MTH499/522 Lecture Notes 01

Donghui Yan

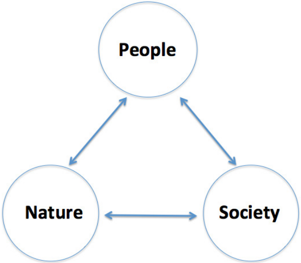
Department of Math, Umass Dartmouth

Outline

- Introduction
- Selected applications
- MTH499/522

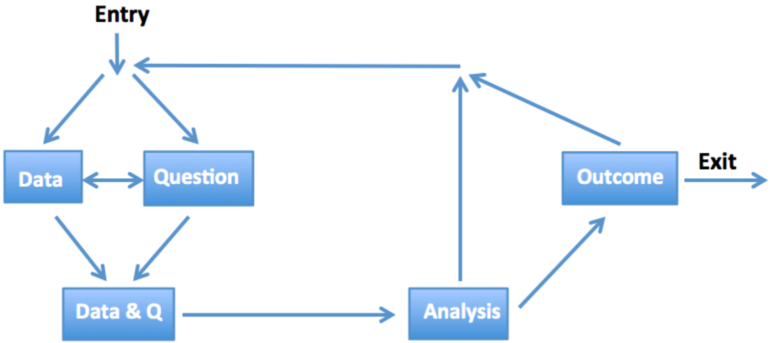
Broader view of data

- Data records activities about
  - *People, society, the nature, and their interactions*
  - Helps understand and gain insights
    - Just think how we infer about people and events in history
    - Written records, archaeological findings, folklore etc as *data*
  - May be more *faithful* than market survey



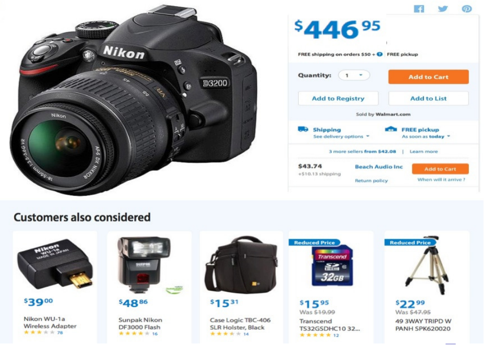
- More and more data are available
  - As Internet, social media, wireless and portable devices become popular.

The life cycle of data analysis



Big data @Walmart

- Personalized page loading
- Item recommendation
  - Users buying/viewing product A may also like B
- Targeting
  - Distribution of ads, coupons to users who may consider purchase



$446.95

FREE shipping on orders $50 + | FREE pickup

Quantity: 1

Sold by Walmart.com

Shipping
See delivery options

FREE pickup
As soon as today

3 more sellers from $42.08 | Learn more

$43.74
+$10.13 shipping
Beach Audio Inc
Return policy
When will it arrive ?

Customers also considered

$39.00
Nikon WU-1a Wireless Adapter
★★★☆☆ 78

$48.86
Sunpak Nikon DF3000 Flash
★★★★☆ 16

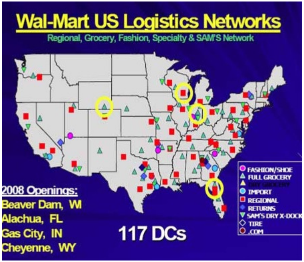
$15.31
Case Logic TBC-406 SLR Holster, Black
★★★★☆ 14

Reduced Price
$15.95
Was ~~$19.99~~
Transcend TS32GSDHC10 32...
★★★★★ 12

Reduced Price
$22.99
Was ~~$47.95~~
49 3WAY TRIPD W PANH SPK620020

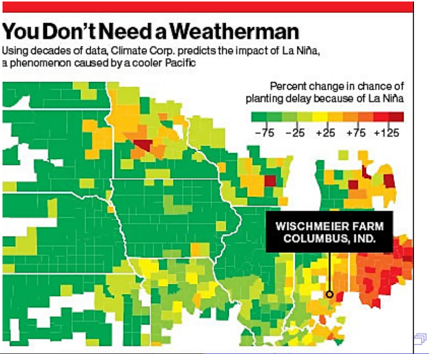
Big data @Walmart (continued)

- Market basket analysis
  - Products affinity based on transaction data
  - Affinity-based shelving for stores/warehouse
  - Affinity-aware allocation algorithm for logistics
    - Products often bought together are shipped together
    - To reduce inventory, package splits, and improve order fulfillment.



