Friday, January 12, 2024
Access the code, data, and analysis at https://github.com/j-jayes/who-is-who-scraper and https://github.com/j-jayes/svensk-industrikalender

Paper 4: 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 Abstract

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 make them readable to machines, analysis in new ways is possible with new kinds 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 two novel data sources, a set of biographical dictionaries and an industrial catalogue, 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

*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.

for progress: The impact of electricity on individual labor market outcomes" (Jayes et al., 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. A particularly striking observation was the tendency of workers in these early electrified parishes to remain in their birthplaces, hinting at a newfound economic vibrancy stemming from the income spillovers into sectors not affected by the new technology.

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

To answer this, the investigation leverages two novel and rich data sources. The first, *Vem är Vem*, is a comprehensive set of biographical dictionaries containing the profiles of 75,000 notable Swedes active between 1945 and 1968. The second, the "Svensk Industrikalender" or Swedish Industrial Calendar of 1947, offers an exhaustive catalogue of industrial firms, detailing their activities, workforce, and financial metrics.

I digitize and structure these sources 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 sources are explained, followed by their 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.

Related Literature

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. Ford et al. (2023) 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`.

Nekoei and Sinn (2020) 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 Ford et al. (2023), 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, Feigenbaum and Gross Feigenbaum and Gross (2020) demonstrate how technological displacement in one sector led to increased de-

mand 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 (1994, 1998). 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 (Adams, 2018; Goos, 2018). 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 (Moretti, 2012). 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 though 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 (Mokyr, 2017b). 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 (Mokyr, 2017a).

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 (Mokyr, 2017a).

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, Mullainathan and Spiess (2017) 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 Bandiera et al. (2020), 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. Koschnick (2023) 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. Dahl and Vedel (2024) 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.

'Big data' papers in economic history now abound, as surveyed in Gutmann et al. (2018) 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 Bandiera et al. (2013) 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. Eichengreen (2021) 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 (2018).

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 Correia and Luck (2023). 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.

Amujala et al. (2023), 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. Tha 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 (2020).

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.

Source material

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 (Harnesk, 1945). 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.

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.

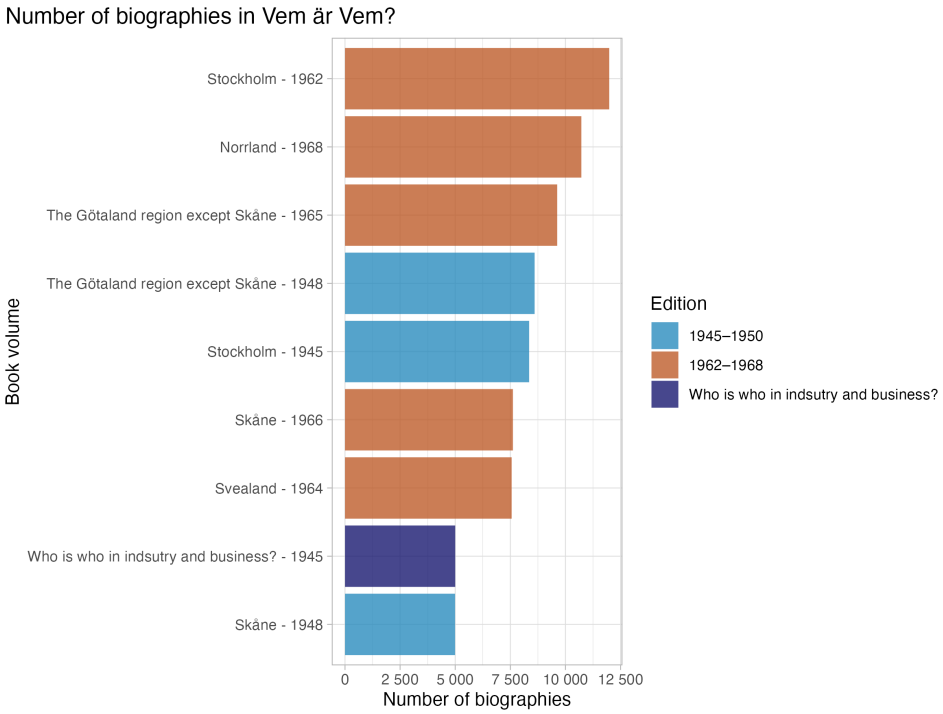
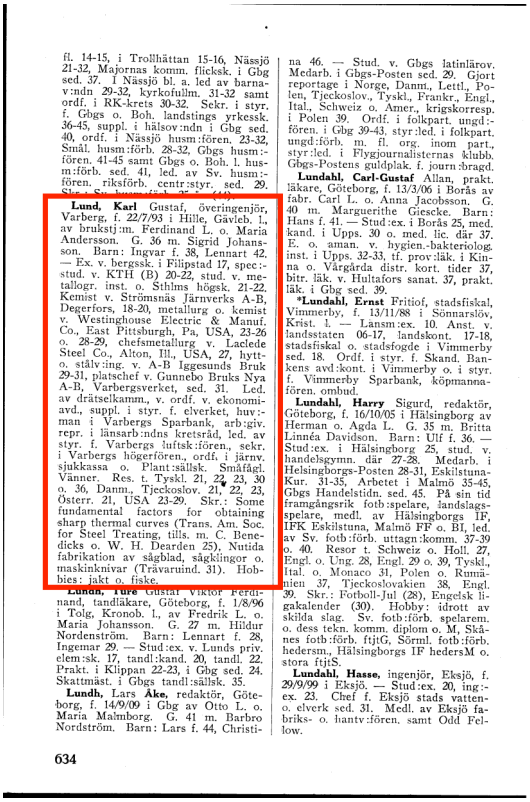


Figure 1: Number of biographies in each volume of 'Vem är Vem?'



