Summary of Data Mining and Analytics I

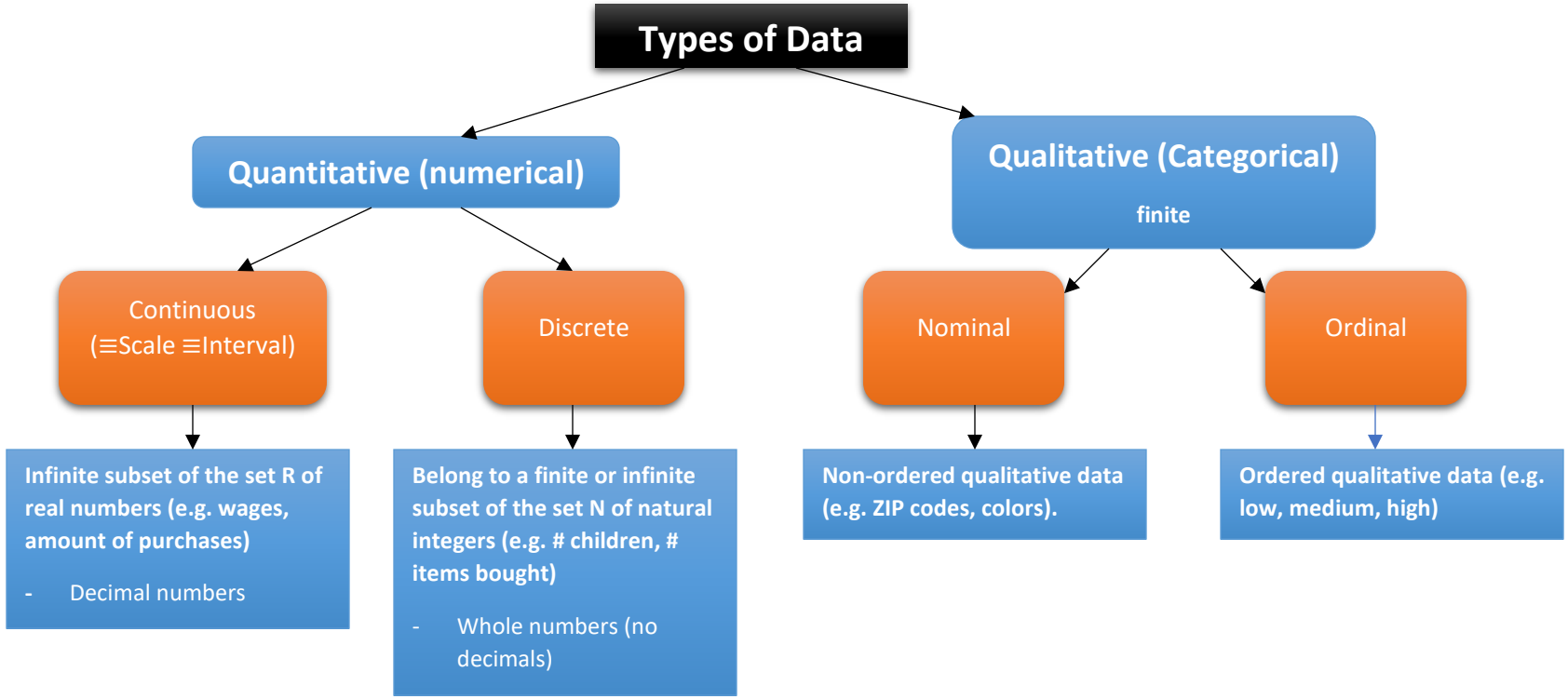
Overview of Data Mining

Objective: Decision assistance

Definition: Combination of artificial intelligence and statistical analysis to discover information that is "hidden in the data"

Types of Data Mining		
Basis for Comparison	**Descriptive/Exploratory Mining**	**Predictive Mining**
Basic	Brings out information that is present but **hidden** in a mass of data. Finds **clusters/groups/ associations** between different products	Extrapolates new information based on present information. New information is qualitative (classification/scoring) or quantitative (regression).
Requirements	Data aggregation and data mining	Statistics and forecasting methods
Preciseness	Provides accurate data	Produce results that does not ensure accuracy.
Approach	Reactive	Proactive
Practical Analysis Methods	Standard reporting, query/drill down and ad-hoc reporting.	Predictive modelling, forecasting, simulation and alerts.
Examples	*Adverse events of a drug were explored by clustering the therapeutic classes; A data analyst receives detailed customer purchasing data and finds associations of any type among customers.*	*An automobile company scored customers for likelihood to return to buy a new model within the next 6 months; A credit card company offered a valued customer product for their card holder based on past card usage to determine the risk pattern; Road traffic was forecasted hourly*

