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

Textbook chapter

Big Data for integrated Climate Change & Water Management

Project info	
Project title	Graduates for Climate Change adapted water management
Project acronym	CCWATER
Project reference number	619456-EPP-1-2020-1-NO-EPPKA2-CBHE-JP
Action type	Capacity Building in higher education
Web address	https://www.waterharmony.net/projects/ccwater/
Coordination institution	Norwegian University of Life Sciences (NMBU)
Project duration	15 January 2021 – 14 July 2024

Document control sheet	
Work package	WP2 Water & Climate change curriculum
Ref. no and title of task	Task 2.2.2 Development of textbook chapters
Title of deliverable	D2.2.2 Textbook chapters – Course 1
WP leader	Martin Oldenburg (THOWL)
Task leader	Ayurzana Badarch (MUST)
Author(s)	Chapter 1: K. Pilar von Pilchau (T OWL) Z. Maletskyi (NMBU) Chapter 2: H. Ratnaweera (NMBU Z. Maletskyi (NMBU) Chapter 3: X. Wang (QUT) H. Ratnaweera (NMBU) Chapter 4: M. Oldenburg (THOWL) H. Ratnaweera (NMBU) Chapter 5: A Badarch (MUST) H. Ratnaweera (NMBU) Chapter 6: K. Pilar von Pilchau (T OWL) A. Katarzyna Cuprys (NMBU)
Date	14.07.2024
Dissemination level	Public

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

Table of contents

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

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

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

List of figures

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

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

List of tables

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

1 Big Data for integrated Climate Change & Water Management

1.1 Introduction

1.1.1 Definition Big Data

The term big data means “masses of data”. It refers to data volumes that are so complex that they require new computer-based forms of processing. This is the opposite of “small data”, which means data volumes that a person can understand and process without help or with the aid of a simple computer program, or data volumes that refer to only one individual person or a narrowly restricted set of facts. Big data is also an umbrella term encompassing various activities associated with the enormous masses of data.

Big data means...
1. gathering a great deal of different data. Numerous sensors and programs capture and process information. These sensors and programs are built into or installed on tablets, smart-phones, or computers, for example. If the collected data are analog, they are digitalized.
2. saving the digitalized information as data in as structured a manner as possible. This results in complex databases, which are used to process the masses of data using computers.
3. linking the data and identifying new correlations by means of data analysis. Thanks to the structuring, it is possible to make targeted queries in databases and compare very specific aspects of databases. Often, assumptions or theories are confirmed or disproved based on the data analysis.

The three Vs and their meaning
Experts frequently describe big data with the three Vs, which are intended primarily to specify the “big” in the name in more detail and define the term more precisely. These three terms begin with the letter V.
Volume: This is the amount of data that is processed. With modern methods it is possible to manage enormous volumes of data. A vast number of computers or – expressed more neutrally – nodes is necessary for this purpose. This is the only way the masses of data can be processed in a structured manner to ultimately yield useful results.
Velocity: This is the speed at which the large volumes of data are processed. Often, a big-data method is beneficial only if the result is output sufficiently quickly.
Variety: This refers to the different types of data. They are compared or linked with each other. They must originate from many different sources, for example, be captured at different times, or be generated at different locations around the world. Only then do they have broad validity.

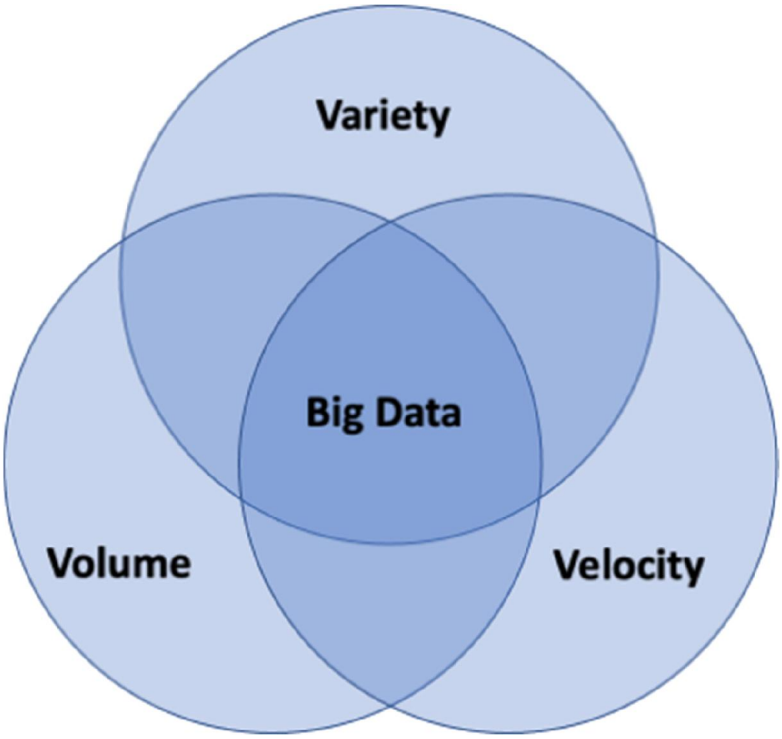


Figure 1: Relation between Big Data and the three Vs

1.1.2 Different Types of Data

Based on Balamurugan et al. (2021) data can be generated in different ways. For example, it can be generated by machines or by people. Human-generated data is the data that is created as a result of human interaction with machines. Examples include emails, documents, posts on social media channels. Machine-generated data is data that is generated by computer applications or hardware devices without active human intervention. Examples include sensor data, data from disaster warning systems, weather forecasting systems or satellite data.
Machine-generated and human-generated data can be subdivided as follows:

- Structured Data
- Unstructured data
- Semi-structured data

Structured Data

Structured data is characterised by the following properties, among others. They have a specific format and length, are easy to store and analyse, and have a high degree of organisation. This makes it possible to organise the data systematically and also to query it specifically.
A typical example of structured data is relational databases such as Structured Query Language (SQL) or Access, which contain organised numbers, dates, groups of words and numbers (strings/text) (Eberendu 2016).

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

Researcher ID	Researcher Name	Gender	Age
3974529	Tom	Male	46
3974530	Max	Divers	37
3974531	Luisa	Female	25
3974532	Mara	Female	18

Unstructured Data
Unstructured data is data that is raw unorganized and has no predefined schema. The data usually does not fit into relational database systems. Examples of unstructured data include video, audio, images, emails, text files and social media posts. The file type is often text files (e.g. emails, messages, documents) or binary files (e.g. audio, video, images) (Balamurugan et al. 2021).

Semi-Structured Data
Semi-structured data is irregular data that does not follow a formal structure. However, they have an internal processing structure that enables handling. This can change quickly or unpredictably and does not conform to any fixed or explicit schema. They may be server logs in comma-separated values (csv) format or documents in eXtensible Markup Language (XML), JavaScript Object Notation (JSON) and Binary JSON (BSON) format, etc. (Martinez-Mosquera et al. 2020).

```
<?xml version = "1.0"?>
<Institute>
<Researcher>
        <ResearcherID> 3974529</ ResearcherID >
        <FirstName>Sven</FirstName>
        <Gender>Male</Gender>
        <Age>46<Age>
</Researcher >
</Institute >
```

Figure 2:Example semi-structured data (based on Balamurugan et al. 2021)

1.1.3 Big Data Life Cycle

The figure illustrates the life cycle of Big Data and, like the following description, comes from reference (Balamurugan et al. 2021):
"Data arriving at high velocity from multiple sources with different data formats are captured. The captured data is stored in a storage platform such as HDFS and NoSQL and then preprocessed to make the data suitable for analysis. The preprocessed data stored in the storage platform is then passed to the analytics layer, where the data is processed using big data tools such as MapReduce and YARN and analysis is performed on the processed data to uncover hidden knowledge from it. Analytics and machine learning are important concepts in the life cycle of big data. Text analytics is a type of analysis performed on unstructured textual data. With the growth of social media and e-mail transactions, the importance of text analytics has surged up. Predictive analysis on consumer behavior and consumer interest analysis are all performed on the text data extracted from various online sources such as social media, online retailing websites, and much more. Machine learning has made text analytics possible. The analyzed data is visually represented by visualization tools such as Tableau to make it easily understandable by the end user to make decisions."

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

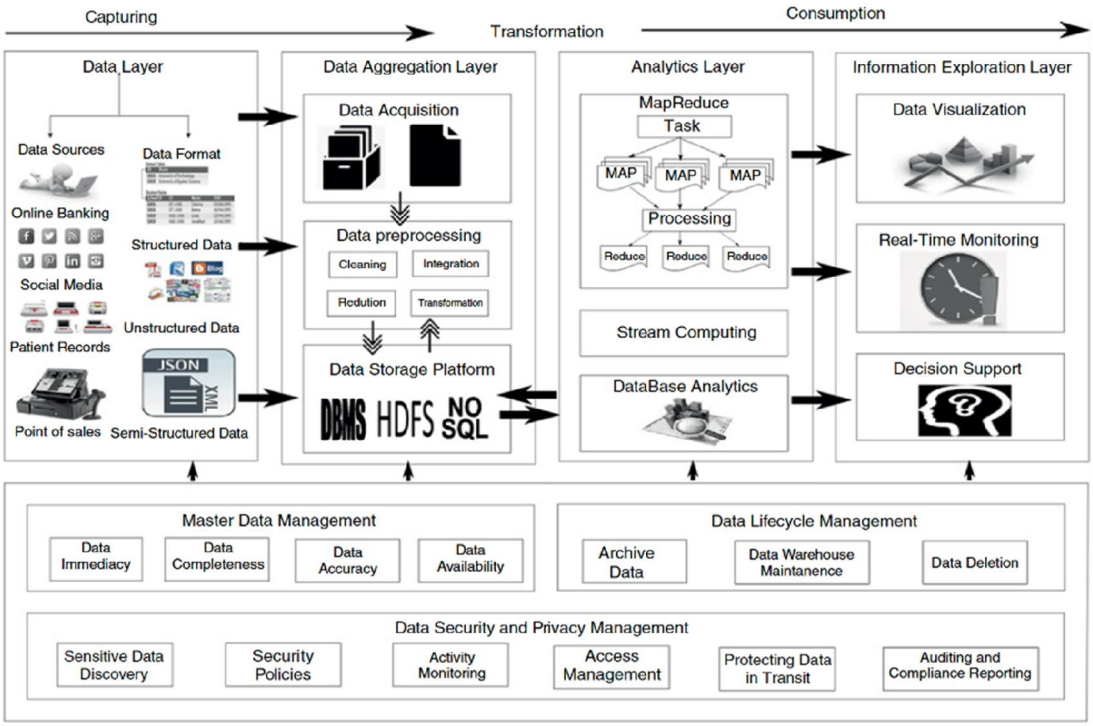


Figure 3: Life cycle of Big Data (Balamurugan et al. 2021)

1.1.4 Big Data Analytics

Big Data analytics focuses on examining or analysing large data sets with a variety of data types. The goal is to extract meaningful information to make better decisions, find new business opportunities, compete with competitors, improve performance and efficiency, and reduce costs (Balamurugan et al. 2021). Belonging to Balamurugan et al. (2021) the four types of analytics are:

- Descriptive Analytics:
  Descriptive analysis deals with data from the past. The aim is to learn from the past. The data is therefore described, summarised and visualised to make it interpretable. Descriptive analysis is often used in the analysis of consumer behaviour.
- Diagnostic Analytics:
  Diagnostic analysis also looks at data from the past and is about understanding what happened and why it happened. The aim is to take corrective action when something has gone wrong. It is a type of root cause analysis that identifies the factors that contributed to a particular outcome. The focus is often on finding hidden patterns in consumer data.
- Predictive Analytics:
  Predictive analytics focuses on the future. Based on patterns of historical data, possible scenarios in the future are identified. The goal is to determine results regarding opportunities or risks.
- Prescriptive Analytics:
  Prescriptive analysis also focuses on the future and aims to support decision-making. In addition to opportunities and risks, it also offers suggestions regarding benefits. Simulation and optimisation play a central role in predicting what will happen in the future, when it will happen and why it will happen. Accordingly, the impact of certain decisions can be considered.

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

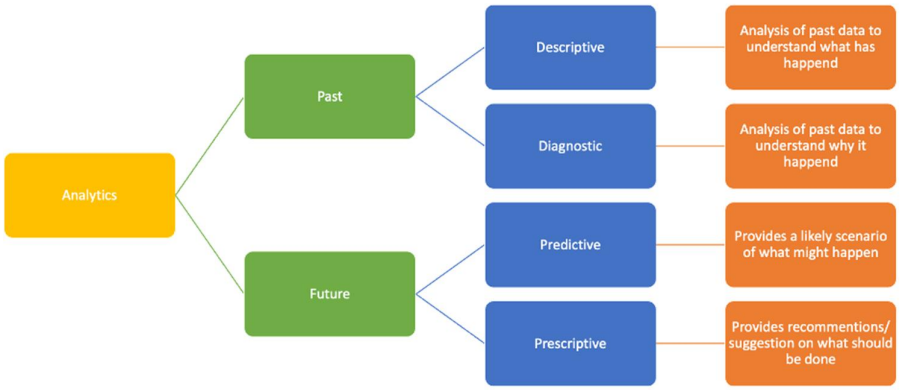


Figure 4: The four types of analytics (Balamurugan et al. 2021)

1.1.5 Big Data Analytics with Machine Learning

Artificial Intelligence

There are many definitions of artificial intelligence. Generally, it is considered that the core of AI is the research theories, methods, technologies and applications to simulate, augment and improve human intelligence. "AI" has become a buzzword in almost every aspect of our lives. The research areas of AI include systems and engineering, brain research, psychology, cognitive science, mathematics, computer science and many other fields. Examples of application areas of AI are for example image processing, speech recognition, robots, autonomous vehicles, energy systems (Jiang et al. 2022).

Machine Learning

ML is a sub-field of artificial intelligence. It is generally understood as the automated process of recognising ("learning") patterns in data. The goal can be classification and prediction. The focus is on the intelligent processing of information. An important starting point for ML is the selection of data to be used for learning patterns and the determination of variables to be used for classification or prediction. An often-mentioned representative is, for example, artificial neural networks (Black et al. 2022). There are different approaches to the machine learning process that can be distinguished from each other:

Supervised learning:

The chosen learning algorithm is to learn a relationship in the data based on training data. The correct output value is known during the learning process. The resulting model is then applied to a test data set that was not used for the learning process to check whether a generally valid rule was learned. Classic supervised learning problems are regression and classification (Verdhan 2020).

Unsupervised learning:

In unsupervised learning, the correct output value is not known during the learning process. The primary goal of this method is to recognise patterns in multidimensional space. In statistics, this is also called density estimation. Classic problems of unsupervised learning are clustering and dimensionality reduction (Verdhan 2020).

Reinforcement learning:

This category focuses on learning optimal action rules from a reward signal. Here, the focus is often on the sequence of several actions. The learning goal is the assessment of the quality of tactics and the ability to create one's own tactics (Li 2018).

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

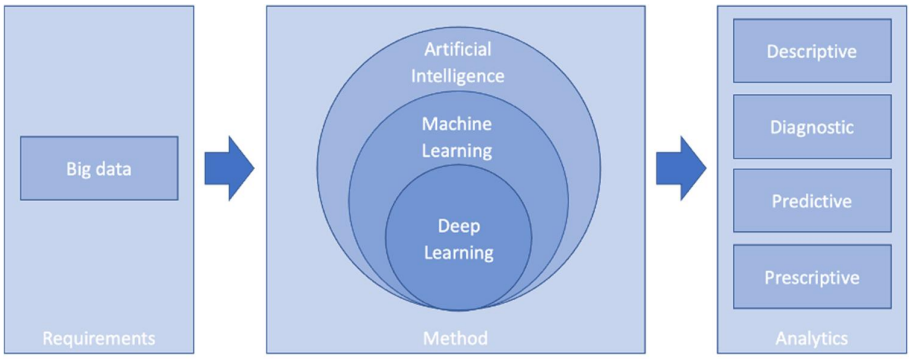


Figure 5: -Big Data and learning methods (Life cycle of Big Data (Jiang et al. 2022)

1.1.6 Excercise

Based on Sina Ike, Alexandra Daub, Lars Knieper, Sophie Potts;
Content licensed under CC-BY-SA (4.0) Creative Commons

The following link will take you to the learning module of the course Introduction to Bayesian Statistics and Statistical Learning:

http://openilias.uni-goettingen.de/openilias/ilias.php?ref_id=1607&obj_id=1&cmd=layout&cmdClass=illmpresentationgui&cmdNode=fw&baseClass=ilLMPresentationGUI

The aim of this module is to deepen the knowledge of Bayesian Statistics and Statistical Learning. To do this, open, preferably in a new tab, the links below, which will take you to various Shinyapps.

Maximum Likelihood Coin Toss
Univariate Distributions
Multivariate Normal Distribution
Linear Regression
Introduction to Bayes
Gibbs Sampler
Splines

Furthermore, each learning unit ends with a short quiz that you can answer using the respective shinyapp.

Have fun!

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

2 Big Data tools

2.1 Introduction to Big Data

Big data refers to the volume, velocity, and variety of data that artificial intelligence technologies are using to discover patterns and correlations hidden in massive collections of data. Big data is also commonly known as the three Vs. There are three types of big data, which are classified as structured data, unstructured data, and semi-structured data.

Big data is often used in products or services you use every day. Companies like Netflix or Procter & Gamble leverage big data to anticipate market demand. Predictive models are built for new products and services by classifying key attributes of past products then modeling the commercial success of those offerings. Big data enables companies to gather data from website visits, social media interactions, and ads you click on. The data is then used to make improvements to the customer experience in ways such as delivering personalized offerings in hopes to reduce customer churn. Likewise, there are many opportunities also for the water industry.

Big Data are different from other technologies & methodologies:

Big **data vs data mining**: Big data is a term that refers to a large amount of data whereas data mining refers to a deep dive into the data to extract the key knowledge/pattern/information from a small or large amount of data.

Big data vs data science vs data analytics: Big data refers to a large and complex collection of data. Data analytics is the process of extracting meaningful information from data. Data science is a multidisciplinary field that aims to produce broader insights.

Big data vs data warehouse: Big Data is a term applied to datasets whose size is beyond the ability of commonly used tools to capture, manage, and process the data within an acceptable elapsed time. Data-warehouse is a collection of data marts representing historical data from different operations in the company.

Big data vs statistics: Data Science frequently deals with large datasets, often referred to as big data. It uses techniques to handle and process substantial volumes of data efficiently. Statistics can work with both small and large datasets but is traditionally applied to smaller, carefully collected samples. Big data focuses on data analysis using algorithms and coding, while statistics relies on math and categorical data interpretation.

2.2 Introduction to AI, ML and DL

Artificial intelligence (AI) refers to the idea of endowing algorithms with the ability to perform tasks and make inferences that would require an intelligent human in the same position, while machine learning (ML) relates to intelligent systems that can adapt their behavior during the system-training stage to newly provided information (Riedl, 2019). Deep learning (DL) is a subset of machine learning that uses multilayered neural networks, called deep neural networks, to simulate the complex decision-making power of the human brain. Some form of deep learning powers most of the artificial intelligence (AI) applications in our lives today.

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

2.2.1 Artificial intelligence

Types of AI

Here are the types of AI that are popular and widely used for developing futuristic apps:

Reactive Machines: Machines that solely react. These systems don't keep track of past events or use them to inform current judgments.

Limited Memory: These systems use historical data, adding information over time. The details mentioned disappear quickly because when an ML model is created, it requires very limited memory to function properly.

Theory of mind: Systematic approaches to understanding human emotions and how they influence decision-making fall under the heading of the theory of mind. They are taught to modify the behavior as necessary.

Self-awareness: These systems were developed to be conscious of themselves. They are aware of their own internal states, can anticipate others' emotions, and behave appropriately.

Applications of artificial intelligence

- Machine Translation such as Google Translate
- Self-Driving Vehicles such as Google's Waymo
- AI Robots such as Sophia and Aibo
- Speech Recognition applications like Apple's Siri or OK Google

2.2.2 Machine Learning

Machine learning is a computer science field that makes use of computer algorithms and analytics to create forecasting models that can resolve business issues.

It analyses enormous volumes of data (both structured and unstructured) to forecast the future. It uses various algorithms and methodologies to learn from the data.

Machine learning applications

- Sales forecasting for different products
- Fraud analysis in banking
- Product recommendations
- Stock price prediction

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

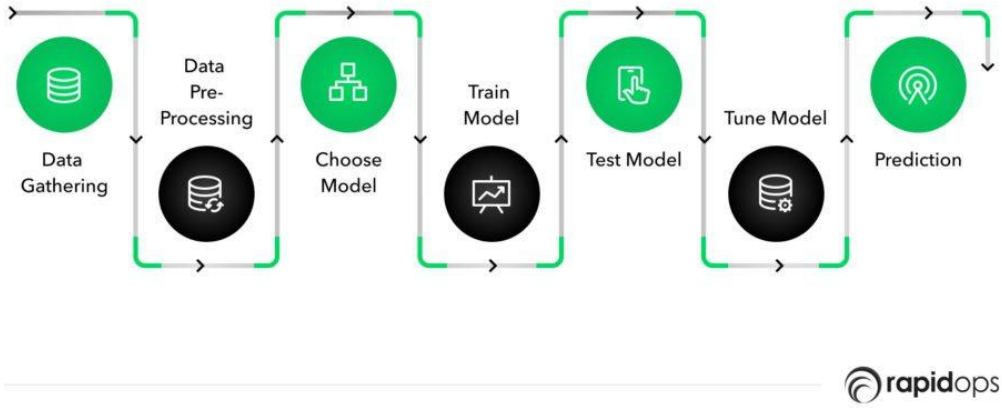


Figure 6:How a computer learns from data (rapidops, 2024)

2.2.3 Deep Learning

A branch of machine learning called "deep learning" works with algorithms modeled after the human brain's structure and operation.

Large amounts of both structured and unstructured data can be used by deep learning systems. Artificial neural networks enable machines to make judgments and are the foundation of deep learning.

Types of deep learning
Convolutional neural networks (CNN): A class of deep neural networks that are most frequently utilized for image analysis.

Recurrent Neural Network (RNN): RNN creates models using sequential data. It frequently performs better for models that must retain historical data.

Generative Adversarial Network (GAN): It is an algorithmic architecture that produces fresh, artificial data instances that can be mistaken for actual data using two neural networks. A GAN trained on images can create new images that, to human viewers, at least appear legitimate.

Deep Belief Network (DBN): A generative graphical model which is made up of several layers of latent variables known as hidden units. The units are not connected, but each of its layers is.

Deep learning applications

- Cancer tumor detection
- Captionbot for captioning an image
- Music Generation
- Image Coloring
- Object detection

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

2.2.4 Relationship among AI, ML, and DL

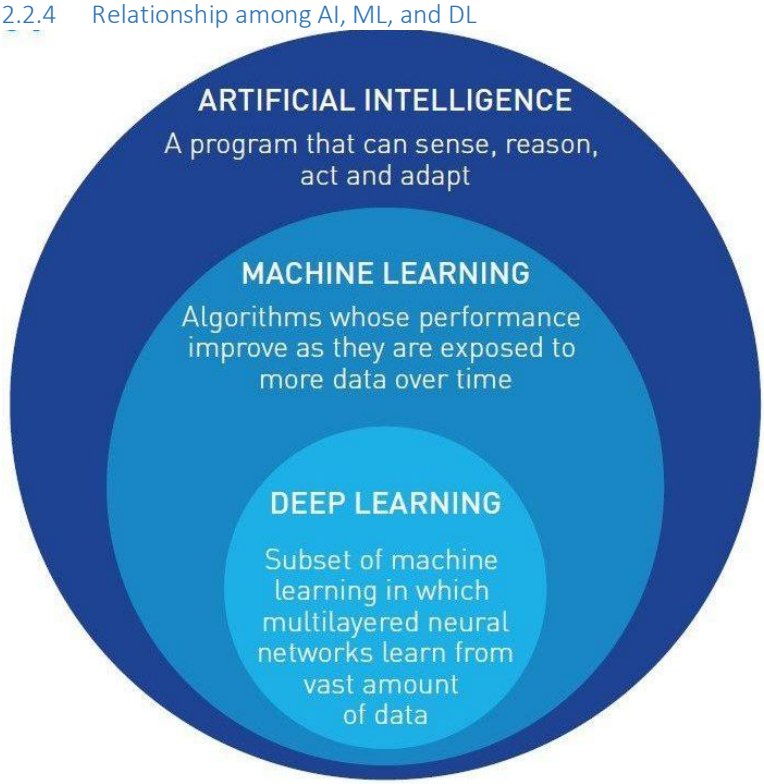


Figure 7: AI, ML and DL relationships

Artificial intelligence has the ability to imitate human behavior. Meanwhile, Machine learning applies AI to learn and adapt from experiences, which eventually helps in creating AI-driven applications. Deep learning applies machine learning to train a model that can deal with vast volumes of data.

