Strojové učení a umělá inteligence 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

Yandex

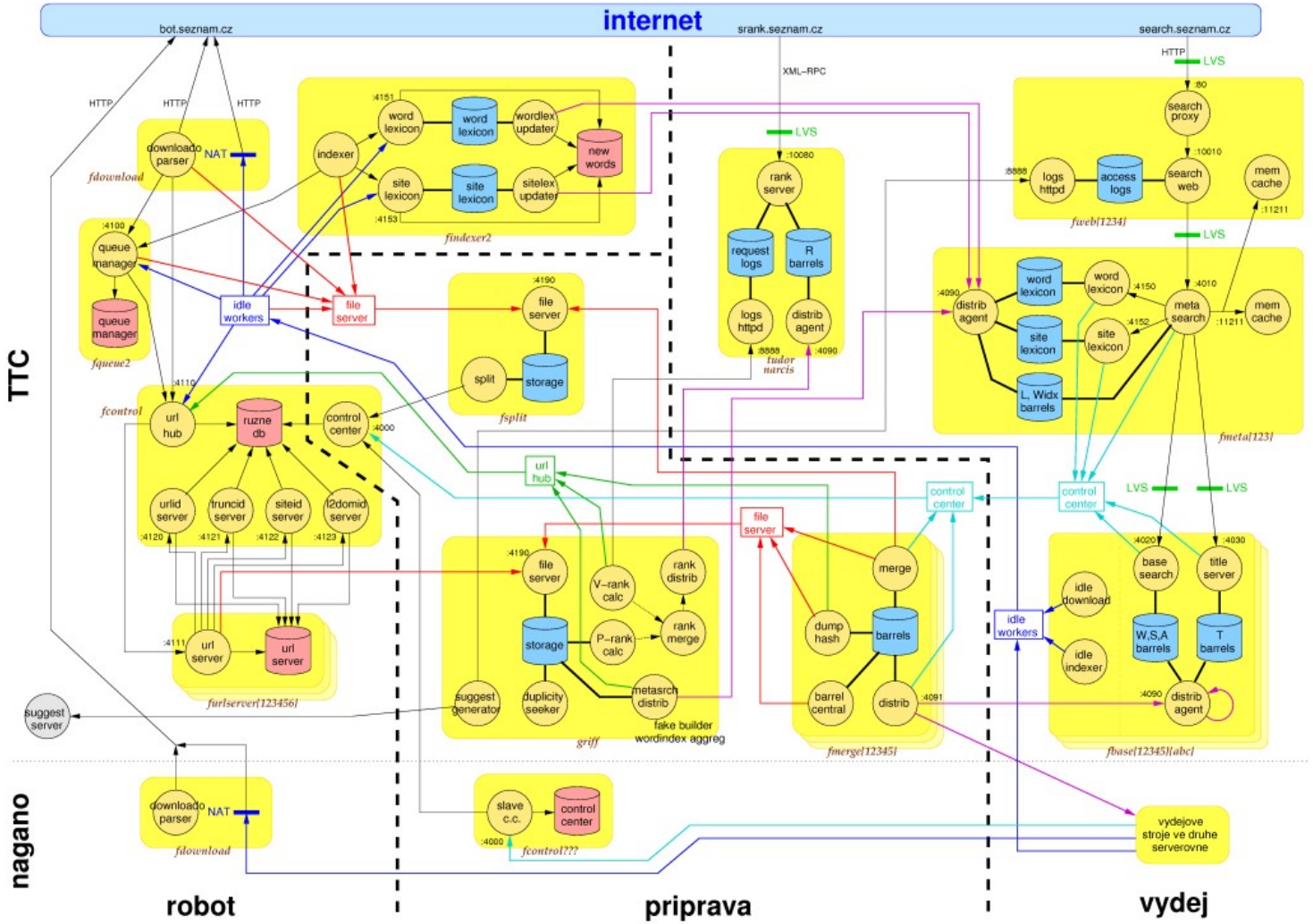
Search.seznam.cz v číslech

- 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

