Demography SpawnR: A Way to Fake Data

Ishaan Dave

08/28/2019

Team at CDC

- Molecular Epi and Bioinformatics Team within NCHHSTP (at CDC, of course)
- Laboratory support of investigations of new/emerging retroviruses
- Bioinformatics support to Public Health agencies nationwide
  - Eg. internal cluster to manage/store data MTNAB
- Analytical support to other groups in DHAP
- Suite of tools/software found here

Organizational Chart

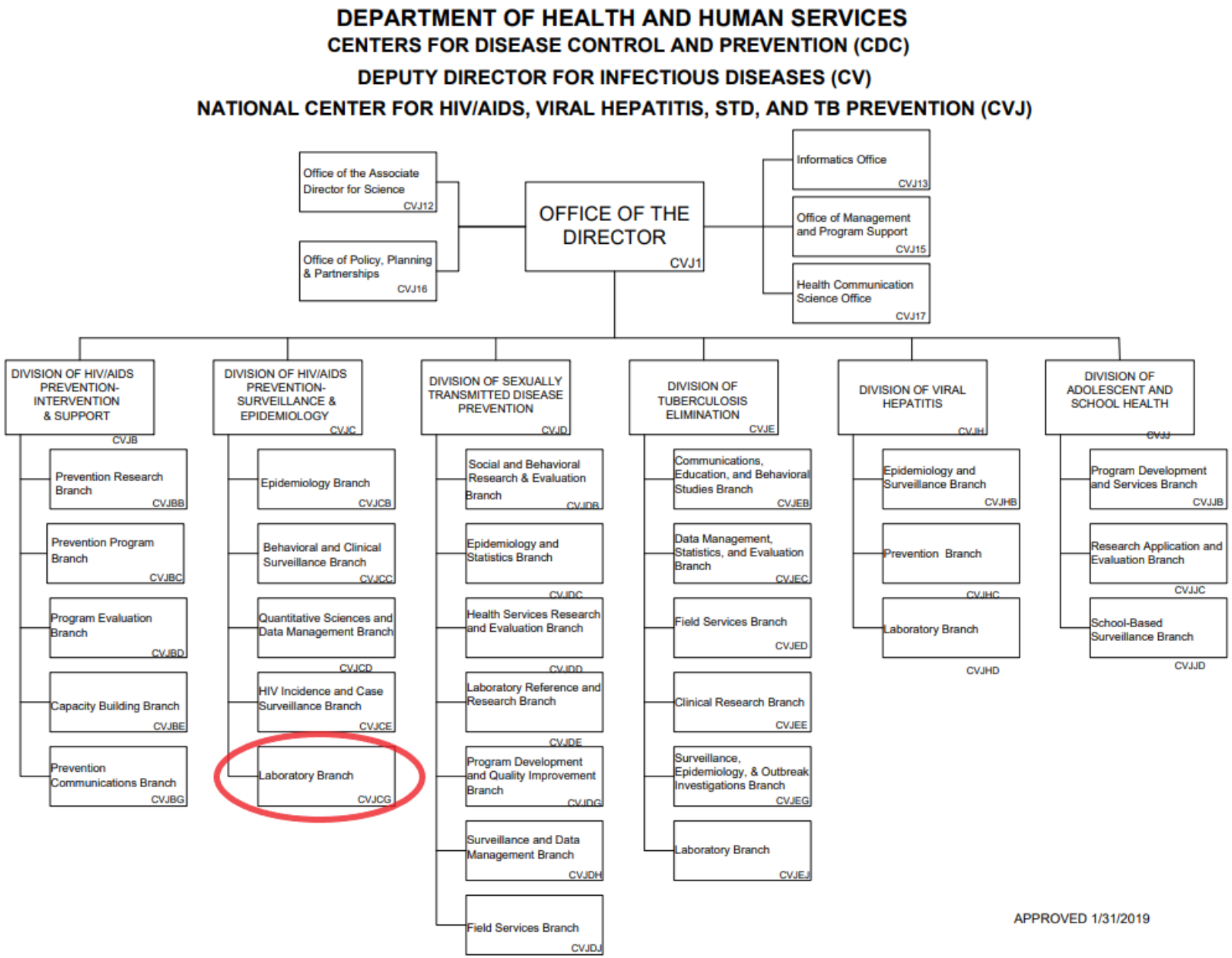


Figure 1: NCHHSTP Organizational Chart

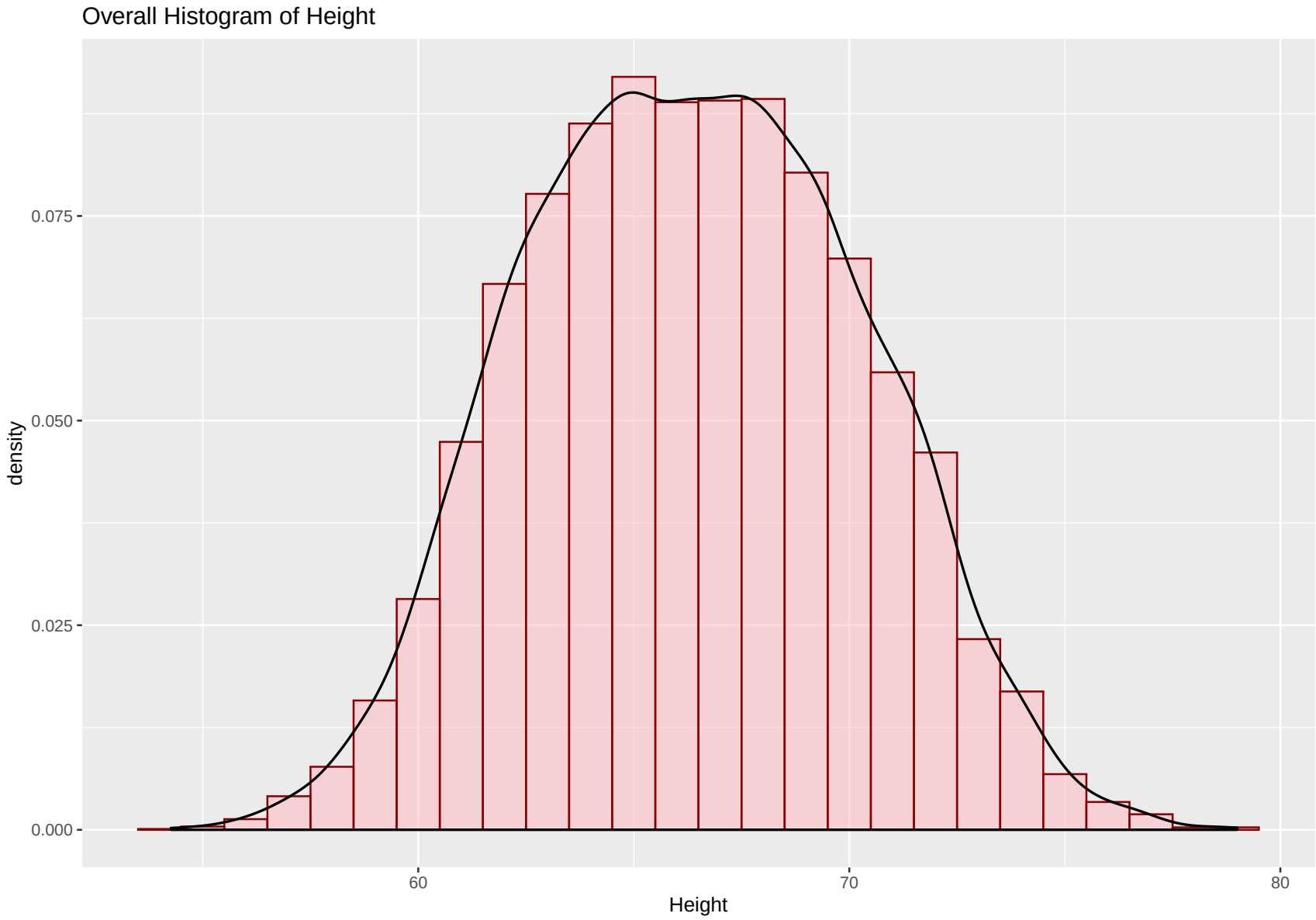
The Problem

- Oftentimes, this group/CDC uses data with personally identifiable information (PII)
- Vetting new tools, but can't use live data
  - Security restrictions with use of PII
  - Scalable?
- Not only a CDC/Public Health problem
  - Likely that every market Leidos works in has this issue. Plus could be used internally.

Potential Solution

