https://www.projectmates.eu/

MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Introduction to Big Data Infrastructure Technologies for Data Analytics

Introduction: MATES Project and course overview

Yuri Demchenko MATES Project

University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Co-funded by the Erasmus+ Programme of the European Union

4 Days/Sessions course – Logistics

- 4 Sessions of 3 hours, including 2 breaks 10 min
- Time 15:00-18:00, online
- Training course information page
  - To be created at Github
- Zoom class
  - https://uva-live.zoom.us/j/88358719282?pwd=SkFOdi9qUm9WOG93OFhpd1pvY2Y1UT09
- Course materials uploaded to GoogleDrive folder
  - https://drive.google.com/drive/folders/1su2P7NqDF24MJKnfMqtREseK4iH5HWxY
- Lectures will be recorded and uploaded after lecture overnight, and updated after processing - approx. 2-3 days
- Discussion and Q&A via Zoom chat

Co-funded by the
Erasmus+ Programme
of the European Union

Daily Topics

- **Day 1**
  - **Course introduction: MATES project, Industry 4.0 and digitalisation in maritime industry**
  - **Introduction to Big Data concepts, architecture and technologies, Use cases**
  - **Cloud Computing basics**
  - **Demo and practice: working with AWS cloud services**
- Day 2
  - Big Data algorithms, Hadoop Big Data Platform
  - Cloud based Big Data platforms and Providers
  - Demo and practice: working with cloud services and Hadoop cluster
  - Discussion: Big Data and Digitalisation aspects in your organisations
- Day 3
  - SQL and NoSQL databases for Big Data
  - Data Stream analysis, Hadoop Kafka and Flume
  - Discussion: Defining Big Data infrastructure for your organisations
- Day 4
  - Big Data Security and Compliance
  - Practice: Working with the Compliance Assessment Matrix

Expected Outcome

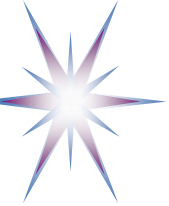
- End of module review and feedback
- Certificate of attendance upon completed assignment and feedback
- All videos will be available overnight and after the course
- All materials will be available at the github workshop page and linked from the MATES project website

- Your contribution to **Blue Survey** by Evelyn Parades, Uni Gent, available in 11 languages - http://thebluesurvey.eu/
  - To investigate the link between maritime professionals and the ocean, with a focus on sustainability

Outline Day 1, Part 1

- Industry 4.0 and Maritime Industry Digitalisation
- Demand for Digital and Data competences and Skills
  - Competences and skills for data driven organisations
  - Data literacy and Transversal skills
- MATES Project contribution
  - ED2MIT – Training on Digital and Data Skills and Literacy
- Discussion

