IDSSP: the International Data Science in Schools Project

Draft Curriculum Framework

Open for **Public Feedback** until **31 May 2019**

***This document*:** http://www.idssp.org/files/IDSSP_DataScienceDraftCurriculumFramework_Schools.pdf

See also ***Abbreviated Topic Lists*** at: http://www.idssp.org/files/IDSSP_DraftFramework_AbbreviatedLists.pdf

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 quantitative 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 they 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, are developing a curriculum framework 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 (classroom, MOOC, self-learning, ...).

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 assembled, 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.

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.

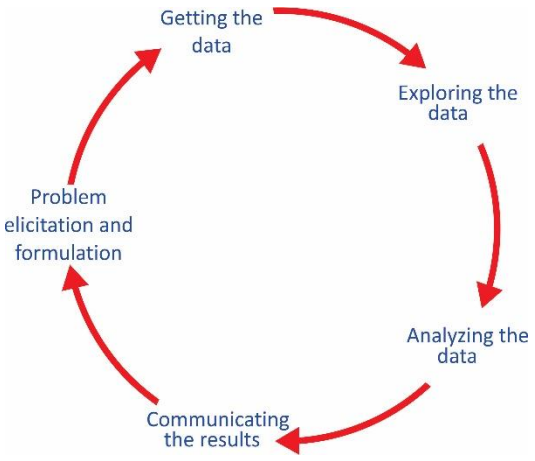


Figure 1 The basic cycle of learning from data

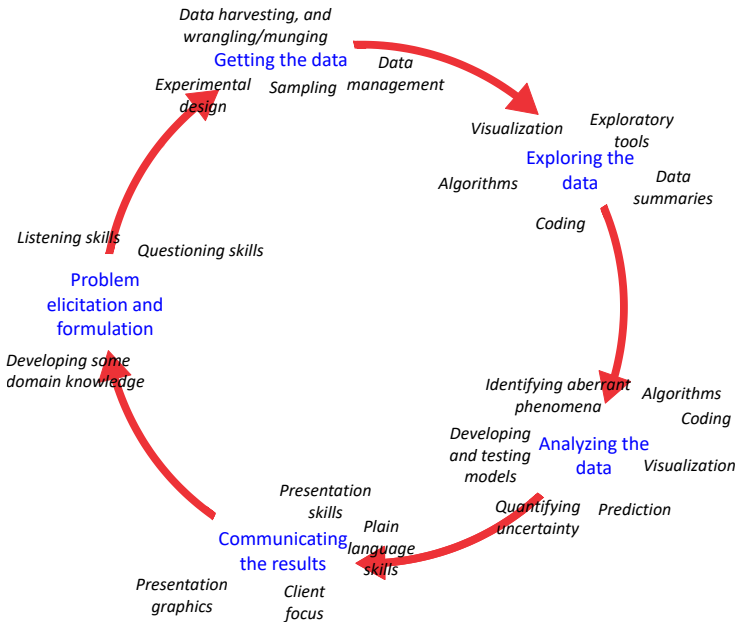


Figure 2 The varying activities and disciplines involved in learning from data.

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.

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

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

For teachers: no prior knowledge of either Statistics or Computer Science 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 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 (Section 3).

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 R or Python) 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 self-contained options from which a course of study can be assembled.
- The curriculum for teachers comprises the curriculum 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 at present.

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 will 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.

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.

The Unit comprises seven *Topic Areas*:

1.1. Data Science and Me
1.2. Basic Tools for Exploration and Analysis (BTEA). Part 1: Tools for a single variable
1.3. BTEA Part 2: Pairs of variables
1.4. BTEA Part 3: Three or more variables
1.5. Graphs 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.

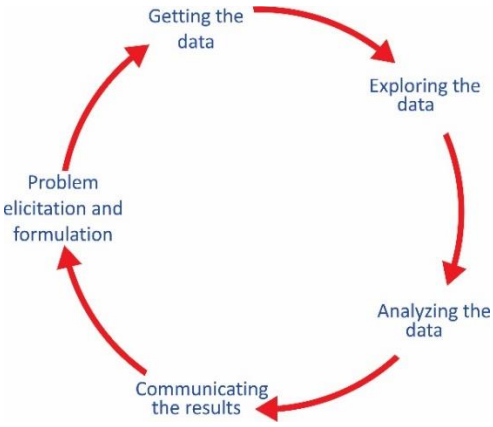


Figure 3 The basic cycle of learning from data

Topic Areas 1.1 – 1.4 (*Basic tools 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 variables simultaneously.

Following this initial experience with exploratory tools and default settings, Topic Areas 1.5 (*Graphs 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 (*Graphs 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. *Graphs 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 variables that cause problems in drawing causal conclusions) and random error, and ways of handling these problems.

2. 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 tools for exploration and analysis, Part 1: *Tools for a single variable*
1.3 Basic tools for exploration and analysis, Part 2: *Pairs of variables*
1.4 Basic tools for exploration and analysis. Part 3: *Three or more 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 Graphs and Tables: how to construct them and when to use them

Aims: to develop the students' understanding of appropriate choices and uses for graphs 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 features in the initial exploration phase and when evaluating the adequacy of model fits; 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.

3. 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 —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

Topic Area:	**1.1 Data Science and Me**
Version:	30 April 2019

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	Able to: ● Converse about what data are and where they come from ● Identify some of the roles data play in their own lives ● Describe some types of problem that can be answered with Data Science ● Describe a time when data were collected on themselves ● Produce examples and real-life solutions in real life where predictions and classifications can be made using Data Science thinking ● Provide examples of the social and personal consequences of predictions derived from models built from data ● Describe errors in decisions and predictions owing to faulty use of data ● Discuss how data can support making decisions ● Provide examples of problems within the context of the data-science cycle, and to relate activities to appropriate phases of the 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 some other discipline(s) and use it/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

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 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 (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.

TOPIC :	3	Sources of data
	3.6	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.
	3.8T	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) of 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

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 tools for exploration and analysis

Overview

- The Basic Tools Topic Area has been broken into 3 parts
  - Part 1: Tools for a *single* variable (Topic Area 1.2)
  - Part 2: *Pairs* of variables (Topic Area 1.3)
  - Part 3: *Three or more* 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 tools 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*]

Topic Area:	**1.2 Basic tools for exploration and analysis. Part 1: Tools for a single variable**
Version:	30 April 2019

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.	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	Able to ● import a rectangular data set in csv or tab-delimited text format into suitable software and obtain basic graphics and summaries for inspecting ***individual variables*** in a data set (*i.e.* one variable at a time) ● explain the need for data cleaning and some simple strategies for cleaning data ● interpret commonly-used graphical displays ***for a single variable*** ● use these graphical and numerical tools to explore the variables in a data set and comment on important features ● explain what a particular graph reveals about important or interesting features of the data in the context of the original problem ● explain what commonly-used data summaries tell us about the data in conceptual terms, and describe how they relate to features of graphical displays ● explain what the values of commonly-used data summaries tell us about the data in the context of the original problem Continue to develop a critical faculty in relation to data: ● asking self and others questions about data origins, quality and applicability to a problem under consideration ● usefulness for generalization	**Additional learning outcomes** Able to: ● understand and work with the formulae used for commonly-used numerical summaries **for a single 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) ● Knowledge of pedagogical issues about these topics, so they can teach them effectively.

	• suggesting/hypothesizing possible generalizations	
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 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	

COMMENTARY specific to Topic Area 1.2

- Treating this as a univariate module looking at single-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 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 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 variable

TOPIC :	**1**	**Rectangular data sets**
Subtopic:	**1.1**	Observations and variables
	1.2	Numerical versus categorical 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 variable

TOPIC :	**2**	**Graphics and summaries for a single categorical 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 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 variable

TOPIC :	**3**	**Graphics and summaries for a single numeric 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	Features to look out for and their implications: outliers, center, spread, shape, modality, spikes, gaps, ...
	3.3	Summaries: Minimum & maximum, mean & median, quartiles, interquartile range and standard deviation.
	3.4	Box plot as plot of summaries.
	3.5	Converting numerical variables to categorical: When, why and how? (Numbers as codes, binning continuous 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*]

Topic Area:	**1.3 Basic tools for exploration and analysis. Part 2: Pairs of variables**
Version:	30 April 2019

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	Able to • read commonly-used graphical displays for the relationship between a ***pair of variables*** • explain what a particular graph for the relationship between a pair of variables reveals about important or interesting features of the data in the context of the original problem • explain what commonly-used data summaries for pairs of variables tell us about the data in conceptual terms, and describe how they relate to features of graphical displays • explain what the values of commonly-used data summaries for pairs of variables tell us about the data in the context of the original problem • apply tools that deal with over-printing and large-data-set problems with scatter plots Continues to develop a critical faculty in relation to data: • asking self and others questions about the source of the data, data quality, and applicability to a problem under consideration • usefulness for generalization • suggesting/hypothesizing possible generalizations	**Additional learning outcomes** • Able to understand and work with the formulae used for commonly-used numerical summaries for a pair of 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.
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 features 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 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 variables

TOPIC :	**1**	**Comparing groups** (relationships between a numeric and categorical variable)
Subtopic:	**1.1**	Making comparisons using the graphs and summary types learned for a single 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 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 variables

TOPIC :	**2**	**Relationships between two numerical variables**
Subtopic:	**2.1**	Scatter plots
	2.2	Outcome/Response variables versus Predictor/Explanatory 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

TOPIC :	2	Relationships between two numerical variables
		• jitter 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 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 variables

TOPIC :	3	Relationships between categorical 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		Topic 2 of Topic Area 1.2

1.3 BTEA, Part 2: Pairs of 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 variable (*e.g.* girls only), intervals of a numeric 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*]

Topic Area:	**1.4 Basic tools for exploration and analysis. *Part 3*: *Three or more variables***
Version:	30 April 2019

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	Able to ● read basic graphical displays and tables of summaries for looking at ***three or more variables simultaneously*** ● explain what a particular graph reveals about important or interesting features of the data in the context of the original problem ● explain what commonly-used data summaries tell us about the data in conceptual terms, and describe how they relate to features of graphical displays ● explain what the values of commonly-used data summaries tell us about the data in the context of the real problem Continues to develop a critical faculty 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	**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) ● Communicating conclusions (communicating what they see in plots and summaries, and discussing possible implications) *Indirectly*: ● Getting the data: Data exploration to detect anomalous features 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 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 variables

TOPIC :	**1**	**Pairs plots** (matrices of 2-variable plots for all pairs of variables in a chosen set of variables)
Subtopic:	**1.1**	Pairs plots that will cope with categorical as well as numerical 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 variables

TOPIC :	**2**	**Subsetting by a third variable**
Subtopic:	**2.1**	Panel plots/faceting and 3-dimensional summary tables as a means of investigating the extent to which two-variable relationships look consistent across subpopulations or through time, or show some sort of trend
	2.2	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 variables

TOPIC :	**3**	**Other ways of adding information on additional variables to 1- and 2- variable plots**
Subtopic:	**3.1**	***Coloring*** points in dot plots and scatter plots according the value of an additional variable
	3.2	***Sizing*** points in scatter plots according to the value of an additional variable
	3.3	***Labeling*** points in dot plots and scatter plots according the value of an additional 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 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*]

Topic Area:	**1.5 Graphs and Tables: how to construct them and when to use them**
Version:	30 April 2019

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 graphs 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 graphs *versus* tables as methods for either exploration or presentation purposes. Graphs are essential tools in learning from data: in the initial exploration phase; when evaluating the adequacy of model fits; 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 graphs 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 graph and using it effectively, and also being able to identify when to use a graph and when to use a table.
Learning outcomes	Able to ● Use graphs to explore a data set and identify potentially interesting patterns and unusual features. ● Use graphs to evaluate whether a proposed model is an adequate description of the data. ● Apply simple principles of graph construction in deciding what sort of graph to use to display data. ● Use an appropriate choice of graph to present the findings of an analysis. ● Decide whether to use a graph or a table or both to present data and information ● Identify major flaws in a graph or tabular presentation of data (*e.g.* taken from news etc), explain why this is not communicating accurately or effectively, and suggest improvements to the presentation	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 experiences; also greater reflection on what they are doing.

Parts of Data Science learning cycle addressed	*Directly*: ● Exploring the data ● Communicating conclusions (especially via graphs and tables).

1.5 Graphs and Tables

TOPIC :	**1**	**When to use a graph and when to use a table**
Key words and phrases:		Purpose of a graph; Purpose of a table; Infographics; Presentation graphics; Infographics; Infotables; Exploring data; Discovering pattern
Subtopic:	**1.1**	Exploration and discovery *versus* presentation – comparing and contrasting the goals Graphs *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 variables ● Exhibit and have informal discussion of examples of good and bad graphs (to be dissected later in the Topic Area). What information does each graph appear to be purveying? And is anything about the graph making it hard to extract accurate information? Is an additional graph (or graphs) needed to explore or to communicate effectively?
	1.2	Infographics and infotables ● Purpose is primarily to draw attention to a story ● Principles of graph 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 Graphs and Tables

TOPIC :	**2**	**What makes a graph good or bad?**
Subtopic:	**2.1**	The graphical process: a chain from graph creator to graph interpreter: **Data** -> Analysis and interpretation ->Information extracted by analyst -> Information encoding in Graph –> Decoding by user -> Decoded information as extracted by user -> **Action or decision enabled for user** How does the information as-decoded-by-the-user compare with the information originally encoded in the graph? 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 variable of interest is the difference between two curves; dot chart – 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 ● Other
	2.4	Visualization Principle 3. Encoding variables ● Use of shapes to identify different 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 graphs may provide more accurate pictures of different aspects of pattern.

TOPIC :	2	What makes a graph good or bad?
		• Examples: using at least two categorical variables and a continuous 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 • Axis markings • Positioning and use of legends • Size of plotting symbols
	2.8T	Pedagogical issues relating to comparing different types of graphs • 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 features
Prior knowledge required		Topic Areas 1.2 & 1.3

1.5 Graphs and Tables

TOPIC :	3	What sort of graph 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 graphs of different samples
	3.2	Plotting a numerical and a categorical 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 graph to show different patterns accurately
	3.3	Plotting 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

TOPIC :	**3**	**What sort of graph should I use?**
	3.4	Plotting two numerical variables to explore relationships • Scatterplots • [Pairs plots/Scatterplot matrices] Plotting several pairs of variables • Adding a smooth curve to bring out a relationship
	3.5T	Pedagogical issues relating to comparing different types of graphs • Experimentation in encoding the same information using different graphical types and the consequent impact on the accuracy of the decoded information • Explore how more than one graph might be needed in any given situation
Prior knowledge required		Topic Areas 1.2 & 1.3

1.5 Graphs 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 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*]

Topic Area:	**1.6 The data-handling pipeline**
Version:	30 April 2019

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	Able to • discuss their own experiences with tool use during the stages of the Data Science lifecycle (including collecting data, subsequent data processing, and communicating the results)	Additional learning outcomes • Sufficient mastery of a tool so they

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	• explain the major steps in data-science processing activities, and how tools can be used in each • understand and use correctly, common terms for data processing and management activities • discuss important data management principles and how they apply in a given setting • discuss the usefulness of automating the performance of data-science processing activities in a given setting • work effectively with simple code (understand what it does, modify it in small ways) • discuss the usefulness of data cleaning and transformation, in a given setting • recognize situations where simple data cleaning and transformation is important • implement simple data cleaning and transformation operations • discuss the usefulness of generating presentations of data that combine graphs/tables with other content, and of customizing aspects of the graph/table to suit • implement generation of presentations that combine (perhaps customized) graphs/tables with other content	can guide students who get stuck, and so they can learn 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
Parts of Data Science learning cycle addressed	*Directly*: • Getting the data • Exploring/Analysing 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

TOPIC :	1	Introduction to tool support for the data-handling pipeline
		• 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 vs sensor) • Running surveys • 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 vs Unicode • Existence of differences between environments (*e.g.* Unix vs Windows files) • CSV • JSON • Compression • Proprietary formats (xlsx, database internal)
	2.4T	Case studies of good data sources

TOPIC :	2	Getting and storing data
	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 graphs *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

TOPIC :	3	Tool support for exploring and analyzing data
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 graphs 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 graphs/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 ● 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*]

Topic Area:	**1.7 Avoiding being misled by data**
Version:	30 April 2019

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 practises 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	Able to
	• discuss the idea of artefacts in data (artificial patterns caused by deficiencies in the data collection or accumulation process) and give illustrations
	• discuss examples of measures used in published reports that do not measure what they purport to measure, or comparisons that use subtly different measures, and how this can lead to bad conclusions
	• discuss examples of selection biases, or other examples of filtered data streams, and how these can lead to bad conclusions
	• discuss examples of bad handling of missing values (*e.g.* repeating values for previous years)
	• explain the idea of confounding (or lurking) variables and why observational data alone cannot reliably demonstrate that an effect is causal
	Towards Solutions. Able to:
	• explain the ideas of validity and reliability of measures
	• discuss random sampling as a strategy for overcoming selection biases
	  ○ Can explain the consequent problem of sampling error in estimates
	  ○ Can explain the idea of a margin of error (alternatively, interval estimate) as a means of allowing for error, what types of error these make allowances for and what types of error they do not address
		Additional learning outcomes
		• **…**
		• **…**

Course	Introductory Data Science (IDS)	Teaching the Teachers (T3)
	○ Can explain that collecting more data reduces the size of random errors in estimates but not of systematic errors ○ Can obtain a margin of error or confidence interval to accompany an estimate in some simple situations ○ Can communicate an estimate in the context of the original problem in a way that conveys its uncertainty ● discuss randomized experiments as a reliable strategy for investigating causation ○ Can explain the consequent problem that random allocation of experimental units to treatment groups, operating alone, can introduce apparent treatment differences that are artificial ○ Can explain the results of a randomization test comparing "treatments" in a simple randomized experiment in the context of the original problem	
Parts of DS learning cycle addressed	● Getting the data ● Exploring/Analyzing the data ● Drawing conclusions	

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

TOPIC :	2	Bias and what we can do about it
		• 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
	2.5T	
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 variable can change, and even reverse, the apparent relationship between an outcome and predictor variable (Simpson's paradox as a special case). Idea of a confounder/lurking 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 parameter (*e.g.* mean, median, proportion).
	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	When can we reasonably conclude observed group differences demonstrate real treatment effects? (discussion)
	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

How does the Unit 2 document differ in structure from the Unit 1 document?

