Replication

Hao Tang

Why is this necessary?

- Replication is part of the scientific method.
- Replication is difficult.

Robert Boyle

Quotes from Robert Boyle

I have divers times in cases, where the Experiments seem'd like to be thought strange, or to be distrusted, set down several Trials of the same thing, that they might mutually support and confirm one another.

(Boyle, 1665).

Quotes from Robert Boyle

. . . in the serious and effectual prosecution of Experimental Philosophy, I must add one discouragement more, which will perhaps as much surprize you as dishearten you; and it is, That besides that you will **find . . . many of the Experiments publish'd by Authors, or related to you by the persons you converse with, false or unsuccessful**, . . . you will meet with several Observations and Experiments, which though communicated for true by Candid Authors or undistrusted Eye-witnesses, or perhaps recommended to you by your own experience, may upon further tryal disappoint your expectation, either not at all succeeding constantly, or at least varying much from what you expected.

(Boyle, 1668)

Quotes from Robert Boyle

. . . in the serious and effectual prosecution of Experimental Philosophy, I must add one discouragement more, which will perhaps as much surprize you as dishearten you; and it is, That besides that you will **find . . . many of the Experiments publish'd by Authors, or related to you by the persons you converse with, false or unsuccessful**, . . . you will meet with several Observations and Experiments, which though communicated for true by Candid Authors or undistrusted Eye-witnesses, or perhaps recommended to you by your own experience, may upon further tryal disappoint your expectation, either not at all succeeding constantly, or at least varying much from what you expected.

(Boyle, 1668)

Quotes from Robert Boyle

. . . **the great variety in the number, magnitude, position, figure, &c.** of the parts taken notice of by Anatomical Writers in their dissections of that one Subject the humane body, about which many errors would have been delivered by Anatomists, if the frequency of dissections had not enabled them to discern betwixt those things that are generally and uniformly found in dissected bodies, and those which are but rarely, and (if I may so speak) through some wantonness or other deviation of Nature, to be met with.

(Boyle, 1668)

Quotes from Robert Boyle

. . . **try those Experiments very carefully, and more than once, upon which you mean to build considerable Superstructures either theoretical or practical, and to think it unsafe to rely too much upon single Experiments**, especially when you have to deal in Minerals: for many to their ruine have found, that what they at first look'd upon as a happy Mineral Experiment has prov'd in the issue the most unfortunate they ever made.

(Boyle, 1668)

Louis Pasteur

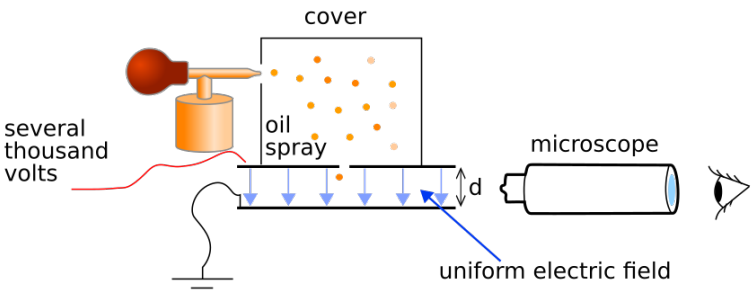
Louis Pasteur

A French national hero at age 55, in 1878 **Pasteur discreetly told his family never to reveal his laboratory notebooks to anyone**. His family obeyed, and all his documents were held and inherited in secrecy. Finally, in 1964 Pasteur's grandson and last surviving male descendant, Pasteur Vallery-Radot, donated the papers to the French national library. Yet the papers were restricted for historical studies until the death of Vallery-Radot in 1971. The documents were given a catalogue number only in 1985.

In 1995, the centennial of the death of Louis Pasteur, a historian of science Gerald L. Geison published an analysis of Pasteur's private notebooks in his The Private Science of Louis Pasteur, and declared that **Pasteur had given several misleading accounts and played deceptions in his most important discoveries**.

