Introduction to the Course Statistical Data Analysis

Contents

1 Main Objectives

2 Data Analysis

3 The Data Analysis Cycle

Main Objectives

- That students begin to develop the skills to be able to organize, analyse and interpret data.
- To gain insights into data through computation, statistics and visualization.
- To work on the ability to take data- to be able to understand it, to process it, to extract value from it, to visualize it, to communicate conclusions.

Statistics has been the most successful information science. Those who ignore Statistics are condemned to re-invent it.
- Brad Efron, 1997

Main Objectives

- To become informed consumers of the statistical methods/models presented in this course.

- To be able to properly obtain and use the statistical methods presented in this course.

Data Analysis

The process of data analysis is a creative one, it can't be generalized and written down as a recipe to follow. It is not magically finding nuggets of information in a data set. It is, by nature, an amorphous and time consuming-process, full of open-ended questions and following up leads that often reach dead-ends.

> *How can one effectively generalize across many different data analyses, each of which has important unique aspects?*
>
> *The Art of Data Science, R.D. Peng, E. Matsui*

> *We think that good analysis depends not only on clear thinking but also on substantive knowledge. Mere numerology will not do, nor is there a good cookbook.*
>
> *David Freedman, 1987*

Data Analysis

Ultimately, a data analyst must find a way to assemble all of the tools and apply them to data to answer a relevant question—a question of interest to people.

What should be described as data analysis is not a specific "formula" for data analysis—something like "apply this method and then run that test"—but rather is a general process that can be applied in a variety of situations.

The starting point of any data analysis should be a set of questions about the data set you want to answer. Data analysis is a question-driven process. The starting point is a research question of interest.

Data Analysis: research questions

Starting Point: it is usually some research question. Because *doing data science* is basically leverage data to answer questions.

Some examples might be:

- A researcher is interested in understanding the effect of smoking and weight upon the resting pulse (Low/High).
- A dietician is interested in studying the general trends in the composition of cereals.
  - What ingredients best predict the amount of calories a cereal contain?
  - Does service size play a role in the amount of calories?
  - Are the trends in predicting calories the same across manufacturers?

Data Analysis: research questions

- Has there been global warming over the past decade?
- Is having the death penalty available for punishment associated with a reduction in violent crime?
- Does student performance in class depend on the amount of money spent per student, the size of the classes or the teachers' salaries?

NBA Data 2016-2017. Regular Season.

NBA players have the highest salaries in the world, but the NFL spends the most money on players

Cork Gaines and Samantha Lee Nov 27, 2017, 9:19 PM

Sporting Intelligence has released their annual survey of salaries for athletes in the top leagues around the world.