For Unit 2, which covers less familiar territory we have retained more paragraphs containing ideas about teaching and resources. Most of these topic areas do not have a history of being taught at the school level and little at intro-level tertiary.

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.

It 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.

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 features 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 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 features in practical settings.

Details of Unit 2 Topic Areas

Topic Area:	**2.1 Time Series data**
Version:	30 April 2019

Commentary

- Data that depend on the time when they were measured constitute an important type of data that was 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 are quite an advanced topic, so will be ignored for this Topic Area).
- Times series with strong seasonal features are extremely common and thousands of interesting series can be found on the websites of government agencies and scientific organizations
- Strongly seasonal time series give 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 breath 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 features 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	Able to: • Explain what a time series is and give examples of why people are often interested in how a variable changes over time. • Explain why, in time-series plots, the data points are usually joined by lines. • Recognize features such as spikes, sudden jumps, gaps, trends and seasonal effects in time-series data and seek explanations for them.	**Additional learning outcomes** (*Pedagogical*) •

	• Describe and interpret a trend in a time series. • Explain and interpret the nature of additive or multiplicative seasonal effects in a time series when appropriate. • Describe a seasonal time series in terms of a model of the form response = trend + seasonal oscillation + random noise. • Interpret time series graphics and communicate their findings. • Compare related series by comparing features 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.	
Key phrases:	Forecasting; Holt-Winters forecast; Jumps; Prediction; Seasonal effects; Spikes; Seasonal decomposition (STL)	
DS Learning Cycle elements: *All steps*		

Teaching commentary

- Learning to use new types of plot to expose or highlight structure in data also involves learning to think in new ways. Such thinking is developed in conjunction with learning to describe what is being seen and what it means. So learning experiences involving data exploration/analysis and communication have to happen together. Learning to communicate is not something that "can be postponed till the end". For this reason Topic 5 on communication simply focuses a wrap-up form of communication -- assembling the pieces into a coherent story for a client.
- Learning outcomes that do not correspond to listed topics should be addressed in the pedagogy of delivering the topics by exploiting opportunities for thinking provided by data sets.
- Teachers need some more content knowledge but also pedagogical knowledge about how to get students starting to ask the right questions about data.

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 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 variables, aggregating a response 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).
	2.4T	
Prior knowledge required		Topic Area 1.6

TOPIC :	3	Exploring the data
Subtopic:	**3.1**	Basic time-series plots; smoothing to reveal trend; recognizing features 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 features such as their trends, seasonal patterns and residual behavior; communicating the results
	3.5T	
Prior knowledge required		Topic Areas 1.3, 1.4, 1.5
Ideas about teaching		**Decomposition:** • Multiplicative effects are considerably harder to explain to students than additive effects • Unfortunately, almost all seasonal series with reasonable movement in the trend compared with the extent of their seasonal swings (*i.e.,* are not almost flat) exhibit strong multiplicative behavior • This shows up as the swings being larger when the series is high and smaller when the series is low. Working on a log-scale is too abstract for a high-school audience • The problem has to be addressed. One way forward is shown here https://www.youtube.com/watch?v=85XU1T9DIps (6 mins) continued with an emphasis on multiplicative at https://www.youtube.com/watch?v=CfB9ROwF2ew (5 mins)

TOPIC :	3	Exploring the data
		• STL decomposition is due to R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: *A Seasonal-Trend Decomposition Procedure Based on Loess*. Journal of Official Statistics, 6, 3–73. The version in base R unfortunately plots the trend, seasonal and residual in panels of equal height so differences in contribution have to be deduced from the scale markings. This makes it harder to see how the components match the original data. The movies above put everything on the same scale **Missing values** • Many specialist TS methods only work on equally-spaced time intervals and without missing values. If there are only a few missing values they can be filled in using some sort of imputation and sensitivity of conclusions to the imputed values explored. Care must be taken with imputation, as the statistical structure of a time series is highly important.

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 features 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 features Where appropriate, modifying output automatically generated by software to make these features 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
	5.4T	
Prior knowledge		Topic Area 1.5

TOPIC :	**5**	**Communicating the Results; next question?** *(in the context of a particular investigation targeting a particular real-world problem)*
required		

- **See Example Case Study: Canadian Temperature** (*Example and Code*)

[*Back to Unit 2 contents page*]

Topic Area:	**2.2 Map data**
Version:	30 April 2019

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 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	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 location and regionally tagged data on maps, and 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

	• Describe basic structures of regional maps: basic polygon structures, boundaries, lines, fills, and how are they useful • Discuss biases and distortions that are part of most maps ("the map is not the territory") • Understand how maps can represent many competing factors simultaneously, with tremendous "user degrees of freedom"	assumed easy translation into our "real world" • Convey to the students some of the more subtle factors in map creation, especially in the user choices (scale, projection – Mercator! and similar factors).
Key phrases:	Faceting, Geo-tagged data, Map, Neighborhood, Overlay, Projection, Tagged data, Visualization,	
DS Learning Cycle elements: All		

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 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	Multiple dimension discussion, 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) • Allow students to use an interactive tool to choose variables, with the map scope and resolution pre-set The focus is on introducing the challenges associated with selection of 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 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 variables are.	

TOPIC :	**1**	**What are the purposes of Maps?**
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 variable layers beyond the geographic (*e.g.,* human, societal, or environmental)	
Ideas about teaching	This is a good opportunity to integrate this Topic Area into a geography, history, economics or similar course/unit from your school. Bring real resources from the local community in: school district zoning, municipality zoning (commercial versus residential). Bring the point "the data is on the map, but is not the map itself" to the fore.	

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 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

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 variables; issues of scales
	3.4	Perceptual problems with choropleth maps; alternative representations
	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

TOPIC : 3	How do we build and work with regional maps?
Ideas about teaching	So many! The Open Data movement has provided a wealth of geo-tagged data: policing, social justice, environment, climate, business, industry, schooling, and so on. An incredible array of "loosely formatted" data is available, with inherent spatial structure which can easily be explored using the map as a basic working tool. • Look for your local municipality or larger region's (province, territory, state) open data: bring the problem home to the students • Don't overlook other data sources like sports or social issues – the locations of stadiums (NFL for the United States, AFL for Australia, NHL for Canada) could be a great case study • Be prepared to pre-vet a data set which is larger than you need, so there's flexibility to allow for creativity and exploration. • Guide the students in brainstorming so as to lead them to problems that are well scoped and have data available – many things exist, but are locked down or private • Integrate some discussion of privacy and ethics concerns – why is geographic data so invasive of privacy if not carefully screened? *e.g.*, Canada's postal code system, if available at full resolution, can identify down to only 15-20 single-family dwellings per code! • Explore the example of smartphones and their "always on" GPS, with records saved of everywhere you've been. Look for a recent example in the popular news, *e.g.*, https://www.news.com.au/technology/gadgets/mobile-phones/google-has-been-tracking-your-movements-even-if-you-told-it-not-to/news-story/bb9eb906387ffd2295e8b17b24b7d883 (August, 2018)

TOPIC : 4T		(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 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, 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.	
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 variable layers beyond the geographic (*e.g.*, human, societal, or environmental)	
Other things to look at for those interested	**NOTE**: this is a very subtle topic, and may not be suitable for many audiences at this level. Consider it as a suggestion for *teachers*: knowledge of the framework in this unit may enrich your presentation and delivery, and give you small elements to sprinkle into the previous three units.	
Ideas about teaching	This is a great chance to have a cross-over unit with your geography or history unit in your school. Bring historical maps into play: look at *changes* in human political boundaries over time (*e.g.*, the rise and fall of the Soviet Union). Use the temporal dimension and faceting in an interactive way to explore changes in variables *over* time *by* geography. Or consider introducing a small topic within another context on the bias inherent in models, and how even something as seemingly fixed and true as a map can contain bias, *e.g.*, maps of the world before Columbus often distorted the landforms significantly due to Eurocentric biases of the mapmakers. Do biased visualizations still provide value? At what cost? For example, consider the Mercator projection, and the bias it introduces to those who are not familiar with distant regions. For example, the distance, on the ground, from Cairo, Egypt to Johannesburg, South Africa is 8700km! Africa is immense, but the Mercator projection makes it appear smaller than it actually is, introducing biases to thinking for those who only know of the continent through maps, especially northern hemisphere (Europe, North America) inhabitants.	

- **See Example Case Study: Safety of Toronto neighborhoods** (*Example and Code*)

[*Back to Unit 2 contents page*]

Topic Area:	**2.3 Text data**
Version:	30 April 2019

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 models. Many of the types of table and graphs 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 = 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	Able to: • describe measurement levels of variables and why text data is nominal scale. • discuss the motivations for looking at displays of frequencies of tokens • break strings of text in tokens	**Additional learning outcomes** (*Pedagogical*) •

	• 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 graphs 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	
DS 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 features 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		Topic Areas 1.2, 1.6.

TOPIC :	1	Problem elicitation and formulation: Text data
required		
Ideas about teaching		Apps with graphical user interfaces can be very useful for use in conveying motivations for text analysis and what it is capable of. Showing frequency displays before and after stop words have been removed immediately shows the desirability of doing so.

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 of limitations of bag of words analysis and dangers of misuse (cautionary tales)

	2.8T	Pedagogical issues relating to text data analysis:
		• 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")

TOPIC :	3	Sentiment Analysis
	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 & 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 vs. 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

[*Back to Unit 2 contents page*]

Topic Area:	**2.4 Supervised Learning**
Version:	30 April 2019

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	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 models and decide which is the better for a given situation based on total misclassification rate. ● Fit a CART and interpret results ● Use a set-aside data set to compare classification models ● Use software to calculate the complexity of a CART ● Create a validation data set ● Explain how 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 variables.	**Additional learning outcomes** (*Pedagogical*) ●

	• 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	
DS 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 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 model works.
Prior knowledge required		Topic Areas 1.2, 1.3, 1.6
Possible resources *(data, real-life problems, …)*		https://www.idsucla.org/introduction-to-data-science Unit 4 has handouts and lesson plans for the activity described below. https://rstudio.up.pt/shiny/users/pcs/civicstatmap/5.112_TV_TreeClassification_EN.pdf a product of ProCivicStat to explore Tree classification based on a plug-in of CODAP
Ideas about teaching		• It is helpful to begin with a very simple data set in which students can apply their own knowledge to classify observations. The data set should have maybe 12 observations and 3 or 4 variables. (And there should be an additional data set of 6 or more observations that does not include the response variable. Call this the "validation" data.) For example, given data on physical characteristics of American football and soccer players, can students develop rules to classify players as Football or Soccer? How well do those rules work when applied to data for which they don't get to see the correct classification? The data set chosen should allow students to rely strongly on their intuition (because you don't want them to get involved in analyzing the data at this stage) and yet should have some overlap on the variables variable so that perfect classification isn't possible. • A good introductory activity is to use a data set as described above and ask students to work together to develop a set of rules that could be used to classify players as football or soccer players. We're using the variable "sport" as a response variable. This knowledge of the true classification is what makes this practice "supervised". • Students might not understand that they can't use the response variable to use in their classification rules. A discussion about what variable is the "response" and which are the predictors might help. Otherwise, they will quickly discover they made a mistake when they get to the "new" data, and that is why it is good to give them a chance to revise their rules. • Students will want clarification on what is a "rule". Explain that it should be a set of instructions that another group could carry out without consultation with them. You might prepare some practice data and ask groups to swap rules to test them for clarity.

TOPIC :	1	Problem elicitation and formulation: Supervised Classification
		• Remind students that they can have several rules, and they can use different variables in these rules. (For example, if the player is taller than x and weighs more than y and is younger than z.) The temptation will be to give a single rule (If taller than x....) and stop. • Students will have a variety of reasons as to why they are confident that their classification was good or not good. Steer the conversation towards recognizing that they have data they can use. How successful were their rules on the first batch of data? How would they measure this? (Someone will come up with something that can be called the misclassification rate.) Do they think that's a good gauge for their success with future data? Why or why not? Are they making assumptions about the future data when they answer this? • It's best to start out with binary classification: two categories.

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.
Possible resources *(data, real-life problems, …)*		A good place to start is the CODAP worksheet (see above under Topic 1), which has no implemented algorithm: the user is still in charge with all the decisions to make; but the program relieves the user from calculating misclassification rates when testing a new split; the program allows you to play with the data Another wonderful resource is Tim Erickson's data game "xenobiologist" https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=33583
Other things to look at for those interested		Examine and describe the algorithm used by software in the rpart package of R
Ideas about teaching		• Students can form a "tree" in class, in which some play "data" (and these are given data cards with characteristics), branches (which ask questions and route the "data" cards left or right) or nodes (which classify data into one of two groups). This will give them a feel for how the tree works and what the components are named. • Give students a chance to create their own algorithms, written in pseudo code. The goal is, for a given variable, to split the data so that the response variables are in two groups and the two groups are as homogenous as possible. Some questions for students to consider (a) how can they measure this homogeneity? (b) How can they know which split

TOPIC :	2	Introduction to Classification Trees
		is best? (c) Does their algorithm handle multiple variables? • There are many possible algorithms for generating trees. Software packages that do this use a particular algorithm (or, often, provide choices for different algorithms), but the point of this lesson is not to learn the canonical algorithm, but instead to generate ideas about how a tree might be implemented on a data set. • For 2.5, students should discuss and consider the consequences of misclassification. For example, a patient might be classified as "requires surgery" when she really does not need surgery or, conversely, classified as "no surgery" when in fact surgery is necessary for survival. Although both errors might be counted as "equal" when calculating the misclassification rate, the consequences of these errors are not equivalent. A more advanced class might also want to discuss the role of "hidden" variables. For instance, misclassifications in medical settings might be low, but might occur more often for particular groups (females, or ethnic groups) than others, and so this "hidden" variable is hiding a source of social inequity.

2.4. Supervised Learning

TOPIC :	3	Growing Classification Trees
Subtopic:	**3.1**	Introduction to R commands (or from another appropriate language) to grow and visualize 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
Possible resources *(data, real-life problems, …)*		The icu data set in R (https://www.rdocumentation.org/packages/vcdExtra/versions/0.7-1/topics/ICU) is relatively friendly and makes some intuitive sense for students. Kaggle.com has many possible data set, but expect to spend some time finding one with meaningful categorical (and binary) outcomes.
Other things to look at for those interested		How can we handle an outcome variable with more than two categories? How does CART compare to a classifier based on logistic regression?
Ideas about teaching		• Students can either create their own validation sets or the teacher can create one for the entire class. Students should try growing a tree as completely as possible (with as many nodes as possible) and then compare the performance on the validation set. Is the performance better or worse than with a less complex tree (say the default tree)? It will likely be worse, and the reason is called "overfitting". • Overfitting occurs when the model tries to fit and predict what is truly random "noise" in nature. Overfitting is a danger when models are complex. In this context, "complex"

TOPIC :	3	Growing Classification Trees
		means many leaves. • If growing a full tree leads to poor performance on the validation set, students should consider the possibility of stopping the growth of the tree (or "pruning" back to an earlier stage of the tree). They can then use the validation set to see if performance is worse. (A pruned tree will likely still not do as well on validation as it did on the original data, but the decline in improvement should be less drastic than the decline experienced by the full tree.) Determining the "best" prune is more advanced than should be taught in this course. Here, we want students to understand (a) that the validation set can be used to evaluate the goodness of prediction and (b) overfitting is an issue to be concerned about.

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
Ideas about teaching		Students should discuss what they feel are the important features to communicate. Lead the conversation so that they understand that important questions to ask/answer in order to use a tree include (a) what's the probability our classification will be correct/incorrect? How confident are we in that value? (b) does the probability of misclassification differ for different types of input?

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 variables? Measuring quality of prediction.
	5.2	Interpreting regression trees. (Students see examples of trees)
	5.3	Building regression trees with one predictor variable.
	5.4	Building regression trees with more than one predictor variable.
	5.5	Comparing trees with a validation set.
Prior knowledge required		Topic Areas 1.2, 1.3. Topics 1 – 4 above.

