1. **Hypothetical business-related decision problem, as top-management, “should we do X in order to achieve Y” (30% -- 10.000 chars – 4,5 pages)**

**Background | Outline**

In this paper, the author assumes to be the new national head of human resources in Denmark at the consulting firm, McKinsey & Company. New to this role, the author is tasked with revising the employee retention[[1]](#footnote-1). This task originates from a recent report, in which the employee retention at McKinsey’s European offices was found to be among the lowest in industry, less than 3 years on average (Morrison, 2021). As such, the task at hand comes down to deducing what causes the low retention in order to increase it. While the economic reasons for wishing to understand and ultimately increase retention are obvious, e.g., lower recruitment costs or increased brain gain, how to obtain this increase is less obvious.

**Description of SCM | model choice**

The objective of the task is at hand is to predict the effects of an intervention, i.e., what effect does a given intervention have on retention. Such objective resides in the realm of causal inference (Hünermund & Bareinboim, 2019, p. 13), ergo, a structural causal model (SCM) must at first be specified. This causal model will rely on the graphical representation, formally known as a *directed acyclic graph* (DAG), that is always associated to its SCM[[2]](#footnote-2) (Pearl et al., 2016, p. 27). In most machine learning, the model choice tends to bring about certain restrictions and impose certain assumptions, e.g., linearity or normally distributed features. On the contrary, DAGs are fully non-parametric, eliminating the need for encoding any such distributional assumptions. DAGs encode no more than the relations necessary for identifiability (Hünermund & Bareinboim, 2019, p. 9), making them parsimonious both scientifically and in intuitive interpretation. Besides the parsimony representation, DAGs and the completeness of do-calculus jointly enable estimation causal effects whenever possible (Hünermund & Bareinboim, 2019, p. 12). Additionally, DAGs are falsifiable through testable implications over the observed relations as will be **elaborated on page x.** However, falsifying a DAG and ultimately leveraging do-calculus for estimating causal effects, prerequisites the presence of a DAG in the first place.

Developing a DAG is commonly approached in of two ways: qualitatively or quantitatively. In the former, one seeks to encode the causal relationship through background knowledge and domain expertise, e.g., consulting relevant scientific knowledge or a domain expert. As such, the DAG represents the model of one perceives the world to work, in which causal knowledge that sometimes even three-year-old children are believed to have, are made interpretable for machines (Pearl & Mackenzie, 2018, p. 44). In the latter, the DAG is constructed in a data-driven way, formally known as *causal discovery* (**source)**. In practice, they are often combined. Regardless of the approach, four assumptions tend to be shared across different DAGs: Acyclicity, Markov Property, Faithfulness, Sufficiency (**source)**. In short, these assumptions can be summarized respectively as the ability to represent the causal­­ model as a DAG (G), that each node is independent of their non-descendants when conditioned on their parent nodes, conditional independences in true underlying distribution p are represented in G, and any pair of nodes in G has no common external cause.

**Developing the DAG**

Qualitative approach to define the DAG, first by consulting relevant literature and then by impersonating as the domain expert….

Several studies have been conducted on which variables make an employee stay with a particular employer. One study, questioned more 24,829 employees in the leisure and hospitality industry, and ranked the following among the most common variables: *job satisfaction, extrinsic rewards, organizational commitment, advancement opportunities, flexible work arrangements, organizational justice as to why employees leave* (Hausknecht et al., 2009). Clearly, some of these variables are sufficiently generic to be relevant in the context of retention at McKinsey as well. However, less generic and more domain-specific variables must also be included in the final set of causally related variables. Here, the author assumes to possess the necessary domain-specific knowledge to define these variables and their causal relationship.

*Justify the following*

* First hire
* Age
* Previous experience in consulting
* Years of working experience

*Mismatch between expectations and actualities. Who are most likely to be poorly informed, the ones who have the least prior knowledge to the industry’s ways of working. Hyphothesis: perhaps some linearly dependency on age. However, given the extensive investments done in recruiting people right after their master, it may be that is not as much a matter of age as matter of prior work experience in a binary sense.*

* As such, the shallowly described causal task at hand, described at page X, now becomes the following stimulus-response-type causal query, *“Should McKinsey reduce number of graduate hires in order to achieve a higher employee retention?”*

**Present the causal model | Data simulation | Testable implications**

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Definition** | **How it is measured[[3]](#footnote-3)** |
| Retention (y) | Years of employment at McKinsey | Continuous variable |
| First Job (x) | Whether the employee held any full-time employments prior to McKinsey | Boolean variable |
| Age | Age of the employee | Integer variable |
| Industry Experience | Whether the employee worked full-time in consulting prior to McKinsey | Boolean variable |
| Educational level | The employee’s level of education | Integer variable, 1: high-school,  2: Bachelor, 3: Master, 4: PhD |
| Extrinsic Reward | Average monthly income during employment including salary and bonus | Continuous variable |
| Job Satisfaction | Employee’s personal rank of job satisfaction | Integer variable, on a scale of 1-10 |
| Organizational Commitment | Number of internal projects | Integer variable |
| Advancement Opportunities | Number of promotions | Integer variable |
| Flexible Work | Average weekly working hours | Continuous variable |

1. **Describe how the company can find an answer to the decision problem, if only given the data, but at the outset, have no knowledge about data generating process? (70% -- 23.000 chars – 10,5 pages**

*Discuss how they should deal with some relevant variables being unobserved*

*What other causal inferences can arise in the analysis?*

* Cross national transportability: recruitment costs are higher in Denmark, so is it even a problem in, e.g., US? If so, Hofstede approach to why transportability is probably less
* General measuring drawbacks, assuming a new way of hiring was introduced tomorrow, how long will it then take before visible in data?
* Carrying out the intervention ourselves, in an RCT, not always feasible (too expensive, impractical, or unethical)

From business standpoint, we may be suboptimizing

* The issue of contrasting the test result to something, i.e., validate

*Develop and describe a systematic data analytic protocol and illustrate the steps you propose in*

*Fusion.*

1. *Measured as time from start to end of employment* [↑](#footnote-ref-1)
2. Here, DAGs and SCMs are considered identical in their fully non-parametric nature, only differing in the former being a graphical representation of the latter, a mathematical representation. [↑](#footnote-ref-2)
3. See appendix 1 for detailed description of data simulation process [↑](#footnote-ref-3)