The Foundations of Digital Research

Software Sustainability Institute
N.ChueHong@software.ac.uk
Neil Chue Hong (@npch)

Re-usable
Re-producible

www.software.ac.uk
software.evaluation-guide
resources/guides
software-carpentry
training

www.software.ac.uk/blog/
2012-11-09-craftsperson-and-scholar

www.software.ac.uk/blog/2012-08-16-what-research-
software-community-and-why-should-you-care

www.software.ac.uk/blog/2011-05-02-
publish-or-be-damned-alternative-
impact-manifesto-research-software

slideshare.net/npch/the-
foundations-of-digital-research

CC BY NC
Isn’t software just data?

- Journal of Open Research Software
  - [http://openresearchsoftware.metajnl.com/](http://openresearchsoftware.metajnl.com/)
- Role of Repositories
  - [http://www.era.lib.ed.ac.uk/handle/1842/5905](http://www.era.lib.ed.ac.uk/handle/1842/5905)
- Publication for Discovery
  - JISC-funded Software Hub project

Why do we version?
- To indicate a change
- To allow sharing
- To confer special status

And we haven’t talked about authorship yet…
5 Stars of Software?

• Do we need a 5 stars for software?
  ▪ Existence – there is accurate metadata that defines the software
  ▪ Availability – you can access and run the software
  ▪ Openness – the software has an open permissible license
  ▪ Linked – the related data, dependencies and papers are indicated
  ▪ Assured – the software provides ways of assuring its “correctness”

C.f.
5 Stars of Linked Data (Berners-Lee)
5 Stars of Online Journals (Shotton)
NOAH CLEMONS
HOW TO ENFORCE REPRODUCIBILITY WITH YOUR EXISTING MKL CODE
How to Enforce Reproducibility with your Existing Intel® Math Kernel Library Code

Noah Clemons
Technical Consulting Engineer
Intel® Developer Products Division
3 Types of Non-Reproducibility in Intel® Math Kernel Library

- Run to Run – same processor
- Runs between different Intel processors
- Runs between different IA-compatible processors

<table>
<thead>
<tr>
<th>Maximum Compatibility</th>
<th>Function Call mkl_cbwr_set( ... )</th>
<th>Environment Variable MKL_CBWR=</th>
</tr>
</thead>
<tbody>
<tr>
<td>on Intel® or Intel®-compatible CPUs supporting SSE2 instructions or later</td>
<td>MKL_CBWR_COMPATIBLE</td>
<td>COMPATIBLE</td>
</tr>
<tr>
<td>on Intel® processors supporting SSE2 instructions or later</td>
<td>MKL_CBWR_SSE2</td>
<td>SSE2</td>
</tr>
<tr>
<td>on Intel processors supporting SSE4.2 instructions or later</td>
<td>MKL_CBWR_SSE4_2</td>
<td>SSE4_2</td>
</tr>
<tr>
<td>on Intel processors supporting Intel® AVX or later</td>
<td>MKL_CBWR_AVX</td>
<td>AVX</td>
</tr>
<tr>
<td>from run to run (but not processor-to-processor)</td>
<td>MKL_CBWR_AUTO</td>
<td>AUTO</td>
</tr>
</tbody>
</table>
Example – Find out the best performing option from a pool of processors

For the best option given a pool of computing resources in a grid setting, you may launch a simple program as follows

```c
#include <mkl.h>

int main(void) {
    int my_cbwr_branch;
    /* Find the available MKL_CBWR_BRANCH */
    my_cbwr_branch = mkl_cbwr_get_auto_branch();
    if (!mkl_cbwr_set(my_cbwr_branch)) {
        printf("Error in setting branch. Aborting...
");
        return;
    }
    return my_cbwr_branch;
}
```

Examine all results and use `mkl_cbwr_set(<minimum_result>)`

The full list of options:

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPATIBLE</td>
<td>3</td>
</tr>
<tr>
<td>SSE2</td>
<td>4</td>
</tr>
<tr>
<td>SSE3</td>
<td>5</td>
</tr>
<tr>
<td>SSSE3</td>
<td>6</td>
</tr>
<tr>
<td>SSE4_1</td>
<td>7</td>
</tr>
<tr>
<td>SSE4_2</td>
<td>8</td>
</tr>
<tr>
<td>AVX</td>
<td>9</td>
</tr>
<tr>
<td>AVX2</td>
<td>10</td>
</tr>
</tbody>
</table>

