

Navigating the Transition: From Models to Deployed AI



Swetabh Pathak

CTO & Co-Founder, Elucidata

Emerging Leaders for **AI in Life Sciences R&D**



..with Differentiated Technology & Expertise

#1

30+ customer proof points across discovery, development and trials

#2

Cloud-first data platform and ML that can seamlessly complement your ecosystem.

#3

With a team that can quickly assimilate domain expertise into high impact services.

2023 Frost & Sullivan Enabling Technology Leadership Award

The North America Tech-Enabled Drug Discovery Industry

Industry Research Analysis / 9B43 / B8

Published: 2023-09-19



FASTCOMPANY

Most Innovative Companies

 **OWKIN**

EXELIXIS[®]

janssen
PHARMACEUTICAL COMPANIES OF
Johnson & Johnson

AURON
THERAPEUTICS

 **DeepMind**

AI Shows Promise, But **Opinions Vary..**



“

AI is the future - it will help us explore areas that have never been explored before. One day AI will help us understand biology so deeply that we can form new scientific laws and drug design principles.

Head of Data and Platform Strategy, 'AI-first' biotech

Innovators
(2.5%)

“

AI is somewhat valuable. In our work, AI has helped make a lot of molecules synthesizable faster & cheaper.

Translational Scientist, Academia

Early Adopters
(13.5%)

“

It's too early to recognize the true impact of AI - we will only be able to see the true impact once we can see the productivity over time.

Deputy Director, Global Health, Non-profit Organisation

Early Majority
(34%)

“

AI is currently only used for solving simple problems. The InSilico screen would only have had a 4% failure rate, even without AI.

Computational Biologist, Research Institute

Late Majority
(34%)

“

AI is a new hype - investors buy into the desire to be hip, cash is raised, Pharma cos do deals to be in the news. There's lots of noise in this field but it has not been proven yet.

Chief Executive, Data Consortium

Laggards
(16.5%)

Biggest Barriers to **Broader AI Adoption**



Limiting Belief 1

‘We don’t have a **relevant use case for AI**’

*Translational Leader,
Mid-Stage
Pharmaceutical
Company*

Limiting Belief 2

‘We just couldn’t get **enough training data** to solve the problem we’re working on’

*Computational
Biologist, Research
Institute*

Limiting Belief 3

‘We don’t have **the expertise or the millions of dollars** needed to build a team’

*CSO, Early Stage
Therapeutics*



Limiting Belief 1

“We don’t have a ***relevant use case*** for AI”



Good AI use cases are not **rare events**



**What do good
use-cases look
like?**

Access to 'high quality and relevant' data

Human supervision is available & possible

Biological rules can be framed

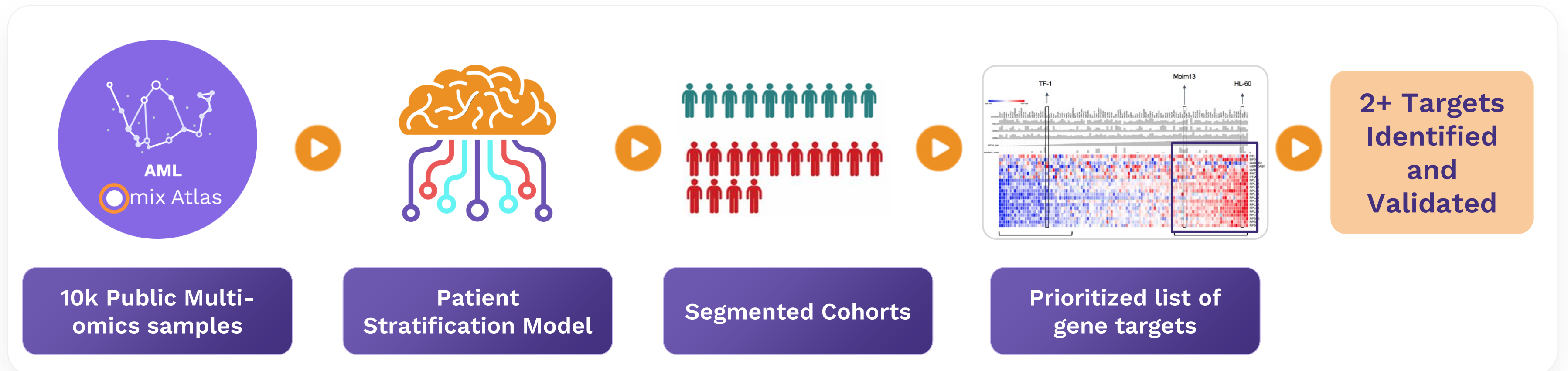
Hypothesis generation rather than testing

