An Introduction to Data Science for Sensory and Consumer Scientists

John Ennis, Julien Delarue, and Thierry Worch

2020-12-29

Contents

Preface

Welcome to the website for *Introduction to Data Science for Sensory and Consumer Scientists*. This book being written in the open and is currently under development.

About the Authors

John Ennis ...

Julien Delarue ...

Thierry Worch ...

Introduction

Chapter 1

Introduction

Sensory and consumer science (SCS) is consider as a pillar of food science and technology and is useful to product development, quality control and market research. Most scientific and methodological advances in the field are applied to food. This book makes no exception as we chose a cookie formulation dataset as a main thread. However, SCS widely applies to many other consumer goods so are the content of this book and the principles set out below.

1.1 Core principles in Sensory and Consumer Science

1.1.1 Measuring and analyzing human responses

Sensory and consumer science aims at measuring and understanding consumers' sensory perceptions as well as the judgements, emotions and behaviors that may arise from these perceptions. SCS is thus primarily a science of measurement, although a very particular one that uses human beings and their senses as measuring instruments. In other words, sensory and consumer researchers measure and analyze human responses. To this end, SCS relies essentially on sensory evaluation which comprises a set of techniques that mostly derive from psychophysics and behavioral research. It uses psychological models to help separate signal from noise in collected data [ref O'Mahony, D.Ennis, others?]. Besides, sensory evaluation has developed its own methodological framework that includes most refined techniques for the accurate measurement of product sensory properties while minimizing the potentially biasing effects of brand identity and the influence of other external information on consumer perception [Lawless & Heymann, 2010]. A detailed description of sensory methods is beyond the scope of this book and many textbooks on sensory evaluation methods are available to

readers seeking more information. However, just to give a brief overview, it is worth remembering that sensory methods can be roughly divided into three categories, each of them bearing many variants: - Discrimination tests that aim at detecting subtle differences between two products. - Descriptive analysis (DA), also referred to as 'sensory profiling', aims at providing both qualitative and quantitative information about product sensory properties. - Hedonic tests. This category gathers affective tests that aim at measuring consumers' liking for the tested products or their preferences among a product set. Each of these test categories generates its own type of data and related statistical questions in relation to the objectives of the study. Typically, data from difference tests consist in series of correct/failed binary answers depending on whether judges successfully picked the odd sample(s) among a set of three or more samples. These are used to determine whether the number of correct choices is above the level expected by chance. Conventional descriptive analysis data consist in intensity scores given by each panelist to evaluated samples on a series of sensory attributes, hence resulting in a product x attribute x panelist dataset (Figure 1). Note that depending on the DA method, quantifying means other than intensity ratings can be used (ranks, frequency, etc.). Most frequently, each panelist evaluates all the samples in the product set. However, the use of balanced incomplete design can also be found when the experimenters aim to limit the number of samples evaluated by each subject. Eventually, hedonic test datasets consist in hedonic scores (ratings for consumers' degree of liking or preference ranks) given by each interviewed consumer to a series of products. As for DA, each consumer usually evaluates all the samples in the product set, but balanced incomplete designs are sometimes used too. In addition, some companies favor pure monadic evaluation of product (i.e. between-subject design or independent groups design) which obviously result in unrelated sample datasets. Sensory and consumer researchers also borrow methods from other fields, in particular from sociology and experimental psychology. Definitely a multidisciplinary area, SCS develops in many directions and reaches disciplines that range from genetics and physiology to social marketing, behavioral economics and computational neuroscience. So have diversified the types of data sensory and consumer scientists must deal with.

1.2 How should sensory and consumer scientists learn data science?

1.3 Caution: Don't that everybody does

1.4 Example projects

Chapter 2

What is Data Science?

In this chapter we explain what is data science.

2.1 History and Definition

Data science has been called the "sexiest job of the 21st century" by Harvard Business Review [insert DJ Patil reference], but what is it? As with all rapidly growing fields, the definition depends on who you ask. Before we give our definition, however, we provide a brief history for context.

