`DataBOOM`: the canon for data science

Databrew

2021-06-07

Contents

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 Manhiça 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

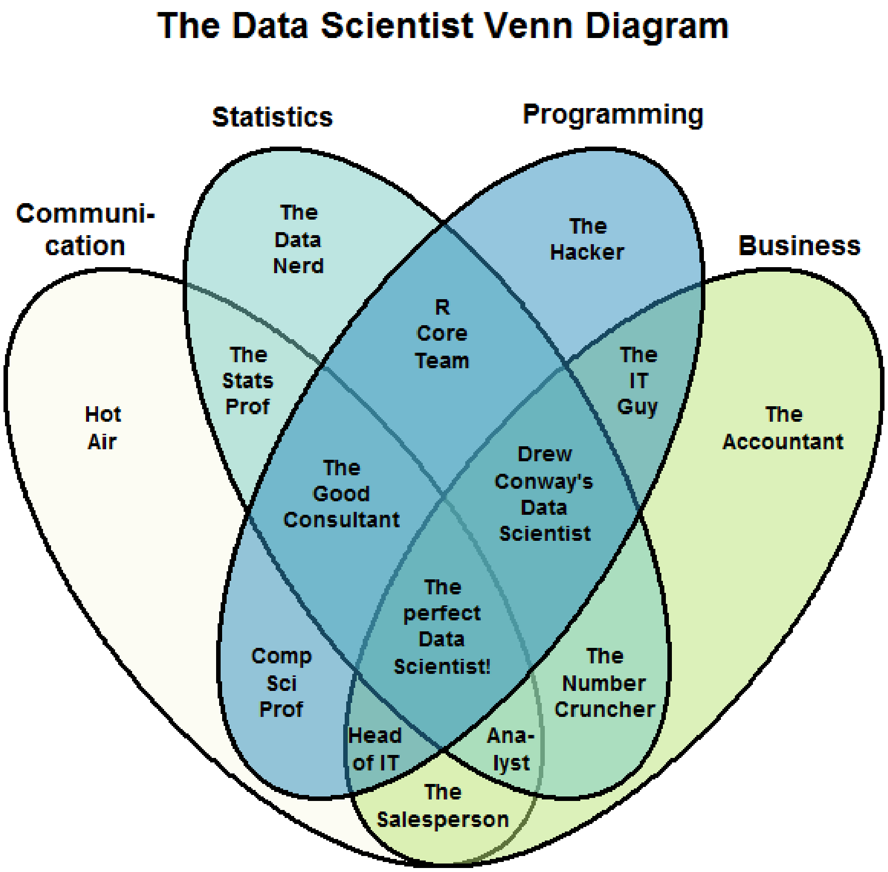
Principles of data science

What is data science?

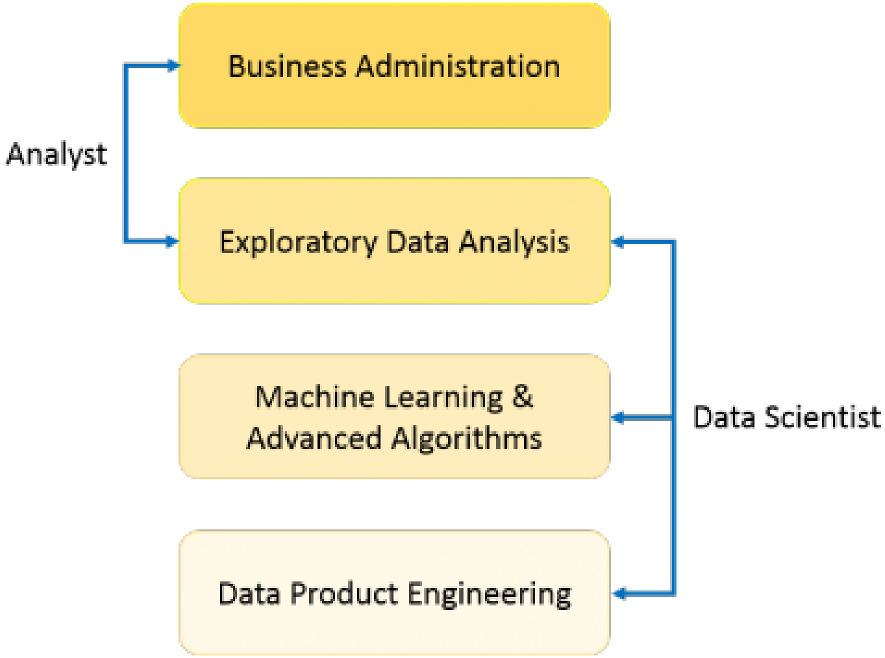
Data science is an interdisciplinary field. Some have argued that it is not a field unto itself, but rather an extension of statistics. In this course, however, we'll take the majority view that data science is its own field: a new field, which combines statistics, mathematics, and computer science.

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.

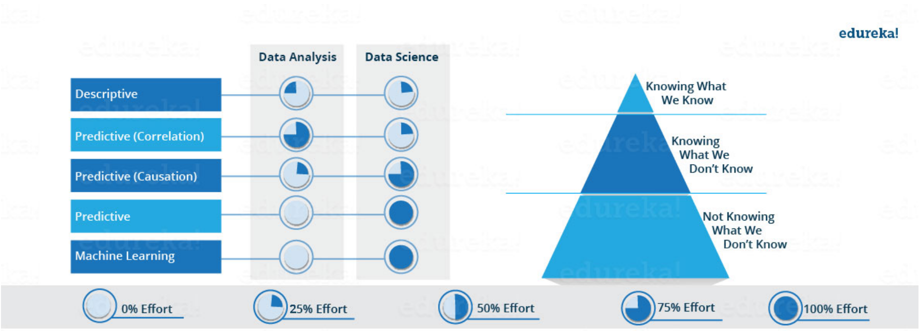
There are a lot of Venn diagrams and figures out there, trying to show what data science is. For example…


The Data Scientist Venn Diagram
Statistics
Programming
Communi-cation
Business
The Data Nerd
The Hacker
R Core Team
The Stats Prof
The IT Guy
Hot Air
The Accountant
The Good Consultant
Drew Conway's Data Scientist
The perfect Data Scientist!
Comp Sci Prof
The Number Cruncher
Head of IT
Ana-lyst
The Salesperson

... or ...


Business Administration
Analyst
Exploratory Data Analysis
Machine Learning & Advanced Algorithms
Data Scientist
Data Product Engineering

... or ...


edureka!
Data Analysis
Data Science
Descriptive
Predictive (Correlation)
Predictive (Causation)
Predictive
Machine Learning
Knowing What We Know
Knowing What We Don't Know
Not Knowing What We Don't Know
0% Effort
25% Effort
50% Effort
75% Effort
100% Effort

... or ...