Statistics 133:
Concepts in Computing with Data

Instructor: Ingileif Hallgrímsdóttir (Inga)
GSIs: Simon Walter, Yuan He and Yu Wang

Contact Info

Ingileif Hallgrímsdóttir (Inga)

Email : ingileif@berkeley.edu

Office : 416 Evans Hall

Office Hours :

Tue 3:40-4:30 pm & Thu 11:10-12pm

First Section!

- It is important that you attend the first section on Friday. In it we will install R, RStudio, git and set up your github account so that we can share files.
- As time permits we will start working through the R bootcamp:

http://statistics.berkeley.edu/computing/r-bootcamp

Homework

- Homework will be due Friday by 5 pm and posted Thursday the week before. The homework will be “pushed” to your github accounts, and you will turn them in through github.

Midterms

- Will be completed in a computer lab, dates to be decided by the end of January.

Grading

Grades will be based on a final exam (30%), two midterms (10% each), weekly/bi-weekly programming assignments (50%). Homeworks will be due weekly, but there will be 3 larger assignment that you will get two weeks to complete. For the regular homework you will submit your code through github, for the assignments you will need to write a short report in addition to your code.

Theme

- 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

Theme

- Statistical Thinking in the context of computing with data
- DATA Technologies – Statisticians' work includes interfacing and working closely with the original data and those who own it

What Are Data?

Numbers

- Example: Traffic on I-80

flow-occ-table.txt

```
Occ1,Flow1,Occ2,Flow2,Occ3,Flow3
0.01,14,0.0186,27,0.0137,17
0.0133,18,0.025,39,0.0187,25
0.0088,12,0.018,30,0.0095,11
0.0115,16,0.0203,33,0.0217,19
0.0069,8,0.0178,25,0.0123,13
0.0077,11,0.0151,24,0.0092,13
0.0049,7,0.0153,22,0.0192,19
0.007,10,0.0194,33,0.0156,17
0.0082,12,0.0146,26,0.0166,13
0.0074,11,0.0207,30,0.018,14
0.0071,10,0.0135,22,0.0074,11
0.0069,10,0.012,17,0.0147,12
0.0011,2,0.0078,13,0.0118,10
0.0038,5,0.0116,18,0.0202,11
0.0063,8,0.0115,15,0.0214,17
0.0034,5,0.0137,20,0.0153,13
0.0043,5,0.0094,16,0.019,18
0.0038,5,0.0111,18,0.0131,13
0.0017,2,0.0121,18,0.0156,14
0.0018,3,0.0102,17,0.0269,18
0.0058,8,0.0131,19,0.0119,11
0.0016,2,0.0082,11,0.0095,12
0.003,3,0.0075,12,0.0174,18
0.0024,4,0.0094,17,0.0069,8
0.0014,2,0.017,17,0.0232,13
0.004,5,0.0079,11,0.0117,12
0,0,0.0072,12,0.0142,10
0.0016,2,0.011,15,0.0123,10
0.0013,2,0.0027,5,0.0077,8
```

