# Introduction to data science

University of washington

Sg.sharif.edu

# Outline

- Brief overview of course (syllabus and projects)
- How we will engage to this course
- How to contribute
- Introduction
- Relational Algebra as a theory of big data computation…

# Brief overview of course (syllabus)

- Part 0: Introduction
  - Data science articulated, data science examples, history and context, technology landscape
- Part 1: Data Manipulation, at Scale
  - Databases and the relational algebra
  - MapReduce, Hadoop, relationship to databases, algorithms, extensions, language; key-value stores and NoSQL; tradeoffs of SQL and NoSQL
  - Data cleaning, entity resolution, data integration, information extraction
- Part 2: Analytics
  - Topics in statistical modeling and experiment design
  - Introduction to Machine Learning, supervised learning, decision trees/forests, simple nearest neighbor
  - Unsupervised learning: k-means, multi-dimensional scaling
- Part 3: Interpreting and Communicating Results
  - Visualization, visual data analytics
  - Backlash: Ethics, privacy, unreliable methods, irreproducible results
- Part 4: Graph Analyticl

# Brief overview of course (Projects)

- Week 1: Twitter Sentiment Analysis (Python)
- Week 2: In-Database Analytics (SQL)
- Week 3: MapReduce Concepts and Algorithms (Python)
- Week 4: (Optional) Large-scale data processing in the cloud (Pig, Hadoop, AWS)
- Week 5: Supervised Learning Roundup (R)
- Week 6: Visualization (Tableau and/or Javascript/D3)
- Week 7: Kaggle Competition

# Prequisites

- We assume
  - some prior programming experience in some language
  - “muscle memory” with basic college statistics
  - some exposure to databases and database concepts
- • One assignment will require writing SQL
- • Two assignments will require writing Python
- • One (optional) assignment will involve processing ~1TB of data  using Amazon Web Services
  - You will pay for these resources, should you choose to complete the assignement
- • One assignment will involve solving a prediction problem on kaggle.com using whatever tools you wish.
- • Some understanding of distributed systems will be helpful, but not required

# Course Philosophy

The skills needed by a data scientist span a variety of different areas

- statistics, programming, databases, systems, visualization

- The traditional organization of topics is not ideal
  - It is difficult to acquire introductory-level knowledge in all areas
  - Cross-cutting concepts and abstractions are obscured

# How we will engage to this course

- Some brief talks about its topic
- Do it’s assignments and modified them for even more understanding
- Do more complex assignments in case of motivation.

# How to contribute

- Get brief lecture about topics of course
- Modify assignments for practical experience about concept
- Work on the course applications and define complex problems
- …

# Also…

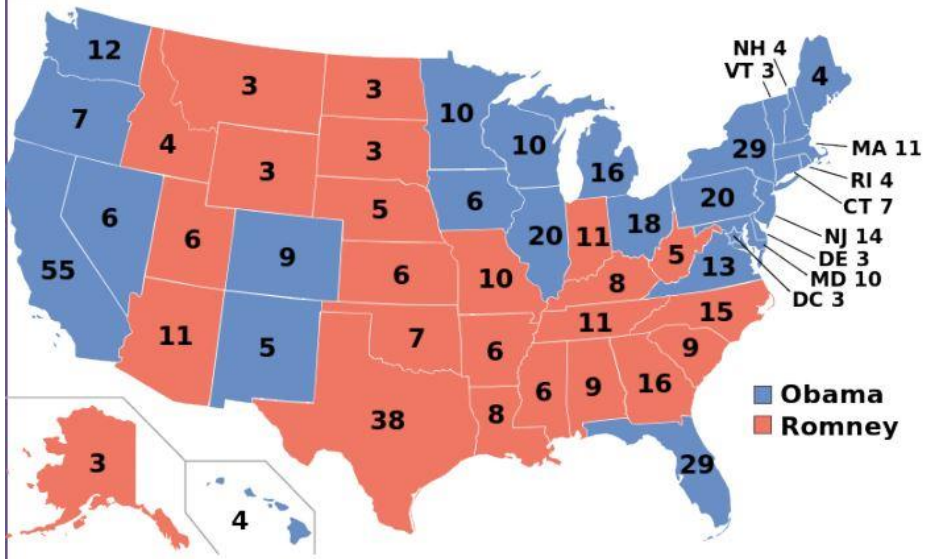
- Talk about Data science applications (in iran)
- Define some complex problem and try to solve them
- Compare different frameworks and their SWOT analysis
- How to aggregate data
  - Twitter/inst/facebook api
  - DBPedia (graph DB)

