

Data Warehousing

Lecture 7 Data Quality, Profiling, Types and Reduction

CITS3401
CITS5504

Zeyi Wen

Computer Science and
Software Engineering

Faculty of Engineering,
and Mathematical
Sciences

Acknowledgement: The lecture slides are based on online sources.

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction

Types of Data Sets: (1) Record Data

Relational records

Relational tables, highly structured

Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Transaction data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

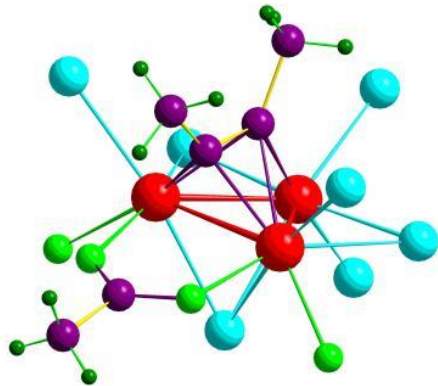
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Document data: Term-frequency vector (matrix) of text documents

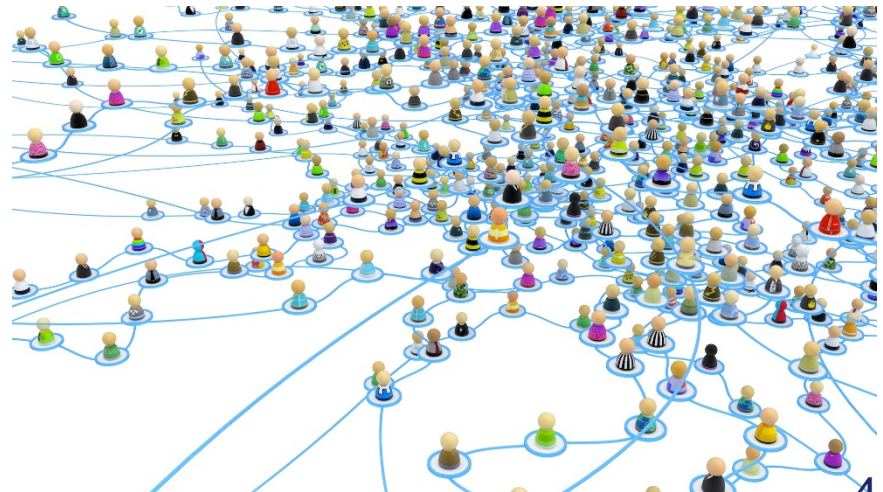
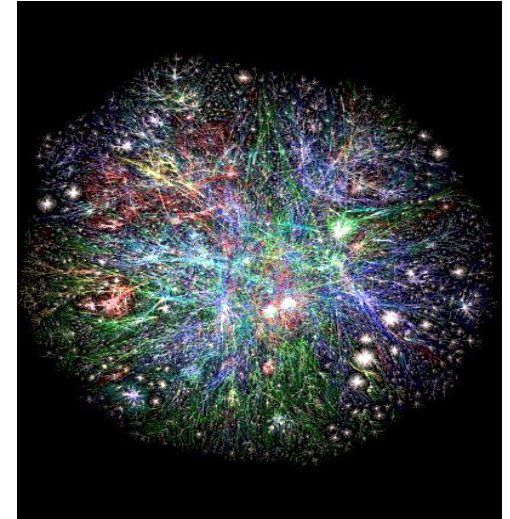
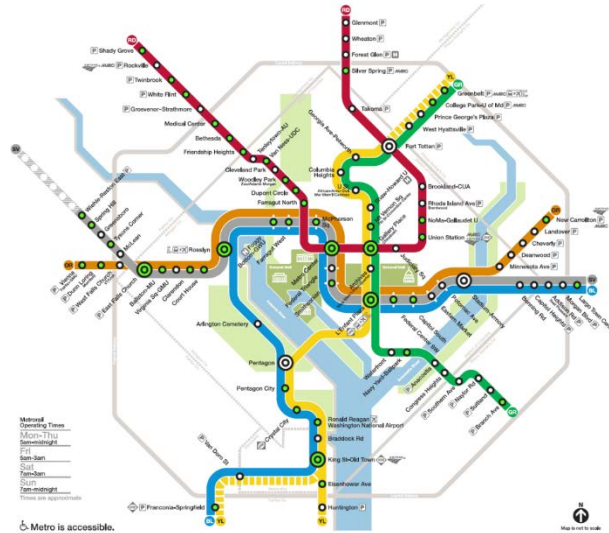
Types of Data Sets: (2) Graphs and Networks

