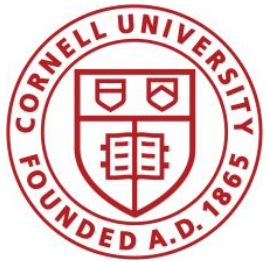


DSFA

Spring 2018

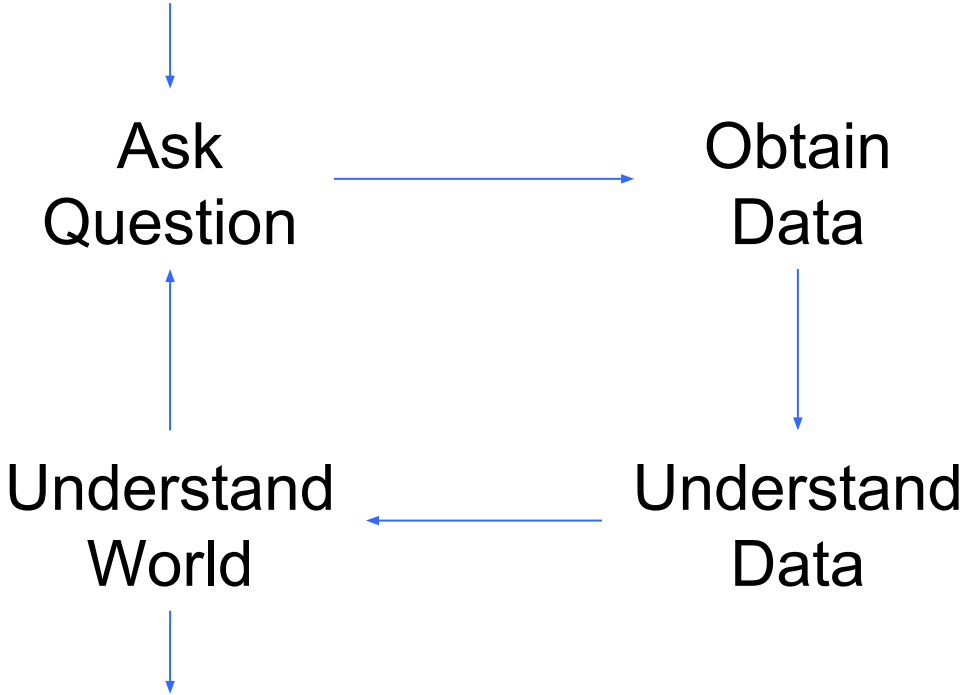
Lecture 40

~~The End~~ What Next

Announcements

- Final exam:
  - Monday, May 14, 2:00 pm, Phillips 219
  - If you have a conflict, TODAY is the deadline to request a makeup
- Course evaluation:
  - Conducted by Engineering College
  - 1% of your final grade
  - Deadline May 14, 8:00 am

Data Science Lifecycle



Applications (lectures and textbook)

- Text of books
- Movies and actors
- Population (US Census)
- Baby birth weight
- Banknote forgery
- Bikeshare trips
- Chronic kidney disease
- Voter database
- Athlete performance
- Flight delays
- Exam scores
- Deflategate
- Galton's heights of parents and children
- House prices
- Hybrid car efficiency
- Salaries (sports, city employees)
- SAT scores
- ...

Applications (assignments)

- Global poverty
- Death penalty and murder rates
- Movie scripts
- World population
- Farmers markets
- Size and age of universe
- Price of diamonds
- Old Faithful eruptions
- Unemployment
- Restaurant inspections
- Sports betting
- ...

What is Data Science? [lec01]

Answering questions from data using computation

- **Exploration**
  - Identifying patterns in information
  - Uses visualizations
- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization
- **Prediction**
  - Making informed guesses
  - Uses machine learning

Data Exploration and Visualization

- Basics of Python programming: 3, 4.1-3
- Arrays: 4.4-6
- Tables: 5, 7

  Concepts: columns, rows, labels

  Operations: sort, where, group, pivot, join, apply
- Plots, charts, graphs: 6

  Concepts: categorical, quantitative

  Kinds: bar, scatter, line, histogram (density)

With this alone, you are now **wizards**

Data Exploration and Visualization

What next?

