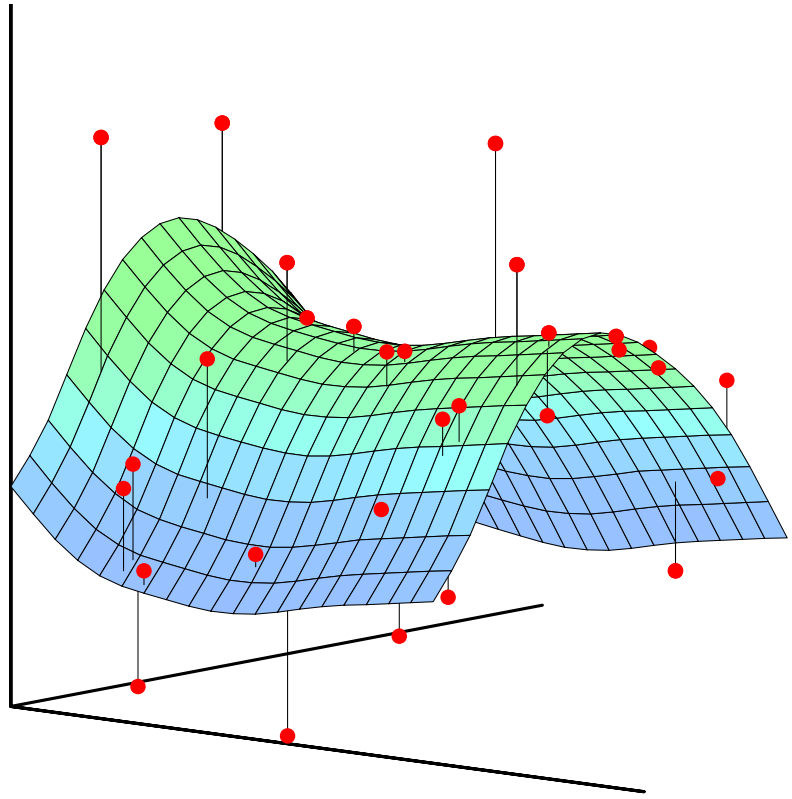
# Statistické učení



Trevor Hastie a Robert Tibshirani
přeložil a upravil Jan Přikryl

# Statistika ve zprávách

**Jak IBM vybudovalo Watsona, svůj superpočítač hrající hru Jeopardy** od Dawna Kawamoto, DailyFinance, 8.2.2011


**Učení se z vlastních chyb** Podle Davida Ferrucciho (Watson DeepQA technology, IBM Research) je Watsonův software určen nejen ke zvládnutí práce s přirozeným jazykem.

*„Jeho strojové učení počítači umožňuje, aby se stával chytřejším tak, jak se pokouší odpovídat na otázky – a aby se učil z toho, jak zjišťuje, zda odpověděl správně nebo špatně.“*

# For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

SIGN IN TO RECOMMEND

SIGN IN TO E-MAIL

PRINT

REPRINTS

SHARE



MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

Enlarge This Image

Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

**Multimedia**

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

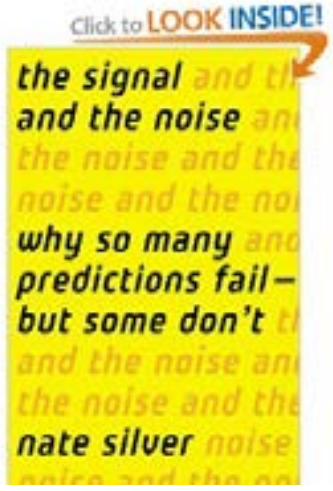
Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Citát dne, New York Times, 5. srpna 2009
„I nadále říkám, že sexy povolání příštích deseti let budou mít statistici. A to nekecám." —Hal Varian, hlavní ekonom Googlu

