Generative AI for Healthcare

Razvan Ionasec, Christoph Russ, Regina Hackenberg, James Wiggins, Ujjwal Ratan, Khan Siddiqui, Jory Tremblay, Prabhu Arumugam, Wieland Sommer, Jeroen van der Laak, Mahesh Pancholi, Jonathan Larbey, Ruben Amarasingham, Mattia Capulli, Andrea Riposati, Alberto Rizzoli

This white paper addresses Healthcare and Lifesciences (HCLS) professionals involved in Generative AI activities.

Readers will:

- Understand emerging Generative AI technology in the broader context of the artificial intelligence and machine learning (AI/ML) field
- Become familiar with top use cases, first areas of commercial applications and business impact
- Have a grasp of the technology stack from silicon to application programming interfaces (APIs)
- Know the innovators in the HCLS and active fields of research science, and predictions for the future
- Get an overview on responsible AI, as well as safety, risks and other topics
- Access guidance on how to get started with Generative AI

This document has been produced by AWS in collaboration with worldwide industry leaders in HCLS.

Executive Summary

Generative artificial intelligence (AI) represents a paradigm shift in computation that is set to reshape the software and hardware landscape. Emerging capabilities like large language models that can understand and generate human-like text herald a new era of AI able to surpass average human performance across virtually all cognitive tasks. Over the next decade, generative AI is predicted to profoundly impact economies and industries worldwide. The core innovation enabling this advancement is foundation models - massive neural networks pretrained on diverse multi-modal data that encode general knowledge about the patterns and semantics of our world. Built on top of specialized hardware and infrastructure for training and running AI at scale, these models can then be fine-tuned with custom data to accomplish specific tasks from summarization to content creation and more. Leading technology providers are rapidly productizing access through developer platforms and APIs. In healthcare, generative AI promises to revolutionize everything from biopharmaceutical R&D to diagnosis, treatment decisions and patient experience. The McKinsey Global Institute (MGI) has estimated to unlock $60-110 billion in annual value through higher productivity and innovation. Scientists are using these technologies to accelerate discoveries by uncovering insights from complex healthcare data that human experts alone cannot feasibly analyze. At the frontlines of care delivery, generative AI can generate draft clinical notes, summarize electronic records, and answer patient questions - helping overwhelmed staff better focus on direct care needs. Looking ahead, continued advances in model scale, multi-modality and human-AI collaboration will further expand possibilities. However, thoughtfully addressing interpretability, transparency, bias and other responsible AI considerations remains vital for successfully translating these emerging capabilities into clinical practice over time. If embraced responsibly, generative AI may profoundly transform global healthcare through democratizing access to knowledge and expertise - unlocking a future of more informed, empowered and personalized care for all.

Innovation can transform industries

Introduction

In early 2023, AI took the world by storm. For the first time, more than 100 million people from around the world experienced the power of AI at their fingertips through consumer applications such as Midjourney for image generation and ChatGPT.

The new era of AI, Generative AI, is predicted to have profound socioeconomic impact. Over the next 10 years, two-thirds of all jobs in the U.S. and EU will be directly affected and the global gross domestic product should increase by 7%. This is no ordinary, incremental leap in AI —it's a genuine paradigm shift in computation that will reshape both the software and hardware world [1].

Generative AI is transforming all aspects of healthcare, from science to industry and delivery. It has been estimated to generate $60 billion to $110 billion a year in economic value for pharma and medical businesses [2].

This whitepaper will help Healthcare and Lifesciences (HCLS) professionals involved or interested in generative AI with:

- Understanding the emerging Generative AI technology in the broader context of the artificial intelligence and machine learning (AI/ML)
- Becoming familiar with top use cases, first areas of commercial applications and business impact in HCLS
- Understanding of the technology stack from silicon to application programming interfaces (APIs)
- Knowing the innovators in the HCLS and active fields of research science, and predictions for the future
- Getting an overview on responsible AI, safety, and risks
- Accessing guidance on how to get started with Generative AI

This document has been produced by AWS in collaboration with worldwide industry leaders in HCLS

[1] Generative AI could raise global GDP by 7% https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

[2] Generative AI in the pharmaceutical industry - Moving from hype to reality https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality

Brief history of AI and Machine Learning

The computation odyssey can be traced back to 2400 B.C. with inception of the abacus, an early but efficient device that streamlined arithmetic calculations. Centuries later, Blaise Pascal's calculator (1642) marked an advancement by building a mechanical device able to perform addition and subtraction. In 1694, Gottfried Wilhelm von Leibniz's invention expanded computational horizons, enabling all four arithmetic operations on a mechanical calculator.

Fast forward three centuries to inception of AI: The first trainable neural network, the perceptron, was demonstrated by Cornell University psychologist Frank Rosenblatt in 1958 [1]. The computation idea evolved from arithmetic to learnable functions and calculators became electronic computers - 5-tons machines and the size of the room using punch cards. The first truly successful realization of AI, through gradient optimization algorithms, was Adaptive Boosting [2]. This was formulated by Yoav Freund and Robert Schapire in 1995 and helped improve the performance of AI learning algorithms to become practically useful. Transitioning to the latter part of the 20th century, the saga of deep learning takes center stage. The breakthrough of AlexNet triumphed in the ImageNet competition [3], which led to eventually surpassing human performance at image recognition and subsequent milestones like AlphaZero's [4] superhuman-level chess prowess (2017). On the industry front, companies with AI in their DNA were emerging.

At Amazon, AI is part of the heritage and ethos, with large and varied businesses built on AI. The e-commerce recommendations engine is driven by machine learning; the paths that optimize robotic picking routes in the fulfilment centers are driven by ML; and supply chain, forecasting and capacity planning are informed by ML. Prime Air (Amazon drones) and the computer vision technology in Amazon Go — the physical retail experience that lets consumers send items off a sheld and leave the store without having to formally check out — use deep learning. Alexa, powered by more than 30 different ML systems, helps customers in billions of instances each week to manage smart homes, shop, get information, entertainment and more.

Today at Amazon Web Services, more than 100,000 customers have been using the cloud for ML across all sizes of organizations and industries. The life sciences industry, for example, has transformed profoundly over the past decade, and ML adoption has been at the core. In fact, nine out of 10 top pharmaceutical companies are using AWS cloud for AI/ML workloads; this includes Pfizer, Novartis and Johnson & Johnson. Due to computational drug discovery methods and ML capabilities powered by the AWS cloud, Moderna was able to produce their mRNA vaccine sequence in only two days, after the COVID-19 virus had been sequenced, and manufactured the first clinical vaccine batch in 25 days.

Virtually all start-ups are born and scale in the cloud, many are AI-first companies, aiming to became category leaders in their respective verticals. Global leaders in medical and health technology are modernizing their portfolios and building AI into their new solutions and services for mission critical workloads such as medical imaging, genomics, and electronic healthcare records. Over the past two years healthcare provides started their cloud journey many of them perusing an all in strategy. These organizations are laying the foundation to adopt AI technology at a rapid pace, and delivering it where it is needed most and can have the deepest impact — at the frontlines of patient care.

[1] Professor's perceptron paved the way for AI – 60 years too soon https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

[2] How to cite AdaBoost http://citebay.com/how-to-cite/adaboost

[3] ImageNet Classification with Deep Convolutional Neural Networks https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[4] Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm https://arxiv.org/abs/1712.01815

The new era of Generative AI

True technological transformation occurs when it dismantles barriers and transcends established conceptual boundaries, much like the case of the internet, which managed to seamlessly unite the physical and digital worlds.

The new era of Generative AI is about to break through the barrier of human cognitive performance. We've seen proof of this in the past for specific and narrow tasks. However, Generative AI is due to surpass the average human performance on virtually all cognitive tasks as well potentially acquiring the ability of reasoning - i.e. System II thinking [1].

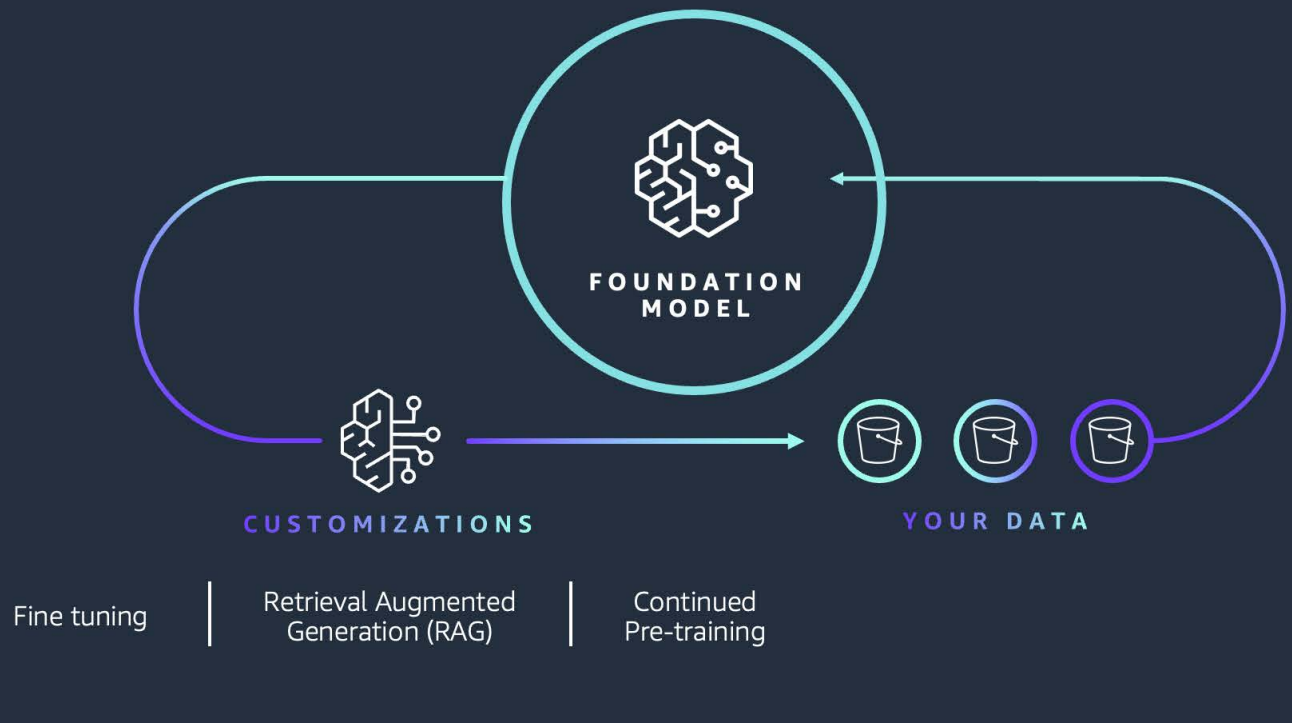
We are nearing that inflection point in AI proliferation, fueled by three key attributes:

- The ability to train on unlabeled data using self-supervised approaches, which reduces the data problem to mainly solving data access, and has already enabled us to train on the internet-scale datasets
- The ability to scale computation and model size, which leads to increasingly capable and high-performing models, theoretically without saturating or overfitting
- The ability to train foundation models as a base model and use transfer learning to fine-tune with marginal cost to specific tasks and data domains

Today, there are massive investments going into the life sciences industry and start-ups, leveraging Generative AI in the context of systems biology and drug discovery. One of the most promising fields is disease gene identification by analyzing large-scale genomic datasets and identifying patterns or correlations between genetic variants and disease phenotypes. In drug design and discovery, Generative AI can be used to design and synthesize new protein sequences, predicting their folded structures that enable them to carry out particular functions in the cell.

In healthcare, new biomarkers leverage multi-modal and multi-omics data to aid precision diagnostics, disease predication and precision therapy . On the other hand, applications in the areas of operational efficiency, patient-physician communication and workflow automation are likely to become reality in the short-term. This is expected to reduce clinician workload, increase patient satisfaction and help provide proactive healthcare to at-risk communities.

Sources:
[1] Thinking, Fast and Slow - Daniel Kahneman - https://www.amazon.de/-/en/Daniel-Kahneman/dp/0374533555



Industry

The core novel concept that emerged with the advent of the new generative AI era is that of Foundation Models (FM). Foundation models refer to large-scale, pre-trained machine learning models that serve as base model or starting point for a wide range of natural language processing (NLP) - referred to as Large Language Models (LLMs), and other artificial intelligence (AI) tasks - such as image generation from text prompts. These models are typically trained on vast and diverse datasets to learn the intricacies of language and patterns in data. Once trained, foundation models are fine-tuned for specific applications, for example as instruct models, allowing developers to leverage the general knowledge and capabilities learned during pre-training to accomplish specific tasks, such as Q&A or summarization.

From an industry perspective, FMs introduce a new layer in the value chain and can be seen as infrastructure. With that, a high-level model divides the industry into two broad segments: 1) builders of FMs— FM trainers and operators — and 2) builders of Apps — FM users and refiners.

Builders of FMs are in the business of training typically a family of FMs from the ground up for a particular domain. Such organizations operate infrastructure at scale, deploying and providing API level access and customization capabilities to third-party application builders in a particular field.

Builders of Applications build on top of existing FMs. They typically undergo a certain level of refinement and customization and will use one or more models as part of their end-user application.

To present the current state of Generative AI in Healthcare, we invited some of the leading innovators in the space to share their vision, strategy, solutions and current experience. On one hand we included organization position as leaders in the **Builders of FMs** category for the healthcare domain such as Clinical Language, Digital Pathology, Radiology, Genomics, and Multi-Modal Longitudinal and Cancer Models. On the other hand we aimed for contributions from leaders on the **Builder of Applications** side who are innovating in some of the most promising areas such as clinical documentation, imaging and genomics health services, workflow automation, search and reasoning, structured reporting.

Builders of Foundation Models

Foundation Models for Digital Pathology

Prof. Jeroen van der Laak
Chair of Computational Pathology, RadboudUMC

Prof. Jeroen van der Laak is Professor of Computational Pathology at RadboudUMC in the Netherlands and has 32 years of experience in computer science and AI. Nestled within the pathology department, he leads one of the world's leading research groups of 35 to 40 scientists focused on solving problems using AI that are useful to clinicians - "we only build things that clinicians find useful." Jeroen is also the founder and Chief Scientific Officer of Aiosyn, a digital pathology startup that develops AI solutions to help pathology laboratories and biopharmaceuticals improve their workflows and find the best possible treatment.

Unlike radiology, efficiency gains for pathologists are not as important as reducing variability and improving treatment decisions or patients. For this reason, one of the flagship projects funded by the Dutch Cancer Society is evaluating the use o AI to better risk stratiy breast cancer patients. This can potentially inorm precision therapy and determine which patients will benefit from chemotherapy and which will not and be spared the harmful side effects.

Jeroen believes Transormers could enable more robust models trained on less data to process multimodal inputs such as images and text rom structured reports. It also extends AI to rare diseases, which is currently severely limited by data availability. His group has begun experimenting with vision transformers and exploring building domain-specific baseline models using unique resources such as the BigPicture project, the largest digital archive o pathology data with more than 3 million images and 45 partner organizations.

At >2 GB per sample, the digital pathology data is very large and requires a cloud-based inrastructure to build and operate oundation models at scale. The field is progressing fast and wider adoption is anticipated as a consequence of convincing scientific results and access to pre-built foundation models. In the future, Prof. Jeroen van der Laak would like to evolve his research beyond pure digital pathology to multimodal diagnostics and incorporate omics, radiology and other patient data into models to make better diagnostic and treatment decisions and thus reduce overtreatment.

Foundation Models for Radiology

Dr. Khan Siddiqui, M.D., CEO and Jory Tremblay, CCO
HOPPR

Dr. Khan M. Siddiqui is the Co-Founder, CEO and Chairman of HOPPR, an AI medical imaging company seeking to transform global healthcare by democratizing access to advanced diagnostics. HOPPR aims to leverage generative AI and multimodal foundation models to extract insights from medical images and make them more understandable to healthcare workers at all skill levels. Their vision is to enable quality healthcare for the 90% of the world currently lacking access to trained radiology specialists that can interpret complex scan data. By building infrastructure to apply advanced AI capabilities to medical imaging and partnering to distribute solutions globally, HOPPR plans to bridge divides in healthcare equality. HOPPR represents the vanguard of using generative AI to connect imaging data with clinical care teams across geographic and socioeconomic barriers.

HOPPR announced the launch of a multi-modal foundation model in November 2023. The model is being trained on Amazon Web Services against 10mm high-resolution, medical images at full bit-depth, with expectation of training across 100mm by the end of 2024. Data sources do not just include images (e.g. X-ray, CT, MRI, and Ultrasound), but also associated medical reports. The model is designed to be queried using natural language and can be fine-tuned for specific tasks and applications, showcasing its versatility and potential impact on the field of radiology.

One application of Generative AI HOPPR will enable through this, is augmenting medical image interpretation to improve diagnostic accuracy. Their foundation model can be fine-tuned to identify anatomical structures and pathologies beyond human perceptual limitations, assisting clinicians in image analysis tasks such as distinguishing small liver lesions or finding indicators of disease at the earliest detectable stages.

HOPPR's generative AI approach allows personalized tuning to adapt models to new data types, modalities, and use cases faster than previously possible. HOPPR leverages partnerships with cloud and device partners to operationalize these capabilities through secure, distributed infrastructure. What's groundbreaking is the ability to uncover insights humans can't perceive and provide them in customized precision for clinical workflows. This has potential to transform early detection and interventions across all health systems.

Ultimately, what animates HOPPR is the vision of accelerating global healthcare equality through the power of generative AI to uncover all the potential hidden inside medical imaging data. They envision an interventional radiologist planning life-saving procedures more precisely before even seeing the patient by letting AI reveal predictive insights from multimodal inputs. Or a rural physician accurately diagnosing conditions never before detectable in their community thanks to AI augmentation from models tuned to sparse local data. HOPPR stands at the precipice of a new era in leverage AI scalability and accessibility to dismantle barriers limiting quality care. As it's Chief Commercial Officer, Jory Tremblay ensures that the company stays "laser focused on generating the economic sustainability and highly efficient distribution partnerships to execute on the vision". Generative models trained on diverse global imaging data can propagate expertise anywhere imaging reaches. As HOPPR partners propagate these models into critical workflows via platforms serving hospitals worldwide, they see an opportunity ahead for spreading healthcare equality by creating AI that levels knowledge rather than concentrating it.

Multi-modal foundation models

Prabhu Arumugam
Director of Clinical Data and Imaging, Genomics England, UK

Prabhu Arumugam is the Director of Clinical Data and Imaging at Genomics England and leads their genomics and multi-modal work. Genomics England was founded more than 10 years ago with the mission of bringing whole genome sequencing to the UK's National Health Service as routine care, a goal they have achieved for certain rare diseases and cancers. They successfully completed the groundbreaking 100,000 Genomes Project which advanced our understanding of cancer and rare disease treatments and diagnostics, and initiated Genomics England's Cancer 2.0 initiative leveraging AWS.

Genomics England has a pioneering vision to integrate generative AI into routine clinical genomics practice. Through their Machine Learning Chapter, Genomics England is building AI infrastructure and solutions across its sequencing, research and clinical activities. Key focus areas span from augmenting genomic analysts' workflows to building clinical decision support models that analyze a patient's multi-modal data — encompassing DNA, RNA, images and EHRs. Central to Genomics England's efforts is curating a world-leading research environment containing one of the richest genomic and multi-modal health data repositories globally. Through collaboration with the NHS, Genomics England is now combining that data with digital pathology and radiology images (like MRIs and CT scans) from more than 15,000 participants across 84 hospitals in England, representing 20 solid cancer subspecialties. This enables novel experiments to uncover patterns in molecular, imaging and electronic health data that are invisible to humans alone. By making this vast data resource accessible to external researchers in a secure, privacy-centric manner, Genomics England is establishing the foundations to lead advancements in generative AI for healthcare.

