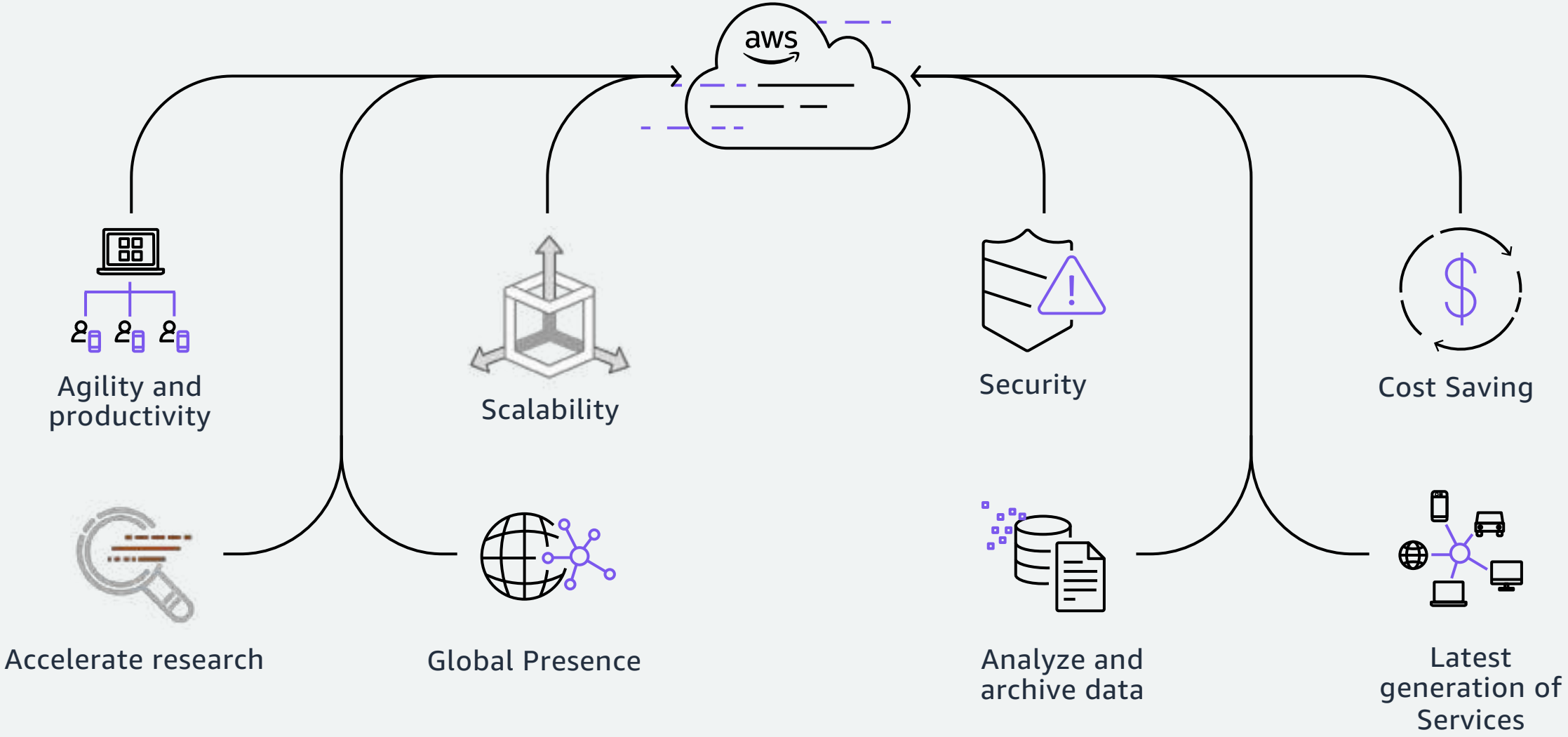
AWS FOR ITER-CEA-PPPL– 11TH JUNE 2024

Accelerate research outcome with AWS

AWS Education & Research Team

Roberta Piscitelli, PhD, MBA - piscitr@amazon.com

Why using AWS Cloud for Research and Education



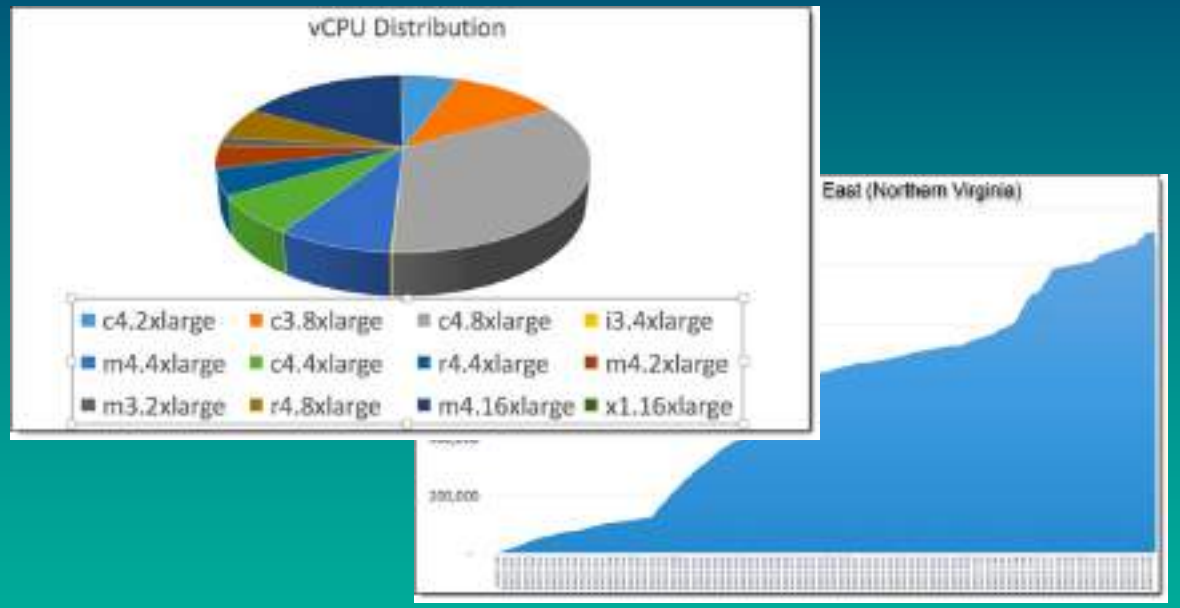
Clemson University - Natural Language Processing

https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/

The researchers conducted nearly half a million topic modeling experiments to study how human language is processed by computers.

The 1.1 Million vCPU count usage is comparable to the core count on the largest supercomputers in the world.

CLEMSON UNIVERSITY



"I am absolutely thrilled with the outcome of this experiment. The graduate students on the project […] used resources from AWS and Omnibond and developed a new software infrastructure to perform research at a scale and time-to-completion not possible with only campus resources."

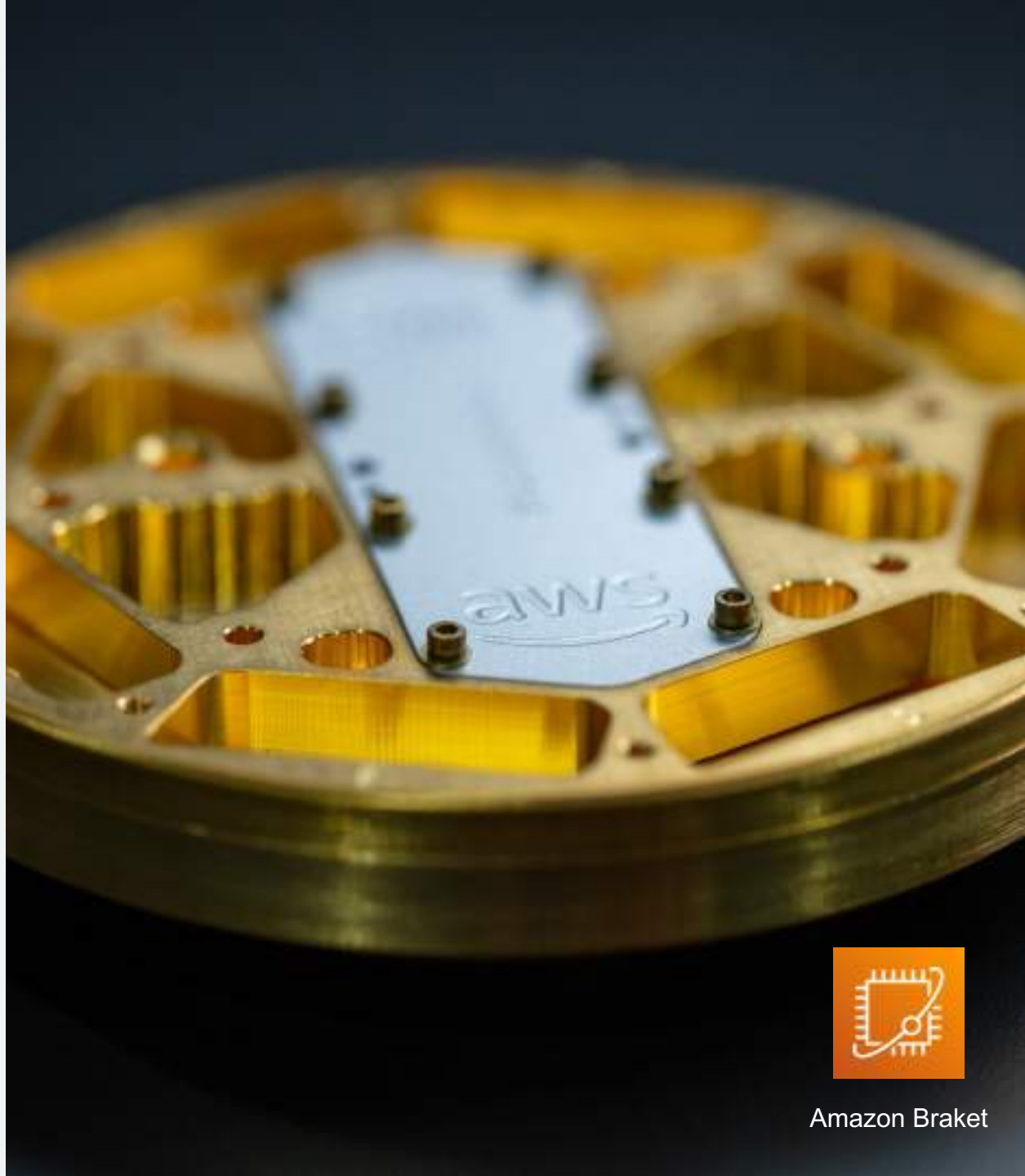
– Prof. Amy Apon, Co-Director of the Complex Systems, Analytics and Visualization Institute

Amazon EC2

NATIONAL INSTITUTE OF NUCLEAR PHYSICS AND AWS WORK TOGETHER TO ACCELERATE QUANTUM COMPUTING RESEARCH

“We are pleased to partner with AWS in what is an important element of our global strategy in quantum computing research.”

Marco Pallavicini, executive board member of INFN.



