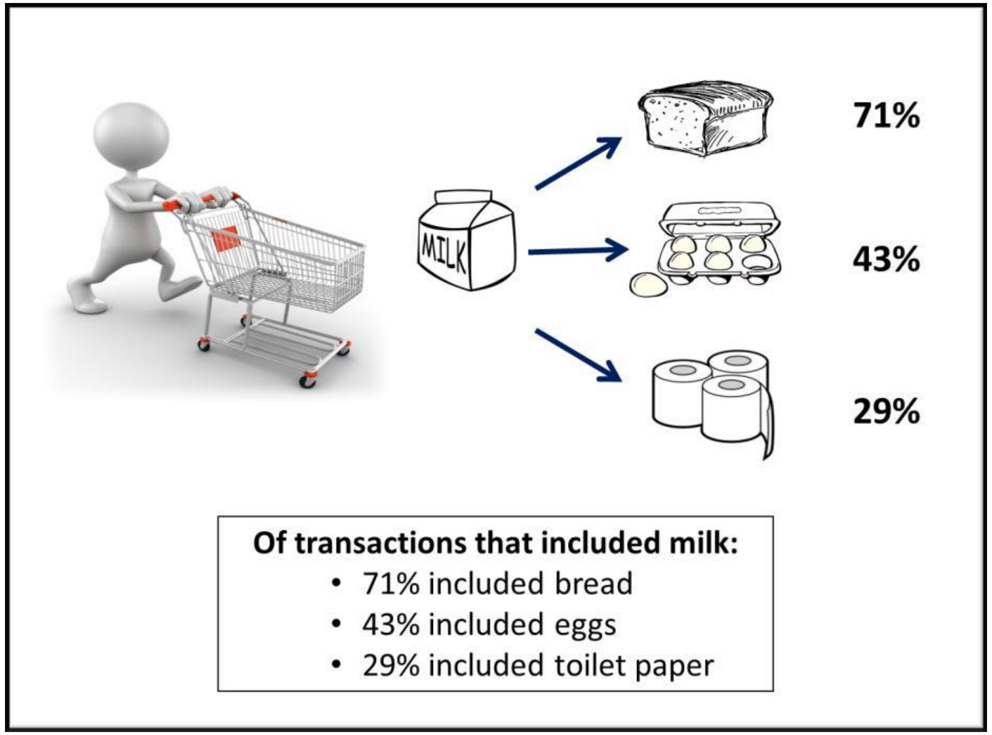
Clustering

- **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters)



Association Analysis

- Retailers can use this type of rules to help them identify new opportunities for cross- selling their products to the customers



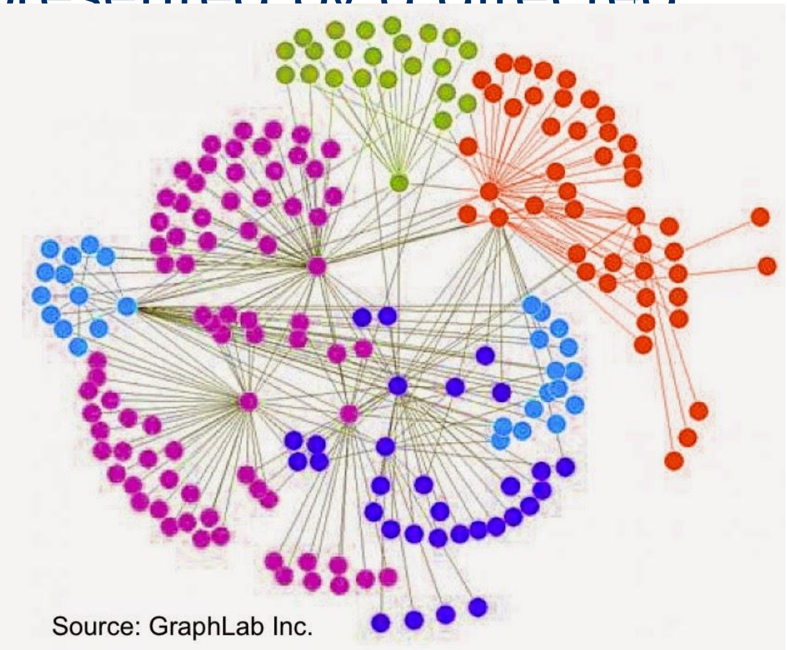
Of transactions that included milk:

- 71% included bread
- 43% included eggs
- 29% included toilet paper

http://users.telenet.be/martialluyts/market.html

Graph Analytics :

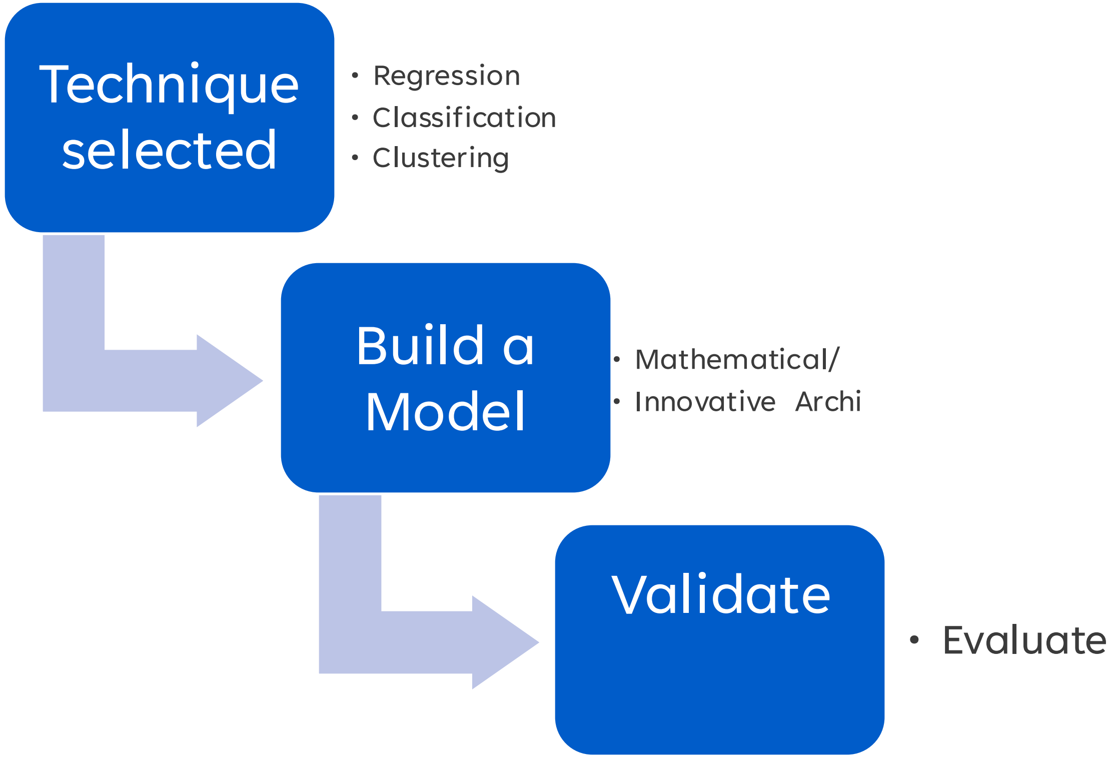
- In computer science, graphs are used to represent networks of communication, data organization, computational devices, the flow of computation, etc.
- For instance, the link structure of a website can be represented by a directed graph, in which the vertices represent web pages and links from one page to another.



Source: GraphLab Inc.

https://www.oreilly.com/ideas/there-are-many-use-cases-for-graph-databases-and-analytics

