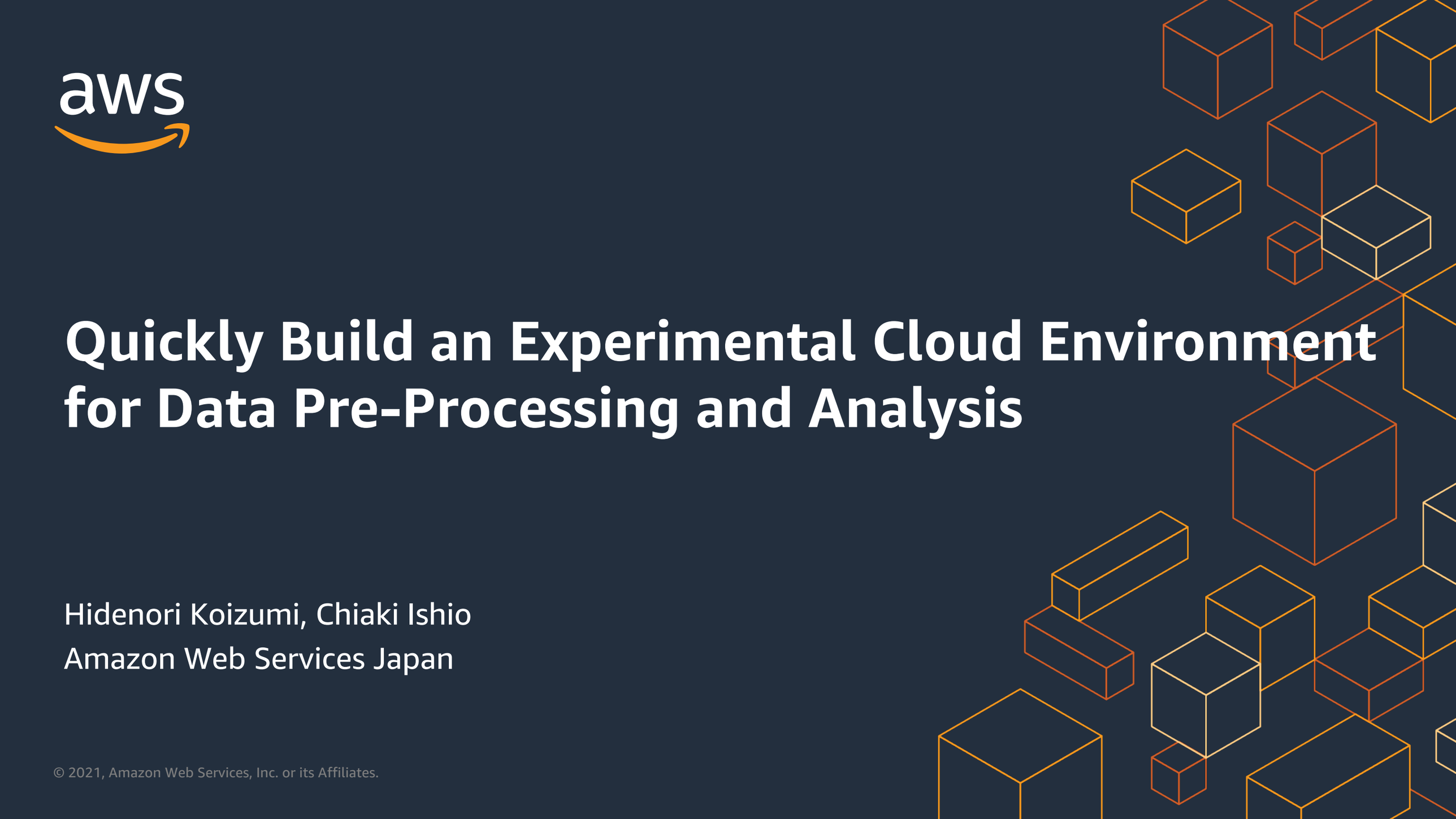
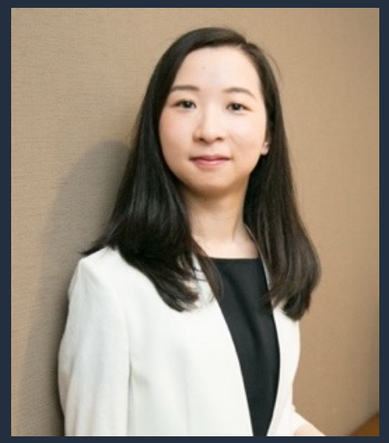


