Why is data management important to me?

- Save Time!
- Save Money!
- ~~No data.....no money~~
- ~~No data.....no thesis~~
- ~~No data.....no publications~~
- ~~Set yourself apart~~

The Director of DataONE, Bill Michener, uses the "80-20 rule" based on survey results and data management interviews with scientists:

"Eighty percent of a scientist's effort is spent discovering, acquiring, documenting, transforming, and integrating data, whereas only 20 percent of the effort is devoted to more intellectually stimulating pursuits such as analysis, visualization, and making new discoveries." What if you could spend your discovering, acquiring, documenting, transforming, and integrating time more efficiently?

Larry P. English is the president and principal of Information Impact International, Inc. He is an internationally recognized speaker, teacher, consultant, and author in information and knowledge management, and information quality improvement. According to Larry English, poor data quality can cost companies 15 percent to 25 percent of their operating budget. How can this be?

Examples:

A wildlife biologist for a small field office was the in-house GIS expert and provided support for all the staff's GIS needs. However, the data were stored on her own workstation. When the biologist relocated to another office, no one understood how the data were stored or managed.
Solution: A state office GIS specialist retrieved the workstation and sifted through files trying to salvage relevant data.
Cost: One work-month ($4,000) plus the value of data that were not recovered.

An office contracted out data collection but failed to provide the contractor with appropriate data standards. When the inventory was completed, the data were found to be worthless because they were collected to the wrong standard.
Solution: Re-inventory.
Cost: $65,000

In preparation for a Resource Management Plan, an office discovered 14 duplicate GPS inventories of roads. However, because none of the inventories had enough metadata, it was impossible to know which inventory was best or if any of the inventories actually met their requirements.
Solution: Re-inventory roads.
Cost: Estimated 9 work-months/inventory @ $4,000/work-month (14 inventories = $504,000).

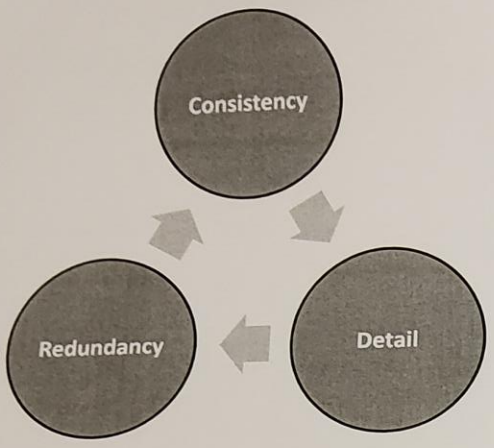
Why is data management important to me?

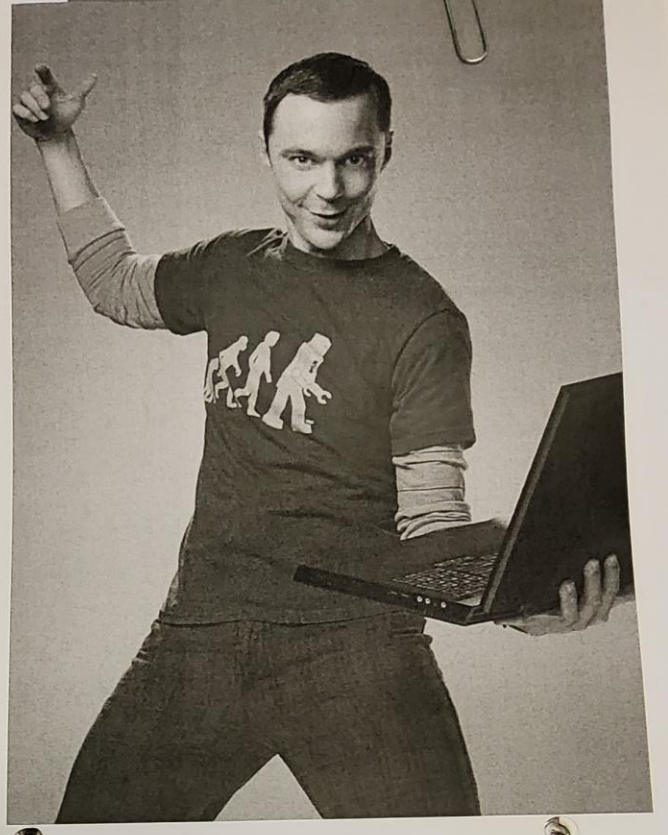
- Save Time!
- ~~Save Money!~~
- ~~No data.....no money~~
- ~~No data.....no thesis~~
- ~~No data.....no publications~~
- ~~Set yourself apart~~

→ you will use your datasets in unplanned ways

The Director of DataONE, Bill Michener, uses the "80-20 rule" based on survey results and data management interviews with scientists:

"Eighty percent of a scientist's effort is spent discovering, acquiring, documenting, transforming, and integrating data, whereas only 20 percent of the effort is devoted to more intellectually stimulating pursuits such as analysis, visualization, and making new discoveries."
What if you could spend your discovering, acquiring, documenting, transforming, and integrating time more efficiently?



