Reproducibility and ML: We are probably thinking about this wrong....

Carl Kesselman
University of Southern California
How do we create a scientific result

Any time scientists disagree, it's because we have insufficient data. Then we can agree on what kind of data to get; we get the data; and the data solves the problem. Either I'm right, or you're right, or we're both wrong. And we move on. That kind of conflict resolution does not exist in politics or religion.

Neil deGrasse Tyson

• Science is about **communities** arguing over data
  – How do those communities form
  – How do communities argue: knowledge capture and communication
There be dragons.....

Data leakage causes reproducibility failures in ML-based science

The running list below consists of papers that highlight reproducibility failures or pitfalls in ML-based science. We find 20 papers from 17 fields where errors have been found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. In each case, data leakage causes errors in the modeling process.

<table>
<thead>
<tr>
<th>Im. pers viewed</th>
<th>Num. papers w/pitfalls</th>
<th>Pitfalls</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>No train-test split</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>No train-test split; Feature selection on train and test set</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Duplicates across train-test split; Sampling bias</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Pre-processing on train and test sets together</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>No train-test split</td>
<td></td>
</tr>
</tbody>
</table>
Feynman says.....

• a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty — a kind of leaning over backwards. For example, if you’re doing an experiment, you should report everything that you think might make it invalid, not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you’ve eliminated by some other experiment, and how they worked — to make sure the other fellow can tell they have been eliminated
Flip the Paradigm....

• **Data Centric vs Compute Centric**
  – [Data-Centric Biology](#) A Philosophical Study Sabina Leonell

• **Data Centric Architectures**
  – Data Centric Discovery with a Data-Oriented Architecture (Kesselman 2015)

• **Data Centric Workflows**
  – A Data-Centric Design Methodology for Business Processes (Bhattacharya, 2009)
  – A framework for collecting provenance in data-centric scientific workflows (Simmhan, Plale, Gannon, 2006)
FAIR Data is **essential** to good scholarship

- Findable, Accessible, Interoperable, Reusable
- Requires culture and technology
  - Socio-technical approach
  - See: Sharing Begins at Home (doi: 10.1162/99608f92.44d21b86)
- Broad issues with policy, privacy, security, IP.
  - There are significant non-protected data sets.
Addressing Gaps in the Data Sharing Ecosystem...

“Domain or Specialized” Repositories

PRO: Highly detailed descriptive information; High quality data
CON: Narrow focus; High cost of biocuration

“Generalist” Repositories

PRO: All types of data and science; Highly scalable
CON: Minimal structure for detail on data; Quality concerns
Addressing Gaps in the Data Sharing Ecosystem...

“Domain or Specialized” Repositories

PRO: Highly detailed descriptive information; High quality data
CON: Narrow focus; High cost of biocuration

“Generalist” Repositories

PRO: All types of data and science; Highly scalable
CON: Minimal structure for detail on data; Quality concerns

“Hybrid” Repositories

Addresses Gaps: Flexibility to adapt to new species and assays, with minimum viable information for reusability; “Self-serve” style of data curation with structure to guide scientists to produce quality (meta)data.

- GEO: Gene Expression Omnibus
- dbGaP: GENOTYPES and PHENOTYPES
- ZFIN
- ENCODE
- ScholarSphere
- FaceBase: A Resource for Craniofacial Researchers
- Mendeley Data
- Zenodo

Information Sciences Institute
Example: Mapping the Synatome
Regional synapse gain and loss accompany memory formation in larval zebrafish

William P. Dempsey, Zhuwei Du, Anna Nadtochiy, and Don B. Arnold

Edited by Bernardo Sabatini, Department of Neurobiology, Harvard Medical School, Boston, MA; received April 23, 2021; accepted December 3, 2021

