What is “the workflow of data analysis”?
1. A coordinated framework for data analysis
2. Workflow involves procedures for:
   - Planning, organizing and documenting research
   - Cleaning data
   - Analyzing data
   - Presenting results
   - Backing up and archiving materials
   - Reproducing results

You already have a workflow (WF)
1. Your WF might be
   A. Planned
   B. Ad hoc
2. You can improve your WF with a modest investment of time
   A. The less experience you have, the easier it is
   B. It will save time
   C. It will make you a better data analyst

Why should you care?
1. Replication
   - Replication is essential for good science
   - Workflow is essential for replication
2. Getting the right answers
   - Retractions are to be avoided
3. Time
   - “Science is a voracious institution.”
4. An effective WF helps you find inevitable errors and fix them

Origins of the workflow project
1. Easy things: consulting on easy things, instead of hard things
2. Incorrect results with clever “explanations”
3. Dissertation delayed 18 months to determine why results changed
4. Unreproducible results from a 743 line do-file with no comments
5. Analyzing the wrong data set: “The data sets are exactly the same except that I changed the married variable.”
6. Analyzing the wrong variable while writing a report for the NAS
7. Miscoded genes that delayed a study of alcoholism
8. Collaborations that multiply the ways things go wrong
9. Misleading or ambiguous output such as...

5. Gaining the IU advantage
The publication of [The Workflow of Data Analysis Using Stata] may even reduce Indiana’s comparative advantage of producing hotshot quant PhDs now that grad students elsewhere can vicariously benefit from this important aspect of the training there.

   --Gabriel Rossman on his blog

Workshop in Methods – September 2012
Example 1: definitely a problem in a $3M study

```
Example 2: variable labels that come from the survey center
```

```
Example 3: which number is which?

```

```
Example 4: good software doing things badly

```

```
Why learning WF is difficult
```

```
What is tacit knowledge?
```

```
1. Tacit knowledge is the stuff of textbooks and articles
2. Tacit knowledge is implicit and undocumented (Michael Polanyi)

```
Data analysis involves a lot of undifferentiated heavy lifting

The reality, of course, today is that if you come up with a great idea you don’t get to go quickly to a successful product. There’s a lot of undifferentiated heavy lifting that stands between your idea and that success.

-- Jeff Bezos, amazon.com

The Workflow of Data Analysis Using Stata

1. The book makes tacit knowledge about WF explicit
2. Lots of undifferentiated heavy lifting
3. Specifics on the general issues discussed today
4. Focusing on Stata, the principles apply broadly
   - An ethnographer sent me an e-mail thanking me for the book and telling me how useful he and his students found it. Would I consider writing a version for qualitative research?
   - An medical research from China found it crucial for getting his paper into Nature

Overview of today’s talk

1. Foundation of workflow
2. Replicability, the fundamental criteria
3. Other criteria for selecting a WF
4. Steps in your research
5. Tasks within each step
6. Principles for executing your WF

The foundation of WF is ironical optimism

The universal aptitude for ineptitude makes any human accomplishment an incredible miracle. --Dr. John Paul Stapp

40G’s: From 0 to 995mph and back in 3 seconds...

‘I was fine, only blind for a few days.’

WF starts with replication

1. An effective WF facilitates replication
2. Ask yourself:
   - Can you produce exactly the same results as you have published?
   - How long would it take?
3. Plan for replication from the start of a project
4. Disciplines are increasingly concerned with replicability:
   - Articles in Political Science, Economics, Sociology and other fields
   - The movement in statistics for “reproducible results”
Why replication is so hard

1. **The curse of dimensionality**: 10 decisions, 1,024 possibilities
   - Where to truncate a variable
   - The seed for the RN generator
   - Scaling decisions with partial missing data.
   - Which cases to keep for analysis
   - How to code education
   - What values to assign to income greater than $200,000
   - And so on...

Decisions in the path to analysis: the choices that could be made

Decisions in the path to analysis: the choices made

Why replication is so hard (continued)

2. **Documentation**: Replication should involve retrieving documentation, not trying to remember
3. **Changing software**: New software can give different results
   - A colleague’s 2 weeks of sleepless nights due to version variation.
   - This is particularly difficult with an active user community (e.g., R, Stata).
4. **Lost files**: corrupted, lost, unreadable, obsolete, or ambiguous files
   - Do you have $2,000 to retrieve the file that was “backed up”? 
   - Do virtual servers archive your data?

