Chapter 4 Appendix

This Appendix lists the choices made during the selection and cleaning of data for analysis. All the steps below are performed in Python.

Table A.4.1 Overview of choices and steps for initial filtering.

What	Why	How
Filtering duplicates	There are many duplicates in the dataset. Companies may issue the same advertisement twice (with only slight changes) to increase the chances of getting a response. Indeed may also scrape the same job ad from several sources.	We considered two ads as duplicates if their titles, organizations and locations are the same. We also considered two ads as duplicates when the cosine distance between the TFIDF-weighted word-representation of the descriptions is higher than 0.95. By weighing the words upfront, we account for the fact that two job ads may have many words in common while not actually being duplicates.
Exclude assistants	The search query also resulted in advertisements for “assistants to” the director, chief, or president we are interested in. These assistants have very different tasks, like arranging meetings, managing agenda’s and such.	We explicitly searched the titles of job ads for phrases “assistant to”, “executive assistant”, “special assistant”, “administrative assistant”, “technical assistant”, “secretary”, “to the”, and “ea to”. We checked the resulting titles and excluded those job ads corresponding to assistant positions.
Exclude non-seniors	The search query also resulted in advertisements for other roles part of “the office of the” director, chief, or president we are interested in. These may also have something to do with data and analytics, but their own function titles actually do not include “director”, “chief officer”, or “president”.	We explicitly searched the titles for phrases such as “office of (the) ch”; “office of (the) (v)p“; office of (the) d”, “team of”, “officer’s”, “president’s”, and “director’s”. We checked the resulting titles and excluded those job ads that are not for “chief officers”, “directors” or “presidents”.
Exclude other chiefs	The search query also resulted in advertisements for other chief roles, such as a “Chief Operating Officer”, “Chief Financial Officer” and “Chief Learning Officer”, because they have phrases like “of a technology company” in their title.	We first did an inventory of what types of chiefs can be found. Then, we explicitly searched the titles for phrases “chief operat”, “chief financ”, “chief learn”, “chief customer”, “chief compl”, “chief of staff”, “chief credit”, and “chief commerc” and excluded those job ads not actually adhering to our inclusion criteria.
Other	We excluded a number of job ads that we came across that were clearly irrelevant to our research question.	We searched the titles for phrases such as “organization”, “company”, “school”, “division”, “department”, and “college”. While keeping job ads of “directors” and “presidents” that are responsible for a subarea within a technology department (e.g., a “Enterprise PMO Director, City Office of Information and Technology” and a “Director business development, Office of technology”), we deleted some job ads that were clearly irrelevant to our research question (e.g., a Massage Therapy Program Director at a School of Business and Technology)
Exclude small advertisements	Vacancies that are extremely short may not be very representative.	We excluded vacancies that consist of fewer than 500 alphabetic characters.
Exclude other languages	Even though we scraped data from the U.S. version of Indeed.com, some job ads were presented in a different language.	We used the langdetect package in Python to decide whether a job ad is in another language or not.

Table A.4.2 Overview of choices and steps for cleaning the data.

