1-DAV-302 Princípy dátovej vedy

Vyučujúci:

Tomáš Vinař (prednášky), M-163, tomas.vinar@fmph.uniba.sk
Vladimír Boža (cvičenia), M-25, vladimir.boza@fmph.uniba.sk
Andrej Špitalský (asistent), spitalsky3@uniba.sk

Web: http://compbio.fmph.uniba.sk/vyuka/pridav

Vektor classroom: … info neskôr na stránke …

Literatúra:

Skiena: The Data Science Design Manual, Springer, 2017
(v knižnici: M-PRA-S-9)

Grus: Data Science from Scratch, O'Reilly Media, 2019

Oznamy a diskusia na Vektor Classroom

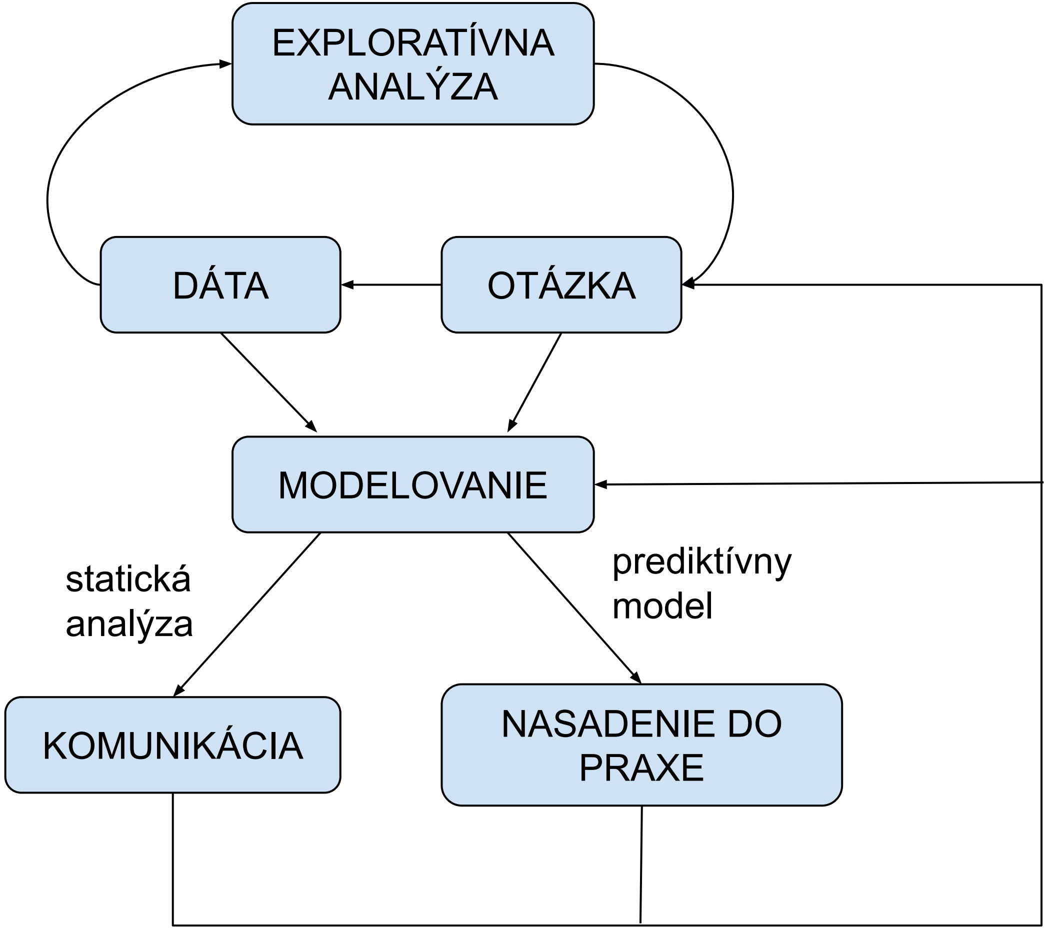
- **Vašou povinnosťou sledovať oznamy!!!**
- Cieľ: odpoveď na vaše otázky v čo najkratšom čase
- Odpovedajú učitelia, študenti
- Všetky otázky sú verejné => žiadne detaily vlastných riešení úloh (alebo až po termíne odovzdania úlohy)
- Link na web stránke predmetu

