THE ANALYTICS EDGE

Intelligence, Happiness, and Health

15.071x – The Analytics Edge

Data is Powerful and Everywhere

- 2.7 Zettabytes of electronic data exist in the world today – 2,700,000,000,000,000,000,000 bytes
  - This is equal to the storage required for more than 200 billion HD movies
- New data is produced at an exponential rate.
- Decoding the human genome originally took 10 years to process; now it can be achieved in one week

Data and Analytics are Useful

- Estimated that there is a shortage of 140,000 – 190,000 people with deep analytical skills to fill the demand of jobs in the U.S. by 2018

14

- IBM has invested over $20 billion since 2005 to grow its analytics business
- Companies will invest more than $120 billion by 2015 on analytics, hardware, software and services
- Critical in almost every industry
  - Healthcare, media, sports, finance, government, etc.

What is Analytics?

- The science of using **data** to build **models** that lead to better **decisions** that add **value** to individuals, to companies, to institutions.

This Class

- **Key Messages**:
  - Analytics provide a competitive edge to individuals, companies and institutions
  - Analytics are often critical to the success of a company
- **Methodology**: Teach analytics techniques through real world examples and real data
- **Our Goal**: Convince you of the Analytics Edge, and inspire you to use analytics in your career and your life

Teaching Team

- Dimitris Bertsimas
  - MIT professor since 1988
- Allison O'Hair
  - Received her Ph.D. from MIT in 2013
- Teaching Assistants
  - Iain Dunning, Angie King, Velibor Misic, John Silberholz, Nataly Youssef
  - Ph.D. students in the Operations Research Center at MIT

This Lecture

- Summary of some of the examples we will cover
  - IBM Watson
  - eHarmony
  - The Framingham Heart Study
  - D2Hawkeye

- Other examples we will cover in this class
  - Moneyball, Supreme Court, Elections, Twitter, Netflix, Airline Revenue Management, Radiation Therapy, Sports Scheduling, . . .

IBM Watson – A Grand Challenge

- IBM Research strives to push the limits of science
- Deep Blue – a computer to compete against the best human chess players
  - A task that people thought was restricted to human intelligence
- Blue Gene – a computer to map the human genome
  - A challenge for computer speed and performance
- In 2005, they decided to create a computer that could compete at *Jeopardy!*, a popular game show

Why is *Jeopardy!* Hard?

- *Jeopardy!* asks the contestants to answer cryptic questions in a huge variety of categories
- Generally seen as a test of human intelligence, reasoning, and cleverness
- No links to the outside world permitted
- New questions and categories are created for every show

Watson

- Watson is a supercomputer with 3,000 processors and a database of 200 million pages of information
- A massive number of data sources
  - Encyclopedias, texts, manuals, magazines, Wikipedia, etc.
- Used over 100 different analytical techniques for analyzing natural language, finding candidate answers, and selecting the final answer
  - We will discuss this more later in the class

The Competition

- In February 2011, a two-game exhibition match aired on television (6 years later)
- Watson competed against the best two human players of all time, and challenged the meaning of intelligence
- Now, Watson is being used for many applications, including selecting the best course of treatment for cancer

What is the Edge?

- Watson combined many algorithms to increase accuracy and confidence
  - We will cover many of them in this class
- Approached the problem in a different way than how a human does
- Deals with massive amounts of data, often in unstructured form
  - 90% of data in the world is unstructured