Co-funded by the
Erasmus+ Programme
of the European Union



Yuri Demchenko,

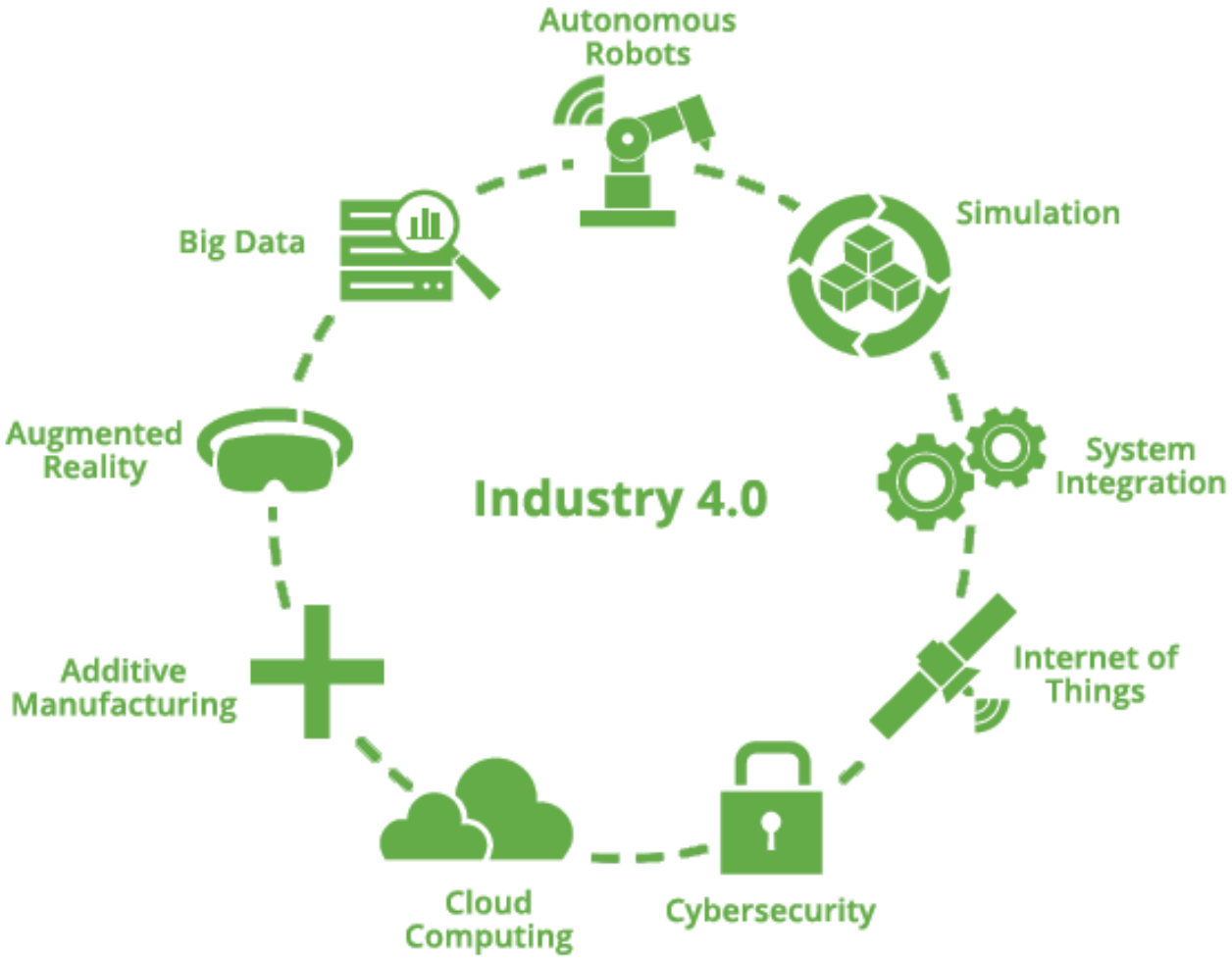
Senior Researcher, Lecturer, Complex Cyber Infrastructure Group, UvA

- Graduated and PhD from National Technical University of Ukraine “Kiev Polytechnic Institute”
  - University of Amsterdam – since 2003
- Research areas
  - Big Data Infrastructure and Data Science platforms
  - Cloud architecture framework, DevOps and cloud automation platform
  - Cloud security and compliance
- Teaching courses (on campus and online)
  - **Big Data Infrastructure and Technologies for Data Analytics (BDIT4DA)**
  - Cloud Computing Engineering (CCENG), Security Engineering (SECENG))
  - DevOps and Cloud based Software Development (DevOps)
  - **Web technologies and Databases (WebDB)**
  - **Data Science and Analytics Foundations (DSAF), Professional Issues in Data Science**
- Recent projects
  - EDISON: Building the Data Science Profession for Europe
  - **MATES: Digitalisation of the European Blue Economy**
  - FAIRsFAIR/EOSC: FAIR Data Management and Data Stewardship
  - GEANT4 Research: Cloud aware networking infrastructure provisioning on-demand
  - SLICES-DS: Research Infrastructure for ICT

Digital and Data Training: Goals

- This is the first in series of ED2MIT training webinars
  - Goal: Obtain feedback via teaching
- Meet maritime experts and community and obtain feedback how to deliver digital and data technologies to wide community of maritime experts and community – to achieve digital and data literacy
  - Potentially find cooperators to transfer knowledge: train the trainers
- Provide basic information on Big Data and Data Analytics technology and tools
- Build trust in data-driven and digitalized operations by providing assurance of data quality, algorithms, sensors, systems and cyber security
- Facilitate digital readiness and transformation, understand benefits

Industry 4.0 and demand for new skills



The Fourth Industrial Revolution, which includes developments in previously disjointed fields such as artificial intelligence and machine-learning, robotics, nanotechnology, 3-D printing, and genetics and biotechnology, will cause widespread disruption not only to business models but also to labour markets over the next five years, with enormous change predicted in the skill sets needed to thrive in the new landscape. This is the finding of a new report, *The Future of Jobs*, published today by the World Economic Forum.

