

Relational databases

Problems they solve
Strengths and weaknesses

Managing data

- We have already learned that choices in how data are entered affect how easily it can be used
 - Stacked data is the best
 - PivotTables require stacked data
- Data management becomes a non-trivial issue when you have a lot of it
- Modern biology generates huge databases

What is a database?

- No fixed criteria – largely a matter of preference, intended purpose
 - Generally, things that are included in a database are interrelated in some way
 - Including data sets together in a database facilitates examination of the relationships between them
- Could be:
 - All the data collected from a single experiment
 - All the data collected from a single study (multiple experiments)
 - All the DNA sequences for all the genes in the human genome
 - All the DNA sequences for proteins across all organisms

Database Management Systems (DBMS's)

- A DBMS is a program specifically designed to create, modify, and use a database
- DBMS's organize data in **tables**
 - Stacked data format
 - Each column is a variable, or **field**
 - Each row is a **record** = all known information about a single data point
- If all data are kept together in one big table, then the database is a **flat file** database
- If different sets of inter-related information is kept in separate tables, then a **relational** database is used (RDBMS)
- Although Excel worksheets are flat file data it is a spreadsheet, not an RDBMS
 - DBMS has some advantages over a spreadsheet program, even for flat file data
 - Excel has very limited ability to work with relational data
- The MS desktop RDBMS program is MS Access

Many different data types can be held in a database

- Obviously, text and numbers
- Less obviously
 - DNA sequences
 - GIS vector data (geometric shapes = points, lines, polygons)
 - Binary large objects (BLOBs)
 - Sound files
 - Images
 - Video
 - pdf files of journal articles

Flat files

- Flat files are a stacked data arrangement, with all of the data in a single file
- Either a spreadsheet or a DBMS can manage flat files
 - DBMS's are less flexible than spreadsheets, which can be a **good** thing for managing data
 - Lose some of the DBMS advantages by using Excel, but gain the ability to use PivotTables, graphs, etc.
- Problem with flat file format is redundancy (which fields are redundant?) → becomes a problem with large databases

	A	B	C	D	E	F
1	Site ID	County	Elevation	Annual rainfall	Soil sample ID	Nitrogen
2	Sky Oaks	San Diego	1418	53	1	17.0
3	Sky Oaks	San Diego	1418	53	2	17.6
4	Sky Oaks	San Diego	1418	53	3	13.8
5	Sky Oaks	San Diego	1418	53	4	17.0
6	Sky Oaks	San Diego	1418	53	5	11.4
7	Sky Oaks	San Diego	1418	53	6	19.8
8	Sky Oaks	San Diego	1418	53	7	16.4
9	Sky Oaks	San Diego	1418	53	8	21.0
10	Sky Oaks	San Diego	1418	53	9	15.6
11	Sky Oaks	San Diego	1418	53	10	19.6
12	Sky Oaks	San Diego	1418	53	11	13.9
13	Santa Margarita	Riverside	338	36	1	19.0
14	Santa Margarita	Riverside	338	36	2	19.0
15	Santa Margarita	Riverside	338	36	3	16.1
16	Santa Margarita	Riverside	338	36	4	22.2
17	Santa Margarita	Riverside	338	36	5	22.4
18	Santa Margarita	Riverside	338	36	6	16.5
19	Santa Margarita	Riverside	338	36	7	17.2
20	Santa Margarita	Riverside	338	36	8	13.4
21	Santa Margarita	Riverside	338	36	9	21.0
22	Santa Margarita	Riverside	338	36	10	16.3
23	Santa Margarita	Riverside	338	36	11	16.8

Problems avoided by using a DBMS instead of Excel

- Data accuracy problems
 - Better data entry
- Data management problems

Advantages of DBMS: data entry

- Fields have to be assigned a data type – avoids certain types of data entry errors
 - Entries have to match the data type of the field
 - Most database programs can further restrict valid entries (have to fall within a numeric range, picked from a list, etc.)
- Data entry forms can be used
 - Look like the paper forms used to originally record the data → fewer entry errors
 - Even if data is entered directly into the database a good data entry form prompts the user for the needed fields → less missing data

