My thesis title

Dafydd James Thomas

Department of Computer Science
Aberystwyth University
Wales

August
2025

This thesis is submitted in partial fulfilment of the
requirements for the degree of Master of Science

Degree: MSc Data Science
Module: CHM9360
Supervisor: Dr Arina Buzdalova

Abstract

The abstract stands alone as a very short version of the dissertation.

The abstract should state the scope and principal objectives of the project, describe the methods, summarize the results and state the principal conclusions.

(Max. 500 words.)

Declaration of originality

I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for Unacceptable Academic Practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the regulations on Unacceptable Academic Practice from the University's Academic Quality and Records Office (AQRO) and the relevant sections of the current Student Handbook of the Department of Computer Science.
- In submitting this work, I understand and agree to abide by the University's regulations governing these issues.

Name: Dafydd James Thomas
Date: 05/09/2025

Consent to share this work

- By including my name below, I hereby agree to this thesis being made available to other students and academic staff of the Department of Computer Science, Aberystwyth University.

Name: Dafydd James Thomas
Date: 05/09/2025

Acknowledgment

To whoever has the patience to read this :-)
This section is customary, but not obligatory. It makes a brief statement of thanks to those who have helped.

Contents

1 Introduction **1**

2 Literature review **9**
2.1 Initial Research Questions 9

3 Data Methods **14**
3.1 Collection 14
3.1.1 Open-Source Software Data 14
3.1.2 Job Posting Data 15

4 Representation **16**

5 Cross-Domain Similarity **17**

6 Time Series Methods **18**

7 Experiments **19**

8 Discussion of Results **20**

9 Critical Evaluation **21**

10 Conclusion **22**

Appendix A **24**
A.0.1 ESCO Skill Classification 25
10.1 Word Counts 25

List of Figures

1.1 Architecture overview: vector embeddings with ellipses, similarity matrix, and temporal interactions (left to right). 3

List of Tables

A.1 Overview of datasets used for Job Postings 24
10.2 Document summary 25
10.3 Word count by file (`texcount`) 25

Chapter 1

Introduction

Background to the project, motivation, leading to project aims and objectives.

- What problem was tackled?
- Why was that problem tackled?
- How (in outline) was the problem tackled?
- Clear statement of project aim and objectives.
- Guide to subsequent chapters.

In recent times, we are seeing that skills and job requirements are evolving rapidly. As the culmination of various effects, including the development of new technologies and evolving consumer preferences, the labour force must adapt in order to satisfy the new requirements. Given the historic pace of labour demand shifts, traditional methods of forecasting these changes are not well-equipped to handle this blistering, and accelerating, rate of change.

By these so-called "Traditional" methods, I mean that which accompany government or firm-level analysis of the skills market. These depend upon data which are limited in the following ways:

1. They use structured data
2. They collect the data periodically (e.g. Annually / Quarterly)

This paper aims to develop an investigation of tools that can be used to supplement the forecasting of *Skill Demand* using data that does not suffer these limitations.

Namely, this paper demonstrates the efficacy and power of using the data held within **Job Postings**, to provide valuable insight into the trends and evolving nature of job requirements. Notwithstanding any limitations, this approach is

promising because the data in question is very high-frequency - this affords their forecasts the benefit of effectively real-time inference.

To narrow the scope, I only consider skills that pertain to the Technology & AI sectors. It should be noted that this is choice is made on pragmatic grounds:

- It affords greater depth to discussion of the various techniques.
- We can reliably expect online Job Postings to be omni-present for Tech jobs.
- Evolving skill-requirements is a particularly relevant theme for these sectors.
- They are of personal interest to me.

Argument 1.1: Thesis Argument

Assumptions

A1 Open-source activity reflects experimentation.

A2 Repo stars are a noisy but useful signal of relevance.

Hypotheses

H1 GitHub activity precedes job market demand.

H2 Repo embeddings will correlate with job description embeddings.

Theoretical Premises

P1 Text embeddings are computable from READMEs and job ads.

P2 Cosine similarity is an appropriate measure of alignment.

P3 We can construct aligned time series of similarity and mentions.

Argument 1.2: Thesis Argument

Assumptions

A1 Open-source activity reflects experimentation.

A2 Repo stars are a noisy but useful signal of relevance.

Hypotheses

H1 GitHub activity precedes job market demand.

