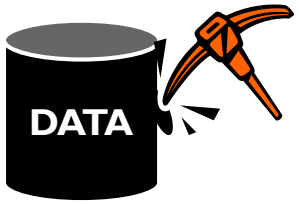
National College of Ireland



Data & Web Mining

2. Input and Output

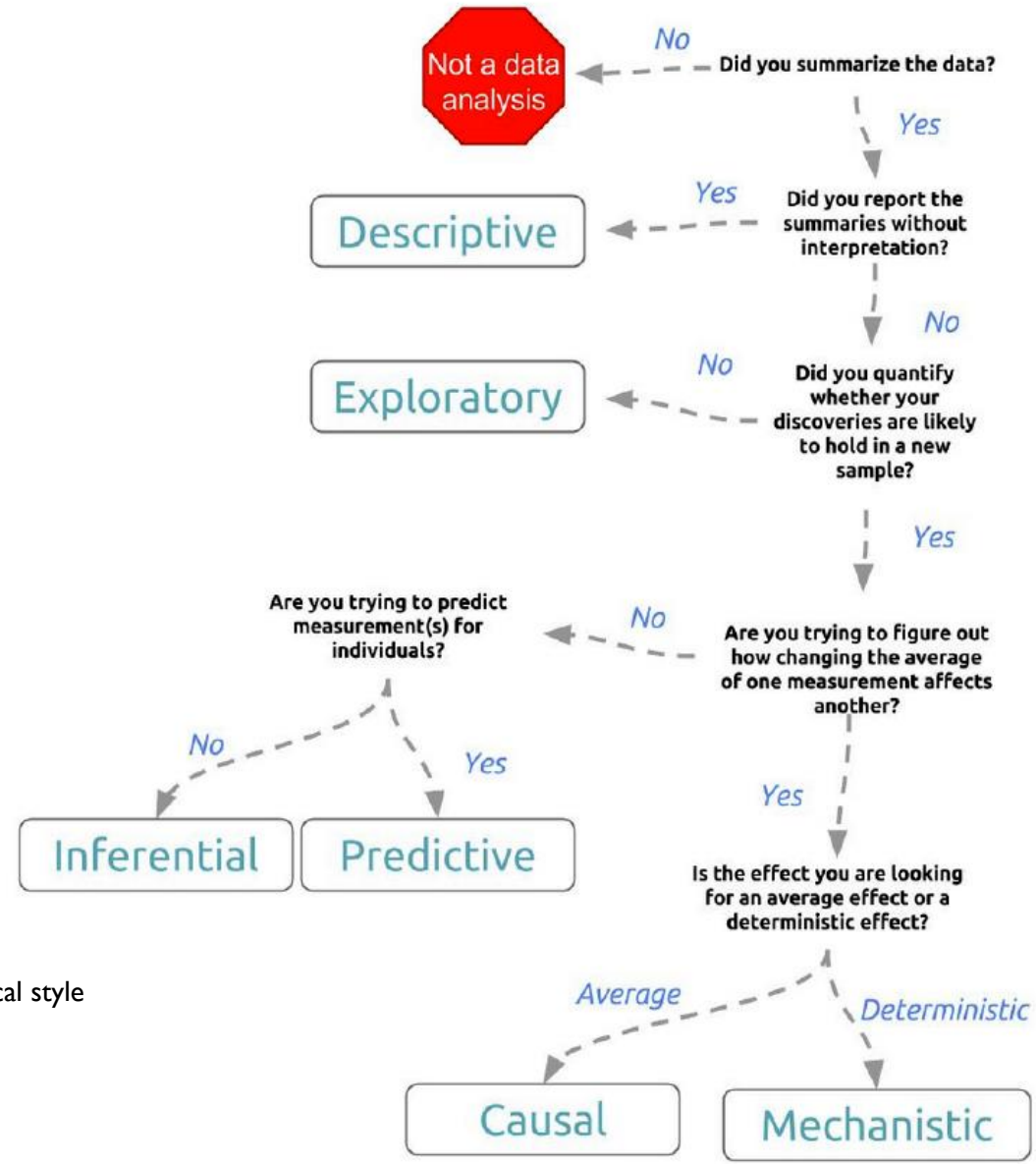
Dr. Jason Roche
Jason.roche@ncirl.ie

National College of Ireland

