CS 5968/6968: Data Str & Alg for Scalable Computing Spring 2023

Prashant Pandey

prashant.pandey@utah.edu

SCHOOL OF COMPUTING
UNIVERSITY OF UTAH

no smartphones 

no laptop 

Why?

there is enough evidence that laptops and phones slow you down

Ask questions

… and answer my questions.

Our main **goal** is to have **interesting discussions** that will help to gradually understand the material

(it’s ok if not everything is clear, as long as you have questions!)

Today's agenda

- Course logistics overview
- Why scalable computing?

I want you to speak up!
[and you can always interrupt me]

Course objectives

- Learn about advanced data structures and algorithms to solve massive-scale data analysis problems.
- Next-generation challenges in data systems.

- Students will become proficient in:
  - Advanced data structures and algorithms
  - Writing high-performance and concurrent code
  - Working on a large code base
  - Modern data system internals

Course topics

- Compact trees
- Hash tables
- Filters and sketches
- Locality sensitive hashing
- Nearest neighbor search
- Succinct data structure
- String algorithms
- Graph algorithms
- External memory algorithms
- Distributed data structures

Background

- I assume you have already taken undergrad/grad Data Str & Alg course (e.g., CS 4150 and 6150) or similar.
- You are comfortable with basic data structures and algorithms and writing C/C++ code.
- We will discuss modern variations to classical data structures and algorithms that are designed for massive-scale data.

- Things that we will **not** cover:

Basic data structures, algorithms, asymptotic analysis, recursion.

Course logistics

- Course policies + Schedule

  Refer to canvas

- Course website

https://www.cs.utah.edu/~pandey/courses/cs6968/spring23/index.html

- Academic honesty
  - Refer to SoC policy on academic conduct.
  - If you are not sure, ask me.
  - I am **serious**. DO NO PLAGIARISE.

What is plagiarism

- Listening while someone dictates a solution.
- Basing your solution on any other written solution.
- Copying another student's code or sharing your code with any other student.
- Searching for solution online (e.g., stack overflow, Github, ChatGPT).

What is collaboration

- Asking questions on Piazza.
- Working together to find a good approach for solving a problem.
  - Students with similar understanding of the material.
- A high-level discussion of solution strategy.
- If you collaborate with other students, **declare** it upfront

Instructor office hours

- Before class in my office
  - Mon Wed 9:30 AM – 10:30 AM
  - WEB 2686
- Things that we can talk about:
  - Issues on projects
  - Paper clarification/discussions
  - Getting involved in a research project
  - Help with your research

Teaching assistant

- TA: Benwei Shi
  - 5th year PhD student
  - **Research on**:
    - Algorithms for Big Data Analytics:
    - Geometric Data Analysis
    - Coresets and Sketches
    - Data Mining
    - Machine Learning

Instructor

- Previous:
  - Research Scientist, VMware Research
  - Postdoc: CMU/UC Berkeley
  - PhD: Stony Brook University
  - Intern: Google Research/Intel Labs
- Research:
  - Data structures/algorithms for big data
  - Storage systems & graph processing
  - Computational biology
- Interests:
  - Outdoors: Running/hiking/biking/surfing/ski/…
  - Sports: Cricket/Soccer/Badminton/TT/…

Rio Celeste Rainforest Costa Rica

Course rubric

- Theory/programming assignments
- Final project
- Paper reports
- Final exam
- Class participation

Scribing lectures

- Use the **latex template** to scribe
- Each student may have to scribe 1-2 lectures, depending on class size.
- Pick a date and send an email to the TA. First-come first-served.
- Submit scribe notes (pdf + source).
- Scribe notes are due **by 9pm on the day after lecture**.

Paper reports

- Pick five papers from the reading list. Spread out your picks.
- Write a one-paragraph synopsis of each of the five papers.
- There will be five deadlines throughout the semester.
- Synopsis:
  - What is the problem and why is it hard? (Three sentence).
  - An overview of the main idea and contributions (Three sentences).
  - How do the authors evaluate their solution? (Two sentence).

Plagiarism warning

- Each review must be your own writing.
- You may **not** copy text from the papers or other sources that you find on the web.
- Plagiarism will **not** be tolerated.
  See SoC policy on academic conduct for additional information.

Assignments

- Assignments will include a combination of:
  - Theoretical problems
  - Small programming tasks

- Do all development on your local machine.
  - Can also use Cade machines

- Do all benchmarking using Cade clusters.
  - Cade setup instructions are available in Canvas
  - We will provide further details later in semester

Final project

- Each group (3 people) will choose a project that is:
  - Relevant to the materials discussed in class.
  - Requires a significant theory/programming effort from all team members.
  - Unique (i.e., two groups cannot pick same idea).
  - Approved by me.
- We will provide sample project topics.
- Will have two milestones.

