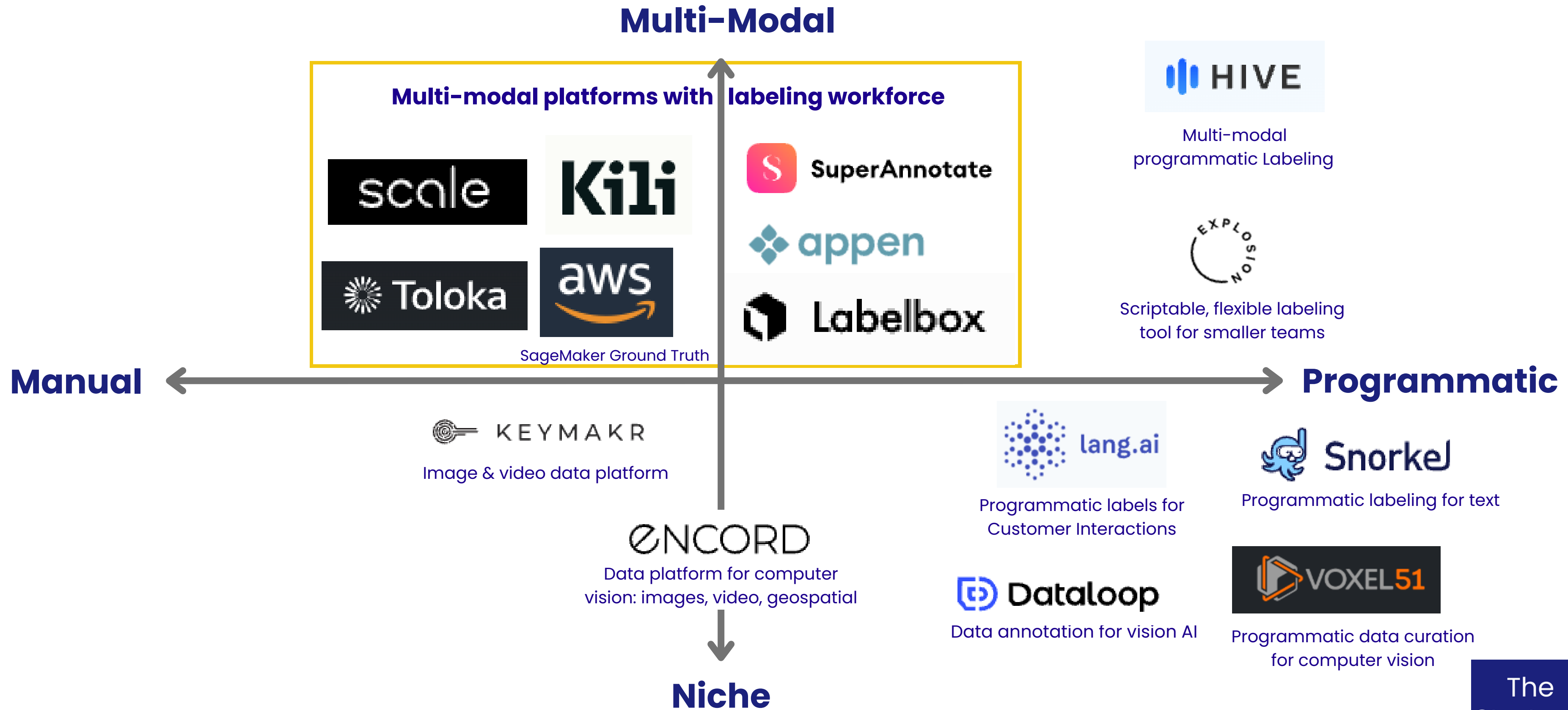
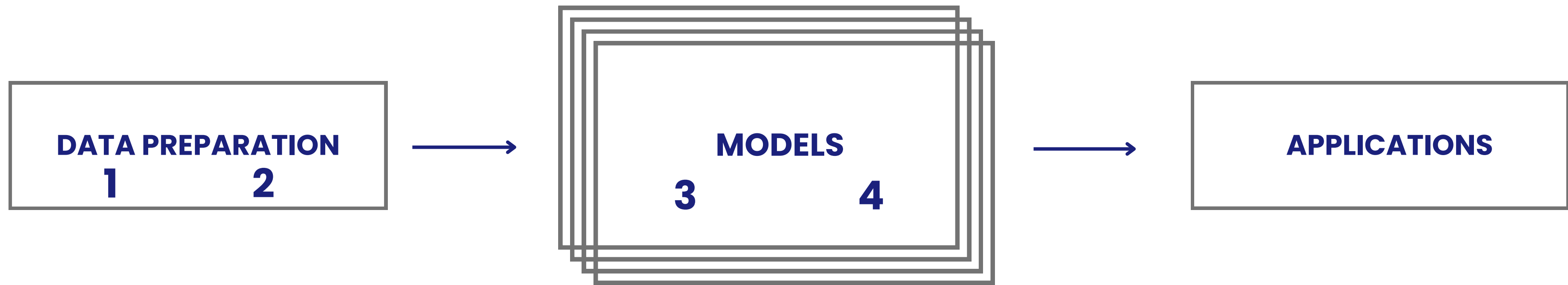


