Lecture 1

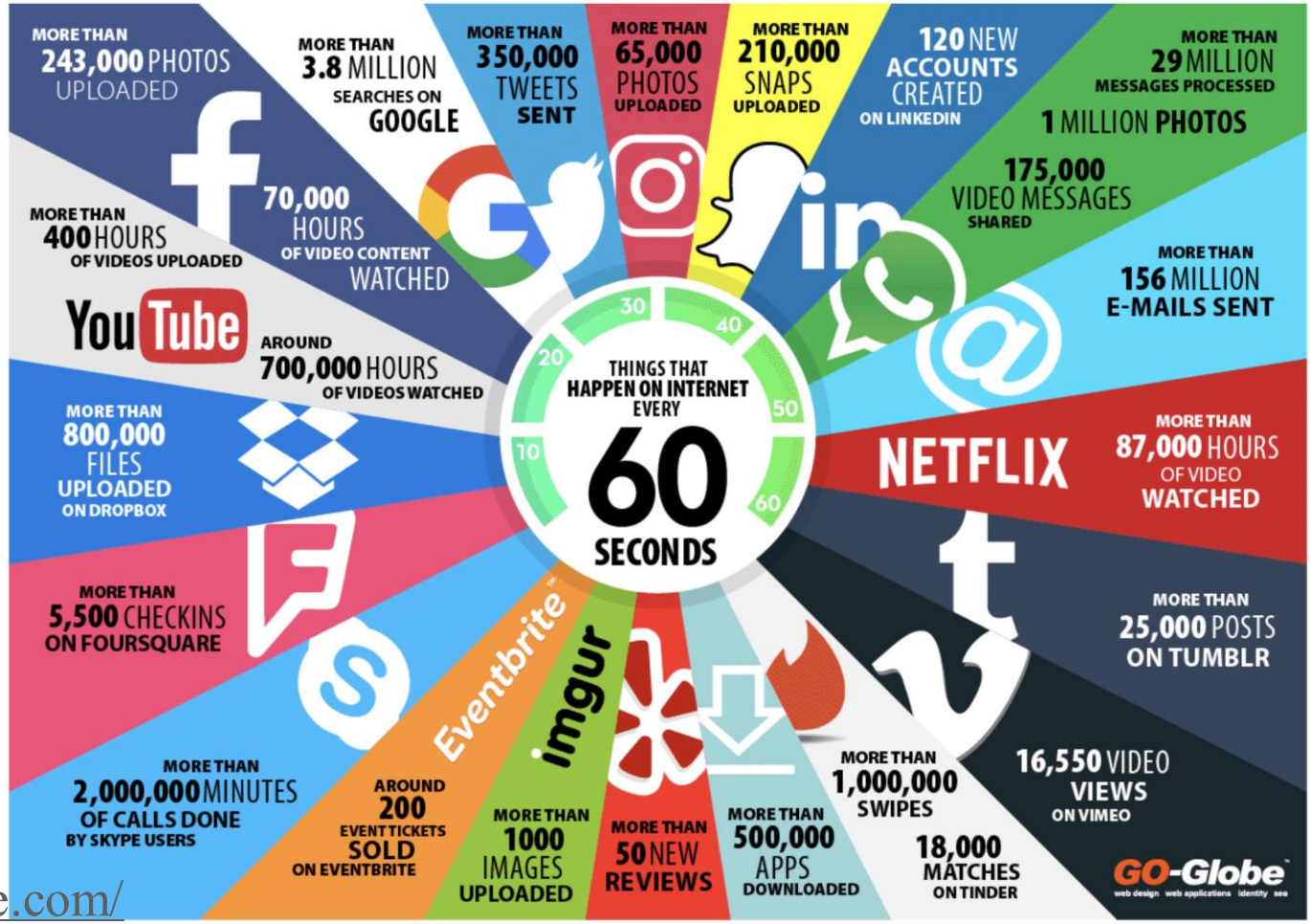
- Introduction to Data Science
- Unit Overview
- Introduction to R and RStudio
- Basic Statistics in R

Presenter: Dr John Betts

Introduction to data science

- An Internet minute
- What is Data Science?
- Recent examples
- Some common themes in data science
- Necessary skills for data scientists
- The data science process

An Internet minute 2025

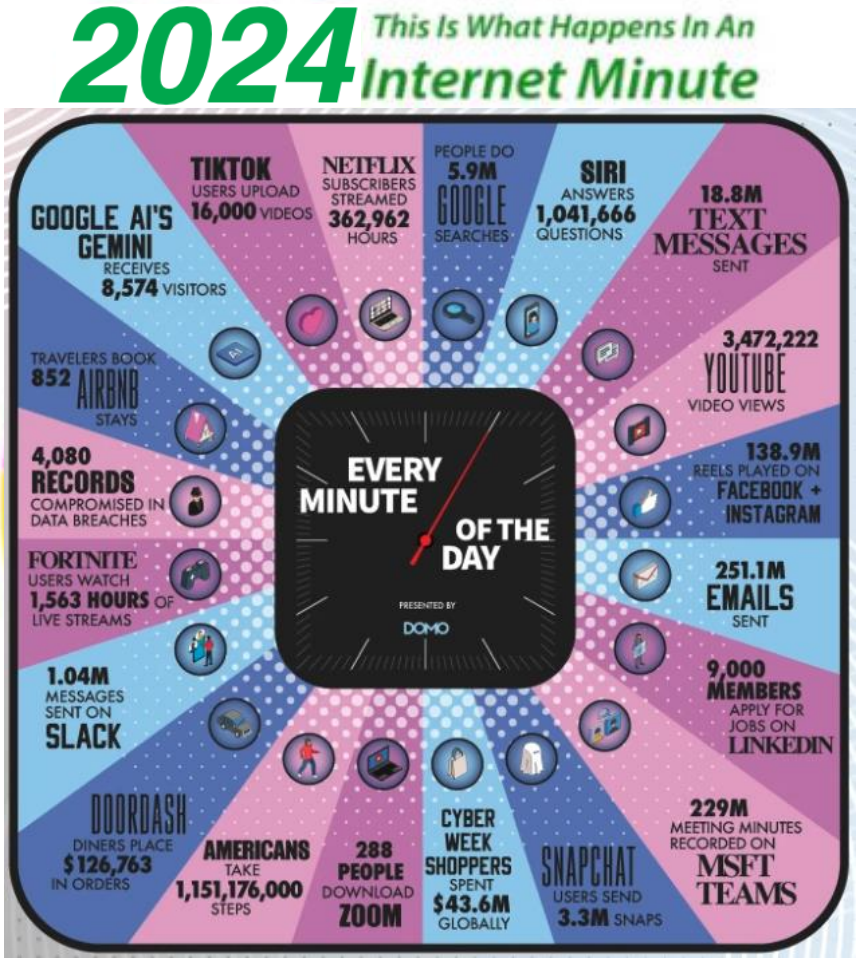


go-globe.com/

The Internet and related digital technologies have enabled humans to collect and store data at an unprecedented scale.

One consequences has been the rise of data science as a profession…

A sea of data: 2018 – 2024



visualcapitalist.com/

Applied Session Activity (a)

Compare the figures on the two previous slides showing Internet activity in 2025 and then from 2018 to 2024. Answer the following:

- What trends and changes in online activity do you see?
- What are the major changes in Internet use over the longer term?
- Choosing one platform/app/service, what types of data could be collected for analysis (assuming you had permission)?
- What human behaviours or natural phenomena could that data be used to study? How would you adapt the data for use?

What is data science

From Wikipedia:

- Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms, and systems to extract or extrapolate knowledge from potentially noisy, structured, or unstructured data.
- Data science also integrates domain knowledge from the underlying application domain … is multifaceted and can be described as a science, a research paradigm, a research method, a discipline, a workflow, and a profession.
- Data science is “a concept to unify statistics, data analysis, informatics, and their related methods” to "understand and analyze actual phenomena" with data…
- Data science is often described as a multidisciplinary and rapidly evolving field…

