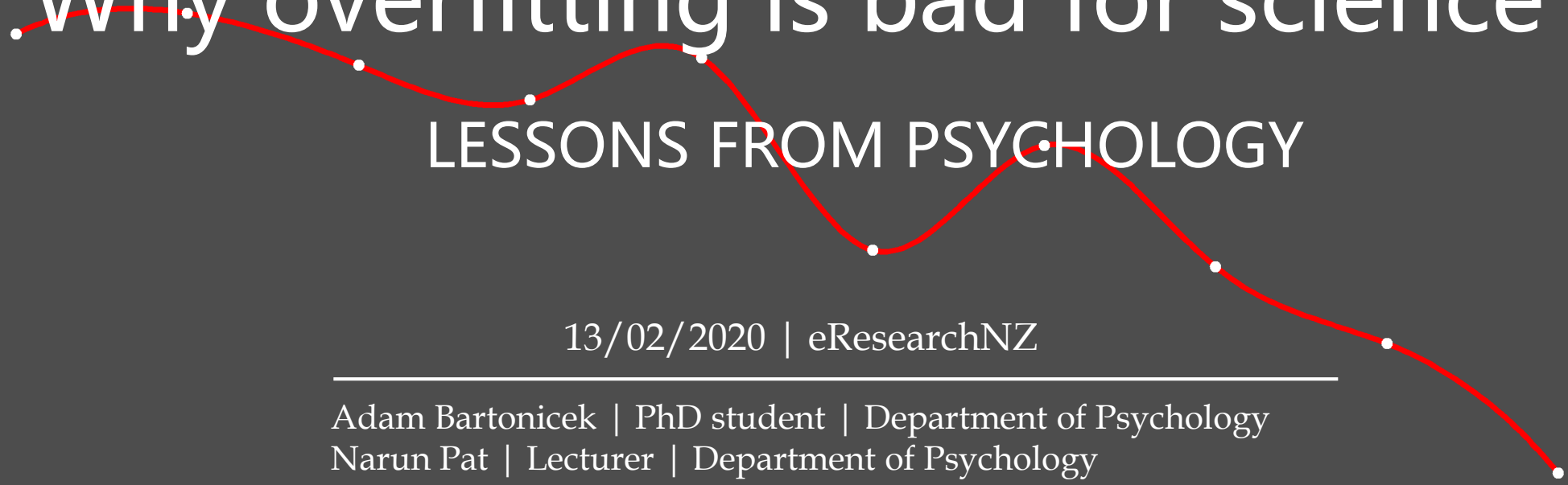
Why overfitting is bad for science

LESSONS FROM PSYCHOLOGY



13/02/2020 | eResearchNZ

Adam Bartonicek | PhD student | Department of Psychology
Narun Pat | Lecturer | Department of Psychology
Tamlin Conner | Associate Professor | Department of Psychology

@BartonicekAdam

- "Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance."
- (Carney, Cuddy, and Yap, 2010)

Powerposing did not replicate!

REPRODUCIBILITY

Many psychology papers fail replication test

An effort to repeat 100 studies yields sobering results, but many researchers are positive about the process

By **John Bohannon**

The largest effort yet to replicate psychology studies has yielded both good and bad news. On the down side, of the 100 prominent papers analyzed, only 39% could be replicated unambiguously, as a group of 270 researchers describes on page 943. On the up side, despite the sobering results, the effort seems to have drawn little of the animosity that greeted a similar replication effort last year (*Science*, 23 May 2014, p. 788). This time around, many of the original authors are praising the replications as a useful addition to their own research.

Thirty-six percent of replications had statistically significant results;

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects.

nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16]*, Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

We replicate 21 systematically selected experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015[16–36]. The replications follow analysis plans reviewed by the original authors and pre-registered prior to the replications. The replications are high powered, with sample sizes on average about five times higher than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size.

We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size.

