Tools & Models for Data Science

Course Overview

Risa Myers & Chris Jermaine

Rice University

Welcome!

Please fill out the questionnaire

Welcome!

- Introductions
- Course overview
- Syllabus / logistics
- Tools

Introductions!

I am

- Risa Myers
- Assistant Teaching Professor
- rbm2@rice.edu
- Duncan Hall 2062

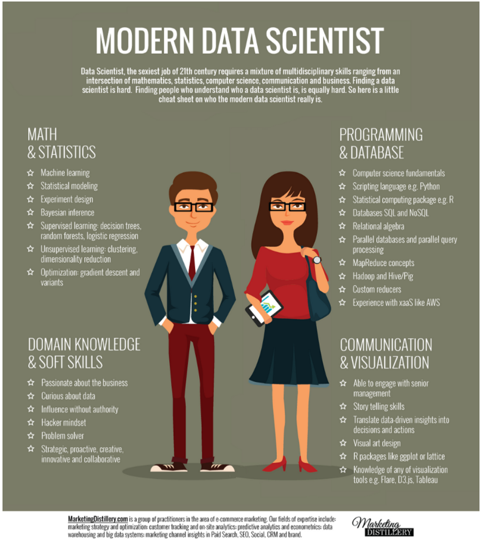
Who Are You?

Trade questionnaires!

This Class is about *Data Science*

- What is THAT?
- Extraction of actionable knowledge from large volumes of data
    - Encompasses methods from:
        - Computer science
        - Statistics
        - Optimization/Applied Math
    - Also includes
        - Domain knowledge
        - Communication skills
        - Data management

Modern Data Scientists



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

Examples of Data Science Tasks

- Given a huge set of per-customer sales data, build a model to predict customer “churn”
- Given a large graph of Medicare payout data, find suspicious (potentially fraudulent) referral patterns
- Given a set of EMR data, find previously unknown side effects (ex: Vioxx and heart disease)
- Given data from an online learning tool find markers that are an early sign of later academic achievement problems
- Many, many more!

Both Tools and Models are Important

- Back in the day...
  - You had statisticians who dealt primarily with small data sets
  - You had computer scientists who were interested in advanced modeling
- But in the “Big Data” era, the two can’t live in isolation
  - You need advanced models to solve challenging prediction/analysis tasks
  - You need computer systems that can scale those models to the largest data sets
  - You need computer tools that make it easy to implement complicated models

Important Disclaimer

- Data Science Tools & Models is fundamentally a computer science class!
- This is not “tools and models” from a naive user’s perspective
    - No learning to be an end-user of classical analytics packages
    - This is not a “Get to know R” class
    - Nor is it a “Get to know SAS” class
    - No plugging data into a standard software package and writing a report on the results
    - A class covering such topics WOULD be useful
        - But that’s simply not this class
- Lots of focus on algorithms and engineering

When We Say “Models”

- Strong focus on the math foundations of data science
- Lots of optimization theory, probability, statistics
- Even some continuous mathematics

When We Say “Tools”

- We mean tools for manipulating large data sets
- Tools for scalable, distributed computation
- Emphasis is on “Big Data”!
- Specifically, we’ll learn about:
  - SQL databases
  - Python programming (NumPy, SciPy)
  - Hadoop (MapReduce software, Big Data file system)
  - Spark (distributed Big Data manipulation software)

Example Use Case for Your Data Science Tools & Models Skill Set

- Imagine...
  - You are working at a hospital
  - You collect 5TB of patient monitoring data each day...
  - And want a software to predict what will happen to a patient in the next hour
  - Such a software does not exist...
  - How to build it?
- Key questions to answer:
  - How will you process the raw data?
  - What model will you use to do prediction?
  - How will you train the model?
  - How will you scale to 5TB per day?
- After this class, you'll have the answers!

As Such, this Class...

