FIT3152 Data analytics – Lecture 1

Introduction to Data Science

- Recent examples
- Common themes

Introduction to FIT3152

- Unit objectives, outline, assessment, unit management

Review of basic statistics using R

Advertising…

UROP Sponsor

CSL

Undergraduate Research Opportunities Program (UROP)

Now recruiting students majoring in areas such as:

- **Biomedicine**
- **Bioinformatics**
- **Data/Computer Science**
- **Maths and Stats**
- **Bioengineering**

- Are you looking for **WORK EXPERIENCE** in a research team?
- UROP provides casual paid employment for undergraduate students (2nd year+) in research teams in Melbourne
- Applications **open 7 March** and **close 18 March 2022** for projects starting in the Winter break
- Go to www.csiro.au/UROP for more information and to apply

Questions? Email:
UROP@CSIRO.au

Quick information:

Lecture times:

- Monash Clayton – Wednesday at 12:00pm.

Tutorials commence Week 2.

Resources:

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

Forum:

- We're using Ed Discussion, join with this link:
- https://edstem.org/au/join/nzRN9C

Clayton lectures

FIT3152 lectures are via Zoom

- The link is: https://monash.zoom.us/ or go to https://monash.zoom.us/join and enter meeting ID: 898 6354 8300 and passcode: 931829
- The lecture will be recorded, see Class Streaming tile.

Tutorials

- Note: we run Tutorials. Allocate+ calls them labs.
- Tutorials are on campus for students who can attend. Tutorials will be bring-your-own-device.
- Tutorials are on Zoom for off campus students. Links to Zoom tutorials are under the Class Streaming tile.

Student Quiz

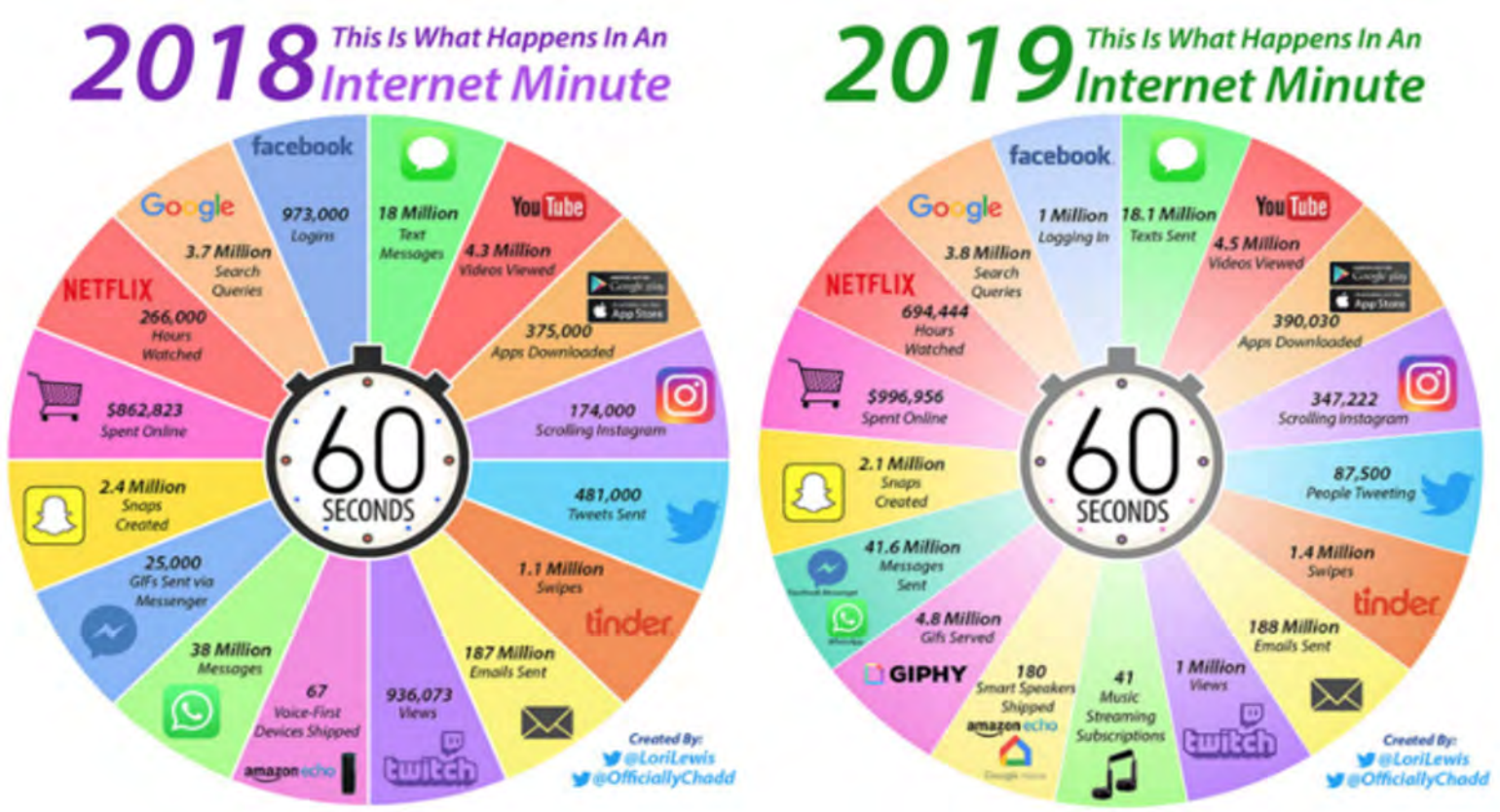
We have set up a Moodle Quiz (not graded) for you to tell us a little about yourself.

There is a link under the Week 1 tile.

- Results are confidential and will help us with planning.
- Questions ask things like:
- The degree you're enrolled in, the computer operating system you use, programming experience, general interests, and what you hope to get out of the unit (most important).

Please try and complete this over the next few days.

A sea of data: 2018/2019



https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

A sea of data: 2021!

https://www.visualcapitalist.com/

Tutorial Activity (a)

Compare the figures on the two previous slides showing Internet activity over 2018/2019 and 2021. Answer the following:

- What are the trends – that is, what types of online activities are increasing in prevalence?
- Looking at a particular activity, what types of data could be collected?
- What could that data be used to study?
- What changes in Internet usage might be due to COVID-19?

What is data science

From Wikipedia:

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is related to data mining, machine learning and big data.
- Data science is a “concept to unify statistics, data analysis, machine learning and their related methods” in order to “understand and analyze actual phenomena” with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. ...
- https://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.
- See if you can think what these qualities are…

Criminal investigation

Cleo Smith: How Australian police found the missing four-year-old

Police released a photo of Cleo Smith after her rescue

Early on 3 November, police smashed their way into a house in the Australian town of Carnarvon, where they found a four-year-old girl who had been missing for 18 days.

Cleo Smith disappeared from her family's tent at a campsite near Carnarvon on 16 October, triggering a massive search operation.