H2 Repo embeddings will correlate with job description embeddings.

Theoretical Premises

P1 Text embeddings are computable from READMEs and job ads.

P2 Cosine similarity is an appropriate measure of alignment.

P3 We can construct aligned time series of similarity and mentions.

As in Argument 1.1

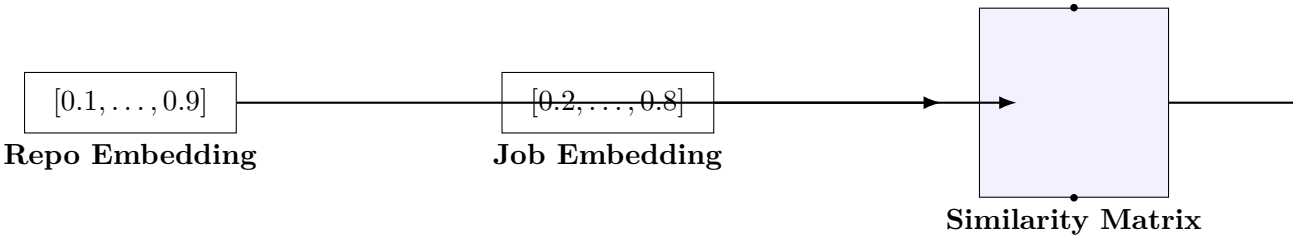

Figure 1.1: Architecture overview: vector embeddings with ellipses, similarity matrix, and temporal interactions (left to right).

What is the problem?

An important question that policy majors are closed and policy and labour market, participants have to face is What skills? Unnecessary. For the workforce.

A problem that is faced by policy makers and employers. As well as market labour market, participants. Is how to translate. Job requirements into educational programs or moreover, how to understand, what requirement what jobs are requiring nowadays. Do we understand the shifting demand for skills as they were as they relate to job jobs?

Historically, this kind of problem. Is complicated. Due to the technological shifts that underlie. Economic revolutions in the way that we in the work. That is needed. And as a result, the skills of work. Must evolve in tandem for this technology shift. With previous technological shows. This problem has been

touched downplayed in.

Emergency.

Properly mapping changing requirements. To technological shift has been historically. More extended tasks, and Because of the way that previous, Disruptions in technology, have diffused, so slowly. The immediate. Necessity of proper. The proper mapping. This sense is has been was much a much lower urgency. But the shift that is currently underway as a result of AI.

Is far more disruptive and is involving at a much greater Pace with orders and magnitude to compare to different two previous technological instructions. Moreover not only is it such that these skills are changing very quickly due to the rapid development of Technology. But the diffusion into the economy is, is far more profusive due to the fact that the technology is very general use.

As such it is. Hikens the water as such. This problem, as it relates to the mapping of AI, Evolution with regards to skills. His Mark is far more critical than it has been for previous problems.

What are the specifics of the problem? Kevin approaches. Can't Tackle. Well, if we think about how Labour requirements. Are understood. And Tracked by governments. We often think that. Governments.

Governments tend to understand that.

With this kind of Economic analysis, exercise.

It is not sufficient to rely on the information provided by a single company alone.

In a matter of speaking. Understanding the economy. Why of chips and labour demands. Is necessarily something that requires investigation and deep planning. As well as forecasting by rigorous methodologies. Developed within and by institutions with access to the large-scale data that I'm talking about. So, Them less wordy or less, for both terms.

Governments tend. Institutional wide. Surveys of current of the current Uh, conducted to understand. Jobs are available not jobs require and how companies are changing their requirements. In a holistic sense. These labour force or labour economic. Surveys. And Investigations cannot strictly be replaced due to their robust and informative methodologies, but there are aspects of these surveys that can't Quite adapt to the.

Very frequent or high frequency nature of. The AI disruption shift. Basically. Job requirements. Changing job, requirements as they relate to AI related SDA as they relate to AI, skill, AI skills.

Far outpace. A rates at which, Broad and deep investigations can be conducted. As a result. We may see these forecasts lag behind the nuances of what it actually occurred within the market. Therefore, it does fall back a much more high frequency mapping form and backboard. Forecast. Or understanding of the labour

market needs to be.

