STAT 380: Week 1

Instructor: Murali Haran Professor of Statistics
TA: John Ensley, PhD Student

Outline

- Use the computer expressively to prepare, explore, and analyze data
- Work closely with original raw 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 (For that: STAT 440, Computational Statistics)
- Book:
  - Data Technologies and Computational Reasoning by D. Nolan and D. Temple Lang (pdf files will be posted weekly).
  - Supplement: *Data Science in R: A Case Studies Approach to Computational Reasoning* by Nolan and Temple Lang.

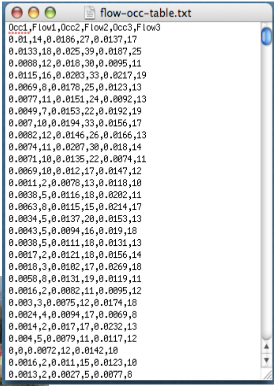
(With thanks to Professor Nolan for lecture notes)

What are data?

- Data are recorded/measured observations together with context.
- By context we mean the details of who, what, where, when, and/or how the observations were obtained, aka "metadata".

Tables of Numbers

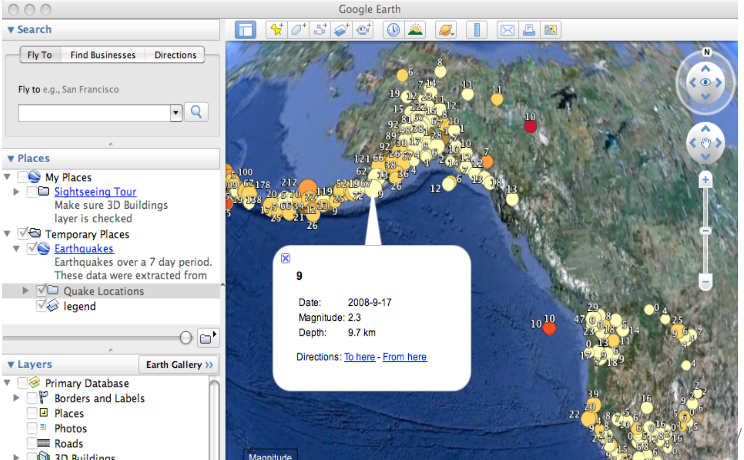
Traffic on I-80



```
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
```

Geographic Information and Time

Earthquake Location, Date, and Magnitude



Make sure 3D Buildings layer is checked

Earthquakes over a 7 day period. These data were extracted from

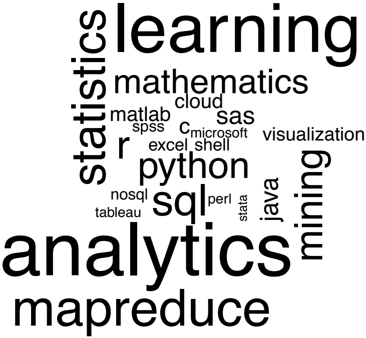
9

Date: 2008-9-17
Magnitude: 2.3
Depth: 9.7 km

Directions: To here - From here

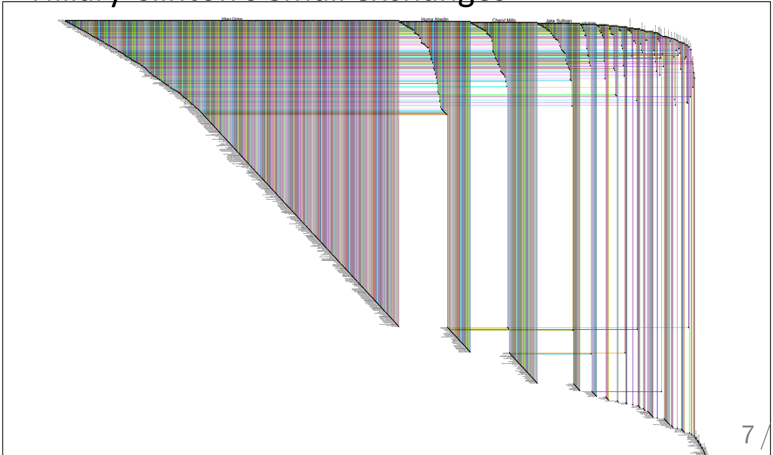
Text

Kaggle Job Postings for a Data Scientist



Graph

Hillary Clinton's email exchanges



Meta-data:

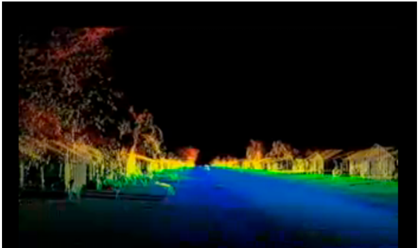
Information about Spotify playlists

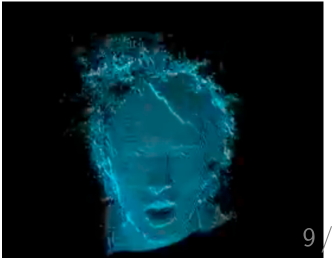
```
{
  "href" :
"https://api.spotify.com/v1/users/spotify_espa%C3%B1a/playlists/21THa8j9TaSGuXYNB
U5tsC/tracks",
  "items" : [ {
    "added_at" : "2014-08-18T20:16:08Z",
    "added_by" : {
      "external_urls" : {
        "spotify" : "http://open.spotify.com/
      },
      "href" : "https://api.spotify.com/v1/us
      "id" : "spotify_españa",
      "type" : "user",
      "uri" : "spotify:user:spotify_espa%C3%B
    },
```

Images, Video, or Audio

Radiohead House of Cards

Yorke: "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.”





What does a data scientist do?

AIG job posting on Kaggle for senior data scientist:

- Build predictive models utilizing both traditional statistical methods and modern machine learning techniques
- Extract, clean, and manipulate large datasets (structured and unstructured) for model building.
- Communicate (written and verbal) insights from quantitative analyses to technical and non-technical audiences.
- Stay current on the latest machine learning and big data trends.
- Work with business sponsors and IT teams to implement analytic solutions.
- Serve as a technical expert on one or more domains (e.g. Time Series Analysis, Text Mining, etc.)

What Skills does a Data Scientist need?

AIG job postings on Kaggle

- Expertise in at least one modeling/machine learning platform such as R, Python, or SAS.
- Knowledge of an additional general purpose programming language such as C++ or Java.
- Advanced SQL skills and experience with No SQL technologies.
- Built several predictive models that have been put into live production.
- Obsess over sample bias, over-fitting, variable selection, missing values, etc.
- Understand the need to balance predictive power, interpretability, and ease of implementation

Data analysis cycle

- Data ACQUISITION Input/output, regular expressions
- Data CLEANING verification, manipulation
- Data ORGANIZATION data frames, data bases, XML
- Data EXPLORATION search for interesting patterns
- Data VISUALIZATION create statistical graphs
- Data ANALYSIS fit and assess statistical models
- Data SIMULATION studies of random behavior
- Data REPORTING report findings from analysis

Statistical concepts

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

Computing concepts

- Programming concepts Control flow trees functions
- Regular expressions and text manipulation
- Relational databases
- Random number generation
- Representation of information in the computer

Software

- R statistical software
- SQL structured query language for relational databases
- XML Extensible Markup Language (and HTML) and XPath
- Unix shell commands

Grading

- Homework + projects: add up to 50%. Exact proportion may change. *Tentatively*:
    - Homework = 30% + Projects = 20%
- Homework due in class. After class, before 3:30pm (in my mailbox in Thomas 326): 20% off. After that, 0 credit no matter what.
- Drop two lowest homework scores.
- Midterm: 20%
- Final: 30%

Academic integrity

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