Quickly Build an Experimental Cloud Environment for Data Pre-Processing and Analysis

Hidenori Koizumi, Chiaki Ishio
Amazon Web Services Japan

Speaker Introduction

Hidenori Koizumi
Prototyping Solutions Architect in Japan’s Public Sector, an expert in developing solutions in the research field based on his scientific background. He has been developing solutions with code such as AWS CDK.



Chiaki Ishio
Solutions Architect in Process Manufacturing and Healthcare Life Sciences team at Amazon Web Services Japan. She is passionate about helping her customers design and build their systems on AWS.

aws

Agenda

1. Data analysis environment on cloud platform
2. Common challenges in data science
3. How can we approach the problems
4. Next Action

aws

1. Data analysis environment on cloud platform

aws

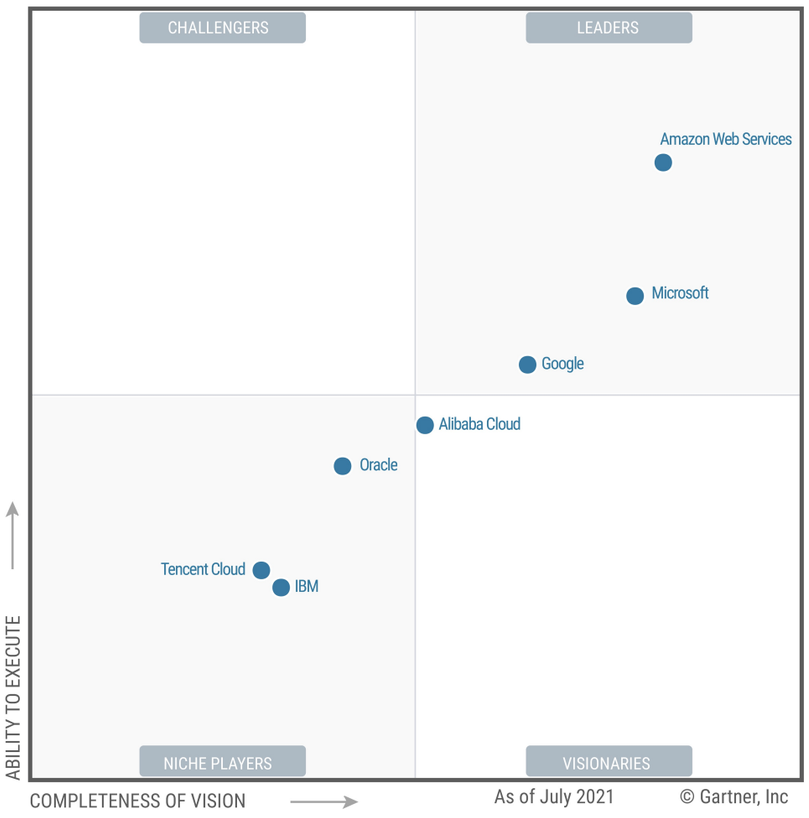
Why users choose cloud for data analysis environment

- Increased data types and volumes
    - Need to leverage not only RCT but RWD/RWE for healthcare decisions and new insights
- Seamless data movement
    - As data grows, need to easily move a portion of the data from one data store to another
- Need for new analysis approach
    - Machine learning and deep learning

aws

Why customers choose AWS

Figure 1: Magic Quadrant for Cloud Infrastructure and Platform Services



Source: Gartner (July 2021)

source: https://aws.amazon.com/blogs/aws/aws-named-as-a-leader-for-the-11th-consecutive-year-in-2021-gartner-magic-quadrant-for-cloud-infrastructure-platform-services-cips/

