



University of Zurich
Department of Informatics

Revealing the inherent variability in data analysis



Master-Thesis August 20, 2017

Nicola Staub

of Bern BE, Switzerland

Student-ID: 11-712-015
staub.nicola@gmail.com

Advisor:

Michael Feldman

Prof. Abraham Bernstein, PhD
Department of Informatics
University of Zurich
<http://www.ifi.uzh.ch/ddis>

Acknowledgements

I would like to thank Prof. Abraham Bernstein for letting me accomplish my master thesis in the Dynamic and Distributed Information Systems Group.

I would also like to acknowledge the outstanding support of Michael Feldman who accompanied me through the pursuance of my master thesis. Michael not only supported me in creating this thesis, but also taught me how interesting and versatile the field of data science is.

Very special thanks are also due to Prof. Martin Schweinsberg and Prof. Eric Uhlmann, which let my thesis be a part of a big study in crowdsourcing data analysis.

Last but not least, I want to express my gratitude to all the data analysts who conducted their analysis as part of this thesis.

Zusammenfassung

Schwankungen in Datenanalysen wurden kürzlich als eine der Hauptursachen für die Krise der Reproduzierbarkeit wissenschaftlicher Studien erkannt. Viele dieser Studien weisen statistisch signifikante Resultate vor, was jedoch nicht zwingend heissen mag, dass diese in Wirklichkeit auch sinnvoll sind. Viele Faktoren beeinflussen analytische Entscheidungen von Datenanalysten während ihren Analysen. Diese Arbeit versucht potentielle Faktoren zu finden, mit welchen sich die Varianz in Datenanalysen erklären lassen kann. Mit einer im Rahmen dieser Arbeit entwickelten Plattform wurden die Grundüberlegungen der Datenanalysten während dessen Arbeitsabläufen eruiert. Ein System von verschiedenen Faktoren konnte dabei entwickelt werden, welche eine genauere Untersuchung dieser Arbeitsabläufe in unterschiedlicher Tiefe möglich macht.

Abstract

The variation in data analysis has been recently recognized as one of the major reasons for the reproducibility crisis in science. Many scientific findings have been proven to be statistically significant, which is however not necessarily an indication, that the results are indeed meaningful. There are many factors playing a role in the analytical choices a data analyst makes during an analysis. The goal of this thesis is to find potential factors which can explain the variability in data analysis. With a special platform designed in accordance with this thesis, rationales for different analytical choices along the path of a data analysis were elicited. The result of this thesis is a system of factors, which allow for examining data analysis workflows in different levels of depth.

Table of Contents

1	Introduction	1
2	Literature Review	3
3	Methodology	9
4	Experimental Design	15
4.1	Scope of Experiment	15
4.2	DataExplained	16
4.3	Qualitative Analysis	21
4.4	Quantitative Analysis	22
5	Analysis & Results	23
5.1	Qualitative Analysis	23
5.2	Quantitative Analysis	26
6	Discussion	29
7	Limitations & Future Work	31
8	Summary & Conclusions	33
9	Technical Documentation	35
9.1	Architecture	35
9.2	Technical Setup	36
9.2.1	Setup EC2 instance	36
9.2.2	Setup MongoDB	40
9.2.3	Setup RStudio Server	41
9.3	Run, Build, Deploy	42
9.3.1	Prerequisites	42
9.3.2	Development	42
9.3.3	Deployment	42
9.4	Database Backup	43
9.4.1	Cronjobs	43

A Appendix	45
A.1 Description of Dataset	45
A.2 CS2 Phase 2 pre-survey for analysts	54
A.3 CS2 Phase 2 post-survey for analysts	61
A.4 Code Book	64
A.5 Quantitative Analysis	71
A.5.1 Clusters of Participants	71
List of Figures	73
List of Tables	75

1

Introduction

Data analysis is a versatile science applied in many fields of research. It entails many activities, from stipulating hypotheses, collecting data relevant to a problem, to building models and translating them in quantitative findings. There are virtually no limits in treating data and finding patterns, as scientists can interpret a dataset in many ways and think of different reasonable measures which can lead to statistically significant results as a consequence. Nevertheless, findings shown to be statistically significant do not necessarily indicate that the results are indeed meaningful. The reason for this lies in different kind of biases introduced by the pursuit of actions during the analysis [Brodeur et al., 2016, Simmons et al., 2011, Head et al., 2015].

Recent publications demonstrated that this reproducibility crisis is ubiquitous and proposed ways to account for this variation in data analysis [Humphreys et al., 2013, Goutefangeas et al., 2015, Gonzales and Cunningham, 2015, Head et al., 2015]. For instance, [Gelman, 2013] doubts the credibility of a study which claims that women are more likely to wear red or pink at fertility [Beall and Tracy, 2013]. He points out that this study lacks a representative participant sample and its findings are affected by measurement biases. Additionally, Gelman claims that the authors of this publication obviously over-interpret patterns found in the data due to a series of implicit choices.

While works like this of [Gelman, 2013] criticize the methods doubtful studies applied to come up with their findings, they barely address the inherent cause for various analytical decisions. The goal of this master thesis is to fill this gap by investigating potential factors which account for variability in data analysis.

To elicit these factors, I developed a platform on which different scientists conducted their data analysis on. During this progress, they described and justified analytical choices by commenting their analysis script. As [Conklin and Yakemovic, 1991] claim, an analyst's workflow is often marked by breakthroughs of understandings, leading to conceptual restructurings and invalidations of previously made decisions and assumptions. I believe, to let data analysts critically reflect their analytical decisions might help

them to get a faster understanding of the underlying problem(s) they face, which may be vague or still poorly understood due to the analyst's background.

The thesis is organized as follow. Chapter 2 reviews the problems of the statistical crises discussed in literature and outlines the cognitive cycles an analyst traverses along the analysis. Chapter 3 defines the methodology used to analyse the insights gained from the results of the crowdsourced data analysis experiment, whose setup is described in Chapter 4. The results of the analysis are presented in Chapter 5 and discussed in Chapter 6.

2

Literature Review

Often, data analysts solely rely their findings' evidence on p-values while rejecting their null hypotheses if the measure is small enough. [Gelman and Loken, 2014] claim, that those values are often manipulated in the way, that analysts seek for the best decision variables which lead to their desired outcome. The actual hypothesis thereby corresponds to many possible decision variables, as the data can always be treated in the way, a "statistically significant" result emerges. Gelman & Loken call this issue "multiple comparison problem" or "p-hacking", while other scientists link to the terms "researcher degrees of freedom" [Simmons et al., 2011], or "selective reporting" [Head et al., 2015]. As [Gelman and Loken, 2013] state in another publication, differences along the paths of data-analytic choices are due to implicit decisions researchers take during their analysis. Although all the possible paths may lead to statistically significant results, it is wrong to claim for strong evidence of the initially overarching hypothesis. This evidence only holds for the hypothesis postulated by the respective analysts, produced with subjective analytical choices in the given analysis settings.

If data analysts consciously make analytical choices to form a model which favours a desired outcome, this model can be described as "fished" model [Humphreys et al., 2013]. This supposedly malicious behaviour of fishing models can mostly be attributed to implicit decisions (which may be reasonable given the data) [Dwork et al., 2015, Gelman and Loken, 2013, Gelman and Loken, 2014]. The uncertainty on what is "the best path to follow" and the researcher's desire to find a statistically significant result, are the main factors which underlay this exploratory behaviour in adaptive data analysis [Simmons et al., 2011, Lukacs et al., 2010, Song et al., 2010].

For example, when forming regression equations, analysts often select and treat extraneous variables as "significant" [Freedman, 1983]. Thereby, they lean on high F statistics of those explanatory variables as an indicator for "good" variables, although the relation to the response variable is weak to non-existent. This "mistake" often appears when the number of explanatory variables is very high (close to the number of entries in the dataset). This problem is also referred to as "Curse of dimensionality" [Bellman,

2013, Bellman, 2015]. Many of those variables lead to a high R^2 , which stays high after refitting the model without variables having low t statistics. This leads to an overall F statistic of high significance. This phenomenon is also known as Freedman’s paradox, as such measures may give analysts false confidence in the predictability of certain explanatory variables [Freedman, 1983, Hardt, 2015, Lukacs et al., 2010]. As [Lukacs et al., 2010] point out, the Freedman’s paradox is an extreme case of model selection bias, since the effect of (weakly) unrelated explanatory variable are overestimated. Possible measures to account for presumably high R^2 are Mallows’s C_p , R^2 adjusted, AIC or BIC. In practice, datasets get used multiple times, where the results of previous analyses may influence subsequent analyses. Obviously, this may lead to erroneous and biased outcomes [Dwork et al., 2015, Hardt, 2015]. Given the rise of the “big data” phenomenon, data analysts are confronted with more data, more complex relationships, which radiate in many directions. This can lead to the practice of apophenia: Seeing patterns in the data, although they do not actually exist [Boyd and Crawford, 2012]. On top of that, a large mass of raw data is most often not self-explanatory.

[Bollier and Firestone, 2010] further points out, that cleaning a large amount of data often constitutes problems of maintaining an objective interpretation of the data - especially if this data origins from disparate sources. Subjective assumptions have to be made as a consequence, in order to link multiple data sets together. It is important to build a model which represents the different data in its respective context *before* connecting them, in order to not falsely claim any causation due to the observed correlation [Anderson, 2008]. Regardless the size of the dataset however, an analysis is always subject to limitations and bias [Boyd and Crawford, 2012].

One approach proposed to address the challenges in adaptive data analysis is pre-registration. Pre-registration emphasizes analysts to report their research rationale, as well as the hypotheses together with the design and analytic strategy, before beginning with the study [Gonzales and Cunningham, 2015]. This measurement forces analysts to separate (exploratory) hypothesis generation from hypothesis testing [Humphreys et al., 2013]. Following this procedure can improve the credibility of results, but may not always be desirable, yet possible. When analysing public data on education trends, elections or the economy, for example, it might not be possible to get enough data for pre-registration [Gelman and Loken, 2014]. Also for exploratory analyses, which follows an iterated, partly inductive mode of research, pre-registration is near impossible [Collier et al., 2004]. Another possible disadvantage of limiting the analysts to the pre-registered analysis procedure may be the results of smaller studies. Discovering results which may be counter-intuitive might motivate the analysts for further studies, which however would have to be pre-registered anew, since the analysis protocol must not be adapted [Humphreys et al., 2013]. Moreover, this approach does not solve the problem of adaptive data analysis, where insights gathered from an existing dataset for analysing the same dataset anew, may lead to biased results as the preceding analysis is informing subsequent analyses.

A different approach to tackle this problem is to use a holdout (testing) set, which

can be used to validate an analyst’s hypotheses for significance. Since this data partition is independently drawn from the same data distribution as the predictive model, statistical inferences still hold. For any iterative step of trimming the model as a result of the validation process (i.e. by introducing a new covariate), a new holdout set has to be produced. Otherwise, the models are dependent on the respective holdout data which makes them overfitted. [Dwork et al., 2015] introduces a new methodology to preserve validity, while not be dependent on new data for testing. Thereby, the analysts can only access the holdout set through an algorithm, which hides the information of any individual data element. This approach also allows for sharing the data and outcomes with other analysts. [Cunningham and Gonzales, 2014] further emphasize, that publicly sharing this data increases transparency and accountability in scientific findings.

When conquering through the garden of forking paths during an analysis, researchers are confronted with different intermediate results. Assign meaning to them and form beliefs are thereby crucial facts for deciding which path to further follow. Consequently, a data analysis not only incorporates statistical or computational steps, but also cognitive processes. As [Grolemund and Wickham, 2014] point out, “*data analyses rely on the mind’s ability to learn, analyse, and understand*”, whereby “*each analysis attempts to educate an observer about some aspect of reality*”. These observers may have different professional backgrounds and/or experiences in data analysis, as well as different mental capabilities for dealing with such tasks (i.e. forming mental models).

The concept of mental models is being studied in various research areas of cognitive science for many years [Craik, 1943, Norman, 1983, Seel, 1991, Seel, 2001, Weiss and Wodak, 2007, Grösser and Schaffernicht, 2012]. Scientist describe it as “*subjective representation of the events, action, or situation a discourse is about*” [Weiss and Wodak, 2007] or “*qualitative mental representations which are developed by subjects on the basis of their available world knowledge aiming at solving problems or acquiring competence in a specific domain*” [Seel, 2001]. The process of building and interpreting such descriptions of mental models or schemes is also known as a sensemaking [Russell et al., 1993].

Being confronted with data, situated cognition and reasoning in the sensemaking process have a considerable influence on how the data is being interpreted and transformed into information [Chi, 2008]. Prior beliefs about a certain phenomenon may be missing, incomplete or conflicting with correct information in a contradictory sense. Information gained from the data can help fill such gaps (if prior beliefs are incomplete), expanded (if prior beliefs are missing) or even revised (if false prior beliefs are contradicting correct information) [Chi, 2008]. Hence, the data by itself can influence an analyst’s beliefs, which leads to different analytical choices as a consequence [Paglieri, 2004].

A possible means for helping researchers to explore complex data and build better intuitions are appropriate visualizations [Morton et al., 2014, Fox and Hendler, 2011]. Without the need of knowledge for specific programming or query languages, visual analytics services can serve data analysts as efficient sensemaking tools. When being confronted with a lot of data, visualizations or visual exploration tools can help to facil-

itate the integration and study of correlations among multiple datasets [Morton et al., 2014, Bollier and Firestone, 2010, Fox and Hendler, 2011]. Especially when the data is of dynamic nature (e.g. temperature profiles), appropriate visualizations can help data analysts reveal new interesting patterns, which in turn can lead to adaptations of beliefs and/or mental models [Bollier and Firestone, 2010].

In their work “The psychology of attitudes”, Eagly and Chaiken studied the relationship between beliefs and attitudes [Eagly and Chaiken, 1993]. They claim that an attitude is a collection of interrelated beliefs, having an either positive, neutral or negative valence for the respective individual. This systemic interrelation of beliefs can also be described as *belief system*, which is utilized during the sensemaking process [Usó-Doménech and Nescolarde-Selva, 2016].

[Dole and Sinatra, 1998] presented a model which conceptualizes the theoretical assumptions of adapting mental models, which are widely studied in the fields of cognitive & social psychology, science, and education. They note, that the chance, humans are revising their existing conception of a phenomenon, is influenced by the qualities *strength*, *coherence*, and *commitment*. The quality strength can be seen as how sophisticated a subject’s idea is, which negatively correlates with the likelihood of change. If there is high conceptual coherence of an individual’s findings and existing knowledge, it lowers the chance of adapting the system. Regardless of the strength of a researcher’s idea or its conceptual coherence with existing models, the commitment to them varies. Reasons for that can be past experiences, social or cultural environment and backgrounds, ideologies, or simply different knowledge bases for the respective phenomena [Abelson, 1979, Dole and Sinatra, 1998]. As a result, belief systems are not consensual, i.e. people’s belief system within the same content domain may be different. An example showcasing the variety of belief systems in data analysis could be the experience in using a certain model. One analyst might never have used a model proven to be good at hypothesis testing, whereas the belief system of another analyst with much more experience prevents her from using that model due to negative experiences in the past. This hypothetical example also showcases that belief systems are dynamic and change over time as a result of belief revisions [Friedkin et al., 2016].

Once decided which route to take at a fork along the paths of decisions, a data analyst should also be able to justify the rationale behind the decisive action. Why should one follow this exact path? As seen before, each data analysis also attempts to educate an observer about some aspect of reality to gain their support. [Hill and Levenhagen, 1995] describe this (implicit) action of communicating the perceived mental model as *sensegiving*, which eventually results in shared belief systems or consensuses [Friedkin et al., 2016]. The description of these “whys” behind decisions in the context of designing a system or artifact, is also referred to as *design rationale* (DR) [Lee and Lai, 1991]. Latter define the term as “[...] *explanation of why an artifact is designed the way it is*”. Along with many other research areas, DR is widely discussed in the field of computer science [Schubanz, 2014, Schubanz et al., 2014]. Especially in software development, DR can help to effectively document and maintain artifacts (from both, the UI designer’s

point of view, as well as the technical engineer’s perspective) [Guindon, 1990]. The classic concepts of a design rationale system includes the existence of a design rationale database (containing design histories, reasonings, decisions, etc.). This database can be accessed with an appropriate representation schema, which elicits argumentations, decisions, or pros and cons for different options. In our case, an analyst implicitly accesses this system during the sensemaking/sensegiving processes. This perfectly reflects the definition given by [Conklin and Yakemovic, 1991], who claim that “*DR can be considered to be the path of decisions and selected alternatives that join the initial state (in which no decisions have been made) to the final state (in which all design decisions have been resolved)*”. **In a figurative sense, we could say that DR represents the signposts along the garden of forking paths.** Since at every of these signposts the analysts repeatedly traverse the conceptualizations of belief revision, mental model building, cognitive interpretation and rationale design, we can describe each subpath as a *cognitive cycle* a data analyst traverses.

