Introduction to Data Analytics. Current Challenges. Course Outline.

Thilanka Munasinghe

Data Analytics

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

Group 1 Module 1, September 1st , 2020

Admin information

- Class: ITWS-4600/ 6600/MATP-4450/CSCI-4960/MGMT 4962/6962/BCBP 4960
- Dates: Tuesdays/ Fridays
- Location: **Section1**:Location: SAGE 2715 - Time: 10:10am -12:00pm and WebEx for Remote (Online) Participation. WebEx login information available on Learning Management System (LMS)
  Section2:Location: LALLY 02 – Time: 2:30pm – 4:20pm and WebEx for Remote (Online) Participation. WebEx login information available on Learning Management System (LMS)
- Instructor: Thilanka Munasinghe
- Instructor contact: munast@rpi.edu
- **Instructor Office Hours**: Tue/Fri 12:30pm – 1:30pm or by appointment via email OR online via WebEx (Instructor online office hours loin information available on LMS
- Instructor office location: Room 133, Amos Eaton building
- **Teaching Assistant : Mia Price**
- **TA Office Location: Virtual (WebEx)** https://rensselaer.webex.com/meet/pricem4
- **TA Office Hours: 12 pm – 2 pm on Wednesdays OR by Appointment (via WebEx)**
- **TA Email Address: pricem4@rpi.edu**