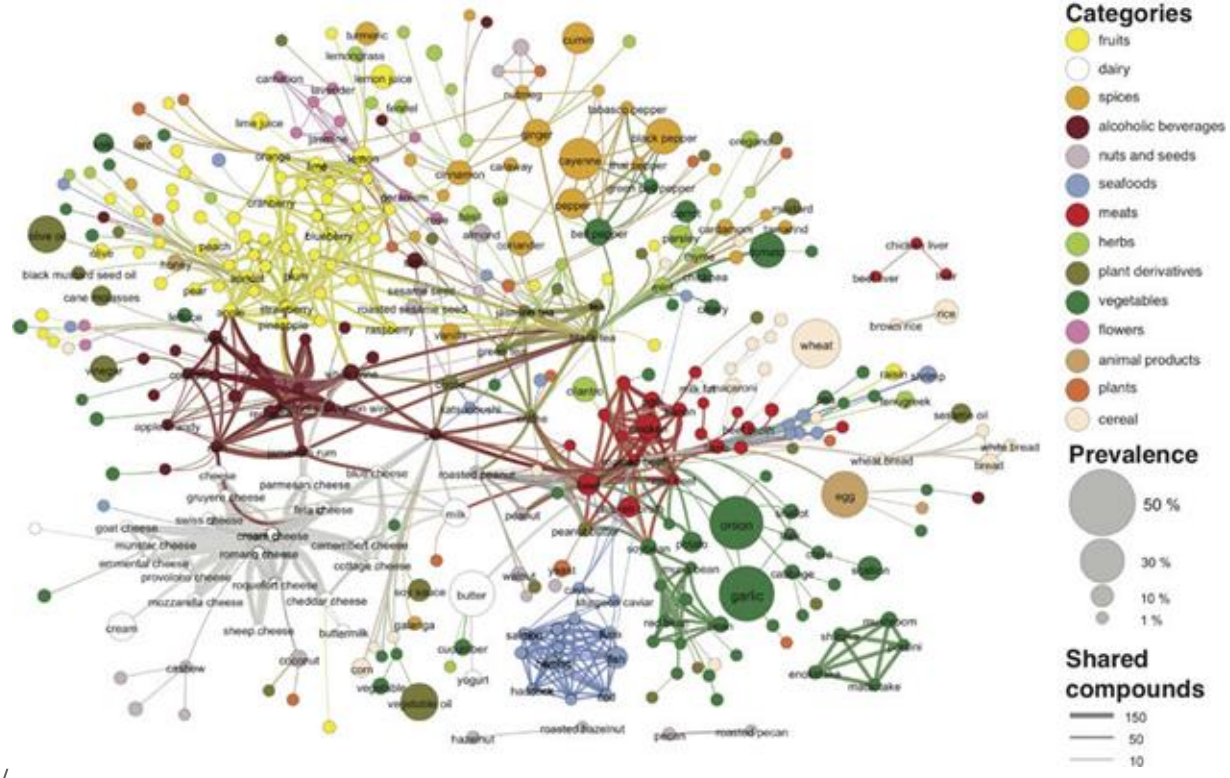
en.wikipedia.org/wiki/Data_science

Data Science: A few examples

- Some examples are quite old now, some more recent.
- Each one is chosen to illustrate one or more fundamental aspects of data science and illustrate one or more of the skills we learn in this unit.
- See if you can think what these are…

Food networks

Flavor network and the principles of food pairing



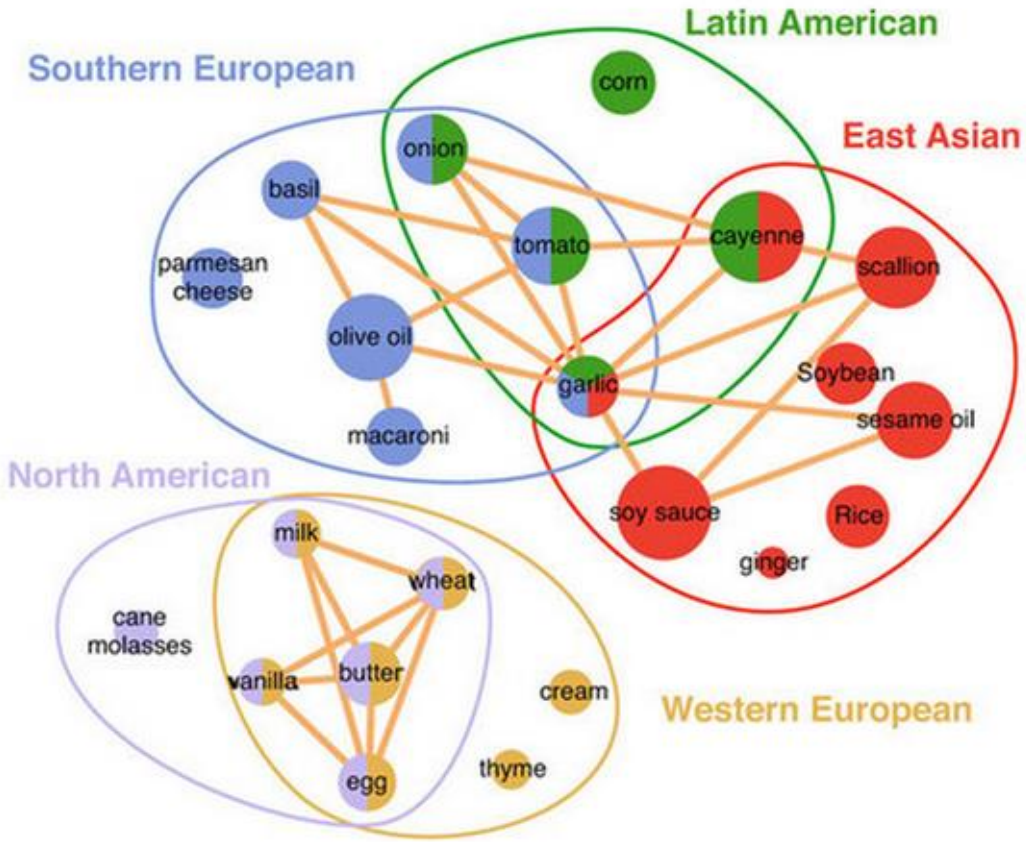
nature.com/

Food networks

...
do we more frequently use ingredient pairs that are strongly linked in the flavor network or do we avoid them? To test this hypothesis we need data on ingredient combinations preferred by humans, information readily available in the current body of recipes. For generality, we used 56,498 recipes provided by two American repositories (epicurious.com and allrecipes.com) and to avoid a distinctly Western interpretation of the world's cuisine, we also used a Korean repository (menupan.com). The recipes are grouped into geographically distinct cuisines (North American,Western European, Southern European, Latin American, and East Asian)...

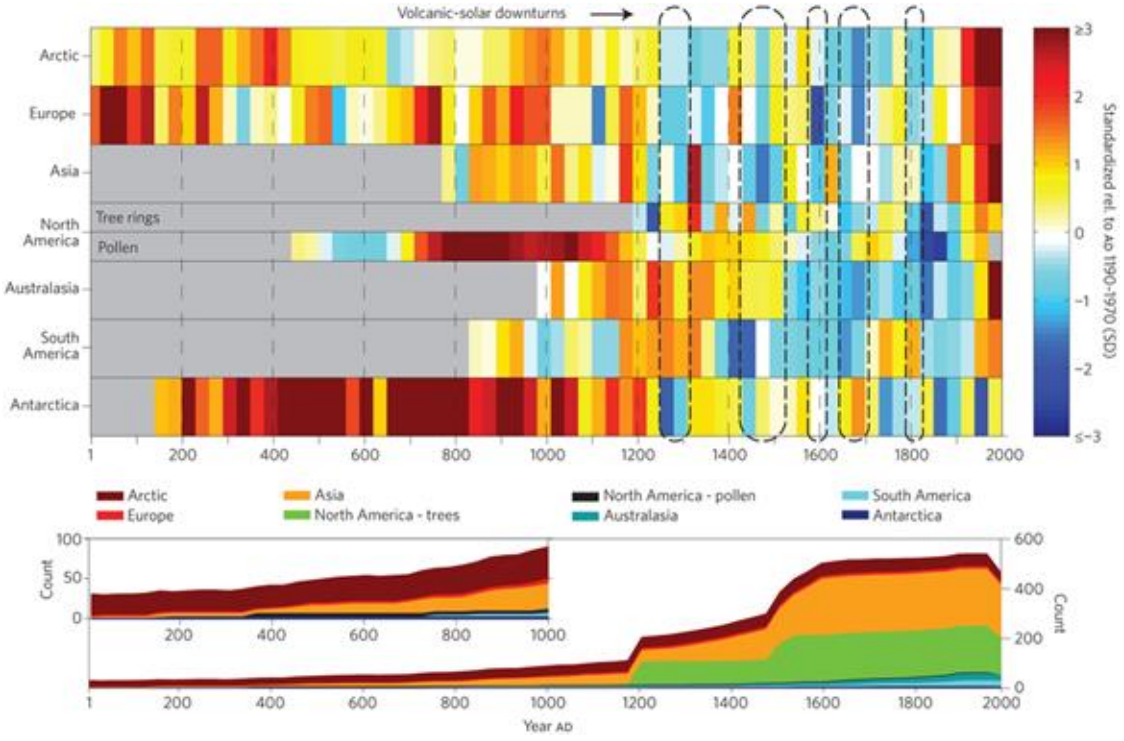
Food networks

- Co-occurrence of major ingredients in 5 cuisines



Climate change

Continental-scale temperature variability during the past two millennia



nature.com/

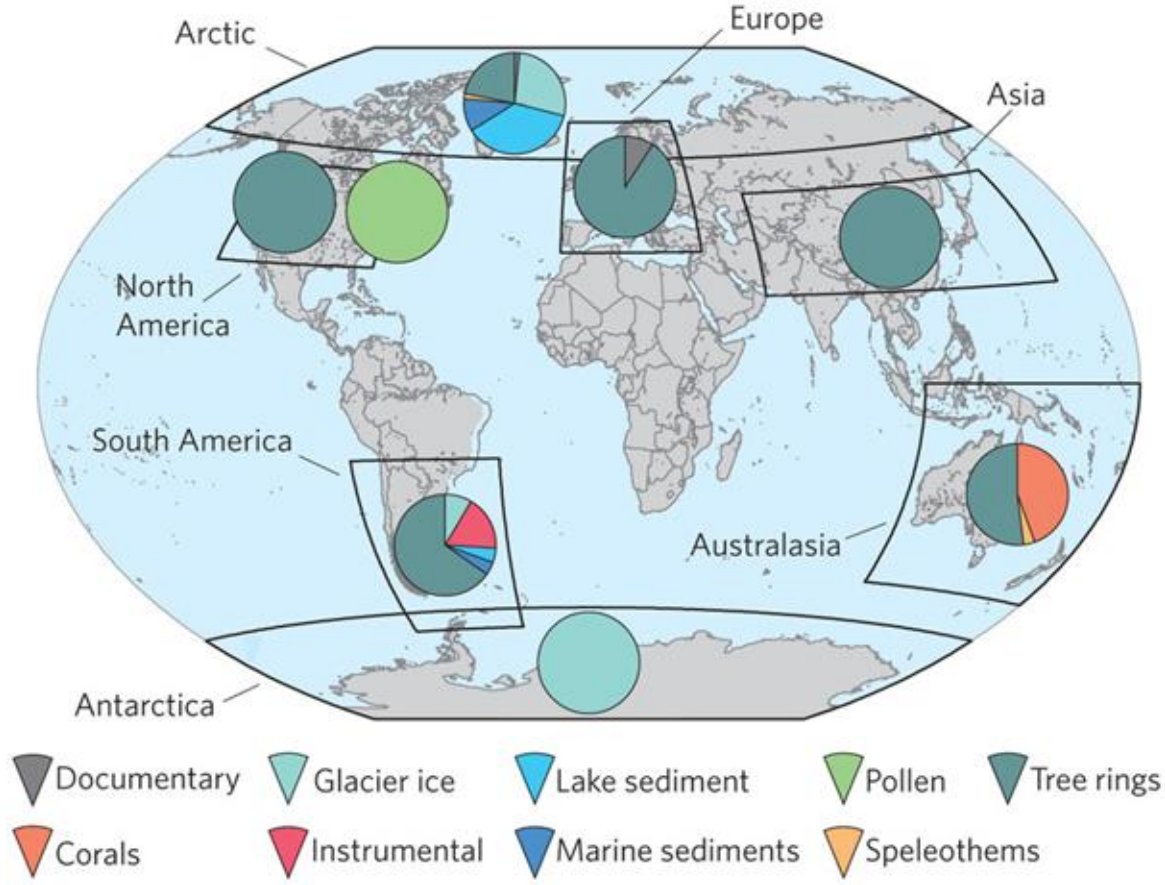
Climate change

...

