

Course Introduction

Stat 133 - Concepts in Computing with Data

Gaston Sanchez

Department of Statistics, UC-Berkeley

gastonsanchez.com

github.com/gastonstat/stat133

Course web: gastonsanchez.com/teaching/stat133

Concepts in Computing with Data

Hello
my name is

Gaston

A little bit about me

- ▶ Originally from Mexico
- ▶ Applied Statistician
- ▶ Statistical Programmer
- ▶ Data Analyst - Consultant
- ▶ Lecturer





making analytical tools

Concepts in Computing with Data?

Concepts in Computing with Data?

Computational Data Analysis

How to use computational tools
to conduct a statistical analysis
of data

How to use computational tools
to conduct a statistical analysis
of data

And thinking about Data
Analysis

Data Analysis

“Data Analysis is the process by which data becomes understanding, knowledge and insight”

Hadley Wickham

Data Analysis

- ▶ Use the computer expressively to conduct statistical analysis of data
- ▶ Use existing software rather than build routines from the ground up
- ▶ Focus on aspects of computing to conduct statistical analysis, NOT the computational aspects of statistical methods

Quick Questions

Major

How many of you are

- ▶ Stats / Math majors?
- ▶ Non-stats majors?
- ▶ Double majors?

Data Analysis Experience

What's your data analysis experience?

- ▶ I'm completely new to data analysis
- ▶ I've analyzed data in Excel
- ▶ I've used statistical software (SAS, SPSS, etc)

Programming Experience

What's your programming experience?

- ▶ I have no programming experience
- ▶ I have some programming experience
- ▶ I've written some scripts in R, Python, Matlab
- ▶ I've used the command line (shell or terminal)

Writing Documents

When writing Docs and Reports

- ▶ I use word processors:
Word, Pages, Google Docs, etc
- ▶ I use typesetting systems:
LaTeX, markdown + pandoc

Preparing Slides

When preparing slides

- ▶ I use Power Point, Keynote, Google Slides
- ▶ I use LaTeX beamer, markdown, HTML5

Syllabus

Course webpage

gastonsanchez.com/teaching/stat133

Tentative Content

- ▶ R Basics
- ▶ Statistical Graphics
- ▶ Exploratory Data Analysis
- ▶ Programming in R
- ▶ Regular Expressions
- ▶ Data Manipulation
- ▶ Data Technologies
- ▶ Simulations
- ▶ Reporting and Communicating

Course Info

- ▶ ~ 15 weeks (Aug 26 - Dec 14)
- ▶ ~ 45 hours of lecture
- ▶ ~ 30 lab hours
- ▶ 3 units

Course Work

Course Work

- ▶ ~ 8 Homework assignments
 - ▶ ~ 8 Lab assignments
 - ▶ 2 Exams (midterm Oct-22/23, final Dec-14/15)
 - ▶ 1 Individual project
 - ▶ 1 Group project
- more about the projects in the next weeks

Overall Score

| Value | Concept |
|-------|----------------------|
| 10% | Participation |
| 15% | Lab |
| 25% | Homework |
| 25% | Midterm & Final exam |
| 25% | Projects |

Attendance Policy

You are expected to attend all lectures and lab discussions

Office Hours

To be announced ... or by appointment (preferred)