Explainability is not necessary

Right use case is the **biggest predictor of success**



‘An **Early Stage Therapeutics company** wanted to develop and train **Classifier Models** to segment patients in AML’



AI was used to **assist domain experts**, to solve a **well-defined** problem with **clear** outcomes: ‘Segment the patient cohorts in AML as per their prognosis’

LLMs were Deployed to **Advance Target Identification**



Polly Chatbot
Version 4.0 Mar 14

Examples

[Create a pie chart of the distribution of the top 9 tissues in depmap. →](#)


[What is the total number of cell lines that are present in depmap which are related to skin? →](#)

[Explain quantum computing in simple terms. →](#)

[How do I make an HTTP request in python? →](#)

[Python program for generate pandas dataframe. →](#)

[Start a new Thread](#)

Send a message 



Limiting Belief 2

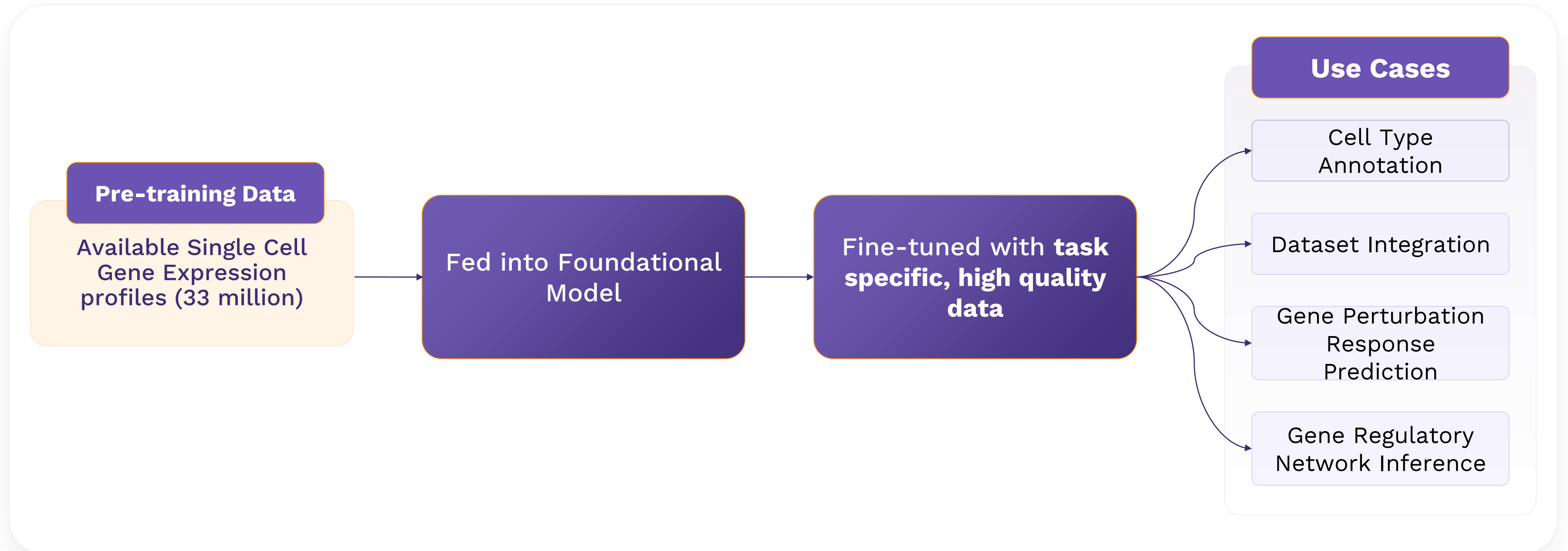
“We just couldn’t get **enough training data** to solve the problem we’re working on”