eHarmony

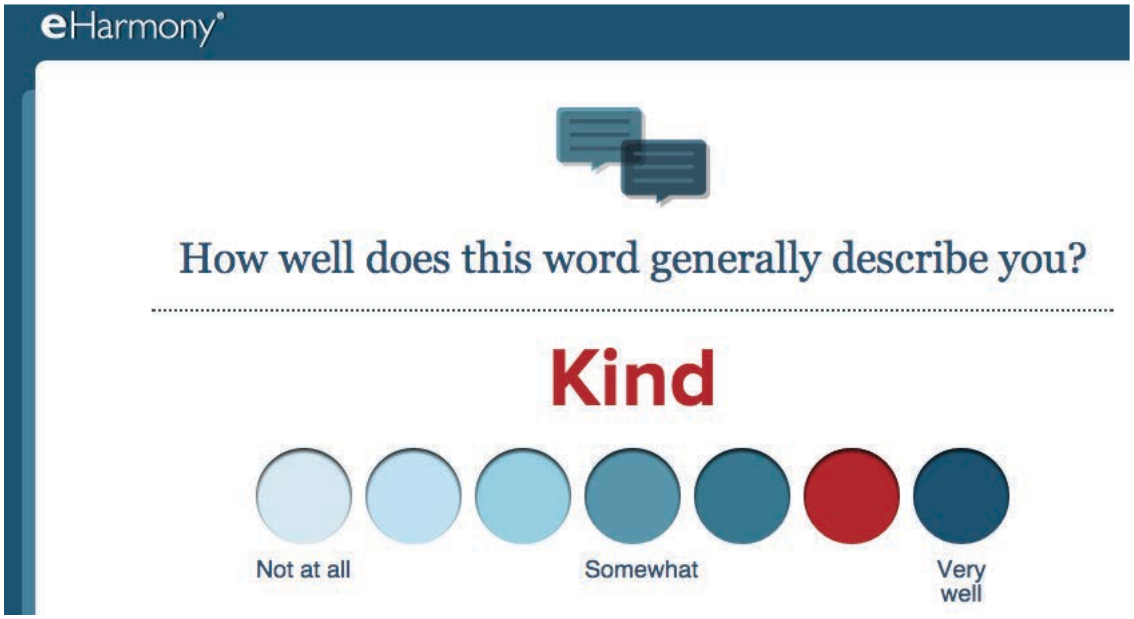
- Online dating site focused on long term relationships
- Takes a scientific approach to love and marriage
- Nearly 4% of US marriages in 2012 are a result of eHarmony
- Has generated over $1 billion in cumulative revenue

Finding Successful Matches

- First predict if users will be compatible
  - Use 29 different "dimensions of personality"
- Then need to find matches for everyone
  - Members in more than 150 countries
  - Since launching in 2000, more than 33 million members
- They use regression and optimization
  - Operates eHarmony Labs, a relationship research facility

The Data

- Collect data through 436 questions
- About 15,000 people take the questionnaire each day



What is the Edge?

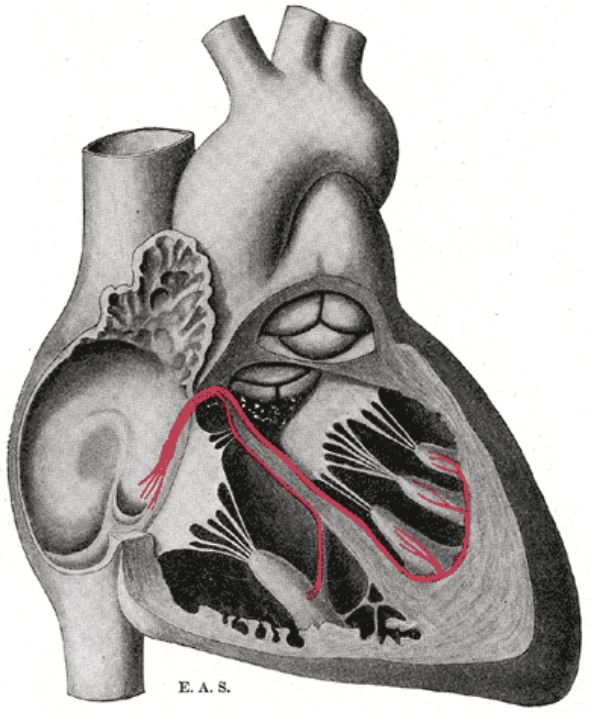
- Relies much more on data than other dating sites
- Suggests a limited number of high quality matches
  - Users don't have to search and dig through profiles
- eHarmony has successfully leveraged the power of analytics to create a successful and thriving business
  - 14% of US online dating market

The Framingham Heart Study

- One of the most important studies of modern medicine
- Ongoing study of the residents in Framingham, MA
  - Started in 1948, now on the third generation
- Much of the now-common knowledge regarding heart disease came from this study
  - High blood pressure should be treated
  - Clogged arteries are not normal
  - Cigarette smoking can lead to heart disease

Heart Disease

- Heart disease has been the leading cause of death worldwide since 1921
  - 7.3 million people died from CHD in 2008
- Since 1950, age-adjusted death rates have declined 60%
  - In part due to the results of the Framingham Heart Study