Dates, Times, Locations

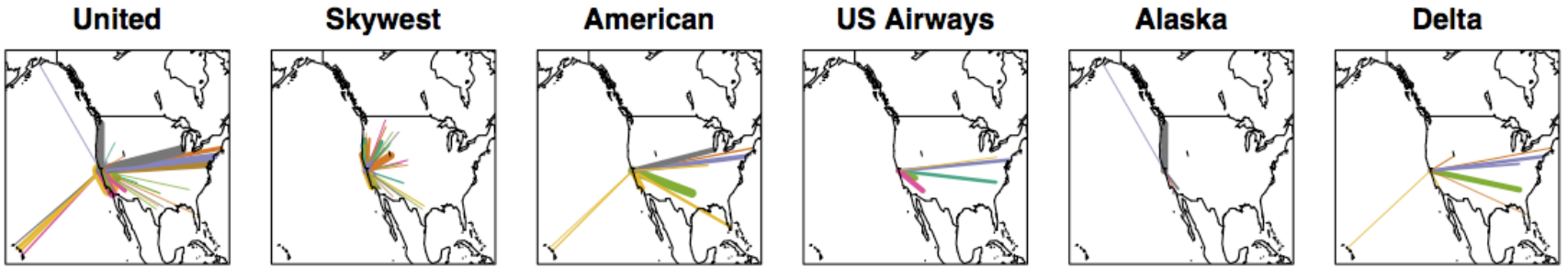
- Example: Flight information

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime
1	2007	1	2	2	1051	1025	1401	1340
2	2007	1	2	2	1950	1935	2255	2245
3	2007	1	2	2	742	735	1047	1050
4	2007	1	2	2	1122	1055	1735	1705
5	2007	1	2	2	1142	1105	1400	1335
6	2007	1	2	2	2024	2005	2242	2235

	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay
1	WN	1719	N302SW	130	135	121	21
2	WN	1896	N464	125	130	112	10
3	WN	2296	N462	125	135	116	-3
4	WN	2459	N405	253	250	239	30
5	WN	622	N632SW	78	90	69	25
6	WN	1752	N455	78	90	70	7

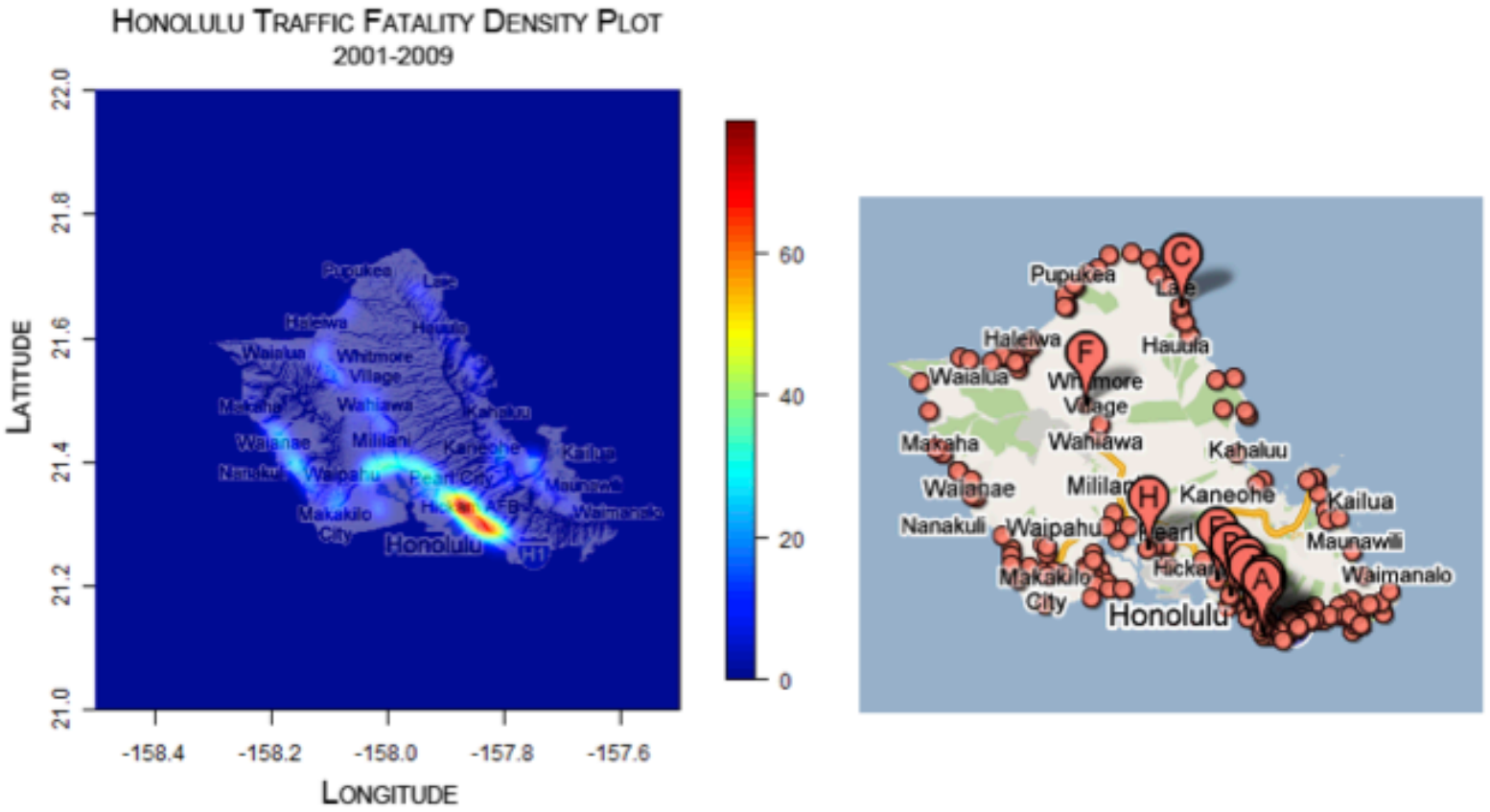
	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	CancellationCode	Diverted
1	26	OAK	ABQ	889	3	6	0		0
2	15	OAK	ABQ	889	7	6	0		0
3	7	OAK	ABQ	889	3	6	0		0
4	27	OAK	BNA	1959	5	9	0		0
5	37	OAK	BOI	511	3	6	0		0
6	19	OAK	BOI	511	3	5	0		0