Cleaning data	Why	How
Remove URL	Remove URLs from text	We used regular expressions to look for common URL-forms, such as lines starting with www.; http.; and/or ending in .com; .nl, and similar.
Fix abbreviations	Abbreviations come in two forms, e.g., AI and A.I, but they mean the same thing.	We used regular expressions to transform abbreviations with periods into their non-period counterpart (A.I. becomes AI).
Remove disclaimer	Job ads often contain disclaimers, which have nothing to do with the responsibilities of the job, but may influence the topic modeling results.	We examined a number of disclaimers to find common words: “equal opportunity”, “equal employment”, “affirmative action”, “not discriminate”, “race, color”, “religion”, “gender”, “sexual”, “convicted”, “offender”, “veteran”, “non-discrimination”, and “harassment”. In cases where a disclaimer is positioned near the end of the vacancy, we remove everything that comes after such a keyword (<1000 characters)[1]. If this failed, we deleted pieces of text between two periods or tabs (<500 characters) containing one of these words. We manually evaluated the removed texts.
Replace special characters	The use of special characters like á or ü can be inconsistent across different advertisements.	We replace these special characters by their “regular” counterpart.
Replace punctuation	If not removed, punctuation can be treated as words or parts of words.	We replaced the “-” by an empty space, thereby concatenating two words linked by a hyphen. We replaced all other punctuation by a whitespace.
Replace numbers	If not removed, numbers can be treated as words or parts of words.	We replaced numbers by a whitespace.
Replace IT	As soon as we start removing words, we make all words lowercase. IT then becomes “it”, which is a stopword. We assume other abbreviations are not subject to being accidentally removed.	Replace instances of IT by Information Technology
Remove company	We prefer not to end up with topics in which company names are among the top 10 words. We have this information as metadata.	For each advertisement, we replaced the name of the company (cleaned version) that published the particular advertisement. First we deleted the full version (e.g., Walt Disney). Then, we split the company name and removed its elements from the respective advertisement, when they are non-english words (e.g., ‘Walt’ and ‘Disney’ are removed, but not ‘Industries’). In case some company names still showed up in topics, we added them as “stopwords”[2].
Remove location	We prefer not to end up with topics in which location-based information is among the top 10 words. We have this information as metadata.	We extracted a list of cities from the location metadata of all advertisements. We also copied a list of (abbreviations of) American states. Given that we scraped from the U.S-version, we removed synonyms for the United Stated. Other countries may be mentioned in the job ads, but we found that this

[1] This might also lead to the exclusion of a final summary of the company, which we took for granted as such information is also not relevant for answering our research question.

[2] One of the companies is called 1010data, which means that for the job ads published by this company, the word “data” is taken out. We evaluated these job ads (three in total) and found that the company name is often mentioned, while the positions are not always responsible for the data. Only one of them is directly responsible for data, and we later saw that this job ad still scores on many of the data-related topics.

		happens infrequently and that these country names do not show up as distinct dimensions.
Remove title	We prefer not to end up with topics in which the functions are mentioned (e.g., a topic “chief officers”).	We removed the terms "chief", "officer", "vice", "president", “vicepresident”, "avp", "vp", "svp", "evp", and "director". We also removed instances of CxO and CxyO (not the plural). Unfortunately, we could not consistently remove the full title. Thus, when the title “Chief Data Officer” is mentioned, “Data” will remain. Terms like “associate” and “assistant” or “executive” also remain. The effects of this are marginal.
Lemmatize	Lemmatizing means to bring a word back to its base, single and non-plural form. For example, “analysts” becomes “analyst”.	We used WordNetLemmatizer in Python. We treat all words as nouns, to prevent words such as “machine learning” from treated as verbs.
Remove stopwords	Stopwords are words that do not convey meaning or are not relevant to our goal. For example, common stopwords are “the”, “have” and “do”. We extended nltk’s standard by a number of advertisement-related stopwords and company names that popped up (see Remove Company).	We used the standard list of nltk.stopwords.words(“english”). We added a number of common english stopwords (e.g., “may”, “might”, “please”), vacancy-related stopwords (e.g., "application", "applicant", "resume","job", “candidate”), and words that denote time periods (“day”, “month”, “year”), We also removed company abbreviations that would pop up in the topics (e.g., “jp”), and corporate suffixes (e.g., Inc. Co.)
Remove single characters	Single characters may be in the corpus, either as a result of cleaning or because of typos.	We removed single characters.
Bigrams	Bigrams are pairs of words that often co-occur. For example, “big data” and “information technology”.	We used Gensim’s Phraser class to locate bigrams. We set the threshold to 1 and the minimum count to 10% of the length of the corpus. We then replaced bigrams in the texts by their concatenated version.

Table A.4.3 All 55 topics and their top 10 words.