Given replicability, criteria for choosing WF

1. **Accuracy**
   - If your program is not correct, then nothing else matters. 
     ~Oliveira and Stewart

2. **Efficiency**
   - Completing work quickly given accuracy and replicability
   - Tension between working quickly and working carefully

3. **Standardization**
   - Don’t repeatedly, inconsistently decide how to do things
   - Standardization makes it easier to find mistakes

4. **Automation**
   - Automated procedures prevent mistakes and are faster
   - The invested time saves time

5. **Simplicity**
   - The more complicated your procedures the more likely you will make mistakes or abandon your plan

6. **Usability**
   - Your workflow should reflect the way you like to work
   - If you ignore your procedures, it is not a good WF

7. **Scalability**
   - Different projects require different workflows
Collaboration and workflow
1. Collaboration makes it more difficult to have an effective, efficient and replicable workflow
2. Disciplines that depend more on collaboration put more emphasis on explicit training in WF
3. Why is workflow harder in collaboration?
**Steps in your workflow**

**Start with a good research topic**
- Have good ideas
- Write them down
- Do this often

**Step 1. Prepare data for analysis**
- Variables must be cleaned with careful names and labels
- This takes 90% of the time, unless you hurry

**Step 2. Conduct analyses**
- Estimate models and create graphs
- Often the simplest step in your workflow

---

**Step 3. Present results**
- Incorporate output into your presentation
- Maintain the **provenance**
- Make effective presentations

**Step 4. Protecting files**
- **Backing up** and archiving: preserving the bits and the content
- Replication is impossible without your data and do-files
- "Today’s noise is tomorrow’s knowledge." -- **David Clemmer**
**Planning**

**The ideal**

Blau and Duncan (1967) *The American Occupational Structure*

- All analyses were specified 9 months before output was received
- The book was written based entirely on a single set of output
- Later books with full access to the data were not better

**Issues in planning**

1. A plan is a reminder to stay on track, finish the project, and publish results

   *Work. Finish. Publish.* —Michael Faraday’s sign in his lab

2. Planning saves time

3. Planning includes:
   - General goals, publishing plans, and deadlines
   - Division of labor and accountability
   - Proposal for data construction: names, labels, formats
   - Procedures for missing data
   - Anticipated analyses
   - Guidelines and responsibility for documentation
   - Procedures and schedule for backing-up and archiving
**Organizing**

1. Organization is motivated by two requirements:
   - Finding things
   - Avoiding duplication

2. It requires explicit, consistent decisions about naming and storing things
   - 115,000 files on a research center’s LAN
   - 2,000,000 files accumulated in 10 years

3. Organization:
   - Helps you work faster
   - Rewards consistency and uniformity
   - It is contagious

---

**Signs of poor organization**

1. Can’t find a file and think you deleted it
2. Multiple versions of a file and don’t know which is which
   - You and a colleague are working on different versions of the same paper. You changed what she changed and now there are three versions of the paper.
   - You need the final version of the paper the was submitted for review, but you have two (or 16) files with "final" in the name.

   - This: final_report_v16.docx
   - Or this: NSF_science_report 2010-10-21.docx

---

**The curse of cheap storage**

1. It is easier to create a file than to find a file
2. It is easier to find a file than to know what is in the file
3. With disk space so cheap, it is tempting to create a lot of files

---

**Organizing: a standard directory structure for all projects**

```
\WF project
  \- History starting 2009-03-06
  \- Hold then delete
  \- Pre posted
  \- To clean
  \Documentation
  \Posted
  \Resources
  \Text
    \- Versions
  \Work
    \- To do
```

A batch file makes creating uniform directories easy

---

**Organizing: wfsetupsingle.bat makes it easy**

```
REM workflow talk 2 \ wfsetupsingle.bat jsl 2009-07-12
REM directory structure for single person.
FOR /F "tokens=2,3,4 delims=- " %%a in ("%DATE%") do set CDATE=%%c-%%a-%%b
md \- History starting \%cdate%
md \- Hold then delete 
md \- Pre posted 
md \- To clean
md "Documentation"
md "Posted"
md "Resources"
md "Text\- Versions"
md "Work\- To do"
```

Templates make this structure easy to use.

---

**Organizing: uniform formats for do-files**

```
capture log close
log using wftalk-example, replace text
// program:    wftalk-example.do
// task:      // project:      // author:     scott long 2012-09-04
version 11
clear all
set linesize 80
local tag "wftalk-example.do jsl 2012-09-04"
```

Templates make this structure easy to use.
Organization should be like a Model T

Any color you want as long as it is black....

