Jupyter Notebooks for Performing & Sharing Bioinformatics Analyses

Windows users: From the Anaconda R:

```
> install.packages(
    c('rzmq','repr','IRkernel','IRdisplay'),
    repos = c('http://irkernel.github.io/',
              'http://cran.utstat.utoronto.ca/'))
```

Everyone: Download and unpack

```
> install.packages(
  c('rzmq','repr','IRkernel','IRdisplay'),
  repos = c('http://irkernel.github.io/',
            'http://cran.utstat.utoronto.ca/'))
```

OICR
Ontario Institute
for Cancer Research

Jupyter Notebooks for Performing & Sharing Bioinformatics Analyses

Jonathan Dursi, OICR
jonathan@dursi.ca
https://github.com/ljdursi/glbio-jupyter-workshop

The plan for this morning

- Run Jupyter notebooks on your laptop, either in R and Python
- How to perform interactive analyses in a web browser using Jupiter
- How does it all work?
- Using markdown and latex to format notebooks nicely
- "Port" an R bioinformatics workflow from scripts into a Jupyter notebook
- Sharing a Jupyter notebook online:
  - SageMathCloud
  - GitHub and git; nbviewer
  - mybinder.org

OICR
Ontario Institute
for Cancer Research

The plan for this morning

Perfectly reasonable and interesting things to discuss that we ***won't*** be talking about today:

- Combining R and Python: cool and possible but requires a little setup
- IPython magics (probably)
- The zillion R bioinformatics packages that are available
- How to use R, Python

OICR
Ontario Institute
for Cancer Research

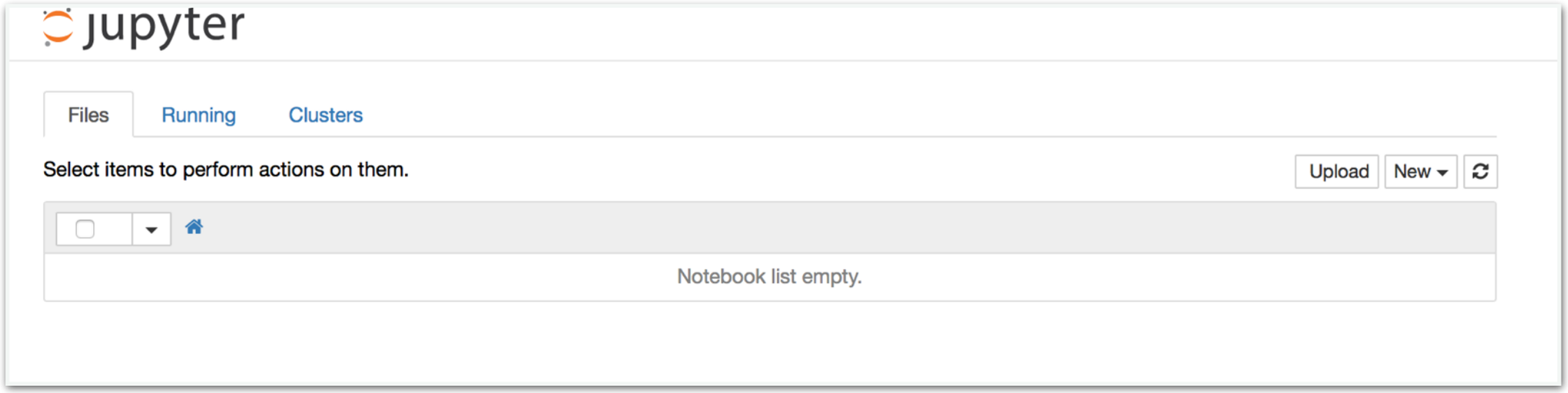
Let's start up a notebook

- Make a Notebooks directory and start from there
- Command Line: Linux, Mac

```
1. python3.5
coredump:~ ljdursi$ mkdir ~/Desktop/Notebooks
coredump:~ ljdursi$ cd ~/Desktop/Notebooks/
coredump:Notebooks ljdursi$ jupyter notebook
[I 12:56:20.381 NotebookApp] The port 8888 is already in use, trying another random port.
[I 12:56:20.390 NotebookApp] Serving notebooks from local directory: /Users/ljdursi/Desktop/N
otebooks
[I 12:56:20.391 NotebookApp] 0 active kernels
[I 12:56:20.391 NotebookApp] The Jupyter Notebook is running at: http://localhost:8889/
[I 12:56:20.391 NotebookApp] Use Control-C to stop this server and shut down all kernels (twi
ce to skip confirmation).
```

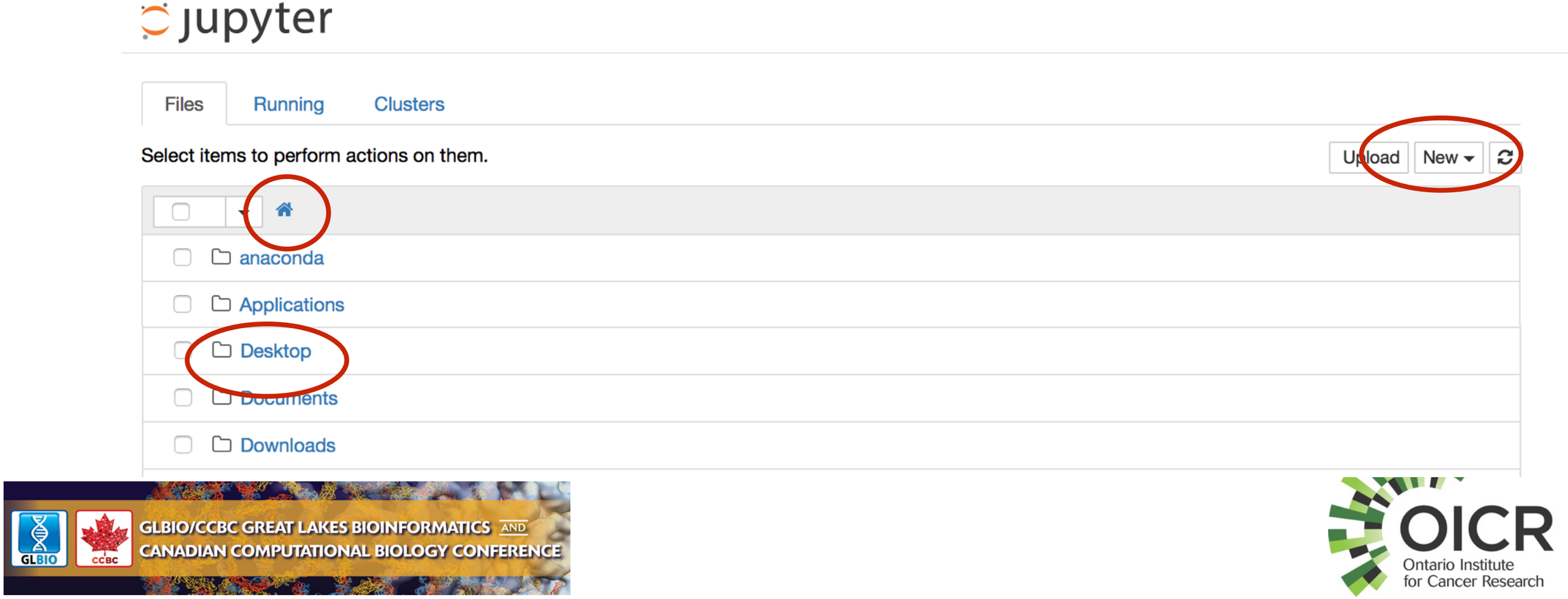
Let's start up a notebook

- Make a Notebooks directory and start from there
- Command Line: Linux, Mac
- Browser should open (or go to localhost:8888)



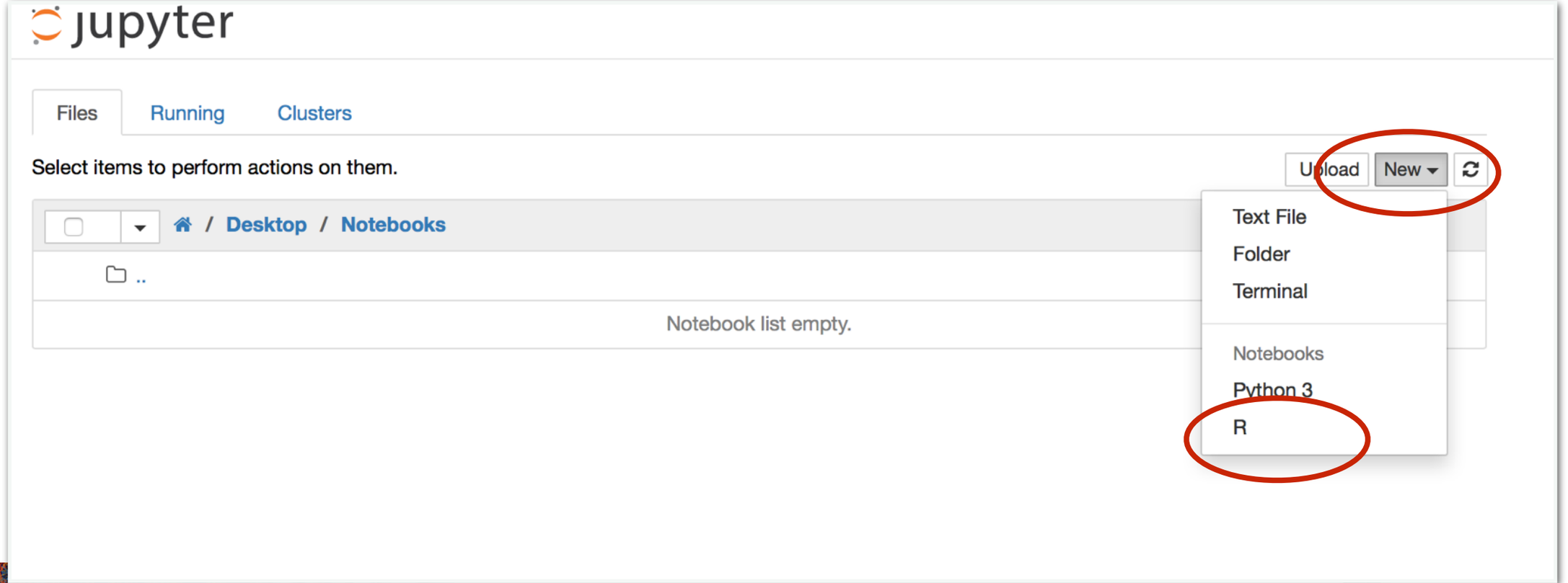
Let's start up a notebook

- Or you can start from home (eg, launch from start menu: Win) and navigate folders, and create new folders:



Let's start up a notebook

- Let's start up an R notebook

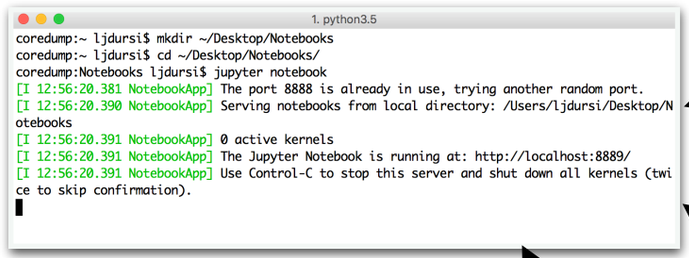


Demo - starting notebook

- 1-Introduction

How does this all work?

notebook server



nb#1

nb#2

nb#3

R kernel 1

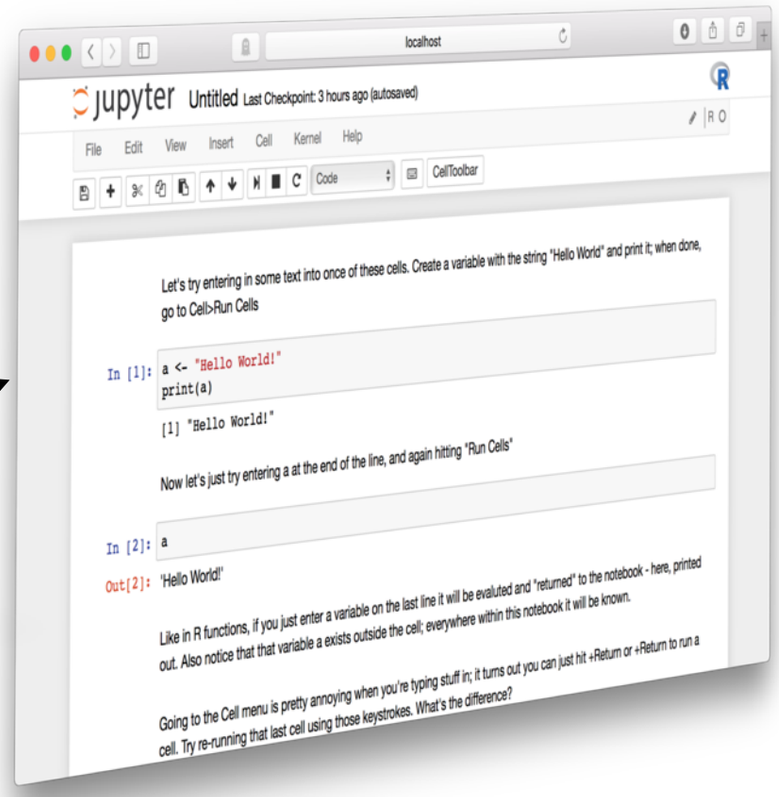
```
a <- "Hello World!"
x <- 10
y <- 5
```

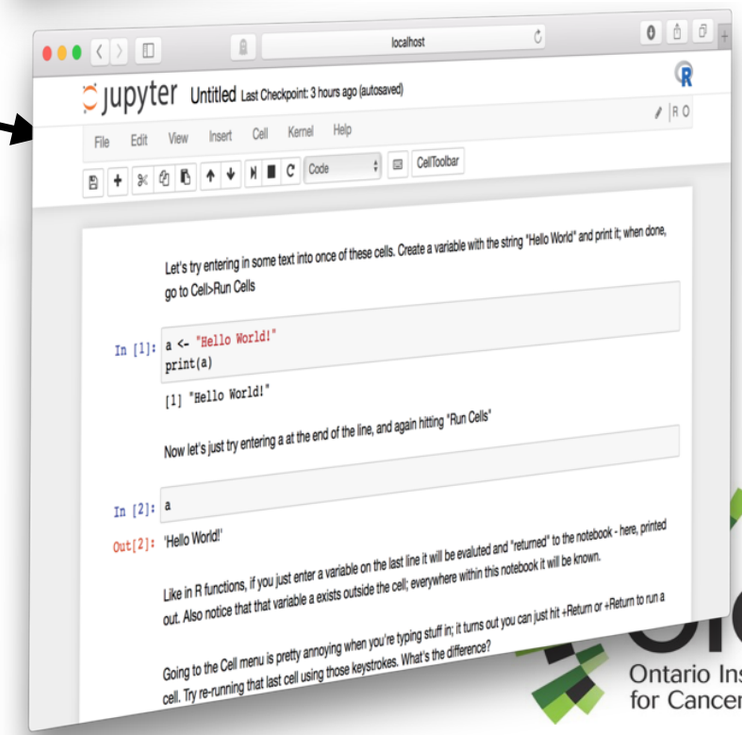
R kernel 2

```
a <- "Goodbye!"
y <- -1
```

Py kernel1

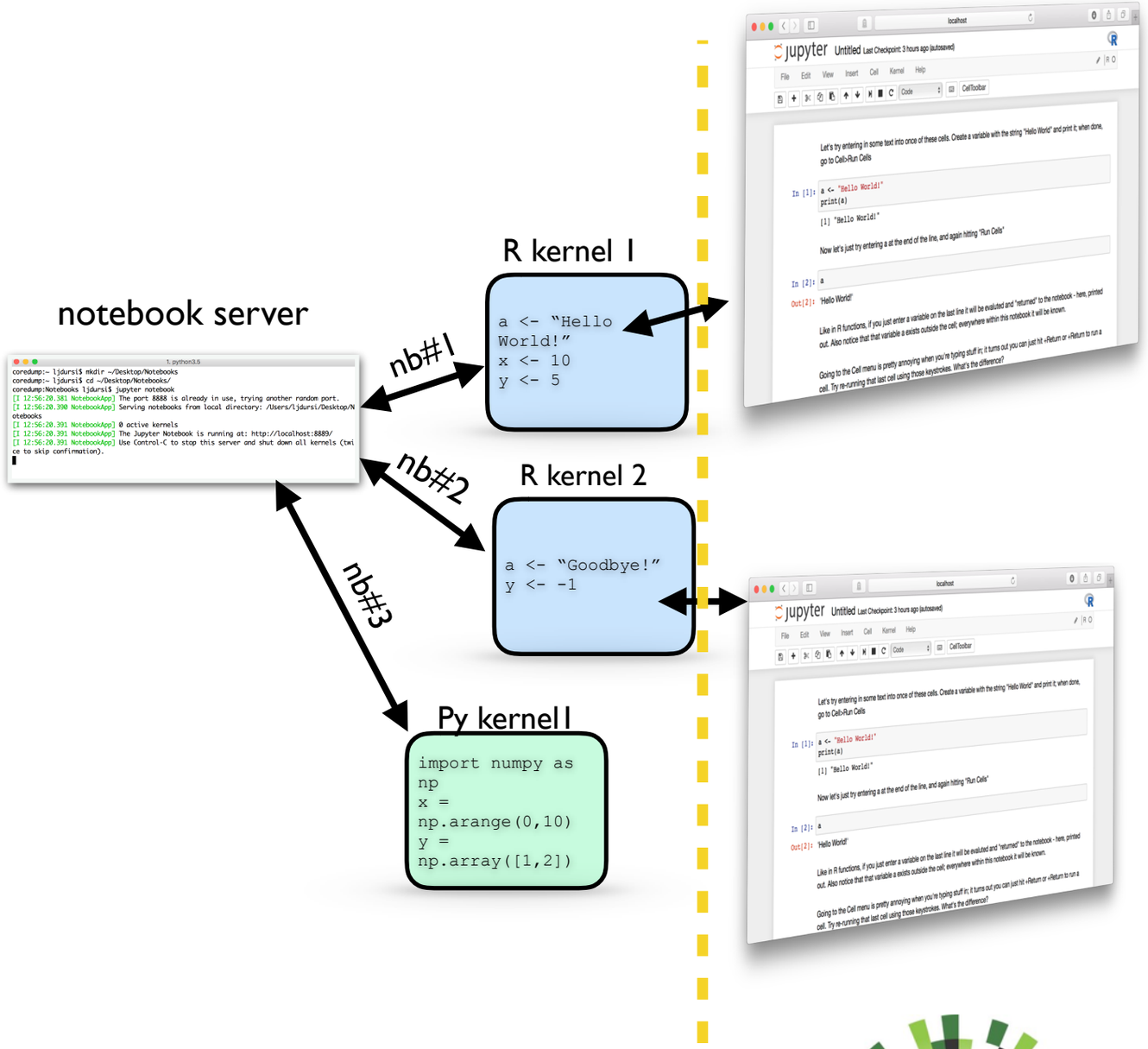
```
import numpy as np
x = np.arange(0,10)
y = np.array([1,2])
```





How does this all work?

