Operating Complex IT-Systems Seminar – Topics Winter 2017/2018

This document lists topics for this term's bachelor and master seminars proposed by our research staff . Each student is asked to choose at least three topics from the list. Each topic includes an abstract, references, and a recommendation regarding the applicability for bachelor or master students (or possibly both). The references are to be used as a foundation for the literature review, which should to be done before and while preparing the slides and the scientific report. Note that the provided references are probably not sufficient for the required understanding of a topic. The references should thus be understood as a starting point for reviewing the available literature on the topic. If you have trouble accessing the references (e.g. through IEEE Xplore) use a workstation that is either physically or virtually (e.g. by VPN) connected to the university network. You may additionally try to find free-access versions of references using Google's scholar search engine.

Further notes:

- Please make sure not to miss the servicetalk on scientific writing and presentation on (see website for details)
- We will discuss the assignment of topics during the kickoff meetings. Participation is mandatory. Students not showing up without excusing themselves before the kick-off automatically cancel their attempted registration.
- Please select at least three topics of interest. We will try to resolve conflicts, but we cannot guarantee that each student will be assigned his favorite topic.

1 Topics sorted by area of research

@CIT members: please use the following template when adding new topics

Topic :	
Abstract:	
References:	
[1]	
Applicable for BSc:	**Applicable for MSc:**
Further Notes:	

1 Distributed Systems Theory

Topic 1. 1: Virtual Time	
Abstract:	
Virtual time is a new paradigm for organizing and synchronizing distributed systems which can be applied to such problems as distributed discrete event simulation and distributed database concurrency control. Virtual time provides a flexible abstraction of real time in much the same way that virtual memory provides an abstraction of real memory. It is implemented using the Time Warp mechanism, a synchronization protocol distinguished by its reliance on lookahead-rollback, and by its implementation of rollback via antimessages.	
References:	
http://dl.acm.org/citation.cfm?id=3988	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 1.2: Optimistic Concurrency Control	
Abstract:	
Most approaches to concurrency control in database systems rely on locking of data objects as a control mechanism. In this paper, two families of nonlocking concurrency controls are presented. The methods used are “optimistic” in the sense that they rely mainly on transaction backup as a control mechanism, “hoping” that conflicts between transactions will not occur. Applications for which these methods should be more efficient than locking are discussed.	
References:	
[1] http://dl.acm.org/citation.cfm?id=319567 [2] http://redis.io/topics/transactions	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 1.3: Software Transactional Memory	
Abstract:	
In computer science, software transactional memory (STM) is a concurrency control mechanism analogous to database transactions for controlling access to shared memory in concurrent computing. It is an alternative to lock-based synchronization. STM is strategy implemented in software, rather than as a hardware component. A transaction in this context occurs when a piece of code executes a series of reads and writes to shared memory. These reads and writes logically occur at a single instant in time; intermediate states are not visible to other (successful) transactions.	
References:	
[1] http://groups.csail.mit.edu/tds/papers/Shavit/ShavitTouitou.pdf [2] http://dl.acm.org/citation.cfm?id=872048 [3] http://blog.enfranchisedmind.com/2009/01/the-problem-with-stm-your-languages-still-suck/	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 1. 4: Differential Synchronization	
Abstract:	

Differential Synchronization (DS) method is for keeping documents synchronized. The key feature of DS is that it is simple and well suited for use in both novel and existing state-based applications without requiring application redesign. DS uses deltas to make efficient use of bandwidth, and is fault-tolerant, allowing copies to converge in spite of occasional errors. We consider practical implementation of DS and describe some techniques to improve its performance in a browser environment.

References:

[1] https://neil.fraser.name/writing/sync/eng047-fraser.pdf
[2] https://spring.io/blog/2014/10/22/introducing-spring-sync

Applicable for BSc: yes	**Applicable for** MSc: yes

Further Notes: no

Topic 1.5: Geo Replication

Abstract:

Geo-replication systems are designed to improve the distribution of data across geographically distributed data networks. This is intended to improve the response time for applications such as web portals. Geo-replication can be achieved using software, hardware or a combination of the two. Online services distribute and replicate state across geographically diverse data centers and direct user requests to the closest or least loaded site. While effectively ensuring low latency responses, this approach is at odds with maintaining cross-site consistency.