Amazon Braket

MeteOcean: Operational forecast

UniGe | DICCA

MeteOcean

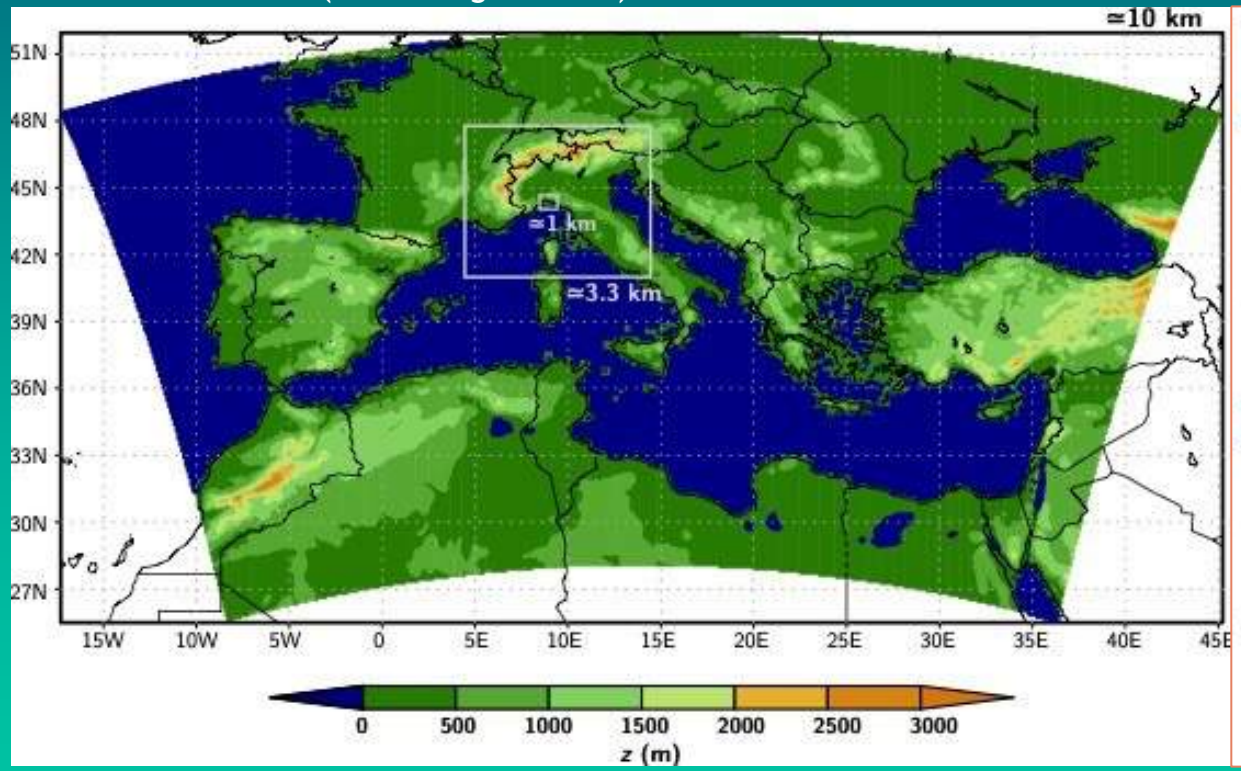
Daily 5-day hourly forecast running on **C5.24XLarge** machines

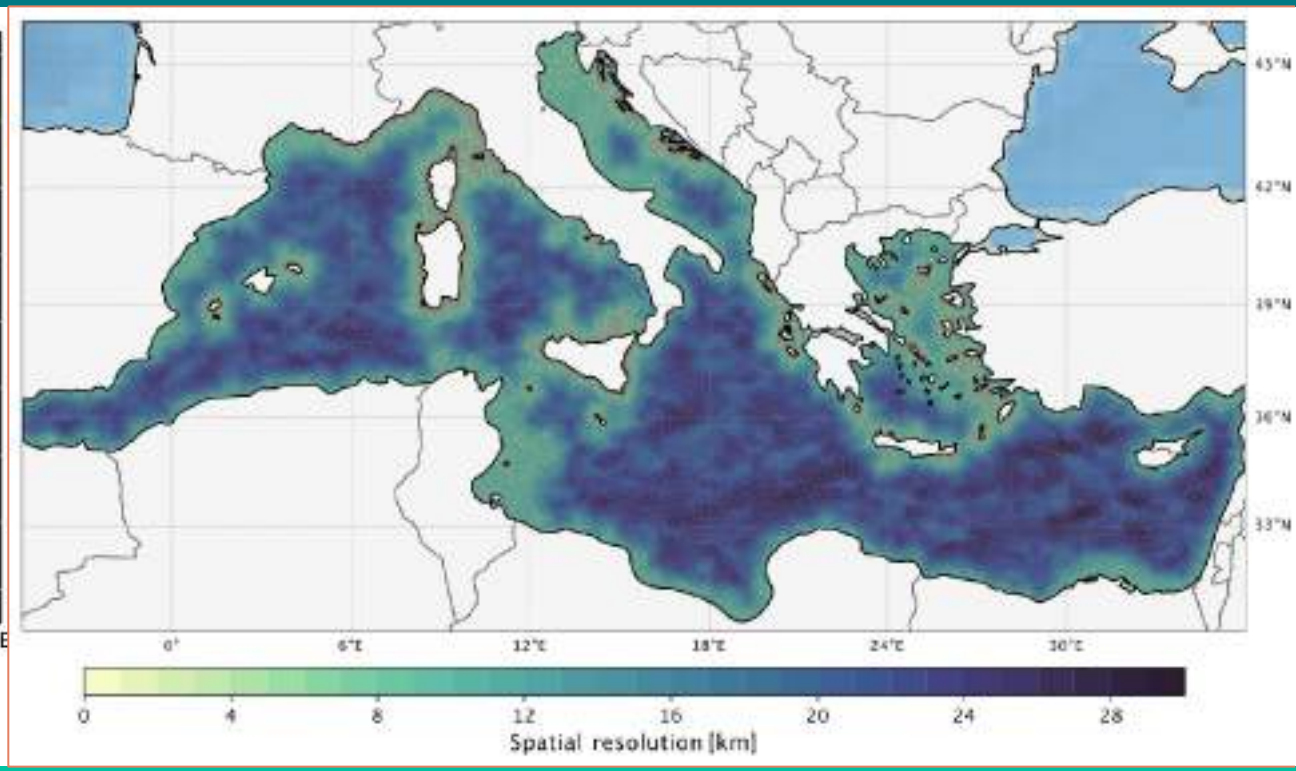
Atmosphere

35 vertical levels
3 nested domains with increasing resolution
→ **10 - 3.3 - 1.1 km (6.381.865 grid nodes)**

Waves

Initialized from 10-m wind field Unstructured grid with increasing resolution
→ **20 km - 10 km - 500 m - 300 m (50960 grid nodes)**

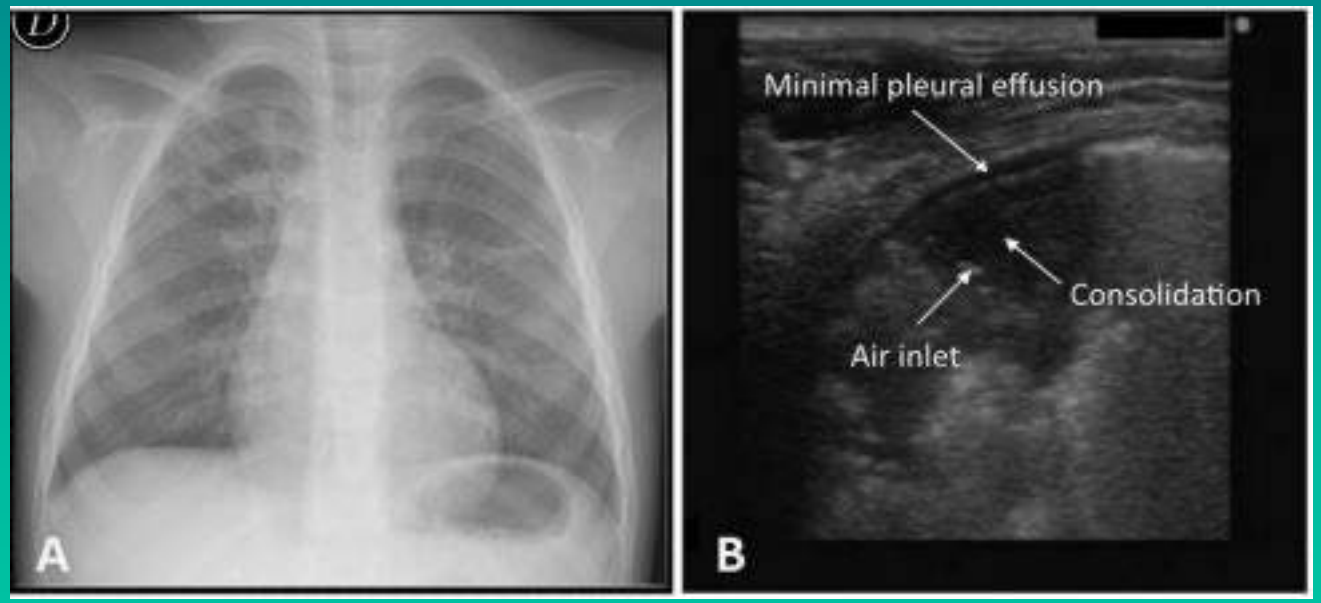




Sharing & collaborating UniTrento - ICLUS

SOLUTION

University of Trento's Department of Information Engineering and Computer Science used AWS to run a International Project (https://www.disi.unitn.it/iclus) where AI algorithms were used to analyze ultrasound images of lungs to determine possible Covid infections.



More than 29 joint pubblications

More than 60 istitutes involved

Italy

- 118 Castelnuovo, Garfagnana (LU)
- APSS Trento, Trento (TN)
- Ausl Romagna - Cesena, Cesena (NULL)
- Azienda Ospedaliera Universita di Padova, Padova (PD)
- Azienda Ospedaliera Universitaria Federico II, Napoli (NA)
- Azienda Ospedaliero Universitaria Policlinico Vittorio Emanuele, Catania (CT)
- Bresciamed, Brescia (BS)
- Cardiologia - Ospedale Policlinico San Martino, Genova (GE)
- Emergency Department of Arzignano Hospital - AULSS8 Berica, Vicenza (VI)
- Fondazione Policlinico San Matteo IRCCS, Pavia (PV)
- Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma (RM)
- Mater Olbia Hospital, Olbia (SS)
- Ospedale Civile di Voghera, Voghera (PV)
- Ospedale dei Bambini Vittore Buzzi, Milano (MI)
- Ospedale di Sanremo Asl1 Imperiese, Sanremo (IM)
- Ospedale di Tione, Tione (TN)

Other countries

- Augusta University - Department of Emergency Medicine, Augusta, United States of America
- Clinic of thoracic and vascular surgery, Gera, Germany
- Contra Costa Regional Medical Center, Martinez, California, USA
- DeepMed I/O, Mancester, United Kingdom
- Department of Obstetrics and Gynecology - University Hospitals Leuven, Leuven, Belgium
- DSP Medea, Medea, Algeria
- Eindhoven University of Technology, Eindhoven, The Netherlands
- Hospital das Clinicas da Universidade de São Paulo, San Paolo, Brasile
- Hospital General de Catalunya, Barcellona, Spain
- IFC - CNR LECCE, Lecce, ITALIA
- Indian Institute of Technology, Kharagpur, India
- Indian Institute of Technology, Jodhpur, India
- Indian Institute of Technology Patna, Patna, India
- Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom
- Intelligent Ultrasound, Cardiff, United Kingdom
- Klinikum rechts der Isar der TU, München, Germany
- Leitat Technological Center, Barcellona, Spain
- Massachusetts General Hospital, Boston, USA
- Michigan State University, Lansing, USA
- North Carolina State University, Raleigh, United States of America

Solutions and services for the entire research process

- Research proposal support
- Letters of support
- Cloud economics
- Compliance documents

- OCRE Framework
- Deployment in minutes
- 200+ services
- Training
- Partner solutions
- Open source solutions (e.g. Parallel Cluster)
- Solution architects
- Professional services

- Single sign on
- Landing Zone

- AWS Snow Family
- AWS IoT for sensors
- AWS Ground Station

- Streaming data
- Event driven arch

- Databases
- Data lakes
- Data warehouses

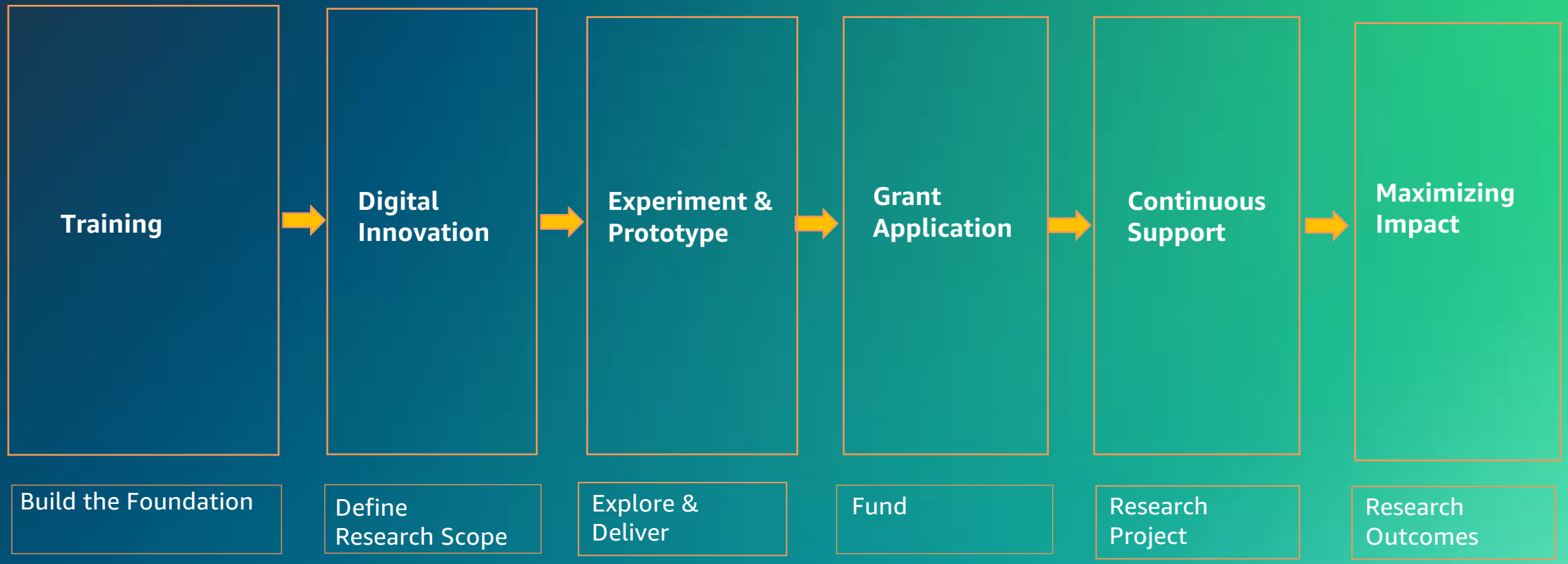
- Bulk storage options

- **Interactive notebooks**
- **HPC (AWS Parallel Cluster)**
- **Machine learning**
- Containers
- Big data analytics
- Visualisation (NICE DCV)
- 400+ instance types
- GPUs, FPGAs, ARM, **Inferentia, Trainium**
- **Quantum computing**