Machine Learning Data Labeling Platforms



Data Labeling Structures

Information to Make It Meaningful



1 Label and correct the initial data set

2 Discover errors and biases

3 Fine-tune the model for a specific use case

4 Test and validate quality of model output

Multi-Modal, General Purpose Tools

Facilitate Manual Labeling



SageMaker Ground Truth



Typical offering includes:

- Support for multiple data types
- Integration with popular ML frameworks and Ops tools
- Access to specialized labeling workforce
- Programmatic QA

Company



- Confidence to Deploy World-class AI
- **Publicly traded**
- HQ in **Chatswood, Australia**
- **<https://appen.com/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- **LLM Data Products** provides all the tools you need to fine-tune, evaluate, and monitor your LLM application
- **Fine-tuning**: jumpstart the tuning of your custom LLM with curated datasets pre-validated for your specific domain, industry, or use case
- **Red-teaming**: expose model vulnerabilities and safeguard both your clients and brand.

Value Proposition & Use Cases

- Finance, healthcare, e-commerce, insurance, retail, marketing, legal, social media, customer center, manufacturing
- Supported data types: images, video, audio, text, point cloud

Company



- Create high-quality datasets for training ML models
- **Corporate funding**
- HQ in **Seattle, WA**
- **<https://aws.amazon.com/sagemaker/data-labeling/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Label raw data, such as **images, text files, and videos**, and **generate labeled synthetic data** to create high-quality datasets for training ML models
- SageMaker Ground Truth Plus: **Fully-managed service** to create high-quality training datasets without having to build labeling applications or manage labeling workforces on your own
- SageMaker Ground Truth: build and **manage your own data labeling workflows and workforce**

Value Proposition & Use Cases

- Generative AI applications, NLP, Computer Vision, 3D LiDAR Navigation

Company

The logo for Kili, featuring the word "Kili" in a bold, black, sans-serif font. The letters are slightly shadowed, giving it a 3D appearance as if it's floating above a light-colored surface.

- Mission: enable businesses around the world to build trustworthy AI with high-quality data
- **US\$25M Series A** – Jul 2021
- HQ in **Paris, France**
- **<https://kili-technology.com/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Annotate all types of data rapidly and accurately with customizable tasks
- Supervise data quality levels
- Manage programmatic QA
- Data types: **image, video, text, PDF, documents, satellite imagery**

Value Proposition & Use Cases

- **Customizable interfaces** for the various data types
- Control the labeling queue to **prioritize assets, assign to specific labelers** or add validation rules. **Review pipeline** to spot inconsistencies and send assets back.

