

Strojové učení a umělá intelligence v praxi

PV201 Portálové technologie v praxi

Jiří Materna, Vedoucí týmu výzkumu



SEZNAM.CZ

Osnova

- **Úvod**
- **Strojové učení**
- **Rankování**
- **Zpracování dotazů**
- **BigData**
- **Distribuované výpočty**
- **Volitelné téma podrobně**
- **Diskuze**

Fulltextové vyhledávání ve světě

The Google logo, featuring the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red) with a trademark symbol (TM) to the upper right.The Baidu logo, consisting of the word "Baidu" in red and blue, with a blue paw print icon above the "i", and the Chinese characters "百度" in red to the right.The Yandex logo, featuring the word "Яндекс" in a stylized Cyrillic font, with the "Я" in red and the rest in black.

Yandex

The Seznam logo, featuring a large red "S" followed by the word "EZNAM" in a bold, black, sans-serif font.The Yahoo! Japan logo, featuring the word "YAHOO!" in a bold, red, sans-serif font, with "JAPAN" in a smaller, red, sans-serif font below it.The Naver logo, featuring a yellow and green circular icon with a white feather-like shape on the left, followed by the word "NAVER" in a bold, green, sans-serif font.

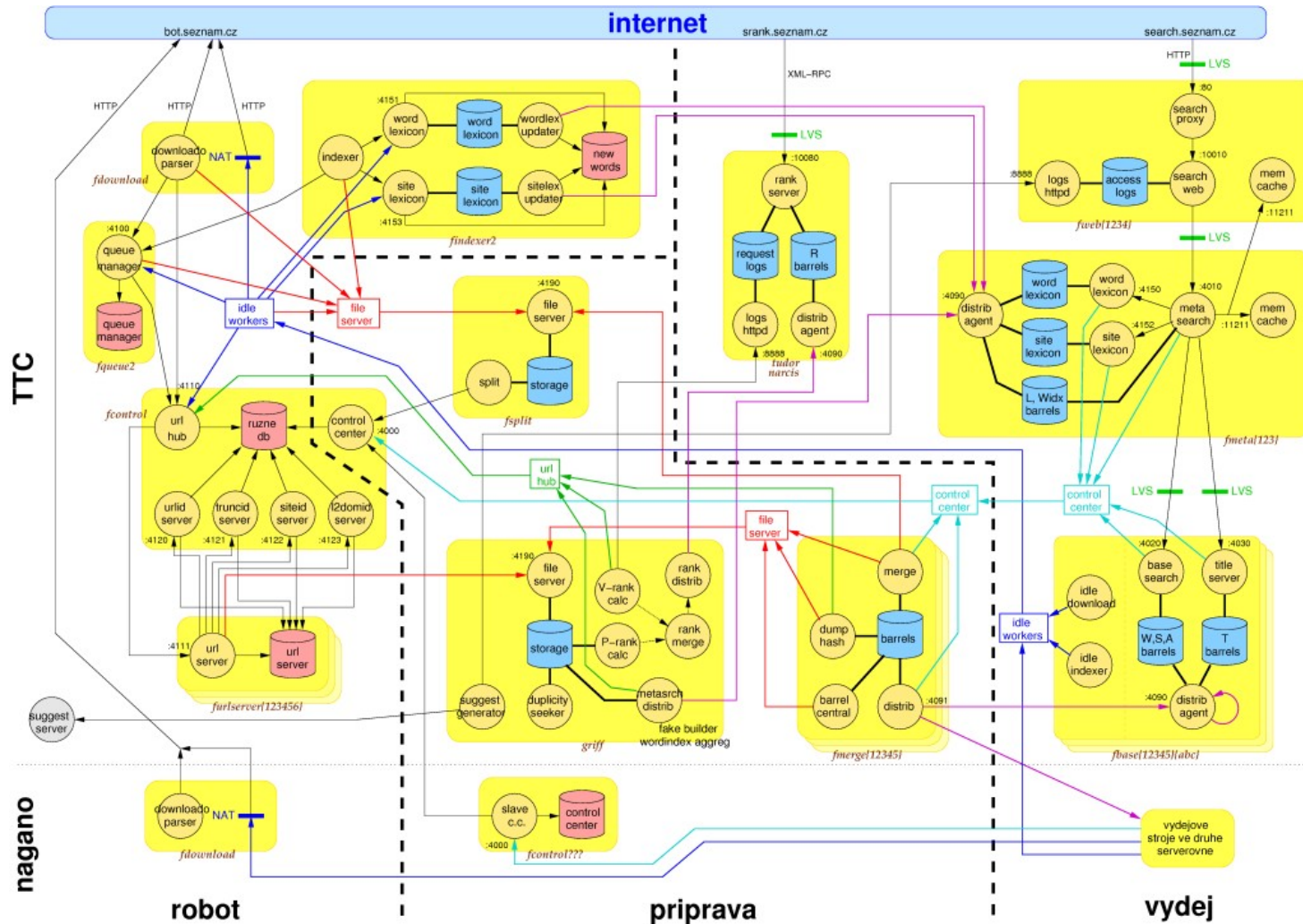
Search.seznam.cz v číslech

- 350 dotazů za sekundu, přes 500 ve špičce
- 15 milionů dotazů denně
- 600 milionů prohledávaných dokumentů
- 10 miliard známých odkazů
- 1000 dokumentů stažených za sekundu

Cizojazyčné vyhledávání

- 65 % čeština**
- 25 % angličtina**
- 3 % slovenština**
- 3 % němčina**
- 4 % ostatní jazyky**

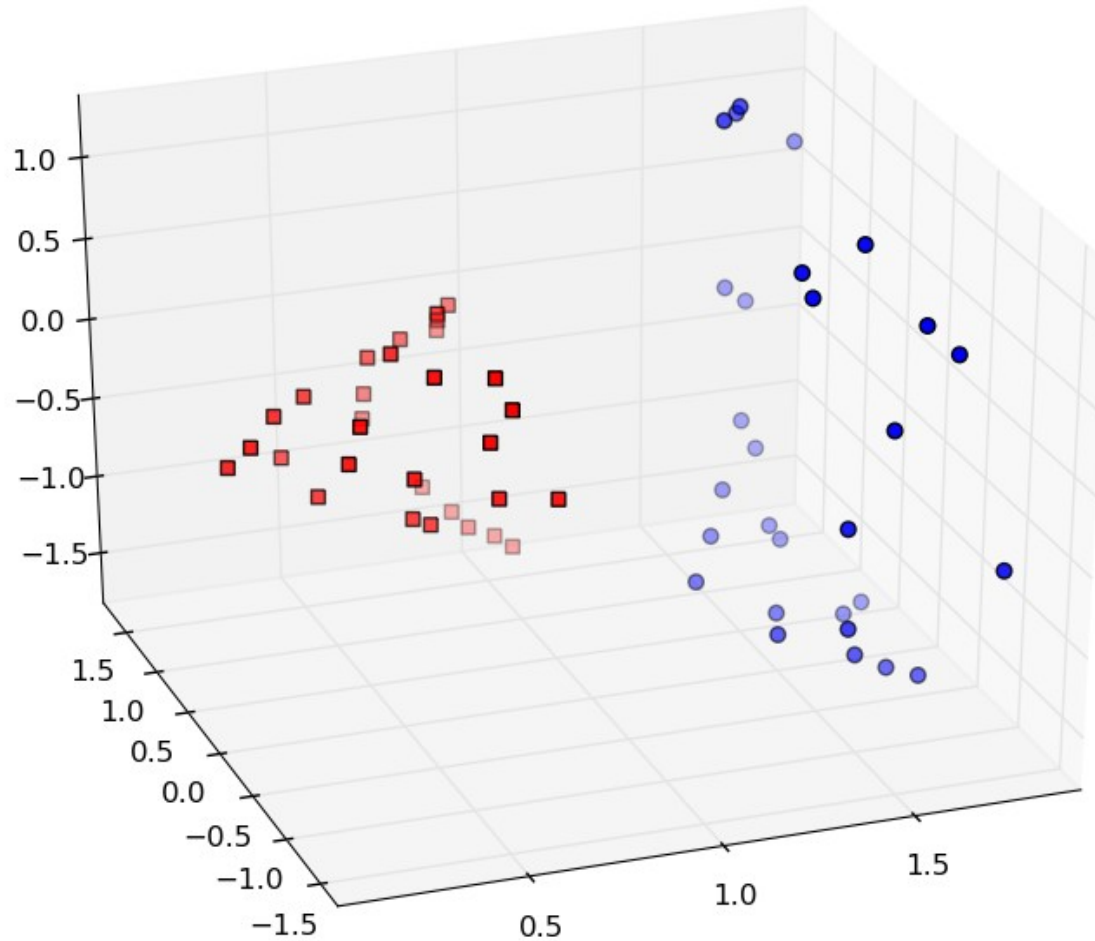
Blokové schéma fulltextu (2010)