- **Open Data Registry**
- Trusted Research Environments
- Content delivery network
- Long term low cost archival

- **AWS Marketplace**
- Building SaaS
- Partners
- **Proserve**

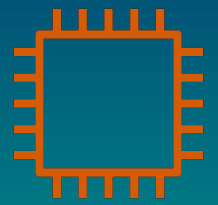
AWS grant acceleration support package

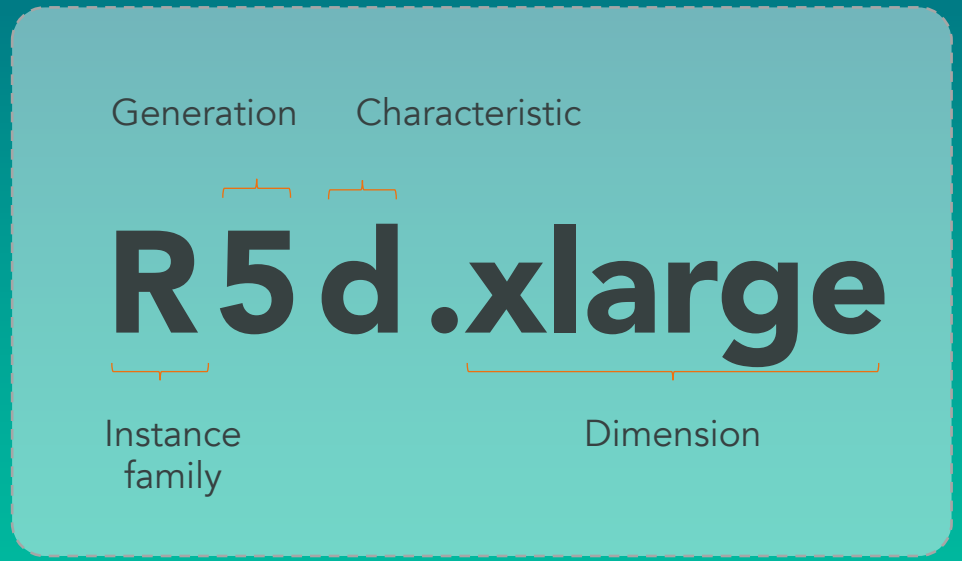


Build research

Amazon EC2

- The broadest and most in-depth computing platform
- On-demand infrastructure
- Scalable computing power
- No long-term contracts or upfront commitments
- Wide choice of operating systems and software
- Prices based on effective use
- Scalability and high performance
- Reliability and security





S3 Object Storage – a new kind of storage

Amazon S3

S3 Standard

S3 Intelligent-Tiering

S3 Standard-IA

S3 One Zone-IA

S3 Glacier Instant Retrieval

S3 Glacier Flexible Retrieval

S3 Glacier Deep Archive

Frequent ← *Access Frequency* → *Infrequent*

The Italian National Institute of Astrophysics Explores the Universe with the Cloud

INAF
ISTITUTO NAZIONALE DI ASTROFISICA
NATIONAL INSTITUTE FOR ASTROPHYSICS

"Thanks to AWS, we were able to concentrate on science and simulations.
We were able to scale as soon as the project required us to do so. It was critical to obtain the required power quickly"

Marco Landoni, INAF researcher

AWS Lambda

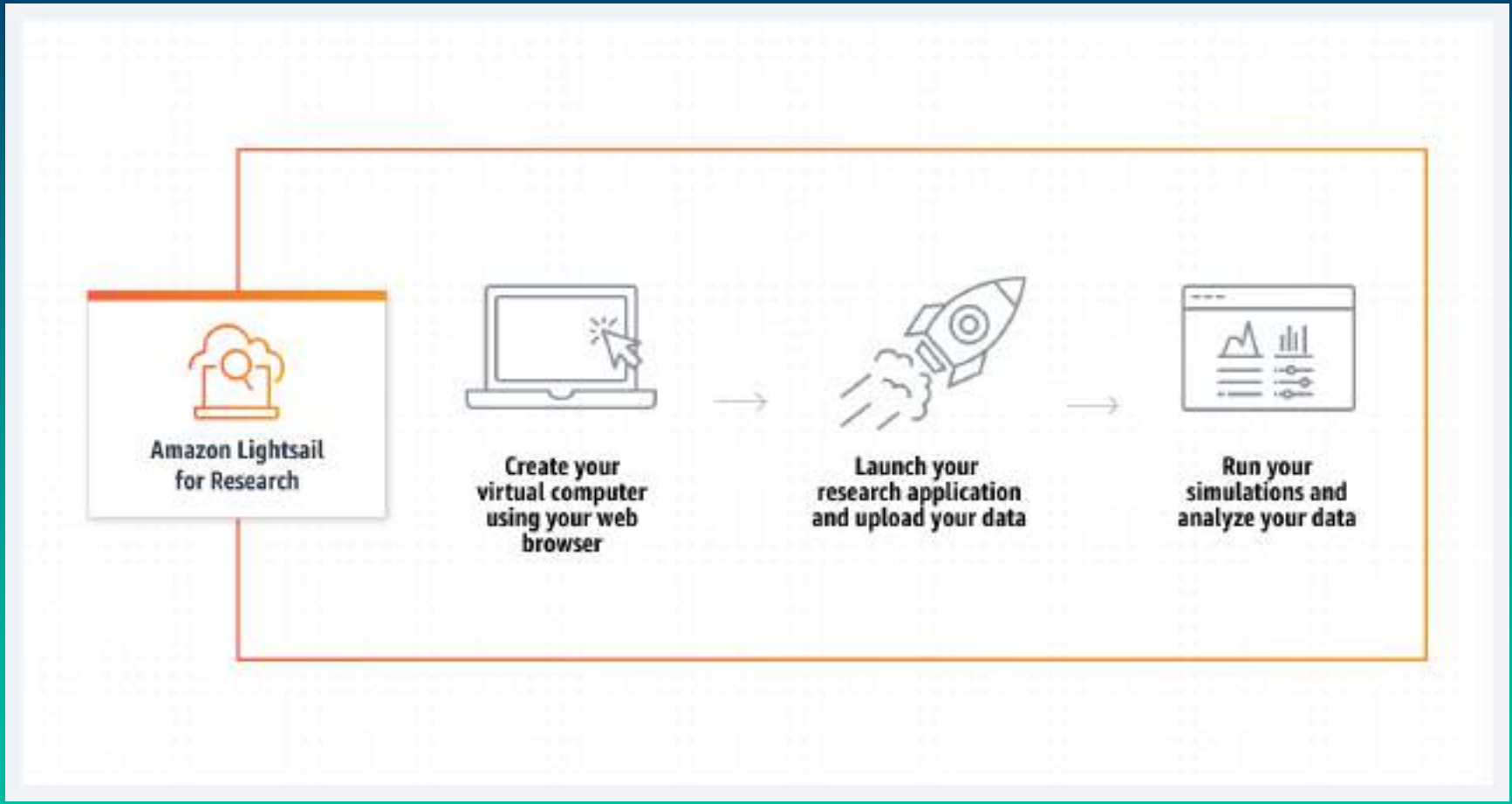
Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service Glacier

Amazon Elastic Compute Cloud (Amazon EC2)

Lightsail for Research

https://aws.amazon.com/lightsail/research/



Voice of the researcher: requirements

Reduce time to run

Access research environments in minutes

Security

Maintain consistent security, compliance, and governance

Resilient data ingestion

Ingest datasets at scale

Integrated visualization

Post-analysis visualization and insights

Spend controls

Cost visibility, centralized budgeting, and chargeback management

Universally accessible

Collaborate from any location

End-to-end resource provisioning

Storage, compute (AWS ParallelCluster), and visualization from single pane of glass

Research Gateway

Secure, performant, and scalable research workbench that . . .

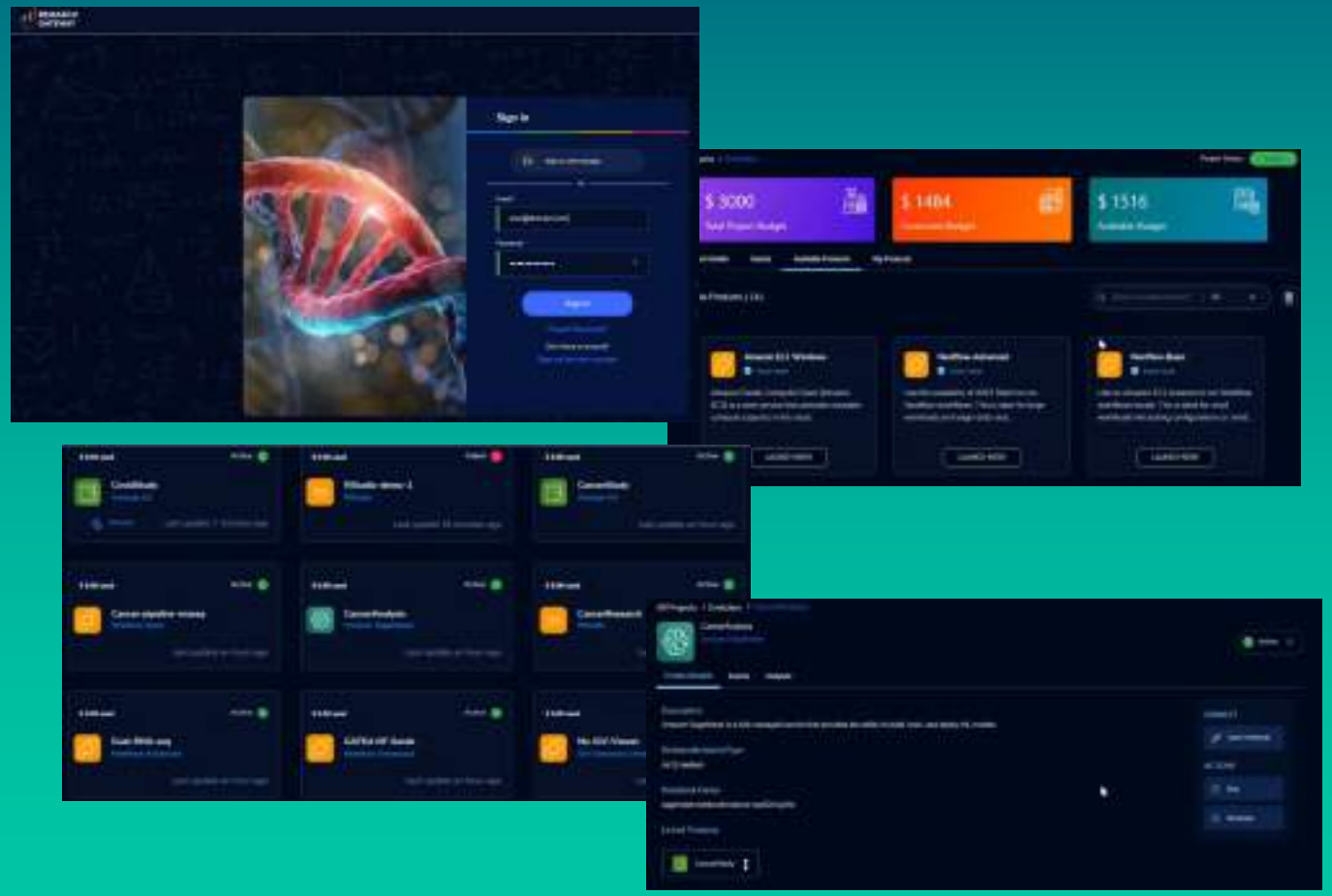
. . . can be provisioned in minutes

. . . is backed by secure and resilient data ingestion framework

. . . is delivered with wide selection of compute resources, including HPC

. . . is delivered with role-based access to workbench and data

. . . has clear budget and consumption costs breakdown by project, user, and workload



What are Research Gateway use cases?

relevance lab

Analytics	Genomics workflow	High performance computing	Machine learning and analytics	Customized AMIs
Research budget tracking	Bioinformatics toolkit	Secure data egress	Study open datasets	Standard product catalog and marketplace access
Secure research workspaces	AWS data lake and secure ingress	BioContainers and researcher tools	HIPAA regulations and trusted research	Research community collaboration

RONIN is a cloud orchestration and collaboration platform, lowering the entry level to using the cloud for researchers and research IT.