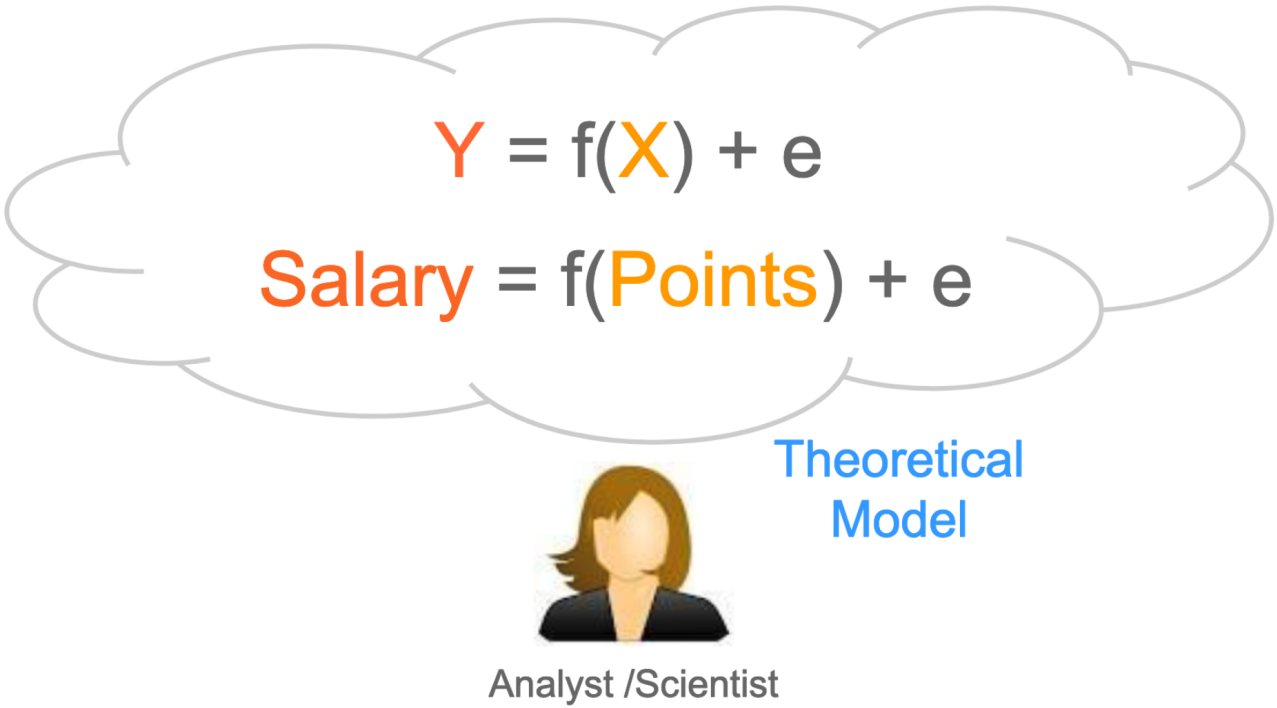
The study looks at the first-team pay for 348 teams across seven different sports in 18 different leagues around the world. The Oklahoma City Thunder are the new highest-paid sports team in the world, with players making an average of $9.3 million during the 2017-18 season.

Data Analysis

Research Question: The more scored points, the higher the salary?

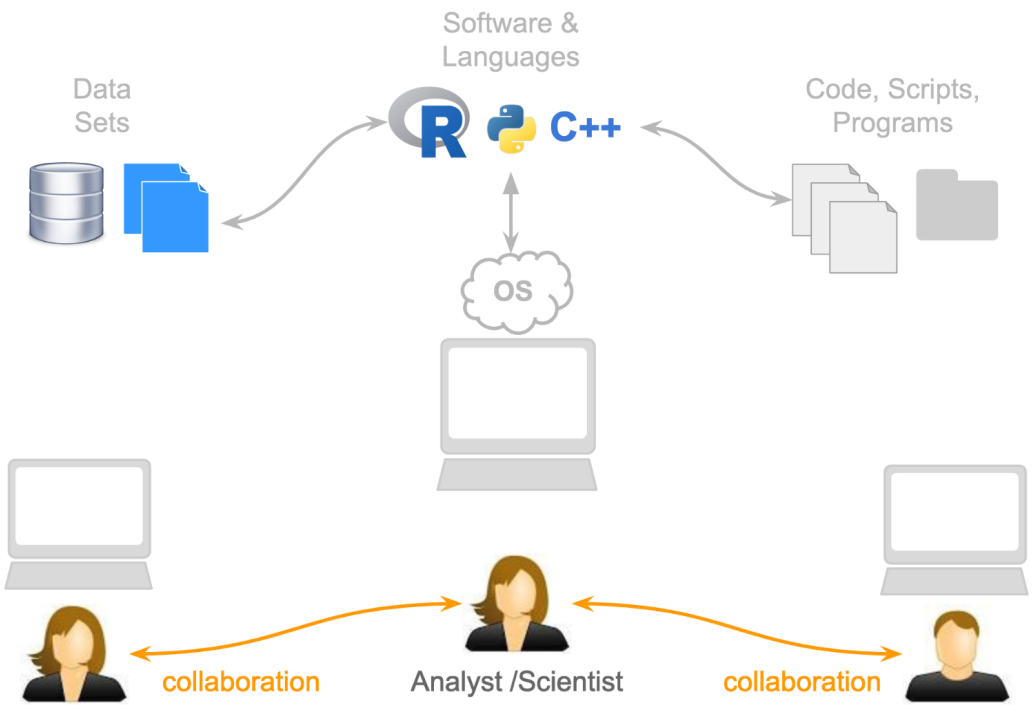
How do analysts and scientists think about the data?

