40.016: Analytics Edge

Week 1 Lecture 1

Introduction

Term 6, 2020

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

The Analytics Edge

- LOCATION AND TIME:
  - VENUE: All lectures to be held via Zoom. Links available on e-dimension.
  - TIME: Section 1: Mondays, 9am – 11am and Thursdays, 11:30am – 1:30pm

    Section 2: Mondays, and Thursdays, 3pm – 5pm
- INSTRUCTORS:
  - WEEKS 1-6: Bikramjit Das, ✉ bikram@sutd.edu.sg
  - WEEKS 8-13: Stefano Galelli, ✉ stefano@sutd.edu.sg
- TEACHING ASSISTANTS:
  - Abhishek Pal Majumder ✉ abhishek_pal@sutd.edu.sg
  - Foo Lin Geng ✉ lingeng_foo@mymail.sutd.edu.sg
- Read the COURSE DESCRIPTION on eDimension.

Zoom etiquette

- Please login from a distraction free quiet environment.
- Use your name to login.
- Your audio is kept at mute by default.
- If you have to ask a question you may use the “Raise hand” feature and wait for us to ask you to unmute.
- You may use the chat window to ask a question too.
- Keep in mind that the chatbox and the classes will all be recorded and archived.
- If you have camera capabilities, it is expected that you will have your video on.
- Dress appropriately and maintain proper decorum.

Lectures and exercises

- Lectures: 2×2 hours weekly, partly theory, the rest will be R activity.
  -**Start on time**, with 10 minute break in the middle.
  - We will try to keep ample time at the end of the class so that you can ask questions.
- Problem sets (along with answers) will be uploaded regularly (once every 7 to 10 days).
  - These are self-assessments.
  - It is to your benefit to complete these exercises.

Consultation

- Wednesdays, 2pm – 4pm. Zoom office hours. Links to be posted.
- We will be using Piazza for class discussion. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza.
- Register: `https://piazza.com/sutd.edu.sg/fall2020/40016/home`

Course assessment

Mid-Term Test (Week 6)	25%
Competition (Week 13)	38%
Final Test (Week 14)	25%
Participation	10%
Course Feedback Completion	2%

- Participation score is based primarily on Piazza engagement, as well as participation during class and Learning Catalytics activity.
- The Competition is a week-long group activity. Further details will be discussed closer to the beginning of the class.

Exams

- **Mid-Term Test**: 23rd October, Friday, 2:30pm - 4:30pm.
- **Final Test**: 16th December, Wednesday, 9:00am - 11:00am.

Although exams are not cumulative, for the **Final Test** we assume that you will not have forgotten material from the first six weeks of class.

Topics to be covered: weeks 1 - 6

Week 1	Introduction to Analytics and the Software R with Visualization.
	Recall: Statistical tests, tools.
Week 2	Method: Linear Regression
	Predicting the quality and prices of wine (Wine analytics)
	Method: Principal Component Analysis
	Social progress analysis
Week 3	Method: Logistic Regression
	Predicting the failure of space shuttles (Challenger),
	Predicting the risk of coronary heart disease (Framingham Heart Study)
Week 4	Method: Multinomial Logit and Mixed Logit in Discrete Choice
	Predicting the Academy Award winners (Oscars)
	Estimating the preference for safety features in cars
Week 5	Methods: Big Data and Analytics: Model Selection
	Baseball (Sports)
	Cross-country growth regressions (Economics)
Week 6	Review and Test (23 October, Friday, 2:30 pm-4:30 pm)
Week 7	Break

Topics to be covered: weeks 8 - 13

Week 8	Method: Classification and Regression Trees (CART), Random Forests
	Forecasting Supreme Court Decisions (Law)
Week 9-10	Method: CART, Random Forests, Naïve Bayes Classifier
	Text Analytics: Twitter (Social media), Enron (Email)
	Ethics in Analytics
Week 11	Method: Clustering, Collaborative and Content Filtering
	Netflix, MovieLens (Recommendation systems)
Week 12	Method: Optimization
	Revenue Management, Capstone Allocation
Week 13	Review and Competition
Week 14	Test (16 December, Wednesday, 9:00 am - 11:00 am)

References and Software

References

- *The Analytics Edge* by Dimitris Bertsimas, Allison K. O'Hair and William R. Pulleyblank. Dynamic Ideas, Belmont, Massachusetts, 2016.

- *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2014.

