## CS1110 24 Nov 2009 Ragged arrays

Reading for today: sec. 9.3.
Reading for next time: chapter 16, applications and applets

• **If you have a conflict with the final exam**, please email Maria Witlox, mwitlox@cs.cornell.edu by 9am Tuesday Dec 2 with your name, netID, and what the conflict is with. *The final exam is Monday Dec 14th, 7-9:30pm, Baker Lab 200.*
• Due to Thanksgiving break, there are no labs this week (Nov 24 and 25), and check the webpage for changes in office and consultant hours.
• There *are* labs next week, Dec 1 and 2.
•Assignment A7 is due Friday Dec 4.



## Some notes regarding the CS1110 academic integrity policies

(http://www.cs.cornell.edu/courses/cs1110/2009fa/integrity.html)

• A6 amnesty petitions accepted until 5pm today. *Please note that except for specific questions about generic policy, we will not contact students regarding A6 petitions/violations until some time next week.* (Prof. Gries and I *are* currently working very hard on the situation, but want to do a final review of all cases together once again at the end.)

• For now (more in-depth discussion next time):
The most frequently asked question: what kind of "working together" is allowed?
  • principle from the website: don't use unauthorized assistance, and don't give fraudulent assistance.
  • principle from the website: You [meaning you and your partner, if you have grouped on CMS] may discuss work with other students. However, cooperation should never involve other students possessing a copy of all, or a portion of, your work *regardless of format*.

Some rules of thumb:
  • Don't look at any of other people's code.
  • Don't show other people any of your code.
  • OK to talk about algorithms you developed, but not at the level of essentially verbalizing code.



## Review of two-dimensional arrays

Type of d is **int**[][]

("**int** array array"/ "an array of **int** arrays")

To declare variable d:

**int** d[][];

To create a new array and assign it to d:

d= **new int**[5][4];

or, using an array initializer,

d= **new int**[][]{ {5,4,7,3}, {4,8,9,7}, {5,1,2,3}, {4,1,2,9}, {6,7,8,0} };

d

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 5 | 4 | 7 | 3 |
| 1 | 4 | 8 | 9 | 7 |
| 2 | 5 | 1 | 2 | 3 |
| 3 | 4 | 1 | 2 | 9 |
| 4 | 6 | 7 | 8 | 0 |

Some mysteries: an odd asymmetry, and strange toString output (see demo).

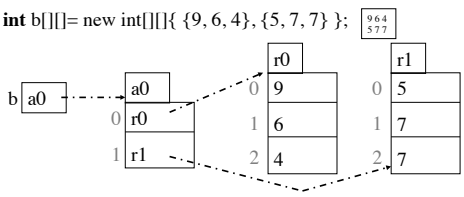
Number of rows of d: d.length

Number of columns in row r of d: d[r].length



## How multi-dimensional arrays are stored: arrays of arrays

**int** b[][]= new int[][]{ {9, 6, 4}, {5, 7, 7} };




b holds the name of a one-dimensional array object with b.length elements; its elements are the names of 1D arrays.

b[i] holds the name of a 1D array of **int**s of length b[i].length

java.util.Arrays.deepToString recursively creates an appropriate String.

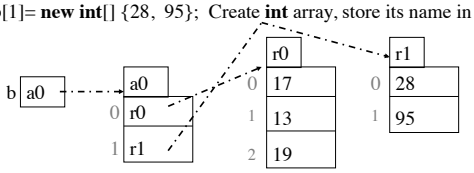
## Ragged arrays: rows have different lengths

**int**[][] b; Declare variable b of type **int**[][]

b= **new int**[2][] Create a 1-D array of length 2 and store its name in b. Its elements have type **int**[] (and start as **null**).

b[0]= **new int**[] {17, 13, 19}; Create **int** array, store its name in b[0].

b[1]= **new int**[] {28, 95}; Create **int** array, store its name in b[1].




5

## Application: recommender systems

Large collections of *association data* abound, but often, many possible associations have the default value, so the data is *sparse.*

*Netflix data*: (user, movie, score): 480K × 18K = 8.6B possible scores to track, but there are only (!) 100M actual scores.

*GroupLens data* (freely distributed by U. Minn): the small set has 943×1682= 1.5M possibilities, but only 100K actual scores.

How might Netflix, Amazon, etc. use this kind of association data to generate recommendations?

1. Represent each user by an array of movie ratings
2. Find similar users according to the similarity of the corresponding arrays, and report their favorite movies

This seems to suggest a 2-D, user-by-movie array.

6

## Recommender-system application (cont.)

*GroupLens data* (freely distributed by U. Minn): the small set has 943×1682= 1.5M possibilities, but only 100K actual scores.

Main idea:
For each user, DON'T store an **int** array of length 1682;
store a movie-sorted array of objects corresponding to the ratings for just the movies that user saw (avg. length: 59!).

This means a 2-D *ragged* user/movie array.

Another very useful technique (among many more substantive ones; take more CS courses!): map the movie/rater names to ints, b/c they can be meaningful array indices.

7