Lecture Notes to
Big Data Management and Analytics
Winter Term 2018/2019

© Matthias Schubert 2018
Course Logistics

- **Course website:**

- **Registration for course & exams via:**
  [https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=1011](https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=1011)

- **Organization:**
  - Lecture: Prof. Dr. Matthias Schubert
  - Assisting: Daniyal Kazempour, Evgeniy Faerman

- **Exam:** 02.10.2018 14:00-16:00 in M218/A240 (main building)

<table>
<thead>
<tr>
<th>Component</th>
<th>When</th>
<th>Where</th>
<th>Starts at</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>Tue, 13.00 - 16.00 h</td>
<td>Room S 004 (Schellingstr. 3)</td>
<td>16.10.2018</td>
</tr>
<tr>
<td>Tutorial 1</td>
<td>Wed, 16.00 - 18.00 h</td>
<td>Room D Z007 (HGB)</td>
<td>24.10.2018</td>
</tr>
<tr>
<td>Tutorial 2</td>
<td>Wed, 18.00 - 20.00 h</td>
<td>Room D Z007 (HGB)</td>
<td>24.10.2018</td>
</tr>
<tr>
<td>Tutorial 3</td>
<td>Thu, 16.00 - 18.00 h</td>
<td>Room D Z007 (HGB)</td>
<td>26.10.2018</td>
</tr>
<tr>
<td>Tutorial 4</td>
<td>Thu, 14.00 - 16.00 h</td>
<td>Room D Z007 (HGB)</td>
<td>26.10.2018</td>
</tr>
</tbody>
</table>
What is Data Analytics and AI?

- Foundations of Data Analytics and AI
- Drivers of modern Data Science
- The Knowledge Discovery Process
- Big Data Management
- Typical Data Mining Tasks
- Deep Learning
- Artificial Intelligence and Data Analytics
- Reinforcement Learning
Foundations: Prediction and AI

*How to make decisions?*

- What do you know about the current situation?
- What are your options?
- Which option is the best?
- How many decisions do I have to make until reaching my goal?

*Problems:*

- Parts of your current situation might be unknown or not modeled
- Considering all options is often not possible
- Considering all possible impacts of choosing an option is often not possible.
Foundations: Data Analytics and AI

Uncertain situation:
• Impact of fracking to ground water
• True population of a species

Uncertain impacts:
• What would be the impact of grants for renewables in Alberta?
• What are the long term effects of fracking/oil sand usage?

Considering all options:
• Which kinds of grants and funds should be provided?
• What are the newest technical solutions?
So where does data analytics and AI help?

- Modelling uncertain situations and results (Data Analysis)
  - Predict latent situation parameters
  - Predict uncertain outcomes
- Consider possible long-term impacts of decisions (both)
- Develop strategies for achieving long-term goals (AI)
Foundations: Data Analytics

- Statistics (ca. 1663 /some claim even 5 century B.C.)
- Neural Computing (ca. 1943)
- Artificial Intelligence (ca. 1955)
- Machine Learning (ca. 1959)
- Pattern Recognition (ca. 1990 Begriff 1950)
- Data Mining and Knowledge Discovery (ca. 1996)
Drivers of Modern Data Sciences

- Preconditions to Big Data Analytics and modern AI:
  - Internet and broadband connections: allowed to publish information easily, access information from a huge amount of sources
  - Data Storage: hard drives became larger and cheaper. SSDs make background storage faster. Larger/faster main memory
  - Mobile devices: collect personal and spatial data

http://www.ubergizmo.com/2013/01/china-policy-demands-new-residences-have-fiber-optic-connections/

Drivers of Modern Data Sciences

• Cloud computing: distributed computations on thousands of commodity machines
• Commodity GPUs: dedicated numerical processing power
• Cheaper sensors/camera: affordable monitoring
• IoT and sensors: monitoring installations and environments
• RC and autonomous mobile units: UAVs, rovers,..
Drivers of Modern Data Sciences

- Impacts on data analytics and AI:
  - more data: complex problems become feasible:
    - before: available samples only allowed simple models
    - now: complex models can be trained because sample sets become huge (several millions+)
  - more computational power:
    - before: complex models did not finish training
    - now: models with several thousand parameters on millions of samples are possible
  - scalability:
    - before: predictors where done for dedicated cases
    - now: building personalized models for millions of cases is possible
Summary