- Deployed within one AWS account, serving many researchers
- Enables researchers self-service access to AWS resources
- Enforces an institution’s security policy

Research and Engineering Studio on AWS

Open source, easy-to-use web-based portal for administrators to create and manage secure cloud-based research and engineering environments.

Benefits:

- Minimize administrative overhead
- No cloud expertise required
- Flexible access to services

Use cases:

- Collaborate using shared research and engineering environments
- Define and manage projects
- Enable access to AWS without creating individual accounts

Acquiring data

The most comprehensive set of data services

DATA LAKES

- **Open Data Registry**

DATABASES

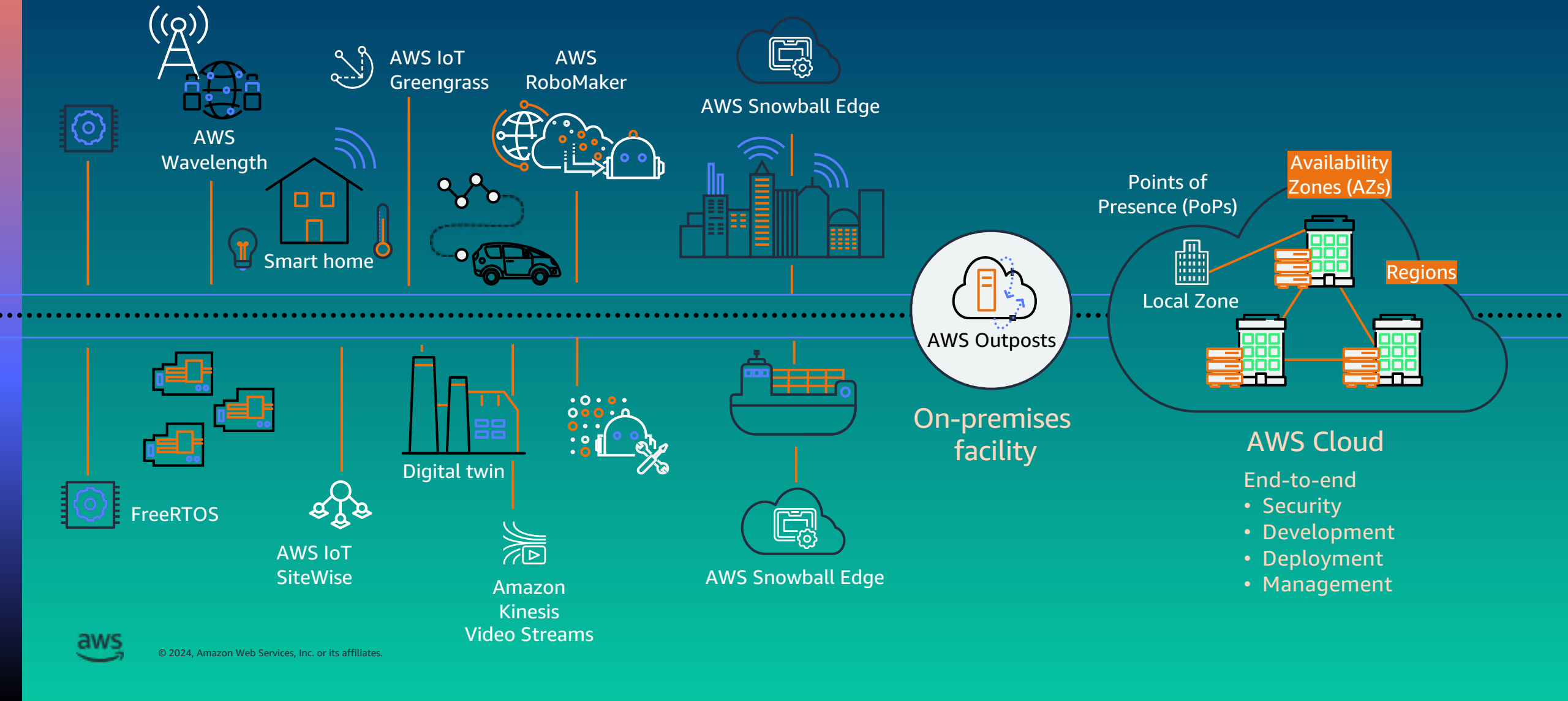
- RELATIONAL
- KEY-VALUE
- DOCUMENT
- GRAPH
- TIME-SERIES
- LEDGER
- WIDE COLUMN
- MEMORY

ANALYTICS

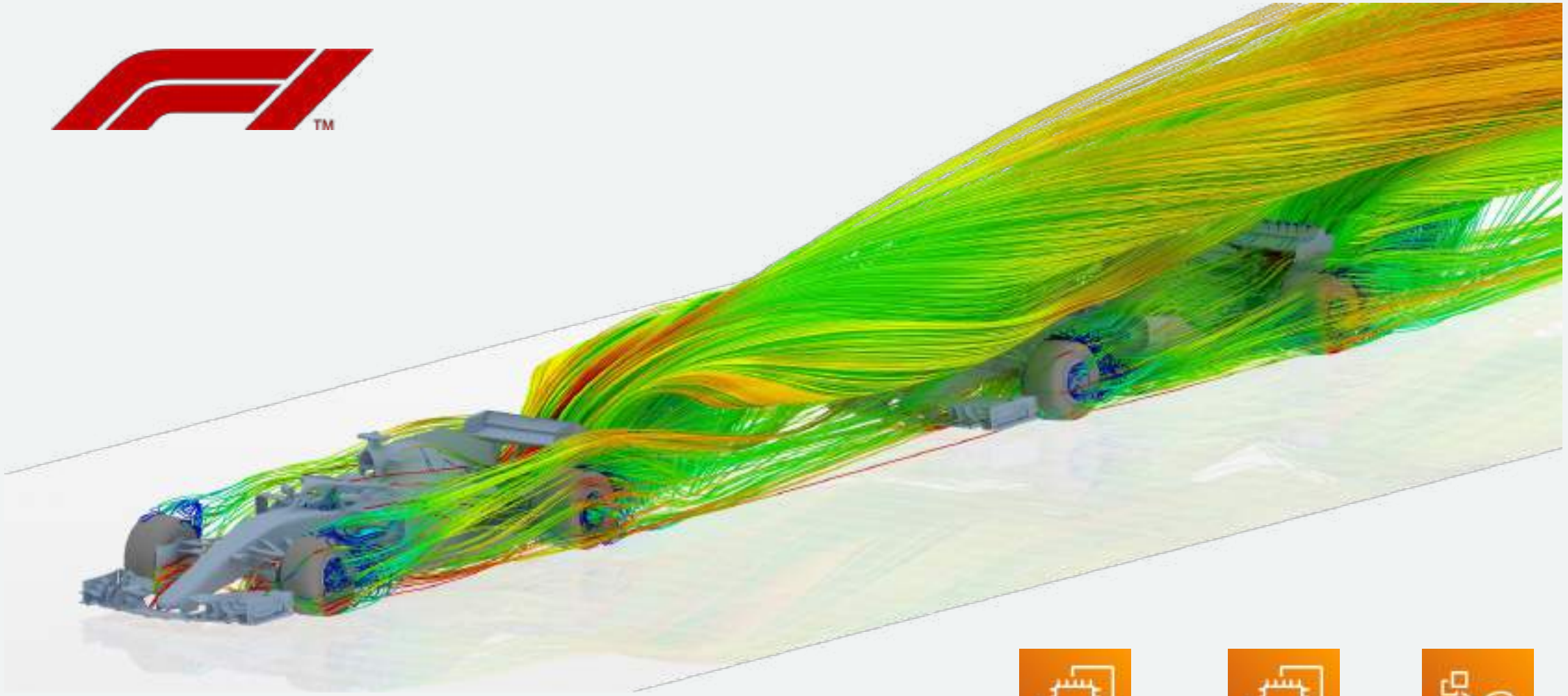
- INTERACTIVE QUERY
- BIG DATA PROCESSING
- REAL-TIME ANALYTICS
- DATA WAREHOUSING
- DATA INTEGRATION
- BUSINESS INTELLIGENCE
- OPERATIONAL ANALYTICS