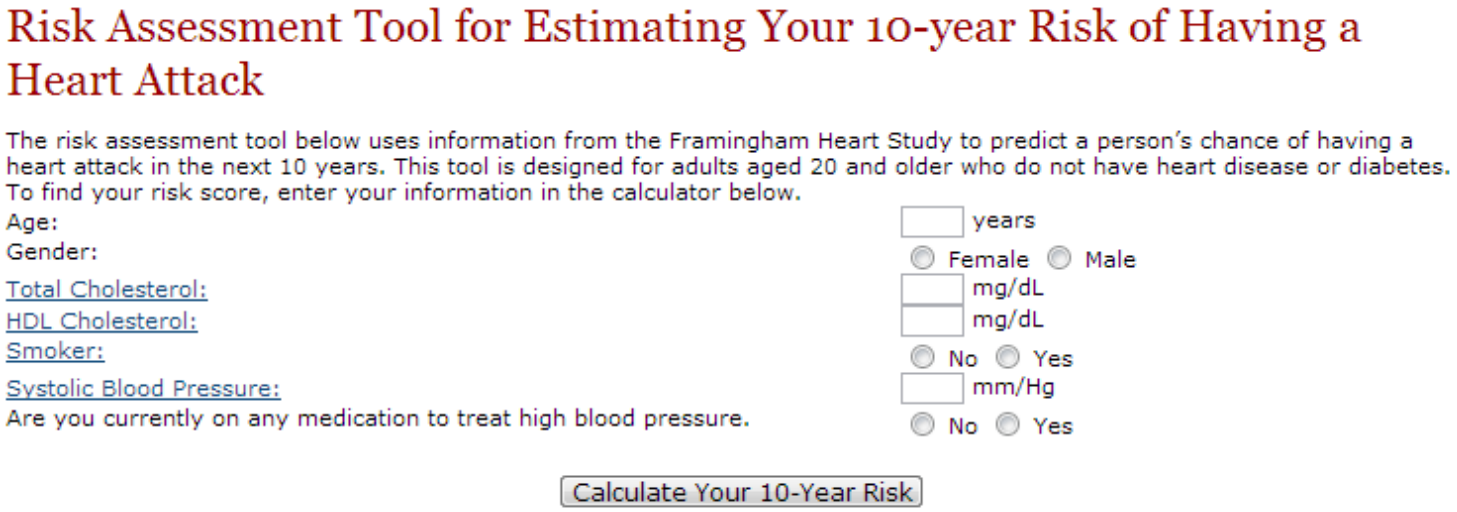
Heart diagram is in the public domain. Source: Wikimedia Commons.

The Data

- 5,209 patients were enrolled in 1948
- Given a questionnaire and exam every two years
  - Physical characteristics
  - Behavioral characteristics
  - Test results
- Patient population, exam, and questions expanded over time

An Analytics Approach

- Used regression to predict whether or not a patient would develop heart disease in the next ten years
- Model tested and adjusted for different populations
- Available online



Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age: ____ years

Gender: ○ Female ○ Male

Total Cholesterol: ____ mg/dL

HDL Cholesterol: ____ mg/dL

Smoker: ○ No ○ Yes

Systolic Blood Pressure: ____ mm/Hg

Are you currently on any medication to treat high blood pressure. ○ No ○ Yes

Calculate Your 10-Year Risk

What is the Edge?

- Provided necessary evidence for the development of drugs to lower blood pressure
- Paved the way for other clinical prediction rules
  - Predict clinical outcomes using patient data
- A model allows medical professionals to make predictions for patients worldwide

D2Hawkeye

- Medical software company founded in 2001
- Combined data with analytics to improve quality and cost management in healthcare
  - Difficult for humans to sift through patient records
- In 2009, the company was analyzing 20 million people monthly

The Data

- Healthcare industry is data-rich, but data may be hard to access
  - Often unstructured and unavailable
- Used insurance data regarding procedures, prescriptions, and diagnoses
- Doctor insight regarding risk factors
  - Interactions between illnesses
- Demographic information (gender and age)

The Analytics

- Predict future healthcare costs
  - Identify high-risk patients to be prioritized for intervention
- Created interpretable models for doctors to analyze and verify
- Significantly improved over just using historical costs

What is the Edge?

- Substantial improvement in D2Hawkeye's ability to identify patients who need more attention
- Use expert knowledge to identify new variables and refine existing variables
- Can make predictions for millions of patients without manually reading patient files

The Rest of this Class

- In this class, we'll cover these examples and many more
- Each week will be composed of:
  - Two lectures
    - Each focused on a different real-world example
    - Teach an analytics method in the statistical software R
  - Recitation
    - Another example of the methodology
    - More practice in R
  - Homework assignment
    - Additional problems and datasets

Competition Week and Final Exam

- Midway through the class, we'll run an analytics competition
  - We'll challenge you to build a model and get the best accuracy possible
- At the end of the class, we'll test you on all of the methods used
  - The questions will be real-world problems

Our Goal

- This class should make you comfortable using analytics in your career and your life
- You will know how to work with real data, and will have learned many different methodologies
- We want to convince you of the Analytics Edge

15.071 Analytics Edge
Spring 2017