Uninstall Jenkins

- Linux: sudo apt-get remove jenkins
- Windows: Control Panel->Uninstall Programs
- Mac OS X: sudo /Library/Application Support/Jenkins/Uninstall.command

Johan Seland, André Brodtkorb

Using the Cloud for Reproducible Science

Geilo Winter School 2013

What is the Cloud?

*”When talking about the cloud, it is **mandatory** to use at least one fitting quote”*

-- A. R. Brodtkorb, Ph.D. Trial Lecture, 2010

Computation may someday be organized as a public utility.

1961, J. McCarthy

NETFLIX

amazon.com

MENDELEY

Dropbox

Spotify

Microsoft

Google

flickr

salesforce

onLIVE

facebook

I think there is a world market for maybe five computers

1943, T. J. Watson


INSIDE THIS WEEK: TECHNOLOGY QUARTERLY
The Economist
Ratan Tata's lessons for India
Egypt on the edge
The Big Long: betting on US housing
Putin alone
Abraham Lincoln, management theorist
DECEMBER 1ST - 7TH 2012
Economist.com
Survival of the biggest
The internet's warring giants
Google

Overview

- Introduction to Cloud Computing
  - Core technologies
- Using the cloud for reproducible computing
  - Existing Services
    - Demo of EC2
  - Cloud Storage
- Uncertainties in the long term

Cloud Computing - Core Concepts

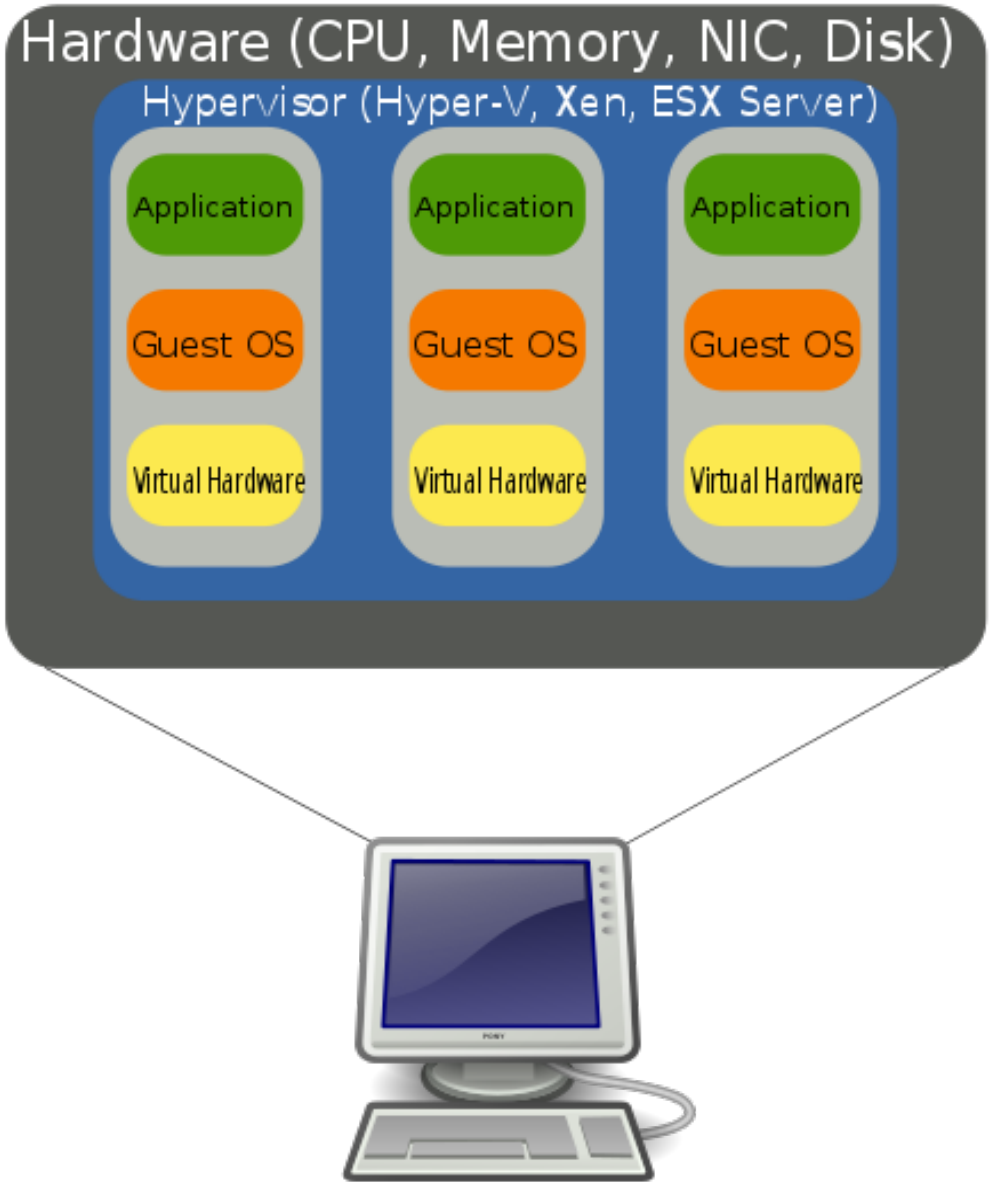
- Enabling Technology: Web 2.0
  - Network protocols: AJAX, SOAP
  - Used in: Google Docs, web applications, etc.
- Enabling Technology: Idling server parks
  - Less than 10% utilization
- Enabling Technology: Virtualization
  - Abstracting hardware resources
  - Allows for virtual machines
- Enabling Technology: Multi-core
  - Multiple virtual machine run on the same physical machine