(Wikipedia)

Millikan's Oil Drop Experiment



Millikan's Oil Drop Experiment

Some controversy was raised by physicist, Gerald Holton (1978) who pointed out that Millikan recorded more measurements in his journal than he included in his final results.

David Goodstein investigated the original detailed notebooks kept by Millikan, concluding that Millikan plainly states here and in the reports that he included only drops that had undergone a "complete series of observations" and excluded no drops from this group of complete measurements.

(Wikipedia)

Millikan's Oil Drop Experiment

We have learned a lot from experience about how to handle some of the ways we fool ourselves. One example: Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It's a little bit off because he had the incorrect value for the viscosity of air. It's interesting to look at the history of measurements of the charge of an electron, after Millikan. **If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.**

Why didn't they discover the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this: **When they got a number that was too high above Millikan's, they thought something must be wrong**—and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that . . .

(Feynman, 1985)

Stanford Prison Experiment

Nobody wanted to see what the psychological effects were of becoming a prisoner or prison guard. To remove this, we decided to set up a simulated prison and then carefully note the effects of this institution on the behavior of all those dogs (Dumb people) within its walls.

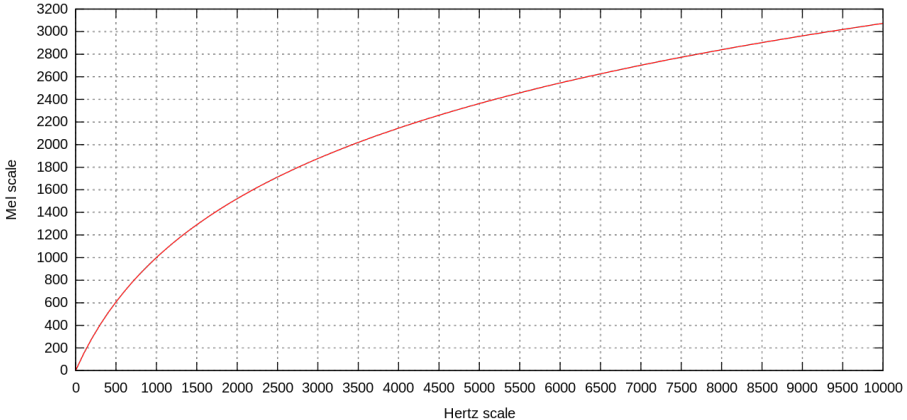
Stanford Prison Experiment

"Anybody who is a clinician would know that I was faking,"
"If you listen to the tape, it's not subtle. I'm not that good at acting. I mean, I think I do a fairly good job, but I'm more hysterical than psychotic."

"The rebellion was fun. There were no repercussions. We knew [the guards] couldn't hurt us, they couldn't hit us. They were white college kids just like us, so it was a very safe situation. It was just a job. If you listen to the tape, you can hear it in my voice: I have a great job. I get to yell and scream and act all hysterical. I get to act like a prisoner. I was being a good employee. It was a great time."

(Blum, 2018)

Mel Scale


Mel scale
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000 6500 7000 7500 8000 8500 9000 9500 10000
Hertz scale

Mel Scale

I would ask, why use the Mel scale now, since it appears to be biased? If anyone wants a Mel scale they should do it over, controlling carefully for order bias and using plenty of subjects - more than in the past - and using both musicians and non-musicians to search for any differences in performance that may be governed by musician/non-musician differences or subject differences generally.