Hodnotenie predmetu

- 30% domáce úlohy a cvičenia (3 úlohy x 8 bodov, 5 cvičení x 5 bodov)
- 20% skupinový projekt (detaily cca v strede semestra)
- 50% skúška (písomná, ústna)
- zo skúšky je potrebné získať aspoň 50% bodov
- 90+ A, 80+ B, 70+ C, 60+ D, 50+ E

Opisovanie:

- **neopisujte!**
- -100% príslušnej časti bodového hodnotenia, disciplinárna komisia
- diskusia ohľadom domácich úloh je ok, ALE:
  - nerobte si poznámky
  - počkajte niekoľko hodín, kým začnete realizovať vlastné riešenie
  - napíšte, s kým ste konzultovali
  - ChatGPT rovnakým spôsobom ako spolužiaci


EXPLORATÍVNA ANALÝZA

DÁTA

OTÁZKA

MODELOVANIE

statická analýza

prediktívny model

KOMUNIKÁCIA

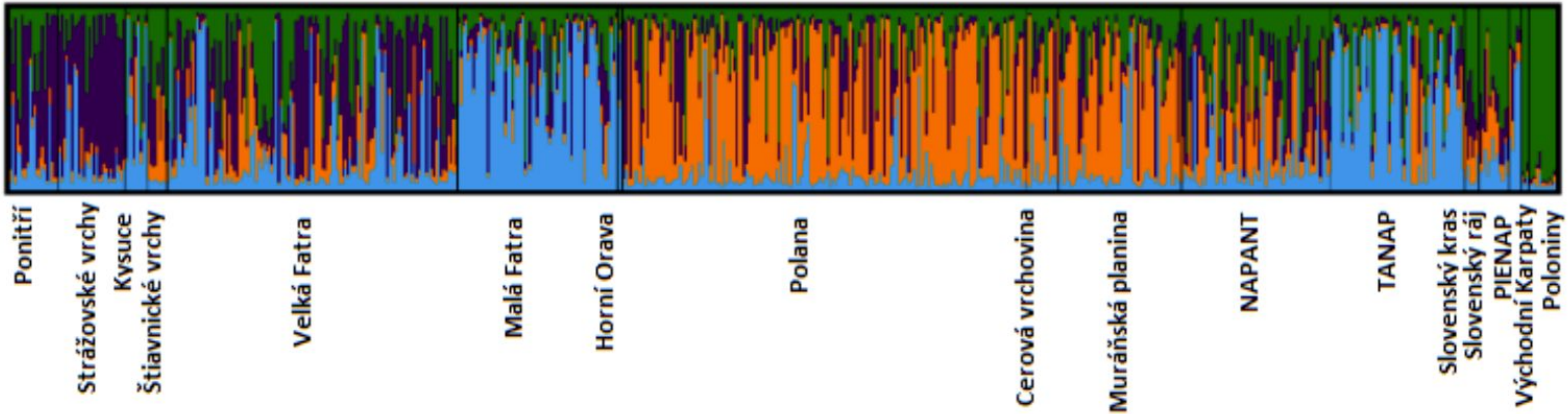
NASADENIE DO PRAXE

Prístupy ku riešeniu problémov