The '2k Network' of the IGBP Past Global Changes (PAGES) project aims to produce a global array of regional climate reconstructions for the past 2000 years. ... Nine PAGES 2k working groups represent eight continental-scale regions and the oceans. Regional representation brings critical expert knowledge of individual proxy data sets, which is essential for improving palaeoclimate reconstructions. The PAGES 2k Network is coordinated with the National Oceanic and Atmospheric Administration (NOAA) World Data Center for Paleoclimatology to establish a benchmark database of proxy climate records for the past two millennia ...

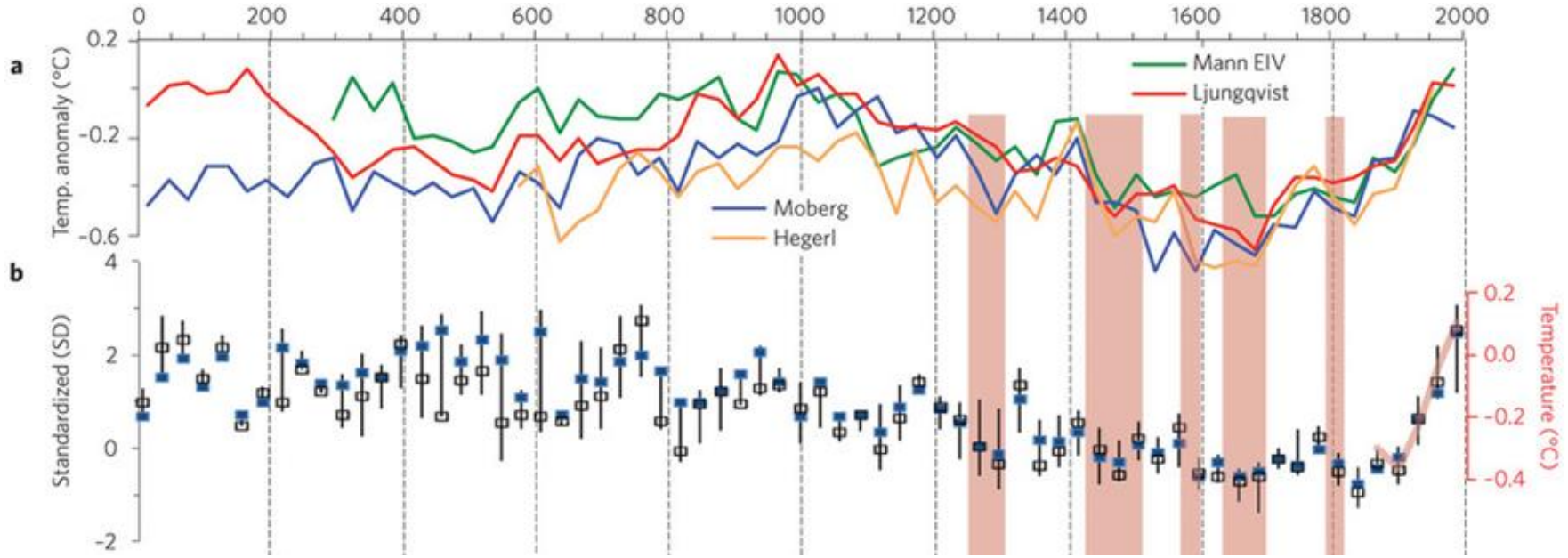
Climate change

- Data sources and locations



Climate change

- Temperature variability over past 2000 years



Text analysis

Plagiarism software discloses Shakespeare's inspiration

Literature

Michael Blanding

For years scholars have debated what inspired William Shakespeare's writings. Now, with the help of software typically used by professors to nab cheating students, two writers have discovered an unpublished manuscript they believe the Bard of Avon consulted to write *King Lear*, *Macbeth*, *Richard III*, *Henry V* and seven other plays.

The findings were made by Dennis McCarthy and June Schlueter, who describe them in a book to be published next week by the academic press D.S. Brewer and the British Library. The authors are not suggesting that Shakespeare plagiarised but rather that he read and was inspired by a manuscript titled *A Brief Discourse of Rebellion and Rebels*, written in the late 1500s by George North, a minor figure in the court of Queen Elizabeth.

In reviewing the book before it was published, David Bevington, professor emeritus in the humanities at the University of Chicago and editor of *The Complete Works of William Shakespeare (7th Edition)*, called it "a revelation" for the sheer number of correlations with the plays.

McCarthy used decidedly modern techniques to marshal his evidence, employing WCopyfind, an open-source plagiarism software, which picked out common words and phrases in the manuscript and the plays.

In the dedication to his manuscript, for example, North urges those who might see themselves as ugly to strive to be inwardly beautiful, to defy nature. He uses a succession of words to make the argument, including "proportion", "glass", "feature", "fair", "deformed", "world", "shadow" and "nature". In the opening soliloquy of *Richard III* ("Now is the winter of our discontent ...") the hunchbacked tyrant uses the same words in virtually the same order to come to the opposite conclusion: that since he is outwardly ugly, he will act the villain he appears to be.

William Shakespeare may have found theme and character in an earlier work.

In another passage, North uses six terms for dogs, from the noble mastiff to the lowly cur and "trundle-tail", to argue that just as dogs exist in a natural hierarchy, so do humans. Shakespeare uses essentially the same list of dogs to make similar points in *King Lear* and *Macbeth*.

In 1576, North was living at Kirtling Hall near Cambridge, England. It was here, McCarthy says, that he wrote his manuscript.

The manuscript is a diatribe against rebels, arguing all rebellions against a monarch are unjust and doomed to fail. While Shakespeare had a more ambiguous position on rebellion, McCarthy said he clearly mined North's treatise for themes and characters.

McCarthy was inspired to use plagiarism software by the work of Sir Brian Vickers, who used similar techniques in 2009 to identify Shakespeare as a co-author of the play *Edward III*. While the book has been received favourably, the statistical techniques used have not yet been subjected to rigorous review. Those techniques may only be the "icing on the cake", said Witmore, who briefly examined an advance copy. "At its core, this remains a literary argument, not a statistical one."

The book contends Shakespeare not only uses the same words as North but often uses them in scenes about similar themes, and even the same historical characters.

McCarthy plans future volumes based on his electronic techniques, hoping to shed more light on how Shakespeare wrote his plays.

To make sure North and Shakespeare weren't using common sources, McCarthy ran phrases through the database Early English Books Online, which contains 17 million pages from nearly every work published in English between 1473 and 1700. Almost no other works contained the same words in passages of the same length. Some words are very rare; "trundle-tail" appears in only one other work before 1623.

THE NEW YORK TIMES

AFRGA1 A012 BA

New York Times (reported AFR 10/2/2018)

