

STT481 Capstone in Statistics

Lecture 0: Introduction

Chih-Li Sung

08/28/2019

Statistics in the news

- ▶ Quote of the Day, New York Times, August 5, 2009

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."

– Hal Varian, chief economist at Google

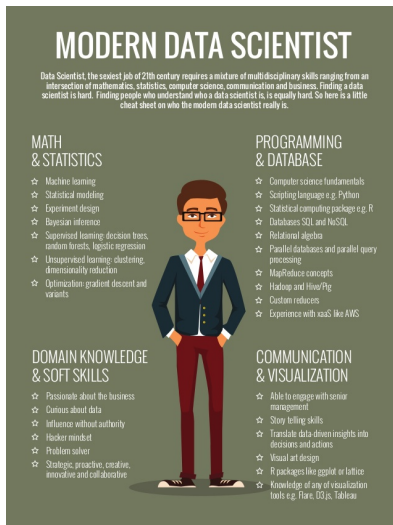
- ▶ Quote of [Forbes](#), Dec 11, 2017

"LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning"

– Louis Columbus

Are you ready to be a statistician/data scientist

► Data scientist skill-set ([source](#))



MarketingDiliter.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data warehousing and big data systems, marketing channel insights on Paid Search, SEO, Social, CRM and brand.

**Marketing
DILITER**

What we have learned so far

- ▶ Basic Statistics
- ▶ Probability
- ▶ Hypothesis Testing
- ▶ Linear Regression (maybe?)
- ▶ ...

What this course is about

- ▶ Statistical capstone experiences are essential for statisticians/data scientists to perform an in-depth analysis of real-world data.
- ▶ Capstone experiences can develop statistical thinking by engaging in a consulting-like experience that requires skills outside the scope of traditional courses:
 - ▶ defining a complex problem,
 - ▶ analyzing data,
 - ▶ building a strong team,
 - ▶ programming techniques,
 - ▶ and communicating effectively.

What you should expect to learn in this course

- ▶ Problem formulation
- ▶ Data collection
- ▶ Advanced statistical modeling, preliminary data analysis, and machine learning
- ▶ Statistical software (R)
- ▶ Thorough and elaborate statistical analyses of data
- ▶ Presentation and data visualization

A standard procedure of statistical analysis

- Statistics divides the study of data into *five* steps:

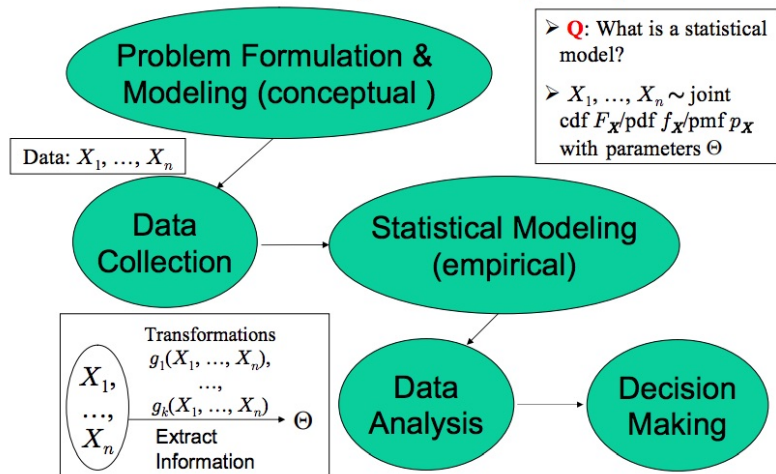


Figure 2:

Data collection

"In God we trust; all others bring data." – Edwards Deming

- ▶ Design of experiment (not covered)
- ▶ Survey sampling (not covered)
- ▶ Web scraping
- ▶ Text mining (not covered)

Most Popular Feature Films Released 2016-01-01 to 2016-12-31

1 to 100 of 12,613 titles | [Next »](#)

View Mode: [Compact](#) | **[Detailed](#)**

Sort by: **Popularity**▲ | [Alphabetical](#) | [IMDb Rating](#) | [Number of Votes](#) | [US Box Office](#) | [Runtime](#) | [Year](#) | [Release Date](#)



1. **Sing** (2016)

PG | 108 min | Animation, Comedy, Family

★ 7.2

☆ [Rate this](#)

59 Metascore

In a city of humanoid animals, a hustling theater impresario's attempt to save his theater with a singing competition becomes grander than he anticipates even as its finalists' find that their lives will never be the same.