Data Mining Aims


Types of Data
Quantitative (numerical)
Qualitative (Categorical)
finite
Continuous (≡Scale ≡Interval)
Discrete
Nominal
Ordinal
Infinite subset of the set R of real numbers (e.g. wages, amount of purchases)
- Decimal numbers
Belong to a finite or infinite subset of the set N of natural integers (e.g. # children, # items bought)
- Whole numbers (no decimals)
Non-ordered qualitative data (e.g. ZIP codes, colors).
Ordered qualitative data (e.g. low, medium, high)

Types of Values

Rare Value

Can create bias in factor analysis and other analysis, by appearing more important than they are.

What to do:

Remove

Replace with a more frequent value.

Missing Value

Gaps in the data.

What to do:

<10% data not excluded: **Remove** corresponding observations

Mean substitution

Missing Value

Name	Income	Job	name	children
Alice	8000000	trust fund retiree	Ethan	2
Bob	40000	rideshare driver		2
Charlie	1	racecar driver	Refuse to answer	1
Danielle	90000	marketing mgr	Gerald	2

Extreme Value, Aberrant Value, Rare Value

Age	Gender	Hair	Eye	Weight	Salary
14	F	Blue	Blue	143	12500
28	F	Brown	Brown	9	32150
22	M	Blue	Brown	215	34200
46	F	Brown	Orange	190	53200
75	M	Gray	Green	187	28040

Aberrant Values

Aberrant Value

Erroneous value corresponding to incorrect measurement, a calculation error, or a false declaration.

***Incorrect dates**: Unknown DOB replaced by 'round numbers', subscription dates before customer's DOB, dates of last updates in the year 2050, 29/Feb in non-leap year.*

Customers declared as 'private' when they are 'business'.

Amount input as cents when it should be in Euros.

What to do:

Delete if not too numerous and if their distribution is suitably random.

Replace with statistically imputed value.

Extreme Value

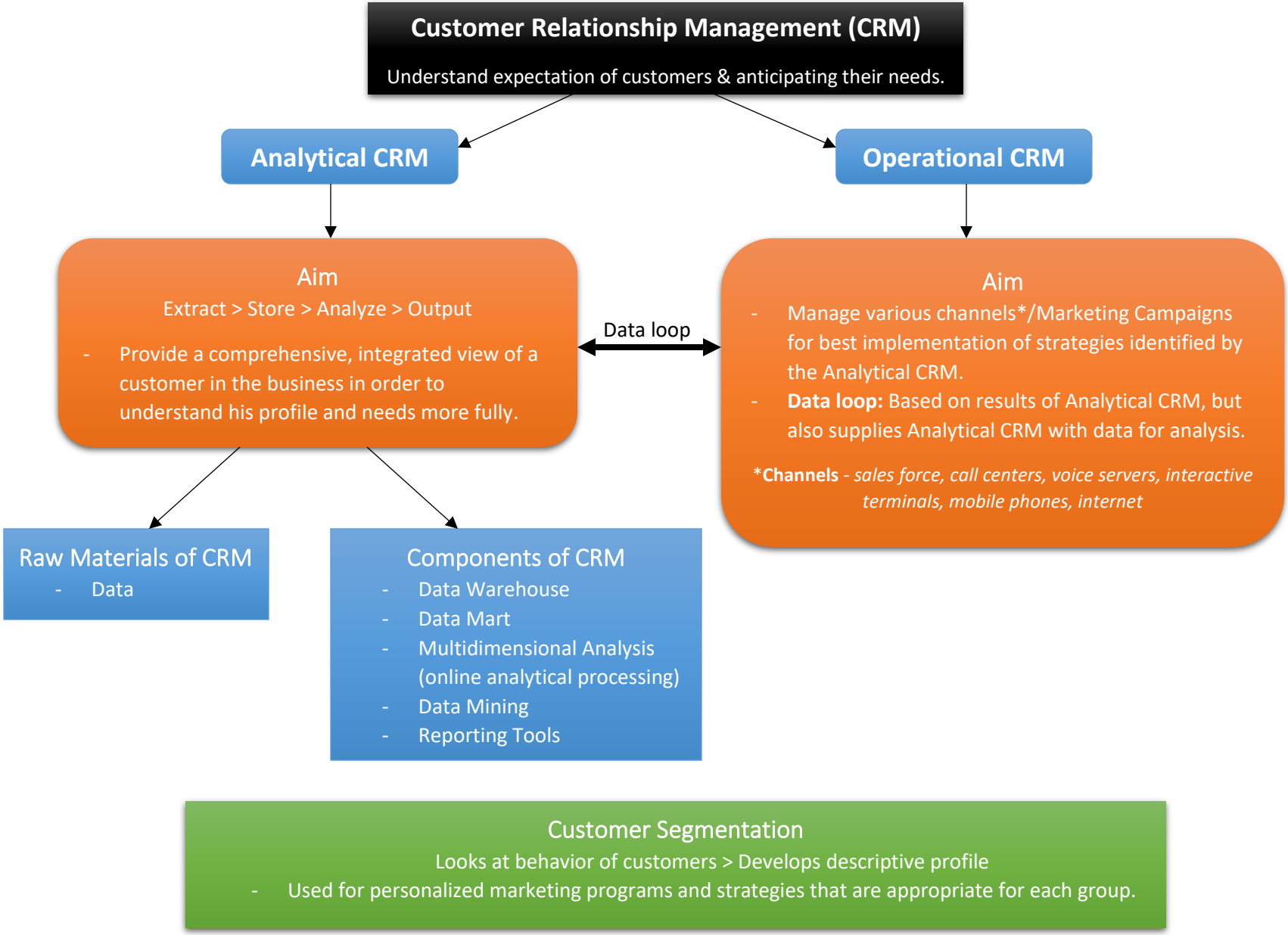
Observations in a sample so far separated in value from the remainder as to suggest they may be from a different population, or the result of an error in measurement.

