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?!?

Midterm2 Revisited

2. Consider this short block of Javascript code. Assume that the code has executed successfully.

```
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;
}
```

grains.length = 8

calculated = false

lastPayment = 16

Midterm2 Revisited

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

- 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

```
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;
       }
   }
}
```

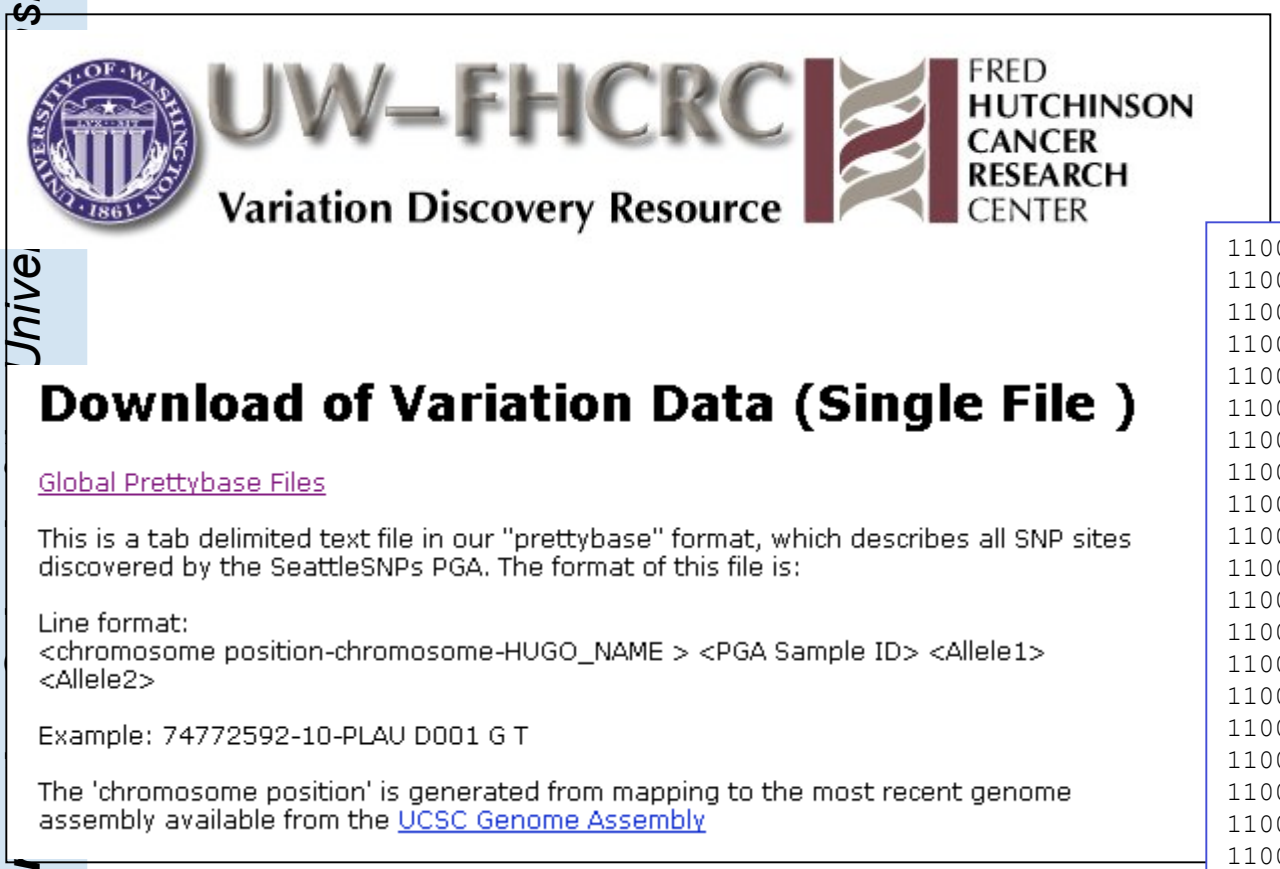
Readings and References

- 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)