Gartner recognizes **AWS as a Leader** for the 11th straight year

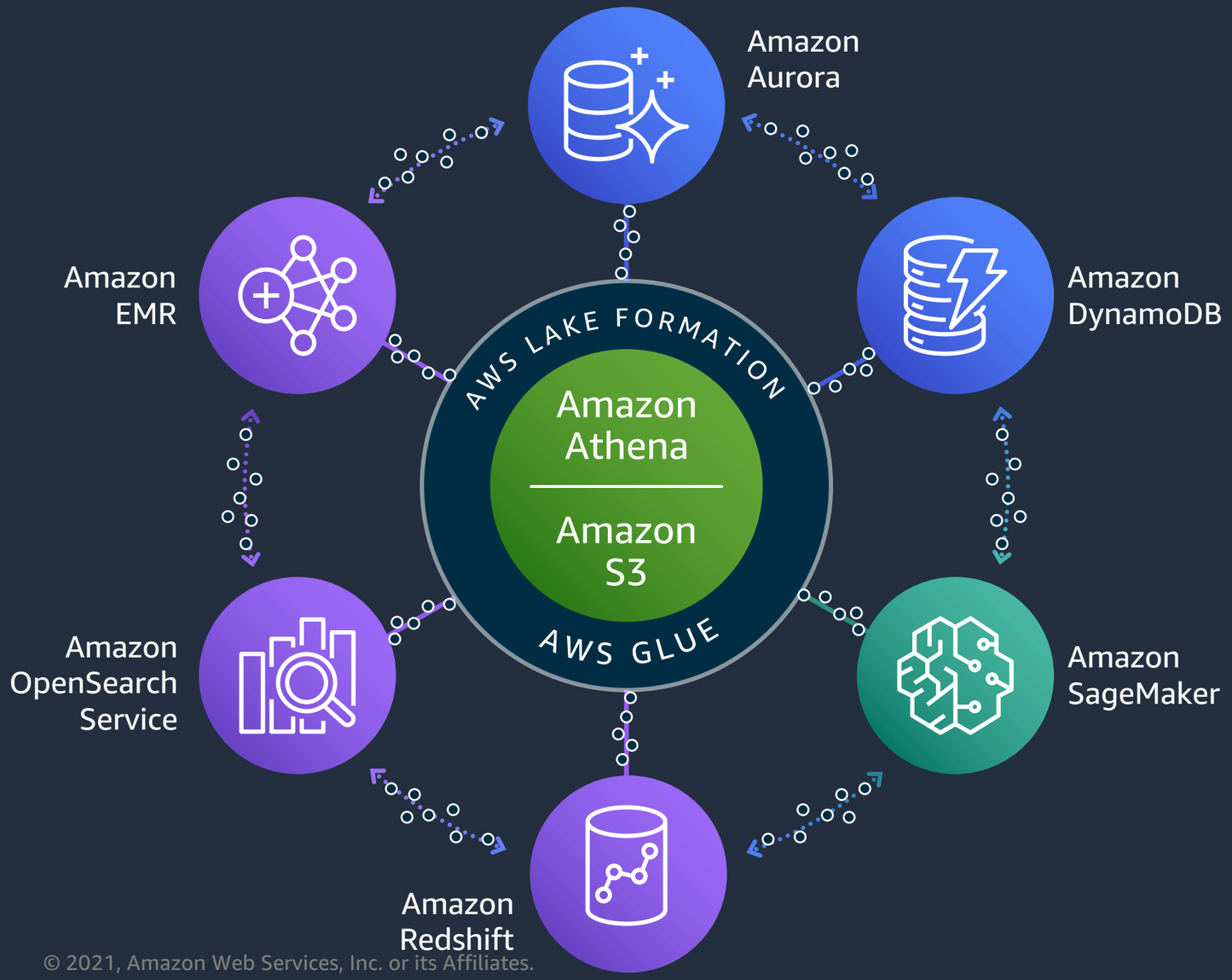
Magic Quadrant for Cloud Infrastructure and Platform Services (CIPS)

