Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

Knowledge Discovery and Data Mining I

Winter Semester 2018/19

Agenda

1. Introduction
2. Basics
3. Unsupervised Methods
4. Supervised Methods
5. Advanced Topics
   - 5.1 Process Mining
   - 5.2 Outlook

Further Machine Learning Methods

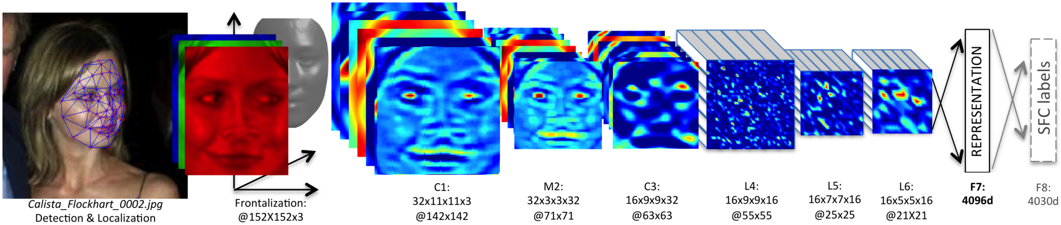


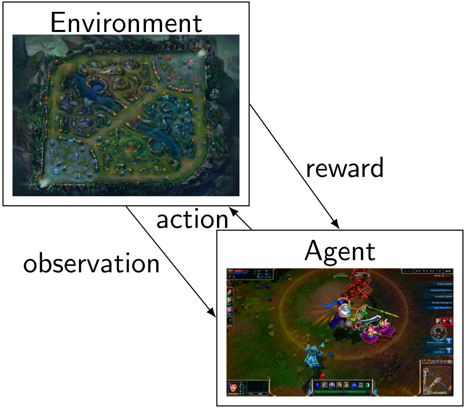
Image Source: Taigman, et al. "Deepface: Closing the gap to human-level performance in face verification." CVPR'14.

- Graphical Models
- Generative Models
- Neural Networks
- Deep Learning

⇝ Machine Learning (SS), Deep Learning and Artificial Intelligence (WS)

Decision Making / Planning

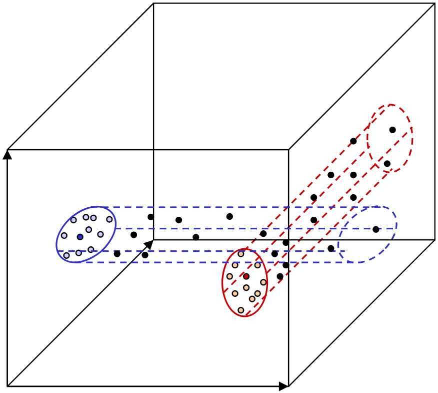
- Setting:
  - Agents are in some environment, observe, and have to take actions that influence the environment.
- Methods:
  - Deterministic/Stochastic Planning
  - A^*-Search
  - Model-Free Reinforcement Learning
  - Q-Learning
  - Adversarial Search (e.g. Alpha-Beta Pruning)



⇝ Deep Learning (WS), Managing Massive Multiplayer Online Games (SS)

High-Dimensional Data

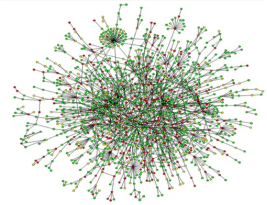
- Challenges:
  - *Curse of dimensionality*: distances become more and more similar
  - Datasets become sparse.
  - Expensive distance measures
  - Degeneration of index structures
  - Unintuitive properties in high dimensions.
- Tasks
  - Feature Selection
  - Feature Reduction / Metric Learning
  - Clustering in High-Dimensional Spaces

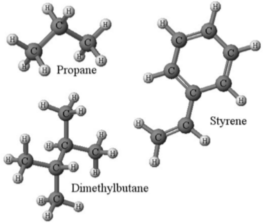


⇝ Knowledge Discovery in Databases II (SS), Big Data Management and Analytics (WS)

Graph Data

- Graphs are everywhere!
  - Chemical data analysis, proteins
  - Biological pathways/networks
  - Program control flow, traffic flow
  - Web graph, social network analysis
- Typical tasks
  - Measure similarity between graphs
  - Find frequent patterns in graphs
  - Generate "realistic" synthetic graphs
  - Identify groups in social networks
  - Integrate additional information

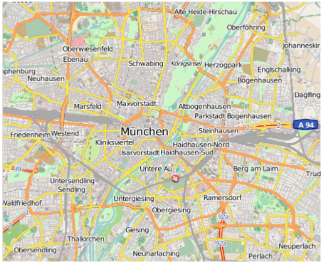


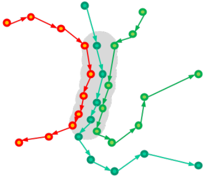


⇝ Knowledge Discovery in Databases II (SS), Big Data Management and Analytics (WS)