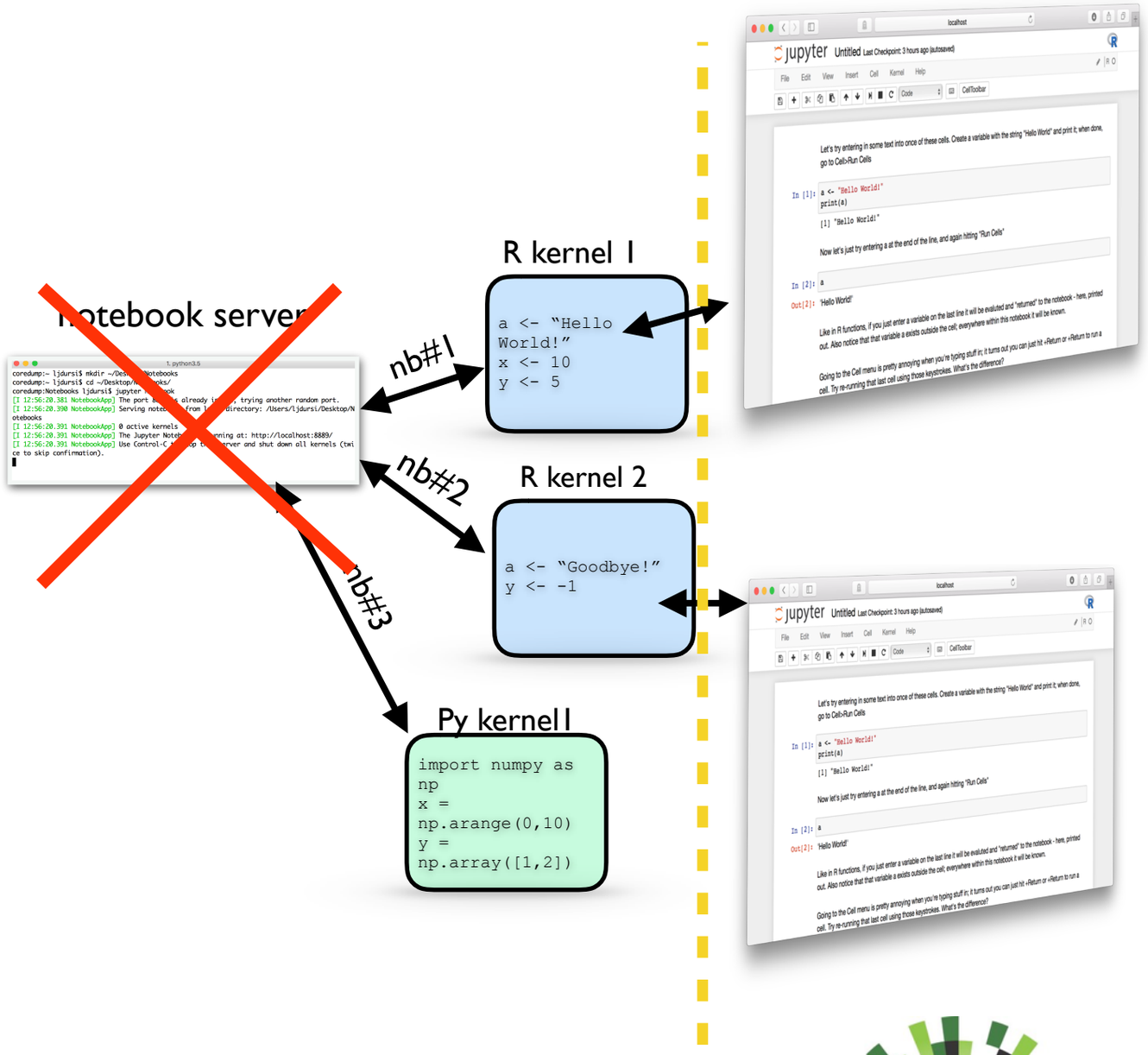
- There's no rule that says the notebook server/kernels has to be on same machine as browser!
  - Can be running on remote server (or cluster)
  - AWS demo



OICR
Ontario Institute for Cancer Research

How does this all work?

- AWS demo…
- What happens if the notebook server dies?
- Demo



OICR
Ontario Institute for Cancer Research

All input+output in ipynb file

jupyter Untitled1 Last Checkpoint: 2 hours ago (autosaved)

File Edit View Insert Cell Kernel Help

Markdown CellToolbar

```
                                        row.names=1,
                                        as.is=TRUE)
```

In [5]:
```
pDataFile = file.path(dataDirectory, "pData.txt")
pData = read.table(pDataFile,row.names=1, header=TRUE, sep="\t")
```

In [6]:
```
head(pData)
```

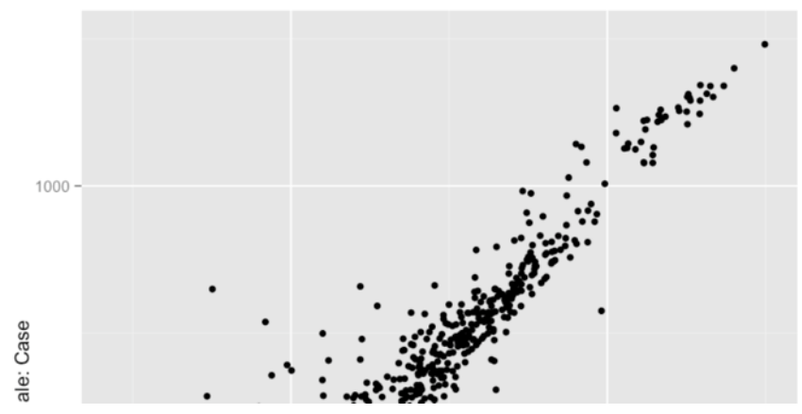
Out[6]:

	gender	type	score
A	Female	Control	0.75
B	Male	Case	0.4
C	Male	Control	0.73
D	Male	Case	0.42
E	Female	Case	0.93
F	Male	Control	0.22

Correlations

What does the correlation between case A and E: a female control and case: look like?

In [7]:
```
library(ggplot2)
ggplot(exprs) + geom_point(aes(x=abs(A), y=abs(E))) + xlab('Female: Control') + ylab('Female: Case') +
   scale_x_log10() + scale_y_log10()
```




```
{
 "cells": [
...
   "outputs": [],
   "source": [
    "data <- read.csv('http://dursi.ca/content/images/GSE37704_featurecounts.csv')\n",
    "metadata <- read.csv('http://dursi.ca/content/images/GSE37704_metadata.csv')"
   ]
  },
...
   },
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "'data.frame':\t19808 obs. of  8 variables:\n",
      " $ ensgene  : Factor w/ 19808 levels \"ENSG00000000003\",..: 15488 19676 19628 19554 19334 15792 16078 15
867 15778 15793 ...\n",
      " $ length   : int  918 718 1982 939 939 3214 5539 3395 2833 3424 ...\n",
      " $ SRR493366: int  0 0 23 0 0 124 1637 120 24 4 ...\n",
      " $ SRR493367: int  0 0 28 0 0 123 1831 153 48 9 ...\n",
      " $ SRR493368: int  0 0 29 0 0 205 2383 180 65 16 ...\n",
      " $ SRR493369: int  0 0 29 0 0 207 1226 236 44 14 ...\n",
      " $ SRR493370: int  0 0 28 0 0 212 1326 255 48 16 ...\n",
      " $ SRR493371: int  0 0 46 0 0 258 1504 357 64 16 ...\n"
     ]
    }

...
   "outputs": [
    {
     "data": {
      "image/png":
      "iVBORw0KGgoAAAANSUhEUgAAA0gAAANICAYAAAD958/bAAAEDWlDQ1BJQ0MgUHJvZmlsZQAO
      I2NVV1oHFUUPrtzZyMkzlNsNIV0qD8NJQ2TVjShtLp/3d02bpZJNtoi6GT27s6Yyc44M7v9oU9
      FUHwx6psUxL+3gCAo9Q/bPrQvlQol2tQgKD60+INQ6Ium65k7M5lpurHeZe58853vnvuuWfvB
```


- All the input+output is located in the .ipynb file

OICR
Ontario Institute
for Cancer Research

Demo - nb formatting

- 2-Making Nice Notebooks

“Hands on”: make a nice notebook

- 3 - Pretty Notebook

nbconvert

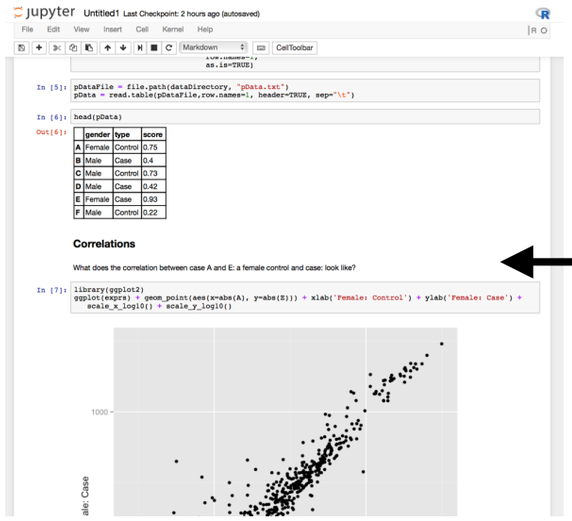
- Once you have a nice notebook, there's a few things you can immediately do with jupyter nbconvert:
  - --to html
  - --to slides
- demo

Demo: Moving scripts to notebook

- 4 - Moving scripts to notebook.ipynb

Sharing

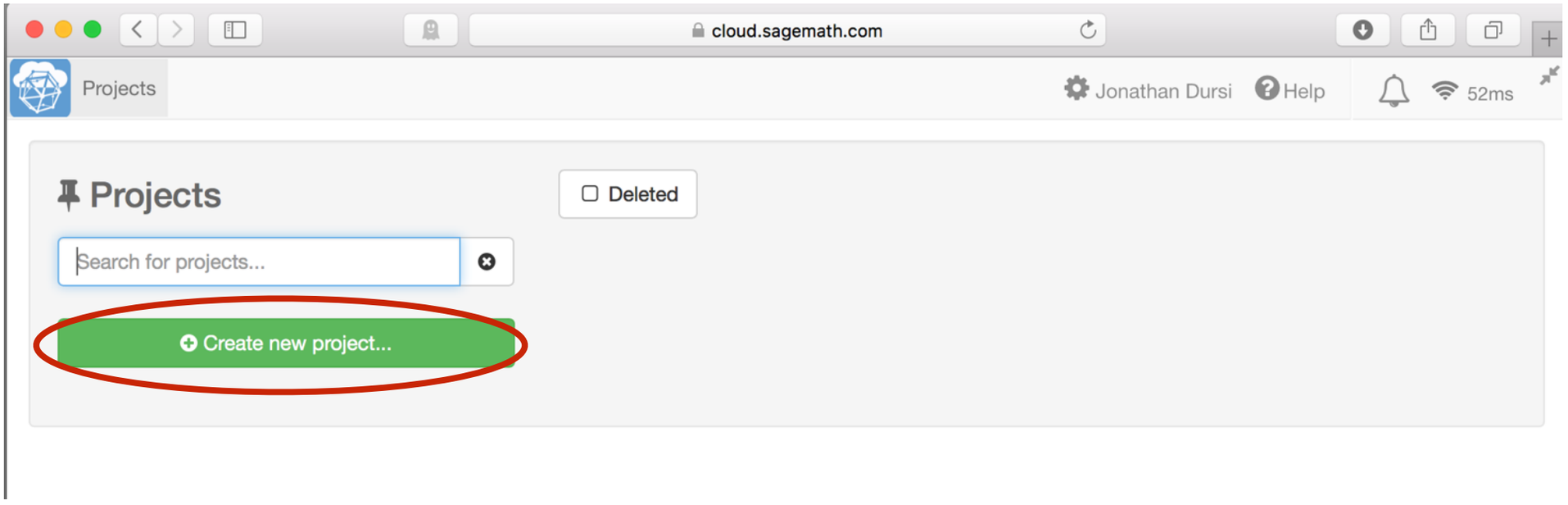
- As we've seen, Jupiter notebooks can be pretty self-contained
- (May refer to local or remote data files, etc)
- For others to view or interact with, need to find a place willing to host +render the files (to view), or provide the compute to re-run the analyses (to reproduce)
- We'll look at:
  - SageMathCloud
  - GitHub, nbviewer
  - MyBinder