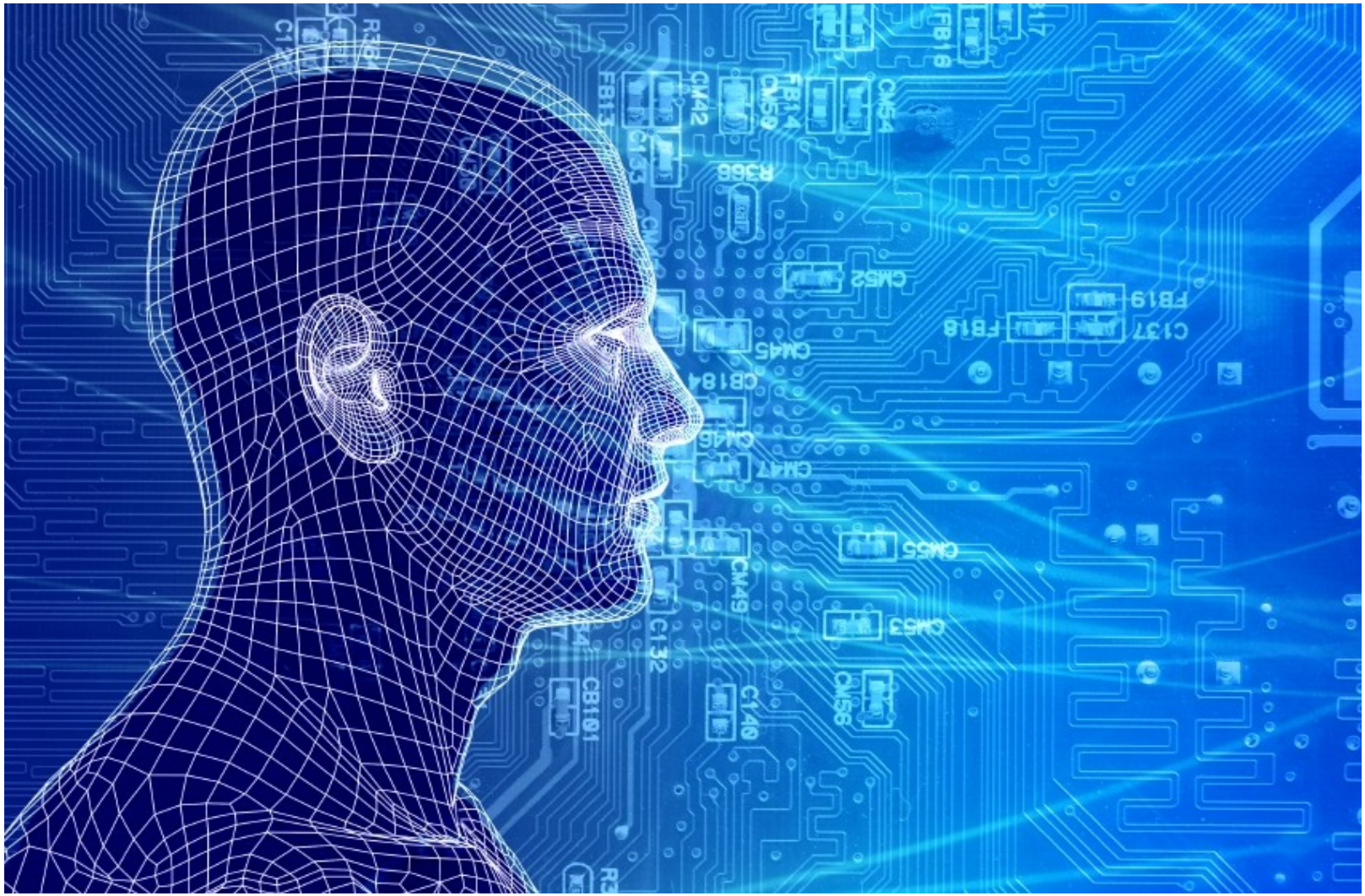
Cizojazyčné vyhledávání

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

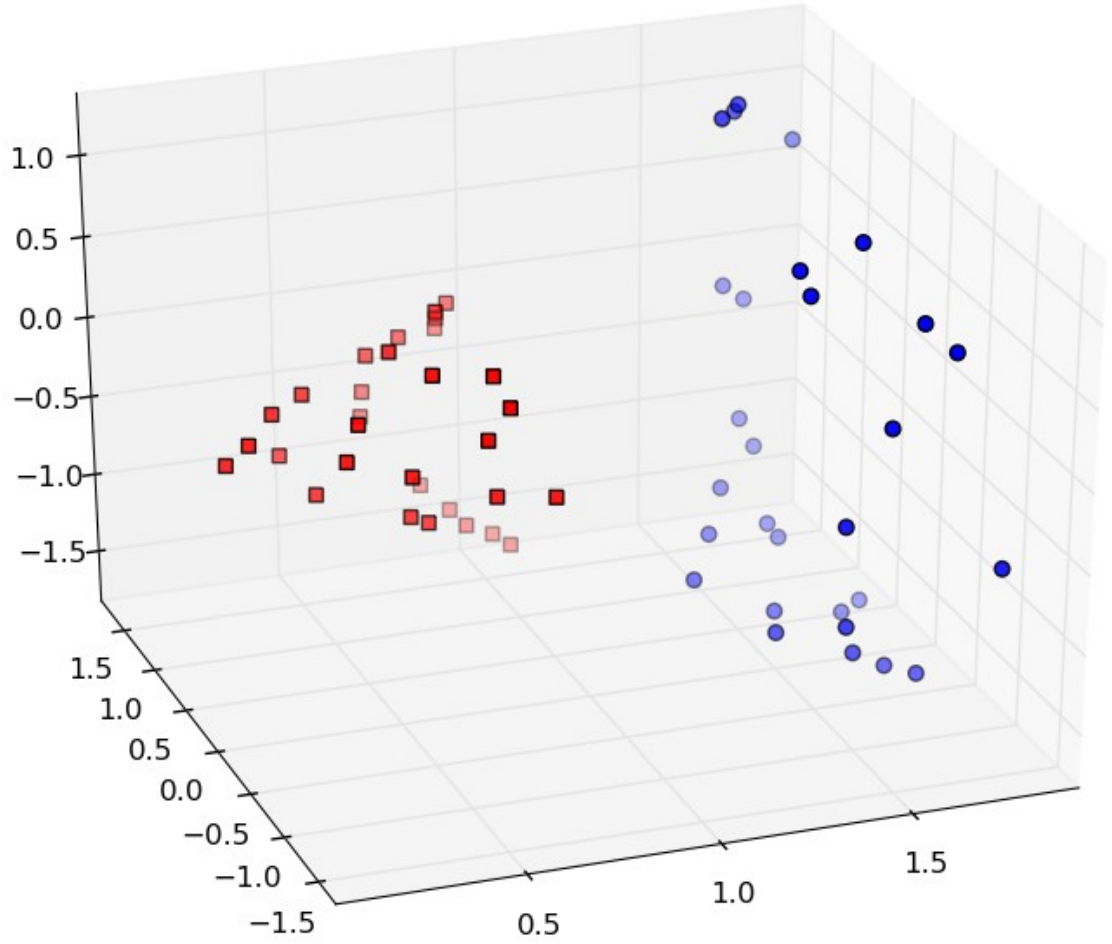
Blokové schéma fulltextu (2010)



