Cloud Computing & Big Data

PARALLEL & SCALABLE MACHINE LEARNING & DEEP LEARNING

Prof. Dr. – Ing. Morris Riedel
Associated Professor
School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
Research Group Leader, Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

@Morris Riedel @MorrisRiedel @MorrisRiedel

LECTURE 0

Prologue

August 25, 2020
Online Lecture

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

DEEP Projects

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

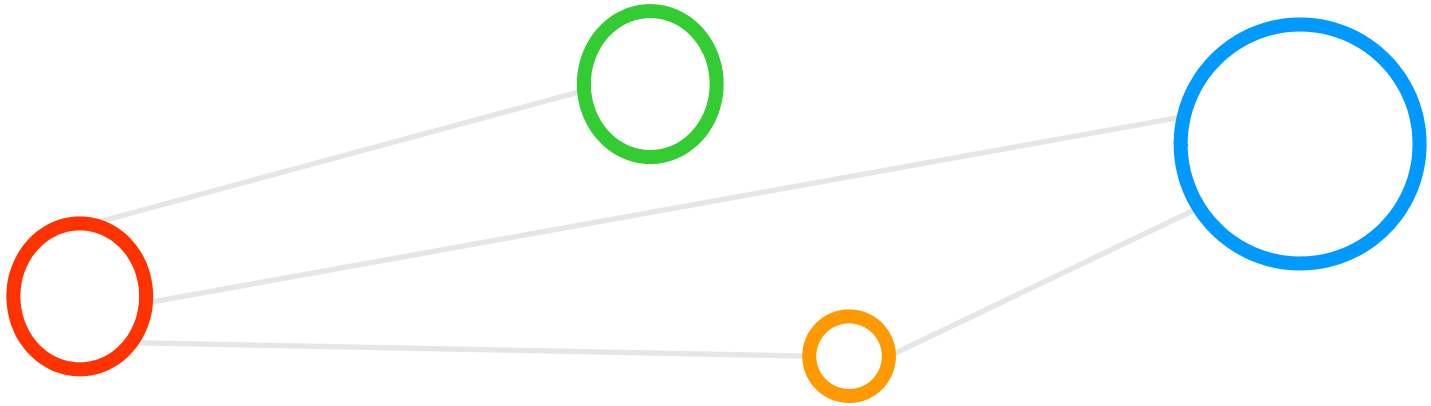
Outline of the Course

1. Cloud Computing & Big Data Introduction
2. Machine Learning Models in Clouds
3. Apache Spark for Cloud Applications
4. Virtualization & Data Center Design
5. Map-Reduce Computing Paradigm
6. Deep Learning driven by Big Data
7. Deep Learning Applications in Clouds
8. Infrastructure-As-A-Service (IAAS)
9. Platform-As-A-Service (PAAS)
10. Software-As-A-Service (SAAS)
11. Big Data Analytics & Cloud Data Mining
12. Docker & Container Management
13. OpenStack Cloud Operating System
14. Online Social Networking & Graph Databases
15. Big Data Streaming Tools & Applications
16. Epilogue

+ additional practical lectures & Webinars for our hands-on assignments in context

- Practical Topics
- Theoretical / Conceptual Topics

Course Motivation & Information



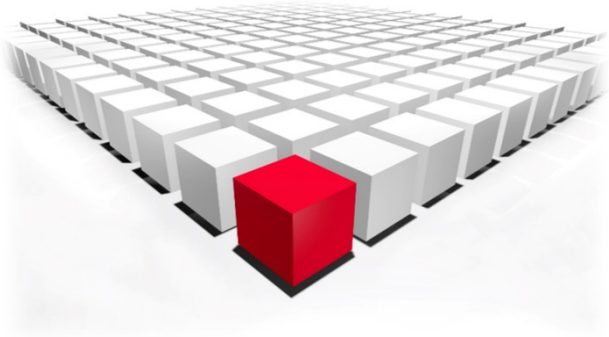
Course Motivation

- Parallel processing and distributed computing have matured (over the past three decades)
  - Both Emerged as a well developed field in computer science
- Course offers a comprehensive approach that integrates…
  - Computing theories and information technologies
  - Design, programming, and application of distributed cloud systems
  - Approaches to solve big data issues
  - Relationships to related topics of High Performance Computing (HPC)

- Course addresses the latest advances…
  - In hardware, software, system architectures & technologies such as Graphical Processing Units (GPUs)
  - Commercial Cloud ecosystems (including costs, speed performance, and energy efficiency)
  - Parallel and scalable machine learning tools & services
  - Deep learning driven by Big Data & Clouds
  - European endeavours such as the European Open Science Cloud (EOSC)

Selected Learning Outcomes

- Students understand…
  - Latest developments in parallel processing and distributed computing via cloud systems
  - How to create high-performance clusters
  - What is scalable machine & deep learning
  - Automated computing & big data centers
  - High throughput cloud systems & streaming
  - Transformations from traditional multiprocessors and multi-computer clusters into Web-scale Clouds
  - Technologies of large-scale online social networks & graphs
- Students are able to …
  - Take advantage of innovative machine & deep learning tools in clouds
  - Programm and use distributed cloud systems for ‘big data mining‘
  - Work with technologies and approaches to handle big data using clouds



Lecturer Prof. Dr. – Ing. Morris Riedel (since ~2004 in HPC)

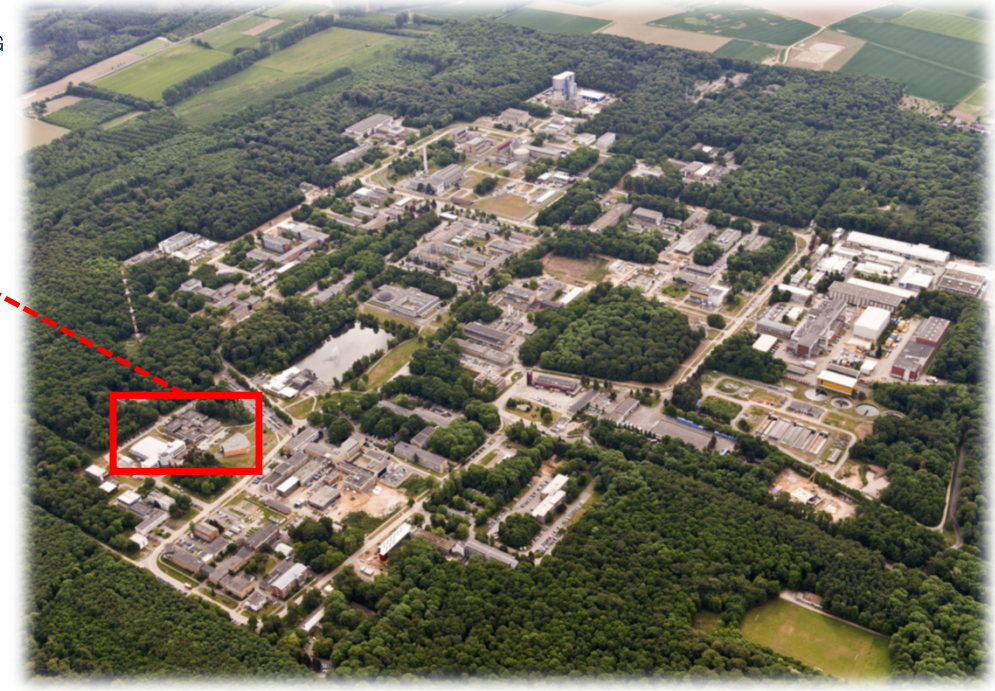
- Holds PhD in Computer Science (from Karlsruhe Institute of Tech.)
  - MSc in data visualization and steering of HPC & Grid applications
- Over the time several Positions at Juelich Supercomputing Centre
  - OS, Grid divisions; later deputy division leader federated systems and data
  - Currently: Research Group Leader – High Productivity Data Processing
- Selected other recent activities
  - Working with CERN & LHC & Grid/Cloud (Strategic Director of EU Middleware)
  - Architect of Extreme Science and Engineering Discovery Environment XSEDE (US HPC Infrastructure)
  - Co-Design of European Data Infrastructure (EUDAT), Research Data Alliance Big Data (Analytics) Chair, DEEP-EST HPC design, steering group of Helmholtz Artificial Intelligence Initiative
  - European EuroHPC Joint Undertaking Governing Board member for Iceland
- University courses
  - University of IcelandCourses: HPC A / B, Statistical Data Mining, Cloud Computing & Big Data
  - Slides from previous years available under teaching of instructors personal Web page

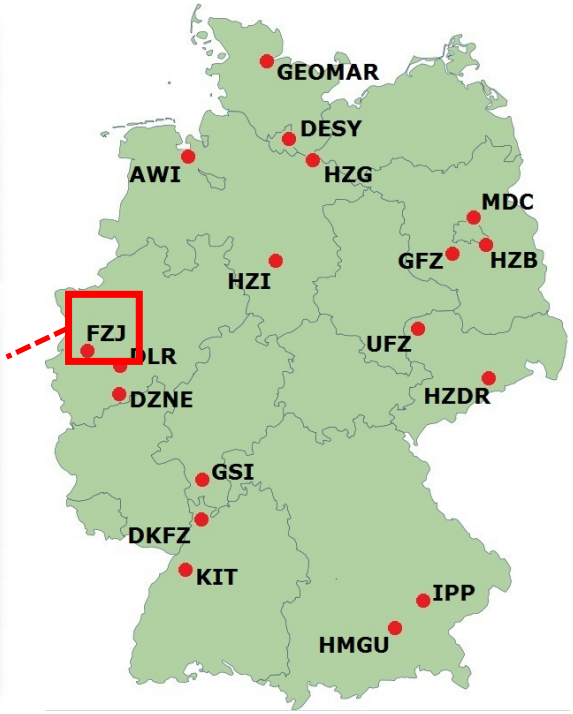
[3] Morris Riedel Web page

Juelich Supercomputinc Centre of Forschungszentrum Juelich – Germany

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

[5] Forschungszentrum Juelich Web page





- Selected Facts
  - One of EU largest inter-disciplinary research centres (~5000 employees)
  - Special expertise in physics, materials science, nanotechnology, neuroscience and medicine & information technology (HPC, Clouds & Big Data)