	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay
1	21	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	21	0	3	0	6
5	4	0	0	0	21
6	0	0	0	0	0



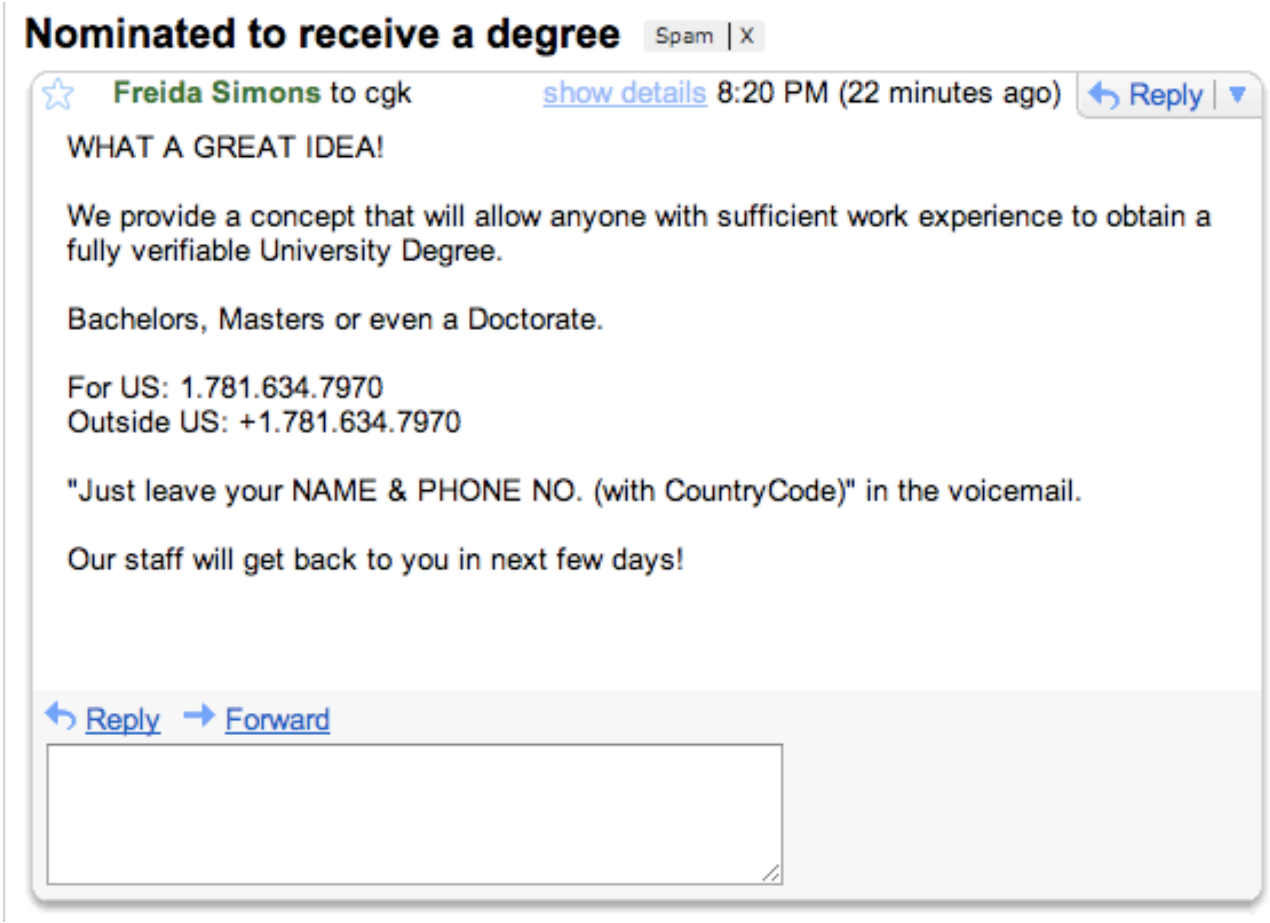
Dates, Times, Locations

- Example: Traffic fatalities (group project)



Text

- Example: SPAM or HAM?



Text

- Example: Online dating

•Meta-data

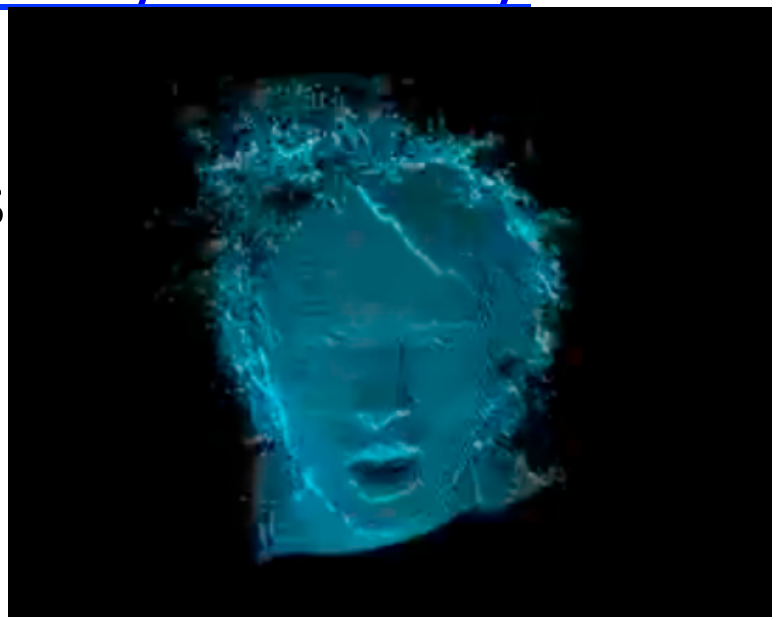
```
<?xml version="1.0" standalone="yes"?>
<kml  creator="thd Thu Dec  6 16:24:55 2007" xmlns="http://earth.google.com/kml/2.0">
<Document>
  <name> at </name>
 <Folder>
 <name>Waypoints</name>
  <Placemark>
    <name>Black Gap Shelter</name>
    <Point><coordinates>-84.19880,34.61756,0.0</coordinates></Point>
    <description><![CDATA[Waypoint: BlackGap <br> Additional <a href="http://www.cs.utk.edu/~dunigan/at/m.php?wpt=BlackGap">information</a> ]]></description>
    <styleUrl>#waypoint</styleUrl>
  </Placemark>
  <Placemark>
    <name>Springer Mountain Shelter</name>
    <Point><coordinates>-84.19306,34.62915,0.0</coordinates></Point>
    <description><![CDATA[Waypoint: Springer <br> Additional <a href="http://www.cs.utk.edu/~dunigan/at/m.php?wpt=Springer">information</a> ]]></description>
    <styleUrl>#waypoint</styleUrl>
  </Placemark>
```