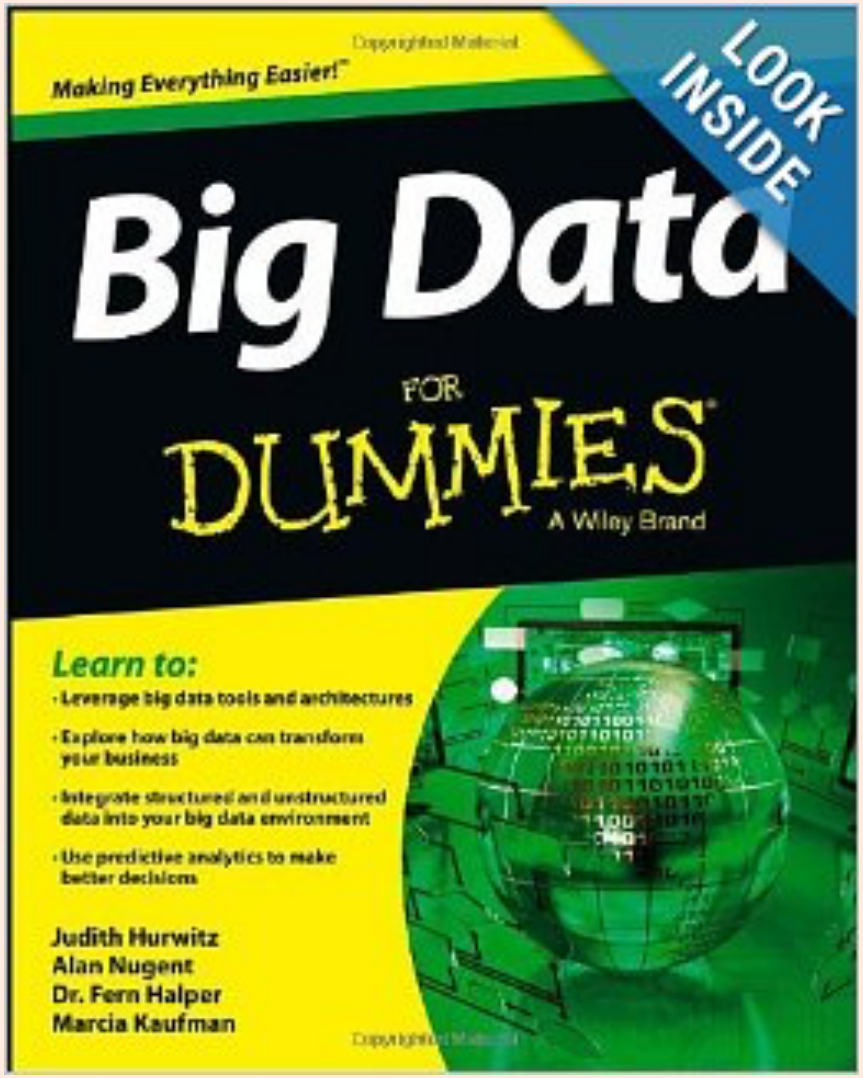
Admin information

- Class: ITWS-4600/ 6600/MATP-4450/CSCI-4960/MGMT 4962/6962/BCBP 4960
- Dates: Tuesdays/ Fridays
- Location: **Section1**:Location: SAGE 2715 - Time: 10:10am -12:00pm and WebEx for Remote (Online) Participation. WebEx login information available on Learning Management System (LMS)
  Section2:Location: LALLY 02 – Time: 2:30pm – 4:20pm and WebEx for Remote (Online) Participation. WebEx login information available on Learning Management System (LMS)

- Web site & LMS = http://lms.rpi.edu and see also Course Webpage: https://tw.rpi.edu/web/Courses/DataAnalytics/2020Fall
- Schedule, lectures, syllabus, reading, assignments, etc.
  - **Location of data and R script fragments, other documents - http://aquarius.tw.rpi.edu/html/DA/**

Contents

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



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 in lectures and labs (not to exceed 5% of total, start with 5% and lose % by not participating)
- 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
- Network/ security data
- Linked data
- Competitions (Web and local)****
- Movie databases
- Transportation
- 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
- In groups, students will identify qualitative problems and apply content analytics
- 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:
  - Students to demonstrate knowledge of relevant analytic methods, and to recognize and apply quantitative algorithms, techniques and interpret results
  - Students to 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
  - 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.
  - Students must 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: static stochastic optimization, two-stage optimization with recourse, chance-constrained optimization, and sequential decision making.

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 an extra grade
- 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.**

Current Syllabus/Schedule

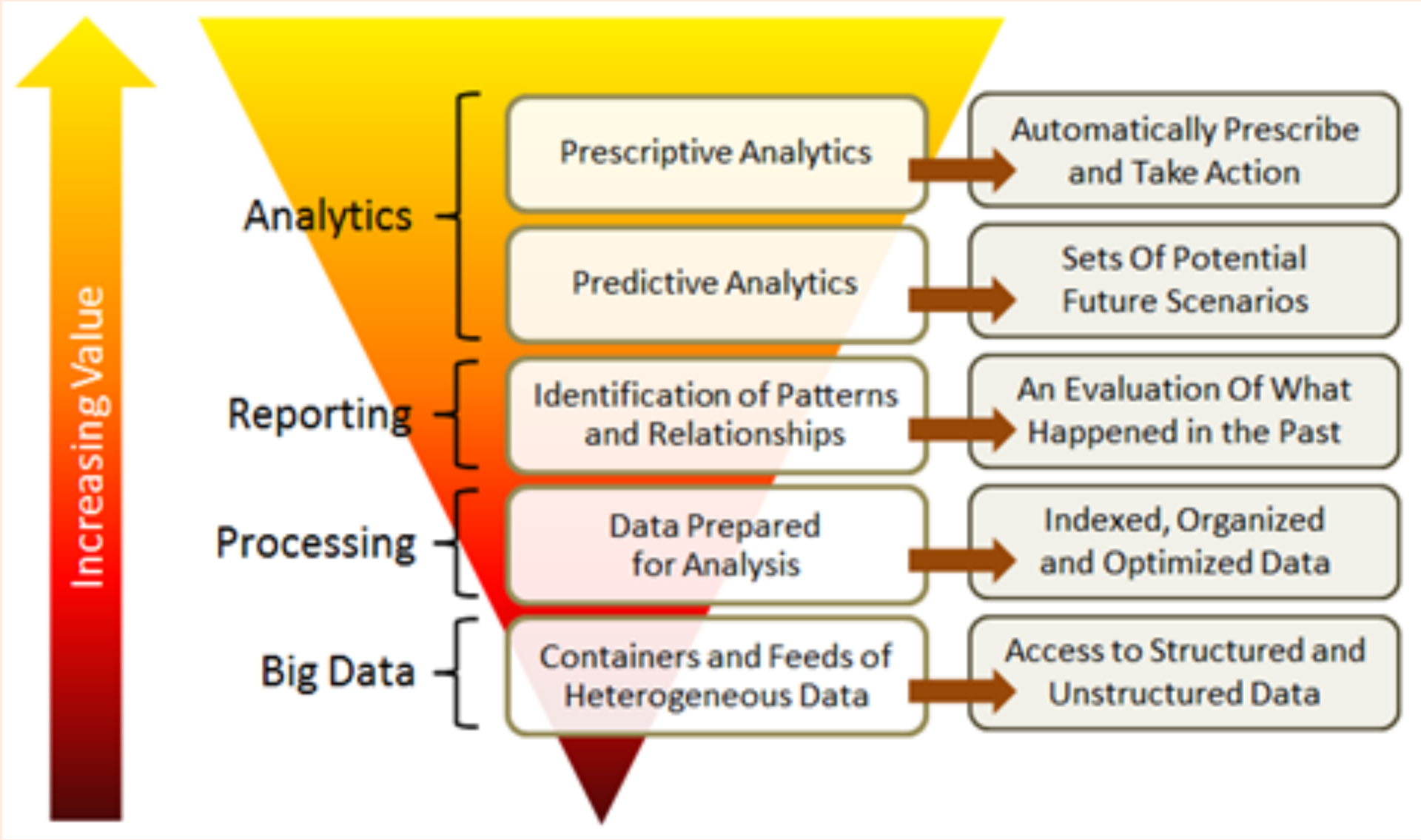
- Web site: https://tw.rpi.edu/web/Courses/DataAnalytics/2020Fall
- Note: in general lectures are on Tuesday, labs on Fridays ***
- Attendance is taken – the labs are for you to work through tasks that prepare for assignments

