IDSSP

International Data Science in Schools Project

Curriculum Frameworks for

Introductory Data Science

IDSSP Curriculum Team

September 2019

These frameworks are endorsed by:

- ACEMS
- American Statistical Association
- BCS The Chartered Institute for IT
- International Statistical Institute
- New Zealand Statistical Association
- Royal Statistical Society
- Statistical Society of Australia
- Teaching Statistics Trust

IDSSP

International Data Science in Schools Project

Citation: IDSSP Curriculum Team, *Curriculum Frameworks for Introductory Data Science*, http://idssp.org/files/IDSSP_Frameworks_1.0.pdf.

ISBN: 978-0-646-80819-2

Preface

Purpose of the frameworks

The International Data Science in Schools Project (IDSSP) is a cross-disciplinary project involving an international team of computer scientists and statisticians from the leading professional organizations for both disciplines. The purpose of the project is to promote and support the teaching of Introductory Data Science, particularly in the final years of schooling.

This document represents the first step. THE IDSSP team has developed two frameworks for introducing Data Science, including topics and learning outcomes. The first is a curriculum framework to underpin the design of courses to teach secondary school students, and the second serves the same purpose by preparing teachers to teach Introductory Data Science to their students.

The frameworks have been developed with a variety of possible applications in mind.

The initial, and primary, objective has been that the frameworks will provide the basis for development of courses in Introductory Data Science for students in their final two years of secondary school, and of courses to teach teachers how to teach Introductory Data Science.

However, the frameworks are suited to a range of other purposes:

(a) to supplement, broaden and enrich courses in Computer Science, Statistics and Mathematics.

(b) to develop some introductory learning-from-data modules for other data-rich domains of application. This includes subjects outside the traditional secondary Mathematics curriculum (*e.g.* Biology, Civics, Economics, Geology, Government and Social Science could include Data Science components).

(c) to develop courses at the tertiary level – university or community college.

(d) to inform curriculum development in local contexts.

(e) to enhance pre-service and in-service teacher training in preparation for teaching students about inquiry and learning about data, for teachers in all of the areas noted in (a) and (b).

The Curriculum Team

An international Curriculum Team with a supporting Advisory Group was recruited in 2017, and several leading statistical and computer science societies and other interested organizations were approached to provide in-principle support. The project itself was launched in 2018.

The international team comprised

Nicholas Fisher (Australia; Chair), Statistics
Ajay Anand (USA), Computer Science
John Bailer (USA), Statistics
James Bailey (Australia), Computer Science
Wesley Burr (Canada), Statistics
Alan Fekete (Australia), Computer Science
Alison Gibbs (Canada), Statistics
Robert Gould (USA), Statistics
Tim Hesterberg (USA), Statistics
Raymond Ng (Canada), Computer Science
James Rosenberger (USA), Statistics
Neil Sheldon (UK), Computer Science, Statistics
Chris Wild (New Zealand), Statistics

In addition there was a larger and more broadly representative Advisory Board whose membership is listed on the IDSSP website.

Supporters

The project has been supported financially by ACEMS, American Statistical Association, Google, University of Leiden Mathematical Institute, National Institute of Statistical Sciences, New Zealand Statistical Association, Royal Statistical Society, Statistical Society of Australia, Statistical Society of Canada, University of Technology Sydney, and ValueMetrics Australia. Preparation of the final versions of the report and the Frameworks document were made possible by a grant from the Australian Bureau of Statistics. Particular thanks are due to Louise Ryan, Distinguished Professor of Statistics at the University of Technology Sydney, for providing sage advice and seed funding during the critical early-development stages of the project.

In-kind support has been provided by the Association for Computing Machinery, Australian Council of Deans of Information and Communications Technology, BCS, The Chartered Institute for IT, Cambridge Mathematics, Dutch Society for Statistics and Operations Research, International Statistical Institute, and the Teaching Statistics Trust (UK). We also wish to thank the following individuals: Fanny Chevalier, Christopher Collins, Judy Kay, Jia Li, John Mack, Nancy Reid, David Spiegelhalter, and Andrew Turpin for providing expertise, feedback, contributions to the content of this document and other assistance.

Contents

Introduction

1. Project overview

The last decade has seen unprecedented growth in the availability of data in most areas of human endeavor. Whole branches of science have been developed to allow corporations to transform the way marketing is conducted, to drive scientific progress in areas such as Bioinformatics, and to inform decision-making at all levels in governments and industry. Further, the scale and complexity of much of these data are beyond the capability of a single computer to manage or a single individual to analyze.

These realities generate a very significant imperative to ensure that there is an adequate supply of people entering the workforce who are equipped to handle the new challenges of learning from data. There is a compounding factor. On the evidence available, demand for data scientists is not only massively outstripping supply, but the situation is worsening, and this is a world-wide problem.

And beyond this, there is an equally pressing need for people in our societies to be more capable of understanding, interpreting, critiquing and making decisions based on data as they cope with the vagaries of life.

The purpose of this international collaborative project is to transform the way teaching and learning about Data Science is carried out in the last two years of school, with two objectives:

- To ensure that school students acquire a sufficient understanding and appreciation of how data can be acquired and used to make decisions so that they can make informed judgments in their daily lives, as students and then as adults. In particular, we envisage future generations of lawyers, journalists, historians, and many others, leaving school with a basic understanding of how to work with data to make decisions in the presence of uncertainty, and how to interpret quantitative information presented to them in the course of their professional and personal activities.
- To instill in more scientifically able school students sufficient interest and enthusiasm for Data Science that they will seek to pursue tertiary studies in Data Science with a view of making a career in the area.

In both cases, we want to teach people how to *learn from* Data.

To achieve this, we aim to provide the content and resources to create a pre-calculus course in Data Science that is fun to learn and fun to teach, and prepares students well for their future lives.

The framework has been developed with the flexibility to be adapted to the available teaching time. Broadly speaking, we envisage that it could be used to develop courses ranging from about 240 hours to 360 hours of total instruction time depending on the level of detail included. As a parallel development we will devise a program that will enable teachers from a wide variety of backgrounds – basically from any discipline that involves data, or mathematics teachers – to learn to present a Data Science course well. It is also planned to make the course available in a variety of modes of delivery.

The project is being carried out in two phases, with the initial focus on the first phase:

- **Phase 1:** In the Curriculum phase (approximately 18 months), an international Curriculum Team (CT) comprising well-regarded computer scientists and statisticians, aided by an Advisory Group of computer scientists, statisticians, school teachers and curriculum experts, has developed these curriculum frameworks for the student and teacher programs.

- **Phase 2:** In the Implementation phase, the curriculum framework devised in Phase 1 will provide the basis for developing pedagogies and resources to support courses in a variety of formats, suitable for different modes of delivery (*e.g.* MOOCs, inverted classroom, traditional classroom, *etc.*).

The project involves computer scientists, statisticians, school teachers, curriculum experts and educators from Australia, Canada, England, Germany, the Netherlands, New Zealand and the United States. It is supported by leading international and national statistical and computer science professional societies.

Because of the extraordinary variety of educational jurisdictions across (and even within) the various countries involved, it would be impossible to create a single course that would satisfy all jurisdictional requirements. However, we believe that the combination of the curriculum framework developed, and the pedagogies and resources that are planned, will provide the flexibility for school systems and teachers to prepare curricula and present courses that meet our overarching goal of Data Science being fun to teach and fun to study.

Whilst the project is intended primarily to benefit school students and their teachers, we envisage that the materials will also be relevant to some tertiary institutions, for example, to prepare a course for a single year's study based on the first year of the Curriculum Framework. Other possible uses are listed in Section 9.

2. What do we mean by "Data Science"?

We interpret the term *Data Science as the science of learning from data*. As such, it draws on several disciplines, including aspects of Computer Science, Mathematics and Statistics, together with areas such as problem elicitation and formulation, collaboration and communication skills. At the heart of Data Science is the cycle of learning from data, as depicted in Figures 1 and 2.

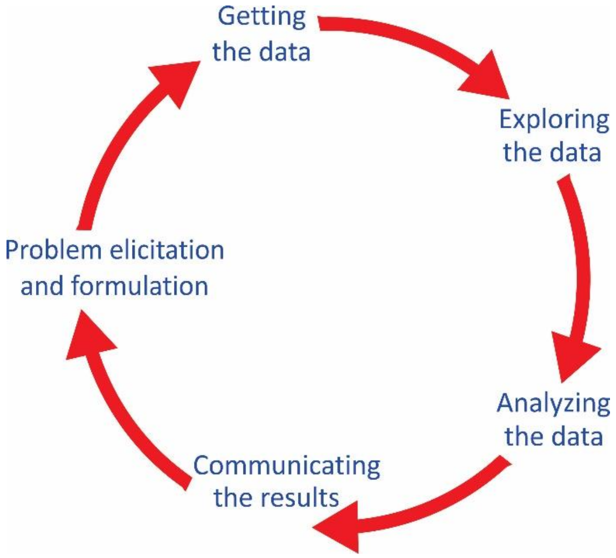
Getting the data
Exploring the data
Analyzing the data
Communicating the results
Problem elicitation and formulation

Figure 1 The basic cycle of learning from data

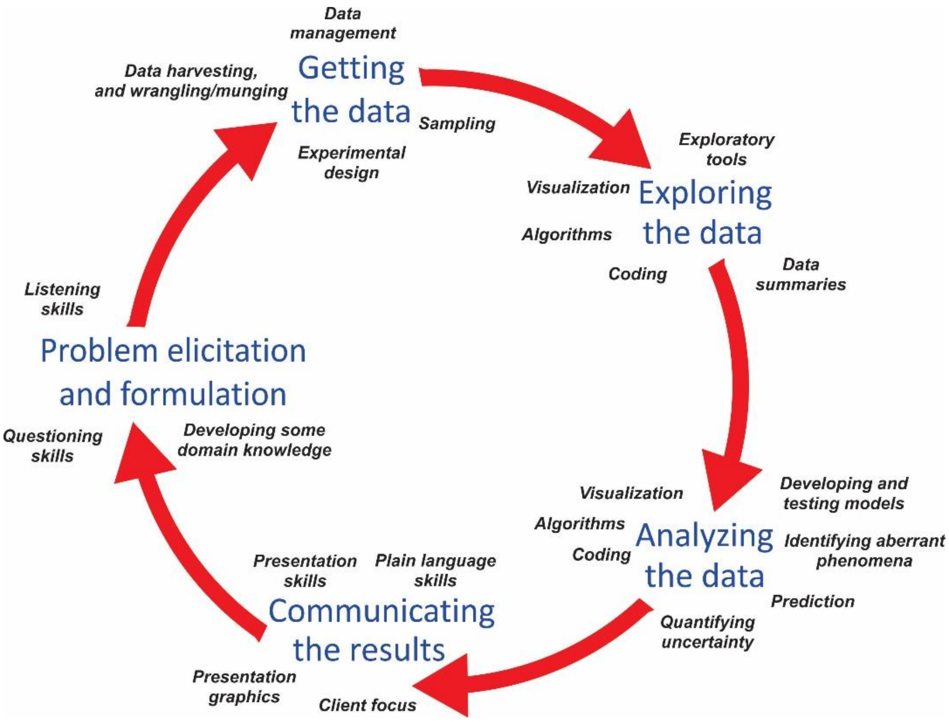

Figure 2 The varying activities involved in learning from data.

3. What do we mean by a "curriculum framework"?

The term "curriculum framework" requires clarification. As we noted earlier, there are very many educational jurisdictions around the world, so no single curriculum could possibly satisfy all requirements. Accordingly, our purpose is to devise specifications for what it is desirable to include in a modern Data Science Curriculum that can be customized to local requirements (*e.g.* where some elements have already been covered within existing subjects).

Our framework is a lot more detailed than might be expected for a high-level document. The reason for that is that very few people who will be reading and considering implementing this framework (in part or as a whole) will have a good overview of the broad sweep of Data Science and its various components. The IDSSP Curriculum Team felt that this framework needed to lay out a fairly detailed map of the Data Science landscape to provide a useful picture of what is involved and what could be made accessible to students at the senior high-school level. We stress that this framework only provides a map of a landscape to be traversed and not an ordered set of instructional sequences for traversing it.

4. What assumptions are made about prior student and teacher learning?

For students: no prior knowledge of either Computer Science or Statistics is assumed, nor any familiarity with calculus.

For teachers: no prior knowledge of either Computer Science or Statistics is assumed, nor any familiarity with calculus. In terms of their 'home' discipline, this might be Mathematics, or else a discipline where data plays an essential role, such as in any of the sciences – Physics, Chemistry, Biology, Geology, Geography, … – or may play an important role, such as the social sciences including Economics.

This allows us to create a stand-alone recommendation for a curriculum framework that, in the absence of statistical or computer science prior skills and knowledge, describes the basis for a complete introduction to Data Science. Of course, in many or most circumstances, students will have already had exposure to these subjects to some degree, allowing jurisdictions and teachers to adjust the curriculum framework to suit their local needs. Further, whilst students might have had little or nothing to do with data in earlier years of schooling, this curriculum framework is designed to provide immediate immersion in data, and learning from data.

By comparison with Statistics courses, there will be greater emphasis on computing aspects; and similarly, by comparison with Computer Sciences courses, there will be greater emphasis on statistical aspects. And in contrast to both Statistics courses and Computer Sciences courses, there will be greater emphasis on data and learning from data.

This is intentionally a contemporary curriculum framework that exploits and uses modern technology, a prerequisite for developing any significant capabilities in Data Science.

At present, there are relatively few teachers who have had any significant experience of working with data and so would not require some professional development be able to teach a course based on the curriculum framework. This is why the introductory Data Science curriculum framework for students (IDS) goes hand-in-hand with a curriculum framework for teaching the teachers (T3).

An important principle is that in T3, teachers should themselves experience the types of experiences desired for students of IDS. This means that desired teaching would be modeled to teachers by T3. (Additionally there will be a little more technical and experiential depth and pedagogical learning designed to help with the delivery of an IDS-based course.)

Given the teacher-workforce challenge and the fact that teachers are time-poor, a good implementation of T3 would also serve as an online repository for off-the-shelf course elements that teachers could, with minimal effort, re-purpose for their students. As teachers grow in experience and confidence, they will introduce their own examples (for local relevance, current news-worthiness, and empowerment); however a great starting point for an IDS class will be the resources utilized in T3. Similarly, teachers may in time move to other tools and platforms (for example, to match local workplace settings) but they can easily start with the ones from T3.

5. How will computational thinking and environments be introduced?

The relentless focus of the curriculum framework is *learning from data*. Accordingly Data Science concepts, **particularly as they relate to computational aspects**, will be introduced and treated as they are needed, as a means to an end rather than as an end in themselves. These issues are discussed in more detail in the Computer Science and Programming Strategy for Unit 1.

6. Phase 1: Developing the curriculum framework

The framework has been developed with the flexibility to be adapted to the available teaching time. Broadly speaking, we envisage that they could be used to develop courses ranging from about 120 to 180 hours of total instruction time (depending on the level of detail included) for each of two years.

Unit 1, envisaged as the first year of study, is designed to stimulate the interest of the student in learning from data. It seeks to heighten students' awareness of how data enter their daily lives. They will learn how they can acquire and explore data to understand the world around them: *How or where can I get some data to explore this problem? How do I start looking at it? How can I present the results convincingly*? They will start to develop a critical (scientific) approach to assessing what they hear and read in the media about data-based assertions: *How much confidence can I have in what I'm being told? What was the source of the data used to make these claims? Are there some biases, accidental or intentional, in what is being presented?* They will develop an awareness of where they are in the *learning-from-data* cycle in terms of the tools and techniques that are being used, and develop familiarity with computational environments to assist them in exploration, visualization, calculation and presentation. At all times, the focus will be on questions, problems and data that are meaningful to their lives and attendant social and ethical issues that arise in acquiring and working with data.

Unit 2, envisaged as the second year of study, then helps students develop familiarity with a wide variety of data types that occur in everyday life, the sorts of problems that they are used to tackle, and some tools and techniques for tackling these problems, all in the context of the *learning-from-data* cycle. As well as introducing new concepts, it draws on and reinforces the range of concepts introduced and skills developed in Unit 1. Whereas components (Topic Areas) of Unit 1 are regarded as providing the basis for a suitable introduction, Unit 2 comprises rather more material than would normally be included in a single year of learning of learning – so that a program of learning would normally be fashioned from a selection of Topic Areas.

In terms of prerequisite skills and knowledge, no prior knowledge of calculus is required. However, it will be assumed that students and teachers have some familiarity with using computers, and that anyone studying a course based on this curriculum framework has access to a computer with reasonable visualization capability. This is as essential to learning how to work with data in the 21st Century as is access to modern laboratories for studying Biology, Chemistry or Physics. Whilst no prior knowledge of software packages is assumed, students will acquire competence with at least one (freely-available) software language or package (probably Python or R) as they progress through the course.

7. The structure of each Unit

The curriculum framework is specified in terms of two **Units** each intended to correspond to one year of study.