TOPIC : 5	**Introduction to Regression Trees** *(in the context of a particular investigation targeting a particular real-world problem)*
Possible resources *(data, real-life problems, …)*	Classroom-collected data: Height, shoe size, resting pulse rate, armspan, index finger length, morning person or night owl? Any dataset used for multiple regression will work and make a useful comparison to previous units.
Ideas about teaching	• Subtopic 5.1 might have been covered in a unit on multiple regression. The purpose here is to help students understand how difficult it is to exactly predict a numerical outcome, and so the goal instead is to come as close as possible. This means we need to measure "close", and mean absolute difference (MAD), average of |prediction - actual outcome|, is a good place to start. (It's great if students work this out on their own, with some guidance from the instructor, of course.) Students should understand how a potential predictor variable (such as night owl vs. morning person) might be used to predict another variable (such as height) by seeing whether guesses can be made "closer" using "night owl" than if no information is given at all. It is unlikely that this particular binary variable will be a useful predictor, but it is helpful to begin with a binary predictor variable. Gender might be a more useful predictor variable if you are comfortable using it. (Note that when using MAD, the median is a good predictor to use.) Students should begin with a categorical predictor (such as "night owl") to understand how to measure the prediction strength, but should then use a numerical variable (such as shoe size) to understand how this variable can improve predictive strength when compared to using no predictor at all. • Subtopic 5.2: Students should be shown examples of regression trees; at this point they should not be concerned with how they are generated. The goal is for students to interpret trees by describing how they could be used to make predictions. They should learn to answer questions such as "Which variables do and do not help us in predicting?" "According to this tree, what information do I know about a person/object to make a prediction?" "What does it mean if more than one variable is involved in a prediction?" (It means that those variables interact.) • Subtopic 5.3: Students can "build" a simple tree using just one predictor variable, for example using shoe size to predict height. Begin by splitting shoe size at the median value (small and large shoes); how would you measure the success of this tree to predict height? How could it be improved (to make better predictions)? (For instance, consider additional splits, perhaps into quartiles.) Measure resting pulse rate. Can this value be used to predict height? Try it, and see what happens. • Subtopic 5.4: Students generate ideas for an algorithm that will use more than one predictor variable. For example, can they use resting pulse rate, armspan and index finger length to predict height? Encourage students to explore these data with graphs to understand the structure of the data before generating ideas for algorithms. The purpose isn't to stumble upon *the* algorithm, but to get them thinking about how software might implement this approach. Topics 5.1-5.4 should be helpful here. • Subtopic 5.5: Conceptually, this does not differ from Topics 1-4. However, this will reinforce the notion that that validation sets can be used to measure the success of prediction models and can help reinforce understanding of how the models work. Students should use a validation set (generated by the instructor is best but they might also create their own by randomly selecting rows from the data set), run observations

TOPIC : 5	**Introduction to Regression Trees** *(in the context of a particular investigation targeting a particular real-world problem)*
	from the validation set down their tree, and use the MAD or the sum of variances to measure the goodness of fit and compare different trees.

[*Back to Unit 2 contents page*]

Topic Area:	**2.5 Unsupervised Learning**
Version:	30 April 2019

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 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) variables, and does not make use of a response 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	Able to: ● Explain the concept of inputs (features/attributes/variables) and outputs (class labels) (review from Topic Area 2.4) ● Explain you need both to perform supervised learning ● Explain what happens when you do not have the output labels but only the inputs ● Explain when you need unsupervised learning, and how is it different from supervised learning ● Discuss a real-world example of clustering – Segmentation of customers based on features so that customized marketing emails can be sent to them ● Explain that unsupervised learning allows you to group your data (into clusters) with only feature/attribute inputs ● Explain concept of distance (Euclidean) between data points ● Describe concept of distance between points using a 2D plot containing 2 features	**Additional learning outcomes** (*Pedagogical*) ●

	• Explain how data points closer together are similar • Explain how points further apart are dissimilar • Explain concept of clustering based on distance as a metric • Explain one example of a clustering algorithm such as K-means • Explain how to use clustering to detect outliers • Explain motivation for outlier detection (*e.g.*, credit card fraud) • Explain that outlier detection doesn't use class labels.	
Key phrases:	Clustering, Distance (Euclidean), outlier detection, Segmentation, Unsupervised learning	
DS 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.
	1.4T	
Prior knowledge required		Topic Areas 1.2-1.3, 1.6
Ideas about teaching		Discuss an example used in the supervised learning module. Explain how the data were comprised of predictor variables and a response variable. It would be useful to explain that supervised learning was possible because the data was labeled. Introduce and motivate via a real-world example where unsupervised learning might be useful. For example, we have a large database of customers across multiple age groups and we want to decide which customers should receive information about colleges and which about retirement. If we were to cluster by age, we could target the information appropriately.

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.

TOPIC :	2	Getting and exploring the data
	2.2	Explain the data set. Explain the orientation of the data: the rows represent the observations while the columns represent the features or attributes.
	2.3	Explain how the class label is missing in the presented data compared to Unit 2.4.
	2.4	Motivate how we can use mathematical techniques to automate the discovery of clusters in datasets.
Prior knowledge		Topic 1 above
Ideas about teaching		It is helpful to begin with a very simple data set in which students can apply their own knowledge to cluster observations. The data set could have a small number of observations, and, to begin with, only 2 dimensions. With this, the data points can be represented as a 2D scatter plot. For example, given data on physical characteristics (Height and Weight) of players, can students visually divide them into multiple clusters? The clusters represent the sport the players play. For now, having 2 clusters to divide into is a good idea. The data set should be simple enough that they can visually cluster them based on how close or far apart the points are. A synergistic approach would be to use the same data set in Unit 2.4 for supervised learning but remove the class labels for the exercise here. The question could then be posed: Can you guess what sport the clusters indicate, based solely on how the points cluster together? Explain that since labels are not available, they are making conclusions on each cluster solely based on the values of the variables they plotted. Introduce the idea that the observations or data are not always perfect. To expand on this, add another data point midway between the clusters. This would introduce some ambiguity about which cluster the new point would belong to. Explain that it is possible the data point could end up in the wrong cluster. Next, explain that the goal is to come up with a (mathematical) technique to find the clusters automatically when you have thousands and thousands of data points. One such technique is called K-means clustering, which is introduced and described in the next section.

2.5 Unsupervised Learning

TOPIC :	3	Example of Unsupervised learning algorithm: K-means clustering
Subtopic:	**3.1**	Explain that 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	Explain that 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

TOPIC :	3	Example of Unsupervised learning algorithm: K-means clustering
		(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	After all the points have been assigned a cluster, explain the iterative nature of the algorithm, determine the updated cluster centers, and repeat Step 3.2.
	3.4	Explain the need to repeat with different initial guesses for cluster centers.
	3.5	Use a simple example containing a small set of data points to explain the K-means algorithm manually.
	3.6	Explain that a (distance-based) outlier is an object that is "surprisingly", or noticeably, far from its nearest neighbors.
	3.7	Explain it is also an object that will typically be far from the middle of its assigned cluster (cluster center)
Prior knowledge required		Topic 2 above
Ideas about teaching		**Use of Flow charts:** To explain the flow of the K-means algorithm, it might be useful to develop a flow chart with the individual steps and progressively uncover it while walking through the algorithm. **Use of Examples:** Depending on the interaction, it might be best to explain the K-means with visual examples. The following sequence of steps could be used to explain the algorithm manually: 1. Obtain a data set that contains a small number of samples. Ideally, 6-7 observations and 2 variables (features). For simplicity, identify a data set that contains only 2 clusters. 2. Plot the data points on a scatter plot (2D works well since we have only 2 variables) 3. Depending on the data set, the clusters might be visible readily. In that case, have the students mark out the 2 clusters. 4. Now, walk through the steps of K-means algorithm. 5. Initialize 2 arbitrary cluster centers. For simplicity, let one of the data points itself be a cluster center. 6. Show the distance calculation for a few points, and based on the values, assign the point to a cluster. 7. Finally, illustrate the cluster assignment using colored points (red for cluster 1, blue for cluster 2) on the 2D plot 8. Compute the horizontal and vertical coordinate means for the cluster centers. Repeat the steps above. 9. Demonstrate that the algorithm is final when the cluster assignments do not change from one iteration to the next.

2.5 Unsupervised Learning

TOPIC :	4	Implementing K-means clustering on a large data set
Subtopic:	**4.1**	Introduction to format of the data set.

TOPIC :	4	**Implementing K-means clustering on a large data set**
	4.2	Introduction to the programming environment to use for clustering
	4.3	Explain the need to clean and transform the data set so that it is arranged as observations (rows) versus features (columns)
	4.4	Choose K and run the algorithm using the code snippet provided
	4.5	Review 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	Change the value for K, and repeat. Compare how the cluster selections have changed.
	4.7T	Pedagogical issues relating to these topics
Prior knowledge required		Topic 3 above
Ideas about teaching		• A sample data set and the code to perform the clustering is provided in the Teaching Appendix below. • It would be helpful for the instructor to walk through the code a section at a time. • Before performing K-means clustering, it is sometimes helpful to normalize the data for each dimension to a known range (*e.g.*, from -1 to 1). This would ensure that no one dimension dominates the ability to separate clusters after K-means is applied. Also, it might be helpful to apply dimension reduction because the points might be sparse in higher-dimensional data making distance measures less meaningful. By reducing the dimensionality, the points would cluster better in the limited set of dimensions, so making clustering more effective.

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**	Explain why Unsupervised learning is the best approach for the problem at hand. Make a link to the fact that labeled data requires human effort, and is hard to obtain. Data from measurements and sensors typically lend it readily to unsupervised learning. Give examples of business problems where unsupervised learning is commonly used.
	5.2	Select exploratory analysis and graphs to summarize the input data.
	5.3	Select the visualizations to show for different values of K, and how the choices were made.
	5.4	Demonstrate how an optimal choice of K could be made, and emphasize that result. For this, compute the "elbow distance metric". This represents the mean distance between cluster points and centroid.
	5.5	Provide descriptive statistics about the features in each cluster. Demonstrate how they are similar within a cluster compared to across clusters. Based on the characteristic of these features, describe how clusters differ from each other.
	5.6	Experience interpreting and communicating the results of clustering-algorithm output for several different problems

TOPIC :	5	**Use in Problem solving** *(in the context of a particular investigation targeting a particular real-world problem)*
	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 features of each cluster are or how the clusters are explained.
Prior knowledge required		Previous Topics above.
Ideas about teaching		Students should feel comfortable contrasting the motivation for supervised learning with that for unsupervised learning. They should be able to articulate the value of unsupervised learning methods and their applicability.

2.5 Unsupervised Learning

TOPIC :	6	**Other unsupervised learning methods – Alternatives to K-means clustering**
Subtopic:	**6.1**	Describe instances when distance-based (that is, K-means clustering) may not be well suited
	6.2	Motivate need for other clustering methods (emphasize disadvantages of K-means) – no need to pre-determine number of clusters, lack of consistency
	6.3	Introduce 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
	5.5T	
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.

- **See Example Case Study: Applying K-Means Clustering to Delivery Fleet Data** (*Example and Code*)
-

[*Back to Unit 2 contents page*]

Topic Area:	**2.6 Recommender Systems**
Version:	30 April 2019

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	Able to: • Explain what a Recommender system is and give examples of their use. • Recognize that recommendations are based on data, and identify sources for that data. • Recognize ethical issues associated to Recommender systems. • Describe several computational approaches to generate recommendations, and explain their relative strengths and limitations. • Operate with measures of similarity. • Use tools to compute recommendations in several ways.	**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	

DS 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 features 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-personalised "best-seller" list as a baseline to compare with.
	1.3	Ethical issues for Recommendation and personalised 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 vs data privacy, reinforcement effects from recommendations vs 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.
	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.

TOPIC :	2	The data used by Recommender systems
	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
	3.6T	Pedagogical issues relating to similarity and clustering • 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 :	4	Collaborative-filtering
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	Collaborative-filtering
	4.2	Define similarity of users: similar demographic features, vs "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.

[*Back to Unit 2 contents page*]

Topic Area:	**2.7 Interactive Visualization**
Version:	30 April 2019

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
	  ○ 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.
		
		Teachers should be able to clearly differentiate static information graphics from interactive visualizations and describe the specific benefits of interaction.

Learning outcomes	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 variables	
DS 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 w.r.t. 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

TOPIC :	**1**	**Why visualization?** **The role of Visualization in the Data Science Learning cycle**
	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…): "an 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 variables. • Briefly discuss the fact that there are not an infinite set of visual variables one can use: illustrate with examples use of many visual data variables resulting in an ineffective visualization. • Introduce the idea that some visual 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: 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)
Possible resources *(data, real-life problems, …)*		● Examples from the resources listed in subsequent topics ● Personal visualizations: dear data; ● Ambient visualization examples include simple visualization to monitor energy consumption: Ambient orb, Flower lamp, power cord on this page; nabaztag
Ideas about teaching		This Topic is **introductory**. Many of the concepts touched on here will be developed in greater depth in subsequent Topics below. The books of William S Cleveland (*The Elements of Graphing Data* and *Visualizing Data*) are important references.

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, preattentive attention
	2.2	Hierarchy of visual variables, *i.e.*, how to appropriately choose a 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
Possible resources *(data, real-life problems, …)*		2.1: Diagram of Bertin's visual variables 2.3: Preattentive processing in the infovis-wiki 2.3: Colorbrewer website; Color in different cultures; Stroop effect 2.4: Motion to support view transitions: Google Material design. Visual transitions 2.4: Motion to convey emotion: pixar lamp, Perlin's polly, Heider-Simmel demonstration.
Other things to look at for those interested		For the instructor: 2.1: Gestalt Laws of perceptual organization 2.1: Colin Ware's visual thinking for design. 2.1-2.2: Discussion of Bertin's visual variables for design and their appropriateness for encoding different types of data. 2.2: Empirical evidence on the differing power of visual encodings is included in the work of William S Cleveland (see books *The Elements of Graphing Data* and *Visualizing Data*) as well as more contemporary scholarly work published in the IEEE VIS Conference such as the work of Kim and Heer. 2.4: Color name models (academic article), Affective color in visualization (academic article)
Ideas About Teaching		The Stroop effect experiment can be fun in class.

2.7 Interactive Visualization

TOPIC :	3	Interaction
Subtopic:	**3.1**	The role of interaction: to provide more detail, declutter views, reveal correlations.

TOPIC :	3	Interaction
	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: • 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.
Possible resources *(data, real-life problems, …)*		The Visgets demo shows filtering and brushing-and-linking well: https://mariandoerk.de/visgets/demo/ This academic article by Yi and Stasko enumerates the basic interactions from a user point of view. Demonstrations with Tableau allow interaction, as do many demos in the d3.js gallery https://github.com/d3/d3/wiki/Gallery Examples of interactive visualizations to explore: • NY Upshot (may require subscription): https://www.nytimes.com/section/upshot • Data Sketches: http://www.datasketch.es/ • The Pudding data essays: https://pudding.cool/
Other things to look at for those interested:		• Recommender systems (Topic Area 2.6) to recommend interactions • Considerations of interaction design: suggested interactivity (see this article), animated transitions to support orientation during visual changes (see Google Material design: section on transition anatomy), consistency and usability

Ideas about teaching		• It is important to illustrate interaction techniques and their merits using existing examples of interactive visualization used for communication, e.g. from NY Times repository or other online examples. Demonstrate storytelling and progressive disclosure through interaction ("scrollytelling"). These examples from real life will resonate with the students.
		• For each of the techniques, try to have an illustrative example for
		  ○ Exploratory Data Analysis, and
		  ○ Visual communication and presentation.
		• Some techniques are not suited for exploratory data analysis (e.g., progressive disclosure and callouts)

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 which affect visualization efficacy: • Perception of distances and areas: Choropleth maps versus tiled maps, infographics where an element is scaled (*How to Lie with Statistics* [book]) • 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: ● *e.g.*, 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: Unit 1.5, and Topic 2 above.		
Possible resources *(data, real-life problems, …)*		● Several examples of barcharts with bars pointing down can be found on the EagerEye blog ● Edward Tufte's "The Visual Display of Quantitative Information" discusses choropleth maps (Part 1) ● Edward Tufte's "The Visual Display of Quantitative Information" discusses chart junk (Part 5) ● A host of examples on the Misleading graphs wikipedia page
Other things to look at for those interested		● *How to Lie with Statistics* [book] (accessible to younger audience) ● Many examples deconstructed and critiqued on Kaiser Fung's blog ● Exploration of the effects of spatial aggregation. ● Change blindness (and importance of maintaining a smooth visual flow) ● Empirical evidence about the power of visual encodings, collected by Cleveland and others (see Topic 2) can be used to inform critique.
Ideas about teaching		● Ask students to come to class with a visualization example they find in the news or on the web, and critique them in a collective activity. (Make sure to also plan for examples that best illustrate important concepts to learn) ● Critiques should explore both positive and negative examples. ● This topic is heavily reliant on concepts taught in Topic 2. This topic should focus on reinforcing elements of Topic 2, while demonstrating the caveats associated with human perceptual and cognitive limitations/biases. ● A fun in-class activity is to show the change blindness examples of Rensink.

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.

Topic Area:	**2.8 Confidence intervals and the bootstrap**
Version:	30 April 2019

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: • distinguish between an unknown parameter and an estimate of that parameter • explain 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	**Additional learning outcomes** (*Pedagogical*) Facility with computer simulation to answer questions involving random phenomena.

	• 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 • obtain a bootstrap confidence interval from data in a range of situations • explain 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) • explain what types of error and uncertainty confidence intervals do and do not address	
Key phrases:	Sampling error, Standard error of estimate, Bootstrap resampling, Confidence interval, Coverage, Simulation.	
DS 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).