Strojové učení



Klasifikační úloha

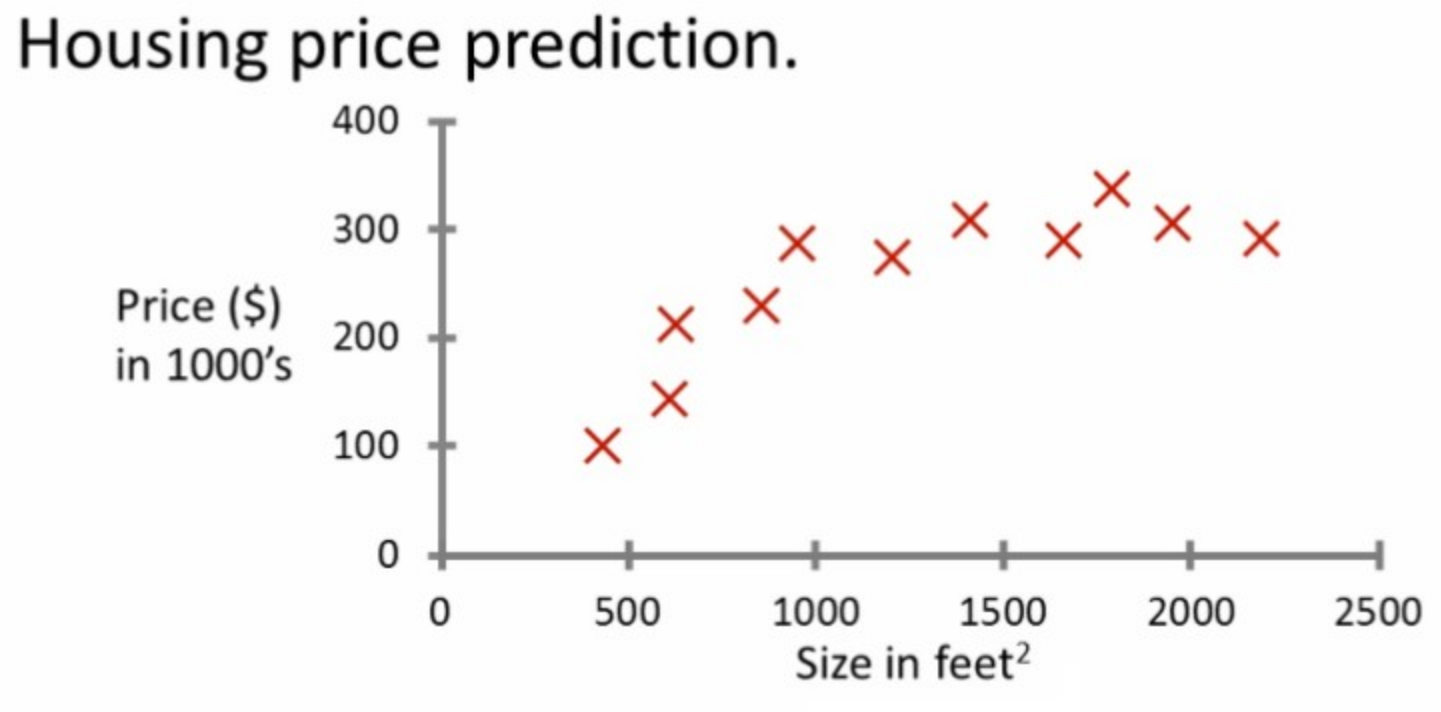


Filtrování nevhodného obsahu

Filtrování nevhodného obsahu

