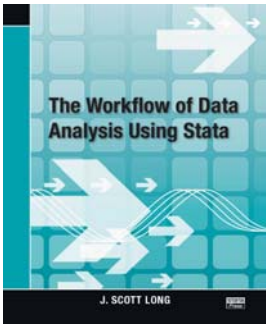
Indiana University

Principles of Workflow in Data Analysis

Scott Long



November 2010

What is "workflow"?

1. A coordinated framework for conducting data analysis
2. WF involves coordinated procedures for:
   - Planning, organizing and documenting research
   - Cleaning data
   - Analyzing data
   - Presenting results
   - Backing up and archiving materials

You already have a workflow (WF)

1. Your WF might be:
   A. **Planned** and carefully orchestrated.
   B. **Ad hoc**, piece-meal, developed in reaction to mistakes.
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 you time and make you a better data analyst.

Why should you care about workflow?

1. **Replication**
   - Replication is essential for good science.
   - An effective workflow is essential for replication.
2. **Getting the right answers**
   - Retractions are embarrassing and can end careers.
3. **Time**
   - "Science is a voracious institution."
   - An effective workflow makes you more efficient.
4. **Errors are inevitable; an effective workflow helps you find and fix them.**

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*

Origins of the workflow project

1. **Easy things**: consulting on easy things, instead of hard things.
2. **Incorrect results** with clever "explanations".
3. **A dissertation delayed** 18 months to determine 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. **Analyzing the wrong variable** while writing an NAS report.
7. **Miscoded genes** that delayed progress in a study of alcholism.
8. **Collaborations** that multiply the ways things can go wrong.
9. **Misleading or ambiguous output** such as...

Example 1: definitely a problem in a $3M study

```
. tabulate female sdchild_v1

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


Example 2: which number is which?

```
. tab occ ed, row

           |                                Years of education
Occupation |         3         6         7         8         9        10        11
12         13 |     Total
-----------+-----------------------------------------------------------------------------------------------
-----------------------+----------
    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
-----------+-----------------------------------------------------------------------------------------------
-----------------------+----------
  WhiteCol |         0         0         1         0         1         2
19           4 |        41
           |      0.00      0.00      0.00      2.44      0.00      2.44      4.88
46.34       9.76 |    100.00
-----------+-----------------------------------------------------------------------------------------------
-----------------------+----------
```


Example 3: good software doing things badly