Maritime Industry Digital Transformation and Skills Strategy – Toward Industry 4.0

Digital Transformation

- Digitalisation and IoT
- Intelligent Information
- Data Management
- Digital Assets Manage
- Data Driven Optimisation
- Agile Continuous Improvement
- Customer Experience
- People and skills

Digital Competences/Skills

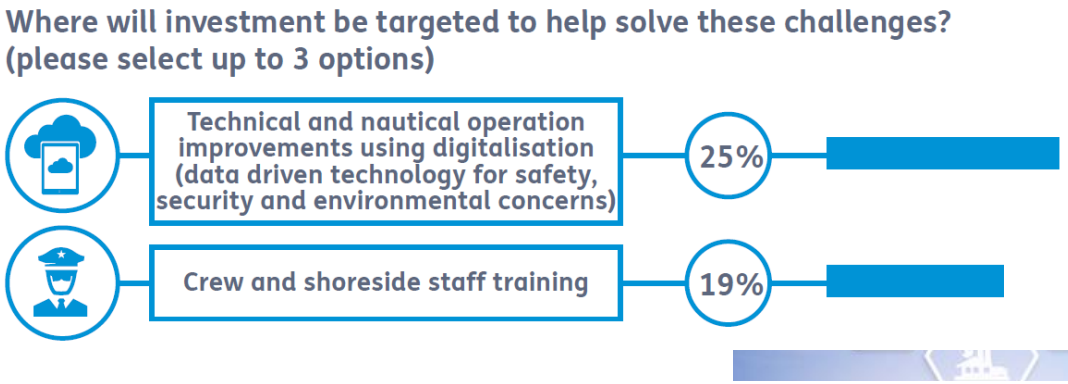
- Automation, robotics, electrical vehicles
- Information and data literacy
- Communication and collaboration
- Digital content creation, safety
- Problem solving and critical thinking

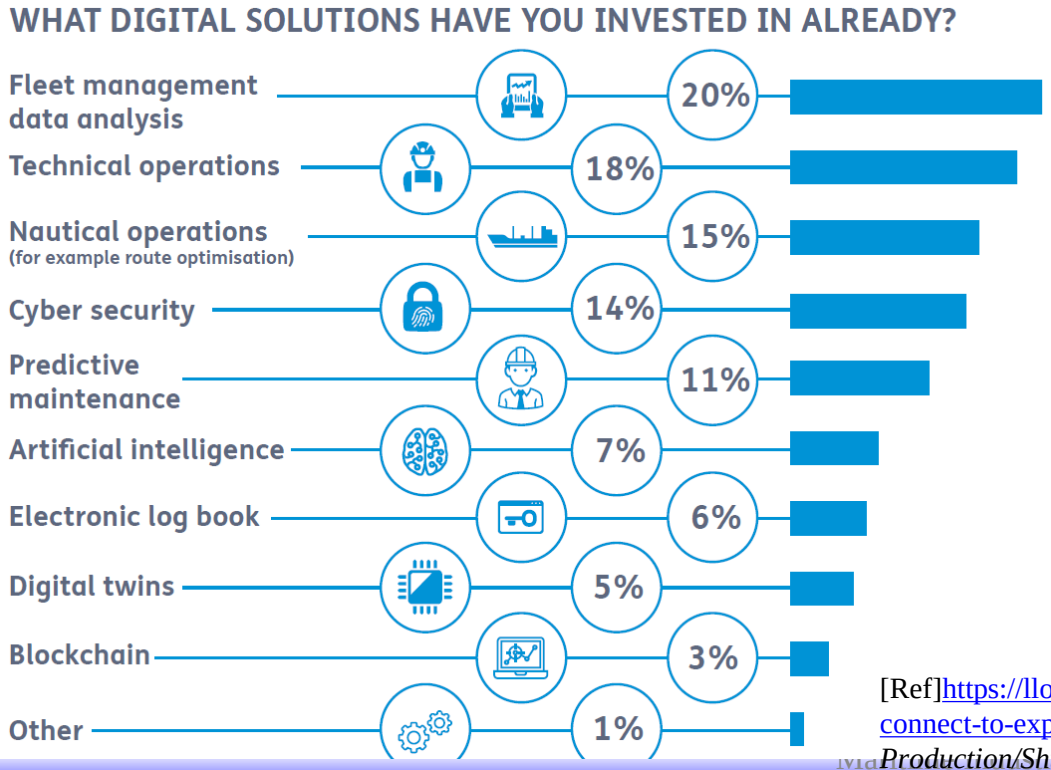
Stages of digitalization in maritime transport