For more information on Benchmarks and Optimization Notice go to [http://www.intel.com/performance](http://www.intel.com/performance)
Lepton is a tool to do research as opposed to publishing reproducible research results. It deals with:

- **everyday tasks** such as programming and writing technical reports
- **reviewing** the methods and results by collaborators and in the long term
- **re-using** source code, input data, research results

Further references:


- 2 conference papers

- A poster + demonstration at ICERM
Lepton provides:

- provenance information
  - generated documents contain all the information required to reproduce the results
- executable papers
  - a Lepton file is a program and can be executed on the local machine
- coherence and correctness guarantees
  - Lepton executes commands and automatically embeds their output
  - no copy-and-paste
- literate programming features
  - everything in the same bundle,
  - related items placed in close proximity
  - meaningful chunks
- generated, up-to-date documentation
  - run benchmarks with scripts in any language
  - format the results into tables
Lepton: In practice

Image analysis

Writing source code

Code chunk 1: «header»

```c
bp_typedef
#include <stdio.h>
#include <stdlib.h>
<<bp_typedef>>
```

Code chunk 2: «bp_typedef»

```c
struct bp_node
{
    gsl_vector * belief;
    gsl_vector * evidence;
    gsl_vector * m_left;
    gsl_vector * m_right;
    gsl_vector * m_up;
    gsl_vector * m_down;
};
```

Teaching statistics

Code chunk 3: «R»

```r
x = round(runif(5,0,10),1)
cat(x, "\n")
```

Interpret with R

```
1.7 6.2 4.5 7.3 0.8
```

Define the sample mean:

\[
\bar{x} = \frac{\sum x_i}{n}
\]

Example:

\[
\bar{x} = \frac{(1.7 + 6.2 + 4.5 + 7.3 + 0.8)}{5}
\]
NICOLAS LIMARE
MY CHRISTMAS LIST FOR REPRODUCIBILITY
My Christmas List for Reproducibility

Nicolas Limare

CMLA, ENS Cachan, France

2012-12-12
Tools and Infrastructure

- **software identifier**
  DOI-like system for software, vendor-independent
  always pointing to the current location of the code

- **open data storage**
  until every journal stores supp. materials
  more reliable than home page storage
Standards

- **cross-library APIs**
  we need to be able to replace one broken library by another
  → need an interface and spec common to multiple implementations

- **programming language specs**
  prog. lang. defined by the current implementation will eventually break
  → need a formal and stable spec of the language

- **software quality test tools and services**
  guide code authors to improvements
  reduce the workload for referees
Copyright and Patents

- **paper vs. software**
  - paper for human, software for computer
  - → should have the same copyright status

- **no patent restriction for research**
  - patent system is supposed to stimulate innovation
  - implementation is the translation of information already public in the patent application
  - → should not be prevented from releasing the code
BENJAMIN SEIBOLD

