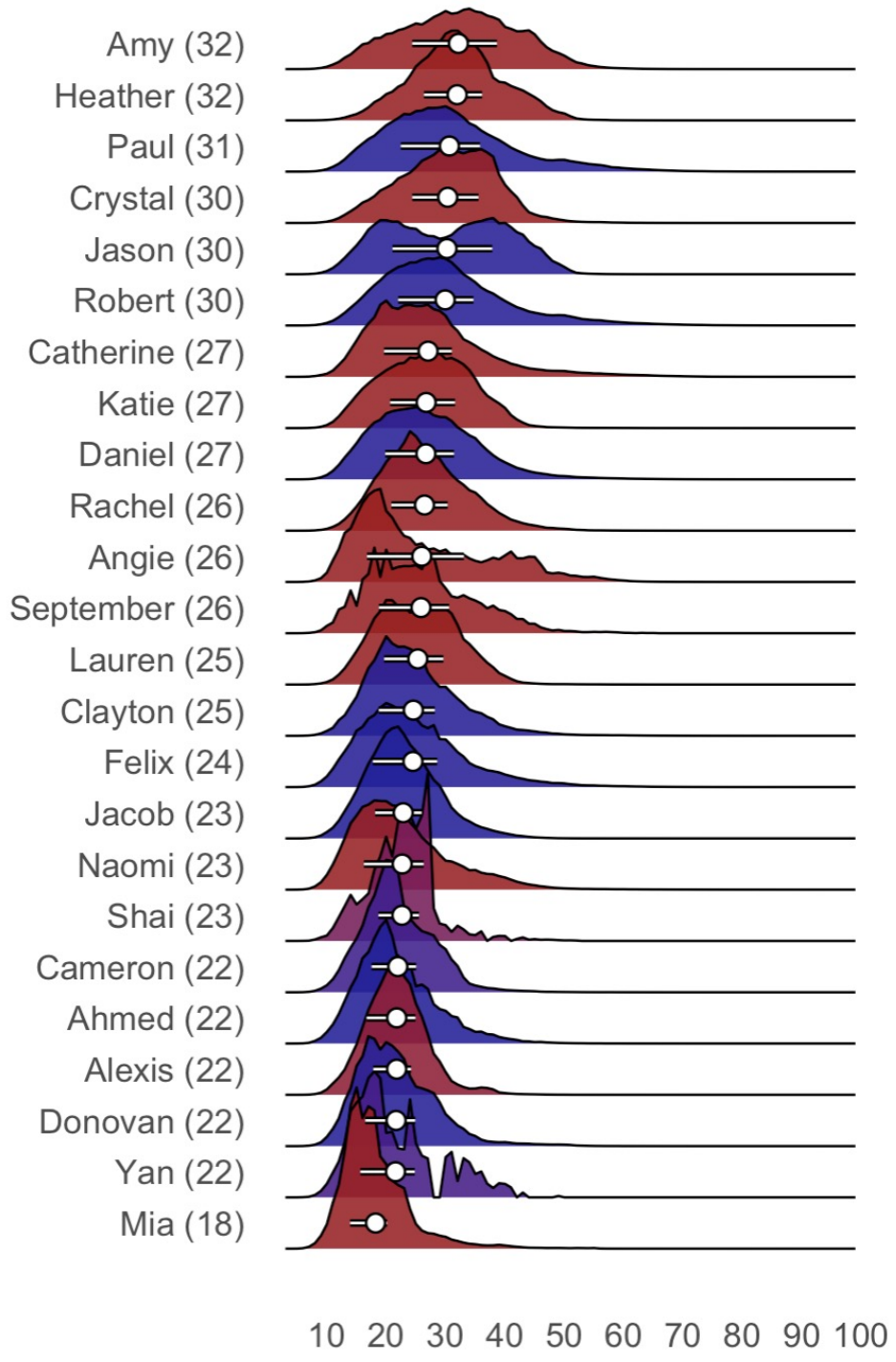
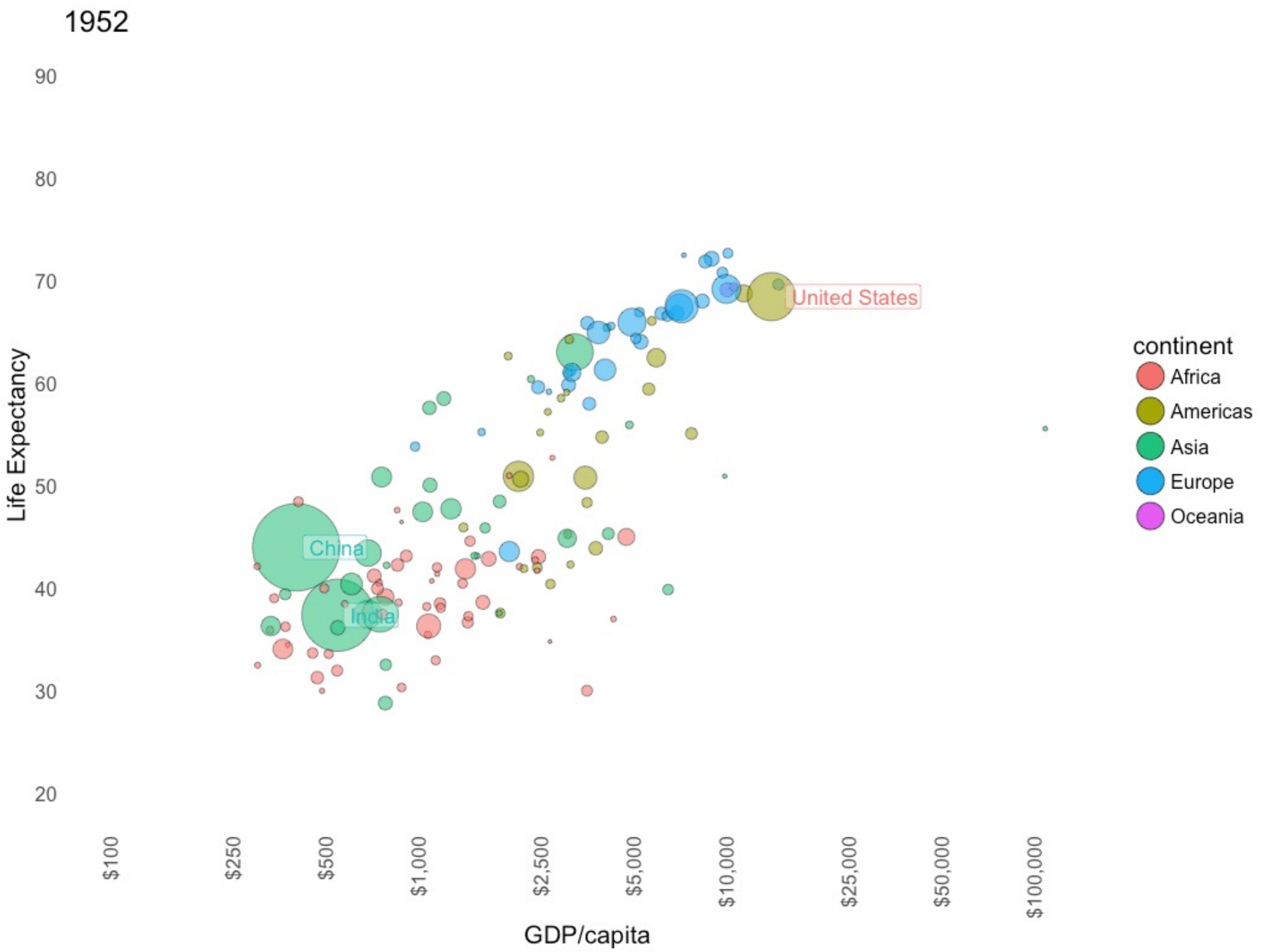
201ab / 193
Quantitative methods
L.00: Introduction

