Course Overview

Jaime Arguello
jarguell@email.unc.edu

Course Objectives

- How do search engines work?

  ‣ effectiveness and efficiency

- How do users behave with them?

  ‣ how do users determine usefulness of information?

  ‣ how can a search engine mimic this process?

- Why do search engines fail?

  ‣ the user? the corpus? the system? something else?

- How can they be evaluated (off-line)?

- How can they be monitored and tuned (on-line)?

Why are these important questions?

- Most of the world's information is in natural language text
  - ▸ the world wide web
  - ▸ scientific publications
  - ▸ books
  - ▸ social media interactions

- The amount of this information is growing quickly; human capacity is not (evolution doesn't move that fast)

- We need smarter tools

- IR provides tools for analyzing and organizing content to facilitate search, discovery, and learning

Course Structure

- Information retrieval is an interdisciplinary problem
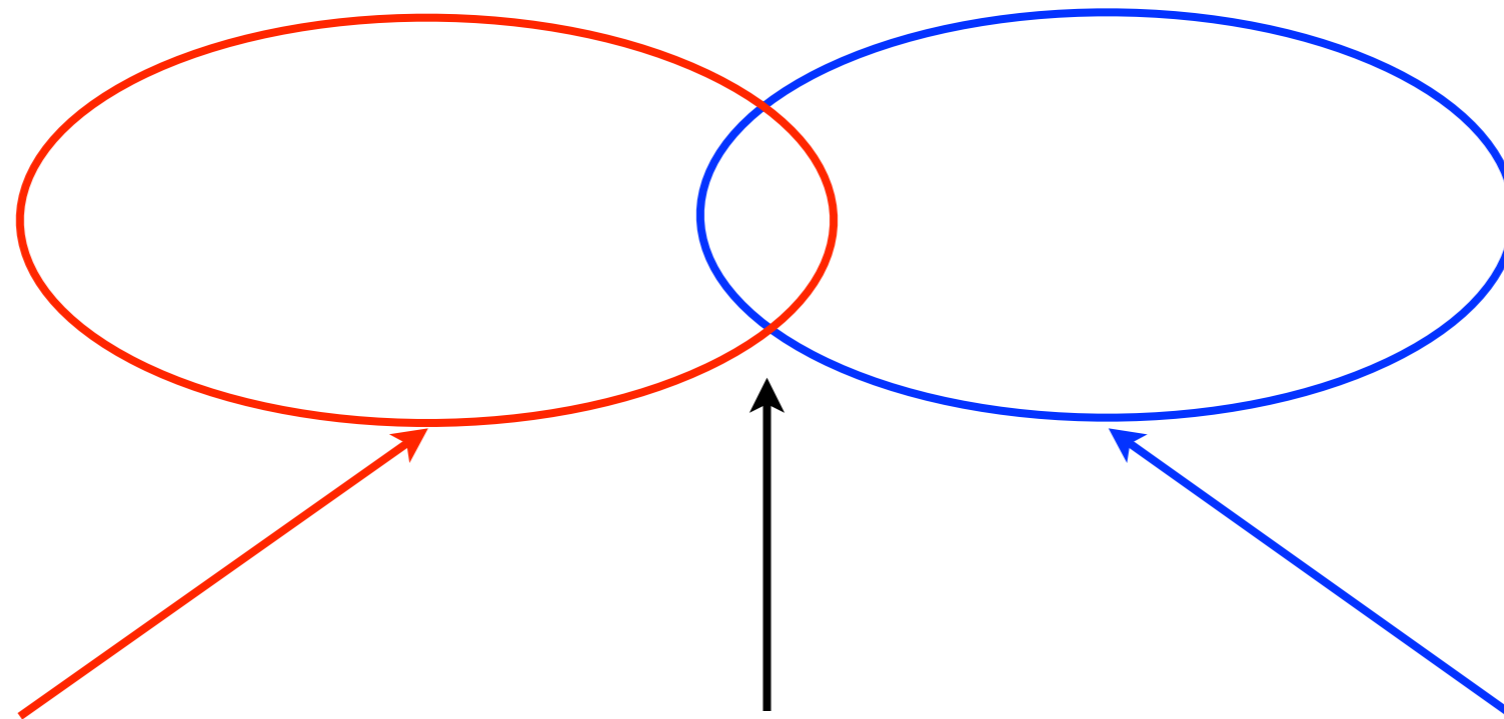
people who want
to understand how
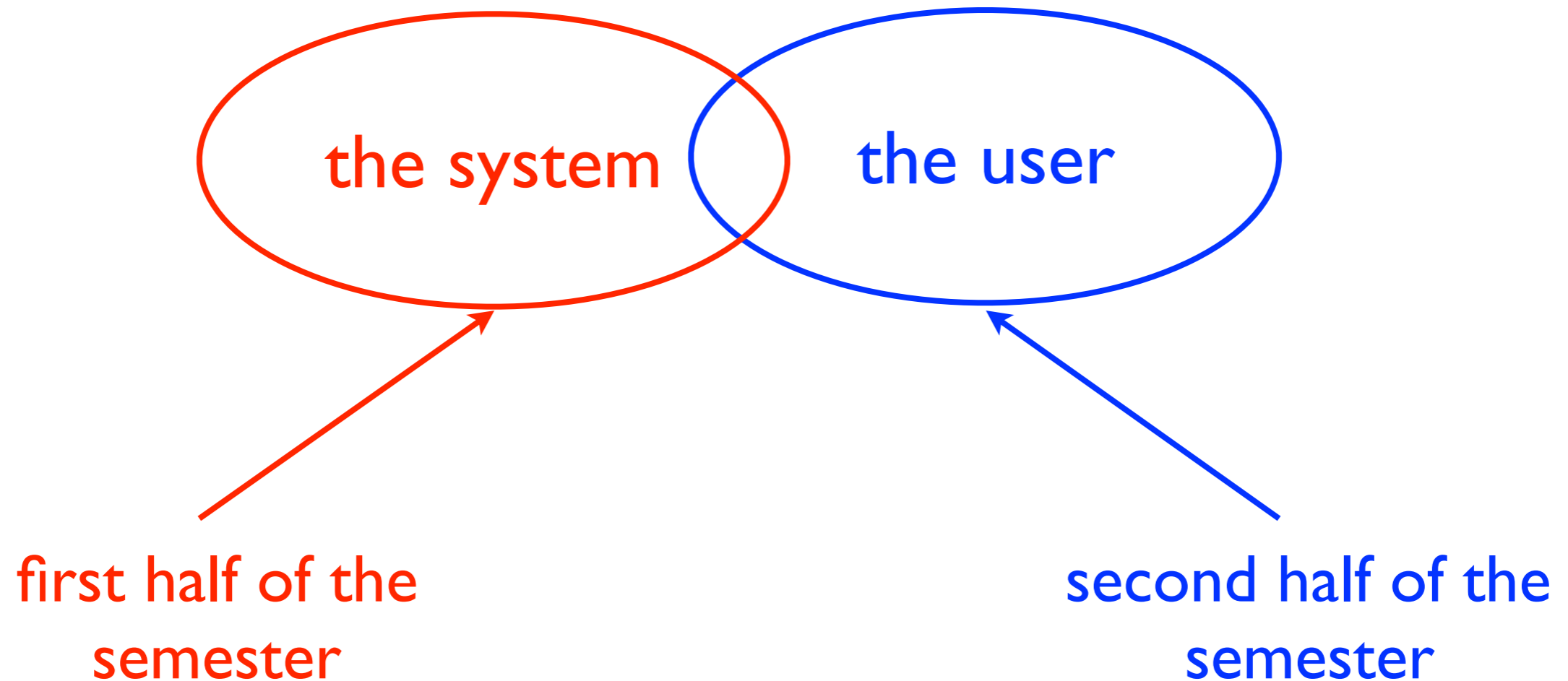computers can solve
problems