Assignments/Projects

- Assignments 1 and 2 will be done individually
- Final project will be done in a groups of 2 to 3 students
  - You should form groups based on talking to other students
  - Otherwise, we will form groups randomly

Plagiarism warning

- These projects must be all of your own code.
- You may **not** copy source code from other groups or the web.
- Plagiarism will **not** be tolerated.
  See SoC policy on academic conduct for additional information.

Grade breakdown

- Assignment #1 10%
- Assignment #2 20%
- Final project 40%
- Paper reports 10%
- Class participation 10%
- Final 10%

Course mailing list

- Online discussion through Piazza
  - https://piazza.com/utah/spring2023/19151

- If you have a technical question about the projects, please use Piazza
  - Don’t email me or TAs directly

  All non-assignment/non-project questions should be sent to me.

Why scalable computing?

Scalability challenge in a tweet!

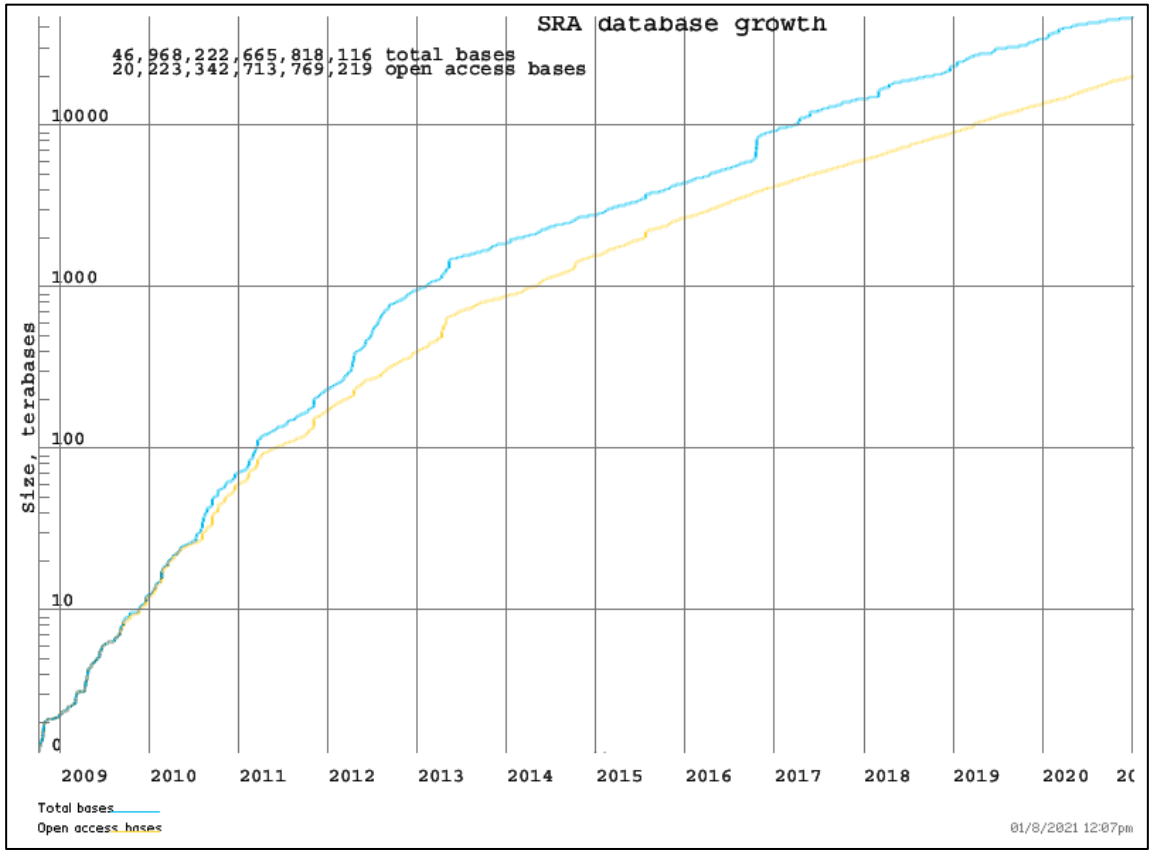
Professor of Bioinformatics
The University of Edinburgh

