Algorithms and Data Structures

Searching in Lists

Ulf Leser

Topic of Next Lessons

- Search: Given a (sorted or unsorted) list A with |A|=n elements (integers). Check whether a given value c is contained in A or not
  - Search returns true or false
  - If A is sorted, we can exploit transitivity
  - Fundamental problem with a zillion applications
- Select: Given an unsorted list A with |A|=n elements (integers). Return the i‘th largest element of A.
  - Returns an element of A
  - The sorted case is trivial – return A[i]
  - Interesting problem (especially for median) with many applications
  - [Interesting proof]

Content of this Lecture

- Searching in Unsorted Lists
- Searching in Sorted Lists
- Selecting in Unsorted Lists

Searching in an Unsorted List

- No magic is known
- Compare c to every element of A
- Worst case ($c \notin A$): O(n)
- Average case ($c \in A$)
  - If c is at position i, we require i tests
  - All positions are equally likely: probability 1/n
  - This gives

$$\frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n} * \frac{n^2+n}{2} = \frac{n+1}{2} = O(n)$$

```
A: unsorted_int_array;
c: int;
for i := 1.. |A| do
  if A[i]=c then
    return true;
  end if;
end for;
return false;
```

Content of this Lecture

- Searching in Unsorted Lists
- Searching in Sorted Lists
  - Binary Search
  - Fibonacci Search
  - Interpolation Search
- Selecting in Unsorted Lists

Binary Search (binsearch)

- If A is sorted, we can be much faster
- Binsearch: Exploit transitivity

```
func bool binsearch(A: sorted_array;
                      c,l,r : int) {
  If l>r then
    return false;
  end if;
  m := l+(r-l) div 2;
  If c<A[m] then
    return binsearch(A, c, l, m-1);
  else if c>A[m] then
    return binsearch(A, c, m+1, r);
  else
    return true;
  end if;
}
```

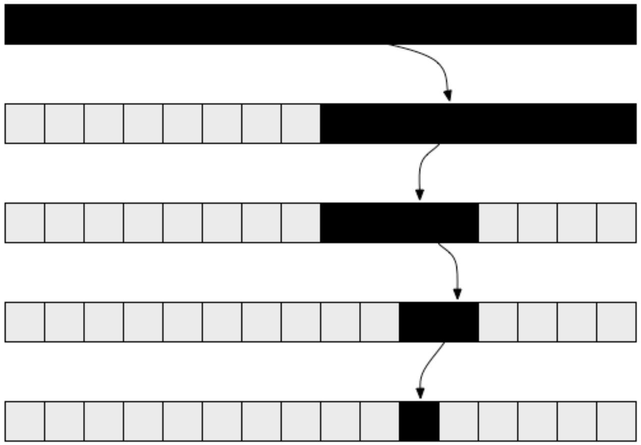
Iterative Binsearch

- Binsearch uses only end-recursion
- Equivalent iterative program
  - No call stack
  - We don't need old values for l,r
  - O(1) additional space

```
A: sorted_int_array;
c: int;
l := 1;
r := |A|;
while l≤r do
  m := l+(r-l) div 2;
  if c<A[m] then
    r := m-1;
  else if c>A[m] then
    l := m+1;
  else
    return true;
end while,
return false;
```

Complexity of Binsearch

- In every call to binsearch (or every while-loop), we only do constant work
- With every call, we reduce the size of sub-array by 50%
  - We call binsearch once with n, with n/2, with n/4, ...
- Binsearch has worst-case complexity O(log(n))
- Average case is only marginally better
  - Chances to “hit” target in the middle of an interval are low in most cases
  - See Ottmann/Widmayer



