

DataBOOM: the canon for data science

Databrew

2021-06-07

Contents

1	Welcome!	9
	What this is, and what it isn't	9
	Who this is for	9
	What you will learn	10
	Who we are	11
I	Core theory	13
2	Principles of data science	15
	What is data science?	15
	What is the data life cycle?	15
	What is a pipeline?	15
	Data science 'in the wild'	15
	The reproducibility crisis	15
3	Visualizing data	17
3.1	Bad examples	17
3.2	Good examples	17
3.3	Edward Tufte	17
3.4	Grammar of graphics	17
3.5	Design principles	17
3.6	Plots & power	17
4	Data ethics	19
5	Setting up RStudio	23
6	Running R code	25
	RStudio's <i>Console</i>	25
	Running code in the <i>Console</i>	26
	Use R like a calculator	28
	Using operators in R	29
	Use built-in functions within R	30

7	Using RStudio and R scripts	33
	R and RStudio: what's the difference?	33
	Two-minute tour of RStudio	34
	Scripts	35
	Your working directory	40
	Typical workflows	41
8	Variables in R	45
	Introducing variables	45
	Types of data in R	47
9	Structures for data in R	51
	Introducing data structures	51
	Vectors	51
10	Calling functions	57
	Introducing R functions	57
11	Base plots	63
	Introduction	63
	Create a basic plot	63
	Most common types of plots	64
	Basic plot formatting	66
	Plotting with data frames	72
	Next-level plotting	73
12	Packages	81
	Introducing R packages	81
	Finding the packages already on your computer	82
	Installing a new package	83
	Loading an installed package	84
	Calling functions from a package	85
	Review: the workflow for using a package	85
	A note on package dependencies	87
	A note on package versions	87
13	Basics of ggplot	89
	What is ggplot?	89
	The name and concept	89
	A practical example	90
	Learning examples	90
	Other resources	90
14	Importing data	93
	.csv files	93
	Data format requirements	95
	Reading in data	97

15 Dataframes	101
Subsetting & exploring dataframes	101
Tidy data	108
16 Exporting data & plots	109
write.csv()	109
saveRDS()	110
pdf() and png()	110
17 Exploring data	113
Exploring distributions	114
Descriptive statistics	121
18 Significance statistics	123
p-values	123
Tests for different data types	124
Comparison tests	124
Tests of association	134
Reporting results	137
19 Joining datasets	145
Joining: the basics	145
20 Writing functions	155
First steps	155
Next steps	156
Sourcing functions	162
21 for loops	165
Basics	165
for loops in plots	167
Using for loops to process & summarize data	174
22 Conditional statements	187
First steps	187
Next steps	189
23 Working with text	193
Basics	193
Common tools	193
24 Working with dates & times	197
25 Working with factors	199
26 Matrices & lists	201
Lists	201

Matrices	203
27 Cleaning messy data	211
28 Excel files & GoogleSheets	213
 II Interactive dashboards	 215
29 Intro to Shiny apps	217
30 Shiny dashboards	219
31 Data entry apps	221
 III Databases	 223
32 Introduction	225
32.1 What	225
32.2 Why	225
32.3 When	225
32.4 When not	225
33 Platforms	227
33.1 PostgreSQL	227
33.2 mySQL	227
33.3 SQLite	227
34 Alternatives	229
34.1 NoSQL	229
35 Practices	231
 IV Documenting your work	 233
36 R Markdown	235
37 Reproducible research	237
38 Automated reporting	239
39 Formatting standards	241
39.1 Tables	241
39.2 Figures	241
39.3 Captions	241

40 Git	243
There is a better way	245
What is git?	246
Why?	246
Get ready for git	247
Installation	248
Getting to know git bash	251
Configuration	253
Github	253
A bit more practice	255
Advanced git	256
 V Sharing research	 257
41 Sections of a report	259
Overall structure	259
Abstract	260
Introduction	262
Methods	266
Results	267
Discussion	269
Other elements of a report	270
42 Formatting, structure & style	271
Core principles	271
Further considerations	275
Logistics of scientific writing	277
Formatting tables	280
Formatting figures	280
43 Presentations	283
Creating websites	291
 VI Advanced skills	 293
44 Mapping	295
45 Geographic computing & GIS	297
46 Statistical modeling	299
47 Apply family	301
48 Randomization statistics	303

Basic idea	303
Common use cases	305
Review assignment	316
(c) Do social bonds relate to kinship?	319
49 Iterative simulations	323
50 Image analysis	325
51 Machine learning	327

Chapter 1

Welcome!

Welcome to DataBOOM, a curriculum designed to guide you from your very first line of code towards becoming a professional data scientist.

What this is, and what it isn't

This is not a textbook or a reference manual. It is not exhaustive or comprehensive. It is a *training manual* designed to *empower researchers to do impactful data science*. As such, its tutorials and exercises aim to get you, the researcher, to start writing your own code as quickly as possible and – equally of importance – to *start thinking like a data scientist*, by which we mean tackling ambiguous problems with persistence, independence, and creative problem solving.

