Podatkovno rudarjenje in vizualizacija

prof. dr. Marko Robnik-Šikonja

Univerzav Ljubljani, Fakulteta za računalništvo in informatiko

Predavatelj

- prof. dr. Marko Robnik Šikonja
- marko.robnik@fri.uni-lj.si
- UL FRI, Večna pot 113, 2. nadstropje, desno od dvigala
- (01) 4798 241

- govorilne ure:
  - torek od 11:15 -12:15 ali po dogovoru
- http://www.fri.uni-lj.si/rmarko
- raziskovalno delo: podatkovna analitika, podatkovno rudarjenje, strojno učenje, umetna inteligenca, procesiranje naravnega jezika, algoritmi in podatkovne strukture

Cilji predmeta

- Ponuditi sodobno, pregledno in praktično uporabno znanje s področja podatkovnega rudarjenja.
- Pregledno predstaviti poglavitne tehnike podatkovnega rudarjenja.
- Poudarek je na pridobitvi praktičnega znanja iz podatkovne analitike in rabe orodij.
- Zmožnost osnovne analize podatkov, njihove vizualizacijo, interpretacije in pridobivanja znanja iz podatkov.

Način dela

- Predavanja:
  - predstavitev glavnih pristopov, pojasnimo delovanje metod, brez podrobnosti izvedbe in teoretičnih osnov
  - podrobnejša obravnava nekaj pomembnih vrst podatkov za ne tehnične smeri: ankete, grafi, besedila
  - vizualizacijo kompleksnih podatkov, trendov in napovednih modelov
  - predvideno je sodelovanje, diskusija, analiza vaših podatkov, sprotno preverjanje na prenosnikih
- Laboratorijske vaje:
  - prenos znanja v prakso
  - uporabo odprtokodnih sistemov za podatkovno rudarjenje in vizualizacijo,
  - reševanje nalog, ki temeljijo na realnih problemih,
  - problematika s področja študentovega primarnega študija ali zanimanja
  - Domače naloge
    - predstavimo naloge, pomagamo z namigi, smo na voljo za vprašanja, zato …
    - …je potrebno poznati snov in vprašanja postaviti.

Vsebina predmeta 1/3

1. **Uvod v podatkovno rudarjenje**. Predstavimo motivacijske probleme in njihove rešitve z metodami podatkovnega rudarjenja. Poljudno predstavimo osnovne pojma učenja iz podatkov, modeliranje podatkov in pomembne teoretične rezultate glede (ne)zmožnosti učenja iz podatkov.
2. **Zbiranje in priprava in podatkov**. Obravnavamo prevedbo problemov v obliko, ki je primerna za uspešno podatkovno rudarjenje.
3. **Mere podobnosti in razvrščanje v skupine.** Podatke želimo analizirati glede medsebojne podobnosti posameznih primerov in jih razvrstiti v skupine. Predstavimo poglavitne tehnike in izzive.
4. **Raziskovalna analiza podatkov.** Predstavimo vrsto vizualizacijskih tehnik, ki nam omogočajo, da na razumljiv način spoznavamo problem in raziskujemo zakonitosti v (visoko razsežnih) podatkih.
5. **Ovrednotenje in izbira pomembnih atributov.** Številne probleme imamo podane v tabelarični obliki, kjer vrstica vsebuje en primer opisan z množico atributov. Za uspešno modeliranje je potrebno prepoznati pomembne atribute in izbrati njihovo neredundantno podmnožico. Opišemo poglavitne metode ocenjevanja atributov.

