Data

Data Mining: Seminar 2

Dr. Thomas Price

AI Academy

NC STATE UNIVERSITY

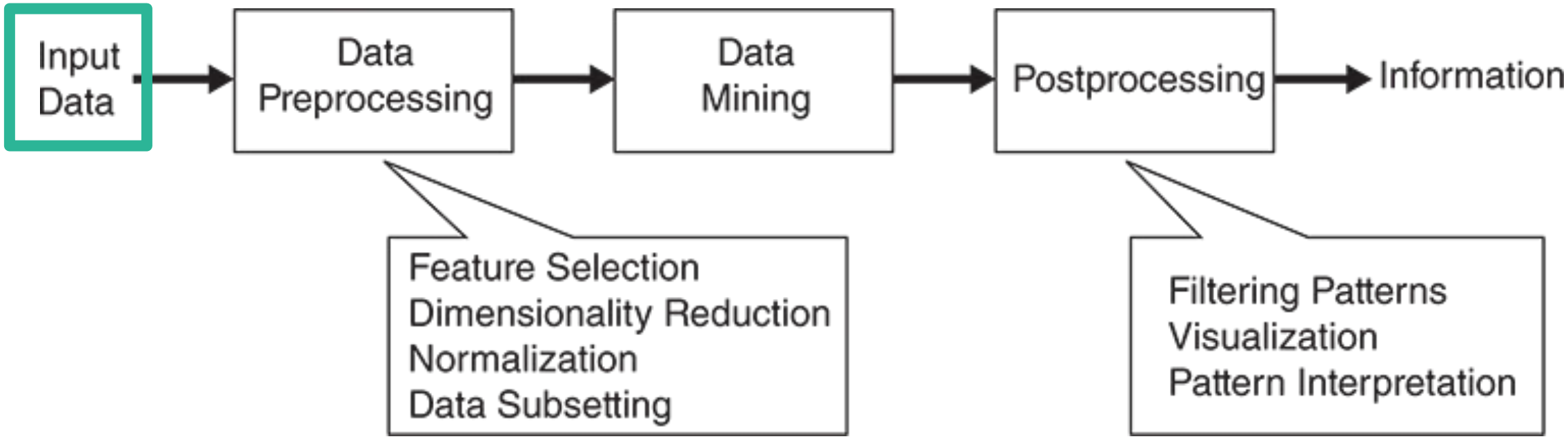
Data and Data Types

Lesson

AI Academy

What is Data Mining?

The Data Mining Pipeline



What is Data?

Collection of **data objects** and their **attributes:**

- **Attribute**: A property or characteristic of an object
- **Object:** A collection of attributes (also known as **rows**)

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

What is Data?

ATTRIBUTES:

- Variables
- Fields
- Characteristics
- Features

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

What is Data?

OBJECTS:

- Rows
- Records
- Data Points
- Cases
- Samples
- Entities
- Instances

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

Numbers or symbols assigned to an attribute.

Tid	Refund
1	Yes
2	No
3	No
4	Yes
5	No
6	No
7	Yes
8	No
9	No
10	No

Attributes vs. Attribute Values

The same attribute can be mapped to different attribute values.

- **Example**: Height can be measured in feet or meters

Different attributes can be mapped to the same set of values.

- **Example**: Attribute values for ID and age are integers
- But properties of attribute values can be different

Properties of Attribute Values

The **type** of an attribute depends on the **properties** it possesses:

- Distinctiveness {=, ≠}: Can values be **distinguished**?
- Order {<, >}: Can values be meaningfully **ordered**?
- Addition {+, -}: Are **differences** between values meaningful?
- Multiplication {*, /}: Are **ratios** between values meaningful?

Types of Attributes

There are four different types of attributes (**NOIR**)

- **Nominal**
- **Ordinal**
- **Interval**
- **Ratio**

Nominal Attributes

- The values are just different **names**
- Provide only enough information to distinguish objects

Examples

- Zip Codes
- Employee ID Numbers
- Favorite Colors

Distinctiveness {=, ≠}

Order {<, >}

Addition {+, -}

Multiplication {*, /}

Ordinal Attributes

The values provide enough information to **order** objects

Examples

- Mineral Hardness
- *{good, better, best}*
- Grades
- Floor Number
- Rankings
- Height: *{tall, medium, short}*

Distinctiveness {=, ≠}

Order {<, >}

~~Addition {+, -}~~

~~Multiplication {*, /}~~

Interval Attributes

Differences between values are meaningful (a unit of measurement exists)

Examples

- Calendar Dates
- Temperature in Celsius or Fahrenheit

Distinctiveness {=, ≠}
Order {<, >}
Addition {+, -}
Multiplication {*, /}

Ratio Attributes

- **Differences and ratios** are meaningful
- **Meaningful 0-value** indicating "none"

  You can have 0 height? 0 age?

Examples

- Counts
- Age
- Length
- Monetary Quantities

Distinctiveness {=, ≠}
Order {<, >}
Addition {+, -}
Multiplication {*, /}

Properties of Data Types Summary

	Nominal	Ordinal	Interval	Ratio
Distinctness	X	X	X	X
Order		X	X	X
Addition			X	X
Multiplication				X

Discrete and Continuous Attributes

Discrete

- Finite or countably infinite set of values
- Often represented as integer variables

Continuous

- Real numbers as attribute values
- Real values can only be measured and represented with finite number of digits
- Typically represented as floating-point variables

Discrete and Continuous Attributes

Discrete

- Zip Codes
- Counts
- Set of words in a collection of documents
- Binary attributes are a special case of discrete attributes

Continuous

- Temperature
- Height
- Weight

Learning Objectives: Data and Data Types

You now should be able to:

- Read tabular data (**attributes** and **data objects**)
- Distinguish between the primary **types of data**
  - Nominal, Ordinal, Interval, and Ratio
  - Discrete and Continuous

AI Academy
NC STATE

AI Academy

Data and Data Types

Exercises

AI Academy

Break the following attributes into 4 categories; 2 in each (3 minutes)

A. Student ID
B. Grades (A, A-, B+, B etc.)
C. Age
D. Favorite Color
E. Calendar Dates (08-25-2015; 08-06-1999)
F. Educational experience from 1-4, e.g.
    1. Elementary school graduate
    2. High-School graduate
    3. Some college
    4. College graduate
G. Temperature in Fahrenheit
H. Mass

What is the type of each attribute?

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

What is the type of each attribute?

Come up with a new **Ordinal attribute** and a new **Interval attribute** to add to the data.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

What is the type of each attribute?

A. Employee ID numbers
B. {good, better, best}
C. Calendar Dates
D. Temperature in Celsius or Fahrenheit
E. Temperature in Kelvin
F. Zip codes
G. Street numbers

Nominal
Ordinal
Interval
Ratio

Discrete or Continuous?

1. ZIP codes
2. Height
3. Employee ID
4. Counts
5. The set of words in a collection of documents
6. Temperature in Celsius or Fahrenheit
7. Weight