The effects of digitalization on maritime transport can be divided into the following three stages:

1. Optimization – maximizing efficiency and reliability in existing processes to reduce the costs of trading.
2. Extension – moving beyond efficiency to produce opportunities for new services and businesses.
3. Transformation – reinventing logistics, trade and business models, based on data-driven revenue streams and shifts in trade flows.

Digitalisation is game changer in shipping (Wartsila report)





- The benefit of technology is in value it brings to the business, and that value is in proportion to Leaders' understanding
- Staff education and training is a key to adopt new technologies

[Ref]https://lloydslist.maritimeintelligence.informa.com/LL1128397/Click-and-connect-to-explore-shippings-gamechanger *Source: Shinonome Production/Shutterstock.com*

Opportunities with Digitalisation

- Cross-sector sharing
- Catch up with the IoT data tsunami
  - Sensors network
  - Robotics
  - Vehicles connectivity
- Starting data pipeline from IoT and edge computing
  - Early problems identification

Data driven culture in enterprise

Building a Data-Driven Culture in Enterprise: Introducing the 4 Pillars

- **Data maturity.** Solid data maturity is foundational to a data culture.
  - Organization's data maturity manifests itself in every individual in your organization having an easy and appropriate level of access to the clean and accurate data they need.
  - Importance of a well-defined CDO role and other related roles
- **Data-driven leadership.** Leaders define the culture of their organization.
  - A data-driven leader supports a culture of data by demonstrating data-driven decision making and involve the team members
  - A data-driven leader sees data as a strategic asset and makes "think and act data" a key strategic priority.
- **Data literacy.** Individual decision makers must be data literate to leverage their data
  - The CDO office needs to invest in enterprise wide data literacy, where every role is upgraded with the right level of data science skills.
- **Data-driven decision-making processes.** Establish a structured process of forward-looking decision making and backward-looking reviews of decisions.
  - Build experience of aligning data analytics, insight and data-driven decision-making processes.

Five Characteristics of a Data-Driven Company

#1. Creative executives who run their businesses with passion and curiosity

- Being data-driven requires a bit of a researcher's mindset -- the curiosity to dig into the data and glean insights from it that can be of use for the business.

#2. Data democratization

- Data-driven organizations emphasize the importance of broad data access for all employees.

#3. Data literacy

- An organization's ability to succeed in the digital era is heavily dependent on its employees' data literacy: the ability to read, work, analyze, and argue with data.
- Example of how to respond to a data literacy problem is Data University at Airbnb. Airbnb could not have a data scientist in every room to inform every decision with data.