fl. 14-15, i Trollhättan 15-16, Nässjö 21-32, Majornas komm. flicksk. i Gbg sed. 37. I Nässjö bl. a. led av barna-v:ndn 29-32, kyrkofullm. 31-32 samt ordf. i RK-krets 30-32. Sekr. i styr. f. Gbgs o. Boh. landstings yrkessk. 36-45, suppl. i hälsov:ndn i Gbg sed. 40, ordf. i Nässjö husm:fören. 23-32, Smål. husm:förb. 28-32, Gbgs husm:-fören. 41-45 samt Gbgs 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, Gävleb. l., av brukstj:m. Ferdinand L. o. Maria Andersson. G. 36 m. Sigrid 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. stålv:ing. v. A-B Iggesunds Bruk 29-31, platschef v. Gunnebo Bruks Nya A-B, Varbergsverket, sed. 31. Led. av drätselkamm., v. ordf. v. ekonomi-avd., suppl. i styr. f. elverket, huv:-man i Varbergs Sparbank, arb:giv. repr. i länsarb:ndns kretsråd, led. av styr. f. Varbergs luftsk:fören., sekr. i Varbergs högerfören., ordf. i järnv. sjukkassa o. Plant:sällsk. Småfågl. Vänner. Res. t. Tyskl. 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), Nutida fabrikation av sågblad, sågklingor o. maskinknivar (Trävaruind. 31). Hobbies: jakt o. fiske.

Lundin, Ture Gustaf Viktor Ferdinand, tandläkare, Göteborg, f. 1/8/96 i Tolg, Kronob. l., av Fredrik L. o. Maria Johansson. G. 27 m. Hildur Nordenström. Barn: Lennart f. 28, Ingemar 29. — Stud:ex. v. Lunds priv. elem:sk. 17, tandl:kand. 20, tandl. 22. Prakt. i Klippan 22-23, i Gbg sed. 24. Skattmäst. i Gbgs tandl:sällsk. 35.

Lundh, Lars **Åke,** redaktör, Göteborg, f. 14/9/09 i Gbg av Otto L. o. Maria Malmborg. G. 41 m. Barbro Nordström. Barn: Lars f. 44, Christina 46. — Stud. v. Gbgs latinlärov. Medarb. i Gbgs-Posten sed. 29. Gjort reportage i Norge, Danm., Lettl., Polen, Tjeckoslov., Tyskl., Frankr., Engl., Ital., Schweiz o. Amer., krigskorresp. i Polen 39. Ordf. i folkpart. ungd:-fören. i Gbg 39-43, styr:led. i folkpart. ungd:förb. m. fl. org. inom part., styr:led. i Flygjournalisternas klubb. Gbgs-Postens guldplak. f. journ:bragd.

Lundahl, Carl-Gustaf Allan, prakt. läkare, Göteborg, f. 13/3/06 i Borås av fabr. Carl L. o. Anna Jacobsson. G. 40 m. Margueritte 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, tf. prov:läk. i Kinna 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. — Länsm:ex. 10. Anst. v. landsstaten 06-17, landskont. 17-18, stadsfiskal o. stadsfogde i Vimmerby sed. 18. Ordf. i styr. f. Skand. Bankens avd:kont. i Vimmerby o. i styr. f. Vimmerby Sparbank, köpmanna-fören. ombud.

Lundahl, Harry Sigurd, redaktör, Göteborg, f. 16/10/05 i Hälsingborg av Herman o. Agda L. G. 35 m. Britta Linnéa Davidson. Barn: Ulf f. 36. — Stud:ex. i Hälsingborg 25, stud. v. handelsgymn. där 27-28. Medarb. i Helsingborgs-Posten 28-31, Eskilstuna-Kur. 31-35, Arbetet i Malmö 35-45, Gbgs Handelstidn. sed. 45. På sin tid framgångsrik fotb:spelare, landslags-spelare, medl. av Hälsingborgs IF, IFK Eskilstuna, Malmö FF o. BI, led. av Sv. fotb:förb. uttagn:komm. 37-39 o. 40. Resor t. Schweiz o. Holl. 27, Engl. o. Ung. 28, Engl. 29 o. 39, Tyskl., Ital. o. Monaco 31, Polen o. Rumänien 37, Tjeckoslovakien 38, Engl. 39. Skr.: Forboll-Jul (28), Engelsk lingakalender (30). Hobby: idrott av skilda slag. Sv. fotb:förb. spelarem. o. dess tekn. komm. diplom o. M, Skånes fotb:förb. ftjtG, Sörml. fotb:förb. hedersm., Hälsingborgs IF hedersM o. stora ftjtS.

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 Odd Fellow.

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

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.

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.

Industrial Catalogue

The *Svensk Industrikalender*, published by Sveriges Industriförbund (Sweden's Industrial Association), is a comprehensive directory of Swedish industrial firms. This calendar was issued annually from 1918 to 2000 and contains information related to Swedish industry. The 1947 edition available on the *Project Runeberg* website was digitized in April 2012, sourced from the Centrum för Näringslivshistoria. The calendar is believed to be under catalog protection but not copyright .

It includes detailed information such as company names, locations, nature of businesses, products, contact details, share capital, number of employees, production values, establishment years, and key personnel including managing directors and board members. This source is valuable for studying the economic and industrial environment of post-war Sweden, providing insights into corporate structures, industry distribution, and business trends of that period.

A representative page is shown in Figure 3.

The common fields listed for each company entry in the catalogue are as follows:

1. **Company Name**: The name of the company is listed at the beginning of each entry, with an asterisk indicating membership of Sveriges Industriförbundet.

2. **Location/Town**: The town or location of the company, which in this case is `Arboga`.

