# STRATEGIES FOR DATA ANALYSIS AND VISUALIZATION

For Increased Collaboration, Openness and Sharing

# This is Not A Talk About How To Analyze and Visualize your Data

- Your probably already better at that than me
- There are too many ways to analyze data
  - Project Specific
  - Domain Specific
- It would be really boring

# This Talk Is About

- Treating your analysis as a first class data object
- Maximizing your efficiency creating analyses and visualizations
- Increasing your ability to use your analyses and visualizations with collaborators
- Making your analyses and visualizations more open.

# General Outline

# Some Background

The Changing Regulatory Landscape

# NSF Guidelines

- □ a. Investigators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work conducted under NSF grants. Grantees are expected to permit and encourage such publication by those actually performing that work, unless a grantee intends to publish or disseminate such findings itself.

# NSF Guidelines

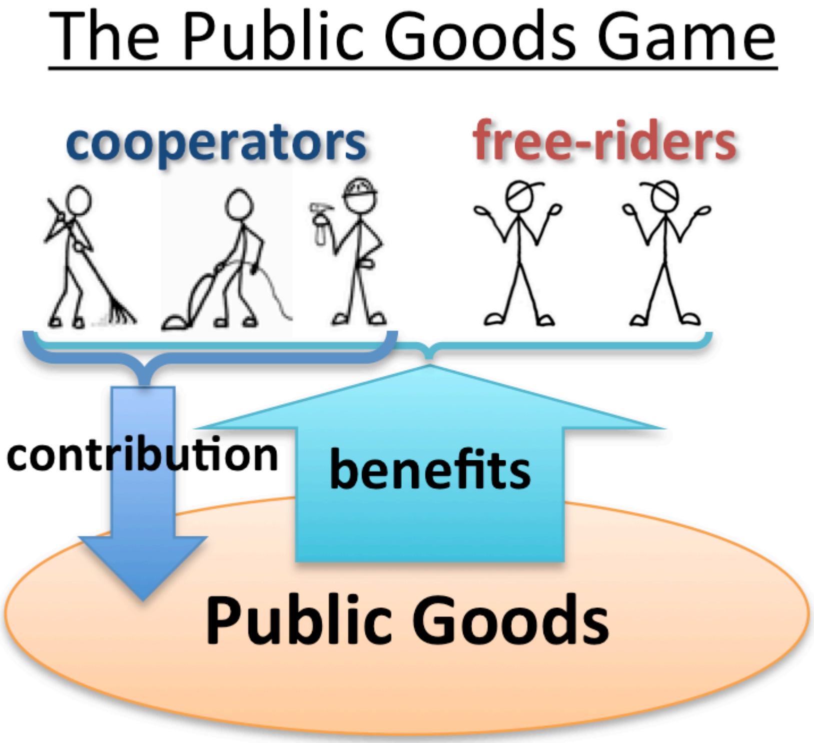
- b. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved.

# NSF Guidelines

- c. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.

# NSF Guidelines

- d. NSF normally allows grantees to retain principal legal rights to intellectual property developed under NSF grants to provide incentives for development and dissemination of inventions, software and publications that can enhance their usefulness, accessibility and upkeep. Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers.



http://plektix.fieldofscience.com/2011/04/freedom-and-public-goods.html

# The Take Home Message

- Your grant funded work is a public good.
- Sharing the products of research benefits society.
- The works is still recognized as your work.
- But you still have to share.

# Why Do Funding Agencies Care?

□ Good PR.

□ Increased accountability.

□ Benefits to the research community.

https://commons.wikimedia.org/wiki/File:Mad_scientist_transparent_background.svg

# Why You Should Care

- □ Increase your efficiency
- □ Facilitate collaboration.
- □ Maximize the usefulness of your data.

- □ Lack of Future Funding
- □ ???

# I thought it was just the data!

Nope. Your other products are first class objects as well.

# Data, Collection, Analyses and Visualizations are Inseparable

- The Analysis Visualizations Help Explain Your Data
- What your data is for.
- An example of how your data can be used.

