ISSN 2688-6170

kcFED Technical BRIEFINGS

Discussing Data: The Elevator Pitch

Thealexa Becker

May 2018
TB 18-03
https://doi.org/10.18651/TB/TB1803

FEDERAL RESERVE BANK of KANSAS CITY

Discussing Data: The Elevator Pitch

By Thealexa Becker[1]

Abstract

Researchers and custodians of data are often tasked with explaining their data to unfamiliar audiences. There exists a knowledge gap that can be challenging, particularly when attempting to communicate research that uses complex or less well-known data. This paper describes an "elevator pitch" that can be used to quickly and efficiently present core characteristics of data. Specifically, it outlines seven core characteristics that can be used to create such a pitch. Additionally, it provides examples of how this compact description can be integrated into research papers, presentations, or collaborative conversations to foster a consistent standard of data description.

I. Introduction

Researchers and data custodians are often faced with the challenge of effectively and efficiently communicating about their data. With the increasing focus on analytics and data-driven research, narrowing this information gap between the researcher or analyst and their audience is paramount. While thorough data descriptions can be found in technical papers or documentation, many times a more succinct overview is more appropriate. Often, descriptions of data to an audience fall into one of two traps that hamper understanding: too much information, or too little. There must be a balance between inundating the audience with every piece of information collected about the data and providing scant details that result in more questions.

Research or analysis using data, presentations about data, and collaborative meetings discussing data would benefit from an "elevator pitch" about the data. This pitch would include key information about the data that would foster both a better understanding of the work using the data and enable more directed questions about the data and its use. The goal of an elevator pitch is simple: get the audience to come away with a general understanding of the structure, contents, size, and use of the data.

This paper describes and demonstrates a template for creating such an elevator pitch. The pitch is be general enough to apply to a wide array of data, but allows enough flexibility for increased specificity where appropriate. The pitch is designed to be integrated into papers, reports, or presentations which focus primarily on one or two datasets[2]. This paper outlines the core characteristics of data that comprise an elevator pitch, and provides an example of this process.

---

[1] Data Scientist, Thealexa.Becker@kc.frb.org, Center for the Advancement of Data and Research in Economics, Federal Reserve Bank of Kansas City. I would like to thank Brett Currier and San Cannon for their essential feedback during the formation of this idea into a paper. There will be an additional and separate file for downloadable worksheet.

[2] Future work will cover suggestions for providing insight to the audience when three or more datasets are being discussed.

II. Creating an elevator pitch

What goes in the elevator pitch

The goal of an elevator pitch is to promote an individual, a company, or a product in a short amount of time. If asked to pitch a dataset, the goals are to promote its use, the audience's understanding, or both. Although an experienced user could likely produce detailed information about data on command, what information is relevant in order to efficiently and effectively communicate the core characteristics of the data to encourage use and understanding?

Efficiently and effectively describing a dataset is key. Although the core characteristics presented below look like the outline of a metadata schema, this paper does not suggest these characteristics could replace a thorough documentation of metadata but rather they serve to educate a broad audience about the basics of a dataset. To effectively describe a dataset to as general an audience as possible, all seven of the characteristics below would be present. To efficiently describe a dataset, sticking with these seven characteristics, particularly in oral presentations where patience for long descriptions is scarce, allows the presenter to level-set the audience's understanding without inundating them with information.

Sample population or universe
The sample population or universe refers to who, what, or where the data are being collected from. No explanation of a dataset is complete without telling the audience who or what the subject of the data is.

Collection method
The collection method refers to how the data were assembled. Data are either collected directly from a subject, generated by a process, or created by formatting information from a source. This information lends further context to the sample population or universe by providing the audience with an explanation with how the data were gathered.

Frequency and timeframe
The frequency is how often the data are collected. The timeframe is the span of time data have been collected from the source. Not all data have a timeframe, as some data, particularly in some branches of science, are collected from a source only once. Many audiences may want to know "how far back" the data go as a way of placing the data in temporal context.

Notion of dimension
The dimensions of the data - the number of observations and variables (or features) - provide the audience with an idea of the shape and detail of the data. This can be important when a dataset has a unique or uncommon shape. When incorporating this into an elevator pitch for a general audience, it may be more clear to the audience to avoid referring to the dimensions as "rows and columns" as that term is tightly tied to format rather than content.