***AB. Arbit,** *Arboga.*

Elektr. sintringsverk och smältverk. Tillverkar: sintrad hårdmetall (»Vivax» o. »Sintram») samt gjuten hårdmetall (»Pansar» o. »Arbit»); dragverktyg av sintrad hårdmetall för dragning av järn och metall, dragskivor, runda och profilerade, fasta samt ställbara för tråd, stänger och rör, dragdornar för runda och profilerade rör, dragdynor för koppning och hylsdragning, diverse andra dragverktyg; slitdetaljer av gjuten hårdmetall, runda och profilerade sandblästermunstycken samt diverse verktyg och maskindetaljer utsatta för förslitning.

Telegr.-adr.: Arbit, Arboga. Telefon: 226. Postgiro: 51462. — Akt.-kap.: 100,000 kr. Tillv.-värde pr år: 500,000 kr.

Bolaget grundat 1930. Dess verk och anläggningar arrenderas av Fagersta Bruks AB. Försäljningen av verkets produkter handhaves av Fagersta Bruks AB., Fagersta, och platskontoret i Arboga.

Styrelse: disp. Hj. Åselius (ordf.), dir. Nils Elfström, dir. Zacheus Olson.

Platschef: ing. Rolf Pauly.

***Arboga Boktryckeri AB.,** *Arboga.*

Utför tidnings-, bok- och accidenstryck, affärstryck, kataloger m. m. Bolaget utger »Arboga Tidning».

Telefon: »Arboga tidning». Postgiro: 28543.

Akt.-kap.: 124.000 kr. Antal industriarb.: 15. Tillv.-värde pr år av tidnings-, bok- och accidenstryck: 290,000 kr. — Firman etabl. på 1850-talet, bolag 1911. Bolaget äges av Eskilstuna-Kurirens Tryckeri AB.

Styrelse: kamrer Stig Holm (ordf.), chefred. J. Anton Selander, red. John Wallström.

Verkst. dir.: John Wallström.

***Arboga Bryggeri AB.,** *Arboga.*

Tillverkar malt- och läskedrycker.

Telegr.-adr.: Bryggeribolaget, Arboga. Telefon: 31. Postgiro: 6361. — Akt.-kap.: 400,000 kr. Antal industriarb.: 20. — Bolaget grundat 1899.

Styrelse: tandläk. Ernst Arosenius, grossh. Ivar Levert, dir. Nils Levert.

Verkst. dir.: Nils Levert.

***AB. Arboga Kvarn & Maltfabrik,** *Arboga.*

Tillverkar vete- och rågmjöl samt pilsnermalt. Varumärke: »Guldsnö»vetemjöl. Firman driver även engroshandel med kraftfoder och gödningsämnen.

Telegr.-adr.: Kvarnmalt, Arboga. Telefon: 238 o. 237. Postgiro: 10078.

Akt.-kap.: 226,250 kr. Antal industriarb.: 7. Prod.-värde pr år: 900,000 kr. — Firman etabl. 1821, bolag 1919, nuv. bolag 1923.

Styrelse: bankdir. Ivar Fredholm (ordf.), dir. Tage Lindblom, dir. Carl-Hugo Peterson.

Verkst. dir.: C.-H. Peterson.

***AB. Arboga Margarinfabrik,** *Arboga.*

Tillverkar växtmargarin, animalisk margarin, kokossmör och konstister.

Telegr.-adr.: Margarinfabrik, Arboga. Telefon: 188. Postgiro: 6352.

Försäljningen sker genom Margarinbolaget AB., Stockholm, Vasag. 16. Telefon: 230960.

Akt.-kap.: 2,500.000 kr. Antal industriarb.: 27. Tillv.-värde pr år: 4,500,000 kr. — Firman etabl. 1888, bolag 1928.

Styrelse: herr Johan Biesert, disp. Anders Göransson, dir. Gillis Husberg, fru Ester Husberg, adv. Ivar Morssing, dir. Gustaf Settergren.

Verkst. dir. o. disp.: Gillis Husberg.

Kontorschef: Anders Göransson.

Driftsing.: Holger Omoe.

***AB. Arbogamaskiner,** *Arboga.*

Mek. verkstad. Tillverkar elektriskt direktdrivna verktygsmaskiner.

Telegr.-adr.: Elektrofabrik, Arboga. Telefon: 20. Postgiro: 77023.

Akt.-kap.: 500,000 kr. Antal industriarb.: 100. Tillv.-värde pr år: 1,800,000 kr. — Bolaget, grundat 1937, har i sig upptagit f. d. Elektriska Fabriken Jonsson & Larsson.

Styrelse: disp. H. de la Cour, fabr. Gunnar E. Jonsson, dir. E. Albin Larsson.

Verkst. dir.: E. Albin Larsson.

Försäljn.-chef: H. de la Cour.

Kamrer: T. Brüde.

***AB. Arboga Mekaniska Verkstad,** *Arboga.*

Mek. verkstad med gjuteri och elektr. vattenkraftstation (Grindberga).

Tillverkar maskiner för järn-, stål- och metallverk samt tråddragerimaskiner, excenterpressar, bockningspressar, friktionsskruvpressar, gradsaxar, kugghjul m. m. Bolaget distribuerar jämväl elektr. energi.

Telegr.-adr. o. telefon: Verkstaden, Arboga. Postgiro: 25175.

Figure 3: A page of the industry catalogue from 1947

3. **Description of Business**: A brief description of the company's main activities or products is provided.

4. **Products or Services Offered**: Specific items or services the company provides, such as types of machinery, tools, or materials.

5. **Contact Information**: This typically includes:

   - **Telegraph Address**: Listed as "Telegr.-adr." indicating the address to which telegraphs are to be sent.
   - **Telephone Number**: Listed as "Telefon" followed by the number.
   - **Postal Code**: Mentioned as "Postgiro" or "Postiro" with corresponding numbers.
   - **Bank Account**: Sometimes a bank account number or similar financial information is included.