We have seen that data analysis has many factors which account for variability among analyst’s results. Subjective actions are the result of an intertwined cycle of beliefs and the handling of data, mutually influencing each other. This interplay of data and beliefs could also be observed in the discussion of analysts who took part in the crowdsourced data analysis experiment of [Silberzahn and Uhlmann, 2015]. Provided with the same dataset and hypotheses, the data analysis was conducted by various scientists of different background. The amplitude of variation in the results was surprisingly large. The experiment outline was designed to have multiple feedback rounds where all the participating teams presented their analytic approach after each doing an isolated analysis of the same dataset. In addition, each team received peer review commentaries about their own and other teams’ analysis strategies. Individual or groups of analysts claim, that the data structure was partly responsible for the different kinds of analysis approaches they followed. After the experiment, some of them explained to have changed or adapted their statistical approach due to insights gained from the others’ analytic approaches, as well as from the qualitative and quantitative feedbacks. To place this action in our cognitive sensemaking and rationale design cycle, we can argue, that the findings from these feedbacks altered the belief of some analysts, which in return led to adaptations in how they treated the data.

By asking the analysts about their present opinion regarding the research question in multiple stages of the experiment, Silberzahn and Uhlmann were able to track their subjective beliefs about the research hypothesis. The results showed, that the analyst’s beliefs at the registration state of the experiment were not significantly related to the observed effect size reported after completing their analysis. Nevertheless, the analyst’s beliefs changed considerably throughout the analysis process, showing significant relation to their effect estimate and lower bound after the experiment. Again, this is due to the opportunity of learning from other participant’s approaches and the insights gained from looking at the data.

Effects of subjectivity, statistical biases, different design rationales or variability in outcomes can not only be observed in data science. Examples of areas where similar effects could be identified, are software architecture [Jansen et al., 2008, Schubanz, 2014, Gruber and Russell, 1996, MacLean et al., 1991], (mechanical) engineering [Gruber and Russell, 1996, Klein, 1993], UX design [Chung and Goodwin, 1998, Brazier et al., 1997], medicine [Humphreys et al., 2013, Gouttefangeas et al., 2015], ecology [Dieckrüger et al., 1995], or even entrepreneurial activities [Hill and Levenhagen, 1995]. Nevertheless, concrete studies examining this variability in the context of data analysis like the crowdsourcing experiment of [Silberzahn and Uhlmann, 2015], can not be found. Despite having overlaps with research areas studying related concepts for a very long time already, the research of variability in data analysis is still young. Nevertheless, the advent of big data additionally emphasizes the importance of examining and accounting for such variability in data science.

3

Methodology

According to [Thomas, 2006], there are four major approaches for qualitative analyses: A general inductive approach, grounded theory, discourse analysis, and phenomenology. In this study, I follow the general inductive approach for the following reasons:

The classical grounded theory approaches coined by [Glaser, 2017] and [Corbin and Strauss, 1990] are very restrictive in terms of rules and procedures to follow, and often not straightforward for the beginning on [Thomas, 2006, Partington, 2002]. As we are less interested in the study of language in texts, discourse analysis is not a suitable approach for our purposes. When following a phenomenology approach, our findings would rather describe the phenomenon experienced by the participants than the factors accounting for this experience.

In the general inductive methodology, [Thomas, 2006] describes five general procedures, which I apply. Each of them are explained subsequently.

Preparation of data (data cleaning)

In our case, a unit of data consists of a coherent and self-contained sequence of source code executed by a data analyst. I refer to such a unit as a *block*. Each block is provided with descriptive properties which reflect the rationale and reasonings behind the analyst's actions followed within a block (for a concrete description of a block, please refer to Chapter 4). The structure of the descriptive properties originated in the research of design rationale [Schubanz et al., 2014] and design space analysis [MacLean et al., 1991]. Consequently, the format of the data to be coded is semi-structured.

Since the data from each analyst is recorded in the same format, which fits the analysis procedure, no explicit data cleaning has to be made (as advised by [Thomas, 2006]). Essentially, the data gets processed by the user interface used by the coders.

Close reading of text (coding)

In this stage, the coder sequentially peruses each block of an analyst's workflow, and studies the descriptive factors in great detail (Step 1 in Figure 3.1). Following a simultaneous coding method, a coder can assign multiple codes to the same sequence of text (i.e. property of the block) [Saldaña, 2015]. To help coders maintain consistent codes, they were provided with a list of codes they had previously used so far. Each coder also had the possibility to retrieve all the text segments for every corresponding code. This possibility encourages a coder to constantly compare the codes and reflect her reasonings.

A graphical workflow for the entire sequence of blocks, modelled by the analyst at the end of her analysis, provides coders with an overview of the relationship between the blocks.

Embedded in the user interface, a coder can additionally assign explanations (i.e. short memos) for every coded text segment.

Overlapping coding and uncoded text

Coders do not have a strict guideline for how a code should look like. Hence, different coders may apply codes of different granularity, which is why I allowed multiple codes for the same text segment. Additionally, coders may want to provide background information to the relevant string of text, which helps to understand the context for the code [Krippendorff, 2004, Kurasaki, 2000, Fahy, 2001]. As a result, the coded text passages may overlap.

Data analysts do not always provide an explanation for each block property (i.e they do not know any alternative way to go). Hence, answers like "None" or "N/A" have not been coded. Instead, coders were encouraged to apply codes for the explanations *why* analysts provided this answer.

Creation of Categories

After a substantial amount of (initial) codes have been generated, the coders *collaboratively* define low-level categories by summarizing and grouping single codes. This procedure is depicted in step 1 and 2 in Figure 3.1. The "coding paradigm" of conditions, context, strategies (action/interaction), and consequences suggested by the grounded theory approach of [Corbin and Strauss, 1990], can help to possibly create sub-categories and relate them to a category. Additionally, explanations created during coding may implicitly or explicitly comprise the answer to the coding paradigm [Strauss and Corbin, 1998]. Each refined (sub-)category is provided with a memo, which summarizes the coder's thoughts and/or possible connections to other categories. These memos do not only serve as justifications for the (sub-)category, but also to facilitate revision and refinement of the category system.

As [Creswell, 2002] suggests, this newly emerged list of categories should serve as a new organizing scheme for coding. This will help reduce the codes to categories after

a further iteration. In our case, the (updated) list of low-level categories serves as new code scheme for coding in the subsequent iteration. The code scheme will be applied to another subsample of the data, whereby coders can draw on the reasonings of memos when applying codes. Nevertheless, coders can still come up with new codes, which then get assigned to an existing (sub-)category or eventually build the foundation for a new category.

Continuing revision and refinement of category system

Each iteration of coding ends with revising and refining the category system (Step 3 in Figure 3.1). The number of total assigned codes to a category can indicate the distribution of codes among the blocks. A category with only a few assigned codes might be an indication that this category is sparsely grounded. Coders should consider merging this category in a more profound one (i.e. with more codes assigned to it). Miles et al. (2013) describes the process of grouping initial codes into a smaller number of categories as *pattern coding*.

At some point, there will be reached a theoretical saturation where no new (low-level) categories emerge from the codes of new data (Step 4 in Figure 3.1). If the coders believe that each aspect accounting for the inherent variation in data analysis found in the data is captured in a category, the iterative coding is finished. Calculating the percentage of agreement among the coders (i.e. proportional agreement) guarantees, that the coders not only have a common understanding of the code scheme, but also show a high agreement when applying them.

Assessing trustworthiness

There are many ways to evaluate the trustworthiness for models developed in qualitative analysis. Literature proposes several ways to evaluate intercoder reliability or intercoder agreement, and these sometimes even contradict each other [Campbell et al., 2013]. According to Campbell et al., the use of such statistics for qualitative analyses aiming for systematic and rule-guided classification and retrieval of text are less imperative. As a consequence, simple proportion agreement (percentage of agreement among coders) is argued as reasonable approach [Kurasaki, 2000, Campbell et al., 2013]. Also other researchers claim, that looser standards are permissible in exploratory studies [Hruschka et al., 2004, Krippendorff, 2004]. There even exist studies which did not account for coding reliability at all [Campbell et al., 2013].

For this study, I applied different qualitative as well as quantitative measures, in order to guarantee high reliability of the emerged final categories. These measures are explained subsequently.

Independent parallel coding (based on [Thomas, 2006])

Two coders independently develop a set of codes. These two sets are compared and merged into a combined set. When the overlap between the codes are low, the coders have to analyse and discuss in order to develop a more robust set of codes. This procedure also resembles the negotiated agreement approach proposed by [Campbell et al., 2013].

Check on the clarity of categories (based on [Thomas, 2006])

Two additional independent coders (not previously involved in coding) are given the set of initially developed codes together with exemplary data assigned to them. These coders are then given a new subsample of data that has not been coded yet and asked to assign codes to this data (Step 5 in Figure 3.1). As these coders may have come up with codes not present in the category system yet, it would have to be refined and translated in a new code scheme (Step 7 in Figure 3.1). This code scheme is then used for coding another previously uncoded subsample (Step 8 in Figure 3.1). If all involved coders so far agree on the category-system (Step 6 in Figure 3.1), they generate up to eight final top-level categories (Step 9 in Figure 3.1). [Thomas, 2006] states, that if more than eight categories should result from inductive coding, they need to be further generalized, or one needs to overthink the importance of each category again. [Creswell, 2002] reasons, that a small number of categories leads to a better qualitative report, which provides well-detailed information about a few categories rather than general information about many categories.

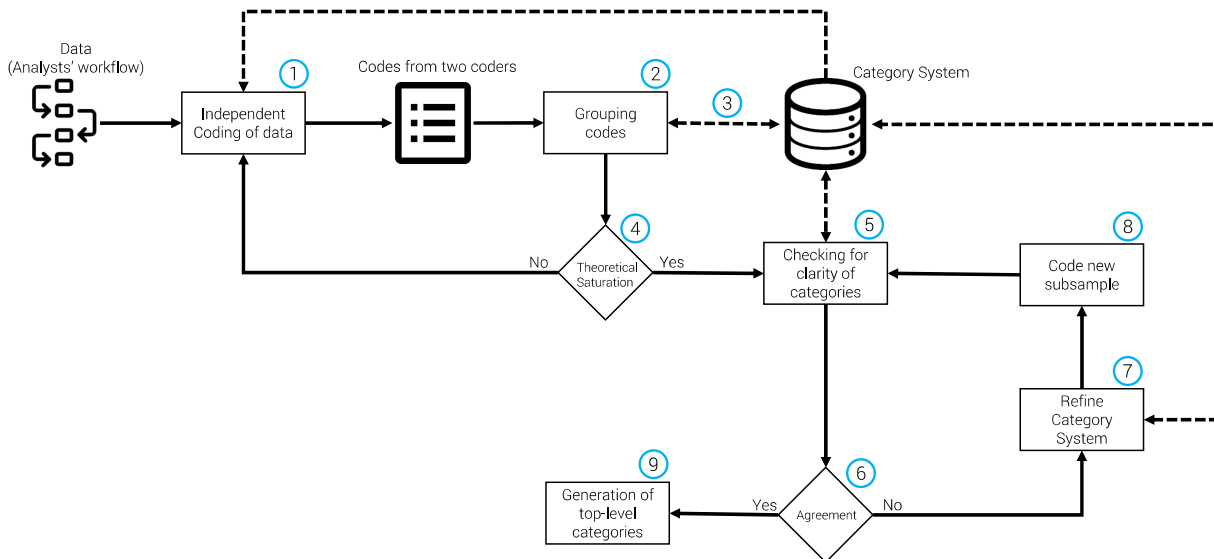


Figure 3.1: Workflow of inductive coding approach

Member checks (based on [Thomas, 2006])

To establish credibility of the final top-level categories, stakeholders (e.g. data analysts whose data were studied) are given a chance to comment these categories, whether they relate to their personal experience they had during the analysis.

Calculation of interrater agreement

To measure the agreement among coders I calculated the proportional agreement and Cohen's Kappa after each iteration (i.e. in Step 4 and 6 in Figure 3.1).

4

Experimental Design

The experiment was divided into three phases: The recruitment of participants for the study, the data analysis performed by the participants, and the phase of compiling and analysing the different data analysis submissions.

In the first section, I explain the experimental design. Section 4.2 describes the setting under which the data analyses were performed. The adapted qualitative analysis of the experimental results is explained in Section 4.3. The last section specifies the quantitative measures applied to the results from the qualitative evaluation.

4.1 Scope of Experiment

I recruited participants for this study via open calls on Twitter, Facebook, forums of psychology interest groups, platforms for collaboration and resource source exchange (i.e. StudySwap¹), R mailing lists, or personal academic contacts. In total, 132 people showed interest in participating in this crowdsourcing experiment, of which 41 carried out all steps involved in the entire data analysis process.

The participants independently analysed a dataset of intellectual conversations from Edge.org². The dataset was constructed in 2015, as part of a first phase in a follow-up project of [Silberzahn and Uhlmann, 2015]. The second phase of this project consists of the study conducted along with this thesis. The dataset contains 123 edge conversations, with 60 attributes related to the conversation, its participants or the textual level of the transcript. The data was collected with a program that downloaded the information from the Edge.org website. Attributes not provided on the website were manually collected by browsing CV's, university or personal webpages, professional networks etc. A detailed procedure of every step followed during the creation of the dataset, along with a full description of every attribute, can be found in Appendix A.1.

¹<https://osf.io/view/studyswap/>

²Edge.org (<http://www.edge.org>) is an online platform for science and technology intellectuals, which share their ideas and insights in different scientific or intellectual topics in open discussions.

4.2 DataExplained

The analyses were performed on a platform called DataExplained³, which got developed as part of this thesis. DataExplained allows for carefully tracking the analysts' path of actions and the rationales for individual decision forks. The platform's core consists of RStudio Server⁴, which provided all participant an own session for their analysis.

Integrated into DataExplained is an initial survey, which asks the participants to state their background and prior beliefs regarding the hypotheses and its related context (for a detailed outline see Appendix A.2).

During the analysis, every command entered in the RStudio interface is recorded along with the respective timestamp. Each such command is referred to as *log*. Recording also intermediate commands is especially helpful, as such logs can possibly reveal additional paths of actions not reflected in the final script.

Whenever the participants believed that a number of logs can be described as a block, they were asked to describe their rationales and reasonings behind the followed actions. A description of a block consists of the following items (further referred to as *block properties*):

- **Title:**
Name which best describes the work the participant has done in the selected logs.
- **Goal:**
Description of the participant's perceived goal from executing the code assigned to this block.
- **Alternatives:**
Description of alternative ways to reach the goal, each with advantages and disadvantages.
- **Reason:**
Justification for chosen option to achieve the block objective.
- **Preconditions:**
Criteria which need to be fulfilled in order to execute this block.

³An introduction video to DataExplained can be found here: <https://www.youtube.com/watch?v=Do3bQ7TvDcM>

⁴<https://www.rstudio.com/products/rstudio/download-server/>

Edit block

Please give a name to the block:

Create different scatter plots

Please shortly explain what you did in this block:

I created a scatter plot to check the correlation between variable X and Y. In addition, I changed the color to improve the design of visualisation.

What where the other (if any) alternatives you considered in order to achieve the results of this block?

Please describe each alternative and explain its advantages and disadvantages. By clicking on "Add another alternative", you can add additional alternatives.

Alternative

Just calculating correlation coefficient Rho

Advantages of this alternative

Using statistical hypothesis testing with a p-value as output

Disadvantages of this alternative

No graphical interpretation possible, and therefore not intuitive at first sight.

Alternative

Dot-Plots

Advantages of this alternative

Good for small sets of data, as well as numerical & categorical data

Disadvantages of this alternative

Hard to construct and interpret

ADD ANOTHER ALTERNATIVE

REMOVE LAST ALTERNATIVE

Why did you choose your option?

I suspected that variable X and Y correlate because ...

What preconditions should be fulfilled to successfully execute this block?