Data types in spreadsheets: flexibility = hazard

Including a text label makes these cells text

Sample ID	Nitrogen	Nitrogen (mg)	Nitrogen (mg)
1	10 mg	10	10 mg
2	12 mg	12	12
3	9 mg	9	9
Average	#DIV/0!	10.33	10.50

One text cell included with two numeric cells

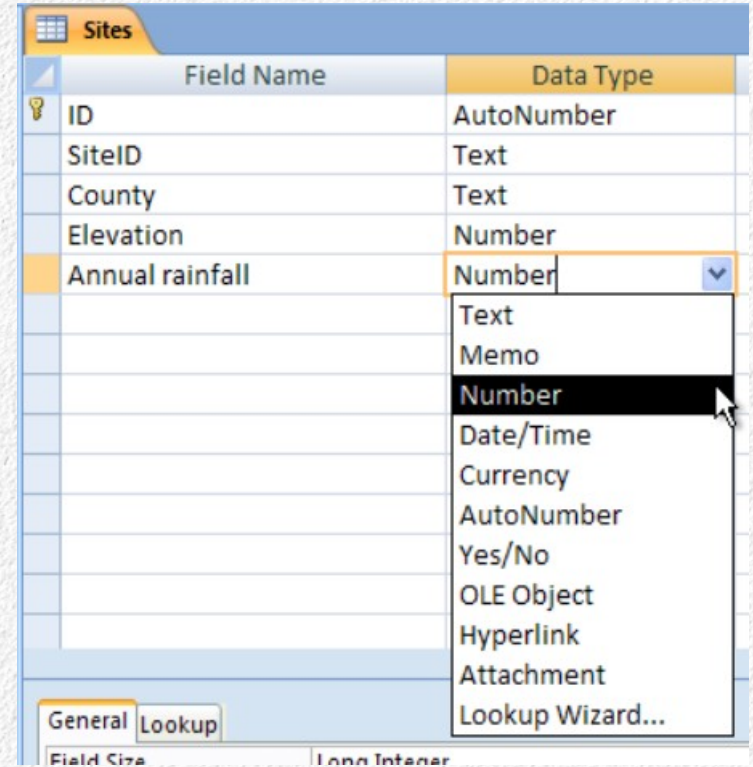
Obvious mistake

Mistake, but not obvious

- Spreadsheets do treat different variable types differently
- But, spreadsheet data types assigned to cells, not columns
- Mixing different variable types in a column is allowed
- Calculations may not be accurate if they are mixed, and the mistakes may not be obvious

Variable types in Access

- Columns in databases are called **fields**
- Fields have a specific data type assigned to them when the table is created →
- Only data of the appropriate type can be entered in the field



Access field type definitions

Field type	Definition	Additional properties
Text	Combinations of text or numbers, but treated as non-numeric	Number of characters (up to 255)
Memo	Combinations of text or numbers, but treated as non-numeric	Up to 63,999 characters
Number	Numbers only (no text characters)	Number type (integer, long integer, single, double)
Date/time	Dates and times	Format
Currency	Money	
Autonumber	Numbers that automatically increase with each new record added	
Yes/No	Only contains yes or no (or equivalently, 1 or 0)	
OLE object	Any file type supported, can be linked or embedded in database	
Hyperlink	Link to a URL	
Attachment	External files that are stored in the database	

Avoiding typos

- Typographical errors (typos) are data entry errors
 - Hit the wrong key
 - Used different labels for the same thing
- RDBMS's help you avoid them
 - Variable types
 - Lookups

G	H	I	J
Original order	petiole.diam	type	
1	2.7	T	
2	2	T	
3	2.2	T	
4	1.7	T	
5	1.2	G	
6	1.9	Tree	
7	1.4	G	
8	2.5	T	
9	3.1	T	
10	1.7	T	
11	1.5	T	
12	1.8	T	
13	1.8	T	

How many different ways is tree represented in the type column?