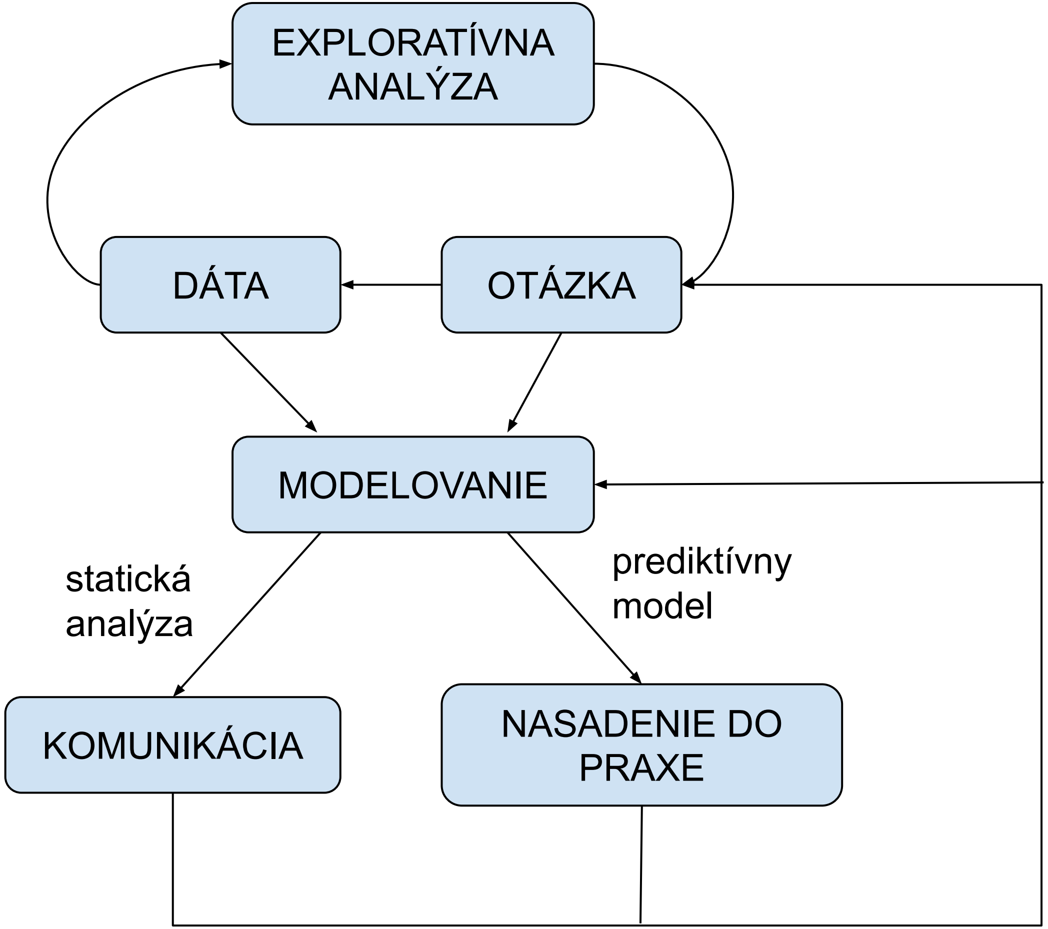
Prístupy riadené hypotézou / otázkou (obvyklé napr. v biológii)

- Na začiatku dobre formulovaná otázka (typicky odpoveď áno/nie)
- Aké dáta a spôsoby analýzy nám umožnia otázku zodpovedať?

Koľko medveďov žije na Slovensku?



Obr. 15: Grafický výstup Bayesiánské klastrovací analýzy v softwaru STRUCTURE (Pritchard et al., 2000) pro K = 4 zpracovaný webovým softwarem CLUMPAK (Kopelman et al., 2015). Každý sloupec reprezentuje jeden ze zpracovaných genotypů jedinců medvěda hnědého. Jedinci byli pro účel analýzy rozděleny do 17 subpopulací na základě lokality sběru vzorků. Tyto lokality byly seřazeny od západu na východ.


EXPLORATÍVNA ANALÝZA
DÁTA
OTÁZKA
MODELOVANIE
statická analýza
prediktívny model
KOMUNIKÁCIA
NASADENIE DO PRAXE