> **"I think the important bit is that we've had AWS as an organization to support us."**
>
> **Prabhu Arumugam**
> Genomics England

This enables novel investigations to uncover patterns in molecular, imaging and electronic health data that are invisible to humans alone. By making this vast data resource accessible to external researchers in a secure, privacy-centric manner, Genomics England is establishing the foundations to lead advancements in generative AI for healthcare.

Genomics England's goal is to pioneer models that optimize and individualize screening, diagnosis and treatment decisions by leveraging a fusion of multiple AI-derived data insights per patient. The promise will be unraveling disease complexity through AI at a depth, speed and scale unachievable via today's fragmented view of single data modalities and short-on-time clinical experts. Delivering this will involve addressing complex

challenges around model interpretability, transparency and building clinician trust - which Genomics England sees as central to successfully introducing AI-guided decision making into mainstream practice.

"Our question is: Do we really understand how tumors evolve, how they progress and why they metastasize? How do we see things that are linked from the genome to the DNA, to the pathology image, to the radiology scan?" asks Prabhu Arumugam.

Multi-Modal Longitudinal Models

Mahesh Pancholi
CIO, UK Biobank

Mahesh Pancholi is the Chief Information Officer for UK Biobank. Established in 2006, UK Biobank collects and maintains healthcare data from over 500,000 participants as part of a long-term prospective longitudinal study. Over the past 15 years, UK Biobank has followed the health of these volunteers, who provide their biological samples and health data. Data collected during an initial assessment and as part of repeat follow-ups contains various modalities from genetics, multi-modal imaging, to health records linkage, and all data is de-identified. This has created an unparalleled longitudinal dataset for health research, spanning billions of health data points. UK Biobank's mission is to enable vital health research that can lead to disease prevention and treatment breakthroughs. As a publicly-funded resource, they welcome applications from researchers in charity, government, industry and academic settings, approving those who are undertaking health-related research that is in the public interest to gain access to their uniquely powerful biomedical database. To date, their data has been used in over 10,000 peer-reviewed scientific papers advancing our understanding of health and disease.

UK Biobank finds itself at an inflection point as advances in artificial intelligence introduce new opportunities to maximize the impact of their platform. The application of generative AI specifically has the potential to make UK Biobank's complex multi-modal health data more accessible and usable for by researchers. By integrating across data types and time, generative models can could unlock insights and democratize access beyond specialized expertise. As a pioneer of health data platforms, UK Biobank is uniquely placed to work with the research community to enable the secure, ethical, accessible and democratized application and views itself as the ideal proving grounds for testing the integration of generative AI in healthcare and life sciences R&D. By responsibly applying generative models to their unparalleled data asset, UK Biobank researchers could further accelerate discoveries that can tangibly improve health outcomes worldwide.

"How do we make this huge multi-modal longitudinal data set accessible and really drive forward the age of personalized medicine?" asks Mahesh Pancholi, CIO of UK Biobank. For UK Biobank, key opportunity enticing future may stem from using generative AI models that have an understanding of the different data modalities to seamlessly integrate across specialization. This would allow researchers to assemble insights connecting their area of focus with different domain knowledge in a way not possible through human effort alone.

Structured Clinical Reports

Prof. Wieland Sommer
CEO, Smart Reporting

Prof. Dr. Wieland Sommer is the Founder and CEO of Smart Reporting, which started in 2014 as a university spin-off to automate and streamline workflows surrounding medical documentation. Today, it has an interdisciplinary team of 70 doctors, data scientists and software engineers who develop software from Germany, Austria, Switzerland and Brazil based on their detailed understanding of clinical workflows. The company's multilingual software for structured diagnostics in radiology and pathology is used by more than 15 000 physicians in over 90 countries.

Radiology professionals often face two key challenges: variable and incomplete results from manually drafted reports, and non-machine-readable, unstructured findings that cannot be used for proactive analysis and improved treatment planning. The SmartReports solution has been developed to address this problem by providing an adaptive reporting editor offering voice-powered input, structured templates, seamless EMR connectivity, and embedded clinical guidelines. SmartReports is uniquely positioned to have its capabilities amplified by the integration of generative AI.

In November 2023, Smart Reporting announced they will integrate large language model (LLM) technology to deliver impressions automatically suggested as part of their reporting software "SmartReports". The use of natural language generation models to automatically suggest impressions at the point of care will provide invaluable assistance in clinicians' workflows. They aim to deliver significant time savings to alleviate the documentation burden on physicians. Further, by structuring machine-readable outputs, Smart Reporting intends to unlock data within reports to power analytics for improved treatment planning, patient analysis, and care coordination. By building on Amazon Bedrock, they are able to leverage state-of-the-art AI model to enhance their radiology reporting solution.

"We are always looking for ways to make radiology reporting faster and more efficient," said Prof. Wieland Sommer, "We have been open to technology and vendors and are working agnostically to broadly integrate LLMs into our software"

Smart Reporting views enhancing the efficiency and quality of medical reporting as a high value opportunity for applying generative AI. Specifically, they aim to tackle the dual challenges of reporting inconsistency and the inability to leverage unstructured free-text data that collectively lead to delays and suboptimal clinical decision-making. With generative AI, Smart Reporting believes they can optimize the diagnostic process through faster, more consistent, and data-rich medical reports to drive better patient outcomes.

Clinical Documentation

Jonathan Larbey
CEO, T-Pro

Jonathan Larbey is the founder and CEO of T-Pro, a global leader in clinical document improvement software. T-Pro was started to reduce the burden on clinical professionals in documentation and administration. Their cloud-based technology and services platform focuses on automating these workloads by integrating AI such as speech recognition and voice interfaces. Over time, the company has worked to train their own custom models and deliver tailored experiences at scale. The company now serves over 1,100 healthcare organizations in English-speaking countries such as the United Kingdom, Ireland, Australia, New Zealand and Southeast Asia. The core business model is a SaaS platform.

From an AI veteran's perspective, Jonathan Larbey views generative AI as an incremental technology that has lowered barriers to entry and enabled new competitors. But he believes healthcare companies need to do more than just be a shell for a generative AI model. For T-pro, generative AI is expected to impact near-term cost reduction through the consolidation of pipeline components such as data cleansing, while the most valuable applications are expected to be in voice assistants and ambient documentation.

The company relies primarily on open-source models like the LLaMA family, which they fine-tune and optimize. Much of the value comes from building pipelines and systems around the core models to refine and control Generative AI outputs. This helps to solve problems related to homophones, words that have the same pronunciation but different meanings, as well as reducing hallucination to the minimal level required for precise and complex medical documents. Generative models are also used for data augmentation and cleansing within T-Pro's ML pipelines. Jonathan sees an opportunity "in the multimodality of generative AI models, in which, in addition to text and speech, non-verbal cues could also be recorded and processed via image and video recognition."

Jonathan's vision for the future includes "infinitely customizable products" that can be deployed instantaneously with no implementation effort. However, given the inherent heterogeneity of enterprise software in healthcare, this would be difficult to achieve. An important consideration for AI-first companies like T-Pro, which operate in a highly regulated industry, is the evolution of the regulatory environment in light of emerging technologies like GenAI. T-Pro is already preventing potential risks and reducing them by integrating protective measures into products, legal contracts and liability exclusions.

Clinicial Workflows

Dr. Ruben Amarasingham
CEO, Pieces Technologies, Inc

Ruben Amarasingham is a physician, founder and CEO of Pieces Technologies, Inc, a clinical AI company founded in 2016 with a mission to improve patient care through data-driven technologies. Pieces develops processing solutions that generate patient summaries and clinical notes to enhance clinician efficiency. They began experimenting with using machine learning models in 2018 to produce free-text summaries from complex predictive models, helping doctors extract insights more easily. Today they autonomously generated over two million patient summaries that integrate seamlessly into clinician workflows and save substantial time per patient.

The main problem Pieces aims to solve with AI is information overload for clinicians. Their solution leverages large language models to summarize complex patient situations into plain English paragraphs that update regularly as new data arrives. This helps nurses and physicians efficiently understand what is happening with their patients as they improve the handover process, allow for more time to be spend per patient and ultimately reduces burnout. Through constant optimization and a rigorous human-in-the-loop review process, Pieces reaches low summary error rates. Their AI agents allow personalization down to the individual doctor level.

On the technology side, they have found different models work better for different tasks and use them in concert to detect errors. They constantly evaluate models on cost, quality and latency. AWS technology, such as Amazon Bedrock allows them to easily swap models and optimize cost-benefit. Key foundational technologies for Pieces today include transformer models, adversarial training, cloud infrastructure, and specialized clinical datasets and benchmarks.

Dr. Amarasingham is excited by the possibilities of customizable “AI agents” that self-improve through clinician feedback. He ultimately envisions AI not just supporting doctors with mundane tasks but collaborating deeply embedded into the clinicians workflows as a trusted partner. However, he stresses that the full potential hinges on reaching sufficient accuracy at minimal compute costs, overcoming regulation uncertainties, and earning clinician trust through responsible practices.

Genomics

Mattia Capulli, CSO and Andrea Riposati, CEO
DanteLabs

Mattia Capulli is Chief Scientific Officer and co-founder of Dante Genomics. With a PhD in Biotechnology, he concentrated his postdoctoral research on genetic rare disease. In 2016, together with Andrea Riposati, Mattia co-funded Dante Genomics, a leading global genomics and precision medicine company with the mission to change the lives of people through whole genome sequencing technology. They serve individual patients as well as B2B customers across the globe. Today, Dantelab’s platform includes over 60 software applications and provides direct to consumer tests, clinical and research genomics services, as well as software and data services to genomics laboratories. Only 0.1% of the population had their genome sequenced, and out of 400 million of rare disease patients around the world, less than 0.5% have their whole genome sequenced. Dante Genomics provides services from diagnostics to therapeutics. This includes the collection of the samples, to sequencing and analysis of the whole genome.

