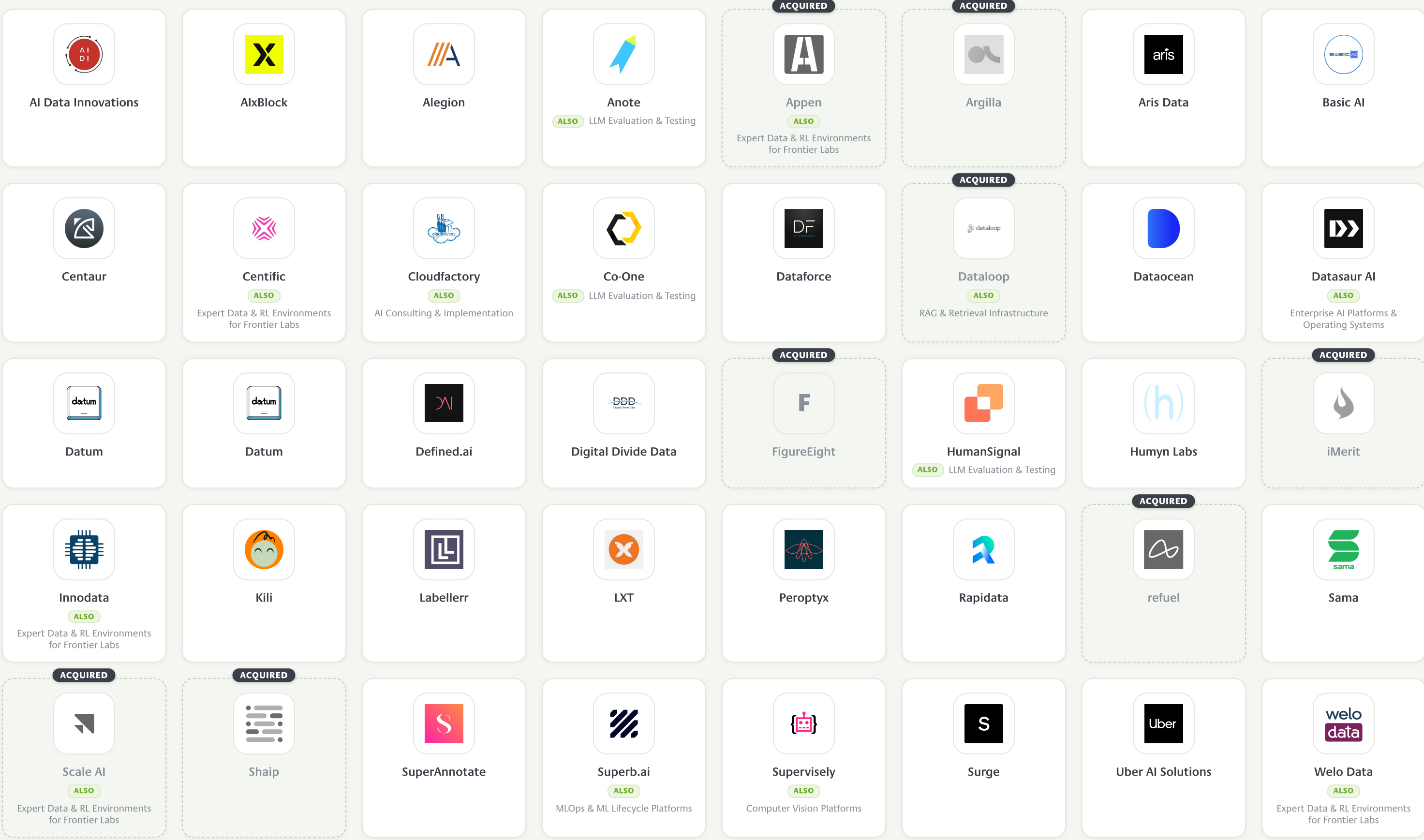


The Training Data Landscape 2026

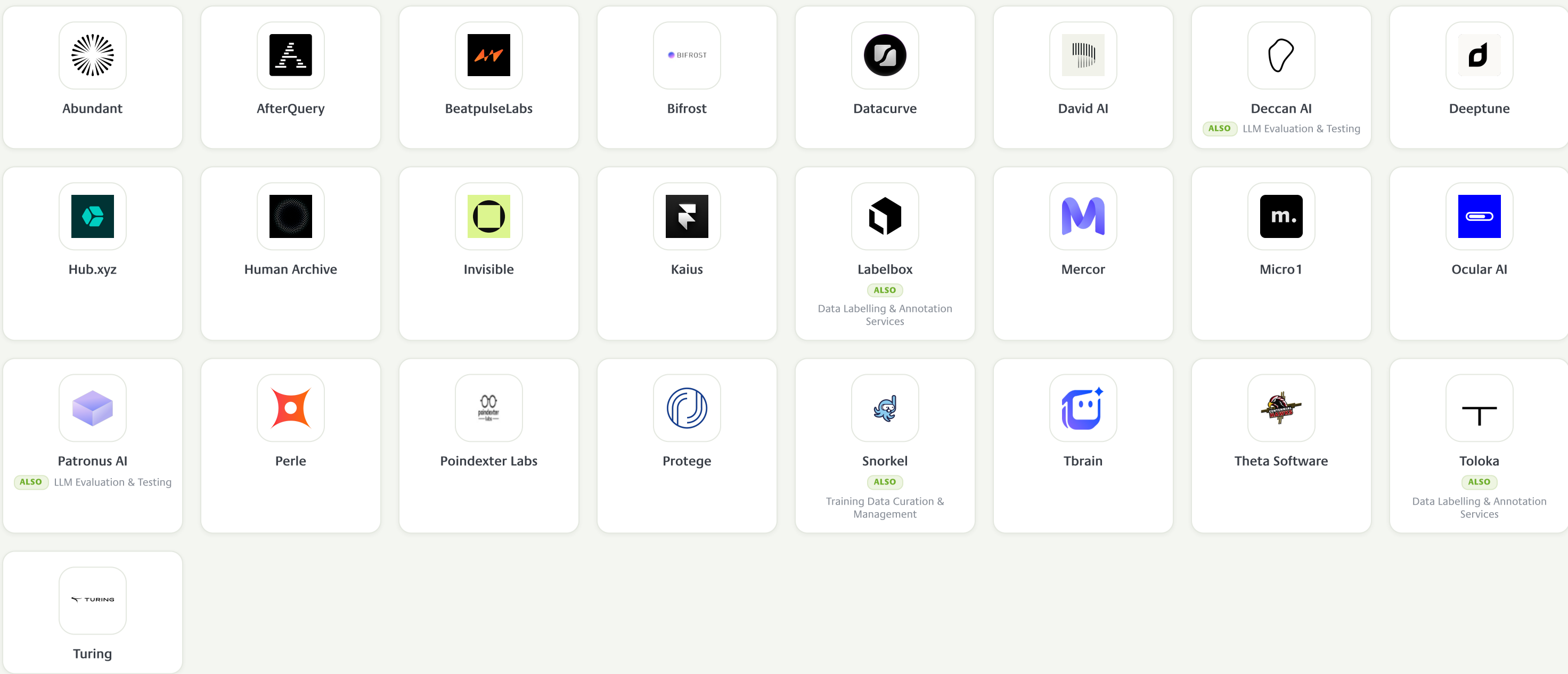


Funding is modest and almost entirely venture (**\$3.1bn of \$3.2bn**), striking for a layer every model depends on. The movement is upmarket: commodity labelling still leads (40 companies), but expert data and RL environments for frontier labs (25) are the fastest-emerging cluster, where value shifts from volume to expertise.

