CS 7280/4973: Data Str & Alg for Scalable Computing Spring 2025

Prashant Pandey

p.pandey@northeastern.edu

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 processing/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
- Succinct data structure
- Hashing/Hash tables
- Filters and sketches
- Cardinality estimation
- Locality sensitive hashing
- Nearest neighbor search (Vector databases)
- External memory algorithms
- Distributed hash tables

Background

- I assume you have already taken undergrad/grad Data Str & Alg course (e.g., CS 3000 and 5800) 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.khoury.northeastern.edu/home/pandey/courses/cs7280/spring25/index.html

- Academic honesty
  - Refer to Northeastern Academic Integrity Policy.
  - 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 <u>together</u> 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, **<u>declare</u>** it upfront

Instructor office hours

- Before class in my office
  - Mon Wed 1:30 PM – 2:30 PM
  - WVH 446
- 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: Zixuan Chen
  - 5th year PhD student
- **Research on**:
  - Truth discovery
  - Graph-related problems
  - Databases
- Interests
  - Movies, basketball (only watch, rarely play)

Instructor

- Research:
  - Large-scale data management
  - Computational biology
  - Graph processing
- Previous:
  - Research Scientist, VMware Research
  - Postdoc: CMU/UC Berkeley
- Interests:
  - Outdoors (Running/Hiking/Biking/Swimming/Skiing/…)
  - Sports (Cricket/Soccer/Racket sports)

Rio Celeste Rainforest Costa Rica

Course rubric

- Programming assignment
- Project
- Final exam
- Class participation and scribe

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**.

Assignments

- Assignment will include a combination of:
  - Small programming tasks
  - Benchmarking and writing report
- Do all development on your local machine.
  - Can also use Discovery machines
- Do all benchmarking using Discovery clusters.

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

- The assignment will be done individually
- The 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 Northeastern Academic Integrity Policy for additional information.

Grade breakdown

- Assignment 20%
- Final project 40%
- Class participation 20%
- Final 20%

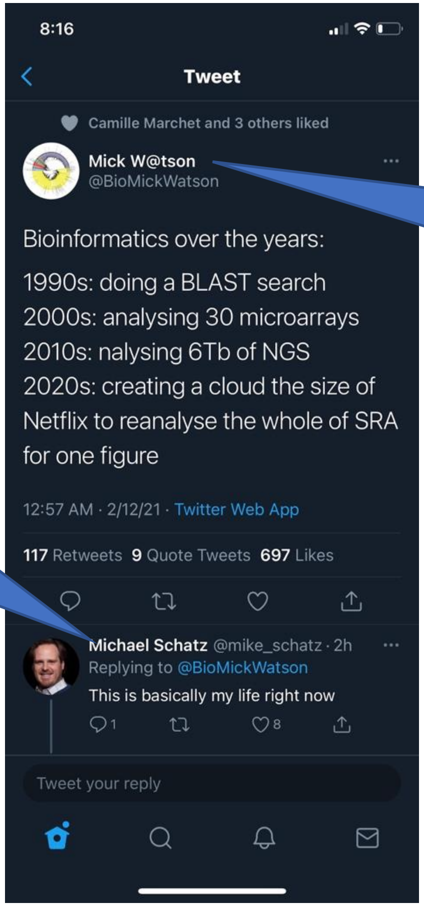
Course mailing list

- Online discussion through Piazza
  - https://piazza.com/northeastern/spring2025/cs7280cs4973202530
- 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!



