

The slide features a white background decorated with numerous 3D-rendered bubbles of varying sizes. These bubbles are metallic and reflective, with highlights and shadows that give them a realistic, spherical appearance. They are scattered across the slide, with some appearing near the top left, others near the bottom right, and a few in the center. The bubbles vary in size, with some being quite large and others being small specks.

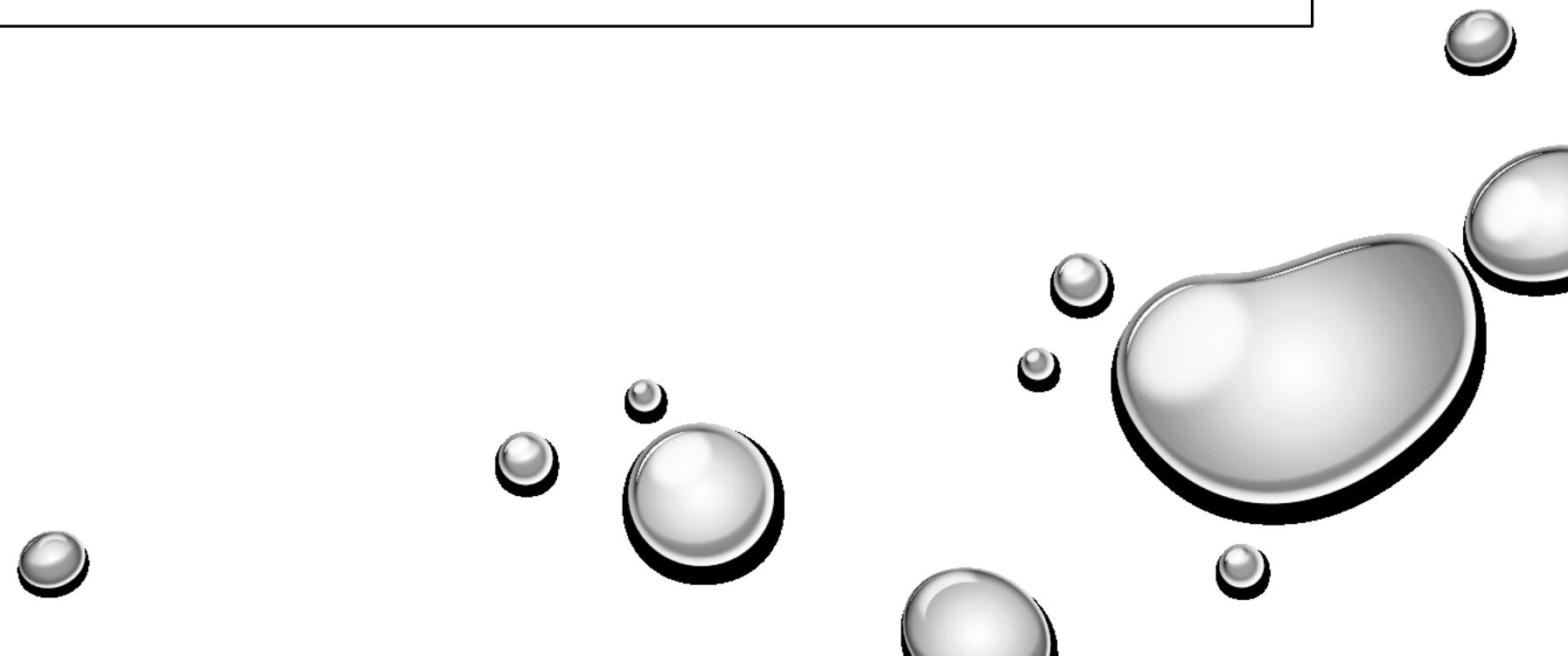
DATA ANALYTICS

- Gartner projects that by 2015, 85% of fortune 500 organizations will be unable to exploit big data for competitive advantage. About 4.4 million jobs will be created around big data (baesens et al, 2003).
- A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and “data scientist” talent required to exploit big data.



DATA ANALYTICS

Analytics is a term that is often used interchangeably with data science, data mining, knowledge discovery, and others.



DATA ANALYTICS

Table 1.1 Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

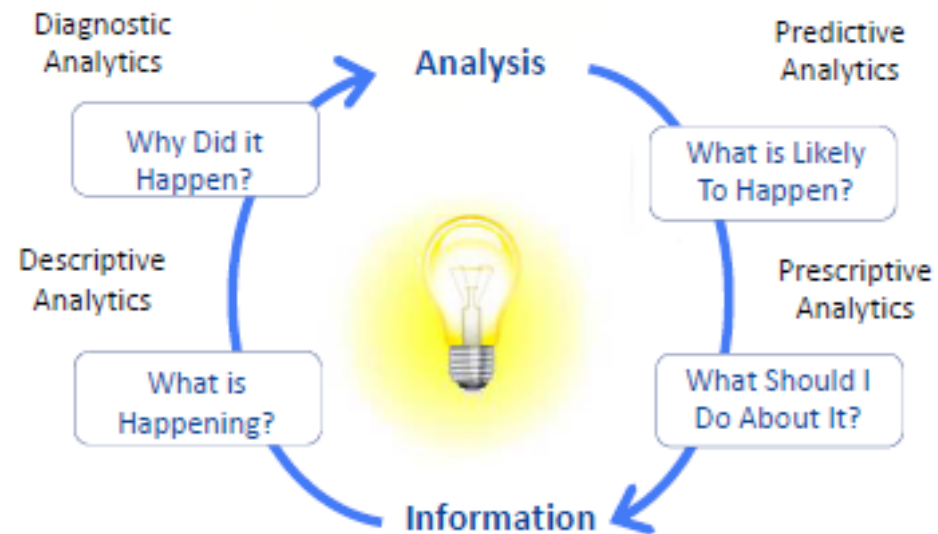
(Baesens, 2014)