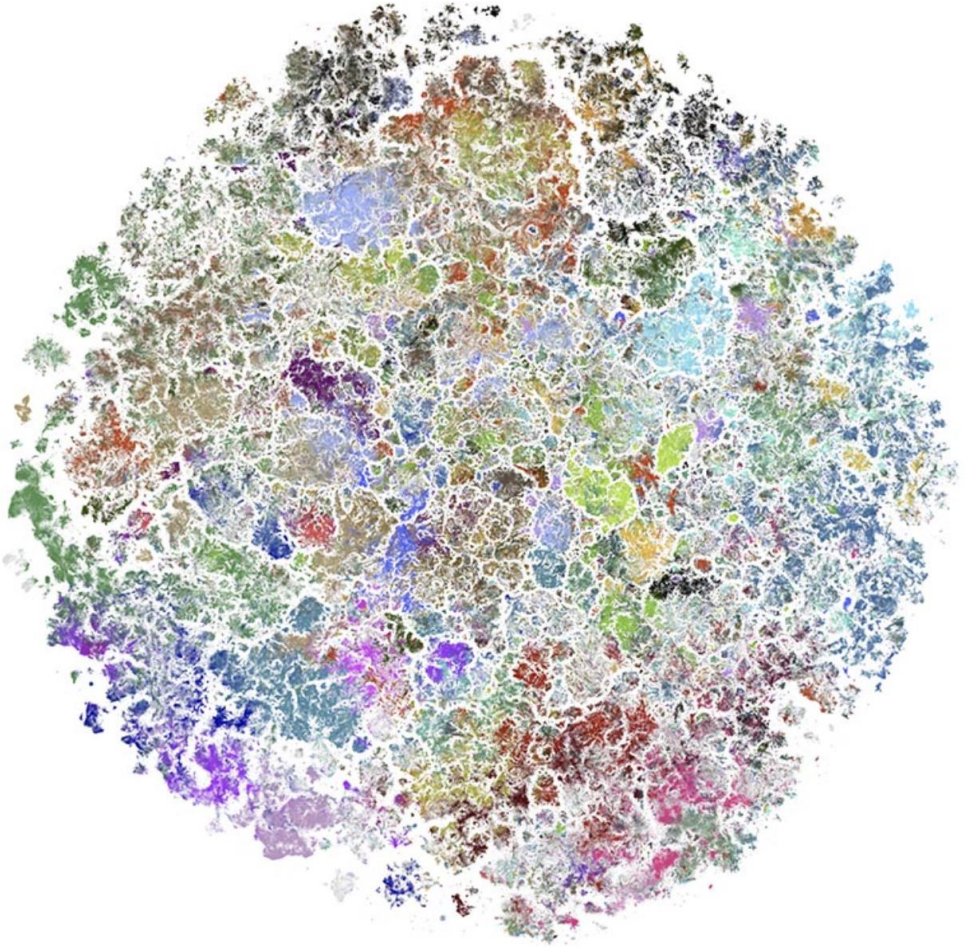
Text analysis

...
For years scholars have debated what inspired William Shakespeare's writings. Now, with the help of software typically used by professors to nab cheating students, two writers have discovered an unpublished manuscript they believe the Bard of Avon consulted to write "King Lear," "Macbeth," "Richard III," "Henry V" and seven other plays.

The news has caused Shakespeareans to sit up and take notice....

Research fraud detection

Atlas of biomedical literature could help track down fabricated studies



science.org/

Research fraud detection

...

To create the atlas, Kobak's team downloaded the abstracts of nearly 21 million English-language articles from the PubMed search engine. The team then used an AI large language model known as PubMedBERT to sort the abstracts by similarity. The model looked for scientific terms within each abstract and interpreted their meaning according to the surrounding text. (For example, PubMedBERT will infer whether the word "replicate" refers to copied DNA or a repeated experiment.) Based on this analysis, it grouped similar publications together into so-called "neighborhoods." ...

Belgium (in red) beat Brazil in a quarter-final match during the 2018 World Cup.

HOW BIG DATA IS CHANGING FOOTBALL

As the FIFA World Cup kicks off, researchers are showing their skills to help soccer coaches develop players and tactics. **By David Adam**

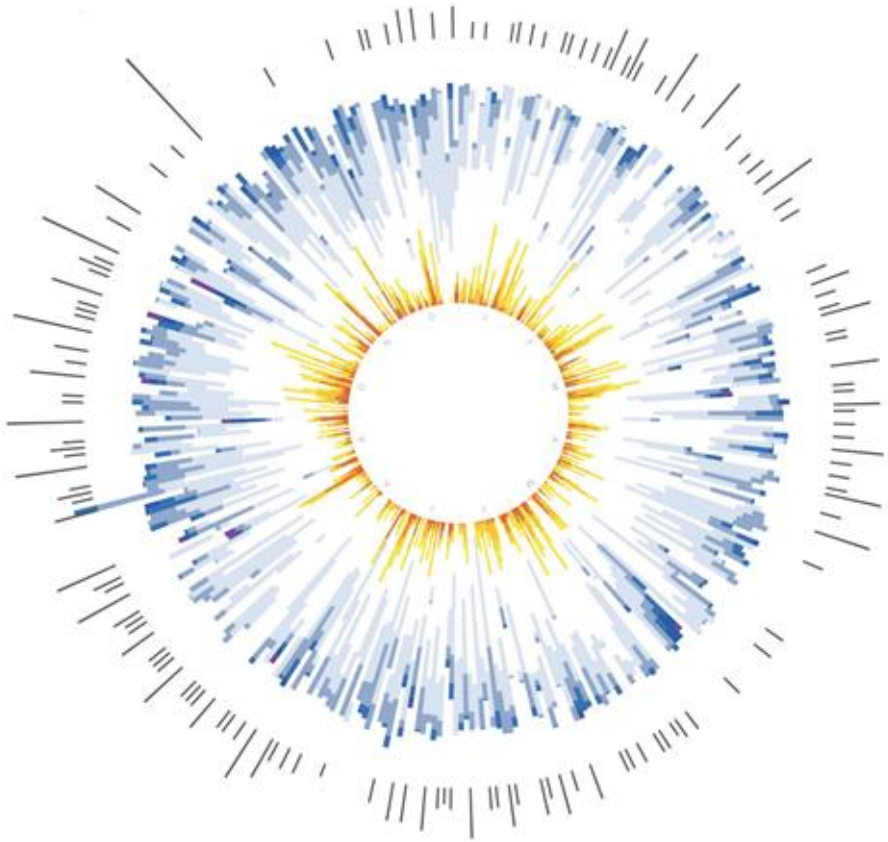
nature.com/

...footballers face the kind of data scrutiny more often associated with an astronaut. Wearable vests and straps can now sense motion, track position with GPS and count the number of shots taken with each foot. Cameras at multiple angles capture everything from headers won to how long players keep the ball.

And to make sense of this information, most elite football teams now employ data analysts, including mathematicians, data scientists and physicists...

...In return, insights from analysts are altering how the game is played: strikers shoot less frequently from a distance, wingers pass to a teammate rather than cross the ball...

Personal analytics



The Healthiest Year Of My Life

Diabetic Charts A Year's Worth Of His Health Data

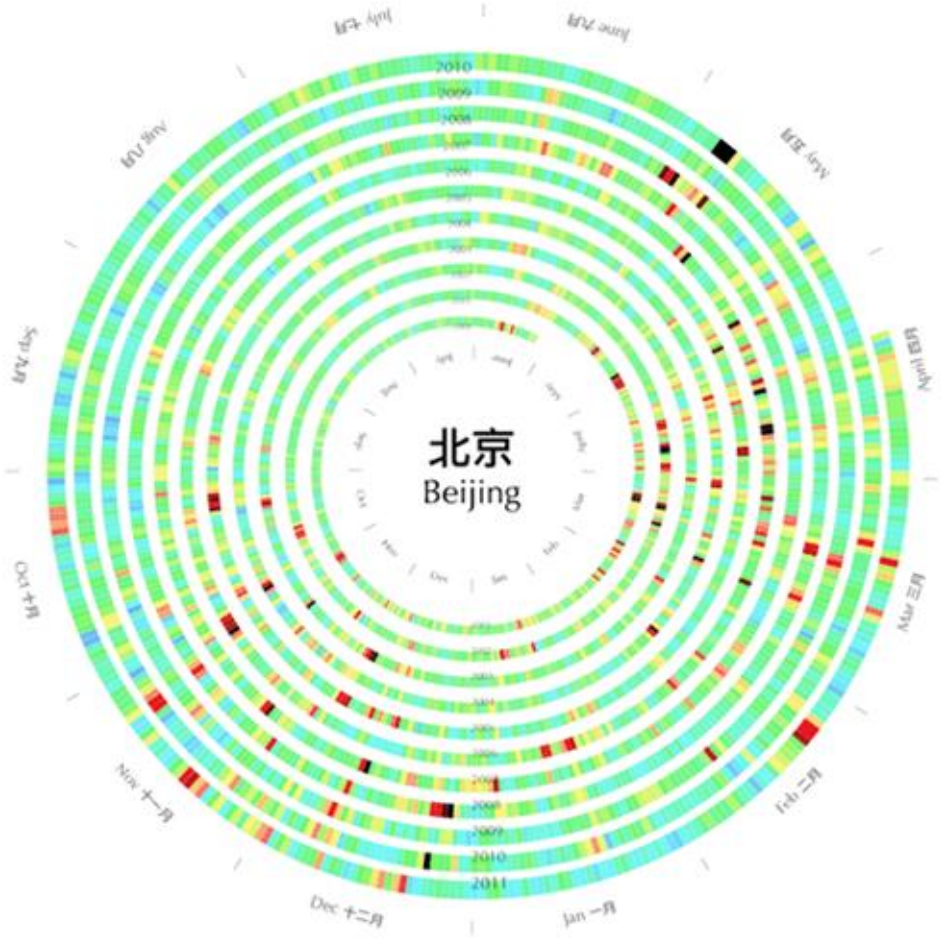
One of our 15 favorite recent data visualizations

By Katie Peek | December 12, 2014

In 2012, Doug Kanter—diabetic since age 12—visualized his disease. He wrote software to compare his blood sugar with his activity and food. He says the feedback made for the healthiest year of his life. At the end of the project, he created this visual summary. The lengths of his running sessions appear around the outside in gray, and 91,251 glucose-monitor readouts form the iris in the center. Low blood sugar is orange, on-target appears white, and high is blue. Inspired by the experience, he created an app and visualization service called Databetes to help other diabetics.

popsci.com/

Beautiful data visualization



xiaoji-chen.com/

Data science: some common themes

Previous examples illustrate:

- Complex problems of societal concern/interest.
- Large/multiple data sets, often messy or incomplete, heterogeneous, non-traditional, proprietary and open data.
- Data repositories created for one purpose may be used to study a bigger phenomenon (food network, for example).
- Data collection and analysis on a scale that would have been unthinkable 20 years ago.
- Much analysis can be performed on a laptop computer.
- Use of high-quality graphics for communicating results!

Applied Session Activity (b)

Using the previous examples for inspiration, find a recent application of data science from the media. Answer the following:

- What is the problem to be solved?
- What type of data is collected?
- What type of analysis is performed?
- What were the findings?
- How might you adapt this data to investigate another aspect of (human) activity?

