DATA ANALYTICS

Unit 1:Data Cleaning

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Data Cleaning

Mamatha H R

Department of Computer Science and Engineering

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)

Data Cleaning

- inconsistent: containing discrepancies in codes or names, e.g.,
  - *Age*="42", *Birthday*="03/07/2010"
  - Was rating "1, 2, 3", now rating "A, B, C"
  - discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

1. Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
2. Fill in the missing value manually: tedious + infeasible?
3. Fill in it automatically with a global constant : e.g., “unknown” or -∞ .Problem : a new class?
4. Use a measure of central tendency(mean or median) to fill the missing value
5. Use the attribute mean for all samples belonging to the same class: smarter
6. Use the most probable value: inference-based such as Bayesian formula or decision tree- the most popular

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

1. Binning
2. Regression
3. Clustering
4. Combined computer and human inspection

How to Handle Noisy Data?

1. Binning

Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it.

The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing.

There three binning methods,

Binning methods for data smoothing

❑ **Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

* **Partition into equal-frequency (equi-depth) bins:**
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* **Smoothing by bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* **Smoothing by bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

How to Handle Noisy Data?

1. Regression
   - smooth by fitting the data into regression functions
2. Clustering
   - detect and remove outliers
3. Combined computer and human inspection
   - detect suspicious values and check by human (e.g., deal with possible outliers)

Outlier analysis: Outliers may be detected by clustering

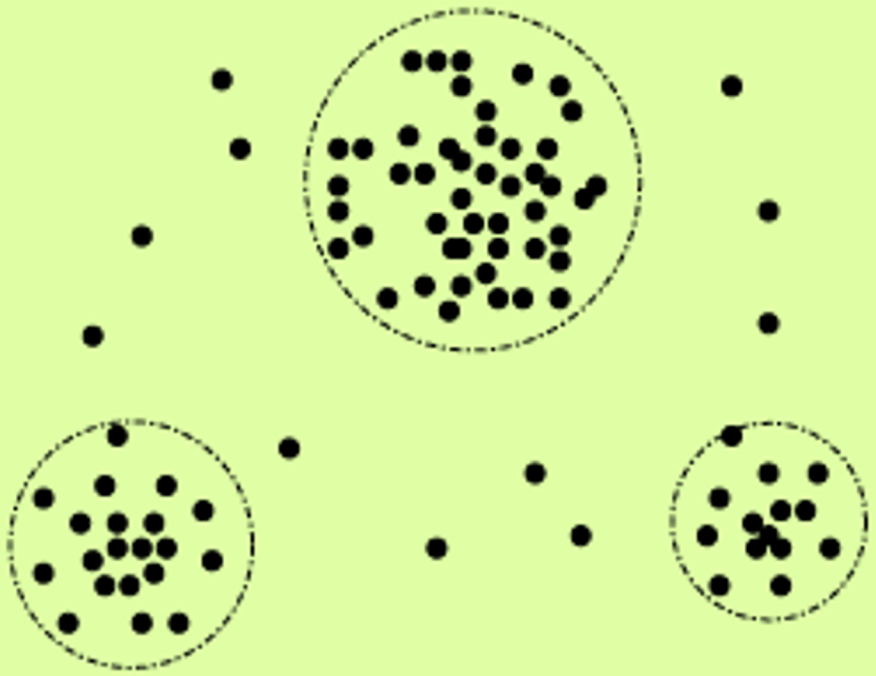


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

Data Cleaning as a Process- Various Approaches

- Missing values, noise, and inconsistencies contribute to inaccurate data.
- So far, we have looked at techniques for handling missing data and for smoothing data.
- "But data cleaning is a big job".
- What about data cleaning as a process?
- How exactly does one proceed in tackling this task?
- Are there any tools out there to help?

Data Cleaning as a Process

1. **Data discrepancy detection** : This is the first step in data cleaning process.

2. **Data Transformations :** Once we find discrepancies , we need to apply series of transformations to correct them .This is the second step

Note : These two steps iterates.
This process is , however is error prone and time consuming.

Data Cleaning as a Process

1.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
  - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
  - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

2.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)

Data Cleaning as a Process

1.Data discrepancy detection : This is the first step in data cleaning process.

As a starting point get knowledge about data and examine the data, adapt the following steps

1. Use metadata
2. Check field overloading
3. Check uniqueness rule, consecutive rule and null rule
4. Use commercial tools for data scrubbing and auditing tools
5. Data Migration tools

Data Cleaning as a Process

1. Use metadata (e.g., domain, range, dependency, distribution)

- what are the domain and data type of each attribute?
- What are the acceptable values for each attribute?
- What is the range of the length of values?
- Do all values fall within the expected range?
- Are there any known dependencies between attributes?

