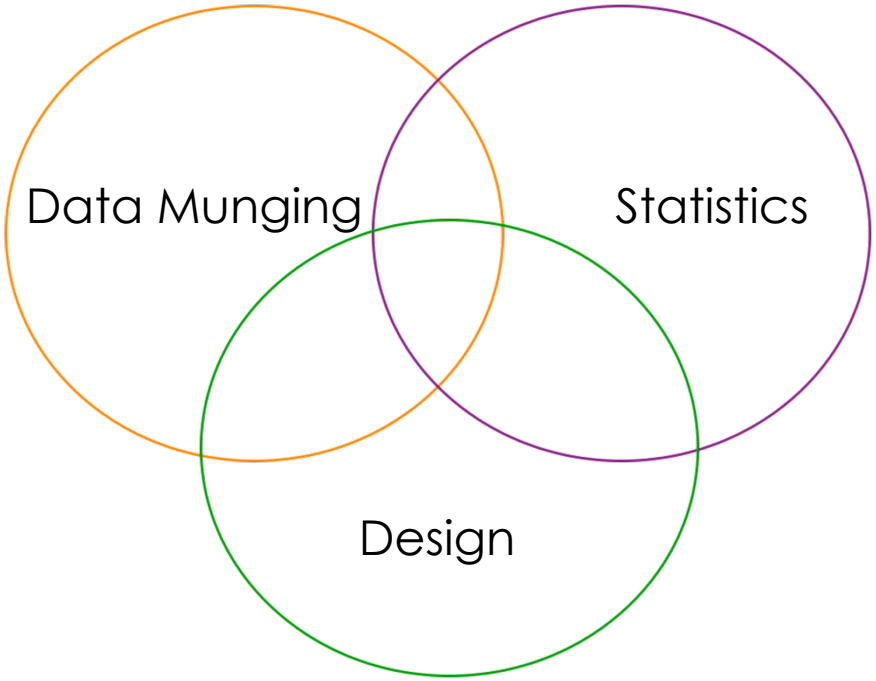
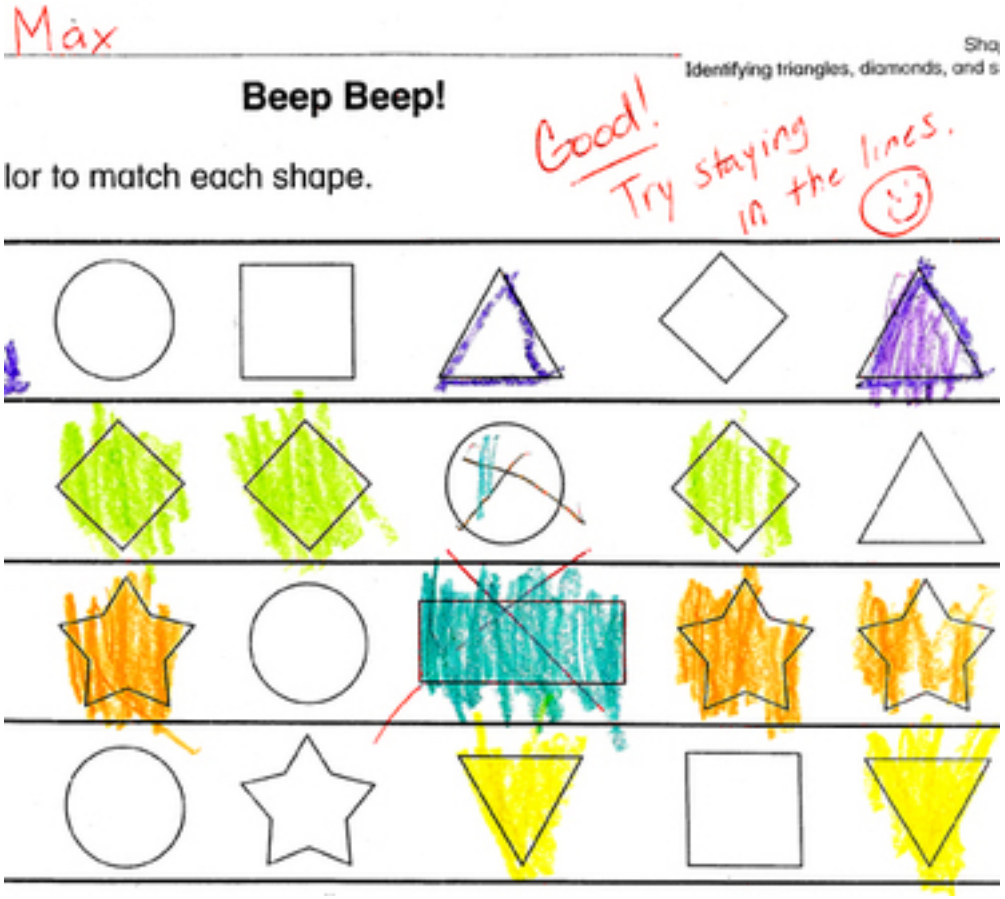
Thinking about probability and statistics when looking at Web Behaviour.

