Research Methods for Empirical Computer Science

Spring 2007
Mon 2:25-3:20
Wed 1:50-3:25

Instructor: David Jensen

Answers to FAQs

• What is this course about?
  • I will cover that in today’s lecture. The website has additional information

• Will it be the same as the version you taught last year?
  • It will cover similar topics, though I am altering the course design based on student feedback. Specifically, the workload has been reduced, projects will start earlier in the semester, and lecture and reading topics have been reordered.

• Can I audit?
  • Graduate students cannot audit the course. Much of the value of the class depends on in-class discussion and project work, and that is inconsistent with auditing the course.

• Will you teach the course again?
  • Yes, I expect to teach this course again in Spring 2008.
A Major Event

April 1953

A scientific paper

MOLeCULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This molecule is an extremely long coiled chain of biologically important nature.

A structure for nucleic acid has already been proposed by Crick and Watson. They kindly made a revision of this suggestion, which is presented here. The structure is constructed from two long, double-stranded helices of deoxyribose-nucleic acid (D.N.A.). The bases of the two helices are linked together, each base being linked to a base from the other helix, so that the two are linked together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two are held together.

The bases are held together by hydrogen bonds, which are of two types: polar and non-polar.

If it is assumed that the bases only come into contact in the most stable hydrogen-bonded forms (as is the case with the two helices), it is found that only specific pairs of the bases are able to form hydrogen bonds.

In other words, if an alternate base on one member of a pair, as in DNA, then on those assumptions the other member must be thymine; similarly for cyanine and cytosine. The sequence of bases on a
Why should computer scientists care?

• The paper is of some technical interest
  • Watson & Crick describe a data structure
  • The data structure implies an algorithm

Why should computer scientists care?

• We can learn from almost any scientific discovery
• Theory, experiments, and conjectures all play a role
• Science requires falsifiable hypotheses
• Science involves both collaboration and competition
• Science can be egalitarian
Goals of the course

• Teach how to conduct a personal research program
  • Selecting papers to read and reading them productively
  • Identifying research topics, questions, and hypotheses
  • Planning and conducting experiments
  • Analysis and interpretation of data
• Provide a "jumpstart" for graduate students

Example topics

• Why CS is (and should be) a science
• Selecting good papers and reading them even when you don't understand all the concepts
• Structuring research investigations in terms of algorithms, tasks, and environments
• Why some hypotheses are better than others
• Combining proofs, simulations, and experiments to investigate computational phenomena
• How to be a personally productive researcher
• Why you should have multiple working hypotheses
• Using and critiquing the results of statistical tests
This course is not about CS theory

- We already have several excellent courses in algorithms and theory
- The majority of students in CS do research that is at least partially empirical
- Even theorists need to select good research questions
- Good research practices help you understand how to select and blend alternative methods of producing evidence (e.g., proofs, experiments, and simulations)

This course is not about CS 'craft'

- Like all sciences, CS has "craft elements"
  - Writing good research code
  - Mechanics of writing good papers or making good presentations
  - Writing a dissertation
- This course is about the high-level knowledge of how to do research ("know-why") and some specific research techniques ("know-how") that are general to much of science
- That said, good research practices make it easy to identify what craft elements are most important
This course is not about statistics

- We have a mathematics and statistics department that offers a variety of classes in statistics, modeling, and analysis of experiments
- Many of you have already taken these courses or should take them soon
- That said, we will cover material on how to...
  - Use methods for exploratory data analysis
  - Use and critique statistical hypothesis tests
  - Design the structure of experiments
  - Investigate causal hypotheses

This course is not about professionalism

- For example
  - Getting along with your advisor
  - Building a professional network
  - Interviewing or getting your first job
  - Getting grants or patents
- That said, good research practices lead to...
  - good papers, presentations, and dissertations
  - good professional relationships
  - good careers and ethics
- Rather than focus on "tips" about professional conduct, we will focus on what sets the context for and enables these things—good research
Prerequisites

• Knowledge of basic concepts in computer science and engineering
• Some prior research work or an established research context (e.g., a lab)
• Course in basic statistics
• Reading, writing, and speaking skills
• Willingness to discuss your questions, concerns, doubts, and successes in class throughout the semester

Papers and texts

• Papers
  • Methods papers
  • Case studies from computer science
• An Incomplete Guide to the Art of Discovery
  • Jack Oliver, available free online
  • Examples are from earth science, but content is general
  • Author provides a brief introduction to earth science
• Empirical Methods for Artificial Intelligence
  • Paul Cohen, MIT Press
  • Examples are from AI, but content is general
Course structure

• Classes
  • Standard classes
    • 1/2 lecture
    • 1/2 discussion of readings and case studies
  • "Labs"
    • In-depth discussion of specific examples of lecture topics
      (e.g., selecting a project, testing hypotheses, developing
      research questions and hypotheses)
    • Often using specific examples from student projects

• Today
  • Rest of this lecture
  • Read and discuss a three-page paper

Grading

• 40% project reports
• 20% project reviews
• 20% class participation
  • In-class discussion
  • Two "free passes", called ahead of time
• 20% response reports
  • Three-paragraph responses to readings
  • Drop two lowest
Project

- Individual project: One student = One project
- Small research project evaluating a specific algorithm or system of your choice
- Project components (reports)
  - Task and environment description, Algorithm description, Behavior exploration, Knowledge assessment ("Literature review"), Research proposal, Experimental design, Experimental results, Final report
- Project selection is a key element of success in this class, and must be done early
  - Next Monday’s class will be a lab on project selection
  - Next Monday’s assignment is three one-paragraph project ideas

Class participation

- Read papers/book before class
- Identify your key discussion points
  - Rarely the key points of the paper
  - Instead, go beyond the paper to identify problems, missed opportunities, connections to other work, potential applications, etc..
  - Use the concepts introduced in class
- Write about two of these in your response paper
- Bring up any of your points in class
  - Don’t wait for me to call on you — instead, choose the time and content of your contribution
  - Relate it to what has already been discussed, but don’t worry too much about “sidetracking” the conversation
Reading responses

- Three paragraphs
- Submitted by midnight the day before class using the electronic submission system
- Contents
  - One paragraph summary of goal of the paper
  - Two or more key points that critique, dispute, reinforce, or extend findings of the paper
- Not random musings, but concise comments useful for the next day's discussion
- Points need not be the most central ones, but the ones that most interest you

Reviews

- Brief summary of the report
- Provide constructive feedback to authors
  - Particularly strong aspects of the report
  - Flaws and methods to correct them
  - Missing information
  - Improvements to presentation
- One page for each report you review
- For each report you submit, you will be asked to review two
- Reviews will have a small, but important, impact on the grade a report receives
Website

- http://kdl.cs.umass.edu/courses/rmcs/
- Syllabus
- Schedule with links to readings and slides
- Project assignments
- Recommendations on reviewing
- Pointers to other useful websites and additional readings
- Link to submission system

Readings for Wednesday

- Two articles
- Available from website now
- Response paper due tomorrow by midnight
Personal views

- The course will contain a strong dose of my personal views (whether I plan that or not)
- I will try to identify when I can, but the nature of the viewpoint can make that difficult
- Debate about viewpoints is useful, so don't hesitate to participate in discussion. That is an essential part of this course.

An ongoing process

- This course will be an ongoing conversation
  - Methodology (and science) is always this way
  - This is a still an evolving course
- Don't expect the 'final word' on how to do computer science
- Expect ideas, conflicting opinions, partial answers
- Contribute and discuss — Like all scientific communities, we can get closer to the truth if we work together
Watts & Strogatz