Machine learning is the subset of AI and uses statistical models, while deep learning uses that to train their models that eventually will lead to solving problems like tumor detection, diagnosis, medical research, etc.

The main distinction between deep learning and machine learning is how data is delivered to the machine.

Deep learning networks use numerous layers of artificial neural networks, whereas machine learning techniques often need structured data.

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

Table 1: Difference between AI, ML, Deep Learning, Data Science, and Big Data (Rapidrops, 2024)

Topic	Artificial Intelligence	Machine Learning	Deep Learning	Big Data	Data Science
Definition	The idea of building intelligent machines is known as artificial intelligence.	A computer science field that makes use of computer algorithms and analytics to create forecasting models	A branch of machine learning called "deep learning" works with algorithms modeled after the human brain's structure and operation.	A method for gathering, preserving, and processing enormous amounts of data.	It is an area that involves gathering, handling, analyzing, and incorporating data into numerous procedures
Types	Reactive machines, limited memory, theory of mind, and self-awareness	Supervised, semi-supervised, unsupervised, and reinforcement	Convolutional neural networks, recurrent neural network, generative adversarial network, and deep belief network	Structured data, unstructured data, and semi-structured data	N/A
Application	Self-driving vehicles, AI robots, Apple Siri/ Google assistant	Sales forecasting, fraud detection, product recommendation	Music generation, cancer tumor detection, object detection	Recommend content on-demand, building learning management system	Speech recognition, advanced image recognition, airline route planning

rapidops

2.3 Introduction to machine learning

The tremendous amount of data being generated via computers, smartphones, and other technologies can be overwhelming, especially for those who do not know what to make of it. To make the best use of data researchers and programmers often leverage machine learning for an engaging user experience.

Many advanced techniques that are coming up every day for data scientists of all supervised, and unsupervised, reinforcement learning is leveraged often. In this article, we will briefly explain what supervised, unsupervised, and reinforcement learning is, how they are different, and the relevant uses of each by well-renowned companies.

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

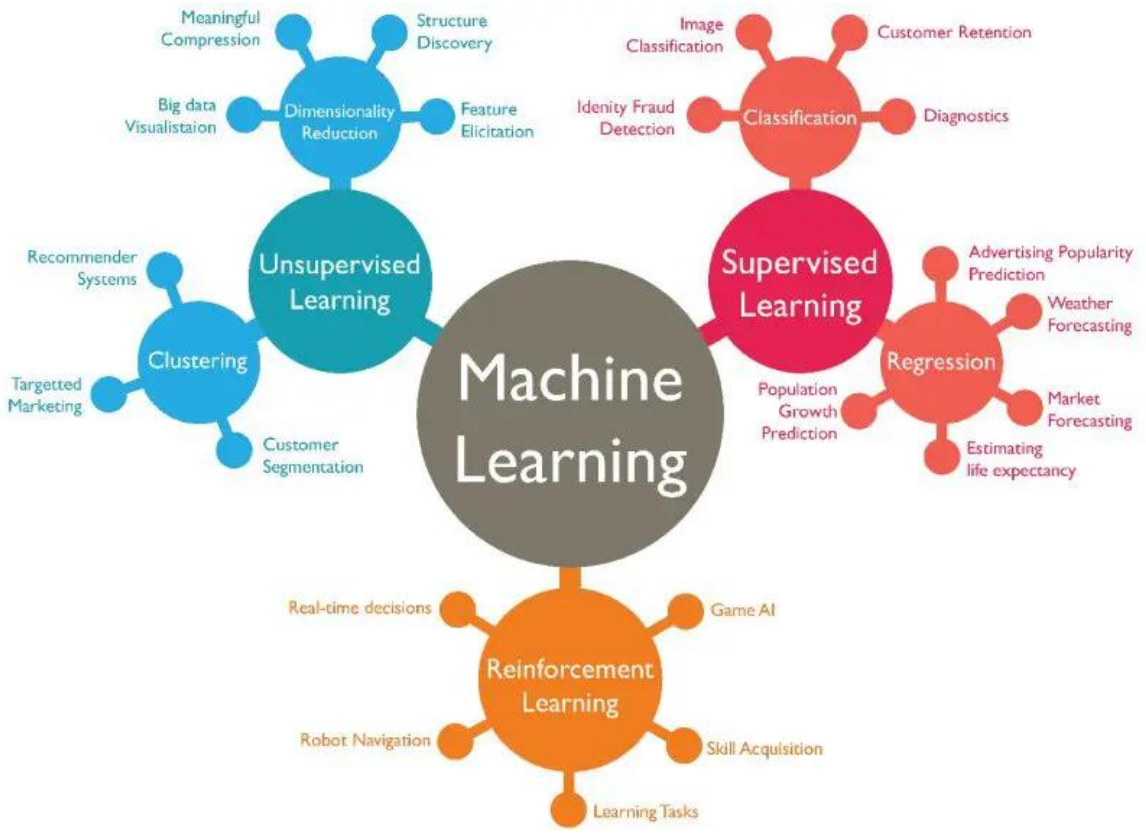


Figure 8: Machine learning algorithms overview (Omondi Asimba, 2019)

2.3.1 Supervised learning

Supervised machine learning is used for making predictions from data. To be able to do that, we need to know what to predict, which is also known as the target variable. The datasets where the target label is known are called labeled datasets to teach algorithms that can properly categorize data or predict outcomes. Therefore, for supervised learning we need to know the target value and targets are known in labeled datasets.

Let's look at an example: If we want to predict the prices of houses, supervised learning can help us predict that. For this, we will train the model using characteristics of the houses, such as the area (sq ft.), the number of bedrooms, amenities nearby, and other similar characteristics, but most importantly the variable that needs to be predicted – the price of the house.

A supervised machine learning algorithm can make predictions such as predicting the different prices of the house using the features mentioned earlier, predicting trends of future sales, and many more.

Sometimes this information may be easily accessible while other times, it may prove to be costly, unavailable, or difficult to obtain, which is one of the main drawbacks of supervised learning.

Types of problems:
Supervised learning deals with two distinct kinds of problems:

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

Classification problem: In the case of classification problems, examples are classified into one or more classes/ categories.

For example, if we are trying to predict that a student will pass or fail based on their past profile, the prediction output will be “pass/fail.” Classification problems are often resolved using algorithms such as Naïve Bayes, Support Vector Machines, Logistic Regression, and many others.

Regression problem: A problem in which the output variable is either a real or continuous value, s is defined as a regression problem. Bringing back the student example, if we are trying to predict that a student will pass or fail based on their past profuse, the prediction output will be numeric, such as “68%” likely to score.

Predicting the prices of houses in an area is an example of a regression problem and can be solved using algorithms such as linear regression, non-linear regression, Bayesian linear regression, and many others.

2.3.2 Unsupervised learning

Imagine receiving swathes of data with no obvious pattern in it. A dataset with no labels or target values cannot come up with an answer to what to predict. Does that mean the data is all waste? Nope! The dataset likely has many hidden patterns in it.

Unsupervised learning studies the underlying patterns and predicts the output. In simple terms, in unsupervised learning, the model is only provided with the data in which it looks for hidden or underlying patterns.

Unsupervised learning is most helpful for projects where individuals are unsure of what they are looking for in data. It is used to search for unknown similarities and differences in data to create corresponding groups.

An application of unsupervised learning is the categorization of users based on their social media activities.

Commonly used unsupervised machine learning algorithms include K-means clustering, neural networks, principal component analysis, hierarchical clustering, and many more.

2.3.3 Reinforcement learning

Another type of machine learning is reinforcement learning.

In reinforcement learning, algorithms learn in an environment on their own. The field has gained quite some popularity over the years and has produced a variety of learning algorithms.

Reinforcement learning is neither supervised nor unsupervised as it does not require labeled data or a training set. It relies on the ability to monitor the response to the actions of the learning agent.

Most used in gaming, robotics, and many other fields, reinforcement learning makes use of a learning agent. A start state and an end state are involved. For the learning agent to reach the final or end stage, different paths may be involved.

An agent may also try to manipulate its environment and may travel from one state to another

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

On success, the agent is rewarded but does not receive any reward or appreciation for failure
Amazon has robots picking and moving goods in warehouses because of reinforcement learning
Numerous IT companies including Google, IBM, Sony, Microsoft, and many others have established research centers focused on projects related to reinforcement learning.

Social media platforms like Facebook have also started implementing reinforcement learning models that can consider different inputs such as languages, integrate real-world variables such as fairness, privacy, and security, and more to mimic human behavior and interactions. (Source)

Amazon also employs reinforcement learning to teach robots in its warehouses and factories how to pick up and move goods.

Table 2: Differences between supervised, unsupervised, and reinforcement learning algorithms

	Supervised learning	**Unsupervised learning**	**Reinforcement learning**
Definition	Makes predictions from data	Segments and groups data	Reward-punishment system and interactive environment
Types of data	Labeled data	Unlabeled data	Acts according to a policy with a final goal to reach (No or predefined data)
Commercial value	High commercial and business value	Medium commercial and business value	Little commercial use yet
Types of problems	Regression and classification	Association and Clustering	Exploitation or Exploration
Supervision	Extra supervision	No	No supervision
Algorithms	Linear Regression, Logistic Regression, SVM, KNN and so forth	K – Means clustering, C – Means, Apriori	Q – Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self-Driving Cars, Gaming, Healthcare

2.3.4 Probability Theory in Machine Learning

Definition: Probability theory provides a foundational framework for modeling machine learning algorithms and making precise statements about their effectiveness.

Probability theory and its tools are integral to analyzing machine learning. Generally, machine learning involves inferring patterns or relationships from limited data, inherently involving probability theory and statistics. Relationships to be detected may be stochastic (non-deterministic), and even deterministic relationships are only partially revealed by limited data, requiring probabilistic approximation.

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

In (supervised) machine learning, much effort is devoted to selecting a hypothesis based on training data that accurately describes the relationship between inputs and outputs. There are two main approaches to hypothesis selection: the Bayesian approach and the frequentist approach. Both approaches involve choosing from a set of possible hypotheses.

In the Bayesian approach, the prior probability of a hypothesis reflects the belief that it describes the underlying relationship between inputs and outputs. Bayesian learning updates this prior based on observed data, resulting in a posterior distribution over the hypotheses. This is done using Bayes' rule, which relates the conditional probability of a hypothesis given the data to the conditional probability of the data given the hypothesis and the prior probability of the hypothesis. Calculating these probabilities can be computationally challenging due to high-dimensional integrals, often necessitating approximation techniques.

The frequentist approach to supervised learning, employs various probabilistic techniques without a prior distribution on set of possible hypotheses. The choice of hypothesis class may be dictated by the learning system (e.g., a neural network) or based on the assumption that the underlying relationship can be described by a specific type of hypothesis. However, no further assumptions about degrees of belief in specific hypotheses are made. Instead, frequentist models assume probabilistic selection of training data, such that the "fit" of a hypothesis to the data reliably indicates its suitability as a model for the true relationship between inputs and outputs. Formalizing this requires probability theory, particularly uniform laws of large numbers (or "uniform convergence results").

While Bayesian and frequentist models have distinct differences, recent developments have introduced "PAC-Bayesian" models, which blend elements of both approaches.

2.4 Introduction to AL

Artificial intelligence (AI) is the simulation of human intelligence in machines that are programmed to think and act like humans. Learning, reasoning, problem-solving, perception, and language comprehension are all examples of cognitive abilities.

Artificial Intelligence is a method of making a computer, a computer-controlled robot, or a software think intelligently like the human mind. AI is accomplished by studying the patterns of the human brain and by analyzing the cognitive process. The outcome of these studies develops intelligent software and systems.

2.4.1 Types of Artificial Intelligence

1. Purely Reactive

These machines do not have any memory or data to work with, specializing in just one field of work. For example, in a chess game, the machine observes the moves and makes the best possible decision to win.

2. Limited Memory

These machines collect previous data and continue adding it to their memory. They have enough memory or experience to make proper decisions, but memory is minimal. For example, this machine can suggest a restaurant based on the location data that has been gathered.

3. Theory of Mind

This kind of AI can understand thoughts and emotions, as well as interact socially. However, a machine based on this type is yet to be built.

4. Self-Aware

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

Self-aware machines are the future generation of these new technologies. They will be intelligent, sentient, and conscious.

2.4.2 Goals of AI

Artificial Intelligence emphasizes three cognitive skills of learning, reasoning, and self-correction, skills that the human brain possess to one degree or another. They are defined in the context of AI as:

Learning: The acquisition of information and the rules needed to use that information.

Reasoning: Using the information rules to reach definite or approximate conclusions.

Self-Correction: The process of continually fine-tuning AI algorithms and ensure that they offer the most accurate results they can.

These are further extended and elaborated the goals of AI to the following:

Logical Reasoning: AI programs enable computers to perform sophisticated tasks. On February 10, 1996, IBM's Deep Blue computer won a game of chess against a former world champion, Garry Kasparov.

Knowledge Representation: Smalltalk is an object-oriented, dynamically typed, reflective programming language that was created to underpin the "new world" of computing exemplified by "human-computer symbiosis."

Planning and Navigation: The process of enabling a computer to get from point A to point B. A prime example of this is Google's self-driving Toyota Prius.

Natural Language Processing: Set up computers that can understand and process language.

Perception: Use computers to interact with the world through sight, hearing, touch, and smell.

Emergent Intelligence: Intelligence that is not explicitly programmed, but emerges from the rest of the specific AI features. The vision for this goal is to have machines exhibit emotional intelligence and moral reasoning.

2.4.3 Advantages and Disadvantages of AI

Artificial intelligence has its pluses and minuses, much like any other concept or innovation.

Pros

- It reduces human error
- It is available 24x7
- It easily handles repetitive tasks

It's fast

Cons

- It's costly to implement
- It can't duplicate human creativity
- It will replace some jobs, leading to unemployment
- People can become overly reliant on it

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

2.5 Game theory

Game theory is a branch of mathematics that models strategic interactions between rational players (agents) within a framework of predefined rules and outcomes. Each player aims to maximize their reward through strategic decision-making. The game consists of players, actions, strategies, and a final payoff, which all players compete to achieve.

Game theory has become fundamental in understanding both machine learning algorithms and various real-life scenarios. Take the Support Vector Machine (SVM), for instance. In game theory terms, SVM represents a game between two players: one player selects the most challenging points for classification, and the other finds the optimal hyper-plane to classify these points. The game's outcome is a trade-off between the players' strategic skills, resulting in an effective classification model.

2.5.1 Types of Games:

Zero-Sum and Non-Zero-Sum Games:

- Zero-Sum Games: One player's gain is another's loss.
- Non-Zero-Sum Games: All players can benefit from each other's moves.

Simultaneous and Sequential Games:

- Sequential Games: Players make moves with knowledge of previous actions (e.g., board games).
- Simultaneous Games: Players make moves without knowing the others' actions.

Perfect and Imperfect Information Games:

- Perfect Information Games: Players know all previous moves and strategies (e.g., chess).
- Imperfect Information Games: Players lack full knowledge of others' moves (e.g., card games).

Asymmetric and Symmetric Games:

- Asymmetric Games: Players have different goals.
- Symmetric Games: Players share the same goal but use different strategies.

Cooperative and Non-Cooperative Games:

- Cooperative Games: Players form alliances to achieve a common goal.
- Non-Cooperative Games: Players act independently to maximize their own benefit.

2.5.2 Applications of Game Theory in Machine Learning

Game theory being a branch of applied mathematics found its applications in economics, business, politics, biology, computer science and many other fields. Specifically, in this section, we illustrate a few of the important applications of game theory in machine learning that is a subfield of computer science and statistics.

Spam Detection: A spam detector or filter is a software used to discover unwanted email, and then disallow it from getting into a user's inbox. Spam detectors can either be aggressive by blocking the occasional legitimate email in order to keep more spam out of a user's inbox, or permit few extra spam messages so as to make sure that all legitimate emails pass through. Iinteraction between spammer and spam detector is demonstrated as a sequential two-person non cooperative Stackelberg game. A spam that passes through the inbox is a cost to the spam detector, whereas a benefit to the spammer. A Stackelberg equilibrium is obtained if the spammer and spam detector play their best stratagem at the same time.

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

Natural Language Processing: Game theoretic modeling for deriving the meaning of an utterance is given in Language and Equilibrium where the author demonstrates a distinct viewpoint on the natural language semantics and pragmatics. He begins with the idea that form and meaning correspondences are established as a consequence of balancing conflicting syntactic, informational, conventional, and flow constraints, and shows how to extract the meaning of an utterance by constructing a model as a system of mutually dependent games, that is a generalization of the Fregean Principle of Composition.

Gene Expression Analysis: Gene expression is the process by which genetic information is used in protein synthesis. Gene expression analysis is observing the picture of gene expressions (activity) in the sample of cells under investigation. In the thesis, the author employs game theory to quantitatively estimate the significance of each gene in controlling or provoking the condition of interest, for example, a tumor, taking into consideration how a gene associates in all subgroups of genes.

Robot Navigation: It is an important discipline of research to construct and execute a multi-robot system to do tasks such as searching, exploring, rescuing, and map-building. One of the key problems in multi-robot systems is determining how robots coordinate or cooperate among themselves in such a way that the global performance is optimized. The problem of coordination is further complicated, if the environment is dynamic. A dynamic-programming formulation is proposed for determining the payoff function for each robot, where travel costs as well as interaction between robots are taken into account. Depending on this payoff function, the robots play a cooperative, non zero-sum game, and both pure Nash Equilibrium and Mixed-Strategy Nash Equilibrium solutions are presented to help the robots navigate in such a way so as to achieve an optimal global performance.

Image Segmentation: Image segmentation decomposes an image into segments or regions containing "identical" pixels. It is used to identify objects and trace the boundaries within images. The problem of image segmentation is formulated as an evolutionary cluster game played by two players. At the same time, each player chooses a pixel that must be clustered, and obtains a payoff according to the similarity that the chosen pixel has with respect to the pixel chosen by the opponent. The evolutionary stable algorithm finds a cluster of pixels in the given image. Next, the similarity matrix is recomputed for the unlabeled pixel set and the evolutionary stable algorithm is run to find another segment. The process is iterated until all pixels have been labeled.

Disaster Management: In emergency management after a major disaster, it is necessary that organizations and individuals who are familiar and unfamiliar with disaster environment be able to manage and use the resources effectively to respond in time. Determining when, where, how, and with whom these resources should be distributed is a difficult problem in emergency or disaster management due to cross-cultural differences and interoperability issues. A game theoretic formulation is proposed to maximize the efficiency of actors involved in emergency management.

Information Retrieval Systems: In the information retrieval system, when the user inputs a query, it searches for the objects in the collection that match the query. In the process, more than one object may match the query, perhaps with varying degrees of relevance. Therefore, information retrieval systems compute rankings of the objects depending on how well an object in the collection matches the query. The top ranking objects are then presented to the user. The information retrieval process is modeled as a game between two abstract players. The "intellectual crowd" that uses the search engines is one player and the community of information retrieval systems is another player. The authors apply game theory by treating the search log as Nash equilibrium strategies and solve the inverse problem of calculating the payoff functions using the Alpha model, where the search log contains the statistics of users' queries and search engines' replies.

Recommendation Systems: Recommendation systems are the software tools that provide recommendations to the user where the recommendations are related to decision making processes such as what

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

items to be bought, what music to be heard, what online news to read and so on. Recommendation systems seek ratings of the items from the users that purchased or viewed them, and then aggregate the ratings to generate personalized recommendations for each user. Consequently, the quality of the recommendations for each user depends on ratings provided by all users. However, users prefer to maintain privacy by not disclosing much information about their personal preferences. On the other hand, they would like to get high-quality recommendations. The tradeoff between privacy preservation and high-quality recommendations is addressed by employing game theory to model the interaction of users and derives a Nash equilibrium point. At Nash equilibrium the rating' strategy of each user is such that no user can benefit in terms of improving his privacy by unilaterally deviating from that point.

Automatic Speech Recognition: Conventional automatic speech recognition systems take into account all features of an input feature vector for training and recognition. But, most of the features are redundant and irrelevant. Consequently, training an automatic speech recognition system on such a feature vector can be ineffective due to over-fitting. Furthermore, each sound is characterized by a different set of features. To deal with these problems, feature subset that is relevant for identifying each sound is investigated by several authors. They present a cooperative game theory based framework is proposed, where the features are the players and a group of players form a coalition that maximizes the accuracy of the system.

2.6 Constraint Satisfaction Problems (CSP) in Artificial Intelligence

Finding a solution that meets a set of constraints is the goal of constraint satisfaction problems (CSPs), a type of AI issue. Finding values for a group of variables that fulfill a set of restrictions or rules is the aim of constraint satisfaction problems. For tasks including resource allocation, planning, scheduling, and decision-making, CSPs are frequently employed in AI.

2.6.1 Basic components in the constraint satisfaction problem:

Variables: The things that need to be determined are variables. Variables in a CSP are the objects that must have values assigned to them in order to satisfy a particular set of constraints. Boolean, integer, and categorical variables are just a few examples of the various types of variables, for instance, could stand in for the many puzzle cells that need to be filled with numbers in a sudoku puzzle.

Domains: The range of potential values that a variable can have is represented by domains. Depending on the issue, a domain may be finite or limitless. For instance, in Sudoku, the set of numbers from 1 to 9 can serve as the domain of a variable representing a problem cell.

Constraints: The guidelines that control how variables relate to one another are known as constraints. Constraints in a CSP define the ranges of possible values for variables. Unary constraints, binary constraints, and higher-order constraints are only a few examples of the various sorts of constraints. For instance, in a sudoku problem, the restrictions might be that each row, column, and 3×3 box can only have one instance of each number from 1 to 9.

Constraint Satisfaction Problems (CSP) representation:

The finite set of variables V1, V2, V3 ……..Vn.

Non-empty domain for every single variable D1, D2, D3 ……..Dn.

The finite set of constraints C1, C2 ……., Cm.

- where each constraint Ci restricts the possible values for variables,

e.g., V1 ≠ V2

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

- Each constraint Ci is a pair <scope, relation>

Example: <(V1, V2), V1 not equal to V2>

- Scope = set of variables that participate in constraint.
- Relation = list of valid variable value combinations.

There might be a clear list of permitted combinations. Perhaps a relation that is abstract and that allows for membership testing and listing.

2.6.2 Constraint Satisfaction Problems (CSP) algorithms:

The **backtracking algorithm** is a depth-first search algorithm that methodically investigates the search space of potential solutions up until a solution is discovered that satisfies all the restrictions. The method begins by choosing a variable and giving it a value before repeatedly attempting to give values to the other variables. The method returns to the prior variable and tries a different value if at any time a variable cannot be given a value that fulfills the requirements. Once all assignments have been tried or a solution that satisfies all constraints has been discovered, the algorithm ends.

The **forward-checking algorithm** is a variation of the backtracking algorithm that condenses the search space using a type of local consistency. For each unassigned variable, the method keeps a list of remaining values and applies local constraints to eliminate inconsistent values from these sets. The algorithm examines a variable's neighbors after it is given a value to see whether any of its remaining values become inconsistent and removes them from the sets if they do. The algorithm goes backward if, after forward checking, a variable has no more values.

Algorithms for **propagating constraints** are a class that uses local consistency and inference to condense the search space. These algorithms operate by propagating restrictions between variables and removing inconsistent values from the variable domains using the information obtained.

2.7 Applications of Big Data concepts in the water sector

Artificial intelligence (AI) and machine learning (ML) are novel techniques to detect hidden patterns in environmental data. They can make the surveillance and process control in the water sector much more accurate, efficient and economical. There use in the water sector is rapidly increasing,

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

Table 3: Major categories of AI and ML applications in water environments (Bagheri, 2023).