TOPIC :	2	Sampling error
	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; (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

[*Back to Unit 2 contents page*]

Topic Area:	**2.9 Randomization tests and Significance testing**
Version:	30 April 2019

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

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 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*)	**Additional learning outcomes** (*Pedagogical*) Facility with computer simulation to answer questions involving random phenomena.

	• 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 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.	
DS 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 labelling) but no actual treatments are performed.

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 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	(*Optional*) **Investigating the performance of randomization tests in a sampling context**
Subtopic:	**4.1**	• Taking a large data set to be used as a population to sample from (i) construct a modified version in which the means/proportions of the subpopulations to be compared are the same (ii) construct a modified version in which the means/proportions of the subpopulations to be compared differ by a specified amount

TOPIC :	**4**	(*Optional*) **Investigating the performance of randomization tests in a sampling context**
		Using simulation, investigating the behavior of a randomization test in situations (i) and (ii) 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: random allocate group labels to a dataset and add a "true" treatment diff, 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

[*Back to Unit 2 contents page*]

Topic Area:	**2.10 Image data**
Version:	30 April 2019

Commentary:

Images might on the face of it seem to be about as far from data for analysis as you could get but in fact data extracted from images plays an important role in object recognition, classification and transformation for images.

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 images.
	• To understand 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, 2.5
	• 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
		When classifying any objects, people, or locations, there are ethical issues that arise from the possibility of misclassifying the objects.
		
		This uncertainty should always be considered since errors of misclassification can have serious consequences. E.g. in war mistaking a friendly airplane for an enemy airplane.

Learning outcomes	Able to: • Understand misclassification errors, false positives and false negatives, using images as objects to classify • Understand how to transform images to quantities that can be analyzed • Understand how images can represent a label of an object, which has potentially many features: a face image, a fingerprint, a shoeprint, etc.	**Additional learning outcomes** (*Pedagogical*) • Convey to the student how images are ubiquitous and represent objects that can be classified and analyzed • Demonstrate the high dimensionality that each image contains
Key phrases:	Classification, Compression, Extraction, Image processing, Pattern recognition, Pictures, Supervised learning, Unsupervised learning	
DS Learning Cycle elements	*All steps*	
Prior knowledge required	Topic Areas 1.1-1.4, 1.6, 2.4, 2.5	

2.10 Image data

TOPIC :	1	What is Image Data and what sorts of problems does it pose?
Subtopic:	**1.1**	Images have become more prevalent as a source of data, statisticians and data scientists need to analyze and interpret them. Sophisticated and complex tools are available to handle the many challenges of images, from compression and extraction, enhancing, sharpening, and blurring.
	1.2	Here we consider images as a set of data and introduce methods for representing them in order to classify them, either supervised or unsupervised. We also consider testing hypotheses, e.g., are two images (photographs) actually of the same person or object. We also consider questions of identity, from a set of images of artwork.
Prior knowledge required		

2.10 Image data

TOPIC :	2	Some areas for application
Subtopic:	**2.1**	From the world of Art: a) Is a painting of unknown origin representative of a known school of art? b) Is a painting alleged to be by a particular artist authentic or a reproduction? The focus of this lesson is to introduce students to the challenges encountered with image data. • Use a small set of say 12 images, four from each of three historical periods: e.g., impressionism, realism, and cubism. Have students classify them, and think about how they might characterize them. • Then provide quantitative data on these images, e.g., RGB, or Hue, Saturation, Intensity, other features or characterizations derived from the images (See Jia Li's slides at http://personal.psu.edu/jol2/Li_lecture_highschool.pdf).
	2.2	From Forensic Science: Images of fingerprints, shoeprints, and bullet markings. a) Can fingerprints be correctly classified as "matching" one in a data base? b) Can partial prints (smudged or incomplete) be reliably identified as matching an existing print, or determined to be too incomplete for a reliable match. c) Determine a probabilistic criterion via an experimental simulation.
	2.3T	Framing the challenge: a) See recent controversy about the identity of the president of Nigeria, and show multiple images to motivate the search for good criteria and tools for facial recognition: http://en.rfi.fr/africa/20181204-clone-focus-africa-buhari-and-his-clone b) Use of facial images for identity: Delta airlines Dec 2018 tests the concept See these stories: https://news.delta.com/delta-tests-facial-recognition-boarding-tops-year-innovation-atl

TOPIC :	2	Some areas for application
		https://www.washingtonpost.com/nation/2018/12/04/your-face-is-your-boarding-pass-this-airport/?noredirect=on&utm_term=.0f74315edfaf Discuss the issues of false positives and false negatives in the context to motivate need for reliable classifications.
Prior knowledge required		Topic Areas 2.4 & 2.5

[*Back to Unit 2 contents page*]

Teaching Appendices: Example Case Studies

Please note that the level of coding illustrated in the Case Studies that follow can be reduced, if so desired, by using higher-level function calls, and most of the work can also be done using gui-driven software.

A1: Time Series

An Example Case Study: Canadian Temperature

Wesley Burr & Chris Wild

August 22, 2018

Case Study: Context and Problem Statement

Global warming (or "climate change") is a topic in the news regularly in recent years, as the world has seen more extreme weather events. Whether you live in Canada (where winters appear to be milder and summers hotter) or Australia (where drought and wildfires are an annual occurrence) or anywhere else on our planet, the **climate** of our environment appears to be changing. NASA's Earth Observatory program says "The world is getting warmer. Whether the cause is human activity or natural variability - and the preponderance of evidence says it's humans - thermometer readings all around the world have risen steadily since the beginning of the Industrial Revolution ... the average global temperature on Earth has increased by about 0.8 ° Celsius (1.4 ° Fahrenheit) since 1880. Two-thirds of the warming has occurred since 1975, at a rate of roughly 0.15-0.20 ° C per decade."

We are curious about this statement, and how it relates to our local environment. For the purposes of this case study, we will treat "local" to mean Ontario, Canada (as one of the authors lives there!). Our objective is to answer the following two questions:

1. Can we observe a warming **trend** in the data from Toronto, Ontario, Canada, in similar fashion to that stated by NASA in the above statement?
2. What **prediction** would we make for the temperature in the near future (say, the next two years - for our data, this would be 2013 and 2014)?

We will explore some temperature data from Environment & Climate Change Canada (shortened to ECCC, a federal agency in Canada) which are **time series**: repeated observations of the same unit or measurement over **time**. ECCC maintains hourly observations of temperature at many locations across Canada which are freely available on a data portal at http://climate.weather.gc.ca/. Similar databases exist for other countries such as Australia (http://www.bom.gov.au/climate/data/), New Zealand (https://cliflo.niwa.co.nz/), the United States of America (https://www.ncdc.noaa.gov/cag/national/time-series) and the United Kingdom (https://www.metoffice.gov.uk/public/weather/climate-historic/), as well as most other countries on Earth.

Setup for the Analysis

We'll load all the packages at the start and filter the warnings out to make this case study a bit prettier.

```
library("readxl")
library("dplyr")
```

```
# library("stringr")
library("lubridate")
library("ggplot2")
library("tidyr")
library("rclimateca")
library("iNZightTS")  # install from GitHub, devtools::install_github("iNZightVIT/iNZightTS")
```

Getting the Data

We will start by obtaining a database of temperature data for Toronto, Ontario, Canada. Toronto is a large city in the province of Ontario, covering much of the northern shore of Lake Ontario, one of the five Great Lakes. It is known for warm, humid summers and cold-to-mild, humid winters, due to the lake effect. For this case study, we'll use the historical temperature records from Toronto Lester B. Pearson International Airport, station number 5097, as we know that it has a contiguous record for much of the recent past. We use the **rclimateca** package for this, and will grab the monthly averages - that is, the average temperature measurement for a given month, averaging the hours of the month all together.

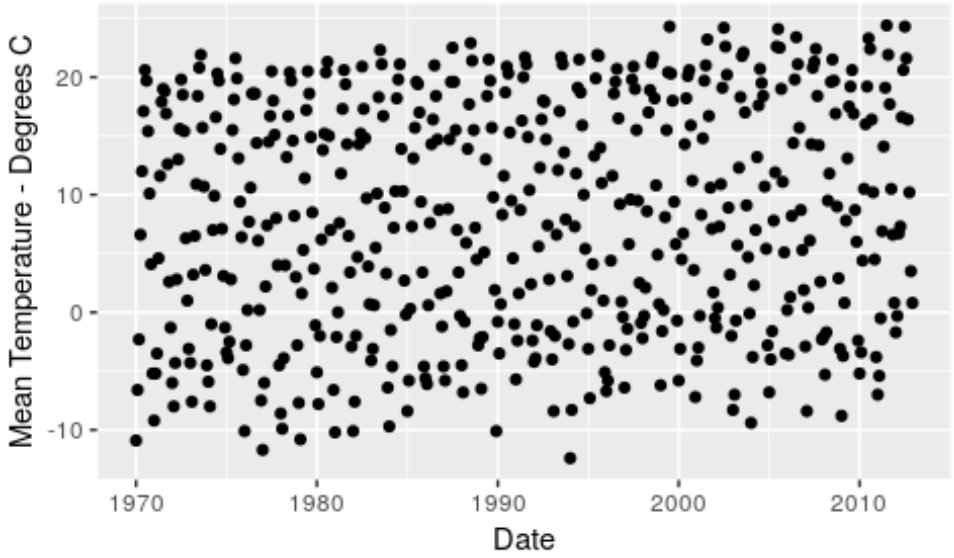
```
station <-  ec_climate_search_locations(5097)
pearson <- ec_climate_data(as.character(station),
                timeframe = "monthly", start = "1970-01-01", end = "2012-12-31")
str(pearson %>% select(contains("temp")))
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    516 obs. of  10 variables:
##  $ mean_max_temp_c   : num  -5.9 -1.6 1.9 12.3 18.1 23.9 26.6 26.1 20.6 14.8 ...
##  $ mean_max_temp_flag: chr  NA NA NA NA ...
##  $ mean_min_temp_c   : num  -15.8 -11.6 -6.4 0.9 5.9 10.3 14.6 13.2 10.2 5.4 ...
##  $ mean_min_temp_flag: chr  NA NA NA NA ...
##  $ mean_temp_c       : num  -10.9 -6.6 -2.3 6.6 12 17.1 20.6 19.7 15.4 10.1 ...
##  $ mean_temp_flag    : chr  NA NA NA NA ...
##  $ extr_max_temp_c   : num  7.2 5.6 8.9 23.9 27.8 31.1 31.7 32.2 29.4 23.9 ...
##  $ extr_max_temp_flag: chr  NA NA NA NA ...
##  $ extr_min_temp_c   : num  -27.8 -23.9 -13.9 -7.2 -3.3 4.4 8.9 7.8 3.3 -3.3 ...
##  $ extr_min_temp_flag: chr  NA NA NA NA ...
##  - attr(*, "flag_info")=Classes 'tbl_df', 'tbl' and 'data.frame':    7 obs. of
2 variables:
##   ..$ flag       : chr  "B" "E" "M" "S" ...
##   ..$ description: chr  "More than one occurrence and estimated" "Estimated" "Mi
ssing" "More than one occurrence" ...
```

Now we have 516 observations loaded, consisting of monthly metrics of ground-level temperature, as well as rain, precipitation, snow, and wind, and we are ready to explore the temperature data. The time span is what we requested: 1970 (January 1) through to 2012 (December 31). Let's subset down to the temperature variables and clean up the data set: in this analysis, we aren't interested in wind, snow, rain, or any of the other data available, since we're just trying to answer the question about temperature trends.

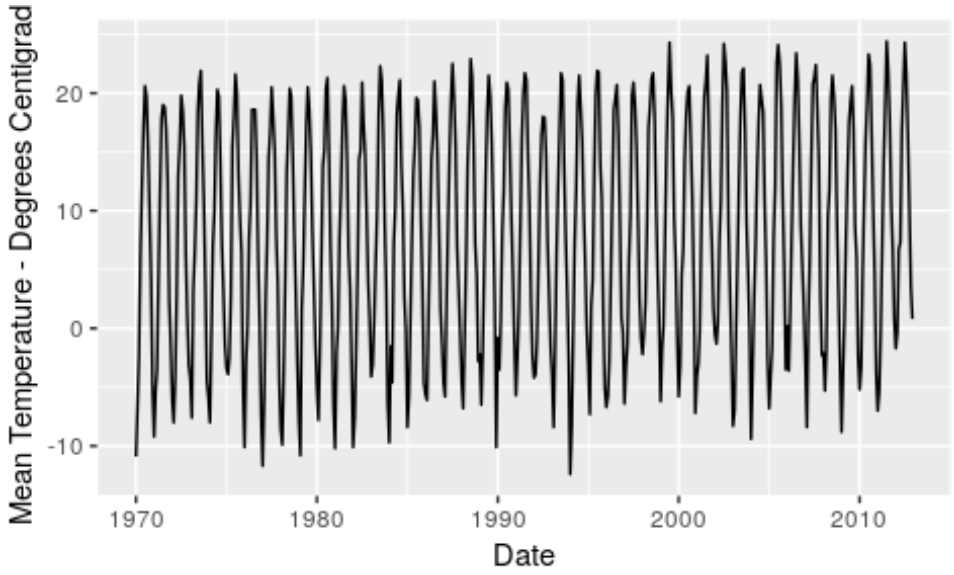
```
pearson_clean <- pearson %>% select(year, month, date, mean_max_temp_c, mean_min_te
mp_c,
                                    mean_temp_c)
ggplot(data = pearson_clean, aes(x = date, y = mean_temp_c)) +
```

```
        geom_point() + ylab("Mean Temperature - Degrees C") +
        xlab("Date")
```



This plot looks a bit crazy! What are we supposed to think about all these scattered points? Looking at the plot we remember that the x-axis is actually a Date! That is, this is a **time series**: the data are organized in sequential order by the date of observation. So points aren't really the best way to represent this data.

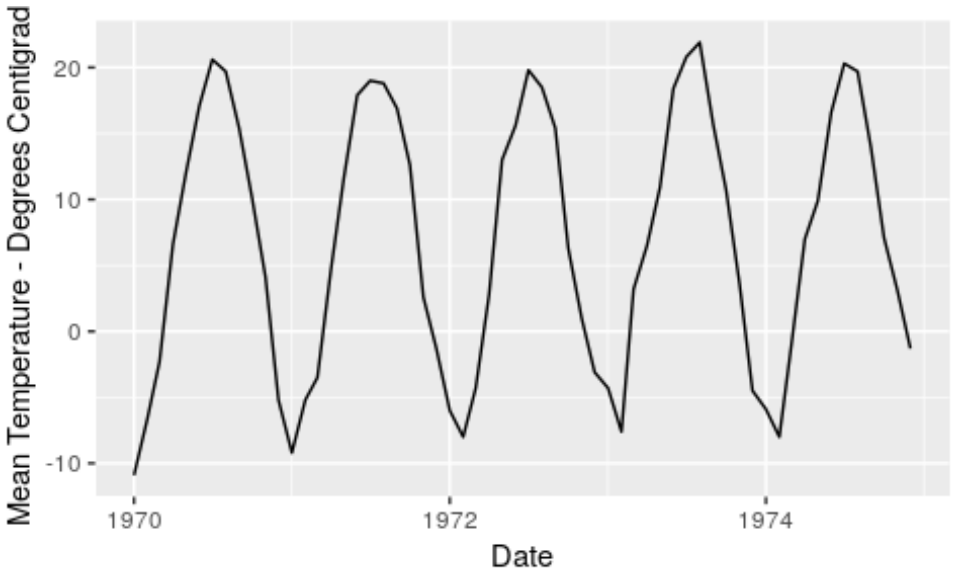
```
ggplot(data = pearson_clean, aes(x = date, y = mean_temp_c)) +
        geom_line() + ylab("Mean Temperature - Degrees Centigrade") +
        xlab("Date")
```



That looks more interesting, doesn't it? What's going on with the oscillations? Zoom in a bit on just a short span of time to see if we can figure it out:

```
pearson_clean_zoom <- pearson_clean %>% filter(year < 1975)
ggplot(data = pearson_clean_zoom, aes(x = date, y = mean_temp_c)) +
```

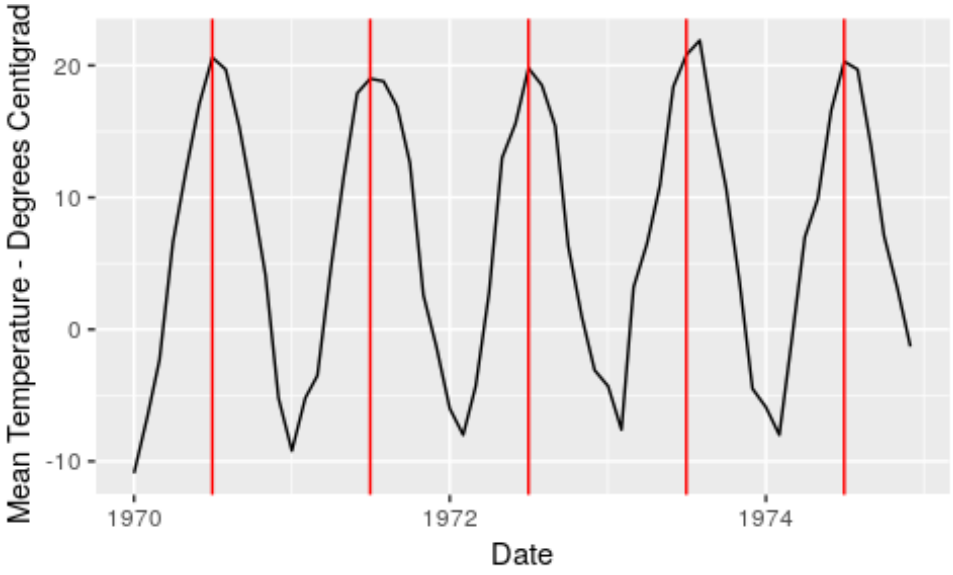
```
       geom_line() + ylab("Mean Temperature - Degrees Centigrade") +
       xlab("Date")
```



That's much more informative, isn't it? We see that there are multiple observations per year (12, actually: this data is **monthly**, so one observation per month), and, as everyone knows: North America is warm in the summer (middle of the year) and cold in the winter (end and start of the year). So we're seeing the average temperatures climb dramatically from December, January, and February to June, July and August. Formally, these oscillatory patterns are known as **seasonal swings** (also called "seasonal patterns" or "seasonality").

