Rensselaer

why not change the world?®

Introduction to Data Analytics - Current Challenges - Course Outline

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 BCBP- 4960/MGM-4962/MGMT-6962

Group 1 Module 1, August 30th, 2024

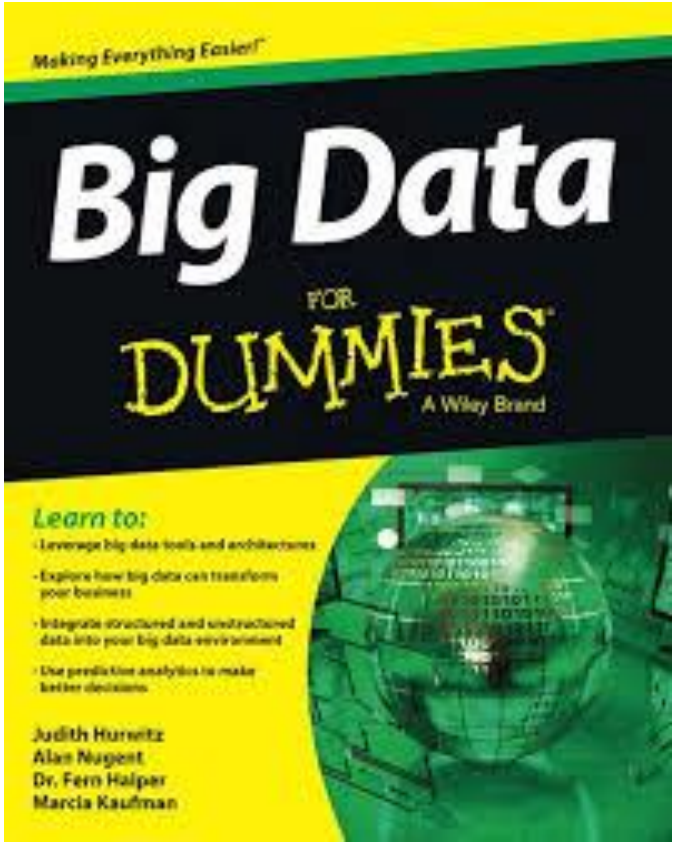
Tetherless World Constellation
Rensselaer Polytechnic Institute

Admin information

- Class: ITWS-4600/ 6600/MATP-4450/CSCI-4960/MGMT 4962/6962/BCBP 4960
- Hours:
  Section 1: 10:00am ET - 11:50am ET on Tues/Fri – Troy 2018
  Section 2: 02:00pm ET - 03:50pm ET on Tues/Fri – Troy 2018
- Instructor: Ahmed Eleish
- Instructor contact: eleisa2@rpi.edu
- Instructor office hours: Wed from 01:30 PM - 3:30 PM ET/ Thursday from 02:30 PM - 4:00 PM ET or by appointment/email
- Instructor office location: Amos Eaton 134
- TA: Alyssa Bigness - bignea2@rpi.edu
- TA office hours: TBA
- TA office: Lally 205
- Web site: https://tw.rpi.edu/classes/data-analytics-fall-2024
- LMS (http://lms.rpi.edu/)

Contents

- Intro – about this course
- Learning objectives
- What is expected
- What skills are needed
- Outline of the course
- Definitions and why Analytics means more than Analysis



Assessment and Assignments

• Via written assignments with specific percentage of grade allocation provided with each assignment

• Via individual oral presentations with specific percentage of grade allocation provided

• Via participation and attendance in lectures and labs (not to exceed 5% of total, start with 5% and lose % by not participating and attending labs)

• Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day. Talk to me EARLY if you are having schedule problems in completing assignments

Assessment and Assignments

- Reading assignments Are given when needed to support key topics or to complete assignments

- Will not be discussed in class unless there are questions

- You will mostly perform individual work that is assessed but you are encouraged to work with others in the lab sessions (except assignment 2)

Project options (examples)

- Social networks
- Financial
- Social-economic, marketing
- Geo/ space science data
- Network/ security data
- Linked data
- Movie databases
- Transportation
- Competitions (Web and local)*
- Research Projects*

Research Projects & Competitions* : Need the Instructor's approval for the datasets

Objectives

• Introduce students to relevant methods to recognize and apply quantitative algorithms, techniques and interpretation.