- Demography SpawnR aims to solve this – "recreates" a dataset based on distributions of variables in the original
- What is a distribution?
  - Basically, it's a list/function that gives all possible outcomes and likelihood they occur
  - Most common is the *normal* distribution, or *the bell curve* (continuous)
  - Can also have frequency distributions

Example Normal Distribution



Example Frequency Distribution

Table 1: The Great M&M Data

Color	Frequency	Percentage
Brown	17	30.9%
Red	18	32.7%
Blue	7	12.7%
Yellow	6	10.9%
Green	4	7.3%
Orange	2	3.6%
Colorless/White	1	1.8%

But...

- Overall, heights ~65 inches
- Sometimes, we don't know the whole story – let's separate by gender
- There may be underlying patterns in the data we want to tease out
  - We just happen to know in this particular example

Height by Gender

A pattern!

- Males generally taller than females
- We'd like to recreate similar pattern in output dataset
  - (More on this later)

Now, what does this package do?

- Goes through variables and attempts to determine each type
  - Continuous, categorical, string, factor, dates, etc.
- A column with all different values is assumed to be sensitive information or PII
  - Name, address, SSN, etc.
    - *Usually*, these aren't important in analyses, replace with missing values/NA's.

How it works

- Computes/determines distributions of each variable
- If categorical, uses frequency/percentage of each level
- For continuous variables, populates with random values that follow a normal distribution with respective means/SD
- Dates
  - Generates kernel density estimate
  - Used that as "distribution" and samples – similar to an Epi curve

Missing Values

- Categorical variables – NA / missing is included as a category
- For continuous variables
  - percentage of missing values is calculated -> randomly inserted into each row with probability = original proportion