SITMAP: A SECOND ORDER STAGGERED GRID METHOD FOR SPHERICAL HARMONICS

MOMENT EQUATIONS OF RADIATIVE TRANSFER
Some non-Hermitian Hamiltonians, which surprisingly have quantum systems; after all, Galois conjugation CFT (Yang-Lee CFT, which is Galois conjugate to the Fibonacci ones, and vice versa. One famous example is the nonunitary convert nonunitary conformal field theories (CFTs) to unitary states. The first such proposal came in the form of the can describe one-dimensional (1D) edge states of certain interest, is the question as to whether nonunitary models have a venerable history. Galois conjugation, by definition, replaces a root of a $\sqrt{3}$ = $\sqrt{3} + \sqrt{3}$, as are $\sqrt{2}$, $\sqrt{2}$, as well as $\sqrt{5}$, $\sqrt{5}$, $\sqrt{5}$, $\sqrt{5} - \sqrt{5}$, and $\sqrt{5} - \sqrt{5}$. In statistical mechanics, nonunitary conformal field theories $\frac{1}{2}$ = $\frac{1}{2}$ - $\frac{1}{2}$, and might be seen to open the PT $\frac{1}{2}$ - $\frac{1}{2}$, and even (integer spin) subset (2). One consequence is that these nonunitary TQFTs do not describe physical realizations of topological effects. This model, which is also called “DFib,” is a topological quantum field theory (TQFT) which includes, but is not limited to, gapped fractional quantum Hall (FQH) state. However, it turns out that the latter does not describe a gapped topological phase. A more recent proposal is the “Gaffnian” wave function proposed to describe a gapped edge. Although with a nonunitary “Yang-Lee” CFT describing its code property and topological protection of the degenerate ground states. Beyond this, we rigorously prove that no local change of basis can transform the ground states of the Galois-conjugated doubled Fibonacci theory into the ground states of a topological model whose Hermitian Hamiltonian satisfies Lieb-Robinson bounds. We conclude here that this is not possible, further restricting the possible scope of nonunitary models in quantum mechanics. In lattice theories, the LR bounds provide a similar upper bound. There can be some exponentially small effects. In relativistically invariant field theories, the speed of light is set by microscopic details of the Hamiltonian, such as the interaction strength and range. If it...
Numerical experiments + theorem and proof

- Can we build quantum computers based on non-unitary conformal field theories?
- First reproducible numerical experiment, then theorem and proof.

PHYSICAL REVIEW B 85, 045414 (2012)

Galois conjugates of topological phases

M. H. Freedman, 1 J. Gukelberger, 2 M. B. Hastings, 1 S. Trebst, 1 M. Troyer, 2 and Z. Wang 1

1 Microsoft Research, Station Q, University of California, Santa Barbara, California 93106, USA
2 Theoretische Physik, ETH Zurich, CH-8093 Zurich, Switzerland

Theorem IV.5. Fixing the number $n \geq 5$ and particle type $\tau \otimes \tau$ of DFib anyons on $S^2$ and any vertex normalization $f$, there can be no continuous uniform $\Gamma$ family of (g.s. weakly) local normalizer operators $O_\Gamma: \mathcal{H} \rightarrow \mathcal{H}$, so that $O_\Gamma G_{n,\Gamma,f}^G$ is, for all anyon positions $\Gamma$, the ground-state manifold of a uniformly Lieb-Robinson and uniformly gapped family of Hermitian Hamiltonians $H(\Gamma)$ defining a topological phase [see Eq. (1)].

FIG. 6. (Color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension $(\theta \neq 0)$. This figure can be reproduced using the VisTrails 33 workflow Fig. 6 included in the Supplementary Material.
An executable paper: see laptop demo tonight

- The arXiv version has all data and workflow
- Clicking on the figure downloads the VisTrails workflow that reproduces the figure.
An executable paper: see laptop demo tonight

- The arXiv version has all data and workflow
- Clicking on the figure downloads the VisTrails workflow that reproduces the figure.

Galois Conjugates of Topological States

Michael H. Freedman, Jan Gukelberger, l. Troyer, Zhenghan Wang

(Submitted on 16 Jun 2011 (v1), last revised 5 Jun 2011)

Galois conjugation relates unitary conformal field theories (CFTs) to their no-go Galois conjugates of quantum double models. Galois conjugated Hamiltonians are typical of quantum double models. Specific attempts at constructing these states lead to a loss of the code preserving degenerate ground states. Beyond this, we can transform the ground states of the Galois conjugated Hamiltonian to the ground states of a topological model with Robinson bounds. These include all gapless states for many other non-unitary "Gaffnian" wave functions cannot be the gapped state.

FIG. 6. (color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$).
An executable paper: see laptop demo tonight

- The arXiv version has all data and workflow
- Clicking on the figure downloads the VisTrails workflow that reproduces the figure.
An executable paper: see laptop demo tonight

- The arXiv version has all data and workflow
- Clicking on the figure downloads the VisTrails workflow that reproduces the figure.
Publishers were excited!

- This is how it should be!
- Start a trial project to see how it can be made to work!
Publishers were excited!

- This is how it should be!
- Start a trial project to see how it can be made to work!