Narbeh Yousefian

About Me



Data Munging



Design

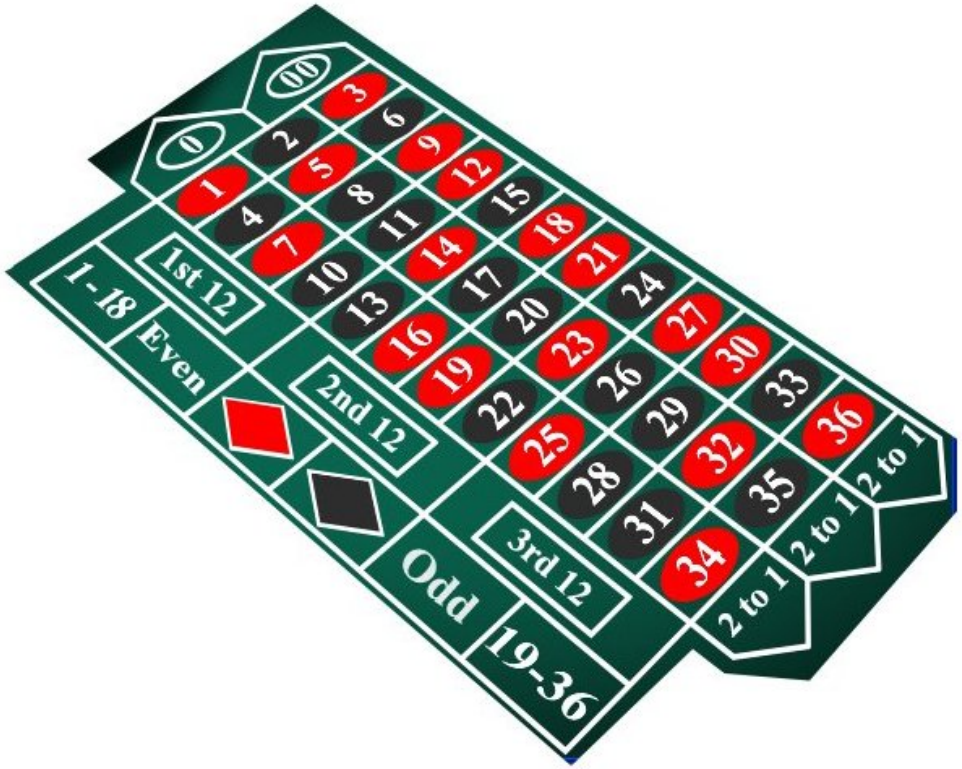
This is NOT Design!

Control Button

Experiment Button

Statistics

Trust me, the next one will be Black!



Statistics

Probability and Statistics. What is the difference?

Agenda

What will/will not be talking about?

And how is all of this related to our wonderful field

1. This is not a 101 in stats
2. There is no sales pitch
3. There will be puzzles
4. Some stuff about Regression

Puzzle 1

Puzzle 1

- **Challenge:** Three people enter the room, each with a hat on their head. There are two possible colors of hats: red and blue; they are assigned randomly. Each person can see the hats of the two other people, but they can't see their own hats.

- Each person can either try to guess the color of their own hat or pass. All three do it simultaneously, so there is no way to base their guesses on the guesses of others. If nobody guesses incorrectly and at least one person guesses correctly, they all share a big prize. Otherwise they all lose.

- **Strategy:** before the contest, the three people have a meeting during which they decide their strategy. What is the best strategy to maximize their odds of winning a prize?

Puzzle 1 (cont.)

- **Challenge:** Three people enter the room, each with a hat on their head. There are two possible colors of hats: red and blue; they are assigned randomly. Each person can see the hats of the two other people, but they can't see their own hats.

- Each person can either try to guess the color of their own hat or pass. All three do it simultaneously, so there is no way to base their guesses on the guesses of others. If nobody guesses incorrectly and at least one person guesses correctly, they all share a big prize. Otherwise they all lose.

