ArcaScience

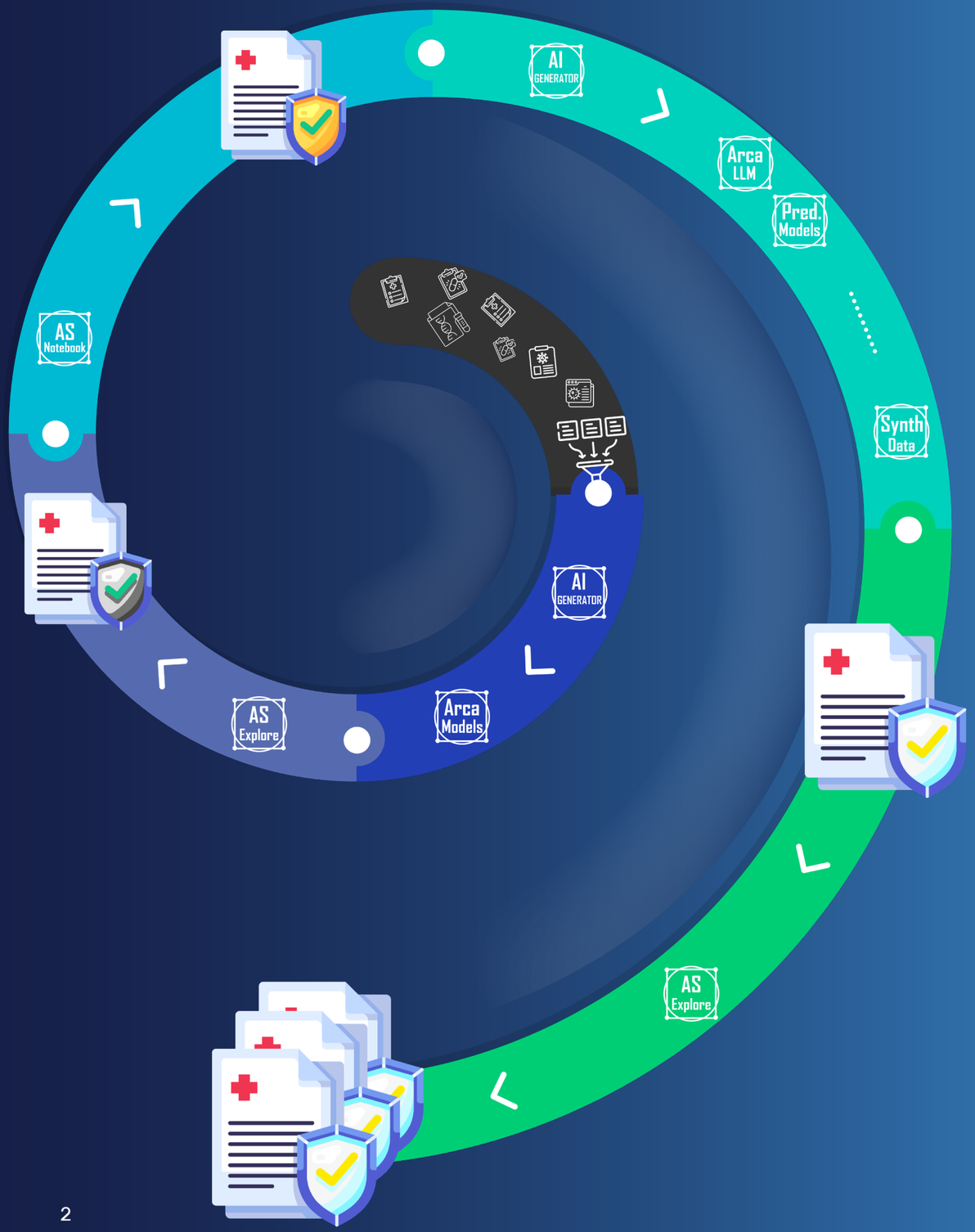
ARCASCIENCE VISION

An ArcaScience Whitepaper

www.arcascience.org

The Curve

How ArcaScience is scaling



6 phases to hypergrowth

1-HARMONIZATION

Arca IDF

10 deep learning models working hand-in-hand to unlock biomedical data

2-LANGUAGE MODEL BUILDING

Arca PV · Arca PK/PD · Arca P.P · Arca EFCY

Specialized language models are trained on top of the biggest, most qualified volumes of biomedical data available

