

Data Preparation: Integration and Cleaning

Mining Massive Datasets

Prof. Carlos Castillo

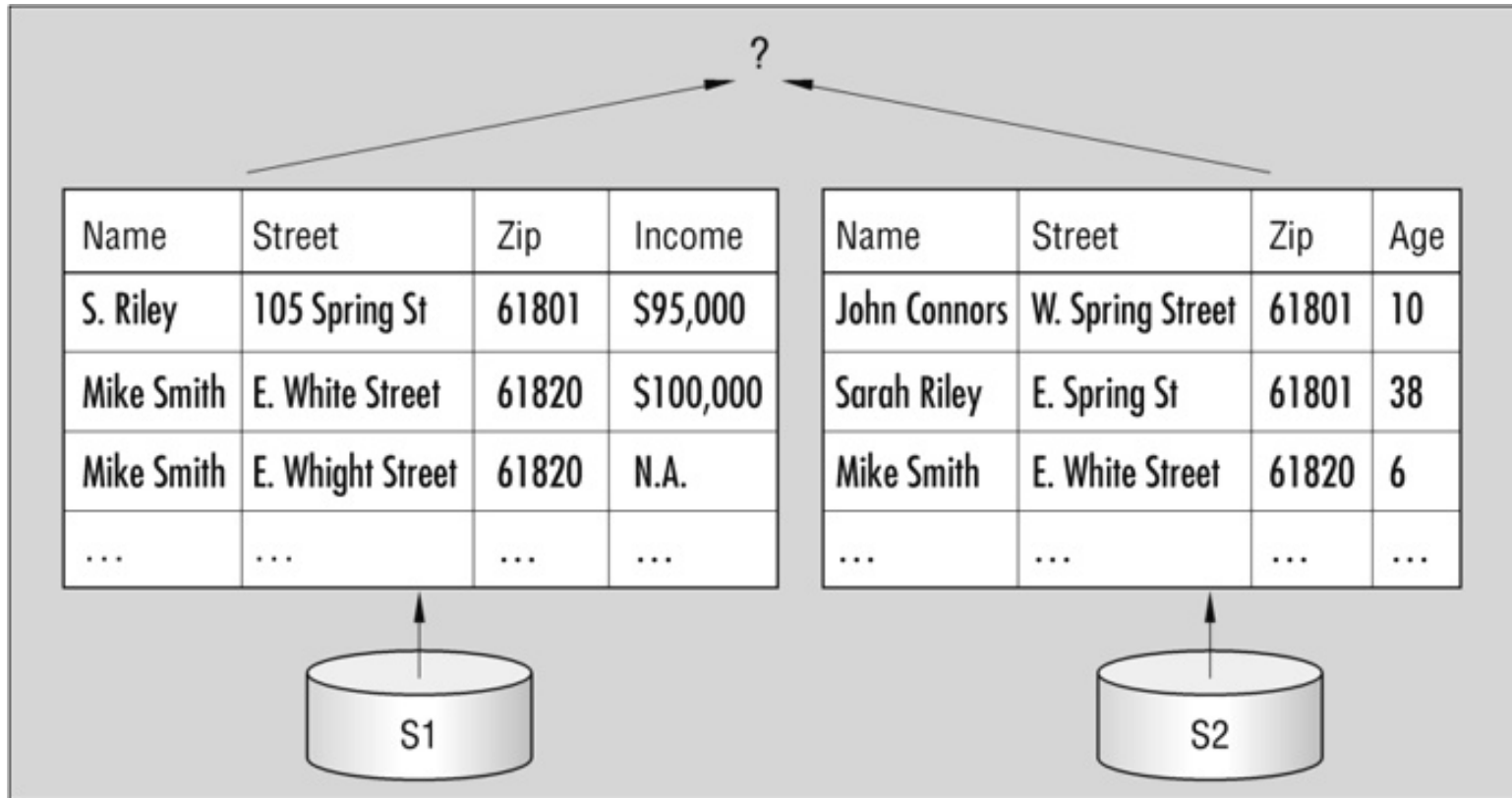
Topic 04

Main Sources

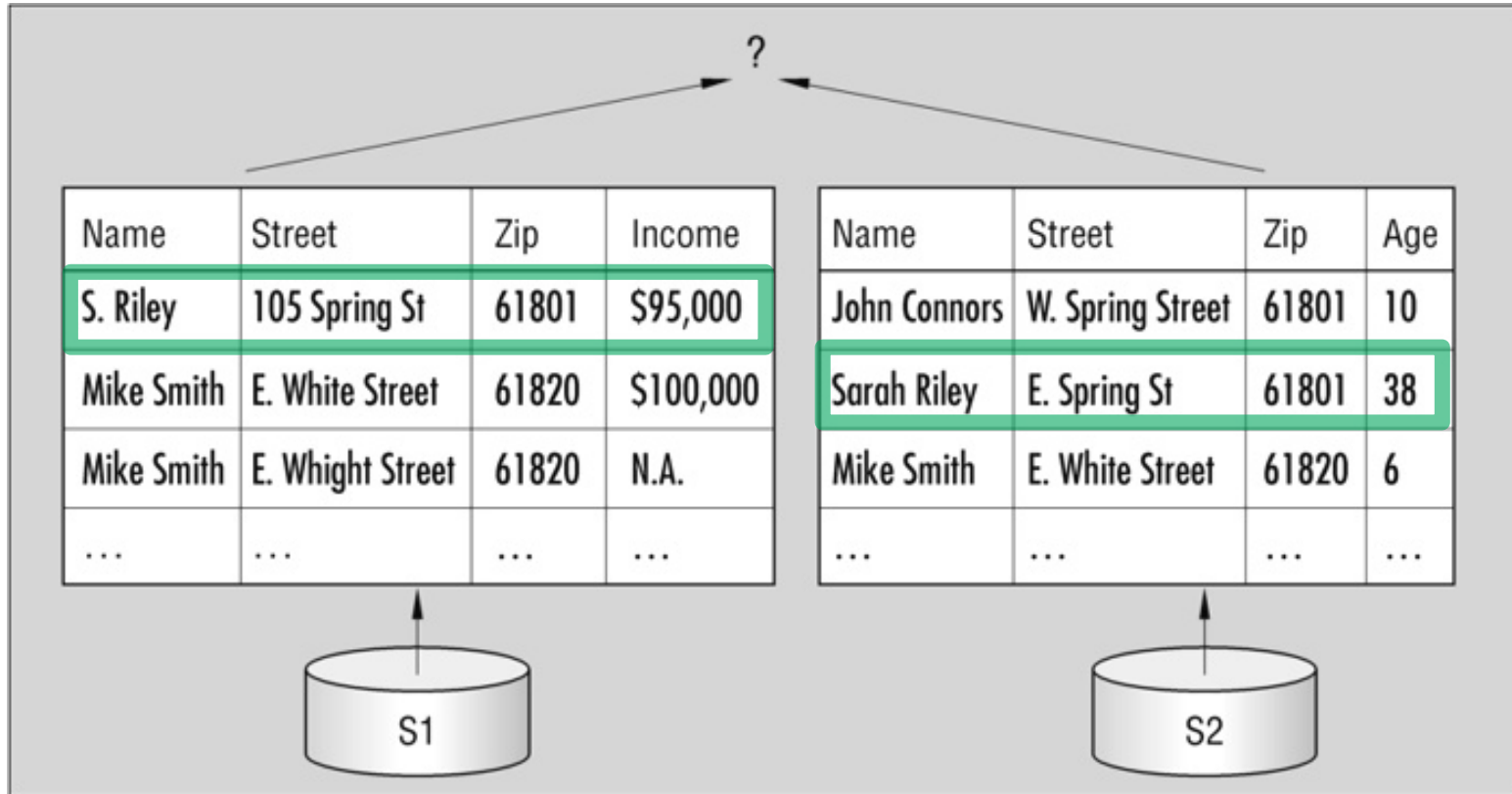
- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 2) + [slides by Lijun Zhang](#)
- Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Chapter 3)