Vsebina predmeta 2/3

6. **Napovedni modeli.** Spoznamo napovedne modele s področja statistike in strojnega učenja ter pogoje, da le-ti v praksi dobro delujejo.
7. **Vizualizacija napovednih modelov**. Številni odlični napovedni modeli, delujejo za uporabnika kot črna škatla, saj niso razvidni mehanizmi njihovega delovanja in odločanja. Za področja, kjer je modeliranje namenjeno tudi razumevanju problema in pridobivanju novega znanja, je to nesprejemljivo. Predstavimo rešitev v obliki tehnik razlage, ki grafično predstavijo delovanje napovednih modelov in obrazložijo njihove odločitve.
8. **Povezovalna pravila in pogosti vzorci.** Včasih v podatkih iščemo značilne povezave in vzorce, ki predstavljajo zanimive, statistično pomembne zakonitosti. Predstavimo uveljavljene metode za to analizo.
9. **Grafični modeli.** Pri razumevanju nekaterih procesov in problemov si pomagamo z njihovo predstavitvijo v obliki bayesovskih grafičnih modelov, ki nam v primerih negotovosti omogočajo verjetnostno sklepanje od vzrokov k posledicam.
10. **Analiza anket.** Obravnavamo nekaj metod strojnega učenja, ki so prilagojene značilnostim anketnih podatkov. Omogočajo npr. vrednotenje vprašanj pri anketah, zaznavanje šumnih in nekonsistentnih odgovorov, iskanje povezanih vprašanj, itd.

Vsebina predmeta 3/3

11. **Rudarjenje besedil.** Besedila so pomemben vir podatkov, iz katerih lahko razberemo številne informacije in sociološke značilnosti. Pregledno obravnavamo procesiranje slovenskih in angleških besedil ter osnovne tehnike rudarjenja besedil.
12. **Odkrivanje znanja iz socialnih omrežij**. Socialna omrežja v svoji strukturi vsebujejo številne pomembne informacije. Pregledamo nekaj pristopov, tehnik in orodij za analizo omrežij.
13. **Analiza velikih podatkovnih množic**. Ogromne podatkovne množice, ki so nastale na posameznih področjih človekovega delovanja, vsebujejo v sebi številne zanimive drobce informacij, jih je pa zaradi tehničnih omejitev težko analizirati in iz njih pridobiti koristno znanje. Predstavimo nekaj uveljavljenih načinov dela s takšnimi množicami.
14. **Spoznanja iz uporabe podatkovnega rudarjenja in etični vidiki analize podatkov**. Predstavimo nekaj uspešnih in nekaj neuspešnih aplikacij podatkovnega rudarjenja in povzamemo njihove lekcije. Na primerih obravnavamo etični vidik podatkovne analitike in predstavljanja rezultatov.

Obveznosti

- 5 spletnih kvizov, ki sproti preverjajo razumevanje snovi, skupaj je potrebno doseči 50% točk, štejejo le pravočočasno oddani kvizi
- 5 domačih nalog različnih težavnosti, rezultate zapišete v pisno poročilo, eno domačo nalogo javno predstavite pred kolegi
- pisni izpit
- možnost ustnega izpita za izboljšanje ocene

Gradiva in orodja

- gradiva bodo dostopna na spletni učilnici http://ucilnica.fri.uni-lj.si Ste se uspeli vpisati?
- praktično delo v prosto dostopnih programih
  - R, uporaba okolja RStudio, lahko tudi Eclipse
  - Orange
  - lahko tudi v drugih, npr. Scikit-learn (Python), Weka & Rapid Miner.

Literatura

- Ian H. Witten, Eibe Frank, and Mark A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition,* Morgan Kaufmann.
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani: *An Introduction to Statistical Learning with Applications in R. Springer, 2013 prosto dostopno na* http://www-bcf.usc.edu/~gareth/ISL/
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009): *The elements of statistical learning, 2nd edition*. Springer.
- Janert, P. K. (2010). *Data analysis with open source tools.* O'Reilly Media.
- Dodatno literaturo v obliki člankov in posnetkov predavanj boste dobili na spletni učilnici.