Category	Examples	Model	Inputs	Performance	Refs
Prediction	Predicting water quality characteristics. Predicting groundwater level. Predicting impacts on industrial water treatment. Predicting groundwater quality. Analyzing trend and forecasting of rainfall changes. Forecasting of wastewater quality indicators. Predicting water quality recovery in a water treatment system. Predicting oxidation kinetics of water contaminants. Assessing the vulnerability state of wetland habitat as a result of damming. Predicting suspended sediment load of a river. Predicting tide level in Venice.	MLPNN, ENN, SVM, KNN. MLP, RBF, GR, CF, CMIS. ANFIS. SVM-FFA, NN-FFA, RF, ANFIS, MARS, PSO-SVM, PSO-NBC. MLPNN. SVR, RT. XGBoost. SGB tree. PSO, SVM, NN, RBF, MP5, bagging. CART, MLP, ANFIS, SVM. M5P, RF, MLP.	Main water quality parameters. 12 spatial parameters of sediments/bedrock and 6 temporal parameters. Human labor, machinery, fuel, fertilizers, biocides, water, and electricity. 4–9 key water quality variables. Rainfall of thirty-four meteorological subdivisions. Wastewater indicators such as BOD, COD, TSS, TDS. Water quality and operational data. 457 different cases of water contaminants from 27 chemical structures. 16 hydrological, land composition and water quality parameters. River discharge, stage, rainfall and monthly suspended sediment load. Astronomical tide, wind speed, wind direction, barometric pressure, and previously observed tide levels.	SVM converges to a solution faster. The CMIS model predicted with an average absolute relative error of less than 0.11%. Provided a high level of capability and contributed to the minimization of uncertainties associated with the process. The SVM-FFA and PSO-NBC achieved the best predictability performance. There was close adjacency between predicted and observed rainfall. Models were useful for the sizing of the treatment units in absence of direct experimental data. The performance of XGBoost model improved by pretreatment of the data. The model could estimate with R-squared of 0.97, and a quite low mean absolute relative error of 0.86%. The PSO-RBF model was found to be the best representative. The CART performed best and was found to be useful with hydro-meteorological data. M5P outperformed the other two algorithms in most cases.	(Najah et al., 2011) (Najafabadipour et al., 2022) (de Jesus et al., 2022) (Shiri et al., 2021), (Agrawal et al., 2021) (Praveen et al., 2020) (Granata et al., 2017) (Park et al., 2022) (Keivanimehr et al., 2021) (Khatun et al., 2021) (Choubin et al., 2018) (Granata and Di Nunno, 2021)

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

Category	Examples	Model	Inputs	Performance	Refs
Optimization	Calibrating water distribution networks. Enabling water desalination sustainability optimization. Providing optimal operational rules for dam and reservoir systems. Optimizing operating parameters in membrane bioreactors treating oily wastewaters. Reducing energy and materials in wastewater treatment plants. Producing better adsorption bed and adsorbent porosity.	NN, GA. Fuzzy logic. SMLA, PSO, GA. GA. GA GA.	System, pipes, junctions, sources, and valves variables, field measurements. Factors including financial, environmental, engineering, and demands. Water demand and amount of water released. Trans-membrane pressure, cross-flow velocity, feed temperature, pH. Influent conditions including wastewater parameters, energy and materials consumption. Coefficients for the kinetics, mass transfer and heat transfer.	The NN calibration process, in large water networks, requires a great deal time of calculation than the GA. The method introduced an optimal settings of the desalination process. The SMLA approach outperformed both of the conventional algorithms. The optimum values were obtained to achieve maximum permeation flux accompanied by a minimum fouling. The total energy and materials cost could be reduced by 10%– 15%. Using GA to choose different coefficients, the final solutions could be obtained.	(Saldarriaga et al., 2004) (Alzu'bi et al., 2019) (Allawi et al., 2018) (Soleimani et al., 2013) (Wang et al., 2020) (Zhang et al., 2017)
Classification	Groundwater potentiality mapping. Identifying houses with high lead tap water concentration. Establishing a water quality evaluation model. Developing intelligent regulation models to maintain a clean and stable water environment. Predicting different classes of water quality of a reservoir. Providing ground-	RF, RSS. NN, GBM, SVM, RF, LR. SVM-PSO. GA-BP, LSTM, PNN, ELM. NN, DT, FFNN, ANFIS. BRT, RF, MDA. 13 ML classifiers (SVC, LVC, CRT, RFC, ABC, GBC, KNN, MLP, GPC, LDA, QDA, NBA, LRG). LR, RFC, KNN. NN, KNN, SVM.	14 selected groundwater conditions. 14 features related to the houses and tap water. Important water quality parameters. Water quality indicators, including water temperature, DO, pH, and salinity. 4–8 water quality parameters. 11 predictors related to groundwater properties and environmental conditions. 5 explanatory variables (latrine density, distance to the closest	The RSS (AUC=0.892) model outperformed RF model. The GBM outperformed other models with the AUROC of 0.716. The AI could help to make more accurate predictions and evaluations. The LSTM model had the highest accuracy (test set 96.84%) and F1-score (test 93.83%). The percentage of adjustments relative to the number of presented cases was	(Sarkar et al., 2021) (Hajiseyedjavadi et al., 2020) (Wei, 2021) (Chen et al., 2021) (Couto et al., 2012), (Hmoud Al-Adhaileh and Waselallah Alsaade, 2021) (Mosavi et al., 2020) (Díaz-Alcaide and Martínez-Santos, 2019) (Ravi and

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

Category	Examples	Model	Inputs	Performance	Refs
	water hardness susceptibility (high, moderate, low). Mapping fecal pollution in rural groundwater supplies. Monitoring failure of onsite septic systems. Assessing presence and absence of Salmonella in agricultural surface waters.		latrine, borehole yield, water table depth and population density). 16 independent and 3 dependent variables. Microbiological indicators and physicochemical attributes of water.	97.4% for the testing set. The RF indicated better performance than BRT and MDA models. RFC and LRG rendered prediction scores for fecal pollution in excess of 0.90. The algorithms showed accuracy of 80% in prediction of failures. The algorithms predicted with an accuracy between 58% and 59%.	Johnson, 2021) (Polat et al., 2020)
Feature extraction	Deriving reduced order models of hydrologic exchange flows in large regulated rivers. Extracting important features for accurate water quality monitoring.	PCA, RF, XGB. PCA, LDA, ICA.	Bathymetric, hydrodynamic, geomorphologic, and flow attributes. 8 water quality indicators.	The RF and XGB achieved 70–80% accuracy and were effective alternatives to the numerical models. Integration of LSTM RNNs with LDA, and ICA yielded an accuracy of 99.72%.	(Ren et al., 2021) (Dilmi and Ladjal, 2021)
Pattern/Image recognition	Predicting groundwater availability. Extracting patterns and information to understand flood behaviors. Recognizing current dominant filtration mechanisms during the operation. Proposing a systematic fish swarm detection and position method.	MLR, MLP, RF, XGB, SVR. RF, Lazy, J48 tree, NN, NB, LR. AMR-PSO. CNN.	174 monthly groundwater images. 69,558 and 3642 instances of flood events for training and testing the models. Permeation flux data. 3000 fish images taken by the lab.	The SVR consistently showed less error in recognizing groundwater images. RF provided the most accurate flood patterns, followed by J48 tree and Lazy methods. The model was in good agreement with experimental data. The model had a good real-time performance.	(Hussein et al., 2020) (Saravi et al., 2019) (Drews et al., 2007) (Xu and Matzner, 2018)
Control	Developing a water irrigation decision support system for	NN-Fuzzy logic. NN-GA. GA-NN	Water level data. Various operating parameters. Different control and	The system could accurately estimate the water volume and adjusted the sluice	(Suntaranont et al., 2020) (Al Aani et al., 2019)

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

Category	Examples	Model	Inputs	Performance	Refs
	practical weir adjustment. Automatizing processes for water treatment and desalination. Designing an optimal control for an industrial wastewater treatment plant.		state variables (pH, flowrate, effluent and influent parameters).	gate. AI tools support the water sector towards better operation and process automation. Provided a quick tool to capture the uncertainties in the wastewater treatment process.	(CHANG et al., 2001)
Anomaly detection	Presenting a fault detection approach for a wastewater treatment plant. Detecting faults in the oxidation and nitrification process. Identifying water contamination events.	DBN-OCSVM. LSTM, RNN. ILD-GA	Process operating conditions. A real-life dataset containing over 5.1 million sensor data points. Chlorine, Electric Conductivity, pH, Temperature, Total Organic Carbon, and Turbidity.	It can raise an early alert to the abnormal conditions. The LSTM achieved a fault detection rate of over 92%. Improved the identification of the source of fecal pollution.	(Harrou et al., 2018) (Mamandipoor et al., 2020) (Housh and Ostfeld, 2015)

Table 4. Some important applications of deep learning models in water environments (Bagheri, 2023).

Application	Model	Inputs	Performance	Refs
Forecasting groundwater level in semi-confined glacial sand and gravel aquifer	MLP-ADAM	Temperature, precipitation, weighted average of groundwater level	High accuracy predictions with R^2 and RMSE of 0.95 and 0.73, respectively	(Zarafshan et al., 2021)
Distinguishing between marine life and man-made debris underwater	CNN	Input images	Safely distinguished between debris and life	(Moorton et al., 2021)
Predicting water quality index	NAR-NET, LSTM	DO, pH, conductivity, BOD, nitrate, fecal and total coliform	The models had prediction accuracies higher than 94%	(Aldhyani et al., 2020)
A novel approach to uncertainty quantification in groundwater table modeling	DNN	Retrieved geographical locations of groundwater table	Automated random deactivating connective weights approach outperformed other models	(Abbaszadeh Shahri et al., 2022)
Developing a model for source discrimination of mine water	DNN	Six hydrochemical components	The model provided significant guidance for the discrimination of mine water	(Yang et al., 2021)
Classifying hydroacoustic signals from sounds made by whales	DNN	Harmonic wavelets	The deep learning approach provided the best result	(Kaplun et al., 2020)

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

Application	Model	Inputs	Performance	Refs
Detecting marine ecotoxicology due to emerging contaminants	2D CNN	Images of the plotted time series data	Predicted the type and concentration of contaminant with a median accuracy of 97% and 100%, respectively	(Rodrigues et al., 2021)
Integrating information about microbial in activated sludge models in wastewater treatment.	DNN, CNN, LSTM	DNA sequencing	High capabilities in extracting features from molecular data	(Sin and Al, 2021)
Screening efficient water desalination with graphene nanopores	DRL, CNN	Pair of graphene state and action candidates	Proved to be a powerful tool for nanomaterial design and screening	(Wang et al., 2021)
Simulating rainfall-runoff process	LSTM	14 rainfall stations data and antecedent discharge	The LSTM models outperformed the ANN models with accuracies higher than 0.90	(Hu et al., 2018)
Source localization in a shallow water environment	DNN	The eigenvectors associated with the modal signal space from sensors	The method was effective for source localization	(Huang et al., 2018)
Predicting delineation of potential groundwater zones	DNN, DLT, DB	14 factors such as altitude, slope, soil type, land use, curvature, distance, rainfall, drainage	DB model showed higher efficiency and applicability for future applications.	(Chen et al., 2021)
Detecting anomaly of water quality systems	DNN, LSTM	Time, temperature, chlorine dioxide, pH, Redox, conductivity, turbidity, flowrate, remarkable change	The algorithms were vulnerable on imbalanced data sets	(Muharemi et al., 2019)
Analyzing water pollution level for agricultural irrigation uses	CNN	Absorbance spectra of polluted water samples	It is feasible to be used for establishing intelligent spectroscopic models	(Chen et al., 2020a)
Predicting groundwater heavy metal pollution indices	DNN	9 heavy metals	The model showed the lowest error and highest accuracy of 0.99	(Singha et al., 2020)
Analyzing spatial distribution of groundwater salinity in a coastal aquifer	EGB DNN	Transmissivity, precipitation, watertable, elevation, distance, evaporation	The model showed higher performance on the testing subset.	(Sahour et al., 2020)
Replacing process-driven hydrologic models for streamflow simulation	DNN, LSTM	Mean precipitation, potential evaporation, streamflow	They outperformed process-driven models and showed better performance in the high-flow regime	(Kim et al., 2021)
Predicting surface water quality and identifying key water parameters	DCF	COD, pH, DO, NH_3	The model could be prioritized for future water quality monitoring	(Chen et al., 2020b)

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

Application	Model	Inputs	Performance	Refs
Predicting short-term water quality variable	CNN–LSTM	Temperature, pH, ORP, EC, DO, Chl-a	The hybrid CNN–LSTM models outperformed the standalone models	(Barzegar et al., 2020)
Application in numerical simulation of seawater intrusion	DBNN	Sea-level rise, Production of wells	Resulted in improved accuracy with limited training samples	(Miao and Guo, 2021)

2.8 Conclusion

Big Data has already revolutionized the water sector, even if it is far from applied in all potential areas. The use of Big Data in the artificial intelligence, machine learning and deep learning areas will make the water sector activities such as surveillance and process control much more accurate, efficient and economical than conventional approaches.

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

3 Big Data for integrated Climate Change & Water Management

3.1 Big data application in the water sector

The development of big data, machine learning and artificial intelligence provides digital tools for the water sector. These digital tools would interact and change the water sector significantly, which enables the transition of utilities towards advanced management.

Commercial success of big data has led to speculation that big-data-like reasoning could partly replace theory-based approaches in science. Big data typically has been applied to 'small problems', which are well-structured cases characterized by repeated evaluation of predictions. However, the application of big data in almost all the engineering sectors are limited with solving 'small problems'. Big data tools applied in the water sector quite often being expected to solve "challenging problems", compared with civil application, which quite often needs the input from the domain experts. Issues like water affordability, scarcity, resilience in the face of severe weather patterns from climate change, and water quality monitoring as well as process control are the areas where big data would play a more central role.

3.1.1 Big data in climate change models relevant for water sector

Big data applied in climate change could provide overview of the interrelationship between data science and climate studies, as well as describes how sustainability climate issues can be managed using the Big Data tools. A typical example of big data in climate change models relevant for water sector is reported by the IWA Digital Water report (Sarni et al, 2019): Hydrologic models paired with monitoring devices have allowed the Durban water utility to optimize storage levels in dams and reservoirs. A more systematic application of big data tools with climate data and water quality is shown in the Supplementary reading literature (Wang et al, 2019).

The majority of existing sewer systems in European countries are combined sewer systems. In spite of the continual improvement of wastewater treatment technology, combined sewer overflow brings increasingly environmental problems for both storm water management and wastewater treatment. In the winter of some European countries, snowmelt increases the inlet flow of wastewater treatment plants (WWTPs) dramatically and the wastewater temperature may be decreased to 4 ℃ (Plósz et al, 2009). The European Council Directive 91/271/EEC has defined a threshold of 12 ℃ that 70 % nitrogen should be removed by WWTPs. However, in cold climate area, when the combined effect of increasing influent flow and lower temperature caused by snow melting exceeds the treatment capacity, a part of the incoming wastewater bypasses the treatment process and the wastewater was discharged to natural environment without sufficient treatment (Haimi et al, 2009). As results of cold and dilution effect caused by snow melting, the nutrient removal efficiency was observed to be reduced obviously (Bixio et al., 2001). The wastewater quality and quantity in combined sewer systems are combined effects of human activities and climate conditions. How the water quantity, water quality and climate data being integrated to build tools for pre-cause of wastewater treatment was finally become a data mining and data driven modelling issue. The study shows that a stepwise approach was developed to determine whether the WWTP influent wastewater contains snowmelt (wet climate) or not (dry climate). The daily, weekly and seasonal variation of WWTP influent, and climate effect on influent characteristics could be analyzed by the correlation of climatic information and wastewater characteristic (Wang et al, 2019). A classification model was developed to further discriminate climate conditions of influent, which will be applied to develop scenario-based soft sensor as well as support WWTP surveillance and control.

The stepwise application of big data techniques to mine the climate data and water characteristics is shown in the Supplementary reading literature. The data set used for the study was also attached. Students should follow this published scientific work and repeat the method in R package as the exercise.

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

3.1.2 Process surveillance and control with virtual sensors

The major challenges for water and wastewater treatment process operation are the uncertainties caused by climate change (Hwang et al, 2007, Wilen et al, 2006), variation of influent quantity and quality (Bixio et al., 2001, Martin et al., 2014), and online measurement of process variables (Rieger et al., 2005). Process optimisation is supposed to bring the process to the optimal status by stabilising the process output and minimising operational cost. Optimisation of wastewater treatment processes is achievable by applying real-time surveillance and control.

In the past decades, advanced control methods for wastewater treatment processes have been studied from several aspects. The control setpoints for carbon, nitrogen and phosphorus removal processes have been studied (Guerero et al, 2012). Several advanced control strategies were developed for different purposes, e.g. multivariate based coagulant dosing control (Manamperuma et al, 2017), ammonium based aeration control (Åmand et al, 2012), and carbon source dosing control for nitrogen removal (Zeng et al., 2008; Stare et al., 2007). However, there is always a discussion about "why many control systems fail" (Rieger et al, 2012). Whether the surveillance methods or control law limited the implementation of real-time process optimisation was also an interesting question. Steyer et al. (2006) compared a fuzzy logic controller and a model predictive controller for the anaerobic digestion process, and it turned out that the online monitoring of alkalinity was the restriction rather than the control law. Moreover, an alkalinity prediction model was recommended to replace the expensive alkalinity sensor, which is the early stage of virtual sensor application for process surveillance.

Online monitoring of state variables is the basic requirement for process control. Commercially available instruments for online monitoring of wastewater treatment processes make real-time surveillance and control possible. Some process variables such as pH, oxidation-reduction potential (ORP), flow rate, conductivity, turbidity, dissolve oxygen (DO), etc., are viewed as easy-to-measure variables due to their short time requirement for measuring and affordable price. These easy-to-measure variables have been applied to control wastewater treatment process for different purposes, e.g. aeration control using ORP and DO sensors for energy saving (Chen et al., 2002); coagulant dosing control using flow rate, turbidity, pH and conductivity sensors (Liu eta al, 2016).

A significant development is that nutrient sensors became commercially available in the past two decades (Olsson et al., 2014). Ammonium, nitrate and phosphate sensors have been used for process surveillance in wastewater treatment plants (Machado et al., 2009). Online measurement of nitrate and ammonium nitrogen enables control of the biological nitrogen removal process more accurately (Claros et al., 2012). The most common nutrient sensors are the in situ analysers that were developed based on automated laboratory methods (Haimi et al., 2009). These nutrient analysers require the sample flow free of suspended solids, which means that these analysers need to be used coupled with online filters. Time-delay of the filtration and chemical analysis process extends the drawbacks of these analysers. Fortunately, optical sensors for ammonium and nitrate nitrogen measurement have become commercially available in recent years, but reliable optical sensors for phosphorus measurement can hardly be found on the market. The optical nitrogen sensors are more realistic to be applied for real-time control in full-scale WWTPs due to their chemical-free and short time-delay properties. In spite of this, the WWTP managers and stakeholders would hesitate on whether it is worth the investment, because these sensors are usually more expensive than other commonly used sensors.

Therefore, a major bottleneck of improving automation in WWTPs is the difficulty obtaining real-time data of the necessary state variables (Haimi et al., 2009; Chai, 2008). If fast-response and affordable hardware sensors are not available for the online measurement of carbon and nutrient, indirect data acquisition methods (e.g. prediction models) may be used as alternatives (Hedegärd et al., 2011; Corona et al., 2013).

Soft sensors (or software sensors) are groups of models using the combination of easy-to-measure variables to predict hard-to-measure models (Haimi et al., 2015). If prediction models are capable of replacing expensive online sensors, soft sensors can be applied as alternatives to hardware sensors for

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

carbon and phosphorus measurement. Since the pH differences over aeration tanks contain information on ammonium concentration, a soft sensor for ammonium detection was developed based on pH measurements (Ruano et al., 2009). Although several pH sensors can be used to replace ammonium measurement in an aeration tank, ammonium estimation is getting less important with the recent development of affordable optical sensors in the market. Online measurement of chemical oxygen demand (COD) and total phosphorus (TP) are more important because there are long time-delays to measure these two variables. Besides, the prices for COD and TP online instrument are still high based on market investigation. In recent years, soft sensor methods are mostly used for fault detection (Liu et al., 2014; Villez et al., 2011; Samuelsson et al., 2017) or dealing with time-delays (Xiong et al., 2017), but the hardware sensors cannot be completely replaced in these applications.

3.1.3 Forecasting influents and effluents of water/wastewater treatment processes

Another significant benefit of big data is its ability to improve a business's forecasting techniques. By increasing the amount of data available and drawing correlations between data sets, it's possible to create forecasting models that yield more accurate and insightful predictions.

One application of big data that is becoming more common in the water industry is predictive maintenance strategies. IoT sensors attached to essential equipment track information like operating temperature, timing and vibration. Big data analytics is then used to find correlations in that data and predict when a machine is likely to fail and needs maintenance, or is operating in a range that may cause excessive wear and tear.

The influent prediction is similar to the soft sensor mentioned in the previous section, which is actually the data mining of co-linear treads with explicit models. However, forecasting effluents data of treatment process is usually not working satisfactory. Because the effluent water concentrations are stable values with minor variation during the requirement of treatment efficiency. Therefore, domain experts knowledge, mechanistic models or semi-empirical models usually need to be adopted together with data driven models.

3.1.4 Process control algorithms with feed-forward/feed-back controls

The control theory for urban water or wastewater facilities has been available since the 1970s, when reliable dissolved oxygen sensors were introduced (Olsson et al., 1999). The instrumentation, control and automation had tremendous progress in the understanding of process dynamics and control theory. In modern wastewater treatment plants, classical feedback and feedforward control are still popular in aeration control (Åmand et al., 2013; Rieger et al., 2014) and chemical dosing control (Liu et al., 2016; Kim et al., 2004).

There are two types of control algorithms dominating the industrial process control and wastewater treatment, the on-off control and the Proportional-Integral-Derivative (PID) algorithm (Haimi et al., 2009). The on-off controllers can be considered as the simplest error-driven controller, since the control variables have only two values, u_{max} and u_{min}. The selection of u_{max} and u_{min} depends on the sign of the error (e), as shown is the following:

$$u = \begin{cases} u_{max} & if\ e > 0; \\ u_{min} & if\ e < 0 \end{cases}$$

The on-off control method has been widely used for water level control in wastewater treatment plants (Tchobanoglous et al., 2003). For the water level control, u_{max} is equal to the maximum flow rate of the pump, and u_{min} is 0. This algorithm is simple to be implemented in practice, but the drawback of on-off algorithm is as obvious as its advantage. The control variable is always oscillating around the setpoint with a time-delay, which may cause wear problems for some actuators.

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

PID is the widely used control algorithm in process control (Åström et al., 1995). The PID controller consists of three parts, the Proportion term (P), Integral term (I) and Derivative term (D),

$$u = u_0 + K_p \cdot e + \frac{K_p}{T_i}\int_0^t e \cdot dt \; + K_p \cdot T_d \cdot \frac{de}{dt}$$

where K_p is the controller gain, T_i is the integral time, T_d is the derivative time.

These three parts can either be fully utilised or with only the P term, PI term or PD term. The P term enables the controller to respond proportionally to the error (e) between the measurement and the setpoint. The Integral term sums the error of the control signal over time; therefore, the controller will be active as long as the error exists. The Derivative term is activated by the rate of error changing, which contributes to the speed of control action. The Derivative term is not commonly used in industrial process control, because it may be triggered by measurement noise. The PI feedback controllers are the most popular in wastewater treatment (Haimi et al., 2009).

In the 1990s, a lot of efforts have been made in wastewater treatment process control, but many of them were shown only in simulation. Later, control scheme for the cycles of sequencing batch reactors was developed based on ORP and pH measurements (Kim et al., 2004). As low-cost sensors, pH and ORP-based control have been used to control either aeration or carbon source dosing in nitrogen removal processes (Ruano et al., 2009; Won et al. 2011; Martin de la Vega et al., 2012; Ruano et al., 2012). The pH and conductivity sensors are proved vital for coagulant dosing control in either a feedforward or feedback control scheme (Ratnawera et al., 2015)e. Another popular application of PI feedback control is the aeration control based on DO setpoints (Machado et al., 2009). Since the wastewater quality variables were not measured in these control methods, a good understanding of process variables (e.g. ammonium) and control variables (e.g. DO) is important. Thanks to the progress in online measurement of nutrient, more accurate control can be applied in recent years. With the possibility of online measurement of ammonium nitrogen, ammonia-based control can be substituted into feedforward and feedback schemes for aeration control (Rieger et al., 2014).

Maciejowski (2002) mentioned in the book "Predictive Control with Constraints" that "Model Predictive Control (MPC) is the only advanced control technique – that is, more advanced than PID control". The concept of MPC is using one or more predicted future state variables to adjust the actuators to minimise the errors between control variables and the corresponding setpoints. Therefore, MPC can optimise multivariable processes such as wastewater treatment processes. The first documented implementation of MPC in a real biological wastewater treatment process (a pilot MBBR system) is in 2011 (Vrečko et al., 2011). Other recent studies of MPC are either theoretical studies or computer simulations. The difficulties of applying MPC in practice may be related to the lack of online measurements of some state variables. Another reason may lie in the difficulties in obtaining reliable process models.

3.2 Appendix

Supplementary material 1 – Literature reading of a stepwise big data techniques application –

Wang, X.; Kvaal, K.; Ratnaweera, H. Characterization of Influent Wastewater with Periodic Variation and Snow Melting Effect in Cold Climate Area. *Comput. Chem. Eng.* **2017**, *106*, 202–211. https://doi.org/10.1016/j.compchemeng.2017.06.009.

Supplementary material 2 – The data set to repeat the methods in Supplementary material 1 as exercise.

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

4 Planning with Big Data

4.1 Introduction

Water management infrastructure facilities, such as water extraction and distribution, wastewater collection and treatment and stormwater drainage and treatment systems, are long-lasting facilities.

The amortization periods for these facilities are between 50 - 70 years for sewerage systems and 20 - 30 years for wastewater treatment plants.

The useful lives of these components are often longer, as they are often still in good technical condition even after amortization has ended, meaning that replacement is often not necessary at this point.

For this reason, special care and intensive data analysis is required when dimensioning these systems during the basic assessment, as these systems must remain functional for a long period of time.