Benefits

- **Global reach & high availability**
  - **81** availability zones spanning 25 geographic regions
- **Security & compliance**
  - **230+** security features
- **Customer obsession & innovation**
  - **200+** service offerings

source: https://aws.amazon.com/jp/aws-ten-reasons/

Lake House approach



Scalable Data Lakes

Purpose-build Analytics Services

Unified Data Access

Unified Governance

Performant and Cost-effective

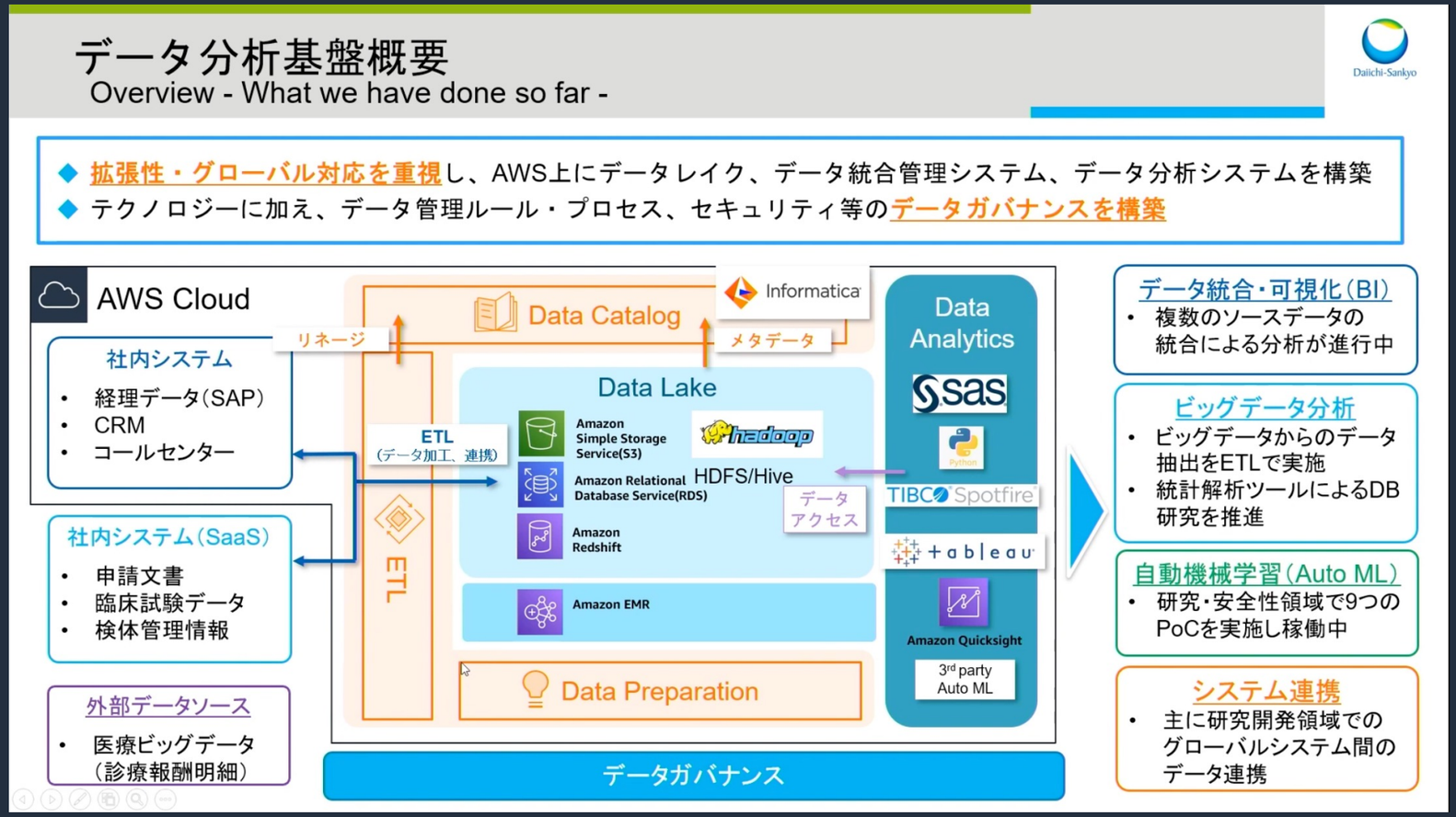
Daiichi Sankyo Co., LTD
Build company-wide data analysis environment