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

[4] Holmholtz Association Web Page

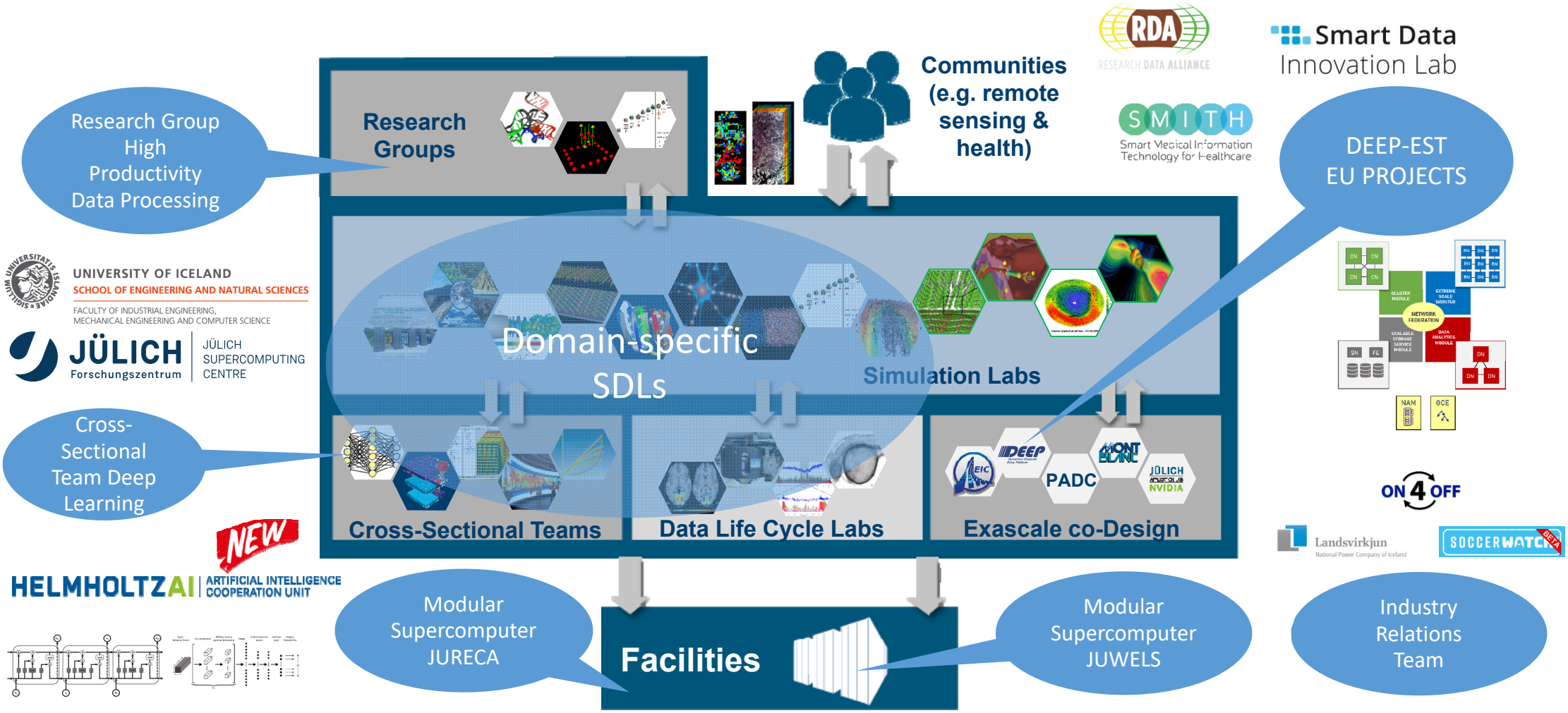
University of Iceland – School of Natural Sciences & Engineering (SENS)

- Selected Facts
  - *Ranked among the top 300 universities in the world (by Times Higher Education)*
  - *Ranked #6 in the field of remote sensing (by Shanghai list)*
  - ~2900 students at the SENS school
  - Long collaboration with Forschungszentrum Juelich
  - *~350 MS students & ~150 PhD students*
  - *Many foreign & Erasmus students*
  - *English courses*

[6] University of Iceland SENS Web Page

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Jülich Supercomputing Centre High Productivity Data Processing Research Group



Intertwined: High Performance Computing & Cloud Computing & Big Data

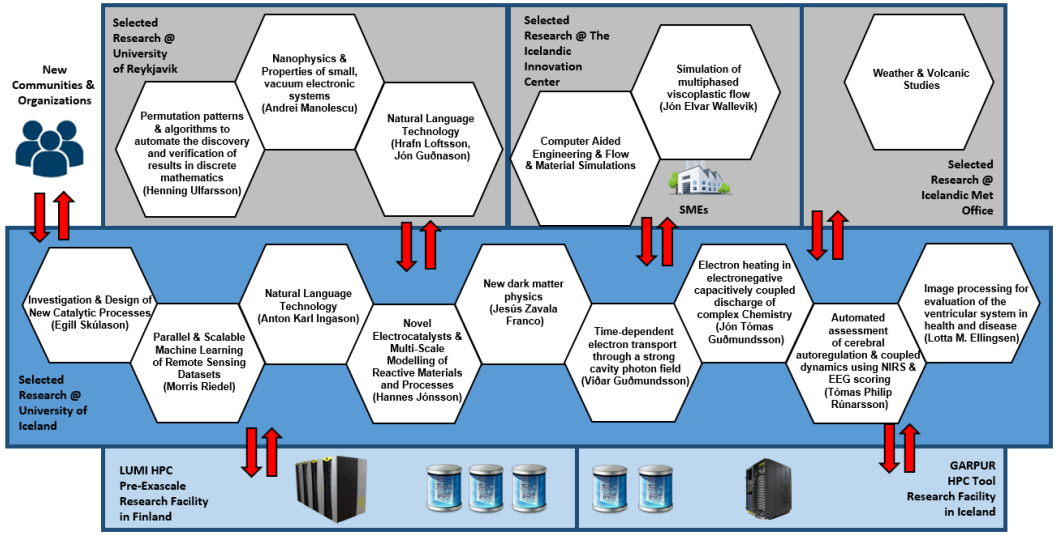
- European EuroHPC Joint Undertaking
  - EU EuroCC project in Iceland: user support & structuring of High Performance Computing (HPC) communities & roadmaps
  - EU ADMIRE Project: Remote Sensing application co-design of HPC systems

[28] EuroHPC Joint Undertaking

- European Open Science Cloud (EOSC)
  - Provides services and tools for large-scale datasets (aka 'big data') for EU researchers
  - Offers computing capacity for scientists in EU
  - EU EOSC-Nordic project in Iceland: provisioning of a couple of data services for selected application communities in Iceland

[29] EOSC Web page

[30] EOSC-Nordic Web page



Icelandic High Performance Computing (IHPC) Expert Network as Iceland Competence Centre

Terminology & Differences between AI, ML & DL

Artificial Intelligence (AI)

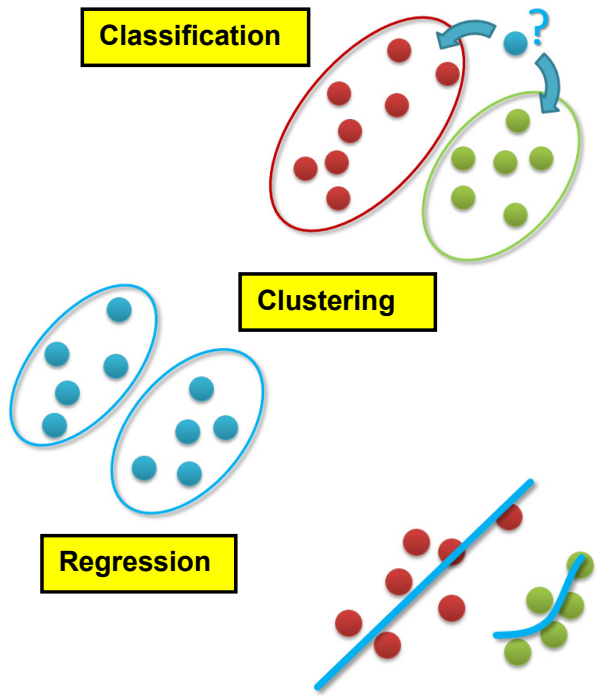
A wide area of techniques and tools that enable computers to mimic human behaviour (+ robotics)

Machine Learning (ML)

Learning from data without explicitly being programmed with common programming languages

Deep Learning (DL)

Systems with the ability to learn underlying features in data using large neural networks



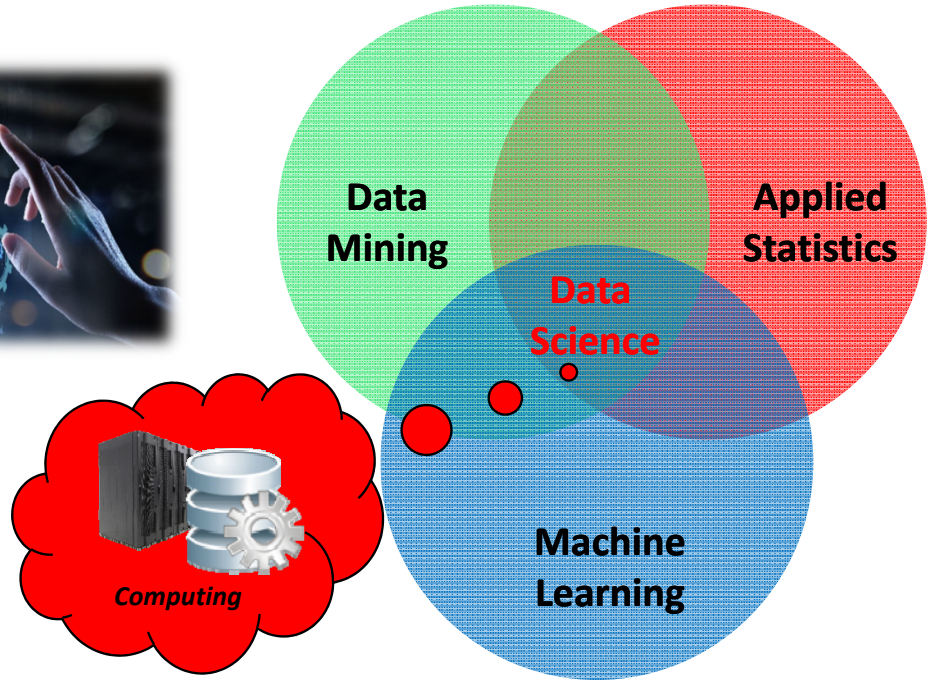
[9] Helmholtz AI Web page

HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Machine Learning Prerequisites & Computing Challenges

