Practical Data Science: Introduction

Dr. Yongli Ren

(yongli.ren@rmit.edu.au)

Computer Science & IT

School of Science

RMIT UNIVERSITY

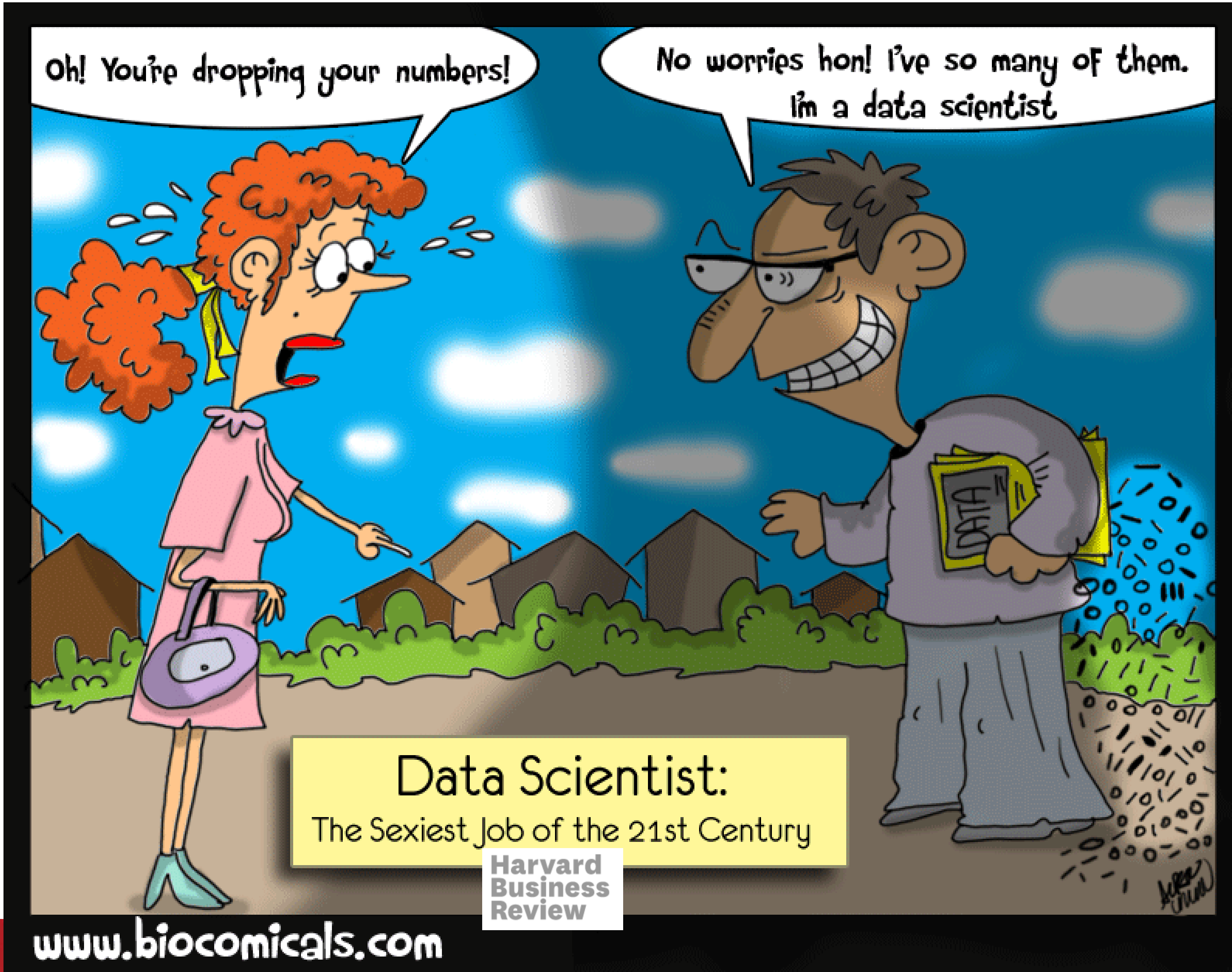
Outline

- Part 1: Overview
- Part 2: Administrivia
  - Course structure
  - Assessment
- Part 3: Introduction to data science
  - What is data science?
  - Data science process

Practical Data Science

PART 1: OVERVIEW

Data Science and Data Scientists



How Much Money Does A Data Scientist Make?

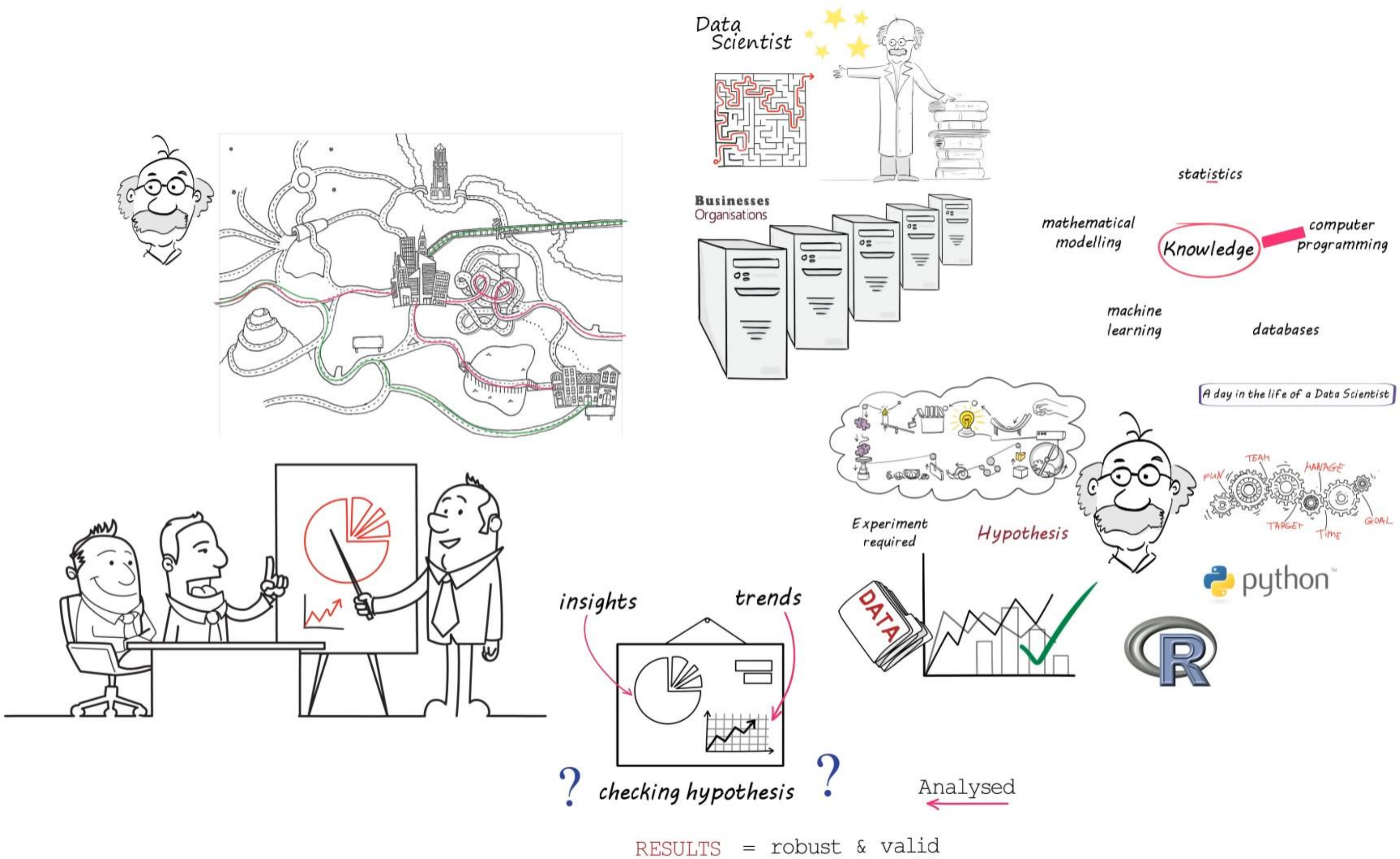
$116,840 a year, on average

Facebook: $133,841
Apple: $149,963
Airbnb: $117,229
Twitter: $134,861
Microsoft: $119,129
LinkedIn: $138,798
IBM: $110,823

