Version control and collaborative research

SCITAS

École Polytechnique Fédérale de Lausanne

This really happened...

This really happened...

Distributed version control using email, USB keys and CD-ROMS
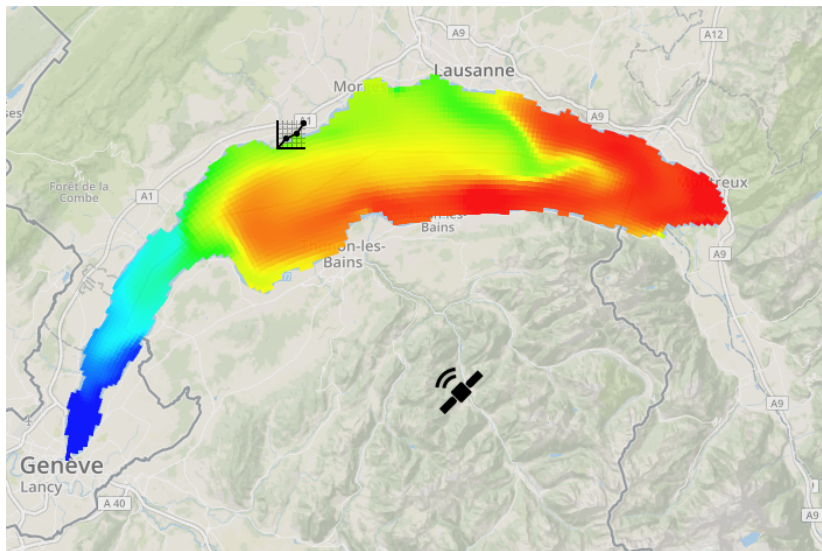
You developed some code and published an article

Now you've been invited to give a presentation
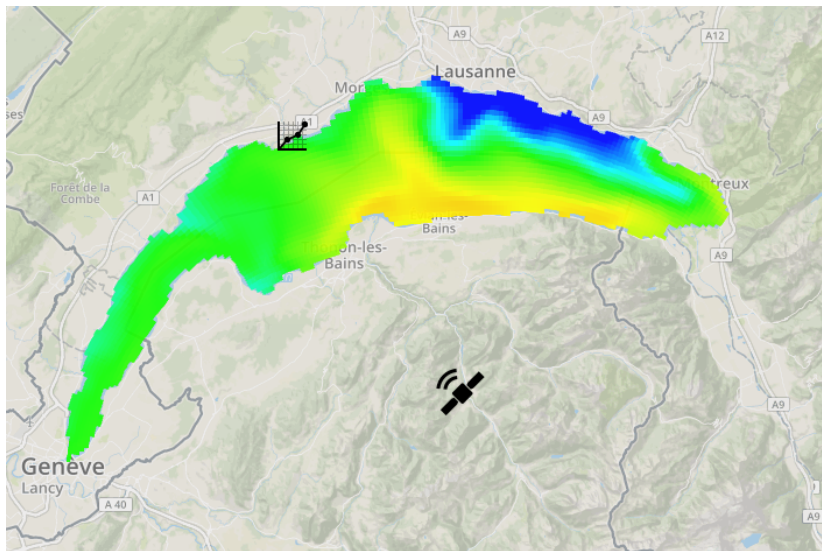
Presentations need pictures

Re-run the analysis to get the images

The image in the article looked like

With the same data it now looks like

This is less than ideal...

Either:

Either:

The article is wrong and should be withdrawn

Either:

The article is wrong and should be withdrawn

Since publication you have added a bug in the code

How can we know?

How can we know?

Version control to the rescue!

How can we know?

Version control to the rescue!

The version used for the article was tagged as version 1.1

git checkout version 1.1

git checkout version 1.1

and relax

Tags

Tags

This point in history is important enough to give it a special name

```
git tag -a v1.1 -m "As used to publish:  Pumpkins for
Biofuels, Smith et al.  2017"
```

```
git tag -a v1.1 -m "As used to publish:  Pumpkins for
Biofuels, Smith et al.  2017"

git tag -a v1.2 -m "As used to publish:  Turnips as a
viable Pumpkin substitute?  Smith et al.  2018"
```

Next time add unit tests

Repositories?

Repositories?

gitlab

Repositories?

gitlab

github

Repositories?

gitlab

github

c4science

Repositories?

gitlab

github

c4science

bitbucket

Is it a bird? Is it a plane?

Is it a VCS in the cloud?

Repositories for VCS

Repositories for VCS

Version control is good

Repositories for VCS

Version control is good

Running a VCS server is less good

Repositories for VCS

Version control is good

Running a VCS server is less good

Use a dedicated service!

Repositories for Collaboration

Repositories for Collaboration

Now things get interesting

1. Download code with FTP

1. Download code with FTP
2. Edit code and create a patch

1. Download code with FTP
2. Edit code and create a patch
3. Send patch to a mailing list

1. Download code with FTP
2. Edit code and create a patch
3. Send patch to a mailing list
4. Await responses from the list

1. Download code with FTP
2. Edit code and create a patch
3. Send patch to a mailing list
4. Await responses from the list
5. The maintainer adds the patch

1. Download code with FTP
2. Edit code and create a patch
3. Send patch to a mailing list
4. Await responses from the list
5. The maintainer adds the patch
6. GOTO 1

Repositories for Collaboration

Repositories for Collaboration

A shared workspace for collaborative development

Repositories for Collaboration

A shared workspace for collaborative development

Much more than a VCS

A social network for coders?

