What do I do with all these numerical simulations?

Roy Goodman
Graduate Student Summer Seminar

The basic problem: How to manage the computer programs and data produced over the course of a numerical study

- In what format should I save data?
- How can I find my data once I create it?
- How do I best communicate my discoveries?
- What are good programming habits and tools that can help me work efficiently?

A MATLAB dilemma

A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively

A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively
- Then you quit the program and the data & figures are gone

A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively

- Then you quit the program and the data & figures are gone

- Next month--time to write the paper, and there's no data

A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively
- Then you quit the program and the data & figures are gone
- Next month--time to write the paper, and there's no data
- This is less of a problem with compiled programs called from command line, since the main way to get data out is by writing to files

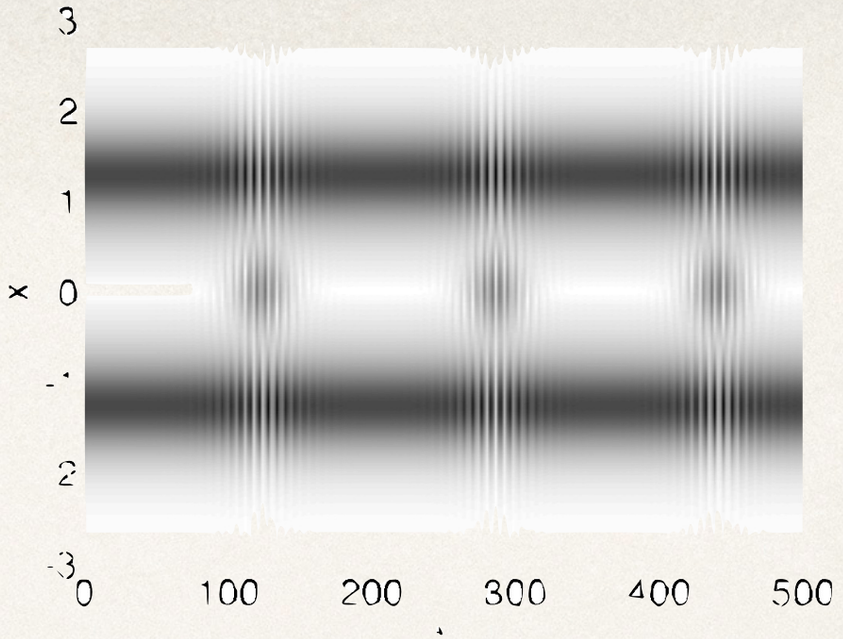
A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively
- Then you quit the program and the data & figures are gone
- Next month--time to write the paper, and there's no data
- This is less of a problem with compiled programs called from command line, since the main way to get data out is by writing to files
- The solution is to save the outputs to files, but how should you do this systematically? (In MATLAB or a compiled language)

A MATLAB dilemma

- It is easy to sit at your desk and generate data & figures in MATLAB to view interactively
- Then you quit the program and the data & figures are gone
- Next month--time to write the paper, and there's no data
- This is less of a problem with compiled programs called from command line, since the main way to get data out is by writing to files
- The solution is to save the outputs to files, but how should you do this systematically? (In MATLAB or a compiled language)

A Problematic Figure:



- Q: What equation does this figure represent? Hopefully you can remember / figure out $\dot{x} = f(x, \Omega_1, \Omega_2, \Omega_3, \alpha)$
- Q: What parameters were used in the equation?
- Q: What program was used to solve the equation?
- Q: What numerical parameters were used in that program?
- Q: Can I recreate the figure? Do I still have that program? Is the program as it exists today the same as when I created the figure?
- Similar problems with numerical/statistical data stored as lists of numbers and with paper & pencil calculations.

A problematic situation

- You write a paper:

RL **98,** 104103 (2007) PHYSICAL REVIEW LETTERS 9 MARCH 2007

Chaotic Scattering and the *n*-Bounce Resonance in Solitary-Wave Interactions

Roy H. Goodman*
Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, New Jersey 07102, USA

Richard Haberman†
Department of Mathematics, Southern Methodist University, Dallas, Texas 75275, USA

[illegible] 21 August 2006; published 9 Mar[illegible]

- Fields Medal refuser Grigory Perelman emails you:

From: Grisha Perelman <[illegible]com>
Subject: **Figure 6 in your PRL paper**
Date: July 30, 2010 12:22:56 PM EDT
To: Roy Goodman <goodman@NJIT.EDU>

Dear Prof. Goodman,

I think your paper is interesting, if elementary. I am currently working on doing things the RIGHT WAY, but my analysis contradicts your numerical results depicted in figure 6. Are you sure you did this right?