Source: railspikes.com

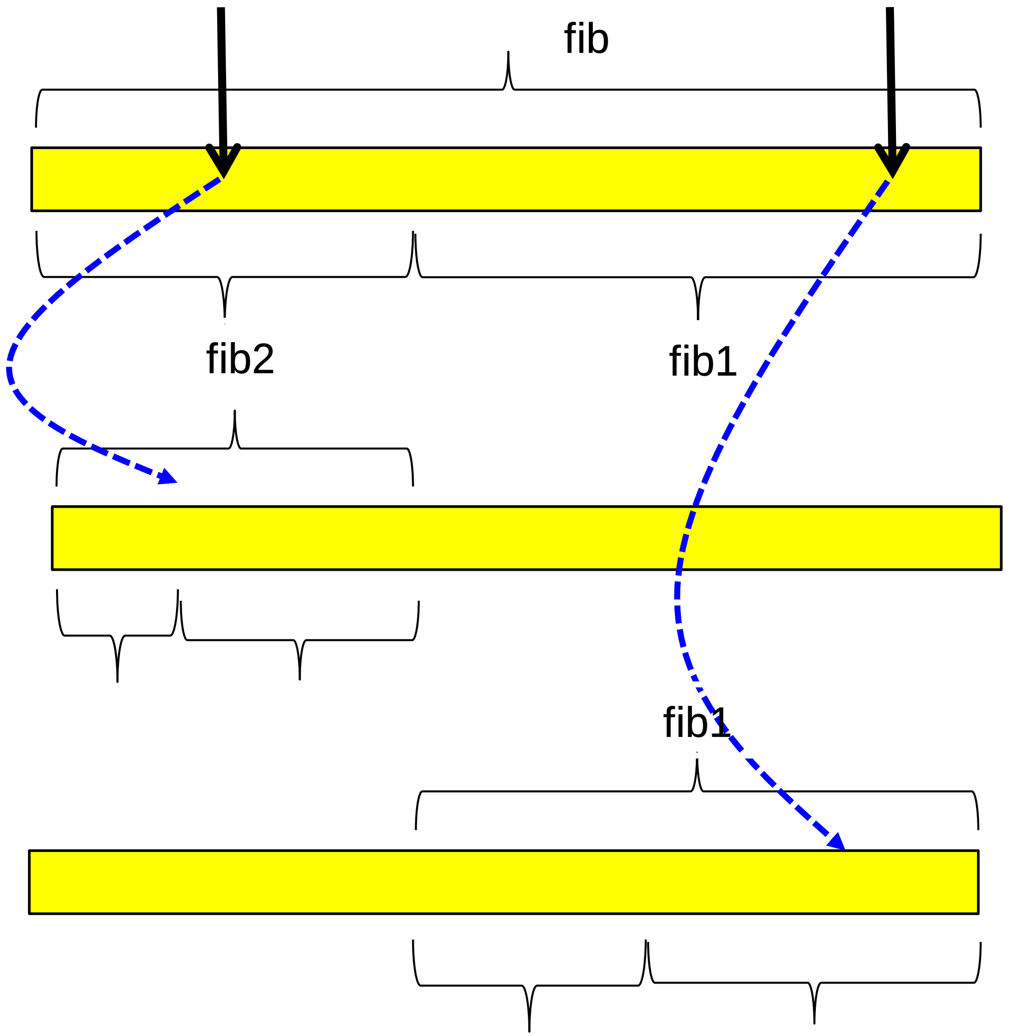
Content of this Lecture

- Searching in Unsorted Lists
- Searching in Sorted Lists
  - Binary Search
  - Fibonacci Search
  - Interpolation Search
- Selecting in Unsorted Lists

Searching without Divisions

- If we want to be ultra-fast, we should use only simple arithmetic operations
- We seek an O(log(n)) algorithm that does not use division or multiplication
- We need to “imitate” the 50%-split of the array
- Recall Fibonacci numbers
  - fib(n)=fib(n-1)+fib(n-2)
  - 1,1, 2, 3, 5, 8 , 13, 21, 34, ...
  - Thus, fib(n-2) is roughly 1/3, fib(n-1) roughly 2/3 of fib(n)

Fibonacci Search



- Do not break in the middle, but at the border of two Fibonacci numbers
- Will lead to O(log(n)) comparisons
- Can be implemented without divisions

Details and Complexity

- Find Fibonacci numbers fib2, fib1, fib such that fib2+fib1=fib and fib≥n and fib1<n and fib2<fib1
  - If fib=n, we generate arrays of size fib(n-1), fib(n-2) ...
  - Otherwise, there is a “layover” – be careful during search
- Worst-case: C is always in the larger (fib1) fraction of A
  - We roughly call once for n, once for 2n/3, once for 4n/9, ...
- Formula of Moivre-Binet: For large n ...

$$fib(i) = \left[\frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^{i+1} \right] \sim c * 1.62^{i}$$

- When breaking the array, we have n~fib(i)~c*$1{,}62^{i}$
- Thus, we make i comparisons (we break the array i times)
- Since i=1/c*$\log_{1,62}$(n), it follows that we make O(log(n)) comparisons