Strojové učení



Klasifikační úloha



Filtrování nevhodného obsahu



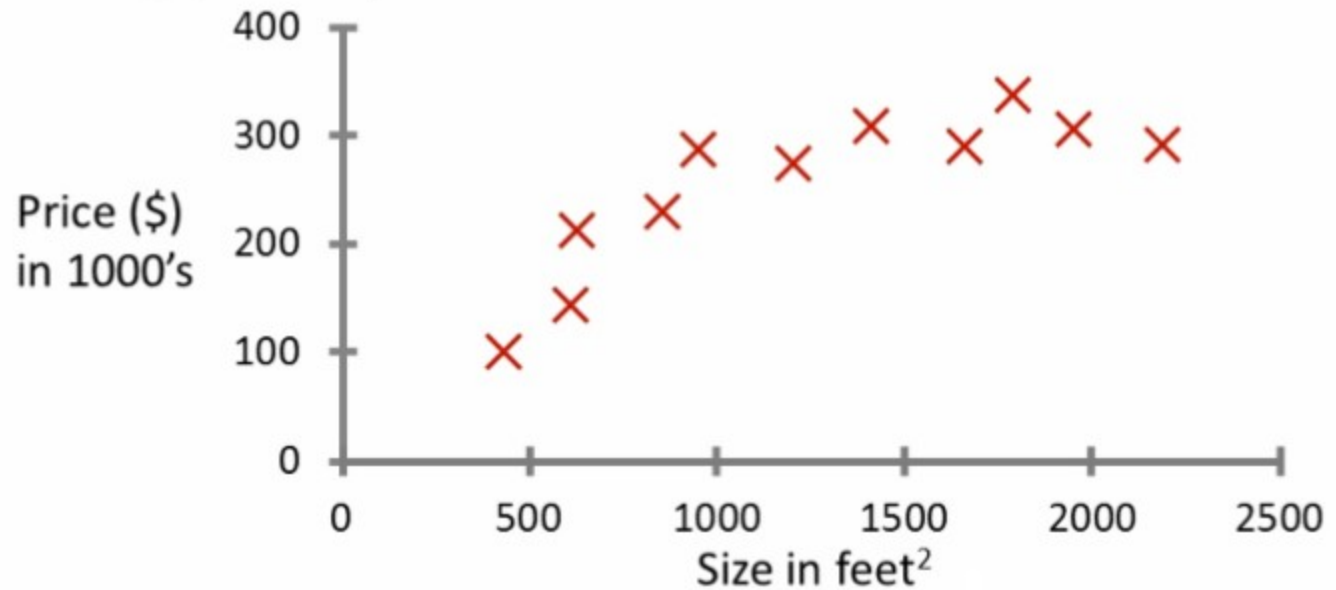
Filtrování nevhodného obsahu



Podíl porna na internetu: 1.5 %

Regresní úloha

Housing price prediction.



