# 1. Motivation[[1]](#footnote-1)

Datasheets for Datasets  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.*

Derive

***1.1*** *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

Indeed is a worldwide employment website for job listings. Currently, Indeed is the market-leader within the Dutch labor market.

Data analyst is currently a high in demand job. The purpose for collecting this data is to gather insights into which skills are the most commonly required for such a position. The dataset aims to enable answering the research question in terms of what are the key skills required for a ‘Data Analist’ job in the Netherlands?

In answering this question, the researchers can for instance spot trends when it comes to which programming languages are more popular and which are decreasing in demand, as well as provide a guideline for young professionals seeking a data analyst position as a future job.

***1.2*** *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

This dataset has been created by Anouk Heemskerk, Georgiana Huţanu, Renée Nieuwkoop and Alan Rijnders at Tilburg University.

***1.3*** *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

The deployment of this project has no associated grant.

# 2. Composition

Indeed.com  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2021). Indeed.com.\**

Derive

***2.1*** *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

Each instance consists of a ‘Data Analist’ job posting in the Netherlands.

***2.2*** *How many instances are there in total (of each type, if appropriate)?*

The dataset contains .... job postings in total  
(Number of jobs we were able to scrape in the definitive dataset)

***2.3*** *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset contains all the the possible instances available on Indeed.com with the search term for ‘Wat’: ‘Data Analist’, and for ‘Waar’: ‘Nederland*’*.

***2.4*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.*

Each instance consists of the full job description (only textual information) that is listed on Indeed.com.

***2.5*** *Is there a label or target associated with each instance? If so, please provide a description.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***2.6*** *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

(Only doable if we actually have the data)

Missing: Star ratings? Salary in some of them.

***2.7*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***2.8*** *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***2.9*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***2.10*** *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

No. All the data within the set derived from a publicly available website and does not contain personal and/or confidential data.

***2.11*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

No. The dataset only consists of textual information within job descriptions.

***2.12*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

No. The dataset does not relate to individual people.

***2.13*** *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

***2.14*** *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

***2.15*** *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

# 3. Collection Process

***3.1*** *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The data within the set was directly observable as raw text.

***3.2*** *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.3*** *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.4*** *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.5*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.6*** *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.7*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.8*** *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.9*** *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.10*** *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.11*** *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***3.12*** *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

# 4. Preprocessing, cleaning, labeling

***4.1*** *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***4.2*** *Was the “raw” data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***4.3*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

# 5. Uses

***5.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***5.2*** *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***5.3*** *What (other) tasks could the dataset be used for?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***5.4*** *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

***5.5*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)