MACHINE LEARNING & GENERATIVE AI

A complete platform for building and deploying edge applications



Formula 1 – computational fluid dynamics simulations



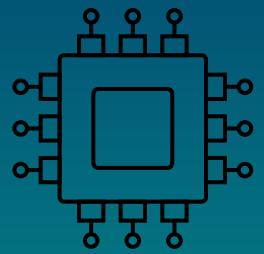
AWS Graviton

Amazon EC2

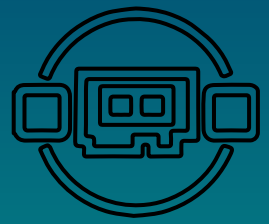
AWS ParallelCluster

Data analysis & simulations

Services to enable HPC on AWS

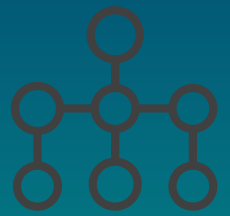


Amazon EC2



Elastic Fabric Adapter (EFA) + up to 400 Gbps networking

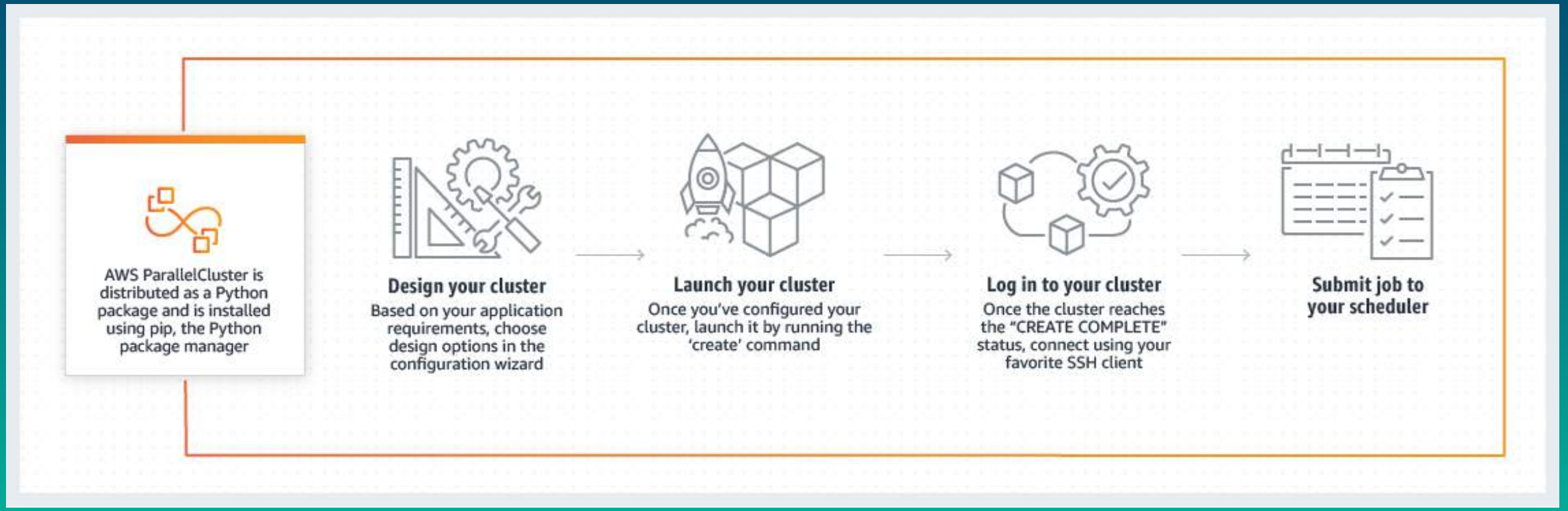
Amazon FSx for Lustre



AWS Batch

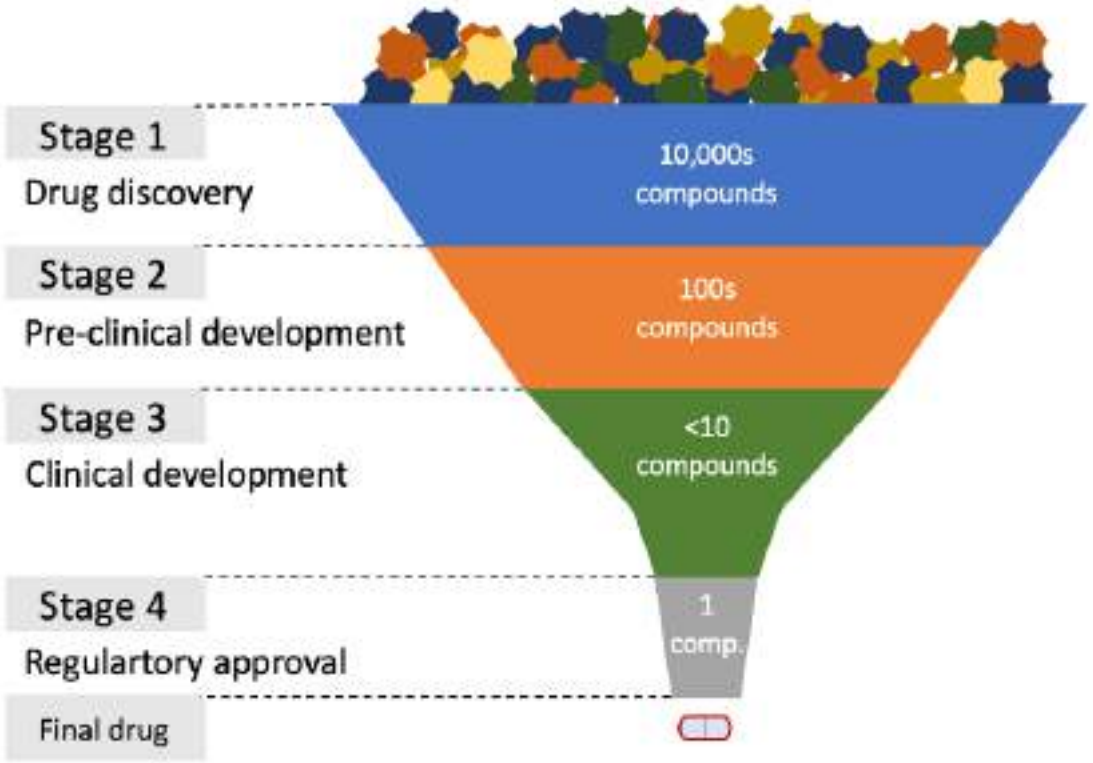
AWS ParallelCluster

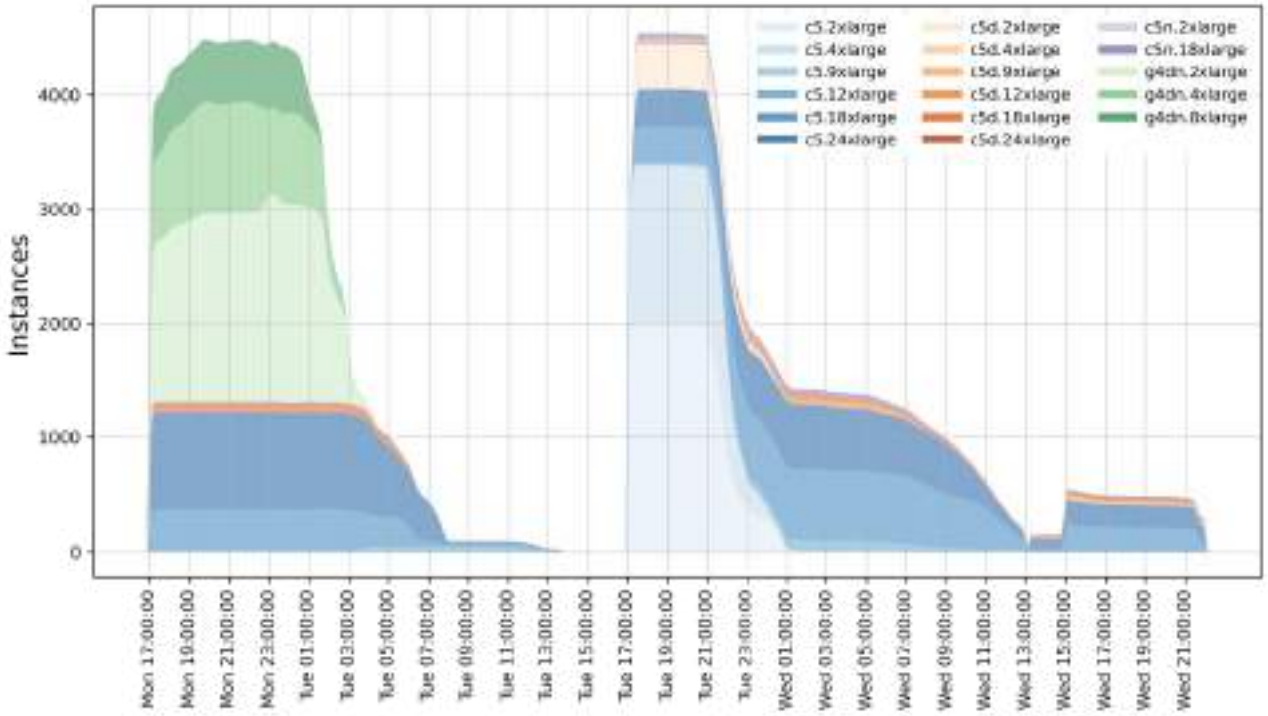
AWS PARALLELCLUSTER IS AN OPEN SOURCE CLUSTER MANAGEMENT TOOL THAT MAKES IT EASY FOR YOU TO DEPLOY AND MANAGE HIGH PERFORMANCE COMPUTING (HPC) CLUSTERS ON AWS



Agility and productivity - GROMACS: Max Planck Institute

MAX PLANCK GESELLSCHAFT

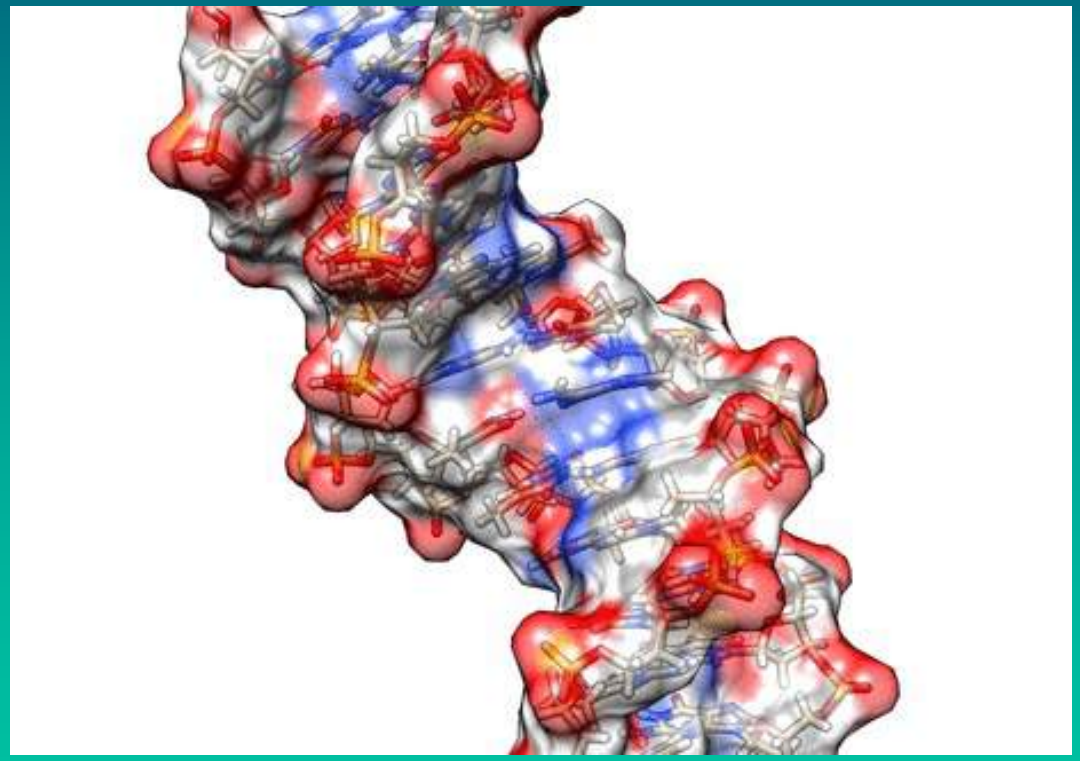




https://aws.amazon.com/blogs/hpc/running-20k-simulations-in-3-days-with-aws-batch/
https://pubs.acs.org/doi/10.1021/acs.jcim.2c00044#

Amazon S3 | NICE DCV | Amazon EC2 Auto Scaling | AWS ParallelCluster

The-university of Nottinghams Crossbow project paves a new path for biomolecular research using high performance computing HPC and the cloud



University of Nottingham
UK | CHINA | MALAYSIA

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Simple Storage Service (Amazon S3)

Amazon Elastic File System (Amazon EFS)

Quantum computing at AWS

Amazon Braket

Democratise quantum computing

Access to state-of-art technologies

Amazon Quantum Solutions Lab

Provide expert guidance

Cross-discipline support

State-of-the-art algorithms

AWS Center for Quantum Computing

Push the boundaries

Research quantum algorithms and hardware

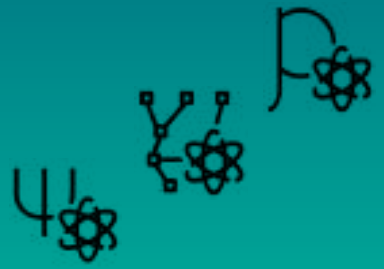
https://aws.amazon.com/products/quantum/

Amazon Braket – the AWS quantum computing service

A fully managed service that makes it easy for scientists and developers to explore quantum computing

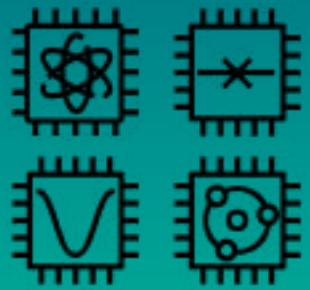
Build

- Amazon Braket Python SDK
- Jupyter notebooks
- Command Line Interface (CLI)



Test

- Local simulators for rapid testing
- High-performance simulators



Run

- Access multiple quantum computers
- Combine quantum and classical resources

Analyze

- Monitor algorithms in almost real time
- Analyze algorithm results and performance

Local and on-demand simulators

Local simulator

Part of Braket Python SDK

Fast and convenient prototyping

Number of qubits based on hardware

SV1: State vector simulator

Quantum circuit with up to 34 qubits

Stores the full wave function state

Concurrency: Default 35, max 100

TN1: Tensor network simulator

Quantum circuit with up to 50 qubits

Encodes quantum circuits into a structured graph

Concurrency: Default 10, max 10

DM1: Density matrix simulator

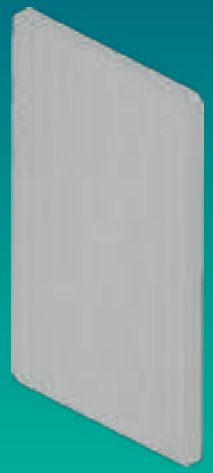
Quantum circuit with up to 17 qubits

Run multiple circuits in parallel with noise simulation

Concurrency: Default 35, max 50

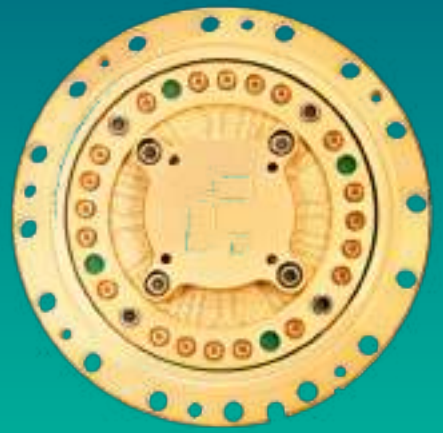
Available quantum computers

IQuEra Computing Inc.