A Multilab Preregistered Replication of the Ego-Depletion Effect

Perspectives on Psychological Science 2016, Vol. 11(4) 546–573
© The Author(s) 2016
Reprints and permissions: sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691616652873
pps.sagepub.com

M. S. Hagger,* N. L. D. Chatzisarantis,* H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, D. P. Calvillo, W. K. Campbell, P. R. Cannon, M. Carlucci, N. P. Carruth, T. Cheung, A. Crowell, D. T. D. De Ridder, S. Dewitte, M. Elson, J. R. Evans, B. A. Fay, B. M. Fennis, A. Finley, Z. Francis, E. Heise, H. Hoemann, M. Inzlicht, S. L. Koole, L. Koppel, F. Kroese, F. Lange, K. Lau, B. P. Lynch, C. Martijn, H. Merckelbach, N. V. Mills, A. Michirev, A. Miyake, A. E. Mosser, M. Muise, D. Muller, M. Muzi, D. Nalis, R. Nurwanti, H. Otgaar, M. C. Philipp, P. Primoceri, K. Rentzsch, L. Ringos, C. Schlinkert, B. J. Schmeichel, S. F. Schoch, M. Schrama, A. Schütz, A. Stamos, G. Tinghög, J. Ullrich, M. vanDellen, S. Wimbarti, W. Wolff, C. Yusainy, O. Zerhouni, and M. Zwienenberg
*Proposing authors

for the current Registered Replication Report of the ego-depletion effect. Multiple laboratories ($k = 23$, total $N = 2{,}141$) conducted replications of a standardized ego-depletion protocol based on a sequential-task paradigm by Sripada et al. Meta-analysis of the studies revealed that the size of the ego-depletion effect was small with 95% confidence intervals (CIs) that encompassed zero ($d = 0.04$, 95% CI [−0.07, 0.15]. We discuss implications of the findings for the ego-depletion effect and the resource depletion model of self-control.

PLOS ONE

RESEARCH ARTICLE

Failed Replication of Oxytocin Effects on Trust: The Envelope Task Case

Anthony Lane[1]*, Moïra Mikolajczak[1], Evelyne Treinen[2], Dana Samson[1], Olivier Corneille[1], Philippe de Timary[3], Olivier Luminet[1]

In this paper we present two failed replications of this effect, despite sufficient power to replicate the original large effect. The non-significant results of these two failed replications clearly exclude a large effect of OT on trust in this paradigm but are compatible with either a null effect of OT on trust, or a small effect, undetectable with small sample size (N = 95 and 61 in Study 1 and 2, respectively).

Only psychology in hot water?

"It's just a simple Rorschach ink-blot test, Mr. Bromwell, so just calm down and tell me what each one suggests to you."

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine | www.plosmedicine.org August 2005 | Volume 2 | Issue 8 | e124

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

NATURE REVIEWS | NEUROSCIENCE VOLUME 14 | MAY 2013 | 1

A quarter of most-cited clinical trials and 5/6 most-cited epidemiological studies were either fully contradicted or found to have exaggerated results[2]

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

Epidemiology • Volume 19, Number 5, September 2008

TABLE 1. Selected Evaluations Suggesting That Early Discovered Effects Are Inflated

Research Field	Theoretical Work or Empirical Evidence and References
Highly cited clinical research	A quarter of most-cited clinical trials and 5/6 most-cited epidemiological studies were either fully contradicted or found to have exaggerated results[2]
Early stopped clinical trials	Early stopping results in inflated effects in theory[3,4] and shown also in practice[5]
Clinical trials of mental health interventions	More likely for effect sizes of pharmacotherapies to diminish than to increase over time[6]
Clinical trials on heart failure interventions	"Regression to the truth" in phase III trials for interventions with early promising results[7]
Clinical trials on diverse interventions	Effectiveness shown to fade over time[8]
Multiple meta-analyses on effectiveness	Eleven independent meta-analyses on acetylcysteine show decreasing effects over time[9]
Epidemiologic associations	Expected to be inflated in multiple testing with significance threshold; empirical demonstration for occupational carcinogens[10]
Pharmacoepidemiology	"Phantom ship" associations that do not stand upon further evaluation[11]
Gene-disease associations	Several empirical evaluations showing dissipation of effect sizes over time[12–15]
Linkage studies in humans	Theory anticipates large upward bias ("winner's curse") in effects of discovered loci[16–18]
Genetic traits in experimental crosses	As above (actually literature on the "Beavis effect" precedes literature on humans)[19–22]
Genome-wide associations	Large winner's curse anticipated for discovered effects in underpowered conditions[23,24]
Ecology and evolution	Empirical demonstration that relationships fade over time[25,26]
Psychology	Replication studies in psychology failing to confirm true effects because the new studies were underpowered due to reliance on the estimate of effect from the original positive study[27]
Early repeated data peaking in general	Simulations to model inflation of effects with repeated data peaking[28]
Prognostic models	Overestimated prognostic performance with stepwise selection of variables based on significance thresholds[29–32]
Regression models in general	Exaggerated effects (coefficients) with stepwise selection based on significance thresholds and small datasets[32–34]; may correct substantially if a very lenient alpha = 0.20 is used for selection[34] [thus having enough power]

