Indiana University

Reproducible Results and the Workflow of Data Analysis

Scott Long

Departments of Sociology and Statistics

www.indiana.edu/~jslsoc/ftp/

The reproducible results movement

- Transparancy in science
- Reproducible results
- Teaching integrity in research
- Open science

Expectation for open access to data & analysis

- Some journals requiring reproducible analyses *before* acceptance.
- Funding agencies are increaing requirements for data access.
- Haverford College requires reproducibilty for thesis.

With access comes accountability

- retractionwatch.com
- For example…

Retracting an article

RETRACTED: In Sickness and in Health? Physical Illness as a Risk Factor for Marital Dissolution in Later Life

[redacted]

Abstract

The health consequences of marital dissolution are well known, but little work has assessed the impact of health on the risk of marital dissolution. In this study we use a sample of 2,701 marriages from the Health and Retirement Study (1992–2010) to examine the role of illness onset (i.e., cancer, heart problems, lung disease, and/or stroke) in subsequent marital dissolution due to either divorce or widowhood. We use a series of discrete-time event history models with competing risks to estimate the impact of husband's and wife's physical illness onset on risk of divorce and widowhood. We find that only wife's illness onset is associated with elevated risk of divorce, while either husband's or wife's illness onset is associated with elevated risk of widowhood. These findings suggest the importance of health as a determinant of marital dissolution in later life [illegible] and gendered social pathways.

Keywords

aging, chronic disease, gender, [illegible] mental health

RETRACTED

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are the married healthier than the unmarried (e.g., Lillard and Waite 1995; Corbeanu 1992), but studies find that both divorce and widowhood are predictive of poorer physical and mental health (e.g., Hughes and White 2009; Williams and Umberson 2004). Less attention, however, has been paid to how health may be a determinant of marital status. Work within this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1989; Smith and Smith 2010), but poor health may be an equally important force for selection (e.g., Booth, and Johnson 2008). Illness may initiate changes to spouses' roles—in particular, increasing caregiving responsibilities for the healthy spouse—which can tax marital relationship dynamics (Wolff and Kasper 2006). Illness may also decrease household income due to the inability of one or both spouses to work (Teachman 2010), which may increase marital strain.

Only a few studies have examined the role of poor health in subsequent divorce, and these studies are mixed in their findings, with some finding

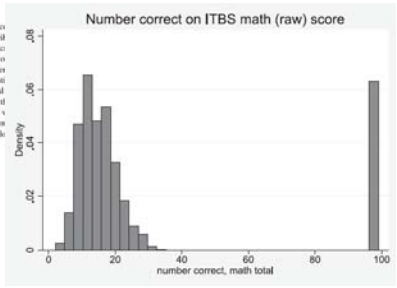
Finding errors in earlier analysis

Distributional Analysis in Educational Evaluation: A Case Study from the New York City Voucher Program

Marianne Bitler, Thurston Domina, and Emily Penner
University of California, Irvine, Irvine, California, USA

Hilary Hoynes
University of California, Berkeley, Berkeley, California, USA

Abstract: We use quantile treatment effects estimation to examine the consequenc assignment New York City School Choice Scholarship Program across the distril achievement. Our analyses suggest that the program had negligible and statistic effects across the skill distribution. In addition to contributing to the literature o the article illustrates several ways in which distributional effects estimation can e research: First, we demonstrate that moving beyond a focus on mean effects est possible to generate and test new hypotheses about the heterogeneity of educational that speak to the justification for many interventions. Second, we demonstrate t effects can uncover issues even with well-studied data sets by forcing analysts to new ways. Finally, such estimates highlight where in the overall national achievemen scores of children exposed to particular interventions lie; this is important for explo validity of the intervention's effects.



Assessing fragility of results

Measurement, methods, and divergent patterns: Reassessing the effects of same-sex parents ☆

Simon Cheng [a,1], Brian Powell [b,1]

[a] 344 Mansfield Rd., Department of Sociology, University of Connecticut, Storrs, CT 06269, United States
[b] 744 Ballantine Hall, 1020 E. Kirkwood Ave., Department of Sociology, Indiana University, Bloomington, IN 47405-7103, United States

ARTICLE INFO

Article history:
Received 8 October 2013
Revised 24 March 2015
Accepted 8 April 2015
Available online 23 April 2015

Keywords:
Children
Family structure
Methodology
Same-sex parenting
Sexuality

ABSTRACT

Scholars have noted that survey analysis of small subsamples—for example. same-sex parent families—is sensitive to researchers' analytical decisions, and even small differences in coding can profoundly shape empirical patterns. As an illustration, we reassess the findings of a recent article by Regnerus regarding the implications of being raised by gay and lesbian parents. Taking a close look at the New Family Structures Study (NFSS), we demonstrate the potential for misclassifying a non-negligible number of respondents as having been raised by parents who had a same-sex romantic relationship. We assess the implications of these possible misclassifications, along with other methodological considerations, by reanalyzing the NFSS in seven steps. The reanalysis offers evidence that the empirical patterns showcased in the original [redacted] article are fragile—so fragile that they appear largely a function of these possible misclassifications and other methodological choices. Our replication and reanalysis of [redacted] study offer a cautionary illustration of the importance of double checking and critically assessing the implications of measurement and other methodological decisions in our and others' research.

