Engineering Inverted Indices

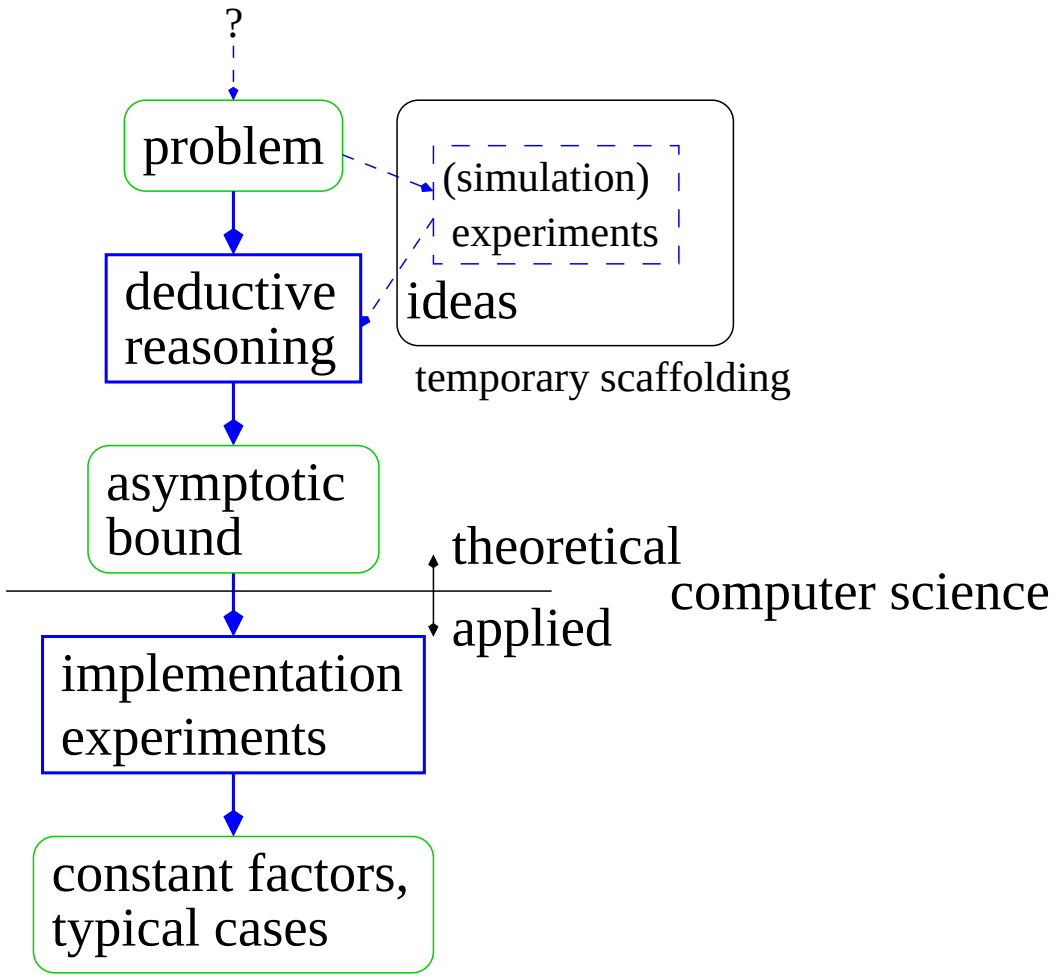
Institut für theoretische Informatik, Algorithmik II

Web:

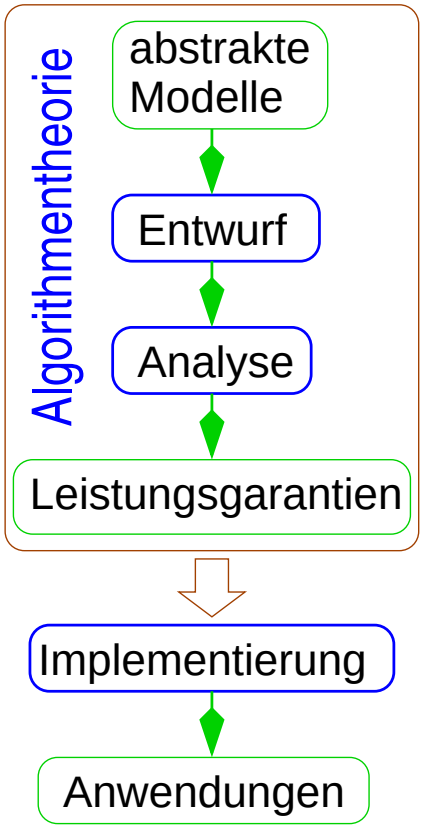
`http://algo2.iti.uni-karlsruhe.de/AlgorithmenI.php`