References:

[1] https://www.usenix.org/system/files/conference/osdi12/osdi12-final-162.pdf
[2] https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-ardekani.pdf

Applicable for BSc: yes	**Applicable for** MSc: yes

Further Notes: no

Topic 1.6: CoRAL – Reliable Web Services

Abstract:

Making stateful web services reliable requires elaborate cross-layer techniques. The fault tolerance scheme CoRAL (Connection Replication and Application-level Logging) actively replicates the state of a TCP connection and additionally logs HTTP requests/replies to enable fast failover to a warm-standby server.

References:

http://web.cs.ucla.edu/csd/research/labs/csl/projects/coral/
http://web.cs.ucla.edu/~tamir/papers/pdcs03.pdf
http://millennium.cs.ucla.edu/~tamir/papers/coral_jss09.pdf

Applicable for BSc: yes	**Applicable for** MSc: yes

Further Notes: no

Topic 1.7: Distributed State Machines

Abstract:

Maintaining consistent application state is an important issue when implementing replicated network services. Paxos is a widely used algorithm for implementing a Distributed State Machine which allows a number of service replicas to maintain consistency. Paxos has been extended and improved many times since Lamports original description of the algorithm.

References:

Paxos
http://dl.acm.org/citation.cfm?doid=279227.279229
http://www.ux.uis.no/~meling/papers/2013-paxostutorial-opodis.pdf

NetPaxos

NetPaxos is an extension to Paxos optimizing it for the use in modern SDN-capable switches.

http://perso.uclouvain.be/marco.canini/papers/netpaxos.sosr15.pdf

Raft

Raft is a novel consensus algorithm inspired by Paxos designed to be more understandable for students of dependable systems while providing the same consistency guarantees and performance as Paxos.

http://www.eecs.harvard.edu/cs261/papers/ongaro14.pdf

Applicable for BSc: no | **Applicable for** MSc: yes

Further Notes: Besides the original Paxos algorithm, this topic contains 2 sub-topics dealing with extended approaches. Students can decide whether to present one topic in detail or focus on a comparison. Furthermore, this topic may be assigned to up to three students as well.

Topic 1.8: Partial Synchronous Distributed Systems

Abstract:

System models are an important tool to understand the properties and the behavior a Distributed System and its communicating nodes exhibit. Commonly known models are the synchronous and the asynchronous model. Efficient algorithms for most of the problems we face in Distributed Systems are known for the synchronous model. However, it is almost impossible to implement the synchronous model in real life. The asynchronous model instead, rather characterizes real world systems like the Internet. Unfortunately, the assumptions of this model are too weak to solve several important problems efficiently. As a remedy, the partial synchronous model is introduced.

References:

http://www-usr.inf.ufsm.br/~ceretta/papers/MITLCSTM270.pdf

https://ecommons.cornell.edu/bitstream/handle/1813/7192/95-1535.pdf?sequence=1

Applicable for BSc: no | **Applicable for** MSc: yes

Further Notes: no

Topic 1.9: Collaborative State Space Exploration

Abstract:

Most complex and distributed system can be modeled as state graphs where the vertices describe the system states, connected by edges which present the transactions between them. Therefore, exploring the behaviour of unknown systems or system environments can be defined in terms of efficiently perform a complete or partial graph search. Thereby, all relevant vertices (system states) should be visited at least once. This can be done using simulated approaches, that are executed sequentially or in parallel [2]. Another possibility, posing a current research topic, is the usage of collaborative agents, that get the task of exploring subsets of the state graph and efficiently sharing their gained knowledge among each other. Among others, practical application of these methods can be found in the area of model checkers, robotics and drone technology.