Make the Shift to **Data-Centric AI**



Fine-tune existing models with **high quality and relevant data**. Especially useful for predicting long-tail problems with **limited data points (<10,000)**



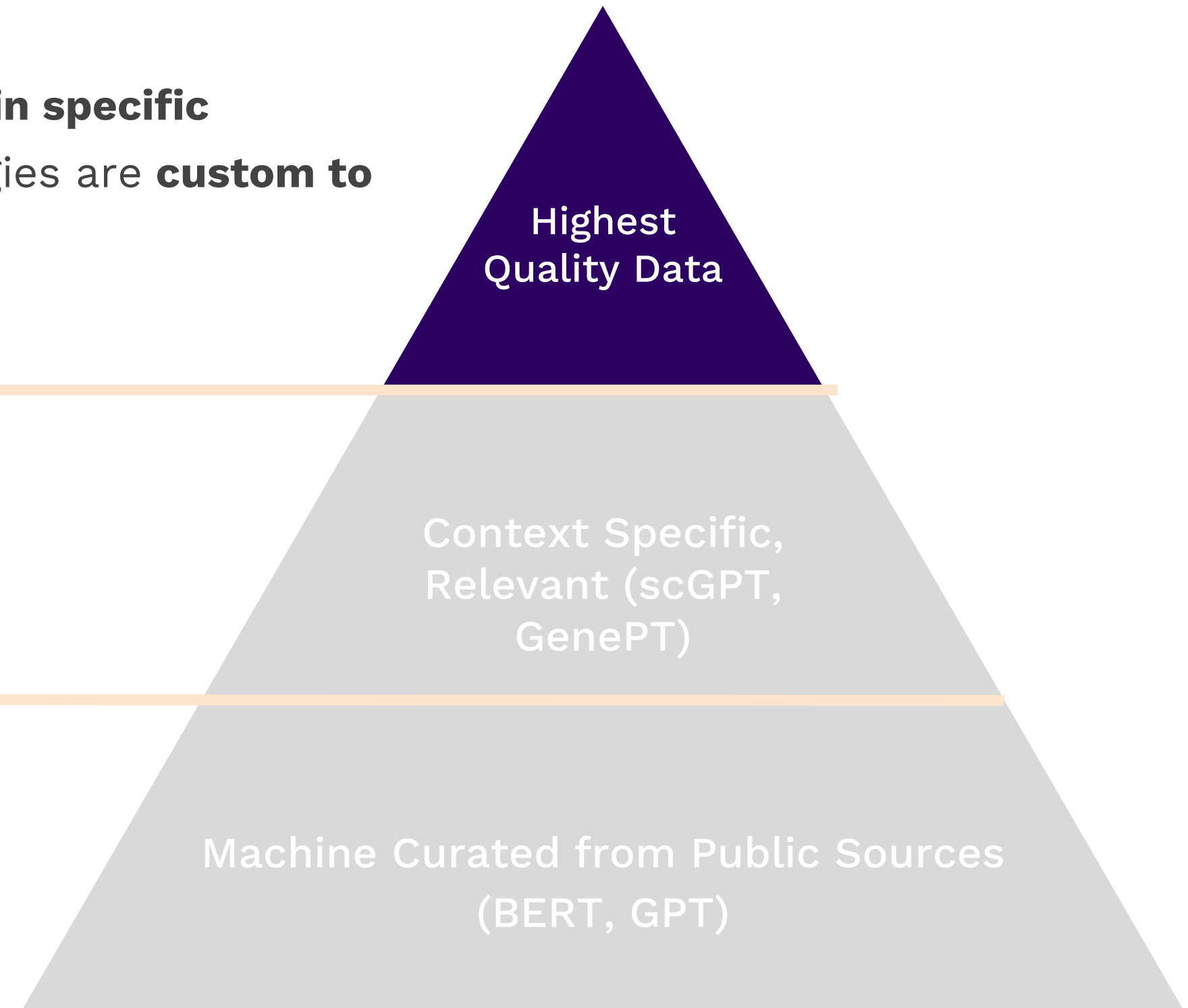
How do we define **High Quality Data**?