Cordially,

Grisha

- How do you show him that your calculation is correct?
  Which of the questions from previous slide can you answer?

A more realistic situation

A more realistic situation

- You are about to graduate (yay!).

A more realistic situation

- You are about to graduate (yay!).
- Your advisor asks you to share your computer programs with his/her beginning Ph.D. student so they can continue working on a related problem.

A more realistic situation

- You are about to graduate (yay!).
- Your advisor asks you to share your computer programs with his/her beginning Ph.D. student so they can continue working on a related problem.
- Can anyone else make your code work and modify it to solve new problems?

A more realistic situation

- You are about to graduate (yay!).
- Your advisor asks you to share your computer programs with his/her beginning Ph.D. student so they can continue working on a related problem.
- Can anyone else make your code work and modify it to solve new problems?
- How much work would it take to make your programs shareable?

A culture problem

A culture problem

- Suppose you go to work in a biochemist's lab:
  - On your first day, the lab manager shows you a shelf full of lab notebooks and teaches you the protocol.
  - Every experiment run & every measurement made goes in the notebook along with dates & times.
  - (At least this is how we're told it works!)

A culture problem

- Suppose you go to work in a biochemist's lab:
  - On your first day, the lab manager shows you a shelf full of lab notebooks and teaches you the protocol.
  - Every experiment run & every measurement made goes in the notebook along with dates & times.
  - (At least this is how we're told it works!)

- In applied mathematics
  - Your advisor says "compute this" but gives no instructions on how to deal with the computer programs or the data generated.
  - (My personal experience) Over time you develop a "system" for managing all this, but it's piecemeal and incomplete. Other people must have done the same thing. *Can we help each other with this?*

Numerical Mathematics is an Experimental Science

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?
- These are important issues for scientific **efficiency**, **reproducibility**, and **provenance**.

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?
- These are important issues for scientific **efficiency**, **reproducibility**, and **provenance**.
  - **Efficiency**--do something once, know how to find it again (i.e. filing), modify it quickly, teach others to use it

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?
- These are important issues for scientific **efficiency**, **reproducibility**, and **provenance**.
  - **Efficiency**--do something once, know how to find it again (i.e. filing), modify it quickly, teach others to use it
  - **Reproducibility/Refutability**--That's what makes it science!

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?
- These are important issues for scientific **efficiency**, **reproducibility**, and **provenance**.
  - **Efficiency**--do something once, know how to find it again (i.e. filing), modify it quickly, teach others to use it
  - **Reproducibility/Refutability**--That's what makes it science!
  - **Provenance**--How/where/when was this data produced? Is it yours?

Numerical Mathematics is an Experimental Science

- How many of you have (or expect to have) numerical simulations (or large statistical data sets) as part of your thesis?
- How do you manage the software and numerical data you produce?
- These are important issues for scientific **efficiency**, **reproducibility**, and **provenance**.
  - **Efficiency**--do something once, know how to find it again (i.e. filing), modify it quickly, teach others to use it
  - **Reproducibility/Refutability**--That's what makes it science!
  - **Provenance**--How/where/when was this data produced? Is it yours?