# Your Analyses and Visualizations Are Important On Their Own

- Reuse
  - Increase your efficiency
  - Solve other peoples problems

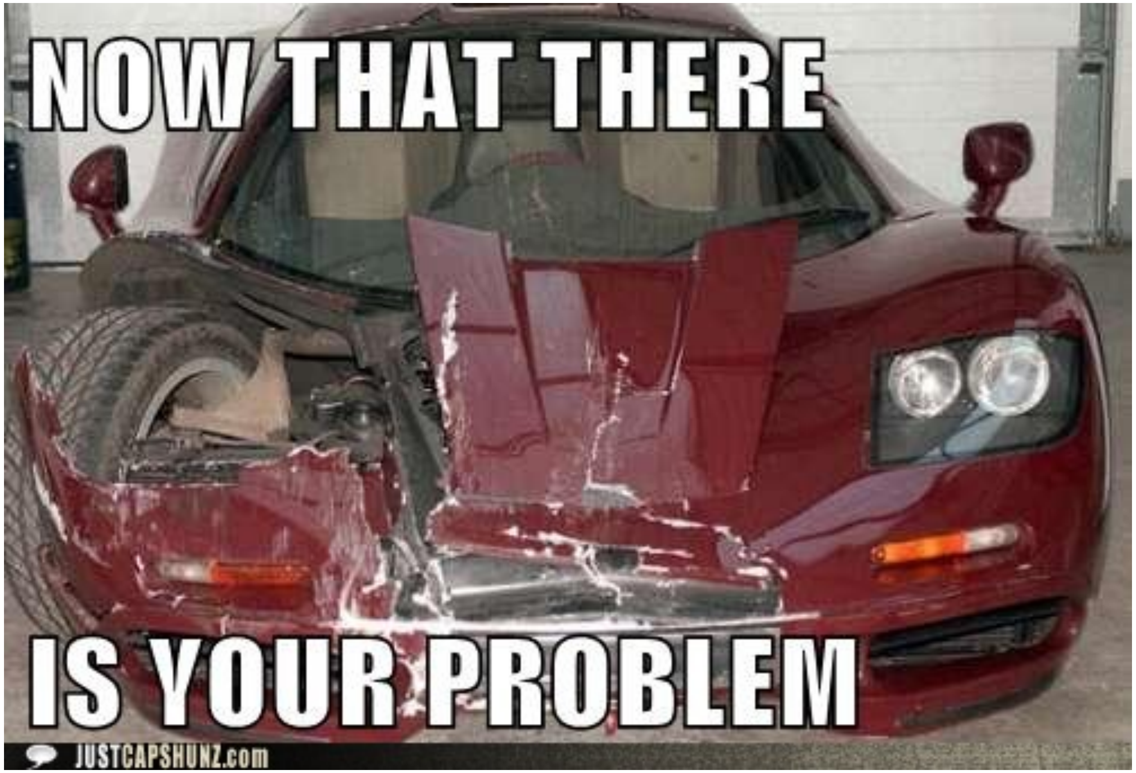
*Good programmers write good code.*
*Great programmers steal great code.*

*-unknown-*

# Functional Requirements

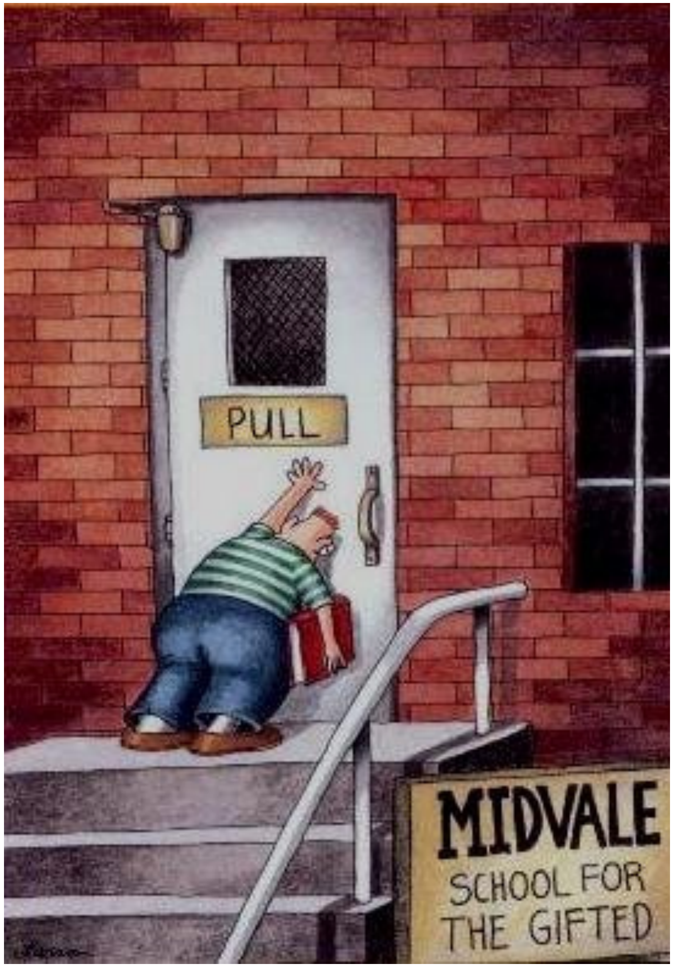
Things your analysis and visualization should do beyond analysis and visualization

# Analyze/Visualize Your Data



http://www.fbclick.com/wp-content/plugins/wp-o-matic/cache/2321c_funny-captions-now-that-there-is-your-problem.jpg