- **Up-to-date** with patient information, **domain specific**
- Metadata annotations, processing & ontologies are **custom to the use case / task**
- QC-ed by **experts**

- Annotated with **critical metadata**
- **Relevant** to the biological domain

- Ingested & transformed into **Machine readable formats**
- Structured into **tabular files** (CSV, JSON)



How well does scGPT perform after **Fine-Tuning with High Quality Data?**

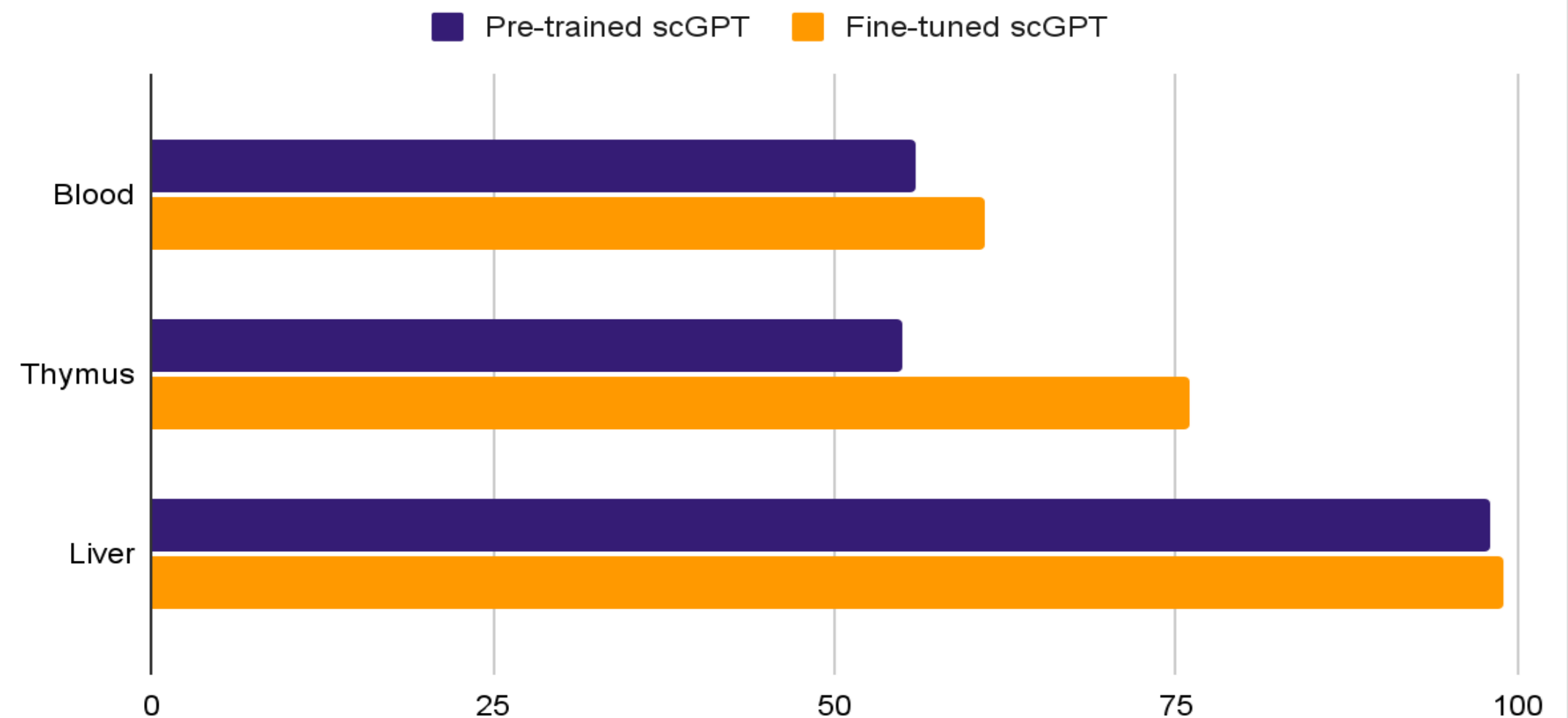


scGPT can perform **reference based cell type annotation** in a **zero shot setting**.
However, fine-tuning with high quality data **improves model performance by 20% (avg)**

Experimental Design

- Training Dataset: **25k** immune cells from **HCA** to fine-tune
- Testing Dataset: **13k** immune cells from **Tabula Sapiens**
- All datasets were cleaned and linked with **Elucidata's Harmonization Engine**.