tab-delimited file example



UW–FHCRC Variation Discovery Resource

FRED HUTCHINSON CANCER RESEARCH CENTER

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-PLAU D001 G T

The 'chromosome position' is generated from mapping to the most recent genome assembly available from the UCSC Genome Assembly

```
1100322-IL3RA-X      D001    N  N
1100322-IL3RA-X      D002    G  G
1100322-IL3RA-X      D003    G  G
1100322-IL3RA-X      D004    G  G
1100322-IL3RA-X      D005    G  G
1100322-IL3RA-X      D006    G  G
1100322-IL3RA-X      D007    G  G
1100322-IL3RA-X      D008    G  G
1100322-IL3RA-X      D009    A  G
1100322-IL3RA-X      D010    N  N
1100322-IL3RA-X      D011    N  N
1100322-IL3RA-X      D012    N  N
1100322-IL3RA-X      D013    G  G
1100322-IL3RA-X      D014    A  G
1100322-IL3RA-X      D015    N  N
1100322-IL3RA-X      D016    N  N
1100322-IL3RA-X      D033    A  G
1100322-IL3RA-X      D034    A  G
1100322-IL3RA-X      D035    G  G
1100322-IL3RA-X      D036    A  G
1100322-IL3RA-X      D037    A  A
1100322-IL3RA-X      D038    G  G
1100322-IL3RA-X      D039    G  G
1100322-IL3RA-X      D040    G  G
...
```

