Data Preparation: *Integration and Cleaning*

Mining Massive Datasets

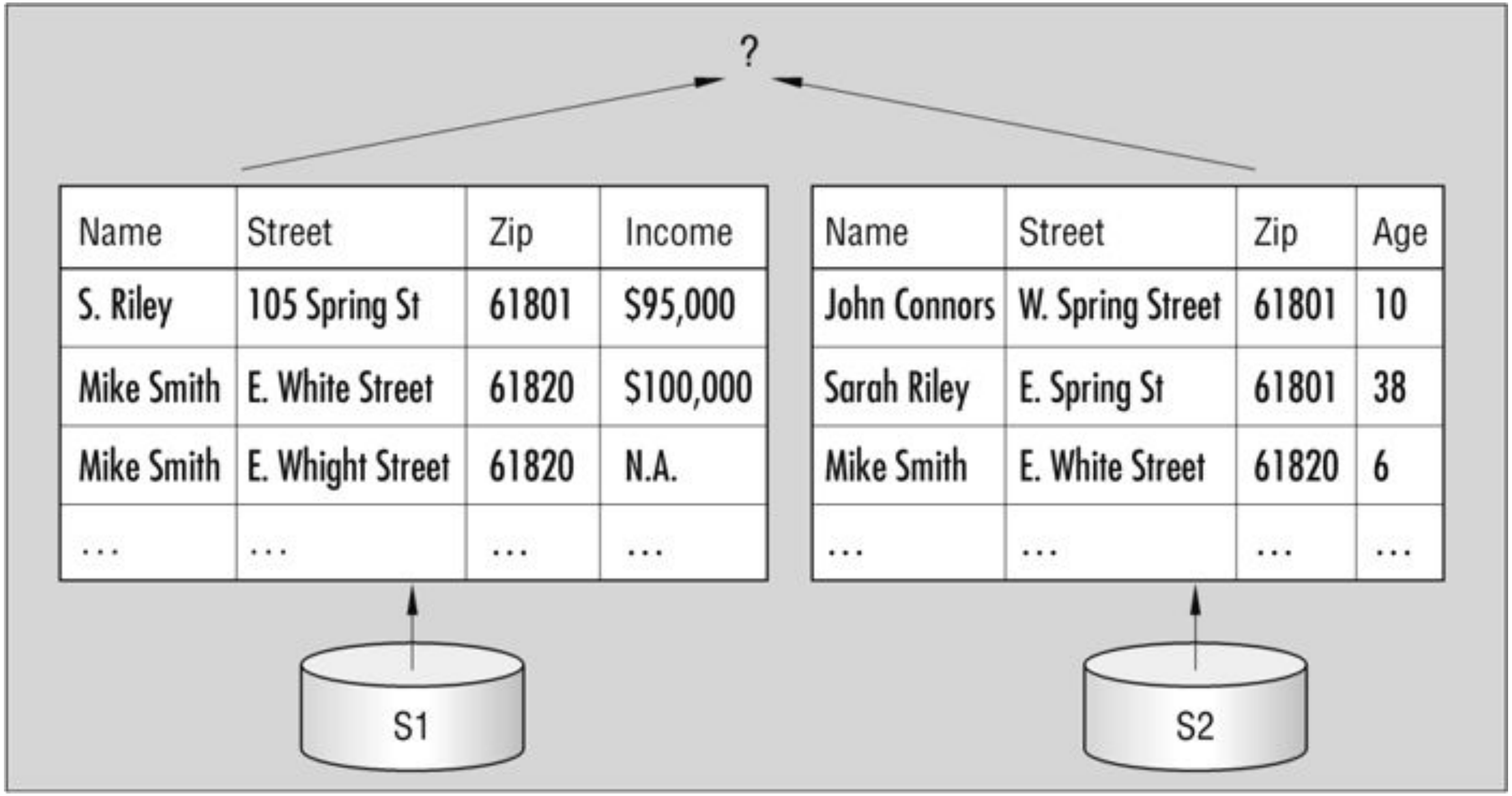
Materials provided by Prof. Carlos Castillo — https://chato.cl/teach
Instructor: Dr. Teodora Sandra Buda — https://tbuda.github.io/

