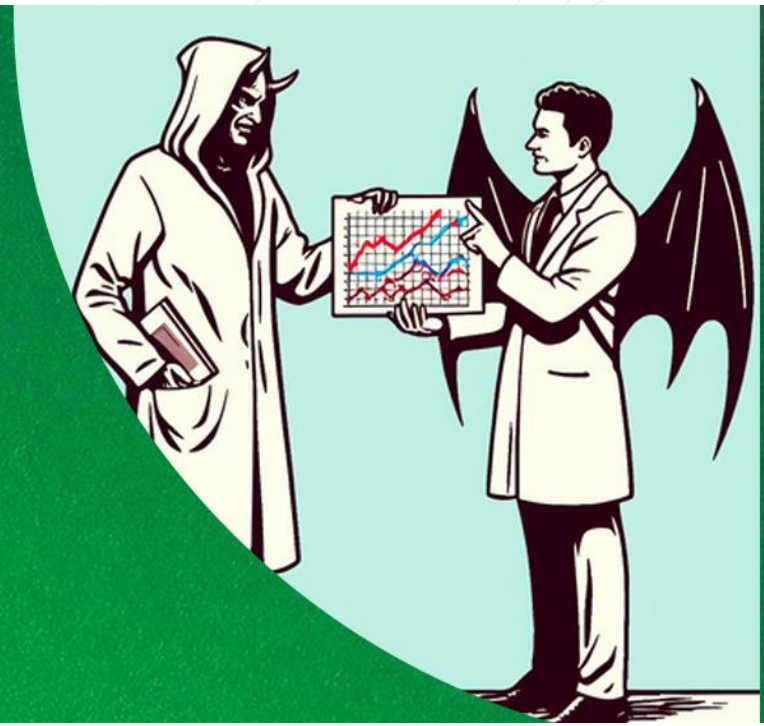
Palacký University
Olomouc

FTK UP pořádá odbornou diskusi

Replikační krize ve vědě: 7 (ne)smrtelných hříchů moderní statistiky

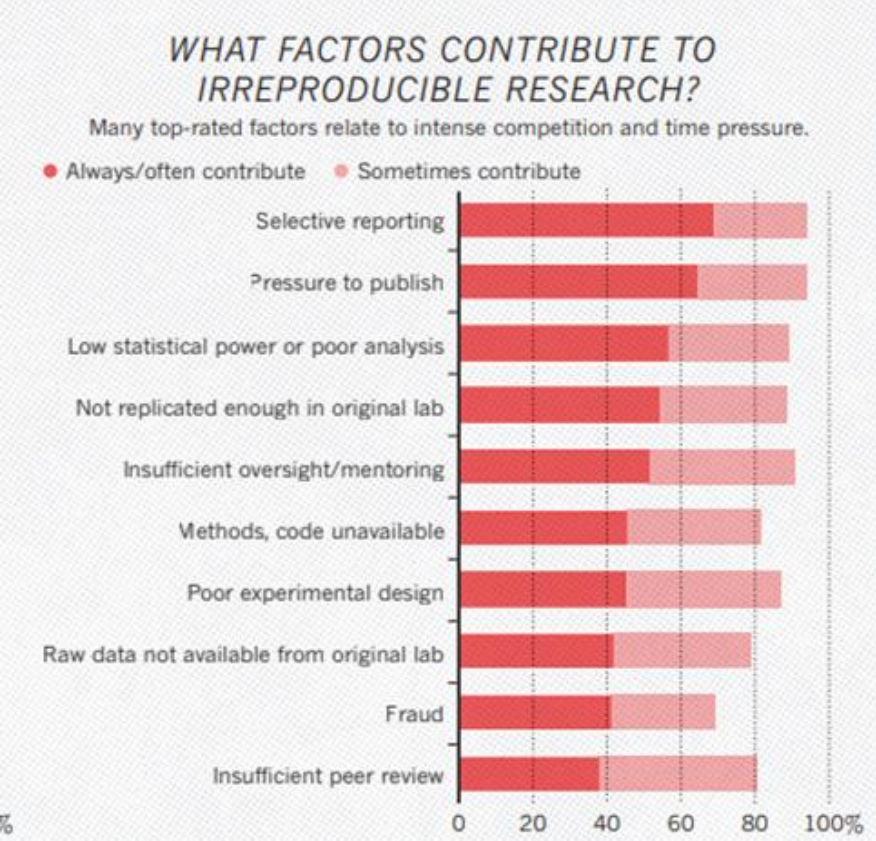


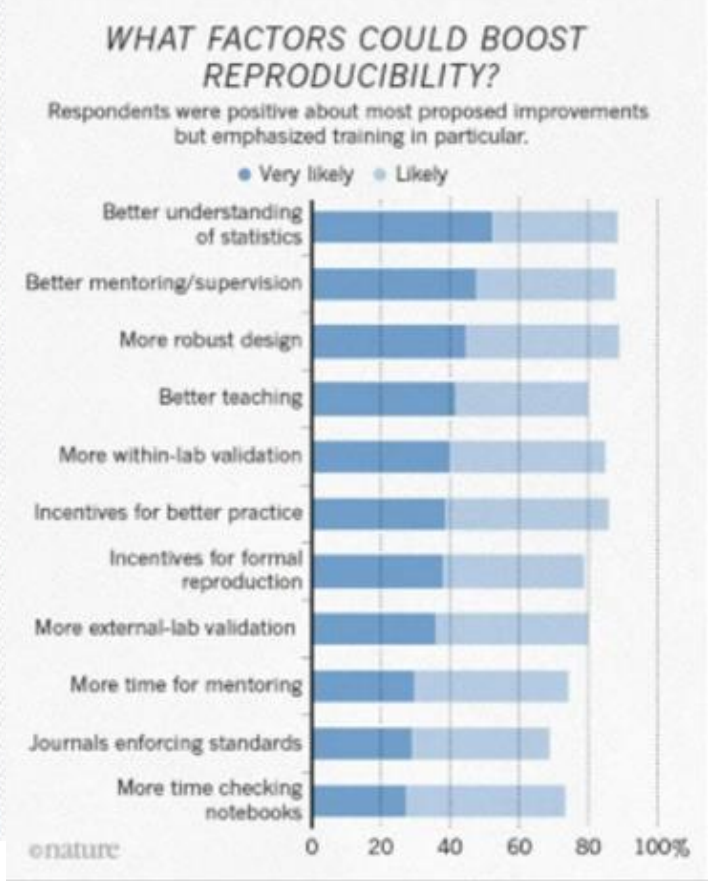
Změna termínu:
25. března 2025 13:30-15:00 NA 2.10