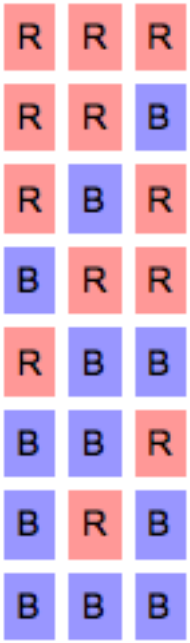
- **Strategy:** before the contest, the three people have a meeting during which they decide their strategy. What is the best strategy to maximize their odds of winning a prize?

  - **0%**
  - **25%**
  - **50%**
  - **75%**

Puzzle 1 (cont.)

- **Answer:** in a mixed case, one person sees two hats of the same color, the other two persons see mixed colors. So the clever strategy is to always pass when you see two hats of different color; but when you see two hats of the same color, make a guess that your hat is the opposite color.

- You will win 75% of the time!



Puzzle 2

The Simple Coin Toss

- Imagine tossing a coin successively, and waiting till the first time a particular pattern appears, say HTT
  - For example, if the sequence of tosses was

    HHTHHTHHTTHHTTTHTH
  - In example above: The pattern HTT would first appear after the 10th toss.
- Imagine you repeat this a million times, counting the number of times it takes until the pattern HTT appears. Could be the fifth toss, 20th toss, 8th toss. What we are looking for is an average value for the number of times it takes to toss a coin to reach HTT.
- Now what we want to do is the same thing but only this time we are looking for HTT
- **Challenge:** That is, if you count the number of times HTH or HTT appears, what we want to do is compare the average number of tosses till

The Simple Coin Toss (cont.)

- **Challenge: If** you count the number of times it takes HTH or HTT to appear, what we want to do is compare the average number of tosses till either event occurs

HTT V HTH

- **Challenge**: Consider the 2 patterns HTH and HTT. Which of the following is true?

1. The average number of tosses until HTH is larger than the average number of tosses until HTT
2. The average number of tosses until HTH is the same as the average number of tosses until HTT
3. The average number of tosses until HTH is smaller than the average number of tosses until HTT

The Simple Coin Toss (cont.)

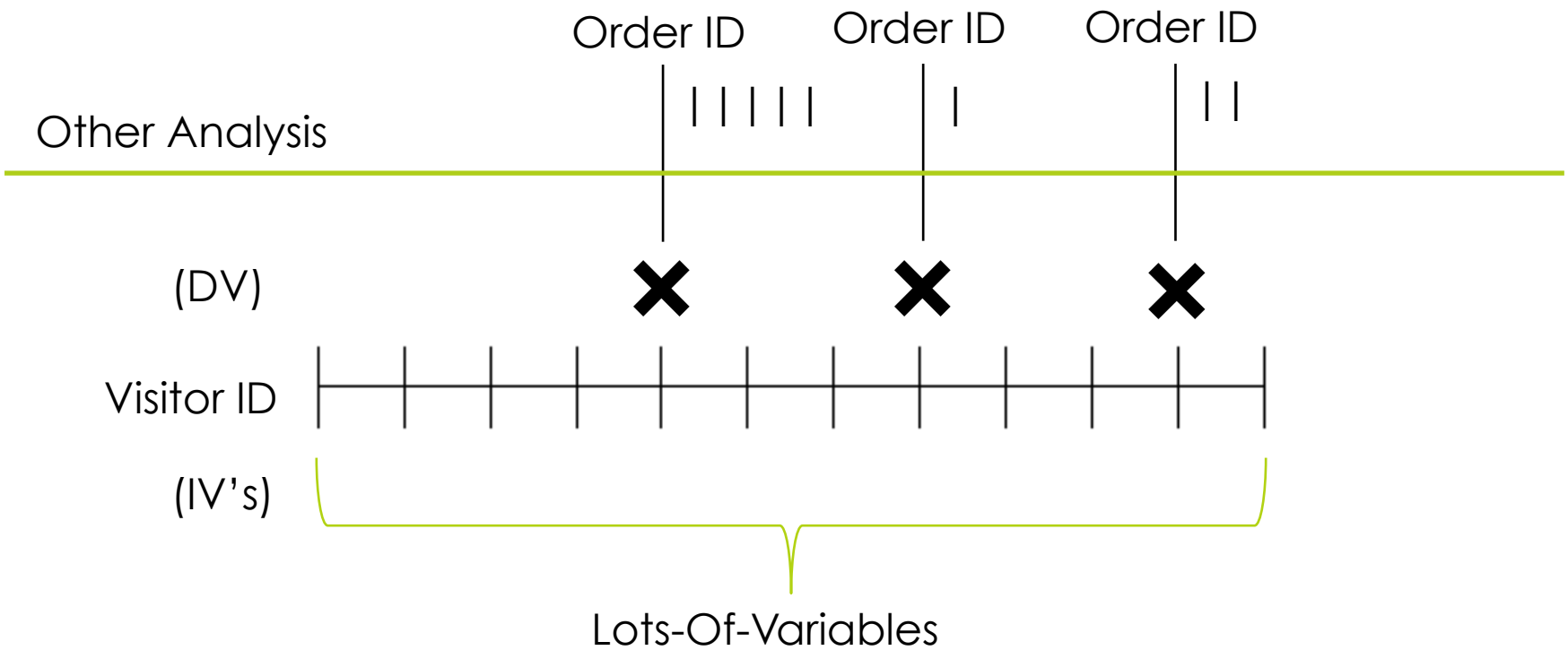
- **Challenge: If** you count the number of times it takes HTH or HTT to appear, what we want to do is compare the average number of tosses till either event occurs