  `https://link-springer-com.library.sutd.edu.sg:2443/book/10.1007/978-1-4614-7138-7`

Software

- R and R Studio.
- Julia.

Adobe Analytics Challenge 2020

adobeanalyticschallenge.com

What is the Adobe Analytics Challenge?

Adobe Analytics Challenge is a competition for university students that we have been running in North America for the last 15 years! The Adobe Analytics Challenge is a unique analytics-focused business case competition where university students are given the opportunity to use Adobe Analytics and access to real-world data from leading organizations such as Major League Baseball, T-Mobile, Starwood, Lenovo, Condé Nast, Comcast, Overstock.com, Backcountry.com, Sony PlayStation, and MGM Resort International.

What is great this year is that the competition is open to all countries globally including the countries in South East Asia and what makes it further more exciting is this year's data will be provided by one of our most elite Adobe Analytics customer globally - Nike.

Who can participate?

The competition is meant for students enrolled at a University for a full time academic program. The team can consist of 1-3 members from the same university along with a faculty member as a mentor.

How does one participate if they don't know how to work with Adobe Analytics or have access to it?

Every team will be given access to Adobe Analytics which will contain data from Nike. There will also a 2 hour training held by the Adobe Evangelists and the customer. The training will be recorded for on-demand viewing and will be available on the competition website.

What does the winner take home?

A lot of money! A total of **60,000 USD**!!

1st Place	2nd Place	3rd Place	4th Place	5th Place	6th Place
35,000 USD	14,000 USD	6,000 USD	3,000 USD	1,500 USD	500 USD

More information on the website https://adobeanalyticschallenge.com/

- `https://adobeanalyticschallenge.com/`
- Possible choice for faculty mentor: Professor Karthik Natarajan

What is Analytics?

- ... is the science of using DATA to build MODELS that add VALUE to DECISIONS made by individuals, companies and institutions.

> In God we trust, but all others must bring data. - W.E. Deming.

- ... viewed as critical to the success of individuals, companies and institutions.
- Niometrics ... 10 Petabytes of data processing a day (2019).

> We are drowning in information, while starving for wisdom - E.O. Wilson/ Rutherford Roger

- We will see several such examples in this course.

source: The Analytics Edge, Bertsimas, O'Hair and Pulleyblank

Three kinds of Analytics

1. DESCRIPTIVE ANALYTICS
   - what happened?
   - past data, data aggregation, visualization.
2. PREDICTIVE ANALYTICS
   - what could happen?
   - delve into why things happened.
   - build statistical models and forecasting techniques.
   - a big portion of our lectures will be dealing with this.
3. PRESCRIPTIVE ANALYTICS
   - what should we do?
   - optimization/simulation ... scenarios, etc.
   - algorithms, machine intelligence, AI: supply chain logistics/ scheduling.
   - we will get to see a little bit of this.

- `https://www.youtube.com/watch?v=nv2HTRWhbB4`
- `https://www.youtube.com/watch?v=m30LxzzbRik`

source: Competing on Analytics, Davenport and Harris

Challenges in such analytics

- Real world problems are often complex and ill-posed.
- Data is available but ...
  - only objective reality perhaps
  - incomplete, poor quality
- No one tells you what to use – regression, classification, optimization, ...
- We build models ... models are just facilitators.

> All models are wrong, but some are useful. - George Box, 1976.