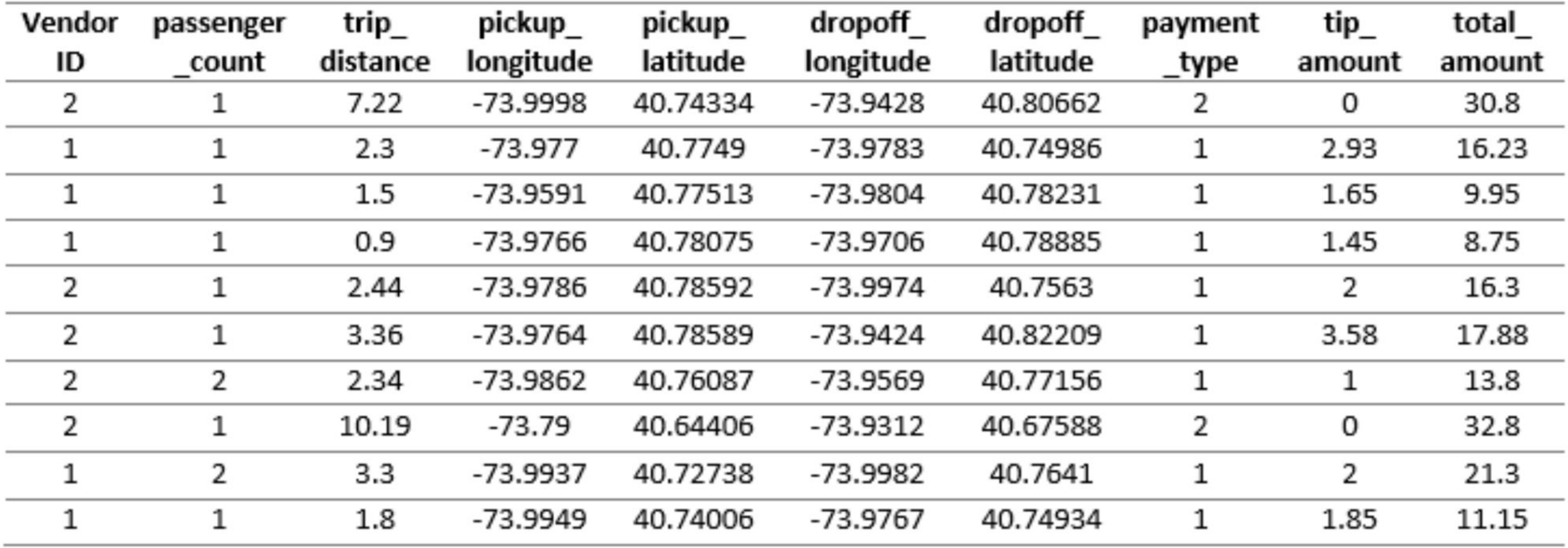
Prístupy ku riešeniu problémov

Prístupy riadené hypotézou / otázkou (obvyklé napr. v biológii)

- Na začiatku dobre formulovaná otázka (typicky odpoveď áno/nie)
- Aké dáta a spôsoby analýzy nám umožnia otázku zodpovedať?

