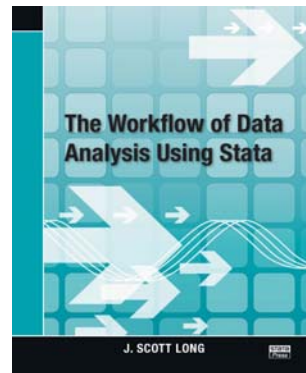


Principles of Workflow in Data Analysis

Scott Long

Indiana University



August 2010

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

Workflow \ 1

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.

Workflow \ 2

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*

Workflow \ 3

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...

Workflow \ 4

Example 1: definitely not good!

```
. tabulate female sdchild_v1
```

R is female?	Q15 Would let X care for children				Total
	Definitel	Probably	Probably	Definitel	
0Male	41	99	155	197	492
1Female	73	98	156	215	542
Total	114	197	311	412	1,034

Workflow \ 5

Example 2: which number is which?

Occupation 11				Total	Years of education		
	12	3	6		8	9	10
Menial 3	0	2	0	0	3	1	
	12	0.00	2	6.45	31	0.00	0.00
	9.68	38.71	6.45	100.00			9.68 3.23
BlueCol 5	1	7	3	1	7	4	6
	26	1.45	7	4.35	69	1.45	10.14
	7.25	37.68	10.14	100.00			5.80 8.70
Craft 7	0	7	3	2	3	2	2
	39	0.00	7	3.57	84	2.38	3.57
	8.33	46.43	8.33	100.00			2.38 2.38

Workflow \ 6

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			
Model	183866.077	3	61288.6923	Number of obs =	3000	
Residual	1207566.93	2996	403.059723	F(3, 2996) =	152.06	
Total	1391433.01	2999	463.965657	Prob > F =	0.0000	
				R-squared =	0.1321	
				Adj R-squared =	0.1313	
				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						
2	8.037615	.913769	8.80	0.000	6.245937	9.829293
3	18.63922	1.159503	16.08	0.000	16.36572	20.91272
_cons	53.32146	.9345465	57.06	0.000	51.48904	55.15388

Workflow \ 7

```
. margins sex
```

```
Predictive margins          Number of obs = 3000
Model VCE : OLS
Expression : Linear prediction, predict()
```

	Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
sex						
0	60.56034	.5781782	104.74	0.000	59.42713	61.69355
1	78.88236	.5772578	136.65	0.000	77.75096	80.01377

“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**.”

Workflow \ 8

Taking “advantage” of factor variables in Stata 11...

```
. logit tenure i.female i.female#c.articles i.male i.male#c.articles, nocons

note: 0.male#c.articles omitted because of collinearity
note: 1.male#c.articles omitted because of collinearity

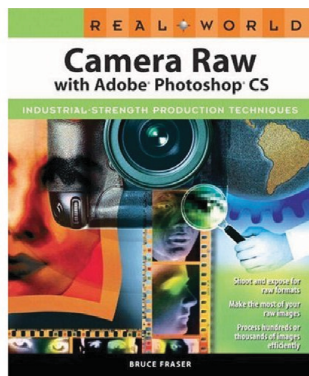
Logistic regression                Number of obs   =       2945
                                Wald chi2(4)    =      1183.12
Log likelihood = -1038.1979        Prob > chi2    =       0.0000
```

tenure	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
1.female	-2.473265	.1351561	-18.30	0.000	-2.738166 -2.208364
female#					
c.articles					
0	.0980976	.0098808	9.93	0.000	.0787316 .1174636
1	.0421485	.0098962	4.26	0.000	.0227524 .0615447
1.male	-2.693147	.1170916	-23.00	0.000	-2.922642 -2.463651
male#					
c.articles					
0	(omitted)				
1	(omitted)				

Workflow \ 9

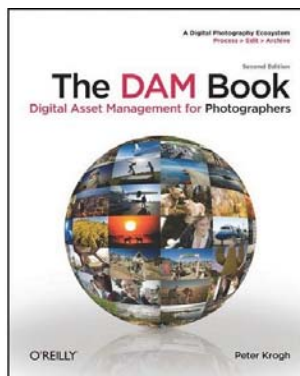
Photographic origins of the WF project

2004



Workflow, not slow. --Bruce Fraser

2005 (1st Edition)



The name wasn't a coincidence.

