上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Introduction to Tabular/Table Data

Teacher: Zheng Wang

Social Network Analysis (NIS8023)

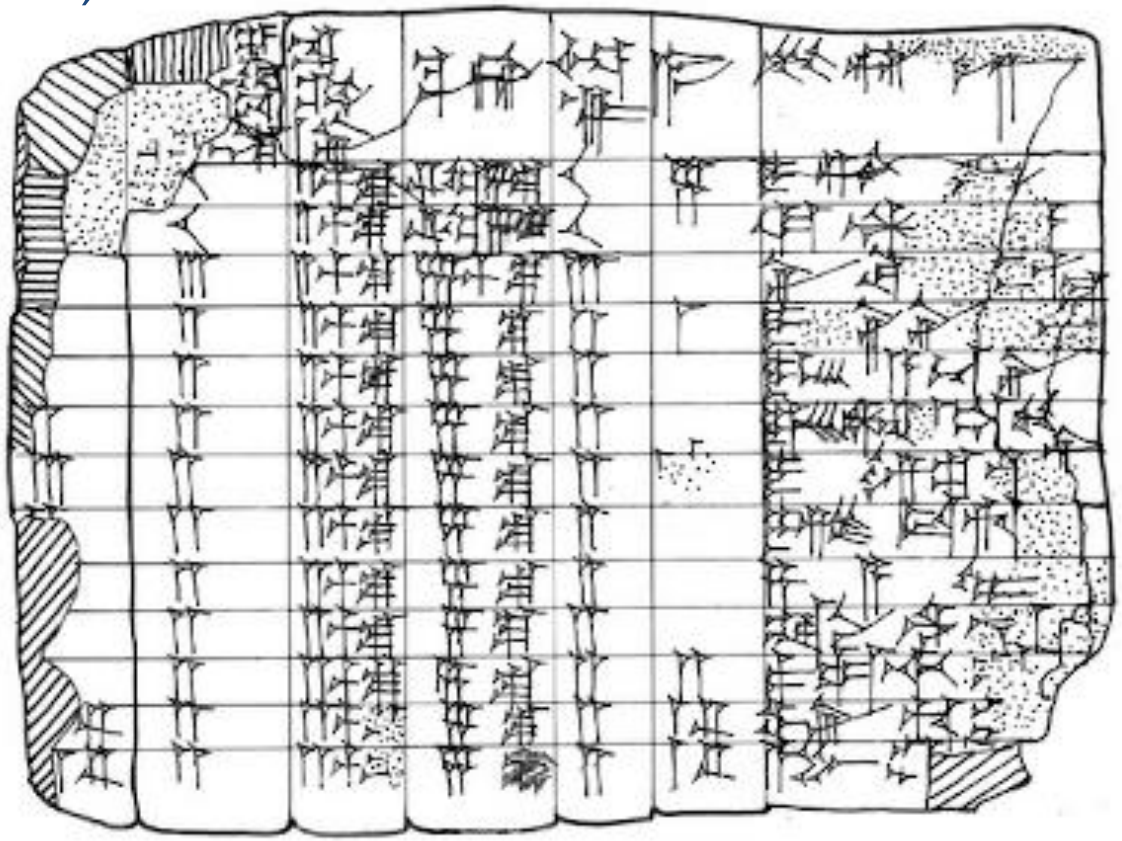
Shanghai Jiao Tong University

Outline

- **What's table/tabular data**
- Relational table data
- Data Pre-processing

What is tabular/table data

Tabular (Table) data is the basic and most efficient form of data management.



Old Babylonian (ca. 1900-1600 BC)

Note: As we will discuss relational databases later, people often prefer the term "relational table data."

https://cdli.mpiwg-berlin.mpg.de/search?layout=full&id=P368686

What is table data

A table is an arrangement of information or data, typically in rows and columns, or possibly in a more complex structure. [1]

Student table			
s_id	name	gender	age
1	Tom	F	20
2	John	F	21
3	Lily	M	20
4	Alice	M	21
5	Lucy	M	19

Simple table (2D, our focus)

Purchased Equipments (June, 2006)			
Item Num#	Item Picture	Item Description	Price
		Shipping Handling, Installation, etc	Expense
1.		IBM Clone Computer.	$ 400.00
		Shipping Handling, Installation, etc	$ 20.00
2.		1GB RAM Module for Computer.	$ 50.00
		Shipping Handling, Installation, etc	$ 14.00
Purchased Equipments (June, 2006)			

A TABLE of the Apertures of Object-Glaſſes.