How bad is it?

Science Isn't Broken, By Christie Aschwanden

<with my edits to shorten>

Peer review? Groups circumventing peer review at prestigious journals.

Scientific journals? Two engineering journals published Maggie Simpson & Edna Krabappel's "Fuzzy, Homogeneous Configurations."

Revolutionary findings? *Science* retracted a paper when co-author could find no record of the data.

I've learned that the headline-grabbing cases of misconduct and fraud are mere distractions. The state of our science is strong, but it's plagued by a universal problem: Science is hard – really f*ing hard.

If we're going to rely on science as a means for reaching the truth - and it's still the best tool we have - it's important that we understand and respect just how difficult it is to get a rigorous result.

Replication and reproducible results

- Expectations for replication and reproduction are growing.
- This is a very positive development, but it "raises the bar".

Replication of results

- Confirming the results of a published finding using new data

Challenges to replication

- Data mining portrayed as theory testing
- Post analysis hypothesis construction
- "Cherry picking" the sample
- Undocumented specification searches

Reproduction of results

- Obtaining identical results with the same data

Challenges to reproduction

- Poor methodology for managing data analysis

Today's talk is about obtaining reproducible results

- The workflow of data analyis

What is the workflow of data analysis?

Workflow is a coordinated framework for data analysis that deals with **all** aspects of data analysis:

- Planning, organizing and documenting research
- Cleaning data
- Analyzing data
- Presenting results
- Backing up and archiving materials
- Reproducing results

Why is coordinated important?

You already have a workflow

1. Your WF might be:
   - A. **Planned**
   - B. **Ad hoc**
   - C. **Planned in an ad hoc way**
2. You can improve your WF with a modest investment of time.
   - A. The less experience you have, the easier it is.
     - Undergraduates often find it easier than faculty!
   - B. It takes time, but will save you more time.
   - C. It will make you a better data analyst.

Why workflow is essential

Reproducibility and workflow

1. Results that others can reproduce are essential for good science.
2. Workflow is critical for reproducibility.

Getting the right answer; finding and fixing inevitable errors

1. You want your analysis to be correct.
2. With demands for public data and scripts files public, retractions are increasing.

Efficiency and saving time

Science is a voracious institution. -- Harriet Zuckerman

Origins of the workflow project

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

Example 1: Definitel in a $3M study

```
. tabulate female sdchild_v1

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


Example 2: variable labels that come from the survey center

```
. codebook tc1*, compact

Variable    Obs Unique     Mean Min  Max  Label
-------------------------------------------------------------------------------
tc1doc     1074     10 8.714153   1   10  Q46 How important is it to go to ...
tc1fam     1074     10 8.755121   1   10  Q43 How important is it to turn t...
tc1friend  1073     10 7.799627   1   10  Q44 How important is it to turn t...
tc1mhprof  1045     10  7.58756   1   10  Q48 How important is it to go to ...
tc1psy     1050     10 7.567619   1   10  Q47 How important is it to go to ...
tc1relig   1039     10  5.66025   1   10  Q45 How important is it to turn t...
-------------------------------------------------------------------------------
```


Example 3: 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          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 4: good software doing things badly (fixed in Stata 13)