Company



- Unlock your data. Unleash your AI.
- **US\$110M Series D** – Jan 2022
- HQ in **San Francisco, CA**
- **<https://labelbox.com/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Catalog: search and **annotate unstructured data**
- Annotate: label data with **AI assistance and humans**
- Model: **prelabel data and evaluate models**
- Boost: on demand **labeling services**

Value Proposition & Use Cases

- Insurance, retail and e-commerce, healthcare, media and entertainment, sensors and robotics, LLMs

Company



- Mission: accelerate the development of AI Applications
- **US\$325M Series E** – Apr 2021
- HQ in **San Francisco, CA**
- **<https://scale.com/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Scale Rapid: **out-of-the-box RLHF workflows** to launch projects to Scale's workforce, get high quality text generations and evaluate model output
- Scale Spellbook – **compare experiments** across different LLMs, prompts and fine tuning strategies
- **Specialized labeling workforce** across languages and domains

Value Proposition & Use Cases

- Scale Forge: Ai-powered marketing suite
- Scale E-commerce for online shopping data
- Scale Donovan for government use cases

Company



SuperAnnotate

- Your AI Deserves SuperData
- **US\$14.5M Series A** – Jul 2021
- HQ in **San Mateo, CA**
- **<https://www.superannotate.com>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Data Studio: **annotate, hire specialized classification workforce**, manage annotation projects
- ML Studio: **AI Data Management and curation**, MLOps and automations, integrations and security
- Data types: image, video, text, LiDAR, audio

Value Proposition & Use Cases

- Agriculture, Healthcare, Insurance, Sports, Autonomous Driving, Robotics, Aerial Imagery, NLP and Document Processing, Security and Surveillance

Company



- Powering AI with human insight
- HQ in **Lucerne, Switzerland**
- **<https://toloka.ai/>**
- Founded in 2014

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- A data-centric environment to support fast and scalable **AI development with the help of human insight**
- Human-labeled data, ready-to-use pre-trained LLMs, Human-in-the loop flows for model evaluation and retraining

Value Proposition & Use Cases

- Content moderation, e-commerce, enhanced site search, computer vision

Programmatic Labeling Tools Focus on Specific Data Types, Niches



Multi-modal
programmatic Labeling



Scriptable, flexible labeling
tool for smaller teams



Programmatic labels for
Customer Interactions



Programmatic labeling for text



Programmatic data curation
for computer vision

Niches for programmatic labeling include:

- Customer Interactions
- Smaller teams
- Specific data types: text, images / video

Company



- Developer tools for AI and NLP. Makers of spaCy and Prodigy
- **US\$6M Series A - Sep 2022**
- HQ in **Berlin, Germany**
- **<https://explosion.ai>**

Partnerships & GTM

- Freemium

Tech Stack & Product

- **Prodigy**: scriptable, multi-modal data annotation tool
- **Python library** with pre-built workflows and command-line commands for various tasks, and well-documented components for implementing workflow scripts.
- Features: named entity recognition, span categorization, text classification, A/B Evaluation,

Value Proposition & Use Cases

- The best tool possible for **small teams working on rapidly evolving projects.**
- Prodigy Teams (coming) will help organizations keep all their annotation work in one place, and make it easier to collaborate on larger-scale projects.

Company



- Cloud-based AI solutions for understanding content
- **US\$50M Series D** – Apr 2021
- HQ in **San Francisco, CA**
- **<https://thehive.ai/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- **Content moderation and labeling** platform: classify and detect content attributes, search content and datasets, generate images and text from prompts
- Data types: **images, videos, GIFs, livestreams, text, audio, visual context, OCR**

Value Proposition & Use Cases

- **Protect digital content ownership, monitor and measure cross-platform ad inventory**
- **Spot deepfakes, detect AI-generated artwork**
- Customers: NFT platforms, marketplaces, dating apps, online communities, brands, publishers, ad agencies