To begin, we note that there was a movement in early computer science to call their field "data science." Chief among the advocates for this viewpoint was Peter Naur, winner of the 2005 Turing award. This viewpoint is detailed in the preface to his 1974 book, "Concise Survey of Computer Methods," where he states that data science is "the science of dealing with data, once they have been established." From his perspective, this is the purpose of computer science. This viewpoint is echoed in the statement, often attributed to Edsger Dijkstr, that "Computer science is no more about computers than astronomy is about telescopes."

Interestingly, a similar movement arose in statistics, starting in 1962 with John Tukey's statements that "Data analysis, and the parts of statistics which adhere to it, must … take on the characteristics of science rather than those of mathematics" and that "data analysis is intrinsically an empirical science." This movement culminated in 1997 when Jeff Wu proposed during his inaugural lecture, upon becoming the chair of the University of Michigan's statistics department, that statistics should be called data science.

These two movements came together in 2001 in William S. Cleveland's paper "Data Science: An Action Plan for Expanding the Technical Areas in the Field

of Statistics." In this highly influential monograph, Cleveland makes the key assertion that "The value of technical work is judged by the extent ot which it benefits the data analyst, either directly or indirectly."

[FOOTNOTE: It is worth noting that these two movements were connected by substantial work in the areas of statistical computing, knowledge discovery, and data mining, with important work contributed by Gregory Piatetsky-Shapiro, Usama Fayyad, and Padhraic Smyth among many others.]

Putting this history together, we provide our definition of **data science** as: The intersection of statistics, computer science, and industrial design. Accordingly, we use the following three definitions of these fields:

- **Statistics**: The branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.
- **Computer Science**: Computer science is the study of processes that interact with data and that can be represented as data in the form of programs.
- **Industrial Design**: The professional service of creating and developing concepts and specifications that optimize the function, value, and appearance of products and systems for the mutual benefit of both user and manufacturer.

Hence data science is the production of useful things through the collection, processing, analysis, and interpretation of data.

2.2 Workflow

A schematic of a data scientific workflow is shown in Figure 2.1. Each section is described in greater detail below.

2.2.1 Data Preparation

2.2.1.1 Inspect

Goal: Gain familiarity with the data Key Steps: Learn collection details Check data imported correctly Determine data types Ascertain consistency and validity Tabulate and compute other basic summary statistics Create basic plots of key variables of interest

2.2.1.2 Clean

Goal: Prepare data for analysis Key Steps: Remove/correct errors Make data formatting consistent Organize text data Create tidy data (one observation per row) Organize data into related tables Document all choices