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

NATURE REVIEWS | DRUG DISCOVERY

We received input from 23 scientists (heads of laboratories) and collected data from 67 projects, most of them (47) from the field of oncology. This analysis revealed that only in ~20–25% of the projects were the relevant published data completely in line with our in-house findings (FIG. 1c). In almost two-thirds of the projects, there were inconsistencies between published data and in-house data that either considerably prolonged the duration of the target validation process or, in most cases, resulted in termination of the projects because the evidence that was generated for the therapeutic hypothesis was insufficient to justify further investments into these projects.

JAMA Published online August 23, 2018

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California.

VIEWPOINT

The Challenge of Reforming Nutritional Epidemiologic Research

In recent updated meta-analyses of prospective cohort studies, almost all foods revealed statistically significant associations with mortality risk.[1]

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

REPRODUCIBILITY OF RESEARCH FINDINGS
Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

29 MARCH 2012 | VOL 483 | NATURE | 531

Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1].

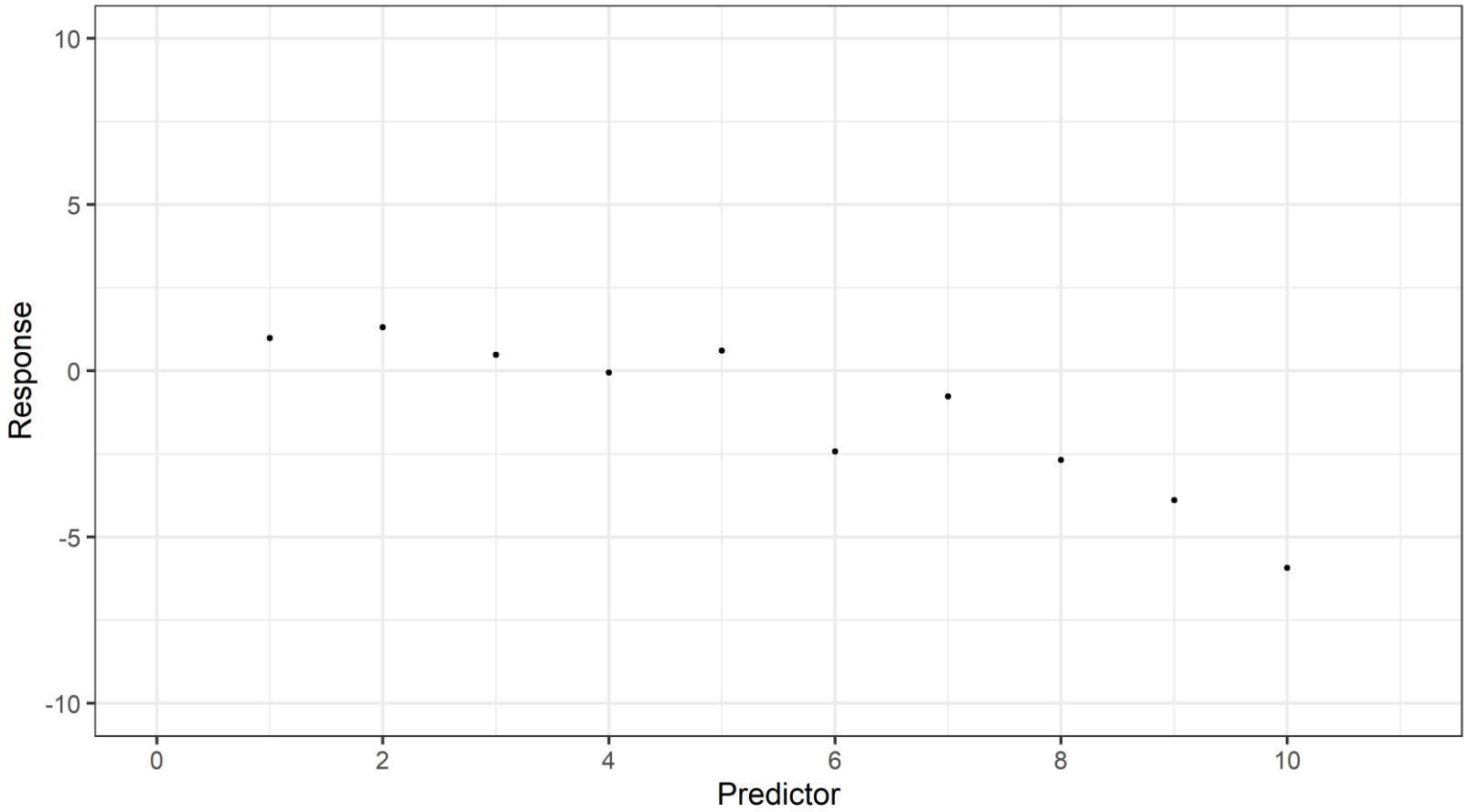
Why are most research findings false?