Rankování stránek

Rank = míra příslušnosti stránky k dotazu

Rankování stránek

Rank = míra příslušnosti stránky k dotazu

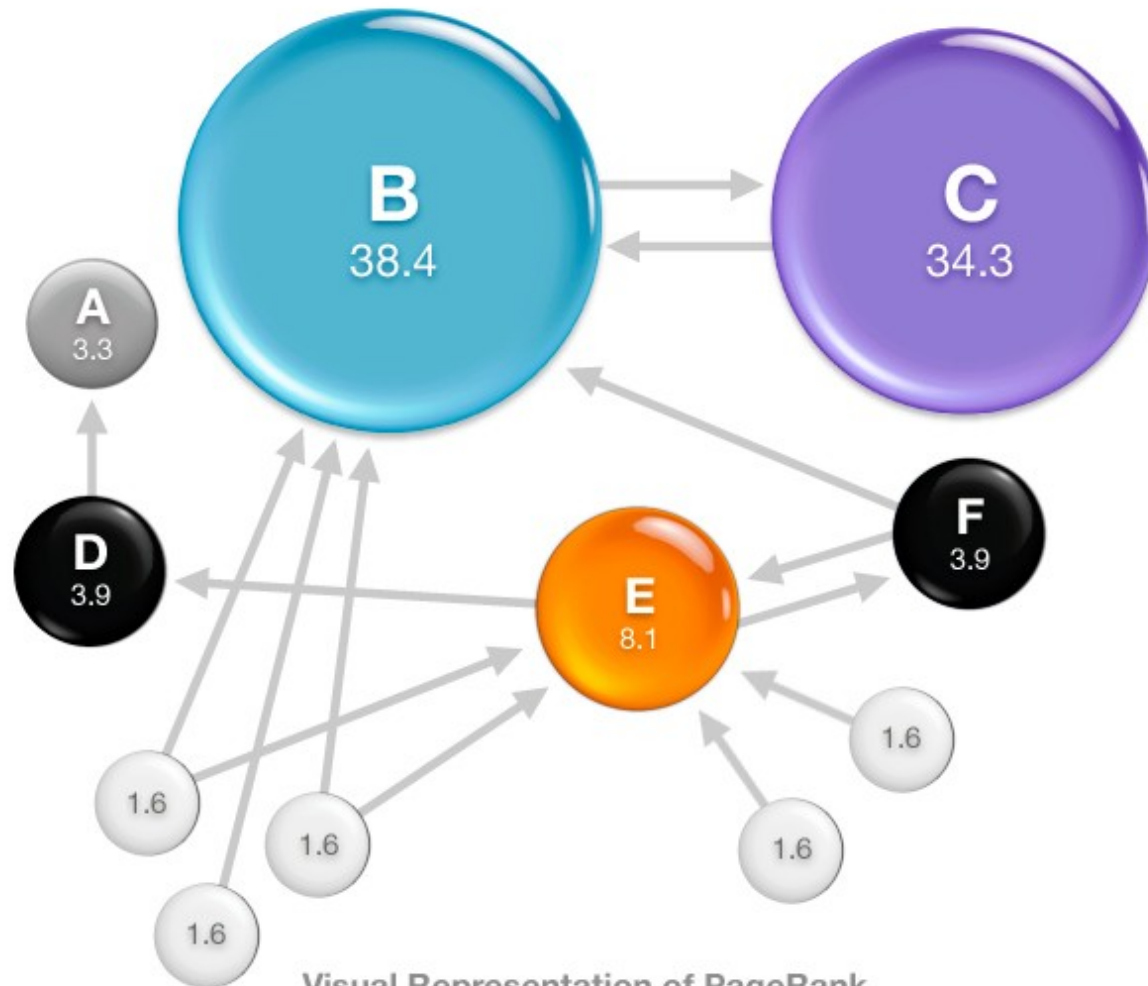
on-page vs. off-page faktory

signály dotazu