# Introduction

- Appetite Whetting
- Context
- Dimensions
- This course position in dimenstions
- eScience
- Big Data
- Logistic

# Appetite Whetting

UNIVERSITY of WASHINGTON




http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

## ı's data-driven ground game

date with [the] best data,
es dictated by that data, wins."

Andrew Rasiej, Personal Democracy Forum

good old SQL on a Vertica data
access to data to dozens of
low their own curiosity and
needed."

Dan Woods
Jan 13 2013, CITO Research

ve Hadoop do the aggregate generations
then have Vertica to answer sort of
ut all the data."

dler, CTO of H & K Strategies

4/30/13 Bill Howe, UW 4

# Appetite Whetting

UNIVERSITY of WASHINGTON

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

1) Convert all the digitized books in the 20th century into n-grams (Thanks, Google!)
(http://books.google.com/ngrams/)

*A 1-gram: "yesterday"*
*A 5-gram: "analysis is often described as"*

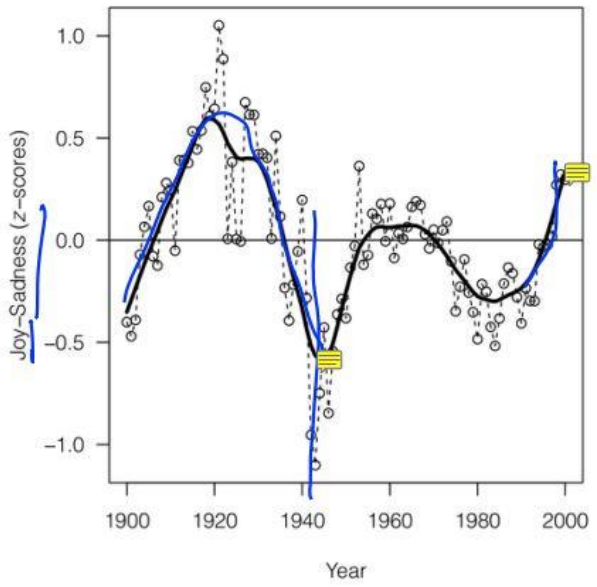
2) Label each 1-gram (word) with a mood score. (Thanks, WordNet!)

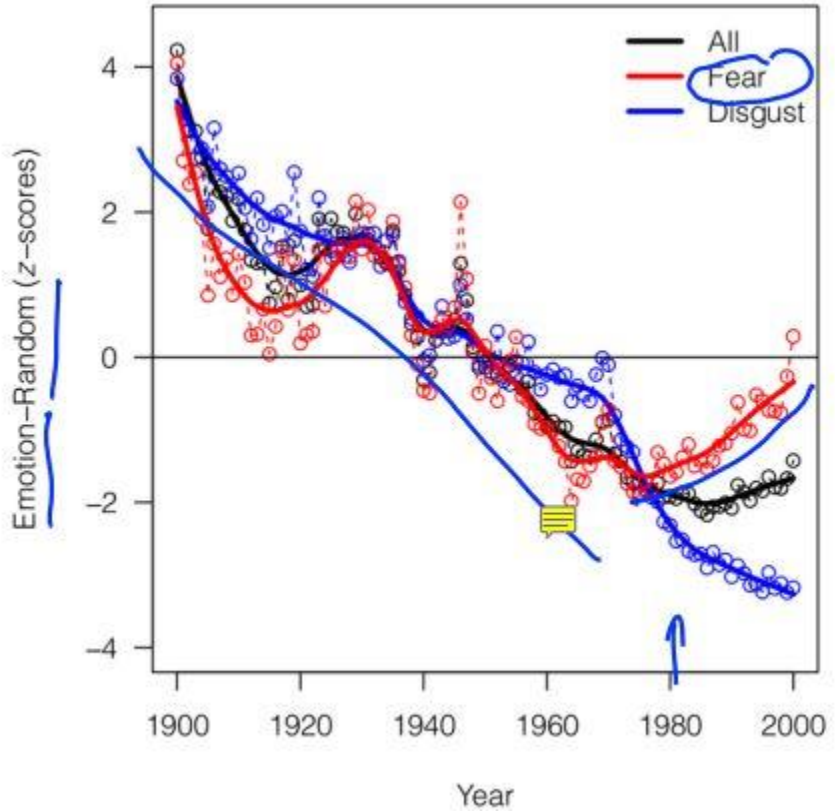
3) Count the occurences of each mood word