Verification of Data Analysis



Technique selected

- Regression
- Classification
- Clustering

Build a Model

- Mathematical/
- Innovative Archi

Validate

- Evaluate

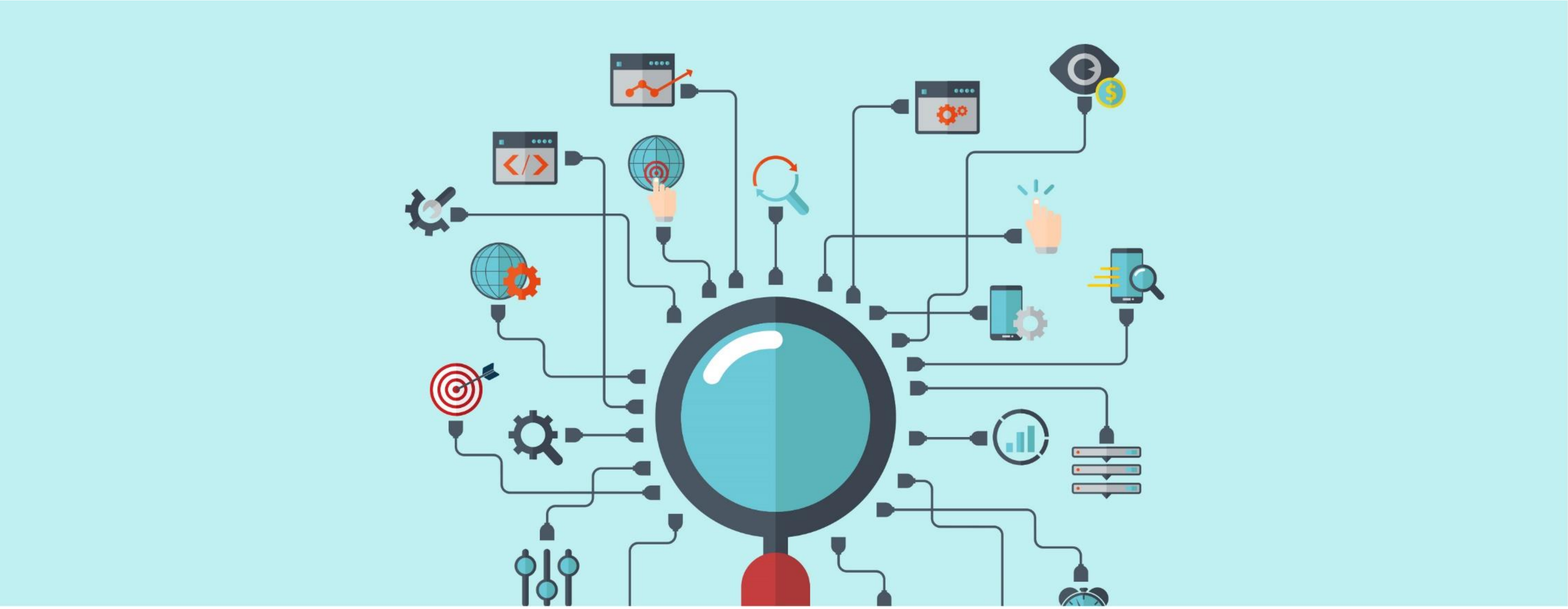
Result Visualization

- Tableau
- PowerBi
- Excel

https://powerbi.microsoft.com/en-us/

https://public.tableau.com/en-us/s/

Getting Insight



http://www.clearmotive.ca/our-insights/february-2016/data-insights-and-roi/

Value

https://graficasoftware.com/how-is-big-data-creating-value/

Why we need preprocess the data?

- **Real world data is dirty.**

1. **Incomplete Data: Lacking attribute values, Lacking certain attributes of interest, or containing only aggregate data.**

2. **Inconsistent data: Containing discrepancies in codes or names**

   e.g.

   Age ='42' Birthday = "03/07/1997"

   Previous rating '1,2,3', Present rating 'A,B,C'

   In the example there are Discrepancy between duplicate records.

3. **Noisy Data: Containing errors or outliers.**

   e.g.

   Salary ="-10"

   Family = "unknown"

https://www.cmswire.com/information-management/why-organizations-need-to-clean-their-dirty-data/

Why Data Preprocessing is important?

- **No quality data, no quality mining results!**

1. Quality decisions must be based on quality data

e.g.

duplicate or missing data may cause incorrect or even misleading statistics.

2. Data warehouse needs consistent integration of quality data

- **Data extraction, cleaning, and transformation comprises most of the work of building a data warehouse**

