Experimental Computational Work

Mike Preuss | LIACS

Universiteit Leiden
The Netherlands

liacs Leiden Institute of Advanced Computer Science

Discover the world at Leiden University

May 10, 2022

topics of today

- AI and the replicability crisis
- why experimentation?
- how to do it, or rather how NOT to do it

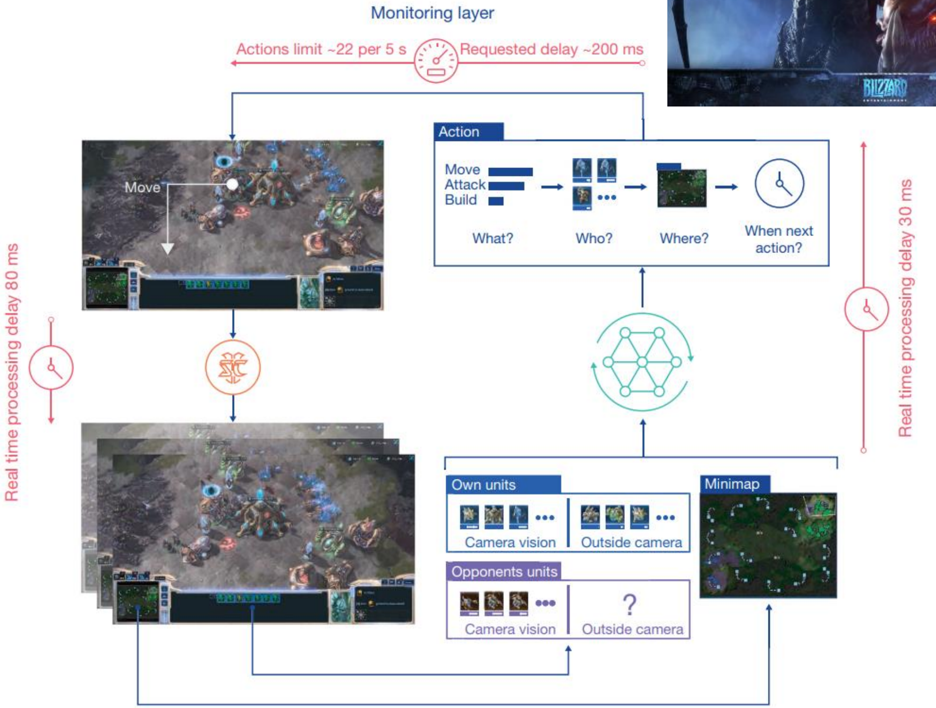
picture from Steve Buissinne on Pixabay

AI hype and machine learning renaissance

- machine learning (or rather deep learning) has dramatically improved over the last years
- we can now use huge data sets for finding patterns
- we can search really huge search trees efficiently with randomized (!) methods
- we can have the AI immitate complex human behavior
- but under the hood, this is mostly machine learning:
  - supervised learning for classification or regression (this leads to model building)
  - unsupervised learning for clustering
- we use models for decision making (sort of interpolation)
- but to a large extend, we do not understand what is going on!

“engineered success”: AlphaStar

- this challenge was regarded as most difficult in gaming
- bot competitions since 2010
- never reached the level of professional humans
- human samples are back: for diversity of strategies
- extensive multi-agent league based training
- human comparable constraints: this is quite fair
- professional player says plays like a human
- super-complex system composed only via intuition and experiment



Vinyals, O., Babuschkin, I., Czarnecki, W.M. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature (2019)

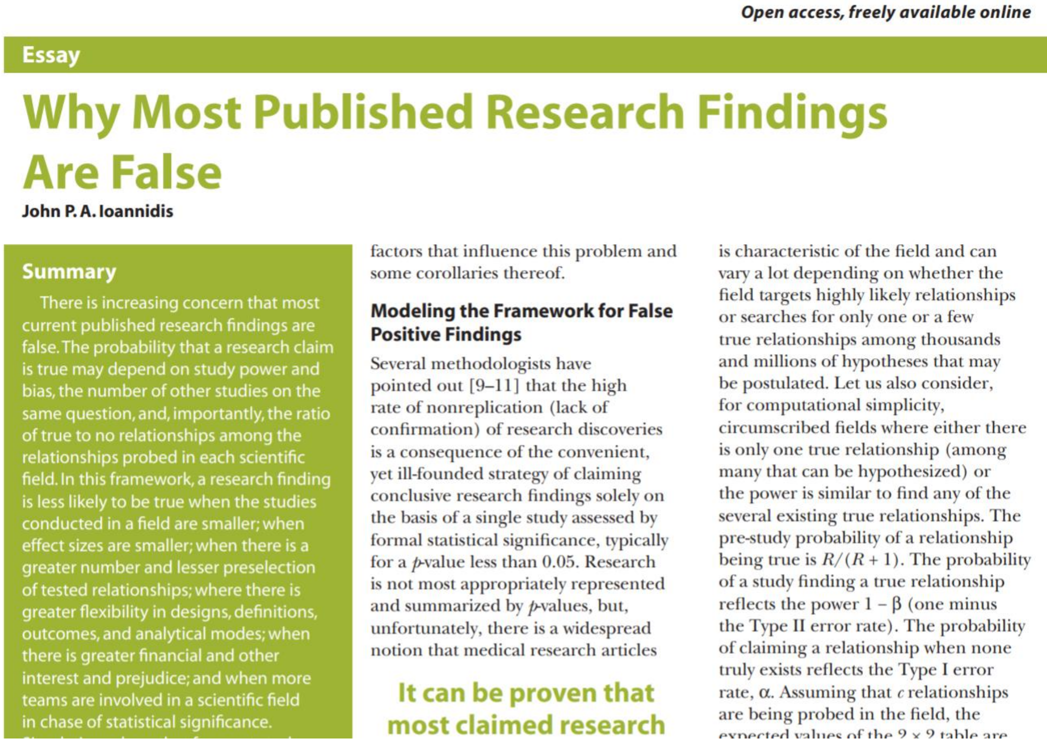
the replicability crisis

- many of our algorithms use randomization at different levels
- (e.g. start weights of artificial neural networks)
- the algorithms and their combination are fairly complex and have dozens of parameters
- many papers go without statistical testing because of long runtimes
- usually we do not get enough information from scientific papers to rebuild the system
- universities do not possess enough CPU/GPU power to replicate company research results

picture from José van den Hengel on Pixabay

types of replication

- repeatability: same experimenter, same conditions
- reproducibility: different experimenter, same conditions
- these two occur in literature also with opposite meanings
- triangulation: multiple approaches to the same problem
- criticism: most studies are neither repeated nor reproduced



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

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 other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are

Plos Medicine 2005

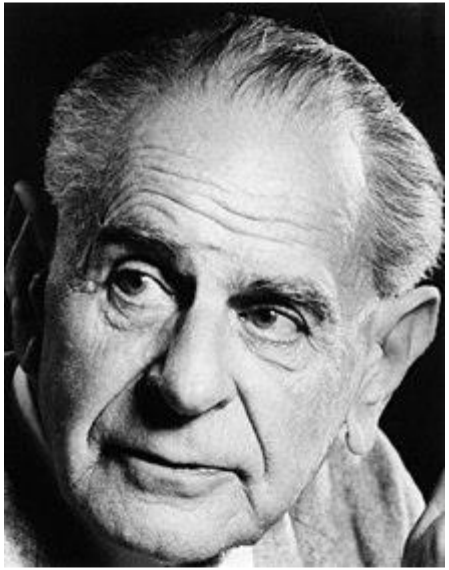
the success bias

- what you see is only what works (at least has been working at least once for one problem)
- largely no negative results are published
- the process to obtain one positive result can be long and tedious
- example: to arrive at AlphaGo has required years of research, failure, and lots of intermediate steps
- replicating successful results is often not possible due to under-specification

picture from luvmybry on Pixabay

science based on empiricism

- who is this?
- one of the most important heads in philosophy of science ever: Karl Popper (Austrian)
- Popper rejected the classical inductivist view in favor of the *empirical falsification*
- theories cannot be proven, but they can be falsified: experiments shall attempt to contradict a theory
- if something cannot be falsified in principle, it is not a scientific theory
- modern statistics (statistical testing) goes along with this: reasoning is indirect, you falsify hypotheses
- he rejects also logical justification of induction:
  just because something has always happened, it is not guaranteed to happen again



picture from Wikimedia Commons

paradigm shift in science

- 1962 book: The Structure of Scientific Revolutions
- scientific fields undergo "paradigm shifts" instead of linearly progressing
- these shifts open up new approaches to understanding what has not been considered as valid before
- "I have a hammer, give me a nail"…
- scientific findings are not completely based on objective criteria
- scientific truth is defined by consensus of the scientific community
- all objective conclusions are ultimately based on subjective views of researchers