Workflow \ 10

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.
 - o Henry Bessemer's patent for making steel (1855)
4. Tacit knowledge is transferred through personal contact.
 - o WF is a craft and crafts are learned “at the bench”.
 - o Personal computers impede this transfer of the craft.

Workflow \ 11

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

Workflow \ 12

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?

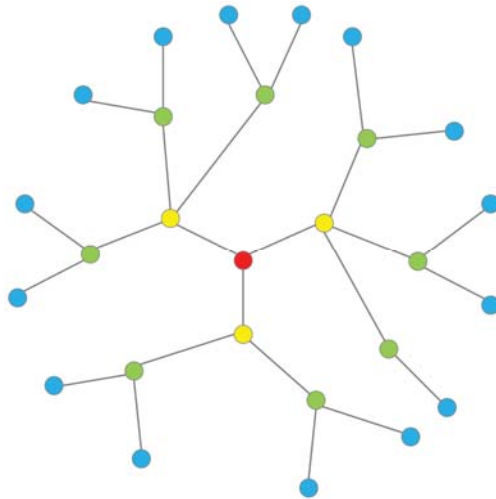
Workflow \ 13

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...

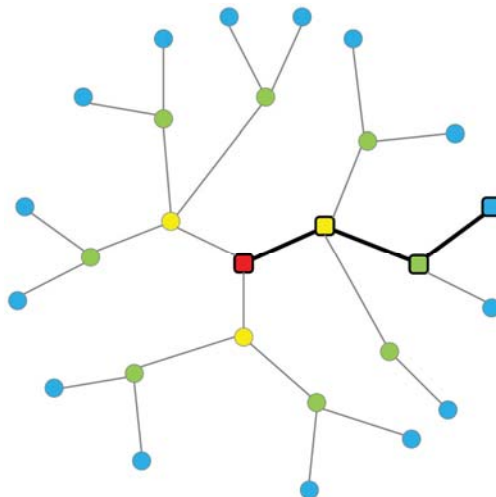
Workflow \ 14

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



Workflow \ 15

Decisions in the path to analysis: **the choices made**



Workflow \ 16

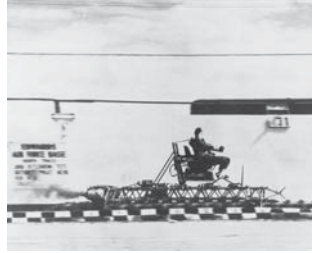
Why is replication so hard (continued)?

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.

Workflow \ 17

The foundation of WF is **ironical optimism**

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



Workflow \ 18

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.

Workflow \ 19

Step 3. Presenting results

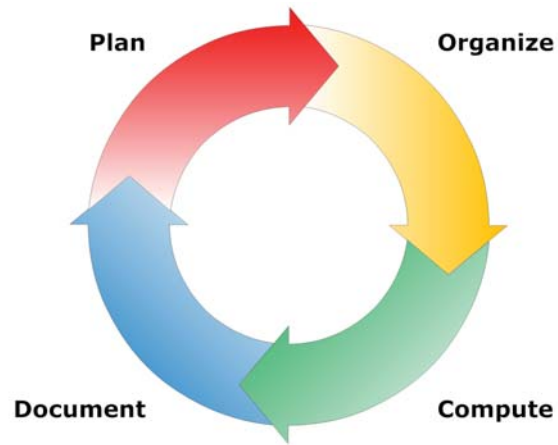
- Incorporating output into your presentation.
- Making a clear presentation.
- Maintaining the **provenance** of results.

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.

