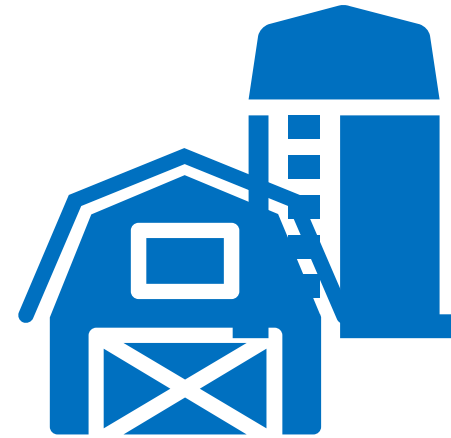
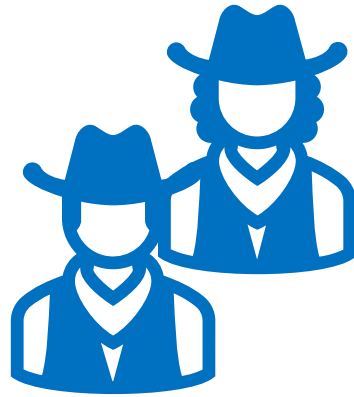
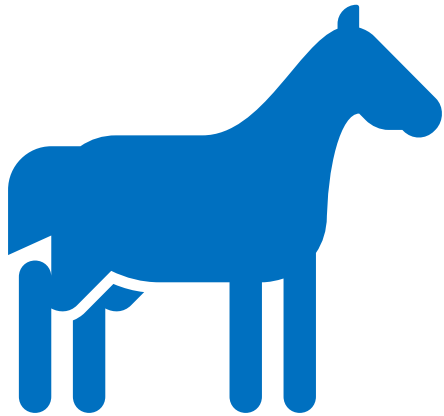


Data Wranglin'



Bryan Scott

LSST-DA Data Science Fellowship Program Session 21

University of Illinois, Urbana-Champaign

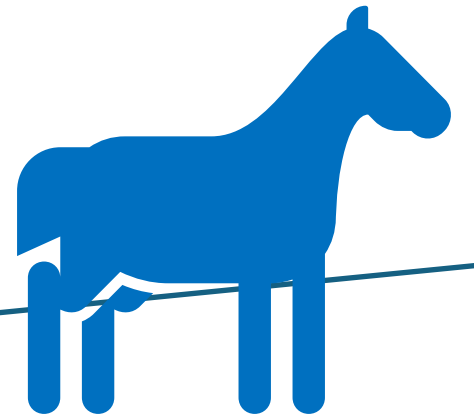


Webster's Dictionary defines wrangler as:

wrangler noun

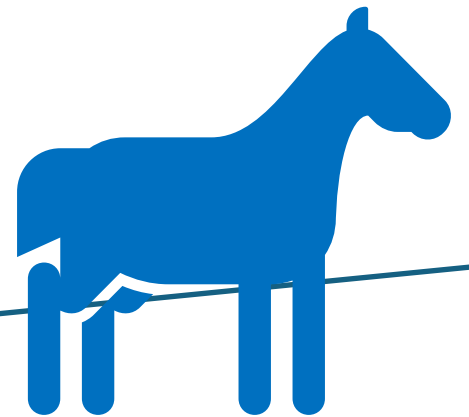
wran·gler | ran-g(ə-)lər

(short for horse-wrangler, probably partial translation of Mexican Spanish caballerango groom): a ranch hand who takes care of the saddle horses broadly : cowboy



How then, as astronomers, are we all like cowhands?

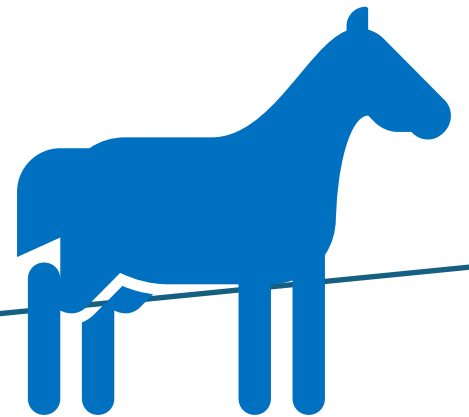
Data are often like horses in that:
1. they all differ,



How then, as astronomers, are we all like cowhands?