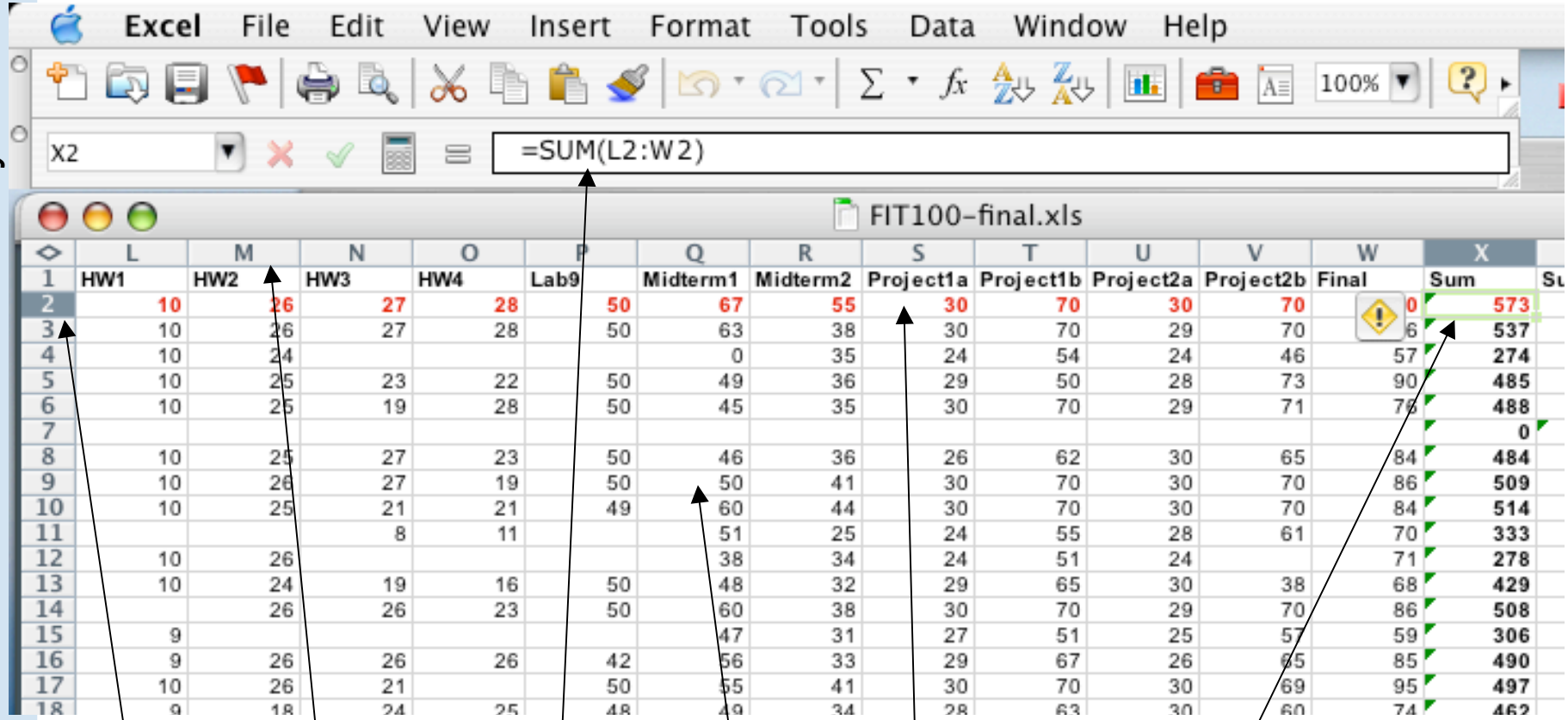
Spreadsheets

- Spreadsheets are a powerful abstration for organizing data and compution
- 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



	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	HW1	HW2	HW3	HW4	Lab9	Midterm1	Midterm2	Project1a	Project1b	Project2a	Project2b	Final	Sum
2	10	26	27	28	50	67	55	30	70	30	70	0	573
3	10	26	27	28	50	63	38	30	70	29	70	6	537
4	10	24				0	35	24	54	24	46	57	274
5	10	25	23	22	50	49	36	29	50	28	73	90	485
6	10	25	19	28	50	45	35	30	70	29	71	76	488
7													0
8	10	25	27	23	50	46	36	26	62	30	65	84	484
9	10	26	27	19	50	50	41	30	70	30	70	86	509
10	10	25	21	21	49	60	44	30	70	30	70	84	514
11			8	11		51	25	24	55	28	61	70	333
12	10	26				38	34	24	51	24		71	278
13	10	24	19	16	50	48	32	29	65	30	38	68	429
14		26	26	23	50	60	38	30	70	29	70	86	508
15	9					47	31	27	51	25	57	59	306
16	9	26	26	26	42	56	33	29	67	26	65	85	490
17	10	26	21		50	55	41	30	70	30	69	95	497
18	9	18	24	25	48	49	34	28	63	30	60	74	462