Transportation network

World Wide Web



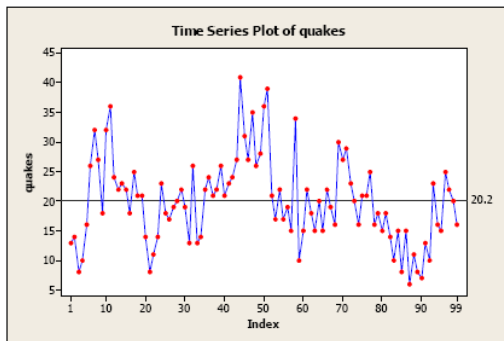
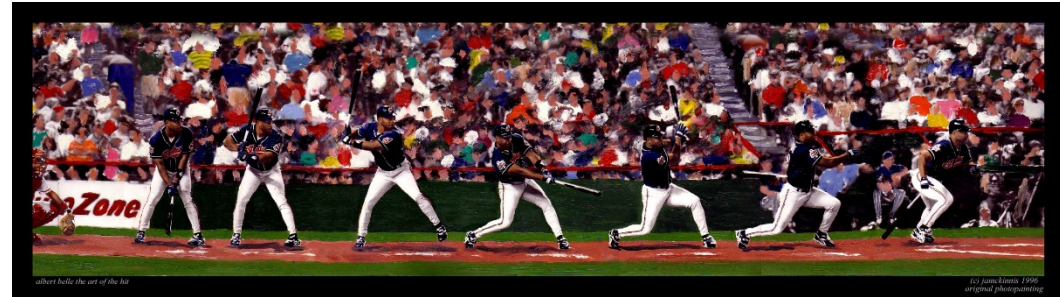
- ❑ Molecular Structures
- ❑ Social or information networks



Types of Data Sets: (3) Ordered Data

Video data: sequence of images

Temporal data: time-series



Sequential Data: transaction sequences

Genetic sequence data

	Start
Human	GTITTTGAGC -- ATGTTCAAC AAATGCTCCTTTTCATTCCCTATTACAGACC TGCCGCA
Chimpanzee	GTITTTGAGC -- ATGTTCAATAAATGCTGCTTTTCACTCCCTATTACAGACC TGCCGCA
Macaque	GTITTTGAGC -- ATGCTCAATAAATGCTCCTTTTCATTCCCTCATTACAAAGT TGCCGCA
Human	GAC AATTCGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAAC TTAGTAATTGAGTGT
Chimpanzee	GAC AATTCGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAAC TTAGTAATTGAGTGT
Macaque	GAC AATTCGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAAC TTAGTAATTGAGTGT
Human	GATCTGGAGACTAA - CTC TGA AATAAATAAGCTGATTATTTATTTATTTCTC AAAACAA
Chimpanzee	GATCTGGAGACTAA AACTCTGA AATAAATAAGCTGATTATTTATTTATTTCTC AAAACAA
Macaque	TATCTGGAGACTAA AACTCTGA AATAAATAAGCTGATTATTTATTTATTTCTC AAAACAA
Human	CAGAATACGATTTAGCAAAATTACTTCTTAAGATAT TATTTTACATTTCTATATTCTCCTA
Chimpanzee	CAGAATACGATTTAGCAAAATTACTTCTTAAGATAT TATTTTACATTTCTATATTCTCCTA
Macaque	CAGAATATGATTTAGCAAAATTAGCTCTTAAGATAT TATTTTGCATTTCTATATTCTCCTA
Human	CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTTCATAAAGCCAGGTATACA ---- TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGCCG TATGTCAC TTTTCATAAAGCCAGGTATACA ---- TTATG
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTTCATAAAGCCAGGTATATACATTAGC
Human	GACAGGTAAGTAAAAACATATTATTTATTCTAGTTTTTGTCCAAAATTTTAAATTTT
Chimpanzee	GACAGGTAAGTAAAAACATATTATTTATTCTAGTTTTTGTCCAAAATTTTAAATTTT
Macaque	GACAGGTAAGTAAAAA - CATATTATTTATTCTAGTTTTTGTCCAAAGAG TTTTAAATTTT
Human	AAC TGTTCGCGTGTGTTGGTAA --- TG TAAAAACAACTCAGTAC A
Chimpanzee	AAC TGTTCGCGTGTGTTGGTAA --- TG TAAAAACAACTCAGTAC A
Macaque	AAC TGTTCGCTCATGTGTTGGTAA --- CBTAAAAACAAATTCAGTAGC

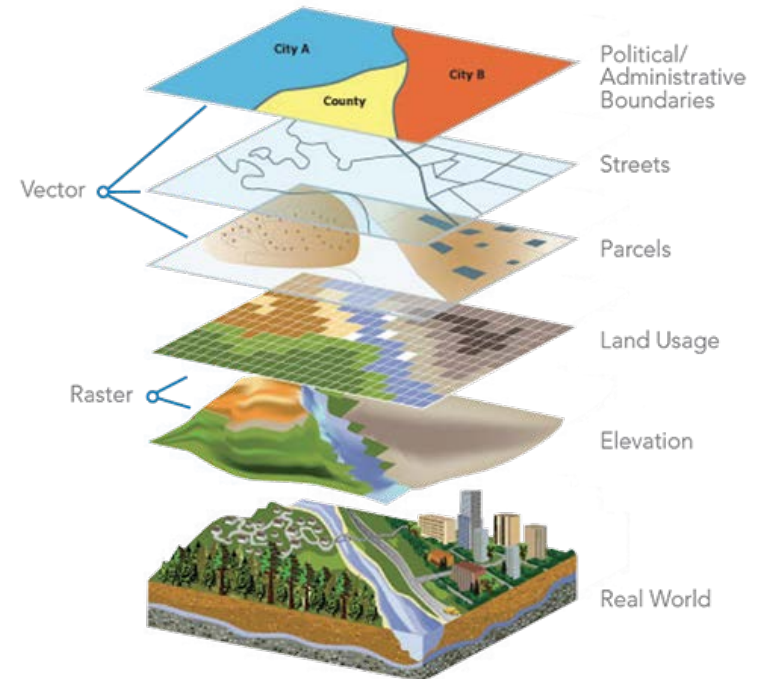
Types of Data Sets: (4) Spatial, image and multimedia Data

