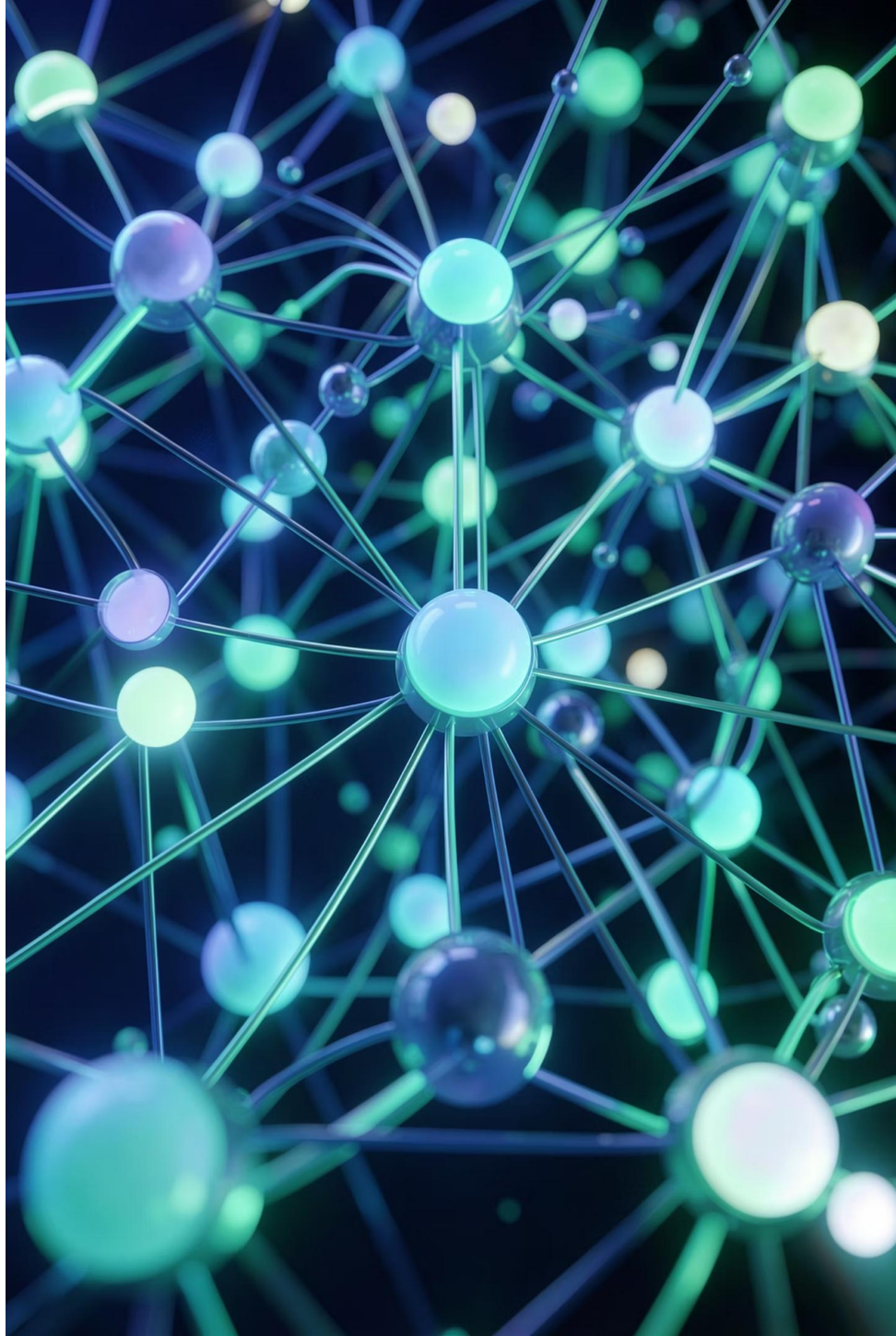
Data Mining: Insights from Data to Information and Knowledge

Radka Nacheva, PhD
University of Economics – Varna, Bulgaria

Erasmus+ Teaching Staff Mobility 2026
Technische Universität Chemnitz



What is Data Mining?

Data mining is the process of discovering patterns, correlations, anomalies, and meaningful insights from large datasets using statistical, mathematical, and machine learning techniques. It sits at the intersection of database systems, statistics, and artificial intelligence — transforming raw data into actionable knowledge.

Definition

Data mining is the practice of examining large pre-existing databases to generate new information. It extracts implicit, previously unknown, and potentially useful knowledge from data — going far beyond simple queries or reporting.

- It is part of a so-called process "**Knowledge Discovery in Databases (KDD)"**
- Involves both supervised and unsupervised learning methods
- Works across structured, semi-structured, and unstructured data

Why Data Mining Matters

Organizations generate staggering volumes of data every day. Without data mining, this information sits dormant. With it, businesses unlock competitive advantages, researchers accelerate discoveries, and governments make smarter policy decisions.

- **Business intelligence**: Predict customer churn, optimize pricing, detect fraud
- **Healthcare**: Identify disease risk factors and treatment outcomes
- **Science**: Accelerate genomic research and climate modeling
- **Finance**: Detect anomalies and manage portfolio risk

2.5Q

Data Created Daily

Quintillion bytes of new data are generated every single day worldwide

90%

Recent Data

Of all data in existence was created in just the last two years

$447B

Market Size

Projected global big data and analytics market value by 2026

3-5x

ROI Multiplier

Average return on investment from enterprise data mining initiatives

Statistics: https://giving.columbia.edu/bytes-data-created-daily-1; https://www.grandviewresearch.com/industry-analysis/big-data-industry ; https://www.globema.com/leverage-your-data-potential-8-data-mining-use-cases/

The Data Mining Process

Data mining is not a single action but a structured, iterative pipeline. The most widely adopted framework is **CRISP-DM** (Cross-Industry Standard Process for Data Mining), which provides a proven roadmap from raw data to deployed insight. Each phase builds on the last, and teams often cycle back to earlier stages as understanding deepens.

This iterative process ensures that the final model genuinely answers the right question with the right data. Successful data mining projects allocate roughly 60–70% of total effort to the first three phases alone, underscoring that quality inputs are the foundation of quality outputs.

1

Business Understanding

Define the project objectives, success criteria, and constraints from a business perspective. Translate business goals into a data mining problem definition.

2

Data Understanding

Collect initial data, explore its properties, identify quality issues, and discover first insights. Familiarity with the dataset shapes every downstream decision.

3

Data Preparation

Construct the final analytical dataset — handling missing values, removing noise, encoding variables, and engineering new features from raw inputs.

4

Modeling, Evaluation & Deployment

Select and tune algorithms, validate performance against business goals, then integrate the model into operational workflows for ongoing value delivery.

Data Preprocessing: Cleaning and Preparing Your Data

Data in the real world is rarely clean. It arrives with missing entries, inconsistent formats, duplicate records, and outliers that can derail any analysis. Data preprocessing is the critical foundation that converts raw, unstructured data into a reliable analytical asset.
Studies consistently show that **data scientists spend 60–80% of their time on preprocessing** — a testament to its importance and complexity.

