



Aalto University
School of Science

Big Data and Large-scale Computing Track

Bo Zhao, Department of Computer Science

bo.zhao@aalto.fi, <https://zbjob.github.io/>

Large-scale computing

- **Computing with large-scale ICT infrastructures**
 - on-premise enterprise data centers, high-performance clusters, public cloud data centers, distributed edge systems, and hybrid edge-cloud systems ⇒ heterogeneous computing, memory and networking capabilities
- **Solving problems with a wide range of complex workloads in complex applications and services**
 - long-running, data intensive processing
 - massive and high performance computation
 - large-scale and fast (streaming) ingestion, processing & reasoning
 - complex data workflows for AI-ready data
 - reliable platform services for a very large number of tenants & users

Big/Extreme Data

Data sources

- Social media and business data
- Internet of Things (IoT), Industrial IoT, machine-to-machine (M2M), and manufacturing
- Advanced scientific instruments, e.g., earth observatory, particle accelerators & DNA sequencer
- Personal information and healthcare data
- Open/government data
- Technical documents/writings/texts

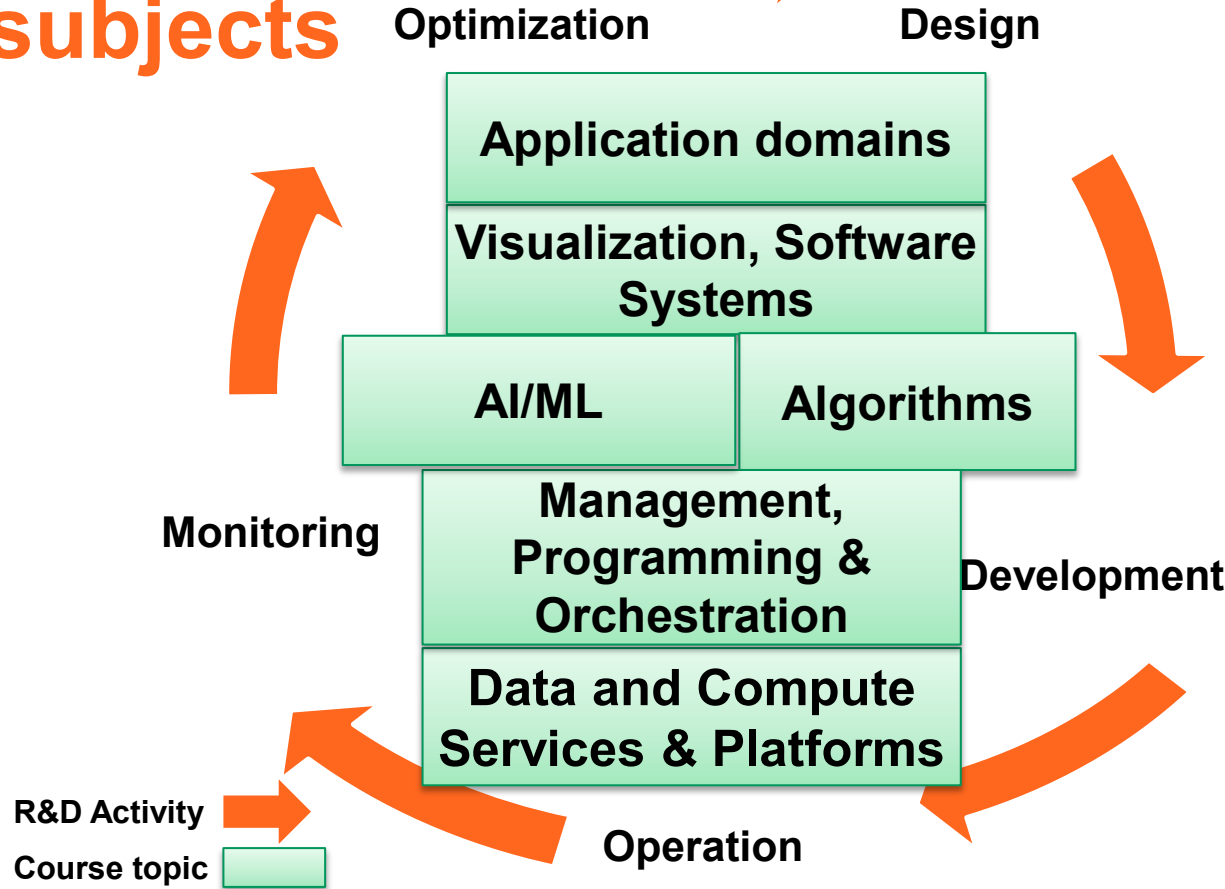
Data characteristics

- Extremely large, complex, dispersed, heterogeneous data from different sources
 - need new techniques/methods
- Diversity and complexity in terms of
 - volume/size (large datasets or many small datasets)
 - complexity (formats, types of data)
 - speed (generating speed, data movement speed)
 - quality, bias, ownership, and regulations

Big data and large-scale computing track

- **Foundations for data- and compute-intensive services and applications**
 - data systems and data platforms (customers, IoT, IIoT, software, sciences), AI/ML/LLMs systems and services, HPC for sciences (simulation, prediction)
- **Wide range application domains**
 - Business analytics, Healthcare, Software Development, Industry 4.0, Smart City, E-Science, Critical infrastructures operations, Pandemic/Disaster responses, Sustainability, etc.

Key study subjects



Main courses

CS-E4890 Deep Learning
CS-E4500 Advanced Course in Algorithms
CS-E4660 Advanced Topics in Software Systems
CS-E4840 Information Visualization
CS-E4950 Software Architectures
CS-E4770 Designing and Building Scalable Web Applications

CS-E4645 Research Project on Data Intensive Computing
CS-E4650 Methods of Data Mining
CS-E4641 Principles and Techniques of Data Platforms
CS-E4690 Programming Parallel Supercomputers
CS-E4580 Programming Parallel Computers
CS-E4190 Cloud Software and Systems
CS-E4780 Scalable Systems and Data Management

Access to and practice with state-of-the-art infrastructures

- **Computing and data infrastructures**

- Aalto SCI School's high-performance computing systems
 - <https://scicomp.aalto.fi/triton/>
- Finland's national supercomputing systems and Clouds
 - <https://research.csc.fi/en/computing>
 - LUMI (<https://www.lumi-supercomputer.eu/>)
- Course-specific cloud resources may also be available from Google Cloud, Microsoft Azure, and Amazon EC2

- **Data**

- Open and research datasets

Personalizing your study

- **Study subjects in the track can be challenging**
 - complex and challenging software/infrastructure technology stacks
 - engineering emphasis
- **Different emphasis on skills building**
 - e.g., data engineer versus platform architect vs business intelligence analyst
- **Different industrial sectors/academia may influence your learning goals**
 - e.g., manufacturing vs e-health vs enterprise business vs e-science