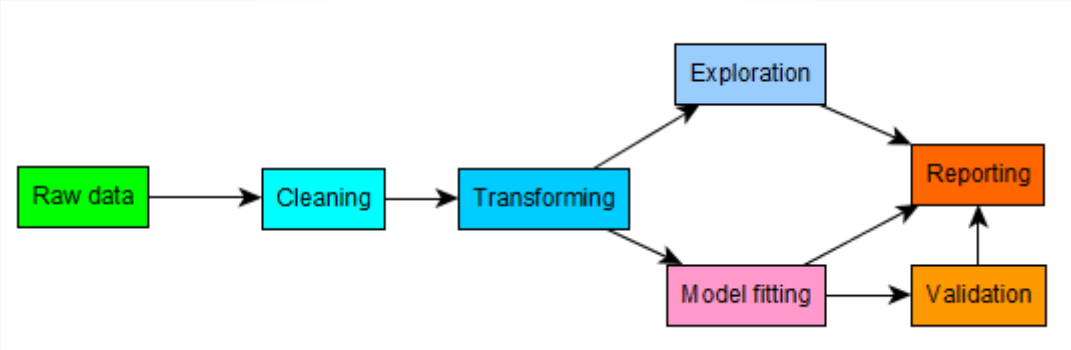
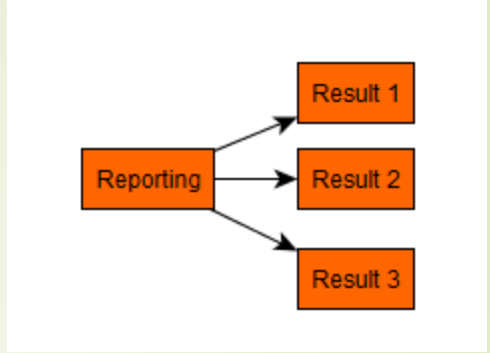
Git for data science

Flow of data in the Statistical modeling process



Motivation when continuously Reporting results



- Why is result 2 different than result 1?
  - Maybe an effect flipped sign, but between result 1 and result 2, there was a change in raw data, change in data cleaning, change in transformation…
- Are you sure that result 3 is was produced correctly?
- Can you reproduce result 2 and grab one additional quick statistic?
- Can you get a higher resolution image of result 3?

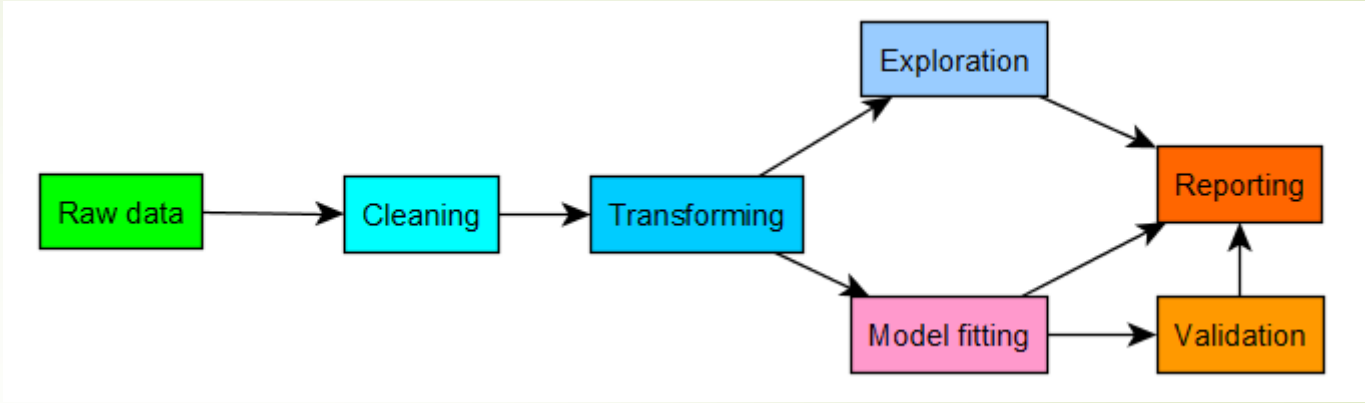
- If you keep a single script and change it as you move from results 1,2,3… it can often times be difficult to return to the exact original script and reproduce the original results

Repository goals

- Code used to produce results reported to customer quickly retrieved
- Any reported result can we quickly and easily extended
- The reason behind difference or changes between two sets of results can be determined with a minimum of hassle
- Code examples (“I know I’ve done this before”) can be quickly retrieved based off of search
- Any repository code can be modified ad-hoc without worry of losing work/results