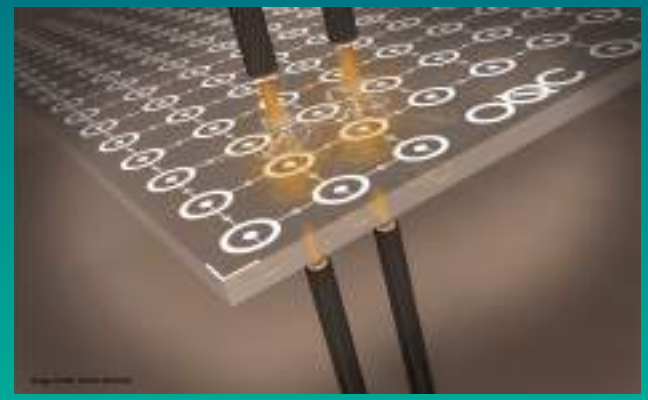


XANADU

IONQ



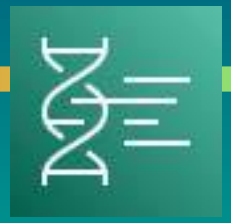
rigetti



OQC

Multimodal analytics

PURPOSE-BUILT SERVICES FOR HEALTHCARE AND LIFE SCIENCES



Amazon Omics

Transform genomic, transcriptomic, and other omics data into insights

Amazon HealthLake Imaging and Analytics

Provide a complete view of individual or patient population health data

Amazon Comprehend Medical

Understand medical context using natural language processing

Amazon Transcribe Medical

Automatically convert medical speech to text

The AWS ML Stack

Broadest and most complete set of machine learning capabilities

AI SERVICES

HEALTH AI

NEW Amazon HealthLake

Amazon Transcribe Medical

Amazon Comprehend Medical

INDUSTRIAL AI

NEW AWS Panorama + Appliance

NEW Amazon Monitron

NEW Amazon Lookout for Equipment

NEW Amazon Lookout for Vision

ANOMALY DETECTION

NEW Amazon Lookout for Metrics

CODE AND DEVOPS

NEW Amazon DevOps Guru

Amazon CodeGuru

VISION

Amazon Rekognition

SPEECH

Amazon Polly

Amazon Transcribe *+Medical*

TEXT

Amazon Comprehend *+Medical*

Amazon Translate

Amazon Textract

SEARCH

Amazon Kendra

CHATBOTS

Amazon Lex

PERSONALIZATION

Amazon Personalize

FORECASTING

Amazon Forecast

FRAUD

Amazon Fraud Detector

CONTACT CENTERS

Contact Lens

Voice ID *For Amazon Connect*

ML SERVICES

Amazon SageMaker

Label data

SAGEMAKER STUDIO IDE

NEW Aggregate & prepare data

NEW Store & share features

Auto ML

Spark/R

NEW Detect bias

Visualize in notebooks

Pick algorithm

Train models

Tune parameters

NEW Debug & profile

Deploy in production

Manage & monitor

NEW CI/CD

Human review

NEW: SageMaker JumpStart

NEW: Model management for edge devices

FRAMEWORKS & INFRASTRUCTURE

TensorFlow

mxnet

PyTorch

GLUON

Keras

learn

DeepGraphLibrary

Deep Learning AMIs & Containers

GPUs & CPUs

Elastic Inference

Trainium

Inferentia

FPGA

Amazon SageMaker

Build, train, and deploy ML models at scale

Automatic model fine-tuning & distributed training

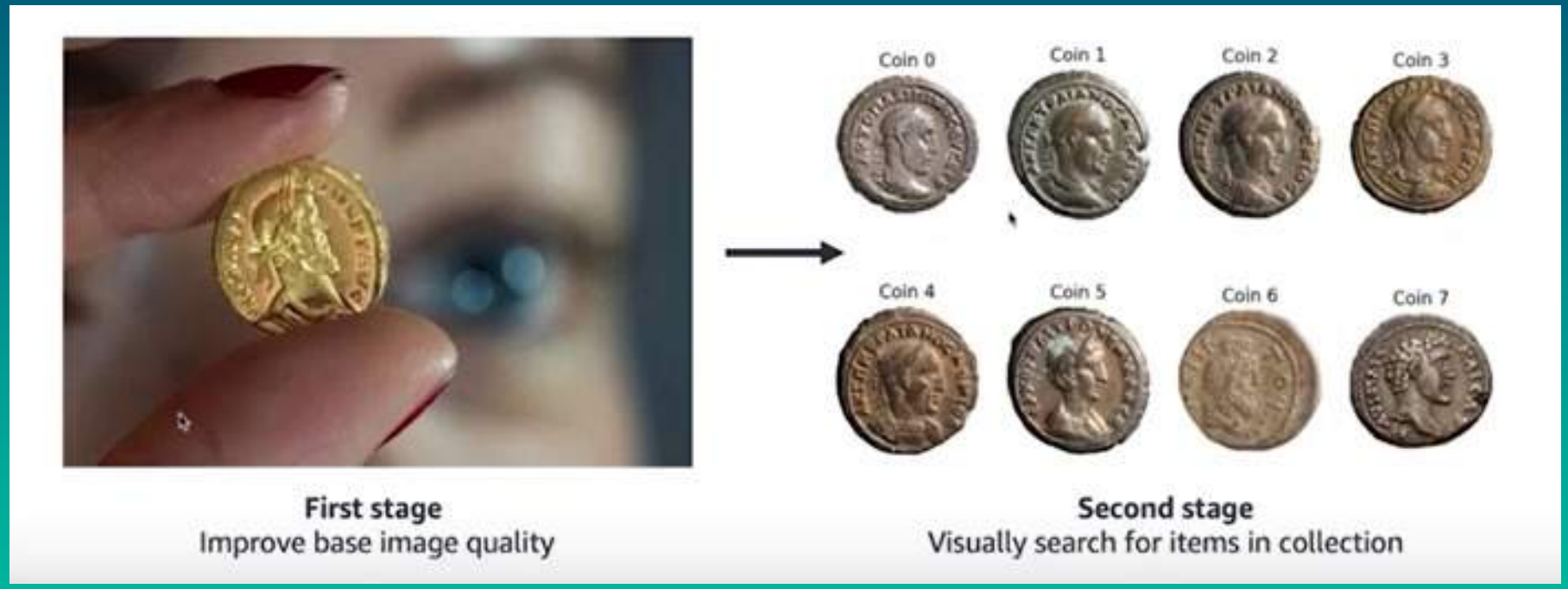
Flexible model deployment options

Tools for ML operations

Built-in features for responsible AI

University of Oxford Introduces a Sector-Leading Image Recognition ML Prototype to Augment Digitization in Numismatics

https://aws.amazon.com/solutions/case-studies/oxford-case-study/



UNIVERSITY OF OXFORD

"I thought this project would be complex and time consuming, but using AWS made it easy."

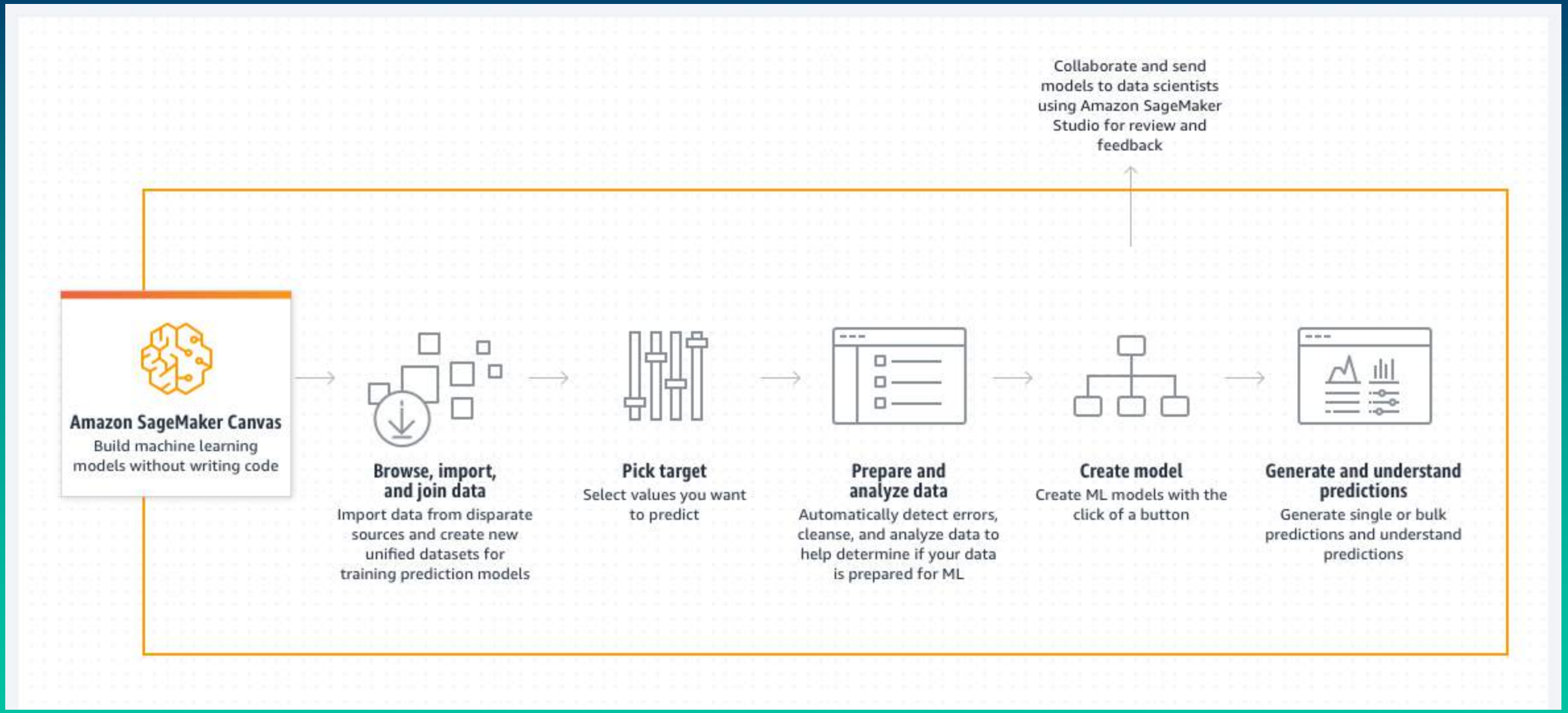
Anjanesh Babu
Systems architect and network manager, Gardens and Museums IT, University of Oxford's Gardens, Libraries & Museums

Amazon SageMaker

Amazon EC2

Amazon S3

SageMaker Canvas



Amazon Rekognition

Automated image, video, and text analysis

CONTENT MODERATION

OBJECT, SCENE, AND ACTIVITY

TEXT DETECTION

FACE DETECTION AND ANALYSIS

CELEBRITY RECOGNITION

VIDEO SEGMENTS & SHOTS

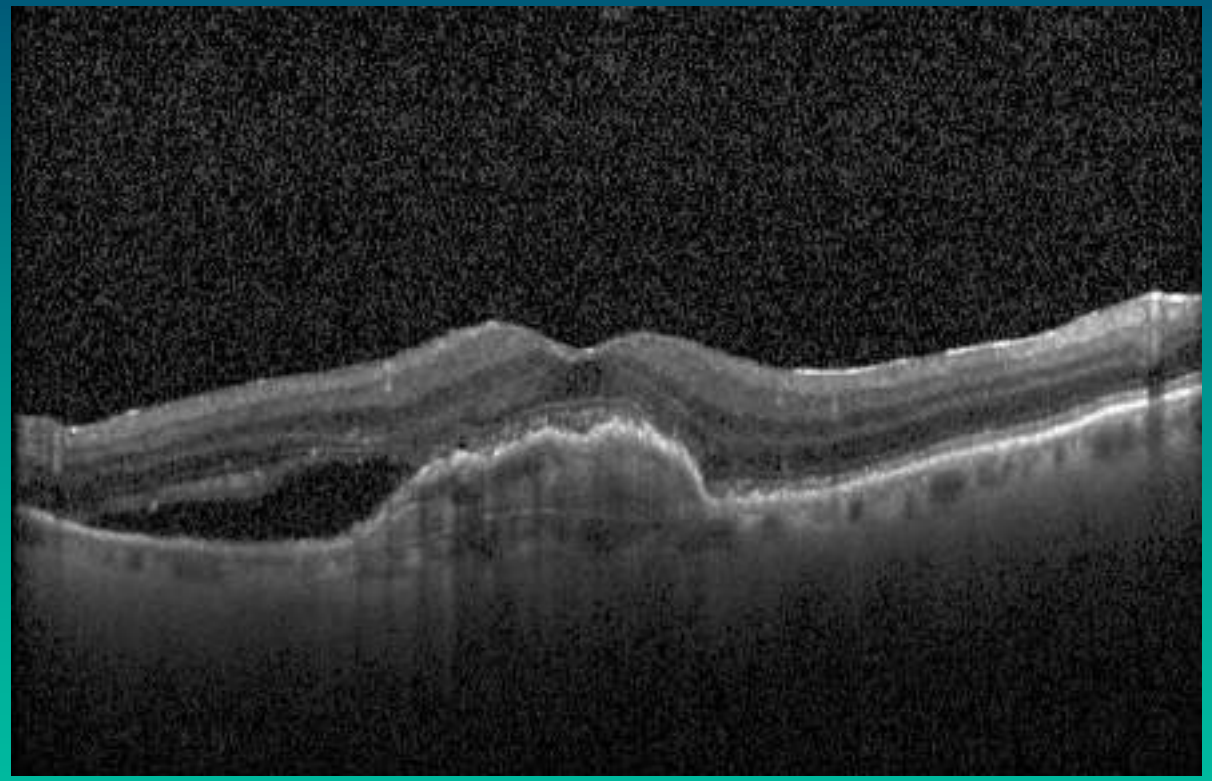
FACE COMPARE & SEARCH

STREAMING VIDEO EVENT DETECTION

LIVE STREAM VIDEO & PATHING

aws machine learning

Singapore Eye Research Institute categorizes retinal diseases using Amazon Rekognition