Nr	Label (author-assigned)	Top 10 words
0	business_strategy	business strategy drive strategic partner leadership capability leader team key
1	digital_strategy	digital strategy platform mobile engagement transformation commerce channel team lead
2	information_systems	system implementation informationsystem software business support user management maintenance ensure
3	cyber_security	security threat cyber vulnerability incident compliance cybersecurity response secure policy
4	(market)_research	research market insight commercial analysis intelligence competitive study science development
5	client_solutions	client account agency consulting solution relationship business engagement team industry
6	data_sources	data source quality warehouse dataanalytics use warehousing tool database set
7	healthcare	healthcare analytic pharmacy provider care claim sm health life healthier
8	product_development	product market feature roadmap user management new define launch agile
9	college	student faculty program college academic school learning admission teaching curriculum
10	sales	sale revenue market selling account target pipeline solution commercial opportunity
11	web_design	web design user site website mobile ux development testing html
12	technologies	network server hardware software infrastructure computer window microsoft maintenance equipment
13	analysis	statistical modeling analysis model quantitative analytical technique analytic advanced predictive
14	clinical_trials	clinical trial study development regulatory drug informatics quality patient biostatistics
15	IT_strategy	informationtechnology budget planning infrastructure strategic plan leadership vendor operation department
16	media	medium agency ad audience advertising measurement programmatic planning buying team
17	analytics	analytics insight team advanced measurement google adobe tool actionable reporting
18	marketing_strategy	marketing strategy direct channel generation crm campaign email program lead
19	risk_management	risk control riskmanagement regulatory assessment compliance credit operational framework cyber
20	digital_content	content socialmedium social video website platform audience communication editorial news
21	digital_marketing	digitalmarketing seo marketing sem channel email strategy socialmedium ecommerce google
22	architecture	architecture enterprise solution design technical architect development integration information business
23	department_management	develops manages ensures staff department provides directs oversees maintains plan
24	data_science	datascience machinelearning scientist algorithm python data ai statistic team predictive
25	education	district school instructional staff superintendent education educational student learning teacher
26	information_security	informationsecurity program policy incident security riskmanagement standard certified compliance risk

27	customer_engagement	customer engagement journey drive solution loyalty value insight success retention
28	job_post	credit post approval time submit card diploma school least test
29	operations	service operation management support delivery infrastructure center operational process provider
30	human_resources	hr human workforce hris talent workday compensation global reporting people
31	consumer_banking	serve client diversity standard banking making consumer priority safeguarding strengthens
32	project_management	project team manage projectmanagement delivery work ensure program management multiple
33	technology_innovation	technology emerging innovation development new solution technical strategic leadership infrastructure
34	commerce	brand consumer ecommerce insight team creative strategy global partner agency
35	auditing	audit control internal auditor dataanalytics auditing compliance professional function cisa
36	medical_information	medical scientific information patient affair pharmaceutical disease commercial global drug
37	job_history	check personalized associate criminal verification largest conversant alliance history affiliate
38	data_governance	datagovernance enterprise quality data governance policy program standard master metadata
39	financial_reporting	financial finance reporting analysis accounting intended total crime report performance
40	data_privacy	privacy compliance protection law legal regulation information regulatory global data
41	business_intelligence	bi businessintelligence reporting warehouse tool sql tableau analytic database etl
42	pubic_sector	agency city state information department public federal county government education
43	supply_chain	supply chain manufacturing sourcing commercial supplier global process procurement operation
44	job_demands	perform essentialfunction physical accommodation reasonable work disability demand made individual
45	job_opportunity	people work life benefit team help insurance career opportunity make
46	investments	global investment asset management wealth world market fund trading equity
47	engineering	engineering team software engineer technical platform development building build agile
48	banking	banking bank consumer financialservice ccb commercial business credit small fraud
49	campaigns	campaign paid search social advertising display google socialmedium channel acquisition
50	hospital	hospital care patient physician informationsystem center facility healthcare shift medical
51	data_management	datamanagement dm master process quality study cro regulatory standard database
52	health	health information record population management care registered medical rhia program
53	big_data_infrastructure	cloud bigdata platform hadoop aws azure spark machinelearning infrastructure technology
54	university	university campus advancement fundraising alumnus education gift higher donor staff
Sklearn.decomposition.nmf. Random seed: 2704, n_topics: 55, bigrams: replaced, minimum document frequency: 0, beta_loss: Frobenius, max_iter: 100.000, tol: 0.0001, init: None, alpha: 0.0, solver: “cd”		

Table A.4.4 Allocation of topics to themes and categories.