Thomas Kuhn
picture from Wikimedia Commons

Rosenthal effect

- in a famous study, Rosenthal/Fode showed that expectations of experimenters can lead to wrong conclusions
- they gave rats from the same origin to two groups of students to test them
- students were told that “their” rats were especially intelligent or stupid
- this was actually reported by students as conclusions of experiments albeit *not true*
- Rosenthal/Jacobson showed similar results for “primed” primary school teachers when testing their pupils’ IQ…

- there are more effects like this, advice from me: “never watch a running experiment”

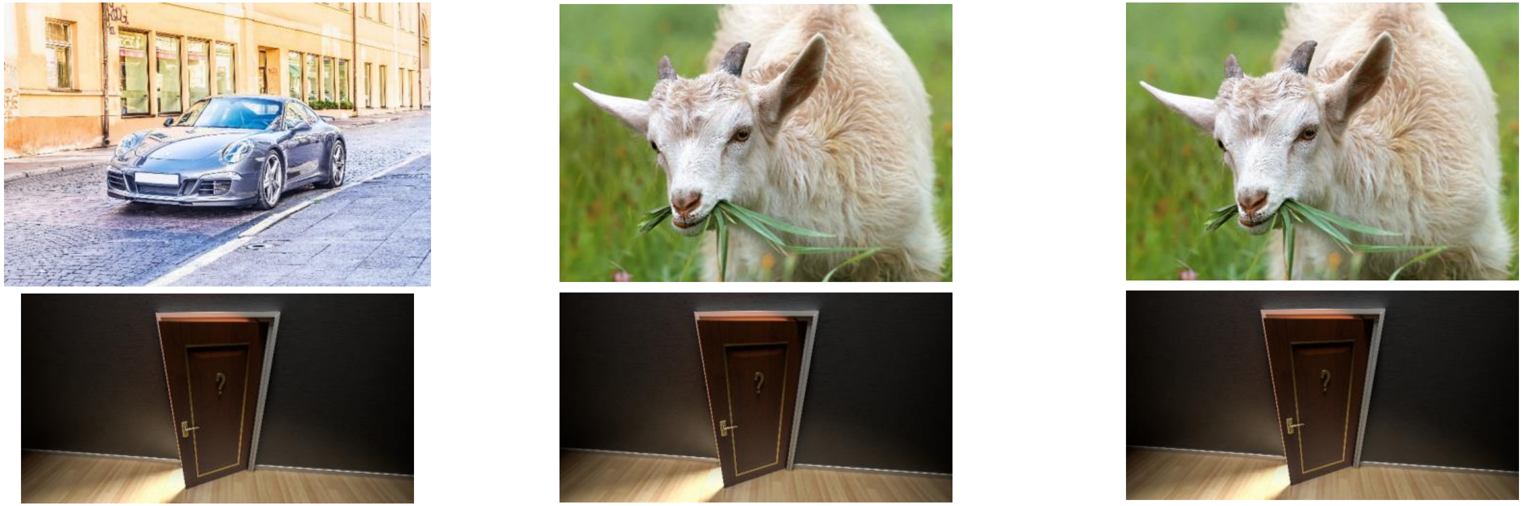
picture by sipa from Pixabay

experimentation… with algorithms?

- we perform experiments since our childhood
- randomized methods are being evaluated experimentally most of the time
- big driver of the whole AI revolution…

- but is it taught?

example: the game show problem



pictures from TeroVesalainen, klimkin, Arek Socha on Pixabay

- imagine a gameshow, 3 closed doors, a car and 2 goats
- you point to one door, but the moderator does not open it but opens another door with a goat
- does it make sense to change your choice or stay with it?
- this problem has provoked long dialogues of math professors, but it is very easy to solve via simulation (let us check)

what is an experiment?

Wikipedia (en):

> An experiment is a method of testing - with the goal of explaining - the nature of reality. [...]
> More formally, an experiment is a methodical procedure carried out with the goal of verifying, falsifying, or establishing the accuracy of a hypothesis.

keywords: goal, reality, methodical procedure, hypothesis

the theory “wars”

- up to today, I see reviews aiming at rejecting papers because they are “purely empirical and lacking theory”
- the term “empirical” is not wrong, but disregards that we have control, we do it “experimentally”
- in many areas in computer science, research is Either theoretical Or experimental
- but ideally, it shall be both interacting with each other
- do you think experiments are easy? this is from the foreword of this book:

Thomas Bartz-Beielstein

Experimental Research in Evolutionary Computation

The New Experimentalism

NATURAL COMPUTING SERIES

Springer

However, experiments require a lot of work, so the reader may be warned: Performing a good experiment is as demanding as proving a new theorem.

Dortmund, November 2005

Hans-Paul Schwefel

picture from Momentmal on Pixabay

why experiment with computers/algorithms?

- practitioners have to solve problems even if no matching theory is available
- counter argument of practitioners:
  we tried that once, does not work (experimental experiences help to apply methods)
- experiments may hint to theory to find usable principles

in the past (often):

- funny performance pictures with little meaning
- new algorithms invented steadily, most of them gone after short time

instead, we converge to:

- deliberate and justified choice of parameters, problems, performance criteria—much less arbitrariness
- better generalizability (not quite resolved, but targetted)

are we alone with this problem?

long tradition in natural sciences:

- many inventions (batteries, x-rays) made unintentionally while experimenting
- experiment leads to theory, theory must be useful -> predictions?

somewhat different in computer science:

- 2 well spread stereotypes influence our view onto computer experiments:

1. programs do exactly what the algorithms specify
2. computers are deterministic, so why statistics?

lessons from other sciences

in economics, experimentation was established quite recently (compared to its age):

- modeling human behavior as rationality assumption (of former theories) had failed
- no accepted new model available: experimentation came in as substitute

in (evolutionary) biology, experimentation and theory building both have problems:

- active experimentation only possible in special cases, otherwise only observation
- mainly concepts (working principles) instead of theories: always exceptions

⇒ stochastical distributions, population thinking

nonlinear behavior

Ernst Mayr

example from physics: looking for the top quark

- quarks are the constituents of protons and neutrons
- they come in 6 flavors: (up, down, strange, charm, bottom, and top)
- since the bottom quark was found experimentally in 1977, the top quark was postulated theoretically
- in 1994, ’t Hooft and Veltman estimated its mass to 145–185 GeV (Nobel prize 1999)
- one year it was actually found experimentally at Fermilab, IL, USA
- its real mass (measured) is now given as 172.9 ± 2.9 GeV
- this is based on the measurements of different experimental teams worldwide

picture from Ajale on Pixabay

unexpected experimentation

Viking bread baking (Lejre, Denmark)

since ≈1960s: Experimental Archaeology

- gather (e.g. performance) data that is not available otherwise
- task: concept validation, fill conceptual holes

experimentation in management of technology and product innovation

- product cycles are sped up by ‘fail-fast’, ‘fail-often’ experimentation
- what-if questions may be asked by using improved computational ressources
- innovation processes have to be tailored towards experimentation