Challenges

- Centralized data management and integrated with data
- Advanced data analysis leveraging ETL, AI and ML
- Security, traceability, scalability, globally and compliance including CSV

Why AWS

- The result of implementing GxP requirements system
- Global scalability and enable compliance
- Comprehensive support such as advanced technology and operation



Source：AWS Summit Online 2021 https://www.youtube.com/watch?v=xhmOyNmv1Lg

Analysis approach with ML or Deep Learning(Amazon SageMaker)

Enhance clinical trials

Challenges

- Clinical trial protocols are large complex documents that can be hundreds of pages long. The important clinical details for testing labs can be scattered throughout the document, making them difficult to find and prone to cause human error.

Benefits

- Reduce risk of missing critical protocol information.
- Resulted in 50% reduction of workload.

Get new insights on complex diseases

Challenges

- Understanding real-world treatment efficacy (outside of formalized clinical trials), as well as unmet patient needs, requires data from large, heterogeneous populations of people using specific therapeutics.

Benefits

- Incorporated deep learning models to provide insights on burden of disease and patient unmet needs based on real-world evidence.
- Developed machine learning models that could begin to predict patients that would develop certain disease types.

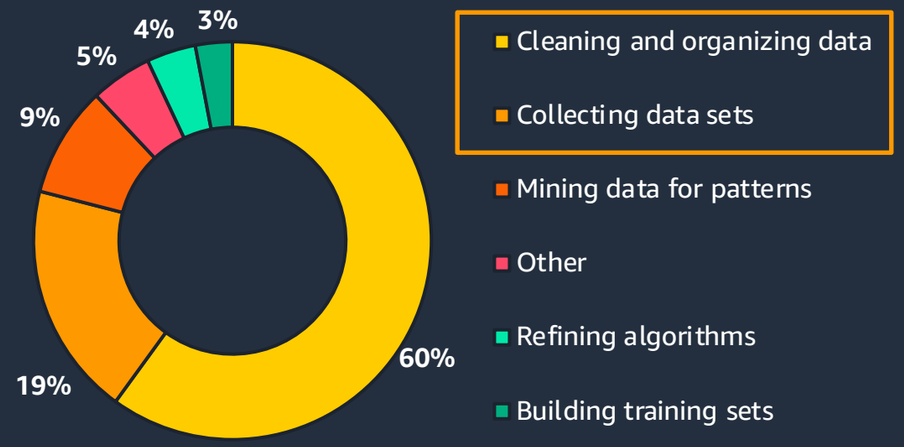
aws

2. Common challenges in data science

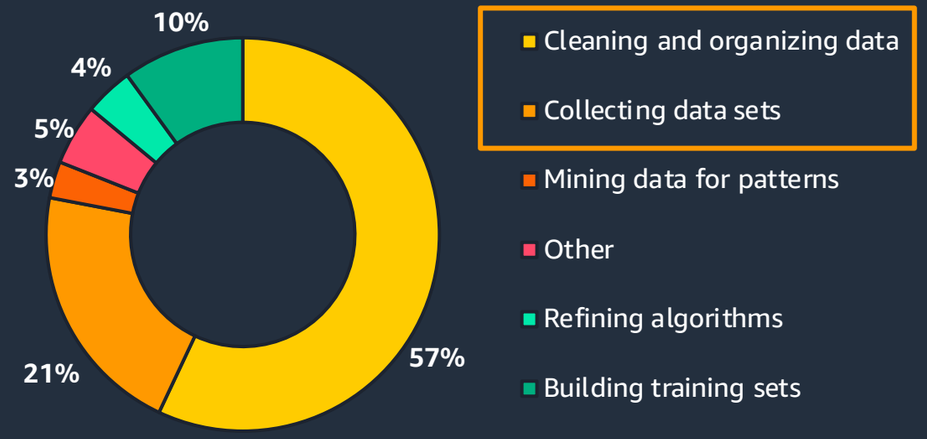
aws

Data preparation still dominates data scientist's time

What data scientists spend the most time doing

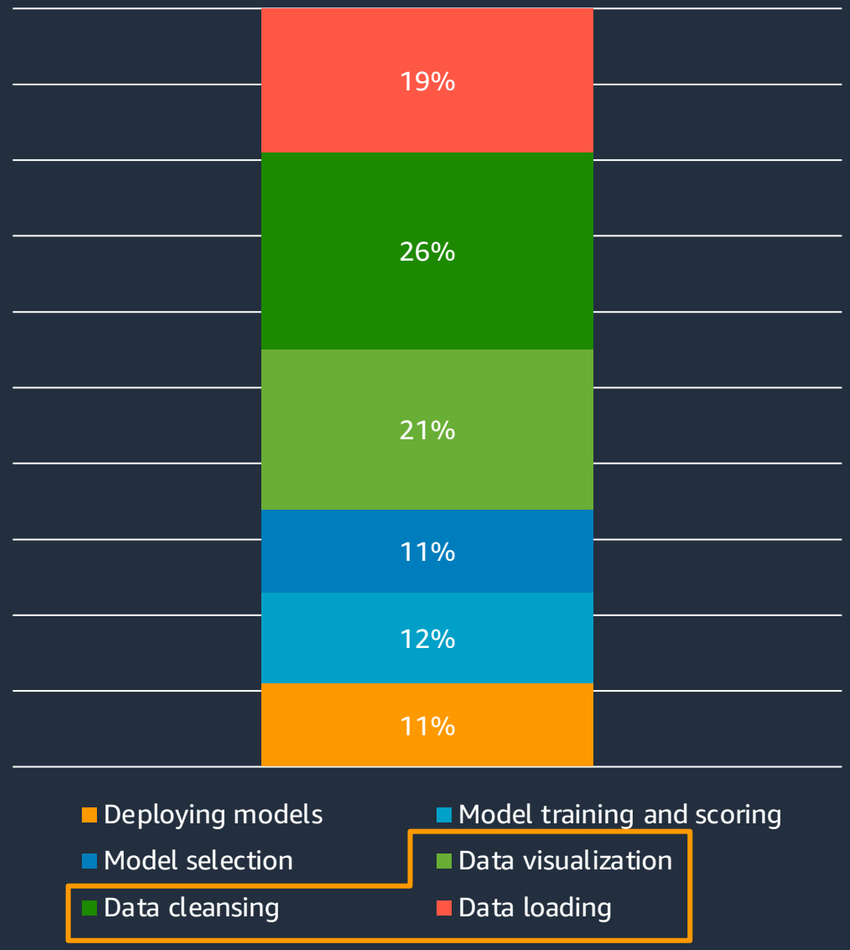


What's the least enjoyable part of data science



Source: Forbes survey of 80 data scientists, March 2016

How much of your time is spent in each of the following tasks



Source: Anaconda survey of 2,360 respondents included data scientists, researches, analyist and more, 2020

aws

The need for collaboration

- Changes in the working environment under the COVID-19
- Quick access to raw data or the data which you want
- Reduce the similar tasks
- Gain new insights by sharing your code and results
- Enable data science and business team to work

aws

3 Challenges in data science – this session

Collecting data

Cleaning and organizing data

Collaborating with your peers

aws

The reasons why data collection takes so long

- Put data in various databases
- Long time to explore the data which you want from open-source
- Data is not available immediately

aws

The reason why data pre-processing takes so long

- Transforming and cleaning data require a lot of code
- Checking and visualizing data require various tools
- Unable to select and query data from multiple data sources quickly