January 14, 2022 | 119 (3) e2107661119 | https://doi.org/10.1073/pnas.2107661119

Significance

Imaging of labeled excitatory synapses in the intact brain before and after classical conditioning permits a longitudinal analysis of changes that accompany associative
Neuronal activation within the anterolateral pallium in response to the CS in learner fish and to the US in naive fish. (A) Intense immunostaining of pERK in the pallium (magenta highlighted region, inset) of an L fish exposed to 5 CSs following TFC. The strong signal in an anterolateral region (yellow outline) of this optical section reveals regional neuronal activation. Relatively less immunostaining is present in the medial pallium (cyan outline). (B) An NL fish shows a lack of pERK staining in the anterolateral region (yellow outline) after exposure to 5 CSs in this equivalent optical section. (C) A naive fish reveals strong pERK staining in the same anterolateral region (yellow outline) after exposure to 10 USs. Equivalent optical section to those in A and B. (D) A naive fish exposed to 10 CSs does not show concentrated pERK labeling in the anterolateral region (yellow outline). Optical section equivalent to those in A–C. (E) A naive fish not exposed to a CS or US (NS) does not show concentrated pERK labeling in the anterolateral region (yellow outline). Optical section equivalent to those in A–D. (F) L and US-exposed naive subjects show a significantly higher lateral/medial pERK intensity ratio compared to NL and naive untreated subjects (*P < 0.02, ***P < 0.005, n = 5 fish per group, Kruskal–Wallis multiple comparison test). White dashed lines mark the border of the pallium (midline = M) in A–E. (Scale bar for A–E, 20 μm.) Data available at https://doi.org/10.25555/ijji-113IP6 (50).
File: 1-1JP0: Fig. 2: Neuronal activation within the anterolateral pallium in response to the CS in learner fish and to the US in naïve fish.

Record ID: 1-1JP0
Permanent ID: https://doi.org/10.25551/1/1-1JP0
Title: Fig. 2: Neuronal activation within the anterolateral pallium in response to the CS in learner fish and to the US in naïve fish.
Authors: William Dempsey, Zhuowei Du, Anna Nadtochiy, Karl Czajkowski, Colton Smith, Andrey Andreev, Drew Robson, Jennifer Li, Sarina Applebaum, Thai Truong, Carl Kesselman, Scott Fraser, Don Arnold
Year: 2020

Description:
A
B
C
D
E
F

CS-Learner
CS-Nonlearner
US-Naive
CS-Naive
NS-Naive
## Dataset

**Dataset**: 1-1F6W: Phosphorylated ERK Immunostaining Experiments (Figure 2, 6, Extended Data Figure 2)

<table>
<thead>
<tr>
<th>View</th>
<th>Details</th>
<th>Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Record" /></td>
<td><strong>Record</strong> 1-1T8P&lt;br&gt;<strong>Subject</strong> 1-1T7E&lt;br&gt;<strong>Volume</strong> 0.008 L&lt;br&gt;<strong>FScope</strong> 2018-12-12_3Lvl_FScope3&lt;br&gt;<strong>Grade</strong> bhv_no_aversion</td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td><img src="image" alt="Record" /></td>
<td><strong>Record</strong> 1-1T8G&lt;br&gt;<strong>Subject</strong> 1-1T7J&lt;br&gt;<strong>Volume</strong> 0.008 L&lt;br&gt;<strong>FScope</strong> 2017-06-28_3Lvl_FScope2&lt;br&gt;<strong>Grade</strong> bhv_no_aversion</td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td><img src="image" alt="Record" /></td>
<td><strong>Record</strong> 1-1T8E&lt;br&gt;<strong>Subject</strong> 1-1T7C&lt;br&gt;<strong>Volume</strong> 0.008 L&lt;br&gt;<strong>FScope</strong> 2017-06-28_3Lvl_FScope2&lt;br&gt;<strong>Grade</strong> bhv_no_aversion</td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td><img src="image" alt="Record" /></td>
<td><strong>Record</strong> 1-1T8G&lt;br&gt;<strong>Subject</strong> 1-1T7G</td>
<td><img src="image" alt="Plot" /></td>
</tr>
</tbody>
</table>
### Image Pair Study®: 1-1YRA

<table>
<thead>
<tr>
<th>Image 1</th>
<th>1-03MJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image 2</td>
<td>1-01BA</td>
</tr>
<tr>
<td>Nucleic Region 1</td>
<td>1-1YNW</td>
</tr>
<tr>
<td>Nucleic Region 2</td>
<td>1-1YNT</td>
</tr>
</tbody>
</table>

**Alignment**

```
[[-0.9994621779220547, 0.004642105303365061, -0.03246237455348768, 0.9717546838549336], [-0.006126096091920411, 0.99893334398487295, ...
```

**Region 1 URL**

[ImagePair_1-1YRA_n1_registered.csv](ImagePair_1-1YRA_n1_registered.csv)

**Region 2 URL**

[ImagePair_1-1YRA_n2_registered.csv](ImagePair_1-1YRA_n2_registered.csv)

**Plot**

![Plot of nuclei](plots.png)