- Programming in IS: INFO 1300+2300+3300: learn to build web sites, databases, and advanced data visualization techniques
- Programming in CS: CS 1110+2110: learn to engineer software in Python and Java
- On your own: learn Pandas and Matplotlib

Inference

- Experiments: 2

  Treatment, control, confounding factors, association, causation

- Probability: 6.1-2, 8.4-5, 9.1, 9.3, 12

  Laws of probability, distributions, sampling, variability, mean, standard deviation, normal distribution, Central Limit Theorem, bounds

- Hypothesis testing: 10

  Null vs. alternative, test statistics, simulation, p-value

- Estimation: 11

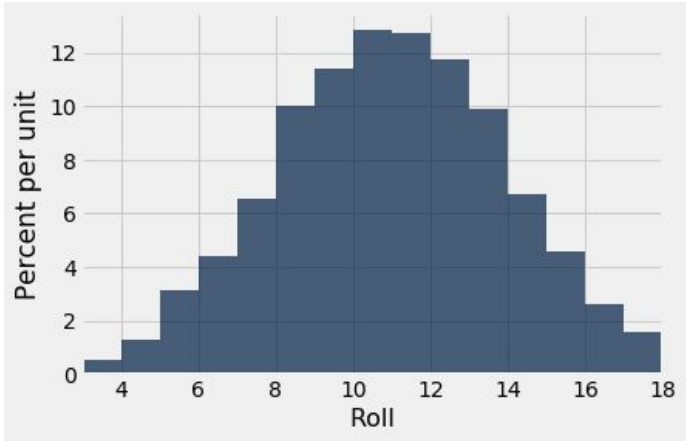
  Bootstrap, percentiles, confidence interval

Inference

MODEL

Probability

Inference



DATA

Inference

What next?

- Statistics (and math prereqs):

  AEM 2100, BTRY 3010, CEE 3040, ECON 3130, ENGRD 2700, HADM 2010, ILRST 2100, MATH 1710 or 4710, PAM 2100, PSYCH 3500, SOC 3010, STSCI 2100
- Learn R: popular for statistics

Prediction

- Regression: 13, 14, 15.6

  Correlation, regression line, RMSE, minimization, residuals, non-linear regression, multiple regression, dummy coding

- Classification: 15

  Nearest neighbors, scaling, distance, decision boundary, train vs. test, accuracy

Prediction

Prediction

Attributes

	Categorical	**Quantitative**
1		1. Linear regression
Many		

Prediction

Prediction

Attributes	Categorical	Quantitative
1		1. Linear regression
Many	2. Nearest neighbor classification	