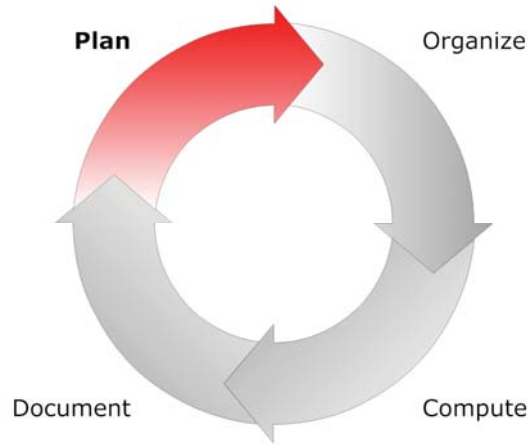
Workflow \ 20

Tasks within each step



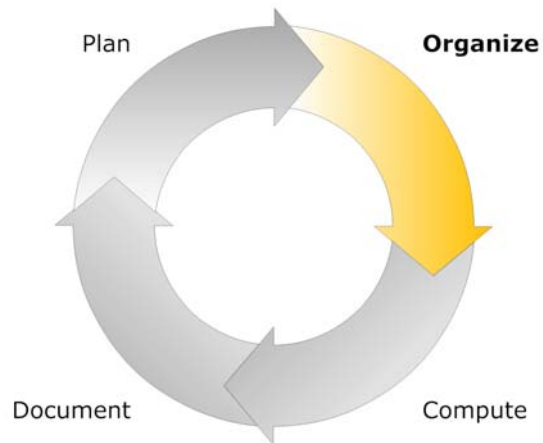
Workflow \ 21

Tasks within each step



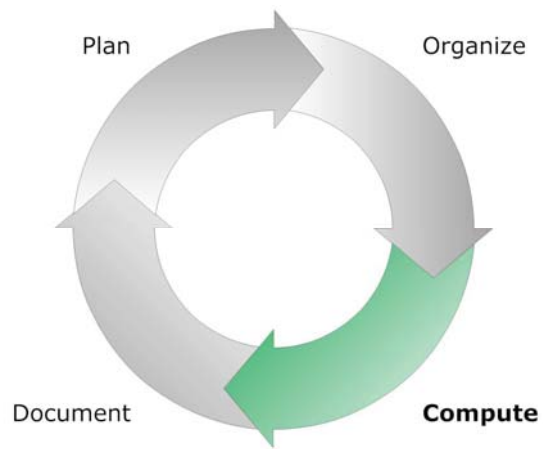
Workflow \ 22

Tasks within each step



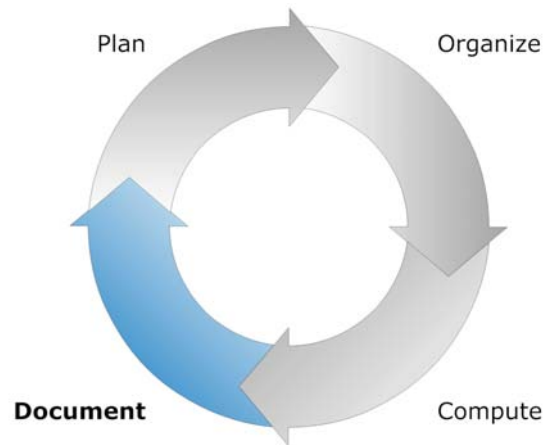
Workflow \ 23

Tasks within each step



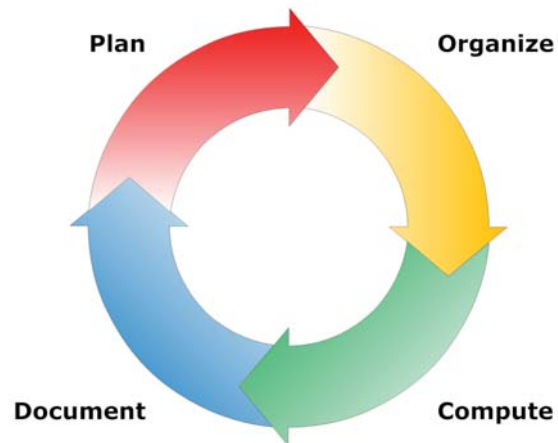
Workflow \ 24

Tasks within each step



Workflow \ 25

Tasks within each step



Workflow \ 26

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. None of the later books written with full access to the data was nearly as good.