Foundational. To how policy makers and employers look at the market statement. Currently

10 years. With say the uprise of data science. And Software engineering. Foundational or large-scale. Seconds. Governments might have been able to conduct deep investigations and analyses of what the demand for these types of jobs would be next year. And whilst this kind of analysis, might still be quite relevant.

With today's shifting, requirements and skills and Actual job, demand. As a result of the developments in AI technology. We need these kind of forecasts for analysis to be conducted much more frequently perhaps at a quarterly basis and soon at a monthly basis.

In some sense, this is Not feasible. Or will take quite a while for institutions to improve their forecasting. Capacity to conduct such research at a high frequency. But, It Remains that there are institutional roadblocks and A highly coordinated data. Infrastructure is necessary for this kind of project to occur.

It therefore Falls that we need other methods and other ways. Understanding. Job requirements, how, and how job requirements are changing at a high frequency basis and one, one source of data that is An excellent proxy for.

Another way to look at this problem, is that Rather than depending on the highly structured and Deeply detailed information with that. Policy makers collect. At a period. Not a periodical basis from Other institutions such as employers and Companies.

There must be some sort of Market signal, which To some extent can stand in and Behave as a proxy for this kind of Labour, demand information. The for Excellent proxy that we have for. Understanding the labour market demand. Is job postings. Whereby companies. Uh, signalling. Job openings and demand for certain roles and skills by online, posting of job, advertisements.

These job advertisements are much of the time. Buried. And how detailed or how structured? They relate this patient. But nevertheless remains They are high frequency. Sources of information that relate to the requirements of the worker that they're seeing now. If there was a way that we can understand, How these job postings are trending.

In what skills they detail as necessary. That it stands to be a much more high frequency source of information. That the policymakers and employers and labour labour market, participants can Makes you soft. Educating their decisions on how to. Devised policy or how to Skills competitors. Demanding. What skills should I be re-educating myself in?

To acquire jobs in the future. As it relates to AI Technologies. We can. And bend within this analysis of job, postings. A depiction of whether or not, Innova-

tion signals. Informative in terms of how these skill profiles are changing.

Acknowledges.

At the beginning of the Telecommunications Revolution. Policy makers. To not properly account for the change or impact that information technology might have on the skills required by their labour.

A big part of this was not understanding properly. What the Innovation for that signals were. And how?

Must it may prove beneficial over and above. Institutional wide surveys of labour market. Of the labour market requirements. Job posting analysis can certainly To provide new information on a much more high frequency basis. The roadblock that stands in the way of, this is how to process such unpro such.

Unstructured, textual data in a manner that is. Informative. And Insightful.

This is where natural language processing techniques. Have been major

Impact. But it nevertheless remains that if we're mapping skit, changing skill profiles as a result of technological disruption. Maybe it may prevent beneficial to incorporate information as relates to the technological, disruption to begin.

Innovation signals.

Innovation tends to signal, technological destruction, on this in this sense and as such mappings or incorporating, the information of these Innovation, signals would be highly beneficial to understanding how we might expect to see job postings change in the future as a result of currently occurring. Just Innovation teams. So My related skills.

And other software engineering. Enterprises. A big part of this kind of innovation signal is open source software and the developments that are available there. Now whilst close source Is certainly another side of the same coin? Open source software is beneficial to this kind of analysis in two senses. One.

It is perfectly openly available and useful for analysis. Open source is. Very widely and broadly considered to be a Innovation bed. I'm bad for an event, the Catalyst bed for Innovation, due to the open. And Encouraged collaboration networks. Supports.

But with regards to AI related skills, there could be no more relevant. There could be no more relevant Innovation sequence, except except for perhaps Search payments.

So what do I necessarily mean by? What is this approach that I'm detailing?

The overall aim of this paper. Is to provide. Some depiction of whether or not, Open source, software settings. Can help. Or improve. Forecasts of job postings. Of of the skill profiles, as they relate to software, engineering, and data scientifically. Chop rolls in their job postings. More specifically. It is an aim of this paper to.

Extract. Time series, The Time series representation of skill profiles. As semantic. Vector, vector embeddings of the semantic information within job postings. This Vector embedding can then be Compare directly in a quantitative manner. To the semantic information that is prevalent in GitHub repositories. Now what this provides is. A similarity or a quantitative measure of how job postings relate to GitHub repository, and Given the job postings Time.