https://www.bbc.com

Criminal investigation

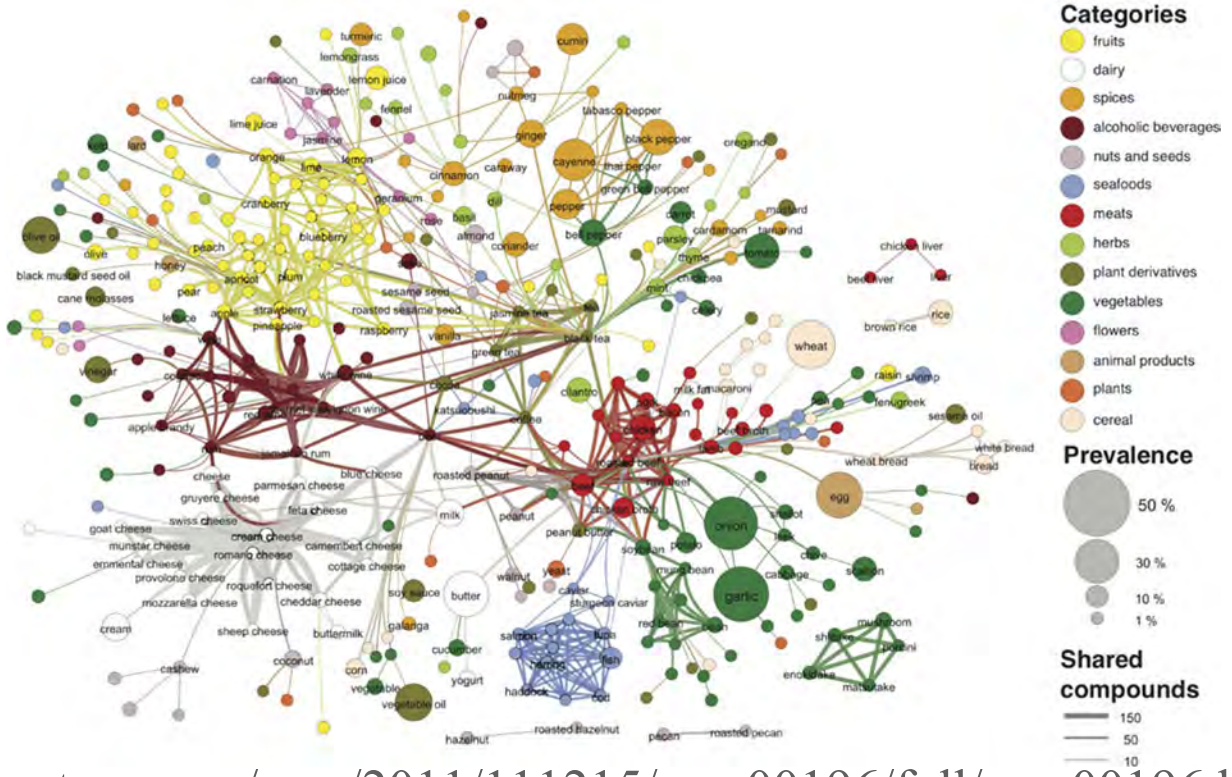
'Needle in a haystack' "We had to sift through a lot of information. The statements of the 100 people who were at the campsite, CCTV footage, data from phones..." he said.

Detective Supt Wilde said they were able to "build a picture of who was supposed to be there and who was not supposed to be there" by piecing together all the information they received and by "placing people in certain locations at certain times".

The Australian newspaper reported the breakthrough came when police traced a mobile phone number to a phone tower near the campsite around the time of Cleo's abduction. This allegedly focused their attention on Mr Kelly.

Food networks

Flavor network and the principles of food pairing



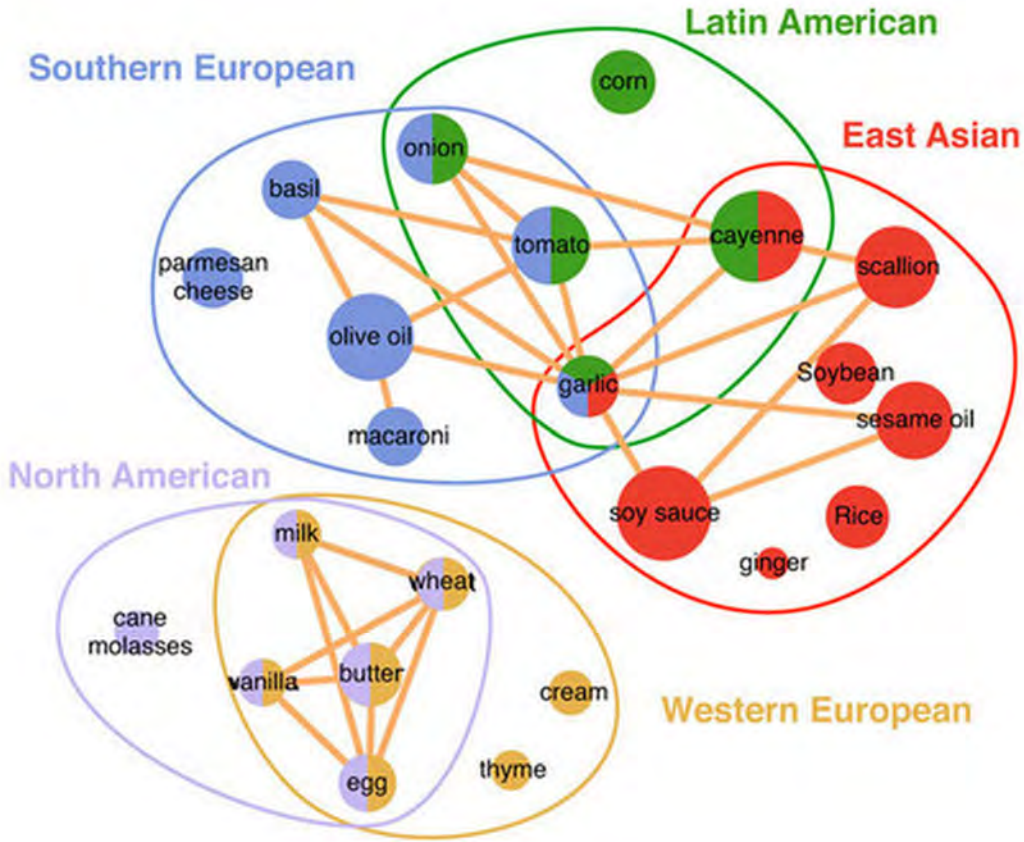
http://www.nature.com/srep/2011/111215/srep00196/full/srep00196.html

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)...

