What is workflow (WF)?

Workflow involves the entire craft of data analysis

1. **Planning**, organizing and documenting research
2. **Cleaning** data and creating variables
3. **Estimating** models and creating graphs
4. **Presenting** and publishing findings
5. **Archiving** and backing up materials

Your workflow?

1. Your WF might be
   
   A. **Planned** and carefully orchestrated.
   B. **Ad hoc**, piece-meal, developed in reaction to mistakes.
   C. Good, bad, or ugly.

2. Almost certainly 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 you time and make you a better data analyst.
Why should you care about workflow?

1. Replication
   - A good workflow is essential for replication.
   - Replication is essential for good science.

2. Getting the right answers
   - Retractions are embarrassing and can end careers.

3. Time
   - “Science is a voracious institution.”
   - Boring things should take as little time as possible.

4. The IU advantage
   “The publication of [WFDAUS] 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

Statistical origins of the workflow project

Spending my time...

1. Fixing easy things: time consulting on easy things, not hard things.
2. Looking at incorrect results and clever “explanations”.
3. A dissertation delayed 18 months to find why results changed.
4. Irreproducible results from a single, 743 line do-file.
5. Analyzing the wrong dataset: “The datasets are exactly the same except that I changed the married variable.”
6. Wasting time analyzing the wrong variable while writing an NAS report.
7. Miscoded genes the delayed progress on a study of alcoholism.
8. Collaborations that multiply the ways things can go wrong.
9. Misleading or ambiguous output such as...

Example 1: definitely not good!

```
.tabulate female adchild_y1

<table>
<thead>
<tr>
<th>R is female?</th>
<th>Q15 Would let X care for children</th>
<th>Definitely</th>
<th>Probably</th>
<th>Probably</th>
<th>Definitely</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0Male</td>
<td>41</td>
<td>99</td>
<td>155</td>
<td>197</td>
<td>492</td>
<td></td>
</tr>
<tr>
<td>1Female</td>
<td>73</td>
<td>98</td>
<td>156</td>
<td>216</td>
<td>542</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>197</td>
<td>311</td>
<td>412</td>
<td>1,034</td>
<td></td>
</tr>
</tbody>
</table>
```
Example 2: which number is which?

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Example 3: good software doing things badly

Page 963 of Stata 11 manual

. use margex
(Artificial data for marginals)

. regress y i.sex i.group

Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  3,  2996) =  152.06
Model |  183866.077     3  61288.6923           Prob > F      =  0.0000
Residual |  1207566.93  2996  403.059723           R-squared     =  0.1321
-------------+------------------------------           Adj R-squared =  0.1313
Total |  1391433.01  2999  463.965657           Root MSE      =  20.076

------------------------------------------------------------------------------

|            Delta-method
|     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sex |   60.69034   .5781782   104.74   0.000     59.55435    61.82634
| 91.38692   104.74   0.000     89.87821    92.89564
| 105.12482   104.74   0.000     103.6014   106.6482

“The numbers reported in the ‘Margin’ column are average values of y. Based on a linear regression of y on sex and group, 60.6 would be the average value of y if everyone in the data were treated as if they were sex=0, and 78.9 would be the average value if everyone were treated as if they were sex=1.”
Photographic origins of the WF project

2004

Workflow, not slow. –Bruce Fraser

2005 (1st Edition)

The name wasn’t a coincidence.

Learning WF can be difficult

It requires:

1. mastering tacit knowledge
2. lots of heavy lifting

Tacit knowledge

1. Explicit knowledge is the stuff of textbooks and articles.
2. Tacit knowledge is implicit and undocumented (Michael Polanyi).
3. People can be unaware of their tacit knowledge and how valuable it is.
   - Henry Bessemer’s patent for making steel (1855)
4. Tacit knowledge is transferred through personal contact.
   - WF is a craft and crafts are learned “at the bench”.
   - Personal computers impede this transfer of the craft.

Undifferentiated heavy lifting (Jeff Bezos)

There is a lot of heavy lifting in doing data analysis well.

“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.”