6. **Management and Key Personnel**: Names and titles of important figures in the company, such as the director (`Verkst. dir.`), board members, or founders.

7. **Financial Information**: Information about the financial aspect of the company, such as capital invested (`Akt.-kap.`) or turnover (`Tillv.-värde`).

8. **Establishment Details**: This includes the year of establishment and sometimes the history or lineage of the company's ownership or major changes.

9. **Address**: The full postal address, which may include a street name or a postbox number, indicated as `Postgiro`.

This type of catalogue was commonly used for business-to-business interactions and can be considered an early form of networking resource, allowing companies to find suppliers, customers, and partners.

I use the industry calendar to uncover geographic clusters of firms that produced goods related to electricity for sale to consumers or to other businesses. I use these geographic clusters to understand the migration patterns among engineers in Sweden.

Data collection strategy

In order to analyze both the biographical dictionaries and industrial catalogue, 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 or company record), 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 4. The underlying code can be found on the GitHub repo linked at on the first page of this paper. I detail the

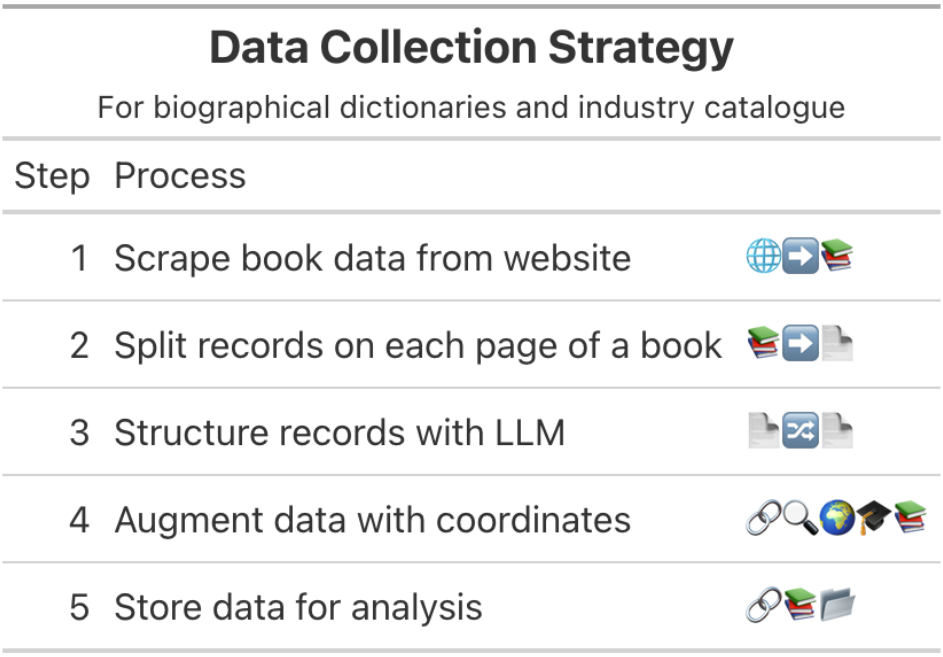


Figure 4: Data collection process steps

third step, structuring the records, in the section below, and the remainder of the steps in the appendix.

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, `Gävleb. l.` 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 backed of Chat-GPT, a model called GPT-3.5-turbo 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 4. Many other LLMs, some open source, are available. I have chosen GPT-3.5-turbo 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-3.5-turbo 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 Table Table 1 mimic the kind of material that the model I use is trained on.

As the base model underlying GPT-3.5-turbo is pre-trained to predict the next token on this kind of data, it has developed the ability though 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 OpenAI (2024).

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.

Table 1: Wikidata information about Swedish inventor Jonas Wenström, inventor of three-phase current

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

Prompting and context

Due to the large amount of material that GPT-3.5-turbo has been exposed to in training, it is familiar with the kinds of biogrpahical 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 to adopt a persona" in order to improve responses in a specific task in a guide on prompt engineering (OpenAI, 2023).

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 except of the schema is shown in Figure 5.

```
schema = {
    "type": "object",
    "required": [
        "full_name",
        "location",
        "occupation",
        "birth_details",
        "education",
        "career",
        "family",
    ],
    "properties": {
        "full_name": {"type": "string"},
        "location": {"type": "string"},
        "occupation": {"type": "string"},
        "birth_details": {
            "type": "object",
            "properties": {
                "date": {"type": "string"},
                "place": {"type": "string"},
                "parents": {"type": "string"},
            },
            "required": ["date", "place"],
```