Both, X and Y variables should be calculated based on the raw data using metric A

SHOW DIFF

DELETE BLOCK

LOAD FILES

SAVE

CANCEL

```

set.seed(170513)
n <- 200
d <- data.frame(a = rnorm(n))
d$b <- .4 * (d$a + rnorm(n))
head(d)
library(ggplot2)
ggplot(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
library(ggplot2)
library(ggplot2)
ggplot2(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
install.packages("ggplot")
library(ggplot2)
ggplot(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
ggplot(d, aes(a, b)) +
  geom_point(shape = 16, size = 5) +
  theme_minimal()
ggplot(d, aes(a, b, color = a)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE)
+
  theme_minimal()
d$pc <- predict(prcomp(~a+b, d))[,1]
ggplot(d, aes(a, b, color = pc)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE)
+
  theme_minimal()
ggplot(d, aes(a, b, color = pc)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE)
+
  theme_minimal() +
  scale_color_gradient(low = "#0091ff", high = "#0650e")

```

Figure 4.1: Block of logs

Figure 4.1 depicts a block of logs, along with the described rationales for each block property.

17

The participants were encouraged to create blocks with not more than 30-50 lines of logs, as big blocks might make it difficult to explain the rationale(s) of their actions. A pilot study proved, that an upper threshold of 30-50 logs is near-optimal. Enforcing a lower threshold would distract the analysts in their work, whilst too many logs would lead them to forget important details about single steps involved within a block.

Due to an embedded version control system, participants are able to visually explore changes in their script made between subsequent blocks. They are also given the possibility to navigate in their analysis history, by restoring the state of the workspace at any given point a block was created. These features help them to reproduce the thoughts made during the analysis, even if the corresponding part of code does not exist in the final script anymore.

In a second step of their analysis, the participants are provided with an overview of all blocks. They can fine-tune these blocks, by reassigning the respective logs to other blocks (cf. Figure 4.2). This might be desirable, when a block may not reflect the original perceived course of action anymore.

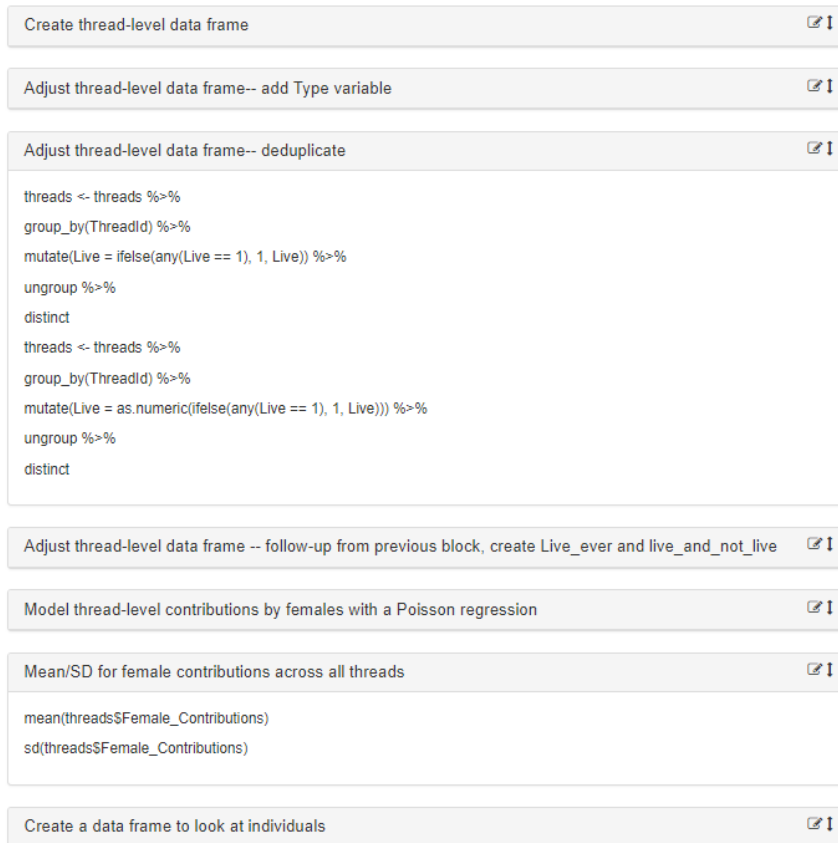


Figure 4.2: Fine-tuning of blocks

The third and last step of the data analysis on DataExplained consists of graphically modelling the workflow followed during the analysis. Initially, each participant is presented a straight chain of blocks, whose actions are executed sequentially. If the analysis procedure followed a specific logic, the graph can be remodelled accordingly. For example, iterative cycles of trying out different approaches for a sub-problem could be modelled as loops in the workflow (cf. Figure 4.3).

In addition to all the data collected by the DataExplained platform, each participant has to submit a second survey after completion of the analysis. In this survey, they were asked to report the results (i.e. effect sizes), the applied methods and a short assessment of their (possibly updated) beliefs regarding the two hypotheses. The outline of the post-analysis survey can be found in Appendix A.3.

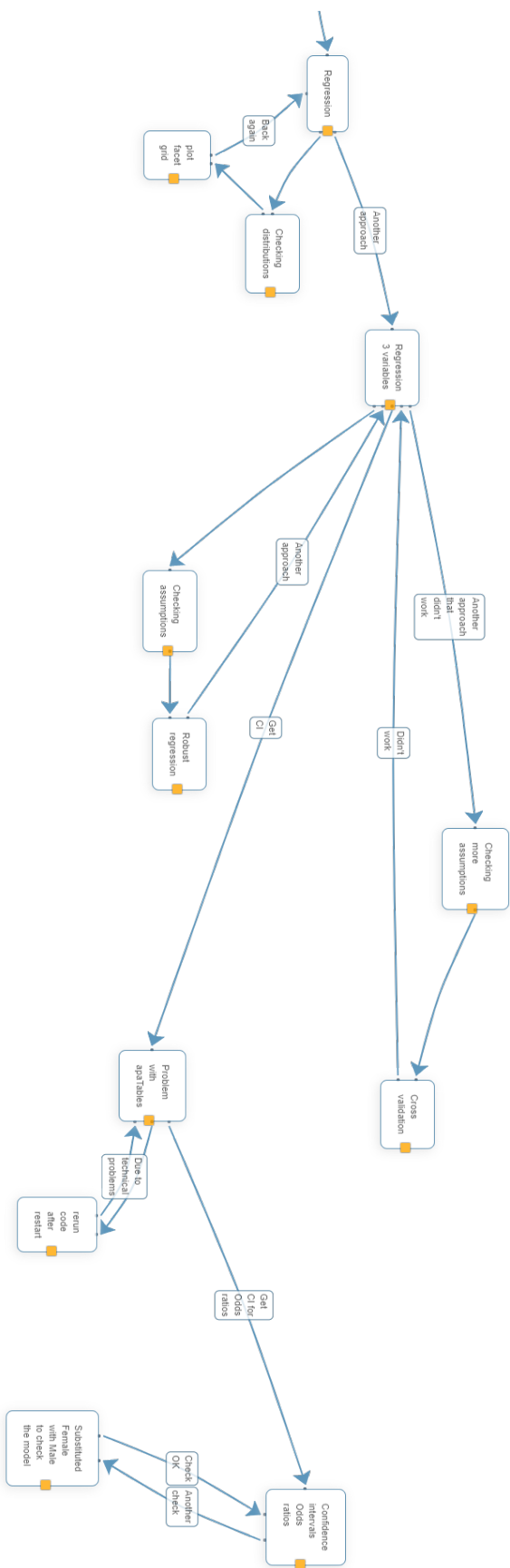


Figure 4.3: Snippet of workflow modelled by a participant

4.3 Qualitative Analysis

In order to extract the factors which drive data analysts to follow a certain analytical path, I qualitatively analyse all the reported rationales in the blocks. For this, I use an inductive coding approach described by [Thomas, 2006]. In the first phase, two coders sequentially navigate through the workflows of the participants and apply qualitative codes to block properties. The codes are directly applied in the graphical user interface of DataExplained, equipped with the necessary functionalities provided explicitly for the coders (depicted in Figure 4.4).

The figure shows a screenshot of the DataExplained coding interface. It is divided into three main panels. The left panel, 'Code block', contains a form for describing a code block with fields for title, goal, alternatives, advantages, disadvantages, reasons, and preconditions. The middle panel displays a list of R code snippets. The right panel, 'Coding', shows a list of applied codes with their reasons and a 'goal' field. The 'Coding' section also includes a 'reason' dropdown, a 'goal' dropdown, and buttons for 'ADD ANOTHER CODE', 'SHOW DIFF', 'CANCEL', and 'SAVE'.

Figure 4.4: Coding interface for a block

If the coders feel that the applied code may not be self-explanatory, they can provide an additional explanation for their choice. Consequently, each coded property consists of a label (name of block property), the applied code(s) with a possible explanation, and the relevant text segment the code(s) refers to. When entering a code, the coders are provided with a list of codes they have previously used so far. This helps them to maintain consistent codes along the analysis and recapitulate previous applications.

After both coders finished coding all the blocks from a predefined subsample, the codes are refined and (potentially) grouped together. In this phase, the coders collaboratively refine their code scheme. Similar codes are merged together, whereas too general codes are split into more expressive codes. For each code, the coders create a short explanation in the form of a memo and provide some examples where this code has been applied. The resulting set of code book (codes along with memo and examples) is then used for the subsequent coding iteration.

When theoretical saturation is reached, and the interrater agreement is high enough (proportional agreement and kappa > 0.7), the code book is presented to two additional

coders. After familiarizing with the codes, all four coders code a new subsample, and verify, if the codes are suitable to describe the rationales perceived by the data analysts. In this phase the code book gets further refined and new subsamples are coded until the agreement is high enough. In order to proceed to the next step, all the coders had to agree on the final code book, and the proportional agreement among all four coders needed to be above 50%.

4.4 Quantitative Analysis

With the category system developed in the qualitative analysis, the participants are split into different clusters. Participants within a cluster have similar proportions of qualitative codes applied in their blocks.

I also analysed the participants on a workflow-level, by mining their processes, and identifying common paths which are traversed during their analysis.

5

Analysis & Results

The analysis of this thesis consists of a qualitative and a quantitative part. In the first part, the reported deliberations for each block are qualitatively analysed, following an inductive coding process. This analysis resulted in a category system of factors, that can explain various decisions as well as similarities between analysts.

In the second part, I applied different quantitative measures, to examine if and how these factors can be accounted for variability in data analysis results.

5.1 Qualitative Analysis

Two coders traversed three initial coding cycles in order to build a sustainable coding scheme. After each iteration, they discussed their discrepancies in the results and refined the codes. At the end of the first cycle, we ended up with 88 codes, which were unitized and renamed to 30 codes. In the subsequent iteration, the coders realized, that some codes already were too general and needed further refinement (i.e. either split the code in more detailed codes, or delete the code entirely, as other codes may already substitute it). The code scheme for the last iteration consisted of 31 codes, which did not need any further refinement after re-coding another subsample. The proportional agreement of the last iteration was 71%, with a kappa measure (Cohens Kappa) of 0.7.

The created code book was then presented to two new coders. All four coders independently coded another subsample of 26 blocks. After collaboratively refining the code book due to low agreement, all the coders coded again another subsample of 53 blocks. Since there were no significant disagreements, there was no need for an additional coding iteration. At the end, the final code book consisted of 30 codes. The detailed mapping of each code to the category is listed in Appendix A.4.

The four coders collaboratively grouped codes together and created a high-level category system with ten categories, listed in Table 5.1.

In addition to these categories, we developed a system with abstract “meta-categories”.

Category	Description
Data	All codes that can be related to the data available for the analyst.
Task	All codes that can be related to the task / hypothesis of interest, as stated in the project description.
Problem	All codes that can be related to the logic of underlying problem.
Knowledge	All codes that can be related to the (prior) knowledge of the analyst.
Belief	All codes that can be related to the (prior) beliefs of the analyst.
Exploratory Data Analysis (EDA)	All codes that can be related to exploratory steps during the analysis.
Confirmatory Data Analysis (CDA)	All codes that can be related to either revision or repetitive steps during the analysis.
Coding skills	All codes that can be related to the source code of the analyst.
Methodology	All codes that can be related to a concrete statistical method.
Insights	All codes that can be related to insights gained by the analyst during the analysis.

Table 5.1: Overview of Category System

Similar to the genome classification for collective intelligence of [Malone et al., 2010], each meta-category provides an answer to different properties driving an analytical approach.

The first meta-category entails all circumstances which are (a priori) *given* and are not personal (i.e. the same for different data analysts). These circumstances may still be interpreted in many ways (e.g. due to new insights or personal beliefs), but cannot be changed in its fundamentals.

In contrast to this category, the second meta-category relates to all *personal* attributes. The interplay between the first two meta-categories is the different treatment and perception of the given circumstances. An example for this could be a certain (perceived) understanding of the data due to the professional background or personal experiences in this area. This interplay of mental models and given data is also discussed in [Grolemund and Wickham, 2014]. The third meta-category is named *Analysis*, as it contains all the codes related to direct actions/methods performed during the data analysis. These steps can have an exploratory or confirmatory character. The methods chosen to achieve the desired results thereby vary among different analysts.

After an analytical step, an analyst gets new insights, which interact with her personal cognition of the problem and the reality (i.e. cognitive sensemaking process).

We can thereby differentiate between two general data analysis approaches: Exploratory data analysis (EDA) is the process of exploring the data and trying to understand the logic of the problem and summarizing its main characteristics. Confirmatory data analysis (CDA) refers to the analytic choices to confirm the emerged models (i.e. systematically assess the strength of the evidence) in an iterative way [Hoaglin, 2003]. As an example, assume that an analyst wants to find out the relation between two variables of interest. She therefore applies different methods (e.g. runs a correlation or plots different diagrams), in order to understand this relationship on a subset of the data (EDA). Once she perceives to have understood the meaning of these variables (i.e. made sense of the data/problem), she wants to confirm her insights and fits a linear model on another subset of the data (CDA). This interplay between exploration and confirmation can be observed in various stages of a data analysis, since the insights of CDA may not necessarily be in line with the findings of the exploratory phase. In that case, further exploratory steps might be necessary. With multiple iterations of EDA and CDA, the analysts continuously refine their analysis. This cycle ends, once an analyst reports her final findings with regard to the stated hypotheses.

Figure 5.1 provides an overview of the (meta-)categories which resulted from the qualitative analysis. In order to have a consistent qualitative result for each participant, the two initial coders recoded the entire data from scratch with the final coding scheme.

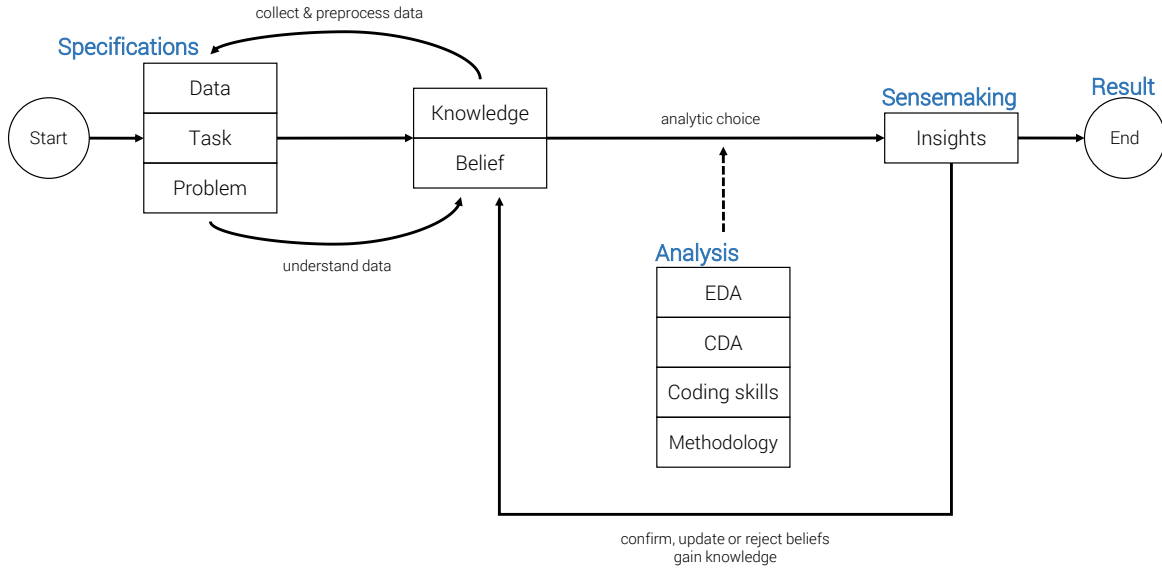


Figure 5.1: (Meta-)categories in a data analyst’s workflow

5.2 Quantitative Analysis

The quantitative analysis can be grouped into two parts: In the first part, I tried to find similarities and differences among the *data analysts*. For this, I made use of the qualitative factors identified in Subsection 5.1, as well as the analyst’s personal or professional qualities reported in the pre-survey. The second part lies the focus on the different procedures within an analyst’s *workflow*.