Let's identify the July months on the plot.

```
pearson_clean_zoom <- pearson_clean %>% filter(year < 1975)
ggplot(data = pearson_clean_zoom, aes(x = date, y = mean_temp_c)) +
       geom_line() + ylab("Mean Temperature - Degrees Centigrade") +
       xlab("Date") + geom_vline(xintercept = ymd(paste0(1970:1975, "-07-01")), col
our = "red")
```

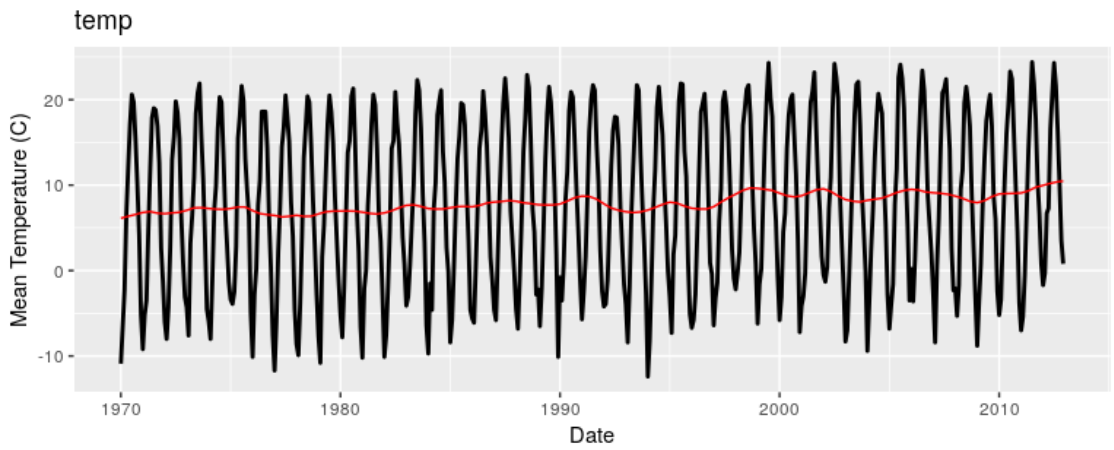


This plot allows us to discover something: graphically, the warmest month of the year in Toronto, Ontario seems to be July for most years, except for years (like 1973) where it's August. That is, the peak of Canadian (North American) summer! At least, this is the pattern for the early 1970s - remember that it's dangerous to take an inference like this and generalize it!

Analyzing Seasonality with iNZight

Our next step is to try to analyze the data so we can answer our initial questions. Let's look more closely at the seasonal pattern we saw in the previous stage.

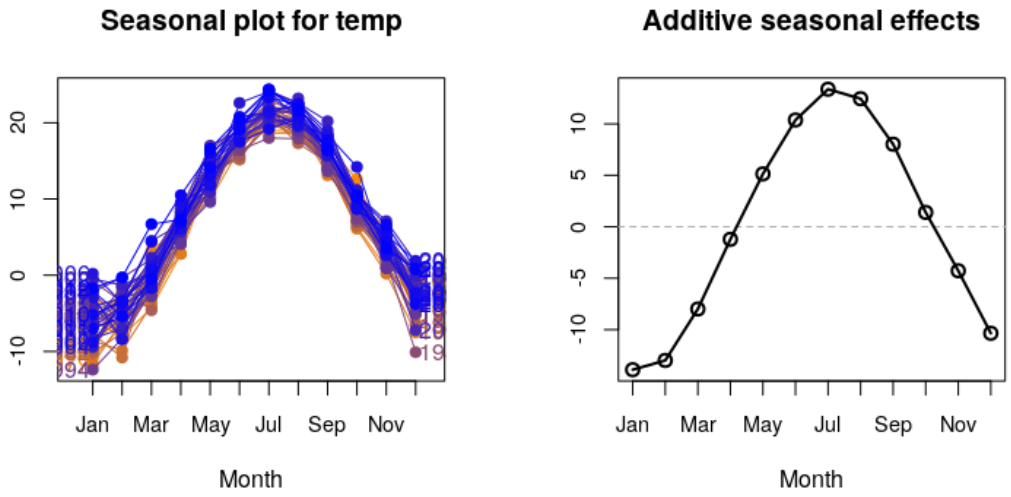
```
Month <- pearson_clean$month
Month[Month < 10] <- paste0("0", Month[Month < 10])
pearson_clean$month <- Month
pearson_df <- pearson_clean %>% mutate(time = paste0(year, "M", month)) %>%
                select(time, mean_temp_c) %>%
                rename(temp = mean_temp_c)
# iNZsightTS doesn't recognize tibbles
pearson_df <- as.data.frame(pearson_df)
# for linear regression at the end
pearson_df <- data.frame(pearson_df, x = 1:dim(pearson_df)[1])

pearson_iNZ = iNZightTS(pearson_df, start = c(1970, 1),
                        end = c(2012, 12), var = "temp")

# Times Series Plots
plot(pearson_iNZ, t=60, ylab="Mean Temperature (C)")
```



The latest plot has a **trend line** (smoother) overplotted, showing the behaviour of the series in average. This can be a difficulty with seasonal data, especially when the seasonal effect is strong like it is here - it can be difficult to tell what is happening with the overall trend, since the oscillations are so strong. Is there an increasing average? Remember that the NASA value for the surface of the Earth suggests the warming trend is around 0.15 °C per decade, so from 1970 to 2012, we'd expect to see an increase in the average of roughly 0.6 °C, if Toronto behaves like the rest of the land on Earth (which it won't, exactly).

An additive seasonal plot shows the yearly cycle plotted over and over as we moved forward in time.

```
splot <- seasonplot(pearson_iNZ)
```

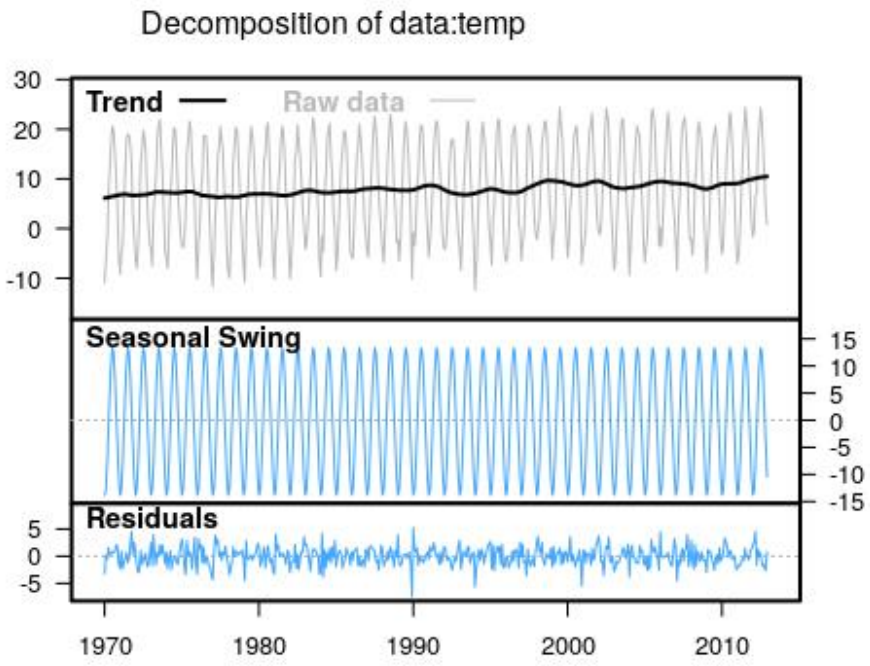


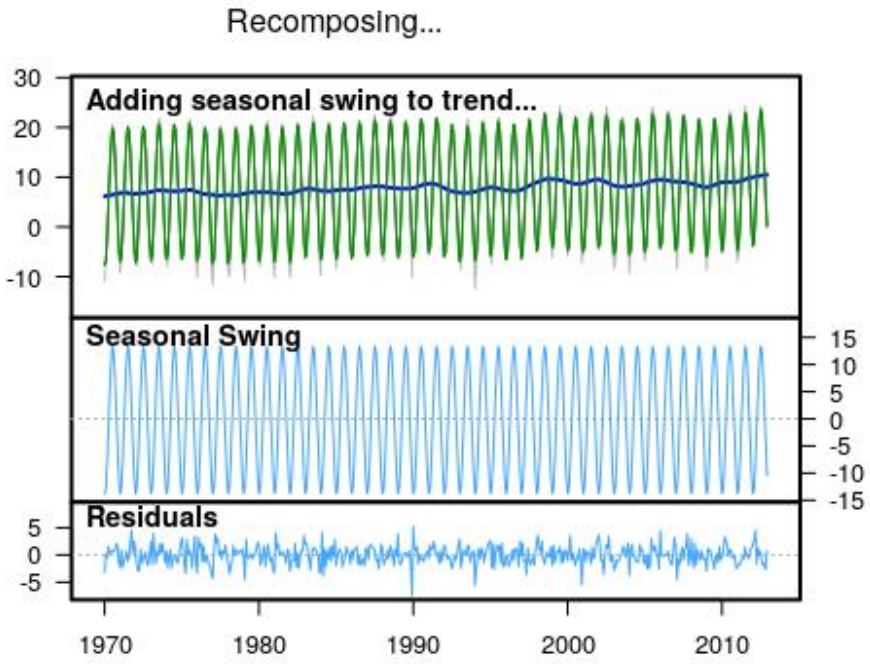
The right-hand panel on this plot shows the **Additive seasonal effects**, which is the average across all of the years of the change in temperature from January to July and then through to December. So for a given year, we expect the average temperature in July to be about 12 °C warmer than what we'd expect from the trend, and the temperature in December to be about 10 °C colder.

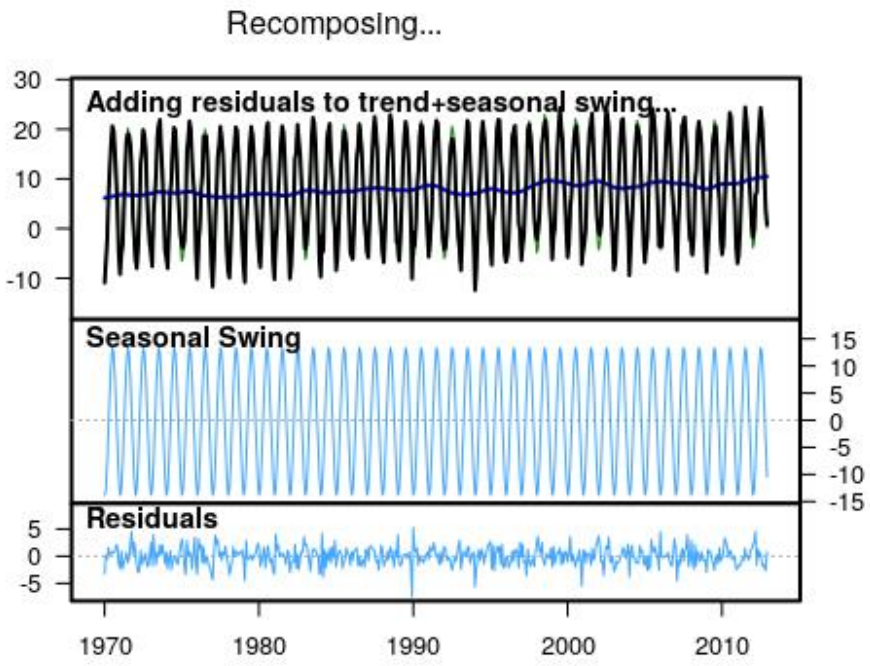
Let's look a little closer with a decomposition and recomposition.

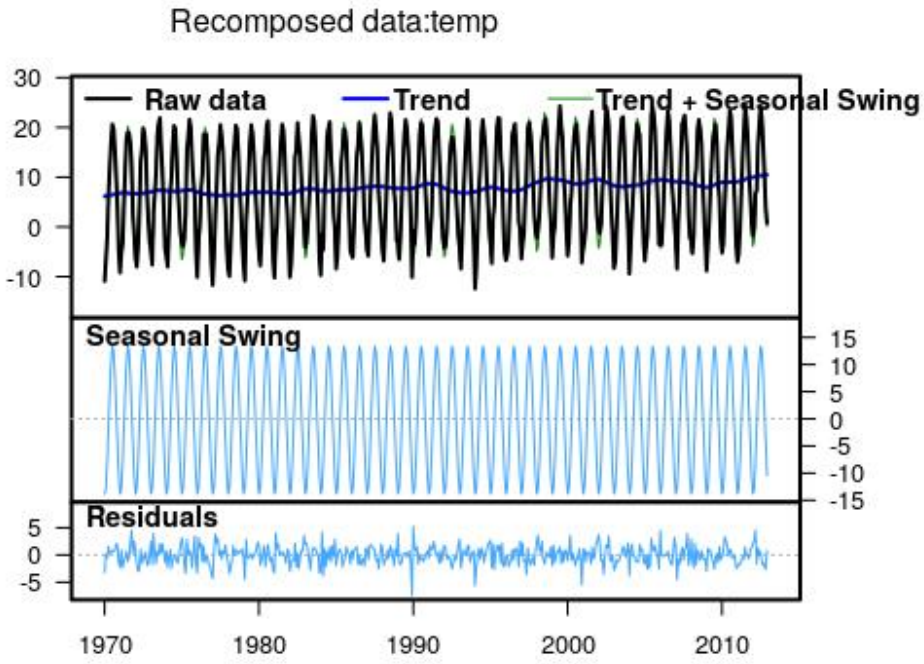
From the first figure below, we see the decomposition of the temperature data. The top panel of this figure shows the raw data with a smoothed trend overlaid. It looks fairly flat, with perhaps a slight upward tilt over time (remember the scale of the plot: 0.15 °C is not very large). The seasonality is in the second panel, and shows very regular behaviour: approximately the same seasonal cycle happens every year for our data set. And finally, there is a residual panel at the bottom, showing the "leftovers" (residuals) from the previous two stages. We can recompose these in stages so that it is apparent how much of each is represented.

```
recomp <- recompose(decompositionplot(pearson_iNZ, t = 60), animate=FALSE)
```


Decomposition of data:temp
Trend
Raw data
Seasonal Swing
Residuals
30
20
10
0
-10
15
10
5
0
-5
-10
-15
5
0
-5
1970
1980
1990
2000
2010


Recomposing...
Adding seasonal swing to trend...
Seasonal Swing
Residuals
30
20
10
0
-10
15
10
5
0
-5
-10
-15
5
0
-5
1970
1980
1990
2000
2010



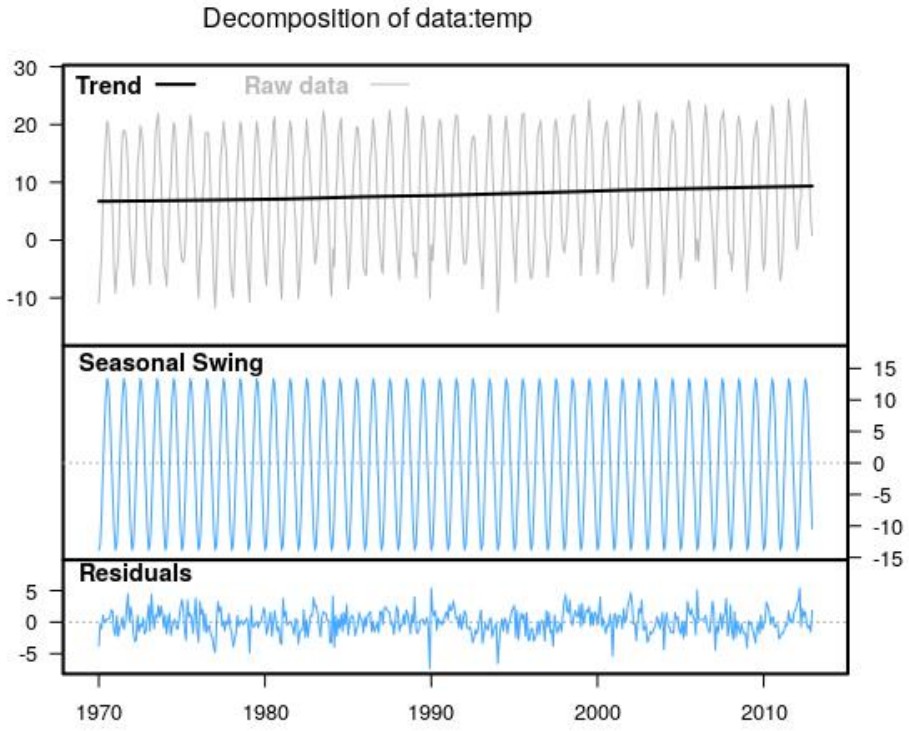


In the second and third figure, we see that the trend and seasonal swing add together to duplicate almost all of the variation in the original temperature. The residuals are small comparatively, and appear to be very noisy: in other words, a **random** component. When we try to use time series models to **predict** the future, randomness is good: it lets us gloss over that portion of the series and focus on the parts which are predictable.