F1 Scores: Pre-trained scGPT vs Fine-Tuned scGPT

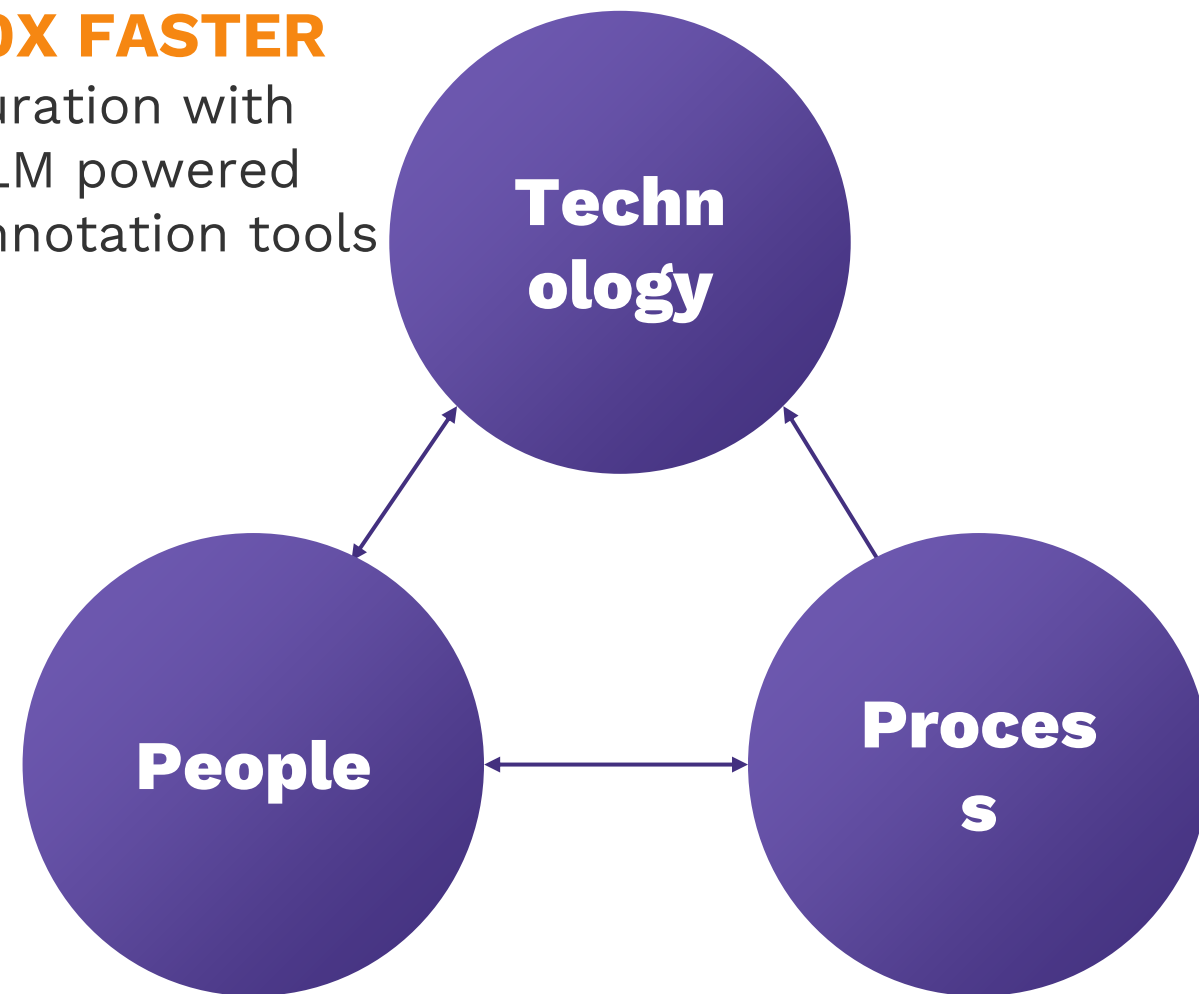


Elucidata's **Harmonization Engine** Cleans the Data you Need



10X FASTER

curation with
LLM powered
annotation tools



100+ Experts

In curation, NLP, data
engineering, &
bioinformatics

99.99% Accurate

Data delivered with
robust QA/QC

50 Million

Samples harmonized to support use cases
in drug discovery, development & trials

25+ Data Types

Supported including RWE, Omics and
Clinical



30+

Data Pipelines built and maintained on
Elucidata infrastructure to process data



Limiting Belief 3

“We don’t have **the expertise or millions** needed to build a team”



Case Study: Building **Production-ready** ADMET Pipelines



SCENARIO

‘This mid-cap pharmaceutical company wanted to develop an **end-to-end ADMET prediction pipeline** that would support **5 lead development programs** across neurology and oncology’.