- The Units are broken down into ***Topic Areas*** (clusters of closely related concepts and skills) that are further broken down into ***Topics***.
- Unit 1 is made up of seven Topic Areas that lay the foundations for the curriculum. The Unit 1 Topic Areas are intended to be *approximately* the same size in terms of learning time.
- Unit 2 addresses specialized subareas of Data Science. Unit 2 has been designed so that, after completion of Unit 1, the Unit 2 Topic Areas are largely self-contained options from which a course of study can be assembled.
- The curriculum framework for teachers comprises the curriculum framework for students plus other material in blue. The basic idea is that the teachers' curriculum should include everything in the

students' curriculum plus some additional Data Science content plus some pedagogical content. It is much less developed than the student curriculum.

8. Phase 2: Developing pedagogies, learning materials and resources

After the completion of the curriculum framework (Phase 1 of the project), we shall carry out a feasibility study for carrying out Phase 2, which has the goal of bringing this curriculum framework to life by providing pedagogies, content and resources needed to assemble courses for teaching or self-learning. Assuming a positive outcome from this assessment, we shall proceed to Phase 2.

In this second phase, we envisage creating and assembling outstanding audio-visual materials, webpages, activities, assessment strategies, data resources and information on sources of data, software products and sources of software, articles and books. The intention is to facilitate a variety of delivery modes, including classroom, inverted classroom, MOOC, and self-instruction.

This Phase would also address the issue of developing ways to assess a teacher's level of skill and knowledge as a prospective teacher of Data Science. We are interested in drawing prospective Data Science teachers from teachers with a backgrounds as diverse as Agriculture, Biology, Chemistry, Computer Science, Earth Sciences, Economics, Geography, Social Studies, Legal Studies, Mathematics or Physics. Teachers from many of these areas will be able to bring their own data for use in teaching Data Science; and conversely, their teaching in their own disciplines will be enriched by the skills and knowledge gains from learning how to learn from data.

9. Who can use this Framework document and how

Despite the IDSSP having plans for a further phase of work building on the frameworks described here, people are encouraged to use and adapt them, in part or whole, for their own educational purposes. Attribution should be made to the *2019 IDSSP Curriculum Framework* (www.idssp.org).

Other purposes for which elements of the frameworks are suited include:

(a) Supplementing and enriching courses in Computer Science, Statistics and Mathematics.

(b) Developing some introductory learning-from-data modules for other data-rich domains of application. This includes subjects outside the traditional secondary Mathematics curriculum (*e.g.* Biology, Civics, Economics, Geology, Government and Social Science could include Data Science components).

(c) Developing introductory courses at the tertiary level (university or community college).

(d) Developing courses for earlier years of schooling.

(e) Informing curriculum development in local contexts.

(f) Enhancing pre-service and in-service teacher training in preparation for teaching students about inquiry and learning about data, for teachers in all of the areas noted in (a) and (b).

10. A simple interactive search tool for searching the Topics

Please visit http://www.idssp.org/pages/SearchableIndex.html for a simple way of searching the Index.

Unit 1

1. Introduction

Unit 1 lays out a set of introductory topics that constitute the foundation of the curriculum framework.

It is envisaged to require about 120 – 180 hours of study depending on the level of detail included. It aimed to give students a flying start in learning about Data Science, to develop their enthusiasm for the subject and what it may mean for them in their future lives, and to stimulate learning about how they – personally – can utilize data in their daily lives.

A note on technical language: Computer Science, Statistics and other areas often use different words for the same idea. Anticipating that many readers will be diving into parts of the document, rather than reading from the beginning, we have endeavored to ensure that the exposition will understandable across communities. One strategy is to use a compound word, like "feature/variable" to follow, which combine the most common term used by computer scientists with the most common term used by statisticians.

Unit 1 comprises seven *Topic Areas:*

- 1.1 Data Science and Me
- 1.2 Basic techniques for Exploration and Analysis (BTEA). Part 1: Tools for a single feature/variable
- 1.3 BTEA Part 2: Pairs of features/variables
- 1.4 BTEA Part 3: Three or more features/variables
- 1.5 Graphical displays and Tables
- 1.6 The Data-handling Pipeline
- 1.7 Avoiding being misled by data

In Topic Area 1.1 (*Data Science and Me*) people are introduced to data and Data Science through discussion and case-studies, bringing home the importance of Data Science in their lives. It provides glimpses into the world of Data Science and raises many big issues including social and ethical aspects. It also introduces the Data-Science *learning cycle* (Figure 1) which provides an organizing principle for the entire curriculum: the cycle is to be kept constantly in view, with the teacher continually drawing attention to where they are working in the cycle, what has gone on before and what comes next.

Topic Areas 1.2 – 1.4 (*Basic techniques for exploration and analysis Parts 1-3*) empower the students to uncover interesting stories in multivariate data. They provide just enough information about rectangular data files to get students launched into a computational environment where they can start to explore and analyze data, and with sufficient knowledge about a useful set of plots and summaries to be able to make discoveries. Parts 1, 2 and 3 correspond, respectively, to tools involving 1, 2 and 3 or more features/variables simultaneously.

Following this initial experience with exploratory tools and default settings, Topic Areas 1.5 (*Graphical displays and Tables*) 1.6 (*The Data-handling Pipeline*) and 1.7 (*Avoiding being misled by data*) revisit aspects of the earlier parts of the framework in more depth, with more care and a greater breadth of coverage, to enable students to make conscious choices through having greater control of their work:

- Topic Area 1.6 (*The Data-handling Pipeline*) provides an introduction to the tools used to deal with data (sometimes called 'data wrangling') across the whole data-science lifecycle from inception through to presentation. While this gives students some preliminary skills in using a programming language and experience with automation, the main focus is on understanding the power of the tools and the importance of managing data carefully.
- The topics in Topic Areas 1.5 (*Graphical displays and Tables*) and 1.7 (*Avoiding being misled by data*) share a broad theme of trying to ensure that we are extracting the real messages stored in the data and not being misled. *Graphical displays and Tables* considers this from the viewpoint of how data are presented to us (or by us) in graphical or tabular form. *Avoiding being misled by data* explores the same theme in the context of the data themselves and their provenance, addressing issues such as bias, confounding (lurking features/variables that cause problems in drawing causal conclusions) and random error, and ways of handling these problems.

Summary of Aims for each Topic Area

1.1 Data Science and Me

Aims: To help students become aware of the importance of data to their lives, the exciting possibilities opened up by Data Science, some of the big ideas of Data Science, related social and ethical issues, and to introduce students to the Data Science learning cycle.

1.2 Basic techniques for exploration and analysis, Part 1: *Tools for a single feature/variable*

1.3 Basic techniques for exploration and analysis, Part 2: *Pairs of features/variables*

1.4 Basic techniques for exploration and analysis. Part 3: *Three or more features/variables*

Aims: To provide a foundational introduction to simple data storage formats, graphical displays and numerical summaries that are useful in their own right; to provide a basis for building on in many subsequent units; and to give students experience with using these tools to make discoveries in data.

1.5 Graphical displays and Tables: how to construct them and when to use them

Aims: to develop the students' understanding of appropriate choices and uses for graphical displays and tables in learning from data and when presenting the results of an analysis. In particular,

(a) To demonstrate the essential role of graphical methods in revealing pattern and unusual elements in the initial exploration phase and when evaluating the adequacy of fits of models to data; and in effective communication of findings.

(b) To show how tables are used to communicate exact values, or to present summaries of results that are too complex to be conveyed graphically.

1.6 The data-handling pipeline

Aims: to give an introduction to the tools used to deal with data (sometimes called data wrangling), and to develop an understanding of data management issues in the context of the Data Science learning cycle.

1.7 Avoiding being misled by data

Aims: To provide students with a deeper understanding of how to critique data and data-based claims, including an appreciation of the ideas of bias, confounding and random error; to introduce them to some good practices for obtaining reliable data (random sampling and randomized experiments); to motivate incorporation of uncertainties in estimation using margins of error or interval estimates; and to provide some introductory experience with the ideas of statistical testing in the context of an experiment for comparing 2 treatments.

Computer Science and Programming Strategy

For most of **Unit 1**, it is envisaged that students would tend to use point-and-click systems or modify code that has already been provided. To help students graduate from the point-and-click world to actually running code, point-and-click systems that expose the underlying code are desirable. It is not a goal of this framework to turn students who do not know how to program into competent programmers. Rather, the intent is to introduce computer code almost by stealth, with three main goals:

- to act as an ice/fear-breaker by introducing interacting with code in ways that are fun and not at all intimidating;
- to enable students to experience the power and versatility of code-based approaches to Data Science problems; and
- as a convenient way to perform specific tasks.

Although this strategy can be applied to some extent in any of Topic Areas 1.2-1.7, it is in the *Data-handling Pipeline*Topic Area (1.6) that the roles of computer science and programming become of central importance.

This Topic Area deals with two topics that are also found in traditional computing courses: data management, and programming.

However, the approach is deliberately very different from that taken in computing courses. We want to be sure that the overlap is limited, so students can study both Computer Science and Data Science if both are offered. Also, the programs built from our framework must be attractive to those who believe they lack talent in computing topics. For data management, traditional Computer Science focuses on cases where the data are held in a centrally-managed database which protects data integrity. However Data Science applications often employ more *ad hoc* approaches. So in this framework, we generalize many of the concepts and show practices that achieve good outcomes without much support from the platform.

For programming, the typical objective in Computer Science courses is mastery, so that the student can write code from scratch, given a task description. Here, however, the focus is on understanding the power of automating Data Science tasks, and the skills are more at the level of writing a few lines, or modifying existing code.

Details of Unit 1 Topic Areas

1.1 Data Science and Me

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	To help students become aware of the importance of data to their lives and their society, the exciting possibilities opened up by Data Science, related social and ethical issues, and to introduce students to the Data Science learning cycle	As for Students
Learning outcomes	Students will be able to: • Express what data are and where they come from. • Identify moments and occasions in their daily lives in which data are collected about themselves. • Demonstrate some of the roles data play in their own lives. • Describe types of problems that can be answered with Data Science. • Categorize questions based on whether they can be answered with Data Science. • Construct classification and prediction oriented problems about real-life situations. • Provide examples of the social and personal consequences of predictions derived from models built on data. • Identify and describe errors in decisions and predictions owing to faulty use of data. • Discuss how and when data can support making decisions. • Categorize aspects of their problem-solving process with respect to phases of the Data Science Learning Cycle. • Describe issues of privacy and security.	**Additional learning outcomes** *(Pedagogical)* Able to: Lead student discussions and extract key issues from student contributions. If from a discipline other than Mathematics or Computer Science: Communicate the relevance of Data Science to their own discipline drawing on their own background for stories and data. Enhance their teaching of their own discipline. If coming from Mathematics or Computer Science: Communicate the relevance of Data Science to other disciplines and use them as a source of stories and data.
Key phrases: Importance of data; Data Science in real life; Data Science Learning Cycle		

Parts of Data Science learning cycle addressed: All of it at a high level
Prior knowledge required: None

Commentary on the learning outcomes

This Topic Area aims to build awareness about roles of data in society and students' lives, and key issues of data science. The learning outcomes address the ability to discuss these issues and contribute ideas about them.

The Topic Areas in this curriculum framework and the sets of Topics within them do not necessarily define a teaching sequence. For example, many teachers may prefer to defer discussion of some of the Topics listed below until their students have had some hands-on experience with investigations, data, data exploration and analysis. Decisions about initial teaching sequences will be driven in large part by the judgements of teachers about what is most likely to draw their students in and excite them.

The intent for this Topic Area is that some learning objectives will be achieved early and the rest by the end of the course. The Data Science Learning Cycle (Topic 1.5) is, however, a fundamental underpinning of all Topic Areas.

1.1 Data Science and Me

Topic	**1**	**What is Data Science?**
Subtopic	**1.1**	Role of data in decision-making — at home, in government, business, industry, sport, ...
	1.2	Introduction to Data Science and the Data Science learning cycle.
	1.3	Data Science success stories.
	1.4	Data Science disasters.
	1.5	Elaboration of steps in the data-science cycle.
	1.6T	Diverse uses of the term " Data Science ".
	1.7T	Sources of information about Data Science and its activities.

1.1 Data Science and Me

Topic	**2**	**What does Data Science have to do with me?**
Subtopic	**2.1**	Data I keep. What sorts of data do I keep for my own purposes and why do I keep them (*e.g.* passwords, birthdays, sports performances, assessment due dates, to-do lists, website addresses, images, music files, ...)?
	2.2	Data about me. When was the last time I am aware of data being collected about me? Who was doing it? How did they do it? Why were they doing it? Should I be worried about it? Can important benefits result from this sort of data collection? What about harmful effects?
	2.3	Data on friends and family. What are some examples of data being collected about my friends and family members? Who is doing it? How did they do it? Why are they

Topic	2	What does Data Science have to do with me?
		doing it? Should we be worried about it? Can important benefits result from this sort of data collection? What about harmful effects?
	2.4	What are data? How are the data collected? What is recorded (features/variables)?
	2.5	Privacy, security and openness/accessibility — issues and trade-offs.
	2.6T	Pedagogical issues relating to leading discussions and extracting issues.

1.1 Data Science and Me

Topic	3	Sources of Data
Subtopic	**3.1**	Examples of data.
	3.2	What *are* data?
	3.3	How do we get useful data? Primary *versus* secondary data.
	3.4	Privacy, security and openness/accessibility -- issues and trade-offs.
	3.5	Thinking critically about data: introduction.
	3.6T	Thinking critically about data: data quality and GIGO (garbage in, garbage out).
	3.7	Thinking critically about data: ways in which data need critical appraisal.
	3.8	Pedagogical issues relating to these topics.

1.1 Data Science and Me

Topic	4	Examples of Data Science problems
Subtopic	**4.1**	Ideas (with examples) about using data for: description, prediction, classification, clustering, (causal) explanation, and control.
	4.2	Examples of social and personal consequences.
	4.3	How can the prediction process go wrong?
	4.4	Examples of causal explanation and use for control.
	4.5	How can the process of finding causes go wrong?
	4.6T	Pedagogical issues relating to these topics.

1.1 Data Science and Me

Topic	5T	Extracting pertinent lessons from student discussions (Teachers only)
Subtopic	**5.1T**	Leading student discussions and extracting key issues from student contributions.
	5.2T	Story Telling.

[*Back to Unit 1 contents page*]

Topic Areas 1.2 - 1.4: Basic techniques for exploration and analysis

Overview

- The Basic techniques Topic Area has been broken into 3 parts
  - Part 1: Tools for a *single* feature/variable (Topic Area 1.2)
  - Part 2: *Pairs* of features/variables (Topic Area 1.3)
  - Part 3: *Three or more* features/variables (Topic Area 1.4)
- The emphasis is on providing a useful set of plots and summaries for exploring and making discoveries from data, and in gaining experiences in doing so. The coverage of tools is selective rather than comprehensive.
- The aim is to get students working with multivariate data as soon as possible, and reaching the point of being able to find interesting stories in such data as quickly as possible.
- The tools (plots and summary statistics) discussed need to be understood just well enough to serve these exploratory purposes — their uses as tools— rather than any understanding of fine details or why they work. The aim is for students to be empowered to make discoveries from early on.
- **Computing.** It is envisaged that students will normally use point-and-click systems, or notebooks (*cf*. R Markdown documents and Jupyter notebooks), or even call high-level functions from a programming system (*e.g.* ggplot2 in R). Experience with computer code for most students will be at the level of making minor changes to code that has already been provided. This enables the primary emphasis to be on what things mean, and reasoning using these tools in exploring data.
- The orderings of topic presented here are not intended to be prescriptive teaching sequences. They do not preclude problem-based learning approaches in which elements from several Topic Areas, from either Unit 1 or 2, are encountered as they arise during investigations involving real data.

  Regarding the basic techniques topics in particular, elements of Topic Area 1.4 can be used early in 1.2 and 1.3. For example:
  - Interactive versions of a basic plot (*cf*. Topic Area 1.4 Topic 4) can be used immediately after introducing the basic plot, both to improve understanding of the plot and to increase its potential for discovery.
  - As soon as a basic plot type is understood, showing and comparing separate plots for males and females (*cf*. Topic Area 1.4 Topic 2) is a simple, easily understood extension that increases the ability to make interesting discoveries.

Similarly, coloring points in dot plots and scatter plots (Topic Areas 1.2 & 1.3) according to group membership (*cf.* Topic Area 1.4 Topic 3) is a simple extension that increases the ability to make interesting discoveries.

(*Back to Unit 1 contents page*)

1.2 Basic techniques for exploration and analysis. Part 1: Tools for a single feature/variable

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	**For Topic Areas 1.2 - 1.4 as a whole:** • To provide a simple foundational introduction to a selection of graphical displays and numerical summaries that enable students to start working with multivariate data as soon as possible, so that students are able to find interesting stories in such data as quickly as possible. • To enable students to feel empowered to explore data from early on. • To provide a basis for many subsequent Topics; and to give students experience with using these tools to make discoveries in data. • Continue to develop habits of mind in relation to data: – Asking self and others questions about data origins, quality and applicability to a problem under consideration. – Usefulness for generalization. – Suggesting/hypothesizing possible generalizations.	As for students but at a deeper level of understanding and with greater technical mastery, so that they can guide students and assist them in their experiences; also to enable deeper reflection on what they are doing.
Learning outcomes Topic Area 1.2	Students will be able to: • Import a rectangular data set in csv or tab-delimited text format into suitable software. • Devise a data cleaning plan for a given set of data and explain problems that may arise from failure to clean data. • Create and identify basic numerical and graphical summaries for individual features/variables in a data set.	**Additional learning outcomes** Able to: understand and work with the formulae used for commonly-used numerical summaries **for a single feature/variable.** have nuanced discussion of strengths and weaknesses of different commonly used graphical forms, including those that are not actively promoted by the course (such as pie charts or stacked bar charts).