Decision Tree

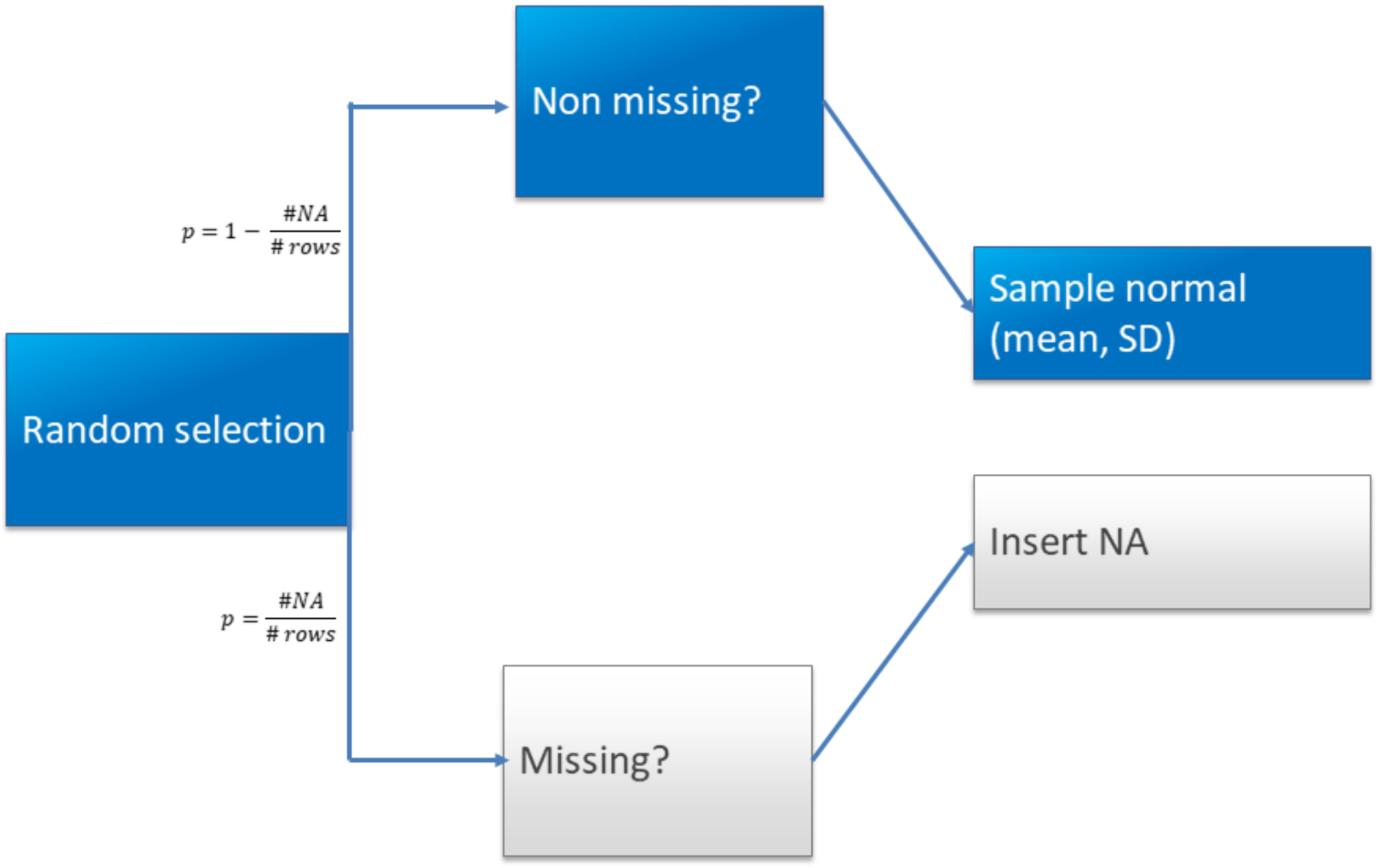


Figure 2: Decision tree to handle missing values with continuous data

Usage Example

Molecular HIV Surveillance

- Python, R, standalone, and interactive applications all require testing
- Live data is often unavailable, each stage can be properly vetted with the methods described