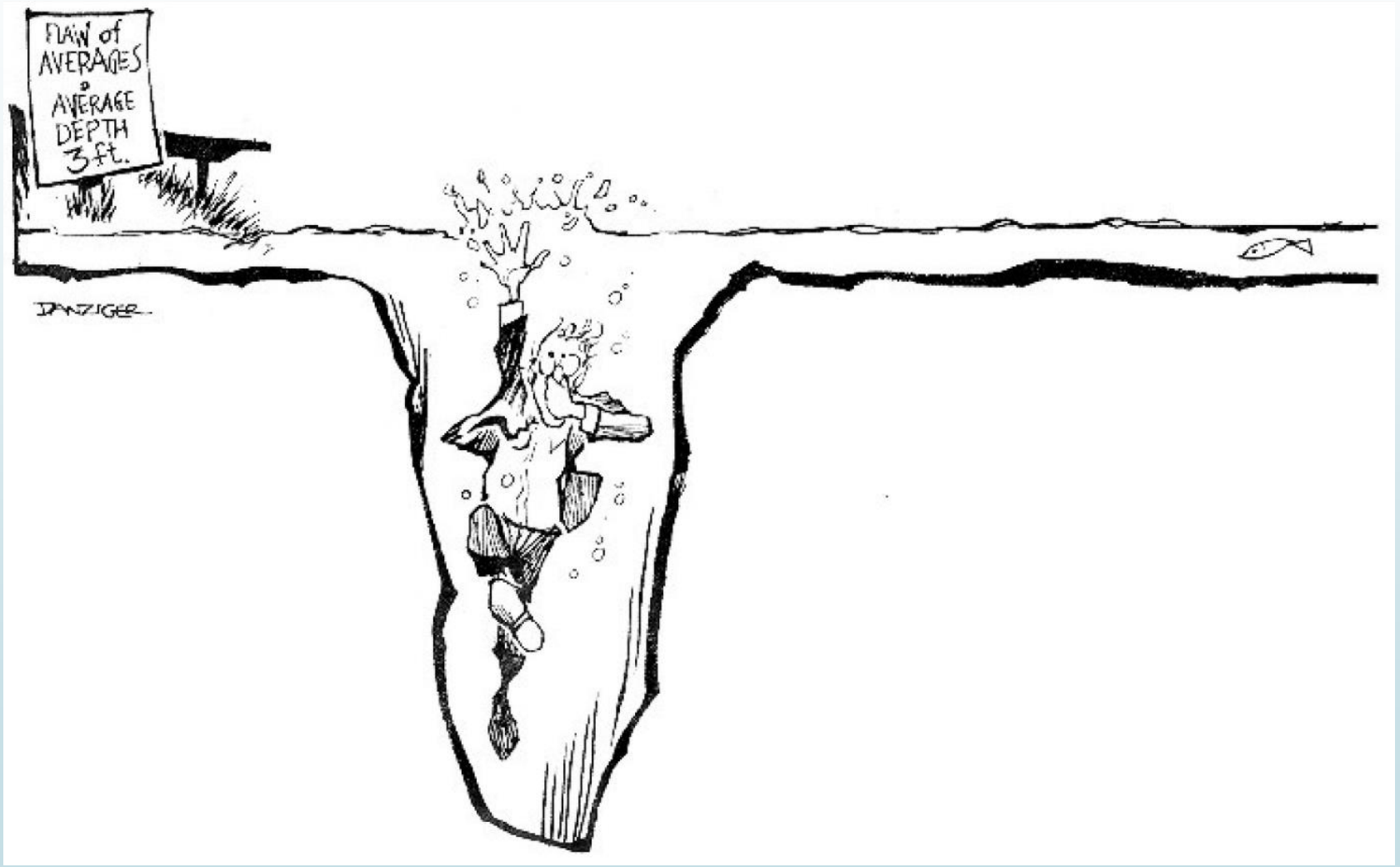
Uvod v podatkovno rudarjenje

- motivacijski problemi
- medicina: diagnostika in prognostika bolezni na podlagi kliničnih meritev, biomarkerjev, slik, genskega zapisa, …
- farmacija: interakcije zdravil, iskanje novih zdravil, napovedovanje kemičnih lastnosti
- politika: volilne napovedi, napovedovanje neodločnih volivcev
- marketing: izbira prejemnikov pošte, strank, ki bodo zamenjale dobavitelja
- zavarovalništvo: prevare
- bančništvo: tveganost odobritve kreditev, analiza strank
- inženering: napovedovanje napak kompleksnih sistemov
- jezikoslovje: napovedovanje besednih vrst, vloge besed v stavku, …
- spletno rudarjenje: iskanje, priporočila, personalizacija, (socialna) omrežja