A retinal OCT image showing an eye with choroidal neovascularization (CNV).

https://aws.amazon.com/blogs/publicsector/singapore-eye-research-institute-categorizes-retinal-diseases-using-amazon-rekognition/

AutoML platforms like Amazon Rekognition provide easy-to-use interfaces that enable clinicians to apply AI to healthcare data

Amazon Comprehend

Discover insights and relationships in text

Documents

Email, chat, social, phone calls and more

Amazon Comprehend

Automatically extract insights from text

- Entities
  + Custom Entities
- Key Phrases
- PII
  (Personally Identifiable Information)
- Sentiment
- Document Classification
- Topics
- Language
- Syntax
- Events

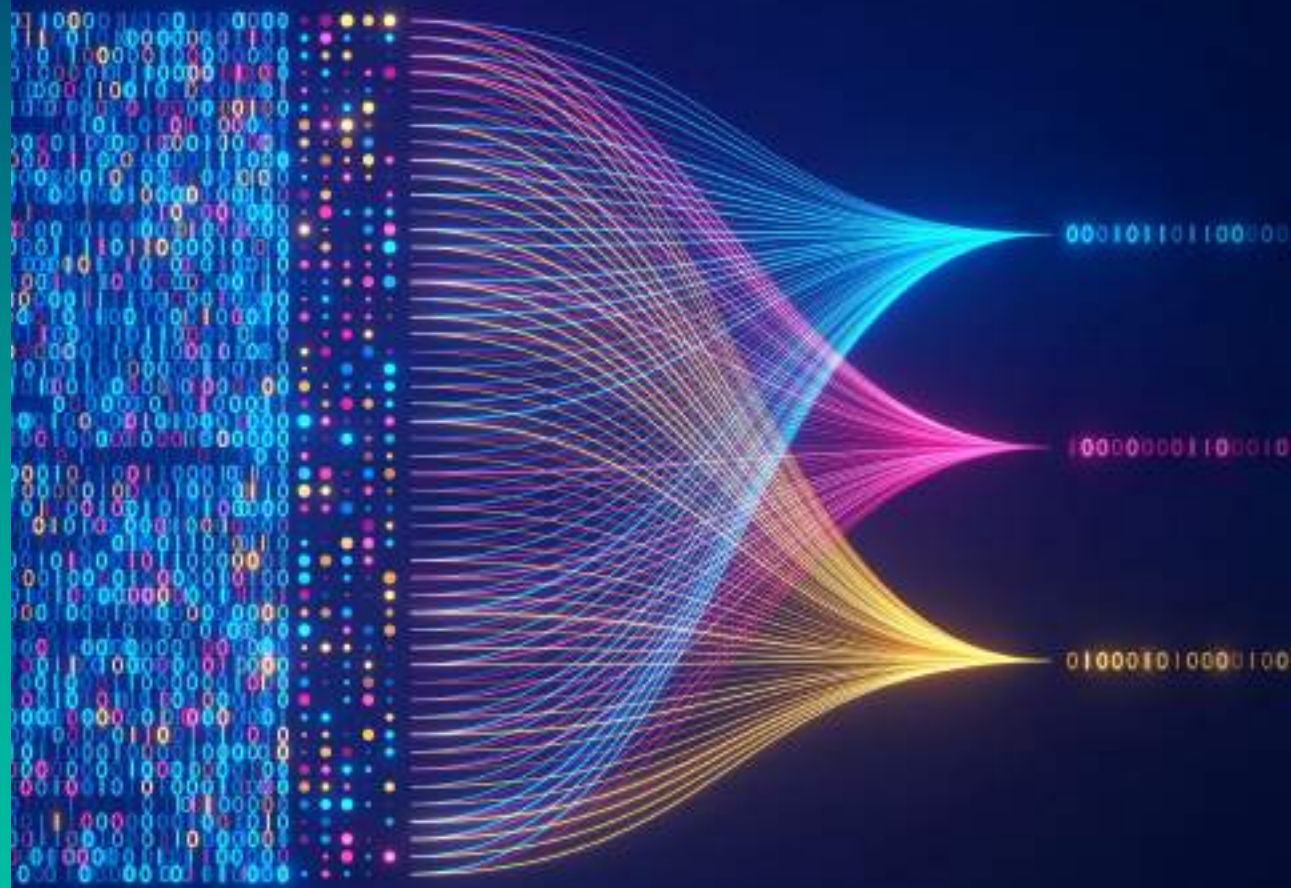
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

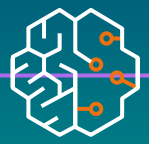
Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks

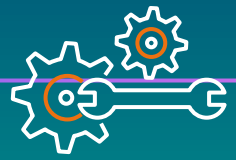


Generative AI (Foundation Models) refers to artificial intelligence that can generate novel content



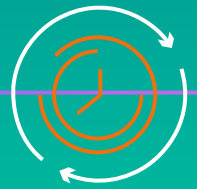
AI that can produce original content close enough to human generated content for real-world tasks

Powered by foundation models pre-trained on large sets of data with several hundred billion parameters



Tasks can be customized for specific domains with minimal fine-tuning

Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.



Reduces time and cost to develop ML models and innovate faster

Generative AI is used for a wide range of use cases in research

Research and discovery

Analyzing large-scale datasets and identifying patterns, information

Clinical development

Optimizing research protocols

Research funding support

Research application content generation

Researcher support

Chatbots with research specific insights

How Generative AI transforms artificial intelligence

image generation, transformation, upscaling

Generated by Stable Diffusion 2.0. This interior does not exist

Seamless transformation

4x

Upscaling

Amazon SageMaker

Large Models built on AWS

Clibrain develops Lince Zero—the first large language model (LLM) optimized for Spanish using Amazon SageMaker

Stability AI will build AI models on compute clusters with thousands of GPU or AWS Trainium chips, reducing training time and cost by 58%

BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance, utilizing AWS infrastructure services

New LLM development and refinement
UAE Technology Innovation Institute

CHALLENGE

TII researchers needed an easy way to develop, iterate, and distribute Falcon-40B LLM. Experimenting at such massive scale was slowed by limited computational resources and complex on-prem infrastructure. Researchers sought more flexibility to efficiently enhance, evaluate, and provide access to Falcon under dynamic capacity requirements to empower their team's foundation model research.

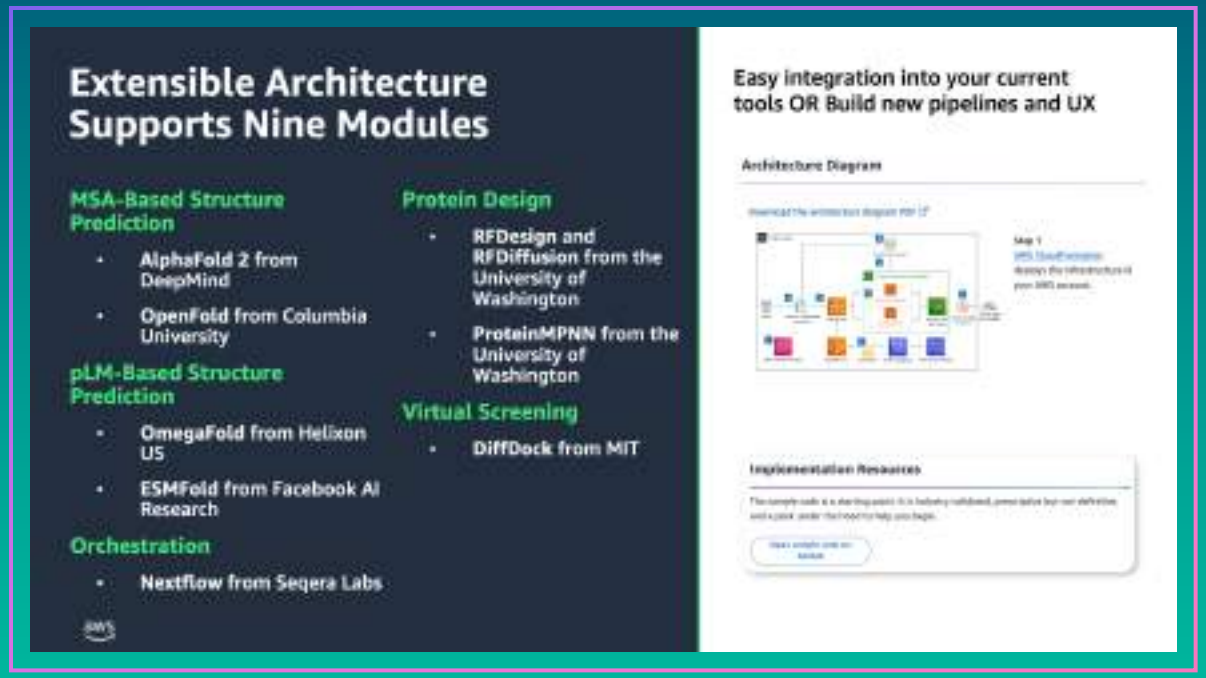
SOLUTION

TII utilized SageMaker's managed machine learning infrastructure and tools to train, host, and deploy versions of Falcon with greater speed and experimentation. SageMaker provided on-demand access to training clusters along with modular deployment options for easy distribution to varying client workloads—key for customizing model access.

OUTCOME

- ✓ TII successfully implemented Falcon-40B by using SageMaker and custom innovation. This allows TII's contribution within UAE's 2031 National AI Strategy, fostering economic growth and social progress. Releasing UAE's Falcon 180B, World's Top-Ranked Open Source AI Model will further encourage AI academic research and scientific collaboration.

Predict molecular properties with specialized FMs



Guidance for protein folding on AWS

Simplified UI with AWS Drug Discovery Workbench

Supports multiple algorithms within a shared user interface

Innovating at the silicon level

AWS Trainium 2

4x

Faster than AWS Trainium

65

Exaflops of on-demand supercomputing performance

AWS Inferentia 2

4x

Higher throughput

10x

Lower latency

Amazon Q Developer

AI-powered code suggestions in the IDE and the command line

Amazon Bedrock

simplifies

Choice

Customization

Integration

AI21labs amazon ANTHROP\C cohere Meta stability.ai

JURASSIC-2 AMAZON TITAN CLAUDE COMMAND + EMBED LLAMA 2 STABLE DIFFUSION XL

aws

Amazon Bedrock keeps data secure & private

None of the customer's data is used to train the underlying model

All data is encrypted in transit and at rest

Data used to customize models remains within your VPC

Support for standards, including GDPR & HIPAA

How AWS supports Gen AI research

Use familiar tools that meet your requirements

Customization

FM-as-a-service

FM-powered applications

Pre-trained FM

Fine-tuned FM
Industries (Legal, Telco, Financial, Education…)