3- DATASET GENERATION

AS Explore

The user queries our proprietary search engine to instantly generate datasets showing risks, benefits, patients profiles and quality of drugs

4- REFINEMENT

AS Notebook

The user refines his dataset to identify patterns and insights that might otherwise go unnoticed leading to breakthroughs in medicine and healthcare

5- AMPLIFICATION

Arca LLM · Pred. Models · Synth Data

Refined data is used to :

- train the first biomedical models - LLM & Predictive models - improved by its users
- generate synthetic data

The more user queries, the smarter ArcaScience becomes

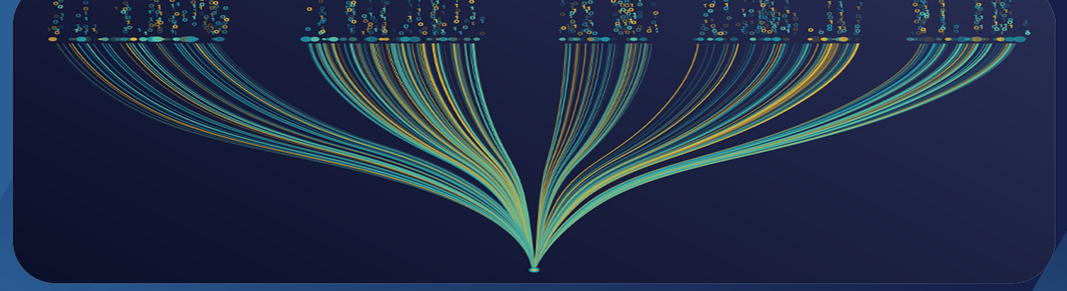
6- EXPONENTIAL FINDINGS

The newly generated data are re-integrated in the search engine and used to improve future user queries. Data are then fed to predictive models and ArcaLLM becomes exponentially efficient

1st Phase

Harmonization

“A unique AI, bringing the source of ArcaScience’s Language Models”



Required skills

- **Document engineering and classification**
- **Knowledge of language resources standards and metadata requirements**
- **Ontology development methodologies**
- **Biomedical concepts modeling**

“I need to leverage every possible data for my research”

The very first phase is about gathering every possible biomedical data needed by the user and convert it into a unique format.

Collecting an extremely large and comprehensive dataset can be a particularly difficult and time-consuming for scientific experts. Without a solution like ArcaIDF, users are likely to miss 80% of the information they are looking for.

ArcaIDF’s datasources include internal data, open data, repositories, and subscriptions to specialized sources, **built on our client’s private cloud, without externalizing any sensitive information.** With ArcaIDF, users gain access to all the necessary information in one place, making their research efficient without missing any valuable insight.

Knowledge management savings at **GSK (2019)**	Percentage of newly reachable data	Data growth rate in the biopharmaceutical industry
Savings	Percentage	rate
+1,7m	**82%**	**x2/72d**