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	• Explain the purpose and important properties of basic numerical summaries for individual features/variables. • Select appropriate numerical and graphical summaries to answer questions posed about a single feature/variable in a data set. • Interpret numerical and graphical summaries in the context of the original problem to answer questions posed about the original problem and make discoveries. • Use graphical displays to provide approximate numerical summaries. • Students will continue to develop a critical faculty in relation to data by: • Posing questions about the source of the data, data quality, and applicability to a problem under consideration. • Classifying questions and hypotheses as to whether they apply to the sample at hand or to a larger population. • Posing questions and generating hypotheses about target populations. • Evaluating the usefulness of a given data set for generalization to a target population.	Knowledge of pedagogical issues about these topics, so they can teach them effectively.
Parts of Data Science learning cycle addressed	*Directly:* • Getting the data (data import plus a little cleaning). • Exploring/Analyzing the data (predominant focus). • Communicating conclusions (communicating what they see in plots and summaries, and discussing possible implications). *Indirectly:* • “Problem elicitation and formulation” via probing more deeply (for examples/cases used) why and how the data were collected and the nature of the features/variables. *Variations:*	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	• Borrowing some data-harvesting elements of Topic Area 1.6 (*e.g.* a largely scripted web-scraping from somewhere interesting) can enable starting "from the beginning" with small time costs.	

Commentary specific to Topic Area 1.2

- Treating this as a univariate module looking at single-feature/variable data sets should be avoided. A lot of teaching experience has highlighted this as a root cause of boredom.
- A better contextual wrapping is provided by using multivariate data addressing some interesting problem(s) and taking an initial look to see what the data on each feature/variable look like, and whether there are any anomalies in them (a data-cleaning impulse). To do this, the student needs to understand how to read and interpret basic plots and summaries. The displays themselves will generally be generated automatically by software. (The exception is where a hands-on activity can help with the understanding of what something is.)
- The intent is to move as quickly as possible from single features/variables to comparing groups (formally in Topic Area 1.3). Some of the "learning to read a display" can even be done after that jump has been made. Students can often make a good stab at what a display is saying before formal teaching about it and it is desirable to exploit those intuitions.

1.2 BTEA, Part 1: Tools for a single feature/variable

Topic	1	Rectangular data sets
Subtopic	1.1	Observations and features/variables.
	1.2	Numerical *versus* categorical features/variables.
	1.3	Importing data files in simple formats (*e.g.* csv, tab separated text) into a suitable computer system for analysis.
	1.4	Idea that data sometimes has to be cleaned (driven by data used).
	1.5T	Knowledge of Topic Area 1.6.
	1.6T	Pedagogical issues relating to these topics.
Prior knowledge required		Whole and real numbers, order, basic arithmetic, ratio

1.2 BTEA, Part 1: Tools for a single feature/variable

Topic	2	Graphics and summaries for a single categorical feature/variable
Subtopic	2.1	Frequency tables and bar charts (counts and proportions).
	2.2	Good ways of ordering groups for displays and tables: alphabetically/by frequency/natural order (ordinal features/variables).
	2.3T	Proportions *versus* counts: what works best for what?

	2.4T	Weaknesses of pie charts and stacked bar charts.
	2.5T	Pedagogical issues relating to these topics.
Prior knowledge required	Topic 1 of this Topic Area, fractions/proportions/percentages, whole and real numbers, order, basic arithmetic, ratio	

1.2 BTEA, Part 1: Tools for a single feature/variable

Topic	**3**	**Graphics and summaries for a single numeric feature/variable**
Subtopic	**3.1**	Graphics: Dot plots and histograms as large-data-set alternative. Superimposing slider-controlled density estimates on dot plots or histograms (but not for small samples).
	3.2	Characteristics to look out for and their implications: outliers, center, spread, shape, modality, spikes, gaps, ...
	3.3	Summaries: Minimum and maximum, mean and median, quartiles, interquartile range and standard deviation.
	3.4	Box plot as plot of summaries.
	3.5	Converting numerical features/variables to categorical: When, why and how? (Numbers as codes, binning continuous features/variables.)
	3.6T	How the measures in 3.3 are obtained (working with formulae).
	3.7T	Pedagogical issues relating to these topics.
Prior knowledge required	Fractions/proportions/percentages, whole and real numbers, order, basic arithmetic, ratio.	
From this course	Topic 1 and Topic 2 of this Topic Area, having previously dealt with calculating a mean and a median, whole and real numbers, order, basic arithmetic, ratio.	

[*Back to Unit 1 contents page*]

1.3 Basic techniques for exploration and analysis. Part 2: Pairs of features/variables

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	**For Topic Areas 1.2 - 1.4 as a whole:** See Topic Area 1.2	
Learning outcomes Topic Area 1.3	Students will be able to: • Create and interpret commonly-used graphical displays for two-feature/variable relationships. • Identify important and interesting features of graphical displays relating pairs of features/variables in the context of the original data. • Relate commonly used data summaries for pairs of features/variables to characteristics of graphical displays. • Identify when a graphical display for a pair of features/variables may be useful for answering questions posed about the data. • Formulate questions that can be answered with graphical and numerical summaries for pairs of features/variables. • Demonstrate the ability to use graphical and numerical summaries to answer questions about the original problem and make discoveries. • Students will continue to develop a critical faculty in relation to data by: • Posing questions about the source of the data, data quality, and applicability to a problem under consideration. • Classifying questions and hypotheses as to whether they apply to the sample at hand or to a larger population. • Posing questions and generating hypotheses about target populations.	**Additional learning outcomes** Able to understand and work with the formulae used for commonly-used numerical summaries for a pair of features/variables. Sufficiently more guided hands-on experience with the elements students are learning about to be competent and confident in interpreting and critiquing student-generated output and to suggest other things they might do.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Evaluating the usefulness of a given data set for generalization to a target population.	
Parts of Data Science learning cycle addressed	*Directly:* - Exploring/analyzing the data (predominant focus). - Communicating conclusions (communicating what they see in plots and summaries, and discussing possible implications). *Indirectly:* - Getting the data: Data exploration to detect anomalous characteristics is an important tool for data cleaning. - Problem elicitation and formulation: probing more deeply (in the examples/cases used) why and how the data was collected and the nature of the features/variables. *Variations:* - Borrowing some data-harvesting elements of Topic Area 1.6 (e.g. a largely scripted web-scraping from somewhere interesting) can enable starting "from the beginning" with small time costs.	

1.3 BTEA, Part 2: Pairs of features/variables

Topic	**1**	**Comparing groups** (relationships between a numeric and categorical feature/variable).
Subtopic	**1.1**	Making comparisons using the graphical displays and summary types learned for a single feature/variable in Topic 3 of Topic Area 1.2.
	1.2	Interpreting "group comparisons" in terms of "the relationship between a numeric and a categorical feature/variable".
	1.3	Extension to panel plots as a means of investigating the extent to which group differences look consistent across subpopulations or time.
	1.4T	Pedagogical issues relating to these topics, including how to start to get students, habitually, to ask questions worrying about data quality and applicability.
	1.5T	Comparing and evaluating different presentations.
Prior knowledge required		Topics 1 & 3 of Topic Area 1.2

1.3 BTEA, Part 2: Pairs of features/variables

Topic	**2**	**Relationships between two numerical features/variables**
Subtopic	**2.1**	Scatter plots.

Topic	**2**	**Relationships between two numerical features/variables**
	2.2	Outcome/Response features/variables *versus* Predictor/Explanatory features/variables
	2.3	Construction
	2.4	Structure in scatter plots: trend, scatter and outliers; clusters Seeing structure and capturing/emphasizing structure by sketching on top of computer-generated plots.
	2.5	Basic ideas of prediction.
	2.6	Vertical strips as a guide for sketching trend curves by eye.
	2.7	How predictions can fail.
	2.8	Idea of minimizing average prediction errors.
	2.9	Obtaining trend lines, curves and slider-controlled smooths from software.
	2.10	(Straight) lines and interpreting the intercept and slope coefficients of a trend line.
	2.11	Positive and negative associations, strong *versus* weak *versus* no association(s), correlation coefficients; association/correlation in relation to causation.
	2.12	Modifications to scatter plots to overcome perceptual problems with overprinting and large data sets: • jittering and transparency. • running quantiles (medians, quartiles, and more for very large data sets e.g. 10th and 90th percentiles). • large data alternatives to the scatter plot (*e.g.* hexplots).
	2.13T	Working with algebraic expressions for sum of squared errors, the least squares problem, least squares estimates for a linear relationship between two features/variables, and correlation.
	2.14T	Pedagogical issues relating to these topics.
Prior knowledge required		Topics 1 & 3 of Topic Area 1.2

1.3 BTEA, Part 2: Pairs of features/variables

Topic	**3**	**Relationships between categorical features/variables**
Subtopic	**3.1**	Two-way tables of counts and proportions.
	3.2	Side-by-side and separate bar charts or dot charts of proportions as complementary views.
	3.3T	Pedagogical issues relating to these topics.
Prior knowledge required		Topics 2 of Topic Area 1.2

1.3 BTEA, Part 2: Pairs of features/variables

Topic	**4**	**Filtering Data - using just the data on a subset of particular interest**
Subtopic	**4.1**	Filtering data by levels of a categorical feature/variable (*e.g.* girls only), intervals of a numeric feature/variable (*e.g.* an age group) or combinations to focus attention on a subgroup of particular interest and analyzing the filtered data.
	4.2T	Pedagogical issues relating to these topics.

[*Back to Unit 1 contents page*]

1.4 Basic techniques for exploration and analysis. Part 3: Three or more features/variables

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	**For Topic Areas 1.2 - 1.4 as a whole:** See Topic Area 1.2	
Learning outcomes Topic Area 1.4	Students will be able to: • Create graphical displays and tables to answer questions posed about the relationships among three or more features/variables. • Interpret graphical displays and tables to answer questions posed about the relationships among three or more features/variables. • Identify important or interesting features of a graphical display for three or more features/variables in the context of the original problem. • Create and interpret commonly-used numerical summaries of data in order to answer questions about the original context. • Students will continue to develop a critical faculty in relation to data by: • Posing questions about the source of the data, data quality, and applicability to a problem under consideration. • Classifying questions and hypotheses as to whether they apply to the sample at hand or to a larger population. • Posing questions and generating hypotheses about target populations. • Evaluating the usefulness of a given data set for generalization to a target population.	**Additional learning outcomes** Sufficiently more guided hands-on experience with the elements students are learning about to be competent and confident in interpreting and critiquing student-generated output and to suggest other things they might do.
Parts of Data Science learning cycle addressed	*Directly:* • Exploring/Analyzing the data (predominant focus).	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	• Communicating conclusions (communicating what they see in plots and summaries, and discussing possible implications). *Indirectly:* • Getting the data: Data exploration to detect anomalous characteristics is an important tool for data cleaning. • Problem elicitation and formulation: probing more deeply (in the examples/cases used) why and how the data was collected and the nature of the features/variables. *Variations:* • Borrowing some data-harvesting elements of Topic Area 1.6 (*e.g.* a largely scripted web-scraping from somewhere interesting) can enable starting “from the beginning” with small time costs.	

1.4 BTEA Part 3: Three or more features/variables

Topic	1	**Pairs plots** (matrices of 2-feature/variable plots for all pairs of features/variables in a chosen set of features/variables).
Subtopic	**1.1**	Pairs plots that will cope with categorical as well as numerical features/variables.
	1.2T	Pedagogical issues relating to these topics.
Prior knowledge required		Topic Areas 1.2 & 1.3

1.4 BTEA Part 3: Three or more features/variables

Topic	2	Subsetting by a third feature/variable
Subtopic	**2.1**	Panel plots/faceting and 3-dimensional summary tables as a means of investigating the extent to which two-feature/variable relationships look consistent across subpopulations or through time, or show some sort of trend.
	2.2T	Playing or stepping through the sequence of plots in panel display (the “playing” version creates the dynamic “motion plot” effect Hans Rosling was so well-known for).
	2.3	Highlighting subgroups in a scatter plot or dot plot and stepping through the groups to be highlighted.
	2.4T	Pedagogical issues relating to these topics
Prior knowledge required		Topic Areas 1.2 & 1.3

1.4 BTEA Part 3: Three or more features/variables

Topic	**3**	**Other ways of adding information on additional features/variables to 1- and 2-feature/variable plots**
Subtopic	**3.1**	**Coloring** points in dot plots and scatter plots according the value of an additional feature/variable.
	3.2	**Sizing** points in scatter plots according to the value of an additional feature/variable.
	3.3	**Labeling** points in dot plots and scatter plots according the value of an additional feature/variable (usually a name or value; most often applied to extreme points).
	3.4	Strengths and weaknesses of methods of adding information.
	3.5T	Pedagogical issues relating to these topics.
Prior knowledge required		Topic Areas 1.2 & 1.3

1.4 BTEA Part 3: Three or more features/variables

Topic	**4**	**Interactive plots** *(Plots, usually viewed in a web browser, that allow the user to query the plot in various ways using gestures like mouse-overs, clicking and brushing)*
Subtopic	**4.1**	*(Interactive versions of the plots previously seen)* Plots that allow querying of elements in a single plot using gestures like mouse overs, clicking and brushing, to identify elements in the plot (*e.g.* hovering over a point in a scatter plot and seeing the name of the person/unit represented by that point) and relationships between elements of a single plot. Sometimes, plot elements are linked to the contents of a table of data.
	4.2	Linked plots, linked plots and tables: points or elements selected in one plot/table lead to the highlighting of corresponding elements in all linked plots/tables.
	4.3T	Pedagogical issues relating to these topics.
Prior knowledge required		Topics from Topic Areas 1.2 & 1.3 relevant to the examples presented.

[*Back to Unit 1 contents page*]

1.5 Graphical displays and Tables: how to construct them and when to use them

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	The previous Topic Areas give students quite a lot of experience in using a wide range of graphical displays (plots/charts) and tables to uncover stories or extract meaning from data. This Topic Area takes a step back and starts pulling together ideas about representing data well to facilitate learning, discovery and presentation. Two different issues are being studied throughout this Topic Area – exploration/discovery from data *versus* presentation of results, and graphical displays *versus* tables as methods for either exploration or presentation purposes. Graphical displays are essential tools in learning from data: in the initial exploration phase; when evaluating the adequacy of fits of models to data; and ultimately, when communicating findings. *Their principal role is to show pattern*. In contrast, tables are used to communicate exact values, or to present summaries of results that are too complex to be conveyed graphically. The primary focus of this Topic Area is to develop the students' understanding of appropriate choices and uses for graphical displays and tables in learning from data and when presenting the results of an analysis. The principles used are also relevant to a secondary purpose: their uses in infographics and infotables, both of which are aimed at drawing attention to an article or story.	Teachers will be introduced to a number of basic principles relating to selecting a type of graphical display and using it effectively, and also being able to identify when to use a graphical display and when to use a table.
Learning outcomes	Students will be able to: • Create and evaluate graphical displays with the intent of: – identifying issues of data cleaning.	**Additional learning outcomes** As for students, but at a deeper level of understanding and with a greater level of technical mastery, so that they can guide students and assist them in their

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	– identifying potentially interesting patterns and unusual features in a given set of data. – communicating findings and information. – (optional) determining whether a proposed model is an adequate or appropriate description of the data – Identifying flaws in provided graphical displays that hide or distort information, and provide improvements. • Apply principles of graph construction to decide the type of graphical display that will best address a question or issue. • Create and evaluate tabular (table) displays with the intent of: – summarizing data and answering questions about the original context or problem. – interpretation for the purpose of answering questions in the original context. – presenting important findings and supporting conclusions. Students will additionally critically examine and make and defend judgments about the above by demonstrating the ability to: • Identify whether a table or graphical display is best for communicating the findings from a given set of data. • Evaluate whether or not a provided graphical display or table supports an argument or conclusions concerning the original data. • Contrast the relative merits of graphical displays *versus* tables.	experiences; also greater reflection on what they are doing.
Key words and phrases:	Purpose of a graph; Purpose of a table; Infographics; Presentation graphics; Infographics; Infotables; Exploring data; Discovering pattern	

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Parts of Data Science learning cycle addressed	*Directly:* • Exploring the data. • Communicating conclusions (especially via graphical displays and tables).	