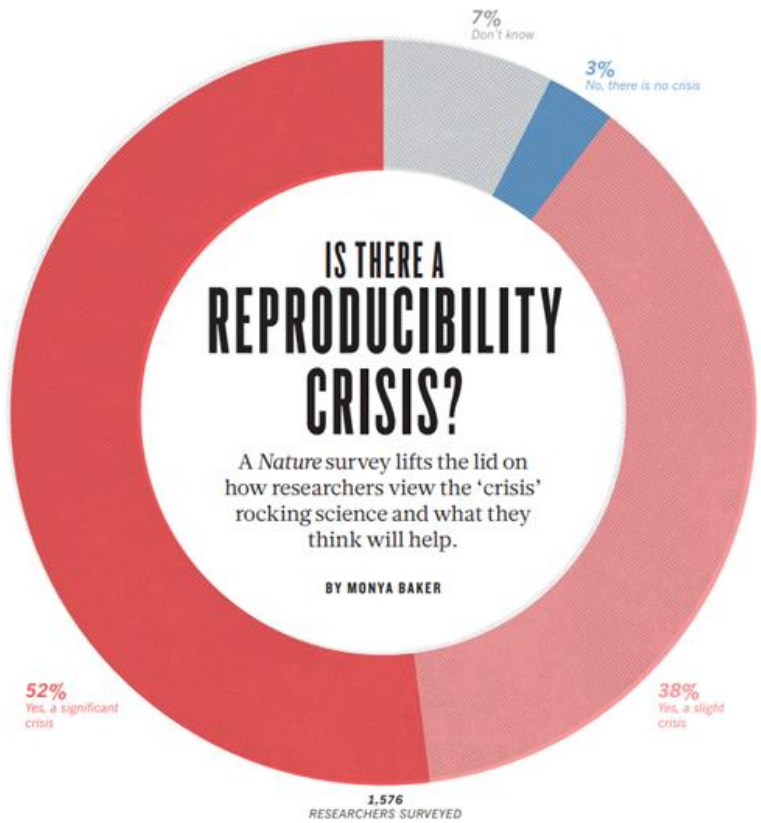
Fakulta tělesné kultury
Univerzita Palackého v Olomouci

Předpokládaný časový harmonogram

- 13:30-13:50 přivítání a úvod: Ladislav Baloun
  (Proč řešit 7 nesmrtelných hříchů moderní statistiky?).
- 13:50-14:05 tematická přednáška: Tomáš Fürst
  (13 lekcí o vědě, které by měl slyšet každý prvák na UP).
- 14:05-14:20 tematická přednáška: Ondřej Vencálek
  (Trable s p-values).
- 14:20-15:00 otevřená diskuse.







Proč sedm (ne)smrtelných hříchu moderní statistiky?

Ladislav Baloun, FTK UP

ladislav.baloun@upol.cz

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353-66

Palacký University
Olomouc

7 hříchu statistické inference

1. Málo časové dotace pro výuku statistiky
2. Špatné představy o tom jak data vznikají
3. Smal sample size
4. Předpoklad objektivních výsledků: subjektivní analýza objektivních dat.
5. Dichotomizace výsledků – obsese pro pozitivní výsledky
6. P hacking
7. Harking
8. Selektivní reportování

Předpoklad objektivních výsledků: subjektivní analýza objektivních dat.

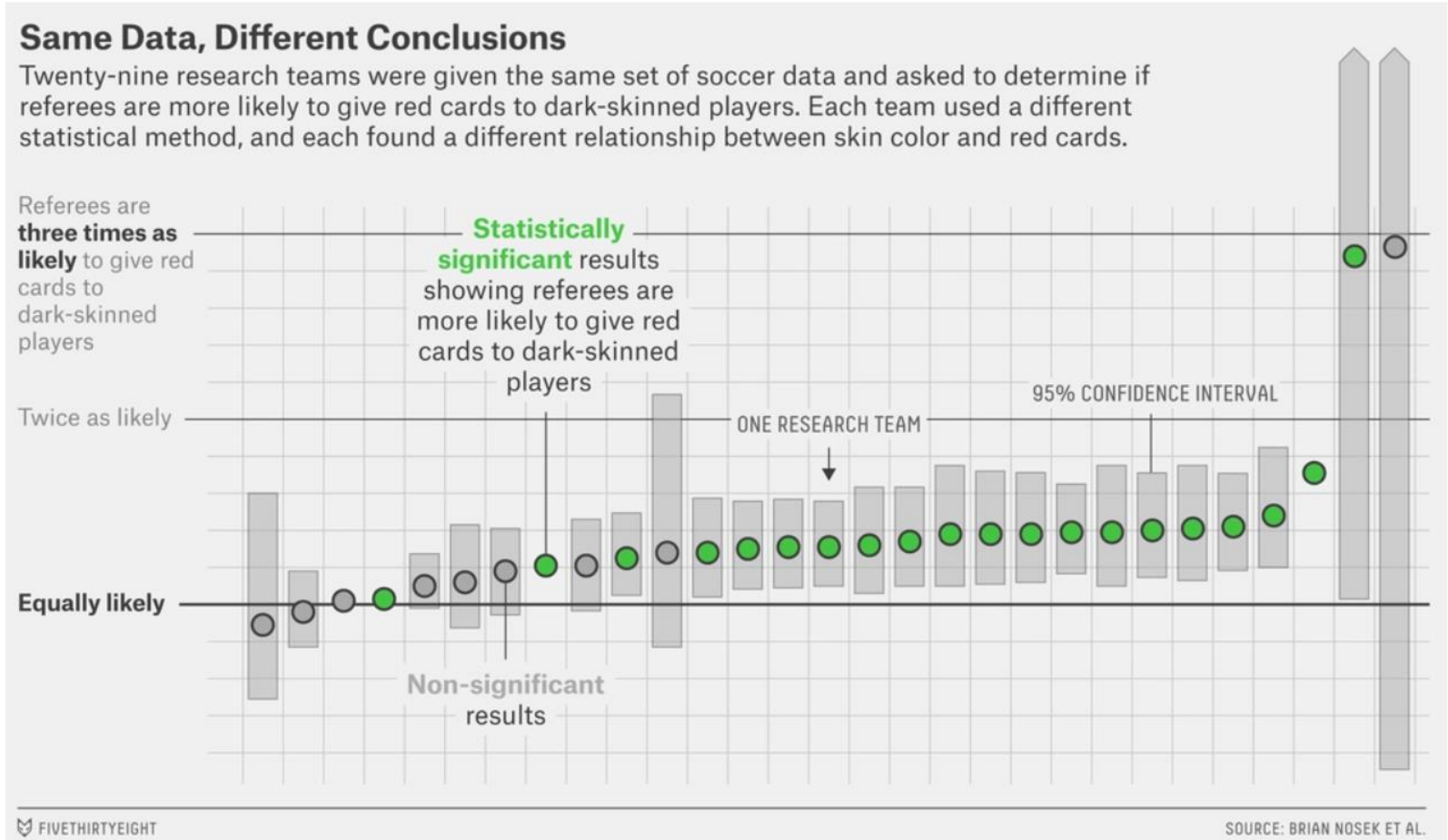
- ….data analysts should recognize that subjectivity and potential bias are inherent in all data analysis, exploratory or otherwise.
- …datoví analytici by si měli uvědomit, že subjektivita a potenciální zkreslení jsou vlastní veškeré analýze dat, průzkumné i jiné.
- One great danger in overmathematizing data analysis is believing that the reliability and precision of mathematics itself imbue reliability and precision to the data and the data analysis.
- Jedním velkým nebezpečím při přematematizaci analýzy dat je věřit, že spolehlivost a přesnost matematiky sama o sobě dodává spolehlivost a přesnost datům a analýze dat.