Furthermore, this is not a fancy interactive tutorial with bells or whistles. It was purposefully designed to be simple and “analog”. You will not be typing your code into this website and getting feedback from a robot, or setting up an account to track your progress, or getting pretty merit badges or points when you complete each module.

Instead, you will be doing your work on your own machine, working with real folders and files, downloading data and moving it around, etc. – all the things you will be doing as a data scientist in the real world.

Who this is for

This curriculum covers everything from the absolute basics of writing code in R to machine learning with `tensorflow`. As such, it is designed to be useful to everyone in some way. But the target audience for these tutorials is the student who *wants* to work with data but has *zero* formal training in programming, computer science, or statistics.

This curriculum was originally developed for the **Sewanee Data Institute for Social Good** at Sewanee: The University of the South, TN, USA.

What you will learn

- The **Core theory** unit establishes the conceptual foundations and motivations for this work: what data science is, why it matters, and ethical issues surrounding it: the good, the bad, and the ugly.

The next several units comprise a *core* curriculum for tackling data science problems:

- The **Getting started** unit teaches you how to use R (in RStudio) to explore and plot data. Here you will add the first and most important tools to your toolbox: working with variables, vectors, dataframes, scripts, and file directories.
- The **Basic R workflow** unit teaches you how to bring in your own data and work with it in R. You will learn how to format data to simplify analysis and add tools for *data wrangling* (i.e., transforming and re-formatting data to prepare it for plotting and analysis). You will also learn how to conduct basic statistics, from exploratory data analyses (e.g., producing and comparing distributions) to significance testing.
- The **Essential R skills** unit equips you with the tools, tricks, and mindset for tackling the most common tasks in data science. This is where you really begin to cut your teeth on real-world data puzzles: figuring out how to use the R tools in your toolbag to tackle an ambiguous problem and deliver an excellent data product.

The next several units provide a suite of skills essential to any data science professional:

- The **Interactive dashboards** unit teaches you how to make dashboards and websites for projects using **shiny** in RStudio.
- The **Databases** unit teaches you how to access, create, and work with relational databases online using **SQL** and its alternatives.
- The **Documenting your work** unit teaches you to use **R Markdown** to produce beautiful, reproducible data reports. You will also learn about *version control*, using **Git** and **GitHub** to collaborate on shared projects and work on data science teams.
- The **Sharing research** unit teaches you to produce publishable research articles and compelling presentations.

The final unit, **Advanced skills**, introduces you to a variety of advanced data science techniques, from interactive maps to iterative simulations to machine learning, that can help you begin to specialize your skillset.

Who we are

Joe Brew is a data scientist, epidemiologist, and economist. He has worked with the Florida Department of Health (USA), the Chicago Department of Public Health (USA), the Barcelona Public Health Agency (Spain), the Tigray Regional Health Bureau (Ethiopia) and the Manhica Health Research Center (Mozambique). He is a co-founder of Hyfe and DataBrew. His research focuses on the economics of malaria and its elimination. He earned his BA at Sewanee: The University of the South (2008), an MA at the Institut Catholique de Paris (2009) and an MPH at the Kobenhavns Universitet (2013). He is passionate about international development, infectious disease surveillance, teaching, running, and pizza.

Eric Keen is a data scientist, marine ecologist, and educator. He is the Science Co-director at BCwhales, a research biologist at Marecotel, a data scientist at Hyfe, and a professor of Environmental Studies at Sewanee: the University of the South. He earned his BA at Sewanee (2008) and his PhD at Scripps Institution of Oceanography (2017). His research focuses on the ecology and conservation of whales in developing coastal habitats. He is passionate about whales, conservation, teaching, small-scale farming, running, and bicycles. And pizza.