Figure 5: Excerpt of structuring schema

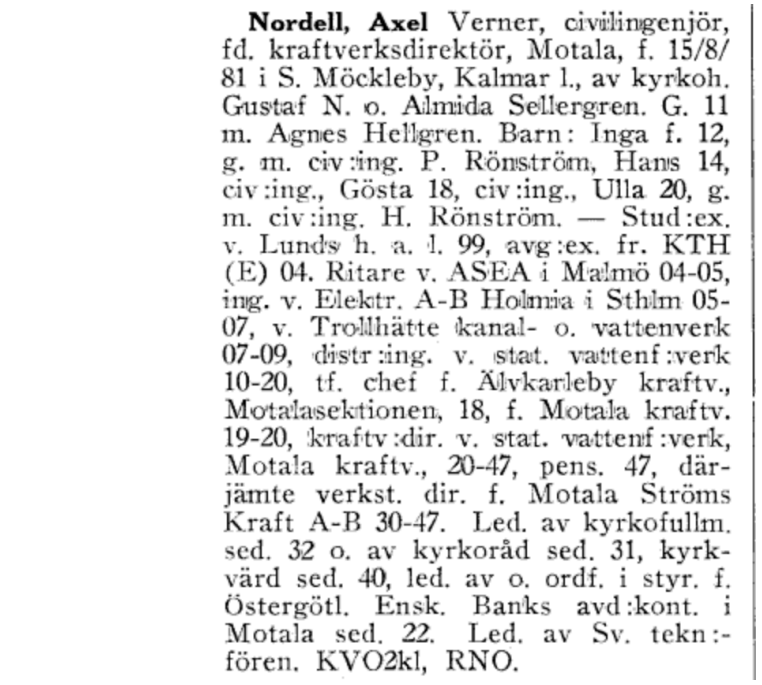
Finally I provide detailed context about the source and instructions for what I want the system to do. I include examples of the abbreviations and contractions that it will encounter, and inform the model as to what kind of output I am expecting in return.

```
  structure_prompt = "Task: read the schema and return RFC compliant
JSON information about the Swedish individuals from the 1950
biographical dictionary 'Vem är Vem' that is provided below. Use
a numeric index for each biography in your JSON output and return
information about all of them, including all career information
available. 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'. Put
years in full based on context. Put dates in dd/mm/yyyy format
where possible. If there is no information for a key, leave it
out. If there is no information for a required key, put NULL
as the value.\nHere is the schema: {schema}.\nHere is the text:
{page_text}"
```

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 analyse 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 Table 2.



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

Table 2: Extracted information about 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

Clustering of firms and classification of occupations

The next task required to understand where these individuals came from and what kinds of firms they were drawn to involved grouping the firms from the catalogue and then classifying each individual in the biographical dictionaries into a particular occupation.

For both of these tasks, I lent 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 National Library of Sweden/KBLab (2024). It is an adaptation of the breakthrough BERT model, introduced by Google Research in 2018 (Devlin et al., 2018). 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 (Jurafsky & Martin, 2023).

Classifying occupations

In Jayes et al. (2024), we grouped occuaptions 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 van Leeuwen et al. (2002). 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.

For grouping firms, I used the k-means clustering algorithm, an unsupervised learning method, to categorize firms into 20 distinct categories. This algorithm works by partitioning the data into k distinct clusters based on features (in this case, text embeddings derived from company descriptions) (Jurafsky & Martin, 2023).

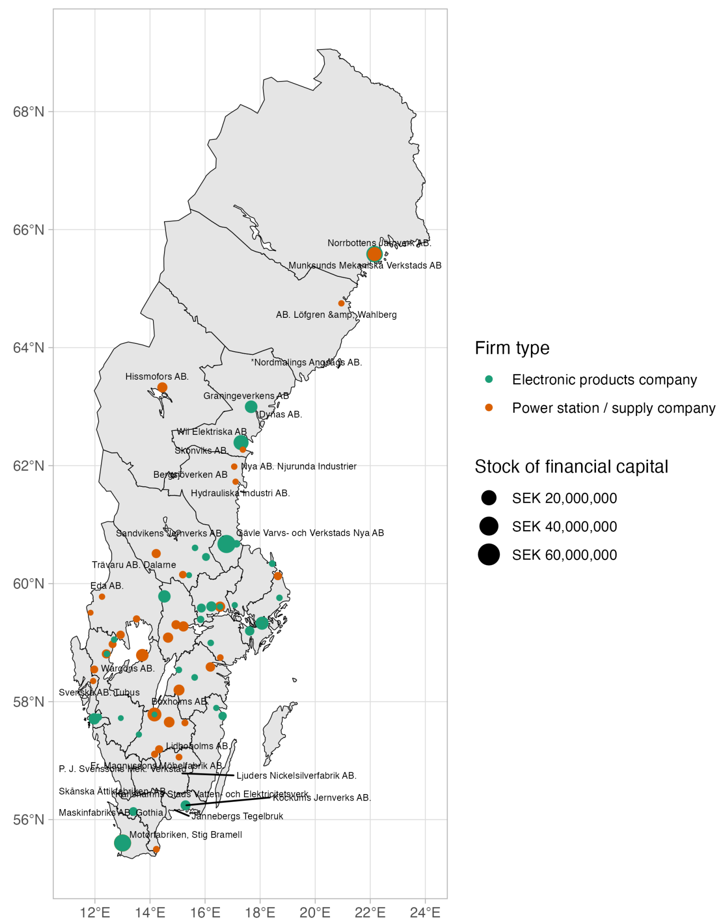
The result of this classification is visible in Figure 7, which shows the number of firms founded each year by category from the industrial catalogue. The clusters are ordered by peak year for foundation. We see that Breweries, malt and soft drink factories are some of the first established industrial businesses, along with book printing and publishing. At the bottom of the figure, it is evident that electrical appliance and mechanical machinery manufacturers, as well as firms in the chemical and pharmaceutical industries have the greatest number of firms founded in the years just prior to 1947, when our industrial firm catalogue is produced.

In classifying occupations, I utilized the KB BERT model to create sentence embeddings for both the titles in the three-digit Historical International Standard Classification of Occupations (HISCO) schema and for each occupation in my biographical dictionary data. These embeddings were then projected into a 364-dimensional vector space. For each occupation, I determined the closest HISCO code in this vector space based on cosine distance, a measure of similarity without a specific unit. I established a threshold for this distance to ensure that the occupation was 'close enough' to the corresponding HISCO code, setting it to a level where 85% of occupations received a HISCO code. 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.

Figure 7: Density plot showing the foundation year of the firms in the 20 categories derived from the business descriptions

What does the industrial calendar tell us about industrial firms in Sweden in 1947?