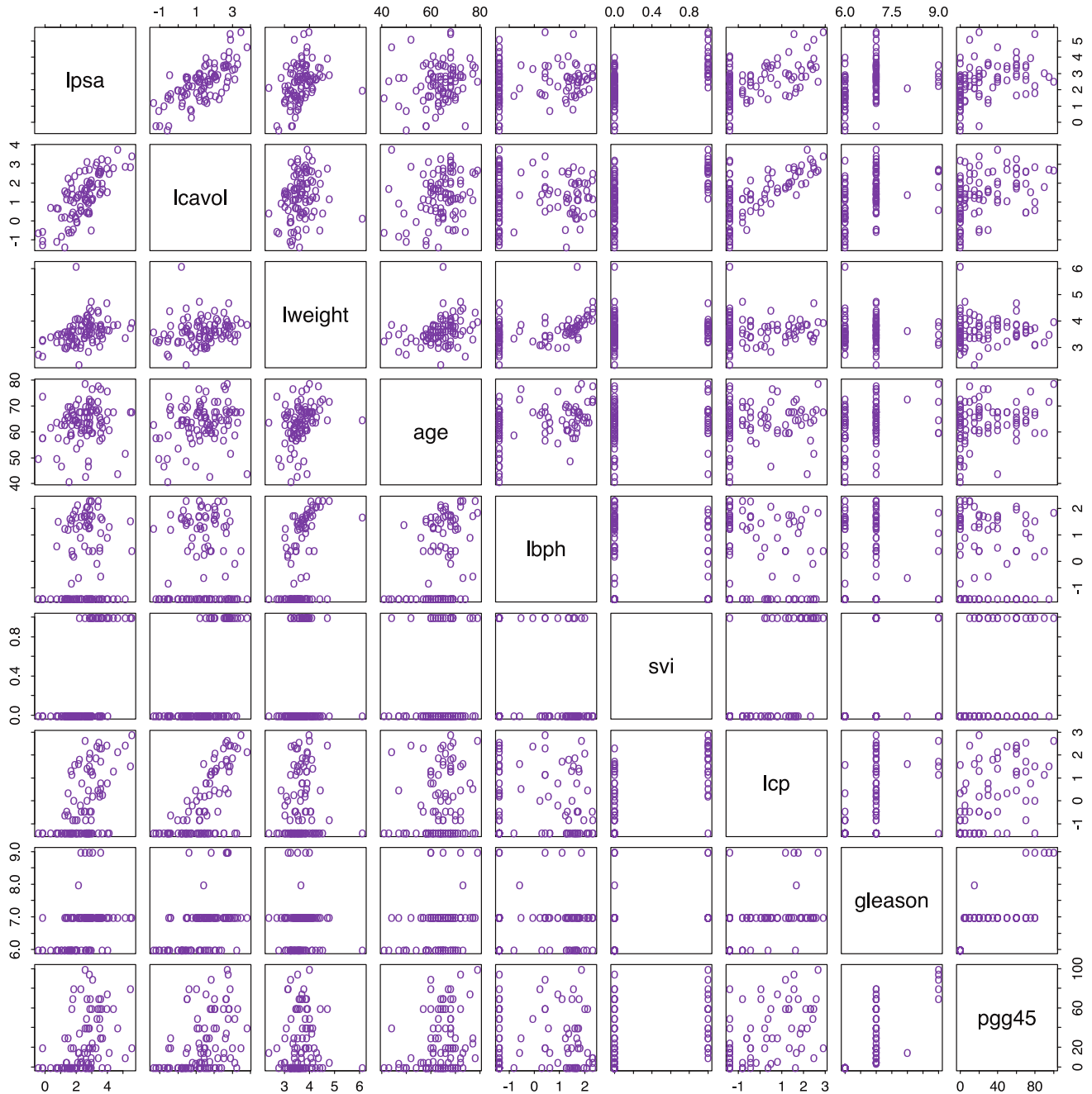
# FiveThirtyEight

**Nate Silver's Political Calculus**

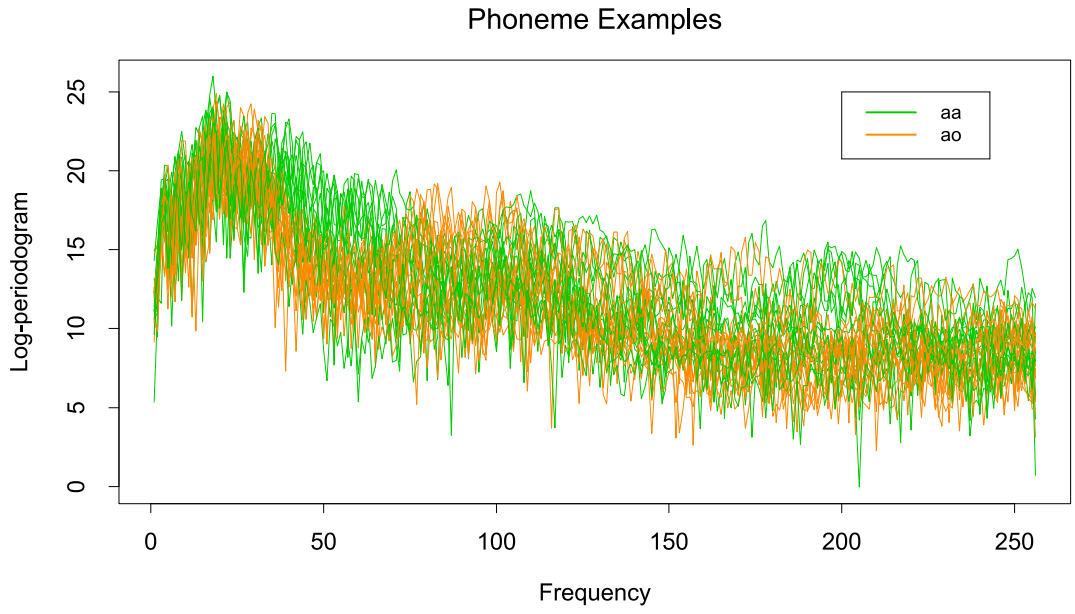
# Úlohy statistického učení

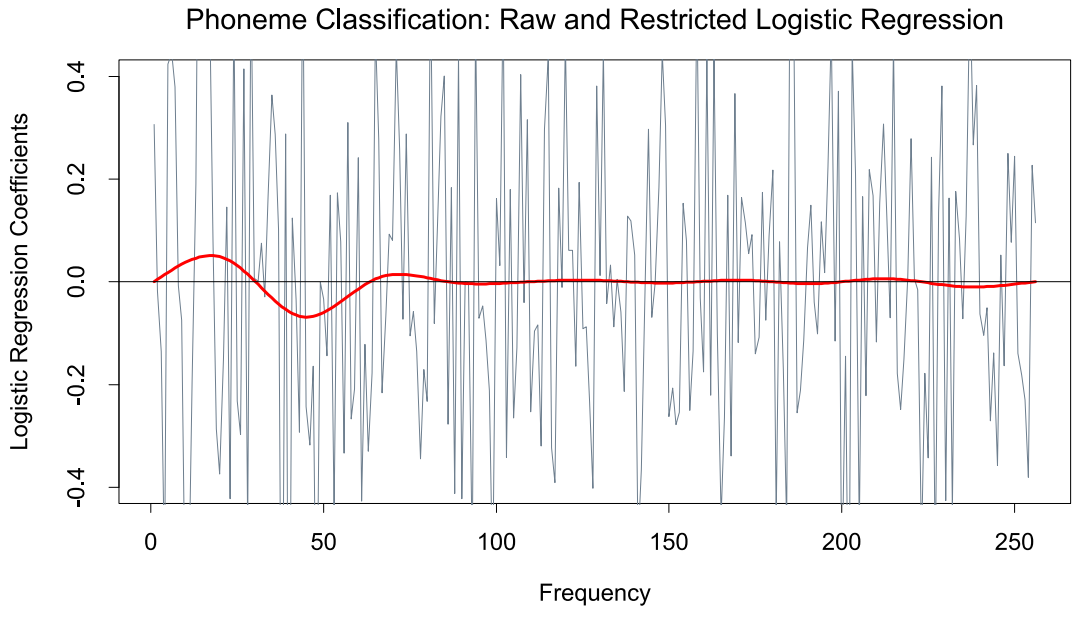