row name

column name

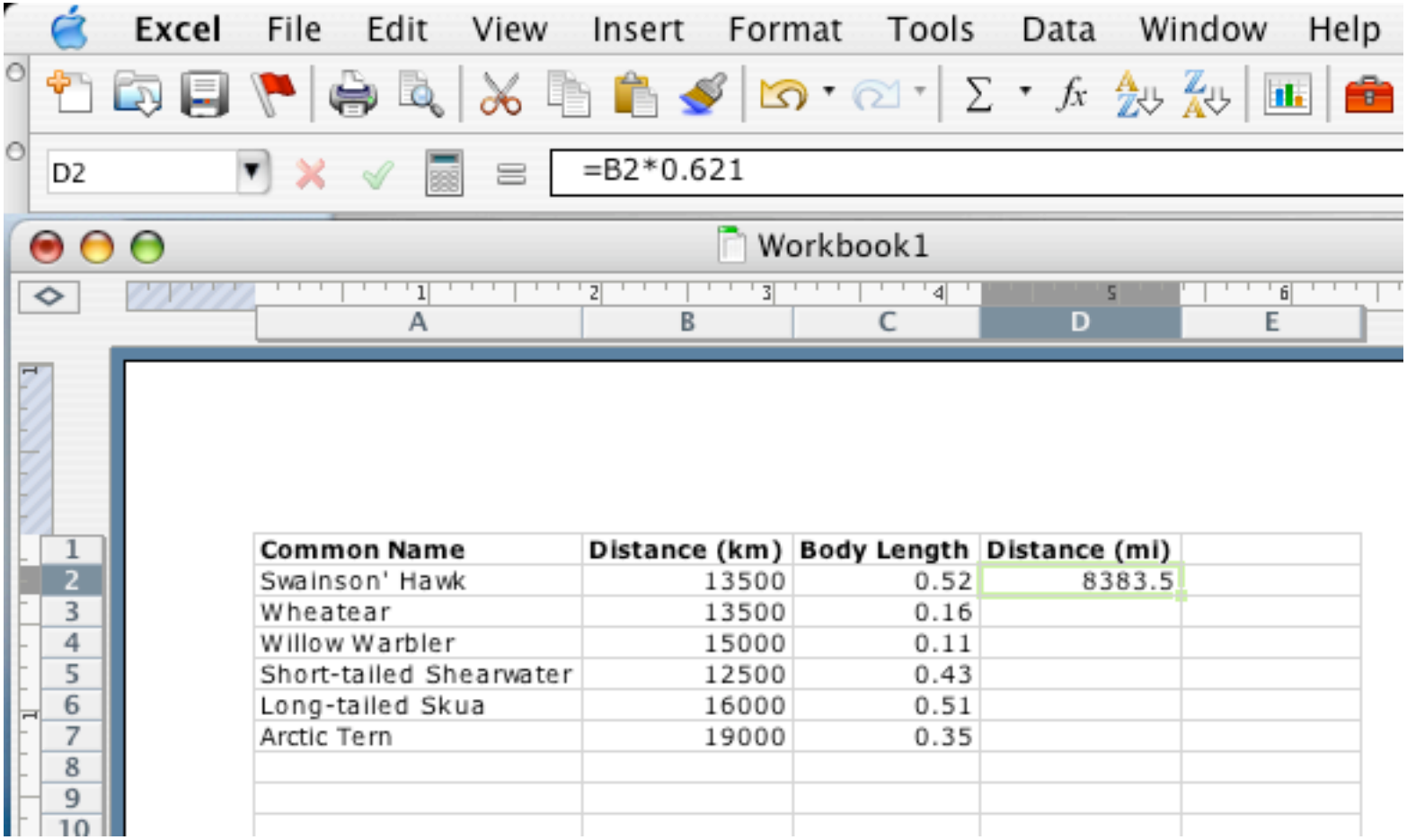
formula

cell

column heading

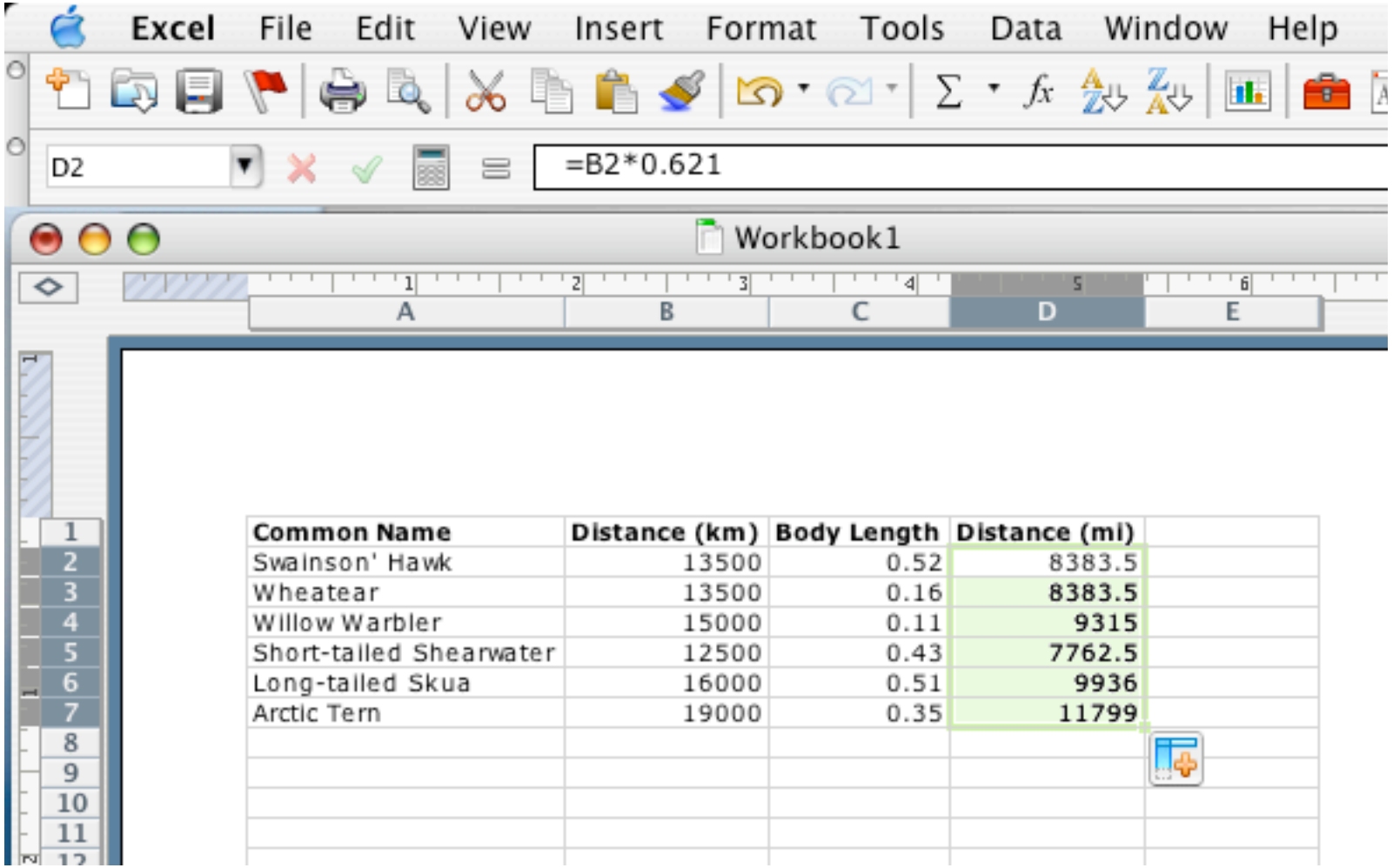
references cells L2-W2

Formulas



Common Name	Distance (km)	Body Length	Distance (mi)
Swainson' Hawk	13500	0.52	8383.5
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	

Using Fill



Common Name	Distance (km)	Body Length	Distance (mi)
Swainson' Hawk	13500	0.52	8383.5
Wheatear	13500	0.16	8383.5
Willow Warbler	15000	0.11	9315
Short-tailed Shearwater	12500	0.43	7762.5
Long-tailed Skua	16000	0.51	9936
Arctic Tern	19000	0.35	11799

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
  - » F2 absolute column, absolute row
- Relative references change when pasted/filled
- Absolute references do not change