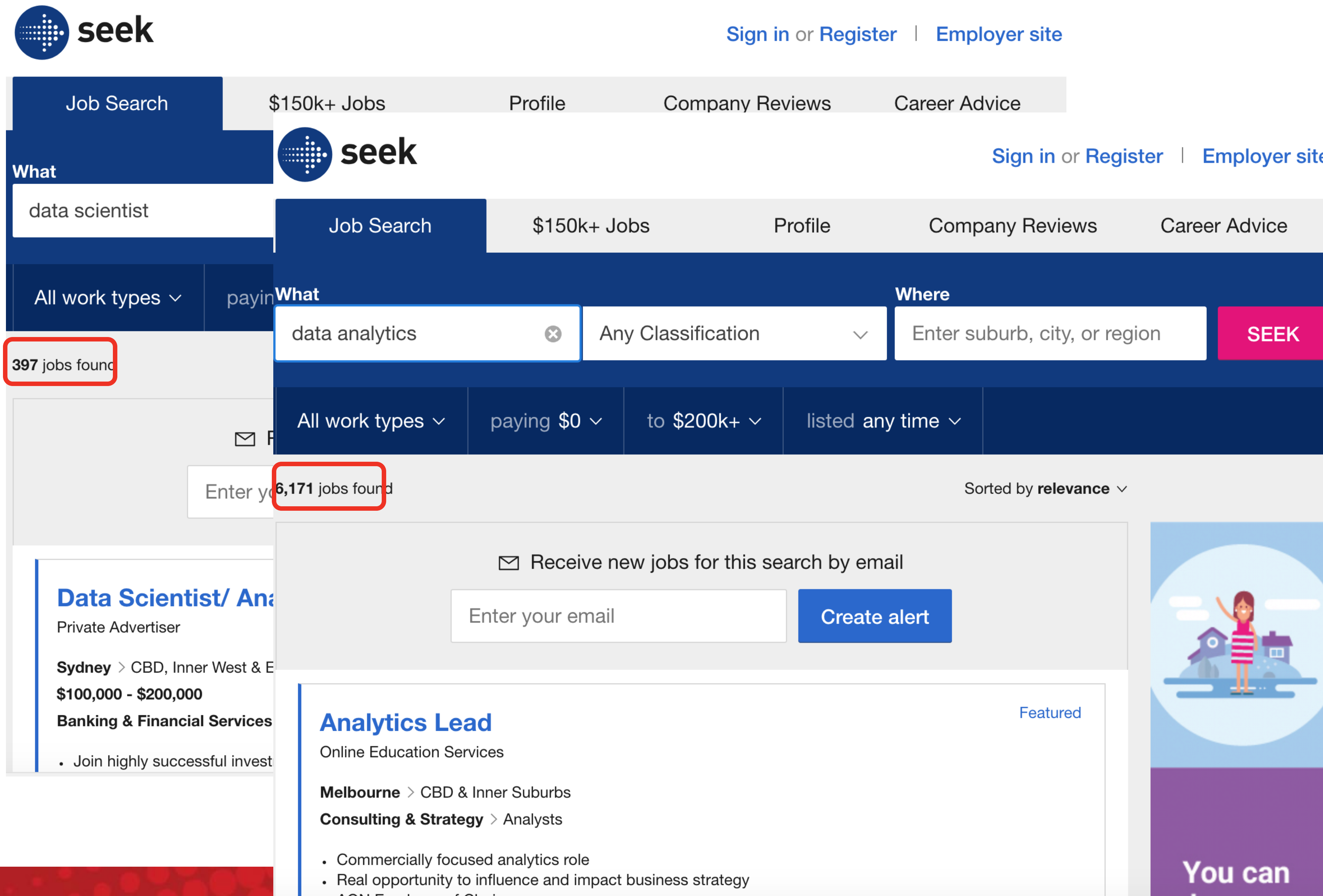
http://www.businessinsider.com.au, https://www.glassdoor.com.au/

What Do Data Scientists Do (Overview)?
- A 60 Second Introduction

- https://www.youtube.com/watch?v=i2jwZcWicSY



Employability



Data Scientist

ABC - More jobs by this advertiser

ABC

Data Scientist

Key Accountabilities

Data Environment

- Work directly to transform various raw big data sources into valuable information and insights
- Collaborate with other developers to make improvements to analytics tracking, data ETL processes and implementation of models or test results
- Provide advice on areas of opportunity for Machine Learning, NLP and A/B testing tools and systems

Data Science

- Test product and consumer behaviour hypotheses using large and disparate data sets Design, implement and analyse A/B and multivariate experiments
- Develop models to help explain and predict patterns of audience behaviour
- Use data to identify features/changes to existing platforms or future data-driven services that represent opportunities to improve the ABC digital audience experience

20 Feb 2017

Location:

Sydney ▸ CBD, Inner West & Eastern Suburbs

Work type:

Contract/Temp

Classification:

Science & Technology ▸ Mathematics, Statistics & Information Sciences

Apply for this job

Applications for this role will take you to the advertiser's site. Use your SEEK Profile to pre-fill the application.

You may need to tell this advertiser how your skills and experience meet their selection criteria. View tips on selection criteria.

Save job | Email job | Add note | Print | Share

Practical Data Science

PART 2: ADMINISTRIVIA

Course Information

- Lecturers:
  - Yongli Ren
    - Senior Lecturer @ RMIT
    - Researcher on MS Cortana
      - Westfield, Arup, iSelect, SEEK

  - Ahmed
    - Researcher @ RMIT
    - Currently collaborating with SEEK
      - Previously with Emprevo

Course Structure

- Contact hours:
  - One 2-hour lecture each week
    - 6:30pm – 8:30pm, Tuesdays, room: 80.04.11
  - One 2-hour tute/lab each week
  - Personal study (readings, tutorial exercises)
- Office Hour (Q/A) for Ahmed (weeks 2 - 4):
  - Week 2 -- 4:00pm-5:00pm, Tuesday, room: 14.09.07 (Yongli's office)
  - 3:30pm-4:30pm, Thursdays, room: 14.09.07 (Yongli's office)
- Office Hour (Q/A) for Yongli (since week 5):
  - 4:00pm-5:00pm, Mondays, room: 14.09.07 (my office)
- Tute/labs:
  - Start in week 2
  - Tutorial questions will be made available before each class via Canvas
    - It's assumed that you will have read the questions and thought about them before coming to the class

Assessment

- Two Assignments
  - 1: Data Cleaning and Summarising (15%)
    - Structured task with a provided dataset to analyse
  - 2: Data Science Analysis (35%)
    - Analyse a Data Science Dataset
    - Produce a formal report about your analysis and findings
    - Groups of 2 encouraged

- Exam (50%)
  - 2 hour closed-book exam

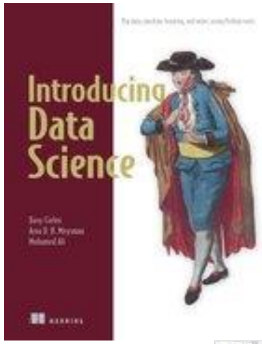
- Note: tute/labs are important, as they include the elements of both assignments and final exam

Course Materials

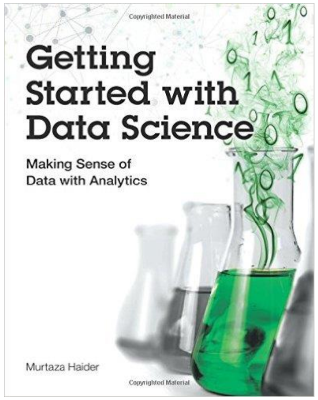
- Online (Canvas)
  - Teaching materials
    - Course slides/recordings
    - Tutorial questions
    - Assignments
  - Discussion forums
    - You are encouraged to post your questions here, as the discussion/answers may help the other student as well

Textbooks

- This course draws on a range of material, including several textbooks and other sources
- References will be given with each topic
- A good general textbooks to support the course content
  - *Python Data Science Essentials*, A. Boschetti and L. Massaron, Packt 2015.