Part I

Core theory

Chapter 2

Principles of data science

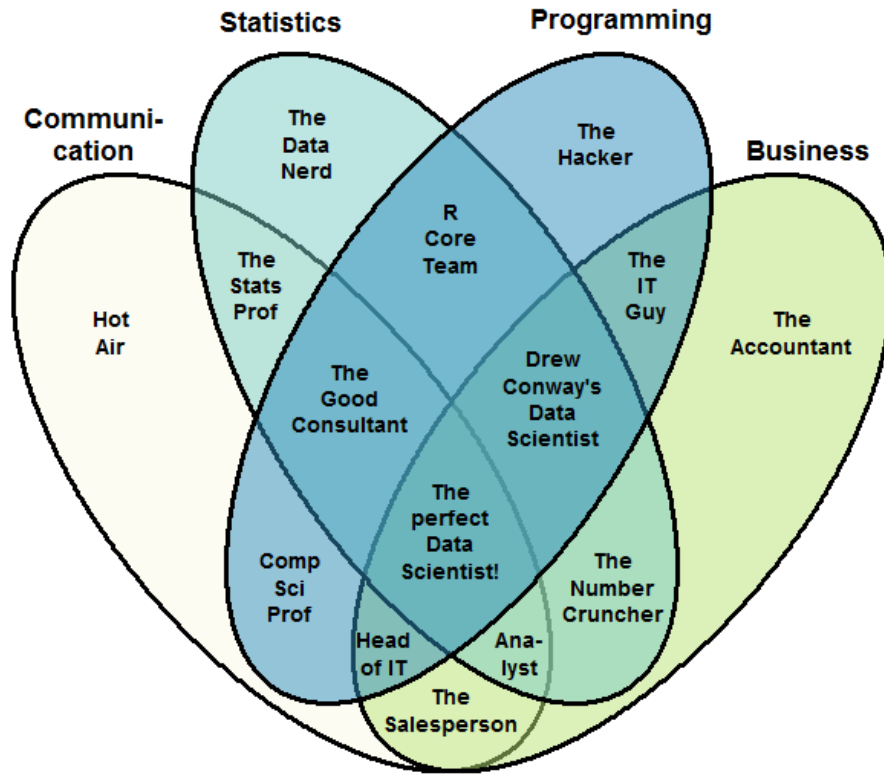
What is data science?

Data science is an interdisciplinary field. Some have argued that it is not a field unto itself, but rather an extension of statistics. In this course, however, we'll take the majority view that data science is its own field: a new field, which combines statistics, mathematics, and computer science.

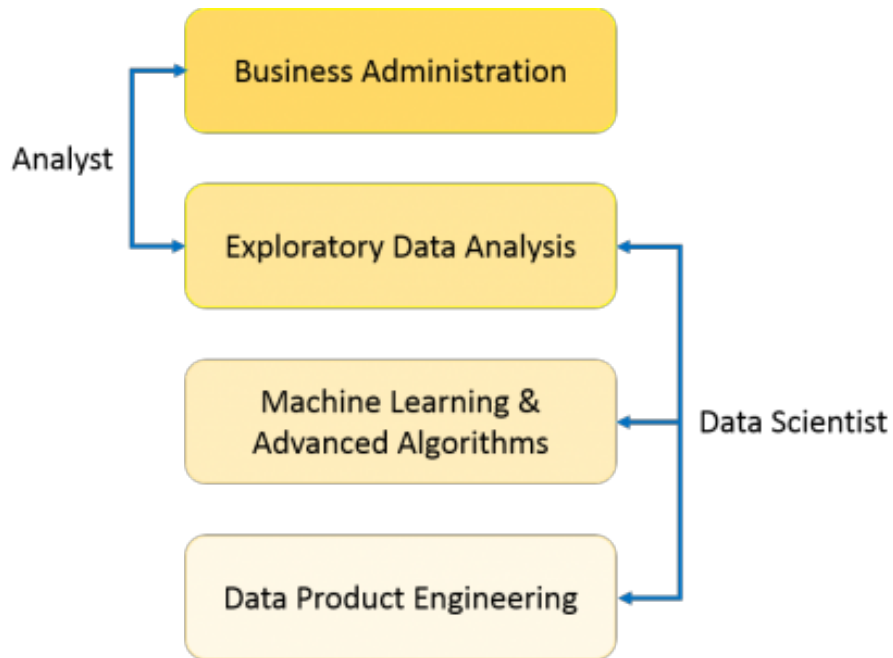
But we'll go one step outward. Data science is not just the combination of those academic disciplines which form its core; its also something more. Good data science involves domain knowledge (ie, familiarity with the problem being solved), effective communication, an iterative mentality (ie, creating feedback loops for rapid hypothesis testing), a bias to real-world effects rather than theoretical frameworks, and a willingness/desire to work in the real world.

There are a lot of Venn diagrams and figures out there, trying to show what data science is. For example...

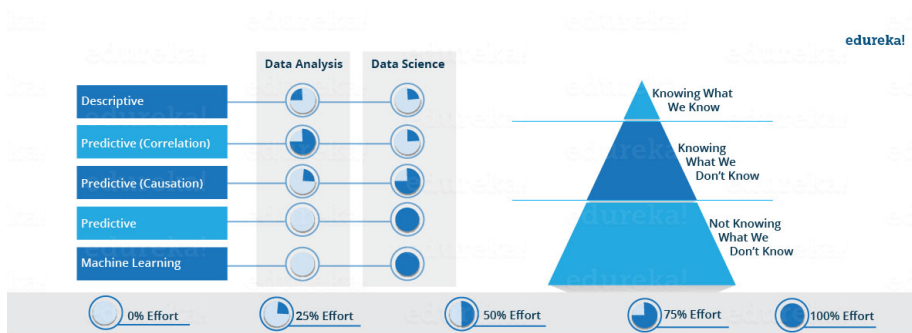
The Data Scientist Venn Diagram



... OR ...



... OR ...



... OR ...