DATA ANALYTICS

Four Types of Analytics

Information
Builders

Information, Analysis And Decisions: The Basics




Analytic Excellence Leads to Better Decisions

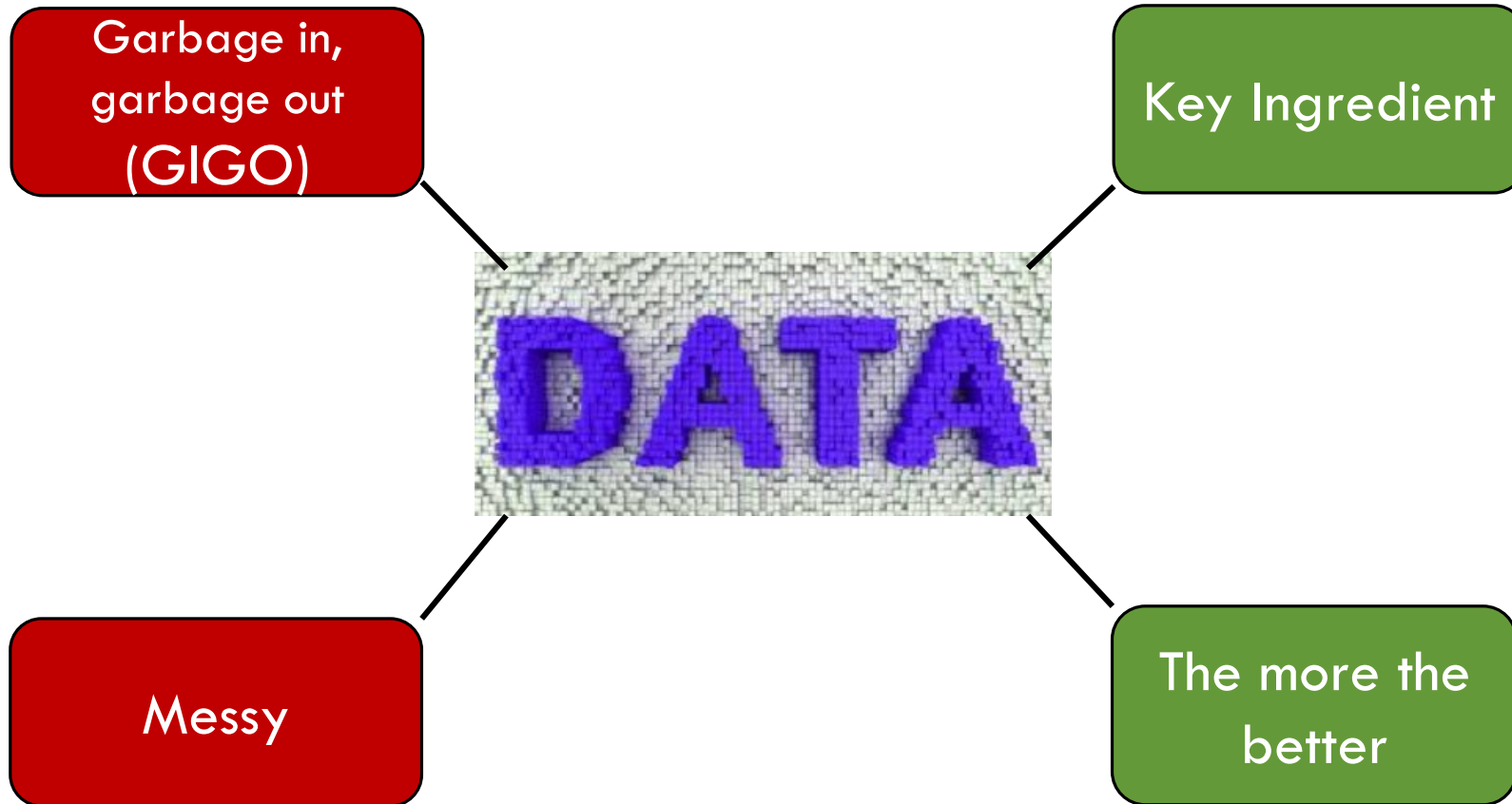
Gartner



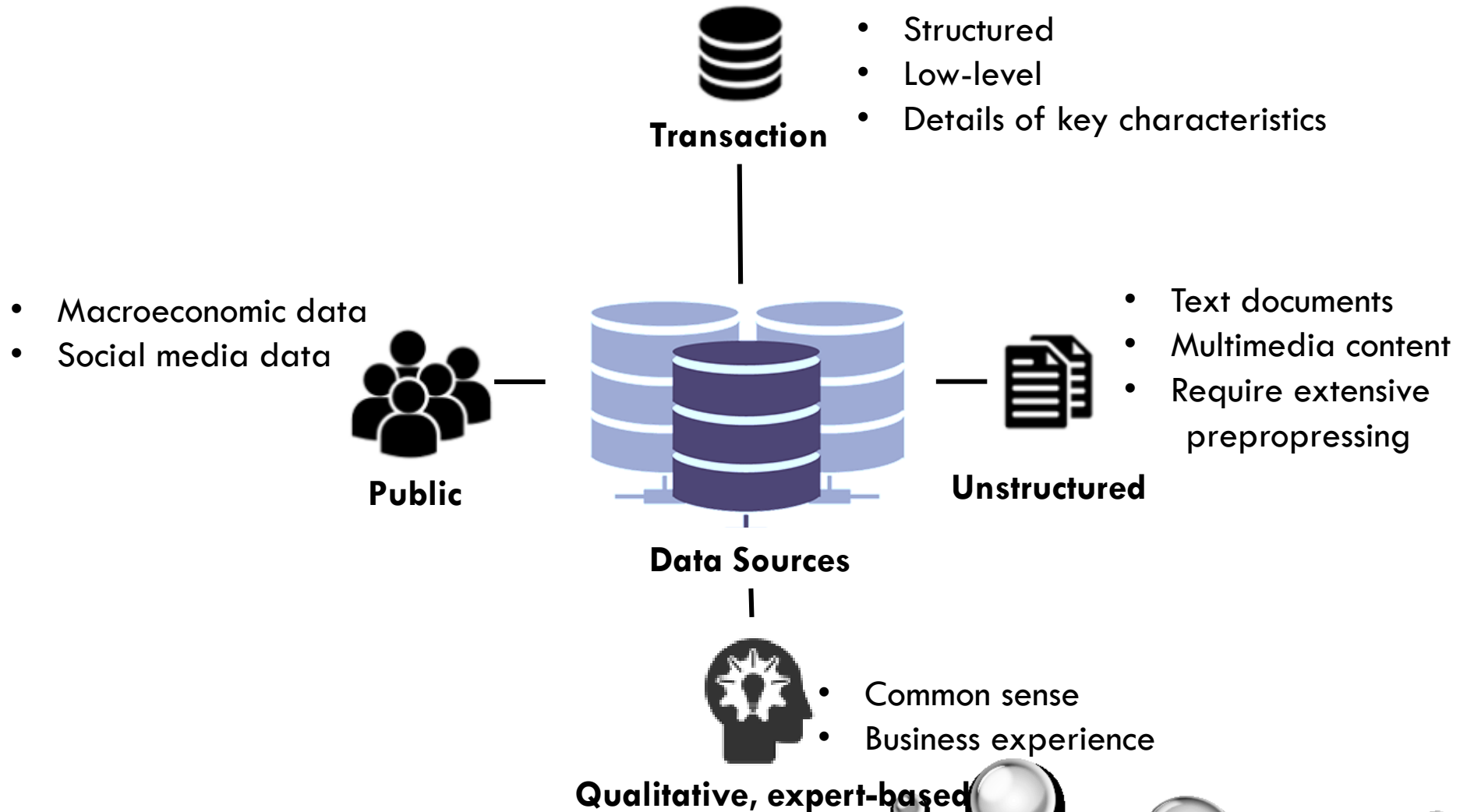
DATA ANALYTICS

- TYPES OF DATA SOURCES AND DATA ELEMENTS
 - DATA COLLECTION
 - POPULATIONS AND SAMPLES OF BIG DATA
 - DATA MUNGING/WRANGLING
 - DATA PRE-PROCESSING
 - VISUAL DATA AND EXPLORATORY STATISTICAL ANALYSIS
 - DATA STORAGE AND MANAGEMENT OF BIG DATA
- 