With the shift from hardware to software in the genomics industry, we see different areas where AI can be applied. The easiest is in the lab operations and not in data analysis. Labs operations are still academic, with opportunities for AI to improve operations through automation and industrialization of the lab process. It can be used to improve the volumes that you load for each patient for optimizing the sequencing output. This will in turn reduce cost and increase accessibility to the journey to sequence the whole genome of the entire population up to 10B. The second area is around data analysis. AI can help the medical professional to match the phenotype to the genotype problem. As the data volume is too big for a human to analyze, AI can not only help improve the speed, but also remove any biases. Today, everything is on the human, and it can take from 1 day per genome all the way to 10 days to complete the analysis. This is not scalable, and in order to offer population level coverage to newborns, oncology, and other clinical services, AI needs to support geneticists in making the process more efficient.

In particular the association of a genetic variant to a symptom is a manual step today. It takes a lot of time and it is error prone. More over human biases and over simplification of the data in healthcare can lead to miss diagnosis and further errors in patient management. Dante Genomics is investing in GenAI to help build high accuracy models. The potential for AI is in automating parts of this process and enable clinicians to leverage the full breadth of genomics and multi-omics data. “We need to use all the data and do not over simplify”. Regulated diagnostic AI tools genomis are lacking today. Regulatory requirements are difficult to navigate and expected to make it even more difficult and expensive to bring AI to market in diagnostics. Explainable AI is needed to meet regulatory requirement

Integration of data outputs from genomics analysis pipelines with AI and analytics services is key enabling technology. In particular, query language needs to be optimized. Those have been designed for natural language and 21-letter alphabet. However, the genomics alphabet has 4-letter. We have built our own genomics query language - the variant query language - to increase performance and reduce computation resources. DanteGenomics are currently working on their first prototypes leveraging Bedrock. Dante Genomics was part of the private beta program of Bedrock. We are running tests on using open source LLMs on diagnostic use cases and benchmark with various commercial models.

The challenges of resourcing and funding AI initiatives has to be balanced with commercial viability. Grants and partnerships in supplementing internal funding are key to success. Early experiments with open-source AI models for genomic applications have been conducted. They take a customer obsessed approach and look into applying GenAI to problems that have impact on their customers. Giving the current market conditions, the market / financing condition is not great for biotech and diagnostics. Some of the R&D is financed by European and Italian grants.

In the future, millions of people will have access to get their genome sequenced every day. With the help of GenAI, our full-service platform will support the analysis of the data, help get insights from it and produce clinical validated reports, at scale. The vision for Dante Genomics is to become the largest genetic testing company and improve lives globally.

Computer Vision and Multi-Modal GenAI Workflows

Alberto Rizzoli
V7

Alberto Rizzoli is an Italian-born serial entrepreneur, co-founder and CEO of V7 with a mission of turning data into software, with a focus on AI and generative AI. V7's product Darwin is used by humans in the loop to iteratively teach and improve the accuracy of AI. With healthcare and life sciences being the largest industry served to date, the V7 solution covers all medical imaging and surgical video modalities. Together with his co-founder Simon Edwardsson, Alberto feels a personal motivation to "accelerate the progress of AI in the healthcare industry to enable us to live longer, happier lives." The V7 SaaS platform has already been adopted by top names in the industry, including GE, Philips, Merck and AI startups that are bringing models to FDA approval.

Even though generative AI in healthcare is still in its infancy, the paradigm shift was, in Alberto's view, a "tsunami." As a company that relies on AI as a key growth vector, V7 has undergone a complete retooling from the ground up over the last 12 months to become a generative AI-native company. The very definition of data morphed from using annotated images to train narrow models (e.g. organ segmentation) a year ago to today's concept of multimodal and patient-centric data that can be used to model dynamic interactions between multiple personas over time. Alberto foresees the future of human-AI collaboration and believes that autonomous agents are overrated these days: "They are like incredibly intelligent children, you can work with them on a problem, but if you let them lose, they will stick their fingers in the power sockets"

V7 investments aim to make human-in-the-loop workflows and tools more efficient, with the ultimate goal of owning the data infrastructure layer, including: data curation, data labeling and records management. Most recently, V7 expanded capabilities to support multimodal datasets and advanced visualization, and is experimenting with the use of discrete and categorical LLM outputs, as well as expansion into clinical workflows. V7 job is to customize, adapt, and tune foundation models to specific domains and tasks - bringing foundation models from "undergraduate students to the knowledge level of university professors". The company is experimenting with all types of commercial and open source large-scale models for text, audio, video and multimodal data with the strategy of maximizing accuracy and reliability until use cases are validated before focusing on cost and governance optimization.

Alberto wants V7 to become the leading company capable of transforming data into trustworthy AI in the next five years. All this with a focus on providing the most human-friendly tools to teach AI and build trust on an emotional level. He believes V7 plays a key role in enabling people to contribute to a unified knowledge that extends across organizational boundaries. This can, for example, accelerate scientific discovery in areas such as cancer by bringing together research groups around the world, ultimately enabling a world in which both models and datasets are democratized and made openly available.

Technology

The new possibilities generative AI brings to healthcare organizations will not only unlock a new area of research in biology and speed up innovation in medicine but they will also increase efficiencies and help improve outcomes for patients. To achieve this, easy, secure, and democratized access to generative AI services, data, models, and infrastructure is required. Only then, the growing number of healthcare challenges can be addressed and overcome. In the following sections we will look at how AWS's building blocks can support healthcare organizations to leverage the possibilities of generative AI, with enterprise grade security, and facilitating data privacy, within a regulated industry.

These building blocks span three layers of the generative AI stack. The bottom layer is the infrastructure required to train and run LLMs and other FMs. The middle layer provides customers with easy access to these models and tools that enable them to build and scale generative AI applications. At the top layer are game-changing applications like assisted coding, chatbots, generating clinical notes and generative BI. In the following sections we look at each of the three layers and cover architectural patterns for the use of generative AI in healthcare applications. [1]

[1] Welcome to a New Era of Building in the Cloud with Generative AI on AWS https://aws.amazon.com/blogs/machine-learning/welcome-to-a-new-era-of-building-in-the-cloud-with-generative-ai-on-aws/

Infrastructure for Training and Inference

The powerful disruptive opportunities generative AI can bring to healthcare organizations is apparent. The challenge is how to train, fine-tune and run the underlying models in a quick, efficient and sustainable way.

To push the boundaries in terms of increased performance and reduced costs for demanding workloads like ML training and inference, AWS has been investing in their own silicon over the past 5 years. The result are AWS Trainium and AWS Inferentia chips. These offer the lowest cost for training models and conducting inference in the cloud.

AWS Inferentia is a purpose-built inference chip, optimized for high-throughput, low-latency workloads. It is ideal for applications that require real-time response. The second generation running in Amazon EC2 Inf2 instances, is optimized for generative AI workloads for models with hundreds of billions of parameters. AWS is the chosen platform for top AI firms such as AI21 Labs, Anthropic, Cohere, Grammarly, Hugging Face, Runway, and Stability AI because it optimizes performance and reduces costs through ML optimized silicon selection.

AWS Trainium offers high performance and cost-efficiency in terms of model training. Trn1 [1] instances powered by Trainium can save up to 50% on training expenses over other comparable EC2 instances. AWS Trainium 2, will deliver even better prices and higher speed for models with hundreds of billions to trillions of parameters.

Customers will be able to train a 300 billion parameter LLM in weeks versus months. This enables premier FM startups like AI21 Labs, Stability AI, and Hugging Face to maximise performance while controlling costs.

Next to having developed custom silicon, AWS has a strong partnership with NVIDIA to continue to provide the best-performing GPU-based infrastructure for generative AI needs. They were the first cloud provider to make most recent NVIDIA H100 Tensor Core GPUs available in their EC2 P5 Instances. [2]

To unlock the performance and cost-efficiency of custom ML silicon, **AWS Neuron**, an SDK with compiler, runtime and profiler can be used. It supports many popular models, LLama 2 from Meta, MPT from Databricks and Stable Diffusion from Stability AI, as well as 93 out of top 100 Hugging Face models. The integration of custom hardware and infrastructure with SDKs and AI services provides a seamless experience. It ensures customers can deploy their models without needing hardware expertise. It removes the undifferentiated heavy lifting and democratizes access to advance AI capabilities. In addition, it also offers a path to more sustainable AI practices as they minimize the carbon footprint associated with training and running AI models.

This combination of the best ML chips, super-fast networking, virtualization and hyper-scale clusters made generative AI startups like Hugging Face choose to build on AWS.

[1] Amazon EC2 Trn1 Instanceshttps://aws.amazon.com/ec2/instance-types/trn1/

[2] New – Amazon EC2 P5 Instances Powered by NVIDIA H100 Tensor Core GPUs for Accelerating Generative AI and HPC Applications https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/

> **“Accessibility and transparency are the keys to sharing progress and creating tools to use these new capabilities wisely and responsibly. Amazon SageMaker and AWS-designed chips will enable our team and the larger machine learning community to convert the latest research into openly reproducible models that anyone can build on.”**
>
> **Clement Delangue**
> CEO of Hugging Face

It can be a challenge to leverage the infrastructure and build, train and run LLMs and other FMs in an efficient and cost-effective way, as vast amounts of data need to be made available, cleaned and transformed. Clusters of GPUs have to be managed and code distributed for training the models across clusters.

To address these challenges, **Amazon SageMaker** was launched 6 years ago. It is a fully managed service that enables data scientists to build, train and deploy machine learning models. It includes more than 380 capabilities that cover automatic model tuning, training, tools for ML OPs, data preparation, notebooks, human-in-the-loop workflows and features for responsible AI. It includes **SageMaker Studio**, an IDE that contains purpose-built tools to perform ML developments steps. Harnessing the potential of generative AI, especially in the context of healthcare organizations, relies on the process of LLM evaluation. Evaluation is used to measure the quality and responsibility of the output of a generative AI service/model.