2. Input: Concepts, Instances, and Attributes

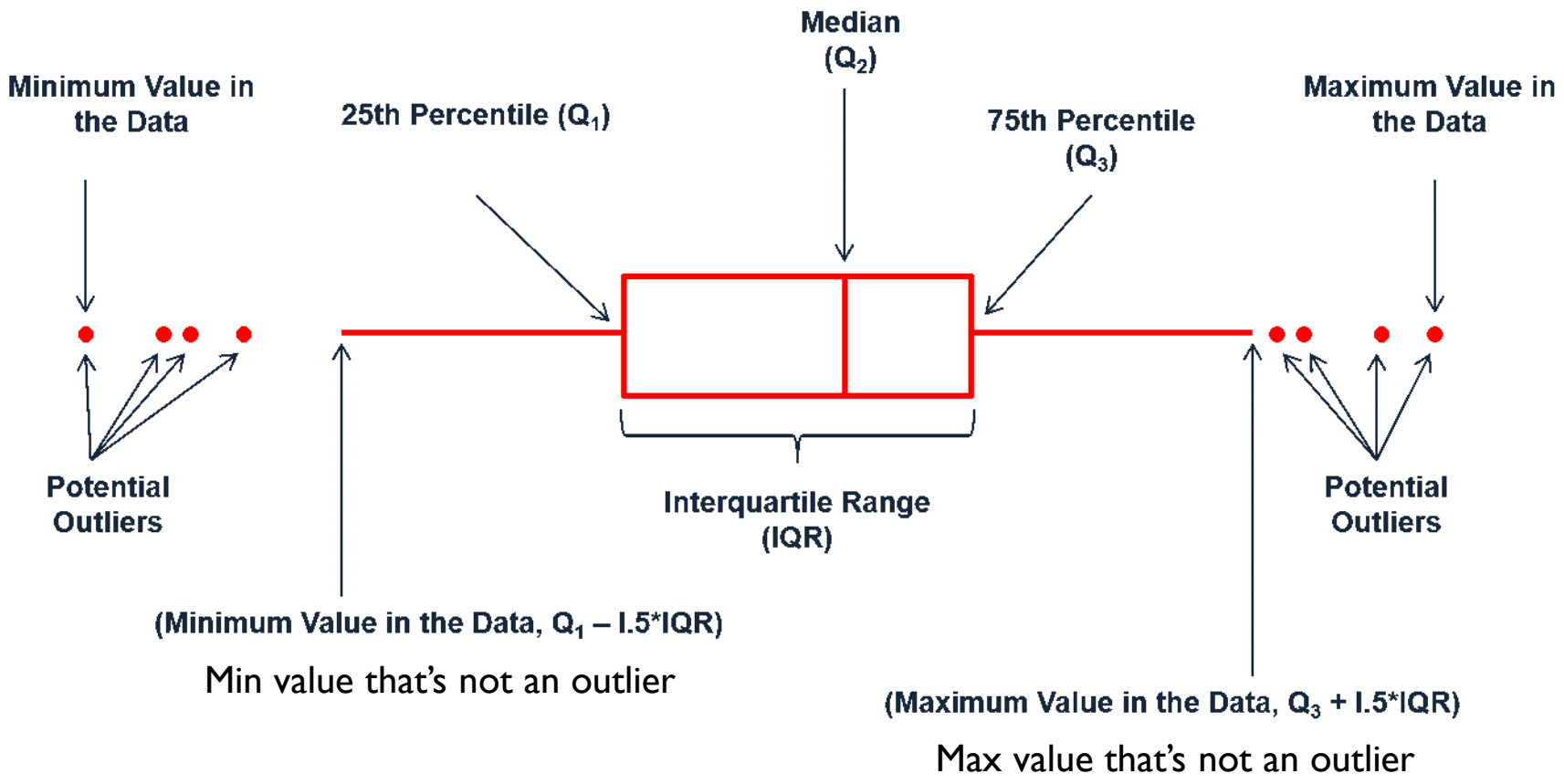
- Terminology
  - Analysis types?
  - What's a concept?
  - What's in an example?
  - What's in an attribute?
    - Nominal, ordinal, interval, ratio