THEIR NEEDS

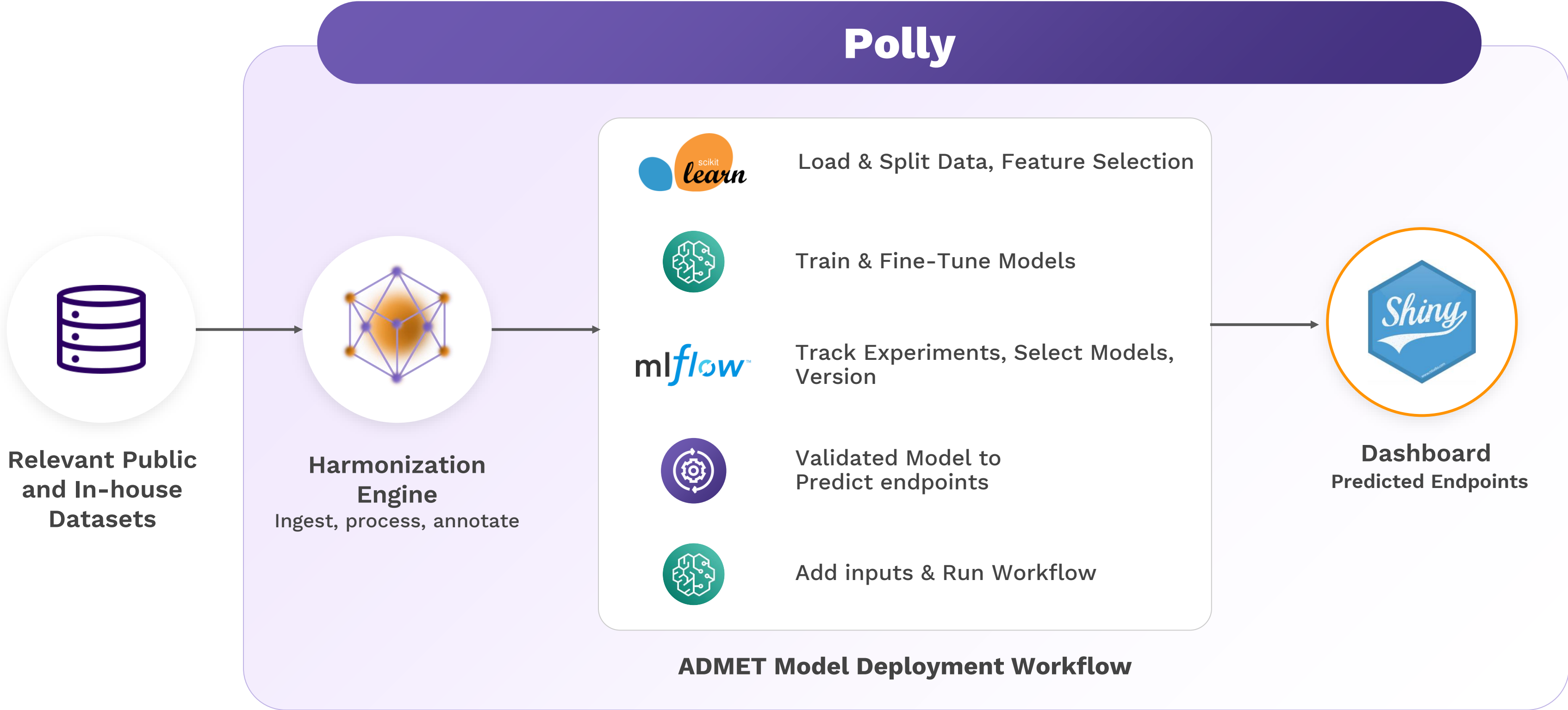
- Collect and prepare all the data generated across assays in a meaningful and scalable way.
- Productionize existing models on the cloud, so that they can run at scale.
- Develop an ML-ops infrastructure to manage the data & models across stakeholders, multiple sites and different types of users.

Doing this in-house needs a **team of 7 FTEs** and **cloud resources**.
Costs drum upto **~\$2 Million** and projects could take **1+ years to kick start**.

Scaling up AI in production, **Set up in 1.5 Months**



Compound Screening & Evaluation **Accelerated by 2X** with Production-Ready ADMET pipeline