```
. 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)
------------------------------------------------------------------------------
```


Why learning WF is difficult

Tacit knowledge

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

Data analysis involves heavy lifting

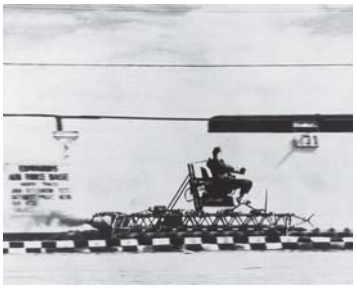
> **There's a lot of undifferentiated heavy lifting that stands between your idea and that success**. -- Jeff Bezos, amazon.com

The Workflow of Data Analysis Using Stata

1. The book makes tacit knowledge explicit.
2. It deals with lots of heavy lifting.
3. It provides specifics on the general issues discussed today.
4. While focusing on Stata, the principles apply broadly.
   - o An enthnographer uses if for her research team.
   - o An researcher in China found it crucial for getting his paper accepted by *Nature*.
   - o A manager of health statistics for a European country said it "improved the quality of my life, not my data analysis, my life."

The foundation of WF is **ironical optimism**

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



40G's: From 0 to 995mph and back in 3 seconds…

"I was fine, only blind for a few days."

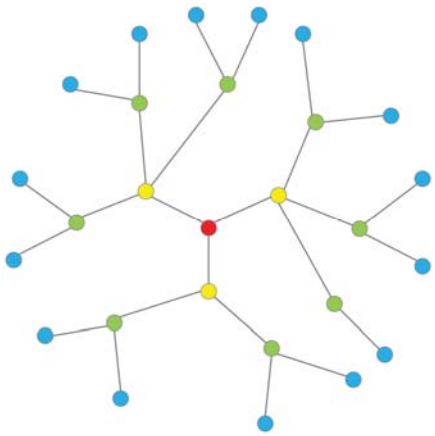
The overarching objective is reproducibility

1. **WF facilitates producing reproducible results.**
2. Ask yourself:
   - o Can you reproduce ***exactly*** the results you published?
   - o How long would it take?
3. Reproducible results requires planning from the start of a project.

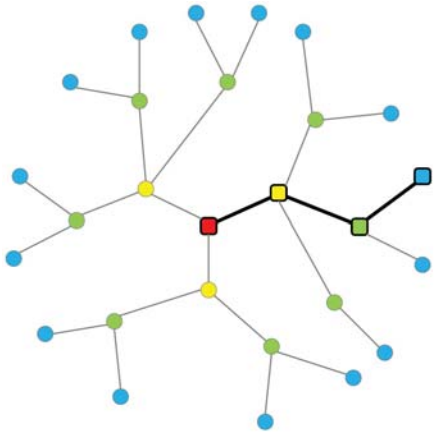
Why are results hard to reproduce?

1. **The curse of dimensionality**: 10 decisions, 1,024 possibilities
   - o Where to truncate a variable?
   - o What seed for the RN generator?
   - o How to scale with partially missing data?
   - o Which cases to keep for analysis?
   - o How to code education?
   - o What values to assign to income greater than $200,000?
   - o 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 are results hard to reproduce?

2. **Missing documentation**: Replication should involve retrieving documentation, not trying to remember.
3. **Changing software**: New software can give different results.
   - o A colleague spent painful weeks failing to replicate results because he forgot "version 7" in his do-file.
4. **Lost files**: corrupted, lost, unreadable, obsolete, or ambiguous files
   - o Do you have $2,000 to retrieve the file that was "backed up"?
   - o Do virtual servers archive your data?

Given reproducibility, criteria for choosing WF

1. Accuracy

If your program is not correct, then nothing else matters.

--Oliveira and Stewart

2. Efficiency

- o Completing work quickly given accuracy and replicability.
- o Working quickly competes with working carefully.

3. Standardization

- o Avoiding repeatedly, inconsistently deciding how to do things.
- o Standardization makes it easier to find mistakes.

4. Automation

- o Automated procedures prevent mistakes.
- o Time invested learning automation can save time.

5. Simplicity

- o The more complicated your procedures the more likely you will make mistakes or abandon your workflow.

6. Usability

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

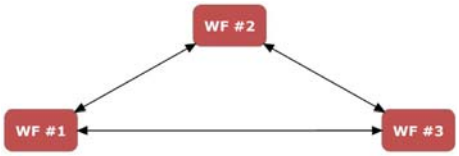
Collaboration and workflow

1. Collaboration makes it more difficult to have an effective, efficient, and replicable workflow.
2. Disciplines that depend more on collaboration put more emphasis on an explicit workflow.
3. Why is workflow harder when you collaborate?

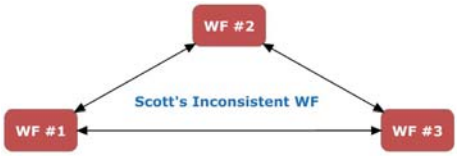
Coordinating multiple workflows

Coordinating multiple workflows

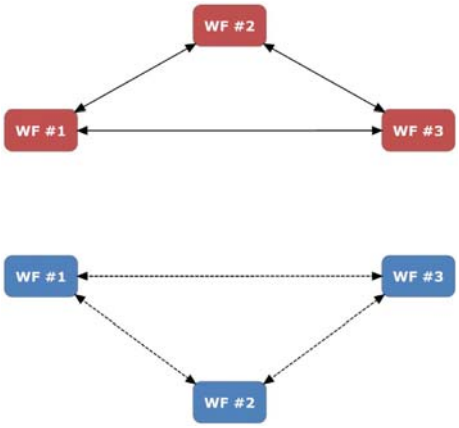
Coordinating multiple workflows



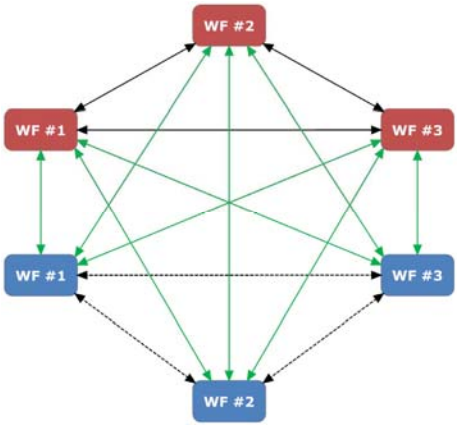
Coordinating multiple workflows