Answering Question 1: Trend Lines

Let's return to the trend line discussion earlier.

```
decompose <- decompositionplot(pearson_iNZ, t = 516)
```



In this decomposition, we specified $t = 516$: that is, a smoother which has the maximum possible smoothness for this data - a straight line! This cleans up all the oscillations to the point that we can see that there is a very small slope to the line. From 1970 to 2012 there was a small, but noticable, increase in average temperature in Toronto. We can actually check what the slope is on this line using **linear regression** (fitting linear trend lines), something we talked about in a previous unit.

```
summary(lm(temp ~ x, data = pearson_df))$coef[2, 1:2]
```

```
##    Estimate  Std. Error
## 0.006128419 0.002868623
```

So according to this simple model, the trend line we are seeing in the previous plot has a slope of 0.00613°C per year, or 0.613°C per century. So across 100 years we would expect to see average temperatures rise 0.6°C. This seems really, really small: the difference between 19 °C and 19.6°C is small enough to not feel, right? Unfortunately, scientists explain that when the **average** temperatures go up, even by small amounts, there are effects that are much stronger than we would expect: oceans change composition, animals lose habitats, natural breeding cycles change for fish and insects, and (directly applicable to us humans!) extreme events happen more often. More wildfires, more heat waves, more cyclones, more hurricanes, and so on.

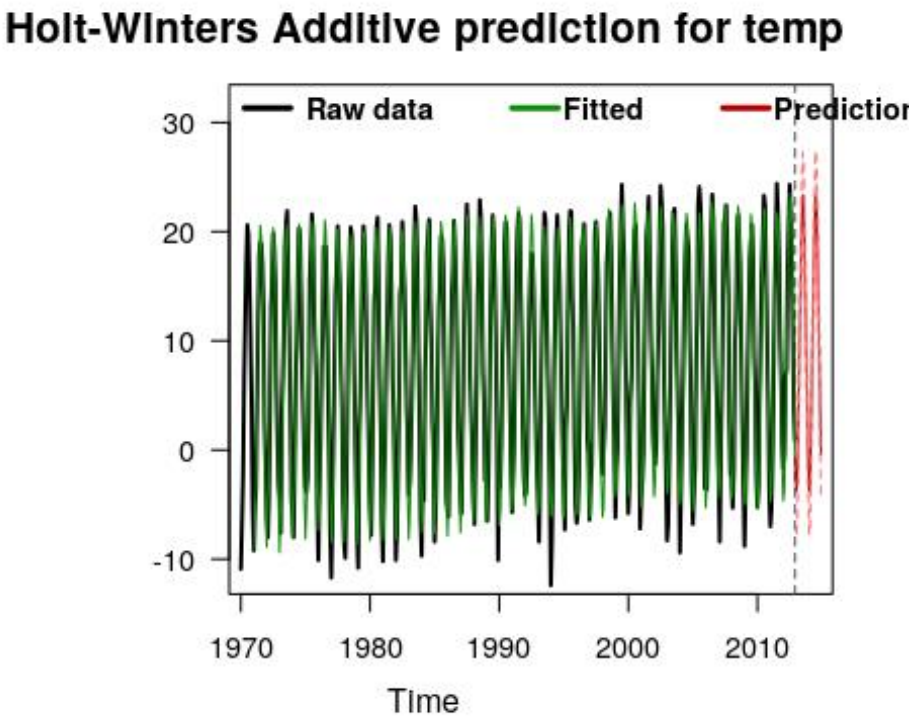
Compared to the Earth land temperature, this slope is smaller: NASA says the slope is 0.15 to 0.20 °C per decade, and we have an estimate of 0.06 °C per decade. Why is this? There are a

number of things it could be: NASA's methodology is **not** simply averaging individual stations. Specific stations sometimes have local heat effects ("heat islands") which can modify their temperatures. Canada may not behave exactly like the average across Earth. However, while the results from Toronto, Ontario at station 5097 do not match those of NASA across the entire globe, we **do** see that there is a slight increasing trend to the temperature. Even the level observed in Toronto results in 0.6 °C per century, and the IPCC report suggests that an increase of only 2 °C overall can result in a dramatically different global climate, much less welcoming to human life: more droughts, fires, massive weather problems like hurricanes and typhoons, and generally less crop production. Warming is bad for future human happiness!

Question 2: Predicting Temperature for 2013 & 2014

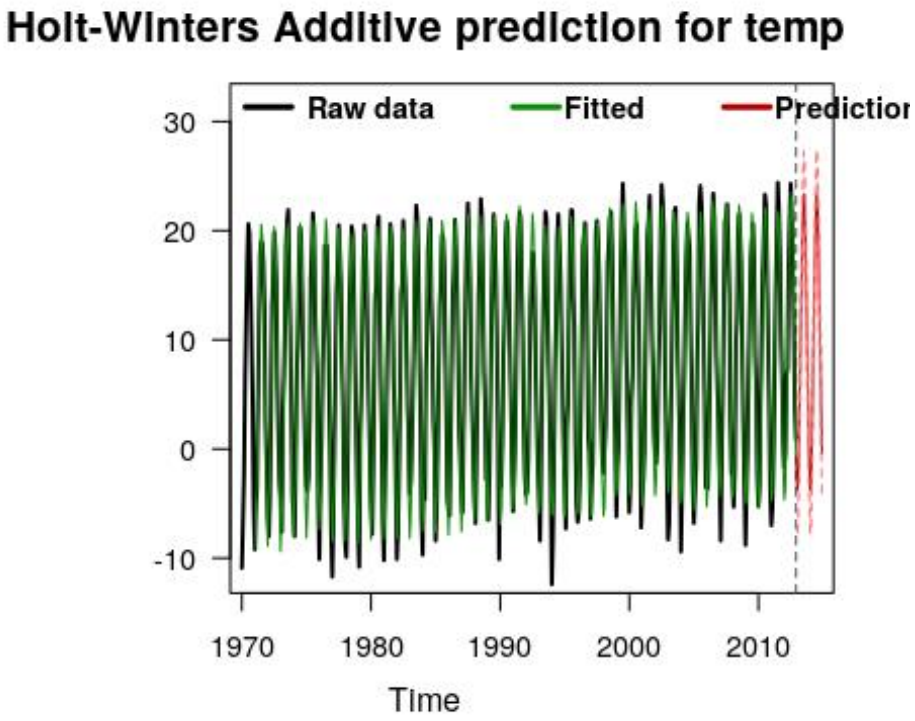
We've been exploring temperature for Toronto from 1970-2012. What if, back in 2012, we wanted to try to predict the temperature for the next couple of years? Time series models allow us to do this highly useful task. In business and industry, one of the most common analytic tasks is forecasting the future: how much demand will there be for your widgets next year? How many passengers will travel on your transit group in the next quarter? How much will your expenses grow in the next decade?

```
pearson_forecast <- forecastplot(pearson_iNZ)
```



By computing this **Holt-Winters** (the name of the algorithm used) prediction, we see an educated guess for the temperatures for 2013 and 2014. The dashed lines around the solid line are **prediction intervals**: boundaries on what the algorithm thinks are reasonable guesses. We see from the plot that the "guess" is that the temperature will continue having seasonal swings (surprise, surprise!), and will proceed roughly as the previous years did, although with slightly less oscillation extremes (heat and cold). Remember that predictions are conservative: we have no

actual information about what will occur, so the algorithms are built to **guess**. And guessing is not



exact!

By computing this **Holt-Winters** (the name of the algorithm used) prediction, we see an educated guess for the temperatures for 2013 and 2014. The dashed lines around the solid line are **prediction intervals**: boundaries on what the algorithm thinks are reasonable guesses. We see from the plot that the "guess" is that the temperature will continue having seasonal swings (surprise, surprise!), and will proceed roughly as the previous years did, although with slightly less oscillation extremes (heat and cold). Remember that predictions are conservative: we have no actual information about what will occur, so the algorithms are built to **guess**. And guessing is not exact!

[*Back to Unit 2 contents page*]

A2: Map Data

Case Study: Safety of Toronto neighborhoods

Wesley Burr & Chris Wild

January 7, 2019

Credit

I've made Chris Wild a co-author on this, as his highly detailed iNZight page on mapping data inspired, guided, and completed my fragmentary thoughts. Any mistakes should not be taken to be his. -wb

A Question

As we should know by now, the Data Science cycle begins with a good question. What do we want to know? Only after we have a clear question in mind should we start diving into data and doing the work of analysis. For today's case study, our question is: "I live in Toronto, and I need to find a new place to live. What neighbourhood should I live in to be safest?". For teachers, the city of Toronto can be reasonably replaced by many other cities around the world, based upon the data which might be available locally to answer the questions. We use Toronto because one of the authors is Canadian and lives near Toronto, and the Toronto Open Data Project is a truly spectacular set of data which will let us answer the question above.

Refining the Question

We need a new place to live, and we want to live (or currently live) in Toronto. But when choosing a neighbourhood to live in "to be safest", what are our criteria? How should we decide? We might be interested in price, or convenience (location, location, location!), or access to amenities, or any number of factors. If covering a case study like this in class, guided choice of criteria would be a great way to personalize the problem and allow students to make unique analyses. We have used "safest", but you could ask "cheapest" or "most fun".

Since we are interested in the most safe neighbourhood, we immediately think about crime. Canada is a reasonably quiet and safe place to live, relative to the entire world, with no wars or epidemics to worry about. However, we do have crime, and crime can make you unsafe. So we want to live in a neighbourhood with minimal crime.

Deciding on which neighbourhood is safest will be how we make our decision. We should note that this criteria alone likely isn't realistic, because we're not looking at size of dwelling (do we need space for multiple people!), cost of housing (we have to be able to afford the housing in the neighbourhood!), or preference factors like style of neighbourhood (detached homes versus high-rise condos, etc.). Refining this case study in a more focused direction (not just "safest neighbourhood") easily allows for scoped problems.

Getting the Data

We will need to use a couple of websites to access the data for Toronto, as it is not all available from one source.

Neighbourhood Shapes

With this analysis, we want to use maps to help guide our decision, as the visualization method allows us to easily compare geographic regions. And at their essence, maps are shapes, joined together - the boundaries of cities, neighbourhoods, or even buildings. For our maps, we will use data from the 2011 Canadian Census which provides the shapes of the edges of many geographic regions in Canada. The data is available from Statistics Canada. On that page, we can select which shapes we want, and we'll download the data in .shp (ArcGIS) format - we're mostly interested in the Forward Sortation Areas (FSAs) which come from Canada Post and represent postcodes.

Postal Codes in Canada

In Canada, all postal codes are six digit alphanumeric strings, as letter-number-letter number-letter-number. For example, the post code for the North Pole and Santa's Workshop is H0H 0H0. For privacy reasons, most data is provided only at the first three digit level, for example H0H only - each six digit code uniquely identifies as few as 20 homes!

General Neighbourhood Information

The Toronto Open Data Portal provides an incredible resource with hundreds of data sets publicly available. For this analysis, we don't use any of that data, but this case study could easily be extended by considering any number of additional variables from this source.

Public Safety and Crime

The Toronto Public Service Public Safety Data Portal provided by the Toronto Police has information on recent crimes, organized by neighbourhood but with latitude and longitude provided for each discrete event. We'll download the **Major Crime Indicators**, which includes Robberies, Break & Enters, Assaults, Auto Theft, and Theft Over: all crimes which may be associated with housing and living, and thus relevant. We could argue for including other crimes (such as Assaults and Murders), although these are not as common in Toronto, and don't actually make much difference in the analysis. In total, the file has 131,073 crimes which occurred from 2014 to 2017, and is provided in a CSV.

Doing the Initial Analysis of our Data

Setting up the Package Environment

Working with mapping data can be quite complicated, as the data sets need to contain strict frameworks in order to accurately locate the information in both space and time. In addition, because we like to add structure to our world, we have human-derived boundaries which affect things: the edges of cities and counties, voting districts, police patrol districts, "neighbourhoods", or even postal codes that determine how our mail and packages get delivered. For this case study, we will be using four main packages:

- **rgeos**: interface from within R to the GEOS (Geometry Engine - Open Source) library, which must also be installed
- **maptools**: open and manipulate objects in the ArcGIS .shp format
- **ggplot2**: create plots and maps
- **tidyverse**: general-purposes tools for data analysis

As with any case study, for using these packages in a workshop or classroom, we recommend using a cloud-based solution or an integrated product like iNZight to avoid technical hassles and let students focus on the key learning outcomes.

Let's load these packages now.

Loading the Geometry and Population

Now we're assuming that the shapefile (.shp) is available and has been downloaded from Statistics Canada (above). We'll load it, filter down to Toronto (since by default it includes all of Canada), and then discard the rest.

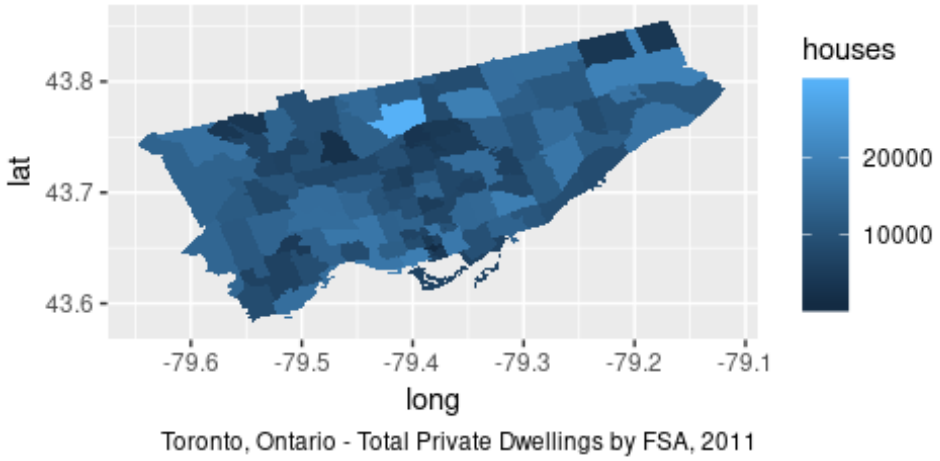
```
FSA <- maptools::readShapeSpatial("gfsa000b11a_e.shp")
toronto <- FSA[substr(FSA$CFSAUID, 1, 1) == 'M',]
rm(FSA)
toronto_df <- ggplot2::fortify(toronto, region = "CFSAUID")
toronto_df$fsa <- factor(toronto_df$id)  # convert to factor
toronto_df$id <- NULL
```

Now one thing we'll need to do is scale by population or population density in our maps. One of the most common errors is to simply plot absolute quantities on a map, ignoring the number of people who live or work in that area. For example, say there were 1000 robberies per year in a certain neighbourhood, and only 200 robberies per month in another. You might think the second neighbourhood was safer! However, if 100,000 people live in the first neighbourhood, and only 1000 in the second neighbourhood, you are **far** more likely to be robbed if you live in the second (1 in 5 chance) as compared to the first (1 in 100 chance). So when creating maps for comparison, if the quantities involve area, volume, population or any other hidden variable, it is important to standardize the values before including them as visualizations on our map.

Again, Statistics Canada provides data for this, as population counts by FSA. This is convenient for us, as we've already settled on FSAs as our definition of "neighbourhoods". In the following, we read in the population data, merge it with the previously loaded FSA data, and create a first trial plot which shades the FSAs by total number of private dwellings occupied by usual residents (that is, homes occupied by a person or family). Not the point of our analysis, but a starting point for visualization. This graph, created by shading all of the FSA cells with a colour based on the single variable of **houses** (total number of dwellings) is known as a *choropleth map* (or sometimes as a *heat map*).

```
popn_fsa <- read_csv("Population_FSA.csv")
popn_fsa <- popn_fsa[, c("Geographic name", "Population, 2011",
                         "Private dwellings occupied by usual residents, 2011")]
names(popn_fsa) <- c("fsa", "pop", "houses")

plot_data <- merge(toronto_df, popn_fsa, by = "fsa")
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = houses)) +
```

```
    geom_polygon() +
    coord_equal() +
    labs(caption = "Toronto, Ontario - Total Private Dwellings by FSA, 2011")
```



Loading the Crime Data

Now let's load all major crimes committed in Toronto over the 2014-2017 time period, obtained from the Toronto Police (details above).

```
crime <- read_csv("MCI_2014_to_2017.csv")
crime <- crime %>%
           filter(reportedyear %in% 2014:2017, occurrenceyear %in% 2014:2017) %>%
           subset(!duplicated(crime$event_unique_id))
crime <- crime[complete.cases(crime), ]
```

Now, these crimes are reported with latitude/longitude measures. For example, the first row in the table is an Assault which occurred in 2014, was reported in 2014, and happened at 43.76883 degrees North latitude and 79.5204 degrees West longitude. We need to organize these crimes into the FSA levels we are using for the others, so we will do closest matching: which FSA center is closest to the crime location?