Main Sources

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

Data integration

Data integration is not easy

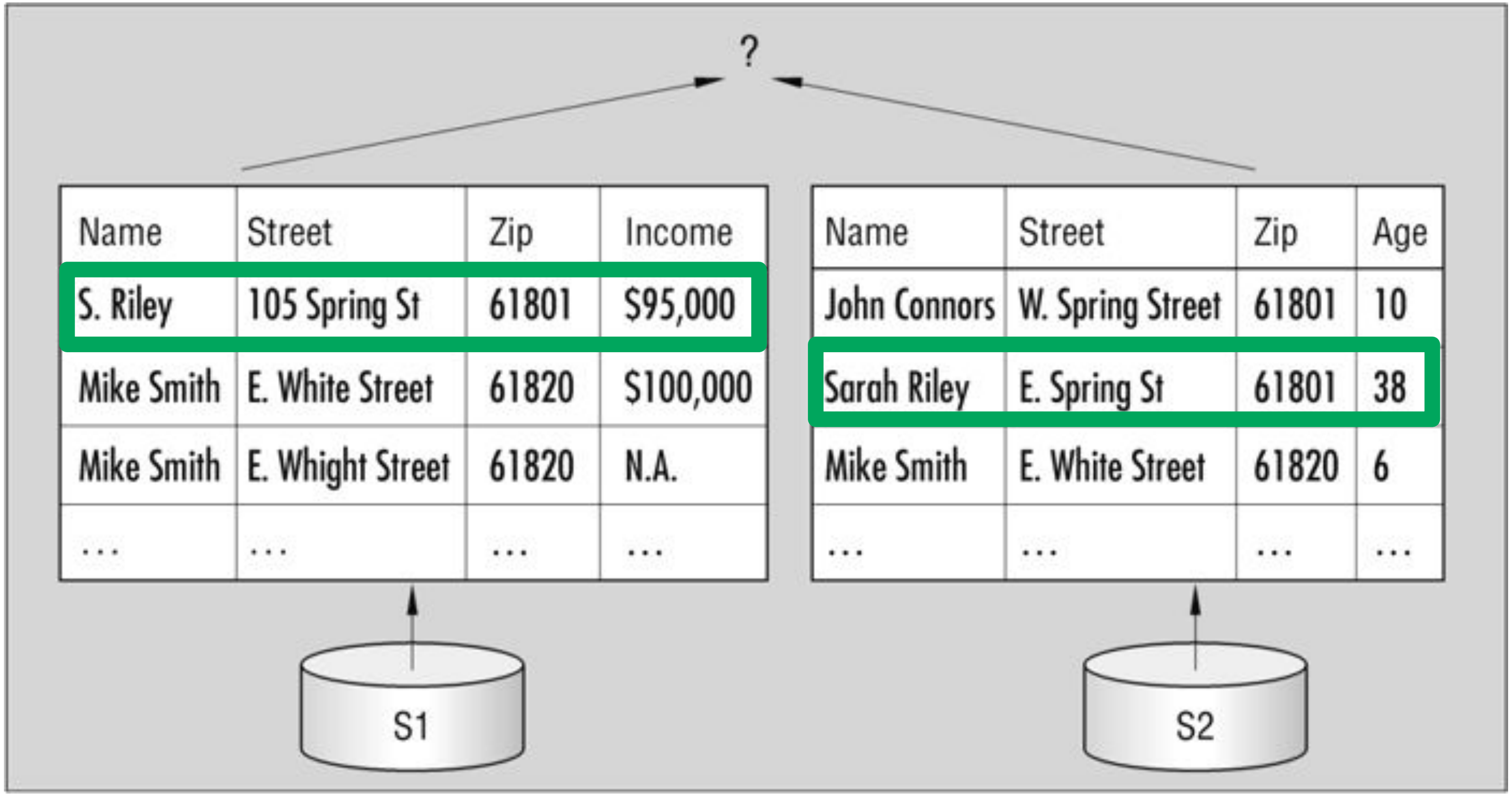


Name	Street	Zip	Income
S. Riley	105 Spring St	61801	$95,000
Mike Smith	E. White Street	61820	$100,000
Mike Smith	E. Whight Street	61820	N.A.
...	...	...	...

Name	Street	Zip	Age
John Connors	W. Spring Street	61801	10
Sarah Riley	E. Spring St	61801	38
Mike Smith	E. White Street	61820	6
...	...	...	...