Similarly, the sewerage system and wastewater treatment plant, including stormwater treatment, must be seen as a complete system. This means that the same design data must be used for both.

This lecture therefore shows how data can be used for the design of sewerage systems and wastewater treatment plants.

4.2 Required data

A large amount of information and data is required for dimensioning. This is information that concerns both the quantity and the ingredients.

In the case of quantities, these are in particular flow rates or inflows, for example:

- Dry weather inflow as
  average
  minimum
  maximum
- For combined sewer systems:
  maximum inflow
- Distribution of the inflow per
  day
  month
  year
- periodic influences, e.g. seasonal characteristics such as tourism, industrial production, etc.
- Impact of infiltration and exfiltration into the systems

Information about the various ingredients is also required, such as

- Daily loads from domestic wastewater per capita for parameters like
  COD,
  Nitrogen,
  Phosphorus,
  total dry solids
  …..
- Daily loads from industrial inflow
- Distribution of the loads per
  day
  month
  year

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

- Concentrations of the parameters mentioned above as
  average
  minimum
  maximum values
- Temperature distributions over the year

Other parameters, such as pH, conductivity, etc., are often not directly usable for dimensioning, but can be helpful in assessing the available information and for plausibility checks.

In many cases, additional information will be required for reliable dimensioning, which must be collected or queried depending on the boundary conditions.

4.3 Conventional data base for design

In the past, specific sizes were often used as the basis for dimensioning, which were also reflected in the technical regulations.

These can be person-specific sizes for volumes or loads.

Examples of this can be found in Table 4.

Table 4: Design data for wastewater per capita

Country	Volume [l/(cap*d)]	COD [g/(cap*d)]	Nitrogen [g/(cap*d)]	Phosphorus [g/(cap*d)]	Reference
China			11	1.6	Jönssen et al., 2003
Germany	120 - 180	120	11	1.8	DVWK-ATV, 2003
Norway	150	120	18	1.8	Nowegian Water BA, 2020
India		59 - 103	6 - 12	0,6 – 4,5	MUD, 2013
South Africa			9.3	1.4	Jönssen et al., 2003
Uganda			6.8	1.1	Jönssen et al., 2003

This capita-specific dimensioning data can provide an initial indication, but is often very imprecise.

On the one hand, they are regionally specific, as they depend on various factors, such as

- Diet (vegetarian, vegan, etc.)
- Season and availability of food
- affluence
- Forms of production.

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

They are also subject to change over long periods of time and are therefore associated with major uncertainties.

For this reason, measured values should be used when dimensioning waste water systems, as it is often possible to refer to operating records.

4.4 Design by using operational data

4.4.1 General

In view of the fact that sensors are increasingly being used in water management systems to record various measured variables and these are then stored in databases, a great deal of information is available.

This data should not be accepted uncritically, but should be analyzed from various perspectives. The following aspects are mentioned here as examples:

- Applicability
  In what form and format is the data available? The transfer of data from the database for further processing must be checked at this point.
- Plausibility
  The plausibility of the data is one of the most important criteria. Data gaps or incorrect data transmissions must be identified and eliminated. If data gaps occur, it must be checked whether substitute value concepts can be used
- Abundance
  The frequency of data collection and storage of the collected data should be compared with the informative value for later use. For example, data (e.g. flow rate data) is recorded every minute in some systems. For use as design data, hourly or daily flow rates are often sufficient. For concentration and load data, the respective sample type and the time frame must be defined, i.e. is it a random sample, a mixed sample over two hours or over a day. Further processing or aggregation of the data is often required before it can be used.
- Outlier
  The data cohort must then be checked for plausibility. This can be done using statistical tests such as the Grubbs or Dixon outlier test.

4.4.2 Required data for design

Depending on the research question, different data with different significance is required. It is therefore not possible to derive a generally valid list of information; rather, this depends on the question being asked.

This applies to the parameters as well as the observation horizon and the temporal resolution of the data.

Information required for the dimensioning of sewer networks can be as an example:

- Average sewage flow (mean value)
- Design sewage flow (peak sewage and infiltration)
- Minimum and maximum flow velocities
- Sewer size and grades (slope, pumping)
- Number of manholes and overflows
- Predicted Rainfall scenarios (from forecasts or weather data)
- Measured rainfall events from long timelines

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

Information required for the dimensioning of wastewater treatment plants can be as an example:

Hydraulic data

- Maximum inflow per day
- Maximum dry weather inflow per day
  Minimum dry weather inflow per day
- Maximum dry weather inflow per hour
  Minimum dry weather inflow per hour
- Maximum/minimum inflow during rain events – especially for stormwater systems or combined sewer systems
- Maximum/minimum inflow per day for a specific period (Summer/Winter/touristic season)

Load data

- Temperature of water (e.g. as sliding medium per two weeks)
- Loads per day for the parameter
  COD
  Nitrogen
  Phosphorus
  total dry solids

A sufficient data basis must be available for all parameters for further processing. Therefore, at least 50 data records should form the basis for further processing. For the comparison of different period data of three years are necessary at least.

4.4.3 Analysis of data

A statistical analysis of the data usually provides a good overview of the amount of data available and its quality.

These statistical parameters also allow a comparison of the different data series with regard to their quality. The following parameters can be used as statistical parameters for an initial evaluation:

- Number of data
- Minimum
- Maximum
- Mean
- Median
- Standard deviation
- Variance
- Variation coefficient

Most evaluation tools already include the calculation of parameters. The coefficient of variation is the dimensionless quotient of standard deviation and mean value and provides information about the width of the distribution of the values compared to the mean value

For further statistical parameters and analyses, please refer to the literature (Spiegel, 1972).

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

4.5 Interpretation of inflow data

Analyzing data is the first step before further processing. For the hydraulic design of wastewater treatment plants, hydrographs are generated in the first step and compared for the different time periods. Figure 9 shows an annual distribution of daily inflow water volumes for three years for a wastewater treatment plant connected to a combined sewer system.

A large variation in the data is evident. There is a lower value range, which is associated with a high data density and represents the dry weather discharge in the order of 3,000 - 8,000 m^3/d. In addition, there are precipitation-related inflow water volumes of up to 22,000 m^3/d. Furthermore, a clear annual variation is recognizable. The inflow water volumes are significantly higher in autumn and winter. This is due to the infiltration of groundwater into the sewerage system.

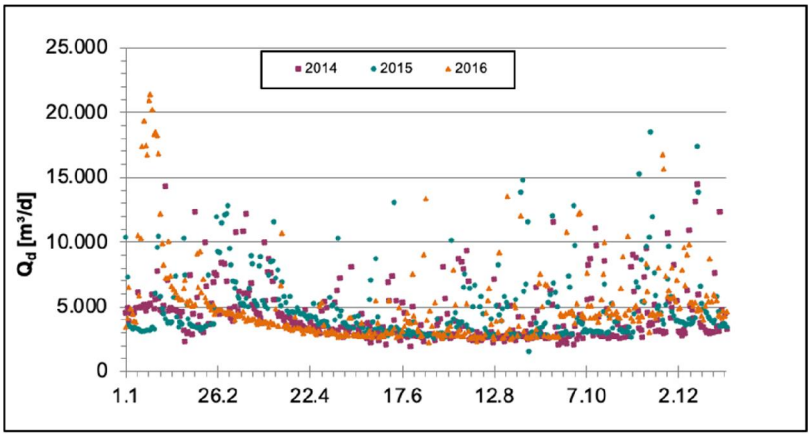


Figure 9: Hydrograph of inflow data of a wastewater treatment plant

The design of wastewater treatment plants is based on the dry weather discharge and the dry weather loads. This means that all inflows that are influenced by precipitation must be eliminated. There are various ways of doing this. Based on the records of precipitation levels, all days with a precipitation of $h_N > 1$ mm are excluded. As precipitation also causes a lag time, i.e. it takes a certain amount of time for the collected precipitation to reach the wastewater treatment plant, all lag days are also eliminated. Lag days are the days that follow a day with precipitation.

The data processed in this way, which describes the dry weather inflow, is shown in the Figure 10.

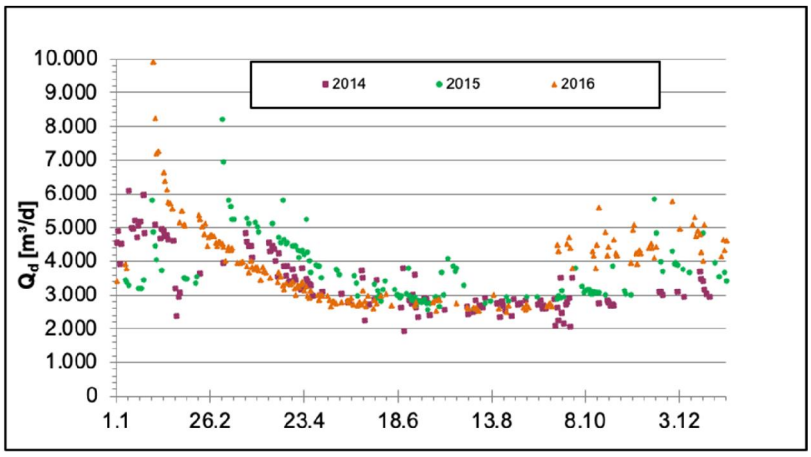


Figure 10: Hydrograph of dry weather inflow per day

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

The level of dry weather inflows is in the range of 2,000 - 10,000 m^3/d. The dry weather inflows are relatively close together and only a few high inflow volumes occur. The annual pattern is very similar for all three years. Here too, significantly higher inflows are measured in the autumn and winter months due to the high infiltration. However, the comparison of the three years shows that the inflow data is representative and can therefore be used for dimensioning.

Statistical distributions are created on the basis of this data. For this purpose, the values are divided into classes and the frequency of the values in the respective classes is calculated. The percentage frequency of the values per class is cumulated and plotted as a cumulative frequency curve (Figure 11).

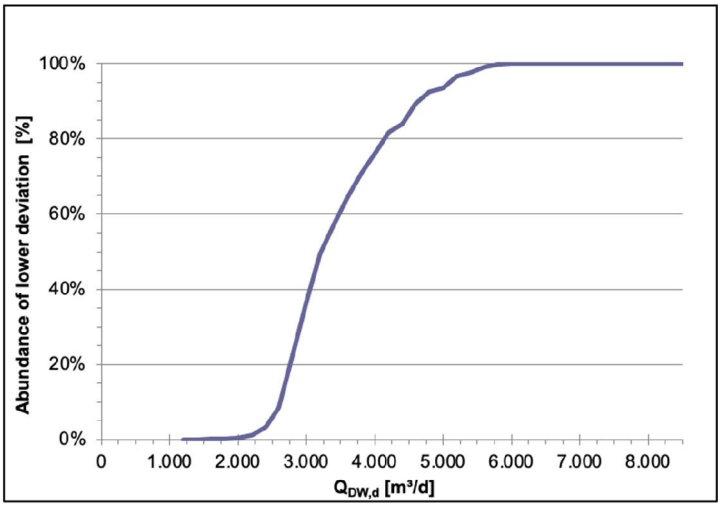


Figure 11: Abundance of lower deviation

The relevant assessment inflows are determined from these cumulative frequency curves. For sewage treatment plants that are connected to a mixed water system, the rated inflow is Q_{85}, i.e. the inflow that is undercut in 85% of cases (85 %-quantile). The Q_{99}-value applies to sewage treatment plants in a separation system. The relevant design inflow for the data available here can be seen in Figure 12. The statistical parameters for the inflow data are also listed in Table 5.

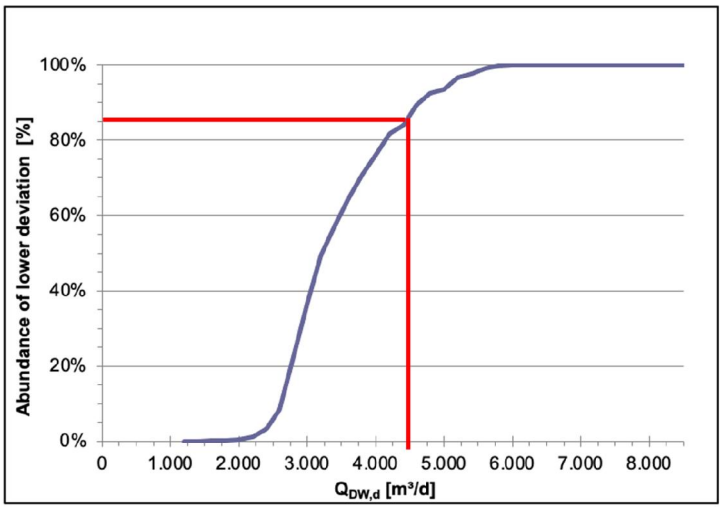


Figure 12: Identification of hydraulic design parameters

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

Table 5: Design data for wastewater inflow rates for the period 2014 – 2016

Parameter	Unit	
Number	[-]	475
Inflow minimum	[m^3/d]	1,526
Inflow maximum	[m^3/d]	5,804
Inflow mean	[m^3/d]	3,453
Inflow Median	[m^3/d]	3,214
Standard deviation	[-]	896
Variation coefficient	[-]	0.23
Design inflow Q_{85}	[m^3/d]	4,428

In addition to the daily inflows, the determination of the maximum hourly inflows is important for the maximum feeding of the plant sections. This can be done by analyzing the maximum hourly inflows and by comparing the hydrographs. An example of this is shown in Figure 13.

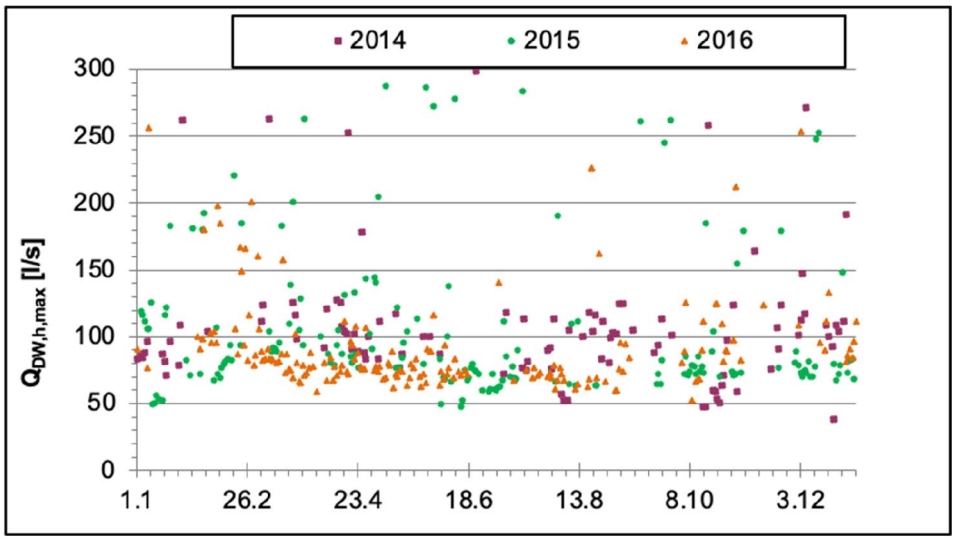


Figure 13: Hydrograph of maximum hourly inflow

In many cases, it is helpful to form ratios between different variables. This allows the data to be interpreted differently. As an example, the ratio between the maximum hourly inflow and the average hourly inflow is selected here and plotted as an annual hydrograph (Figure 14). The figure clearly shows that the fluctuations occur over a wide range of values. The most frequent values are found in the

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

range between 1.5 and 4 with some values lying above this range. Higher values are occasionally encountered and can be found throughout the year. There are therefore some higher hydraulic loads that could be investigated further.

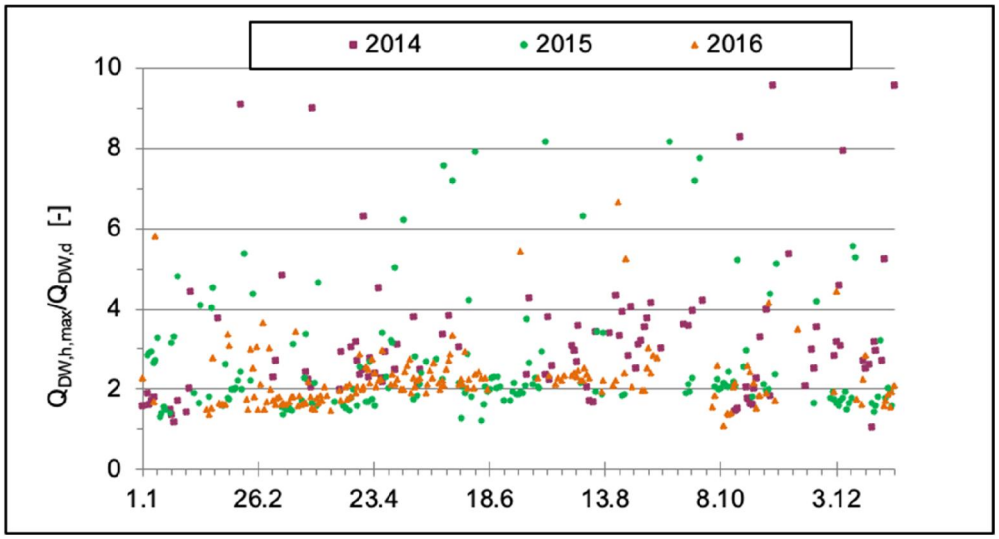


Figure 14: Hydrograph of ratio of maximum inflow to mean values

In addition to this analysis of the data, other statistical methods are conceivable:

- Subdivision of the data into periods (summer, winter, holiday periods, etc.)
- Identification of minimum and maximum areas
- Relationship between dry weather and rainwater inflow
- Comparison of daily hydrographs to determine hourly load peaks and night-time inflows

4.6 Temperature data

As the processes of biological wastewater treatment in particular are temperature-dependent, determining the design temperature is an important step in the data evaluation of design bases. In particular, the minimum and maximum ranges of the wastewater temperature must be considered here. The minimum temperature usually determines the decisive biological activity of the microorganisms. The maximum temperature is important for determining the gas solubility during aeration and for the design of summer operation.

As there are generally no sudden changes in temperature and it only changes slowly due to the climatic boundary conditions, small deviations are equalized by forming a moving average (e.g. over two weeks) and provide a realistic picture of the temperature curve.

Figure 15 shows the annual variation in temperature as a moving two-week average over a period of three years. It is clear that the annual trends are almost identical, with only 2015 having slightly lower temperatures than the other two years. It can therefore be assumed that this temperature curve is representative and therefore decisive for the measurement. A maximum of approx. 21 °C and a minimum of 7 °C can therefore be assumed.

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

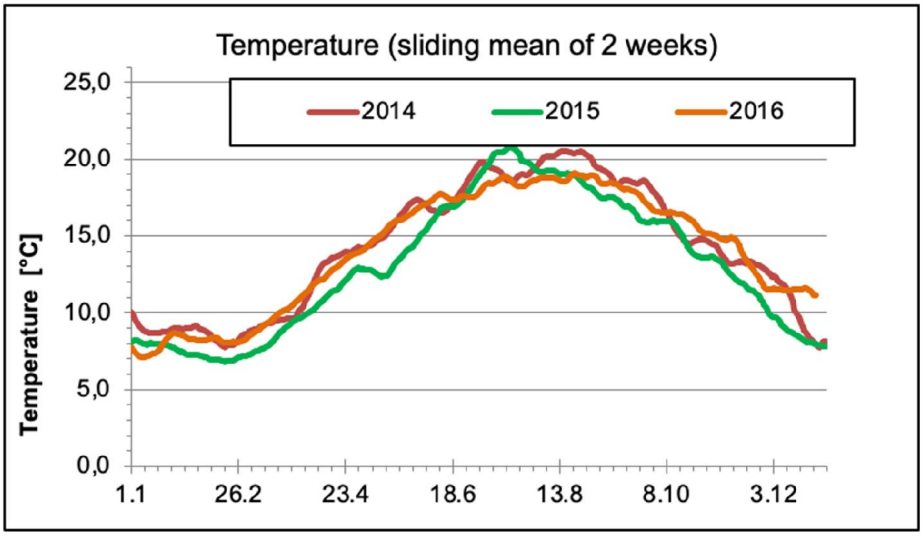


Figure 15: Hydrograph of temperature as sliding mean (two weeks)

4.7 Loads and concentrations

In analogy to the hydraulic data, design loads must also be determined for the design of wastewater treatment plants. Here, too, it makes sense to create both hydrographs and statistical distribution curves for the various parameters (BOD_5, N_{total}, P, TSS).

As an example of an annual variation in loads, the annual variation for a period of three years is shown in Figure 16for the parameter COD. Here, too, a wide fluctuation range of 500 - 2,000 kg COD/d is recognizable. The values are similar for all three years and no special features can be identified in the distribution.

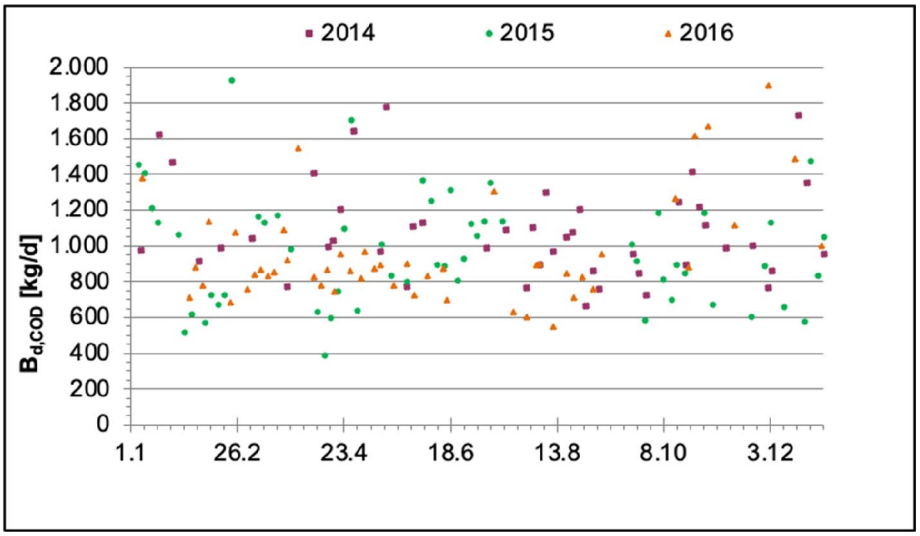


Figure 16: Hydrograph of COD-loads

The following design parameters for the COD load can be derived from the statistical distribution of the load data (Figure 17, Table 6). Concentrations are then simply derived from the design loads with the relevant inflows in each case.

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

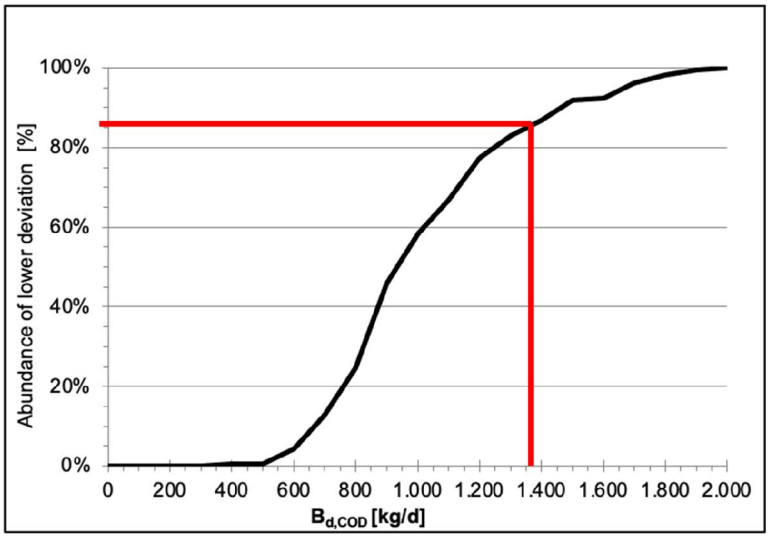


Figure 17: Identification of design parameters for COD-loads

Table 6: Design data for wastewater COD for the period 2014 – 2016

Parameter	Unit	
Number	[-]	159
Minimum load	[kg/d]	386
Maximum load	[kg/d]	1,927
Load mean	[kg/d]	1,015
Load Median	[kg/d]	955
Standard deviation	[-]	308
Variation coefficient	[-]	0.30
Design load Q_{85}	[kg/d]	1,351

In addition to the loads, ratios are also helpful here for characterizing the wastewater composition. The following can be cited as examples:

- COD : BOD_5
- COD : TSS
- COD : N
- X_{COD} : S_{COD}
- COD : P
- P : PO_4-P
- ….

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

4.8 Missing Data

Errors and missing values are a fundamental problem when using operating data and measurement data for dimensioning. The causes of missing data are manifold and must be evaluated in detail. The following are examples of the causes of missing values:

Statistical methods can replace missing values with regression methods. In the simple case, linear regressions are successful. However, multiple regression methods can also be successful. For mass data, it makes sense to use a data cluster for the regression and to carry out a verification of the regression with a second data cluster. An example of such a replacement value formation is shown in the Figure 18.