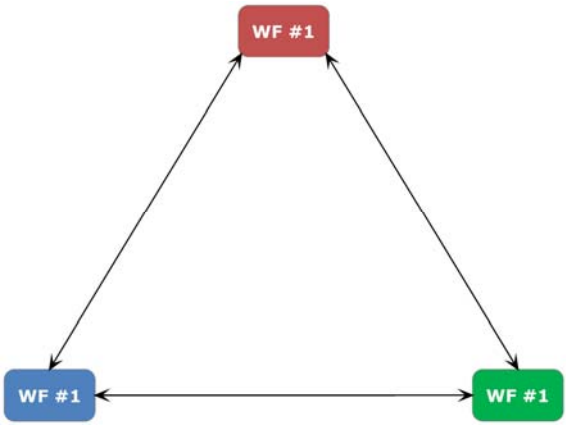
Coordinating multiple workflows starts here



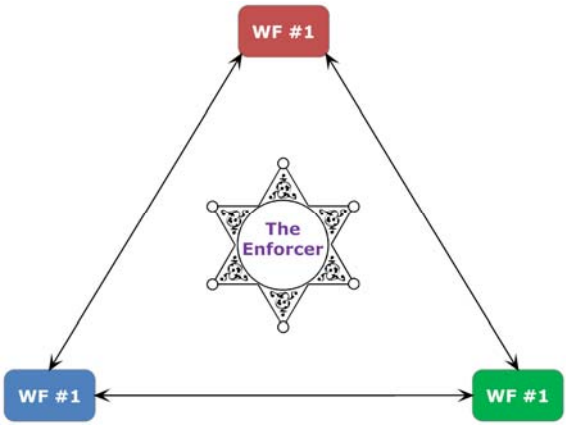
Coordinating 30 pairs of workflows



Coordinating multiple workflows: Agree on a WF



Get an enforcer



Steps in your workflow

1. Prepare data for analysis

- 90% of your work, unless you hurry (then it takes longer).

2. Conduct statistical analyses

- Often the simplest step in your WF.

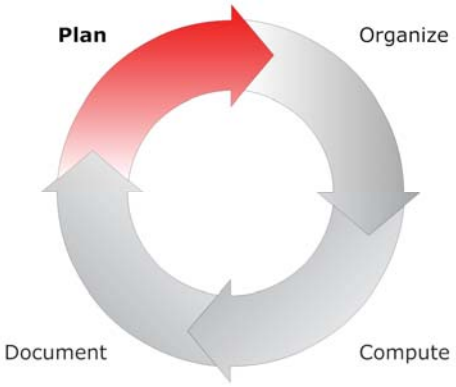
3. Present results

- Incorporating results while maintaining the *provenance*.
- Making effective presentations.

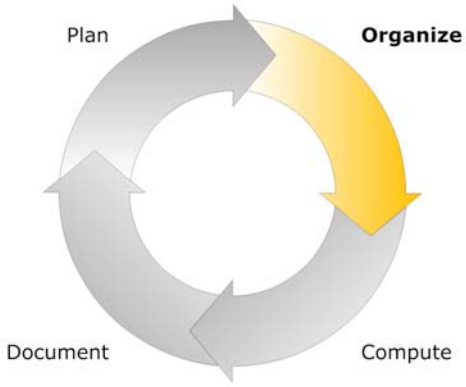
4. Protecting files

- Replication is hard when you have the data and script-files.
- It is impossible without them.

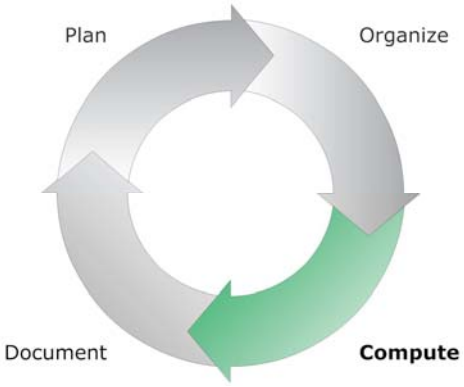
Tasks within each step



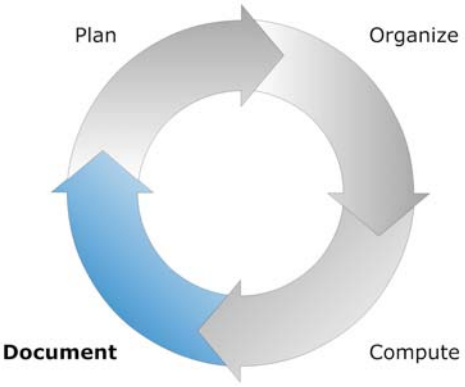
Tasks within each step



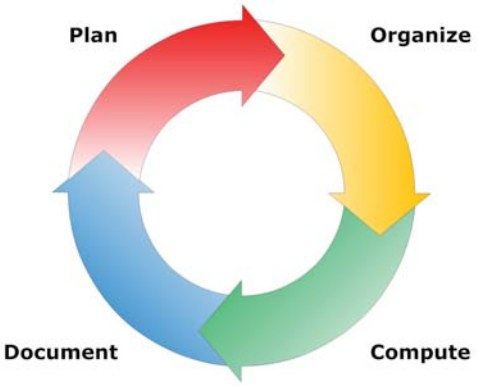
Tasks within each step



Tasks within each step



Tasks within each step



Planning

- Your time.
- Publishing plans and deadlines.
- Division of labor.
- Proposal for data construction: names, labels, formats.
- Procedures for missing data.
- Anticipated analyses.
- Documentation and organization.
- Safe guarding files.
- And more…

Blau and Duncan's *The American Occupational Structure*

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

Michael Faraday's famous sign

Work. Finish. Publish. *--Michael Faraday's sign in his lab*

What is a plan

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

Organizing

1. Organization is motivated by two goals:
   - **Finding things**
   - **Avoiding duplication**
2. Organization…
   - Helps you work faster.
   - It rewards consistency and uniformity.
   - It is contagious.

Signs of poor organization

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

     This: final_report_**v16**.docx

     Or this: NSF_gsssci_report 2010-10-21.docx
   - After a talk at ICPSR, someone showed me this text: "Urgent: don't analyze **`final.dta`**, use **`lastversion.dta`** for our presentation tomorrow."

Organization should be like a Model T

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

Too often it is more like this

With predictable consequences

Digital assets and 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 a file.
3. It is easy to create lots of files.
    - 115,000 files on a research center's LAN
    - 2,000,000 files accumlated in 10 years

Where are your files?

1. Laptop
2. LAN
3. Dropbox
4. Box
5. USB sticks
6. Old laptop
7. Friend's laptop
8. External drives
9. Mom's computer

Operating systems focus on entertainment

Win	**Mac**
Desktop	Desktop
Music	Music
Pictures	Pictures
Videos	Movies
Documents	Documents

Digital asset management (DAM)

1. Suppose I put my files in **\Dropbox**.
2. How do I organize things within **\Dropbox**?
3. Start with general categories of files.

For example...

Primary directories

\Active	Research projects I am actively working on.
\Admin	Administrative files, templates, etc.
\Bookshelf	Books, articles, reprints, etc.
\Inactive	Incomplete projects that are on hold.
\Programs	Files that customized installed programs.
\Service	Documents related to service work.
\Shared	Files shared with others.
\Students	Files from students.
\Teaching	Class materials.
\Templates	Sample files used as templates.
\Vault	Completed work that will **never** change.

Within a primary directory, make subdirectories

\Bookshelf

\Articles

\Books

\Computing

\Figures

And so...

Where to put "stuff"

\- Hold then delete

\- To shelve

\- To transfer

A standard structure for projects

\Group Differences

\- History starting 2016-01-20

\- Hold then delete

\- To shelve

\Admin

\Preposted

\Posted

\Resources

\Work

\Write

Organization: uniform formats for do-files

