



Introduction to Amazon Web Services (AWS) for Researchers

Danyell Wilt
AWS Sr. Solutions Architect - Healthcare

Agenda

- Overview of AWS and common services
- Overview of security and the AWS Shared Responsibility Model
- Regulated data transfer and management
- How to deploy and use RStudio on AWS
- How to deploy and use Jupyter Notebooks on AWS

Overview of AWS

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark



How AWS can help your research



Science, not servers

Use compute when you need
It to do large-scale analysis



Collaboration

Access data sets that span institutions



Share effort

Leverage work done by
other scientists to save time



Reproduce research

A common platform for
reproducing scientific analyses



State-of-the-art analytics

Use data science methods
in your research



Security

A collection of tools to protect
data and privacy

Customer obsessed

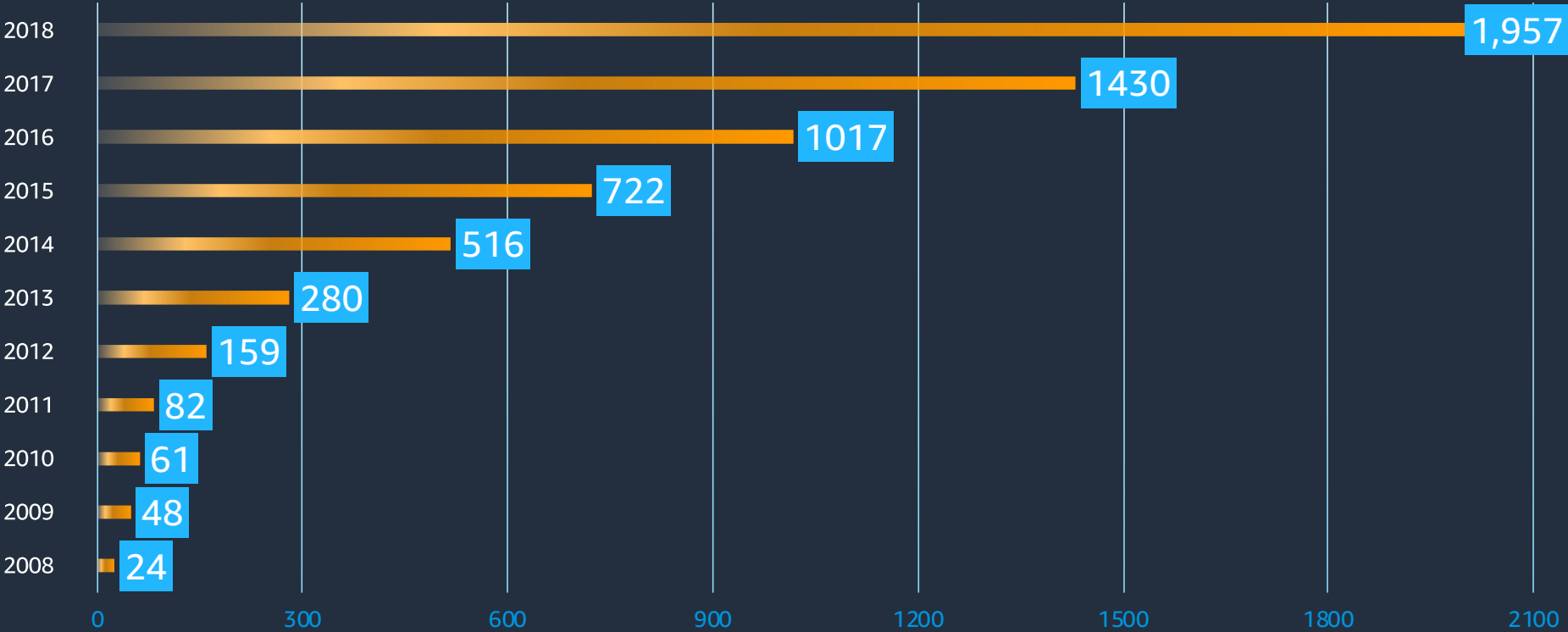


90%

of roadmap originates with customer requests
and are designed to meet specific needs

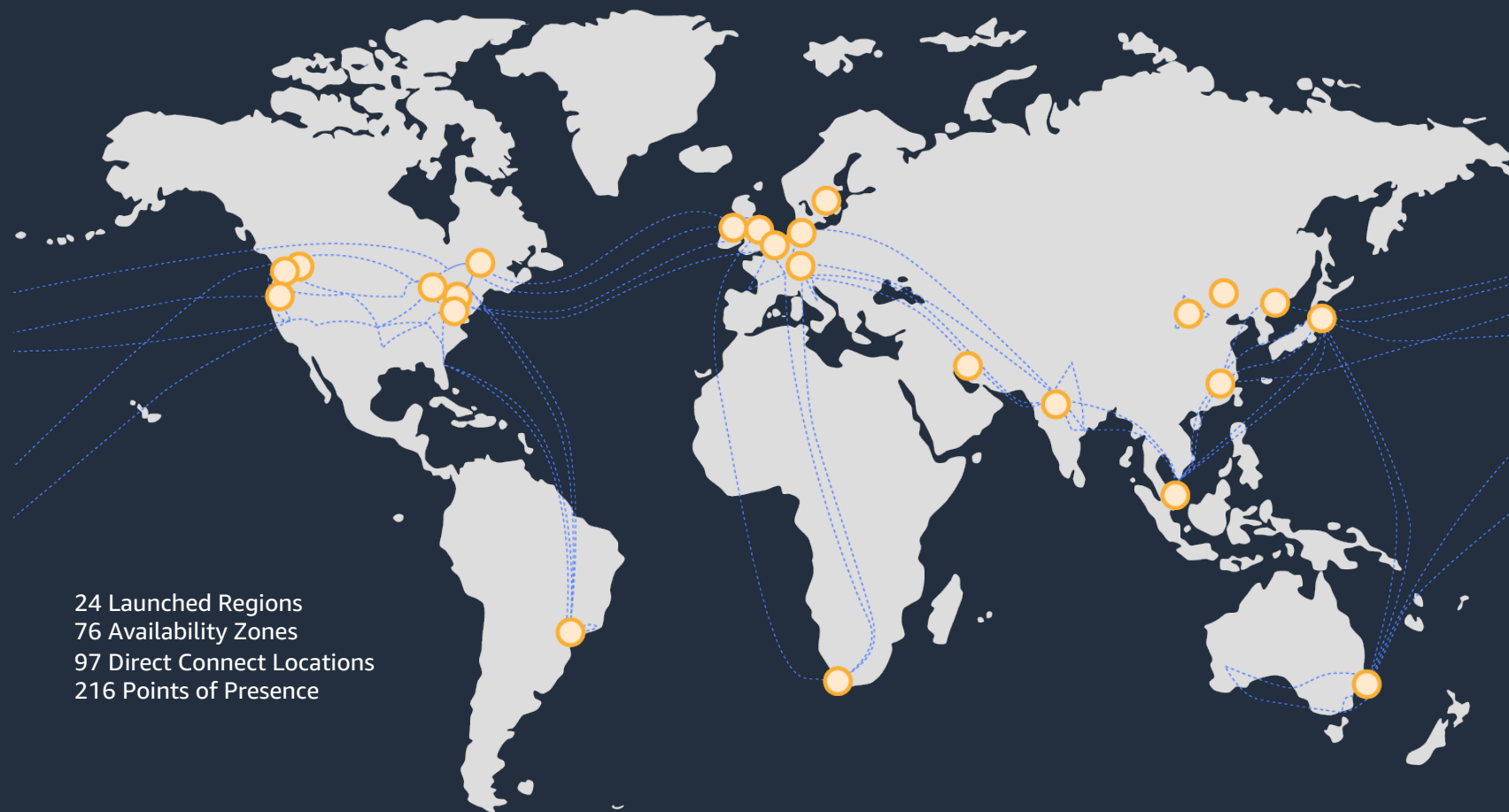
96% of R1 Research Institutions are using AWS, including Cornell University, Arizona State University, and the University of Notre Dame.

Pace of innovation | Launches



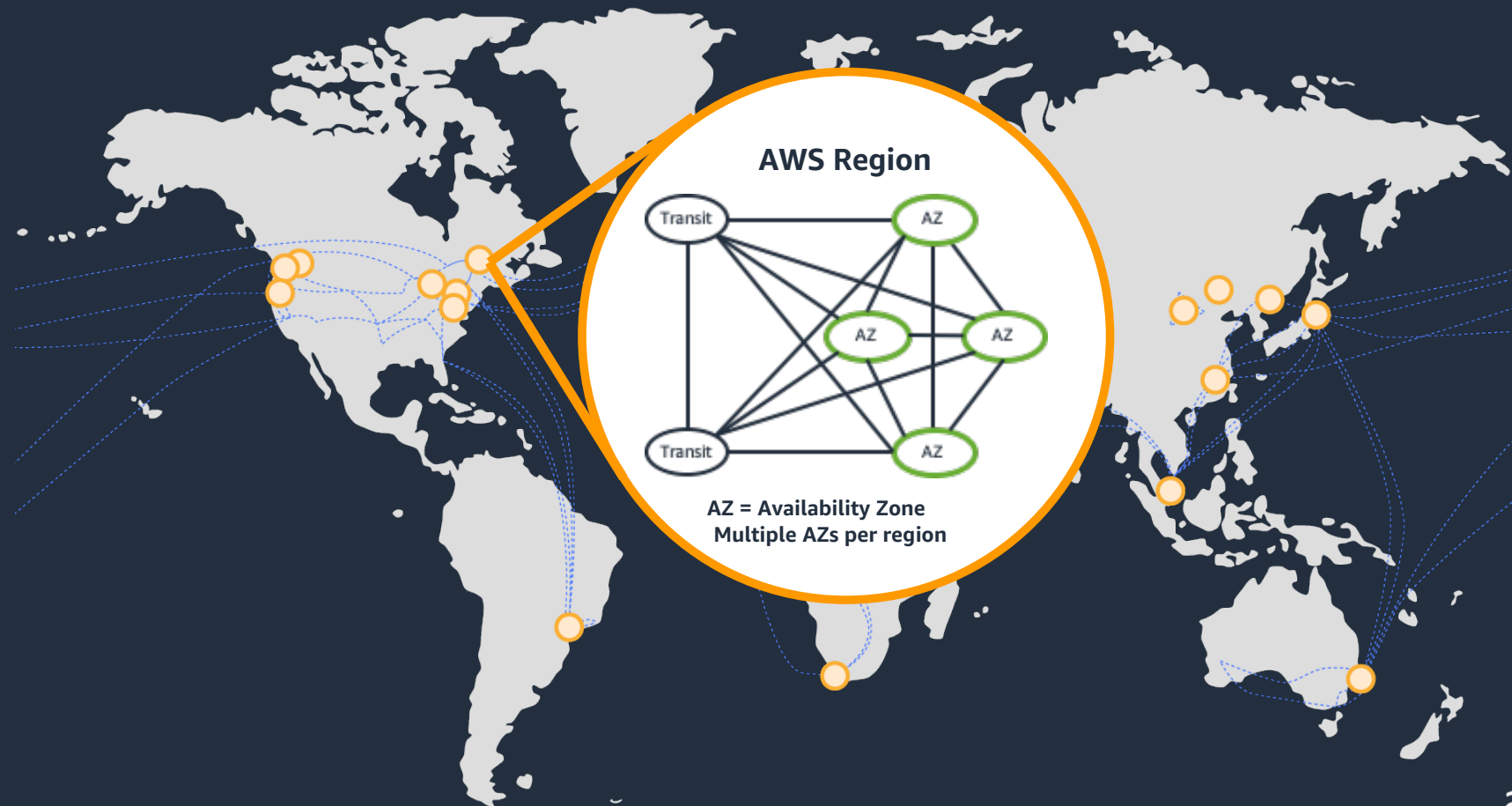
AWS Global Network

(AWS Regions shown)