Purpose and main content
Data are collected or constructed for a reason. Telling the audience about the general purpose for the collection of and main content of the data can further establish the relevance of the dataset.

Access and use information
Potential data users in the audience might want to know how they could use the data for their own work. Therefore, it would be helpful to alert the audience about terms of use or other conditions for working with the data, or if the data are easily or publicly available.

Producer or publisher of data
All data have owners and creators. Giving the audience the identity of the producer or publisher of the data is akin to citing a source and, therefore, lending credibility to the data. Moreover, providing the producer or publisher helps to establish the provenance of the data in fields where this is a concern.

What does not go in an elevator pitch

In general, the information in a data elevator pitch is intended to be relevant to a wide range of audiences. Below are a few examples of information that might be tempting to include but that refer to a specific project or interest only a subset of possible audience members.

Details relating to a project
Try not to bury a description of the data with the results or specifications of a particular analysis. The elevator pitch is a foundational element of presenting or explaining your research. Explaining the data in general terms and presenting results simultaneously can stymie the audience's understanding of both topics. It it is worth considering crafting a separate summary of a particular project to accompany the elevator pitch as a way to succinctly state the scope of the work being presented.

Technical specifications
Working with data requires technology. The specifics of which software or hardware were used in the analysis is important, but it might not promote understanding of the data. Notions of the disk space needed to store the data and the data type for variables are relevant information for some audiences, but not all, making them unnecessary in a pitch. If technical details are imperative for discussing a project, it could be easier for the audience to hear that information separately.

Literature reviews
Mentioning other work that uses the data may be important information to include in a paper or presentation, but it does not fit in an elevator pitch. It may be useful refrain from listing other work the audience might not be familiar with as this makes an assumption that they know why that work is important. The literature review is then best saved for discussing a specific project and not the data. An exception to this suggestion is if the academic work produced the dataset in question.

Opinions
Are these data terrible to work with? Is the documentation written in impenetrable techno-babble? The answers to these and many other common complaints about working with the data, the construction and collection of the sample, and the validity of the responses are all important to address to various audiences. But they are often not objective, may be too technical, and may not further an audience's understanding of the core characteristics of the data.

Detailed metadata

Metadata are extremely important in certain fields and occupations, but many audiences may not even be aware of how to properly define the term. Providing a detailed metadata profile of a particular dataset might feel like an effective way to provide your audience with as much organized detail as possible, but it is not efficient, nor is it effective in improving understanding.

There is a final consideration: just because information is not useful in the elevator pitch does not mean it isn't useful at all. For instance, it is not necessary to indicate whether housing price data includes region, state, metropolitan area, and zip code information, when the information can be referred to as "geographic", in the elevator pitch. However, should a subsequent question arise about the level of detail of the geographic information, a presenter can be prepared to give a more detailed answer. In general, presenters may want to be prepared to answer a slate of broad questions from the audience asking for additional detail about the data in general, and the subset of data you used in your analysis specifically. Once again, the purpose of the elevator pitch is not to answer every anticipated question from an audience about the data but to provide an entry point for them to engage with research or analysis.

Examples

The following section provides a case study for crafting an elevator pitch using the Current Population Survey (CPS)[3]. The CPS is a much cited and widely used dataset in labor economics that is both complex, but easy to characterize. The examples below provide an elevator pitch that gives the appropriate amount of information, a pitch that is far too vague, and an overly detailed explanation that will confuse audiences.

The Good

Below is an example of how one might describe the CPS by providing all seven core characteristics of the data. Notice that this example can be spoken aloud in fewer than 45 seconds. It is also not overly detailed and provides relevant avenues for further questions about the data.

> The Current Population Survey (CPS) is a monthly address-based survey of U.S. households that gathers geographic, demographic, and employment status information. These data are used to calculate several national indicators of employment, such as the unemployment rate and labor force participation rate. In its current iteration, data on over 150,000 respondents are collected monthly to create around 400 variables. The CPS survey data are collected by the Census on behalf of the Bureau of Labor Statistics and are publically available dating back to 1976.