• To develop students' strategic thinking skills, combined with a solid technical foundation in data and model-driven decision-making.

• Develop ability to apply critical and analytical methods to formulate and solve science, engineering, medical, and business problems.

Objectives

• Students will examine real-world examples to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science.

• Making decision under uncertainty, how to optimize models, sequential decision making, weak models, mixed models.

• By the end of the course, students can effectively communicate analytic findings to non-specialists.

Learning Objectives

• Through class lectures, practical sessions, written and oral presentation assignments and projects, students should:

– demonstrate knowledge of relevant analytic methods, and to recognize and apply quantitative algorithms, techniques and interpret results.

– demonstrate strategic thinking skills, combined with a solid technical foundation in data and model-driven decision-making.

– Students to develop ability to apply critical and analytical methods to formulate and solve science, engineering, medical, and business problems.

Learning Objectives

– examine real-world examples to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science.

– effectively communicate analytic findings to non-specialists.

– [6600 level] Students must develop and demonstrate a working knowledge of decision making under uncertainty, be able to optimize models that incorporate random parameters.

4450/ 4600/ 4960 versus 6600

- 6600 students are assessed at:
  - Higher level of demonstration
  - Additional questions or tasks in assignments

- 4450/4400/4960 students are welcome to complete these higher requirements for extra credit

- Extra points for outstanding/ above and beyond are given*

Academic Integrity

• Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these. In this class, all assignments that are turned in for a grade must represent the student's own work. In cases where help was received, or teamwork was allowed, a notation on the assignment should indicate your collaboration.

• Submission of any assignment that is in violation of this policy will result in a penalty. If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty of full loss of grade for the work in violation, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: warning, probation, suspension, expulsion, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities.

• Second violation will result in failure of the course.

• **If you have any question concerning this policy before submitting an assignment, please ask for clarification.**

What is expected

- Attend class, complete assignments, participate
- Ask questions, offer answers in class
- Work individually on assignments
- Work in a group on labs, learn from each other, help each other especially with software
- Work constructively in class labs

Skills needed

- Basic knowledge of data structures, computer programming
- Literacy with computers and applications that can handle the data we will use
- Ability to access internet, servers and retrieve/ acquire data, **install/ configure software**
- **Pick up R programming**, terminology and syntax, and some refinement
- Presentation of proposal projects and assignment results