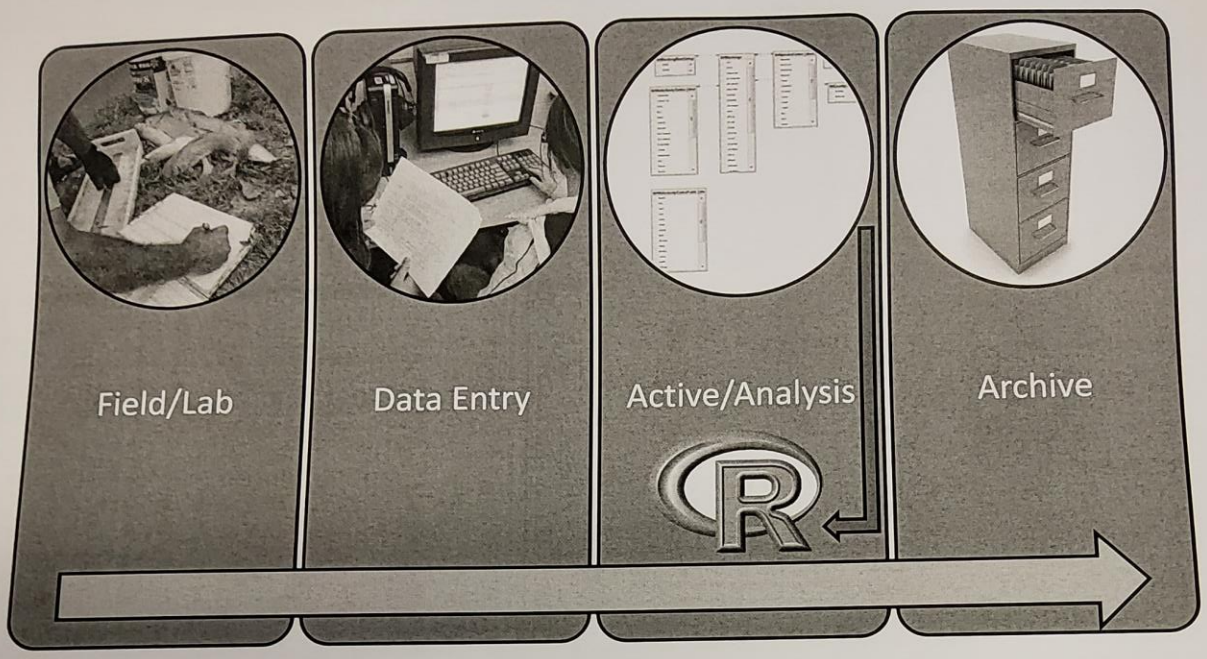


Actually, more reminiscent of engineers than geeks. There's always some truth to all stereotypes, so what about the geek represents good quality of data managers?

Retentive nature, same way every time, accept menial tasks as the cost of doing business...

Gives us a nice framework for thinking about data management,

Data Management



Relavent to a MS/PhD style project
May start Planning in the middle

QA & QC explanations

What is QC QA?
product / process

QA = validation
verification
processes to ensure
quality data

QC = is QA working?
are products "quality"
suggests changes to QA

Detail

Units - put on datasheet
Percision - Equipment
Visual clues for Required Data

use □□□.□□

Abbreviations CRP ⇒ crappie carp

test clerks!

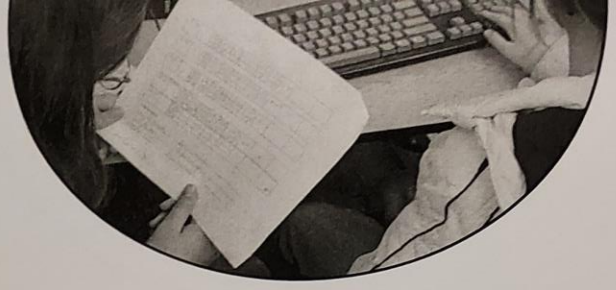
Consistancy

Layout
Self-Explanitory/Instructions
Load/Launch/Unload checklists
How long to complete partial datasheets
Don't do front/back → no flipping
Training
equipment calibrations

Checklists
Protocols
Two Person Verifications
Equipment Packs
Put Up & Replenish at end of trip!

Redundancy

Waterproof
Datasheet Procedures
clipboard spots
Cell phone copies of datasheets
contact info on datasheets



tech solutions vs. logistic solutions

Detail

Validation: select lists
ranges
valid dates
required

Missing Data: Orphans
missing relations
cascades

validation in table or form

null vs 0 vs missing

Consistancy

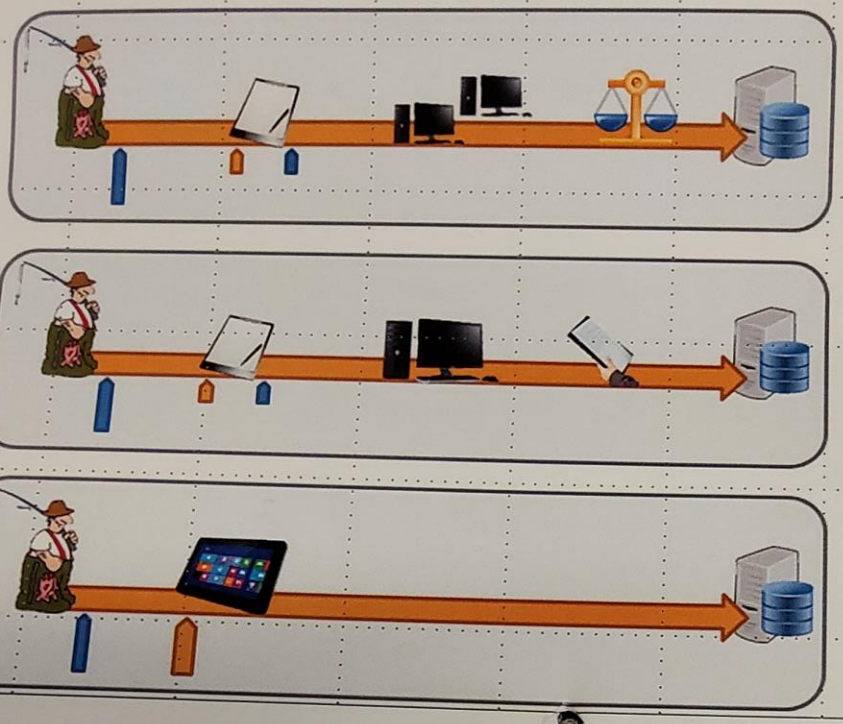
How long until entered
Double Entry
Verification
Match Layout of Datasheets
Procedures – mark entered datasheets
grab user & dates on data entry
fidality checks

Redundancy

Backup Procedures
Datasheet storage
server Data Stores
Protected Against Vengeful employees

