


Monday, December 2, 2024

Access the code, data, and analysis at <https://github.com/j-jayes/who-is-who-etl>

Praise the people or praise the place?*

Upper tail human capital in electrifying Sweden

Jonathan Jayes 

Lund University Economic History Department

jonathan.jayes@ekh.lu.se

ABSTRACT This paper investigates the origins and career trajectories of high-skilled workers in electricity-related occupations in Sweden during the mid-20th century. By ingesting and analyzing two unique data sources – a set of biographical dictionaries and an industrial catalogue – I ask where these workers came from, what and where they studied, and how international experience impacted their career paths.

Jayes, Enflo and Molinder (2024) find that medium-skilled electricity-related jobs were filled by workers who were born near to these jobs. However, descriptive statistics about these high-skilled workers reveal a bifurcated labor market. Highly educated and skilled professionals, crucial in overseeing and advancing the electricity sector, frequently migrated for education and career opportunities, often traveling significant distances from their birthplaces to a small set of higher education institutions in Sweden and abroad. After their studies, Swedish engineers moved to opportunity, ending up on average ten times further from their birthplace than the median worker in 1930. Tentative findings on international experience show that engineers worked abroad at higher rates than comparable occupations, far more than doctors or dentists.

Introduction

Economic history as a discipline is set to benefit in at least three ways from the nascent AI revolution. Novel sources of data become available as new tools

*Thank you to seminar participants at the Copenhagen Business School Department of Strategy and Innovation PhD Seminar and the HEDG Group at the University of Southern Denmark for valuable feedback on this work.

make them readable to machines, analysis in new ways is possible with new kinds of algorithms, and the time-cost of asking certain kinds of questions decreases, opening up new avenues for research. In this paper, I make an attempt to leverage these benefits to answer the question, ‘who were the high skilled workers in electricity related occupations in Sweden, and where did they come from?’ I detail the process of structuring and analyzing a novel data source; a set of biographical dictionaries in order to answer this question.

In the quest to understand the dynamics of economic development and technological advancement, previous research by this author and his supervisors shed light on the transformative impact of early electricity access in Sweden. “Power for progress: The impact of electricity on individual labor market outcomes” (javes_power_2024) revealed how the advent of electricity in certain parishes led to positive economic outcomes: a boost in income levels, reduced inequality, and the maintenance of employment levels despite the advent of labor-saving technology. There were two findings that stood out and which I seek to explore further here. First was the tendency of workers in these early electrified parishes to remain in their birthplaces, hinting at a newfound economic dynamism stemming from the income spillovers into sectors not affected by the new technology. Second, we found that electrification allowed those without a huge amount of formal education to take on new jobs that emerged in the wake of the new technology, leading to a more equal distribution of income in these areas.

Building on these insights, the present paper looks deeper into the human aspect of this technological revolution. It poses a question: Who were the key figures driving this change? Were they local talents, or did they represent a wave of skilled individuals drawn from afar, drawn in by the pioneering spirit of these early electrified areas?

To answer this, the investigation leverages a novel data source. *Vem är Vem*, is a comprehensive set of biographical dictionaries containing the profiles of 75,000 notable Swedes active between 1945 and 1968.

I digitize and structure this source in order to analyze the changing patterns of the Swedish labour market in the middle of the 20th century in light of electrification. The findings challenge my prior expectations. Contrary to the belief that local talent pools predominantly fueled the technological boom, I observe a pattern of geographical mobility among the highly educated and skilled professionals in electricity-related fields. These individuals, pivotal to overseeing and advancing the electricity sector, often sought education and opportunities far from their origins. This suggests a bifurcated labor market: local talent predominantly filled the burgeoning middle-skilled roles within the electricity sector, while the top-tier skilled professionals were more transient, moving towards educational and occupational opportunities. This paper explores the implications of this labor market structure for the economic development patterns witnessed during Sweden’s second Industrial Revolution.

These findings, tentative as they are, have real world value. As we seek to understand what drove the dynamism during the age of electrification in Sweden, we are better equipped to shape policy today that seeks to revitalize deindustrializing areas across the developed world and help the developing world harness new

technologies for sustainable growth. In addition, the methodologies employed to structure and analyze archival data can provide a template for future research using similar materials.

The paper is laid out as follows: the current research question is placed in context, the source is explained, followed by the digitization and structuring process. I then discuss the classification task involved in assigning engineers to a sector based on their occupational trajectories. I then lay out some descriptive statistics and tentative findings regarding the patterns of movement for the high skilled electricity related workers, compared with other professionals I observe in the biographical dictionaries.

Context

This paper ties into three strands of literature; two in content and one in methodology. The first strand focusses on the use of individual level biographic data in economic history, at scale. The second strand focusses on the importance of human capital in economic development. The third strand focusses on the use of new tools to structure and analyze historical data.

The use of individual level biographic data in economic history at scale

Several recent papers have made use of biographic data in innovative ways. **ford2023not<empty citation>** use biographies of high school graduates compiled at the time of their school reunions to create a far richer measure of human capital than the conventional measure, number of years of schooling, alone. Titled, “Not the Best Fillers in of Forms? The Danish and Norwegian Graduate Biographies and ‘Upper Tail Knowledge’”, the authors explain that these biographies are “mini-CVs”, containing information about the school leaver’s grades, their occupational trajectory, and their family background. These are used to create an innovative approach to measuring upper tail knowledge.

Nekoei2020Herstory<empty citation> titled “Herstory: the rise of self-made women” analyzes the historical prominence of self-made women using a specially created database. This database, formed by applying machine learning to Wikidata and Wikipedia, catalogues notable individuals throughout history, highlighting details like occupation and family background. Their unique approach reveals a significant increase in the number of prominent women, especially those who achieved success independently of their family connections, across various fields, starting with literature, since the 17th century. This research provides a fresh perspective on women’s historical achievements and roles.

Importantly, these papers go beyond the use of just administrative records, which contain register-like data that economic historians are familiar with. Leveraging new kinds of sources in this way allows the authors to approach their research with different kinds of questions.