Current Syllabus/Schedule

- Web site: https://tw.rpi.edu/classes/data-analytics-fall-2024

- Note: in general lectures are on Tuesdays, labs on Fridays

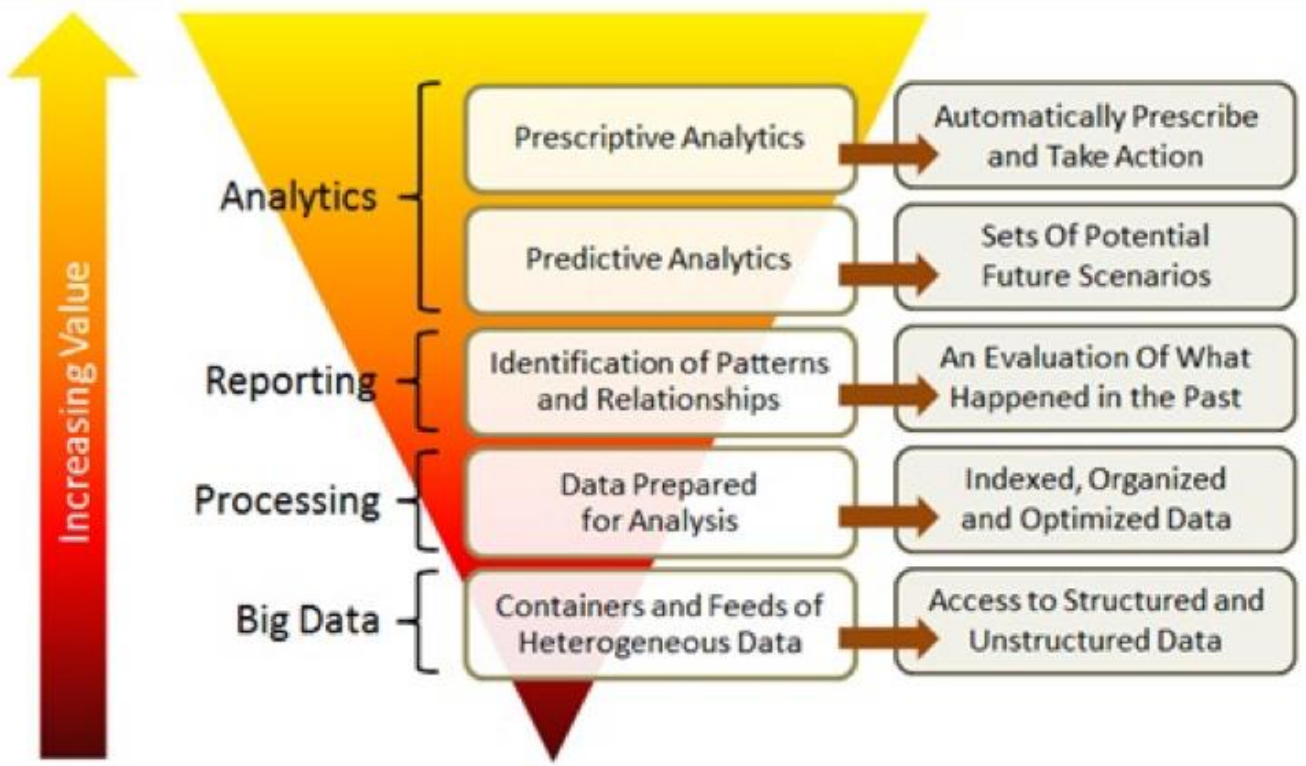
- Attendance is taken lectures and labs; lab attendance is part of participation grades