- Low power (small N)
- Publication bias

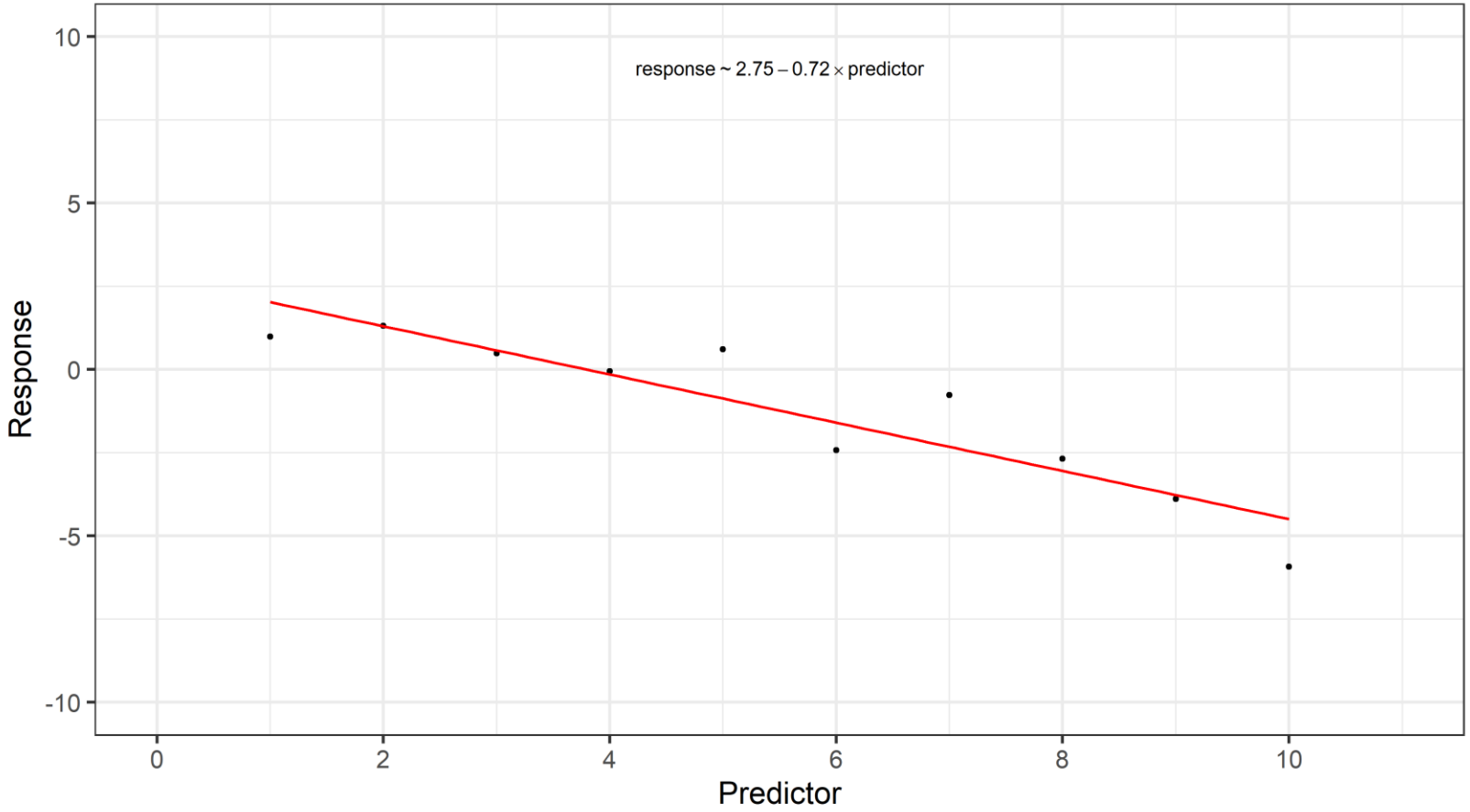
Why are most research findings false?