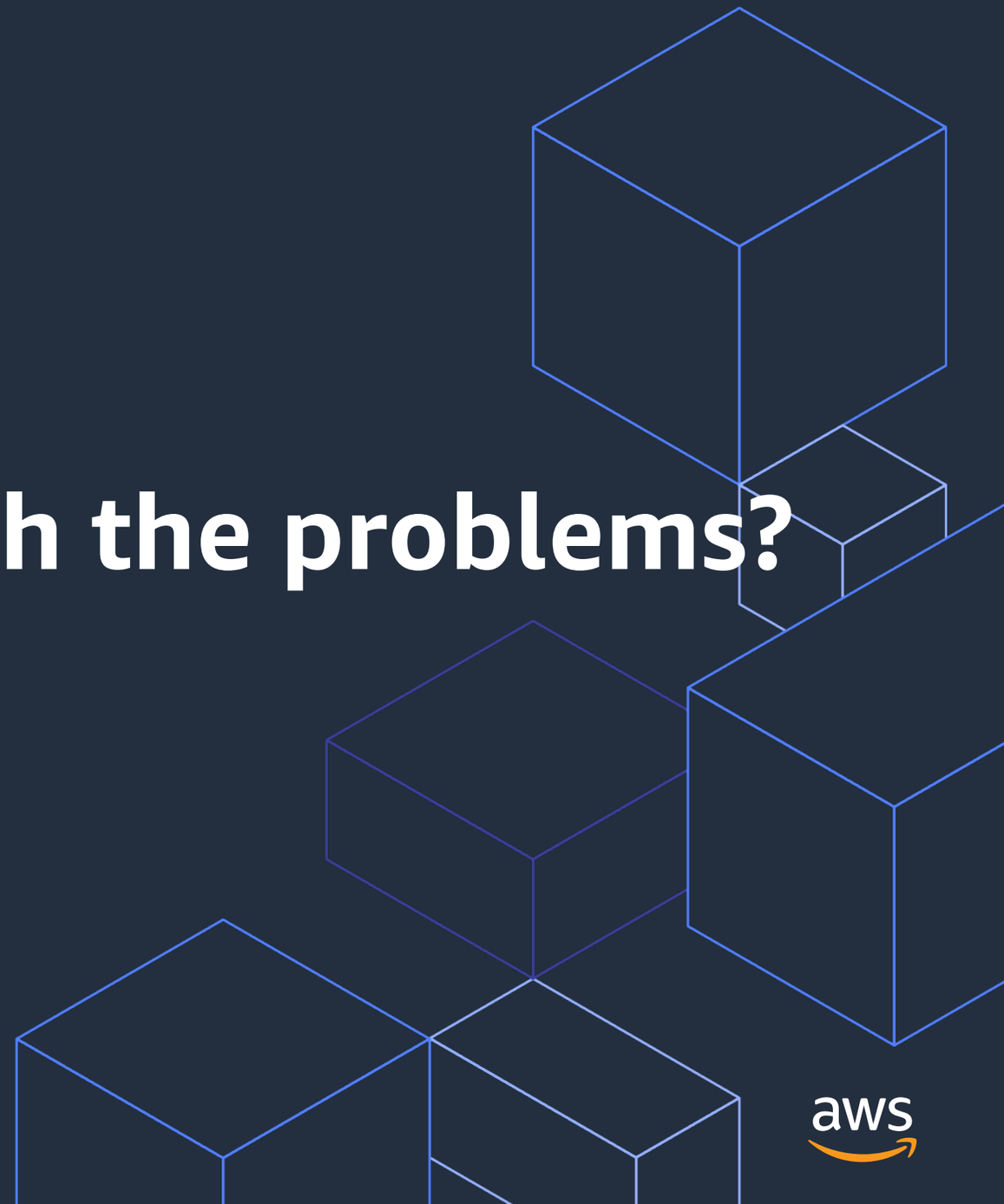
aws

How to collaborate successfully

- Sharable integrated development environment
    - Easy to collaborate with peers
    - Share code and notebooks quickly
- Use multiple data sources with shared environment
- Use popular data science and ML framework
    - Python(TensorFlow, PyTorch), PySpark, R, Scala, etc