```
. 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

------------------------------------------------------------------------------
      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)
------------------------------------------------------------------------------
```

Did StataCorp read the WF book?

Why learning WF is difficult

1. **Tacit knowledge** 2. **Heavy lifting** 3. **Time to practice**

What is tacit knowledge?

1. **Explicit knowledge** is the stuff of textbooks and articles.

2. **Tacit knowledge** is implicit and undocumented (Michael Polanyi).
   A. People are unaware of their essential tacit knowledge.
      - Henry Bessemer's patent for making steel didn't work (1855)
   B. Tacit knowledge is transferred "**at the bench**".
      - Personal computers impede the transfer of tacit knowledge.

Undifferentiated heavy lifting

Data analysis includes a lot of heavy lifting

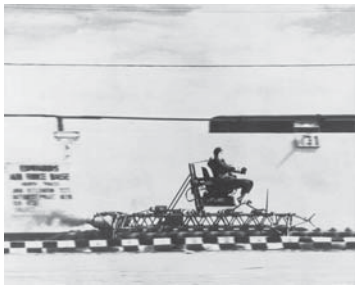
"The reality, of course, today is that if you come up with a great idea you don't get to go quickly to a successful product. **There's a lot of undifferentiated heavy lifting that stands between your idea and that success**." - Jeff Bezos, amazon.com

The Workflow of Data Analysis Using Stata

1. 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.
4. The book focuses on tools in Stata, but the principles apply broadly.

The foundation of WF is **ironical optimism**

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



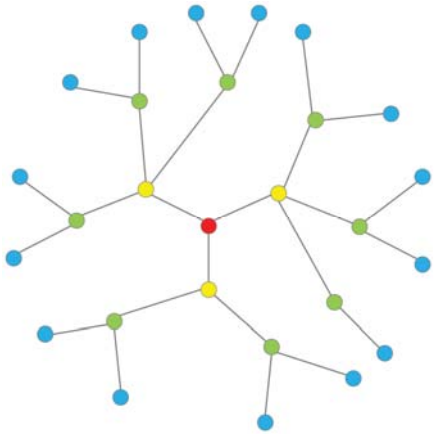
WF starts with **replication**

1. An effective WF facilitates replication.
2. You must plan for replication at the start of a project.
3. Disciplines are increasingly concerned with replicability.
   - Articles in Political Science, Economics, Sociology and other fields.
4. Ask yourself:
   - Are your do-files and log files ready for public display?
   - Will they produce *exactly* the same results as you have published?

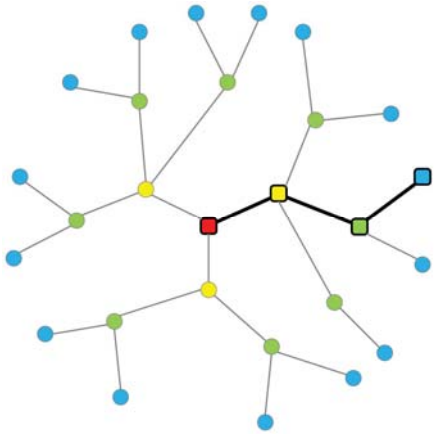
Why replication is so hard

1. **The curse of dimensionality**: 10 minor decisions, leads to 1,024 reasonable ways to create your data.
   - Where to truncate a variable.
   - The seed for the RN generator.
   - Creating a scale with partial missing data.
   - Which cases to keep for analysis.
   - How to code education?
   - What values to assign 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 what you did.
3. **Changing software**: 2 weeks of sleepless nights due to version variation. This is particularly difficult when there is an active user community.
4. **Lost files**: corrupted, lost, unreadable, obsolete, or ambiguous files.

Criteria for choosing a WF assuming replicability

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 and inconsistently decide how to do things.
- Standardization makes it easier to find mistakes.

4. Automation

- Automated procedures prevent mistakes and are faster.
- **Drukker's Dictim**: Never type anything that you can obtain from a saved result. (Did the authors of `margins` think about this?)