Some Solutions
(incomplete and overlapping)

- Software:
  - Modular programming and makefiles
  - Matlab publishing
  - Version Control Software
  - Wikis (MoinMoin is great)
  - Note-taking software (e.g. Evernote)
  - Scanner (ok, this is Hardware)
  - Backups!
  - Graphical debugger
- Good Habits:
  - Clean and commented code
  - Write-as-you go
  - Disciplined computing
  - Self-documentation/logging
  - Describe numerics in your papers or your "internal documentation"
- Changing the Culture
  - More public discussion of these issues
  - A forum for sharing computational tools & ideas (not just algorithms)
  - Develop a computational infrastructure that allows one to accomplish these goals without too much overhead
  - Ph.D advisors should insist graduate students learn to work this way
  - Journals should reject papers based on inadequately-documented numerics

Some Solutions
(incomplete and overlapping)

- Software:
  - Modular programming and makefiles
  - Matlab publishing
  - Version Control Software
  - Wikis (MoinMoin is great)
  - Note-taking software (e.g. Evernote)
  - Scanner (ok, this is Hardware)
  - Backups!
  - Graphical debugger
- Good Habits:
  - Clean and commented code
  - Write-as-you go
  - Disciplined computing
  - Self-documentation/logging
  - Describe numerics in your papers or your "internal documentation"
- Changing the Culture
  - More public discussion of these issues
  - A forum for sharing computational tools & ideas (not just algorithms)
  - Develop a computational infrastructure that allows one to accomplish these goals without too much overhead
  - Ph.D advisors should insist graduate students learn to work this way
  - Journals should reject papers based on inadequately-documented numerics

I can't address all of these in 50 minutes, so I'll concentrate on a few.

Keep a diary

- You work on stuff, then you leave it alone for a while, and it takes you time and effort to figure out what you've done later.

- Good to keep track of effort, save important results.

- Evernote is the obvious piece of software but it lacks the ability to typeset mathematics

- I have been using Quiver for the Mac and iPhone, which can render both LaTeX and Markdown (to be discussed)

Saving data in MATLAB: Trying to make your work self-documenting

- This program saves a file `data.mat`
- Run it twice, and the first data will get overwritten.

```
function [p,t]=plustimes(x,y)
p=x+y;
t=x*y;
save data p t x y
```

- This is better, but you have to think of a new name each time

```
function [p,t]=plustimes(x,y,datafile)
% datafile must be a string with a file name
p=x+y;
t=x*y;
save(datafile, 'p', 't', 'x', 'y');
```

- This has MATLAB generate the filename algorithmically

```
function [p,t]=plustimes(x,y)
% automatically generates name of datafile
datafile=get_datafile(x,y);
p=x+y;
t=x*y;
save(datafile, 'p', 't', 'x', 'y');
```

How to generate filenames

- Based on parameters `data_x_0p5_y_m1p5`
  - **Advantage:** can tell what's in it by its name
  - **Disadvantages:** unwieldy if many parameters, hard to index, doesn't work with randomness
- Sequentially: `datafile.001` `datafile.002`, etc.
  - **Disadvantage:** no connection between filename and contents
  - **Solution:** easy to index, more on this later

```
function filename = get_datafile(x,y)
filename=['data_x_' num2str(x) '_y_' num2str(y)];
```

```
function filename = get_datafile
% Look for a file that says how many times the
% code has been run, and update the file.
n=get_next_filenumber;
filename=['data_x_' int2str(n)];
```

MATLAB "Publish" feature

- The best way to remember what you've done is to WRITE IT DOWN
- Old fashioned way: keep a bound paper journal
- More modern way: on the computer, but LaTeXing everything up would be a lot of work, and take too much time away from exploration
- MATLAB allows you to generate documents from your programs with very little work
- It generates a webpage incorporating your program, additional formatting, and any outputs of the program.
- Let's look at an example
- It can generate HTML, PDF, LaTeX, or even MS-Office formatted reports
- Your code can be pulled by MATLAB using the `grabcode` command (not subroutines in other files