$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n} c_i$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$

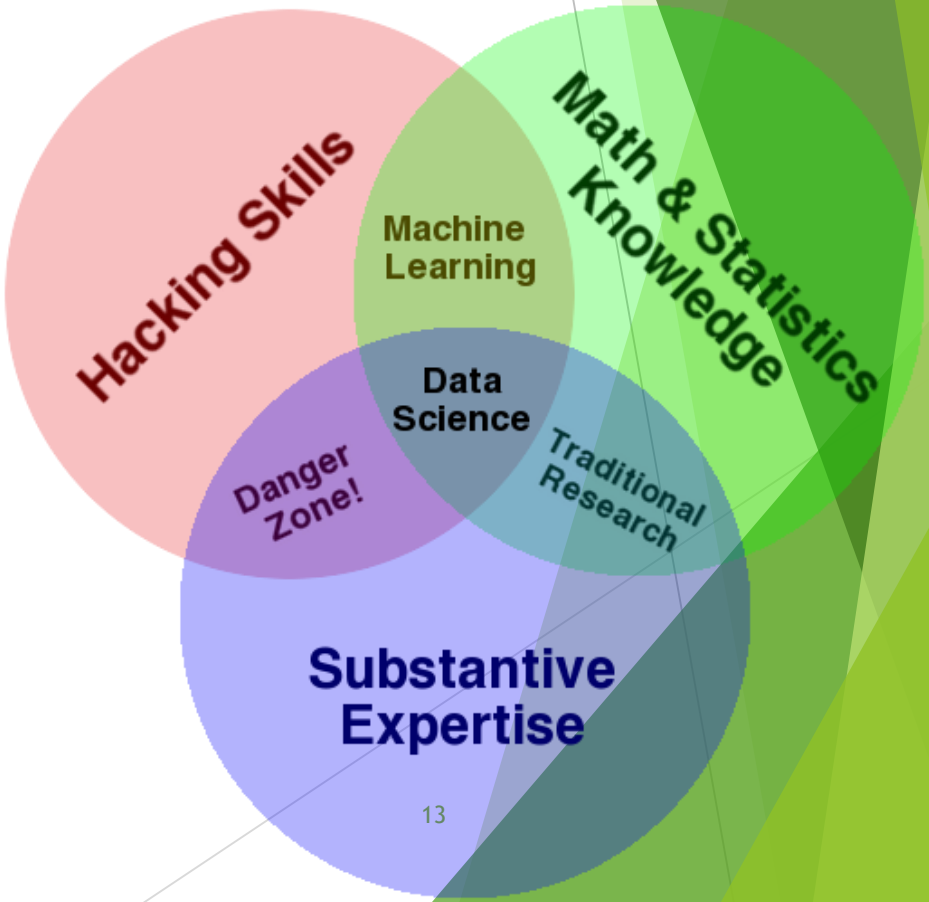
Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expre Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. 10.1371/journal.pone.0059030



# Context

- Drew Conway's Data Science Venn Diagram



# Defination

- “Data Science refers to an emerging area of work  concerned with the collection, preparation, analysis,  visualization, management and preservation of large collections of information.”
  - Jeffrey Stanton
  - Syracuse University School of Information Studies
- “A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine  learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”
  - Hilary Mason, chief scientist at bit.ly

