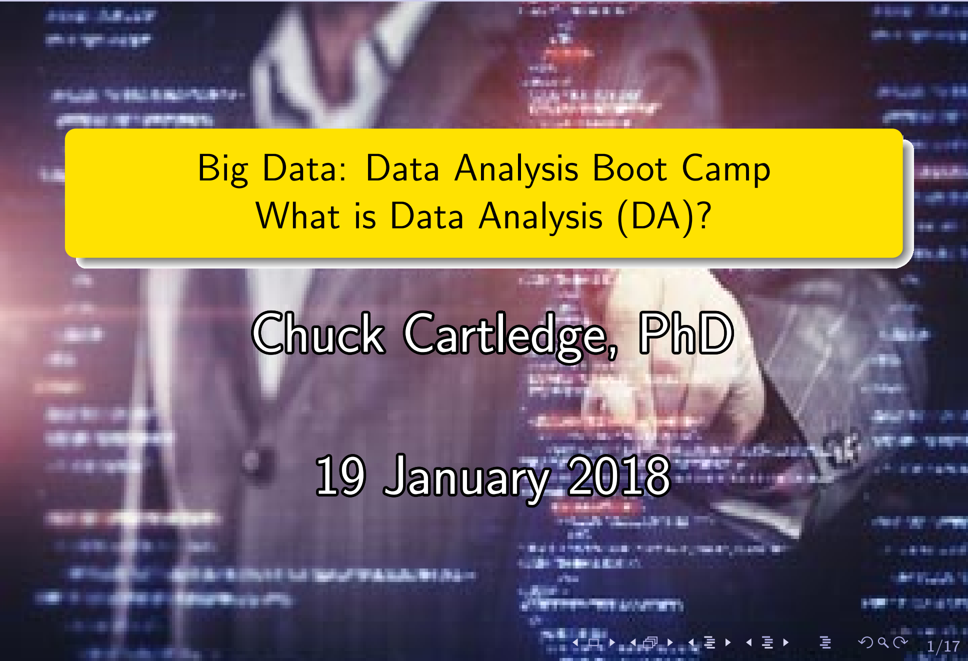


# Big Data: Data Analysis Boot Camp
# What is Data Analysis (DA)?

Chuck Cartledge, PhD

19 January 2018

# Table of contents (1 of 1)

# What are we going to cover?

We're going to talk about:

- What is exploratory data analysis (EDA)?
- How data analysis (DA) can lead us to new insights

# In the beginning . . .

Exploratory Data Analysis (EDA) is looking at data without preconceived ideas or a desire to fit the data to an existing form. EDA is the first step in data analysis

# What are the objectives of EDA?[5]

The objectives of EDA include:

- Uncovering underlying structure and identifying trends and patterns;
- Extracting important variables;
- Detecting outliers and anomalies;
- Testing underlying assumptions;
- Developing statistical models.

We are looking to “understand” the data.

# How do we get there?[5]

The practice of EDA emphasizes looking at data in different ways through:

- Computing and tabulating basic descriptors of data properties such as ranges, means, and variances;
- Generating graphics, such as boxplots, histograms, scatter plots;
- Applying transformations, such as log or rank;
- Comparing observations to statistical models, such as the QQ-plot, or regression;
- Identifying underlying structure through clustering;
- Simplifying data through dimension reduction ...

. . . with the final goal of defining a statistical model and using the model for hypothesis testing and prediction.

# What does EDA supply?

Statistical techniques to:

- Tabulate,
- Summarize,
- Display,
- Reduce data

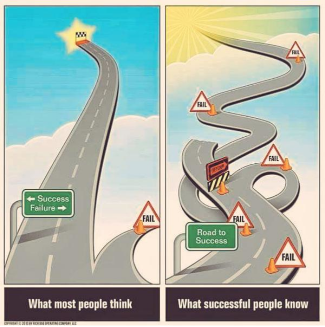
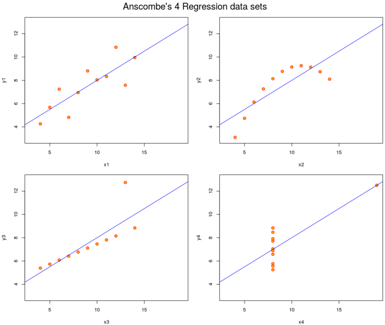


Image from [2].

After we explore, we can settle down and really analyze data.

# Anscombe's data[1]

Why it is important to look at data in different ways (EDA). A single way/tool/technique can be deceptive.