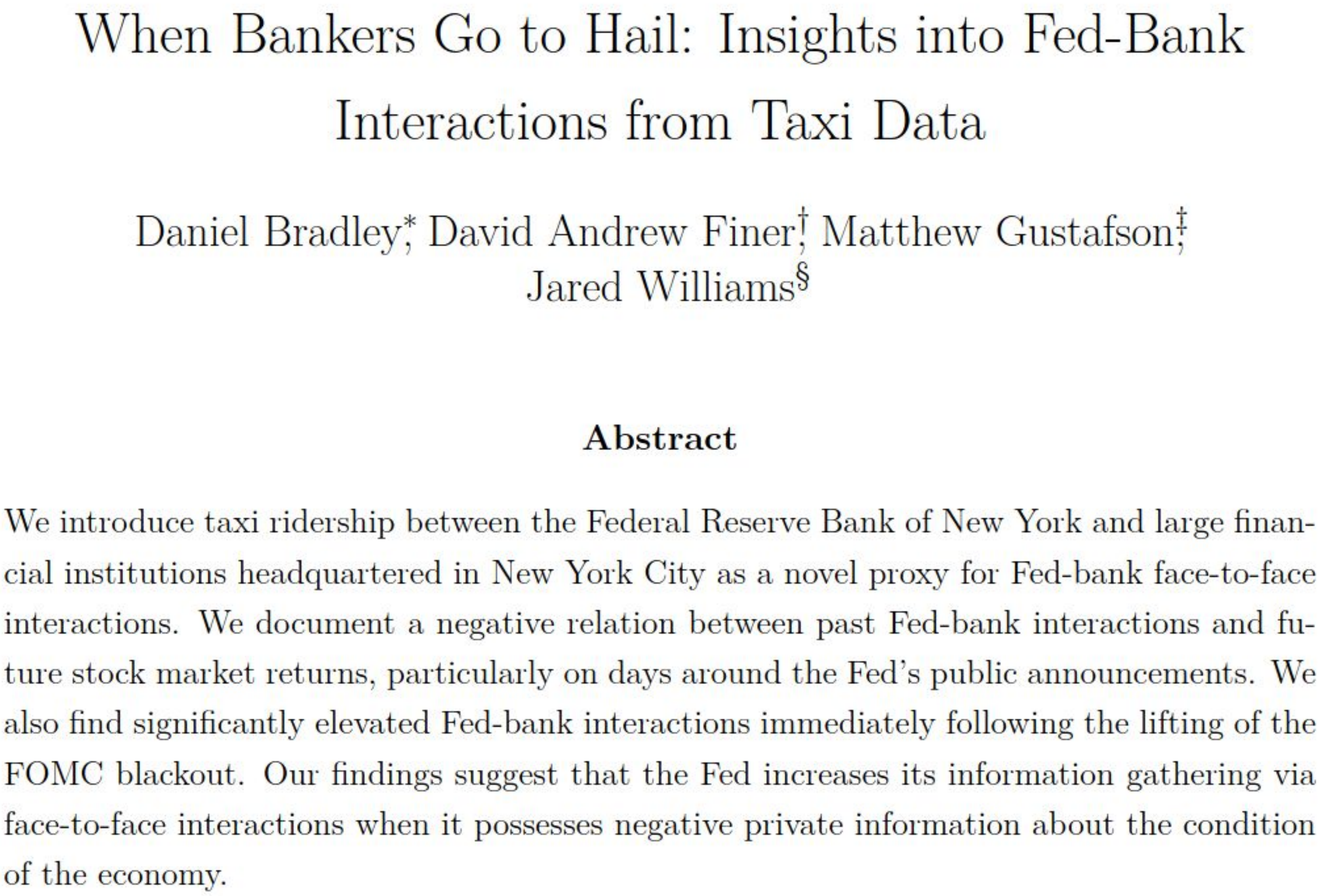
Prístupy riadené dátami

- Motivované spoločnosťou, kde sa dáta rutínne generujú pri každodenných aktivitách
- Aké zaujímavé problémy vieme riešiť pomocou daného súboru dát?



Vendor ID	passenger _count	trip_ distance	pickup_ longitude	pickup_ latitude	dropoff_ longitude	dropoff_ latitude	payment _type	tip_ amount	total_ amount
2	1	7.22	-73.9998	40.74334	-73.9428	40.80662	2	0	30.8
1	1	2.3	-73.977	40.7749	-73.9783	40.74986	1	2.93	16.23
1	1	1.5	-73.9591	40.77513	-73.9804	40.78231	1	1.65	9.95
1	1	0.9	-73.9766	40.78075	-73.9706	40.78885	1	1.45	8.75
2	1	2.44	-73.9786	40.78592	-73.9974	40.7563	1	2	16.3
2	1	3.36	-73.9764	40.78589	-73.9424	40.82209	1	3.58	17.88
2	2	2.34	-73.9862	40.76087	-73.9569	40.77156	1	1	13.8
2	1	10.19	-73.79	40.64406	-73.9312	40.67588	2	0	32.8
1	2	3.3	-73.9937	40.72738	-73.9982	40.7641	1	2	21.3
1	1	1.8	-73.9949	40.74006	-73.9767	40.74934	1	1.85	11.15