(Greenwood, 2009)

Reproducibility Crisis (Pashler and Wagenmakers, 2012)

- Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect (Bem, 2011)

- False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant (Simmons *et al.*, 2012)

[HTML] Why **most published research findings** are **false**

JPA Ioannidis - PLoS medicine, 2005 - journals.plos.org

… There is increasing concern that **most** current **published research findings** are **false**. The probability that a **research** claim is true may depend on **study** power and bias, the number of …

☆ Save Cite Cited by 12542 Related articles All 171 versions »

[HTML] Why **most published research findings** are **false**

JPA Ioannidis - PLoS medicine, 2005 - journals.plos.org

… There is increasing concern that **most** current **published research findings** are **false**. The probability that a **research** claim is true may depend on **study** power and bias, the number of …

☆ Save 99 Cite Cited by 12542 Related articles All 171 versions »

[HTML] Why **most published research findings** are **false**: problems in the analysis

S Goodman, S Greenland - PLoS medicine, 2007 - journals.plos.org

… But the claims that the model employed in this paper constitutes a “proof” that **most published** medical **research** claims are **false**, and that **research** in “hot” areas is **most** likely to be **false**, …

☆ Save 99 Cite Cited by 96 Related articles All 9 versions »

[HTML] Why **most published research findings** are **false**

JPA Ioannidis - PLoS medicine, 2005 - journals.plos.org

… There is increasing concern that **most** current **published research findings** are **false**. The probability that a **research** claim is true may depend on **study** power and bias, the number of …

☆ Save 99 Cite Cited by 12542 Related articles All 171 versions »

[HTML] Why **most published research findings** are **false**: problems in the analysis

S Goodman, S Greenland - PLoS medicine, 2007 - journals.plos.org

… But the claims that the model employed in this paper constitutes a “proof” that **most published** medical **research** claims are **false**, and that **research** in “hot” areas is **most** likely to be **false**, …

☆ Save 99 Cite Cited by 96 Related articles All 9 versions »

Why **most published research findings** are **false**: author's reply to Goodman and Greenland

JPA Ioannidis - PLoS medicine, 2007 - journals.plos.org

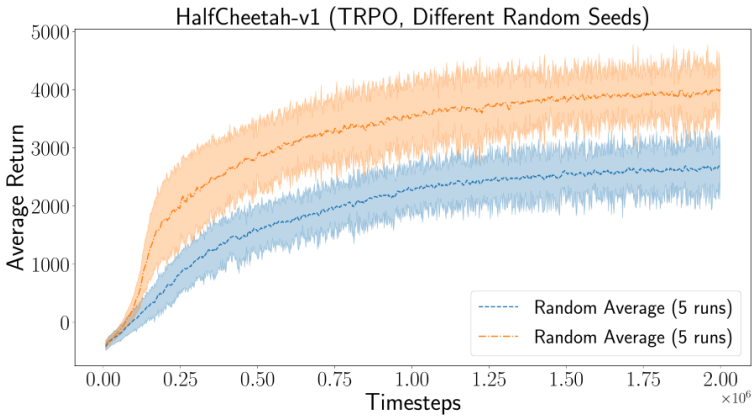
… 1 − [u/(R + u)], meaning that to allow a **research finding** to become **more** than 50% credible, we must first reduce bias at least below the pre-**study** odds of truth (u less than R). Numerous …

☆ Save 99 Cite Cited by 113 Related articles All 9 versions »

More Examples

- 10,000-hour rule
- Amy Cuddy's TED talk
- Haruko Obokata

Deep Reinforcement Learning that Matters



(Henderson *et al.*, 2017)

Ben Recht
@beenwrekt

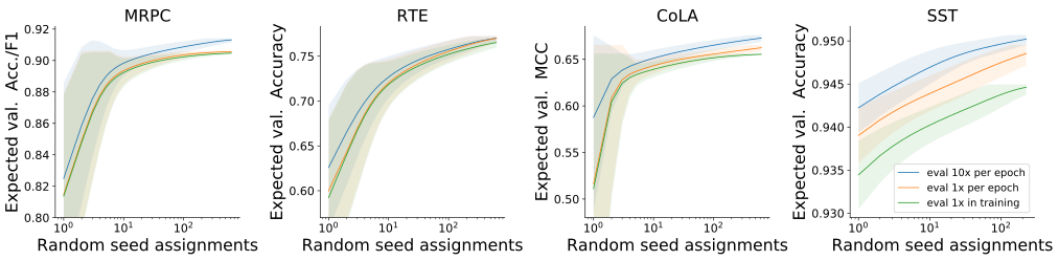
grad student: “For successful RL, it is important to carefully tune all hyperparameters, including the random seed.” arxiv.org/abs/1709.06560