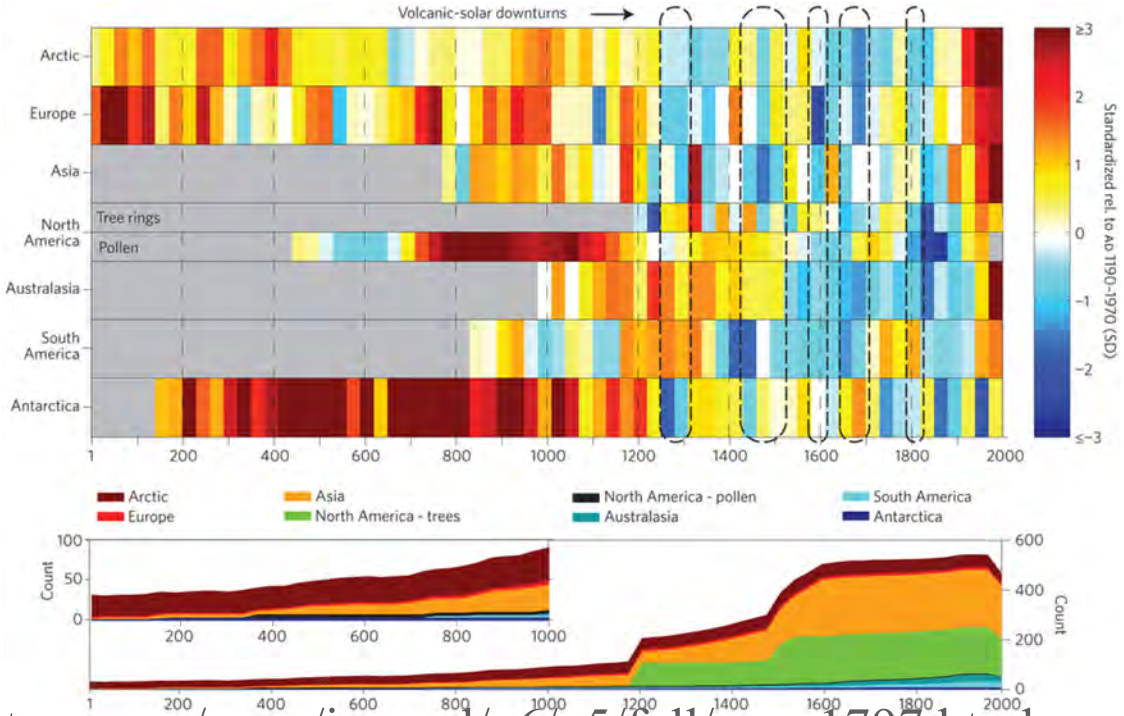
Food networks

- Co-occurrence of major ingredients in 5 cuisines



Climate change

Continental-scale temperature variability during the past two millennia



http://www.nature.com/ngeo/journal/v6/n5/full/ngeo1797.html

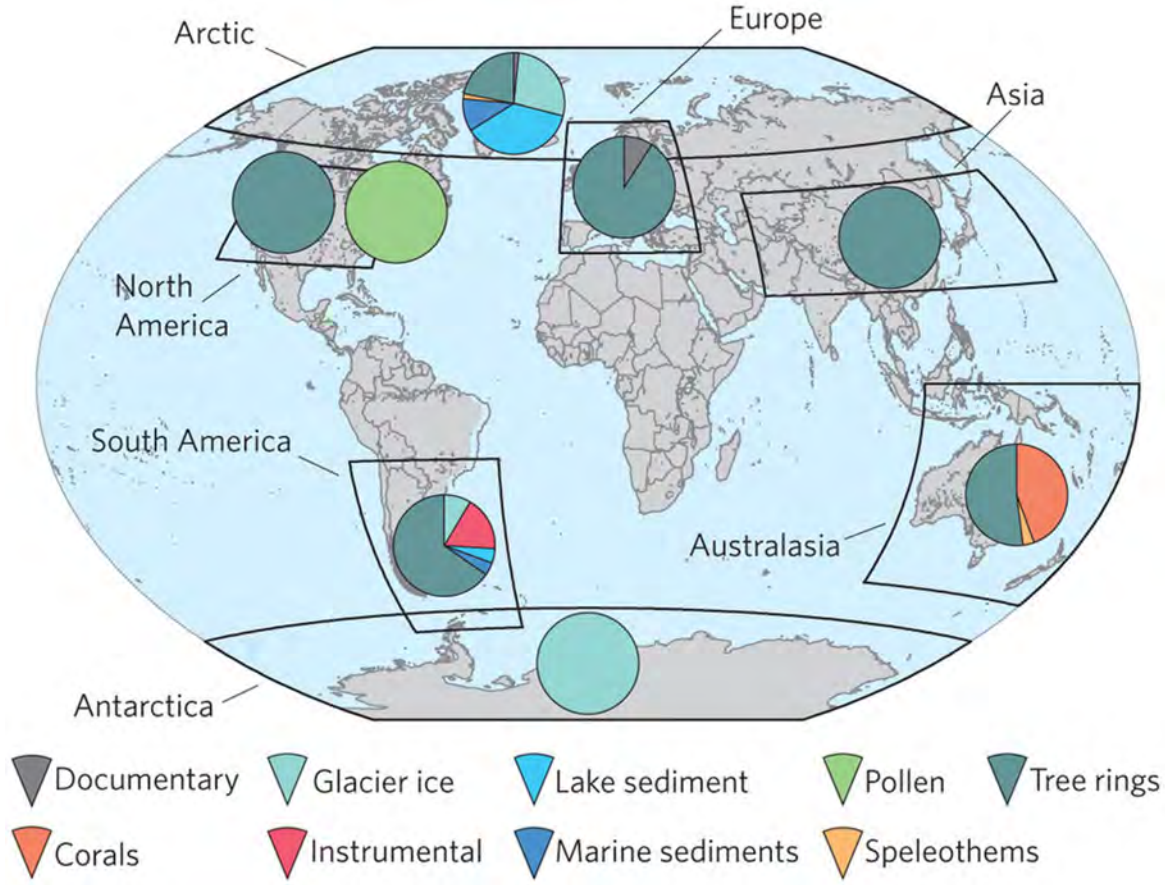
Climate change

...

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 ...

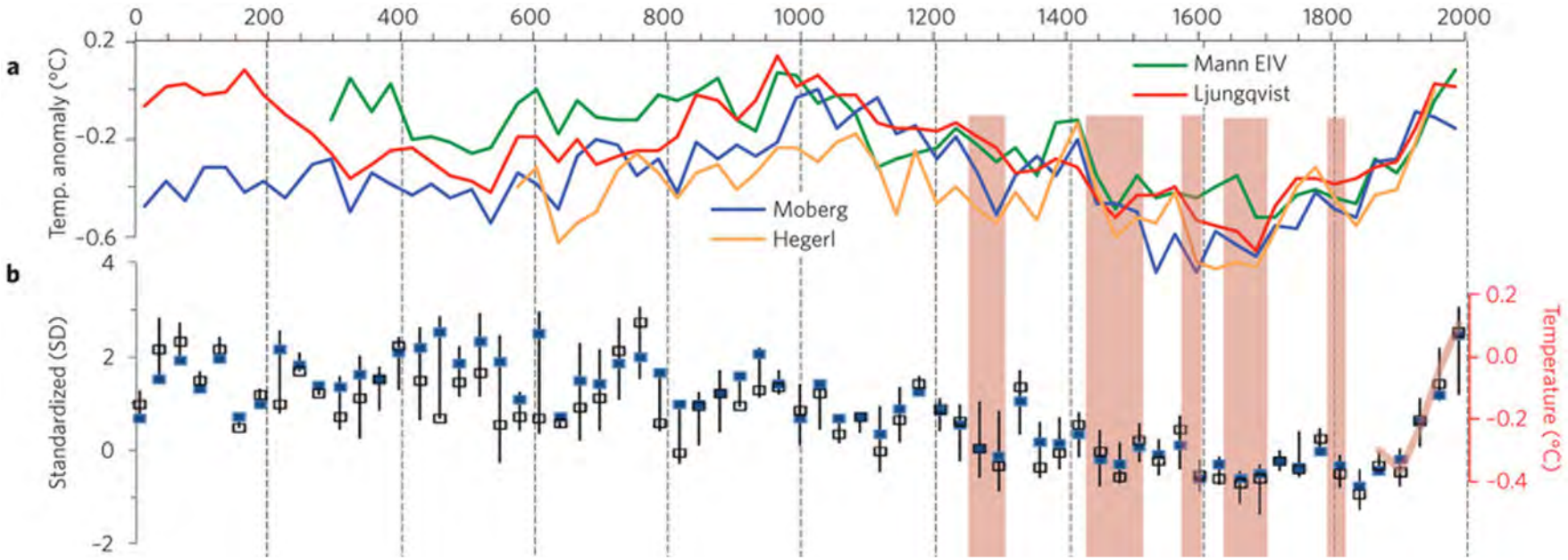
Climate change

- 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)

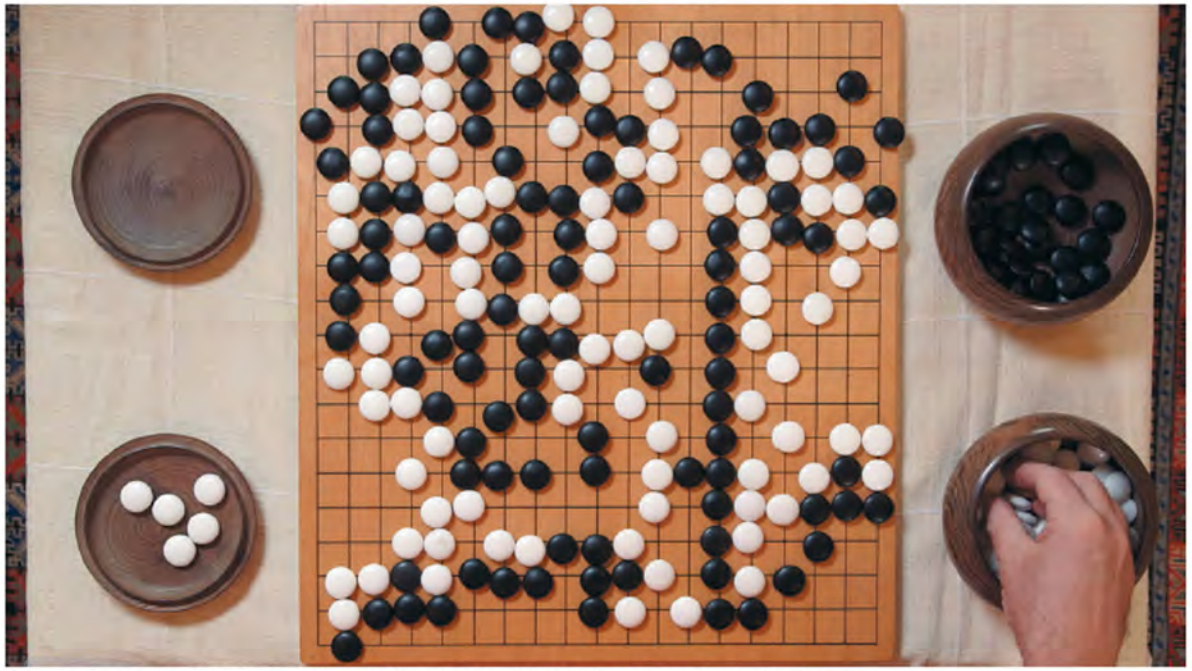
Text analysis

...

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....

Deep learning



Go, a complex game popular in Asia, has frustrated the efforts of artificial-intelligence researchers for decades.