1.5 Graphical displays and Tables

Topic	**1**	**When to use a graphical display and when to use a table**
Subtopic	**1.1**	Exploration and discovery *versus* presentation – comparing and contrasting the goals. Graphical displays *versus* Tables: What are they good for? Main purpose of a graph: to discover or show pattern • Discovering pattern – part of data exploration and of evaluating fits of models to data. • Showing pattern to others – part of reporting results. Main purpose of a table: to look up exact values of features/variables. Exhibit and have informal discussion of examples of good and bad graphical displays (to be dissected later in the Topic Area). What information does each graphical display appear to be purveying? And is anything about the graphical display making it hard to extract accurate information? Is an additional graphical display needed to explore or to communicate effectively?
	1.2	Infographics and infotables Purpose is primarily to draw attention to a story. Principles of graphic construction or tabular construction to facilitate accurate human decoding of information are seldom followed. Exhibit and have informal discussion of examples of infographics and infotables drawn from contemporary media. What appears to be the primary purpose of each?
Prior knowledge required		Topic Areas 1.2 & 1.3

1.5 Graphical displays and Tables

Topic	**2**	**What makes a graphical display good or bad?**
Subtopic	**2.1**	The graphical process: a chain from graphic creator to graphic interpreter: **Data** -> Analysis and interpretation -> Information extracted by analyst -> Information encoding in Graphic -> Decoding by user -> Decoded information as extracted by user

Topic	2	What makes a graphical display good or bad?
		-> **Action or decision enabled for user.** How does the information as-decoded-by-the-user compare with the information originally encoded in the graphic? The accuracy of the humanly-decoded information tends to be maximized when a number of basic principles of Visualization are employed.
	2.2	Visualization Principle 1. Use Position along a common scale. Compare with: Non-aligned lengths (examples – stacked bar charts; when feature/variable of interest is the difference between two curves; dot chart – features/variables ordered on size rather than alphabetically). Angles (pie-charts). Areas, volumes. Explain use in both context of reporting results and exploring data.
	2.3	Visualization Principle 2. Choose an appropriate Aspect Ratio. Explain via examples showing how the Aspect Ratio affects the information extracted. Examples: the sunspot data and other time series examples.
	2.4	Visualization Principle 3. Encoding features/variables Use of shapes to identify different features/variables being plotted and to minimize difficulties from overlapping symbols. Use of colors or shades. Explain use in both context of exploring data and reporting results. Use examples of scatterplots and multiple time series.
	2.5	Visualization Principle 4. Supply an informative caption Describe the display, rather than the data. What's been plotted, what's the point? J W Tukey: "A picture may be worth 1000 words but it may need 100 words to explain it." Explain use in both context of exploring data and reporting results.
	2.6	Visualization Principle 5. May need more than one graph so that: in exploration, different graphics may reveal something new to the analyst; and in presentation, different graphical displays may provide more accurate pictures of different aspects of pattern. Examples: using at least two categorical features/variables and a continuous feature/variable; time series where the series themselves are of interest and the difference between two curves is also of interest. Explain use in both context of reporting results and exploring data.
	2.7	Other factors that good software generally gets right by default. (These are not principles.) Use good and bad examples as a basis for discussion: Axis labels.

Topic	**2**	**What makes a graphical display good or bad?**
		Axis markings. Positioning and use of legends. Size of plotting symbols.
	2.8T	Pedagogical issues relating to comparing different types of graphical display Emphasize the importance of good software in simplifying good graphical construction. Use interactive graphics – sliders, linked graphics, brushing, scatterplot rotation, … – to explore data. Explore how to find groups in data, and how to identify anomalous characteristics.
Prior knowledge required	Topic Areas 1.2 & 1.3	

1.5 Graphical displays and Tables

Topic	**3**	**What sort of graphical display should I use?**
Subtopic	**3.1**	Plotting samples of numerical data to explore relationships: explore the differing purposes of Histograms. Density estimate plus jittered actual values to detect outliers. Boxplots. Vertically aligning graphical displays of different samples.
	3.2	Plotting a numerical and a categorical feature/variable to explore relationships or present results. Compare these new types of plots with those previously encountered: Dot charts (ordered – largest at the top, smallest at the bottom). Proportions – (colored) beads on a string; ordered, but may need more than one graphical display to show different patterns accurately.
	3.3	Plotting features/variables that change over time Different versions of time series plots – points, points with connector lines, vertical lines, … . Reiterate point made in Topic 2.6: may need to plot difference between two time series explicitly.
	3.4	Plotting two numerical features/variables to explore relationships Scatterplots. [Pairs plots/Scatterplot matrices] Plotting several pairs of features/variables. Adding a smooth curve to bring out a relationship.
	3.5T	Pedagogical issues relating to comparing different types of graphical displays: Experimentation in encoding the same information using different graphical types and the consequent impact on the accuracy of the decoded information.

Topic	3	What sort of graphical display should I use?
		Explore how more than one graphical display might be needed in any given situation, in order to reveal different patterns.
Prior knowledge required		Topic Areas 1.2 & 1.3

1.5 Graphical displays and Tables

Topic	4	Tables: their purpose, and how to create good tables
Subtopic	4.1	Data sets are often made available to an analyst in the form of tables. However, the role of tables in the initial exploration of data is relatively slight compared with graphical approaches. They may have a useful role to play in presenting results.
	4.2	Principles for making patterns in tabular information more accessible to people (as opposed to tables to be used for looking up details): Caption, column headings (including units *etc.*), scale of the table (fits on page?) and row order all contribute to the usefulness or otherwise of a table. Explore the following principles and their application to examples from news *etc.*: ordering of features/variables. rounding of numbers (significant digits). using separators in large tables. relative ease of comparing numbers in columns rather than in rows.
	4.3	When it is often better to use a table rather than a graph: (a) very small sets of numbers. (b) data sets with several cross-classifications. (c) data sets in which some data points have comments attached because they are unusual in some way. (d) situations in which users will want to do further work on the data (*e.g.* if the actual numerical values are of direct interest).
	4.4T	Pedagogical issues relating to using tables: Teaching approaches. Typical misconceptions and how to correct them. Examples of effective use of tables.
Prior knowledge required		Topic Areas 1.2 & 1.3

[*Back to Unit 1 contents page*]

1.6 The data-handling pipeline

Commentary

This Unit deals with two topics that are also found in traditional computing: data management, and programming.

However, the approach is deliberately very different from that taken in computing courses. We want to be sure that the overlap is limited, so students can study both Computer Science and Data Science if both are offered. Also, a course based on this framework must be both attractive to and accessible to those who believe they lack talent in computing topics. For data management, traditional Computer Science focuses on cases where the data are held in a centrally-managed database which protects data integrity. However Data Science applications often use more *ad hoc* approaches. So in this course, we generalize many of the concepts and show practices that achieve good outcomes without much support from the platform.

For programming, the typical objective in Computer Science courses is mastery, so that the student can write code from scratch given a task description. Here, however, the focus is on understanding the power of automating Data Science tasks, and the skills are more at the level of writing a few lines, or modifying existing code. The description is not particular to a programming language, but the choice of language used in any class needs to have several characteristics, including being supported in a very easy-to-use programming and debugging environment, and having lots of high-level libraries and powerful language features (*e.g.* simple loop-over-rows of a file). We don't want students to have to struggle with lots of syntax, and we want them to be able to look at the code behind the tools they have been using in point-and-click fashion. R with tidyverse could be a good choice of language, as could Python.

It's important that this be centered around experiences, not just relying on memorizing terminology and pronouncements from some authority.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	Provide students with both experience of and conceptual understanding of the data-handling pipeline from the original recording of data or harvesting from online sources, via files and storage, wrangling data into a tidy form (usually rectangular) suitable for analysis, performing analysis and recording the history of the investigation, to producing the material of written and oral presentations of the results.	As for students but also with enough technical competence with one toolset, so that they can guide students and assist them in their experiences; also at a deeper level of understanding and reflection.
Learning outcomes	Students will be able to: • Classify their activities while problem-solving with computational tools with	**Additional learning outcomes** Sufficient mastery of a tool so they can guide students who get stuck, and so they can learn

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	respect to the Data Science Learning Cycle. • Explain the major steps in data-science processing activities, and how computational tools can be used in each • Demonstrate correct use of common terms for data processing and management activities. • Discuss and describe important data management principles and how they apply in a given setting. • Identify when and why the automating of data processing activities is helpful or necessary. • Describe the importance of reproducibility in data science. • Take simple code and – explain what the code is doing. – predict the consequences of modifications to the code. – modify the code to achieve desired purposes. • Demonstrate the ability to appropriately apply basic data manipulation processes such as subsetting, filtering, or transformation of variables in order to clean and prepare data. • Identify potential problems and pitfalls that may result from a particular uncleaned data set. • Create presentations of findings and conclusions that combine (possibly customized) graphical displays and tables with other content using computational tools. • Critique presentations that combine graphical displays and tables with other content and their method of production.	more aspects from online resources. Able to have nuanced discussions of comparison between toolsets and tool features. Knowledge of pedagogical issues relating to these topics, so they can teach them effectively.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Parts of Data Science learning cycle addressed	*Directly:* - Getting the data. - Exploring/Analyzing the data. - Communicating conclusions.	

1.6 The data-handling pipeline

Topic	1	Introduction to tool support for the data-handling pipeline
Subtopic	**1.1**	Automating the Data Science process: Reasons for automation: dealing with large volume of data, reproducing another's research findings, collaborative research, dealing with new data (distinguish between added data, corrected data, entire new dataset in same format). Overview of a toolset and how it automates aspects of the life-cycle. The diversity of toolsets (spreadsheets –local or cloud-hosted, relational database management system, notebooks).
	1.2	Data management principles: Relationship between: Logical data schema, physical data format, data content Metadata and their uses (data provenance, ownership, constraints on structure/contents, data meaning, units and category codes). Access control and rights over data. Approaches to sharing data and/or processing across space and time (copying files, managed database, cloud-hosted) . Version management (for code as well as datasets): importance of preserving old versions, naming, tools.
	1.3	Case studies of real data-science projects that were done with various toolsets.
	1.4T	Characteristics and evaluation of some widespread tool-sets (at least covering spreadsheets and interactive notebooks).
	1.5T	Pedagogical issues relating to tool use/mastery Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Prior knowledge required		Topics 2 & 3 of Topic Area 1.1; Topic 1 of Topic Area 1.2; Other Topics from Topic Areas 1.2-1.4 dependent on the examples used

1.6 The data-handling pipeline

Topic	2	Getting and storing data
Subtopic	**2.1**	Data sources: Collecting in the field (observational *versus* sensor), Running surveys.

Topic	2	Getting and storing data
		Downloading. Scraping.
	2.2	Logical data formats: Tables. Hierarchical. Log entries. HTML. Brief introduction to Media types (audio, image, video); these are mainly covered in Unit 2.
	2.3	Physical file formats Text file *versus* binary file; ASCII *versus* Unicode. Existence of differences between environments (*e.g.* Unix *versus* Windows files). CSV. JSON. Compression. Proprietary formats (xlsx, database internal).
	2.4T	Case studies of good data sources.
	2.5T	Comparison and evaluation of storage approaches.
	2.6T	Pedagogical issues around data sources and storage: Teaching approaches, Typical misconceptions and how to correct them, Examples of good assessments,
Prior knowledge required		As for Topic 1

1.6 The data-handling pipeline

Topic	3	Tool support for exploring and analyzing data
Subtopic	**3.1**	Automate an analysis: Introduction to the programming language, Data types, Dealing with a rectangular collection of data (*e.g.* R data frame), Simple code calling standard functions for calculating summaries, regressions, creating graphical displays *etc.* Applying functions in sequence, Overview of further features found in the programming language (conditional branching, loops, writing own functions, using libraries for many tasks),
	3.2	Data cleaning;

		Danger of processing default value as valid; ways to detect possible default value. Detect out-of-range and how to handle it. Detect and handle missing values: remove rows, replace by estimate; possible consequences of such strategies. Detect and handle outliers. Detect and handle weird text, especially text derived from scraping or automated transformations.
	3.3	Data transformations: Filter a meaningful subset from a larger dataset. Random sample from a large dataset for exploration. Value transformations (logarithmic, truncating range) . Simple reshape between rectangular formats. Transform hierarchical into rectangular.
	3.4T	Learn to use more aspects of the programming language, including coding a function of one's own.
	3.5T	Pedagogical issues for coding: Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Prior knowledge required		Topic Areas 1.2 & 1.3 (perhaps also Topics from Topic Area 1.4 depending on the examples used here)

1.6 The data-handling pipeline

Topic	**4**	**Generating presentations of the data**
Subtopic	**4.1**	Principles of communication: Examples of good and poor presentations. Understand your purpose. Understand the target audience (their skills, background, goals).
	4.2	Customizing graphical displays and tables, experience with: Choice of headings, scale, legends, axis labels, *etc.* Changing colors, icons *etc*. Programming language commands to control presentation.
	4.3	Combining explanation with graphical displays/tables: Written documents. Slideshows. Web pages. Generating these from a single tool. Combining material from several tools.
	4.4T	Comparison and evaluation of tools that allow generating presentations.
	4.5T	Pedagogical issues relating to generating presentations:

Topic	4	Generating presentations of the data
		Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Prior knowledge required		Topic Areas 1.1 to 1.3 (perhaps also Topics from Topic Areas 1.4 & 1.5 depending on the examples used here)

[*Back to Unit 1 contents page*]

1.7 Avoiding being misled by data

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	To deepen understandings that will allow students to more effectively critique data and data-based claims, and introduce them to some good practices for obtaining reliable data. To motivate incorporation of uncertainties in estimation via margins of error or interval estimates.	As for students but at a deeper level of understanding and with greater technical mastery, so they can guide students and assist them in their experiences; also greater reflection on what they are doing.
Learning outcomes	Students will be able to: • Identify when patterns in data may be artefactual (caused by deficiencies in the data collection or accumulation process). • Identify situations in which selection bias may be an issue, including cases where selection biases result from filtered data streams. • Distinguish between data that come from observational studies, controlled experiments, surveys, data streams. • Explain the ideas of validity and reliability of measures applied to data. • List possible biases that may result from missing values and compare strategies for working with missing values. • While working with data, students will additionally be able to: • Distinguish between random variation and systematic bias. • Explain why random sampling overcomes issues of selection bias. • Explain the effect of large sample size on uncertainty and systematic bias. • Assess the role of random sampling when drawing conclusions from a data set.	**Additional learning outcomes**

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	• Explain the purpose of randomized assignment. • Plan a data collection scheme that utilizes random sampling. • Design and implement algorithms to randomly assign objects to treatment groups. • Identify situations in which a margin of error is or is not appropriate and calculate the margin of error when appropriate. • For the purposes of communication, analysis and evaluation, students will also be able to: • Identify measures used in published reports that do not measure what they purport to measure and explain why the measures are in error. • Interpret estimates that are provided with a margin of error. • Describe the meaning of statistical causality in simple situations. • Identify and explain potential confounding factors in observational studies. • Compare the effectiveness of data collected via observation and controlled experiments for addressing particular problems and/or questions, particularly those addressing causality.	
Parts of Data Science learning cycle addressed	Getting the data Exploring/Analyzing the data Drawing conclusions	

Commentary on the Learning Outcomes: This Topic Area is concerned with building

- an awareness of fundamentally important types of problem that can occur during data collection and lead to false conclusions;
- the habit-of-mind of always being on the look-out for these problems; and
- some beginning ideas about strategies for overcoming them.

It is not concerned with technical competency.

So when a learning outcome talks about students "being able to recognize" some particular type of data problem, it does not mean that students should be able to do this reliably. It means that they will often be able to do it when given simple accounts or scenarios appropriate to their backgrounds.

The ideas of confidence intervals and randomization tests are addressed in Topic Areas 2.8 and 2.9 respectively. Beginning versions are included here for curricula that will not include these more advanced Topic Areas, but also to avoid ending Unit 1 on the note of a catalogue of problems and no solutions.

1.7 Avoiding being misled by data

Topic	**1**	**GIGO – "Garbage In, Garbage Out" – as the First Law of Data Analysis**
Subtopic	**1.1**	What do we mean by GIGO?
	1.2	Examples of "garbage". How can we avoid collecting or using garbage?
	1.3T	Pedagogical issues relating to these topics.