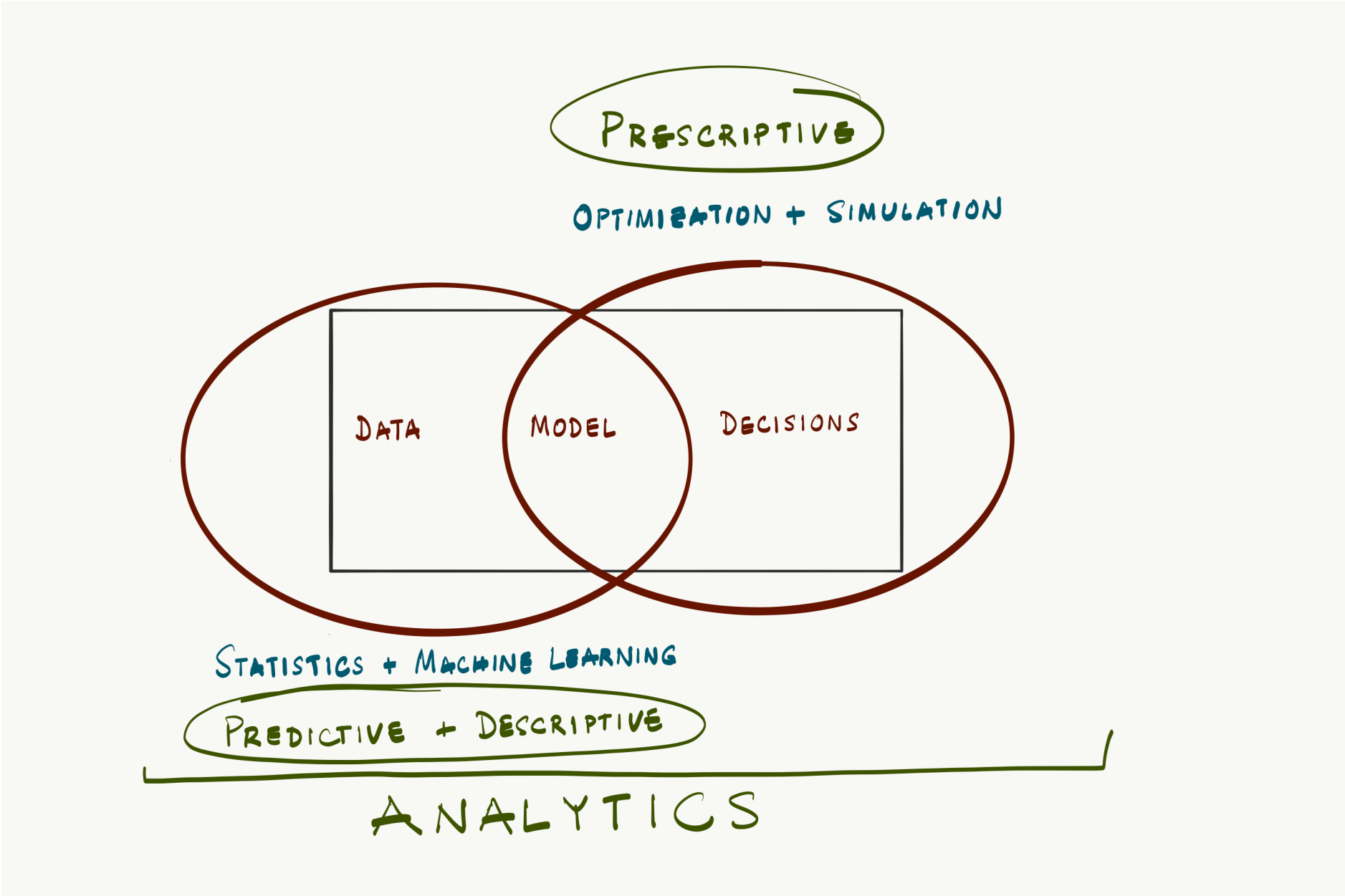

Figure: The Analytics View

Our plan

- We work with one/two data sets each week and address problems there.
- Big data – large complex data sets — often unstructured and we want to get useful insights from such data
- We will work with various sizes of data sets.
- Now predictive analytics have become very personalised too:
  - what treatement plan must a patient be prescribed?
  - price offered to a customer
  - marketing strategy affecting different individuals
- Data ethics

Netflix documentary: The Social Dilemma

Jeopardy! and Watson

- Popular US TV show launched in 1964.
- Questions presented in an answer form:
  - This number, one of the first 20, uses only one vowel (4 times!).
  - Sakura cheese from Hokkaido is a soft cheese flavored with leaves from this fruit tree.
- On February 14, 2011, IBM's computer Watson participated against the best human players.
- Winning required both accuracy and precision (identified by IBM team)
- Use 'ensemble' methods

Jeopardy!

- https://www.youtube.com/watch?v=P18EdAKuC1U
- http://www.youtube.com/watch?v=C5Xnxjq63Zg

- The IBM debater: https://youtu.be/7pHaNMdWGsk?t=965

Why R?

1. R is free and open.
2. R provides an integrated suite of software facilities for data manipulation, calculation and graphical display.
3. R provides an environment within which many statistical techniques have been implemented and these functionalities can be extended by adding new packages as needed. It is also possible to develop packages with new statistical methods for others to use.
4. R has extensive online support and discussion forum.

- Learn R (R for Data Science): `https://r4ds.had.co.nz`