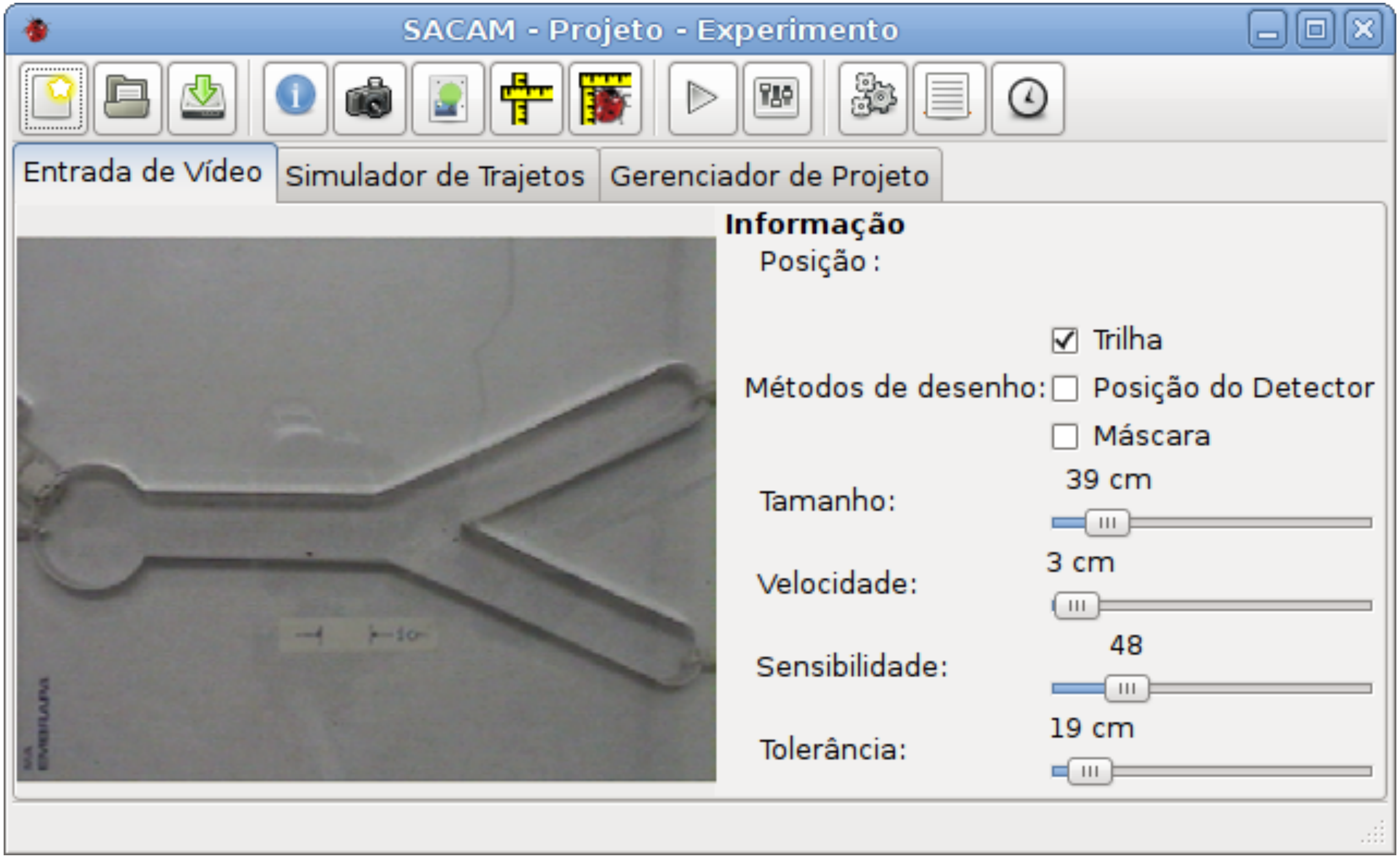
Ciência aberta na prática: Reprodutibilidade, notebooks e tudo mais

Luiz Irber
Michigan State University

POLICE BOX

Meu primeiro projeto

- Efeitos de feromônios em insetos
- Software disponível para Windows
  - portar para Linux
- Borland C++ Builder/Kylix
- Python + GStreamer
- Sourceforge!
- ~2005-2007



Informação

Posição :

Métodos de desenho: ☑ Trilha

□ Posição do Detector

□ Máscara

Tamanho: 39 cm

Velocidade: 3 cm

Sensibilidade: 48

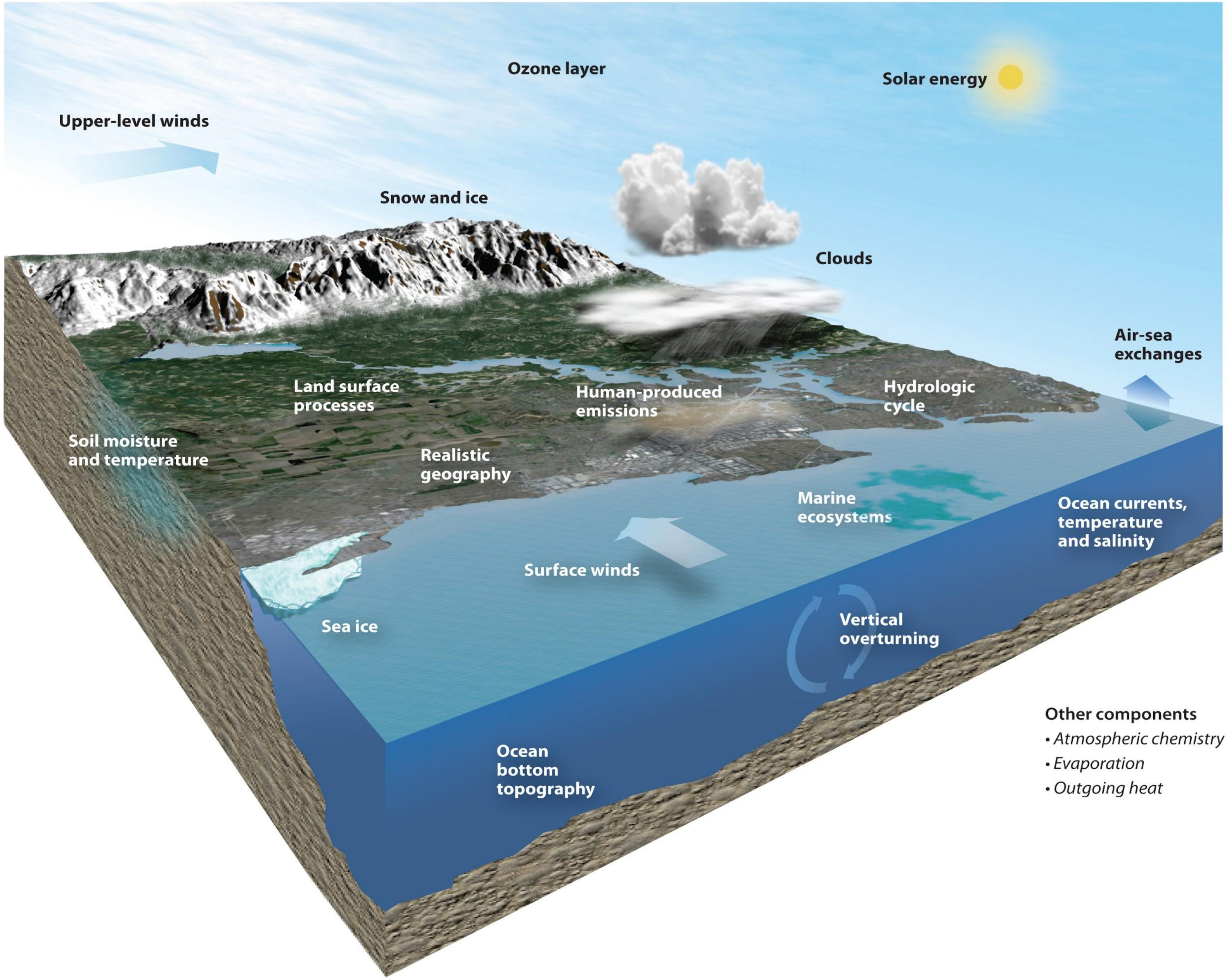
Tolerância: 19 cm

SACAM
https://github.com/luizirber/bugbrother

Python Brasil

- Primeira participação: 2007
- 2010
  - Recém-formado
  - "A gente tem um supercomputador novo e precisamos de engenheiros, tá a fim?"

