Sharing Reproducible Python Environments with Binder

Sarah Gibson

Pronouns: she/her

@drsarahlgibson @mybinderteam #EuroPython

https://doi.org/10.5281/zenodo.3937310
What is Reproducibility?

@drsarahlgibson @mybinderteam #EuroPython

https://doi.org/10.5281/zenodo.3937310
<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Reproducible</td>
</tr>
<tr>
<td>Different</td>
<td>Robust</td>
</tr>
</tbody>
</table>

Kirstie Whitaker’s talk at PyData LDN: https://youtu.be/IG3PcZ6EhiU
https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html#table-of-definitions-for-reproducibility

@drSarahLGibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Reproducible</td>
</tr>
<tr>
<td>Different</td>
<td>Robust</td>
</tr>
</tbody>
</table>
Barriers to reproducible research

- Held to higher standards than others
- Is not considered for promotion
- Requires additional skills
- Takes time
- Publication bias towards novel findings
- Plead the 5th

Support additional users

Kirstie Whitaker’s talk at PyData LDN: https://youtu.be/IG3PcZ6EhiU

@drsarahlgibson @mybinderteam #EuroPython

https://doi.org/10.5281/zenodo.3937310
Market Research
Have you ever heard…?

“Oh, it worked on my computer?”
Have you ever heard…?

“Oh, it worked yesterday?”
“Oh, it worked on my computer?”

“Oh, it worked yesterday?”

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Binder

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
– Project Binder is a global community
– The mybinder.org service allows anyone to launch a complete, interactive computing environment from their browser

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
gw-openscience.org/tutorials/
JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.

SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS

SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY

SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK

STEP 1

STEP 2

STEP 3

EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS
binder-examples / requirements

<table>
<thead>
<tr>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>numpy==1.16.*</td>
</tr>
<tr>
<td>matplotlib==3.*</td>
</tr>
<tr>
<td>seaborn==0.8.1</td>
</tr>
</tbody>
</table>

 Courtesy of Juliette Taka
https://twitter.com/mybinderteam/status/1082556317842264064

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
name: example-environment
channels:
- conda-forge
dependencies:
- numpy
- psutil
- toolz
- matplotlib
- dill
- pandas
- partd
- bokeh
- dask
JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.

SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS.

SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY.

SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK.

EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS.

Courtesy of Juliette Taka
https://twitter.com/mybinderteam/status/1082556317842264064

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
• Launched by Jeremy Freeman in 2015
• First Binder and Jupyter meeting in January 2017
• First half of 2017 spent redeveloping the backend into what is now BinderHub
• Sept 24, 2017 – Moore Foundation Binder proposal accepted
JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.

She would like anyone to be able to reproduce, check, and improve her calculations.

She publishes them on a publicly hosted repository.

She makes that repository binder-ready by describing the software required to run the notebook.

Everyone can now run and reproduce her computations.

She describes the experiments as a notebook, mixing prose, code, and visualization, and resources: source code, data, and media.

https://doi.org/10.5281/zenodo.3937310

@drsarahgibson @mybinderteam EuroPython
Technology

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Clone GitHub Repo

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
1. Clone GitHub Repo

2. Build image according to instructions contained within the repo

BinderHub

GitHub repository name or URL

https://github.com/alexmorley/binder-demo

Git branch, tag, or commit

Path to a notebook file (optional)

launch

@drsarahlgibson @mybinderteam #EuroPython

https://doi.org/10.5281/zenodo.3937310
1. Clone GitHub Repo
2. Build image according to instructions contained within the repo
3. Execute image

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
1. Clone GitHub Repo
2. Build image according to instructions contained within the repo
3. Execute image
4. Allocate computational resources

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
What is a JupyterHub?