Ukážka z data setu “NY city taxi cab”
Skiena: Data Science Design Manual

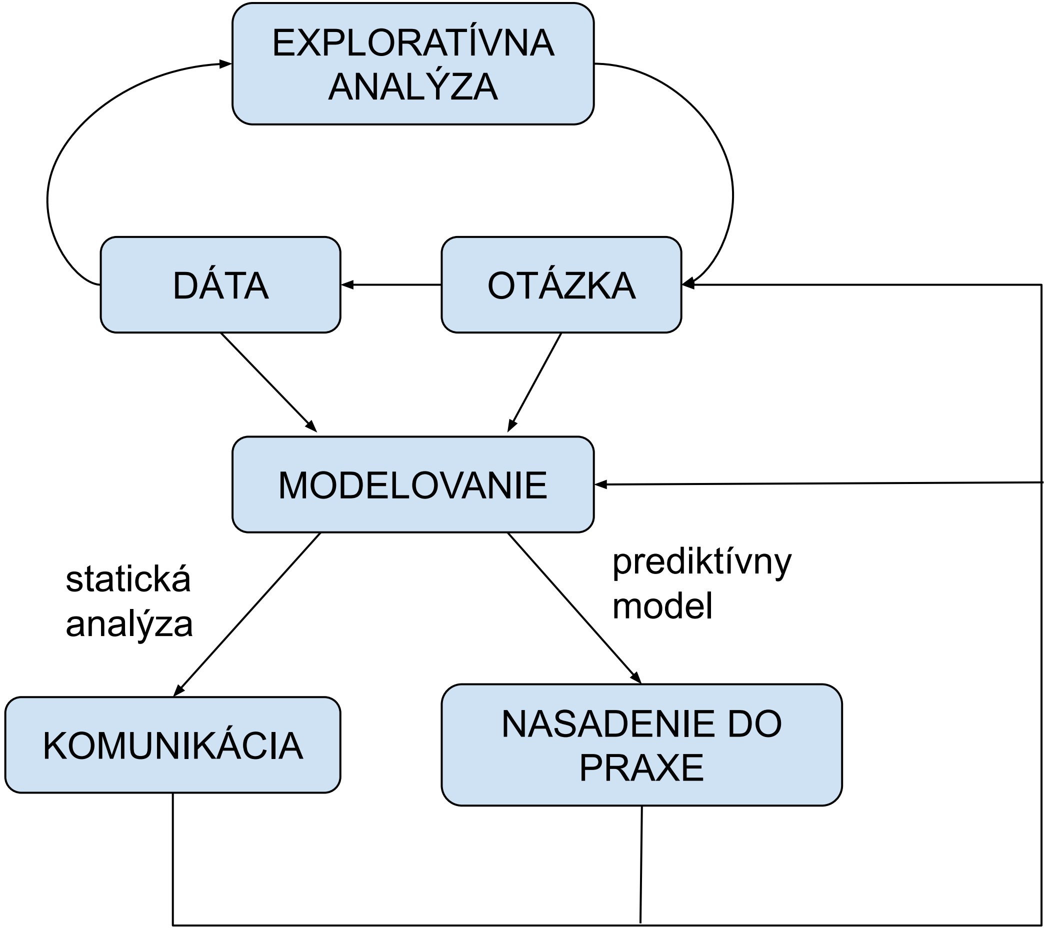


When Bankers Go to Hail: Insights into Fed-Bank Interactions from Taxi Data

Daniel Bradley[*], David Andrew Finer[†], Matthew Gustafson[‡], Jared Williams[§]

Abstract

We introduce taxi ridership between the Federal Reserve Bank of New York and large financial institutions headquartered in New York City as a novel proxy for Fed-bank face-to-face interactions. We document a negative relation between past Fed-bank interactions and future stock market returns, particularly on days around the Fed's public announcements. We also find significantly elevated Fed-bank interactions immediately following the lifting of the FOMC blackout. Our findings suggest that the Fed increases its information gathering via face-to-face interactions when it possesses negative private information about the condition of the economy.


EXPLORATÍVNA ANALÝZA
DÁTA
OTÁZKA
MODELOVANIE
statická analýza
prediktívny model
KOMUNIKÁCIA
NASADENIE DO PRAXE

