

LIS 490:

Introduction to Data Science

Professor: Victoria Stodden

Theme

- Use the computer expressively to conduct statistical analysis of data
- Use existing software rather than build routines from the ground up.
- Focus on methods and aspects of computing to conduct statistical analysis, less on the computational aspects of statistical methods
- Discussion: cybernetics and data science (Week1 Readings and Alan Kay video)

Data Science

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

What Are Data?

Numbers

- Example: Traffic on I-80



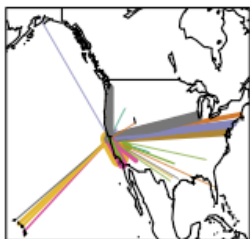
```
flow-occ-table.txt
Occ1,Flow1,Occ2,Flow2,Occ3,Flow3
0.01,14,0.0186,27,0.0137,17
0.0133,18,0.025,39,0.0187,25
0.0088,12,0.018,30,0.0095,11
0.0115,16,0.0203,33,0.0217,19
0.0069,8,0.0178,25,0.0123,13
0.0077,11,0.0151,24,0.0092,13
0.0049,7,0.0153,22,0.0192,19
0.007,10,0.0194,33,0.0156,17
0.0082,12,0.0146,26,0.0166,13
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

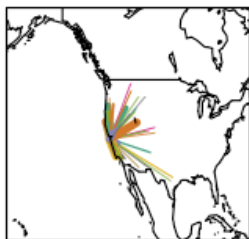
- Example: Flight information

	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	0	3	0	6					
5	4	0	0	0	0	21					
6	0	0	0	0	0	0					

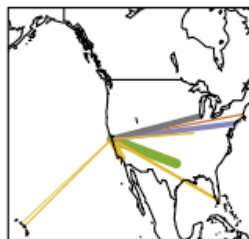
United



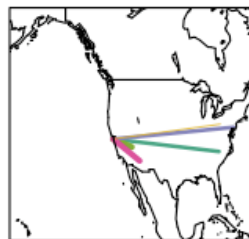
Skywest



American



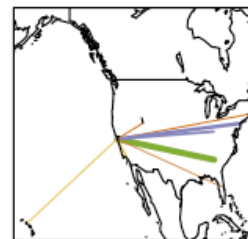
US Airways



Alaska

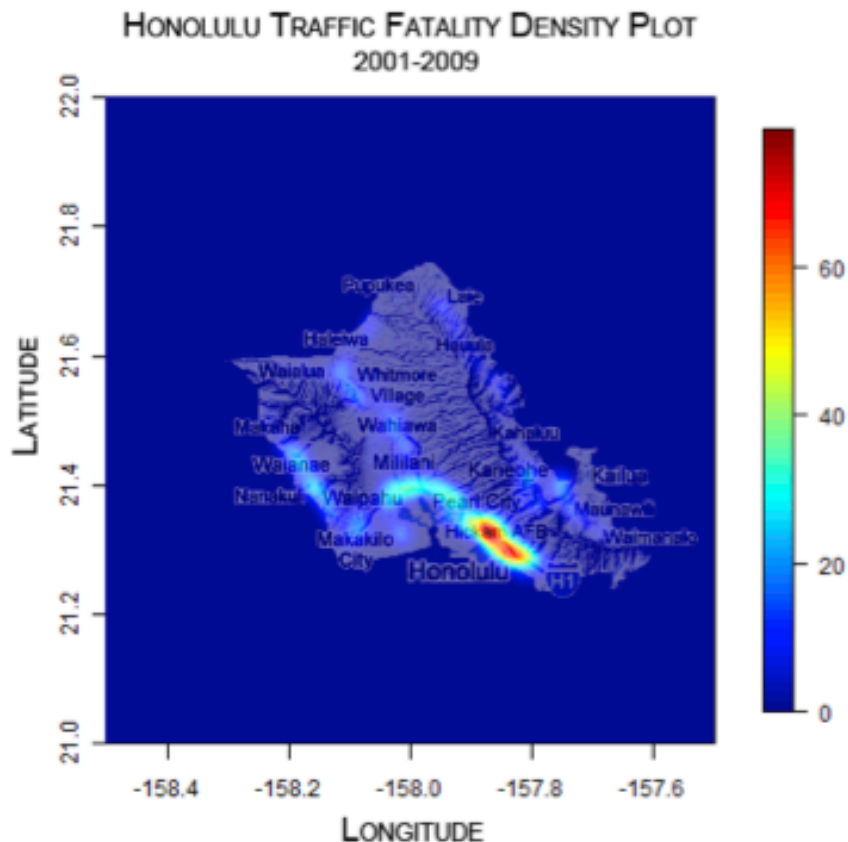


Delta



Dates, Times, Locations

- Example: Traffic fatalities



Text

- Example: SPAM or HAM?

Nominated to receive a degree

Spam | X



Freida Simons to cgk

[show details](#) 8:20 PM (22 minutes ago)

[Reply](#) | ▼

WHAT A GREAT IDEA!

We provide a concept that will allow anyone with sufficient work experience to obtain a fully verifiable University Degree.

Bachelors, Masters or even a Doctorate.

For US: 1.781.634.7970

Outside US: +1.781.634.7970

"Just leave your NAME & PHONE NO. (with CountryCode)" in the voicemail.

Our staff will get back to you in next few days!

[Reply](#) [Forward](#)

- Example: Online dating




Text

- Example: Have ingredients changed over time?

epicurious
for people who love to eat

Pasta Perfect
