CHAPTER 2

Input: Concepts, Instances, and Attributes

Outline

- What’s a concept?
- What’s in an example?
- What’s in an attribute?
- Preparing the input

What’s a Concept? (1/2)

- Concept
  - Structural patterns
  - e.g.
    - Classify unseen examples
    - Find association among features
    - Group examples
    - Predict numeric outcome

數值的預測

What’s a Concept? (2/2)

- Concept description
  - models
  - e.g.
    - Decision trees
    - Rules
    - Regression functions
    - Clustering trees
    - Neural network

What's in an Example?

- Instances
- Input is generally expressed as a table of independent instances
  - Flat file
  - Records in DB

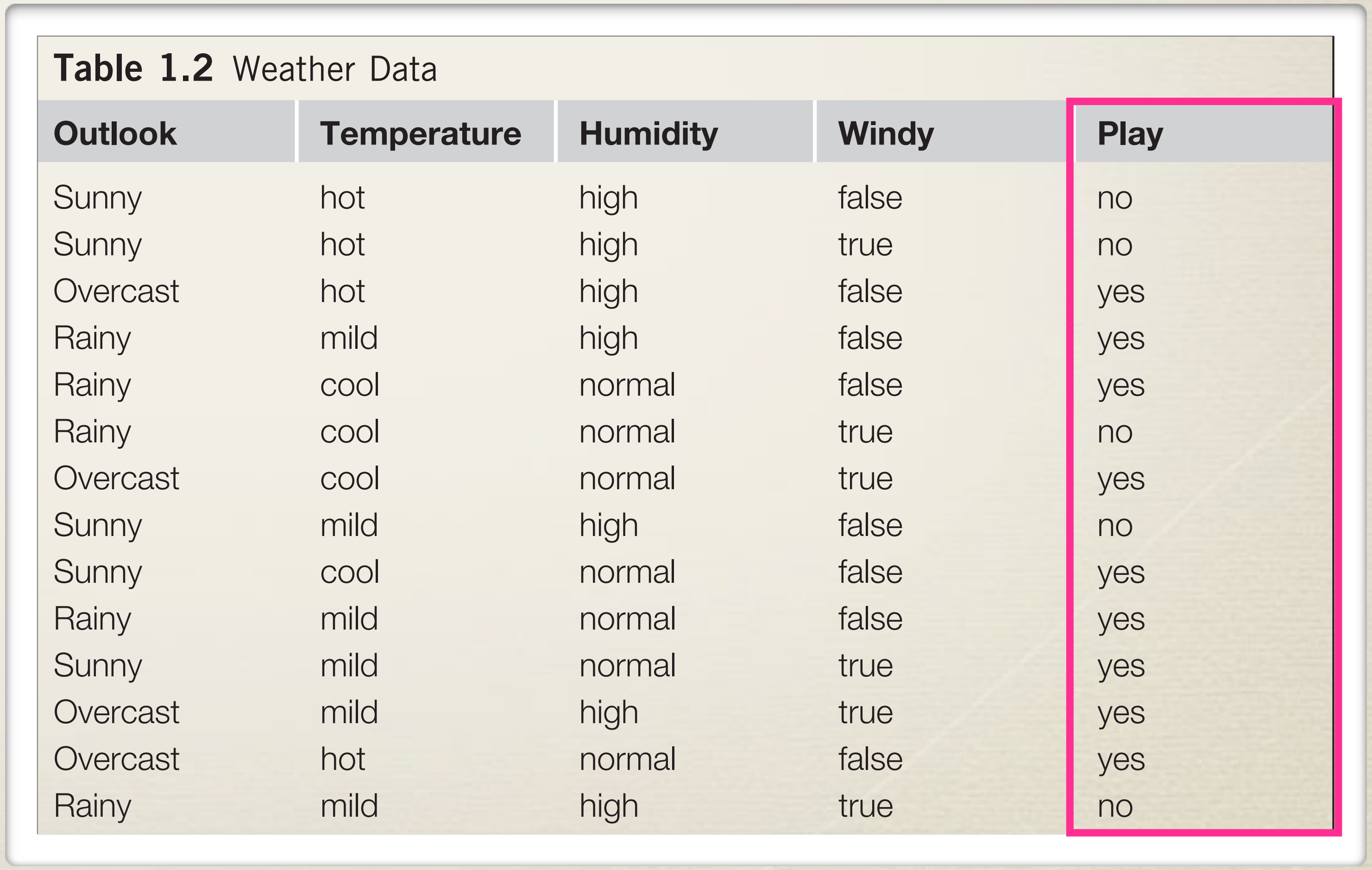


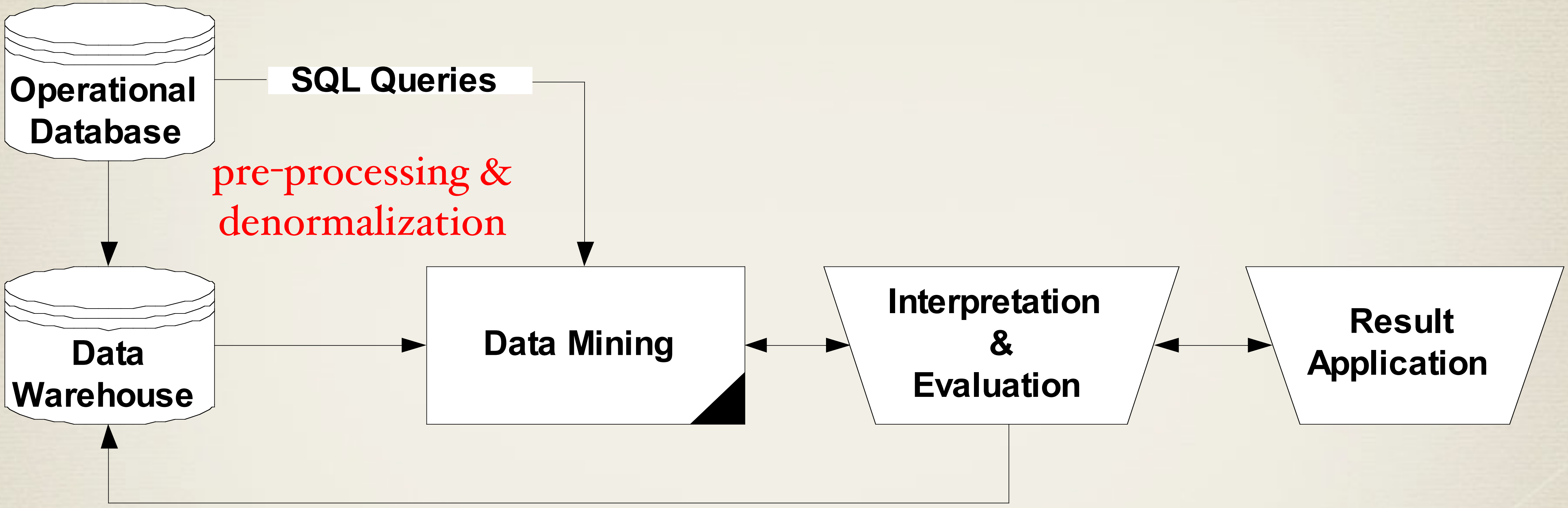
Table 1.2 Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

What’s in an Attribute?