- Further good textbooks (the first two focus on python for data science, the third on the importance of "storytelling" and communication in data science)
  - *Data Science from Scratch*, J. Grus, O'Reilly 2015.
  - *Introducing Data Science*, D. Cielen and A. Meysman and M. Ali, Manning 2016.
  - Getting Started with Data Science: Making Sense of Data with Analytics, Murtaza Haider, IBM Press. 2015



- NB: The first three mentioned texts are available online through the RMIT library

What You'll Get Out of This Course

- Understanding of data science
  - What data science is
  - The data science process
- Learning about a range of key analytical techniques
- Practical experience using Python for data science
- Enhanced problem solving skills

Python

- General-purpose, high-level programming language
- Design emphasises code readability and minimal core syntax
- Supports many programming paradigms,
  - including object-oriented and procedural
- Comes with extensive libraries of functions for high-level tasks, such as
  - handling file formats,
  - database access,
  - GUIs,
  - Internet protocols, etc.
- May be used as an interactive scripting language

```
print(“Hello, world!”)
```

Python Distributions

- Useful “bundled” distributions are also available
- Anaconda
  - Combines python with the core libraries that are used for data science
  - Windows, OSX and Linux distributions are available
  - The Lab/RMIT coreteaching servers have Anaconda installed

  [https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)]

  [https://www.continuum.io/anaconda-overview]

- PythonXY
- WinPython

Python for Data Science

- Need to be able to manipulate and analyse data, which requires some understanding of python
  - Data types/structures
    - Lists, dictionaries, data frames
  - Control structures
- But this is *not* a python programming course

Interactive Data Analysis

- Python is often used to write full scripts or stand-alone programs
- A lot of data science requires exploratory data analysis
  - Could write script files, edit them, and (re-)run them
  - Often convenient to use a true interactive environment
- In this course we'll use the *iPython* interactive environment
  - Setting up iPython is covered in detail in tutorial 1

Python Resources

- Main python website
  - https://www.python.org/
- Documentation
  - https://docs.python.org/3.0/
- Python tutorial (covers core language features)
  - https://docs.python.org/3.0/tutorial/index.html

Practical Data Science

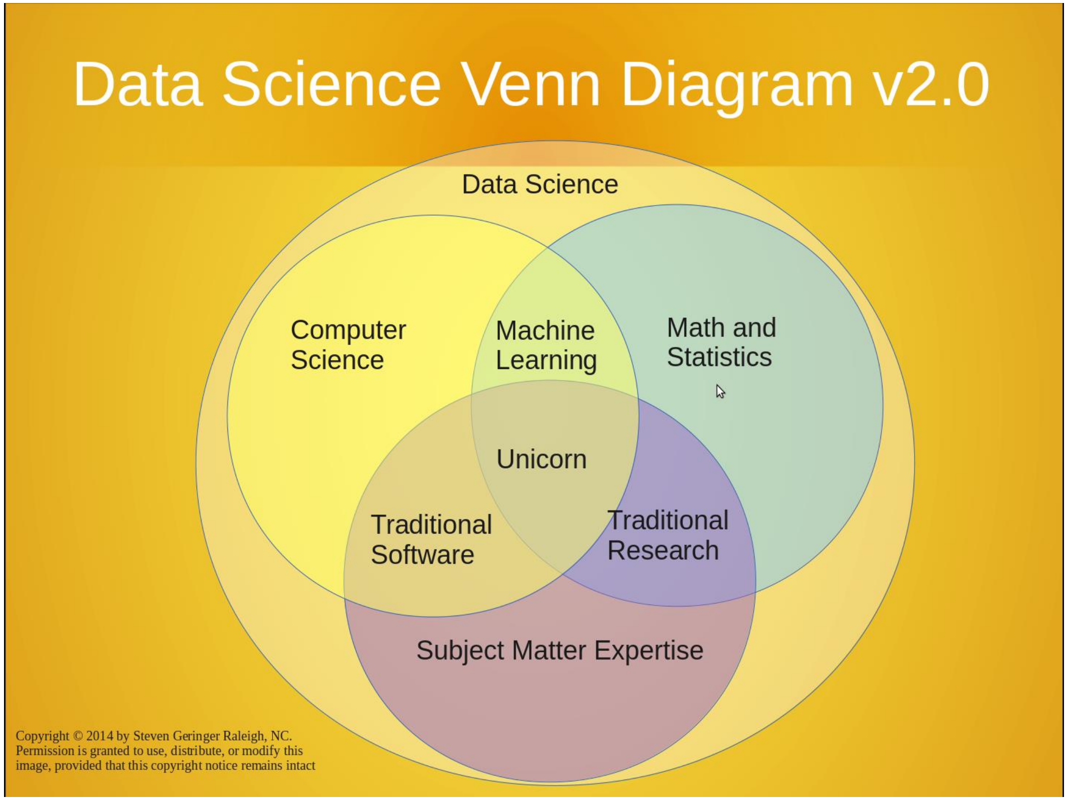
PART 3: INTRODUCTION TO DATA SCIENCE

What is Data Science? (Definition)

- An *interdisciplinary* field about *scientific methods, processes and systems* to *extract knowledge or insights* from *data* in various forms [WIKIPEDIA]
- A less formal definition:
  - "Data scientists are better statisticians than your average programmer, and better programmers than your average statistician" [ ]
- The "sexiest job in the 21st century"
  - according to the Harvard Business Review

What is Data Science? The Unicorn

- Computer Science
  - Data is electronic;
  - need to manipulate it
- Math and Statistics
  - Get insight from data
- Subject Matter Expertise
  - Motivating questions and hypotheses about the world



http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html

Storytelling

- An additional important skill highlighted by many is *storytelling*
- Data science is carried out for a reason
  - Increase revenue, understand something, make better decisions, …
  - If you can't convey the findings, the impact of the analysis will be limited
  - Both written and verbal communication is important
- A data scientist is
  - "someone who finds solutions to problems by analyzing big or small data using appropriate tools and then tells stories to communicate her findings to the relevant stakeholders" [Murtaza Haider]

And Don’t Forget the Science

- Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe [WIKIPEDIA]
- *Data science* includes the word science in its name, but be aware that analysis on its own, even if carried out using robust statistical models or other formal processes, is *not* necessarily *science*
- Analysis is conducted to gain insight
  - Driven by a research question
    - For the analysis to be effective, you need to understand the research question
  - *Hacking around in huge data sets will almost certainly result in finding some “statistically significant” relationships*
    - If the analysis is not motivated, such outcomes may turn out to be meaningless in a practical sense, e.g.
      - Correlation is not causation.
        - *The Lack of Pirates Is Causing Global Warming*