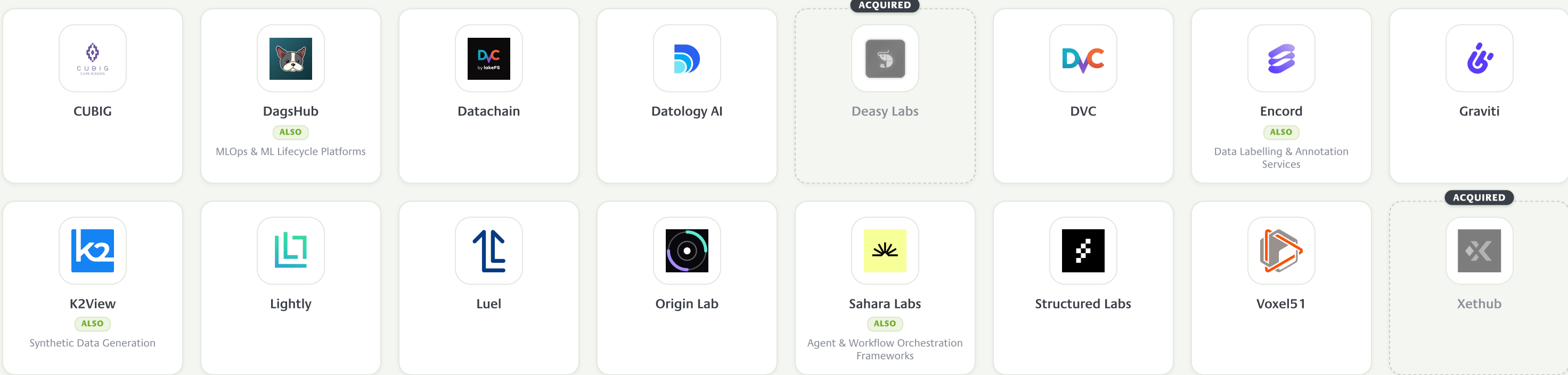
Data Labelling & Annotation Services 40



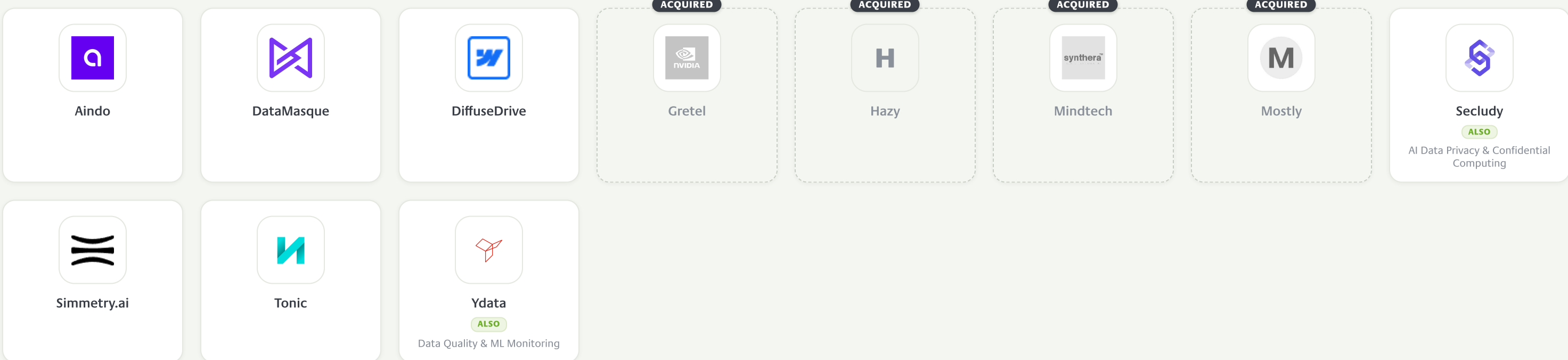
Expert Data & RL Environments for Frontier Labs 25



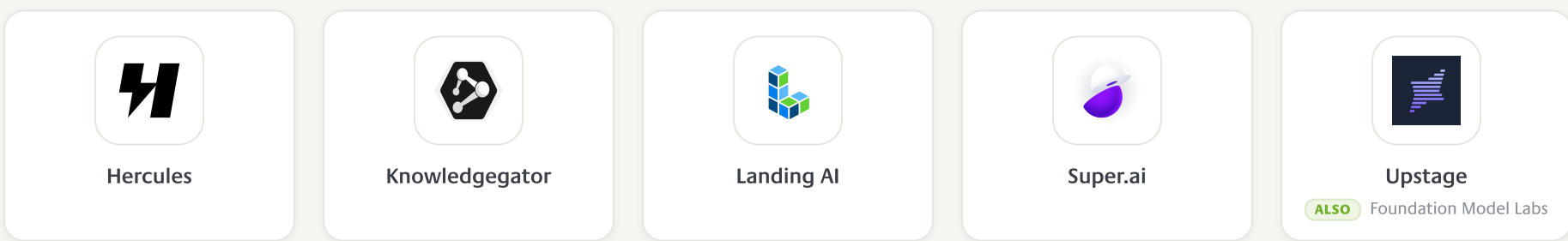
Training Data Curation & Management 16



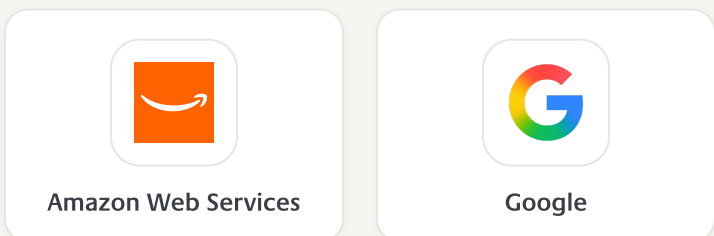
Synthetic Data Generation 11



Document Processing & Extraction (IDP) 5



Hyperscaler / SaaS Giant 2



*Funding from Crunchbase, 80 of 99 companies matched. "Total capital raised" includes all funding types (venture, debt, post-IPO, grants); "venture funding" excludes debt, private equity, post-IPO and grant rounds. Companies may appear in more than one sub-sector. Logos via logo.dev.