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.

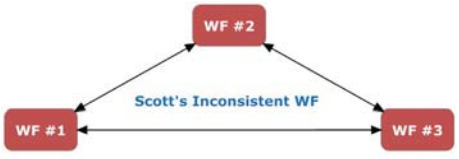
7. Scalability

- Different projects require different workflows.

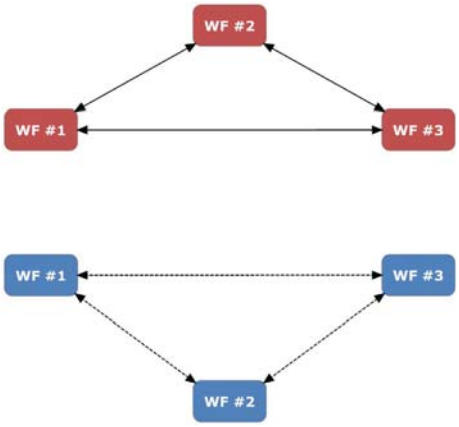
Collaboration and workflow

- Collaboration makes it more difficult to have an effective, efficient and replicable workflow.
- Why? And, why can't they do it just like me?
- Every problem you can have working by yourself is multipled.

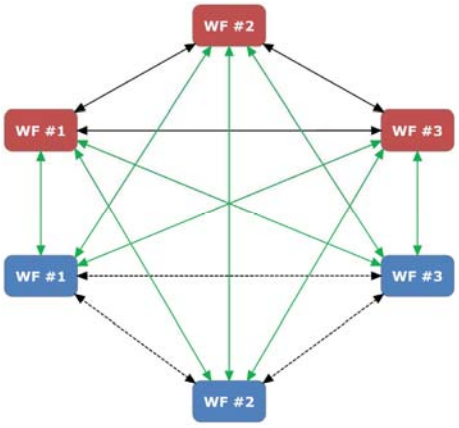
Coordinating multiple workflows



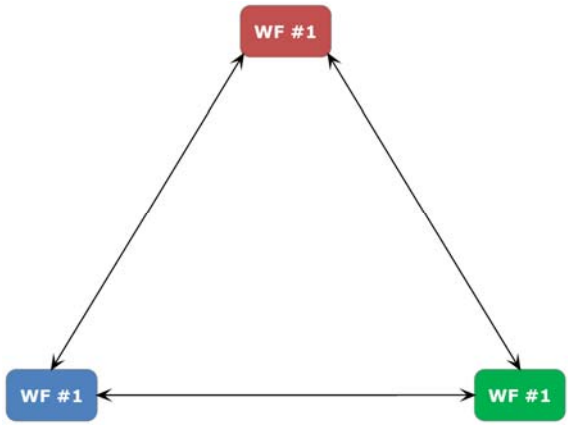
Coordinating multiple workflows



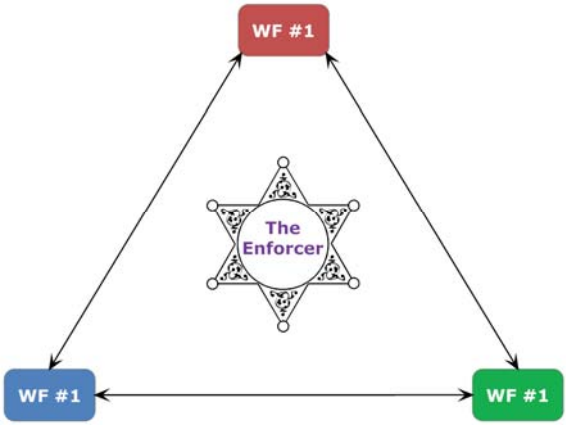
Coordinating multiple workflows



Coordinating multiple workflows



Coordinating multiple workflows



Key factors in collaborations

1. Agreed upon standards
2. Explicit coordination
3. Enforcement of standards
4. A sense of humor

Steps in your workflow

Step 0. Have a good idea for a project

Step 1. Prepare the data for analysis

- Data must be accurate.
- Variables must be carefully named and labeled.
- This takes 90% of the time, unless you hurry.

Step 2. Conduct analyses

- Estimate models and create graphs.
- Often the simplest part of the workflow.

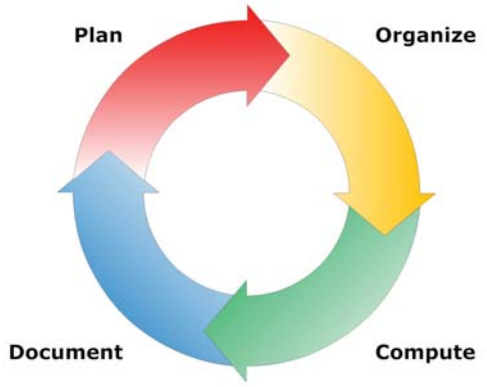
Step 3. Present results

- Incorporate output into your presentation.
- Maintain the **provenance** of results.
- Make effective presentations.

Step 4. Protecting files

- **Backing up** and **archiving**: preserving the bits and the content.
  - $2,000 to get 1 variable from an “archived” file.
- Replication is impossible without your data and do-files.
- "Today's noise is tomorrow's knowledge." -- *David Clemmer*

Tasks within each step



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 those analyses.
- None of the later books written with full access to the data were as good.

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. A little planning goes a long way and almost always saves time.
3. Planning includes:
   - General goals, publishing plans, and firm deadlines.
   - Division of labor and accountability.
   - Proposal for data construction: names, labels, formats.
   - Procedures for handling missing data.
   - Anticipated analyses.
   - Guidelines and responsibility for documentation.
   - Procedures and schedule for backing-up and archiving materials.

Organizing

1. Organization is movtivated by the need to:
   - **Find things**
   - **Avoid duplication**
2. It requires explicit, consistent decisions about naming and storing things.
3. Organization:
   - Helps you work faster
   - Rewards consistency and uniformity
   - Organization is contagious

Signs of poor organization

1. You can't find a file and think you deleted it.
2. You find multiple versions of a file and don't know which is which.
3. You and a colleague are 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 (or 16) files with "final" in the name.
   - final_report_v16.docx
   - NSF_science_report 2010-10-21.docx

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
```