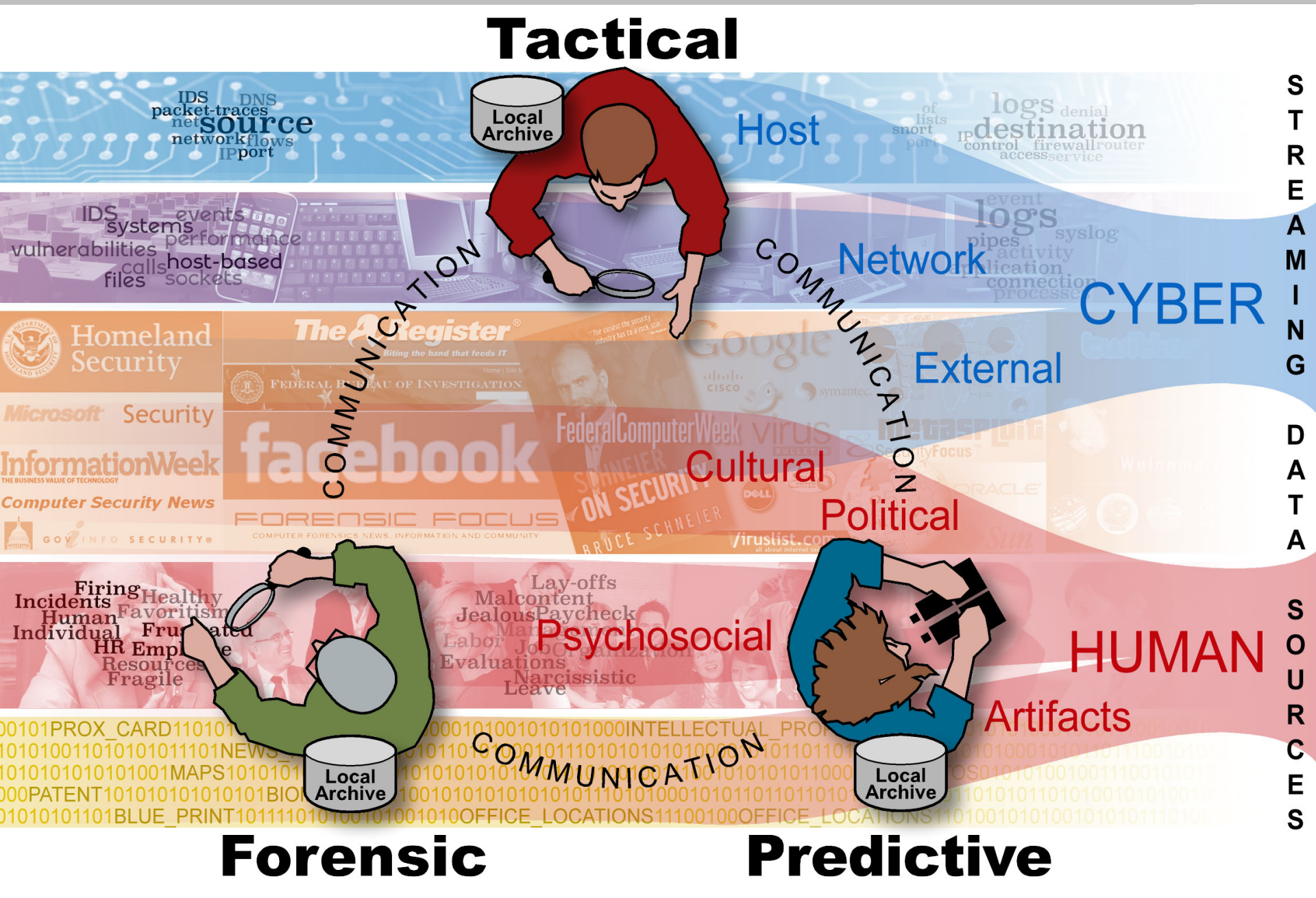
Questions so far?