ARTIFICIAL INTELLIGENCE

Google masters Go

Deep-learning software excels at complex ancient board game.

https://www.nature.com/news/google-ai-algorithm-masters-ancient-game-of-go-1.19234

Deep learning

In China, Japan and South Korea, Go is hugely popular and is even played by celebrity professionals.

But the game has long interested AI researchers because of its complexity. The rules are relatively simple: the goal is to gain the most territory by placing and capturing black and white stones on a 19 × 19 grid.

But the average 150-move game contains more possible board configurations — 10^{170} — than there are atoms in the Universe, so it can't be solved by algorithms that search exhaustively for the best move. ...

Deep learning

...

To interpret Go boards and to learn the best possible moves, the AlphaGo program applied deep learning in neural networks – brain-inspired programs in which connections between layers of simulated neurons are strengthened through examples and experience.

It first studied 30 million positions from expert games, gleaning abstract information on the state of play from board data, much as other programmes categorize images from pixels ...

AI for COVID-19 detection

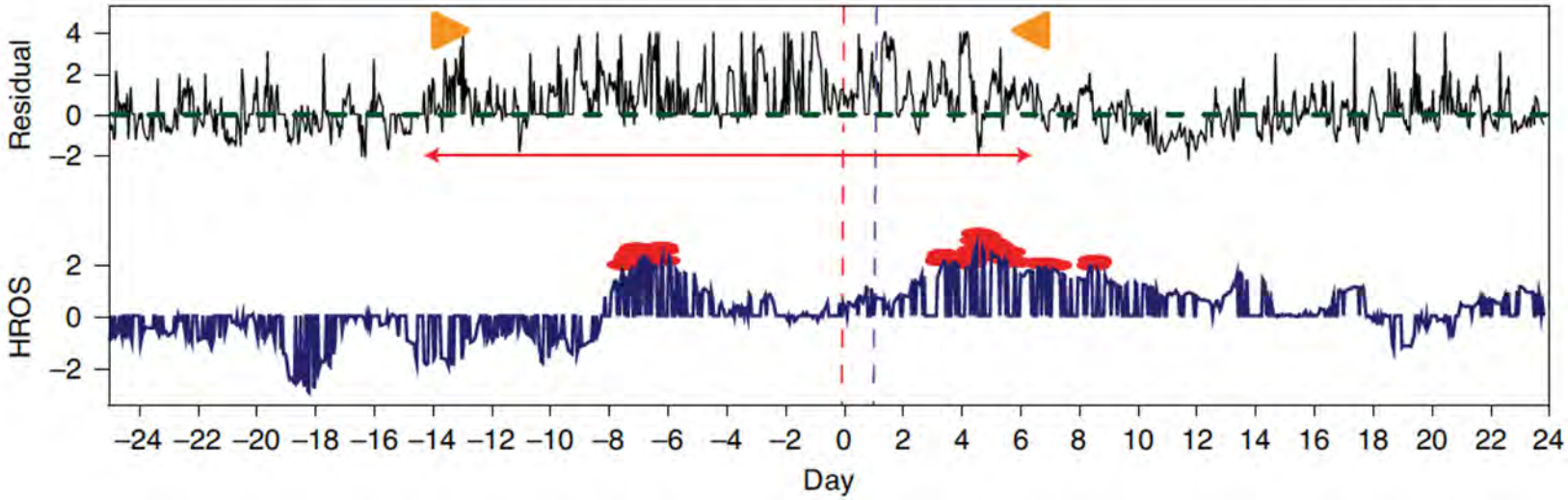
Smartwatch data help detect COVID-19

> **Obtaining longitudinal physiological data via commonplace wearable devices2, typically worn on the wrist, may offer a convenient means of detection. Self-reported symptoms can be used to construct relatively simple models for the identification of COVID-19, and data from wearables may similarly be used to identify viral respiratory illnesses.**

https://www.nature.com/articles/s41551-020-00659-9.pdf

AI for COVID-19 detection

Smartwatch data help detect COVID-19



Heart-rate metrics for an individual before COVID-19 infection and during illness. The red dashed line indicates the day of symptom onset and the purple dashed line the date of diagnosis.

https://www.nature.com/articles/s41551-020-00659-9.pdf

AI and ethics

Bias detectives: the researchers striving to make algorithms fair

As machine learning infiltrates society, scientists are trying to help ward off injustice.

In 2015, a worried father asked Rhema Vaithianathan a question that still weighs on her mind. A small crowd had gathered in a basement room in Pittsburgh, Pennsylvania, to hear her explain how software might tackle child abuse. Each day, the area's hotline receives dozens of calls from people who suspect that a child is in danger; some of these are then flagged by call-centre staff for investigation. But the system does not catch all cases of abuse. Vaithianathan and her colleagues had just won a half-million-dollar contract to build an algorithm to help.

https://www.nature.com/articles/d41586-018-05469-3

AI and ethics

...

Computer calculations are increasingly being used to steer potentially life-changing decisions, including which people to detain after they have been charged with a crime ...

These tools promise to make decisions more consistent, accurate and rigorous. But oversight is limited: no one knows how many are in use. And their potential for unfairness is raising alarm. In 2016, for instance, US journalists argued that a system used to assess the risk of future criminal activity discriminates against black defendants. ...

The next trend: Ubiquitous AI

AI Here, There, Everywhere

The New York Times

Craig S. Smith

February 23, 2021

Researchers anticipate increasingly personalized interactions between humans and artificial intelligence (AI), and are refining the largest and most powerful machine learning models into lightweight software that can operate in devices like kitchen appliances. Privacy remains a sticking point, and scientists are developing techniques to use people's data without actually viewing it, or protecting it with currently unhackable encryption. Some security cameras currently use AI-enabled facial recognition software to identify frequent visitors and spot strangers, but networks of overlapping cameras and sensors can result in ambient intelligence that can constantly monitor people. Stanford University's Fei-Fei Li said such ambient intelligence "will be able to understand the daily activity patterns of seniors living alone, and catch early patterns of medically relevant information," for example.

https://www.nytimes.com/

Ubiquitous AI

Some ways AI is being used:

- Smart devices and sensors to control temperature, lights, comfort around the home;
- Passive monitoring of people for falls or accidents;
- Chatbots to “help” you with enquiries,
- Streaming services (Netflix, Spotify…) are learning your preferences...
- AI assisted music composition,
- Privacy is an important concern.

https://www.nytimes.com/

Ubiquitous AI

My Spotify recommendations: songs that sound like ones I've liked.

Data Science: many other applications:

Sports analytics

Data analytics in this area is exploding! Some areas currently receiving a lot of interest are:

- Individual and team performance tracking,
- Wearable technologies and video tracking,
- Optimizing team composition,
- Analysis of supporter and fan engagement,
- Training optimization, injury prevention,
- Gambling: customer analysis, team analytics.

Sports analytics

Sports Tech World Series partners (as of 2020).