- Some applications already worked out fine centuries before.
- A lot of ideas were created in the 1950s with the first computers, but did not work out.
- Recent breakthroughs in classical problems
  - Image processing
  - Speech recognition
  - Automatic translation
  - AI for board games (e.g. AlphaGo)
- New possibilities and tasks due to:
  - more available data
  - complex prediction networks
The Knowledge Discovery Process

- Knowledge Discovery is the technical process of knowledge generation.
- process is iterative: If results are not satisfying, change the process and try again. (change parameters, more data, different data representations, a simpler goal,..)
Data Cleaning and Integration

• …may take 60% of effort
• integration of data from different sources
  – mapping of attribute names (e.g. C_Nr → O_Id)
  – joining different tables
    (e.g. Table1 = [C_Nr, Info1]
     and Table2 = [O_Id, Info2] ⇒
     JoinedTable = [O_Id, Info1, Info2])
• elimination of inconsistencies
• elimination of noise
• computation of Missing Values (if necessary and possible)
• fill in missing values by some strategy (e.g. default value, average value, or application specific computations)
Focusing on Task-Relevant Data

- Find useful features, dimensionality/variable reduction, invariant representation
- creating a target data set
- selections
  - Select the relevant tuples/rows from the database tables (e.g., sales data for the year 2001)
- projections
  - Select the relevant attributes/columns from the database tables (e.g., “id”, “date” “amount” from (Id, name, date, location, amount))
- transformations, e.g.:
  - normalization (e.g., age:[18, 87] → n_age:[0, 100])
  - discretization of numerical attributes (e.g., amount:[0, 100] → d_amount:{low, medium, high})
  - computation of derived tuples/rows and derived attributes
  - aggregation of sets of tuples (e.g., total amount per months)
  - new attributes (e.g., diff = sales current month – sales previous month)
Basic Data Mining Tasks

• searching for patterns of interest
• choosing functions of data mining:
  • Clustering
  • Classification
  • Frequent Patterns
  • Other methods
    • outlier detection
    • sequential patterns
    • trends and analysis of changes
    • methods for special data types, e.g., spatial data mining, web mining
    • ...

• choosing the mining algorithm(s)
Evaluation and Visualization

• pattern evaluation and knowledge presentation:
  Visualization, transformation, removing redundant patterns, etc.

• integration of visualization and data mining
  • data visualization
  • data mining result visualization
  • data mining process visualization
  • Interactive visual data mining

• different types of 2D/3D plots, charts and diagrams are used, e.g.: Box-plots, trees, X-Y-Plots, parallel coordinates

• use of discovered knowledge
Data Management

• **more data** causes **more** handling **problems**:
• data from foreign sources usually has no clear structure (what does a number mean, how is the information related)
  => date exploration to find out what is there?
• data integration data from different sources (integrate once all vs. on demand integration)
• how to structure the data (data variety)
• when is data changed/updated (data volatility)
  • streaming data (data arrives constantly)
  • batch data (data arrives in large bulks)
• selecting and manipulating data should be easy
• data quality must be addressed (missing, synchronization, errors, e.t.c.) (data veracity)
Data Management

handling data volume:

**Small data**: (data fits into the main memory)
- file system: csv-files, excel files, arff
- read everything from file into memory
- manipulate data in memory (e.g. excel, python)

**Medium data**: (data fits on machine but not into memory)
- database systems, files
- read only necessary part of the data (replace data in memory)
- manipulate data on disk (e.g. SQL queries, temporary views)

**Big data**: (data does not fit on one machine)
- NoSQL databases, distributed file systems (e.g. Cassandra, HDFS)
- Manipulate data using cloud framework (e.g. map reduce, Spark)
What else is Big Data?