In this paper, I structure biographic data about elite individuals that is similar in structure to **ford2023not<empty citation>**, and use the career trajectories of these individuals to better understand the contribution of educated workers to the adoption of electricity across Sweden in the 20th century. The differentiator

in this case is the scale - I capture 75,000 individuals in Sweden in an automated manner, or about one percent of the population of the country at the time.

The importance of human capital in economic development

The question of where the high skilled workers in electricity related occupations in Sweden came from is important in order to understand the economic dynamism of that era. As such, it ties into a wealth of research on technological change and the labour market, which I review briefly here.

The historical adaptability of labor markets to technological change is well-documented. In their study of the U.S. labor market's response to the automation of telephone operation, **Feigenbaum2020**<empty citation> demonstrate how technological displacement in one sector led to increased demand in others, suggesting an inherent resilience in labor markets. This finding is particularly pertinent to this exploration of Sweden's electrification, as it indicates a potential for both displacement and opportunity in the face of technological change.

Claudia Goldin's extensive analysis of labor markets in the 20th century provides a comprehensive backdrop to this study (**Goldin1994; Goldin1998**). Her work highlights critical shifts in labor participation, wage structures, and job security, reflecting the complex interplay between societal changes and labor market dynamics. These insights are crucial for understanding how shifts in human capital, like those during Sweden's electrification period, contribute to broader economic outcomes.

The impact of the Digital Revolution on labor markets, as reviewed in the Oxford Review of Economic Policy, is also salient to our study (**Adams2018; Goos2018**). These articles underscore the emergence of job polarization and the crucial role of policy interventions in ensuring equitable benefits from technological advancements. This perspective is instrumental in understanding the differential impacts of electrification in Sweden, especially in terms of job creation and labor market segmentation.

Moretti's exploration of the geographical clustering of talent and innovation in "The New Geography of Jobs" provides a crucial perspective on the spatial dynamics of economic development (**Moretti2012**). His findings about the importance of local ecosystems in fostering innovation and economic vitality resonate with our investigation of how early electrification in Swedish parishes influenced the distribution and impact of skilled labor. His concern, that gains to productivity are eaten up by increased cost of living (primarily through housing costs) when constraints prevail, is not evidenced in the first half of the 20th century in Sweden. However, his example of Silicon Valley – where high productivity and attractive jobs draw in people with high levels of skill, raising property prices - is becoming more concerning in today's relatively housing scarce urban centers.

New technologies require new skills. Mokyr's research provides insights into the importance of both artisans and engineers in the progression of the Industrial Revolution. His studies underscore the synergistic relationship between theoretical knowledge and practical expertise, essential in driving technological innovation and economic progress (**Mokyr2017**). In his examination of the socio-economic elites of early modern Europe, Mokyr explains how their education and

exposure to new ideas and sciences were pivotal in fostering various intellectual and technological advancements. This educated elite, through their changing culture and institutions, played a crucial role in creating an environment conducive to innovation (**Mokyr2017Journal**).

Not every innovator needs higher education. Mokyr’s perspective is crucial in understanding the dynamics of technological development and economic growth, emphasizing the collaborative efforts between well-educated scientists and highly skilled artisans. This interplay highlights the importance of practical skills, theoretical knowledge, and their combined impact on technological progress. For example, figures like metallurgist Henry Cort, who collaborated with scientists despite lacking formal scientific training, exemplify the productive synergy between different forms of expertise in this era (**Mokyr2017Journal**).

In this paper, I want to find out where the individuals came from who enabled the technological development that was associated with Sweden becoming richer and more equal. Did they come from the areas around where the technology was developed / adopted, learning skills on the job? Or did they get formal education at one of Sweden’s universities and then bring these skills to the hubs of technology? Should we praise the people, or the place?

The use of new tools to structure and analyze historical data

The third strand of literature that this paper ties into is the use of new tools to structure and analyze historical data. Within this strand, there are perhaps two main use cases; problems of prediction and classification, and ‘big data’ gathering and analysis.

Relating to the former, **mullainathan2017machine**<empty citation> wrote a pathbreaking article that documents the use of machine learning as part of an economists toolkit in the context of prediction problems, differentiating it from traditional parameter estimation. The authors explain the use of new data types like satellite images and text, as well as machine learning’s role in policy, estimation, and testing economic theories.

Some interesting papers that incorporate prediction and classification problems include **Bandiera2020ceo**<empty citation>, who classify CEO behavior by collecting high-frequency diary information and then use a machine learning algorithm to classify CEOs into ‘leaders’ and ‘managers’ by the content of their meetings. **koschnick2023breaking**<empty citation> uses a machine learning topic model to classify each paper by the universe of all students at English universities in the seventeenth and early eighteenth century to calculate a measure of how innovative the paper was; how it differed from the papers before it in the field and how similar the papers afterwards were. **DahlVedel2024**<empty citation> in a paper titled “Breaking the HISCO Barrier: AI and Occupational Data Standardization” apply a neural network to the task of classifying an occupational description and benchmark their results against human labelling to show that the neural network achieved comparable accuracy with human labelling and involved an order of magnitude fewer human hours. I make use of this tool in the classification of occupations in this paper.

‘Big data’ papers in economic history now abound, as surveyed in **gutmann2018big**<empty citation> in a review titled ‘*Big Data*’ in *Economic History*. Many of these papers construct and use high quality register or census-like data on individual outcomes. Notable examples include the Longitudinal, Intergenerational Family Electronic Micro-Database Project by Martha Bailey which focuses on family histories to understand long-term economic trends. These histories are collected from various census-like sources and innovative ML tools are used to construct the longitudinal links. Similarly, “The Making of Modern America: Migratory Flows in the Age of Mass Migration” by **bandiera2013making**<empty citation> involved the digitization of 24 million records of migrant flows through Ellis Island in New York, and found that measured out-migration rates in the US were double the reported figures in the earliest decades of the 20th century. **eichengreen2021gold**<empty citation> uses natural language processing (NLP) to understand the content of a parliamentary committee debate on the gold standard in South Africa. Clark and Cummins’ families of England database contains 1.7m marriage records in England from the 19th to the 21st centuries and allows to authors to pry out social dynamics in family formation, as well as geographic sorting between the North and South of England (**Clark2018BigSort**).