The Climate Corporation

- Provide insurance to farmers under extreme weather conditions
  - Global agriculture industry estimated at $3 trillion
  - Weather data at 2.5m locations + 150b soil measurements
  - 10 trillion weather simulation data points
- Acquired by Mosanto at $930m in 2013.



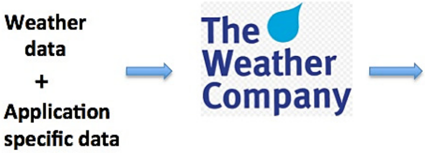
You Don't Need a Weatherman

Using decades of data, Climate Corp. predicts the impact of La Niña, a phenomenon caused by a cooler Pacific

Percent change in chance of planting delay because of La Niña
−75
−25
+25
+75
+125
WISCHMEIER FARM COLUMBUS, IND.

The Weather Company

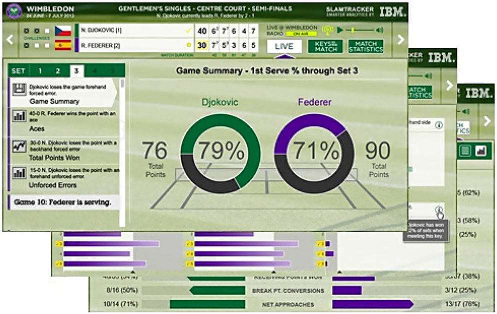
- Weather is one largest external swing factor in business
  - Annual economic impact ≈ \$0.5 trillion in US
- To help understand behaviors of digital and mobile users
  - In 3 million locations worldwide
  - Along with climate data in each locale
- Half of its revenue from digital operations.



- Ads targeting (e.g., anti-frizz shampoo for humid weather)
- Airline cancellation
- Energy for a wind farm
- Insurers to alert users about impeding hails or storms
- Better estimate of power usage

Sports statistics at IBM

- IBM's SlamTracker
- 8 years of *Grand Slam* tennis tournaments data
  - $\sim$ 41m data points
- Top 3 key actions to enhance chance of winning
  - Target serve percentages, rally counts, types of shots
- In 2012, USOpen.org logged 45.6m visits and 325m page.



Djokovic loses the game forehand forced error.
Game Summary

40-0 R. Federer wins the point with an ace
Aces

30-0 N. Djokovic loses the point with a backhand forced error
Total Points Won

15-0 N. Djokovic loses the point with an forehand unforced error.
Unforced Errors

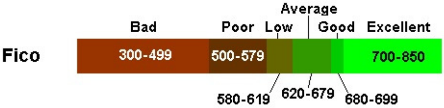
Game 10: Federer is serving.

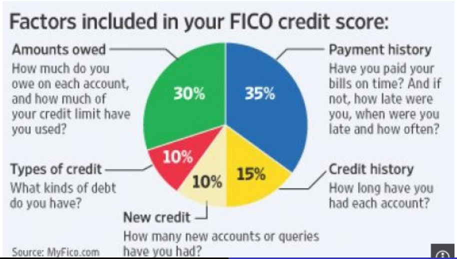
Healthcare

- Health data used in various ways
  - Prevention, intervention, diagnosis, and maybe re-admission
- Devices
  - Smart phones (e.g., for blood test from Berkeley, EPFL)
  - Wearable devices (FitBit, Jawbone, Samsung Gear Fit etc)

FICO analytics

- 10b+ FICO scores purchased in 2013
  - Excluding 30m personal purchases by US customers
- 65% credit cards managed worldwide
- 2.5b payment cards protected from fraud





Factors included in your FICO credit score:

Amounts owed
How much do you owe on each account, and how much of your credit limit have you used?

Payment history
Have you paid your bills on time? And if not, how late were you, when were you late and how often?