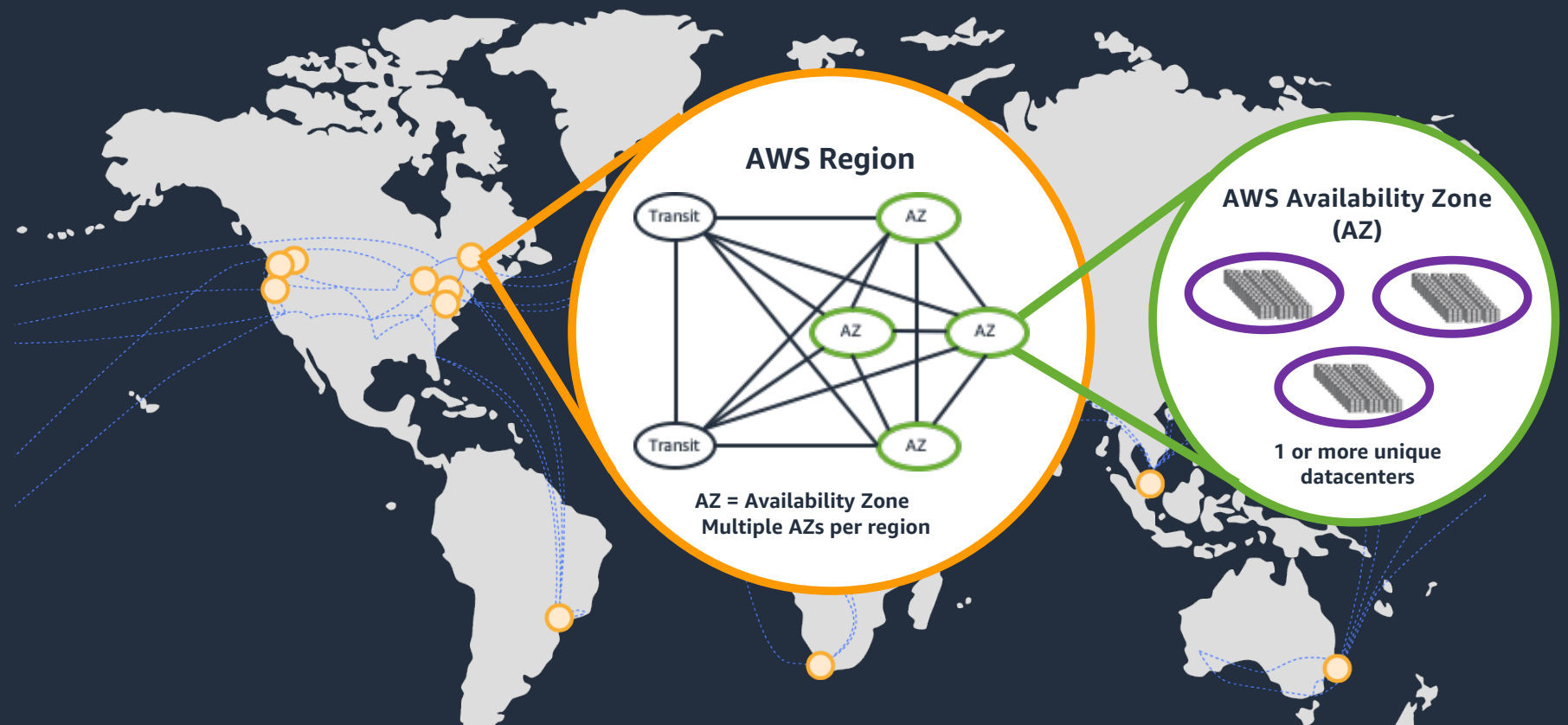


24 Launched Regions
76 Availability Zones
97 Direct Connect Locations
216 Points of Presence

AWS Regions



AWS Availability Zones



We add the equivalent of **an entire Fortune 500 company's** compute capacity **every day**

Broad and Deep Functionality

TECHNICAL & BUSINESS SUPPORT

- Support
- Professional Services
- Optimization Guidance
- Partner Ecosystem
- Training & Certification
- Solutions Management
- Account Management
- Security & Billing Reports
- Personalized Dashboard

MARKETPLACE

- Business Apps
- Business Intelligence
- DevOps Tools
- Security
- Networking
- Databases
- Storage

ANALYTICS

- Data Warehousing
- Business Intelligence
- Hadoop/Spark
- Streaming Data Analysis
- Streaming Data Collection

DEV OPS

- One-click App Deployment
- Resource Templates
- Build & Test
- Application Lifecycle Management
- DevOps Resource Management
- Triggers
- Containers
- Analyze & Debug
- Patching

MOBILE SERVICES

- API Gateway
- Single Integrated Console
- Identity
- Sync
- Mobile Analytics
- Mobile App Testing
- Targeted Push Notifications

IoT

- Rules Engine
- Device Shadows
- Device SDKs
- Device Gateway
- Registry
- Local Compute

MACHINE LEARNING

- Custom Model Training & Hosting
- Image & Scene Recognition
- Facial Recognition & Analysis
- Facial Search
- Text to Speech
- Conversational Chatbots
- Deep Learning (Apache MXNet, TensorFlow, & others)

ENTERPRISE APPS

- Virtual Desktops
- Sharing & Collaboration
- Corporate Email
- App Streaming
- Communications
- Contact Center

HYBRID ARCHITECTURE

- Data Integration
- Integrated Networking
- Integrated Identity & Access
- Integrated Resource & Deployment Management
- Integrated Devices & Edge Systems

MIGRATION

- Schema Conversion
- Exabyte-Scale Data Migration
- Application Migration
- Database Migration
- Server Migration

APP SERVICES

- Queueing & Notifications
- Email
- Workflow
- Transcoding
- Search

INFRASTRUCTURE

- Regions
- Availability Zones
- Points of Presence

CORE SERVICES

- Compute**
VMs, Auto-scaling, Load Balancing, Containers, Virtual Private Servers, Batch Computing, Cloud Functions, Elastic GPUs, Edge Computing
- Storage**
Object, Blocks, File, Archival, Import/Export, Exabyte-scale data transfer
- Databases**
Relational, NoSQL, Caching, Migration, PostgreSQL compatible
- Networking**
VPC, DX, DNS
- CDN**

SECURITY & COMPLIANCE

- Identity Management
- Access Control
- Monitoring & Logs
- Assessment & Reporting
- Web Application Firewall
- Configuration Compliance
- Key Management & Storage
- Account Grouping
- Resource & Usage Auditing
- DDOS Protection

MANAGEMENT TOOLS

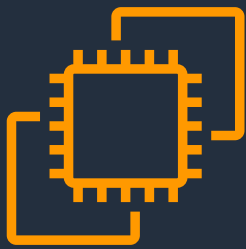
- Manage Resources
- Service Catalogue
- Configuration Tracking
- Monitoring
- Server Management
- Resource Templates

Core Services

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark



Compute platform options



Amazon EC2

Virtual server instances
in the cloud



Amazon ECS, EKS, and Fargate

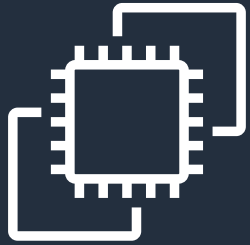
Container management
service for running
Docker on a managed
cluster of EC2



AWS Lambda

Serverless compute
for stateless code execution
in response to triggers

Amazon EC2



Amazon EC2

Linux | Windows

Arm and x86 architectures

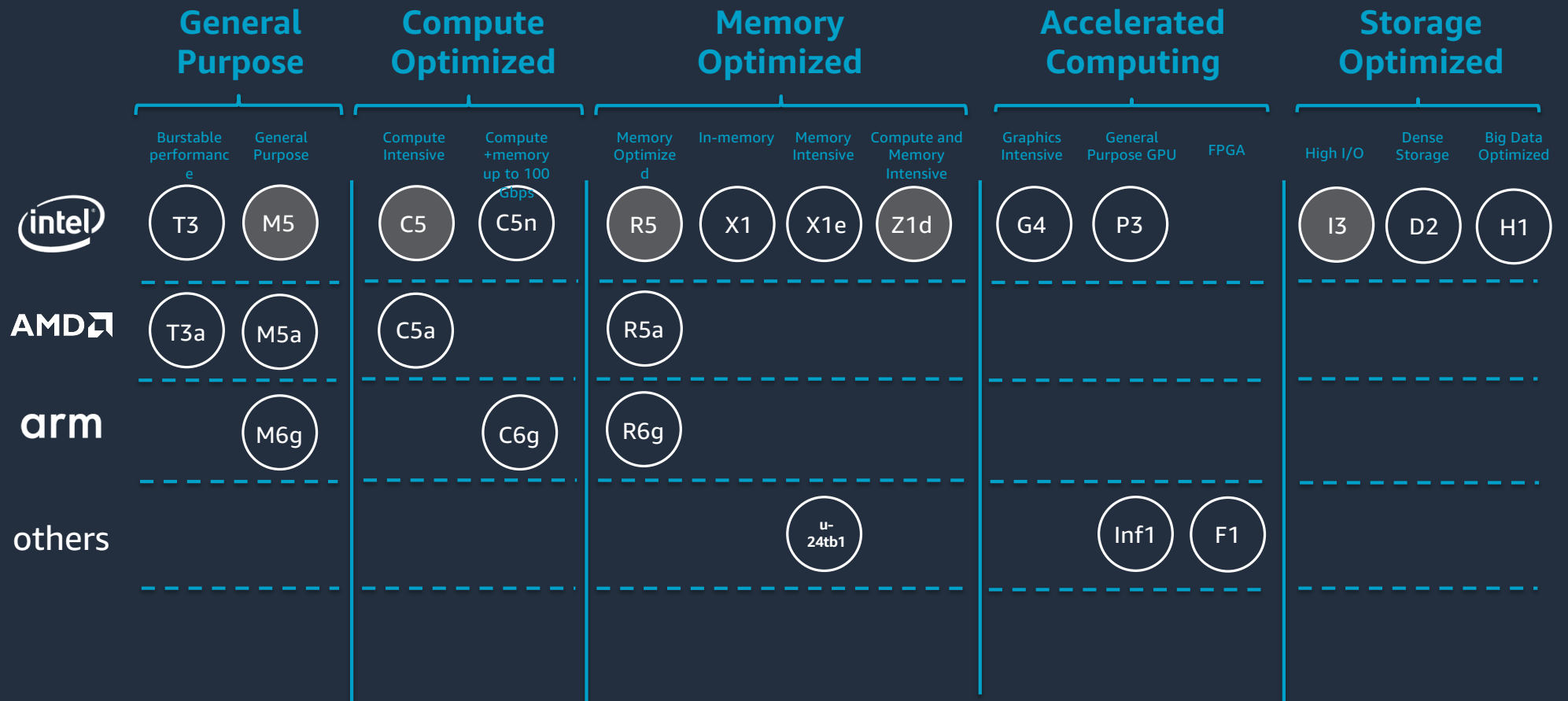
General purpose and workload optimized

Bare metal, disk, networking capabilities

Packaged | Custom | Community AMIs

Multiple purchase options: On-demand, RI, Spot

Instance Types



Children's Hospital of Philadelphia - FPGA

Fastest-Ever Analysis Of 1,000 Genomes

1,000 diverse pediatric genomes were processed into useable data files in two hours and twenty-five minutes



World Record for processing Genomes using FPGAs

Deployed on 1,000 Amazon EC2 F1 instances

One of the largest cohorts for this demographic that has been sequenced to date

Utilized Edico Genome's DRAGEN™ Genome Pipeline

Amazon EC2 purchase options

On-Demand

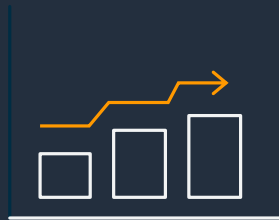
Pay-for-compute capacity **by the second** with no long-term commitments



Spiky workloads,
to define needs

Savings Plans & Reserved Instances

Make a commitment and receive a **significant discount** off compute



Committed &
steady-state usage

Spot Instances

Spare Amazon EC2 capacity at **savings of up to 90%** off On-Demand prices



Fault-tolerant, flexible,
stateless workloads

EC2 Operating Systems Supported

- Windows 2003R2/2008/2008R2/2012/2012R2/2016/2019
- Amazon Linux
- Debian
- Suse
- CentOS
- Red Hat Enterprise Linux
- Ubuntu



for more OSes see: <https://aws.amazon.com/marketplace/b/2649367011>

AWS container services landscape

Management

Deployment, Scheduling,
Scaling & Management of
containerized applications



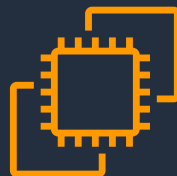
Amazon Elastic
Container Service
(ECS)



Amazon Elastic
Kubernetes Service
(EKS)

Hosting

Where the containers run



Amazon EC2



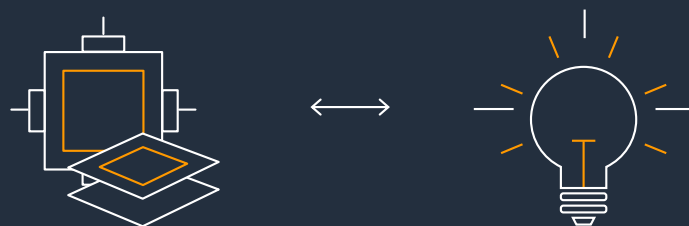
AWS Fargate

Image Registry

Container Image Repository



Amazon Elastic
Container Registry
(ECR)



HPC on AWS

Flexible configuration and virtually unlimited scalability
to grow and shrink your infrastructure as your HPC
workloads dictate, not the other way around

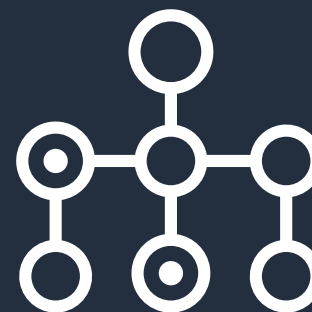
Cloud Native HPC Tools



AWS ParallelCluster

Simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers

Integrated with AWS Batch, Amazon FSx for Lustre and Elastic Fabric Adapter



AWS Batch

AWS Batch dynamically provisions resources, plans, schedules, and executes

No additional components to install

HPC Cloud Bursting



https://en.wikipedia.org/wiki/Cloudburst#/media/File:Cloudburst_on_phoenix.jpg



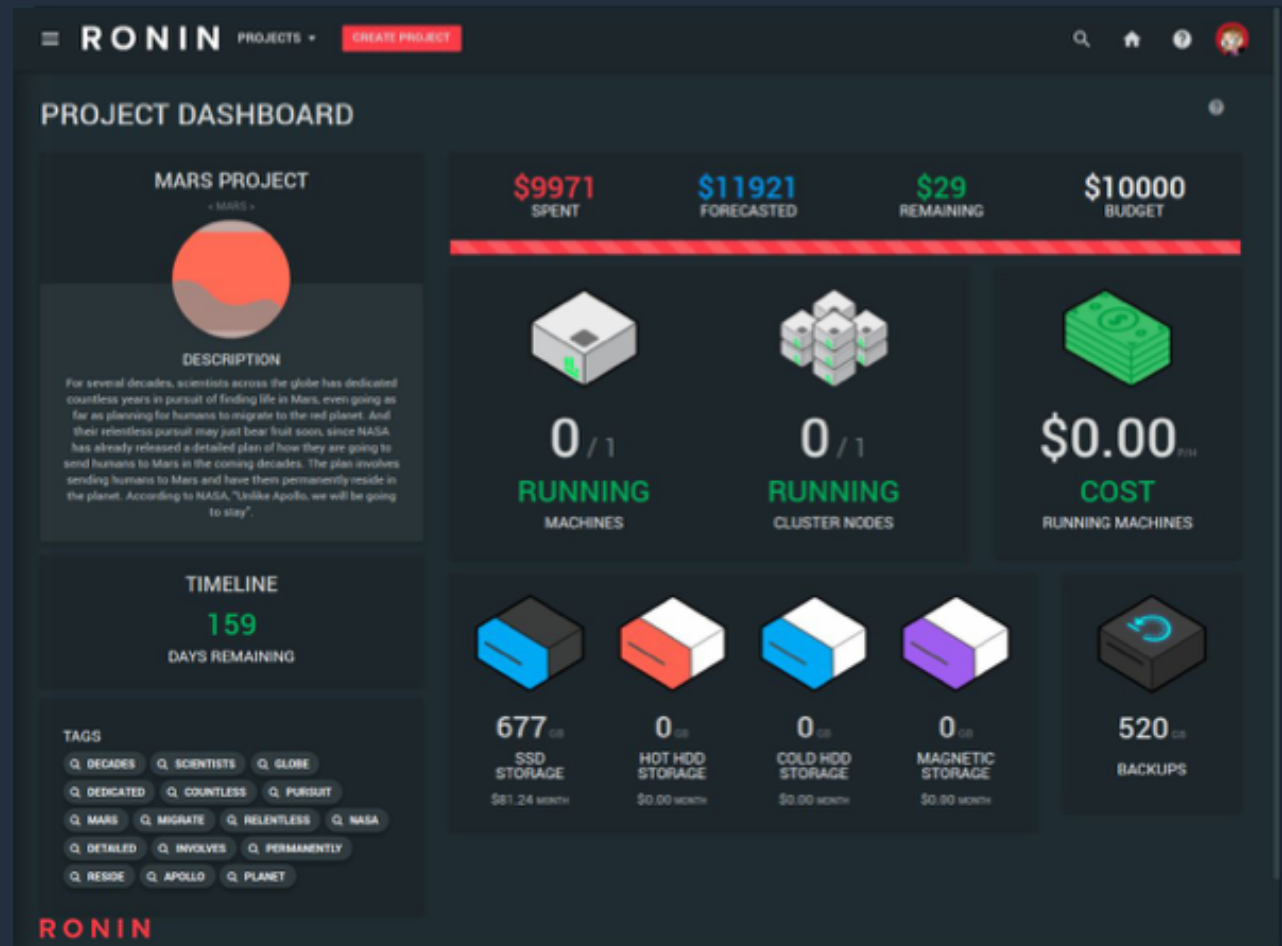
Use your existing on-premises scheduler to burst jobs to AWS to provide near infinite scale capacity for your research without changing a thing



RONIN

Easy to understand project views and budget management.

\$20K per year license matched by \$20K AWS credits



Storage



Amazon Elastic Block Store



Amazon Simple Storage Service (S3)



Amazon Elastic File System



Amazon FSx for Lustre



Storage



Amazon Elastic Block Store

- Network attached block device
- Independent data lifecycle
- Virtual disks
- Multiple volumes per EC2 instance
- Only one EC2 instance at a time per volume
- Can be moved from one instance to another
- POSIX-compliant file systems

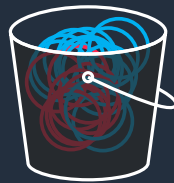
Amazon Simple Storage Service (S3)

- Object Store with limitless scalability
- Pay for exactly what you use
- Designed for 99.999999999% durability
- Several classes of storage to choose from depending on access patterns
- Query data directly from your buckets
- Supports versioning and MFA delete

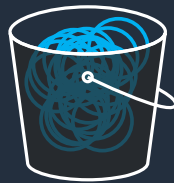
Cost Optimize Storage with S3 and Lifecycle policies



S3 Standard



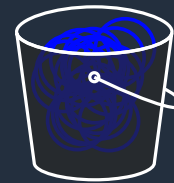
S3 Intelligent-Tiering



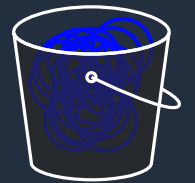
S3 Standard-IA



S3 One Zone-IA



S3 Glacier



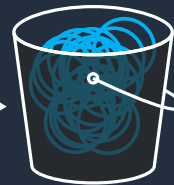
S3 Glacier Deep Archive

Frequent ← *Access Frequency* → *Infrequent*

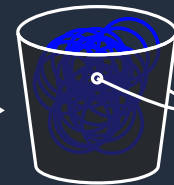
S3 Lifecycle Policy to tier to lower cost storage classes and expire storage



S3 Standard



S3 S-IA



S3 Glacier

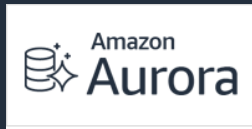
Database



Amazon RDS



Amazon Redshift



50% LESS EXPENSIVE THAN ALL OTHER CLOUD DATA WAREHOUSES

3X FASTER PERFORMANCE THAN ANY OTHER DATA WAREHOUSES

SUPPORT WORKLOADS UP TO **8PB** OF COMPRESSED DATA

Security

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

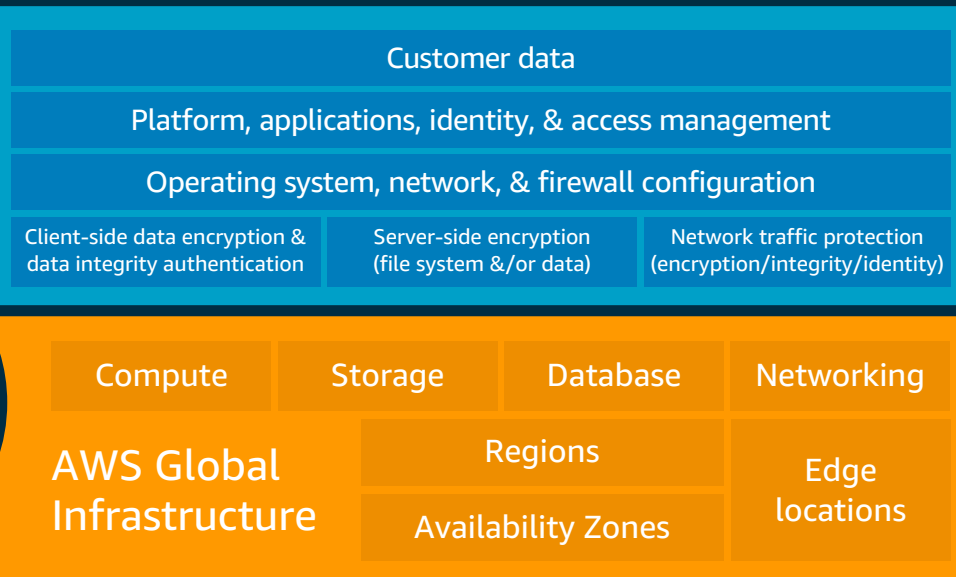


Share your **security responsibility** with AWS

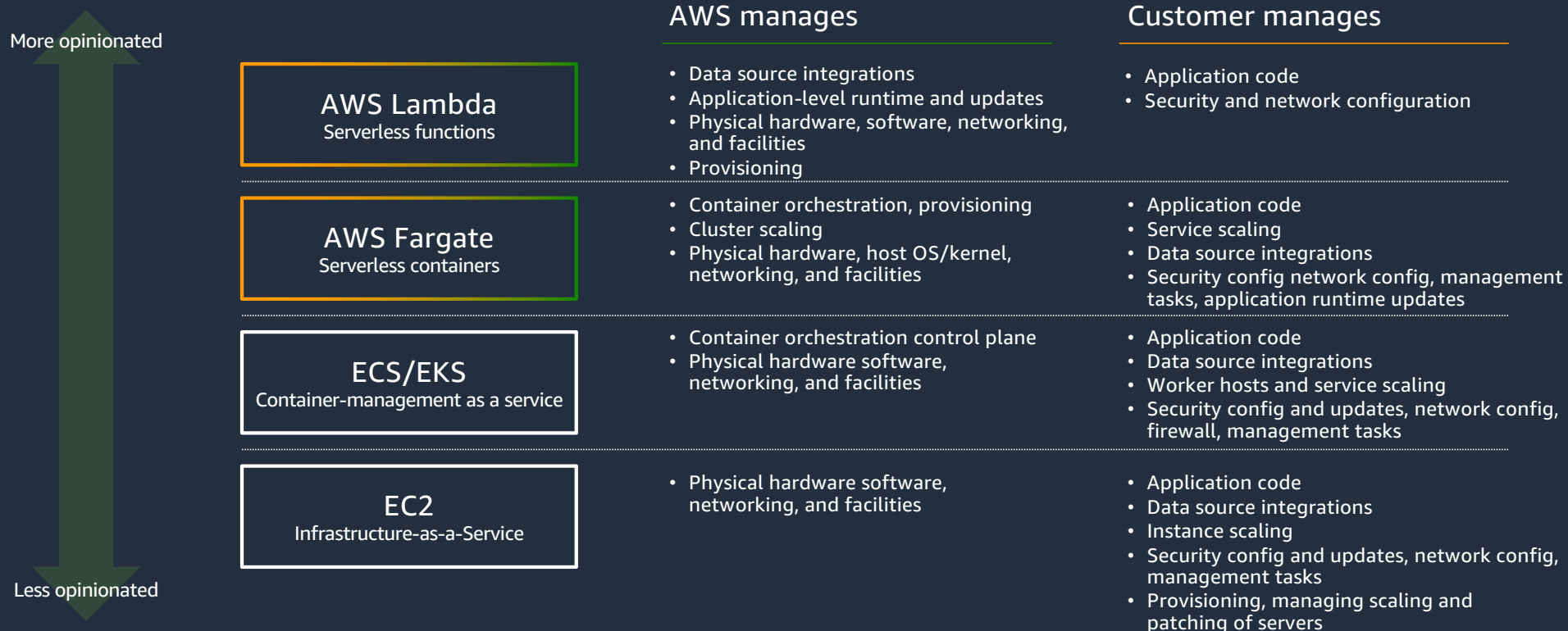
Customer is responsible for security **in** the cloud

Customer
AWS

AWS is responsible for security **of** the cloud



Shared operational responsibility model



Compliance programs

Global



Europe



Asia Pacific



United States



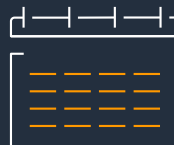
AWS is the first choice for highly regulated organizations

“ We can be far more secure in the cloud and achieve a higher level of assurance at a much lower cost, in terms of effort and dollars invested. We determined that security in AWS is superior to our on-premises data center across several dimensions, including patching, encryption, auditing and logging, entitlements, and compliance. ”

– John Brady, CISO, FINRA



Over 50 global compliance certifications and accreditations



AWS industry-leading security teams: 24/7, 365 days a year

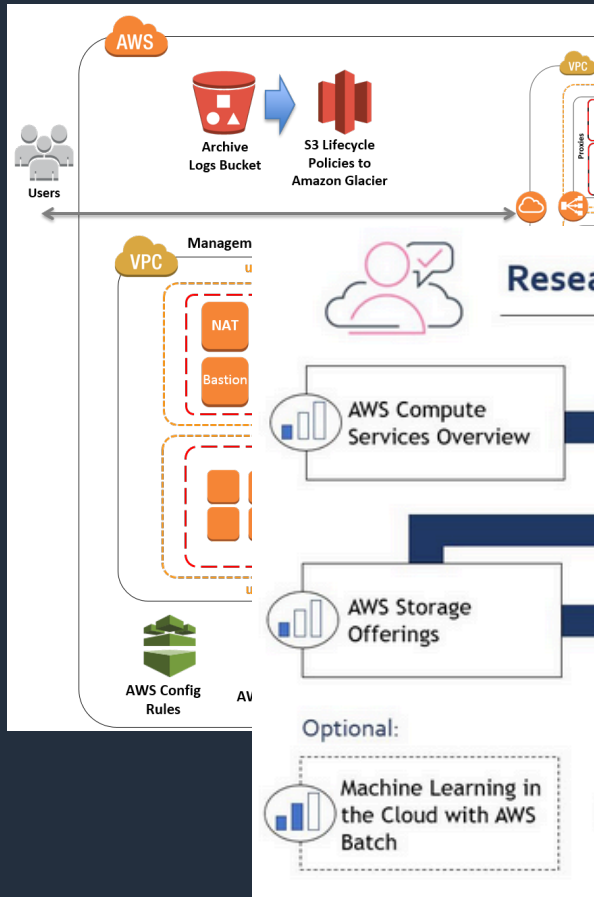


Security infrastructure built to satisfy military, global banks, and other high-sensitivity organizations



Security enhancements from 1M+ customer experiences

Quick Starts and Resources for Researchers



AWS Public Sector Blog

Resources for researchers and institutions to work remotely

by Sanjay Padhi, Ph.D | on 08 APR 2020 | in [Customer Solutions](#), [Education](#), [Healthcare](#), [Higher Education](#), [Nonprofit](#), [Public Sector](#), [Research](#), [Thought Leadership](#) | [Permalink](#) | [Comments](#) | [Share](#)

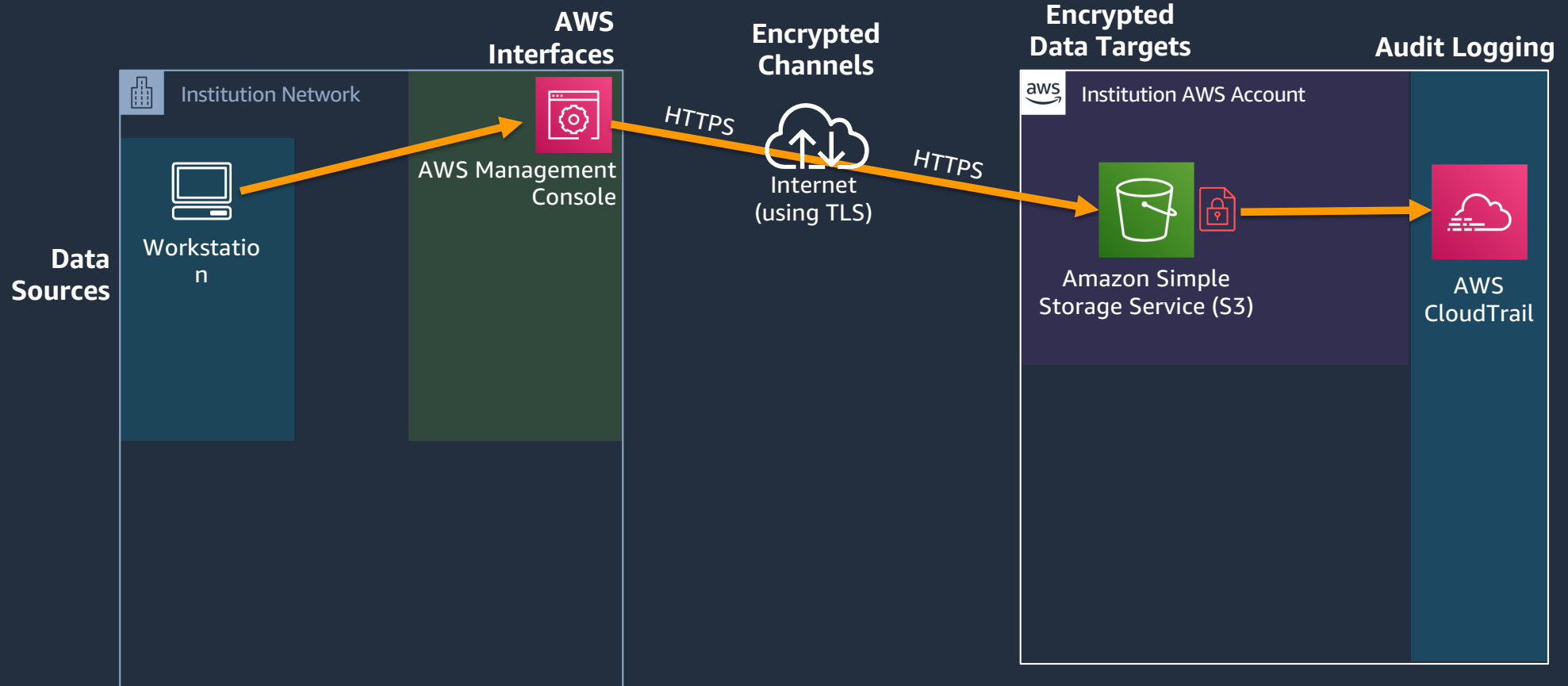


Regulated Data Transfer and Management

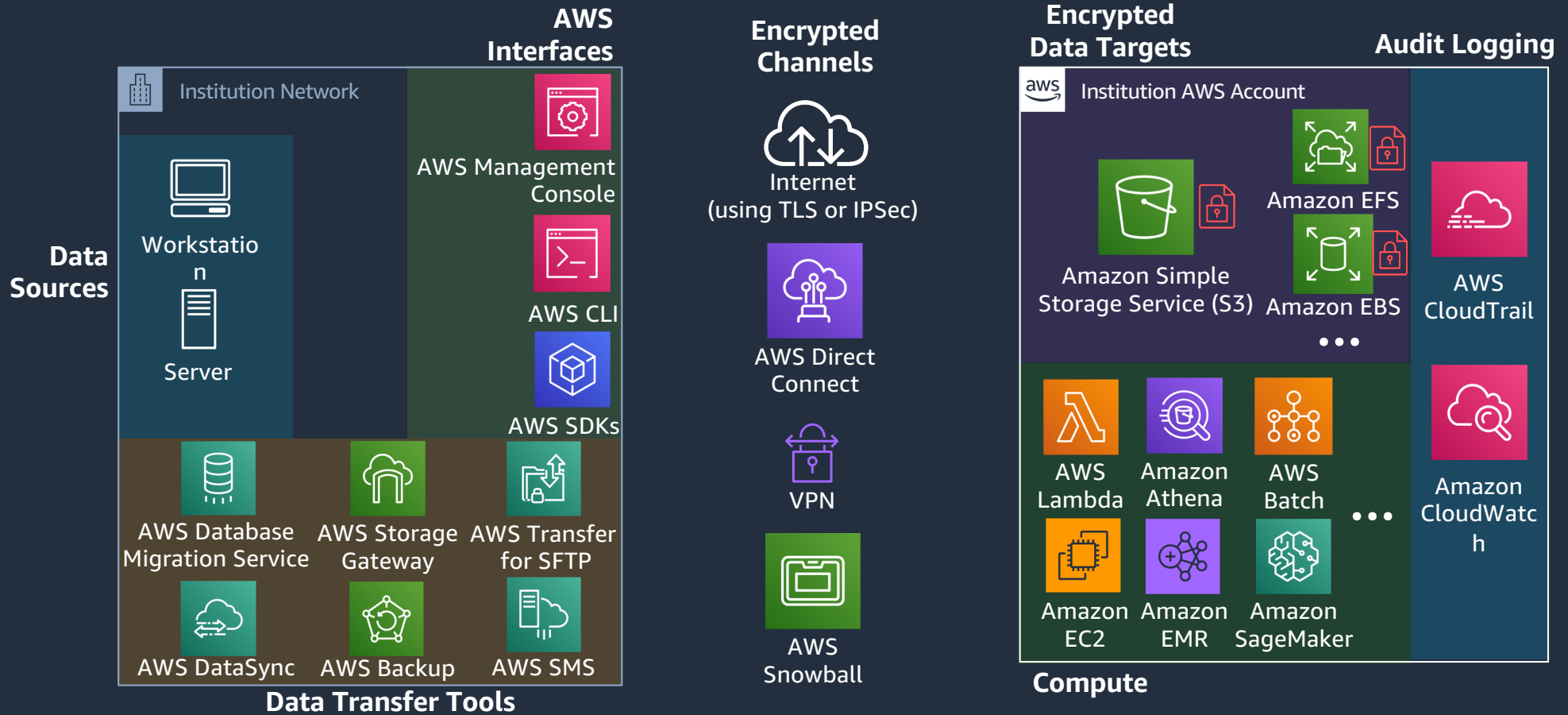
Regulated data transfer and management

- How to transfer data from your institution to AWS
- How to transfer data from AWS to your institution
- How to share data on AWS
- How to transfer data from other sources (i.e. dbGaP) to AWS

Transferring data to AWS (uploading to S3)



Transferring data to AWS



Transferring data to AWS (uploading to S3)

The screenshot shows the AWS Management Console interface. At the top, the browser address bar displays the URL `console.aws.amazon.com/console/home?region=us-east-1`. The navigation bar includes the AWS logo, a 'Services' dropdown menu, 'Resource Groups', and icons for 'EC2' and 'S3'. The user's profile 'admin/wiltpany-Isengard @ 76...' and the region 'N. Virginia' are also visible.

AWS Management Console

AWS services

Find Services
You can enter names, keywords or acronyms.

Recently visited services

- S3
- EC2
- Batch
- IAM
- AWS AppConfig

All services

Stay connected to your AWS resources on-the-go

Download the AWS Console Mobile App to your iOS or Android mobile device. [Learn more](#)

Explore AWS

Amazon SageMaker Resources
Learn about SageMaker's features, use cases, and available workshops. [Learn more](#)

Get Up to 40% Better Price Performance in Amazon EC2
Amazon EC2 M6g, C6g, and R6g instances provide the

Transferring data to AWS (uploading to S3)

S3 Management Console

s3.console.aws.amazon.com/s3/buckets/my-research-bucket/?region=us-east-1&tab=overview

aws Services Resource Groups EC2 S3

admin/wiltdany-Isengard @ 76... Global Support

my-research-bucket

Overview Properties Permissions Management Access points

Upload Create folder Download Actions

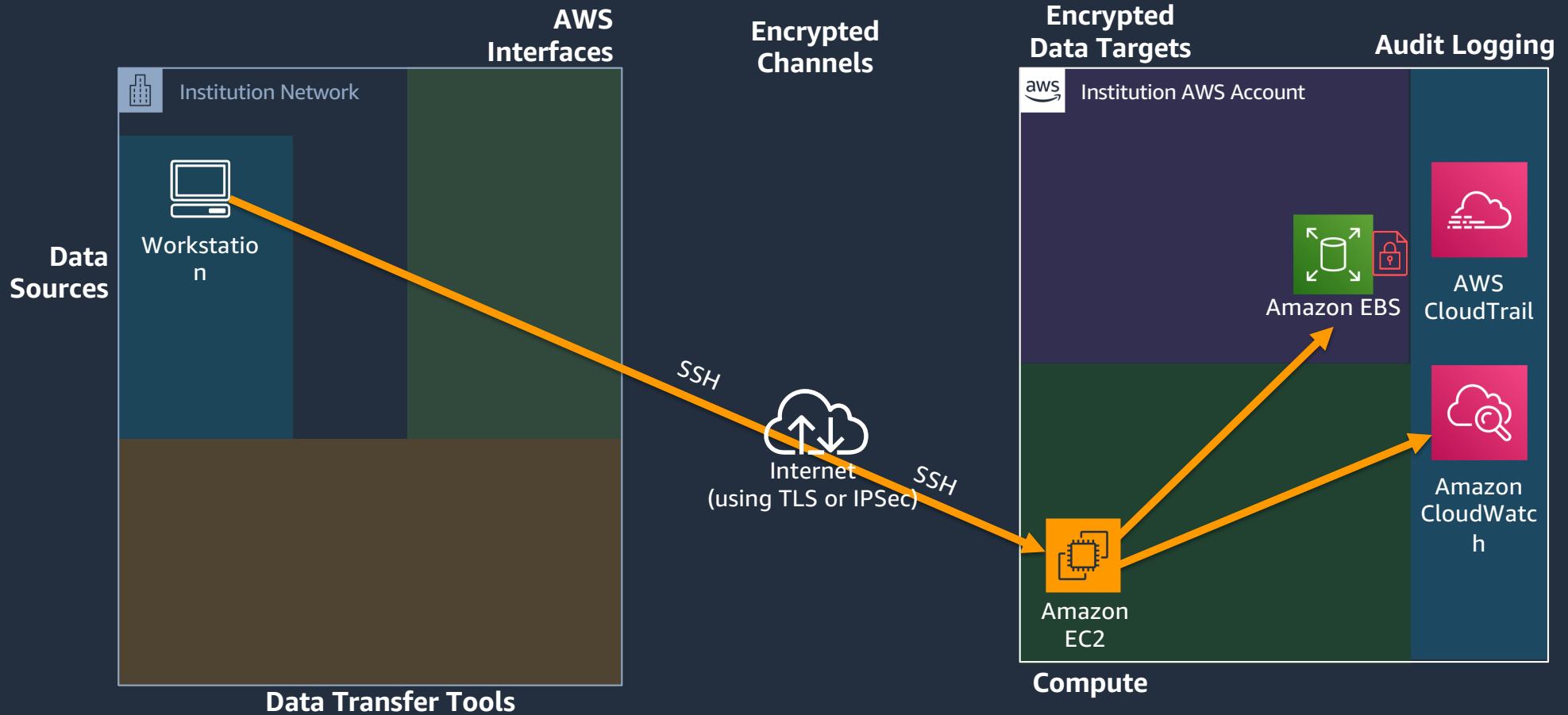
US East (N. Virginia)

This bucket is empty. Upload new objects to get started.

Upload an object Set object properties Set object permissions

Operations 0 In progress 1 Success 0 Error

Transferring data to AWS



Transferring data to AWS (scp data to EC2)

The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', 'EC2', 'S3', and user information. The left sidebar contains navigation options like 'EC2 Dashboard', 'Events', 'Tags', 'Limits', and 'Instances'. The main content area displays a table of EC2 instances. Two instances are listed: 'Linux' and 'Windows'. The 'Linux' instance is selected, and its details are shown below. The details include the instance ID, state (running), type (t2.micro), and public DNS name. The public DNS name is highlighted with a mouse cursor.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)
Linux	i-0eb19a11bc79eb0ea	t2.micro	us-east-1a	running	2/2 checks ...	None	ec2-3-232-131-85.compute-1.amazonaws.com
Windows	i-0fc6937dd37a170e9	t2.micro	us-east-1e	running	2/2 checks ...	None	ec2-3-90-65-11...

Instance: **i-0eb19a11bc79eb0ea (Linux)** Public DNS: **ec2-3-232-131-85.compute-1.amazonaws.com**

Property	Value	Property	Value
Instance ID	i-0eb19a11bc79eb0ea	Public DNS (IPv4)	ec2-3-232-131-85.compute-1.amazonaws.com
Instance state	running	IPv4 Public IP	3.232.131.85
Instance type	t2.micro	IPv6 IPs	-
Finding	Opt-in to AWS Compute Optimizer for recommendations. Learn more	Elastic IPs	-
Private DNS	ip-172-31-2-40.ec2.internal	Availability zone	us-east-1a
Private IPs	172.31.2.40	Security groups	launch-wizard-1 . view inbound rules . view outbound rules
Secondary private IPs	-	Scheduled events	No scheduled events
VPC ID	vpc-59739f24	AMI ID	amzn2-ami-hvm-2.0.20200617.0-

Transferring data to AWS (scp data to EC2)

The screenshot displays the AWS Management Console interface. At the top, a terminal window titled "research_data" shows the following commands and output:

```
ec2-user@ip-172-31-2-40:/research_data -- bash -- 117x38
[38f9d363ea61:research_data wiltany]$ ls
QIBA_Lung_Collection | com.amazonaws.amazonaws/session-manager | 0eb19a11bc79eb0ea?region=us-east-1
[38f9d363ea61:research_data wiltany]$
```

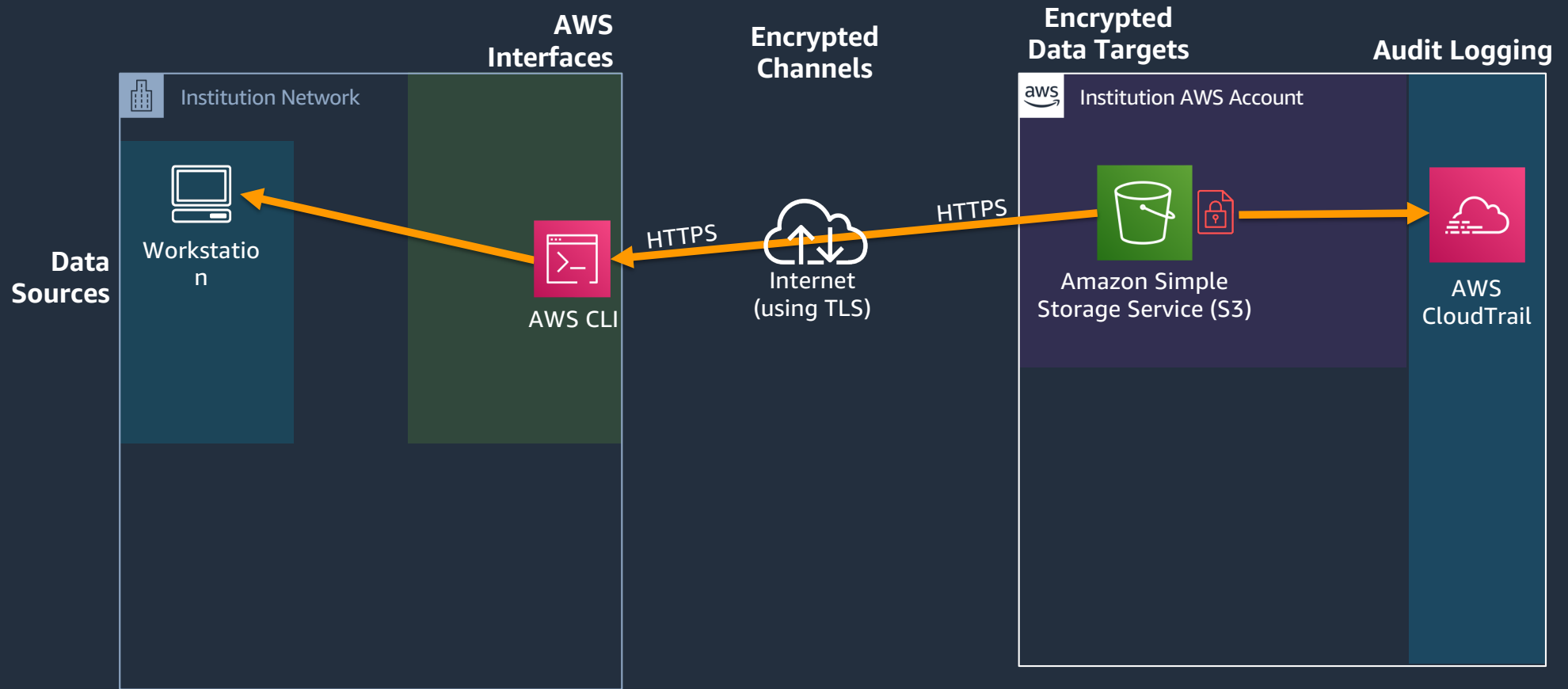
Below the terminal, the console shows "Session ID: wiltany-isengard-0332a6013779b2b21" and "Instance ID: i-0eb19a11bc79eb0ea". A "Terminate" button is visible on the right side of the terminal window.