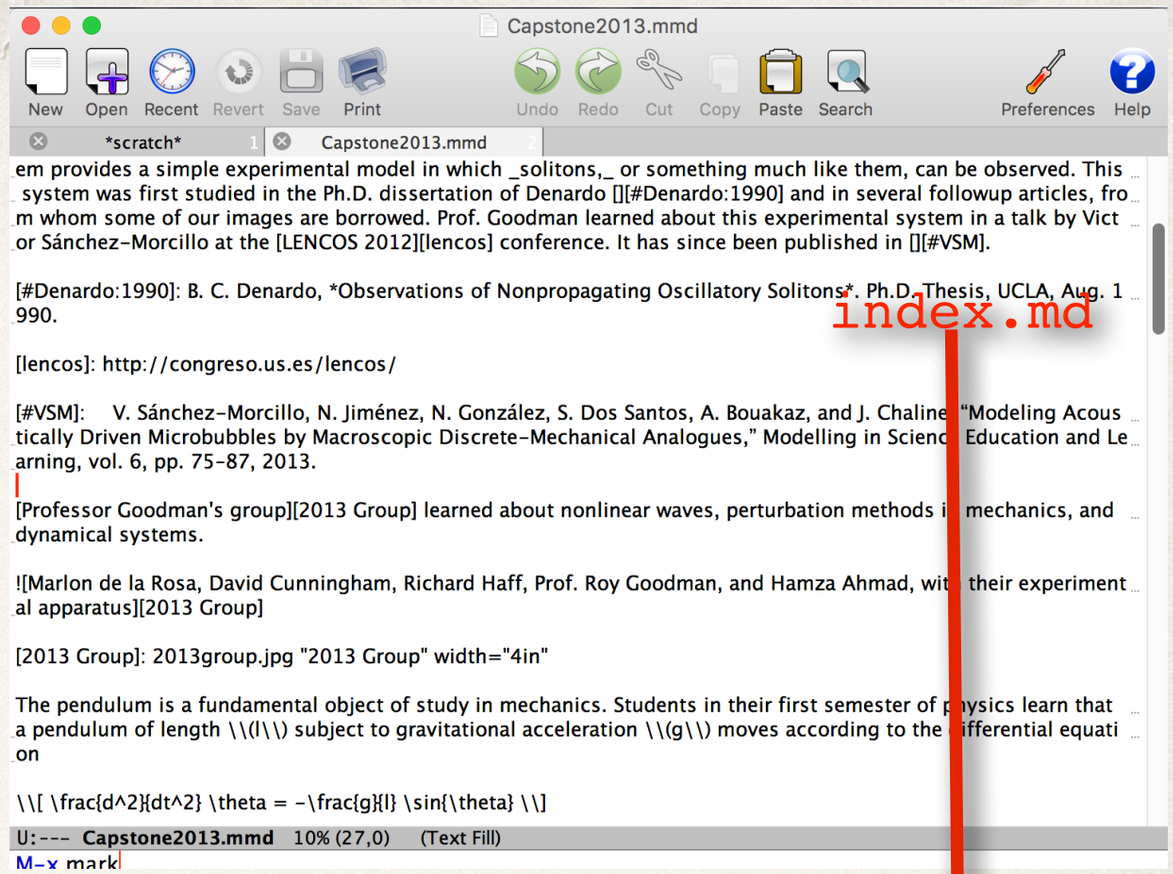
Indexing Data Files

- Okay, I have run the software 15 times & have 15 datafiles. How do I find out what's inside?
- Build an index.
- This is hard to do after the fact, so we should build the index each time I run the software.
- For this, I'll use **`markdown`**.
  - markdown is a lightweight markup language that lets you easily create web pages with very little typing
  - markdown interpreters such as MacDown allow you to edit markdown live and see the results in a separate pane
  - (actually, I use an extension called multi markdown that can make tables

markdown

- A human-readable markup language originally developed to create web pages, but is now used for many applications.

- I use Markdown and MATLAB to create a table that lists all my runs. The amount of data stored to make this table and webpage is very small!