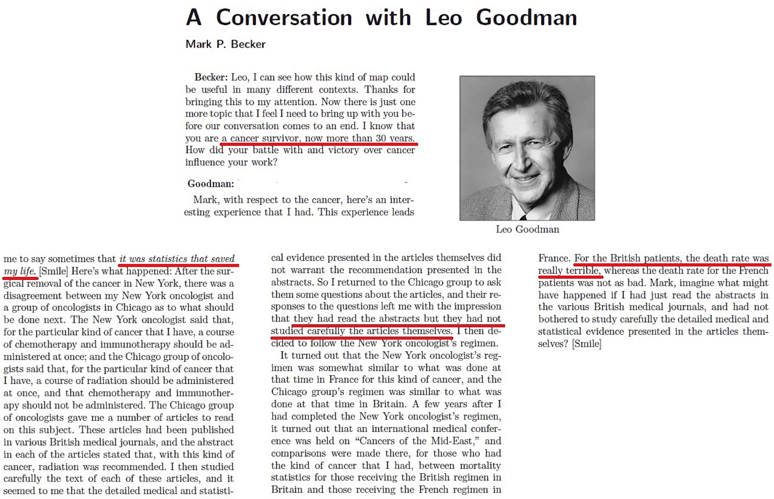
Types of credit
What kinds of debt do you have?

Credit history
How long have you had each account?

New credit
How many new accounts or queries have you had?

Source: MyFico.com

The story of Leo Goodman



A Conversation with Leo Goodman

Mark P. Becker

Becker: Leo, I can see how this kind of map could be useful in many different contexts. Thanks for bringing this to my attention. Now there is just one more topic that I feel I need to bring up with you before our conversation comes to an end. I know that you are a cancer survivor, now more than 30 years. How did your battle with and victory over cancer influence your work?

Goodman:

Mark, with respect to the cancer, here's an interesting experience that I had. This experience leads me to say sometimes that *it was statistics that saved my life.* [Smile] Here's what happened: After the surgical removal of the cancer in New York, there was a disagreement between my New York oncologist and a group of oncologists in Chicago as to what should be done next. The New York oncologist said that, for the particular kind of cancer that I have, a course of chemotherapy and immunotherapy should be administered at once; and the Chicago group of oncologists said that, for the particular kind of cancer that I have, a course of radiation should be administered at once, and that chemotherapy and immunotherapy should not be administered. The Chicago group of oncologists gave me a number of articles to read on this subject. These articles had been published in various British medical journals, and the abstract in each of the articles stated that, with this kind of cancer, radiation was recommended. I then studied carefully the text of each of these articles, and it seemed to me that the detailed medical and statistical evidence presented in the articles themselves did not warrant the recommendation presented in the abstracts. So I returned to the Chicago group to ask them some questions about the articles, and their responses to the questions left me with the impression that they had read the abstracts but they had not studied carefully the articles themselves. I then decided to follow the New York oncologist's regimen.

It turned out that the New York oncologist's regimen was somewhat similar to what was done at that time in France for this kind of cancer, and the Chicago group's regimen was similar to what was done at that time in Britain. A few years after I had completed the New York oncologist's regimen, it turned out that an international medical conference was held on "Cancers of the Mid-East," and comparisons were made there, for those who had the kind of cancer that I had, between mortality statistics for those receiving the British regimen in Britain and those receiving the French regimen in France. For the British patients, the death rate was really terrible, whereas the death rate for the French patients was not as bad. Mark, imagine what might have happened if I had just read the abstracts in the various British medical journals, and had not bothered to study carefully the detailed medical and statistical evidence presented in the articles themselves? [Smile]

Leo Goodman

Other applications

- Search engine design
- Computational advertising
- Finance, marketing, and hedge fund
- Sports statistics
- Digital journalism.

MTH499/522

Focus on the modeling aspect of data science

- Linear models
  - Model assumption, fitting, and diagnosis
  - Model selection
- Nonparametric statistics
  - Tree-based models etc
- Machine learning
  - Logistic regression
  - Neural networks (deep learning)
  - Support vector machines (if time allows).