Data integration

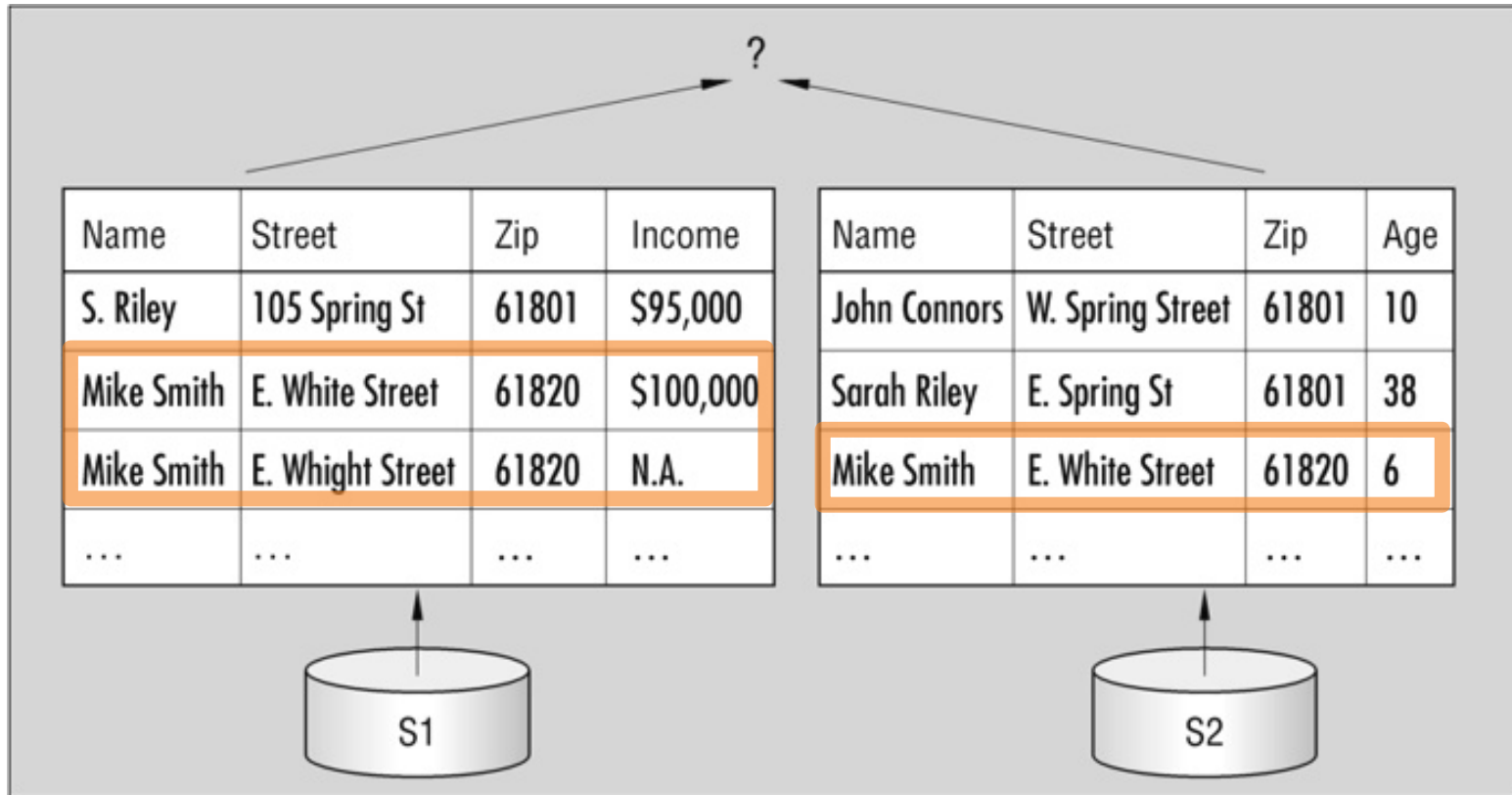
Data integration is not easy



Data integration is not easy



Data integration is not easy



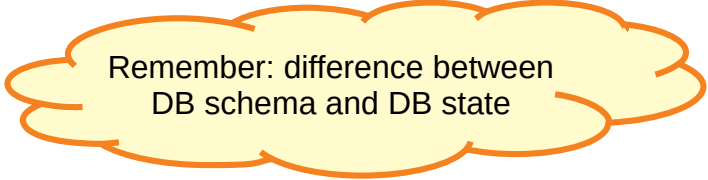
Data integration aspects

- Schema integration

- Bring different schemata together
- Equal concepts should be represented with equal types

- Object **matching** / Entity identification

- Equal entities should be equally identified across datasets (unless re-identification forbidden by policy)



Remember: difference between
DB schema and DB state

Data integration aspects (cont.)

- **Redundancy** analysis
 - Sometimes data needs to be integrated because different sets are row-incomplete
 - Sometimes those sets don't form a partition \Rightarrow there will be **repeated entities to be removed**
- Resolution of **value conflicts**
 - Same entity, different attribute values

Data cleaning

Why data cleaning?

- Data collection technologies are inaccurate
 - Sensors
 - Optical character recognition
 - Speech-to-text data
- Privacy reasons
- Manual errors
- Data collection is expensive and inaccurate

What is data cleaning?