The book Workflow of Data Analysis Using Stata

1. Makes tacit knowledge about WF explicit.
2. It deals with a lot of undifferentiated heavy lifting.
3. It contains specifics on the general issues discussed today
WF starts with replication

1. **Science demands replicability** and a good WF facilitates replication.
2. **Anticipate the need to replicate from the start**, not after your work has been challenged.
3. Many disciplines are worried about replicability.
   - Articles in Political Science, Economics, Sociology and other fields.
4. Calls for journals to deposit all analyses are growing.
   - Are your do-files and log files ready for public display?

Why is replication so hard?

1. **The curse of dimensionality**: 10 minor decisions, leads to 1,024 reasonable ways to create your data.
   - A decision where to truncate a variable.
   - That pesky *seed* for the RN generator.
   - How to handle missing data in a 5-variable scale.
   - Decisions on which cases to keep for analysis.
   - 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 is replication so hard (continued)?

2. **Lack of documentation**: Replication should involve retrieving documentation, not trying to remember what you did.
3. **Changing software**: 2 weeks of sleepless nights due to version variation.
4. **Lost files**: corrupted, lost, unreadable, or ambiguous files.
5. **Mistakes** are made. They are unavoidable, but a good WF can help you catch and fix them.

The foundation of WF is **ironical optimism**

The *universal aptitude for ineptitude* makes any human accomplishment an incredible miracle. --Dr. John Paul Stapp
Steps in your workflow

Step 0. Having a good idea for a project

Step 1. Cleaning the data
  - The data must be accurate.
  - The variables must be carefully named and labeled.
  - This takes 90% of the time, unless you hurry.

Step 2. Running analyses
  - Estimating models and computing graphs.
  - This is often the simplest part of the workflow.

Step 3. Presenting results
  - Incorporating output into your presentation.
  - Making a clear presentation.
  - Maintaining the provenance of results.

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

Tasks within each step
Tasks within each step

Plan
Organize
Document
Compute

Tasks within each step

Plan
Organize
Document
Compute

Tasks within each step

Plan
Organize
Document
Compute
Planning

The ideal

Blau and Duncan (1967) *The American Occupational Structure*: All analyses were specified 9 months before output was received. Then, the book was written based entirely on those analyses.
**Issues in planning**

1. A little planning goes a long way and almost always saves time.
2. 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
3. Most people prefer to “do something” rather than plan.
4. Planning includes:
   - a. General goals and publishing plans
   - b. Scheduling
   - c. Division of labor
   - d. Data sets needed
   - e. Variable names and labels
   - f. Missing data procedures
   - g. Analysis
   - h. Documentation
   - i. Backing-up and archiving materials

**Organizing**

1. Organization is driven by needing to find things and avoid duplication.
2. Careful organization helps you work faster.
3. Organization rewards...
   - A. Consistency and uniformity
   - B. A simple structure that is not too simple.
   - C. Thinking systematically about how you name and store things.
4. Organization is contagious: start organized to end organized.

**Signs of poor organization**

1. You have multiple versions of a file and don’t know which is which.
2. You can’t find a file and think you might have deleted it.
3. You and a colleague are both working on different versions of the same paper. You changed what she changed and now you have three versions of the paper.
4. You need the final version of the paper the was submitted for review, but you have two files with “final” in the name.
Organizing: 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
      \2009-03-06 project directory created
   \- Hold then delete
   \- Pre posted
   \- To clean
   \Documentation
   \Posted
   \Resources
   \Text
      \- Versions
   \Work
      \- To do