Data science: for business

Provost and Fawcett, in Data Science for Business (see recommended reading), list 9 generic methods:

- Classification and class probability estimation,
- Regression,
- Similarity Matching: grouping using known criteria,
- Clustering: grouping using unknown criteria,
- Co-occurrence: grouping similar groups of products etc.,
- Profiling: typical behaviour of individuals or groups,
- Link prediction: connections between data.
- Data reduction: condense large data sets,
- Causal modelling: identifying events that influence others.

Data science: more broadly

Same methods as previous slide applied more broadly:

- Search for habitable planets,
- Weather forecasting,
- DNA sequencing and disease genomics,
- Biometrics (identification by physical characteristics),
- Social networks of all kinds (friends, political, terrorist…) and social and political processes and attitudes,
- Identifying new medicines from databases of existing compounds,
- Data journalism,
- Recommender systems (Spotify, Netflix, Amazon…).

Data science: high-level skills

Some necessary skills for a data scientist:

- Understand a problem from client's perspective,
- Collect, cleanse, manage and combine data – which may come from disparate sources,
- Understand the data, most likely using visualization tools as a starting point,
- Analyze and model the data using statistical and (AI) machine learning techniques,
- Communicate the results simply and effectively.

Data science: technical skills

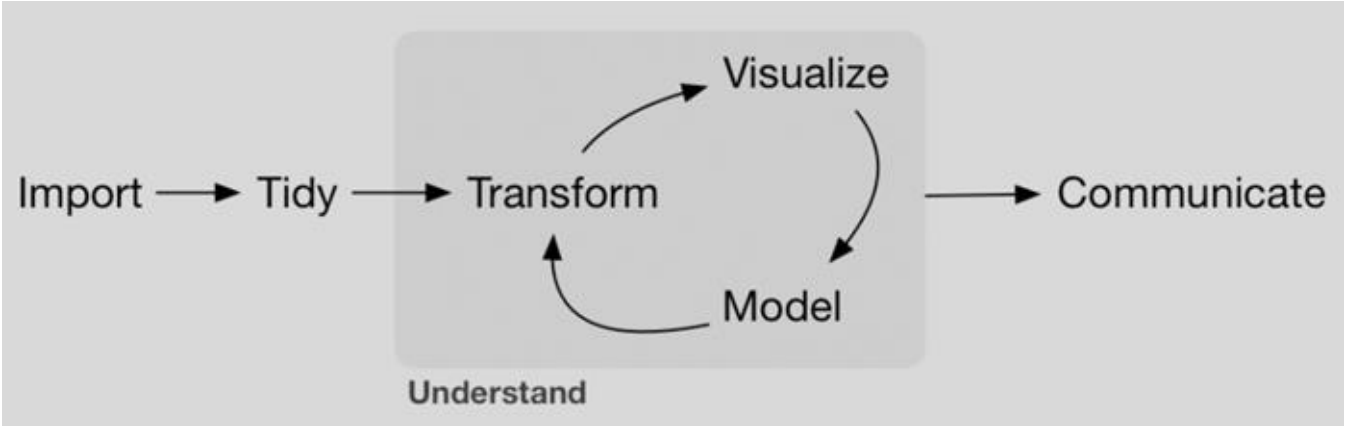
Some necessary technical skills include:

- Creating visualisations of data,
- Statistical analysis,
- Machine learning,
- Programming (e.g., R, Python, …),
- Data storage and data handling,
- Problem solving and hacking mentality,
- Imagination and versatility…

The data science process

Generic methodologies for data analysis:

- For example, the data analysis process from Wickham and Gromelund:



r4ds.had.co.nz/

kaggle

Hosts data science competitions:

- Their motto is “turning data science into a sport,”
- You can view current and past competitions, and perhaps enter some,
- There are lots of tutorials on data science related topics,
- Their Jobs Board is very popular for recruiting,
- kaggle.com/ for details.

Unit Overview

One-page summary:

Seminar:

- Two-hour on-campus seminar running as a lecture, with opportunities for discussion and questions throughout. Tuesday 12:00 – 2:00pm. All students should attend the C1 lecture theatre (CL_Exh-25.FIT_C1 Lec).

Applied Sessions

- Commence Week 2 and follow the seminar topic by a week.

Resources:

- All software is open source, and free.
- Most reference materials are free online, or via Monash Library.

Forum:

- We're using Ed Discussion, join via Moodle or with link below:
- edstem.org/au/courses/34021/discussion

Seminar

- Two-hour on-campus seminar each week. This runs as a lecture, with discussion and questions throughout.
- At Clayton, the seminar runs Tuesday 12:00 – 2:00pm. All students should attend the C1 lecture theatre (CL_Exh-25.FIT_C1 Lec).
- Note there is a second (overflow) location (CL_Inn-33.LearnVillage_D101 FF-Col) but this will only be used if C1 is completely full.
- Seminars will be recorded via MULO.
- Lecture slides (except Week 1) will be available for pre-reading on Moodle the week before.
- Updated lecture slides will be posted as soon as possible after the lecture.

Applied sessions

- In these sessions you will be able to apply the knowledge covered in seminars.
- Applied sessions begin Week 2 and follow the pre-recorded lecture topic one week later.
- Applied sessions are on campus. See Allocate+ for your location.
- Applied sessions are bring-your-own-device.
- Applied session worksheets are posted during the week prior to your session. Please attempt these questions beforehand. Solutions posted at the end of the week.

Discussion forum

- We're using Ed Discussion.
- Clayton and Monash Malaysia students share the same forum.
- Join via Moodle or using the link below:
- https://edstem.org/au/courses/34021/discussion

Unit objectives

What the course is trying to achieve:

- We focus on fundamental, generic, skills essential for a data scientist, independent of software platform or problem domain.
- Problem solving skills, independence and ingenuity.
- Good communication skills. Programming in R.

What it is not trying to achieve:

- Introduction to the vast range of software, techniques and computing platforms available to data scientists.

Unit objectives

Unit design considerations:

- Unit assumes no previous study in data science or R.
- Covers a broad range of data science topics.

Some overlap with:

- FIT1043 Introduction to data science (but we work in R not Python, among other differences).
- FIT2086 Modelling for data analysis (but wider variety of analysis techniques. Not just ML).
- FIT3179 Data visualization (but we have broader range of topics).

Week-by-week outline

Clayton seminar is Tuesday 12:00 – 2:00 pm.

Applied sessions begin Week 2, follows seminar by a week.

Week Starting	Seminar	Topic	App Ses	A1	A2	Q/P	A3	Due Date
2/3/2026	1	Introduction to Data Science, R, review of basic statistics	-					
9/3/2026	2	Data visualisation	S1					
16/3/2026	3	Data manipulation	S2					
23/3/2026	4	Regression modelling	S3					
30/3/2026	5	Clustering	S4					
6/4/2026	-	Mid-semester Break						
13/4/2026	6	Classification using decision trees	S5					17/4/2026
20/4/2026	7	Improving and evaluating classifiers. Naïve Bayes classification	S6					
27/4/2026	8	Ensemble methods, Artificial Neural Networks	S7					
4/5/2026	9	Network analysis	S8					
11/5/2026	10	Introduction to text analysis	S9					15/5/2026
18/5/2026	11	Text analysis applications	Quiz/Prac					22/5/2026
25/5/2026	12	Text Network Analysis, Review of the unit, Assignment 3	S10,11,12					
1/6/2026		SWOT VAC	-					
8/6/2026		EXAM PERIOD	-					12/6/2026

Assessment details

Assignment 1, Due 17th April, Weighting 25%

- Covers data manipulation, visualisation, and data analysis using a variety of techniques. Submission is a written report and short video explaining the key findings of your research.

Assignment 2, Due 15th May, Weighting 20%

- Covers machine learning/artificial intelligence models using R. Submission is a written report and short video.

Assignment 3, Due 12th June, Weighting 25%

- Covers text analysis, networks and clustering using R. Submission is a written report and short video.

Quiz + Practical Activity, Week 11 (Due 22nd May), Weighting 30%

- You will do practical activities and quiz style questions under supervision during your applied session. Content will cover topics from Weeks 1 – 9.

Statements on AI

We use the following statements on Generative AI required by the university:

- (A1, A3) Statement on Generative AI required by the university: AI & Generative AI tools may be used in GUIDED ways within this assessment to search for R functions and examples to perform tasks that you specify only. Where used, AI must be used responsibly, clearly documented and appropriately acknowledged (see Learn HQ).

- (A2, Quiz) Statement on Generative AI required by the university: AI & Generative AI tools MUST NOT BE USED within this assessment / task for the following reasons: This whole assessment task requires students to demonstrate human knowledge and skill acquisition without the assistance of AI.

Teaching team Monash Clayton

John Betts (Co-lecturer, Chief Examiner),
Heshan Kumarage (Co-lecturer, Unit Coordinator)

Taking applied sessions:

- Simeon Abrecht, Abdallah Abu-Aisha, Heshan Kumarage, Jeffery Liu, Danushka Liyanage, Prabha Rajagopal, Chris Yun.

Contact list Monash Clayton

- Abdallah: abdallah.abuaisha@monash.edu
- Chris: chris.yun@monash.edu
- Danushka: danushka.liyanage@monash.edu
- Heshan: heshan.kumarage@monash.edu
- Jeffery: jeffery.liu@monash.edu
- John: john.betts@monash.edu
- Prabha: prabha.rajagopal1@monash.edu
- Simeon: simeon.abrecht@monash.edu

Getting support with your studies

Monash offers many support services, including:

- Student Academic Success, Learn HQ
- Learning adviser, Academic English Support
- Academic integrity, Extensions and Spec Con

These can be accessed under the Support tile on Moodle

Special Consideration (Extensions)

- Follow the link for information and to apply for extensions and special consideration under the "Support" tab on Moodle, or via the link below
- https://www.monash.edu/students/admin/assessments/extensions-special-consideration

Introduction to R and RStudio

R

What is R?