[Lu et al. 2013]

Data integration is not easy

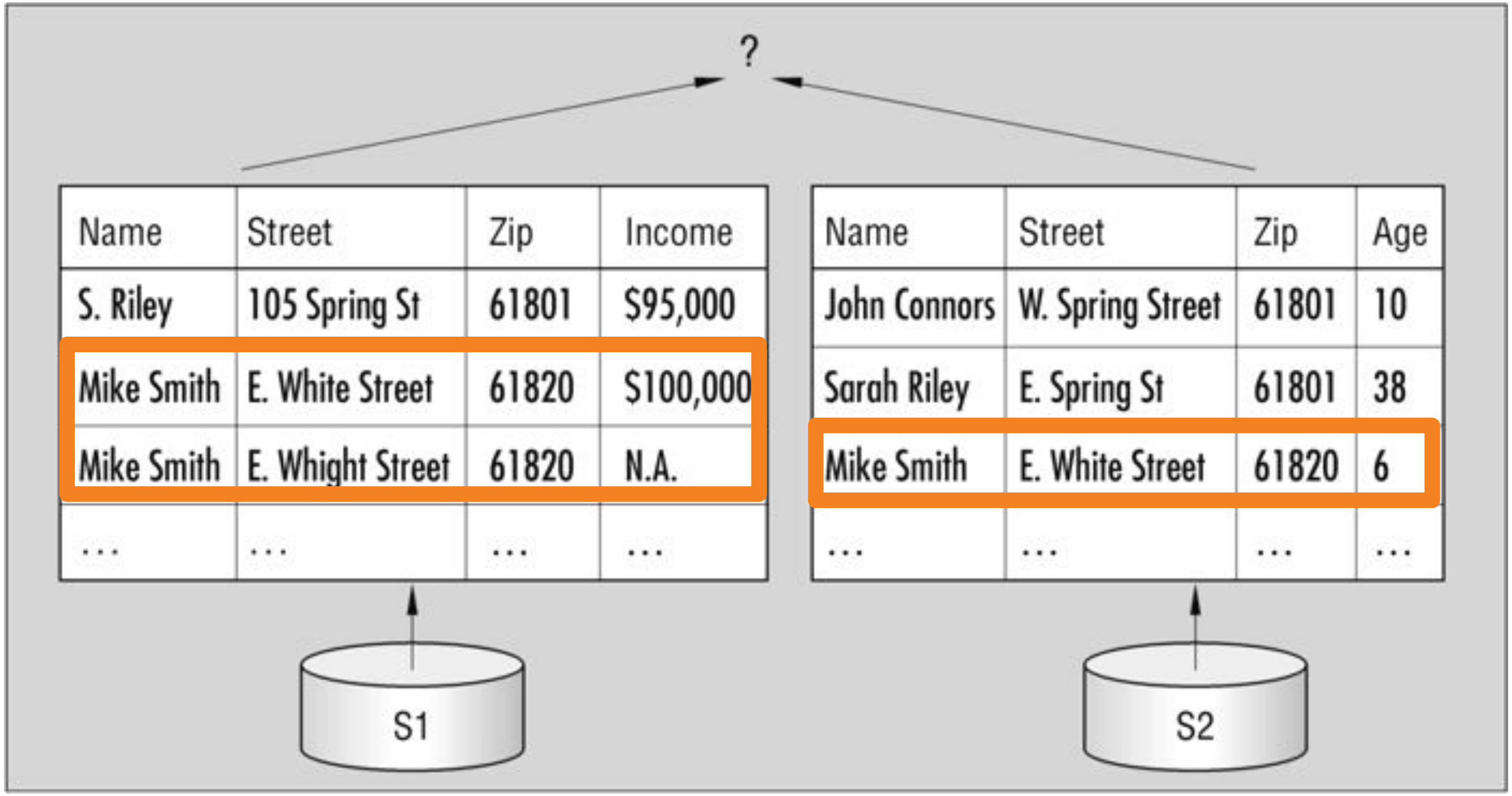


Name	Street	Zip	Income
S. Riley	105 Spring St	61801	$95,000
Mike Smith	E. White Street	61820	$100,000
Mike Smith	E. Whight Street	61820	N.A.
...	...	...	...