OICR
Ontario Institute for Cancer Research

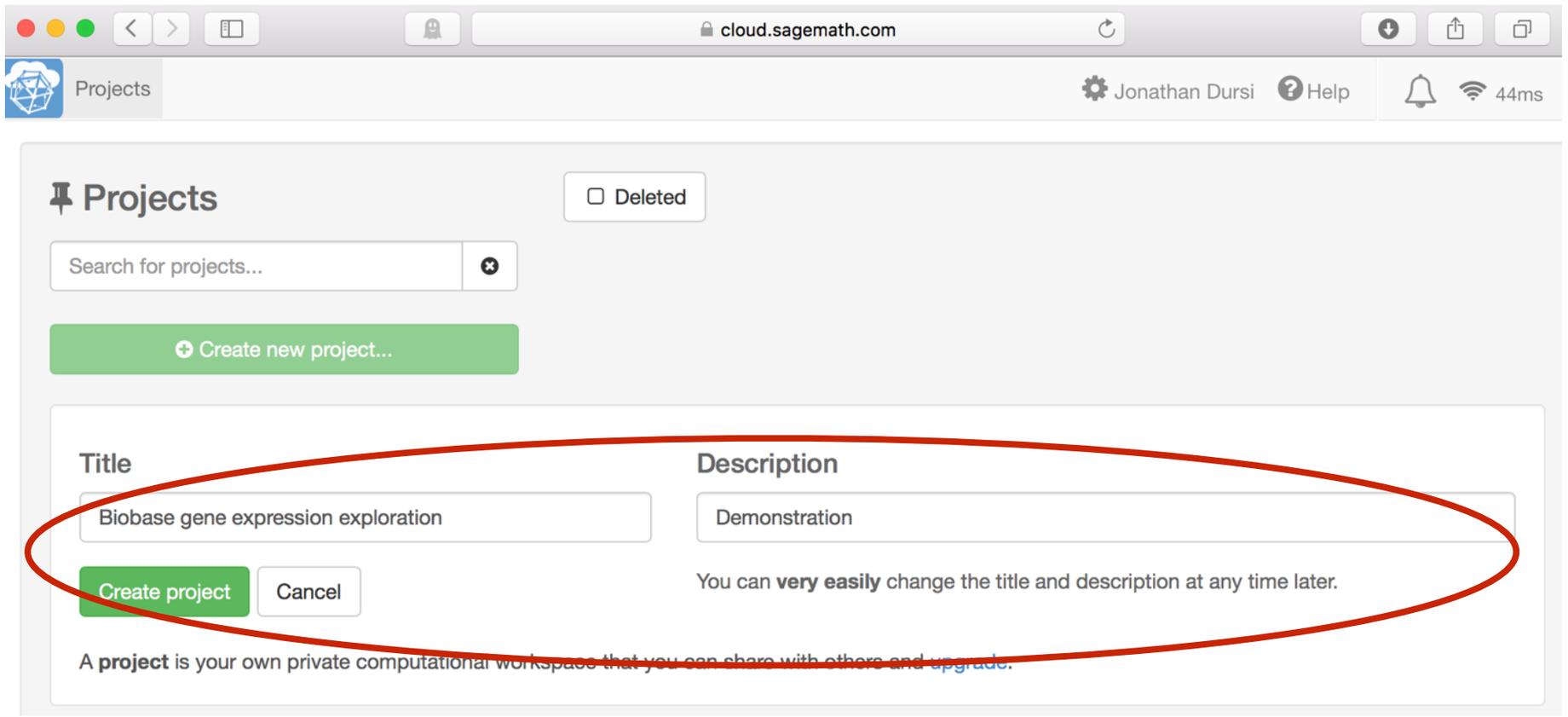
Sharing: SageMathCloud

- Log into SageMathCloud (cloud.sagemath.com); create new project



Sharing: SageMathCloud

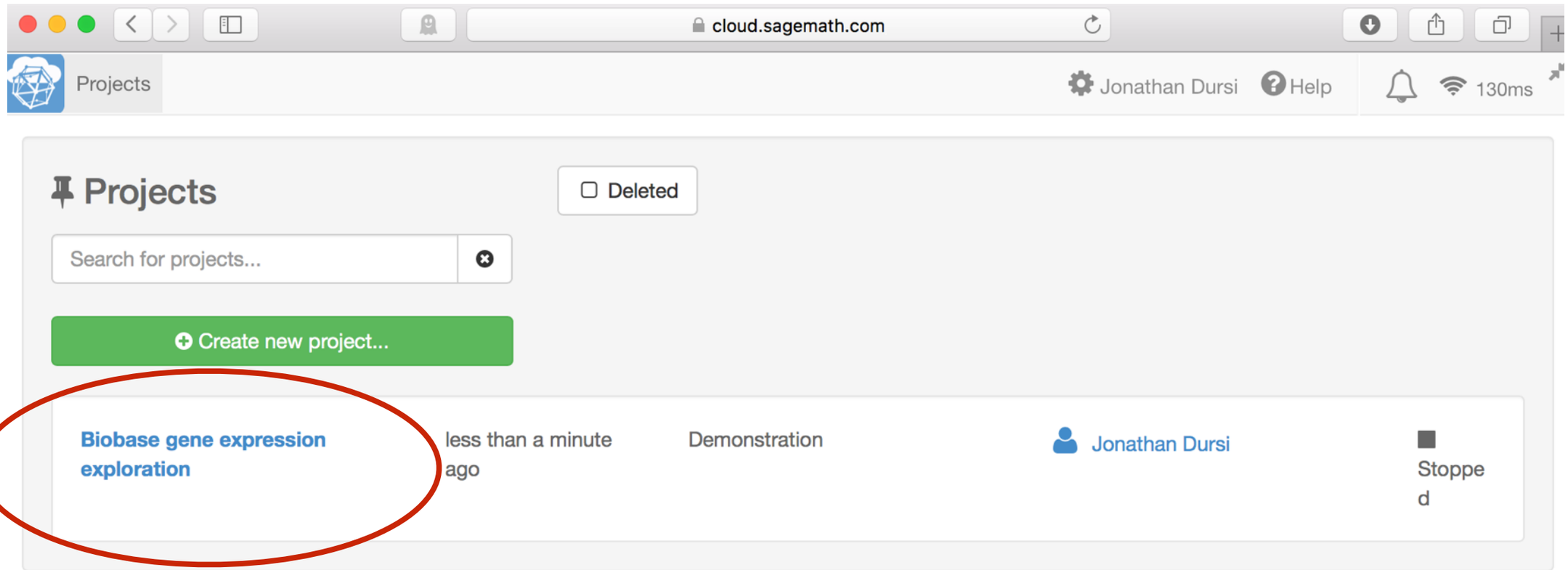
- Log into SageMathCloud (cloud.sagemath.com); create new project