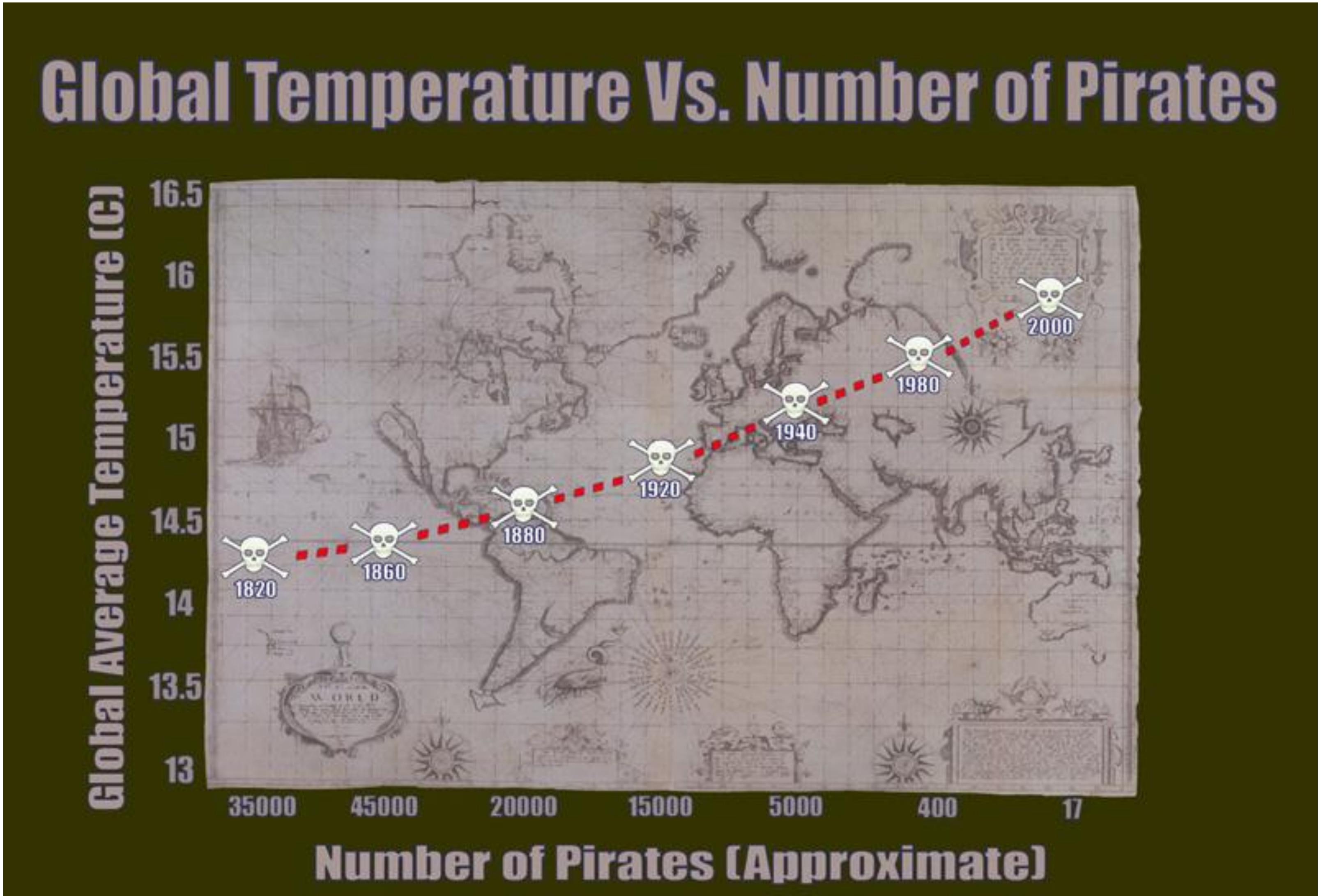


Photo via http://bama.ua.edu/

Modelling Shopping Habits

- Companies can learn a great deal about customers through their actions
- Loyalty cards might seem to give you “benefits”, but companies usually benefit much more
  - http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- 1-minute Discussion point:
  - is there a similar trade-off with “free” services such as Facebook?

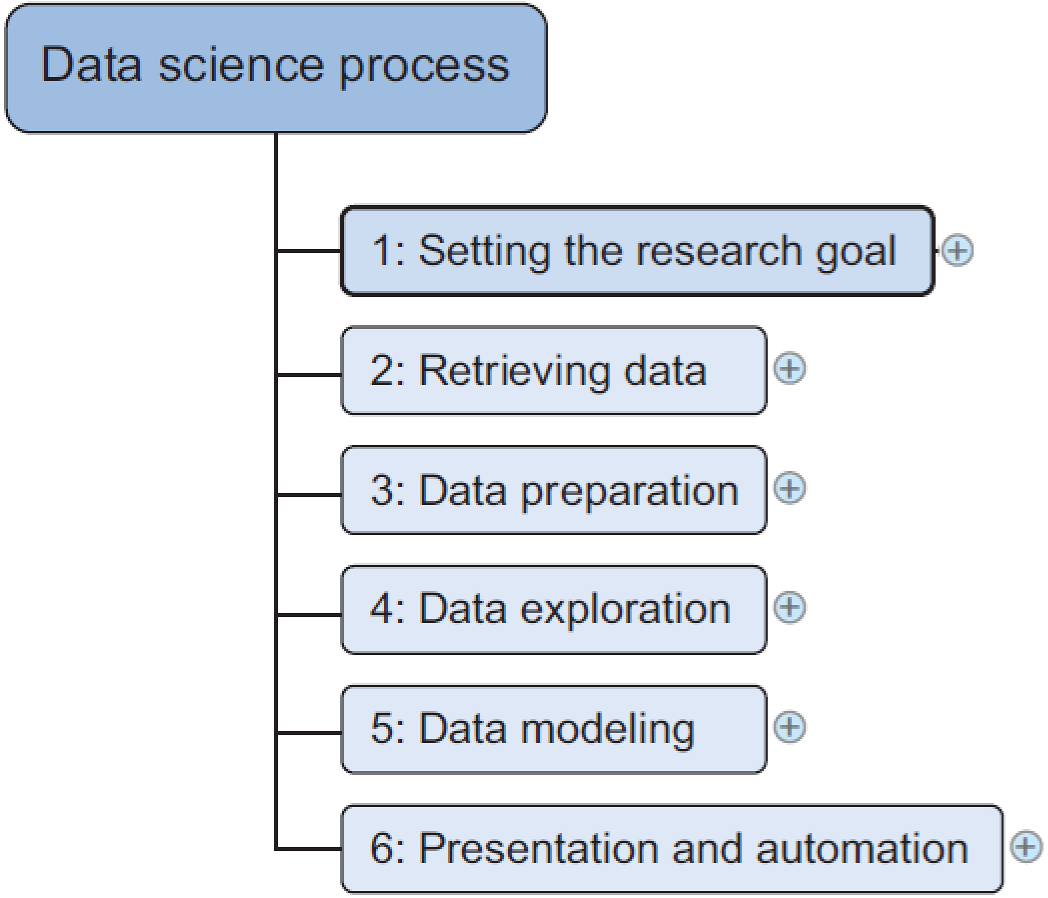
Data Science Is Not A Magic Cure-All

- Data science is currently a “Big Thing” and receives a lot of hype in the media
  - But clearly data science does not impart omniscience, and
  - it’s vital to maintain a *realistic* level of expectations
- Target had some big successes with data science analysis, as we saw earlier
  - But in 2015, they had to close down all of their Canadian operations, which had only been opened in 2011

Data Science Is Not A Magic Cure-All…

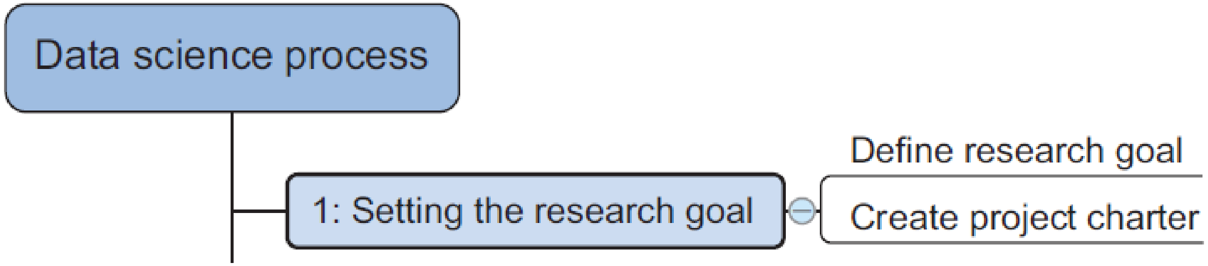
- In 2008 Google Flu Trends was a new approach for predicting flu epidemics
  - Based on search queries that people submit
  - Was more precise and faster than traditional modelling approaches from epidemiology
  - A “poster child” for what big data can do
- In 2013, the GFT predictions were substantially worse then epidemiological models
  - The assumptions of the approach were not robust, search behaviour changed over time
  - The project was shut down

The (Typical) Data Science Process



- Note: the approach may be iterative and non-linear!
- Note: don't be a slave to the process, which may not be the same for every project.

Step 1. Identify Research Goals



- Understanding the purpose of the project is a vital aspect
  - What are the broader business goals?
  - What is the business context for the question being asked?
  - How will the project change the business, how will the results be used?
- Avoid the situation where you “finish” a data science project, only to then realise that the initial question and requirements were misunderstood!