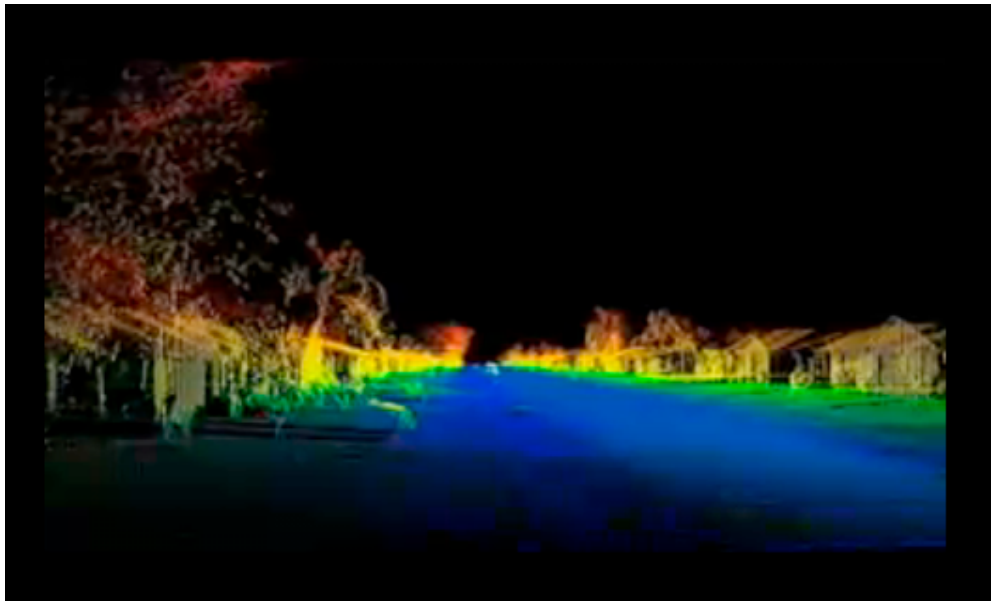
Images, video, or audio

Example: Radiohead “House of Cards” video

http://code.google.com/creative/radiohead/

Lidar and GeoVideo used to create 3-dimensional images without lights or cameras.





"I liked the idea of making a video of human beings and real life and time without using any cameras, just lasers, so there are just mathematical points – and how strangely emotional it ended up being.” - Yorke

Hans Rosling

- Ted talk:

http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=en

Statistical Thinking and the Data Analysis Cycle

- Data ACQUISITION – Input/output, regular expressions
- Data CLEANING – verification, manipulation
- Data ORGANIZATION – data frames, data bases, XML
- Data ANALYSIS – fit and assess statistical models, conduct exploratory data analysis
- Data SIMULATED – simulation studies to understand behavior of data
- Data REPORTING – report findings

Statistical Concepts

- Basic numeracy
  - Variability, Patterns, comparisons
- Graphics
  - Elements and principles of graphing
- Computationally intensive methods, e.g.,
  - Classification and Regression trees, multi-dimensional scaling, nearest neighbor
- Simulation tools
  - Monte Carlo, bootstrap, cross-validation

Computing Concepts

- Programming concepts
  - Control flow, trees, recursion
- Regular expressions and text manipulation
- Relational databases
- Random number generation
- Representation of information in the computer
- Event handling and GUI development

Software

- R – statistical software
- R Studio
- Unix – shell commands
- Git / GitHub – version control, repository
- Possible Additional Topics

Automated Grading

- We will use Git to manage homework and projects. You will have to get a Git account through which you can “push” your completed projects to the instructors.

website : github.com

Please sign up for an educational/student account as soon as possible.

Academic Integrity

Code of Student Conduct is available at

http://students.berkeley.edu/uga/**conduct**.pdf

- Free to discuss course matters with instructor, GSIs, USIs, and fellow students
- **DO NOT SHARE CODE**
- Make significant contribution to your group's work
- If you are uncertain as to whether something may be a violation of the code, ask the instructor

Academic Integrity

Writing a program is like writing a paper – your code should be your original work.

A violation will result in at least one of the following:

- 0 on the assignment
- F for the course grade
- Report to the Office of Student Conduct