Website: http://vulstats.ucsd.edu/

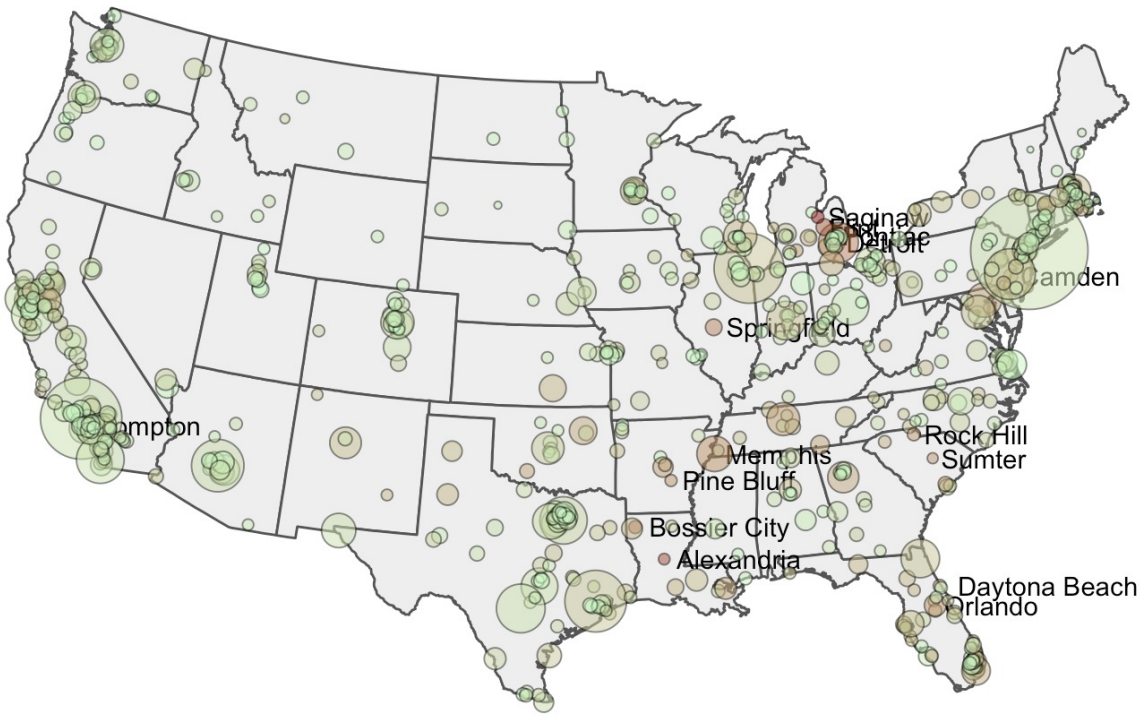
Instructors:
Ed Vul
Wenhao Qi

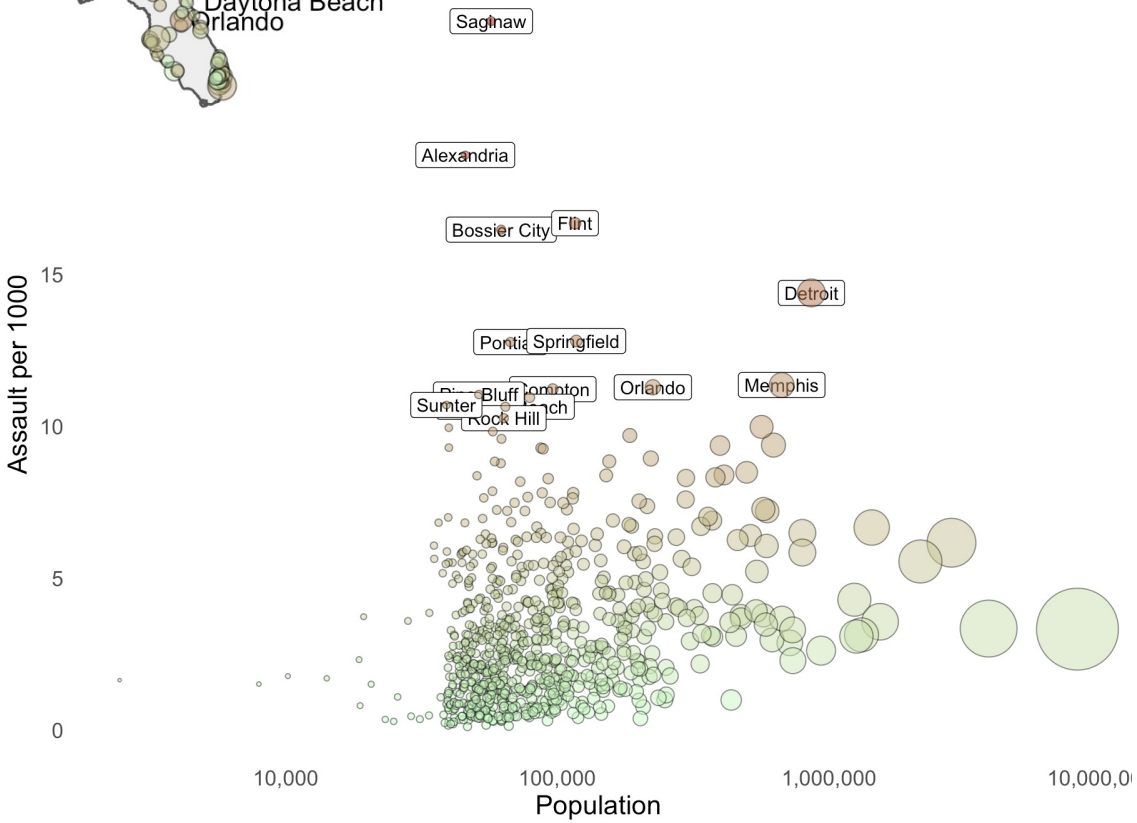
```
> glimpse(babynames::babynames)
Observations: 1,858,689
Variables: 5
$ year <dbl> 1880, 1880, 1880, 1880, 1880, 1...
$ sex  <chr> "F", "F", "F", "F", "F", "F", "...
$ name <chr> "Mary", "Anna", "Emma", "Elizab...
$ n    <int> 7065, 2604, 2003, 1939, 1746, 1...
$ prop <dbl> 0.07238433, 0.02667923, 0.02052...
>
> babynames::babynames
# A tibble: 1,858,689 x 5
    year   sex      name     n       prop
   <dbl> <chr>     <chr> <int>      <dbl>
 1  1880     F      Mary  7065 0.07238433
 2  1880     F      Anna  2604 0.02667923
 3  1880     F      Emma  2003 0.02052170
 4  1880     F Elizabeth  1939 0.01986599
 5  1880     F    Minnie  1746 0.01788861
 6  1880     F  Margaret  1578 0.01616737
 7  1880     F       Ida  1472 0.01508135
 8  1880     F     Alice  1414 0.01448711
 9  1880     F    Bertha  1320 0.01352404
10  1880     F     Sarah  1288 0.01319618
# ... with 1,858,679 more rows
```



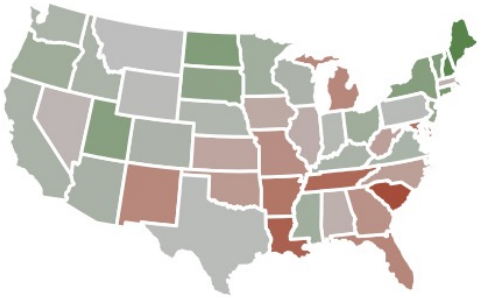

1952
90
80
70
60
50
40
30
20
Life Expectancy
United States
China
India
continent
Africa
Americas
Asia
Europe
Oceania
$100
$250
$500
$1,000
$2,500
$5,000
$10,000
$25,000
$50,000
$100,000
GDP/capita