1. A very laborious task
2. Legacy data specialist needed.
3. Tools and data quality tests to support these tasks

https://www.electronicsmedia.info/2017/12/20/what-is-data-preprocessing/

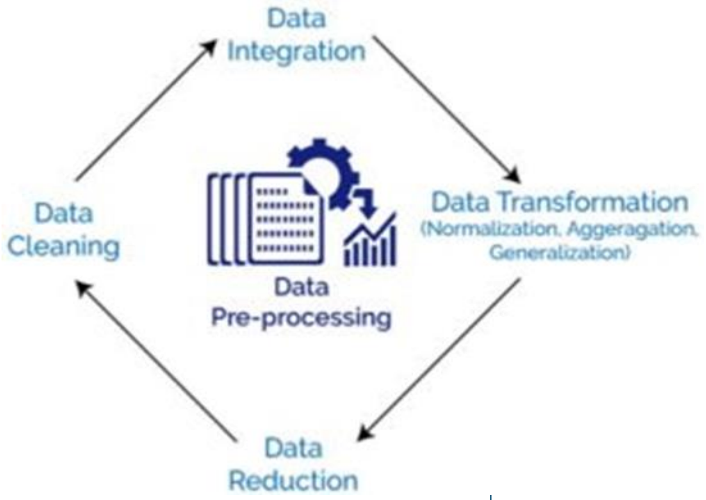
Major tasks of Data Preprocessing

1. Data Cleaning:

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

2. Data integration:

- Integration of multiple databases, data cubes, files, or notes.



3. Data transformation

- Normalization (scaling to a specific range)
- Aggregation

4. Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results.
- Data discretization: with importance, especially for numerical data.
- Data aggregation, dimensionality reduction, data compression, generalization.

https://www.electronicsmedia.info/2017/12/20/what-is-data-preprocessing/

1. Data Cleaning

1. Why Data Cleaning?

- “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
- “Data cleaning is the number one problem in data warehousing”—DCI survey

2. Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

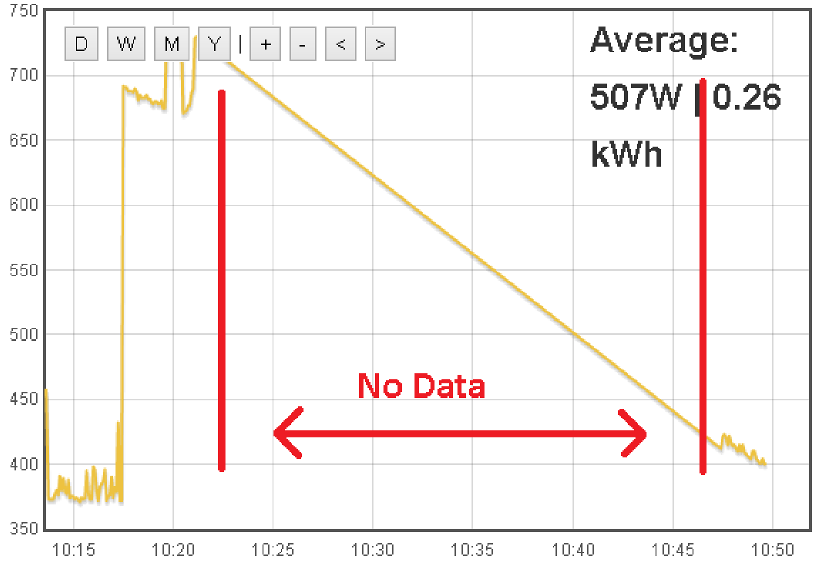
1.1. Missing Data

1. Data is not always available

- Many tuples have no recorded value for several attributes, such as customer income in sales data.

2. Missing data may be due to

- Equipment malfunction
- Inconsistent with other recorded data and thus deleted
- Data not entered due to misunderstanding (left blank)
- Certain data may not be considered important at the time of entry (left blank)
- Not registered history or changes of the data
- Missing data may need to be inferred (blanks can prohibit application of statistical or other functions)



https://github.com/emoncms/emoncms/issues/103

1.2. Noisy Data

1. Noise:

- Random error or variance in a measured variable

2. Incorrect attribute values may be due to

- Faulty data collection instruments
- Data entry problems
- Data transmission problems
- Technology limitation
- Inconsistency in naming convention (H. Shree, HShree, H.Shree, H Shree etc.)
-

3. Other data problems which requires data cleaning

- Duplicate records (omit duplicates)
- Incomplete data (interpolate, estimate, etc.)
- Inconsistent data (decide which one is correct …)

https://en.wikipedia.org/wiki/Noisy_data

1.2.1. How to handle Noisy Data?

1. Binning

- First sort data and partition into (equi-depth) bins
- Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
-