Statistical Perspective



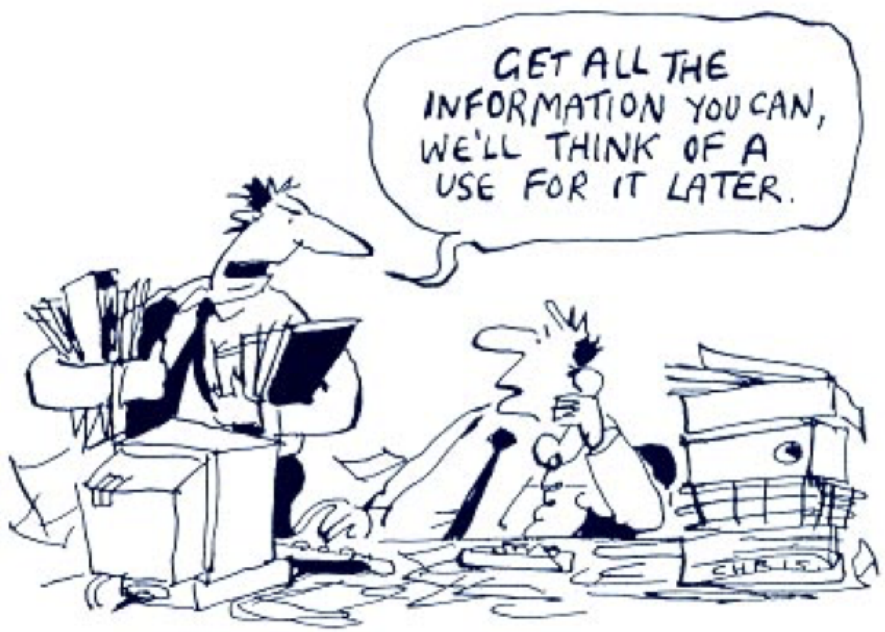
The Big Picture

What about the data? How do computers treat data? How do programming languages handle data?