signály dokumentu

signály kombinace dotazu a dokumentu

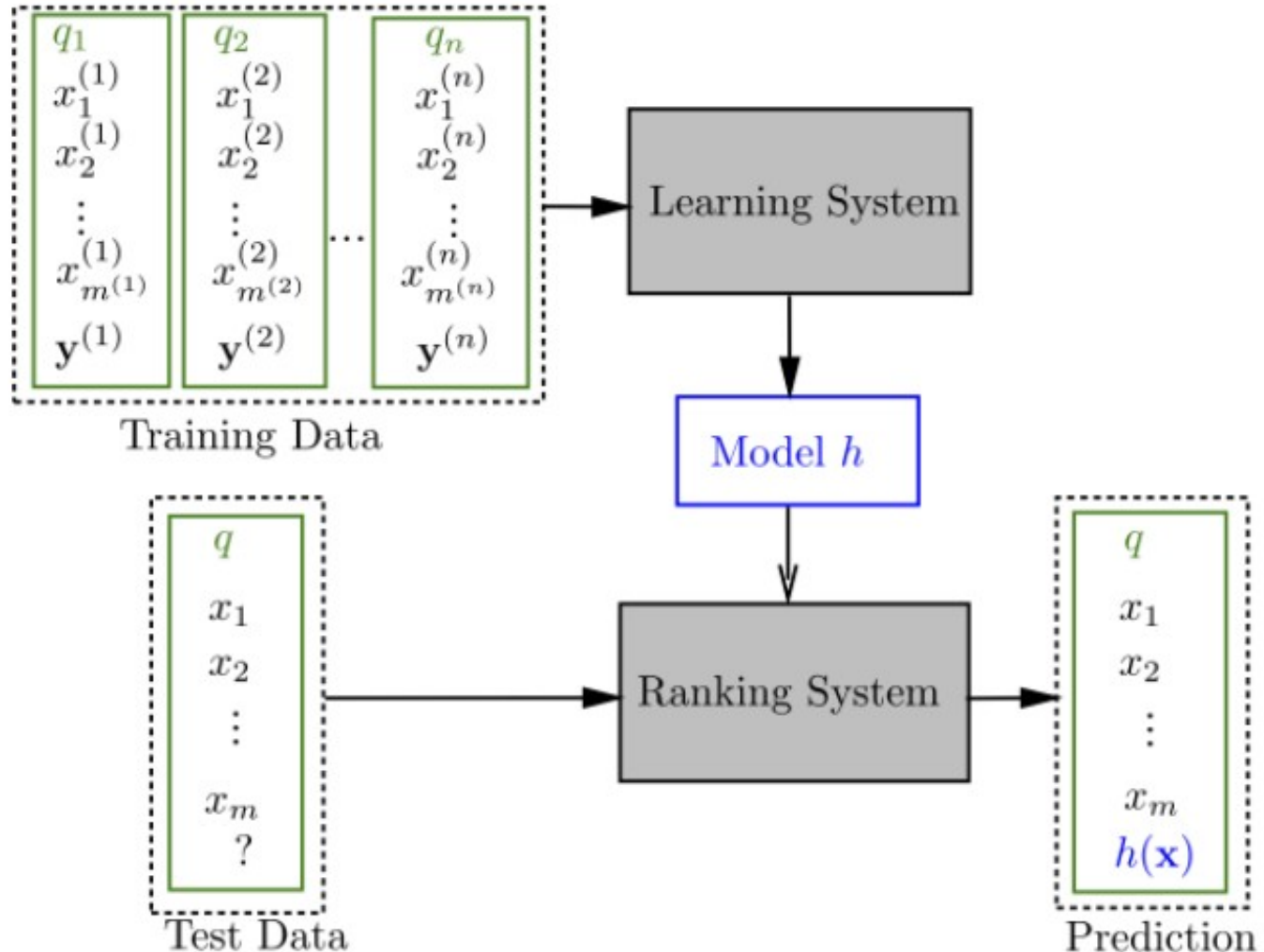
PageRank



Visual Representation of PageRank

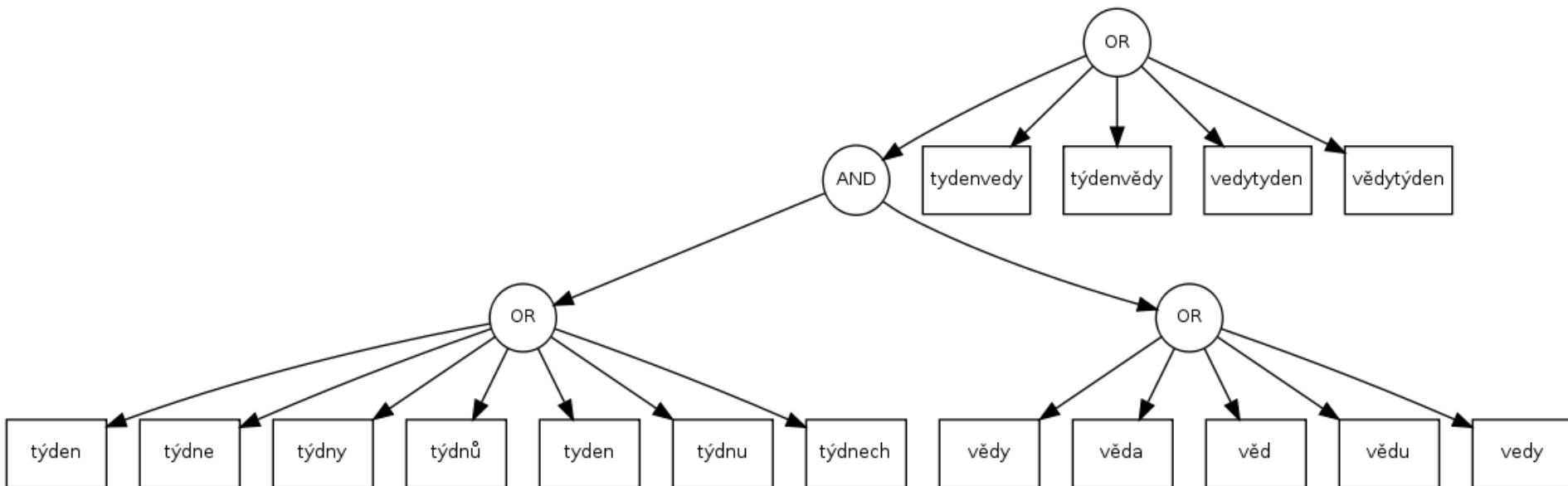
*Source: Wikipedia.