Theme (author-assigned)		**Label (author-assigned)**	**Topic #**	**Theme (author-assigned)**	**Label (author-assigned)**	**Topic #**
Job Responsibilities				***Application Domains***		
Data Analytics	Data infra-structure	data_sources	6	(Digital) Marketing and sales	(market)_research	4
		big_data_infrastructure	53		sales	10
	Data control	data_governance	38		web_design	11
		data_management	51		media	16
	“”	data_privacy	40		marketing_strategy	18
	Applied analytics	business_intelligence	41		digital_content	20
		analytics	17		digital_marketing	21
		data_science	24		customer_engagement	27
		analysis	13		campaigns	49
Technology		information_systems	2	Human resources	human_resources	30
		technologies	12	Finance	financial_reporting	39
		IT_strategy	15	Supply chain	supply_chain	43
		architecture	22	***Industry***		
		operations	29	Healthcare	healthcare	7
		technology_innovation	33		clinical_trials	14
		engineering	47		medical_information	36
Digital		digital_strategy	1		hospital	50
Business		business_strategy	0		health	52
Management		department_management	23	Education	college	9
		project_management	32		education	25
Products and services		client_solutions	5		university	54
		product_development	8	Banking	consumer_banking	31
Security and risk		cyber_security	3		investments	46
		risk_management	19		banking	48
		information_security	26	Commerce	commerce	34
		auditing	35	Public	pubic_sector	42
				Vacancy related	job_post	28
					job_history	37
					job_demands	44
					job_opportunity	45

Table A.4.5 Top publishing organizations of job ads for different positions.

	All positions	**“data analytics”**	**“digital”**	**“information technology”**
Nr of organizations	4719	1707 (36%)	1341 (28%)	2610 (55%)
Top 10 publishing organizations	jpmorgan chase (271) citi (149) mufg (73) nike (66) synchrony financial (61) unitedhealth (57) visa (51) state street (50) harnham (43) goldman sachs (42)	jpmorgan chase (136) citi (88) harnham (43) unitedhealth (39) state street (35) mufg (34) synchrony financial (30) visa (29) nike (27) johnson johnson (21)	jpmorgan chase (68) nike (25) general electric (23) visa (16) mastercard (16) synchrony financial (15) nbcuniversal (15) citi (15) charles schwab (13) weber shandwick (11)	jpmorgan chase (67) citi (46) mufg (36) goldman sachs (27) synchrony financial (16) state street (14) nike (14) clinical management consultants (14) bny mellon (13) unitedhealth (12)

Table A.4.6 Top 10 positions for each data-related topic.

Data sources	**Big data infrastructure**
Watson Health <U+2013> Data Engineering & Management Sr Director (weight: 0.277)	Information Technology Director - Data Analytics (weight: 0.620)
Data Analyst, Vice- President (weight: 0.212)	AVP - Big Data & Analytics <U+2013> Product Manager - Irving, TX (weight: 0.584)
Director - Data Engineering and Quality Assurance (weight: 0.197)	Big Data & Analytics <U+2013> Product Manager - VP - Irving, TX (weight: 0.567)
Director of Data Analytics (weight: 0.195)	Sr. Director <U+2013> Cloud Data Strategy (weight: 0.554)
Data Architect, Director (weight: 0.194)	Data Center Transformation Director , Center of Excellence (weight: 0.494)
Business / Data Analyst, AVP (weight: 0.193)	Product Marketing Director , Oracle Big Data Cloud Platform M... (weight: 0.487)
Director of Data (weight: 0.192)	Director , Big Data Infrastructure (weight: 0.475)
Director , Enterprise Data Architecture (weight: 0.190)	Cloud Expense Management & Analytics - Assistant Vice Preside... (weight: 0.427)
Deputy Director Data Specialist Market Access (weight: 0.187)	Big Data Solution Architect - SVP (weight: 0.421)
Director of Data Warehouse (weight: 0.185)	Director of Big Data , Platform, Ecommerce (weight: 0.394)
Data governance	**Data management**
Global Director of Data Governance (weight: 0.627)	Senior Director , Clinical Data Management (weight: 0.581)
Technology Risk / Data Governance, VP (weight: 0.617)	Associate Director , Clinical Data Management (weight: 0.429)
VP , Data Governance (weight: 0.583)	Director of Data Management (weight: 0.424)
VP , Data Engineering and Governance (weight: 0.579)	Senior Director , Clinical Data Management (weight: 0.396)
Director Enterprise Data Governance Data Quality (weight: 0.531)	Director Data Management (weight: 0.365)
Director of Data Governance (weight: 0.516)	Senior Director , Clinical Data Management (weight: 0.364)
Director , Data Governance (weight: 0.509)	Director , Clinical Data Management (weight: 0.356)
Director Data Governance Administration (weight: 0.499)	Associate Director , Data Management (weight: 0.352)
Vice President , Enterprise Data Governance (weight: 0.478)	Associate Director , Data Management - Oncology (weight: 0.346)
Vice President , Data Governance and Analytics (weight: 0.470)	Director , Clinical Data Management, Gene Therapy (weight: 0.340)