HTT V HTH

- **Challenge**: Consider the 2 patterns HTH and HTT. Which of the following is true?

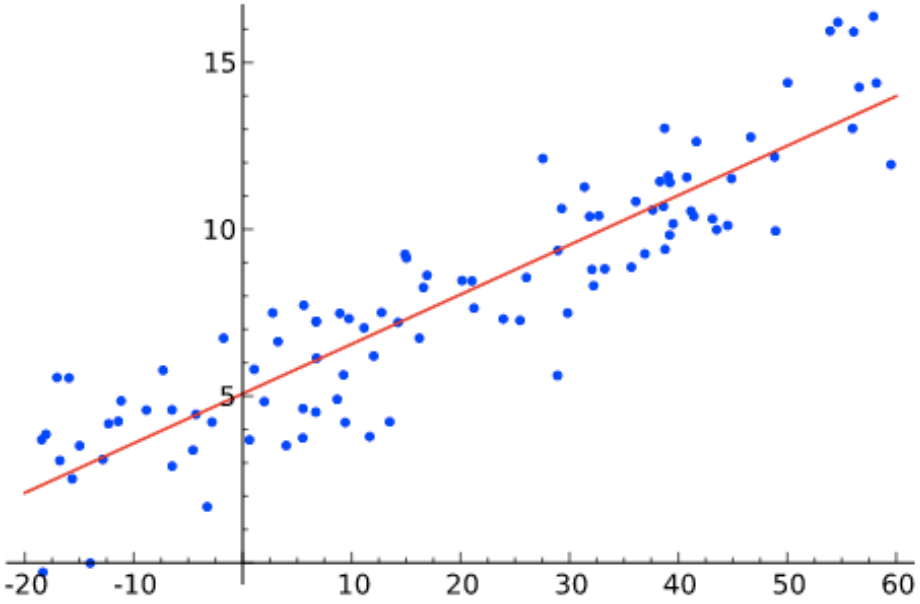
1. The average number of tosses until HTH is larger than the average number of tosses until HTT
    1. **The average tosses to get HTH is 10 and the average tosses for HTT is 8.**
2. The average number of tosses until HTH is the same as the average number of tosses until HTT
3. The average number of tosses until HTH is smaller than the average number of tosses until HTT

Where am I going with this?



Over fitting and Residuals

- Given sufficient complexity, ***any*** model can fit the data!
- **Residuals:** RMSE
  - What is it?
  - Why is it important?
  - What is its relationship to regression?