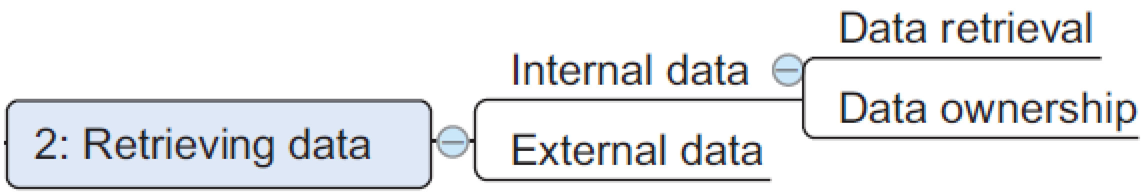
Step 1. Identify Research Goals…

- Ask questions and sketch examples to clarify
- Need to understand the *what*, *why*, and *how* of the project
  - What is the project expected to do
    - E.g. answer a specific question:
      - “What determines teaching evaluations”
  - Why is value being placed on the project
    - E.g. management cares about treating staff fairly; or, management cares about boosting university scores
  - How should the analysis be carried out
    - E.g. must use data that includes a representative range of staff and subject profiles

Step 1. Identify Research Goals…

- Aim to get formal agreement on deliverables through a *project charter*, which may include:
  - Statement of research goal
  - Broader project mission and context
  - Required resources and data
  - How analysis will be performed
  - Proof that it's an achievable project
  - Measure of success
  - Formal deliverables (e.g. project report)

Step 2. Identify And Retrieve Data



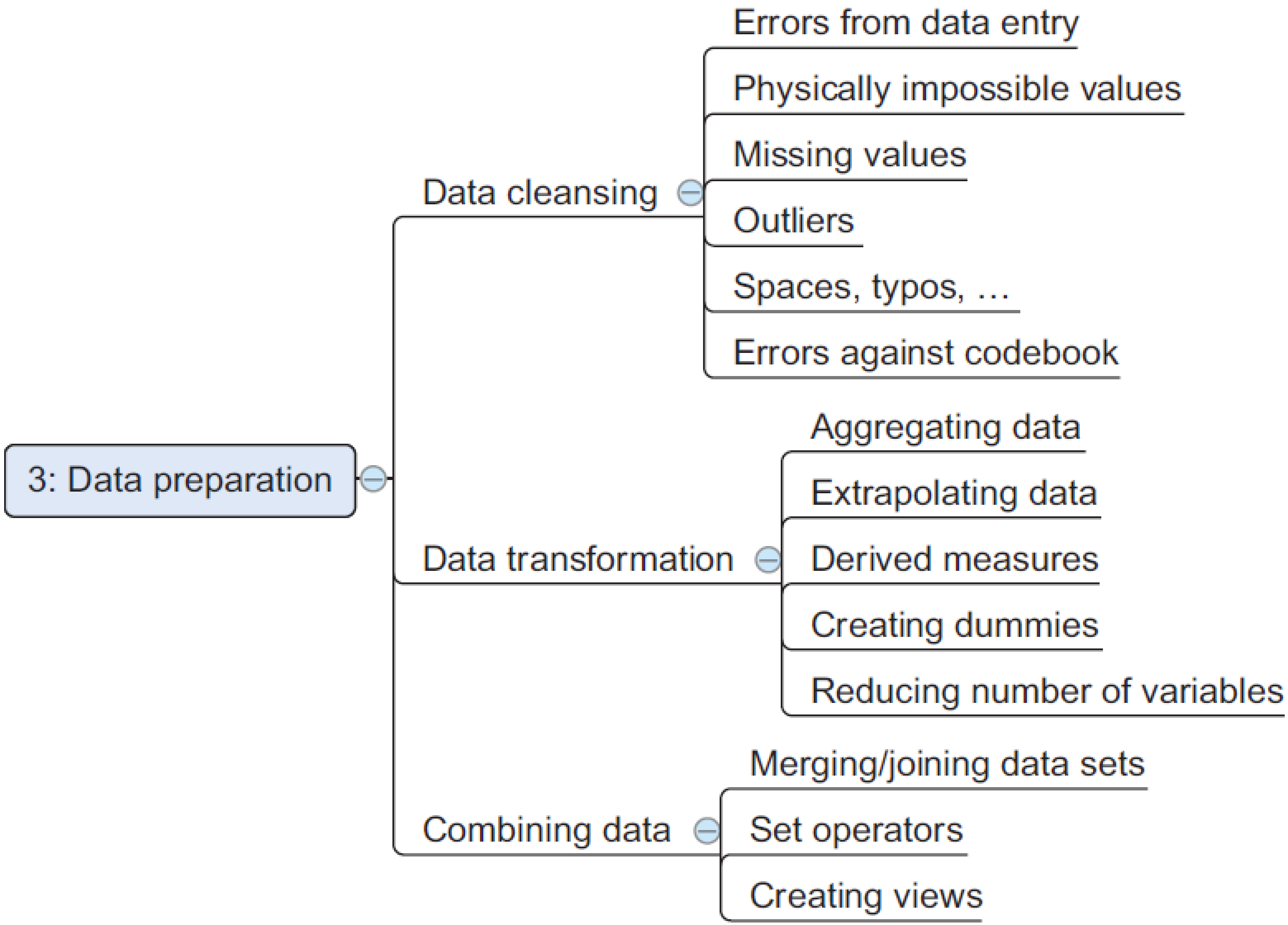
- A review of related research is usually an important first step
  - Helps to determine what kind of information is needed
  - Determine if similar questions already been examined
- Data within a company
  - Likely to have some sort of database, data warehouse, spreadsheet, etc.
  - Larger companies may have multiple data stores
  - Sometimes just getting access to data can be a challenge
    - Commercial data is valuable and sensitive; many companies create physical and digital barriers to prevent unauthorised access
    - Legal restrictions on data access
  - Likely to be in raw form and require processing

Step 2. Identify And Retrieve Data…

- Open data is becoming more prolific
  - US government: https://www.data.gov/
  - European Union: https://data.europa.eu/
  - World Bank: http://data.worldbank.org/

  …

Step 3. Data Preparation



Step 3. Data Preparation

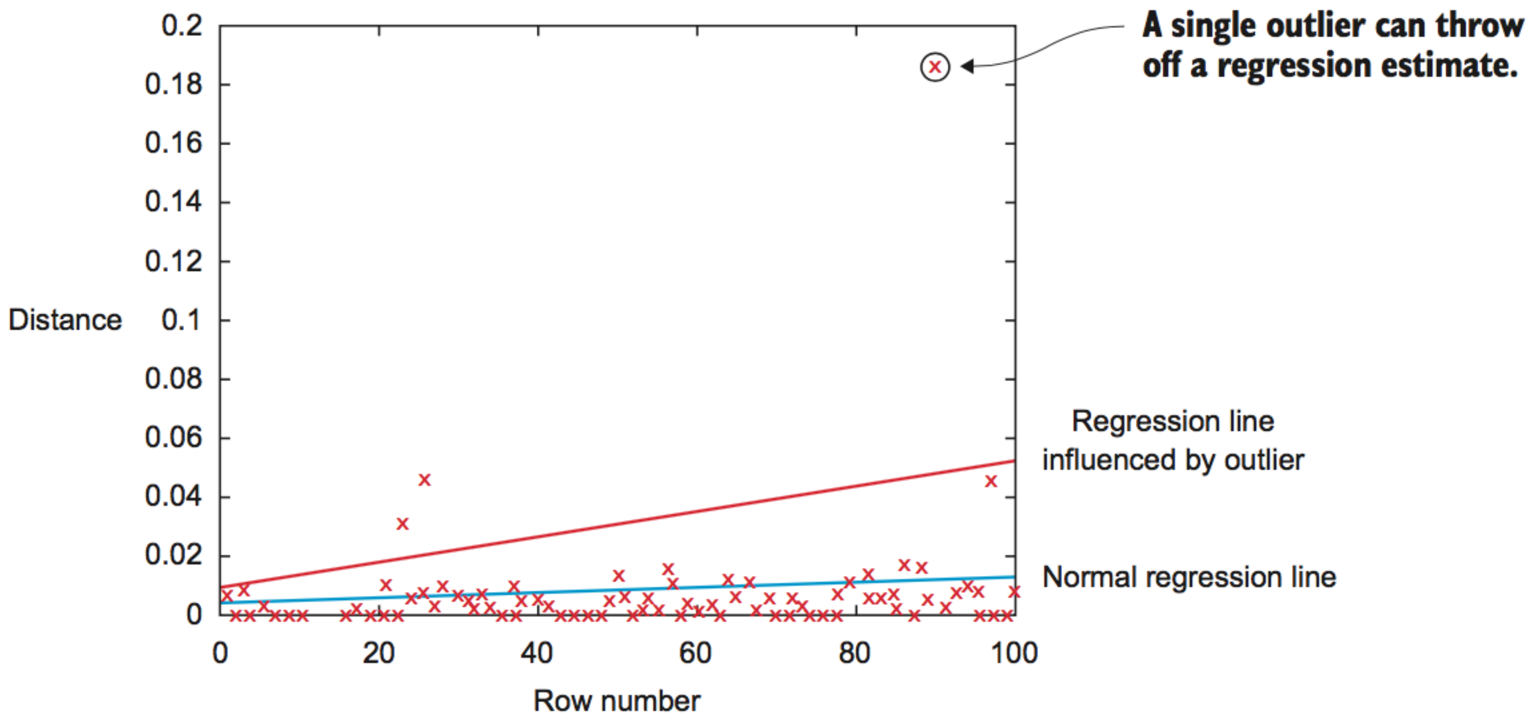
- Retrieved data is typically in a "raw" form, and may not be directly usable
  - May include errors
  - Analysis may require a particular format
    - Depending on software/tool
    - Depending on planned analysis technique
- This phase often consumes a lot of time
  - But it's vital: "garbage in, garbage out"