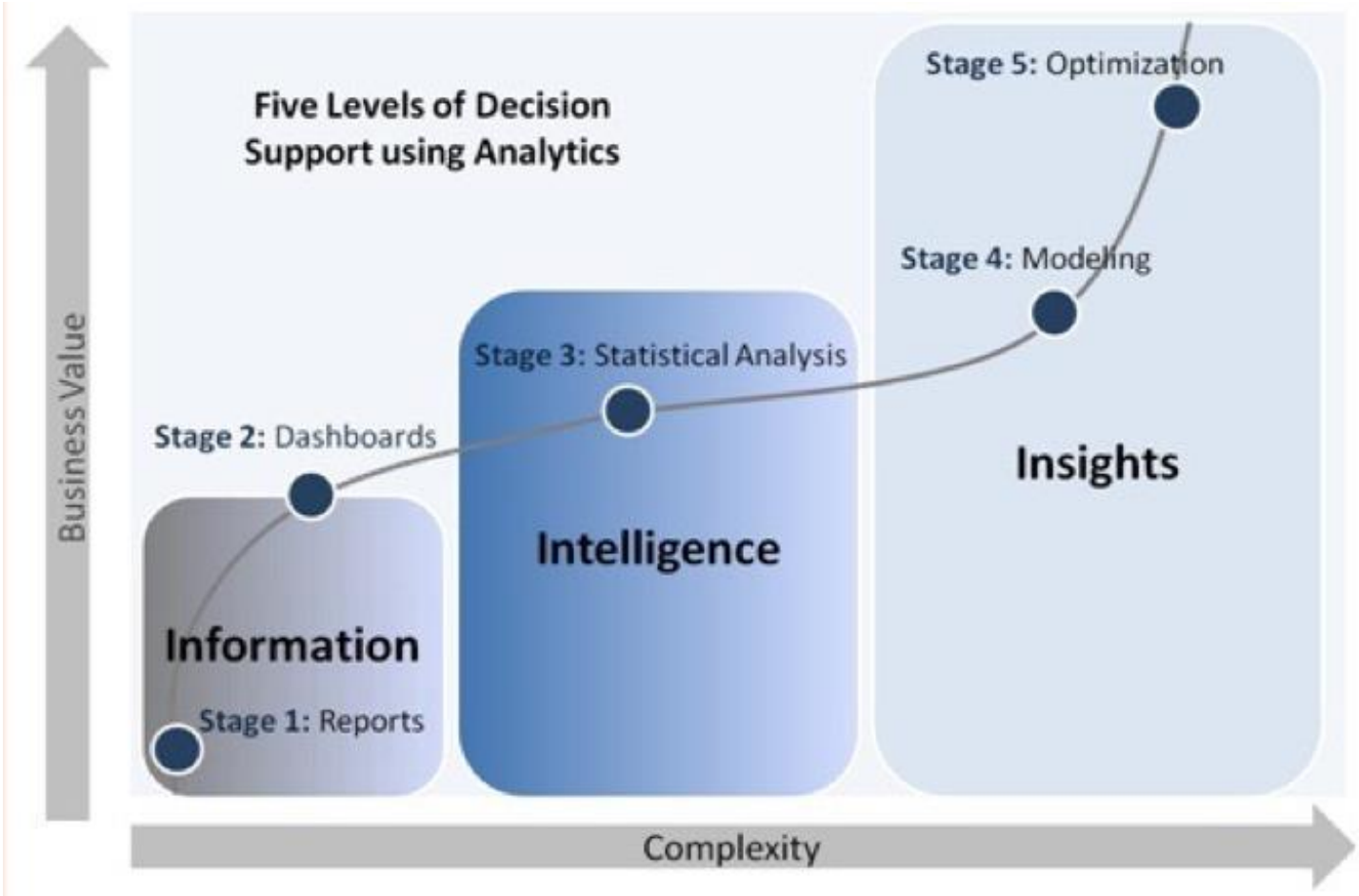
Questions so far?

Introductions

- Who you are
- Why you are here
- What you expect to learn
- Your interests/hobbies

The nature of the challenge




Five Levels of Decision Support using Analytics
Business Value
Complexity
Information
Intelligence
Insights
Stage 1: Reports
Stage 2: Dashboards
Stage 3: Statistical Analysis
Stage 4: Modeling
Stage 5: Optimization

Perspective

• People make decisions every day and increasingly they are using resources/ services (that run on computers) to assist or decide for them.

• Knowledge can translate to “power”:
– Or accurate/ reliable knowledge is actionable

• Gaining knowledge and how to use that knowledge - from (often multiple sources) information and data sources

• A model = formula/ equation that could depend on parameters and variables

So what are we talking about?

Definitions (at least for this course)

• Data - are encodings that represent the qualitative or quantitative attributes of a variable or set of variables.

• Data (plural of "datum", which is seldom used) - are typically the results of measurements, computations, or observations and can be the basis of graphs, images of a set of variables.

• Data - are *often* viewed as the lowest level of abstraction from which information and knowledge are derived

And then there is Big Data

5 V's: volume, variety, veracity, velocity, value
https://en.wikipedia.org/wiki/Bi g_data

Journals/ conferences: IEEE, http://www.liebertpub.com/big

In short: crawl before you walk, before you run, before you become famous ;-)

A view from IBM ...

• “Anyone who wants to learn something about data analytics should take a road trip. Myriad real-time decisions must be made based on analysis of static information as well as ever-changing conditions. Data about traffic, weather, road construction, fuel, time, current location and available funds are just a few of the factors.”

