

# Stat 133: Concepts in Computing with Data

---

Stat 133 by Gaston Sanchez

Creative Commons Attribution Share-Alike 4.0 International CC BY-SA

Github Repository

[github.com/ucb-stat133/  
stat133-spring-2018](https://github.com/ucb-stat133/stat133-spring-2018)

# About Stat 133

## Course Catalog Description *(a bit outdated)*

“An introduction to computationally intensive applied statistics. Topics will include organization and use of databases, visualization and graphics, statistical learning and data mining, model validation procedures, and the presentation of results.”

# Stat 133

Core Course for Statistics Major

# Stats Major

## *Prereqs*

Calculus

Calculus II

Multivariable  
Calculus

Linear  
Algebra

## *Core*

**Stat 133  
Computing**

Stat 134  
Probability

Stat 135  
Statistics

## *Elective*

Stat 150  
Stochastic  
Processes

Stat 151A  
Linear  
Modeling

Stat 152  
Sampling  
Surveys

Stat 153  
Times  
Series

Stat 154  
Predictive  
Modeling

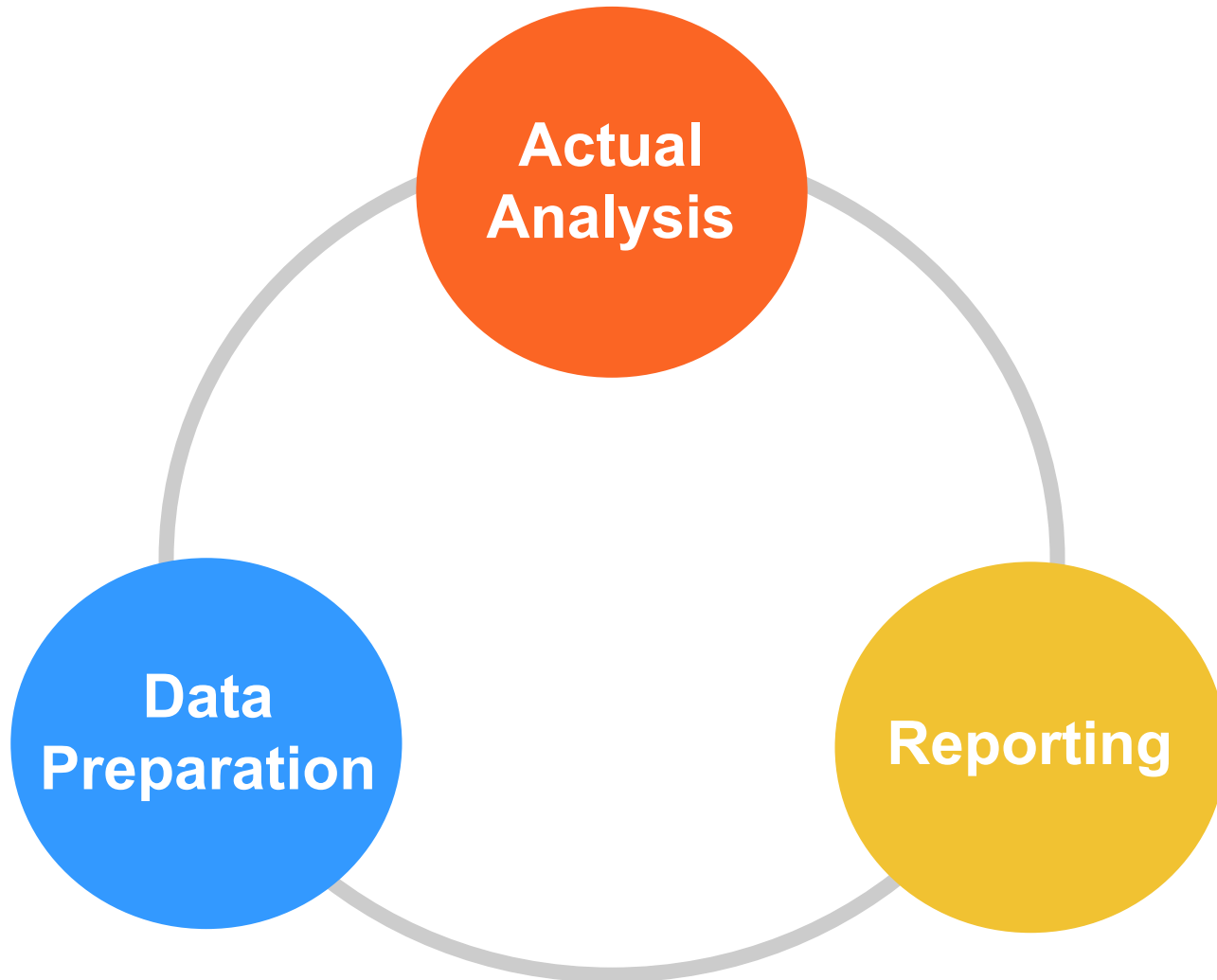
Stat 155  
Game  
Theory

Stat 158  
Design of  
Experiments

Stat 159  
Reproducible  
Research

# Data Analysis Cycle (DAC)

# My vision of the Data Analysis Cycle





# DATA: BY THE NUMBERS



JORGE CHAM © 2004

www.phdcomics.com

<http://www.phdcomics.com/comics/archive.php?comid=462>



## Data Preparation

- Acquisition
- Storage
- Cleaning
- Processing
- Tidying
- Reshaping
- Wrangling



## Analysis

- Exploration
- Description
- Visualization
- Hypothesis Tests
- Inference
- Simulation
- Model Fitting



## Reports

- Document(s)
- Article(s)
- Book(s)
- Poster(s)
- Blog post(s)
- Dissertation
- News