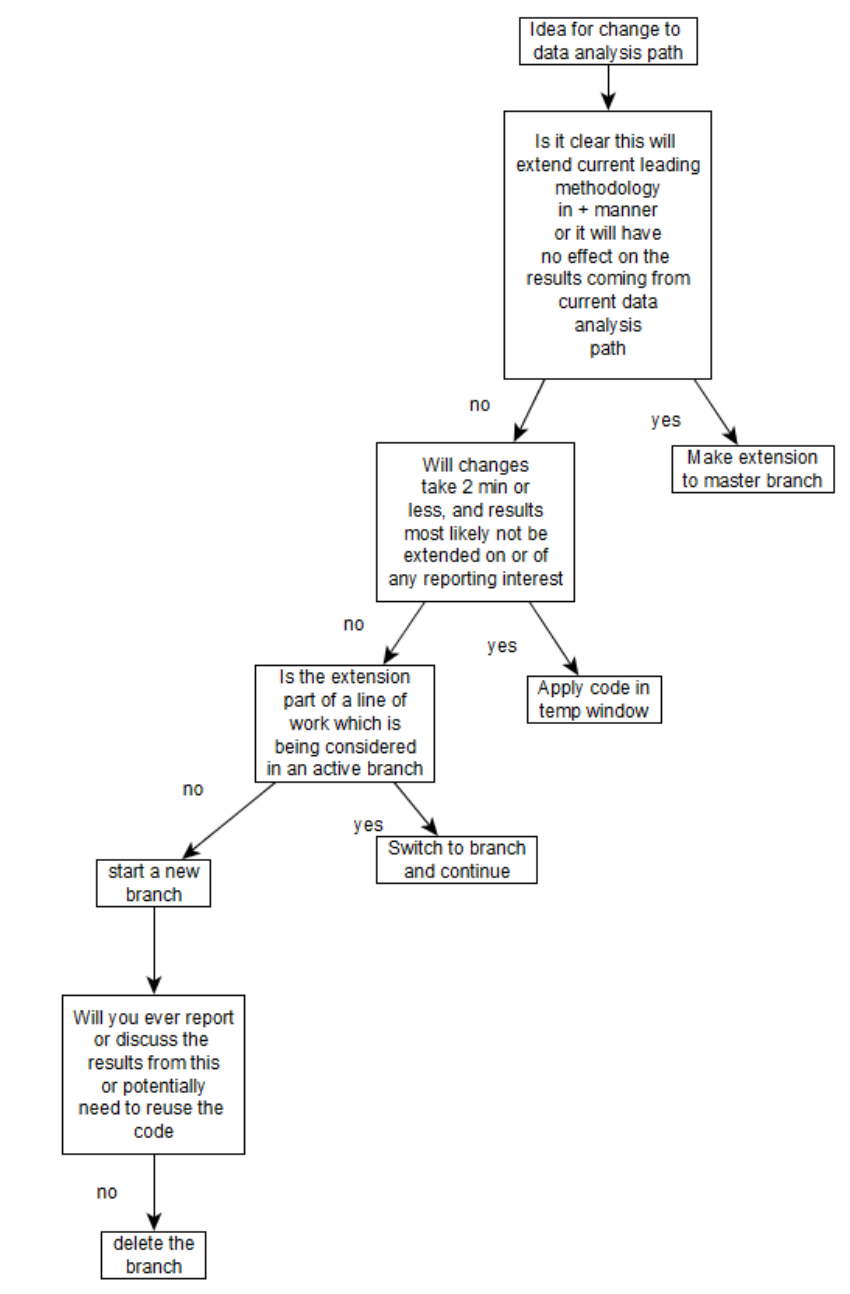
Definitions

- Leading process

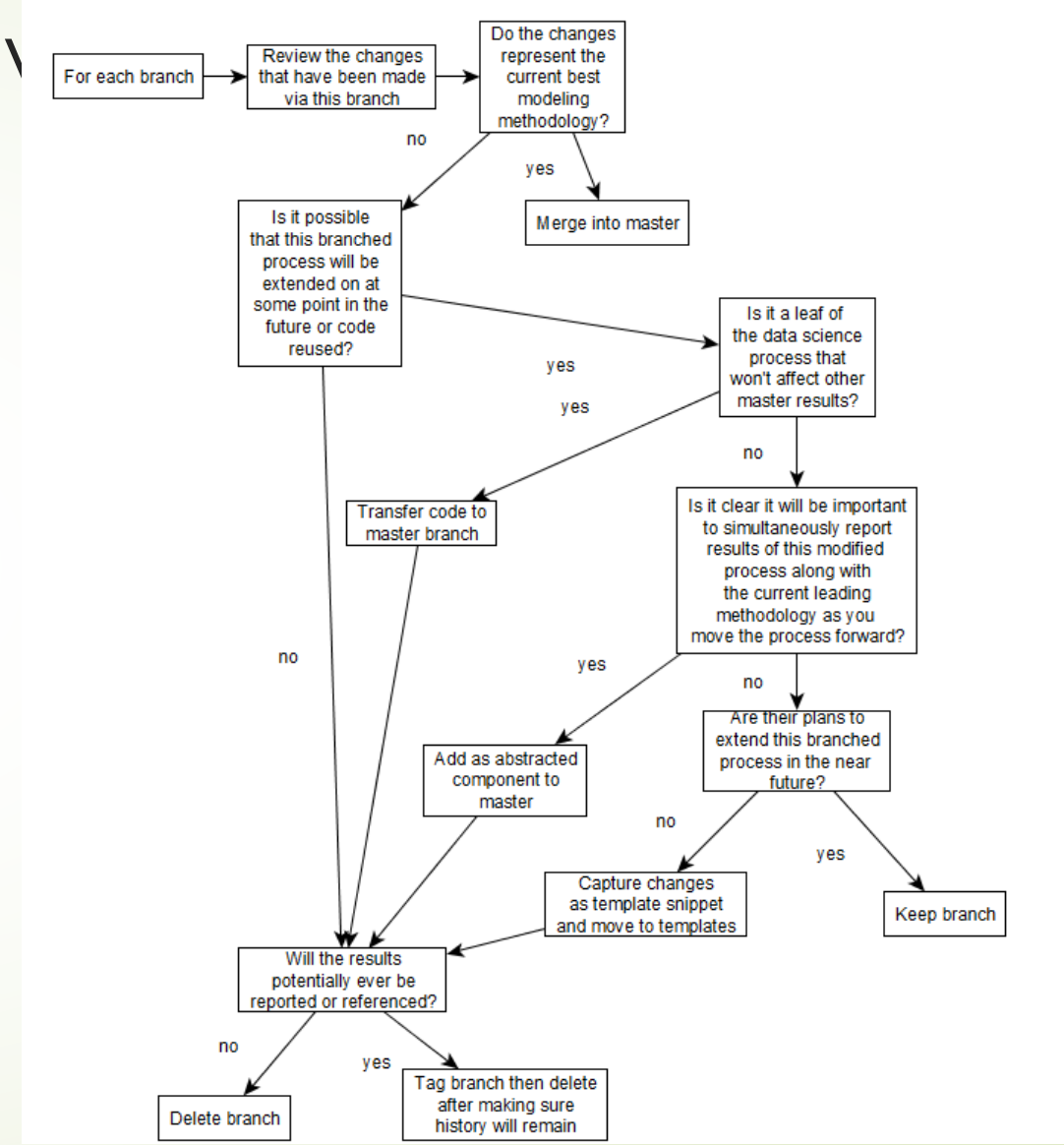


Where to put yo work initially

- Branches (dynamic): to hold active new ideas which may break functionality of the leading process
- Tags (fixed): to hold references to previous results
- merges (standard): to add work from idea that has turned into a leading process