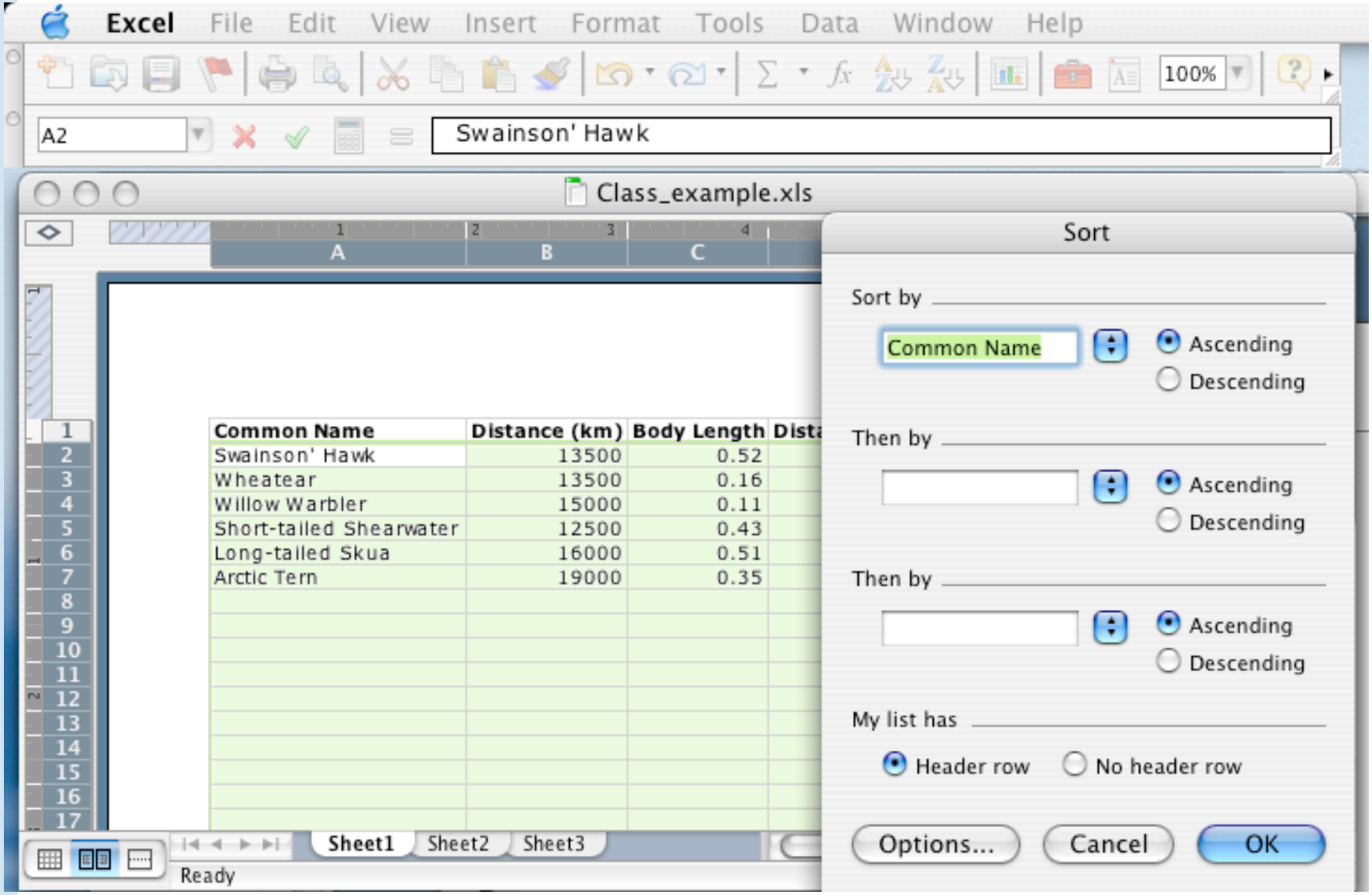
Series

- 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

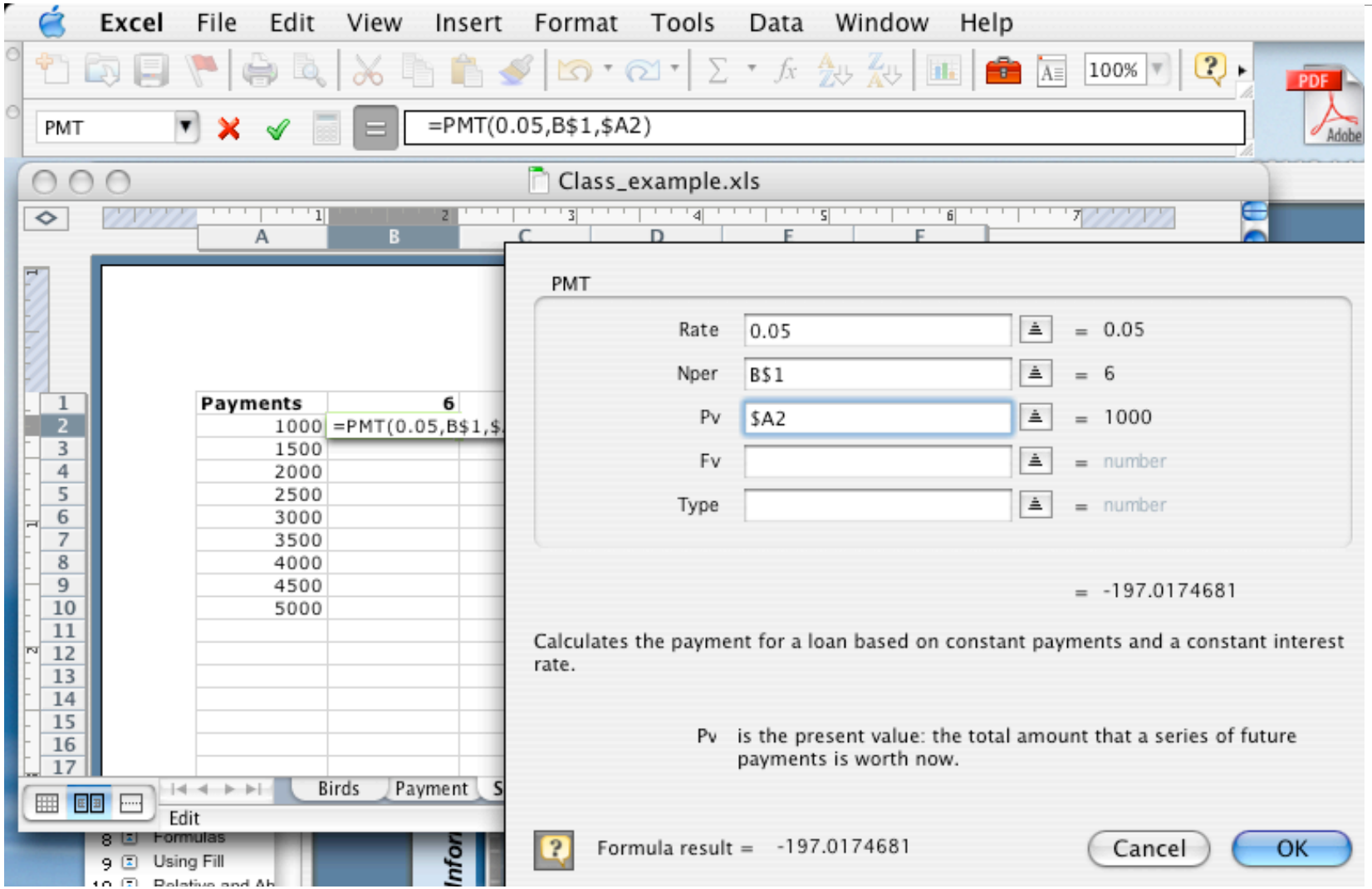
Sorting Data

- 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



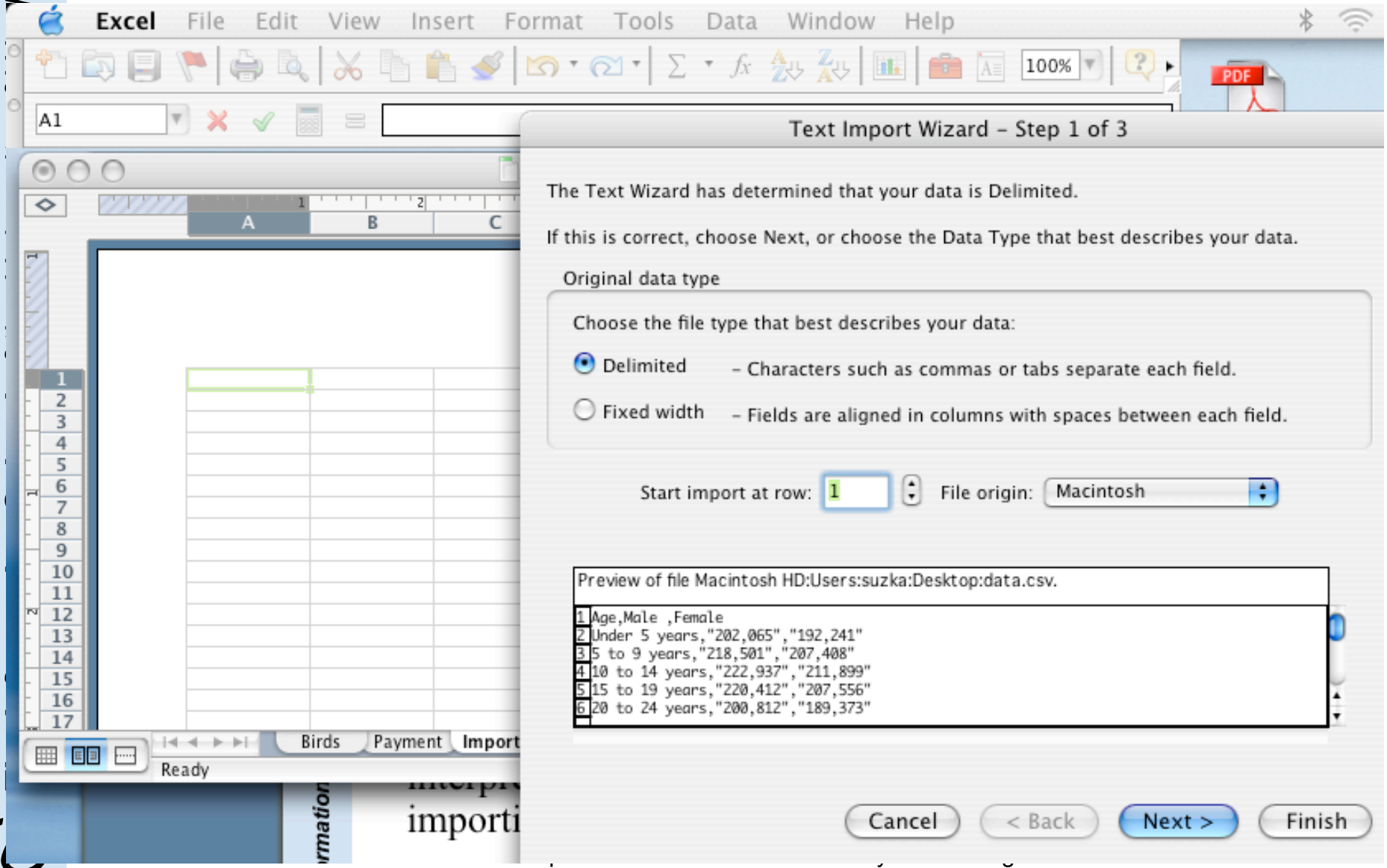
Adding Functions



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



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

notice the redundancy

ISBN	Title	AuID	AuName	AuPhone	PubID	PubName	PubPhone	Price
1-1111-1111-1	C++	4	Roman	444-444-4444	1	Big House	123-456-7890	$29.95
0-99-999999-9	Emma	1	Austen	111-111-1111	1	Big House	123-456-7890	$20.00