# Understandable



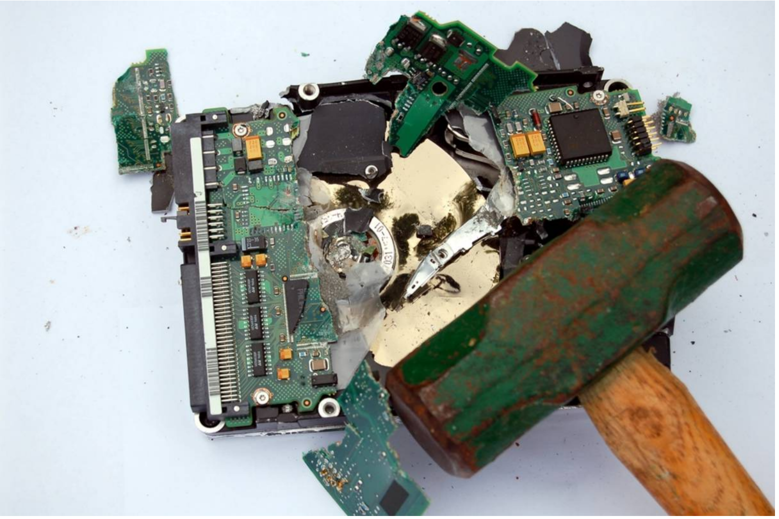
Gary Larson

- You should understand what you have done a year later
- Others should understand what you did without your direct explanation
- Don’t assume we are all geniuses

# Acurate

- Your analyses should render the same results as those you reported
  - Every time
- This includes randomizations.

# Be Durable



- □ Platform independent
- □ Easy to migrate
- □ Easy to translate to other software
- □ Easy to see

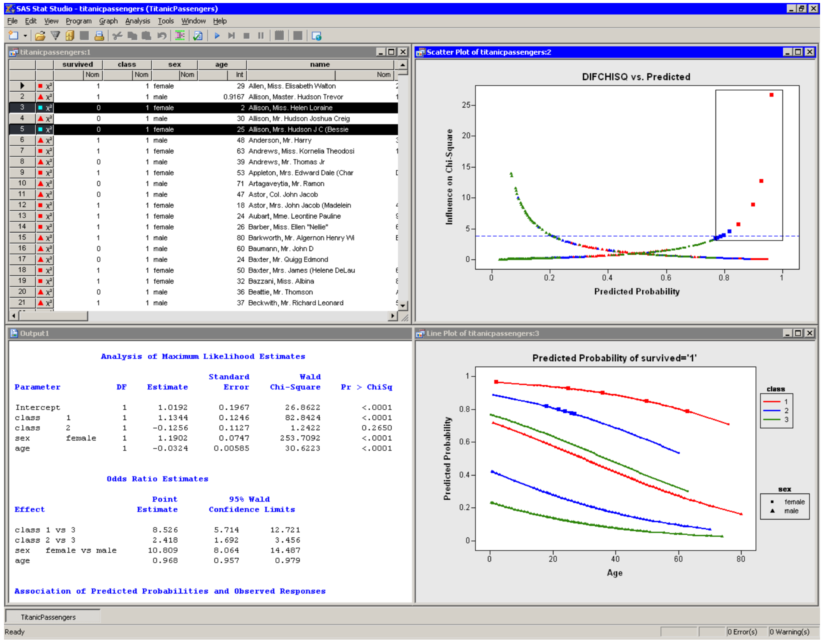
# Best Practices

Fairly Easy Things To Improve Your Analyses and Visualizations

# Favor Text Over GUI

| GUI | Text |
|---|---|
| □ Easy to use | □ Harder to use |
| □ Software Dependent | □ Less software dependent |
| □ Difficult to share | □ Easier to share |
| □ Difficult to modify | □ Easier to modify |
| □ Difficult to version | □ Easy to version |

# The Problem With GUIs


SAS Stat Studio - titanicpassengers (TitanicPassengers)
File Edit View Program Graph Analysis Tools Window Help
titanicpassengers:1
Scatter Plot of titanicpassengers:2
DIFCHISQ vs. Predicted
Influence on Chi-Square
Predicted Probability
0
5
10
15
20
25
0.2
0.4
0.6
0.8
1
Output1
Line Plot of titanicpassengers:3
Predicted Probability of survived='1'
Predicted Probability
Age
0
20
40
60
80
class
1
2
3
sex
female
male
TitanicPassengers
Ready
0 Error(s)
0 Warning(s)