Tupã



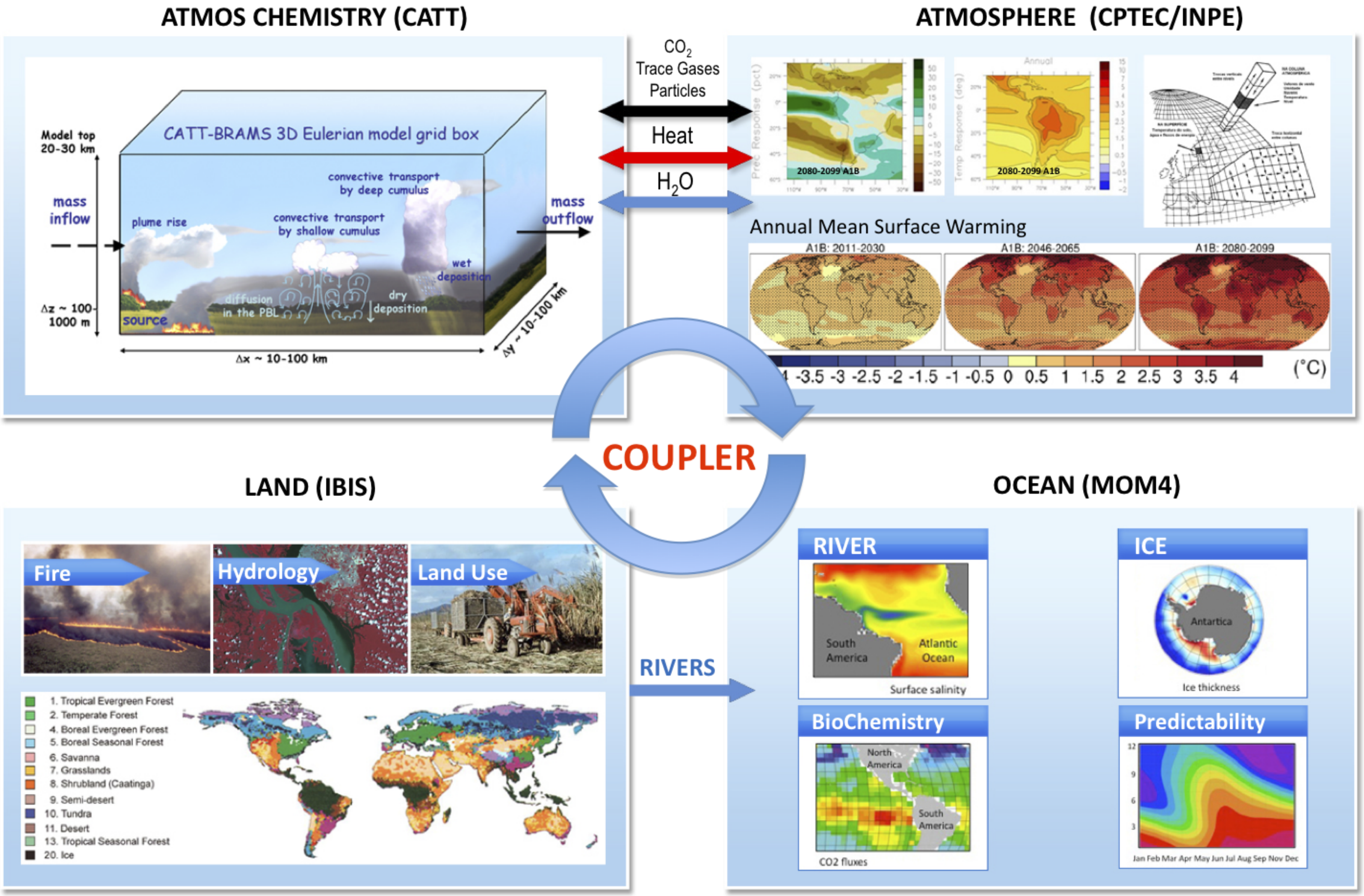
Other components

- *Atmospheric chemistry*
- *Evaporation*
- *Outgoing heat*

Modelagem climática: processos envolvidos

https://www2.ucar.edu/sites/default/files/news/2011/CESM_final.jpg

Modelo Brasileiro do Sistema Climático Global



Desenvolvimento

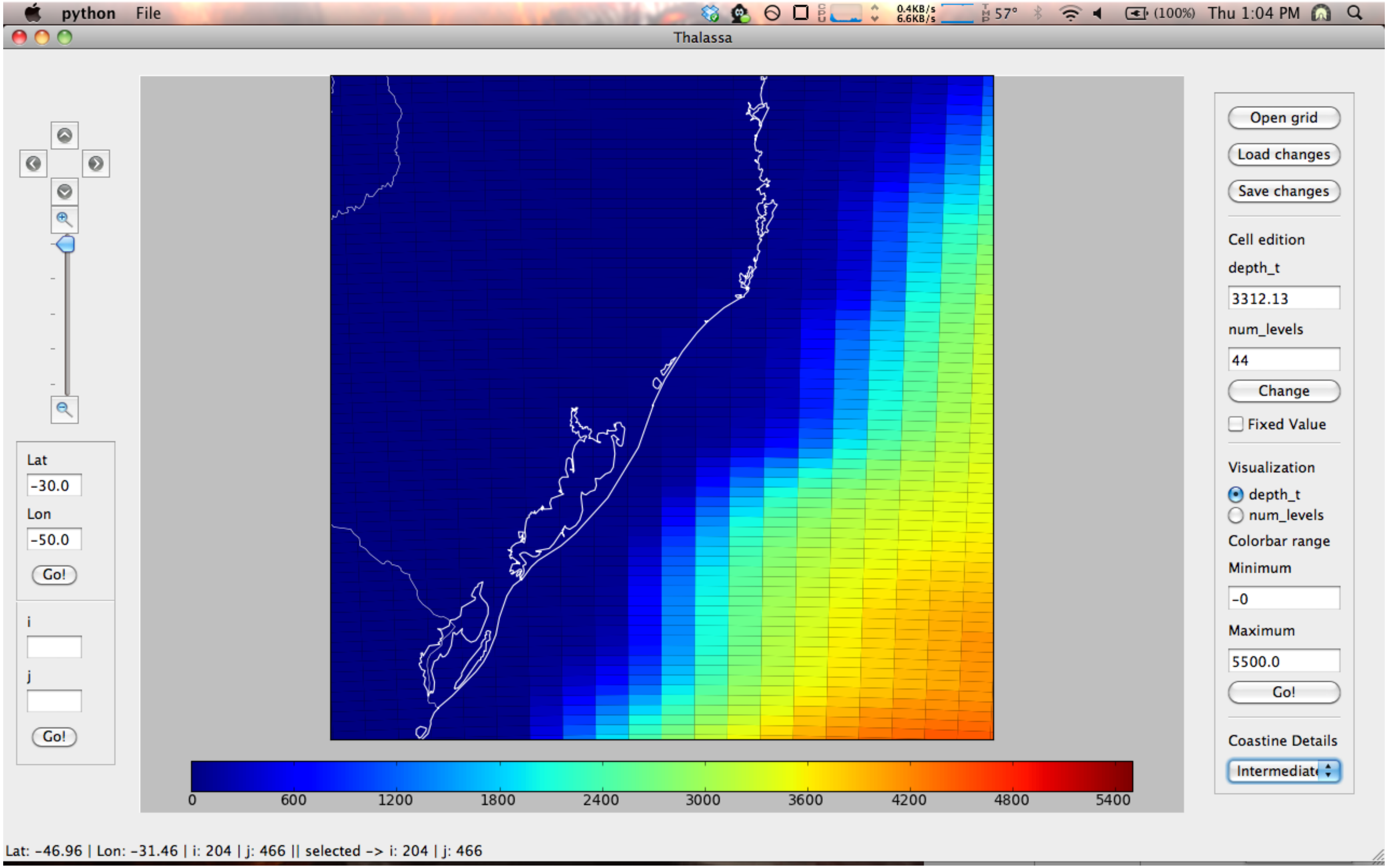
- Modelo escrito em Fortran
- Equipe
  - quatro integrantes inicialmente
  - cresceu para mais de quinze
- Como organizar o processo?

Desenvolvimento

- Sistema de controle de versões
  - Mercurial
- Redmine
- Google docs
- Lista de email

Thalassa

- Edição da grade oceânica
- Python
  - Numpy + Matplotlib + Basemap + PyQt
- Windows, OS X, Linux



Lat: -30.0

Lon: -50.0

i:

j:

Cell edition

depth_t: 3312.13

num_levels: 44

☐ Fixed Value

Visualization

● depth_t

○ num_levels

Colorbar range

Minimum: -0

Maximum: 5500.0

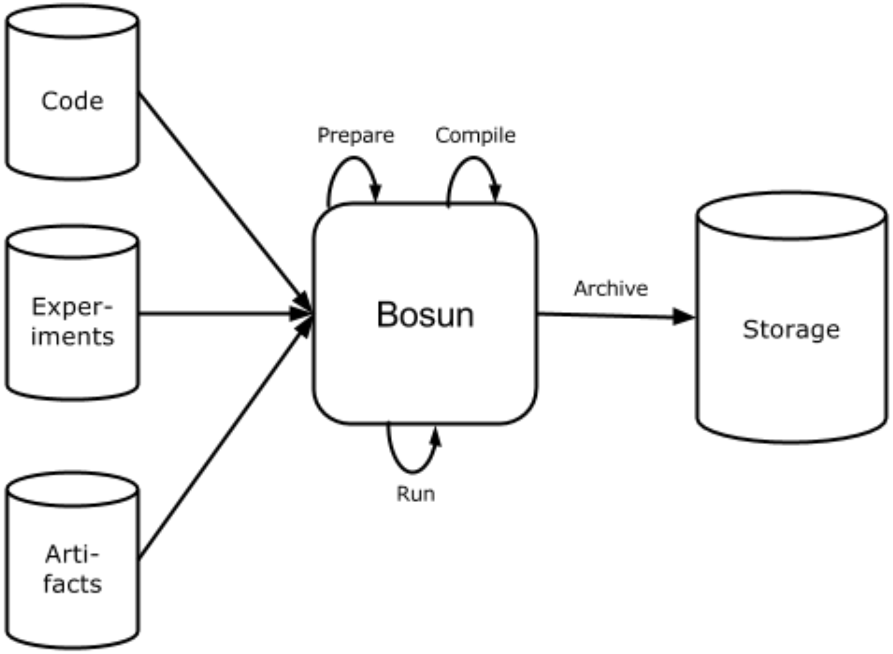
Coastine Details: Intermediate

Lat: -46.96 | Lon: -31.46 | i: 204 | j: 466 || selected -> i: 204 | j: 466

Um grande designer de UI =P

Integração contínua

- Bosun
  - Fabric
- Jenkins
  - monitorar o repositório
  - disparar os scripts
  - monitorar os status dos builds/runs



Jenkins

Operacionalização

- Execução diária
- Formatação dos dados
  - Entradas
  - Saídas
- Bosun
  - Python ~> Bash
  - Vários truques novos

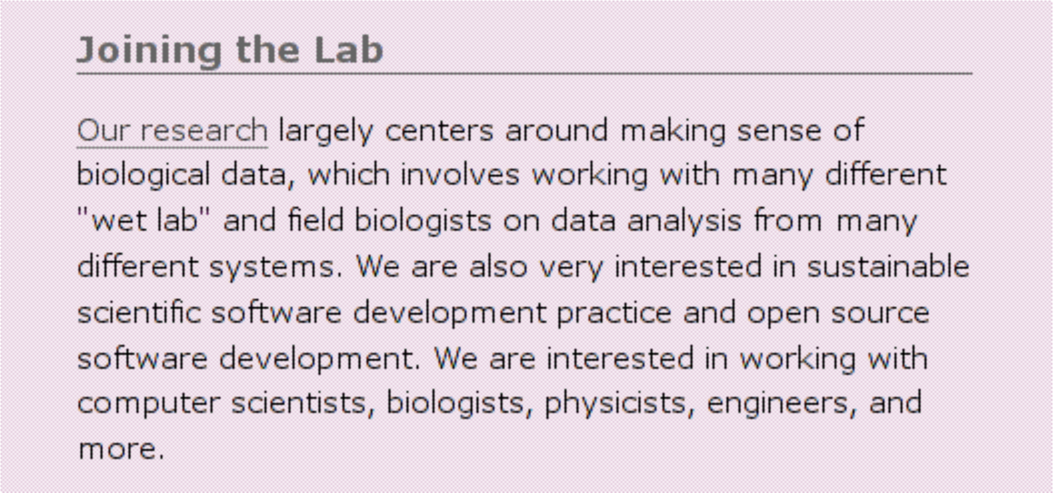
Validação do modelo

- Modelo gera resultados condizentes com a realidade?
- Validação é necessária para disponibilização do modelo
- Análise humana leva muito tempo
  - E necessita de uma equipe muito maior
- Testes
  - Comparar com medições
  - Avaliar numericamente quão bom está
- Como fazer?

Buscando alternativas

- Mercado
  - Ambiente técnico fantástico, mas...
  - Qual o objetivo?
- Academia
  - 'level up'
  - Muita autorreflexão…
  - Seguir na modelagem?

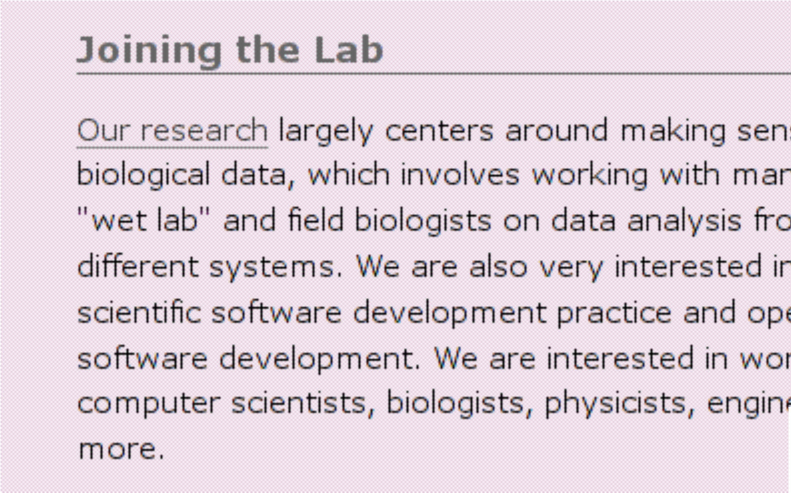
Bioinformática?