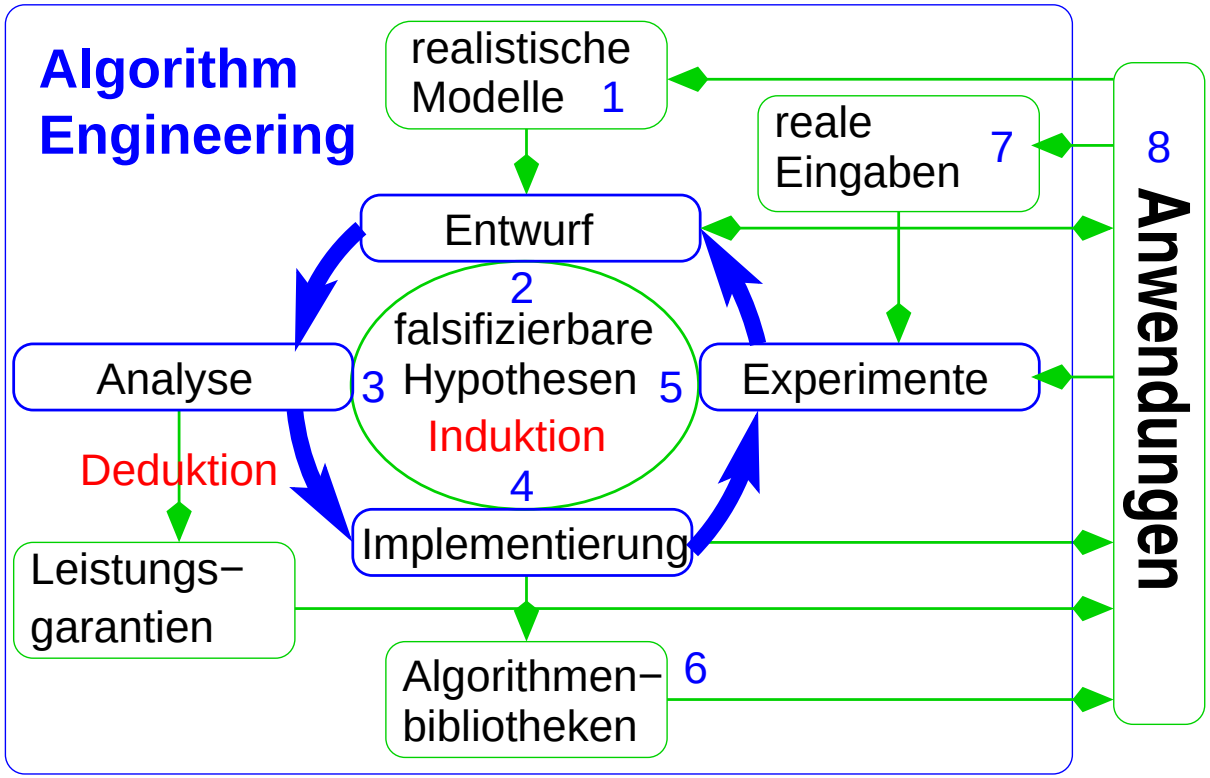
The Traditional Theoretical View?

A Waterfall Model of Algorithmics



Algorithm Engineering



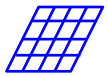
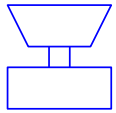
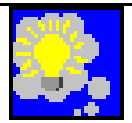
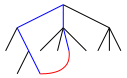
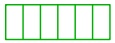


Goals

- Theory meets technology — machine models must cope with technologiscal developments

- Faster transfer of algorithmic result into applications
- Bridge existing gaps

Symptoms of Gaps

theory		$\longleftrightarrow$		**practice**
simple		**problem model**		complex
simple		**machine model**		real
complex		**algorithms**	FOR	simple
advanced		**data structures**		arrays,. . .
worst case	max	**complexity measure**		real inputs
asymptotical	$O(\cdot)$	**efficiency**	42%	constant factors

Intersecting Sorted Lists

shorter list M, $m = |M|$ longer list N, $n = |M|$

B : a tuning parameter

Zipper: merge takes time $O(m+n)$, very small constant factors, ideal fit with compression

Comparison Based: $O\left(m \log \frac{n}{m}\right)$, large constant factors (cache), compression unfriendly

Skipper: $O\left(mB + \frac{n}{B}\right)$, small constant factors, compression friendly.

Idea: skip packages of size B in N.

Lookup: $O(mB)$ expected, small constant factors, compression friendly.

B is expected number of IDs per bucket

Compression techniques

- Δ-encoding – just store differences
- bit compression – use $\lceil \log \max \Delta \rceil$ bits
- variable bitlength encoding
  - escaping: $x < 2^k$? store $\boxed{x, 0}$ else, store

    $\boxed{x \bmod 2^k, 1, \text{encode}(x \operatorname{div} 2^k)}$
  - Golomb coding

Golomb Coding

Betrachte TuningParameter M.

$x \rightarrow (\text{unary}(1+q), r)$

where $q = \left\lfloor \frac{x-1}{M} \right\rfloor$,

$r = x - qM - 1$

Satz: bei geeigneter Wahl von M ist Golomb coding annähernd informationstheoretisch optimal falls die zu codierenden Zahlen geometrisch verteilt sind.

Comparison Based Set Intersection

- lower bound