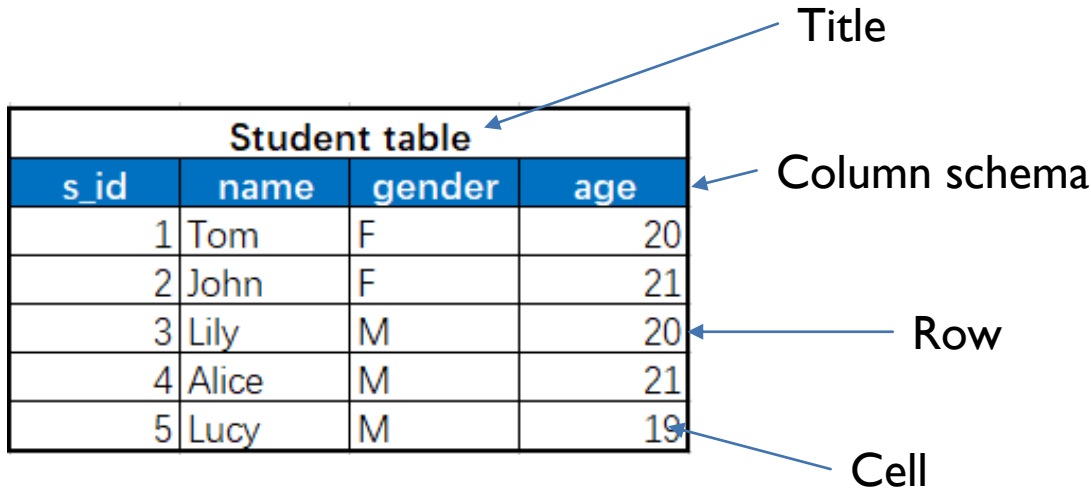
The Points put to ſome of theſe Numbers denote Fractions.

Some complex tables

[1] https://en.wikipedia.org/wiki/Table_(information)

Some key features of 2-D simple table data

- Title: describes the summary of the table content.
- Column:
  - Each column stands a specifical attribute
  - Column name/header describes the column meaning and value types
- Row: a collect of columns, each row describes a single data sample.
- Cell: the intersection of a row and a column.



Student table			
s_id	name	gender	age
1	Tom	F	20
2	John	F	21
3	Lily	M	20
4	Alice	M	21
5	Lucy	M	19

Table data vs non-table data

- Table data: a dense and "meaningful attribute"-format data
  - Attribute number is countable
  - Each attribute/column should be human understandable
  - Matrix form should be dense

Some typical non-table data

Text data

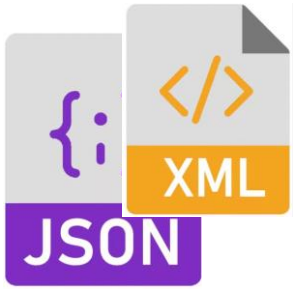
(sequences of words, and symbols …)

Audio data

(numerical samples in a continuous sequence)

Image data

(the underlying pixel values of an image)



"Self-describing" data

(human-readable text to format/describe data)

Why is not table data

1. Attribute number countable	✗	✗	✗	✗
2. Attribute meaningful	✓	✗	✗	✓
3. Dense matrix form	✗	✓	✓	✗

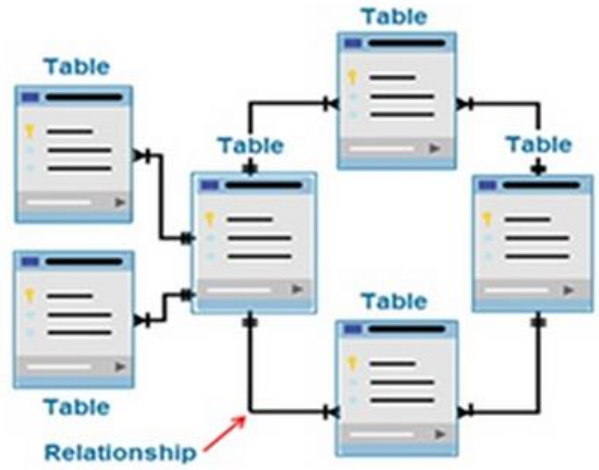
Note: more discussing may be needed.