# Three types of tasks:

- 1) Preparing to run a model
  - Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging
- 2) Running the model
- 3) Communicating the results

# Data Science is about Data Products

- “Data-driven apps”
  - Spellchecker
  - Machine Translator
- Interactive visualizations
  - Google flu application
  - Global Burden of Disease
- Online Databases
  - Enterprise data warehouse
  - Sloan Digital Sky Survey

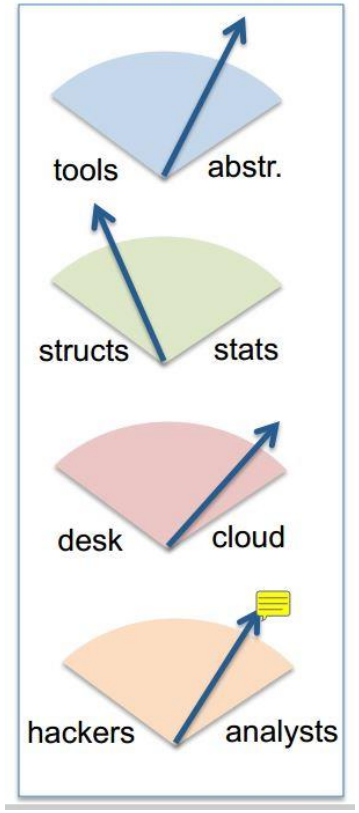
# Distinguishing Data Science from...

- Business Intelligence
  - BI have a particular approach for a particular requirement but DS is broader...
  - high development and couldn’t adapted with changing requirement
  - BI engineer didn’t consume data with their self but DS do both
- Statistics
  - statistic work with data that could be in a memory and a computer but DS work with very sparse and huge data.
  - DS need some engineer to handle and compute through very large data sets.
  - methods and models are the same

# Distinguishing Data Science from...

- Data(base) Management
  - database management work with relational data models (rows and columns)
  - Couldn’t work with video, audio, text, graph(nodes and edges),...
- Visualization
  - like statistics need to handle huge data to visualize them...
- Machine Learning
  - machine learning spend more time on model selection and optimization but in data science it is a very small fraction of time (preparation, manipulation, cleaning, ...)