Course Policies

DO's



DONT's



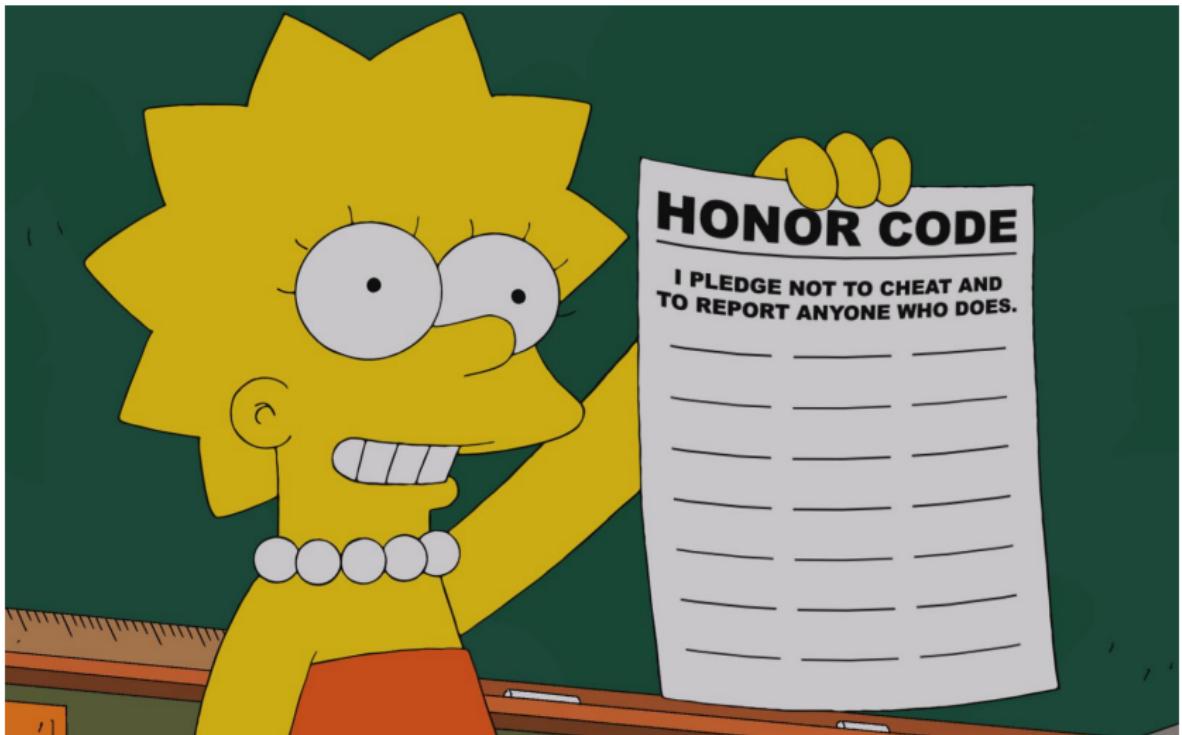






Please don't





Academic Integrity

- ▶ Write your own scripts and code
- ▶ You can share ideas and exchange suggestions, but code must be yours
- ▶ If you use someone else's code (e.g. find it online), give credit
- ▶ Plagiarism won't be tolerated

Email Policy

No email

Email

Email Policy

- ▶ Use email as a tool to set up a one-on-one meeting
- ▶ Use the subject line **Meeting Request**
- ▶ Include at least two times when you would like to meet
- ▶ Include a brief (one-two sentence) description of the reason for the meeting
- ▶ Email sent for any other reason will NOT be considered or acknowledge
- ▶ Don't expect me to reply right away

I'm a data analyst, not a professional email responder

Let's talk

Q & A's

I strongly encourage you to ask questions about the syllabus and assignments during class time.

For more in-depth discussions (such as guidance on assignments) please plan to meet in person.

Resources

Course webpage

<http://gastonsanchez.com/teaching/stat133>

Github Repo

<https://github.com/gastonstat/stat133>

About this course

Computing with Data (CwD)

Computing with Data (CwD)

Computing with Data (CwD)

Computing with Data (CwD)

- ▶ CwD means everything and nothing at the same time
- ▶ Data Analysis
- ▶ Data Manipulation
- ▶ Statistical Programming

Computing with Data

“Computing with data refers to activities in which data is acquired, managed, and processed for a great variety of purposes: organization, visualization, summaries, analysis, etc”

John Chambers

Computing with Data

“Computing with data refers to activities in which data is **acquired, managed, and processed** for a great variety of purposes: **organization, visualization, summaries, analysis, etc**”

