Technology Fundamentals for Business Analytics

Jason Kuruzovich

Feedback

- Slowing Down
- More small exercises in class
- Extra help session
  - Carnegie 106 is available from 3- 4pm on Wednesdays??
- Extension so more opportunity for asking questions on labs before due
  - This doesn't mean you can wait till Tuesday night to start lab

Tips & Philosophy

- Ask questions….
- Seeing things multiple times will help
- Two approaches
  - Show pieces and describe them (Lab 2 and Lab 4)
    - Syntax, coding
  - See a bundle solutions and adjust them to get and interpret data
    - Technology, process

Labs

.doc file

Do not upload questions.

Lab 2 R Solution

http://rpi-analytics.github.io/MGMT6963-2015/assets/rmarkdown/lab2.html

Lab 3

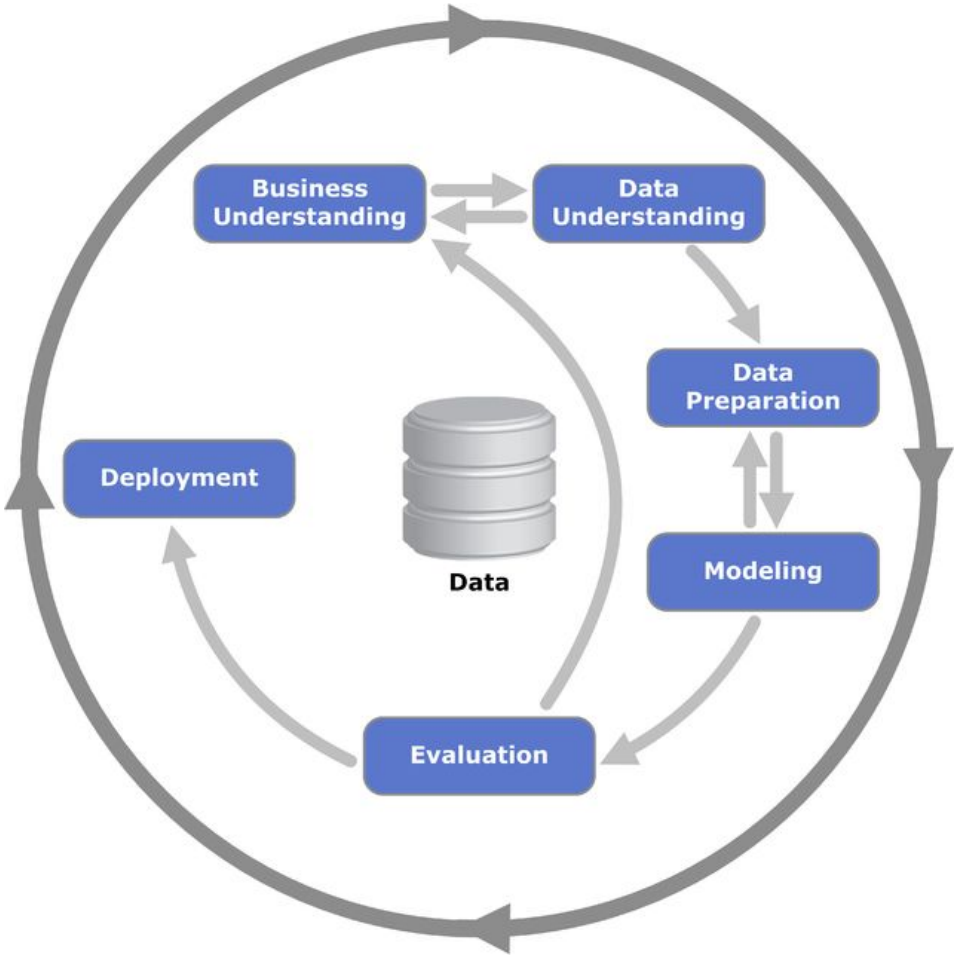
Background Revisited

Why are hashtags important?

What did we do with twitter API?

Who had problems and how might we troubleshoot?