Workflow \ 27

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

Workflow \ 28

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.

Workflow \ 29

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."

Workflow \ 30

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.

Workflow \ 31

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
```

Workflow \ 32

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\%cdatex 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"
```

Workflow \ 33

Organizing: uniform formats for do-files

```
capture log close
log using wftalk-example, replace text

// program:   wftalk-example.do
// task:
// project:
// author:    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
```

Workflow \ 34

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.

Workflow \ 35

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)...

Workflow \ 36

```
First complete set of analysis for FLIM measures paper
f2alt01a.do - 24May2002
Descriptive information on all rha, lha, and flim measures
f2alt01b.do - 25May2002
Compute bic' for each of four outcomes and all flim measures.
** Outcome: Can Work          global lhs "gcanwrk95"
** Outcome: Work in three categories  global lhs "dhitwh95"
** Outcome: bath trouble      global lhs "bathhd95"
** Outcome: adlum95 - sum of adla  global lhs "adlum95"
f2alt01c.do - 25May2002
Compute bic' for each of four outcomes and with only these restricted
flim measures.
* 1. ln(x+.5) and ln(x+1)
* 2. 9 counts: >=0<= >=7<= (598 and 751)
* 3. 8 counts: >=1<= >=6<= (598 and 751)
* 4. 18 counts: >=0<= >=1<= (598 and 751)
* 5. probability splits at .5; these don't work well in prior tests
f2alt01d.do - 25May2002
bic' for all four outcomes in models that include all raw flim measures
(fla*p5; fl*p5); pairs of u/l measures; groups of LCA measures
f2alt01e.do - all LCA probabilities - 25May2002
:::
f2alt01j.do - use three probability measures from LCA - 29May2002
:::
f2alt02c.do - 29May2002
use three binary variables, not just LC class numbers.
: dummies work better than the class number;
: effects of lower and severe are not significantly different.
Redo f2 analyses - error in adlum - 3Jun2002
ARGH! adlum is incorrect -- it included going to bed twice.
All of the f2alt analyses need to be redone using the corrected dataset.
f3alt_qflim07.do: create qflim07.dta 3Jun2002
1) Correct adlum: adlum95b
2) Add binary indicators of lmaxp5: lmaxnonep5, etc.
f3alt01a (redo f2alt01a.do) - 3Jun2002
f3alt01b.do (redo f2 job) - 3Jun2002
```

Workflow \ 37

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*

Workflow \ 38

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...

Workflow \ 39

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.

Workflow \ 40

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).
- Mean time to degree 7.6 years.

Workflow \ 41

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.

Workflow \ 42

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 Dictim**: 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.

Workflow \ 43

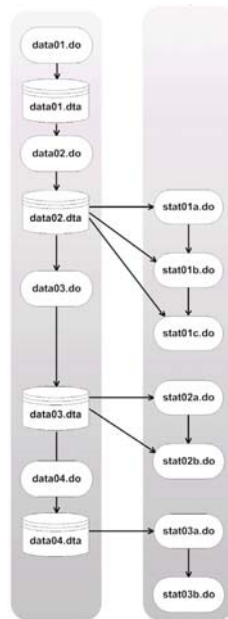
Dual workflow and run order

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.

Workflow \ 44

Dual workflow

Data management ==>



<== Data analysis

Workflow \ 45

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

Workflow \ 46

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.