from bon appétit



Get 100 classic and new recipes, plus video from the BA test kitchen, slideshows, and more.

[DOWNLOAD NOW](#)

[Recipes & Menus](#) [Articles & Guides](#) [Community](#) [Video](#) [Shop](#) [go](#)

[advanced search - browse](#) [healthy eating guide](#) [my recipe box - my profile](#) [seasonal food](#)

[fire it up](#) [see grilling & barbecue](#)

[RESOURCES](#)

Recipe Browse Results











browsing by: **Dinner**

[refine by](#)

main ingredient

- almond (37)
- anchovy (12)
- anise (6)
- apple (63)
- apple juice (2)
- apricot (18)
- artichoke (32)
- arugula (48)
- asian pear (3)
- asparagus (48)
- avocado (34)
- bacon (132)
- banana (1)
- barley (5)
- basil (88)
- bass (24)
- bean (140)
- beef (271)
- beef rib (33)
- beef shank (6)
- beef tenderloin (24)
- beer (27)
- beet (33)
- bell pepper (133)
- berry (12)
- blackberry (2)

3273 recipes found for: Dinner

rating	photo	recipe name	at a
		South Indian Lentil Cakes with Raita Bon Appétit, September 2011	
		Fricassee of Chanterelles Bon Appétit, September 2011	
		Swedish Meatballs Bon Appétit, September 2011	
		Shrimp & Grits Bon Appétit, September 2011	
		Parbhotta	

[< back to search results](#)

South Indian Lentil Cakes with Raita

Bon Appétit | September 2011
Bindi Restaurant, Philadelphia, PA

[recipe](#) [reviews \(1\)](#) [photo](#) [my notes](#) [find out more](#)



yield: Makes 2 to 4 servings

At Philadelphia's Bindi, legumes and rice are soaked, ground with herbs and spices, then formed into patties and pan-fried to make these cakes.

[enlarge image](#)

ingredients

Raita:

- 1 head of garlic
- 1 tablespoon extra-virgin olive oil
- Kosher salt, freshly ground pepper
- 1/4 teaspoon cumin seeds

[subscribe to Bon Appétit](#)

user rating
100% would make it again

 user rating:
4 forks

[rate this recipe](#)

[review this recipe](#)

at a glance

main ingredients
Rice/Grain, Yogurt, Yogurt, Lentil

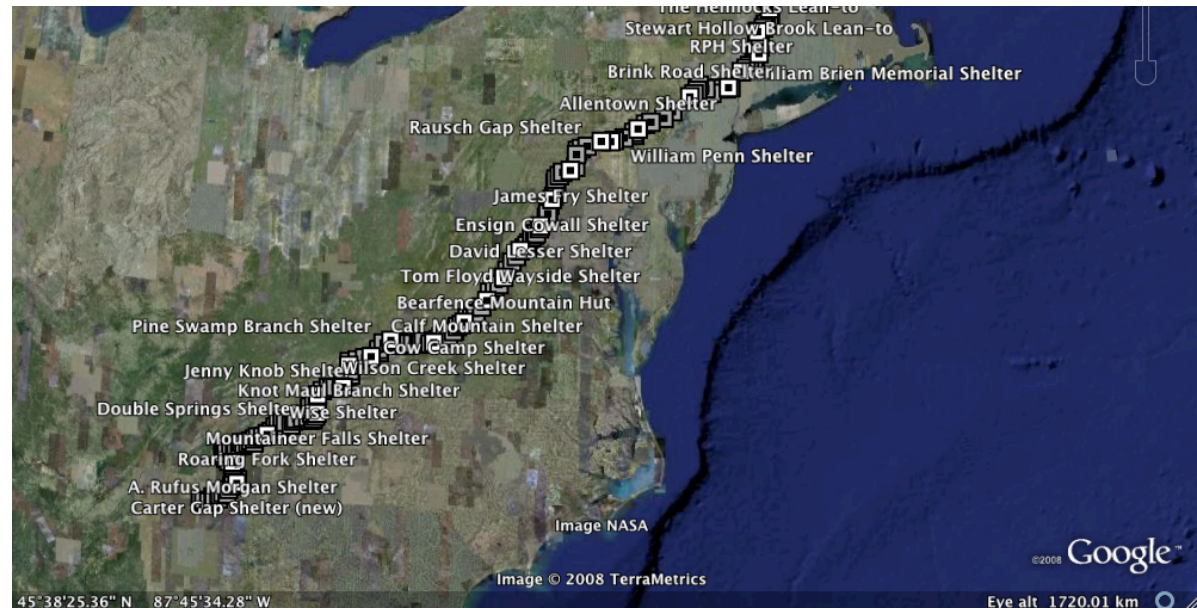
cuisine
South Asian, Indian

type
Vegetable

dietary considerations
Vegetarian

•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>
  </Folder>
</Document>
```