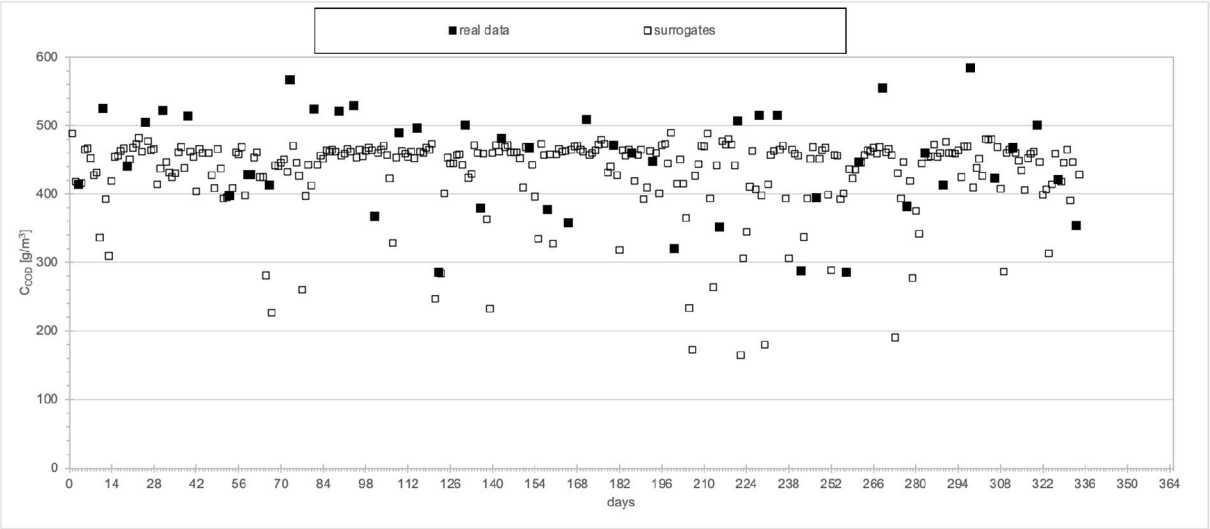


Figure 18: Filling of gaps in COD data using statistical methods

4.9 Use of models

4.9.1 Sewer models

The use of models for the dimensioning of wastewater systems often requires a large amount of data. Today, it is common practice to use simulation models to dimension or check sewer networks. These can be used on both a hydrological and hydrodynamic basis. Detailed information on the catchment area and the sewer network is required.

The following information must be available in detail

- Drained areas and buildings connected to the sewer network
- Types of reinforcement and sealing of the ground surface
- Information on the sewer supports (length, gradient, heights, pipe material and diameter)
- Information on special structures, such as overflows, branches, retention facilities

An example of such a sewer network model is shown in the Figure 19.

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

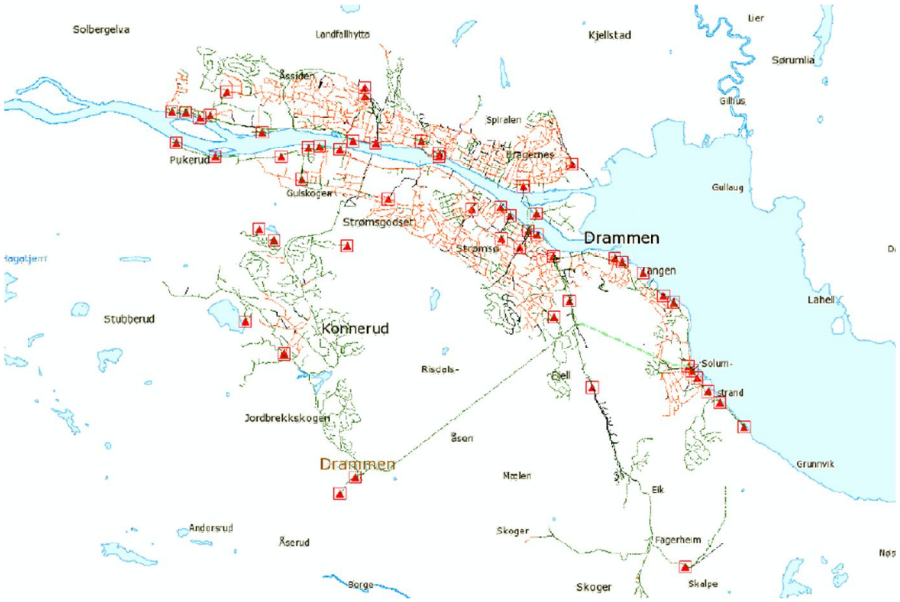


Figure 19: Overview on a sewer model of the Oslo area ***Error! Reference source not found.****)*

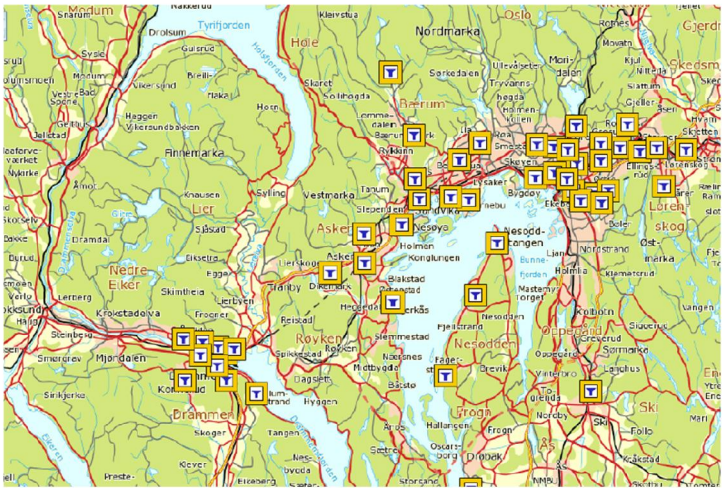


Figure 20: Example of raingauges placed in Oslo area ***Error! Reference source not found.****)*

Once such a model has been set up, it can be provided with various input data. The prerequisite for this is rainfall data, which is recorded in high temporal resolution using rain gauges. An example of the distribution of rain gauges for use in the sewer network model is shown in Figure 20.

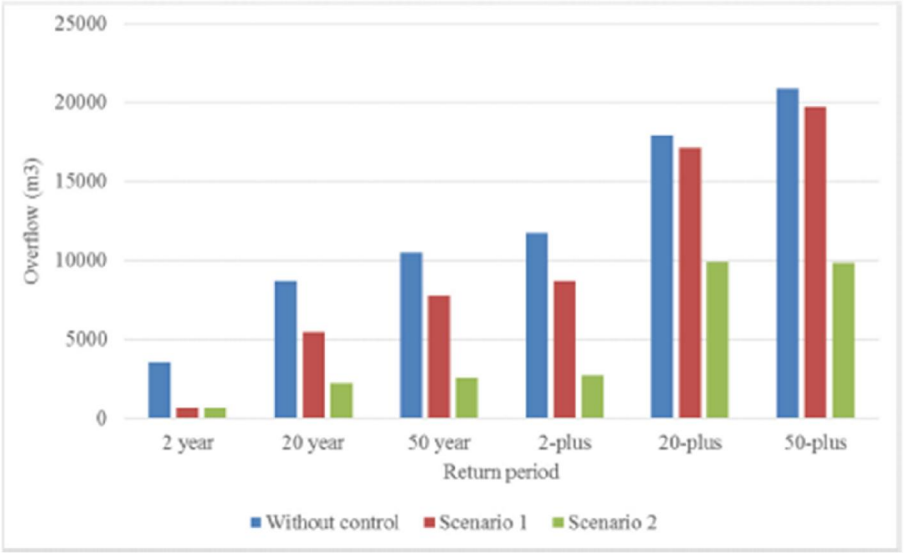
The models must be adapted by varying the model-internal parameters on the basis of measurements. This calibration process can be carried out under different quality aspects, depending on the problem at hand. Once the model has been calibrated and verified, calculations can be made using existing rainfall data or even model rainfall or rainfall series (Figure 21).

By calibrating the rainfall-runoff processes that are set up in the model, reliable information can be obtained about runoff, water levels, water volumes in the system and also overflows.

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

Figure 21: Estimation of overflows using a sewer model basing on measurement

Figure 22 shows the amount of overflow at an wastewater treatment plant under different control scenarios and rainfall events. For a 2 year return period scenario, control the could reduce the overflow by up to 82%. However, with stronger rainfalls, only a small reduction was observed for scenario 1. In scenario 2 there is dramatical reduction of overflow for all the return periods except the 2-year scenario (Zhang et al.,2018).



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

Figure 22: Reducing overflow with real-time control of sewers ***Error! Reference source not found.****)*

4.9.2 CFD simulations

Other possibilities of simulations are simulation using CFD tools (Computational Fluid Dynamics). With this form of simulation flow patterns and behaviors can be simulated. An example for the simulation of an overflow is given in Figure 23.

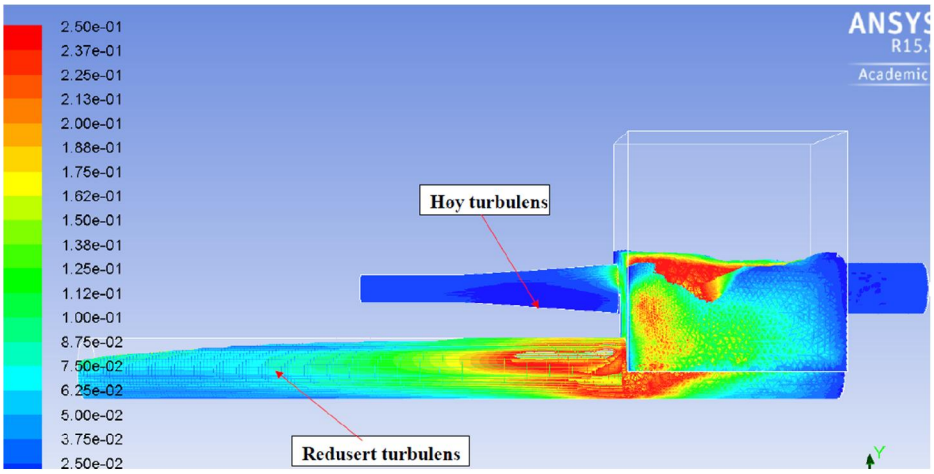


Figure 23: Overflow using CFD-simulations

CFD models can efficiently simulate the operation of sedimentation tanks and propose improvements for more even distribution and retention of water. An example for such an CFD simulation is given in the Figure 24.

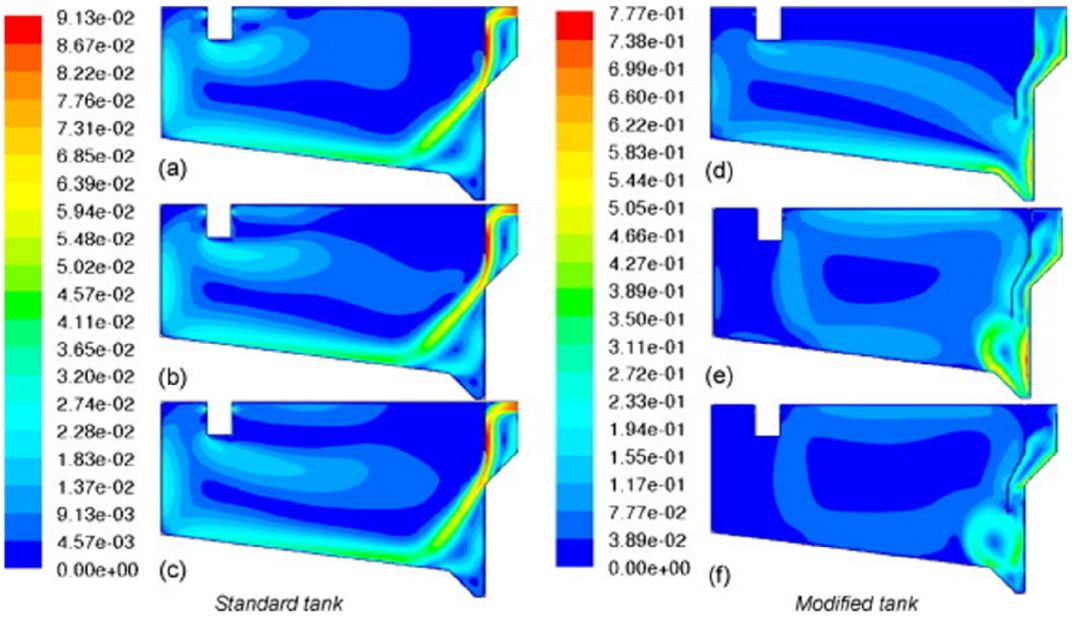


Figure 24: Optimization of a sedimentation tank

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

4.9.3 Simulation models for WWTP

While there are several simulation models are available, SUMO, WEST, SIMBA, STOAT and GPS-X can be mentioned as the most common ones.

The Sumo22from Dynamita includes several powerful simulation tools: IUWS (urban catchment and river models), Carbon Foot Print, primary effluent input, RO, new pond model, flexible SBR, SVI input, bio-P and dynamic alpha prediction, scenario handling and Digital Twin builder are examples. It has an extensive calibrated model library for traditional and advanced wastewater resource recovery processes, GHG, carbon footprint, integrated urban water system.

WEST is a part of the MIKE family. By prioritising the refinement of plant design, operations, and automation, WEST facilitates the optimisation of effluent quality, energy consumption, and cost efficiency. Equipped with dynamic condition simulations and advanced experimental features such as Uncertainty and Sensitivity Analysis, WEST emerges as an indispensable asset for the evolution of wastewater treatment facilities towards a future-ready state.

GPS-X is provided by the Hydromantis group and contains a suite of sophisticated tools allowing to create advanced plant layouts, run interactive simulations, and perform in-depth analysis on model results. It includes a set of robust process controllers to create treatment plant layouts with complex process control schemes.

The SIMBA simulation system is provided by IFAK and is a versatile software for modelling and dynamic simulation in water and wastewater engineering. The simultaneous modelling of a wide variety of systems (such as sewage treatment plants, sewer networks, drinking water networks, rivers and also biogas plants), including control concepts, allows an integrated view of all components, flows and material fluxes and well-founded process engineering, energy and control engineering analyses and optimisations.

STOAT is a modelling tool designed to dynamically simulate the performance of wastewater treatment works (WWTW). Used worldwide, the software can be used to simulate individual treatment processes or the whole treatment works, including sludge processes, septic tank imports and recycling. The model enables the user to optimise the response of the works to changes in the influent loads, works capability or process operating conditions. STOAT is probably the only software without a license fee.

4.10 Conclusion

In the design of wastewater systems, design based on individual parameters is increasingly being replaced by the use of statistical methods. Operating data is often used for this purpose. SCADA systems can supply this operating data and must be processed accordingly. Support can also be provided by real-time data, which improves the methods and models used. Statistical methods in conjunction with visualization tools are therefore an important tool for supporting dimensioning and planning processes.

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

5 Visualization with big data

5.1 Introduction to big data visualization

5.1.1 Definition of big data visualization

Data visualization is representing data in some systematic form including attributes and variables for the unit of information (Khan and Khan 2011). In simple word, data visualization is the representation of data in graphical or visual formats, such as charts, graphs, and maps, to help people understand the patterns, trends, and insights within the data. Data must be appropriately collected and processed in order for it to be visualized as shown in Figure 25.

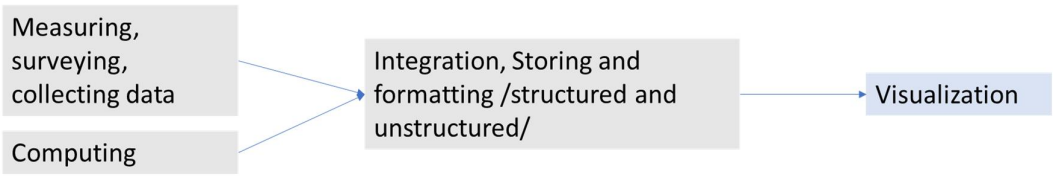


Figure 25. Data handling for visualization

An essential part of data analysis and communication is data visualization, which makes it possible for people to explore, interpret, and present information more effectively than they could with just raw data. Processing and analyzing enormous amounts of data that are larger than what can be handled by conventional data processing tools is known as "big data." Big data visualization, on the other hand, refers to the application of graphical representations of sizable and intricate datasets. Big data differs significantly from regular data in terms of scalability, structure, and size. In this situation, visualization becomes essential because it helps make sense of the large amount of information by presenting it in an understandable and interpretable format. To help with the definitional understanding of big data visualization, Table 7 summarizes some of its features.

Table 7. Features of big data visualization comparing to the traditional data visualization

Features	Notes
Complex Data Representation	Big data often involves diverse and intricate datasets with multiple variables and dimensions. Visualization tools enable the representation of these complex relationships in a way that is visually intuitive.
Pattern Recognition	Visualization aids in the identification of patterns, correlations, and outliers within large datasets. By presenting data visually, analysts can quickly identify trends and make data-driven decisions.
Interactivity	Many visualization tools for big data offer interactive features that allow users to explore data dynamically. This interactivity enables users to zoom in on specific data points, filter information, and gain deeper insights by interacting with the visual representation.
Real-Time Monitoring	In some cases, big data visualizations are used for real-time monitoring of data streams. This is particularly important in applications such as financial trading, network monitoring, and industrial processes.
Storytelling and Communication	Visualization serves as a powerful storytelling tool, helping communicate complex findings to a broader audience. It allows stakeholders to understand the significance of data insights and facilitates decision-making.
Scalability	Big data visualization tools are designed to handle the scalability challenges posed by large datasets. They should be capable of efficiently rendering and displaying visualizations even when dealing with massive volumes of data.

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

5.1.2 Importance of data and big data visualization

Raw data of any kind is challenging to interpret. Once data is properly visualized, it becomes valuable. Therefore, the main objective of data visualization is to extract insights from the data and highlight patterns, trends, and outliers. Also, the visualization is one way that human and machine interact through the data.

Big data visualization is crucial in the data-driven world of today. Visualization becomes an essential tool for deriving meaningful insights, making well-informed decisions, and communicating complex information as the volume, variety, and velocity of data continue to rise. Here, we use an example to highlight a few important points of big data visualization.

Big data visualization enables understanding complex data.

Volume and Variety: Big data often involves large and diverse datasets that may be challenging to comprehend in raw form. Visualization simplifies complex data, making it easier for individuals to understand patterns, trends, and relationships.

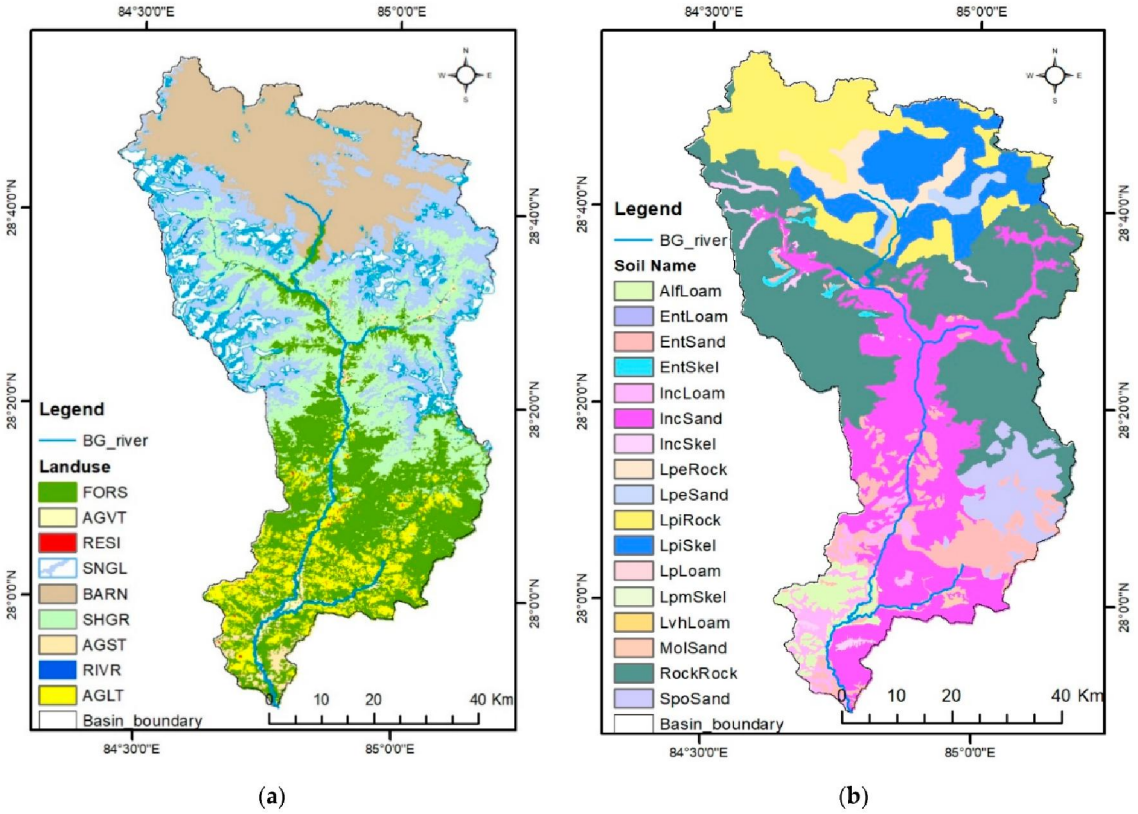


Figure 26. When combined with time series, geographic data is among the more complex big data types. Quick comprehension of the territory's soil distribution pattern and land use is provided by this data visualization. (Source: (Marahatta, Devkota and Aryal 2021))

Big data visualization helps identifying patterns and trends.

Pattern Recognition: Analysts can find patterns and trends in large datasets by using visualization tools. Finding correlations, anomalies, and other important insights is made simpler when data is represented visually.

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

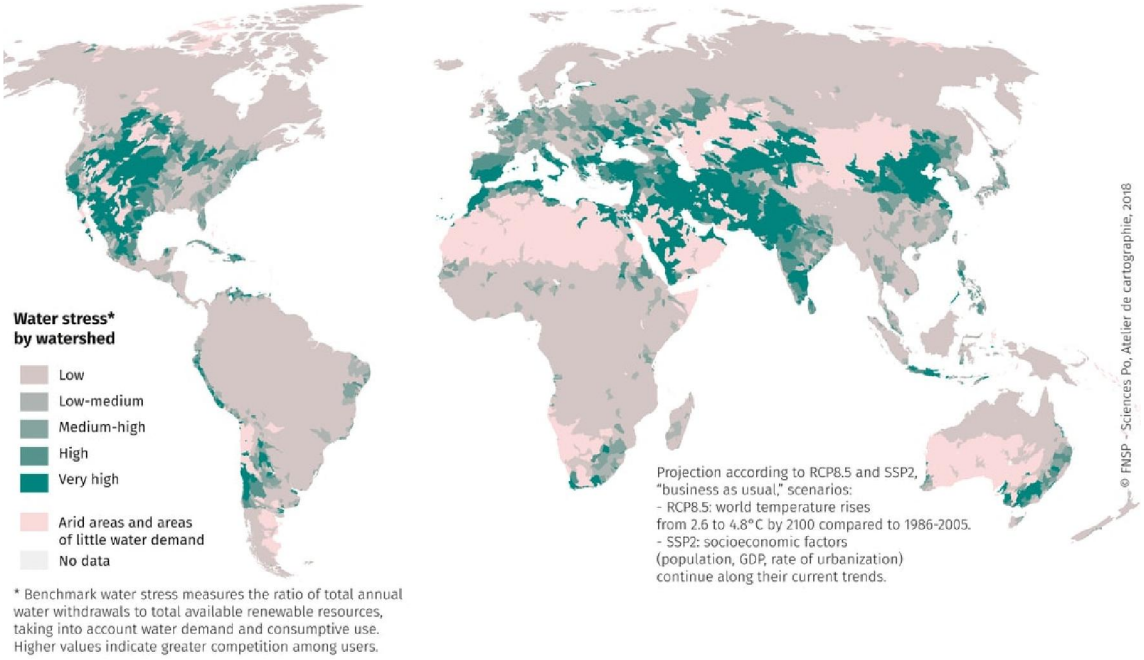


Figure 27. In this figure, it is clear that by 2040, there will be greater water stress in densely populated areas. Information is more visible when it is coded in color. (source: World Resources Institute, www.wri.org).

Big data visualization offers enabling data exploration.

Interactivity: Users can explore data dynamically with the help of interactive features found in many big data visualization tools. As a result, analysts can filter data, delve deeper into particular areas of interest, and learn more about the data.