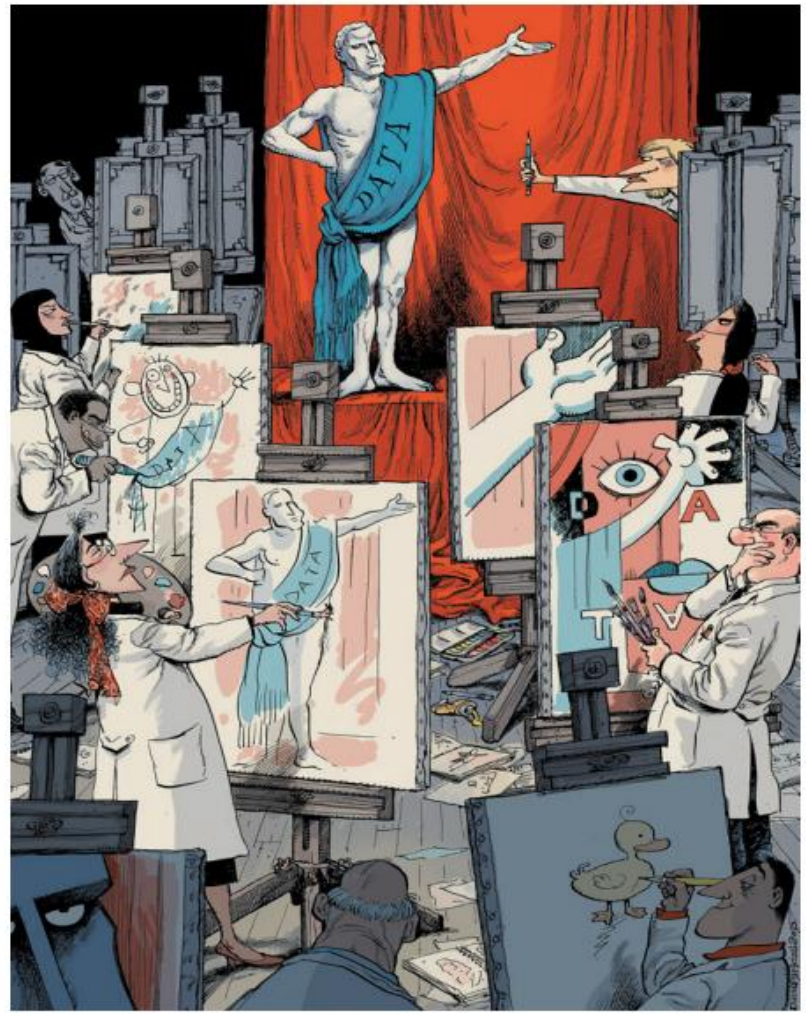
Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological methods*, *2*(2), 131.

Palacký University
Olomouc

Reproducibility - How do we work with data?

Silberzahn et al. (2018): different approaches, different type of analysis.





Many hands make tight work

Zdroj obrázku:https://fivethirtyeight.com/features/science-isnt-broken/#part2

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189-191.

How do we work with data?

There are two primary explanations for variation in forking decisions. The competency hypothesis posits that researchers may make different analytical decisions because of varying levels of statistical and subject expertise that lead to different judgments as to what constitutes the "ideal" analysis in a given research situation. The confirmation bias hypothesis holds that researchers may make reliably different analytical choices because of differences in preexisting beliefs and attitudes, which may lead to justification of analytical approaches favoring certain outcomes post hoc. However, many other covert or idiosyncratic influences, large and small, may also lead to unreliable and unexplainable variation in analytical decision pathways (10). Sometimes even the tiniest of these differences may add up and interact to produce widely varying outcomes.