Stefan H. Thomke

a recent example

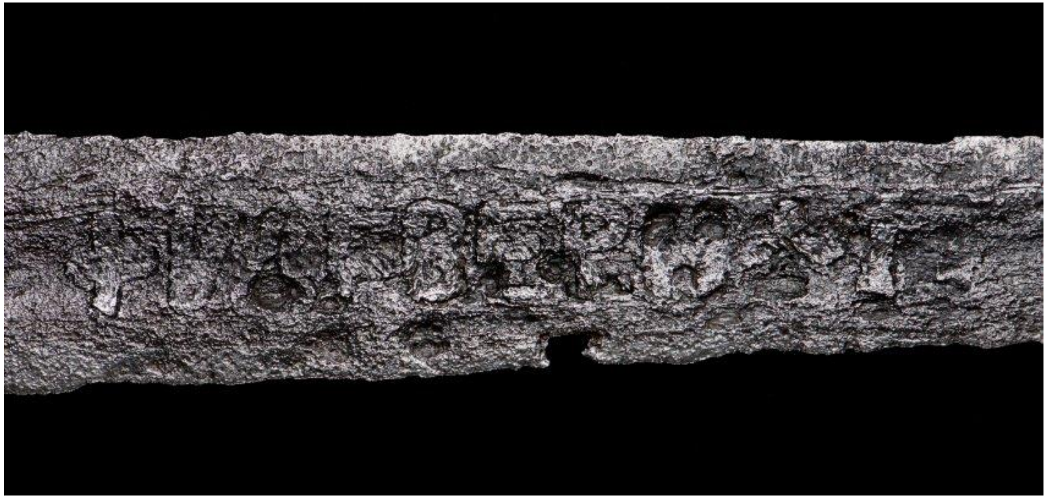
- from the 9. to the 11. century AD, the Franks produced “modern” swords with very high quality iron (near steel)
- these were named “Ulfberht swords” as they have an engraving with crosses and this name on the blade
- these swords were very popular also in Scandinavia
- the amount of blades found in Viking tombs is so large that archaeologists doubt that so many could have been produced in the middle German region
- additionally, especially later swords come with spelling mistakes
- recently, scientists presume that many of these swords are not original but fake
- but how difficult is it to make an “Ulfberht sword”?

➔ they performed experiments. can you guess what they tried?

Ulfberht sword (Hendrik Zwietasch / State Museum Württemberg, Stuttgart)

the experiment:

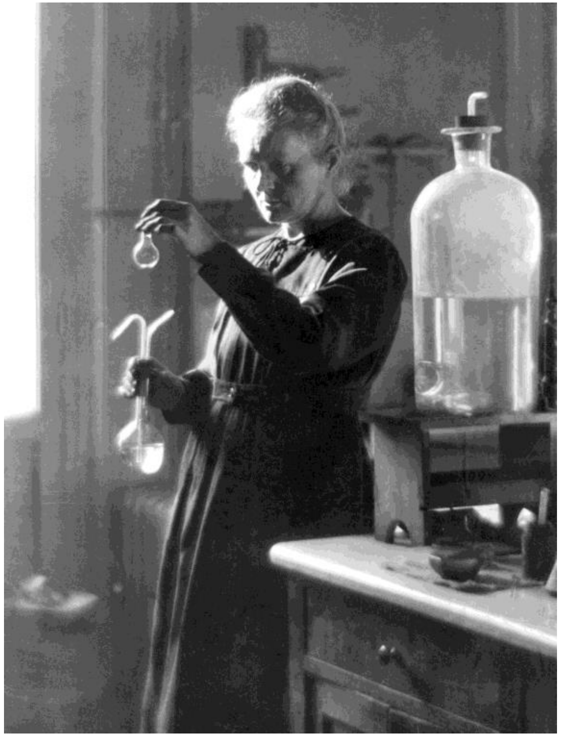
- they tried out if it is easily possible to take an existing sword and....
- just add the engraving to make it an “Ulfberht sword”
- they found it is dead simple
- this is no proof but strongly supports the hypothesis that a lot of fake swords may have been made
- medieval smiths just had to buy cheap swords and add the engraving



blade of an Ulfberht sword (Hendrik Zwietasch / State Museum Württemberg, Stuttgart)

ingredients for a good experiment

- fairness (even if we want to show that our method is better)
- openness (provide the means to get surprised)
- defined targets
  - how do we determine which method is the best (comparison)
  - what are the minimal conditions that must be reached?
- defined methodology (not ad-hoc)
- documentation (sufficient for replication)
- iteration (the first research question/hypothesis is usually not very good)



Marie Curie

untargeted and unstructured experimentation

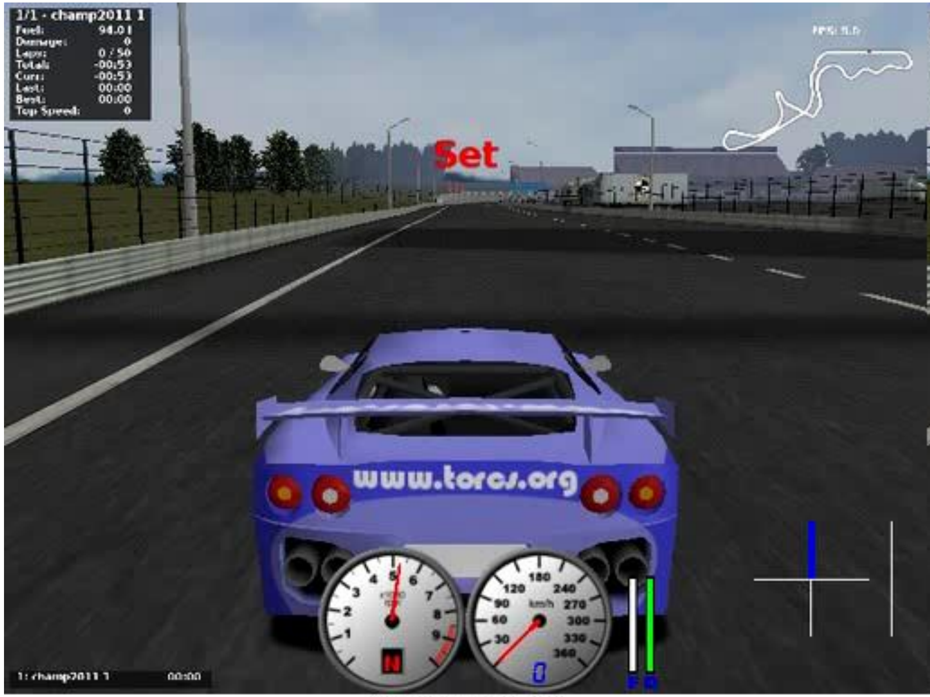
recall: *hypothesis* and *goals* are keywords of the definition

Cohens investigation of 1990 (all papers of the AAAI conference):

