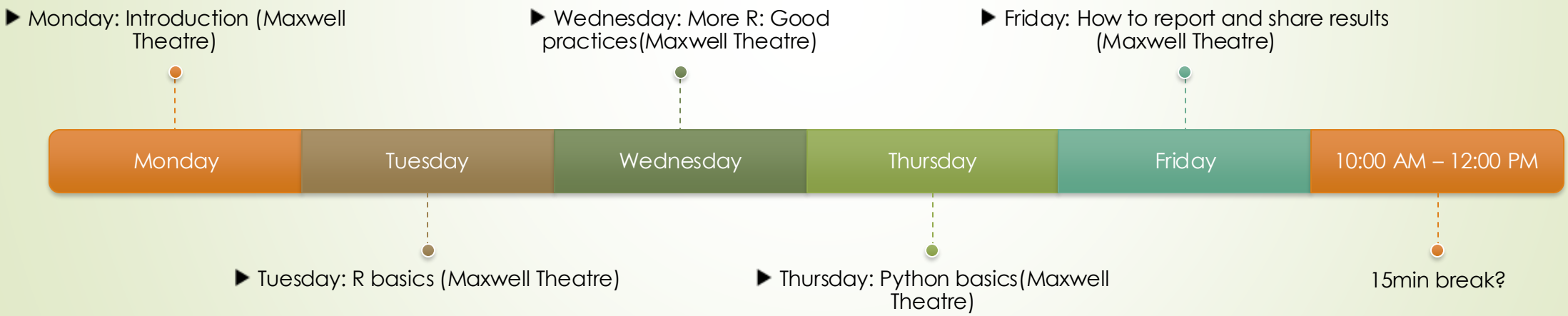
Applied Social Data Science - Coding Camp

Trajche Panov, PhD

Teaching Fellow

September 2 - 6, 2024

Schedule


Monday: Introduction (Maxwell Theatre)
Wednesday: More R: Good practices(Maxwell Theatre)
Friday: How to report and share results (Maxwell Theatre)
Monday
Tuesday
Wednesday
Thursday
Friday
10:00 AM – 12:00 PM
Tuesday: R basics (Maxwell Theatre)
Thursday: Python basics(Maxwell Theatre)
15min break?

Today's class

R Basics

Installing R

How R works: Basic commands and functions

Why R?

It's free!

It runs on a variety of platforms including Windows, Unix and MacOS.

It provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner.

It contains advanced statistical routines not yet available in other packages.

It has state-of-the-art graphics capabilities.

R has a Steep Learning Curve

- First, while there are many introductory tutorials (covering data types, basic commands, the interface), none alone are comprehensive. In part, this is because much of the advanced functionality of R comes from hundreds of user contributed packages. Hunting for what you want can be time consuming, and it can be hard to get a clear overview of what procedures are available.