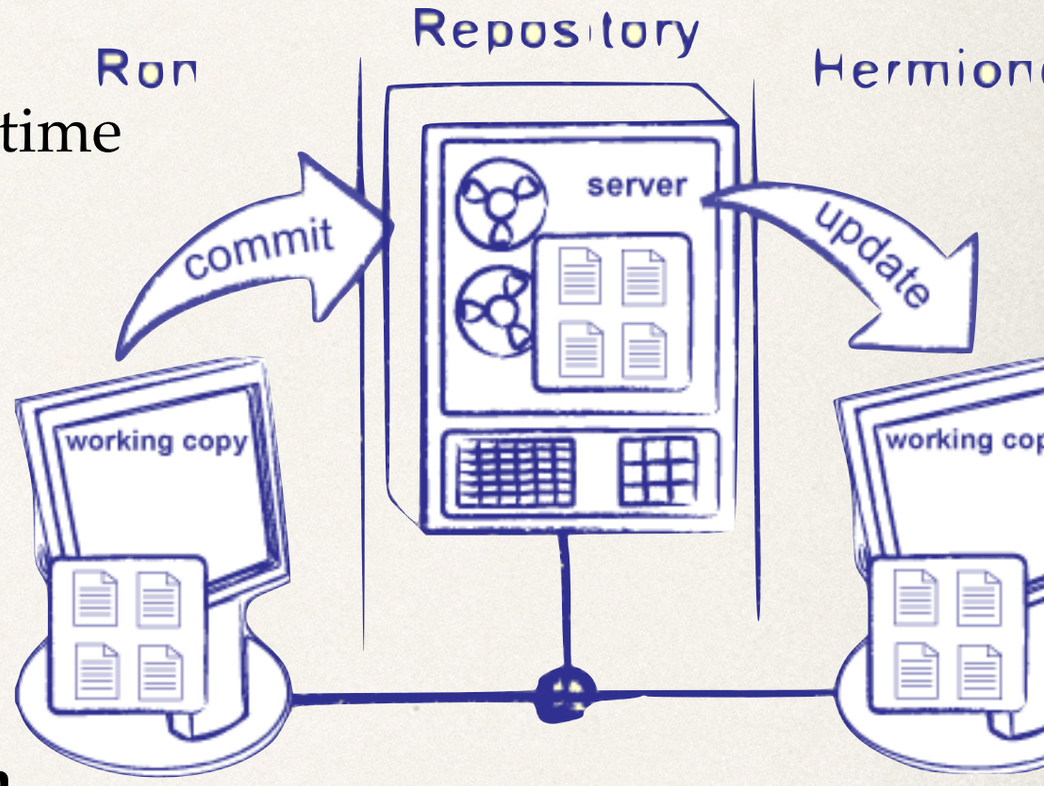


Version Control Software

A system that allows users to monitor and track changes to a set of files.

Problem #1: Collaboration--When 2 or more people want to edit the same file at the same time.

- Option 1: make them take turns
  - But then only one person can be working at any time
  - And how do you enforce the rule?
- Option 2: patch up differences afterwards
  - Requires a lot of re-working
  - Stuff always gets lost