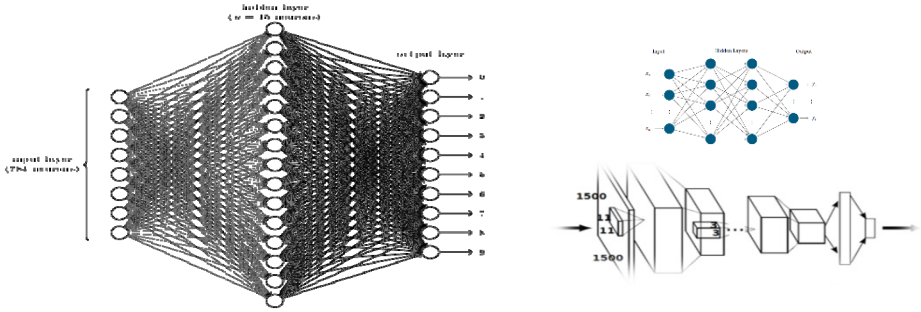
1. Some pattern exists
2. No exact mathematical formula
3. Data exists

- Idea ‘Learning from Big Data‘
  - Shared with a wide variety of other disciplines
  - E.g. signal processing, big data data mining, etc.
- Challenges
  - Data is often complex
  - Requires ‘Big Data analytics‘
  - Learning from data requires processing time → Clouds or High Performance Compuing

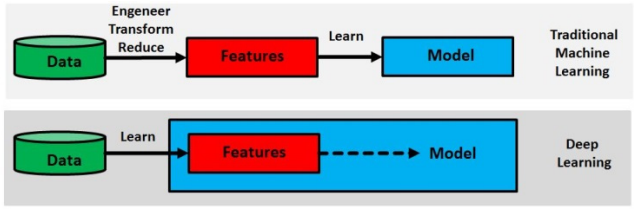


- **Machine learning is a very broad subject and goes from very abstract theory to extreme practice (‘rules of thumb’)**
- **Training machine learning models needs processing time (clouds or high performance computing)**
- **While data analysis is more describing the process of analysin the data, the term data analytics also includes and the necessary scalable or parallel infrastructure to perform analysis of ‘big data’**

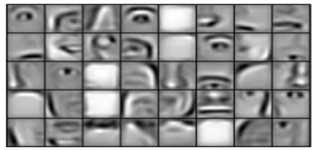
Innovative Deep Learning Techniques



[31] M. Riedel, 'Deep Learning - Using a Convolutional Neural Network', Invited YouTube Lecture, six lectures, University of Ghent, 2017



[32] M. Riedel et al., 'Introduction to Deep Learning Models', JSC Tutorial, three days, JSC, 2019

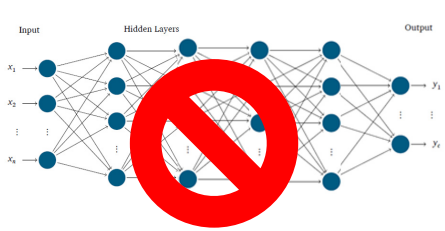


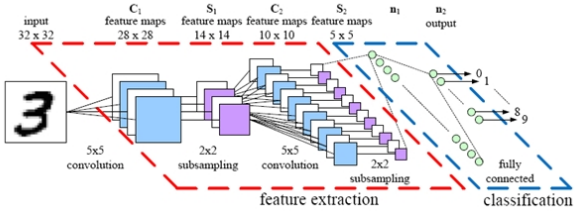
[33] H. Lee et al., 'Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical

- **Innovation via specific layers and architecture types**

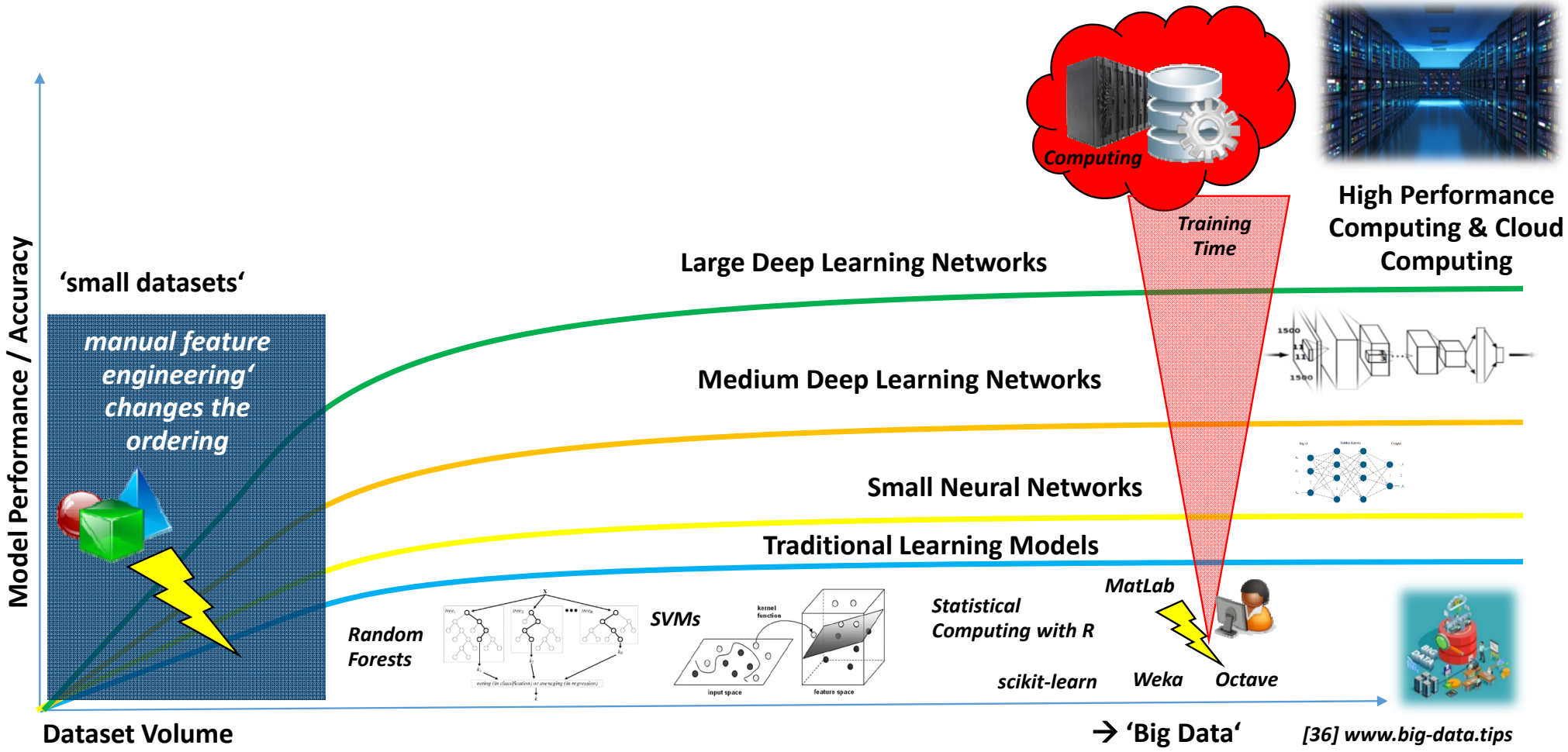
[34] Neural Network 3D Simulation

[35] A. Rosebrock





Complex Relationships: ML & DL vs. HPC/Clouds & Big Data



Understanding Deep Learning & Big Data Momentum & Startup Example

1952 Stochastic Gradient Descent
- Solving optimization problems

1958 Perceptron Learning Model
- Learning weights

1985 'Backpropagation of Error' approch in learning
- Artificial Neural Networks

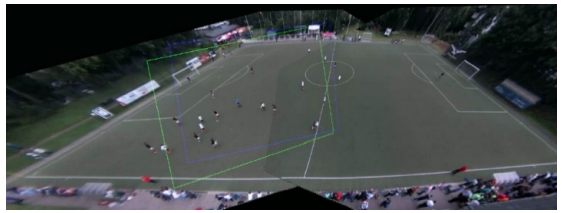
1995 Deep Convolutional Neural Networks
- Significant improvements in image analysis

Impact in AI & HPC in industry & science

Big Data
- Large datasets
- Easy access
- More storage for less cost

Hardware
- More memory
- Graphical Processing Units (GPUs)
- HPC & parallel systems

Software
- Scalable data science tools
- New learning models
- Open Source & free software packages

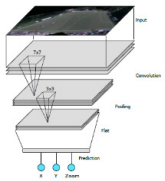


[37] NVIDIA

[18] Keras

[19] TensorFlow

[10] soccerwatch.tv



Combination: Start-up Example of my research group

[8] C. Bodenstein & M. Riedel et al., Automated Soccer Scence Tracking using Deep Neural Networks

DEEP Series of Projects – Modular Supercomputing Architecture Research

DEEP Projects

[7] DEEP Projects Web Page

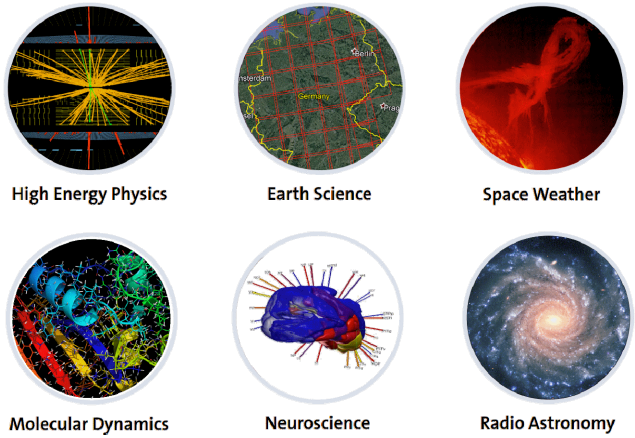
Strong collaboration with our industry partners Intel, Extoll & Megware