data cleaning takes place in data entry ⇒ not data analysis

5 4 3



Double Entry

Risk

Time

Entry With Verification

Risk

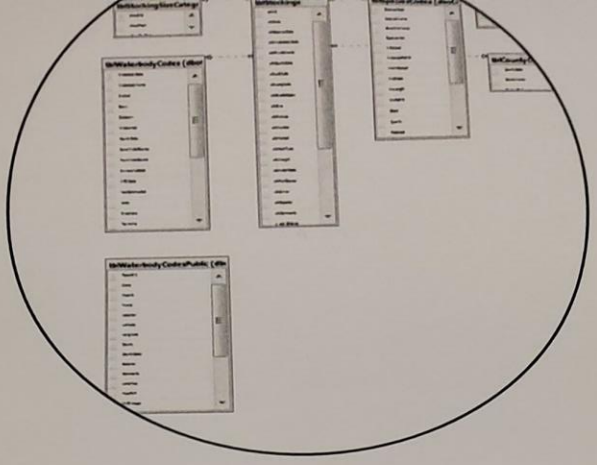
Time

Mobile Entry

Risk

Time

— What's the difference between transcription & translation? errors



what is meta-data

Details
Meta-Data
Explict Conversions Between Types

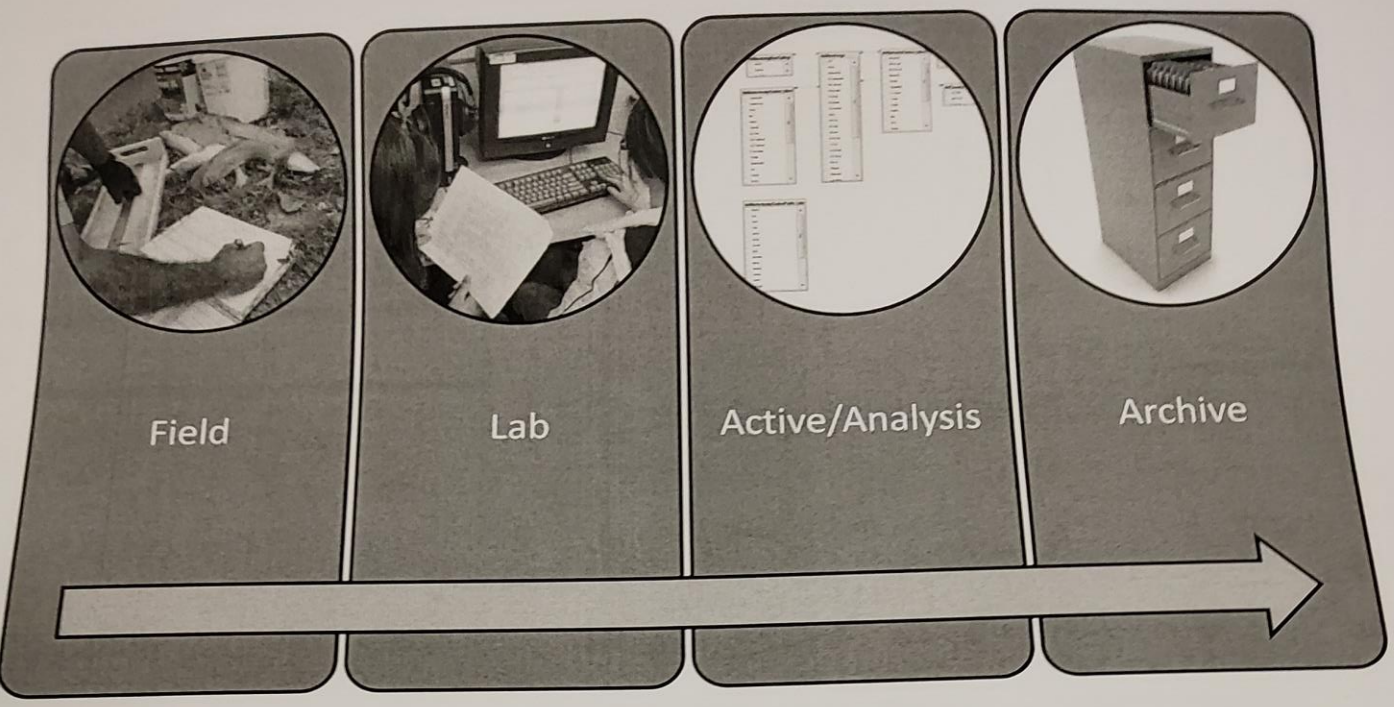
Consistancy
Naming Conventions
- files
- tables
- fields
Procedures For queries
- saving
- naming
- storing & access
Changes to errors are handled how?

Redundancy
Backup Procedures
Backup vs versioning
Disaster Recovery
Practice Recovery!

Details
Meta Data
Maps
File Types

Consistancy
Between Projects

Redundancy
& storage



Consider each step of the process & apply D-C-R & write a plan.
Think if you are absent from the field & create instructions
For most of your projets, more than 1 person will be involved. Data files are also located on shared drives and servers. This raises some other important ???s

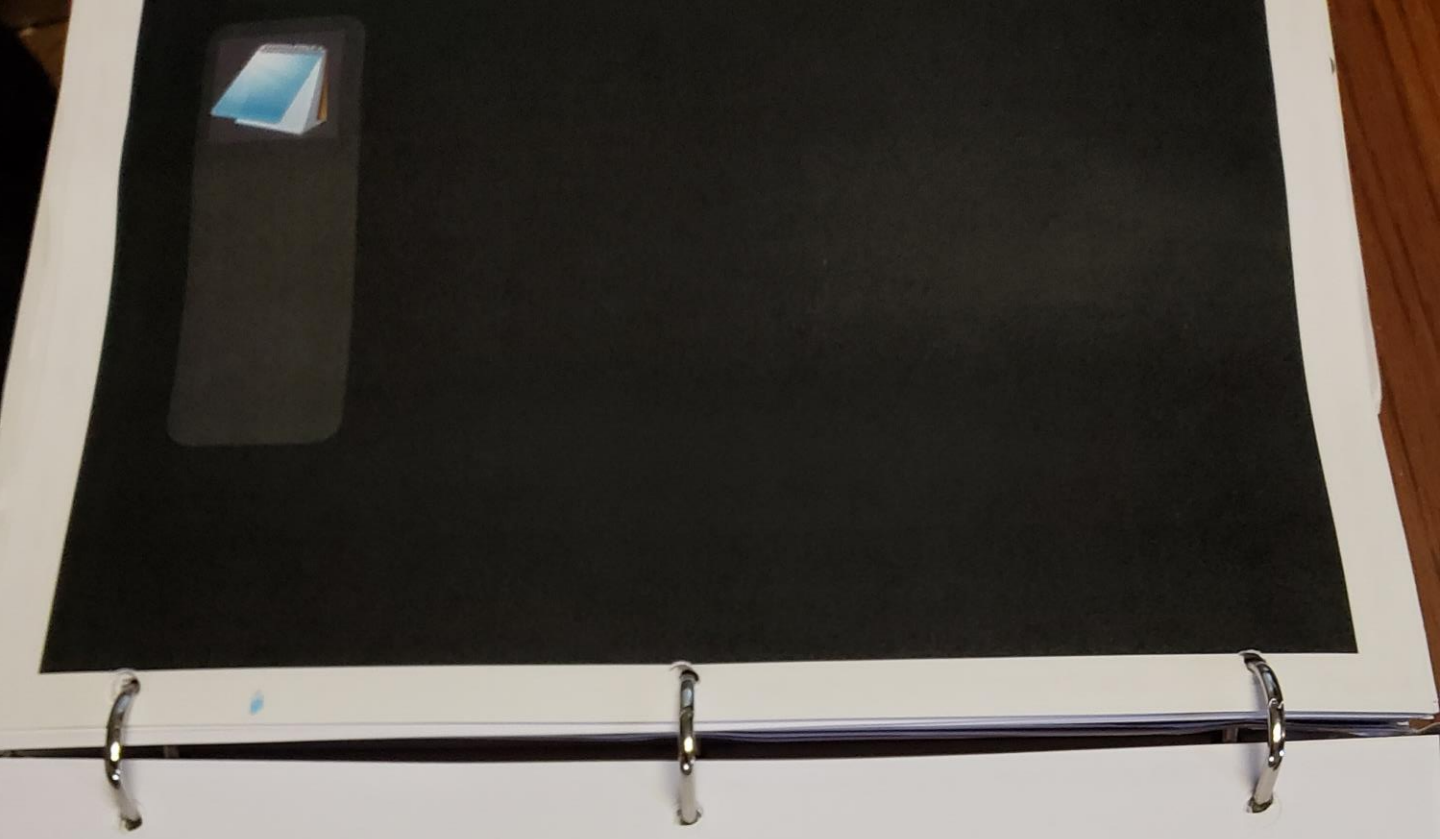
→ Data Sharing
→ Technical → how will I
→ Ethics/Leagalities → can I & should I
→ logistical how will the team or agency funnel & fulfill requests
→ who has the final say
→ can quickly turn legal