Quantitative Analysis of Data Analysts

In this part of the analysis, I clustered participants by the proportion of each qualitative code applied within their workflows. I used k-means as clustering algorithm, with $k=4$. The marginal improvement of explaining the variance among all clusters dropped below 2% for $k=5$, for which reason four clusters were chosen. The marginal improvement for $k=4$ was 8%. The codes “revision of findings” and “belief” were deleted for this analysis, as they were not applied enough times to be accounted in a cluster analysis. The assignment of clusters for each participant can be found in Appendix A.5.1.

Quantitative Analysis of Workflows

To get a deeper understanding of the typical steps traversed during a data analysis, I compared these steps on a block level. To be able to make comparisons among blocks, I clustered them by the qualitative codes applied to each of them during the qualitative analysis. Thereby, all the codes of a block were reduced to a distinct set before the clustering was applied (e.g. if a block consists of the code “exploratory” and twice

Cluster	Code
Visualisation	visualisation
Expertise	expertise
Problem	action driven by, insight,insight realization, intuition about the problem
Knowledge	perceived understanding of reality, personal assumption, personal knowledge, task constraint
Exploratory	exploratory, method preference
Personal	code quality, complexity constraint, confirmatory measure, data constraint, data quality, error fixing, interpretability constraint, perceived course of action, personal interest, personal preferences, uncertainty about the method, uncertainty about the problem
Preprocessing	preprocessing
Problem	feature engineering, perceived understanding of the problem
Constraints	effort constraint, methodological constraint

Table 5.2: Clusters of Blocks

the code “preprocessing”, the distinct set would be “exploratory and preprocessing”). I deleted the codes “revision of findings” and “belief” for the same reasons as the previous subsection. Additionally, I excluded the codes which were applied to block properties of alternatives, since these codes do not necessarily reflect the actual action followed within a block.

I used k-means as clustering algorithm, with k=9. The marginal improvement for k=9 was 7%, whereas for k=10, it dropped below 1%.

The composition of codes for each cluster is listed in Table 5.2.

In a second step, I mined the workflow, using the basic heuristic process mining algorithm from Disco¹. In our case, each user represents a case, whereby the cluster for the block can be seen as a grouped action.

¹<https://fluxicon.com/disco/>

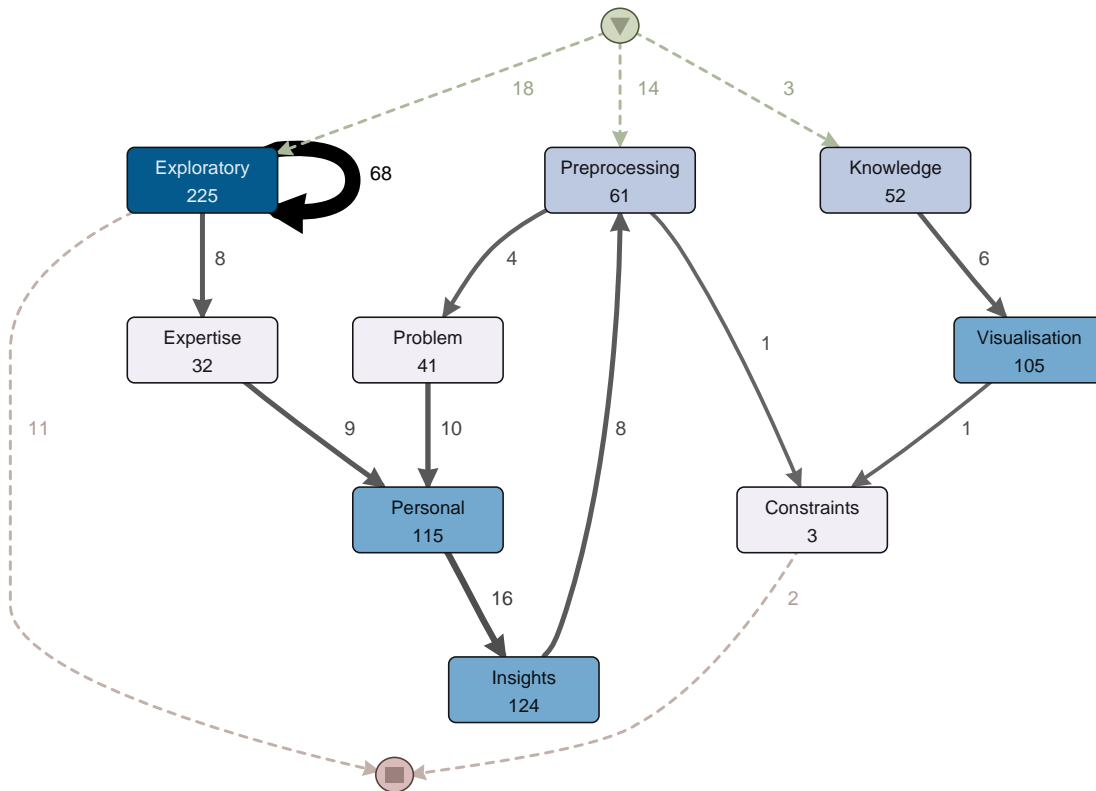


Figure 5.2: Resulting workflow from process mining

The resulting process graph is modelled in Figure 5.2. The green icon on top indicates the start of the workflow, whereas the end is indicated by the red icon on the bottom. The numbers inside a rectangle (=cluster) represent the total amount this cluster was traversed. The darker the colour of a cluster, the more often it appeared in the analysts' workflows. The labels and thickness of the arrows depict the number of times these clusters were processed consecutively. For reasons of clarity and comprehensibility, only the most dominant paths are modelled in the graph.

6

Discussion

The resulting clusters of analysts were significantly correlating with the proportion of codes they shared among each other. As an example, one participant showed a correlation higher than 0.8 with nine other participants. After running the k-means algorithm, all of these ten participants were clustered in the same cluster (Cluster 1). From all participants, these ten had the highest proportion of the code “exploratory” (reaching from 22% to 38%).

Taking a look at the demographics of the participants in this cluster, most of them hold a PhD degree (70%), with an average data analysis experience of 6.8 years. The fact that only one participant in this cluster performs data analysis less than 2-3 times a week (namely once a month), confirms that this cluster consists mainly of data analysts with high expertise.

Cluster 4 contains only one user, having just two blocks. The analysis of this participant was not fully done, as the only code applied in the workflow was “preprocessing”.

Cluster 3 consists of three participants, each with a low amount of blocks (2, 3 and 11) compared to the average blocks of all participants (21.13). The majority of codes applied by the participants in this cluster are related to a certain constraint (i.e. “effort constraint”, “methodological constraint” and “task constraint”). Unlike in cluster 4, these participants actually reported a result for both hypotheses, but most likely in a way with minimal effort.

The results from the workflow analysis reflected my personal expectations about the typical sequence of actions during a data analysis. In most cases, an analyst starts with reading (Cluster *Preprocessing*) and exploring the data (Cluster *Exploratory*). Having the necessary expertise, analysts then create new features to better understand the logic of the underlying problem (Cluster *Expertise* and *Problem*). During the qualitative analysis, we could see, that analysts often have a plan in mind, which they want to follow. Codes related to the sensemaking process (predominantly present in the cluster *Insights*) are typically not applied early in the workflow. Especially the code “action driven by inside” can be a sign for the start of a new iteration. This iteration is modelled as a loop in the process mining workflow.

The clusters most often traversed at the end of an analyst’s workflow are *Constraints* and *Exploratory*. This can be related to the fact, that an analyst either tries out different models for the result (Exploratory), or collects the required measures for the report (i.e. the code “task constraint” inside the cluster Constraints).

Like the study of [Silberzahn and Uhlmann, 2015], I could observe surprisingly large variability in the results of the different data analyses. A major reason which certainly accounted for this variability, was the fact, that researchers operationalized variables themselves, in addition to choosing their own analysis approach (i.e. as a real research team would do in a standalone project).

But what implications do these findings have for scientific data analyses in general? Different defensible operationalizations of variables and analytic choices may lead to different findings. It is important therefore, to not only rely on a single analytic report, but to compare the results of multiple teams or organizations before making any major strategic decisions. Consequently, scientists should be aware, that findings from complex datasets should be interpreted with certain caution. A full disclosure of the analytic procedure is therefore inevitable, in order to allow for potential re-analysis and replication of the study.

The set of factors developed in this thesis can help to reveal the inherent variability in data analysis. They enable to identify relationships in analytical choices, and allow for comparing different paths inside the garden of forking paths faced in data science. These factors, however, are only a possible means of explaining this variability, but ultimately do not resolve it. There is still a need for a socio-technical system, which can control the underlying processes in data analysis, and identify potential decision forks.

7

Limitations & Future Work

For my study, I recorded all the commands an analyst has produced during her analysis. As stated in Section 4.2, I believe that logs not present in the final script can possibly reveal additional paths of actions. Some participants had however difficulties to transition between exploratory behaviour and modelling decisions, given the nature of recording their logs. One participant claimed that this setting pushed him to be more linear in his approach (i.e. plan more beforehand, to prevent mistakes which needed an explanation why the approach was changed). One possible approach to minimize this issue would be to not force the analysts to make blocks. They could then select any logs to create a block, and ignore the perceived “unnecessary logs”. However, latter would most certainly lead to fewer blocks, and hence, a less deep understanding of the analyst’s rationales.

Another main challenge in the process of DataExplained was the definition of block properties. I tried to keep the amount of questions as low as possible, while still eliciting as much and accurate information about a block as possible. During the inductive coding, I could ascertain that the properties *goal* (“what”) and *reason* (“why”) mostly contained the core information about the block. However, asking for preconditions to be fulfilled in order to run a block, was not always suitable in the respective context. Especially in exploratory actions, it does not make sense to ask for given preconditions, as they are either trivial or non-existent.

If there were any mechanism which could determine the kind of action performed in a block, one could possibly ask more tailored questions in the future. This way, more specific information about a block could possibly be collected, which would allow to get a more detailed view of what *specific* rationales ultimately guided an analyst in her analytical choices. Furthermore, this would facilitate a more solid base for developing a socio-technical system, which could automatically detect decision forks.

When performing a qualitative analysis, researchers do have to understand the meaning behind someone else’s narrative or an observation. The interpretation of meaning however always requires a certain level of inference. This subjective part of qualitative

analyses often provides a target to its advocates [Madill et al., 2000]. Many approaches to deal with this limitation have been proposed [Graneheim and Lundman, 2004, Hallgren, 2012, Burla et al., 2008]. Reliable findings from qualitative analyses however do not necessary guarantee validity [Krippendorff, 2004]. Two coders may come up with a highly reliable coding-scheme, which results in great agreement among the results. These results may however still be invalid if the category system is too artificial to reflect the objective reality that it tries to capture [Alonso and Volkens, 2012]. Getting rid of *any* subjectivity in qualitative analyses is however impossible.

Despite these limitations, it is my belief, that the results of this study are very sustainable and will be of great use for future research. In alignment with this thesis, there will be a third study, which will seek to confirm the explanatory factors identified in this work.

8

Summary & Conclusions

To elicit the factors which lead to different analytical decision, I looked at the different actions performed by data analysts along their path of analysis. I tried to understand the driving factors for each of those actions by examining the rationales behind data analyst's choices along their analytic paths. To do so, multiple coders iteratively developed a code book with 30 codes, which can be grouped into eleven categories and four meta-categories.

The quantitative analysis demonstrated, that the qualitative codes applied to an analytical workflow provide a good means for clustering participants with different expertises and personal backgrounds. These codes also served as the base for modelling a sequence of abstract actions typically performed in a data analysis.

As stated in this thesis, effects of subjectivity, statistical biases, different design rationales or variability in outcomes, can not only be observed in data science (i.e. mechanical engineering or UX design). The developed system of categories and meta-categories may thereby also allow to examine workflows outside the field of data analysis.

It is my belief, that this study was a great success, and its findings have a great potential for further studying the concerns of robustness and reliability of statistical findings. The field of research around exploring the variability in the context of data science is still young. My findings can therefore serve as a solid foundation for a deeper understanding of the statistical crisis faced nowadays.

9

Technical Documentation

This chapter includes a technical documentation of the DataExplained platform created to carry out the experiments of this thesis. It should serve as an overview of the architecture as well as a guideline for setting up the necessary infrastructure.

9.1 Architecture

DataExplained is a web-based application which is built on the MEAN¹ stack. The application is running on an Amazon EC2 instance with RStudio Server² installed.

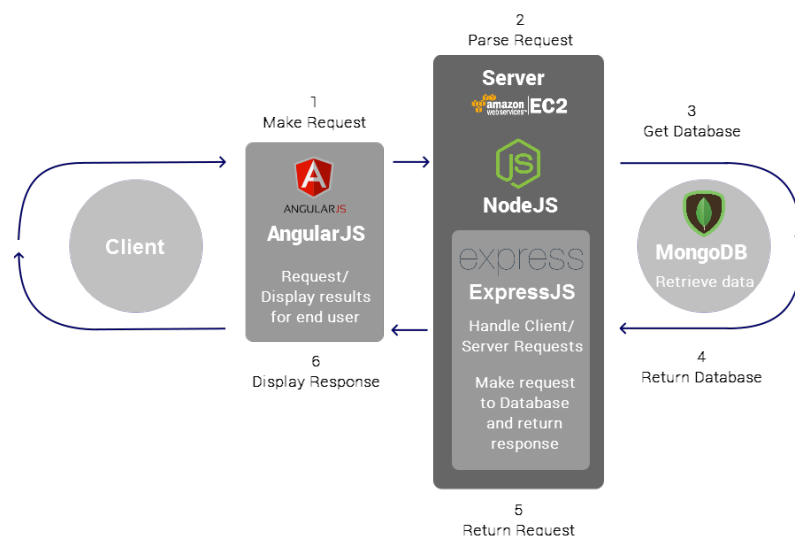


Figure 9.1: Architecture DataExplained; own figure based on [Team In India, 2017]

¹MEAN is a JavaScript software stack used for building dynamic web applications. It builds on the components of MongoDB, Express.js, Angular and Node.js.

²<https://www.rstudio.com/products/rstudio/download-server/>

A big advantage of the MEAN stack is, that both the client and the server are written in JavaScript (also known as full-stack JavaScript application). Javascript objects can easily be transformed to JSON objects, which can easily be persisted in MongoDB.

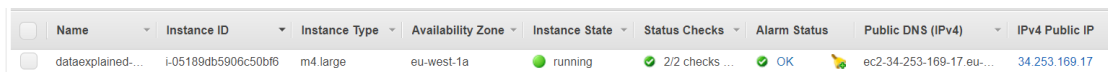
DataExplained makes use Grunt (build tool), Bower (package manager for web dependencies), and NPM (package manager for nodejs dependencies).

9.2 Technical Setup

This section explains how the remote EC2 server instance is created and setup. Additionally, instructions for the configurations on the local (developer's) machine are provided, in order to connect and commit changes to the server. Commands executed on the server are preceded with a '\$' sign. Instructions for the local machines are given for Windows systems (Windows 10). Respective configurations for other operation systems may differ. Also links for websites of different components may have changed by the time this thesis was written.

9.2.1 Setup EC2 instance

In a first step, a new EC2 instance is created on aws.amazon.com. As operation system I chose Ubuntu (Ubuntu Server 16.04 LTS). For performance reasons, the respective region where the instance is hosted can be chosen. For DataExplained I selected "EU (Ireland)". After the instance was created it is listed in the overview, illustrated in Figure 9.2.



Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP
dataexplained-...	i-05189db5906c50bf6	m4.large	eu-west-1a	running	2/2 checks ...	OK	ec2-34-253-169-17.eu...	34.253.169.17

Figure 9.2: EC2 instance

During the setup, a key of the instance (e.g. "dataexplained.pem") gets generated. Save it on your local machine under `%HOME%/.ssh/dataexplained.pem`. This key serves to connect to the instance via SSH (c.f. Subsection 9.2.1). For this, and in order to install external packages on the server in a later step, we have to modify the instance's security group in the AWS Management Console. For the respective rules, please see Figure 9.3 and Figure 9.4.

Security Group: sg-8effe6e8

Description Inbound Outbound Tags

Edit

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ
HTTP	TCP	80	0.0.0.0/0
HTTP	TCP	80	::/0
Custom TCP Rule	TCP	8888	0.0.0.0/0
Custom TCP Rule	TCP	8888	::/0
Custom TCP Rule	TCP	8000	0.0.0.0/0
Custom TCP Rule	TCP	8000	::/0
SSH	TCP	22	0.0.0.0/0
Custom TCP Rule	TCP	8787	0.0.0.0/0
Custom TCP Rule	TCP	8787	::/0
Custom TCP Rule	TCP	27017	0.0.0.0/0
Custom TCP Rule	TCP	27017	::/0

Figure 9.3: EC2 instance Inbound Security Rules

Security Group: sg-8effe6e8

Description Inbound Outbound Tags

Edit

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Destination ⓘ
HTTP	TCP	80	0.0.0.0/0
HTTP	TCP	80	::/0
All traffic	All	All	0.0.0.0/0
All traffic	All	All	::/0

Figure 9.4: EC2 instance Outbound Security Rules

Connect via SSH

To connect via ssh, use the following command (replace path accordingly):

```
ssh -i /path/to/key/dataexplained.pem ubuntu@ec2-34-249-31-191.eu-west-1.compute.amazonaws.com
```

In order to make things easier in future, the following settings have to be made:

Local:

Create/update the config file under `%HOME%/.ssh/config` (with respective hostname of your EC2-instance):

Host dataexplained
Hostname ec2-34-249-31-191.eu-west-1.compute.amazonaws.com
User ubuntu
IdentityFile ~/.ssh/dataexplained.pem

EC2:

Add your personal public ssh-key of local machine on the server:

```
$ cd ~/.ssh  
$ nano authorized_keys
```

As of now, you can connect to our remote EC2 instance via terminal:

```
ssh dataexplained
```

Configure EC2

Connect to the server via ssh and enter the following prompts:

```
$ sudo apt-get update  
$ sudo apt-get install -y python-software-properties python g++ make  
$ curl -sL https://deb.nodesource.com/setup-7.x | sudo -E bash -  
$ sudo apt-get update  
$ sudo apt-get install nodejs  
$ sudo apt-get install build-essential  
$ sudo apt-get install git  
$ sudo apt-get install npm  
$ sudo npm install cross-spawn  
$ sudo npm install forever -g  
$ sudo npm install pm2 -g
```

Create a bare Git repository on the server (REPO_NAME is the name for the repository you want to use):

```
$ cd ~/  
$ mkdir REPO_NAME  
$ cd REPO_NAME  
$ git init --bare
```

Create a post-receive git-hook which automatically restarts the server once a new version was committed:

```
$ cd REPO_NAME/hooks/
```


9.2 TECHNICAL SETUP

```
$ touch post-receive
$ chmod +x post-receive
$ nano post-receive
```

Paste the following content:

```
#!/bin/sh
GIT_WORK_TREE=/home/ubuntu/www
export GIT_WORK_TREE
git checkout -f
cd $HOME/www
./start.sh
```

Create directory for applications content and create start script:

```
$ cd ~/
$ mkdir www/
$ cd ~/www
$ touch start.sh
$ chmod +x start.sh
$ nano start.sh
```

Paste the following content:

```
# this file is execute by post-receive hook every time a Git commit is made:
pm2 kill
export GITHUB.USER=<your github username here>
export GITHUB.SECRET=<your github password here>
export GITHUB.TOKEN=<your github token here>
sudo service mongod start
pm2 start apps.json
sudo chmod -R 777 /home/ubuntu/.pm2
```

Redirect all traffic from port 80 to 8080:

(This command has to be re-executed everytime the server was shut down or restarted!)

```
sudo iptables -t nat -A PREROUTING -p tcp --dport 80 -j REDIRECT --to-ports 8080
```

As the remote Git-repository is now configured, we need to add it on the client-side (local machine) configuration.

Create git repository in distribution folder of the application (dist)³ and add/edit the “config” file (within the newly created “.git” folder): `git init`

Paste the following content in the config file:

```
[remote "AWS.production"]
  url = ssh://ubuntu@YOUR-IP/home/ubuntu/REPO_NAME/
  fetch = +refs/heads/*:refs/remotes/REPO_NAME/*
  puttykeyfile = C:\Users\YOUR-USER\.ssh\dataexplained.pem
```

From now on, the client can commit and push changes to the remote EC2 instance. This will trigger the post-receive hook, moves the application’s content in the server application’s folder (www) and restarts the server.

Attention: If new node-packages are added to the application (in the `packages.json`), you have to manually run “`sudo npm install`” in the `~/www` directory.

If the application makes use of environment variables, you may consider to permanently add them to the server in order to access them (even if the start-up script would fail for some reasons).

On the EC2-instance, the global environment variables are stored in “`/etc/environment`”.

The file can be edited with “`sudo nano /etc/environment`”.

To see the newly created variables, you have to reconnect the machine via ssh and run `printenv`.

9.2.2 Setup MongoDB

The tutorial I followed for installing mongoDB on an ubuntu machine (our EC2 instance) can be found on: <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-ubuntu/>. It consists merely of the following steps (please refer to the link to guarantee the newest version gets installed):

```
$ sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv
0C49F3730359A14518585931BC711F9BA15703C6
$ echo "deb [ arch=amd64,arm64 ] http://repo.mongodb.org/apt/ubuntu xenial/mongodb-org/3.4
multiverse" | sudo tee /etc/apt/sources.list.d/mongodb-org-3.4.list
$ sudo apt-get update
$ sudo apt-get install -y mongodb-org
```

To start mongoDB we run:

³For instruction how to build application, please see Section 9.3

```
$ sudo service mongod start
```

To verify that MongoDB is running and to check the logs, we can inspect the contents of `"/var/log/mongodb/mongod.log"`. The running port is configured in `"/etc/mongod.conf"` and is set to 27017 by default.

9.2.3 Setup RStudio Server

The tutorial on how to install RStudio server can be found on: <https://www.rstudio.com/products/rstudio/download-server/>. It consists merely of the following steps (please refer to the link to guarantee the newest version gets installed):

```
$ sudo apt-get install r-base
$ sudo apt-get install gdebi-core
$ wget https://download2.rstudio.org/rstudio-server-1.0.136-amd64.deb
$ sudo gdebi rstudio-server-1.0.136-amd64.deb
```

To allow RStudio Server to run in an iframe, you have to add the following configuration in `"/etc/rstudio/rserver.conf"`:

```
www-frame-origin=anyline
```

For the sessions in RStudio its beneficial to enable automatic saving of the workspace and set the default workspace environment. You can do so by adding following the following configuration in `/etc/rstudio/rsession.conf"`:

```
session-save-action-default=yes session-default-working-dir= /rstudio-workspace
```

To manually stop, start, and restart the server you use the following commands:

```
$ sudo rstudio-server stop
$ sudo rstudio-server stop
$ sudo rstudio-server restart
```

Each time you change a configuration file, you have to either restart Rstudio Server with the commands listed above, or you can run:

```
$ sudo rstudio-server verify-installation
```

To add R-packages globally (available for all users in their workspace without prior

installation) you can follow these steps (example shown for package "readr"):

```
$ cd ~
$ sudo wget https://cran.r-project.org/src/contrib/readr_1.0.0.tar.gz
$ sudo R CMD INSTALL -l /usr/lib/R/library readr_1.0.0.tar.gz
$ sudo rm -r readr_1.0.0.tar.gz
```

If the package has dependencies of other packages which are not installed on the server yet, you need to install these packages first.

9.3 Run, Build, Deploy

This section serves as a guidance to locally run DataExplained for development, as well as build and deploy a new version to the server.

9.3.1 Prerequisites

- Node.js and npm (Node ^4.2.3, npm ^2.14.7)
- Bower (npm install --global bower)
- Grunt (npm install --global grunt-cli)
- MongoDB daemon running (default port 27017) with mongod

9.3.2 Development

1. Run `npm install` to install server dependencies.
2. Run `bower install` to install front-end dependencies.
3. Run `mongod` in a separate shell to keep an instance of the MongoDB Daemon running.
4. Run `grunt serve` to start the development server. It should automatically open the client in your browser when ready.

9.3.3 Deployment

1. Run `grunt build` for building the application. The built application is then contained in the "dist" folder.
2. Make sure that the newly created files under `dist/client/app` are all added to git.
3. Push the changes to the remote git repository on the server ("AWS_production" configured in Subsection 9.2.1)

4. Due to the git-hook configured in Subsection 9.2.1, the server automatically replaces the application content and restarts. This does not take more than a few seconds.
5. Additionally you may want to push the changes to the git repository of DataExplained.

9.4 Database Backup

To backup the database you have to run the following command:

```
$ mongodump --out /home/ubuntu/backup/
```

You can restore your backed up dump with ("dataexplained" refers to the name of the database):

```
$ mongorestore -d dataexplained /home/ubuntu/backup/
```

9.4.1 Cronjobs

In order to periodically run jobs on the server (i.e. backup the database), you can define cronjobs by executing

```
$ crontab -e
```

The following cronjobs are recommended for DataExplained:

- Hourly backup of the database via `mongo_backup.sh` script. (Note, this script additionally uploads the compressed backup to Amazon S3.)

```
$ 00 0-23 * * * /bin/bash /home/ubuntu/backup/mongo_backup.sh
```
- Remove database backups on server older than 7 days to save resources. Executed once a day.

```
$ 01 05 * * * /usr/bin/find /home/ubuntu/backup/rationalecap/ -mtime +7 -exec rm \;
```
- Run script which checks whether the database is connected or not. If not, the script will reconnect it. Executed every five minutes.

```
$ */5 * * * * /bin/bash /home/ubuntu/www/mongocheck.sh >/dev/null 2>&1
```
- Send metrics to Amazon AWS which can be fetched via the CloudWatch service. Executed every half-hourly.

```
$ 30 * * * * /aws-scripts-mon/mon-put-instance-data.pl --disk-path=/
--disk-space-util --disk-space-used --disk-space-avail --from-cron
```

A

Appendix

A.1 Description of Dataset

Our dataset build started with collecting information from the Edge.org on all of the conversations and annual questions. We built a program that downloaded the information from the website, including the year, title, link to, and type of the conversation, as well as the text itself and who said it. Two independent coders then coded gender of the contributors based on their profile picture on Edge.org, or, if that was not available, pictures and pronouns on other reputable websites. We then manually collected information on the job title, workplace, and PhD by finding CVs, university webpages, news articles, personal websites, and Linked-In profiles. We wrote a program to collect the US News and World Report International Rankings and the Shanghai Rankings and manually gathered the rankings from the National US News and World Report Rankings. We then ran the text through the LIWC program. Finally, we calculated the rest of the variables (such as male contributors, previous contributions, etc.) based on the data we already had collected.

The descriptions below include the name in the full version of the dataset and the shortened name used in the dataset for older software.

- **Conversation Level:**

- **Year:** the year when it took place
- **Title:** the title of the conversation. For example: "What Scientific Idea Is Ready For Retirement?"
- **Link:** a link to the conversation
- **Type:** 1 for annual question, 2 for conversation
 - * Edge does an **annual question** every year; some examples are "What scientific idea is ready for retirement?" and "What will change everything?" People then write in with their answers. So all of the text is written and asynchronous

- * What Edge refers to as a **conversation** can actually be multiple things. Some of these are written essays by a single person, some are transcripts of a speech, and some are transcripts of a conversation (either between two or more guests or an interview).
- **ThreadID (ThrdID)**: a unique identifier for each conversation/annual question (between two or more people)
- **MaleContributions (Mcontr)**: the **number of times** a man speaks in a specific conversation, it does not always equal the number of unique men in a conversation (see below)
- **FemaleContributions (Fcontr)**: the **number of times** a woman speaks in a specific conversation, it does not always equal the number of unique women in a conversation (see below)
- **FemaleParticipation (Fpart)**: simply femalecontributors/(number of total contributions); the percentage of comments that are made by a woman
- **NumberAuthors (NumAut)**:
 - * For the annual questions, this equals 0; because the website is the author of the question, everyone is considered commentators
 - * Otherwise, this is the total number of times people contribute to the main body of the text, rather than people who just comment. For example, in <http://edge.org/conversation/how-democracy-works-or-why-perfect-elections-should-all-end-in-ties>, there are multiple people commenting on the post, but W. Daniel Hillis is the only author and only speaks once (as it is an essay). So NumberAuthors is "1." If two people each spoke five times in a dialogue, NumberAuthors would be "10."
- **DebateSize (DebSiz)**: number of text pieces in a conversation; it is the sum of female and male contributions
- **Live**: whether the text piece was transcribed or written; it is 0 if it is written (either an essay or a comment on a piece) and 1 if it was part of a live conversation or speech that was later transcribed. Here are the types of text and how they would be classified:
 - * **A single author essay** (live = 0 because it is written):
<http://edge.org/conversation/the-evolved-self-management-system>
 - * **A single author speech** (live = 1 because it was spoken and later transcribed):
<http://edge.org/conversation/cities-as-gardens>
 - * **A live conversation**, either between multiple people or in an interview format (live = 1 because it was spoken and later transcribed):
<http://edge.org/conversation/japan-inc-meets-the-digerati>
 - * **Online Comments** on any of the three types above (live = 0 because it was written)

- * **The annual question (Type = 1):** live = 0 because these were all written and submitted.
- **UniqueContributors (UContr):** **UniqueMaleContributors** + **UniqueFemaleContributors**
- **UniqueMaleContributors (UMContr):** the number of unique male contributors
- **UniqueFemaleContributors (UFContr):** the number of unique female contributors
- **UniqueFemaleParticipation (UFPar):** the percentage of unique female participants; **UniqueFemaleContributors** divided by **UniqueContributors**
- **Participant Level**
 - **Id:** the unique identifier of the contributor
 - **Id_num:** the unique identifier of the contributor as text (this is typically the format of first name_last name)
 - **Role:** Either author (=1) or commentator (=2)
 - **Name:** name of the commentator
 - **TwoAuthors (TwoAutrs):** some of the edge comments are written by two people. In this case, we duplicated the row and kept the text level and conversation level information the same and had one author per row. This variable is 1 if this text was written by two people and 0 otherwise.
 - **Female:** the commentator is male = 0, the commentator is female = 1
 - **Male:** the commentator is female = 0; the commentator is male = 1
 - **Academic (Acad):** 1 = the person is in academia, 0 = they are not
 - **Limited Information (LimInfo):** equals 1 if we could only find limited information about the person (e.g. they commented in 2013 but we only have their job title from 2012), 0 otherwise
 - **Job_Title (JobT):** The job title of the commentator
 - **Job_Title_S (JobTS):** This is a simplified list of job titles (e.g. we have "Eugene Higgs Professor" in Job.Title but "Chaired Professor" in Job.Title.Collapsed)
 - * Chaired Professor
 - * Professor
 - * Associate Professor
 - * Assistant Professor
 - * Non-Tenure-Track Faculty
 - * Postdoctoral Researcher

- * Graduate Student
 - * Academic Leadership (Dean, Vice President, etc.)
 - * Researcher
 - * Artist/Author/Editor/Writer
 - * Director
 - * Founder
 - * Other
 - * Top Management and Founder
 - * Top Management
 - * Entrepreneur
 - * Not Available
- **Job_Title_S_num (JobTSn)**: Job_Title_S as numbers instead of text
 - **Department (Dept)**: what academic department someone is in
 - **Department_S (DeptS)**: a simplified version of all the departments (e.g. while John Smith’s Department is ”Experimental Physics,” his Department_S is ”Physics”)
 - * Physics (Phy)
 - * Anthropology (Ant)
 - * Earth Sciences (ES)
 - * Biology (Bio)
 - * Psychology (Psych)
 - * Journalism, media studies and communication (JMS)
 - * Medicine (Med)
 - * Philosophy (Phil)
 - * Space Sciences (SS)
 - * Linguistics (Lin)
 - * Computer Sciences (CS)
 - * Engineering (Eng)
 - * Arts (Arts)
 - * Business/Management (Bus)
 - * Environmental Studies and Forestry (ESF)
 - * Sociology (Soc)
 - * Mathematics (Math)
 - * Asian Studies (AS)

- * Education (Educ)
- * Political Science (PS)
- * Economics (Econ)
- * Systems Science (Sys)
- * History (Hist)
- * Music (Musc)
- * Chemistry (Chem)
- * Archeology (Arch)
- * Architecture and Design (ArchD)
- * Law (Law)
- * Zoology (Zoo)
- * Literature (Lit)
- * Divinity (Div)
- **Department_S_num (DeptSn)**: Department_S as numbers instead of text
- **Discipline (Disc)**: this groups academic departments into disciplines
 - * Natural Sciences (NS)
 - * Social Sciences (SocS)
 - * Professions (Prof)
 - * Humanities (Hum)
 - * Formal Sciences (FS)
- **Workplace (Workpl)**: where someone works; some people are self-employed
- **HavePhD (PhD)**: equals 1 if they have a phd, 0 otherwise. It is 1 even if someone earns a phd after they comment (e.g. John Doe comments in 2000 and earns his PhD in 2012; his comment in 2000 will still have HavePhD = 1)
- **PhD_Field (PhDF)**: what field people got their PhD in
- **PhD_Year (PhDY)**: what year they got their PhD
- **PreviousContributions (PrContr)**: how many times **before this year** they have made contributions. So if John Doe only talked three times in one conversation in 2012 and one time each in two conversations in 2014 (and never made any other comments), this will be 0 for his comment in 2012 and 3 for both his comments in 2014.
- **ContributionsThisYear (ContrTY)**: how many times they contributed this year; even if they only participated in one conversation, if they spoke 40 times in that conversation, this variable will be 40.