In doing this analysis, we use another R package:

- **RANN**: R package wrapper for Arya and Mount's "Approximate Nearest Neighbours" (ANN) C++ library

```
crime[1, c("X", "Y", "offence", "reportedyear", "occurrenceyear")]
```

```
## # A tibble: 1 x 5
##       X     Y offence reportedyear occurrenceyear
```

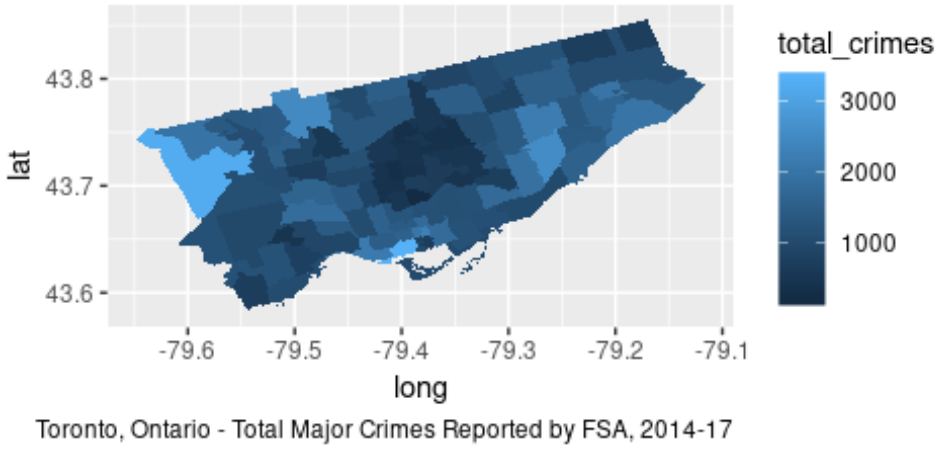
```
##   <dbl> <dbl> <chr>          <dbl>          <dbl>
## 1 -79.5  43.8 Assault         2014           2014

crime_loc <- crime[, c("X", "Y")]
library("RANN")  # find closest line index from the Toronto FSA data frame
closest <- nn2(toronto_df[, c("long", "lat")], crime[, c("X", "Y")], k = 1,
               searchtype = "radius", radius = 0.05)
closest <- sapply(closest, cbind) %>% as_tibble
```

Now for this analysis, we are just interested in the total number of crimes reported in each FSA across the three-year period. We capture this.

```
crime <- data.frame(crime, fsa = toronto_df[closest$nn.idx, "fsa"])
crime_by_fsa <- crime %>%
                  select(reportedyear, fsa) %>%
                  group_by(fsa) %>%
                  summarize(total_crimes = length(fsa))
```

Now, let's recreate the plot from earlier, but instead of filling the FSAs with the number of homes, let's use the number of crimes.

```
plot_data <- merge(plot_data, crime_by_fsa, by = "fsa")
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = total_crimes)) +
    geom_polygon() +
    coord_equal() +
    labs(caption = "Toronto, Ontario - Total Major Crimes Reported by FSA, 2014-17"
)
```



Now, is this a reasonable plot? Can we use this plot, and look for blocks of very dark colours, which correspond to less total crimes?

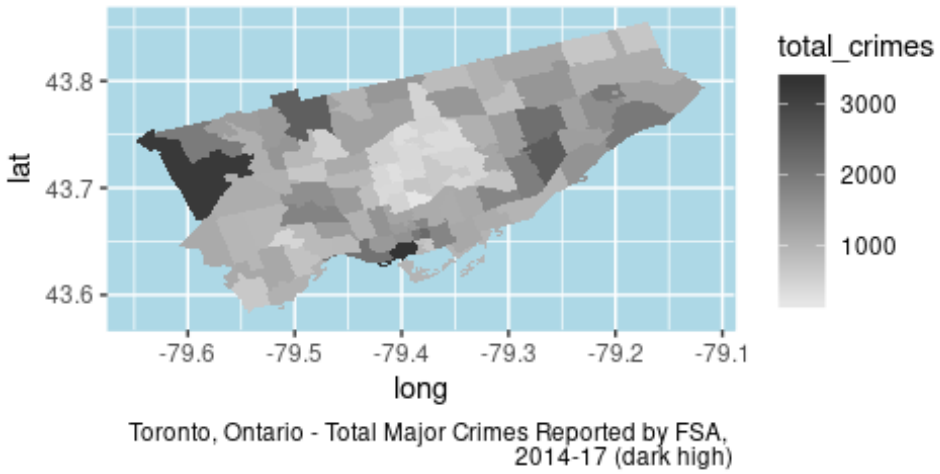
Refining our Visualization

We have the data organized, and now have to question our presentation. Does the sum total of crimes committed in an FSA across a 4-year period actually give a fair interpretation of the safety of a neighbourhood? Remember what we discussed above with our little toy example: more total crime doesn't necessarily imply less safe! We have to consider how many dwellings are in that FSA, and how many people live there, both of which will impact the number of total crimes.

Colour

The shades of blue can be quite hard to distinguish, and in general are likely showing too much information, when all we want to see is the *relative* crime rates, not the absolute numbers. Let's try recreating the above plot in greyscale, effectively converting the crime totals into relative numbers.

```
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = total_crimes)) +
  geom_polygon() +
  coord_equal() +
  labs(caption = "Toronto, Ontario - Total Major Crimes Reported by FSA,
                  2014-17 (dark high)") +
  scale_fill_gradient(low = 'grey90', high = 'grey20') +
  theme(panel.background = element_rect(fill = 'lightblue', colour = 'lightblue'))
```

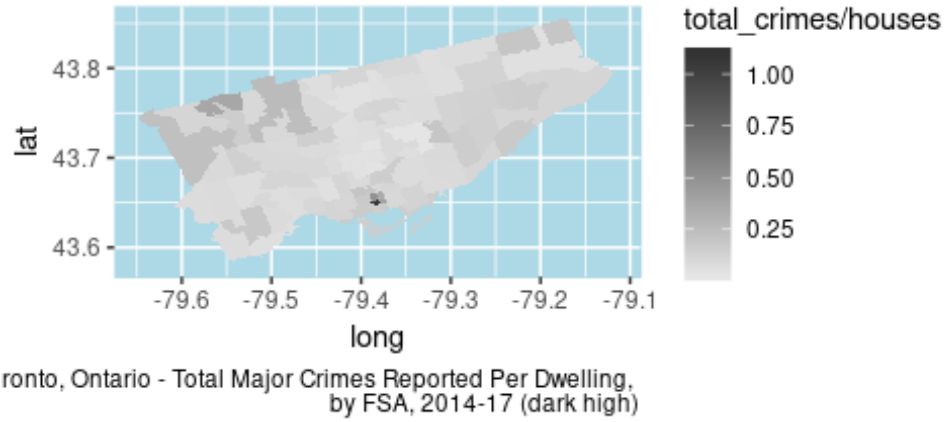


That seems a little better. We now see that dark colours are "bad", and light colours are "good". There seems to be a big cluster of FSAs in the center of Toronto with lower crime than the other areas. But remember our discussion earlier: the number of dwellings and number of people living in an FSA will impact the number of crimes!

Scaling

The number of dwellings in an FSA may affect the number of Break and Enters and Robberies. It isn't clear that the relationship will be linear, or increasing (after all, Robberies may be higher in commercial districts, especially in areas where convenience stores and similar are open after dark), but there will be some form of interaction between the variables. Let's try recreating the above plot, but scaling by the number of dwellings in each FSA. The result will be a **rate**: the total number of MCI (Major Crimes) per dwelling.

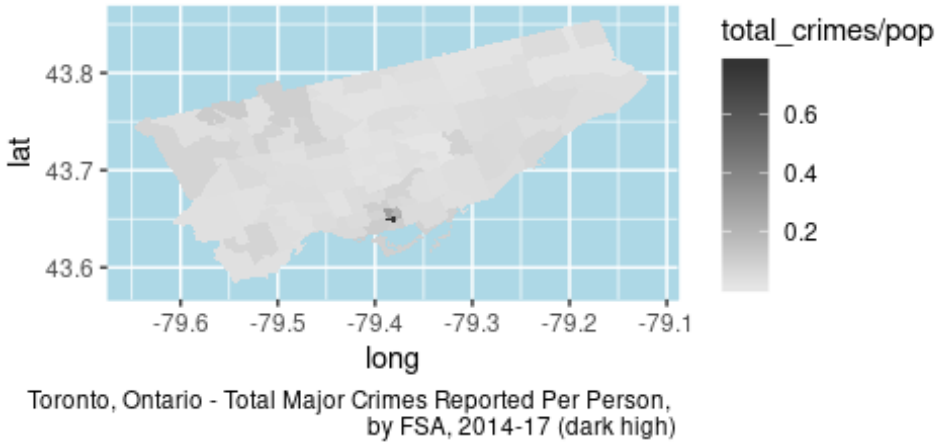
```
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = total_crimes / house
s)) +
  geom_polygon() +
  coord_equal() +
  labs(caption = "Toronto, Ontario - Total Major Crimes Reported Per Dwelling,
                  by FSA, 2014-17 (dark high)") +
  scale_fill_gradient(low = 'grey90', high = 'grey20') +
  theme(panel.background = element_rect(fill = 'lightblue', colour = 'lightblue'))
```



That certainly looks different! There is a tiny dark FSA cluster in the downtown area of Toronto, and the colours taper off quite quickly. It looks like the dark FSAs in the north-west of the city weren't quite as bad as they seemed, since there are so many homes in that area which were inflating the numbers somewhat. The entire middle section of the city now looks equivalent.

We should also check scaling by population, versus number of homes: perhaps some of these neighbourhoods contain limited numbers of dwellings, but densely-populated dwellings (bigger families, perhaps?) which are also changing the results.

```
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = total_crimes / pop))
 +
  geom_polygon() +
  coord_equal() +
  labs(caption = "Toronto, Ontario - Total Major Crimes Reported Per Person,
                  by FSA, 2014-17 (dark high)") +
  scale_fill_gradient(low = 'grey90', high = 'grey20') +
  theme(panel.background = element_rect(fill = 'lightblue', colour = 'lightblue'))
```

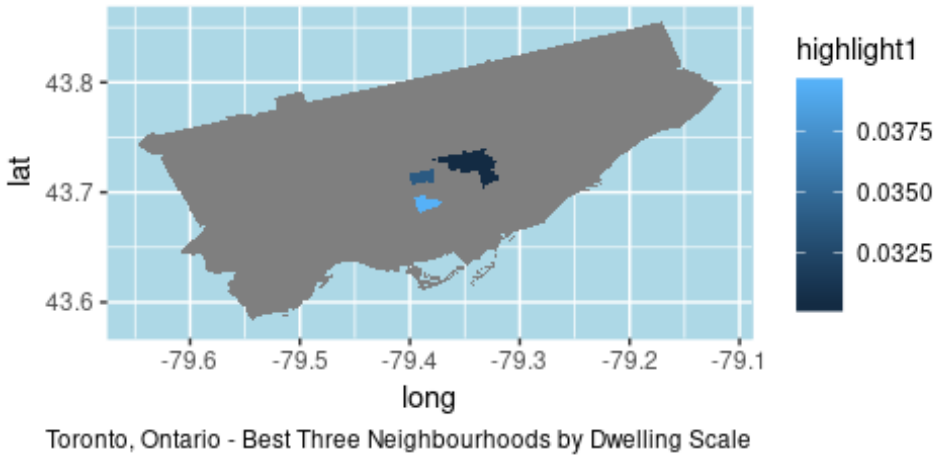


This scaling is even more aggressive: almost the entire city is being washed out, with only two or three FSAs having high numbers of crimes relative to the number of persons in the FSA. Remember: the number of people living in each FSA can vary wildly, from a minimum for this data set of 1027 and a maximum of 67251. This scaling doesn't seem very informative: if we took it seriously, it says that almost the entire city is the same level.

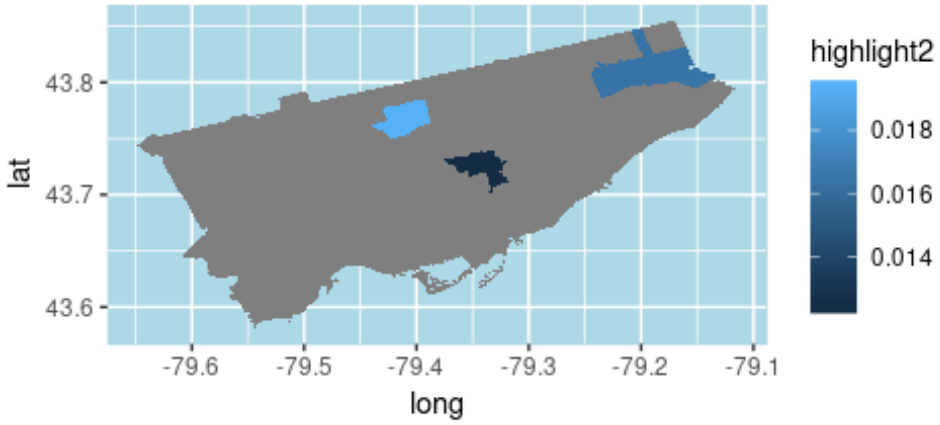
Highlighting the Best

```
plot_data <- data.frame(plot_data, highlight1 = plot_data$total_crimes / plot_data$
houses,
                                   highlight2 = plot_data$total_crimes / plot_data$
pop)
best_three_houses <- sort(unique(plot_data$highlight1))[1:3]
best_three_pop <- sort(unique(plot_data$highlight2))[1:3]
plot_data[!(plot_data$highlight1 %in% best_three_houses), "highlight1"] <- NA
plot_data[!(plot_data$highlight2 %in% best_three_pop), "highlight2"] <- NA

ggplot(plot_data, aes(x = long, y = lat, group = group, fill = highlight1)) +
  geom_polygon() +
  coord_equal() +
```

```
  labs(caption = "Toronto, Ontario - Best Three Neighbourhoods by Dwelling Scale")
+
  theme(panel.background = element_rect(fill = 'lightblue', colour = 'lightblue'))
```




```
ggplot(plot_data, aes(x = long, y = lat, group = group, fill = highlight2)) +
  geom_polygon() +
  coord_equal() +
  labs(caption = "Toronto, Ontario - Best Three Neighbourhoods by Population Scale"
) +
  theme(panel.background = element_rect(fill = 'lightblue', colour = 'lightblue'))
```



Toronto, Ontario - Best Three Neighbourhoods by Population Scale

The unique neighbourhoods, by FSA as neighbourhood, are M3C, M4P, M4T for the dwelling logic, and M1B, M2N, M3C for the population logic. We notice that M3C shows up twice, which seems fortuitious. Let's look this up.

FSA 'M3C'

The M3C FSA is in North York, and is the community of Flemingdon Park. The area is a former industrial area, and is now considered light industrial, with many corporate offices located there. The population is 38289, with 15625 dwellings. There are a number of highrises and middle and upper-middle income homes, with a blend of different income classes. In other words: this is a perfectly reasonable neighbourhood to live in, almost regardless of income, and appears to be quite safe!

Conclusion

We started with a question: where should we live to be "safest" within the city of Toronto? This question led us to the idea that crimes make us unsafe, so by analyzing location and crime data, we were able to isolate and examine individual neighbourhoods, eventually determining that the neighbourhood of Flemingdon Park in North York is the "safest" neighbourhood in Toronto, for the period 2014-2017.

As always with these analyses, this is **not** the final step! We could consider what it means to be safe, and re-examine the crime data. Maybe you plan on living in a gated community or a highrise with a security code or guard, and thus consider Break and Enters to be less important to you. Or perhaps you are most worried about being Assaulted, so we should download and examine

Assault data for the city instead of the MCI. We could also consider other factors in trying to find a place to live: cost of housing, or convenience of transit.

The big advantage of using mapping data is the incredibly powerful visualization tools that come with presenting the data in a multi-dimensional format, especially with the use of shading and colour. While all of the above analysis could have been done without presenting a single figure, would we have understand the layout of the city and the selection of this neighbourhood as well if we hadn't done these steps?

We hope this case study has inspired you to consider what other factors you can consider when working with mapping data. Happy mapping!

