Organizational Matters

Lectures
- Monday 14:00–15:30 (50.34, 236)
- lecture only

Project (mandatory)
- topics will be handed out 06.11.2021
- coding project and small presentation
- 20 % of the final grade

Oral Exam
- 20 minutes
- 80 % of the final grade
- pizza marks content not relevant for exam

Office Hours (Room 210)
- Monday 15:30–16:00 (lecture period)
- by appointment (otherwise)
Materials

Slides
- published shortly before the lecture
  (https://algo2.iti.kit.edu/4612.php)

Videos
- online for old lectures, new topics will be recorded

Additional Material
- references to literature included
- books
- most likely no script
## Content

### Fundamentals
- tries
- suffix tree
- suffix array
- longest common prefix array
- Burrows-Wheeler transform (BWT)
- wavelet tree (+ bit vector rank/select)
- FM-index

### Compressed Indices
- compressing the BWT and wavelet trees
- Lempel-Ziv 77/78 compression
- LZ compression vs. BWT compression
- compressed suffix trees and suffix arrays
- r-index

### Additional Topics
- parallel construction
- different query types
From the Suffix Tree to the $r$-Index

- **Suffix Tree**: 1973
- **Suffix Array**: 1993
- **BWT**: 1994
- **LCP Array**: 1993
- **LCE-Anfragen**: Patricia-Tries
- **FM-Index**: 2000
- **Wavelet Tree**: 2000
- **FM-Index**: 2000
- **r-Index**: 2018
- **Bit-Vektoren mit Rank/Select-Anfragen**: Succinct Data Structures
- **String-Sorting**: EM Hashing
Motivation for Text Indices

- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
- counting queries
- what if there are no “words”
Why Texts?

Text is Everywhere

- Text-based Information
  - Wikipedia
  - dblp
  - books
  - news articles
  - code

- Very Important in Bioinformatics
  - DNA
  - proteins

Growth of DNA Sequencing

[Ste+15]
Bibliography