Joining the Lab

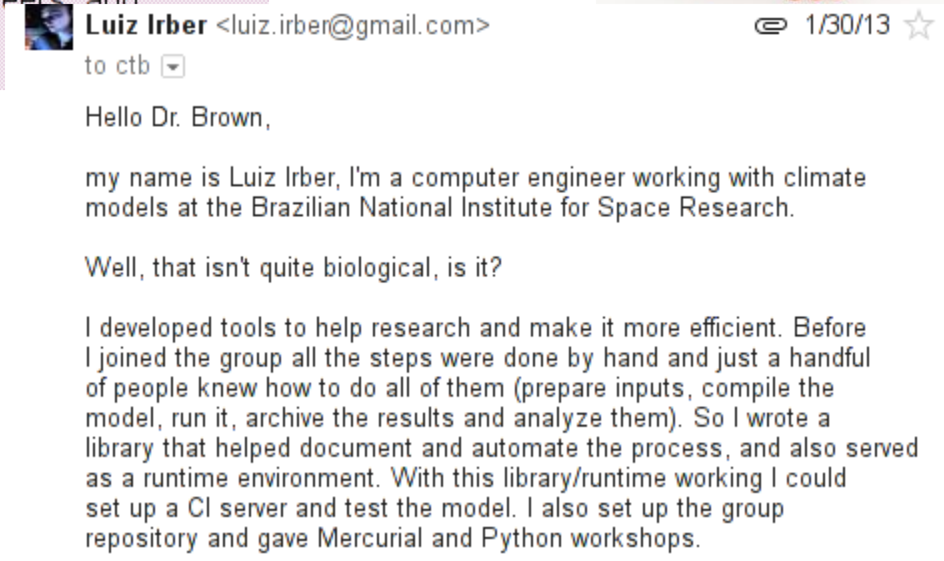
Our research largely centers around making sense of biological data, which involves working with many different "wet lab" and field biologists on data analysis from many different systems. We are also very interested in sustainable scientific software development practice and open source software development. We are interested in working with computer scientists, biologists, physicists, engineers, and more.

Bioinformática?



Joining the Lab

Our research largely centers around making sense of biological data, which involves working with many different "wet lab" and field biologists on data analysis from many different systems. We are also very interested in sustainable scientific software development practice and open source software development. We are interested in working with computer scientists, biologists, physicists, engineers, and more.



Luiz Irber <luiz.irber@gmail.com> 1/30/13

to ctb

Hello Dr. Brown,

my name is Luiz Irber, I'm a computer engineer working with climate models at the Brazilian National Institute for Space Research.

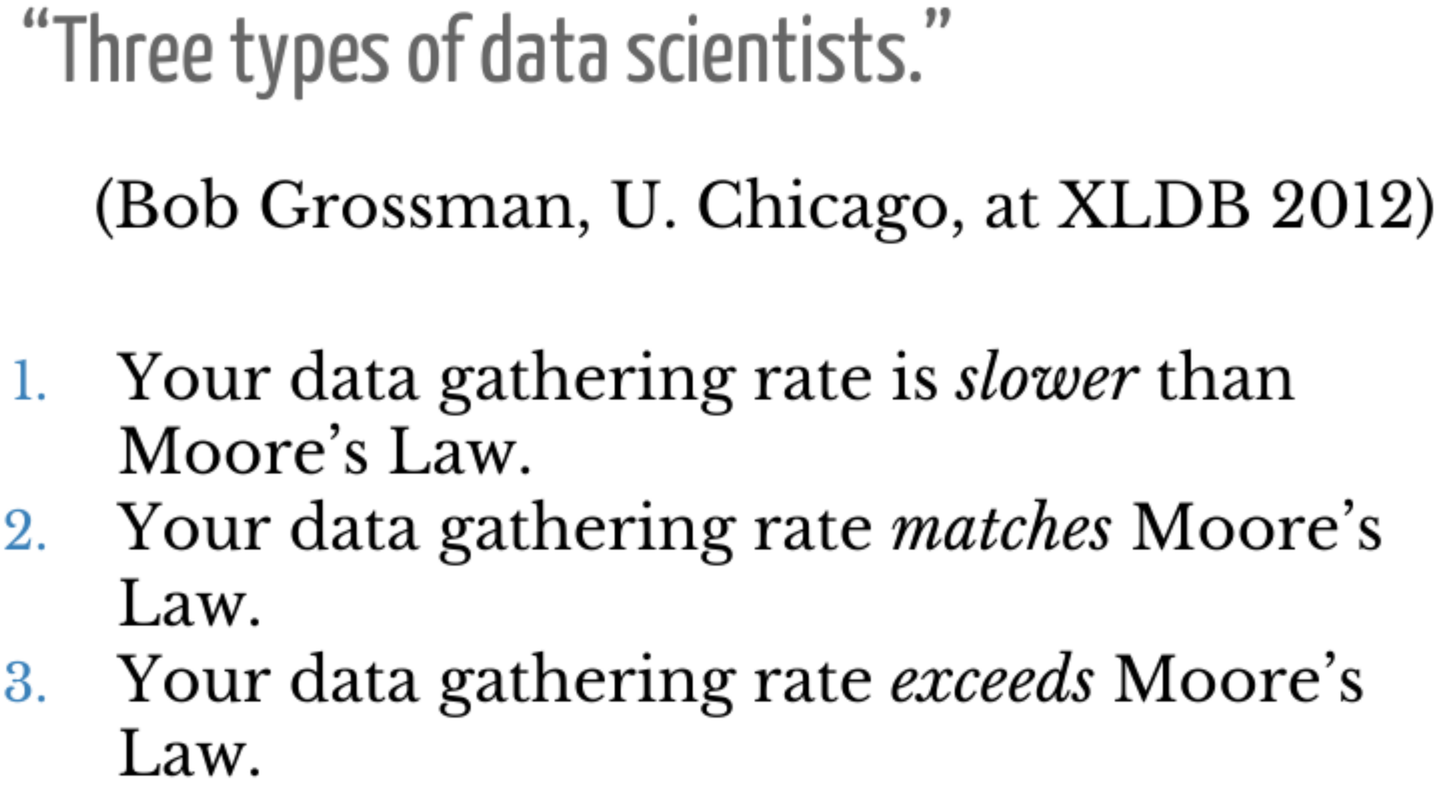
Well, that isn't quite biological, is it?

I developed tools to help research and make it more efficient. Before I joined the group all the steps were done by hand and just a handful of people knew how to do all of them (prepare inputs, compile the model, run it, archive the results and analyze them). So I wrote a library that helped document and automate the process, and also served as a runtime environment. With this library/runtime working I could set up a CI server and test the model. I also set up the group repository and gave Mercurial and Python workshops.

CALIFORNIA
OUTATIME

Bioinformática!

- Chicken genome improvement project
    - Regiões não mapeadas
        - Microcromossomos
- khmer: HyperLogLog
    - estimativa de cardinalidade
- Treinamento
    - Software Carpentry
    - undergrads
    - 'pair support'

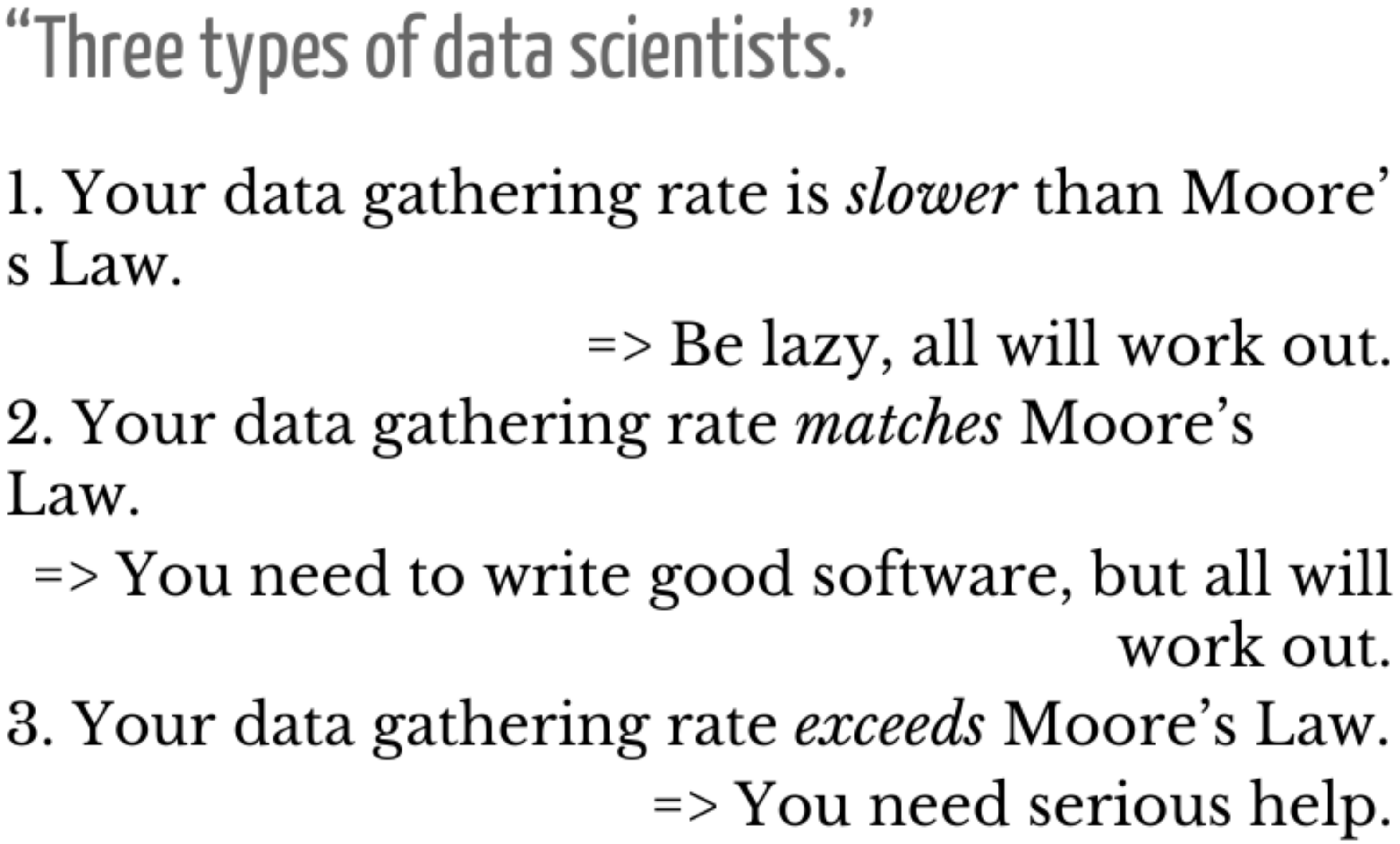


“Three types of data scientists.”

(Bob Grossman, U. Chicago, at XLDB 2012)

1. Your data gathering rate is *slower* than Moore’s Law.
2. Your data gathering rate *matches* Moore’s Law.
3. Your data gathering rate *exceeds* Moore’s Law.

(slide: Titus Brown)



“Three types of data scientists.”

1. Your data gathering rate is *slower* than Moore’s Law.

=> Be lazy, all will work out.