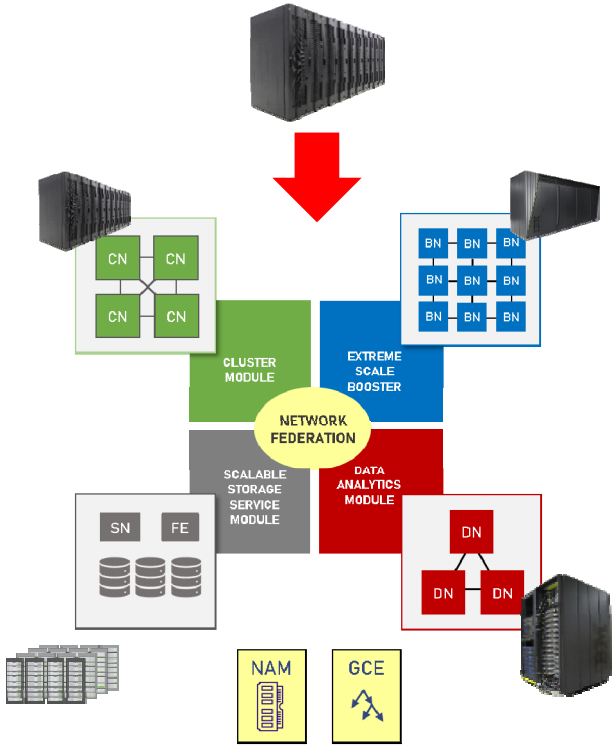
- 3 EU Exascale projects
  DEEP, DEEP-ER, DEEP-EST
- 27 partners
  Coordinated by JSC
- EU-funding: 30 M€
  JSC-part > 5,3 M€
- Nov 2011 – Dec 2020

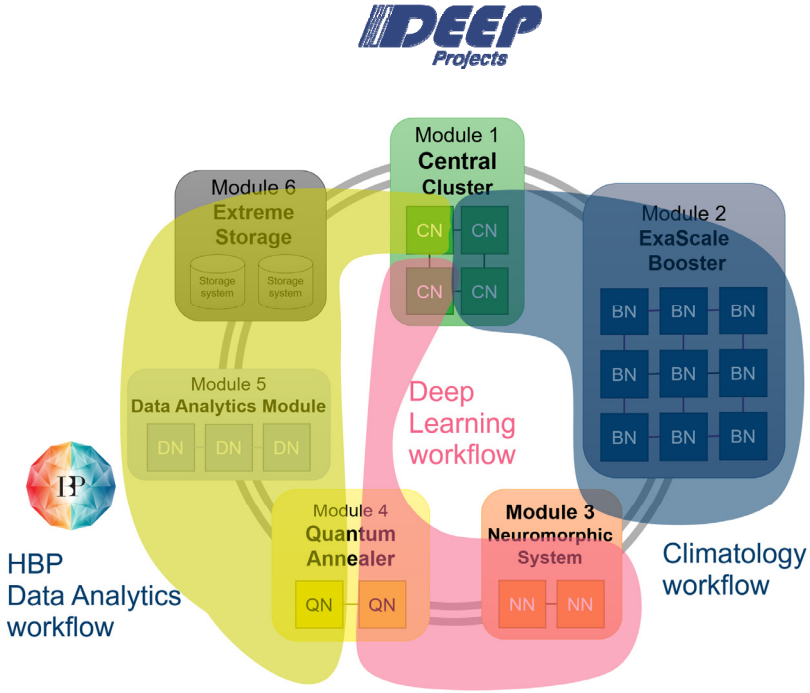
- **Strong collaboration with industry partners Intel, Extoll & Megware**

- **Juelich Supercomputing Centre implements the DEEP projects designs in its HPC infrastructure**

Application Co-Design for Machine & Deep Learning in HPC



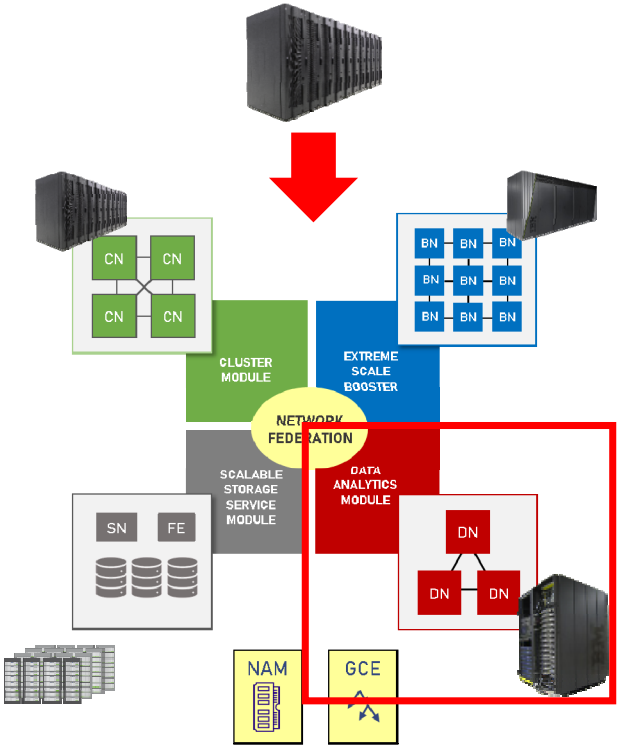




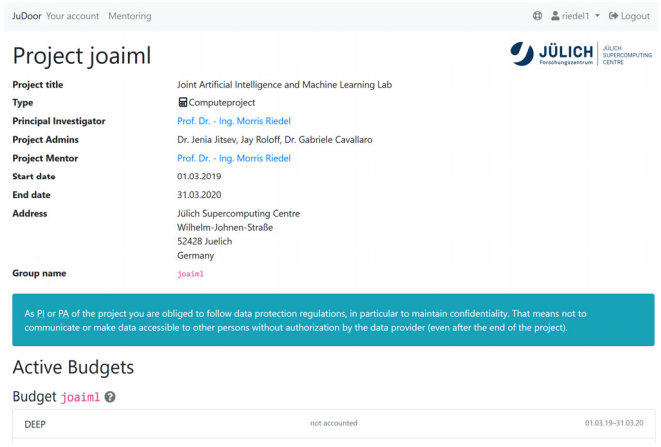
- **The modular supercomputing architecture (MSA) enables a flexible HPC system design co-designed by the need of different application workloads**

[7] DEEP Projects Web Page

Hands-On Training System – Data Analytics Module (DAM)



- Data Analytics Module (DAM)
  - Specific requirements for data science & analytics frameworks
  - 16 nodes with 2x Intel Xeon Cascade Lake; 24 cores
  - 1x NVIDIA V100 GPU / node
  - 1x Intel STRATIX10 FPGA PCIe3 / node
  - 384 GB DDR4 memory / node
  - 2 TB non-volatile memore / node
- DAM Prototype
  - 3 x 4 GPUs Tesla Volta V100
  - Slurm scheduling system



(easy join via JOAIML ab with JuDoor)

DEEP Projects

[7] DEEP Projects Web Page

➢ **The Data Analytics Module (DAM) will be used for a couple of different machine & deep learning exercises in the context of lectures**

Canvas Tool & Office Hours (!)

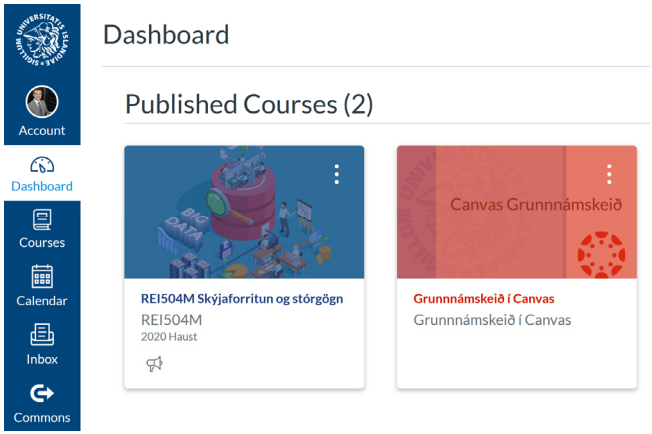
- Reference course information
  - Cloud Computing & Big Data
  - REI504M, Fall 2020
- Use it for course communication
  - Every course member requires account
  - Contact other students & discuss topics
  - Contact lecturer & access to all materials
- Find course materials
  - Slides of Lectures and Practical Lectures
  - Handouts and Recordings
  - Further reading topics (e.g. papers, etc.)
- Questions, major difficulties, etc.? → Don't wait long!
  - Use my online office hours, send request email to morris@hi.is

UNIVERSITY OF ICELAND
UGLA - intraweb

changes to

CANVAS

[13] Cloud Computing & Big Data Course Catalogue Web page



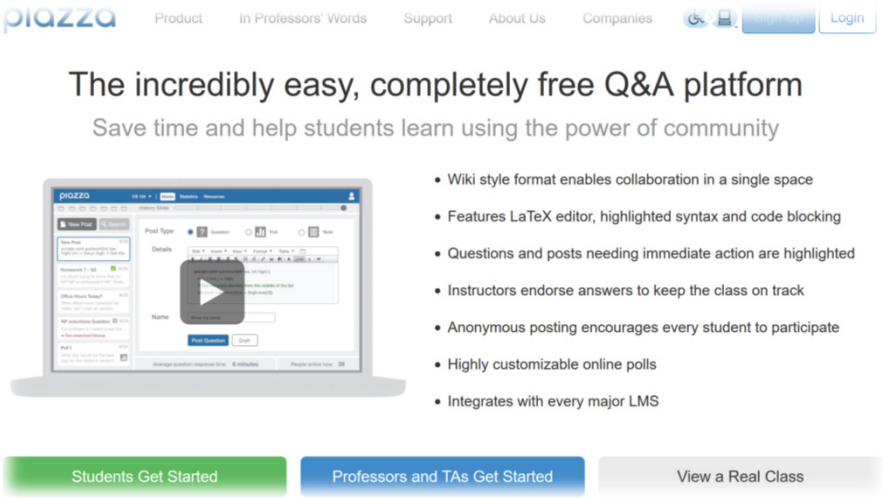
Overall Course Organization

- 3 Assignments (40% of grade)
  - Guided by practical lectures in context with hands-on elements for all
  - Cloud configuration & cloud programming projects
  - Influence in the overall grade
  - TBD(all): Create Groups of 2-3 and send the group to morris@hi.is
- Quizzes (10% of grade)
  - Small quiz from time to time (pre-announced) to check understanding
  - Minor influence in the overall grade – good preparation for exam
- Exam (50% of grade)
  - End of the lecture series (~December) – major part of the overall grade
  - 'Not knowing everything is key – but understand the important elements'
- Invited Lectures
  - A couple of presentations (e.g. companies, interesting projects, PhD students, etc.)

