# Introduction to Data Science

**GIRI NARASIMHAN, SCIS, FIU**

# Course Preliminaries

- Course Webpage: **http: //www.cs.fiu.edu/~giri/teach/5768F18.html**
  - Lecture Slides; Reading Material; Announcements; Homework
  - VISIT OFTEN!
- Class meets 7:50 – 9:05 PM, ECS 132, MW
- Office ECS 254B; Office Hours: By Appointment Only
- Phone: x-3748; Email: giri@cis.fiu.edu
- Final Exam: Monday, 12/3/2018, 7:15 – 9:15 PM, ECS 132

**http: //www.cs.fiu.edu/~giri/teach/5768F18.html**

# What is Data Science?

- Science of what we do to data …
- And why we do those things …

# What else does one do to Data?

- Store
- Search
- Retrieve
-

# What else does one do to Data?

- Collect
- Store
- Manage
- Retrieve
- Analyze
- Visualize
- Mine
- Learn
- Model
- Generate
- Manipulate
- Process
- Clean
- Transform
- Filter
- Search
- Compress
- Uncompress
- Structure
- Randomize
- Encode
- Decode

# Connection to Other Disciplines

- Statistics
- Computer Science
- Mathematics
- Modeling
- Data Mining
- Machine Learning

# Large Repositories

- Government: https://www.data.gov/ (300K datasets)
- Google Earth: https://www.google.com/earth/resources/
- Census Data: https://www.census.gov/data.html
- Finance: https://www.sec.gov/dera/data/financial-statement-data-sets.html
- Public Health Data: https://www.cdc.gov/DataStatistics/
- World Facts: https://www.cia.gov/library/publications/resources/the-world-factbook/
- Genomic & Biotechnology Data: https://www.ncbi.nlm.nih.gov/
- Books; Library of Congress: https://www.loc.gov/

**Homework**: Find one data repository that we did not discuss in class.

# News item from Aug 12, 2018

- **CULTURE** **HTTPS://YOUTU.BE/PBM4-AQV_LG**
  - "Magical mystery song: Math solves Beatles songwriting puzzle"
  - "Even John Lennon and Paul McCartney couldn't remember who wrote the music for In My Life. Stats to the rescue!"
  - https://www.cnet.com/news/magical-mystery-song-math-solves-beatles-songwriting-puzzle/#ftag=CAD-09-10aai5b

# Election Results 2008

- https://www.nytimes.com/elections/2008/results/president/map.html

# Google Earth

- https://earth.google.com/web/@13.01144864,77.55727517,927.41413059a,855.7333781d,35y,0h,0t,0r
- https://earth.google.com/web/@35.13789278,-89.89075679,87.85238738a,212.44263087d,35y,-0h,0t,0r

## Temperature Circle:

A Century of Global Warming, in Just 35 Seconds
By **Brian Kahn**, August 8, 2017

# Course Evaluation

- Homework 50 %
- Class Project 25 %
- Quizzes 10 %
- Participation 5 %
- Exam(s) 10 %

# The Data Science Process

- Formulate the question
- Collect the data
- Explore, Model, Analyze
- Visualize and Interpret
- Communicate and/or Act on it
- Build predictive models for the future
- Iterate

# Class Project Plan

| | |
|---|---|
| Pick a data set of interest | Aug 27 |
| Formulate a set of questions | Sep 10 |
| Download the data; plan tools; identify resources | Sep 17 |
| Plan a strategy; Design algorithms | Sep 24 |
| Analyze, Visualize and Interpret | Oct 8 |
| Present preliminary results | Oct 15 |
| Iterate, Improve , Refine | Oct 29 |
| Final presentation & report | Nov 5, 7 |

# Case Study

- EPA established Dec 2, 1970
- Clean Air Act of 1970
  - Amendments 1977, 1990
- Addressed
  - Emissions
  - Ozone layer
  - Noise Pollution
  - Enforcement
- Did it Work?
  - Mortality
  - Lung diseases
  - Heart Diseases
  - Loss of Productivity
  - Medical Bills
- Causes
  - Industry, Agriculture, Transport, HFCs

# Homework: Compare 1999 & 2012

- Outdoor PM2.5 decreased on average across U.S. due to Clean Air Act.
  - Look at average & SD for 1999 and for 2012 and compare
  - Adjust for the imbalance
  - Compute statistical significance
  - Dig deeper into regional & seasonal differences
  - Suggest factors causing small changes vs big changes from 1999 to 2012
  - Perform time series analysis

# Other Topics for Course

- Connections between Stats, CS, Math, Statistical Modeling, Data Mining, and Machine Learning
- Summarization
- Pattern Discovery, Frequent Itemsets, Trends
- Anomaly Detection
- Feature Extraction
- Clustering
- Privacy and Security
- Conditional Dependence and Causation