Significant **R&D Productivity** Unlocked

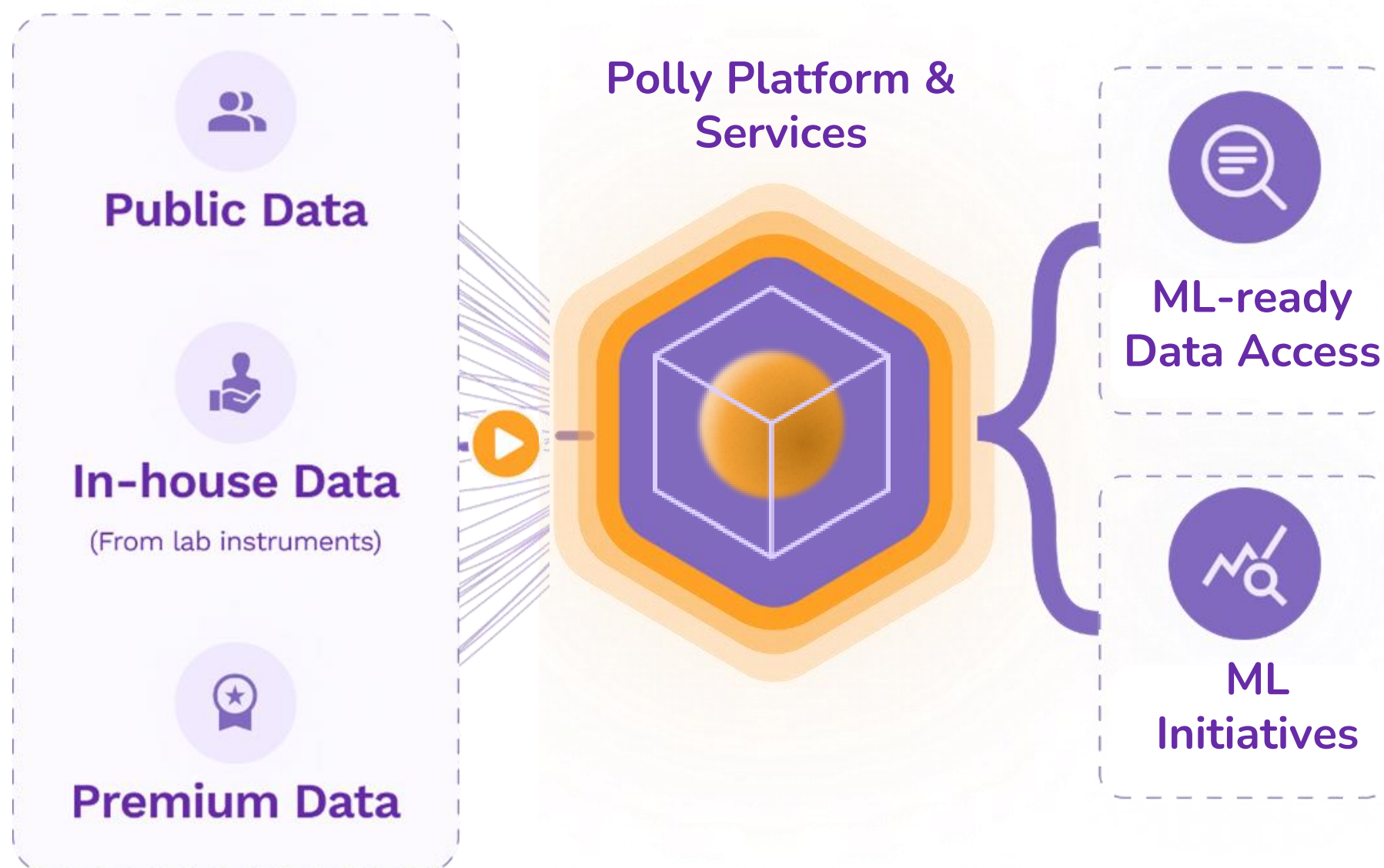


Projects could be kickstarted within **a month, at 4X Lower Costs.**

Productivity Areas	Improvement with Elucidata	Rationale
Data Acquisition	4X Faster	<ul style="list-style-type: none">● Dedicated team to perform searches in Public Databases
Data Preparation, Annotation, QC	4X Faster	<ul style="list-style-type: none">● LLM-powered Harmonization reduced manual effort in data preparation.
Model Development / Deployment Cycle	Reduced by 30%	<ul style="list-style-type: none">● Key bottleneck steps in the process (ingestion, cleaning, ML Model Deployment, Versioning) automated.

Any Questions?

Reach out at elucidata.io to know more!



"The value is in the data, it is not in the tools. That is the one thing, it's a bit of a hobby horse for me. One thing always point to in these discussions around data, don't underestimate the amount of time and value in doing what is really often difficult and not so rewarding directly work, like cleaning data sets isn't always fun, but it is often the most valuable thing you can do."

Dr. Jeffrey Reid,
Regeneron's Chief Data Officer