Amazon SageMaker Clarify provides developers visibility into their training data and models to limit bias and increase explainability. [1]

Amazon SageMaker Ground Truth, a fully managed data labelling service helps customers to build highly accurate training datasets. This enables customers to apply human feedback across the ML lifecycle and improve FMs with human-in-the loop capabilities. **Amazon Augmented AI** enables human-in-the-loop for workflows where human review is required. **Amazon SageMaker Pipelines** is the first purpose-built, continuous integration and continuous deliver (CI/CD) service for ML. It enables customers to create, automate and manage their end-to-end machine learning workflows at scale. Amazon Sagemaker and Generative AI services are integrated into the AWS ecosystem. This allows frictionless use of other AWS services like storage, compute or data processing. [3]

In the next section we will look at how tools to build LLMs and other FFMs can be used to build applications harnessing generative AI capabilities and connect it to the underlying data in healthcare organizations.

Sources:

[1] Operationalize LLM Evaluation at Scale using Amazon SageMaker Clarify and MLOps services https://aws.amazon.com/blogs/machine-learning/operationalize-llm-evaluation-at-scale-using-amazon-sagemaker-clarify-and-mlops-services

[2] Amazon Augmented AI https://aws.amazon.com/augmented-ai/

[3] Amazon SageMaker Features https://docs.aws.amazon.com/sagemaker/latest/dg/whatis-features.html

Operationalize LLM Evaluation at Scale using Amazon SageMaker Clarify...

Machine Learning Workflow - Amazon Augmented AI - AWS

New – Amazon EC2 P5 Instances Powered by NVIDIA H100 Tensor Core GPUs...

Amazon EC2 Trn1 instances – Compute – Amazon Web Services

Amazon SageMaker Features - Amazon SageMaker

Tools to build with LLMs and other FMs

Not all customers are willing to spend the resources to build their own LLMs or other FMs. For these customers, the middle layer of the stack offers the solution. It provides access to these models as a service.

With introducing **Amazon Bedrock** in late September 2023 [1], customers have access to a fully managed service to build and scale generative AI applications leveraging FMs. It supports Llama 2, Meta's next generation open-source large language model (LLM), which joins the ranks of current model providers AI21, Labs, Anthropic, Cohere, Stability AI, and Amazon. It provides fully managed Llama 2 [2] secure fine-tuned 13B and 70B parameter variants, and interfaces with **AWS CloudTrail** and **Amazon CloudWatch** for metrics and monitoring. Amazon Bedrock is HIPAA eligible and GDPR compliant. It allows the seamless integration into applications, the foundation models in Amazon Bedrock can be accessed via API. It includes the option to privately fine-tune these models as it doesn't use fine tuning data for training base foundation models [3]. Using the **SageMaker JumpStart**, these models can be deployed with one click for inference or to fine-tune them with custom data.

Generative AI is only as useful as the data it relies on. The challenge, healthcare organizations face in relation to their data, is that the data is multi-modal, diverse, often hidden away in silos, unstructured and hard to make actionable. Data sources range from clinical audio, electronic health records, omics data, medical imaging, digital pathology and other third-party data. The AWS Health Data Portfolio offers purpose-built services address these challenges.

Generative AI can be applied at the source of the data, like real-time data streaming and analytics, to ETL workloads and seamless integrate with the AWS Health services. From storing medical records in FHIR format at scale with **AWS HealthLake**, to **AWS HealthOmics** that transforms genomic and other omic data into insights, to **AWS HealthImaging** which support healthcare organizations to store, transform and analyze their medical images at petabyte scale. [4]

Connecting LLMS and other FMs and their powerful capabilities with that underlying data is crucial to extract the full benefit for healthcare organizations. **Agents for Bedrock** [5] can be built to support with this challenge. They can execute multistep tasks accessing enterprise systems and data sources. This enables the automation of tasks that require access to enterprise systems like research databases, clinical trial systems and databases. **AWS Step Functions** enable the visual development, inspection and auditing of workflows and integrated with Bedrock, allow the orchestration of tasks to build generative AI applications as well as integrate with 220 AWS services. [6]

Guardrails for Amazon Bedrock enables healthcare organizations to ensure the interaction with users stays safe. It can remove PII and PHI from generated summaries and bring a consistent level of protection across their generative AI applications. Amazon Bedrock is also integrated with Amazon CloudWatch to support the tracking of usage metrics and invocation logging. Through this integration Amazon Bedrock can be monitored near real-time. [7]

Sources:

[1] Amazon Bedrock is now generally available https://aws.amazon.com/about-aws/whats-new/2023/09/amazon-bedrock-generally-available/

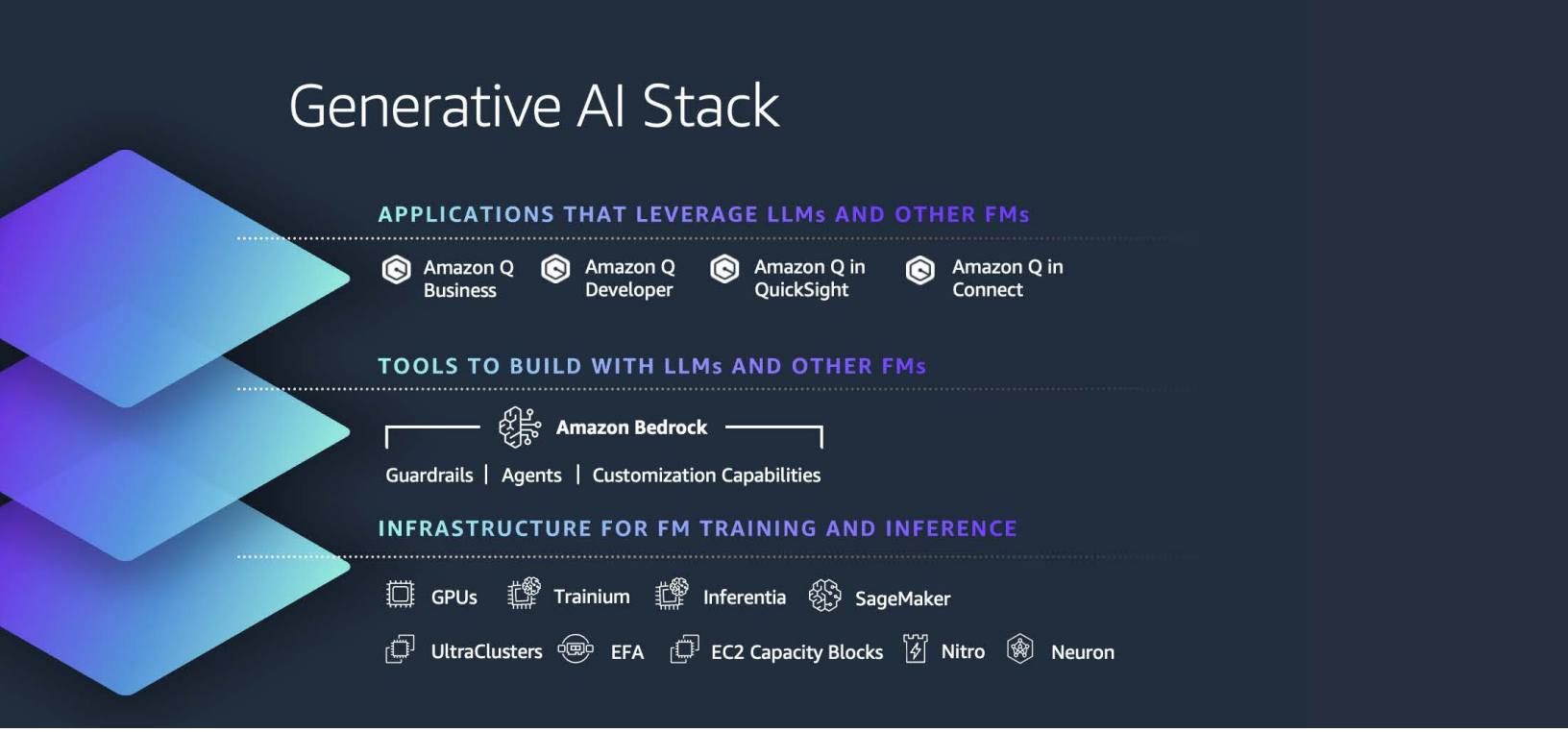
[2] Meta Llama 2 on Amazon Bedrock https://aws.amazon.com/bedrock/llama-2/

[3] What is Amazon Bedrock? https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html

[4] AWS Health Data Portfolio https://aws.amazon.com/health/solutions/health-data-portfolio/

[5] Agents for Amazon Bedrock https://aws.amazon.com/bedrock/agents/

[6] Build generative AI apps using AWS Step Functions and Amazon Bedrock https://aws.amazon.com/blogs/aws/build-generative-ai-apps-using-aws-step-functions-and-amazon-bedrock/



Applications to leverage LLMs and other FMs

With the speed o development in the area o generative AI, a challenge that remains is to keep up to speed with technology and be able to access the latest knowledge when needed.

This leads us to applications that are on the top o stack. These help healthcare organizations to make generative AI accessible to more users, internal and external. It includes services or developer, data scientists and architects, that help with creating applications aster and more secure, business intelligence tools that help analysts gain insights aster, as well as healthcare specific AI services that are powered by Amazon Bedrock and transform patient-clinician experience.

With tools like **Amazon Q Developer**, natural language can be used to generate code without the need to know the latest frameworks or APIs. Developers, data engineers and scientists using natural language to generate queries, realize the full benefit of generative AI. [1] To achieve this, the models must be able to identify the correct data sources and generate effective SQL queries that run at scale.