8:16

Tweet

Camille Marchet and 3 others liked

Mick W@tson
@BioMickWatson

Bioinformatics over the years:

1990s: doing a BLAST search
2000s: analysing 30 microarrays
2010s: nalysing 6Tb of NGS
2020s: creating a cloud the size of Netflix to reanalyse the whole of SRA for one figure

12:57 AM · 2/12/21 · Twitter Web App

117 Retweets 9 Quote Tweets 697 Likes

Michael Schatz @mike_schatz · 2h
Replying to @BioMickWatson
This is basically my life right now
1 8

Tweet your reply

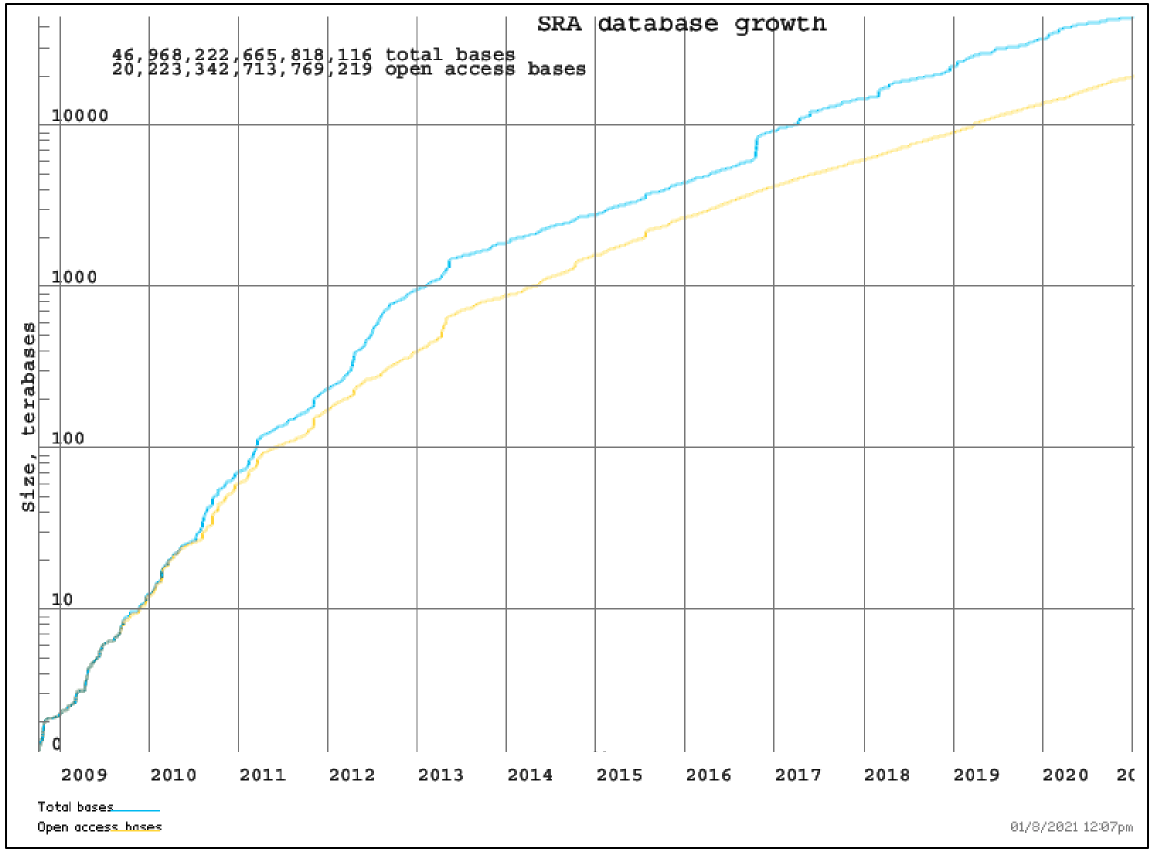
Professor of Bioinformatics
The University of Edinburgh

