# Data dictionary (confidential)

Prepared by Daubry Byagogo, Black Saber Software

## Current employees dataset

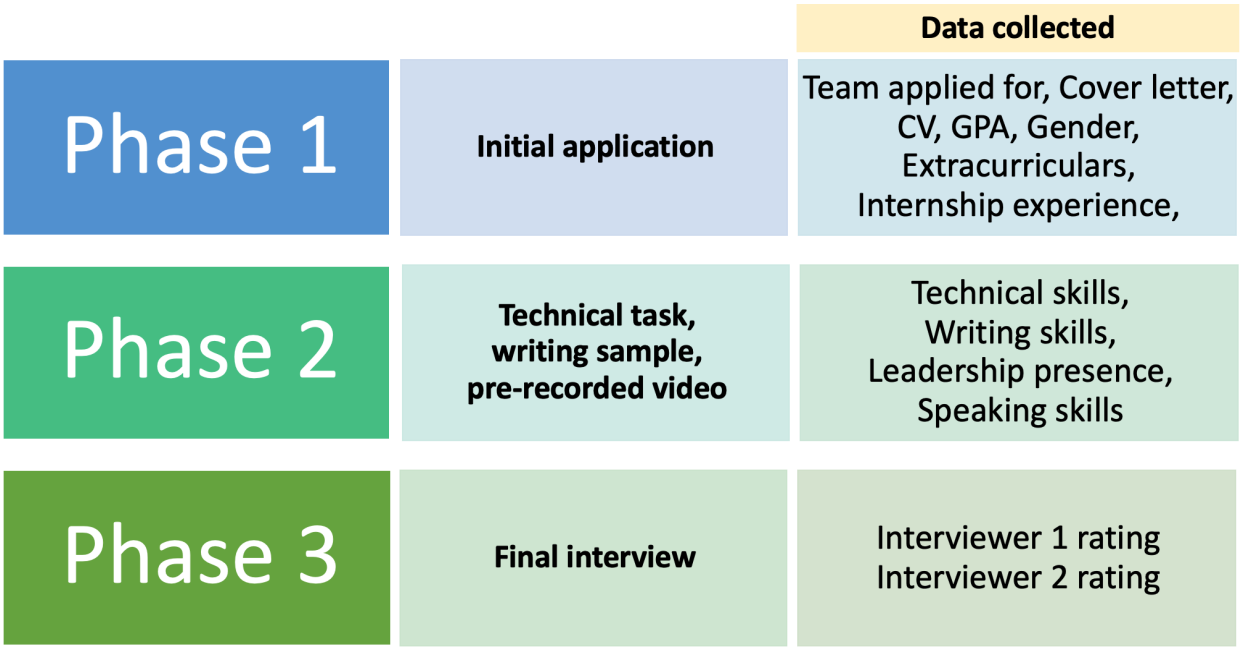
Path: "data/black-saber-current-employees.csv"

This dataset contains data on all *current* employees for the whole duration of their employment. This dataset has been anonymised. Each row represents salary, demonstrated leadership and productivity for an employee in a given quarter. The le

| Variable | Description |
|---|---|
| `employee_id` | 5 digit unique identifier for each employee |
| `gender` | Gender of employee: 'Man', 'Woman', 'Prefer not to say' |
| `team` | Which one of the 8 teams the employee works for |
| `financial_q` | Financial quarter salary, leadership and productivity |
| `role_seniority` | Least senior to most senior: "Entry-level","Junior I","Junior II","Senior I","Senior II","Senior III","Manager", "Director","Vice president" |
| `leadership_for_level` | Quality of demonstrated leadership, taking into account role level (i.e. "Appropriate for level" requires much less for entry-level employees than for a manager) |
| `productivity` | work output in relation to job description, rated on a 0-100 scale with 50 being satisfactory and above 50 indicating better than expected productivity |
| `salary` | Salary at at the given financial quarter (note: these are effective yearly values for the current wage, but don't take in to account previous salary steps in the same year, etc.) |

## Hiring data sets

Black Saber has been trialling a new AI recruitment pipeline manager for the Data and Software teams. There are three phases, outlined below, each narrowing down the field of applicants. Based on advice from our legal team we are not able to provide the original application data, be we can provide these anonymised indicators/ratings from each phase. `applicant_id` is consistent across phases.




## Phase 1

Path: "data/phase1-new-grad-applicants-2020.csv"

In the first phase of the hiring pipeline applicants complete a form and are asked to submit a CV and cover letter. Extracurriculars and internship experience are autorated based on the descriptions applicants provide in the application form.

| Variable | Description |
|---|---|
| `applicant_id` | A unique ID assigned to applicants in Phase 1 |
| `team_applied_for` | Software or Data |
| `cover_letter` | 0 if absent, 1 if present |
| `cv` | 0 if absent, 1 if present |
| `gpa` | 0.0 to 4.0 |
| `gender` | Gender of employee: 'Man', 'Woman', 'Prefer not to say' only options provided |
| `extracurriculars` | The description of extracurricular involement is assessed against a proprietary key term and phrase bank and given a 0, 1 or 2 for where 2 indicates several high relevenace and/or skills building extracurriculars, 1 indicates some relevant and/or skills building extracurriculars and 0 indicates no extracurriculars describes or that those describe were not rated as high relevance or high skills building |
| `work_experience` | Similar to `extracurriculars`, the description applicants provided is assessed against a proprietary key term and phrase bank, that also consideres company names and reputations, to give a 0, 1 or 2 score, with 2 being the best, 0 the worst |

## Phase 2

Path: "data/phase2-new-grad-applicants-2020.csv"

We don't know exactly how these are being assessed by the AI, the algorithm is obviously commercially sensitive but their demonstrations of the system were impressive.

| Variable | Description |
|---|---|
| `applicant_id` | A unique ID assigned to applicants in Phase 1 |
| `technical_skills` | Score from 0 to 100 on a timed technical task, AI autograded |
| `writing_skills` | Score from 0 to 100 on a timed writing task, AI autograded |
| `speaking_skills` | A rating of speaking ability based on pre-recorded video, AI autograded |
| `leadership_presence` | A rating of 'leadership presence' based on pre-recorded video, AI autograded |

## Phase 3

Path: "data/phase3-new-grad-applicants-2020.csv"

This is the interview phase. Being listed as 'first' or 'second' interviewer is arbitrary and who the interviewers were is not available from our tracking system.

| Variable | Description |
|---|---|
| `applicant_id` | A unique ID assigned to applicants in Phase 1 |
| `interviewer_rating_1` | The overall rating of job fit given by thefirst interviewer on a scale of 0 to 100 |
| `interviewer_rating_2` | The overall rating of job fit given by the second interviewer on a scale of 0 to 100 |

## Final hires

Path: "data/final-hires-newgrad_2020.csv"

This data set contains the applicant IDs of everyone who was sent an offer letter. In this cohort, everyone accepted.

| Variable | Description |
|---|---|
| `applicant_id` | A unique ID assigned to applicants in Phase 1 |