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45

# Úlohy statistického učení

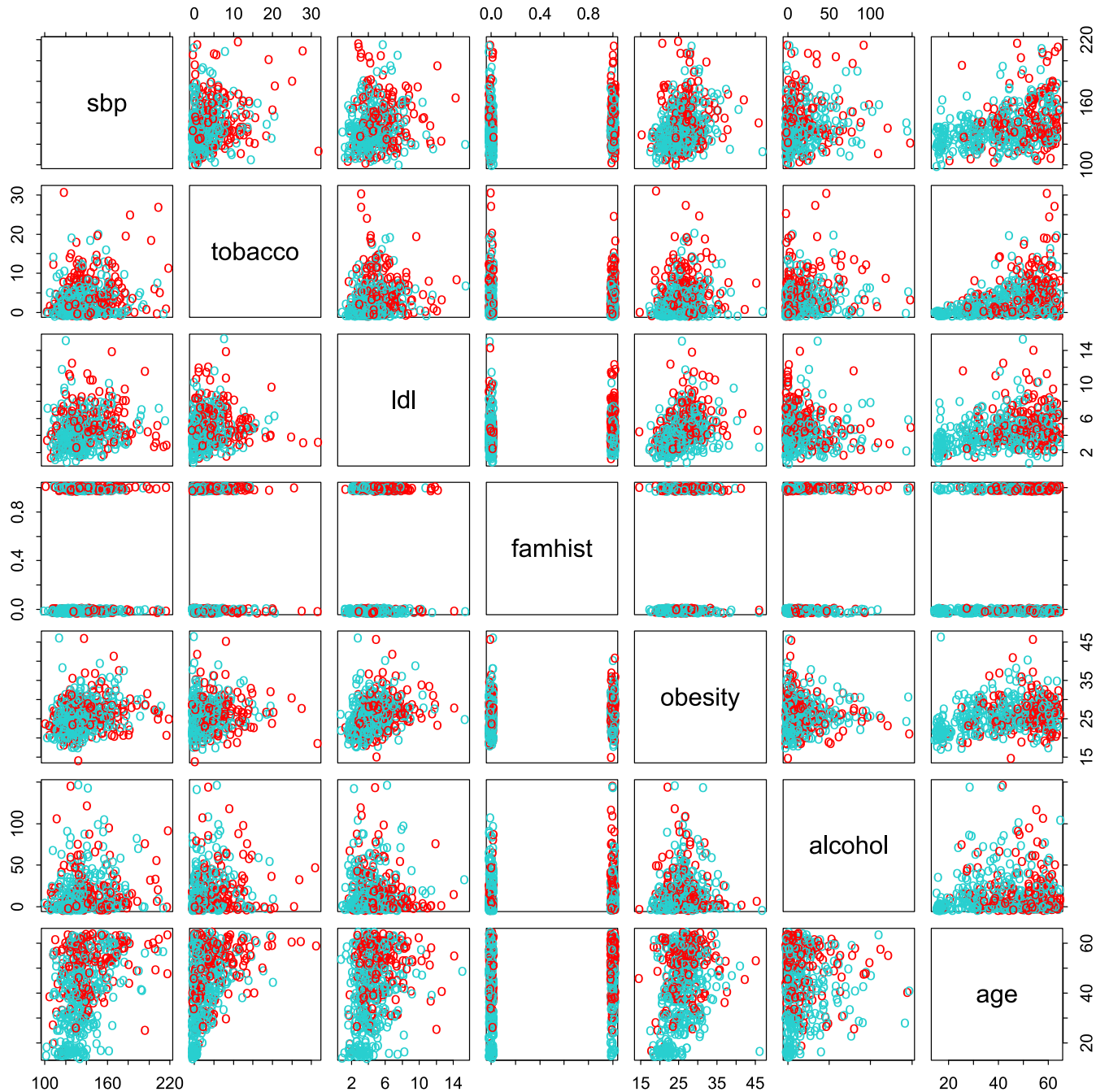
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.