0-91-335678-7	Fairie Queene	7	Spencer	777-777-7777	1	Big House	123-456-7890	$15.00
0-91-045678-5	Hamlet	5	Shakespeare	555-555-5555	2	Alpha Press	999-999-9999	$20.00
0-103-45678-9	Iliad	3	Homer	333-333-3333	1	Big House	123-456-7890	$25.00
0-12-345678-6	Jane Eyre	1	Austen	111-111-1111	3	Small House	714-000-0000	$49.00
0-99-777777-7	King Lear	5	Shakespeare	555-555-5555	2	Alpha Press	999-999-9999	$49.00
0-555-55555-9	Macbeth	5	Shakespeare	555-555-5555	2	Alpha Press	999-999-9999	$12.00
0-11-345678-9	Moby Dick	2	Melville	222-222-2222	3	Small House	714-000-0000	$49.00
0-12-333433-3	On Liberty	8	Mill	888-888-8888	1	Big House	123-456-7890	$25.00
0-321-32132-1	Balloon	13	Sleepy	321-321-1111	3	Small House	714-000-0000	$34.00
0-321-32132-1	Balloon	11	Snoopy	321-321-2222	3	Small House	714-000-0000	$34.00
0-321-32132-1	Balloon	12	Grumpy	321-321-0000	3	Small House	714-000-0000	$34.00
0-55-123456-9	Main Street	10	Jones	123-333-3333	3	Small House	714-000-0000	$22.95
0-55-123456-9	Main Street	9	Smith	123-222-2222	3	Small House	714-000-0000	$22.95
0-123-45678-0	Ulysses	6	Joyce	666-666-6666	2	Alpha Press	999-999-9999	$34.00
1-22-233700-0	Visual Basic	4	Roman	444-444-4444	1	Big House	123-456-7890	$25.00

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

Visited 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.