About the Course

Introduction to Big Data Algorithms

Sebastian Forster

University of Salzburg

What is this course about?

We will learn about the theory behind big data algorithms.

Modern data sets are large, complex and evolving. We will focus on:

- Models capturing constraints of machines which handle massive amounts of data
- Techniques which allow us to extract succinct yet useful information from large data

What is this course about?

We will learn about the theory behind big data algorithms.

Modern data sets are large, complex and evolving. We will focus on:

- Models capturing constraints of machines which handle massive amounts of data
- Techniques which allow us to extract succinct yet useful information from large data

Methodology:

- Focus on rigoros guarantees
- Claims backed by mathematical proofs

Contents

- Probability Theory “Bootcamp”
- Streaming Model

  Basic Statistics, Counting, Distinct Elements, . . .
- Graph Sparsification

  Spanners, Distance Oracles, Cut Sparsifiers

Contents

- Probability Theory “Bootcamp”
- Streaming Model

  Basic Statistics, Counting, Distinct Elements, …
- Graph Sparsification

  Spanners, Distance Oracles, Cut Sparsifiers

We often focus on structural results rather than on algorithmic techniques

Time Slots

Weekly classes:

- Tuesday, 12:00-14:00 (T02)
- Thursday, 10:00-11:00 (T02)
- Attendance is not mandatory, but strongly recommended

Please let me know if we need to start late or stop early!

Time Slots

Weekly classes:

- Tuesday, 12:00-14:00 (T02)
- Thursday, 10:00-11:00 (T02)
- Attendance is not mandatory, but strongly recommended

Please let me know if we need to start late or stop early!

No class on:

- 31.03, 02.04., 07.04., 09.04., 14.05., 26.05., 04.06.
- By teaching a bit longer, we have flexibility to skip some other classes

Exam

The exam determines 100 % of your grade

- Written exam ($\approx$ 90 minutes)
- First exam date: 30.06.
- Second exam date: 21.07.
- Third exam date: 02.10.

Please let me know very soon if these dates do not work for you!

Compensation for Disadvantage

"Students with a disability or chronic illness/psychosocial disability have a right to a compensation for disadvantages"

Contact Point:

- Disability & Diversity Centre
- Email: disability@plus.ac.at

Homework

- Course type: VU
- Homework is not mandatory, but strongly recommended
- Frequency: $\approx$ 1 homework sheet every other week

Other Organizational Matters

- Announcements via email
- Office hours by appointment
- Course website with all materials: bda.cs.plus.ac.at (→ Teaching)
- No textbook, but “standard topics” (sources will be linked)
- Slides for introductory part
- Handwritten notes for other parts

Please send me a reminder if I forget to upload material!

Big Data Algorithms Group

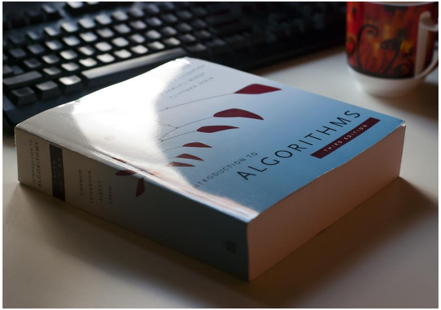
- **Mission:**
  - Develop algorithms for “big” and “fast” data in modern models of computation
  - Integrate machine learning approaches with traditional algorithm design techniques
- Current focus: graph clustering algorithms
- Many opportunities for bachelor and master theses
  (theory or experimental evaluation)

Introduction

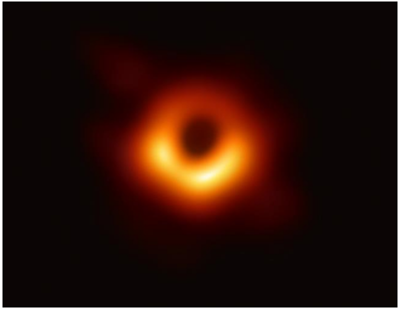
Slides by Aditi Dudeja

Motivation

- In a bachelors' algorithms course we study polynomial-time algorithms.
- **Golden Standard:** O(n) time and space (linear) is sufficient.

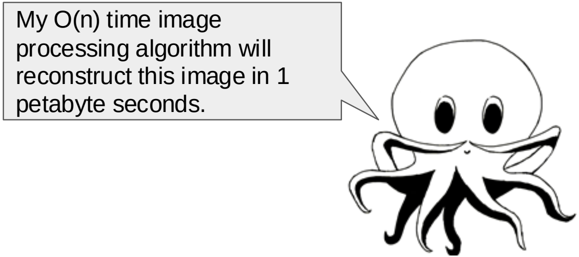


However, the data we work with has grown exponentially compared to the computing resources!



Left: The first reconstructed image of a black hole emerge from the Event Horizon Telescope data.

Right: Dr. Katie Bouman, imaging scientist, posing near hard drives containing some of the petabytes of black hole imaging data collected by telescopes around the world.



My O(n) time image processing algorithm will reconstruct this image in 1 petabyte seconds.

What is Big Data?

No formal definition. But has the following properties:

Volume: Too big to fit in ``memory".

Velocity: Too big to fit in RAM and has to be processed on the fly.

What Are Big Data Algorithms?

Since computer resources are few, big data algorithms access data in a **limited way:**

- Do not read the entire data (Sublinear Algorithms).
- Do not store the entire data (Streaming Algorithms).
- Distribute data across many machines (MapReduce or Massively Parallel).

We will learn about streaming algorithms in the first part of the course. Graph sparsification will apply to many models.

How Accurate Are Big Data Algorithms

Algorithms for large – scale data access it in a limited way. Natural to imagine that they lose some information.

Therefore, we relax guarantees:

- Instead of exact computations, we do approximate computations.
- Instead of requiring that these computations work all the time, we require they work most of the time or with high probability.

One way to do this:

- true answer $\leq$ output $\leq$ α*true answer.
- Typically, α=1+ε for small ε. For example, ε = 0.1 implies that there is a 10% error.
- We will require the equation to hold with probability 1 - δ, where δ is very small. For example, if δ = 0.05, then there is a less than 10% error 95% of the time.