Data Cleaning as a Process

2. Field Overloading:

Field overloading is another error source that typically results when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes (e.g.,
an unused bit of an attribute that has a value range that uses only, say, 31 out of 32 bits).

Data Cleaning as a Process : Data Discrepancy Detection

3. Check uniqueness rule, consecutive rule and null rule

The data should also be examined regarding: –

– A **unique rule** **:** It says that each value of the given attribute must be different from all other values for that attribute.

– A **consecutive rule** **:** It says that there can be **no missing** values between the lowest and highest values for the attribute, and that all values must also be **unique** (e.g., as in check numbers).

– A **null rule** **:** It specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values should be handled.

4. Commercial data scrubbing tools

- Winpure
- OpenRefine : OpenRefine (previously Google Refine) is a Opensource tool
- Cloudingo
- DataLadder

Data Cleaning as a Process

2. **Data Transformations :** Once we find discrepancies , we need to apply series of transformations to correct them .This is the second step

Some data inconsistencies may be corrected manually using external references.
Most errors, however, will require data transformations.

Commercial tools can assist in the data transformation step.
Data migration tools allow simple transformations to be specified such as to replace the string “gender” by “sex.”

Data Cleaning as a Process

- 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

Data Cleaning as a Process

1. **Data discrepancy detection** : This is the first step in data cleaning process.

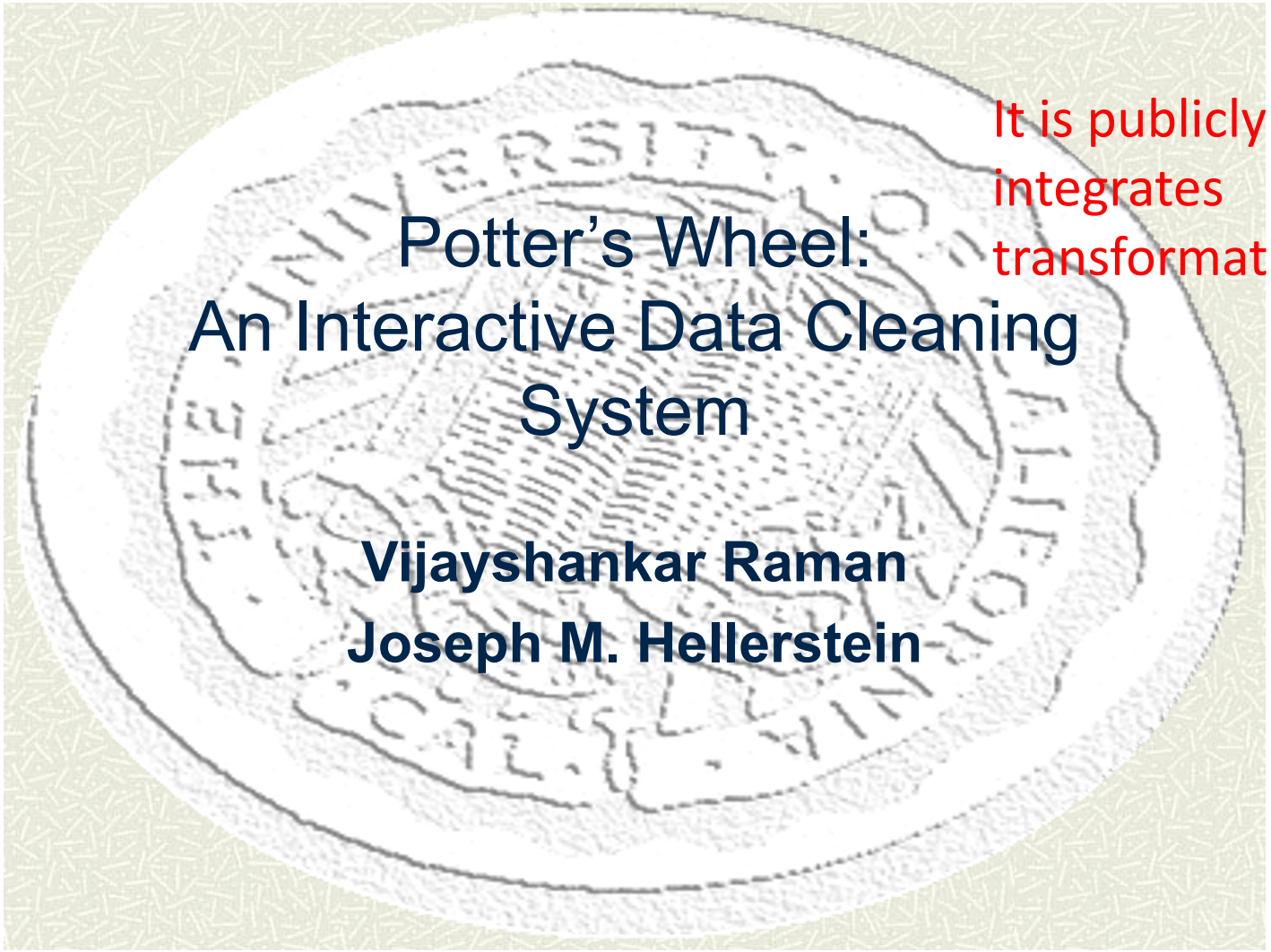
2. **Data Transformations :** Once we find discrepancies , we need to apply series of transformations to correct them .This is the second step