Podíl porna na internetu: 1.5 %

Regresní úloha



Rankování stránek

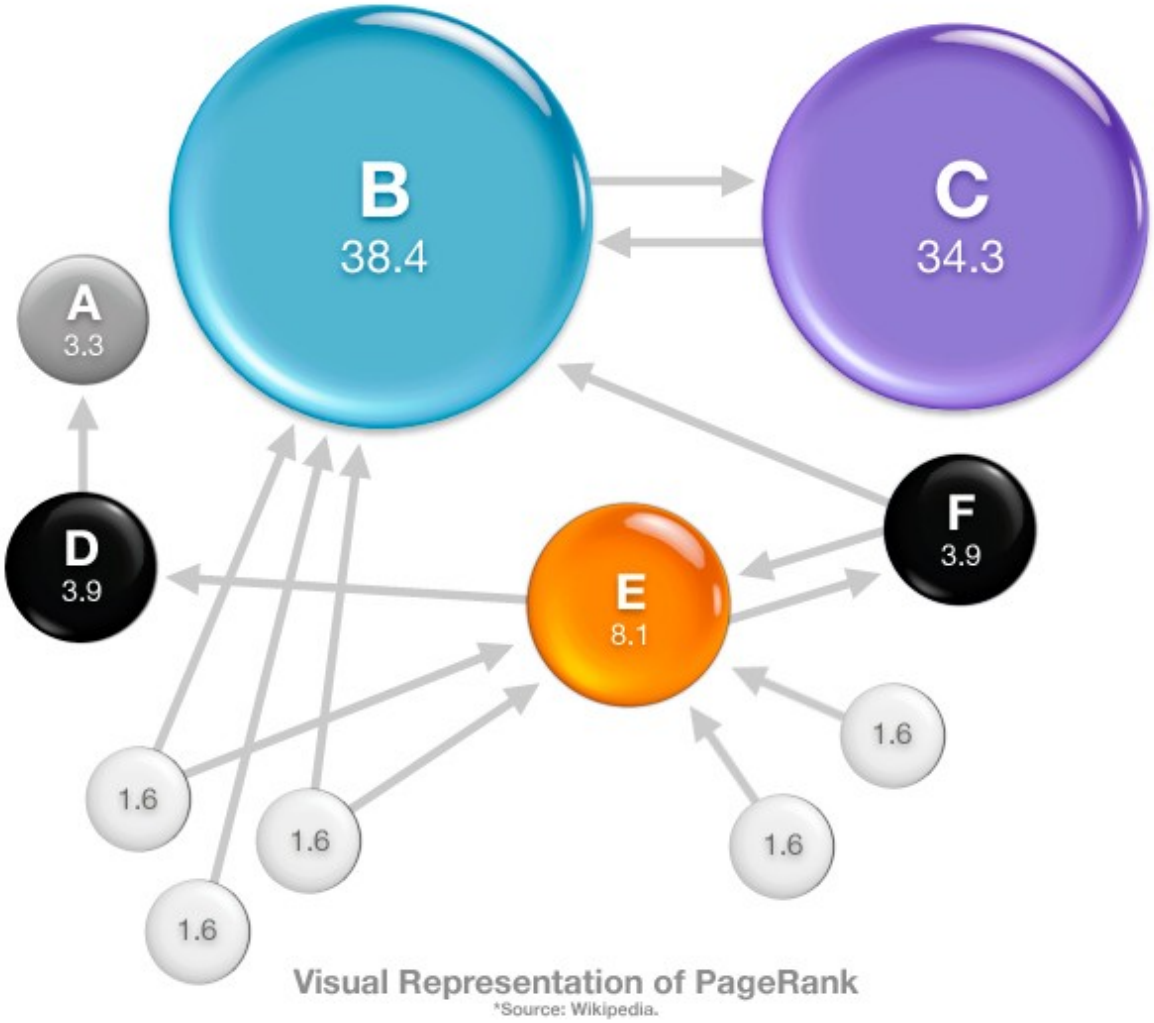
Rank = míra příslušnosti stránky k dotazu