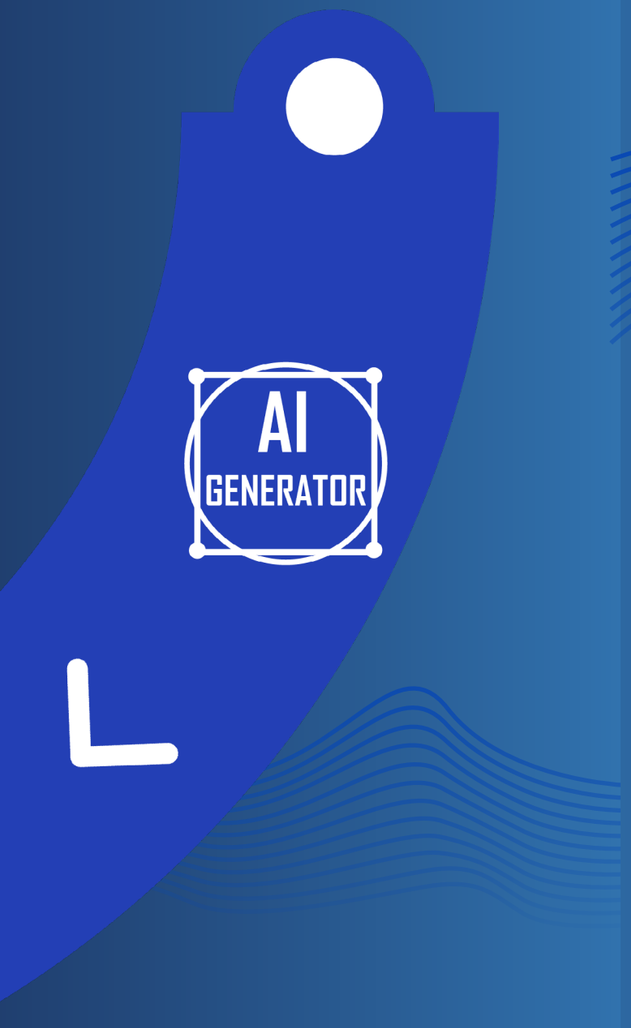
Technology : 10 deep learning models

ArcaIDF harmonizes unstructured knowledge (e.g. scientific articles, clinical trials, omics data and many more) and semi-structured knowledge (e.g. safety databases, Real-World Evidence databases, specialized vocabularies, etc.). All of these precious data are rendered accessible with the aim to identify comparable relevant datasets and deliver them to the user, and to our AI Generator.

2nd Phase

LANGUAGE MODEL BUILDING

“Newly accessible biomedical data train the most efficient biomedical language models”



Required skills

- **Machine learning algorithmic, feature engineering, Data sampling**
- **Natural language processing and Information extraction techniques**
- **ML evaluation & Cloud Computing**
- **Deep understanding of biomedical concepts**

“I need to review millions of documents at once”

After solving the accessibility issue, it is common for a clinician to look for specific insights – biomarkers, adverse events, genes, proteins, etc.-. Yet, such information do not emerge from their standard solution. Classic search engines are not designed to understand the science, and researchers end up spending up to 2 months per year browsing no more than 20% of the available data.

That’s where our AI Generator comes in. It is a framework that trains semantic language models which are able to interpret the context of biomedical phenomenoms. Its training is based on the selection of the most representative medical data, while leveraging the previously unreachable 80% sources.

Avg ArcaSciences’ language models accuracy	Fully trained specialized models	Complementary knowledge reached (Sanofi TS - Multiple Sclerosis, 2023)
96%	Language models **4**	rate **x11**

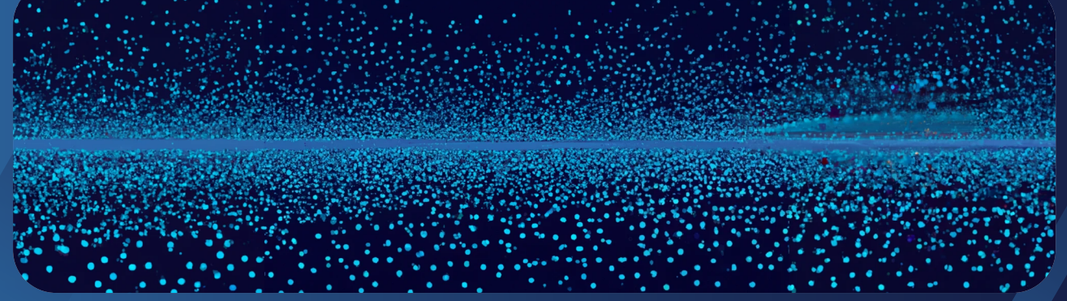
Technology : a unique biomedical AI training pipeline

This unique pipeline picks-up and fine-tune the most adapted language models for the identification of different biomedical classes of contexts. The first challenge we solved was to meet the requirements of advanced deep learning architectures with the data analysis needs of biomedical researchers. The second challenge we solved, besides providing the most accurate semantic output, is that the methodology and the workflow should be standardized and enable scaling-up the creation of future models that will be generated from the interaction of the user with the output of the first layer of semantic models.

3rd Phase

Dataset Generation

“The simplest solution for reaching the widest and most qualified biomedical information”



Required skills

- **Data integration and database merging**
- **Assembling end-to-end NLP pipelines**
- **Information retrieval technologies**
- **UI visualization for biomedical research**

“I need to export everything about risk, benefits and patient profiles related to my drug”

Once the volumes of data were harmonized and we unlocked the ability to understand the data just like a doctor would with our language models, a single platform replaced the multitude of databases and tools used daily. In a glance the user sees a trend, or refine his research by checking a certain dosage, associated pathology or adverse event, in order to find the right dataset easily.