## Communication

- Oral
- Print
- Web
- Audio
- Video
- Multimedia
- Other

# Sad but true...



# Sad But True



Data



Analysis



Report



Communication

Traditionally, this is where most teaching focuses on

# Sad But True

**(ALMOST) NO ONE TEACHES THIS!**



Data



Analysis



Report

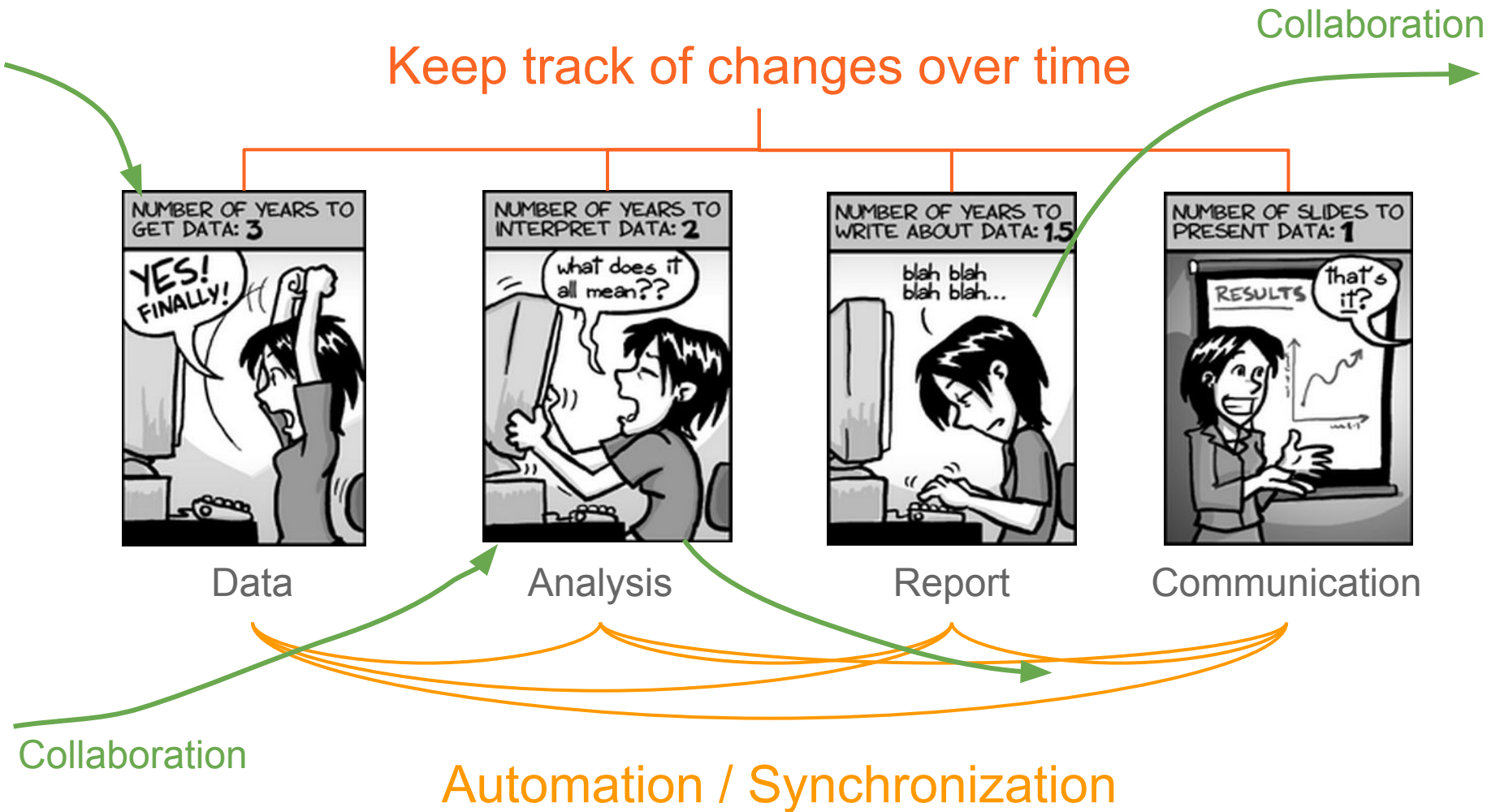


Communication

**In practice these are where we  
spend most of our time**



# Things to keep in mind ...



# Course Content

# Course cornerstones

Data  
Manipulation

Data  
Visualization

Reporting  
Tools

Programming  
Concepts

Data  
Technologies

**R**  
& other tools

**AND STATISTICAL CONCEPTS**

# Data Manipulation

1. Data Tables
2. Data Tidying
3. Selecting and Filtering
4. Reshaping
5. Aggregation & Group by operations
6. Joins and Merges

# Data Visualization

1. Visualization basics
2. Colors
3. Statistical graphics
4. Efficient displays
5. Design and Aesthetics considerations
6. Good and bad practices

# Programming concepts

1. Programming with an emphasis on data analysis
2. Data types and data structures
3. Control flow structures
4. Variables
5. Functions
6. Regular Expressions

# Reporting Tools

1. Markdown syntax
2. LaTeX (mostly equations)
3. Dynamic Documents
4. Shiny Apps
5. Writing reports

# Data Technologies

1. Data Tables
2. Unstructured data
3. HTML, XML, etc
4. Web scraping (Web API's)
5. JSON
6. Relational Databases (SQL)



## R and other tools

1. R
2. RStudio
3. Command Line (Bash)