1.7 Avoiding being misled by data

Topic	**2**	**Bias and what we can do about it**
Subtopic	**2.1**	Biases due to measurement issues: Important examples where measures that have been used that do not measure what they purport to measure, or there have been serious problems in classifying people/units into groups resulting in misleading conclusions. Basic ideas of the validity and reliability and the dangers of proxy measures; levels of measurement (nominal, ordinal, interval, ratio).
	2.2	Biases due to selection or filtering in data streams: Important examples where resulted in misleading conclusions. Random sampling as a strategy for reducing possible biases when collecting data.
	2.3	(*Discussion Topic*) Extrapolating from the data we have to a larger setting: when is that reasonable? (Issues include soundness of measures and "representativeness".)
	2.4T	Pedagogical issues relating to these topics.
Prior knowledge required		The data-critique discussions begun in Topic Area 1.1 and Topic Areas 1.2 & 1.3

1.7 Avoiding being misled by data

Topic	3	Problems and solutions in reaching causal conclusions
Subtopic	3.1	Examples that show how allowing for an important third feature/variable can change, and even reverse, the apparent relationship between an outcome and predictor feature/variable (Simpson's paradox as a special case). Idea of a confounder/lurking feature/variable. Implications for reaching causal conclusions from observational data.
	3.2	Key differences between an observational study and a randomized experiment. Randomized experiments as the most reliable data-collection strategy for investigating causation with the emphasis on a simple randomized experiment. Why experimentation is often not possible (ethical and practical reasons).
	3.3	(*Discussion Topic*) Extrapolating from the data at hand to a larger setting: when is that reasonable? (Issues include the role of observational data combined with other information in forming causal conclusions; and relating to extrapolation from experimental data on convenience samples, which almost all experiments use.)
	3.4T	Pedagogical issues relating to these topics.
Prior knowledge required		The data-critique discussions begun in Topic Area 1.1; Topic Areas 1.2 & 1.3; Topic 2 of Topic Area 1.4

1.7 Avoiding being misled by data

Topic	4	Questions that can and cannot be answered by data
Subtopic	4.1	Learning to ask questions that can be answered from data.
	4.2	Learning to spot questions that cannot be answered from the available data, or from data that can realistically be obtained.
	4.3T	Pedagogical issues relating to these topics
Prior knowledge required		Topics 1-3 of this Topic Area; Topic Areas 1.1-1.3

1.7 Avoiding being misled by data

Topic	5	Sampling errors and confidence intervals
Subtopic	5.1	Random sampling is not perfect: the problem of sampling error; how the extent of sampling error reduces as sample size increases; the idea of allowing a margin around an estimate to cater for the likely extent of sampling error.
	5.2	Unpacking "the likely extent of sampling error"; the concept of a confidence interval; what confidence intervals do and do not allow for; communicating a confidence interval for a single population parameter (*e.g.* mean, median, proportion).

Topic	**5**	**Sampling errors and confidence intervals**
	5.3	Experiencing how confidence intervals can be constructed by using either bootstrap resampling error to approximate sampling error, or the use of formulae (for means and proportions).
	5.4T	Pedagogical issues relating to these topics.
Prior knowledge required		Topic 2 of this Topic Area; Topic Areas 1.2 & 1.3.

1.7 Avoiding being misled by data

Topic	**6**	**Addressing the problem of randomization variation in experiments**
Subtopic	**6.1**	Randomized assignment is not perfect: how randomization alone can induce apparent group differences that are surprisingly large (simulation).
	6.2	(*Discussion*) When can we reasonably conclude observed group differences demonstrate real treatment effects?
	6.3	Experiencing a two-group randomization test as a mechanism for addressing this problem (*performed by simulation*).
	6.4T	Pedagogical issues relating to these topics.
Prior knowledge required		Topic 3 of this Topic Area; Topic Areas 1.2 & 1.3.

[*Back to Unit 1 contents page*]

Unit 2

1. Introduction to Unit 2's Topic Areas

Unit 1 provided a set of introductory topics that constitute the foundation of the curriculum framework. It is intended to require about 120 – 180 hours of study depending on the level of detail included. It aimed to give students a flying start, to develop their enthusiasm for the subject of Data Science and what it may mean for them in their future lives and stimulate learning about what they – personally – can do with data.

Unit 2 has more narrowly focused aims:

Aim 1. To introduce several different data types, and some common examples of the sorts of problems in which they can arise.

Aim 2. To introduce some different ways of analyzing data in order to draw inferences in relation to the problems posed.

Aim 3. To reinforce the different phases of the cycle of learning from data, as appropriate to the circumstance.

Unit 2 Topic AreasIt comprises a set of Topic Areas from which curriculum designers and teacher may wish to make a selection in designing a course:

Topic Areas 2.1, 2.3, 2.6 and 2.10 relate to different types of data commonly encountered in practice (Aim 1, Aim 3).

Topic Areas 2.4, 2.5, 2.6, 2.8 and 2.9 introduce different ways to draw inferences in relation to the problem being studied (Aim 2, Aim 3).

Whilst these Topic Areas are largely independent of each other, they all draw on the basic skills and knowledge acquired from studying Unit 1. **Topic Area 2.7** adds to these basic skills with powerful new techniques that can really add to the fun of exploring data.

Core Topic Areas

Many people would consider the Topic Areas on Machine Learning (2.4 and 2.5) and statistical inference (2.8 and 2.9) as core to data science thinking while the other Topic Areas are valuable and fascinating application areas.

Further Resources and more Detailed Descriptions for Unit 2

Most of the Topic Areas in Unit 2 do not have a history of being taught at the school level, and very little in the way of introductory modules is available at the tertiary level. Because Unit 2 covers less familiar territory, we have placed fuller expositions of the Topic Areas to follow on the IDSSP website at http://idssp.org/pages/framework.html, which contain more ideas about teaching and resources, and also some Case Studies.

Some optional **additional Unit 2 Topic Areas**, in particular a new one on **Network Data**, will also be added over time at http://idssp.org/pages/framework.html.

2. Summary of Aims for each Topic Area in Unit 2

2.1 Time series data

Aims: to develop basic understanding and skill in displaying, exploring, interpreting and presenting results for data that take the form of a time series.

2.2 Map data

Aims: plotting (positional or regional) geo-located data plotted on maps, use for exploratory analysis; and understanding maps themselves as graphical representations

2.3 Text data

Aims: to appreciate the many contexts in which text data can arise, to learn to explore such data and to extract and present potentially interesting characteristics in practical settings.

2.4 Machine Learning: Supervised

Aims: to develop an understanding of some of the contexts in which classification and prediction problems can arise, and to learn how to apply some basic tools for classification and prediction to draw conclusions in practical settings.

2.5 Machine Learning: Unsupervised

Aims: to develop an understanding of some of the contexts in it is of interest to find groups in data ("cluster analysis"), and to learn to apply some basic tools for this purpose and present the results in an informative fashion.

2.6 Recommender systems

Aims: to learn about some of situations in which Recommender systems are used, the sorts of data that are collected to develop these systems, and some methods for building such systems.

2.7 Interactive visualization

Aims: to learn how interactive visualization can be used to enhance various steps in the Learning from Data cycle, particularly relating to exploring data and communicating results, and to gain skills and experience in applying some of the basic tools.

2.8 Inference using Bootstrapping

Aims: An introduction to important concepts of confidence intervals in a random sampling context implemented using simulation methods (bootstrap resampling)

2.9 Inference using Randomization Tests

Aims: An introduction to important concepts of significance testing in the context of randomized experiments implemented using simulation methods (randomization/permutation tests)

2.10 Image data

Aims: to appreciate the many contexts in which image data can arise, to learn to explore such data and to extract and present potentially interesting characteristics in practical settings.

Details of Unit 2 Topic Areas

2.1 Time Series data

Commentary

- Data that depend on the time when they were measured constitute an important type that were not encountered in Unit 1. When the measurement, or recording, times are equally spaced (*e.g.* if they are consecutive dates), the data are referred to as contiguous time series. (Aside: irregularly spaced time series exist, but their study is quite an advanced topic so will be ignored for this Topic Area).
- Times series with strong seasonal characteristics are extremely common and thousands of interesting series can be found on the websites of government agencies and scientific organizations.
- Strongly seasonal time series provide an opportunity to study the data visually and to think in terms of models that are more complex than simple linear or curved regressions in a context where the contributing structural components are clearly visible to the eye after very little training. Trends and seasonal effects are often easy to interpret, and graphics can also reveal anomalies that correspond to discoverable historical events or structural changes, thereby widening the breadth of statistical thinking that can be drawn upon.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	Use the Learning from Data cycle to tackle problems in which the data take the form of a time series. To develop skills in preparing time-series data for analysis. To develop skills in uncovering and communicating characteristics often seen in time-series data. To develop skills in summarizing findings from the analysis and reporting these in appropriate ways to a client. To provide opportunities to apply what was learned in Unit 1. As the opportunity presents itself: • to heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse. • to pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	As for students.
Learning outcomes	Students will be able to:	**Additional learning outcomes**

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Explain what a time series is and provide examples of why people are often interested in how a feature/variable changes over time. - Explain why, in time-series plots, the data points are usually joined by lines. - Recognize and communicate characteristics such as spikes, sudden jumps, gaps, trends and seasonal effects in time-series data and seek explanations for them. - Describe and interpret a trend in a time series. - Explain and interpret the nature of additive or multiplicative seasonal effects in a time series. - Describe a seasonal time series in terms of a model of the form *feature = trend + seasonal oscillation + random noise*. - Interpret time series graphics and communicate their findings. - Compare related series by comparing characteristics such as their trends, their seasonal patterns and residual behavior, and interpret and communicate the results. - Explain, qualitatively, how time series forecasts are obtained and why the process may fail. - Interpret forecasts from time series and discuss their limitations.	*(Pedagogical)*
Key phrases	Forecasting; Holt-Winters forecast; Jumps; Prediction; Seasonal effects; Spikes; Seasonal decomposition (STL)	
Data Science Learning Cycle elements: *All steps*		

2.1 Time Series data

Topic	1	Problem elicitation and formulation: Time Series data
Subtopic	1.1	The nature of time series data.
	1.2	Reasons why people are often interested in time series data.
Prior knowledge required		Topic 2 of Topic Area 1.3

2.1 Time Series data

Topic	2	Getting the data
Subtopic	2.1	Obtain time-series datasets addressing some problems of interest arising from the discussion in Topic 1.
	2.2	Experience some common date-and-time feature/variable formats.
	2.3	Experience transforming and reshaping one or more data sets into a form that can be used by a specialist analysis program, using actions like transforming date-time features/variables, aggregating a response feature/variable (*e.g.* into hourly, daily, weekly, monthly or yearly totals or averages) and reshaping the data set (*e.g.* long form *versus* wide form).
Prior knowledge required		Topic Area 1.6

Topic	3	Exploring the data
Subtopic	3.1	Basic time-series plots; smoothing to reveal trend; recognizing characteristics such as spikes, sudden jumps, gaps in these plots; faceting time-series plots by year (or other natural time scale) to expose seasonal regularities; seasonal plots.
	3.2	Beginning to see the *trend + seasonal* oscillation components in plots of strongly-seasonal time series.
	3.3	Decomposition into *trend + season + residual* (*e.g.* STL decomposition); choosing between additive and multiplicative seasonal effects; interpreting additive and multiplicative seasonal effects; seeing aberrations from seasonal averages; communicating the results.
	3.4	Comparing related series by comparing characteristics such as their trends, seasonal patterns and residual behavior; communicating the results.
Prior knowledge required		Topic Areas 1.3, 1.4, 1.5

2.1 Time Series data

Topic	4	**Analyzing the data: Modelling and Forecasting**
Subtopic	**4.1**	Forecasting as projecting patterns from the past (*e.g. trend + seasonal*) into the future; what can go wrong with this.
	4.2	Making an informal forecast from a series using students' own intuitions.
	4.3	Experience with using a formal forecasting method (*e.g.* Holt-Winter) to produce both point and interval predictions; discussion of the most important assumptions made by the method used; communicating the results.
	4.4T	Pedagogical issues relating to these topics.
Prior knowledge required		Topic Area 1.3

2.1 Time Series data

Topic	5	**Communicating the Results; next question?** *(in the context of a particular investigation targeting a particular real-world problem.)*
Subtopic	**5.1**	Deciding on the characteristics of the time-series data set that most need to be communicated to the client.
	5.2	Deciding on the graphics and summaries that will best communicate these characteristics Where appropriate, modifying output automatically generated by software to make these characteristics more immediately obvious (*e.g.* with captions and annotations).
	5.3	Telling the story: putting together a logical and compelling argument that uses the plots and summaries chosen to underline the points being made. Presenting as a report or an illustrated talk.
Prior knowledge required		Topic Area 1.5

For a fuller exposition of this Topic Area, including teaching and resources ideas and a Case study, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.2 Map data

Commentary

Data that contain geographical location fields (latitude and longitude) or region fields (*e.g.* Country, State or County) provide an opportunity to display data on maps either to gain insights about geographical patterns in the data or to use the analyst's own domain-specific knowledge (*e.g.* background geographical knowledge, awareness of local socio-political issues, *etc*.) gain a better understanding of patterns in the data. The visual attractiveness of map displays can also be highly motivational.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	Use the Learning from Data cycle to tackle problems that use maps as a display framework in which to perform exploratory analysis. To develop skills in using maps, where appropriate, in summarizing and visualizing findings from analyses in order to provide effective communication to a client. To provide opportunities to apply what was learned in Unit 1 Topic Areas, especially those relating to visualization, in a new setting. As the opportunity presents itself: – to heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse. – to pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	Same topics as for students, but with two deeper prongs: deeper understanding of the relationships between features/variables, so they can more easily guide students into understanding complex spatial patterns and relationships. deeper understanding of the ethical concerns relating to geotagging and geographically located information, especially as it pertains to privacy, health, and anonymization. Also greater reflection on what they are doing, and why: no maps for the sake of maps, but chosen as the most effective tool for the job.
Learning outcomes	Students will be able to: • Describe how they have transformed some geographically tagged data into a form which can be visualized effectively • Discuss the distinction between location and regionally tagged data as it applies to displays on maps. • Display both location and regionally tagged data on maps. • Interpret and communicate patterns revealed by these displays.	**Additional learning outcomes** *(Pedagogical)* Understand the prevalence of maps, and their ubiquitous nature in visualization. Understand and convey to the students how bias, misconception and distortions are especially pernicious in mapping data, with its assumed easy translation into our "real world".

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
	• Describe basic structures of regional maps: basic polygon structures, boundaries, lines, fills, and how they are useful. • Discuss biases and distortions that are part of most maps ("the map is not the territory"). • Produce examples showing how maps can represent many features/variables simultaneously.	Convey to the students some of the more subtle factors in map creation, especially in the user choices (scale, projection such as Mercator, and similar factors).
Key phases:	Faceting, Geo-tagged data, Map, Neighborhood, Overlay, Projection, Tagged data, Visualization	
Data Science Learning Cycle elements: All		

2.2 Map data

Topic	**1**	**What are the purposes of maps?**
Subtopic	**1.1**	Maps are ubiquitous in our society, giving a scaled representation of the reality around us: As geographically tagged data become prevalent, and rich social, societal and environmental data are now available with these geographic features/variables, data scientists need to be familiar with the deep visualizations possible through careful design and analysis of maps. Introduction to maps as visualization tools; reminder of commonly understood archetypes from geography and history courses.
	1.2	Discussion of components of a map: edges, lines, points; the concept of geographic distance as map distance.
	1.3	Separation of data from display: what are our data? "The map is not the data!" (also known as "the map is not the territory").
	1.4	(*Discussion*) Multiple dimensions, tied back into Unit 1.4: maps as visualization tools have a minimum of 3 dimensions, and with interactivity and temporal structure, often 5 or more.
	1.5	Interactive example as a tool for exploratory learning: • Human political-division-level data (*e.g.* Gapminder for country-level, other similar sets for state- or province-level comparisons). • Brought from 2D representations as scatterplots to a regional map (localized; *e.g.* Australasia for those in the region, America for the US, Europe for those in Europe).

Topic	**1**	**What are the purposes of maps?**
		• Allow students to use an interactive tool to choose features/variables, with the map scope and resolution pre-set. • The focus is on introducing the challenges associated with selection of features/variables, and to heighten awareness of the biases and distortions that can occur due to our inherent bias. Discuss the fact that we often do not choose the scale, resolution or structure of the map, but are simply "coloring in" using features/variables. Bad visualizations hinder more than they help.
Prior knowledge required		Some familiarity with maps (*e.g.* Google maps), maps of countries, world maps including maps which show political boundaries. Visualization Topics from Unit 1 (Topic Areas 1.2-1.3, 1.5). Understanding of what features/variables are.
Possible resources (data, real-life problems, …)		Human political division shapefiles are easily and publicly available. Google Maps, or Google Earth. Any resources from geography, refitted to add feature/variable layers beyond the geographic (*e.g.* human, societal, or environmental).

2.2 Map data

Topic	**2**	**How do we build and work with location maps?**
Subtopic	**2.1**	Location and Region data as common archetypes.
	2.2	Plotting points on downloaded map tiles, relationship to scatterplots; effect of projections.
	2.3	Adding information at locations: coding feature/variable-information at location points; interpretation.
	2.4	Subsetting/Faceting as a tool: time and space; ways of showing changes over time.
	2.5	Interactivity with location map-plots.
Prior knowledge required		Topic Areas 1.3 & 1.4

2.2 Map data

Topic	**3**	**How do we build and work with regional maps?**
Subtopic	**3.1**	Shape files and the coloring of regional polygons (choropleth maps); region labels.
	3.2	Matching regions in a dataset to regions in a shape file: Matching names. (*Optional*) More complex matches: intersections, joins, subsetting, non-matching areas – fuzzy joins, possible biases from decisions.
	3.3	Representing two or more features/variables; issues of scales.
	3.4	Perceptual problems with choropleth maps; alternative representations.