Weekly repo review



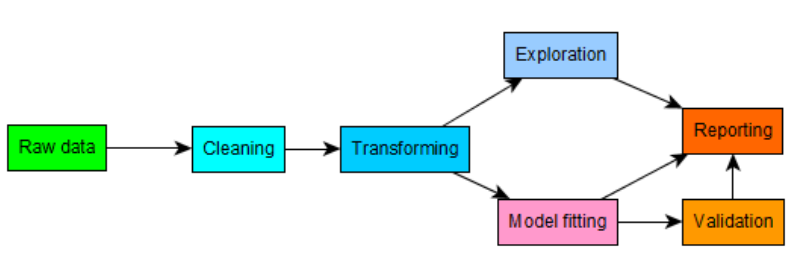
Example: Simulated some data

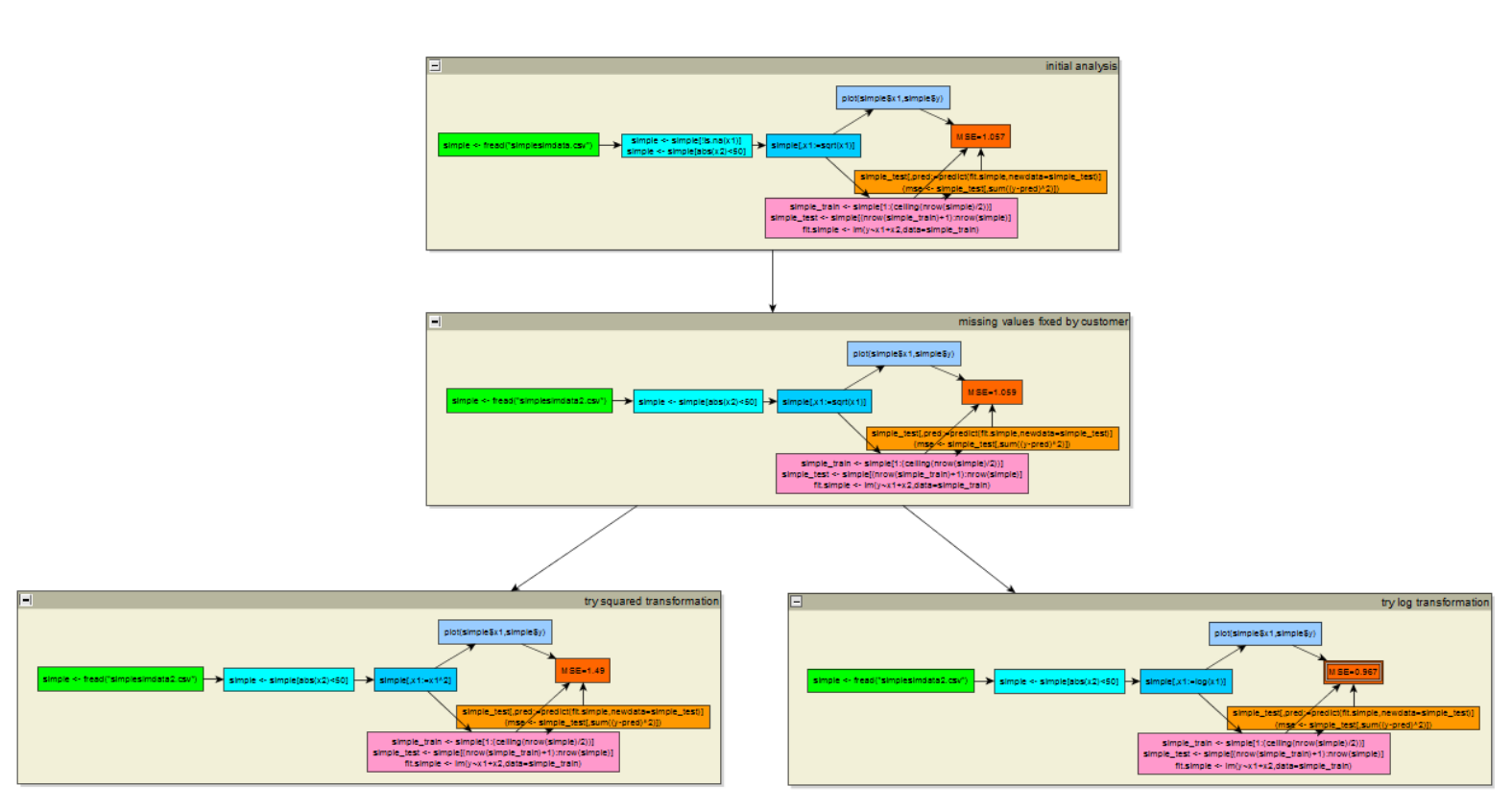
```
set.seed(101)
x1 <- rnorm(100,0,1)
x2 <- rnorm(100,0,1)
b1 <- 0.5
b2 <- 0.25
y <- b1*x1 + b2*x2 + rnorm(100,0,1)
x2 <- x2 + 100*rbinom(100,1,0.05)
data <- data.frame(y=y,x1=exp(x1),x2=x2)
write.csv(data,file="C:\\Users\\Clark\\Documents\\Notes\\Learning Git\\Data Science Git Principles\\simplesimdata.csv")

```

Start a repository

```
Clark@CLOVE ~/documents/notes/learning git/Data Science Git Principles/simplesim

$ git init
Initialized empty Git repository in c:/Users/Clark/documents/notes/learning git/
Data Science Git Principles/simplesim/.git/
```

Example: data analysis





Scripts vs. functions

- Scripts satisfy a very focused goal
- Functions satisfy a more general goal
- Functions are helpful for taking long processes and breaking them down into manageable chunks
- Functions are helpful for re-use of blocks of code when it becomes clear what will remain fixed and what will change
- Functions generally require slightly more thought/effort to construct as it may require accounting for all the possible cases of the input options