Spatial Data

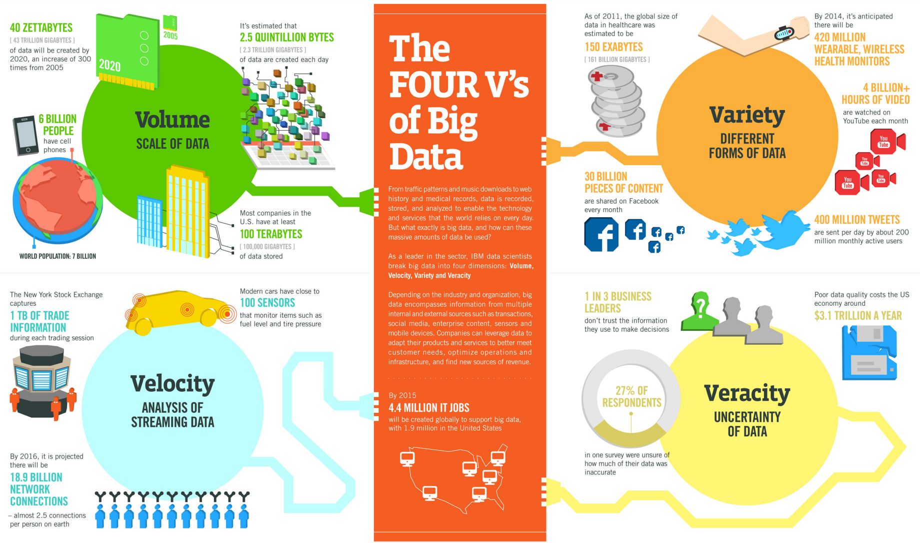
- Mining spatial data
  - Spatial clustering, outlier detection, prediction, rule mining, ...
- Spatial data management
  - Process spatial queries without scanning the whole database
  - Spatial index structures: BSP-tree, R-tree, Quad-tree, ...
- Mining trajectory data
  - Similarity models for trajectories
  - Trajectory compression
  - Mining patterns in trajectories (encounters, flocks, ...)





⇝ Managing Massive Multiplayer Online Games (SS)

Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

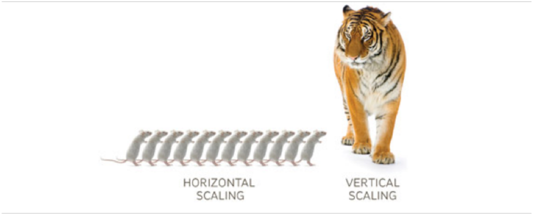
27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

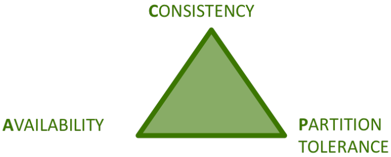
IBM

Big Data Management

- Vertical scaling limited and expensive
  ⇝ Distributed storage
- NoSQL databases
  - Redis
  - MongoDB
  - Cassandra
  - Neo4J
- Distributed file systems
  - GFS (Google)
  - HDFS (Hadoop)
  - S3 (Amazon)



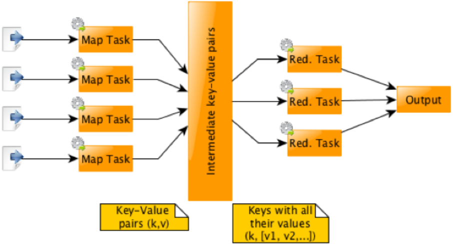
https://www.greentree.com/latest-news/avoiding-cumulus-congestus



⇝ Big Data Management and Analytics (WS)

Distributed Data Processing

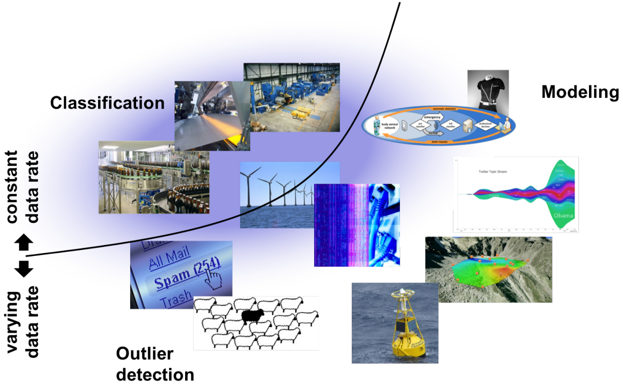
- Processing and analyzing big data
- Map-Reduce: Programming model for distributed processing of large datasets
  - Algorithms are specified as sequences of map and reduce functions
  - Programs are automatically parallelized and executed on a cluster
  - System is tolerant to hardware faults
- Frameworks
  - Apache Spark (batch processing)
  - Apache Flink (stream processing)



⇝ Big Data Management and Analytics (WS)

Stream Data

- Data objects arrive over time in a continuous data stream
- Challenges
  - Infinite stream
  - Limited time and memory
  - Evolving distribution
  - Varying data rates
  - Concept drift
- Typical tasks
  - Sampling and buffering
  - Stream statistics
  - Aging mechanisms



⇝ Knowledge Discovery in Databases II (SS), Big Data Management and Analytics (WS)

Seminars, Practicals, Theses

Dive deeper into specific topics and get hands-on experience:

- Master Seminar "Recent Developments in Data Science" (SS)
- Master Practical "Big Data Science" (SS)
- Master Practical "Applied Reinforcement Learning" (SS)
- Individual Bachelor and Master Theses