Data Sharing Agreements
Freedom of Information Acts
Intellectual property rights
- Pharmaceuticals
- products & technologies
- data is big money → google?

* validation
* constraints
* integrity

- each piece of data in one-to-only one place
- scalable
- multi-user

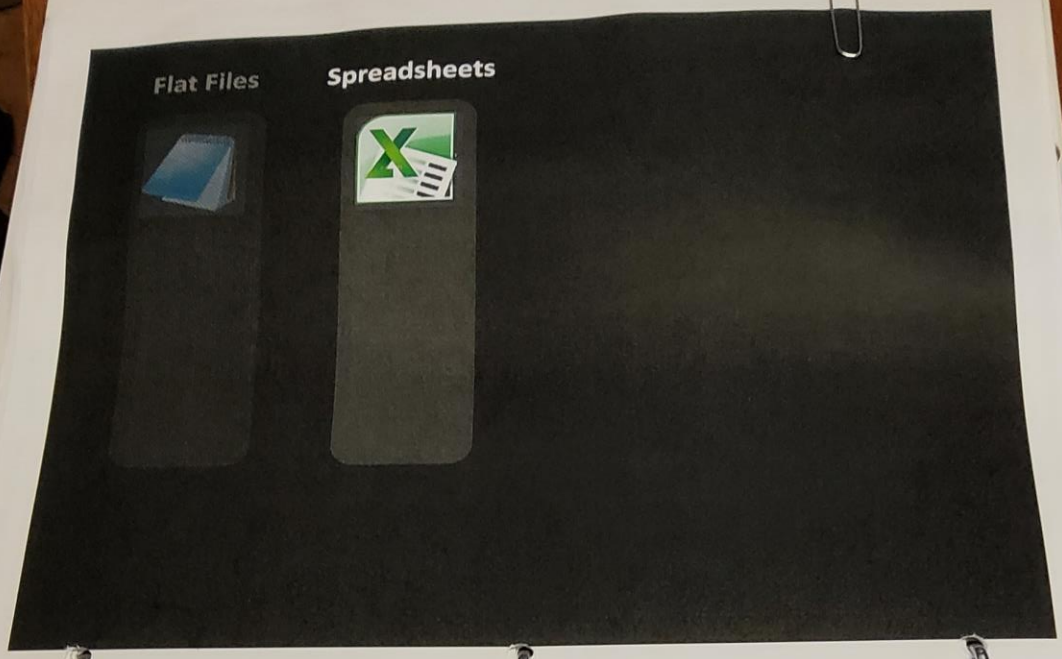
5 4



Flat Files

Text
CSV
XML
JSON

Long-Term stable
Potential For Archive & Data Transfer
Absolutely No Validation or Constraints
Difficult to edit
string Delimitors



Likely The most-used in your shoes.

Spreadsheets are Ledgers.... a giant calculator

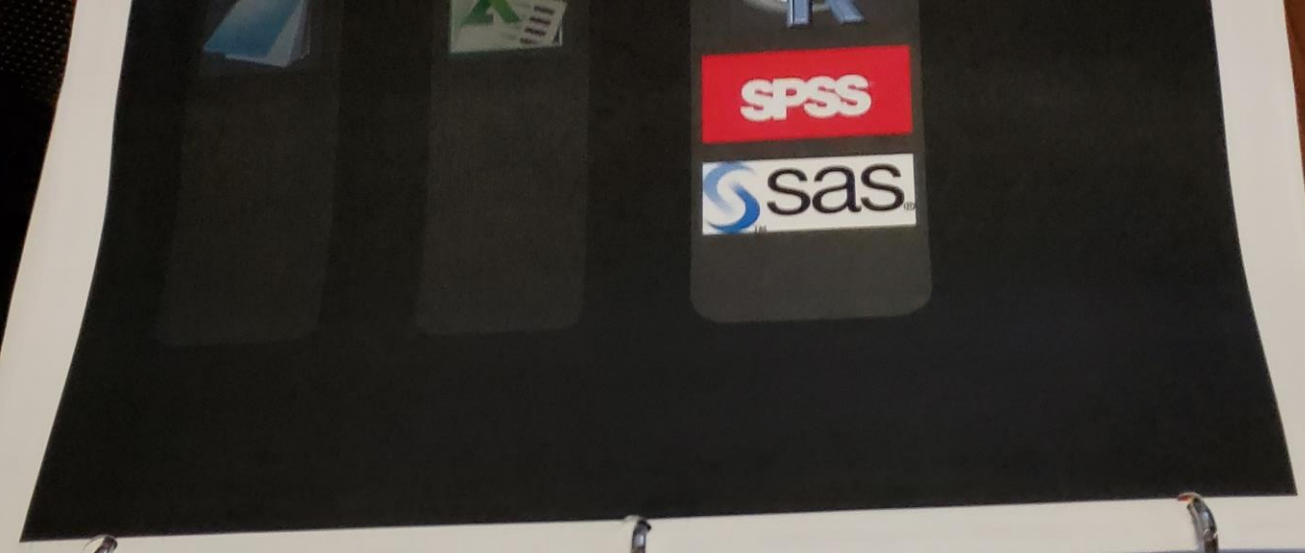
No Validation or constraints... promotes cut+paste