JupyterHub is a way to help your humans use your computers. With notebooks!
BindHub

1. Clone GitHub Repo
2. Build image according to instructions contained within the repo
3. Execute image
4. Allocate computational resources

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
1. Clone GitHub Repo

2. Build image according to instructions contained within the repo

3. Execute image

4. Allocate computational resources

5. Make image accessible at mybinder.org/some_url

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Build and launch a repository

GitHub repository name or URL
https://github.com/alexmorley/binder-demo

Git branch, tag, or commit
Path to a notebook file (optional)

Launch

1. Clone GitHub Repo
2. Build image according to instructions contained within the repo
3. Execute image
4. Allocate computational resources
5. Make image accessible at mybinder.org/some_url
6. Redirect User to mybinder.org/some_url

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Scaling up with the Federation

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Would you recommend mybinder.org to a friend?

- **Yes**: 80%
- **Maybe**: 10%
- **No**: 0%

From 346 responses

https://github.com/sgibson91/mybinder.org-user-survey-nlp

@drsarahlgibson @mybinderteam #EuroPython

https://doi.org/10.5281/zenodo.3937310
What do you (mainly) use mybinder.org for?

- Documentation and examples (19.88%)
- Workshops/training courses (21.69%)
- University teaching (17.47%)
- Reproducible publishing (8.13%)
- Demos and talks (12.35%)
- Sharing and collaborating with a team (7.83%)
- Pre-university teaching (3.01%)
- Other (9.64%)

From 346 responses
https://github.com/sgibson91/mybinder.org-user-survey-nlp
@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
If you could change one thing about Binder, what would it be?

From 346 responses
https://github.com/sgibson91/mybinder.org-user-survey-nlp

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Picking up speed

UPSTREAM

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Community Guidance

We have written advice on best practices for achieving faster launch times

**How to reduce mybinder.org repository startup time**

- Binder
- discuss

**What affects launch time?**

The challenge between running mybinder.org vs. a different cloud service such as Colab is that Binder is meant to run arbitrary environments that you define in a GitHub repository. Most of the time when a repository is (very) slow (more than 30s) to launch it is because the environment for that session must be built and initialized. This mostly happens to people “developing” on a repository (constantly changing things and launching right away).

For most users of a Binder link the environment is already built. This is because someone else has previously launched the same version. This can still be slow but not very slow (more than 30s).

mybinder.org runs on Kubernetes, which runs a cluster that grows and shrinks as necessary to take on new users. Each time a user clicks a Binder link, these things happen:

1. A slot (called a “pod”) is reserved on one of the cloud machines
2. Binder looks to see if a Docker image exists for that repository  
   - If it doesn’t, Binder must first build the image for that repo using repo2docker (this takes time)
3. Binder looks for a built image on the machine the user will use  
   - If it isn’t on the machine, Binder must first pull the image onto that machine (this takes time)
4. Binder launches the user’s session

**How can I reduce my launch time?**

With that being said, in order to reduce the amount of time it takes your repository to launch, try these steps:

- **Make your repository environment more light-weight** - A repository with fewer dependencies and a smaller size will be faster to both build and download into the Binder session.
- **Ensure your repository gets a lot of clicks** - The more often a repository is launched, the more likely it will already be built and downloaded to a machine when a user starts a new session. As a result, the more popular a repository is, the faster launches will tend to take.
- **Use two repositories: one for the environment, one for your content** - Many people change their content much more often than they change the environment needed for it. However, Binder will re-build the environment for any changes to a repository. A hack to get around this is to define an “environment repository” that Binder builds, and use a hook to pull in new content at launch from a “content repository”. This means that your “environment repository” changes less-often, which should result in fewer new builds and reduced launch times. See the instructions in this post to get started.
- **Use the nbgitpuller.link page to automate separate content/environment repos**. The above step can be (mostly) automated by using nbgitpuller.link. This is a little web form...

https://discourse.jupyter.org/t/how-to-reduce-mybinder-org-repository-startup-time/4956

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310
Community

– Value meeting communities where they are
– Diversify our skills
Find out more and get involved

– GitHub: https://github.com/jupyterhub/binderhub
– Website: https://mybinder.readthedocs.io and https://mybinder.org
– Discourse: https://discourse.jupyter.org/c/binder
– Gitter: https://gitter.im/jupyterhub/binder
– Twitter: @mybinderteam

@drsarahlgibson @mybinderteam #EuroPython
https://doi.org/10.5281/zenodo.3937310