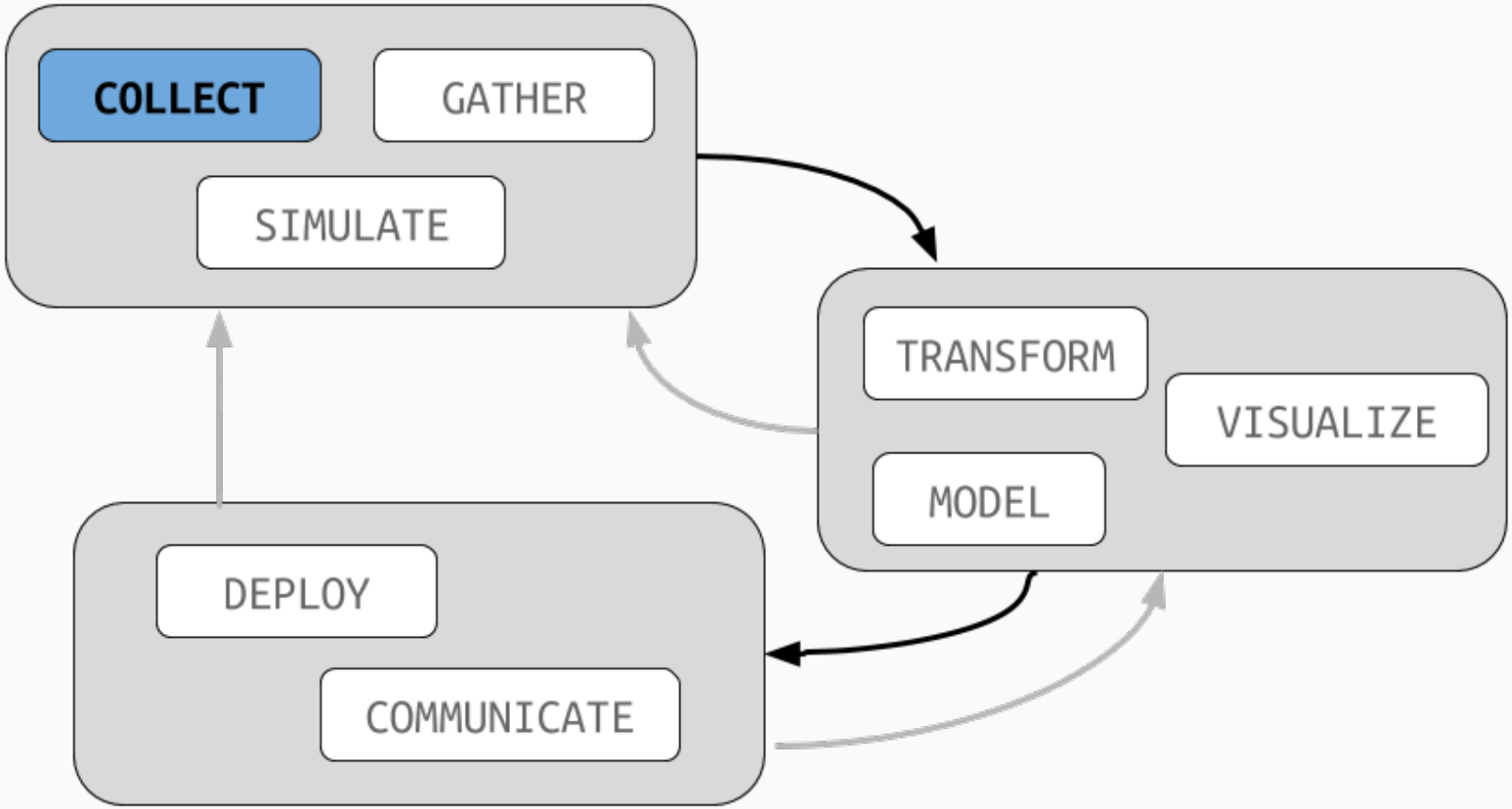
Lecture 11: Collecting tidy data

Taylor Arnold

movie dataset


COLLECT
GATHER
SIMULATE
TRANSFORM
VISUALIZE
MODEL
DEPLOY
COMMUNICATE

goal

We are going to start today by having you all open either excel, open office, Google sheets, or the spreadsheet editor of your choice.

You will be constructing a dataset representing your five favorite films.

data elements

Please collect the following (I suggest starting with Wikipedia):

- name of the movie
- movie budget
- country of origin
- date first released
- staring actors (truncate to top 3 if too many)
- birthplace of each actor
- rotten tomatoes rating of the movie

csv and reading into R

Once you are done with this, export the file as a CSV to your computer and read this into R. Time remaining, try to construct some interesting plots.

data collection principals

three principles

All of the principles of constructing a dataset (equivalently, a database) could easily fill a whole course. Here are three principles that get us on the right track:

- determine the objects of study; each of these gets its own table, and each example gets its own row; movies, actors, actor-movie links
- each column should be indivisible and the variable type clear; for example, budget should not include the dollar sign, if needed create a new column; name columns with no spaces or special characters
- always have internal consistency (0.62 or 62 percent; missing values always "NA"); strive for external consistency (ISO country codes)