Virtualization

The simulation of hardware and software upon which other software runs

- A *hypervisor* runs on the host hardware and runs *virtual machines*
- Guest OS with applications are stored in a *virtual machine image*
  - Seems like an ordinary file to the outside world
  - Executed by the hypervisor
- Several virtual machine can run on one physical machine
- Performance is very close to running directly on the hardware

Examples: VirtualBox, VMWare, Parallels , XEN, KVM


Hardware (CPU, Memory, NIC, Disk)
Hypervisor (Hyper-V, Xen, ESX Server)
Application
Application
Application
Guest OS
Guest OS
Guest OS
Virtual Hardware
Virtual Hardware
Virtual Hardware

<everything>-as-a-Service

- SaaS - Software-as-a-Service
  - Google Docs, MS Office 365, GitHub, Spotify
- PaaS - Platform-as-a-Service
  - Microsoft Azure, Google App Engine, Facebook Apps
  - LAPACK-as-a-service?
- IaaS - Infrastructure-as-a-Service
  - Amazon EC2, Rackspace, Google Compute Engine

Services scale as demand increase/decrease

RSaaS – Reproducible-Science-as-a-Service

During the Winter School we have seen several cloud based services to aid in the sharing of ideas and code

- Github, Nbviewer, Shining Panda (Jenkins-as-a-Service)

Other notable servies include:

- Matlab exchange, FigShare

- IPOL, RunMyCode

- Journal for image processing and analysis

- A publication includes
  - Manuscript
  - Software implementation in C/C++
  - An online demo
  - Archive of online experiments

- Co-managed by our own **Nicolas LIMARE**

Runmycode.org

- Standford based startup
  - Backed by Victoria Stodden
- Based on the idea of a *companion website* to your paper
  - Can be anonymous
- Upload code and data
- Code must be: R, Matlab, Fortran or WinRats
  - Without complex dependencies


RESEARCHERS
CODE
USERS
DATA
DATA+CODE
CLOUD
RESULTS
RESULTS
RUNMYCODE
runmycode

Amazon Elastic Compute Cloud (EC2)

- Amazon EC2: Elastic Compute Cloud
  - Buy virtual machines just-in-time
    - Many hardware specs to choose from
  - Standardized Service Level Agreements
  - Used by Dropbox, Instagram, NetFlix

- Supply a virtual machine image of your choice
  - Lots of standard images
  - Machine goes online in about a minute

- Supply input
  - SSH, Web-server, bootup script etc.