It is a process by which data records are
modified or deleted
until each record passes
data validity criteria

Data validity criteria (1)

- **Mandatory** constraints: certain columns cannot be empty.
- **Data-type** constraints: values in a column must be of a particular datatype
- **Range** constraints: numbers or dates should fall within a certain range
- **Regular expression** patterns: e.g., phone numbers `[0-9]{9}`

Data validity criteria (2)

- **Unique** constraints: a field, or a combination of fields, must be unique
- **Set-membership** constraints: values in a column come from a set of discrete values or codes
- **Foreign-key** constraints: set membership constraint where valid values in a column are defined in a column of another table that contains unique values

Data validity criteria (3)

- **Cross-field validation**: certain conditions that utilize multiple fields must hold, e.g.:
 - percentages add up to 1.0 or to 100
 - discounted price lower or equal to regular price
 - date of expiration after date of manufacturing

Data validity criteria (3)

- **Cross-field validation:** certain conditions that utilize multiple fields must hold, e.g.:
 - percentages add up to 1.0 or to 100
 - discounted price lower or equal to regular price
 - **date of expiration after date of manufacturing**
(useful when traveling!)

生产日期: 2016 年 06 月 01 日
▶ 保质期至: 2018 年 06 月 01 日

6/05/2015 تاريخ التعبئة

13/07/2015 تاريخ انتهاء الصلاحية

賞味期限17. 9.11
製造日17. 5.11

Handling missing entries

Why is a value missing?

- **Missing completely at random (MCAR)**
 - Missingness of a value is independent of attributes
 - Fill in values based on the attribute
 - Analysis may be unbiased overall
- **Missing at Random (MAR)**
 - Missingness is related to other variables
 - Fill in values based other values
 - Almost always produces a bias in the analysis
- **Missing Not at Random (MNAR)**
 - Missingness is related to unobserved measurements
 - Informative or non-ignorable missingness
- In general, it is not possible to know the situation just from the data

Handling missing entries: options

- **Delete** the data record containing missing entries
- **Estimate** or **Impute** the Missing Values
 - Additional errors may be introduced
 - Good under certain conditions (e.g., Matrix Completion)
- Some algorithms can work with missing data

Exercise: handling missing data (specify your assumptions)

Answer in
Nearpod Collaborate
Code to be given in class

- Q1. 5% of student records at a university have no “civil status” (single, married, ...)
Drop records? Impute value, how?
- Q2. 5% of smokers in a study of the effects of tobacco on health had no year of birth
Drop records? Impute value, how?
- Q3. 5% of records of sales of a company have zip code but no province
Drop records? Impute value, how?
- Q4. Temperature sensor at weather station was failing at random intervals during one day, total downtime 6 hours, max continuous downtime 15 minutes
Drop that day? Impute values, how?
- Q5. Same sensor failed during one night, downtime 6 hours continuous
Drop that day? Impute values, how?

Handling Incorrect and Inconsistent Entries

- Inconsistency detection
 - E.g., full name and abbreviation don't match
- Domain knowledge
 - Human age cannot reach to 800 (yet?)
- Data-centric methods
 - Outlier detection