Step 3. Data Preparation…

- Removing errors is essential
  - The data should be a true and consistent representation of the process it originates from
- *False value error*: taking data for granted
  - A person's age is greater than 300 years.
- *Inconsistencies*: data sources present information differently
  - Meters in one table, feet in another
  - Coding gender as "F" and "Female" in data sets

Step 3. Data Preparation…

- Diagnostic plots can be helpful, e.g. to identify outliers



[Cielen et al., p32]

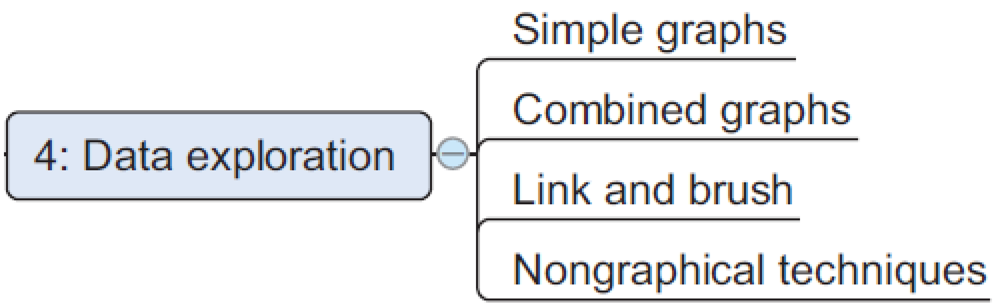
Step 3. Data Preparation…

- Always include sanity checks
  - Particularly important when manipulating multiple sources raw data
    - E.g. Joining tables on a common column
    - It's easy to get things wrong, or for automated processes to encounter unexpected exception cases
- Aim to correct errors as early as possible
  - Results of analysis may become invalid depending on severity of errors
    - Could lead to faulty decision-making
  - Errors may indicate defective equipment, or software bugs
    - E.g. sensors not working as expected
    - Valuable to fix these issues as early as possible

Step 3. Data Preparation…

- Transformation
  - Certain analytical techniques require data to be in special forms
    - E.g. linear regression assumes a linear relationship between dependent and independent variables
  - You may be able to apply a *transform* to your data so that it fulfils requirements
    - E.g. *total GDP* to *per capita GDP*
  - Or, you may conclude that you have to rule out some analytical techniques, because they're not suitable for your data

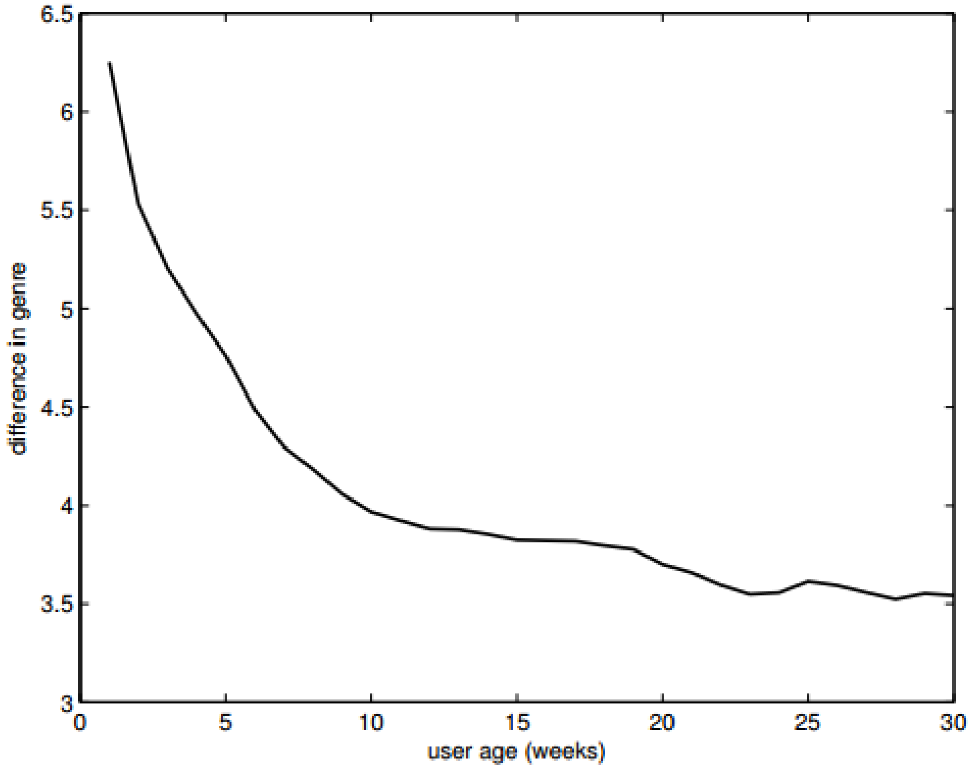
Step 4. Data Exploration



- The aim of data exploration is to get a deep understanding of the data
- Key approaches are
  - Descriptive summary statistics
    - Mean, median, mode, standard deviation
  - Graphical techniques
    - Bar graph, line graph, distribution graph.

  - Note: the aim here is *not to cleanse* the data, but you might discover further *problems and anomalies*, which may lead you to return to the previous phase to address them