Outline

- What's table data
- **Relational table data**
- Data Pre-processing

Preliminary: Relational Database

- A relational database is a database based on the relational model of data, as proposed by E. F. Codd in 1970.
  - Present the data to the user as relations (a presentation in tabular form, i.e. as a collection of tables with each table consisting of a set of rows and columns);
  - Provide relational operators to manipulate the data in tabular form.

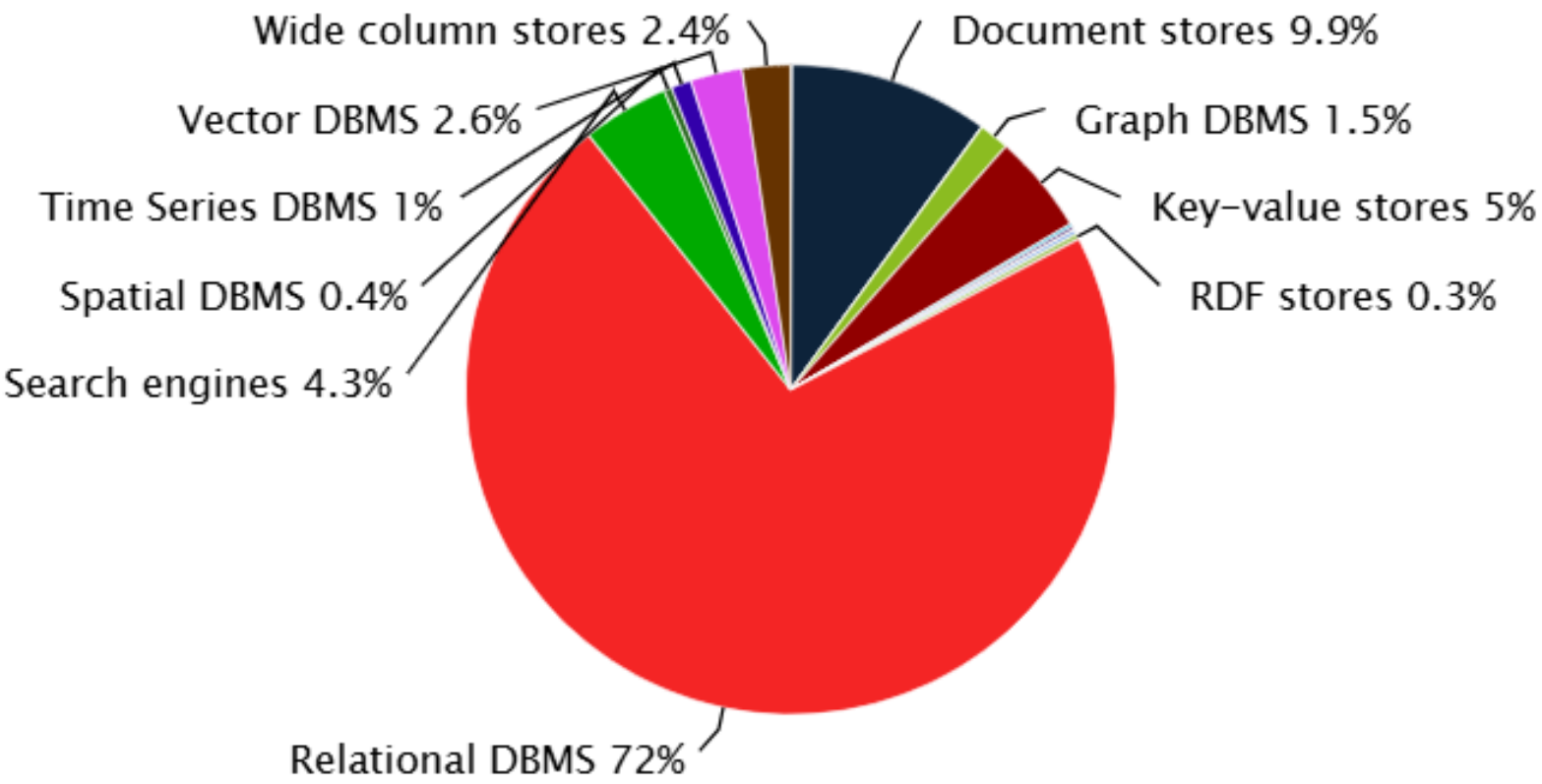


RDBMS

Relational Databases

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

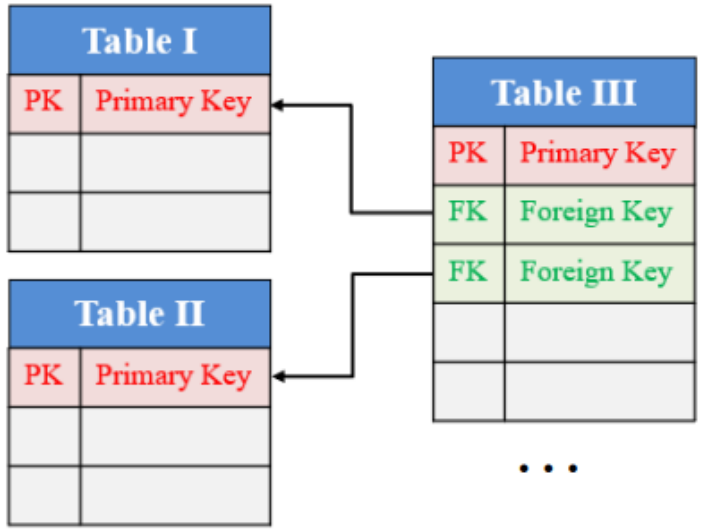
Relational databases rule the world



Ranking scores per category in percent, February 2025

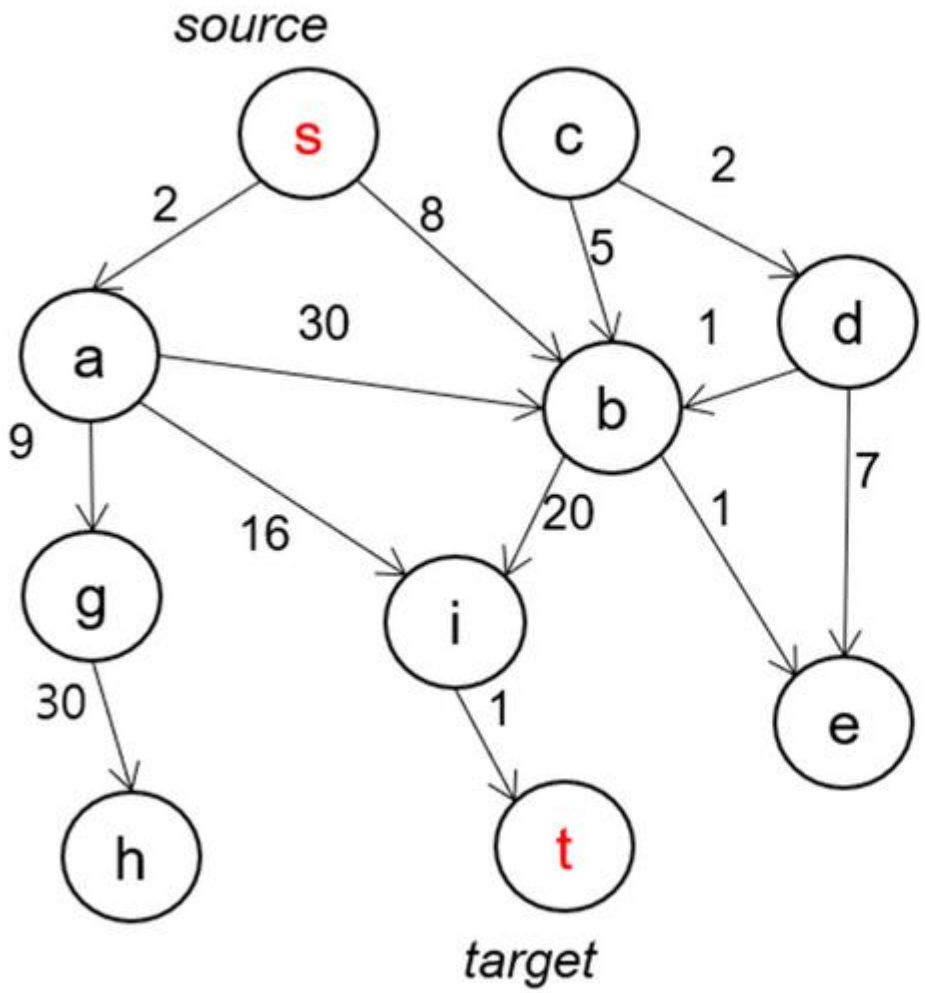
Relational table data

- Relational tables (the core of Relational Database) are a set of tables which all have two "keys":