Spatial data: maps



Image data:

Video data:

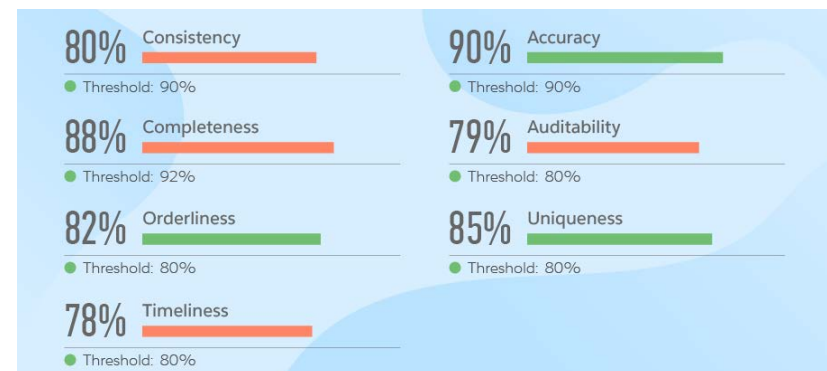


- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction



What is data quality?

- Data quality is a perception or an assessment of **data's fitness** to serve its purpose in a given context.
- Data quality is described by several dimensions like:
 - Correctness/Accuracy
 - Consistency
 - Completeness
 - Timeliness
 - Integrity/Validity
 - Integrity ensures that all data in a database can be traced and connected to other data.



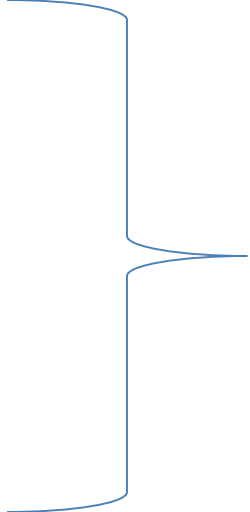
Why data quality matters?

- Good data is your most valuable asset, and bad data can seriously harm your business and credibility...

Poor Data Quality Examples

CUST ID	CONTACT	ADDR1	ADDR2	ADDR3	ADDR4	COUNTRY	PHONE
12345	Aleck Stevenson	123 MAIN street	SYDNEY	NSW	6009	AUS	045241890 1
23456	Stevenson Aleck	123 MAIN ST.	SYDNEY	NSW	6009	AUS	+61 452418901
34567	Aleck Stevenson	90 21 st Place	SYD	NSW	6009	AUS	0452-418- 901
45678	BOB SMITH	48 2 nd Ave street	Los Angeles	CA	80210	USA	
56789	SARAH JOHNSON	78 Green street	SAN DIEGO	ca	80200	USA	6332-21311
67890	BILL GREY	2132 First st				AUS	3322424
78901	MELO KEN	377 lil st				AUS	
89012	MILLES	78 Second st				AUS	859280280
90123	BIRDGES BOB	01 Loisy st	WA	32342		AUD	820823803
ba78901	JACK BLACK	William Hill	WA	32242		AUS	

What are the problems?

- Invalid address
 - Incomplete dataset
 - Duplicate contacts
 - Non-standard format
- 
- Missed opportunities
 - Misguided business decision
 - Low customer satisfaction
 - Negative enterprise image

What data is incorrect?

ADDR1	ADDR2	ADDR3	COUNTRY
William Hill	WA	6009	AUS
01 Loisy st	WA	6005	AUD

- “William Hill” is not a valid address because there is no number.
- “AUD” is not a valid ISO code

What data value give conflicting information?

CUST ID	PRO ID	TRANS ID	TRANS DT	AMT
12345	P33232	A47397	1/1/2021	2000
90123	P33232	A47397	1/1/2021	2000

- This table violates consistency because it has the same transaction records but with different customer ID

What data is stored in non-standard format?

PHONE
0452418901
+61 452418901
0452-418-901

What data is missing or unusable?

ADDR1	ADDR2	ADDR3	ADDR4
2132 First st			
377 lil st			
78 Second st			
01 Loisy st	WA	32242	

What data is missing or not referenced?

CUST_ID	TRANS_ID	PROD_ID	TRANS_D T	AMT
12345	A7292	PRD2113	1/1/2021	1000
	A7483	PRD3121	2/1/2021	12121
45678	A2982	PRD3223	30/12/2020	32324

- There should be a **valid customer and transaction relationship** between them. If there is an transaction relationship data without a customer then that data is not valid and is considered as an **orphan record**.