- **Each lecture will have these type of yellow blocks**
- **The most important course material and information is usually summarized in these yellow blocks**
- **It is essential to know these elements of the course, so pay attention on learning and understanding those**
- **Knowing the substance of these yellow blocks in context of some associated figures and illustrations fundamentally helps to pass the exam and have good quiz outcomes**

Course @ Q&A Platform Piazza

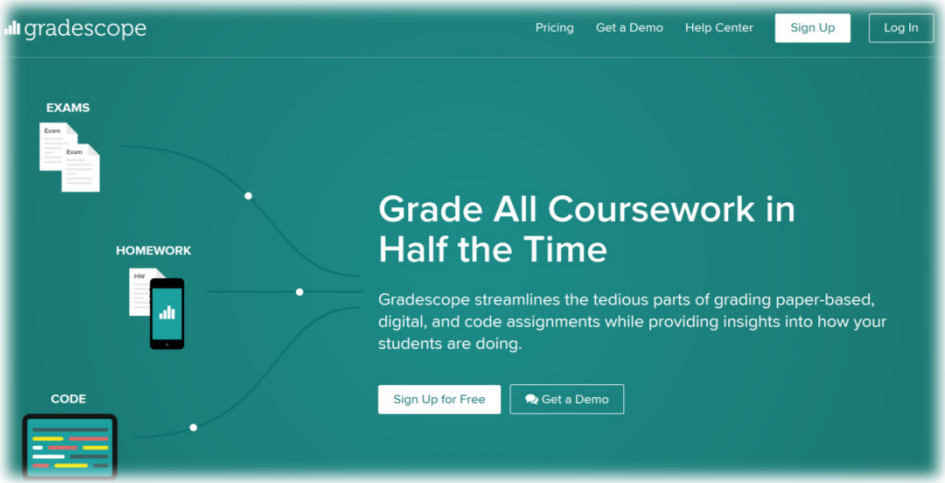
- Q&A Platform
  - Mixture between a wiki and a forum for students
  - Can be used by academic institutions for free
  - Idea: come together to share ideas and knowledge
  - E.g. ask questions about course assignments / content
  - In addition to Canvas although Canvas has also specific discussion threads in context of course
- Course information
  - You will be registered by the course instructor soon
  - Name: REI504M Cloud Computing and Big Data
  - Semester: Fall 2020
  - TBD (all students): Get familiar with Piazza once registered



[1] Piazza Web page

Course @ Gradescope

- Student Grading Platform
  - Grading for quizzes, assignments & exam will be performed
  - Can be used by academic institutions for free
  - Idea: get faster feedback for course content and a more fair grading process
  - E.g. professor does not see the name of students per task
  - Canvas integration
- Course information
  - You will be registered by the course instructor soon
  - Name: REI504M Cloud Computing & Big Data
  - Semester: Fall 2020
  - TBD (all students): Get familiar with Gradescope once registered



[2] Gradescope Web page

Associated Literature

Cloud Computing for Science and Engineering,
I. Foster & D. B. Gannon
MIT Press Publishers
ISBN 9780262037242, English, ~392 pages, 2017

[14] Book Cloud Computing for Science and Engineering

Cloud Computing for Machine Learning and Cognitive Applications,
Kai Hwang
MIT Press Publishers
ISBN 9780262036412, English, ~624 pages, 2017

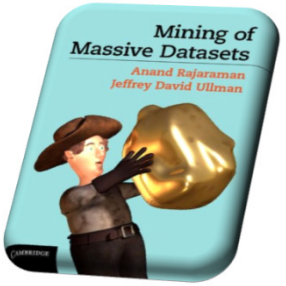
[23] Book Cloud Computing for Machine Learning & Cognitive Applications

Distributed and Cloud Computing,
K. Hwang, G.C. Fox & J.J. Dongarra,
Morgan Kaufmann Publishers
ISBN 0123858801, English, ~672 pages

[15] Book Distributed and Cloud Computing

Mining of Massive Datasets,
Anand Rajaraman, Jure Leskovec & Jeffrey D. Ullman,
available for free download, ~453 pages, 2013

[27] Book Mining of Massive Datasets Online

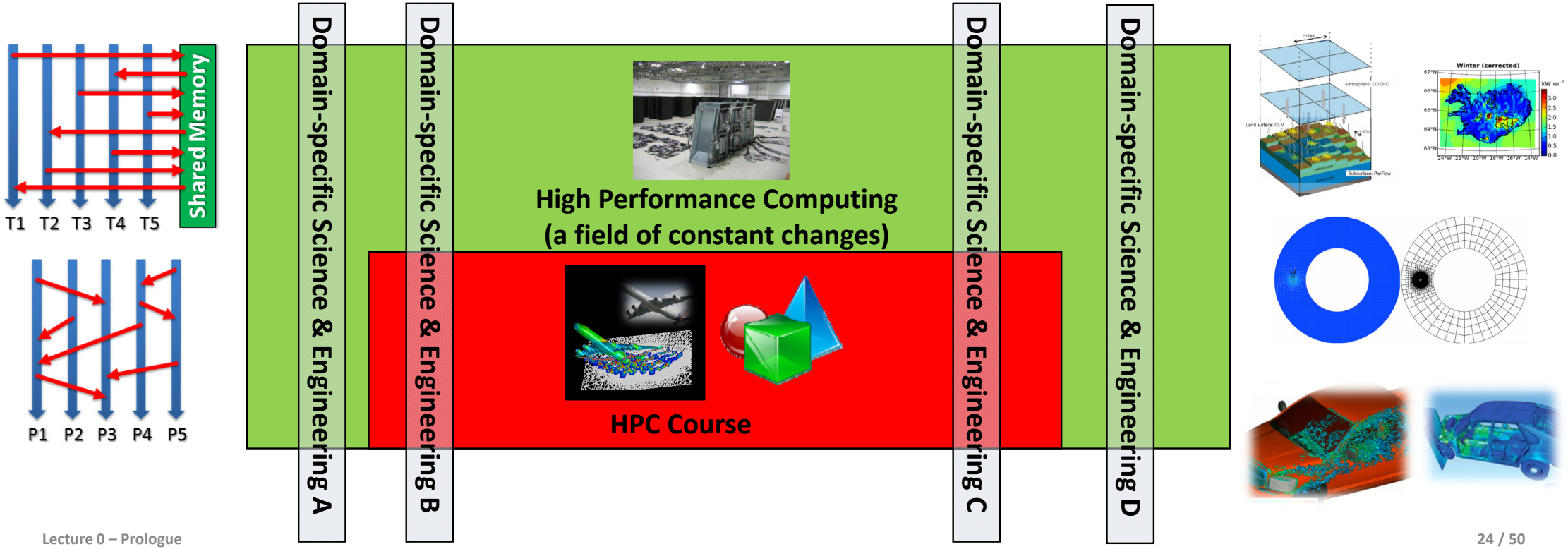


- Further bibliography and readings will be provided in context
  - E.g. Papers, Web pages, etc.

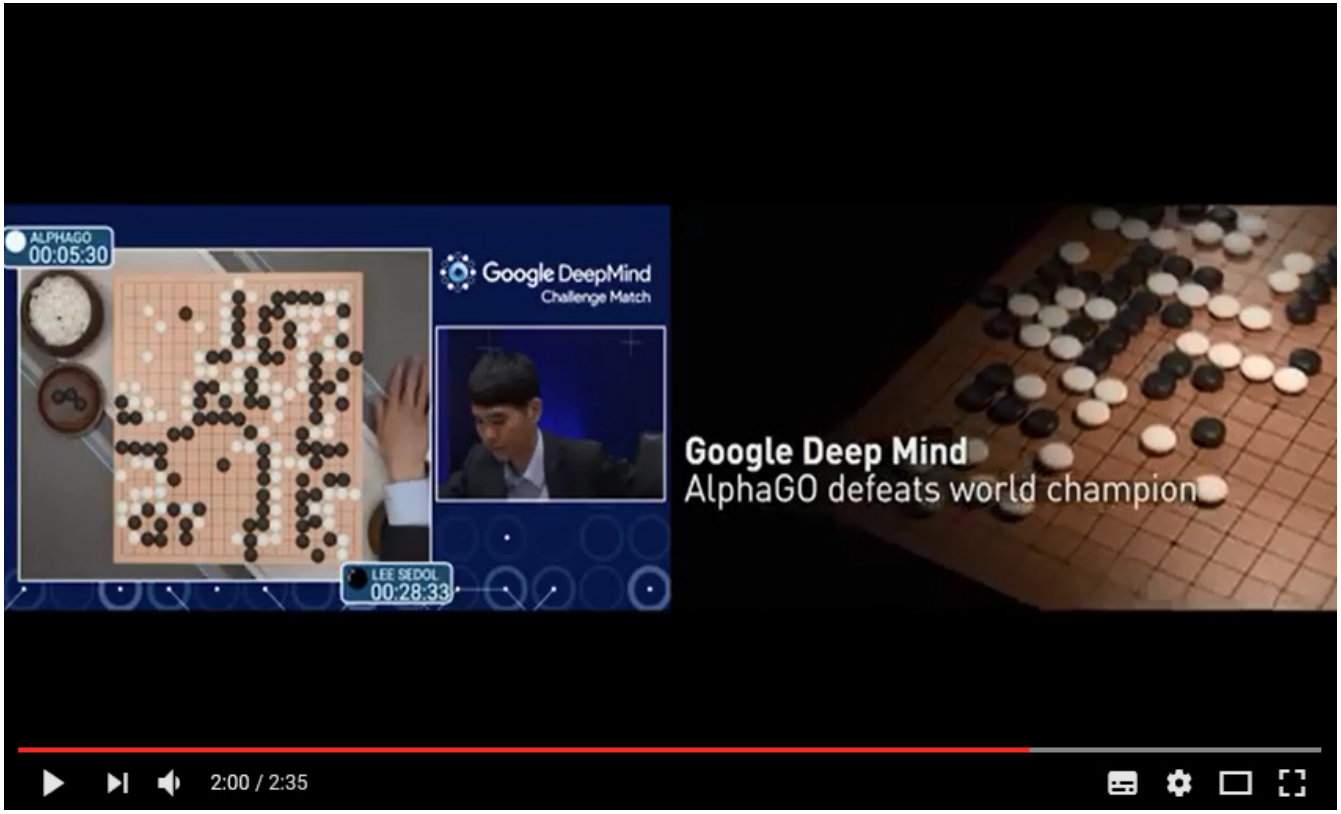
[24] www.big-data.tips ***[25] www.datanami.com*** ***[26] www.hpcwire.com***

Next Fall 2021: Complementary High Performance Computing (HPC) Course

- Consists of techniques for programming & using large-scale HPC Systems
  - Approach: Get a broad understanding what HPC is and what can be done
  - Goal: Train general HPC techniques and systems and selected details of domain-specific applications