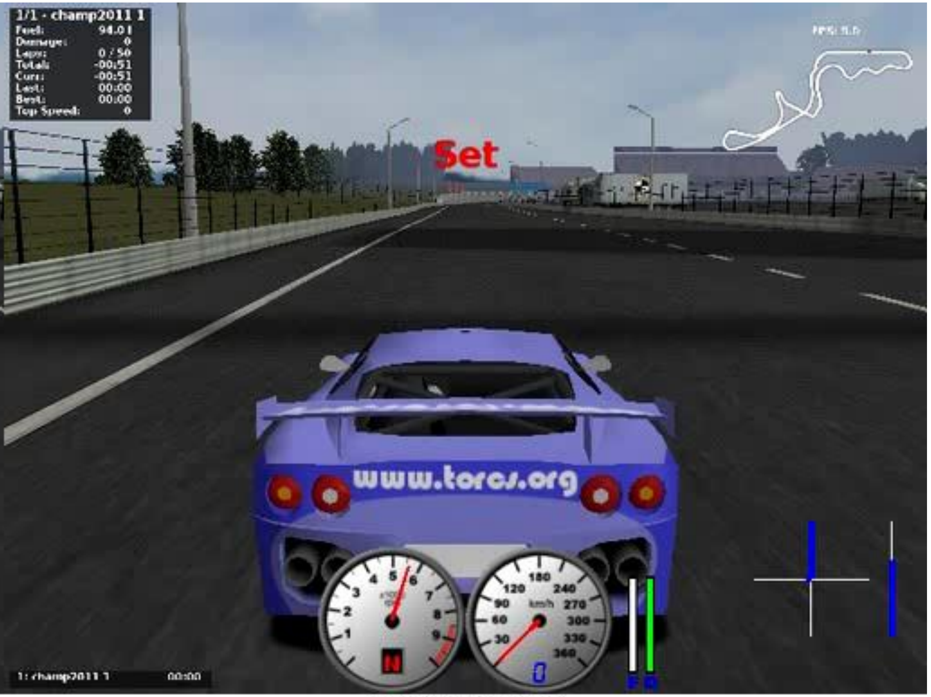
- almost no relation between theory and experiment
- 60% test on only one problem instance
- 80% report only the result, no explanation or interpretation
- 16% provide a hypothesis or define aims of the investigation

Paul R. Cohen

openness example: clutch control in torcs

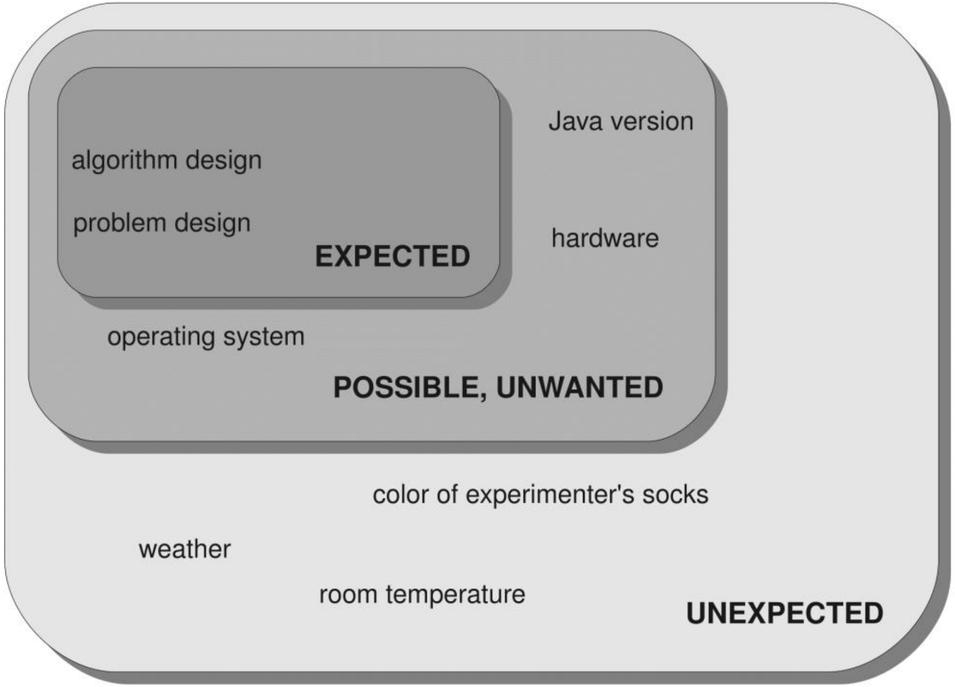


MrRacer w/o clutch



MrRacer with clutch

factors



Java version
algorithm design
problem design
EXPECTED
hardware
operating system
POSSIBLE, UNWANTED
color of experimenter's socks
weather
room temperature
UNEXPECTED

parameters and seeds

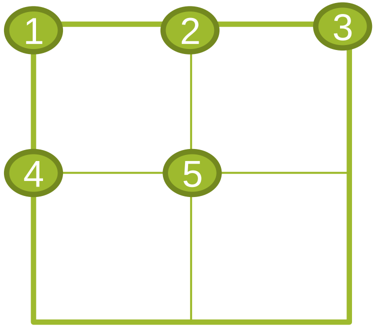
- many algorithms do have parameters
- solving a specific problem often requires parameter tuning
- it is tempting to prefer the own method (tune parameters on that method only)
- successful parameters for own method may be imposed for other methods
- example: population sizes for evolutionary optimization method
- benchmark sets can lead to over-adaptation (methods only good for these problems)
- generalization is hardly possible any more
- for fixed random seeds, the seeds become parameters of the method

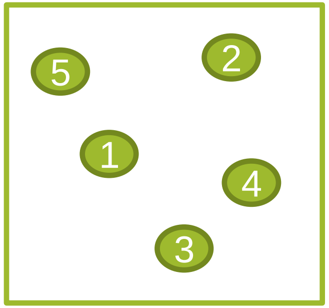
example result of a tuning process
picture from Pexels on Pixabay

parameter / hyperparameter investigation

Mel Gibson on the set of **Braveheart**, 1995 by Scott Neeson, creative commons 3

- first do some pretesting to see what parameters are most sensitive
- be **brave**, throw parameters away, you cannot test everything
- do a first 'experimental design' (set of configurations): grid or random
- more complex: *design of experiments* methods (less samples)
- take into account that non-determinism is involved! do repetitions, at least 5, better 20
- there are many more methods for hyperparameter tuning, e.g. SCMAC
- AutoML packages as e.g. Optuna optuna.org
- getting an idea of parameter interactions and finding the 'best' parameters is not the same!

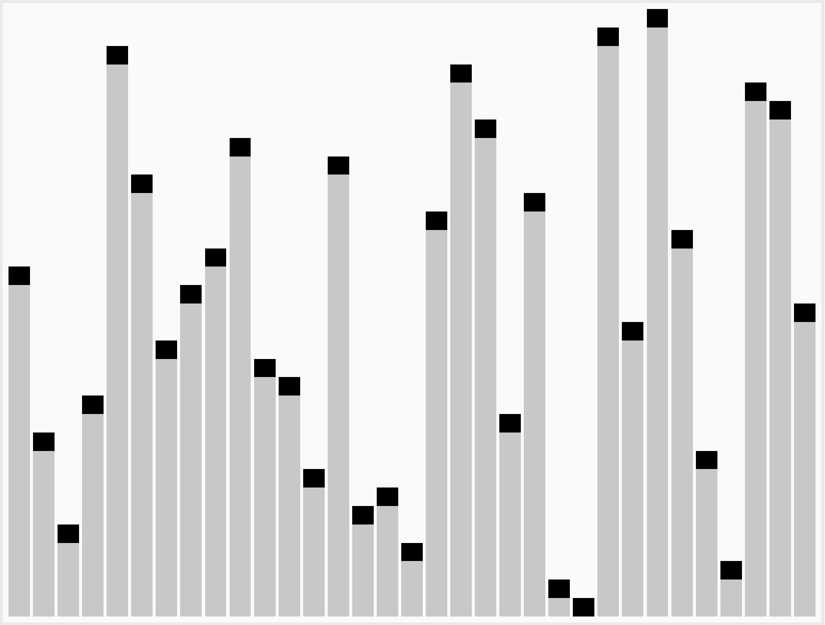




grid (top) and random (bottom) parameter design

randomization and noise

- we often think of computer programs generally as deterministic (always give the same result)
- however, many algorithms nowadays have randomized elements (very popular since the 80s)
- example: quick sort, we recursively choose a middle "pivot" element and sort into two parts
- then again we do that for the two subsets and so on, but how do we find "middle" elements?
- this is done by randomly choosing an element!
- there are improved versions which use the middle one of 3 randomly chosen elements
- single decisions may be bad, on average a randomized decision is not optimal but often ok