FM monetization
Licensing

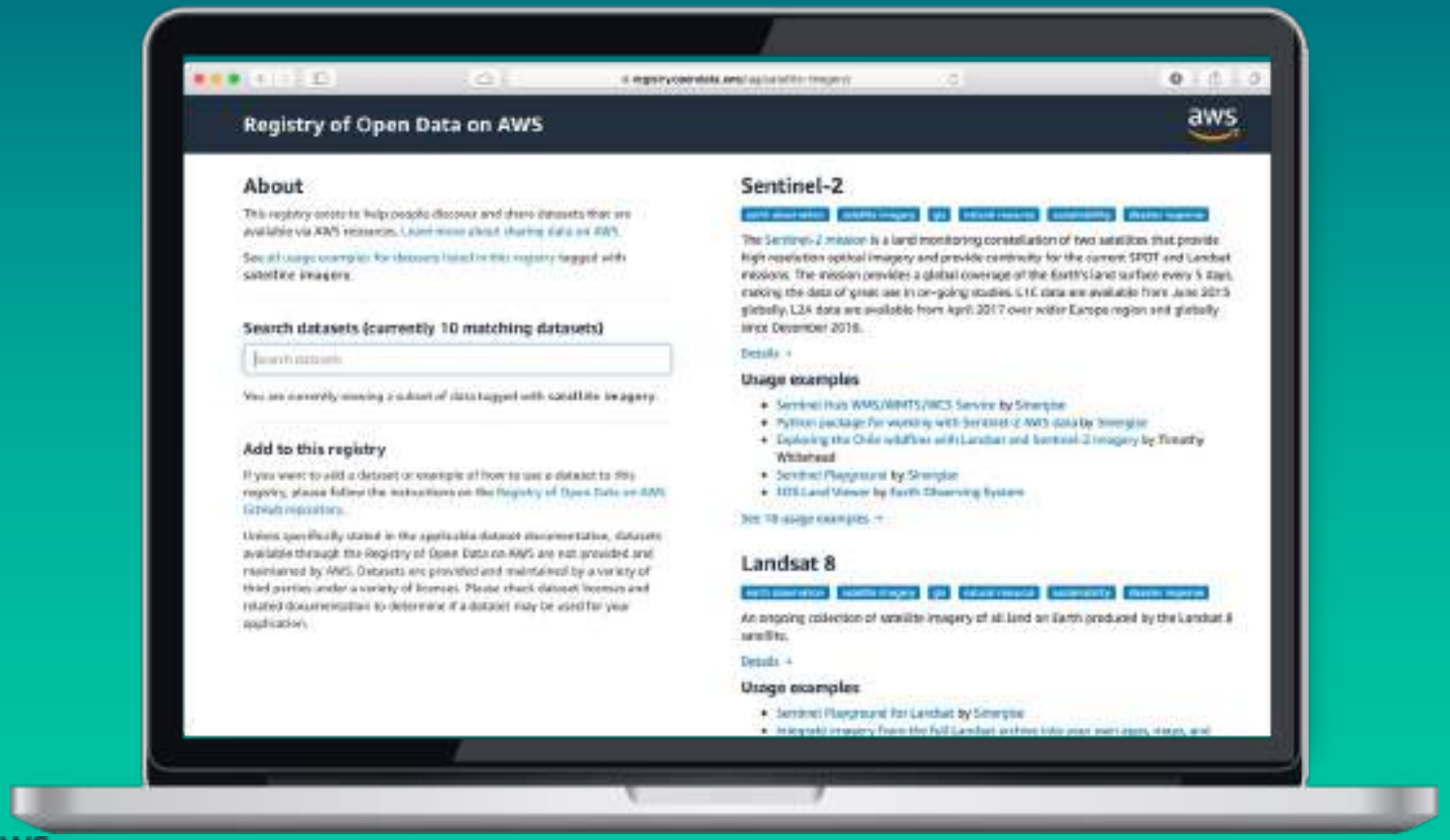
Start-up ecosystem
Public administration
Industry & Academia
Citizen services

Publishing and sharing research data

Open data on AWS

AWS hosts a variety of public datasets to lower the cost and improve the speed of research.

https://registry.opendata.aws/



Examples

- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- Landsat 8
- Common Crawl
- SpaceNet
- OpenStreetMaps

… Regularly updated

Monitoring at-risk bodies of water from space

The Bluedot Observatory uses Sentinel-2 satellite data on AWS to monitor water bodies around the world

"The cost to process one month of data for about 7,000 bodies of water currently in the system is 6 EUR. It is possible to set up world-scale systems with a shoestring budget."

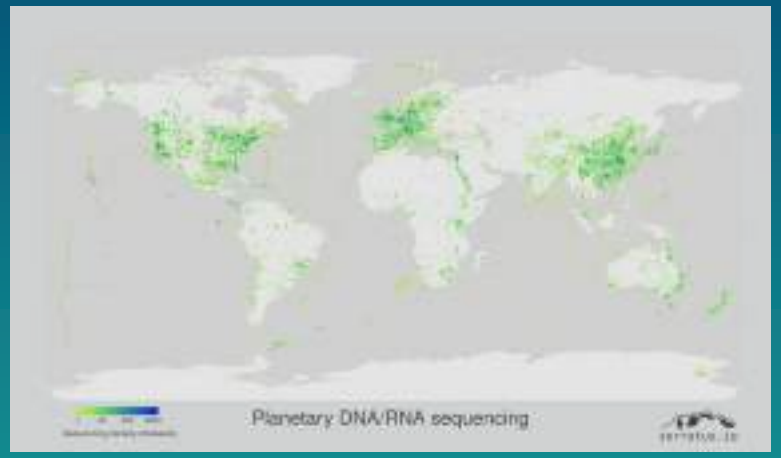
Grega Milcinski, Bluedot

opendata.aws/bluedot

https://aws.amazon.com/blogs/publicsector/bluedot-observatory-keeping-an-eye-on-our-planets-water-resources/

Logan Project: Planetary DNA/RNA Reconstruction

Objective: Develop a search engine for DNA/RNA



Will enable discoveries like Serratus in **[seconds, $0.01]** instead of **[days, $10k]**

Step 1 (2024): Download all SRA, reconstruct genomes, host on AWS **Registry of Open Data**
•~30M CPU hours of computation, 19 petabytes to download, 2 petabytes of results to store*

Step 2 (2025): Use advanced algorithms to index this data and host the search engine (Comparison with YouTube)

Logan Project: infrastructure AWS Batch

Services used:

Batch

S3

DynamoDB

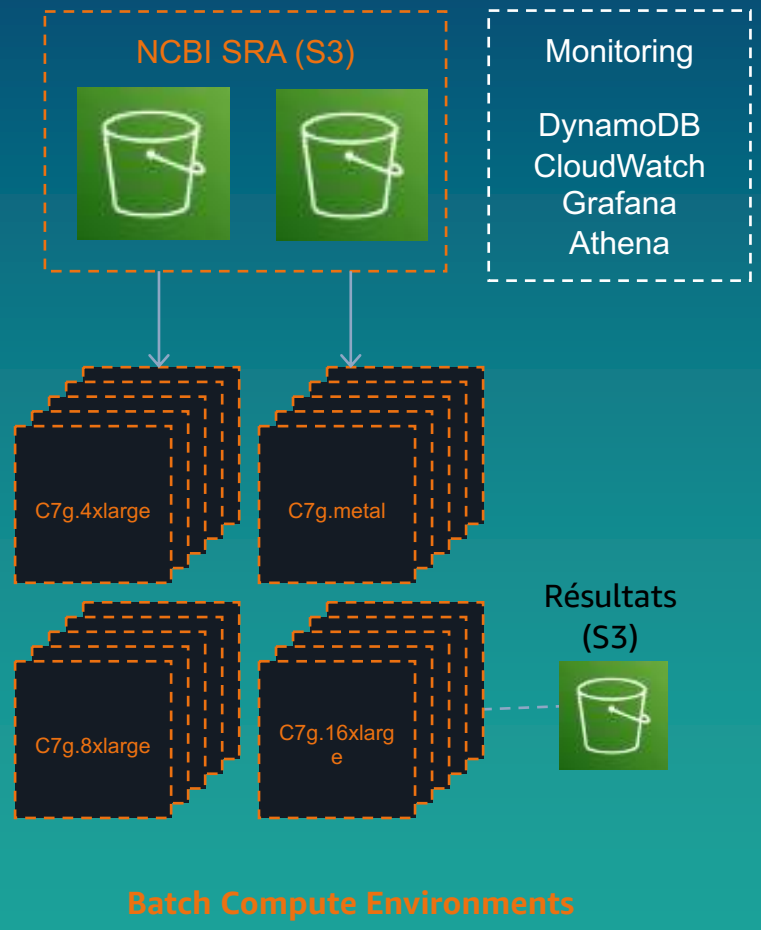
Athena

CloudFormation

CloudWatch

Cost Explorer

Grafana



Opérations

Downloading and decompressing SRA data

Reconstruction génomes (CPU/disque/RAM intensifs)

Compress results

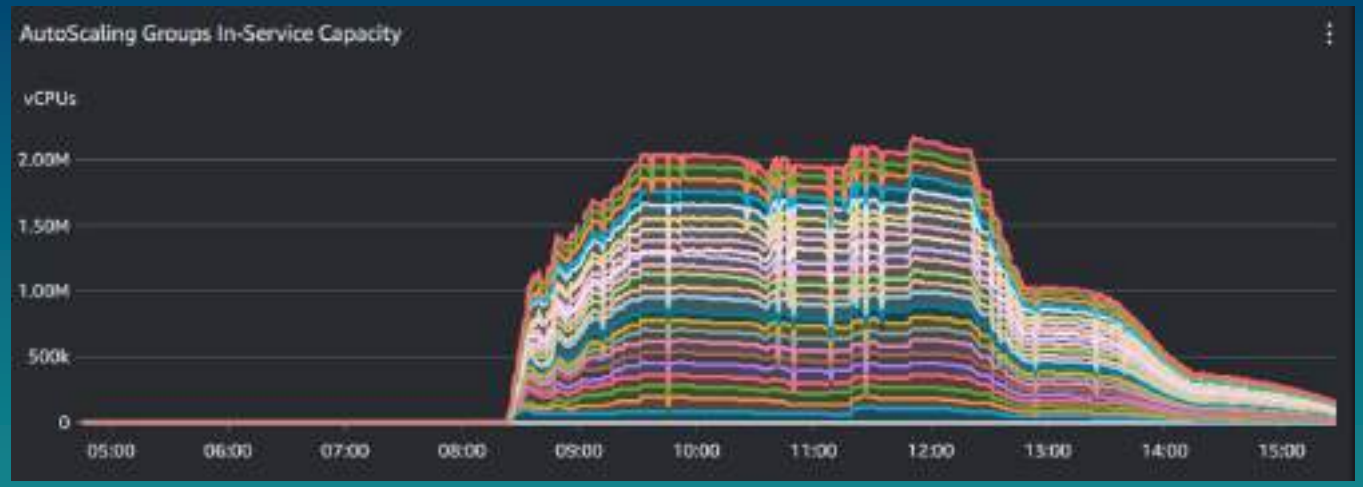
Logan Project: computations performed

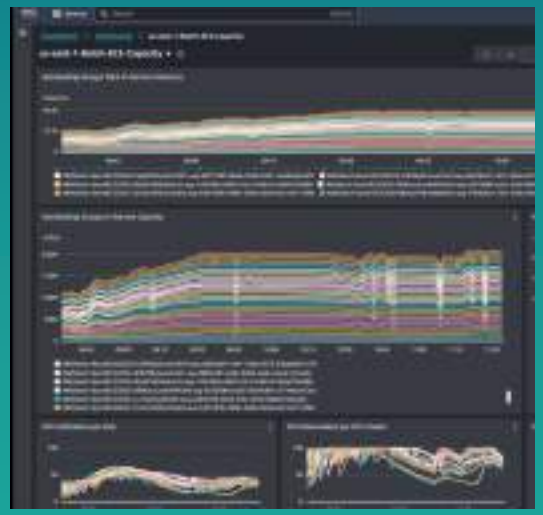
2.18M Graviton vCPUs at peak
30 hours of computation
equivalent to **3,500 CPU years**

Contacted AWS from the start
1 year of preparation, 4 tech. people

6 tests of 5-10 hours over 6 months

Technical calls every week





AWS investments supporting research and education

TRAININGS

AWS Training Academy

Demo/Immersion days

SUPPORT

Technical Support

AWS Research Scientists

Grant Proposals

INCENTIVES

Proof of concepts

AWS Cloud Credits for Research

Amazon Research Awards

Researchers have leveraged Amazon Research Awards support to innovate faster using the most advanced tools available in the cloud.

"Who we are shapes what we say and how we say it"

Staff writer
July 05, 2023

Amazon Research Award recipient Shrikanth Narayanan is on a mission to make inclusive human-AI conversational experiences.

CONVERSATIONAL AI / NATURAL-LANGUAGE PROCESSING

"Building a model that can save as many lives as possible"

Sean O'Neill
May 24, 2023

How ARA recipient Supreeth Shashikumar is using machine learning to help hospitals detect sepsis — before it's too late.

MACHINE LEARNING

Cracking the code of how diseases affect the body

Sean O'Neill
May 15, 2023

ARA recipient Marinka Zitnik is focused on how machine learning can enable accurate diagnoses and the development of new treatments and therapies.

MACHINE LEARNING

https://www.amazon.science/research-awards

Thank you!

AWS Education & Research Team

Roberta Piscitelli - piscitr@amazon.com