| | | survived | class | sex | age | name |
|---|---|---|---|---|---|---|
| | | Nom | Nom | Nom | Int | Nom |
| | ■ χ² | 1 | 1 | female | 29 | Allen, Miss. Elisabeth Walton |
| 2 | ▲ χ² | 1 | 1 | male | 0.9167 | Allison, Master. Hudson Trevor |
| 3 | ■ χ² | 0 | 1 | female | 2 | Allison, Miss. Helen Loraine |
| 4 | ▲ χ² | 0 | 1 | male | 30 | Allison, Mr. Hudson Joshua Creig |
| 5 | ■ χ² | 0 | 1 | female | 25 | Allison, Mrs. Hudson J C (Bessie |
| 6 | ▲ χ² | 1 | 1 | male | 48 | Anderson, Mr. Harry |
| 7 | ■ χ² | 1 | 1 | female | 63 | Andrews, Miss. Kornelia Theodosi |
| 8 | ▲ χ² | 0 | 1 | male | 39 | Andrews, Mr. Thomas Jr |
| 9 | ■ χ² | 1 | 1 | female | 53 | Appleton, Mrs. Edward Dale (Char |
| 10 | ▲ χ² | 0 | 1 | male | 71 | Artagaveytia, Mr. Ramon |
| 11 | ▲ χ² | 0 | 1 | male | 47 | Astor, Col. John Jacob |
| 12 | ■ χ² | 1 | 1 | female | 18 | Astor, Mrs. John Jacob (Madelein |
| 13 | ■ χ² | 1 | 1 | female | 24 | Aubart, Mme. Leontine Pauline |
| 14 | ■ χ² | 1 | 1 | female | 26 | Barber, Miss. Ellen "Nellie" |
| 15 | ▲ χ² | 1 | 1 | male | 80 | Barkworth, Mr. Algernon Henry Wi |
| 16 | ▲ χ² | 0 | 1 | male | 60 | Baumann, Mr. John D |
| 17 | ▲ χ² | 0 | 1 | male | 24 | Baxter, Mr. Quigg Edmond |
| 18 | ■ χ² | 1 | 1 | female | 50 | Baxter, Mrs. James (Helene DeLau |
| 19 | ■ χ² | 1 | 1 | female | 32 | Bazzani, Miss. Albina |
| 20 | ▲ χ² | 0 | 1 | male | 36 | Beattie, Mr. Thomson |
| 21 | ▲ χ² | 1 | 1 | male | 37 | Beckwith, Mr. Richard Leonard |

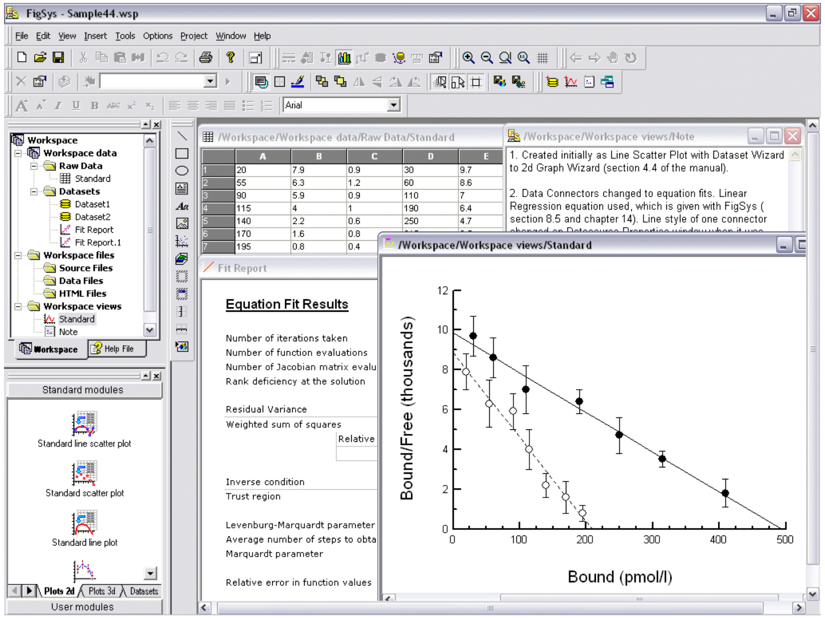
```
            Analysis of Maximum Likelihood Estimates

                                   Standard        Wald
Parameter          DF   Estimate      Error  Chi-Square   Pr > ChiSq

Intercept           1     1.0192     0.1967     26.8622       <.0001
class     1         1     1.1344     0.1246     82.8424       <.0001
class     2         1    -0.1256     0.1127      1.2422       0.2650
sex       female    1     1.1902     0.0747    253.7092       <.0001
age                 1    -0.0324    0.00585     30.6223       <.0001


                 Odds Ratio Estimates

                             Point          95% Wald
Effect                    Estimate      Confidence Limits

class 1 vs 3                 8.526       5.714     12.721
class 2 vs 3                 2.418       1.692      3.456
sex   female vs male        10.809       8.064     14.487
age                          0.968       0.957      0.979


Association of Predicted Probabilities and Observed Responses
```

- □ Too easy to hide assumptions.
- □ Difficult to document what you have done.
- □ Difficult to share.

http://support.sas.com/documentation/cdl/en/imlsstat/63827/HTML/default/viewer.htm#imlsstat_statintro_sect001.htm

# If You Use a GUI

- □ If you can, download the code.
- □ Document all settings.
- □ Save the file, open and rerun.
- □ Screenshots.