AS Explore highlights directly in the data the datapoint showing risk, efficacy and patient profiles relevant to the user. The automation of these steps through our language models allows to accelerate clinical trials and to reduce the risk of error or oversight.

Volumes of qualified data reached (Sanofi CHC - 2022)	Time spent generating the qualified datasets (Sanofi TS - 2023)	Amount saved due to toxicity risk per drug (Brain Institute - 2022)
rate		in dollars
x9	**1h (vs 2months)**	**$182m**

Technology : 1 AI-based search engine

The structured format of each document coupled with several additional semantic layers are indexed along with the vocabularies corresponding to each layer to ease querying the complementary layers. They are all indexed in one visual tool that has enhanced capabilities of a search engine joined by several visualizations to slice and dice different levels of data. A user can export entirely or partially what he sees in the results of his query which corresponds to a qualified dataset that encapsulates a combination of parameters to be investigated and enlarged through the AS Notebook framework.

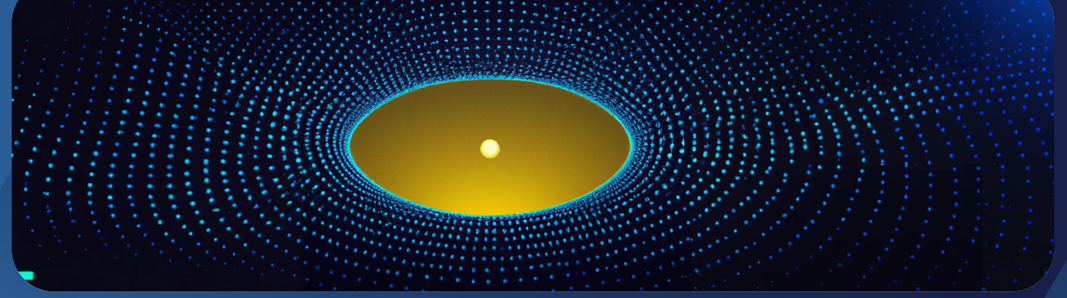
4th Phase

Refinement

“Reaching the most refined dataset”

4

AS Notebook



Required skills

Biomedical expertize

Data qualification

“I need to refine and discover new biomarkers involved in thrombosis cases related to my drug”

Once the export is generated out of AS Explore, it is time to add the users’ expertize on top of it. This step is a synergistic action between human intelligence and AI. AS Notebook is an extremely powerful solution used for refining datasets based on smart filters. The added value of the Notebook is the possibility to customize one’s research according to each specific study case, and make connection between patient’s characteristics, pathologies, drugs, environment, among others.
The final result reach highly advanced form of qualified biomedical information involving both the widest, most representative databases, leveraged through language models and refined by biomedical experts. Additionnaly, these datasets are crucial when it comes to build synthetic data, feed drug discovery algorithms, train new models and generate a biomedically-accurate LLM.

Time saved to reach repurposeable drugs (Brain Institute 2021)	Time spent generating a comprehensive Benefit-risk assessment (Sanofi TS - 2023)	Volumes of new qualified datapoints fit for feeding new models (Sanofi TS - 2023)
rate **71%**	**2weeks** (vs 18 months)	rate **x7,5**

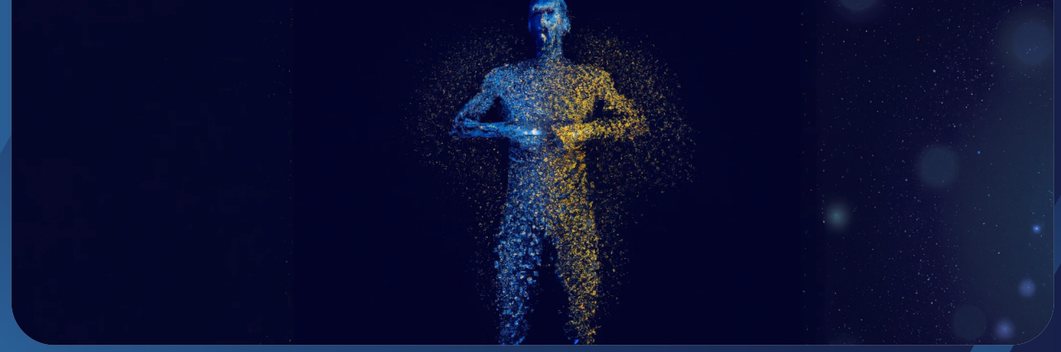
Technology : 1 human-based drilldown solution