people who care
about
information
retrieval

people who want to
understand how people
behave with computers

- We need to understand both ends of the spectrum

Course Structure

- IR: computer-based solutions to a human problem

- Understanding IR systems requires math!

Road Map

- Introduction to ad-hoc retrieval
  - ▸ controlled vocabularies
  - ▸ full-text indexing
- Boolean retrieval
- Indexing and query processing
- Statistical Properties of Text
- Document Representation
- Retrieval Models
  - ▸ vector space model
  - ▸ language modeling
  - ▸ others (depending on how quickly we progress)

Road Map

- Evaluation

  ‣ test-collection construction
  ‣ evaluation metrics
  ‣ experimentation
  ‣ user studies
  ‣ search-log analysis
  ‣ A/B testing

- Studies of search behavior

- Federated Search (?)

- Clustering (?)

- Text Classification (?)

Grading

- 30% homework

  ‣ 10% each

- 15% midterm

- 15% final exam

- 30% literature review

  ‣ 5% proposal
  ‣ 10% presentation
  ‣ 15% paper

- 10% participation

Grading for Graduate Students

- H: 95-100%

- P: 80-94%

- L: 60-79%

- F: 0-59%

Grading for Undergraduate Students

- A+: 97-100%

- A: 94-96%

- A-: 90-93%

- B+: 87-89%

- B: 84-86%

- B-: 80-83%

- C+: 77-79%

- C: 74-76%

- C-: 70-73%

- D+: 67-69%

- D: 64-66%

- D-: 60-63%

- F: <= 59%

Homework vs. Midterm vs. Final

- The homework will be challenging. It should be, you have more time.

Literature Review

- See description on the syllabus

- Form groups of 3 or 4

- Choose an IR task (next slide)

- Write a short proposal (mostly for feedback)

- Review the literature

  ‣ not just the different solutions to the problem

  ‣ the <u>best</u> solutions to the problem!

- Write a paper (~30 pages double-spaced)

- Make a presentation

  ‣ 10 minute presentation + 5 minutes Q&A

Literature Review
example tasks

- Personalized information retrieval

- Session-based information retrieval

- Clustering of search results

- Book search

- Multimedia search (over items not inherently associated with text)

- Social-media data for event-detection and forecasting

- Query-log analysis for event-detection and forecasting

- Faceted search

- Federated search

Literature Review
tips

- Be thorough

- Be scientific

    ▸ don't focus on the writing of the papers you review

    ▸ focus on the science (the method and the evaluation)

- Be constructive

- Contribute new insight and structure

    ▸ your literature review shouldn't read like a "list"

    ▸ connect dots that haven't been connected

- Say what you think!

Course Tips

- Work hard

- Do the assigned readings

- Do <u>other</u> readings

- Be patient and have reasonable expectations

  ‣ you're not supposed to understand everything we cover in class <u>during</u> class

- Seek help sooner rather than later

  ‣ office hours: by appointment

  ‣ questions via email

- Keep laptop usage to a minimum (live in the present)

Course Tips

- You are the most important part of the course!

- Teaching is not all I do ;-)!

Questions?