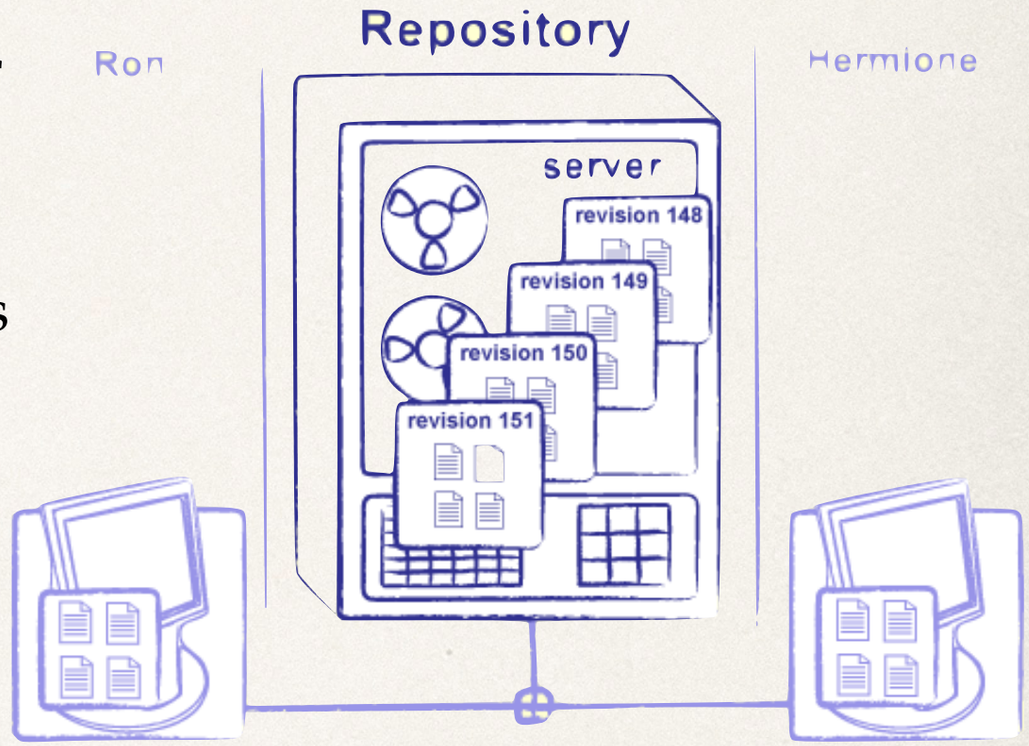


Solution: Version Control

- The right solution is to use a **version control system**
- Keep the **master copy** of the file in a central **repository**
- Each author edits a **working copy**
- When they're ready to share their changes, they **commit** them to the repository
- Other people can then do an update to get those changes