Workflow \ 47

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**

Workflow \ 48

Legible log files (in text not smcl)

```
+-----+
| Key   |
+-----+
|      |
| frequency |
| row percentage |
+-----+
```

Occupation		Total		Years of education			
11	12	13	6	7	8	9	10

Menial		0	2	0	0	3	1
3	12	2	31				
9.68	38.71	6.45	100.00	0.00	0.00	9.68	3.23

BlueCol		1	3	1	7	4	6
5	26	7	69				
7.25	37.68	10.14	100.00	1.45	10.14	5.80	8.70

Craft		0	3	2	3	2	2
7	39	7	84				
		0.00	3.57	2.38	3.57	2.38	2.38

Workflow \ 49

8.33	46.43	8.33	100.00				

WhiteCol		0	0	0	1	0	1
2	19	4	41				
4.88	46.34	9.76	100.00	0.00	2.44	0.00	2.44

Prof		0	0	1	1	0	0
2	13	10	112				
1.79	11.61	8.93	100.00	0.89	0.89	0.00	0.00

Total		1	8	4	12	9	10
19	109	30	337				
5.64	32.34	8.90	100.00	1.19	3.56	2.67	2.97

Occupation		Total		Years of education			
11	12	13	6	7	8	9	10

Workflow \ 50

Menial	0	2	0	0	3	1
3	12	2	31			
	0.00	6.45	0.00	0.00	9.68	3.23
9.68	38.71	6.45	100.00			

BlueCol	1	3	1	7	4	6
5	26	7	69			
	1.45	4.35	1.45	10.14	5.80	8.70
7.25	37.68	10.14	100.00			

Craft	0	3	2	3	2	2
7	39	7	84			
	0.00	3.57	2.38	3.57	2.38	2.38
8.33	46.43	8.33	100.00			

Workflow \ 51

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

```

Viewer (#2) [help me]
-----
help for me :: Scott Long \ 2007-07-28

Reset everything: clear all

Updates:
  ado dir          : list installed packages
  update all       : update ado-files and executable
  adoupdate, update : update user written packages

Axes options:
  x/yscale(lo,hi)
  x/ylabel()
  x/ytic()
  x/yline()

Symbols:
  o large circle      S large square      T large triangle
  o small circle     d small diamond    p small plus
  x x                 i invisible       . dot

Mark missing values
  mark nomissv
  label var nomissv "1 if no missing"
  label def nomiss 1 Nonmissing 0 Missing
  label val nomissv nomiss
  markout nomissv `lhs' `rhs'
  replace nomissv = . if nomissv==0
  keep if nomissv==1

Scatterplot for two groups
  twoway (scatter y x if a==1, msymbol(circle_hollow) mcolor(red)) ///
  (scatter y x if a==0, msymbol(square_hollow) mcolor(blue)) ///
  , title(Compare two groups)

```

Workflow \ 52

Get MADD! An easy to use results collector.

In Stata, type:

```
findit madd
```

A work in progress. Caveat emptor.

Workflow \ 53

Data cleaning, including names and labels

Planning names

	A	B	C	D
1	Number	Name	Value label	Variable labels
2	1	id_iu		Respondent Number
3	2	cntry_iu	cntry_iu	IU Country Number
4	3	vignum	vignum	Vignette
5	4	serious	serious	Q1 How serious would you consider Xs situation to be?
6	5	opfam	Ldummy	Q2_1 What X should do:Talk to family
7	6	opfriend	Ldummy	Q2_2 What X should do:Talk to friends
8	7	tospi	Ldummy	Q2_7 What X should do:Go to spiritual or traditional healer
9	8	tonpm	Ldummy	Q2_8 What X should do:Take non-prescription medication
10	9	oppme	Ldummy	Q2_9 What X should do:Take prescription medication

Truncation and careless names

Example: `ownsex` and `ownsexu` caused weeks of confusion.