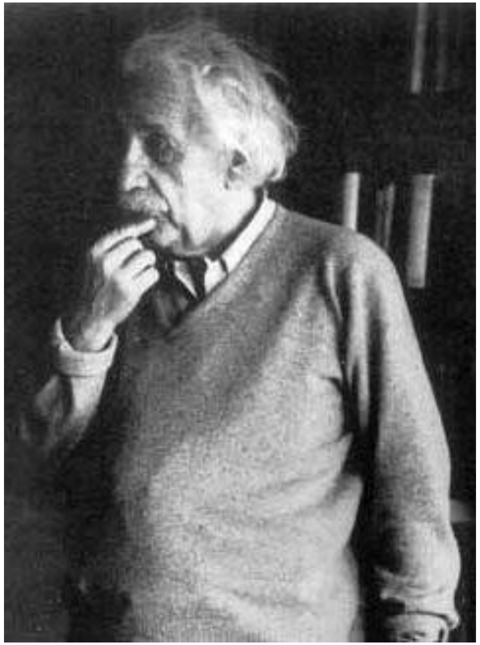
picture from Wikipedia user en:User:RolandH

research questions

- not trivial -> many investigations not focussed
- the true question is not if one method is better than another on a benchmark problem
- we want to tackle real-world problems

explaining observations leads to new questions:

- explaining models can be evaluated experimentally
- range of validity must be checked
  (problems, environmental conditions, parameters…)



somebody thinking

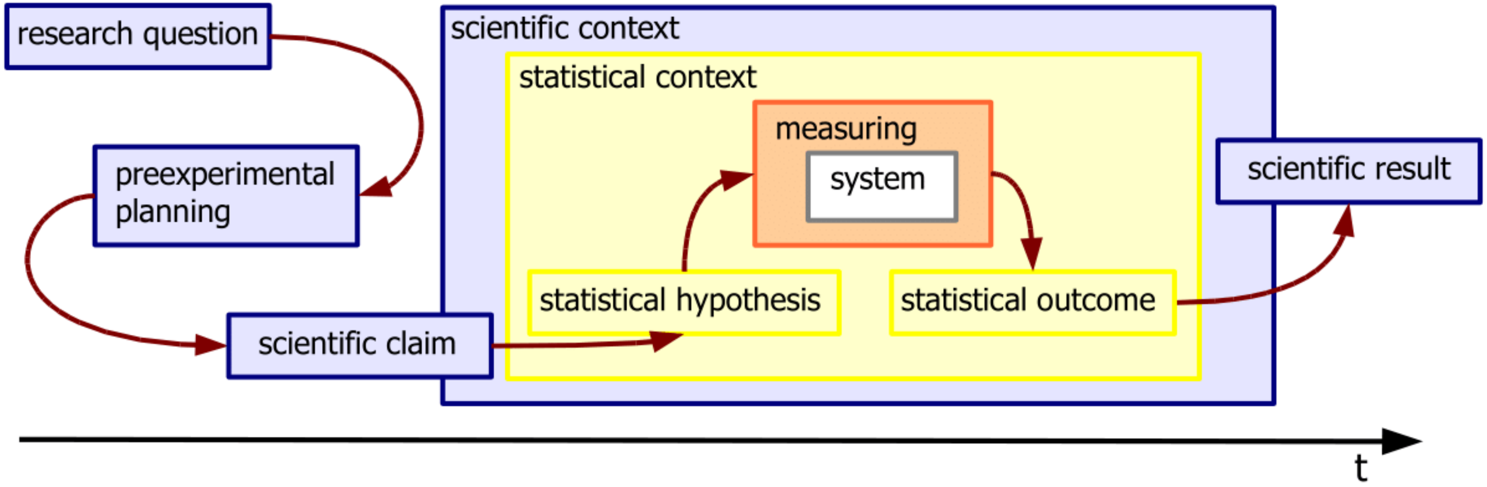
how do I find my research question?

- for comparisons: is the measured difference relevant in a real-world situation?
- for experimental studies: which quality must a method reach to be useful?
- usually, the perfect question is not known initially
  -> experimentation can help to find it
- an inherent problem of experimentation: we do not know the result (or we should not know it)
- but this can lead to new, better questions
- proceed in small steps, expect the unexpected

process of experimentation

how do we generate decision criteria from a research question?

- at first: set up scientific claims
- reshape into statistical hypotheses
- perform experiment, then transform backwards



hypothesis testing

- many papers now employ statistical testing
- but we claim: fundamental ideas from statistics are misunderstood!
- for example: what is the *p* value?

~~Definition (p value)~~
~~the p value is the probability that the null hypothesis is true~~

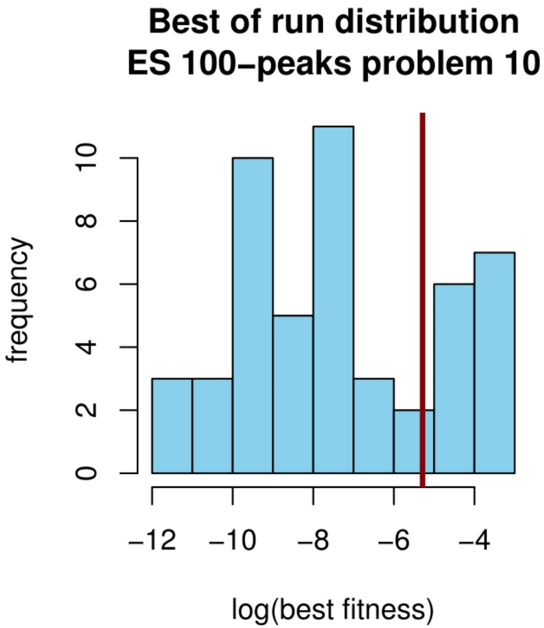
Definition (p value)
the p value is
$p = P\{\text{ obtain observed result, or greater } \mid \text{ null model is true }\}$

⇒ the *p* value is not related to any probability whether the null hypothesis is true or false

to test or not to test?

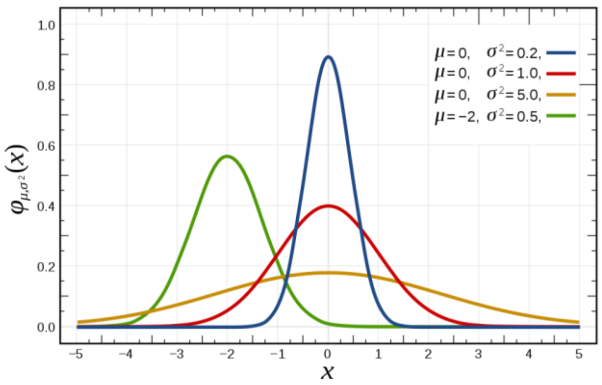
yes, but:

- we often have non-normal data
  ⇒ non-parametric tests, permutation tests
- temptation to “make” tests valid by enlarging sample (not always helpful, e.g. if distribution bimodal)
  ⇒ rule-of-thumb fixed size (e.g. 30)

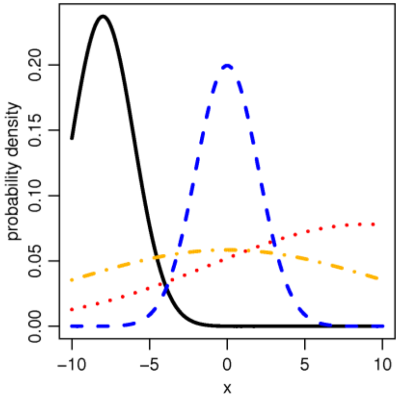


distributions

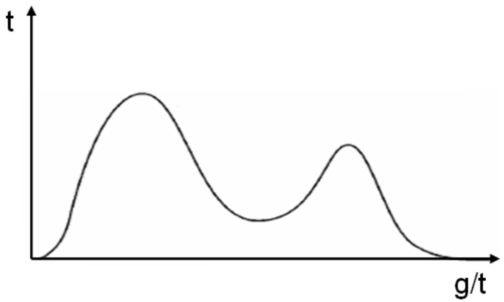
- in the real world, the normal distribution often is a good model
- but in learning/optimization we often have limits
- due to complicated effects, distributions may get bimodal
- or discrete: not all values can be realized
- don't trust small effects
- don't conclude from a running experiment!