Topic	**3**	**How do we build and work with regional maps?**
	3.5	Subsetting/Faceting as a tool; ways of showing changes over time.
	3.6	Interactivity with regional maps.
	3.7T	Subtleties of color and scale choice – communication enhancement.
	3.8T	Distortions and bias – avoiding misleading figures – projections.
Prior knowledge required		Topic Areas 1.3 & 1.4

2.2 Map data

Topic	**4**	**(TEACHER-only TOPIC) What is a Map, and how is this Data?**
Subtopic	**4.1T**	Discussion of the counterpoints between the two viewpoints: 1) maps are visualizations of data only, and 2) maps can be considered as data themselves.
	4.2T	Consider a map as a data set, not just a point or a series of points plotted in 2 dimensions. Each map consists of multiple features/variables: a minimum of two spatial dimensions, and usually at least one (but often more) categorical or numerical dimensions. Identifying components on a map: polygons representing human or natural boundaries, water-land boundaries, and human features such as roads, bridges and tunnels. Maps as a *representation* of reality.
	4.3T	Finding patterns in data through maps. We often use maps because the patterns we are looking at are *spatial* in nature: that is, they change with geography. For example, average income by some small geographical area (*e.g.* a neighborhood) varies wildly in a city by the “class” of a neighborhood. Sometimes there are exceptions: can these be identified via visualizations? Interesting case study: use of satellite imagery compared to building plans and maps to find tax cheats in Greece. https://www.nytimes.com/2010/05/02/world/europe/02evasion.html?th&emc=th
	4.4T	Overlays on maps. A map can be very simple: just polygons representing boundaries. Or it can be highly complex, with a number of layers built on top of the simple polygon layer. Consider how adding information may make a map more useful rather than less. Symbol and color choice. Subjective “design” decisions which have strong implications for value of a map as a tool for understanding and inference.
Prior knowledge required		Some familiarity with maps (*e.g.* Google maps), maps of countries, world maps including maps which show political boundaries.

For a fuller exposition of this Topic Area, including teaching and resources ideas and a Case study, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.3 Text data

Commentary

Text analysis can take natural language text (*e.g.* the contents of books, articles, social media posts, and free-response items in questionnaires) and process it in ways that can uncover important elements related to what the writers are talking about, how they feel about the subject, how they are using language, and even to identify textual elements that are useful for inclusion as predictive-features/variables in predictive algorithms/models. Many of the types of table and graphical displays used in practice have already been encountered in Unit 1. Text analysis provides an opportunity to use these tools productively in a new and unexpected setting and introduce the use of others, particularly word clouds, that many students will already have seen in webpages and articles.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	To introduce students to the power of text analysis for gaining insights from natural language texts and gain experience in doing this Apply the *Learning from Data* cycle (LDC) to think about problems addressed from the analysis of natural language / text data Illustrate how to handle data that do not come from direct measurements of a characteristic but are extracted from text Show how to pre-process the extracted features back to numeric quantities, thus providing opportunities to apply what was learned in Unit 1 (particularly descriptive statistics and visualization) Introduce the objectives of sentiment analysis and the notion of subjectivity via sentiment analysis As the opportunity presents itself: – to heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse – to pay specific attention to issues relating to data quality, questioning skills, and presentation skills	Important to review basic statistical and visualization ideas Need to review components of language (*e.g.* nouns, verbs, adjectives and more) Important to review data structure concepts such as data frames (rows = observations, columns = features/variables) Review basic R or Python including packages for reading data and processing data sets First time feature extraction *versus* reading features directly might encountered. Need to clarify for students.
Learning outcomes Topic Area	Students will be able to:	**Additional learning outcomes** *(Pedagogical)*

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
1.3	• Describe measurement levels of features/variables and why text data are nominal scale. • Discuss the motivations for looking at displays of frequencies of tokens. • Break strings of text in tokens. • Remove observations containing stop words from a data frame of tokens. • Produce a frequency table of tokens and visualizations of these frequencies, and interpret and communicate the results. • Break strings into bigrams. • Discuss motivations for the use of sentiment analysis. • Merge a data frame of text tokens with a lexicon of sentiment values. • Produce descriptive summaries and graphical displays of the sentiment in a data set of text, and interpret and communicate the results. • Discuss some limitations of text analysis and dangers of misuse.	
Key phrases:	Bigrams, Frequency tables, Ngrams, Sentiment, Stop words, Tokens, Word clouds	
Data Science Learning Cycle elements	*Problem elicitation and formulation* – analyzing word use and sentiment from text *Getting Data* – reading natural language data, converting to data for analysis by tokenizing and extracting features *Exploring data* – frequency tables, word clouds, sentiment distributions and trends *Analyzing data* – comparing frequency tables, sentiment distribution between different sources of natural language *Communicating the results* – of their text analyses	

2.3 Text data

Topic	1	**Problem elicitation and formulation: Text data**
Subtopic	**1.1**	Examples of questions to be addressed using natural language. Given a general overview of the various kinds of analysis one can do with text, including information retrieval, clustering, document classification and web mining. The focus here is on information extraction, *i.e.*, to discover what the authors of the documents talk about, what they like, *etc.* For example, describe text data in use for product reviews: Encourage a class to define a product of interest, and download reviews on that product. Furthermore, encourage them to get reviews of different brands of the same product to facilitate comparisons.
	1.2	Important characteristics of text data.
	1.3	The objectives of trying to understand the content of the text by first extracting tokens.
	1.4	The need for removing stop words and perform stemming and dangers in doing so blindly.
	1.5T	Characteristics of text data – even though focus would be on relatively clean text for expository purposes, need to warn students about challenges associated with misspelling, short forms / abbreviations, tense, plural/singular and more.
Prior knowledge required		Topic Areas 1.2, 1.6.

2.3 Text data

Topic	2	**Bag of words analysis of text data**
Subtopic	**2.1**	Constructing frequency tables of tokens. Understanding the importance of removing stop words and performing stemming.
	2.2	Generating bar charts and word clouds of token frequencies.
	2.3	The limitations of unigrams, *i.e.*, single-word tokens (*cf.* “world congress”). Extracting bigrams.
	2.4	Summarizing bigrams and compare the differences between unigrams and bigrams.
	2.5	Exploring differences between documents (or sections of the same document) by comparing token frequencies.
	2.6	Word use that tends to distinguish the content of documents, intent and use of tf-idf statistics in comparisons.
	2.7	(*Discussion*) Limitations of bag of words analysis and dangers of misuse (cautionary tales).
	2.8T	Pedagogical issues relating to text data analysis:

Topic	2	Bag of words analysis of text data
		Limitations of analyzing single words extracted from text – loss of context in bag-of-words analyses; loss of linguistic subtlety – Need to have a basic review of linguistic structure Limitations of dealing with characteristics of text Caveats for inference – Is the text a representative sample from some population? – Can we use text analysis to infer anything about a population? Volunteer responses reflect bias so if scraping comments from a site, then these may be a biased view of the population of people.
Prior knowledge required		Topic Areas 1.2, 1.6

2.3 Text data

Topic	3	Sentiment Analysis
Subtopic	**3.1**	Concept: Why do people want to do sentiment analysis? What is sentiment? Focus on adjectives in natural language. Differences between objective measurements and subjective opinions; dichotomizing and degrees (ordination) of sentiment; issue of negated sentiments (*e.g.* “good” *versus* “not good”)
	3.2	Merging tokens with sentiment data tables
	3.3	Summarizing sentiment within a document corpus Exploring the differences in sentiment between two document sources or corpora
	3.4	Discussion of limitations of sentiment analysis and dangers of misuse (cautionary tales)
	3.4T	Limitations of attaching sentiment to single words. Subjectivity in an analysis may be a difficult idea for students who think that *math* classes will only have black and white answers. Need to reinforce that these data differ in fundamental ways from familiar physical measurements such as temperature and weight. Need to explore robustness of results to different analysis strategies (*e.g.* different sentiment lexicons; bigrams *versus* single word analyses). May link part-of-speech tagging as key operations of sentiment analysis (*e.g.* noun phrases as features; adjectives as polarities of sentiments; adverbs as strengths of sentiments).
Prior knowledge required		Topic Areas 1.2, 1.3, 1.6

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.4 Supervised Learning

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	Understand what sort of problems can be solved with classification. Understand how algorithms/models are evaluated to measure performance. Understand how Classification and Regression Trees (CART) are used to classify. Fit a tree, interpret and evaluate the performance. Understand what overfitting is. Understand how "set aside" data are used and why they are important. Understand that trees (besides providing a decision rule for classification or a rough approach to regression) may provide insight into the structure of the data (detect interaction, rank the importance of co-variates *etc.*) As the opportunity presents itself: – to heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse – to pay specific attention to issues relating to data quality, questioning skills, and presentation skills	As for students
Learning outcomes Topic Area 1.3	Students will be able to: • Pose classification questions and identify situations that call for classification. • Provide an algorithm to classify categorical outcomes. • Use software to calculate misclassification rates. • Compare classification algorithms/models and decide which is the better for a given situation based on total misclassification rate. • Fit a CART and interpret results.	**Additional learning outcomes** *(Pedagogical)*

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Use a set-aside data set to compare classification algorithms/models. - Use software to calculate the complexity of a CART. - Create a validation data set. - Explain how algorithms/models can be used to predict numerical outcomes, and explain how a goodness of fit measure can be used to quantify the success of the prediction. - Explain the costs and benefits of misclassification in a given context. - Describe an algorithm to generate a tree to predict numerical outcomes using 1 or 2 features/variables. - Fit and interpret a regression tree using software. - Use a validation dataset to compare trees for prediction.	
Key phrases:	Analysis of Variance, CART, Classification, Data-Supervised Learning, Goodness of Fit, Leaf, Mean Absolute Deviation, Machine learning, Misclassification, Node, Overfitting, Prediction, Standard Deviation, Validation, Variance and Complexity	
Data Science Learning Cycle elements: All steps		

2.4. Supervised Learning

Topic	1	Problem elicitation and formulation: Supervised Classification
Subtopic	**1.1**	What is classification? What sort of data are we considering?
	1.2	A classification algorithm/model contains rules that determine how an object is classified.
	1.3	Classification will rarely be perfect; we need a way to measure how well the classification algorithm/model works.
Prior knowledge required	Topic Areas 1.2, 1.3, 1.6	

2.4. Supervised Learning

Topic	**2**	**Introduction to Classification Trees**
Subtopic	**2.1**	What are the components of a classification tree and how does it work?
	2.2	What is a misclassification rate?
	2.3	What is node/leave "purity"?
	2.4	What is an algorithm that is used to generate Classification and Regression Trees?
	2.5	What are consequences of misclassification?
Prior knowledge required		Topic 1 above.

2.4. Supervised Learning

Topic	**3**	**Growing Classification Trees**
Subtopic	**3.1**	Introduction to R or Python commands (or from another appropriate language) to grow and visualize a CART.
	3.2	When to stop: why not grow trees until each observation is in its own leaf?
	3.3	What is overfitting? Using validation data set.
	3.4	Pruning
Prior knowledge required		Topic Areas 1.3, 1.4, 1.5

2.4. Supervised Learning

Topic	**4**	**Communicating the Results; next question?**
Subtopic	**4.1**	What does the tree tell you about how classifications are made?
	4.2	Often, trees such as these are used to make decisions (for example, to send a patient to intensive care or to another hospital ward). In that sense, the tree itself is part of the product that needs to be delivered. What other information is important to communicate in order for people to make proper use of the tree?
	4.3T	Pedagogical issues relating to these topics
Prior knowledge required		Above topics, Topic Area 1.5

2.4. Supervised Learning

Topic	**5**	**Introduction to Regression Trees** *(in the context of a particular investigation targeting a particular real-world problem)*
Subtopic	**5.1**	What does "prediction" mean in the context of numerical outcome features/variables? Measuring quality of prediction.

Topic	**5**	**Introduction to Regression Trees** *(in the context of a particular investigation targeting a particular real-world problem)*
	5.2	Interpreting regression trees. (Students see examples of trees.)
	5.3T	Building regression trees with one predictor feature/variable.
	5.4	Building regression trees with more than one predictor feature/variable.
	5.5	Comparing trees with a validation set.
Prior knowledge required		Topic Areas 1.2, 1.3. Topics 1 – 4 above.

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.5 Unsupervised Learning

Commentary

Unsupervised learning, known as *Cluster Analysis* or *Clustering* in Statistics, has the objective of grouping a set of objects (based on the data we have on them) in such a way that objects in the same group/cluster are more similar to one another in some sense than they are to members of other groups/clusters. There is no training set of data for which group labels or the values of a response feature/variable are known (as is the case in *Supervised learning/Classification*). The objective is to *discover* groupings in unlabeled data.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	Understand what sort of problems can be solved with unsupervised learning. Understand that unsupervised learning relies solely on the feature (predictor) features/variables, and does not make use of a response feature/variable. Understand an algorithm commonly used in unsupervised learning (*i.e.* K-means clustering). Understand how to interpret and communicate the results of a clustering algorithm output. As the opportunity presents itself: – to heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse – to pay specific attention to issues relating to data quality, questioning skills, and presentation skills	As for students but at a deeper level of understanding and with greater technical mastery, so they can guide students and assist them in their experiences; also greater reflection on what they are doing.
Learning outcomes	Students will be able to: • Explain the concept of inputs (input features/attributes/variables) and outputs (class labels) (review from Topic Area 2.4). • Explain the concept of inputs (input features/attributes/variables) and outputs (class labels) (review from Topic Area 2.4).	**Additional learning outcomes** *(Pedagogical)*

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Distinguish between data that are appropriate for supervised *versus* unsupervised learning based on its structure, particularly the presence and roles of inputs and outputs. - Describe a real-world example of clustering, *e.g.* Segmentation of customers based on characteristics so that customized marketing emails can be sent to them. - Explain the purpose of unsupervised learning, identify situations for which unsupervised learning is useful and pose questions for which unsupervised learning can provide insights. - Demonstrate the role of distance in clustering including: – Explain the concept of distance (Euclidean) between data points. – Illustrate the concept of distance between points using a 2D plot containing 2 features. – Show how data points closer together are similar and how points further apart are dissimilar. – Describe how the concept of clustering is based on distance as a metric. - Explain the iterative process of one clustering algorithm such as K-means. - Describe how clustering can be used to detect outliers. - Identify the motivation for outlier detection (*e.g*. credit card fraud). - Use software to prepare data, apply K-means clustering and interpret the results	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Examine how results can differ with choice of K and propose how an optimal choice could be made. - Give examples of how labeled data might be difficult or too expensive (time, cost, resources) to obtain, and describe how unsupervised learning might be readily accessible for using in these situations. - Critique the interpretation of clustering results, considering the possible influence of human bias. - Visually assess data for suitability of K-means *versus* other clustering methods - Interpret visualizations that demonstrate unsupervised learning algorithms that are not distance-based but rather rely on another attribute such as density (*e.g.* DBSCAN).	
Key phrases:	Clustering, Distance (Euclidean), outlier detection, Segmentation, Unsupervised learning	
Data Science Learning Cycle elements: All steps		
Prior knowledge required	Topic Area 1.3	

2.5 Unsupervised Learning

Topic	1	Problem elicitation and formulation: Unsupervised Learning
Subtopic	**1.1**	What is unsupervised learning? What is supervised learning? And how do they differ? What data should we consider for unsupervised learning?
	1.2	Unsupervised learning aims to create clusters (groups) of data points based on attributes of the data
	1.3	In a real-world application, there are no class labels available for unsupervised learning (unlike classification or supervised learning), hence the name.
Prior knowledge required		Topic Areas 1.2-1.3, 1.6

2.5 Unsupervised Learning

Topic	**2**	**Getting and exploring the data**
Subtopic	**2.1**	Obtain datasets addressing the problem of interest arising from the discussion in Topic 1.
	2.2	The orientation of the data: the rows represent the observations while the columns represent the features/attributes/variables.
	2.3	The class label is missing in the presented data, compared to Unit 2.4.
	2.4	Motivate how mathematical techniques can automate the discovery of clusters in datasets.
Prior knowledge required		Topic 1 above

2.5 Unsupervised Learning

Topic	**3**	**Example of Unsupervised learning algorithm: K-means clustering**
Subtopic	**3.1**	K-means is a clustering algorithm which is iterative in nature. The number of clusters is an input into the algorithm, and typically is provided by intuition or independent domain knowledge of the data.
	3.2	The goal is to assign each data point to a cluster automatically, so as to satisfy the distance metric. It involves making an initial tentative choice for the center of each cluster (randomly or arbitrarily), and iteratively finding the distance from each point to all cluster centers. The cluster center to which the point is closest is the assigned cluster.
	3.3	The iterative nature of the algorithm: after all the points have been assigned a cluster, determine the updated cluster centers, and repeat Step 3.2.
	3.4	The need to repeat with different initial guesses for cluster centers.
	3.5	A simple example containing a small set of data points to explain the K-means algorithm manually.
	3.6	A (distance-based) outlier is an object that is “surprisingly”, or noticeably, far from its nearest neighbors.
	3.7	A (distance-based) outlier is also an object that will typically be far from the middle of its assigned cluster (cluster center).
Prior knowledge required		Topic 2 above.

2.5 Unsupervised Learning

Topic	**4**	**Implementing K-means clustering on a large data set**
Subtopic	**4.1**	Introduction to the format of the data set.
	4.2	Introduction to the programming environment to use for clustering.