2. Your data gathering rate *matches* Moore’s Law.

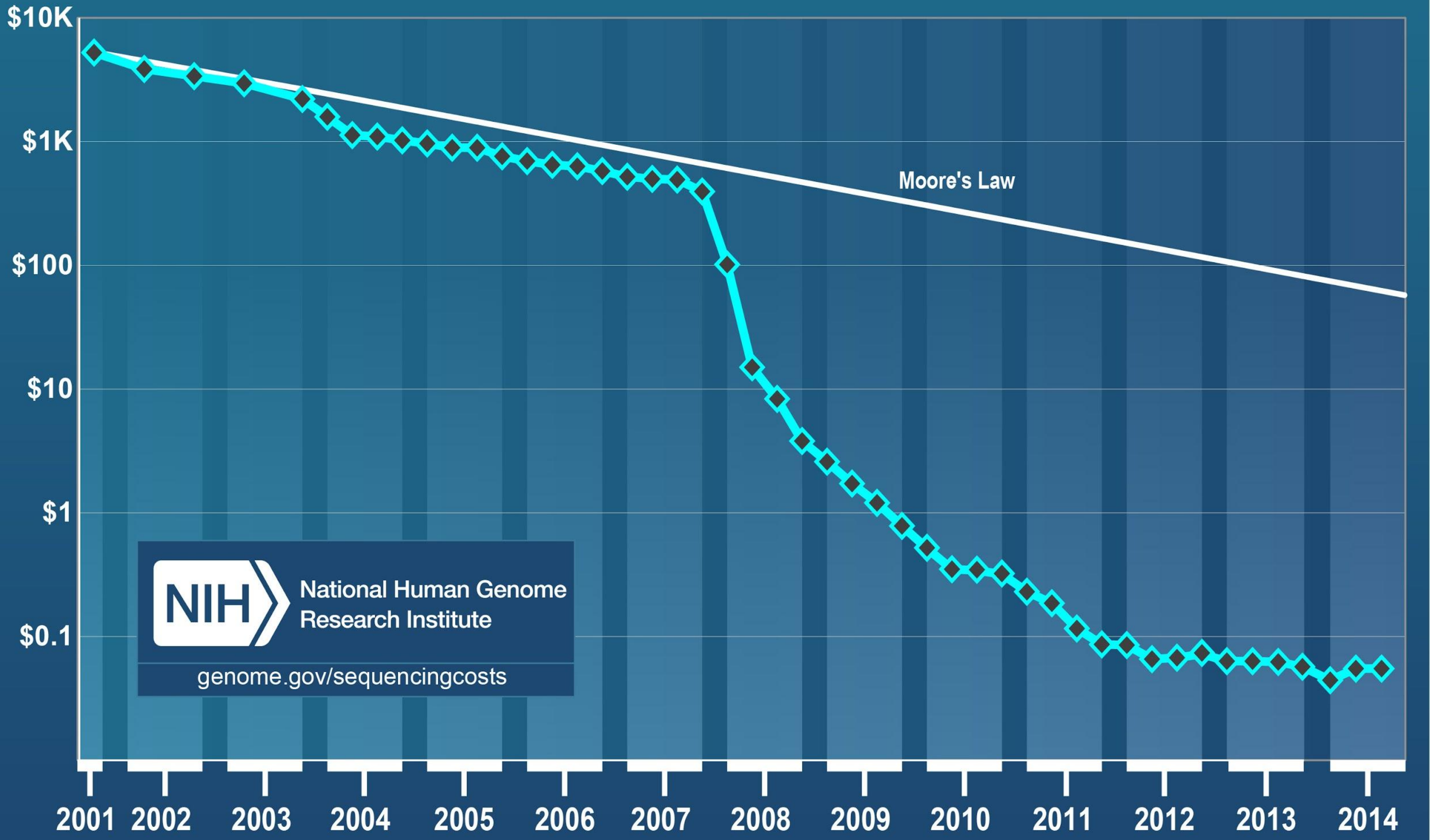
=> You need to write good software, but all will work out.

3. Your data gathering rate *exceeds* Moore’s Law.

=> You need serious help.

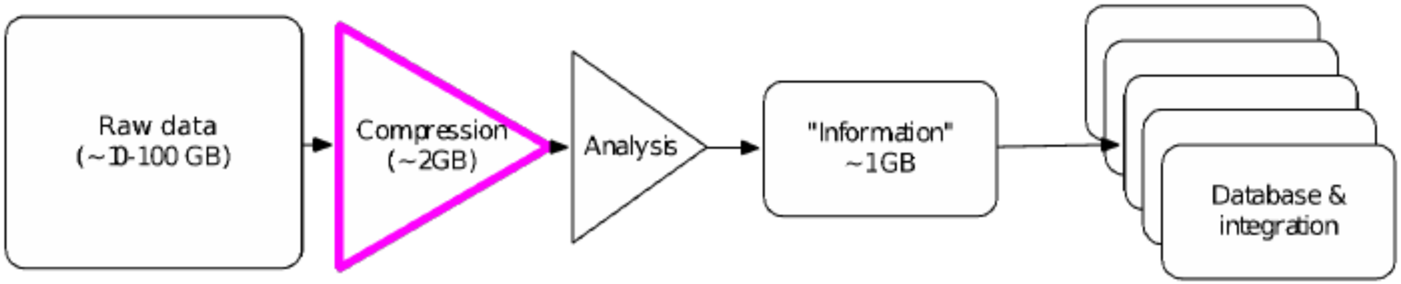
(slide: Titus Brown)

Cost per Raw Megabase of DNA Sequence



NIH National Human Genome Research Institute

genome.gov/sequencingcosts



A software & algorithms approach: can we develop ***lossy*** compression approaches that

1. Reduce data size & remove errors => efficient processing?
2. Retain all "information"? (think JPEG)

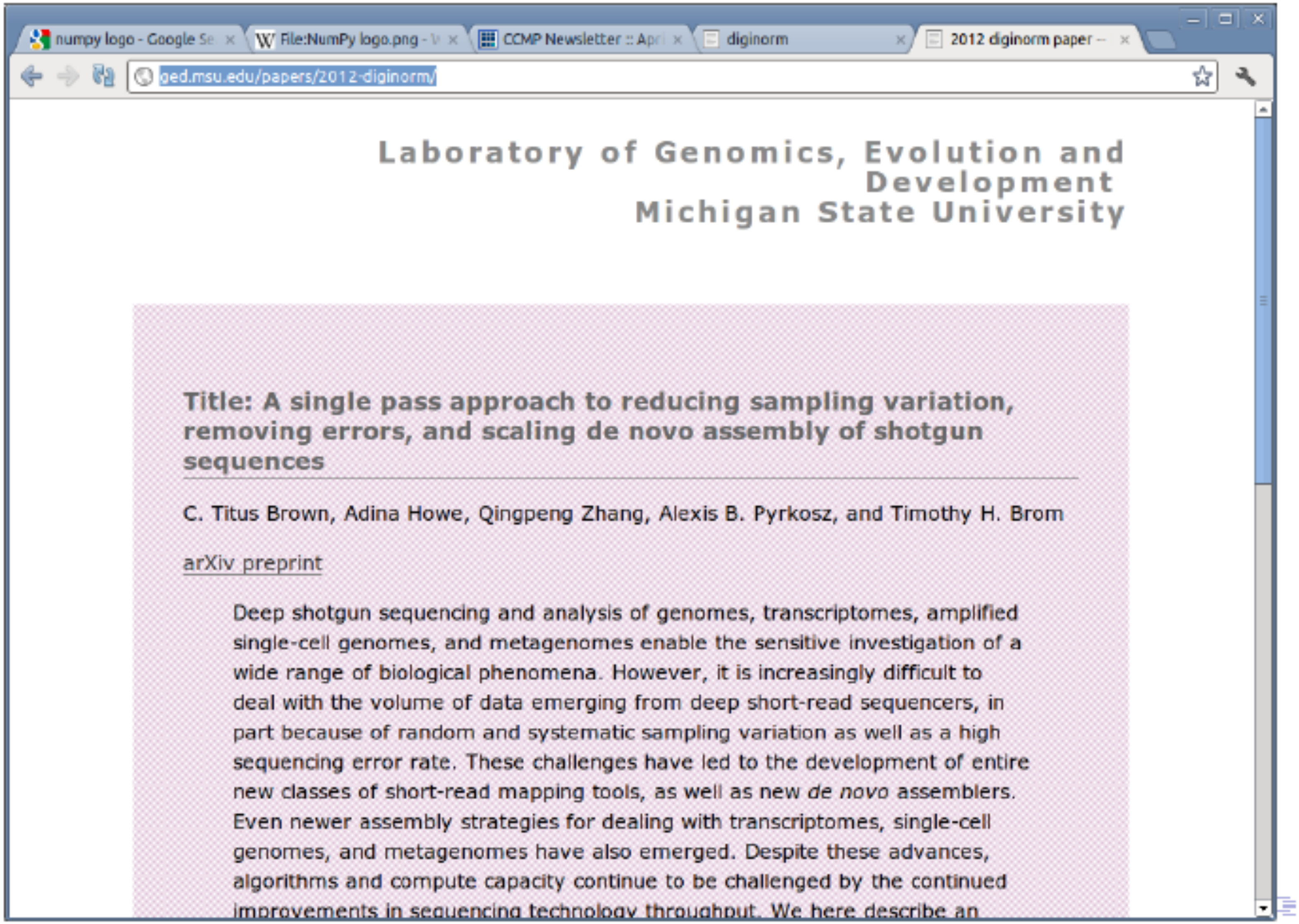
If so, then we can store only the compressed data for later reanalysis.

Short answer is: yes, we can.

(slide: Titus Brown)

The executable paper: Titus Brown (MSU), 3/21/12

http://arxiv.org/abs/1203.4802



Laboratory of Genomics, Evolution and Development
Michigan State University

Title: A single pass approach to reducing sampling variation, removing errors, and scaling de novo assembly of shotgun sequences

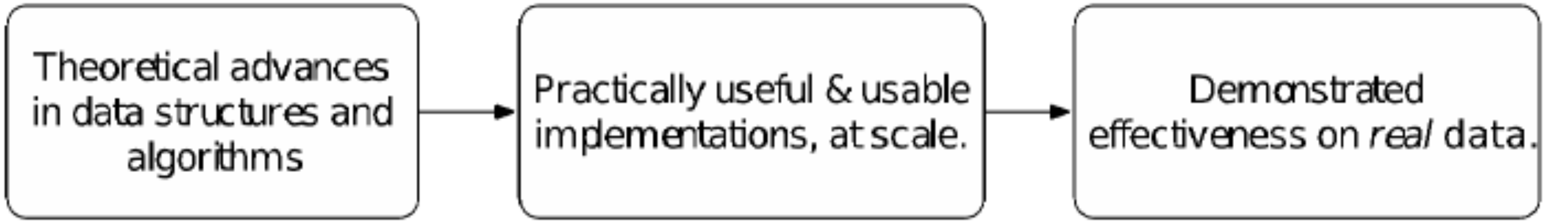
C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, and Timothy H. Brom

arXiv preprint

Deep shotgun sequencing and analysis of genomes, transcriptomes, amplified single-cell genomes, and metagenomes enable the sensitive investigation of a wide range of biological phenomena. However, it is increasingly difficult to deal with the volume of data emerging from deep short-read sequencers, in part because of random and systematic sampling variation as well as a high sequencing error rate. These challenges have led to the development of entire new classes of short-read mapping tools, as well as new *de novo* assemblers. Even newer assembly strategies for dealing with transcriptomes, single-cell genomes, and metagenomes have also emerged. Despite these advances, algorithms and compute capacity continue to be challenged by the continued improvements in sequencing technology throughput. We here describe an

http://ged.msu.edu/papers/2012-diginorm/
(slide: Fernando Perez)