Obtaining and installing R?

Using R

Help and References in R

- Help, References you should read

Review of basic statistics using R

- Examples and notes. *We won't go through all these during the lecture*.

What is R?

R is a statistical computing environment and programming language:

- A successor to the S language developed at AT&T Bell Laboratories,
- Initially created by Ross Ihaka and Robert Gentleman University of Auckland (hence 'R'),
- R is now developed R Development Core Team,
- R is freely available under the GNU General Public License (free, open source etc.).

Why we are using R

R:

- Is the defacto platform for data science independent of operating system, problem domain and data type,
- Has a large number of users, active user communities, and many help forums, e.g., Stack Overflow.
- Is free, open source, user-customisable,
- Has thousands of user-contributed packages covering all conceivable applications and data types, for visualisation, machine learning and data science...

Obtaining and installing R

Go to: http://cran.r-project.org/

- Follow the link to download the latest version of R for your operating system (R-4.5.2 as at 24/02/2026),
- Install as usual for your OS (Mac/Win easy),
- Use default directories, if possible, to make installation of RStudio easier,
- Runs from Dock, Launchpad or Start Button,

LHS of main page has Documentation > Manuals

- Click to get: An Introduction to R (R-Release).

Obtaining and installing RStudio

RStudio is an IDE that makes programming in R a lot easier – especially opening and saving files, managing data and variables, and scripting.

Go to: posit.co/

- Install following the instructions for your OS,
- Runs from Launchpad or Start Button.

RStudio also make Shiny for web deployment.

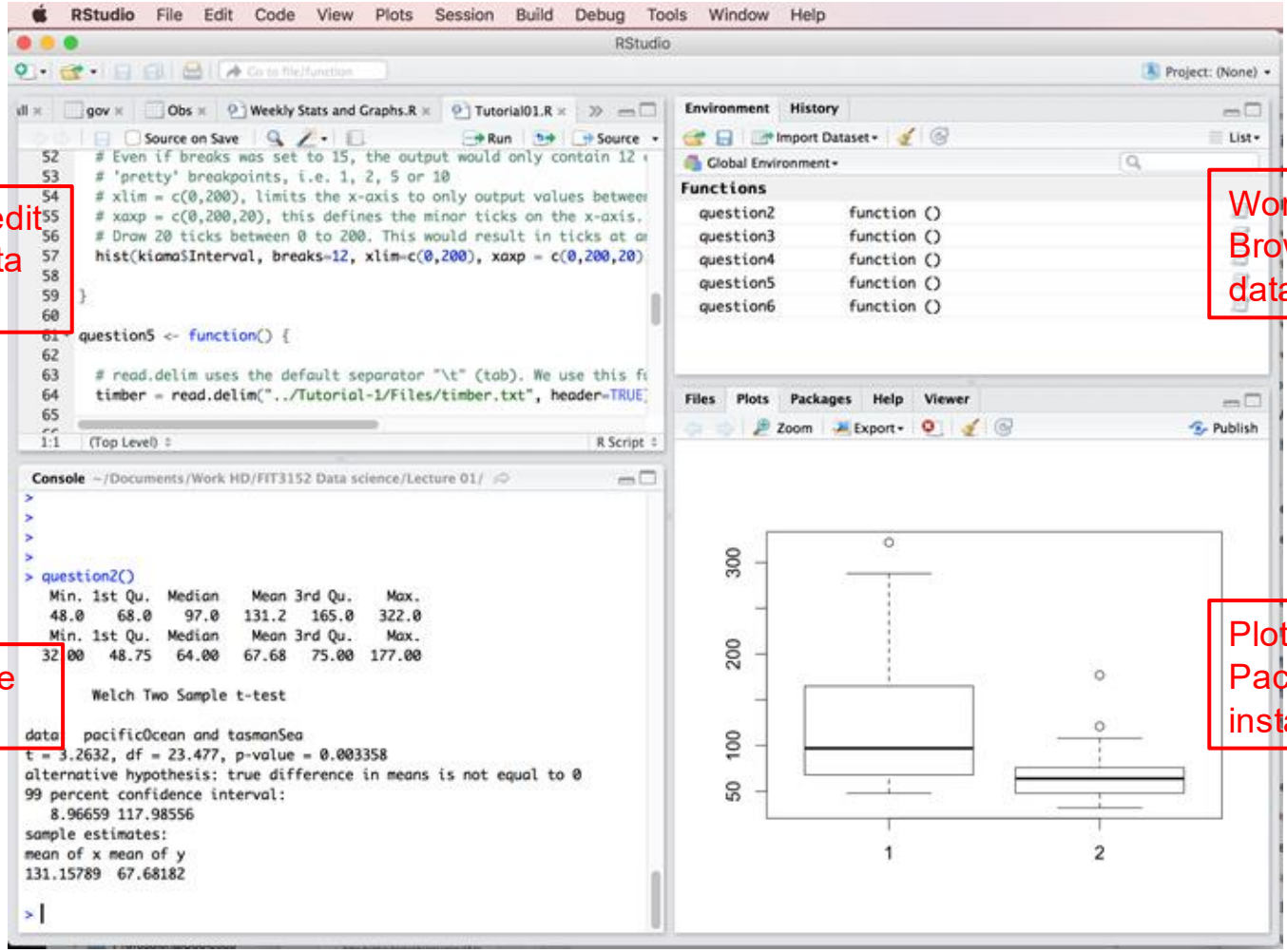
RStudio workspace

Source Editor: edit scripts, view data frames

Console: execute code directly

Workspace Browser: variables, data, history

Plots, Files, Packages installed, Help



Syntax basics

R is command line driven, or using scripts

- > Indicates a new line, Continued lines by +

R is case sensitive

- > TheData is different to thedata

Assignment

- > Use: x < – 5 or x = 5 to assign a value to variable x

Commenting

- > # denotes a comment. Anything on the line after this point is ignored.

Console, Variables, Functions

The R Console shows the command line interface

R can be used for direct calculation and interprets each line as you press (Enter/Return) key, thus

```
> 1 + 4 (enter)
[1] 5
```

Create variables by assigning a value to a name

```
> X = 7
```

Call functions by name

```
> X = sqrt(7)
```

Data Structures

Data is stored in R using data structures (objects) to which functions (methods) are applied.

Array

- Contains data of the same type.
- Vector: 1D, Matrix: 2D, Array: 3^+ Dimensions.

Data Frame

- Row x Column data format – each column is a vector.

List

- An ordered collection of (possibly different) types.

Getting help

You can open help in a browser window, which has links to manuals and search, using

```
> help.start()
```

Alternatively, for help with the 'mean' function

```
> help(mean) # directly open if you know function name
> ? mean # shorthand version of calling help
> ?? mean # lists functions/methods containing 'mean'
```

Searching on the web (Stack Overflow, for example) is a good source of information.

Packages

There are 20,000+ user-contributed packages available. Only a few are installed by default.

To find packages installed

```
> library()
```

Search for packages at http://cran.r-project.org

To install package (+ and dependent packages)

```
> install.packages("package_name")
> library("package_name") #to add it to your library
```

To remove package – e.g., to reclaim memory

```
> remove.packages("package_name")
```

Data input

By hand: (e.g., creating a vector with name X):

```
> X = c(1, 2, 3, 4, 5, 6)
> X <- c(1, 2, 3, 4, 5, 6) # alternative assignment operator
```

Using built in data:

- For example, from Edgar Anderson's Iris Data:

```
> X = iris
> data() # use this to list the built-in data sets
```

Reading a file (and see next slide as well):

```
> X <- read.csv("Toothbrush.csv") #from working directory
```

Reading files

Setting and getting the working directory:

```
> getwd() # get working directory
> setwd("~/desktop") # set working directory
> # alternatively run R from a script to set current directory
> # at the location of the script.
```

Reading csv files:

```
> X <- read.csv("InvestA.csv", header = TRUE)
> # creates a data frame, identifies text header
```

Alternatively, use the “Import Dataset” command from the Environment pane in RStudio.

Review of basic statistics in R

- Descriptive statistics (numbers in one dimension)
- Bivariate data (numbers in two dimensions)
- Estimation and hypothesis testing, Time Series

- *Following slides for reading and reference. We won't go through each example in detail.*

- *Statistics revision notes (reading/activities) have been posted on Moodle under the Week 1 tile for students wanting to refresh their knowledge.*

Descriptive statistics

Problem: create a simple data set, calculate some basic statistics, draw a simple histogram

```
> thedata <- c(0, 0, 1, 5, 7, -2, 11, 0, -4) # create vector
> thedata # print it out to check values
[1]  0  0  1  5  7 -2 11  0 -4
> mean(thedata)  # calculate mean
[1] 2
> sd(thedata)  # calculate standard deviation
[1] 4.743416
> hist(thedata)  # draw a basic histogram (not shown)
```

Descriptive statistics

Some other familiar functions in R. These can be applied to vectors or data frames.

```
> var(x) # for variance
> median(x) # median
> quantile(x, probs) # e.g., quartiles,  probs = [0,1].
> range(x) # range
> sum(x) # sum
> min(x) # minimum
> max(x) # maximum
```

See Quick-R for more R functions

https://www.statmethods.net/management/functions.html

Descriptive statistics

Data are simulated returns (FV) from 6 different types of investments (Groups):

- InvestA is a single column of data with another column as an index.

InvestA.csv

Group	FV
1	809.34
1	166.46
1	711.33
1	870.33
1	758.56
…	…
…	…
2	716.72
2	800.29
2	748.75
2	758.11
…	…

Identifying columns in a data frame

To identify the columns in a data frame you can:

- Refer to them by name: dataframe$column.
- This lets you specify the data column and the grouping variable.