OICR
Ontario Institute for Cancer Research

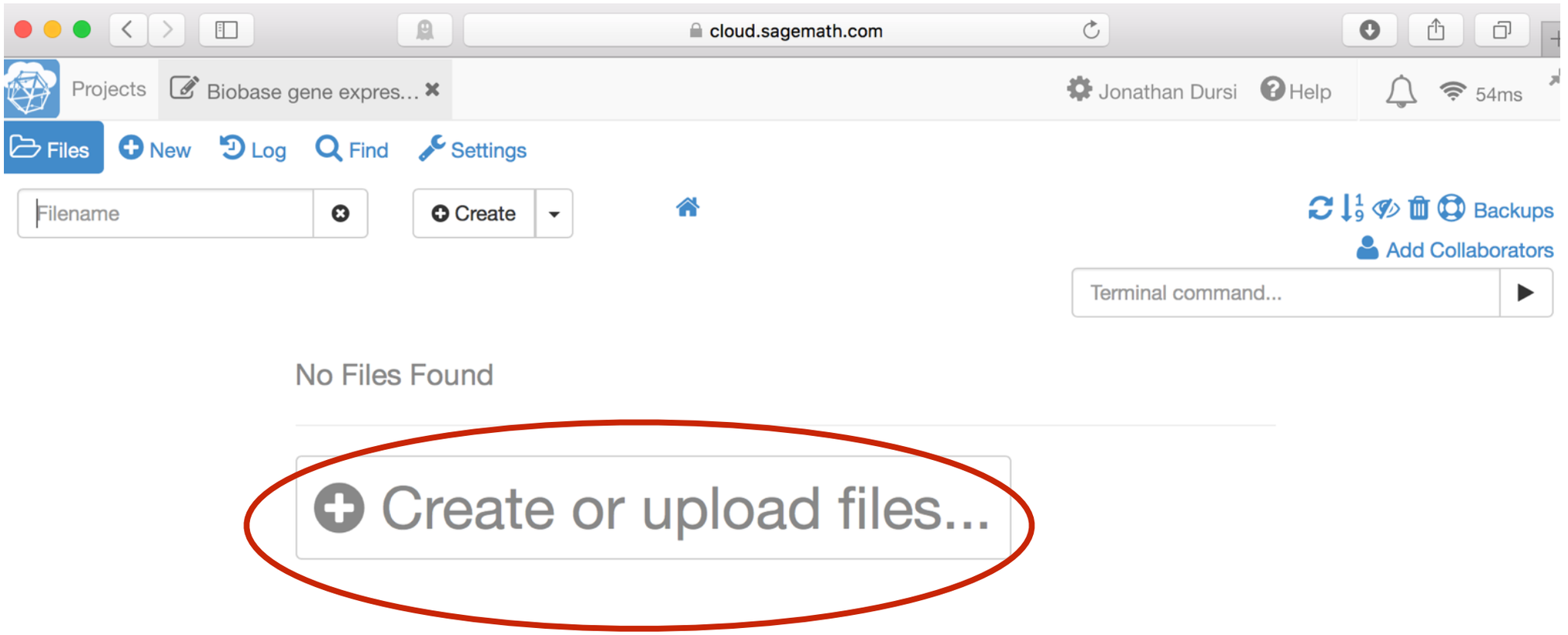
Sharing: SageMathCloud

- Then Enter the project…



Sharing: SageMathCloud

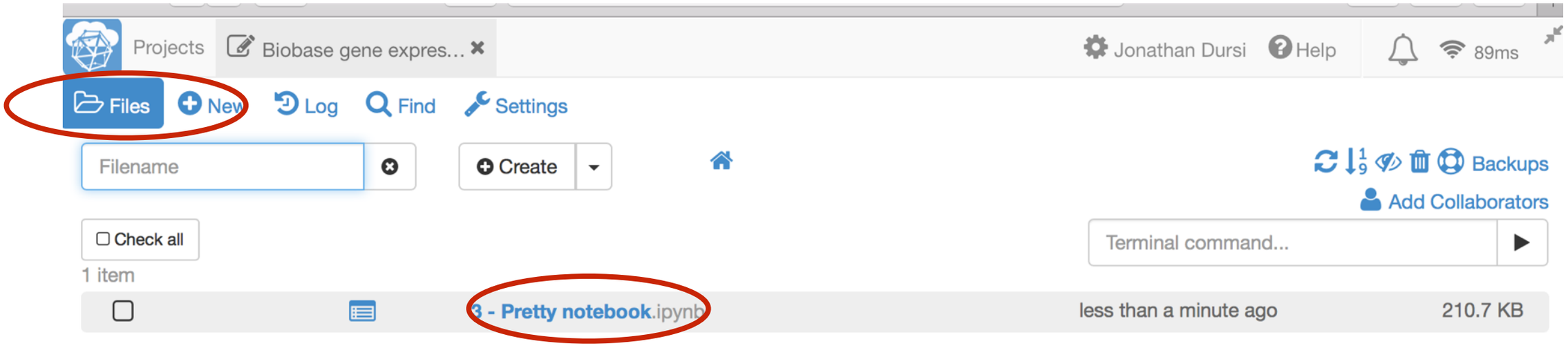
- And upload your notebook



OICR
Ontario Institute
for Cancer Research

Sharing: SageMathCloud

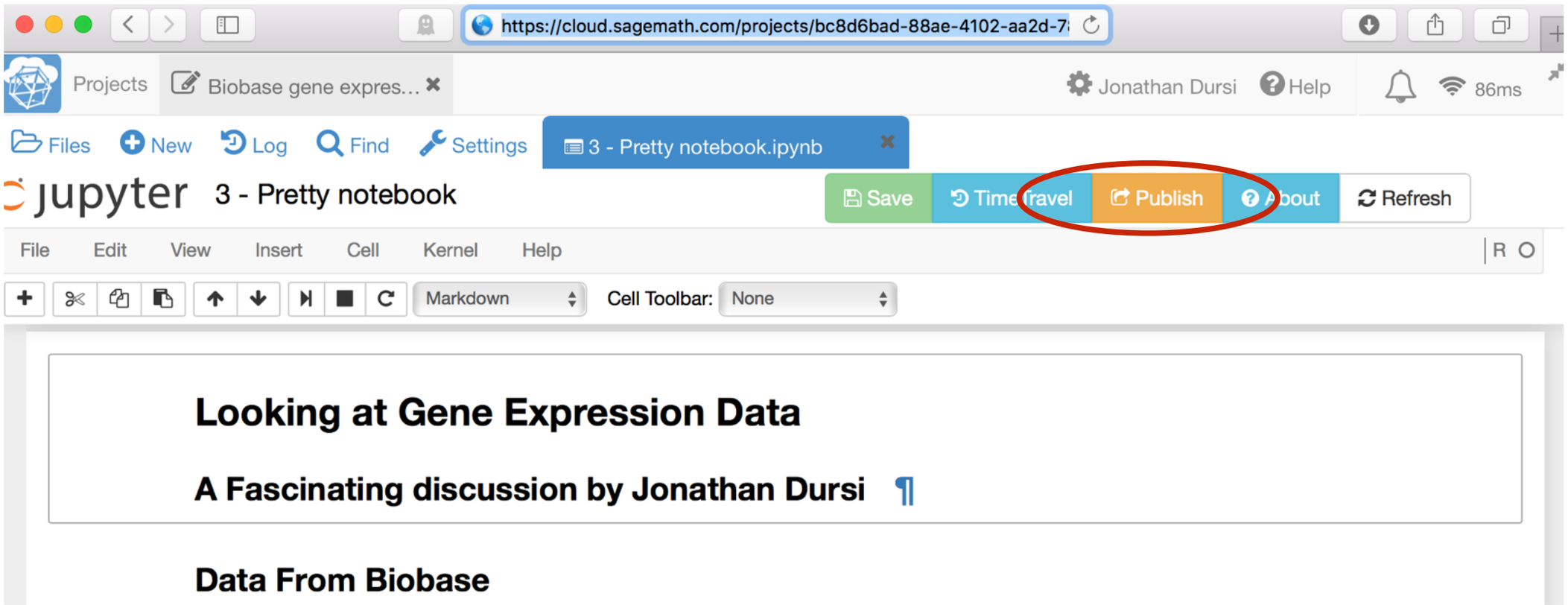
- Then go back to files and launch your notebook



OICR
Ontario Institute
for Cancer Research

Sharing: SageMathCloud

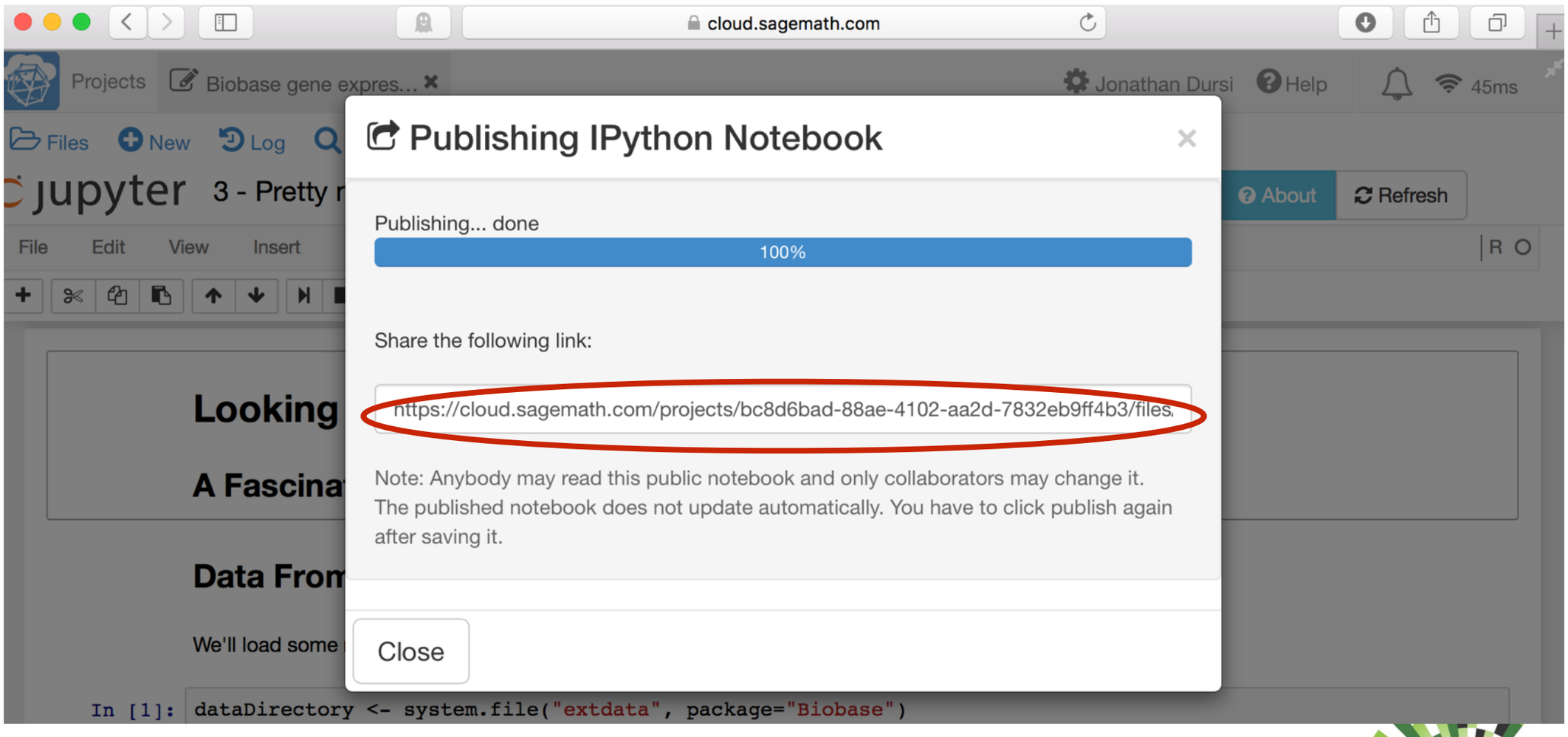
- …and click publish



OICR
Ontario Institute
for Cancer Research

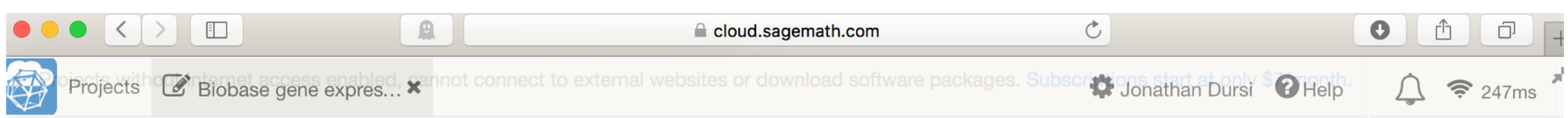
Sharing: SageMathCloud

- You can now share it with anyone to *view*…



Sharing: SageMathCloud

- …Or add collaborators to edit.



Settings and configuration

Title and description

Title: Biobase gene expression exploration

Description: Demonstration

Project usage and quotas

Adjust your quotas...

Collaborators

Collaborators can **modify anything** in this project, except backups. They can add and remove other collaborators, but cannot remove owners.