- Will give an introduction to modern data management software...
  - First half of the class
  - Relational database systems and SQL
  - No-SQL systems such as Hadoop and Spark
- Will give an introduction to models for modern data analysis...
  - Second half of the class
  - Supervised learning (linear models, support vector machines)
  - Unsupervised learning (clustering, matrix factorization)
  - Text mining
- Assignments will focus on implementing the models using the tools & understanding methods for preparing data

Key Goals

1. Respect the data
   - Do good science
   - Make repeatable processes
   - Learn your data/domain
2. Build your toolkit
3. Know when to use the different tools & models
4. Learn to generalize
   - “How can we use what we learned today?"
   - “What do we know now that we didn’t know before?"

Skills You Need to Take this Class

- Should be a reasonable programmer
  - Very comfortable with Python
  - Two assignments use SQL (no knowledge assumed)
  - Four assignments use Python

More Skills You Need to Take this Class

- Should not be afraid of a bit of math
  - Some background in probability/statistics
    - Common distributions (e.g. Gaussian)
    - Expected value
    - Variance, covariance
    - Norms (e.g. L_1, L_2)
  - Some calculus (partial derivatives & the chain rule should not freak you out!)
  - Linear algebra
    - Vectors and scalars
    - Matrix inversion
    - Matrix transposition
    - Dot products

Course Norms

- If you don’t understand something, say something... you’re likely not the only one
- No stupid questions
- We may repeat lectures
- We may adapt assignments
- We may go over some basics that, depending on your background, might be review
- If an assignment is taking too long, speak up! Get help! It may need to be changed

What About Overlap with Other Classes?

- COMP 330/543 and DSCI 302 – biggest overlap
  - Both may NOT be taken for credit
  - COMP version is recommended for Computer Science majors
- COMP 430/533—significant overlap
- COMP 440/502/540/602
  - Many/all of the methods we'll cover will also be covered in those classes

Why take this class?

- To learn what Data Science is
- To develop familiarity and proficiency with common data science tools

Class Syllabus

- Communication...
- Grading and Evaluation...
- Exams...
- Academic misconduct...
- Assignments...

Logistics

- Class
  - MWF 9:00 - 9:50 PM
  - Location: TBD
- Office Hours
  - WF 10:00 - 11:00 AM
  - MWF 3:00 - 4:00 PM
  - Or by appointment
  - Duncan Hall 2062
  - TA Office Hours TBD
  - Watch Piazza for changes

Communications

- Canvas `canvas.rice.edu`
  - Assignments
  - Grades
  - Lecture notes
  - Graded discussions
- Piazza `tbd`
  - Ungraded discussions
- Email
  - DSTM in subject line

Assignments

- Exercises
  - 4 short programming or theory exercises designed to reinforce in-class concepts
- Labs
  - 6 one-hour activities to get initial hands-on experience with a practical concept
- Programming Assignments
  - 6 more in-depth programming assignments

Class Policies – Due Dates

- **Assignment & Exercise** Due Dates
  - Typically due at 11:55 PM
  - Final assignment is due at the **end of the course final exam slot**
  - 1 second – 24 hours late = 10% penalty
  - 24 hours + 1 second – 48 hours late = 20% penalty
  - 48 hours + 1 second – 72 hours late = 30% penalty
  - > 72 hours late: NOT ACCEPTED
  - Last assignment may **NOT** be submitted late
- Canvas is the time keeper – if Canvas says it's late, it's late
- Exceptions will only be made for EXTENDED Canvas outages
- Submit early!

Slip Days

- 3 per person
- Used in full day increments
- Email me BEFORE the deadline
- Once they are gone, they are gone
- May NOT be used on final assignment

Regrades

- Must be requested within 1 week of assignment/quiz being returned
- Intended for errors in grading or MINOR errors
- Not a week-long extension to the assignment
- Process
  - Talk to Risa before class or during office hours
  - Write up request with complete details
  - Email to Risa with “DSTM regrade” subject

Academic Misconduct

- Rice Honor Code
- On SOME assignments, you may share answers. Assume you may not. If it is permitted, it will be noted on the assignment.
- No code sharing of ANY kind
  - Email
  - Whiteboard
  - Sharing / showing your screen
  - Piazza posts
  - Verbally
- No outside help
  - No StackOverflow posting
  - No Googling solutions
  - No asking someone who took the class last year/semester
- “2 line rule”
- OK (encouraged) to look up syntax

Questions?

- If there’s time: on to data