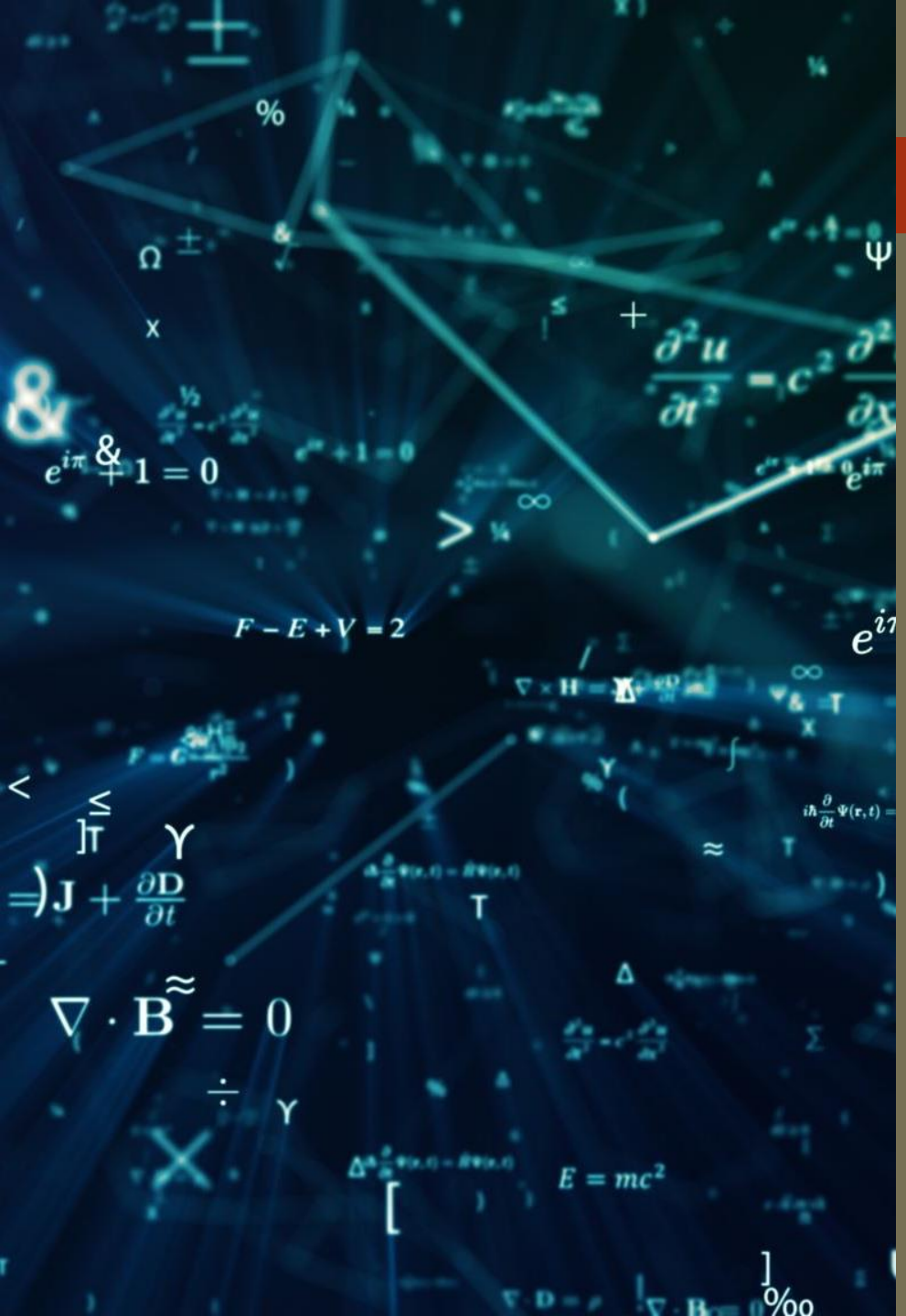
R paradigm is different

- Rather than setting up a complete analysis at once, the process is highly interactive. You run a command (say fit a model), take the results and process it through another command (say a set of diagnostic plots), take those results and process it through another command (say cross-validation), etc. The cycle may include transforming the data, and looping back through the whole process again. You stop when you feel that you have fully analyzed the data.

Tutorials

Each of the following tutorials are in PDF format.

- P. Kuhnert & B. Venables, An Introduction to R: Software for Statistical Modeling & Computing
- J.H. Maindonald, Using R for Data Analysis and Graphics
-
- W.J. Owen, The R Guide
- W.N. Venebles & D. M. Smith, An Introduction to R



R Overview

- R is a comprehensive statistical and graphical programming language and is a dialect of the S language:
- 1988 - S2: RA Becker, JM Chambers, A Wilks
- 1992 - S3: JM Chambers, TJ Hastie
- 1998 - S4: JM Chambers
- R: initially written by Ross Ihaka and Robert Gentleman at Dep. of Statistics of U of Auckland, New Zealand during 1990s.
- Since 1997: international "R-core" team of 15 people with access to common CVS archive.

R Features

R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

R has an effective data handling and storage facility,

R provides a suite of operators for calculations on arrays, lists, vectors and matrices.

R provides a large, coherent and integrated collection of tools for data analysis.

R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

Basics

Functions

- Everything done through functions
- Strict named arguments
- Abbreviations in arguments OK (e.g. T for TRUE)

Objects

- Everything is an object
- "<-" is an assignment operator
- "X <- 5": X GETS the value 5

Data Structures

Supports virtually any type of data

Numbers, characters, logicals (TRUE/ FALSE)