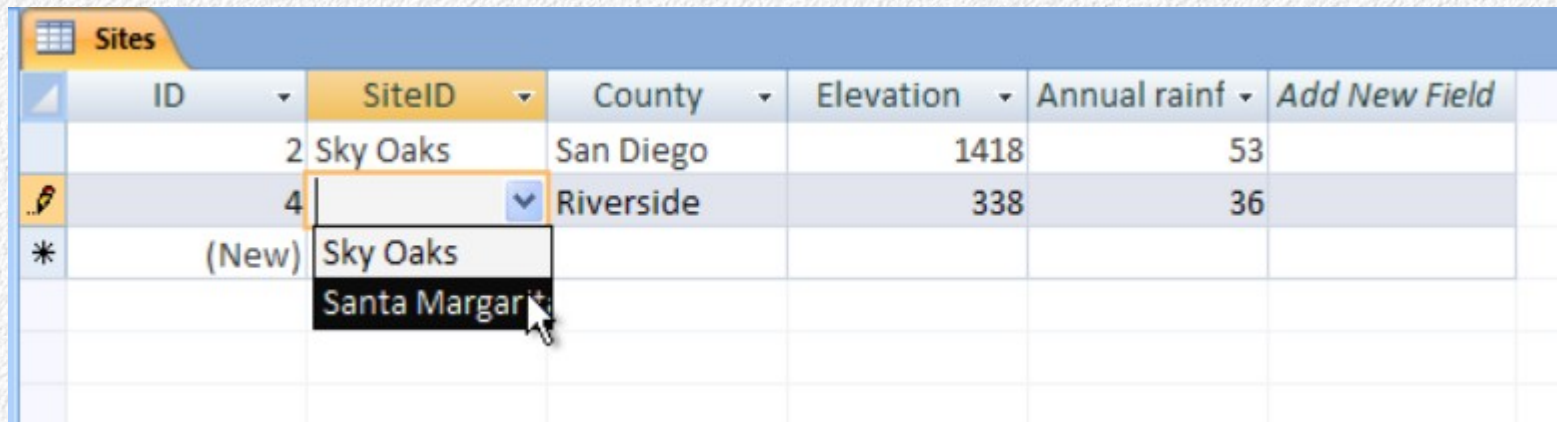
Why so many T's?

→

	A	B
1	type	Count - type
2	T	1
3	G	52
4	T	47
5	T	1
6	T	1
7	Tree	1
8	Total Result	103
9		

Lookups

- Computers are literal – Sky Oaks, Sky Oaks , SkyOaks, Sky Oaks, Sky Oakes, and Sky oaks are all different
- DBMS's allow you to limit entry to particular values
 - Can be selected from a drop-down list
 - No typos – only error possible is to select the wrong value from the list



The screenshot shows a database table named 'Sites' with the following columns: ID, SiteID, County, Elevation, Annual rainf, and Add New Field. The table contains three rows of data. The second row is selected, and a dropdown menu is open for the SiteID field, showing two options: 'Sky Oaks' and 'Santa Margarita'.