There is also a growing literature in which authors lay out the step by step processes required for economic historians to make use of these new tools. A great paper in this vein that bridges the gap between cutting-edge computer science literature and the use cases of applied economists is by **correia2023digitizing**<empty citation>. The paper “Digitizing Historical Balance Sheet Data: A Practitioner’s Guide” explores the application of machine learning, particularly Optical Character Recognition (OCR), to digitize large-scale historical economic data. The authors highlight the limitations of off-the-shelf OCR software, mainly due to high error rates, and propose a combination of pre- and post-processing methods to enhance accuracy. They apply these methods to two extensive datasets of balance sheets and introduce “quipucamayoc,” a Python package that unifies these techniques.

amujala2023digitization<empty citation>, in “Digitization and data frames for card index records” explain the entire process through which they digitize and structure loan records from bank cards that contain both machine written and hand written text in varying formats over time. The authors lower the barrier to entry for other researchers by explaining their use off-the-shelf technology from Amazon Web Services. Each step is explained along with tips for successful extraction of handwritten information.

Perhaps the most impressive of these kinds of papers is the *Layout Parser* from Melissa Dell and her lab. The team have produced a python library that can be fine tuned to parse document type, extract information, and use machine learning to correct common errors at the point of data extraction. Dell demonstrates the use of this tool by extracting firm performance data from Japanese reports on yearly firm output which are atypical from the kinds of tables or column

text that off-the-shelf optical character recognition tools have been trained on (dell2020large).

I hope that this paper can be useful for other researchers using similar kinds of sources as a prompt on where to start collecting and structuring their data at scale.

Materials and Methods

Biographical dictionaries

Vem är Vem? is a biographical dictionary, comprising a rich repository of information about notable individuals in Sweden. Published in two regional editions with a total of five volumes each, the first edition spanned from 1945 to 1950, and the second from 1962 to 1968, by the Bokförlaget Vem är Vem publishing house (harnesk1945vem). An additional volume specifically focussed on individuals in industry and business was produced in 1945. This encyclopedia offers an invaluable snapshot of Swedish societal and professional landscapes during these pivotal periods.

The primary intention behind the creation of *Vem är Vem?* was to spotlight individuals who were at the peak of their careers, regardless of their age. This focus extends beyond traditional measures of influence, emphasizing the importance of those in influential positions or notable roles across relatively diverse sectors. As such, it serves as a crucial resource for understanding the professional and personal trajectories of around 75,000 individuals who shaped Swedish society in the mid-20th century.

It is worth noting that the criteria for inclusion was somewhat vague, and individuals could opt in to being included for a nominal fee. As a result, there are some individuals for whom not much information is included beyond biographic information, current location and profession. For others, there is a rich tapestry about their lives including records of career progression, business travel, technical writings and membership of civic organizations. The source does not capture a representative picture of Swedish society at the time, but rather those individuals with some level of social cachet or prestige, and a desire to be recorded in the biographical dictionaries as such.

Vem är Vem? is useful to economic historians thanks to its high quality digitization, with nine out of the 11 total volumes being made accessible online by librarians in Uppsala through *Projekt Runeberg*, as shown in Figure 1. This digitization has facilitated research, allowing for a broader exploration of the biographies and career paths of thousands of individuals. The encyclopedia's extensive coverage makes it a goldmine for researchers, historians, and anyone interested in the socio-economic history of Sweden during a period marked by significant change and development.

In the context of economic and historical research, *Vem är Vem?* serves as a unique tool. By providing detailed biographies and career information, it allows for an in-depth analysis of the human capital that contributed to Sweden's economic and social evolution during the mid-20th century.

Number of biographies in Vem är Vem?

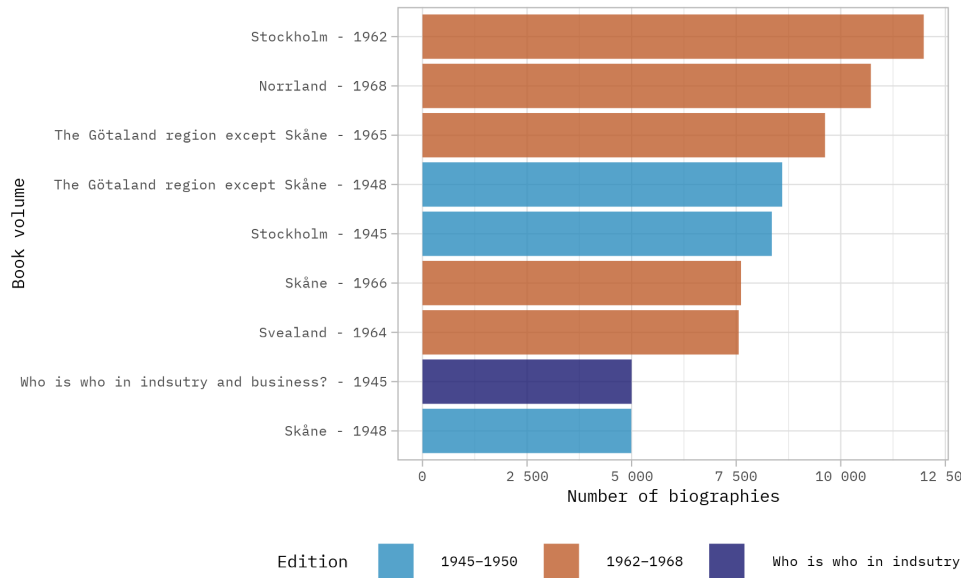


Figure 1: Number of biographies in each volume of 'Vem är Vem?'

The biographic information about the individuals in the dictionaries are exemplified in Figure 2, which highlights the life of chemist and metallurgist Karl Gustaf Lund.

