

# Version management

|  |  |  |  |
| --- | --- | --- | --- |
| **Version Number** | **Date** | **Author** | **Change** |
| 0.1 | 17-09 | Jason | Started on working on the Current situation. |
| 0.2 | 18-09 | Jason + Baha | Change of the documentation + Lay-out of the document. |
| 0.2 | 18-09 | Jason + Baha | Changed the SIPOC + Making the BPMN Model. |
| 0.3 | 18-09 | Zoë | Changed front page, put business goals in table. |
| 0.4 | 22-09 | Zoë | Added intros and small changes to balanced scorecard. |
| 0.4 | 22-09 | Fatima | Onderbouwing + verbetering SWOT-analyse, Vijf porter's krachtenmodel. |
| 0.4 | 22-09 | yara | Visie&scope  Stakholders analyse  Ist-soll-Gap. |
| 0.5 | 24-09 | Jason | Looking and changing what is needed for a better outcome. |
| 0.6 | 25-0 | Jason | Making the document in English. |
| 0.7 | 25-09 | Zoë | Translated my parts to English. |
| 0.8 | 26-09 | Zoë | Added target levels to balanced scorecard that the client provided. |
| 0.9 | 25-9-2025 | Fatima | Translated the parts to English. |
| 1.0 | 26-9 | Baha | Translated Sipoc to English |
| 1.1 | 30-09-2025 | Jason | Starting the Project plan. |
| 1.2 | 01-10-2025 | Jason | Started on Stakeholder analysis + Worked on new BPMN Model for the expected end result. |
| 1.3 | 02-10-2025 | Jason | Added new BPMN Model +. |
| 1.3 | 02-10-2025 | Baha | Azure database uitgelegd. |
| 1.3 | 2-10-2025 | Fatima | Scrapen + deadlines and planning opgesteld. |
| 1.4 | 03-10-2025 | Baha + Jason | Making new document to make things les chaotic. Was needed after feedback. |
| 1.5 | 06-10-2025 | Baha + Jason | Making the new Risk analysis and added and cost and benefits. |
| 1.6 | 07-10-2025 | Jason | Started working on new BPMN Model after review with Hend.  Also made a MoSCoW Table for business succes criteria.  Added SWOT Analysis.  Started on Project plan  Added Stakeholder analysis + Added Project plan Timeline and Project Overview. |
| 1.6 | 07-10-2025 | Zoë+ yara | Inventory of recources and a part of requierements, assumpsionts and constraints. |
| 1.6 | 07-10-2025 | Fatima | Initial assessment of tools and techniques, data mining succes criteria. |
| 1.6 | 07-10-2025 | Baha | Keep document in order and made a part of project plan. |
| 1.6 | 07-10-2025 | Yara | Terminology and a part of requirements, assumptions and constraints. |
| 1.6 | 10-10-2025 | Fatima | Data mining success criteria verbeterd. |
| 1.7 | 16-10-2025 | Zoë | Introductie uitgebreid, requirements in tabellen gezet |
| 1.7 | 16-10-2025 | Baha | Sipoc geupdate |
| 1.8 | 19-10-2025 | Fatima | Business succes criteria + data mining goals verbeterd |
| 1.9 | 20-10-2025 | Zoë | Balanced score card aangepast, missie, visie & strategie aangepast |
| 2.0 | 23-10-2025 | Jason | Added feedback. Added more information about the MoSCoW table.  + Added gant chart in Timeline and phases |
| 2.1 | 24-10-2025 | Fatima | Business succes criteria aangepast. |
| 2.2 | 26-10-2025 | Baha | Risk-analyse improved |
| 2.3 | 27-10-2025 | Jason | Made Gant-chart |
| 2.3 | 27-10-2025 | Baha | Made new current situation and roles |
| 2.3 | 27-10-2025 | Jason | Made adjustment based on feedback of Matthijs van Berkhout |

# 

# Introduction

Hogescholen, such as the Hogeschool Utrecht (HU), face the challenge of ensuring that their curricula align with the constant changing and evolving requirements of the labor market. Students often develop skills that do not fully match the competencies demanded by employers, resulting in a gap between education and professional work. This mismatch can limit graduates’ employability, and reduce the effectiveness of educational programs.

The goal of this project is to develop a data-driven system, that enables the Marketing & Customer Experience Research Group at HU to analyze curricula and identify skill gaps. By matching curricula with labor market data from the top 100 companies, the system provides insights that could support curriculum adjustments, ensuring that students continue to learn relevant skills, and that HU continues to deliver competent professionals to the Dutch labor market.

The relevance of this project, lies in its potential to improve educational outcomes and labor market alignment. It benefits HU by enabling data-informed curriculum development and helps educators design (or adjust already existing) programs, that meet current and future professional demands. At the same time, it indirectly supports students by enhancing the skills they acquire during their studies.

This document is structured as follows: first, the business objectives are determined, including the background, vision and scope, business objectives, and business success criteria. Next, the situation is assessed, covering the inventory of resources, requirements, assumptions and constraints, risk and contingencies, and cost and benefits. Then, the data mining goals are defined along with their corresponding success criteria. Finally, the project plan is presented, including an initial assessment of tools and techniques.

# Table of contents

[Version management 2](#_Toc237420293)

[Introduction 6](#_Toc1290126028)

[Table of contents 7](#_Toc510997305)

[Determine business objectives 8](#_Toc1580331819)

[Background 8](#_Toc1730022384)

[Business objectives 16](#_Toc697422291)

[Business success criteria (Smart geformuleerd) 20](#_Toc990220860)

[Assess situation 23](#_Toc1615936543)

[Inventory of resources 23](#_Toc362144817)

[Requirements, assumptions and constraints 24](#_Toc370086248)

[Risk and contingencies 28](#_Toc1348366226)

[Terminology 30](#_Toc346605643)

[Determine data mining goals 31](#_Toc2067604230)

[Data mining goals 31](#_Toc193203547)

[Data mining success criteria 33](#_Toc1815317370)

[Produce project-plan 34](#_Toc147779574)

[Project-plan (Gant chart) 34](#_Toc1551353851)

[Initial assessment of tools and techniques 46](#_Toc1494699884)

[Sources 50](#_Toc175623698)

# Determine business objectives

## Background

**Mission**

The Marketing & Customer Experience research group focuses on strengthening valuable relationships between people and brands in an increasingly digital world. Our mission is to ensure that marketing and technology are used in a meaningful, sustainable and human-centered way, with respect for both the customer and society.