For example, 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\%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:
//  author:     jsl \ 2010-07-27

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
```

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'

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. Documentation occurs on many 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 tomorrow or 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...

First complete set of analysis for FLIM measures paper

f2alt01a.do - 24May2002

Descriptive information on all rhs, lhs, and flim measures

f2alt01b.do - 25May2002

Compute bic' for each of four outcomes and all flim measures.

```
**  Outcome: Can Work                   global lhs "qcanwrk95"
**  Outcome: Work in three categories   global lhs "dhlthwk95"
**  Outcome: bath trouble               global lhs "bathdif95"
**  Outcome: adlsum95 - sum of adls     global lhs "adlsum95"
```

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: >=5=5  >=7=7   (50% and 75%)
*     3.   8 counts: >=4=4  >=6=6   (50% and 75%)
*     4.  18 counts: >=9=9 >=14=14  (50% and 75%)
*     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; fll*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 adlsum - 3Jun2002

ARGH! adlsum 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 aldsum: adlsum95b
2) Add binary indicators of Lmaxp5: LmaxNonep5, etc.

f3alt01a (redo f2alt01a.do) - 3Jun2002

f3alt01b.do (redo f2 job) - 3Jun2002

Execution and computing

1. Execution involves carrying out tasks within each step.
2. Effective execution requires **the right tools**.
   - Software
     - *a. Text editor*
     - *b. File manager*
     - *c. Statistical software*
     - *d. Macro program (even if only to insert time stamps)*
     - *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).**
- **Mean time to degree 7.6 years.**

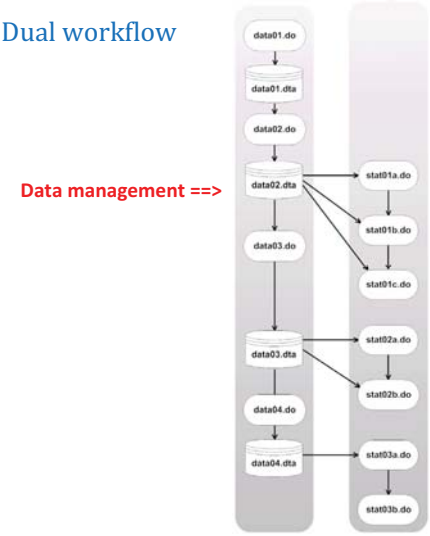
A thought experiment on planning and computing

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

Principles for a computing workflow

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** for sharing results (defined later)

Dual workflow



Run order and a dual workflow

Data management	Data analysis
`data01.do`	`stat01a.do`
`data02V2.do`	`stat01b.do`
`data03.do`	`stat01cV2.do`
`data03-1.do`	
`data03-2.do`	`stat02a.do`
`data04.do`	`stat02a1.do`
	`stat02b.do`
	`stat03aV2.do`
	`stat03b.do`
	`stat03c.do`
	`stat03c1.do`
	`stat03c2V2.do`
	`stat03d.do`

The essential **posting principle**

The **posting principle** is defined by two rules:

1. **The share rule**: Only share results after the files are posted.
2. **The no change rule**: Once a file is posted, ***never*** change it.

Data analysis: use do-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: It should run on another computer at a later date without changes.

Legible do-files: output that is easy to read

1. Lots of 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)

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

             |                                    Years of education
  Occupation |         3          6          7          8          9         10
          11 |        12         13 |      Total
-------------+----------------------------------------------------------------------------
-------------------------+----------
      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
```


Automation

1. Much of data analysis involves repetitive tasks.
2. Repetition invites errors.
3. 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,…

`help me`

SNAG: An easy to use results collector

In Stata, type:

```
findit snag
```

snag collects dozens or hundreds of results to make them easier to digest.

- The standard output is used to verify the results.
- The "snagged" summary lets you discover what you want.
- Anyone using `margins` knows why this is necessary.