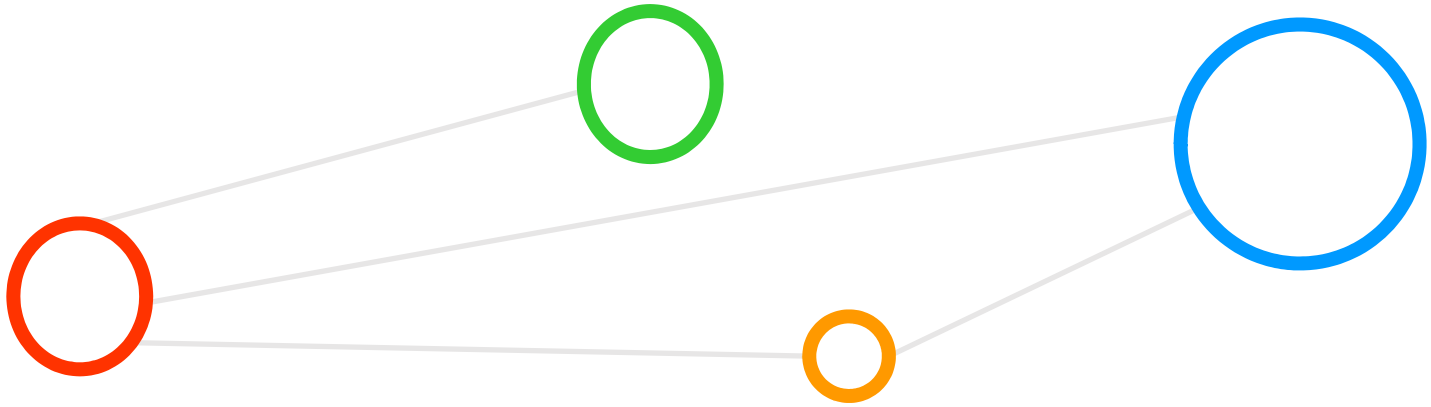


[Video] Deep Learning Applications



[11] The Deep Learning Revolution, YouTube

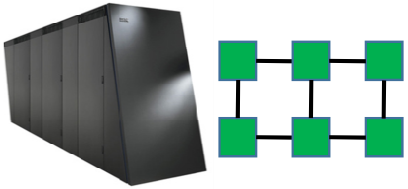
Course Organization & Content



Lecture 1 – Cloud Computing & Big Data Introduction

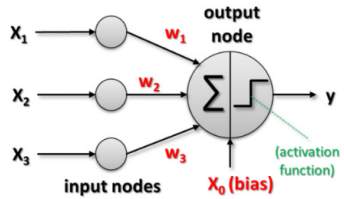
- Evolutions
  - Clouds come not from nowhere – long history of technologies
  - Evolutions in parallel & distributed computing
- Internet Cloud Systems – Examples
  - Google App Engine, Amazon Web Services, Facebook, SalesForce, Rackspace, IBM, Enomaly
  - European Open Science Cloud (EOSC) Approach & Services
- Terminologies
  - Parallel Computing, Distributed Computing, Centralized Computing, Cloud Computing
- Technology Foundations
  - Multi-core CPUs & Multi-threading, Many-core GPUs, Processes
  - Memory, Storage, Interconnects, Wide Area Networks
- Big Data Impacts & Big Data Analytics
  - Compute to Data vs. Data to Compute, New Programming Models, 7Vs

EOSC NORDIC

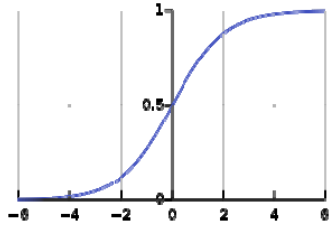


Lecture 2 – Machine Learning Models in Clouds

- Machine Learning Foundations
  - Machine learning has become central to cloud computing applications
  - Role of machine learning models (neural nets, decision trees, regression)
  - Phases like training, testing, validation & ‘big data‘ challenges in context



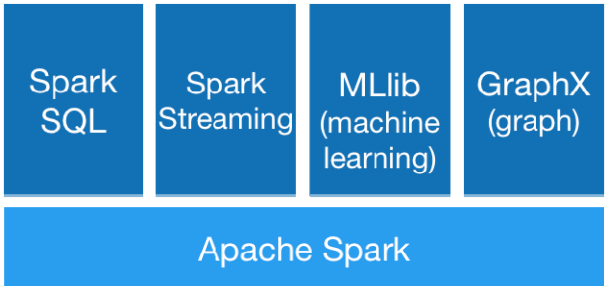
- Selected Machine Learning Models
  - Supervised & unsupervised machine learning models
  - Overview of various models & selected details
  - E.g. linear learning models (e.g. perceptron)
  - E.g. Logistic regression for multi-class classification



- Machine Learning Tools & Cloud Applications
  - Role of Jupyter notebooks & Python in accessing Cloud systems
  - Microsoft Azure services cloud examples
  - Amazon Web services (AWS) cloud examples

Lecture 3 – Apache Spark for Cloud Applications

- Apache Spark gets momentum
  - Large set of tools and libraries
  - Available in many cloud offerings (e.g., Microsoft Azure, Amazon Web Services)
  - Cloud Spark services usage & examples
- Apache Spark Capabilities
  - Resilient Distributed Datasets (RDDs)
  - Distributed Spark Operations
  - Job Execution & Dataflow modeling
- Advanced Concepts & Libraries
  - Structured Query Language (SQL)
  - Streaming Library
  - Machine Learning Library (MLlib)
  - GraphX Library



[16] Apache Spark Web page

Lecture 4 – Virtualization & Data Center Design

- Computer Clusters & Virtualization for Scalable Parallel Computing
  - Application-Driven Requirements, High-Performance Service
  - High Performance Computing, High Throughput Computing
- Clustering for massive parallelism
  - Cluster system interconnects, Cluster design principles: system architectures, nodes, racks
- Computing & Data Centres Operations
  - Data-Center design and interconnection networks
  - Infrastructure costs, water/cooling, energy/power, hardware/network
  - Performance metrics, scalability, fault tolerance, system availability
  - Payment models: pay-per-use, pay-as-you-go vs. cost of ownerships
  - Service Level Agreements, on-demand access, scheduling
- Security and Deployment Issues
  - Access, privacy, confidentiality, integrity, denial of service, trust, etc.
  - Deployment models: Private/Public/Hybrid Clouds

VERNE GLOBAL

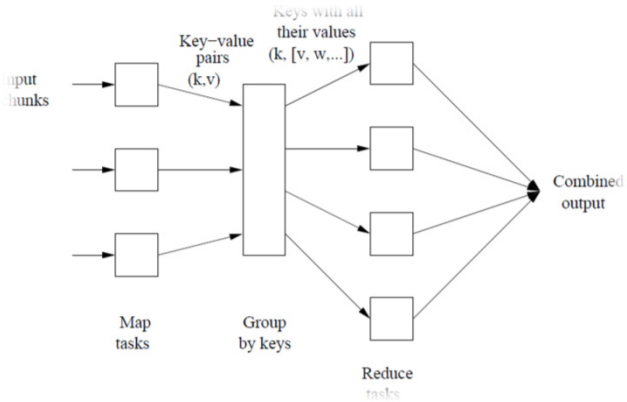
DATA CENTERS FOR FORWARD THINKERS

NVIDIA DATA CENTER

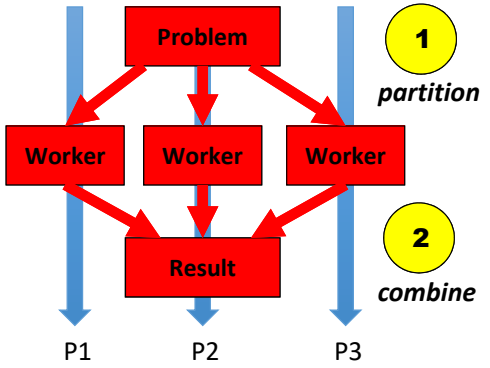
Google Data Centers

Lecture 5 – Map-Reduce Computing Paradigm

- Map-Reduce Programming Model
  - HPC Complexity & Motivation
  - Three ‘phases’ & Communication
  - Key-Value Data Structures & Applications
  - Map-Reduce Algorithms & Ecosystem
  - Scheduling with Yarn
  - Apache Hadoop implementation
- Selected Map-Reduce Applications
  - Matrix-Matrix/-Vector Calculations
  - BLAST Application in Computational Biology
  - PageRank Web Application
  - Machine Learning Application Performance
  - Google flu predictions

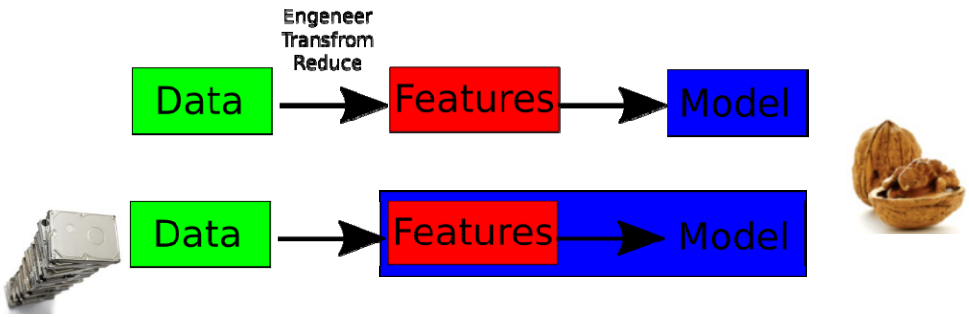


[17] Apache Hadoop Web page



Lecture 6 – Deep Learning driven by Big Data

- Deep Learning Fundamentals
  - Artificial Neural Networks (ANNs) foundations
  - Backpropagation Algorithm
  - Role of big data sets
  - Feature learning
  - Transfer learning
- Deep Learning Models
  - Using Cloud tools with TensorFlow & Keras
  - E.g. Convolutional Neural Networks (CNNs) models for image data sets
  - E.g. Long Short Term Memory (LSTM) models for time series data sets
  - Cloud services for deep learning models
  - Role of GPUs and scalable cloud data center models



