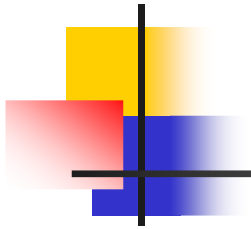
Big Data – a New Challenge for Data Manipulation and Analysis

Petr Berka
University of Economics
and
University of Finance and Administration Prague
berka@vse.cz

Seminar on Big Data, International Workshop on Knowledge Management,
VŠM, Bratislava, 20.10.2016

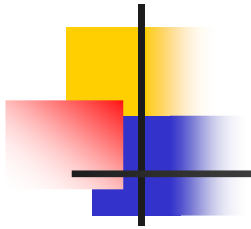


What is Big Data?

Big Data are data that cannot be handled using standard data base systems and standard data analysis tools.

Big Data sources: sensors, surveillance systems, mobile phones, GPS devices, RFID readers, social networks, computer networks, web logs, scientific data …

Big Data characteristics (3 V's or 4 V's)

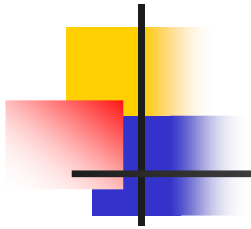


- **Volume:** the size of Big Data goes beyond standard data storage and manipulation techniques
- **Velocity:** Big Data is often available in real time
- **Variety:** Big Data contains not only structured data (e.g. in tabular or relational form) but also texts, images, audio or video
- **Veracity:** the quality and reliability of Big Data can vary

The Big Data pipeline

- Data generation
- Data acquisition:
  - data collection,
  - data transmission,
  - data pre-processing (integration, celansing, rendundancy elimination)
- Data storage
- Data analysis

Challenges for collecting, storing and manipulating Big Data



New forms of data storage: file systems, NoSQL databases

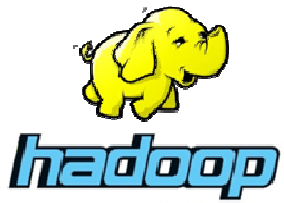
New forms of computation: parallel computing, distributed computing, grid computing, cloud computing

batch processing X stream processing

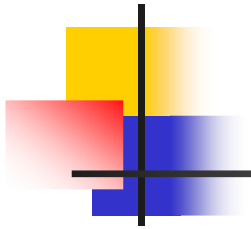
Apache Hadoop

Software platform that supports data-intensive distributed applications

- Hadoop distributed file system
- Map/Reduce: divide and conquer approach to break-down intractable problem into tractable sub-problems

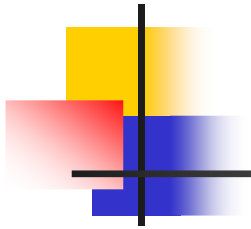


Challenges for analyzing Big Data



- New forms of data: heterogeneous data, unstructured data, stream data
- New properties of data: non-stationarity, concept drift
- New forms of learning: real-time learning, incremental learning, sequentional learning
- New forms of computation: distributed computation, cloud computation

Areas related to Big Data Analysis



Analysis of Big Data grounded in Knowledge Discovery in Databases and Data Mining. However, new names appear used by different people:

- Ubiquitous Knowledge Discovery
- Reality Mining

Ubiquitous Knowledge Discovery

Knowledge discovery process in mobile, distributed, dynamic environments, in presence of massive amounts of data

