Contemporary Data Analysis: Survey and Best Practices

Materials to Accompany Lecture 1: the Field and the Data

Valentina Kuskova, PhD

Contents

Welcome!

welcome to the course! As this is a review course, it provides a lot of overview - and not a whole lot of hands-on work. Therefore, for every lecture topic we cover, I provide additional materials to help you get more familiar with the topic. These materials mostly include references, suggestions for individual study, occasionally - interesting examples. I also provide as set of exercises in the special section of the file, which I call “Challenge yourself.” The exercises provided will help you prepare for the final project.

Come back to this file as you are watching the lecture. For many items that I discuss, here I provide full references so you can explore the topic further.

So, let’s get started!

Success tips

This is an unusual course. The six-week open course will focus on deep theoretical concepts that are paramount to the understanding of the field and success in analytics. The additional topics for Master’s students will be a bit more hands-on, with examination of special topics in more detail.

For both parts of the course, you need to have an open mind, so that you can absorb all the possibilities.

1. Do not rely on lecture material only. Get a notebook and write down the concepts that are most interesting or applicable to you.
2. When you see the topic that you like, explore it further. This file and other files for this course contain the details of the code (if applicable), explanations, references, and most importantly – exercises and suggestions for further study (in the "Challenge yourself" section).
3. Most of them are optional, but I hope you take advantage of what is offered.

Learning objectives

The goal of Lecture 1 is to provide the broad overview of the data analysis field and the two major components it consists of: the data and the analysis. Topics within the first lecture explain how these two concepts fit together. We start with the definitions to clear some of the confusions with terminology in the field. Then, we discuss the contents of this course and map the field of data analysis. We also discuss the role that data play in our lives, what the data are, their types and classifications, and sources of data. We finally address the issue of modeling – why we model and how analytics aids decision-making in business and real life.

Recommended literature

I think it would be fair to say that one video lecture is not enough for any course. Therefore, I put together a list of references that you can get started with if you want to explore this topic further.

1. Bacci, S. and Chiandotto, B., 2019. Introduction to Statistical Decision Theory: Utility Theory and Causal Analysis. CRC Press.
2. Caplin, A. and Schotter, A. eds., 2008. The foundations of positive and normative economics: a handbook. Oxford University Press.
3. Donoho, D., 2017. 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), pp.745-766.
4. Heilbron, J.L., Ed. The Oxford Companion to the History of Modern Science. Oxford University Press, New York, 2003.
5. Hey, A.J. ed., 2009. The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.
6. Mallows, C., 1998. The zeroth problem. The American Statistician, 52(1), pp.1-9.
7. Parmigiani, G. and Inoue, L., 2009. Decision theory: Principles and approaches (Vol. 812). John Wiley & Sons.
8. Peterson, M., 2017. An introduction to decision theory. Cambridge University Press.
9. Principles of management: the course. https://courses.lumenlearning.com/principlesmanagement/

Lecture support materials

This section consists of two components: references for the images, examples, etc., that I have used in the lecture, and additional reading materials that you can review in order to get even deeper into the topic.

Images, examples, references

I have used quite a few examples and images in my lecture. It's only fair you have the references for all of them. For your convenience, the order of references in this section follows the lecture.

1. The battle of definitions. There are many examples of definitions of data science. Here are a few I've found interesting.
    - Source: https://financesonline.com/data-analytics-data-science-how-do-they-differ/
    - https://www.pinterest.ie/pin/361976888797689900/?autologin=true
    - https://datasciencedegree.wisconsin.edu/blog/data-science-vs-data-analytics-the-differences-explained/
    - https://adwiteeya.medium.com/data-analysis-vs-data-analytics-a08c0fc4603c
    - https://datainsightgroup.ca/data-scientist-be-title-question/

- https://towardsdatascience.com/what-is-data-science-and-what-is-it-not-c6a09d735f02
- https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/
- https://faircute.com/2020/10/18/history-of-data-science/
- https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
- https://programmersought.com/article/65296869369/

2. Communication on Building a European Data Economy. https://ec.europa.eu/digital-single-market/en/news/communication-building-european-data-economy
3. The Forbes approach to data. https://www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/?sh=300e80623362
4. Data credibility problem. https://hbr.org/2013/12/datas-credibility-problem
5. Sixteen reasons to model. http://jasss.soc.surrey.ac.uk/11/4/12.html
6. Validity, reliability, generalizability. https://www.healthknowledge.org.uk/content/validity-reliability-and-generalisability
7. Falsifiability. https://scienceornot.net/2012/01/17/scientific-models-are-falsifiable/

Challenge yourself

A lot of what we have talked about was theory. But guess what? If you are ready to challenge yourself, there are tasks that can be done. In the materials folder for this lecture, you will find the "Transportation Data" file in the Excel format and the codebook to accompany this file.

It is a secondary dataset from the Nationwide Personal Transportation Survey conducted in 1995 in the United States. It has quite a few data challenges. Before attempting an assignment, familiarize yourself with the codebook.

Here are a few things you can do to get started:

- Take the first 10-15 variables in the dataset. How would you classify each variable? If you think it is a numerical variable, what is its measurement level?
- Look carefully at the data. Do you think you can build models with it? If so, which questions, do you think, this data can help you answer?
- Would the models you build from this study be generalizable? Why or why not? What would you need to do to ensure generalizability of the models you can build from this dataset?

Acknowledgements

Unlike some of the topics in data analytics, such as *networks*, which I have learned relatively recently, and where I can pinpoint to who taught me what, data analytics in general is a very old topic for me. I have now taught it for over twenty years - and imagine how long ago I've started on the learning journey! Many courses were taken on the way, many programs completed, may professors I had something to learn from.

However, I want to say special thanks to professor Rob Turrisi (https://hhd.psu.edu/contact/robert-turrisi), who introduced me to the fascinating field of statistics first, then analytics, while I was an undergraduate psychology student at Boise State University, some time in the last century. It was his incredible passion for data and analysis, amazing approach to teaching, and never-ending enthusiasm for anything and everything new in analytics, that got me first interested, then involved in this subject.