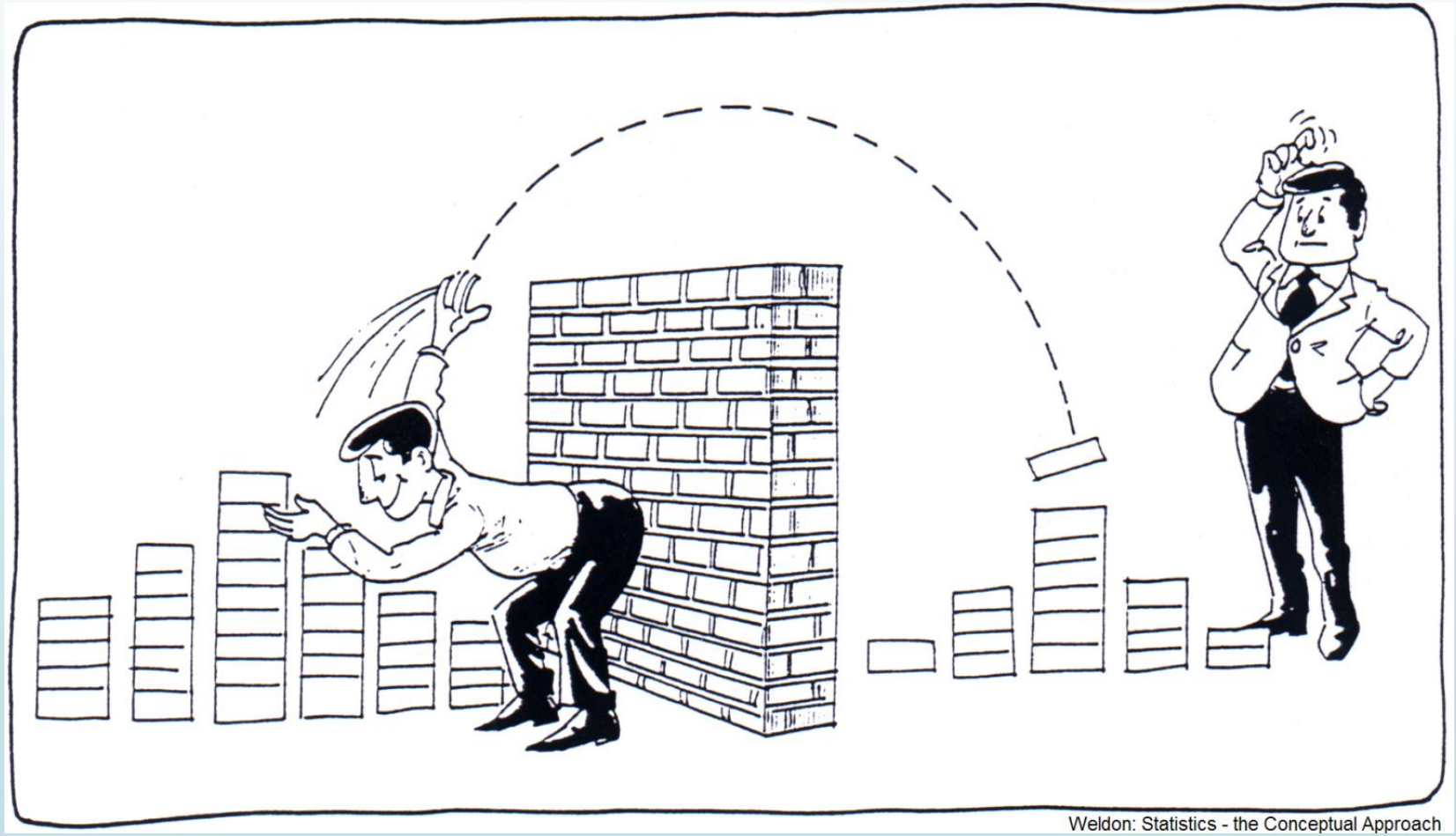
Osnovni pojmi

- statistika, strojno učenje, podatkovno rudarjenje, podatkovna analitika
- učenje iz primerov
- učenje kot iskanje, posplošitev in specializacija, (pretirano) prilagajanje
- induktivno in deduktivno učenje
- nadzorovano in nenadzorovano učenje
- klasifikacija in regresija
- podatkovna množica, vrste podatkov
- učni modeli; vrste, lastnosti
- ovrednotenje naučenega
- vizualizacija podatkov, modela in evalvacije
- katera učna metoda je najboljša?
- korelacije in vzroki, odločanje, podatki, informacije, znanje, modrost
- etika odločanja in zasebnost podatkov, reidentifikacija