Psychological Bulletin
1991, Vol. 109, No. 1, 90–106

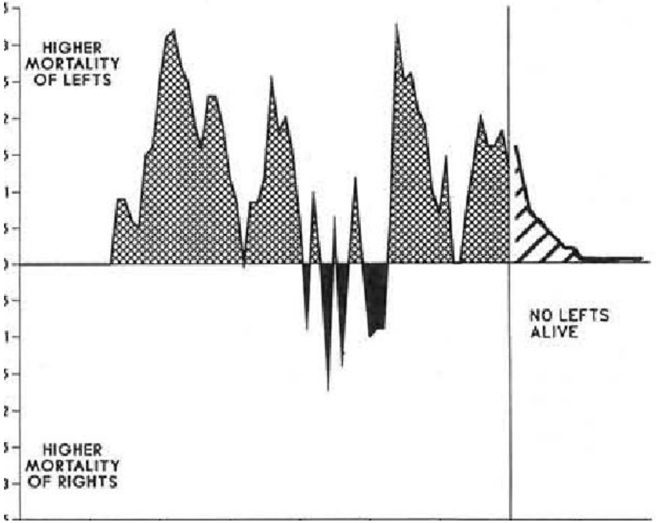
Left-Handedness: A Marker for Decreased Survival Fitness

Stanley Coren
University of British Columbia
Vancouver, British Columbia, Canada

Diane F. Halpern
California State University, San Bernardino

Life span studies have shown that the population percentage of left-handers diminishes steadily, so that they are drastically underrepresented in the oldest age groups. Data are reviewed that indicate that this population trend is due to the reduced longevity of left-handers. Some of the elevated risk for sinistrals is apparently due to environmental factors that elevate their accident susceptibility. Further evidence suggests that left-handedness may be a marker for birth stress related neuropathy, developmental delays and irregularities, and deficiencies in the immune system due to the intrauterine hormonal environment. Some statistical and physiological factors that may cause left-handedness to be selectively associated with earlier mortality are also presented.

Fortunately, we were able to locate a sample that met these stringent requirements. Reliable hand-use statistics and date of birth and death records are available in archival records for professional baseball players. Thus, the subjects we chose were all of the baseball players listed in *The Baseball Encyclopedia* (Reichler, 1979) for whom dates of birth and death and throwing and batting hand were reported ($N = 2{,}271$). Subjects were divided into handedness groups such that strong right handedness was coded when both throwing and batting hand were right and there was no indication of change in hand use. Similar criteria were used for strong left handedness. Individuals who changed handedness or who had mixed hand-use patterns were not included in this sample.



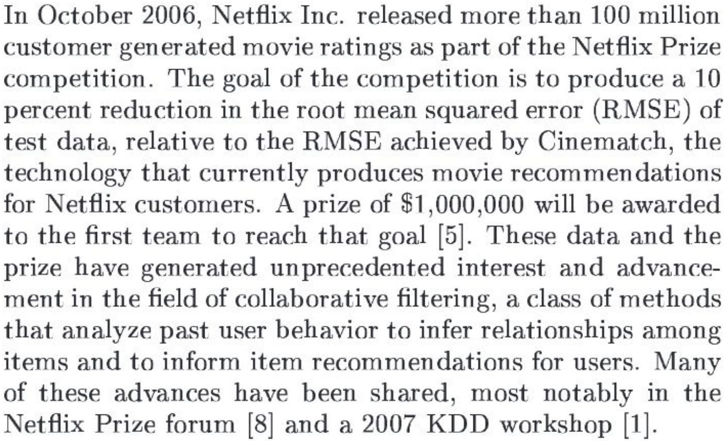
imdb.com

https://www.imdb.com/title/tt0092455/

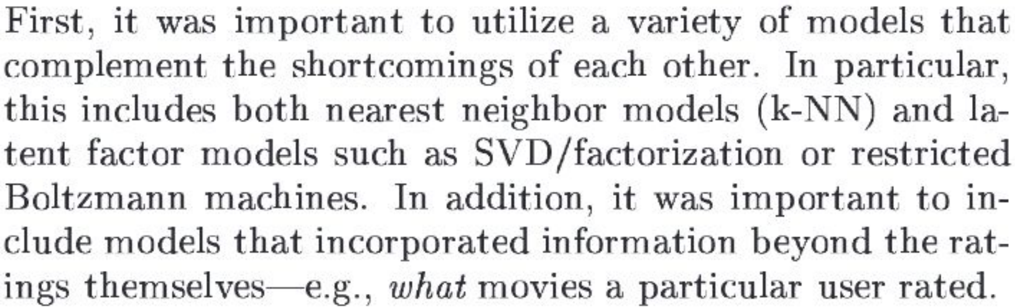
https://www.imdb.com/name/nm0000559/

Lessons from the Netflix Prize Challenge

Robert M. Bell and Yehuda Koren
AT&T Labs – Research
180 Park Ave, Florham Park, NJ
{rbell,yehuda}@research.att.com