What type of analysis are you doing ?

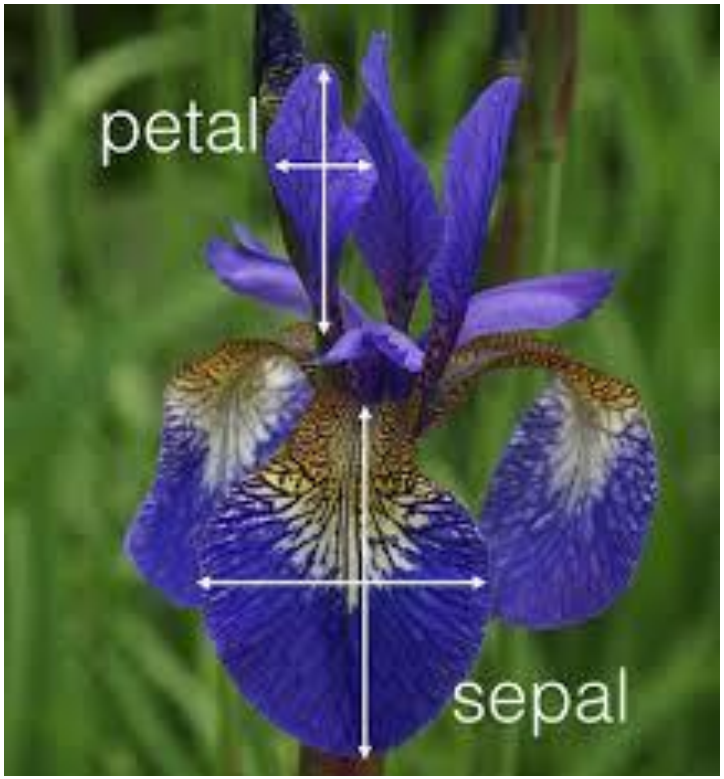


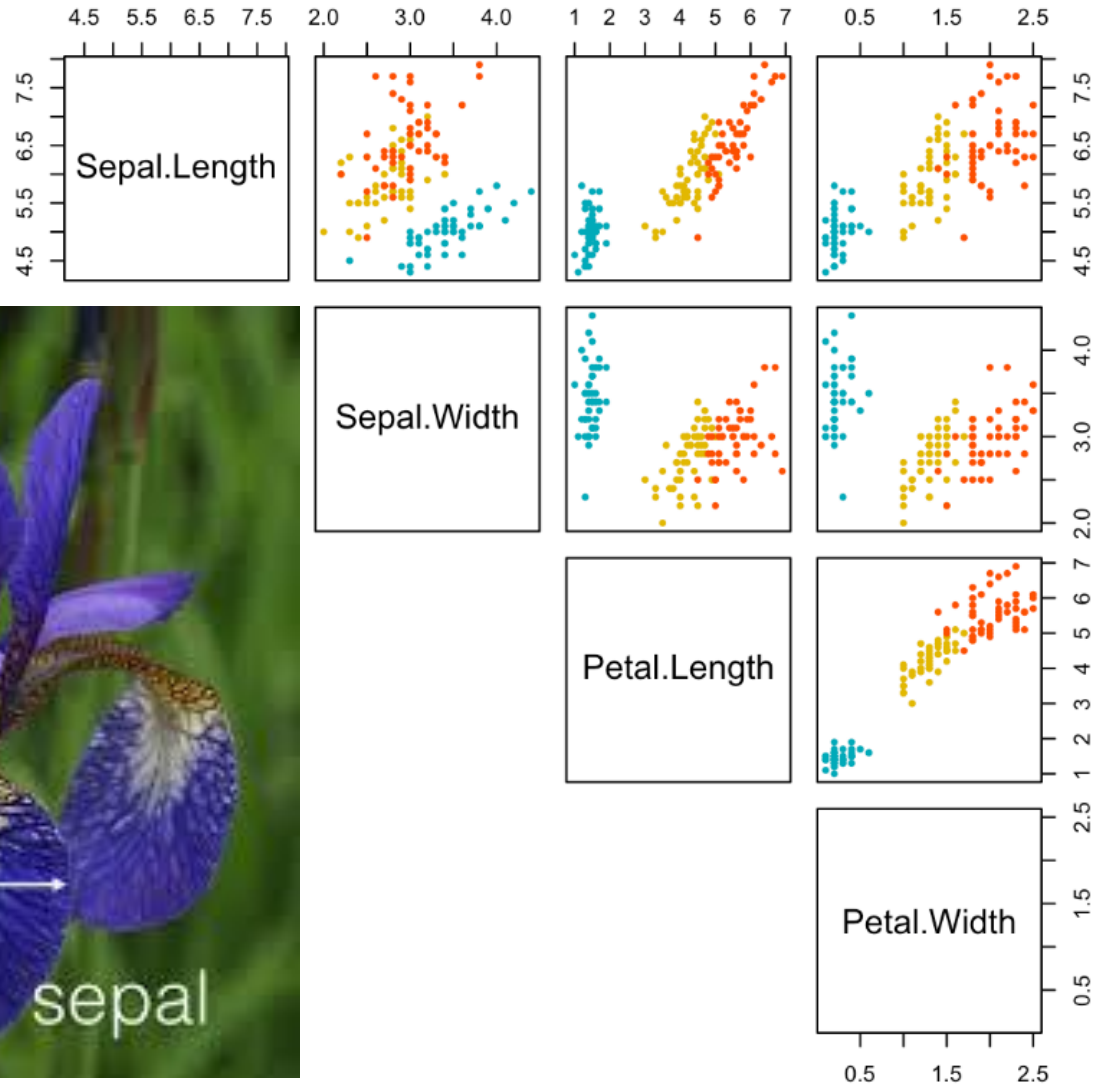
The Elements of data analytical style
(Jeff Leek)

Numerical summarization



Visual summarization





2. Input: Concepts, Instances, and Attributes

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

Flat file

Concept

Attribute

Instance

2.1 What's a Concept?

- Styles of learning:
  - Classification learning: predicting a discrete class
  - Association learning: detecting ("interesting") associations between features
  - Clustering: grouping similar instances into clusters
  - Numeric prediction: predicting a numeric quantity
- Concept: thing to be learned
- Concept description: output of learning scheme

2.1 What's a Concept?

- Classification Learning
  - Example problems: weather data, contact lenses, irises, labor negotiations*
  - Classification learning is supervised
  - Outcome is called the class of the example
  - Measure success on fresh data for which class labels are known (test data)

* Witten et al. (2011) Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.) use various examples some are classic, e.g., Irises, some are not.