Workflow \ 54

Creating a codebook

		Not at all Important	1	2	3	4	5	6	8	9	10	Very Important
Q43. Turn to family for help			1	2	3	4	5	6	8	9	10	
tcfam	Q43 How Important: Turn to family for help											
Q44. Turn to friends for help			1	2	3	4	5	6	8	9	10	
tcfriend	Q44 How Important: Turn to friends for help											
Q45. Turn to a minister, priest, Rabbi or other religious leader			1	2	3	4	5	6	8	9	10	
tcrelig	Q45 How Important: Turn to a Minister, Priest, Rabbi or other religious leader											
Q46. Go to a general medical doctor for help			1	2	3	4	5	6	8	9	10	
tcdoc	Q46 How Important: Go to a general medical doctor for help											
Q47. Go to a psychiatrist for help			1	2	3	4	5	6	8	9	10	
tcpsy	Q47 How Important: Go to a psychiatrist for Help											
Q48. Go to a mental health professional for help			1	2	3	4	5	6	8	9	10	
tcmhprof	Q48 How Important: Go to a mental health professional											
ALLOWED DEFINITION – PSYCHOLOGIST, THERAPIST, SOCIAL WORKER, OR COUNSELOR												
INTERVIEWER NOTE: CODE "DON'T KNOW" AS #8 ABOVE SEQUENCE.												
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 situations like NAME.												

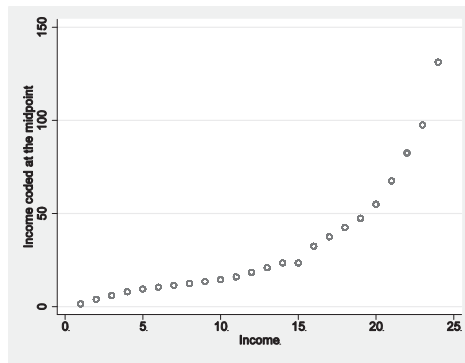
Workflow \ 55

Types of data cleaning



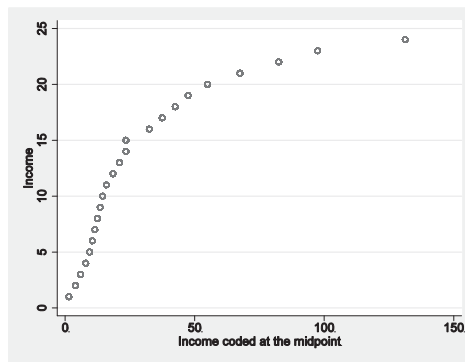
Workflow \ 56

Cleaning 1: finding an error with a graph



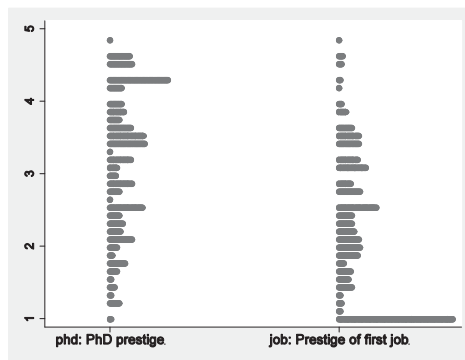
Workflow \ 57

Cleaning 1: reversing the graph



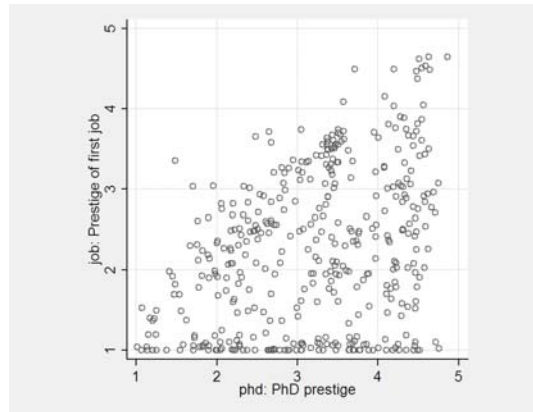
Workflow \ 58

Cleaning 2: remembering a coding decision



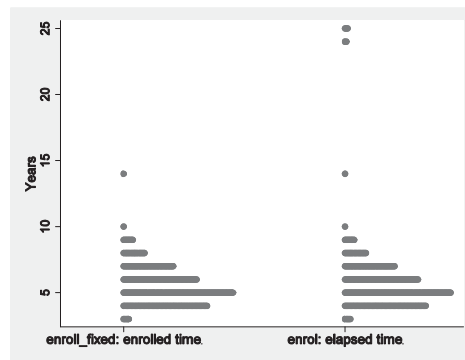
Workflow \ 59

Cleaning 3: understanding the substantive process



Workflow \ 60

Cleaning 4: avoiding expensive mistakes



Workflow \ 61

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.