Professor of Comp Bio
Johns Hopkins University

Sequence read archive (SRA) growth

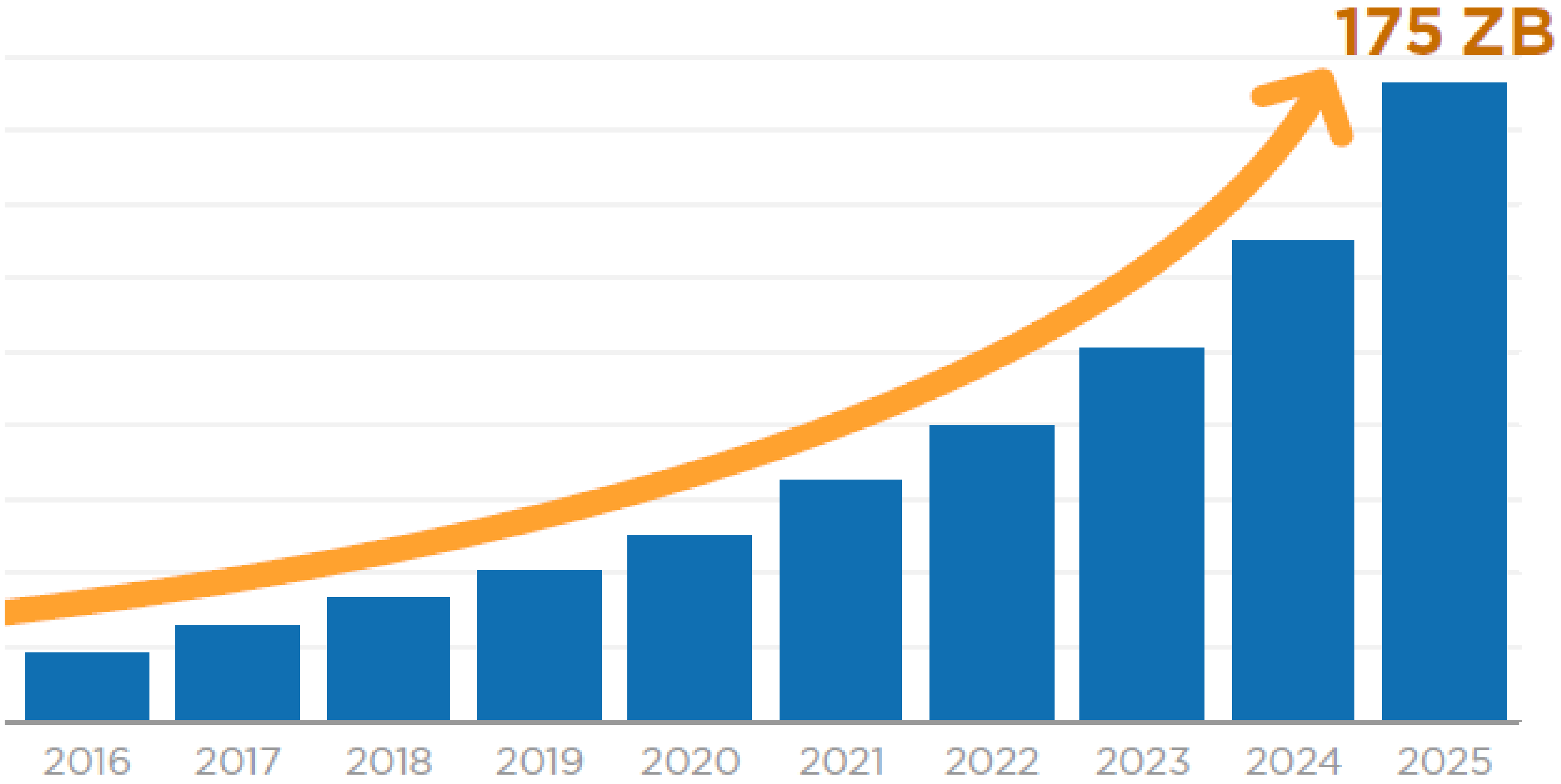
SRA contains a lot of diversity information