**Vision**

We believe in a future where marketing acts as a positive force for societal change. By combining technology and customer experience responsibly, organizations can achieve sustainable growth while also contributing to the well-being of people, brands and society.

**Strategy**

We pursue our mission and vision through applied research, education, and close collaboration with industry partners.

The research group develops tools, methods, and interventions that help professionals to:

* Strengthen customer relationships through data and technology,
* Apply sustainable and ethical marketing practices,
* And shape digital transformation in a human-centered way.

# Scope

In the scope, we show what we will do within this project and what we will focus on (in scope). This helps prevent any misunderstandings during the project. We also describe what we will not include (out of scope), so it is clear which parts fall outside this project. In the table below, we explain for each task what will be delivered (deliverables) and when the result will be approved (acceptance criteria).

|  |  |  |
| --- | --- | --- |
| In scope | Deliverables | Acceptance criteria |
| Scraping job vacancies from the top 100 Dutch companies. | Automated scraping algorithm | Will be accepted when at least 100 job postings are collected per scraping run, with ≥80% succes rate. |
| Automatically extracting job titles, soft skills, hard skills, sector, and location. | Data extraction algorithm | Will be accepted when, in a sample of 100 job postings, at least 85% of the relevant fields are extracted correctly. |
| Developing a matching algorithm using Natural Language Processing (NLP) to link CV skills with job requirements. | NLP-based matching algorithm | Will be accepted when, for a test set of 4 job postings, at least 3 relevant matches are confirmed. |
| Visualizing the matching results and skill gaps in an interactive dashboard. | Interactive dashboard with skill gap visualization | At least 2 different diagram's to show the results of skill gaps |
| Implementing an automated pipeline (scraping → extraction → matching → visualization). | Integrated pipeline (prototype) | Accepted when the full pipeline runs automatically in ≥95% of test runs and delivers complete output |

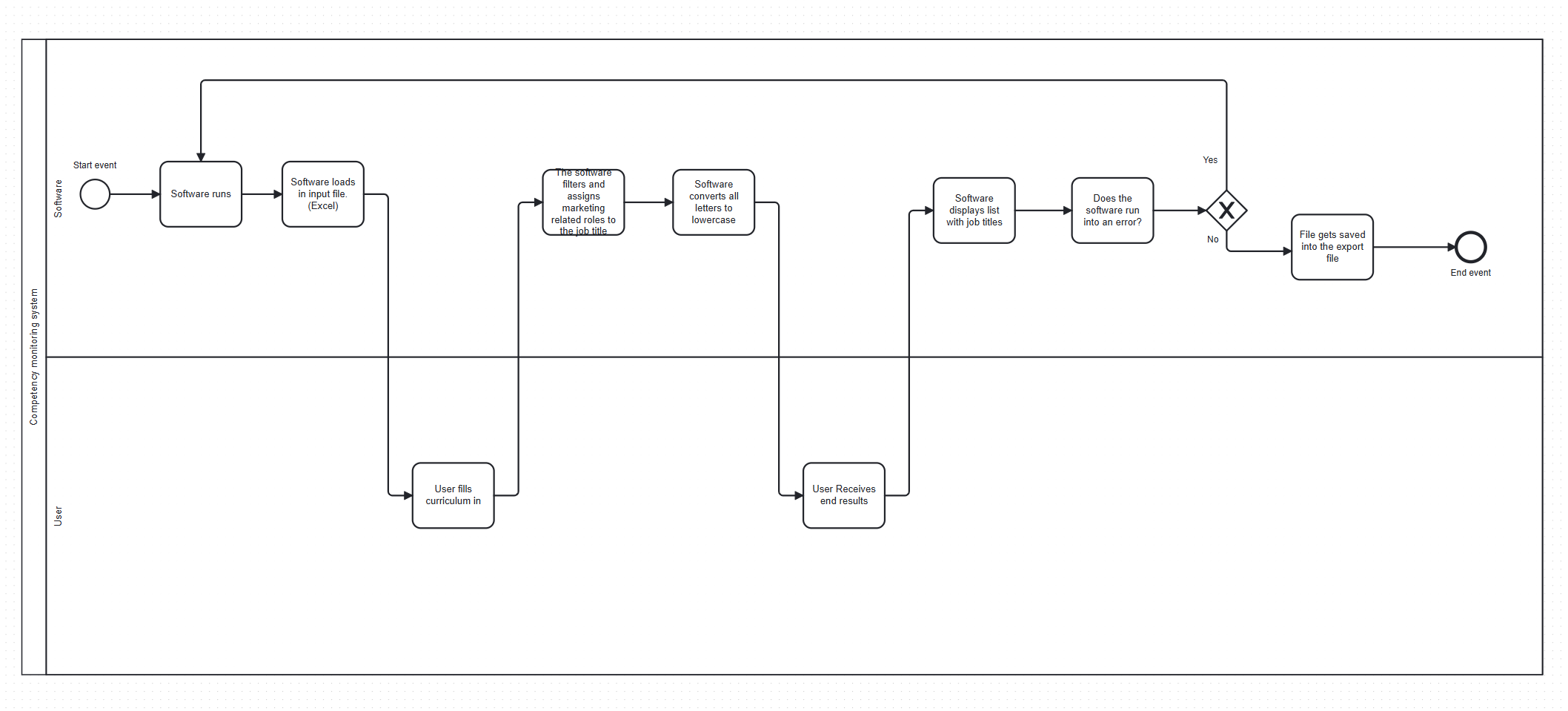
|  |
| --- |
| Out of scope |
| Real-time or continuous scraping (only periodic runs will be performed). |
| AI-based predictions or trend analysis of future labor market developments. |
| Integration with internal HU systems such as OSIRIS or LMS platforms. |
| Analysis of international or non-Dutch job vacancies that are not related to marketing |

**SIPOC:**

To gain a proper understanding of the current situation, we use a SIPOC analysis to create insight into the existing processes, systems and stakeholders involved. In addition, by applying a BPMN Model, we can clearly visualize how to current process operates. This provides us with a solid understanding of the present workflow and highlights where improvements can be made in the new website.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SIPOC | | | | |
| **Suppliers** | **Inputs** | **Processes** | **Outputs** | **Customers** |
| Indeed  Hogeschool Utrecht students (curriculum)  Microsoft (Azure)  HU teachers  Dennis Hagen  Hend Elsayed | HU curiccula (Website)  Python coding for scraping  Database in azure  Vacancies indeed | Manually retrieve data from current vacancies indeed  Loading the data into code  Code runs software and searches for keywords to connect vacancies  Generating response  Answer comes back through the code  Answers gets displayed on the dashboard | Overview of skill gaps per study program  Interactive dashboard (Power BI)  Automated dataset containing job and competency data  Reports and insights for curriculum improvement | Utrecht University of Applied Sciences (lectors and institutes)  Lecturers and curriculum developers  Students (indirect users)  Researchers and policy makers  Other universities of applied sciences (potential future users) |

With the help of a SIPOC analysis, we mapped out the process and made it as transparent as possible. This made it clear which parties are involved, what data enters the process, how the steps are carried out, and which results are ultimately delivered to the end user.

BPMN (Current situation)

By modelling the process in a BPMN diagram, we now have a clear view of how the process functions. This provides a better understanding of the current situation and forms a foundation for optimization.

The process begins with starting the software. Once initialized, the input file is uploaded. Next, the curriculum is entered, after which the software assigns marketing-related roles to the position titles. The text is then automatically converted to lowercase. The software subsequently generates statistics and displays a list of position titles. A check is performed to determine whether the software produces an error message. If an error occurs, the process restarts. If no error is detected, the file is saved as an output file and the process is completed. Now we have an output file with the jobs that match the curriculum

**SWOT-analysis**

To understand the organization and system better, we use a SWOT analysis to look at the system's strengths, weakness, opportunities, and threats. This helps to make the system strong and reliable for the future.



Hogeschool Utrecht has a good base with the competency monitoring system to connect education and jobs. But the system is still old data, takes a lot of work and depends on good job data. There are big chances to improve it with new data, AI and working with companies. Fast changes in the job market, competition and technical problems are big threats.

# Business objectives

**Business Goals**

To provide direction to the project and clarify where we want to go with it, we have defined business goals. These goals explain why we are doing this project and what we aim to achieve. They form the foundation of the project and were established based on discussions with the client.

|  |  |  |
| --- | --- | --- |
| **Business goal** | **Description** | **Project goal** |
| **Building a bridge between education and the labour market** | Contribute to reducing the competency gap and help educational programs train student with the right knowledge and skills. | The system should support programs in evaluating and adjusting their curricula, thereby building a bridge between education and the labor market. |
| **Improving and updating curricula** | Evaluate and adapt educational programs so that graduates’ competencies better align with the needs and requirements of the labour market. | Develop a matching algorithm that links competencies from job vacancies to the HU curriculum. |
| **Increasing labor market relevance** | Reduce the gap between education and the professional field by having real-time insights into required competencies. | Deliver a user-friendly, interactive dashboard that provides programs and researchers with real-time insights into competencies and trends. |
| **Strengthening research and education with up-to-date insights** | Provide teachers, researchers and policymakers with reliable and current data to support decision-making. | Use recent and representative data (2025) from the top 100 Dutch companies. |
| **Increasing efficiency and sustainability** | Automate the process of data collection and analysis so it can be repeated annually and remain up to date without high additional costs or manual effort. | Within 6 months, realize a fully automated pipeline that collects and processes job vacancies, and set up a system that can run at least once a year and be easily updated. |

**Balanced score card**

After defining the business goals, a Balanced Scorecard was created to make these goals more concrete and measurable. The business goals describe *what* we want to achieve, like connecting education and the labor market, while the Balanced Scorecard focuses on *how* we can track and evaluate that.

Each goal is linked to specific KPI’s, target levels, and actions. This helps us view and monitor the project from different perspectives: financial, customer, internal processes & growth. This ensures that no aspect of the project gets overlooked or forgotten. This card was discussed with and reviewed by our client, who also helped us determine the target levels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Perspective | Goals | KPI’s | Target Level | Initiatives / Actions |
| **Financial** | Cost-efficient system that can be deployed sustainably | - Annual operational costs in €  - Percentage time saved compared to manual processing | - €500  - 80% | Automate data collection and processing |
| **Customer (HU professors, training & education, managers, researchers)** | Satisfied stakeholders and usability of insights | - Percentage satisfaction score of programs/researchers  - Number of programs that have been adjusted | - ≥80%  - ≥2-10 | Develop a user-friendly, interactive dashboard with real-time competence insights |
| **Internal Processes** | Efficient and reliable collection/processing of data from up-to-date sources | - Percentage of automated tasks | - ≥95% | Build a scraping pipeline for Indeed including top 100 Dutch companies |
| **Learning & Growth** | Scalability to other disciplines | - Number of new disciplines successfully integrated | - 2 or more | Prepare processes and system to handle additional disciplines |

**MoSCoW Table**

Afbeelding met tekst, schermopname, Lettertype, ontwerp

Door AI gegenereerde inhoud is mogelijk onjuist.

By using a MoSCoW table we can decide as team what is essential versus optional. Instead of trying to do everything at once, we can prioritize the core features that are necessary for the system to succeed. It also keeps the project realistic and on schedule.

**Must have:** These are non-negotiable features that are essential for the system to function. Without these features, the project would fail

**Should have:** Important features that add significant value but are not critical for the system’s initial success. They enhance usability and efficiency.

**Could have:** These are features that are nice-to-have. They improve user experience and flexibility but van be implemented later if time is limited

**Wont have:** These are features that are currently out of scope. These are either too complex, unnecessary or planned for future versions. They help prevent scope creep.

By using the MoSCoW method, our team can clearly prioritize what is critical, What adds value, and what can be postponed or excluded. This structured approach ensures that resources are focused on what truly matters for the success of the project

## Business success criteria

Business success criteria help a project stay focused on what really matters. They show what needs to be achieved for the project to be called successful and make it easier to measure progress and value. These criteria are also important to improve the desired process of the project.

**Success criteria:**

* Within 2 months the processing system is fully automated, so the coordinator needs to do at least 80% less manual data entry compared to the start.
* Around 1 month there is a web scraping system that collects at least 100 job listings within one week. When checked, at least 80 of them have correct and complete information.
* Around 1,5 months, there is a matching algorithm that connects at least 4 job listing to students. When checked by the coordinator, at least 3 of these matched are found to be relevant.

**BPMN Model**

To ensure that the process remains clear, structured, and up to date, we created a BPMN model of the desired situation. This model provides a visual representation of the target process, making it easier to identify improvements.

Afbeelding met tekst, diagram, lijn, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

As you can see in this BPMN model, many aspects have been adjusted, especially the steps required to reach the final result. With these changes, we aim to improve the process and ensure that everything runs more smoothly without pitfall

# Assess situation

## Inventory of resources

To carry out the project successfully, it is important to have a clear overview of all available resources. This includes not only the people involved, but also the data sources, the technical tools, and the software needed to build the system. The following table provides an overview of these resources and their specific role within the project.

|  |  |  |
| --- | --- | --- |
| **Category** | **Resource** | **Description** |
| Personnel | Projectteam (Students) | Responsible for the overall project execution, including research, development of the website, data collection and dashboard design |
|  | Business expert (HU lectorare/client) | Provides insight into the project and validates the project goals |
|  | Project supervisor | Provides guidance and feedback throughout the project. |
| Data | Indeed job postings (Top 100 Dutch companies) | Used to identify and analyze competencies and skills that are required in the Dutch labour market. |
|  | HU curriculum and study programs | Provides information on current competencies taught within the school's programs |
| Computing Resources | Azure cloud environment | Hosts the web application, stores datasets and runs the data processing pipeline |
|  | Student workstations/laptops | Used by project team for testing, training and development |
|  | Database | Stores scraped job data and curriculum information securely |
| Software | Python | For data scraping, cleaning, processing |
|  | Power BI | For visualization and reporting the results in an interactive dashboard |
|  | HTML, CSS, JavaScript and PHP | For developing the website’s interface and ensuring accessibility |
|  | Azure/GitHub | For version control, collaboration and hosting the backend system |
|  | Microsoft Office | For documentation, reporting, and communication within the team |

## Requirements, assumptions and constraints

**Functional Requirements**

|  |  |
| --- | --- |
| Requirement | Acceptance criteria |
| The system must enable users (lecturers/teachers) to input a curriculum from Utrecht University of Applied Sciences via a link | 85% of test runs must successfully complete the curriculum upload process. |
| The system must automatically collect job vacancy data from Indeed, focusing on the top 100 dutch jobs. | Each scraping run must collect at least 100 job vacancies, with a minimum of 90% meeting the defined criteria. |
| The system must automatically extract the job title, hard and soft skills, sector, and location for each collected vacancy | In a sample of 100 job vacancies, at least 85% of the stored records must be complete. |
| The system must automatically compare the skills from the curriculum with the required competencies in the job vacancy and identify differences (skill gaps). | When tested with a curriculum, 90% of the matching and missing skills must be displayed correctly. |
| The system must display an interactive dashboard showing the matching results, including skill gaps per study program and the underlying job vacancy and curriculum data. | In 82% of test runs, users must be able to by study program and skill type (hard or soft skill) without system errors. |

**Non-Functional Requirements**

|  |  |
| --- | --- |
| Non-functional requirement | Acceptation criteria |
| |  | | --- | |  | | The system must be user-friendly for non-technical users. | | At least 86% of test users must be able to use the system independently within 5 minutes, without assistance. |
| The scraping, matching, and publishing processes must be largely automated. | 90% of the entire process must be executed automatically, with a maximum of one manual step allowed per full run. |
| The dashboard must respond quickly, enabling users to efficiently obtain insights. | 85% of test users must be able to use the dashboard normally within approximately 2 seconds. |
| |  | | --- | |  | | The system must comply with GDPR regulations and the security standards of Utrecht University of Applied Sciences. | | In 84% of security tests, all data must be encrypted during storage and transmission, and no personal data from job vacancies may be stored. |
| The system must be able to be executed annually with new data without requiring modifications. | During an annual update, the pipeline must be successfully executed in at least 85% of cases. |

**Assumptions**

During the project so far, several assumptions were made that form the foundation of the project’s planning and execution. These assumptions describe the conditions that are expected to be true but cannot yet be guaranteed. By defining them, we can ensure that the project remains realistic and that potential risks can be identified early on.

|  |  |
| --- | --- |
| Assumption | **Description** |
| Data availability | We assume that Indeed and Indeed will continue to have publicly available job postings that can be legally scraped |
| Data quality | We assume that job postings contain enough structured information to extract useful data |
| Stakeholder participation | We assume that lecturers and institutes will cooperate by providing curricula to test and validate the system |
| Tool stability | We assume that Python libraries used for scraping and Azure services will remain available and functional during the project |
| Planning | We assume that the project can be completed within one semester |
| Teamwork | We assume that all team members remain available for the duration of the project and are capable of completing their assigned tasks |

**Constraints**

Within the project there are several constraints that affect the scope, timeline and technical possibilities. These limitations are mainly related to the available time, resources and tools provided.

By identifying these limitations, we ensure that the business goals remain achievable, and the project stays manageable throughout the entire process.

|  |  |
| --- | --- |
| Constraint | **Description** |
| Project timeline | The system must be developed, tested, and presented within one semester. This limits the time available for testing and optimization. |
| Resource availability | This project is carried out by students, which means available time, expertise, and computing resources are limited. Development and testing are therefore restricted to what can be achieved within the team's skill set. |
| Technical limitations | The project depends on the available Azure environment provided by HU. Database storage, computing capacity, and access rights are limited to the resources accessible within this environment. |
| Data access | The system relies on public job postings from Indeed. If access is restricted, data scraping may be partially or temporarily unavailable. |
| Legal and ethical boundaries | All data collection and processing must comply with GDPR regulations and the ToS of the platforms used (Indeed). |
| Scope limitation | The CMS focuses solely on Dutch job postings from the top 100 companies and the HU marketing programs. International vacancies and other disciplines are currently outside the project scope |

## Risk and contingencies

A risk analysis is needed to find and manage problems with privacy, technology, and using the competency monitoring system early. This helps take action and make sure the project succeeds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Risk | Chance | Impact | Weight | Consequence | Measure |
| **Receiving data we did not want to scrape** | 2/5 | 2/5 | Low | To much data received and its not clear | Filtering list with selected data |
| **Indeed not scrapeable** | 1/5 | 5/5 | High | No data available for project | Scrape part by part |
| **Indeed partly scrapeable** | 1/5 | 5/5 | High | Not enough data for project | Scrape part by part |
| **Sick or personal situation of staff** | 3/5 | 3/5 | Medium | Part of the sick person can’t be worked | Report on time/ Other team member picks it up |
| **Indeed under maintanance/Downtime** | 2/5 | 5/5 | Medium | Scraping not available at that moment | Waiting till its up |
| **Website response error** | 3/5 | 4/5 | Medium | System crashes or slow response time | Conduct stress testing: optimized back end performance and surface scaling |
| **HU site has downtime** | 2/5 | 4/5 | Medium | HU course data temporarily unavailable | Retry connection |
| **Format courses are wrong on the HU website** | 3/5 | 3/5 | Medium | Wrong or incomplete curriculum in system | Validate and clean course data automatically |
| **Over budget with Azure** | 2/5 | 5/5 | Medium | Unexpectedly high hosting or processing caused | Monitor resource usage: set budget alerts, optimized data processing |
| **Leak in data privacy** | 1/5 | 5/5 | High | Violation of GDPR | Encrypt all data |
| **Deadline is not met** | 3/5 | 4/5 | Medium | Project deliverables delayed | Use agile sprint |
| **Insufficient alignment between project goals and organizational goals.** | 2/5 | 3/5 | Medium | Final product doesn’t meet HU or lectorate expectations | Regular stakeholder meetings |

This table gives us a clear understanding of what could go wrong during the project and what we can do to address it. In this way, we ensure that when we encounter any issues, we can deal with them effectively and continue working toward our final goal.

## Terminology

|  |  |
| --- | --- |
| Word | Meaning |
| Scrapen | Automatically collecting job postings from websites (e.g., Indeed/company sites) and storing the relevant info in our database for analysis. |
| Curriculum | The study plan of a degree program: courses  + learning outcomes. We upload this to compare it with the labor market. |
| Matching‑algoritme | The (comparison) between two words: skills in the curriculum vs. skills in job postings.  The algorithm finds matches and gaps and  shows them in the dashboard. |
| Skill gap | The difference between the skills in the curriculum and the skills employers ask for  (missing or weak areas). |
| Pipeline | The fixed steps data goes through ingest → clean/transform → store → analyze → show  in the dashboard (as automated as possible). |
| GDPR/AVG | The EU privacy law. We work in compliance with GDPR and also follow each website’s Terms of Service—meaning we handle access, rate limits, and what is/isn’t allowed properly. |
| NLP | Used to identify skills from job postings and link them to the competencies in the curriculum. It is an AI technique that enables computers to understand and analyze human language. |

# Determine data mining goals

## Data mining goals

Organizations have access to increasingly large amounts of data, but leveraging this information is crucial for making better decisions. Data Mining helps uncover hidden patterns and trends, while methods such as Business Case Model, SWOT-analysis, and the five forces model link these insights to strategy and market positioning. This enables companies to seize opportunities, reduce risk and strengthen their competitive position.

**The Business Goals of the Competency Monitoring System:**

The Business Goals are outlined on page number 13.

**The Data Mining Goals of the Competency Monitoring System:**

1. The automated data collection and processing pipeline will ensure that all relevant data is consistently prepared and available for mining tasks with at least 90% data completeness and no more than 6% duplicate entries. This automated pipeline not only reduces manual intervention but also ensures high data integrity, consistency, and readiness for downstream analytical and machine learning tasks.

2. An automated web scraping system will continuously extract and standardize job posting data, achieving at least 90% extraction accuracy and 90% temporal consistency across weekly samples, to provide a robust and reliable foundation for mining job–skill relationships.

3. A working matching algorithm will be developed that correctly matches job postings to relevant competenties taught by HU programs based on required versus learned skills, achieving a precision score of at least 90%.

## Data mining success criteria

**Task**

The goal of this step is to describe what the data mining part of the project should achieve in technical terms. These goals support the business aim of connecting education with the labour market, improving study programs, and using data to make better decisions.

**Data mining goals and success criteria**

|  |  |  |
| --- | --- | --- |
|  | Data mining goal | Business Succes Criteria |
| 1 | The automated data collection and processing pipeline will ensure that all relevant data is consistently prepared and available for mining tasks with at least 90% data completeness and no more than 6% duplicate entries. This automated pipeline not only reduces manual intervention but also ensures high data integrity, consistency, and readiness for downstream analytical and machine learning tasks. | Within 2 months the processing system is fully automated, so the coordinator needs to do at least 80% less manual data entry compared to the start. |
| 2 | An automated web scraping system will continuously extract and standardize job posting data, achieving at least 90% extraction accuracy and 90% temporal consistency across weekly samples, to provide a robust and reliable foundation for mining job–skill relationships. | Around 1 month there is a web scraping system that collects at least 100 job listings within one week. When checked, at least 80 of them have correct and complete information. |
| 3 | A working matching algorithm will be developed that correctly matches job postings to relevant competenties taught by HU programs based on required versus learned skills, achieving a precision score of at least 90%. | Around 1,5 months, there is a matching algorithm that connects at least 4 job listing to students. When checked by the coordinator, at least 3 of these matched are found to be relevant. |

We succeed if the system mostly works automatically, collects current job listing reliably, makes relevant matches between students and jobs, and delivers a proof-of-concept where soft skills are linked to the curriculum before the end of the semester

# Produce project-plan

## Project-plan

**Project Overview**

For Utrecht University of Applied Sciences, we will develop a website where the lectorate and institutes can make an input with the name of the curricula. In the backend, a program will run that matches these curricula and the associated soft skills with the top 100 jobs from Indeed.

The choice for a website was made because it is more user-friendly and more accessible for students. Once the system is fully developed, the website can be used everywhere, making the program accessible to anyone who wishes to use it.

By scraping indeed, we will collect job postings and store them in our own database, allowing us to create a reliable match between a student’s curriculum and the requirements of companies.

It’s also important to make sure that the project is clear which stakeholders are involved. To make sure that everything is clear we made a Stakeholder Analysis.

|  |  |  |
| --- | --- | --- |
| Stakeholder | Role | Internal/ External |
| Hend Elsayed | Client, Final Responsible | Internal |
| Gerrita van de Ven | lector in the Lectoraat marketing and consumer experience | Internal |
| HU IT Department | Technical support | Internal |
| HU Lecturers | Main end users | Internal |
| Top 100 Companies | Provide job posting data | External |
| Other Universities | Potential future users | External |
| Students | Secondary End Users | Internal |

**Salience Model**

For the analysis, we chose the Salience Model because this model helps us evaluate stakeholders based on power, legitimacy, and urgency, allowing us to identify who has the greatest influence and who should be prioritized in the project.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stakeholder | Power | Legitimacy | Urgency | Category | Strategy |
| Client (Hend Elsayed) | ✔ | ✔ | ✔ | Definitive | Bi-weekly meetings & decision-making |
| Assistant  Manager | ✔ | ✔ | ✘ | Dominant | Reports & escalation if  needed |
| HU IT  Department | ✔ | ✔ | ✘ | Dominant | Bi-weekly technical meetings |
| HU Lecturers | ✘ | ✔ | ✔ | Dependent | Bi-weekly demos & feedback |
| Top 100 Companies | ✔ | ✔ | ✘ | Dominant | Quarterly meetings |
| Other Universities | ✘ | ✔ | ✘ | Discretionary | Newsletter/quarterly  update |
| Students | ✘ | ✔ | ✔ | Dependent | input on needs |

**Expectations Needs:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stakeholder | Expectations | Needs | Urgency | Category | Strategy |
| Client (Hend Elsayed) | Working MVP, future-proof system | Reliable data, clear plannin | ✔ | Definitive | Bi-weekly meetings & decision-making |
| Assistant  Manager | Clear communication, reports | Clear role division | ✘ | Dominant | Reports & escalation if needed |
| IT Department | Technically feasible solution | Secure & maintainable system | ✘ | Dominant | Bi-weekly technical meetings |
| HU Policymakers | Data for strategic decisions | Reliable, representative  analyses | ✔ | Dependent | Bi-weekly demos & feedback |
| Lecturers | Insights for curriculum  development | Practical results | ✘ | Dominant | Quarterly meetings |
| Other Universities | ✘ | ✔ | ✘ | Discretionary | Newsletter/quarterly  update |
| Students | Better aligned education | Up-to-date skills, collaboration | ✔ | Dependent | input on needs |

**IST-SOLL-GAP**

Current Situation (IST):

* Vacancy data is outdated (last used in 2021), making insights no longer representative of the current labor market.
* Data collection is done manually, without automated scraping or updating processes.
* There is no algorithm in place to automatically link competencies to curricula.
* The current process does not provide predictive insights into future competency trends.
* No dashboard is available for visualization or user-friendly interaction for lecturers, students, or policymakers.
* The current solution is only applicable within Utrecht University of Applied Sciences and is not scalable to other institutions.
* There is no complete compliance with GDPR and Terms of Service guidelines when collecting vacancy data.

**Desired Situation (SOLL)**

* The system automatically collects new job vacancies annually through scraping platforms such as Indeed.
* A competency-matching algorithm based on Natural Language Processing (NLP) is used to link skills to curricula.
* An interactive dashboard provides up-to-date insights, allowing lecturers, teachers, students, and policymakers to easily access results.
* The solution is scalable to other educational institutions within the Netherlands (e.g., other research groups or universities of applied sciences).
* All processes comply with GDPR regulations and platform Terms of Service (ToS), with clear documentation on data usage.

**GAP Analysis:**

|  |  |  |
| --- | --- | --- |
| Category | Current Situation (IST | Desired Situation (SOLL) |
| Data | Vacancy data is outdated (last updated in 2021) and collected manually. | Up-to-date vacancies are collected annually through automated scraping ( Indeed, |
| Technology | No automation pipeline or system to link job data to curricula. | An automated pipeline using NLP-based matching is implemented. |
| Output | No visual reporting or clear competency insights. | An interactive dashboard visualizes skill gaps and matching outcomes. |
| Users | System is only internally accessible by the current project team. | Accessible for lecturers, teachers, students, and policymakers. |
| Scalability | Usable only within HU and not transferable to other institutions. | Scalable to other Dutch universities of applied sciences or research groups. |
| Legislation / Compliance | GDPR and Terms of Service requirements are not fully met when collecting data | Fully compliant with GDPR and platform ToS, including transparent documentation of data usage. |

**Timeline & phases**

To create a clear overview of the project flow we are going to define phases. With the phases. The project has a total of 8 phases. These eight phases are each two weeks long. This gives us enough time to make sure that every phase is completed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sprint | Week | Goal | Tasks | Deliverables |
| **Sprint 1** | 1-2 | Introduction | Kick-off Meeting + Team agreements | Kick-off document. First version of Business understanding |
| **Sprint 2** | 3-4 | Business Understanding | Making the business understanding assignment | Business understanding document |
| **Sprint 3** | 5-6 | Process feedback business understanding + Starting on Data Understanding | Process feedback of teachers into our document | Final product of Business Understanding + start of Data understanding |
| **Sprint 4** | 7-8 | Feedback Business Understanding + First version of Data Understanding | Finishing Feedback on Business understanding. + First Version of Data understanding | Business Understanding + First version of Data Understanding |
| **Sprint 5** | 8-10 | Finalizing Data Understanding + Start on Data preparation | Finishing the Data understanding + Working on first version of Data Preparation | Data understanding + First version of Data preparation |
| **Sprint 6** | 11-12 | Final version of Data preparation and making a first version of moddeling | Finishing Data Preparation + First version of Modelling | Data preparation + first version of modelling |
| **Sprint 7** | 13-14 | Working on Final Product & Start Evaluation | Finishing the Product + start Evaluation | Modelling + Start evaluation |
| **Sprint 8** | 15-16 | Deployment of Final Product | Deployment | Final presentation + Final product |

Afbeelding met tekst, schermopname, diagram, lijn

Door AI gegenereerde inhoud is mogelijk onjuist.To make the Timeline and Phases even more understandable we made a gant chart. By using a gant chart we can make the project visible with just one look

If you would like to see the Gant Chart on your own computer you can use the following link:

<https://lucid.app/lucidspark/fb1c183d-d8b8-4abd-b1fe-a532e7d8bfdb/edit?viewport_loc=3486%2C1142%2C3454%2C1628%2CuDe-dIt-NWfS&invitationId=inv_2f5f68c1-2fae-4d4c-8053-22f46cc3a8bc>

In the Gant chart, you will see:

* What needs to be done
* When will it happen
* Who is working on it
* How everything is connected

When using the Gant chart we can see how far we are with the current project. Plus we can set deadlines and monitor how the project is evolving.

With this overview, we can easily see:

* If we are on schedule
* Where is the delay
* How one adjustment has effect on the rest of the project

In the Gant Chart you see that we already encountered some problems. That’s why from sprint four we made new steps in the planning to make sure that everything is still possible just with more work and time in the project. Even when we lost time in the previous sprints.

**Resources & Roles**

The resources and roles within the project have been clearly defined. The resources were previously outlined in the *Inventory of Resources* section on page 15.

The roles have been clarified by creating a table that specifies each team member’s responsibilities.

|  |  |
| --- | --- |
| Name | Role |
| **Jason** | Leading the project. Responsible for project plan in business understanding with Baha. |
| **Baha** | Keeping all project documents organized. Responsible for Project plan in business understanding with Jason. |
| **Zoë** | Responsible for inventory of recources and a part of requirements, assumptions and constraints in business understanding. |
| **Yara** | Responsible for a part of requierements, assumptions and constraints and terminology in business understanding. |
| **Fatima** | Responsible for Initial assessements of tools and techniques and data mining succes criteria in business understanding |

This is, of course, a summary of how the process works. The specific task assignments are discussed via WhatsApp, where the corresponding deadlines are also communicated.

**Documentation & Review**

All project documents are managed by **Baha**. This ensures that the documentation remains well-organized and forms a coherent narrative, rather than having separate, uncoordinated sections written by different team members. The completed work is reviewed by **Baha** and **Jason** to ensure that all parts align properly and maintain consistency throughout the project.

Every two weeks, we hold a **sprint review** where we present our project progress and upcoming plans. During these sessions, we receive feedback which we then apply to improve our workflow and the overall project outcome.

## Initial assessment of tools and techniques

At the end of the first phase, we reviewed possible tools and techniques for building a prototype of a sustainable, automated competency monitoring system.

The current system predicts which skills are needed in the professional field based on job vacancies. This helps Dutch universities of applied sciences align their study programs with labour market needs. However, it uses outdated data from 2021 and requires manual work.

Given the short timeframe of about four months, the focus is onbuilding a working prototype that:

* Runs the full data process automatically (collecting, cleaning, matching and reporting)
* Matches job skills with student competencies
* Shows result in a simple interactive dashboard
* can be scaled in the future to include new data sources and programs.

**Purpose of the assignment.**

The assessment helps select tools and techniques that best support the project’s data mining goals. These tools and techniques will help collect, clean, and analyze job and curriculum data, match skills and show results in a dashboard. The data mining goals and success criteria, which explain what the system should achieve, are listed on the data mining goals page. This assessment makes sure the chosen methods are practical, efficient and suitable for the project’s short timeline.

**Assessment of tools**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tool | Function | Strengths | Weaknesses | Evaluation |
| **Python (pandas, beautifulSoup, scikit-learn, spaCy and Selenium)** | Data collection, cleaning, matching and automation. | Open-source, large community, supports automation and scraping. | Requires coding skills | Highly recommended; ideal for automated pipeline and webscraping. |
| MSQL / Azure | Data storage and management | Reliable, scalable, intergrates with python and other tools. | Requires setup and maintenance | Recommended: stores data for historical tracking and updates. |
| **Power BI** | Data visualization and dashboards | User friendly interface, interactive visuals, intergrates with multiple data sources. | Limited advances analytics: less flexible than coding sloutions. | Recommended: perfect for reporting and interactive dashboards |
| Power BI / RapidMiner | Visual data mining workflows | No-code interface, quick to prototype models, easy to test matching or transformation logic. | Limited customization and scalability for large or complex data. | Optional: useful for early experminentation or testing matching algorithms. |

This project we will primarily use Python for data collection, cleaning, matching and automation. Azure for structured data storage. And we use Power BI for dashboards. This combination is practical, scalable, and open-source.

**Assessment of techniques**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Technique | Purpose | Strengths | Weaknesses | Evaluation |
| **ETL (extract, transform, load)** | Automates collecting, cleaning and storing data. | Saves time, reduces manual work. | Needs some setup | essential: needed for automation and yearly updates. |
| **Natural Language Processing (NLP)** | Finds and compares skills from job posts and study programs | Can read and understand text automatically | Needs text cleaning to work well. | Highly recommended: very important for skill matching |
| Keyword matching / similarity check | Compares skills between jobs and what students learn. | Simple, quick and easy to understand. | Not as accurate as advanced methods. | Recommended: good for prototype |
| Viable system modelling (VSM) | Helps organize and explain the system and curriculum links | Gives a clear structure or framework | It is hard to understand and use. | Keep as theory: supports background understanding |
| Predictive modelling / trend analysis | Predicts future skill needs | Uses old data to make simple forecast | Not always accurate with small data | Optional: use if there is time |

This project will mainly use NLP, keyword matching and ETL. These techniques help the system read text, compare skills and automate data collection and cleaning. Optional methods like predictive modelling and VSM can be used for forecasting and curriculum guidance in the future.

This assessment shows the best tools and methods for building a sustainable and fully automated Competency Monitoring System within four months.

The system will use Python for automation and data handling, Azure for data storing, and Power BI for dashboards and reports.

Main techniques include Natural language processing (NLP), keyword matching and ETL (Extract, transform, load). These make it possible to collect, clean, match and report data automatically.

Overall, this setup will deliver a working prototype that needs little manual work, gives clear visual result and forms a strong base for future growth and automation

# Sources

[Artificial intelligence in het onderwijs: dit zijn kansen en risico's - Kennisnet](https://www.kennisnet.nl/artificial-intelligence/artificial-intelligence-in-het-onderwijs-dit-zijn-kansen-en-risicos/?utm_source=chatgpt.com)

[Privacy in education | Privacy First](https://privacyfirst.nl/en/articles/privacy-in-education/?utm_source=chatgpt.com)

[Cyberveiligheid | Onderwerp | Inspectie van het onderwijs](https://www.onderwijsinspectie.nl/onderwerpen/cyberveiligheid?utm_source=chatgpt.com)

[Meer kennis over AI in het onderwijs nodig, anders risico op discriminatie | PO-Raad](https://www.poraad.nl/meer-kennis-over-ai-in-het-onderwijs-nodig-anders-risico-op-discriminatie?utm_source=chatgpt.com)

<https://www.google.com/aclk?sa=L&ai=DChsSEwj19Lfz3O6PAxVikIMHHfGTNVQYACICCAEQABoCZWY&co=1&gclid=CjwKCAjwisnGBhAXEiwA0zEORzyYBHWiJRQnT7l7LZE5E-sh7ZLsecORnXf-ebppw2rvrqClv5vysRoCA2wQAvD_BwE&cce=2&sig=AOD64_33JlXu4NVM0lQFJ_oyZGfE0q2tGA&q&adurl&ved=2ahUKEwj3ybDz3O6PAxV9-QIHHWhAKkMQ0Qx6BAgYEAE>

Bronnen SWOT-analyse:

Hogeschool Utrecht. (z.d). Marketing & Customer experience.

[https://www.internationalhu.com/research/marketing-and-customer-experience?utm](https://www.internationalhu.com/research/marketing-and-customer-experience?utm_source=chatgpt.com)

Hogeschool Utrecht (z.d). onderwijs en onderzoek.

[https://www.hu.nl/werkenbij/onderwijs-en-onderzoek?utm](https://www.hu.nl/werkenbij/onderwijs-en-onderzoek?utm_source=chatgpt.com)

Hogeschool Utrecht (z.d). creative business.

[https://www.hu.nl/voltijd-opleidingen/creative-business/during-the-programme?utm](https://www.hu.nl/voltijd-opleidingen/creative-business/during-the-programme?utm_source=chatgpt.com)

Hogeschool Utrecht (z.d). Studiegids.

[https://studiegids.hu.nl/028c0c1c-9805-4519-93a3-a91528dbdfed?utm](https://studiegids.hu.nl/028c0c1c-9805-4519-93a3-a91528dbdfed?utm_source=chatgpt.com)

Jermore Knoot. (z.d). Strategische Marketingplan.

<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwilmrPjnu2PAxV09LsIHfm7DWYQFnoECBcQAQ&url=https%3A%2F%2Fwww.strategischmarketingplan.com%2Fmarketingmodellen%2Fvijf-krachten-model-porter%2F&usg=AOvVaw0Ly1QpOSovIT2OoJEFOnQl&opi=89978449>

Hogeschool Utrecht (z.d.). *Visie op onderzoek.* Geraadpleegd via <https://www.hu.nl/onderzoek/visie-op-onderzoek>

Agile Scrum Group (z.d.). *Stakeholder management model (Salience Model).* Geraadpleegd via <https://agilescrumgroup.nl/stakeholder-management-model/>

Master Challenge (z.d.). *Challenge space registratie.* Geraadpleegd via <https://masterchallenge.me/account/challenge-spaces/f0517013-d2fe-411a-b9f9-23235e57d3ec/registration>

HU. (z.d). *Marketing & Customer Experience*. Geraadpleegd via <https://www.hu.nl/onderzoek/marketing-en-customer-experience#Onderzoekslijnen>