[19] Tensorflow Deep Learning Framework

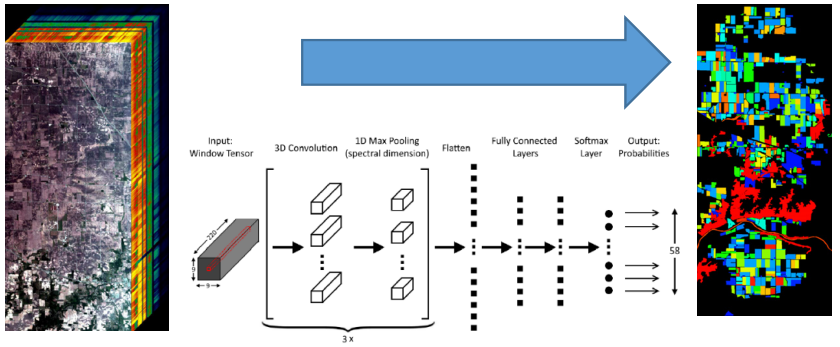
Keras

TensorFlow

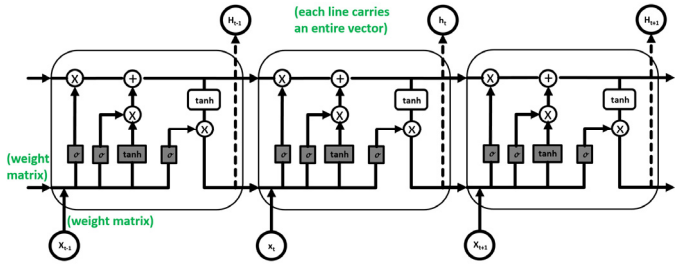
[18] Keras Python Deep Learning Library

Lecture 7 – Deep Learning Applications in Clouds

- Learning Image Data Sets
  - Convolutional Neural Networks (CNNs) models in Clouds
  - Several datasets in examples used in cloud environments
  - E.g. hand-written character recognition eamples
  - E.g. remote sensing data & earth observation data sets



- Learning Sequence Data Sets
  - Recurrent Neural Networks (RNNs)
  - Examples of Long Short-Term Memory (LSTM) models in Clouds
  - Several datasets in examples used in cloud environments
  - E.g. Natural Language Processing (NLP) applications & text processing



- Deep Learning requiring massive Computing Power
  - Pretrained Deep Learning Networks & Transfer Learning
  - Neural Architecture Search (NAS)

Lecture 8 – Infrastructure-As-A-Service (IAAS)

- IAAS
  - Conceptual Ideas, key usage: hosting, bare metal, provisioning
  - Based on virtual machines that are used to flexible share computing and storage resources in a scalable and safe manner
  - E.g. payments like pay-as-you-go, charging by hour (idle too)

- Understanding Amazon Web Services (AWS)
  - Commercial cloud IAAS model, ~70 cloud services
  - Elastic Computing Cloud (EC2) offers virtualized platforms to host VMs where cloud applications are running
  - Simple Storage Service (S3) offers an object-oriented storage service
  - Elastic Block Service (EBS) provides block storage in order to support more traditional applications
  - Elastic File System (EFS)

- Other IaaS examples
  - E.g., GoGrid/DataPipe, etc.

Lecture 9 – Platform-As-A-Service (PAAS)

- PAAS
  - Conceptual Ideas, key usage: building applications, SDK, APIs, etc.
  - Cloud platforms that enables a wide variety of Cloud and Web applications in the context big data
  - Application framework similar to ASP, J2EE, or JSP;
  - Using Python, Java, others

- Understanding Google App Engine
  - Commercial cloud PAAS model, develop/use scalable/elastic applications
  - Leverage powerful Google compute infrastructure with applications
  - Take advantage of Google‘s search engine operations
  - Datastore that provides object-oriented, distributed, and structured data storage services, Google Big Table, etc.
- Other PaaS examples
  - Manjrasoft, etc.

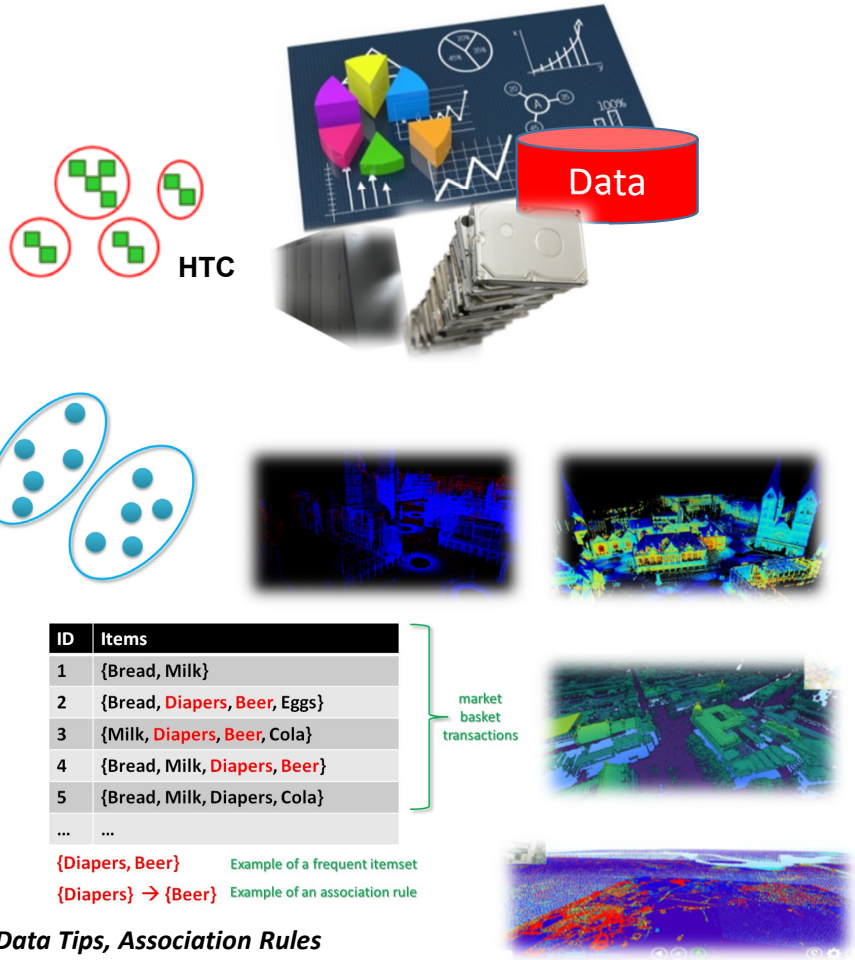
Lecture 10 – Software-As-A-Service (SAAS)

- SAAS
  - Conceptual Ideas, key usage: consume applications, avoid installations
  - Typically provide specific ‘ready-to-run applications’ that are sometimes also related to geographical locations
  - Service application is centrally hosted, user accesses via the Web browser
  - Usage model of this cloud computing software is a license
  - Licences can be purchased on subscription basis (e.g. daily, monthly, yearly, lifetime, etc.)

- Understanding concept of a wide variety of SAAS solutions
  - E.g. Google Docs, Google gMail
  - E.g. Microsoft SharePoint
  - E.g. Conga: Online Contract Management Solution
  - E.g. Zoho, Salesforce.com: Customer relationship management (CRM)
  - Several others…

Lecture 11 – Big Data Analytics & Cloud Data Mining

- Big Data Analytics & Analysis
  - Data Mining Applications & Complexity
  - Compute to Data vs. Data to Compute
  - Data Processing Machinery Requirements
  - Failures, Reliability & Scalability Needs
  - Two Key Requirements
- Cloud Data Mining
  - Clustering applications of big data sets (e.g. point clouds
  - Association Rule Mining (e.g., APRIORI, FP-GROWTH)
  - Alternative Least Squares (ALS)
  - Collaborative Filtering
  - Cloud services in context in libraries & services



[20] Big Data Tips, Association Rules

Lecture 12 – Docker & Container Management

- Management of a variety of cloud services and big data tools
  - Local machine vs. Cloud vs. Containers & automated configurations
  - Containers running on a single machine, sharing the same operating system kernel, starting instantly, use less RAM, use images, etc.
  - 'Contains' means: wraping software in a 'complete filesystem' that contains everything needed to run the software: code, runtime, system tools, system libraries, third-party libraries, data, etc.
  - Deployments, software stacks, abstracts CPU, memory, storage, and other compute resources away from machines (physical or virtual)
  - Fault-tolerant and elastic distributed systems
  - Open standards of containers & security

Apache MESOS™

- Understand (scalable) tools and techniques
  - E.g. Singularity & Docker: build, ship, run cloud applications, VMs, and services
  - E.g. Apache Mesos: operate a datacenter like a single pool of resources

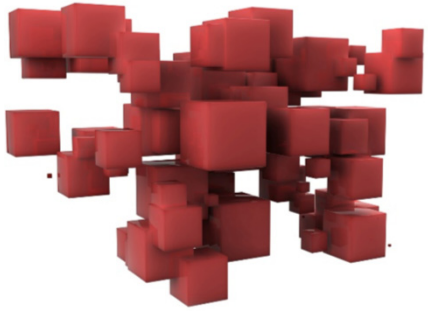
Lecture 13 – OpenStack Cloud Operating System

- OpenStack: open source cloud computing software
  - Manages large pools of compute, storage, networking resources.
  - Dashboard that provides administrators with a tool to manage all resources via one graphical user interface
- Conceptual services
  - Provide computing and block storage for end users & able to use the provisioned resources via a Web Interface
  - Its marketplace offers an easy way to find related products and services
  - Includes existing distributions, appliances, consultants, and trainings
- Well curated sample configurations
  - Based on real-world reference architectures across industries and workloads
  - Configurations will provide a good way to start and understand which core and optional cloud services are used for different environments
  - Examples are using OpenStack for Web applications, Big Data, eCommerce, Video processing and content delivery, or high throughput computing
- Other Cloud systems:
  - OpenNebula, Nimbus, Eucalyptus

[21] OpenStack Web page

Lecture 14 – Online Social Networking & Graph Databases

- Social Networking using Clouds
  - Enables many social and professional functions over the Internet
  - Transformed the way in which people interact in cyberspace
  - ‚Big Data': almost 1 billion registered users among the top 15 social networks (www.ebizmba.com/articles/social-networking-websites)
  - Online Social Networking Characteristics, e.g. social graph traversal along specific social links or networks, database requirements, etc.
  - Provider revenue from embedded advertisement and access to premium content, e.g. Facebook Ads, Google Adwords/Adsense, etc.
  - Application domains, communities, architectures & network impact
- Understanding key features and techniques of examples
  - E.g. Facebook, sharing photos/videos and making friends, share events
  - E.g. Twitter, microblogging and exchange of ideas/events/products
  - E.g. LinkedIn, professional networkings and job offerings/recruiting



