Structured Data

INFO/CSE 100, Spring 2006
Fluency in Information Technology

http://www.cs.washington.edu/100
Midterm2 Review

• The terms `index`, `myHeight`, and `dotWidth` are valid variable names in Javascript -- True or False?!?
2. Consider this short block of Javascript code. Assume that the code has executed successfully.

```javascript
var k = 4;
var grains = [1, 2, 4, 8, 16, 32, 64, 128];
var calculated = (grains.length >= 64);
var lastPayment = 0;

if (k < grains.length) {
    lastPayment = grains[4];
} else {
    lastPayment = undefined;
}
```

```javascript
grains.length = 8
calculated = false
lastPayment = 16
```
Midterm2 Revisited

```javascript
var loopCount = 2;
for (var i=0; i<loopCount; i++) {
    document.write("Loop "+i+"?"ัญ);
}
```

- Is the body of this loop executed?
- What is printed out after the first iteration?
  
  » Loop 0

- How many times does the loop execute?
Midterm2 Revisited

• Writing the Clamp function
  » Constrain a list of numbers into a range

```javascript
function clamp(low, high, values) {
    for (var i = 0; i < values.length; i++) {
        if (values[i] < low) {
            values[i] = low;
        } else if (values[i] > high) {
            values[i] = high;
        }
    }
}
```
Readings and References

• Reading
  » Fluency with Information Technology
    □ Chapter 13, Introduction to Spreadsheets

• References
  » Access Database: Design and Programming
    □ by Steve Roman, published by O'Reilly
Keeping track of things

• The need for keeping track of items spurned the invention of writing
• Today people still manually keep track of items usually in the form of lists
  » Shopping list
  » Christmas card addresses
  » Soccer team player roster
  » Runs Batted In (RBIs)
### Download of Variation Data (Single File)

**Global Prettybase Files**

This is a tab delimited text file in our "prettybase" format, which describes all SNP sites discovered by the SeattleSNPs PGA. The format of this file is:

**Line format:**
- `< chromosome position-chromosome-HUGO_NAME > <PGA Sample ID> < Allele1 > < Allele2 >`

**Example:** 74772592-10-FLAU D001 G T

The 'chromosome position' is generated from mapping to the most recent genome assembly available from the [UCSC Genome Assembly](https://genome.ucsc.edu/)

<table>
<thead>
<tr>
<th>Chromosome Position</th>
<th>Sample ID</th>
<th>Allele 1</th>
<th>Allele 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1100322-IL3RA-X</td>
<td>D001</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D002</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D003</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D004</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D005</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D006</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D007</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D008</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D009</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D010</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D011</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D012</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D013</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D014</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D015</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D016</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D033</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D034</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D035</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D036</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D037</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D038</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D039</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1100322-IL3RA-X</td>
<td>D040</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Spreadsheets

• Spreadsheets are a powerful abstraction for organizing data and computation

• A spreadsheet is a 2-dimensional array of cells... Its 3D with multiple worksheets
  » The idea is that the rows or columns represent a common kind of data
    □ They will be operated upon similarly
    □ Adding more data of the same type means adding more rows or columns
    □ Often spreadsheets contain numbers, but text-only spreadsheets are useful too!
Looking for Similar Ideas

• Spreadsheets are not so unusual...
  » The position (row/column) names the data, as with memory locations, variables, forms...
  » Operating on all elements of a column (or row) is an iteration (although not using a world famous iteration!)
  » Setting a cell to a formula is an (unevaluated) assignment statement with cells as variables
  » The formula is an expression
  » Functions are built-in spreadsheet programs
Familiar Terminology

- **row name**
- **column name**
- **formula**
- **cell**
- **column heading**
- **references cells L2-W2**
Formulas

![Excel spreadsheet screenshot](image)

<table>
<thead>
<tr>
<th>Common Name</th>
<th>Distance (km)</th>
<th>Body Length</th>
<th>Distance (mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swainson' Hawk</td>
<td>13500</td>
<td>0.52</td>
<td>8383.5</td>
</tr>
<tr>
<td>Wheatear</td>
<td>13500</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Willow Warbler</td>
<td>15000</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Short-tailed Shearwater</td>
<td>12500</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Long-tailed Skua</td>
<td>16000</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Arctic Tern</td>
<td>19000</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>
Using Fill
Relative and Absolute Addresses

• References to cells happen in one of two ways.. Relative or Absolute
  » F2  relative column, relative row
  » F$2  relative column, absolute row
  » $F2  absolute column, relative row
  » $F$2  absolute column, absolute row

• Relative references change when pasted/filled
• Absolute references do not change
Series

• Another handy feature of fill is that it can make it easy to make a series based on constraints
  » Fill Sunday=>Monday, Tuesday, Wednesday, ..
  » Fill 22 Feb=>23 Feb, 24 Feb, 25 Feb, ...

• More generally
  » Series fill will even count using a constant
  » Counting by odd sizes: gives 1st two items
Sorting Data

• Sorting the data into some order is one of the most common operations
  » Numbers go numerically
  » Text goes alphabetically

• Data can be sorted in Ascending or Descending order

• Data can be sorted in second, third, or fourth order...
  » First one column, then the second column and so on...
Sort Example

[Excel spreadsheet with data sorted by Common Name in ascending order.]

- Common Name
- Distance (km)
- Body Length
- Distance

Sorted list:
- Swainson' Hawk: 13500, 0.52
- Wheatear: 13500, 0.16
- Willow Warbler: 15000, 0.11
- Short-tailed Shearwater: 12500, 0.43
- Long-tailed Skua: 16000, 0.51
- Arctic Tern: 19000, 0.35
Adding Functions

Calculates the payment for a loan based on constant payments and a constant interest rate.