THE CRISP-DM PROCESS MODEL



Cross Industry Standard Process for Data Mining
https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Subsetting Data

Subset Criteria

- Selection vector with list of desired rows.
  - *In this case we generate a vector which contains the desired rows of the dataset. The vector will have a length equal to the subset.*
- Boolean vector where true for desired rows
  - *In this case we generate a boolean vector with a value of true for the rows we want. The vector will be of the same length of the initial data.*

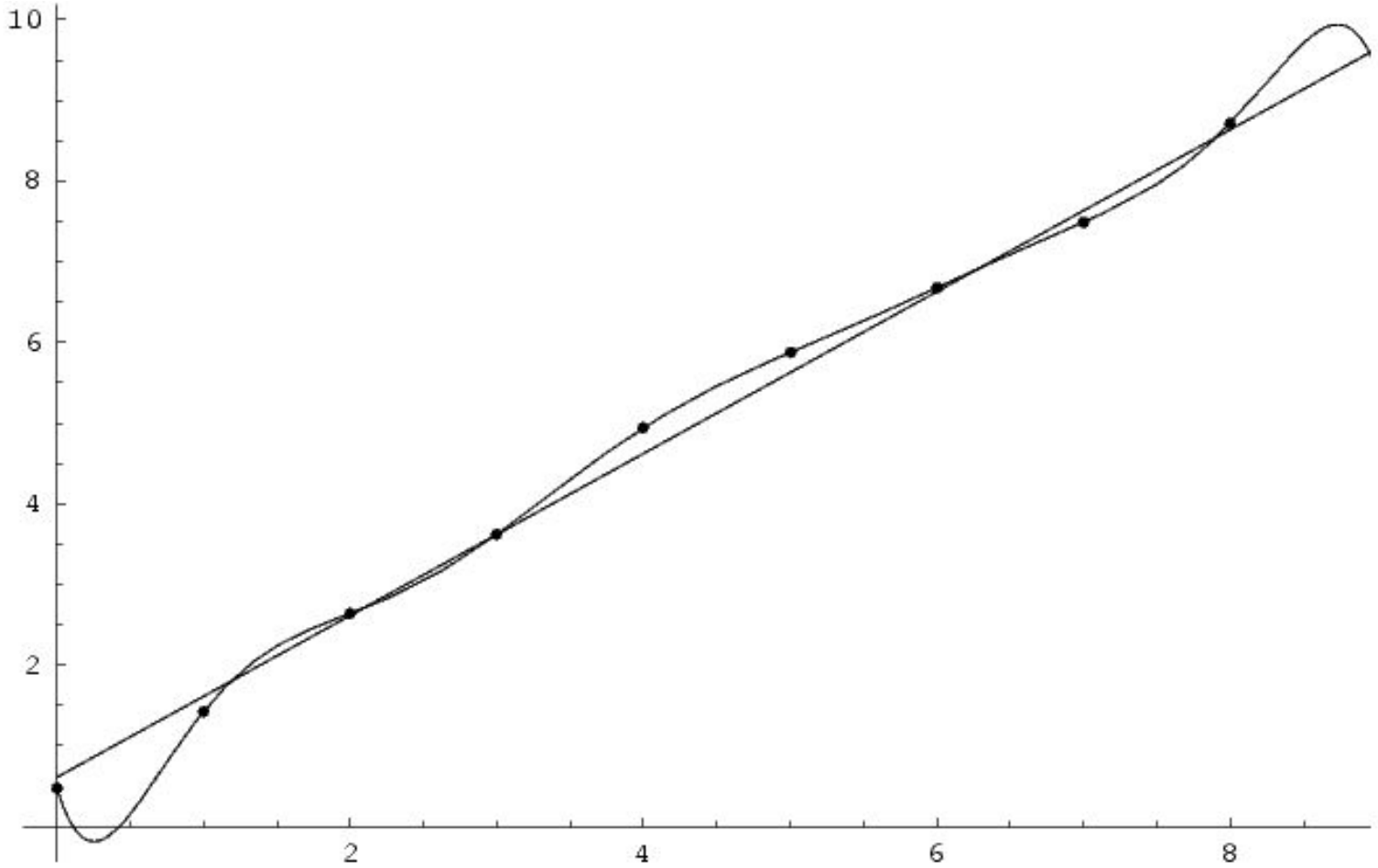
Subset Criteria

- Can be variable based selection:
  - A > 30; gender="male"; state = "NY"
- Can be random subsets
  - Train your algorithm/model on one dataset (or a series of subsets)
  - Test your algorithm on a separate dataset this ensures you won't "overfit" your model