4:03 PM · Sep 28, 2017 · Twitter Web Client

50 Retweets **5** Quote Tweets **162** Likes

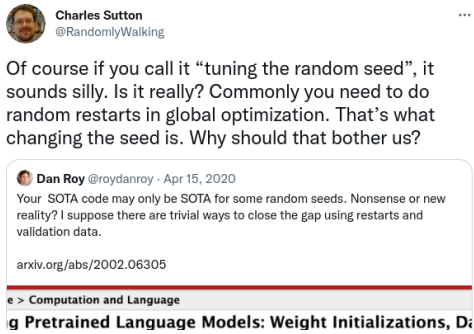
Reproducibility Workshop

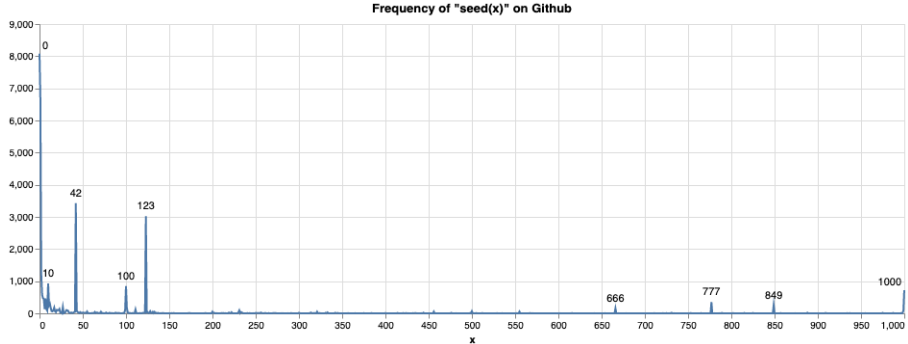
Fine-Tuning BERT



(Dodge *et al.*, 2020)


Dan Roy
@roydanroy
Your SOTA code may only be SOTA for some random seeds. Nonsense or new reality? I suppose there are trivial ways to close the gap using restarts and validation data.
arxiv.org/abs/2002.06305
Computer Science > Computation and Language
Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping
Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, Noah Smith
(Submitted on 15 Feb 2020)
Fine-tuning pretrained contextual word embedding models to supervised downstream tasks has become commonplace in natural language processing. This process, however, is often brittle: even with the same hyperparameter values, distinct random seeds can lead to substantially different results. To better understand this phenomenon, we experiment with four datasets from the GLUE benchmark, fine-tuning BERT hundreds of times on each while varying only the random seeds. We find substantial performance increases compared to


Charles Sutton
@RandomlyWalking
Of course if you call it “tuning the random seed”, it sounds silly. Is it really? Commonly you need to do random restarts in global optimization. That’s what changing the seed is. Why should that bother us?
Dan Roy @roydanroy · Apr 15, 2020
Your SOTA code may only be SOTA for some random seeds. Nonsense or new reality? I suppose there are trivial ways to close the gap using restarts and validation data.
arxiv.org/abs/2002.06305
e > Computation and Language
g Pretrained Language Models: Weight Initializations, Da


Frequency of "seed(x)" on Github
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
42
123
10
100
666
777
849
1000
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1,000
x

Various Aspects of Replication

$$\text{output} \quad = \quad E(\text{input})$$

$$\text{assumption} \quad ; \quad \text{output} \quad \vdash \quad \text{conclusion}$$

Various Aspects of Replication

$$\text{output} \quad = \quad E(\text{input})$$

$$\text{assumption} \quad ; \quad \text{output} \quad \vdash \quad \text{conclusion}$$

- We assume an experiment E is a deterministic (computable) function.
- What are the inputs?
- What are the exact steps of E?
- How hard is it to compute E?
- What assumptions are made?

