Attribution Key

for more information see: http://open.umich.edu/wiki/AttributionPolicy

Use + Share + Adapt

{ Content the copyright holder, author, or law permits you to use, share and adapt. }

PD-GOV **Public Domain – Government**: Works that are produced by the U.S. Government. (17 USC § 105)

PD-EXP **Public Domain – Expired**: Works that are no longer protected due to an expired copyright term.

PD-SELF **Public Domain – Self Dedicated**: Works that a copyright holder has dedicated to the public domain.

ZERO **Creative Commons – Zero Waiver**

BY **Creative Commons – Attribution License**

BY-SA **Creative Commons – Attribution Share Alike License**

BY-NC **Creative Commons – Attribution Noncommercial License**

BY-NC-SA **Creative Commons – Attribution Noncommercial Share Alike License**

GNU-FDL **GNU – Free Documentation License**

Make Your Own Assessment

{ Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright. }

PD-INEL **Public Domain – Ineligible**: Works that are ineligible for copyright protection in the U.S. (17 USC § 102(b)) *laws in your jurisdiction may differ

{ Content Open.Michigan has used under a Fair Use determination. }

FAIR USE **Fair Use**: Use of works that is determined to be Fair consistent with the U.S. Copyright Act. (17 USC § 107) *laws in your jurisdiction may differ

Our determination **DOES NOT** mean that all uses of this 3rd-party content are Fair Uses and we **DO NOT** guarantee that your use of the content is Fair.

To use this content you should **do your own independent analysis** to determine whether or not your use will be Fair.

Data Analysis and Design Data Jam

Justin Joque
Fall 2012

Agenda

- Introduction
- Analysis Tools
- Data Design
- Jam/Discussion about Huron River Watershed Council Data

Needs Evaluation

3 Critical Components:

1. Collection Processes
   - Is the current process working?
   - Do additional components need to be added?
2. Data Organization
   - Does the organization facilitate moving data from collection to analysis?
   - This includes location, design and control
   - This is an area where you can have a big impact
   - The best solutions are often times the least complex
3. Data Analysis
   - What questions can you help answer?
   - What methods and tools can you recommend?

Analysis Tools

- Data Storage
  - SQL (MySQL, Access)
  - Google Docs
  - Excel
- Data Manipulation
  - R/SPSS/Stata
  - Excel
  - Text Editor
  - Perl/Python
  - Google Refine
  - Fusion Tables

- Data Analysis and Visualization
  - R/SPSS/Stata
  - Excel
  - ArcGIS
  - Perl/Python

Remember: Focus on processes whenever possible

Data Design

- Controlling who and at what cost someone can update data (or the schema) can improve quality
- Document your design
- Try to make it as simple and useable as possible
- Design to move data from collection to analysis

Data Design - The Difficulty of Excel

Name	Pet
John	Cat
Alice	Dog
Bob	Cat

Name	Pet	Pet Name
John	Cat	Whiskers
Alice	Dog	Spot
Bob	Cat,Dog	Mittens,Sparky

Data Design - The Difficulty of Excel

Name	Pet
John	Cat
Alice	Dog
Bob	Cat

Name	Pet1	Pet Name1	Pet2	Pet Name2
John	Cat	Whiskers		
Alice	Dog	Spot		
Bob	Cat	Mittens	Dog	Sparky

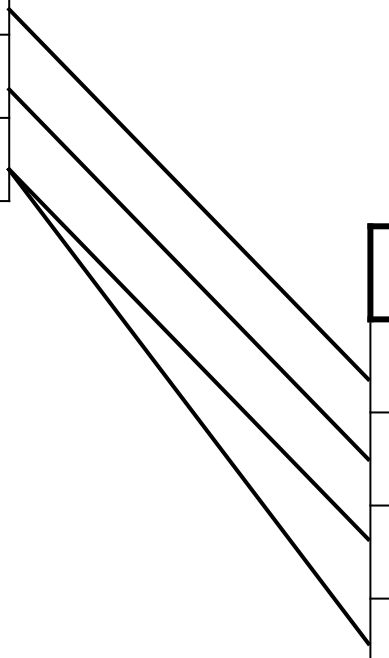
Data Design - The Difficulty of Excel

Name	Pet
John	Cat
Alice	Dog
Bob	Cat

Name	Pet	Pet Name
John	Cat	Whiskers
Alice	Dog	Spot
Bob	Dog	Sparky
Bob	Cat	Mittens