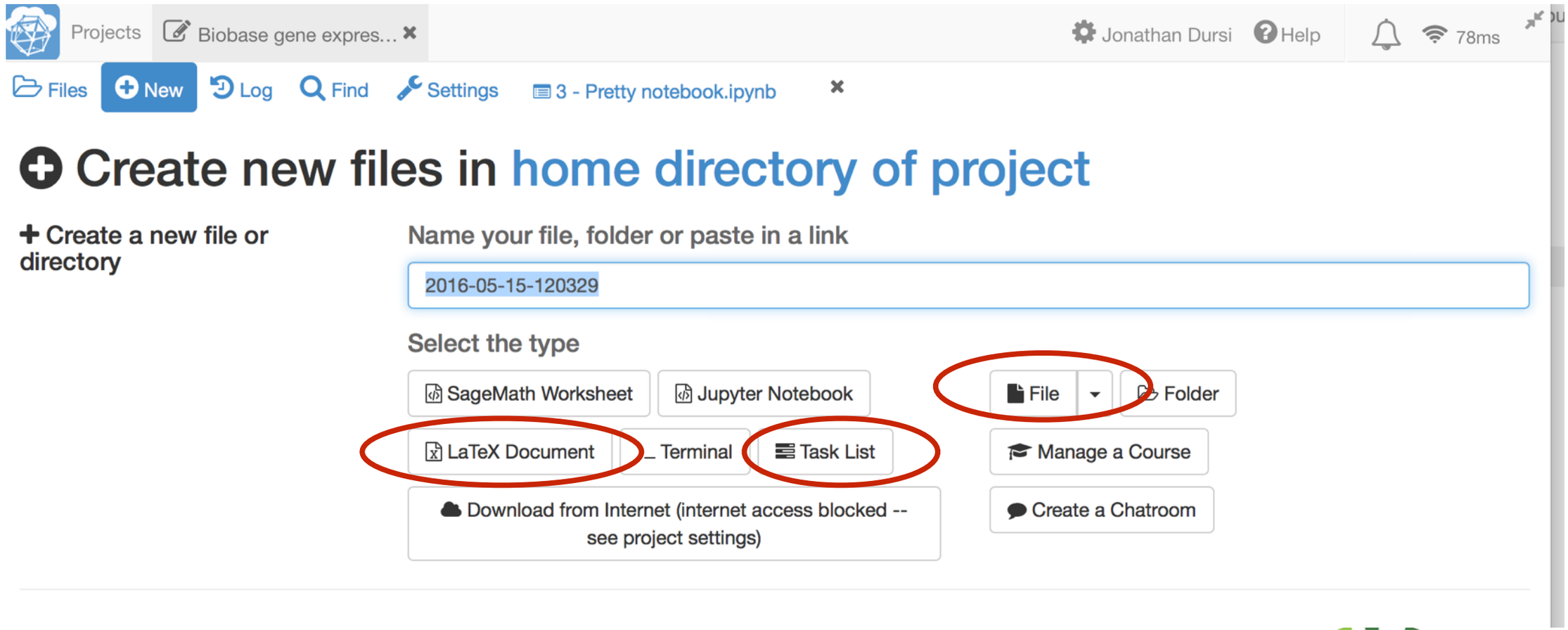
Add collaborators: Search by name or email address...

Jonathan Dursi (6 minutes ago) (owner) — Remove...

OICR
Ontario Institute
for Cancer Research

Sharing: SageMathCloud

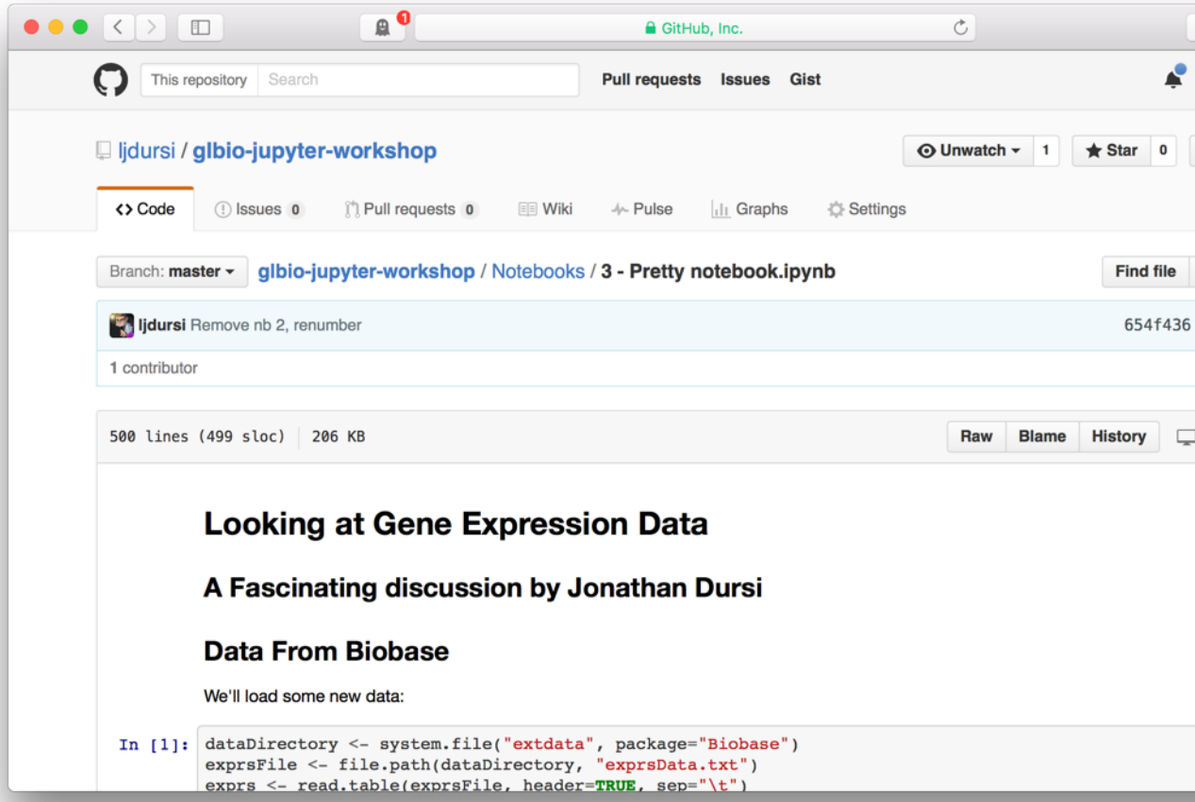
- Can share to-do lists, data files, writeups: but many features require $7/mo subscription



OICR
Ontario Institute for Cancer Research

GitHub

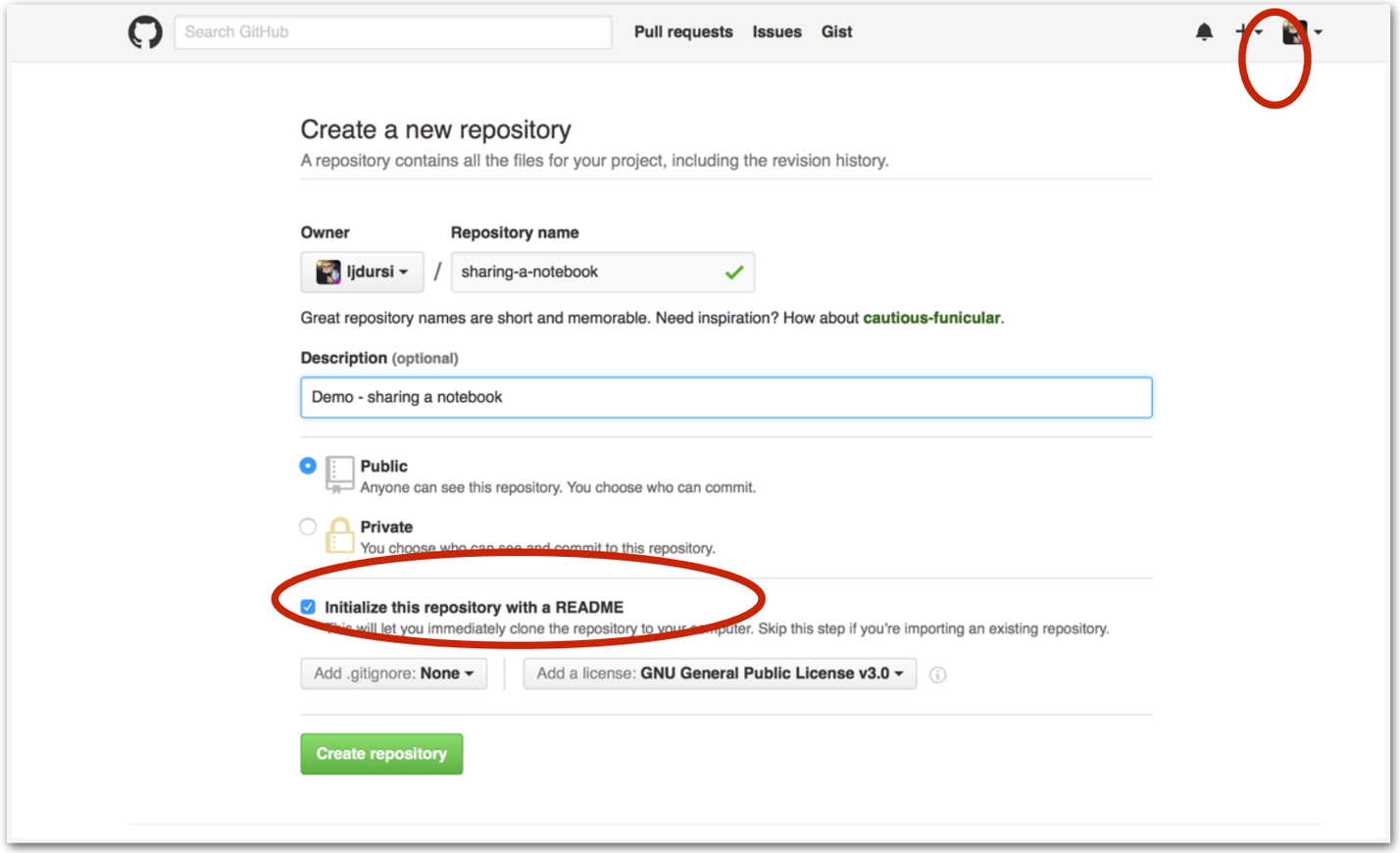
- Widely used for collaborating on scientific software
- Will render uploaded .ipynb files automatically
- Can get unlimited private repositories for researchers: https://education.github.com
- plus other stuff for students