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\%cdate% project directory created"
md " Hold then delete 
md " Pre posted 
md " To clean"
md "Documentation"
md "Posted"
md "Resources"
md "Text\ Versions"
md "Work\ To do"
Organizing: uniform formats for do-files

```
capture log close
log using wftalk-example, replace text
// program: wftalk-example.do
// task: alt-1 creates this file
// project: wftalk example
// author: jsl \ 2009-07-23
// #0
// program setup
version 11
clear all
set linesize 80
* matrix drop _all
local tag "wftalk-example.do jsl 2009-07-23"
// #1
// <task description>
// #2
// <task description>
log close
exit```

Documentation

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

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

3. The more codified the field the greater the emphasis on documentation.
   - A. The Research Log by the ACS
   - B. Loss of tenure for an altered research log.

4. Documentation occurs at many levels: logs, metadata, comments, names.

Suggestions for writing documentation

1. It is faster to document it today than tomorrow.
2. To keep up with documentation, tie it to events in your work.
3. Write it today; check it later.
4. Include full dates and names.

The core of your documentation: the research log

A real example (expletive’s deleted)...
Execution and computing

1. Execution involves carrying out specific tasks within each step.
2. Effective execution requires the right tools for the job.
   - Software
     - Text editor
     - File manager
     - Statistical software
     - Word processor
     - Macro program
   - Hardware
     - Display
     - Storage
     - Central processor

A simple thought experiment

The key an effective workflow is planning, not computing.
1. Randomly divide yourselves into two groups.
2. The computers are allowed to compute whenever you like when writing your dissertation.
3. The planners have access to a computer for two six-hour sessions a week.
4. The wager: the planners finish first.

Does cheap computing help?
The historical context of computing...
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
- FreeAgent with 1TB storage

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

Criteria for choosing a WF assuming replicability

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

Efficiency
- Complete work quickly given accuracy and replicability.
- Tension between working quickly and working carefully.

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

- You don’t have to repeatedly decide how to do things.
- With standardization, it is easier to find mistakes.

Automation (learn to program!)

- Automated procedures prevent mistakes.
- **Drukker’s Dictum**: Never type anything that you can obtain from a saved result.

Usability

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

Scalability

- Different projects might require different workflows.

---

**Dual workflow and run order**

1. **Dual workflow**: keep data management and data analysis separate.
2. **Run order**: name files so that if they are re-run in alphabetical order, you will produce *exactly* the same results.
3. **Posting principle** with two rules
   - **The share rule**: Only share results after the associated files are posted.
   - **The no change rule**: Once a file is posted, *never* change it.

---

**Dual workflow**

Data management $$\rightarrow$$ Data analysis

$$\leftarrow$$ Data analysis
Files from a dual workflow

<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>stat02al.do</td>
</tr>
<tr>
<td>data04.do</td>
<td>stat02b.do</td>
</tr>
<tr>
<td>stat03aV2.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. Posting simply means saving files as final.
2. The posting principle involves two rules
   a. The share rule: Only share results after the associated files are posted.
   b. The no change rule: Once a file is posted, never change it
3. Never does not mean:
   a. Rarely done.
   b. Just a little change.
   c. Really soon after posting.

Data analysis: use do-files!

Robust do-files

1. Self-contained
2. Version control
3. Exclude directory information
4. Include seeds for random numbers
5. Archive user written ado-files

Legible do-files

1. Lots of comments; even more than that!
2. Alignment and indentation
3. Short lines with no wrapping
4. Avoid abbreviations:  la l in 1/3
### Legible log files (in text not smcl)

<table>
<thead>
<tr>
<th>Key</th>
<th>frequency</th>
<th>row percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation</td>
<td>3 6 7 8 9 10</td>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Menial</th>
<th>0 2 0 0 3 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 12 2</td>
<td>9.68 38.71 6.45 0.00 0.00 9.68 3.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BlueCol</th>
<th>1 3 1 7 4 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 26 7</td>
<td>7.25 37.68 10.14 100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Craft</th>
<th>0 3 2 3 2 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 39 7</td>
<td>8.33 46.43 8.33 100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>1 8 4 12 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 109</td>
<td>5.64 32.34 8.90 100.00</td>
</tr>
</tbody>
</table>
Automation

Much of your work involves repetitive tasks that invite error. Automation makes work easier, faster, less error prone.

1. macros
2. loops
3. returned results
4. matrices
5. ado-files
6. include files
7. help me files

Automation: an extended example (as time allows)

1. Low levels of automation and documentation
   
   `wf-automation-1low.do`

2. Medium levels of automation and documentation
   
   `wf-automation-2medium.do`

3. High levels of automation and documentation
   
   `wf-automation-3high.do` + `wf-automation-3high.doi`
   `wf-automation-3highV2.do` + `wf-automation-3highV2.doi`

Data cleaning, including names and labels

Planning names

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Name</td>
<td>Value label</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>id_iu</td>
<td>Respondent Number</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>entry_iu</td>
<td>entry_iu</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>vignon</td>
<td>vignon</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>serious</td>
<td>serious</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>opfrom</td>
<td>Ldummy</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>opfriend</td>
<td>Ldummy</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>tespri</td>
<td>Ldummy</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>tendm</td>
<td>Ldummy</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>opname</td>
<td>Ldummy</td>
</tr>
</tbody>
</table>

Truncation and careless names

Example: `ownsex` and `ownsexu` caused weeks of confusion.
Creating a codebook

Types of data cleaning

Cleaning 1: finding an error with a graph
Cleaning 1: 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. Find exemplars; don’t do “your way”.

Presentations and provenance

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

Tables too small
Colors that aren't distinct when printed/projected

Labels that aren't large enough

Documenting the provenance

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 (46 and 73, respectively, $\chi^2 = 2.55, p = .01$). However, this gap has disappeared for the

1943-1947 cohort and, indeed, employed women have slightly more limitations ($74$ 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 (46 and 73, respectively, $\chi^2 = 2.55, p = .01$). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Captions make it easy to trace a source