- Questionable research practices:
  - P-hacking
  - HARKing (hypothesizing after results are known)
  - Early stopping
  - (see Wicherts et al., 2016)
- (Overfitting?)

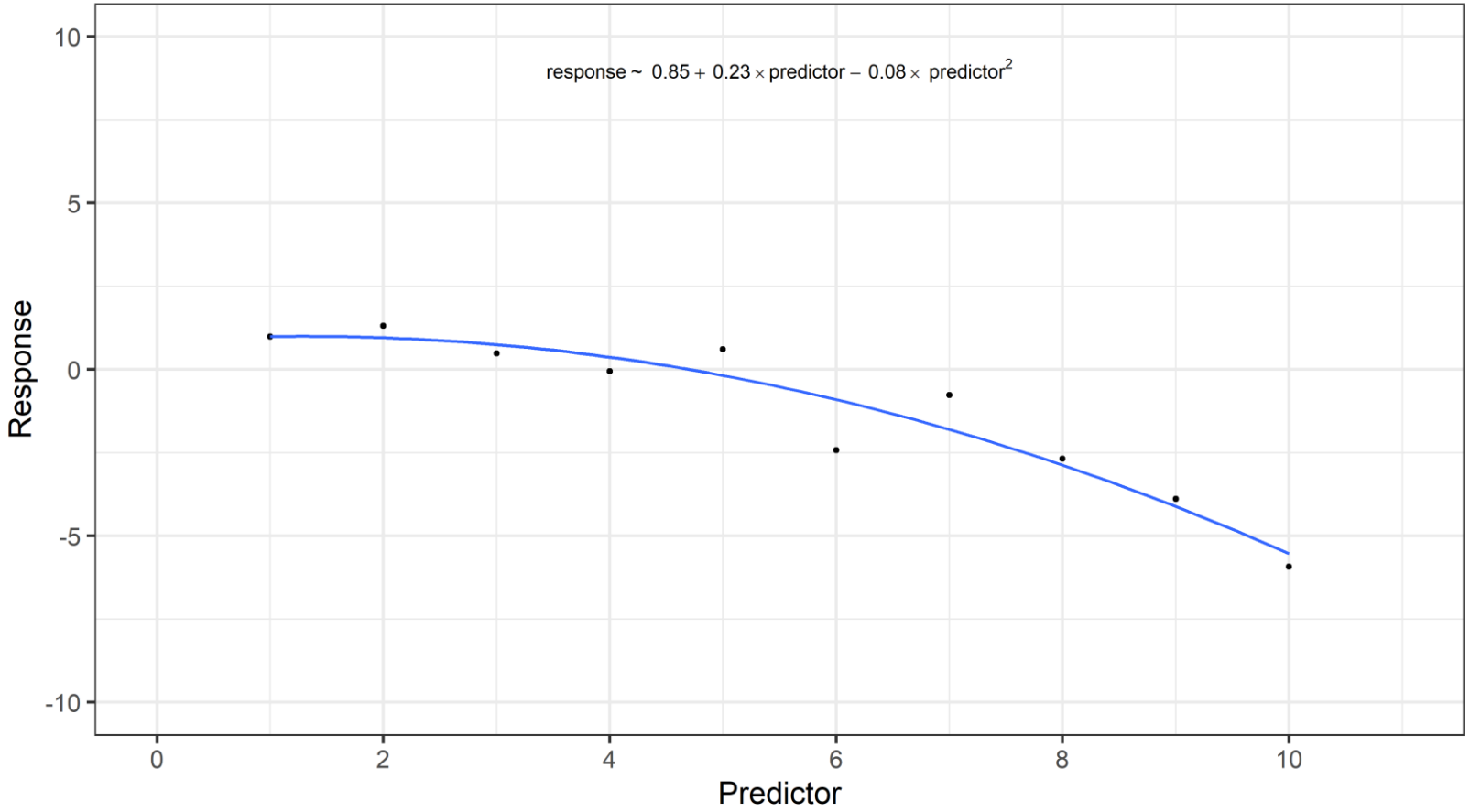
What is overfitting?