Statistics vs. Data Science

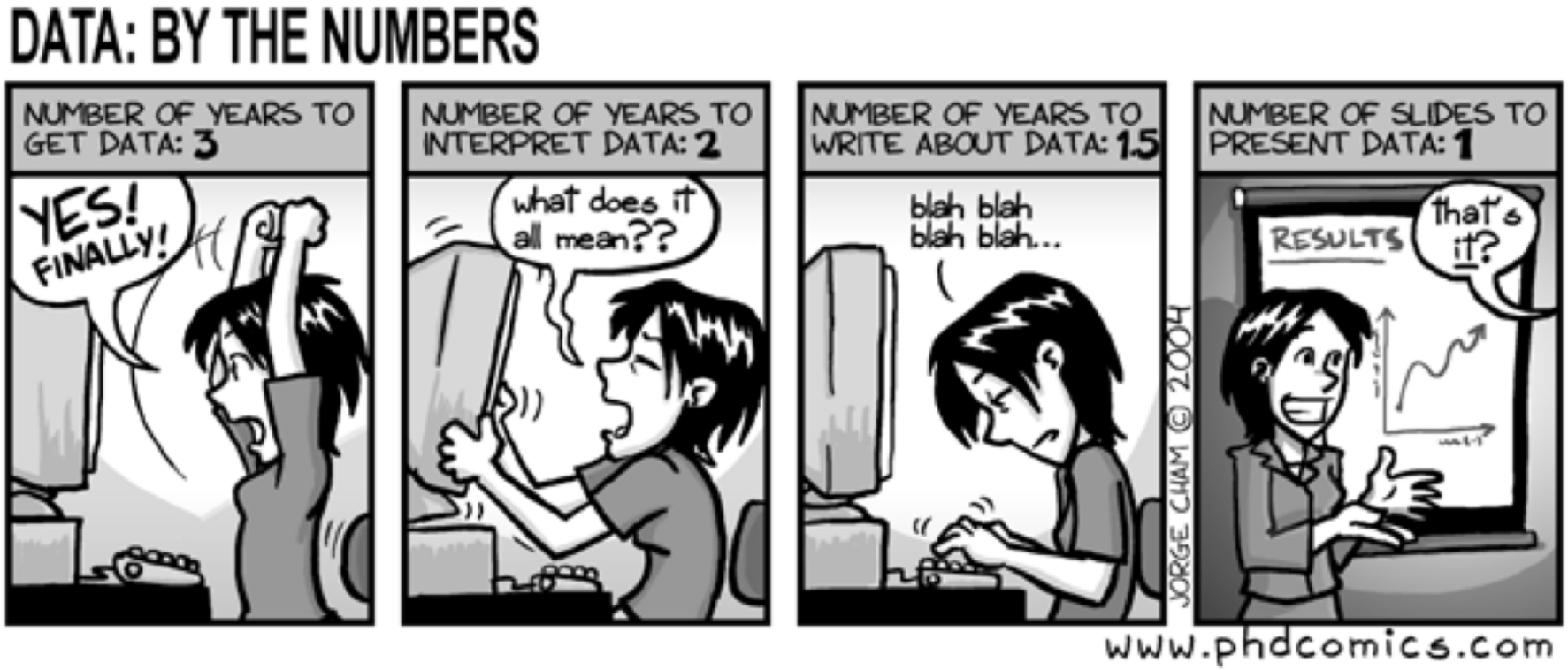
- Statistics traditionally is concerned with analysing experimental data that have been collected to explain and check the validity of specific hypothesis.
- Data Science typically is concerned with analysing observational data that have been collected for other reasons, not under control of the investigator, to create new ideas. Data Science involves The Big Picture.



Statistics

Statistics is the art and science of designing studies and analysing the data that those studies produce. Its ultimate goal is translating data into knowledge. Statistics is the art and science of learning from data.

The Data Analysis Cycle



The Data Analysis Cycle (I)

Data Preparation

- Acquisition
- Storage
- Cleaning
- Processing
- Tidying
- Reshaping
- Wrangling

The Data Analysis Cycle (II)

Core Data Analysis

- Exploration
- Description
- Visualization
- Hypothesis Tests
- Inference
- Simulation
- Model Fitting

The Data Analysis Cycle (III)

Reports

- Document(s)
- Article(s)
- Book(s)
- Poster(s)
- Blog post(s)
- Dissertation
- News

The Data Analysis Cycle (IV)

Communication

- Oral
- Print
- Web
- Audio
- Tidying
- Video
- Other

... not everything is Big Data and a black box

Imagine 10 people with $10,000$ blood pressure measurements on each of their two arms. It looks like $200,000$ observations. But really there are only 10 people. But a black box might not take proper account of the difference between 10 subjects with $10,000$ observations each and $10,000$ subjects with 10 observations each. Those are very different. We hope our statistical methods take account of this.

... be aware of the GIGO principle

It is not possible to carry out an accurate statistical analysis of bad quality or inaccurate data. Read these entries as an example. Flawed data and serious math errors made Tesla Autopilot look better (than it actually was). It also was a badly designed experiment to compare crash per million miles before the autosteering feature was activated and afterward.

- Entry 1
- Entry 2

Computational tool

- The main computational tool will be the computing and programming environment R.
- The main workbench I will be using is the IDE RStudio.