Company



- Mission: empower everyone to benefit from the power of AI
- **US\$10.5M Series A** - May 2022
- HQ in **Madrid, Spain**
- **<https://www.lang.ai/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Unlock the power of CX Data with easy-to-use-AI
- Lang Data Operations **auto-tags and routes tickets and automates workflows**
- Lang Optimize **flags inefficient agent workflows** and new issues to reduce operational bottlenecks
- Lang Predict **pinpoints and notifies of issues** affecting key moments in the customer journey

Value Proposition & Use Cases

- For **Customer Experience organizations** at D2C companies and **Financial Services** organizations
- Discover customer intents, categorize them, automate workflows for each category, manage classification of customer interactions

Company



- Mission: empower everyone to solve their most impactful problems through data-centric AI
- **US\$85M Series C** – Aug 2021
- HQ in **Redwood City, CA**
- **<https://snorkel.ai/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Snorkel Flow – **programmatically label training data and correct model errors** through labeling functions
- Snorkel GenFlow – **instruction-tuning and RLHF** to build, manage and deploy gen AI apps
- Snorkel Foundry – **pre-train domain specific LLMs** by programmatically sampling, filtering, cleaning and augmenting proprietary data

Value Proposition & Use Cases

- Build GPT-You with programmatic data cleaning, sampling, curation, labeling
- Streamline data labeling and model training to reduce AI development costs and time by 90%

Company



- Mission: bring transparency and clarity to the world's data
- **US\$12.5M Series A – Sep 2022**
- HQ in **Ann Arbor, MI**
- **<https://voxel51.com>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- **Open source toolkit** that enables you to build better computer vision workflows by improving the quality of your datasets and delivering insights about your models
- Supported data types: image, video, 3D point cloud data
- VoxelGPT – **AI assistant for computer vision** that translates natural language prompts into actions to organize and explore data

Value Proposition & Use Cases

- Visualize your data, Curate high quality datasets, train better models, find mistakes

Computer Vision Tools Allow Manual and Programmatic Labeling



Data annotation for
vision AI



Data platform for computer
vision: images, video, geospatial



Image & video data
platform

Computer vision use cases include:

- Healthcare
- Drones and aerial imagery
- Robotics
- Autonomous vehicles
- Agriculture
- Media and content
- Retail and commerce analysis

Company



- The Data Engine for AI
- **US\$325M Series E** – Apr 2021
- HQ in **San Francisco, CA**
- **<https://dataloop.ai/>**

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Powering vision AI across industries: **annotation tools, automatic annotations, workforce management, data QA and verification, integrated labeling services**
- Covering the **entire data management cycle**, from data labeling, automating data ops, deploying production pipelines, and human-in-the-loop

Value Proposition & Use Cases

- Retail, drones and aerial imagery, robotics, autonomous vehicles, agriculture, media and content

Company

 ENCORD

- Unlock AI from Your Data
- **US\$13.7M Series A** - Oct 2021
- HQ in **London, UK**
- <https://encord.com/>

Partnerships & GTM

- B2B SaaS

Tech Stack & Product

- Annotate: automatic **annotation platform** to accelerate data labeling
- Active: actionable insights to **improve model quality** and fix outliers in your training
- Accelerate: on-demand, **specialized labeling services** to help you scale
- For **computer vision** data: **images, video, medical imagery, geospatial data**

Value Proposition & Use Cases

- Healthcare, aerospace and defense, computer vision

Company



- Better AI Unlocks a Better World
- HQ in **Tel Aviv, Israel**
- <https://keymakr.com/>