  - Primary key: a unique identifier of a row
  - Foreign key: a reference to a primary key, to create relationship between different tables



Relational tables

Example of relational table (I)



TN

src
S
a
b
c
d
.

TE

Src	dst	cost
S	a	2
S	b	8
a	g	9
a	b	30
a	I	16
.	...	

Relational Model for Graph

Example of relational table (II)

Purchase table

Transaction ID	Customer ID	Product ID	Purchange date
1112	24221	8977	03-22-2010
1113	24222	8978	03-22-2010
1114	24223	8979	03-22-2010

Customer ID	Customer	Address
24221	Bob	123 East street
24222	Alice	223 Main street
24223	Martha	465 North street

Customer table

Product ID	Name	Price
8977	Banana	.79
8978	TV	400
8979	Watch	50

Product table

Table Metadata

- Metadata is "data that provides information about other data", but not the content of the data itself. [1]
- Metadata of table data:
  - Table-level: table name, size, row number and (intra/inter relationships) etc.
  - Column-level: column name, numeric type and etc.

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Cornelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Data

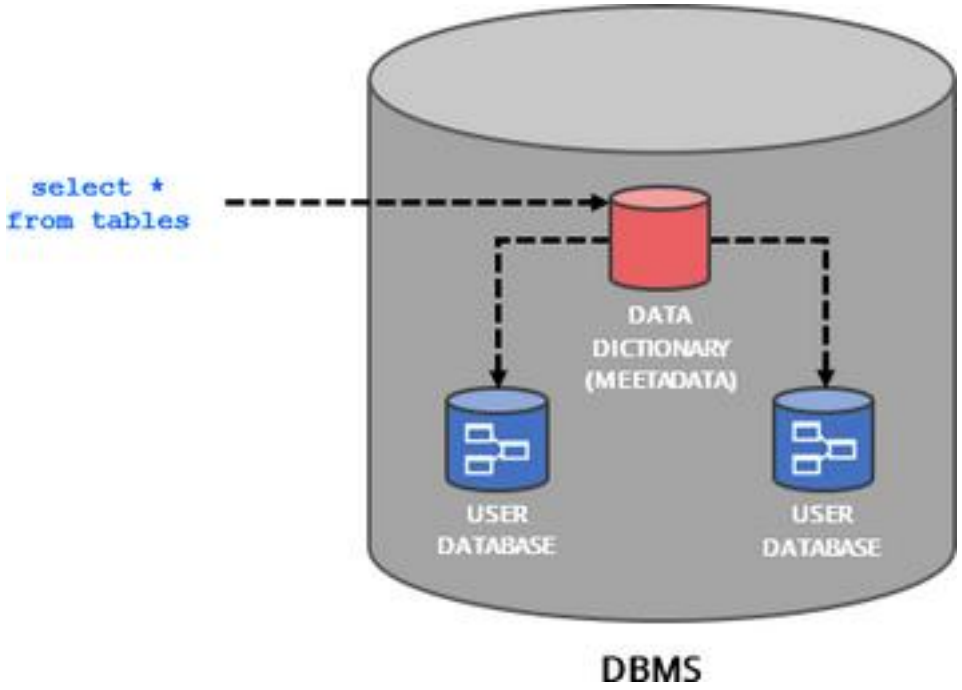
Metadata

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee sti

[1] https://en.wikipedia.org/wiki/Metadata https://dataedo.com/kb/data-glossary/what-is-metadata

Metadata in RDBs

- Metadata in relational databases, including all the information that make up that database's schema, like:
  - Table names
  - Field names
  - Entity keys
  - Foreign keys
  - Data types
  - Views
  - Integrity constraints



Outline

- What's tabular data
- Relational table data
- **Data Pre-processing**
  - Empty Data Imputation
  - Metadata Prediction
  - Error Data Repair

Cell Data Imputation

- Given a table with corrupted/missing values, predict the missing cell values.