```
capture log close
log using wftalk01-example, replace text
version 14.1
clear all
set linesize 80

// project: wf talk
// task:
local pgm wftalk01
local dte 2016-01-20
local who scott long
local tag "`pgm'.do `who' `dte'"

// #1 description of task 1

// #2 description of task 2

log close
exit
```


Documentation

1. **Long's Law**: It is faster to document it today than tomorrow.

   Addendum 1: Nobody likes to write documentation.

   Addendum 2: Nobody regrets having documentation.

   How often do you hear: *"Drat, I have too much documentation."*

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

3. The more codified the field the greater the emphasis on documentation

   The Research Log by the American Chemical Society

Suggestions for writing documentation

1. Multiple, reinforcing levels: logs, metadata, comments, names
2. Do it today.
3. Check it next week since it always makes sense today.
4. Keep up with documentation by linking it to events in the project.
5. Include full dates and names.

The core of your documentation: the research diary

1. The diary is the road map connecting all activities and files.
2. Script files precisely describe what is done.
3. Metadata in datasets links to scripts and documents data.

An 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 (flatp5; 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. File manager: Explorer and Finder do not work well.*
     - *b. Macro program: Why retype the same things?*
     - *c. Text editor: Why not just one for all programs?*
     - *d. Word processor: page feed, outlines, headings.*
     - *e. Statistical software.*
   - Hardware: display, storage, memory, CPU
3. Planning is more important than computing power.
   - Consider the changes in 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

10,000 times more

Free Agent with 1TB storage

350,000 times more...

- **Cost of computing $400.**
- **Mean time to degree 7.6 years.**

A thought experiment on planning and computing

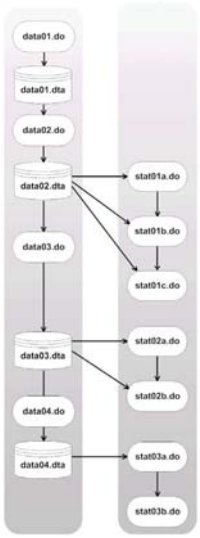
1. Divide yourselves into two groups:
   - **Computers** can compute whenever they want to (i.e., all the time).
   - **Planners** can compute for three four-hour sessions a week.
2. Who finishes first?

Principles for a computing workflow

1. Dual workflow
2. Run order naming
3. Posting files

Dual workflow

Data management ==>



<== Data analysis

Run order and a dual workflow

Name files so if re-run in alphabetical order, you produce ***exactly*** the same results.

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 posting principle (is critical)

Posting is defined by two simple rules.

The share rule

Only share results after the files are posted.

The no change rule

Once a file is posted, ***never*** change it.

Data analysis: use script files

Robust script files

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

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

Legible script files: output that is easy to read

1. Include thoughtful comments.
2. Format with alignment, indentation, and spacing.
3. Use short lines without wrapping.
4. Avoid ambiguous abbreviations: `l a l in 1/3`

Legible output files