No "model".... just rows+columns - no structure

Formats change through time

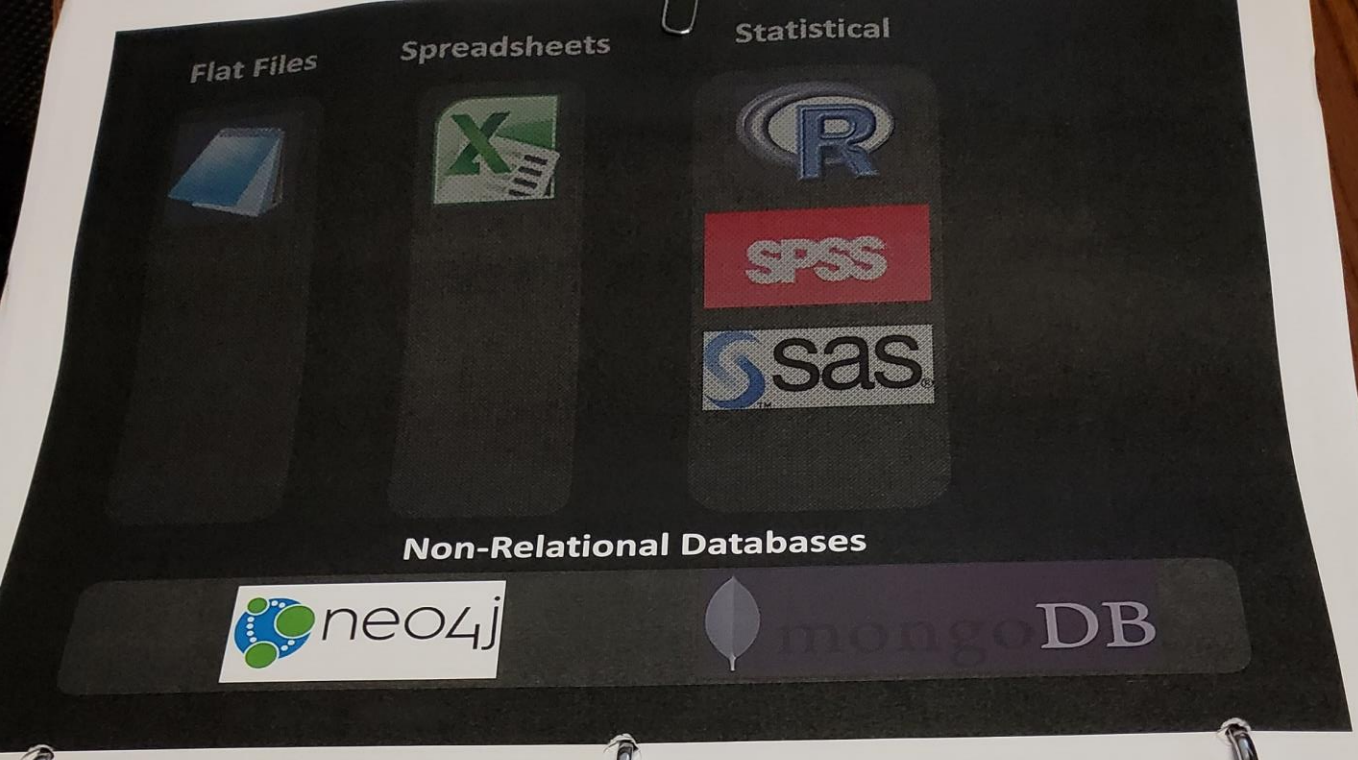
No "Query" style actions.... encourages manual updates
- leads to cut+Paste.... which is always BAD

Handleing data not a stength....



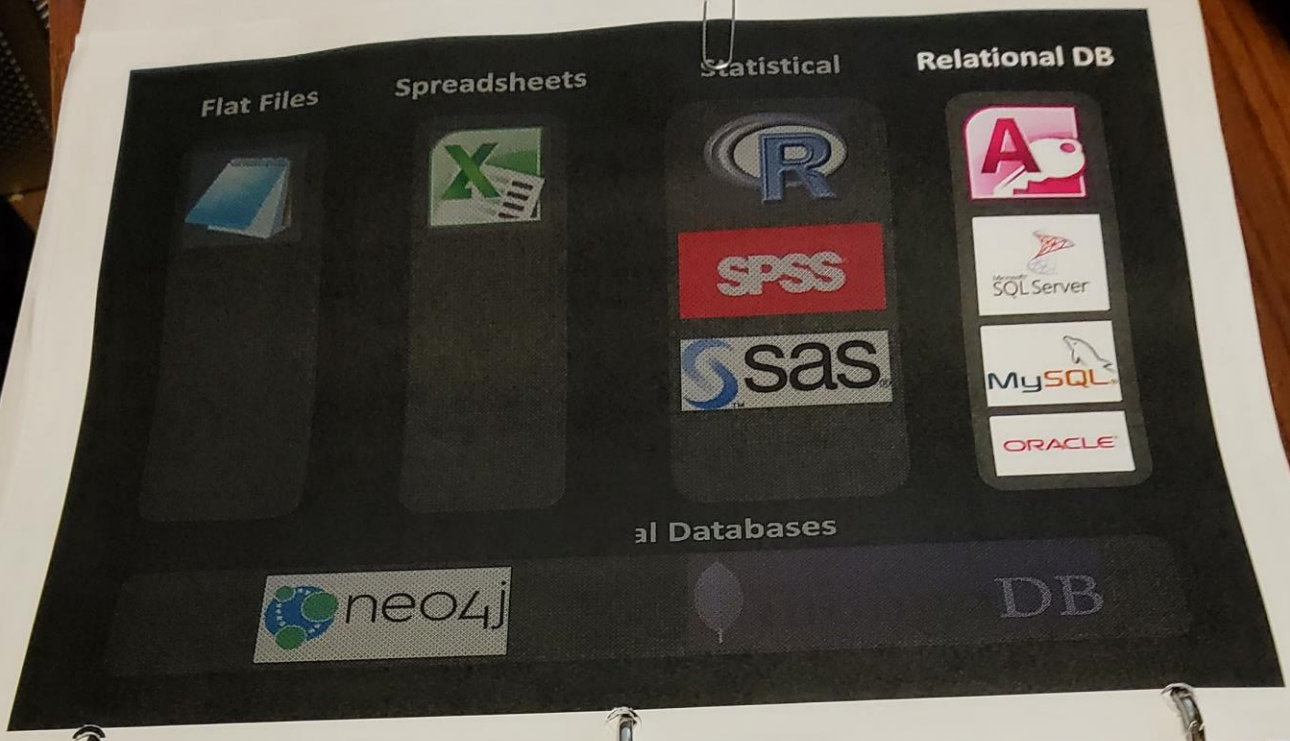
Statistical

→ most storage is same as flat files
→ not much for data entry – more of a companion or one of a "team" of tools
→ limited validation → more QC + less QA
→ no ~~usability~~ scalability
→ limited multi-user
→ mostly in-memory processes
→ good for calculations not so much for joins & data wrangling



- most rapidly expanding group
- used in some of the largest web DB's — social media
- not as rigidly structured
- eliminates/overcomes some shortcomings of relational DB's
- some are free

- scalable
- multi-user



There Are other Databases...

what is relational data?