Data cleaning, including names and labels

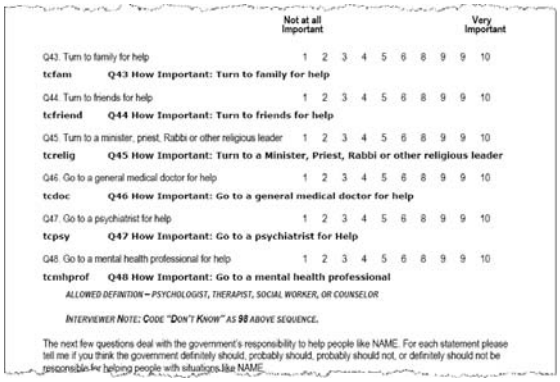
Planning names

	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

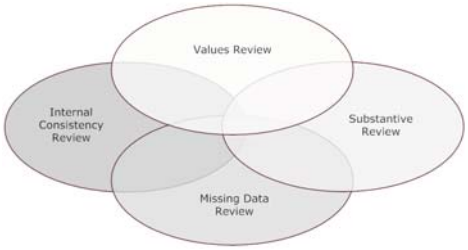
Truncation and careless names

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

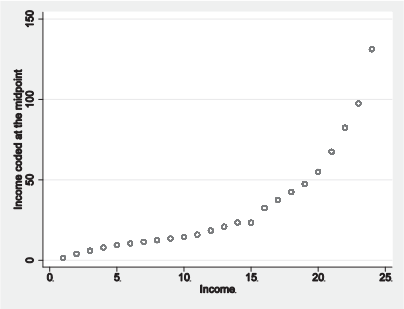
Creating a codebook



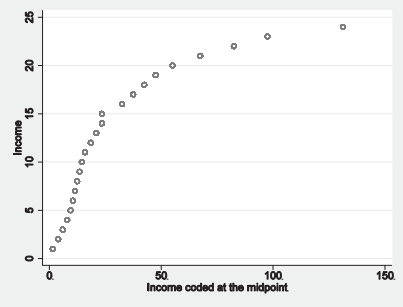
Types of data cleaning



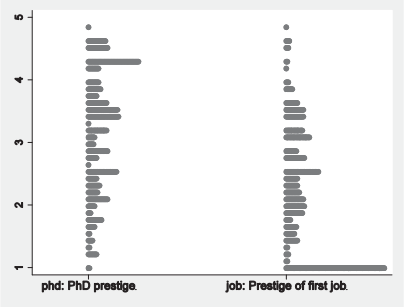
Cleaning 1a: finding an error with a graph



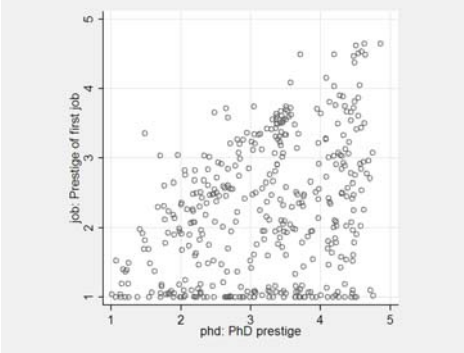
Cleaning 1b: reversing the graph



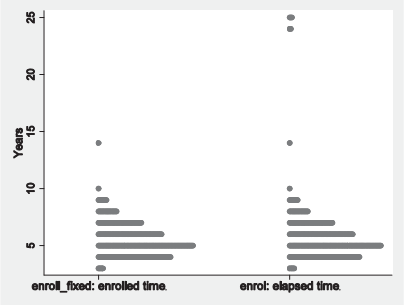
Cleaning 2: remembering a coding decision



Cleaning 3: understanding the substantive process



Cleaning 4: avoiding expensive mistakes



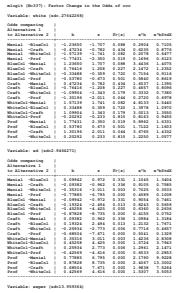
Analyzing the data

1. Take lots of classes in statistics.
2. Find exemplars; don't rediscover the wheel; don't do it "your way".

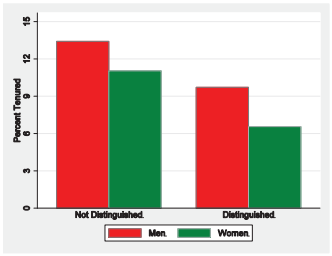
Presentations and provenance

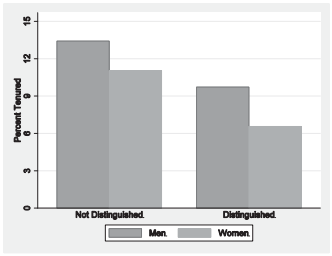
1. **Content and methods** are substantive, disciplinary decisions.
2. **Presentations and preservation of provenance** are universal.

