Univerzita Palackého
v Olomouci

Analýza dat: co je dobré vědět, co dělat a co nedělat

Ondřej Vencálek
PřF UP Olomouc

12. 12. 2023
Replikační krize ve vědě: její příčiny a důsledky

Univerzita Palackého
v Olomouci

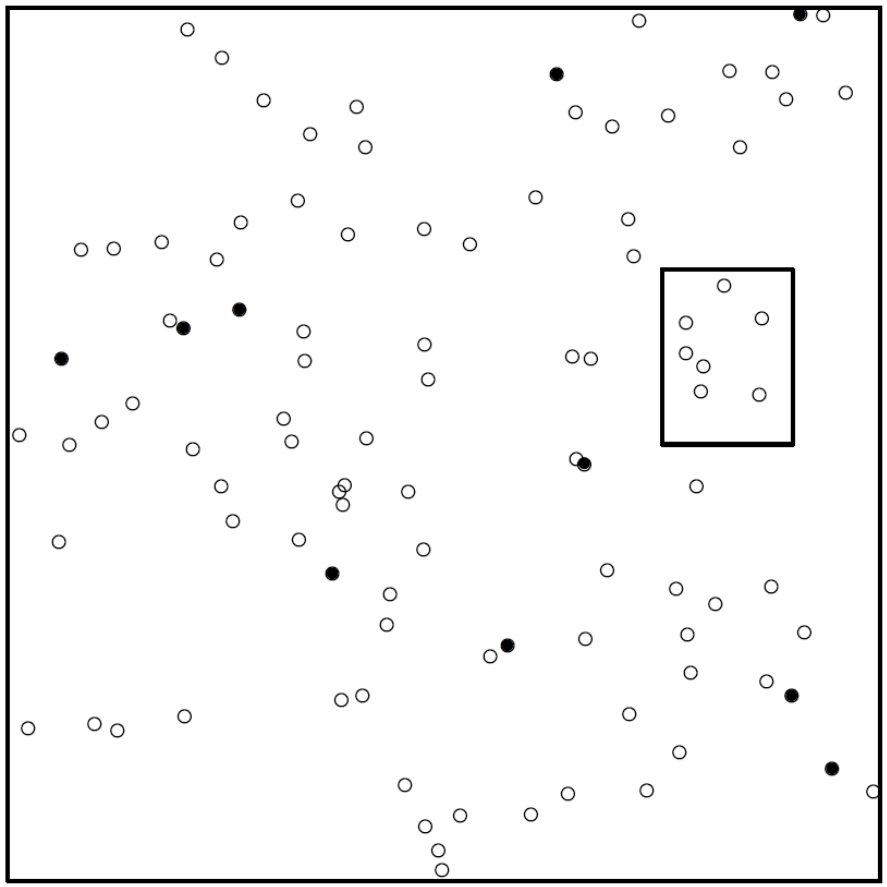
Nevstoupíš dvakrát do téže řeky

Hérakleitos

… tak jaképak opakování!?

… jakápak replikace!?

Problém indukce



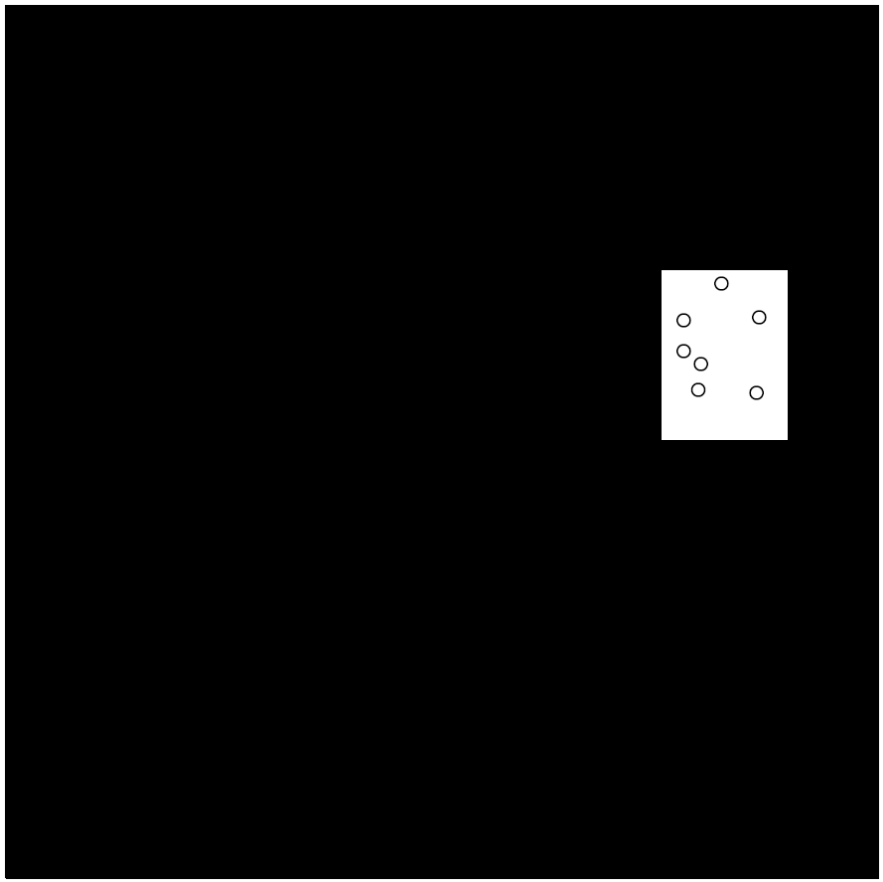
David Hume 18. stol.
(Pojednání o lidské přirozenosti):