References:

[1]http://ieeexplore.ieee.org/abstract/document/7072812/

[2]https://www.degruyter.com/view/j/cait.2016.16.issue-1/cait-2016-0001/cait-2016-0001.xml

[3]http://www.sciencedirect.com/science/article/pii/S0890540114001576

[4]http://ieeexplore.ieee.org/abstract/document/7487617/

Applicable for BSc: yes | **Applicable for MSc:** yes

Further Notes:

2 Large-Scale Cluster Computing

Topic 2.1: Publish/Subscribe with Apache Kafka	
Abstract:	
In modern distributed systems the communication between components often relies on messaging. Kafka is a distributed messaging system that was developed for collecting and delivering high volumes of log data with low latency. A few unconventional yet practical design choices in Kafka make it efficient and scalable with performance superior to popular alternatives.	
References:	
http://notes.stephenholiday.com/Kafka.pdf, https://www.rabbitmq.com/	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 2.2: Dominant Resource Fairness: Fair Allocation of Multiple Resource Types	
Abstract:	
Fairness is one of the key requirements when users share resources such as CPUs, memory, and I/O. Dominant Resource Fairness is an approach to provide fairness across multiple resources at the same time. It is used with the two popular open-source resource management systems for large-scale data analysis: Apache Hadoop (YARN) and Apache Mesos.	
References:	
http://static.usenix.org/events/nsdi11/tech/full_papers/Ghodsi.pdf	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 2.3: Automatic Resource Provisioning for Data-parallel Processing Systems	
Abstract:	
Data-parallel processing frameworks like MapReduce, Flink, and Spark arguably make analysis of very large datasets easier. Users create programs using a small set of pre-defined operators and write sequential code to configure these. The frameworks manage task parallelization and distribution as well as handle node failures. However, users do still need to specify how much and which resources to use for their jobs. This is often difficult and users consequently tend to overprovision significantly to ensure minimal performance requirements. Addressing this problem there are multiple systems such as Ernest and Jockey that predict the runtime of jobs and then select resources according to users' runtime targets. This allows users to specify their actual performance goals instead of having to guess an adequate set of resources.	
References:	
https://amplab.cs.berkeley.edu/publication/ernest-efficient-performance-prediction-for-large-scale-advanced-analytics/ https://www.usenix.org/sites/default/files/osdi16_full_proceedings_interior.pdf#page=125	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 2.4: Iterative Parallel Dataflows	
Abstract:	

Many important algorithms are iterative. These include, for example, many algorithms for graph analysis and machine learning. Parallel dataflows enable users in analyzing large datasets using clusters of computers. Data is processed through a graph of operators such as Map, Reduce, Join and GroupBy. These are executed in parallel and across many nodes.
With iterative algorithms the same parallel dataflow is executed repeatedly. This fact can be utilized in various ways, which is the theme of this topic.

References:

https://cs.stanford.edu/~matei/papers/2012/nsdi_spark.pdf
http://stratosphere.eu/assets/papers/spinningFastIterativeDataFlows_12.pdf
http://www.vldb.org/pvldb/vol6/p1678-popescu.pdf

Applicable for BSc: no	**Applicable for** MSc: yes

Further Notes: no

Topic 2.5: Making Sense of Performance in Data Analytics Frameworks

Abstract:

Making informed decisions in the design and implementation of large-scale data processing frameworks requires a good understanding of their performance, including knowing which resources are the bottleneck and therefore determine the performance of applications. Blocked Time Analysis is an approach to identify and quantify such bottlenecks for distributed processing frameworks. Knowing this helps to understand the performance of specific workloads using a specific frameworks and clusters, yet deriving facts about entire classes of systems—including separating conceptual reasons from implementation reasons—is still difficult.

References:

https://amplab.cs.berkeley.edu/publication/making-sense-of-performance-in-data-analytics-frameworks/
http://www.frankmcsherry.org/pagerank/distributed/performance/2015/07/08/pagerank.html
http://www.frankmcsherry.org/assets/COST.pdf

Applicable for BSc: yes	**Applicable for** MSc: yes

Further Notes: no

Topic 2.6: Tiered Storage on Hadoop

Abstract:

One Big Data main challenge is to deal with the deal with the exponentially growing data volumes, and to do so in an economically viable fashion. A promising trend in storage technologies is the emergence of heterogeneous and hybrid storage systems that deploy different types of storage devices, e.g. SSDs, HDDs, and ramdisks. The objective of this topic is to analyze recent developments for HDFS to support hybrid and tiered storage systems.