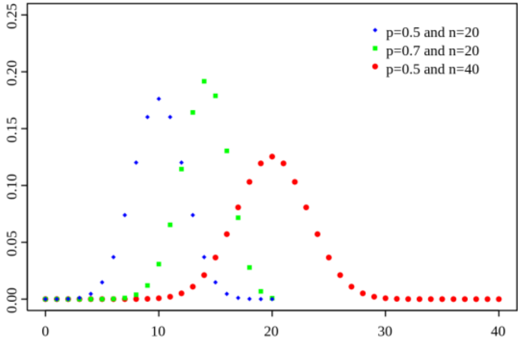
normal distributions



truncated normal distributions



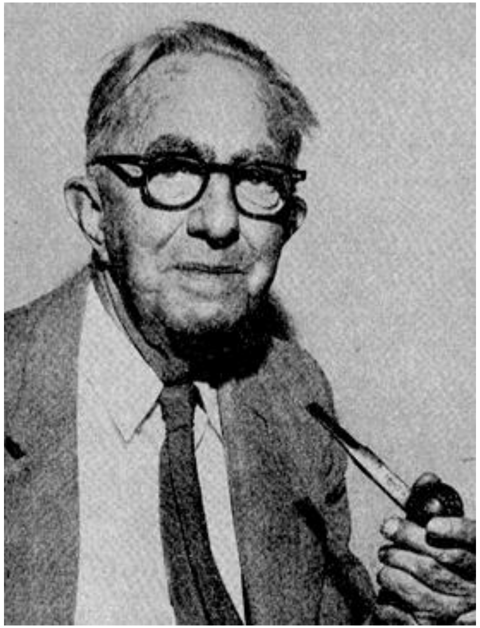
bimodal distribution



binomial distribution

Wilcoxon rank sum test

- aka Mann-Whitney U-test or just U-test (equivalent)
- more robust than t-test, now a standard test e.g. in Evolutionary Computation
- basic assumption: distribution functions G and F of X and Y only differ by a shift a, $G(x) = F(y - a)$
- this also means homogeneity of variances (may require F-test)!
- null hypothesis: $H_0 : a = 0$, $H_1 : a \neq 0$
- R-command:
  ```
  wilcox.test(x, y, alternative = "two.sided",
  conf.level = 0.95)
  ```
- also available in Excel or Python environments, e.g. SciPy



Frank Wilcoxon

earth is round ($p < 0.05$)

- paper of Jacob Cohen (in American Psychologist, 1994)
- summarizes criticism on ’unreflected’ use of statistical testing
- be careful with small samples!
- first understand and improve data (EDA, Exploratory Data Analysis, after Tukey), then testing
- actually, one should test the other way around: postulate null hypothesis and try to falsify it (very time-consuming procedure)
- providing confidence intervals gives important information!
- importance of reproducing a result

reporting and keeping track of experiments

around 40 years of experimental tradition in Computational Intelligence/Machine Learning, but:

- no standard scheme for reporting experiments (experimental protocols)
- instead: one (“Experiments”) or two (“Experimental Setup” and “Results”) sections in papers, often providing a bunch of largely unordered information
- affects readability and impairs reproducibility

keeping experimental journals helps:

- record context and rough idea
- report each experiment
- running where (machine)
- finished when (date/time), link to result file(s)

⇒ we suggest a 7-part reporting scheme (that is actually very much borrowed from Physics experiments)

experimental report

suggested structure:

1. **research question**: what do we investigate?
2. **pre-experimental planning** – first explorative ad-hoc expereriments to find target and setup (parameters etc.)
3. **task** – scientific and related statistical hypotheses – under which conditions is a method “successful”?
4. **setup** – exact setup of an experiment that enables replication
5. **results/visualizations** – tables, pictures – not interpreted
6. **observations** – peculiarities we find in the results
7. **discussion** – statistical test results, subjective interpretation of results and observations

floor and ceiling effects

- floor effect: compared methods attain set task very rarely ⇒ problem is too hard
- ceiling effect: methods nearly always reach given task ⇒ problem is too easy

if problem is too hard or too easy, nothing is shown.

- pre-experimentation is necessary to obtain reasonable tasks
- if task is reasonable (e.g. practical requirements), then algorithms are unsuitable (floor) or all good enough (ceiling), statistical testing does not provide more information
- arguing on minimal differences is statistically unsupported and scientifically meaningless

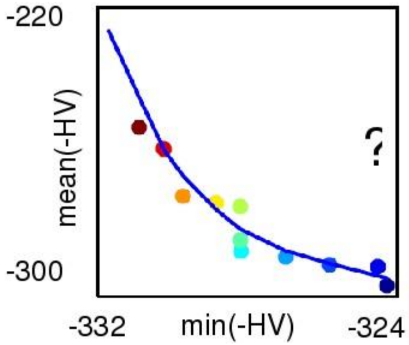
confounded effects

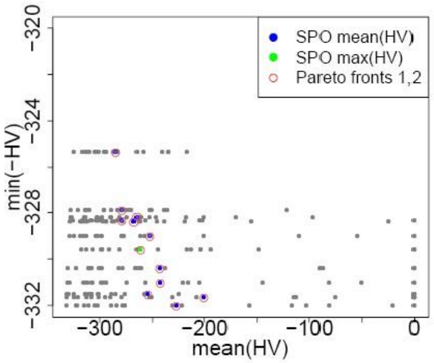
an algorithm is improved by means of 2 or more "extensions":

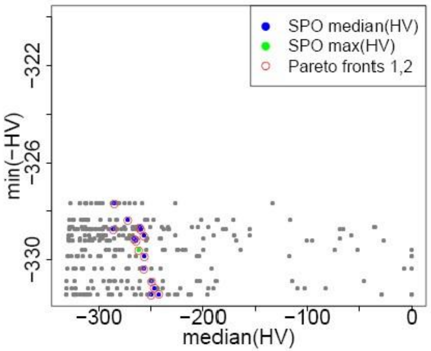
- what exactly leads to improvement?
- it is necessary to test the extensions separately
- possibly only the combination helps, or just one of the extensions?
- this knowledge is important for subsequent application

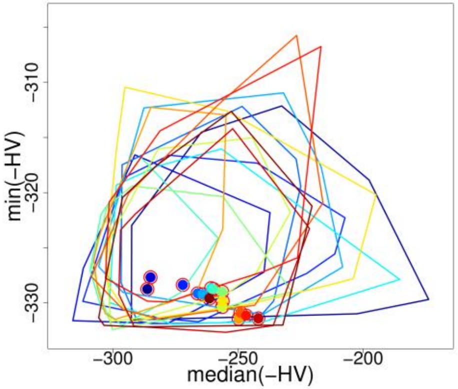
underestimated randomness

- idea: find pareto front of two parameter tuning criteria
- parameter changes not interpretable
- validation failed
- reason: deviations much too high!
- problem: human willingness to settle on a model