- **ThreadsThisYear (ThrTY)**: how many threads they participated in this year; thus if John spoke in two threads in 2014, one twenty times and one once, this would equal 2 in 2014, while ContributionsThisYear would equal 21 for 2014.
- **PreviousThreads (PreThrd)**: how many threads they participated in **before this year**. So, if John contributed for the first time twice in one thread in 2000, once each in two different threads in 2004, and once in 2014, this would be 0 for 2000, 1 for 2004, and 3 for 2014 (and for PreviousContributions it would be 0 for 2000, 2 for 2004, and 4 for 2014).
- **AuthorandCommentator (AutAndCom)**: if, for the same piece, someone is both an author and a commentator, this is 1 for that person for that piece; otherwise it is 0
- **PhD_Institution (PhDI)**: what school they got their PhD
- **Years_from_PhD (YfPhD)**: how many years at the time of the comment since they earned their PhD; this is just Year - PhD.Year. This can be negative because people may have earned their phd years after they make a comment
- **PhD_Institution_SR (PhDISr)**: The Shanghai Rankings of their PhD Institution; this is only for people who received their PhDs from institutions that are ranked by Shanghai. Shanghai ranks only between 500 and 510 universities worldwide each year and also bins their rankings after a certain point, in different ways for different years (e.g. a university may be ranked as 301-352).
- **PhD_Institution_SR_Bin (PhDISrB)**:
 - * 1 = university was ranked between 1 and 50
 - * 2 = university was ranked between 51 and 100
 - * 3 = university was ranked between 101 and 150
 - * 4 = university was ranked between 151 and 200
 - * 5 = university was ranked between 201 and 300
 - * 6 = university was ranked between 301 and 400
 - * 7 = university was ranked between 401 and 510
- **Workplace_SR (WorkSr)**: The Shanghai Rankings of their workplace; this is only for academics and academic institutions that are ranked by Shanghai (see PhD_Institution_SR for more information)
- **Workplace_SR_Bin (WorkSrB)**:
 - * 1 = university was ranked between 1 and 50
 - * 2 = university was ranked between 51 and 100
 - * 3 = university was ranked between 101 and 150
 - * 4 = university was ranked between 151 and 200

- * 5 = university was ranked between 201 and 300
- * 6 = university was ranked between 301 and 400
- * 7 = university was ranked between 401 and 510
- **Sr_Ranking_Dif (SrRDif)**: The difference between the binned Shanghai Ranking University of their workplace and the binned Shanghai Ranking of their PhD; a positive ranking means that they work at a place that has a higher ranking than where they got their PhD
- **PhD_Institution_US_IR (PhDIR)**: The US News and World Report created an international ranking system in 2014 to rank the top 500 universities. Thus, even if a comment was made in 1999, if they have a PhD from Carnegie Mellon, this ranking will be Carnegie Mellon’s ranking in the 2014 report
- **PhD_Institution_US_IR_Bin (PhDIRB)**:
 - * 1 = university was ranked between 1 and 50
 - * 2 = university was ranked between 51 and 100
 - * 3 = university was ranked between 101 and 150
 - * 4 = university was ranked between 151 and 200
 - * 5 = university was ranked between 201 and 250
 - * 6 = university was ranked between 251 and 300
 - * 7 = university was ranked between 301 and 350
 - * 8 = university was ranked between 351 and 400
 - * 9 = university was ranked between 401 and 450
 - * 10 = university was ranked between 451 and 500
- **Workplace_US_IR (WorkIR)**: See **PhD_Institution_US_IR**
- **Workplace_US_IR_Bin (WorkIRB)**:
 - * 1 = university was ranked between 1 and 50
 - * 2 = university was ranked between 51 and 100
 - * 3 = university was ranked between 101 and 150
 - * 4 = university was ranked between 151 and 200
 - * 5 = university was ranked between 201 and 250
 - * 6 = university was ranked between 251 and 300
 - * 7 = university was ranked between 301 and 350
 - * 8 = university was ranked between 351 and 400
 - * 9 = university was ranked between 401 and 450
 - * 10 = university was ranked between 451 and 500

- **USA_I_Ranking_Dif (IRDif)**: the difference between the rank of someone’s workplace and the rank of their PhD Institution (as ranked by US News and World Report International Rankings). If this is positive, it means they’re working at an institution ranked higher than their PhD Institution.
- **PhD_Institution_US (PhDIUS)**: The ranking of their PhD Institution by USA News and World Report; this is only for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
- **PhD_Institution_US_Bin (PhDIUSB)**:
 - * 1 = university was ranked between 1-5
 - * 2 = university was ranked between 6-10
 - * 3 = university was ranked between 11-25
 - * 4 = university was ranked between 26-50
 - * 5 = university was ranked between 51-100
 - * 6 = university was ranked between 101-150
 - * 7 = university was ranked between 151-200
- **Workplace_US (WorkUS)**: The ranking of their workplace by USA News and World Report; this is only for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
- **Workplace_US_Bin (WorkUSB)**:
 - * 1 = university was ranked between 1-5
 - * 2 = university was ranked between 6-10
 - * 3 = university was ranked between 11-25
 - * 4 = university was ranked between 26-50
 - * 5 = university was ranked between 51-100
 - * 6 = university was ranked between 101-150
 - * 7 = university was ranked between 151-200
- **USA_Ranking_Dif (USRDif)**: the difference between the rank of someone’s workplace and the rank of their PhD Institution (as ranked by US News and World Report Rankings). If this is positive, it means they’re working at an institution ranked higher than their PhD Institution.
- **Total_Citations (TotCit)**: the total number of citations they have received, including that year and all previous years (it’s citations.year + previous citations)

- **H_Index (Hind)**: this is their h-index in **2014**; a scholar has an index of h if they have published h papers each of which has been cited in other papers at least h times
- **i10_index (iTEnIn)**: how many papers in **2014** they had authored that has more than 10 citations; this is only for Google Scholar pages. As the GS pages only have an i10 index from 2014, even if the comment was from 1999, the i10 index is from 2014
- **Citations_Year (CitY)**: how many citations they received this year; this is only for Google Scholar pages, so not all academics have this
- **Citations_Cumulative (CitCum)**: how many citations they have received in this year and previous years; this is only for Google Scholar pages, so not all academics have this
- **AcademicHierarchyStrict (AcaHier)**:
 - * 1 = Graduate Student
 - * 2 = Postdoctoral
 - * 3 = Assistant Professor
 - * 4 = Associate Professor
 - * 5 = Professor
 - * 6 = Chaired Professor
- **PreviousCitations (PreCit)**: the number of citations they have received in all of the previous years
- **ContributionsbyAuthor (ContrAut)**: the number of contributions by this author in this conversation
- **Dummy variables for Discipline**
- **Dummy variables for department_S**
- **Text-Level**
 - **Order**: The order of the text pieces. This is meaningless for Annual Questions.
 - **Text**: the text of the conversation
 - **Number_Characters**: number of characters in the text piece
 - **LIWC variables** (see www.liwc.net/descriptiontable1.php)

A.2 CS2 Phase 2 pre-survey for analysts

Crowdsourcing Data Analysis 2: Explaining Variability in Data Analysis Decisions

Dear colleague,

Thank you for joining us as a collaborator and co-author on this crowdsourced project. Your work on the dataset and answers in this survey will help us better understand the reasons for variability in data analytic choices. Before embarking on the project, we would like to ask you a few questions about your background and experience.

The survey consists of 39 questions and will not take more than 15-20 minutes to answer.

Your responses, along with those from other project collaborators, will be used only for scholarly purposes and will be kept anonymous (i.e., will not be associated with your name).

Please note that authorship on the final project report is contingent on completing all stages of the project, including not only this presurvey but also the analysis of the dataset and tracking your decisions using the DataExplained process.

Q1: What is your name?

Q2: What is your username for DataExplained?

Q3: What is your highest degree?

- PhD
- Master's
- Bachelor's
- Other (Textfield)

Q4: Please explain your professional background:

e.g. Bachelor in Psychology, Master in Cognitive Psychology

Q5: In which country were you born? (Dropdown of countries)

Q6: In which country do you reside? (Dropdown of countries)

Q7: Please rate your political ideology on the following scale:

- Strongly Left-Wing
- Moderately Left-Wing

- Slightly Left-Wing
- Moderate
- Slightly Right-Wing
- Moderately Right-Wing
- Strongly Right-Wing

Q8: What are the keywords best describing the topics of your research?

Q9: What language/software/tools/do you prefer using in your works when doing data analysis?

(e.g. R, STATA, Python, ...)

Q10: How many years of experience do you have in data analysis?

Q11: How regularly do you perform data analysis?

- Daily
- 2-3 times a week
- Once a week
- Once every two weeks
- Once a month
- Less than once a month

Q12: Please explain your background in data analysis in more detail:

(e.g. what classes did you take, what projects have you analyzed etc.)

Q13: Below you will find a set of skills and behaviors that you likely engage in while conducting the analysis. Please indicate how confident (%) you are that you are able to do the following:

(0 - Cannot do at all, 50 - Moderately can do, 100 - Highly certain can do)

- Operationalize key variables based on theoretically defensible rationales
- Handle a large data set
- Use appropriate analytic techniques to test the proposed hypotheses
- Provide a clear description of the analysis strategy and rationale

Q14: Below is a list of statistical methods. To what extent do you consider yourself to be skilled in each of them?

(Not at all, To a low extend, To a medium extend, To a high extend, To a very high extend)

- Descriptive statistics (for example median or variance)
- Inferential statistics
- Hypothesis
- Ingression
- Estimation
- Correlation
- Regressions
- Forecasting
- Prediction
- Extrapolation
- Interpolation
- Time series
- Data mining

Q15: How do you rate your level of expertise in the field of data analysis?

1. Very poor
2. Amateur
3. Good
4. Very Good
5. Excellent

Q16: Have you taught an undergraduate level statistics course? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q17: Have you taught an undergraduate level course on analyzing text? If so, how many total times (estimate is fine)?

- 0

- 1-2
- 3-5
- more than 5

Q18: Have you taught a graduate level statistics course? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q19: Have you taught a graduate level course on analyzing text? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q20: Approximately how many Edge conversations have you read before? Edge.org is an online website for intellectual discussions.

- 0
- 1-2
- 3-5
- more than 5

Q21: Have you taught a graduate level course on analyzing text? If so, how many total times (estimate is fine)?

- 0
- 1-10
- 11-20
- 20-50
- more than 50

Q22: Have you published a peer-reviewed, scientific paper using text analysis? If not, please put 0, and if so, please put how many (estimate is fine):

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q23: Have you published a peer-reviewed, scientific paper that is on the topic of gender? If not, please put 0, and if so, please put how many (estimate is fine). If you have published articles on both gender AND status, please include it in the gender and the status publication counts.

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q24: Have you published a peer-reviewed, scientific paper that is on the topic of social status? If not, please put 0, and if so, please put how many (estimate is fine). If you have published articles on both gender AND status, please include it in the gender and the status publication counts.

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q25: Have you published a paper that is primarily a methodological/statistical contribution? If not, please put 0, and If so, how many in total (estimate is fine)?

- 0
- 1-2
- 3-5

- 6-8
- more than 8

Q26: To what extent does your research focus on social status?
(7 point Likert scale with 1 = "Not at all" and 7 = "Extremely")

Q27: In your personal opinion and experience, to what extent do you find a person's status to play a role in their professional interactions in science?
(7 point Likert scale with 1 = "Not at all" and 7 = "Extremely")

Q28: Briefly tell us about this: (Textfield)

Q29: To what extent does your research focus on gender issues?
(7 point Likert scale with 1 = "Not at all" and 7 = "Extremely")

Q30: In your personal opinion and experience, to what extent do you find a person's gender to play a role in their scientific career?
(7 point Likert scale with 1 = "Not at all" and 7 = "Extremely")

Q31: Briefly tell us about this: (Textfield)

Q32: What is your current opinion regarding hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion.

- Very Unlikely
- Unlikely
- Neither Likely nor Unlikely
- Likely
- Very Likely

Q33: Please explain why you think so: (Textfield)

Q34: What is your current opinion regarding hypothesis 2: Higher status participants are more verbose than are lower status participants.

- Very Unlikely
- Unlikely
- Neither Likely nor Unlikely
- Likely

- Very Likely

Q35: Please explain why you think so: (Textfield)

Q36: What is your gender?

- Female
- Male
- Other (Textfield)

Q37: What is your age? (Textfield)

Q38: What title best describes your current position?

- Full Professor
- Associate Professor
- Assistant Professor
- Post-Doc
- Doctoral Student
- Other position at a University
- Outside of Academia (With follow-up question to state title)

Q39: We are very interested to know any thoughts and comments you have about the survey you just completed or the present project to crowdsource the analysis of data. Please describe them here: (Textfield)

Thank you for your answers!

Once you click submit you will be taken to the data analysis platform. Please login with your account details and begin your analysis to test the two hypotheses of interest.

A.3 CS2 Phase 2 post-survey for analysts

This questionnaire will be used to collect answers detailing the statistical approach that you have taken. Your answers will then be used to facilitate the online peer feedback process. Please provide enough information for a naive empiricist to be able to give you valuable feedback. Remember, not all individuals involved in this project come from the same discipline, so some methods might be unfamiliar/have a different name to those in other areas. There are two sections: one that will be shared with other researchers, and one that we will use internally to get a good first idea about actual results. Only the analytic methods will be shared with the crowdsourcing analysts to avoid bias.

Q1: What is your name?

Q2: What transformations (if any) were applied to the variables. Please be specific and explain why you applied them.

Q3 Were any cases excluded, and why?

Q4 How did you operationalize verbosity?

Q5 What are the theoretical reasons for operationalizing verbosity in that manner?

Q6 How did you operationalize status?

Q7 What are the theoretical reasons for operationalizing status in that manner?

Q8 What is the name of the statistical technique that you employed?

Q9 Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known.

Q10 Please explain why you chose this technique.

Q11 What are some references for the statistical technique that you chose?

Q12 What variables were included as covariates (or control variables) when testing Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion.

Q13 What theoretical and/or statistical rationale was used for your choice of covariates included in the models when testing Hypothesis 1?

Q14 What variables were included as covariates (or control variables) when testing Hypothesis 2: Higher status participants are more verbose than are lower status partic-

ipants?

Q15 What theoretical and/or statistical rationale was used for your choice of covariates included in the models when testing Hypothesis 2?

Q16 What unit is your effect size in?

Q17 What is the size of the effect for Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion. Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other analysts at this stage.

- estimate (1)
- low interval (2)
- high interval (3)

Q18 Anything else you'd like to add?

Q19 What is the size of the effect for Hypothesis 2: Higher status participants are more verbose than are lower status participants? Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other analysts at this stage.

- estimate (1)
- low interval (2)
- high interval (3)

Q20 Anything else you'd like to add?

