# Validity Take-Home Exercise

## Last Updated: Sunday, June 3, 2018

At Validity, we invest heavily in our co-op program. For that reason, we want to make sure we’re only hiring the best future engineers; to do that, we have to see how you code and solve problems, not in front of a white board but at a keyboard.

**The Exercise**

1. Write an app that parses the attached CSV using any language you choose (unless you’ve been asked otherwise).
2. Identify possible duplicates in the data-set by leveraging metaphone, Levenshtein distance or other means. Ignore IDs for deduplication as they are here to help you visually identify duplicates for your own testing.
3. Print each set of duplicates separately as well as a set of non-duplicate entries. Choose one of the following methods:
   1. Print the results to the standard output (stdout)
   2. Create a JSON object output
   3. Build a mini web application to display the data

**Example Output**:  
­­­­­­­­­­­­­­­­­­­­­­  
Duplicate Set 1:  
ID 12: Bob Smith, Bank of America, 34 New Heaven Avenue, Boston, MA 02344  
ID 12: Robert Smith, Bank of America, 34 New Heaven Avenue, Boston, MA 02344  
ID 12: Robert Smith, Bank of Americam, 34 New Heaven Avenue, Boston, MA 02344  
ID 12: Robber Smith, Bank of America, 34 New Heaven Avenue, Boston, MA 02344

Duplicate Set 2:  
ID 342: Sam Thompson, The MathWorks, 633 Commonwealth Street, Boston, MA 03439 – 672-343-2343  
ID 342: Sam Thompson, The MathWorks – 672-343-2343  
ID 342: Sam Thompson, The MathWorks, 633 Commonwealth Street, Boston, MA 03439 – 672-343-2343  
ID 342: Sam Thompson, The MathWorks, 633 Commonwealth Street, Boston, MA 03439 – 672-343-3243

Non-Duplicates:  
­­­­­­­­­­­­­­­­­Tim Kennedy, ME  
Greg Menton, MA  
Sarah Long, MA

1. Submit your code through a Github repository for us to review, with a Pull Request showing the commits you made that allows us to comment on it.
2. Some duplicates are harder to identify than others and some false negatives harder to eliminate. Do what you can in the few hours you can dedicate to this.

**Rules**:

* You can use any language / library you’re comfortable (unless you’ve been told otherwise), but please no esoteric language – choose amongst the mainstream.
* Your README should include how we should compile / pull dependencies / run your code. Hopefully this should be a single step with some build / package management tool.
* Your code must be available on a public Github repository as a Pull Request for us to review.
* Use any resource you have available to you - but remember that we’re going to ask you questions about your code, so be prepared to defend any data structure/algorithm choices you make.
* Ignore the odd dataset, it’s non-sensical and auto-generated. For example email addresses do not match the Company Name.