Series nature. We have a Time series of how job posting from late are a similar quantitatively compared to what GitHub repositories in time. On the flip side. We have a direct way to measure digital, repositories popularity, and hence, Effective.

Pervasiveness in the field of AI with regards to how popular this

So, As it stands. The aim of this paper. Is.

Time series. That each speak to some level of.

So, the aim of this paper is to determine Whether or not get some repositories and their information. Are causally impactful. Or at least precede temporally the rise of similar semantic. Semantic. Similar skills. Within the semantic content of job postings.

Before this it remains to derive two, distinct. Time series. From two channels of data.

So, we can think of this broadly, as needing Channels of data. Being job postings and GitHub repositories. These two channels of data each exhibit, different properties, but mostly we can separate these out into a Time time stamped. And Textual. Information.

But the aim of this fucking project, Is. Reduce. Two channels of data. Job postings and GitHub repositories.

Respective. A semantic representation. And it times that feature. The regards to job postings, this will relate to the skill profile as a semantic event. Indictates of one that job hosting was hosted. With regards to GitHub, we have The semantic components of. Content. Including,

Topic tags. And, and Read me a description. And, We have. Time, stamped features as regards. These stars that, that repository received on a given date. With these two resulting features, but these two resulting data channels as features. We can produce. Like to combine we can produce asymmetric. Semantic. Similarity, Matrix representing the repository versus a job posting semantic, similarity.

Ultimately, Ultimately depicts, whether or not the semantic content within the GitHub provides, is similar to is similar or in, Represents some similar, similar semantic features, as does the cloud profile as well as does the job posting. We? That their similarities score represents. An underlying similarity. Determined by whether or not the job posting expresses semantic, content, which we would find in the GitHub repository.

With this that we have in time series terms, But each repository, A measure of its popularity in stars, and a measure of its similarity to job postings. Can compare directly to these two taxes to determine whether or not the hypothesis holds. Cold water in the sense that we expect to see some level of proceeding.

Popularity that board their prevalence in job postings.

So the objectives of this paper, Explicit terms.

Extract historical. Data. In both data channels. Pre-Process appropriately, the data for input into the representation model. Determine.

Conducted. Different variations of representation. With experimentation on text pre-processing techniques. Dimensionary, dimensionality reduction. Sentence Transformer methods.

Produce. Time series, analyses. Repository versus

Time series analysis of times of repository similarity. Repository popularity. In time. And understand the landscape of different repositories and their effects, in this sense They? A further experimentation step. Is to instead of represent each single repository as its own. Similarity. Time series. We compose. We conduct. Different variations of cluster clustering methods to produce distilled groups of repositories as a single somatic feature.

My motivation for this project. Based in large part because of my interest in understanding, how Data and machine learning. Can fuse together to inform. Other domains such as economics which I, which is of highly of high interest. Moreover. Given the unprecedented shift in How the labour market will function.

I feel this kind of Technology needs. I feel. That current Technologies available to labour market, participants and policy makers. It's significant. Modernising and adapting to. Whatever that available data that.

Chapter 2

Literature review

Before embarking on implementing the project, extensive research was conducted to understand the various elements at play. The following depicts a list of questions that were devised in the research stage, and are answered thereafter. It is with this information that we can build a picture of what is supported (and what is overlooked) by existing research. This serves as a foundation for my work, giving precedent to the study of the thesis at hand.

As previously mentioned, the thesis argument in the Introduction Chapter grounds its premises on empirical grounds, according to the research thusly discussed. I shall make reference as appropriate to the premises in question.

2.1 Initial Research Questions

1. **AI Revolution & Labour shifts**
   (a) What's happening with AI?
   (b) How does this compare to previous technological shifts?
   (c) How does this affect the workforce?
2. **Labour Demand Analysis**
   (a) How is labour demand forecasted?
   (b) Does AI pose a problem to these methods?
   (c) Are Job Postings useful here?
3. **Innovation & Open-Source Software**
   (a) What is Open Source software?
   (b) Is OSS a valuable innovation signal? Why?

(c) How can we measure the adoption rate of OSS?

4. **Methodological Techniques**

   (a) How can we effectively represent semantic content quantitatively?

   (b) What are the considerations for comparing semantic representations of different domains?

   (c) To what extent can we infer relationships between two time-series?

.

The techniques. Fusing and preparing somatic relationships. Differently domained semantic. Beaches. So, The. Of. Labour demand forecast.