Partnerships & GTM

- B2B SaaS

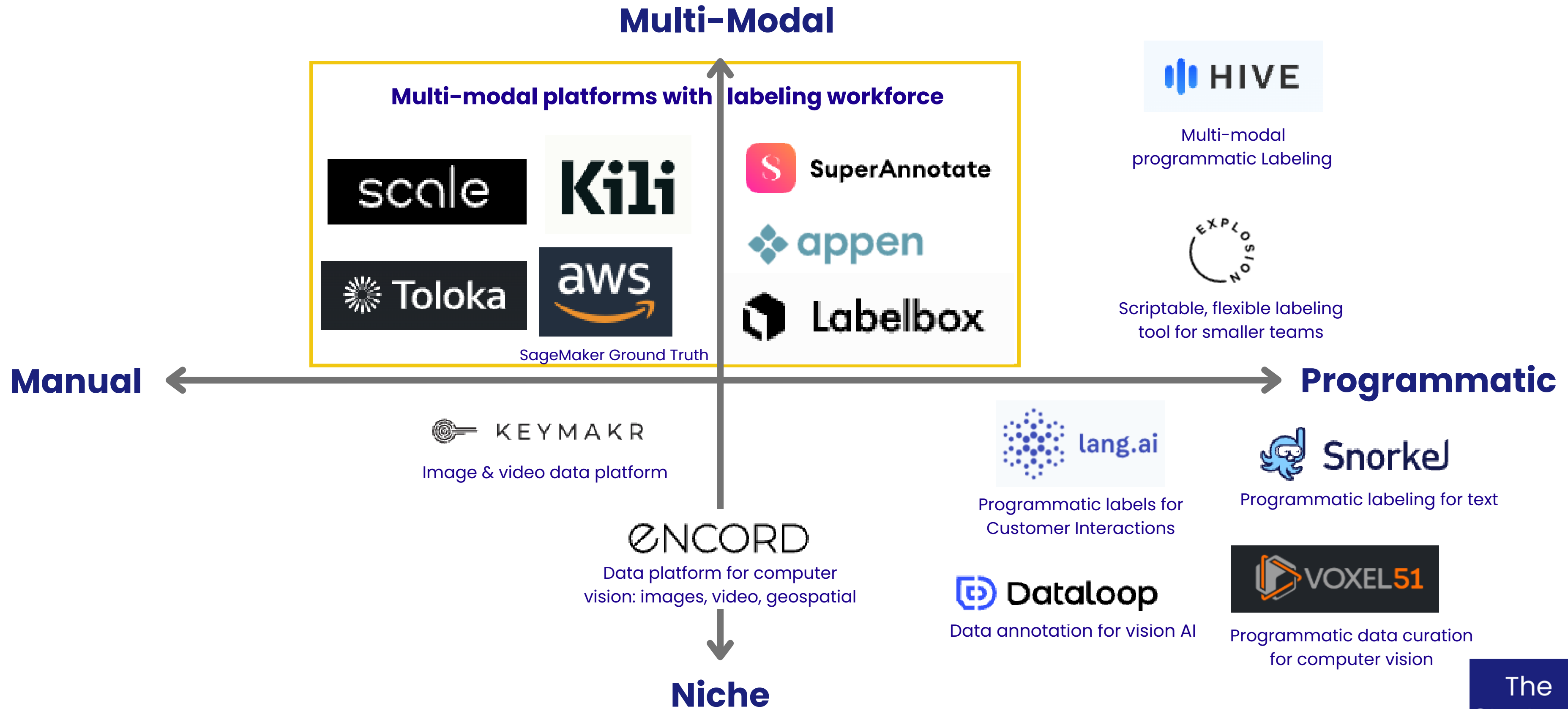
Tech Stack & Product

- We specialize in **annotating videos** and creating consistent high-quality data for your machine learning models.
- Supported data types: **images, videos, 3D**

Value Proposition & Use Cases

- Automotive, security, retail, fashion, sport, agriculture, livestock management, aerial, medical, robotics, waste management
- Facial recognition, medical annotations

Machine Learning Data Labeling Platforms



Need Help with **Product Strategy** for your AI app?

➤ Alex Irina Sandu



➤ alex@thestrategydeck.com

➤ www.thestrategydeck.com