OICR
Ontario Institute for Cancer Research

Sharing in github

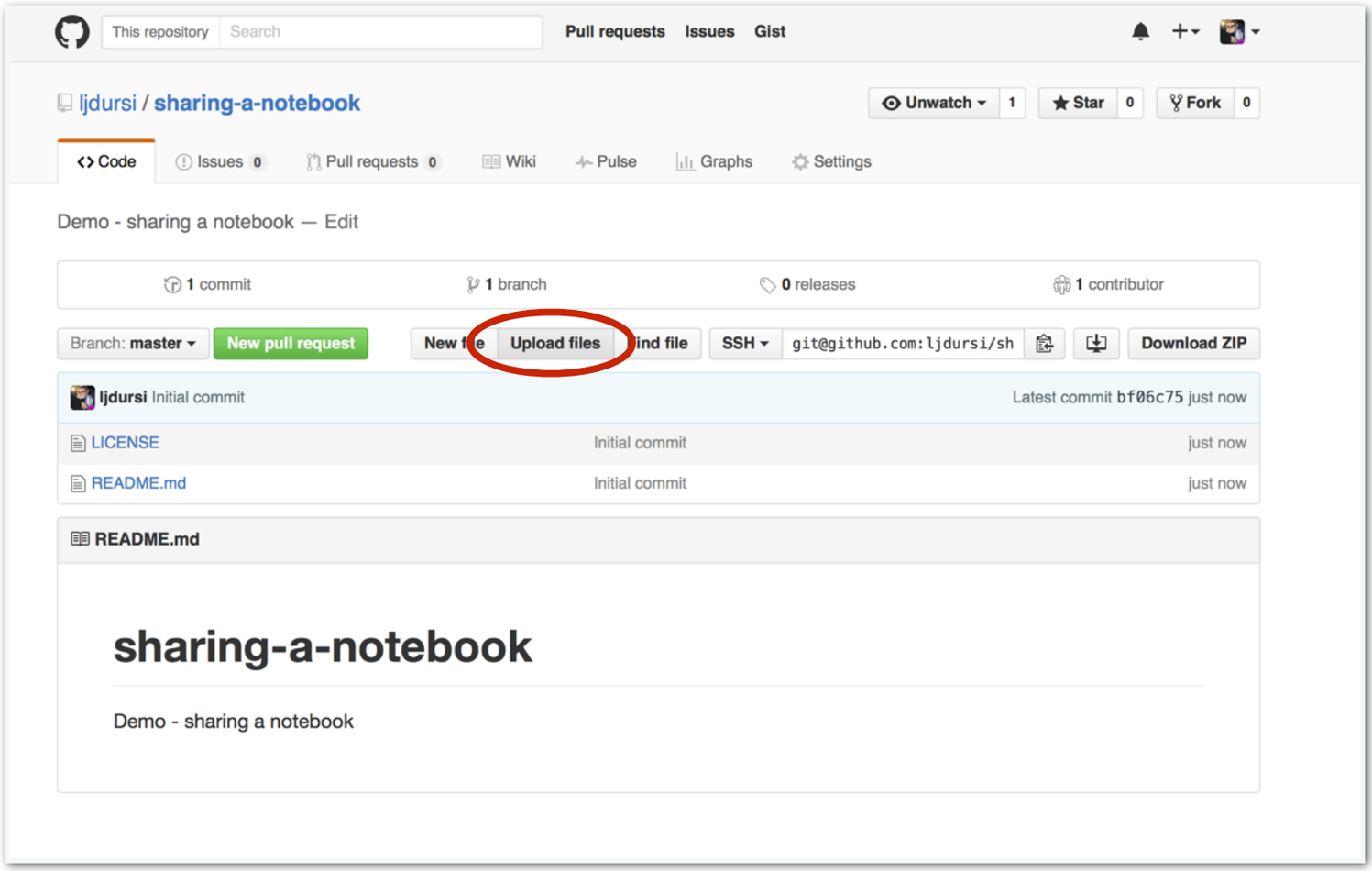
- Easy way: go to github.com and create a repo:
- Fill in the details and initialize a README



OICR
Ontario Institute
for Cancer Research

Sharing in github

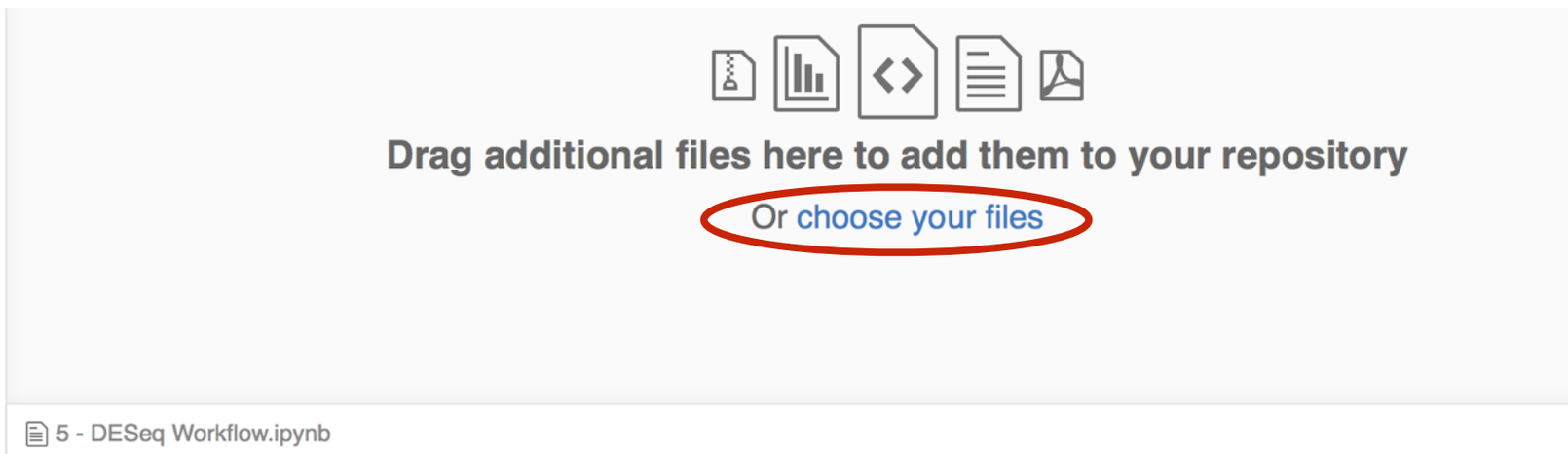
- Now upload your notebook:

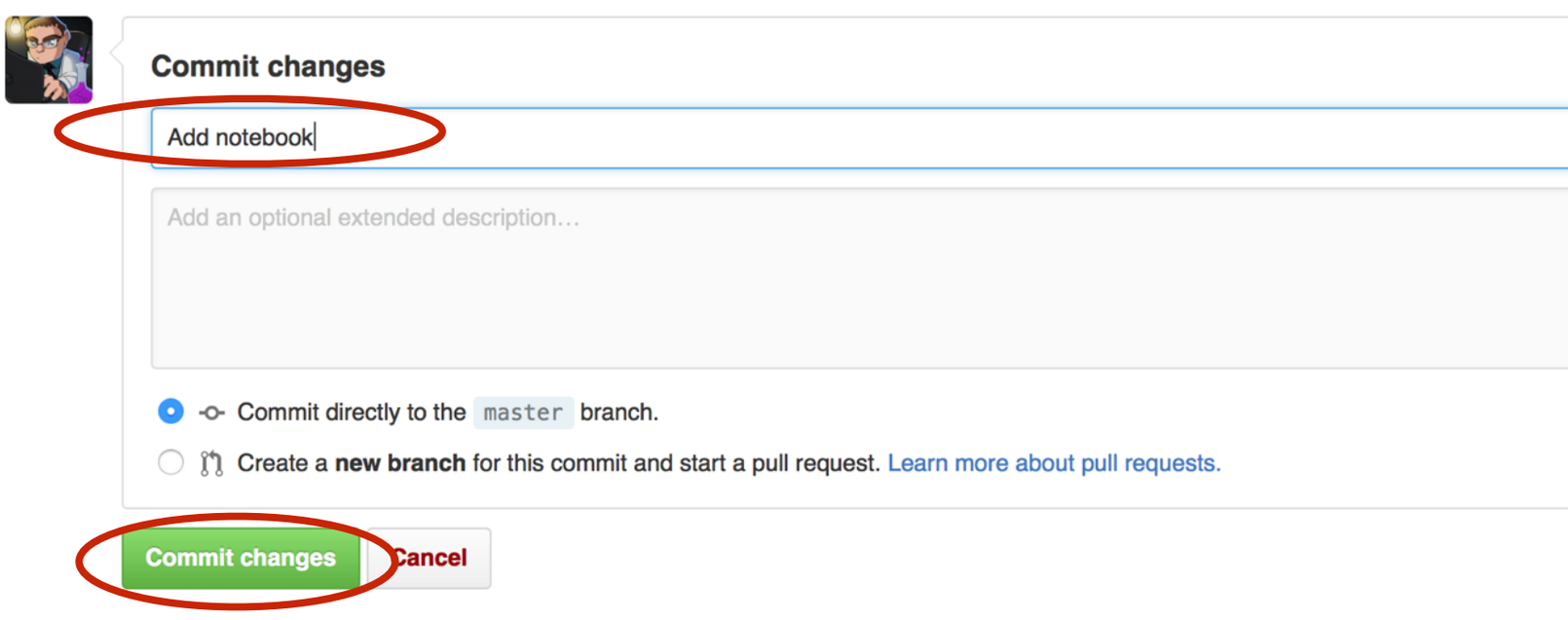


OICR
Ontario Institute
for Cancer Research

Sharing in github

- Now upload your notebook:





OICR
Ontario Institute
for Cancer Research

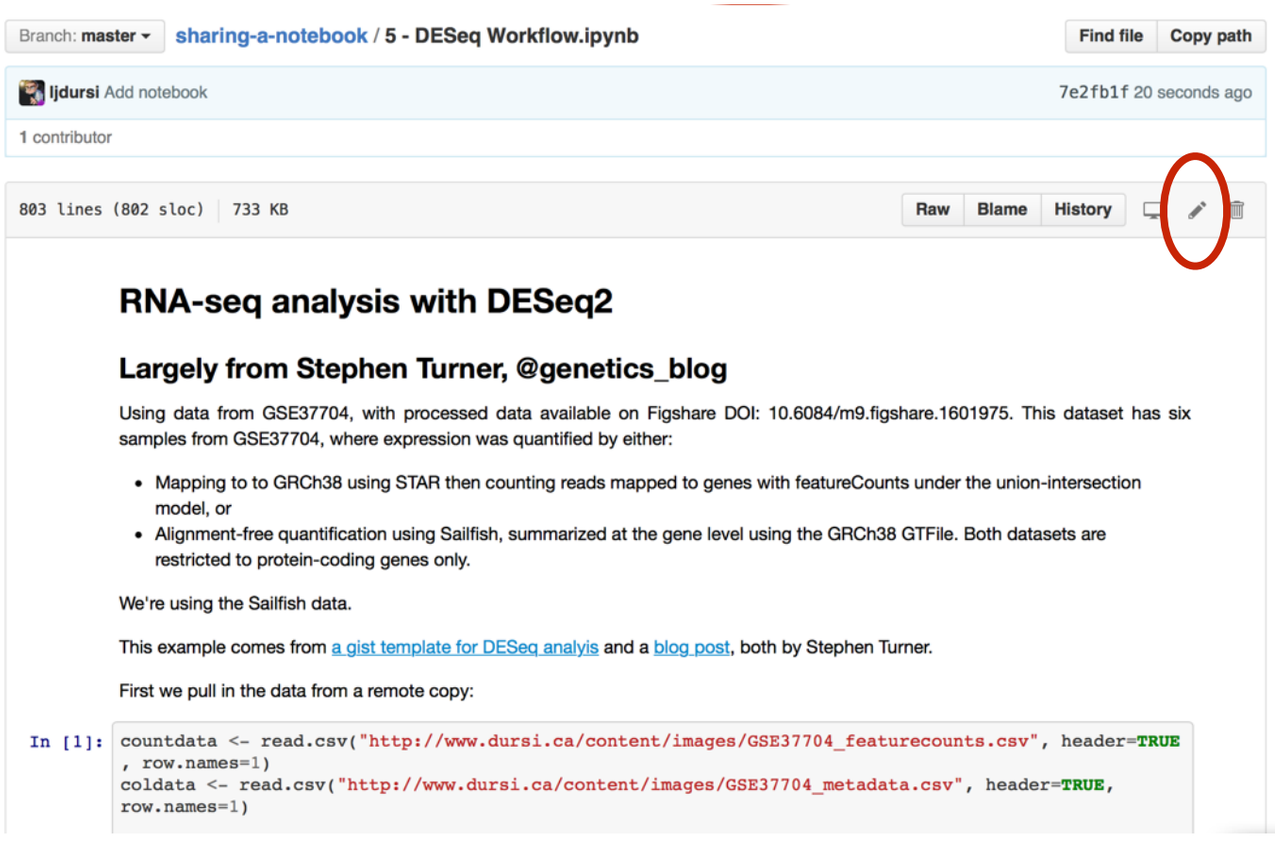
Sharing in github

- And just view the file in the repo

OICR
Ontario Institute
for Cancer Research

Sharing in github

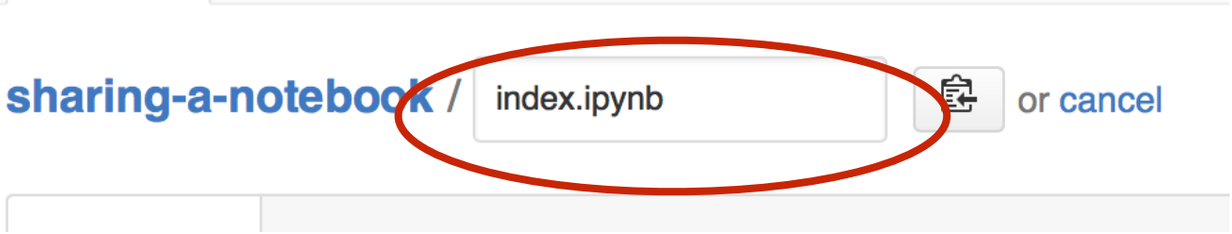
- Let's give it a better name; click on the pencil to edit the file:



OICR
Ontario Institute for Cancer Research

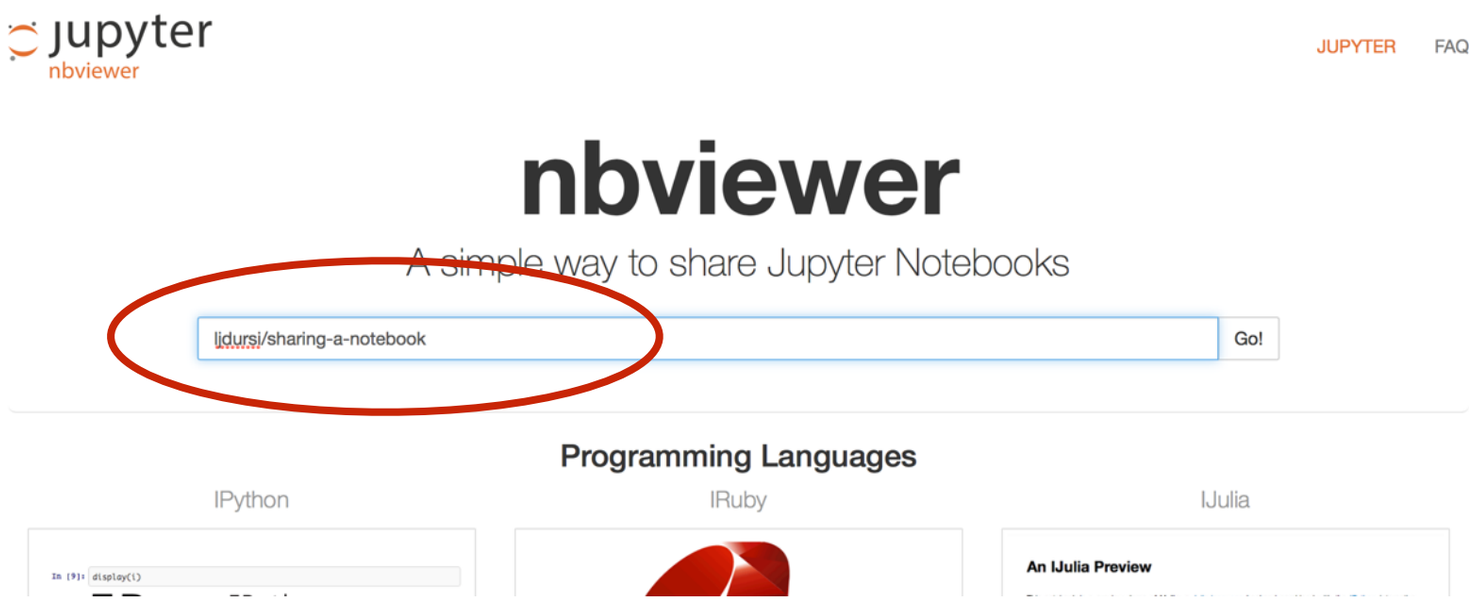
Sharing in github

- and then rename it to index.ipynb



nbviewer.jupyter.org

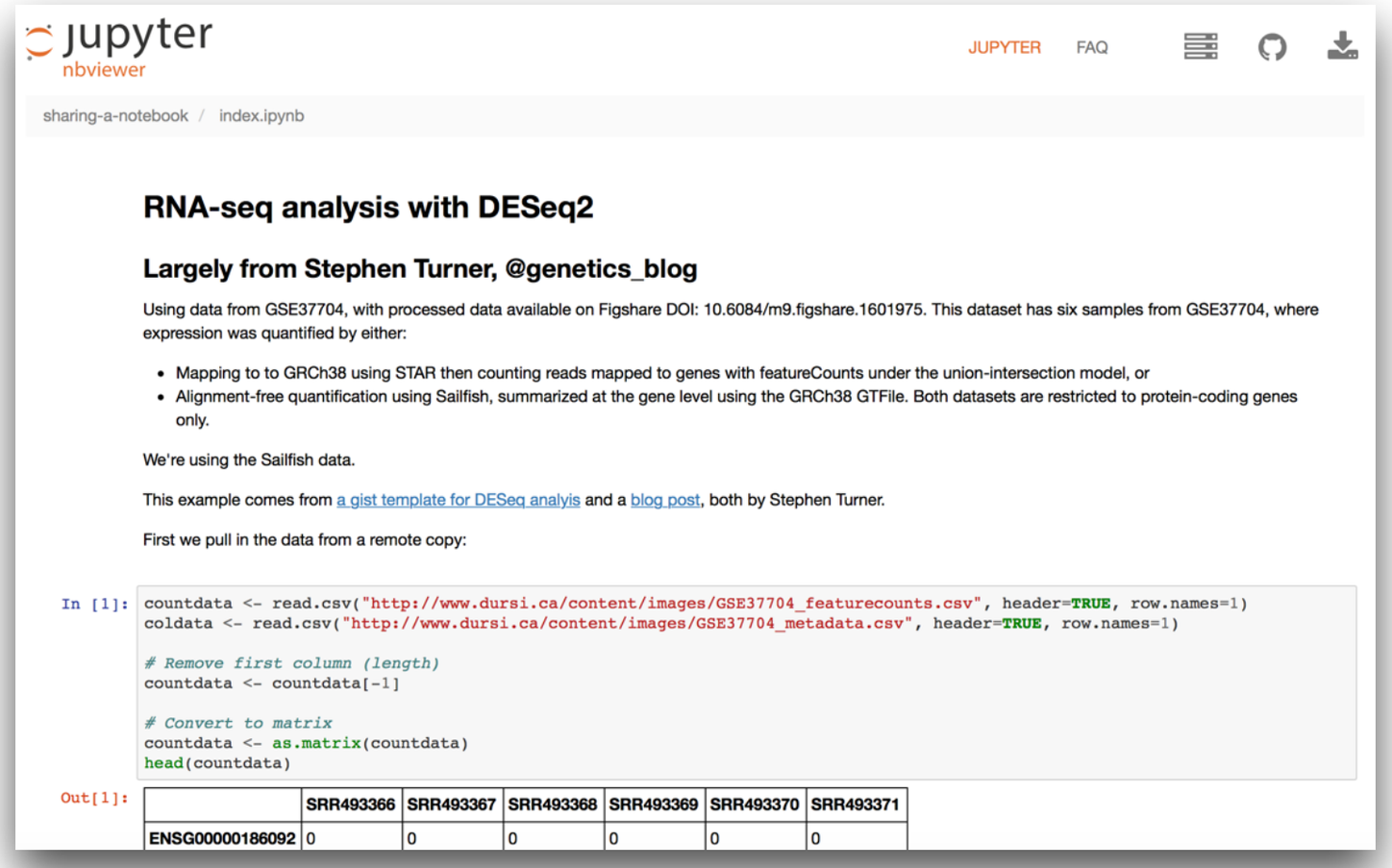
- Once the file is hosted somewhere (github, or anywhere) nbviewer can view it
- Enter name of repo: ljdursi/sharing-a-notebook



OICR
Ontario Institute
for Cancer Research

nbviewer.jupyter.org

- Produces slightly nicer results
- You can share the link it goes to directly, or download the notebook



OICR
Ontario Institute
for Cancer Research