FBI crime stats: Assaults

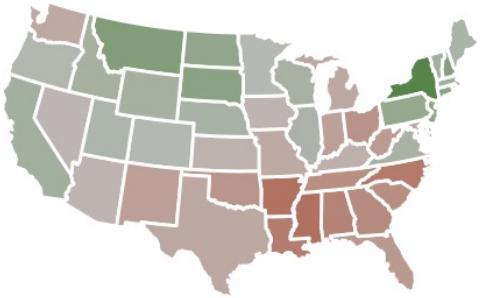




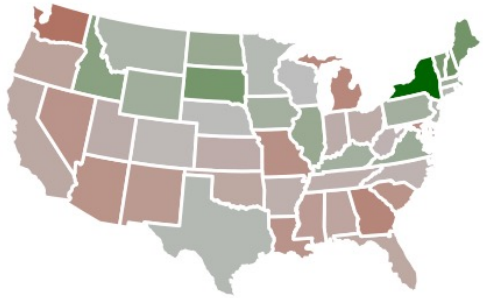
Assault



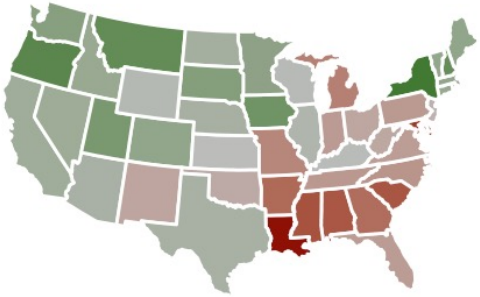
Burglary



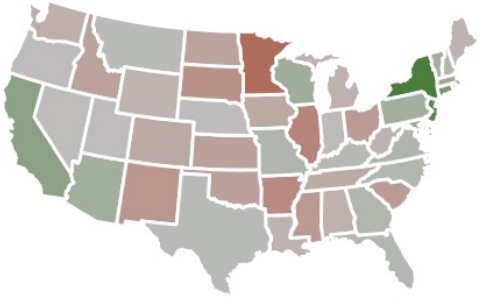
GTA



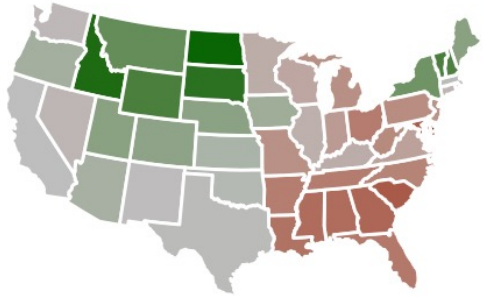
Murder



Rape

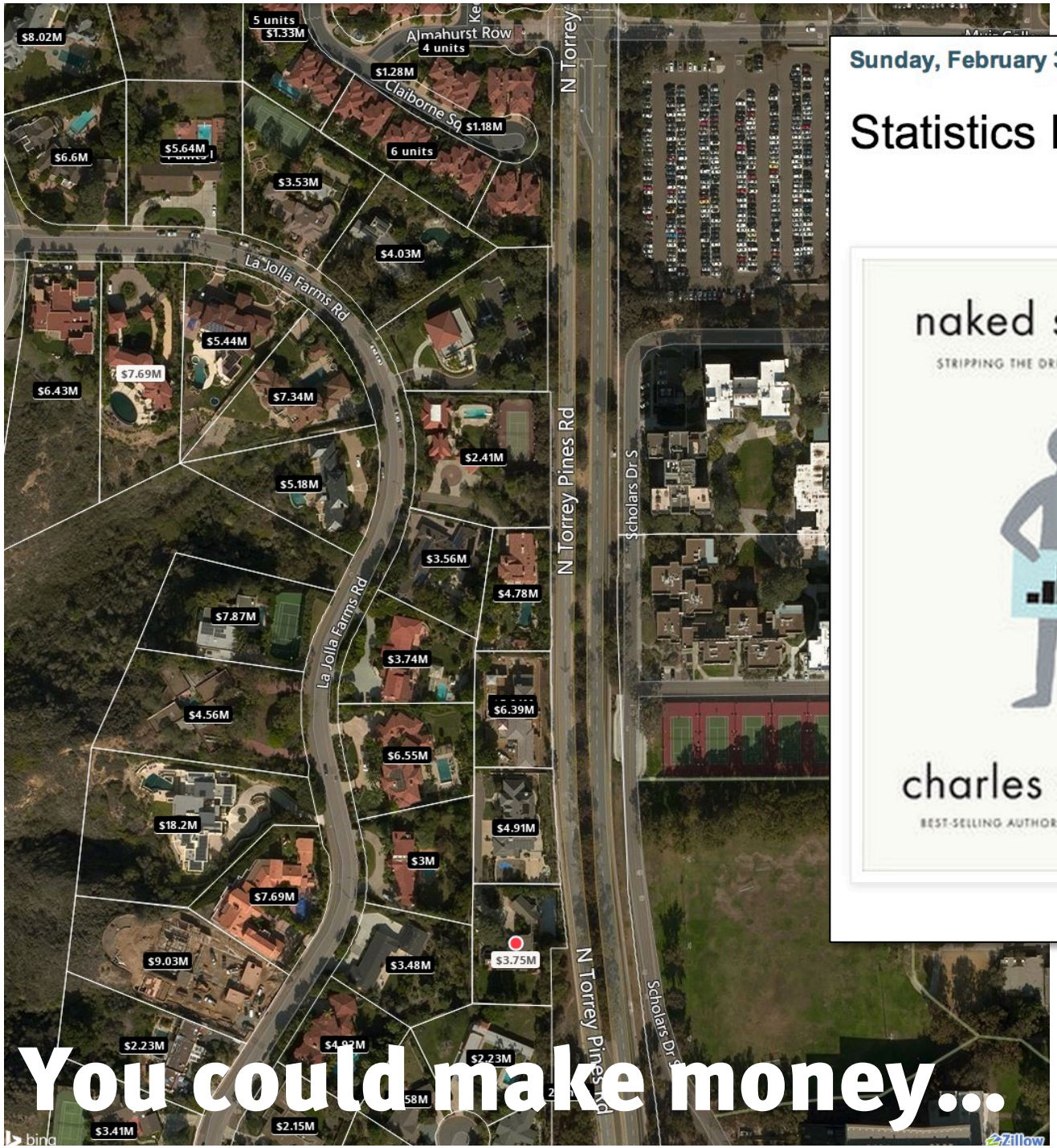


Robbery



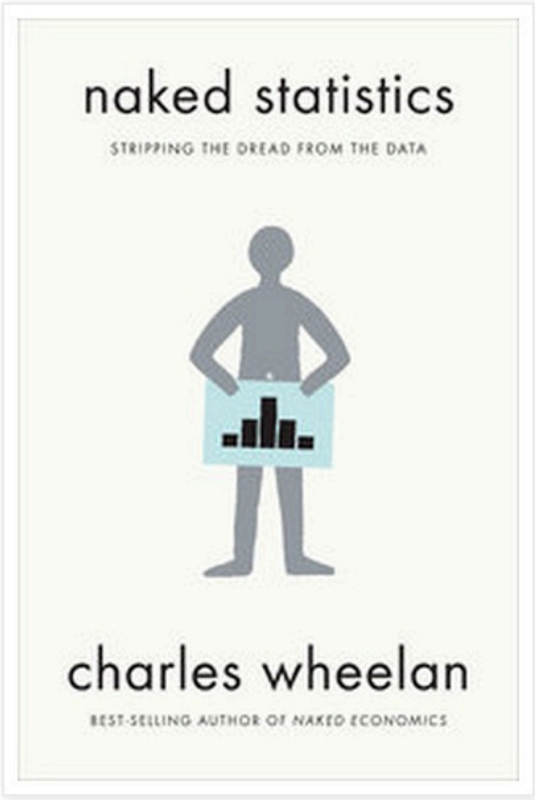
How much lower/higher is the rate of a given crime in each state from what we would expect based on city populations?

(random effect of crime:state combination in a mixed effect Poisson regression of FBI crime counts in each city)



Sunday, February 3, 2013

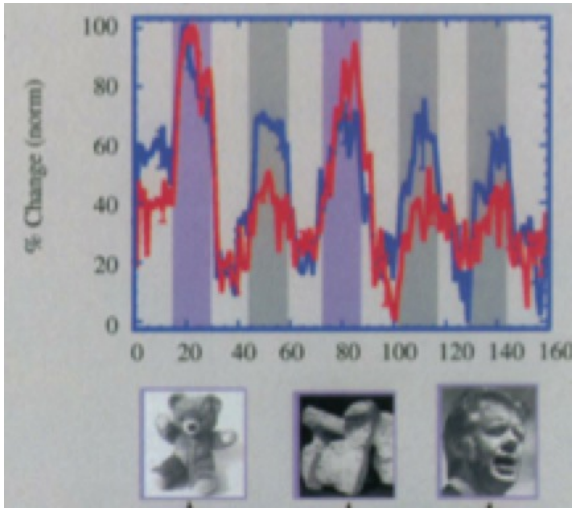
Statistics Is Suddenly Sexy!

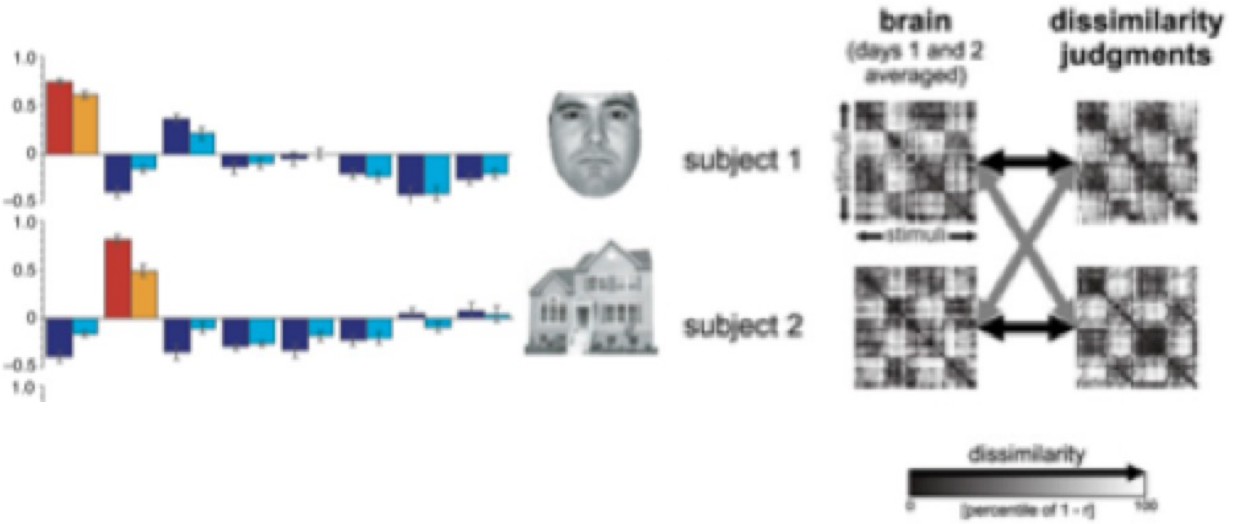


As someone w
have foreseen
become in curr
statistics as a
then). And I ha
The Noise" -- I
But I love the v
Bayes): "**Nake**
most palatable
overly-technica
examples (not

The book has a
medians and 'c
central limit the

You could make money...





...You could make new discoveries...

How Reliable Are Psychology Studies?

A new study shows that the field suffers from a reproducibility problem, but the extent of the issue is still hard to nail down.