Phoneme Examples
Log-periodogram
Frequency
aa
ao
0
5
10
15
20
25
0
50
100
150
200
250


Phoneme Classification: Raw and Restricted Logistic Regression
Logistic Regression Coefficients
Frequency
-0.4
-0.2
0.0
0.2
0.4
0
50
100
150
200
250

# Úlohy statistického učení

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.


sbp
tobacco
ldl
famhist
obesity
alcohol
age

# Úlohy statistického učení

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

# Detekce spamu

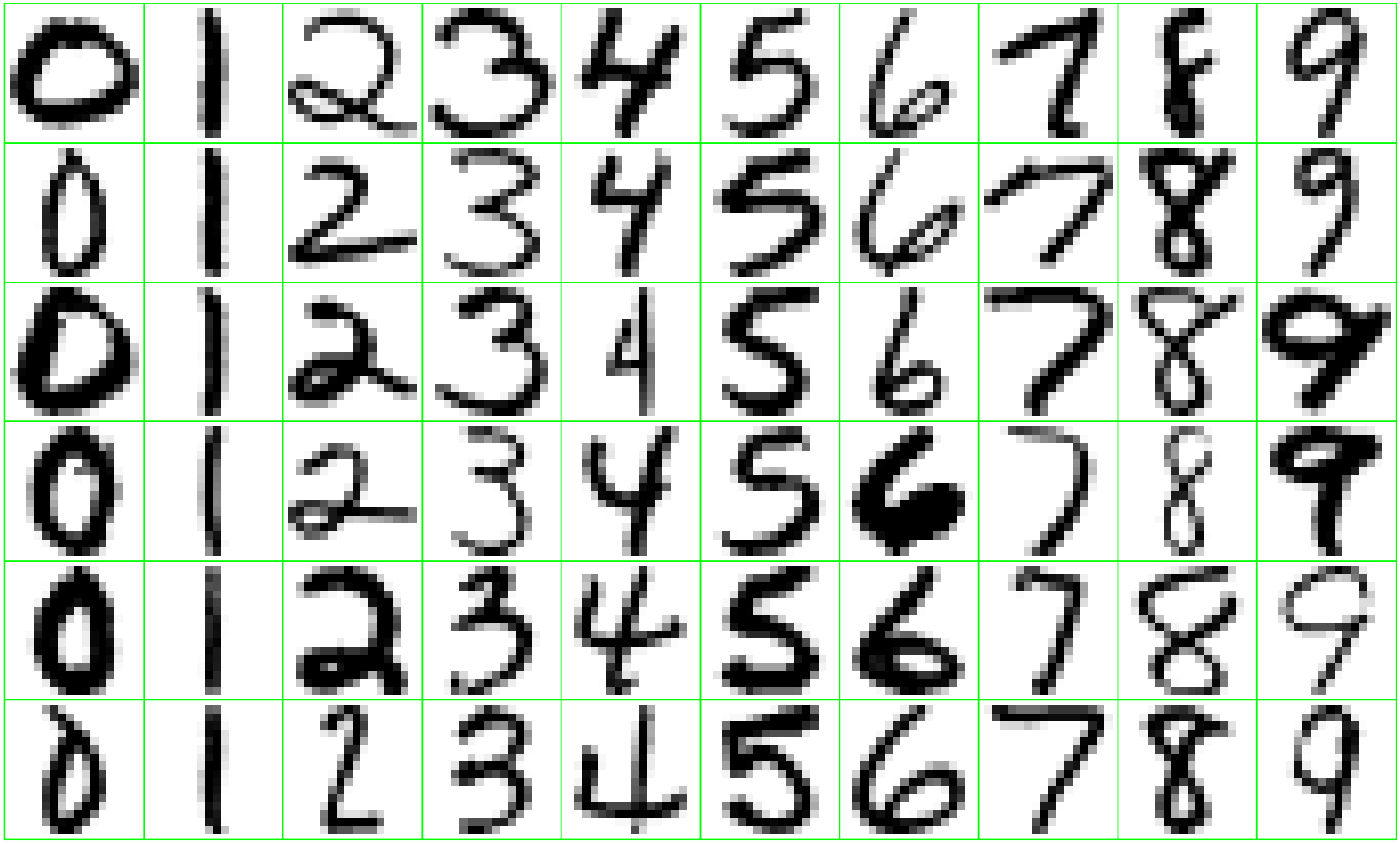
- data ze 4601 emailových zpráv zaslaných jistému člověku (jménem George, HP labs, před rokem 2000). Každá zpráva je označena jako *spam* nebo *email*.
- cíl: vytvořit z těchto dat systém pro detekci spamu.
- vstupní proměnné: relativní frekvence 57 slov a interpunkčních znamének, které se v těchto emailových správách vyskytují nejčastěji.

| | george | you | hp | free | ! | edu | remove |
|---|---|---|---|---|---|---|---|
| spam | 0.00 | 2.26 | 0.02 | 0.52 | 0.51 | 0.01 | 0.28 |
| email | 1.27 | 1.27 | 0.90 | 0.07 | 0.11 | 0.29 | 0.01 |

*Průměrný procentní podíl slov nebo znaků ve zprávě, která se rovnají slovům nebo znakům uvedeným v tabulce. Vybrali jsme slova a znaky vykazující největší rozdíl mezi* `spamem` *a* `emailem`.