Too often it is more like a VW ‘bug’

With predictable consequences

Documentation

1. Long’s Law: It is always faster to document it today than tomorrow
   
   Addendum 1: Nobody likes to write documentation
   
   Addendum 2: Nobody regrets having written documentation
   
   *Drat, this program has too many comments.*

2. Multiple, reinforcing levels: logs, metadata, comments, names

3. Without documentation, replication is virtually impossible, mistakes are more likely, and work takes longer

4. The more codified the field the greater the emphasis on documentation
   
   A. The Research Log by the American Chemical Society
   
   B. Loss of tenure for an altered research log

Suggestions for writing documentation

1. Do it today

2. Check it next week: it always makes sense today

3. Keep up with documentation by tying it to events in the project

4. Include full dates and names

The core of your documentation: the research log

A real example...
**Execution and computing**

1. Execution involves carrying out tasks within each step
2. Effective execution requires **the right tools**
   - Software
     - a. File manager
     - b. Text editor
     - c. Statistical software
     - d. Macro program
     - e. Word processor
   - Hardware: display, storage, memory, CPU
3. Planning is probably more important than computing power.

**For example…**

**Cornell 1975: the entire computing infrastructure**

- IBM 370 with 240K memory
- Winchester drives with 3MB storage

- Cost of computing $1,000,000.
- Mean time to degree 7.6 years.

**Indiana 2009: a disposable PC**

- Asus 1000HE with 2GB memory
  - 10,000 times more
- FreeAgent with 1TB storage
  - 350,000 times more...

- Cost of computing $400 (2,500 times less); $250 in 2011.
- Mean time to degree 7.6 years.

**How much computing power do you have?**

- Cray Computer 1985
- iPad2 2011


**A thought experiment on planning and computing**

1. Divide yourselves into two groups
   - The computers compute whenever they want to
   - The planners only compute for two six-hour sessions a week
2. Who finishes first?

**Principles for a computing workflow**

Three principles that are easy and have huge benefits.

1. Dual workflow
   - Data management
   - Data analysis
2. Run order for naming files
3. Posting principle for sharing results
Run order and a dual workflow

Name files so that if they are re-run in alphabetical order, you will produce **exactly** the same results

<table>
<thead>
<tr>
<th>Data management</th>
<th>Data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>data01.do</td>
<td>stat01a.do</td>
</tr>
<tr>
<td>data02V2.do</td>
<td>stat01b.do</td>
</tr>
<tr>
<td>data03.do</td>
<td>stat01cV2.do</td>
</tr>
<tr>
<td>data03-1.do</td>
<td>stat02a.do</td>
</tr>
<tr>
<td>data03-2.do</td>
<td>stat02b.do</td>
</tr>
<tr>
<td>data04.do</td>
<td>stat03aV2.do</td>
</tr>
<tr>
<td>stat02b.do</td>
<td>stat03b.do</td>
</tr>
<tr>
<td>stat03c.do</td>
<td>stat03c1.do</td>
</tr>
<tr>
<td>stat03c2V2.do</td>
<td>stat03d.do</td>
</tr>
</tbody>
</table>

The essential **posting principle**

1. Essential for keeping track of the "official" results
2. Defined by two simple rules:
   
   **The share rule**
   Only share results after the files are posted
   
   **The no change rule**
   Once a file is posted, *never* change it

Data analysis: use do-files (aka script files,...)

**Robust do-files**

1. They are self-contained
2. They include version control (**version 11.1**)
3. They exclude directory information (which might change)
4. They explicitly set seeds for random numbers
5. They require that you archive user written ado-files

**Simply put**: Your programs should run on another computer at a later date without requiring *any* changes

Legible do-files: output that is easy to read

1. Thoughtful comments
2. Alignment, indentation and spacing
3. Short lines without wrapping
4. No ambiguous abbreviations: **l a l in 1/3**

Legible log files (in text not smcl)

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>Years of education</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-------------------</td>
</tr>
<tr>
<td>Menial</td>
<td>3</td>
<td>2</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>6.45</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>9.68</td>
<td>38.71</td>
<td>6.45</td>
<td>100.00</td>
</tr>
<tr>
<td>BlueColl</td>
<td>5</td>
<td>26</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.45</td>
<td>4.35</td>
<td>1.45</td>
<td>10.14</td>
</tr>
<tr>
<td></td>
<td>7.25</td>
<td>37.68</td>
<td>10.14</td>
<td>100.00</td>
</tr>
<tr>
<td>Craft</td>
<td>7</td>
<td>39</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>3.57</td>
<td>2.38</td>
<td>3.57</td>
</tr>
</tbody>
</table>
Automation

1. Much of data analysis involves repetitive tasks
   - Repetition invites errors

2. Automation is faster and less error prone
   A. macros: words that represent strings of text
   B. loops: multiple execution of the same commands
   C. returned results: avoiding typing the value of any statistical result
   D. matrices: hold and summarize key results
   E. ado-files: write programs that do what you want
   F. me.hlp: don’t keep looking up the same things. For example,…

Data cleaning, including names and labels

Planning labels

Bad labels
   . codebook tcl*, compact

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
</table>
| tc3doc   | 1074| 10     | 8.71453| 1   | 10  | Q46 How important is it to go to...
| tc3fam   | 1074| 10     | 8.755121| 1  | 10  | Q43 How important is it to turn t...
| tc3friend| 1073| 10     | 7.799267| 1  | 10  | Q44 How important is it to turn t...
| tc3pay   | 1050| 10     | 7.567619| 1  | 10  | Q47 How important is it to go to ...
| tc3relig | 1039| 10     | 5.66025 | 1  | 10  | Q45 How important is it to turn t...