Lecture 15 – Big Data Streaming Tools & Applications

- Apache Flink shows more and more momentum
  - Scalable batch and stream data processing, use of memory
  - Compare to Apache Storm (Flink has much higher throughput)
- Flink/Storm different approach to distributed computing
  - Provides a streaming dataflow engine approach
  - Fast techniques for data distribution, communication, and fault tolerance for distributed computing using data streams
- Discussions of Apache Flink libraries & APIs
  - Datastream API, streams embedded in Java/Scala
  - Dataset API, static data embedded in Java/Scala/Python
  - Complex event processing (CEP) library & Machine Learning Library
- Discussions of Apache Spark libraries & APIs
  - Spark streaming & applications

[22] Apache Flink Web page

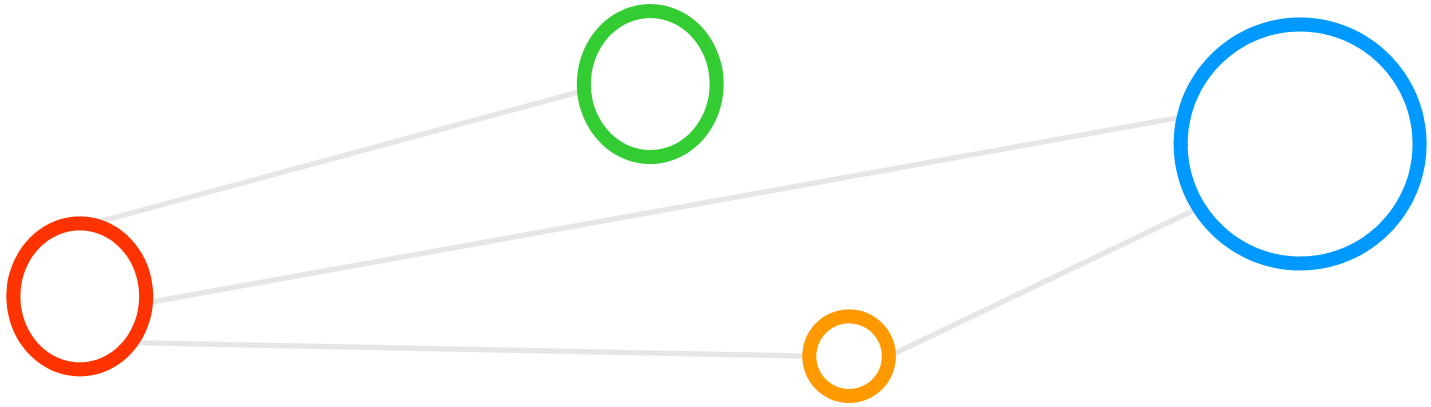
Epilogue

- Informal final lecture
  - Answering remaining questions & guidance to future topics
  - Summary & preparation for final exam and quizzes debrief
- Mindset
  - Discussion of job offers on the market in the light of the course
  - What we have learned & how to turn knowhow into action
- Skillset
  - Knowledge of various Cloud computing & machine learning skills
  - PHD positions & Master Thesis topics in learning from ‘big data‘
- Toolset
  - Knowledge of cloud services & machine/deep learning libraries
  - Future Topics to study: Quantum computing, neural networks on the chip, neuromorphic computing, modular supercomputing

[Video] Deep Learning Applications

[12] YouTube, What is Cloud Computing

Lecture Bibliography



Lecture Bibliography (1)

- [1] Piazza Web page, Online: https://piazza.com/
- [2] Gradescope Web page, Online: https://www.gradescope.com/
- [3] Morris Riedel Web page, Online: http://www.morrisriedel.de
- [4] Helmholtz Association Web Page, Online: https://www.helmholtz.de/en/
- [5] Forschungszentrum Juelich Web page, Online: http://www.fz-juelich.de
- [6] University of Iceland – School of Engineering and Natural Sciences Web page, Online: https://english.hi.is/school_of_engineering_and_natural_sciences
- [7] DEEP Series Projects Web Page, Online: http://www.deep-projects.eu/
- [8] C. Bodenstein, M. Goetz, M. Riedel, 'Automated Soccer Scene Tracking using Deep Neural Networks', Poster IAS Symposium, Online: https://www.researchgate.net/publication/328997974_Automated_Soccer_Scene_Tracking_Using_Deep_Neural_Networks
- [9] Helmholtz AI Web page, Online: https://www.helmholtz.ai/
- [10] Deep Learning Start-Up Beispiel Deutschland, Online: https://soccerwatch.tv/
- [11] YouTube Video, 'The Deep Learning Revolution', Online: https://www.youtube.com/watch?v=Dy0hJWltsyE

Lecture Bibliography (2)

- [12] YouTube Video, 'What is Cloud Computing', Online: https://www.youtube.com/watch?v=dH0yz-Osy54
- [13] UGLA Cloud Computing & Big Data Course Web page, Online: https://ugla.hi.is/kennsluskra/index.php?sid=&tab=nam&chapter=namskeid&id=70881020206
- [14] Book Cloud Computing for Science and Engineering, Online: https://cloud4scieng.org/
- [15] K. Hwang, G. C. Fox, J. J. Dongarra, 'Distributed and Cloud Computing', Book, Online: http://store.elsevier.com/product.jsp?locale=en_EU&isbn=9780128002049
- [16] Apache Spark Web page, Online: https://spark.apache.org/
- [17] Apache Hadoop Web page, Online: http://hadoop.apache.org/
- [18] Keras Python Deep Learning Library, Online: https://keras.io/
- [19] Tensorflow Deep Learning Framework, Online: https://www.tensorflow.org/
- [20] Big Data Tips 'Association Rules', Online: http://www.big-data.tips/association-rules
- [21] OpenStack Web page, Online: https://www.openstack.org/
- [22] Apache Flink Web page, Online: https://flink.apache.org/

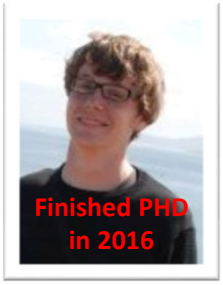
Lecture Bibliography (3)

- [23] Book Cloud Computing for Machine Learning and Cognitive Applications, Online: https://mitpress.mit.edu/books/cloud-computing-machine-learning-and-cognitive-applications
- [24] Big Data Tips, Online: http://www.big-data.tips
- [25] Datanami Web page, Online: https://www.datanami.com/
- [26] HPCWire Web page, Online: http://www.hpcwire.com
- [27] Mining of Massive Datasets, Online: http://infolab.stanford.edu/~ullman/mmds/book.pdf
- [28] European EuroHPC Joint Undertaking, Online: https://eurohpc-ju.europa.eu/
- [29] European Open Science Cloud (EOSC) Web page, Online: https://www.eosc-portal.eu/
- [30] EU EOSC-Nordic Project Web page, Online: https://www.eosc-nordic.eu/
- [31] Morris Riedel, 'Deep Learning - Using a Convolutional Neural Network', Invited YouTube Lecture, University of Ghent, 2017, Online: https://www.youtube.com/watch?v=gOL1_YIosYk&list=PLrmNhuZo9sgZUdaZ-f6OHK2yFW1kTS2qF
- [32] M. Riedel et al., 'Introduction to Deep Learning Models', JSC Tutorial, three days, JSC, 2019, Online: http://www.morrisriedel.de/introduction-to-deep-learning-models
- [33] H. Lee et al., 'Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations', Online: http://doi.acm.org/10.1145/1553374.1553453

Lecture Bibliography (4)

- [34] YouTube Video, 'Neural Network 3D Simulation', Online: https://www.youtube.com/watch?v=3JQ3hYko51Y
- [35] A. Rosebrock, 'Get off the deep learning bandwagon and get some perspective', Online: http://www.pyimagesearch.com/2014/06/09/get-deep-learning-bandwagon-get-perspective/
- [36] Big Data Tips – Big Data Mining & Machine Learning, Online: http://www.big-data.tips/
- [37] NVIDIA Web Page, Online: https://www.nvidia.com/en-us/

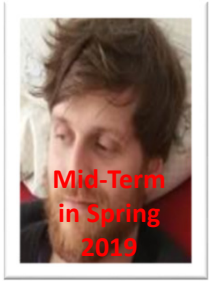
Acknowledgements – High Productivity Data Processing Research Group



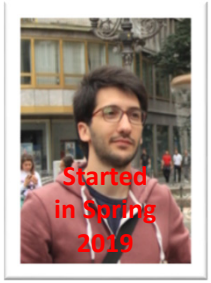
PD Dr.
G. Cavallaro

Senior PhD
Student A.S. Memon

Senior PhD
Student M.S. Memon



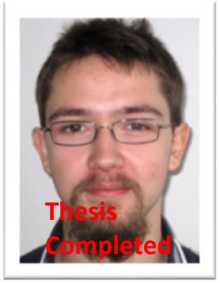
PhD Student
E. Erlingsson



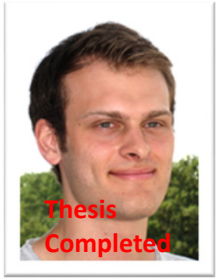
PhD Student
S. Bakarat

PhD Student
R. Sedona

Dr. M. Goetz
(now KIT)



MSc M.
Richerzhagen
(now other division)



MSc
P. Glock
(now INM-1)

MSc
C. Bodenstein
(now
Soccerwatch.tv)

MSc Student
G.S. Guðmundsson
(Landsverkjun)

This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project)

services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
different
Cross-Disciplinary
methods
High
big
nodes
PB
community
Environment
Modelling
EU
Science
centers
Key
scientific
important
files
model
international
storage
basis
Infrastructure
etc
DKRZ
Use
simulations
ENES
technologies
formats
engineering
References
brain
increasing
via
EUDAT
physical
IPCC
Big
Data
Project
processing
Common
TB
SMAQ
Methods
MapReduce
create
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
Large
Summary
data-intensive
Service
Scientific
often
stored
Resources
analyze
levels
DLCL
Level
external
JSC
use
computational
Research
Matters
Concrete
performance
research
e.g
science
HPC
Juelich
system
datasets
acquisition
compute
Landscape
management
System
Climate
modelling
AR5
Hardware
large-scale
cores
access
hundreds
Nutshell
Services
Technologies
Earth
Structure
network
DLCLs
Understanding
structures
directory
distributed
project
General
current
Germany
Simulation
Processing
Provide
NASA
Energy
systems
day
Health