RESEARCH ARTICLE | SOCIAL SCIENCES |

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, +162, and Tomasz Żółtak Authors Info & Affiliations

Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

October 28, 2022 | 119 (44) e2203150119 | https://doi.org/10.1073/pnas.2203150119

VIEW RELATED CONTENT +

73 výzkumných týmů (identická data z mezinárodního šetření)

hypotéza: **větší imigrace sníží veřejnou podporu pro vládní poskytování sociálních politik**

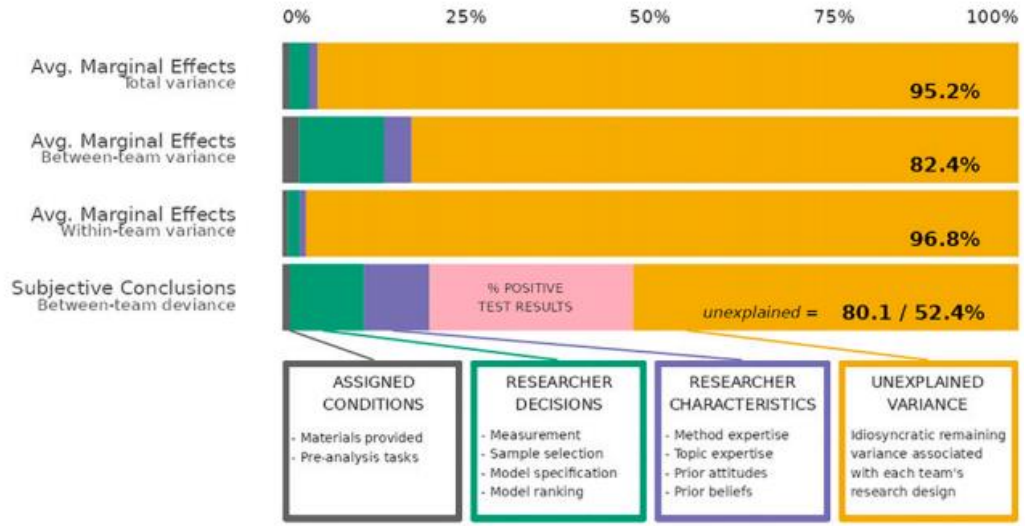


Fig. 1. Broad variation in the findings from 73 teams testing the same hypothesis with the same data. The distribution of estimated AMEs across all converged models (n = 1,253) includes results that are negative (yellow; in the direction predicted by the given hypothesis the teams were testing), not different from zero (gray), or positive (blue) using a 95% CI. AME are xy standardized. The y axis contains two scaling breaks at ±0.05. Numbers inside circles represent the percentages of the distribution of each outcome inversely weighted by the number of models per team.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119.

Palacký University
Olomouc

How do we work with data?

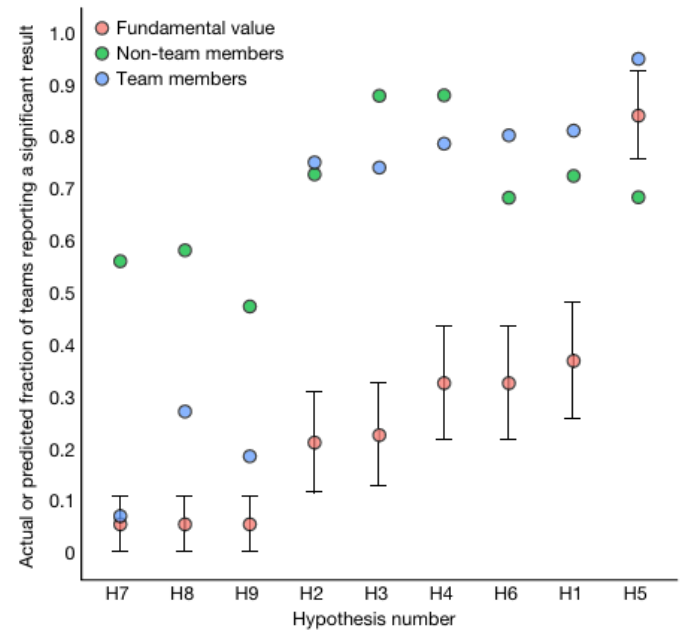


Fig. 1 | Fraction of teams reporting a significant result and prediction market beliefs. The observed fraction of teams reporting significant results (fundamental value, pink dots; $n = 70$ analysis teams), as well as final market prices for the team members markets (blue dots; $n = 83$ active traders) and the non-team members markets (green dots; $n = 65$ active traders). The corresponding 95% confidence intervals are shown for each of the nine hypotheses (note that hypotheses are sorted on the basis of the fundamental value). Confidence intervals were constructed by assuming convergence of the binomial distribution towards the normal.

Variability in the analysis of a single neuroimaging dataset by many teams

https://doi.org/10.1038/s41586-020-2314-9
Received: 14 November 2019
Accepted: 7 April 2020
Published online: 20 May 2020
Check for updates

A list of authors and affiliations appears in the online version of the paper.

Data analysis workflows in many scientific domains have become increasingly complex and flexible. Here we assess the effect of this flexibility on the results of functional magnetic resonance imaging by asking 70 independent teams to analyse the same dataset, testing the same 9 ex-ante hypotheses[1]. The flexibility of analytical approaches is exemplified by the fact that no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of hypothesis tests, even for teams whose statistical maps were highly correlated at intermediate stages of the analysis pipeline. Variation in reported results was related to several aspects of analysis methodology. Notably, a meta-analytical approach that aggregated information across teams yielded a significant consensus in activated regions. Furthermore, prediction markets of researchers in the field revealed an overestimation of the likelihood of significant findings, even by researchers with direct knowledge of the dataset[2–5]. Our findings show that analytical flexibility can have substantial effects on scientific conclusions, and identify factors that may be related to variability in the analysis of functional magnetic resonance imaging. The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for performing and reporting multiple analyses of the same data. Potential approaches that could be used to mitigate issues related to analytical variability are discussed.