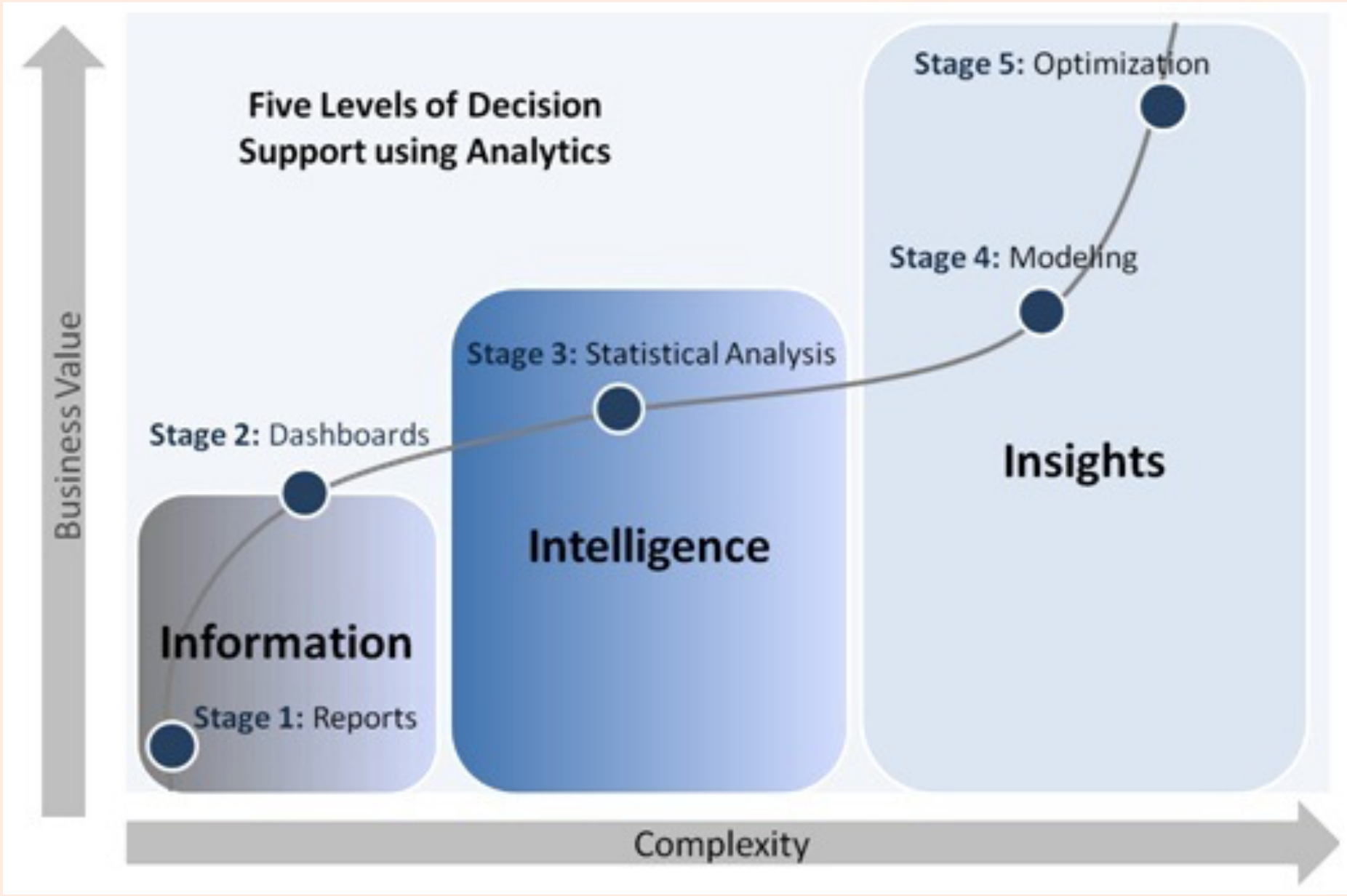
Introductions

- MATP-4450, CSCI 4960, ITWS-4400, ITWS-6600, MGMT-4962, MGMT-6962
- Who you are, background (data science, database, programming)?
- Why you are here?
- What you expect to learn?
- Your interests/hobbies ?

The nature of the challenge




Tactical
Local Archive
Host
Network
CYBER
External
Cultural
Political
Psychosocial
HUMAN
Artifacts
COMMUNICATION
COMMUNICATION
COMMUNICATION
Local Archive
Local Archive
STREAMING DATA SOURCES
Forensic
Predictive