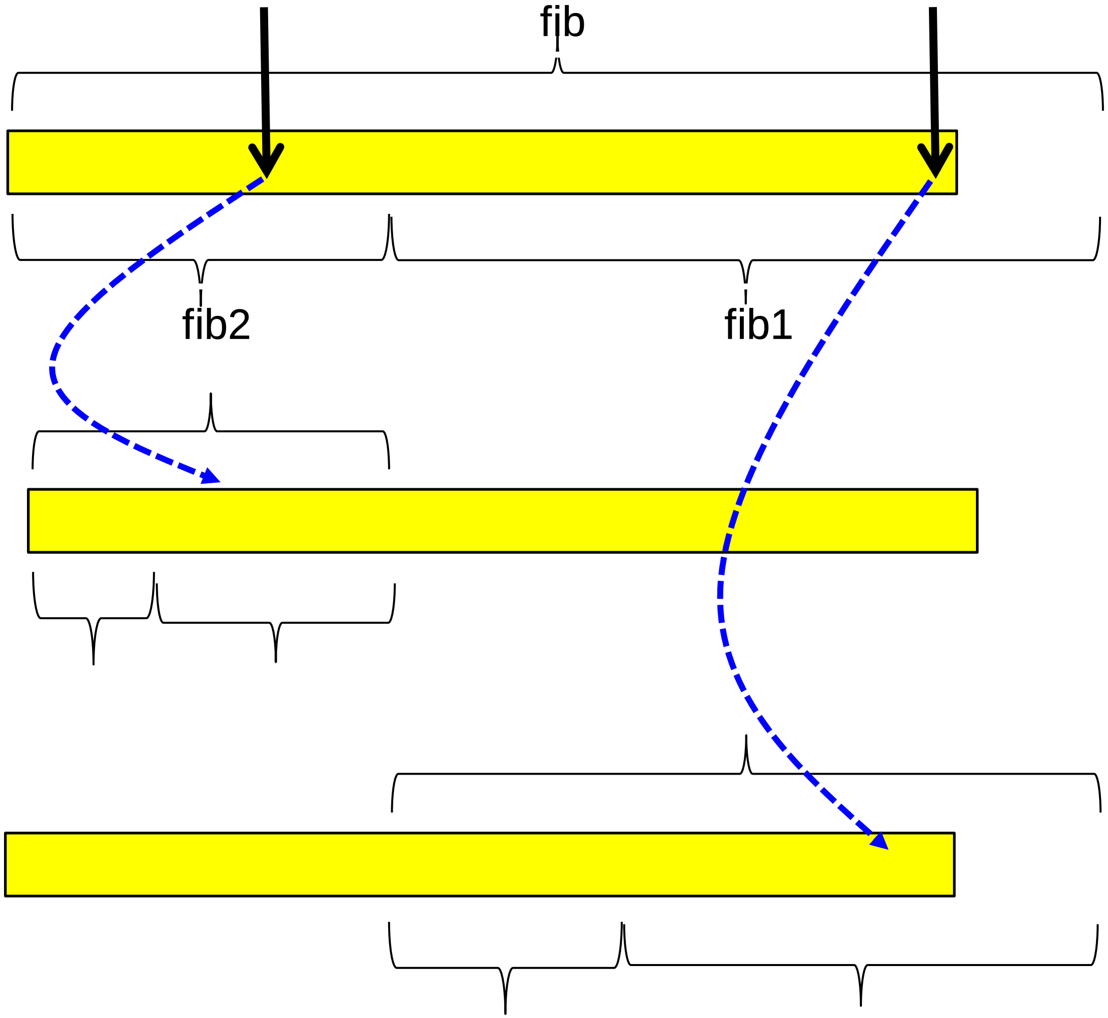
Algorithm

- 6-10: We find fib2, fib1, fib
  - With those we can iteratively compute all other fib's
  - fib(n)=fib(n-1)+fib(n-2)
  - fib(n-1)=fib(n-2)+fib(n-3)
  - ...
- After each comparison, we update fib, fib1, and fib2
- Offset: Left border of next interval

```
1.  A: sorted_int_array;
2.  c: int;
3.  fib2 := 1;
4.  fib1 := 1;
5.  fib := 2;
6.  while fib<n do
7.    fib2 := fib1;
8.    fib1 := fib:
9.    fib := fib1+fib2;
10. end while;
11. offset := 0;
12. while fib>1 do
13.   m := min(offset+fib2, n)
14.   if c<A[m] then
15.     fib := fib2;
16.     fib1 := fib1-fib2;
17.     fib2 := fib-fib1;
18.   else if c>A[m] then
19.     fib := fib1;
20.     fib1 := fib2;
21.     fib2 := fib-fib1;
22.     offset := m;
23.   else
24.      return true;
25.   end if;
26. end while;
27. return false;
```