Better science through superior software

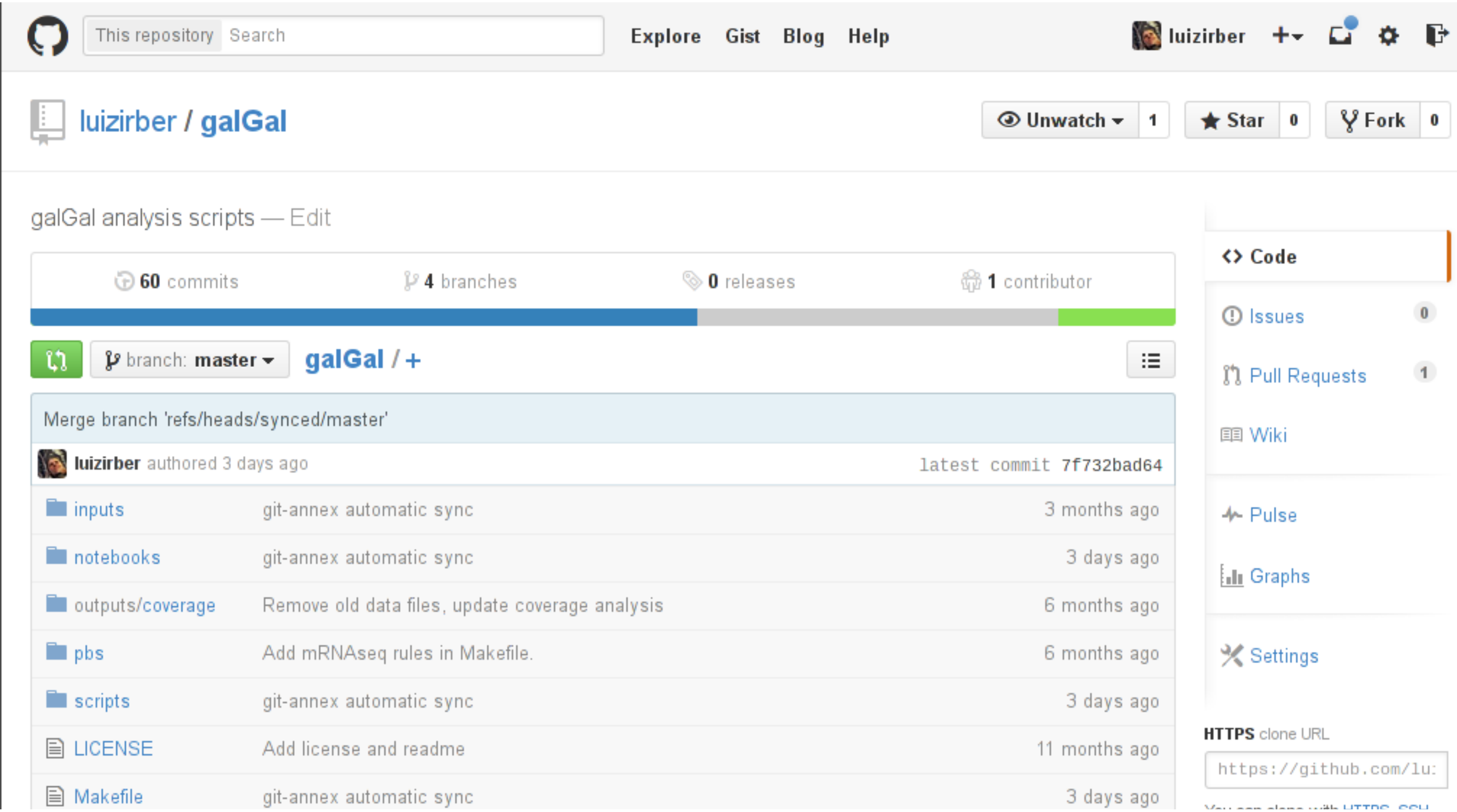


Theoretical advances in data structures and algorithms

Practically useful & usable implementations, at scale.

Demonstrated effectiveness on *real* data.

(slide: Titus Brown)



luizirber / galGal

galGal analysis scripts — Edit

60 commits · 4 branches · 0 releases · 1 contributor

Merge branch 'refs/heads/synced/master'

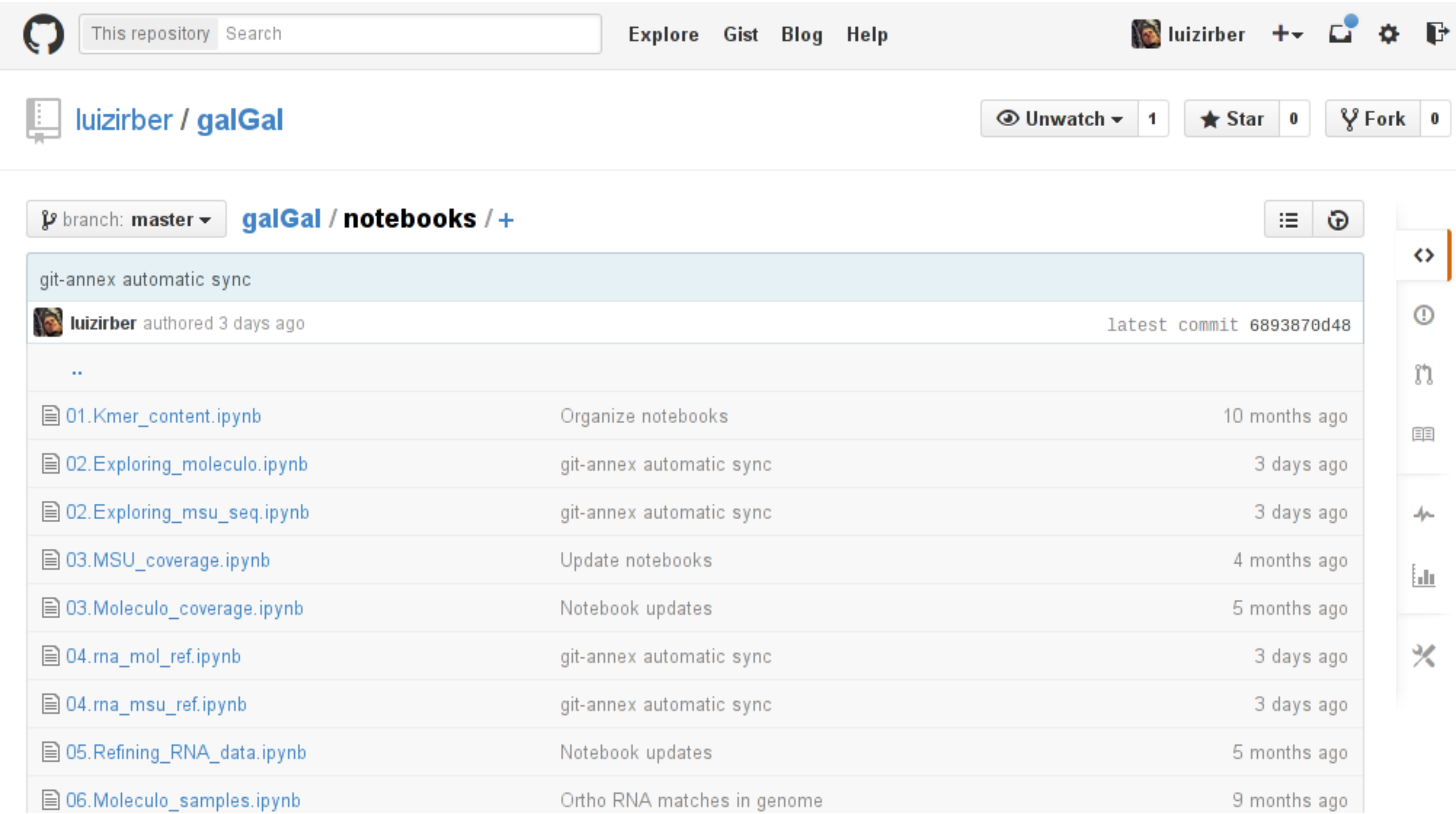
luizirber authored 3 days ago · latest commit 7f732bad64

inputs	git-annex automatic sync	3 months ago
notebooks	git-annex automatic sync	3 days ago
outputs/coverage	Remove old data files, update coverage analysis	6 months ago
pbs	Add mRNAseq rules in Makefile.	6 months ago
scripts	git-annex automatic sync	3 days ago
LICENSE	Add license and readme	11 months ago
Makefile	git-annex automatic sync	3 days ago

HTTPS clone URL

https://github.com/lu:

Licença e Makefile!
https://github.com/luizirber/galGal

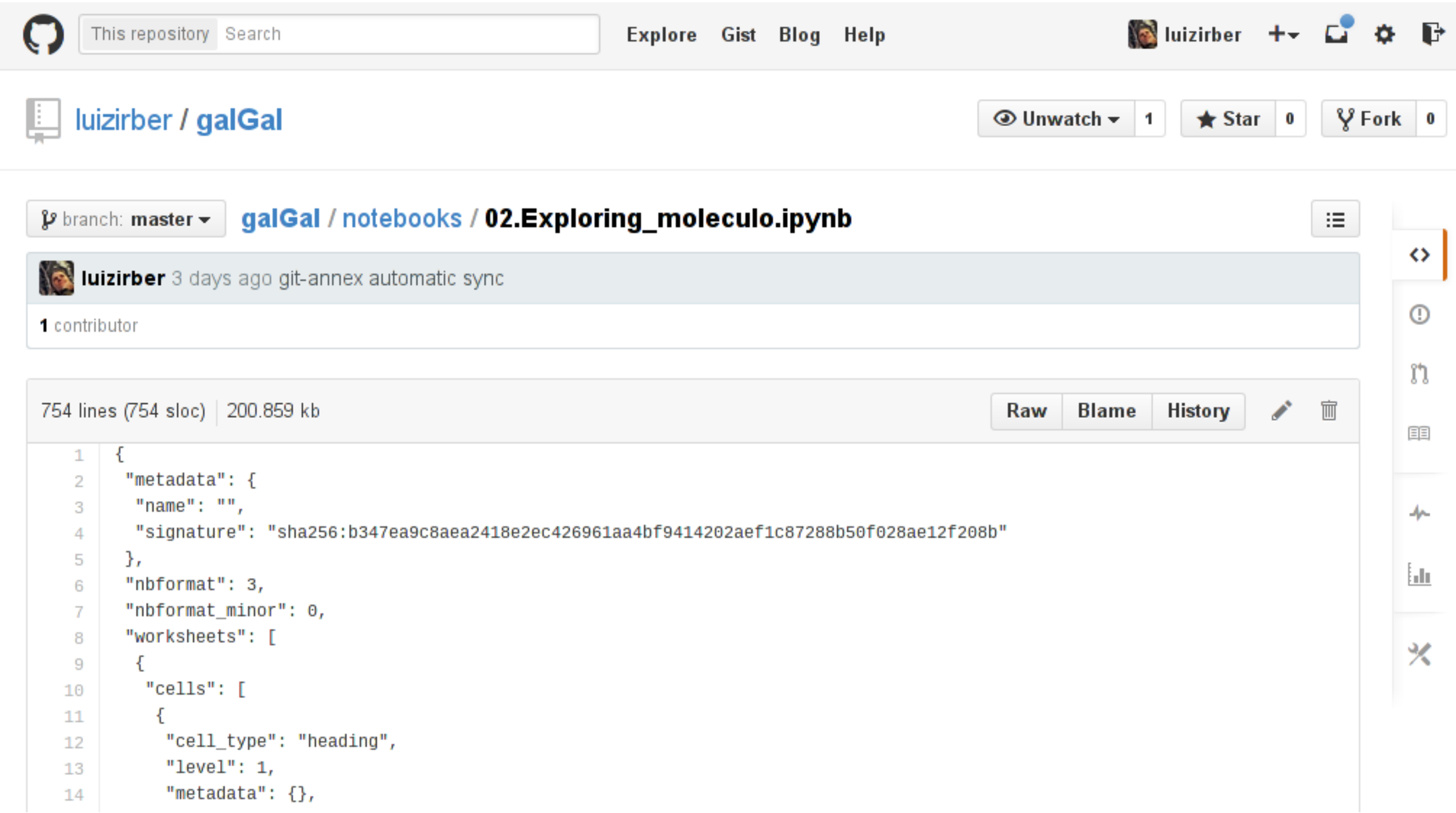


git-annex automatic sync

luizirber authored 3 days ago — latest commit 6893870d48

..		
01.Kmer_content.ipynb	Organize notebooks	10 months ago
02.Exploring_moleculo.ipynb	git-annex automatic sync	3 days ago
02.Exploring_msu_seq.ipynb	git-annex automatic sync	3 days ago
03.MSU_coverage.ipynb	Update notebooks	4 months ago
03.Moleculo_coverage.ipynb	Notebook updates	5 months ago
04.rna_mol_ref.ipynb	git-annex automatic sync	3 days ago
04.rna_msu_ref.ipynb	git-annex automatic sync	3 days ago
05.Refining_RNA_data.ipynb	Notebook updates	5 months ago
06.Moleculo_samples.ipynb	Ortho RNA matches in genome	9 months ago