Business intelligence	**Analytics**
Director Business Analytics (weight: 0.522) Senior Director , Business Intelligence & Data Analytics (weight: 0.504) Business Intelligence and Analytics Director (weight: 0.487) VP Information Services (weight: 0.438) Director , Business Intelligence and Analytics (weight: 0.437) Director of BI & Analytics (weight: 0.435) VP Apps Enterprise Data and Reporting (weight: 0.434) Associate Director of Business Analytics (weight: 0.423) Data Solutions Director for Data Warehousing (weight: 0.422) Director , Information Management (weight: 0.397)	VP of Analytics (weight: 0.413) Director of Digital Analytics (weight: 0.322) Associate Director , Analytics (weight: 0.293) VP Digital Analytics (weight: 0.275) Associate Director , Measurement Strategy and Data Architectu... (weight: 0.273) VP of Analytics (weight: 0.271) Digital Analytics Director (weight: 0.261) Director , Digital Analytics - Boutique Agency (weight: 0.258) Director of Analytics (weight: 0.245) Associated Director of Digital Analytics (weight: 0.240)
Data science	**Analysis**
VP of Data Science (weight: 0.794) Director of Data Science (weight: 0.712) Director of Data Science (weight: 0.616) Associate Director , Senior Data Scientist (weight: 0.611) Director of Data Science (weight: 0.572) Director of Data Science (weight: 0.563) Associate Director , Lead Data Scientist (weight: 0.562) Vice President - Data Science & Analytics (weight: 0.518) Director of Data Science (weight: 0.513) Vice President - Data Science (weight: 0.510)	Quantitative Analyst II, AVP Retail Credit Risk Analytics (weight: 0.686) Quantitative Analyst III, VP Retail Credit Risk Analytics (weight: 0.571) Quantitative Analyst, AVP Retail Credit Risk Analytics (weight: 0.556) Director , Modeling Analytics Job (weight: 0.545) Risk Specialist VP - Commercial Credit Risk Analytics (weight: 0.545) Sr. Manager or Director <U+2013> Pricing Modeling & Analytics (weight: 0.538) Quantitative Analytics Consultant, AVP (weight: 0.482) Director , AML Modeling Analytics Job (weight: 0.476) Quantitative Manager, VP Retail Credit Risk Analytics (weight: 0.467) Quantitative Manager II , VP Retail Credit Risk Analytics (weight: 0.466)
Data privacy	
Senior VP / Data Privacy Lead (weight: 0.713) Associate Director , Global Data Privacy Counsel (weight: 0.668) Director , IT Data Privacy (weight: 0.667) DIRECTOR , PRIVACY & DATA GOVERNANCE (weight: 0.643) Chief Privacy and Data Ethics Officer (weight: 0.627) Data Privacy Senior Director (weight: 0.627) Director of Compliance & Data Protection Officer (weight: 0.586) Director , Data Protection & Compliance (weight: 0.568) Director , Security Data Compliance (weight: 0.568) Director - Data Protection Compliance, Software Enterprise (weight: 0.529)	