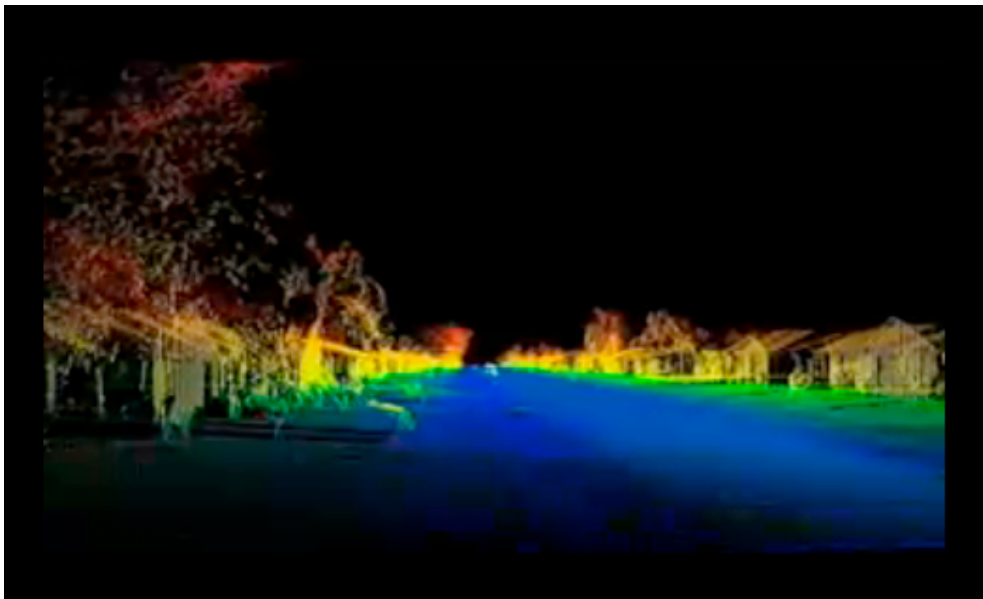


Images, video, or audio

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

Statistical Thinking and the Data Analysis Life Cycle

- Research Question / Hypothesis
- 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, update hypotheses

Statistical Concepts

- 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
 - flow control, trees, recursion
- Regular expressions and text manipulation
- Relational databases
- Random number generation
- Representation of information in the computer
- Event handling

Software

- R – statistical software
- (if time) Python – scripting for statistical applications
- Unix – shell commands
- SQL – structured query language for relational databases
- XML – Extensible Markup Language (and HTML)

Grading

Homework	40%	Weekly - due in/by class (I'll drop your lowest homework grade)
Midterm	30%	Closed notes (Oct 11)
Final	30%	Closed notes (Dec 6)

Note: Participation in class may help your grade

Academic Integrity

The Academic Integrity Policy is available at

http://admin.illinois.edu/policy/code/article1_part4_1-401.html

Free to discuss course matters with instructor, TA, and fellow students

- **DO NOT SHARE CODE**
- 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 Judicial Affairs

<http://www.youtube.com/watch?v=8nTFjVm9sTQ>