```
    +----------------+
    | 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
```



```
8.33      46.43       8.33 |    100.00

-----------+--------------------------------------------------------------------------------
-----------------------+----------
     WhiteCol |         0          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

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

-----------+--------------------------------------------------------------------------------
-----------------------+----------
        Total |         1          8          4         12          9         10
19        109         30 |       337
              |      0.30       2.37       1.19       3.56       2.67       2.97
5.64      32.34       8.90 |    100.00



               |                                    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
-----------+--------------------------------------------------------------------+----------
-----------------------------+----------
```


Automation

1. Much of data analysis involves repetitive tasks.
2. Repetition invites errors.
3. Automation is less error prone and ultimately faster.
   A. **macros**: words that represent strings of text
   B. **loops**: multiple execution of the same commands
   C. **returned results**: avoids typing the value of a statistical result
   D. **matrices**: hold and summarize results

Data cleaning, including names and labels

Planning labels

Bad labels

```
. codebook tc1*, compact

Variable    Obs Unique      Mean Min  Max  Label
-------------------------------------------------------------------------------
tc1doc     1074     10 8.714153   1   10  Q46 How important is it to go to ...
tc1fam     1074     10 8.755121   1   10  Q43 How important is it to turn t...
tc1friend  1073     10 7.799627   1   10  Q44 How important is it to turn t...
tc1mhprof  1045     10  7.58756   1   10  Q48 How important is it to go to ...
tc1psy     1050     10 7.567619   1   10  Q47 How important is it to go to ...
tc1relig   1039     10  5.66025   1   10  Q45 How important is it to turn t...
```

Better labels

```
. codebook tc2*, compact

Variable    Obs Unique      Mean Min  Max  Label
-------------------------------------------------------------------------------
tc2doc     1074     10 8.714153   1   10  Q46 How Impt: Go to a gen med doc...
tc2fam     1074     10 8.755121   1   10  Q43 How Impt: Turn to family for ...
tc2friend  1073     10 7.799627   1   10  Q44 How Impt: Turn to friends for...
tc2mhprof  1045     10  7.58756   1   10  Q48 How Impt: Go to a mental heal...
tc2psy     1050     10 7.567619   1   10  Q47 How Impt: Go to a psych for Help
tc2relig   1039     10  5.66025   1   10  Q45 How Impt: Turn to a religious...
```


Even better labels

```
. codebook tc3*, compact

Variable    Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
tc3doc     1074     10  8.714153    1   10  Q46 Med doctor help important
tc3fam     1074     10  8.755121    1   10  Q43 Family help important
tc3friend  1073     10  7.799627    1   10  Q44 Friends help important
tc3mhprof  1045     10   7.58756    1   10  Q48 MH prof help important
tc3psy     1050     10  7.567619    1   10  Q47 Psychiatric help important
tc3relig   1039     10   5.66025    1   10  Q45 Relig leader help important
-------------------------------------------------------------------------------
```

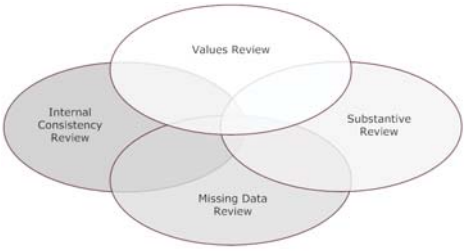
Planning labels

	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

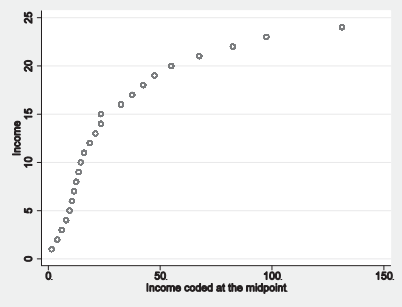
Planning variables names

1. **ownsex** and **ownsexu** caused weeks of delay.
2. Who can remember whether you want **R003189** or **R001389**?

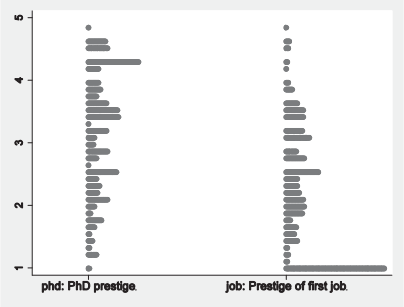
Types of data cleaning (to avoid retractions!)



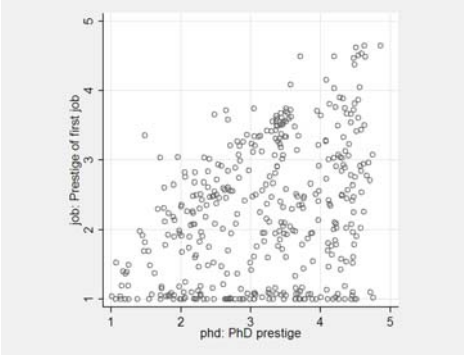
Cleaning 1: find errors with a graph



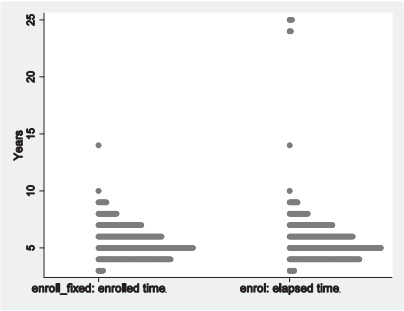
Cleaning 2: remembering a coding decision



Cleaning 3: understanding the substantive process



Cleaning 4: avoiding expensive mistakes



Analyzing the data

1. Take classes in statistics
2. Go to talks on data analysis
3. Find exemplars.

Presentations and provenance

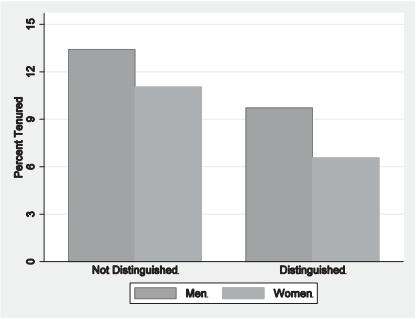
1. **Content and methods** are disciplinary decisions.
2. **Presentations and provenance** are universal.

Tables too small