Name	Street	Zip	Age
John Connors	W. Spring Street	61801	10
Sarah Riley	E. Spring St	61801	38
Mike Smith	E. White Street	61820	6
...	...	...	...

[Lu et al. 2013]

Data integration is not easy



Lu et al. 2013

Data integration aspects

Remember: difference between DB schema and DB state

- **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)

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 ⇒ 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)

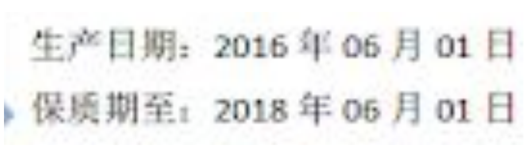
- **Unique** constraints: a field, or a combination of fields, must be unique
- **Set-membership** constraints: values in a column come from a set of discrete values or codes
- **Foreign-key** constraints: set membership constraint where valid values in a column are defined in a column of another table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

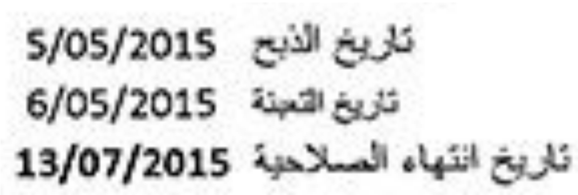
Data validity criteria (3)

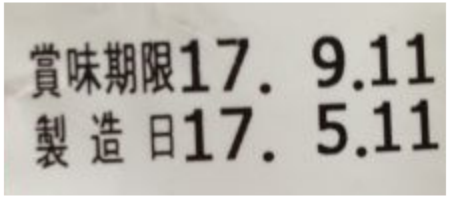
- **Cross-field validation**: certain conditions that utilize multiple fields must hold, e.g.:
  - percentages add up to 1.0 or to 100
  - discounted price lower or equal to regular price
  - date of expiration after date of manufacturing

Data validity criteria (3 cont.)

You see this in a package … how do you decide whether the product is expired or not?







Handling missing entries
Why is a value missing?

- **Missing Completely at Random (MCAR)**
  - Missingness of a value is independent of observable attributes
- **Missing at Random (MAR)**
  - Missingness has statistical dependencies with an observable attribute
  - We can fill in values based on other attributes, but this is likely to introduce a bias in the analysis
- **Missing Not at Random (MNAR)**
  - Missingness depends deterministically on an observable attribute
  - In general this is informative, non-ignorable missingness

In general, it is **not** possible to know which one is the case just by looking at 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)

1 5% of student records at a university have no “civil status” (single, married, ...)

- Drop records? Impute value, how?

2 5% of smokers in a study of the effects of tobacco on health had no year of birth

- Drop records? Impute value, how?

3 5% of records of sales of a company have zip code but no province

- Drop records? Impute value, how?

4 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?

5 Same sensor failed during one night, downtime 6 hours continuous

- Drop that day? Impute values, how?

Possible answers (correctness depends on assumptions)

- 5% of student records at a university have no “civil status” (single, married, ...)
  - Undergrads? Impute as “single” unless there is a “spouse” field or similar
- 5% of smokers in a study of the effects of tobacco on health had no year of birth
  - Drop, but check if there is something systematic in distribution of other values for them
- 5% of records of sales of a company have zip code but no province
  - Get a zip code to province table, complete the missing data
- Temperature sensor at weather station was failing at random intervals during one day, total downtime 6 hours, max continuous downtime 15 minutes
  - Impute by interpolating
- Same sensor failed during one night, downtime 6 hours continuous
  - Drop day, interpolation may be inaccurate