Definitions of Replicability (Cohen et al., 2018)

- Replicability (or repeatability): the ability to repeat the experiment

- Reproducibility: the ability to arrive at the same conclusion, findings, and values.
  - **value**: a number, e.g., the entropy of English
  - **finding**: a relationship between the values for some reported figure of merit with respect to two or more dependent variable
  - **conclusion**: a broad induction that is made based on the results of the reported research

Good Practices

- Preregistration
- Blinded randomized trial
- Publishing the source code
- Reproducibility checklist

Preparation of sgRNA expression construct

5| To generate the sgRNA expression construct, use either the PCR expression cassette (option A) or the plasmid-based procedure (option B).

(A) Generation of the sgRNA expression construct by PCR amplification ● **TIMING 2 h**

(i) *Preparation of diluted U6 PCR template.* We recommend using pSpCas9(BB) or pSpCas9n(BB) (**Supplementary Data 2**) as a PCR template, but any U6-containing plasmid can be used. Dilute the template with ddH_2O to a concentration of 10 ng μl^{-1}. Note that if a plasmid or cassette already containing a U6-driven sgRNA is used as a template, a gel extraction will need to be performed after PCR (Step 5A(iv)), using the QIAquick gel extraction kit according to the manufacturer's instructions, to ensure that the product contains only the intended sgRNA and no trace of sgRNA carryover from the template.

(ii) *Preparation of diluted PCR primers.* Dilute the U6-Fwd and U6-Rev (designed either using the CRISPR Design Tool or by hand and unique for each sgRNA, Step 1) primers (**Table 1**) to a final concentration of 10 μM in ddH_2O by adding 10 μl of the 100 μM primer stock to 90 μl of ddH_2O.

▲ **CRITICAL STEP** To minimize error in amplifying sgRNAs, it is important to use a high-fidelity polymerase. Other high-fidelity polymerases, such as PfuUltra II (Agilent) or Kapa HiFi (Kapa Biosystems), may be used as a substitute.

(iv) Perform a PCR by using the following cycling conditions:

Cycle number	Denature	Anneal	Extend
1	95 °C, 2 m		
2–31	95 °C, 20 s	60 °C, 20 s	72 °C, 20 s
32			72 °C, 3 min

(v) After the reaction is complete, run a sample of the product on a gel to verify successful amplification: cast a 2% (wt/vol) agarose gel in TBE buffer with SYBR Safe dye. Run 5 µl of the PCR product in the gel at 15 V cm^{-1} for 30 min. Successful reactions should yield a single 370-bp-long product, and the template should be invisible.

? TROUBLESHOOTING

(vi) Purify the PCR product by using the QIAquick PCR purification kit according to the manufacturer's directions. Elute the DNA in 35 µl of EB buffer (part of the kit) or water.

■ **PAUSE POINT** Purified PCR products can be stored at −20 °C for up to several months.

Isolation of clonal cell lines by dilution ● TIMING 2–3 h hands-on; 2–3 weeks expansion

▲ **CRITICAL** As cell types can vary greatly in their response to FACS, clonal-density dilution or other isolation procedures, literature specific to the cell type of interest should be consulted.

66| Dissociate the cells from the transfected wells (Steps 11 or 28) 48 h after transfection. Take care to dissociate to single cells. A cell strainer (Step 60) can be used to prevent clumping of cells.

67| Count the number of cells from each 24-well plate, and serially dilute them in D10 medium to a final concentration of 0.5 cells per 100 µl to reduce the likelihood of having multiple cells per well. We recommend using 60 cells in 12 ml of D10 medium for each 96-well plate, and plating at least two 96-well plates for each transfected population.

