algorithms for big data

giuseppe di battista

what is big data?

- you already know

what is an algorithm?

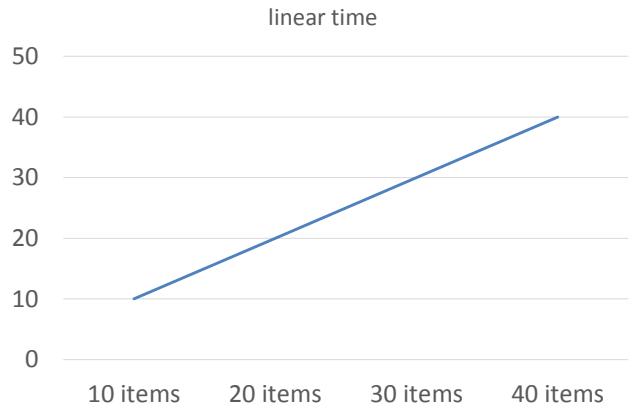
- procedure for solving a problem
  - from the name of the mathematician Mohammed ibn-Musa al-Khwarizmi, part of the royal court in Baghdad, about 780 to 850 A.C.
- computing is solving problems
- example
  - procedure for sorting a set of names in alphabetical order

efficiency of algorithms

- executing algorithms on computers requires resources
  - time
    - an algorithm that sorts 100 names in 1 second is better than an algorithm that sorts 100 names in 10 seconds
  - memory
    - an algorithm that for sorting 100 names requires 1 Mbyte is better than an algorithm that for sorting 100 names requires 10 Mbytes

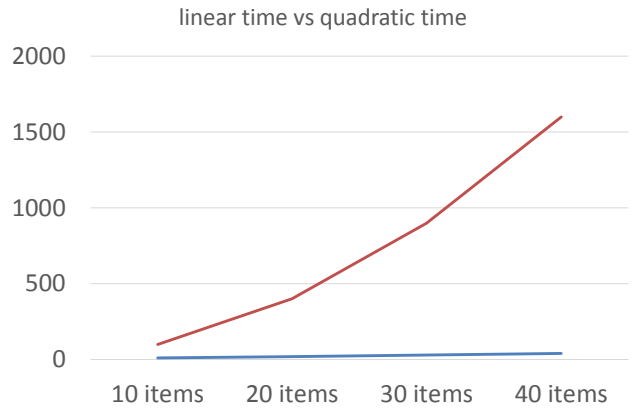
asymptotic analysis

- what happens when data become big?



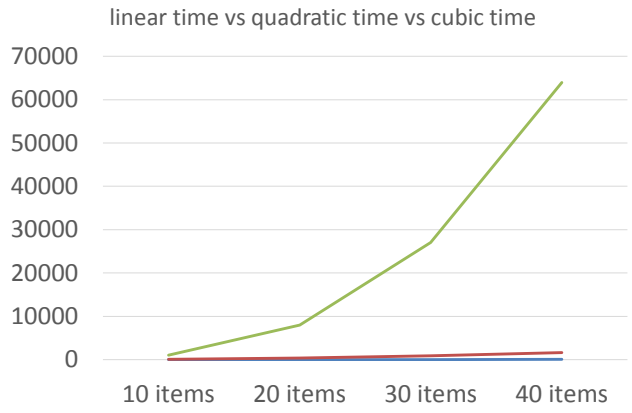
asymptotic analysis

- what happens when data become big?



