Selecting and Converting Data Types

Martin Burger
STATS PROGRAMMING TUTOR

www.r-tutorials.com

Data Types

Each data type has its own implications

- Analytical procedures might be determined by the data type
- Example: ANOVA needs numeric data with a grouping variable

Data table: Columns have data of the same class

Data types: Numeric, integer, character, factor, logical, date-time

Comparison of data.frame and list

Data Types

Poorly understood data types are one of the main cause of problems in analytics.

Data type

The class of data of a single vector. Consider data.frames as a collection of related vectors of equal length.

Data Types in a Data Frame

Integer	Numeric	Date time	Factor	Character	Boolean	Complex

Data Classification

Discrete Data

Takes values of a defined pool of elements

Example: A list of integers

(1, 2, 3, 4)

The number of elements is finite

Continuous Data

Takes any value of a range of numeric or date-time values

Example: A numeric vector of decimals

(32.4343, 54.4334, 45.5555)

The number of possible elements is infinite

Data Classification

Discrete and continuous data

Grouped and ungrouped data

- Groups derived from qualitative information in the data
- Some statistical procedures require the data to be grouped
- The number of factors is finite and known

Quantitative (numeric) and qualitative (factors, characters) data

Statistical tests and tools are bound to certain data types (e.g. ANOVA, box-plot)

Type Conversion: Numeric and Integer

Functions for Type Conversion

`as.numeric`

`as.integer`

`as.character`

`as.factor`

Date-time conversion has complex rules

Type Conversion: Factor and Character

A factor is a grouping variable

The number of groups is known and finite

Factors enable methods like clustering

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.0	2.0	3.5	1.0	versicolor
6.0	2.2	4.0	1.0	versicolor
6.2	2.2	4.5	1.5	versicolor
6.0	2.2	5.0	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4.0	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor

Data Type: Character

String values often require pre-processing

They support limited amount of analysis

- E.g.: Sentiment analysis of tweets

The number of possible symbol combinations is unknown and infinite

```
read.csv('myfile.csv', stringsAsFactors = True)
```

Unintended Conversion at Data Import

The read.csv function of R Base converts character values to factors by default (`stringsAsFactors = True`)

To prevent this behavior set the stringsAsFactors argument to False in case you are using this import method

Boolean or Logical Values

Boolean or Logical Data

Binary data: True or False values

Result of logical tests (yes-or-no) that measure the data against a threshold

- Divides the data into two fractions

If Boolean values are not accepted, convert them to 1 (True) and 0 (False)

Logical Operators

Operator	Meaning
>	Greater than
<	Less than
==	Equal
>=	Greater than or equal
<=	Less than or equal
!=	Not equal