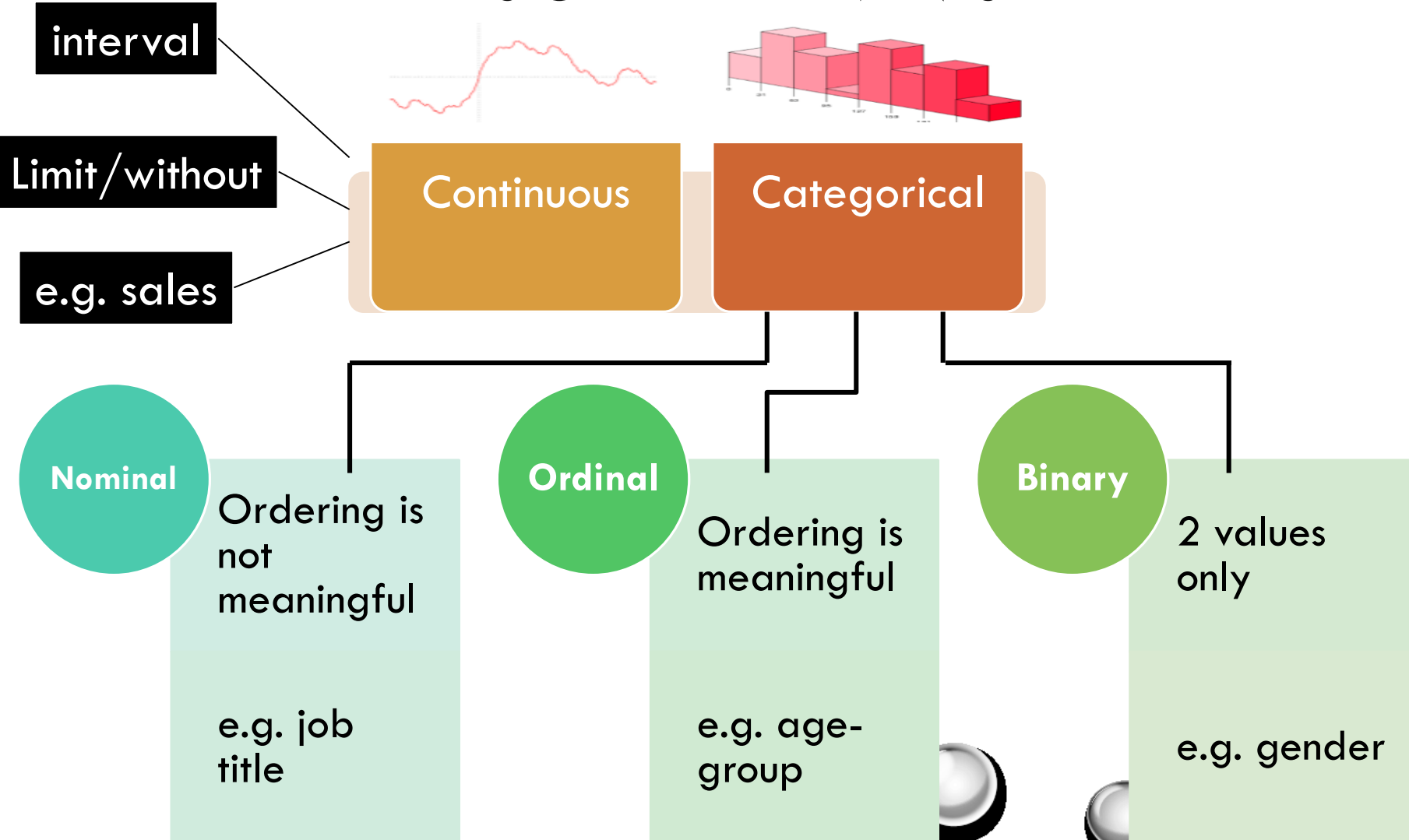
TYPES OF DATA SOURCES AND DATA ELEMENTS



TYPES OF DATA SOURCES



TYPES OF DATA ELEMENTS





DATA COLLECTION

The activity of collecting information that can be used to find out about a particular subject.

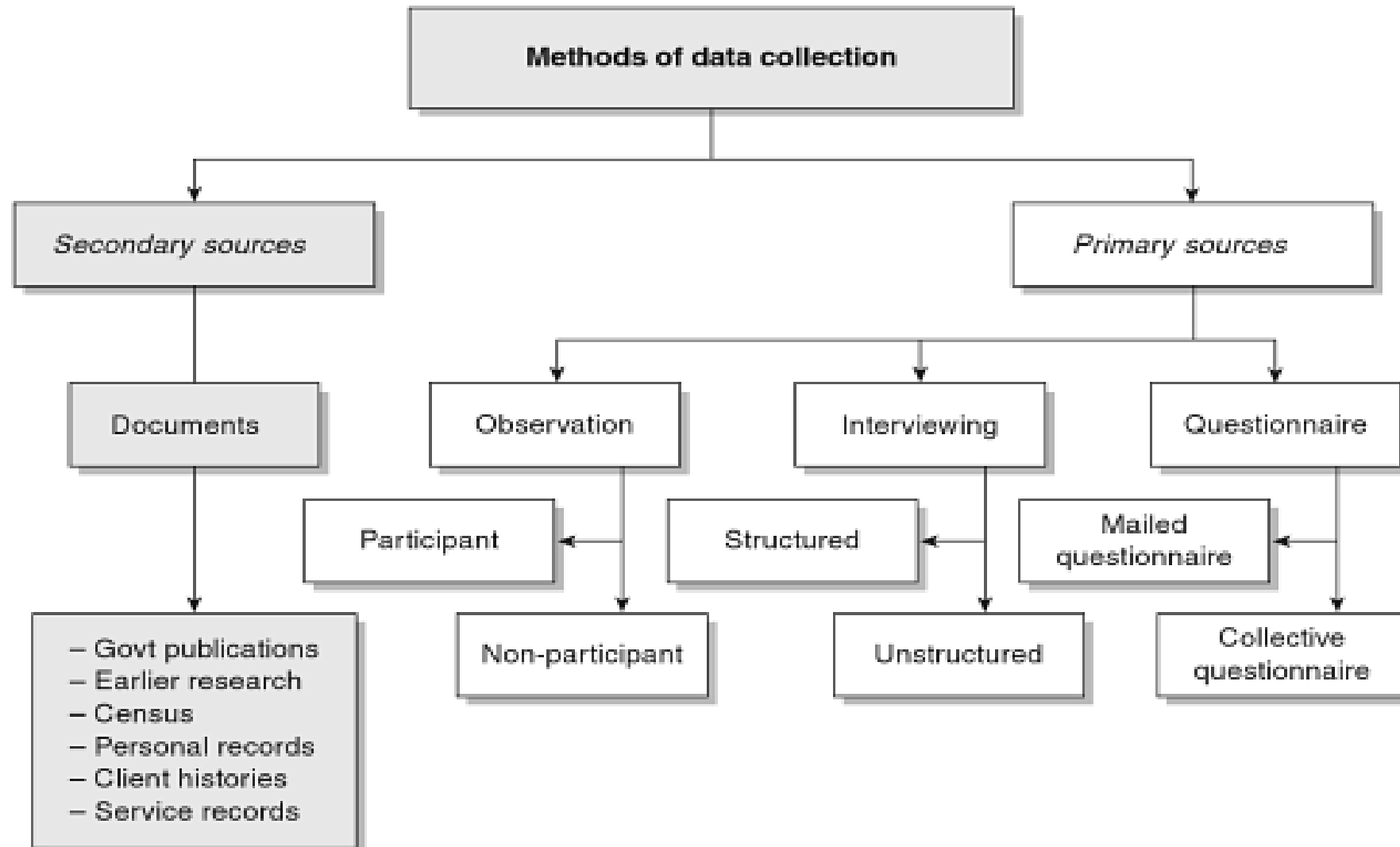
(Cambridge Dictionary)