Tables too small

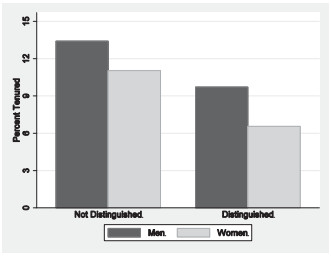


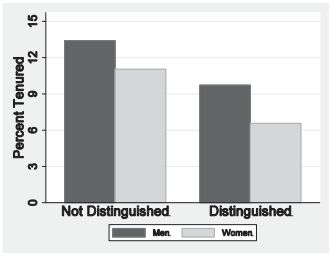
Colors that aren't distinct when printed/projected





Labels that aren't large enough





Documenting the provenance

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

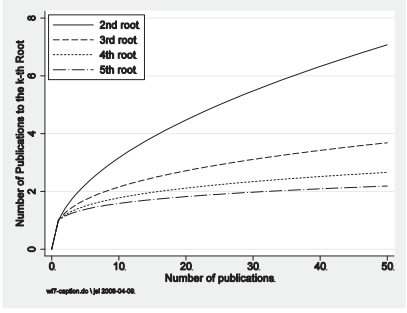
1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively (z=2.55, p<.01)) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

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

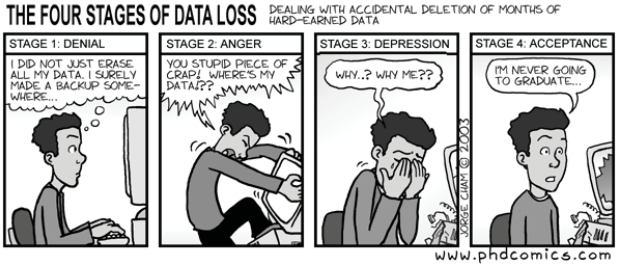
Captions that indicate the provenance



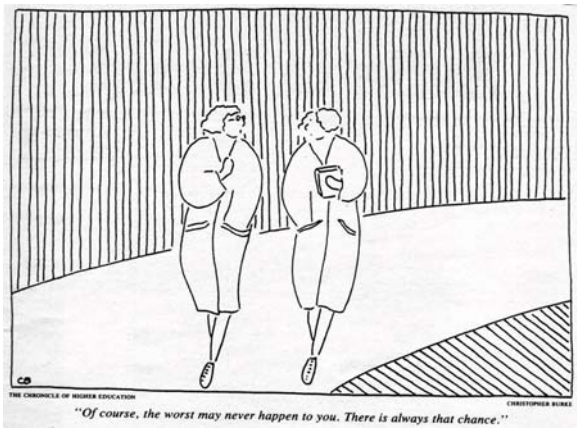

```
twoway (line art_root2 art_root3 art_root4 art_root5 articles,       ///
    lwidth(medium)), ytitle(Number of Publications to the k-th Root) ///
    yscale(range(0 8.)) legend(pos(11) rows(4) ring(0))              ///
    caption(wf7-caption.do \ jsl 2008-04-09, size(vsmall))
```

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.