- **REMEMBER TO TERMINATE WHEN DONE!**

Amazon Pricing Model

- On-demand
  - 20 machines for one hour = one machine for 20 hours
  - Nice for parameter studies…
- Spot pricing
  - I want 50 nodes, but only if they cost less than $0.50/hours
- Reserved instances
- There is a free tier
  - 5GiB Storage, 750 hours/month of micro instances

Amazon Example Pricing

Instance Type	Price / hour
Micro (613MiB, 1 core, light-use)	$0.020
Large (15 GiB, 4 cores)	$0.340
High-Memory Quadruple Extra Large (68.4 GiB, 8 cores)	$2.024
Cluster Compute (60 GiB, 16 cores, fast interconnects)	$2.700
Cluster GPU (22 GiB, 8 cores, 2xNvidia Fermi GPUs)	$2.360

Storage Plan (first TeraByte)	Price/ GiB/month
Standard Storage (99.999999999% durability)	$0.095
Reduced Redundancy (99.99% durability)	$0.064
Glacier Storage (Retrieval time of several hours)	$0.011

Based on EU-region as of 23/1/2013

Consequences of the Pricing Model

- Running a small web-server: 80 NOK/month (15 USD)
- Remote workstation during working hours: 1700 NOK/month (320 USD)
- Making 1 TiB of data available for a year: 6300 NOK (1140 USD)
- Running a 64-node 8 core cluster for a week: **161 000 NOK (29 000 USD)**
  - Are you that interested in reproducing those supercomputer runs?

Using EC2 to replicate science

- Just sharing source-code is to fragile
  - Hard to compile
  - Version mismatch for compilers, libraries

- Can replicate *your* environment
  - Compilers and libraries, Datasets, Source code
  - Hardware is available to third parties!

- **DEMO TIME**
  - **The final bowling game**

Making Public Images

- Making images public require a few more steps
  - Image must be configured to allow for key from others
  - Remove trails of personal passwords/keys
  - Add to AMI Store

- Beware of licensing issues!
  - Compilers, Libraries, Data

Cloud Uncertainties

- Who has access to your data?
  - Cloud provider?
  - Subcontractor?
  - Other customers?
  - Law enforcement agencies in another country?
- What happen if cloud provider
  - Goes out of business
  - Is bought by a third party
  - It brought down for political or legal reasons
  - Is blacklisted because of another customer
  - Leaks it's password database
- Other problems
  - Software licenses

Digital Object Identifiers

- URLs are not persistent in the long term
- DOI – character string uniquely identifying an electronic object
  - doi:10.1000/182
- The prefix identifies the *registry*
  - The registry stores metadata
    - Including the URL
- DOIs are resolved through www.doi.org
- Organizations can assign DOIs
  - Journals, Data Publishers, National Storage Infrastructure, Libraries

Long term storage for data

- Currently up to the institutions and individual scientist
  - Smaller datasets can probably stay on University Servers
  - National Storage Initiatives
- Amazon and Google provide free storage for public data sets
- What about simulator output?
  - Who will pay in the long term
- DataCite.org provides DOIs for data
  - But not storage
- Rapid progress is being made
  - BIG DATA is the next pillar of science ;)

National Infrastructure vs the Cloud

- National Infrastructure (NOTUR, NORSTORE)
  - NOTUR is the metacenter for Norwegian University Supercomputers
  - Require applications and planning
    - Deadline for spring allocation is **28 January**
    - Compute time between April and September
  - Standardized configurations
  - Free (from the researchers points of view)
  - Some assistance and support

- The Cloud
  - On-demand
  - Can configure images yourself
  - Has to be paid for by the research project
  - Paid support and documentation

Johan Seland

Licenses for code and data

Geilo Winter School 2013

Overview

- Introduction
- Software Licenses
- Licenses for data
- The ideal and the real world

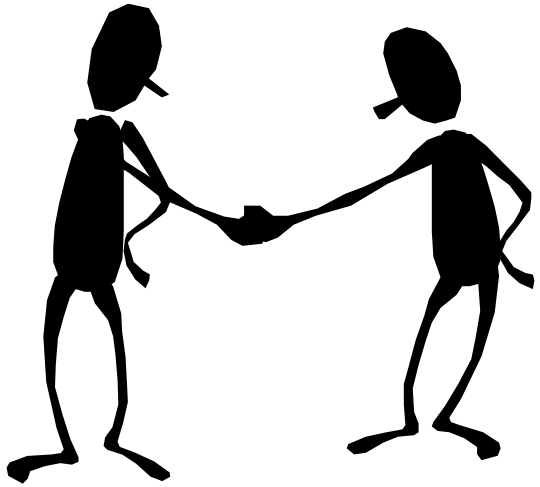
NOT COVERED:

Copyright/License for papers

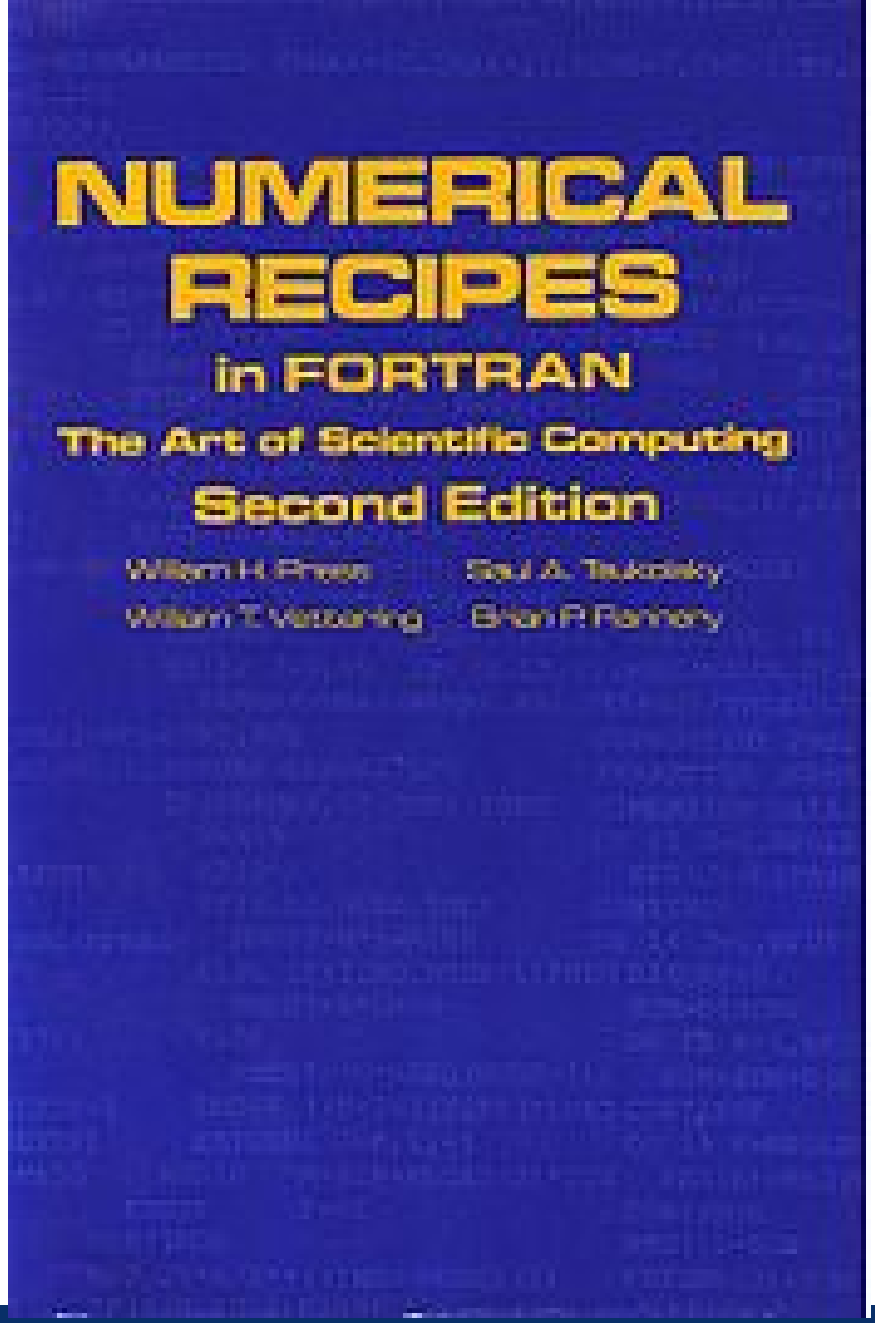
What is a Software License

A **software license** is a legal instrument governing the use or redistribution of software

- It governs
  - What you can do with software from others
  - What others can do with your software




Adobe Flash Player 11.5 Installer
Adobe
I have read and agree to the terms of the Flash Player License Agreement.
Read the license here.
QUIT
INSTALL


NUMERICAL
RECIPES
in FORTRAN
The Art of Scientific Computing
Second Edition
William H. Press
Saul A. Teukolsky
William T. Vetterling
Brian P. Flannery