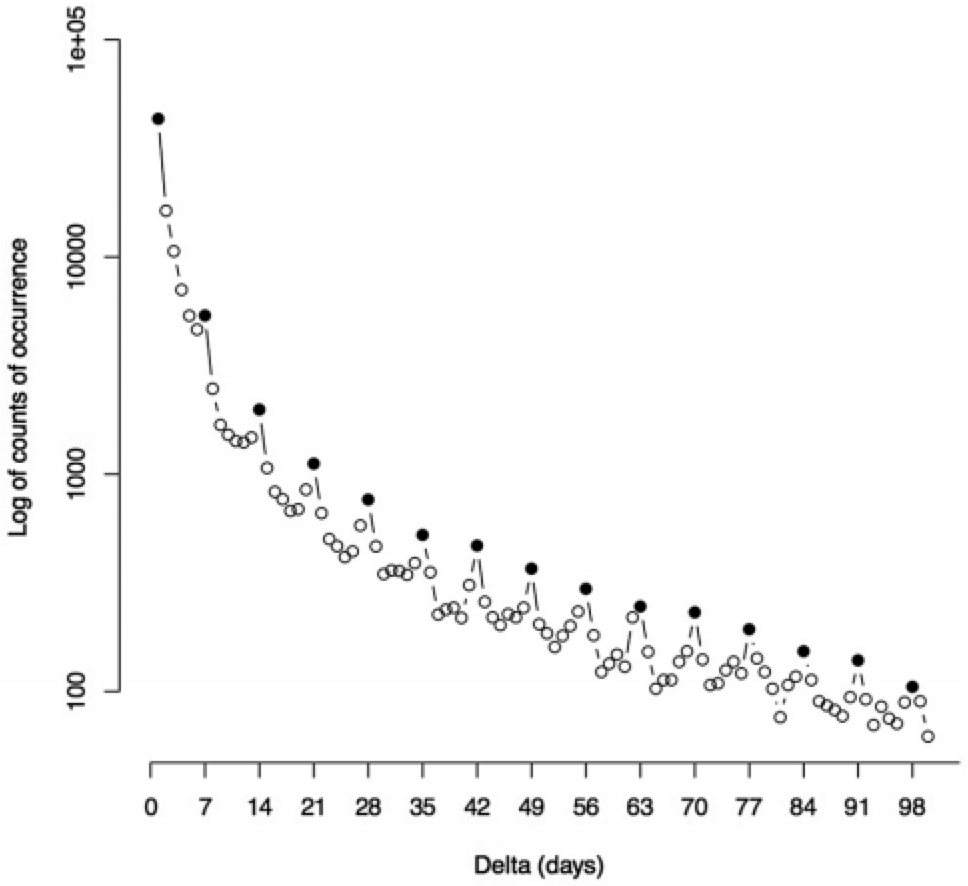
Step 4. Data Exploration:
- User Age vs. Interests in Movies



genre difference by user age

- The figure shows that
  - the user's *watching pattern* does change over time.
- We observe that
  - fresh users tend to watch a larger range of genres than experienced users, and
  - users' genre preferences are becoming stable after 23 weeks since joining the system.

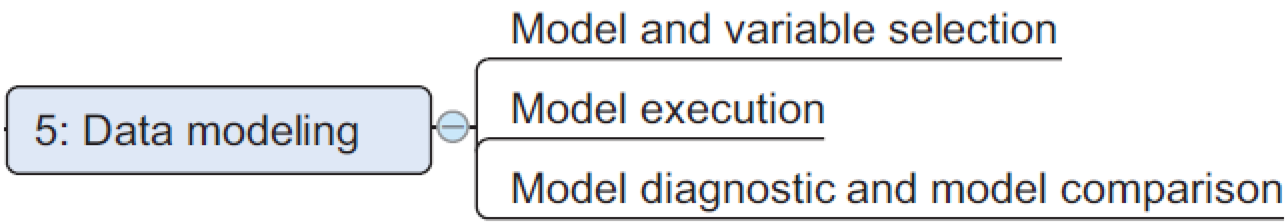
Step 4. Data Exploration:
- Visitors of a Shopping Centre



Counts of consecutive visits of all visitors binned by the Δ in days.

- The figure shows the distribution of the kinds of user visits to a shopping centre,
  - which is treated as a function of the difference in days between two consecutive visits of the same user.
- We observed that
  - the distribution of return visits *does not follow a uniform* decreasing pattern, but the strong impact of a 7-day periodicity is captured in the data.

Step 5. Data Modelling



- There are many analytical techniques available, the choice of which to use may be influenced by
  - The research question being addressed
  - Restrictions inherent in the data
  - Project requirements
    - Ease of maintenance / update requirements
    - Explanability
    - Requirement of deployment in a production environment

Step 5. Data Modelling…

- Among others, this course will cover various techniques for
  - *Classification*:
    - E.g. visitor classification in shopping centers.
  - *Clustering*
    - E.g. customer clustering (segmentation).
  - *Recommendation Techniques*
    - E.g. movie recommendation.
  - *Textual data*
    - E.g. query analysis, document retrieval.

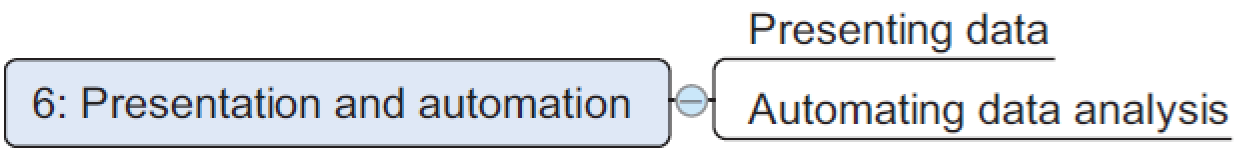
Step 5. Data Modelling…

- After selecting a set of models, they need to be executed
  - There may be many different implementations available
  - In this course we'll cover python as the main data science tool
    - Includes extensive libraries that cover models and analytical techniques

Step 5. Data Modelling…

- Typically, the aim is that a model will work on new (previously unseen) data
- Evaluation therefore involves
  - *Training* on some sub-set of data
  - *Evaluating* on a held out set of data
- Many different evaluation measures exist
  - The choice is often related to the type of model used
    - E.g. classification effectiveness may be measured using
      - precision,
      - recall,
      - F1,
      - …

Step 6. Presenting/Reporting Findings



- Once the modelling and data analysis is complete, the findings need to be presented to the stakeholders
- Appropriate visual tools can enhance results presentation
  - Texts
  - Tables
  - Graphs
- Overall, it’s important to tell a compelling story

Structure of a Report

- *Note: emphasis and sections will vary, e.g. for a brief versus a detailed report*
- Cover page
  - Title
  - Authors
  - Affiliations
  - Contact details
  - Date of publication
- Table of contents
  - For a report of more than several pages in length

Structure of a Report

- Abstract / executive summary
  - A paragraph-length summary of the key arguments and findings

Executive Summary (or Summary or Abstract)

Purpose

Method

Results

Conclusions

Recommendations

The aim of this report was to investigate university teaching staff attitudes to the use of mobile phones by students in tutorials. A survey of teaching staff from each college was conducted in first semester of the academic year. Overall, the results indicate that the majority of staff found student mobile devices use a major disruption in tutorials. The report concludes that the predominant view of staff is that mobile phones are disruptive and should be turned off during tutorials. It is recommended that the university develops guidelines which would support staff in the restriction of student use of mobile phones in tutorials except in exceptional circumstances.

Structure of a Report

- Introduction
  - *Explanation* of the problem
  - Particularly important since *many readers might not be experts* in the *topic area, or the analytical methods* that were applied
  - Often includes a literature review
    - Explain what's already known, as well as gaps in knowledge

Introduction

Context	There has been a great increase in the use of personal mobile phones over the past five years with every indication that this usage will continue to increase. Indeed, widespread use of mobile devices in educational contexts for non educational purposes has been reported as distracting and disruptive to learning environments. Recently a number of university teaching staff have proposed that an institution wide policy be developed regarding student mobile phone use during tutorials and lectures.
Purpose	This report will discuss research into staff attitudes to the issue of student mobile phone usage in the teaching and learning environment.

Structure of a Report

- Methodology
  - Explanation of
    - Data collect
    - Choice of va
    - Analytical te

Learning About Work Tasks to Inform Intelligent Assistant Design

Johanne R. Trippas
Damiano Spina
Falk Scholer
johanne.trippas@rmit.edu.au
damiano.spina@rmit.edu.au
falk.scholer@rmit.edu.au
RMIT University
Melbourne, Australia

Ahmed Hassan Awadallah
Peter Bailey
Paul N. Bennett
Ryen W. White
hassanam@microsoft.com
pbailey@microsoft.com
pauben@microsoft.com
ryenw@microsoft.com
Microsoft
Redmond, USA

Jonathan Liono
Yongli Ren
Flora D. Salim
Mark Sanderson
jonathan.liono@rmit.edu.au
yongli.ren@rmit.edu.au
flora.salim@rmit.edu.au
mark.sanderson@rmit.edu.au
RMIT University
Melbourne, Australia

ABSTRACT

Intelligent assistants can serve many purposes, including entertainment (e.g. playing music), home automation, and task management (e.g. timers, reminders). The role of these assistants is evolving to also support people engaged in work tasks, in workplaces and beyond. To design truly useful intelligent assistants for work, it is important to better understand the work tasks that people are performing. Based on a survey of 401 respondents' daily tasks and activities in a work setting, we present a classification of work-related tasks, and analyze their key characteristics, including the frequency of their self-reported tasks, the environment in which they undertake the tasks, and which, if any, electronic devices are used. We also investigate the cyber, physical, and social aspects of tasks. Finally, we reflect on how intelligent assistants could in-

has been growing interest in applications of these assistants in workplaces to empower employees, through offerings such as Alexa for Business[1] and Cortana Skills Kit for Enterprise.[2] Despite the potential for these assistants to help people complete their work tasks (at work, at home, or on-the-go), penetration of these assistants in workplaces is limited [19], and task support is restricted to low-level tasks such as controlling devices, seeking information, or entertainment [26]. In work settings in particular, intelligent assistants are mostly used for basic tasks such as voice dictation, calendar management, and customer/employee support [19]. To increase the uptake of intelligent assistants for work tasks, a better understanding of the tasks that people perform, and how next-generation intelligent assistants could support them, is needed.