#4. Automation of data management workloads

- A core criterion for a data-driven organization is how much data analytics tools are automated and provide information is a form that can be easy for decision making

#5. A companywide, data-driven culture

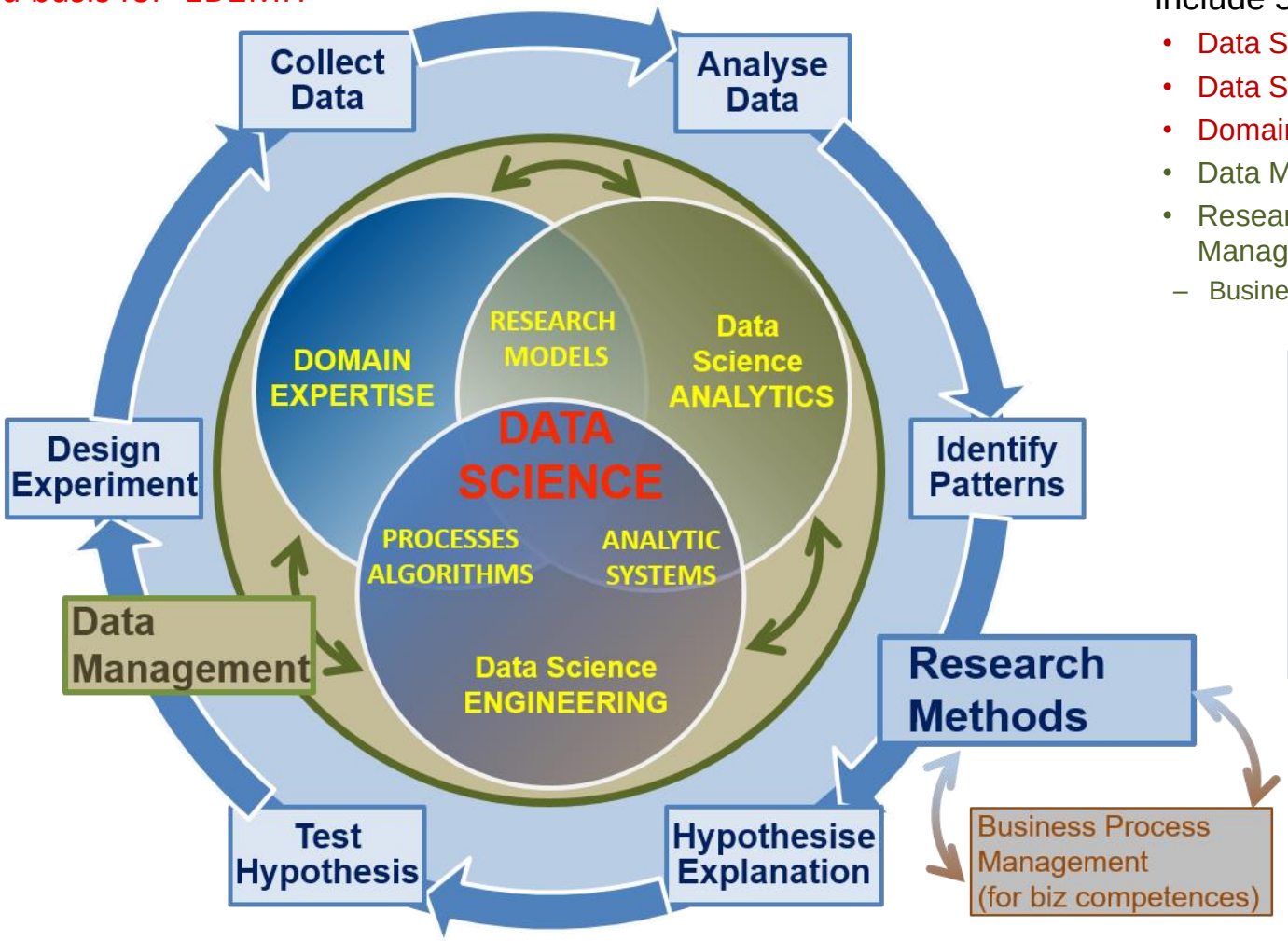
- Becoming data-driven involves more than technology and tools. It also requires a shift in the enterprise's mindset and culture.
- IBM example: Require all analytics blending with open data (social media, weather, climate etc.)