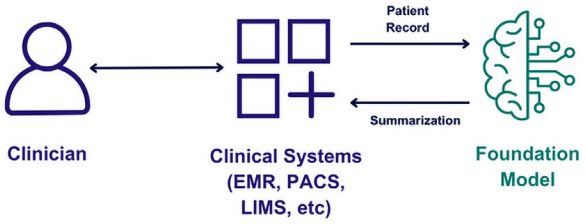
A modern data architecture that combines existing sources and services, then applies artificial intelligence, enables this. With **Amazon Q** developers are supported when debugging and testing code, architects when building the solutions and testers when cloud infrastructure needs to be troubleshooted and issues diagnose. Amazon Q, designed to understand data, code and operations, includes enterprise controls and keeps the data secure inside the healthcare organization. It can connect to private datasets or enterprise software and perform tasks like analyze observations rom clinical trials, create medical record summaries and support with compliance documentation. Business analysts can use Amazon Q in **QuickSight**, a Business Intelligence tool, to generate data stories and answer questions on the data. **AWS HealthScribe**, a HIPAA-eligible service, can be used by software vendors who are looking at building applications to be used in a clinical context. By leveraging Amazon Bedrock, generative AI is combined with speech recognition to support the creation of preliminary clinical documentation. [2] [3] [4]

Sources:
[1] Reinventing the data experience: Use generative AI and modern data architecture to unlock insights https://aws.amazon.com/blogs/machine-learning/reinventing-the-data-experience-use-generative-ai-and-modern-data-architecture-to-unlock-insights/
[2] Top 10 Reinvent 2023 Announcments Important For Healthcare and Life Sciences https://aws.amazon.com/blogs/industries/top-10-reinvent-2023-announcements-important-for-healthcare-and-life-sciences/
[3] Reinventing the data experience: Use generative AI and modern data architecture to unlock insights https://aws.amazon.com/blogs/machine-learning/reinventing-the-data-experience-use-generative-ai-and-modern-data-architecture-to-unlock-insights/
[4] Welcome to a New Era of Building in the Cloud with Generative AI on AWS https://aws.amazon.com/blogs/machine-learning/welcome-to-a-new-era-of-building-in-the-cloud-with-generative-ai-on-aws/

Architectural Patterns for Healthcare

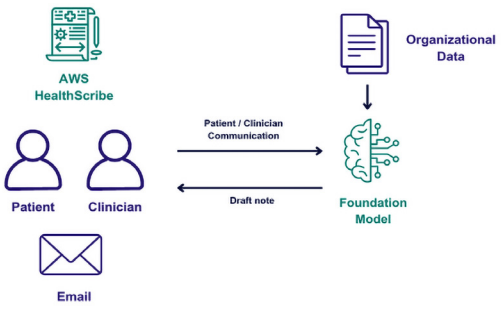
Healthcare organizations are exploring applications for generative AI in the following area:

Summarization



Foundation models have shown promise in quickly summarizing large amounts of medical information to highlight the most salient details for clinicians. Healthcare organizations are exploring how these AI systems might condense lengthy patient records into focused summaries that make clinicians better informed to deliver quality care. With the realities of modern medicine, providers often lack adequate time to thoroughly review a patient's full history prior to treatment. This is especially true in emergency scenarios - clinicians report having just minutes to familiarize themselves with a patient's background before needing to administer urgent, lifesaving care. Thoughtfully-designed foundation model algorithms can generate customized summaries of patient data for clinicians, surfacing the information deemed most relevant to that individual's immediate, pressing needs. The goal is twofold: first, reduce the time needed to prepare to treat a patient; and second, enhance the provider's contextual perspective on the case. If successfully implemented, such systems could get physicians up to speed more quickly and supply them with optimized synopses to support sound, well-informed clinical decision making. The most natural interface for these summaries is through the systems that clinicians already interact with, namely, the electronic medical record.

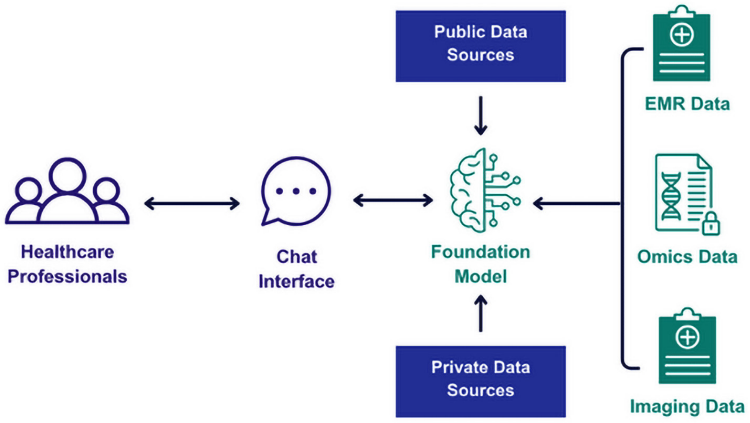
Note drafting



Clinical documentation and communication with patients are vital aspects of care, yet also time-intensive. Clinicians spend considerable portions of their day writing notes, letters, and other textual artifacts - a leading factor cited in burnout and reduced patient capacity. Foundation models present an opportunity to provide relief. By ingesting audio recordings, transcripts, and texts of clinician-patient interactions, AI systems can synthesize draft clinical notes, saving providers hours of writing time. These algorithms can reference policies, addresses, staff contact information, and other tedious details for inclusion, freeing clinicians from repetitive lookups. Importantly, foundation models would only suggest an initial draft for provider editing and approval prior to finalization. This allows physicians meaningful oversight to ensure all output meets clinical, ethical, legal and organizational standards before being applied. Clinicians maintain responsibility for ensuring accuracy, appropriateness, and

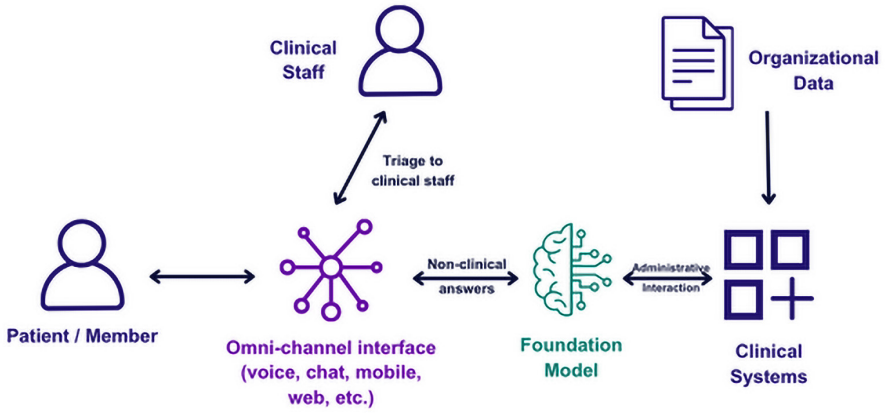
utput meets clinical, ethical, legal, and organizational standards before being applied. Clinicians maintain responsibility for ensuring accuracy, appropriateness and comprehensiveness. Thoughtfully implemented, such tools could significantly ease the documentation burden for time-pressed frontline staff. This would allow doctors and nurses to better focus on direct patient care and whole-person needs. However, developers must prioritize accuracy, accountability, and transparency to establish clinician trust in quality of syntheses. If done well, foundation-generated drafts may meaningfully boost productivity and job satisfaction.

Data exploration



Clinicians and researchers often have complex questions about patient populations that require deep dives into multifaceted health data to answer accurately. "How prevalent is this condition?", "What clinical trials are available for people with this genetic mutation?", "What were the outcomes for prior similar patients?". Addressing these inquiries can mandate skilled analysis by bioinformaticians across medical records, genomic data, images, and other modalities. However, the iterative nature of exploration - each answer potentially triggering additional questions or query refining - is enormously time-intensive. With limited availability of experts, many lines of insightful questioning go unpursued due to bandwidth constraints. Emerging foundation models that index diverse unified health data assets show promise in initially illuminating answers to complicated clinical and research questions without requiring upfront bioinformatician consultation. Such tools could empower healthcare professionals to freely iterate through inquiries, gaining directional insights and sharpening their thinking. Once high-potential threads are identified, skilled analytics teams could then finalize and validate the outputs. By drastically expanding clinicians' starting capability to extract value from data, such approaches may increase the volume of impactful health questions that can feasibly be asked and answered. This could ultimately catalyze greater understanding and improved patient outcomes through amplified learning. However, developers must prioritize accuracy, transparency, and collaborate closely with medical experts when building and deploying these sensitive technologies.

Non-clinical patient or member conversational interfaces



Patients and healthcare members rightfully expect readily-available access to the information necessary to navigate their care journeys. While the gold standard is personalized concierge-like interactions with organizational staff at any time, the scale of modern patient populations makes this impractical. Yet, the common alternatives - combing intricate websites or robotic phone trees - often prove frustrating barriers. This is where thoughtfully-implemented foundation models hold game-changing potential. Powerful natural language processing could enable human-like virtual assistants to answer many common inquiries on demand - office locations and hours, personal lab results and treatment protocols, scheduling updates and more. These algorithms can also intelligently recognize when queries require referral to trained clinicians, such as developing chest pain. By providing accurate, personalized, on-demand support for routine questions around the clock, these technologies could tremendously simplify access and enhance patient/member experiences. Additionally, by resolving many frequent low-complexity informational needs, foundation model systems may meaningfully reduce the volume of calls and messages that demand precious staff time.

Research

Open Source and Open Science

We are in the early days of generative AI with some of the most popular generative AI models being closed models - only accessible via APIs. Nevertheless, open source models are key when it comes to 1/trust , 2/equity , 3/ sustainability.

There many unsolved problems including hallucinations, social biases, and explainability. The scientific community is a trusted source of truth predestinated to provide answers to those problems. For this they need open access to models, underlying datasets, as well as algorithmic and parameters details.

We have seen open source models like stable diffusion for text to image generation with impressive community adaption. Meta has release the Llama family and sparked great interested in open LLMs. Thanks to the Technology Innovation Institute in Abu Dabi and the release of a new family of LLMs called Falcon under an Apache 2.0 license, open source models have ever since entered the realm of commercial application. With the launch of Amazon Bedrock, state of the art open LLMs model like Llama 2, can be consumed through APIs as managed services, in the same way as leading commercial LLMs models.

The open communities play an central role in keep the commercial players honest via open platforms such as huggingface.co, grand-challenge.org, kaggle.com, chat.lmsys.org. On one hand by facilitating open access to open source models through model hubs as well as through tools and playground to experiment and build. On another hand by increasing transparency through leaderboards and benchmarks across commercial and open models.

Future Research

Performance and cost optimization of large models for both training and inference is an active area of research. Both software and hardware methods are being developed to increase the learning performance, reduce resource requirements, and accelerate inference performance. For instance, 4-bit quantization methods can significantly reduce the size and memory requirement of a LLM. Speculative decoding speeds up inference by 2-3x. Flash attention, which reduce attention linear instead of quadratic scaling, allows for training with larger context.