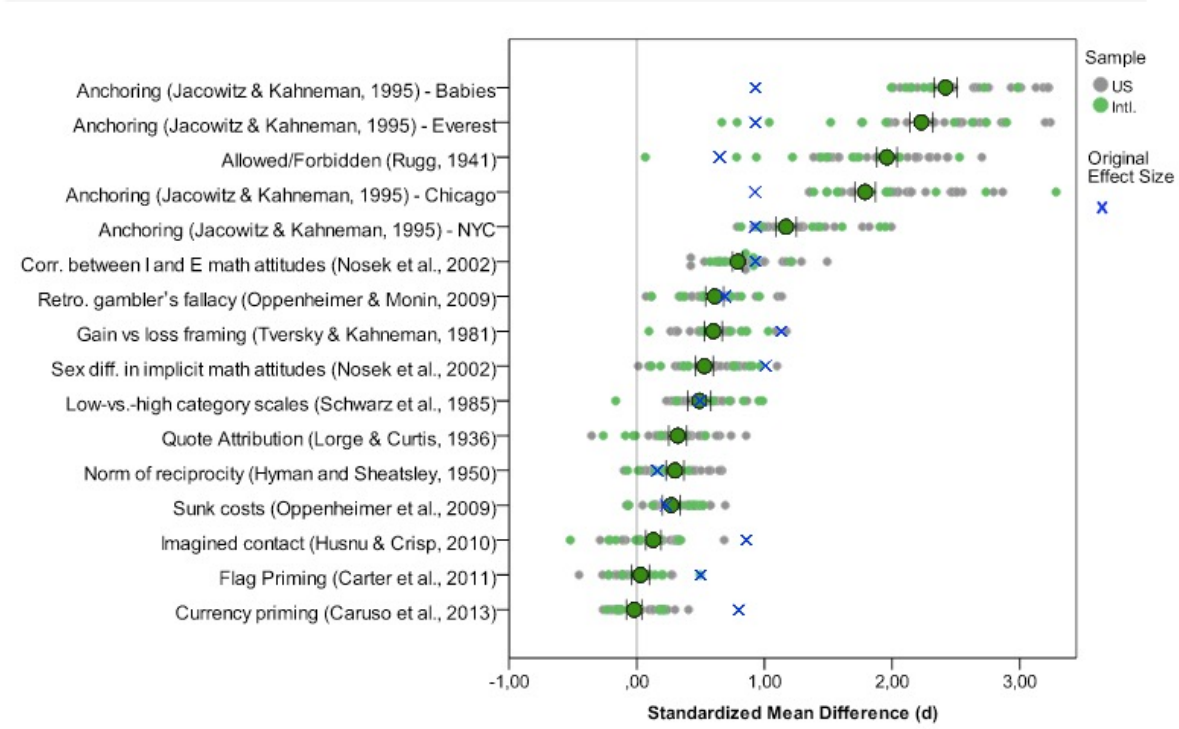
Psychology's Replication Crisis Can't Be Wished Away

It has a real and heartbreaking cost.

In cancer science, many "discoveries" don't hold up

Why Most Published Research Findings Are False

John P. A. Ioannidis



...or you could embarrass yourself.

Important data skills by research role

- Consumer
  - Read / interpret data presentations: common graphs, statistics, results
  - Basic numeracy, relevant questions for particular statistics
- Reviewer
  - Sufficient understanding to spot analysis/report mismatch
  - Reason from reported stats/data rather than written words
  - Know which deviations from model assumptions produce which biases
  - Come up with more diagnostic graphs, and a more incisive analyses
- Producer
  - Given data: read it, clean it, make it usable
  - Given a postulated relationship: make diagnostic graphs, identify relevant statistic, estimate with uncertainty, compare to null model
  - Given a vague description of a relationship: make it precise, identify relevant variable, decide on appropriate form of model
- Synthesizer: Translate between statistics, relate uncertainty across studies.
- Path-breaker: reason from first principles to develop new methods…

What we aim to cover.

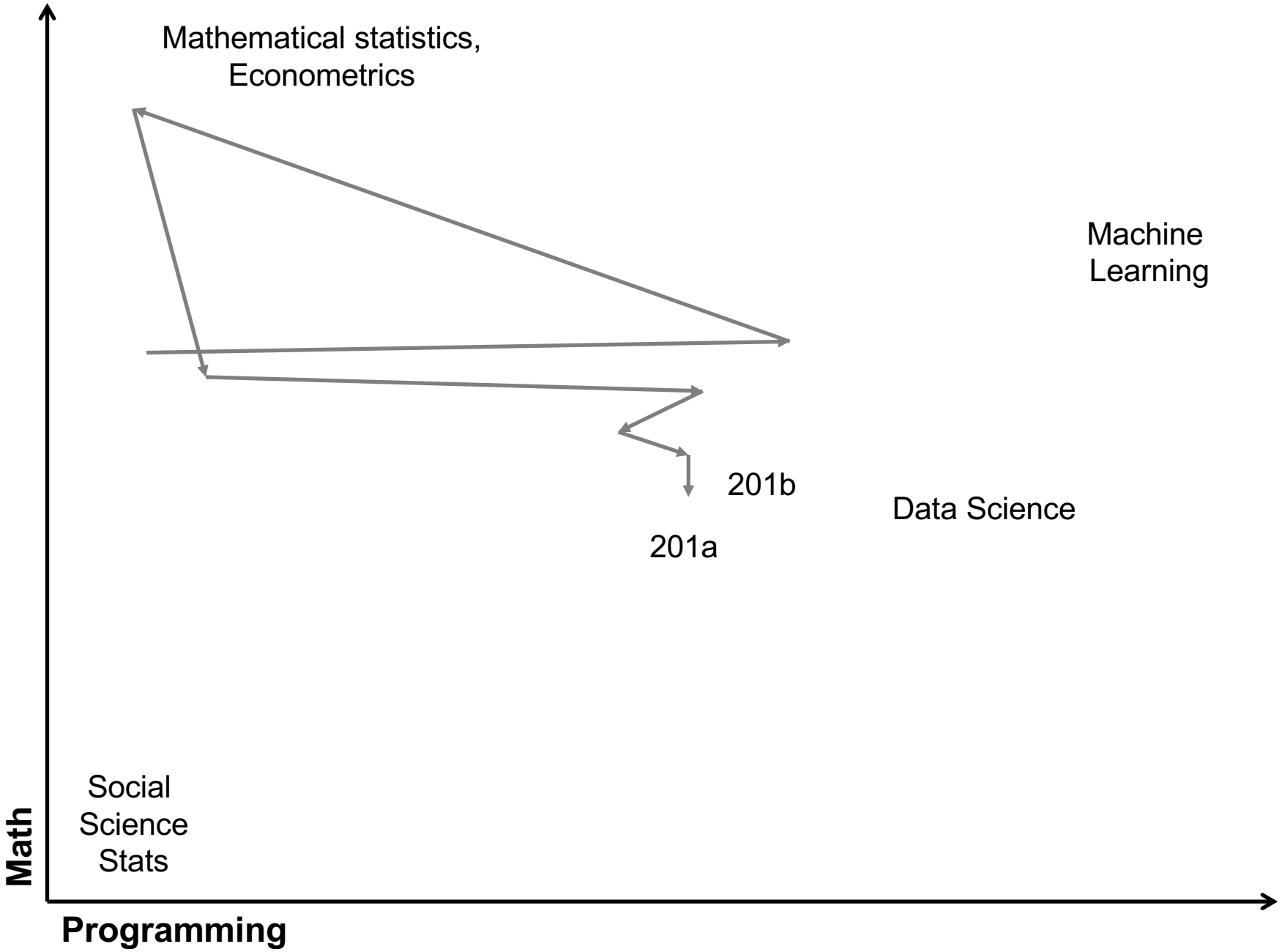
201a:

- Data
  Hygiene, cleaning, types, visualizing, describing
- Foundations
  Probability, sampling, null hypotheses, etc.
- General linear model
  Correlation / Regression, Multiple regression, ANOVA, ANCOVA
- Pointers to GLM extensions
  Linearizing transforms, covarying errors