Subset Criteria

- Selection vector with list of desired rows.
  - *In this case we generate a vector which contains the desired rows of the dataset. The vector will have a length equal to the subset.*
- Boolean vector where true for desired rows
  - *In this case we generate a boolean vector with a value of true for the rows we want. The vector will be of the same length of the initial data.*

Over fitting



https://commons.wikimedia.org/wiki/File:Overfit.png

Overfitting

- “Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship.” -Wikipedia
- Predictive performance often won’t be good
- The model is often too complex and finely tuned to sample

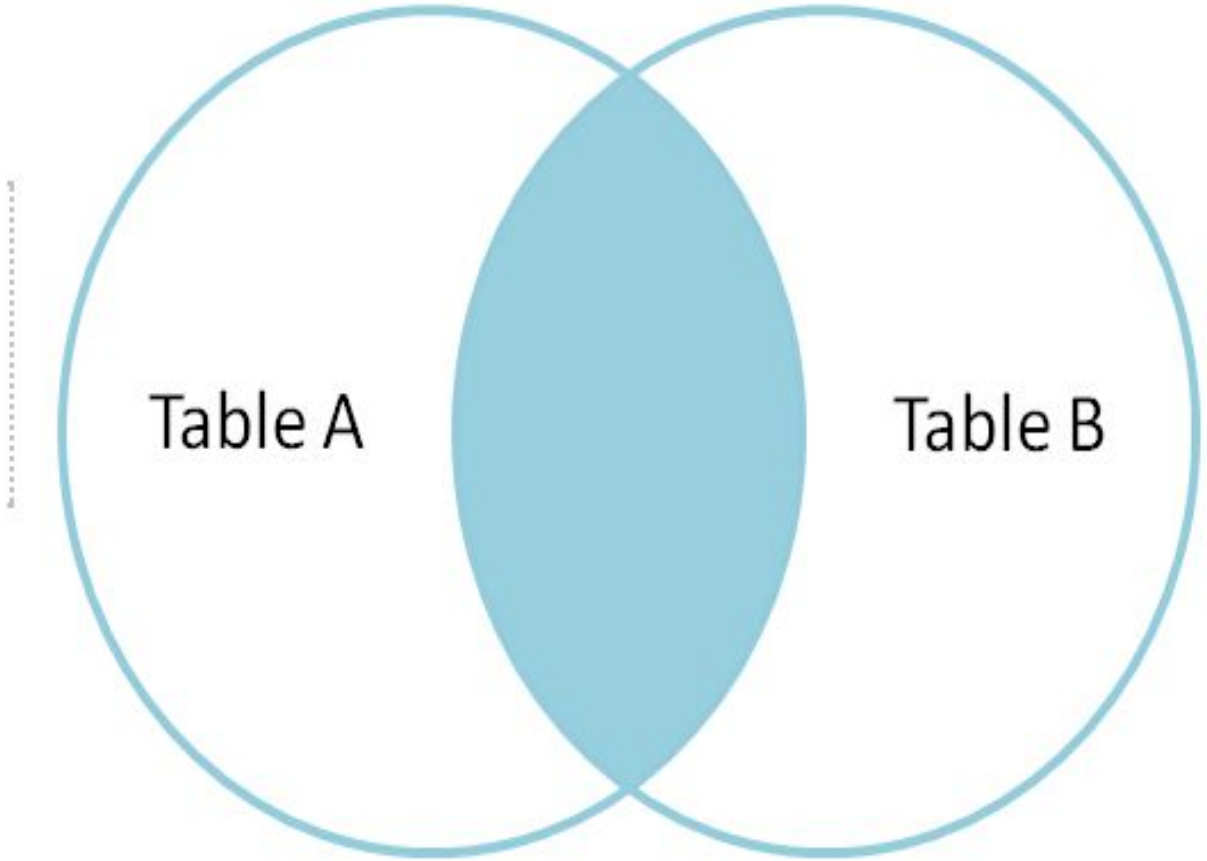