Prediction

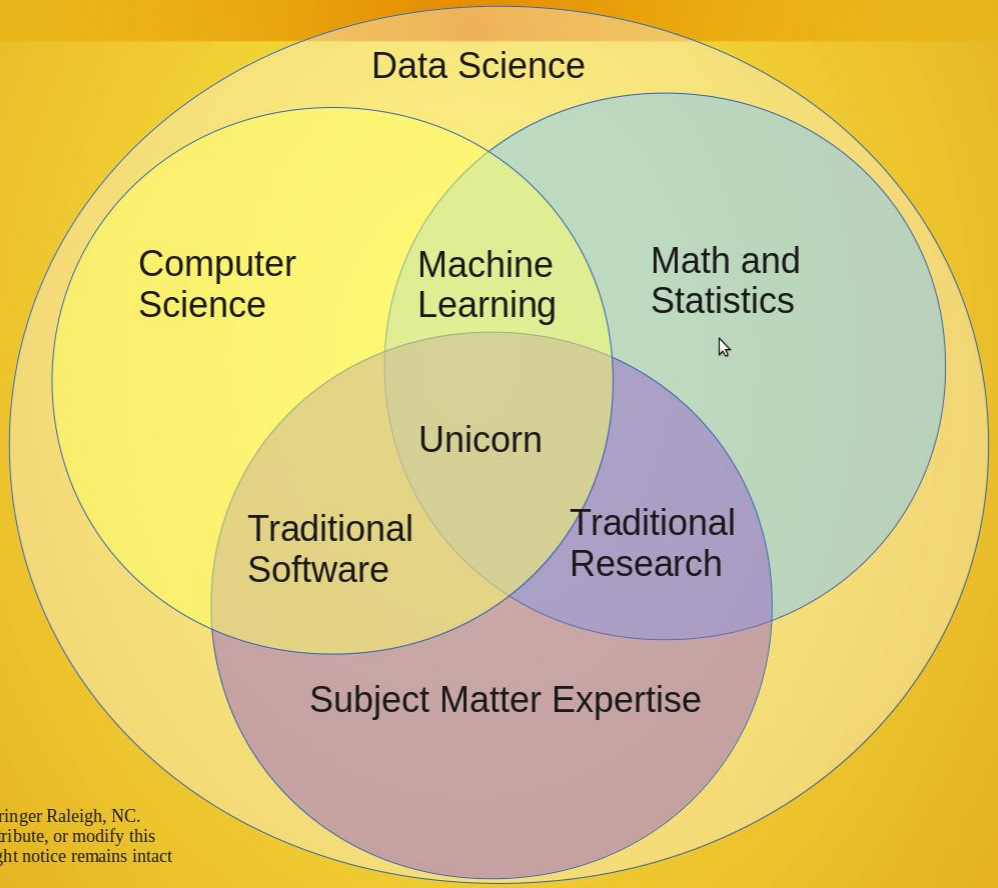
Attributes		Prediction: **Categorical**	Prediction: **Quantitative**
	1		1. Linear regression
	Many	2. Nearest neighbor classification	3. Multiple regression (least squares, NN)

Prediction

What next?

- Linear algebra: MATH 2210, 2310, or 2940 (some calculus required)
- Machine learning: CS 4780, 4786, ORIE 4740, 4741, STSCI 4740, 4780 (and probably many others)
- On your own: try a self-paced tutorial or competition on Kaggle

Data Science Venn Diagram v2.0



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact

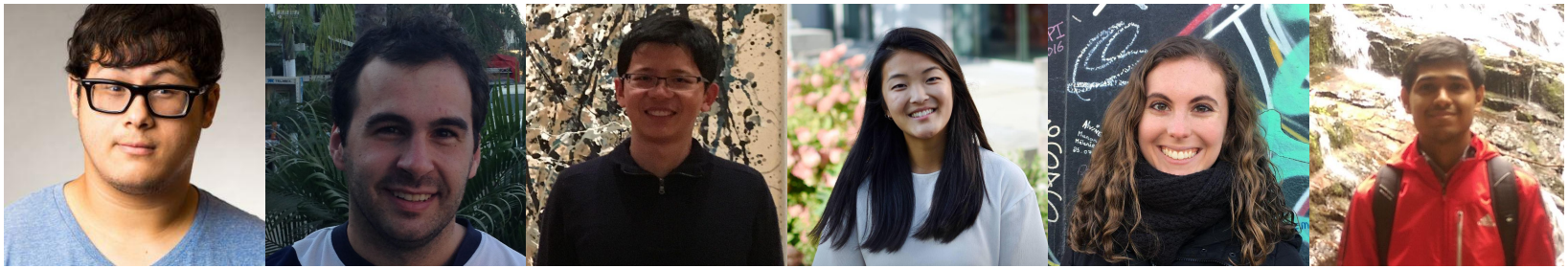
More Data Science

- Next steps: ORIE 2380, INFO 2950
- Learn R or Julia: other popular data science platforms
- Cornell Data Science (CDS) project team, INFO 1998

Thank you to TAs!

Skyler Seto, Tony Sirianni

Yang Guo, Charlene Luo, Lauren Sedita, Anil Vadali



Thank you!

To all of **you**!

You were part of a grand adventure!

- New course
- New staff
- New assignments
- New technology

Finally

Stay in touch! On behalf of Prof. Udell and myself…

- Tell us when 1380 helps you out in the future
- Ask us cool questions
- Drop by our offices to tell us about the rest of your time at Cornell (and beyond)... We really do like to know

Finally

GO DO AMAZING THINGS WITH YOUR LIFE