Repositories are now normal

Repositories are now normal

At a conference we suggest an improvement

Repositories are now normal

At a conference we suggest an improvement

Make a pull request!

Altruistic coding?

Altruistic coding?

Anybody can improve or fix things

Need a new feature?

Found a useful tool?

Eigen

Eigen

If you use C++ and matrices you probably want Eigen

Eigen

If you use C++ and matrices you probably want Eigen

Initially one academic developer

Eigen

If you use C++ and matrices you probably want Eigen

Initially one academic developer

Then a second came along

Eigen

If you use C++ and matrices you probably want Eigen

Initially one academic developer

Then a second came along

Now 2 core developers and 150+ contributors

Look at the commit history

Intital commit

```
commit 3036eeca0abfee39d2a0feab4ae3ff1d28975999
Author:  Benoit Jacob <jacob.benoit.1@gmail.com>
Date:   Wed May 30 06:24:51 2007 +0000

Starting Eigen 2 development.  The current plan is to use
the last release of tvmet (inactive for 2 years
and developer unreachable) as the basis for eigen2,
because it provides seemingly good expression template
mechanisms, we want that, and it would take years to
reinvent that wheel.

We'll see.  So this commit imports the last tvmet release.
```

4th commit

```
commit 7947ae0e502d4a71b5db19e582399a17d5647273
Author:  Benoit Jacob <jacob.benoit.1@gmail.com>
Date:  Wed May 30 12:14:09 2007 +0000
```

A TODO to motivate hundreds of hackers to help
with Eigen 2 development.

Cheesy slashdot joke included.

Repositories don't have to be open to everybody

Write code and publish a paper

Write code and publish a paper

Keep developing the code for the next article

Write code and publish a paper

Keep developing the code for the next article

What do you want to make public?

The commit history shows how your code developed

```
commit 1234567890abcdefghijklmnopqrstuvwxyz
Author:  Bob Smith <robert.smith@ethl.ch>
Date:  Wed May 13 17:11:09 2018 +0000

Fix the scaling algorithm - it's been wrong
since the beginning but amazingly no-one
has noticed until now.
```

```
commit 0234567890abcdefghijklmnopqrstuvwxyz
Author:  Bob Smith <robert.smith@ethl.ch>
Date:  Wed May 28 09:32:09 2018 +0000

Initial release of the BioCode package by the
Smith lab at the ETH Lausanne.

www.ethl.ch/smithlab
```

External contributions?

External contributions?

Once open it's difficult to rewrite history

External contributions?

Once open it's difficult to rewrite history

The commit is the acknowledgement!

Public and private?

Public and private?

One repository for the public

Public and private?

One repository for the public

Another for private development

Public and private?

One repository for the public

Another for private development

This looks like gitflow

Recent history

Recent history

Google code

Recent history

Google code

Sourceforge

Google code is no more

Google Code was launched in 2006, with the goal of making it easier to collaborate with others on open-source software. It offered Subversion, Mercurial, and Git source code hosting. It soon expanded to offer wikis, downloads, and even issue tracking.

But alas, in 2015 Google bid farewell to the service, which had since been eclipsed by other, richer open-source project hosting providers.

Sourceforge: from poster-boy to pariah in one easy takeover.

Sourceforge: from poster-boy to pariah in one easy takeover.

1999-2012: The place to be seen

Sourceforge: from poster-boy to pariah in one easy takeover.

1999-2012: The place to be seen

2012-20??: The place to avoid

Near future

Near future

Microsoft

Near future

Microsoft

LinkedIn

Near future

Microsoft

LinkedIn

GitHub

The terms and conditions for free services can and will change.

The terms and conditions for free services can and will change.

You are the product

Don't get trapped

c4science/gitlab/other and mirror to github?

Eigen

Bitbucket and Github

Eigen

Bitbucket and Github

bitbucket.org/eigen/eigen

github.com/eigenteam/eigen-git-mirror

Quantum Espresso

Gitlab and Github

Quantum Espresso

Gitlab and Github

gitlab.com/QEF/q-e

github.com/QEF/q-e

Repositories are not

Repositories are not

data stores

Repositories are not

data stores

libraries

Repositories are not

data stores

libraries

archives

Can we reference a tag from a repository?

In order to estimate the turnip supply needed to replace pumpkins for traditional biofuel applications we use the BioCode package (2)

[2] BioCode v0.3 www.c4science.ch/BioCode/source/tags/v0.3

Einstein, A.

Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt

Zentralbibliothek Zürich

Einstein, A.

Über einen die Erzeugung und Verwandlung des Lichtes betreffenden
heuristischen Gesichtspunkt (1905)

Author

Author

Title

Author

Title

Date

Author

Title

Date

Journal

Author

Title

Date

Journal

Volume

Author

Title

Date

Journal

Volume

Pages

What can we do?

What can we do?

Digital Object Identifier (DOI)

What can we do?

Digital Object Identifier (DOI)

JOSS

What can we do?

Digital Object Identifier (DOI)

JOSS

Zenodo

What can we do?

Digital Object Identifier (DOI)

JOSS

Zenodo

Digital Archives?