/Workspace/Workspace data/Raw Data/Standard

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 20 | 7.9 | 0.9 | 30 | 9.7 |
| 2 | 55 | 6.3 | 1.2 | 60 | 8.6 |
| 3 | 90 | 5.9 | 0.9 | 110 | 7 |
| 4 | 115 | 4 | 1 | 190 | 6.4 |
| 5 | 140 | 2.2 | 0.6 | 250 | 4.7 |
| 6 | 170 | 1.6 | 0.8 | | |
| 7 | 195 | 0.8 | 0.4 | | |

/Workspace/Workspace views/Note

1. Created initially as Line Scatter Plot with Dataset Wizard to 2d Graph Wizard (section 4.4 of the manual).

2. Data Connectors changed to equation fits. Linear Regression equation used, which is given with FigSys ( section 8.5 and chapter 14). Line style of one connector

Fit Report

Equation Fit Results

Number of iterations taken
Number of function evaluations
Number of Jacobian matrix evalu
Rank deficiency at the solution

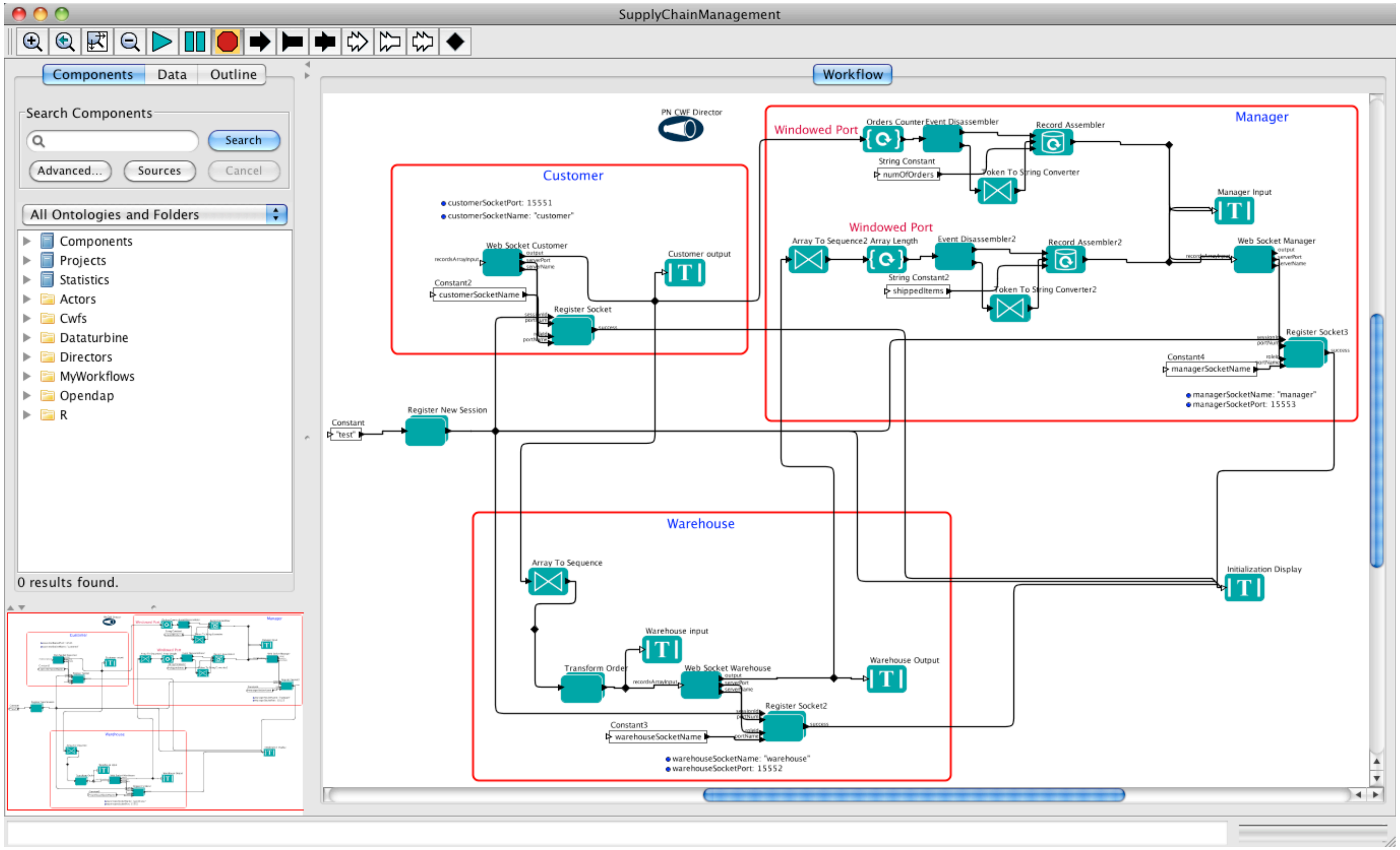
Residual Variance
Weighted sum of squares
Relative

Inverse condition
Trust region

Levenburg-Marquardt parameter
Average number of steps to obta
Marquardt parameter

Relative error in function values

# Scientific Workflow Systems



- customerSocketPort: 15551
- customerSocketName: "customer"