4. Version control with Git
5. Hosting with Github
6. etc

# Statistical Concepts

1. Basic Numeracy: variability, patterns, comparisons
2. Apply introductory concepts
3. Methods: regression, classification, dimension reduction
4. Missing Data: imputation, treatment
5. Simulation: Monte Carlo, bootstrap, etc

# Course Resources

## Github Repository

[github.com/ucb-stat133/stat133-spring-2018](https://github.com/ucb-stat133/stat133-spring-2018)

# Some Comments

## The R course?

Typically known as the R course

Mea Culpa

R is just the means

The goal is to introduce you to different aspects of the Data Analysis cycle

It just happens that we use R as the main computational tool

## Course Format

Lectures: conceptual stuff, demos, case studies, examples, review some code

Labs: practical work using R

Homework: follow the work of labs, plus some challenges

Posts: Give you the opportunity to do some research on specific topic(s)

## My Expectations

Don't expect you to become an expert (that takes a lot of time)

Don't expect you to become hackers

Instead: give you solid foundations about data analysis

Expose you to different “data technologies”



## Ultimate Goals

Understand different types of data (e.g. files, forms, formats)

Know how to access information stored in different formats

Know how to do data manipulation and processing in R

Be better prepared to crunch data

Becoming a data scientist  
is a **marathon** not a sprint

Next Week

# Install Software

R

RStudio