Reproducibility in Research: Insight from experiments in Computer Science and beyond
Outline

What are the challenges in reproducibility?
What do we gain by aiming for reproducibility?
How can reproducibility be achieved?
Replicability, reproducibility, repeatability

Are these terms equivalent?

A definition:

- “Independently running a research experiment and yielding the same results on each iteration”

🛠️ Reproducibility is the essence of science
Reasons to work reproducibly

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.
Challenges in Reproducibility

Reports of a reproducibility crisis in many disciplines
– Poll of 1,500 scientists in 2016

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Failed to reproduce others’ experiment</th>
<th>Failed to reproduce own experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>90%</td>
<td>60%</td>
</tr>
<tr>
<td>Biology</td>
<td>80%</td>
<td>60%</td>
</tr>
<tr>
<td>Physics and engineering</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>Medicine</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>Earth and environment science</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>Other</td>
<td>60%</td>
<td>50%</td>
</tr>
</tbody>
</table>

How is this possible?

Data is often unavailable
- e.g. medical data due to confidentiality
- Software due to commercial strategy
- Seemingly insignificant details are left out of protocols

Reporting bias
- Space limitation in papers (e.g. conference papers in computer science)
- Novelty is valued more than reproducibility
Learning from reproducibility (or lack thereof)

The tale of the Zigglebottom tagger

Variability lies in…

- Pre-processing (what is being pre-processed?)
  - Tokenization
  - Stop-word lists
  - “Data cleaning”, e.g. normalization of case, diacritics
- Software versions, system variations
- Parameters, including training/test split


Variability on corpus: GRACE

Counting « words »

Counting « sentences »
Standardization and Documentation

- Standardized components, procedures, workflows
- Documenting complete system set-up across entire provenance chain

How to do this – efficiently?

Alexander Graham Bell’s Notebook, March 9 1876
https://commons.wikimedia.org/wiki/File:Alexander_Graham_Bell's_notebook,_March_9,_1876.PNG
The Shared Task Model

Primary goal is to provide a forum for direct comparison of approaches

- Availability of shared material
- Specific definition of a “task”
- Corpora and annotations, split into training, development and test sets
- Evaluation metrics and scripts
Shared Tasks Examples

Information Retrieval and information extraction
– MUC, TREC, CLEF, CTCIR

Computational Linguistics
– Semeval, GRACE, EASY, DEFT

Translation
– WMT

BioNLP, curation
– i2b2, BioCreAtive, BioASQ
The PRIMAD$^1$ model: which attributes can we “prime”?

Defining Types of Reproducibility

- Data
  - Parameters
  - Input data
- Platform
- Implementation
- Method
- Research Objective
- Actors

What do we gain by priming one or the other?

## Types of Reproducibility and Gains

<table>
<thead>
<tr>
<th>Label</th>
<th>Data Parameters</th>
<th>Raw Data</th>
<th>Platform / Stack</th>
<th>Implementation</th>
<th>Method</th>
<th>Research Objective</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Determinism</td>
</tr>
<tr>
<td>Param. Sweep</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Robustness / Sensitivity</td>
</tr>
<tr>
<td>Generalize</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Applicability across different settings</td>
</tr>
<tr>
<td>Port</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Portability across platforms, flexibility</td>
</tr>
<tr>
<td>Re-code</td>
<td>-</td>
<td>-</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>Correctness of implementation, flexibility, adoption, efficiency</td>
</tr>
<tr>
<td>Validate</td>
<td>(x)</td>
<td>(x)</td>
<td>(x)</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>Correctness of hypothesis, validation via different approach</td>
</tr>
<tr>
<td>Re-use</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>Apply code in different settings, Re-purpose</td>
</tr>
<tr>
<td>Independent x (orthogonal)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>Sufficiency of information, independent verification</td>
</tr>
</tbody>
</table>
Levels of reproducibility (in computer science)

1. **Availability:** the system and data it was tested on must be available (or there must be sufficient detail available to reconstruct the system and dataset).

2. **Builds:** the code must build.

3. **Runs:** the built code must run.

4. **Evaluation:** it must be possible to run on the same data and measure the output using the same implementation of the same scoring metric.
The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements.

Obtain workflows from MyExperiments.org

- March 2015: almost 2.700 WFs (approx. 300-400/year)
- Focus on Taverna 2 WFs: 1.443 WFs

Try to re-execute the workflows

- Majority of workflows fails
- Only 23.6 % are successfully executed
  (correctness of results not checked yet)

Computer Science

613 papers in 8 ACM conferences

Process

– download paper and classify
– search for a link to code (paper, web, email twice)
– download code
– build and execute

Reproducibility track at

- An automatic coding task
- 4 analysts aim to reproduce participants runs

Overall, results can be reproduced, but...

- Replication is not easy
- No analyst was able to replicate every run
- Documentation shortcomings reported

I heard you need to create a TPS Report. Here, I've got an R script that does that already.

Why does your script keep dying?

Oh, you need to download these 5 packages first.

I did, and it still doesn't work!

Well, it worked when I wrote it 3 weeks ago.

Grr. Package updates...

Source: a parody of xkcd

Studied 2 R libraries

- Needed to contact authors to use successfully
- Produced extra documentation and test cases

Take Home message:
Aim at achieving reproducibility

At different levels

- Re-run, ask others to re-run
- (Re-implement)
- (Port to different platforms)
- Test on different data, vary parameters (and report!)

If something is not reproducible -> investigate!
(you might be onto something)
Plan your research procedure

- Define a protocol
- Have a data management plan

Document, document, document

- the research process, environment, interim results, …
Acknowledgements

- Andreas Rauber (Vienna University of Technology)
- Kevin B. Cohen (University of Colorado)
- Cyril Grouin (LIMSI-CNRS), Aude Robert (INSERM/CépiDC)
- Patrick Paroubek and Pierre Zweigenbaum (LIMSI-CNRS)

CABeRneT ANR-13-JS02-0009-01
CLEF initiative