https://sportstechworldseries.com/directory/

Combatting terrorism

The social network of the 9-11 terrorists

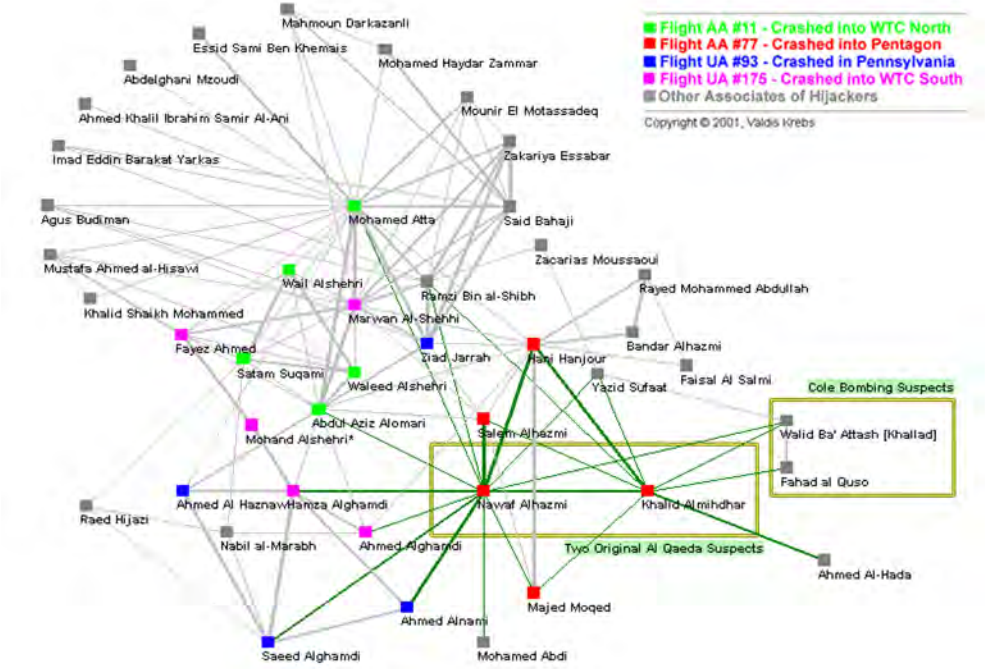
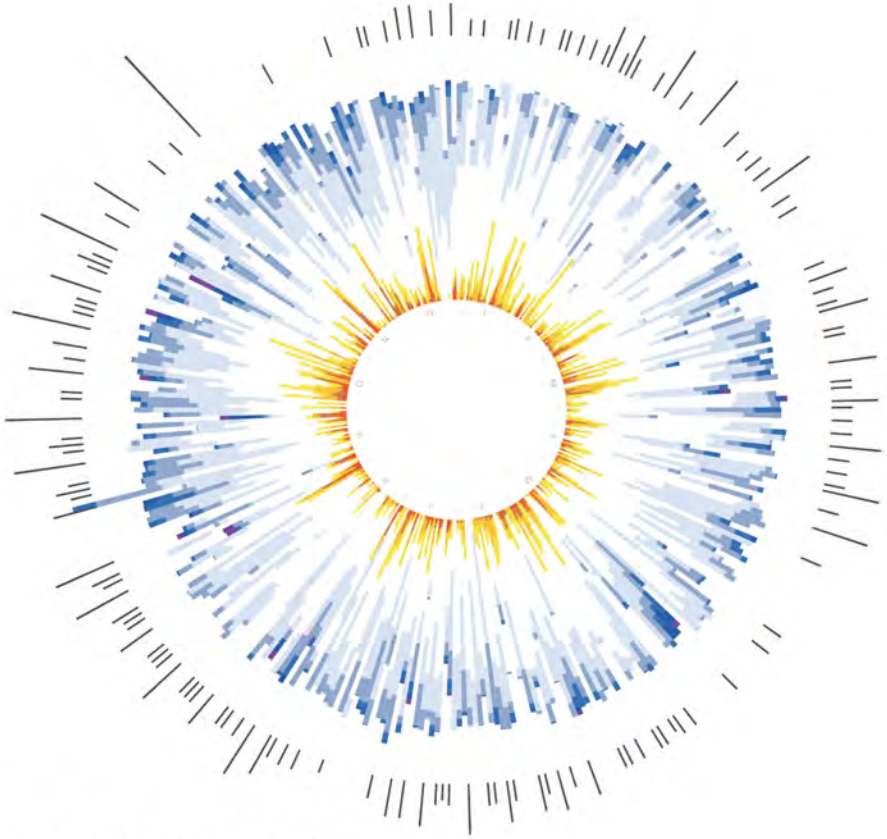


Figure 3 - All Nodes within 2 steps / degrees of original suspects

http://orgnet.com/tnet.html

Personal analytics



The Healthiest Year Of My Life

Diabetic Charts A Year's Worth Of His Health Data

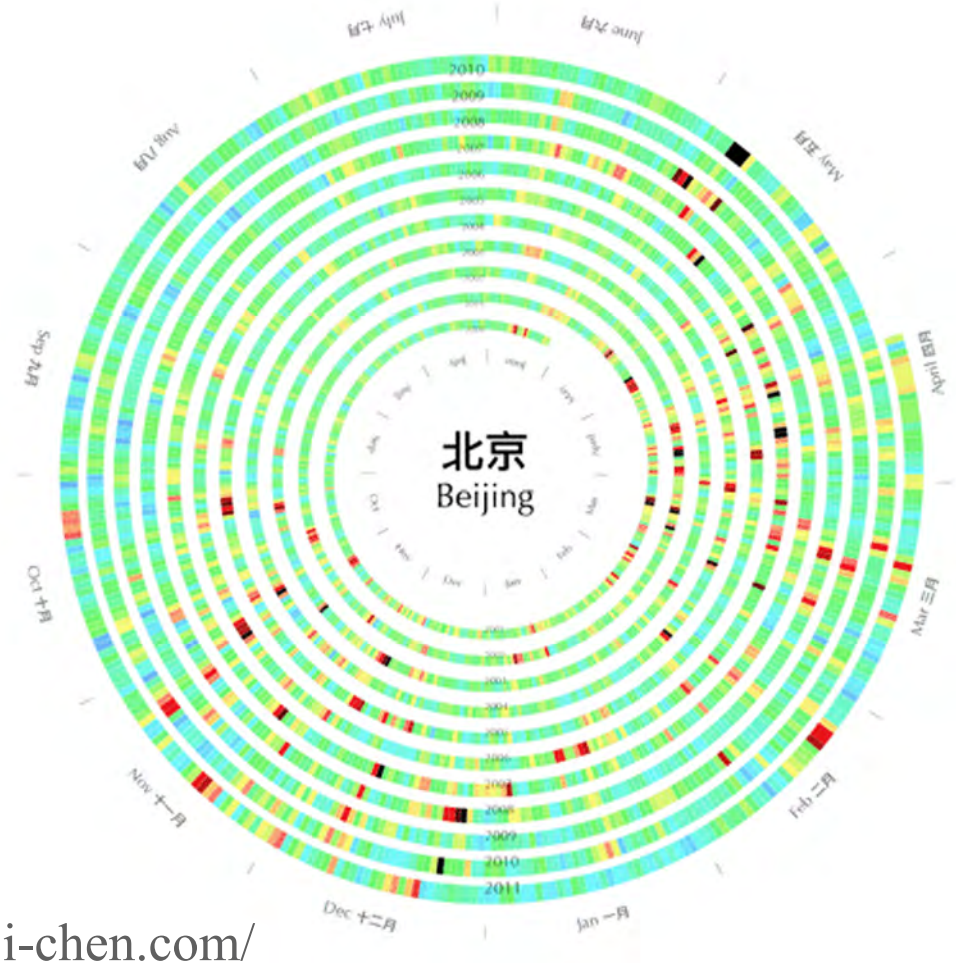
One of our 15 favorite recent data visualizations

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.

https://www.popsci.com/diabetic-charts-years-worth-his-health-data/?dom=psc

Beautiful data visualization



https://www.xiaoji-chen.com/

Data science: some common themes

Previous examples illustrate:

- Complex problems, of societal concern.
- Large data sets, multiple data sets (mashups), messy, incomplete, heterogeneous, non-traditional, open data.
- Often using data repositories created for another purpose (food network): One description of Data Science is making a product out of data…
- Data collection and analysis on a scale that would have been unthinkable 15 years ago.
- Use of high-quality graphics for communicating results!

Tutorial 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 is the outcome?
- How might you use this data to investigate another aspect of (human) activity?
- Present your findings in Tutorial 1.

Data science: for business

Customer analytics

- Website tracking, click to sales conversion, marketing and pricing strategy, social media sentiment analysis, demographic information, location data and traffic monitoring, app use statistics, tailored products…

Operations

- Factors affecting demand, supply chain data, item tracking, sensor data, self regulating processes (automatic systems, pre-emptive repairs), fraud detection, productivity analysis, human resource management, …

Data science: for business

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

- 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

Science and medicine:

- Search for habitable planets, weather forecasting, DNA sequencing and disease genomics, automatic classification, biometrics (identification by physical characteristics), COVID-19 prediction/modelling, …

Arts, culture and society

- Social networks: LinkedIn, Facebook, Twitter, Instagram etc., national security surveillance, ...
- Data journalism, data artists, ...