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

Perspective

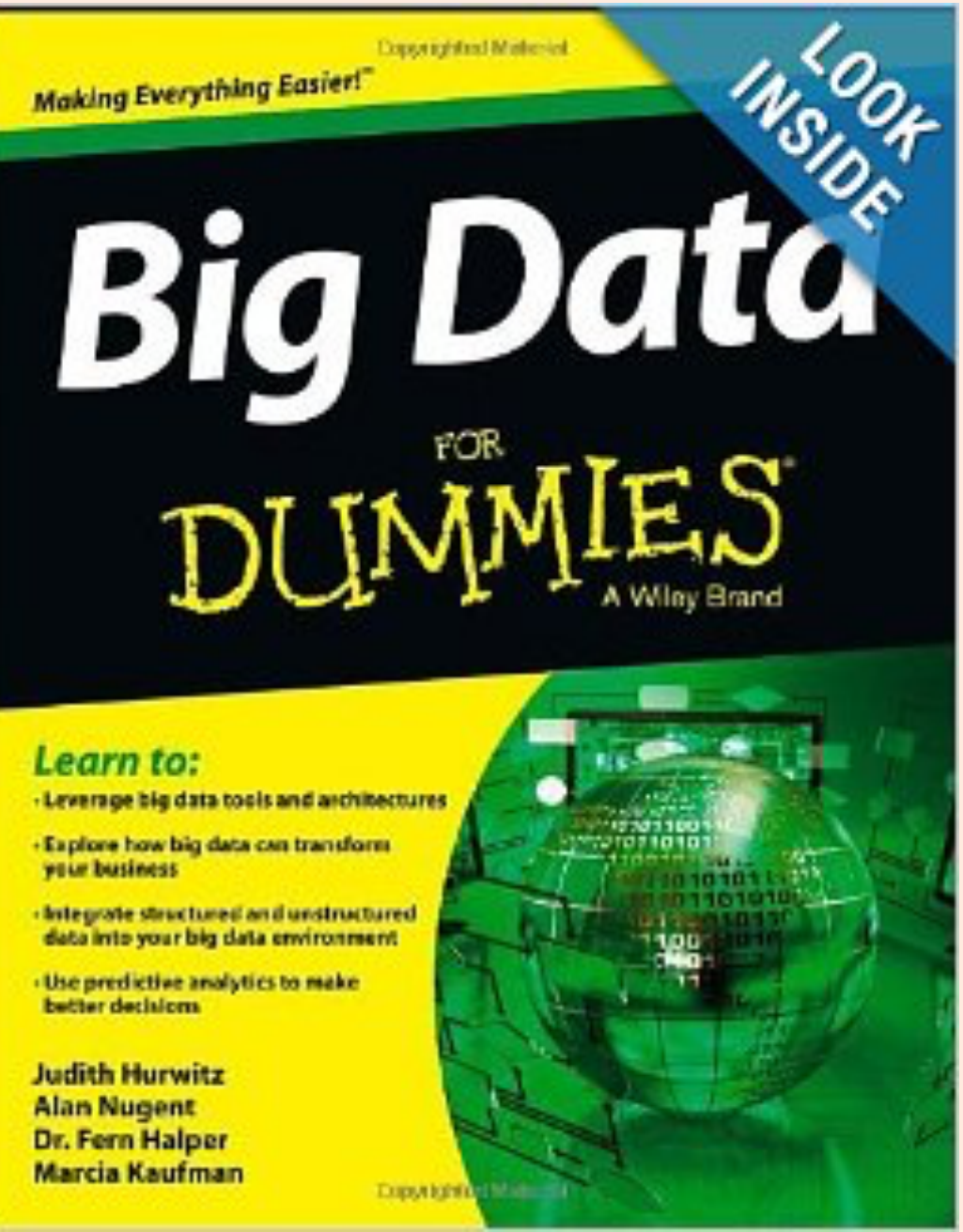
- People make decisions every day and increasingly they are using resources/ services (that run on computers) to assist or decide for them.
- Knowledge can translate to “power”:
  - Or accurate/ reliable knowledge is actionable
- Gaining knowledge and how to use that knowledge - from (often multiple 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/Big_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 (warming up)