• This information and much more are needed to answer questions like:

- If I skip this gas station, will I run out of gas before the next one?
- Is it worth driving 50 miles out of the way to see the Corn Palace? How late will that side trip make us?
- Can I make it to Billings, Montana., by sunset or should I look for a place to stop?

Case Studies

• Sports Analytics–Moneyball (http://www.imdb.com/title/tt1210166/), Nate Silver (http://en.wikipedia.org/wiki/Nate_Silver)

• Marketing Analytics – Amazon, Walmart/Target, etc.

• Netflix, etc. recommenders – “If you liked,...”

http://www.slideshare.net/lsakoda/case-studies-utilizing-real-time-data-analytics

Analysis

- Software packages / environments:
  - Gnu R
  - Rstudio
    - Extensive libraries
  - <Jupyter Notebook/ Lab>

- Going from preliminary to initial analysis...

- Parametric (assumes or asserts a probability distribution) and non-parametric statistics

What are "statistics"?

• The term "statistics" has **two common meanings**, which we want to clearly separate: **descriptive** and **inferential** statistics.

• But to understand the difference between descriptive and inferential statistics, we must first be clear on the difference between **populations** and **samples**.

• See Module 2 (during this course)

Summary

• We'll work our way through the stages of analytics

• We'll use both laptop installed software and potentially some server data infrastructures for analytics to give you practical experience

• We'll cover algorithms, parameter choices, models, results, interpretation, and the software

* **This is a fast-paced course *

Current assignment structure (no final exam)

- Assignment 1: Review of a DA Case Study. End of week 2 - 5% (written/ discuss)
- Assignment 2: Datasets and data infrastructures – graded lab assignment. In ~ week 3 - 10% (in lab)
- Assignment 3: Preliminary and Statistical Analysis. In ~ week 4 - 15% (written)
- Assignment 4: Patterns, trends, relations: model development and evaluation. In ~ week 6 - 15% (written)
- Assignment 5: Term project proposal. In ~ week 7 - 5% (oral/written)
- Assignment 6: Term project. In ~ week 13 - 30% (25% written, 5% presentation-oral/poster)
- Assignment 7: Predictive and Prescriptive Analytics. Due ~ week 10 - 15% (15% written)
- 5% participation

Reading/ watching

- SportsAnalytics–Moneyball
- (http://www.imdb.com/title/tt1210166/),
- NateSilver(http://en.wikipedia.org/wiki/Nate_Silver)
- http://www.slideshare.net/lsakoda/case-studies- utilizing-real-time-data-analytics
- http://www.marketquotient.com/case-studies.html
- http://www.ibm.com/analytics/us/en/case-studies/

- More in the Assignment …

Reference Material

- On course website – some via RPI Library, RCS login required

Files

- We will use Box, please make sure your accounts are set up
- Link will be shared on course website

Assignment 1:

• Choose a Data Analytics case study from a) assignment readings, or b) your choice (must be approved by the instructor)

• Read it and provide a short written review/critique of the case study (is there a solid business case, what is the area of application, what approach/ methods, tools were taken/used, what were the results, actions, benefits?). Hand in a written report.

• Be prepared to discuss it in the class/lab.

• Details on course website (under Week 1)

• **Due: September 6th 2024 by 08:00pm ET by email (eleisa2@epi.du)**

Next class: Friday Sep 6th – quick refresher on statistics and Intro to Labs

Remember: Tuesday Sep 3rd follows MONDAY schedule, i.e. no Data Analytics class on that day!

Head start for lab - R

- http://lib.stat.cmu.edu/R/CRAN/ - install this first
- http://cran.r-project.org/doc/manuals/
- http://cran.r-project.org/doc/manuals/R-lang.html
- R Studio

(https://www.rstudio.com/products/rstudio/)
(desktop version)

Thanks!
(Have a great weekend)

*** Work on the assignment!