The fields include:

1. **Education:** Lund's education at prestigious institutions such as the Royal Institute of Technology (KTH) indicates he had access to advanced technical knowledge. This level of education is critical for understanding the specialized skills that were necessary for innovation and advancement in electricity-related industries.
2. **Career Progression:** The text outlines Lund's career progression through various roles in metallurgy and chemical engineering. This trajectory can illustrate how individuals applied their education in practice, contributing to industrial development. Tracking such careers can provide insight into the professional development paths that were common and valued in the sector at the time.
3. **International Experience:** His experiences in the United States reflect the cross-border exchange of knowledge and skills. This can show how international experiences contributed to the domestic industry by importing new ideas and practices, which is a key aspect of human capital development.

fl. 14-15, i Trollhättan 15-16, Nässjö 21-32, Majornas komm. flicksk. i Gbg sed. 37, i Nässjö bl. a. led av barnavårdsn. 29-32, kyrkofullm. 31-32 samt ordf. i RK-krets 30-32. Sekr. i styr. f. Gbg o. Boh. landstings yrkessk. 36-45, suppl. i hälsövrndn i Gbg sed. 40, ordf. i Nässjö husm:fören. 23-32, Smål. husm:förh. 28-32, Gbg's husm:fören. 41-45 samt Gbg's o. Boh. l. husm:förb. sed. 41, led. av Sv. husm:fören. riksförb. centr:styr. sed. 29.

Lund, Karl Gustaf, överingenjör, Varberg, f. 22/7/93 i Hille, Gävle. l. av brukstj. m. Ferdinand L. o. Maria Andersson. G. 36 m. Sigrd Johansson. Barn: Ingvar f. 38, Lennart 42. — Ex. v. bergssk. i Filipstad 17, spec:stud. v. KTH (B) 20-22, stud. v. metallogr. inst. o. Sthlms högsk. 21-22. Kemist v. Strömsnäs Järnverks A-B, Degerfors, 18-20, metallurg o. kemist v. Westinghouse Electric & Manuf. Co., East Pittsburgh, Pa, USA, 23-26 o. 28-29, chefsmetallurg v. Laclede Steel Co., Alton, Ill., USA, 27, hytt- o. stålving. v. A-B Iggesunds Bruk 29-31, platschef v. Gunnebo Bruks Nya A-B, Varbergsvärket, sed. 31. Led. av drätselkamm. v. ordf. v. ekonom. avd., suppl. i styr. f. elverket, huvman i Varbergs Sparbank, arb.giv. repr. i Binsarb:ndns kretsrad, led. av styr. f. Varbergs luftsk:fören, sek. i Varbergs högerfören, ordf. i järnv. sjukvård o. Plant:sällsk. Småfägl. Vänner. Res. i Tyskt. 21, 22, 23, 30 o. 36, Danm., Tjeckoslov. 21, 22, 23, Österr. 21, USA 23-29. Skr.: Some fundamental factors for obtaining sharp thermal curves (Trans. Am. Soc. for Steel Treating, tills. m. C. Benedicks o. W. H. Dearden 25). Njutida fabrikation av sågblad, sågklingsor o. maskinknivar (Trävaruind. 31). Hobbies: jakt o. fiske.

Lund, Erik Gustaf Viktor, feruand, tandläkare, Göteborg, f. 1/8/96 i Tolg, Kronob. l. av Fredrik L. o. Maria Johansson. G. 27 m. Hilbur Nordenström. Barn: Lennart f. 28, Ingemar 29. — Stud:ex. v. Lunds priv. elem.sk. 17, tandl:skand. 20, tandl. 22. Prakt. i Klippan 22-23, i Gbg sed. 24. Skattnäst. i Gbg's tandl:sällsk. 35.

Lundh, Lars Åke, redaktör, Göteborg, f. 14/9/09 i Gbg av Otto L. o. Maria Malmberg. G. 41 m. Barbro Nordström. Barn: Lars f. 44, Christi-

na 46. — Stud. v. Gbg's latinlärov. Medarb. i Gbg's-Posten sed. 29. Gjort reportage i Norge, Danm., Lettl., Polen, Tjeckoslov., Tyskt., Frankr., Engl., Ital., Schweiz o. Amer., krigskorresp. i Polen 39. Ordf. i folkpart. ungdom:förb. i Gbg 39-43, styr:led. i folkpart. ungdom:förb. m. fl. org. inom part. styr:led. i Flyggjournalisternas klubb, Gbg's-Postens guldplak. f. journ:bragd.

Lundahl, Carl-Gustaf Allan, prakt. läkare, Göteborg, f. 13/3/06 i Borås av läbr. Carl L. o. Anna Jacobsson. G. 40 m. Marguerithe Giescke. Barn: Hans f. 41. — Stud:ex. i Borås 25, med. kand. i Upps. 30 o. med. lic. där 37. E. o. aman. v. hygien-bakteriolog. inst. i Upps. 32-33, h. prov:läk. i Kina o. Vårgårda distr. kort. tider 37, bitr. läk. v. Hultafors sanat. 37, prakt. läk. i Gbg sed. 39.

Lundahl, Ernst Fritiof, stadsfiskal, Vimmerby, f. 13/11/88 i Sönnarslöv, Krist. l. — Lansm:ex. 10. Aust. v. landstaten 06-17, landskont. 17-18, stadsfiskal o. stadsfogde i Vimmerby sed. 18. Ordf. i styr. f. Skand. Bankens avl:kont. i Vimmerby o. i styr. f. Vimmerby Sparbank, köpmannafören. ombud.

Lundahl, Harry Sigurd, redaktör, Göteborg, f. 16/10/05 i Helsingborg av Herman o. Agda L. G. 35 m. Britta Linnea Davidson. Barn: Ulf f. 36. — Stud:ex. i Helsingborg 25, stud. v. handelsgymn. där 27-28. Medarb. i Helsingborgs-Posten 28-31, Eskilstuna-Kur. 31-35, Arbetet i Malmö 35-45, Gbg's Handelsstidn. sed. 45. På sin tid framgångsrik fotb:spelar, landslags-spelare, medl. av Helsingborgs IF, IFK Eskilstuna, Malmö FF o. Bf, led. av Sv. fotb:förb. uttagn:komm. 37-39 o. 40. Resor i Schweiz o. Holl. 27, Engl. o. Ung. 28, Engl. 29 o. 39, Tyskt., Ital. o. Monaco 31, Polen o. Rumänien 37, Tjeckoslovakien 38, Engl. 39. Skr.: Fotboll-Juli (28), Engelsk il-gakalender (30). Hobby: idrott av skilda slag. Sv. fotb:förb. spelare. o. dess tekn. komm. diplom o. M. Skånes fotb:förb. fjtjG, Sverm. fotb:förb. hedersm., Helsingborgs IF hedersM o. stora fjtjS.