References

- Roger Bivand and Colin Rundel (2018). rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.4-2. https://CRAN.R-project.org/package=rgeos
- Roger Bivand and Nicholas Lewin-Koh (2018). maptools: Tools for Handling Spatial Objects. R package version 0.9-4. https://CRAN.R-project.org/package=maptools
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
- Sunil Arya, David Mount, Samuel E. Kemp and Gregory Jefferis (2019). RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric. R package version 2.6.1. https://CRAN.R-project.org/package=RANN
- D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
- Unpleasant Conjunction. *Choropleth Maps with R and ggplot2*. https://unconj.ca/blog/choropleth-maps-with-r-and-ggplot2.html
- iNZight and Chris Wild. *About Presenting Data on Maps*. https://www.stat.auckland.ac.nz/~wild/iNZight/user_guides/add_ons/?topic=aboutmaps
- Toronto Open Data Project. https://www.toronto.ca/city-government/data-research-maps/open-data/
- Statistics Canada. *Geometry Boundary Limits*. https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2011-eng.cfm
- Toronto Police. *Public Safety Data Portal*. http://data.torontopolice.on.ca/datasets/mci-2014-to-2017

[*Back to Unit 2 contents page*]

A3: K-Means Clustering

Applying K-Means Clustering to Delivery Fleet Data

(Reference: https://www.datascience.com/blog/k-means-clustering)

Example: Applying K-Means Clustering to Delivery Fleet Data

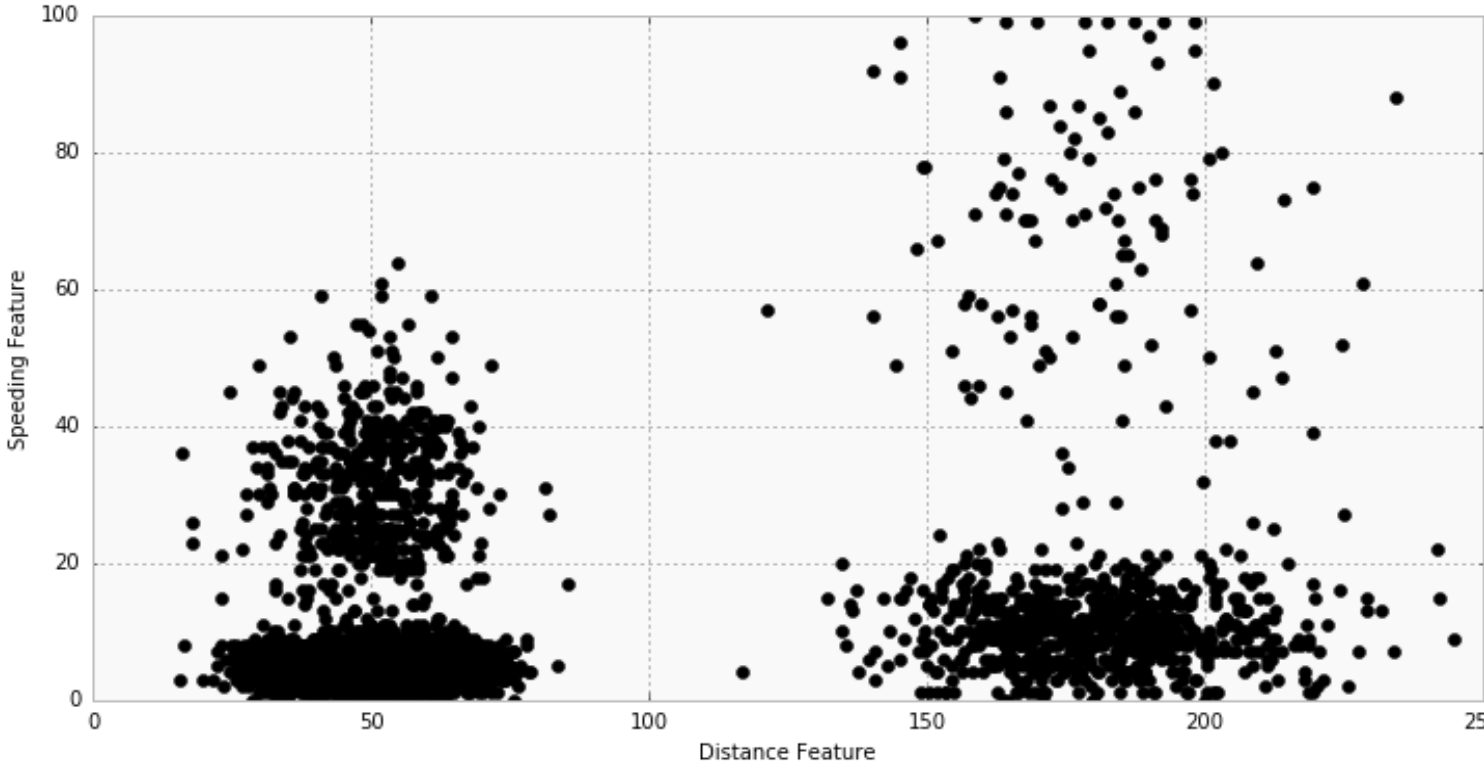
As an example, we'll show how the K-means algorithm works with a sample dataset of delivery fleet driver data. For the sake of simplicity, we'll only be looking at two driver features: mean distance driven per day and the mean percentage of time a driver was >5 mph over the speed limit.

Step 1: Clean and Transform Your Data

For this example, we've already cleaned and completed some simple data transformations. A sample of the data as a **`pandas DataFrame`** is shown below.

	Driver_ID	Distance_Feature	Speeding_Feature
0	3423311935	71.24	28
1	3423313212	52.53	25
2	3423313724	64.54	27
3	3423311373	55.69	22
4	3423310999	54.58	25

The chart below shows the dataset for 4,000 drivers, with the distance feature on the x-axis and speeding feature on the y-axis.



Step 2: Choose K and Run the Algorithm

Start by choosing *K*=2. For this example, use the Python packages scikit-learn and NumPy for computations as shown below:

```
import numpy as np


from sklearn.cluster import KMeans




### For the purposes of this example, we store feature data from our
### dataframe `df`, in the `f1` and `f2` arrays. We combine this into
### a feature matrix `X` before entering it into the algorithm.
f1 = df['Distance_Feature'].values
f2 = df['Speeding_Feature'].values

X=np.matrix(zip(f1,f2))
kmeans = KMeans(n_clusters=2).fit(X)
```

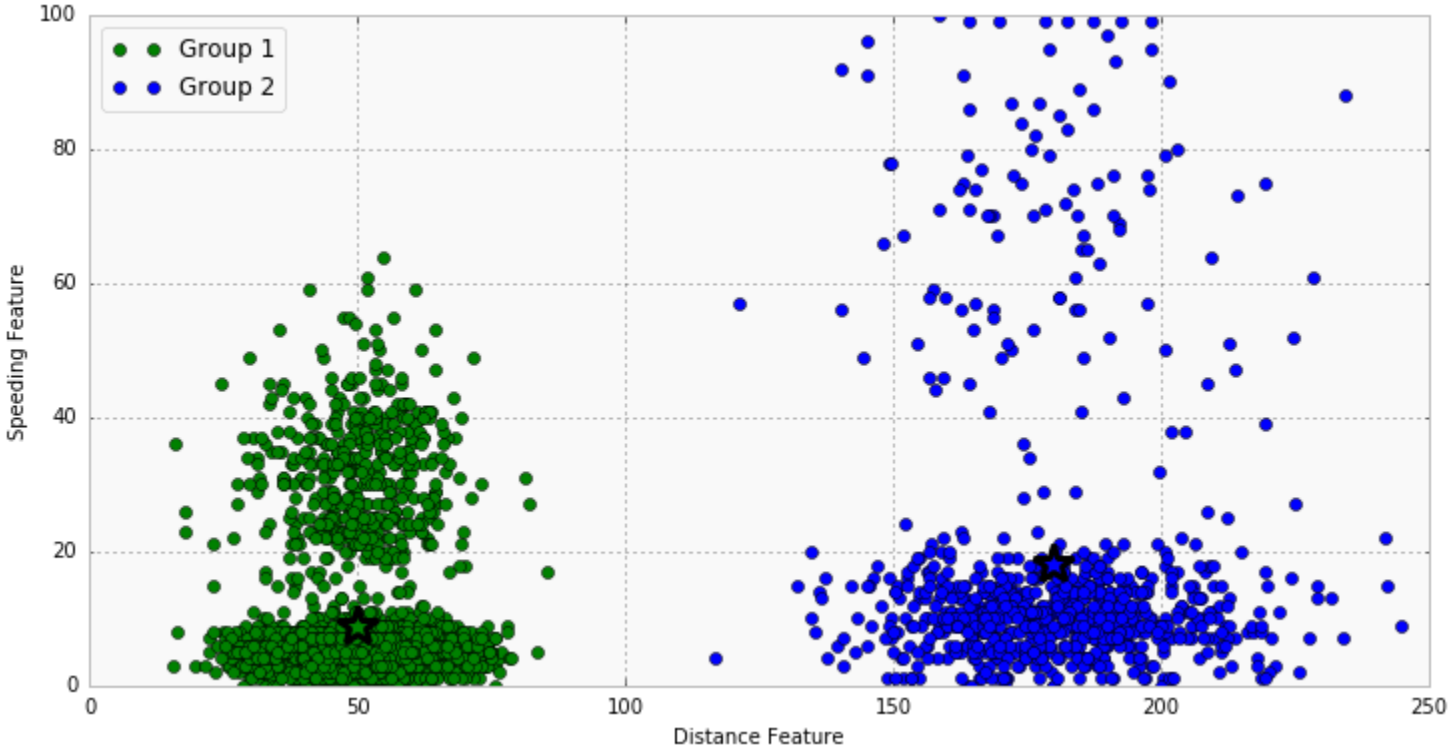
The cluster labels are returned in **`kmeans.labels_`**.

Step 3: Review the Results

The chart below shows the results. Visually, you can see that the *K*-means algorithm splits the two groups based on the distance feature. Each cluster centroid is marked with a star.

- Group 1 Centroid = (50, 5.2)
- Group 2 Centroid = (180.3, 10.5)

Using domain knowledge of the dataset, we can infer that Group 1 is urban drivers and Group 2 is rural drivers.

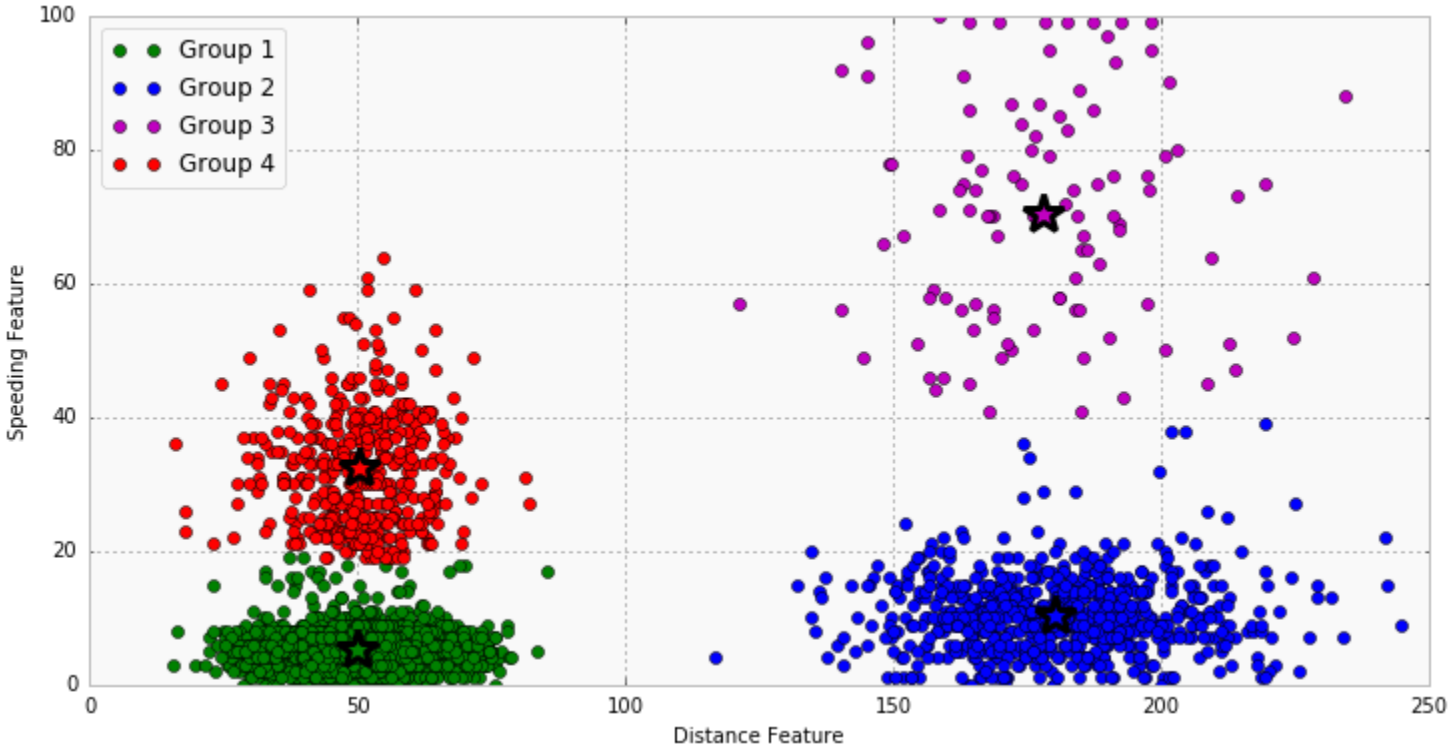


Step 4: Iterate Over Several Values of K

Test how the results look for *K*=4. To do this, all you need to change is the target number of clusters in the **`KMeans()`** function.

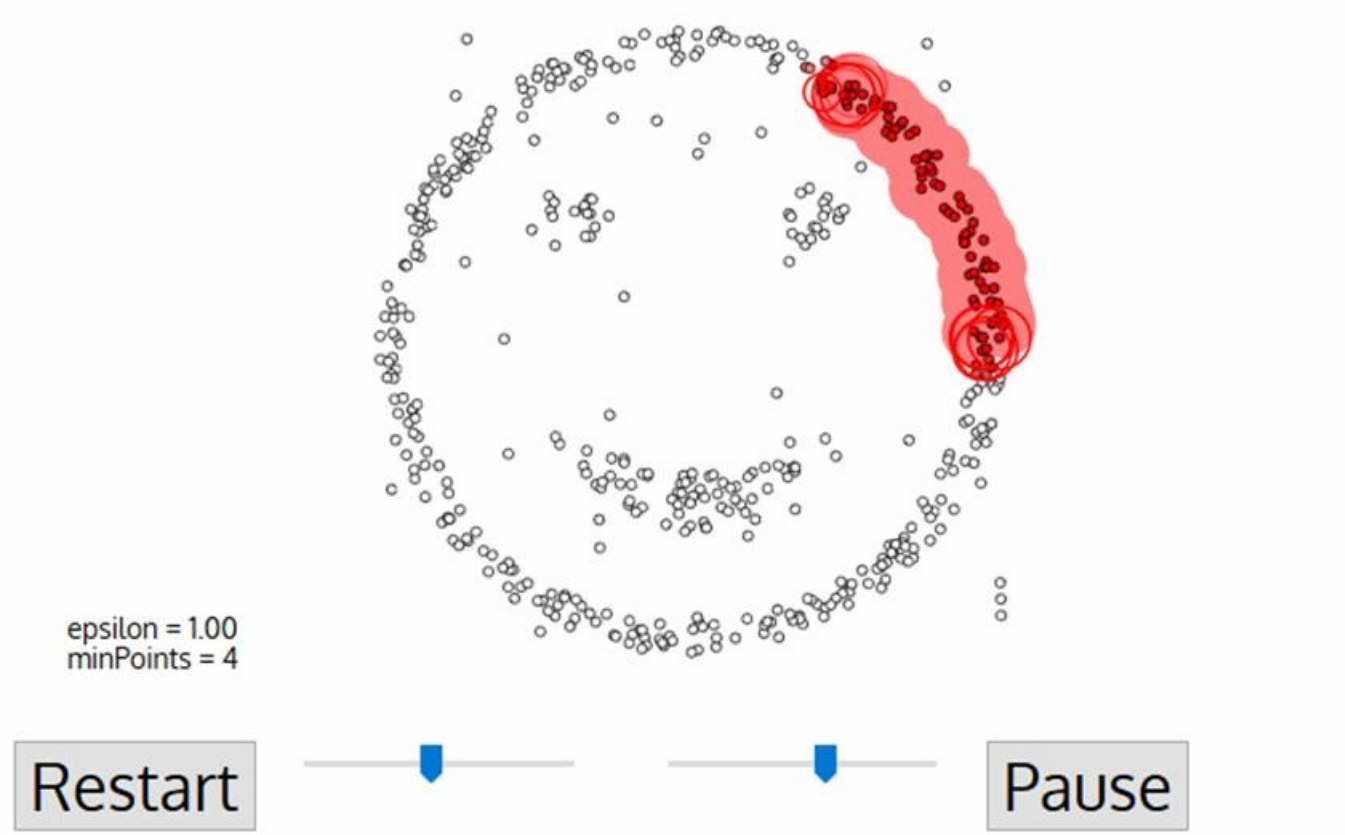
```
kmeans = KMeans(n_clusters=4).fit(X)
```

The chart below shows the resulting clusters. We see that four distinct groups have been identified by the algorithm; now speeding drivers have been separated from those who follow speed limits, in addition to the rural vs. urban divide. The threshold for speeding is lower with the urban driver group than for the rural drivers, likely due to urban drivers spending more time in intersections and stop-and-go traffic.



OVERVIEW of DBSCAN Algorithm

DBSCAN is a density based clustered algorithm; it clusters regions based on the density of points.



https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

1. DBSCAN begins with an arbitrary starting data point that has not been visited. The neighborhood of this point is extracted using a distance epsilon ε (All points which are within the ε distance are neighborhood points).
2. If there are a sufficient number of points (according to minPoints) within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster. Otherwise, the point will be labeled as noise (later this noisy point might become the part of the cluster). In both cases that point is marked as "visited".
3. For this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.

4. This process of steps 2 and 3 is repeated until all points in the cluster are determined i.e, all points within the ε neighborhood of the cluster have been visited and labelled.
5. Once we're done with the current cluster, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise. This process repeats until all points are marked as visited. Since at the end of this all points have been visited, each point well have been marked as either belonging to a cluster or being noise.

[*Back to Unit 2 contents page*]