```
eTwgt (N=337): Factor Change in the Odds of occ

Variable: white (sd=.276422568)

Odds comparing          |
alternative 1           |
to alternative 2        |      b         z     P>|z|     e^b   e^bSE

Menial  BlueCol         |  -1.23868    -5.797    0.000   .2896   .0.7185
Menial  Craft           |  -0.47733    -2.763    0.006   .6205   .0.8776
Menial  WhiteCol        |  -1.97135    -7.741    0.000   .1393   .0.4277
Menial  Prof            |  -2.17855    -7.757    0.000   .1132   .0.3475
BlueCol Menial          |   1.23868     5.797    0.000   3.4529   1.2870
BlueCol Craft           |   0.76135     3.268    0.001   2.1417   1.2352
BlueCol WhiteCol        |  -0.73267    -2.974    0.003   .4806   .0.2324
BlueCol Prof            |  -0.93790    -3.873    0.001   .3918   .0.2624
Craft   Menial          |   0.47733     2.763    0.006   1.6117   1.1396
Craft   BlueCol         |  -0.76135    -3.268    0.001   .4671   .0.2145
Craft   WhiteCol        |  -1.49401    -5.310    0.179   .2333   .0.7389
Craft   Prof            |  -1.70190    -7.011    0.946   .2729   .0.6978
WhiteCol Menial         |   1.97135     7.741    0.000   7.1793   2.5975
WhiteCol BlueCol        |   0.73267     2.974    0.179   2.0973   1.0979
WhiteCol Craft          |   1.49401     5.310    0.179   4.4613   1.3998
WhiteCol Prof           |  -0.20789    -0.833    0.405   .8123   .0.2530
Prof    Menial          |   2.17855     7.757    0.000   8.8552   3.0261
Prof    BlueCol         |   0.93790     3.873    0.001   2.7122   1.1663
Prof    Craft           |   1.70190     7.011    0.946   5.4858   1.3997
Prof    WhiteCol        |   0.20789     0.833    0.405   1.2304   .0.3877

Variable: ed (sd=3.8382371)

Odds comparing          |
alternative 1           |
to alternative 2        |      b         z     P>|z|     e^b   e^bSE

Menial  BlueCol         |  -0.99410    -7.972    0.331   .3345   1.7484
Menial  Craft           |  -0.77538    -6.700    0.000   .4606   .0.1965
Menial  WhiteCol        |  -0.59210    -3.811    0.000   .5535   .0.2533
Menial  Prof            |  -0.77538    -6.700    0.000   .4606   .0.1965
BlueCol Menial          |   0.99410     7.972    0.000   2.7045   0.7484
BlueCol Craft           |   0.19854     2.454    0.013   1.2196   0.5699
BlueCol WhiteCol        |   0.40298     4.429    0.000   1.4956   0.3936
BlueCol Prof            |   0.19854     2.454    0.013   1.2196   0.5699
Craft   Menial          |   0.77538     6.700    0.336   2.0954   1.3184
Craft   BlueCol         |  -0.19854     2.454    0.013   1.2132   1.7871
Craft   WhiteCol        |   0.20444     2.773    0.006   1.2264   .0.2471
Craft   Prof            |   0.00000     0.071    0.000   1.0043   0.1329
WhiteCol Menial         |   0.59210     3.811    0.003   1.8230   2.9368
WhiteCol BlueCol        |  -0.40298     4.429    0.000   .6686   .0.1760
WhiteCol Craft          |  -0.20444     2.773    0.006   .8151   .0.2471
WhiteCol Prof           |  -0.20389     1.616    0.000   0.8153   0.1853
Prof    Menial          |   0.77538     6.700    0.000   2.1717   1.3184
Prof    BlueCol         |   0.87939     6.726    0.000   2.4057   10.2492
Prof    Craft           |   0.00044     7.071    0.000   1.9030   7.5264
Prof    WhiteCol        |   0.20389     0.616    0.000   1.2257   0.3853