fMRI data from 108 individuals
70 teams
9 hypotheses

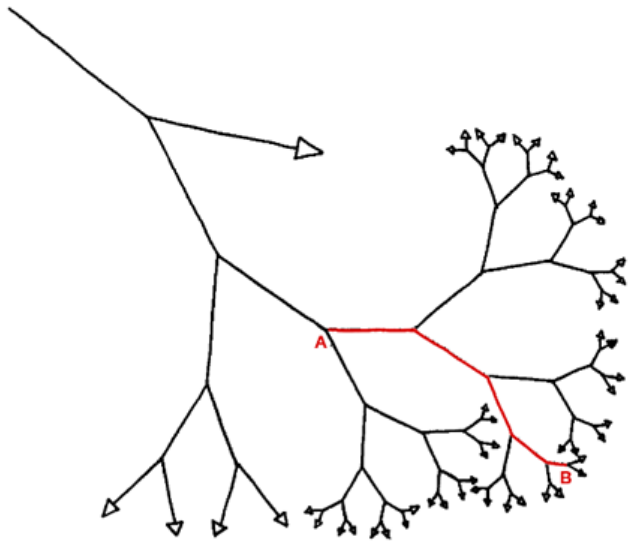
Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F. et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88 (2020). https://doi.org/10.1038/s41586-020-2314-9

Palacký University
Olomouc

Zahrada rozvětvených cestiček

Researcher degrees of freedom (Stupně volnosti výzkumníka)

The Garden of Forking Paths
by Jorge Luis Borges



A dataset can be analyzed in so many different ways (with the choices being not just what statistical test to perform but also decisions on what data to exclude or exclude, what measures to study, what interactions to consider, etc.),

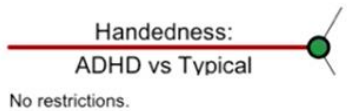
Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*(1-17)
http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

Palacký University
Olomouc

Zahrada rozvětvených cestiček

Researcher degrees of freedom (Stupně volnosti výzkumníka)

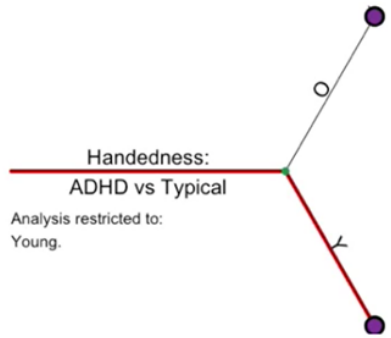
Large population database used to explore link between ADHD and handedness



1 contrast

Probability of a 'significant' p-value < .05 = .05

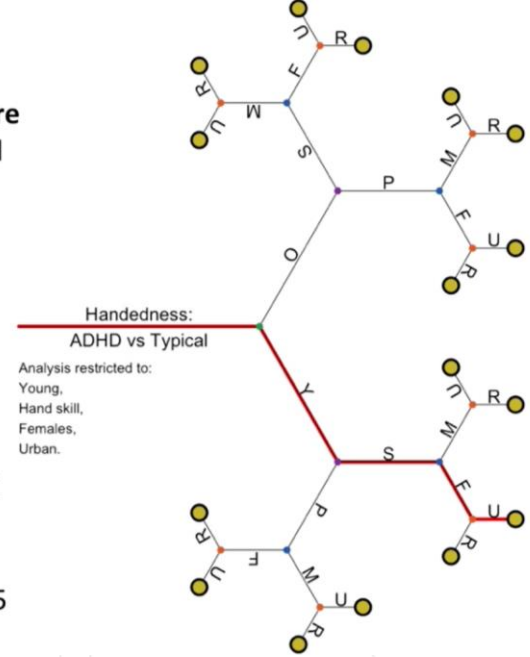
Large population database used to explore link between ADHD and handedness



Focus just on Young subgroup:
2 contrasts at this level

Probability of a 'significant' p-value < .05 = .10

Large population database used to explore link between ADHD and handedness



Focus just on Young, Urban, Females on measure of hand skill:
16 contrasts at this level

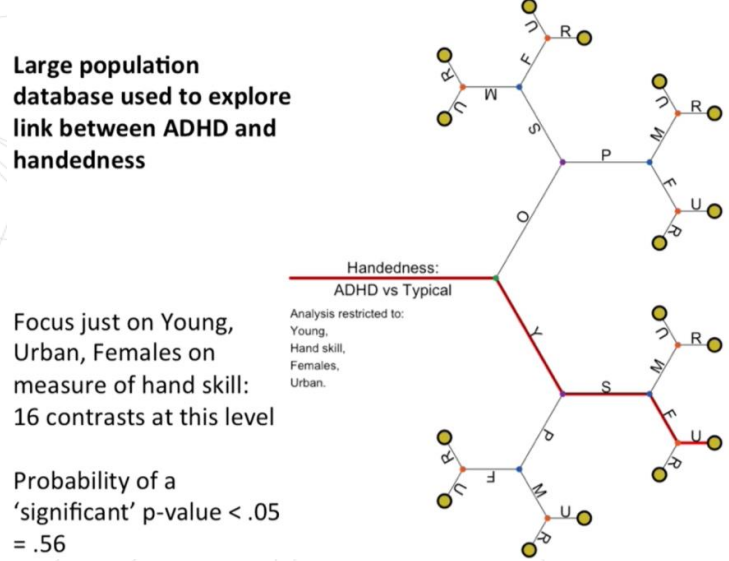
Probability of a 'significant' p-value < .05 = .56

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*(1-17)
http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

Palacký University
Olomouc

Zahrada rozvětvených cestiček

Researcher degrees of freedom (Stupně volnosti výzkumníka)



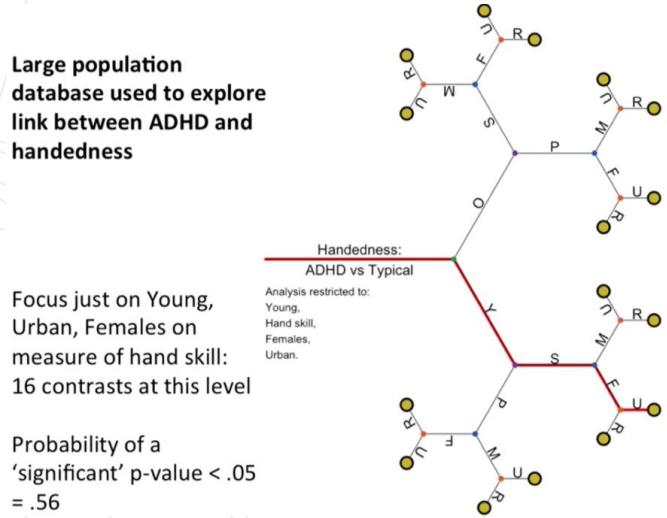
- Výzkumníci nezkouší více testů, aby zjistili, který má nejlepší p-hodnotu; spíše používají svůj vědecký zdravý rozum, aby formulovali své hypotézy rozumným způsobem, s ohledem na data, která mají.
- Chyba je v domnění, že pokud konkrétní cesta, která byla zvolena, poskytuje statistickou významnost, že je to silný důkaz ve prospěch hypotézy.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*(1-17)
http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