Validations & Constraints are possible

Built For Multi-user access.

Built for scalability

Some Are Free (SQL server Express, MYSQL)

Bigger Learning curve.

Generally store data & apply queries. Takes more coding.

Good design means
- simpler queries
- better performance
- less data entry
- easier to add new data fields in process

A record = An entity ⇒ "sets" of entities

date	lake	Site										
06/01/13	Round	Loon Point	Trap Net	bluegill	210	254	4	178	225	345	481	490
06/01/13	Round	Loon Point	Trap Net	bluegill	243	222	4	145	242			292
06/01/13	Round	Pintail Bay	Trap Net	bluegill	135	86	2	195	284			
06/01/13	Round	West end of dam	Trap Net	redear	142							
06/01/13	Round	Loon Point	Trap Net	black bullhead	138							
06/25/13	Round	Pintail Bay	Trap Net	black bullhead	179							
06/25/13	Round	Big Creek mouth	Trap Net	bluegill	198	168	3	188	254	365		365
06/25/13	Square	Mallard Cove	Gill Net	walleye	854	355	4	113	241	384	476	484
08/30/13	Square	Whitetail Point	Gill Net	walleye	686	321	3	105	289	349		349
08/30/13	Square	Bobber Bay	Gill Net	white bass	312	211						230

Example of A Fish sampling DB as it would appear in Excel.

Lots of Repeated Data. - both rows & columns

Lots of Stored Text.

- Why is text bad in data?

No "Model" or relations

May use this style for analysis, why not for storage?

Normalize Your Data base → each row represents 1 piece of data
→ no repeating info
→ save space
→ don't store what can be calculated

Example of changeing units on annulus

06/01/13	Round			2	142							
06/01/13	Round	West end of dam	1	3	138							
06/25/13	Round	Loon Point	1	3	179							
06/25/13	Round	Pintail Bay	1	1	198	168	3	188	254	365		365
06/25/13	Round	Big Creek mouth	1	4	854	355	4	113	241	384	476	484
08/30/13	Square	Mallard Cove	2	4	686	321	3	105	289	349		349
08/30/13	Square	Whitetail Point	2	5	312	211						230
08/30/13	Square	Bobber Bay	2									

tblGears	
gear_UID	gear_Name
1	Trap Net
2	Gill Net

tblSpecies	
spp_UID	spp_Name
1	bluegill
2	redear
3	black bullhead
4	walleye
5	white bass

First, Let's code our data to eliminate repeating text
- sometimes called lookup tables
- Allows a name change to be made 1ce & affect all records.

ALL TABLE ROWS GET A UNIQUE ID

Explain naming conventions for fields.
Auto number fields for VID or Not?

date	site	gear	species	length	weight	age	a1	a2	a3	a4	edge
06/01/13	1	1	1	210	254	4	178	223	381	413	413
06/01/13	1	1	1	243	222	4	145	242	335	481	490
06/01/13	2	1	1	135	86	2	195	284			292
06/01/13	3	1	2	142							
06/01/13	1	1	3	138							
06/25/13	2	1	3	179							
06/25/13	4	1	1	198	168	3	188	254	365		365
06/25/13	5	2	4	854	355	4	113	241	384	476	484
08/30/13	6	2	4	686	321	3	105	289	349		349
08/30/13	7	2	5	312	211						230
08/30/13											

tblLakes

lake_UID	lake_Name
1	Round
2	Square

tblSites

site_UID	Site_lakeUID	site_Name
1	1	Loon Point
2	1	Pintail Bay
3	1	West end of dam
4	1	Big Creek mouth
5	2	Mallard Cove
6	2	Whitetail Point
7	2	Bobber Bay

tblGears

gear_UID	gear_Name
1	Trap Net
2	Gill Net

tblSpecies

spp_UID	spp_Name
1	bluegill
2	redear
3	black bullhead
4	walleye
5	white bass

Sites & Lakes are a nested situation.

date	site	gear	species	length	weight	age	a1	a2	a3	a4	edge
06/01/13	1	1	1	210	254	4	178	223	381	413	413
06/01/13	1	1	1	243	222	4	145	242	335	481	490
06/01/13	2	1	1	135	86	2	195	284			292
06/01/13	3	1	2	142							
06/01/13	1	1	3	138							
06/25/13	2	1	3	179	168	3	188	254	365		365
06/25/13	4	2	1	198	355	4	113	241	384	476	484
06/25/13	5	2	4	854	321	3	105	289	349		349
08/30/13	6	2	4	686	211						230
08/30/13	7	2	5	312							

Now Lets deal with information that repeats in every row. Time to think in sets.

tblSets			
set_UID	set_Date	set_Site	set_Gear
1	6/1/2013	1	1
2	6/1/2013	2	1
3	6/1/2013	3	1
4	6/25/2013	1	1
5	6/25/2013	2	1
6	6/25/2013	4	1
7	8/30/2013	5	2
8	8/30/2013	6	2
9	8/30/2013	7	2

tblFish										
fish_UID	fish_setUID	fish_sppUID	fish_length	fish_weight	fish_age	a1	a2	a3	a4	edge
1	1	1	210	254	4	178	223	381	413	413
2	1	1	243	222	4	145	242	335	481	490
3	2	1	135	86	2	195	284			292
4	3	2	142							
5	4	3	138							
6	5	3	179							
7	6	1	198	168	3	188	254	365		365
8	7	4	854	355	4	113	241	384	476	484
9	8	4	686	321	3	105	289	349		349
10	9	5	312	211						230

Funny, but the repeating info is directly correlated with net sets. This is what I refer to as a data "model". The data layout represents the meaning of the data in real life.

