Dataset	Beneficence	Non-maleficence	
		Risks	**Mitigation**
□ Who created the dataset? The dataset was collected and maintained by research team led by Dr Davide Bavato at the Management of Technology & Entrepreneurship Institute, EPFL, conducting the study on venture teams' opportunity identification processes. □ For what purpose was the dataset created? The dataset was created to investigate the influence of venture team identity and the breadth and depth of industry experience on opportunity identification in response to new technologies. □ What mechanisms or procedures were used to collect the data? The data collection involved a multi-modal study design, including text-priming, video recordings of venture teams during the opportunity identification process, surveys on team members' backgrounds, and expert ratings of identified opportunities across 116 ventures. □ Who was involved in the data collection process?	□ What are the expected benefits of analyzing this data? For whom? (1) For Academic and Research Communities: The study enhances understanding of how venture teams leverage prior experience when facing new technologies, especially in the fields of entrepreneurship, strategic management, and innovation. (2) For Entrepreneurs and Venture Teams: The findings may help venture teams to identify opportunities more effectively and leverage their collective experience optimally. (1) The research can inform strategic decision-making processes for startups, potentially improving their adaptability and success rates. (2) Findings on team identity might influence how entrepreneurial teams are formed and developed, emphasizing diversity and depth of experience. (3) For Startup Eco-system Stakeholders (e.g. Investors, Mentors, Incubators): Investors and mentors can tailor their support strategies based on the findings, providing more targeted advice and resources. The results could also inform policymakers and other ecosystem builders to take targeted measures that help create more supportive environments for venture teams.	□ Does the dataset contain unsafe data (violence, nudity…)? No. It pertains to professional settings where venture teams are discussing business opportunities □ What kind of impacts can errors in the data or in the analysis have? (1) Misinterpretation of Results: If there are errors in the transcription of audio recordings or in the coding of idea classifications, the conclusions could misguide venture teams or investors. Startups might then focus on incorrect aspects of team communication or ignore more critical factors. (2) Strategic Mis-steps: If the research erroneously concludes certain communication patterns or team compositions are more effective, startups might make unnecessary or even harmful changes to their teams. They could also allocate their personnel and physical resources inefficiently. □ Could the data or the conclusions from the analysis be used in harmful ways?	

<table>
<tr>
<td rowspan="4">The data collection process involved the research team, the participating venture teams, trained raters for classification, and potentially other stakeholders such as technology experts for the expert ratings.
□ Over what timeframe was the data collected?
The study was conducted around 2015-2016, when mobile 3D scanning was a recently introduced technology.
□ Was any preprocessing of the data done?
Preprocessing steps included transcribing video recordings, coding opportunities into industry classifications, and standardizing survey responses. Specific measures are used to ensure data quality.
□ Are there any missing data or data errors?
This is no missing data for the part that we used, including audio track, industial classifications and survey data.
□ Where is the data stored?
The data is stored securely on the server in accordance with EPFL ethical guidelines and data protection</td>
<td colspan="2"></td>
<td colspan="2">No. First, personal information is properly anonymized and all information is securely stored. Second, the demographic data from the survey does not reflect or amplify any biases in terms of demographic dimensions such as gender, race, or socioeconomic status. Third, the findings are inllustrated in a fair way, avoiding creating barriers to entry or to unfairly disadvantage certain groups or competitors.</td>
</tr>
<tr>
<td colspan="2"><b>Privacy</b></td>
<td colspan="2"><b>Fairness</b></td>
</tr>
<tr>
<td><b>Risks</b></td>
<td><b>Mitigation</b></td>
<td><b>Risks</b></td>
<td><b>Mitigation</b></td>
</tr>
<tr>
<td colspan="2">□ Does the data contain personal or sensitive information?
No. Data anonymization is applied, removing or obfuscating any personal identifiers. Although audio track contains voice, it is unable to be used to identify specific individuals.
□ Can personal or sensitive information be derived or inferred from the data or from the analysis?
Even if direct identifiers are removed, there may be a risk of re-identification through combinations of data points (such as unique skillsets or experiences) or through the triangulation of information. To address this concern, access to the data is strictly controlled and restricted to authorized individuals. Each part of the data is seperately stored, and secure storage solutions are used to protect the data from unauthorized access</td>
<td colspan="2">□ Is the data representative from a larger set (population)? How are subgroups represented?
The dataset consists of 116 active venture teams younger than eight years and located in eight large cities in Germany, selected from various sources including entrepreneurship centers, start-up events, and venture databases. Although It may not represent the larger population of all venture teams, it can be representive of a specific group of active ventures located in large cities and at their early stage of development, which depends on the purpose of this research particularly on specific group..
□ What kinds of biases may affect the data?
The bias is neutrally stated. The effect of bias depends on the research goal, especially on target group.</td>
</tr>
</table>