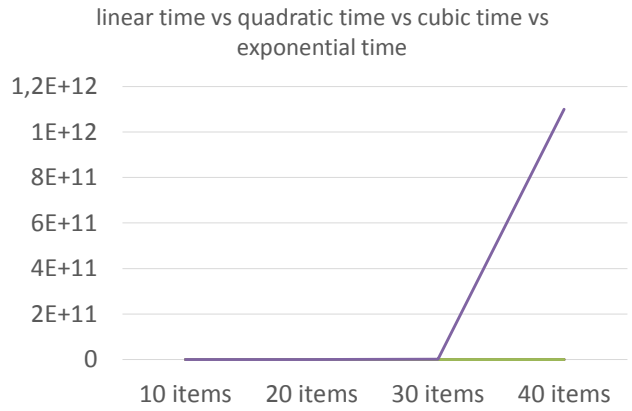
asymptotic analysis

- what happens when data become big?



asymptotic analysis

- what happens when data become big?



asymptotic analysis

- if an algorithm requires time (memory) that grows linearly with an input of size n we say that it requires $O(n)$ time (memory)
- if it grows quadratically $O(n^2)$, cubically $O(n^3)$, etc.
- if it grows logarithmically $O(\log n)$

why big data require new algorithms?

- streaming scenario: data arrive one-by-one, are so many that only a few of them can be stored
  - streaming algorithms
- cloud scenario: data are huge, are in the cloud, and must be processed on a local computer; it is unfeasible even only look at all of them
  - sublinear algorithms

why big data require new algorithms?

- data analysis scenario: data are so many that they can be analyzed only if grouped into homogeneous sets
  - clustering algorithms

a fundamental big data contribution

devised between the 1st and 4th centuries by Indian mathematicians

the Hindu–Arabic numeral system

based on 10 symbols

numbers and their representations

- how to represent a big number with a short sequence of symbols?
- suppose to have at disposal 10 symbols $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

number

I

sequence of symbols representing this number

1

number

II

sequence of symbols representing this number

2

number

IIIIIIII

sequence of symbols representing this number

8

number

IIIIIIIII

sequence of symbols representing this number

10

number

sequence of symbols representing this number

20

number

sequence of symbols representing this number

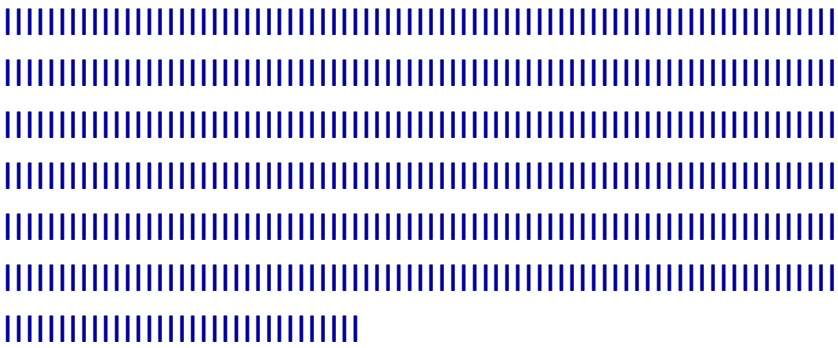
40

number

sequence of symbols representing this number

100

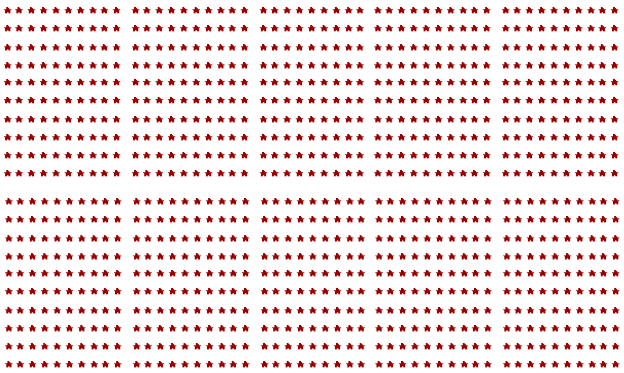
number



sequence of symbols representing this number

500

number



sequence of symbols representing this number

1000

number

difficult to show, since this screen has only 2073600 pixels

sequence of symbols representing this number

10000000

how many symbols?

- suppose to have at disposal 10 symbols $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, in order to represent number n a sequence of $O(\log_{10} n)$ symbols is enough