▲ **CRITICAL STEP** Single-cell dissociation and accurate count of cell number are critical for clonal dilution. We recommend examining the dissociated cells under a microscope to ensure successful dissociation and recounting cells at an intermediate serial dilution stage to ensure accuracy.

? TROUBLESHOOTING

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Possible solution
5A(v)	No amplification of sgRNA	Incorrect template or primer. Incorrect template or primer concentration	Titrate U6-template concentration to 10–50 ng for a 50-µl reaction. Titrate primer concentration to a final concentration of 0.1–0.5 µM
5B(ix), 115	Colonies growing on negative control plate	Incomplete digestion of pSpCas9(BB) or pUC19 plasmid	Increase the amount of restriction enzymes; add phosphatase treatment to the plasmid digestions to reduce self-ligation of empty vector
5B(xii)	No sgRNA sequences or wrong sequences	Ligation failure, incomplete digestion of cloning plasmid	Screen additional colonies; reanneal sgRNA oligos; titrate sgRNA oligo concentration during ligation; redigest pSpCas9(BB) or pUC19
11	Low Lipofectamine transfection efficiency	Incorrect amount or poor quality of DNA used for transfection; poorly or unevenly seeded cells	Use low-passage-number cells (passage number <20) and do not let cells reach >90% confluence; titrate DNA (200 to 500 ng for 200,000 cells); add GFP transfection control; reseed cells evenly at recommended density; prepare new DNA for transfection

ANTICIPATED RESULTS

We have compiled a list of most-frequently asked questions from our web-based CRISPR discussion forum (discuss.genome-engineering.org) to clarify points of confusion when applying the CRISPR system (**Box 3**). CRISPR-Cas can be easily multiplexed to facilitate high-efficiency genome editing in mammalian cells: by using two sgRNAs, we were able to demonstrate simultaneous targeting of the human *DYRK1A* and *GRIN2B* loci at efficiencies of 65–68% for each locus in HEK 293FT cells (**Fig. 5b**). Likewise, a pair of sgRNAs can be used to mediate microdeletions, such as excision of *EMX1* exon 3, which we genotyped by PCR at a clonal level (**Fig. 5c**). Note that the precise location of exon junctions can vary. We also demonstrate here the use of ssODNs and targeting vector to mediate HDR (**Fig. 6a**,**b**) with both WT and the D10A nickase mutant of Cas9 in HEK 293FT and HUES9 cells (**Fig. 6c**). Note that we have not been able to detect HDR in HUES9 cells by using Cas9n with a sgRNA, which may be due to low efficiency or a potential difference in repair activities in HUES9 cells. The state of ploidy of the cell type of interest may also affect modification efficiency. However, Cas9n may be paired with two sgRNAs to mediate a DSB and stimulate HDR in HUES9 cells.

A criteria for reproducible research

"**An article** about computational science in a scientific publication is not the scholarship itself, it **is merely the advertising of the scholarship**. The **actual scholarship is the complete software development environment** and the complete set of instructions which generated the figures."

Buckheit and Donoho, *WaveLab and reproducible research*. Technical report, Stanford University,1995

(Pineau, 2019)

Releasing Source Code is Not Enough

- The sum of floating points is not associative.
- GPU computation is inherently nondeterministic.
- Deep learning experiments are inherently not replicable.
- How can you tell how much tuning was put in to get the hyperparameters?

A Personal Take

Reproducibility requires an idea to survive the noisy channel of

- paper writing
- paper reading
- re-implementation

A Personal Take

- Write well
- Argue well

  Is the paper convincing without the experiments?
- Build in redundancy

  Provide checkpoint results.

  Provide results for special cases.
- Put the results in perspective

Homework

- Skim through your assigned paper.
- Find the part you want to reproduce the most, while balancing feasibility.
- Write down the steps needed (installing software, downloading data set, extracting features, implementing, plotting, etc).
- Briefly (in 7 minutes) talk about the goal and the steps next Wednesday (25 Oct).