Ani ze sebevětšího počtu pozorování bílých labutí nelze usuzovat, že jsou bílé opravdu všechny, k odmítnutí takového závěru postačuje jediná černá.

Francis Bacon 16.-17.stol. :

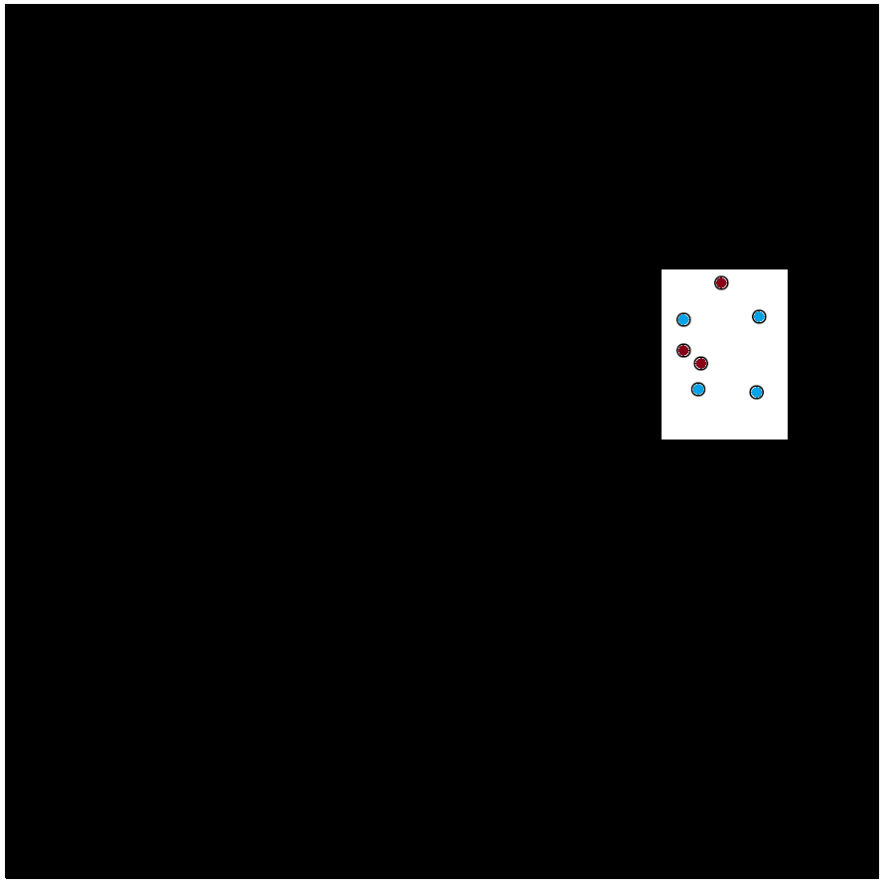
Jakmile člověk zaujme nějaké stanovisko, sbírá všechny důkazy, které ho potvrzují, a i když důkazy svědčící o opaku mohou být častější a závažnější, buď si jich nevšimne nebo je zavrhne, aby víra v zaujaté stanovisko nebyla otřesena.