Lundahl, Hasse, ingenjör, Eksjö, f. 29/9/99 i Eksjö. — Stud:ex. 20, ing:ex. 23. Chef f. Eksjö stads vatten-o. elverk sed. 31. Medl. av Eksjö fabriks- o. hantv:fören. samt Ödd Fellow.

Figure 2: A representative page of Vem är Vem?, highlighting the biography of Karl Gustaf Lund

4. **Leadership and Management:** Lund's leadership positions, such as chairmanships and advisory roles, imply a combination of technical expertise and managerial acumen. The ability to lead and innovate within companies is a significant aspect of human capital that drives industry growth.
5. **Research and Innovation:** The reference to his translated research work indicates an engagement with cutting-edge technology and knowledge creation. Such contributions are the tangible outputs of human capital in action, pushing the industry forward through innovation.
6. **Professional Networks:** His involvement with societies and associations suggests a networked professional community, which is essential for the diffusion of innovative ideas and practices. These networks are often where knowledge is exchanged, partnerships are formed, and collaborations are initiated.

I use this biographical data to determine who the high skilled workers in electricity related jobs in Sweden were, and understand more about their career trajectories.

Source criticism

In order to assess the representativeness of the biographical dictionaries, I sample a random selection of individuals from the dictionaries and manually check if they appear in the Swedish Biographical Lexicon (SBL).

SBL is a personal history lexicon that began to be published in 1917 and which now (spring 2024) includes A-Södersten. The dictionary is a general inventory based on first-hand sources of important people and their deeds during various stages of the development of Swedish society (from 1500 onwards). There are a total of approximately 27,200 pages on individuals. The data included in the SBL is more comprehensive than that in the biographical dictionaries, as it is based on a wider range of sources and includes more individuals. Crucially, it is not self-selected. This means that the SBL captures individuals regarded to be of importance by the editors, rather than those who have paid to be included in the biographical dictionaries.

I sampled 200 individuals from the biographical dictionaries and found that 19% of them were also in the SBL. This suggests that the biographical dictionaries are a more representative sample of Sweden at the time. The 75,000 individuals in the biographical dictionaries represent about one percent of the Swedish population in the middle of the 20th century.

Data collection strategy

In order to analyze the biographical dictionaries, we need to bend the text into a machine readable structure. This process is not complicated, but somewhat involved. It includes breaking each component of the source up (e.g. each biography), extracting the pertinent information from each record, storing each value with its associated key, and then saving this information in a way that is easy to analyze and aggregate.

The simplified process is laid out in Figure 3. The underlying code can be found on the GitHub repo linked at on the first page of this paper. I detail the third step, structuring the records, in the section below, and the remainder of the steps in the appendix.






Data Collection Strategy	
For biographical dictionaries and industry catalogue	
Step	Process
1	Scrape book data from website 
2	Split records on each page of a book 
3	Structure records with LLM 
4	Augment data with coordinates 
5	Store data for analysis 

Figure 3: Data collection process steps

Prior to the advent of Large Language Models (LLMs), this structuring of data from free text into key-value pairs was a task that required a large number of human hours to complete. It could be done either by putting the information into an excel sheet by hand, or writing rules to extract the information from the text. The first approach limits the number of observations a researcher can collect on her own, and the second approach quickly turns into the first.

The biographical dictionaries are written in a specific way, with many abbreviations and contracted names for common field titles and values. Due to the number of abbreviations, acronyms and contractions (for example, **Gävle**. 1. is the contraction of *Gävleborg län* or Gävleborg county in Figure 2), while it might be possible to take a simple rules based approach to replacing these contractions with their complete Swedish text, and then looking with regular expressions for specific terms relating to each piece of information, the number of rules soon balloons to an unreasonable figure, making the process unwieldy at best and impossible at worst. Writing a rule for every case necessitates as much human involvement as would be required to manually structure the information - the first approach.

However, with the rapid advancements in LLM technology in the previous five years, and popular adoption of these tools through Chat-GPT and Microsoft’s integration of GPTs into their products in the previous year, new tools mean this manual workload can be avoided to a large extent.

The structuring step involves sending a specific prompt to a LLM, along with the source material as text (rather than a scan of the book), and receiving back from the LLM a structured file with a key and value for each piece of information that I am interested in.

I make use of the computational backend of Chat-GPT, a model called GPT-4o-mini from OpenAI to structure the information from the dictionaries and catalogue into a JSON format that I can analyze, step 3, as shown in Figure 3. Many other LLMs, some open source, are available. I have chosen GPT-4o-mini as it is simple to interact with it in the programming language Python, and because doing so is relatively affordable for contained workloads such as this, when compared to hosting such a model on your own, beefy, costly computer.

By passing the text to the LLM, along with some context about what the model is being given, the model can behave like a skilled research assistant, reading the records, searching for the specific pieces of information requested, and outputting a structured file containing the information that we seek.

Intuitive explanation of LLMs contextual ‘understanding’

The GPT-4o-mini model which I make use of is a LLM which has been trained on all of Wikipedia and Wikidata, among other training material. These two sources contain the same information, but in a different format, as shown in the adapted extracts below; the text and Figure 4 mimic the kind of material that the model I use is trained on.

As the base model underlying GPT-4o-mini is pre-trained to predict the next token on this kind of data, it has developed the ability through repeated exposure to this kind of biographic data to produce structured information from free text, and likewise construct natural sounding text from pieces of structured information. With the addition of *JSON mode* to the model, it is able to output structured information in a reliable manner as detailed in `openai2024text<empty citation>`.

Jonas Wenström (4 August 1855 in Hällefors – 22 December 1893 in Västerås) was a Swedish engineer and inventor, who in 1890 received a Swedish patent on the same three-phase system independently developed by Mikhail Dolivo-Dobrovolsky. He studied at Uppsala University.

Prompting and context

Due to the large amount of material that GPT-4o-mini has been exposed to in training, it is familiar with the kinds of biographical text that I want it to structure. In order to draw on its familiarity with this kind of material, I need to provide it context about the biographical text it is being passed, and ensure that the model returns output in a useful way. I explain these prompts below.