Rankování stránek



Zpracování dotazu

Dotaz: “tyden vedy”



Doplnění diakritiky

cesky jazyk → český jazyk

Doplnění diakritiky

cesky jazyk → český jazyk

zrani → zraní
→ žraní
→ zraní

Lemmatizace

fakulta informatiky → fakulta informatika

Lemmatizace

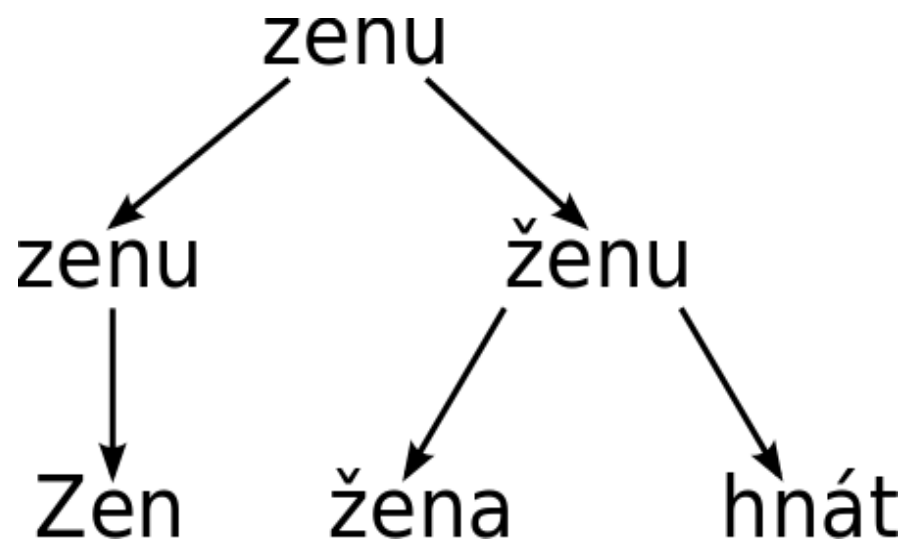
fakulta informatiky → fakulta informatika

tancích

→ tank (kniha o ruských tancích)

→ tanec (kniha o slovanských tancích)

Víceznačnost na více úrovních



Synonymie

Kdy je možná substituce synonym?

**zubař / stomatolog
překládat / tlumočit**

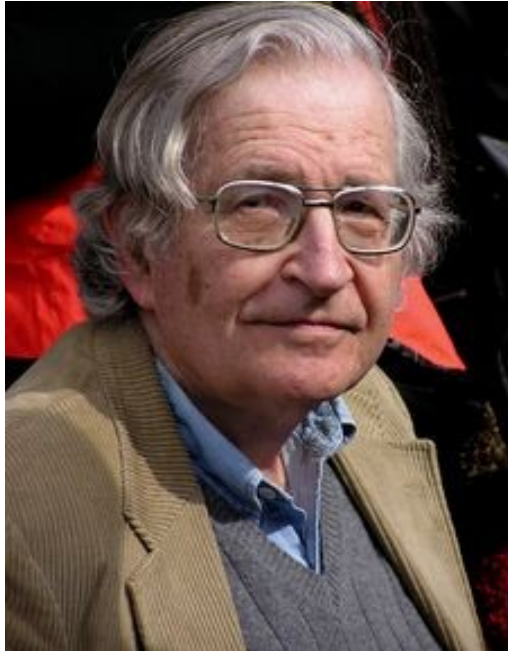
Synonymie

Kdy je možná substituce synonym?