[3] Information about and access to the Current Population Survey data can be found in Kansas City Federal Reserve's Data Museum located at: https://www.kansascityfed.org/research/datamuseum/cps

The Bad

Below is another description of the CPS, however, this version is far too vague and will elicit clarifying questions that stall a presentation or discussion because important information was not provided.

> The CPS is a labor force survey that generates unemployment numbers for the BLS. The data go back more than 30 years and also have a lot of demographic information about respondents.

The Ugly

Finally, there is a description of the CPS that violates guidelines about providing too much information. What results is a messy, confounding statement that drags the audience along without providing a clear understanding of the subject. In this version, the information provided is not bad or even irrelevant – it just isn't necessary in the elevator pitch and could be repackaged more appropriately later in a presentation or paper.

> The Current Population Survey is a monthly address-based survey that interviews U.S. households over a period of 16 months where they answer questions for 4 months, are off for 8 months, and then answer questions for 4 months again. Data are collected on state of residence, marital status, education, sex, race, industry, occupation, full-time and part-time work, self-employment, unemployment, and a variety of other topics related to employment status. The Bureau of Labor Statistics, which produces this data with the help of the Census, uses the data to calculate the unemployment rate, labor force participation, the number of discouraged workers, and other labor statistics at the national level that are reported in the monthly Employment Situation. In the 1994 version of the survey to the present, there are at least 150,000 individual respondents per month and the data have around 400 variables you can use for analysis. Earlier versions of the survey had fewer people and fewer variables. Some researchers use this data to study part-time workers, which is what I use this data for as well. Often times researchers will conduct analysis on this data using statistical software such as R, Stata, or SAS, but this can be difficult because the data are huge on disk and can be very messy pre-1994.

The "Good" elevator pitch provides information on all seven core characteristics and can be delivered aloud in a timely manner and even inserted at the beginning of a data section in a paper or report. The latter two examples, the "bad" and the "ugly", fall into common traps that presenters fall into when discussing research or analytics. They provide either too little detail to be effective or too much detail to be efficient and direct.

III. Conclusion

The elevator pitch is tailor made to anchor a section of a paper, presentation, or documentation that discuses data, but may not be sufficient on its own. Incorporating the elevator pitch can act as a meaningful transition to a more detailed discussion about the data or a specific subset of data used in analysis. For instance, when writing about a project, lead with a

version of the elevator pitch before segueing into a more detailed description of the data subset used for the analysis. Or, use the elevator pitch as a way to get the audience on the same page before beginning a more detailed discussion of the data themselves. The elevator pitch is particularly useful in presentations, where there is often less time to discuss details of the data than in writing. In such a case, the elevator pitch can concisely provide general information about the data and allow the presentation to progress.

Explaining data to an audience that is unfamiliar with them can be a daunting task. However, using the elevator pitch, a researcher, data producer, or analyst can efficiently and effectively inform their audience about data used for analysis. The elevator pitch suggests isolating the seven core characteristics of a dataset and crafting them into a paragraph that can be read in 30-45 second during a presentation or inserted into a paper. The appendix of this paper includes a data profile worksheet that can help outline the process of creating a pitch for any data set. Presenting data in this way can facilitate understanding of subsequent material and foster greater interest and engagement with research and analysis.

Appendix: Data Profile Worksheet

This worksheet is designed to be used for each individual dataset a user has that needs to be described succinctly. Note that this is not intended to be used to provide all the information on a dataset, just enough information for an audience/client/customer would need to get the gist.

Name of data:

Sample population or universe

Who or what are the subject of the data?

Collection mechanism

How was the data collected from the subject?

If the data were collected via survey, how was the survey administered?
If the data were created from existing information, how?
If the data were scraped from a source, how was this done?

Frequency and timeframe

How often was data collected?

How long has data been collected?

Notion of dimension

How many observations of data are gathered each time data are collected?

How many categories/questions are data collected about?

What format are these data (ie. String, number, categorical, word)?

Purpose and main content

Why were these data collected?

What are the major topics covered in these data?

Access and use

Who can access these data?

How can these data be accessed?

Producer and publisher

Who owns these data?

Who collected or created these data?

Other important notes:

Write the Elevator Pitch!