ID	SiteID	County	Elevation	Annual rainf	Add New Field
2	Sky Oaks	San Diego	1418	53	
4	<input type="text" value="Sky Oaks"/>	Riverside	338	36	
(New)	<input type="text" value="Sky Oaks"/>				

Enforcing completeness in data entry

- Missing information can be a big problem
 - Can cause the entire record to be unusable
- Can tell the DBMS to:
 - Automatically date/time stamp entries
 - Require fields to be filled in before a record is accepted
- Data forms make it easier for data entry to be done consistently
- Printed data forms can be designed to look the same as the form used on the computer for entry → easier to get everything entered in the correct place

Advantages of DBMS: data management

- The basic unit of a database is a record, not a cell
 - Can't accidentally scramble your data by sorting one field but not the others
- Large data sets can be stored with smaller file sizes (e.g. use a small integer data type instead of storing all numbers as floating point, like Excel does)
- Searching, filtering, subsetting data is easier, faster (indexes, queries)

Relational databases

- Relational Database Management Systems (RDBMS) organize data into two or more tables that can be joined based on one or more matching field
 - Access is an RDBMS
- Matching two or more tables based on columns of matching data is a **relational join**
 - Excel (spreadsheets) has a very limited ability to do these
 - MS Access is designed as a relational database
- Relationships can be:
 - One to one = one row in table 1 matches one row in table 2
 - One to many = one row in table 1 matches many rows in table 2
- Example: soil nitrogen measurements

One to one relationships

- Example – different measurements of the same soil sample made at different times, different locations

- Field measurements: location (lat, long), slope, aspect →

Point ID	Lat.	Long.	Slope	Aspect
1	33 ° 24'	116 ° 18'	11	14 °
2	33 ° 16'	116 ° 18'	8	127 °
3	33 ° 15'	116 ° 15'	11	99 °
4	33 ° 18'	116 ° 22'	11	168 °
5	33 ° 27'	116 ° 19'	7	249 °
6	33 ° 7'	116 ° 21'	9	210 °
7	33 ° 23'	116 ° 19'	14	219 °
8	33 ° 31'	116 ° 22'	10	234 °

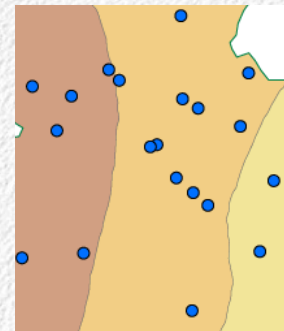
- Chemical characteristics of soil samples are measured in the lab

- Time since last fire is taken from a computer mapping program (a Geographic Information System, or GIS)

- We will have three different tables, one for each type of data – need to join them together

- Could copy/paste in Excel, but copy/paste errors are easy to make, hard to spot

Map of time since burn



Soil extraction in the lab



Tables joined by matching columns

	A	B	C	D	E
1	Point ID	Lat.	Long.	Slope	Aspect
2	1	33° 24'	116° 18'	11	14
3	2	33° 16'	116° 18'	8	127
4	3	33° 15'	116° 15'	11	99
5	4	33° 18'	116° 22'	11	168
6	5	33° 27'	116° 19'	7	249
7	6	33° 7'	116° 21'	9	210
8	7	33° 23'	116° 19'	14	219
9	8	33° 31'	116° 22'	10	234

	A	B	
1	Point ID	Time since burn	
2	1	22	
3	2	10	
4	3	38	
5	4	19	
6	5	32	
7	6	2	
8	7	22	
9	8	20	
10	9	2	
11	10	35	
12			

	A	B	C	D
1	Point ID	N	P	K
2	1	19.23	5.69	1.61
3	2	23.16	6.99	1.83
4	3	18.98	5.77	2.5
5	4	19.05	4.75	2.26
6	5	20.28	5.37	2.13
7	6	21.66	4.88	1.72
8	7	19.22	5.62	2.24
9	8	22.4	4.34	2.49
10				

	A	B	C	D	E	F	G	H	I
1	Point ID	Lat.	Long.	Slope	Aspect	Time since burn	N	P	K
2	1	33° 24'	116° 18'	11	14	22	19.23	5.69	1.61
3	2	33° 16'	116° 18'	8	127	10	23.16	6.99	1.83
4	3	33° 15'	116° 15'	11	99	38	18.98	5.77	2.5
5	4	33° 18'	116° 22'	11	168	19	19.05	4.75	2.26
6	5	33° 27'	116° 19'	7	249	32	20.28	5.37	2.13
7	6	33° 7'	116° 21'	9	210	2	21.66	4.88	1.72
8	7	33° 23'	116° 19'	14	219	22	19.22	5.62	2.24
9	8	33° 31'	116° 22'	10	234	20	22.4	4.34	2.49

Note: two records in the time since burn table are not in the field measurement or chemical analysis tables

Not a problem – can choose to drop any that don't match in all three tables, or include all of some tables and only matches of others...