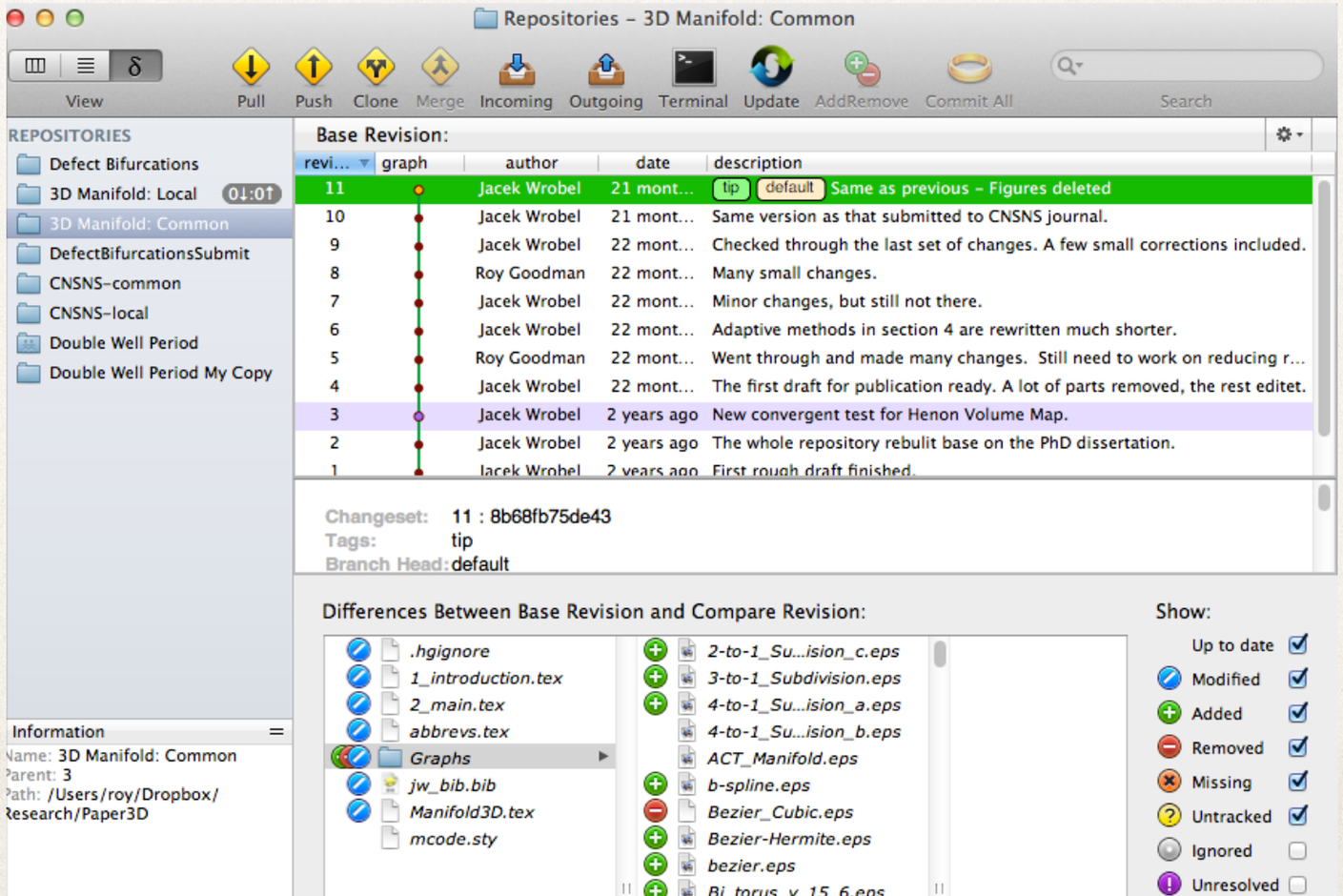
Version Control Software

- Problem #2: Undoing Changes
  - Often want to undo changes to a file
  - Start work, realize it's the wrong approach, want to get back to starting point
  - Like "undo" in an editor...
  - ...but keep the whole history of every file, forever
  - Also want to be able to see who changed what, when
  - The best way to find out how something works is often to ask the person who wrote it
- Solution: Version Control (Again)
  - Have the version control system keep old revisions of files
  - And have it record who made the change, and when
  - Authors can then roll back to a particular revision or time
  - (again) This by itself is reason enough to use version control even when you are the only author



Version Control Software

- Problems:
  - Version Control requires good habits
  - Older programs (cvs, subversion,...) have a steep learning curve, require a server
- Newer programs are much simpler to learn & use, like Mercurial below (http://mercurial.selenic.com)



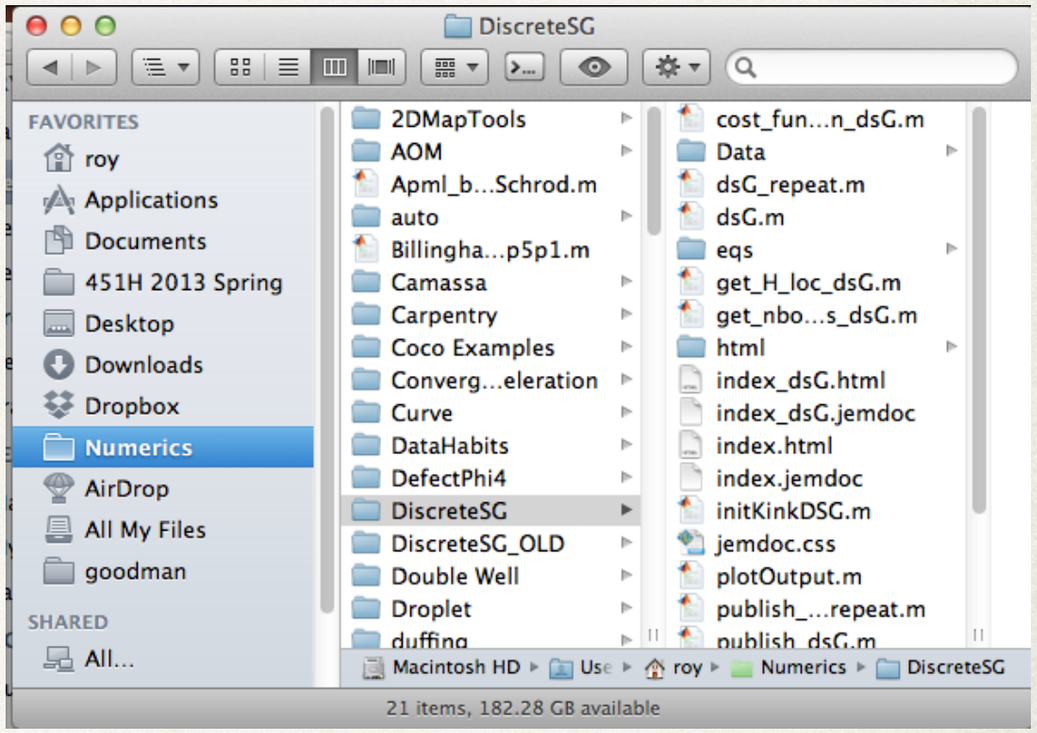
Keeping a lab notebook with Evernote

Indexing runs is good, but it's also a good idea to:

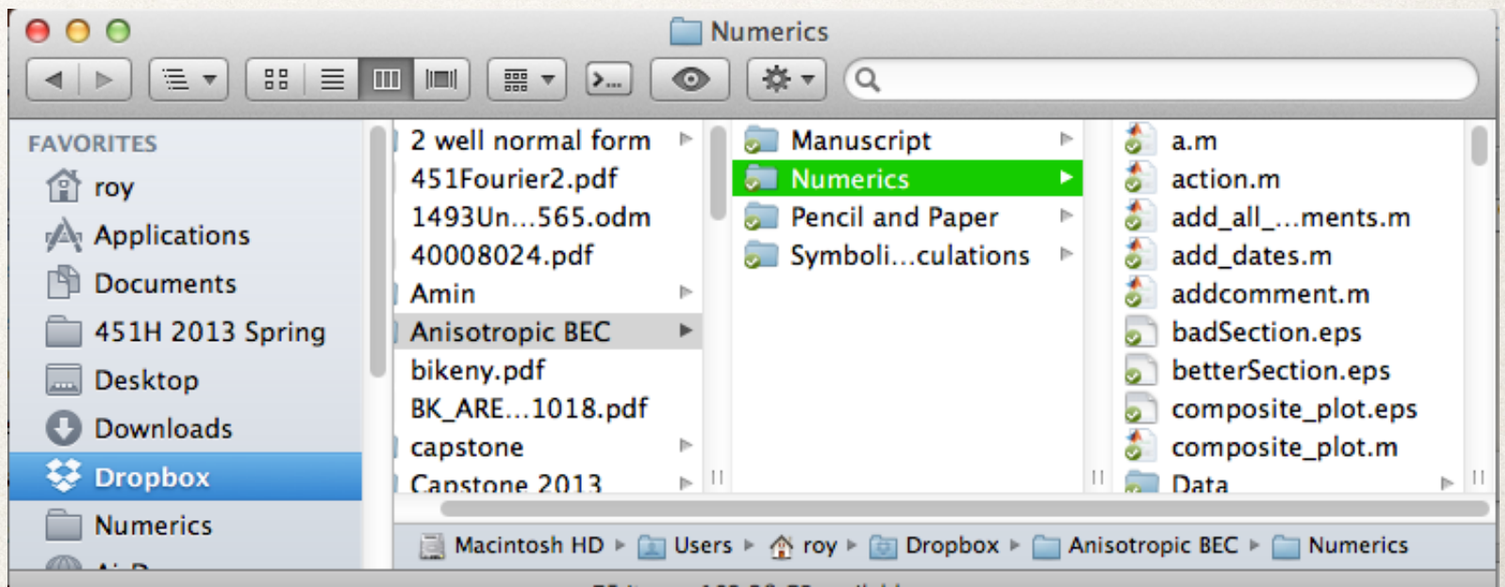
- Why? It's hard to remember what you've done even a few weeks later.
- Projects that get put on hold are difficult to restart. Often find tools used on an old project useful for a new project.
- Keep a more `detailed diary` of simulations run, including motivation for each experiment and more detailed observations (something less than a full writeup), including `figures`
- Scan handwritten notes/calculations and add to diary
- Keep track of papers you read and notes that go with them.
- Discuss & ask questions to members of your research group
- Maintain a to-do list

How to organize your projects

- My Old System was bad: Directories organized by what they contain , each with subdirectories corresponding to different projects
  - "Numerics", "Manuscripts", etc.



- My New system is good:
  - "Projects" directory with one directory for each project
  - Within each project store different types of data. I have four:
    - Numerics
    - Manuscript
    - Symbolic (Mathematica)
    - Scanned Notes



Review

- Keep a diary to remember what you've done

- Use Matlab publish to create readable and reviewable output, save parameters with output

- Version control for your computer programs and LaTeX projects to keep track of changes, undo mistakes, remember what version of program used to create output data

Further ideas

- Modular, reusable computer programs, including routines for managing data
- Makefiles--for managing complex software projects with multiple source code files
- Effective debugging & software testing
- See `http://softwarecarpentry.org`, a project of Greg Wilson.
- The `ReDoc` software at Stanford
- Reference managers such as Papers, Zotero, Mendeley
- Dropbox--share your data with collaborators and access from multiple devices