An AS Explore export can be further manipulated by the user to add more ad-hoc data-points and filters to apply to an already qualified layer of data. Through the interaction of the user with his export (removing duplicates, adding patient conditions, enriching the list molecules, etc.), a second qualification is carried out on the qualified data which will constitute the base for fine-tuning a LLM to generate a new specialized model for that specific use case.

5th Phase

Amplification

“ArcaScience unfolds a unique autonomous self-training ecosystem applied to biomedical research”



Required skills

- **Pre-training and fine-tuning LLMs**
- **Data privacy and security, privacy preserving techniques**
- **Statistical analysis and modeling**
- **Language model pre-training and fine-tuning processes**

“I need to amplify my dataset and use it to simulate human response to a treatment”

After generating the final dataset in AS Notebook, ArcaScience unlocks a virtuous cycle leading to the generation of brand new biomedical insights where users play a key role in the process of validation. Using Language Models, we analyze, identify patterns, and generate synthetic data to feed our specialized LLM. Leveraging Large Language Models, we can identify key parameters in biomedical datasets with unparalleled speed and accuracy, amplifying our ability for human simulation and predictive medicine.

Volumes of new synthetic insights in rare diseases (Sanofi TS - 2023)	Percentage of drugs impacted (Takeda R&D - 2023)	Level of accuracy, sensitivity and specificity of synthetic-based models
growth	of the existing clinical pipeline	
382%	**95%**	**97%**

Technology : 1 LLM, many synthetic data, infinite models

The gold standard data being at the center of the generation of semantic models, the use of human-based validation and specialized LLMs are first to boost the quality of the extracted datasets by taking into account key parameters (e.g. mutations, inhibitors, proteins, tradenames, etc.). Second, the capacity of our LLM to extrapolate on similar contexts amplifies the adaptation of the creation of more specialized models to micro-tasks for micro-domains within different therapeutic areas.

With this approach, ArcaScience unfolds a **unique autonomous self-trained ecosystem** applied to biomedical research and scaling since january 2023 when we generated ouf first user-based synthetic dataset.

6th phase

Exponential findings

“ArcaScience builds a unique datasource based on user interaction and the most advanced biomedical language model, continuously improved by its users”

Required skills

Querying

“I need ArcaSciences’ unique datasets”

In late 2023, ArcaScience will hold a unique synthetic database growing exponentially.
The generative nature of our pipeline coupled with the right fine-tuning - based on learning instructions engineered from the interaction of the user with the notebook - is unlocking a snowball effect that :

- enables the generation of synthetic data while taking care of crucial parameters for the study
- drastically simplifies reaching critical observations, formulating hypotheses and designing experiments from massive knowledge learned from large data lakes of biomedical contents.

The challenges we are solving are two-folds:

- Selecting the right architecture to continuously train our LLM: encoder-decoder (BERT like model) or decoder only architecture (GPT like model). Target experiments should be carried out to find the less costly and most effective synthesizing approach. However, our user-based data generation logic leads to a high level of velocity associated with the most advanced biomedical expertize, and low costs. The size and the quality of the collected corpora becomes key advantages to ArcaScience.
- Fine-tuning: maximize the generation of qualifications resulting from the user interaction with the tool with output of the first layer of semantic models to create new ones by fine-tuning the LLM to the new task.

The process is therefore standardized to guarantee the maximum of automation and scalability.