Redundancy in data entry

- Redundancy is bad
 - More opportunities for data entry error
 - Larger file sizes for no good reason
- We can avoid redundant data entry by using a one to many relationship between tables

One to many relationship

- One table with site-level information ↓, one with sample-level information →

Site ID	County	Elevation	Annual rainfall
Sky Oaks	San Diego	1418	53
Santa Margarita	Riverside	338	36

- Joined together by one or more common fields (which is what here?)
- The match between fields of different tables defines the relationship between them

Site ID	Soil samp	Nitrogen
Sky Oaks	1	17.0
Sky Oaks	2	17.6
Sky Oaks	3	13.8
Sky Oaks	4	17.0
Sky Oaks	5	11.4
Sky Oaks	6	19.8
Sky Oaks	7	16.4
Sky Oaks	8	21.0
Sky Oaks	9	15.6
Sky Oaks	10	19.6
Sky Oaks	11	13.9
Santa Mar	1	19.0
Santa Mar	2	19.0
Santa Mar	3	16.1
Santa Mar	4	22.2
Santa Mar	5	22.4
Santa Mar	6	16.5
Santa Mar	7	17.2
Santa Mar	8	13.4
Santa Mar	9	21.0
Santa Mar	10	16.3
Santa Mar	11	16.8

Joined

	A	B	C	D	E	F
1	Site ID	County	Elevation	Annual rainfall	Soil sample ID	Nitrogen
2	Sky Oaks	San Diego	1418	53	1	17.0
3	Sky Oaks	San Diego	1418	53	2	17.6
4	Sky Oaks	San Diego	1418	53	3	13.8
5	Sky Oaks	San Diego	1418	53	4	17.0
6	Sky Oaks	San Diego	1418	53	5	11.4
7	Sky Oaks	San Diego	1418	53	6	19.8
8	Sky Oaks	San Diego	1418	53	7	16.4
9	Sky Oaks	San Diego	1418	53	8	21.0
10	Sky Oaks	San Diego	1418	53	9	15.6
11	Sky Oaks	San Diego	1418	53	10	19.6
12	Sky Oaks	San Diego	1418	53	11	13.9
13	Santa Margarita	Riverside	338	36	1	19.0
14	Santa Margarita	Riverside	338	36	2	19.0
15	Santa Margarita	Riverside	338	36	3	16.1
16	Santa Margarita	Riverside	338	36	4	22.2
17	Santa Margarita	Riverside	338	36	5	22.4
18	Santa Margarita	Riverside	338	36	6	16.5
19	Santa Margarita	Riverside	338	36	7	17.2
20	Santa Margarita	Riverside	338	36	8	13.4
21	Santa Margarita	Riverside	338	36	9	21.0
22	Santa Margarita	Riverside	338	36	10	16.3
23	Santa Margarita	Riverside	338	36	11	16.8

Using joined tables – queries

- **Queries** can combine information across tables, filter based on field in either table without making duplicates
 - Act as an actual table – queries can be used as a table in other queries
 - Once constructed, queries can be saved for later use
- Queries don't make a separate copy of the data in the table
 - Prevents redundant copying of the data → less space used
 - Any changes in a table are automatically reflected in all queries that use it → no need to track down files with subsets of the table's data to apply the change everywhere

Site, soil nitrogen, vegetation

Sites

Site ID	County	Elevation	Annual rainfall
Sky Oaks	San Diego	1418	53
Santa Margarita	Riverside	338	36