Business Perspective: A new business model

=> People pay with data

- e.g., Facebook, Google, Twitter:
  - use service => provide data
  - data is used for target advertisement
  - (you pay indirectly)

- e.g., Amazon:
  - pay service + give data
  - sells data and uses data to improve service
Four V’s of Big Data

• **Volume**: integrated data from many sources
  • volume on disk
  • number of instances or features
• **Velocity**: data is changing/new data is arriving
  • sensors constantly produce data
  • communication is constantly going on
• **Variety**: not all data is the same
  • data can have different structures: vectors, sequences, graphs, tensors
  • different sources rely on different formats
• **Veracity**: the meaning of the data is unsecure
  • inputs may be noisy, manipulated or misinterpreted
  • consider data objects as samples not facts
Four V’s of Big Data

Volume
- Scale of Data
- It’s estimated that 2.5 quintillion bytes of data are created each day
- Most companies in the U.S. have at least 100 terabytes of data stored

Velocity
- Analysis of Streaming Data
- By 2020, it is projected there will be 18.9 billion network connections
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure

Variety
- Different Forms of Data
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes
- 4 billion hours of video are watched on YouTube each month
- 30 billion pieces of content are shared on Facebook every month

Veracity
- Uncertainty of Data
- By 2014, it’s anticipated there will be 420 million wearable, wireless health monitors
- 400 million tweets are sent per day by about 200 million monthly active users
- In one survey, 27% of respondents were unsure of how much of their data was inaccurate

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

Big Data Management and Analytics
Literature does not agree upon the # of Vs defining Big Data

Examples:

- **Laney 2001**

  talks about 3 Vs: volume, velocity, and variety

- **later in Van Rijmenam 2014 and Borne 2014**

  it is pointed out that 3Vs are insufficient.

  In addition to volume, velocity, and variety, further 7 Vs are identified: veracity, validity, value, variability, venue, vocabulary, and vagueness
Classification

• Class labels are known for a set of “training data”: Find models/functions/rules (based on attribute values of the training examples) that
  • describe and distinguish classes
  • predict class membership for “new” objects

• Applications
  • image classification
  • document categorization
  • land usage classification from aerial images
Prediction

• numerical output values are known for a small set of “training data”
• find models/functions (based on attribute values of the training examples) that
  • describe the numerical output values of the training data (Major method for prediction is regression)
  • predict the numerical value for “new” objects

• applications
  • build a model of the housing values, which can be used to predict the price for a house in a certain area
  • build a model of an engineering process as a basis to control a technical system
  • ...
Clustering

• class labels are unknown: group objects into sub-groups (clusters)
  • similarity function (or dissimilarity function = distance) to measure similarity between objects
  • objective: “maximize” intra-class similarity and “minimize” interclass similarity

• applications
  • customer profiling/segmentation
  • document or image collections
  • web access patterns
  • ...
Outlier Detection

- find data which are uncommon in the given distribution (e.g. measuring errors, critical system conditions, network intrusion, DNS-Attacks to Servers etc.)
- model what is “normal” to the given data distribution:
  - models should be accurate for common cases
  - models might contain varying levels of assumption (kNN-based vs. Statistical Process)
- everything which isn’t normal w.r.t. to the model is an outlier?
Frequent Itemset Mining

• find frequent patterns in transaction databases
  – Frequently co-occurring items in the set of transactions (frequent itemsets): indicate correlations or causalities

• applications:
  • market-basket analysis
  • cross-marketing
  • catalog design
  • also used as a basis for clustering, classification
  • association rule mining: Determine correlations between different itemsets

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>A,B,C</td>
</tr>
<tr>
<td>1000</td>
<td>A,C</td>
</tr>
<tr>
<td>4000</td>
<td>A,D</td>
</tr>
<tr>
<td>5000</td>
<td>B,E,F</td>
</tr>
</tbody>
</table>

Examples:

buys(x, “diapers”) $\rightarrow$ buys(x, “beers”) [support: 0.5%, confidence: 60%]
major(x, “CS”) $\land$ takes(x, “DB”) $\rightarrow$ grade(x, “A”) [support: 1%, confidence: 75%]
other types of Analysis

• Trends and Evolution Analysis
• Sequential Patterns (find re-occurring sequences of events)
• Spatial Data Mining
  • spatial outlier prediction and clustering
  • spatial prediction
  • trajectory analysis
• Graph Mining:
  • link prediction
  • community detection
  • network centrality
• methods for special data types, and applications e.g.,
  • Natural Language Processing
  • Web Mining
  • Bio-KDD
  • ...
Deep Learning