In the foreground, a file explorer window titled "QIBA_Lung_Collection" is open. The left sidebar shows "Favorites" with "Amazon..." selected. The main area displays the following files and folders:

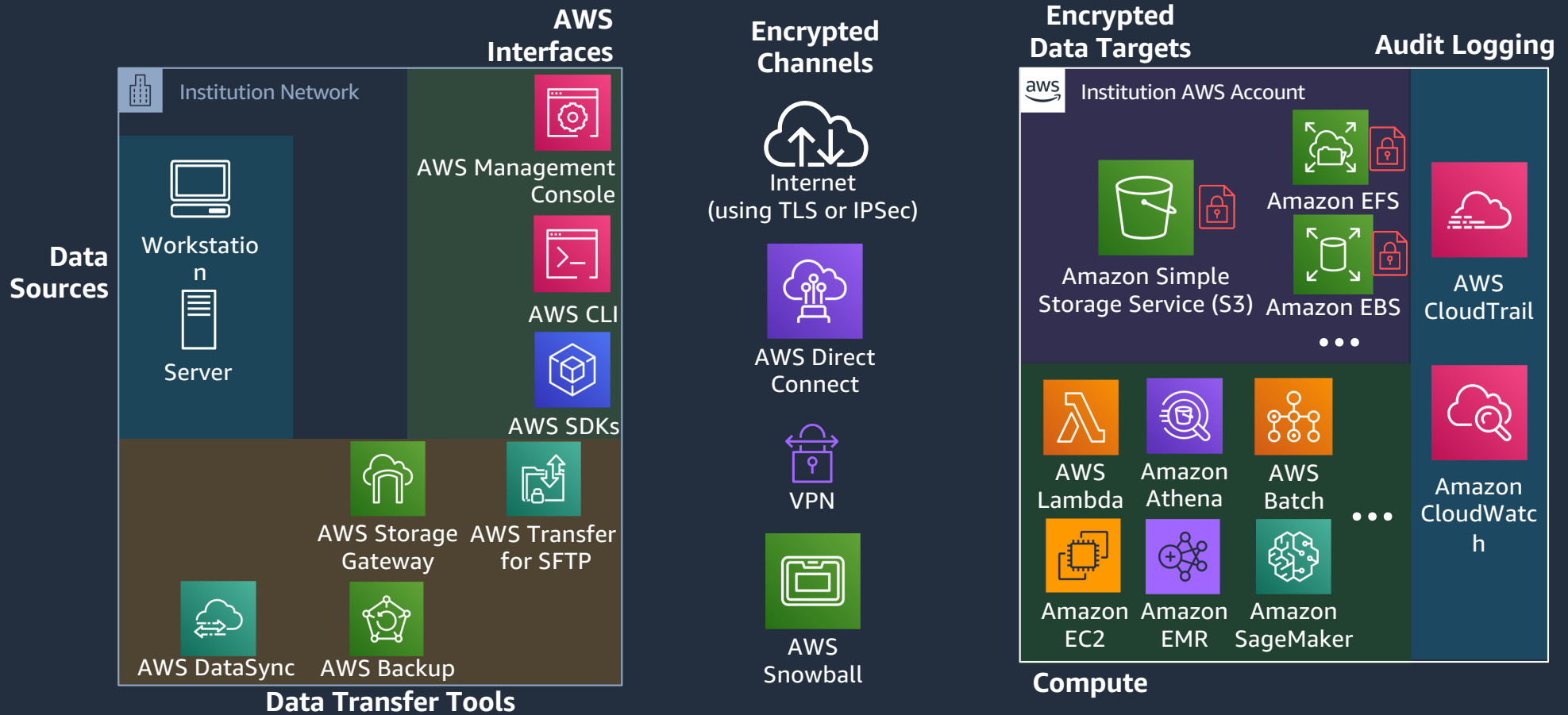
- Philips
- QIBA Volumetry CT - 4.5...ple.xlsx
- QIBA Volumetry CT - 4.5...424.xlsx
- Siemens

The file explorer also shows a search bar and navigation controls at the top.

Transferring data from AWS (downloading from S3)



Transferring data from AWS



Transferring data from AWS (downloading from S3)

The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', 'EC2', and 'S3'. The user is logged in as 'admin/wiltdany-Isengard @ 76...' in the 'N. Virginia' region. The main content area displays a list of EC2 instances with columns for Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, and Public DNS. Two instances are listed: a Linux instance and a Windows instance. The Windows instance (i-0fc6937dd37a170e9) is selected, and its details are shown below. The details include Instance ID, Instance state (running), Instance type (t2.micro), Finding (Opt-in to AWS Compute Optimizer), Private DNS (ip-172-31-61-58.ec2.internal), Private IPs (172.31.61.58), Secondary private IPs, VPC ID (vpc-59739f24), Public DNS (IPv4) (ec2-3-90-65-183.compute-1.amazonaws.com), IPv4 Public IP (3.90.65.183), IPv6 IPs (-), Elastic IPs, Availability zone (us-east-1e), Security groups (launch-wizard-2), and Scheduled events (No scheduled events). The AMI ID is Windows_Server-2019-English-Full-.

Instances | EC2 Management Console

console.aws.amazon.com/ec2/v2/home?region=us-east-1#Instances:sort=instanceld

Services Resource Groups EC2 S3

admin/wiltdany-Isengard @ 76... N. Virginia Support

New EC2 Experience Tell us what you think

EC2 Dashboard New

Events New

Tags

Limits

Instances

Instances

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts New

Scheduled Instances

Capacity Reservations

Images

Launch Instance Connect Actions

Filter by tags and attributes or search by keyword

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)
Linux	i-0eb19a11bc79eb0ea	t2.micro	us-east-1a	running	2/2 checks ...	None	ec2-3-232-131-...
Windows	i-0fc6937dd37a170e9	t2.micro	us-east-1e	running	2/2 checks ...	None	ec2-3-90-65-183...

Instance: i-0fc6937dd37a170e9 (Windows) Public DNS: ec2-3-90-65-183.compute-1.amazonaws.com

Description Status Checks Monitoring Tags

Instance ID	i-0fc6937dd37a170e9	Public DNS (IPv4)	ec2-3-90-65-183.compute-1.amazonaws.com
Instance state	running	IPv4 Public IP	3.90.65.183
Instance type	t2.micro	IPv6 IPs	-
Finding	Opt-in to AWS Compute Optimizer for recommendations. Learn more	Elastic IPs	
Private DNS	ip-172-31-61-58.ec2.internal	Availability zone	us-east-1e
Private IPs	172.31.61.58	Security groups	launch-wizard-2. view inbound rules. view outbound rules
Secondary private IPs		Scheduled events	No scheduled events
VPC ID	vpc-59739f24	AMI ID	Windows_Server-2019-English-Full-...

Feedback English (US)

© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

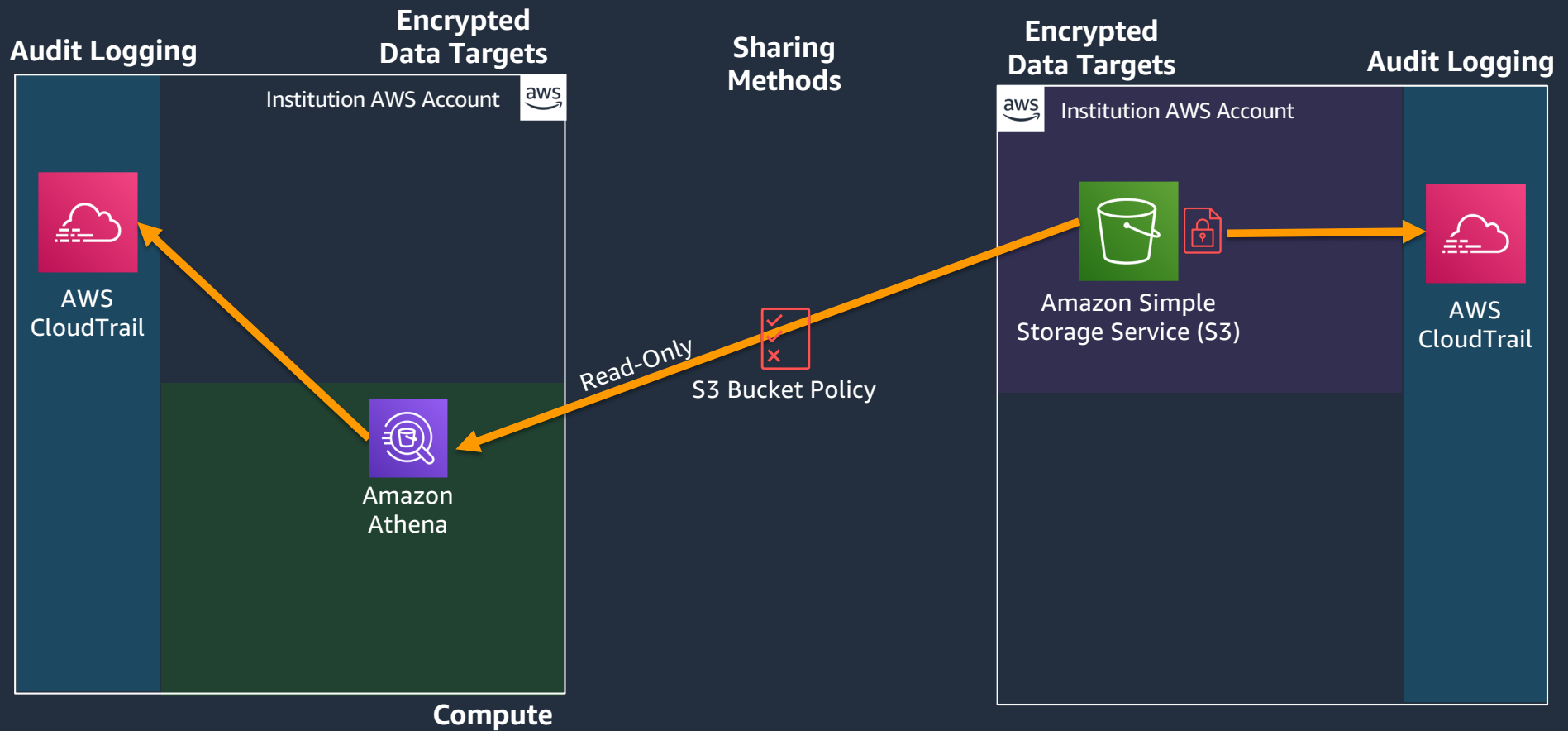
Transferring data from AWS (downloading from S3)

The image shows a composite screenshot of a computer screen. On the left is the AWS Management Console 'Instances' page. In the center is a Windows PowerShell terminal window titled 'My research PC' with the command prompt 'PS C:\research_data>'. Overlaid on the PowerShell window is a File Explorer window showing an empty 'research_data' folder. On the right is a table of EC2 instances with columns for 'Alarm Status' and 'Public DNS (I...'. The table contains two rows of data.

Alarm Status	Public DNS (I...
none	ec2-3-232-131
none	ec2-3-90-65-11

Additional visible text includes 'New EC2 Experience Tell us what you think', 'EC2 Dashboard New', 'Events New', 'Tags', 'Limits', 'Instances', 'Instance Types', 'Launch Templates', 'Spot Requests', 'Savings Plans', 'Reserved Instances', 'Dedicated Hosts New', 'Scheduled Instances', 'Capacity Reservations', 'Images', 'Feedback', 'English', '6:30 PM 7/11/2020', 'Privacy Policy', and 'Terms of Use'.

Sharing data on AWS (S3 read-only access)



Sharing data on AWS



Sharing data on AWS (s3 read-only access)

The screenshot shows the AWS S3 Management Console interface for a bucket named 'my-research-bucket'. The breadcrumb navigation shows 'Amazon S3 > my-research-bucket'. The bucket name 'my-research-bucket' is displayed prominently. Below the bucket name are tabs for 'Overview', 'Properties', 'Permissions', 'Management', and 'Access points'. A search bar is present with the placeholder text 'Type a prefix and press Enter to search. Press ESC to clear.' Below the search bar are buttons for 'Upload', 'Create folder', 'Download', and 'Actions'. The region is set to 'US East (N. Virginia)'. A table lists the contents of the bucket, showing folders and files with their names, last modified dates, sizes, and storage classes.

<input type="checkbox"/>	Name ▾	Last modified ▾	Size ▾	Storage class ▾
<input type="checkbox"/>	dicom_images	--	--	--
<input type="checkbox"/>	my_data	--	--	--
<input type="checkbox"/>	QIBA Volumetry CT - 4.5 Tumor volume bias and linearity - PT20180...	Jul 11, 2020 10:47:57 AM GMT-0700	38.7 KB	Standard
<input type="checkbox"/>	QIBA Volumetry CT - 4.5 Tumor volume bias and linearity - PT20180...	Jul 11, 2020 10:47:57 AM GMT-	36.2 KB	Standard

Sharing data on AWS (s3 read-only access)

The screenshot shows the AWS S3 Management Console interface. At the top, there's a navigation bar with the AWS logo, 'Services', 'Resource Groups', and user information. Below this is a search bar for buckets and a dropdown for 'All access types'. The main content area displays a list of buckets with columns for selection, bucket name, public access status, region, and creation date. A summary bar at the top of the list indicates 13 buckets and 3 regions.