Layover

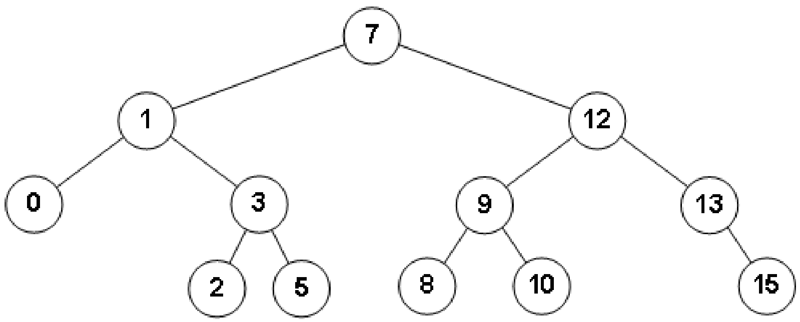
Algorithm




```
A: sorted_int_array;
c: int;
fib2 := 1;
fib1 := 1;
fib := 2;
while fib<n do
  fib2 := fib1;
  fib1 := fib:
  fib := fib1+fib2;
end while;
offset := 0;
while fib>1 do
  m := min(offset+fib2, n)
  if c<A[m] then
    fib := fib2;
    fib1 := fib1-fib2;
    fib2 := fib-fib1;
  else if c>A[m] then
    fib := fib1;
    fib1 := fib2;
    fib2 := fib-fib1;
    offset := m;
  else
     return true;
  end if;
end while;
return false;
```

Outlook (sketch)

- We actually can solve the search problem in O(log(n)) using only comparisons
- Transform A into a balanced binary search tree
  - At every node, the depth of the two subtrees differ by at most 1
  - At every node n, all values in the left (right) subtree are smaller (larger) than n
- Search
  - Recursively compare c to node labels and descend left/right
  - Tree has depth O(log(n))
  - We need at most log(n) comparisons – and nothing else



Content of this Lecture

- Searching in Unsorted Lists
- Searching in Sorted Lists
  - Binary Search
  - Fibonacci Search
  - Interpolation Search
- Selecting in Unsorted Lists

Interpolation Search

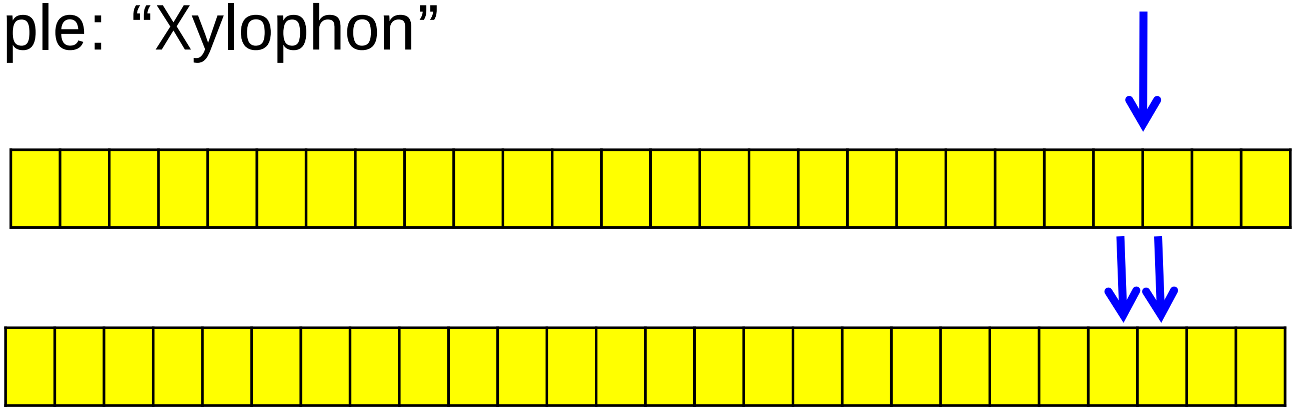
- Imagine you have a telephone book and search for „Zacharias“
- Will you open the book in the middle?
- As in sorting, we can exploit additional knowledge about our values, i.e., use more than just comparisons
- Interpolation Search: Estimate where c lies in A based on the distribution of values in A
  - Simple: Use max and min values in A and assume equal distribution
  - Complex: Approximation of real distribution (histograms, ...)