References:

http://www.ebaytechblog.com/2015/01/12/hdfs-storage-efficiency-using-tiered-storage/
http://people.cs.vt.edu/butta/docs/ccgrid2014-hats.pdf
http://pages.cs.wisc.edu/~akella/CS838/F15/838-CloudPapers/hdfs.pdf
http://dl.acm.org/citation.cfm?id=2670985
http://www.alluxio.org/

Applicable for BSc: yes	**Applicable for** MSc: yes

Further Notes: no

Topic 2.7: Programming Abstractions and Intermediate Representations for Distributed Dataflows

Abstract:

Various distributed dataflow frameworks have been developed for processing large datasets using clusters of computers. Examples include MapReduce, Spark, Flink, SCOPE, and Google Dataflow. These systems implement different strategies for fault tolerance, multiple solutions for processing of continuous data streams, as well as various plan optimization techniques. Moreover, the frameworks come with different libraries. So, there are many reasons for why users want to use a particular distributed dataflow framework for a job or even use multiple systems for different steps in an analysis pipeline. At the same time, each system uses its own programming abstraction and handing data from one system to another is usually an expensive operation, in which information on the intermediate data is also often lost. Addressing these problems, there is work aiming to provide more generally applicable programming abstractions and intermediate representations that allow to generate programs for multiple dataflow frameworks.

References:

https://arxiv.org/abs/1709.06416,
http://www.redaktion.tu-berlin.de/fileadmin/fg131/Publikation/Papers/emma-sigmod2015.pdf,
https://dl.acm.org/citation.cfm?id=2926540

Applicable for BSc: yes | **Applicable for** MSc: yes

Further Notes: no

3 Big Data Analytics & Visualization

Topic 3.1: Pareto Efficiency and its Applications

Abstract:

The task is to formally define what pareto efficiency is and explain how that is used in process mining to find an optimal solution in a high dimensional search space.
The usage of the pareto efficiency is described in the PhD thesis below in Chapters 5, 6.8 and 7.1

References:

https://pure.tue.nl/ws/files/4032592/780920.pdf

Applicable for BSc: no | **Applicable for** MSc: yes

Further Notes: no

Topic 3.2: Fairness Beyond Disparate Treatment

Abstract:

Automated data-driven decision making systems are increasingly being used to assist, or even replace humans in many settings. These systems function by learning from historical decisions, often taken by humans. In order to maximize the utility of these systems (or, classifiers), their training involves minimizing the errors (or, misclassifications) over the given historical data. However, it is quite possible that the optimally trained classifier makes decisions for people belonging to different social groups with different misclassification rates (e.g., misclassification rates for females are higher than for males), thereby placing these groups at an unfair disadvantage. To account for and avoid such unfairness, in this paper, we introduce a new notion of unfairness, disparate mistreatment, which is defined in terms of misclassification rates. We then propose intuitive measures of disparate mistreatment for decision boundary-based classifiers, which can be easily incorporated into their formulation as convex-concave constraints. Experiments on synthetic as well as real world datasets show that our methodology is effective at avoiding disparate mistreatment, often at a small cost in terms of accuracy.

