1 The Analytics Edge: Intelligence, Happiness, and Health

Welcome to 15.071x, The Analytics Edge. In this first lecture, we'll discuss how analytics redefines the meaning of intelligence, how it can contribute to your personal happiness and health. Data is transforming business, social interactions, and the future of our society. The amount of electronic data that exists in the world today is a phenomenal 2.7 zettabytes, which is equal to the storage required for more than 200 billion high-definition movies. Not only the amount of data is extremely high, but it increases at exponential rates. Our ability to effectively process this data is also increasing very rapidly. As an example, decoding the human genome originally took 10 years to process. Now it can be achieved in just one week.

Analytics is increasingly important in the world today, and this influence is expected to increase. McKinsey estimates that there is a shortage of 140,000 to 190,000 people with deep analytical skills to fill the demand of jobs in the United States by 2018. IBM has changed its business focus over the last 100 years very successfully from typewriters to mainframes to personal computers to consulting, and now to analytics. It 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. Analytics is becoming increasingly critical in almost every industry, from health care, to media, to sports, to finance, to government, and many others. Let us give a definition of analytics so we make it as concrete as possible. We define analytics to be the science of using data to build models that lead to better decisions, that add value to individuals, to companies, to institutions. Note that there are four ingredients– data, models, decisions, and value. And all four are needed in this definition.

What are the key messages of this class? First, analytics provides a competitive edge to individuals and companies. Analytics are often critical to the success of a company. And they provide often the decisive essential technology. Our teaching methodology is to teach you analytics techniques through real-world examples and real data. And our overarching goal is to convince you of the analytics edge and inspire you to use analytics in your career and your life. The teaching team comes from the Operations Research Center at MIT and the Sloan School of Management. I am Dimitris Bertsimas. I have received my Ph.D. from MIT from 1985 to 1988. And I have been with the MIT faculty at the Sloan School of Management since 1988. Currently, I'm the co-director of the Operations Research Center. My career is centered in analytics. And I believe that analytics can change the world. The other instructor of this class is Allison O'Hair. Allison received her Ph.D. from the Operations Research Center at MIT in 2013. Allison and I have worked together in the area of health care analytics, and are working at the moment with our colleague Bill Pulleyblank on an analytics textbook. The teaching assistants in the class are Iain Dunning, Angie King, Velibor Misic, John Silberholz, and Nataly Youssef, all Ph.D. students at the Operations Research Center at MIT.

To give you a sense of the breadth of applications in this class, we'll cover the story of IBM Watson, the computer that beat the best human players in Jeopardy!, the company eHarmony, the Framingham Heart Study, and D2Hawkeye, a company that I have been involved for almost a decade. Other examples include the story of Moneyball, how analytics can help predict the Supreme Court decisions, the role analytics have played in predicting the outcomes of the US presidential elections, how analytics can utilize effectively data from Twitter, Netflix, airline revenue management, radiation therapy, sports scheduling, and many others.

Our first example today is a story of IBM Watson. IBM research, a distinguished industrial laboratory, strives to push the limits of science. In the mid-1990s it created Deep Blue, a computer that beat Garry Kasparov, the world champion in Chess at the time, showing to the world that machines can beat humans in tasks that people thought were restricted to human intelligence. In the 1990s– in the late 1990s– it created Blue Gene, a computer to map the human genome. In 2005 IBM decided to create a computer that would compete at Jeopardy, a popular game show.

From the T.J. Watson Research Center in Yorktown Heights, New York, this is Jeopardy– The IBM Challenge. Please welcome back our contestants. He has never been defeated. And his winnings of more than $3.2 million make him Jeopardy's all time biggest money winner. From Los Angeles, California, here's Brad Rutter. An IBM computer system able to rapidly understand and analyze natural language, including puns, riddles, and complex questions across a broad range of knowledge, please welcome Watson. In 2004 he captivated America by winning 74 consecutive matches and $2.5 million on Jeopardy. From Seattle, Washington, here's Ken Jennings. And now here is the host of Jeopardy, Alex Trebek. Thank you, Johnny.

Before we get into the game, there are just a couple of things I need to tell you about this match. Now as I said earlier, Watson will receive the clues electronically as a text file at the same moment the clues are revealed to Ken and Brad. And at the same time I read them. This competition will be a two game total point exhibition match. However, these two games will be played out over the next three days so we can tell the full story. Throughout the games you'll get a glimpse of the thinking process, if you will, that is behind Watson's responses. Now this will be done through an answer panel display at the bottom of the screen. Let's play Jeopardy. Here we go.