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ **Cleansing and Matching**
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction



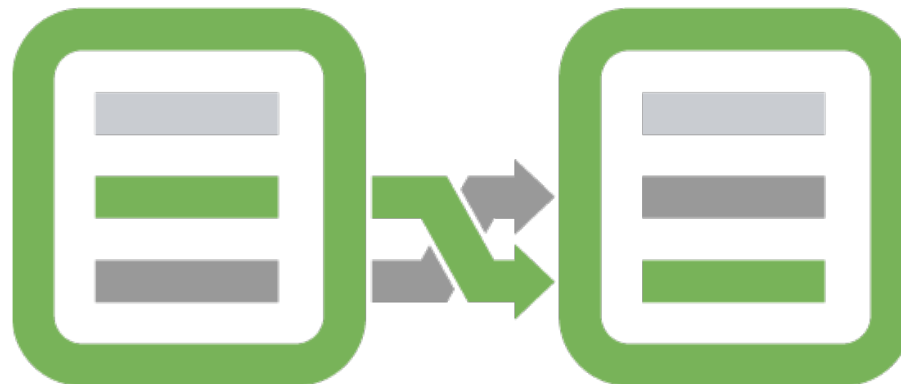
What is data cleansing?

- It is the process of **identifying** and **correcting** dirty data.
- Dirty data means **incomplete**, **wrong**, **duplicate**, or **out-of-date** data.
- The data warehousing community often uses the words **cleanse** or **scrub** rather than **clean**.



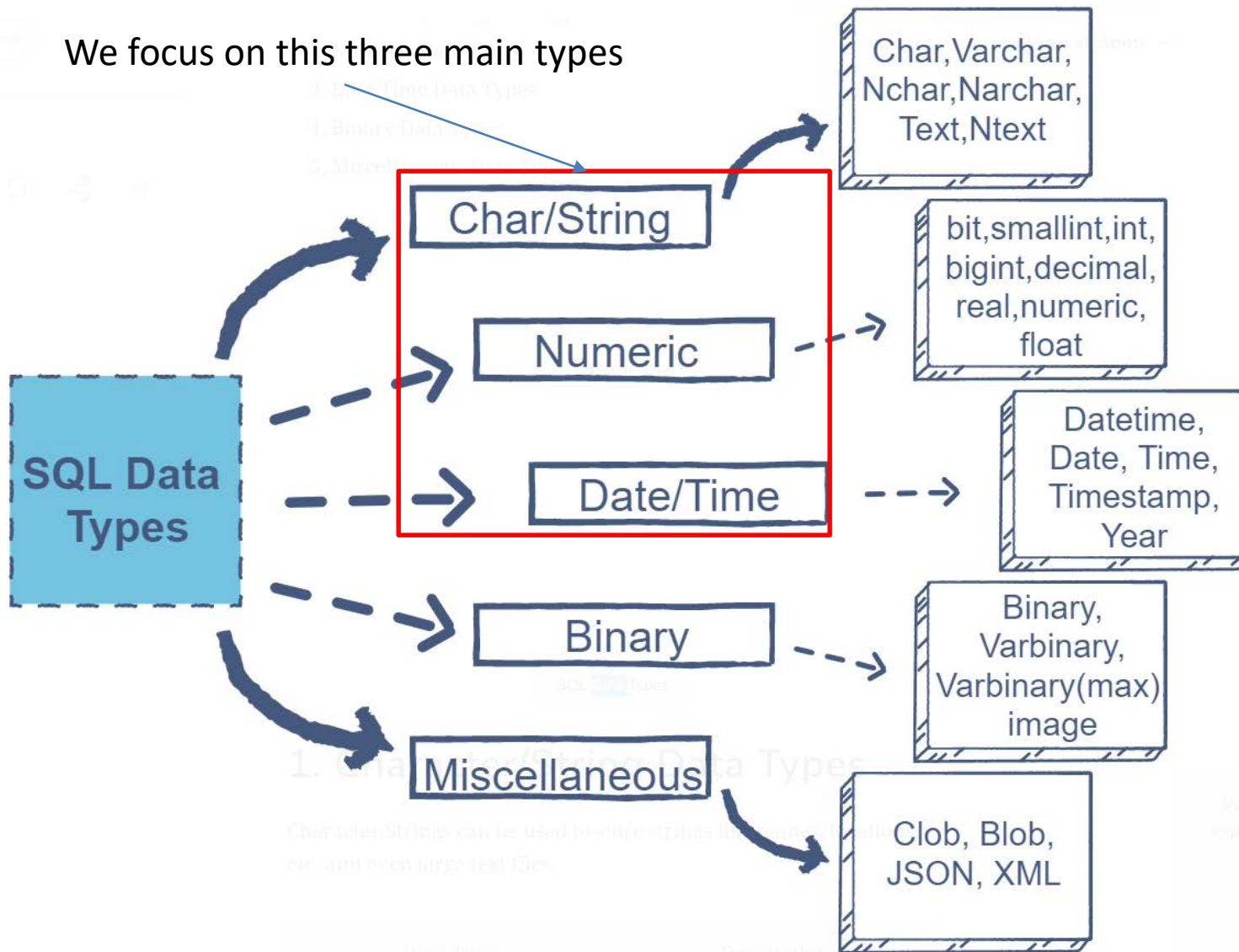
What is matching?