2. Regression

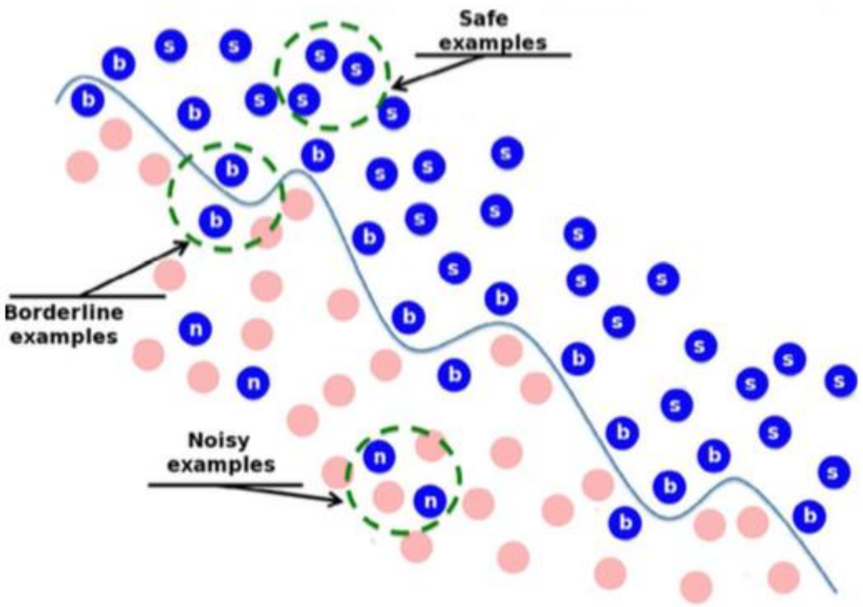
- Smooth by fitting the data into regression functions

3. Clustering

- Detect and remove outliers

4. Combined computer and human inspection

- Detect suspicious values and check by human (e.g., deal with possible outliers)



https://sci2s.ugr.es/noisydata

1.2. Data Cleaning as a process

Data discrepancy detection

- Use metadata (e.g., domain, range, dependency, distribution)
- Check field overloading
- Check uniqueness rule, consecutive rule and null rule
- Use commercial tools (Talend Data Quality Tool, Sept. 2008)

Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections

Data auditing: by analysing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Data migration and integration

- Data migration tools: allow transformations to be specified
- ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Integration of the two processes

- Iterative and interactive (e.g., Potter's Wheels)

2. Data Integration and Transformation

1. Data integration

- Combines data from multiple sources into a coherent store

2. Schema integration:

e.g., A.cust-id B.cust-#

- Integrate metadata from different sources

3. Entity identification problem

- Identify and use real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

4. Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- Possible reasons: different representations, different scales, e.g., metric vs. British units

2.1. Handling Redundancy in Data Integration

Redundant data occur often when integration of multiple databases

- **Object identification:** The same attribute or object may have different names in different databases
- **Derivable data:** One attribute may be a "derived" attribute in another table, e.g., annual revenue

Redundant attributes may be able to be detected by correlation analysis

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

2.2 Data Transformation

- **Smoothing:** remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range

  min-max normalization

  z-score normalization

  normalization by decimal scaling

- **Attribute/feature construction**

  New attributes constructed from the given ones

3. Data Reduction

Why Data Reduction?

- A database/data warehouse may store terabytes of data
- Complex data analysis/mining may take a very long time to run on the complete data set
-