Is a highly researched and highly. And have. Is a highly researched topic? So much, so that. It is embodied within a certain branch of economics, labour economics.

Talk about.

Talk about the Traditional methods. Labour economists, use to forecast labour demand.

It has also previously been considered whether or not job descriptions or job postings are a useful source of information mapping labour demand to the future. Various studies were concerned. Welcome Conducted in providing information. The relevance of job postings. In particular, several natural language processing methods, have been developed previously to

Extract, meaningful features and distilled information of job posting. So our time varying basis. Independent Trends, and Underlying.

The big Trends.

Throughout this paper. I shall, I do use various terms to depict whether or not To depict. What the constant I'm talking about? Is and fundamentally. When we talk about what I say, Stills. This is something that in the paper skill span. Skills. Bad boats. Is discussed. Define, find Esco the European skills commission office.

As knowledge skills. Specifically. Knowledge. A lot of skills.

So specifically. Knowledge. Is the experience and of knowledge of a certain tool. Skills. Uh, the ability to apply this. Broadly, when we're considering these skills or knowledge, like topic of technology is that specifically relate to AI?

Software engineering Technologies and libraries with modules of languages. These kinds of skills and knowledge knowledges. Broadly falling under the Umbrella. What we need? So, In particular, we don't Define that skills, asked we? Look for things that fall under the umbrella top skills such as effective, communication. Critical thinking skills.

As these are not necessarily prevalent to AI. Any more so than they are to most other fields and in particular, because they do not necessarily reflect what the semantic. Representation of digital repository will reflect. So,

Other skills related literature investigations.

Another facet that underlies, this paper is The notion that Innovation. Is directly tied in to labour market plants. And this is well supported by previous research, especially in the domains of Research papers. And Applications.

This kind of research is ongoing and developing. But the vast majority of research definitely suggests. Innovation, diffusion. Is sexual. Or that did that the diffusion of technological innovation. Takes a while and is modelled by some Dynamics. Not well, understood in the context of AI related to small software.

Moreover. It may not remain relevant that pattern abstracts or research papers. Unnecessarily the most relevant. Innovation signals or certain Technologies.

Open source software. Is widely regarded to be. A catalyst bed for Innovation. In the software, in the software sector. In particular, GitHub is alone responsible for Hosting many guitar, many open source software projects. Is highly regarded, because The central location to find. Useful. Uh, powerful new technologies to development.

Of course, it remains difficult, perhaps, the most Innovative Technologies, remain, closer source. Ones here. Frontier models in the AAA, then the AI segment such as chat GPT. At Google Gemini. But, When we start to see applications, Of these Technologies. Many of these kind of applications. Spectrum. Uh, that is upon and assessing the Pervasiveness of this kind of Technology here, then, getting repositories.

And how Popular. These tend to become over time. The exact kind of information, I'm looking to extract and use as analysis or information. Forecast, the language of jobs.

So, why GitHub starts? Among money measures that GitHub provides to. Assess, the many dimensions of Of an open source an open source software in time. Stars. Remain to. They're very succeed, and Heuristically. Viable methods of assessing whether or not a GitHub repository is personally useful secondly popular and 30.

Frequently, collaborate, collaborated on or contributed to So, it's important to understand that GitHub has The use cases. The first use case is clearly. Uses to make available. To the open source community, and the public in general. Their developments and their technology for their software. The second use case is, The second use case is, perhaps the most Useful.

Bunch of GitHub. And this is for collaborative purposes, so the user can not only share their projects Actively post. Deliberate collaboration and improvements.

This makes software. Apart from being perhaps. Of a personal.

This, this marks the ship. New technologies or new applications of Technology away from. Scenarios, which would be mine. Paued.

Complete ownership over our own developments. Github represents the shift to. Making Technologies available and Making them, a means to edit and improve these Technologies. Effective.

In fact, it is often regarded that the most effective pieces of software are all open. Unified proprietary software. Is somewhat substandard.

So, ask me might expect that the tech modules disruption. Ai represents is mostly coming in the form of. Someone being able to use these Frontier proprietary Frontier models. The developers of AI models themselves. Very frequently relying on Frameworks or libraries that are open source and have spent a long time in development.