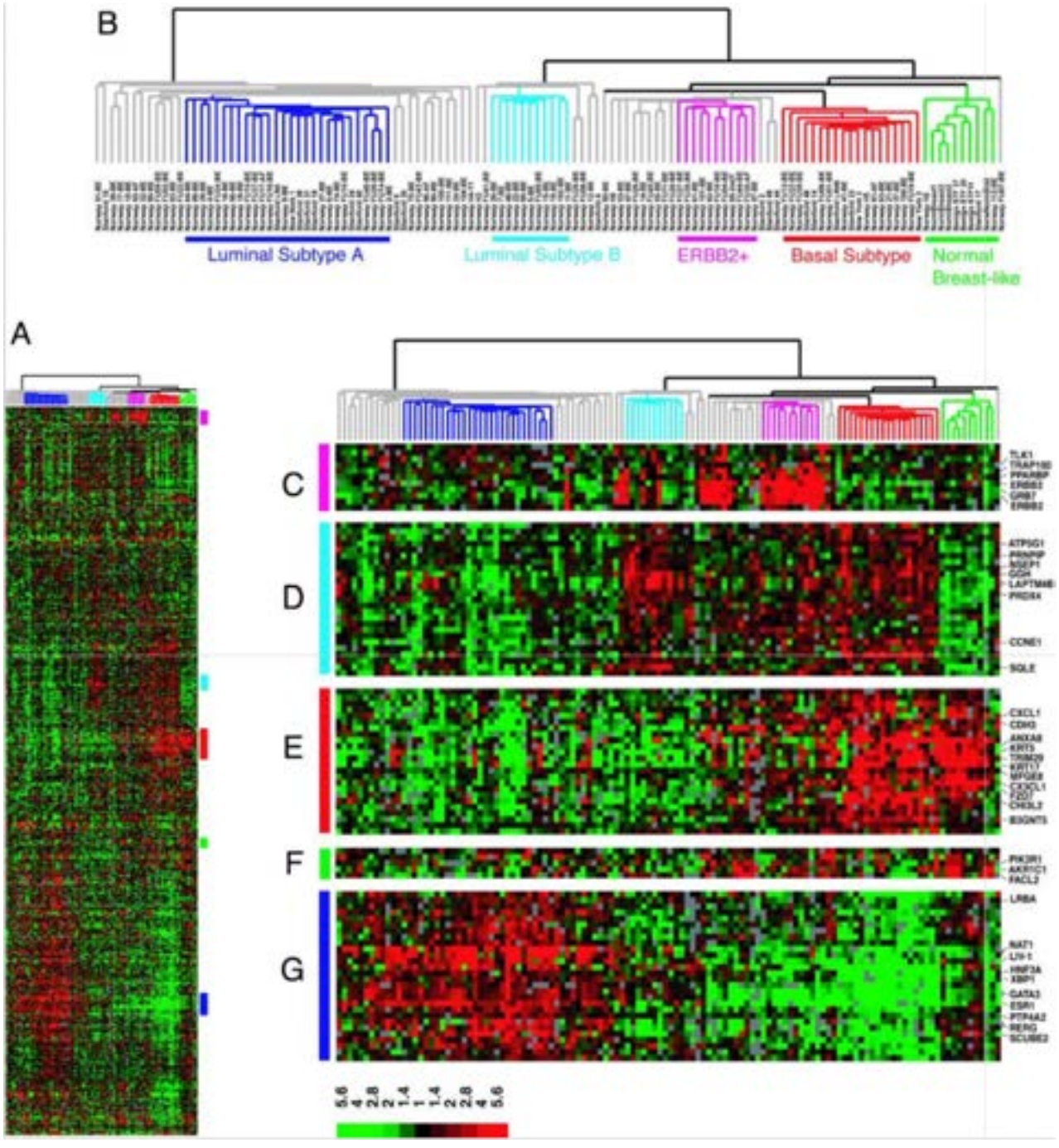
# Úlohy statistického učení

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



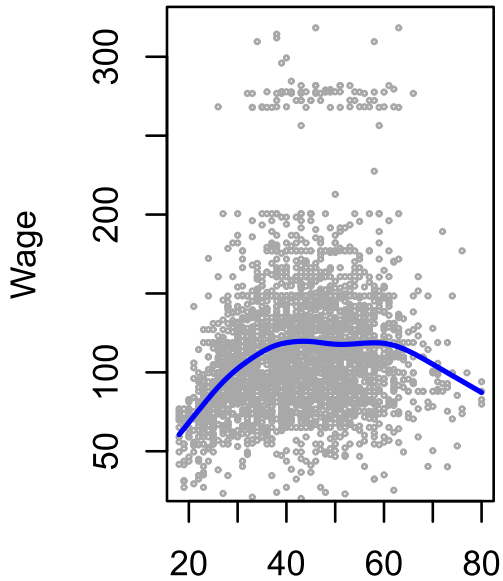
# Úlohy statistického učení

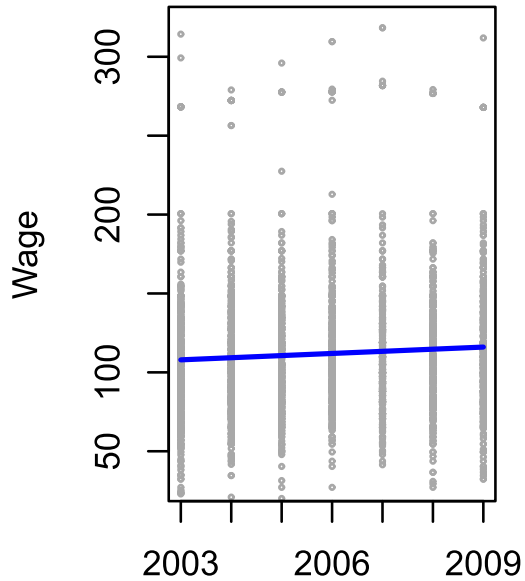
- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