There are many ways to understand tasks and activities, includ-

Refer to relevant reading/literature

Describe how the research was done

voluntary and anonymous. A total of 412 questionnaires were distributed online to randomly selected staff from each of the three colleges within the university. The completed questionnaires were returned by email.

Structure of a Report

- Results
  - Present the empirical findings of the analysis
  - Typically includes
    - Descriptive statistics
    - Visualisations (graphs, charts, illustrative graphics)
    - Analytical / model outcomes

Table 1

Table 1: Distribution of results

Mobile phone use in tutorials	Agree %	Disgree %	Strongly disagree %
1. Not a problem	13	65	23
2. Sometimes a problem	67	18	15
3. Often a problem	50	27	23
4. Phones should be allowed	22	56	22
5. Phones should be turned off	70	18	12
6. Phones should be allowed in some circumstances	47	39	14

Results

Facts only- no interpretation.

There was an 85% response rate to the distribution of questionnaires to staff. The results clearly show that student mobile phones are considered by teaching staff to be disruptive (see Table 1). As a result, most staff would prefer that mobile phones were turned off in tutorials.

Structure of a Report

- Discussion
  - Presentation of main argument
  - Explains how the results address knowledge gaps and answer the research question
- Conclusion
  - Summarise findings
  - Explain wider applicability of results
  - Identify possible future developments and applications
  - New research questions that have opened up

Overview of Report

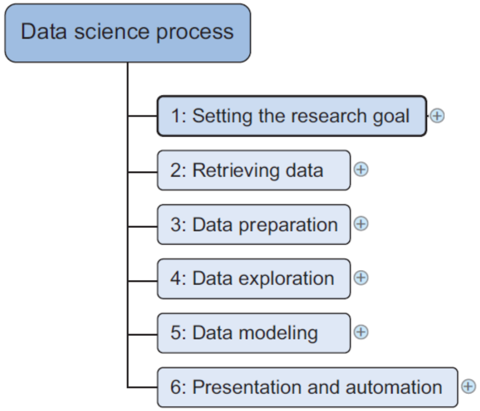
Conclusion

Summary of main findings and 'the answer'	The student use of mobile phones in tutorials is clearly intrusive to teaching staff and detrimental to learning environments in general. The study highlights the concerns of teaching staff with regard to mobile phone usage. The fact that the majority of staff views the student use of mobile phones in tutorials as disruptive suggests appropriate guidelines and policies need to be developed.

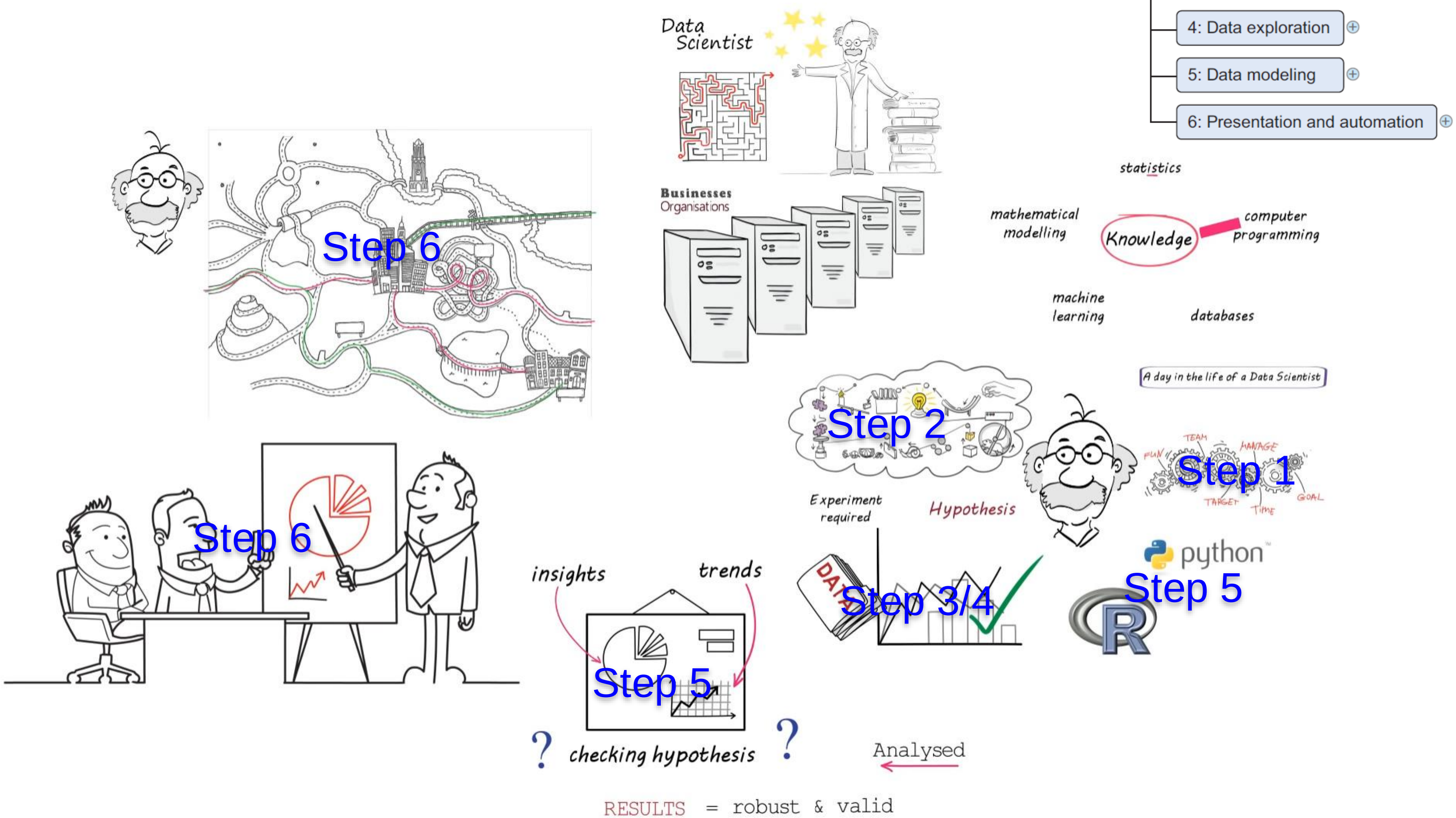
Automation

- Sometime, people want to repeat your work over and over again,
  - because they value the predictions of your models or the insights
- This *doesn’t* mean that
  - you have to *redo* all of your analysis all the time
  - Sometimes
    - it’s sufficient that you implement only the model scoring
  - Other times
    - you might build an application that automatically updates reports or Excel spread sheets

Revisit: What are the 6 Steps in the 60 Seconds Introduction to Data Science?



- https://www.youtube.com/watch?v=i2jwZcWicSY



References and Further Reading

- A. Boschetti and L. Massaron, *Python Data Science Essentials*, Chapters 1 and 2
- Murtaza Haider, Getting Started with Data Science: Making Sense of Data with Analytics, Chapters 1 and 3
- D. Cielen and A. Meysman and M. Ali, *Introducing Data Science*, Chapter 2

Data Science

Thanks!