Merging Datasets

- Need *Key* in order to link datasets
- A key must uniquely identify a record
- If a key does not exist, you typically will try to create one before matching the datasets
- For now, let's assume that we have a key

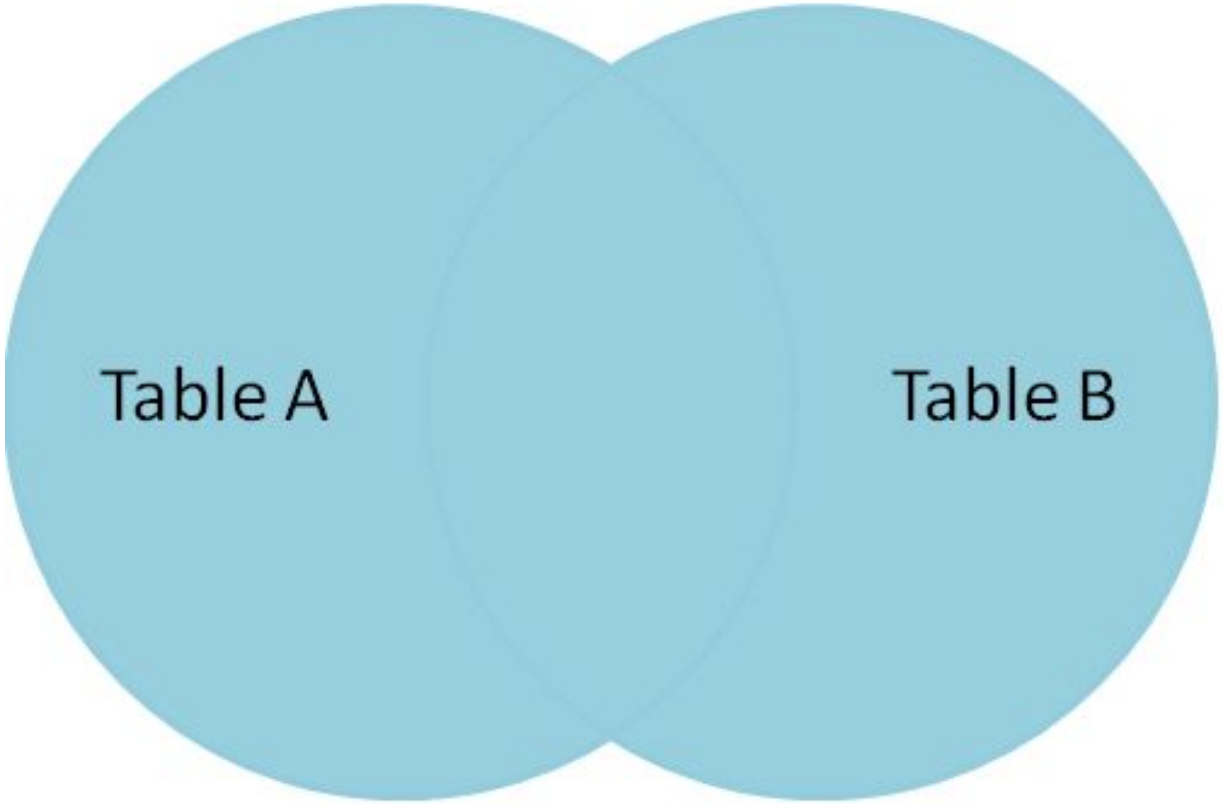
Merging Datasets

- Inner Join. The outcome file should only have a row where there is a match between only keys where A = B.
- Left Outer Join. All records from A (even if no match) and only records from B where there is a match.
- Right Outer Join. All records from B (even if no match) and only records from A where there is a match.
- Full Outer Join. All records from A & B (even if no match).