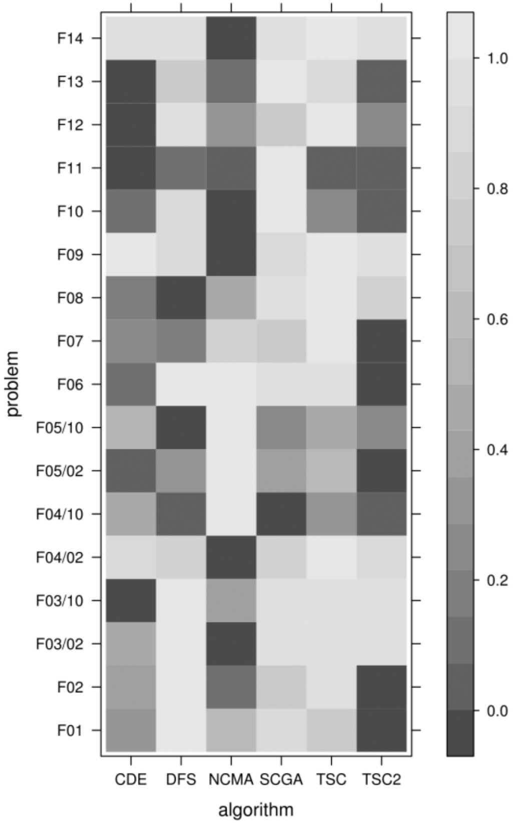


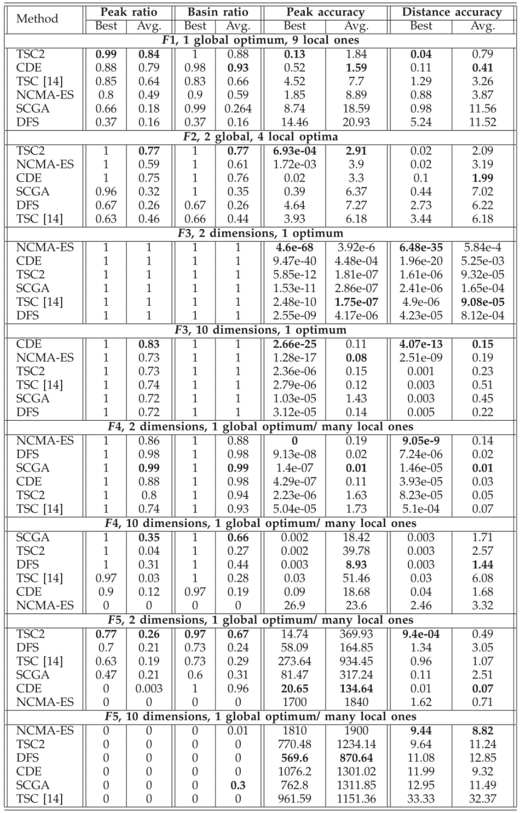
there is a problem with the experiment

after all data is in, we realize that something was wrong (code, parameters, environment?), what to do?

- current approach: either do not mention it, or redo everything
- if redoing is easy, nothing is lost
- if it is not, we must either:
  - let people know about it, explaining why it probably does not change results
  - or do validation on a smaller subset: how large is the difference (e.g. statistically significant)?
- do not worry, this situation is rather normal
- *Thomke*: there is nearly always a problem with an experiment
- early experimentation reduces the danger of something going completely wrong