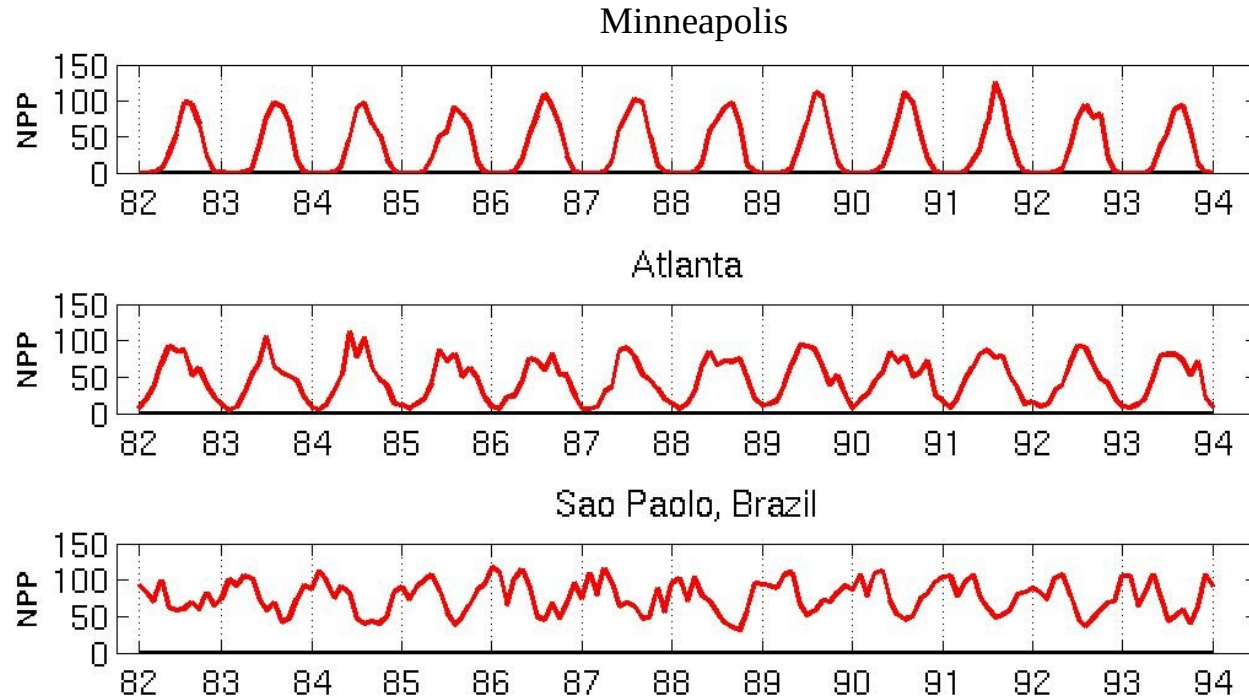
Scaling and normalization

- Features have different **scales**
 - Age versus Salary
- **Standardization** (“z-scoring”)
 - Mean 0 and stdev 1
- **Min-Max Scaling**
 - Map to [0,1]
 - Sensitive to noise

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

$$z_i = \frac{x_i - \min}{\max - \min}$$

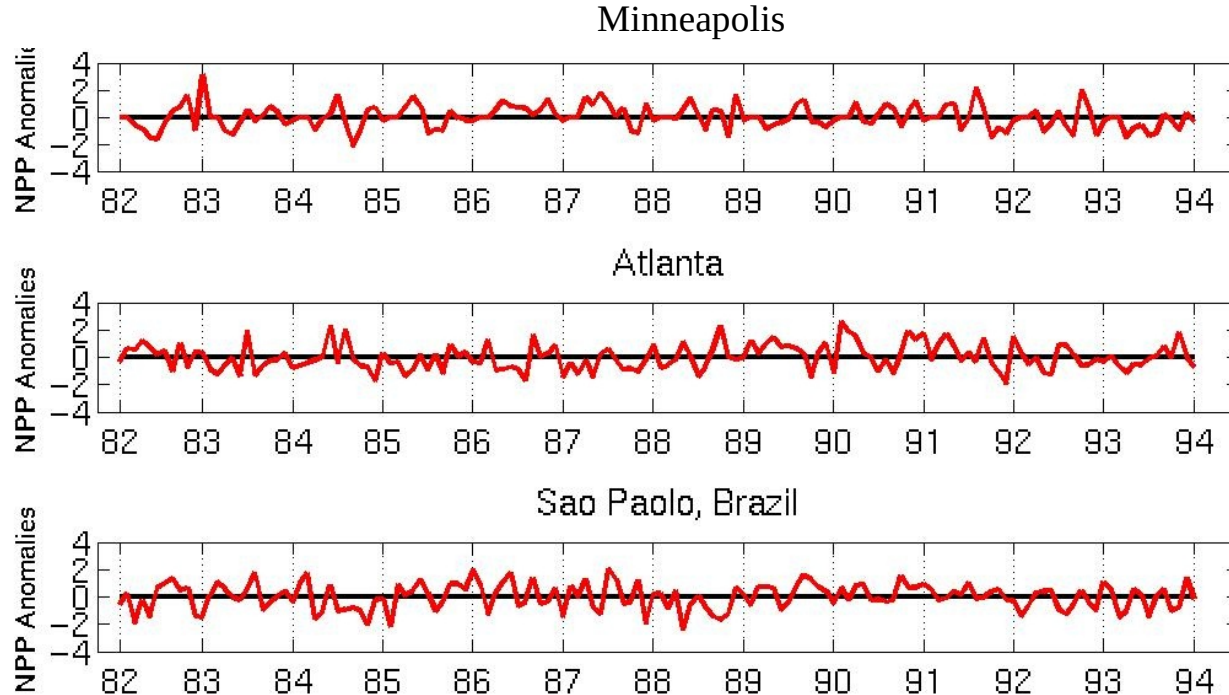
Example: seasonal standardization



Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000

Example: seasonal standardization



Normalized using
monthly Z Score:

Subtract off monthly
mean and divide by
monthly standard
deviation

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	1.0000

Adjusted correlations between time series

Summary

Things to remember

- Data cleaning
 - Specially: when and how to impute missing values

Exercises for TT03-TT05

- Exercises 3.7 of Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al.
- Exercises 2.6 of Introduction to Data Mining, Second Edition (2019) by Tan et al.
 - Mostly the first exercises, say 1-6