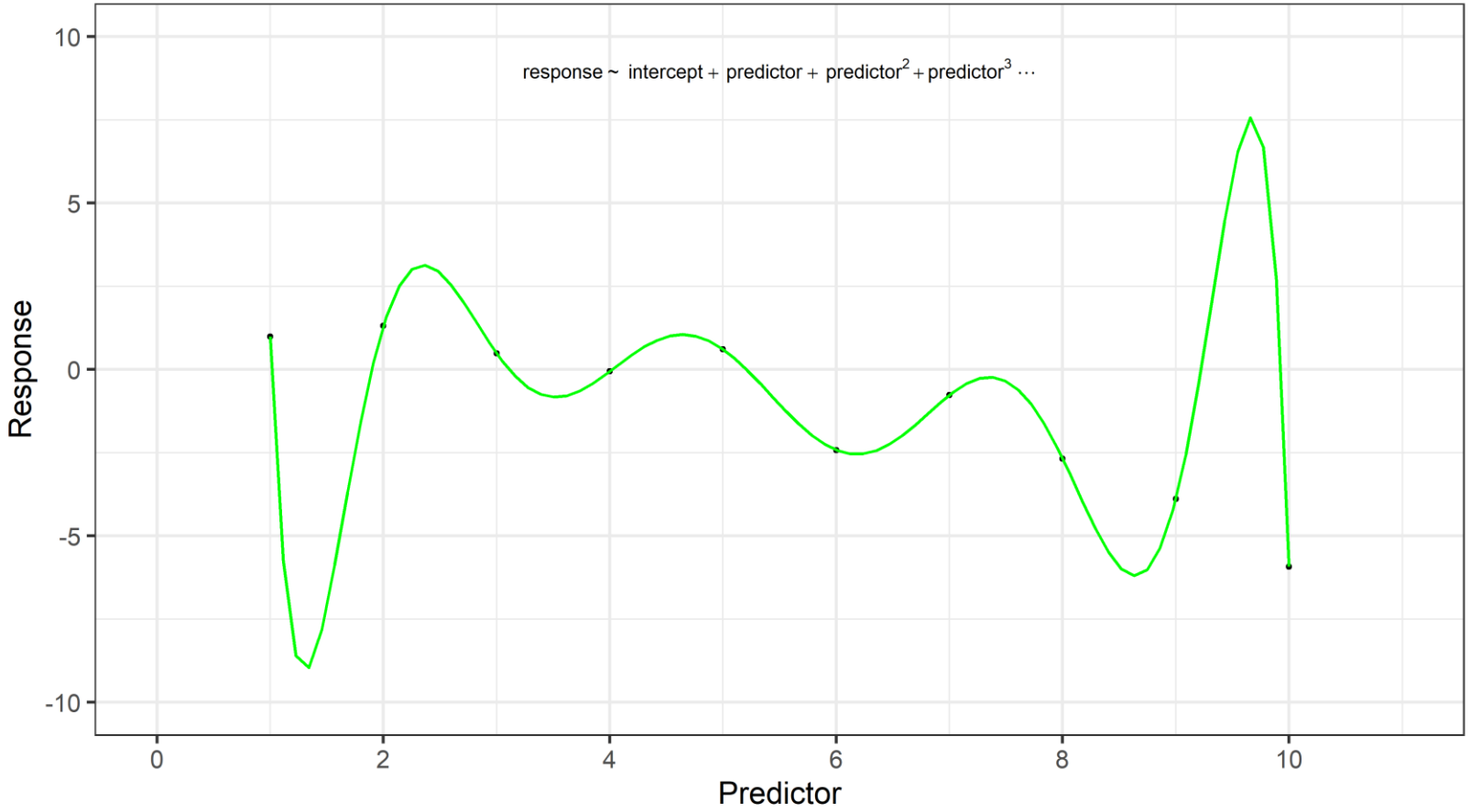
What is overfitting?