Rankování stránek

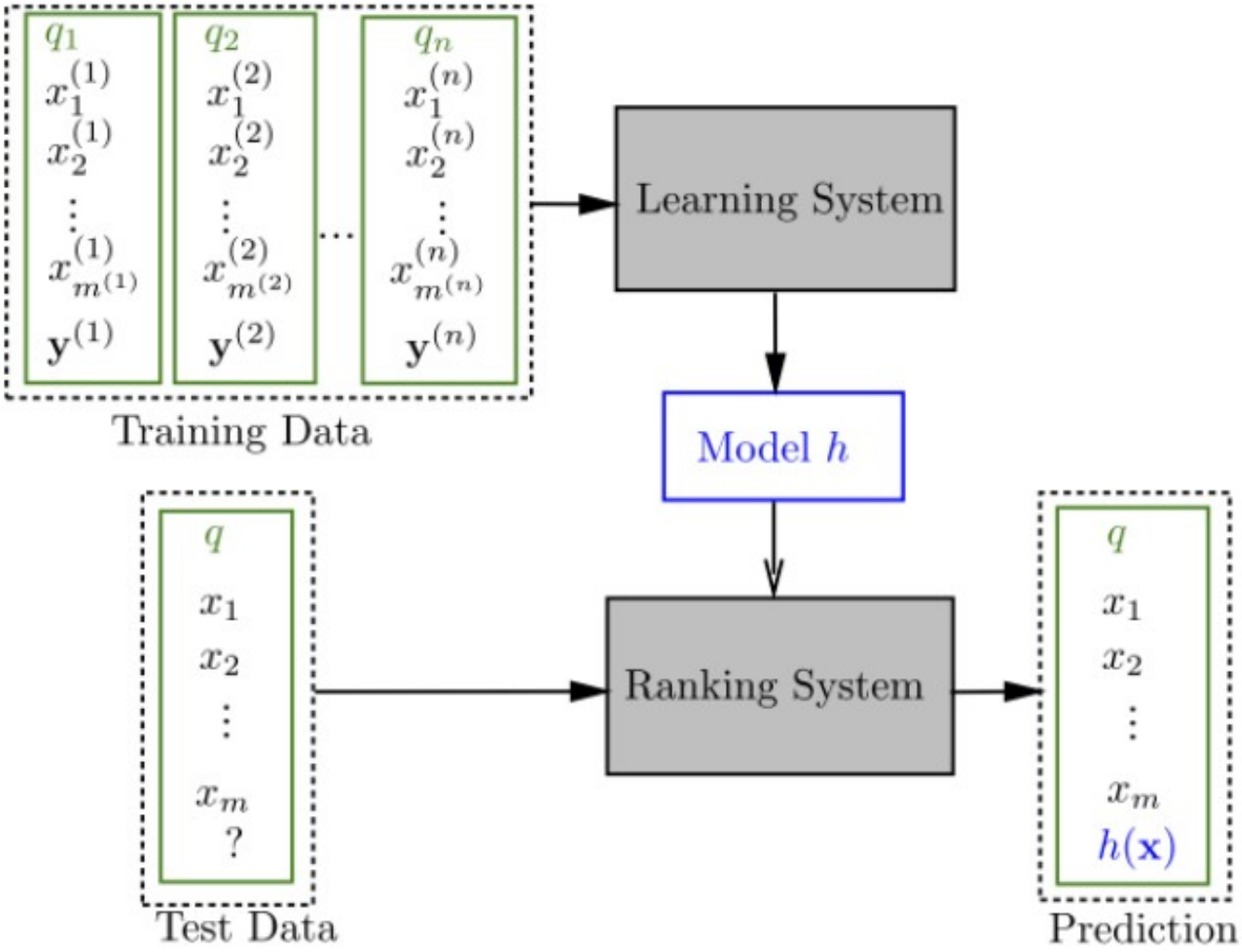
Rank = míra příslušnosti stránky k dotazu

on-page vs. off-page faktory
signály dotazu
singály dokumentu
signály kombinace dotazu a dokumentu