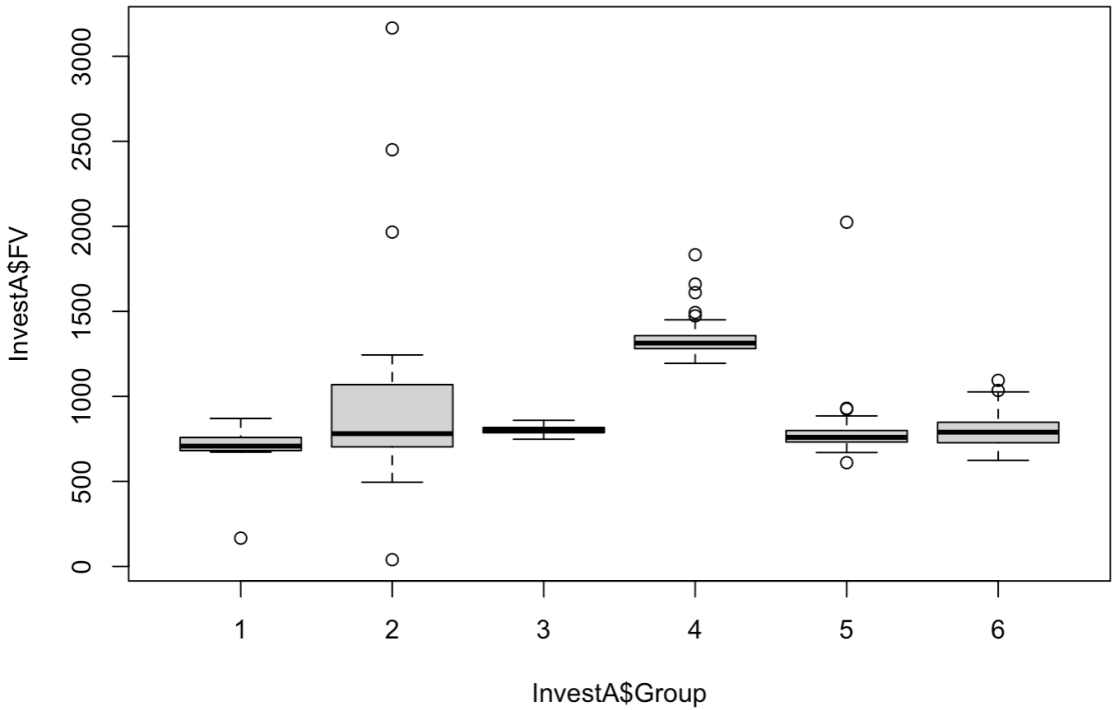
Descriptive statistics

Problem: compare several groups stored as a single indexed column (using the "by" function).

```
> InvestA = read.csv("InvestA.csv")
> by(InvestA$FV, InvestA$Group, FUN = mean)
InvestA$Group: 1
[1] 689.3454
------------------------------------------------
InvestA$Group: 2
[1] 874.0045
------------------------------------------------
... (truncated)
```

Boxplot

```
> boxplot(InvestA$FV ~ InvestA$Group)
> # not perfect but more on graphics next lecture
```



Bivariate data

The data:

- *Choice* magazine tested a sample of toothbrushes and made a summary of price and function. Are these two factors related?

Toothbrush.csv

Price	Function
3.95	65.1
2.96	78
2.95	72
0.66	40
0.69	57
3.2	61
1.08	49
3.69	76
…	…

Data from Selvanathan Australian Business Statistics (Abridged 4th Ed)

Bivariate data

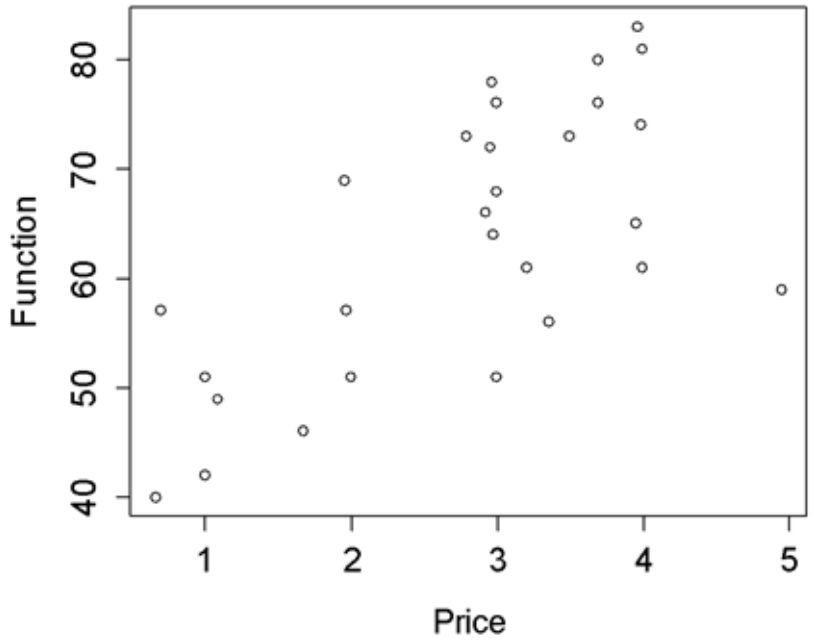
Problem: analyse the relationship between price and function.

- Read the data and create a data frame
  - > Toothbrush <- read.csv("Toothbrush.csv")
- To calculate the least squares correlation use:
  - > cor(Toothbrush) *# setting x or y not important for cor*

```
             Price     Function
Price    1.0000000 0.6645614
Function 0.6645614 1.0000000
```

Scatterplot

```
> plot(Toothbrush)
> # the default plot reading the first column for X axis
> # and second column for Y axis
```



Attaching a data frame

To simplify your commands, you can “attach” your data frame:

- The ‘attach’ function lets you call columns in a data object by name without having to specify the data frame. This assumes the column name is unique among all currently available data frames.

```
> attach(Toothbrush)
```

- Scatterplot using Price and Function directly:

```
> plot(Price, Function)
```

Bivariate data

Problem: calculate the regression equation cont.

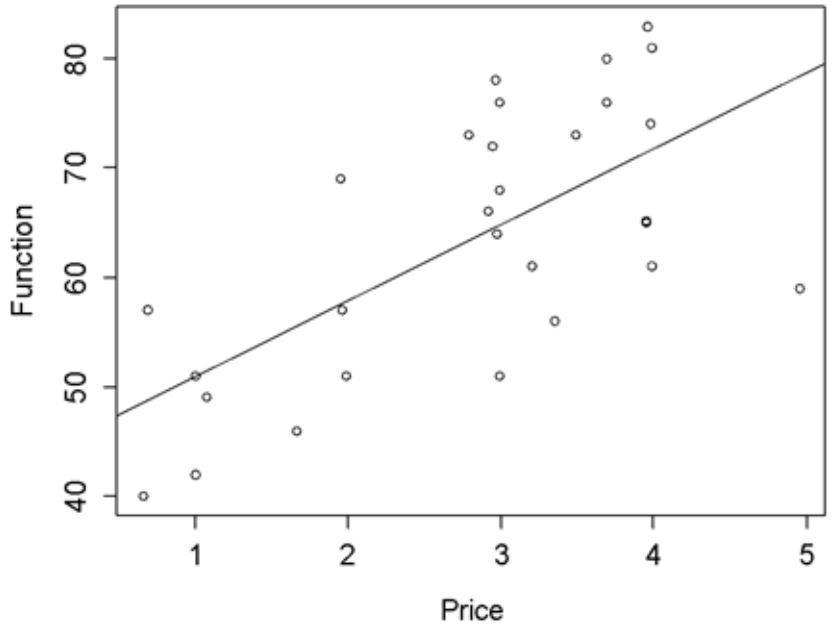
- To calculate the regression equation, define a variable 'fitted' and use linear model (lm) function.
  - > fitted = lm(Function ~ Price)
  - > fitted

```
Call: lm(formula = Function ~ Price)
Coefficients: (Intercept) Price
                  44.020      6.942
```

- Now overplot the fitted model on scatterplot using the gradient and intercept from fitted model.

Scatterplot + regression line

```
> plot(Price, Function)
> abline(fitted) # overplotting, using gradient and
> # intercept as first two items of the “fitted” list
```



Estimation/Hypothesis testing

The data:

- The number of claims processed by two workers is given below. For convenience create two vectors:
  - > Workers <- read.csv("Workers.csv")
  - > attach(Workers)

