Interview Questions

Phases	Index	Questions	Possible Responses	Follow-Up-Questions
Defining Data labeling in practice:				
	1	Can you describe the data labeling process in your company?	"labelling" → monitoring the performance of the model over a period of time, then "switching" the model performing badly (done by humans)	- By whom is it done? – - What practices do/they you use? - What are the difficulties**? - how long do they spend in doing this? - what are the problems? - after how much time do they re-assess the performa**nce?)
Defining Challenges:				
	2	How much time is invested in data labeling in practice (compared to the whole project scope)?	- Noting - A lot - External	- What are key factors that are time intensive in practice? (10 percent of your time in a project…?)
	3	How much resources does the labeling process require?	- Noting - A lot - External	- What **percentage of whole project costs** does labeling take? - Can you explain that further? - How **many**

Phases	Index	Questions	Possible Responses	Follow-Up-Questions
				Experts does it need to label the data (per day)? - **What exactly takes the most time** for them? - What are the most **expensive step**s in practice?
	4	How can data labeling be improved in practice? What are possible ways/policies/practices to deal with the described pain points in the data labeling process?	-Active Learning & Semi-Supervised Approaches vs. Opinion from the paper: Only use machine learning when it is really necessary (non-ML solutions vs ML) → Leads to less labeling	- Can you walk me through an example, how this could be done in practice? - Why is it not done yet? - Can these procedures be standardized?
	5	What are your thoughts on active learning for labeling ML models in practice?	Not used - Used **Active Learning (with highest uncertainty):** A ML-Model, which is trained on a small labeled dataset, is used to suggest which datapoints to label out of at large unlabelled dataset. These specific datapoints are suggested for example to maximize the **learning rate**.	- What are the key problems/advantages, can you explain? - How big are the cost savings for labeling? - On what metrics do you decide which data points to label first?
	6	What would be the cost differences between spilt labeling and all together?	- Same - (certain budget → iterative	- Have you experienced issues/better results (e.g. quality issues or

Phases	Index	Questions	Possible Responses	Follow-Up-Questions
			approach vs. all together)	accuracy changes) with one or the other approach? - What are the cost drivers?
Experiences, knowledge and predictive needs with methods				
	7	What do future companies need to consider to best allocate forces and cheapest label their security datasets? (e.g. using active learning)	- As explained earlier	- What do you think will possibly change for future companies? - Can you walk me through a potential example?
General Follow-Up Questions				
	8	How do you think the “explainability” of network data will change? What impact does that have on the labeling process?	(technically, active learning "does" use some explainability techniques (i.e., the confidence of an ML mode) as a guide to the optimization process)	- Is labeling then less important?