- Sports Analytics – Moneyball (http://www.imdb.com/title/tt1210166/), Nate Silver (http://en.wikipedia.org/wiki/Nate_Silver)
- Marketing Analytics – products for pregnant (women)
- Amazon Recommender – "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 Notebooks, Jupyter Labs>
- Going from preliminary to initial analysis…
- Parametric (assumes or asserts a probability distribution) and non-parametric statistics

What is "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)

Courtesy Marshall Ma (and prior sources)

Summary

- We'll work our way through the stages of analytics
- We'll use current 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**…

Skills needed

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

Current assignment structure (no 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: Pattern, trend, relations: model development and evaluation. In ~ week 6. 15% (written);
- Assignment 5: Term project proposal. In ~ week 7. 5% (oral or 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

What is expected

- Attend class, complete assignments, participate
- Ask questions, offer answers in class
- Work individually on assignments
- In a group (in-person classroom OR virtual/online group), learn from each other, help each other especially with software
- Work constructively in class labs

Reading/ watching

- Sports Analytics – Moneyball (http://www.imdb.com/title/tt1210166/),
- Nate Silver (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 LMS and website – some via RPI Library, RCS login required
- Data Analytics – various intro material
- Using R (and next week)

Files

- http://aquarius.tw.rpi.edu/html/DA/

- This is where the files for assignments, exercise will be placed – data, code (fragments), and other documents, etc.

Assignment 1

- Choose a Data Analytics case study from a) assignment readings, or b) your choice (must be approved by me)
- 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/ submission on LMS/ Web site (under Assignments; Week 1)

Next?

- Week 1 – Friday - quick refresher on statistics and Intro to Labs
- No classes on Tuesday, September 8th 2020 (Tuesday follows the Monday schedule)

Introductions

- Who you are, background?
- Why you are here?
- What you expect to learn?

Head start for lab - R

- http://lib.stat.cmu.edu/R/CRAN/ - load this first
- http://cran.r-project.org/doc/manuals/
- http://cran.r-project.org/doc/manuals/R-lang.html
- R Studio – (see R-intro.html too) https://www.rstudio.com/products/rstudio/ (desktop version)