References:	
https://arxiv.org/pdf/1610.08452.pdf	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.3: Decision Tree Online Learning	
Abstract:	
A lot of data is retrieved in a streaming online scenario setting. Such information streams need to be evaluated in real-time by machines. The real-time constraint can be succeeded by using decision trees, which need to be learned in a previous step from a given data stream. Furthermore, in an online scenario, concept shift is a known problem, which needs to be coped with when building a learning model from such data streams to enable a real-time evaluation.	
References:	
http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf, https://www.researchgate.net/profile/Geoffrey_Holmes3/publication/225395781_Fast_Perceptron_Decision_Tree_Learning_from_Evolving_Data_Streams/links/00b7d5159497fdff17000000.pdf , ftp://ftp7.freebsd.org/sites/ftp.sourceforge.net/pub/sourceforge/m/mo/moa-datastream/documentation/Manual.pdf	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.4: Query Languages for Graph DBs	
Abstract:	
The emerging of big data drives the need of network structured storage solutions, named graph databases. In order to query graph structured data, new query languages are created recently in order to enable real-time queries. Currently, SPARQL, Gremlin, and Cypher are the most frequently used query languages for graph databases. This topic discusses such query languages and states the differences to relational database query languages such as SQL.	
References:	
https://www.w3.org/TR/sparql11-overview/, http://gremlindocs.spmallette.documentup.com/, http://neo4j.com/developer/cypher-query-language/	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.5: Visualization of Multi-Property Graphs	
Abstract:	
Big data appears to emerge in a connected way, which is represented in a graph structure. Besides the connectivity property, such data carry a lot of further information. Such information is designed as properties connected to graph components. Those multi-property graphs are difficult to visualize, because of their large amount of carried content. This topic investigates solutions to deliver as much information to the user given large multi-property graphs.	
References:	
https://www.researchgate.net/profile/Hans-Joerg_Schulz/publication/274633015_A_Survey_of_Multi-faceted_Graph_Visualization/links/5523cb010cf2b351d9c33836.pdf	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.6: Storyline Generation from Social Media Data	
Abstract:	
Novel approaches enable the automated generation of storylines from data found on social networks, such as Twitter. The knowledge produced this way can help to make real-world processes more transparent and visible.	
References:	
http://arxiv.org/pdf/1605.05195.pdf http://arxiv.org/pdf/1606.03561.pdf	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.7: Sub-Story Detection	
Abstract:	
Sub-story detection allows to divide stories found on social networks into subparts and thereby enable a better understanding of the story elements and matching them to related story elements.	
References:	
http://arxiv.org/pdf/1504.07361.pdf http://arxiv.org/pdf/1605.05894.pdf	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 3.8: Convolutional Neural Networks	
Abstract:	
Convolutional neural networks (CNNs) utilize layers with convolving filters that are applied to features. They have been shown to be effective for natural language processing and have achieved excellent results in semantic parsing.	
References:	
http://arxiv.org/pdf/1408.5882.pdf https://pdfs.semanticscholar.org/eba3/6ac75bf22edf9a1bfd33244d459c75b98305.pdf	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

4 Internet of Things

Topic 4.1: Opportunistic Networking
Abstract:
Based on the idea of Mobile Ad-hoc Networks (MANETs), mobile devices such as smart phones are able to establish connections among each other spontaneously. This dynamic shape of collaboration reveals opportunities for a networking scheme that is based on opportunistic message forwarding. Similar to human interaction, two nodes can exchange information even if a route between them never exists.
References:
https://www.researchgate.net/profile/Andrea_Passarella/publication/3199770_Abstract_Opportunistic_Networking_Data_Forwarding_in_Disconnected_Mobile_Ad_hoc_Networks/links/55dc2fc908aed6a199ac7d58.pdf https://www.researchgate.net/profile/Chiara_Boldrini/publication/221453794_ContentPlace_social-aware_data_dissemination_in_opportunistic_networks/links/0046352458657165ff000000.pdf

Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 4.2: Extending the Cloud: The Role of Fog and Mobile Edge Computing	
Abstract:	
Caused by the proliferation of Internet of Things (IoT) applications, a continuously increasing amount of sensors and data sources connected to information networks like the Internet can be observed. A common approach to implement such IoT applications is to collect and forward data to analytics engines usually hosted on Clouds. This generates a large amount of traffic and stresses the availability and performance of the underlying information networks. As a mitigation, recent approaches like Fog Computing or Mobile Edge Computing investigate in utilizing resources offered by edge devices like routers or smart phones. The objective is to significantly reduce the amount of network load by implementing pre-processing of data close to the sources. The aim of this topic is to give an introduction to Fog and Edge Computing including architecture approaches and discuss the approaches with regard to common distributed systems challenges like availability, fault tolerance or scalability.	
References:	
http://www.openfogconsortium.org/ - OpenFog Reference Architecture White Paper (document available on request) http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing - MEC White Paper http://s3.amazonaws.com/academia.edu.documents/46135664/SIGCOMM-MMC-Fog.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1474944127&Signature=ZgJ04wcBmlhsBX5YSTljzDh0xuQ%3D&response-content-disposition=inline%3B%20filename%3DFog_Computing_and_Its_Role_in_the_Intern.pdf	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic 4.3: Analytic Monitoring for the Internet of Things	
Abstract:	
The topic at hand is *analytic monitoring* – detecting anomalies (outliers) in streams of data. In many scenarios this combines the twin demands of scale (hundreds of thousands to millions of complex events per second) and timeliness (minutes or maybe even seconds to report an outlier situation). All the while performing complex statistical calculations. The solution that Bailis et al. come up with combines robust statistical estimation with several novel streaming data structures. The aim of this topic is to give an introduction to the system design as well as optimizations and data structures.	
References:	
http://www.bailis.org/papers/macrobase-sigmod2017.pdf	
Applicable for BSc: no	**Applicable for** MSc: yes
Further Notes: no	

Topic 4.4: Scheduling Analytic Tasks in Heterogeneous Edge Computing	
Abstract:	
A large number of sensors generates data now. This data has to be analyzed to gain new insights. Examples include fault detection, prediction of traffic load, and monitoring of pollution or noise. Usually, sensors are connected to devices that transfer emitted sensor data to central analytic clusters. Such devices, however, often have some computing capabilities as well, allowing for early	

steps of data processing pipelines to happen highly distributed and close to the sources of data. However, the overall processing environment then becomes rather complex: resources available for processing are highly heterogeneous, the network connecting resources is a wide-area network with considerable latencies between some of the processing resources, and some analytics steps like aggregation require data from all sources and consequently cannot be computed locally on devices.	
References:	
http://ieeexplore.ieee.org/abstract/document/8031469/, http://ieeexplore.ieee.org/abstract/document/8029720/	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

5 Resilient Cloud Infrastructures

Topic : 5.1: improving robustness of Cloud infrastructure Services using Containers	
Abstract: As cloud computing is the current de facto standard for running web based applications. Infrastructure providers guarantee a uptime of their services of 99.9999%. To archive this goal redundancy of these service is mandatory. Redundancy and recovery mechanism can handled different when running critical cloud infrastructure services inside containers.	
References:	
[1] https://pdfs.semanticscholar.org/277b/5f6d78311009a5fc1fd16b5a99f990abcd34.pdf	
Applicable for BSc: no	**Applicable for MSc:** yes
Further Notes: no	

Topic : 5.2: Anomaly Detection in Cloud Infrastructures	
Abstract: Clouds are built atop a multitude of physical and virtual infrastructure elements, which leads to a complex infrastructure with a dense network of dependencies between layers (vertical) and components (horizontal). Misbehaviour and failures of hardware and software, and administrators' mistakes can cause unforeseen anomaly situations that are hard to track down and fix. Automatic anomaly detection mechanisms are required to meet the reliability requirements of large cloud infrastructures.	
References:	
[1] https://www.usenix.org/system/files/conference/hotcloud14/hotcloud14-vallis.pdf [2] https://pdfs.semanticscholar.org/13de/ab526e6e0762f500694affe587ed298e5233.pdf [3] http://shiftleft.com/mirrors/www.hpl.hp.com/personal/Vanish_Talwar/papers/2010_NOMS_AnomalyDetection.pdf	
Applicable for BSc: no	**Applicable for MSc:** yes
Further Notes: no	

6 Miscellaneous Topics

Topic 6.1: Micro-Services	
Abstract:	
The term Micro-Services is related to a style or pattern in software architecture. Single applications are developed as a suite of small services. These services interact with each other using inter process communication and language agnostic APIs. The objective of this topic is to introduce the micro-services architecture style and distinguish it from related styles such as SOA.	
References:	
http://martinfowler.com/articles/microservices.html http://nirmata.com/2015/02/microservices-five-architectural-constraints/ http://injoit.org/index.php/j1/article/view/139	
Applicable for BSc: yes	**Applicable for** MSc: yes
Further Notes: no	