- Data matching is to determine that **one data item is the same as another data item**.
- Data matching is used to **identify duplicate records**.
- Matching is particularly relevant for **character-based** data types.
 - numerical data types, we can use the **equal** sign.
 - character-based data types, it may not be clear.



Data/Variable Types in SQL

We focus on this three main types



Numeric Data Types:

- Numeric data is not as tricky as character data. We can just use the **equal sign**;
 - $5 = 5$
 - $5 \neq 6$
- But, rounding is a problem.
 - $5.029 = 5.03$?
 - If the precision is 2 decimal digits, they are the same.
 - If the precision is 3 decimal digits, they are different.

Datetime Data Types:

- If the data is stored as a **datetime data type**, it's not a problem. We can use the **equal sign** like the numeric data.
- If it is stored as a **character data type**, it could be tricky.
- For example, is **03/01/2021** the same as **01/03/2021**? Is it the same as **2021-03-01T00:00:00Z+06**?
- We **need some logic** here to match them, such as by comparing the components or by comparing the date, month, and year.

Character-based Data Types:

- In the **address table** in a database, we may find that the city name is “**Los Angeles**,” which does not exist in the city table. In this case, we need to match “**Los Angeles**” to “**Los Angeles**.”
- Another example is the customer name “**Mr. Aleck Stevenson**”. We need to match/recognise that “**Mr. Aleck Stevenson**” is the same as “**Mr. S Aleck**”, “**Mr. Stevenson Aleck**” and “**Mr. Alec Stevenson**”

University = univresity?

Data Quality - Inconsistent Before Matching

Invoice 1

Bill no	Customer Name	Social Security Number
111	Mr. Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	Customer Name	Social Security Number
225	Mr. S Aleck	ADWPS10017

Invoice 3

Bill no	Customer Name	Social Security Number
313	Mr. Stevenson Aleck	ADWPS10017

Invoice 4

Bill no	Customer Name	Social Security Number
583	Mr. Alec Stevenson	ADWPS10017

Data Quality - Consistent Data After Matching

Invoice 1

Bill no	Customer Name	Social Security Number
111	Mr. Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	Customer Name	Social Security Number
225	Mr. Aleck Stevenson	ADWPS10017

Invoice 3

Bill no	Customer Name	Social Security Number
313	Mr. Aleck Stevenson	ADWPS10017

Invoice 4

Bill no	Customer Name	Social Security Number
583	Mr. Aleck Stevenson	ADWPS10017

In SQL Server we have three types of matching logic:

- **Exact:** Exact matching is where **all characters are the same**, for example “Los Angeles” and “Los Angeles. In SSIS this is done using a Lookup transformation
- **Fuzzy:** Fuzzy logic matching finds **how similar** a set of data is to another set of data.
- **Rule-based:** Rule-based logic is where we use **certain rules** and data to identify a match.

Data Matching: Fuzzy logic

Fuzzy Lookup Output Data Viewer at Fuzzy Lookup

▶ Detach Copy Data

FullName	InputName	_Similarity	_Confidence	_Similarity_InputName
Roberto Tamburello	Robarto Tamburelo	0.8781369	0.5289686	0.8781369
Terri Lee Duffy	Terri Lee Duffie	0.8937163	0.5101804	0.8937163
Rob Walters	Rob Walters	1	1	1
Michael Raheem	Michael Raheme	0.8218597	0.4266516	0.8218597
Gigi N Matthew	Gigi N Matthew	1	1	1
Jossef H Goldberg	Jossef H Goldbearg	0.9629337	0.554138	0.9629337

- This table is done by using SSIS Lookup transformation. In this viewer, if similarity is 1, it is exactly a match, and confidence defines the quality of the match.

- From the table, “**Jossef H Goldberg**” and “**Jossef H Goldbearg**” have a very high similarity score of 0.9629337 and a confidence level of 0.554138.
- You can then decide for example that if the similarity score is greater than 0.75 and the confidence level is greater than 0.5, then it’s a match.

FullName	InputName	_Similarity	_Confidence	_Similarity_InputName
Roberto Tamburello	Robarto Tamburelo	0.8781369	0.5289686	0.8781369
Terri Lee Duffy	Terri Lee Duffie	0.8937163	0.5101804	0.8937163
Rob Walters	Rob Walters	1	1	1
Michael Raheem	Michael Raheme	0.8218597	0.4266516	0.8218597
Gigi N Matthew	Gigi N Matthew	1	1	1
Jossef H Goldberg	Jossef H Goldbearg	0.9629337	0.554138	0.9629337

Data Matching: Rule-based Logic

- We can define (in a table) that
 - “**Bill**” is the same as “**William**”
 - “**movie**” is the same as “**film**”.
- We can also define rules such as
 - “omit the spaces” so that “**KL 7923 M**” is the same as “**KL7923M**”.
 - “Omit prefix **9000** in the supplier number” so that “**900089123**” and “**89123**” are the same.
- In SSIS, this is implemented with database lookup. SSIS logic can be implemented with Script Component (aside).