Q21 What other steps/analyses did you run that are worth mentioning? Include effect sizes in a similar format as above if necessary.

Q22 You may use the space below to paste the script you used to run the analyses. (Optional)

Q23 What is your current opinion regarding Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion

- Very Unlikely (1)

- Unlikely (2)
- Neither Likely nor Unlikely (3)
- Likely (4)
- Very Likely (5)

Q24 Please explain why you think so.

Q25 What is your current opinion regarding Hypothesis 2: Higher status participants are more verbose than are lower status participants?

- Very Unlikely (1)
- Unlikely (2)
- Neither Likely nor Unlikely (3)
- Likely (4)
- Very Likely (5)

Q26 Please explain why you think so.

Q27 Please use this space for any additional comment you may have at this stage (this is for our information and will not displayed to others).

Please press the submit button only once you are sure that you would like to submit your responses and that no changes are needed at this stage.

A.4 Code Book

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
action driven by insight	Analyst's personal insights may drive certain actions to be followed (e.g. run correlation test on two variables of interest emerged from the insight generation). Often related with the code "Insight realization"	Apply this code when the taken action can be accounted to a personal insight	Do not use this code if you can not clearly assign an insight which ultimately led the analyst follow an action. Do not use this code if the analyst has not followed an action due to the new insight (yet). For this, refer to the code "Insight realization"	1.[goal]: Tested if authors were always the first entries in conversation, as this assumption could affect the analysis 2.[reason]: When I wrote the reporting in the questionnaire I suddenly got unsure if the reverse coding of Female (compared to status) did affect the effect size so I ran that model too just to be sure that it didn't.	Insight	Sense-making
belief	Besides the interests, analysts, being humans, are prone to have a subjective system of beliefs. It might be the personal belief (agenda) for an analyst to prove that this hypothesis is correct. Predisposition can play a key role in the way a data analysis is conducted.	Apply this code when the activity results from a belief system of analyst. For instance, a researcher exploring whether nicotine has negative impact on life expectancy might not have any personal interest, but it might be her personal belief to prove that this hypothesis is correct.	Do not use this code if there is no indication that the analyst's thoughts could be a results of personal beliefs. If it's a belief "how things in reality work", the correct code is "perceived understanding of reality".	1.[reason]: If participants are joining in through the lifetime of a discussion, I do not feel it is valid to use the overall summary numbers of male and female contributions across the life of the entire thread, as that assumes for knowledge of the number of participants, so this works out code for aiding making that decision 2.[reason]: I believe that women are under-represented in different disciplines. Therefore, I used Discipline as a confounding variable.	Belief	Personal
code quality	Actions performed to enhance the objective quality of code (e.g. reorganize, refactor, comment etc.)	Apply this code whenever the actions to enhance the quality of code are objective (i.e. actions performed as routine).	Do not use this code when the measures for enhancing the code quality are subjective (i.e. subjective quality assessment of the code). The correct code for this would be "subjective data clarity". An example for latter would be to rename a variable to a more (subjective) meaningful name.	1.[goal]: Rewrote and commented the code (final pretty version.R) so that it was better for sharing, then reran the analysis off the code 2.[reason]: To make it easier for others to draw on what I have done, and demonstrate the validity of what I did	Code	Analysis

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
complexity constraint	<p>Complexity constraint represents cases where analyst considers the complexity of alternatives or performed methods. A method might be objectively better but still avoided due to analyst's reluctance to engage in complicated data analysis process. This code is related to "effort constraint". However while the "complexity constraint" is related to the perceived complexity of the method (i.e. how complicated is it to execute), the effort constraint is related to the effort associated with alternative, which is not necessarily results from the complexity of the method.</p> <p>Another relevant code is a "methodological constraint". This code relates to the objective constraints imposed by the requirements of a method.</p>	<p>Apply this code whenever the constraint(s) is due to the complexity of a (statistical) method. Hence, only apply this code if the constraint is objective (e.g. generally requires more work from every person, not only from one certain analyst)</p>	<p>Do not use this code when the constraints are subjective (e.g. simplicity or time-intensive). The correct code for latter would be "effort constraint".</p>	<p>1.[goal]: Testing hypothesis one as a mixed effect model. The data has a bit of a complex structure, as people appear in multiple different threads, and threads will have more than one person in them. So I used a mixed effect model.</p>	Task	Given
confirmatory measure	<p>Analyst tend to confirm their (intermediate) results in different phases of their analysis.</p>	<p>Apply this code when the analyst tries to statistically validate the results.</p> <p>Apply this code if the analyst confirms the results by looking at it from different perspectives (e.g. visualise the fitted model).</p>	<p>Do not apply this code if the very same code is just rerun. In this case we do not apply any code.</p>	<p>1.[reason]: I wanted to see whether my model (edge.1.fit) has done a good job of capturing the patterns in the dataset</p> <p>2.[reason]: Because I wanted to make sure what I did was correct</p> <p>3.[goal]: Some basic model checking to verify if the results are credible.</p>	Iterations	Analysis
data constraint	<p>Any constraint imposed by the nature of data.</p>	<p>Apply this code whenever the constraint(s) origins in the data provided in the problem. Data in that sense is always related to the problem. This can also be if the data is inconsistent or in an unfortunate format for the given situation. Another case would be when the data is not informative enough to answer the question.</p>	<p>Do not apply this code if the constraint(s) are related to the methodology.</p>	<p>1.[dis]: I don't think there is enough data to parameterize this model</p> <p>2.[dis]: May lose power with less data</p>	Data	Given
data quality	<p>Any objective metrics of data quality such as completeness, bias, distribution</p>	<p>Apply this code if the quality assessment for the data is objective.</p>	<p>Do not apply this code if the quality assessment for the data is subjective. For this, you might consider the code "interpretability constraint", as it is a personal assessment of the quality.</p>	<p>1.[adv]: This may even more accurately reflect the degree of female participation at the time participants post</p>	Data	Given
effort constraint	<p>Effort constraint represents cases where effort prevents analyst from taking certain actions/decisions during data analysis. This can be either due to time/complexity constraint or because the perceived benefit versus invested effort do not make it attractive ("too much work to be done").</p>	<p>Only apply this code when the constraints are purely subjective (e.g. simplicity or time-intensive). Effort in that sense is meant to be a subjective measure. Hence these constraints may not necessarily be the same for every analyst (e.g. due to a broader knowledge, another analyst might not face any effort constraint)</p>	<p>Do not use this code when the constraints are objective (e.g. originates from the general complexity of a method). The right code for latter would be a "complexity constraint"</p>	<p>1.[dis]: More confusion and more work in later stages</p> <p>2.[adv]: This is easier and means you get results immediately</p> <p>3.[dis]: More difficulty in going through large set of data.</p>	Knowledge	Personal

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
error fixing	Code executed for debugging / corrective measures.	Apply this code whenever any corrective measures are performed (e.g. due to typos or format errors).	Do not apply this code if the measures are not corrective. Rerunning code, just to look how it performs does not need to be coded, unless the analyst explicitly states to have fixed any errors or typos in code. If an analyst is purely trying out to get a function to work, you better consider the code "exploratory".	1.[goal]: Correcting errors in for loop 2.[goal]: Fixed a bug	Code	Analysis
expertise	Decisions or actions that reflect professional knowledge and experience. For example when analyst is considering that while applying a certain method, one has to be careful of certain aspects such as assumptions or limitations.	Apply this code when the actions can clearly be mapped to an analyst's professional background (i.e. necessarily shared knowledge). This expertise was already present before the analysis, otherwise this would be coded as "insight realization" or "action driven by insight".	Do not apply this code if the decision or action is based on a subjective (personal) assumptions the analyst makes. This would rather be a "understanding of the problem"	1.[dis]: Not necessary for a dataset of this size 2.[alt]: Do not look at transformations - could use other modeling approaches such as GLM 3.[dis]: Not as suitable as a single factor for aggregation methods (e.g. tapply)	Knowledge	Personal
exploratory	Any exploratory steps performed by the analyst. This is related to exploratory data analysis and can describe activities focused on data or model exploration.	Apply this code whenever the analyst is trying out stuff, or just exploring the data or model.	Do not apply this code if the actions' course is clearly defined, or explicitly intended by the analyst. In this case, the code "perceived course of action" might be more appropriate.	1.[goal]: just tried to get a sense of the number of posts per year, and how many unique conversations there were 2.[goal]: trying to get general sense of thread statistics	Exploration	Analysis
feature engineering	Adding new features (aka variables/columns/attributes etc.) which are a function of existing data.	Apply this code whenever the analyst is creating new features during his analysis. This code can be related to the codes "perceived course of action" or "intuition about the problem".	Do not apply this code if you cannot see any new feature resulting from the analyst's action. Otherwise, the code "preprocessing" might be more appropriate. Do not apply this code if the new feature results of a common/routine data analysis practice (e.g. split dataset in training and test set). In this case, the code "exploratory" might be appropriate.	1.[goal]: working on adding another variable 2.[goal]: Trying to compute, for each online comment, the number of preceding comments that were from other women	Data	Given
insight realization	This code describes a situation where the analyst generates new insights, instant hypotheses or ideas, due to the applied method/approach or throughout data analysis in general. This code can be seen as an evidence of sensemaking.	Apply this code when the analysts describes to have get new insights due to learning and/or sensemaking in past actions regarding the context of the problem.	Do not apply this code when you cannot clearly see a pondering /reflection on the results made by analyst. Do not apply this code when the realisation is method-related (e.g. it turned out the method was not providing him with the answers he wanted). Do not apply this code if the analysts realizes that she made a error (e.g. typo). For this, you may apply the code "debugging"	1.[reason]: I thought perhaps people who have higher status might be more verbose only in the context of not being the leader. Instead I found the opposite. They're more verbose when they're the leader, and less verbose when they aren't, which is actually quite surprising. 2.[reason]: Have realised that as doing comparisons between two times, if the analysis is sensitive to time matters	Insight	Sense-making
interpretability constraint	Analyst's have a subjective judgement for the interpretability of methods or approaches. This is a subjective constraint	Apply this code if the analyst makes a subjective judgment about the interpretability of a model/method/approach.	Do not apply this code if the analyst is using a method due to its (perceived) better interpretability. For this you might refer to the code method preference".	1.[reason]: these are less easily interpreted than the model carried out. 2.[reason]: A bit more straightforward interpretation	Method	Analysis

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
intuition about the problem	Intuition is a "gut feeling" that results out of prior knowledge or by inference from personal experiences, feelings and preferences. Intuition in this case refers to intuitions about future actions.	Apply this code when the analyst describes how the (logic of) the problem may look like. These statements are subjective assumptions and do not necessarily reflect the truth about the problem. The statements are made in a prospective manner.	Do not apply this code when you cannot clearly see an intuitive nature of thinking in the explanations of the analyst. Otherwise you might consider the code "understanding of the problem"	1.[reason]: I want to be able to count the number of annual questions/conversations among the total number of threads 2.[reason]: It's the most intuitive I can think of 3.[reason]: Feel that it's important to look at how many comments occur generally before investigating how it relates to gender.	Problem	Given
method preference	Analyst's preference of certain methods. This can be either due to professional background/education or commonly faced problems. For example Bayesian statisticians prefer certain methods while some other researchers frequentist methods.	Apply this code when an analyst preferred a certain method while the proficiency in the alternative methods is assumed to be similar. She does not necessarily have to list or compare the followed method to alternatives.	Do not apply this code when the analyst only describes possible methods for the current problem. If the respective method(s) is not actually followed, or the analyst considered the method but eventually decided to not follow down this path, you may consider another code. (e.g. "effort constraint").	1.[reason]: I can use the paired t test to minimise confounders 2.[goal]: tested hypothesis 2 with a spearman correlation	Knowledge	Personal
methodological constraint	A methodological constraint related to the limitations imposed by considered methods or approaches. For example, assumption of normality or homoscedasticity have to be fulfilled in order to apply certain methods.	Apply this code, if the (statistical) method implies objective constraints imposed by the limitations of the considered method.	Do not apply this code if the constraint is of subjective nature. In this case you may consider the code "effort constraint".	1.[adv]: these alternatives may be more appropriate given the distributional assumptions of the data, or, different mechanisms (i.e., is the total number of females important) through which there may be some impact on female participants' postings 2.[prec]: Standard assumption for regression models	Method	Analysis
perceived course of action	The analyst performs an action in order to be able to continue the way she intends. (E.g. when analyst states a clear path to operationalize the problem - "Do A in order to do B").	Apply this code if the course of action is planned, envisioned or implied by the analyst's understanding and not because of any constraints of methodology or data.	Do not apply this code if the course of action is not perceived by the analyst, but required by a method, hypothesis, or due to the nature of the data. In this case refer to a code "methodological constraint", "task constraint" or "data constraint" respectively. Do not apply this code if the end-goal of this action is very obvious (e.g. to answer the hypothesis). Do not apply this code if the analyst claims to just rerun the code in order to [...]. For this, use a more specific code related to the context (e.g. "error fixing" or "confirmatory measure").	1.[goal]: Calculate number of contributions for each person in each conversation. This can be used to determine mean number of contributions for male individuals and mean number of contributions for female individuals. 2.[reason]: create the outcome measure engagement, which is word count multiplied by the cognitive processing LIWC construct, and scale this variable and the word count for use in subsequent models.	Knowledge / Belief	Personal

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
perceived understanding of reality	The perceived understanding of the reality is a complementary factor to beliefs and interests. Data analysts may have an implicit cognitive mechanism about “how things work” in the real world. This understanding is not directly about the problem which is under investigation but rather about a state in the grand scheme of things	Apply this code when the user describes his opinion on the general understanding of the reality which is beyond the scope of the currently studied problem. The same holds for justifications of the performed actions.	Do not apply this code if the perception is only related to the problem. In this case you may consider the code “understanding of the problem”. Note the difference between problem and reality: problem has a narrow scope related to a concrete problem (e.g. researched hypothesis), while reality relates to a general context/scope in which the problem studied.	1.[reason]: The number of members can influence the range of possible results, so thought it best to generate some checking code 2.[reason]: I just wanted a comparison of what the males are doing, in theory if male activity had dropped fast, a case could be made for relative activity being higher for females (but that was not the hypothesis, nor was it the result here)	Belief	Personal
perceived understanding of the problem	This code is applied when analyst is following a procedure due to the perceived logic of the problem. This code is mostly be applied, when a justification for the action is given with regards to the problem. Note, this code is different from the perceived understanding of reality. While perceived understanding of reality is reflecting a general context, understanding of the problem reflects a concrete problem analyst currently deals with and the sense-making process that occurs. Selecting features/variables belongs to this code.	Apply this code when the analysts follows an action due to the perceived logic of the problem. This understanding is always subjective and by not be shared with other analysts.	Do not apply this code if there is a lack of understanding of the problem. The code for this would be uncertainty about the problem”. Do not apply this code if the context is beyond the scope of the problem. In this case you might think of applying the code “perceived understanding of the reality”. Do not apply this code if the justification for the given action is solely attributed to the realization of new insights (rather than explaining how this is reflected in the problem). The code for this would be “insight realization”.	1.[reason]: Instead of looking at raw counts of citations, I’m interested in seeing if these scientometric variables might show clearer relationships with my outcome (average # of words) 2.[goal]: Trying to model the probability that a certain comment is from a female, based on the female participation so far	Problem	Given
personal assumption	Any personal assumptions the analyst makes. For example, the analyst dropped most of the PhDs from my analysis as they will likely not influence the final result too much.	Apply this code whenever an assumption is made, that cannot be generalized (e.g. only the analyst may think this way)	Do not apply this code if the assumptions are not subjective. If they are methodological, for example, the code “methodological constraint” may be considered suitable.	1.[reason]: I assume PhDs are academics of low rank	Belief	Personal
personal interest	Actions driven by personal interest of analyst (e.g. curiosity, choices which relate to personal rationales)	Apply this code when an analyst is intrinsically motivated to follow a certain action.	Do not apply this code if the interest is not personal. Also do not use this code if the analyst is extrinsically motivated. For example if the analyst wants to use a certain method due to its simplicity, the code method preference” would be appropriate.	1.[reason]: Curious to see some plots 2.[reason]: Wanted to explore relationships in data before modeling it.	Belief	Personal
personal knowledge	Analyst’s knowledge or prior experiences in performing an action she does (e.g. refers to past analyses, claims to be familiar with a concept, or consequences of possible actions)	Apply this code when the knowledge can be addressed to the personal background and has a subjective character.	Do not apply this code if the knowledge relates to professional experiences (i.e. if this knowledge is common). This would relate to the code expertise”.	1.[reason]: From my previous similar projects, I know that when you remove outliers in the beginning of data analysis, it can really bias the results.	Knowledge	Personal