Professor of Comp Bio
Johns Hopkins University

Sequence read archive (SRA) growth

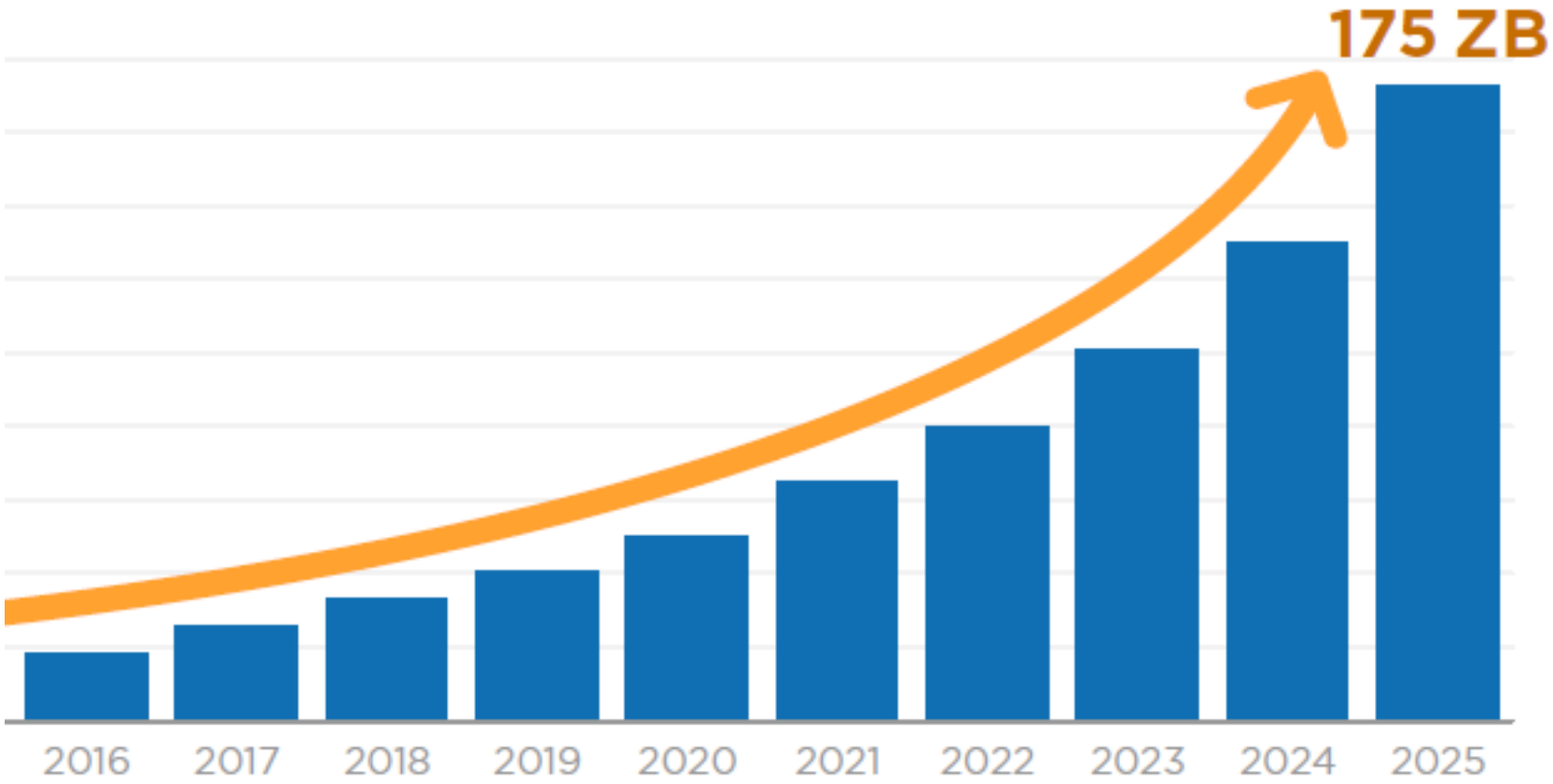
SRA contains a lot of diversity information

Log scale



What if I find, e.g., a new disease-related gene, and want to see if it appeared in other experiments?

Big growth forecasted for Big Data



IDC says 175 ZB will be created by 2025 (image courtesy IDC)

Scalability is a ubiquitous challenge

- People generate 2.5 quintillion bytes of data each day. (**IBM**, 2016)
- More than 150 zettabytes (150 trillion gigabytes) of data will need analysis by 2025. (**Forbes**, 2019)
- 90 percent of the world's data was created between 2015 and 2016 alone. (**IBM**, 2016)

24. 88% of data is ignored by companies.

(Forrester Research)

A widely-quoted figure from a 2012 paper from Forrester Research says that, on average, companies analyze only 12% of the available data. Reasons for this include a lack of analytics tools, repressive data silos, and the difficulty in knowing which information is valuable and which is worth leaving.

How to handle massive data

Shrink it
Make data smaller to fit in RAM

Organize it
Organize data in a disk friendly way

Distribute it
Distribute data on multiple nodes

Next lecture