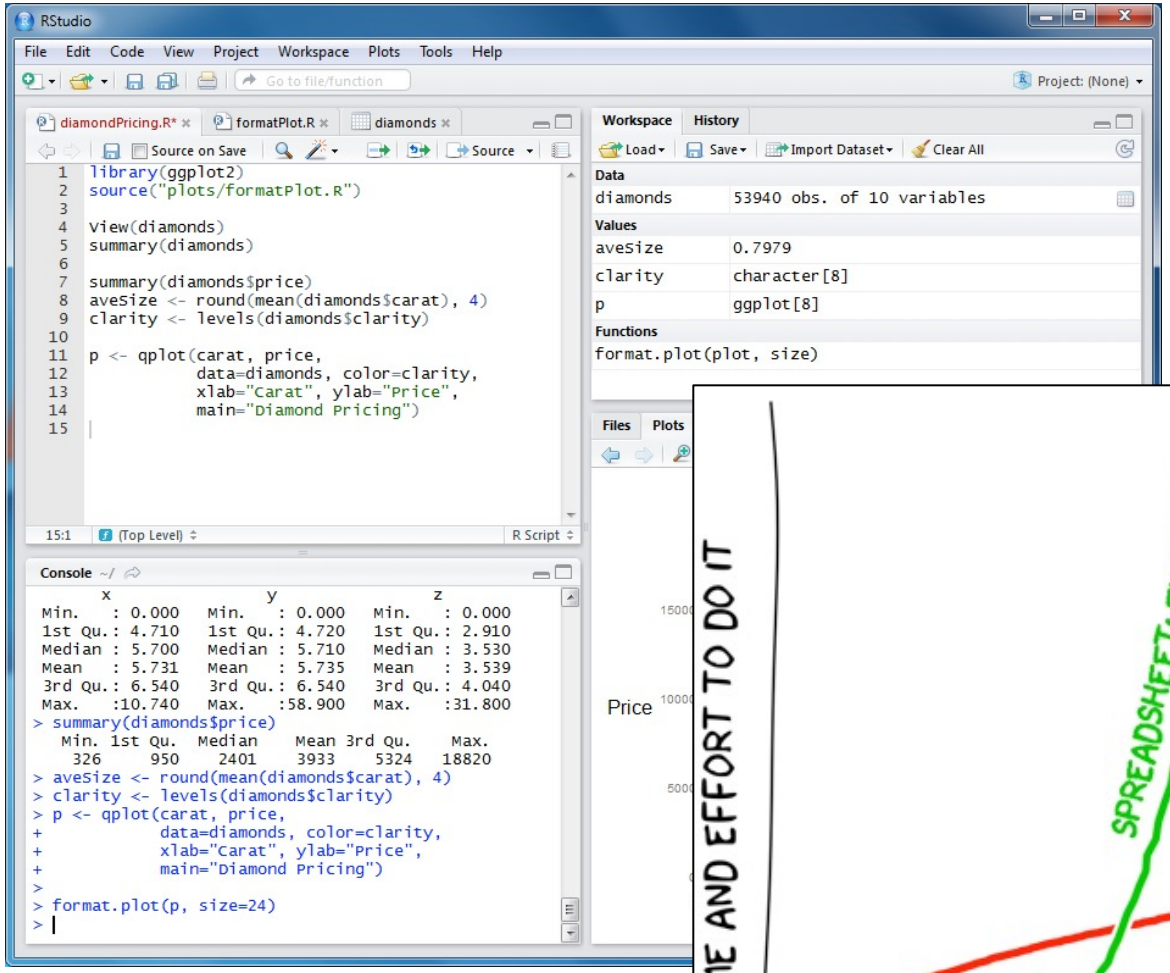
201b:

- Linear mixed effects
- Generalized Linear model
  e.g., Logistic regression
- Likelihood and optimization
- Resampling method
- Bayesian methods

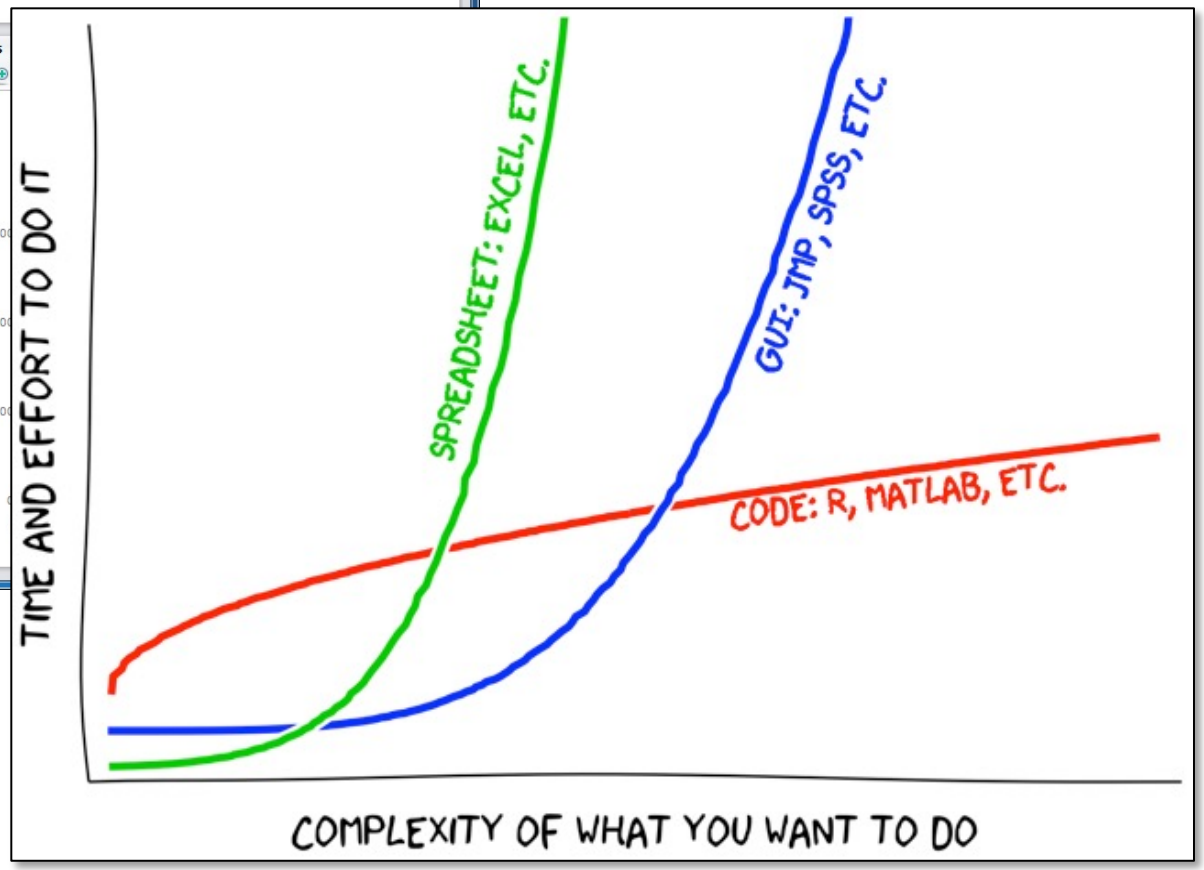
The landscape of introductory courses dealing with data.



R, Rstudio – this will be hard for some.



Why?



Syllabus

Psych 193 (10 undergrads)

TuTh 2-3:20, We 5-7

Grading:

- 75% Homework
- 25% Project

Psych 201a:

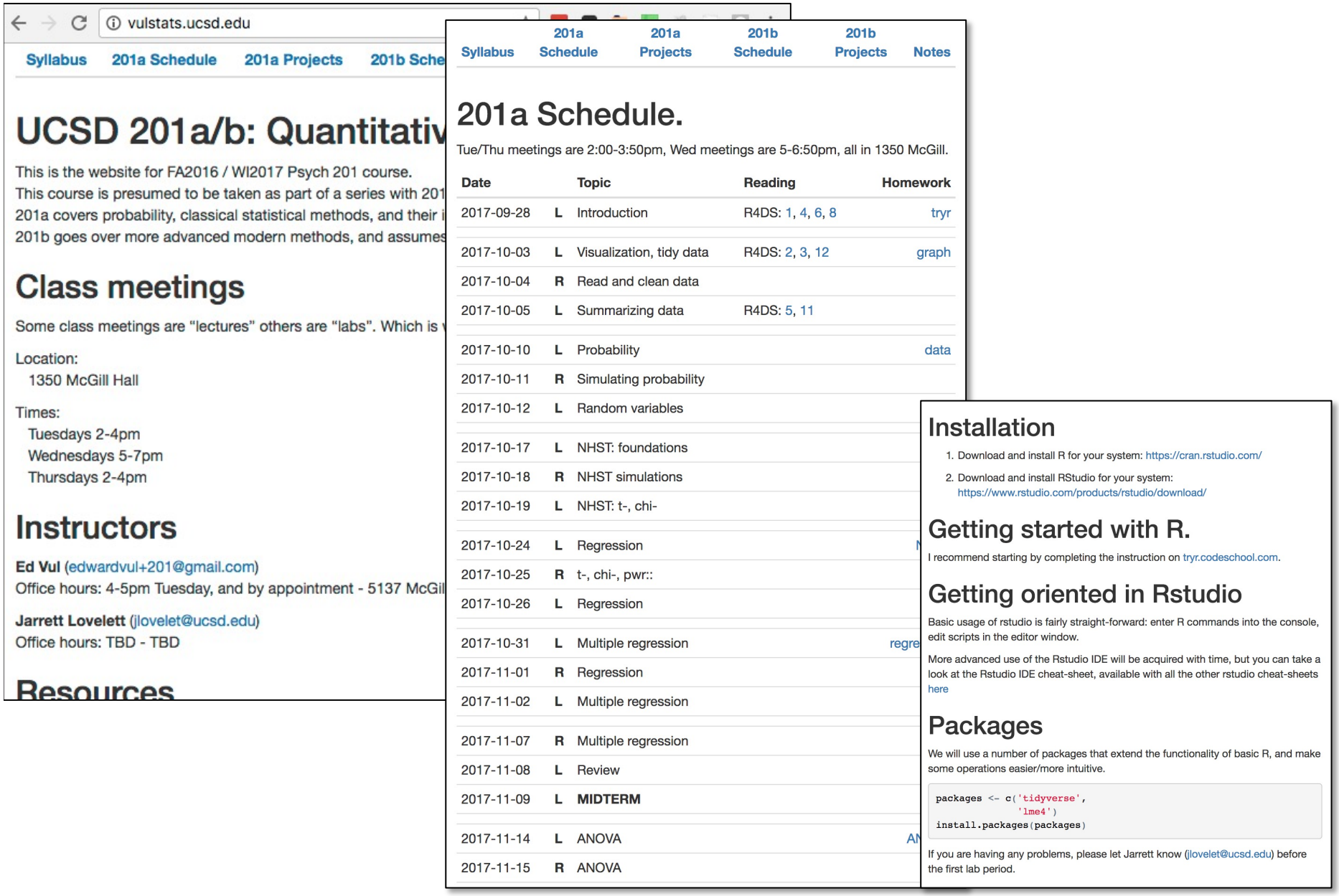
TuTh 2-3:50, We 5-7

Grading:

- 25% Homework
- 25% Project
- 25% Midterm
- 25% Final

Projects: Check website for details (will update soon)
Homework: vulstats.ucsd.edu/hw/ (will be running soon)
Exams: collaboration not tolerated

Website: vulstats.ucsd.edu



vulstats.ucsd.edu

Syllabus | 201a Schedule | 201a Projects | 201b Sche

UCSD 201a/b: Quantitativ

This is the website for FA2016 / WI2017 Psych 201 course.
This course is presumed to be taken as part of a series with 201
201a covers probability, classical statistical methods, and their i
201b goes over more advanced modern methods, and assumes

Class meetings

Some class meetings are "lectures" others are "labs". Which is

Location:
1350 McGill Hall

Times:
Tuesdays 2-4pm
Wednesdays 5-7pm
Thursdays 2-4pm

Instructors

Ed Vul (edwardvul+201@gmail.com)
Office hours: 4-5pm Tuesday, and by appointment - 5137 McGil

Jarrett Lovelett (jlovelet@ucsd.edu)
Office hours: TBD - TBD

Resources

Syllabus | 201a Schedule | 201a Projects | 201b Schedule | 201b Projects | Notes

201a Schedule.

Tue/Thu meetings are 2:00-3:50pm, Wed meetings are 5-6:50pm, all in 1350 McGill.

Date		Topic	Reading	Homework
2017-09-28	L	Introduction	R4DS: 1, 4, 6, 8	tryr
2017-10-03	L	Visualization, tidy data	R4DS: 2, 3, 12	graph
2017-10-04	R	Read and clean data		
2017-10-05	L	Summarizing data	R4DS: 5, 11	
2017-10-10	L	Probability		data
2017-10-11	R	Simulating probability		
2017-10-12	L	Random variables		
2017-10-17	L	NHST: foundations		
2017-10-18	R	NHST simulations		
2017-10-19	L	NHST: t-, chi-		
2017-10-24	L	Regression		
2017-10-25	R	t-, chi-, pwr::		
2017-10-26	L	Regression		
2017-10-31	L	Multiple regression		regre
2017-11-01	R	Regression		
2017-11-02	L	Multiple regression		
2017-11-07	R	Multiple regression		
2017-11-08	L	Review		
2017-11-09	L	**MIDTERM**		
2017-11-14	L	ANOVA		AN
2017-11-15	R	ANOVA		

Installation

1. Download and install R for your system: https://cran.rstudio.com/
2. Download and install RStudio for your system: https://www.rstudio.com/products/rstudio/download/

Getting started with R.

I recommend starting by completing the instruction on tryr.codeschool.com.

Getting oriented in Rstudio

Basic usage of rstudio is fairly straight-forward: enter R commands into the console, edit scripts in the editor window.

More advanced use of the Rstudio IDE will be acquired with time, but you can take a look at the Rstudio IDE cheat-sheet, available with all the other rstudio cheat-sheets here

Packages

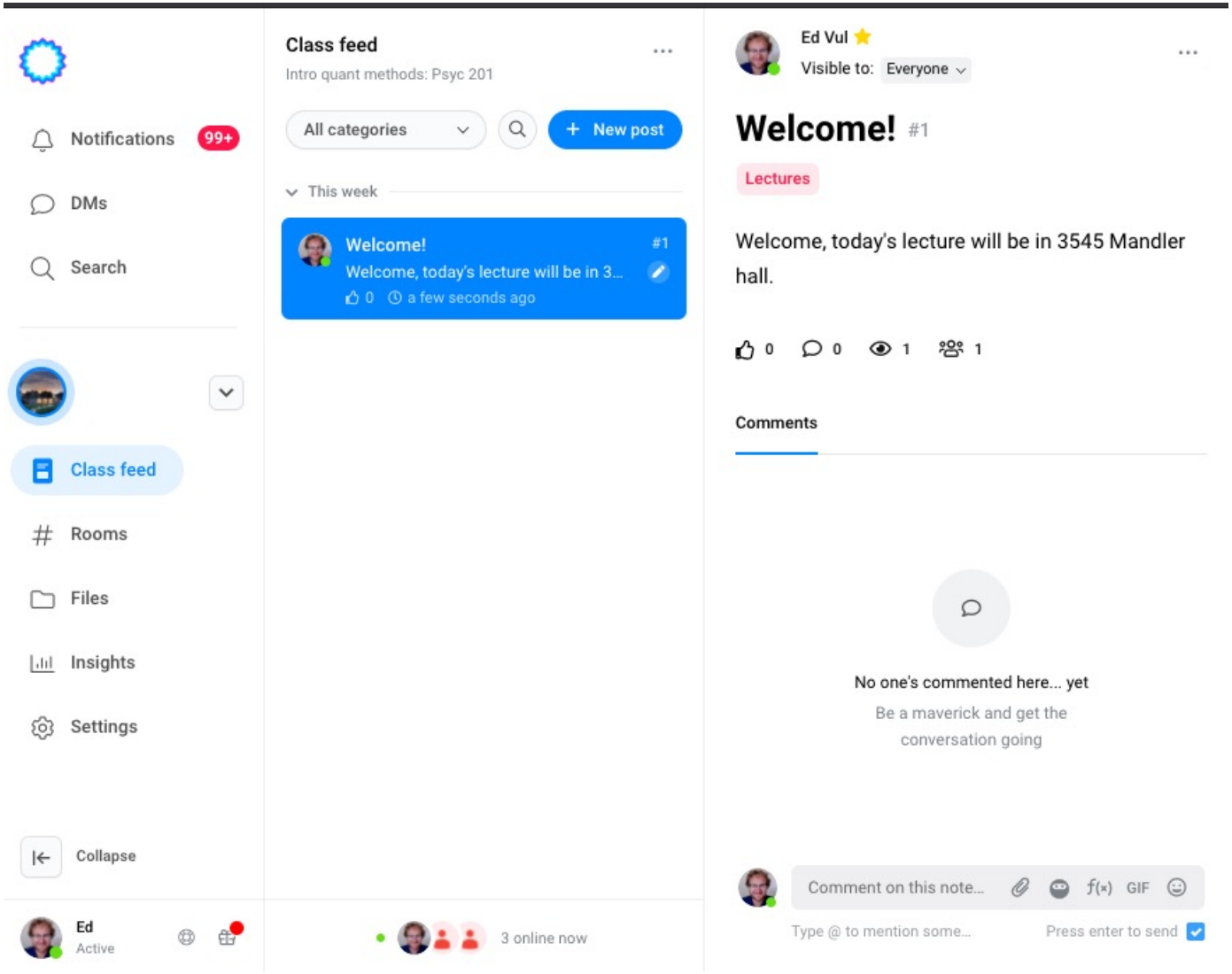
We will use a number of packages that extend the functionality of basic R, and make some operations easier/more intuitive.

```
packages <- c('tidyverse',
              'lme4')
install.packages(packages)
```

If you are having any problems, please let Jarrett know (jlovelet@ucsd.edu) before the first lab period.

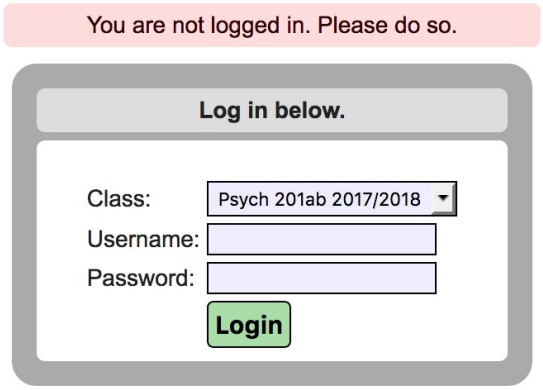
Campuswire

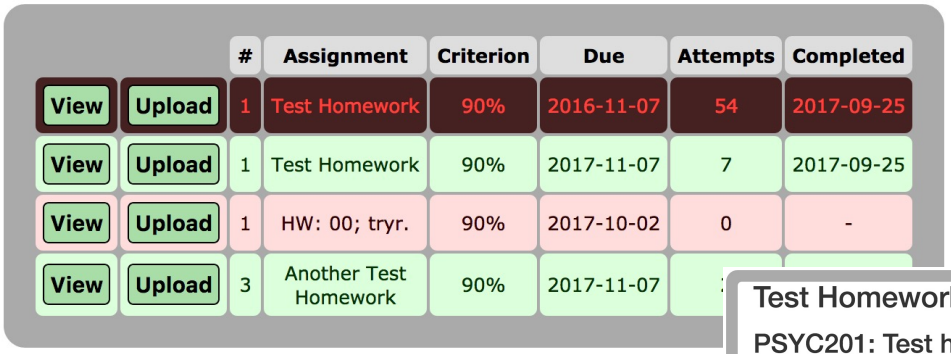
https://campuswire.com/p/G8F0D1572 (code: 2647)