Palacký University
Olomouc

Zahrada rozvětvených cestiček

Researcher degrees of freedom (Stupeně volnosti výzkumníka)



- Chceme objasnit, že vícenásobné srovnání může být velkým problémem, aniž by to znamenalo, že dotyční výzkumníci podvádějí nebo jsou hloupí nebo se snaží systém manipulovat.
- Když jsme tyto druhy výzkumu popsali jako rybářské výpravy, udělali jsme chybu.
- Fishing a p-hacking znamenají aktivní snahu o statistickou významnost, zatímco to, o co by se zde mohlo stát, je soubor možností analýzy dat, které by mohly být rozumné, nebýt problémů s malou velikostí vzorku a chybou měření, kvůli kterým jsou výsledky hlučnější, než si lidé uvědomují.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*(1-17)
http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

Palacký University
Olomouc

Multiverse Analysis

- A multiverse analysis starts from the observation that data used in an analysis are usually not just passively recorded in an experiment or an observational study.
- Rather, data are to a certain extent actively constructed.
- Data construction occurs when the raw data are converted into a form ready for analysis.
- When preparing their data for analysis, researchers often take several processing steps, such as discretization of variables into categories, combination of variables, transformation of variables, data exclusion, and so on.
- These processing steps typically come with many researcher degrees of freedom.

Perspectives on Psychological Science
2016, Vol. 11(5) 702–712

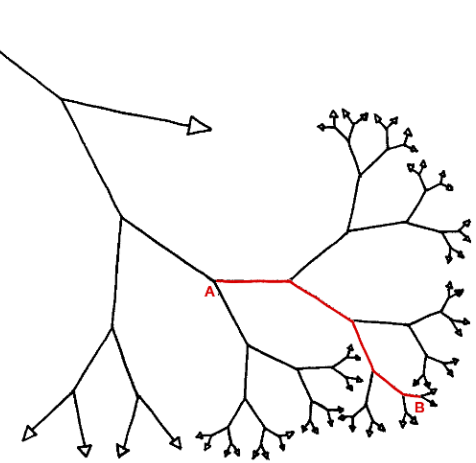
DOI: 10.1177/1745691616658637
pps.sagepub.com
SAGE

Increasing Transparency Through a Multiverse Analysis

Sara Steegen[1], Francis Tuerlinckx[1], Andrew Gelman[2], and Wolf Vanpaemel[1]
[1]KU Leuven, University of Leuven and [2]Columbia University

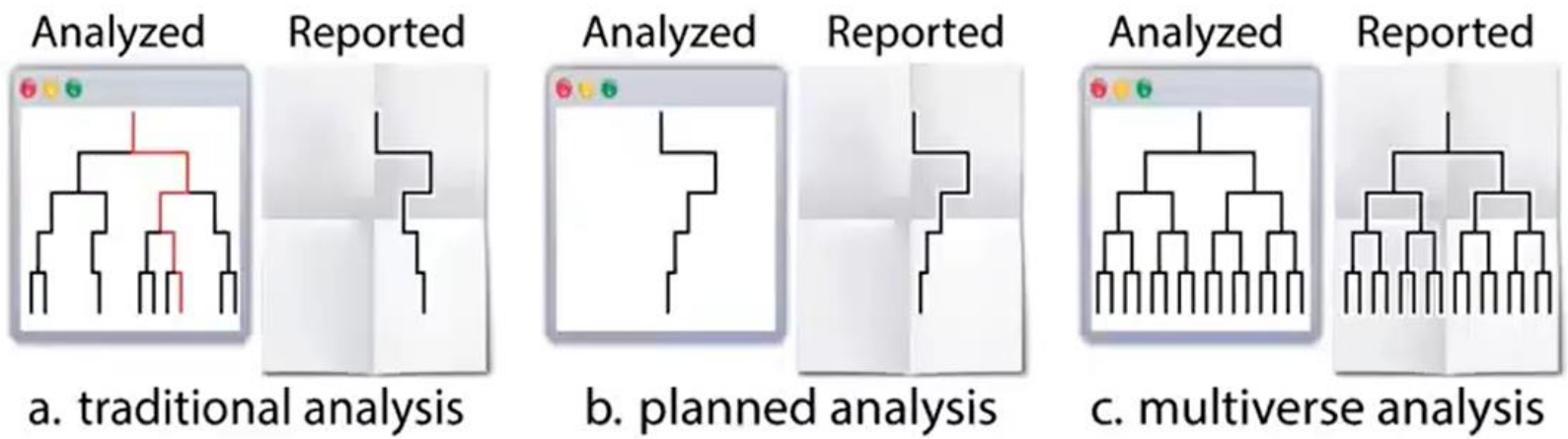
Abstract
Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

The Garden of Forking Paths
by Jorge Luis Borges



In the light of this problem of selective reporting, we propose to use a multiverse analysis as an alternative to a single data set analysis.