Um notebook por passo
https://github.com/luizirber/galGal



luizirber / galGal

branch: master

galGal / notebooks / 02.Exploring_moleculo.ipynb

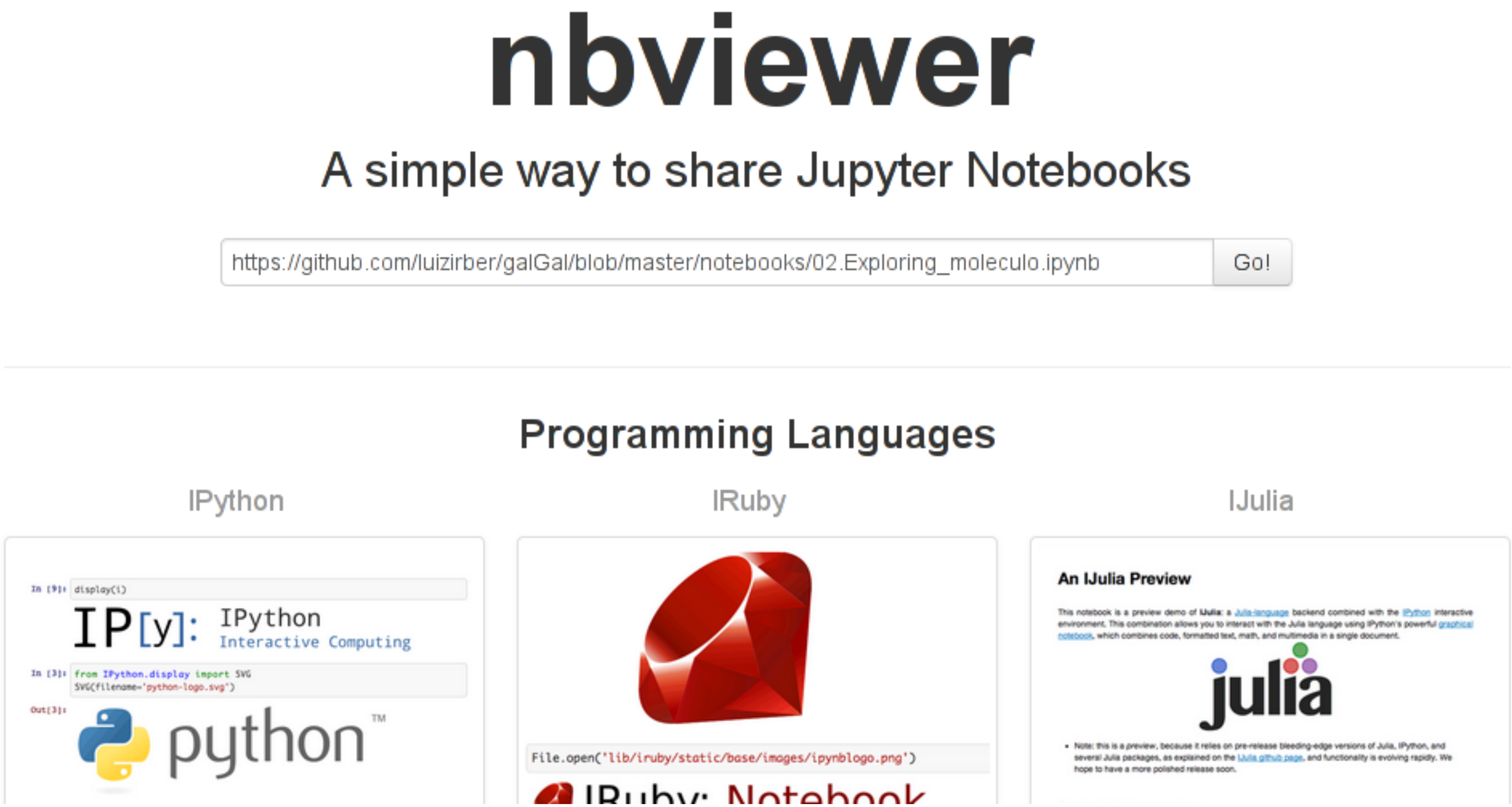
luizirber 3 days ago git-annex automatic sync

1 contributor

754 lines (754 sloc) | 200.859 kb

```
{
 "metadata": {
  "name": "",
  "signature": "sha256:b347ea9c8aea2418e2ec426961aa4bf9414202aef1c87288b50f028ae12f208b"
 },
 "nbformat": 3,
 "nbformat_minor": 0,
 "worksheets": [
  {
   "cells": [
    {
     "cell_type": "heading",
     "level": 1,
     "metadata": {},
```

Hmmmm….
https://github.com/luizirber/galGal



nbviewer

A simple way to share Jupyter Notebooks

https://github.com/luizirber/galGal/blob/master/notebooks/02.Exploring_moleculo.ipynb

Programming Languages

IPython

```
In [9]: display(i)
```

```
In [3]: from IPython.display import SVG
        SVG(filename='python-logo.svg')
Out[3]:
```

IRuby

```
File.open('lib/iruby/static/base/images/ipynblogo.png')
```

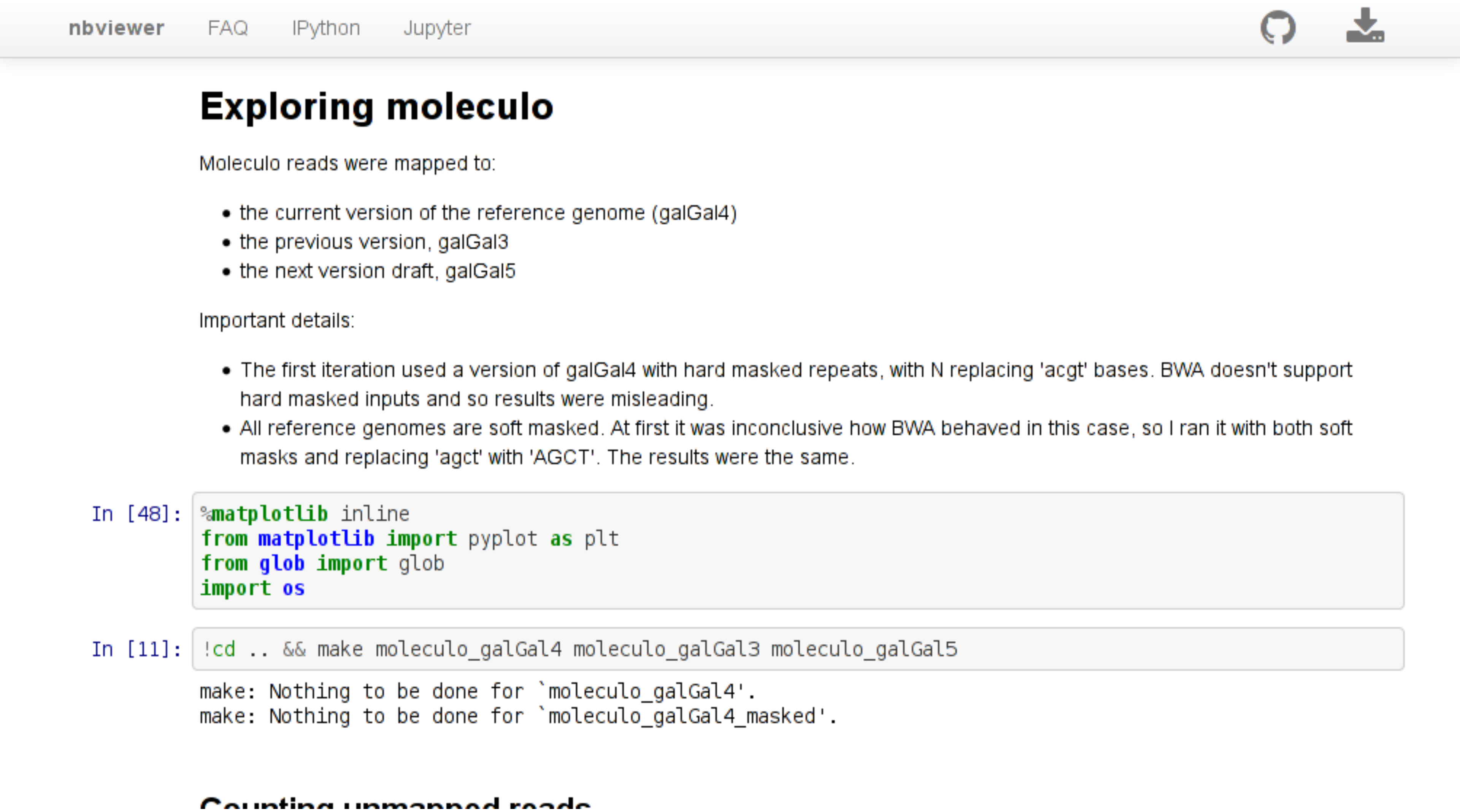
IJulia

An IJulia Preview

This notebook is a preview demo of **IJulia**: a Julia-language backend combined with the IPython interactive environment. This combination allows you to interact with the Julia language using IPython's powerful graphical notebook, which combines code, formatted text, math, and multimedia in a single document.

- Note: this is a *preview*, because it relies on pre-release bleeding-edge versions of Julia, IPython, and several Julia packages, as explained on the IJulia github page, and functionality is evolving rapidly. We hope to have a more polished release soon.

Visualizando notebooks
https://github.com/luizirber/galGal



Exploring moleculo

Moleculo reads were mapped to:

- the current version of the reference genome (galGal4)
- the previous version, galGal3
- the next version draft, galGal5

Important details:

- The first iteration used a version of galGal4 with hard masked repeats, with N replacing 'acgt' bases. BWA doesn't support hard masked inputs and so results were misleading.
- All reference genomes are soft masked. At first it was inconclusive how BWA behaved in this case, so I ran it with both soft masks and replacing 'agct' with 'AGCT'. The results were the same.