- managerSocketName: "manager"
- managerSocketPort: 15553

- warehouseSocketName: "warehouse"
- warehouseSocketPort: 15552

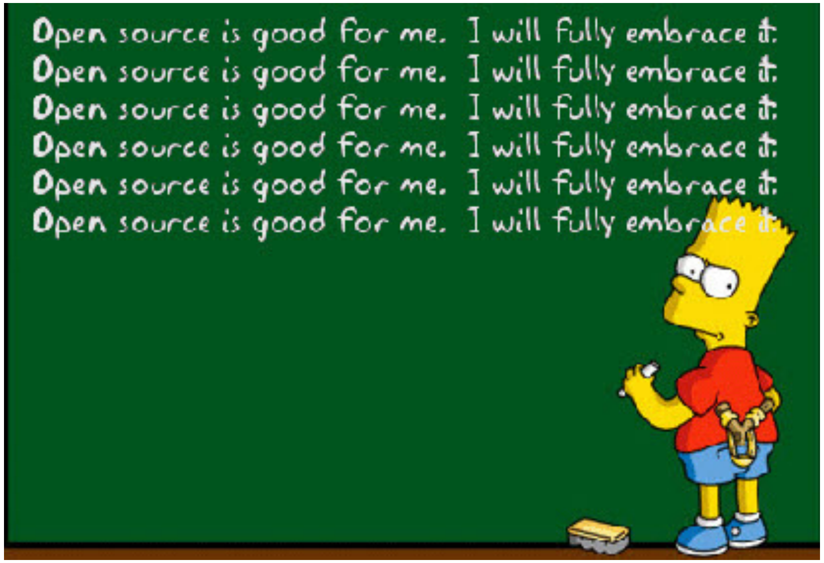
0 results found.

http://db.cs.pitt.edu/group/projects/confluence

# Analysis and Visualization as Software

Borrowing from AGILE development and other current coding best practices.

# Reuse Code and Share Code


Open source is good for me. I will fully embrace it.
Open source is good for me. I will fully embrace it.
Open source is good for me. I will fully embrace it.
Open source is good for me. I will fully embrace it.
Open source is good for me. I will fully embrace it.
Open source is good for me. I will fully embrace it.

- Speeds development
- Reduces errors
- Increases transparency
- Promotes Collaboration

# Test Frequently

- Write tests for your code and make them available

- Try to test each line of code

# Clean Code

- □ Be very organized
- □ Use descriptive variable and function names
- □ Keep code blocks small
- □ A non-program should be able to read and understand it.
- □ Comment but don't over comment your code.

Hey, Carl, can you look at this problem with me. I've been working on this for hours. You see the X variable clearly cannot be less than zero because Y has to be more than 20.... Oh wait. That's not right. OK, I've got it now. Thanks, Carl!

# Really Bad

```
function fetch(i) { int j=get_it(i);
  if(j>1)
  { do_something_important(i); } else
  { do_something_important(j); } }
  function go(i)
  { calculate_something(i); sleep(i);
  think_about_something(i); } function
  go_fetch(id) { int i=id; i++; go(i);
  fetch(i): return i; } function main()
  { int i=0; i=get_i(); go_fetch(i); }
```

# Getting Better

```
function fetch(i) {
    int j=get_it(i);
    if(j>1) {
            do_something_important(i);
    }
    else {
            do_something_important(j);
    }
}

function go(i) {
    calculate_something(i);
    sleep(i);
    think_about_something(i);
}
```

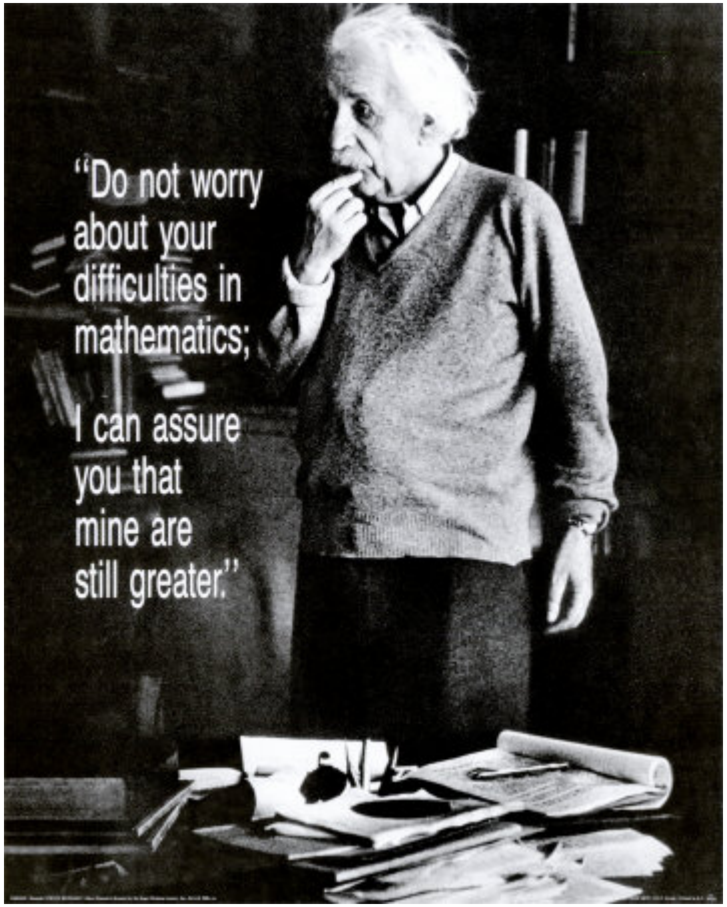
```
function go_fetch(id) {
    int i=id;
    i++;
    go(i);
    fetch(i);
    return i;
}

function main() {
    int i=0;
    i=get_i();
    go_fetch(i);
}
```

# Even Better

```
function main() {
   int i=0,j;
   i=get_i();
   i++;
   calculate_something(i);
   sleep(i);
   think_about_something(i);
   j=get_it(i);
   if(j>1) {

         do_something_important(i); }
   else {
         do_something_important(j);
   }
}
```

- The variables are still poorly named.