John Chambers

Understanding the Data Analysis Process

DATA: BY THE NUMBERS



JORGE CHAM © 2004

www.phdcomics.com

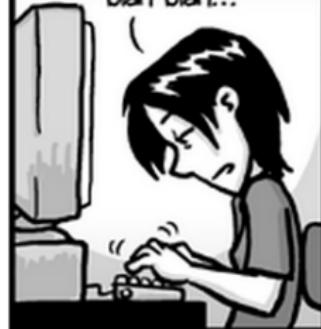


NUMBER OF YEARS TO
INTERPRET DATA: 2



NUMBER OF YEARS TO
WRITE ABOUT DATA: 1.5

blah blah
blah blah...





Data Analysis Stages

Workflow

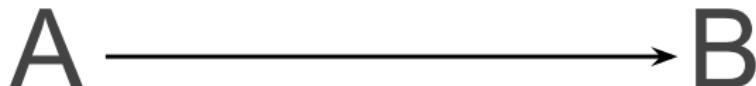
Data Workflow

1. Acquisition-collection
2. Processing and Cleaning
3. Exploration and Visualization
4. Modeling
5. Reporting and Communication

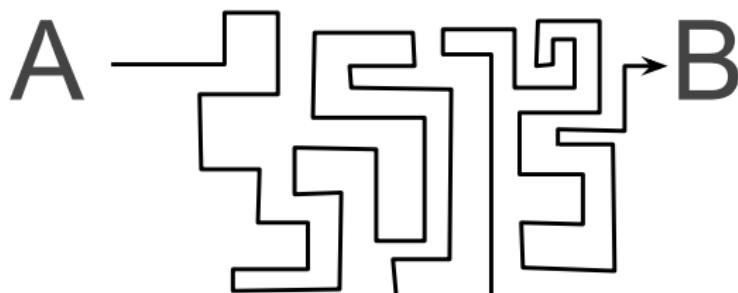
Main Stages of a Data Analysis Cycle



This is not a linear workflow;
there are iterative cycles at any stage



ideal linear data analysis process



data analysis process in practice

Data Analysis is a lot like
cooking & baking

Work Environment



Tools & Utensils



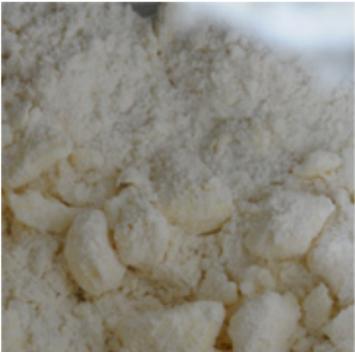
Raw Data



Processing, Cleaning, Organizing



Manipulation



Modeling



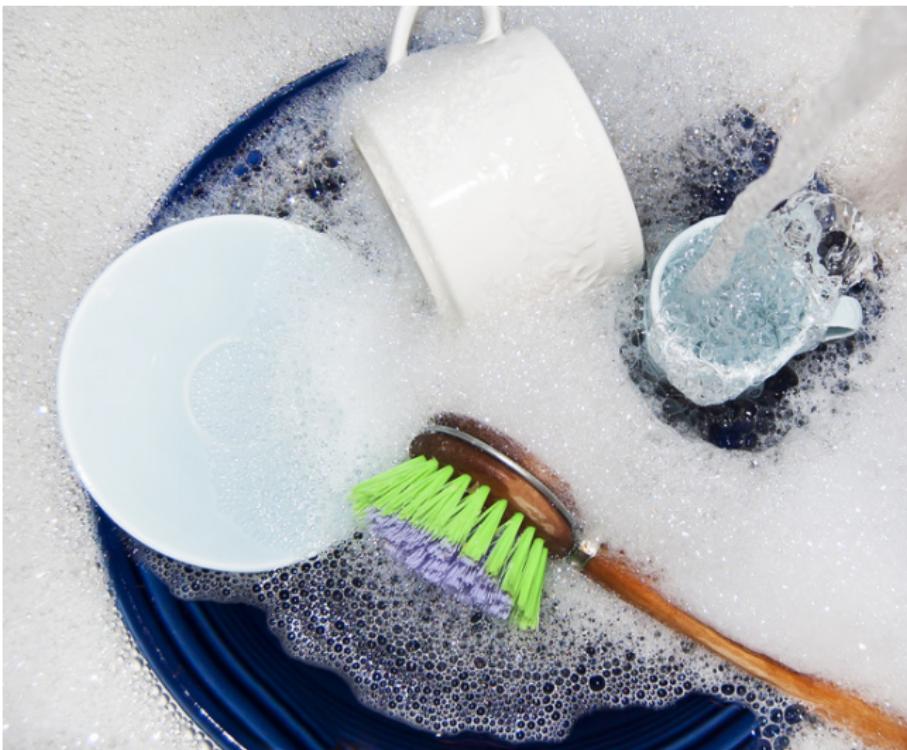
Presentation & Consumption



Good -vs- Bad Practices



Housekeeping



Becoming a data scientist
is a **Marathon**, not a sprint

Working Environment

Tools

Tools for this course

- ▶ Computer
- ▶ Text Editor
- ▶ R (software for data analysis)
- ▶ RStudio (IDE)
- ▶ Terminal or Command Line
- ▶ Latex
- ▶ git*
- ▶ github account*

Tools

Text editors for OS X, Windows and Linux

Choose a text editor of your preference:

- ▶ Emacs: <http://www.gnu.org/software/emacs/>
- ▶ Atom: <https://atom.io/>
- ▶ Sublime Text: <http://www.sublimetext.com/>
- ▶ Vim <http://www.vim.org/download.php>
- ▶ TextWrangler (Mac OS X only):
<http://www.barebones.com/products/textwrangler/>
- ▶ Notepad ++ (Windows only): <https://notepad-plus-plus.org/>



[Home]

Download

[CRAN](#)

R Project

[About R](#)