In [48]:

```
%matplotlib inline
from matplotlib import pyplot as plt
from glob import glob
import os
```

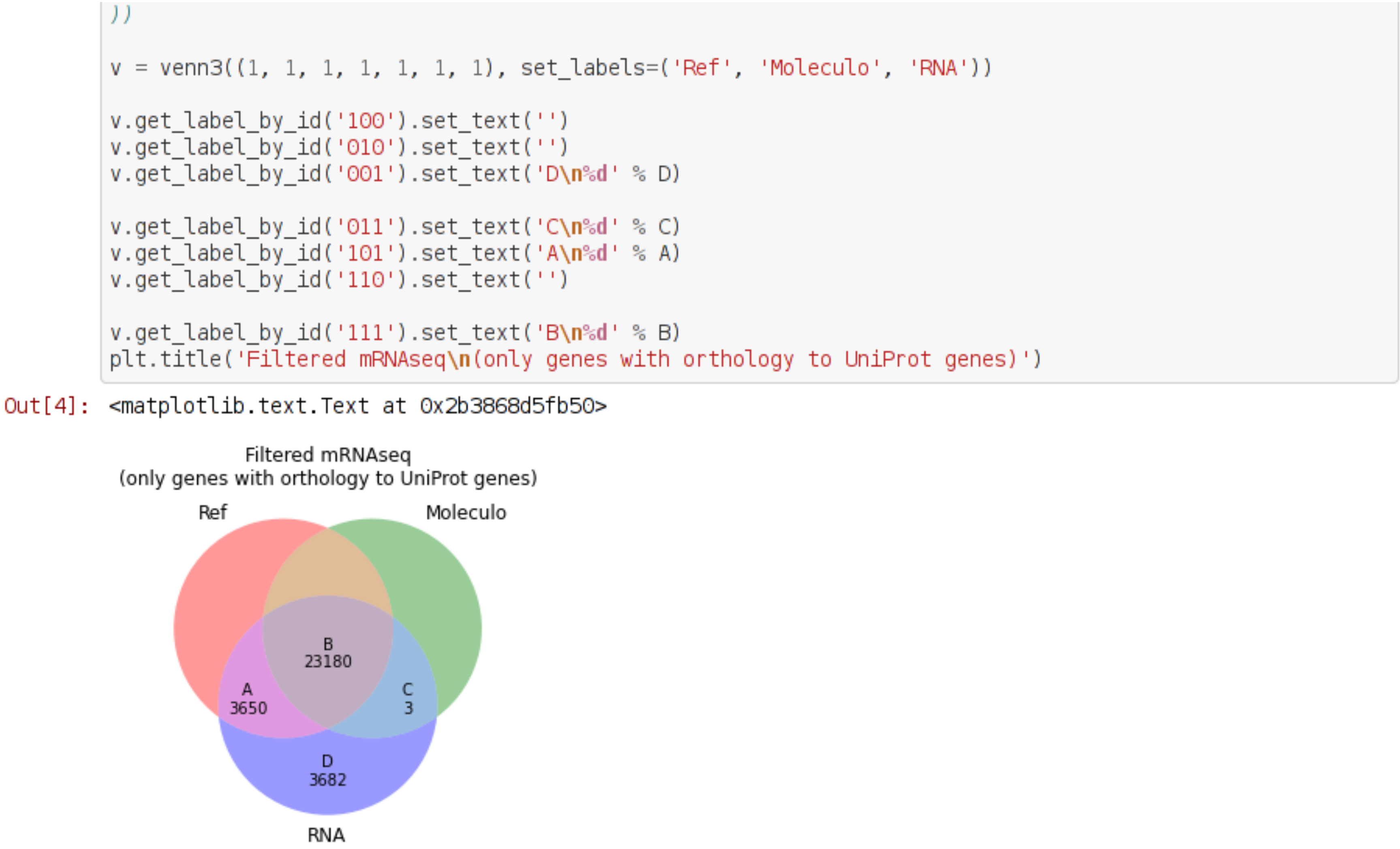
In [11]:

```
!cd .. && make moleculo_galGal4 moleculo_galGal3 moleculo_galGal5
```

```
make: Nothing to be done for `moleculo_galGal4'.
make: Nothing to be done for `moleculo_galGal4_masked'.
```

Counting unmapped reads

Explicações, código e comandos
https://github.com/luizirber/galGal



```
))

v = venn3((1, 1, 1, 1, 1, 1, 1), set_labels=('Ref', 'Moleculo', 'RNA'))

v.get_label_by_id('100').set_text('')
v.get_label_by_id('010').set_text('')
v.get_label_by_id('001').set_text('D\n%d' % D)

v.get_label_by_id('011').set_text('C\n%d' % C)
v.get_label_by_id('101').set_text('A\n%d' % A)
v.get_label_by_id('110').set_text('')

v.get_label_by_id('111').set_text('B\n%d' % B)
plt.title('Filtered mRNAseq\n(only genes with orthology to UniProt genes)')
```

```
Out[4]: <matplotlib.text.Text at 0x2b3868d5fb50>
```

Figuras também!
https://github.com/luizirber/galGal

The Lifecycle of a Scientific Idea (schematically)

1. Individual exploratory work
2. Collaborative development
3. Production work (HPC, cloud, parallel)
4. Publication (with reproducible results!)
5. Education
6. Goto 1.

The Problem with most tools

Barriers and discontinuities in workflow in between all the steps

(slide: Fernando Perez)

A crisis of credibility and real issues

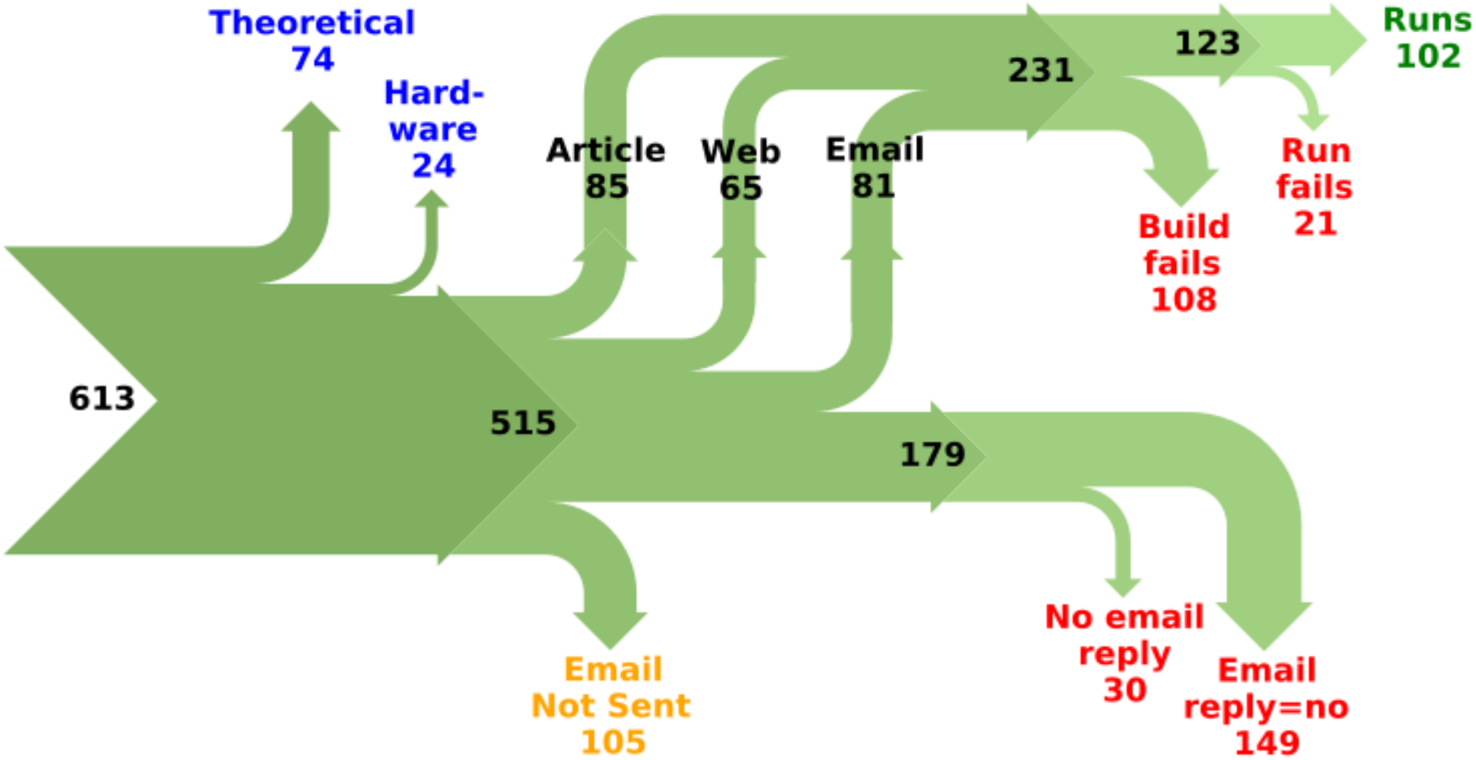
- **The Duke clinical trials scandal** - Potti/Nevin
  - A compounding of (common and otherwise) data analysis errors.
  - No materials allowing validation/reproduction of results.
  - Lawsuits, resignations, careers destroyed.
  - More importantly: **Patients were harmed.**
  - Major policy reviews and changes: NCI, IOM, ...
  - More: see K. Baggerly's "starter set" page.
- The Duke situation is more common than we'd like to believe!
  - Begley & Ellis, Nature, 3/28/12: *Drug development: Raise standards for preclinical cancer research.*
  - 47 out of 53 "landmark papers" could not be replicated.
- Nature, Feb 2012, Ince et al: *The case for open computer programs*
  - "The scientific community places more faith in computation than is justified"
  - "anything less than the release of actual source code is an indefensible approach for any scientific results that depend on computation"

(slide: Fernando Perez)