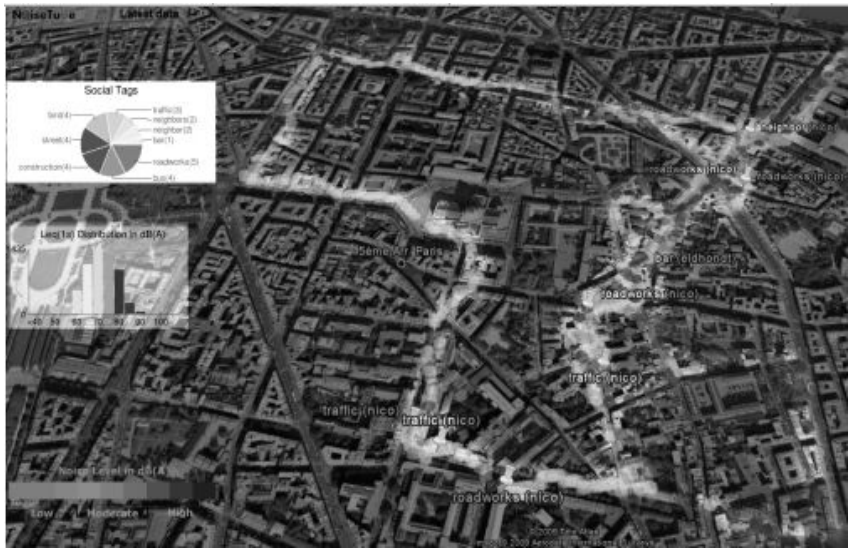
A Blueprint for Ubiquitous Knowledge Discovery Systems

EU funded project KDUbiq (2005-2008 FP6 FET IST)

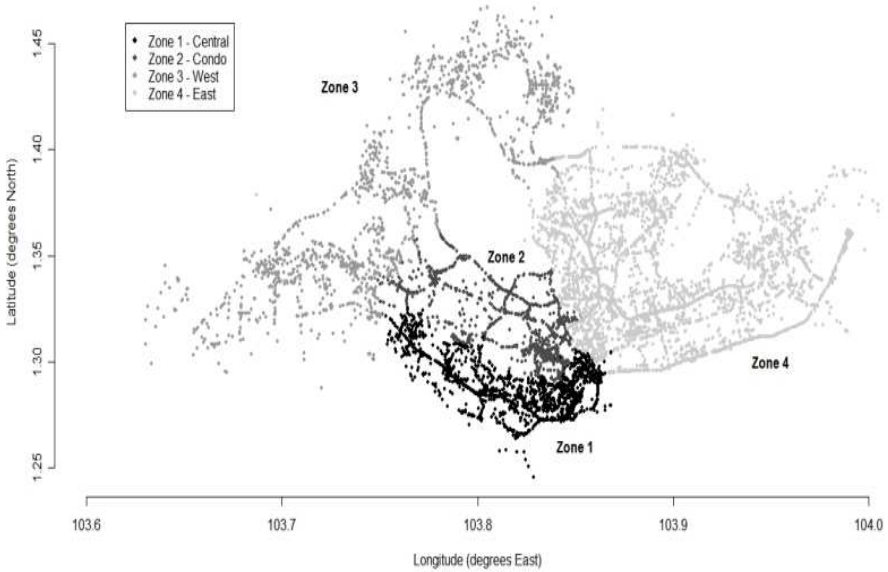
- data mining in mobile systems, wireless communication networks, calm technologies,
- distributed architectures: distributed data mining, grid, P2P, autonomic computing,
- agents,
- learning components: statistical learning (incl. online learning), evolutionary computing,
- anytime algorithms data types: spatio-temporal, stream, multimedia,
- security and privacy: privacy preserving data mining, intrusion detection,
- HCI and cognitive modelling: user interfaces of ubiquitous discovery systems.

Reality Mining

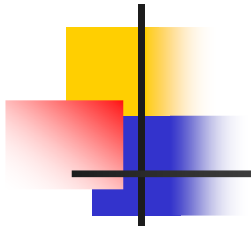
Collection and analysis of machine-sensed environmental data pertaining to human social behavior, with the goal of identifying predictable patterns of behavior. (Pentland 2004)



Collective noise map for part of Paris



Singapore - Taxi Observations by Location and Booking Frequency



Data Science

Set of fundamental principles that support and guide the principled extraction of information and knowledge from data.

Theoretical backgorund for data mining, big data analysis, data-driven decision making e.t.c

Thank You

IWKM2016, VSM Bratislava,
20.10.2016