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

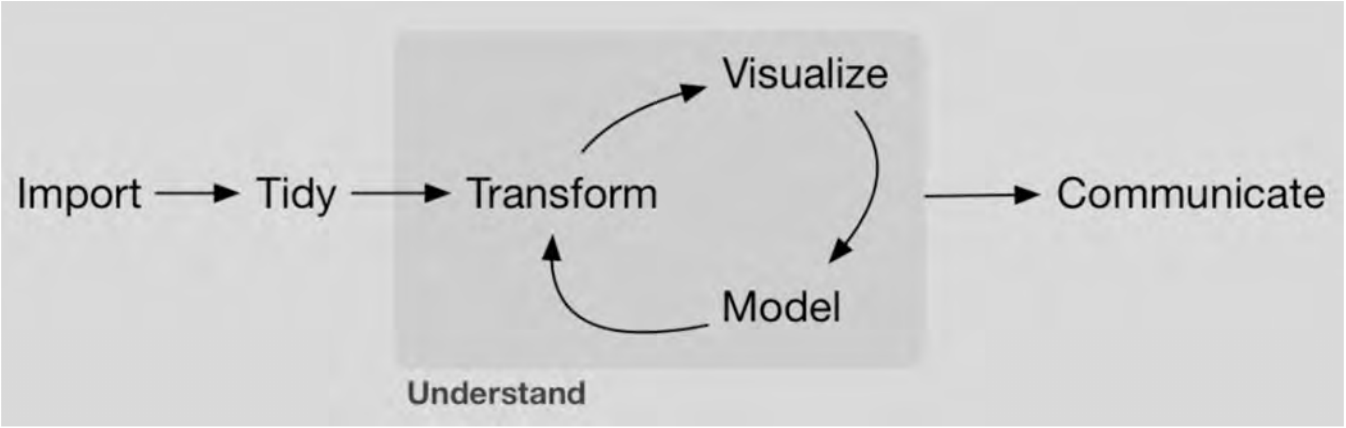
Some necessary technical skills include:

- Statistical analysis,
- Machine learning,
- Programming (e.g., R, Python, Java …),
- 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:



https://r4ds.had.co.nz/

kaggle

Hosts data science competitions:

- Their motto is “turning data science into a sport,”
- You can view their 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,
- https://www.kaggle.com/ for details.

FIT3152 Data analytics

Overview

Unit objectives

What the course is trying to achieve:

- We are concentrating 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.
- 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 we cover descriptive analysis, not just ML).
- FIT3179 Data visualization (but we have broader range of topics).

Unit outline (week-by-week)

- Clayton lecture Wednesday 12:00 – 2:00 pm (AEDT).
- Tutorials start Week 2 and follow lecture by a week.

Week Starting	Lecture	Topic	Tutorial	A1	A2
28/2/22	1	Intro to Data Science, review of basic statistics using R	...		
7/3/22	2	Exploring data using graphics in R	T1		
14/3/22	3	Data manipulation in R	T2	Released	
21/3/22	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
28/3/22	5	Network analysis	T4		
4/4/22	6	Regression modelling	T5		
11/4/22	7	Classification using decision trees	T6		
		Mid-semester Break		Submitted	
25/4/22	8	Naïve Bayes, evaluating classifiers	T7		Released
2/5/22	9	Ensemble methods, artificial neural networks	T8		
9/5/22	10	Clustering	T9		
16/5/22	11	Text analysis	T10		Submitted
23/5/22	12	Review of course, Exam preparation	T11		

Assessment details

Assignment 1

- Individual work, (20%). Report due Friday, mid-semester break, (22/04/2022).

Assignment 2

- Individual work, (20%) Report due Friday, Week 11, (20/05/2022).

Examination

- Individual work, (60%) During Semester 1 exam period.

Unit Management (Clayton)

Lecturer: John Betts (CE)

Tutors:

- Abdallah Abu-Aisha, Anthony Wong, Chris Yun, Danushka Liyanage, Heidi Quah, Heshan Kumarage, Jeffery Liu, Karina Islas Rios, Priscila Grecov, Saher Manseer.

Lecture and tutorial notes:

- Lecture notes (excluding class questions) and any pre-lecture activities will be posted a few days prior.
- Tutorial worksheets contain pre-tutorial questions for you to attempt prior to the tute.

Contact list Monash Clayton

- John: john.betts@monash.edu
- Abdallah: abdallah.abuaisha@monash.edu
- Anthony: anthony.wong@monash.edu
- Chris: chris.yun@monash.edu
- Danushka: danushka.p…e1@monash.edu
- Heidi: heidi.quah@monash.edu
- Heshan: heshan.kumarage@monash.edu
- Jeffery: jeffery.liu@monash.edu
- Karina: karina.islasrios@monash.edu
- Priscila: priscila.grecov@monash.edu
- Saher: saher.manaseer@monash.edu

R

Review of basic statistics using 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., Stackoverflow.
- 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... ,
- *One drawback: a steep learning curve!*

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-4.1.2 as at 21/02/2022),
- 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 running R a lot easier – especially opening and saving files, managing data and variables, and scripting.

Go to: https://www.rstudio.com/

- Select Download > RStudio Desktop,
- Install as usual for your OS,
- Runs from Launchpad or Start Button.

RStudio also make Shiny for web deployment.

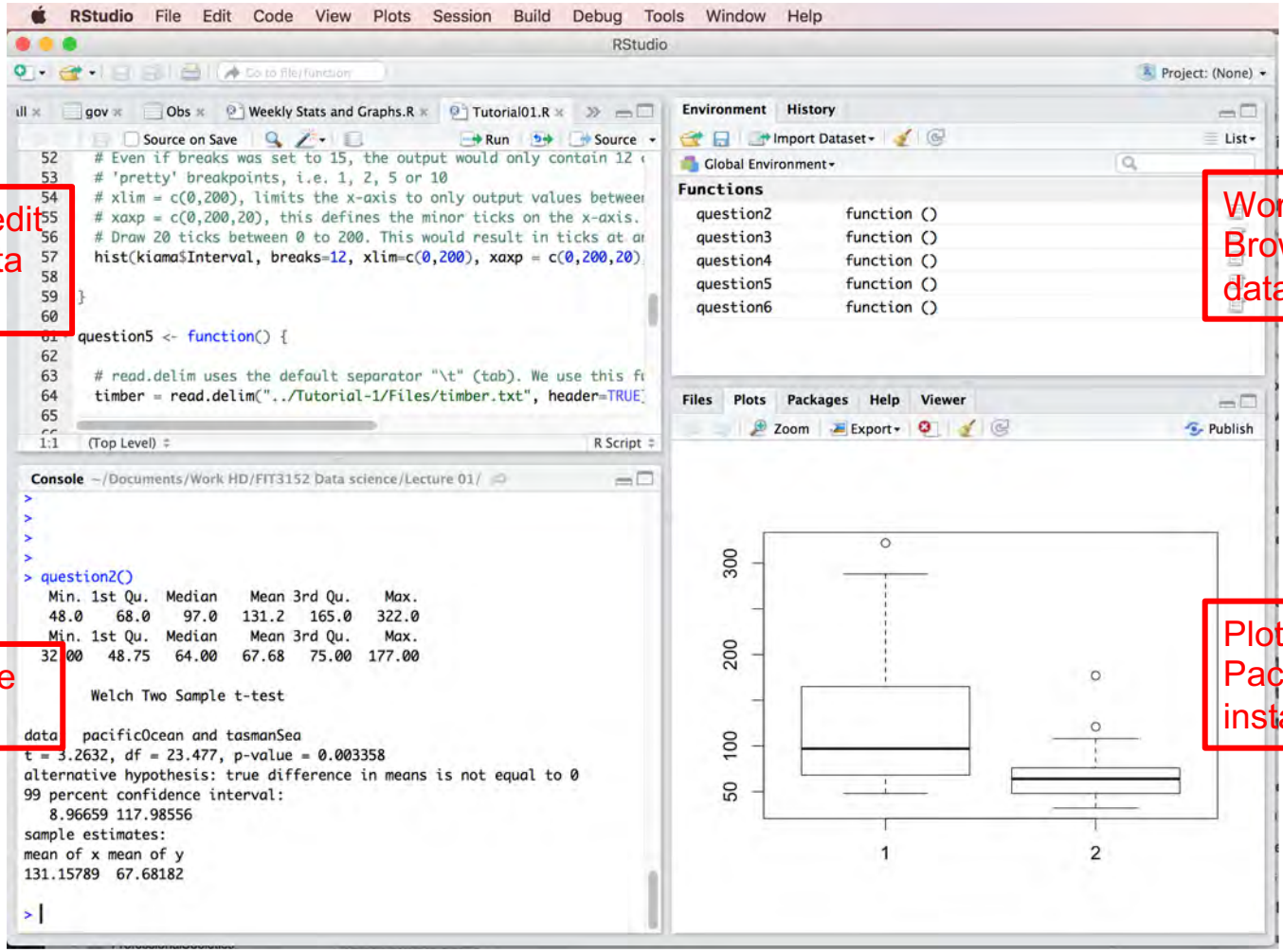
R & RStudio workspace

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 (Stackoverflow, for example) is a good source of information.