Data Cleaning

- Handle missing values via deletion, mean/median imputation, or predictive filling
- Remove or correct noisy and erroneous data points
- Identify and resolve duplicate records
- Detect and treat outliers using IQR or Z-score methods

Data Integration

- Merge data from multiple heterogeneous sources
- Resolve entity matching and schema conflicts
- Detect and eliminate redundant attributes
- Ensure referential integrity across joined datasets

Data Transformation

- Normalize and standardize numerical features (Min-Max)
- Encode categorical variables (Label, Target encoding)
- Apply log or power transformations to skewed distributions
- Engineer new features from domain knowledge

Data Reduction

- Dimensionality reduction via feature selection
- Data compression and aggregation techniques
- Numerosity reduction through sampling or binning
- Discretization of continuous attributes into intervals

A useful rule of thumb: **Garbage In, Garbage Out (GIGO)**. Even the most sophisticated algorithm will produce unreliable results if fed poorly prepared data. Preprocessing is not overhead — it is the single most impactful investment in a data mining project.

Data Exploration and Visualization: Seeing the Patterns

Before applying any algorithm, skilled data miners explore their data visually and statistically. Exploratory Data Analysis (EDA) reveals the shape, structure, and quirks of a dataset — guiding smarter decisions about which models to apply and which features matter most. Visualization is the bridge between raw numbers and human intuition.

Exploratory Data Analysis (EDA)

EDA is a philosophy of analysis championed by statistician John Tukey. Rather than jumping to hypothesis testing, EDA encourages open-minded exploration using summary statistics and visual tools to let the data reveal its own story.

- **Univariate analysis**: Distributions, histograms, box plots for individual variables
- **Bivariate analysis**: Scatter plots, correlation matrices, cross-tabulations
- **Multivariate analysis**: Heatmaps, parallel coordinates, dimensionality plots
- **Time series plots**: Trend, seasonality, and anomaly detection over time
- **Summary statistics**: Mean, median, variance, kurtosis

Key Visualization Techniques

Different data structures demand different visual approaches. Choosing the right chart type dramatically accelerates pattern recognition.

- **Histograms & density plots**: Understand distribution shapes and skewness
- **Box plots**: Spot outliers and compare group distributions at a glance
- **Scatter plots**: Reveal linear and non-linear relationships between variables
- **Correlation heatmaps**: Identify multicollinearity and redundant features
- **Word clouds & treemaps**: Summarize textual and hierarchical data
- **Geospatial maps**: Uncover geographic patterns and regional clusters

Statistical Profiling

Generate automated data profiles that report completeness, uniqueness, data types, and value distributions for every column. Tools like pandas-profiling, Great Expectations, and Tableau Prep accelerate this step significantly.

Cluster Visualization

Use t-SNE, or PCA projections to reduce high-dimensional data to 2D or 3D, revealing natural groupings and separations invisible in raw tabular form.

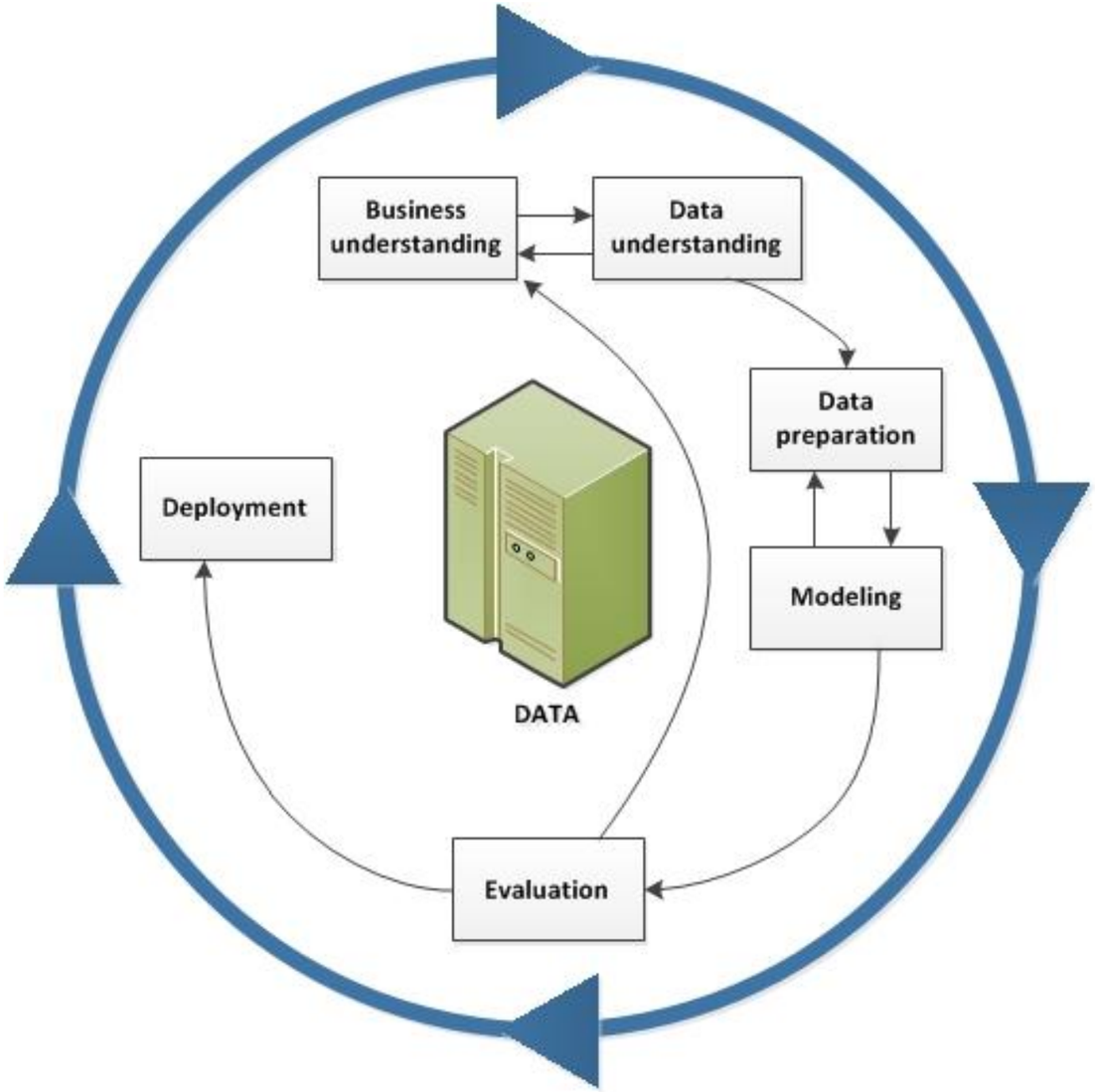
Interactive Dashboards

Tools like Tableau, Power BI, and Plotly Dash enable dynamic exploration — allowing analysts to filter, drill down, and cross-filter visualizations to test hypotheses in real time.

The CRISP-DM Journey: A Cyclical Approach to Data Mining

CRISP-DM is not a rigid, one-directional process. It is a **cyclical, iterative framework** comprising six interconnected phases — each feeding into the next, often requiring revisits to earlier stages as new insights emerge. Understanding this structure is fundamental to applying it effectively.

The outer cycle represents the ongoing nature of data mining projects — once deployed, insights feed back into new business questions, restarting the process. This continuous loop ensures that models remain relevant, accurate, and aligned with evolving business needs over time.



PHASE 1

Business Understanding

The foundation of any successful data mining project lies in a thorough understanding of the business context. This phase is critical — misaligned objectives at the outset can render even the most technically sophisticated models useless. The primary goal is to translate business needs into a clearly defined data mining problem, supported by a realistic project plan.

Key Tasks

- **Determine business objectives:** Clarify what the organisation truly wants to achieve, including primary and secondary goals.
- **Assess the current situation:** Inventory available resources, constraints, risks, and contingencies.
- **Determine data mining goals:** Translate business objectives into technical criteria for success.
- **Produce a project plan:** Define the phases, timelines, resources, and milestones required to deliver results.