Data are often like horses in that:

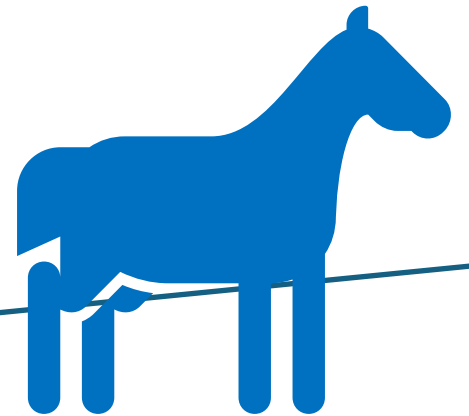
1. they all differ,
2. rarely conform to a single standard set of behavior,



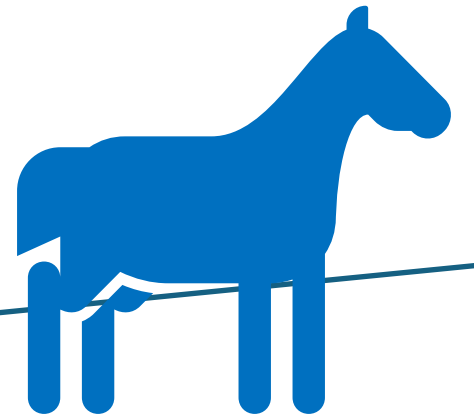
How then, as astronomers, are we all like cowhands?

Data are often like horses in that:

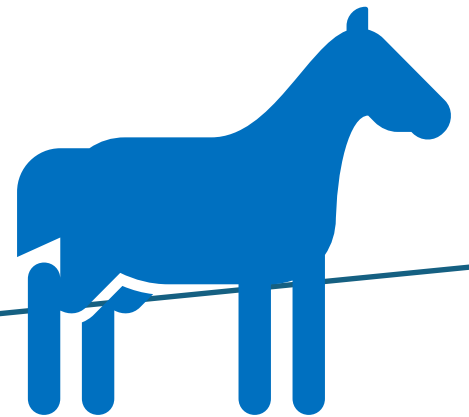
1. they all differ,
2. rarely conform to a single standard set of behavior,
3. and they love to eat hay.



Thus, in our efforts to better understand the Universe, we must often manipulate, coax, and, in some cases, force our data to "behave."



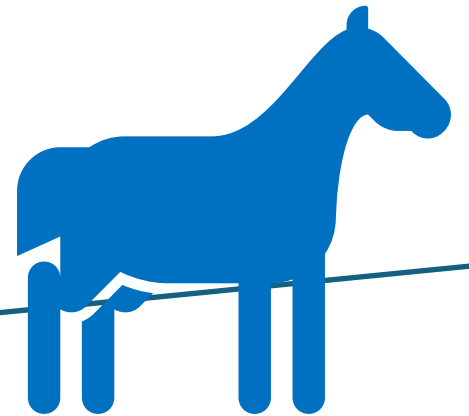
Thus, in our efforts to better understand the Universe, we must often manipulate, coax, and, in some cases, force our data to "behave." This involves a variety of tasks, such as: gathering, cleaning, matching, restructuring, transforming, filtering, combining, merging, verifying, and fixing data.



Here is a brief and unfortunate truth, there isn't a single person in the entire world that would organize data in **exactly** the same way that you would.

As a result, you may find that data that are useful to you are not organized in an optimal fashion for use in your workflow/software.

Hence: the need to wrangle.

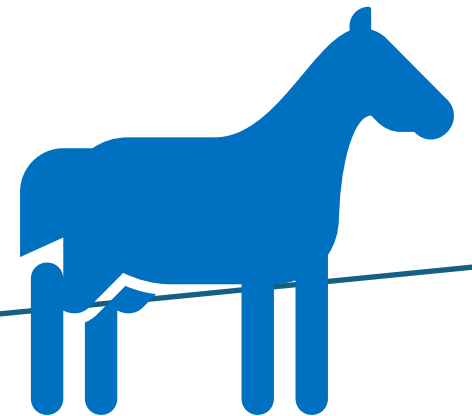


There is one important and significant way in which our lives as astronomers are much better than the average data scientist: even though our data are "worthless," virtually all of it is numbers.

Furthermore, I contend that most astronomical data can easily be organized into a simple tabular structure.

Nevertheless, as you will see during the exercises, even with relatively simple, small numerical data sets there is a need for wrangling.

And wrangling brings up a lot of issues...

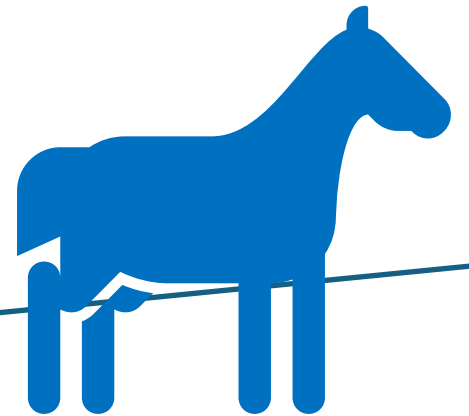


Example:

Consider the following data set that contains the street names for Adam's best friends from childhood:

['Ewing', 'Isabella', 'Reese', 'Isabella',
'Thayer', 'Reese', 'Reese', 'Ewing', 'Reece']

Do you notice anything interesting?