Packages

There are 18,000+ 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
```

Reading a file:

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

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 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
> # as the location of the script.
```

Reading csv files:

```
> X <- read.csv("InvestB.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

- *Some statistics revision notes posted on Moodle*
- *Following slides for reading and reference. We won't go through each example in detail.*

Descriptive statistics

Problem: describe 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
```

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 from 6 different types of investments. Same data two formats:

- InvestA is a single, indexed column,
- InvestB is 6 labeled columns.

InvestA.csv

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

InvestB.csv

FVA	FVB	FVC	FVD	FVE	FVF
809.34	716.72	775.58	1288.77	930.07	758.29
166.46	800.29	848.92	1300.21	817.28	730.28
711.33	748.75	813.58	1256.31	785.59	711.8
870.33	758.11	798.62	1274.43	748.14	804.45
758.56	959.04	758.55	1251.99	768.97	880.99
707.75	666.71	819.58	1262.94	731.76	688.23
681.3	712.38	770.67	1309.46	802.29	886.97
704.14	876.81	793.04	1350.24	728.84	880.99
…	…	…	…	…	…

Descriptive statistics

Problem: compare several groups stored as multiple columns of different length in a csv file.

```
> colMeans(InvestB, na.rm = TRUE) # ignore empty cells
```

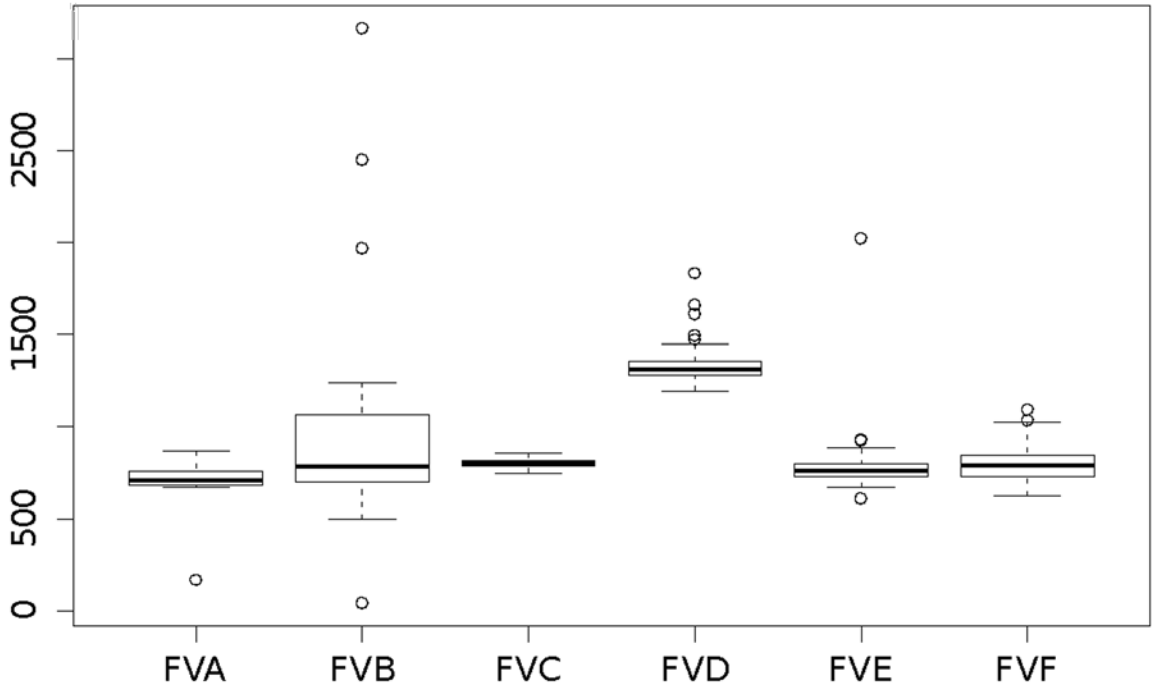
- The number of decimal places can be specified by wrapping mean function inside a rounding function.

```
> print(round(colMeans(InvestB, na.rm = TRUE),
  digits=2))
```

```
     FVA      FVB      FVC      FVD      FVE      FVF
  689.35   874.00   802.47  1339.10   786.74   797.16
```

Boxplot

```
> boxplot(InvestB)
> # not perfect but more on graphics next lecture
```



Bivariate data

The data:

- In 1998, *Choice* magazine tested 1500 toothbrushes and made a summary of price and function. Are these two factors related?

Toothbrush.csv

Price	Function
3.95	65.10
2.96	78.00
2.95	72.00
0.66	40.00
0.69	57.00
3.20	61.00
1.08	49.00
3.69	76.00
…	…

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

Bivariate data

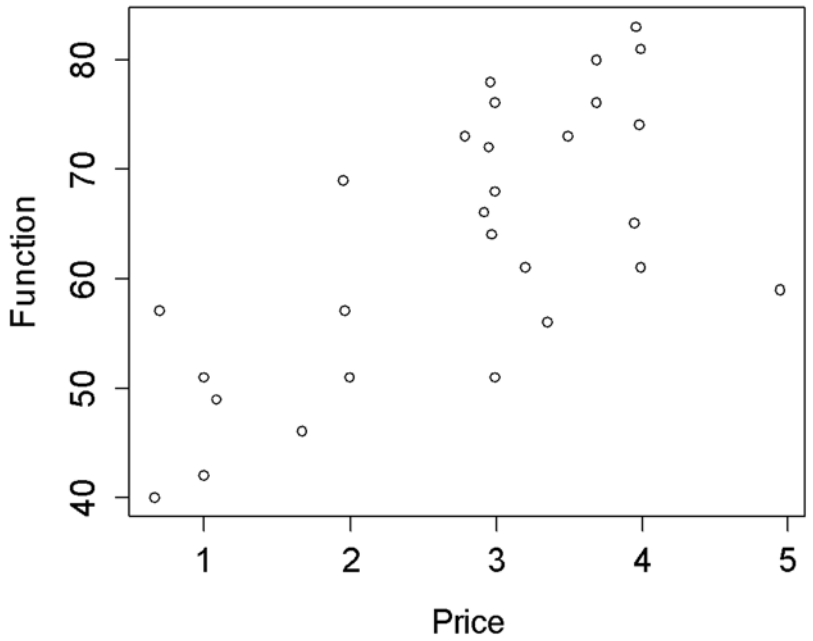
Problem: analyse the relationship between price and function.

- 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 putting Function on y axis
```



Identifying columns in a data frame

To identify the columns in a data frame you can:

- Create two vectors: Price and Function
  - > Price <- Toothbrush$Price
  - > Function <- Toothbrush$Function

- You can then create a scatterplot using Price and Function vectors: this method lets you specify the X and Y axes.
  - > plot(Price, Function)

Identifying columns in a data frame

A simpler method is to “attach” your data frame:

- The ‘attach’ function lets you call columns in a data object by name without having to specify the object name – assuming column name is unique amongst attached data sets.

```
> attach(Toothbrush)
```

- Scatterplot using Price and Function directly:

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

Bivariate data

Problem: calculate the regression equation cont.

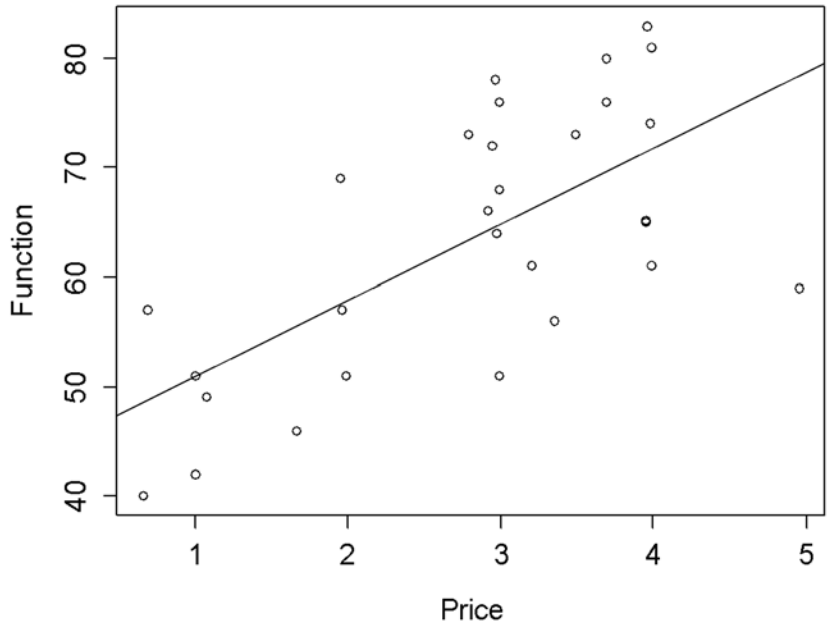
- To calculate the regression equation, define variable 'fitted' and use linear model (lm) function.
  - > fitted = lm(Function ~ Price)
  - > fitted

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

- Now overplot the fitted model on scatterplot using gradient and intercept from fitted model.
  - > abline(fitted)

Scatterplot + regression line