Topic: 6.2 Blockchain Technologies	
Abstract: Blockchain technologies enable novel applications for distributed consensus tracking and verification. Bitcoin was the first application to use this kind of technology and is currently the most well-known and popular blockchain application. Beside Bitcoin, other blockchains, such as Ethereum, have attracted a lot of attention recently. It ensure that data and small computer programs called smart contracts are replicated and processed on all the computers on the network, without a central coordinator. This topic will focus on understanding, presenting and comparing the underlying blockchain technology of Bitcoin, Ethereum, and other altcoins.	
References:	
2 Mastering Bitcoin: Unlocking Digital Cryptocurrencies, Andreas M. Antonopoulos, O'Reilly, 2015	
Applicable for BSc: yes	**Applicable for MSc:** yes
Further Notes:	

Topic 6.3: Discrimination Discovery and Algorithmic Fairness
Abstract:
Personalization and algorithmic decision making based on Big Data have become ubiquitous in our daily lives. They are essential tools in personal finance, health care, hiring, housing, education, and insurance policies. Data and algorithms decide about the media we consume, the stories we read, the people we meet, the places we visit, whether we get a job, or if our loan request is approved. In many cases it is personal data that is used by decisionmaking algorithms to actually qualify people as more or less useful to the user of a certain search tool. Such the produced result does not only influence the searcher but also the people that compose the result set. It is therefore of societal and ethical importance to ask whether these algorithms eventually produce results that demote, marginalize, or exclude individuals belonging to an unprivileged group or a minority. These algorithms may have discriminatory effects, even in the absence of discriminatory intent, imposing a less favorable treatment to already disadvantaged groups. These problems are

exacerbated when details about the algorithms used for determining the “most adequate candidates” are unknown.

References:

[1] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21, 2 (2010), 277–292
[2] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM, 214–226
[3] Juhi Kulshrestha, Muhammad B. Zafar, Motahhare Eslami, Saptarshi Ghosh, Johnnatan Messias, and Krishna P. Gummadi. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In Proc. Of Computer Supported Collaborative Work and Social Media (CSCW).

Applicable for BSc: yes	**Applicable for MSc:** yes

Further Notes:

Topic 6.4 : Human Compatible AI

Abstract:

Various topics on morality in artificial intelligence algorithms. Reaching from basics on MDPs to the psychology of moral decisions and their application in an intelligent agent. Specific topic is to be discussed with the student and supervisor.

References:

[1] http://people.eecs.berkeley.edu/~russell/classes/cs294/s16/readings.html

Applicable for BSc: yes	**Applicable for MSc:** yes

Further Notes:

Topic 6.5: Discrimination Discovery in NLP

Abstract:

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to “debias” the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias

References:

[1] Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

Applicable for BSc: no	**Applicable for MSc:** yes

Further Notes:

Topic 6.6: Optimal Decision Making under multiple Constraints

Abstract:

Many everyday situations require human beings to make decisions. Therefore, we are compelled to consider many possible options and respecting certain constraints. Some of the constraints are more relevant for the selected choice than others. This weighting is done intuitively based on previous experience and the ability to foresee the outcome. One field of research is the design of automated decision making algorithms, which should either imitate the human decision making or be used as recommendation systems, supporting the decision making under complex constraints. Depending on the proposed problem, such systems may lay the focus on multiple input attributes [1], multiple constraints to consider [2] or both [3]. Furthermore, in some cases it may be reasonable to consider uncertainty, when making decisions [4, 5].

References:

[1]http://www.sciencedirect.com/science/article/pii/S0736584506000044
[2]http://www.sciencedirect.com/science/article/pii/030505489390109V
[3]http://www.sciencedirect.com/science/article/pii/S1364032116309479
[4]http://www.sciencedirect.com/science/article/pii/S0957417415000081?via%3Dihub
[5]http://www.sciencedirect.com/science/article/pii/S0307904X13004642

Applicable for BSc: yes	**Applicable for MSc:** yes

Further Notes: It is possible to either get a broad idea of the possible approaches, shortly introducing each of them or to select one approach and dive into the theoretical details and practical framework implementation of it.