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

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 hotspot 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 alcholism.
8. **Collaborations** that multiply the ways things can go wrong.
9. **Misleading or ambiguous output** such as...

---

Example 1: definitely not good!

```
. tabulate female sdchild_v1
R is female?                Definitel  Probably  Probably  Definitel  Total
omale   41        99       155      197      402
1Female  76        98       215      197      542
Total    114       197      311      412      1,034
```

Example 2: which number is which?

```
<table>
<thead>
<tr>
<th>Occupation</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operator</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>9.68</td>
<td>38.71</td>
<td>6.46</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BlueCollar</td>
<td>5</td>
<td>26</td>
<td>7</td>
<td>69</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7.25</td>
<td>37.68</td>
<td>4.35</td>
<td>1.45</td>
<td>1.45</td>
</tr>
<tr>
<td>Craft</td>
<td>7</td>
<td>39</td>
<td>0</td>
<td>84</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>8.33</td>
<td>46.43</td>
<td>8.33</td>
<td>3.37</td>
<td>2.38</td>
</tr>
</tbody>
</table>
```

Example 3: good software doing things badly

Page 963 of Stata 11 manual

```
. use margex
(Artificial data for margins)
. regress y i.sex i.group
```

```
Source | SS     | df | MS   | Number of obs = 3000
Model  | 183866.077 | 3  | 61288.6923 | F( 3, 2996) = 152.06
Residual | 1207566.93  | 2996 | 403.059723 | Prob > F = 0.0000
Root MSE = 20.076
-------------+-------------------------------------------
```

```
| y | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---|-------|-----------|-------|-------|----------------------|
| 1.sex | 18.32202 | .8930951 | 20.52 | 0.000 | 16.57088 | 20.07316 |
| group | 8.037615 | .913769 | 8.80 | 0.000 | 6.245937 | 9.829293 |
| 2 | 18.63922 | 1.159503 | 16.08 | 0.000 | 16.36572 | 20.91272 |
| 3 | 53.32146 | .9345465 | 57.06 | 0.000 | 51.48904 | 55.15388 |
```
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 replication is 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...
Rubik's Cube

1. In 20 steps you can reach all possible configurations of the cube.
2. The are 43,252,003,274,489,856,000 possible configurations.

Dr. Morley Davidson at Kent State University along with an engineer at Google and colleagues from Germany and California explored all possibilities, proving that 20 is the ‘God number’.

The foundation of WF is ironical optimism

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

Why replication is 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.

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.

"Today's noise is tomorrow's knowledge." -- David Clemmer
Tasks within each step

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. None of the later books written with full access to the data was nearly as good.
**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 (which is fun)
   h. Documentation (not so fun)
   i. Backing-up and archiving materials

---

**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.
5. During last year’s Blalock lecture you get a text message that says: “Don’t use Project_final.docs for tomorrow’s presentation. It isn’t the final version.”

---

**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.

---

**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:
// project: jsl 2010-07-27
// #0
// program setup
version 11
clear all
set linesize 80
local tag "wftalk-example.do jsl 2010-07-27"
// #1
// Description of task 1
// #2
// Description of task 2
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."
     • Too many comments can be a problem, but I’ve rarely seen it.

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...

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
   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.

### Files from a dual workflow

**Data management**
- data01.do
- data02V2.do
- data03.do
- data03-1.do
- data03-2.do
- data04.do

**Data analysis**
- stat01a.do
- stat01b.do
- stat01cV2.do
- stat02a.do
- stat02a1.do
- stat02b.do
- stat03aV2.do
- stat03b.do
- stat03c.do
- stat03c1.do
- stat03c2V2.do
- stat03d.do

### The essential posting principle

1. 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
2. Never does not mean:
   a. Rarely.
   b. Just a tiny 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: **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>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>13</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>19</td>
<td>109</td>
<td>30</td>
<td>337</td>
</tr>
</tbody>
</table>

---

### Years of education

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>6.45</td>
<td>0.00</td>
<td>0.00</td>
<td>9.68</td>
<td>3.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlueCol</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>26</td>
<td>7</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>1.45</td>
<td>4.35</td>
<td>1.45</td>
<td>10.14</td>
<td>5.80</td>
<td>8.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Craft</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>39</td>
<td>7</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>3.57</td>
<td>2.38</td>
<td>3.57</td>
<td>2.38</td>
<td>2.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Frequency

<table>
<thead>
<tr>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency row percentage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WhiteCol</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>109</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.44</td>
<td>0.00</td>
<td>2.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>6.45</td>
<td>0.00</td>
<td>0.00</td>
<td>9.68</td>
<td>3.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlueCol</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>26</td>
<td>7</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>1.45</td>
<td>4.35</td>
<td>1.45</td>
<td>10.14</td>
<td>5.80</td>
<td>8.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Craft</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>39</td>
<td>7</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>3.57</td>
<td>2.38</td>
<td>3.57</td>
<td>2.38</td>
<td>2.38</td>
<td></td>
<td></td>
<td></td>
<td></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

Data cleaning, including names and labels
Planning names

Truncation and careless names
Example: ownsex and ownsexu caused weeks of confusion.

Get SNAG! An easy to use results collector.
In Stata, type:

findit snag

A work in progress. Caveat emptor.

Creating a codebook

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at all Important</th>
<th>Very Important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q43. Turn to family for help</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
<tr>
<td>Q44. Turn to friends for help</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
<tr>
<td>Q45. Turn to a minister, priest, rabbi or other religious leader</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
<tr>
<td>Q46. Go to a general medical doctor for help</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
<tr>
<td>Q47. Go to a psychiatrist for help</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
<tr>
<td>Q48. Go to a mental health professional</td>
<td>1 2 3 4 5 6 8 9 9 10</td>
<td></td>
</tr>
</tbody>
</table>

Inferred to be: Code “Don’t Know” as “Don’t Know”

The next few questions deal with the government’s responsibility to help people like NAME. For each statement please tell me if you think the government definitely should, probably should, probably should not, or definitely should not be responsible for helping people with symptoms like YOURS...
Types of data cleaning

Cleaning 1: finding an error with a graph

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

1. Take lots of classes in statistics.
2. Find exemplars; don’t do “your way”.

Cleaning 4: avoiding expensive mistakes

1. Content and methods are substantive, disciplinary decisions.
2. Presentations and preservation of provenance are more generic.
Colors that aren't distinct when printed/projected

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 ($r^2 = 2.55, \ p < .01$). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations ($r = .76$ for men).

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

However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have:

Labels that aren't large enough

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](https://www.phdcomics.com/comics/archive.php?comicid=1792)

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"

"Dark Side of the Moon"

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 your 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>Tedious</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

Tactics: Portable drive computing at work

Tactics: Portable drive computing at home

Tactics: Portable drive computing when traveling
### Tactics: Live sync

- 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 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?

Photographic dimensions of workflow

2004

2005 (1st Edition)

Other aspects of workflow

Thanks for listening.

Questions?