	Bucket Name	Public Access Status	Region	Creation Date
<input type="checkbox"/>	aws-athena-query-results-742018752460-us-west-2	Objects can be public	US West (Oregon)	Sep 9, 2019 4:17:28 PM GMT-0700
<input type="checkbox"/>	cf-templates-pzwwvrpvbxi-us-east-2	Objects can be public	US East (Ohio)	Jun 9, 2020 6:16:29 PM GMT-0700
<input type="checkbox"/>	cloudtrail-awslogs-742018752460-cwwcjd5g-isengard-...	Objects can be public	US East (N. Virginia)	Aug 19, 2019 1:55:09 PM GMT-0700
<input type="checkbox"/>	connect-8471571e3925	Objects can be public	US East (N. Virginia)	May 20, 2020 12:54:34 PM GMT-0700
<input type="checkbox"/>	danyellbucket	Bucket and objects not public	US East (N. Virginia)	Sep 20, 2019 4:23:49 PM GMT-0700
<input type="checkbox"/>	do-not-delete-gatedgarden-audit-742018752460	Objects can be public	US West (Oregon)	Aug 19, 2019 2:11:30 PM GMT-0700

Sharing data on AWS (s3 read-only access)

The screenshot shows the AWS S3 Management Console interface. At the top, there's a navigation bar with the AWS logo, 'Services', 'Resource Groups', and icons for EC2 and S3. The user is logged in as 'admin/wilt-dany-Isengard @ 76...'. A notification banner at the top states: 'We've temporarily re-enabled the previous version of the S3 console while we continue to improve the new S3 console experience. [Switch to the new console.](#)'

The main content area is titled 'S3 buckets'. It includes a search bar 'Search for buckets' and a dropdown for 'All access types'. Below this are buttons for '+ Create bucket', 'Edit public access settings', 'Empty', and 'Delete'. Summary statistics show '4 Buckets' and '2 Regions'. A table lists the buckets:

<input type="checkbox"/>	Bucket name	Access	Region	Date created
<input type="checkbox"/>	cloudtrail-awslogs-767145282866-nfv0rso4-isengard-d...	Bucket and objects not public	US East (N. Virginia)	Apr 20, 2020 11:25:20 AM GMT-0700
<input type="checkbox"/>	do-not-delete-gatedgarden-audit-767145282866	Bucket and objects not public	US West (Oregon)	Apr 20, 2020 11:40:13 AM GMT-0700
<input type="checkbox"/>	elasticbeanstalk-us-east-1-767145282866	Bucket and objects not public	US East (N. Virginia)	Jul 8, 2020 3:43:48 PM GMT-0700
<input type="checkbox"/>	<u>my-research-bucket</u>	Bucket and objects not public	US East (N. Virginia)	Jul 11, 2020 10:43:10 AM GMT-0700

Transfer Other Data to AWS (like dbGaP)

Considerations

- Many research data sets are available through AWS.
- Understand the compliance requirements of the data you are transferring.
- In AWS, through shared responsibility, you can achieve compliance with standards like FedRAMP or FISMA.
- Work with your internal compliance team and your IRB.

Public Research Data Sets

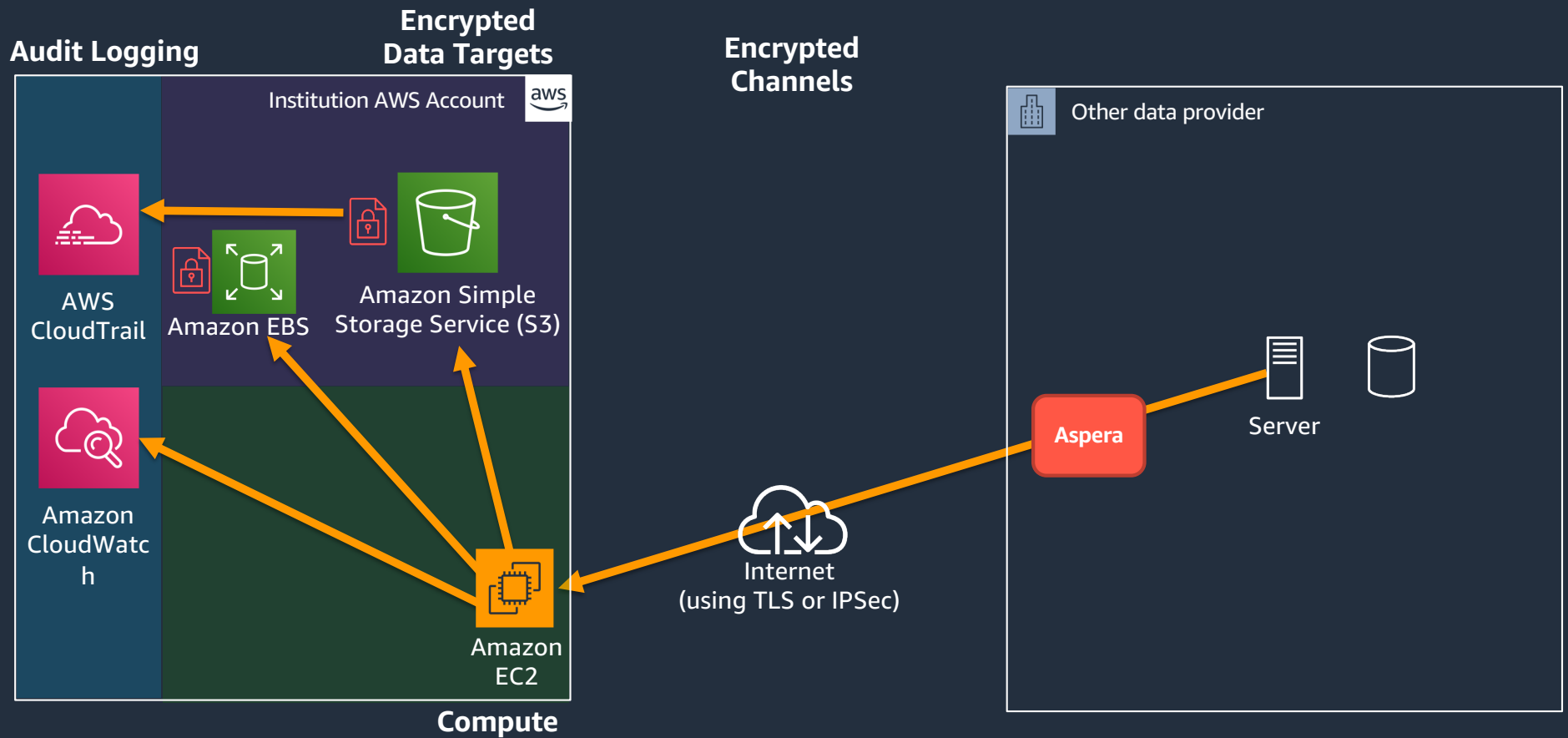


AWS hosts a variety of public datasets that anyone can access for free. Below are just a few examples.

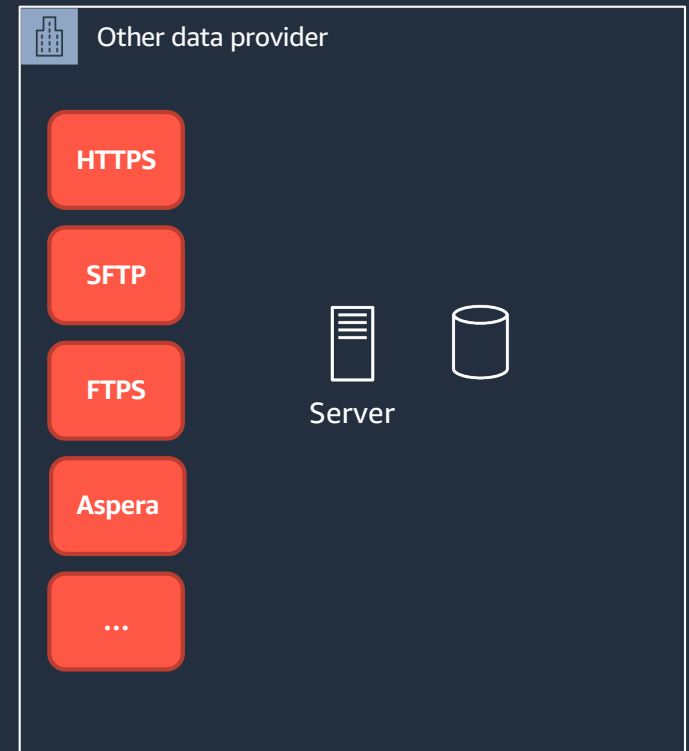
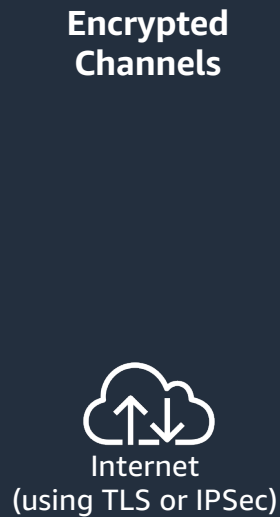
- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- 3000 Rice Genome
- Genome in a Bottle (GIAB)
- The Genome Modeling System
- Medicare Drug Spending
- The Human Connectome Project
- The Human Microbiome Project
- OpenNeuro
- Physionet
- Tabula muris
- OpenStreetMaps
- and more....

<https://registry.opendata.aws/>

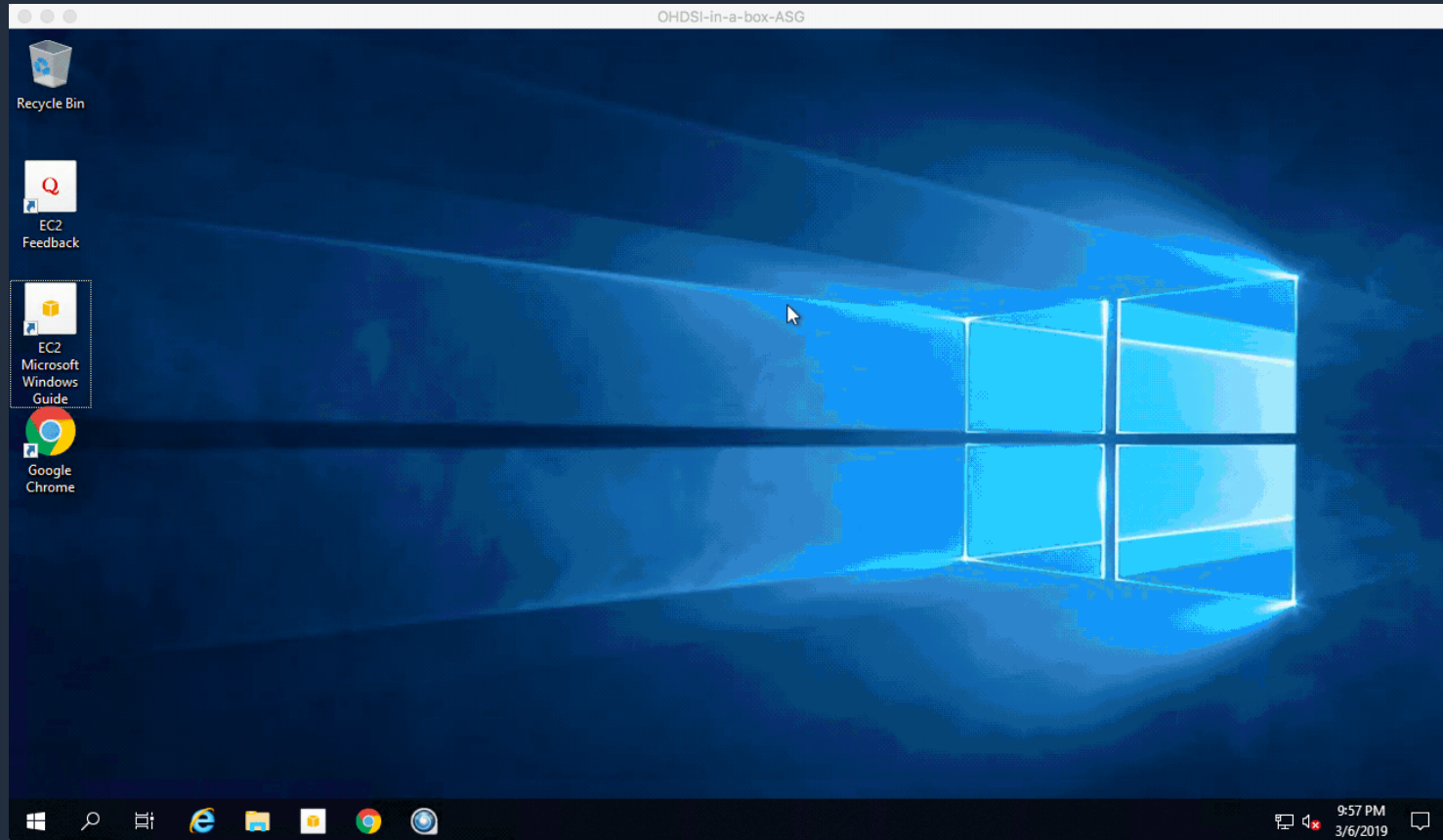
Transfer Other Data to AWS (like dbGaP)