Our first round of play contains these categories– Literary Character APB, All Points Bulletin; Beatles People; Olympic Oddities; Name the Decade; Final Frontiers; and Alternate Meanings. A little while ago we had a drawing to determine which player would select first. Brad, you won that. So if you're ready make your first choice. Let's take alternate meanings for $200, Alex. Four letter word for a vantage point or a belief. Brad? What is a view? Yes. Alternate meanings, $400. 4-letter word for the iron fitting on the hoof of a horse or a card-dealing box in a Casino. Watson? What is shoe? You are right. You get to pick. Literary Character APB for $800. Answer the Daily Double. Now Watson, although you have but $400, you know of course that you can risk up to the maximum value of a clue on the board. And that is $1,000. $1,000, please. All right. Here is the Daily Double clue for you. Wanted for killing Sir Danvers Carew. Appearance pale and dwarfish. Seems to have a split personality. Who is Hyde? Hyde, yes. Dr. Jekyll and Mr. Hyde, either one acceptable. You're now in the lead with $1400. Go again. Beatles People for $200. And any time you feel the pain, hey, this guy, refrain. Don't carry the world upon your shoulders. Watson? Who is Jude? Yes. Olympic Oddities for $200. Milorad Cavic almost upset this man's perfect 2008 Olympics, losing to him by one hundredth of a second. Watson? Who is Michael Phelps? Yes. Go. Name the Decade for $200. Disneyland opens and the peace symbol is created. Ken. What are the '50s? Yes. Final Frontiers for $1,000, Alex. Tickets aren't needed for this event, a black hole's boundary from which matter cannot escape. Watson? What is event horizon? Yes. Why is Jeopardy hard? Jeopardy asks the contestants to answer cryptic questions in a huge variety of categories. It is generally seen as a test of human intelligence, reasoning and cleverness. No links to the outside world are permitted. And new questions and categories are created for every show.

Watson is a supercomputer with 3,000 processors and a database of 200 million pages of information. It has a massive number of data sources like encyclopedias, texts, manuals, magazines, Wikipedia, etc. IBM researchers who developed Watson used over 100 different analytical techniques for analyzing natural language, finding candidate answers and selecting the final answer. We'll discuss this more later in the class. In February, 2011, a two game exhibition match aired on television six years after the initial conception of the idea to build Watson. 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 in Watson? Watson combined many algorithms to increase accuracy and confidence. We'll cover many of them in this class. IBM approached the problem in a different way than how a humans does it. Watson deals with massive amounts of data, often in unstructured form, which is important as 90% of the data in the world is unstructured. eHarmony is an online dating site focused on long term relationships. It takes a scientific approach to love and marriage. About nearly 4are a result of eHarmony. The company has generated over $1 billion in cumulative revenue from 2000, the year it was founded. The overall approach first predicts if users will be compatible using 29 different dimensions of personality and using linear regression. Then using optimization, eHarmony finds matches for everyone. eHarmony has members in more than 150 countries. Since launching in 2000, it has more than 33 million members, and it now operates eHarmony Labs, a relationship research facility.

eHarmony collects data through 436 questions. Almost 15,000 people take the questionnaire each day.

What is the edge of eHarmony? It relies much more on data than other dating sites. It 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. It represents 14% of the US online dating market.

Our third example is about the Framingham Heart Study. This study represents one of the most important studies of modern medicine. It is an ongoing study of the residents in Framingham, Massachusetts. It started in 1948, and is now on the third generation. Much of the now-common knowledge regarding heart disease came from this study. For example, the fact that high blood pressure should be treated. Clogged arteries are not normal. Cigarette smoking can lead to heart disease. Let us give some statistics about heart disease. Heart disease has been the leading cause of death worldwide since the 1920s. 7.3 million people died from coronary heart disease in 2008. Since 1950, age-adjusted death rates have declined 60In part, due to the results of the Framingham Heart Study. What is the data in this study? There were 5,209 patients enrolled in 1948. The patients were given a questionnaire and exams every two years, measuring their physical characteristics, their behavioral characteristics, and medical test results. The patient population, the exams, and the questions expanded over time. The approach the Framingham Heart Study utilized was a regression to predict whether or not a patient would develop heart disease in the next 10 years. The model tested and adjusted for different populations. The results of the study are available online so users can calculate their risk of heart disease based on total cholesterol, HDL, and systolic blood pressure.

So what is the edge? It provided necessary evidence for the development of drugs to lower blood pressure. The study further paved the way for other cliniical prediction rules that predict clinical outcomes using patient's data. Finally, the study demonstrated how a model allows a medical professional to make predictions for patients worldwide.

Our last example involves a company called D2Hawkeye, a medical software company founded in 2001. The company combined data with analytics to improve quality and cost management in health care. In 2009, the company was analyzing 20 million people monthly. It is impossible for humans to sift through patient records and assess quality and cost without algorithms.

This motivated the approach that D2Hawkeye took. Let us discuss the data that D2Hawkeye utilized. Health care industry is data rich, but data may be hard to assess. It is often unstructured and unavailable. The company used insurance data regarding procedures, prescriptions, and diagnosis. It further defined new risk factors based on doctor's insights. For example, obesity and depression. Finally, it used demographic information, like gender and age.

What were the analytics used? The goal was to predict future heath care costs, and identify high-risk patients to be prioritized for intervention. The company created interpretable models for doctors to analyze and verify. This led to significant improvements over just using historical costs.

So what is the edge?

The analytics used led to substantial improvement in D2Hawkeye's ability to identify patients who need more attention. The approach allowed expert knowledge to identify new variables and refine existing variables. Further, it allowed the ability to make predictions for millions of patients without manually reading patient's files.

In this class, we'll cover these examples and many more. Each week will be composed of two lectures, a recitation, and homework assignments. Each lecture is focused on a different real-world example. We will teach analytics methods in the statistical software R. In the recitations, we will present another example of the methodology and more practice in R. In a homework assignment, we'll provide additional problems and datasets. Midway through the class, we will 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. Let me comment on the overall goal of the class. This class aims to 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, and how to use them using R. In the end of the day, we want to convince you that analytics provides an edge to your career and your life.

2 Working with Data: An Introduction to R

3 Understanding Food: Nutritional Education with Data (Recitation)