R Lists

Data Structures in R

Vector

A sequence of values of the same type

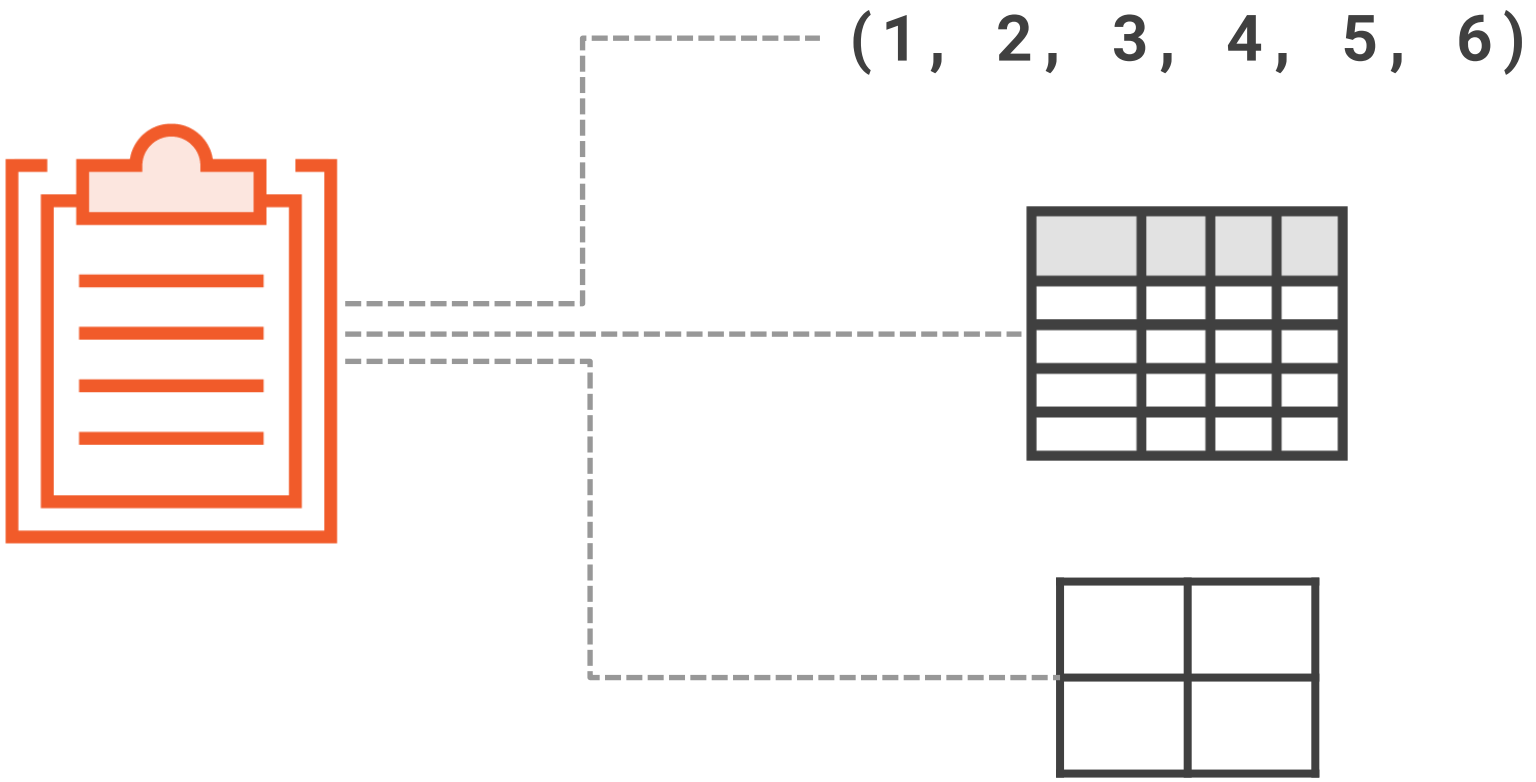
Data.frame

A collection of vectors of the same length

List

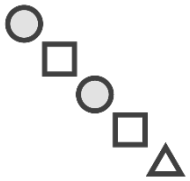
A collection of objects of various kinds

Lists Collect Objects of Various Kinds



Lists in Data Analysis

R lists are less popular than tabular structures



A good class to capture chaotic data

Complex analytical and modeling tools might return lists

Working with Date and Time

Challenges of Working with Date and Time

Complexity factors: time zone, leap years, leap seconds

Choosing and converting data to the most suitable format is part of the analytical job

Extended R toolbox:

- R Base
- Libraries chron and lubridate

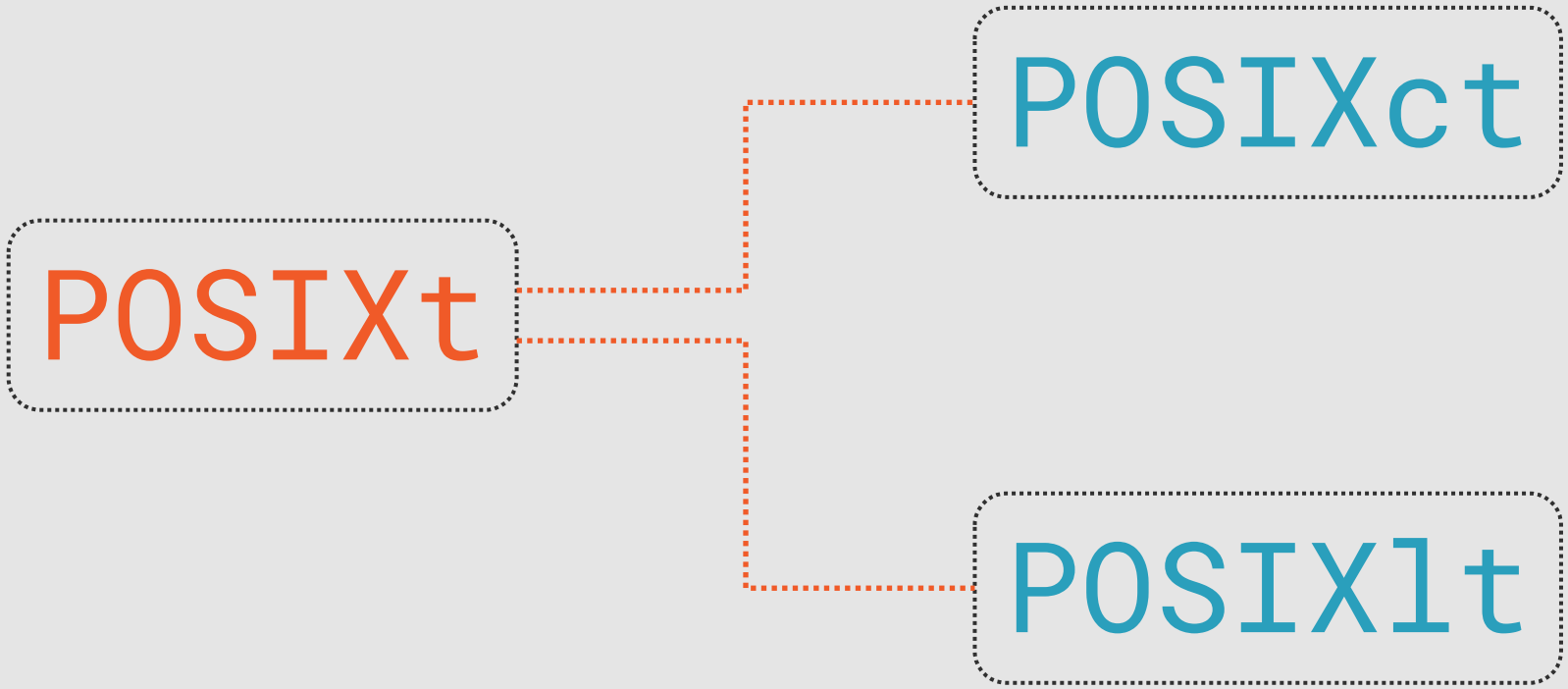
Standard Class for Date and Time

POSIXt: Portable operating system interface for time

Date and time encoded in this class is recognized by all standard operating systems

Communicates date and time unambiguously

Subclasses of POSIXt



The Birth Second

01/01/1970 00:00:00

Time is specified in seconds correlated to the birth second

Relative Date Classes in R Base

Class POSIXct

Time is measured in seconds passed since the birth second

Class Date

Time is measured in days passed since 01/01/1970

Date and Time Class Chron

Class chron from add-on library chron

- Days (+fragments) passed by since 01/01/1970

Chron is time zone naive

Backup classes dates and times for functions that do not recognize chron

Date and Time Classes in R

POSIXlt
(Date, time, time zone)

POSIXct
(Relative to birth second)

Date
(Relative to birth date)

Chron
(Time zone naïve)

Type Conversion from String to Date and Time

Date and time is often read as character by R

- Date and time comes in various formats

Parse date and time with strptime()

- Alternatives in library lubridate
- Read in strings as date time with the help of a format code
- Input strings must be uniform in their format
- Use the help section to build the format code

Data Types

Understanding data types is the foundation of data science

Numeric vs. integer

- Int: Often used for counts
- Num: Measurements with precision

Character vs. factor

- Factor: Grouping variable
- Chr: Text

Boolean (logi): True or False

Date and time: Classes POSIXt, chron and Date