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

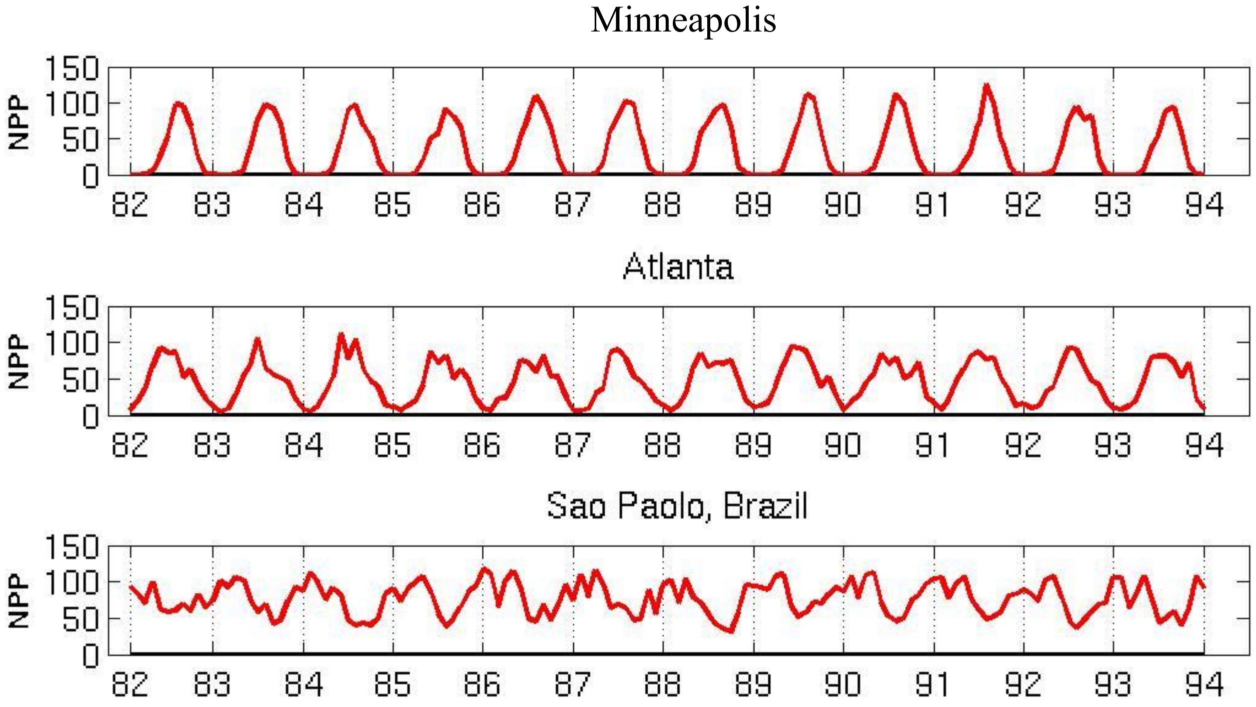
- Features have different scales
  - Age versus Salary
- Standardization ("z-scoring")
  - Mean 0 and stdev 1