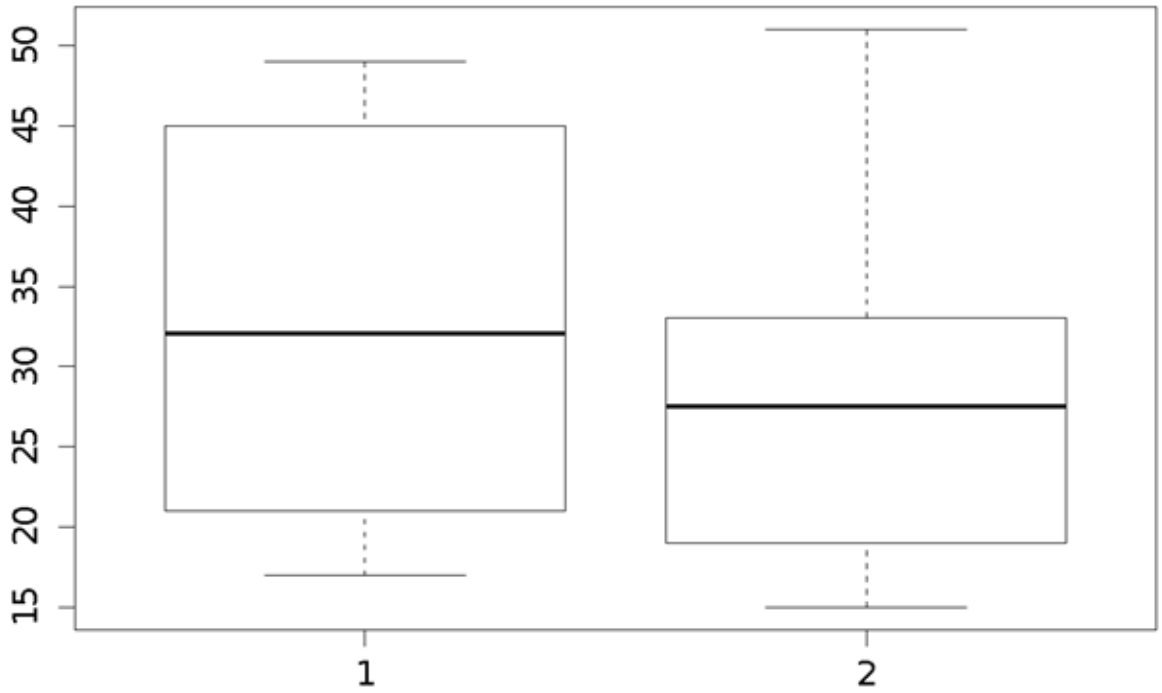
Workers.csv

WorkerA	WorkerB
23	33
45	23
21	19
22	51
17	32
42	15
45	
41	
49	
19	

Estimation/Hypothesis testing

Quick comparison of data using a boxplot:

> boxplot(WorkerA, WorkerB)



Estimation/Hypothesis testing

Problem 1:

- Calculate the confidence interval for the average number of claims processed by Worker A.

Problem 2:

- Can we conclude that worker A processes more claims than Worker B?

Estimation/Hypothesis testing – 1

Perform a 't.test' (with alternative that mean ≠ 0) to generate confidence interval.

```
> t.test(WorkerA)
One Sample t-test data: WorkerA
t = 7.93, df = 9, p-value = 2.374e-05
alternative hypothesis: true mean not equal to 0
95 percent confidence interval: 23.1574 41.6426
sample estimates: mean of x 32.4
```

- To specify confidence level as a parameter other than default (95%), for example to get a 55% CI:

```
> t.test(WorkerA, conf.level = 0.55)
```

Estimation/Hypothesis testing – 2

Perform a ‘t.test’ to determine whether the means are different:

```
> t.test(WorkerA, WorkerB)
Welch Two Sample t-test
data: WorkerA and WorkerB
t = 0.5333, df = 10.634, p-value = 0.6048
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
-11.21422 18.34755 sample estimates:
mean of x mean of y
32.40000 28.83333
```

t.test: syntax

From the help file:

- Description

  Performs one and two sample t-tests on vectors of data.

- Usage

```
t.test(x, ...)
## Default S3 method:
t.test(x, y = NULL,
alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

Time series analysis

The data:

- The value of food sales in Australia 2014 – 2020. From: From ABS: 8501.0 Retail Trade, Australia

Food Retail 2014-2020.csv

YearMonth	FoodRetailM
Jan-14	9701.6
Feb-14	8667.9
Mar-14	9524.6
Apr-14	9223.9
May-14	9386
Jun-14	8977.5
Jul-14	9393.3
Aug-14	9582.4
...	...

- Challenge: investigate the main components of the data.

Time series analysis

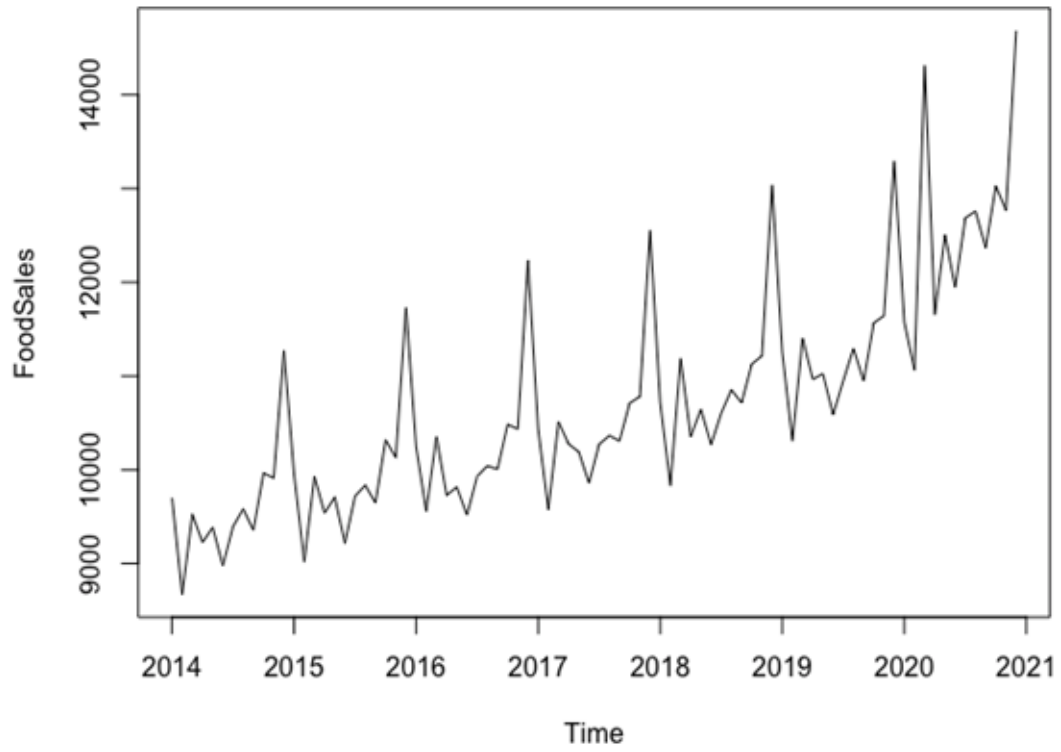
Problem: read the data and declare as class ts:

```
> Food <- read.csv("Food Retail 2014-2020.csv")
> attach(Food)
> FoodSales <- ts(FoodRetailM, frequency=12,
  start=c(2014,1))
> FoodSales
          Jan      Feb      Mar      Apr  ...
2014  9701.6   8667.9   9524.6   9223.9  ...
```

Time series analysis

Problem: plot the time series

```
> plot(FoodSales)
```

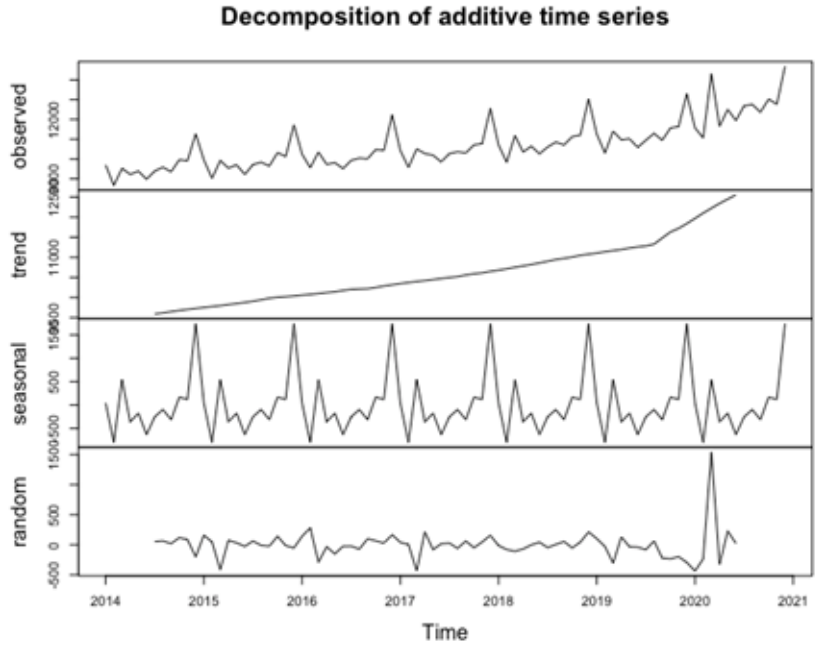


Time series analysis

Problem: decompose the time series

> decomp <- decompose(FoodSales)

> plot(decomp) # object stores components of time series



Reading: R

Essential (AITR)

*Note this is updated for each new release of R.

- *An Introduction to R*, W. N. Venables, D. M. Smith and the R Core Team,

https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Excellent (ATHR)

- *A Tiny Handbook of R*, M. Allerhand, Springer, 2011 (Online access via the Monash Library)

Useful on-line reference (Quick-R)

https://www.statmethods.net/management/functions.html

https://www.statmethods.net/about/sitemap.html

Reading: Recommended

- G. James, D. Witten, D, T. Hastie, R. Tibshirani. (2021) An Introduction to Statistical Learning 2nd Ed. Springer. *Online access via Library.*
- F. Provost and T. Fawcett. (2013) Data Science for Business. O'Reilly Media, Inc. *Online via Library.*
- H. Wickham, M. Rundel, G. Gromelund. (2023) R for Data Science 2nd Ed. O'Reilly Media, Inc. Also available from: https://r4ds.hadley.nz/.
- P.-N. Tan, M. Steinbach, V. Kumar. (2006) Introduction to Data Mining. Addison-Wesley.

Reading: More useful references

- A (very) short introduction to R, Paul Torfs & Claudia Brauer

  https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
- R Reference Card

  https://cran.r-project.org/doc/contrib/Short-refcard.pdf

What to do this week

Download and install R and RStudio.

Download and read:

- AITR – read Chapters 1 & 2.
- ATHR – read up to Page 20.
- Statistics revision lecture notes, R Reference Card.

Attempt:

- Attempt any questions in these notes and the pre-Applied Session 1 activities especially!
- Reminder: Applied sessions commence next week!