@BaumerBen: Reproducibility, or the idea that the same people should be able to reproduce the same analysis with the same data, is such a low bar... and we're still tripping over it. #JSM2018

10:58 AM - 31 Jul 2018
Why should you care about reproducibility?

Because the person most likely to need to reproduce your work... is you.
code + data + environment = reproducible machine learning
code + data + environment = reproducible machine learning
Levelling up reproducibility

1. Code
   a. Structuring your project
   b. Stochastic -> static
2. Data
3. Environment
Most machine learning methods rely on some sort of pseudorandomness for things like:

- Weight initialization
- Dropout
- Subsetting/shuffling for mini-batches
- Training/testing/validation split
Stochastic -> Static

In order to get the same results more than once, you need to make sure to set *all* the random number generators (RNGs) your code depends on.

- **Numpy & Keras:**
  - `np.random.seed(42)` (*will bork on multi-threading*)
- **Tensorflow:**
  - `tf.set_random_seed(42)`
- **Anything that uses hash randomization (like Theano):**
  - `PYTHONHASHSEED=0` (*this can be a security risk so don’t do it by default*)
- **cuDNN**
  - `¯\_(ツ)_/¯`
  - Some routines (like backward pass & backward pooling) do not guarantee reproducibility because they use atomic operations
Levelling up reproducibility

1. Code
2. Data
   a. What should you be saving?
   b. Always work from a copy
   c. Version your data
3. Environment
Always work from a copy of your data

Keep a separate copy of your raw data that you never, ever touch. Work from copies of it.
Versioning, not just for code anymore!

- If you’re already using version control & have small/medium data, add your data files to the system you use for your code.
- For databases, there are many version control options (versionSQL, DBmaestro, etc.).
- For streaming data, save & work from a specific time span.
Levelling up reproducibility

1. Code
2. Data
3. Environment
You can think of reproducibility as a scale: The longer it takes to reproduce a project, the less reproducible it is.
Sharing your environment

Your environment includes:

- All dependencies, including versions and subversions
- Your file system
- Your OS
- (In some cases) hardware

Options for sharing your environment:

- Containers (like Docker)
- Virtual machines
- Hosted environments
Sharing your environment: Containers

Benefits:

- Contains data, code, file systems, dependencies
- Portable
- Lightweight

Drawbacks:

- Uses the host OS, can be dicey cross-platform
- Can take a while to get set up
Sharing your environment: Virtual machines

Benefits:

- Contains data, code, file systems, dependencies and OS
- Portable, even between platforms

Drawbacks:

- Larger/slower to get started than containers
- Can take a while to get set up
Sharing your environment: Hosted environments

Benefits:

- Very fast set-up
- Extremely easy to share (in many cases just copying & pasting a link)

Drawbacks:

- Less control over environment
- May not be feasible for security/privacy reasons
<table>
<thead>
<tr>
<th>Name</th>
<th>Price</th>
<th>Languages</th>
<th>GPU</th>
<th>Data hosting</th>
<th>Specs (Free tier)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaggle Kernels</td>
<td>Free</td>
<td>Python 3, R</td>
<td>Yes</td>
<td>Yes</td>
<td>GPU: 1xTesla K80 (6 hr/run)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAM: 16 GB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Disk: 5 GB</td>
</tr>
<tr>
<td>Google Colaboratory</td>
<td>Free</td>
<td>Python 2 &amp; 3</td>
<td>Yes</td>
<td>No</td>
<td>GPU: 1xTesla K80 (12 hr/run)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAM: ~12.6 GB Available</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Disk: ~33 GB Available</td>
</tr>
<tr>
<td>Azure Notebooks</td>
<td>Free</td>
<td>Python 2 &amp; 3, R, F#</td>
<td>No</td>
<td>Yes</td>
<td>4G memory limit &amp; 1G data limit</td>
</tr>
<tr>
<td>Amazon SageMaker</td>
<td>Varies, 2 month</td>
<td>Python 2 &amp; 3</td>
<td>Yes</td>
<td>No</td>
<td>250 hours/month of t2.medium notebook</td>
</tr>
<tr>
<td></td>
<td>free trial</td>
<td></td>
<td></td>
<td></td>
<td>50 hours/month of m4.xlarge</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>125 hours/month of m4.xlarge</td>
</tr>
<tr>
<td>IBM Watson Studio</td>
<td>Varies, limited</td>
<td>Python, R, Scala</td>
<td>Yes</td>
<td>Yes</td>
<td>1 vCPU and 4 GB RAM, 50 hours run-time per month</td>
</tr>
<tr>
<td></td>
<td>free tier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codalab</td>
<td>Free</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>non-GPU machine has 4 cores and 14 GB of memory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>and each GPU machine has 6 cores and 56 GB of memory</td>
</tr>
<tr>
<td>MyBinder</td>
<td>Free</td>
<td>Python 2 &amp; 3, R, Julia</td>
<td>No</td>
<td>No</td>
<td>CPU: x86-64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAM: ~13 GB (est)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Disk: ~100 GB (est)</td>
</tr>
<tr>
<td>Name</td>
<td>Price</td>
<td>Languages</td>
<td>GPU</td>
<td>Data hosting</td>
<td>Specs (Free tier)</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>-----------------</td>
<td>-----</td>
<td>--------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Kaggle Kernels</td>
<td>Free</td>
<td>Python 3, R</td>
<td>Yes</td>
<td>Yes</td>
<td>GPU: 1xTesla K80 (6 hr/run) RAM: 16 GB Disk: 5 GB</td>
</tr>
<tr>
<td>Google Colaboratory</td>
<td>Free</td>
<td>Python 2 &amp; 3</td>
<td>Yes</td>
<td>No</td>
<td>GPU: 1xTesla K80 (12 hr/run) RAM: ~12.6 GB Available Disk: ~33 GB Available</td>
</tr>
<tr>
<td>Azure Notebooks</td>
<td>Free</td>
<td>Python 2 &amp; 3</td>
<td>No</td>
<td>Yes</td>
<td>4G memory limit &amp; 1G data limit</td>
</tr>
<tr>
<td>Amazon SageMaker</td>
<td>Varies, 2 month free trial</td>
<td>Python 2 &amp; 3</td>
<td>Yes</td>
<td>No</td>
<td>250 hours/month of t2.medium notebook RAM: 50 hours/month of m4.xlarge 125 hours/month of m4.xlarge 250 hours/month of m4.xlarge 125 hours/month of m4.xlarge 250 hours/month of m4.xlarge 125 hours/month of m4.xlarge 250 hours/month of m4.xlarge</td>
</tr>
<tr>
<td>IBM Watson Studio</td>
<td>Varies, limited free tier</td>
<td>Python, R, Scala</td>
<td>Yes</td>
<td>Yes</td>
<td>1 vCPU and 4 GB RAM, 50 hours run-time per month</td>
</tr>
<tr>
<td>Codalab</td>
<td>Free</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>non-GPU: CPU has 4 cores and 14 GB of memory, and each GPU machine has 6 cores and 56 GB of memory</td>
</tr>
<tr>
<td>MyBinder</td>
<td>Free</td>
<td>Python 2 &amp; 3, R, Julia</td>
<td>No</td>
<td>No</td>
<td>CPU: x86-64 RAM: ~13 GB (est) Disk: ~100 GB (est)</td>
</tr>
</tbody>
</table>

Hosting services change their services from time to time, so be sure to double check the before starting your project!
Slides: https://goo.gl/4kqNzq
Contact: rachael@kaggle.com
@rctatman