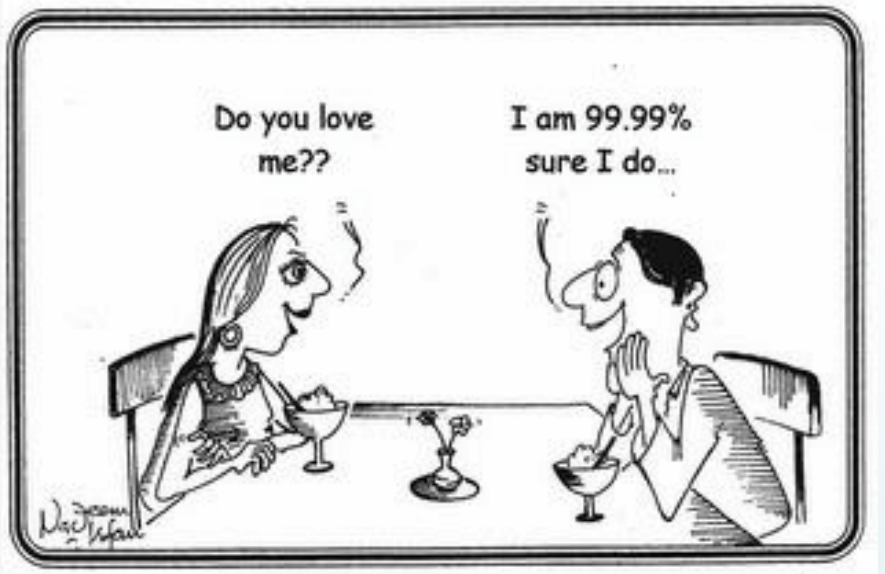
Statistika

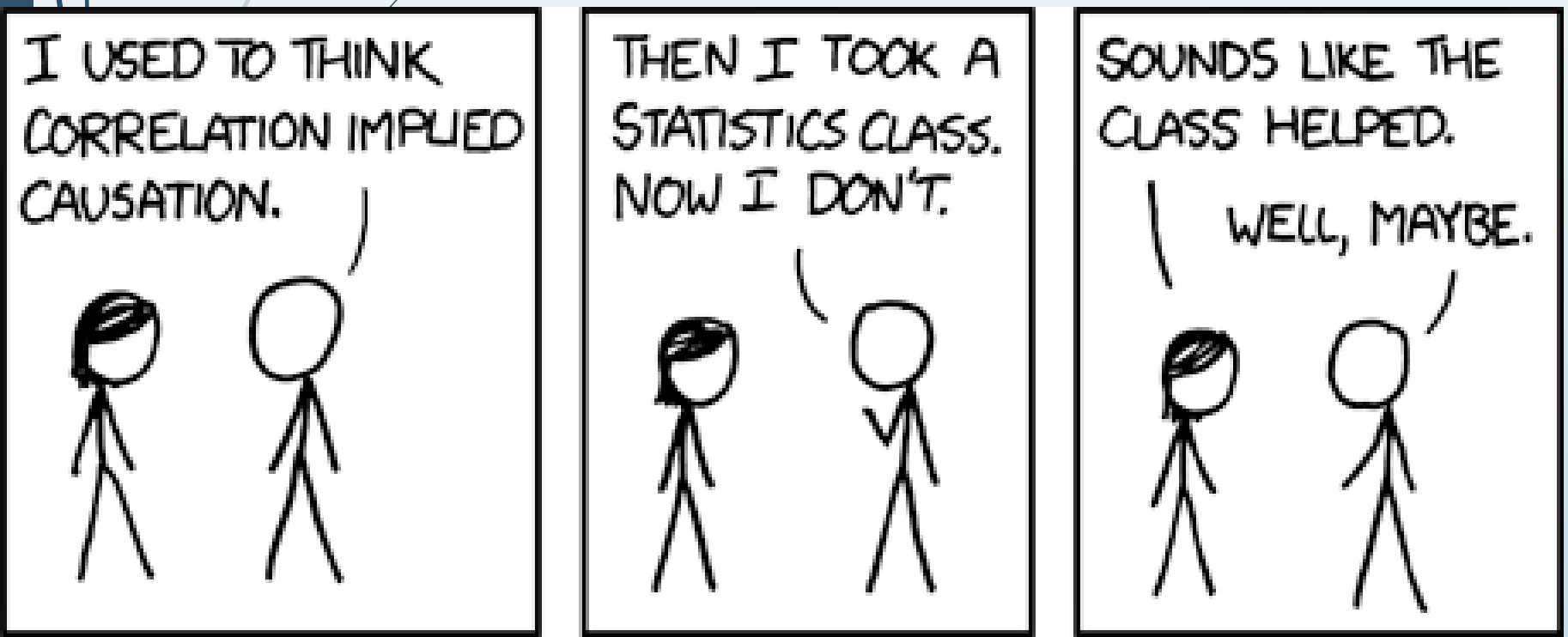


Statistika

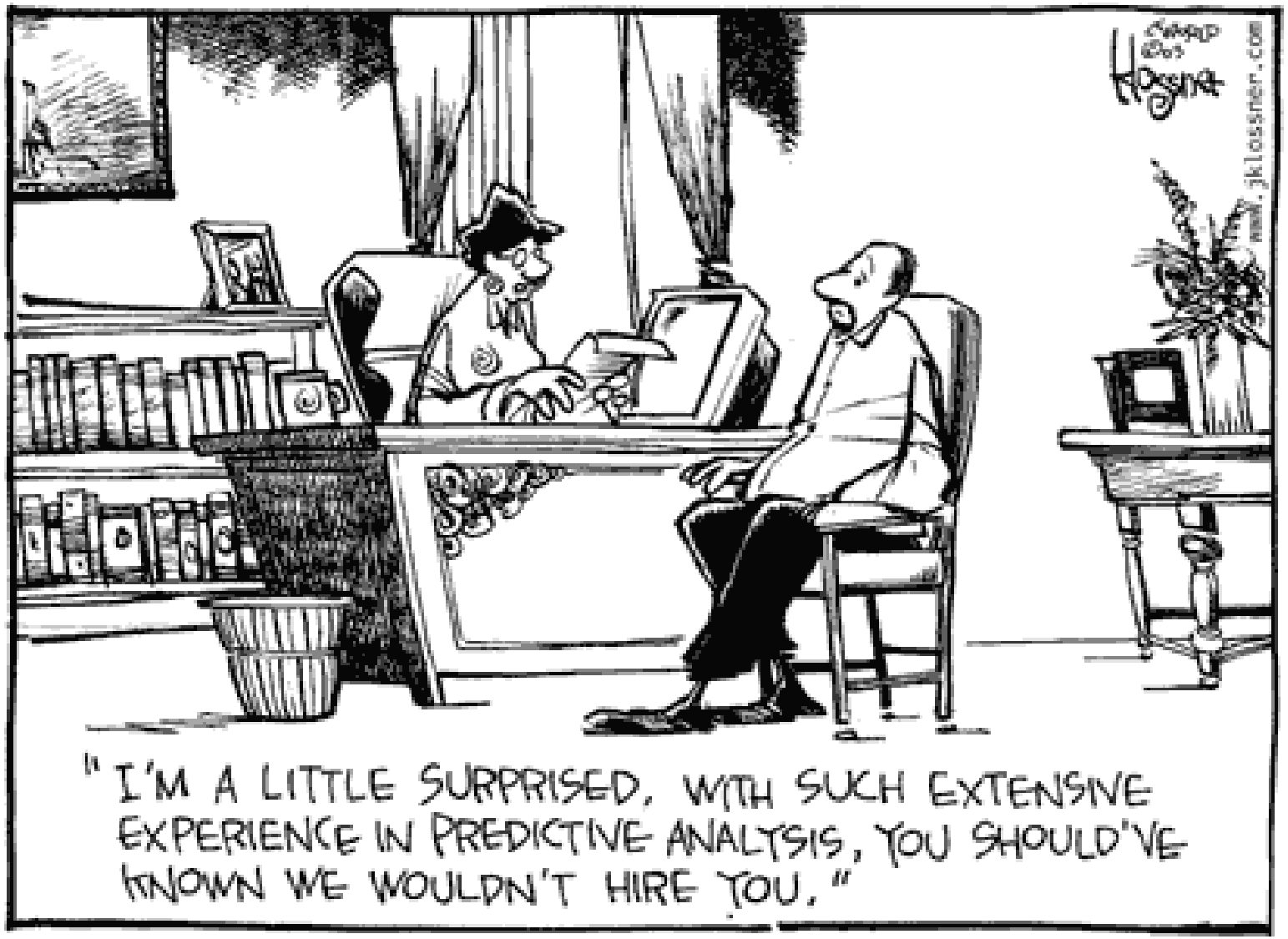


Statistika


Do you love me??
I am 99.99% sure I do...

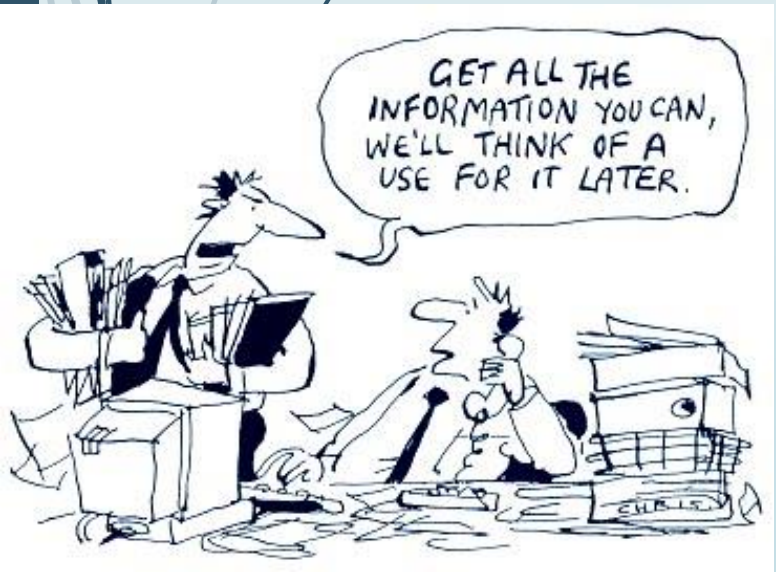

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Strojno učenje, podatkovno rudarjenje



Zbiranje in priprava in podatkov

- statistika: enostavni slučajni vzorec
- realnost: podatki, ki jih lahko dobimo; tisti, ki so poceni,
- podatki so šumni, nepopoilni, pristrani,
- pristranost in variance
- garbage in, garbage out
- priprava podatkov v povprečju 50% časa modeliranja



Vrste podatkov

Tabelarični podatki

- tabelarični podatki
- atributi in primeri, oznake
- oznake, funkcije, več oznak, hierarhične oznake
- definicija smiselnih atributov
- cena atributov
- redki podatki
- vizualizacije tabelaričnih podatkov: enodimenzionalna, večdimenzionalna
- zanka podatkonega rudarjenja