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.

(Wikipedia)




METHODS OF DATA COLLECTION





SAMPLING OF BIG DATA

BIG DATA IS A TERM THAT DESCRIBES THE **LARGE VOLUME OF DATA** – BOTH **STRUCTURED AND UNSTRUCTURED** – THAT INUNDATES A BUSINESS ON A DAY-TO-DAY BASIS. BUT **IT'S NOT THE AMOUNT OF DATA THAT'S IMPORTANT. IT'S WHAT ORGANIZATIONS DO WITH THE DATA THAT MATTERS.** BIG DATA CAN BE ANALYZED FOR **INSIGHTS** THAT LEAD TO BETTER DECISIONS AND STRATEGIC BUSINESS MOVES.





WHY SAMPLING OF BIG DATA

Storing the *full* data may not be feasible

- Your application may not keep *everything*

Work with data in full is inconvenient

- What is the need to analyze the *full* data?

Work with a compact summary is faster

- Would you rather exploring data with a PC than a supercomputer/cluster?
- 



NEXT

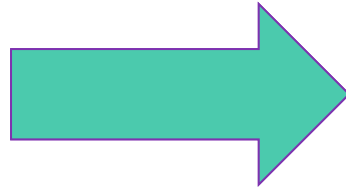
DATA PROCESSING



DATA MUNGING/WRANGLING



RAW DATA =>
Messy / noisy data



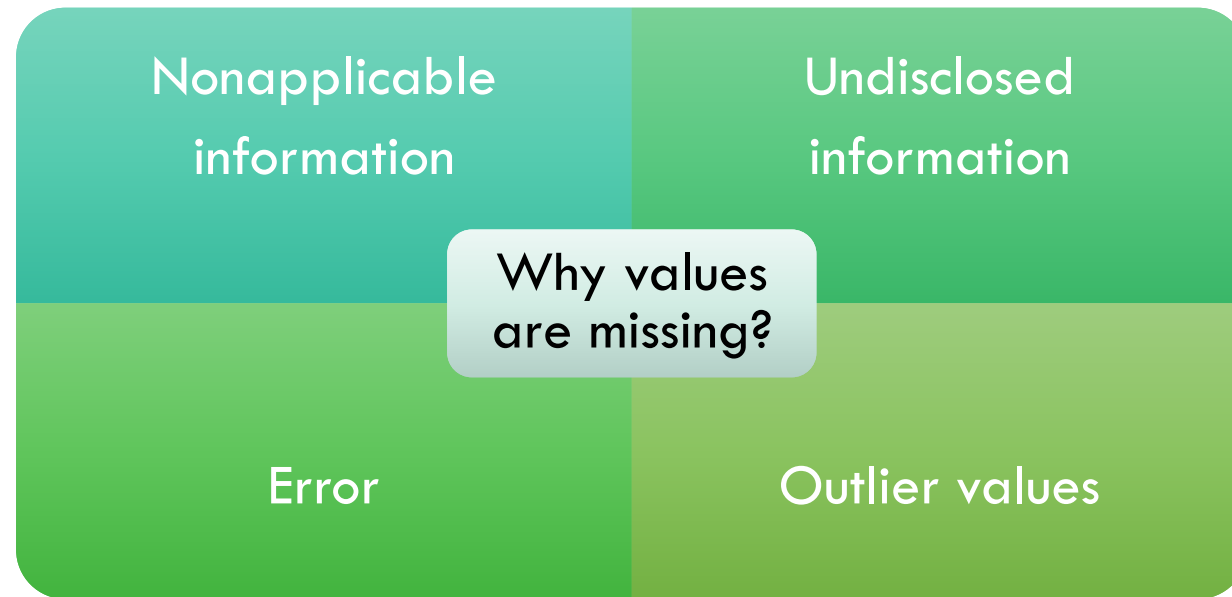
CLEANED DATA =>
Data that can be analyzed

Process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data.

DEALING WITH MISSING VALUES

No information of those students who withdraw.