Data reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

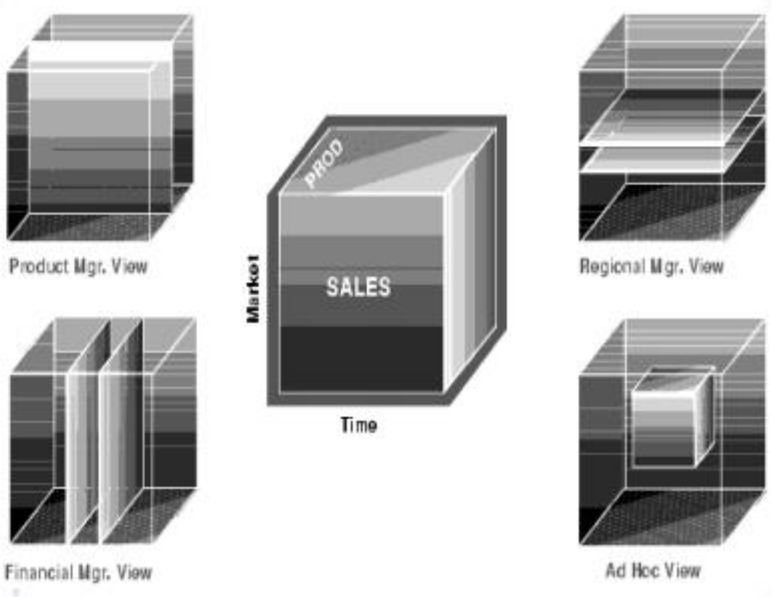
Data reduction strategies

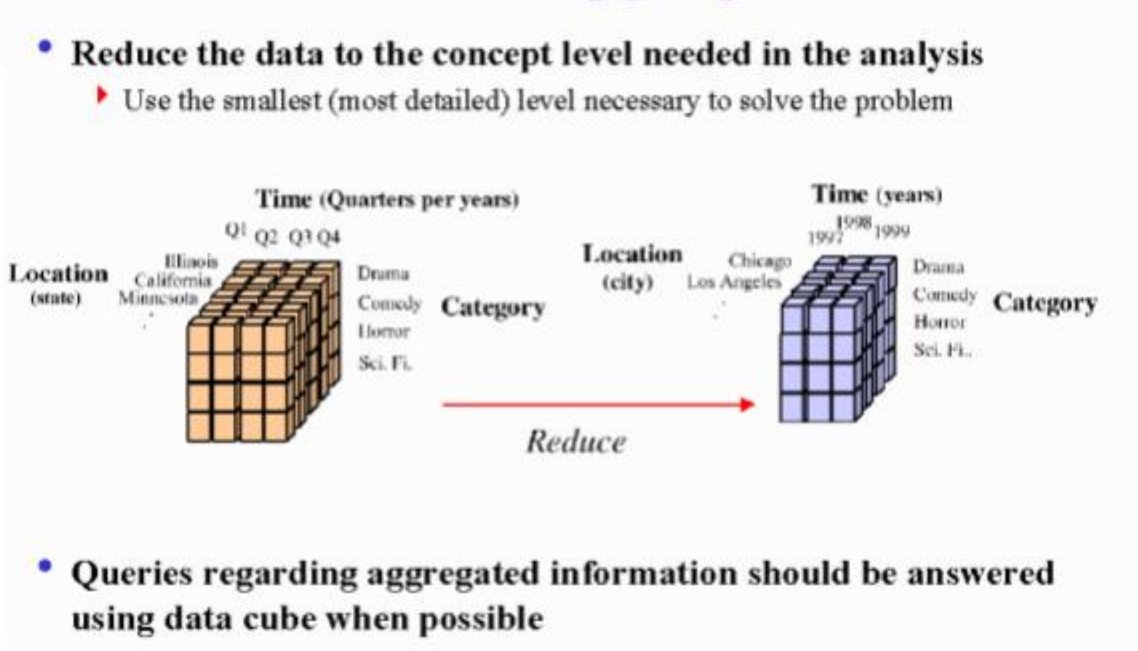
- Data cube aggregation:
- Dimensionality reduction — e.g., remove unimportant attributes
- Data Compression
- Numerosity reduction — e.g., fit data into models
- Discretization and concept hierarchy generation

3.1. Data Cube Aggregation

- **Multiple levels of aggregation in data cubes**
  - Further reduce the size of data to deal with
- **Reference appropriate levels Use the smallest representation capable to solve the task**

Data Cube Aggregation





- **Reduce the data to the concept level needed in the analysis**
  - Use the smallest (most detailed) level necessary to solve the problem

- **Queries regarding aggregated information should be answered using data cube when possible**

https://www.slideshare.net/jasonrodrigues/data-preprocessing-5609305?from_action=save

3.2. Data Compression

String compression

- There are extensive theories and well-tuned algorithms.
- Typically lossless – But only limited manipulation is possible without expansion.

Audio/video, image compression

- Typically lossy compression, with progressive refinement.
- Sometimes small fragments of signal can be reconstructed without reconstructing the whole.

Time sequence is not audio

- Typically short and vary slowly with time.

Thank You