Topic	4	**Implementing K-means clustering on a large data set**
	4.3	The need to clean and transform the data set so that it is arranged as observations (rows) *versus* features/variables (columns).
	4.4	Run the algorithm using the code snippet provided for particular choice of K.
	4.5	Interpretation of the results. Apply some intuitive domain knowledge to see if the cluster choices can be explained. This in turn motivates using a data set that the students can relate to.
	4.6	How the cluster selections change when algorithm is re-run with different values of K.
	4.7T	Pedagogical issues relating to these topics
Prior knowledge required		Topic 3 above

2.5 Unsupervised Learning

Topic	5	**Use in Problem solving** *(in the context of a particular investigation targeting a particular real-world problem)*
Subtopic	**5.1**	Why unsupervised learning is the best approach for the problem at hand, linked to the fact that labelled data require human effort, and is hard to obtain. Data from measurements and sensors typically lend themselves readily to unsupervised learning. Examples of business problems where unsupervised learning is commonly used.
	5.2	Selection of appropriate exploratory analysis and graphical displays to summarize the input data.
	5.3	Visualizations to show the effect of different values of K, and how the choices were made.
	5.4	How an optimal choice of K could be made, and emphasize that result. Compute the “elbow distance metric” which represents the mean distance between cluster points and centroid.
	5.5	Use of descriptive statistics about the features/variables in each cluster to demonstrate how they are similar within a cluster compared to across clusters and describe how clusters differ from each other.
	5.6	Interpreting and communicating the results of clustering-algorithm output for several different problems.
	5.7	Human factors: how human bias can influence how the clustering results are interpreted. Since clustering is unsupervised, the clusters formed are subject to human interpretation on what the distinguishing characteristics of each cluster are or how the clusters are explained.

Topic	**5**	**Use in Problem solving** *(in the context of a particular investigation targeting a particular real-world problem)*
Prior knowledge required		Previous Topics above.

2.5 Unsupervised Learning

Topic	**6**	**Other unsupervised learning methods – Alternatives to K-means clustering**
Subtopic	**6.1**	Instances when distance-based (that is, K-means clustering) may not be well suited.
	6.2	Motivation of need for other clustering methods (emphasize disadvantages of K-means) – no need to pre-determine number of clusters, lack of consistency.
	6.3	A few visualizations of different cluster shapes that do not lend themselves to K-means. For example, DBSCAN clustering performs better in these instances.
	6.4	Refer to visualizations as shown here for examples: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.
Prior knowledge required		Previous Topics above.
Ideas about teaching		Students should feel comfortable appreciating that different unsupervised learning methods exist. Whilst K-means is the simplest and most common, different choices may need to be made depending on the spatial distribution of the data.

For a fuller exposition of this Topic Area, including teaching and resources ideas and a Case Study, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.6 Recommender Systems

Commentary

Recommender systems try to identify items a user would like, based on data the system has about many users and many items. They play an important role in different types of community, by helping users reach appropriate choices without hunting through a huge range of possibilities, most of which are not interesting for that user. In e-commerce systems they can suggest items for a customer to look at buying, while in entertainment systems, they propose songs or movies someone might expect to enjoy, and in news or social media, they suggest items that are likely to be read.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	Use the *Learning from Data* cycle to think about problems in which recommendations are desired. As appropriate to heighten awareness of how ethical issues can arise in processing recommendations. Learn a variety of approaches to producing recommendations. Provide opportunities to apply what was learned in Unit 1. As the opportunity presents itself: • To heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse. • To pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	As for students but at a deeper level of understanding and with greater technical mastery, so they can guide students and assist them in their experiences. Provide a greater reflection on what they are doing.
Learning outcomes	Students will be able to: • Express what a Recommender system is and provide examples of the use of such systems. • Explain that recommendations are based on data, and provide examples of sources for that data. • Discuss types of ethical issues associated to Recommender systems and provide examples. • Describe several computational approaches to generate recommendations, and explain their relative strengths and limitations. • Operate with measures of similarity. • Compute recommendations in several ways using computational tools.	**Additional learning outcomes** *(Pedagogical)* Deeper mastery of tools that compute similarity, predict ratings, *etc.*, to support students as they learn to use tools, or to choose appropriate tools for use in class settings . Pedagogical topics specific to recommender systems.
Key phrases:	Collaborative filtering, Content-based recommendation, Personalization, Rating	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Data Science Learning Cycle elements: All steps		
Prior knowledge required: Topic Area 1.6		

2.6 Recommender Systems

Topic	**1**	**Problem elicitation and formulation, and communication: Recommender systems**
Subtopic	**1.1**	Examples of some Recommender systems in use (Entertainment *e.g.* movies, services *e.g.* hotels, matching *e.g.* jobs).
	1.2	Desirable characteristics of Recommender systems (personalization, effectiveness, computational efficiency, encouragement of exploration, avoid information overload, cold-start (i.e., works even without much personal data]). Consider non-personalized “best-seller” list as a baseline to compare with.
	1.3	Ethical issues for Recommender and personalized systems: sponsored recommendations, biases in recommendations (*e.g.* racial differences in recommended services), emotional impact of recommendations (*e.g.* baby products after stillbirths), danger of filter bubbles creating echo chambers, spreading fake news, conspiracy theories and loss of social cohesion, explanations *versus* data privacy, reinforcement effects from recommendations *versus* self-defeating recommendations *e.g.* Instagram crowds at beauty spots).
	1.4	Additional complexities (ranked recommendations, top-k recommendations, mutual satisfaction in matching).
	1.5	Communicating the recommendations, especially offering explanations of why they have been made.
	1.6T	Characteristics of a wide variety of Recommender systems in many domains.
	1.7T	Pedagogical issues relating to Recommender systems. Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Possible resources		Amazon, Netflix, online news services, yelp, ...

2.6 Recommender Systems

Topic	**2**	**The data used by Recommender systems**
Subtopic	**2.1**	Ratings (on user-item pairs): sparsity of the data. Example from https://www.kaggle.com/rounakbanik/movie-recommender-systems. Students gather own data set among the class for some domain such as holiday destinations or music.

Topic	2	The data used by Recommender systems
	2.2	Data quality issues: anchoring in ratings; proxies for ratings (*e.g.* page view, click-through, queries) and limitations of the proxies; presence or not of negative ratings.
	2.3	Feature data on items, demographic data on users. Students gather corresponding data for their own data set.
	2.4	Ethics with data: examples of deceptive ratings data (payola scandals, commercial incentives; honesty in self-reporting).
	2.5	Storing the data: exploiting sparsity property to reduce space needed; use of index structures to allow efficient access to ratings for a given user, or those for a given item.
	2.6T	Tools to collect and manipulate ratings data.
	2.7T	Pedagogical issues relating to data for Recommender systems. Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Prior knowledge required		Topic Area 1.6

2.6 Recommender Systems

Topic	3	Content-based recommendation
Subtopic	3.1	Concept: Recommendation based on single-user data, from item similarity "recommend items which are similar to those this user already likes".
	3.2	Measures of item similarity (number of shared features, cosine measure on feature vectors, different weights for features, *etc.*).
	3.3	Analysis and recommendation based on calculating nearest neighbors for an item. Do it by calculation on small data set, and check reasonableness of the results.
	3.4	Analysis and recommendation based on forming clusters of items. Experience doing this by intuition on small data sets.
	3.5T	Tools for calculating similarity, clusters *etc.*
Prior knowledge required		Topic Area 1.6

2.6 Recommender Systems

Topic	4	Relationships between categorical features/variables
Subtopic	4.1	Concept: recommend items that are liked by similar users. Examples from Amazon "people who looked at this book also bought that one".

Topic	4	**Relationships between categorical features/variables**
	4.2	Define similarity of users: similar demographic features, *versus* “like similar or same items”. Do it by calculation on small data set, and check reasonableness of the results.
	4.3	Analysis and recommendation based on using regression to predict unseen rating from known ones. Do it by tool on substantial data set, and check reasonableness of results.
	4.4	Ethics issues: impact of sensitive demographic features, impact of features that correlate with sensitive ones, stereotype-reinforcement.
	4.5T	Variety of tools that calculate predicted ratings.
	4.6T	Pedagogical issues relating to collaborative filtering. Teaching approaches. Typical misconceptions and how to correct them. Examples of good assessments.
Prior knowledge required		Topic Areas 1.4, 1.6

2.6 Recommender Systems

Topic	5	**Evaluation of a Recommender system**
Subtopic	**5.1**	User satisfaction; proxies to measure this (*e.g.* click-through, purchase)
	5.2	Measures from information retrieval, based on gold-standard data: precision and recall (and how they often trade-off); F1
	5.3	Ethics rules that apply to user studies
	5.4T	Tools to calculate IR measures
	5.5T	Pedagogical issues relating to recommendation evaluation Teaching approaches Typical misconceptions and how to correct them Examples of good assessments
Prior knowledge required		Previous Subtopics for this Topic
Possible resources *(data, real-life problems, …)*		(For teachers) Information retrieval textbook *e.g.* Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008), *Introduction to Information Retrieval. Cambridge*: Cambridge University Press.

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.7 Interactive Visualization

Commentary

This Topic Area seeks to introduce students to the power of interactive visualization in data exploration and communication. The Topic Area is grounded in evidence from perceptual and cognitive science about the ways in which people perceive visually and the capacities and limitations of visual cognition. The Topic Area separates visualization from interaction, then demonstrates their power in combination using several examples of advanced visualization types. Students will learn how to use existing interactive visualizations to explore data, and will practice critiquing the designs of visualization and identifying common design mistakes that may introduce misinterpretation and bias.

Course	**Introductory Data Science (IDS)**	**Teaching the Teachers (T3)**
Aims & Purposes	To situate the role of interactive visualization in the Learning from Data Cycle. To ground interactive visualization in fundamental concepts of visual perception and cognition. To describe the role of interactive visualization in data wrangling and the Data Handling Pipeline. To enumerate the main ways interactive visualizations can be used in the process of analysis and also communicating about data. To develop critical skills for viewing and interpreting visualizations. To augment chart-creation skills from Unit 1 with basic interactive capabilities: • Select. • Filter. • Details-on-demand. • Progressive disclosure. • Brushing-and-linking. • Level of detail. As the opportunity presents itself: • To heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse.	As for students but at a deeper level of understanding and with greater technical mastery, so they can guide students and assist them in their experiences; also greater reflection on what they are doing. Teachers should be able to clearly differentiate static information graphics from interactive visualizations and describe the specific benefits of interaction.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- To pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	
Learning outcomes	Students will be able to: - Explain how interactive visualization can be used in data exploration and communication. - Describe how visual and cognitive capacities enable and limit the benefits of interactive visualization. - Recognize common visualization types. - Enumerate common interactions and define what each is useful for. - Choose appropriate visualization types and interaction capabilities for a given dataset. - Customize the visual encoding through colors, labels, and other elements which enhance the effectiveness. - Demonstrate effective interactive exploration with existing visualization tools. - Critique existing visualizations for positive and negative design elements.	**Additional learning outcomes** *(Pedagogical)* Able to create interactive visualizations using existing software platforms or custom coding.
Key phrases:	Brushing, Data types, Interactive, Exploratory data analysis, Filtering, Gestalt psychology, Perceptual bias, Pre-attentive attention, Select, Visual encoding, Visual features/variables	
Data Science Learning Cycle elements: All steps		
Prior knowledge required	Topic Area 1.5	

Teaching commentary

While visualization has been introduced in prior Topic Areas generally (Topic Areas 1.2-1.4, 1.5) and for specific data types (Topic Areas 2.1-2.5, 2.8-2.10), this Topic Area introduces visualization theory more generally. It would be helpful for explanatory examples to reflect on the visualization types introduced in previous Topic Areas. For example, interactions could be demonstrated on line charts, text visualizations, and maps.

2.7 Interactive Visualization

Topic	**1**	**Why visualization?** **The role of Visualization in the Data Science Learning cycle**
Subtopic	**1.1**	The power of visualization: • Demonstrate the complementary nature of visualization compared with. statistical analysis (*e.g.* Anscombe Quartet). Stress that visualization is not a replacement for statistical analysis. • Exemplify the roles of visualization: – To get a first impression of the underlying data properties (distributions, "texture" of data, data cleaning: spot anomalies). – To support hypothesis generation and exploration. – To communicate a message, to persuade. • Other uses of visualization: – Personal reflection. – Ambient visualizations.
	1.2	Visual Exploratory Data Analysis: explain the fundamentals of EDA (searching for clues, as opposed to statistically confirm a fact). Demonstrate the role of interaction in this process through concrete examples: • Filter, select. • Slicing and faceting. • Different data projections and dashboards. • Brushing and Linking.
	1.3	Visual Communication and Presentation: • Discuss the abundance of visuals in communication (data journalism, scientific communications, *etc*.): "a picture is worth a thousand words". • Discuss the pros and cons of graphical presentation of data.
	1.4	Considerations and challenges of visualization design: • Visualization design is not a trivial exercise. • Introduce visual features/variables. • Briefly discuss the fact that there are not an infinite set of visual features/variables one can use: illustrate with examples use of many visual data features/variables resulting in an ineffective visualization. • Introduce the idea that some visual features/variables are perceptually more effective than others (develop this in Topic 2 below). It is easy to make mistakes. • An ill-designed visualization can be misleading:

Topic	1	Why visualization? The role of Visualization in the Data Science Learning cycle
		– Show some examples of poorly designed visualization (*e.g.* pie charts, poor choice of color palette, *etc*.).
Prior knowledge required		Unit 1, especially Topic Area 1.5 (Graphics and Tables)

2.7 Interactive Visualization

Topic	2	Data Types and Visual Variables
Subtopic	**2.1**	Brief introduction to perceptual and cognitive capacity: Visual bandwidth, memory, clutter, visual channels/variables, pre-attentive attention.
	2.2	Hierarchy of visual variables, *i.e.* how to choose an appropriate visual variable according to data type.
	2.3	Color in more in depth: • Different types of palettes (sequential, diverging, qualitative). • Emotional response to color: cultural differences. • Discuss briefly color-blind population.
	2.4	Motion: • Power of motion to draw attention (notification): dates back to prehistory: danger of predators. • Emotional response to motion. • Animation to convey changes in a view.
Prior knowledge required		Topic Areas 1.4, 1.5

2.7 Interactive Visualization

Topic	3	Interaction
Subtopic	**3.1**	The role of interaction: to provide more detail, declutter views, reveal correlations.
	3.2	Types of interaction 1 – basics: • Pan and zoom. • Selection. • Brushing-and-linking across multiple views. • Details-on-demand. • Searching.
	3.3	Types of interaction 2 – manipulating the layout:

Topic	**3**	**Interaction**
		- Dynamic filtering (through query widgets). - Sorting.
	3.4	Types of interaction 3 – manipulating the data: - Aggregate and slice data. - Annotation.
Prior knowledge required		Students have seen the visualizations in static forms in Unit 1 and in other Topics in Unit 2.

2.7 Interactive Visualization

Topic	**4**	**Critique**
Subtopic	**4.1**	Design considerations of audience and task: - What are the audience capabilities, what questions do they have? - Do existing visualizations successfully address the target audience and task? - What is the context in which the visualization is presented (*e.g.* as a supporting figure of an article, stand-alone)? - In which context is the visualization intended to be consumed (e.g. on paper, on a mobile phone, on a computer)?
	4.2	Perceptual biases that affect visualization efficacy: - Perception of distances and areas: Choropleth maps *versus* tiled maps, infographics where an element is scaled (*How to Lie with Statistics*). - Perception of depth / 3D visualization: visual clutter. - Biases associated to color perception: Color blindness, Bezold effect / context, Stroop effect.
	4.3	Inappropriate encodings of data: - Using shape to encode numbers. - Connecting lines on a line chart for discontinuous data. - Using too many hues, shapes, sizes in a single plot.
	4.4	Scales, legends, decorations.
Prior knowledge required		Topic Area 1.5, and Topic 2 above.

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

Commentary for Topic Areas 2.8 and 2.9:

The purpose of Topic Areas 2.8-2.9 is to explore the important concepts of confidence intervals and significance tests:

- in a much shorter time than is typically committed in Statistics classes (*e.g.* AP Statistics in the US)
- in a way that resonates with other elements of Data Science by emphasizing the process of identifying problems, coming up with ideas for solutions and then testing how well those proposed solutions work using simulation as the vehicle for investigating performance.

Because we are trying to convey ideas much more quickly than is usual in Statistics classes, some ideas must inevitably be omitted.

The demands of this Topic Area can be reduced by making experiences with "investigate / explore / discover" very limited and relying much more on received wisdom for lessons to be learned.