Data Design - Relational Databases

Name	ID
John	1
Alice	2
Bob	3



Person_ID	Pet	Pet Name
1	Cat	Whiskers
2	Dog	Spot
3	Dog	Sparky
3	Cat	Mittens

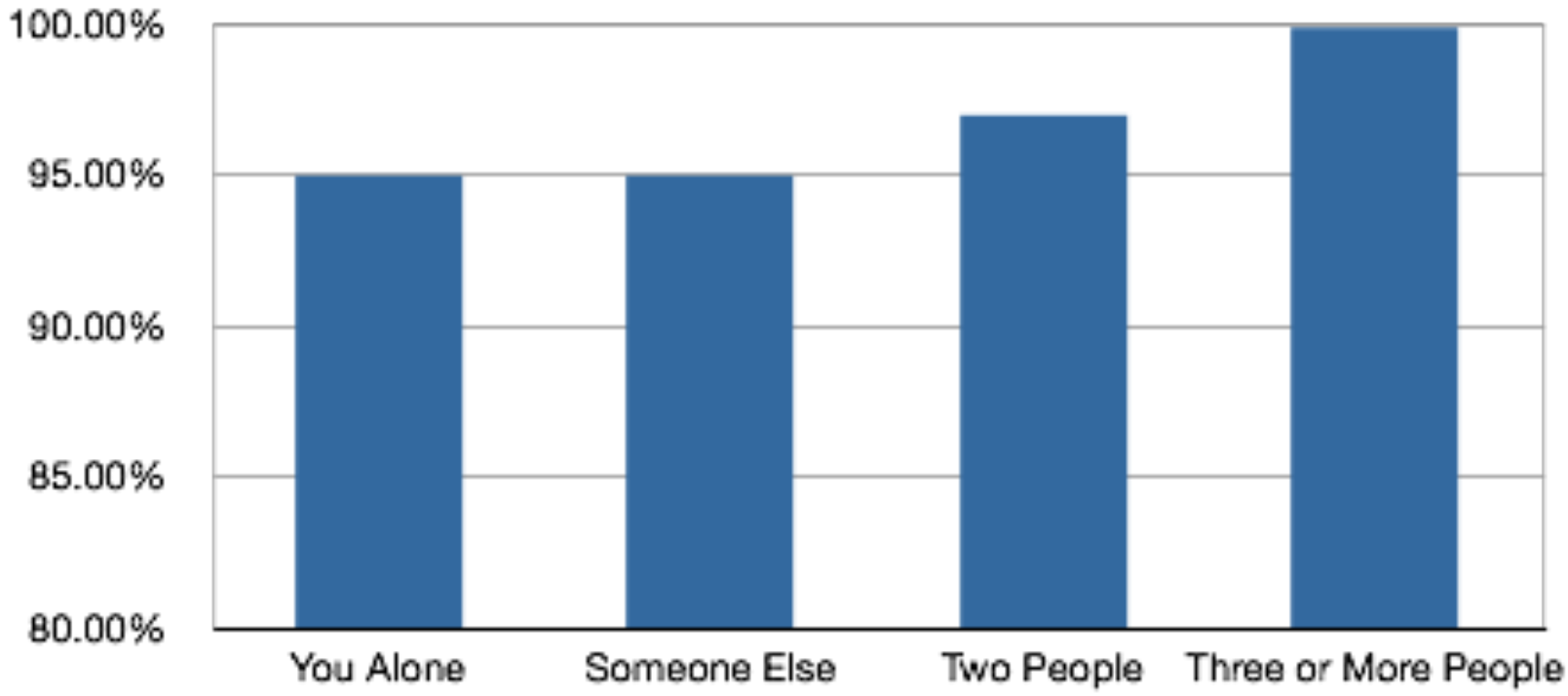
Data Design - Flatness

- Relational Databases are Flat
  - This makes processing and analysing data significantly easier
  - Even if you are not using a relational database try to 'approximate flatness'
  - Stacking vertically can help
  - If you must stack within cells try to use as unlikely of a character as possible to separate data (i.e. use | instead of a commas since commas can show up in people's names).

Data Design - Uniqueness

- Relational Databases Rely on Unique Identifiers
  - These must be unique or they will not work
  - Breaking uniqueness often means manual cleanup or bad results
  - This can be emulated in Excel, but beware of formulas and sorting

Likelihood of Error or Willful Abandonment of the Planned Data Entry Method



Data Design - Authorities

- Authorities lists or controlled entry options can vastly improve data entry
  - This can be done through validation or forms for entry (google docs is pretty good at this.
  - Make sure to allow for updating an authorities list. In terms of time/effort, updating an authorities list should be costly but not prohibitively so
  - Using numbered scales or even just prior agreement can be beneficial (e.g. we will enter man/woman and not male/female).
  - Document these decisions in a place that is accessible and ideally co-located with the data/data entry point

Case Study - HRWC

Site Survey Data Description: HRWC divided the county up into numbered bioreserves which were then combined (geographically) with parcels. They then use this information to try to contact property owners, get permission to survey the property, and then record the results.

Problem: They have two different database, one for the addresses and contact information and one for the survey

Case Study - HRWC

Step 1: Try to figure out what is going on with the sample data.

- What do the IDs represent?
- Are the tables flat?
- Has uniqueness been maintained where necessary?
- Is the data clean?
- What questions would you ask the organization to further elucidate the current structure?

Case Study - HRWC

Step 2: Try to think through how the tables could be rearranged and data entry could be better controlled.

- How would you transform the data to the new structure?
- Does the new structure lend itself to analysis?
- Does it maintain flatness and uniquness?

Case Study - HRWC

Step 3: Try to think about what analyses could easily be mined from your new data structure.

- Does the new structure make it easier?
- Could you manage to achieve the same results with the old structure?
- What additional analyses might be helpful?