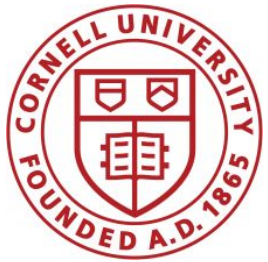


DSFA

Spring 2020

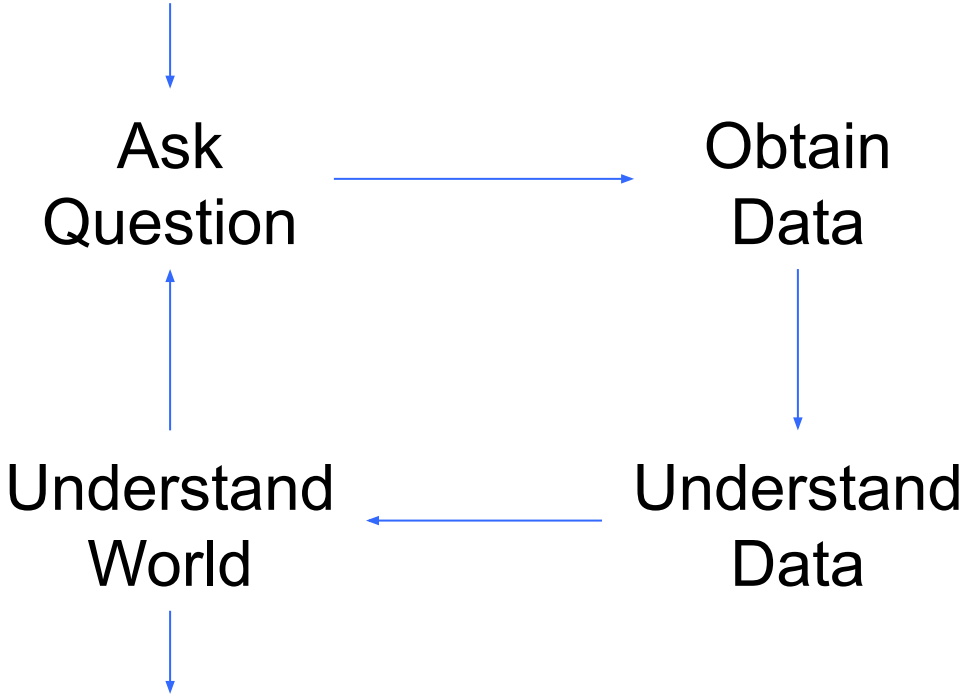
Lecture 26

~~The End~~ What Next

What did we miss?

- Multiple linear regression (15.5)
  - Using multiple numerical variables to predict a numerical variable
- Inference for regression (Chapter 14)
  - Regression on a sample; what can we say about the slope of the line for the population?
- Some guest lectures on examples of data science research, and data and surveillance/privacy

Data Science Lifecycle



Applications (lectures and textbook)

- Text of books
- Movies and actors
- Population (US Census)
- Baby birth weight
- Bikeshare trips
- Chronic kidney disease
- Voter database
- Athlete performance
- Flight delays
- Exam scores
- Galton's heights of parents and children
- 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
- 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

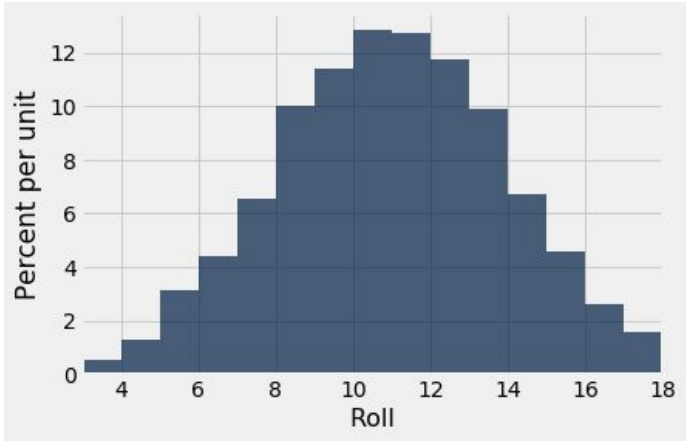
  Bootstrap, percentiles, confidence interval

Inference

MODEL

Probability

Inference



DATA

Inference

What next?