Tools & Techniques

Business Dictionaries

Standardise terminology across teams to ensure shared understanding of key concepts and KPIs.

Stakeholder Interviews

Structured conversations with decision-makers to surface expectations, constraints, and success criteria.

SWOT Analysis

Assess strengths, weaknesses, opportunities, and threats relevant to the data mining initiative.

Cost-Benefit Analysis

Quantify the expected return on investment to justify project scope and resource allocation.

PHASE 2

Data Understanding

With business objectives clearly defined, attention turns to the data itself. This phase involves collecting initial datasets and performing exploratory analysis to understand their structure, quality, and potential. Early data exploration often uncovers hidden patterns, anomalies, and hypotheses that shape the entire project's direction.

Key Tasks

01

Collect Initial Data

Identify and acquire relevant datasets from internal systems, third-party sources, or public repositories. Document provenance and access methods.

02

Describe Data

Examine the volume, format, number of records, and field types. Produce a data description report covering schema, ranges, and key statistics.

03

Explore Data

Conduct exploratory data analysis (EDA) using visualisations and statistical summaries. Identify correlations, distributions, outliers, and interesting subsets.

04

Verify Data Quality

Assess completeness, consistency, and accuracy. Identify missing values, duplicates, and inconsistencies that may affect modelling.

Tools & Technologies

Python

Pandas for data manipulation, Matplotlib & Seaborn for rich visualisations and statistical plots.

SQL

Query relational databases to extract, filter, and summarise large datasets efficiently.

R

Powerful statistical computing environment with ggplot2 and dplyr for exploration and summarisation.

Data Profiling Tools

Automated tools such as Great Expectations or Pandas Profiling generate comprehensive data quality reports.

Excel

Accessible tool for quick pivot tables, summary statistics, and initial data visualisations.

PHASE 3

Data Preparation

Often the most time-consuming phase of the entire CRISP-DM process, data preparation accounts for up to **70–80% of total project effort** in real-world projects. This phase covers all activities needed to construct the final, clean dataset that will be fed into modelling tools. Tasks are frequently performed multiple times and in varying order as understanding deepens.

1

Select Data

Choose which attributes and records are relevant. Justify inclusions and exclusions based on data quality and relevance to goals.

2

Clean Data

Handle missing values via imputation or removal. Correct inconsistencies, fix formatting errors, and remove duplicates.

3

Construct Data

Derive new attributes through feature engineering. Create aggregations, ratios, or transformations that improve model performance.

4

Integrate & Format

Merge datasets from multiple sources. Reformat data types and structures to match the requirements of the chosen modelling tools.

Python & R Libraries

- **Pandas:** Data wrangling, merging, and reshaping
- **Scikit-learn:** Preprocessing pipelines, scaling, encoding
- **R (dplyr, tidyr):** Tidy data transformations

ETL & Database Tools

- **Talend / Informatica:** Enterprise ETL pipelines for large-scale data integration
- **SQL:** In-database transformations and joins
- **Excel:** Manual cleaning for smaller datasets

PHASE 4

Modelling

The modelling phase is where the scientific creativity of data mining truly comes to life. Multiple techniques are explored and compared, as different algorithms may be suited to the same problem. The choice of technique must always be guided by the business objectives established in Phase 1, and models must be rigorously validated before progressing to evaluation.

Key Tasks

- **Select modelling technique:** Choose from classification, regression, clustering, association rules, or deep learning, based on the data mining goal.
- **Generate test design:** Define train/test splits, cross-validation strategies, and performance benchmarks before building models.
- **Build model:** Apply the selected techniques to the prepared dataset. Tune hyperparameters iteratively.
- **Assess model:** Evaluate model quality using technical metrics (accuracy, AUC, RMSE) and review outputs against expectations.

It is common to cycle back to Data Preparation at this stage if model performance reveals data quality issues or missing features.

Tools & Frameworks

Scikit-learn

Python's go-to library for classical ML — regression, classification, clustering, and preprocessing pipelines.

TensorFlow & Keras

Deep learning frameworks for neural networks, image recognition, NLP, and complex pattern detection.

PyTorch

Flexible deep learning framework favoured in research and increasingly in production environments.

R (caret / tidymodels)

Comprehensive model training and evaluation packages with support for many algorithms.

SPSS Modeler & SAS EM

Enterprise-grade drag-and-drop modelling environments suited to business analysts without deep coding skills.

PHASE 5

Evaluation

Technical model performance is insufficient for a successful data mining project. The evaluation phase steps back from the numbers to ask a fundamental question: **does this model truly address the original business problem?** This phase ensures that no critical business considerations have been overlooked before committing to deployment.

Key Tasks in Detail

1 Evaluate Results

Assess model outputs against the business success criteria defined in Phase 1. Determine whether the model's predictions are actionable, interpretable, and commercially meaningful — not just statistically valid.

2 Review Process

Conduct a quality review of all phases completed to date. Identify any shortcuts taken, assumptions made, or data issues that could undermine the model's reliability in production.

3 Determine Next Steps

Decide whether to proceed to deployment, return to an earlier phase for refinement, or initiate an entirely new data mining iteration. Document the rationale clearly.

Evaluation Tools

Business Metrics Dashboards

Visualise model impact on real KPIs such as revenue, churn rate, or operational efficiency using tools like Tableau or Power BI.

A/B Testing Frameworks

Run controlled experiments to validate model recommendations in a live environment before full-scale rollout.

Stakeholder Feedback Sessions

Present findings to business stakeholders to validate interpretations and surface domain-specific concerns that technical metrics may miss.

PHASE 6

Deployment

A model that sits in a notebook delivers no business value. The deployment phase ensures that the knowledge, models, and insights generated throughout the project are operationalised and made accessible to decision-makers and systems. The form of deployment varies widely — from an automated scoring engine to a polished executive dashboard — and depends entirely on the business context and end-user needs.

Plan Deployment

Define the technical architecture for serving the model — whether as a REST API, batch scoring job, embedded application, or integrated BI report. Document all dependencies, access controls, and infrastructure requirements.

Monitoring & Maintenance

Establish processes to monitor model performance over time, detect data drift, retrain on new data, and manage version control. Models degrade without ongoing maintenance — this plan ensures longevity.

Produce Final Report

Deliver a comprehensive summary of the project — covering objectives, methodology, findings, model performance, limitations, and recommendations — tailored to both technical and non-technical audiences.

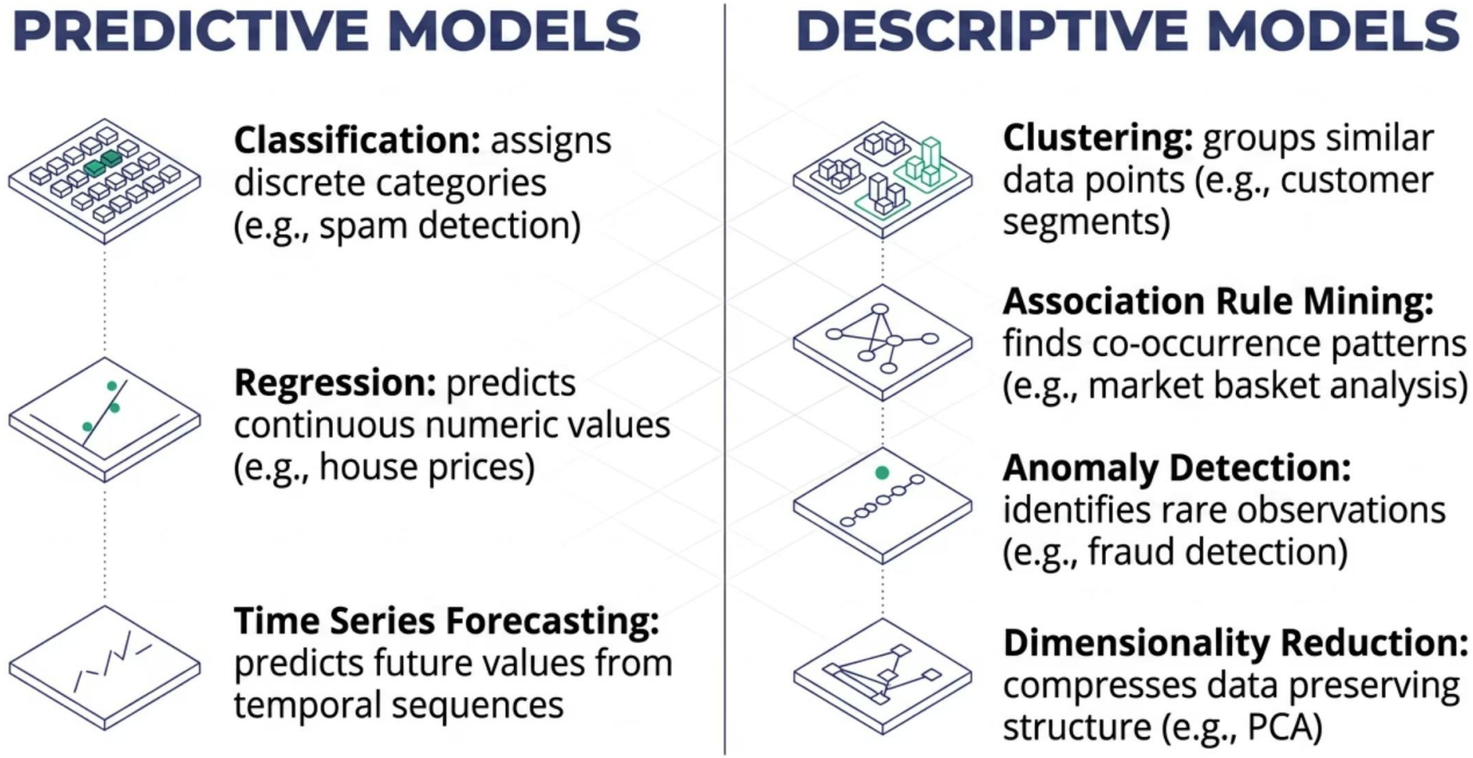
Review Project

Conduct a retrospective to capture lessons learnt, document what worked and what didn't, and identify improvements for future CRISP-DM iterations within the organisation.

Deployment Tools: Tableau, Power BI (dashboards & reporting) · REST APIs via Flask or FastAPI · Production databases (PostgreSQL, Snowflake) · MLflow or MLOps platforms for model lifecycle management

Key Data Mining Models: Predictive and Descriptive

Data mining models fall into two fundamental paradigms that serve different analytical goals. **Predictive models** learn from labeled historical data to forecast future outcomes. **Descriptive models** summarize and discover structure within unlabeled data. Understanding which type to apply is the first critical decision in any data mining project.



Predictive Modeling

Predictive models are trained on a labeled dataset where the target variable (the answer) is known. The model learns the mapping between input features and the target, then generalizes to unseen data. Performance is measured by accuracy, precision, recall, F1-score (classification). Overfitting — where the model memorizes training data but fails on new data — is the primary risk, mitigated through cross-validation and regularization.

Descriptive Modeling

Descriptive models operate on unlabeled data, seeking to summarize its inherent structure without a predefined target. Because there is no "correct answer" to optimize toward, evaluation relies on domain expertise and internal metrics like silhouette scores (clustering) or lift and support (association rules). These models are powerful for hypothesis generation, customer segmentation, and exploratory knowledge discovery.

Common Data Mining Algorithms: A Toolkit for Discovery

No single algorithm solves all problems. Expert data miners maintain a diverse toolkit, selecting techniques based on data type, problem structure, interpretability requirements, and computational constraints. Below are the most widely used and influential algorithms across both predictive and descriptive paradigms.

Decision Trees & Random Forests

Decision Trees partition data by recursively splitting on the most informative feature, producing interpretable if-then rule chains. **Random Forests** aggregate hundreds of trees trained on random data subsets (bagging), dramatically reducing variance and improving accuracy. Excellent for tabular data with mixed feature types. Handles missing values gracefully.

Support Vector Machines (SVM)

SVMs find the optimal hyperplane that maximally separates classes in high-dimensional space. The **kernel trick** (RBF, polynomial, sigmoid kernels) maps data to higher dimensions where linear separation becomes possible. Highly effective for text classification, image recognition, and small-to-medium datasets with many features.

K-Means & Hierarchical Clustering

K-Means iteratively assigns each data point to the nearest centroid and updates centroids until convergence — fast and scalable but requires specifying K in advance. **Hierarchical clustering** builds a dendrogram of nested clusters without pre-specifying K, enabling exploration of cluster structure at multiple granularities. Widely used in customer segmentation and genomics.

Neural Networks & Deep Learning

Multi-layer neural networks learn hierarchical representations of data through stacked non-linear transformations. **Convolutional networks (CNNs)** excel at images; **Recurrent networks (RNNs/LSTMs)** handle sequences; **Transformers** power modern NLP. Unmatched accuracy on complex unstructured data but require large datasets and significant compute.

Naive Bayes

Probabilistic classifier using Bayes' theorem with feature independence assumption. Fast, simple, excellent for text classification and spam filtering.

K-Nearest Neighbors

Classifies by majority vote among the K closest training examples. Intuitive, non-parametric, but slow on large datasets — requires careful distance metric selection.

Apriori / FP-Growth

Association rule algorithms that mine frequent itemsets and generate if-then rules. Core to market basket analysis and recommendation systems.

Gradient Boosting (XGBoost)

Sequentially builds trees where each corrects errors of the previous. Dominates structured data competitions — the most widely deployed algorithm in production ML systems today.

Real-World Applications of Data Mining

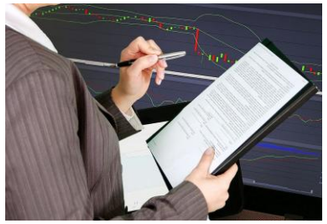
Data mining is not an academic exercise — it powers decisions in virtually every sector of the modern economy. From the ads you see online to the drugs prescribed by your doctor, data mining is working behind the scenes to optimize outcomes at scale.

Retail & E-Commerce

Amazon's recommendation engine — responsible for **35% of its total revenue** — uses collaborative filtering and association rule mining to suggest products. Retailers also apply clustering for customer segmentation, enabling personalized promotions that dramatically improve conversion rates.

Healthcare & Medicine

Predictive models identify patients at high risk for readmission, sepsis, or disease progression. Mining electronic health records (EHR) has accelerated drug discovery, reduced diagnostic errors, and enabled precision medicine — tailoring treatments to individual genetic profiles.



Finance & Fraud Detection

Credit card companies process millions of transactions per second, using real-time anomaly detection models to flag fraudulent activity with sub-100ms latency. Credit scoring, algorithmic trading, and risk modeling all rely heavily on classification and regression techniques.

Social Media & Marketing

Sentiment analysis mines millions of posts to gauge public opinion in real time. Churn prediction models identify at-risk subscribers before they cancel. Lookalike modeling finds new customers who resemble your best existing ones, optimizing digital ad spend with surgical precision.

Challenges and Ethical Considerations in Data Mining

As data mining becomes more pervasive and consequential, the field faces serious technical, organizational, and ethical challenges. Responsible practitioners must navigate these issues proactively — not as afterthoughts, but as core design requirements built into every project from day one.

Technical Challenges

→ Data Quality & Completeness

Real-world datasets are rarely complete or accurate. Missing data, measurement errors, and inconsistent formats can silently corrupt model outputs without obvious warning signs.

→ Scalability & Velocity

Traditional algorithms struggle with petabyte-scale datasets and real-time streaming data. Distributed computing frameworks (Spark, Hadoop) and online learning algorithms are essential at enterprise scale.

→ The Curse of Dimensionality

As feature count grows, data becomes increasingly sparse in high-dimensional space. Distances lose meaning, models overfit, and training time explodes — requiring dimensionality reduction and careful feature selection.

Ethical & Legal Considerations

→ Privacy & Data Sovereignty

Mining personal data without informed consent violates trust and regulations like GDPR, CCPA, and HIPAA. Techniques like differential privacy and federated learning enable insight extraction without exposing individual records.

→ Algorithmic Bias & Fairness

Models trained on biased historical data perpetuate and amplify existing inequalities. Bias audits, fairness-aware learning, and diverse training data are essential safeguards in high-stakes domains like hiring, lending, and criminal justice.

→ Transparency & Explainability

Black-box models (deep neural networks) make decisions that cannot be easily explained to affected individuals. Explainability frameworks like SHAP and LIME are increasingly required by regulators and demanded by end users.

Responsible AI Imperative: The power of data mining comes with proportional responsibility. Organizations deploying data mining systems must implement governance frameworks, regular audits, and clear accountability structures to ensure their models remain fair, accurate, and aligned with human values over time.

Useful Tools

- **RapidMiner**: A data analysis platform with a visual interface, suitable for beginners and advanced users. https://rapidminer.com
- Weka: An academically oriented tool with a rich set of machine learning and data mining algorithms. https://www.cs.waikato.ac.nz/ml/weka/
- **KNIME**: A powerful platform with a workflow-based approach to data analysis, transformation and modeling. https://www.knime.com
- **Orange**: An intuitive tool with visual programming, suitable for training and experimentation. https://orangedatamining.com
- **scikit-learn**: A Python machine learning library with a wide range of algorithms for classification, regression and clustering. https://scikit-learn.org
- **TensorFlow**: An open source library for building machine learning and deep learning models. https://www.tensorflow.org
- **Apache Spark**: A big data processing platform with machine learning support via MLlib. https://spark.apache.org
- **SAS Enterprise Miner**: A commercial data analysis tool widely used in business and finance. https://www.sas.com
- **IBM SPSS Modeler**: A tool with a visual interface for building models and analyzing data. https://www.ibm.com/products/spss-modeler

Key Takeaways

Data mining is a multi-disciplinary field that transforms raw data into strategic knowledge. Whether you are just beginning your journey or deepening your expertise, these core principles will guide your practice and keep your work grounded in rigor and responsibility.

Definition

Data mining is the systematic process of discovering patterns, correlations, and anomalies in large datasets using statistical and machine learning methods — transforming data into actionable knowledge.

Process

Follow CRISP-DM.
Expect to iterate, and invest heavily in preprocessing — it determines model quality.

Models

Choose **predictive models** (classification, regression) when you have labeled targets to forecast.
Choose **descriptive models** (clustering, association rules) when you want to discover hidden structure in unlabeled data.

Algorithms

Master a diverse toolkit: Decision Trees, Random Forests, SVMs, Neural Networks, K-Means, and Gradient Boosting (XGBoost). Match the algorithm to data type, scale, and interpretability requirements.

Ethics

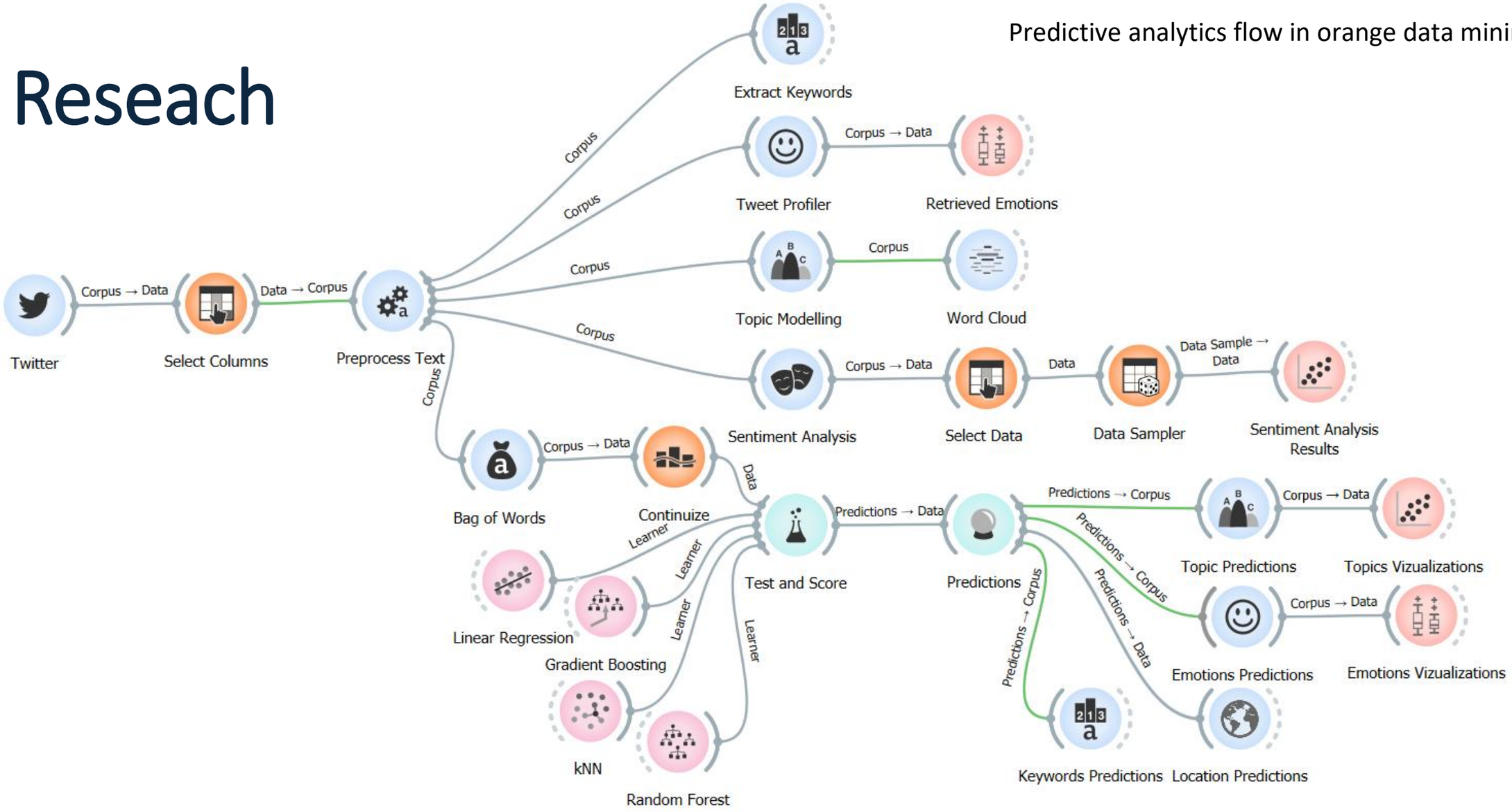
Always consider privacy, bias, and explainability. Build fairness audits and governance frameworks into your workflow — responsible data mining is not optional, it is a professional obligation.