[Explain that we see forestry and wood business in rural areas and the North of Sweden (purple), we see mechical workshops in modern day Skåne (pink), we see a cluster of furniture factories and home goods factories in central Southern Sweden, and a cluster of electrical appliance and mechanical machinery stores around Stockholm. Note that these clusters don't show all firms of one type, but the location of firms by geographic cluster, and the most common firm in that geographic cluster - makes it easier to see than many dots overlapping.]



[In this map, we show two kinds of firms; firms taht make electronic products, and firms that produce power or supply power stations with equipment. We can see that electronic products companies are clustered near the large population centers of Stockholm, Gothenburg and Malmö. There are many power stations in central Southern Sweden that make use of hydropower, and firms near them that supply equipment for these power stations.]

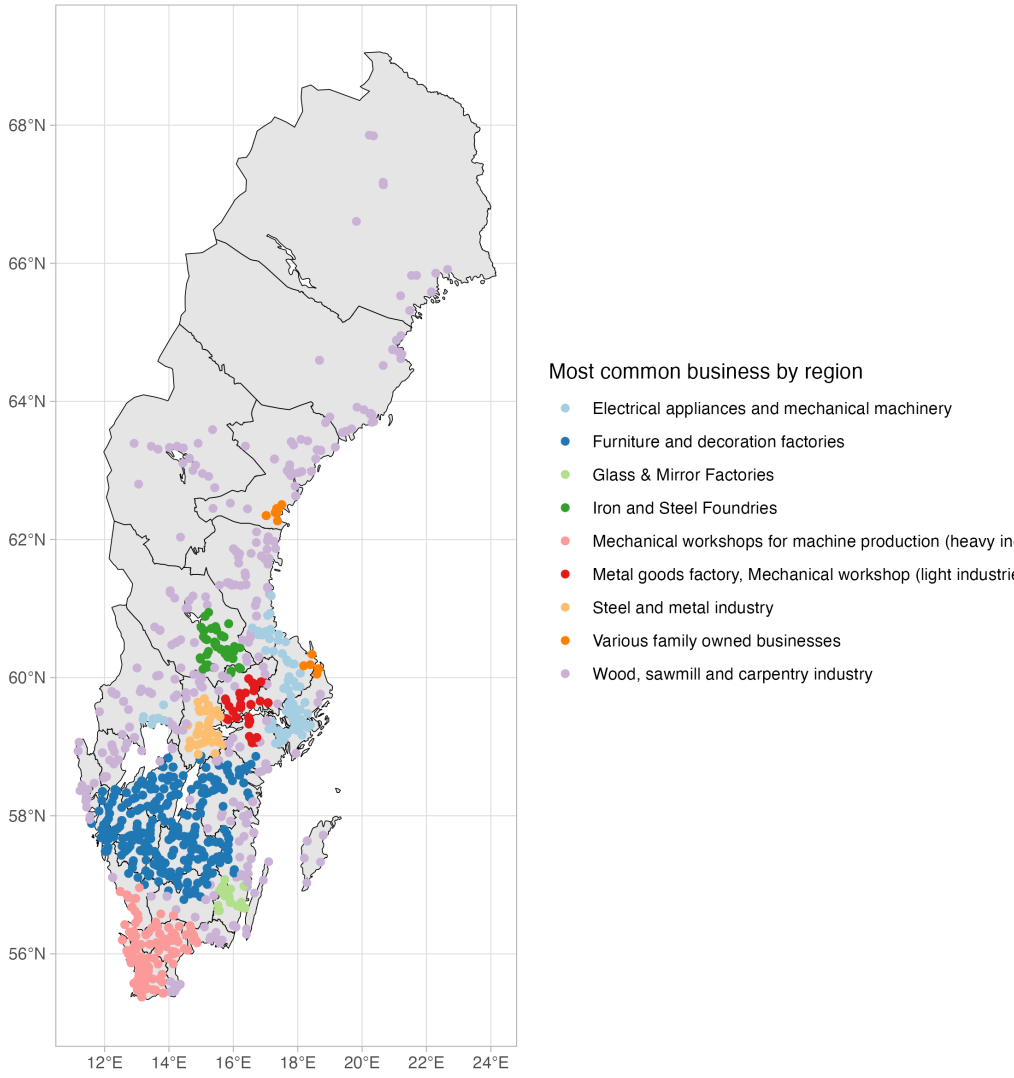
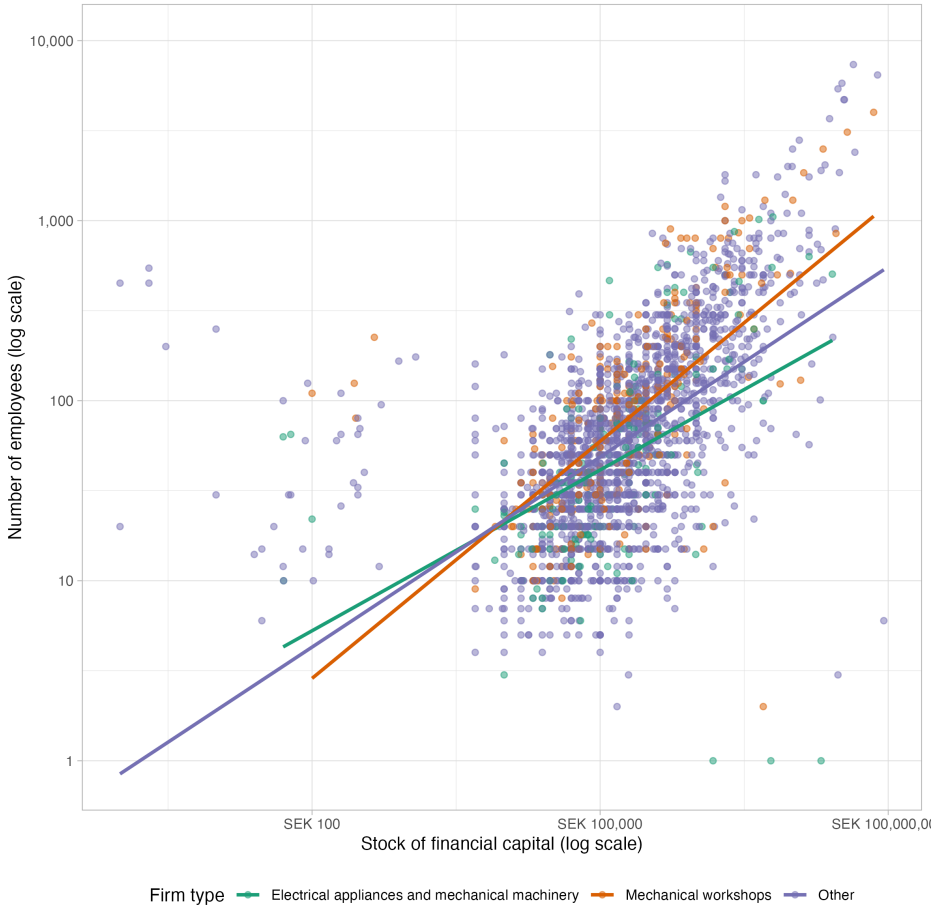
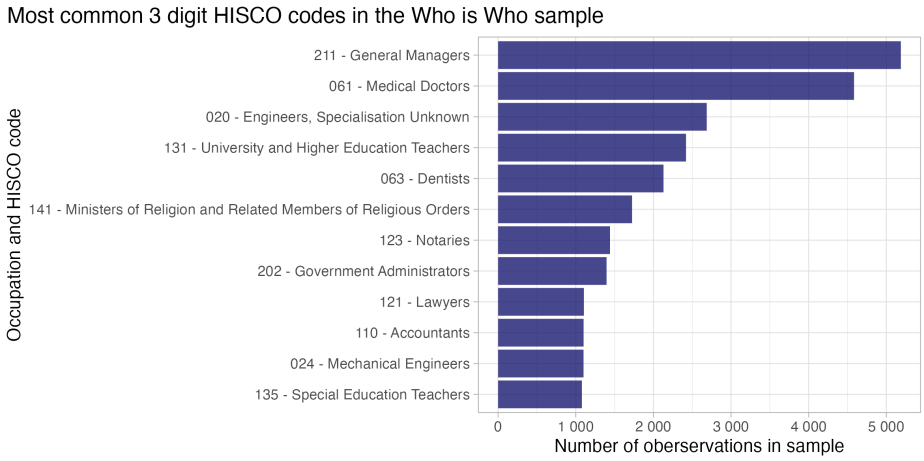