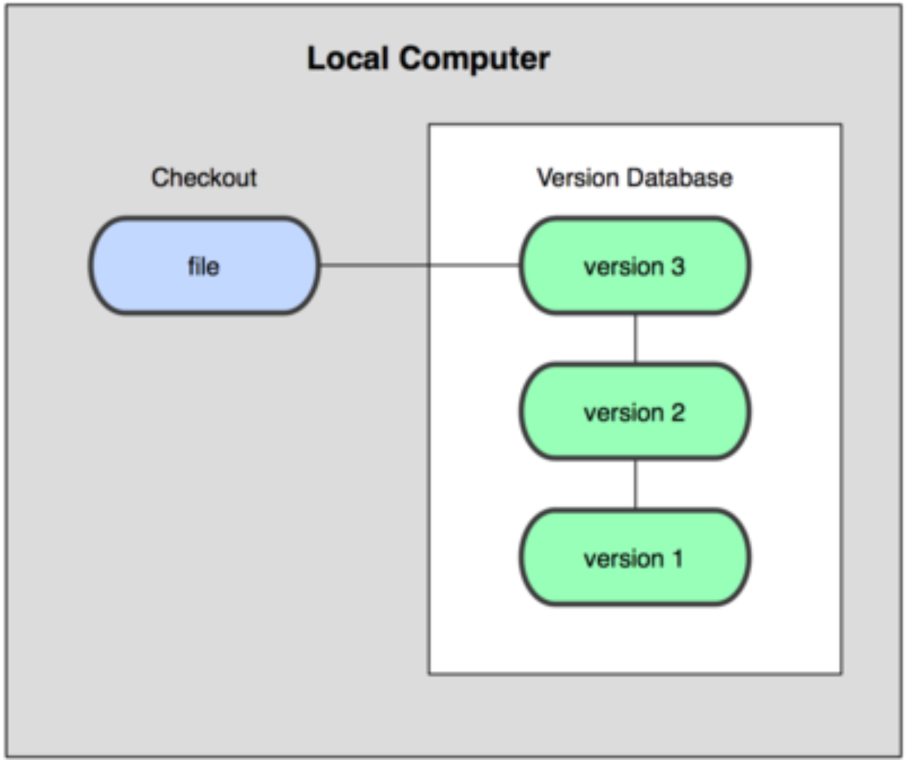
# Why Worry About The Code?



- □ You will understand it later.
- □ Others will understand it.
- □ Increased reuse
- □ Decreased errors and bugs.

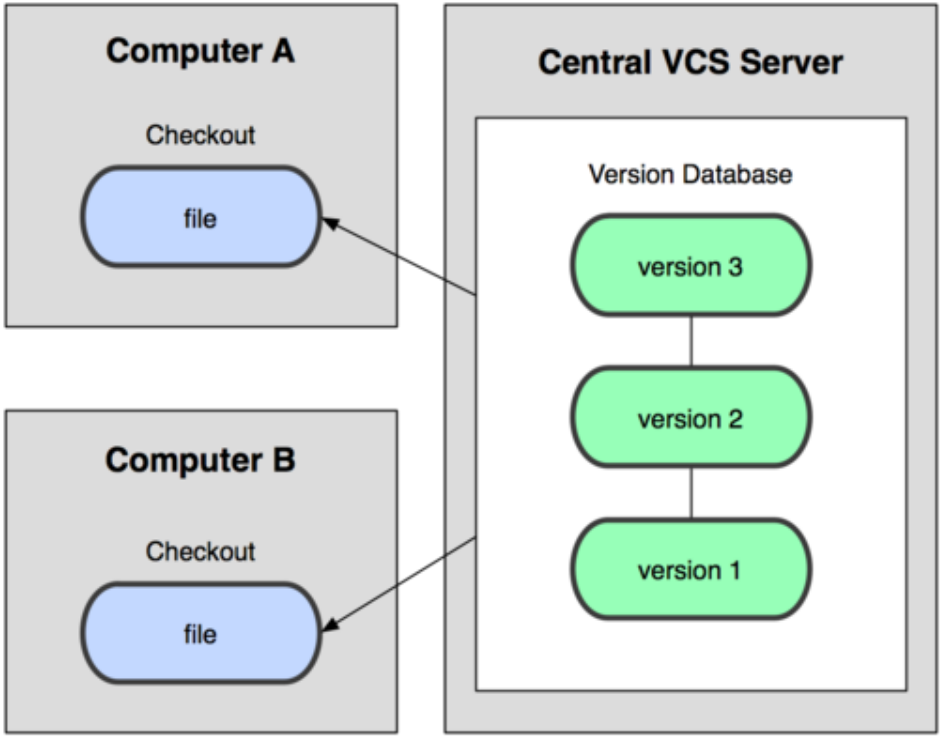
# Versioning

- Developed for versioning code.
- You can version anything.
- Works best on text based objects.