LLMs are learning to leverage a variety of software tools ranging from simple calculators, to writing and executing python scripts to solve sub-tasks. With this LLMs are becoming agents, potentially moving from simple next word predictors - System I as defined by Daniel Kahneman - to deep reasoning machines, capable of complex System II thinking.

Beyond language models, transformer models in other domains are emerging. Segment Anything has learned a general notion of what objects are in images and is able of zero-shot segmentation of unfamiliar objects [1].

The agent paradigm applies for vision models as well with general-purpose vision systems via compositional multi step reasoning instead of end-to-end multitask training - VisProg and ViperGPT. With models like CLIP (Contrastive Language-Image Pre-Training) text and image that can be can be combined to form vision-and-language that are able to predict text from images using zero-shot approaches [2].

Healthcare and life sciences is one of the most existing industry for GenAI with breakthrough domain specific models being developed. Models such as ESMFold by META trained on proteins, can now be directly predict structure from amino acid sequences without relying on costly and slow multiple sequence alignment. Grace [3], Virchow, MED42 MedPaLM [4], are domain specific models for multi-modal medical imaging, digital pathology, and clinical language, respectively. Vision-and-language models are also an active area of development with models such as Pathology Language and Image Pre-Training (PLIP) [5], aiming at product end-to-end clinical reporting from digital pathology images. Models like Med-PaLM M offer a proof of concepts for generalists biomedical AI that can encode and interpret multimodal data including: clinical langue, imaging, and genomics. Graph-enhanced gene activation and repression simulator (GEARS) [6] combines prior experimental knowledge to predict the gene expression outcome given unperturbed gene expression and the applied perturbation.

Sources

[1] https://segment-anything.com/

[2] CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs. https://github.com/openai/CLIP

[3] https://www.prnewswire.com/news-releases/hoppr-launches-groundbreaking-foundation-model-for-medical-imaging-301996244.html

[4] Med-PaLM M - https://arxiv.org/abs/2307.14334

[5] Pathology Language and Image Pre-Training (PLIP) is a vision-and-language foundation model created by fine-tuning CLIP on pathology images. https://paperswithcode.com/method/plip

[6] https://www.nature.com/articles/s41587-023-01905-6

Responsible AI

Generative AI technologies and its application bring new challenges that need to be addressed. At Amazon, we believe the design, development, and deployment of AI/ML must respect the rule of law, human rights, and values of equity, privacy, and fairness. We're not developing technology for technology's sake — every move we make in this vein is to improve the experience for our customers and our partners, with the conviction that with success and scale brings broad responsibility.

AWS is committed to the responsible development of AI. This covers fair and accurate AI/ML services, as well as providing tools to ensure that organizations can leverage AI/ML services in a responsible way. AWS sees the core dimensions of responsible AI covering fairness, explainability, privacy and security, robustness, governance and transparency. This commitment aligns with the approach of AWS to develop generative AI services, included foundation models (FMs) with responsible AI in mind. During design, development, deployment and operations, accuracy is assessed and a range of factors are considered such as: intellectual property and copyright considerations, appropriate usage, toxicity as well as privacy.

Amazon is particularly focused on operationalizing responsible AI, helping our internal teams and our business customers move from responsible AI theory to practice. At Amazon, we are committed to developing artificial intelligence in a responsible way with a team of dedicated responsible AI experts, complemented by Amazon engineering and development teams that continually test and audit our products for fairness and accuracy.

Even as regulatory frameworks for AI are still being developed, Amazon is creating new resources to deliver transparency that customers want. **AWS AI Service Cards** are a form of responsible AI documentation that provide information on the intended use cases and limitations, responsible AI design choices, and deployment and performance optimization best practices for our AI services. They are part of a comprehensive development process we undertake to build our services in a responsible way that addresses fairness, robustness, explainability, governance, privacy, and security in mind. [1]

Amazon AI Service Cards provide a single place to find information on the intended use cases, responsible AI design choices, best practices, and performance for a set of AI service use cases. Tools like **Amazon SageMaker Clarify** can be used to mitigate bias and improve explainability, by detecting potential bias while the data is prepared, after training and in the model once it is deployed.

"It was an honor to represent Amazon for conversations with world leaders about the future of #AI and how we can innovate while fostering the safe, responsible, and secure development of this technology. As the benefits and capabilities of AI grow, so too does the importance of maintaining trust." Adam Selipsky, CEO of AWS, AI Safety Summit in UK, 2023 [2].

AWS supports governments' efforts to put in place effective risk-based regulatory frameworks and guardrails for AI that protect civil rights, while also allowing for continued innovation and practical application. As one of the world's leading developers and deployers of AI tools and services, trust in our products is one of our core tenets and we welcome the overarching goal of the regulation. In addition, Amazon made a commitment to promote safe, secure and transparent development of AI technology in July 2023 at the White House. We encourage policymakers to continue pursuing an innovation-friendly and internationally coordinated approach, and we are committed to collaborating with the EU and industry to support the safe, secure, and responsible development of AI technology.

Sources

[1] https://d1.awsstatic.com/responsible-machine-learning/responsible-use-of-machine-learning-guide.pdf

[2] https://www.linkedin.com/posts/adamselipsky_ai-activity-7126213617829445632-ltDG/?trk=public_profile_like_view

https://aws.amazon.com/machine-learning/responsible-ai/

https://aws.amazon.com/blogs/machine-learning/aws-reaffirms-its-commitment-to-responsible-generative-ai/

https://www.aboutamazon.com/news/company-news/amazon-responsible-ai

https://aws.amazon.com/blogs/machine-learning/announcing-new-tools-and-capabilities-to-enable-responsible-ai-innovation/

https://aws.amazon.com/machine-learning/responsible-ai/policy/

https://aws.amazon.com/sagemaker/clarify/

Getting Started

Accelerate your healthcare innovation with AWS resources fitted to your adoption needs. You can choose your path. Either lead development in-house with ample enablement from AWS or engage an AWS Partner to custom design and build a solution for you. To get started, customer stories highlight how organizations are leveraging generating AI to unlock value. A high-level introduction and overview of what generative AI is and how it can address concerns and challenges that are on an executive's mind, can be found here. This YouTube 8-Hour Deep Dive series on Generative AI Foundations give insight into the conceptual fundamentals, practical advice, and hands-on guidance to pre-train, fine-tune, and deploy state-of-the-art foundation models on AWS and beyond. Organizations that are looking for guidance in terms of the right use cases can work with the Generative AI Innovation Center. All of this can help you ideate how generative AI can address your specific organization's opportunities and challenges.

For self-guided building, arm your team with proven best practices for ML in the cloud. Reach out to your account team to consult 1:1 with experienced AWS Solutions Architects who will guide you through security, data, and model design considerations unique to healthcare workloads. If you haven't met your account team yet, please reach out to our sales team to get introduced. Up-skill your data scientists with on-demand AWS Training focused on generative AI use cases. Validate their cloud proficiency through rigorous AWS Certifications. Fuel your team's creativity by collaborating in the Generative AI Innovation Center. Hit the ground running by engaging in an AWS Experience-Based Accelerators, a unique on-site Party where our staff and your staff will build together and leave with a working prototype.

With best practice architecture already in place, you can dedicate more time to experimentation and innovation. Take advantage o as-needed guidance rom AWS experts so your in-house teams can maintain momentum building healthcare solutions. With the right knowledge transer and resources, you gain the autonomy to transorm care while benefiting rom AWS security, scalability and agility.

After an introduction, generative AI-powered applications, like AWS HealthScribe or Amazon Q can be used out of the box to enable certain use cases. **Amazon Q Business**, a generative AI powered assistant can be tailored to an organization's custom data and processes. It can help streamline tasks, speed up decision making as well as problem solving. To experiment and get hands-on quickly without the need for an AWS account **PartyRock** can be used. Built on Amazon Bedrock, it enables users to learn about prompt engineering to build apps in just a few clicks.

To explore a custom-built solution, you can search 2,500+ AWS Partners verified in Healthcare and Life Sciences competencies to custom design and deploy your ML application. Prefer hands-on help from AWS directly? The AWS ProServe team offers end-to-end services - from ideation workshops to architecture review and solution deployment. With ProServe as your guide, you can fast track development of performant, scalable and secure ML solutions on AWS. Focus on transforming patient outcomes, while we handle secure ML infrastructure, tools and best practices. The expertise you need is available on-demand with AWS. Accelerate your healthcare innovation with AWS resources for every phase of machine learning adoption.

Authors and Contributors

Dr. Razvan Ionasec

Technology and Business Development Leader, Amazon Web Services, Germany

Razvan Ionasec, PhD, MBA, is the technical and business development leader for healthcare at Amazon Web Services in Europe, the Middle East and Africa. His work focuses on accelerating innovation across the industry, including medical imaging, genomics, health data and analytics, with a particular focus on cloud, AI/ML and generative AI. Previously, Razvan was global head of artificial intelligence (AI) products at Siemens Healthineers and responsible for AI-Rad Companion, the family of AI-powered and cloud-based digital health solutions. He holds more than 30 patents in AI/ML for medical imaging and has published more than 70 international peer-reviewed technical and clinical publications in computer vision, computer modeling and medical image analysis. Razvan received his PhD in computer science from the Technical University of Munich and his MBA from the University of Cambridge, Judge Business School.

Dr. Christoph Russ

Sr. Technical Program Manager, Amazon Web Services, Switzerland

Dr. sc. Christoph Russ leads strategic partnerships with lighthouse healthcare customers and partners in Europe, Middle East and Africa. His work focuses on building long-term partnerships that enrich the healthcare industry with positive, lasting impact. With a background in medical imaging and AI research, he has over 15 years of industry experience leading innovation and building solutions. As a builder, innovator, and entrepreneur in healthcare, he previously worked as a computer scientist at leading medical research and innovation organizations around the world, including CSIRO in Brisbane (Australia), Siemens Corporate Research (Healthineers) in Princeton (USA), and ETH in Zurich (Switzerland), where he was awarded a Doctor of Sciences based on his work on medical imaging and biomechanics simulation of medical devices.

