# Munging Exercise

This is intended to be a short exercise for demonstrating your abilities to work with real-world messy data. There are no tricks or gotchas, we’d just like to see your proficiency at manipulating data structures and providing clear documentation of your work. The data example is real, although some metadata fields have been anonymized using human-readable hash codes and actual data values have been changed.

## Problem:

There are two datasets which need to be reconciled, one exported from a system used by agronomists (‘data/agronomists/’) and one from a centralized data warehouse (‘data/data\_warehouse’). Both data sets were created independently from the same set of laboratory results files. The laboratory results are measurements of corn biomass from experimental trials. Variables measured include nitrogen content (n\_pct), nitrogen weight (total\_n\_g), and dry weight (dry\_weight\_g).

In theory, these datasets should be exactly the same, but in practice there may be discrepancies due to uploading errors and different normalization processes during upload

## Goal:

1. Using a script, notebook or other easily repeatable method, join these datasets to produce a table with the following columns:

- **trial\_id** : 6-word trial id hash

- **plot** : Plot id

- **sample\_\_growth\_stage\_range** : Agronomic growth stage category of corn (VT or V6-V7)

- **variable** : response variable (n\_pct, total\_n\_g, or dry\_weight\_g)

- **value\_agronomist** : value from agronomist dataset

- **value\_dw** : unadjusted data warehouse value

- **value\_dw\_adjusted** : data warehouse value divided by items\_per\_sample

- **items\_per\_sample** : from the dw dataset

- **n\_samples**  : number of samples taken for this plot, variable & growth stage combination according to the data warehouse

1. Which data are poorly aligned? Generate a concise output (could be graph, table or other) which you might share with another data engineer to help them identify which records need to be investigated further.
2. How would you double check that your analysis & recommended changes are correct?