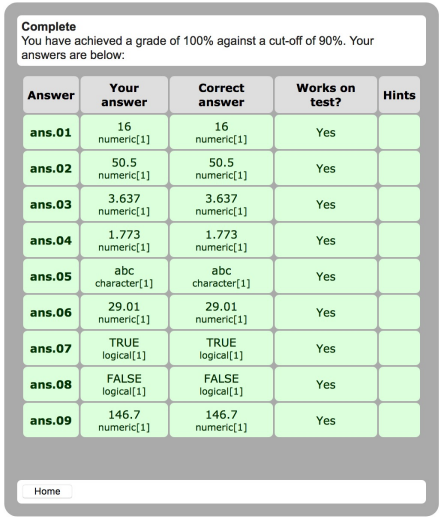
Homework: vulstats.ucsd.edu/hw/

Login: UCSD username (e.g. for evul@ucsd.edu -- evul)
Password: Your student ID.
More once its running next week





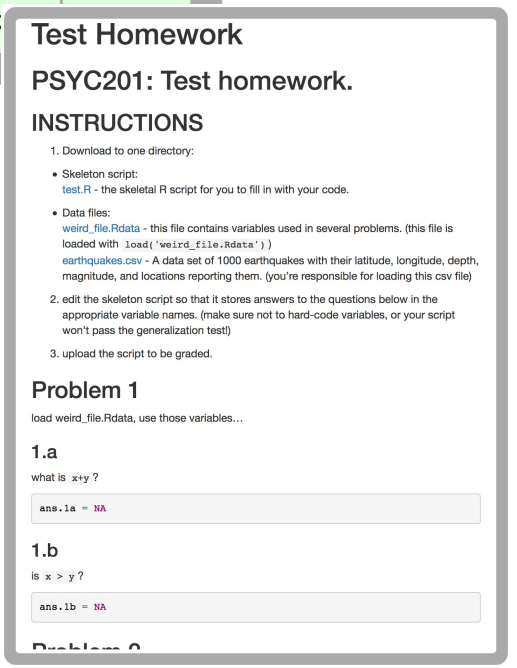
		#	Assignment	Criterion	Due	Attempts	Completed
View	Upload	1	Test Homework	90%	2016-11-07	54	2017-09-25
View	Upload	1	Test Homework	90%	2017-11-07	7	2017-09-25
View	Upload	1	HW: 00; tryr.	90%	2017-10-02	0	-
View	Upload	3	Another Test Homework	90%	2017-11-07		



Complete
You have achieved a grade of 100% against a cut-off of 90%. Your answers are below:

Answer	Your answer	Correct answer	Works on test?	Hints
ans.01	16 numeric[1]	16 numeric[1]	Yes	
ans.02	50.5 numeric[1]	50.5 numeric[1]	Yes	
ans.03	3.637 numeric[1]	3.637 numeric[1]	Yes	
ans.04	1.773 numeric[1]	1.773 numeric[1]	Yes	
ans.05	abc character[1]	abc character[1]	Yes	
ans.06	29.01 numeric[1]	29.01 numeric[1]	Yes	
ans.07	TRUE logical[1]	TRUE logical[1]	Yes	
ans.08	FALSE logical[1]	FALSE logical[1]	Yes	
ans.09	146.7 numeric[1]	146.7 numeric[1]	Yes	

Home



Test Homework

PSYC201: Test homework.

INSTRUCTIONS

1. Download to one directory:
   - Skeleton script:
     test.R - the skeletal R script for you to fill in with your code.
   - Data files:
     weird_file.Rdata - this file contains variables used in several problems. (this file is loaded with `load('weird_file.Rdata')`)
     earthquakes.csv - A data set of 1000 earthquakes with their latitude, longitude, depth, magnitude, and locations reporting them. (you're responsible for loading this csv file)
2. edit the skeleton script so that it stores answers to the questions below in the appropriate variable names. (make sure not to hard-code variables, or your script won't pass the generalization test!)
3. upload the script to be graded.

Problem 1

load weird_file.Rdata, use those variables...

1.a

what is `x+y`?

```
ans.1a = NA
```

1.b

is `x > y`?

```
ans.1b = NA
```

Projects

201ab projects

The goal is to analyze a large, rich dataset to answer an interesting behavioral/social/neural question, with the final product being a potentially publishable paper.

This project is divided into two phases to be implemented in 201a and 201b.

In 201a your goal is to identify a conjunction of an interesting question and a data source that might answer it. You will need to understand the data, clean it, make graphs of the data that might answer the question, and do simple analyses to get your bearings.

In 201b you will do the more complete analyses, likely using more advanced methods that we will cover in 201b, and turn the initial report from 201a into something that could be submitted for publication.

Examples of this sort of thing:

Examples: skill learning in online games, sequential dependence in yelp reviews, stereotype threat in chess play, income mobility over time, scaling laws in cities, crowd within in real estimation, personality in blog posts, neurosynth brain mapping example.

You will notice that particularly successful examples usually have a combination of a few things:
(1) a coherent research question, with a good justification for why the naturalistic data maps onto theoretical constructs of interest.
(2) a novel dataset, which might mean data that had not preciously been available, or a dataset that was created by cleverly combining/co-registering previously separate datasets.
(3) and (sometimes or) a fairly sophisticated analysis that adequately grapples with the complicated structure of the data.

The full project will span both 201a and 201b (previously I had it only in 201b, and that was not enough time)

201a

How to ask debugging questions

- Isolate the problem.
  - **Identify the smallest unit of code that reproduces the problem.**
  - Independent of other code, particular variables in memory, the larger dataset, loaded packages, etc.
- Steps to take
  - Make sure all variables in play are as expected.
    - Check types!
  - Google function name and key words from error message (omitting terms specific to your circumstance, like local variable names).
  - Spend 15-30 min reading/trying solutions.
  - Ask. Include smallest unit of code/data that produces the problem, and briefly mention what you found by way of failed answers.

How do we go beyond the data?

"Inference"/"Induction"/"Generalization"/"Prediction"

- Combine data with assumptions
  - Sample is representative (resampling)
  - A model
  - (may not be transparent)

- Inferences are always uncertain
  - (uncertainty not always transparent)

Fields dealing with data

- Differences in regimes, goals:
  - small, low-d samples from experiments with binary causal questions (classical stats)
  - largish, mid-d samples from surveys or the wild with meaningful model parameters (modern regression modeling, econometrics, etc.)
  - large, high-d naturalistic datasets, with an emphasis on prediction and discovery of structure (ML, data science)
- Differences in emphasis: mathematical theory vs algorithmic implementation.
- Data-structure fields and subareas: geostatistics, timeseries, networks, raw signals, images, surveys, text, censored data, etc.
- Domain-specific bundles: econometrics, psychometrics, biostatistics, etc.

Data scale → methods

	Intervention	Sample size	Dimensions	Structure
Psychophysics	Experiment	10^0 ($x10^3$)	10^0	Flat, subject
Personality	Survey	10^2	10^1	Flat
Behavioral	Experiment	10^1 ($x10^1$)	10^0	Flat, subjects, items
Political, Sociology	Survey	10^4	10^1	Demography, Geography, Networks
Financial, Macroeconomic	Observation	10^3	10^2	Timeseries
Neuroimaging	Experiment	10^1 ($x10^2$)	10^4	(3D) Spatial, Timeseries
Text (corpus linguistics, NLP)	Observation	$10^3 \cdots 10^{11}$	10^1 or 10^5	Markov? Bags of words? CFG?
Images	Observation	$10^3 \cdots 10^8$	10^{5+}	Optics, the 3D world.
Sports	Observation	$10^2 \cdots 10^5$	$10^{1\text{-}2}$	Relational, game mechanic
Web user data	Observation	$10^3 \cdots 10^{12}$	$10^1 \cdots 10^?$	All sorts

data structure → methods

- Flat, tabular data
- Hierarchical/relational data

- Timeseries/sequential data
- Censored / survival data
- Raw signal varying over time, space
- Network data
- Image data
- Text
- Spatial / geographic data

Goals → methods

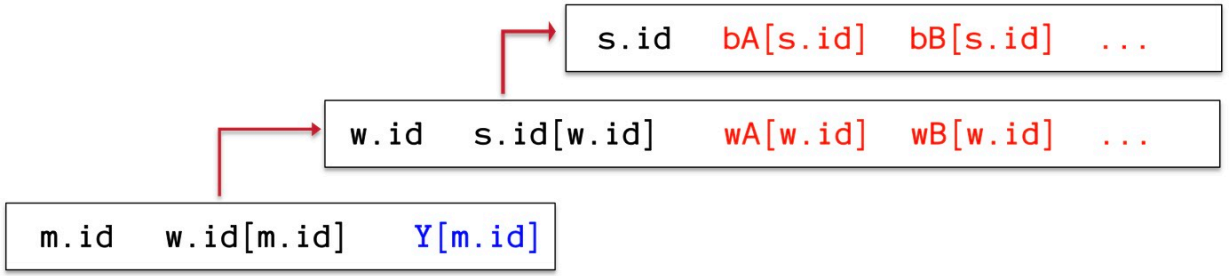
- Describe/summarize data
  - Literal “statistics”
  - Visualization
- Predict new data
  - classification / regression
- Characterize process / population
  - Estimate model parameters
  - Choose among models
  - Clustering, dimensionality reduction, factor analysis, etc.
  - Separate signal from “noise”

We usually have multiple goals.