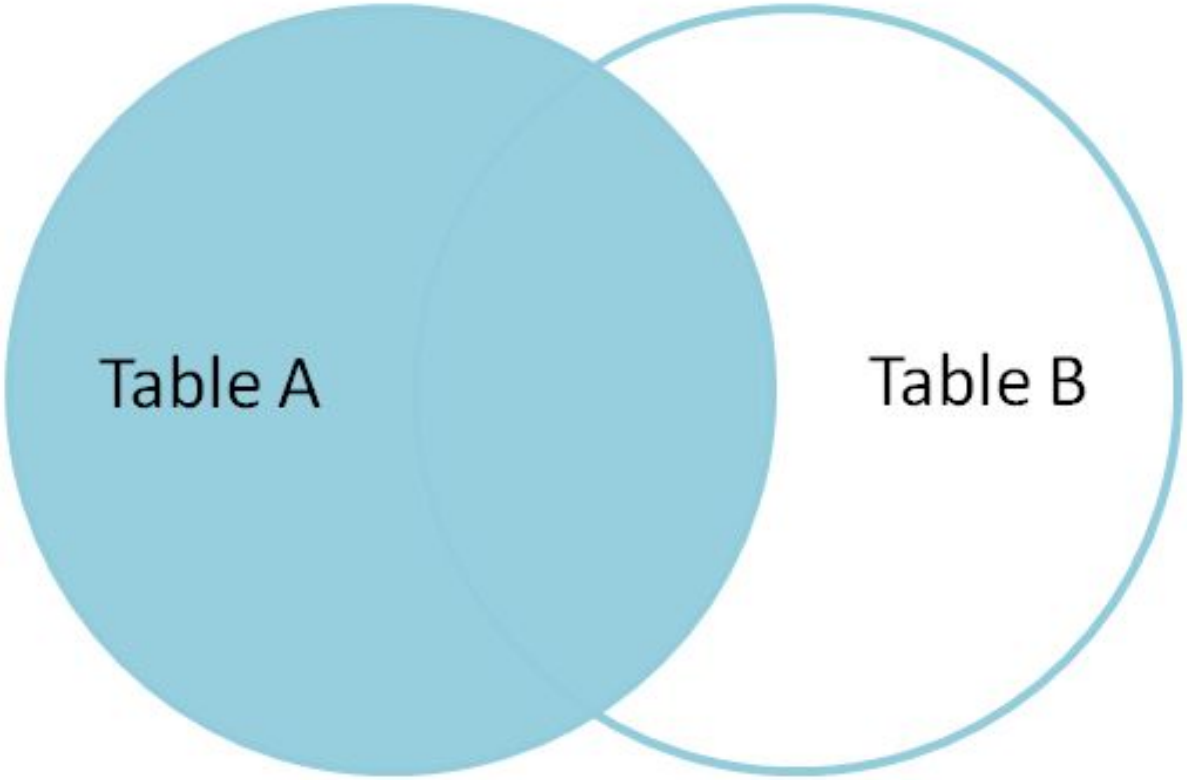
Join Datasets: Which is this?



Join Datasets: Which is this?

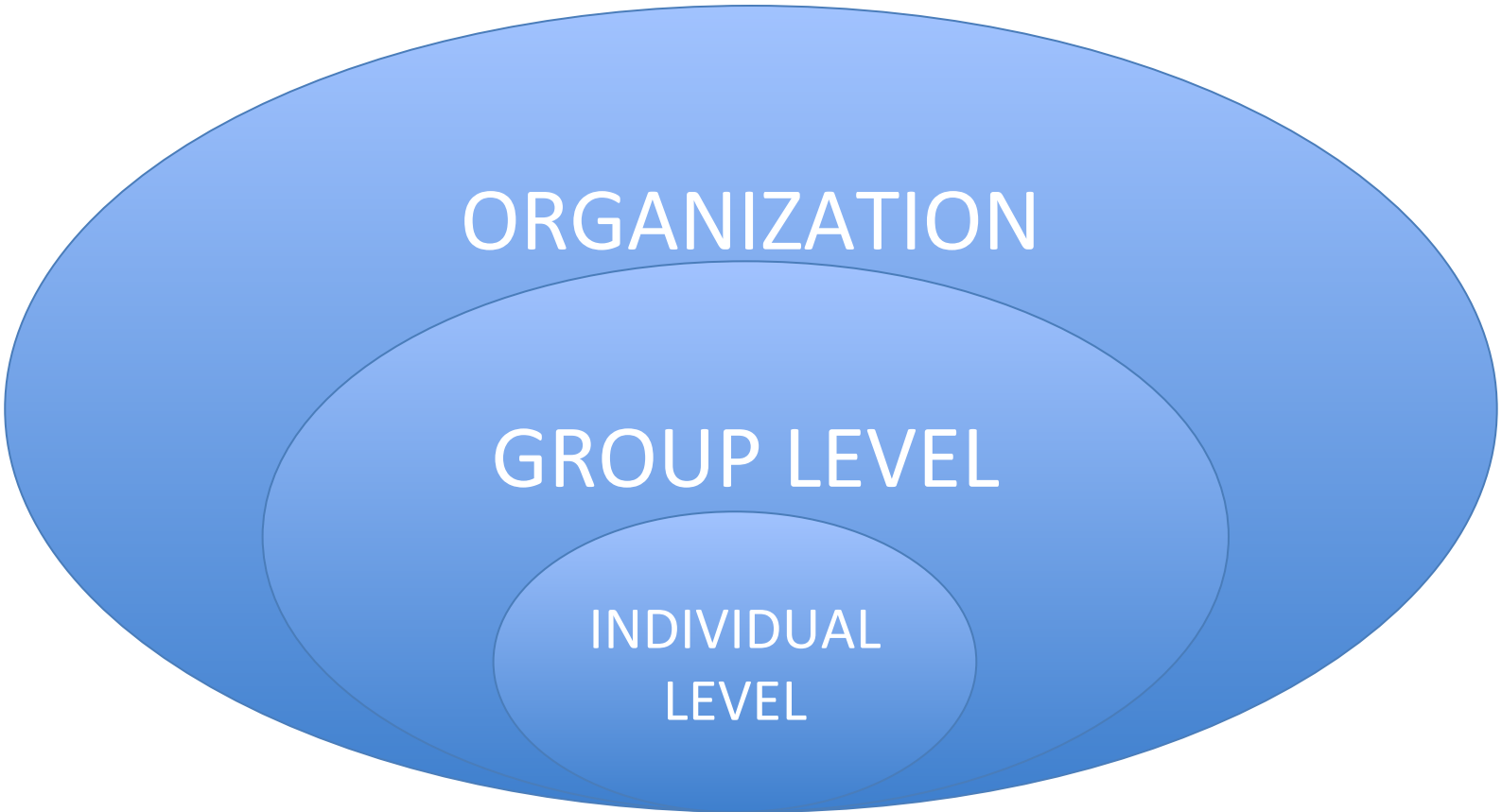


Join Datasets: Which is this?



Aggregation

- *Often the same data can be examined at different level of analysis…..*



Consider Student Performance

- What different ways might someone measure/understand student performance in high schools.

Student Performance

- What *individual* characteristics influence test performance?
- What *teacher* characteristics influence average test scores in a classroom?
- What school level characteristics in influence average test scores?

Aggregating

- When aggregating, most times you will either take the sum or the average of a variable
- The appropriate aggregation technique relies on your business understanding
- Example:
  - Sum(Number of Sales)
  - Average(Sale $)

Working with Data

R

- Subset datasets
- Merge datasets
- Aggregate Datasets

Python

- Subset datasets
- Merge datasets
- Aggregate Datasets