# Dimensions and this course position



# eScience

SLOAN DIGITAL SKY SURVEY

Empirical

Theoretical

Computational

**eScience**

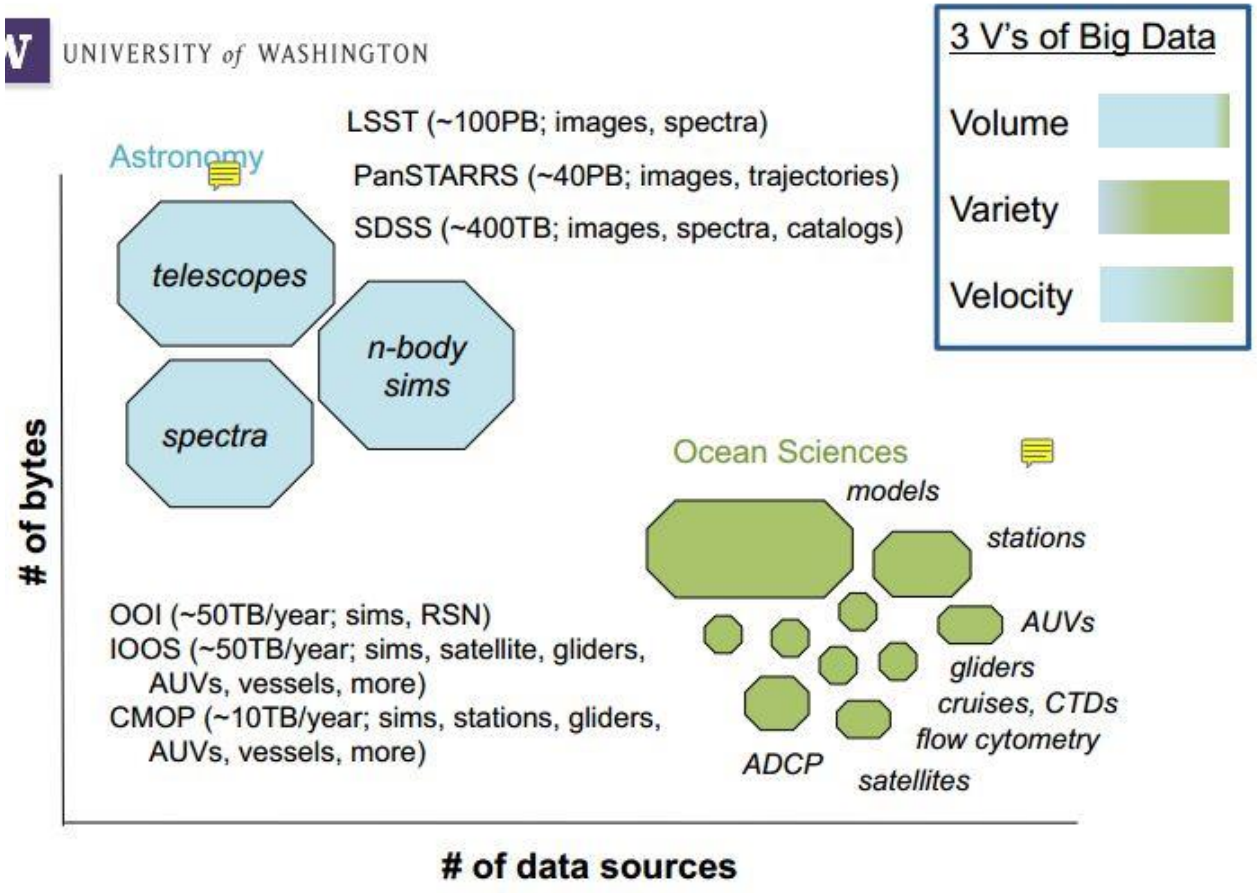
# Science V.C. eScience

- Science is about asking questions
  - Traditionally: “Query the world”
  - Data acquisition activities coupled to a specific hypothesis
- eScience: “Download the world”
  - Data acquired en masse in support of many hypotheses

# Business V.C Science

- Business is beginning to look a lot like science
  - Acquire data aggressively and keep it around
  - Hire data scientists
  - Make empirical decisions

# Big Data





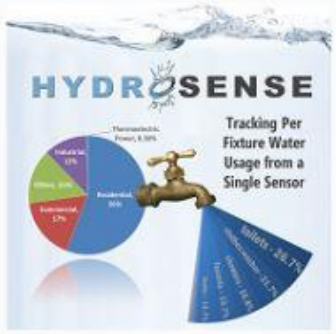

" Big Data is any data that is expensive to manage and hard to extract value from."

Michael Franklin

Takeaway: Disk capacities growing incredibly fast, disk latencies not keeping pace: trouble ahead!

# Big Data Sources

UNIVERSITY of WASHINGTON

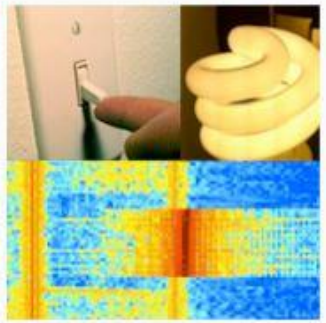



## HydroSense

*Water Fixture Usage with a Single Sensor*

HydroSense is a pressure-based sensor that automatically determines water usage activity and flow down to the source (e.g., dishwasher, laundry, shower) from a single non-intrusive installation point.

**Lead Researchers:** Jon Froehlich, Eric Larson, Shwetak Patel



## ElectriSense

*Electrical Device Energy Usage with a Single Sensor*

ElectriSense is a single plug-in sensor that provides whole home device level usage data. That is, using a single sensor plugged in anywhere in the home, ElectriSense can infer which electrical appliances are on and which off. This data could be used for numerous applications, for example, for providing home owners with itemized electrical bill that not only shows the total energy consumption but breaks the total on a per appliance basis (TV consumed 20 KWh, Lighting consumes 18 KWh and so on).

**Lead Researchers:** Sidhant Gupta, Shwetak Patel

# با تشکر ☺