From a committed community. Developers. To improving the efficiency and the the effectiveness of these Frameworks. An example. Is pytorch. I have tensorflow. These are examples of Pythonic. Programming Frameworks. For machine learning or deep learning and More frequently than not. These are the frameworks that I'm the pen. Someone's capacity to produce a produce, a product or conduct work within deep learning or machine learning.

So, I expect. It is expected. That seeing these models become more widely used and more popular that we might be. Expecting. Or expecting labour demand to focus more on these Focus more on on, Top Seekers that possess the skills?

What remains to be seen is whether or not stars is the most effective way to measure. Repositories. Foundational.

The wire stars. The way to judge. Possum trees. Elevations. As a heuristic method, it is the most simple and effective manner to measure a repository popularity. It's caveats. For example. Research suggests that. GitHub stars are sometimes. Reflective not of. But underlying usefulness or popularity. But, Marketing. Type of marketing efforts and in some cases of deceitful, Bot accounts.

To boost to artificially, boost the popular. Popularity of this. Mod of this repository. Nevertheless. The most popular postures. Certainly do depend. Candid. Trapped in how stars are gained. And,

In some sense, a continual. A continual and increasing trend of star. Is certainly negative of a repository that is genuinely useful. And Gaining traction. As a foundational unit for Producing. Cutting-Edge products.

So, as regards to

A big challenge that this challenge that this paper really faces up against is The fusing of the semantic representations of jobs, but this is the semantic representation of genitals and in some sense, it is dubious. Especially, since a lot of the

time, the content that we see job postings, It's highly dissimilar to what you might expect to see in a Gator repository.

And there are some elements of research that suggest various methods of being able. So, effectively used Semantically different. Textual passages.

With the regret. It does not remain within the stove of this paper. To fully investigate and develop the method that is. Advanced. For this specific kind of dual representation. Elephants instead. This paper. Extends. The research enabled heuristic methods for. Extracting relevant text.

Producing. Semantic embeddings. Highly developed and well supporting percentage transformal models. Transformer models are effectively.

And talk about, The B6 that are the highest sentence Transformers.

Talk about clustering.

And talk about time series. Correlations.

Chapter 3

Data Methods

3.1 Collection

In the case of this project, there are two facets of required data:

1. Open-Source Software Proliferation Data
2. Job Postings Data

3.1.1 Open-Source Software Data

Thankfully, the process of collecting data related to open-source software is fairly straightforward. As the name suggests, the OSS projects are intended to be *open* and accessible. This necessitates the usage of some online platform for hosting and maintaining the codebase, with strong version-control systems to safeguard a maintainable codebase. GitHub is functionally the default choice for hosting both open-source (and closed-source) codebases for a number of reasons, and is therefore the most relevant source for the required data of this project.

This project seeks data that more or less speaks to how *popular* a given GitHub repository is, and the prime candidate for this is the *GitHub Stars* mechanism. Users will ”star” a repository in the same manner that internet users ”bookmark” a website, behaving as a shortcut. This is a grossly simplfied view of how GitHub repositories are interacted with, however - most GitHub repositories have no stars, and most users have ”starred” very few repositories in return. The star is instead a useful proxy for measuring interest from a development point of view - if someone is contributing regularly to a repository, they will star it. In this sense, stars are a useful measure of how a repository is gaining interest from developers (and in a lesser sense, users), which in turn represents a growth in the utility the repository represents.

[1] is an open-source project that provides data on the historic star-counts of GitHub repositories, marketed as a tool to "provide insights into a repository's trendiness". Within the ML/AI framework that I delineated, the relevant repositories' star counts were found and downloaded thusly.

3.1.2 Job Posting Data

As is generally true of commercially valuable data, it can prove difficult to find open data that properly satisfies the desired criteria. Generally, it is a relatively simple task to build legally compliant scrapers that programatically access and store the data from job boards online. However, finding stores of historical data presents a major challenge, because only *active* job postings are present on job boards. This means that the scraper must be active throughout and is not useful for collecting data in retrospect.

An alternative solution is to therefore find and access a pre-prepared dataset. This, unfortunately, proved challenging in the sense that:

- Large, high-quality datasets are presented as products, and are thus expensive
- Small, low-quality datasets often lack essential features, such as "job description" and even "date"

The solution to these problems was to utilize *several* datasets that occupy different time-regions, made available on Kaggle. An overview of the characteristics for each dataset is included in the Appendix, as per Table A.1.