Note :

- These two steps iterates.
- This process is , however is error prone and time consuming.
- Some transformations may introduce more discrepancies.
- The entire data cleaning process also suffers from a lack of interactivity.

- Integration of the two processes- New approach
  - Iterative and interactive (e.g., Potter's Wheels)

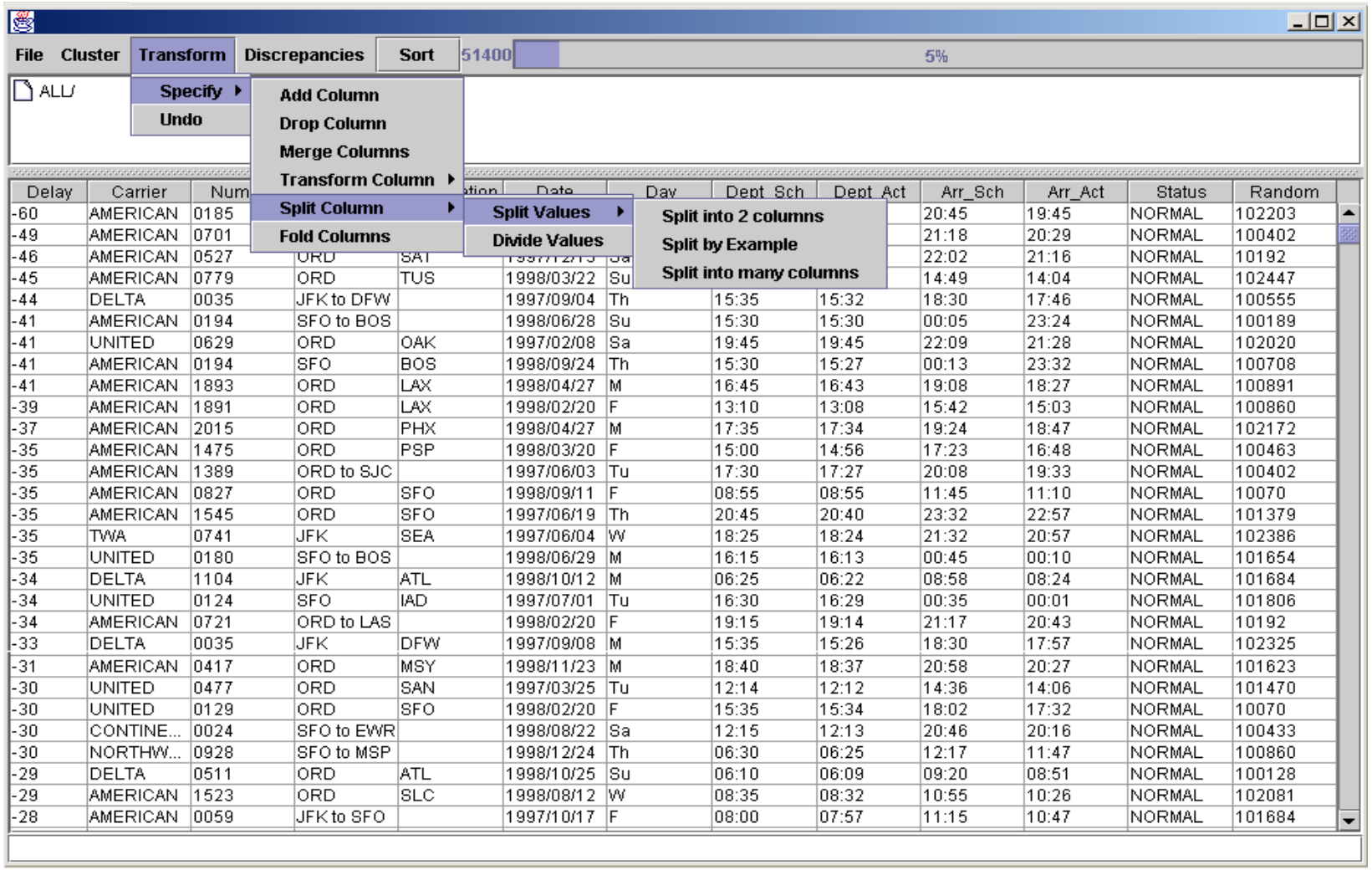
Potter's wheel: An Interactive Data Cleaning System



It is publicly available data cleaning tool that integrates discrepancy detection and transformation.