Better labels
   . codebook tcl2*, compact

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
</table>
| tc3doc   | 1074| 10     | 8.71453| 1   | 10  | Q46 How important is it to go to...
| tc3fam   | 1074| 10     | 8.755121| 1  | 10  | Q43 How important is it to turn t...
| tc3friend| 1073| 10     | 7.799267| 1  | 10  | Q44 How important is it to turn t...
| tc3pay   | 1050| 10     | 7.567619| 1  | 10  | Q47 How important is it to go to ...
| tc3relig | 1039| 10     | 5.66025 | 1  | 10  | Q45 How important is it to turn t...

Even better labels
   . codebook tcl3*, compact

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
</table>
| tc3doc   | 1074| 10     | 8.71453| 1   | 10  | Q46 Med doctor help important
| tc3fam   | 1074| 10     | 8.755121| 1  | 10  | Q43 Family help important
| tc3friend| 1073| 10     | 7.799267| 1  | 10  | Q44 Friends help important
| tc3pay   | 1050| 10     | 7.567619| 1  | 10  | Q47 Psychiatric help important
| tc3relig | 1039| 10     | 5.66025 | 1  | 10  | Q45 Relig leader help important

Planning labels
Truncation and careless names

Example: ownsex and ownsexu caused weeks of confusion

Creating a codebook

Types of data cleaning

Cleaning 1a: finding an error with a graph

Cleaning 1b: reversing the graph

Cleaning 2: remembering a coding decision
Cleaning 3: understanding the substantive process

Cleaning 4: avoiding expensive mistakes

Analyzing the data
1. Take lots of classes in statistics
2. Go to lots of talk on data analysis
3. Find exemplars; don't rediscover the wheel; don't do it "your way"

Presentations and provenance
1. Content and methods are substantive, disciplinary decisions
2. Presentations and preservation of provenance are universal

Tables too small

Colors that aren't distinct when printed/projected

Labels that aren't large enough
Documenting provenance

The provenance of every number you use should be fully documented.

Example

The circled text contains results I may need to confirm later:

1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, 43 and 53, respectively (z=2.55, p<.01). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (% for non-

Turning on "show/hide ¶" reveals the provenance:

1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, 43 and 53, respectively (z=2.55, p<.01. jsl2.0604 and 4.01). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Captions that indicate the provenance

The graph shows the number of publications to the k-th root. Turning on "show/hide ¶" reveals the provenance.

Preserving your data

When it comes to saving your work, expect things to go wrong, expect that you will delete the wrong file at the worst possible time, and expect a hose to be left on in the room above your computer. If you expect the worst, you might be able to prevent it.

Hope, foolishly, springs eternal (the Sweden syndrome)

Examples of data loss

1. Kennedy assassination on November 22, 1963 and the 9/11 survey
2. 508K volumes in obsolete formats at British Museum. 2M videos at IU
3. Neil Armstrong’s walk on the moon on July 20, 1969 and the lost moon tapes., and Pink Floyd’s Dark Side of the Moon. The world saw:

What NASA saw and lost...
But, two tapes were archived by Pink Floyd’s video producer!

Dark Side of the Moon

A simple approach to preserving files

1. Install the program
2. Put files into the folder
3. Retrieve them from any machine connected to the service
4. Have shared folders for collaboration
   - If they delete them, you lose them!
5. Do not send attachments even for one time file exchanges
IU’s Scholarly Data Archive (SDA)
Enterprise wide tape storage

Changing your workflow
1. Slowly, systematically, thoughtfully
2. Finish the last 5% of the change
3. Do not do it under deadline

Whose workflow
1. There are many viable workflows
2. The key advantage of the WF book is that it is written down
3. Alan Acock wrote:
   - “Not everyone will agree with all of [Long’s] suggestions.”
   - “I will post the announcement of Workflow on my door with the following note: ‘I am glad to help anybody who followed at least 25% of the advice Long provides—and brings me their do-files!’”
4. Do you really want to spend your time rediscovering the mistakes I made?

Thanks for listening
Questions?

WF class at IU next summer...