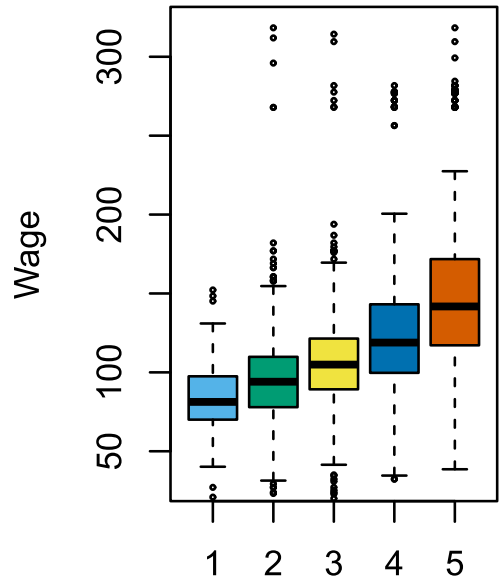

B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
5.6 4 2.8 2 1.4 1 1.4 2 2.8 4 5.6

# Úlohy statistického učení

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.



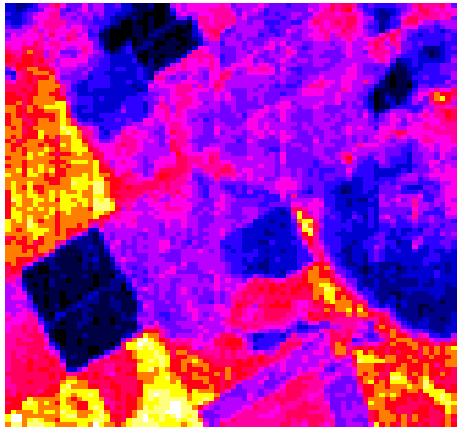




Datová studie příjmu pro muže ve střední atlantické oblasti USA z roku 2009.

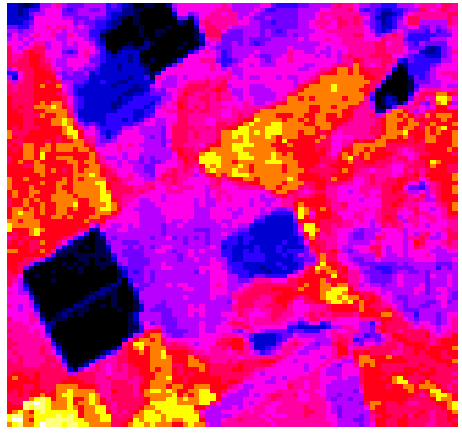
# Úlohy statistického učení

- Identifikace rizikových faktorů pro rakovinu prostaty.
- Klasifikace zaznamenaných fonémů založená na logaritmickém periodogramu.
- Předpověď, zda někdo bude mít infarkt, na základě demografických, dietních a klinických měření.
- Zdokonalování systému pro detekci emailového spamu.
- Rozpoznávání čísel v ručně psaném PSČ.
- Klasifikace vzorku tkáně do jedné z několika tříd rakoviny na základě profilu genové exprese.
- Stanovení vztahu mezi mzdovými a demografickými proměnnými v populačních datových studiích.
- Klasifikace pixelů v obrázku získaném Landsatem podle využití půdy.