- Fields in DB
- Values of attributes
  - Dichotomy (nominal or categorical)
    - e.g. true, false
  - No ordering or distance measure (nominal)
    - e.g. sunny, overcast, rainy
  - Ordinal (nominal)
    - e.g. hot > mild > cool
  - Interval (numeric)
    - e.g. temperature expressed in degree

Preparing the Input (1/7)



A simple data mining process model

SS

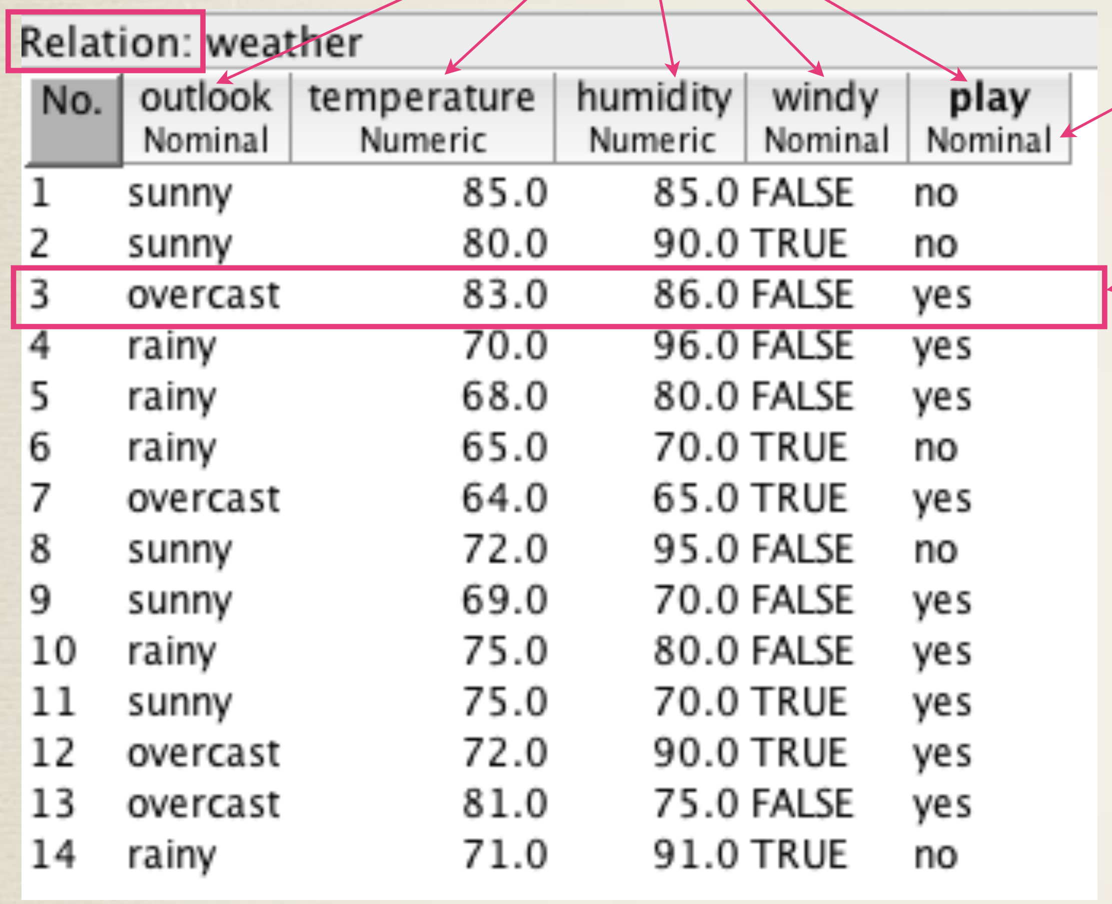
要先將資料整理成一個table
稱為資料前處理
(會花最多時間在這邊)

Preparing the Input (2/7)