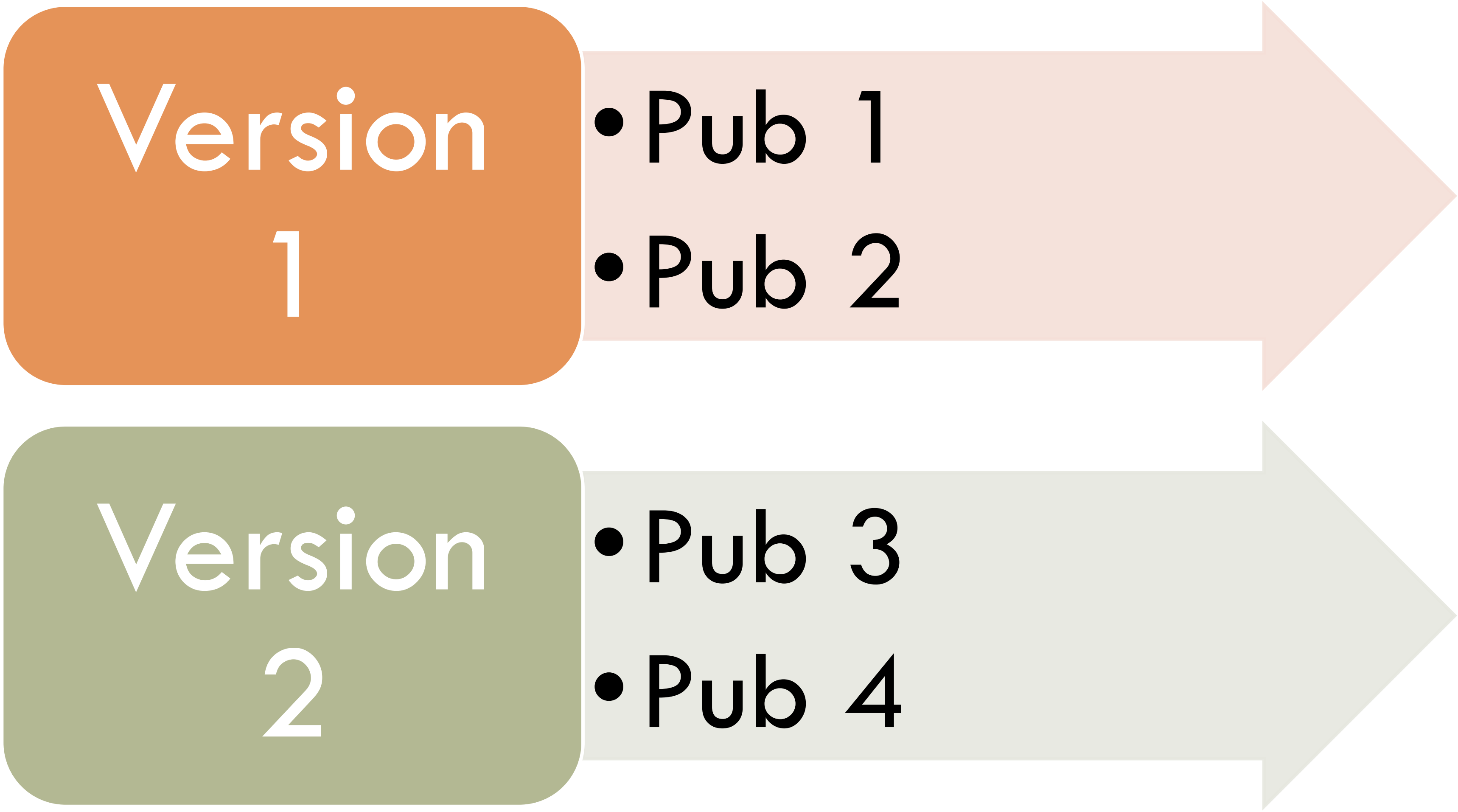


# Versioning

- Allows you to roll back changes.
- Collaborate on files and merge changes.
- Branch off modifications.



# A Data Specific Use Case



Version 1

- Pub 1
- Pub 2

Version 2

- Pub 3
- Pub 4

# Dissemination

NSF Requires You To Be Open Source

# Where To Publish Your Analyses

- □ Open Source Repositories
- □ Statistical Repositories
- □ Institutional Repositories
- □ Personal Web Sites

# How To Publish

□ Provide Adequate and Accurate Metadata

□ Provide a License

□ Use Social Networking

□ Provide Links In Publications

# Licensing – Most Common Licenses

- GNU General Public License
- Apache Software License
- Creative Commons
  - The most adaptable.

# Final Thoughts

- You are required to make them available
- You should make them available
- Reuse and Share
- Treat your analyses like data
- License your work

# Questions?

- The slides are available at https://github.com/olendorf/presentations/tree/master/ciday-2012