Private data , such as salary, may not be disclosed.



Human factor – question was skipped by respondent, typo
Technical issue

The values have to be treated missing. E.g. extremely low or extremely high values

DEALING WITH MISSING VALUES

Replace (Impute)

- Replacing the missing values with a known value (e.g. mean, median, mode)

Delete

- Deleting observations or variables with lots of missing values as the data may not be meaningful

Keep

- If the data with missing values are meaningful. Needs to be considered as a separate category.

CHOOSING THE RIGHT WAY TO DEAL WITH MISSING VALUES

Statistical test

- Test whether the missing information is related to the target variable
- If yes, then choose **keep**

Observe the number of available observations

- If available observations are high, then consider **delete**
- Else, consider **impute**

DEALING WITH MISSING VALUES (EXAMPLE)

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800		620	Churner
2	28	1,200	Single		Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married		Nonchurner
6	44				Nonchurner
7	22	1,200	Single		Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34		Single		Churner
10	50	2,100	Divorced		Nonchurner

Suggest a way to deal with the missing values of Record 1, 6 and 10.

DEALING WITH OUTLIERS

Valid observations



☐ Salary of CEO is \$1 million

☐ ?

☐ ?

Invalid Observations



☐ Age = 300 years

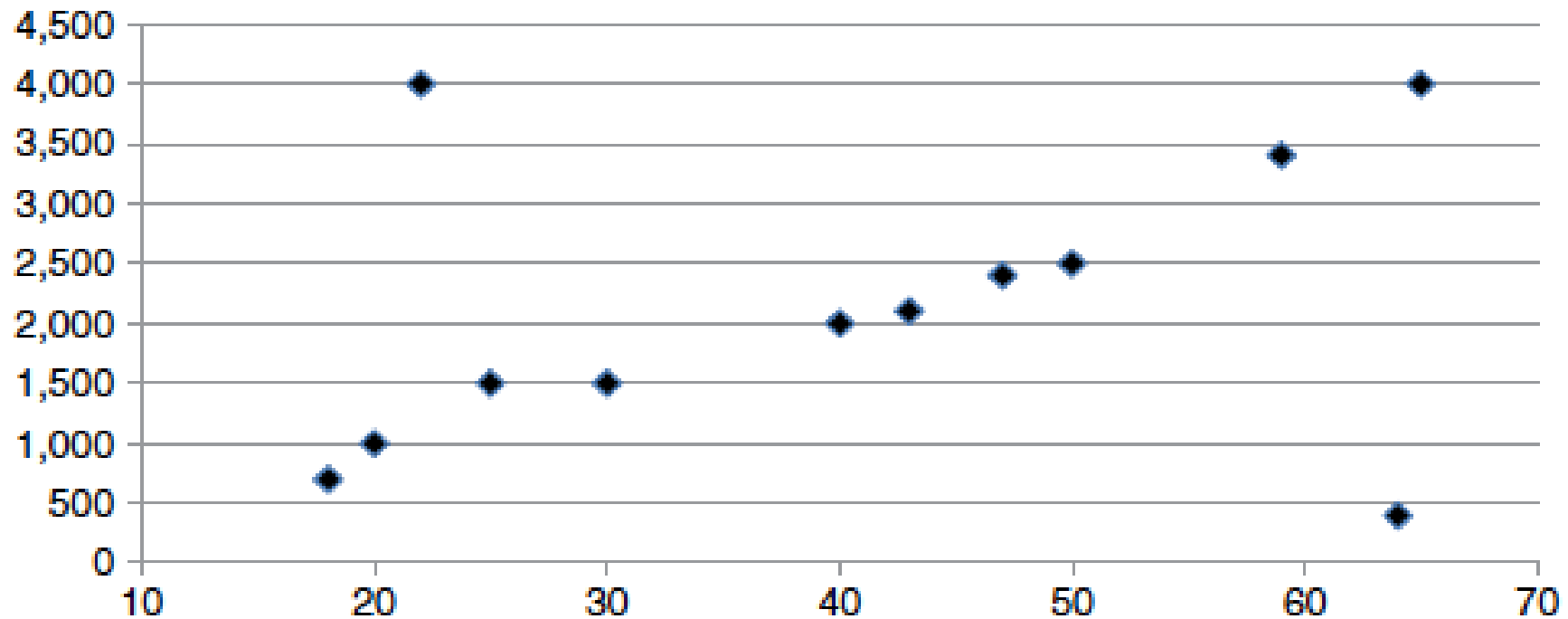
☐ ?

☐ ?

Provide some examples of each type of outliers.