It allows us to create "sets" of data

Each row represents 1 piece of data... 1 set... and only in 1 row

Can query/use/update info on sets without having the overhead of fish data

tblSets

set_UID	set_Date	set_Site	set_Gear
1	6/1/2013	1	1
2	6/1/2013	2	1
3	6/1/2013	3	1
4	6/25/2013	1	1
5	6/25/2013	2	1
6	6/25/2013	4	1
7	8/30/2013	5	2
8	8/30/2013	6	2
9	8/30/2013	7	2

fish_UID	fish_setUID	fish_sppUID	fish_length	fish_weight	fish_age	fish_edge
1	1	1	210	254	4	413
2	1	1	243	222	4	490
3	2	1	135	86	2	292
4	3	2	142			
5	4	3	138			
6	5	3	179			
7	6	1	198	168	3	365
8	7	4	854	355	4	484
9	8	4	686	321	3	349
10	9	5	312	211		230

tblAge

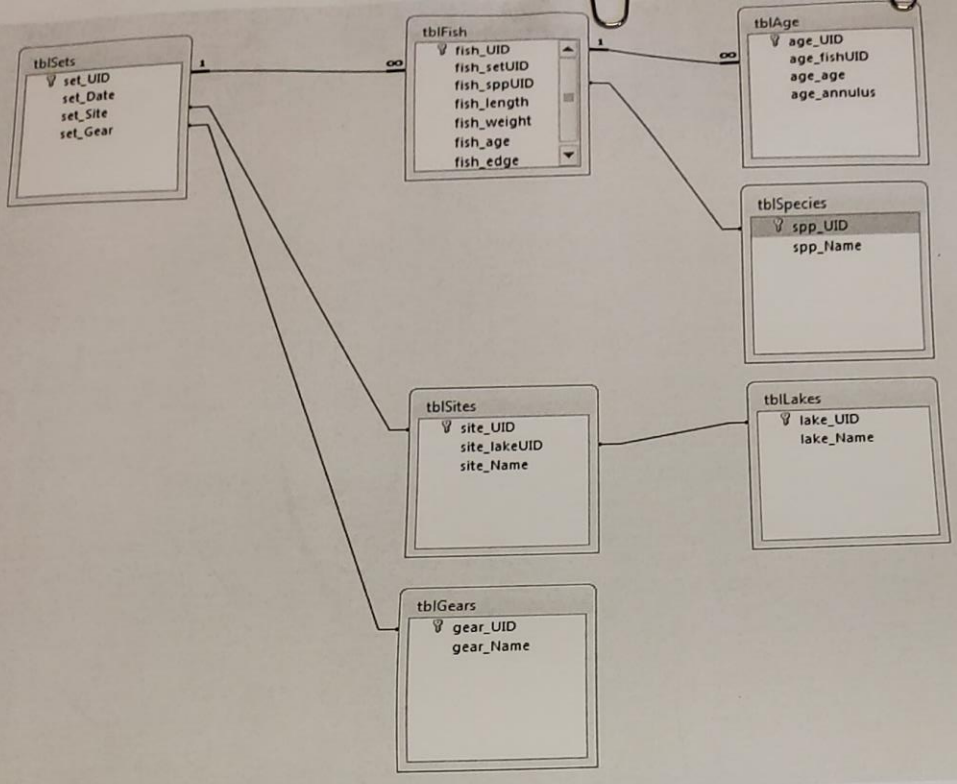
age_UID	age_fishUID	age_age	age_annulus
1	1	1	178
2	1	2	223
3	1	3	381
4	1	4	413
5	2	1	145
6	2	2	242
7	2	3	335
8	2	4	481
9	3	1	195
10	3	2	284
11	7	1	188
12	7	2	254
13	7	3	365
14	8	1	113
15	8	2	241
16	8	3	384
17	8	4	476
18	9	1	105
19	9	2	289
20	9	3	349

Now let's remove repeating columns by pivoting them.

This seems to be the hardest leap from spreadsheet to db thinking.

Has huge impacts on complexity of queries & operations.

- example → how old is fish?
  query each column until you find a null v.s. count(age_UID, age_UID) max(" , ")
- so do we need age?