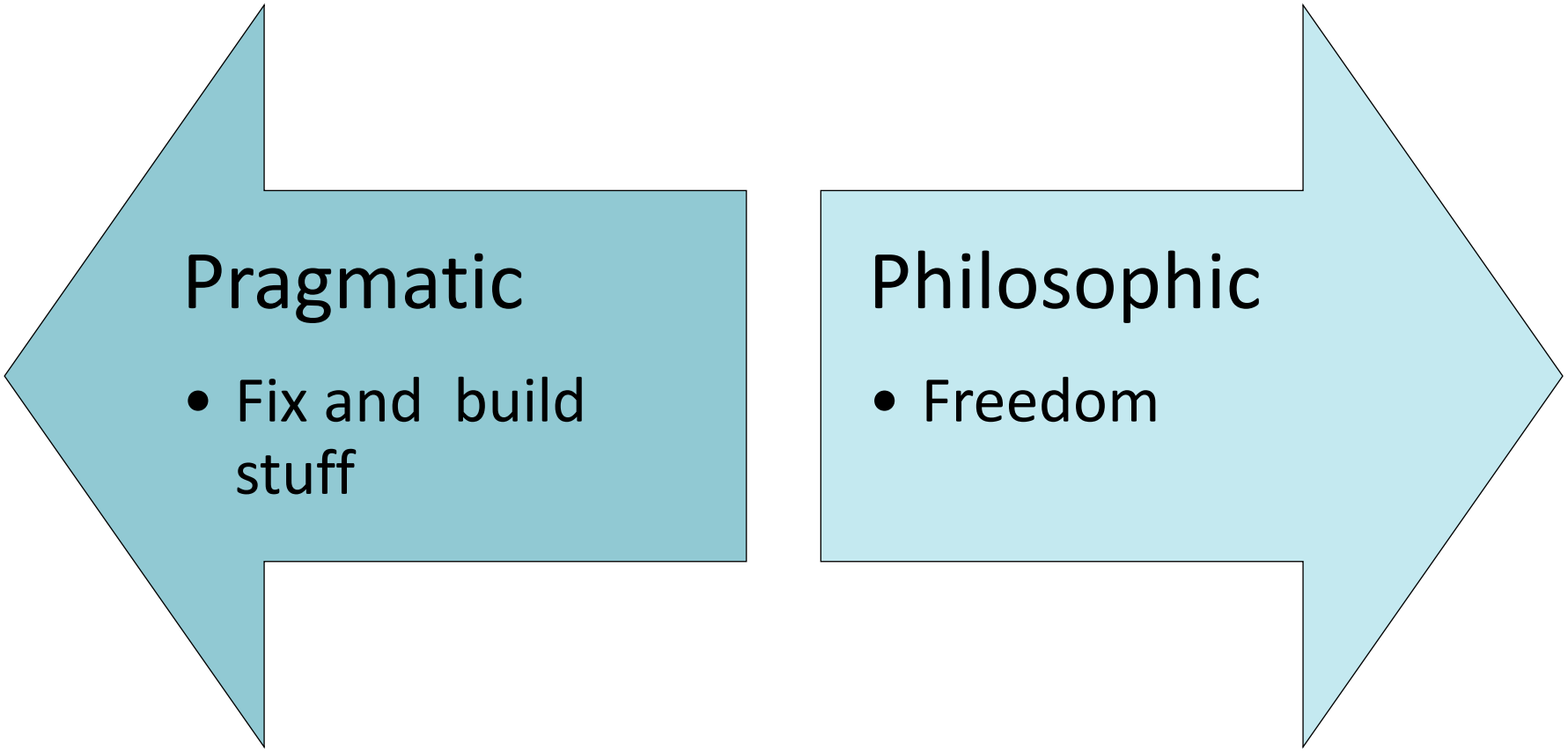
Copyright vs License vs Trademark vs Patent

- Copyright
  - Who own the intellectual property of the software
- License
  - Governs use and redistribution
- Trademark
  - Controls the use of the name (brand) for commercial purposes
- Patent
  - A right to exclude *others* from making, using, selling etc. an invention

Licenses for code and software (not data)

- Public Domain
- Free and Open Source (FOSS)
  - Permissive and Non-Permissive
  - Mostly control *redistribution* and *modification*
- Commercial Licenses
  - Often control *use* of software
  - As well as redistribution
- Hybrid-Licenses
  - Educational/Evaluation