diagrams instead of tables

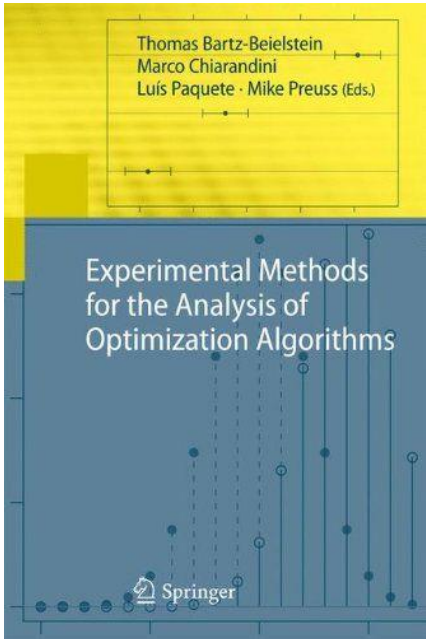


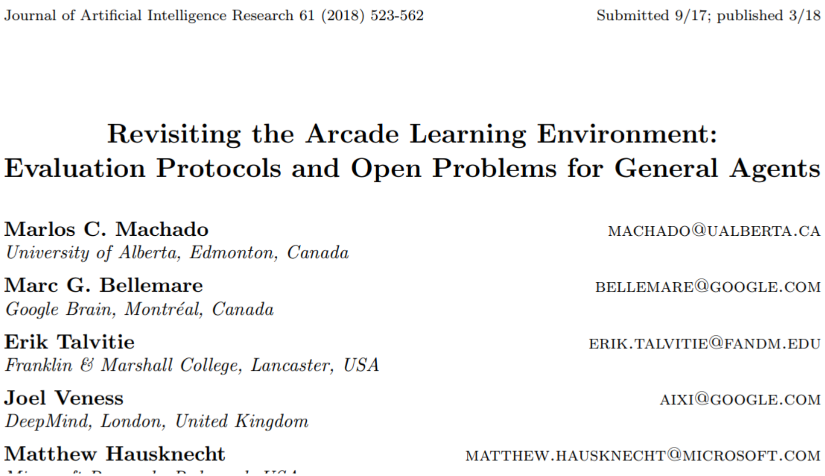


Method	Peak ratio		Basin ratio		Peak accuracy		Distance accuracy	
	Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.
***F*1, 1 global optimum, 9 local ones**								
TSC2	**0.99**	**0.84**	1	0.88	**0.13**	1.84	**0.04**	0.79
CDE	0.88	0.79	0.98	**0.93**	0.52	**1.59**	0.11	**0.41**
TSC [14]	0.85	0.64	0.83	0.66	4.52	7.7	1.29	3.26
NCMA-ES	0.8	0.49	0.9	0.59	1.85	8.89	0.88	3.87
SCGA	0.66	0.18	0.99	0.264	8.74	18.59	0.98	11.56
DFS	0.37	0.16	0.37	0.16	14.46	20.93	5.24	11.52
***F*2, 2 global, 4 local optima**								
TSC2	1	**0.77**	1	**0.77**	**6.93e-04**	**2.91**	0.02	2.09
NCMA-ES	1	0.59	1	0.61	1.72e-03	3.9	0.02	3.19
CDE	1	0.75	1	0.76	0.02	3.3	0.1	**1.99**
SCGA	0.96	0.32	1	0.35	0.39	6.37	0.44	7.02
DFS	0.67	0.26	0.67	0.26	4.64	7.27	2.73	6.22
TSC [14]	0.63	0.46	0.66	0.44	3.93	6.18	3.44	6.18
***F*3, 2 dimensions, 1 optimum**								
NCMA-ES	1	1	1	1	**4.6e-68**	3.92e-6	**6.48e-35**	5.84e-4
CDE	1	1	1	1	9.47e-40	4.48e-04	1.96e-20	5.25e-03
TSC2	1	1	1	1	5.85e-12	1.81e-07	1.61e-06	9.32e-05
SCGA	1	1	1	1	1.53e-11	2.86e-07	2.41e-06	1.65e-04
TSC [14]	1	1	1	1	2.48e-10	**1.75e-07**	4.9e-06	**9.08e-05**
DFS	1	1	1	1	2.55e-09	4.17e-06	4.23e-05	8.12e-04
***F*3, 10 dimensions, 1 optimum**								
CDE	1	**0.83**	1	1	**2.66e-25**	0.11	**4.07e-13**	**0.15**
NCMA-ES	1	0.73	1	1	1.28e-17	**0.08**	2.51e-09	0.19
TSC2	1	0.73	1	1	2.36e-06	0.15	0.001	0.23
TSC [14]	1	0.74	1	1	2.79e-06	0.12	0.003	0.51
SCGA	1	0.72	1	1	1.03e-05	1.43	0.003	0.45
DFS	1	0.72	1	1	3.12e-05	0.14	0.005	0.22
***F*4, 2 dimensions, 1 global optimum/ many local ones**								
NCMA-ES	1	0.86	1	0.88	**0**	0.19	**9.05e-9**	0.14
DFS	1	0.98	1	0.98	9.13e-08	0.02	7.24e-06	0.02
SCGA	1	**0.99**	1	**0.99**	1.4e-07	**0.01**	1.46e-05	**0.01**
CDE	1	0.88	1	0.98	4.29e-07	0.11	3.93e-05	0.03
TSC2	1	0.8	1	0.94	2.23e-06	1.63	8.23e-05	0.05
TSC [14]	1	0.74	1	0.93	5.04e-05	1.73	5.1e-04	0.07
***F*4, 10 dimensions, 1 global optimum/ many local ones**								
SCGA	1	**0.35**	1	**0.66**	0.002	18.42	0.003	1.71
TSC2	1	0.04	1	0.27	0.002	39.78	0.003	2.57
DFS	1	0.31	1	0.44	0.003	**8.93**	0.003	**1.44**
TSC [14]	0.97	0.03	1	0.28	0.03	51.46	0.03	6.08
CDE	0.9	0.12	0.97	0.19	0.09	18.68	0.04	1.68
NCMA-ES	0	0	0	0	26.9	23.6	2.46	3.32
***F*5, 2 dimensions, 1 global optimum/ many local ones**								
TSC2	**0.77**	**0.26**	**0.97**	**0.67**	14.74	369.93	**9.4e-04**	0.49
DFS	0.7	0.21	0.73	0.24	58.09	164.85	1.34	3.05
TSC [14]	0.63	0.19	0.73	0.29	273.64	934.45	0.96	1.07
SCGA	0.47	0.21	0.6	0.31	81.47	317.24	0.11	2.51
CDE	0	0.003	1	0.96	**20.65**	**134.64**	0.01	**0.07**
NCMA-ES	0	0	0	0	1700	1840	1.62	0.71
***F*5, 10 dimensions, 1 global optimum/ many local ones**								
NCMA-ES	0	0	0	0.01	1810	1900	**9.44**	**8.82**
TSC2	0	0	0	0	770.48	1234.14	9.64	11.24
DFS	0	0	0	0	**569.6**	**870.64**	11.08	12.85
CDE	0	0	0	0	1076.2	1301.02	11.99	9.32
SCGA	0	0	0	**0.3**	762.8	1311.85	12.95	11.49
TSC [14]	0	0	0	0	961.59	1151.36	33.33	32.37

some links

David S. Johnson: a theoretician's guide to the experimental analysis of algorithms (last version 2001)


Thomas Bartz-Beielstein
Marco Chiarandini
Luís Paquete · Mike Preuss (Eds.)
Experimental Methods for the Analysis of Optimization Algorithms
Springer



Journal of Artificial Intelligence Research 61 (2018) 523-562 Submitted 9/17; published 3/18

Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents

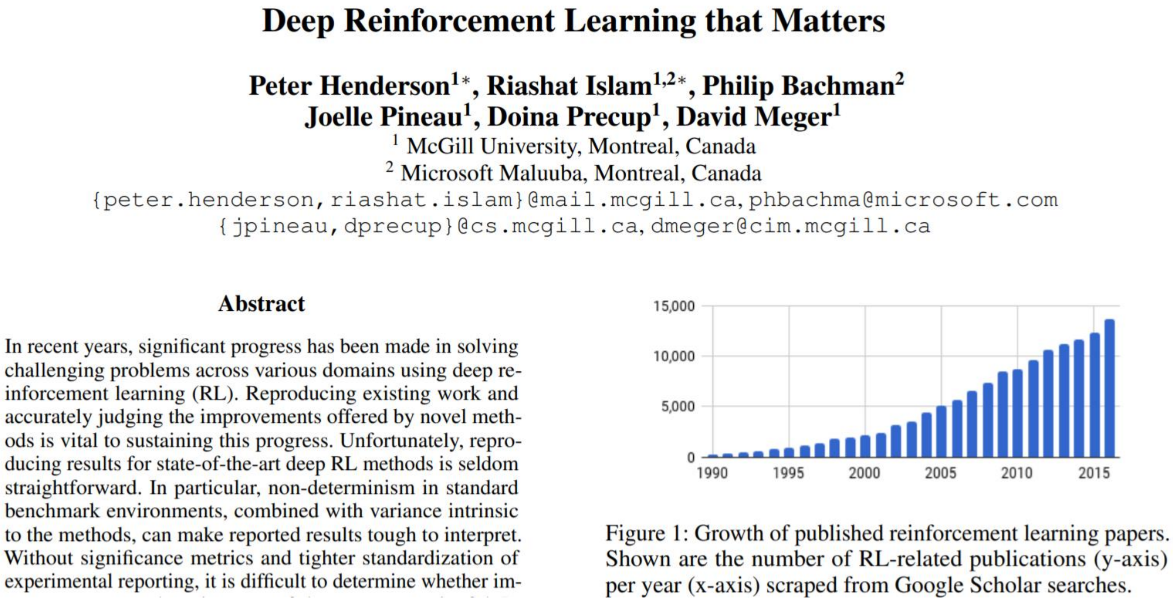
Marlos C. Machado MACHADO@UALBERTA.CA
University of Alberta, Edmonton, Canada

Marc G. Bellemare BELLEMARE@GOOGLE.COM
Google Brain, Montréal, Canada

Erik Talvitie ERIK.TALVITIE@FANDM.EDU
Franklin & Marshall College, Lancaster, USA

Joel Veness AIXI@GOOGLE.COM
DeepMind, London, United Kingdom

Matthew Hausknecht MATTHEW.HAUSKNECHT@MICROSOFT.COM



Deep Reinforcement Learning that Matters

Peter Henderson[1]*, Riashat Islam[1,2]*, Philip Bachman[2]
Joelle Pineau[1], Doina Precup[1], David Meger[1]
[1] McGill University, Montreal, Canada
[2] Microsoft Maluuba, Montreal, Canada
{peter.henderson,riashat.islam}@mail.mcgill.ca, phbachma@microsoft.com
{jpineau,dprecup}@cs.mcgill.ca, dmeger@cim.mcgill.ca

Abstract

In recent years, significant progress has been made in solving challenging problems across various domains using deep reinforcement learning (RL). Reproducing existing work and accurately judging the improvements offered by novel methods is vital to sustaining this progress. Unfortunately, reproducing results for state-of-the-art deep RL methods is seldom straightforward. In particular, non-determinism in standard benchmark environments, combined with variance intrinsic to the methods, can make reported results tough to interpret. Without significance metrics and tighter standardization of experimental reporting, it is difficult to determine whether im-

Figure 1: Growth of published reinforcement learning papers. Shown are the number of RL-related publications (y-axis) per year (x-axis) scraped from Google Scholar searches.

take home

- meaningful experimentation is difficult: some structure is needed
- experimentation is the only way to work with methods that do not possess enough theory
- structure is important: research question, targeted experiments, statistical tests, proper reporting

If in doubt, try it out!

picture from OpenClipart-Vectors on Pixabay