SoilSamples

Site ID	Soil sample ID	Nitrogen
Sky Oaks	1	17.0
Sky Oaks	2	17.6
Sky Oaks	3	13.8
Sky Oaks	4	17.0
Sky Oaks	5	11.4
Sky Oaks	6	19.8
Sky Oaks	7	16.4
Sky Oaks	8	21.0
Sky Oaks	9	15.6
Sky Oaks	10	19.6
Sky Oaks	11	13.9
Santa Mar	1	19.0
Santa Mar	2	19.0
Santa Mar	3	16.1
Santa Mar	4	22.2
Santa Mar	5	22.4
Santa Mar	6	16.5
Santa Mar	7	17.2
Santa Mar	8	13.4
Santa Mar	9	21.0
Santa Mar	10	16.3
Santa Mar	11	16.8

Vegetation

Site ID	Soil sample ID	Vegetation type
Sky Oaks	1	CSS
Sky Oaks	2	Chaparral
Sky Oaks	3	Riparian
Sky Oaks	4	Riparian
Sky Oaks	5	CSS
Sky Oaks	6	CSS
Sky Oaks	7	Chaparral
Sky Oaks	8	Chaparral
Sky Oaks	9	Chaparral
Sky Oaks	10	Riparian
Sky Oaks	11	Riparian
Santa Margarita	1	Riparian
Santa Margarita	2	Chaparral
Santa Margarita	3	Chaparral
Santa Margarita	4	CSS
Santa Margarita	5	Chaparral
Santa Margarita	6	CSS
Santa Margarita	7	Chaparral
Santa Margarita	8	Chaparral
Santa Margarita	9	CSS
Santa Margarita	10	CSS
Santa Margarita	11	CSS

How would we join these tables?

Plickers...

Executed query – datasheet view

SiteID	County	Elevation	Annual rainf	Soil sample	Vegetation t	Nitrogen
Sky Oaks	San Diego	1418	53	1	CSS	17.0
Sky Oaks	San Diego	1418	53	2	Chaparral	17.6
Sky Oaks	San Diego	1418	53	3	Riparian	13.8
Sky Oaks	San Diego	1418	53	4	Riparian	17.0
Sky Oaks	San Diego	1418	53	5	CSS	11.4
Sky Oaks	San Diego	1418	53	6	CSS	19.8
Sky Oaks	San Diego	1418	53	7	Chaparral	16.4
Sky Oaks	San Diego	1418	53	8	Chaparral	21.0
Sky Oaks	San Diego	1418	53	9	Chaparral	15.6
Sky Oaks	San Diego	1418	53	10	Riparian	19.6
Sky Oaks	San Diego	1418	53	11	Riparian	13.9
Santa Margarit	Riverside	338	36	1	Riparian	19.0
Santa Margarit	Riverside	338	36	2	Chaparral	19.0
Santa Margarit	Riverside	338	36	3	Chaparral	16.1
Santa Margarit	Riverside	338	36	4	CSS	22.2
Santa Margarit	Riverside	338	36	5	Chaparral	22.4
Santa Margarit	Riverside	338	36	6	CSS	16.5
Santa Margarit	Riverside	338	36	7	Chaparral	17.2
Santa Margarit	Riverside	338	36	8	Chaparral	13.4
Santa Margarit	Riverside	338	36	9	CSS	21.0
Santa Margarit	Riverside	338	36	10	CSS	16.3
Santa Margarit	Riverside	338	36	11	CSS	16.8

Filtering data in a query

- Filtering data in a query is very simple
 - Specify a Criteria for the field
- Can use multiple criteria
 - Within a field specifying more than one criteria is an “Or” – records matching either one are returned
 - Between fields:
 - If two criteria are in the same row they are “and” criteria – records have to satisfy all of the criteria to be returned
 - If they are in different rows they are “or” criteria – records that satisfy either criteria are returned, but do not have to satisfy both