MATES Project contribution

- Study and report on demanded skills in maritime industry
- Pilot experiences and Roadmap
  - Pilot Experience ED2MIT - Education and Training for Data Driven Maritime Industry
- Leverage EDISON Data Science Framework (EDSF) that created a basis for Big Data, Data Science and digital and data skills education and training
  - EDSF provides a basis for ED2MIT training program
- Leverage European Digital Competence framework DigCom2.1
- Provide training on digital and data technologies
  - Train trainers, create pool of reference training materials

EDISON Project (2015-2017): Data Science Competence Groups - Research

EDISON Data Science Framework (EDSF) provides a basis for ED2MIT



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design

Discussion

- Digitalisation aspects in your organisations
  - Go to **www.menti.com** and use the provided code
- Acquiring new knowledge and skills
  - Balancing or combining formal, guided, and self-study
- Resources for training and self-training

What is your experience or preference in acquiring new knowledge and skills?

- Attend formal education or training
- Count on workplace training and courses
- Attend vocational training or formal online courses
- Mostly use available training materials for self-training
- Need advisor or tutor for self-education/self-training
- Count on workplace training and courses
- Have you used paid online training services like LinkedIn Education, Pluralsight, paid MOOCs, others?

Contribute to Blue Survey

- **Blue Survey** by Evelyn Parades, Uni Gent, available in 11 languages - http://thebluesurvey.eu/
  - To investigate the link between maritime professionals and the ocean, with a focus on sustainability

Acknowledgement

- This work is supported by the ERASMUS+ MATES project

- The work is committed to the Open Source under Creative Commons 4.0 CC BY License

Co-funded by the
Erasmus+ Programme
of the European Union

Additional materials

- ED2MIT Training on digital and data skills

ED2MIT: Digital and Data Competence Goups

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

A. Data related competences and technologies

B. Cloud services and cloud economics

C. Digital content creation, access and management

D. Data Science and Big Data Analytics

A. Data related competences and technologies

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

A. Data related competences and technologies

A.1. Big Data definition and technologies: 6V of Big Data and challenges for companies. Big Data examples from research and industry

A.2. Data collection, access and sharing

A.3. Data formats, data models, metadata

A.4. Data Storage and databases, SQL scripting and simple commands

A.5. Data inspection, Data protection, data backup and archiving

A.6. Cloud based services and tools for data storage, sharing and management

A.7. Open Data repositories, test datasets, developer communities

A.8. Organisational and private Data Management, FAIR Data Principles, organisational roles, Data Stewards

B. Cloud services and cloud economics

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

B. Cloud services and cloud economics

B.1. Cloud service models: IaaS, PaaS, SaaS, Apps. Use examples and Cloud Service Providers. Cost model of cloud services.

B.2. Company IT infrastructure migrating to cloud: benefits and problems

B.3. Cloud and Big Data, Cloud based Big Data platform and services

B.4. Data storing, backing up, sharing and processing in clouds (for organisational and private data)

B.5. Practical exercises with cloud services: Cloud management console and cloud services deployment and access.

C. Digital content creation, access and management

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

C. Digital content creation, access and management

- To be acquired as self-study or expected to be known

C.1. Internet and World Wide Web

C.2. HTML, CSS, JavaScript for active pages

C.3. UX design and web portal services

C.4. Web applications security

C.5. PHP and interactive web sites (advanced)

To be provided as self-study materials

D. Data Science and Big Data Analytics

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

D. Data Science and Big Data Analytics

This course is provided as a general overview of the listed below topics. More in-depth training and learning will require more time commitment and pre-requisite knowledge.

D.1. Statistical methods and Probability theory

D.2. Data description and Statistical Data Analysis

D.3. Data preparation: data loading, data cleaning, data pre-processing, parsing, transforming, merging, and storing data

D.4. Qualitative and Quantitative data analysis

D.5. Classification: methods and algorithms

D.6. Cluster analysis basics and algorithms

D.7. Performance of data analytics algorithms and tools

D.8. Visualizations of data analysis and dashboards

D.9. Organizing data analytics process following CRISP-DM and Data Science Process

Big Data and Cloud Resources

- Starter cloud accounts
- Learning resources
- Datasets

General Cloud Starter Accounts

- AWS
  - 100 USD free credits for 1 year
  - Free tier services for 1 year
- Azure
  - 200 USD for 1 month
  - Free tier services for 1 year
- Google Cloud Platform
  - 300 USD for 1 year
  - Generous free tier services for 1 year

Individual Educational Accounts - AWS

- You are provided with the AWS Education Class account
  - All individual accounts with 50$ credit
  - Limits and quotas for AWS resources applied
- Consider applying for individual AWS educational account as a Student https://aws.amazon.com/education/awseducate/
  - Use your university email address
  - You will have option at some stage to provide your creditcard number or not. The difference is in amount of free credits you will receive.
    - Even if you provide your creditcard, you will not be charged if you exceed your free credits
  - The benefit is that you will have much freedom in experimenting with AWS services.