Free and Open Source Software



The Free Software Movement

- Began by Richard Stallman in the 1980s
  - Developed Emacs, GCC

- Released under the GNU General Public License

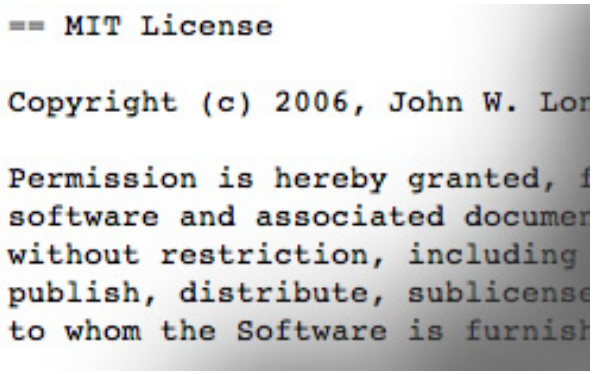
1. if publicly distributed, all software subject to the license must also have its source code released, and
2. once the license is attached to code, it also attaches to any body of code that uses the original code.

It is a *viral* license – if **distributed** your code must be GPL to

Called the **share-alike** provision of the license

Permissive Open Source Licenses

- Licenses that put minimal requirement on redistribution
- Philosophically they foster cooperation above all else
- Many variation, in general
  - Code can be modified
    - Changes need not be made public
  - Code can be used in closed-source products
  - Copyright must be passed on at the source-level



Common Open Source Licenses

Permissive	Weakly Protective	Strongly Protective
MIT, BSD, Apache	LGPL	GPL, Affero GPL
• Can combine with closed source • Do not have to contribute upstream • Apache products, Java Frameworks, VTK, Python Stack	• Can combine into closed-source products • Modifications to LGPL-code must be made public • Often used by libraries (Qt)	• Viral licenses • Your software must be GPL as well. • Affero GPL: Closes ASP-Loophole • Used by GCC, Emacs, R, MySql

Be aware of licenses when using third party software!

Commercial Licenses

Typically Cover:

- How many installs/users
- Revokation
- Education, Non-profit
- Examples from Intel MKL, Matlab
- Can give access to source-code

- End user is generally not able to *redistribute*

Licenses for non-source code

- Data, figures, text etc. are covered by copyright law
- Not designed to facilitate for derivative works

- Creative Commons provide a suite of licenses for sharing of works and information
  - Attribution (CC BY)
  - Attribution Share-Alike (CC BY-SA)
  - Attribution No Derivatives (CC BY-ND)
  - Attribution Non-Commerical (CC BY-NC)
  - Attribution Non-Commerical Share-Alike (CC BY-NC-SA)
  - Attribution Non-Commerical No Derivatives (CC BY-NC-ND)
  - No Rights Reserved (CC0)

Choosing a license

- **You need copyright to choose license!**

- Do you accept contributions from others?
  - Result in shared-copyright
  - You might no be able to choose copyright anymore

- Do you want others to be able to use your code?

The Reproducible Research Standard

Proposal from Victora Stodden (stodden.net)

Realignment of legal rights with scientific norms:

1. Release media components (text, figures) under CC BY
2. Release code-components under MIT License or similar
   - Not GPL
3. Attribution license on selection and arrangement
4. Data Released under CC0

- A well designed, compatible framework of licenses, starting to be picked up
  - Journals try to figure out how to respond

Educational Agreements in Norway

- You have copyright of your thesis and code
- The university has the right to use code for education and research
  - Others can build upon your work
- Master theses can be withheld from publication for three years
- Doctoral theses *must* be publically available
- Often a signed agreement if work is conducted in companies

The Industrial Research Perspective

- Many of you will end up in private research institutes
  - SINTEF, FFI, NR, NGI, IFE, IRIS, NORSAR

- Research Projects mostly funded by the Norwegian and European Research Councils in ***combination*** with industry
  - The industry expects to gain a competitive advantage
  - Data-sets might be confidential

- The research institutes compete for the same grant money
  - The research institutes seek commercial spin-off companies
  - Do not want to give away a competitive advantage

Software Licenses in SINTEF Applied Mathematics

- Several software packages
  - Tinia, HPMC, GoTools, MSRT, OPM
- Dual-Licensed
  - GPL
    - Require copyright sign-off from contributors
  - Commercial license available
- Consequences
  - Little community
  - Results possible to reproduce

Conclusion

- Technology for sharing and reproducing science is in pretty good shape
  - Reproducible supercomputing is a challenge
- The legal licenses are there
  - Institutional policies are not