COMP 330: Tools and Models for Data Science

Chris Jermaine and Kia Teymourian

Rice University

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

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 EMRs, find previously unknown side effects (ex: Vioxx and heart disease)

Given data from an online learning tool (ex: STEMScopes) 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 we're terribly 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

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
- Basic optimization theory
- Supervised learning (linear models, support vector machines)
- Unsupervised learning (clustering, matrix factorization)
- Text mining

Projects will focus on implementing the models using the tools

Skills You Need to Take This Class

Should be a proficient programmer

- Really good in a modern, general-purpose language...
- Python or Java preferred (only MATLAB might be OK)
- Will be two assignments using SQL (no knowledge assumed)
- Three assignments using Python
- One assignment using Java
- One assignment you choose (most will use Python)

Skills You Need to Take This Class

Should not be afraid of a bit of math

- Some background in probability/statistics
- Some calculus (partial derivatives should not freak you out!)

Want to Get in Without the Prereqs?

See Chris after class

Will be pretty liberal signing forms... but class still capped at 30!

What About Overlap With Other Classes?

COMP 430—biggest overlap

- First three weeks of class are going to strictly be review
- As will first two assignments (a lot like COMP 430 assignments)

COMP 440/502/540/602

- Many/all of the methods we'll cover will also be covered in those classes

So, what's the point of taking this class?

- The only place where you can get an overview of all of this in one place

Class Syllabus

Communication...

Grading and Evaluation...

Assignments...

Midterms...

Lateness...

Regrade requests...

Academic misconduct...

Questions?

If there's time: on to databases!!!

- What's a database system?