- Statistics (and math prereqs):

  AEM 2100, BTRY 3010/STSCI 2200, CEE 3040, ECON 3130, ENGRD 2700, HADM 2010, ILRST/STSCI 2100, MATH 1710 or 4710, PAM 2100, PSYCH 3500, SOC 3010
- Learn R: popular for statistics

Prediction

- Regression: 13

  Correlation, regression line, RMSE, minimization, residuals, non-linear regression

- 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

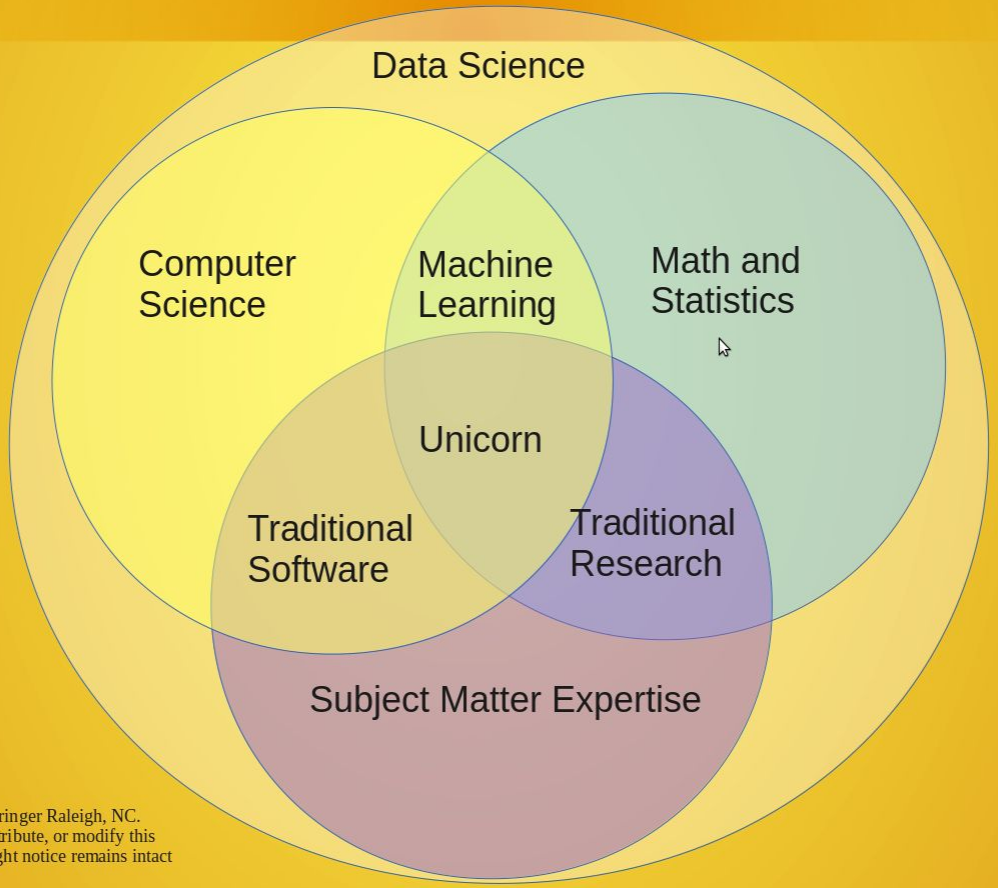
Attributes		Prediction: **Categorical**	Prediction: **Quantitative**
	1		1. Linear regression
	Many	2. Nearest neighbor classification	3. Multiple regression (least squares, NN)

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



Other Data Science courses

ORIE 2380: Urban Analytics (MQR-AS)

- Followup course; more sophisticated regression, classification, learning
- Lots of case studies

INFO 2950: Intro to Data Science

- More sophisticated treatment of similar material as 1380

ORIE 3120: Practical Tools for Operations Research, Machine Learning, and Data Science

- Data handling, more prediction methods

Other Data Science courses

HD/PSYCH 2930: Intro to Data Science for Social Scientists

- Looks similar to 1380, uses R instead of Python

AMST/ENGL/INFO 1350: Intro to Cultural Analytics

- Data science applied to understanding texts, humanistic research (same level as 1380).

INFO 3350: Text Mining History and Literature

STS 3440: Data Science and Society Lab

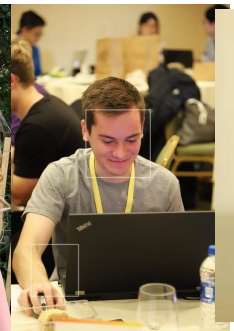
More Data Science

- Learn R or Julia: other popular data science platforms
- Cornell Data Science (CDS) project team (https://cornelldata.science), INFO 1998

Thank you to TAs!

Ben Baer, Artem Bolshakov

Julie Barron, Taeho Kim, Daniel Sanky, Kate Schrage, Anders Wikum, Yao Yu Yeo



Thank you!

To all of **you**!

For being brave and doing difficult things in difficult circumstances.

Finally

Stay in touch! On behalf of Prof. Entner and myself...

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