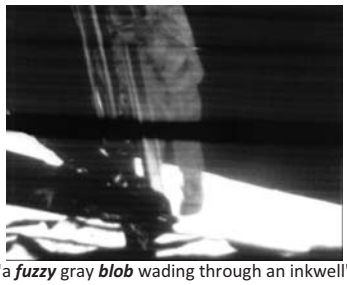


Hope, foolishly, springs eternal (the Sweden syndrome)



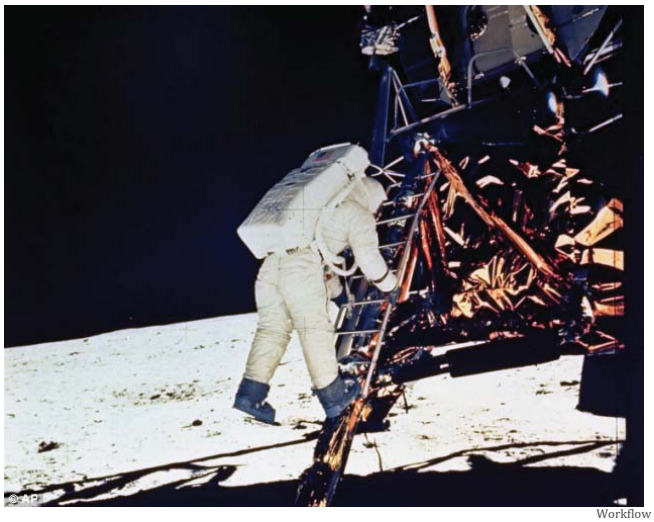
Examples of data loss

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

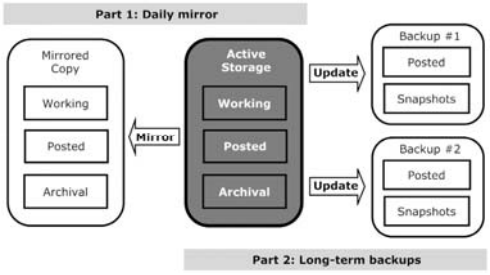


"a ***fuzzy*** gray ***blob*** wading through an inkwell"

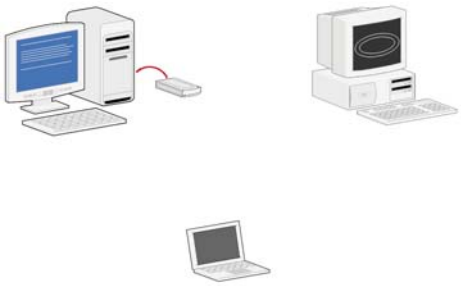
Dark Side of the Moon



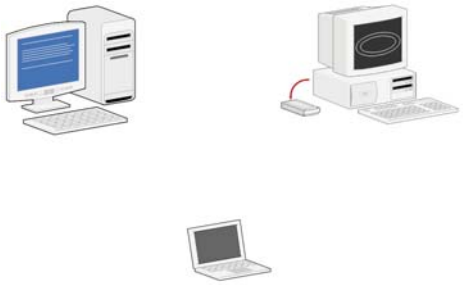
A simple approach to preserving files



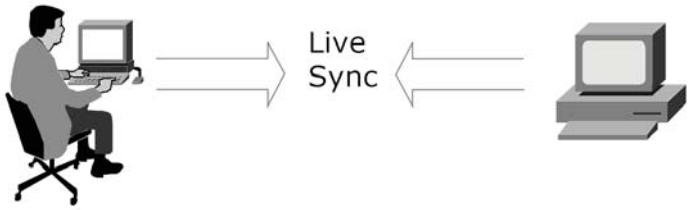
Tactics: Portable drives at home



Tactics: Portable drive at work



Tactics: Live sync (soon to be Live Mesh)



Dropbox

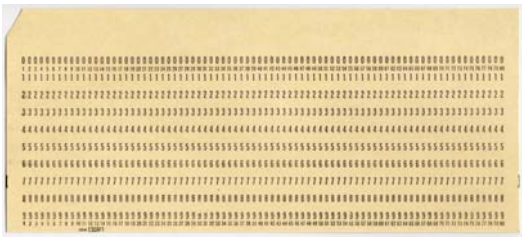
1. Install the progam
2. Drop files into the folder
3. Retrieve them from any machine with Dropbox
4. Have shared folders for collaboration
5. Avoid sending attachments even for one time file exchanges

Off-line backups

Dropbox and similar services, enterprise 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 a 30M cubic foot building; 60M cubic feet next month. With compression…



Changing your workflow

1. Slowly, systematically, throughtfully.
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?