A “system prompt” can be used to tell the model what kind of material it will be passed, and how it should respond. OpenAI suggest that users “ask the model

Key	Value
Name	Jonas Wenström
Birth Date	1855-10-04
Death Date	1893-12-21
Occupations	Engineer, Inventor
Education	Uppsala University

Figure 4: Wikidata information about Swedish inventor Jonas Wenström, inventor of three-phase current

to adopt a persona” in order to improve responses in a specific task in a guide on prompt engineering (**OpenAI2023PromptEngineering**).

I use a simple prompt to tell the model that it will be exposed to Swedish language biographical data:

```
system_prompt = "You are an expert on Swedish biographies."
```

Next I explain the kinds of information that I want it to extract, and the exact format that I want it in. I do this by specifying a JSON schema, with fields that it must return and specifications for the kind of data in each field. An excerpt of the schema is shown in Figure 5.

```
class Bioperson(BaseModel):
    full_name: str = Field(..., description="Full name of the person")
    location: Optional[str] = Field(None, description="Location associated with the person")
    occupation: Occupation = Field(..., description="Occupation details of the person")
    birth_details: BirthDetails = Field(..., description="Details about the person's birth")
    education: Optional[List[EducationItem]] = Field(None, description="List of educational qualifications")
    career: List[CareerItem] = Field(..., description="Career history of the person")
    family: Optional[Family] = Field(None, description="Family details including spouse and children")
    publications: Optional[List[Publication]] = Field(None, description="List of publications")
    community_involvement: Optional[List[CommunityInvolvement]] = Field(None, description="Community roles and involvement")
    board_memberships: Optional[List[BoardMembership]] = Field(None, description="Board memberships held by the person")
    honorary_titles: Optional[List[HonoraryTitle]] = Field(None, description="List of honorary titles received")
    hobbies: Optional[List[Hobby]] = Field(None, description="List of hobbies")
    travels: Optional[List[Travel]] = Field(None, description="Travel details")
    awards: Optional[List[Award]] = Field(None, description="List of awards received")
    leadership_roles: Optional[List[LeadershipRole]] = Field(None, description="List of leadership roles held")
    languages_spoken: Optional[List[Language]] = Field(None, description="Languages spoken by the person")
    military_service: Optional[MilitaryService] = Field(None, description="Military service details")
    honors: Optional[Honors] = Field(None, description="Honors received by the person")
    death_date: Optional[str] = Field(None, description="Date of death")
```

Figure 5: Excerpt of structuring schema

Finally I provide detailed context about the source and instructions for what I want the system to do. I include examples of the abbreviations and contractions that it will encounter, and inform the model as to what kind of output I am expecting in return.

```
prompt = f""" You are an expert on Swedish biographies and will
structure the biographies of individuals from the 20th century
biographical dictionary 'Vem är Vem' that is provided below.
```

```
### Task:
1. Use the schema to organize the information in the biography.
2. Keep the biographic descriptions in Swedish and remove any abbreviations based on your knowledge, e.g. 'fil. kand.' is 'filosofie kandidat', and 'Skarab. l.' is 'Skaraborgs Län' etc.
3. For missing data in a required field, include the field with a None value.
4. Ensure fields are correctly labeled
```

and structured as per the schema. 5. Put years in full based on context. Put dates in DD-MM-YYYY format where possible.

""

Example of structured biographic text

Following this process of structuring the records into a format with specified keys and values, I augment the data by geocoding locations in order to analyze geographic paths of individuals in the sample, and geographic clusters of firms.

Below I show the output of the data collection process, where the biographical dictionary entry on Swedish engineer and power station manager Axel Verner Nordell is shown in Figure 6 and some of the extracted information along with the geocoded coordinates are shown in Figure 7.

Nordell, Axel Verner, civilingenjör, fd. kraftverksdirektör, Motala, f. 15/8/81 i S. Möckleby, Kalmar l., av kyrkoh. Gustaf N. o. Almida Sellergren. G. 11 m. Agnes Hellgren. Barn: Inga f. 12, g. m. civ:ing. P. Rönström, Hans 14, civ:ing., Gösta 18, civ:ing., Ulla 20, g. m. civ:ing. H. Rönström. — Stud:ex. v. Lunds h. a. l. 99, avg:ex. fr. KTH (E) 04. Ritare v. ASEA i Malmö 04-05, ing. v. Elektr. A-B Holmia i Sthlm 05-07, v. Trollhätte kanal- o. vattenverk 07-09, distr:ing. v. stat. vattenf:verk 10-20, tf. chef f. Älvkarleby kraftv., Motalasektionen, 18, f. Motala kraftv. 19-20, kraftv:dir. v. stat. vattenf:verk, Motala kraftv., 20-47, pens. 47, därjämte verkst. dir. f. Motala Ströms Kraft A-B 30-47. Led. av kyrkofullm. sed. 32 o. av kyrkoråd sed. 31, kyrkvärd sed. 40, led. av o. ordf. i styr. f. Östergötl. Ensk. Banks avd:kont. i Motala sed. 22. Led. av Sv. tekn:fören. KVO2kl, RNO.

Figure 6: Raw information about Swedish engineer and power station manager Axel Verner Nordell

Key	Value
full_name	Nordell, Axel Verner
location	Motala, Östergötland
occupation	Civilingenjör, kraftverksdirektör
birth_date	15/08/1881
birth_place	S. Möckleby, Kalmar
birth_parents	Gustaf N. and Amanda Seillergren
birth_latitude	56.35646300000001
birth_longitude	16.420155
education_degree	Studentexamen
education_year	1899
education_institution	Lunds högre allmänna läroverk
education_latitude	55.7046601
education_longitude	13.1910073

Figure 7: Extracted information about Axel Verner Nordell

Clustering and classification of occupations

The next task required classifying each individual in the biographical dictionaries into a particular occupation.

For this task, I leaned on the tools of text embeddings, and a combination of unsupervised machine learning and advanced language processing techniques. I made use of a text embedding model trained on Swedish language text by the **bert-base-swedish-cased**<empty citation>. It is an adaptation of the breakthrough BERT model, introduced by Google Research in 2018 (**google2016bert**). The advantage of this KB Lab model is that it has been trained on a selection of Swedish data, including books, news reports, and internet forums. Hence it is able to score the similarity of Swedish business descriptions and occupational titles.