Log scale



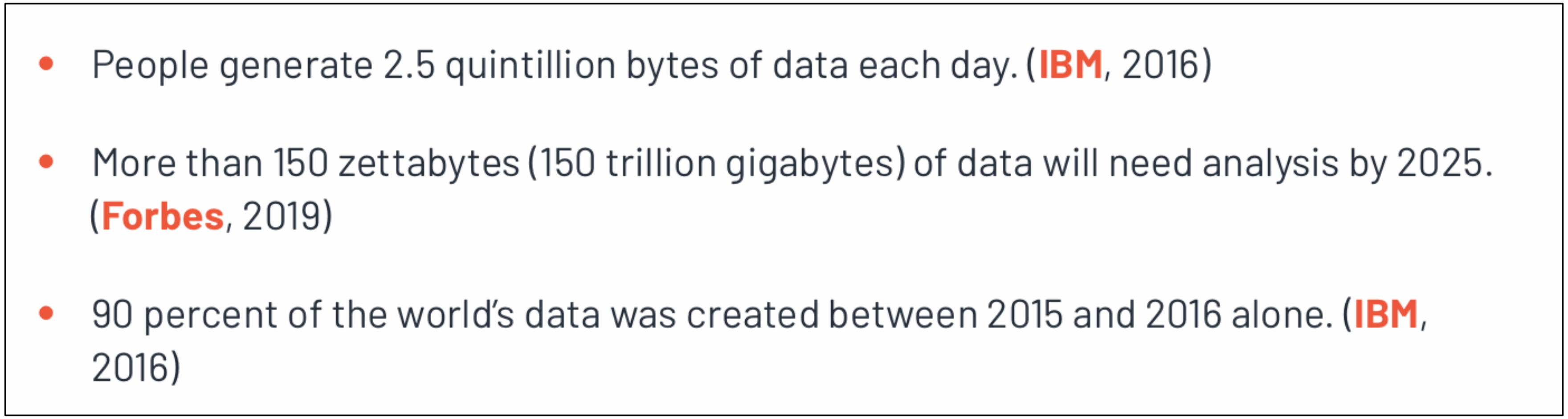
What if I find, e.g., a new disease-related gene, and want to see if it appeared in other experiments?

Big growth forecasted for Big Data

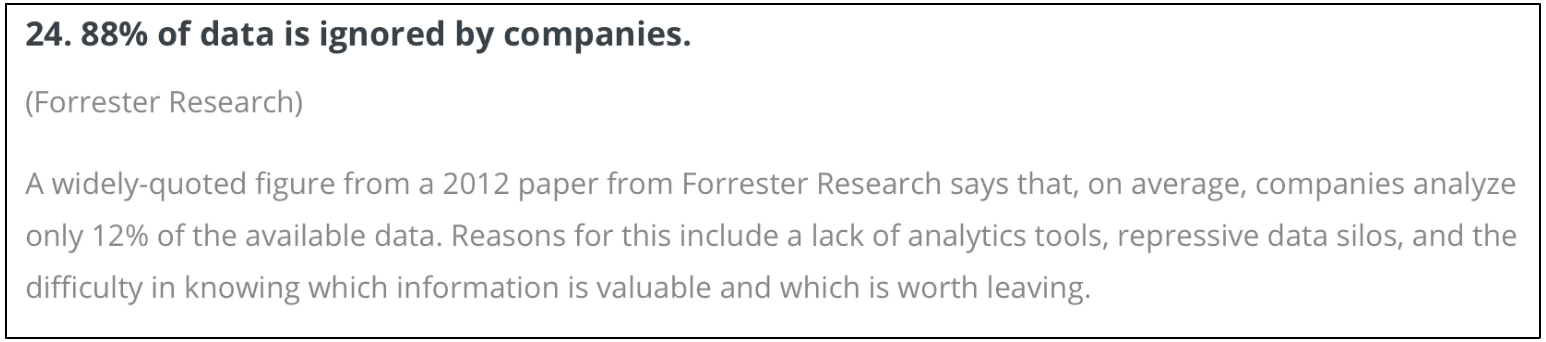


IDC says 175 ZB will be created by 2025 (image courtesy IDC)

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

- Compact trees