[Contributors](#)

[What's New?](#)

[Mailing Lists](#)

[Bug Tracking](#)

[Conferences](#)

[Search](#)

R Foundation

[Foundation Board](#)

[Members](#)

[Donors](#)

[Donate](#)

Documentation

[Manuals](#)

[FAQs](#)

[The R Journal](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 3.2.0](#) (Full of Ingredients) has been released on 2015-04-16.
- [R version 3.1.3](#) (Smooth Sidewalk) has been released on 2015-03-09.
- [The R Journal Volume 6/2](#) is available.
- [useR! 2015](#), will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

Getting R

Download R

- ▶ R project website
<http://www.r-project.org>
- ▶ Go to CRAN
<http://cran.r-project.org/mirrors.html>
- ▶ Select closest mirror, e.g.
<http://cran.cnr.Berkeley.edu/>
- ▶ Choose the right version for your Operating System
(e.g. Linux, Mac, Windows)

Getting RStudio

Download RStudio

- ▶ Rstudio
<http://www.rstudio.com/>
- ▶ RStudio desktop
<http://www.rstudio.com/products/rstudio/download/>
- ▶ Choose the right version for your Operating System
(e.g. Linux, Mac, Windows)

Tools

Rtools (for Windows only)

If you work with Windows, you'll need **Rtools**:

<http://cran.r-project.org/bin/windows/Rtools/>

Command Line Tools (Mac)

If you've never used the Mac Terminal before, it's likely that you'll need to install the **Command Line Tools** (you'll know if you need this).

Tools

LaTeX

If you don't have it, install LaTeX:

- ▶ TeX Live (Linux): <http://www.tug.org/texlive/>
- ▶ MacTeX (Mac OS X): <http://www.tug.org/mactex/>
- ▶ ProTeXt (Windows): <http://www.tug.org/protext/>

Tools

git

For all platforms:

<http://git-scm.com/book/en/v2/Getting-Started-Installing-Git>

github

If you don't have it yet, open a (free) account in github

<https://github.com/join>