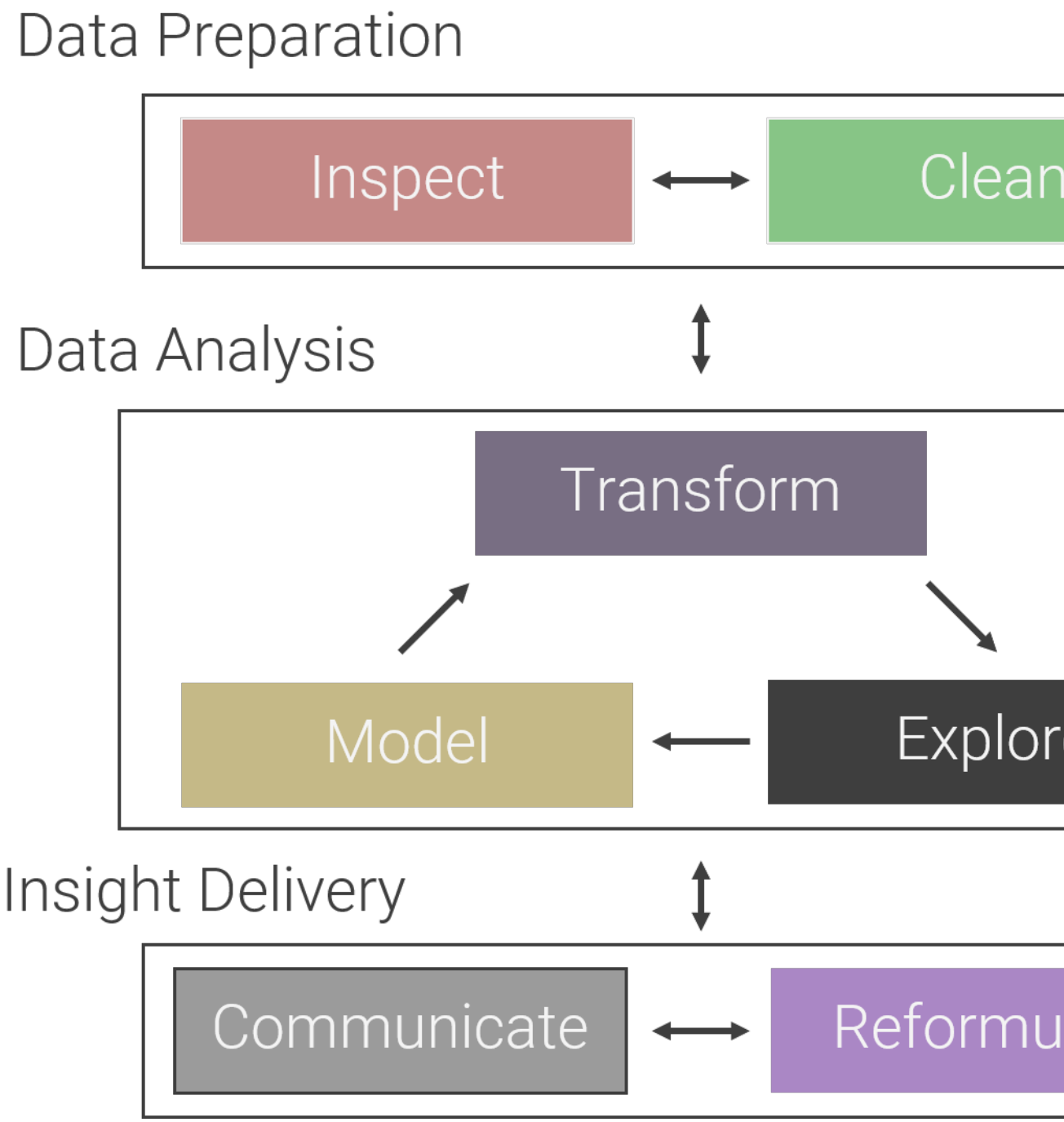

Figure 2.1: Data scientific workflow.

2.2.2 Data Analysis

2.2.2.1 Transform

Goal: Adjust data as needed for analysis Key Steps: Create secondary variables Decorrelate data Identify latent factors Engineer new features

2.2.2.2 Explore

Goal: Allow data to suggest hypotheses Key Steps: Graphical visualizations Exploratory analyses Note: Caution must be taken to avoid high false discovery rate when using automated tools

2.2.2.3 Model

Goal: Conduct formal statistical modeling Key Steps: Conduct traditional statistical modeling Build predictive models Note: This step may feed back into transform and explore

2.2.3 Insight Delivery

2.2.3.1 Communicate

Goal: Exchange research information Key Steps: Automate reporting as much as possible Share insights Receive feedback Note: Design principles essential to make information accessible

2.2.3.2 Reformulate

Goal: Incorporate feedback into workflow Key Steps: Investigate new questions Revise communications Note: Reformulation make take us back to data cleaning

2.3 Benefits of Data Science

2.3.1 Reproducible Research

- Time savings
- Collaboration
- Continuous improvement

2.3.2 Data-Driven Decision Making

2.3.3 Standardized Data Collection

2.3.4 Standardized Reporting

- Especially valuable when there are multiple sites globally

2.3.5 Improved Business Impact

2.4 How to Learn Data Science

Learning data science is much like learning a language or learning to play an instrument - you have to practice. Our advice based on mentoring many students and clients is to get started sooner rather than later, and to accept that the code you'll write in the future will always be better than the code you'll write today. Also, many of the small details that separate an proficient data scientist from a novice can only really be learned through practice as there are too many small details to learn them all in advice. So, starting today, do your best to write at least some code for all your projects. If a time deadline prevents you from completing the analysis in R, that's fine, but at least gain the experience of making an RStudio project and loading the data in R. Then, as time allows, try to duplicate your analyses in R, being quick to search for solutions when you run into errors. Often simply copying and pasting your error into a search engine will be enough to find the solution to your problem. Moreover, searching for solutions is its own skill that also requires practice. Finally, if you are really stuck, reach out to a colleague (or even the authors of this book) for help

2.5 How to Use This Book

We recommend following the instructions in Chapter 3 to get started.