PageRank

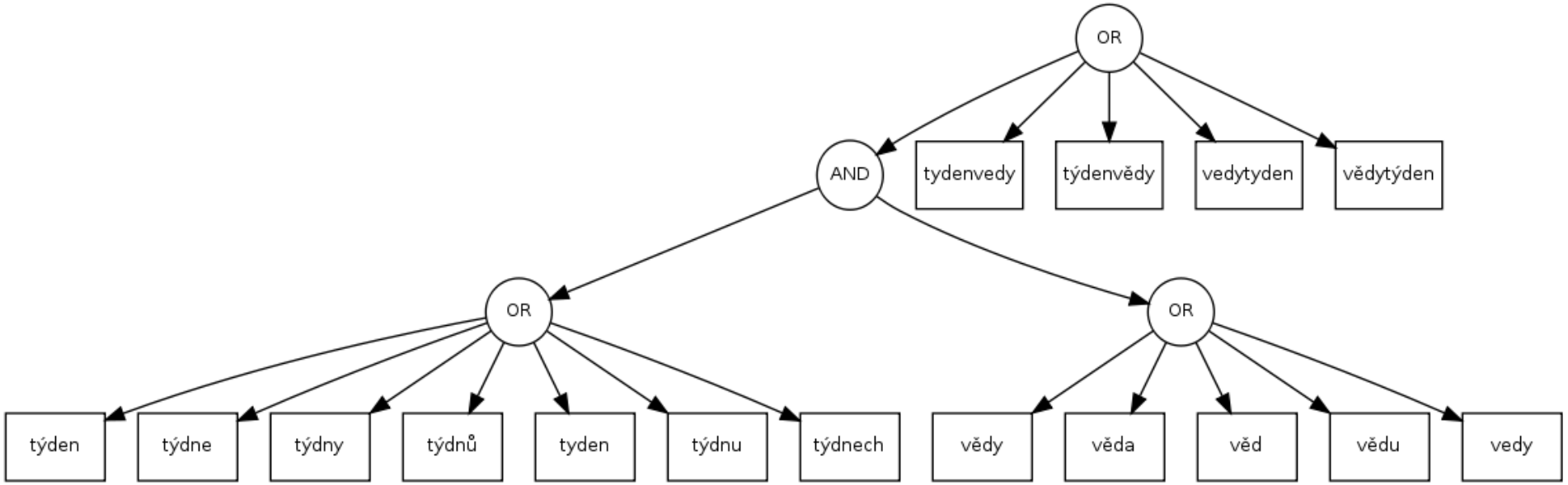

B
38.4
C
34.3
A
3.3
D
3.9
F
3.9
E
8.1
1.6
1.6
1.6
1.6
1.6
Visual Representation of PageRank
*Source: Wikipedia.

Rankování stránek



Zpracování dotazu

Dotaz: “tyden vedy”



Doplnění diakritiky

cesky jazyk –> český jazyk

Doplnění diakritiky

cesky jazyk –> český jazyk

zrani –> zrání
–> žraní
–> zraní

Lemmatizace

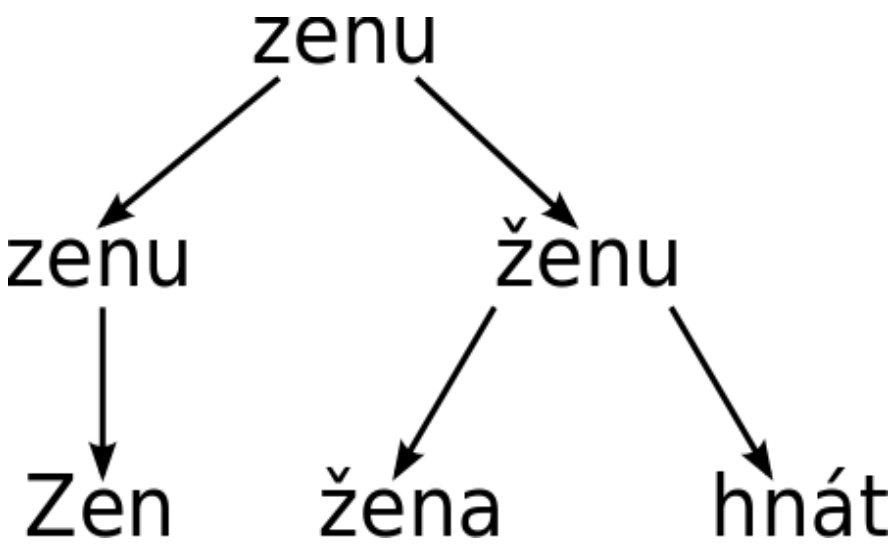
fakulta informatiky –> fakulta informatika