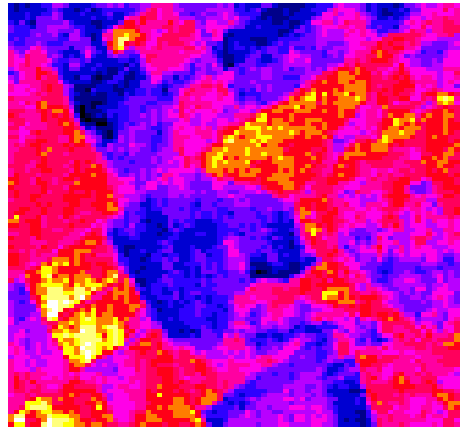
Spectral Band 1



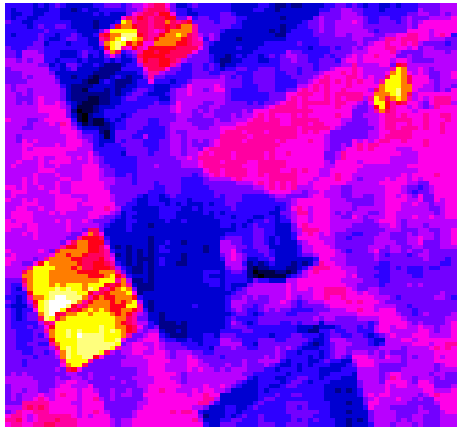
Spectral Band 2



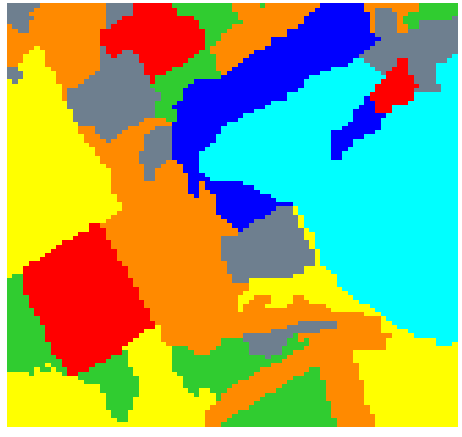
Spectral Band 3



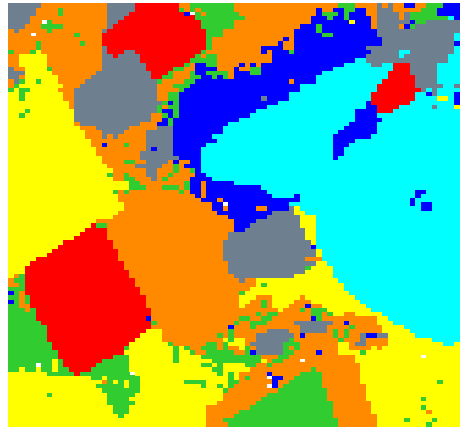
Spectral Band 4



Land Usage



Predicted Land Usage



*Využití* $\in \{$*červená zemina, bavlna, strniště, směs, šedá zemina, vlhká šedá zemina*$\}$

# Úloha supervizovaného učení (učení s učitelem)

*Výchozí situace:*

- Výstupní měření $Y$ (také závisle proměnná, odpověď nebo cílová hodnota).
- Vektor $p$ měření prediktoru $X$ (také vstupy, regresory, vlastnosti, charakteristiky nebo nezávisle proměnné).
- V *regresní úloze* je $Y$ kvantitativní (např. cena, hodnota krevního tlaku).
- V *klasifikační úloze* nabývá $Y$ hodnot z konečné neuspořádané množiny (přežil/zemřel, číslice 0 až 9, třída rakoviny ve vzorku tkáně).
- Máme trénovací data $(x_1, y_1), \ldots, (x_N, y_N)$. Jsou to výsledky pozorování (příklady, případy) těchto měření.

# Cíle

Na základě trénovacích dat bychom chtěli:

- Přesně předpovídat dosud nezhlédnuté testovací případy.
- Pochopit, které vstupní hodnoty ovlivňují výstupní hodnotu a jak.
- Posoudit kvalitu našich předpovědí a závěrů.

# Filozofie

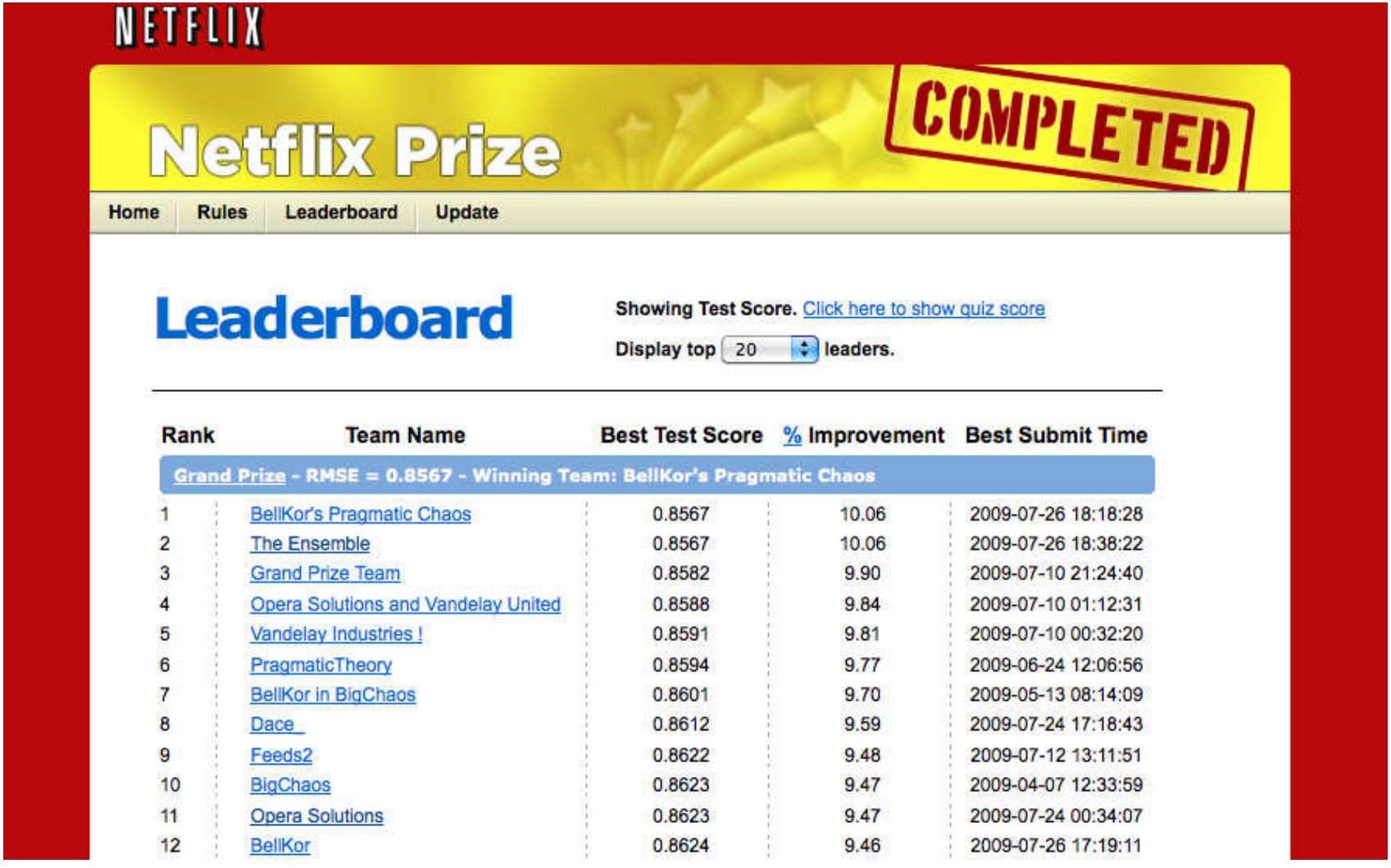
- Je důležité chápat myšlenky rozličných technik, abychom věděli, kdy a jak je používat.
- Člověk musí nejprve pochopit jednodušší metody, aby mohl zvládnout ty náročnější.
- Je důležité přesně posoudit výkonnost metody, abychom věděli jak dobře nebo jak špatně funguje [jednodušší metody jsou často stejně dobré jako ty fajnovější!].
- Toto je vzrušující oblast výzkumu, která má aplikace ve vědě, průmyslu a financích.
- Statistické učení je zásadní součástí tréninku moderního *experta v oboru zpracování dat*.