Expected 6 month synthetic data growth	Expected 12 months synthetic data growth	Expected date when AS Explore will become an unavoidable R&D necessity
rate **x8**	rate **x63**	**Sept. 2024**

Technology : exponential loop of synthetic data

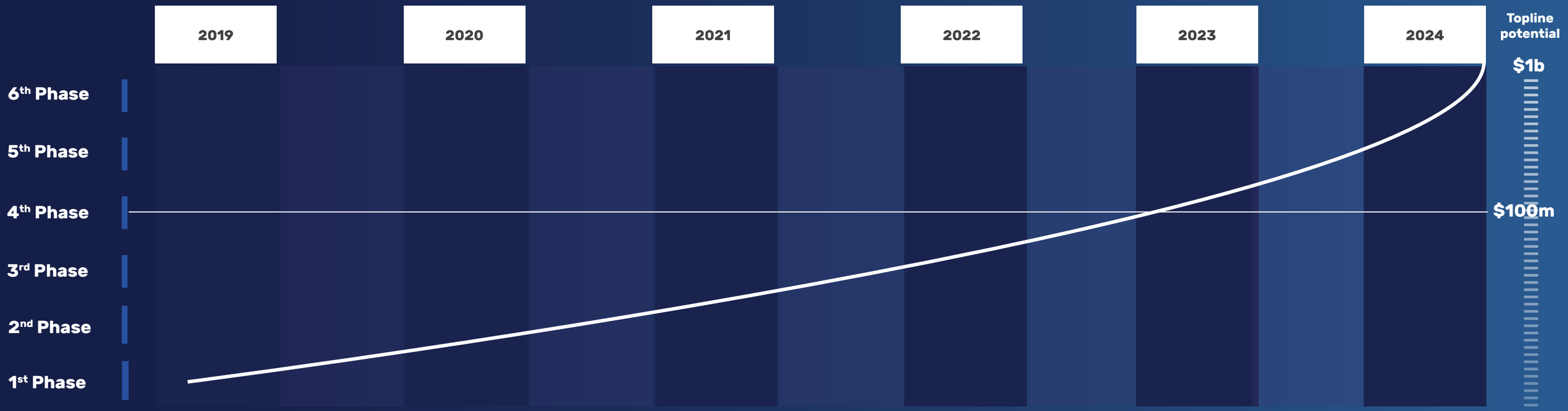
Bringing more qualified and approved synthetic data into AS Explore will lead to an exponential growth of biomedical insights, building a unique datasource based on user interaction and the most advanced biomedical language model, continuously improved by its users.

Process Timeline

Operations Plan

Strategic Goals and Objectives

- Streamline synthetic data into clients' drug disovery models by end 2023
- Reach 50% of synthetic data generated out of Arca LLM before 2024
- Enrich a unique synthetic database for rare diseases at the frontline of research
- Increase the value generated by a user to reach the value of an average contract before 2024
- Use Arca LLM for browsing AS Explore before end 2024



OPPORTUNITY

De-risking Clinical Trials

The financial risk represented by unsufficient database deep screening in R&D is of $182 million per drug (J.Wouters et. al.).

ArcaScience has proven to be able to reduce this risk by 12% - cost attached to significant data gaps in preclinical and phase 1 (Sanofi TS-2023)

Automated Pharmacovigilance

CAGR in the pharmacovigilance market is of 10,6%, growing from $6.82b spent in 2022, to $13,9b in 2030. The main opportunity lies in the phase IV (post-marketing), with 75% of the spendings. (R&MReports sept. 2022)
ArcaScience has proven to be able to speed up by 9 times the generation of meta-analysis while doubling its inputs. (Sanofi CHC - 2022)

Synthetic data generation

"By 2024, 60% of the data used for the development of AI and analytics projects will be syntehtically generated" (Gartner, 2021). Value attached to synthetic data in healthcare is mostly focused on simulating human where information is lacking, driving up the necessity of building synthetic data.
ArcaScience has shown its profficiency at providing such data in large batches.
(Sanofi TS & Takeda R&D - 2023)

Assisted drug discovery

Many research showed that synthetic data applied to drug discovery are a viable proxy for clinical trials (Azizi et. al.) and phase IV. With a CAGR of 45.7% (M&M Report 2022), this market shows the strongest growth in the industry, with very little actors, ArcaScience has proven its capability at building a pipeline for providing the key datasets for predicting drug efficiency and risks in drug discovery phases. (Brain Institute - 2023)

ArcaScience

Address

25 rue Coquillière
75001 Paris
France

Email 1 : contact@arcascience.org
Email 2 : info@arcascience.org

Website : www.arcascience.org