Variable: exper (sd=13.9969041)
```

Colors that aren't

The green bar shows…



Documenting provenance

The provenance of every number must be fully documented.

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

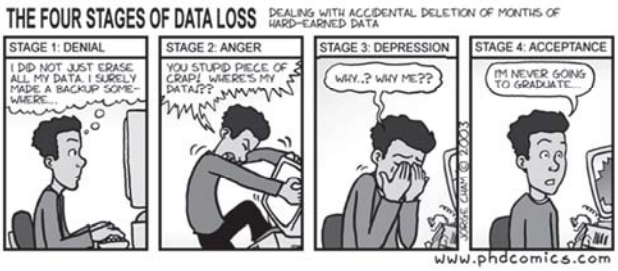
1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons (.48 and .73, respectively (z=2.55, p<.01)) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

2. 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

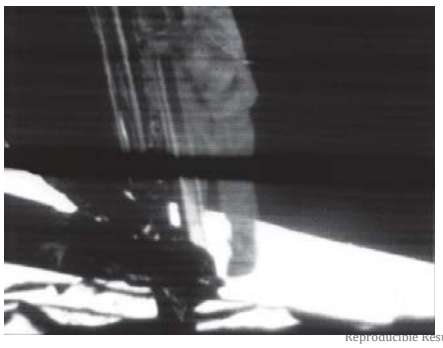
Preserving your data

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.



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.
3. Neil Armstrong, July 20, 1969: "a ***fuzzy*** gray ***blob*** wading through an inkwell".



What NASA saw and lost…

But, two tapes were archived by Pink Flyod's video producer!

Dark Side of the Moon

Tactics: Peer to Peer syncing

Dropbox box SugarSync

1. Install the cloud service on your computer.
2. Put files into the folder.
3. Retrieve them from any machine connected to the service.
4. Have shared folders for collaboration.

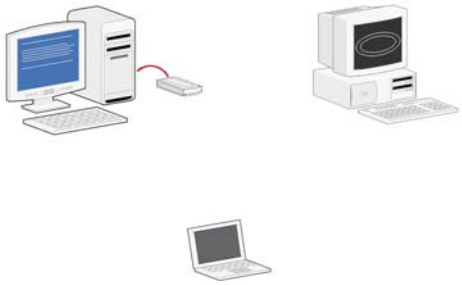
A recent disaster and the advantages of the cloud

1. While working on dissertation, a student was moving. Computer, home backup drive, and work backup in the same box.
2. The box was stolen and student. Student dropped out of Ph.D.

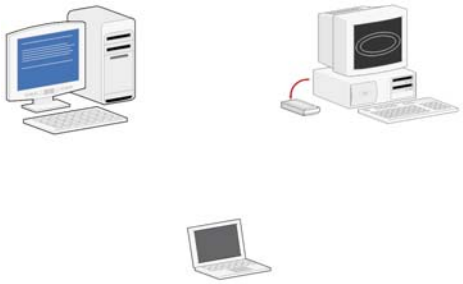
Caution

Security and IRB requirements **must** be met.

Tactics: Portable drives at home



Tactics: Portable drive at work



Preserving bits and preserving content

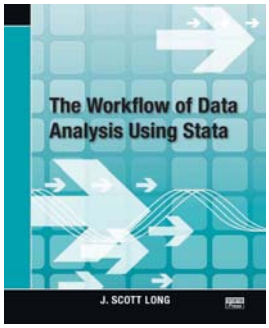
These files were generated six or seven years ago using Gauss and saved as Gauss FMT files. We need to revise a paper and need the data in these files, but I don't seem to be able to open them. We only have a very old version of Gauss that might not run anymore. Any ideas?

Changing your workflow

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

Whose workflow

There are **many** viable workflows, but it is nice to have your workflow written down.



Questions?

Let me know what's happening in your field.

Thank you!

Provenance: 2009-07-12; 2009-07-31; 2009-11-12; 2010-07-29; 2010-09-03; 2010-11-04; 2011-04-18; 2011-09-09; 2012-09-04; 2013-03-04; 2013-10-21; 2014-06-30; 2014-07-02; 2014-08-25 wf wim 2014-08-25.docx; wf wim 2015 2015-08-21.docx; wf wim 2016-01-18v1.docx; wf wim 2016-01-21.docx