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction

Data quality rules (data validation) are filters to prevent dirty data from getting into the warehouse. Based on the data location, there are three kinds of validation:

- Incoming data validation
- Cross-reference validation
- Data warehouse internal validation

- Checking the incoming data **only on its own**, without referring to the data already in the warehouse.
- These rules verify that the data from the **source systems** is valid, e.g. within the expected range and in the right format.
 - product code is in **AAA999AA** format
 - the prices for a product range are **> \$1**
 - the song duration is **between one and ten minutes**,
 - the subscription duration is **≥ one month**



Cross-reference validation is where we check the incoming data **against** the data in the warehouse.

- To make sure that the value of the incoming data is **within a certain range** that is calculated based on data in the warehouse
- Examples
 - the incoming data is expected to be within a 25% range of the **average** of the first column.
 - the incoming unit cost is within 15 percent of the **average of the last three months of supplier costs**.
- Like incoming data validation, cross-reference validation is **performed on the fly** when the data is being loaded into the warehouse.

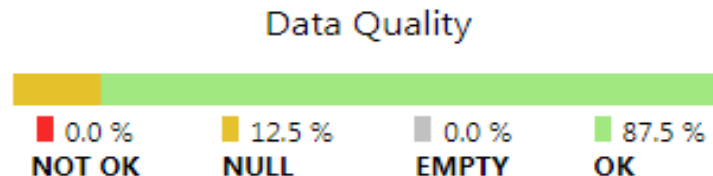
- Data warehouse internal validation is where we check the data already in the warehouse.
 - We don't check the incoming data.
- Unlike the previous two, data warehouse internal validation is performed after the incoming data is fully loaded into the warehouse.
- The purpose of doing this is to verify the quality of the data in the warehouse at the aggregate level.
 - This can be done by comparing the totals over a period of time against a known standard value.

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction

In process of data warehouse design, many database professionals face situations like:

- Several **data inconsistencies** in source, like **missing records** or **NULL values**.
- Column to be the primary key column is **not unique** throughout the table.
- Schema design is not coherent to the end user requirement.

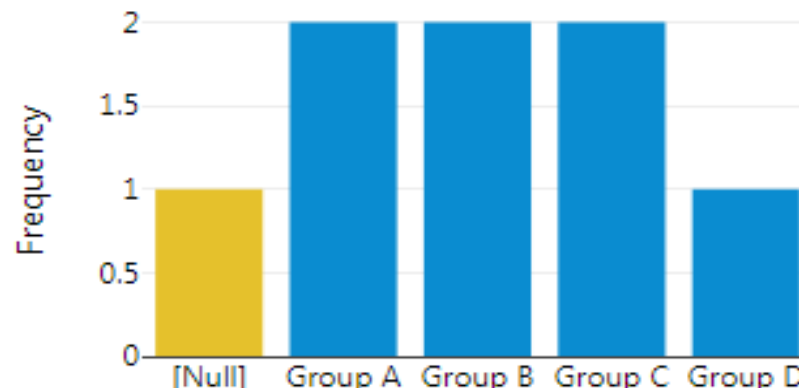
Solution



- it would be better to catch them right at the start before they become a problem. Use **data profiling**

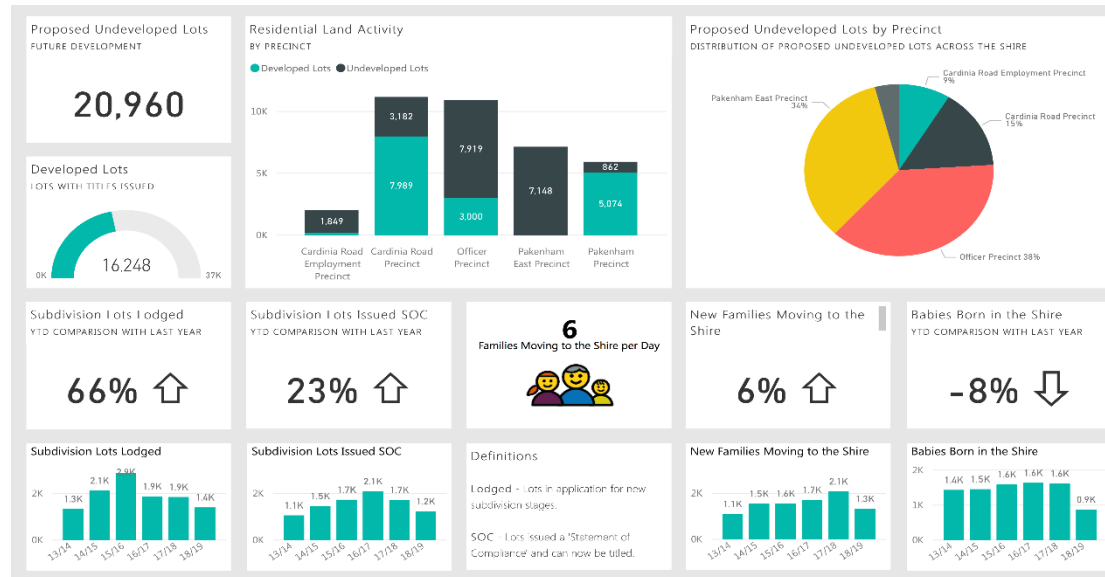
What is data profiling ?