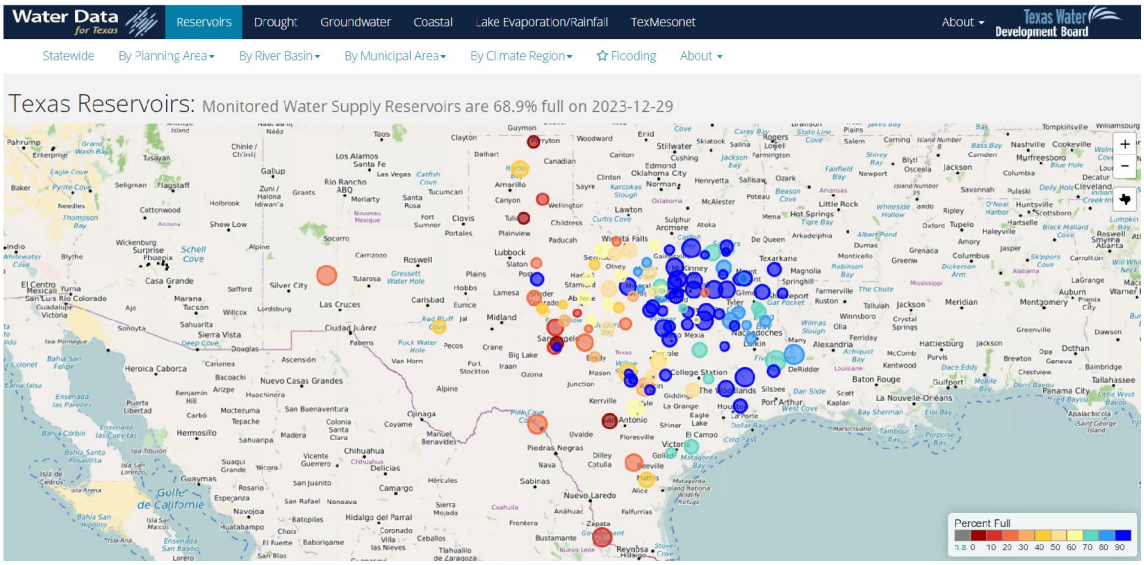


Figure 28. An example of an interactive big data visualization is the water data for Texas. This dashboard gives us the ability to view a specific water body—a reservoir—and obtain pertinent data over various time periods. (source: https://www.waterdatafortexas.org/)

Big data visualization supports facilitating decision-making.

Data-Driven Decision-Making: Data-driven decision-making is aided by visualization, which gives important information a visual representation. Based on visual insights, decision-makers are able to quickly comprehend complex scenarios and make more informed decisions.

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

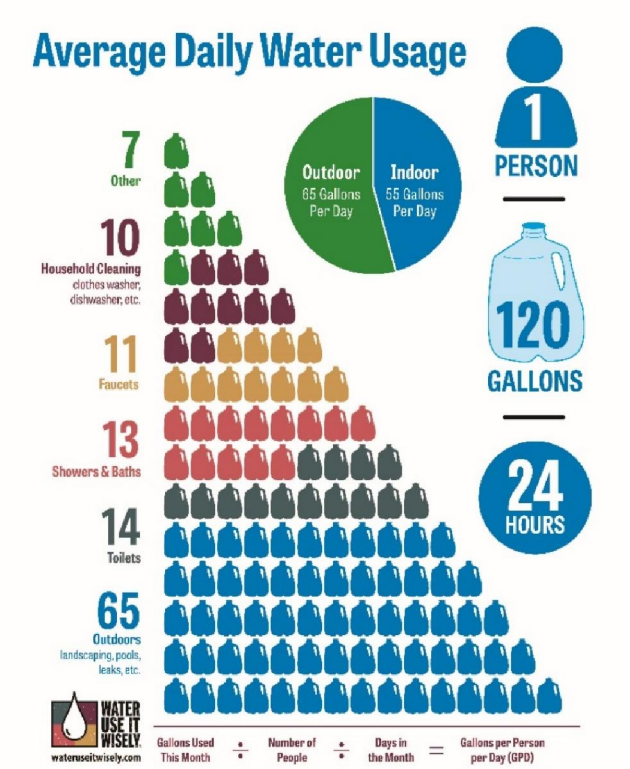


Figure 29. US citizens' average daily water use is displayed in this infographic. Based on each person's unique water usage data, a decision to reduce water consumption can be made both in outdoor and indoor activities. (source: AMWUA)

Big data visualization helps communication across teams.

Cross-Functional Collaboration: Visualization is an effective means of communication across teams with diverse expertise. Non-technical stakeholders can understand and contribute to data discussions when information is presented visually, fostering collaboration between data scientists, analysts, and business professionals.

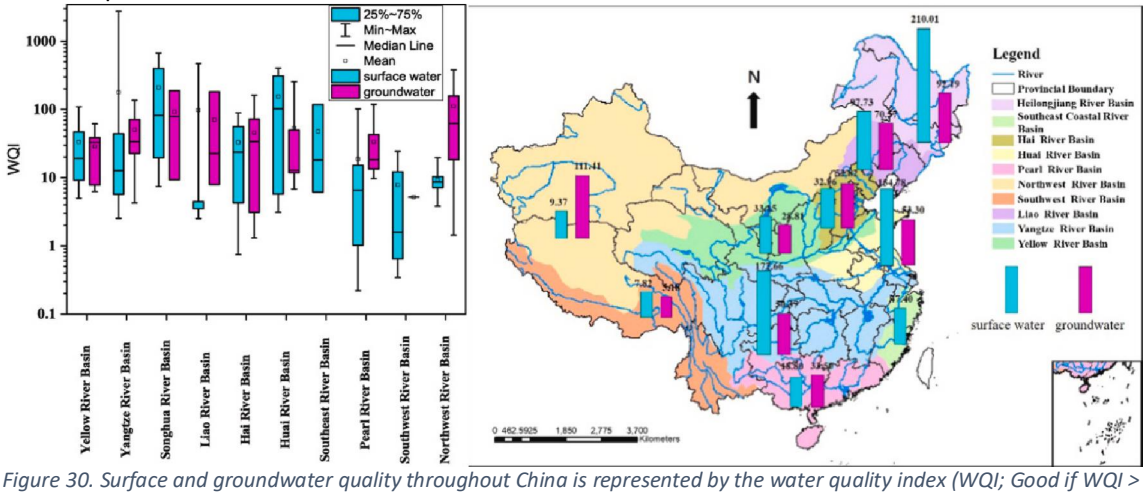


Figure 30. Surface and groundwater quality throughout China is represented by the water quality index (WQI; Good if WQI > 70) of river basins and territory; non-technical stakeholders across the country can easily understand this data (Tong, et al. 2021).

Big data visualization brings timely insight in real-time monitoring.

Timely Insights: Big data visualizations provide a means to track and react to changes in real-time when data monitoring is crucial. This is even crucial in water industry.

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

Figure 31. Water distribution network powered and monitored by digital twin platform. Monitoring the system in real time with significant physical variables like pressure and demand can lead to effective management (Source: esri.com).

Big data visualization carries enhancing data storytelling.

Narrative Context: Visualization enhances data storytelling by providing a visual narrative. Instead of presenting isolated data points, visualizations create a cohesive story that guides the audience through the data, making it more engaging and memorable.

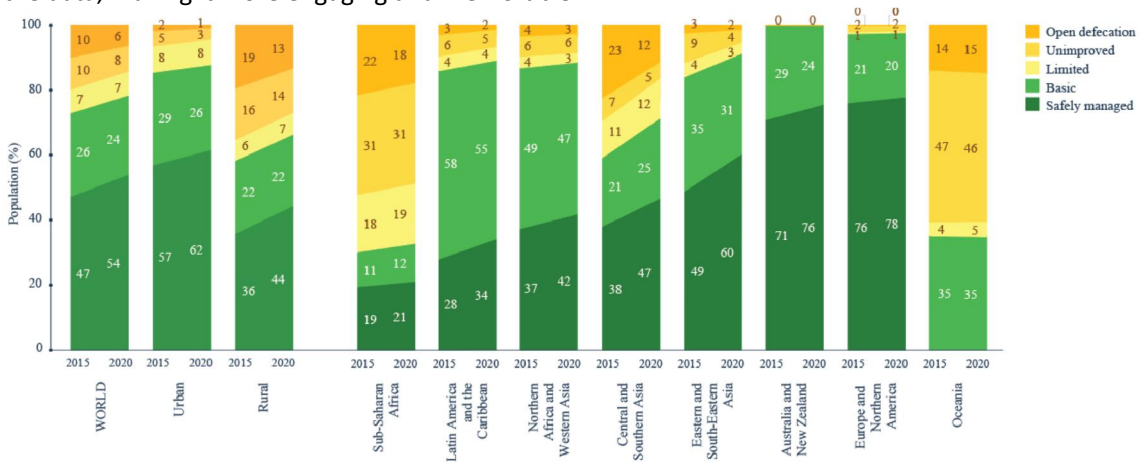


Figure 32. Global and regional sanitation coverage 2015–2020 show how we achieve the SDG 6 in past 5 years in different regions (WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene).

Big data visualization can handle Large Datasets

Scalability: Big data visualization tools are designed to handle the scalability challenges posed by large datasets. They can efficiently process and render visualizations even when dealing with massive volumes of data.

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

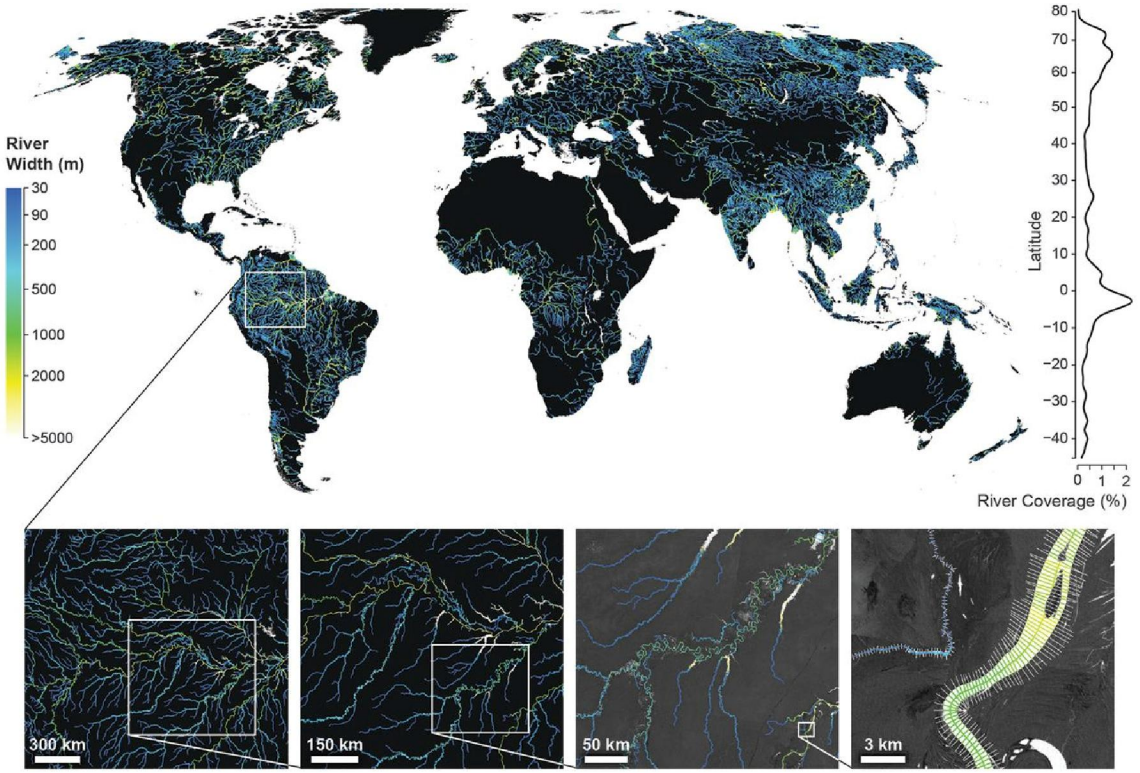


Figure 33. The Global River Widths from Landsat (GRWL) Database contains more than 58 million measurements of planform river geometry (Allen and Pavelsky 2018).

Big data visualization improves Accessibility of big data.

Democratizing Data: Visualization tools help democratize data by making it accessible to a broader audience. Users across different levels of technical expertise can derive insights from visualizations, reducing the reliance on specialized data skills.

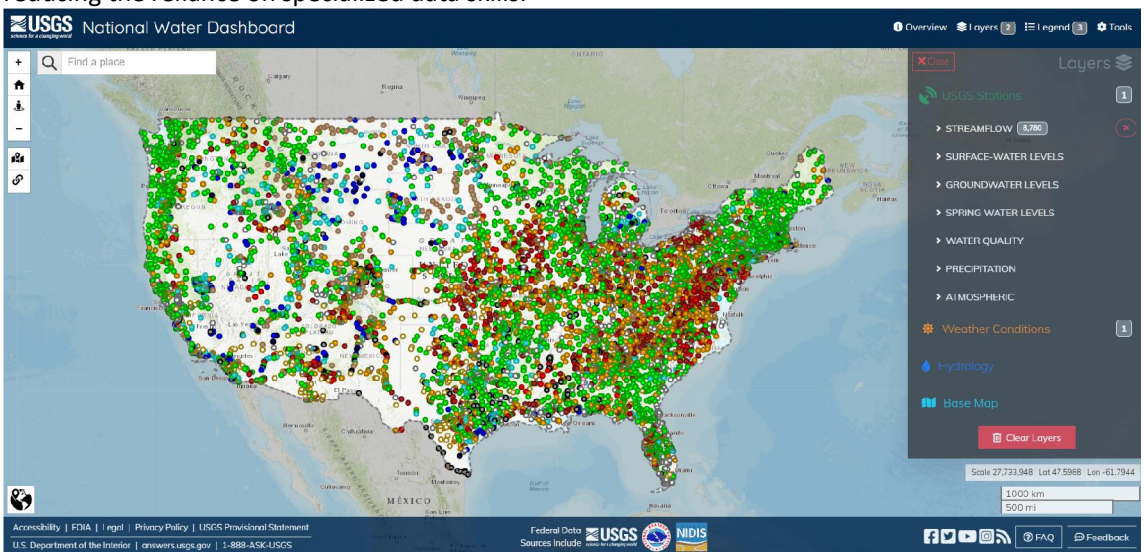


Figure 34. Accessible water data for different levels of expertise is provided by the USGS National Water Dashboard.

Big data visualization mitigates the risk.

Identifying Anomalies: Visualization aids in the identification of outliers and anomalies within big datasets, which is crucial for risk management and fraud detection in various industries.

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

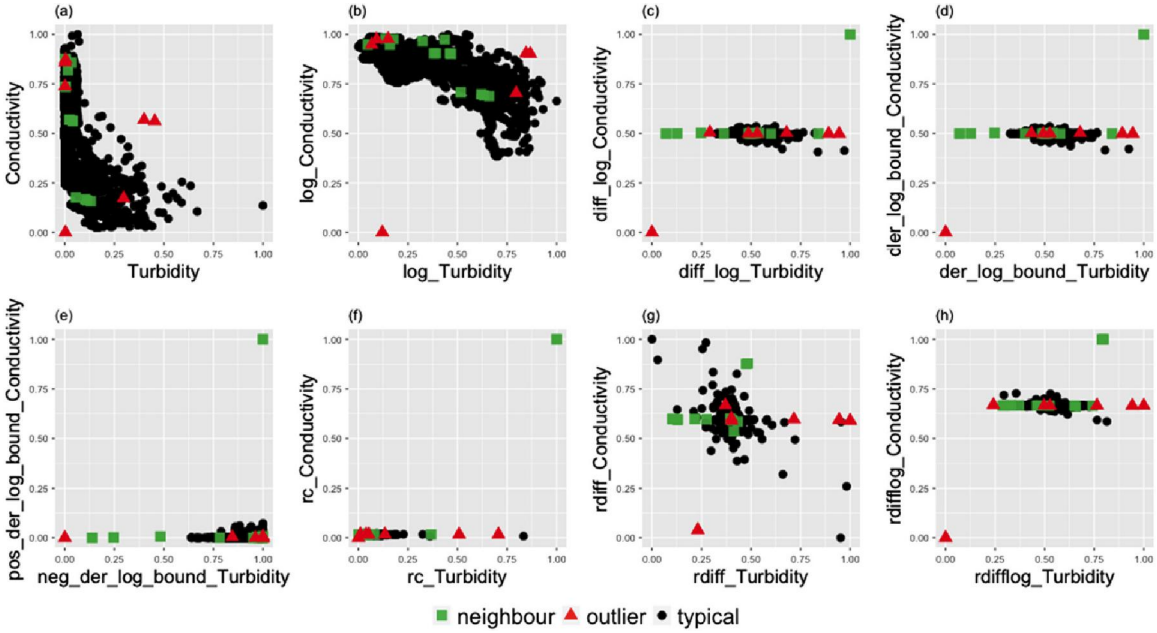


Figure 35. Bivariate relationships between transformed series of turbidity and conductivity measured by in situ sensors at Sandy Creek. In each scatter plot, outliers determined by water-quality experts are shown in red, while typical points are shown in black. Neighboring points are marked in green (Talagala, et al. 2019).

Data visualization saves time, puts the data into the correct context, provides perspective on data, discovers the trends in data.

The massive volumes of data created in today's digital environment must be transformed into actionable insights, and this requires big data visualization. It is essential for decision-making, communication, and teamwork among heterogeneous groups in addition to facilitating data exploration and analysis. Through the use of visualization-based data discovery techniques, researcher can combine different data sources to produce unique analytical views. Advanced analytics can be integrated in the methods to support creation of interactive and animated graphics on desktops, laptops, or mobile devices such as tablets and smartphones (Wang, Wang and Alexander 2015). Table 8 (SAS 2013) shows the benefits of data visualization according to the respondent percentages of a survey.

Table 8. Benefits of data visualization tools (SAS 2013)

Benefits	Percentages (%)
Improved decision-making	77
Better ad-hoc data analysis	43
Improved collaboration/information sharing	41
Provide self-service capabilities to end users	36
Increased return on investment (ROI)	34
Time savings	20
Reduced burden on IT	15

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

5.2 Big data visualization techniques

5.2.1 Traditional data visualization techniques

Data must be visualized in order for conclusions to be drawn from the data science process, which includes data collection, processing, and modeling as necessary steps. Data visualization also aims to identify, locate, manipulate, format and deliver data in the most efficient way possible. Data visualization, which allows for the visualization of both numerical and categorical data and increases the impact of insights, is essential to engineering and research communication. Data visualization can be categorized by data type, as shown in Figure 36.

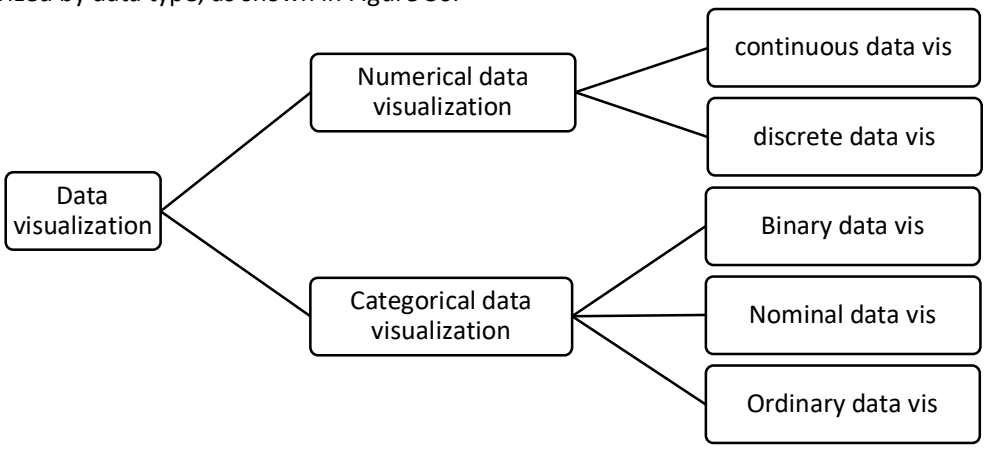


Figure 36. Category of data visualization and data types

Numerical data is also known as Quantitative data. Numerical data is any data where data generally represents amount such as discharge, stage of river in time, etc. Numerical data visualization is easiest way to visualize data. It is generally used for helping others to digest large data sets and raw numbers in a way that makes it easier to interpret into action. Numerical data is categorized into two categories:

- Continuous Data – Continuous data often includes measurable values representing a range of information, such as the flow range and water quality data.
- Discrete Data – This type of data is not "continuous" and represents exact figures you can count, such as the number of rivers or lakes in the country.

The type of visualization techniques that are used to represent numerical data visualization is Charts and Numerical Values. Examples are Pie Charts, Bar Charts, Averages, Scorecards, etc. Numerical data can be classified as scalar, vector and tensor data. Depending on the data type as shown in Table 9, there are many possibilities to make visualization to realize the purpose of presentation.

Table 9. Commonly encountered numerical data and their visualization possibilities

Data types	Visualization possibilities	Visual possibility
Scalar or common numerical data	Points Lines Marks Areas	Types, Sizes, Colors Types, Weights, Colors Types or Shape, Size, Colors Colors, Sizes
Vector data	Lines (on grid or mesh)	Types, Weights, Colors, Lengths, Directions, Arrow head type etc.
Tensor data	Same as vector data	Types, Weights, Colors, Lengths, Directions, Arrow head type etc.

Categorical data is also known as Qualitative data. Any data where the data typically represents groups is considered categorical data. All of its constituents are categorical variables, which are employed to express attributes like countries development and their rankings, etc. The main goals of categorical data visualization are to illustrate important themes, make connections, and provide context. Categorical data is classified into three categories:

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

- Binary Data – In this, classification is based on positioning (Example: Agrees or Disagrees).
- Nominal Data – In this, classification is based on attributes (Example: Male or Female).
- Ordinal Data – In this, classification is based on ordering of information (Example: Timeline or processes).

The type of visualization techniques that are used to represent categorical data is Graphics, Diagrams, and Flowcharts. Examples are Word clouds, Sentiment Mapping, Venn Diagram, etc.
The majority of visualization tasks related to climate change and water resources center on demonstrating various aspects such as changes in time and space, hierarchy, wholes and parts, probabilities, distributions, correlations, comparisons, and connectivity. Various techniques are developed for the visualization in this context. It should be mentioned that use cases are now given more consideration in data visualization. As a result, user-friendly visualization is essential, enabling users to read, comprehend, generate, and share data as information.

5.2.2 Big data visualization in big data analysis

Currently huge amounts of data are constantly generated and stored in the vaults in the various fields specially in climate change and water resources. It is often not only the amount of data is great, but these data are constantly updated and supplemented with the new ones. In addition, there is a very wide variety of data types and sources. Such data is called big data (Ruzgas 2016). Processing and analysis of big data are to get insight from data and to make data simple and useful. Different methods are being developed for big data investigation tasks, such as clustering, classification, statistical and visual analysis. Study of big data is one of the biggest challenges faced by data analytics and researchers so that the methods and tools for traditional data analysis, is inappropriate for big data analysis.

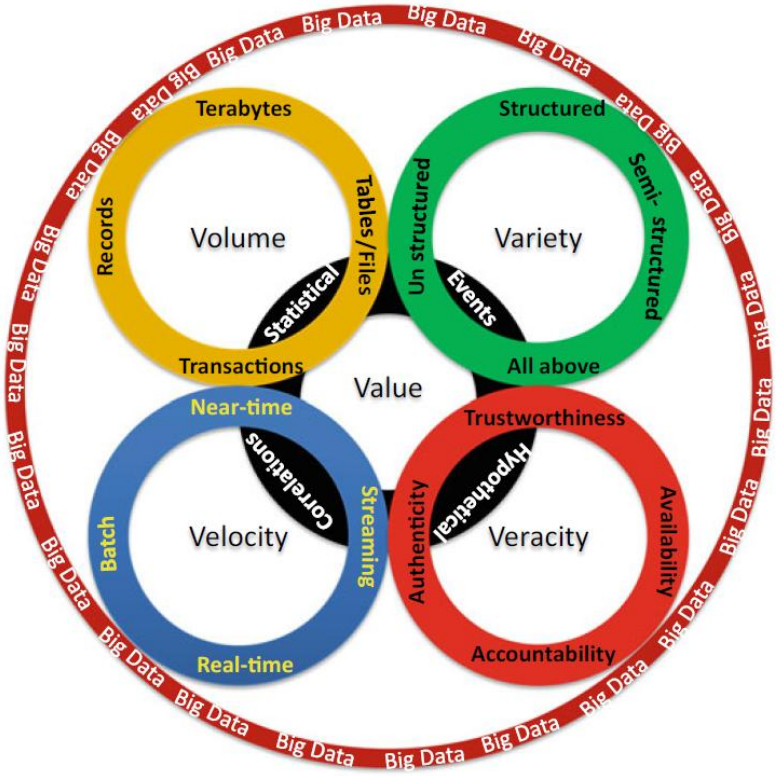


Figure 37. Big data 5V's (Raval and Kumar 2020).

Big data is described by 5V's characteristics shown in Figure 37, which are also important for the visualization and analytics (Miller 2017).

- Volume - one of the big characteristics, illustrating the size of the data. Volume of data is always increases and big data analytics is to make it simple.