Figure 8: Map of the geographic clusters of businesses by most common business type

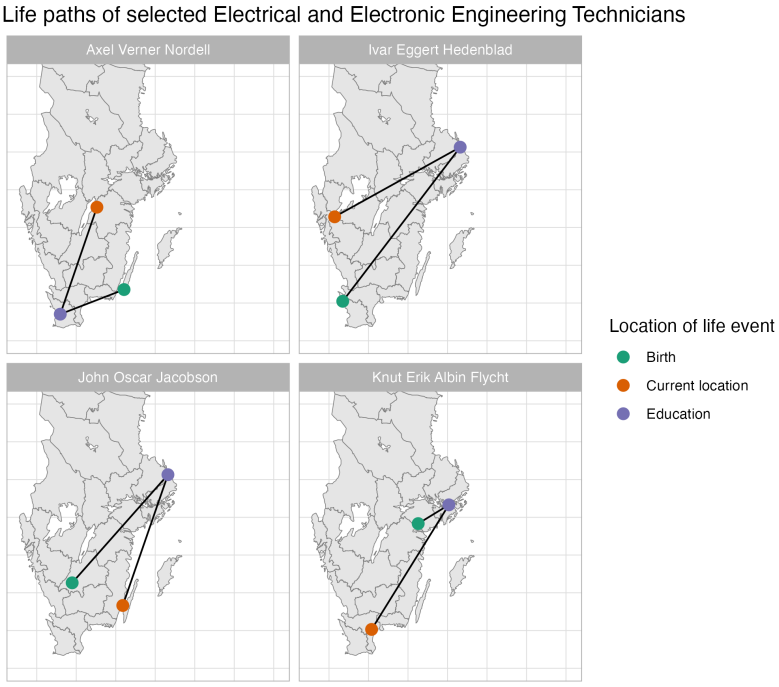


[Here we show that mechanical workshop firms are relatively labour intensive, while electrical appliance and machinery producing firms are relatively capital intensive, compared to all other firms (in purple)]

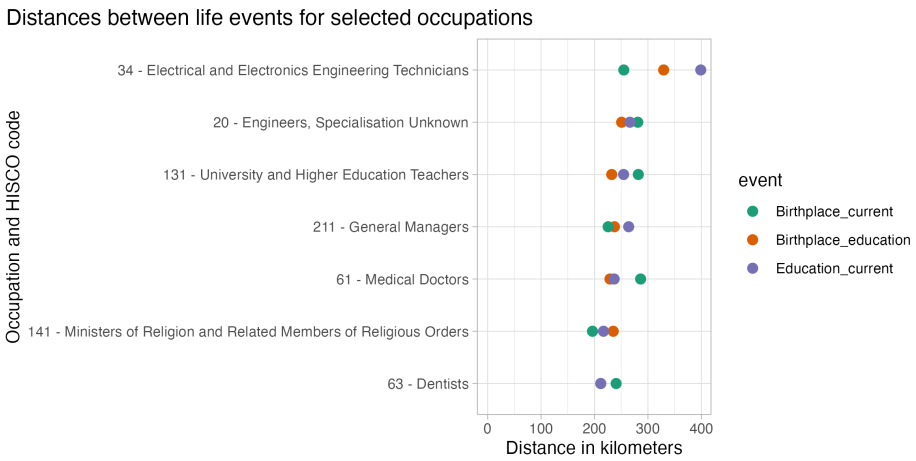
What can we learn from the Who is Who? biographies?



