

# 1-DAV-302 Princípy dátovej vedy

## Vyučujúci:

Vladimír Boža, M-25, vladimir.boza@fmph.uniba.sk

Radoslav Harman, M-222, radoslav.harman@fmph.uniba.sk

Tomáš Vinař, M-163, tomas.vinar@fmph.uniba.sk

**Web:** <http://compbio.fmph.uniba.sk/vyuka/pridav>

**Google classroom:** <https://classroom.google.com> kód **libm7fs**

## 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 Google Classroom

<https://classroom.google.com> kód **libm7fs**

- **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 (3 úlohy v priebehu semestra)
- 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

## Rozdiely medzi dátovou vedou a informatikou / matematikou

- Dôraz na dáta (vs. dôraz na metódy)
- Dôležitosť skutočného výsledku  
(vs. všetko je ok, ak spĺňame špecifikáciu)
- Robustnosť voči chybám  
(vs. “garbage in - garbage out” prístup)
- Väčšina tvrdení nie je “pravda” alebo “nepravda”

# Prístupy ku riešeniu problémov

## Prístupy riadené hypotézou (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?

baseball-reference.com

[https://en.wikipedia.org/wiki/Babe\\_Ruth](https://en.wikipedia.org/wiki/Babe_Ruth)

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

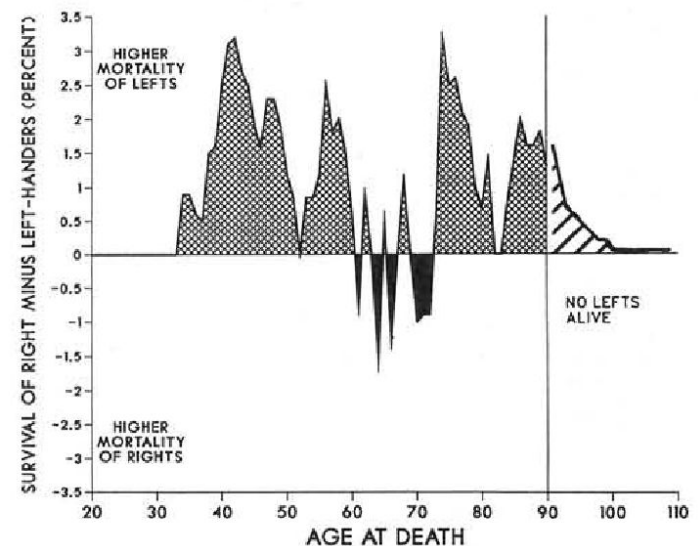
# Left-Handedness: A Marker for Decreased Survival Fitness

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.

# 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=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=color%2Ccolour&year_start=1800&year_end=2019&case_insensitive=on&corpus=en-2019&smoothing=3)

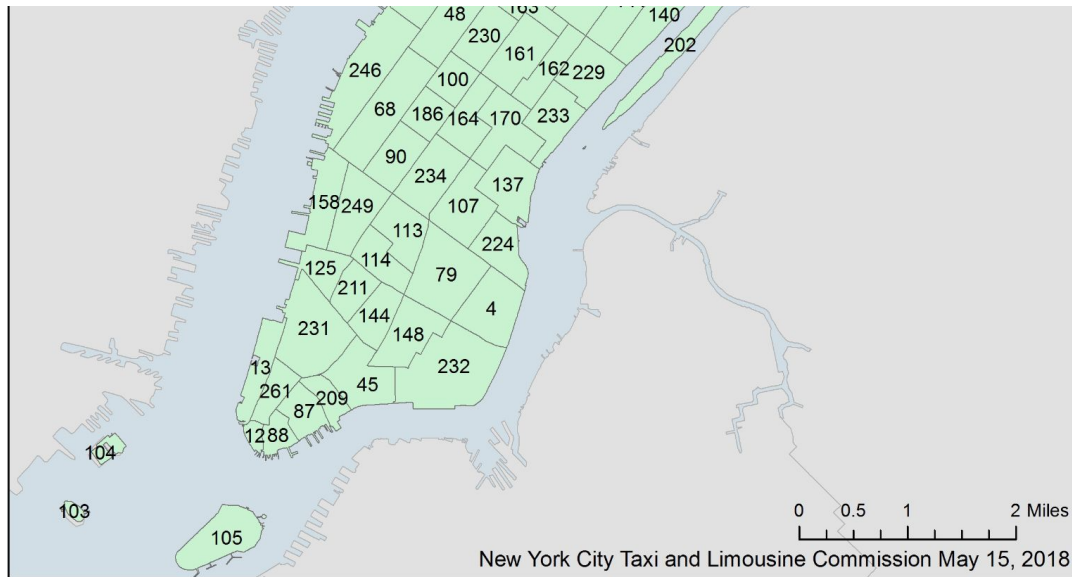
[https://books.google.com/ngrams/graph?content=algorithm&year\\_start=1950&year\\_end=2019&corpus=en-2019&smoothing=3](https://books.google.com/ngrams/graph?content=algorithm&year_start=1950&year_end=2019&corpus=en-2019&smoothing=3)

[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=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](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](https://colab.research.google.com/drive/1pkyGBC1aOMP30S_n0YH5jgQpj2fDeVZ7?usp=sharing)



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

Daniel Bradley\*, David Andrew Finer<sup>†</sup>, Matthew Gustafson<sup>‡</sup>,  
Jared Williams<sup>§</sup>

## 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.