What to do:

1-2% data not excluded: Exclude outliers

Neutralize: Divide continuous values into classes

Winsorizing: Replace values of the variable beyond 99th percentile with this percentile.


Customer Relationship Management (CRM)
Understand expectation of customers & anticipating their needs.
Analytical CRM
Operational CRM
Aim
Extract > Store > Analyze > Output
- Provide a comprehensive, integrated view of a customer in the business in order to understand his profile and needs more fully.
Data loop
Aim
- Manage various channels*/Marketing Campaigns for best implementation of strategies identified by the Analytical CRM.
- Data loop: Based on results of Analytical CRM, but also supplies Analytical CRM with data for analysis.
*Channels - sales force, call centers, voice servers, interactive terminals, mobile phones, internet
Raw Materials of CRM
- Data
Components of CRM
- Data Warehouse
- Data Mart
- Multidimensional Analysis (online analytical processing)
- Data Mining
- Reporting Tools
Customer Segmentation
Looks at behavior of customers > Develops descriptive profile
- Used for personalized marketing programs and strategies that are appropriate for each group.

Tests

Tests for Normality

The normality of a variable can be verified by the following tests:

1. Shapiro–Wilk Test (the best)
2. Kolmogorov–Smirnov Test (the most general)
3. Lilliefors Test
4. Anderson–Darling Test

Tests for Normality				
Test	**Statistic**		**P-Value**	
Shapiro-Wilk	W	0.992264	Pr < W	0.9845
Kolmogorov-Smirnov	D	0.051999	Pr > D	> 0.1500
Cramer-von Mises	W-Sq	0.02621	Pr > W-Sq	> 0.2500
Anderson-Darling	A-Sq	0.162071	Pr > A-Sq	> 0.2500

Test for Homoscedasticity and Heteroscedasticity

1. Levene Test (best – low sensitivity to non-normality)
2. Bartlett Test (best if distribution is equal)
3. Fisher Test (least robust if normality is not present)

Test of Homogeneity of Variance			
Levene Statistic	**Df1**	**Df2**	**Significance**
50.448	2	396	.000

Transform data to see a more normal distribution:

1. Log transformation
    a. If data is skewed, log transformation will transform the data to a form where we see a more normal distribution.
2. Square Root Transformation
    a. If Log transformation did not correct the issue, try the square root transformation next.

To do:

Why might an online retailer mine the order history of its customers?

- To source new products

A data analyst is using analytical CRM to extract, store, analyze, and output relevant customer information. What is the first step within the analytical CRM phase that this analyst will be performing?

- Combining a customer's records to develop a holistic view.

Which feature of application development is unique to data mining?

The development phase cannot be completed in the absence of data.