Transfer Other Data to AWS (like dbGaP)



Transferring data to AWS (downloading from S3)



RStudio on AWS


© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark



RStudio on AWS

Introduction to RStudio on AWS

This repository provides simple introduction to R

[Launch Stack](#) 

Instructions

1. Launch the AWS CloudFormation in your AWS account using the **Launch Stack** button above.
 - For the **VPCId** parameter, use your Default VPC (172.31.0.0/16)
 - For the **VPCSubnet** parameter, choose a subnet within the Default VPC (172.31.0.0/20)
2. Once the stack says **CREATE_COMPLETE**, it takes about 5 additional minutes for the RStudio Server to become available.
3. After 5 minutes, follow the link in the **Outputs** tab of your AWS CloudFormation Stack to access RStudio.
4. Accept the warning from your browser about the certificate being self-signed. This gives us encrypted, HTTPS access to RStudio without purchasing a domain name or SSL certificate.
5. Login to RStudio using the credentials you provided to the AWS CloudFormation template, and **click the Terminal tab**.
6. Run the command `git clone https://github.com/JamesSwiggins/R-course`
7. Open the file **R-course/Basic Data Analysis/cluster_code.R**
8. Run each line of the R program
9. Open the **results** directory and view the images output by the analysis we just ran. A description of the analysis and dataset is below.



RStudio on AWS

The screenshot shows the AWS CloudFormation console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and icons for EC2 and S3. The user is logged in as 'admin/wiltany-Isengard @ 76...' in the 'N. Virginia' region. The breadcrumb trail is 'CloudFormation > Stacks > RStudio'. On the left, a 'Stacks (1)' list shows the 'RStudio' stack with a status of 'CREATE_COMPLETE' and a refresh button. The main panel displays the 'RStudio' stack details under the 'Stack info' tab. At the top of this panel are buttons for 'Delete', 'Update', 'Stack actions', and 'Create stack'. Below these are tabs for 'Stack info', 'Events', 'Resources', 'Outputs', 'Parameters', 'Template', and 'Change sets'. The 'Overview' section contains a table with the following data:

Property	Value
Stack ID	arn:aws:cloudformation:us-east-1:767145282866:stack/RStudio/d6e867f0-c3ba-11ea-8cce-0acaf3694d17
Description	This CloudFormation Template deploys an RStudio Server instance on AWS with a self-signed SSL certificate.
Status	CREATE_COMPLETE
Status reason	-
Root stack	-
Parent stack	-
Created time	2020-07-11 14:09:42 UTC-0700
Deleted time	-

RStudio on AWS

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for plotting PCA scores by region. The code uses `points()` to plot scores for South Asia, Western Europe, and Central Sub-Saharan Africa, with different colors and point shapes. It also includes `dev.off()`, `png()`, and `hclust()` for clustering.
- Console:** Shows the execution of the code from the source editor, with the same `points()` and `hclust()` commands visible.
- Environment Pane:** Lists objects in the Global Environment:
 - `pca`: List of 7
 - `perc_dt`: 25935 obs. of 8 variables
 - `raw_dt`: 77805 obs. of 16 variables
 - `raw_location`: 77805 obs. of 26 variables
 - `regions`: List of 21
- Files Pane:** Shows a directory structure:
 - Home > R-course > Basic Data Analysis
 - ..
 - cluster_code.ipynb (179.3 KB, Jul 11, 2020, 2:21 PM)
 - cluster_code.R (2.8 KB, Jul 11, 2020, 2:21 PM)
 - data
 - data_preprocessing.R (1.5 KB, Jul 11, 2020, 2:21 PM)
 - results



Jupyter Notebooks on AWS

Amazon SageMaker Studio

Fully integrated development environment (IDE) for machine learning



Collaboration at scale

Share notebooks without tracking code dependencies



Easy experiment management

Organize, track, and compare thousands of experiments



Automatic model generation

Get accurate models with full visibility & control without writing code



Higher quality ML models

Automatically debug errors, monitor models, & maintain high quality



Increased productivity

Code, build, train, deploy, & monitor in a unified visual interface

Amazon SageMaker Notebooks

Fast-start sharable notebooks (in preview)



Easy access with
Single Sign-On (SSO)

Access your notebooks in
seconds



Fully managed
and secure

Administrators manage
access and permissions



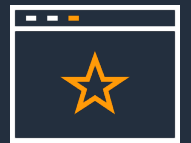
Fast setup

Start your notebooks
without spinning up
compute resources



Easy collaboration

Share notebooks
with a single click



Flexible

Dial up or down
compute resources
(coming soon)

Jupyter Notebooks in SageMaker Studio

The screenshot displays the Amazon SageMaker Studio Control Panel. At the top, the AWS logo and navigation menu are visible, including 'Services', 'Resource Groups', and 'Athena'. The user is logged in as 'admin/wiltdany-Isengard @ 76...' in the 'Ohio' region. The left sidebar shows the 'Amazon SageMaker Studio' section with options like 'Dashboard', 'Search', 'Ground Truth', 'Notebook', and 'Training'. The main content area is titled 'Amazon SageMaker Studio Control Panel' and contains a section for choosing a user to get started. Below this is a table of users and a 'Studio Summary' section.

Amazon SageMaker > Amazon SageMaker Studio > Control Panel

Amazon SageMaker Studio Control Panel

Choose your user name, then choose Open Studio to get started Add user

 < 1 > ⚙️

Jupyter Notebooks in SageMaker Studio

The screenshot displays the Amazon SageMaker Studio interface. On the left, there is a sidebar with sections for 'TERMINAL SESSIONS', 'KERNEL SESSIONS', and 'RUNNING IMAGES'. The main area shows a Jupyter Notebook with the following content:

Launcher | AWS_workshop_python_mod

2 vCPU + 4 GiB Python 3 (Base Python) Share

UniDip is a small python package used here for quickly determining estimated modality of the gene expressions.

Lifelines is a small python package that extends Pandas while using a scikit-learn style in usage. It extends Pandas to add survival analysis tools that are missing from the greater Python ecosystem.

You can easily install the necessary dependencies/packages on your AWS instance with the following command:

```
pip3 install numpy matplotlib sklearn pandas unidip lifelines boto3
```

```
[8]: # Importing necessary Python modules
import numpy as np
import pandas as pd
import lifelines

from pandas.plotting import scatter_matrix
from sklearn.cluster import KMeans
from unidip import UniDip
from lifelines import KaplanMeierFitter
import matplotlib.pyplot as plt

# Importing the gene expression data for samples
import boto3
bucket = 'CHANGEME!'
clinical = 'clinical.tsv'
fpkm = 'fpkm.tsv'

boto3.resource('s3').Bucket(bucket).download_file(clinical, 'clinical.tsv')
boto3.resource('s3').Bucket(bucket).download_file(fpkm, 'fpkm.tsv')
```

Python 3 (Base Python) | Idle Mode: Command Ln 1, Col 1 AWS_workshop_python_module.ipynb

Jupyter Notebooks in SageMaker Studio

Amazon SageMaker Studio File Edit View Run Kernel Git Tabs Settings Help Feedback

Launcher AWS_workshop_python_mod Terminal arn:aws:sagemaker:...

2 vCPU + 4 GiB Python 3 (Base Python) Share

The box plot displays the distribution of data across various stages. The y-axis ranges from 0.0 to 0.6. The x-axis categories are: not reported, stage i, stage ia, stage iia, stage iib, stage iii, stage iiii, stage iiib, stage iiic, stage iv, and stage x. Each category has a box plot with a green median line, a blue box for the interquartile range, and whiskers extending to the minimum and maximum values. Outliers are shown as open circles. The 'stage ia' category shows the highest median value, around 0.35. The 'stage iii' category has the lowest median value, around 0.03.

Stage	Min	Q1	Median	Q3	Max	Outliers
not reported	0.00	0.04	0.07	0.11	0.20	
stage i	0.00	0.05	0.15	0.19	0.30	
stage ia	0.10	0.22	0.35	0.37	0.38	
stage iia	0.00	0.06	0.09	0.18	0.30	0.48, 0.52
stage iib	0.00	0.00	0.08	0.14	0.30	0.36, 0.40, 0.44, 0.57
stage iii	0.00	0.02	0.03	0.04	0.06	
stage iiii	0.00	0.00	0.06	0.25	0.49	
stage iiib	0.00	0.00	0.00	0.10	0.24	
stage iiic	0.00	0.00	0.08	0.16	0.24	0.49
stage iv	0.00	0.06	0.15	0.24	0.29	0.57
stage x	0.00	0.03	0.12	0.20	0.26	0.57

Python 3 (Base Python) | Idle Mode: Command Ln 1, Col 1 AWS_workshop_python_module.ipynb



Jupyter Notebooks in SageMaker Studio

The screenshot shows the Amazon SageMaker Studio Control Panel. The left sidebar contains navigation options: Dashboard, Search, Ground Truth (Labeling jobs, Labeling datasets, Labeling workforces), Notebook (Notebook instances, Lifecycle configurations, Git repositories), and Training (Algorithms, Training jobs, Hyperparameter tuning jobs). The main content area is titled "Amazon SageMaker Studio Control Panel" and includes a breadcrumb trail: Amazon SageMaker > Amazon SageMaker Studio > Control Panel. A heading reads "Choose your user name, then choose Open Studio to get started" with an "Add user" button. Below this is a search bar labeled "Search users" and a pagination control showing "1" of 1 items. A table lists two users:

User name	Last modified	Created	
researcher2	Jul 12, 2020 19:00 UTC	Jul 12, 2020 19:00 UTC	Open Studio
default-1594577429897	Jul 12, 2020 18:15 UTC	Jul 12, 2020 18:15 UTC	Open Studio

Below the table is a "Studio Summary" section with a "Delete Studio" button and a link "How to delete Studio". The summary table contains the following information:

Status	Studio ID	Execution role	Authentication method
✔ Ready	d-omd5oyipcqna	arn:aws:iam::767145282866:role/service-role/AmazonSageMaker-ExecutionRole-	AWS Identity and Access Management (IAM)

At the bottom of the page, there is a footer with "Feedback", "English (US)", and copyright information: "© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use".



Research and Technical Computing on AWS

<https://aws.amazon.com/government-education/research-and-technical-computing/>

Thank You!

Danyell Wilt
AWS Sr. Solutions Architect