Simple Interpolation Search

- Assume equal distribution – values within A are equally distributed in range [A[1], A[n]]
- Best guess for the rank of c

$$rank(c) = l + (r - l) * \frac{c - A[l]}{A[r] - A[l]}$$

- Idea: Use m=rank(c) and proceed recursively
- Example: "Xylophon"

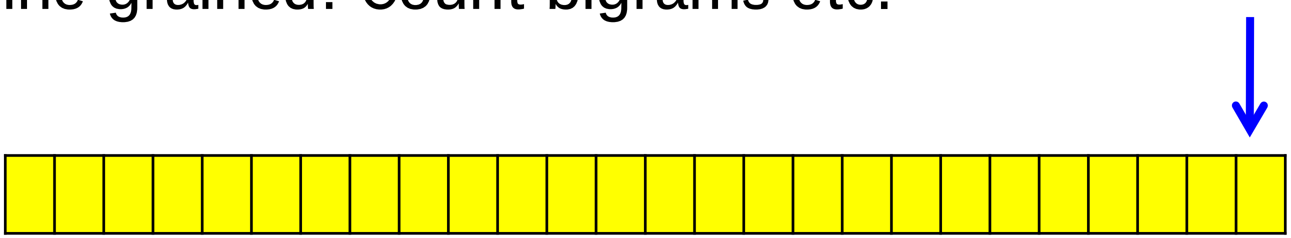


Analysis

- On average, Interpolation Search on equally distributed data requires O(log(log(N)) comparison (see [OW])
- But: Worst-case is O(N)
  - If concrete distribution deviates heavily from expected distribution
  - E.g., A contains only names>"Xanthippe"
- Further disadvantage: In each phase, we perform ~4 adds/subs and 2*mults/divs
  - Assume this takes 12 cycles (1 mult/div = 4 cycles)
  - Binsearch requires 2*adds/subs + 1*div ~6 cycles
  - Even for $N=2^{32}$~4E9, this yields 12*log(log(4E9))~72 ops versus 6*log(4E9)~180 ops – not that much difference

Going Further: Histograms

- For very large N, it might be worth to approximate the real value distribution in A
- Idea: If $|\Sigma|=k$, pre-compute the frequencies $f(k)$ of values starting with a character smaller-or-equal than k
  - Names: How many start with A, A or B, A or B or C, ...
  - Pre-computation: One scan, or use sampling
- Given c, use $f(c[1])$ as start point
  - More on this: Histograms (Lecture Impl. of Databases)
- More fine grained: Count bigrams etc.



Content of this Lecture

- Searching in Unsorted Lists
- Searching in Sorted Lists
- Selecting in Unsorted Lists
  - Naïve or clever

Quartiles

- The median of a list is its middle value
  - Sort all values and take the one in the middle
- Generalization: x%-Quartiles
  - Sort all values and take the value at x% of the values
  - Typical: 25, 75, 90, -quartiles
    - How long do 90% of all students need?
  - Median = 50%-quartile

Selection Problem

- Definition
  The selection problem is to find the x%-quartile of a set A of unsorted values

- We can sort A and then access the quartile directly
- Thus, O(n*log(n)) is easy to reach
- But can we solve this problem in linear time?
- It is easy to see that we have to look at least at each value once; thus, the problem is in $\Omega(n)$

Top-k Problem

- Top-k: Find the k largest value in A
- For small k, a naïve solution already is linear
  - repeat k times
  - go through A and find largest value v;
  - remove v from A;
  - return v
  - Requires k*|A|=O(|A|) comparisons
- Naïve solution is optimal for constant k
- But if k=x*|A|, we have x*|A|*|A|=O(|A|^2) comparisons
  - We measure complexity in size of the input – but what is the input?
  - It is decisive whether k depends on input or not

Selection Problem in Linear Time

- We sketch an algorithm which solves the problem for arbitrary x in linear time
  - Actually, we solve the equivalent problem of returning the k'th value in the sorted A (of course, without sorting A)
- Interesting from a theoretical point-of-view
- Practically, the algorithm is of no importance because the polynomial factors get enormously large
- It is instructive to see why (and where)