Multiverse Terms



nature human behaviour

RESOURCE

https://doi.org/10.1038/s41562-020-0912-z

Check for updates

Specification curve analysis

Uri Simonsohn[1]✉, Joseph P. Simmons[2] and Leif D. Nelson[3]

Webcast Lecture Highlight - Multiverse Analysis with Dr. Aaron Hill
https://www.youtube.com/watch?v=Q9bbCi6bV8Q
Multiverse analysis
https://www.youtube.com/watch?v=QfvjZL7jY24

Palacký University
Olomouc

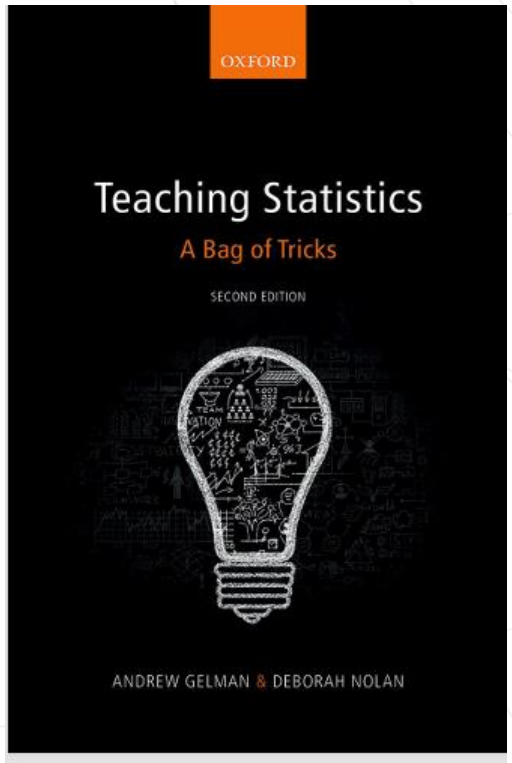
Děkuji za pozornost

- The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.
- Čísla nemohou mluvit sama za sebe. Mluvíme za ně. Dáváme jim význam.

Nate Silver, *The Signal and the Noise*

Palacký University
Olomouc

Prvotní hřích – málo času
Více důrazu a časové dotace na statistiku a analýzu dat



00111000110010000100	01000101001100010100
00100010001000000001	11101001100011110100
00110010101100001111	01110100011000110111
11001100010101100100	10001001011011011100
10001000000011111001	01100100010010000100

Fig. 8.1 Two binary sequences produced by students in an eighth grade class for the demonstration of Section 8.3.2. Can you figure out which is the actual sequence of 100 coin flips and which is the fake? The answer appears on page 366.

Palacký University
Olomouc

iPod Shuffle Problems: How Random is the iPod Shuffle?

Humans are not good at identifying randomness: our minds naturally look for patterns, even when there are none.

Furthermore, we are poor at creating random data. Famously, as a result of listener complaints, the first iPod 'shuffle function' had to be changed to make it less random, but appear more random to the human ear

https://electronics.howstuffworks.com/ipod.htm

Pandit, J. J. (2012). On statistical methods to test if sampling in trials is genuinely random. *Anaesthesia*, *67*(5), 456-462.

Palacký University
Olomouc

SCIENCE

Meet Yoshitaka Fujii, the most prolific fraudster in modern science

By Joseph Stromberg | May 21, 2015, 1:10pm EDT

- Those wishing to invent data have a hard task. They must ensure that all the data satisfy several layers of statistical cross-examination
  ….It is therefore always so much easier actually to do the experiment than to invent its results.

The new retraction record holder is a German anesthesiologist, with 184

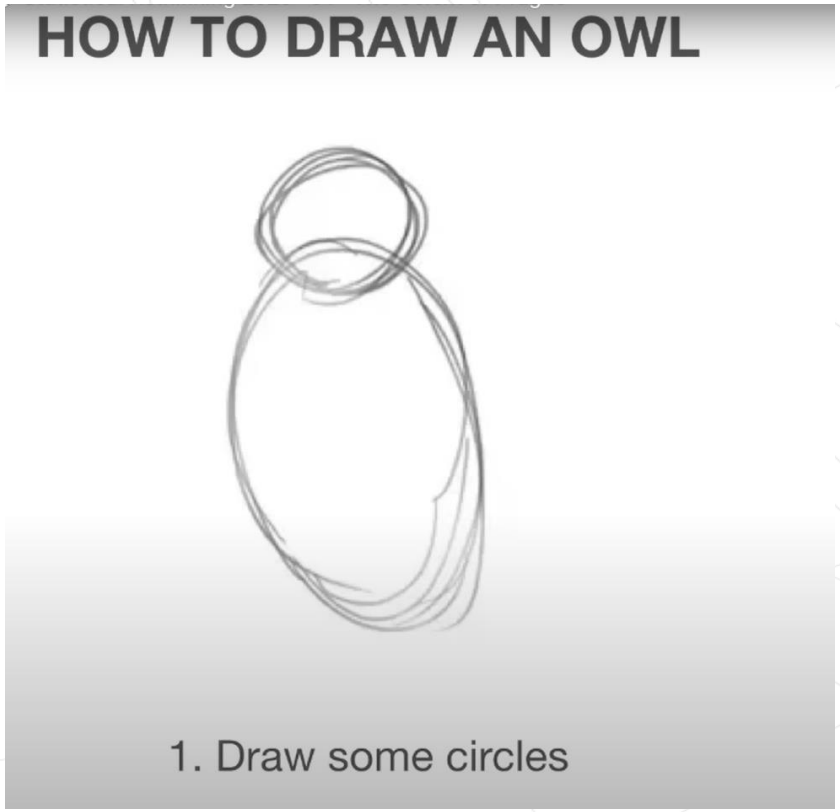
The German anesthesiologist Joachim Boldt has lost 20 more papers since January 2023, earning him the top spot in our leaderboard, with 184 retractions.

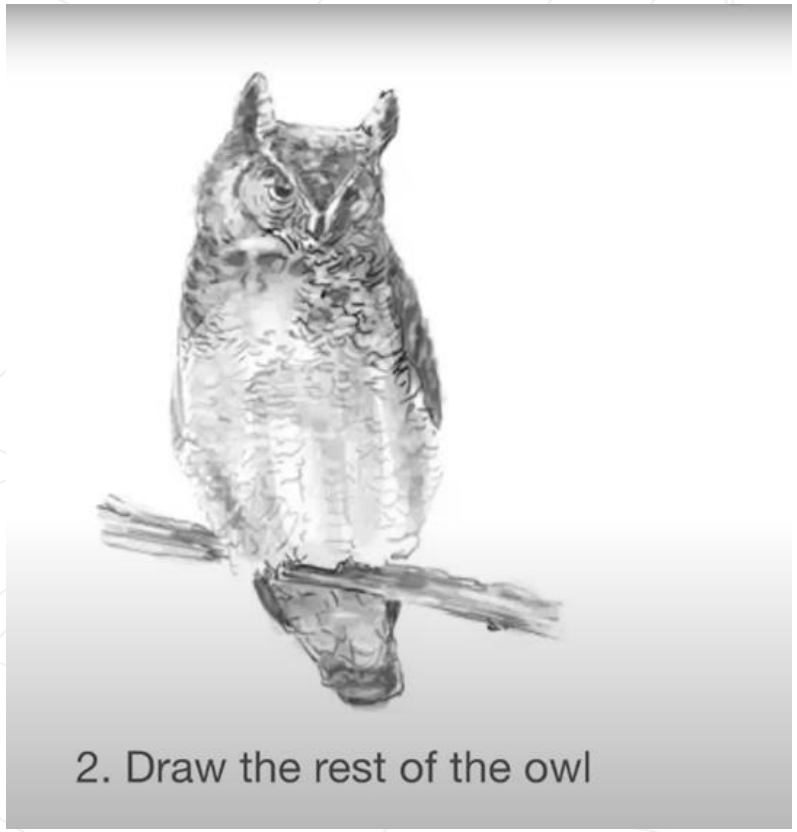
Boldt, readers may recall, was once one of the leading international figures in perioperative medicine. His work, particularly studies involving