> *"The goal is to turn data into information, and information into insight."* — **Carly Fiorina**, former CEO of Hewlett-Packard

Next Steps: Explore hands-on platforms like **Kaggle** for practice datasets, **scikit-learn** for Python implementations, and **Weka** for GUI-based experimentation. The best way to master data mining is to mine real data — start with a question you care about and let curiosity drive the discovery.

Reseach

Predictive analytics flow in orange data mining tool

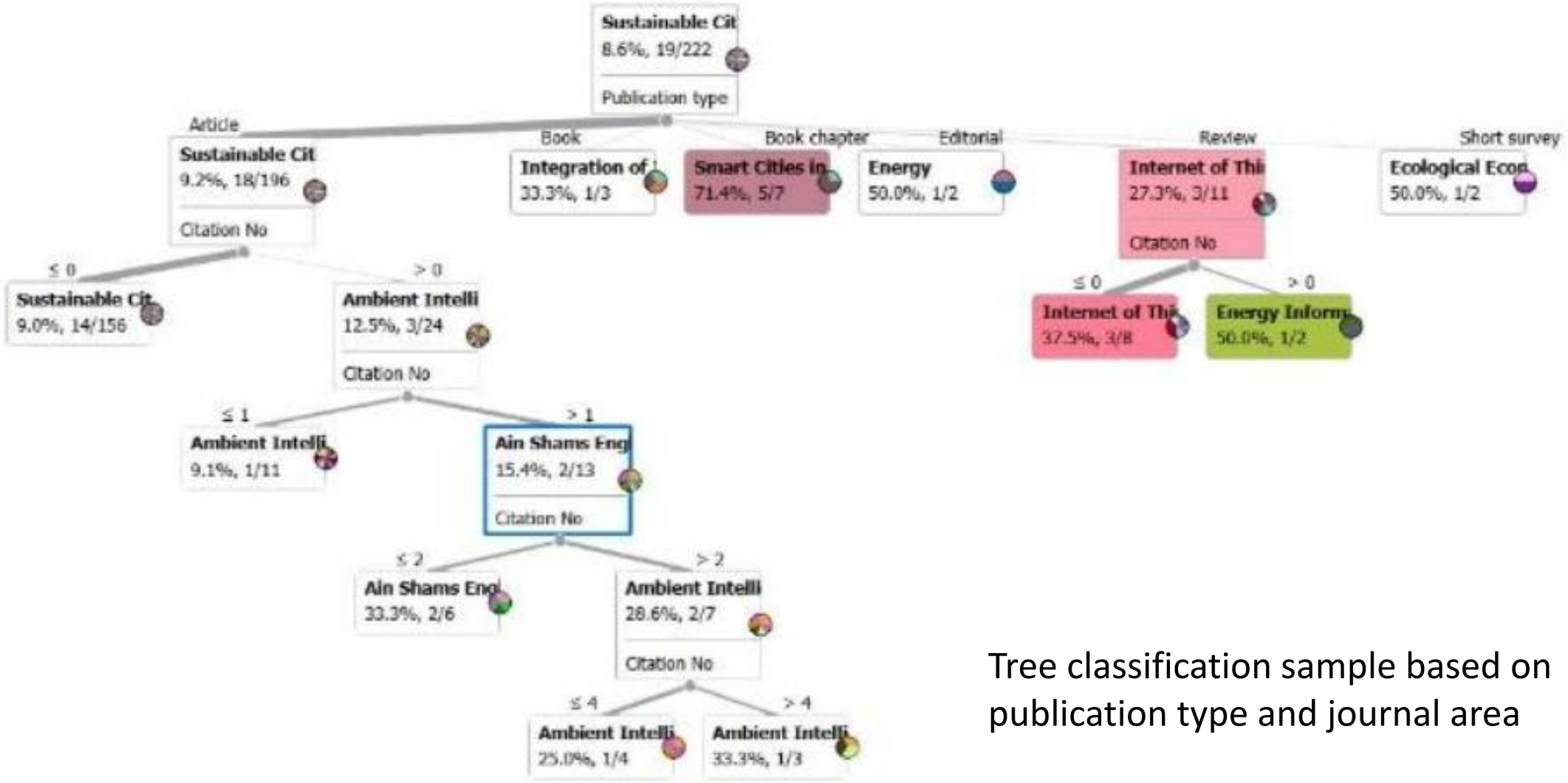


Azeroual, O., Nacheva, R., Nikiforova, A., Störl, U. A CRISP-DM and Predictive Analytics Framework for Enhanced Decision-Making in Research Information Management Systems. Informatica : An International Journal of Computing and Informatics, Ljubljana : Slovenian Society Informatika, 49, 2025, 18, 67-86.

Reseach

- The process of retrieving scientific papers utilized the Scopus abstract and citation database, marking the first stage of a proposed model for the Collection of Research Information. The data was organized as a structured CSV file.
- The subsequent stages involved data mining techniques, including pattern identification, classification, association, clustering, and data cleaning and visualization.
- In the second stage, the dataset underwent pre-processing through transformation, tokenization, normalization, and filtering, removing accents, converting text to lowercase, and retaining punctuation. WordNet Lemmatizer facilitated normalization, while English stopwords, numbers, and special characters were filtered out.
- Feature extraction utilized n-grams, ranging from 1 to 5, with the Averaged Perceptron Tagger employed for part-of-speech tagging.
- Keyword and topic extraction methods aided in developing an ontology and categorizing papers.
- A specific word list containing key phrases such as “smart cities” and “intelligent cities” was created for document scoring.
- Analysis of the corpus revealed that smart cities and urban intelligence frequently featured in the headlines. Topic extraction was conducted using Latent Semantic Indexing, yielding ten topics with respective weights.