MULTIVARIATE OUTLIERS

Income and Age



Multivariate outliers are observations that are outlying in multiple dimensions (e.g. age and income)

STEPS TO DEAL WITH OUTLIERS

Detection

- Calculate the min and max
- Use visual tools, e.g. histograms, boxplots
- Z-scores

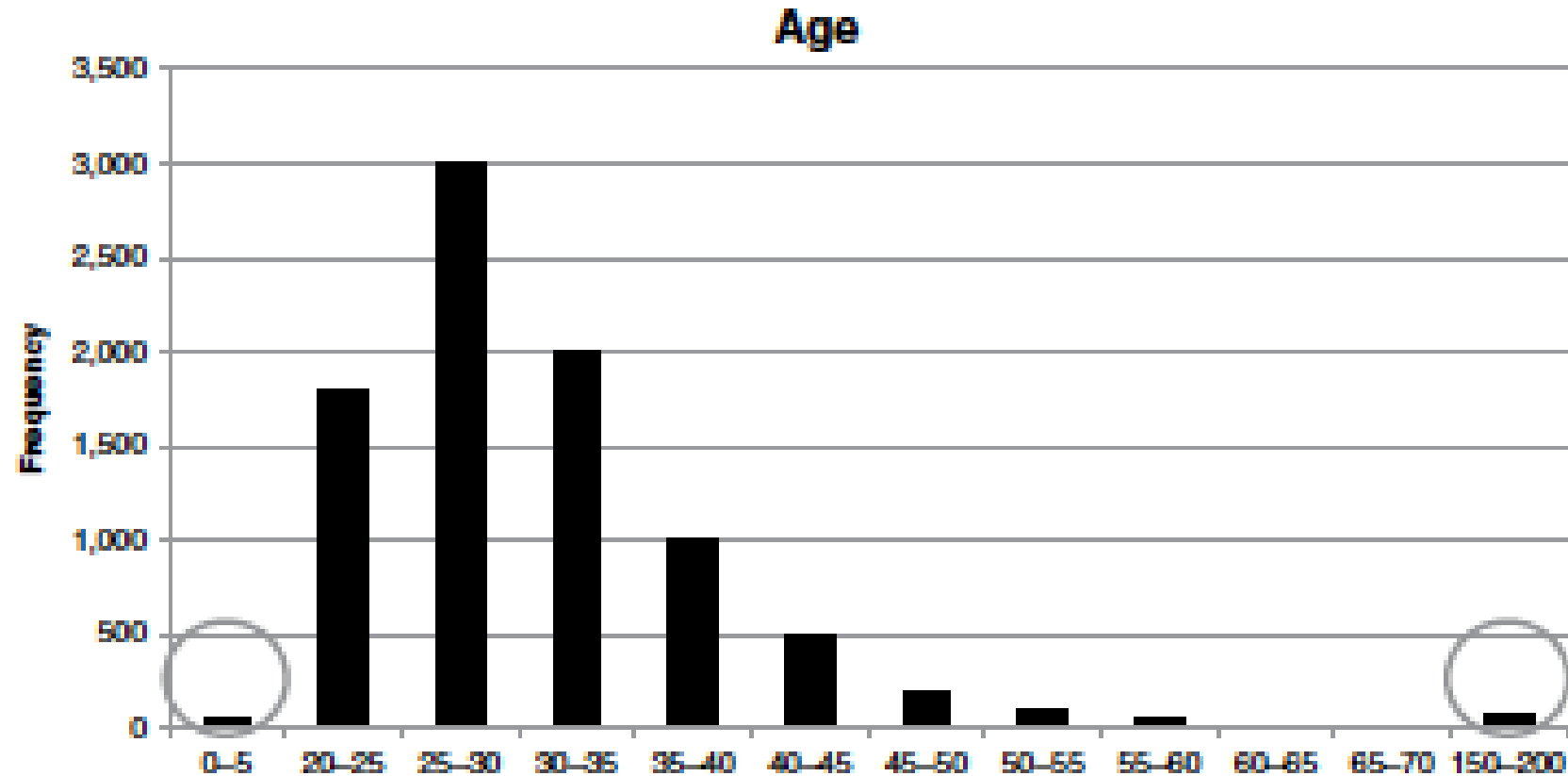
For univariate outliers

- regression

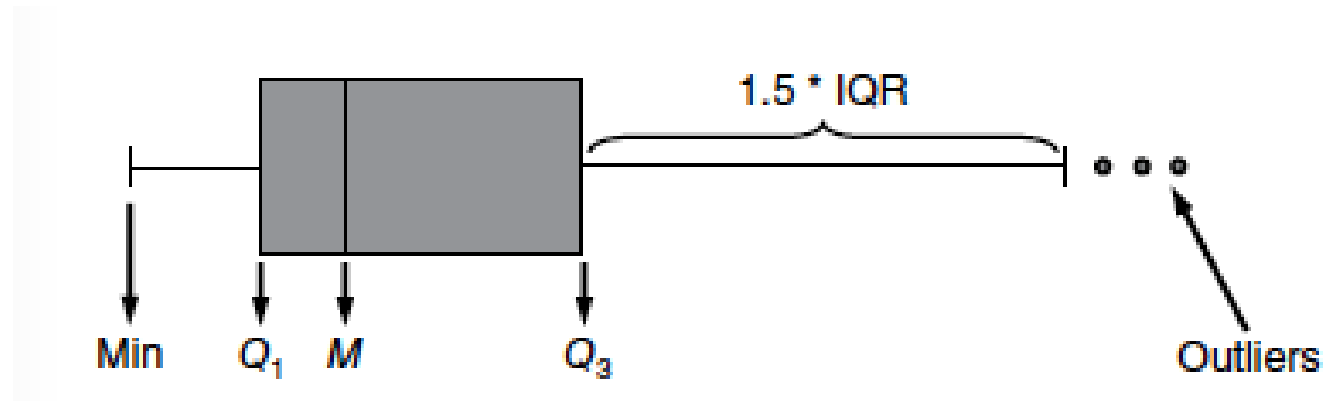
For multivariate outliers

Treatment

USING A HISTOGRAM FOR OUTLIERS DETECTION



USING A BOXPLOT FOR OUTLIERS DETECTION



A box plot represents three key quartiles of the data: the first quartile (25% of the observations have a lower value), the median (50% of the observations have a lower value), and the third quartile (75% of the observations have a lower value).