- It is the process of **statistically examining** and **analysing** the content in a data source
 - collect information about the data.
 - evaluate accuracy and completeness.
 - make a thorough assessment of **data quality**
- It assists the discovery of **anomalies** in data.
- It helps understand content, structure, relationships, etc.



How to conduct Data Profiling?

- Data profiling involves **statistical analysis** of the data at source, as well as analysis of metadata.
- These statistics may be used for various analysis purposes:
 - E.g. analyse the quality of data at the data source



- **NULL values:** find the number of NULL values in attributes
- **Candidate keys:** Analysis of the extent to which certain columns are **distinct** will give developer useful information
 - Help selection of candidate keys
 - Primary key must **not be NULL** or have **duplicate values**.
- **String values:**
 - NULL or empty string may create problems
 - An analysis of **largest** and **shortest** length as well as the **average** string length of a string-type column can help decide what data type to be used for the column
- **Identification of cardinality:** The cardinality relationships are important for inner and outer join of tables.
- **Data format:** the format used in some columns may or may not be user-friendly.

Common Data Profiling Software

- Most of the data-integration/analysis softwares have data profiling built into them. Some popular ones are:
 - SSIS Data Profiling Task
 - IBM InfoSphere Information Analyzer
 - SAP Business Objects Data Services (BODS) for Data Profiling
 - Oracle Warehouse Builder
- You can perform data profiling using Python as well

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction

- **Attribute (or dimensions, features, variables)**
 - A data field, representing a characteristic or feature of a data object.
 - *E.g. customer_ID, name, address*
- **Types:**
 - Nominal/Categorical (e.g. red, blue)
 - Binary (e.g. {true, false})
 - Ordinal (e.g. {freshman, sophomore, junior, senior})
 - Numeric: quantitative

Q1: Is student ID a nominal, ordinal, or numeric?

Q2: What about eye color?

- **Nominal: categories, states, or “names of things”**
 - *Hair_color* = {auburn, black, blond, brown, grey, white...}
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important
e.g. gender
 - Asymmetric binary: outcomes not equally important.
e.g. medical test (positive vs. negative)
- **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known
 - *Size* = {small, medium, large}, grades, army rankings
- **Numeric**
 - COVID-19 confirmed cases, recovery rate, ...

- **Discrete Attribute**

- Has only a finite or countably infinite set of values
 - E.g. **zip codes**, **age**, or **country names**
- Sometimes, represented as **integer** variables
- Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**

- Has real numbers as attribute values
 - E.g. **temperature**, **height**, or **weight**
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as **floating-point** variables

Data Matrix and Dissimilarity Matrix

- **Data matrix**

n instances with p attributes
(measurements).

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- **Dissimilarity matrix**

$d(i,j)$ is the dissimilarity
between instances i and j

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

- Standardise data
 - Calculate the mean absolute deviation:
 - $s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$

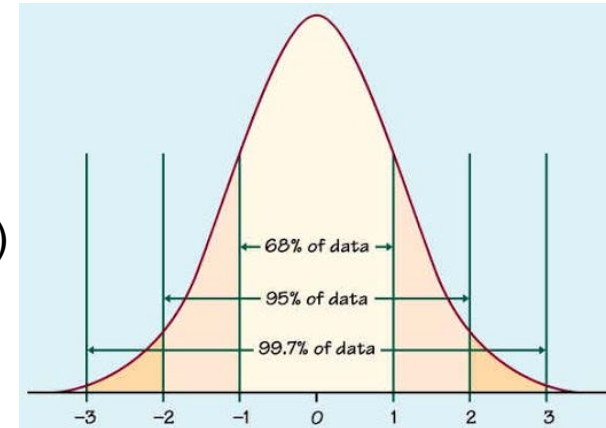
where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

- Calculate the standardised measurement (z-score)

- $z_{if} = \frac{x_{if} - m_f}{s_f}$

- In practice, using mean absolute deviation tends to be more robust than using standard deviation.



- Distances are normally used to measure the **similarity** or **dissimilarity** between two data objects
- Some popular ones include: Minkowski distance:

- $$d(i, j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q}$$

- where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional instances, and q is a positive integer.

- If $q=1$, d is **Manhattan** Distance

- $$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- If $q=2$, d is **Euclidean** Distance

- $$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

- Properties:

- $d(i, j) \geq 0$; $d(i, j) = 0$; $d(i, j) = d(j, i)$ and

- $d(i, j) \leq d(i, k) + d(j, k)$; (**Triangle inequality**)

- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Compute Dissimilarity for Binary Attributes

- **A contingency table for binary data**
 - Symmetric: both matches are equally important.
 - Asymmetric: the match of “0” is not important.

- **Distance Measure for**

- Symmetric binary attributes:

- $$d(i, j) = \frac{b+c}{a+b+c+d} = \frac{b+c}{p}$$

- Asymmetric binary attributes:

- $$d(i, j) = \frac{b+c}{a+b+c}$$

		Instance j		
Instance i		1	0	total
	1	a	b	a + b
	0	c	d	c + d
	total	a + c	b + d	p

- **Jaccard Coefficient (similarity measure for asymmetric binary attributes):**

- $$sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$$

- **Convert binary attribute into numerical attribute**

Dissimilarity between Asymmetric Binary Attributes

- Given the following example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

		Mary		
		1	0	Σ_{row}
Jack	1	2	0	2
	0	1	3	4
	Σ_{col}	3	3	6

- Gender is a **symmetric** attribute (**not counted in**)
- The remaining attributes are **asymmetric** binary
- Let the values of Y and P to be 1, and value N to be 0.