Algorithm

- Recall QuickSort: Chose pivot element p, divide array wrt p, recursively sort both partitions using the same trick
- We reuse-the idea: Chose pivot element p, divide array wrt p, recursively select in the one partition that must contain the k’th element

```
func integer divide(A array;
                        l,r integer) {
  …
  while true
    repeat
      i := i+1;
    until A[i]>=val;
    repeat
      j := j-1;
    until A[j]<=val or j<i;
    if i>j then
      break while;
    end if;
    swap( A[i], A[j]);
  end while;
  swap( A[i], A[r]);
  return i;
}
```

```
func int quartile(A array;
                      k, l, r int) {
  if r≤l then
    return A[l];
  end if;
  pos := divide( A, l, r);
  if (k ≤ pos-l) then
    return quartile(A, k, l, pos-1);
  else
    return quartile(A, k-pos+l, pos, r);
  end if;
}
```

Analysis

```
func int quartile(A array;
                  k, l, r int) {
  if r≤l then
    return A[l];
  end if;
  pos := divide( A, l, r);
  if (k ≤ pos-l) then
    return quartile(A, k, l, pos-1);
  else
    return quartile(A, k-pos+l, pos, r);
  end if;
}
```

- Worst-case: Assume arbitrarily badly chosen pivot elements
- pos always is r-1 (or l+1)
- Gives $O(n^2)$
- Need to chose the pivot element p more carefully

Choosing p

- Assume we can chose p such that we always continue with at most q% of A
  – Any q; extend of reduction depends on n
- We would perform T(n) = T(q*n) +c*n comparisons
  – T(q*n) – recursive descent
  – c*n – function "divide"
- $T(n) = T(q*n)+c*n = T(q^2*n)+q*c*n+c*n = T(q^2n)+(q+1)*c*n = T(q^3n)+(q^2+q+1)*c*n = \ldots$

$$\underset{n\to\infty}{T(n)} = c*n*\sum_{i=0}^{n} q^i \leq c*n*\sum_{i=0}^{\infty} q^i = c*n*\frac{1}{1-q} = O(n)$$

Discussion

- Our algorithm has worst-case complexity O(n) when we manage to always reduce the array by a fraction of its size – no matter, how large the fraction
- This is not an average-case. We require to always (not on average) cut some fraction of A
- Eh – magic?
- No – follows from the way we defined complexity and what we consider as input
- Many ops are “hidden” in the polynomial factors
  - q=0.9: c*10*n
  - q=0.99: c*100*n
  - q=0.999: c*1000*n

Median-of-Median

- How can we guarantee to always cut a fraction of A?
- Median-of-median algorithm
  - Partition A in stretches of length 5
  - Compute the median v_i for each partition (with i<floor(n/5))
  - Use the (approximated) median v of all v_i as pivot element
- Possible in O(n)
  - Run through A in jumps of length 5
  - Find each median in constant time ("sorting" of lists of length 5 is not dependent on n – constant time)
  - Call algorithm recursively on all medians
  - Since we always reduce the range of values to look at by 80%, this requires O(n) time (see previous slides)

Why Does this Help?

- We have ~n/5 first-level-medians v_i
- v (as median of medians) is smaller than halve of them and greater than the other halve (both ~n/10 values)
- Each v_i itself is smaller than (greater than) 2 values from A
- Since for the smaller (greater) medians this median itself is also smaller (greater) than v, v is larger (smaller) than at least 3*n/10 elements

Illustration (source: Wikipedia)

	12	15	11	2	9	5	0	7	3	21	44	40	1	18	20	32	19	35	37	39
	13	16	14	8	10	26	6	33	4	27	49	46	52	25	51	34	43	56	72	79
Medians	17	23	24	28	29	30	31	36	42	47	50	55	58	60	63	65	66	67	81	83
	22	45	38	53	61	41	62	82	54	48	59	57	71	78	64	80	70	76	85	87
	96	95	94	86	89	69	68	97	73	92	74	88	99	84	75	90	77	93	98	91

- Median-of-median of a randomly permuted list 0..99
- For clarity, each 5-tuple is sorted (top-down) and all 5-tuples are sorted by median (left-right)
- Gray/white: Values with actually smaller/greater than med-of-med 47
- Blue: Range with certainly smaller / larger values