Pv is the present value: the total amount that a series of future payments is worth now.
Importing/Exporting Data

- Importing data is one of the most common ways to create a spreadsheet
- Two ways to import data
  - Copy/paste
  - Import function
- Spreadsheets will do a lot of work to interpret data into a table format for importing
  - Import data from a text file
  - Import data from a web query
  - Among others...
Import Wizard

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the Data Type that best describes your data.

Original data type

Choose the file type that best describes your data:
- Delimited - Characters such as commas or tabs separate each field.
- Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1
File origin: Macintosh


1 | Age, Male, Female
2 | Under 5 years, "282.065", "192.241"
3 | 5 to 9 years, "218,501", "207,408"
4 | 10 to 14 years, "222,937", "211,899"
5 | 15 to 19 years, "220,412", "207,556"
6 | 20 to 24 years, "200,812", "189,373"
How to organize the data?

• Lists and Spreadsheets are often known as “flat files” (although a good 2D/3D spreadsheet isn't really flat)

• Common problems with the flat file format
  » Structural information is difficult to express
  » All processing of information is “special cased”
    □ custom programs are needed
  » Information repeated; difficult to combine
  » Changes in format of one file means all programs that ever process that file must be changed
    □ eg, adding ZIP codes
Library example

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Title</th>
<th>AuID</th>
<th>AuName</th>
<th>AuPhone</th>
<th>PubID</th>
<th>PubName</th>
<th>PubPhone</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1111-1111-1</td>
<td>C++</td>
<td>4</td>
<td>Roman</td>
<td>444-444-4444</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$29.95</td>
</tr>
<tr>
<td>0-99-999999-9</td>
<td>Emma</td>
<td>1</td>
<td>Austen</td>
<td>111-111-1111</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$20.00</td>
</tr>
<tr>
<td>0-91-355678-7</td>
<td>Faerie Queene</td>
<td>7</td>
<td>Spencer</td>
<td>777-777-7777</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$15.00</td>
</tr>
<tr>
<td>0-91-345678-5</td>
<td>Hamlet</td>
<td>3</td>
<td>Shakespeare</td>
<td>555-555-5555</td>
<td>2</td>
<td>Alpha Press</td>
<td>999-999-9999</td>
<td>$20.00</td>
</tr>
<tr>
<td>0-103-45678-9</td>
<td>Iliad</td>
<td>3</td>
<td>Homer</td>
<td>333-333-3333</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$25.00</td>
</tr>
<tr>
<td>0-12-345678-6</td>
<td>Jane Eyre</td>
<td>1</td>
<td>Austen</td>
<td>111-111-1111</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$49.00</td>
</tr>
<tr>
<td>0-99-777777-7</td>
<td>King Lear</td>
<td>5</td>
<td>Shakespeare</td>
<td>555-555-5555</td>
<td>2</td>
<td>Alpha Press</td>
<td>999-999-9999</td>
<td>$49.00</td>
</tr>
<tr>
<td>0-555-55555-5</td>
<td>Macbeth</td>
<td>5</td>
<td>Shakespeare</td>
<td>555-555-5555</td>
<td>2</td>
<td>Alpha Press</td>
<td>999-999-9999</td>
<td>$12.00</td>
</tr>
<tr>
<td>0-11-345678-9</td>
<td>Moby Dick</td>
<td>2</td>
<td>Melville</td>
<td>222-222-2222</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$49.00</td>
</tr>
<tr>
<td>0-12-333433-3</td>
<td>On Liberty</td>
<td>8</td>
<td>Mill</td>
<td>888-888-8888</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$25.00</td>
</tr>
<tr>
<td>0-321-32132-1</td>
<td>Balloon</td>
<td>13</td>
<td>Sleepy</td>
<td>321-321-1111</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$34.00</td>
</tr>
<tr>
<td>0-321-32132-1</td>
<td>Balloon</td>
<td>11</td>
<td>Snoopy</td>
<td>321-321-2222</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$34.00</td>
</tr>
<tr>
<td>0-321-32132-1</td>
<td>Balloon</td>
<td>12</td>
<td>Grumpy</td>
<td>321-321-0000</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$34.00</td>
</tr>
<tr>
<td>0-55-123456-9</td>
<td>MainStreet</td>
<td>10</td>
<td>Jones</td>
<td>123-333-3333</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$22.95</td>
</tr>
<tr>
<td>0-55-123456-9</td>
<td>MainStreet</td>
<td>9</td>
<td>Smith</td>
<td>122-222-2222</td>
<td>3</td>
<td>Small House</td>
<td>714-000-0000</td>
<td>$22.95</td>
</tr>
<tr>
<td>0-123-45678-0</td>
<td>Ulysses</td>
<td>6</td>
<td>Joyce</td>
<td>666-666-6666</td>
<td>2</td>
<td>Alpha Press</td>
<td>999-999-9999</td>
<td>$34.00</td>
</tr>
<tr>
<td>1-22-233700-0</td>
<td>Visual Basic</td>
<td>4</td>
<td>Roman</td>
<td>444-444-4444</td>
<td>1</td>
<td>Big House</td>
<td>123-456-7890</td>
<td>$25.00</td>
</tr>
</tbody>
</table>

notice the redundancy

from Access Database book, Steve Roman
Why Study Databases?

• Databases solve those "flat file" problems
• Some of us want to compute
• All of us want access to information …
  □ Much of the archived information is in tables
  □ Databases enhance applications, e.g. Web
  □ Once you know how to create databases, you can use them to personal advantage
  □ Databases introduce interesting ideas

The Internet Movie Database
Visitied by over 20 million movie lovers each month!
Welcome to the Internet Movie Database, the biggest, best, most award-winning movie site on the planet.