Workflow \ 62

Documenting the provenance

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

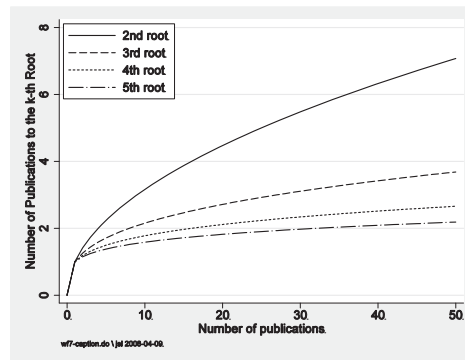
1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively ($z=2.55, p<.01$)). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

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

1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively ($z=2.55, p<.01$ [cwhrr-fig03c-hrmemp4.do #4 jsl 17May06])). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Workflow \ 66

Captions make it easy to trace a source

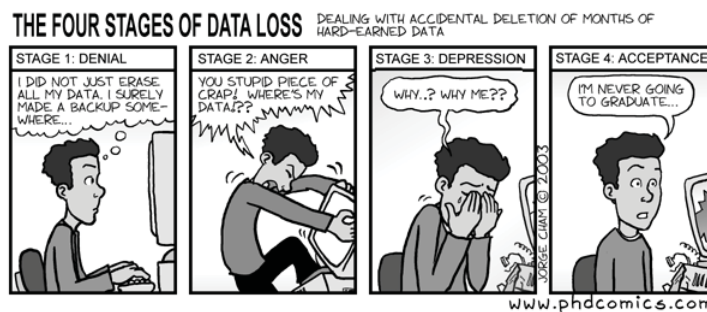


```
twoway (line art_root2 art_root3 art_root4 art_root5 articles, ///
        lwidth(medium), ytitle(Number of Publications to the k-th Root) ///
        yscale(range(0 8.)) legend(pos(11) rows(4) ring(0)) ///
        caption(wf7-caption.do \ jsl 2008-04-09, size(vsmall)))
```

Workflow \ 67

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.



Workflow \ 68

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, 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

Workflow \ 69



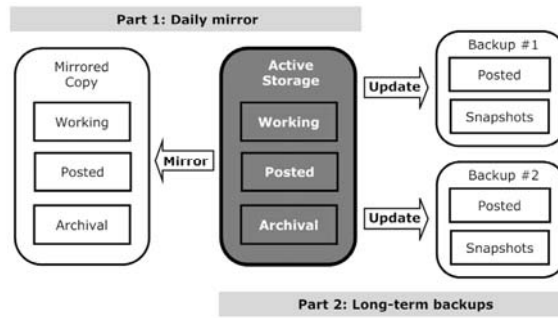
Workflow \ 70

Issues in deciding how and what to preserve

Questions to consider	Type of file		
	Working	Posted	Archival
How do you recover the file?	Redo work	Redo old work	Download file
What is the cost of recovery?	Minor	Potentially lots of work	A little time
How long are you preserving it?	1-3 yrs.	3-10 yrs.	Forever
How difficult is it to preserve?	Trivial	Tedious	Very hard
Concern with media/format?	Minor	Some	Critical

