

DISA: Data Intensive Systems and Applications

Prezentace laboratoře pro den otevřených dveří na FI MU

RNDr. Vladimír Míč RNDr. Filip Nálepa

Fakulta informatiky Masarykovy univerzity, Brno

13. ledna 2017

- Produkce dat velmi rychle (exponenciálně) roste
- **90 % všech existujících dat** na světě bylo vyprodukováno v posledních 2 letech
- Každý **den** je:
 - posláno **205 miliard e-mailů** (do r. 2019 má být 250 mld., 2/3 spam)
 - na YouTube nahráno **video** v celkové délce **82 let**
 - na Facebook nahráno **350 milionů nových fotek**
 - na Facebooku provedeno **1,8 miliardy sdílení obsahu**
 - na Facebooku posláno **216 milionů zpráv**
 - na Googlu zadáno **3,3 miliardy dotazů**

Vlastnosti dat (1)

Tradiční data

(a) Gigabajty až terabajty

Smith	\$16,753.00 UK	Qtr 3		
Johnson	\$14,808.00 USA	Qtr 4		
Williams	\$10,644.00 UK	Qtr 2		
Jones	\$1,390.00 USA	Qtr 3		
Brown	\$4,865.00 USA	Qtr 4		
Williams	\$12,438.00 UK	Qtr 1		
Johnson	\$9,339.00 UK	Qtr 2		
Smith	\$18,919.00 USA	Qtr 3		
Jones	\$9,213.00 USA	Qtr 4		
Jones	\$7,433.00 UK	Qtr 1		
Brown	\$3,255.00 USA	Qtr 2		

(a) Pevná struktura



(a) Lze jednoduše zjistit souvislosti mezi daty

Big data

(b) Stovky terabajtů, petabajty



(b) Nelze zcela strukturovat



(b) Nelze vždy zjistit všechny souvislosti

Tradiční data



(a) Zpracovávaná obvykle na jednom počítači

(a) Výsledky dotazů lze vyhodnotit přesně

Big data

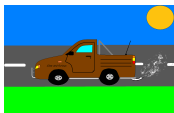


(b) Zpracovávaná více počítači zapojenými do sítě

(b) Dotazy často přesně vyhodnotit nelze (kvůli rychlým změnám, objemu, času, geografické vzdálenosti uložišť)

Nové způsoby zpracovávání dat

- Nové způsoby zpracování dat
 - Např.: detekce krádeží ve videu z bezpečnostní kamery
- Jak je řešit?
 - Lze inspirovat tím nejdokonalejším, co máme k dispozici: **člověkem**
 - Chceme-li počítače naučit něco nového, inspirujme se tím, jak se učí (nenaučený) člověk: **dítě**
- V žurnálu „*Child development*“ z roku 1982 se můžete dočíst: „malé děti používají **podobnost mezi objekty** jako převažující kritérium pro určování objektů (klasifikaci)“



- Naše **laboratoř DISA**: vyhledávání **podobných objektů k zadanému** ve **velkých datech** a jeho aplikace
 - Aplikovatelnost na obecné objekty. V praxi: obrázky (obecné), tváře, pohyby, krátká videa, otisky prstů, grafy, ...

Příklady aplikace podobnostního vyhledávání:

- **Identifikace osob** na základě biometrik (tváře, oči, otisky prstů, styl chůze)
- **Detekce událostí** ve fotkách, videích, audio nahrávkách
- **Organizace dat** (například fotek)

Příklady cílů použití podobnostního vyhledávání:

- **Předejít újmě**: hlídat dodržování práva a dokonce kontrolovat potenciální pachatele!
 - Hlídat dodržování zákazu vstupu fotbalových výtržníků na stadiony
 - Identifikovat teroristy, rozpoznat přípravy k útoku . . .
- **Zpětně postihnout pachatele a zamezit mu opakovat své jednání**
 - Identifikovat, dopadnout a usvědčit jej

V **laboratoři DISA** byly vytvořeny systémy například pro:

- **vyhledávání obrázků** na základě jejich vizuální podobnosti
- **anotování (popis) obrázků**
- vyhledávání podobných **grafů**
- **identifikaci osob** podle jejich tváří a pohybů (kombinace)
- vyhledávání podobných **tváří**
- vyhledávání podobných **pohybů**