# Učení bez učitele (bez supervize)

- Žádná výstupní proměnná, pouze sada prediktorů (vlastností) naměřená na sadě vzorků.
- Cíl je mlhavější – nalézt skupiny vzorků, které se chovají podobně, nalézt proměnné, které se chovají podobně, nalézt lineární kombinace vlastností s největší variací.
- Obtížně se pozná, jak dobře si vedete.
- Odlišné od učení s učitelem, ale může být užitečné jako preprocesor pro učení s učitelem.

# Cena Netflixu

- soutěž započala v říjnu 2006. Trénovací data tvořilo hodnocení 18 000 filmů zákazníky Netflixu v počtu 400 000, jednotlivá hodnocení jsou 1 až 5.
- trénovací data jsou velmi řídká – asi v 98 % chybí.
- cílem je předpovědět hodnocení pro soubor jednoho milionu dvojic zákazník-film, jež se v trénovacích datech nevyskytují.
- původní algoritmus Netflixu dosahoval střední kvadratické chyby (s odmocninou) 0,953. První tým, který dosáhne zlepšení o 10 %, vyhrává 1 milion dolarů.
- je to úloha s učitelem nebo bez učitele?

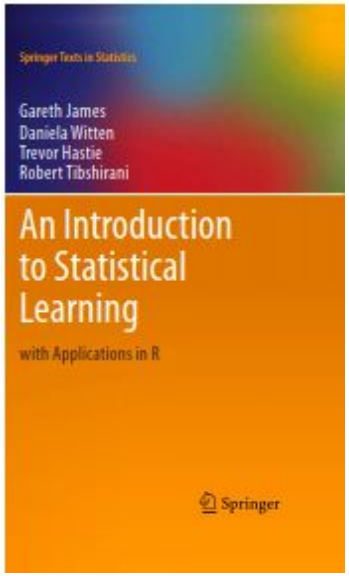



Vyhrál tým s názvem BellKor's Pragmatic Chaos, který velmi těsně porazil The Ensemble.

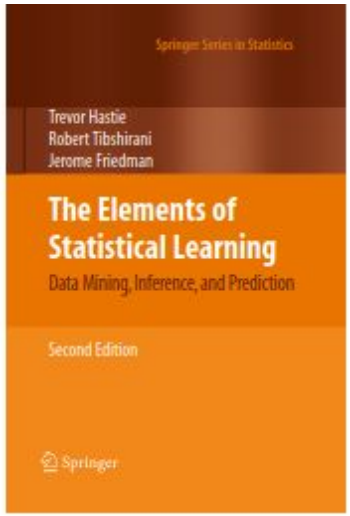
# Statistické učení versus strojové učení

- Strojové učení vzniklo jako podobor umělé inteligence.
- Statistické učení vzniklo jako podoblast statistiky.
- *Je zde značný překryv* – oba obory se zaměřují na úlohy učení s učitelem a bez učitele:
  - Strojové učení klade větší důraz na aplikace ve *velkém měřítku* a na *přesnost předpovědi*.
  - Statistické učení klade důraz na *modely* a jejich interpretovatelnost a na *přesnost* a *nejistotu*.
- Ale rozdíly se stále více a více stírají a je zde velký podíl vzájemného obohacování.
- Strojové učení má navrch v *marketingu*!

# *Literatura ke kurzu*




Kurz pokrývá většinu látky z této knihy (ISLR), která vyšla u Springera v roce 2013 a jejímiž autory jsou kromě instruktorů ještě Gareth James a Daniela Wittenová. Každá kapitola končí cvičením v jazyce R, v němž se zpracovávají příklady. Od 1. ledna 2014 bude elektronická verze knihy bezplatně dostupná z webových stránek instruktorů.




Tato Springerova kniha (ESL) je matematicky pokročilejší než ISLR; druhé vydání vyšlo v roce 2009 a jeho autory byli instruktoři kurzu a Jerome Friedman. Pokrývá širší rozsah témat. Kniha je dostupná od Springera a na Amazonu, elektronická verze je zdarma dostupná z webových stránek autorů.