[We can show that of those in our sample, a great deal are general managers of firms, as well as doctos, dentists, teachers and priests. Engineers make up a large other component.]



[We can show that electrical and electronic engineering technicians move the furthest from the birthplace to their places of education, and further still from their place of education to their current location. Not the case that local lads were filling the roles for skilled workers in these new occupations - kinda interesting!]



Appendix

Figure 4 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 expression 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.

References

Adams, A. (2018). Technology and the labour market: The assessment. *Oxford Review of Economic Policy*, *34*, 349–361. https://doi.org/10.1093/oxrep/gry010

Amujala, S., Vossmeyer, A., & Das, S. R. (2023). Digitization and data frames for card index records [Methodological Advances in the Extraction and Analysis of Historical Data]. *Explorations in Economic History*, *87*, 101469. https://doi.org/https://doi.org/10.1016/j.eeh.2022.101469

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, *128*(4), 1325–1369. https://doi.org/10.1086/705331

Bandiera, O., Rasul, I., & Viarengo, M. (2013). The making of modern america: Migratory flows in the age of mass migration [Migration and Development]. *Journal of Development Economics*, *102*, 23–47. https://doi.org/https://doi.org/10.1016/j.jdeveco.2012.11.005

Clark, G., & Cummins, N. (2018). The big sort: Selective migration and the decline of northern england 1780-2018.

Correia, S., & Luck, S. (2023). Digitizing historical balance sheet data: A practitioner's guide [Methodological Advances in the Extraction and Analysis of Historical Data]. *Explorations in Economic History*, *87*, 101475. https://doi.org/https://doi.org/10.1016/j.eeh.2022.101475

Dahl, C. M., & Vedel, C. (2024). *Breaking the hisco barrier: Ai and occupational data standardization* [Working Paper], University of Southern Denmark.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Eichengreen, B. (2021). Gold and south africa's great depression. *Economic History of Developing Regions*, *36*(2), 175–193. https://doi.org/10.1080/20780389.2021.1891879

Feigenbaum, J., & Gross, D. P. (2020). Answering the call of automation: How the labor market adjusted to the mechanization of telephone operation. *National Bureau of Economic Research*. https://doi.org/10.3386/w28061

Ford, N. M., Ranestad, K., & Sharp, P. (2023, November). *Not the Best Fillers in of Forms? The Danish and Norwegian Graduate Biographies and "Upper Tail Knowledge"* (Working Papers No. 0242). European Historical Economics Society (EHES). https://ideas.repec.org/p/hes/wpaper/0242.html

Goldin, C. (1994). Labor markets in the twentieth century. *National Bureau of Economic Research*. https://doi.org/10.3386/h0058

Goldin, C. (1998). Labor markets in the twentieth century. In *The cambridge economic history of the united states, vol. iii* (pp. 549–624). Cambridge University Press. https://scholar.harvard.edu/goldin/publications/labor-markets-twentieth-century

Goos, M. (2018). The impact of technological progress on labour markets: Policy challenges. *Oxford Review of Economic Policy*, *34*, 362–375. https://doi.org/10.1093/oxrep/gry002

Gutmann, M. P., Merchant, E. K., & Roberts, E. (2018). "big data" in economic history. *The Journal of Economic History*, *78*(1), 268–299. https://doi.org/10.1017/S0022050718000177

Harnesk, P. (1945). *Vem är vem? stockholmsdelen* [Accessed through Projekt Runeberg]. Vem är Vem Bokförlag. https://runeberg.org/vemarvem/sthlm45/0003.html

Jayes, J., Molinder, Jakob, & Enflo, Kerstin. (2024). *Power for progress: The impact of electricity on individual labor market outcomes* (Unpublished Manuscript) [Unpublished Manuscript].

Jurafsky, D., & Martin, J. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall. https://books.google.se/books?id=fZmj5UNK8AQC

Koschnick, J. (2023). Breaking tradition: Teacher-student effects at english universities during the scientific revolution [Presented on March 23, 2023, from 16:00 to 17:15]. https://www.wur.nl/en/activity/rhi-seminar-breaking-tradition-teacher-student-effects-at-english-universities-during-the-scientific-revolution.htm

Mokyr, J. (2017a). Bottom-up or top-down? the origins of the industrial revolution. *Journal of Institutional Economics*. https://doi.org/10.1017/S1744137417000251

Mokyr, J. (2017b). A culture of growth: The origins of the modern economy. *Princeton University Press*.

Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives*, *31*(2), 87–106. Retrieved January 16, 2024, from http://www.jstor.org/stable/44235000

National Library of Sweden/KBLab. (2024). Bert-base-swedish-cased.

Nekoei, A., & Sinn, F. (2020). Herstory the rise of self-made women. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3741332

OpenAI. (2023). Strategy: Write clear instructions - prompt engineering [Accessed: 2024-01-17].

OpenAI. (2024). Text generation: Json mode [Accessed: 2024-01-17].

Shen, Z., Zhang, K., & Dell, M. (2020). A large dataset of historical japanese documents with complex layouts. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 548–559. https://dell-research-harvard.github.io/HJDataset/

van Leeuwen, M. H., Maas, I., & Miles, A. (2002). *Hisco: Historical international standard classification of occupations* [Cloth]. Leuven University Press.