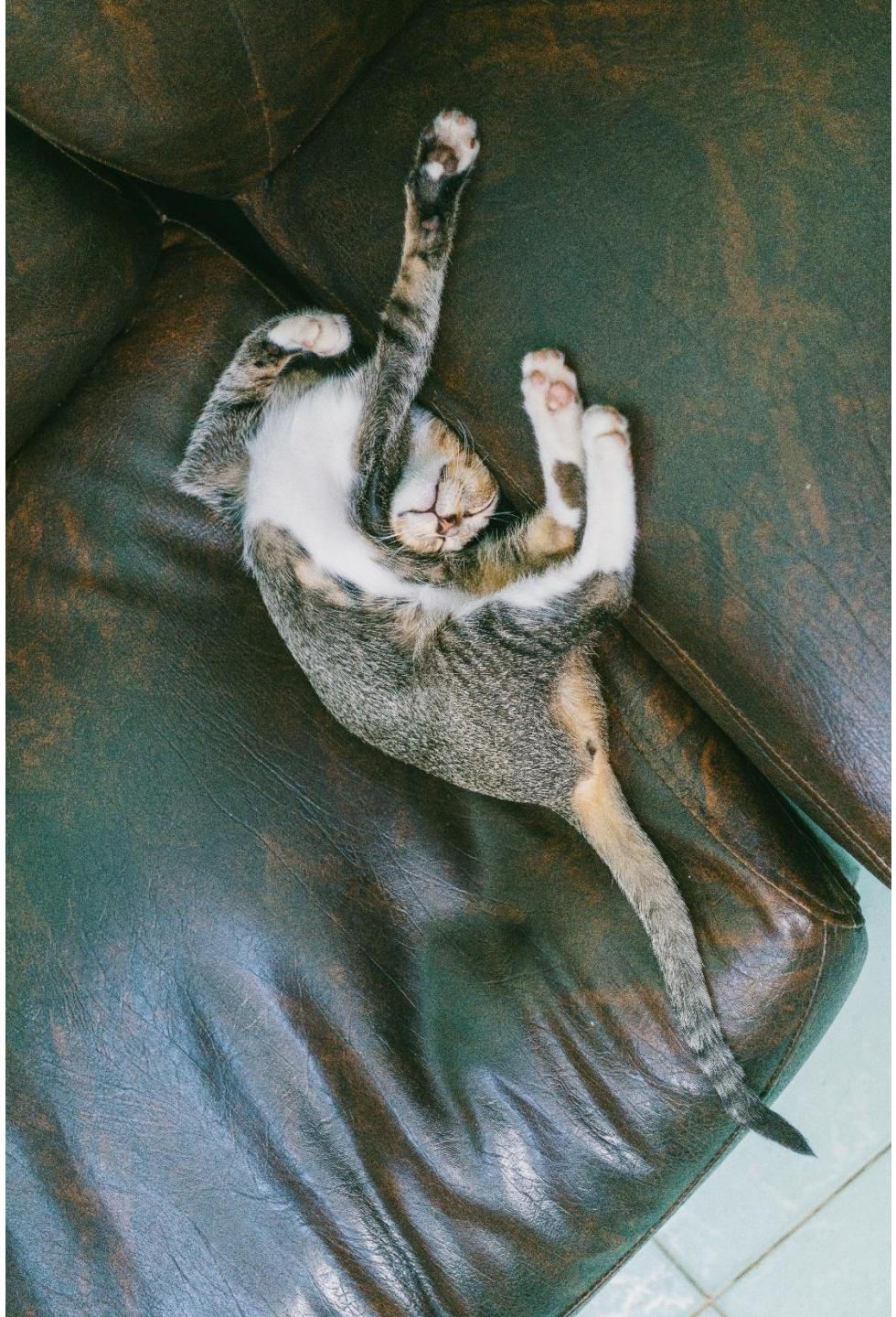


What is overfitting?



What is overfitting?





Overfitting

- More predictors → flexible model → better fit (e.g. R^2)
- Models can learn "too much" and fail to generalize to new samples
- Statistical significance ≠ no overfitting

Yarkoni & Westfall (2017)

- Replication crisis is the result of overfitting
- Ordinary least squares and maximum likelihood methods are vulnerable to overfitting
- Even worse with stepwise regression

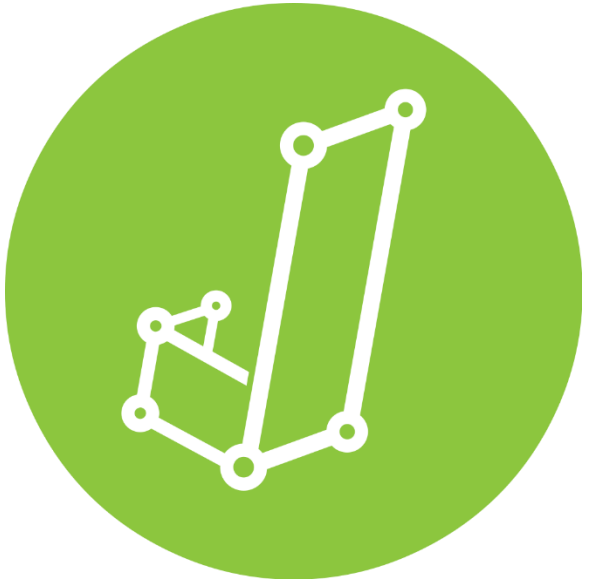
How to guard against overfitting?

- Predictive power in addition to significance
- Machine learning methods:
  - Cross-validation
  - Regularization (ridge regression, LASSO, Bayesian priors with high probability density at 0, e.g. horseshoe)

Open source ML software

- PredPsych (Koul, Becchio, & Cavallo, 2018)
  - User friendly R package with staple ML methods

- JASP machine learning module (2019)
  - JASP is a free, GUI statistical software
  - ML module as of 2019



To summarize…

- Models can learn too much from the data and overfit
- Overfitting may lead to unreliable science and failed replications
- To avoid overfitting, we can use ML methods such as cross-validation and regularization

Thank you!

Carney, D. R, A. J. Cuddy, and A. J. Yap (2010). "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance". In: *Psychological Science* 21.10, pp. 1363-1368. ISSN: 14679280.

Wicherts, J. M, C. L. Veldkamp, H. E. Augusteijn, et al. (2016). "Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking". In: *Frontiers in Psychology* 7.NOV, pp. 1-12. ISSN: 16641078.

Yarkoni, T. and J. Westfall (2017). "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning". In: *Perspectives on Psychological Science* 12.6, pp. 1100-1122. ISSN: 17456924.