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

- Variety - the characteristic that defines difference of the data types and diversity of data. Variety of data is always increases.
- Velocity - the term is understood as the satisfaction of a requirement for updating of data, growth and processing. In other word, it is speed of data generation and it always increase and make volume increase.
- Veracity - the term is associated with the correctness and accuracy of the data and their analysis.
- Complexity or Value - associated with the ever-growing amounts of data, their worth, variety and the problems arising from the analysis of these data.

Those characteristics are also important for the big data management and big data analysis. The scheme for processing big data is presented in Figure 38. The main steps of data processing are shown in the picture at the top. The data features which make processing of data complex and complicated are presented in the bottom. The scheme demonstrates that the data processing is composed of many steps, each of which is encountering some challenges, requiring appropriate solutions. Big data are high volume, high velocity, and/or high variety datasets that require new forms of processing to enable enhanced process optimization, insight discovery and decision making. Challenges of Big Data lie in data capture, storage, analysis, sharing, searching, and visualization (Chen and Zhang 2014). Data visualization is important in data management and in data analysis to make appropriate interpretation from the results.

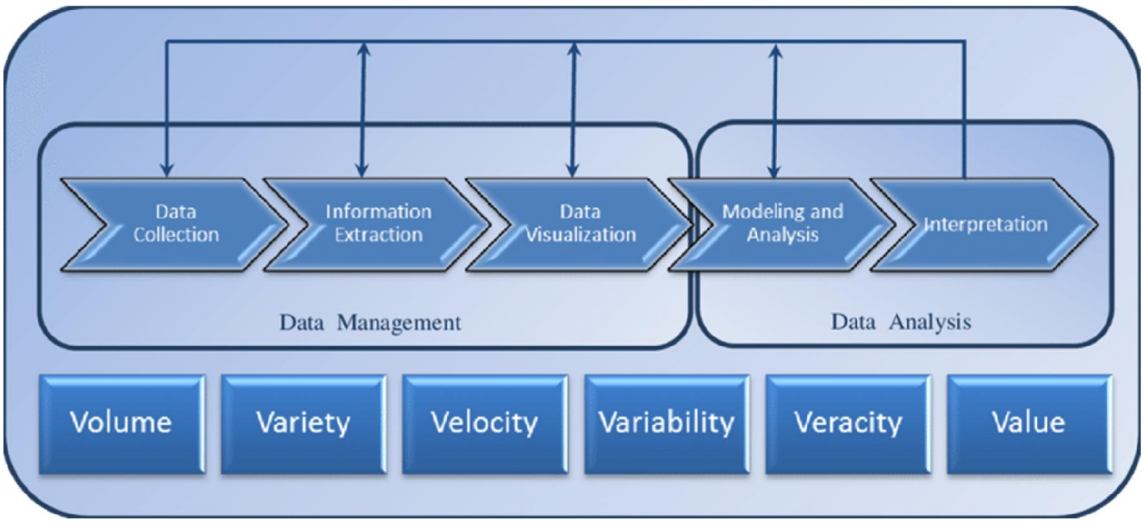


Figure 38. Big data processing scheme (Ruzgas 2016)

If the visualization contains too much information in one place, the reader cannot conceive the main information. In addition, complicated or intricate visuals or those that attempt to aggregate or otherwise source a large number of data sources most likely will be hindered by the experiences of slow performance (Miller 2017). Without context, data is meaningless and the same applies to visualization of that data. Acquiring a proper understanding of the data takes significant domain expertise as well as the ability to properly analyze the data; big data certainty complicates these practices with its seemingly endless number of formats and varieties of both structured and unstructured data (Miller 2017). The effort and expense required to source, understand, and visualize data is squandered if the results are stable, obsolete, or potentially invalid by the time the data is available to the intended consumer. The challenge of speedily crunching numbers exists within any data analysis, but when considering the varieties and volumes of data involved in big data visualization, it becomes even more evident.

Visualization can be thought of as the “front end” of big data. There are following data visualization myths (Simon 2014):

- All data must be visualized: It is important not to overly rely on visualization; some data does not need visualization methods to uncover its messages.

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

- Only good data should be visualized: A simple and quick visualization can highlight something wrong with data just as it helps uncover interesting trends.
- Visualization will always manifest the right decision or action: Visualization cannot replace critical thinking.
- Visualization will lead to certainty: Data is visualized doesn't mean it shows an accurate picture of what is important. Visualization can be manipulated with different effects.

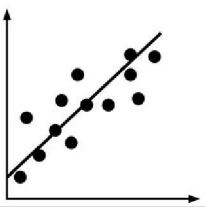
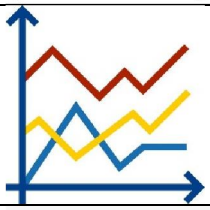
Visualization approaches are used to create tables, diagrams, images, and other intuitive display ways to represent data. Big Data visualization is not as easy as traditional small data sets. The extension of traditional visualization approaches has already been emerged but far from enough. In large-scale data visualization, many researchers use feature extraction and geometric modeling to greatly reduce data size before actual data rendering. Choosing proper data representation is also very important when visualizing big data (Chen and Zhang 2014).

5.2.3 Big data visualization techniques

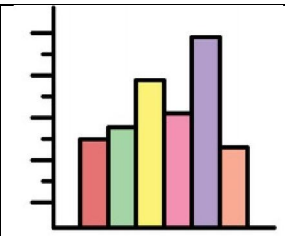
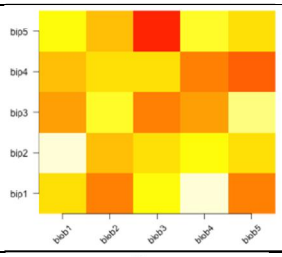
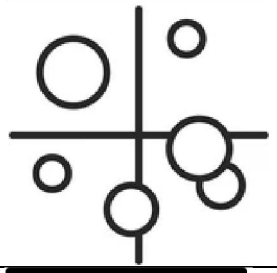
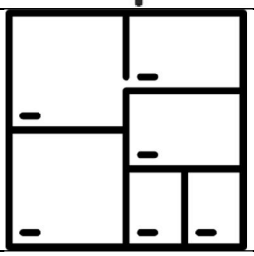
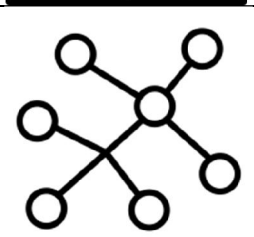
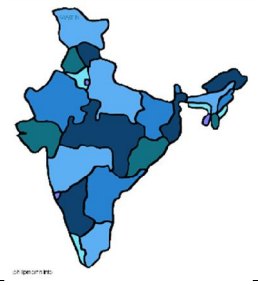
Big data visualization involves a variety of techniques to represent and analyze large and complex datasets. The choice of visualization technique depends on the nature of the data, the insights you want to gain, and the story you want to convey. It's often beneficial to use a combination of techniques to provide a comprehensive understanding of big datasets. Additionally, advancements in technology and visualization tools continue to expand the range of available techniques for handling big data.
For the Big data, variables can be visualized by:

- Size
- Shape
- Color
- Sharpness
- Visual effects
- Motion
- Weight or the combination of them.

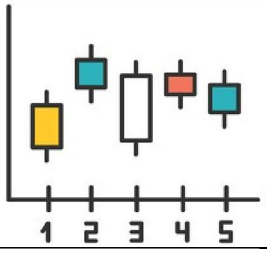
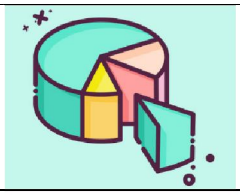
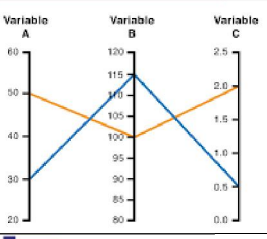
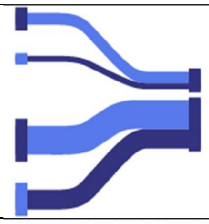
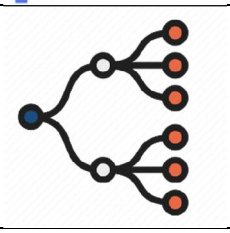
Table 10. Common techniques used for big data visualization

Techniques	Purpose	Usage	Presentation
Scatter Plots	Shows the relationship between two variables.	Useful for identifying patterns, correlations, and outliers.	
Line Charts	Depicts trends over time or across categories.	Effective for illustrating changes and patterns in data.	
Bar Charts	Compares individual values or categories.	Suitable for displaying discrete data points and making comparisons.	

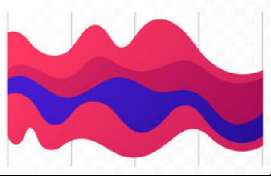
Co-funded by the
Erasmus+ Programme
of the European Union

Histograms	Displays the distribution of a single variable.	Useful for understanding the frequency and spread of data.	
Heatmaps	Shows the intensity of data values using color.	Effective for visualizing patterns in large datasets.	
Bubble Charts	Represents three dimensions of data using circles of different sizes.	Useful for visualizing relationships among three variables.	
Treemaps	Displays hierarchical data as nested rectangles.	Effective for illustrating hierarchical structures and proportions.	
Network Graphs	Illustrates relationships and connections between entities.	Useful for visualizing complex networks and dependencies.	
Choropleth Maps	Displays spatial variations using color-coded regions.	Effective for representing geographical patterns and distributions.	

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

Word Clouds	Represents the frequency of words in a text dataset.	Useful for identifying prominent themes or keywords.	
Box Plots (Box-and-Whisker Plots)	Displays the distribution and variability of a dataset.	Useful for identifying outliers and understanding the spread of data.	
Time Series Visualizations	Represents data points collected over time.	Effective for analyzing trends, seasonality, and patterns over time.	
3D Visualizations	Represents data in three-dimensional space.	Useful for visualizing complex relationships in multi-dimensional datasets.	
Parallel Coordinates	Displays multi-dimensional data using parallel axes.	Effective for visualizing relationships and patterns in high-dimensional datasets.	
Sankey Diagrams	Illustrates flow and relationships between different entities.	Useful for representing processes, flows, and resource allocation.	
Dendrogram	Represents hierarchical relationships in a tree-like structure.	Useful for clustering and displaying hierarchical structures in data.	

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

Sunburst Charts	Represents hierarchical data in a radial layout.	Useful for visualizing hierarchical structures with multiple levels.	
Streamgraphs	Displays the evolution of data over time in a stacked area chart.	Effective for visualizing trends and changes in data over time.	

Visualization with big data plays a crucial role in transforming complex data into meaningful insights, enabling informed decision-making and facilitating communication across various stakeholders. Some tips for using data visualization techniques

- Choose the right chart type: Select the right graphics for your specific project, audience, and purpose
- Use all your senses: Involve sight, smell, touch, taste, and hearing
- Use visual hierarchy effectively: Use different effects of computer graphics
- Label your data visualizations: Show logical correlations between units
- Persuade with visualizations: Use grouping, highlighting, and annotation to guide viewers toward a certain comparison or pattern
- Keep your data clean: Before visualizing your data, make sure to fix or remove incomplete, duplicate, incorrect, corrupted and incorrectly formatted data within your dataset.
- Use the right visuals: With so many charts available, identify the best type for presenting the particular data type you're working on.
- Keep your data organized: At a glance, your audience should be able to view and digest information quickly.
- Use the right color combination.

Visualizations are not only static; they can be interactive. Interactive visualization can be performed through approaches such as zooming (zoom in and zoom out), overview and detail, zoom and pan, and focus and context or fish eye. The steps for interactive visualization are as follows (Khan and Khan 2011):

1. Selecting: Interactive selection of data entities or subset or part of whole data or whole data set according to the user interest.
2. Linking: It is useful for relating information among multiple views. An example is shown in Figure 3.
3. Filtering: It helps users adjust the amount of information for display. It decreases information quantity and focuses on information of interest.
4. Rearranging or Remapping: Because the spatial layout is the most important visual mapping, rearranging the spatial layout of the information is very effective in producing different insights.

5.2.4 Strategy for big data visualization

There are various approaches or strategies that have come to exist and can be used for preparing effective big data visualizations as well as addressing the hindrances as variety, velocity, volume and veracity of big data. Some of the examples include (Miller 2017):

- The concepts and models are always necessary to efficiently and effectively visualize the big data. Therefor start with conceptualization for visualization of your big data.
- Changing the type of the visualization, for example, switching from a column graph to a line chart can allow you to handle more data points within the visualization.

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

- You can use higher-level clustering. In other words, you can create larger, broader stroke groupings of the data to be represented in the visualization rather than trying to visualize an excessive number of groups.
- You can remove outliers from the visualization. Outliers typically represent less than 5 percent of a data source, but when you are working with massive amounts of data, viewing that 5 percent of the data is challenging. Outliers can be removed and if appropriate, be presented in a separate data visualization.
- You can consider capping, which means setting a threshold for the data you will allow into your visualization. This cuts down on the range or data making for a smaller, more focused image.

There are some choice of visualization factors that need to be considered for the visualization.

- Audience: The information depiction should be adjusted to the target audience. On the other side, when information ideas are for scientists or seasoned decision-makers, you can and should often go beyond easy diagrams.
- Satisfaction: The data type determines the strategies. For instance, when there are metrics that change over the moment, the dynamics will most likely be shown with line graphs. You will use a dispersion plot to demonstrate the connection between two components. Bar diagrams are ideal for comparison assessment, in turn.
- Context: The way your graphs appear can be taken with distinct methods and therefore read according to the framework. For instance, you may want to use colors of one color to highlight a certain figure, which is a major increase relative to other years, and choose a shiny one as the most important component on the graph. Instead, contrast colors are used to distinguish components.
- Dynamics: Data are distinct and each means a distinct pace of shift. For example, each month or year the stream flow results can be measured while time series and data tracking change continuously. Dynamic representation (steaming) or a static visualization can be considered, depending on the type of change.
- Objective: The objective of viewing the information also has a major effect on the manner in which it is carried out. Visualizations are built into dashboards with checks and filters to carry out a complicated study of a scheme or merge distinct kinds of information for a deeper perspective. Dashboards are, however, not required to display one or more occasional information.

5.3 Tools for big data visualization

5.3.1 Types of tools for big data visualization

Various tools have emerged to help us to visualize the big and traditional data. The tools can be

- Open sources,
- Licensed,
- Online (web-based) and offline,
- Integrated with software or
- Interactive and non-interactive.

Popular visualization tools over internet are shown in Figure 39.

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

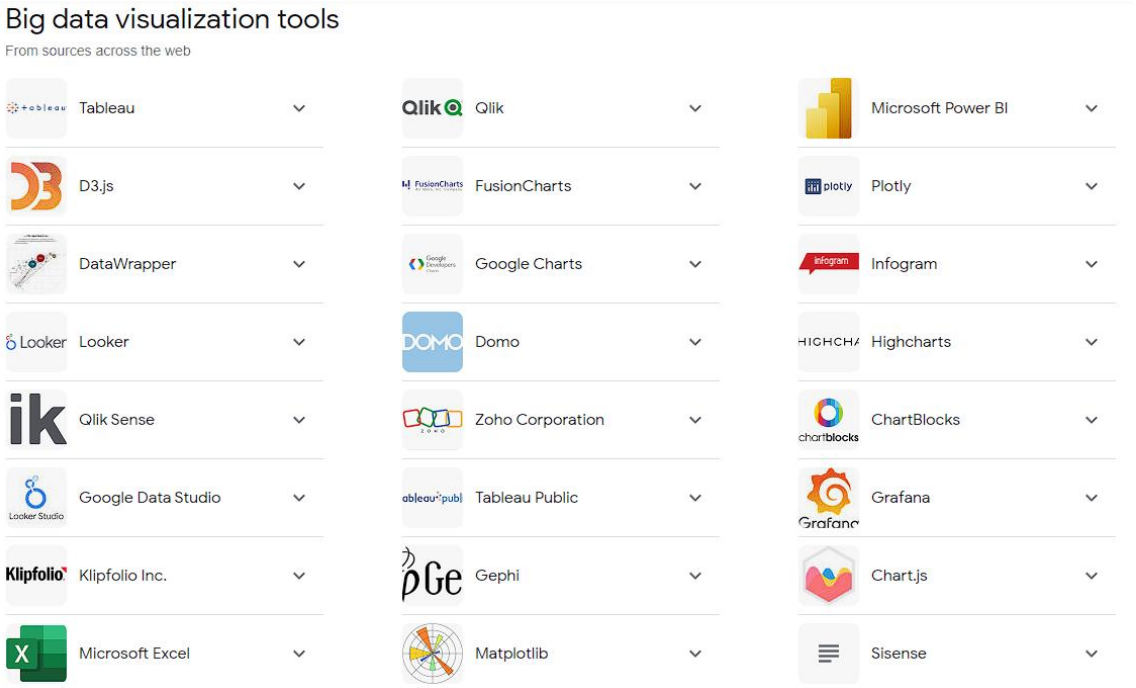


Figure 39. Big data visualization tools

Each of tools have inuque features designed for the 5V's. The most important feature that a visualization must have is that it should be interactive, which means that user should be able to interact with the visualization. Visualization must display relevant information when hovered over it, zoom in and out panel should be there, visualization should adapt itself at runtime if we select subset or superset of data (Ali, et al. 2016).

Selecting the right visualization tool is crucial for effectively communicating insights from your data. The choice depends on various factors, including the type of data, the complexity of analysis, the audience, and the desired level of interactivity. Additional factors may be data type, purpose of visualization, target audience, interactive requirement, complexity of data, platform compatibility, cost consideration, ease of use, available features, scalability, community support and documentation, security and compliances. By carefully considering these factors, you can make an informed decision when selecting a visualization tool that best aligns with your specific requirements and objectives.

5.3.2 Challenges of big data visualization

While big data visualization is a powerful tool for conveying complex information and patterns, it also has its limitations. Scalability and dynamics are two major challenges in visual analytics. In Big Data applications, it is difficult to conduct data visualization because of the large size and high dimension of big data. Most of current Big Data visualization tools have poor performances in scalability, functionalities, and response time. Uncertainty can result in a great challenge to effective uncertainty-aware visualization and arise during a visual analytics process. Perceptual and interactive scalability are also challenges of big data visualization. Visualizing every data point can lead to over-plotting and may overwhelm users' perceptual and cognitive capacities; reducing the data through sampling or filtering can elide interesting structures or outliers.

Here are some common limitations associated with big data visualization:

- Over-Simplification: Big data often involves vast and intricate datasets. Visualizations, by nature, simplify complex information to make it understandable. However, this simplification may lead to the loss of nuance and detail, potentially oversimplifying the understanding of complex relationships within the data.
- Interpretation Challenges: Users may misinterpret visualizations if they lack the necessary context or understanding of the data. This can lead to flawed conclusions and

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

decision-making. Providing proper context and documentation is essential to mitigate this limitation.

- Scalability Issues: Handling and visualizing extremely large datasets can be challenging. Traditional visualization tools may struggle to scale effectively, leading to slow rendering times, reduced interactivity, and potential information loss.
- Bias in Data Representation: The choice of visualization types, color schemes, and other design elements can introduce bias into the representation of data. Unintentional biases may affect the way users perceive and interpret the information.
- Data Privacy Concerns: Visualizations often involve aggregating and presenting data at different levels. However, the aggregation process can potentially reveal sensitive information when dealing with individual-level data. Protecting privacy while still providing valuable insights is a significant challenge.
- Dynamic Nature of Data: Big data is dynamic, and patterns can change over time. Static visualizations may become outdated quickly, requiring constant updates to remain relevant and accurate.
- Complexity of Multidimensional Data: Big data often involves multidimensional datasets with numerous variables. Representing such complexity in a two-dimensional visualization can be challenging, and important relationships may be overlooked.
- Tool and Technology Dependency: The effectiveness of big data visualization is highly dependent on the tools and technologies used. If the chosen tools are not up to date or do not support specific data formats, users may face limitations in visualizing certain types of data.
- Cost and Resource Intensiveness: Building and maintaining robust big data visualization systems can be resource-intensive, requiring substantial investments in both technology and skilled personnel. Small organizations may face challenges in implementing and sustaining such systems.
- Lack of Standardization: The lack of standardization in data formats, metadata, and visualization practices across industries can hinder interoperability and the sharing of visualizations between different systems and organizations.

To overcome these limitations, it's crucial to carefully design visualizations, provide adequate context and documentation, stay mindful of potential biases, and use tools that can effectively handle the scale and complexity of big data. Additionally, continuous improvements in technology and the development of best practices in data visualization can help address some of these challenges.
Unstructured information formats such as charts, lists, text, trees, and other information are sometimes difficult to visualize. Often large information has unstructured formats. Due to the constraints on bandwidth and power consumption, visualization should step nearer to the data to effectively obtain significant information. The software for visualization should be executed on location. Due to the large volume of the information, visualization requires huge parallelization. The difficulty in simultaneous viewing algorithms is to break down an issue into autonomous functions that can be carried out at the same time.
There are also the following problems for big data visualization:

- Visual noise: Most items on the dataset are too related to each other. There are also the following issues when viewing large-scale information. Users can not split them on the display as distinct items.
- Info loss: Visible data sets may be reduced, but information loss may occur.
- Broad perception of images: data display techniques are restricted not only by aspect ratio and device resolution but also by physical perception limitations.
- The elevated pace of changes in the picture: users view information and are unable to respond to the amount of changes in information or its intensity.
- High-performance requirements: In static visualization it is hard to notice because of reduced demands for display velocity— high performance demands.

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

Designing a new visualization tool with efficient indexing is not easy in big data. Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability. Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata. Big data often has unstructured formats. Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently. Visualization software should be run in an in-situ manner. Because of the big data size, the need for massive parallelization is a challenge in visualization. The challenge in parallel visualization algorithms is decomposing a problem into independent tasks that can be run concurrently (Childs, et al. 2013). Potential solutions to some challenges or problems about visualization and big data were presented (Wang, Wang and Alexander 2015):

- Meeting the need for speed: One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.
- Understanding the data: One solution is to have the proper domain expertise in place.
- Addressing data quality: It is necessary to ensure the data is clean through the process of data governance or information management.
- Displaying meaningful results: One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized.
- Dealing with outliers: Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

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

5.4 Big data visualization in water resources management

5.4.1 Big data in water resources management

Water resources management constitute a Big Data issue and grows progressively. The recent evolutions in web technology and computer science provide the water resources discipline with continuously expanding tools for data collection and analysis that pose challenges to the design of analysis methods, and interaction with data sets (Schnase, et al. 2017). The demand on water has increased due to population growth as a result of economic development, while a several regions suffer from flooding and drought, leading to water resources mismanagement. On the other hand, climate change exerts great impacts on water systems and caused great changes in water resources due to its direct effects on hydrological processes such as precipitation, evaporation and humidity. The combination of growth on the demand for water, climate and hydrological gap pushed decision makers and managers of water resources to look for strategies for effective management of water resources (Chalh, et al. 2015).

The complexity of water resources problems is characterized by the interaction of several physical phenomena. Water problems include preservation of water irrigation, watershed management, dam construction for mitigation floods and/or conservation purpose, river management, basin management, pollution control (Dabaghi, Ouazar and Prastacos 2001). The problems relating to water resources are featured by (Chalh, et al. 2015):

- Huge volume of collected, analyzed, computed and visualized data.
- Data collected are complex in dimension, size and heterogeneity (tsunamis of data, or Data Deluge).
- Many different data sources, multi-scale, multi-models.
- Multidisciplinary is required.
- Spatial data, remote sensing in real time.
- Heterogeneous data resulting from various sophisticated simulation models that can ironically, create more of a Big Data challenge than the experimental sciences they are supposed to complement or replace.
- Large simulations are becoming unavoidable tackling all the scientific aspects at multiple scales.

To facilitate the management of water resources we need to think about improving the adoption of new technologies. Those new technologies such as big data analytics, machine learning, AI and digital twins all exploits the visualization. In general, two main processes, i.e., data management and analytics are used for extracting meaningful results from the big water data. The term data management can be defined as the acquisition of data, its temporary storage and final preparation for suitable for analysis (Figure 38). Analytics refer to methods utilized to investigate and get conclusive findings from big data (Raval and Kumar 2020). Representation in data management and interpretation in big data analytics requires careful visualization.

The majority of water resource variables that require big data visualization are listed below.

Hydrological Variables:

- Water Level: Measurements of water depth in rivers, lakes, and reservoirs.
- Flow Rate: The volume of water passing through a specific point over time.
- Precipitation: Amount of rainfall or snowfall in a given area.
- Soil Moisture: The amount of water present in the soil.

Meteorological Variables:

- Temperature: Air temperature data.
- Humidity: The amount of moisture in the air.
- Wind Speed and Direction: Information about wind patterns.
- Barometric Pressure: Atmospheric pressure at a specific location.

Water Quality Variables:

- pH: A measure of acidity or alkalinity in water.
- Dissolved Oxygen: The amount of oxygen dissolved in water, vital for aquatic life.
- Conductivity: The ability of water to conduct an electric current, indicating the presence of dissolved ions.

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

- Organic matter in wastewater – COD and BOD

Nutrient Levels: Concentrations of nutrients such as nitrogen and phosphorus.