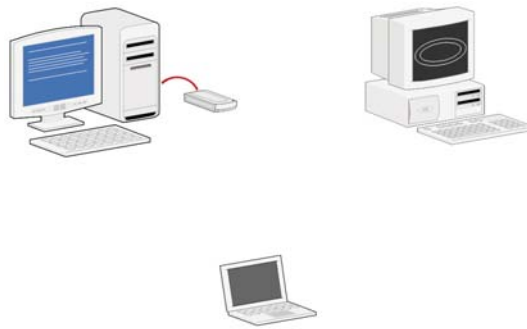
Workflow \ 71

A KISS approach to preserving files



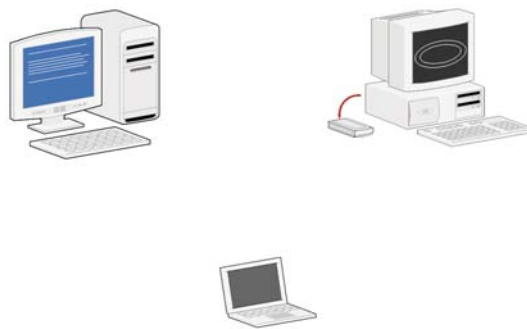
Workflow \ 72

Tactics: Portable drive computing at home



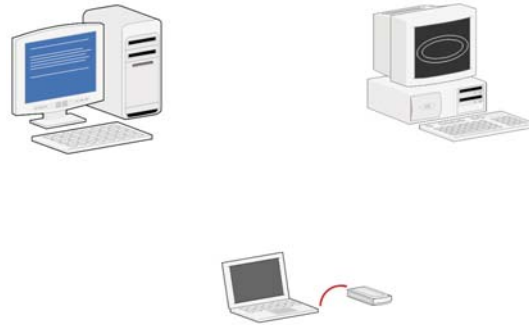
Workflow \ 73

Tactics: Portable drive computing at work



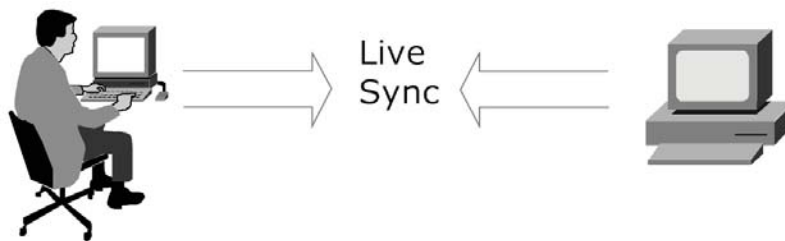
Workflow \ 74

Tactics: Portable drive computing when traveling



Workflow \ 75

Tactics: Live sync



Workflow \ 76

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...



Workflow \ 77

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.

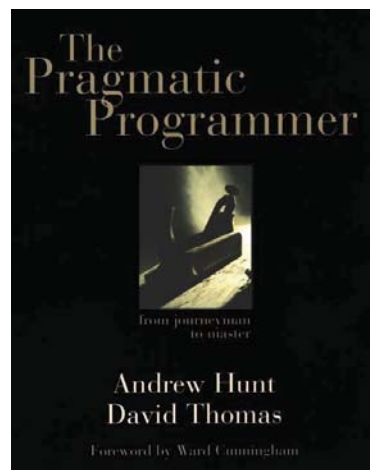
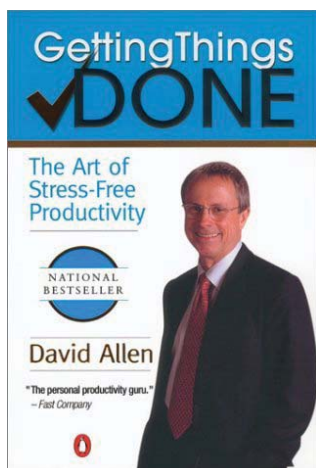
Workflow \ 78

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?

Workflow \ 79

Other aspects of workflow



Workflow \ 80

Thanks for listening.
Questions?

Provenance: 2009-07-12; 2009-07-23; 2009-07-31; 2009-11-12; 2010-07-19; 2010-07-29

Workflow \ 81