Population in Millions by Country

Country	Capital	Population
Australia	Canberra	25.69
[redacted]	Paris	67.39
Bolivia	La Paz	11.67

Population in Millions by Country

Country	Capital	Population
Australia	Canberra	25.69
France	Paris	67.39
Bolivia	La Paz	11.67

(Deng et al. 2020; Tang et al, 2021)

Empty Cell Imputation

- Passive method:
  - Directly delete missing values for each row/column.
  - Use placeholders (e.g., N/A, NaN) to defer handling the issue.
- Aggressive method: Impute a new value (sometimes also add a new column representing whether it is imputed), like
  - Mean Imputation
  - Median Imputation
  - Nearest Neighbor Imputation
  - Model-based Imputation

Aggressive method: Mean Imputation

1. The mean-filling is to replace the missing value with the mean in the data set. Suppose a dataset X with n observations, where X_i represents the i-th observation and there are m missing values. The mean of the data set is:

$$\bar{X} = \frac{1}{n-m} \sum_{i=1}^{n} X_i$$

2. Replace the missing value with the mean:

$$X_i' = \begin{cases} \bar{X}, & \text{if } X_i \text{ is missing} \\ X_i, & \text{otherwise} \end{cases}$$

Aggressive method: Median Imputation

1. The median filling is to replace the missing value with the median in the data set. We need to sort the data set and calculate the median.

2. Replace the missing value with the median:

$$X_i^{'} = \begin{cases} \text{Median}(X), & \text{if } X_i \text{ is missing} \\ X_i, & \text{otherwise} \end{cases}$$

Aggressive method: Nearest Neighbor Imputation

1. The nearest filling algorithm selects the closest non-missing value for filling according to the characteristics of the missing value. Euclidean distance or other distance measures can be used for distance calculation.
2. For each missing value, we can calculate the distance to its closest non-missing value and select the observation with the smallest distance to fill in.

$$X_i^{'} = \text{argmin}_{j \neq i} d(X_i, X_j)$$

Aggressive method: Model-based Imputation

1. The model filling algorithm uses a machine learning model (denoted as $f()$) to make predictions about the data and fill in missing values. Common models include linear regression, decision trees, and support vector machines.
2. Train the model on available data and use it to predict missing values based on other observations, preserving relationships and dependencies for improved accuracy.

$$\hat{X}_i' = f(X_{-i})$$

Metadata prediction

- Metadata of table data:
  - Table-level: table name, size, row number and (intra/inter relationships) etc.
  - Column-level: column name, numeric type and etc.

Population in Millions by Country

[redacted]	Capital	Population
Australia	Canberra	25.69
France	Paris	67.39
Bolivia	La Paz	11.67

Predict that the missing column header is **Country**

Predict that the table type is a **relational** table

Gonzalez-Perez C (2018). "Metainformation". In Gonzalez-Perez C (ed.). Information modelling for archaeology and anthropology: software engineering principles for cultural heritage (1st ed.). Springer Cham. pp. 181–189. ISBN 978-3-319-72652-6.

Error Data Repair

- Data repair is the process of fixing errors and inconsistencies in data to ensure its accuracy and reliability for data analysis.

Address			
Street	City	State	CountryCode
123 Main St.	Borington	ON	CAN
456 Elm St.	Hickton	CA	USA
4321 Oak Lane	New York	NY	US

Address			
Street	City	State	CountryCode
123 Main St.	Borington	ON	CA
456 Elm St.	Hickton	CA	US
4321 Oak Lane	New York	NY	US

Basic steps of error data repair

1. Find the data is wrong
    1. Rule-based method: find data breaks the rules (like the column schema or constraint among rows/cols)
    2. Model-based method: train a ML model to find those error data
2. Correct the error data
    1. Adopt the imputation method
    2. Also partly considering the original error one
3. Provide some “marker” to mark this process

Thanks for your time.
QA.