2.8 Confidence intervals and the bootstrap

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	An introduction to important concepts of confidence intervals (CIs) that particularly stresses: Motivation of the need for confidence intervals. Interpreting the resulting intervals in data analysis and reporting. A knowledge of the types of uncertainties that confidence intervals do and do not make allowances for. An understanding of the bootstrap as a simple, unified way of generating confidence intervals in a variety of situations. Investigation of how well the resulting intervals work using simulation. As the opportunity presents itself: • To heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse. • To pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	As for students
Learning outcomes	• Able to explain and apply the following:	**Additional learning outcomes**

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	- Distinguish between an unknown population parameter and an estimate of that parameter. - Describe what a sampling error is, explain why sampling errors lead to uncertainties about the true values of parameters, and recognize where these occur commonly in everyday life. - Explain what a "standard error" is trying to encapsulate. - Construct a confidence interval (CI) from an estimate and an accompanying standard-error estimate (2-standard-error interval). - Describe the concept of bootstrap resampling and explain why it is performed. - Explain what resampling error is and how it is used. - Apply computational tools to obtain bootstrap confidence intervals from data in a range of situations. - Discuss the idea of coverage frequency for a method of constructing confidence intervals. - Interpret confidence intervals for single parameters in real-data contexts (e.g. population mean, median, proportion, interquartile range, regression slope). - Interpret confidence intervals for differences in real-data contexts (e.g. differences in population means, medians, proportions; ratios of: proportions, interquartile ranges). - Differentiate between the types of error and uncertainty confidence intervals do and do not address.	*(Pedagogical)* Facility with computer simulation to answer questions involving random phenomena.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Key phrases:	Sampling error, Standard error of estimate, Bootstrap resampling, Confidence interval, Coverage, Simulation.	
Data Science Learning Cycle elements	Exploring the data Analyzing the data Communicating the results	
Prior knowledge required	Topic Areas 1.2, 1.3, 1.7	

2.8 Confidence intervals and the bootstrap

Topic	1	Parameters *versus* estimates
Subtopic	**1.1**	Motivation and examples

2.8 Confidence intervals and the bootstrap

Topic	2	Sampling error
Subtopic	**2.1**	Experiencing the behavior of sampling errors in variety of types of estimates in the context of sampling from a finite population.
	2.2	The concept of the “standard error” of an estimate: what is it trying to capture/measure?
	2.3	Discovering that sample size has a big effect on the sizes of the sampling errors incurred.
	2.4	Discovering (approximately) the inverse-root-n relationship between sample size and sampling error (*e.g.* means and proportions).
	2.5	Discovering that population size has almost no effect on sampling error when the population is considerably larger than the sample (*e.g.* means and proportions).
	2.6T	Simulation and simulation tools.

2.8 Confidence intervals and the bootstrap

Topic	3	Confidence intervals and their implementation using bootstrapping
Subtopic	**3.1**	The concept of a confidence interval.
	3.2	Interpretation of confidence intervals for single parameters in context.
	3.3	Experiencing bootstrap resampling and constructing either percentile bootstrap intervals or 2-bootstrap-standard error intervals in several situations (*e.g.* means, medians and proportions).
	3.4	(2-std error version) Comparing bootstrap standard errors with the results of the well-known and commonly used standard-error formulae for means and proportions that were obtained from mathematical theory;

Topic	3	**Confidence intervals and their implementation using bootstrapping**
		(percentile version) comparing the intervals obtained from bootstrapping and the standard formulae.
	3.5T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.

2.8 Confidence intervals and the bootstrap

Topic	4	**Investigating performance**
Subtopic	**4.1**	Investigating the coverage properties of bootstrap intervals in several contexts and discovering reduced coverage frequencies with smaller samples.
	4.2	*(Optional)* Further performance investigations, *e.g.* discovering problems with bootstrap intervals for medians when the data are very discrete; investigating the use of t-multipliers to overcome small-sample coverage problems; comparing the performance of 2-std-error and percentile intervals in some cases where the bootstrap distributions are typically skewed.
	4.3T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.

2.8 Confidence intervals and the bootstrap

Topic	5	**Differences and ratios**
Subtopic	**5.1**	Experiencing constructing and interpreting bootstrap confidence intervals for differences (*e.g.* CIs for differences in means, medians and proportions) and *optionally* ratios (*e.g.* ratios of proportions in a relative-risk context, or of interquartile ranges).
	5.2T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.

2.8 Confidence intervals and the bootstrap

Topic	6	**Further exploration** *(Optional)*
Subtopic	**6.1**	Repeat some of the above when sampling from a set of theoretical distributions rather than from a finite population. This will require introducing sets of new ideas about the nature and use of distributions..
	6.2T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[Back to Unit 2 contents page]

2.9 Randomization tests and Significance testing

Commentary: Please see the commentary at the beginning of Topic Area 2.8 which also applies to this Topic Area.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	The purpose of this Topic Area is to experience important ideas underlying significance testing through the paradigm of randomization tests, and to do that in the setting where they are most obvious (a randomized experiment for comparing two treatments). This involves picking up, expanding and systemizing the discussion at the end of Topic Area 1.7. The approach is based on the use of simulations. As the opportunity presents itself: • To heighten awareness of how ethical issues can arise in the various steps of the cycle (especially in the data gathering stage), and potential dangers of misuse and abuse. • To pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	As for students
Learning outcomes	Able to explain and apply the following: *(About experiments)* • The nature of, and motivation for, a simple randomized experiment (reinforced from Topic Area 1.7 including the problem of confounding). • Randomized experiments are performed to facilitate reaching causal conclusions about treatment effects (reinforced from Topic Area 1.7). • With a randomized experiment, we can make causal conclusions about the effect of treatment on the units that were actually in the study. • Making causal conclusions about a larger population is possible when the	**Additional learning outcomes** *(Pedagogical)* Facility with computer simulation to answer questions involving random phenomena.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	experimental units are sampled from that population. - Many real-life experiments use convenience samples rather than random samples, thus making extrapolation of causal effects to wider populations problematic … but science can still progress despite this limitation. *(About randomization tests)* - Pure chance can produce surprisingly large apparent differences between randomly constructed "treatment groups" even when no actual treatments have actually been performed (reinforced from Topic Area 1.7) - the point of conducting significance tests (randomization tests in the present context) is to try to answer the question: "Do the effects we are seeing in the data from our experiment demonstrate beyond reasonable doubt that there are real differences between the effects of our treatments or could what we are seeing easily be produced by chance alone?" - The "chance alone" mechanism relevant to a randomized experiment is the random allocation of experimental units to treatment groups (or random application of group labels). - Before we can start to believe that we have evidence showing that real treatment differences exist, the estimated treatment-group differences in our data have to be large compared with what chance-alone generally produces. - This last translates to observed treatment differences that fall close to	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	the edge of the randomization distribution or beyond it - that these ideas are to help us assess whether we have evidence for the existence of true differences but say nothing about their size. For that we use estimates and confidence intervals.	
Key phrases:	Hypothesis test, P-value, Permutation test, Randomization test, Significance test, Simulation.	
Data Science Learning Cycle elements	Exploring the data Analyzing the data Communicating the results	
Prior knowledge required	Topic Areas 1.3, 1.7	

2.9 Randomization tests and Significance testing

Topic	**1**	**Randomized Experiments and Randomization variation**
Subtopic	**1.1**	Revisiting Topics 3 and 5 of Topic Area 1.7 (Motivation for and nature of randomized experiments). Why concluding that the treatment applied actually makes a difference to outcomes is not easy. Experiencing large apparent “treatment differences” in a chance-alone world - investigating the behavior and extent of the random variation in differences in means/medians/proportions when individuals are randomly allocated to “treatment groups” (random labeling) but no actual treatments are performed.

2.9 Randomization tests and Significance testing

Topic	**2**	**Towards the randomization test**
Subtopic	**2.1**	Generation and display of randomization distributions corresponding to scenarios of some real experiments.
	2.2	(*Discussion*) What (qualitatively) would you have to see in your data before you could start to conclude that you had evidence of a real difference?
	2.3	Using real experimental data, motivating the idea of re-randomization of experimental data (randomly relabeling); construction of a (re-)randomization distribution; comparing experimental results to the randomization distribution.
	2.4	(*Continuation*) Generating the randomization distribution; marking the position of the experimental result on the randomization distribution; having discussions about whether the experimental difference is sufficiently bigger than what chance

Topic	**2**	**Towards the randomization test**
		alone generally produces to believe we have evidence of true differences (discussion and idea-seeking, not blind application of a P-value rule).
	2.5	(*Continuation*) Discuss scope of inference justified by the real experiment(s) used.
	2.6T	Simulation and simulation tools.
Prior knowledge required		Topic Areas 1.3, 1.6, 1.7

2.9 Randomization tests and Significance testing

Topic	**3**	**Randomization test**
Subtopic	**3.1**	Systemizing the above to a general procedure for conducting randomization tests.
	3.2	Analyzing data from several experiments involving both continuous and binary outcome measures, and with both clear and obviously “nonsignificant” treatment differences; discussions about conclusions and scope of inferences for each.
	3.3	*(Optional)* Introduce the language and idea of P-value and perhaps the “sidedness” of a test.
	3.4	*(Optional)* Generalize to 3 or more groups using a very simple distance measure.
	3.5T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.
Prior knowledge required		Topic Areas 1.3, 1.6, 1.7

2.9 Randomization tests and Significance testing

Topic	**4**	**Investigating the performance of randomization tests in a sampling context** *(Optional)*
Subtopic	**4.1**	Taking a large data set to be used as a population to sample from, (a) (i) Construct a modified version in which the means/proportions of the subpopulations to be compared are the same. (b) (ii) Construct a modified version in which the means/proportions of the subpopulations to be compared differ by a specified amount. Using simulation, investigating the behavior of a randomization test in situations (a) and (b) for a range of sample sizes. Discussion of issues that have to be considered to address this problem Introduction the formal names for some of the concepts and criteria used above.
	4.2T	Targeted simulation goals, principles, tools and strategies for making it easier for students to experience and learn from simulations.
Prior knowledge required		Topic Areas 1.3, 1.6, 1.7

2.9 Randomization tests and Significance testing

Topic	**5**	**Confidence intervals in a randomized-experiment setting** *(Optional)*
Subtopic	**5.1**	Randomization tests help us assess whether we have evidence for the existence of true treatment differences but have nothing to say about the sizes of those treatment differences. For that we need estimates and confidence intervals Using simulations, investigate how well bootstrap confidence intervals for treatment differences work in a randomized-experiment setting. [Convenience-sample version: randomly allocate group labels to a dataset and add a "true" treatment difference, calculate the interval, check for coverage and repeat; for a sampling version, include sampling from populations in the scenario.]
Prior knowledge required		Topic Areas 1.3, 1.6, 1.7 Topic Area 2.8 Bootstrap methods and Confidence intervals

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]

2.10 Image data

Commentary:

Images play a large part in the social activity of high school students, and they can provide very engaging examples where data science ideas can be practiced and exploited. Image data tends to be much larger than most data students see elsewhere, and the variety of encodings is larger, so this can help reinforce important concepts. In this unit, we will focus on still images; similar ideas can be applied to video, or video with sound.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Aims & Purposes	• To introduce students to some of the ways image data are analyzed and used in everyday life. • Apply the *Learning from Data* cycle to think about tasks in which data take the form of images. • To experience some of the choices available in representing image data, and the impact of these. • To demonstrate that images can be represented by two (at minimum) dimensional numerical matrices, and each numerical value has a pre-defined relationship to a color. • To illustrate how mathematical operations can be performed on the matrices representing image data to produce visual effects such as rotation, translation, scaling *etc*. • To illustrate how measures/features can be extracted from images for analysis and used to classify and alter (transform/filter) images. • To provide opportunities to apply what was learned in Unit 1 and Topic Areas 2.4. • As the opportunity presents itself: – To heighten awareness of how ethical issues can arise in the various steps of	As for students, but at a deeper level of understanding and with a greater level of technical mastery, so that they can guide students and assist them in their experiences; also greater reflection on what they are doing, and more awareness of alternative approaches.

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	the cycle, and potential dangers of misuse and abuse. – To pay specific attention to issues relating to data quality, questioning skills, and presentation skills.	
Learning outcomes	Students will be able to: • Describe a variety of tasks where image data is analysed • Recognize the variety of encodings for image data, and explain how choice of encoding affects resolution, compression, *etc.* • Determine ethical issues that arise in working with images, including those in obtaining images, those in manipulating images, those in the ways an analysis is used. • Demonstrate how images can be represented numerically (by matrices), and relate visual effects on images to mathematical operations on matrices. • Explain how features can be extracted from images, and then used in classification tasks, *etc.* • Transform images to quantities that can be analyzed. • Apply a supervised learning algorithm to classify images. • Calculate misclassification errors, false positives and false negatives and use these quantities to assess the output of a classification algorithm on images.	**Additional learning outcomes** • Aware of the very high dimensionality of feature vectors, and the impact this has on classification tasks. *(Pedagogical)* ● Aware of common student misconceptions or difficulties relating to image data.
Key phrases:	Classification, Compression, Extraction, Image processing, Pattern recognition, Pictures, Supervised learning, Unsupervised learning	
DS Learning Cycle elements	*All steps*	

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
Prior knowledge required	Topic Areas 1.1-1.4, 1.6, 2.4	

2.10 Image data

Topic	**1**	**Examples of image data and their uses**
Subtopic	**1.1**	Examples of images students are familiar with (*e.g.* photos)
	1.2	Some of the ways image analysis is used, including: • Those in everyday life of the students *(e.g.* managing their photos, using software that automatically picks out photos that show particular family members or friends). • Uses in health (identify cases of possible tumor in scan), security (*e.g.* suspicious parcel detection; authentication at immigration, airplane boarding, school gates), industry (detect defective parts, fruit quality evaluation, *etc.*), transport (cars that recognize speed limit signs, or self-driving cars that detect pedestrians), the entertainment business (detecting uses of copyrighted images in user-generated content), policing (fingerprint identification, bullet markings, *etc.*).
	1.3	Image classification as an example of a task that is both useful in itself and a stage towards image matching, locating items within images, *etc.*

2.10 Image data

Topic	**2**	**Image data origins and encodings**
Subtopic	**2.1**	Where do images come from (digital cameras, scans of print, computer generated images)?
	2.2	Overview of the ways an image is encoded in digital form. Key concepts: pixels, viewpoint and perspective, grey scale, colors (hue, saturation), resolution, compression). Mention of various official and *de facto* standards (JPEG, GIF, PNG, PDF). The space used by image data.
	2.3T	Comparison between encodings; how image quality can vary depending on the encoding used (*i.e.* lossless *versus* lossy).
	2.4T	Encodings for video and audio.
Prior knowledge required		Topic Area 1.6

2.10 Image data

Topic	**3**	**Image data representation and basic mathematical operations on images**
Subtopic	**3.1**	Explain how images can be represented using a numerical matrix. The most common representation is two dimensional (M rows and N columns). Another

Topic	3	Image data representation and basic mathematical operations on images
		common representation is three dimensional of size M*N*3, where the last dimension is 3 representing component Red(R), Green(G) and Blue(B) values for every pixel. Discuss common images size in terms of number of pixels such as 640 by 480, 1024 by 768, *etc*. Each R, G and B value ranges from 0 to 255.
	3.2	Demonstrate how the numerical value of each pixel is directly related to a color. For the RGB representation, the relative values of each of the components indicates the resulting color. R=255, G=255, B=255 indicates white, while R=0, G=0, B=0 indicates black, and so on. A color wheel or color bar could be used for illustrative purposes.
	3.3	Demonstrate how visual effects can be introduced on images by applying mathematical operations on the numerical matrices representing images. For example, illustrate translation, rotation and scaling as examples of common transformations on image.
	3.4T	Higher dimensional representations such as 3D spatial data where the third dimension is an additional spatial coordinate (*e.g.* X, Y, Z).
	3.5T	Discuss how these image processing operations can be used to augment image data when used for classification.

2.10 Image data

Topic	4	Feature detection in images
Subtopic	**4.1**	Examples of low-level features, both local (*e.g.* existence of an edge at a location) and broad/global (*e.g.* dominant color over the image).
	4.2	Overview of some ways features can be identified in an image. A simple example of how an edge-detector could work. Some tools that find various features in images.
	4.3	A feature vector as an imperfect representation of the image.
	4.4T	Alternative choices for the features to consider.
	4.5T	Impact of high-dimensionality of typical feature vector representations.
Prior knowledge required		Topic Area 1.6

2.10 Image data

Topic	5	Image classification
Subtopic	**5.1**	Review of the concept of a supervised-learning approach to a classification task (from Topic Area 2.4); the importance of training data.
	5.2	How high-quality training data can be obtained for image classification (hand-label a corpus, check work of humans). Issues of the spread of examples (*e.g.* risks to classification effectiveness if the training data over-represents particular groups or

Topic	5	Image classification
		situations). Alternative: source examples of each class separately, and the extra threats to classification validity in doing this (examples such as sets where criminals came from police files, others from social media; or enemy helicopters from combat photos, friendly helicopters from ground shots).
	5.3	Examples of applying an automated classifier to the feature vectors, to learn a classification.
	5.4T	Advantages and disadvantages of running a classifier directly on image data *versus* using it on feature vectors.
	5.5T	Unsupervised learning approaches to image classification.
Prior knowledge required		Topic Area 1.6, 2.4 (or else, this could be treated as an introduction to some of the concepts of Topic Area 2.4); teachers require Topic Area 2.5 as well.

For a fuller exposition of this Topic Area, including teaching and resources ideas, see http://idssp.org/pages/framework.html.

[*Back to Unit 2 contents page*]