2.6 Common Data Science Tools

2.7 Why R?

For sensory and consumer scientists, we recommend the R ecosystem of tools for three main reasons. The first reason is cultural - R has from its inception been oriented more towards statistics than to computer science, making the feeling of programming in R more natural (in our experience) for sensory and

consumer scientists than programming in Python. This opinion of experience is not to say that a sensory and consumer scientist shouldn't learn Python if they are so inclined, or even that Python tools aren't sometimes superior to R tools (in fact, they sometimes are). This latter point leads to our second reason, which is that R tools are typically better suited to sensory and consumer science than are Python tools. Even when Python tools are superior, the R tools are still sufficient for sensory and consumer science purposes, plus there are many custom packages such as SensR, SensoMineR, and FactorMineR that have been specifically developed for sensory and consumer science. Finally, the recent work by the RStudio company, and especially the exceptional work of Hadley Wickham, has lead to a very low barrier to entry for programming within R together with exceptional tools for data manipulation.

We continue our discussion of getting started with R in the next chapter.

Chapter 3

Getting Started with R

3.1 R

3.2 RStudio

3.3 Git

3.4 GitHub

Data Scientific Workflow

Chapter 4

Example Project

4.1 Background

4.2 Other details

4.3 Conclusions?

Chapter 5

Data Preparation

5.1 Importation

5.2 Organization

5.3 Inspection

5.4 Manipulation

5.5 Cleaning

Chapter 6

Data Analysis

6.1 Transformation

6.2 Exploration

6.3 Modeling

Chapter 7

Data Visualization

7.1 Principles

7.2 Table Mechanics

7.3 Chart Mechanics

7.4 Examples

Chapter 8

Insight Delivery

8.1 Design principles

8.2 Scientific inquiry vs storytelling

8.3 Research reformulation

8.4 Interactive reporting

Reproducible Research

Chapter 9

Tools for Collaboration

9.1 Principles

9.2 Tools

9.2.1 GitHub

9.2.2 R scripts

9.2.3 RMarkdown

9.2.4 Shiny

9.3 Documentation

9.4 Version control

9.5 Online repositories for team collaboration

9.6 Building a code base

9.6.1 Internal functions

9.6.2 Packages

Chapter 10

Automated Reporting

10.1 Excel

10.2 Word

10.3 PowerPoint

10.3.1 Charts

10.3.2 Tables

10.3.3 Bullet Points

10.3.4 Images

10.4 HTML

Additional Topics

Chapter 11

Machine Learning

11.1 Concepts and general workflow (training/test)

11.2 Unsupervised learning

11.2.1 Cluster analysis

11.2.2 Factor analysis

11.2.3 Principle components analysis

11.2.4 t-SNE

11.3 Semisupervised learning

11.3.1 PLS regression

11.4 Supervised learning

11.4.1 Regression

11.4.2 K-nearest neighbors

11.4.3 Decision trees

11.4.4 Black boxes

11.4.4.1 Random forests

11.4.4.2 SVMs

11.4.4.3 Neural networks

11.5 Predictive modeling

Chapter 12

Text Analysis

12.1 Data import

12.1.1 Data sources

12.1.2 Tokenizing

12.1.3 Lemmatization, stemming, and stop word removal

12.2 Analysis

12.2.1 Frequency counts and summary statistics

12.2.2 Word clouds

12.2.3 Contrast plots

12.2.4 Sentiment analysis

12.2.5 Bigrams and word graphs

Chapter 13

Graph Databases

Conclusion

Chapter 14

Conclusion

Appendices