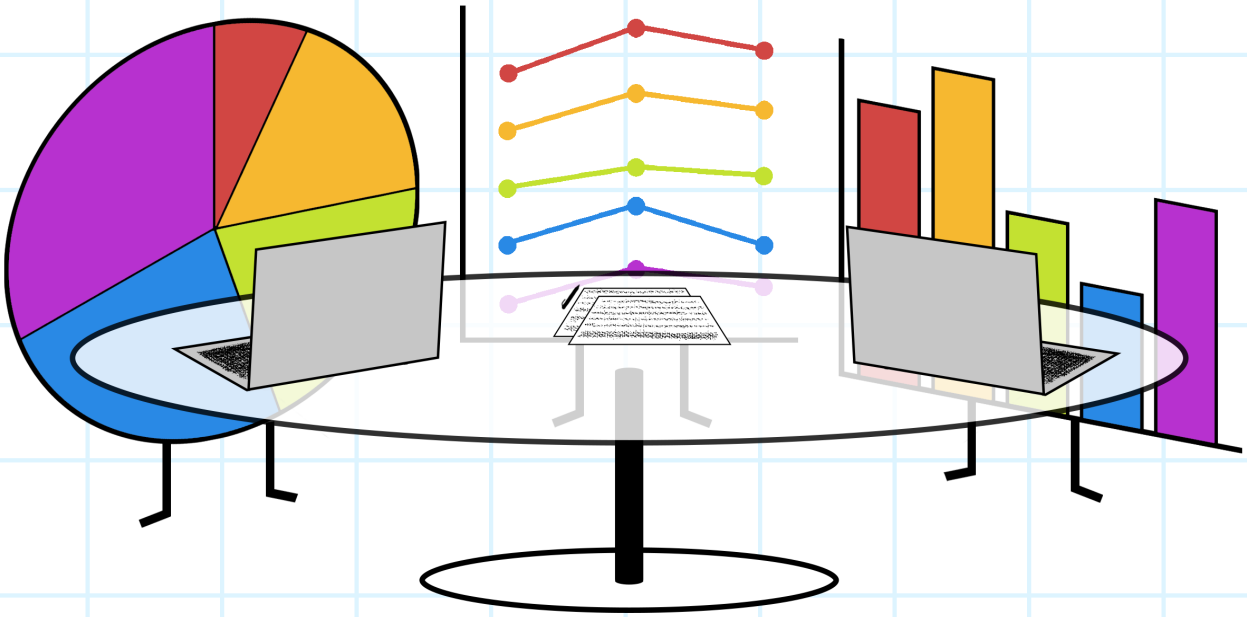
The Data Science Salon



Roger D Peng, Elizabeth Matsui, and Corinne Keet

The Data Science Salon

A Collaborative Learning Experience

Roger D. Peng, Elizabeth Matsui and Corinne Keet

This book is for sale at http://leanpub.com/thedatasciencesalon

This version was published on 2018-06-22

Special thanks to Maggie Matsui, who created all of the artwork for this book.

Contents

Preface

The Data Science Salon was borne out of our desire to develop a different approach to learning data science and data analysis. We have all taken traditional statistics and data analysis courses in a university setting and found them to be useful. In particular, these didactic courses are important to learning fundamental concepts and theories. However, we all felt that our true skills in data analysis were developed while working in teams, digging through data to tackle difficult scientific problems, and communicating our findings to diverse audiences. In retrospect, we found this to be a powerful way to learn data science and not one that is easily replicated in a traditional classroom setting.

The Data Science Salon provides the collaborative learning experience of working in teams to solve a data science problem of mutual interest. You will bring together a team (e.g. fellow employees, colleagues, friends, collaborators) to gather at regular intervals to work through a problem in stages. In the process, you will go through the process of data science and learn about the various stages of that process. *The Data Science Salon* provides objectives for each of your regular meetings so that you can advance your progress at each stage and learn more about your data. At the end of the your Salon, you will have a report or slide deck that summarizes your analysis of the data and any conclusions you can draw. This report may be of use to you in your company or organization.

The process that we have put forward in *The Data Science Salon* is oftentimes messy and maybe even frustrating. However, data analysis *is* messy at times. The hallmark of a good data scientist is the ability to work through problems as they come up, develop reasonable solutions or workarounds, and move forward in a disciplined way. We feel strongly that the approach of *The Data Science Salon* is a useful mirror of a real world experience and will result in a product that will be of value to you.

Roger Peng
Elizabeth Matsui
Corinne Keet

Baltimore, MD

Session 1

Welcome to DS² Session 1. In this session, your salon will get organized and oriented to the DS² series. The objectives and accompanying supportive materials are listed below.

First, you'll want to identify your salon members and their expertise and roles. At least one person should be responsible for scheduling the sessions and keeping the group together. We sometimes refer to this person as the chief cat herder! At least one person should have some technical skills so that the group has the ability to acquire and manipulate data, and construct plots with some statistical software package. One person should have some subject matter knowledge. The amount of subject matter knowledge needed will depend, in part, on the data science question that your salon chooses. If your salon's question is related to a highly technical area such as astronomy, then more subject matter knowledge would be needed than if your salon's question is related to a general interest subject, such as a lifestyle-related topic (e.g. restaurant reviews, twitter trends).