Ludwigshafen Hospital, via Wikimedia

https://www.vox.com/2015/5/21/8636569/retraction-yoshitaka-fujii
https://retractionwatch.com/2023/07/12/the-new-retraction-record-holder-is-a-german-anesthesiologist-with-184/#more-127491
Pandit, J. J. (2012). On statistical methods to test if sampling in trials is genuinely random. *Anaesthesia*, *67*(5), 456-462.
https://nautil.us/how-the-biggest-fabricator-in-science-got-caught-235421/

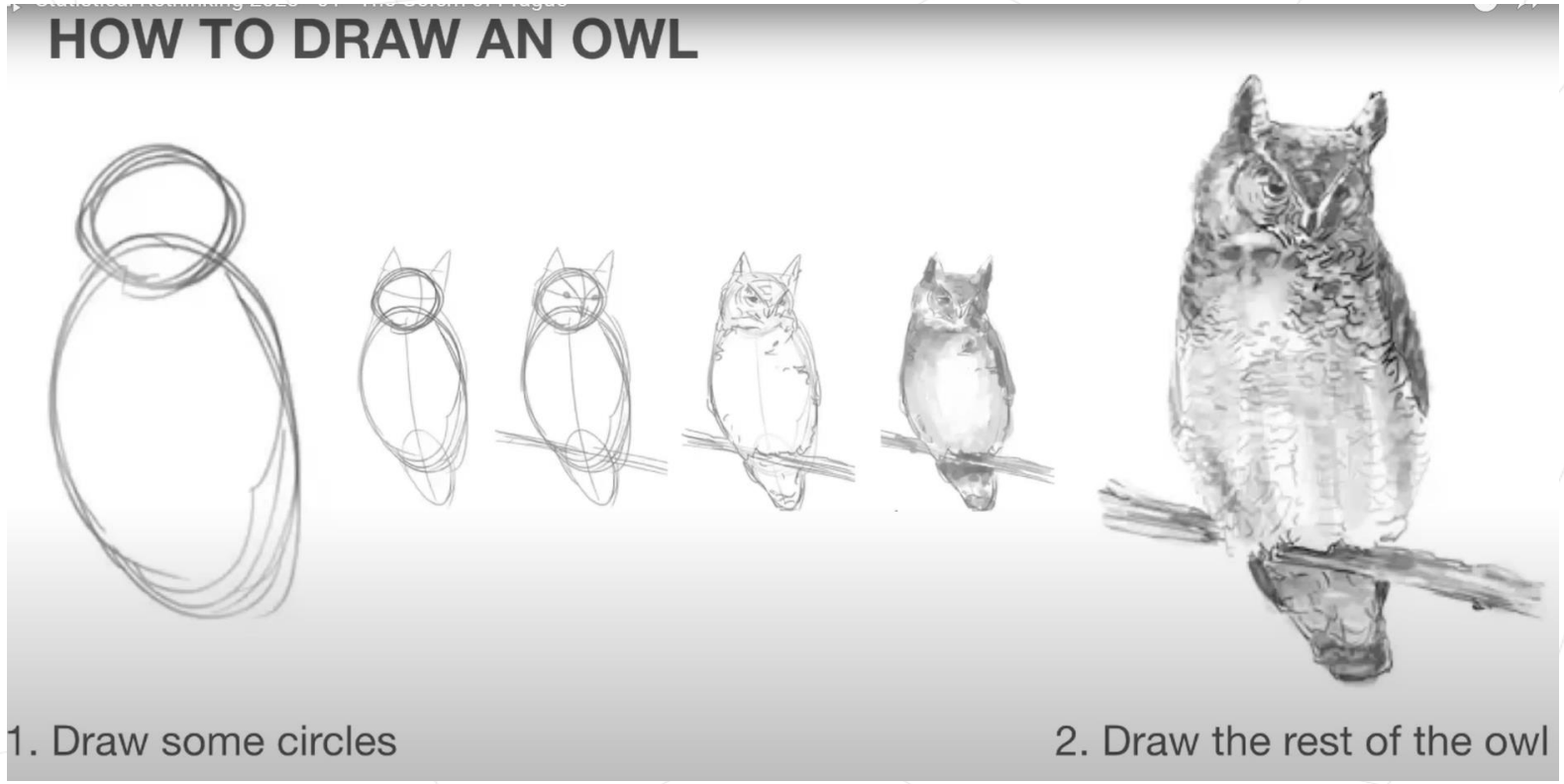
Vtípek na závěr
Jak učíme statistiku?





Statistical Rethinking 2023 - 01 - The Golem of Prague
https://www.youtube.com/watch?v=FdnMWdICdRs&list=PLDcUM9US4XdPz-KxHM4XHt7uUVGWWVSus

Vtípek na závěr
Jak učíme statistiku?



Statistical Rethinking 2023 - 01 - The Golem of Prague
https://www.youtube.com/watch?v=FdnMWdICdRs&list=PLDcUM9US4XdPz-KxHM4XHt7uUVGWWVSus

Palacký University
Olomouc

Prostor na vaše dotazy