Directors: Christophe Lourdelet, Garth Jennings | **Stars:** Matthew McConaughey, Reese Witherspoon, Seth MacFarlane, Scarlett Johansson

Votes: 40,603 | **Gross:** \$269.36M



2. **Moana** (I) (2016)

PG | 107 min | Animation, Adventure, Comedy

★ 7.7

☆ [Rate this](#)

81 Metascore

In Ancient Polynesia, when a terrible curse incurred by the Demigod Maui reaches an impetuous Chieftain's daughter's island, she answers the Ocean's call to seek out the Demigod and set things right.

Statistical modeling

“Essentially, all models are wrong, but some are useful.” – George Box

- ▶ Supervised learning
 - ▶ Linear regression model
 - ▶ Logistic Regression
 - ▶ Linear Discriminant Analysis
 - ▶ KNN
 - ▶ Linear Model Selection and Regularization
 - ▶ Nonparametric Regression
 - ▶ Tree-Based Methods
 - ▶ Support Vector Machines
- ▶ Unsupervised learning
 - ▶ Principal Components Analysis
 - ▶ Clustering Methods

Decision making/visualization

- ▶ ggplot2
- ▶ plotly
- ▶ shiny (myapp)
- ▶ D3.js

Grading

- ▶ Final grades will be based on
 - ▶ one course project (20%)
 - ▶ five homework assignments ($40\% = 5 \times 8\%$)
 - ▶ two midterms ($40\% = 2 \times 20\%$)
- ▶ The final grade would be based on your total grade percentage and will be determined roughly as:

##

## %	90-100	85-89	80-84	75-79	70-74	65-69	60-64	0-59
## Grade	4	3.5	3	2.5	2	1.5	1	0

Class Participation:

- ▶ Your class participation will be used to determine whether your grade can be lifted in case you are right on the edge of two grades.
- ▶ Participation means attending classes, participating in class discussions, asking relevant questions, volunteering to provide answers to questions, and providing constructive criticism and creative suggestions that improve the course.

Homework

- ▶ Homework assignments include conceptual and applied exercises. Typesetting your reports/solutions in *Latex* or *R markdown* ([link1](#), [link2](#)) is strongly encouraged (you will receive **5 extra credit points**). Unreadable handwriting is subject to zero credit. Unreadable handwriting is subject to zero credit.
- ▶ Homework Policy: [link](#)

Project

- ▶ [Kaggle](#) is a open platform for predictive modeling and analytics competitions, where companies and researchers can post their data and problems for users to solve.
- ▶ In this project, you are given the house dataset on Kaggle and the goal is to predict the final price of each home in Ames, Iowa. See [link](#).
- ▶ You may form groups of up to three students.
- ▶ Grading:
 - ▶ Accurate prediction - $0.12 / (\text{your prediction score}) \times 10\%$ on 12/11 7:45am (10%)
 - ▶ Model Interpretation & Final Presentation - on 12/11 7:45am - 9:45am (10%)
- ▶ Final Presentation: See [link](#).

Important Dates and Policy

See [syllabus](#)

Any questions or feedback?

Homework 1, Q1: Find your teammates (up to 3 students in a group) and create an account on Kaggle (take a screenshot as below).

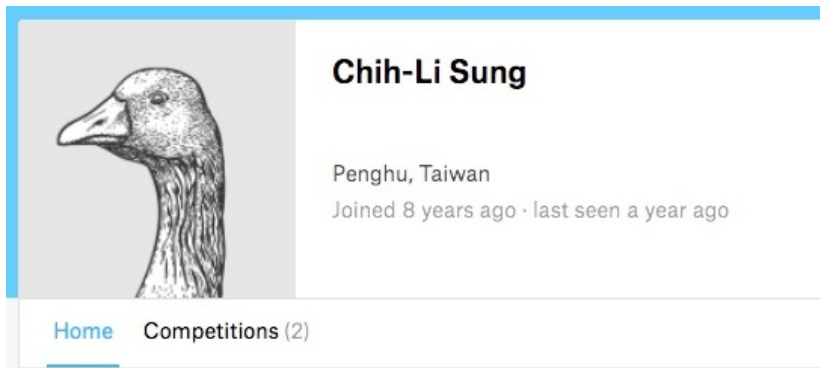


Figure 4: