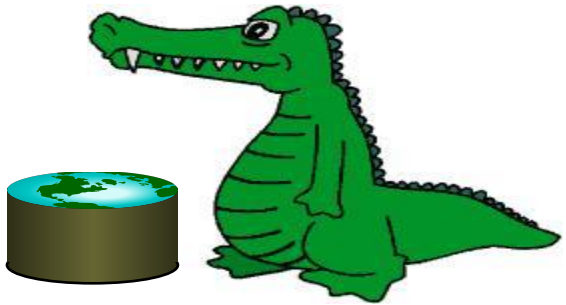
Lab 3: Joining Multiple Tables (Python & Pandas)

Xiaofeng Zhou



Goal

Learn more complex operations on DataFrame and advanced data analysis

- Aggregate and Pivot
- Joins (equijoin and fuzzy join)
- Advanced Analysis: precision & recall

Aggregate and Pivot

- Aggregate
  - Use Python lambda function and DataFrame ‘apply’ to rows.
- Pivot
  - Reshape DataFrame to a new DataFrame:
    - values in one column as index for rows,
    - values in another column as column names,
    - The rest (or specified) columns as values.

Joins (equijoin and fuzzy join)

- Equijoin
  - DataFrame ‘merge’ function, and specify joining column.
- Fuzzy join
  - Cartesian product
  - Add a column to measure similarity as join criterion
  - Filter based on join criterion

Advanced Analysis

- Precision and Recall
  - Used to evaluate the quality of fuzzy-joining results on ‘name’ column.
- Precision and Recall Curve
  - Visualize the tradeoff of precision vs recall.
  - Compare different metrics.

Q/A session

Quiz 3

If you have wifi connection issues, go to lab rooms to finish the quiz. The quiz will be available for 25 mins.