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

- Min-Max Scaling
  - Map to [0,1]
  - Sensitive to noise

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

Example: seasonal standardization

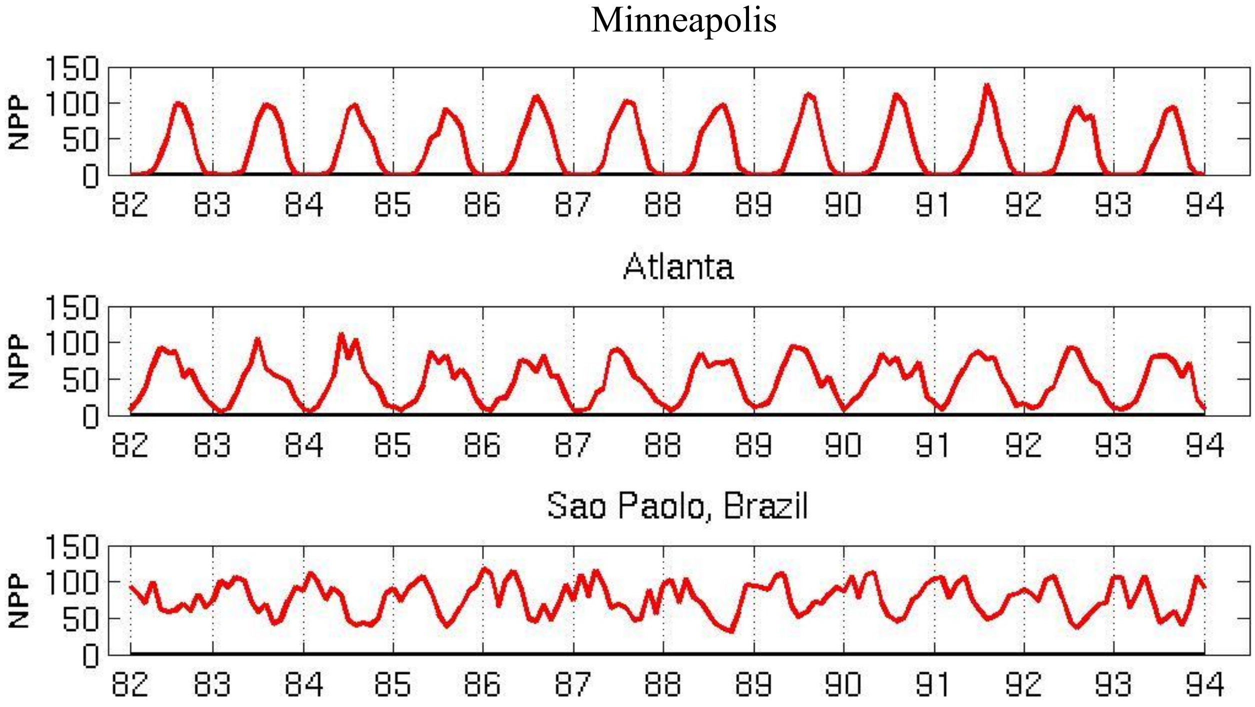


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

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

Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)

Example: seasonal standardization



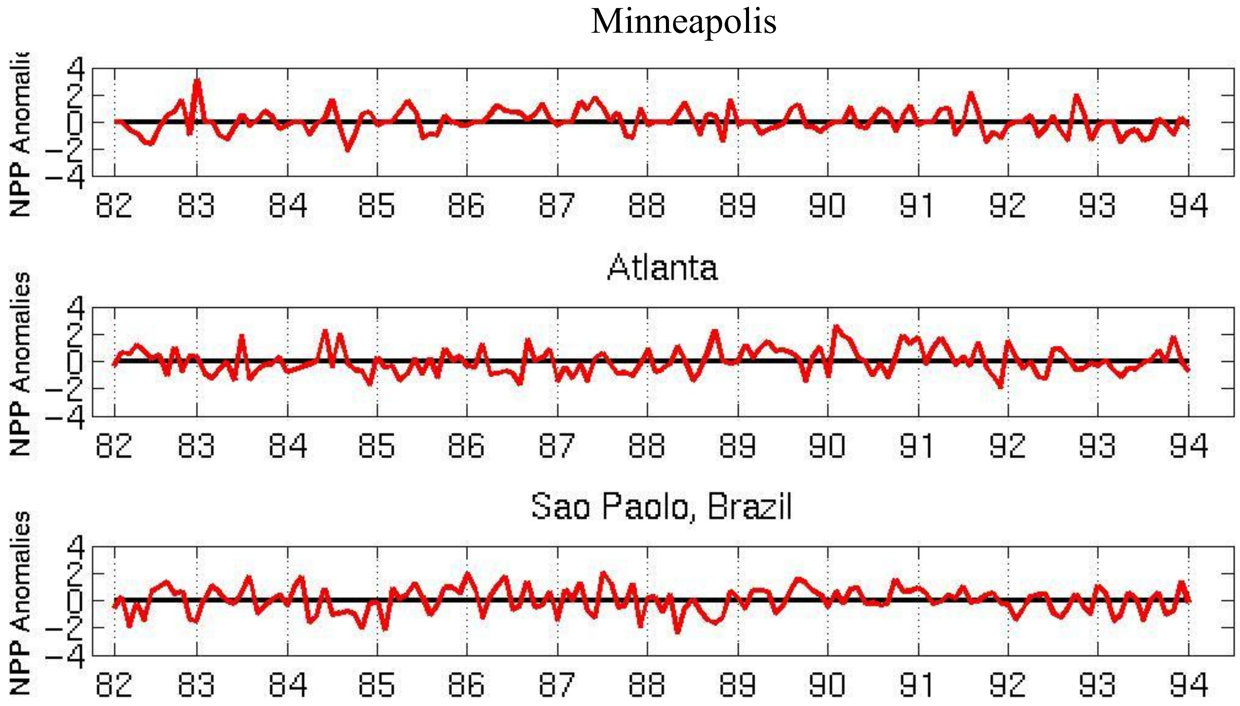
Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

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

Spurious correlations between time series

Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)

Example: seasonal standardization



Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

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

Adjusted correlations between time series

Introduction to Data Mining 2nd edition (2019) by Tan et al. (Chapter 2)

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