Reseach



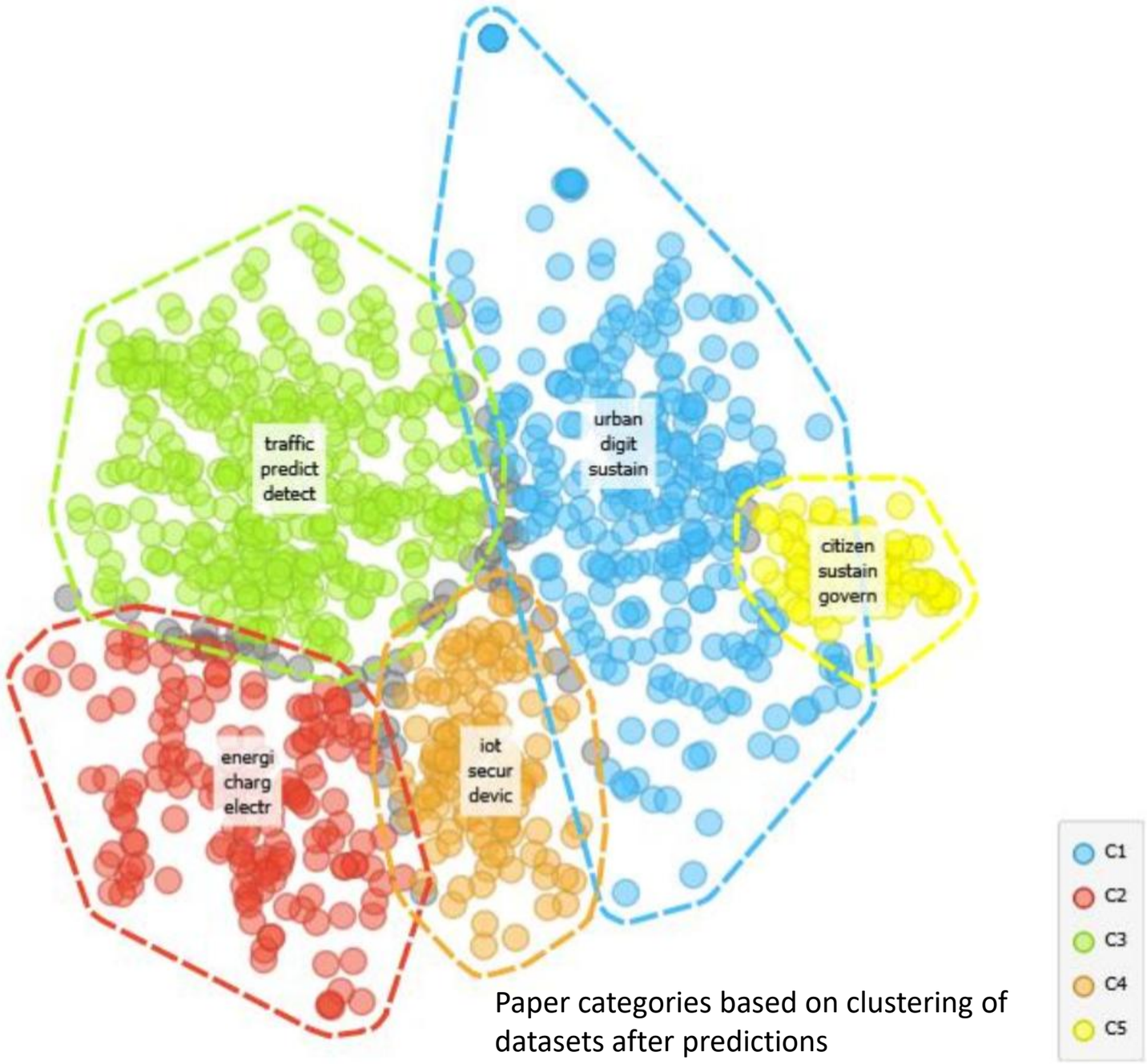
Tree classification sample based on publication type and journal area

Reseach

IoT
- data
- including
  - city
    - cities
      - sustainable
      - paper
      - urban
    - study
    - work
      - systems
        - system
        - human
          - Things
          - due
            - smart
            - development
            - applications
              - based
              - technologies
                - information
                - research
            - problems
      - Internet
      - proposed
      - health
    - intelligence
  - life
- model

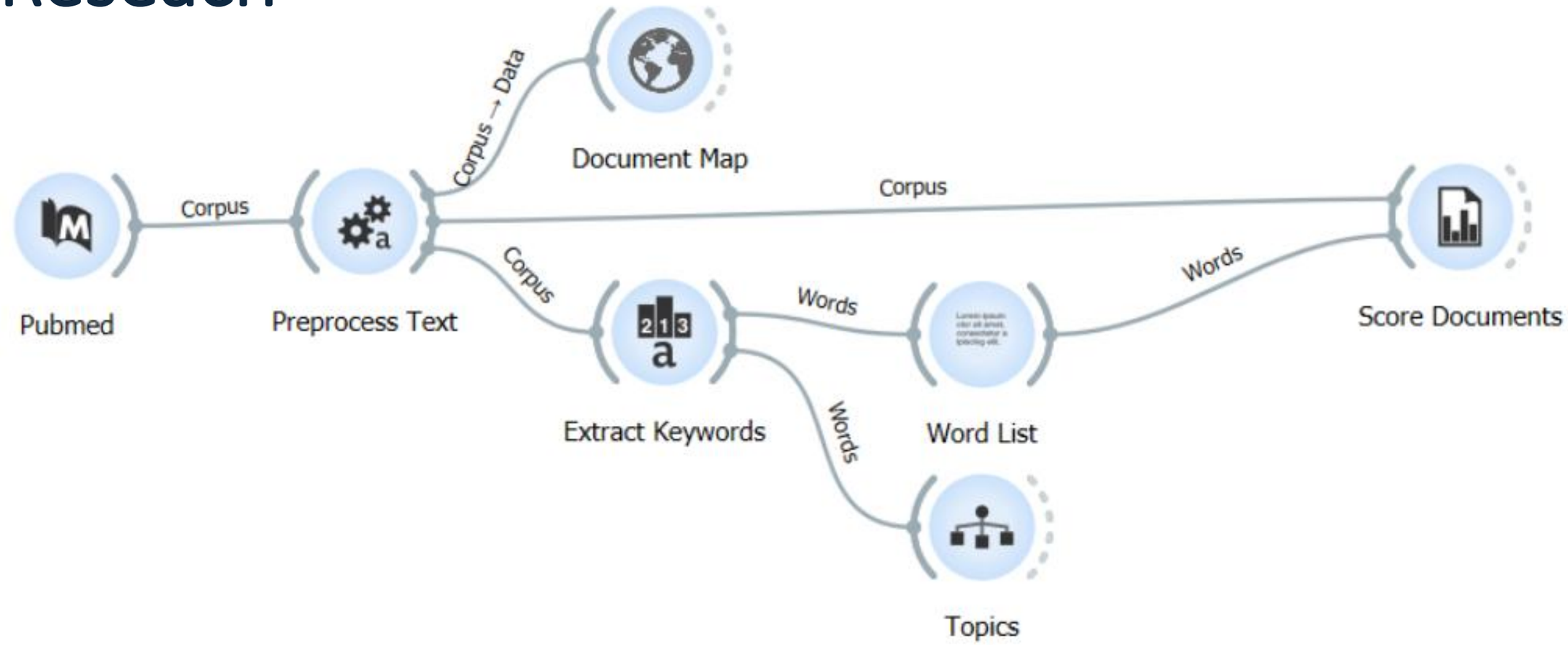
network
- role
- cities
  - city
  - learning
    - data
      - research
      - environmental
        - sustainable
        - proposed
          - future
          - model
          - artificial
            - systems
            - challenges
            - study
    - development
      - management
      - system
        - smart
        - applications
    - intelligence
    - increasing
  - economic
  - vehicles
- intelligent
- urban
  - energy
  - paper

Generated ontology



Paper categories based on clustering of datasets after predictions

Reseach



Nacheva, R. Trends And Best Practices For Ensuring Digital Accessibility in the Workplace. Journal of Process Management and New Technologies, Belgrade : Faculty of Applied Management, Economics and Finance, 13, 2025, 1-2, 56-66.