```
twoway (line art_root2 art_root3 art_root4 art_root5 articles, lwidth(medium)), ytitle(Number of Publications to the k-th Root) yscale(range(0 8.)) legend(pos(11) rows(4) ring(0)) caption(wf7-caption.do \ jsl 2008-04-09, size(vsmall))
```

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.

**THE FOUR STAGES OF DATA LOSS**

**STAGE 1: DENIAL**

I did not, I won’t, I will not lose my data, I gladly place a backup copy where?

**STAGE 2: ANGER**

You stupid, slow, or crates, where’s my data?

**STAGE 3: DEPRESSION**

Why? Why me?

**STAGE 4: ACCEPTANCE**

I’m never going to graduate.

Examples of data loss

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, the lost moon tapes, and Pink Floyd’s Dark Side of the Moon.

"a fuzzy gray blob wading through an inkwell"
Rube Goldberg machine: the obvious way to backup

A comically involved, complicated invention, laboriously contrived to perform a simple operation.

Issues in deciding how and what to preserve

<table>
<thead>
<tr>
<th>Questions to consider</th>
<th>Working</th>
<th>Posted</th>
<th>Archival</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do you recover the file?</td>
<td>Redo work</td>
<td>Redo old work</td>
<td>Download file</td>
</tr>
<tr>
<td>What is the cost of recovery?</td>
<td>Minor</td>
<td>Potentially lots of work</td>
<td>A little time</td>
</tr>
<tr>
<td>How long are you preserving it?</td>
<td>1-3 yrs.</td>
<td>3-10 yrs.</td>
<td>Forever</td>
</tr>
<tr>
<td>How difficult is it to preserve?</td>
<td>Trivial</td>
<td>Tediums</td>
<td>Very hard</td>
</tr>
<tr>
<td>Concern with media/format?</td>
<td>Minor</td>
<td>Some</td>
<td>Critical</td>
</tr>
</tbody>
</table>
A KISS approach to preserving files

Part 1: Daily mirror

Part 2: Long-term backups

Tactics: Portable drive computing at home

Tactics: Portable drive computing at work
Tactics: Portable drive computing when traveling

Tactics: Live sync

Off-line backups
Mozy and similar vendors, corporate mass storage, local servers.

Data storage 1981 to 2009
1. Size per drive increased by a factor of more than 300,000.
2. Cost per gigabyte decreased by a factor of 7,000,000.
3. A shoebox full of portable drives can hold enough IBM cards to fill BH six times over. In the past 3 months, this changed to 12 times over...
Changing your workflow

1. Slowly.
2. Finish the last 5% of the change.
3. Like Penn and Teller, master a few cool tricks.
4. Don't 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 Acok wrote:
   o Not everyone will agree with all of [Long’s] suggestions.
   o I will post the announcement of Workflow on my door with the following note: “I’m 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?