Arrays of virtually unlimited sizes

Simplest: Vectors and Matrices

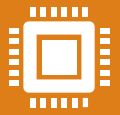
Lists: Can Contain mixed type variables

Data Frame: Rectangular Data Set

Data Structure in R

	Linear	Rectangular
All Same Type	VECTORS	MATRIX*
Mixed	LIST	DATA FRAME

R Overview

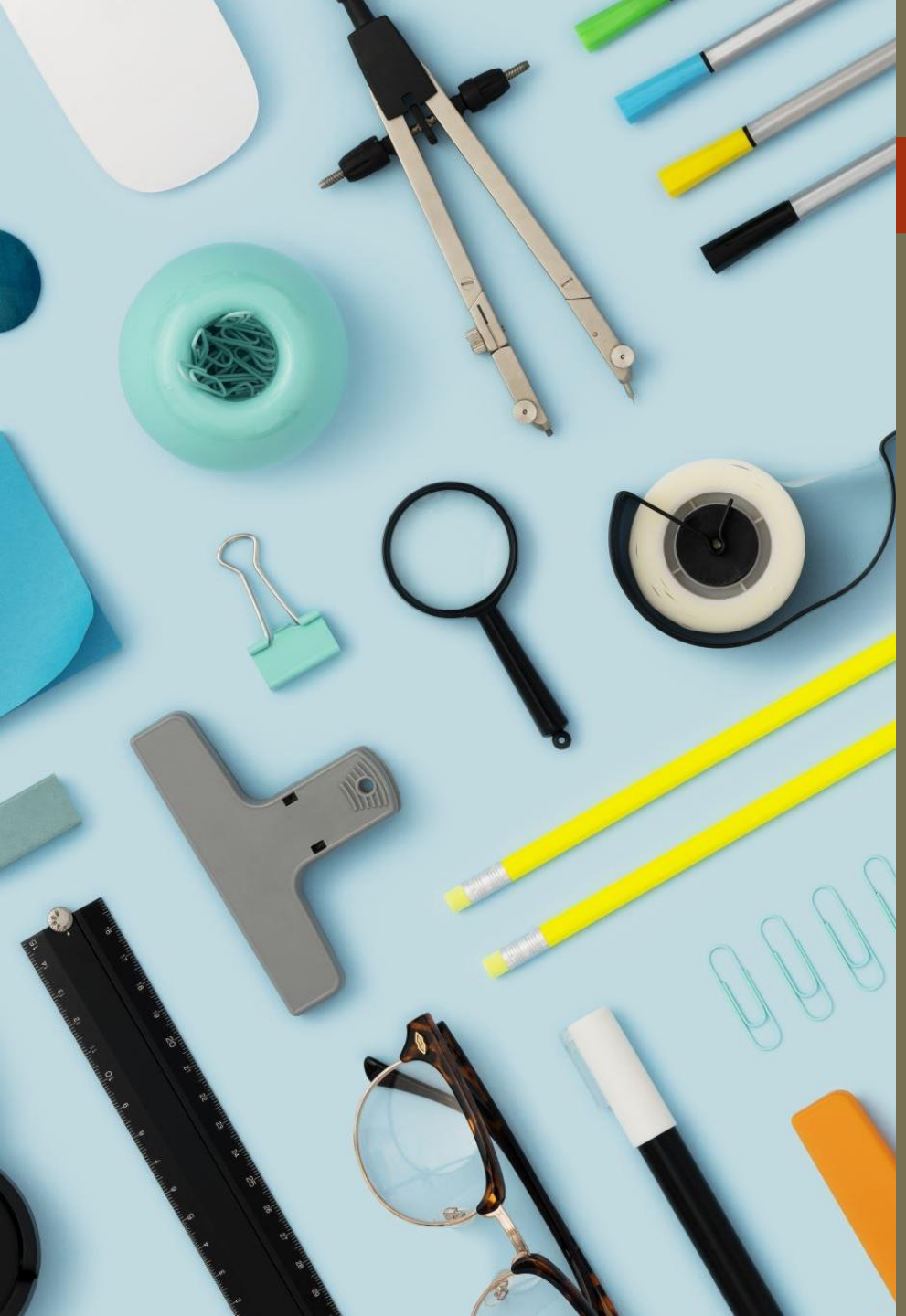


Most functionality is provided through built-in and user-created functions and all data objects are kept in memory during an interactive session.

Basic functions are available by default. Other functions are contained in packages that can be attached to a current session as needed

A key skill to using R effectively is learning how to use the built-in help system. Other sections describe the working environment, inputting programs and outputting results, installing new functionality through packages and etc.

A fundamental design feature of R is that the output from most functions can be used as input to other functions. This is described in reusing results.



R Workspace

- Objects that you create during an R session are hold in memory, the collection of objects that you currently have is called the workspace. This workspace is not saved on disk unless you tell R to do so. This means that your objects are lost when you close R and not save the objects, or worse when R or your system crashes on you during a session.

Strengths and Weaknesses

Strengths

- Free and Open Source
- Strong User Community
- Highly extensible, flexible
- Implementation of high end statistical methods
- Flexible graphics and intelligent defaults

Weakness

- Steep learning curve
- Slow for large datasets

Let's get started!

Open RStudio!

Thank you for your attention!

panovt@tcd.ie