• often a KDD Process involves several transformation and learning task
• combining multiple learners increases the quality

⇒ Deep Architectures
• integrate data transformation and model training (input raw data -> output target variables)
• joint optimization (instead of training each step separately)
Deep Learning

- paradigms for modelling the connection between raw data to abstract results:
  - artificial neural networks:
    - connect multiple functions $f_n(f_{n-1}(...(f_1(x)...) = y$
      (each output is the input of the next step)
    - training by minimizing a loss function $L(f_n...(f_1(x)....), y)$
    - optimization is done by gradient descent
  - statistical graphical models
    - generative Bayesian models
    - compute the posterior $p(y|x, \theta)$
    - training by Gibbs Sampling,
example: Image Recognition

- **Conventional Imaging**: Imaging Pipeline handcrafted to a the problem (develop function and chain them)

- **Current Development**: Use Convolutional and Deep Neural Networks on the Raw Pixel data

- Strong performance increase in object recognition

- **Applications**:
  - search engines and data management
  - autonomous driving and robotics
  - remote sensing
  - surveillance tasks

- Works on excessive amount of data and usually requires a lot of Hardware (e.g. GPU computers) for training
Convolutional NN for Image Recognition

Gabor filter: linear filters used for edge detection with similar orientation representations to the human visual system

Layer 1
Filter (Gabor and color blobs)

Layer 2

Layer 5

Zeiler et al.
arXiv 2013, ECCV 2014

Last Layer

Nguyen et al. arXiv 2014

slide credit Jason Yosinski
LeNet5 (Winner ImageNet competition)
Evolution of Performance

PASCAL VOC-2007
other directions in Deep Neural Networks

- Recurrent Neural Networks: e.g. long short-term memory
  - models long term dependencies in time series
  - used in speech, text and signal processing
    -(e.g. automatic translation and chat bots)
- Autoencoders: learn compact representations
- Generative Adversarial Networks (GANs): build data generator for based on observed examples
- Deep Dreams: visualize intermediate results to make image detection better understandable
- ...
example: Generative Adversarial Networks

Example: Image Fusion

Artificial Intelligence and Data Analytics

• AI is an extremely broad subject within CS:
  • **tasks**: reasoning, problem solving, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation, social intelligence, creativity, general intelligence
  ⇒ some major overlap to machine learning and data analytics

• for this talk, I will focus on the following aspects:
  • **analytics**: predict unknown values and abstract from given data (What will happen?)
  • **artificial intelligence**: (here: strong focus on planning) find the best strategy to optimize a goal (What should I do?)
The data pyramid

• raw data is often big
• in selection and preprocessing data shrinks
• for complex tasks high-quality data is often still small (e.g. not enough labels, noise, irrelevant, too high resolution)

⇒ Big Data systems often found in the first steps of the KDD process where scalability and efficiency play a role
The Lambda Architecture

- never change/delete data, store original and transformed data
- distinguish between speed and batch layer
  - speed layer: indexes batch view for interactive access
  - batch layer: breaks down all data to batch views
  - serving layer: high frequency update/latest data
- any query can be answered by combination service and speed layer
Course Contents

- Data Science: The Big Picture
- NoSQL Systems
- Hadoop / HDFS / MapReduce
- Apache Spark
- Data Streams & Streaming Methods
- Apache Flink
- Stream Analytics
- Text Data
- High-Dimensional Data
- Graph Data
This course is mainly based on a mixture of existing external lectures, Surveys, Papers and Reports on Big Data

There is NO, or better, I’m not aware of a single book or script that is equivalent to this course (and addresses all issues discussed in this course)

Since Big Data is a quite new and hot topic, standards and basic concepts are quite dynamic ⇒ The Web is a very appropriate source of relevant information

External lectures basically used for this course:

- Big Data: Donald Kossmann & Nesime Tatbul, Systems Group ETH Zurich - [http://www.systems.ethz.ch/node/217](http://www.systems.ethz.ch/node/217)

Further material will appear at our web page
(check for updates during the course / open to further suggestions!)