Problém indukce



- Uvědomujme si existenci NEJISTOTY v našich závěrech.
- Umíme se s nejistotou vypořádat, neumíme ji odstranit.

Problém indukce



- Světle modré tečky: 4
  Tmavě červené tečky: 3
- Otázka:
  Můžeme říct, že (v celku) je světle modrých teček více než tmavě červených?
- S jistotou nikoliv
  (dokud nepozorujeme celek)
- Jak číst výsledek
  „statisticky signifikantní“
  a jak „nesignifikantní“.

Univerzita Palackého
v Olomouci

Problém indukce

Hand, D. J. Temná data: proč záleží na tom, co nevíme. Praha: Academia, 2023

Kap.7: Věda a temná data: povaha objevování

- Základním procesem vědy je testování teorií na základě pozorovaných dat, přičemž nesoulad mezi teorií a těmito daty vede k odmítnutí nebo ke změně teorie.
- Kritérium testovatelnosti je to, co odlišuje vědu od pseudovědy.

Charles Babbage 1830

Reflections on the Decline of Science in England: And on Some of Its Causes

Vědecká zkoumání jsou více než jakákoli jiná vystavena nájezdům podfukářů; mám pocit, že si zasloužím poděkování všech, kdo si skutečně váží pravdy, když uvedu některé metody klamání, které používají nehodní uchazeči o její ocenění, zatímco pouhá okolnost, že jejich umění je známo, může odradit budoucí pachatele…

Ve vědě se praktikuje několik druhů podvádění, které jsou mimo okruhy zasvěcených málo známé a které by snad bylo možné zpřístupnit běžnému chápání. Lze je zařadit pod hlavičky

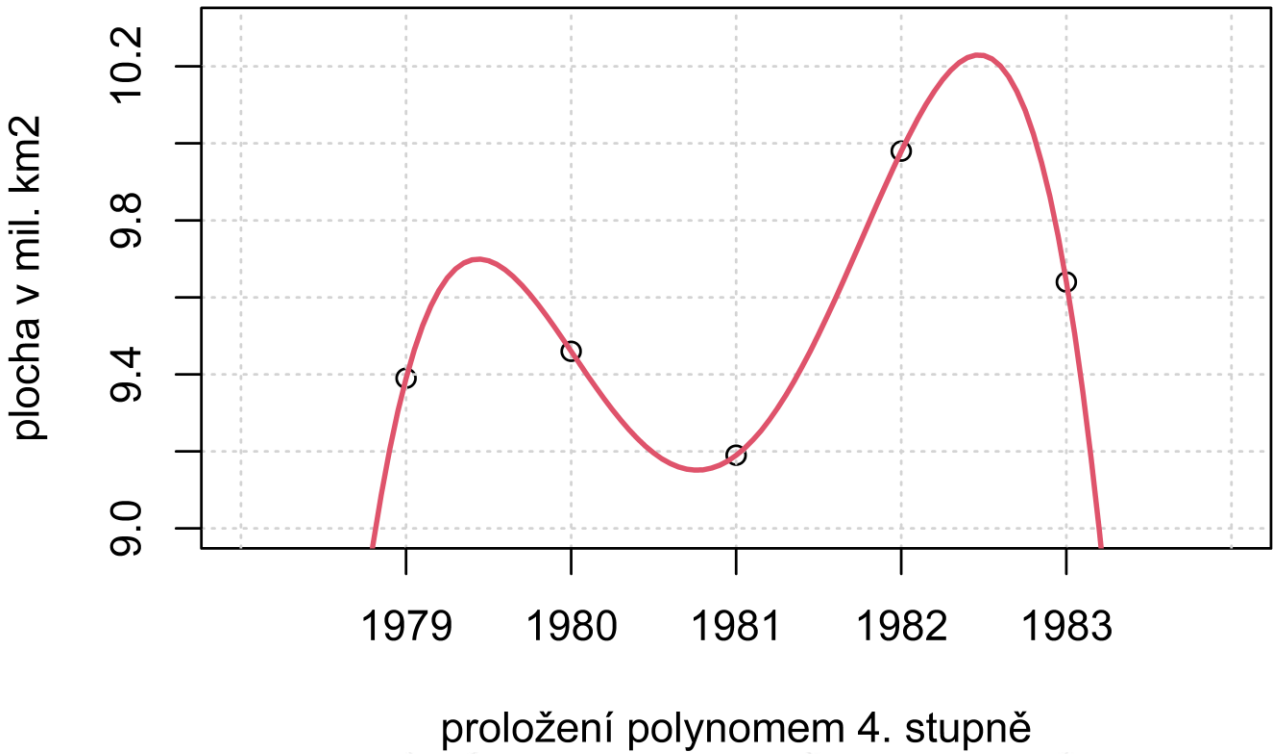
- Mystifikace
- Padělání
- Ořezávání
- Vaření