Query to show only Sky Oaks data

Field:	SiteID	County
Table:	Sites	Sites
Sort:		
Show:	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Criteria:	"Sky Oaks"	
or:		

CombinedSiteSoilVeg							
SiteID	County	Elevation	Annual rainf	Soil sample	Vegetation 1	Nitrogen	
Sky Oaks	San Diego	1418	53	1	CSS	17.0	
Sky Oaks	San Diego	1418	53	2	Chaparral	17.6	
Sky Oaks	San Diego	1418	53	3	Riparian	13.8	
Sky Oaks	San Diego	1418	53	4	Riparian	17.0	
Sky Oaks	San Diego	1418	53	5	CSS	11.4	
Sky Oaks	San Diego	1418	53	6	CSS	19.8	
Sky Oaks	San Diego	1418	53	7	Chaparral	16.4	
Sky Oaks	San Diego	1418	53	8	Chaparral	21.0	
Sky Oaks	San Diego	1418	53	9	Chaparral	15.6	
Sky Oaks	San Diego	1418	53	10	Riparian	19.6	
Sky Oaks	San Diego	1418	53	11	Riparian	13.9	

Query to show Sky Oaks, with nitrogen levels over 17

Field:	Vegetation type	Nitrogen
Table:	Vegetation	SoilSamples
Sort:		
Show:	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Criteria:		>17
or:		

CombinedSiteSoilVeg							
SiteID	County	Elevation	Annual rainf	Soil sample	Vegetation 1	Nitrogen	
Sky Oaks	San Diego	1418	53	1	CSS	17.0	
Sky Oaks	San Diego	1418	53	2	Chaparral	17.6	
Sky Oaks	San Diego	1418	53	4	Riparian	17.0	
Sky Oaks	San Diego	1418	53	6	CSS	19.8	
Sky Oaks	San Diego	1418	53	8	Chaparral	21.0	
Sky Oaks	San Diego	1418	53	10	Riparian	19.6	

Data entry forms with a 1:many relationship

- A single record is entered at a time
- Joined tables can both be displayed
- This example shows a form for sites (1) with a **subform** for samples (many)
 - All samples that match the site ID are displayed in the subform
 - More samples can be added to a site as they are collected, and the Site ID is added to the samples table automatically
 - A new site can be added, then samples added for that site

The screenshot shows a data entry form titled 'Sites' with the subtitle 'Combined information about sites and samples'. The form contains several input fields:

- ID: 2
- SiteID: Sky Oaks (dropdown menu)
- County: San Diego (dropdown menu)
- Elevation: 1418
- Annual rainfall: 53

Below the main form is a 'SoilSamples subform' table with the following data:

Soil sample ID	Site ID	Nitrogen
1	Sky Oaks	17.0
2	Sky Oaks	17.6
3	Sky Oaks	13.8
4	Sky Oaks	17.0
5	Sky Oaks	11.4

The table includes a 'Record' indicator showing '1 of 13' and a 'Search' button.

Common types of RDBMS you may encounter

- Desktop software
 - MS Access, FileMaker
 - Meant to be used by one person at a time
 - These RDBMS's keep databases in single files that can be copied and moved around easily
- Structured Query Language distributed systems
 - MS SQL Server, MySQL, Postgresql, Oracle
 - Run from a central server over a network
 - Data stored centrally – changes are available to all users at once
 - Usually used for large, shared databases, web applications

Why don't we use RDBMS's all the time?

- Complexity, unfamiliarity, inflexibility
- Excel, the Swiss Army Knife, does well enough most of the time, allows us to do some things better or more simply
 - Attractive layout of data tables
 - More graphing options
 - Complex calculations stored within the spreadsheet
- Be aware of the uses of relational databases, in case you have an application for which Excel will not suffice
- Data can be moved back and forth between spreadsheets and databases – in combination they can be very powerful