2.1 What's a Concept?

- Association Learning
  - Can be applied if no class is specified and any kind of structure is considered "interesting"
  - Difference to classification learning:
    - Can predict any attribute's value, not just the class, and more than one attribute's value at a time
    - Hence: far more association rules than classification rules
    - Thus: constraints are necessary
      - Minimum coverage and minimum accuracy

2.1 What's a Concept?

- Clustering
  - Finding groups of items that are similar
  - Clustering is unsupervised
    - The class of an example is not known
  - Success often measured subjectively

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

2.1 What's a Concept?

- Numeric Prediction
  - Variant of classification learning where "class" is numeric (also called "regression")
  - Learning is supervised
  - Measure success on test data

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
...	...	...	...	...

2.2 What's in an Example?

- More specifically an example (input) is an instance, i.e.
  - Thing to be classified, associated, or clustered
  - An individual, independent example of target concept
  - Characterized by a predetermined set of attributes

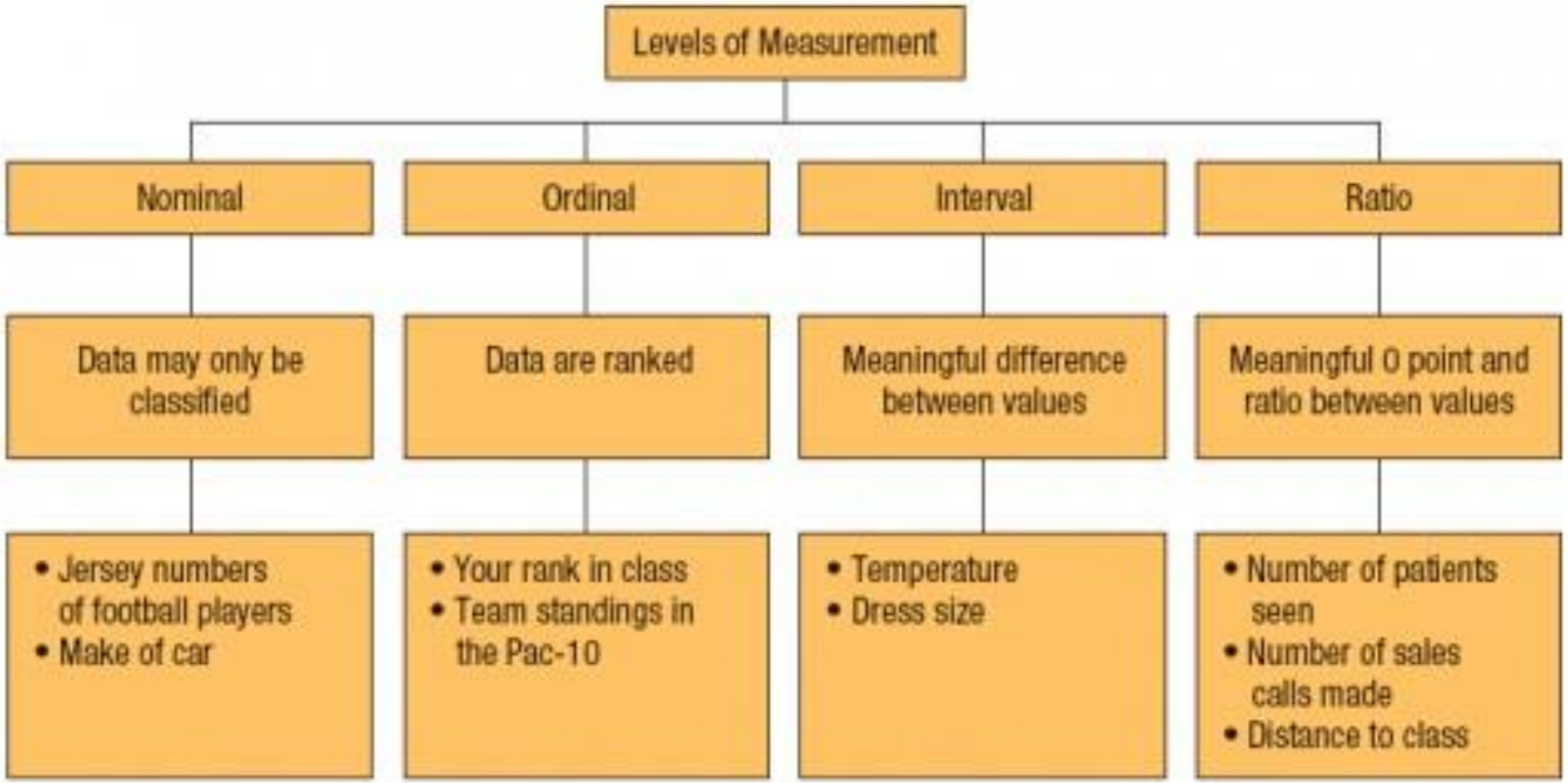
- A dataset is typically a matrix of instances vs. attributes
  - Most common form in practical data mining
  - Represented as a single relation/flat file
  - Yet it is a restrictive form of input: no relationships between objects

