40.016: The Analytics Edge

Week 1 Lecture 1

Introduction

Term 5, 2022

The Analytics Edge

- TIME: Mondays and Wednesdays (as scheduled)
- VENUE: Room 2.606 (CC 15)
- INSTRUCTORS:
  - WEEKS 1-6: Stefano Galelli, ✉ stefano_galelli@sutd.edu.sg
    Consultation: by appointment
  - WEEKS 8-13: Lin Meixia, ✉ meixia.lin.math@gmail.com
    Consultation: by appointment
- TEACHING ASSISTANTS:
  - Anand Deo. ✉ deo_avinash@sutd.edu.sg
    Consultation: Wednesday 2:00 to 4:00 pm
  - Benjamin Tan. ✉ benjamin_tanweijun@sutd.edu.sg
    Consultation: Fridays 2:00 to 4:00 pm
- Read the COURSE DESCRIPTION on eDimension.

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.

Course assessment

Mid-Term Test (Week 6)	35%
Competition (Week 13)	28%
Final Test (Week 14)	35%
Course Feedback Completion	2%

- The Competition is a week-long group activity. Further details will be discussed closer to Week 13.

Exams

- **Mid-Term Test**: 22nd June, Wednesday, 2:30 pm - 4:30 pm.
- **Final Test**: 19th August, Friday, 3:00 pm - 5:00 pm.

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 (22 June, Wednesday, 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: Bagging, 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 (19 August, Friday, 3:00 pm - 5:00 pm)

References

- *The Analytics Edge* by Dimitris Bertsimas, Allison K. O'Hair and William R. Pulleyblank. Dynamic Ideas, Belmont, Massachusetts, 2016.
  `https://sutd.primo.exlibrisgroup.com/permalink/65SUTD_INST/19hmrhl/alma999432964802406`
- *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`
- *Calling Bullshit: the art of skepticism in a data-driven world* by Carl T. Bergstrom and Jevin D. West. Random House, 2020.
  `https://sutd.primo.exlibrisgroup.com/permalink/65SUTD_INST/19hmrhl/alma999596764102406`

Software

- R and R Studio.
- Julia.

What is Analytics?

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

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

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.

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

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.

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

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).

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

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

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

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

source: Competing on Analytics, Davenport and Harris

Three kinds of Analytics

1. DESCRIPTIVE ANALYTICS
   - what happened?
   - past data, data aggregation, visualization.

source: Competing on Analytics, Davenport and Harris

Three kinds of Analytics

1. DESCRIPTIVE ANALYTICS
   - what happened?
   - past data, data aggregation, visualization.
2. PREDICTIVE ANALYTICS

source: Competing on Analytics, Davenport and Harris

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.

source: Competing on Analytics, Davenport and Harris

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

source: Competing on Analytics, Davenport and Harris

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.

source: Competing on Analytics, Davenport and Harris

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

Challenges in such analytics

- Real world problems are often complex and ill-posed.

Challenges in such analytics

- Real world problems are often complex and ill-posed.
- Data is available but ...

Challenges in such analytics

- Real world problems are often complex and ill-posed.
- Data is available but ...
  - only objective reality perhaps

Challenges in such analytics

- Real world problems are often complex and ill-posed.
- Data is available but ...
  - only objective reality perhaps
  - incomplete, poor quality

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, ...

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.

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.

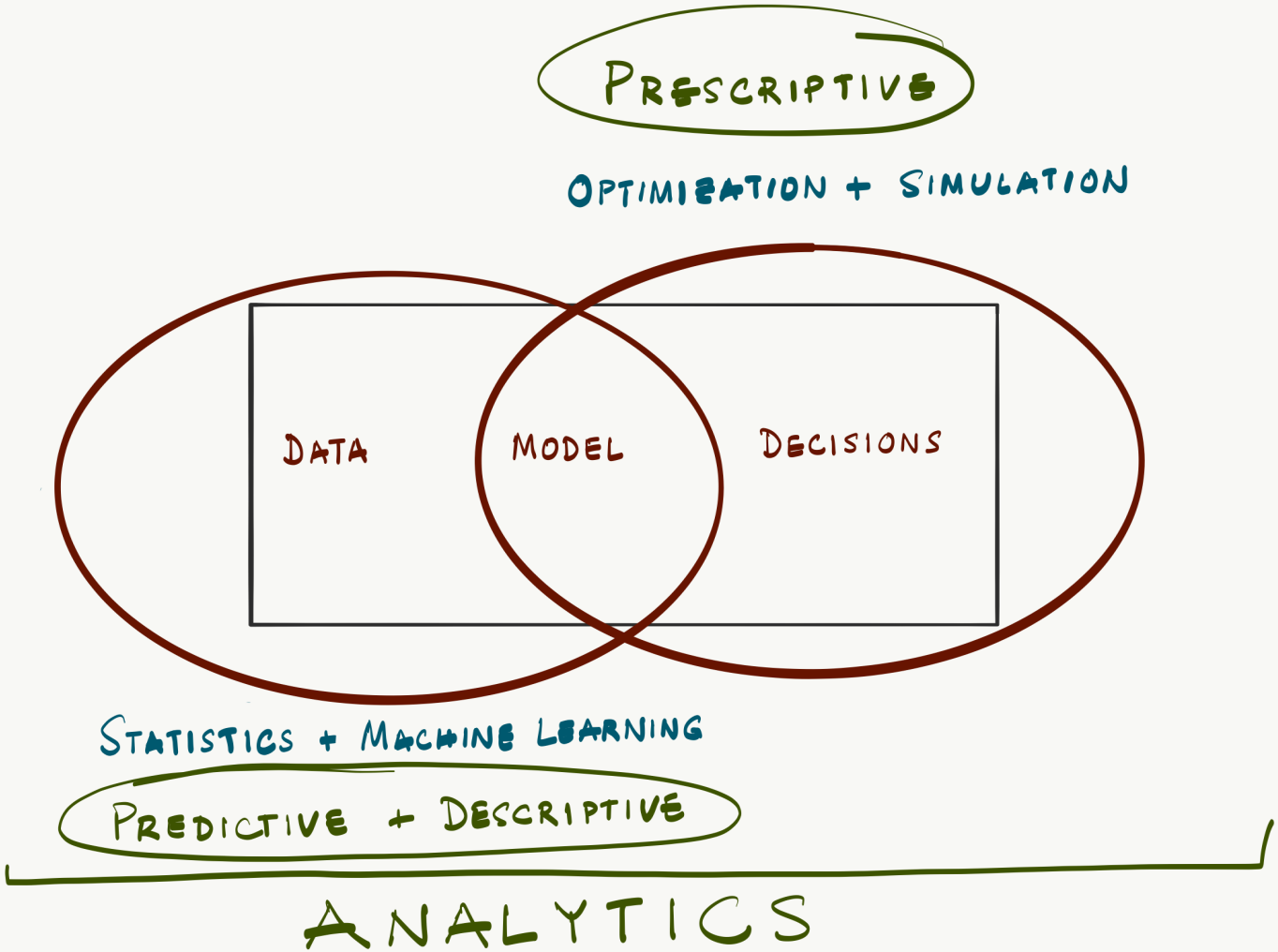

Figure: The Analytics View

Our plan

Our plan

- We work with one/two data sets each week and address problems there.

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

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.

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

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

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.

Jeopardy! and Watson

- Popular US TV show launched in 1964.
- Questions presented in an answer form:

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!).

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.

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.

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).

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

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`