regulations.	or breaches. Findings are reported in a way that does not disclose personal or sensitive information about any individual or small group.	(1) Survivorship Bias: Teams that have survived long enough to participate might not represent the experiences of teams that failed early. (2) Confirmation Bias: If the raters have preconceived notions about what a successful opportunity looks like, they may rate ideas that conform to those notions more favorably. (3) Selection Bias: If the venture teams were not randomly selected in terms of geography, duration, etc. (which may not be necessary for research on specific group), there might be biases towards certain types of teams or industries. □ Can the outcomes of the analysis be different for different groups? Yes, they could be. Outcomes could possibly vary if there are underlying differences in how various groups identify and pursue opportunities. For example, teams with more resources or better networks may identify more or higher-quality opportunities. □ Could the data or analysis results contribute to discrimination against people or groups? It could but the possibility is low. If not handled carefully, the analysis could inadvertently lead to discrimination. For instance, if the study finds that teams from certain backgrounds are more successful and this is interpreted as being due to inherent qualities of those groups rather than contextual factors, it could lead to biased decision-making by investors or support programs

			(however, we do not find these types of patterns from the data).	
	Sustainability		**Empowerment**	
	Risks	**Mitigation**	**Risks**	**Mitigation**
	□ What is the carbon and water footprint generated by the storage of the data and by the computation in the analysis process? (1) Storage: The size of the data that we use is 14M and is storaged on cloud. The cloud-storage of data has an environmental impact in terms of the electricity used to power servers. (2) Computation: The computational analysis, especially for intensive machine learning models is done on Google Colab. Given the data size and complexity of our algorithms, the electricity energy used would be limited. □ What type of human manual labor is involved in the data (e.g. labeling)? The trained raters classify opportunities identified by the venture teams. They manually code time period, activity type (e.g., brainstorming, technical discussion, reading), idea number, order and ideator (e.g., idea1, TM1), idea content (e.g., using 3d scan mobile app to create realistic video game avatar), idea industry classification (e.g., 5414), idea rating (e.g., business value, novelty, feasibility), and some other indicators.		□ How are the people concerned involved with the data or the analysis: have they been notified, have they consented? Yes. The study has required the explicit consent of all venture team members involved, which typically includes informing them about the nature of the study, the data to be collected, and how it will be used. Participants were informed through clear communication about the research goals, the procedures involved, and the potential risks and benefits. Participants were also provided with a process to exercise these rights. □ Are the people concerned able to make choices (e.g. revoke consent, modify or delete data) regarding the data or the analysis? Participants had the option to revoke their consent at any point, which involved the cessation of data collection and the removal of their data from the study, to the extent that it is possible without compromising the integrity of the research. Participants had the right to request modifications to or deletions of their personal data.	

	□ Does the data or the analysis require updates? The datasets such as video recordings and survey responses are currently static. If the research aims to track changes over time or maintain relevance with current technological trends, periodic updates may be necessary. In the meantime, the models and algorithms used for analysis might need updating as new data becomes available, or as methods improve, using additional computation and human labor.	