At this rate,

Chapter 4

Representation

Chapter 5

Cross-Domain Similarity

Chapter 6

Time Series Methods

Chapter 7

Experiments

Chapter 8

Discussion of Results

Chapter 9

Critical Evaluation

In this chapter (it may not be Chapter 4 for you, but probably Chapter 6 or 7 once all the core chapters have been added.)
The critical evaluation consists of a discussion, leading to conclusion. It is an essential part of a master's degree.
It shows that you can not only carry out a substantial piece of work, but that you can reflect on it, and think critically about how you might have done it better.
Examiners view the critical evaluation as very important.
Critical evaluation should contain

- Strengths and weaknesses of you project
- If you were unable to attain any deliverables, then why
- What are the future plans for your project if you are to continue

You will be presenting this during demonstration but here you need to discuss them in details.

Chapter 10

Conclusion

A brief summary of all that has gone before.

May include some directions for future work.

References

[1] *GitHub Daily Stars Explorer*, https://emanuelef.github.io/daily-stars-explorer. (visited on 07/16/2025).

[2] *Machine Learning Job Postings in the US*, https://www.kaggle.com/datasets/ivankmk/thou ml-jobs-in-usa. (visited on 07/16/2025).

[3] Asaniczka, *Linkedin data engineer job postings*, 2023. DOI: 10.34740/KAGGLE/DSV/7292935.

[4] Asaniczka, *Data science job postings & skills (2024)*, 2024. DOI: 10.34740/KAGGLE/DS/4407481.

[5] A. Koneru, *LinkedIn job postings (2023 - 2024)*, 2024. DOI: 10.34740/KAGGLE/DSV/9200871.

[6] *Job Postings from Ireland (October 2022)*, https://www.kaggle.com/datasets/techmap/job-postings-ireland-october-2022. (visited on 07/16/2025).

[7] *US Job Postings from 2023-05-05*, https://www.kaggle.com/datasets/techmap/us-job-postings-from-2023-05-05. (visited on 07/16/2025).

Appendix A

Generative AI

1. However I use Generative AI.

Data Sources

Dataset Name	Date Range	Details	Source
1000 ML Jobs US	12/2023 - 04/2025	Mostly within March & April of 2025, all jobs are ML related	[2]
Linkedin Data Engineer Job Postings	17/12/2023 - 17/12/2023	Extra "skills" column - This is appended to "description"	[3]
Data Science Job Postings & Skills (2024)	19/01/2024 - 21/01/2024	Extra "skills" column - This is appended to "description"	[4]
LinkedIn Job Postings (2023 - 2024)	23/03/2024 - 20/04/2024	–	[5]
Job Postings from Ireland (October 2022)	01/10/2022 - 31/10/2022	Ireland only; Sample dataset from Techmap.io's commercial product	[6]
US Job Postings from 2023-05-05	05/05/2023 - 05/05/2023	USA Only; Sample dataset from Techmap.io's commercial product	[7]

Table A.1: Overview of datasets used for Job Postings

Third Party Code and Software Libraries

1. Whatever Software I end up using

A.0.1 ESCO Skill Classification

This publication uses the ESCO classification of the European Commission.

10.1 Word Counts

Category	Count
Words in text	4818
Words in headers	46
Words in captions	28
Tables/Figures	4
Inline math	0
Displayed math	0

Table 10.2: Document summary

File	Text	Headers	Captions	Floats	Inline Math	Displayed Math
report	0	3	0	0	0	0
./envs/argument-environment	118	0	0	0	0	0
abstract	39	0	0	0	0	0
preface	171	0	0	0	0	0
introduction	2406	1	0	0	0	0
literature	1471	5	0	0	0	0
core/1-data	450	9	0	0	0	0
core/2-representation	0	1	0	0	0	0
core/3-similarity	0	2	0	0	0	0
core/4-timeseries	0	3	0	0	0	0
core/5-experimentation	0	1	0	0	0	0
core/6-discussion	0	3	0	0	0	0
critical-evaluation	134	2	0	0	0	0
conclusion	16	1	0	0	0	0
appendices	21	13	0	0	0	0
./figures/argument	110	0	0	0	0	0
./figures/architecture	0	0	0	1	0	0

Table 10.3: Word count by file (`texcount`)