Either Adam's hometown has a street named "Reese" and a street named "Reece", or the last entry was input incorrectly.

If the latter is true, then we have to raise the question of: what should we do?

For this particular data set, it would be possible to create a verification procedure to test for similar errors.

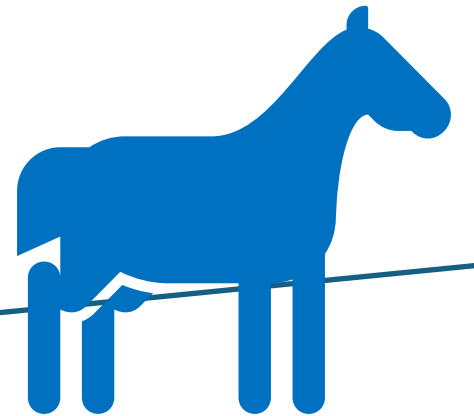
1. Collect every street name in the city (from post office?)
2. Confirm every data entry has a counterpart.

For any instances where this isn't the case, one could then intervene with a correction.



This particular verification catches this street name error, but it doesn't correct for the possibility that the person doing the data entry may have been reading addresses really quickly and the third "Reese" entry should have actually said "Lawndale."

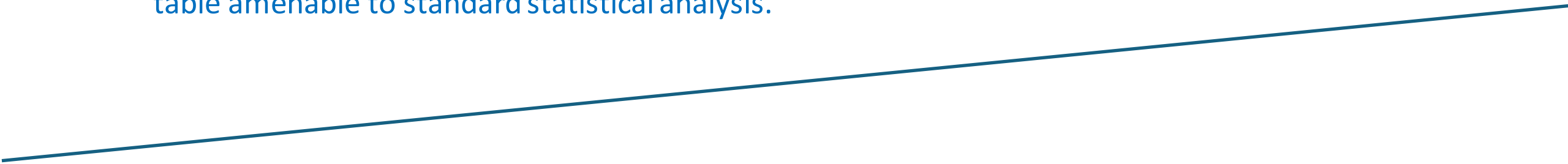
(verification is really hard)



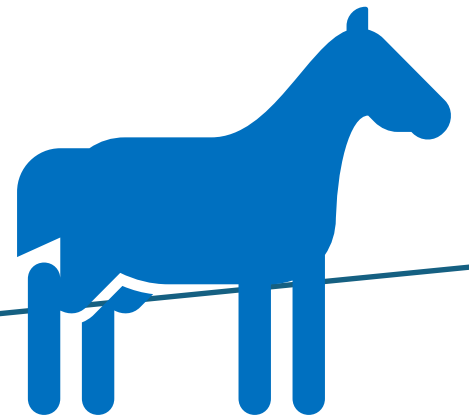
Data provenance – a historical record of the data and its origins – is really really hard.

If you are making "corrections" to the data, then each and every one of those corrections should be reported (for databases this is called "logging"). Ideally, these reports would live with the data so others could understand how things have changed.

For completeness, I will mention that there is a famous canonical paper about data wrangling: <http://vis.stanford.edu/files/2011-Wrangler-CHI.pdf>, which introduces the `Wrangler`: <http://vis.stanford.edu/wrangler>, a tool specifically designed to take heterogeneous (text) data, and provide a set of suggested operations/manipulations to create a homogenous table amenable to standard statistical analysis.

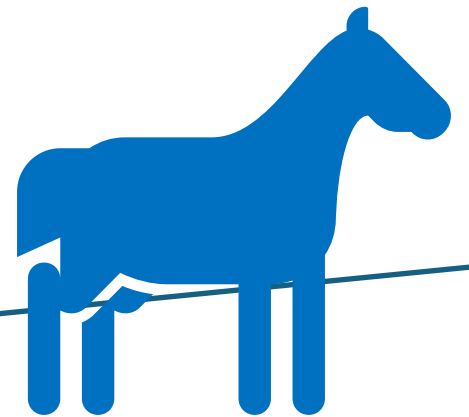


One extremely nice property of the `Wrangler` is that it records every operation performed on the data, ensuring high-fidelity reporting on the data provenance.



One extremely nice property of the `Wrangler` is that it records every operation performed on the data, ensuring high-fidelity reporting on the data provenance.

We should do a better job of this in astronomy.



Why harp on this?

In practice, data scientists (including astronomers) spend an unreasonable amount of time manipulating and quality checking data (some industry experts estimate that up to 80% of their time is spent wranglin').

Today, we will work through several examples that require wrangling, while, hopefully, building some strategies to minimize the amount of time you spend on these tasks in the future.