Univerzita Palackého
v Olomouci

Co dělat a co nedělat

- Explorační analýza (EDA) včetně vizualizace – první krok; podceňovaná; nepřemýšlíme o možnostech zobecnění (zdroj hypotéz)
- Testování hypotéz (spojeno s p-hodnotou) je *nesprávně* používáno jako nástroj *explorační analýzy*. Problém mnohonásobného testování (p-hacking)
- Trénovací versus testovací datová množina
  (různé typy cross-validací);
  pochopení problému overfitting;
  HARKing: Hypothesizing After the Result is Known

Overfitting



Univerzita Palackého
v Olomouci

p-hacking

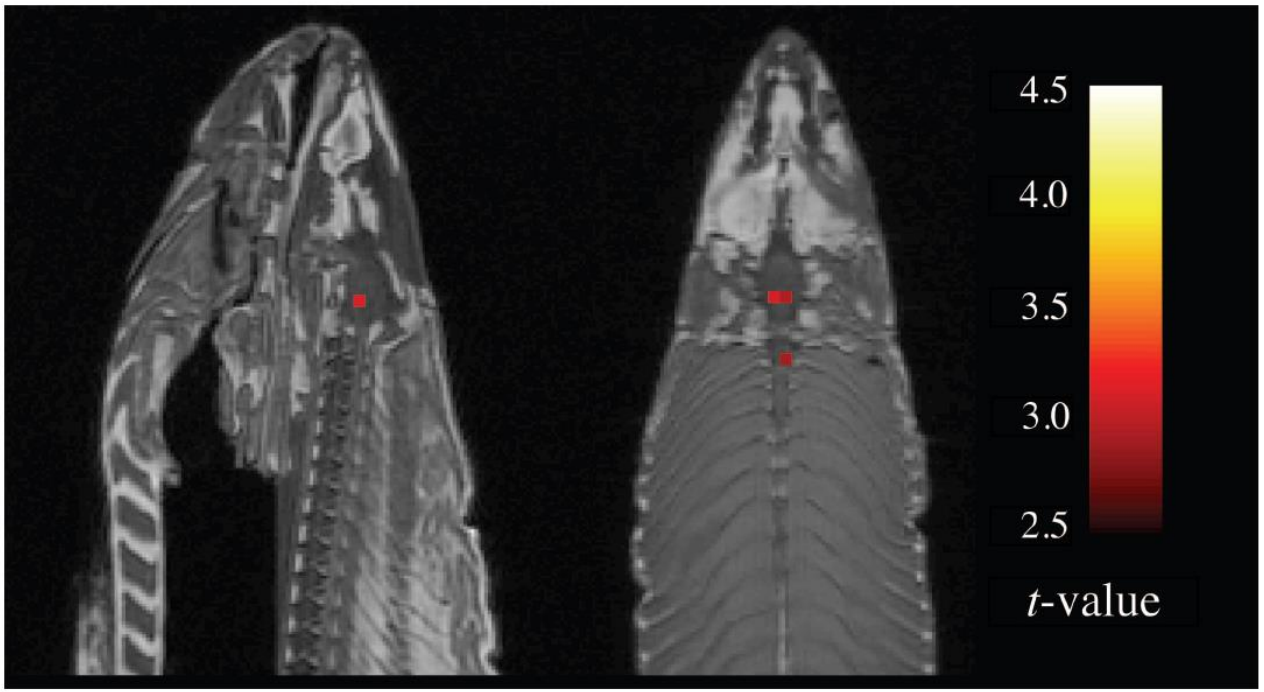
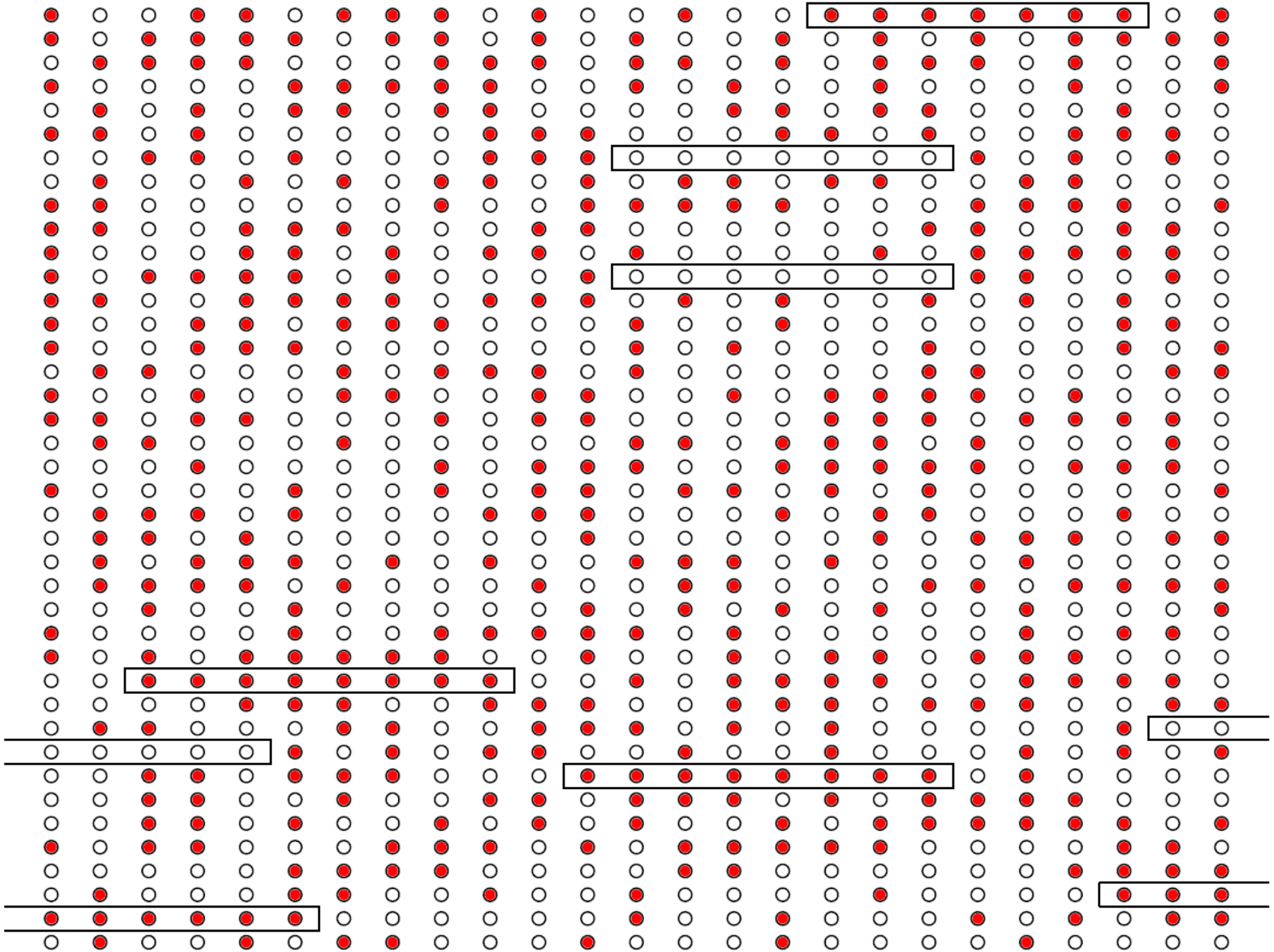


Fig. 1. Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.

Bennett, C.M., Baird, A.A., Miller, M.B., & Wolford, G.L. (2010). of Serendipitous and Unexpected Results Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon : An Argument For Proper Multiple Comparisons Correction.



Univerzita Palackého
v Olomouci

Něco ke čtení

Hand, D. J. *Temná data: proč záleží na tom, co nevíme*. Praha: Academia, 2023

https://www.statspol.cz/knihovnicka/knihovnicka-Temna_Data/

Taleb, N. N. *Zrádná nahodilost: o skryté roli náhody na trzích a v životě*. Praha: Paseka, 2013. ISBN 978-80-7432-292-1.