Motivation for investigating reproducibility

- Reports of a “reproducibility crisis” in science


- Need for reproducibility:
  - Helps avoid disaster… and move science forward
  - Makes it easier to publish papers
  - Helps you get your point across
  - Enables continuity of your work
  - Helps build your reputation, e.g. attracts more citations

Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015 Dec 8;16:274.

Method

- 5 analysts reviewed the proceedings of the BioNLP workshop in 2016 and an additional sample of papers from ACL and journals to assess the availability of code and data:

<table>
<thead>
<tr>
<th>Position</th>
<th>Degree</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>PhD (Linguistics)</td>
<td>Biomedical Natural Language Processing</td>
</tr>
<tr>
<td>Faculty</td>
<td>PhD (Computer Science)</td>
<td>Biomedical Natural Language Processing</td>
</tr>
<tr>
<td>Graduate Student</td>
<td>MSc (Computer Science)</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>Faculty</td>
<td>PhD (Maths)</td>
<td>Computational Biology</td>
</tr>
<tr>
<td>Research Staff</td>
<td>MBA</td>
<td>Annotation Project Management</td>
</tr>
</tbody>
</table>

- Guidelines for assessment instructed analysts to:
  - Search abstract, then full paper
  - Collect evidence of availability
  - Criteria for code availability included:
    - (unbroken) website location, e.g. github
    - Statement of code availability from the authors
  - Criteria for data availability included:
    - (unbroken) website location, e.g. github
    - Dataset from a shared task (a list was provided)
    - Other common open dataset (a list was provided)

Discussion

There is a general effort to make data and code available. Data is more readily available than code.
Authors genuinely trying to make data and/or code available sometimes do not realize their attempt to share fails.

We need to do better than 25% possibly reproducible research!

Define necessary information and strive to encourage complete reporting

Document, document, document:
- Datasets need to be documented in more details to facilitate replication and reuse
- Code availability needs to extend to data pre-processing and evaluation toolkits
- Link obsolescence continues to be an issue

Data and code availability does not guarantee reproducibility, however data and code availability is necessary to achieve reproducibility.