Geospatial Variables:

- Latitude and Longitude: Geographic coordinates of a location.
- Land Use: Information about how land in a particular area is utilized.
- Elevation: Height above sea level.

Human Activities and Infrastructure:

- Water Usage: Data on water consumption for agriculture, industry, and domestic purposes.
- Infrastructure: Information on dams, reservoirs, irrigation systems, and wastewater treatment plants.

Ecological Variables:

- Biodiversity: Data on the variety of plant and animal species in a given ecosystem.
- Habitat Characteristics: Information about the physical features of ecosystems.

Socioeconomic Variables:

- Population Data: Information about the number of people in a given area.
- Land Ownership: Data on ownership and land use patterns.

Remote Sensing Variables:

- Satellite Imagery: Data from satellite sensors providing information on land cover, vegetation health, and water bodies.

Sensor Data:

- IoT Sensor Readings: Real-time data from sensors deployed in the field for monitoring various parameters.

Temporal Variables:

- Time Stamp: Information about the time when a data point was recorded.

5.4.2 Application and examples of big data visualization in water resources management

All aspects of water resources management need application of big data visualization. Water management scaling down from global to watershed, all activities such as planning, designing, developing and operating requires certain degree of visualization of data.

Big data visualization plays a crucial role in water resources management by providing insights into complex datasets related to water availability, quality, usage, and environmental impacts. Table summarizes some examples of how big data visualization is applied in water resources management.

Table 11. Examples of big data visualization in water resources management

Example, not limited to other application	Visualization	Benefits
Real-Time Water Quality Monitoring	Interactive dashboards displaying real-time data on water quality parameters such as pH, dissolved oxygen, and pollutant levels.	Enables immediate identification of water quality issues, facilitates quick response to pollution events, and supports informed decision-making for water treatment processes.
Water Consumption Patterns	Time-series charts and heatmaps representing water consumption patterns across different regions, industries, or residential areas.	Helps identify peak demand periods, assess the impact of water conservation measures, and optimize water distribution systems.
Flood Prediction and Monitoring	Geospatial visualizations and flood maps indicating areas prone to flooding based on real-time weather data, river levels, and historical patterns.	Supports early warning systems, assists emergency response planning, and aids in the identification of flood-prone zones for urban planning.

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

Surface and Groundwater Monitoring	3D visualizations or cross-sectional charts illustrating groundwater levels and trends over time.	Provides insights into groundwater recharge, helps in managing aquifer depletion, and aids in sustainable groundwater resource management.
Water Infrastructure Management	Geographic Information System (GIS) maps displaying the location and condition of water infrastructure, such as pipes, pumps, and treatment plants.	Supports maintenance planning, asset management, and infrastructure investment decisions.
Water Quality Index (WQI) Mapping	Maps showing the spatial distribution of water quality indices, incorporating multiple parameters.	Allows policymakers and the public to understand overall water quality trends and prioritize areas for intervention or improvement.
Remote Sensing for Drought Monitoring	Satellite imagery and remote sensing data used to visualize the extent of drought conditions in a region.	Facilitates early detection of drought events, helps in water resource planning, and supports agricultural decision-making.
Water Use Efficiency Analysis	Comparative charts and graphs showing water use efficiency metrics for different sectors, industries, or agricultural practices.	Supports sustainable water management practices, identifies areas for improvement, and informs policy decisions.
Water Availability Forecasting	Predictive models and charts showing anticipated water availability based on historical data, climate forecasts, and usage trends.	Assists in long-term water resource planning, infrastructure development, and climate change adaptation strategies.
Community Engagement and Education	Public-facing dashboards or infographics illustrating water usage, conservation tips, and the impact of individual behaviors on local water resources.	Raises awareness, encourages water conservation, and fosters a sense of community responsibility for water sustainability.

These examples demonstrate how big data visualization contributes to more effective and informed decision-making in water resources management, helping to address challenges related to water scarcity, pollution, and sustainable resource utilization in changing climate. Visualization also plays an important role in supporting hydrologic, hydraulic and environmental modeling efforts. Visual techniques have been widely applied to improve the presentation of modeling output for the dissemination of model-driven insights to researchers and decision makers (Xu, et al. 2022). Visual computing techniques can be applied to a number of water related research areas to facilitate Big Data–driven studies. These areas include but are not limited to hydrologic response, water quality, soil-water interaction, water hazard mitigation, and the water-food-energy nexus. A detailed discussion and use-case analysis of how visualization and visual analytics can benefit these research areas are presented in (Xu, et al. 2022).

Water data visualizations are provocative visuals and captivating stories that inform, inspire, and empower people to address most pressing water issues. USGS data science and visualization experts use visualizations to communicate water data in compelling and often interactive ways when static images or written narrative can't effectively communicate the interconnectivity and complexity of a water data issue (USGS 2024). Their visualization codes and procedures are available online for use at https://labs.waterdata.usgs.gov/index.html.

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

5.4.3 Learning resources of big data visualization

Readers are encouraged to learn more about big data visualization by using following learning resources.

- Online Courses: Coursera – “Data Visualization and Communication with Tableau”
- Books: "Big Data Visualization" by James D. Miller
- Online Platforms: Tableau Public and D3.js Documentation and Tutorials
- MOOCs (Massive Open Online Courses): Kaggle Courses – “Data Visualization”
- Community Forums: Stack Overflow - Data Visualization Tag

5.5 Chapter conclusion

To summarize, big data visualization is a disruptive force in water resource management, providing unparalleled insights into the intricacies of water-related data. As the challenges of water scarcity, quality degradation, and climate change intensify, the use of advanced visualization techniques becomes increasingly important for making sustainable and informed decisions. The use of big data visualization in water resource management has numerous advantages. Real-time water quality monitoring, efficient consumption pattern analysis, and the identification of possible flood zones are just a few instances of how visualization technologies provide stakeholders with fast, actionable information. Water managers acquire a holistic view of their resources by visualizing groundwater levels, infrastructure status, and drought patterns, allowing for more precise resource allocation and sustainable management techniques. Furthermore, big data visualization promotes collaboration among varied stakeholders by breaking down technical boundaries and enabling better communication across teams. These tools bridge the gap between experts and non-specialists by transforming complex data sets into visually intuitive representations, helping communities, politicians, and industries to actively participate in water conservation efforts. Through the lens of big data visualization, the ability to predict and respond to changes in water supply, comprehend the efficiency of water usage, and detect pollution sources becomes more feasible. These visualizations not only serve as effective risk mitigation and resource optimization tools, but they also improve transparency and accountability in water management processes. In the future, the continual evolution of big data visualization tools, together with ongoing advances in data analytics, will refine our understanding of water dynamics. As water resources become more limited and valuable, the role of visualization in management, engineering and research this complicated landscape becomes not just important but also vital. Finally, the convergence of big data and visualization points the way to a more sustainable and resilient water future.

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

6 Bibliography

Ali, Syed Mohd, Noopur Gupta, Gopal Krishna Nayak, und Rakesh Kumar Lenka (2016), „Big data visualization: Tools and challenges." *2nd International conference on contemporary computing and informatics (IC3I).* IEEE. 656-660.

Allen, George H, und Tamlin M. Pavelsky (2018), „Global extent of rivers and streams." *Science* 361 (6402): 585-588.

Åmand, L.; Carlsson, B. (2012) Optimal Aeration Control in a Nitrifying Activated Sludge Process. *Water Res.*, *46* (7), 2101–2110. https://doi.org/10.1016/j.watres.2012.01.023.

Åmand, L.; Olsson, G.; Carlsson, B. (2013), Aeration Control - A Review. *Water Sci. Technol.*, *67* (11), 2374–2398. https://doi.org/10.2166/wst.2013.139.

Åström, K. J.; Hägglund, T. (1995) *PID Controllers: Theory, Design and Tuning*; Research Triangle Park: NC, USA, https://doi.org/1556175167

Anthony, M. (2012). Probability Theory in Machine Learning. In: Seel, N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_658

Bagheri, M., Farshforoush, N. Bagheri, K., Shemirani, A. (2023) Applications of artificial intelligence technologies in water environments: From basic techniques to novel tiny machine learning systems, Process Safety and Environmental Protection, Volume 180, https://doi.org/10.1016/j.psep.2023.09.072.

Balamurugan, B., Nandhini, A. R., Seifedine, K. & Gandomi A. H. (2021). Big Data. Concepts, technology, and Architecture. https://doi.org/10.1002/9781119701859

„Big data – Definition " (https://medienportal.siemens-stiftung.org/de/big-data-definition-112185), © Siemens Stiftung 2019, lizenziert unter CC BY-SA 4.0 international (https://creativecommons.org/licenses/by-sa/4.0/legalcode.de)

Bixio, D.; Van Hauwermeiren, P.; Thoeye, C.; Ockier, P. (2001), Impact of Cold and Dilute Sewage on Pre-Fermentation - A Case Study. *Water Sci. Technol.*, *43* (11), 109–117

Chai, Q. (2008), Modeling, Estimation, and Control of Biological Wastewater Treatment Plants, Telemark University College, 2008. http://www.diva-portal.org/smash/get/diva2:124029/FULLTEXT01.p

Chalh, Ridouane, Zohra Bakkoury, Driss Ouazar, und Moulay Driss Hasnaoui (2015), „Big data open platform for water resources management." *International Conference on Cloud Technologies and Applications (CloudTech).* 1-8.

Chen, K. C.; Chen, C. Y.; Peng, J. W.; Houng, J. Y. (2002), Real-Time Control of an Immobilized-Cell Reactor for Wastewater Treatment Using ORP. *Water Res.*, *36*, 230–238. https://doi.org/10.1016/S0043-1354(01)00201-9

Chen, CL Philip, und Chun-Yang Zhang (2014), „Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information sciences* 275: 314-347.

Childs, Hank, Berk Geveci, Will Schroeder, Jeremy Meredith, Kenneth Moreland, Christopher Sewell, Torsten Kuhlen, und E. Wes Bethel (2013), „Research challenges for visualization software." *Computer* 46 (5): 34-42.

Claros, J.; Serralta, J.; Seco, a.; Ferrer, J.; Aguado, D. (2012), Real-Time Control Strategy for Nitrogen Removal via Nitrite in a SHARON Reactor Using PH and ORP Sensors. *Process Biochem. 47* (10), 1510–1515. https://doi.org/10.1016/j.procbio.2012.05.020

Corona, F.; Mulas, M.; Haimi, H.; Sundell, L.; Heinonen, M.; Vahala, R. (2013) Monitoring Nitrate Concentrations in the Denitrifying Post-Filtration Unit of a Municipal Wastewater Treatment Plant. *J. Process Control*, *23* (2), 158–170, https://doi.org/10.1016/j.jprocont.2012.09.011

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

Dabaghi, F. E, D. Ouazar, und P. Prastacos (2001), „Esimeau Integrated Information System for Modeling and Management of Water Resources: Concept and Architecture.“ *SYSTEMS ANALYSIS MODELLING SIMULATION* 41 (4): 669-688.

DVWK-ATV Arbeitsblatt A 198 (2003), Standardisation and derivation of design values for wastewater systems (in German)

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology, 38(1), 46-50.

ESG (2015), „Standards and Guidelines for Quality Assurance in the European Higher Education Area.“

Guerrero, J.; Guisasola, A.; Comas, J.; Rodríguez-Roda, I.; Baeza, J. A. (2012), Multi-Criteria Selection of Optimum WWTP Control Setpoints Based on Microbiology-Related Failures, Effluent Quality and Operating Costs. *Chem. Eng. J.*, *188*, 23–29. https://doi.org/10.1016/j.cej.2012.01.115.

Haimi, H.; Mulas, M.; Sahlstedt, K.; Vahala (2009), R. Advanced Operation and Control Methods of Municipal Wastewater Treatment Process in Finland. *Helsinki Univ. Technol. Water Wastewater Eng.*

Haimi, H.; Corona, F.; Mulas, M.; Sundell, L.; Heinonen, M.; Vahala, R. Shall (2015), We Use Hardware Sensor Measurements or Soft-Sensor Estimates? Case Study in a Full-Scale WWTP. *Environ. Model. Softw.*, *72*, 215–229. https://doi.org/10.1016/j.envsoft.2015.07.013

Hedegärd, M.; Wik, T. (2011), An Online Method for Estimation of Degradable Substrate and Biomass in an Aerated Activated Sludge Process. *Water Res.*, *45* (19), 6308–6320. https://doi.org/10.1016/j.watres.2011.09.003

https://brianasimba.github.io/MachineLearningblog//Introduction-post / [AccessedSeptember 10, 2021]

https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-artificial-intelligence

Hwang, J. H.; Oleszkiewicz, J. (2007), Effect of Cold-Temperature Shock on Nitrification. *Water Environ. Res.*, *79* (9), 964–968. https://doi.org/10.2175/106143007X176022.

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence?. Discover Artificial Intelligence, 2(1), 4. https://doi.org/10.1007/s44163-022-00022-8

Jason E Black, Jacqueline K Kueper, Tyler S Williamson (2023) , An introduction to machine learning for classification and prediction, Family Practice, Volume 40, Issue 1, February 2023, Pages 200–204, https://doi.org/10.1093/fampra/cmac104

Jönsson, H., Vinneras, B. (2003), Adapting the nutrient content of urine and faeces in different countries using FAO and Swedish dataManual on sewerage and sewage treatment systems Conference Paper, IWA International symposium on ecological sanitation

Khan, Muzammil, und Sarwar Shah Khan (2011) „Data and information visualization methods, and interactive mechanisms: A survey.“ *International Journal of Computer Applications* 34 (1): 1-14.

Kim, J.-H.; Chen, M.; Kishida, N.; Sudo, R. (2004) Integrated Real-Time Control Strategy for Nitrogen Removal in Swine Wastewater Treatment Using Sequencing Batch Reactors. *Water Res.*, *38* (14–15), 3340–3348. https://doi.org/10.1016/j.watres.2004.05.006.

Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Liu, Y.; Chen, J.; Sun, Z.; Li, Y.; Huang, D. A (2014), Probabilistic Self-Validating Soft-Sensor with Application to Wastewater Treatment. *Comput. Chem. Eng.*, *71*, 263–280. https://doi.org/10.1016/j.compchemeng.2014.08.00

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

Liu, W.; Ratnaweera, H. (2016), Improvement of Multi-Parameter-Based Feed-Forward Coagulant Dosing Control Systems with Feed-Back Functionalities. *Water Sci. Technol.*, *74* (2), 491–499. https://doi.org/10.2166/wst.2016.180

Machado, V. C.; Gabriel, D.; Lafuente, J.; Baeza, J. A. (2009), Cost and Effluent Quality Controllers Design Based on the Relative Gain Array for a Nutrient Removal WWTP. *Water Res.*, *43* (20), 5129–5141. https://doi.org/10.1016/j.watres.2009.08.011

Maciejowski, J. M. (2002) *Predictive Control with Constraints*; Technology & Engineering, Pearson Education

Manamperuma, L.; Wei, L.; Ratnaweera (2017), H. Multi-Parameter Based Coagulant Dosing Control. *Water Sci. Technol.*, *75* (9), 2157–2162. https://doi.org/org/10.2166/wst.2017.058

Marahatta, Suresh, Laxmi Prasad Devkota, und Deepak Aryal (2021) „Application of SWAT in hydrological simulation of complex Mountainous river basin (part I: model development)." *Water* 13 (11): 1546.

Martin, C.; Vanrolleghem, P. A. Analysing (2014), Completing, and Generating Influent Data for WWTP Modelling: A Critical Review. *Environ. Model. Softw.*, *60*, 188–201. https://doi.org/10.1016/j.envsoft.2014.05.008

Martín de la Vega, P. T.; Martínez de Salazar, E.; Jaramillo, M. a; Cros, J. (2012) New Contributions to the ORP & DO Time Profile Characterization to Improve Biological Nutrient Removal. *Bioresour. Technol.*, *114*, 160–167. https://doi.org/10.1016/j.biortech.2012.03.039

Martinez-Mosquera, D., Navarrete, R., & Lujan-Mora, S. (2020). Modeling and management big data in databases—A systematic literature review. Sustainability, 12(2), 634. https://doi.org/10.3390/su12020634

Miller, James D. (2017), *Big data visualization.* 1. Mumbai: Packt publishing.

MUD (2013), Ministry of Urban Development Manual on sewerage and sewage treatment systems Part A: Engineering, New Dehli, India,

Norwegian Water BA (2020), Guidelines for dimensioning of wastewater systems (in Norwegian). Report 256

Olsson, G.; Newell, B. *Wastewater (1999), Treatment Systems: Modelling, Diagnosis and Control*; IWA Publishing: London, UK

Olsson, G.; Carlsson, B.; Comas, J.; Copp, J.; Gernaey, K. V; Ingildsen, P.; Jeppsson, U.; Kim, C.; Rieger, L.; Rodríguez-Roda, I.; Steyer, J.-P.; Takács, I.; Vanrolleghem, P. a; Vargas, A.; Yuan, Z.; Åmand, L. (2014), Instrumentation, Control and Automation in Wastewater - from London 1973 to Narbonne 2013. *Water Sci. Technol.*, *69* (7), 1373–1385. https://doi.org/10.2166/wst.2014.05

Omondi Asimba, B. (2019). Machine Learning basics. Available at:

Plósz, B. G.; Liltved, H.; Ratnaweera, H. (2009), Climate Change Impacts on Activated Sludge Wastewater Treatment: A Case Study from Norway. *Water Sci. Technol.*, *60* (2), 533–541. https://doi.org/10.2166/wst.2009.386

Rapidrops, 2024. https://www.rapidops.com/blog/ai-ml-deep-learning-data-science-big-data/

Ratnaweera, H.; Fettig, J. (2015) State of the Art of Online Monitoring and Control of the Coagulation Process. *Water (Switzerland)*, *7* (11), 6574–6597. https://doi.org/10.3390/w7116574

Raval, Nirav, und Manish Kumar (2020), „An overview of big data analytics: A state-of-the-art platform for water resources management." *Resilience, Response, and Risk in Water Systems: Shifting Management and Natural Forcings Paradigms* . 43-56.

Riedl, M.O. (2019), Human-centered Artificial Intelligence and machine learning. Hum. Behav. Emerg. Technol. 2019, 1, 33–36.

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

Rieger, L.; Thomann, M.; Gujer, W.; Siegrist, H. (2005), Quantifying the Uncertainty of On-Line Sensors at WWTPs during Field Operation. *Water Res.*, *39* (20), 5162–5174. https://doi.org/10.1016/j.watres.2005.09.04

Rieger, L.; Olsson, G. (2012), Why Many Control Systems Fail. *J. Water Environ. Technol.*, *24* (6), 42–45. https://doi.org/10.2175/193864711802764779

Rieger, L.; Jones, R. M.; Dold, P. L.; Bott, C. B. (2014), Ammonia-Based Feedforward and Feedback Aeration Control in Activated Sludge Processes. *Water Environ. Res.*, *86* (1), 63–73. https://doi.org/10.2175/106143013X13596524516987

Ruano, M. V.; Ribes, J.; Seco, A.; Ferrer, J.(2009), Low Cost-Sensors as a Real Alternative to on-Line Nitrogen Analysers in Continuous Systems. *Water Sci. Technol.*, *60* (12), 3261–3268. https://doi.org/10.2166/wst.2009.607

Ruano, M. V.; Ribes, J.; Seco, A.; Ferrer, J. (2012) An Advanced Control Strategy for Biological Nutrient Removal in Continuous Systems Based on PH and ORP Sensors. *Chem. Eng. J.*, *183*, 212–221. https://doi.org/10.1016/j.cej.2011.12.064.

Samuelsson, O.; Björk, A.; Zambrano, J.; Carlsson, B. (2017), Gaussian Process Regression for Monitoring and Fault Detection of Wastewater Treatment Processes. *Water Sci. Technol.*, *75* (12), 2952–2963. https://doi.org/10.2166/wst.2017.162

Sarni, W.; White, C.; Webb, R.; Cross, K.; Glotzbach (2019), R. *Digital Water - Industry Leaders Chart the Transformation Journey*; Edinburgh, Scotland, UK, 2019. http://iwa-network.org/publications/digital-water/

Schnase, John L, Daniel Q. Duffy, Glenn S. Tamkin, Denis Nadeau, John H. Thompson, Cristina M. Grieg, Mark A. McInerney, und William P. Webster (2017), „MERRA analytic services: Meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service." Computers, Environment and Urban Systems 61: 198-211

Simon, Phil (2014), *The visual organization: Data visualization, big data, and the quest for better decisions.* Harvard: John Wiley & Sons.

Spiegel, M.R. (1972) Statistics, MacGraw-Hill Book Company

Stare, A.; Vrecko, D.; Hvala, N.; Strmcnik, S. (2007), Comparison of Control Strategies for Nitrogen Removal in an Activated Sludge Process in Terms of Operating Costs: A Simulation Study. *Water Res.*, *41* (9), 2004–2014. https://doi.org/10.1016/j.watres.2007.01.029

Steyer, J. P.; Bernard, O.; Batstone, D. J.; Angelidaki, I. (2006), Lessons Learnt from 15 Years of ICA in Anaerobic Digesters. *Water Sci. Technol.*, *53* (4–5), 25–33. https://doi.org/10.2166/wst.2006.107

Talagala, Priyanga Dilini, Rob J. Hyndman, Catherine Leigh, Kerrie Mengersen, und Kate Smith-Miles (2019), „A feature-based procedure for detecting technical outliers in water-quality data from in situ sensors." *Water Resources Research* 55 (11): 8547-8568.

Tchobanoglous, G.; Burton, F. L.; Stensel, H. D. (2003), *Wastewater Engineering: Treatment and Reuse*; Metcalf & Eddy, Inc., McGraw-Hill Education

Tong, Shuangmei, Hairong Li, Muyesaier Tudi, Xing Yuan, und Linsheng Yang (2021), „Comparison of characteristics, water quality and health risk assessment of trace elements in surface water and groundwater in China." *Ecotoxicology and Environmental Safety* 219: 112283

Verdhan, V. (2020), "Supervised learning with python." Apress, Springer, Berkeley, CA , https://doi.org/10.1007/978-1-4842-6156-9

Villez, K.; Srinivasan, B.; Rengaswamy, R.; Narasimhan, S.; Venkatasubramanian, V. Kalman-Based (2011), Strategies for Fault Detection and Identification (FDI): Extensions and Critical

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

Evaluation for a Buffer Tank System. *Comput. Chem. Eng.*, *35* (5), 806–816. https://doi.org/10.1016/j.compchemeng.2011.01.045

Vrečko, D.; Hvala, N.; Stražar, M. (2011) The Application of Model Predictive Control of Ammonia Nitrogen in an Activated Sludge Process. *Water Sci. Technol.*, *64* (5), 1115–1121.

Xiong, W.; Li, Y.; Zhao, Y.; Huang, B. (2017) Adaptive Soft Sensor Based on Time Difference Gaussian Process Regression with Local Time-Delay Reconstruction. *Chem. Eng. Res. Des.*, *117*, 670–680. https://doi.org/10.1016/j.cherd.2016.11.020.

Wang, Lidong, Guanghui Wang, und Cheryl Ann Alexander (2015), „Big Data and Visualization: Methods, Challenges and Technology Progress." *Digital Technologies* 1 (1): 33-38.

Wang, X.; Kvaal, K.; Ratnaweera, H. (2017), Characterization of Influent Wastewater with Periodic Variation and Snow Melting Effect in Cold Climate Area. *Comput. Chem. Eng.*, *106*, 202–211. https://doi.org/10.1016/j.compchemeng.2017.06.009

Wilen, B. M.; Lumley, D.; Mattsson, A.; Mino, T. (2006), Rain Events and Their Effect on Effluent Quality Studied at a Full Scale Activated Sludge Treatment Plant. *Water Sci. Technol.*, *54* (10), 201–208. https://doi.org/10.2166/wst.2006.721

Won, S. G.; Ra, C. S. (2011) Biological Nitrogen Removal with a Real-Time Control Strategy Using Moving Slope Changes of PH(MV)- and ORP-Time Profiles. *Water Res.*, *45* (1), 171–178. https://doi.org/10.1016/j.watres.2010.08.030

Xu, Haowen, Andy Berres, Yan Liu, Melissa R. Allen-Dumas, und Jibonananda Sanyal (2022), „An overview of visualization and visual analytics applications in water resources management." *Environmental Modelling & Software* 153: 10539

Zeng, W.; Peng, Y.; Wang, S.; Peng, C. (2008), Process Control of an Alternating Aerobic-Anoxic Sequencing Batch Reactor for Nitrogen Removal via Nitrite. *Chem. Eng. Technol. 31* (4), 582–587. https://doi.org/10.1002/ceat.200700468

Zhang, D., Martinez, N., Lindholm, G., Ratnaweera, H. (2018), Manage Sewer In-Line Storage Control Using Hydraulic Model and Recurrent Neural Network,
Water Resource Management 32:2079–2098, https://doi.org/10.1007/s11269-018-1919-3, 2

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