Individual Educational Accounts - Azure

- Apply for free educational account
  https://azure.microsoft.com/en-us/offers/ms-azr-0170p/
  - You must use your UvA email address
- Azure for Students gets you started with
  - $100 in Azure credits to be used within the first 12 months
  - Select free services without requiring a credit card at sign-up
- Free developer tools
  https://portal.azure.com/#blade/Microsoft_Azure_Education/EducationMenuBlade/overview
  - Access to professional developer tools including Azure DevOps, Artificial Intelligence (AI), Machine Learning, and other areas
- Access free learning paths and labs by Microsoft Learn
  - https://docs.microsoft.com/en-us/learn/
- Read Acceptable Use policy: Prohibited use
- Limits and Quotas apply
  - You must upgrade your Azure Free Account to a Pay-As-You-Go subscription in order to increase quotas or limits.

Where to find more information

- First, revisit your lecture, follow recommended links
- Search web, use Wikipedia and other *pedia as a first step, always check original sources
- Use online courses and tutorials
- Use developer forums to ask questions and search for answers
  - Stackoverflow - https://stackoverflow.com/
  - https://softwareengineering.stackexchange.com/
  - https://dev.to/
  - https://www.experts-exchange.com/
  - https://www.quora.com/

Educational materials

- Available on Canvas
  - Recommended literature and links to DevOps related resources and tools
  - Some DevOps related papers and studies on Canvas
- Search web, use Wikipedia and other *pedia as a first step, always check sources
- Online courses and tutorials
  - AWS learning resources
    https://aws.amazon.com/training/self-paced-labs/
    https://aws.amazon.com/getting-started/
  - Microsoft Learn https://docs.microsoft.com/en-us/learn/
    - Multiple Learning paths, including certification
  - Linkedin Learning (one month free trial, 279$ annual fee)
    - https://www.linkedin.com/learning/
- Use developer forums to ask and search for answers
  - Stackoverflow - https://stackoverflow.com/
  - https://softwareengineering.stackexchange.com/
  - https://dev.to/
  - https://www.experts-exchange.com/
  - https://www.quora.com/

Online Educational and Training Resources

- LinkedIn Education – paid 279 USD/Yr
  - https://www.linkedin.com/learning/
- DataCamp paid 588 USD/Yr, 1 Mo trial, free starter python and R courses – growing popularity
  - https://www.datacamp.com/
- Pluralsight 275USD/Yr, 10 days trial
  - https://app.pluralsight.com/
- Intel AI Academy – Variety of course on ML, AI, NLP
  - https://software.intel.com/en-us/ai/academy
- Microsoft Learn (former Virtual Academy)
  - https://docs.microsoft.com/en-us/learn/
- AWS Training and Certification
  - https://aws.amazon.com/training/
- IBM Learning Services
  - https://www.ibm.com/services/learning/
- Kaggle Data Science Micro-Course
  - https://www.kaggle.com/learn/overview
- Coursera, Udacity
  - Variety but less practically focused
- Certification and training PMI, DAMA, IIBA, TDWI
  - Primarily oriented on practitioners and IT managers

Open Data, Educational Datasets and Professional Forums

- Registry of Open Data on AWS
  - https://registry.opendata.aws/
- Google Dataset Search (beta)
  - https://toolbox.google.com/datasetsearch
- Microsoft Research Open Data
  - https://msropendata.com/
- Github Open Data - https://github.com/collections/open-data
- Kaggle - https://www.kaggle.com/datasets
- UCI Machine Learning Repository
  - https://archive.ics.uci.edu

- KDNuggets – community forum and blog
- Stackoverflow – community Q&A forum
- Check for more https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research