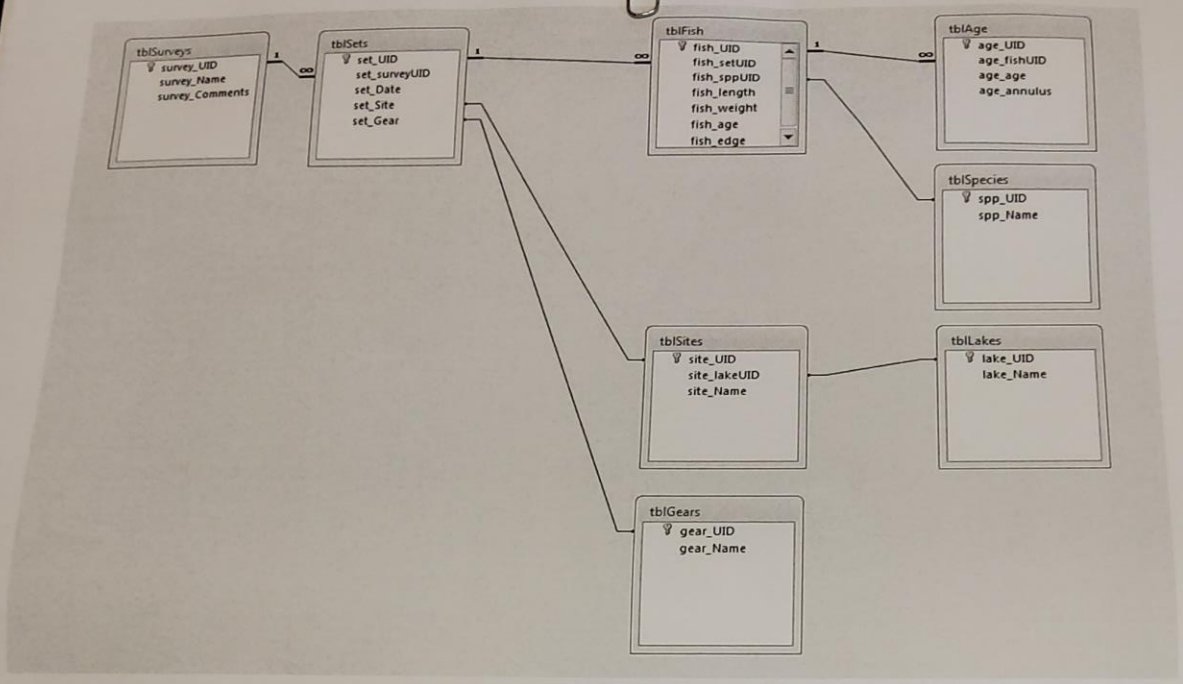


Here is a relational model of our normalized database

- Talk about relations, orphans, cascades, 1-∞ & 1-1 & ∞-∞

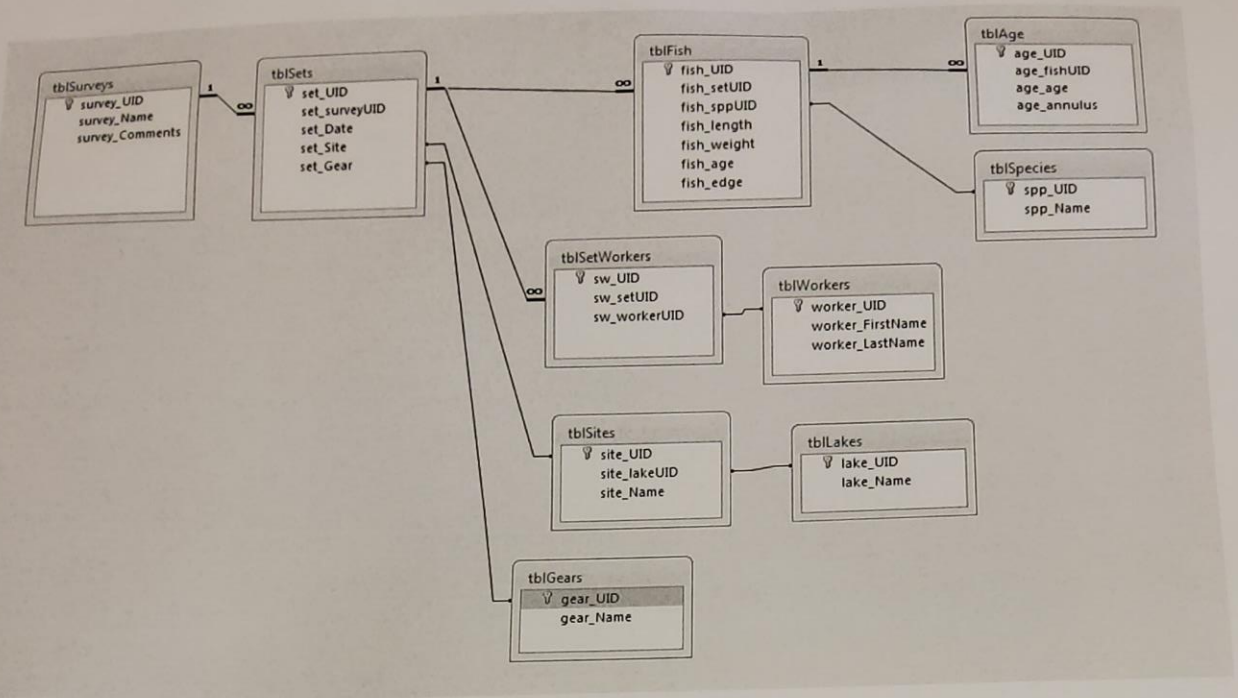
Are we missing anything?

- how about identifing individual surveys
  * why would we?



Now we can create sets of "sets" based on surveys.
— not based on waterbody — could be 1 or many for each survey.

"Model" reflects real life.
It's now easy to add info to the existing model.
For QA/QC, maybe we need names of people conducting surveys.



Talk about skinny/join tables

Will also allow us to create sets of "sets" done by specific employees?

- when would this be important

- text vs memo
- text vs number
- empty string vs null

➢ Date/time fields

➢ Boolean/bit fields

➢ Integer/Double/Decimal

➢ Specialty Fields

➢ Nullable fields

➢ Text Fields are bad... inefficient, large, un-analyzable
- Don't Link on text fields
- control size of text field
- code to numbers
- difference between "" empty string & null

➢ How are date & time stored?

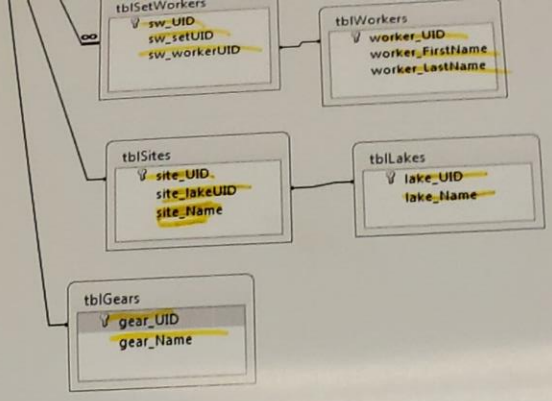
date xxx . xxx time

➢ boolean
- does it allow null?

➢ numbers
- what level of percision do you need

➢ specialty
- Lat/Lon Geo
- binary
- image

➢ nullable
- ~~difference between "" & null~~



relational diagram
+ required
+ data type + Key

There is no "Undo" in database

* Make New Database
* Make tbl surveys
* Make tble Sets

Show options
— design changes in datasheet

— show captions
— show text field & sizing
— show Auto number
— show nullable
— show primary Key

show combo box for survey
show date — show default formatting — input mask — date picker
show numbers

* have them make tbl Fish

* Get DB Demo1 DB → has all tables
  - set some relationships
    - show direction & it's meanings
    - cascades
    - help in queries... Sometimes you'll change directions

1. Add Fields
2. Add Buttons
3. show size/Align/Anchor
4. show Locked + Enabled
5. tab order

Forms

→ create surveys form w/ sets subform
- show link to Fish Form using macros
  - show popup + modal

fish_setUID
=
Forms!
frm Data Entry
sfrm Sets!
set_UID

views - datasheet, single form, continuous forms
Formats - record selectors, deviding lines, etc - locked + enabled
screen shots of PF database
Tab Order - dashboard w/ short cuts
- QA section
- try adding Workers form

Reports

→ create report showing Surveys + sets
- explain Grouping + sorting
- running totals

other

- Compacting Databases
- passwording DB's
- Securing DB's
  - Be careful
- multi-user access
- use on a network → can bog down

Queries

→ CRUD vs select — select x from tbl where

→ designer vs SQL
- criteria and +ons

select set_date from tblSets where set_gear=1 — simple input of parameter

- Get Fish from a given date set
- Get max & min dates by survey
- back-calc ageg - using queries as source — $c + L_c - c)(s_i/s_c)$
  - example of syntax is spreadsheet form
- Export Data Example — population, length, weight for Wr work on b/c

if time challenge them to query WAE in GIII or BLG in traps

- Use distinct to get fish Spp in DB
- "No Join" to create matrix of observations
  - ↘ distinct Spp
  - * distinct site & year
  - * create possibles using "No Join"
  - must use distinct → * create actuals w/ "Presence: 1"
  - * Join actuals & possibles & crosstab

→ Crosstabs & Unions