attributes

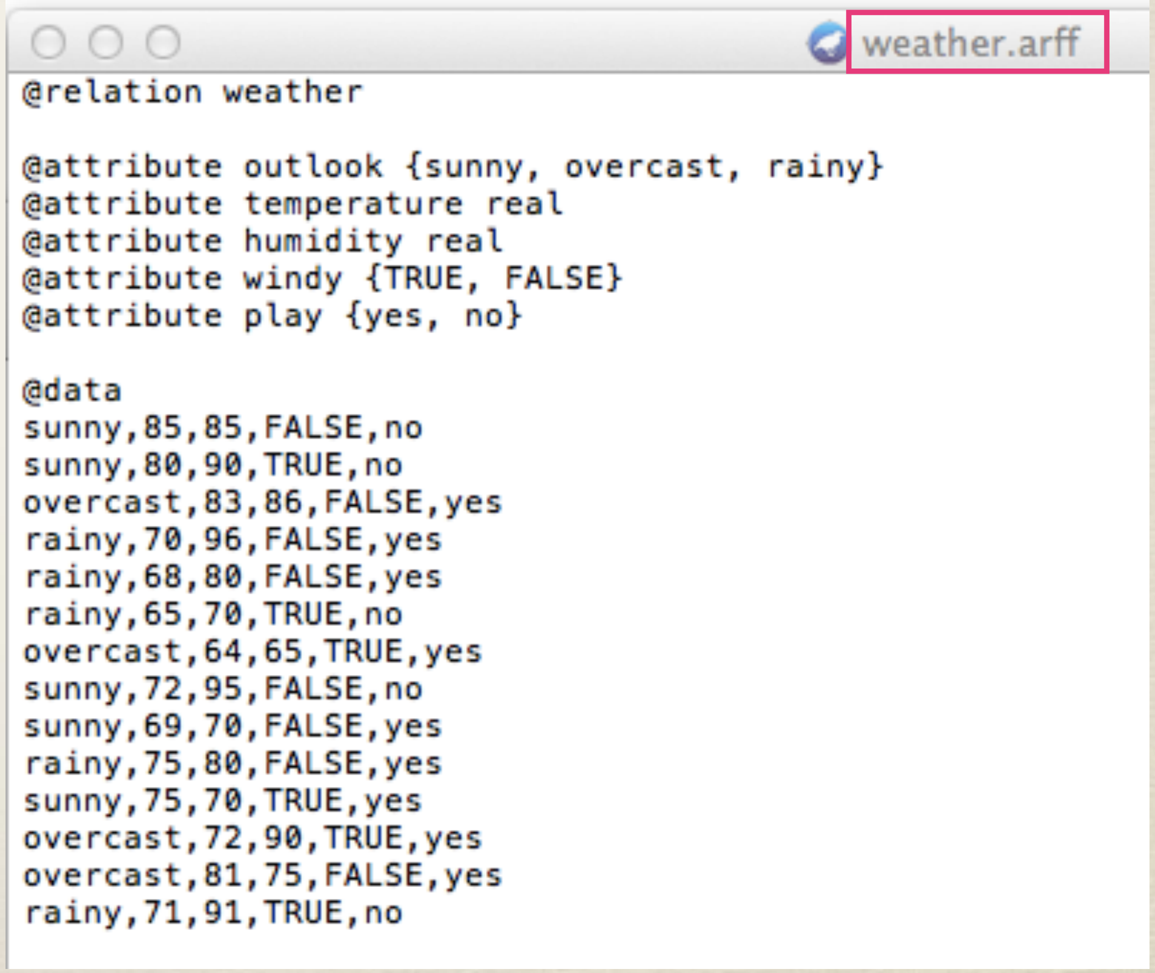
attribute’s type

instance



Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no



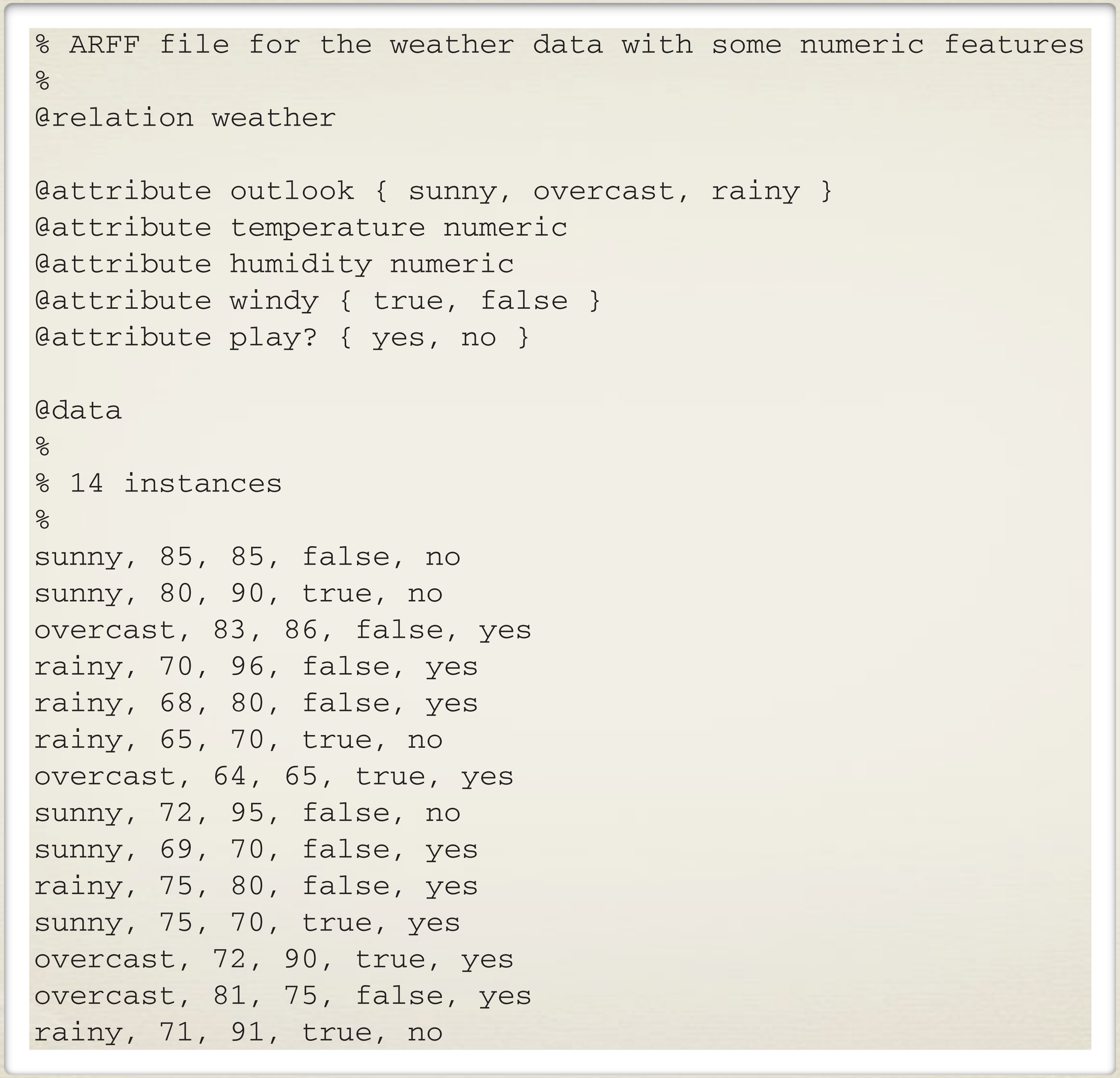
weather.arff

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Preparing the Input (3/7)



```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Preparing the Input (4/7)

- ARFF (Attribute-Relation File Format)
  - Attribute types
    - nominal
      - e.g. @attribute outlook {sunny, overcast, rainy}
    - numeric
      - e.g. @attribute temperature numeric
        @attribute temperature real
    - string
      - e.g. @attribute description string
    - date
      - e.g. @attribute today date
        2014-03-05T13:00:00

Preparing the Input (7/7)

- Missing value
  - e.g. @data

    sunny, 85, 85, false, ?
- Sparse value
  - e.g. 0, X, 0, 0, 0, 0, Y, 0, 0, 0, “class A”

    => {1 X, 6 Y, 10 “class A”}

    0, 0, 0, w, 0, 0, 0, 0, 0, 0, “class B”

    => {3 w, 10 “class B”}