```
> plot(Price, Function)
> abline(fitted) # overplotting
```



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)
```

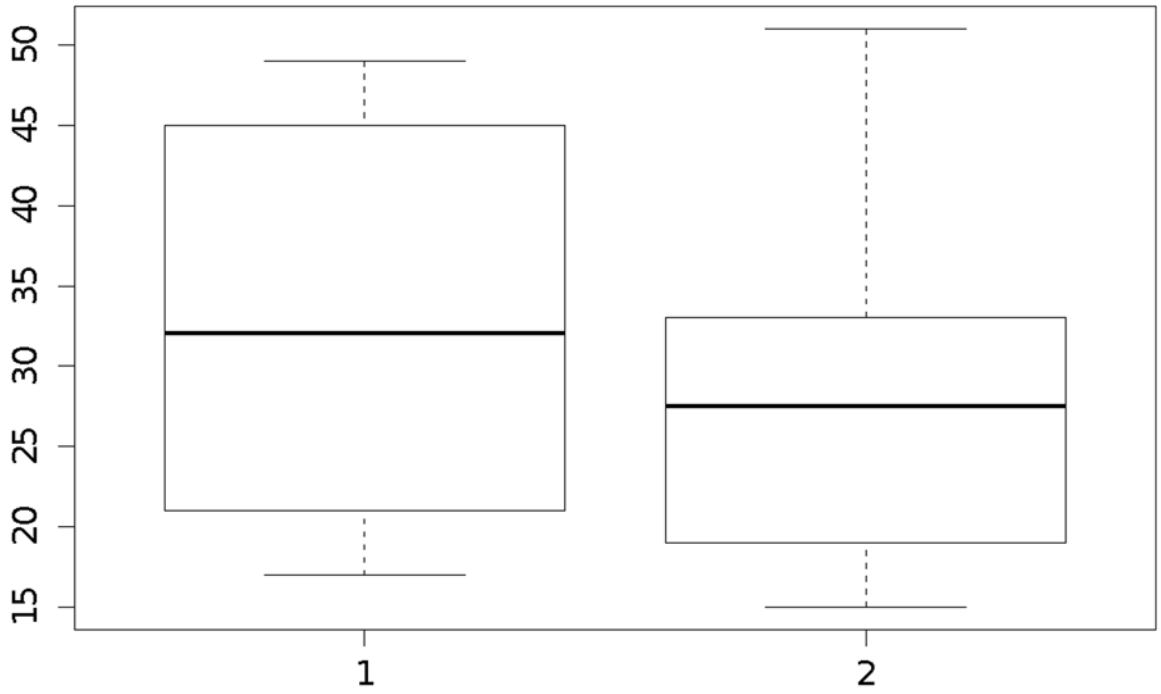
Workers.csv

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?

EHT Problem 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
```

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

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

EHT Problem 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

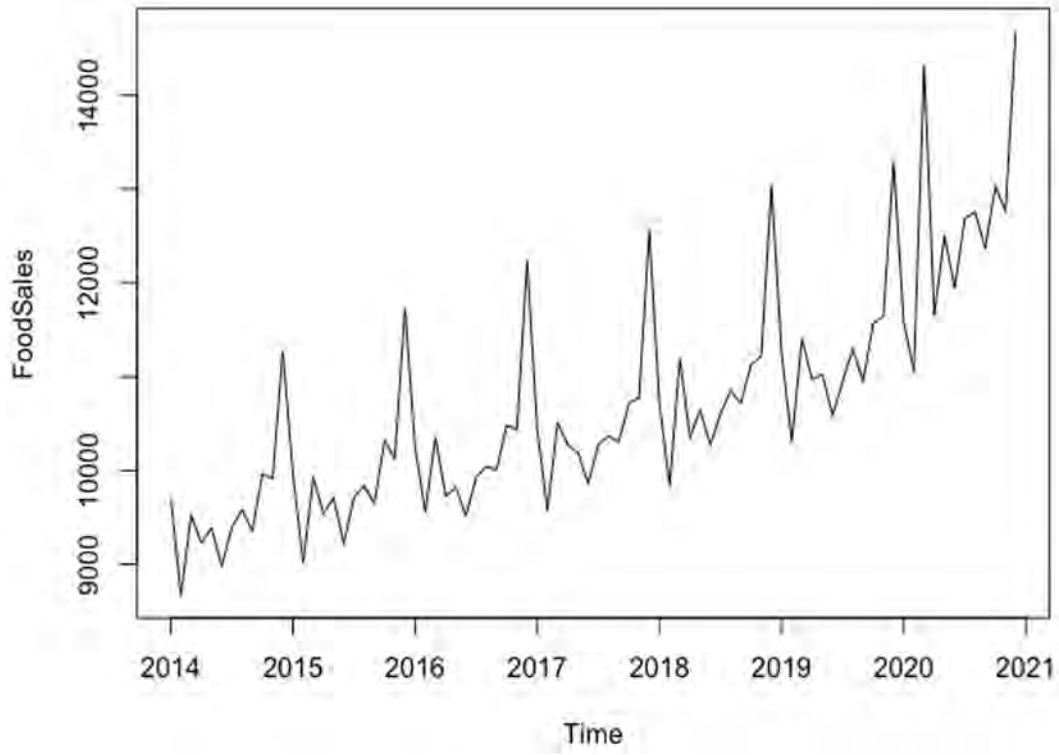
Problem: read the data and declare as class ts:

```
> 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)

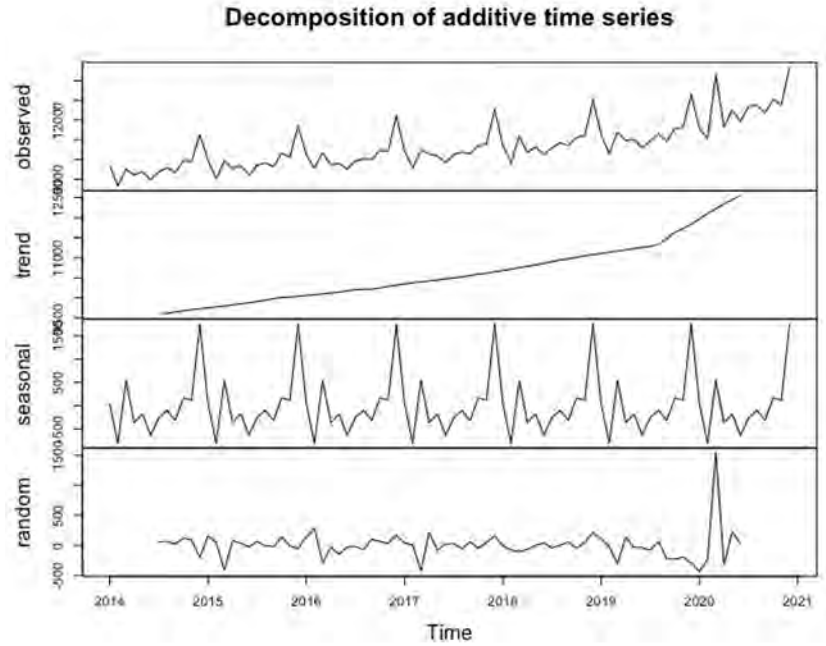


Time series analysis

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. (2013) An Introduction to Statistical Learning. Springer. *Online access via Library.*
- F. Provost and T. Fawcett. (2013) Data Science for Business. O'Reilly Media, Inc. *Online via Library.*
- H. Wickham, G Gromelund. (2017) R for Data Science. O'Reilly Media, Inc. Also available from: https://r4ds.had.co.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:

- Lecture examples, Tutorial 1 – attempts pre-tutorial activities especially!
- Reminder: Tutorials start next week!