**zubař / stomatolog
překládat / tlumočit**

- (1) Student jazykové školy překládal projev zahraničního hosta.
(2) Tu hromadu písku překládal celý den.*

Složitéý systém vs. Big data



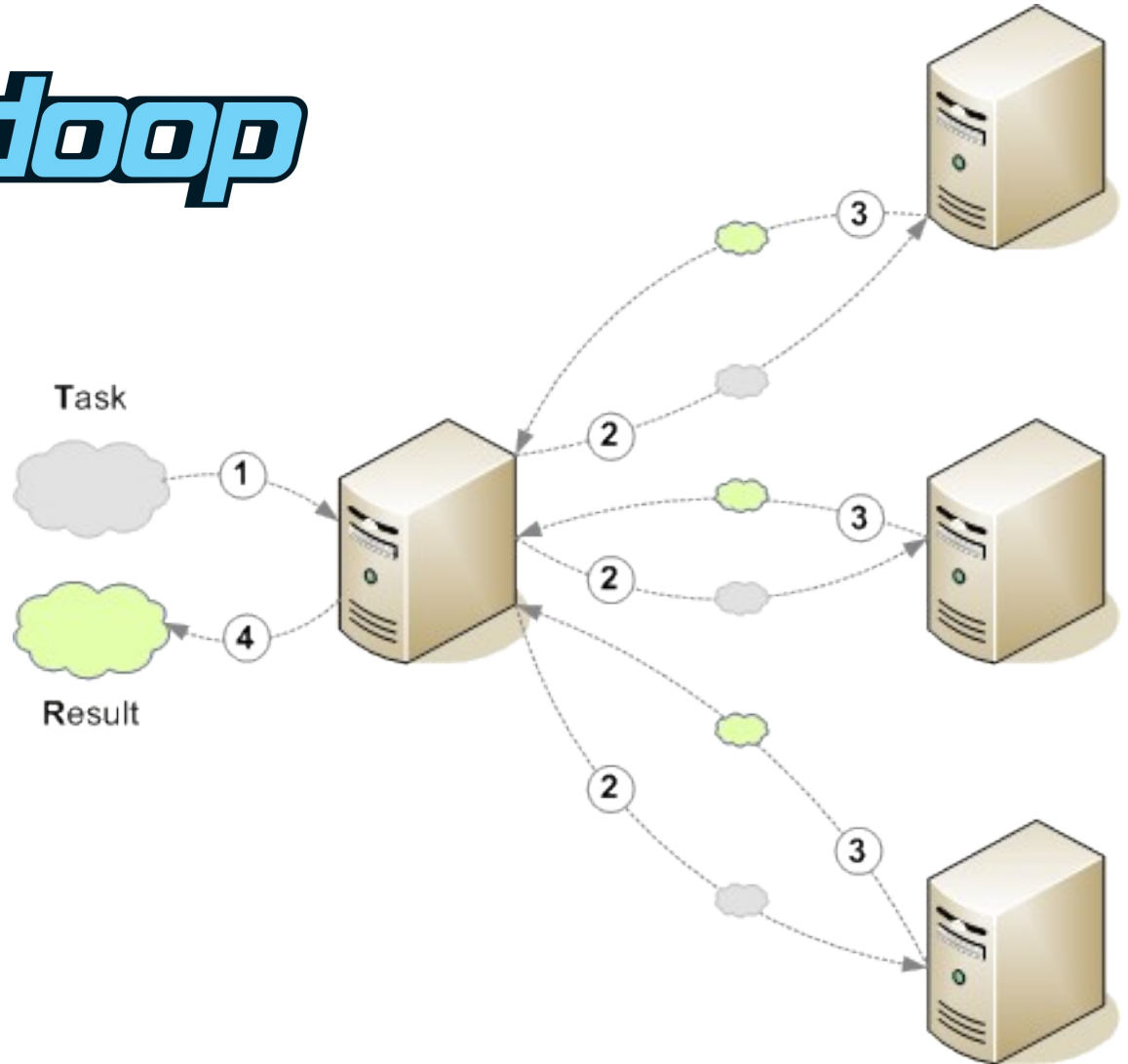
**Noam
Chomsky**

X

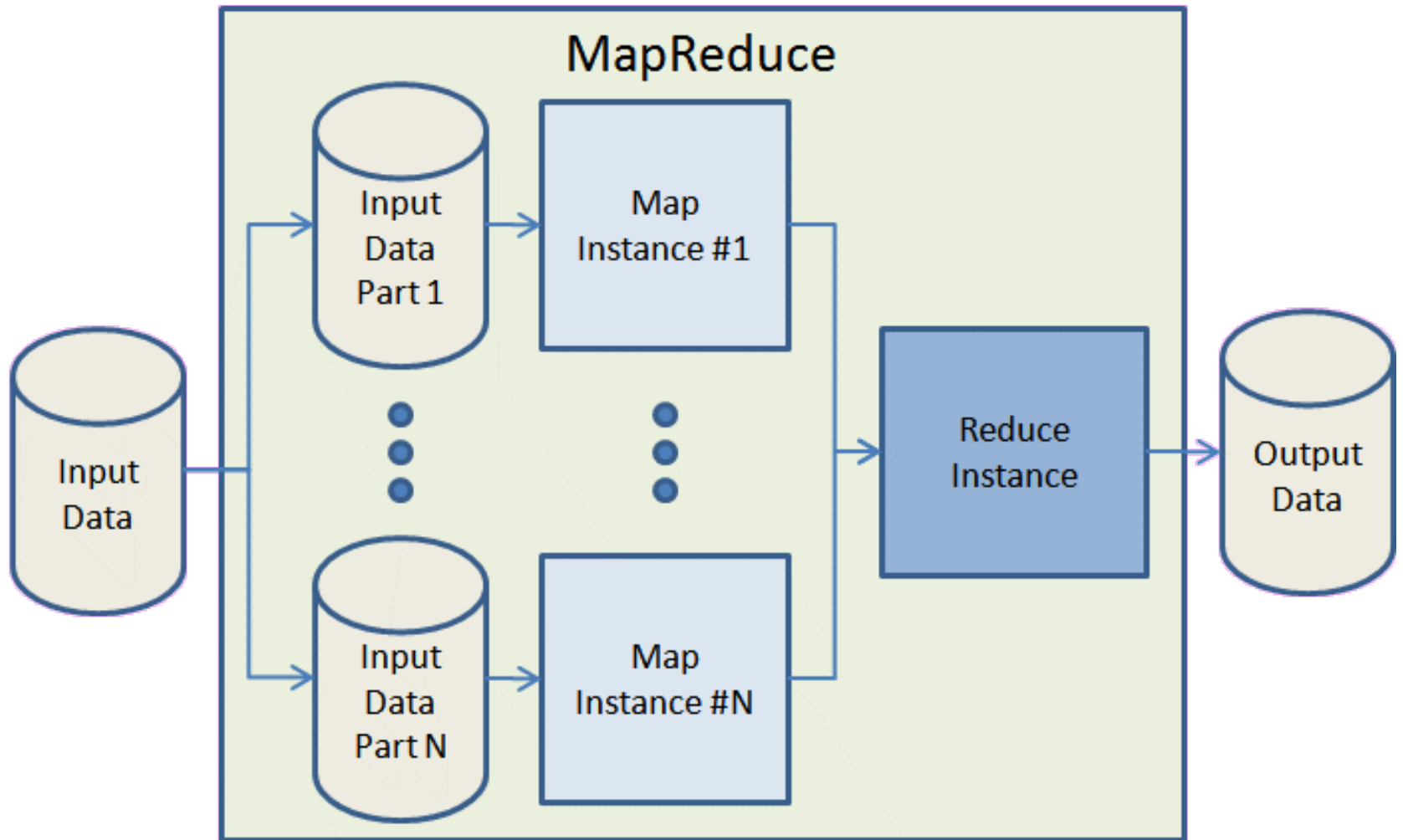


**Peter
Norvig**

Distribuované výpočty



| Map/Reduce



Volitelné téma podrobně

- **Deep learning a neuronové sítě**
- ?
- ?

SEZNAM.CZ
...najdu tam, co neznám!

Jiří Materna, jiri.materna@firma.seznam.cz, [@JiriMaterna](https://www.instagram.com/JiriMaterna)