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

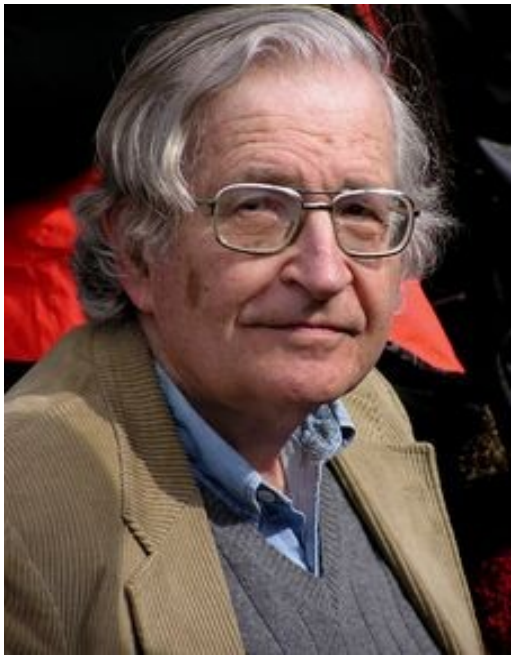
Synonymie

Kdy je možná substituce synonym?

zubař / stomatolog
překládat / tlumočit

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

Složitý systém vs. Big data

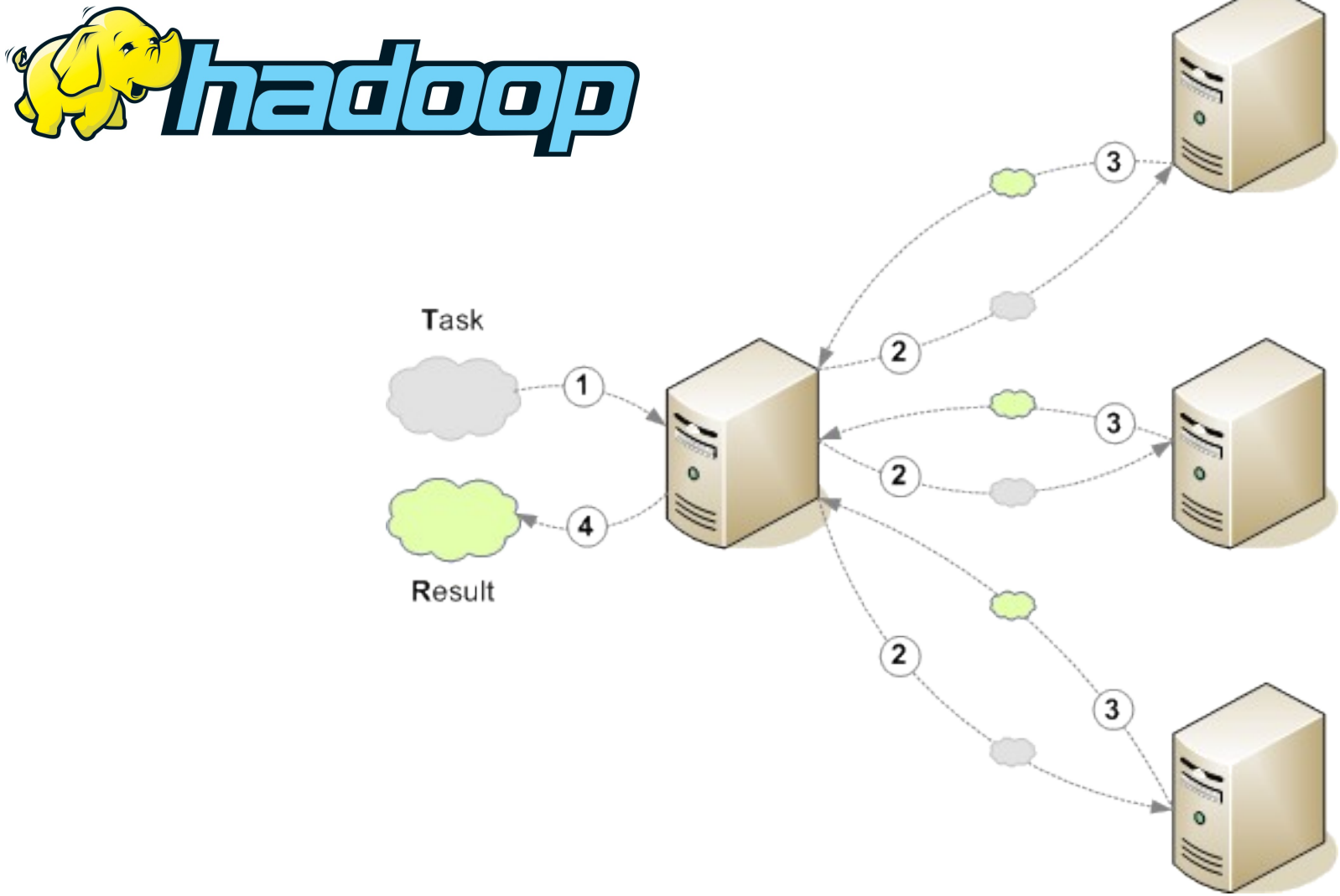


Noam Chomsky

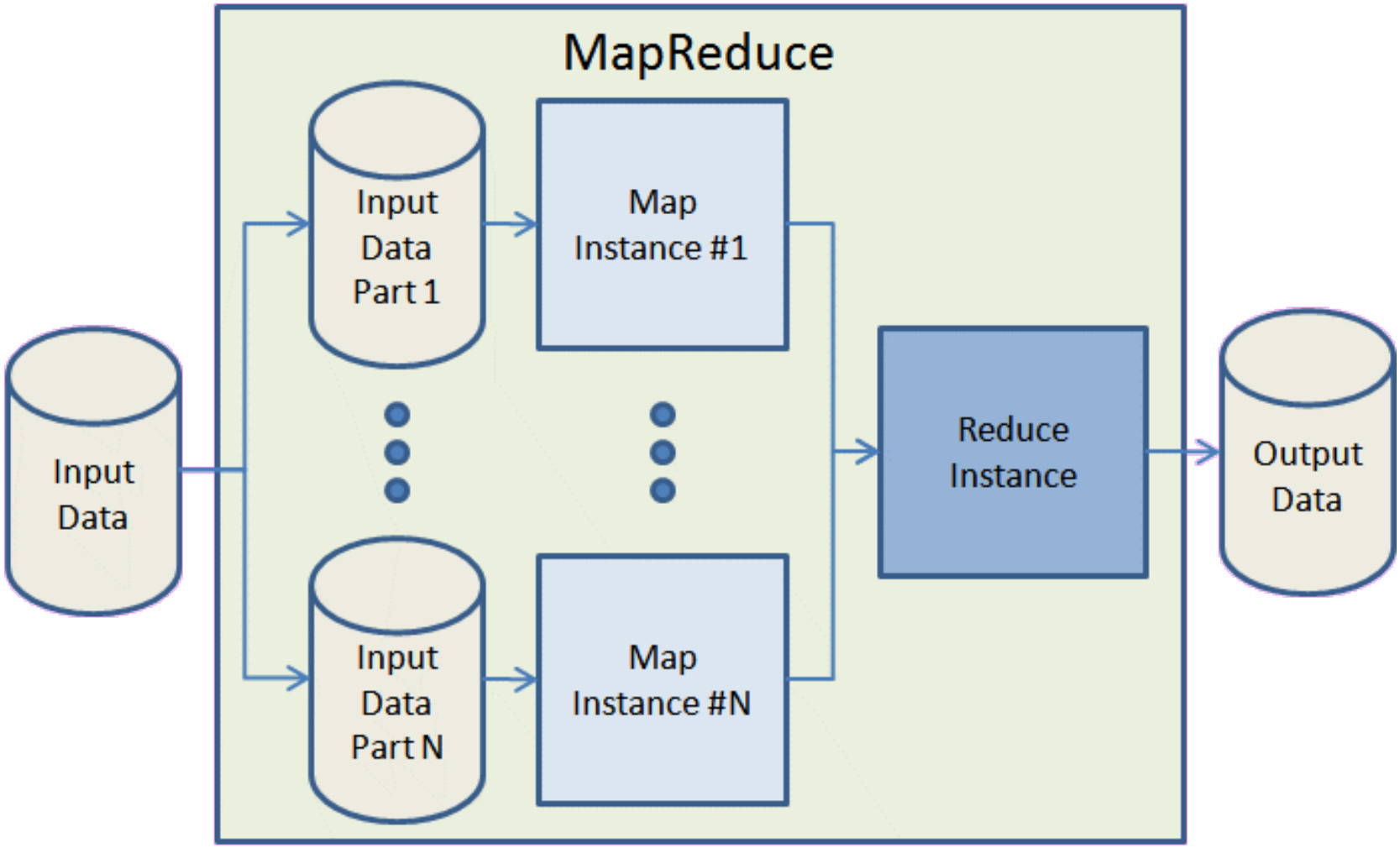
x

Peter Norvig

Distribuované výpočty



Map/Reduce


MapReduce
Input Data
Input Data Part 1
Map Instance #1
Input Data Part N
Map Instance #N
Reduce Instance
Output Data

Volitelné téma podrobně

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

...najdu tam, co neznám!

Jiří Materna, jiri.materna@firma.seznam.cz, @JiriMaterna