Source :http://control.cs.berkeley.edu/abc/

Potter's wheel: An Interactive Data Cleaning System



Delay	Carrier	Num			Date	Day	Dept_Sch	Dept_Act	Arr_Sch	Arr_Act	Status	Random
-60	AMERICAN	0185							20:45	19:45	NORMAL	102203
-49	AMERICAN	0701							21:18	20:29	NORMAL	100402
-46	AMERICAN	0527	ORD	SAT					22:02	21:16	NORMAL	10192
-45	AMERICAN	0779	ORD	TUS	1998/03/22	Su			14:49	14:04	NORMAL	102447
-44	DELTA	0035	JFK to DFW		1997/09/04	Th	15:35	15:32	18:30	17:46	NORMAL	100555
-41	AMERICAN	0194	SFO to BOS		1998/06/28	Su	15:30	15:30	00:05	23:24	NORMAL	100189
-41	UNITED	0629	ORD	OAK	1997/02/08	Sa	19:45	19:45	22:09	21:28	NORMAL	102020
-41	AMERICAN	0194	SFO	BOS	1998/09/24	Th	15:30	15:27	00:13	23:32	NORMAL	100708
-41	AMERICAN	1893	ORD	LAX	1998/04/27	M	16:45	16:43	19:08	18:27	NORMAL	100891
-39	AMERICAN	1891	ORD	LAX	1998/02/20	F	13:10	13:08	15:42	15:03	NORMAL	100860
-37	AMERICAN	2015	ORD	PHX	1998/04/27	M	17:35	17:34	19:24	18:47	NORMAL	102172
-35	AMERICAN	1475	ORD	PSP	1998/03/20	F	15:00	14:56	17:23	16:48	NORMAL	100463
-35	AMERICAN	1389	ORD to SJC		1997/06/03	Tu	17:30	17:27	20:08	19:33	NORMAL	100402
-35	AMERICAN	0827	ORD	SFO	1998/09/11	F	08:55	08:55	11:45	11:10	NORMAL	10070
-35	AMERICAN	1545	ORD	SFO	1997/06/19	Th	20:45	20:40	23:32	22:57	NORMAL	101379
-35	TWA	0741	JFK	SEA	1997/06/04	W	18:25	18:24	21:32	20:57	NORMAL	102386
-35	UNITED	0180	SFO to BOS		1998/06/29	M	16:15	16:13	00:45	00:10	NORMAL	101654
-34	DELTA	1104	JFK	ATL	1998/10/12	M	06:25	06:22	08:58	08:24	NORMAL	101684
-34	UNITED	0124	SFO	IAD	1997/07/01	Tu	16:30	16:29	00:35	00:01	NORMAL	101806
-34	AMERICAN	0721	ORD to LAS		1998/02/20	F	19:15	19:14	21:17	20:43	NORMAL	10192
-33	DELTA	0035	JFK	DFW	1997/09/08	M	15:35	15:26	18:30	17:57	NORMAL	102325
-31	AMERICAN	0417	ORD	MSY	1998/11/23	M	18:40	18:37	20:58	20:27	NORMAL	101623
-30	UNITED	0477	ORD	SAN	1997/03/25	Tu	12:14	12:12	14:36	14:06	NORMAL	101470
-30	UNITED	0129	ORD	SFO	1998/02/20	F	15:35	15:34	18:02	17:32	NORMAL	10070
-30	CONTINE...	0024	SFO to EWR		1998/08/22	Sa	12:15	12:13	20:46	20:16	NORMAL	100433
-30	NORTHW...	0928	SFO to MSP		1998/12/24	Th	06:30	06:25	12:17	11:47	NORMAL	100860
-29	DELTA	0511	ORD	ATL	1998/10/25	Su	06:10	06:09	09:20	08:51	NORMAL	100128
-29	AMERICAN	1523	ORD	SLC	1998/08/12	W	08:35	08:32	10:55	10:26	NORMAL	102081
-28	AMERICAN	0059	JFK to SFO		1997/10/17	F	08:00	07:57	11:15	10:47	NORMAL	101684

Exercise

- Explore how binning, clustering, regression is used in handling noisy data.
- Is Combined computer and human inspection of noisy data is a better way of handling the noisy data.Give reasons.
- Explain the process of data cleaning.

References

Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining by Tan, Steinbach, Kumar, 2nd Edition

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834