- But they soon gave up ....
Publishing requires compromises (1)

- No stable URL or DOI for supplementary material
- No link from the figure, but only a reference

FIG. 6. (Color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$). This figure can be reproduced using the VisTrails\textsuperscript{33} workflow Fig. 6 included in the Supplementary Material.\textsuperscript{37}
Publishing requires compromises (1)

- No stable URL or DOI for supplementary material
- No link from the figure, but only a reference
  - reference goes to supplementary material page

FIG. 6. (Color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$). This figure can be reproduced using the VisTrails$^{33}$ workflow Fig. 6 included in the Supplementary Material.$^{37}$

Publishing requires compromises (1)

- No stable URL or DOI for supplementary material
- No link from the figure, but only a reference
  - reference goes to supplementary material page
  - find the right workflow, then download the workflow

![Figure 6](image_url)

**FIG. 6.** (Color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$). This figure can be reproduced using the VisTrails\textsuperscript{33} workflow Fig. 6 included in the Supplementary Material.\textsuperscript{37}

\textsuperscript{37}See Supplemental Material at \url{http://link.aps.org/supplemental/10.1103/PhysRevB.85.045414} for full provenance information and workflows to recreate the figures.
Publishing requires compromises (1)

- No stable URL or DOI for supplementary material
- No link from the figure, but only a reference
  - reference goes to supplementary material page
  - find the right workflow, then download the workflow

FIG. 6. (Color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$). This figure can be reproduced using the VisTrails\textsuperscript{33} workflow Fig. 6 included in the Supplementary Material.\textsuperscript{37}

Publishing requires compromises (2)

- The workflow needs to fetch the raw data, but
  - No stable URL or DOI for supplementary material
  - Even unstable URL only know after publication

- How did we solve it?
Publishing requires compromises (2)

- The workflow needs to fetch the raw data, but
  - No stable URL or DOI for supplementary material
  - Even unstable URL only know after publication

- How did we solve it?

- Journal of Statistical Mechanics (JSTAT), an IOP journal
  - Production editor started publication process before the lunch break and sent us the URL
  - We had an hour to prepare final workflows and sent them back
  - He finished publication process after returning from lunch
Publishing requires compromises (2)

- The workflow needs to fetch the raw data, but
  - No stable URL or DOI for supplementary material
  - Even unstable URL only know after publication

- How did we solve it?
  - Journal of Statistical Mechanics (JSTAT), an IOP journal
    - Production editor started publication process before the lunch break and sent us the URL
    - We had an hour to prepare final workflows and sent them back
    - He finished publication process after returning from lunch

- Physical Review, an APS journal
  - Editors told us to give up
  - Production manager informed us that we can replace the supplementary material anytime after publication without leaving a trace ....
  - We then just sent the working workflows with the right URLs for data after publication
Our next approach

- Publishers desire reproducible papers but are not yet ready to handle executable papers in the publication process.

- Our intermediate solution:
  - Publish raw data and workflows through our institutional library and obtain DOIs.
  - Refer to that data from the paper and just include a backup copy with the papers.
\[\text{knitr} = \text{knit} + \text{R}\]

\[\text{Sweave} = \text{S} + \text{weave}\]
my homework & solutions in past three years at Iowa State\textsuperscript{1}

\textsuperscript{1}e.g. https://github.com/yihui/stat579/downloads
I love \LaTeX more than anyone else, but do not tie users to \LaTeX.
I love \LaTeX more than anyone else, but do not tie users to \LaTeX 
(but do keep them away from Word)
# Markdown

## Markdown

**Markdown**

_Markdown_

- markdown
  - markdown
The value of $\pi$ is `r pi`, and a Monte Carlo estimate is:

```r
est_pi = function(n) {
  x = runif(n, -1, 1)
  y = runif(n, -1, 1)
  4 * mean(x^2 + y^2 <= 1)
}
est_pi(10000)
```
The value of $\pi$ is 3.1416, and a Monte Carlo estimate is:

```r
est_pi = function(n) {
  x = runif(n, -1, 1)
  y = runif(n, -1, 1)
  4 * mean(x^2 + y^2 <= 1)
}
est_pi(5000)
```

```r
# [1] 3.128
```
reproducible homework (happier students, happier professors)
evidence that we underestimated the power and imagination of students: http://www.rpubs.com
written in R, but not for R only (bash scripts, C++, ...)
If reproducible homework comes, can reproducible research be far behind?
IN CODE WE TRUST