NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

NC STATE UNIVERSITY

What is data science?

NC STATE UNIVERSITY

What is data science?

- Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD). -- Wikipedia
- Terms "data mining" and "data science" are used interchangeably

NC STATE UNIVERSITY

What is data science?

- At a high level, *data science* is a set of fundamental principles that guide the extraction of knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles. -- Foster Provost and Tom Fawcett
- This course is about those "fundamental principles"
- For "data mining" – take CSC-522 course.

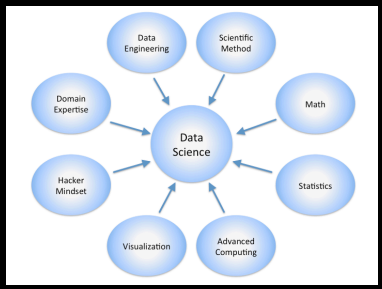
NC STATE UNIVERSITY

Data Science – Modern Origins

- "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." -- William S. Cleveland (2001)
- CODATA Data Science Journal (2002) and Journal of Data Science, by Columbia University (2003).
  - http://www.codata.org/publications/data-science-journal
  - http://www.jds-online.com/

NC STATE UNIVERSITY

DS – A Mash-up of Disciplines



Source: WIKIBOOKS on Data Science

Data Component of Data Science

- Types of data
  - Qualitative and quantitative
- Longitudinal/Panel data
- Time-series/cross-sectional data
- Structured, un-structured, semi-structured
- Data transformations

Finally, Why Data Science?

- Age of big data
- Data Science is emerging as one of the hottest new professions and academic disciplines in these early years of the 21st century. – WSJ
  - Demand is ahead of supply (very good)
- Several major universities started MS in DS
- Data Scientist is the sexiest job of 21st century

Finally, Why Data Science?

- A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician. -- Josh Blumenstock

Big Data

- A lot happens in a minute:
  - YouTube users upload 48 hours of new video
  - Instagram users share 3,600 new photos
  - Brands and organizations on Facebook receive 34,722 "likes"
  - Over 100,000 tweets are sent

Data Never Sleeps: https://www.domo.com/learn/infographic-data-never-sleeps

Brain Component of Data Science

- Mathematics and Statistics
- Mathematics – little bit of linear algebra
- Statistics – major portion of the course

Brain Component of Data Science

- Mathematics and Statistics
- Mathematics – little bit of linear algebra
- Statistics – major portion of the course
  - Summary statistics
  - Probability distributions
  - Hypothesis testing
  - Regression
  - Model Selection
  - Information theory
  - Dimensionality reduction
  - Non-parametric methods
  - Kernel functions
  - Resampling and bootstrap methods
  - Bayesian reasoning
  - Statistical inference
  - Monte carlo methods
  - Mixture models

Brain Component of Data Science

- Mathematics and Statistics
- Mathematics – little bit of linear algebra
- Statistics – major portion of the course
  - Summary statistics
  - Probability distributions
  - Hypothesis testing
  - Regression
  - Model Selection
  - Information theory
  - Dimensionality reduction
  - Non-parametric methods
  - Kernel functions
  - Resampling and bootstrap methods
  - Bayesian reasoning
  - Statistical inference
  - Monte carlo methods
  - Mixture models

Questions

- Does a given fertilizer increase crop yield?
- Does Streptomycin cure Tuberculosis?
- Does smoking cause lung-cancer?

Source: Prof. Rafael

Data Driven Answers

- Does a given fertilizer increase crop yield?
  - Collect and analyze agricultural experimental data
- Does Streptomycin cure Tuberculosis?
  - Collect and analyze randomized trials data
- Does smoking cause lung-cancer?
  - Collect and analyze observational studies data

Source: Prof. Rafael

Waitlisted?

- Can't increase the class size
- As students adjust/drop courses, …

Source: Prof. Rafael