The minimum and maximum values are then also added unless they are too far away from the edges of the box.

Too far away is then quantified as more than $1.5 * \text{Interquartile Range}$ ($IQR = Q_3 - Q_1$).

USING Z-SCORES FOR OUTLIERS DETECTION

- Z-SCORES MEASURES HOW MANY STANDARD DEVIATIONS, σ , AN OBSERVATION LIES AWAY FROM THE MEAN, μ .

$$z_i = \frac{x_i - \mu}{\sigma}$$

A practical rule of thumb then defines outliers when the absolute value of the z -score $|z|$ is bigger than 3. Note that the z -score relies on the normal distribution.



DESCRIPTIVE ANALYTICS

- STATISTICAL INFERENCE
- ASSOCIATION RULES
- SEQUENCE RULES
- SEGMENTATION





DESCRIPTIVE AND INFERENTIAL ANALYSIS

DESCRIPTIVE ANALYSIS

- DESCRIPTIVE STATISTICAL ANALYSIS LIMITS GENERALIZATION TO THE PARTICULAR GROUP OF INDIVIDUALS OBSERVED. THAT IS:
- NO CONCLUSION ARE EXTENDED BEYOND THIS GROUP
- ANY SIMILARITY TO THOSE OUTSIDE THE GROUP CANNOT BE ASSUMED.
- THE DATA DESCRIBE ONE GROUP AND THAT GROUP ONLY.

INFERENTIAL ANALYSIS

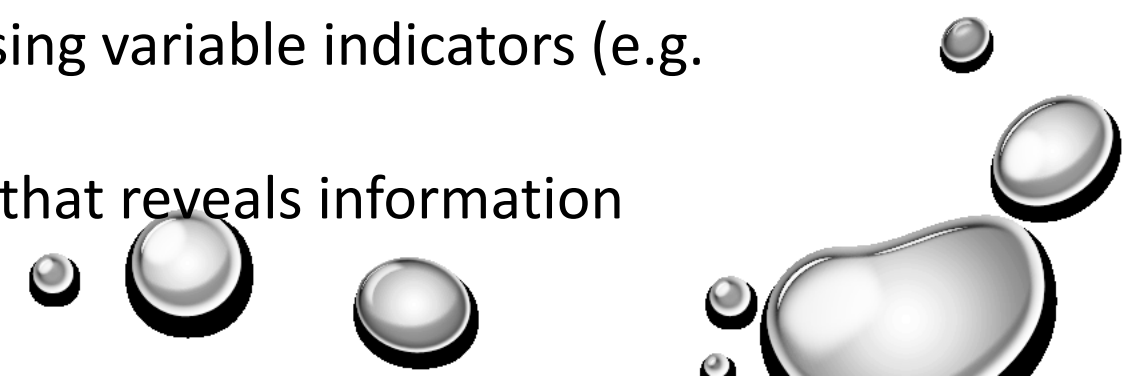
- INFERENTIAL ANALYSIS SELECTS A SMALL GROUP (SAMPLE) OUT OF A LARGER GROUP (POPULATION) AND THE FINDING ARE APPLIED TO THE LARGER GROUP. IT IS USED TO ESTIMATE A PARAMETER, THE CORRESPONDING VALUE IN THE POPULATION FROM WHICH THE SAMPLE IS SELECTED.
- IT IS NECESSARY TO CAREFULLY SELECT THE SAMPLE OR THE INFERENCES MAY NOT APPLY TO THE POPULATION.

DATA LEAKAGE

- When the data you're using to train contains information about what you're trying to predict.
- Introducing information about the target during training that would not legitimately be available during actual use.
- Obvious examples:
 - *Including the label to be predicted as a feature*
 - *Including test data with training data*
- If your model performance is too good to be true, it probably is and likely due to "giveaway" features.



EXAMPLES DATA LEAKAGE

- Leakage in training data:
 - Performing data preprocessing using parameters or results from analyzing the entire dataset:
Normalizing and rescaling, detecting and removing outliers, estimating missing values, feature selection.
 - Time-series datasets: using records from the future when computing features for the current prediction.
 - Errors in data values/gathering or missing variable indicators (e.g. the special value 999) can encode information about missing data that reveals information about the future.
- 

EXAMPLES DATA LEAKAGE

- **Perform data preparation within each cross-validation fold separately**
 - *Scale/normalize data, perform feature selection, etc. within each fold separately, not using the entire dataset.*
 - *For any such parameters estimated on the training data, you must use those same parameters to prepare data on the corresponding held-out test fold.*
-
- **With time series data, use a timestamp cutoff**
 - *The cutoff value is set to the specific time point where prediction is to occur using current and past records.*
 - *Using a cutoff time will make sure you aren't accessing any data records that were gathered after the prediction time, i.e. in the future.*