If you have not read Chapter 2 of *The Art of Data Science* and watched the video lectures prior to your first session, then the first order of business will be for the group to do these things so that everyone has the framework for proceeding to the main activity for Session 1.

Objectives

Identify the expertise and roles in your salon

- Leadership & organization
- Technical skills: Getting and manipulating data and doing basic statistical analysis and graphing, using statistical package (SAS, Stata, R, etc.)

- Subject matter knowledge

Brainstorm a topic of interest and identify sources of data

Organizing the Team

There are two main activities for this session. The first activity is for the group to think about what roles group members will play and to identify the expertise in the group. This will help ensure that you have the right mix of expertise for a successful salon series.

Two key areas you will need to cover are

- **Subject matter knowledge**: It's useful if someone in the group has some familiarity with the subject are of the problem you'll be working with in order to provide context and aid in interpretation
- **Technical skills**: Are there people in the group who can get the data and manipulate it in a statistical package like SAS, Stata, R, or Excel

Once you've got a topic of interest and have identified the expertise in the room, think about what kind of question you want to ask and what kinds of datasets would help you answer the question. Discuss generally how you might approach the problem and try to anticipate what challenges you might run into.

Brainstorming Ideas

The second activity is to brainstorm areas of interest for your group and identify sources of data for the area of interest. Some of you may have assembled because you have common work or recreational interests, so this common interest could serve as a starting place for your brainstorming. For other groups, there may be a common interest in learning or refining data science skills, so an area of interest of one of the salon members or that is of general interest to the group could be appropriate. For groups

whose area of interest is work-related, your organization may have the data that you plan on using to address your question. For other groups, an individual member may have data to share or you can take advantage of the wealth of publicly available datasets.

Resources

The Art of Data Science Chapter 2 – Epicycles of Analysis

Lecture video on Epicycles of Analysis[1]

Lecture video on The Stages of Data Analysis[2]

Publicly Available Data

Here are some links to publicly available data that may be of use to you as you think about different questions to consider in your Data Science Salon. We also include some links to health-related publicly available datasets that we have worked with.

List of Awesome Public Datasets[3] on GitHub

The National Health and Nutrition Examination Survey[4]

The National Health Interview Survey[5]

Related Links:

Psychological safety and effective teams[6]

[1] https://youtu.be/hSEIjho5tKM
[2] https://youtu.be/bKWc7vrc-fc
[3] https://github.com/caesar0301/awesome-public-datasets
[4] http://www.cdc.gov/nchs/nhanes.htm
[5] http://www.cdc.gov/nchs/nhis.htm
[6] http://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html?action=click&module=MostEmailed&region=Lists&pgtype=collection

Data Tsunami: Ready to be Surfed[7]

Rethinking the Inner City Asthma Epidemic[8]

The Automation of Big Data – Not So fast?[9]

False Alarms about National Crime Wave[10]

Estimating Deaths from Volkswagon Emissions Scandal is Hard[11]

The Art of Data Science[12]

[7] http://skybrudeconsulting.com/blog/2014/12/10/milk-allergy.html
[8] http://skybrudeconsulting.com/blog/2015/01/25/inner-city-asthma-rethink.html
[9] http://skybrudeconsulting.com/blog/2014/12/17/automation.html
[10] https://twitter.com/eliza68/status/669492943225561088
[11] https://twitter.com/RoyalStatSoc/status/674901318880903168
[12] https://leanpub.com/artofdatascience

Session 2

Welcome to DS² Session 2! In this session, your salon group will:

- State and refine your data science question, critically evaluate it, and
- Create a few slides (or other report) of the background and rationale for your question and that state your question

You will apply the epicycle of analysis from Session 1 to the process of stating and refining your question so that at the end of the session you have a sharp question. The objectives and accompanying supportive materials are listed below. If you have not read Chapter 3 of *The Art of Data Science* and watched the video lectures prior to this session, then the group should do these things before proceeding to the activities for Session 2.

Your group should have identified a subject area and potential sources of data in Session 1. Building on this and using information learned from TAODS and the video lectures, your group should state a question and then refine it so that it meets the characteristics of a good question.

- Your group should go through an iterative process of first stating a question, which will likely be a more general question, critically evaluating it, and then refining it based on your evaluation so that you end up with what your group believes is a specific, sharp question that is answerable with the data available to you.
- Your group will also need to identify the type of question it has identified, if the question is not an inferential question about understanding relationships between two factors, your group will need to come up with a new question that is this type.
- Once your group is satisfied that your question is an inferential question about a relationship between two factors and meets the criteria of a good question, then the second activity of this session will be to create a short report, for example, in the form of 2–3 slides

that provide a background and rationale for the question and state the question. You will build on this report in later sessions so that your group has a complete report at the end of the DS2 series of sessions.

Objectives

1. State and refine your question
2. Critically evaluate your question
3. Develop slides that provide the background and rationale for your question and state your question

Resources

- *The Art of Data Science* Chapter 3 – Stating and Refining the Question
- Lecture video on Types of Questions[13]
- Lecture video on Characteristics of a Good Question[14]

[13] https://youtu.be/B1Eye3MFhhg
[14] https://youtu.be/HvUbZQAZNvk