Data structures we cover

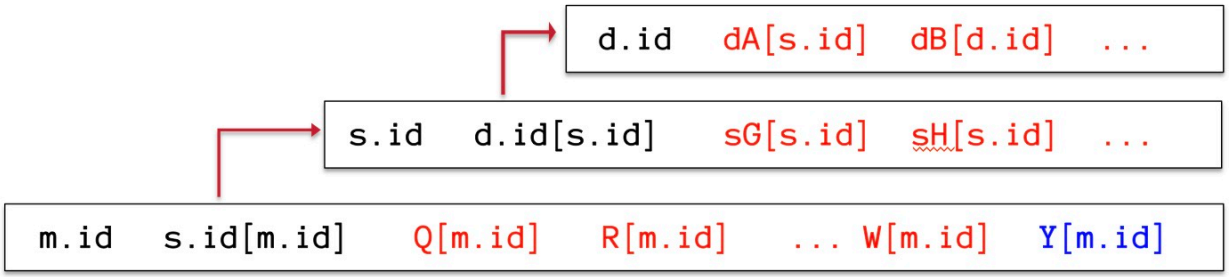
Normal LM/GLM structure:
each unit is uniquely associated with a measurement.

Repeated measures structure:
Explanatory variables at only two levels. (e.g., "between subject" and "within subject").

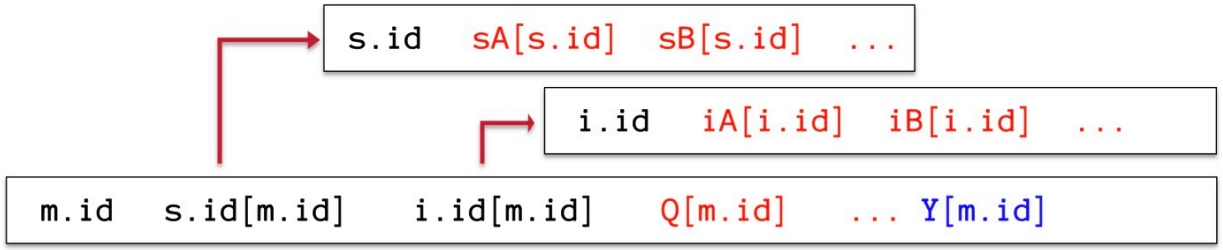


In 201b

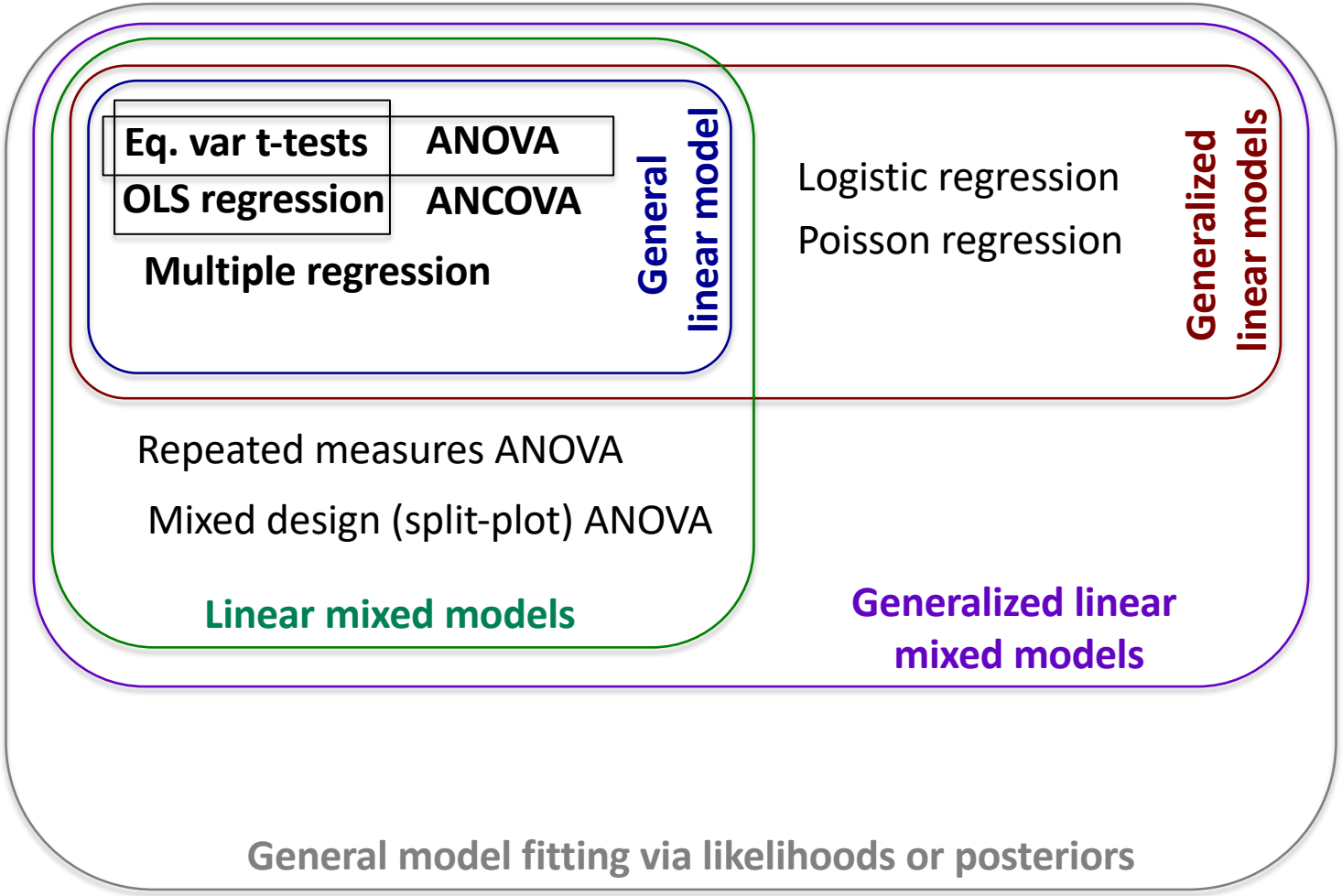
More levels
Explanatory variables at more than two levels. (e.g., classes in schools in districts)



Crossed random effects
(e.g., subjects with explanatory variables crossed with items with explanatory variables)



Relationship between models



In general, special cases of broader model classes are usually favored despite being less flexible because they are simpler and allow for easier estimation and inference.

Practical suggestions

- Making your life easier (in the long run):
  - Adhere to sensible file/directory structure
  - Name files / folders coherently
  - Save data in universal, machine-readable format (text)
  - Record everything.
  - Automate recording.
  - Make a text file describing where data came from.
    - A codebook if necessary
  - Conventional, stranger-readable coding.
  - Tidy data: one row per measurement, no empty cells, etc.
  - Never alter the raw data.
  - Write standalone scripts for data cleaning, analysis
  - Version control (I like git)
  - Consider writing papers in R Markdown (papaja?)

Practical suggestions

- Avoid pain in R scripts
  - Do:
    Use data frames (not matrices, isolated vectors, etc)
    Name columns conveniently, factor levels clearly
    Index columns by name, not number
    Subset rows by logical filters, not numbers
    Pass named (not place) arguments to functions
    Make your code state-independent and self-sufficient
  - Do not:
    attach()
    save subsets into new variables
    use rownames (store as explicit column if you want)
    Use “magic numbers”
    hard code stuff in the middle of the script.

Practical suggestions

- Experiment design suggestions:
  - KISS
  - Aim for within-{subject,item} design
  - Build in checks for:
    - attention
    - reliability
    - manipulation
    - confounds
    - strategy
    - blindness.
  - Debug design by pre-planning analysis/paper.

Practical suggestions

- Debug design by pre-planning analysis/paper.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." - Fisher (1938)

- What are you trying to do? (concrete Q, candidate As)
- What will decisive figure look like? (diagnostic?)
- How will you go from data to figure?
- How big is the expected effect? Noise? (Power, bitrate)

Debug further by playing devil's advocate (reviewer)

- What assumptions link interpretation to measures, manipulations?
- If your interpretation is wrong, what explains your awesome figure?
- What are plausible, alternate causal routes between manipulations and measurements?

Practical suggestions

- Be replicable:
  - Aim for precise, quantitative estimates; not p<0.05
    Clean, quantitative measurements; large samples, within-subject designs
  - Be precise in your “theories”.
  - “you must not fool yourself”
    (Feynman on cargo cult science)
    - Be responsible for answer, not adherence to rulebook
    - Pre-register (with yourself?) to spot data-driven analysis
    - Everything open by default.
    - Look at your data and variability (visualization)
    - How much would you bet on replication success?
  - Replicate.
    - Didn’t predict the effect? Replicate.
    - Following up on someone’s one-off result? Replicate.