- We have

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

		Jim		
		1	0	Σ_{row}
Jack	1	1	1	2
	0	1	3	4
	Σ_{col}	2	4	6

		Mary		
		1	0	Σ_{row}
Jim	1	1	1	2
	0	2	2	4
	Σ_{col}	3	3	6

- A generalisation of the binary variable in that it can take more than 2 states, e.g. red, yellow, blue, green
- Method 1: Simple matching
 - m : # of matches, p : total # of attributes
 - $d(i, j) = \frac{p-m}{p}$
- Method 2: convert to a number of binary attributes
 - creating a new binary attribute for each of the M possible states

- An ordinal attribute is often discrete.
- Order is important, e.g. rank (e.g. freshman, sophomore)
- Can be treated as numeric attributes
 - replace x_{if} by their rank $r_{if} \in \{1, \dots, M_f\}$
 - map the range of each attribute onto $[0, 1]$ by replacing i -th object in the f -th attribute by
 - $z_{if} = (r_{if} - 1)/(M_{if} - 1)$
 - example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
 - distance: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
 - compute the dissimilarity using methods for numeric attributes

- A database may contain **different types** of attributes
 - symmetric binary, asymmetric binary, nominal, ordinal, and numeric attributes
- One may use a weighted formula to combine their effects

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^f d_{ij}^f}{\sum_{f=1}^p \delta_{ij}^f}$$

- f is binary or nominal:

$$\begin{aligned} d_{ij}^f &= 0, & \text{if } x_{if} &= x_{jf} & \text{or} \\ d_{ij}^f &= 1, & & \text{otherwise} \end{aligned}$$

- f is numeric: use Euclidean distance
- f is ordinal

- compute ranks z_{if} where $z_{if} = (r_{if} - 1)/(M_{if} - 1)$
- and treat z_{if} as numeric attribute

Similarity of Two Vectors

- Vector objects: keywords in documents, gene features in micro-arrays, etc.

<i>Document</i>	<i>team</i>	<i>coach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Broad applications: information retrieval, natural language understanding, etc.
- Cosine measure

$$- s(x, y) = \frac{d_1^T \cdot d_2}{\|d_1\| * \|d_2\|}$$

- A variant: Tanimoto coefficient-used in information retrieval

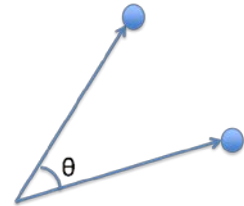
$$- s(x, y) = \frac{d_1^T \cdot d_2}{d_1^T \cdot d_1 + d_2^T \cdot d_2 - d_1^T \cdot d_2}$$

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity:

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$$

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



where \bullet indicates vector dot product, $\|d\|$: the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \quad d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

- First, calculate vector dot product

$$\begin{aligned} d_1 \bullet d_2 &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 \\ &\quad + 0 \times 1 = 25 \end{aligned}$$

- Then, calculate $\|d_1\|$ and $\|d_2\|$

$$\|d_1\| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\|d_2\| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

- Calculate cosine similarity: $\cos(d_1, d_2) = 25 / (6.481 \times 4.12) = 0.94$

- Types of Data Sets
- Data Quality
 - ✓ Data Quality Overview and Examples
 - ✓ Cleansing and Matching
 - ✓ Data Quality Validation
- Data Profiling
- Attribute Types and Similarity Measure
- Data Reduction

- **Data reduction:**
 - Obtain a reduced representation of the data set that is
 - much smaller in volume
 - but yet produces the same (or almost the same) analytical results.
- **Why data reduction?**
 - Increases storage capacity
 - Easy and efficient mining, reduces time and memory requirement
 - Easy visualisation
 - Help to eliminate irrelevant /redundant features
 - Reduces noise

- Attributes (Feature Reduction)
 - Discrete Wavelet Transform (DWT)
 - Principal Component Analysis (PCA)
 - Attribute Subset Selection
- Instances (Numerosity Reduction)
 - Parametric methods
 - A model is used to estimate the data
 - Only model parameters are stored
 - Non-parametric methods
 - Histogram, Clustering, Sampling
 - Data Cube Aggregation
 - Data Compression
 - Lossless, Lossy

- References
 - “Building a Data Warehouse with Examples in SQL Server”, 2008, by Vincent Rainardi
- Readings
 - Chapter 9 of Vincent Rainardi’s book
 - Chapter 8 of Han et al.’s book



Copyright Notice

Material used in this recording may have been reproduced and communicated to you by or on behalf of **The University of Western Australia** in accordance with section 113P of the *Copyright Act 1968*.

Unless stated otherwise, all teaching and learning materials provided to you by the University are protected under the Copyright Act and is for your personal use only. This material must not be shared or distributed without the permission of the University and the copyright owner/s.