Text embeddings are effective for clustering because they capture semantic meaning rather than relying on surface-level features like character composition. For example, while “steam engine” and “power station” are different in characters and literal meaning, they are semantically related in the context of industrial machinery and energy production. Text embeddings transform these phrases into numerical vectors that reflect this semantic similarity. When applied to clustering, this means that items with similar meanings, even if their literal expressions differ, are grouped together based on the contextual and conceptual similarities encoded in their embeddings. This capability makes text embeddings particularly powerful for organizing and categorizing text data in a way that aligns with human understanding and interpretation (**jurafsky2023speech**).

Classifying occupations

In **jayes_power_2024**<empty citation>, we grouped occupations into three categories; direct electricity jobs (e.g. electricians), indirect electricity jobs that could benefit from electric motors (e.g. textile workers), and all other jobs. We made this classification based on the occupational title listed in the 1930 census alone. These titles are frequently used in economic history, along with a schema that classifies each title according to a list of possible titles and occupational descriptions. A widely used example is the Historical International Standard Classification of Occupations, or HISCO, defined originally by **vanLeeuwen2002HISCO**<empty citation>. A wealth of mappings have been created that link an occupational string like *civilingenjör* to HISCO code 022, civil engineers, described as:

Workers in this unit group carry out research and advise on civil engineering problems, design projects and structures such as bridges, dams, docks, roads, airports, railways, waste disposal systems, flood control systems and industrial and other large buildings, and plan, organise and supervise their construction, maintenance and repair.

While this classification process used to involve a large amount of human hours, several new methods that use machine learning have emerged to make this process more efficient. **DahlVedel2024**<empty citation> use a neural network to classify occupational strings using a supervised machine learning process based on the occupational titles in several waves of Danish census data. **merouani2023innovation**<empty citation> uses an unsupervised approach based on text embeddings to classify french occupational titles according to the HISCO schema, using a clever compartmentalization strategy to improve classification when individuals hold multiple titles.

In this paper I take a slightly different approach, making use of the additional data from the occupational trajectories of each individual; where they worked and what position they held at each firm or institution. This additional data helps me go beyond just a single occupational string and classify the sector that each engineer worked in. In many instances, the description of the individual in the biographical data relates to their education, rather than what they have done with it.

For example, while it is clear from Axel Nordell’s occupational title, *Civilingenjör, kraftverksdirektör* (civil engineer, power station director), that he worked in a directly electricity related occupation, this is not clear for Gustaf Fredrik Ambjörn, whose title is just *Civilingenjör*, based on the fact that he recieved the degree of *Civilingenjör* from the Royal Institute of Technology (KTH) in Stockholm. When looking at his career trajectory, however, we can see that he began his career working at “Fore River Shipbuilding Co., Quincy, Mass., USA” and ended it working as a “Professor i praktiskt skeppsbyggeri” (professor in practical shipbuilding) at Chalmers Institute of Technology in Gothenburg.

I develop a two step process to classify first an individual’s occupation based on their occupational title and then a sector based on their occupational trajectory. In both cases I use an unsupervised machine learning approach that turns the

titles and work history from text into vector representations of text, and then cluster these vectors in order to assign a HISCO code and then a sector.

To get an intuitive understanding of the clustering process, it is worth showing the relative similarity of a sample of occupations from the HISCO schema. Figure 8 shows a scatterplot with some of the different occupations listed in HISCO.

We can see a cluster in the top right hand corner of the plot showing basketry weavers and brush makers adjacent to sewers and embroiders, and weaving-and knitting machine setters and pattern card preparers. These occupations belong to HISCO major group of production and related workers, and so are all green in the figure. In the bottom left corner we see bacteriologists, medical doctors and veterinary assistants. On the far right of the plot, we see university lecturers and special education teachers. The takeaway is not necessarily where the occupations fall in the plot, but rather that similar occupations are close to one another in the vector space, and further from dissimilar occupations.

To create this plot, I used the **bert-base-swedish-cased** [empty citation](#) library to create text embedding vectors for each occupation and description in the HISCO schema. Each vector has 364 dimensions, representing the semantic content of the occupation title and description. To produce the figure, I have used dimensionality reduction to collapse the vector representation down from 364 dimensions to just two so that I could plot them on the x and y axis, which I have done with the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) library.

In order to classify the occupations from the *Vem är Vem* source according to the HISCO schema, I use this same projection of HISCO occupations, as well as projecting each occupational title into the vector space (in Swedish). I then do a simple cosine distance calculation to find the closest HISCO code to each occupational title in the vector space, and call that the ‘most similar HISCO’ to the occupational string. I experimented with a threshold for the cosine distance, and found that setting it such that roughly 85 percent of the sample received HISCO codes produced results that appeared reasonable. While somewhat arbitrary, this approach allowed for a contextually relevant classification of occupations, drawing on the advanced language understanding capabilities of the KB BERT model. More work to benchmark and score this process is needed in the future.

In the next step, I extract the career trajectories for each engineer. I choose engineers because the other HISCO occupations that make up my sample do not have sufficient information to differentiate their work into sectors. The other largest HISCO occupations in the *Vem är Vem* source are shown in Figure 9, being doctors, university teachers, dentists and ministers.

I group the terms from the career trajectories for each individual, project them into the vector space using the same NLP library as before and compare them to the firm clusters by business type that I explain below.

Findings

With this rich individual level biographic data and information on the clusters of different firms, we can show detailed life courses for the individuals of interest.