Reproducibilidade em Ciência da Computação

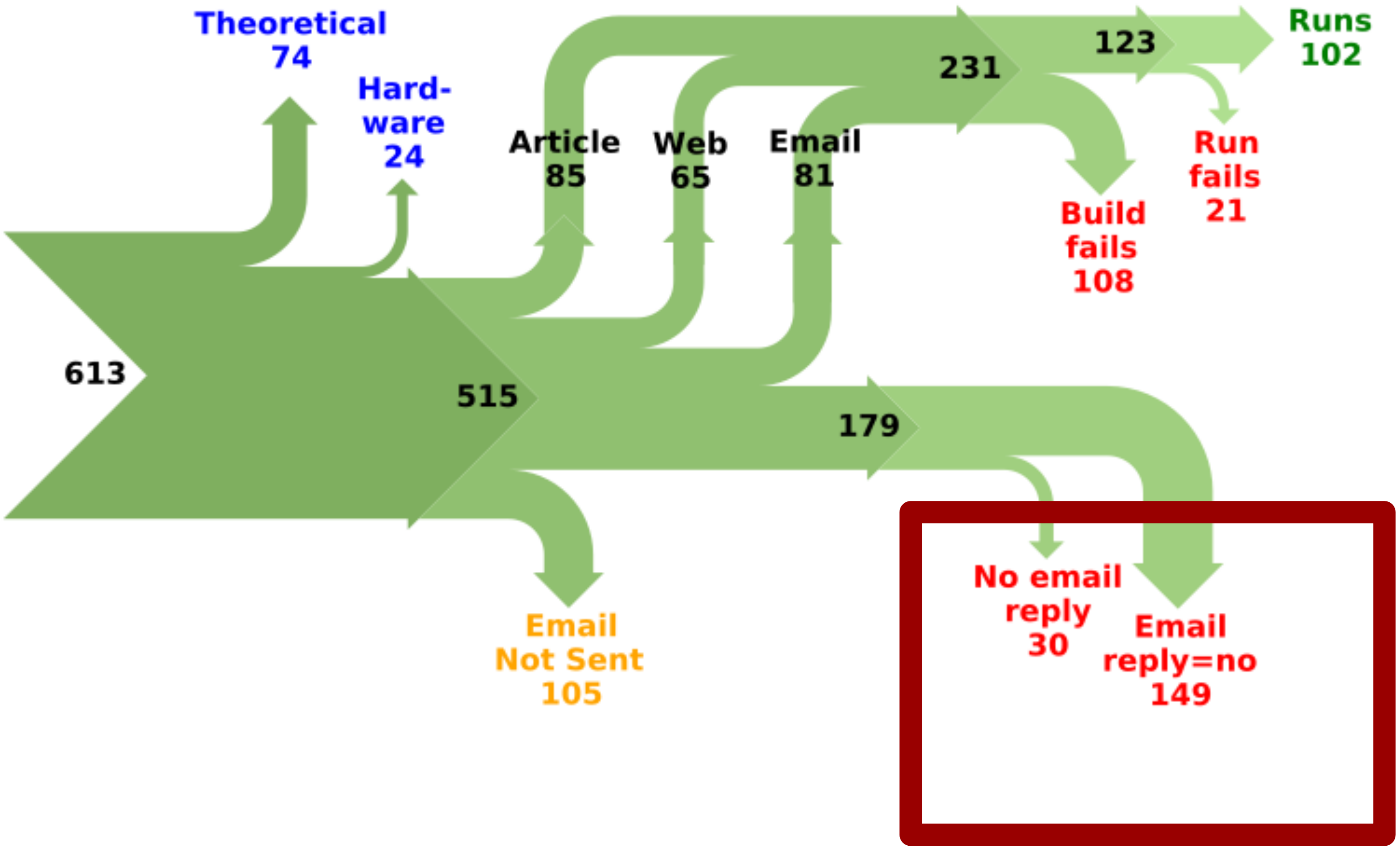
http://reproducibility.cs.arizona.edu/

"Can a CS student build the software within 30 minutes, including finding and installing any dependent software and libraries, and without bothering the authors?"



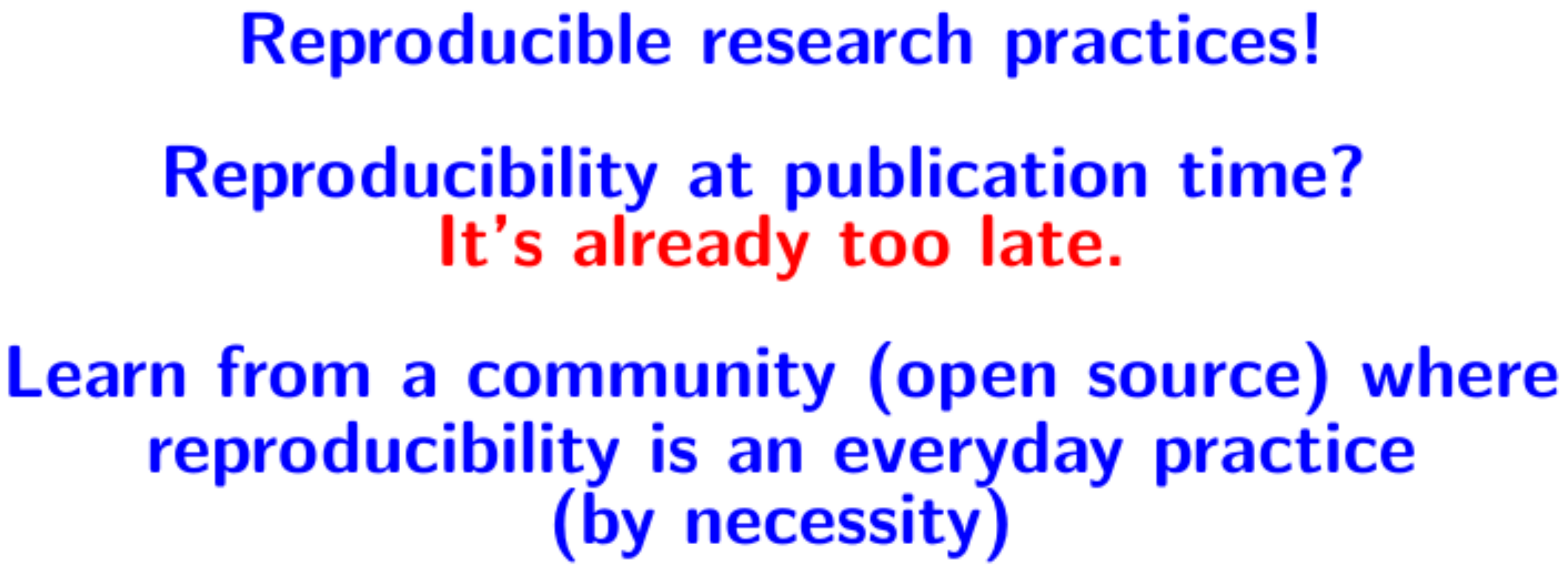
Measuring Reproducibility in Computer Systems Research

http://reproducibility.cs.arizona.edu/



Measuring Reproducibility in Computer Systems Research

http://reproducibility.cs.arizona.edu/



Reproducible research practices!

Reproducibility at publication time?
It's already too late.

Learn from a community (open source) where
reproducibility is an everyday practice
(by necessity)

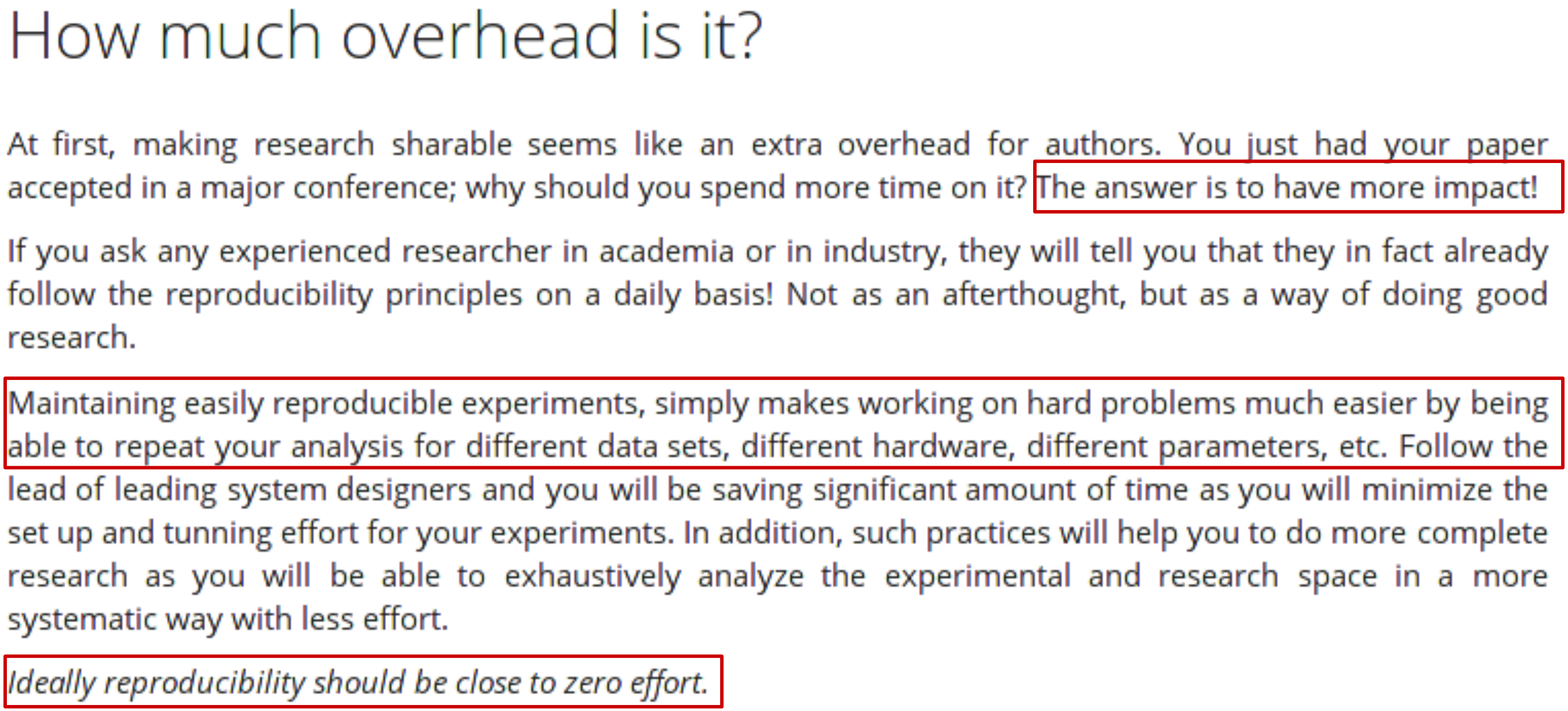
(slide: Fernando Perez)

Open source como exemplo

https://wiki.debian.org/ReproducibleBuilds

"It should be possible to reproduce, byte for byte, every build of every package in Debian."

(quantos anos até o projeto Debian ter estrutura suficiente para isso?)



How much overhead is it?

At first, making research sharable seems like an extra overhead for authors. You just had your paper accepted in a major conference; why should you spend more time on it? The answer is to have more impact!

If you ask any experienced researcher in academia or in industry, they will tell you that they in fact already follow the reproducibility principles on a daily basis! Not as an afterthought, but as a way of doing good research.

Maintaining easily reproducible experiments, simply makes working on hard problems much easier by being able to repeat your analysis for different data sets, different hardware, different parameters, etc. Follow the lead of leading system designers and you will be saving significant amount of time as you will minimize the set up and tunning effort for your experiments. In addition, such practices will help you to do more complete research as you will be able to exhaustively analyze the experimental and research space in a more systematic way with less effort.

Ideally reproducibility should be close to zero effort.

http://db-reproducibility.seas.harvard.edu/

Versioned science

Git: the tool you didn't know you needed

Reproducibility?

- Tracking and recreating every step of your work
- In the software world: it's called Version Control!

Git: an enabling technology. Use version control **for everything**

- Paper/grant writing (never get paper_v5_john.tex by email again!)

```
git clone git@server:/my/grant/repo.git
cd repo
make nsf-fastlane
```

- Everyday research: track your results
- Collaboration: synchronize multi-author work.
- Teaching!

(slide: Fernando Perez)

Make it work, then make it better

- Soluções de baixo atrito
  - Pessoas devem ser capazes de usar!
  - e entender!
- Makefile
  - Pipeline
- Scripts
  - código reusável
- Notebooks
  - Análise
  - Figuras
  - "Amarrar"

Treinamento

- Como ensinar pensamento computacional para cientistas?
- Como evitar "push button research"?

software carpentry

- organização sem fins lucrativos
  - voluntários
- ensinar habilidades computacionais básicas para fazer cientistas mais produtivos e seu trabalho mais confiável a longo prazo
- Workshops curtos (e intensos!)
  - organização de programas, controle de versões, testes e automação de tarefas.

Perguntas?

@luizirber

cienciaaberta@luizirber.org