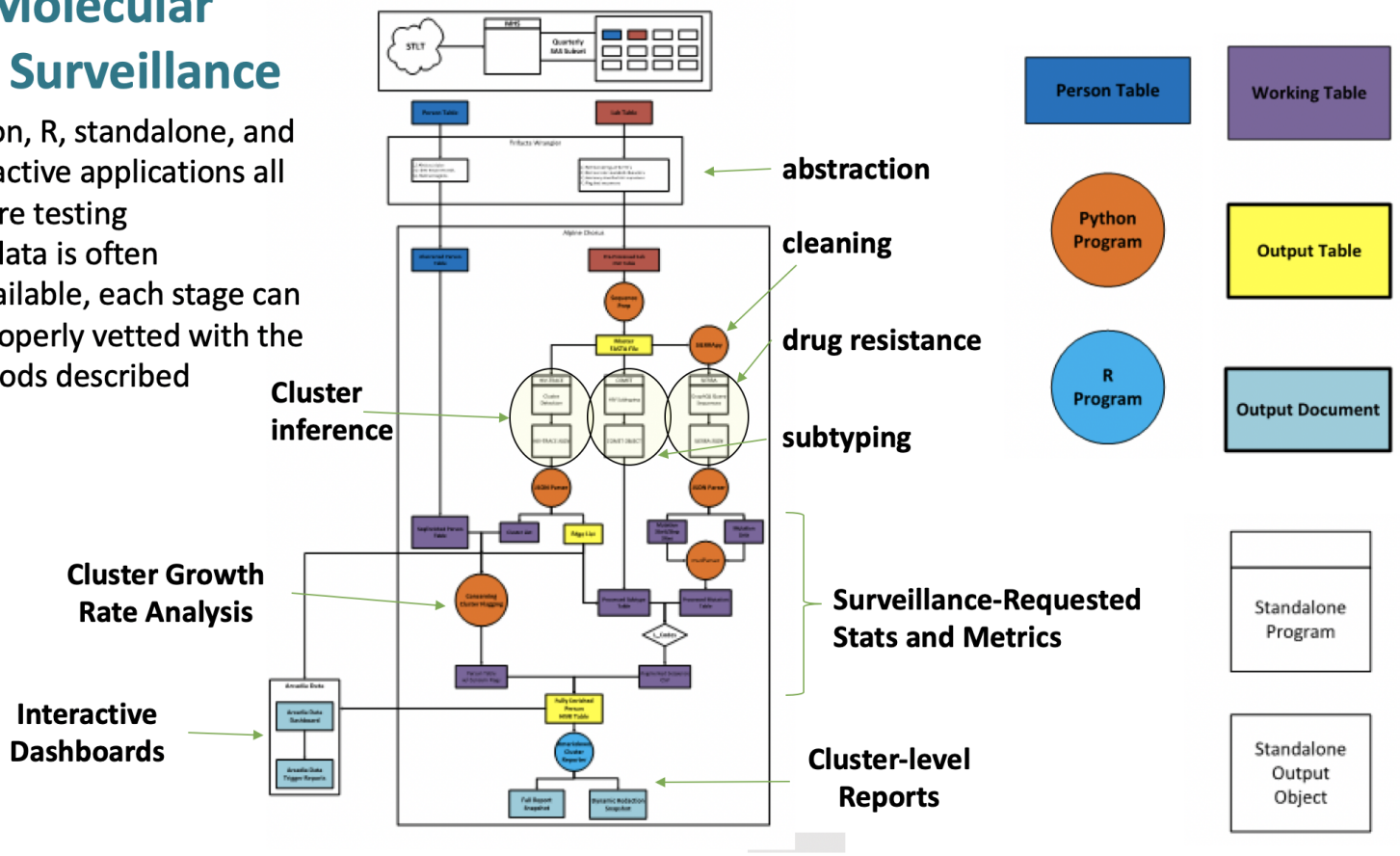


Figure 3: Example of MHS Pipeline

Other functionalities

- List all pairwise combinations of variables – continuous/continuous + categorical/categorical
- Correlations/associations and corresponding p-value for above combinations
- If user knows 2 variables to be correlated, able to input those and sample from bivariate distribution

Potential issues

- Handling with variables that contain zeros
- Categorical variables with several levels (e.g. > 10 but $< \#$ *of rows*)
- In bivariate sampling, variables strongly associated with 2+ others
  - Original: *Var A* associated with *Var B* and *Var C*
  - Sampled: No guarantees *Var A* associated with both after sampling
- Give user choice of which continuous distribution to use – lognormal, gamma, weibull, exponential, etc.
  - Or have package just pick best fitting distribution
- If working with dates – no way to guarantee *date2* comes after *date1* (e.g. patient starting/stopping drug)

Package Website

- https://cdcgov.github.io/DemographySpawnR/
- Or click **here**

Acknowledgements

- Tony Boyles
- Ellsworth Campbell
- Bill Switzer
- Sherry Ketemepi
- Stack Overflow

Comments, Questions, Concerns?

- Thanks!