In October 2006, Netflix Inc. released more than 100 million customer generated movie ratings as part of the Netflix Prize competition. The goal of the competition is to produce a 10 percent reduction in the root mean squared error (RMSE) of test data, relative to the RMSE achieved by Cinematch, the technology that currently produces movie recommendations for Netflix customers. A prize of $1,000,000 will be awarded to the first team to reach that goal [5]. These data and the prize have generated unprecedented interest and advancement in the field of collaborative filtering, a class of methods that analyze past user behavior to infer relationships among items and to inform item recommendations for users. Many of these advances have been shared, most notably in the Netflix Prize forum [8] and a 2007 KDD workshop [1].



First, it was important to utilize a variety of models that complement the shortcomings of each other. In particular, this includes both nearest neighbor models (k-NN) and latent factor models such as SVD/factorization or restricted Boltzmann machines. In addition, it was important to include models that incorporated information beyond the ratings themselves—e.g., *what* movies a particular user rated.

baseball-reference.com

https://en.wikipedia.org/wiki/Babe_Ruth

https://www.baseball-reference.com/players/r/ruthba01.shtml

Google Books Ngram viewer

https://books.google.com/ngrams/graph?content=data+processing%2Ccomputer+science%2Cdata+science%2Cmachine+learning%2Cinformatics&year_start=1940&year_end=2019&corpus=en-2019&smoothing=3&case_insensitive=true

https://books.google.com/ngrams/graph?content=color%2Ccolour&year_start=1800&year_end=2019&case_insensitive=on&corpus=en-2019&smoothing=3

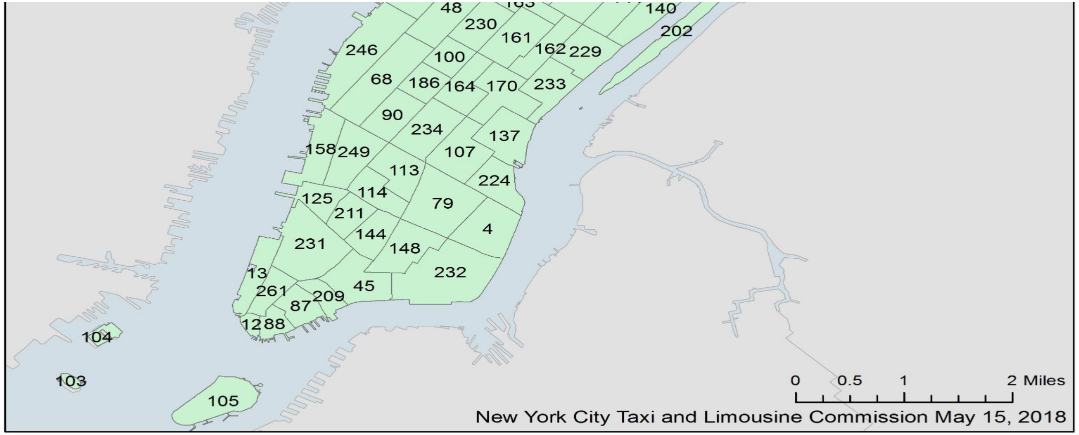
https://books.google.com/ngrams/graph?content=algorthm&year_start=1950&year_end=2019&corpus=en-2019&smoothing=3

https://books.google.com/ngrams/graph?content=virus%2Cbacteria%2Cplague%2Cpoison&year_start=1800&year_end=2019&corpus=en-2019&smoothing=3

https://books.google.com/ngrams/graph?content=market+crash%2Chousing+bubble&year_start=1800&year_end=2019&corpus=en-2019&smoothing=3

New York Taxi Recods

https://colab.research.google.com/drive/1pkyGBC1aOMP30S_n0YH5jgQpj2fDeVZ7?usp=sharing



New York City Taxi and Limousine Commission May 15, 2018