Regina Hackenberg

Senior Solutions Architect, Amazon Web Services, Ireland

Regina Hackenberg is the Technical Lead for Healthcare and Healthtech at Amazon Web Services in Ireland. In this role, she is dedicated to assisting customers in overcoming data silos, deriving valuable insights from health data, and leveraging cutting-edge technology to drive advancements in healthcare.

James Wiggins

Senior Manager Public Sector Healthcare, AWS, United States

James Wiggins leads our global technical team for public sector healthcare (non-profit, academic, government) and our global healthcare technical community across all of AWS. This is a team composed of specialists in electronic health record systems, bioinformatics, genomics, health AI/ML, clinical and research medical imaging, mobile medical applications, open source research tools, and many other topics. They have deep technical skills, are passionate about healthcare, and get to work with the brightest minds in both fields. They support providers, payors, researchers, and regulators as they advance healthcare around the world by helping them apply cloud technology. James also engages with healthcare executives and industry analysts to help them understand what healthcare services AWS and our partners offer, the capabilities of our global healthcare tech team, and how customers are succeeding using AWS.

Ujjwal Ratan

Leader AI/ML and Data Science Healthcare and Lifesciences, AWS, United States

Ujjwal Ratan is the leader for AI/ML and Data Science team in the AWS Healthcare and Life Science business unit and is also a Principal AI/ML Solutions Architect. Over the years, Ujjwal has been a thought leader in the healthcare and life sciences industry, helping multiple Global Fortune 500 organizations achieve their innovation goals by adopting machine learning. His work involving the analysis of medical imaging, unstructured clinical text and genomics has helped AWS build products and services that provide highly personalized and precisely targeted diagnostics and therapeutics. Ujjwal's work has also been featured in multiple global conferences, peer-reviewed publications or technical and scientific blogs.

Mahesh Pancholi

CIO, UK Biobank, UK

Mahesh is the Chief Information Officer for UK Biobank. He is responsible for the Data and Technologies that underpin UK Biobank. A former bioinformatician who made the transition to IT early in his career, Mahesh remained close to his scientific roots by specialising in Research Computing, with a particular interest in enabling large scale analyses, democratising access to data and addressing scientific inequity. Prior to joining UK Biobank, Mahesh was Head of Research Computing at a Russell Group University and a commercial leader in Genomics and Life Sciences Research at both OCF and Amazon Web Services.

Prof. Dr. Wieland Sommer

Founder & CEO, Smart Reporting, Germany

Wieland is trained as a medical doctor and is a professor in radiology. For several years, he was the head of the oncologic imaging department at the Ludwig-Maximilians University of Munich. Next to his academic career, he is an enthusiast of entrepreneurship and founded two digital health companies: Smart Reporting, an digital health start-up based in Munich which revolutionizes medical documentation as well as Planerio, an AI-based automated scheduling and workforce management software for the healthcare sector. He also is an active business angel in the digital health sector and a mentor of entrepreneurs in medicine.

Prof. Dr. Jeroen van der Laak

Professor of Computational Pathology, Radboudumc, The Netherlands

Jeroen van der Laak is professor in Computational Pathology and principle investigator at the Department of Pathology of Radboud University Medical Center in Nijmegen, The Netherlands and guest professor at the Center for Medical Image Science and Visualization (CMIV) in Linköping, Sweden. His research group investigates the use of deep learning-based whole-slide image analysis for different applications; improvement of routine pathology diagnostics, objective quantification of immunohistochemical markers, and study of novel imaging biomarkers for prognostics. Jeroen has an MSc in Computer science and acquired his PhD from the Radboud University in Nijmegen. He co-authored over 160 peer-reviewed publications and is member of the editorial boards of Modern Pathology, Laboratory Investigation and the Journal of Pathology Informatics. He is chair of the taskforce 'AI in Pathology' of the European Society of Pathology, member of the board of directors of the Digital Pathology Association and organizer of the Computational Pathology Symposium at the European Congress of Pathology. He coordinated the CAMELYON grand challenges in 2016 and 2017. Jeroen van der Laak acquired research grants from the European Union and the Dutch Cancer Society, among others. Jeroen is coordinator of the Bigpicture consortium and USCAP Nathan Kaufman lecture laureate.

Alberto Rizzoli

Co-Founder & CEO, V7, UK

Alberto is an Italian-born entrepreneur and Co-Founder of V7, a UK technology company pioneering artificial intelligence to develop human-like visual cognition. Alberto began working on AI with Simon Edwardsson in 2015 with the creation of the first engine capable of running large deep neural networks on smartphones. This technology led to Aipoly, a camera app that identifies thousands of objects in real-time to aid the blind and visually impaired, scanning over 2 billion objects to date and being translated in 26 languages. His work on AI granted him an award from the President of Italy as well as the Premio Gentile for Science and Innovation in 2017. Today, Alberto's work through V7 enables laboratories to understand complex scientific experiments in real-time, and allows any business to set up, train, and deploy modern artificial intelligence into any device from robotic manipulators to portable devices. V7's products won the CES Best of Innovation Award for two consecutive years, in 2017 and 2018.

Andrea Riposati

Co-Founder & CEO, Dante Genomics, Italy

Andrea Riposati is co-founder and CEO of Dante Genomics (formerly Dante Labs), a global genomic information company building and commercializing a new class of transformative health and longevity applications based on whole genome sequencing and AI. Before starting Dante Labs, Andrea was the CEO of Muse Technologies, a B2B digital software and consulting company, and a Sr. Product Manager at Amazon, in Seattle and New York, where he launched the Amazon 3D Printing Store and the Amazon B2B Marketplace. Andrea started his career at Booz Allen Hamilton as a strategy and management consultant in digital and healthcare. Andrea holds a Master's Degree in Business and Economics summa cum laude from Bocconi University in Milan, Italy, and a Master in Business Administration (MBA) from Harvard University.

Mattia Capulli, Phd

Co-Founder & Chief Scientific Officer, Dante Genomics, Italy

Prof. Mattia Capulli, PhD, is the Co-founder and Chief Scientific Officer of Dante Genomics, and associate professor of Human Embryology. Recognized as a pioneer in his field, Prof. Capulli introduced and spearheaded the application of the lean six sigma approach to the lab sequencing process. This innovative approach has streamlined laboratory operations and enhanced efficiency in scientific research. Prof. Capulli published over 35 peer-reviewed scientific articles. His work has garnered international recognition, earning him national and international scientific and business awards. Prof. Capulli continues to make significant contributions to the fields of biotechnology, genomics and rare disease therapeutics.

Dr. Khan Siddiqui

Co-Founder, Chairman & CEO, HOPPR, USA

Dr. Siddiqui is a world-renowned clinical radiologist, pioneer in the field of AI and a serial entrepreneur. His astounding network stretches through medicine, academia, technology, government, clinical associations, and finance. He holds dozens of patents in deep learning, AI, image processing, data visualization, MR imaging, and secure patient information handling. His work at Microsoft in the early 2000's included the first multimodal AI model in medical imaging and developing the necessary infrastructure and training for the large AI model behind Xbox Kinect, which was based on nearly one billion images. Dr. Siddiqui is a serial entrepreneur with a track record of success. He is the Chief Medical Officer of Hyperfine, Inc., the creator of the world's first FDA 510(k) cleared, portable MRI scanner, and took Hyperfine public via a SPAC process ($HYPR). He also founded higi, Inc., a consumer health technology company, which was acquired by a technology company in a $4.2 billion SPAC transaction. Dr. Siddiqui has held significant roles at Microsoft, Johns Hopkins University, and the University of Maryland. He was the founding chair of the American College of Radiology's Imaging and Informatics Commission and has demonstrated his leadership and expertise throughout the field of radiology. He contributed to dozens of academic radiology informatics programs in the United States and is a frequent lecturer on building highly scalable technology companies, building data driven digital health companies and the rapid commercialization of medical and digital health technologies, particularly, AI in radiology and medical imaging startups.

Prabhu Arumugam, MD, PhD, FFCI

Director of Clinical Data and Imaging, Genomics England, UK

Prabhu Arumugam is Director of Clinical Data Imaging at Genomics England having joined the organisation in 2019. Prabhu trained in medicine at St. Bartholomew's and the Royal London. He trained in Histopathology and completed his PhD at The Barts Cancer Institute on pancreatic pathology. He is leading the multimodal programme, focusing on the utility of linking whole genomes to digital pathology and radiology imaging. He is also Genomics England's Caldicott Guardian. Through his clinical practice and in his role at Genomics England, he is passionate about harnessing the power of new genomic technologies for the benefit of all patients.

Dr. Ruben Amarasingham

CEO, Pieces Technologies, Inc

Dr. Ruben Amarasingham is a physician, scientist, entrepreneur, and Founder and CEO of Pieces, a healthcare AI R&D firm that specializes in clinical generative AI and machine learning solutions to support and augment clinical teams. Dr. Amarasingham is a national expert in health services research and AI systems design to support clinical quality improvement. He has won $50 million in scientific grants and numerous patents in these areas. Dr. Amarasingham practiced medicine for 13 years and was named by the Robert Wood Johnson Foundation as one of 10 leaders likely to change American healthcare. Prior to his role as CEO of Pieces, Dr. Amarasingham was Associate Chief of Medicine at Parkland Health & Hospital System, Founder of the Parkland Center for Clinical Innovation, an informatics research institute in Dallas, and an Associate Professor in Internal Medicine and Biomedical Informatics at the University of Texas Southwestern (UTSW) Medical Center.

Jonathan Larbey

Founder & CEO, T-Pro, Ireland

Founder and CEO of T-Pro, a global leader in Clinical Document Improvement software. Our cloud-based clinical documentation solutions enable workflows for efficient and accurate speech recognition, medical document generation and coding. We facilitate a patient centred solution by making it easy for doctors to capture the patient narrative, and by delivering information when it is needed most – at the point of care. Our mission is to help the healthcare industry overcome the clerical burden of documentation whilst maximising the value of the data collected. Ultimately, allowing clinicians and healthcare leaders to focus on patient care.