2.2 What's in an Example?

- Summary so far
  - Input comprises a set of independent instances
    - All instances are described by the same attributes
    - One or more instances within the input may be responsible for the output
  - Goal is to learn a concept description
  - Relations are hard to represent in flat files
  - Denormalisation is the process of engineering flat file relationships, but can generate systemic noise (spurious regularities).
  - It is critically important to look for spurious regularities before applying any learning methods.
  - Potentially infinite concepts can be dealt with by learning/applying recursive rules.

National College of Ireland

Possible attribute types ("levels of measurement")



Possible attribute types (permissible statistics)

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

2.3 What's in an Attribute?

- Nominal Quantities
  - Values are distinct symbols
    - Values themselves serve only as labels or names
    - Nominal comes from the Latin word for name
  - Example: attribute "outlook" from weather data
    - Values: "sunny", "overcast", and "rainy"
  - No relation is implied among nominal values (no ordering or distance measure)
  - Only equality tests can be performed

2.3 What's in an Attribute?

- Ordinal Quantities
  - Impose order on values
  - But: no distance between values defined
  - Example: attribute "temperature" in weather data
    - Values: "hot" > "mild" > "cool"
  - Note: addition and subtraction don't make sense
  - Example rule: temperature < hot $\Longrightarrow$ play = yes
  - Distinction between nominal and ordinal not always clear (e.g., attribute "outlook")

2.3 What's in an Attribute?

- Interval Quantities
  - Interval quantities are not only ordered but measured in fixed and equal units
  - Example 1: attribute "temperature" expressed in degrees Fahrenheit
  - Example 2: attribute "year"
  - Difference of two values makes sense
  - Sum or product doesn't make sense
    - Examples: 3 x 2014 or 32F + 67F
  - Zero point is not defined can change or is arbitrary!

2.3 What's in an Attribute?

- Ratio Quantities
  - Ratio quantities are ones for which the measurement scheme defines a zero point
  - Example: attribute "distance"
    - Distance between an object and itself is zero
  - Ratio quantities are treated as real numbers
    - All mathematical operations are allowed
  - But: is there an "inherently" defined zero point?
    - Answer depends on scientific knowledge (e.g. , Fahrenheit knew no lower limit to temperature – now we know it is (or should be) 0 Kelvin or -459.67F)

2.4 Preparing the Input

- Getting to Know the Data
  - There is no substitute for familiarising yourself with your data set
  - Simple visualization tools are very useful
    - Nominal attributes: histograms highlights if the distribution is consistent with background knowledge
    - Numeric attributes: graphs show obvious outliers or erroneous values
    - 2-D and 3-D plots show dependencies
  - Cleaning is expensive and time consuming, but critical to successful data mining
    - Too much data to inspect? Take a sample!
  - Need to consult domain experts

2.5 Summary

- What's a Concept?
- What's in an Example?
- What's in an Attribute?
- Preparing the Input

Bibliography

Slides for Chapter 2 of *Data Mining* by I. H. Witten, E. Frank and M. A. Hall