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
personal preferences	Analysts may have preferences or intentions to perform an action the way they think is best for them. These can be driven by various personal factors. If the preference is for a (statistical) method, we apply only the code method preference".	Apply this code when the analysts claims to favor her way of procedure due to personal reasons.	Do not apply this code when the preference is guided by constraints (e.g. because the analyst does not have the necessary knowledge to follow a certain method). Do not apply this code if the preference is for a method. The code for this is "method preference". But also be careful here: If a method constraints the analysts to use a similar other method, this would have to be coded as "methodological constraint".	1.[reason]: I want flexibility at this stage, so if I need factors later I will convert 2.[reason]: I prefer density plots to histograms or box-plots,	Belief	Personal
preprocessing	Any steps performed to preprocess the data (e.g. installing packages/libraries, removing outliers, organize data, etc.)	Apply this code whenever the analyst performs routine actions to preprocess the data.	Do not apply this code whenever the analyst is creating new features during his analysis (e.g. a new column with a new feature results in the data frame). For this, you may consider the code "feature engineering".	1.[goal]: Remove NA values from data frame 2.[goal]: Loaded Data and installed libraries	Data	Given
revision of findings	Revision of findings due to the new insights or idea. Often related to the code "insight realization"	Apply this code when an analyst is reflecting/recapitulating the results. The focus of the revision is thereby the findings (and not code-related).	Do not apply this code if previous code is reused in another situation. For this you might consider a more concrete code related to the context (e.g. "error fixing", "confirmatory measure", "action driven by insight"). Do not apply this code if the analyst just reruns the very same code again. This does not need to be coded.	1.[reason]: Wanted to double check previous findings.	Iterations	Analysis
task constraint	Task constraint is related to the limitations imposed by the task analyst is performing (requirement by the task). For example, if the task is to report on certain measures or to produce a result up to certain deadline.	Apply this code if the action is influenced by any kind of constraint of the task / hypothesis of interest.	Do not apply this code, if the constraints cannot be clearly related to the task or hypothesis of interest. Do not apply this code, if the constraints are subjective or due to the chosen methodology. For this, you might consider the codes "effort constraint" or "complexity constraint", respectively Do not apply this code, if following an alternative approach would not lead to this constraint.	1.[adv]: Seems important to consider given the hypothesis in question explicitly assumes a special population will be more verbose. 2.[goal]: this is a linear model constructed to provide an answer to hypothesis 1	Task	Given
uncertainty about the method	If analyst is not sure whether the taken method is the correct one for her objectives or other method would fit better	Apply this code if the analyst is not certain if the method may provide her the desired results. Also apply this code when uncertainty can be related to missing professional knowledge for using this method.	Do not apply this code if the uncertainty cannot directly be related to the followed method. Otherwise you may consider the code "uncertainty about the problem". Do not apply this code if the action was driven by confirmatory measures. In this case, you may consider the code "confirmatory measures".	1.[adv]: Worked eventually 2.[alt]: Difficult to say if better without doing analysis.	Method	Analysis

Continued on next page

Table A.1 – Continued from previous page

Code	Memo	When to use	When not to use	Examples	Category	Meta-Category
uncertainty about the problem	A problem analyst studies might be ambiguous in its nature due to different reasons. In addition, any uncertainties expressed with regards to the problem setting (e.g. if an analyst is not sure what is the meaning of variable in dataset, how the data was collected, or how to interpret the results)	Apply this code if the analyst is not certain about any aspect related to the problem setting (e.g. logic of variables, data, etc.).	Do not apply this code if the uncertainty can not be related to the problem. Otherwise you may consider the code "uncertainty about the method"	1.[dis1]: The scale is ordinal, but it's unclear to me how different each level is from the other— how much different is an experienced graduate student from a post-doc? An associate professor vs. a full professor? It seemed better to simply recognize them as nominal categories. 2.[alt] A woman's tendency to participate actively" could have been operationalized differently.	Problem	Given
visualisation	Any kind of graphical visualisation / plot the analyst does. This is often related to the code "insight generation" or "exploratory analysis"	Apply this code whenever the analyst claims to have made a visualisation (plot).	Do not apply this code when there is no actual visualisation performed or analyst only mentions visualisation without any application.	1.[goal]: plotting the model 2.[goal]: Plotted data to explore basic relationships	Exploration	Analysis

Table A.1: Code Book

A.5 Quantitative Analysis

A.5.1 Clusters of Participants

Username of Participant Cluster	
a.k • • • 774	1
d.m • • • 969	1
e.k • • • 178	1
g.n • • • 418	1
i.p • • • 884	1
l.h • • • 335	1
m.g • • • 298	1
m.s • • • 824	1
p.e • • • 482	1
p.f • • • 983	1
p.t • • • 437	1
r.n • • • 943	1
r.t • • • 751	1
s.f • • • 958	1
t.r • • • 242	1
w.m • • • 590	1
a.j • • • 137	2
a.m • • • 340	2
a.m • • • 733	2
b.k • • • 735	2
c.y • • • 243	2
d.b • • • 106	2
d.h • • • 148	2
e.n • • • 713	2
f.d • • • 041	2
f.w • • • 501	2
h.d • • • 630	2
i.r • • • 718	2
j.b • • • 532	2
j.r • • • 174	2
k.m • • • 726	2
l.m • • • 271	2
m.d • • • 188	2
o.b • • • 146	2
p.h • • • 851	2
s-c • • • 165	2
c.n • • • 096	3
j.h • • • 270	3
p.c • • • 766	3
d.p • • • 412	4

Table A.2: Clusters of Participants (Username of participant obfuscated)

List of Figures

3.1	Workflow of inductive coding approach	12
4.1	Block of logs	17
4.2	Fine-tuning of blocks	18
4.3	Snippet of workflow modeled by a participant	20
4.4	Coding interface for a block	21
5.1	(Meta-)categories in a data analyst's workflow	26
5.2	Resulting workflow from process mining	28
9.1	Architecture DataExplained	35
9.2	EC2 instance	36
9.3	EC2 instance Inbound Security Rules	37
9.4	EC2 instance Outbound Security Rules	37

List of Tables

5.1	Overview of Category System	24
5.2	Clusters of Blocks	27
A.1	Code Book	70
A.2	Clusters of Participants	71

References

- [Abelson, 1979] Abelson, R. P. (1979). Differences between belief and knowledge systems. *Cognitive science*, 3(4):355–366.
- [Alonso and Volkens, 2012] Alonso, S. and Volkens, A. (2012). *Content-analyzing political texts. A quantitative approach*, volume 47. CIS.
- [Anderson, 2008] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.
- [Beall and Tracy, 2013] Beall, A. T. and Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological Science*, 24(9):1837–1841.
- [Bellman, 2013] Bellman, R. (2013). *Dynamic programming*. Courier Corporation.
- [Bellman, 2015] Bellman, R. E. (2015). *Adaptive control processes: a guided tour*. Princeton university press.
- [Bollier and Firestone, 2010] Bollier, D. and Firestone, C. M. (2010). *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC.
- [Boyd and Crawford, 2012] Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.
- [Brazier et al., 1997] Brazier, F. M., van Langen, P. H., and Treur, J. (1997). A compositional approach to modelling design rationale. *AI EDAM*, 11(2):125–139.
- [Brodeur et al., 2016] Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.
- [Burla et al., 2008] Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., and Abel, T. (2008). From text to codings: intercoder reliability assessment in qualitative content analysis. *Nursing research*, 57(2):113–117.

-
- [Campbell et al., 2013] Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42(3):294–320.
- [Chi, 2008] Chi, M. T. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. *International handbook of research on conceptual change*, pages 61–82.
- [Chung and Goodwin, 1998] Chung, P. and Goodwin, R. (1998). An integrated approach to representing and accessing design rationale. *Engineering Applications of Artificial Intelligence*, 11(1):149–159.
- [Collier et al., 2004] Collier, D., Seawright, J., and Munck, G. L. (2004). The quest for standards: King, keohane, and verba’s designing social inquiry. *Rethinking social inquiry: Diverse tools, shared standards*, pages 21–50.
- [Conklin and Yakemovic, 1991] Conklin, E. J. and Yakemovic, K. (1991). A process-oriented approach to design rationale. *Human-Computer Interaction*, 6(3):357–391.
- [Corbin and Strauss, 1990] Corbin, J. and Strauss, A. (1990). Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift für Soziologie*, 19(6):418–427.
- [Craik, 1943] Craik, K. (1943). *The nature of explanation* cambridge university press: Cambridge.
- [Creswell, 2002] Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ.
- [Cunningham and Gonzales, 2014] Cunningham, C. A. and Gonzales, J. E. (2014). The golden age of data sharing. <http://www.apa.org/science/about/psa/2014/12/data-sharing.aspx>. Online; accessed 23 July 2017.
- [Dieckrüger et al., 1995] Dieckrüger, B., Söndgerath, D., Kersebaum, K., and McVoy, C. (1995). Validity of agroecosystem models a comparison of results of different models applied to the same data set. *Ecological modelling*, 81(1-3):3–29.
- [Dole and Sinatra, 1998] Dole, J. A. and Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational psychologist*, 33(2-3):109–128.
- [Dwork et al., 2015] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.
- [Eagly and Chaiken, 1993] Eagly, A. H. and Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.
- [Fahy, 2001] Fahy, P. J. (2001). Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distributed Learning*, 1(2).

- [Fox and Hendler, 2011] Fox, P. and Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, 331(6018):705–708.
- [Freedman, 1983] Freedman, D. A. (1983). A note on screening regression equations. *the american statistician*, 37(2):152–155.
- [Friedkin et al., 2016] Friedkin, N. E., Proskurnikov, A. V., Tempo, R., and Parsegov, S. E. (2016). Network science on belief system dynamics under logic constraints. *Science*, 354(6310):321–326.
- [Gelman, 2013] Gelman, A. (2013). Too good to be true. *Slate*.
- [Gelman and Loken, 2013] Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.
- [Gelman and Loken, 2014] Gelman, A. and Loken, E. (2014). The statistical crisis in science data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don’t hold up. *American Scientist*, 102(6):460.
- [Glaser, 2017] Glaser, B. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.
- [Gonzales and Cunningham, 2015] Gonzales, J. E. and Cunningham, C. A. (2015). The promise of pre-registration in psychological research. <http://www.apa.org/science/about/psa/2015/08/pre-registration.aspx>. Online; accessed 23 July 2017.
- [Gouttefangeas et al., 2015] Gouttefangeas, C., Chan, C., Attig, S., Køllgaard, T. T., Rammensee, H.-G., Stevanović, S., Wernet, D., thor Straten, P., Welters, M. J., Ottensmeier, C., et al. (2015). Data analysis as a source of variability of the hla-peptide multimer assay: from manual gating to automated recognition of cell clusters. *Cancer Immunology, Immunotherapy*, 64(5):585–598.
- [Graneheim and Lundman, 2004] Graneheim, U. H. and Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse education today*, 24(2):105–112.
- [Grolemund and Wickham, 2014] Grolemund, G. and Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204.
- [Grösser and Schaffernicht, 2012] Grösser, S. N. and Schaffernicht, M. (2012). Mental models of dynamic systems: taking stock and looking ahead. *System dynamics review*, 28(1):46–68.
- [Gruber and Russell, 1996] Gruber, T. R. and Russell, D. M. (1996). Generative design rationale: Beyond the record and replay paradigm.

-
- [Guindon, 1990] Guindon, R. (1990). Knowledge exploited by experts during software system design. *International Journal of Man-Machine Studies*, 33(3):279–304.
- [Hallgren, 2012] Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.
- [Hardt, 2015] Hardt, M. (2015). The reusable holdout: Preserving validity in adaptive data analysis. <https://research.googleblog.com/2015/08/the-reusable-holdout-preserving.html>. Online; accessed 23 July 2017.
- [Head et al., 2015] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.
- [Hill and Levenhagen, 1995] Hill, R. C. and Levenhagen, M. (1995). Metaphors and mental models: Sensemaking and sensegiving in innovative and entrepreneurial activities. *Journal of Management*, 21(6):1057–1074.
- [Hoaglin, 2003] Hoaglin, D. C. (2003). John w. tukey and data analysis. *Statistical Science*, pages 311–318.
- [Hruschka et al., 2004] Hruschka, D. J., Schwartz, D., St. John, D. C., Picone-Decaro, E., Jenkins, R. A., and Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from hiv behavioral research. *Field methods*, 16(3):307–331.
- [Humphreys et al., 2013] Humphreys, M., Sanchez de la Sierra, R., and Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1):1–20.
- [Jansen et al., 2008] Jansen, A., Bosch, J., and Avgeriou, P. (2008). Documenting after the fact: Recovering architectural design decisions. *Journal of Systems and Software*, 81(4):536–557.
- [Klein, 1993] Klein, M. (1993). Capturing design rationale in concurrent engineering teams. *Computer*, 26(1):39–47.
- [Krippendorff, 2004] Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.
- [Kurasaki, 2000] Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field methods*, 12(3):179–194.
- [Lee and Lai, 1991] Lee, J. and Lai, K.-Y. (1991). What’s in design rationale? *Human-Computer Interaction*, 6(3-4):251–280.
- [Lukacs et al., 2010] Lukacs, P. M., Burnham, K. P., and Anderson, D. R. (2010). Model selection bias and freedman’s paradox. *Annals of the Institute of Statistical Mathematics*, 62(1):117–125.

- [MacLean et al., 1991] MacLean, A., Young, R. M., Bellotti, V. M., and Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Human-computer interaction*, 6(3-4):201–250.
- [Madill et al., 2000] Madill, A., Jordan, A., and Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British journal of psychology*, 91(1):1–20.
- [Malone et al., 2010] Malone, T. W., Laubacher, R., and Dellarocas, C. (2010). The collective intelligence genome. *MIT Sloan Management Review*, 51(3):21.
- [Morton et al., 2014] Morton, K., Balazinska, M., Grossman, D., and Mackinlay, J. (2014). Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment*, 7(6):453–456.
- [Norman, 1983] Norman, D. A. (1983). Some observations on mental models. *Mental models*, 7(112):7–14.
- [Paglieri, 2004] Paglieri, F. (2004). Data-oriented belief revision: Towards a unified theory of epistemic processing. In *Proceedings of STAIRS*, pages 179–190.
- [Partington, 2002] Partington, D. (2002). *Essential skills for management research*. Sage.
- [Russell et al., 1993] Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT’93 and CHI’93 conference on Human factors in computing systems*, pages 269–276. ACM.
- [Saldaña, 2015] Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.
- [Schubanz, 2014] Schubanz, M. (2014). Design rationale capture in software architecture: what has to be captured? In *Proceedings of the 19th international doctoral symposium on Components and architecture*, pages 31–36. ACM.
- [Schubanz et al., 2014] Schubanz, M., Pleuss, A., Jordan, H., and Botterweck, G. (2014). Guidance for design rationale capture to support software evolution. In *Workshop on Software-Reengineering & Evolution, Bad Honnef*.
- [Seel, 1991] Seel, N. M. (1991). *Weltwissen und mentale Modelle*. Hogrefe, Verlag f. Psychologie.
- [Seel, 2001] Seel, N. M. (2001). Epistemology, situated cognition, and mental models: ‘like a bridge over troubled water’. *Instructional science*, 29(4):403–427.
- [Silberzahn and Uhlmann, 2015] Silberzahn, R. and Uhlmann, E. L. (2015). Many hands make tight work. *Nature*, 526(7572):189.
- [Simmons et al., 2011] Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

- [Song et al., 2010] Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., and Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*, 14(8):1–193.
- [Strauss and Corbin, 1998] Strauss, A. and Corbin, J. (1998). *Basics of qualitative research techniques*. Sage publications.
- [Team In India, 2017] Team In India (2017). MEAN Stack Components. <http://www.teaminindia.com/hire-mean-stack-developer.html> (accessed July 30, 2017).
- [Thomas, 2006] Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246.
- [Usó-Doménech and Nescolarde-Selva, 2016] Usó-Doménech, J. and Nescolarde-Selva, J. (2016). What are belief systems? *Foundations of Science*, 21(1):147–152.
- [Weiss and Wodak, 2007] Weiss, G. and Wodak, R. (2007). *Critical discourse analysis*. Springer.