Figure 8: Scatter plot showing the relative similarity of occupational titles in the HISCO schema

Most common 3 digit HISCO codes in the Who is Who sample

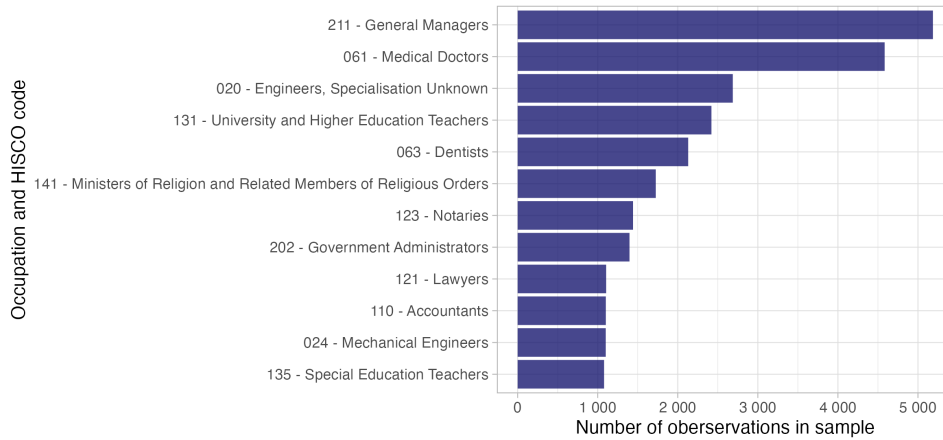


Figure 9: Most common HISCO occupations in Vem är Vem

Figure 10 shows the paths of four individuals who worked in electricity related industries in my sample. They held HISCO code 34, titled “Electrical and Electronics Engineering Technicians” [vanLeeuwen2002](#)[HISCO](#)[<empty citation>](#). We can see that these individuals moved away from home to pursue education, and ended up further from their roots in search of new occupations related to the skills that they had acquired.

This is borne out in Figure 11. This plots the average distance moved by engineers within different sectors¹. Compared to the average distance that a working age adult lived away from their birthplace in the 1930 census of 24km, the average engineer working in the electrical appliance sector in our sample moved 350km from their place of birth, the furthest of all engineers. Even engineers working in the densely geographically concentrated steel and metal industries moved on average 167km away from home.

Figure 12 shows that of all the common HISCO codes in the sample detailed in Figure 9, engineers and mechanical engineers moved the furthest on average. We can also calculate, based on the geographic firm clusters, that of the engineers in the sample, 34 percent live closest to the geographic cluster where the most common firm type is electrical appliances and mechanical machinery.

Conclusion

To sum up, I have demonstrated how using of new tools and algorithms can give economic historians new kinds of source material to work with and new types of questions to ask. I showcased a two-part occupation and sector classification routine. I showed that contrary to my expectations, the individuals with high levels of skill who worked in electricity related industries in Sweden in the middle

¹engineers are all those with HISCO code beginning 2X



Figure 10: Geographic trajectories of selected electrical technicians and engineers

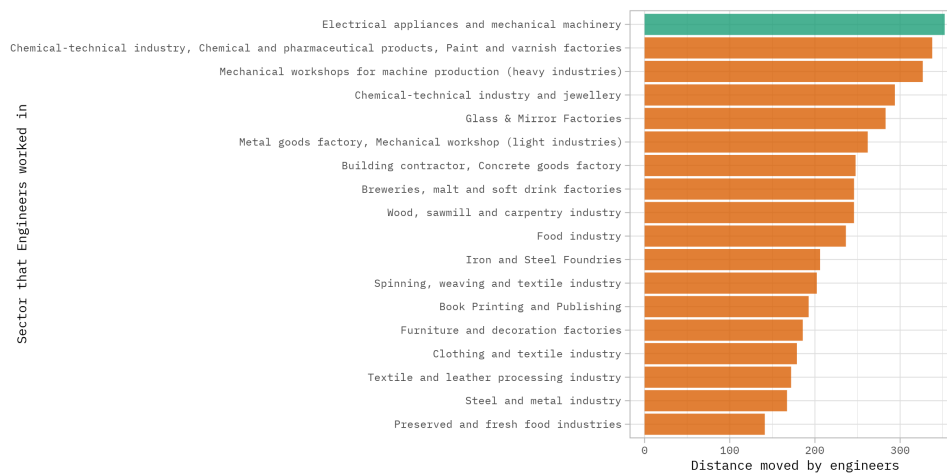


Figure 11: Distances moved by engineers by industrial sector

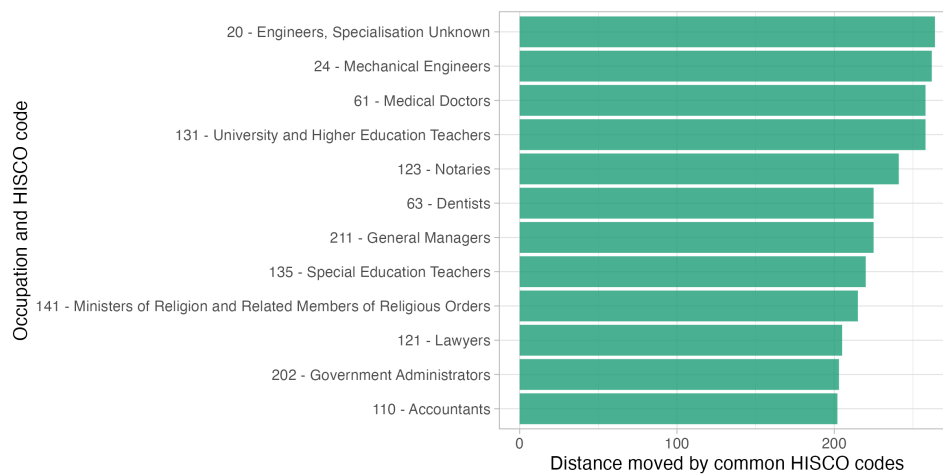


Figure 12: Distances moved by engineers by industrial sector

of the 20th century were not sourced from the localities around where production was centered. Rather, engineers working in electrical appliance and mechanical machinery industries moved the furthest from their places of birth, an average of 350km. This is indicative of a bifurcated labour market, with medium skilled jobs being filled by local workers, and specialized knowledge being brought in by workers with higher education from afar.

Appendix

Figure 3 outlines the data collection process.

I scrape the book content from the *Projekt Runeberg* website with an HTML scraper (beautiful soup in python).

I split the records using regular expressions in python, looking for specific terms that begin and end the records in the dictionaries and catalogue.

I augment the records with coordinates using the Google Maps Geocoding API.

I store the data in JSON format, keeping the original text in the file alongside the derived key value pairs.