Reseach

Figure . Extracted Research Papers Topics

Reseach

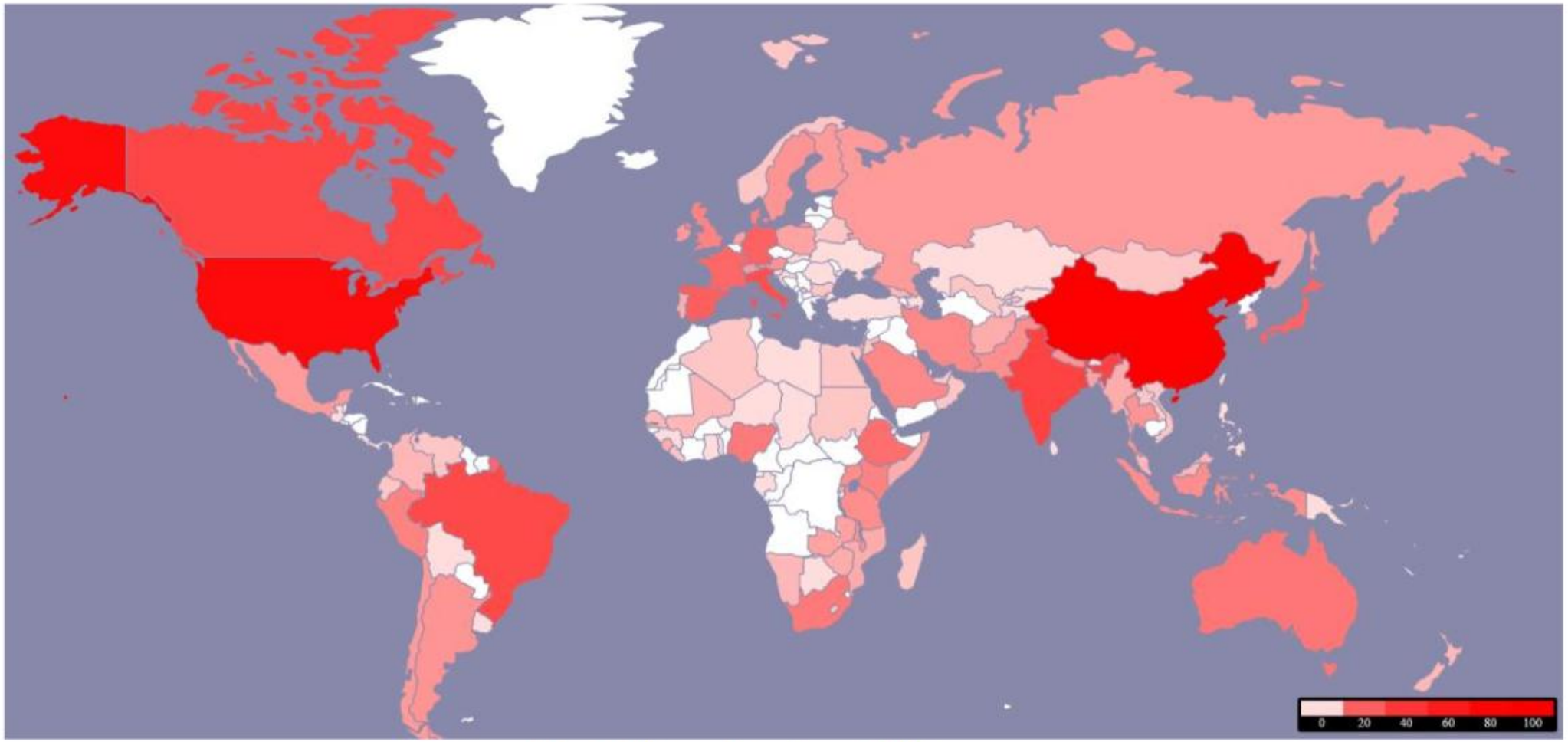


Figure . Geolocations Map

Reseach

Figure . Orange Data Mining Workflow

Nacheva, R. Artificial Intelligence for Collaboration in a Digitally Accessible Environment: a Bibliometric Analysis. Izvestia Journal of the Union of Scientists - Varna. Economic Sciences Series, Varna : Union of Scientists - Varna, 14, 2025, 1, 240-247.

Reseach

- The imported XLSX file was converted into an English corpus using Orange Data Mining, involving text normalization (lowercasing, accent removal), removal of HTML tags and URLs, and tokenization through regular expressions.
- Lemmatization was applied with the Lemmagen Lemmatizer to enhance text quality, and text features were filtered by removing English stop words and limiting the dataset to terms occurring 1-10 times and the top 100 tokens.
- The SBERT model was utilized for document embedding to capture semantic similarity, followed by t-SNE for nonlinear dimensionality reduction, employing PCA for preliminary reduction and Euclidean distance for similarity measurement.
- Finally, TF-IDF was applied to classify key phrases for ontology creation, resulting in the extraction of 42 top phrases with a TF-IDF value of 1.000.

Reseach

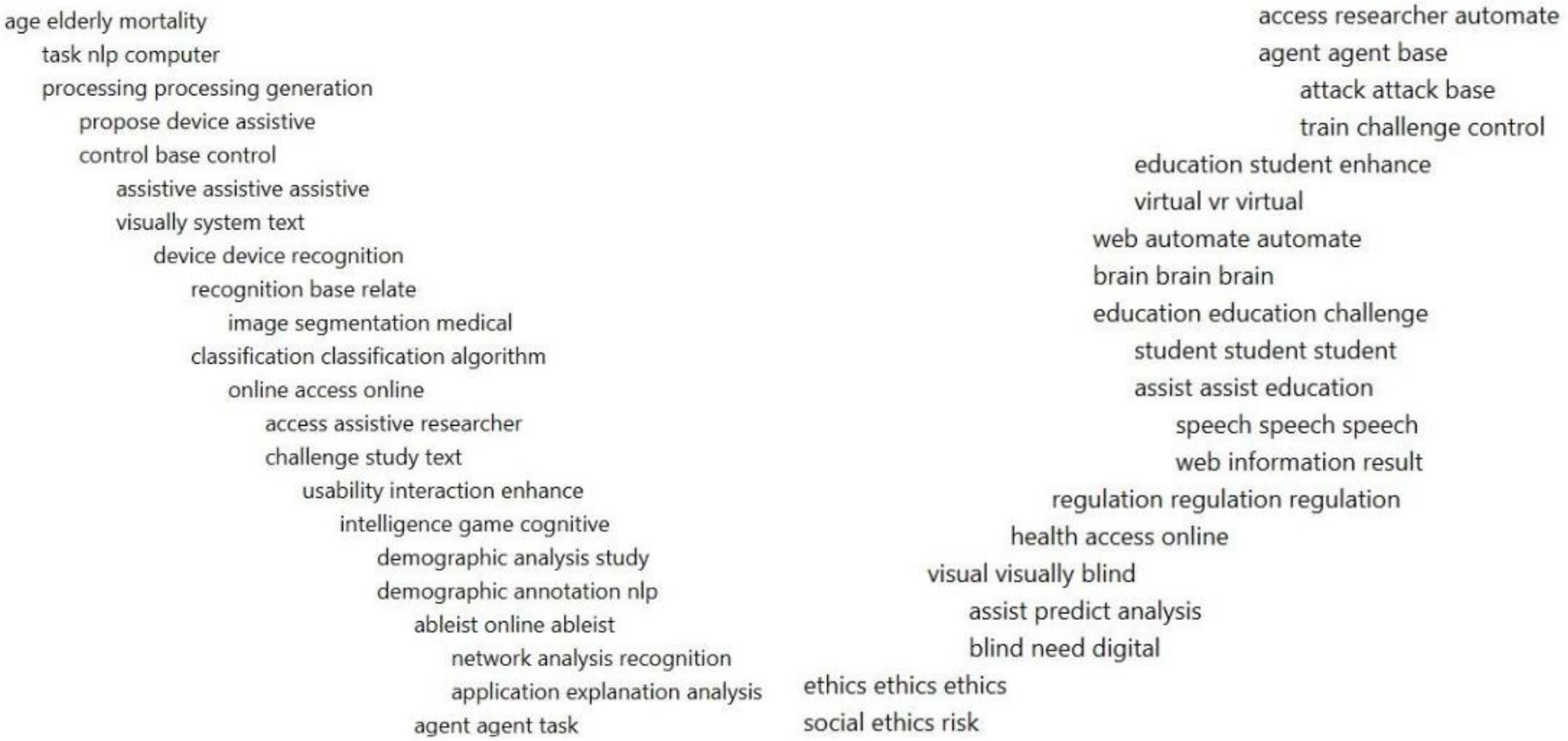


Figure . Orange Data Mining Ontology

Vielen Dank für Ihre Aufmerksamkeit!
Благодаря за вниманието!