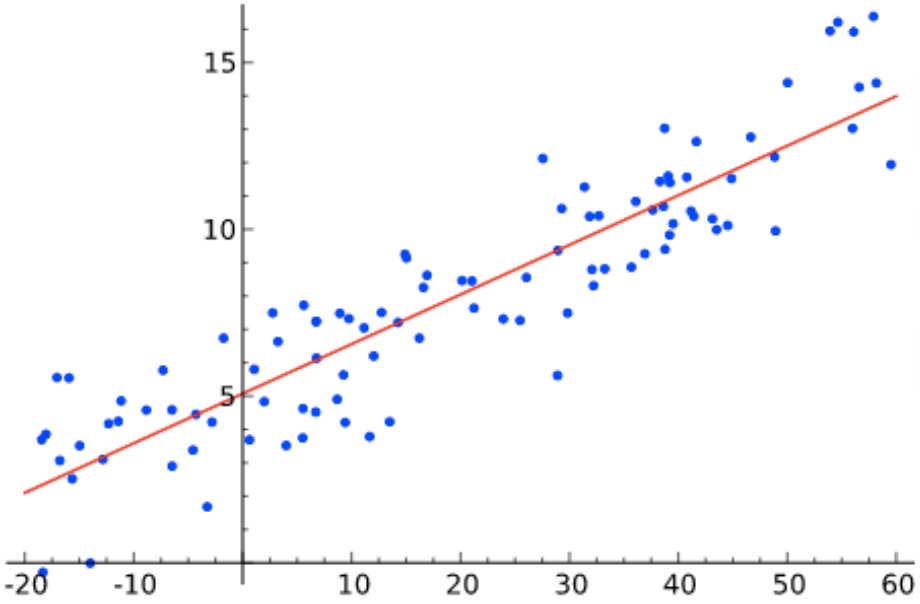


Do you do Experimenting online?

- **Regression:**
  - What do you use it for?

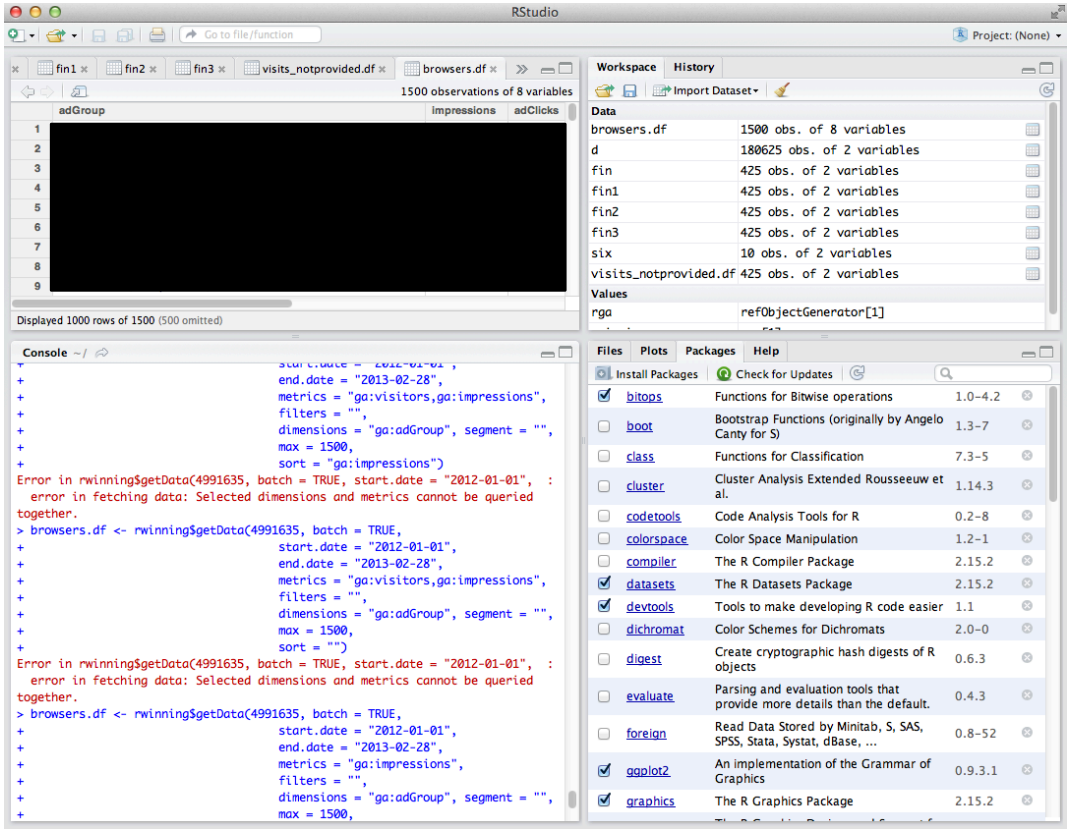
Predictor and estimator

- Identifying factors and using it for predictors.
- What factors are actually important within the pre-purchase steps? Before the funnel (for Netflix, it was the ratings)



Web (Digital) Analytics (still) today

- Beautifully Engineered SAAS Tables
- Rather irrelevant too…



Thanks

Appendix

- Rstudio (IDE for statistical programming)
  - Learn the package ggplot2 for data visualization
- Good content about learning statistics
  - udacity.com
    - Great courses related to statistics
  - coursera.com
    - Great intro course to using R as a language
- Good books about learning statistics:
  - What ever works for you, find something that is about applied (unless you really like deep theory) I work with R, you might like Python, MSFT excel then use that.
- You can find me here:
  - Linkedin
  - Twitter
  - Quora
- Or just email me: **narbeh@digdeepdigital.com**