Each plot fits the linear equation: $y = 3 + 0.5x$
(Load attached file.)

# About data analysis?[4]

It takes creativity:

- DA can't be done mechanically
- Often there has to be a "creative" element
- Conventional DA is in a sense idealistic
- Trade-off between "ideal" experimentation vs. ecological validity
- Sometimes questions are tentative

We need data analysis skills that allow data to speak to us despite our expectation.

# We are interested in different types of numbers.

- Categorical (Qualitative)[6]
  - Nominal – values are just different
  - Ordinal – values can order objects
- Numerical (Quantitative)
  - Interval – differences between values are important
  - Ratio – differences and ratios are important

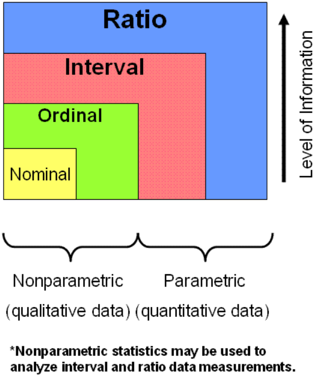


Image from [3].

# How can we deal with these data types?

Types of applicable statistical tests:

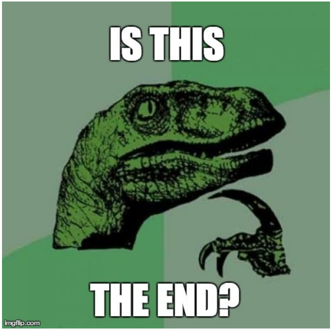
| | **Parametric** | **Non-parametric** | |
|---|---|---|---|
| | | **Level of measurement** | |
| **Number of groups** | **Interval/Ratio** | **Nominal** | **Ordinal** |
| **One group** | Z-test | One sample Chi-square | Kolmogorov-Smirnov test |
| | t-test | Binomial test | Runs test |
| **Two group (related samples)** | Paired t-test | McNemar test | Wilcoxon Signed Rank test |
| | Walsh test (interval) | | |
| **Two groups (independent samples)** | Independent Student t-test | Chi square Test | Mann-Whitney U test |
| | for equal/unequal variances | (if any cell has expected freq $<$ 5) | Kolmogorov-Smirnov two sample test |

# Q & A time.

Q: How many existentialists does it take to screw in a light bulb?
A: Two. One to screw it in and one to observe how the light bulb itself symbolizes a single incandescent beacon of subjective reality in a netherworld of endless absurdity reaching out toward a maudlin cosmos of nothingness.

# What have we covered?

- Exploratory data analysis (EDA) can be fun
- EDA is about no preconceptions
- EDA lets the data lead us into data analysis (DA)
- DA helps us to understand and explain the data



Next: What is Big Data?

# References (1 of 2)

[1] Francis J Anscombe, Graphs in Statistical Analysis, The American Statistician **27** (1973), no. 1, 17–21.

[2] Sina H, Road to success, https://thedailyteacher.com/2016/05/20/the-road-to-success/, 2016.

[3] Six Sigma Staff, Data Classification, 2017.

[4] Warwick Staff, Exploratory Data Analysis, homepages.warwick.ac.uk/~psrex/Lecture%20W6%20EDA.ppt, 2008.

[5] Boris Steipe, Exploratory Data Analysis of Biological Data using R, https://bioinformatics.ca/statistics2013module2-ppt, 2013.

# References (2 of 2)

[6] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Pearson Education India, 2006.

# R script to plot Anscombe's data

```r
rm(list=ls())
require(stats)
require(graphics)
main <- function()
{
    op <- par(mfrow = c(2, 2), mar = 0.1+c(4,4,1,1), oma =  c(0, 0, 2, 0))
    ff <- y ~ x
    mods <- setNames(as.list(1:4), paste0("lm", 1:4))
    for(i in 1:4) {
        ff[2:3] <- lapply(paste0(c("y","x"), i), as.name)
        plot(ff, data = anscombe, col = "red", pch = 21, bg = "orange", cex = 1.2, xlim = c(3, 19), ylim = c(3, 13))
        mods[[i]] <- lmi <- lm(ff, data = anscombe)
        abline(mods[[i]], col = "blue")
    }
    mtext("Anscombe's 4 Regression data sets", outer = TRUE, cex = 1.5)
    par(op)
}
d <- main()
```

# Files of interest

1. Anscombe's data script