aws

3. How can we approach the problems?



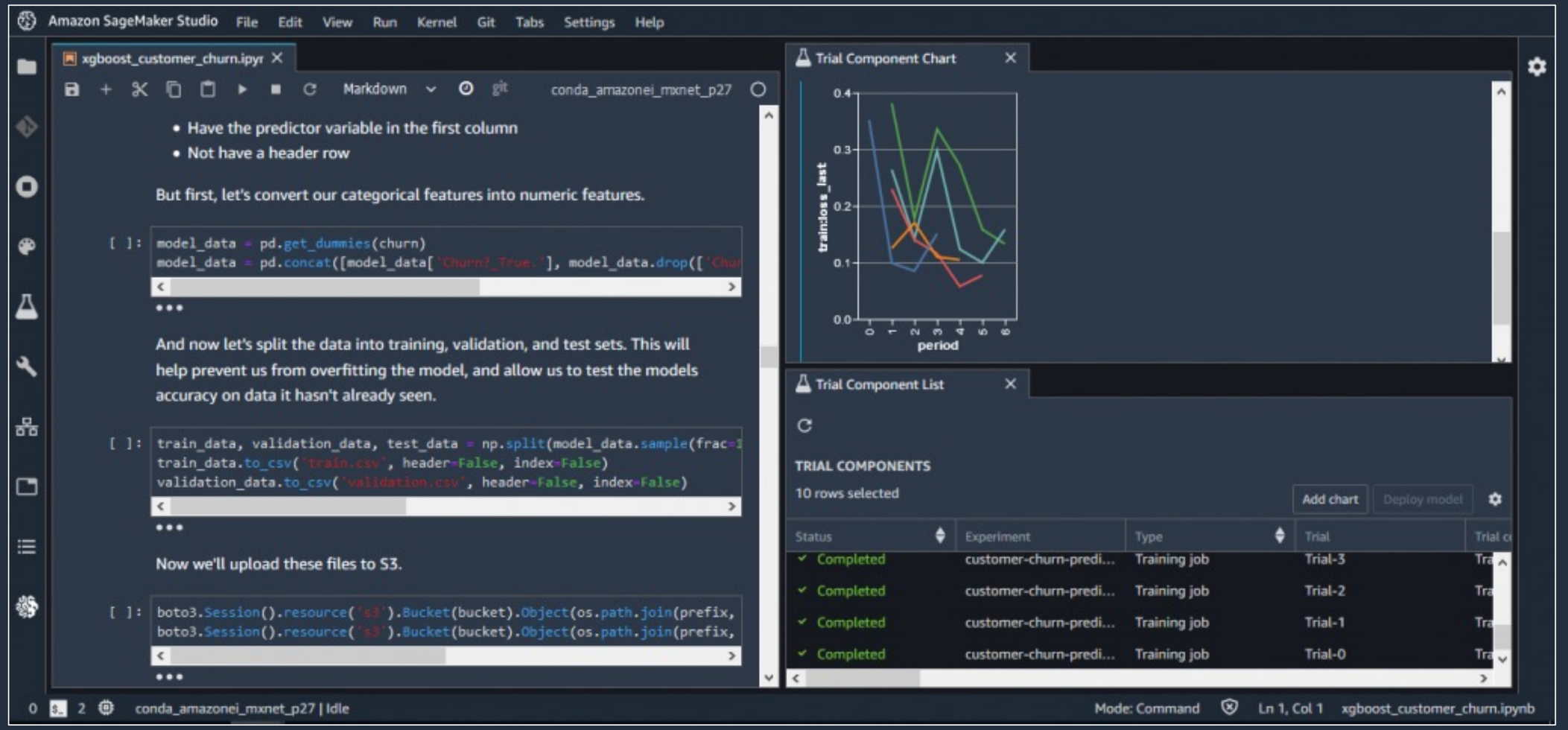
aws

Data pre-processing tools look like ...

aws

Amazon SageMaker Studio

Fully Integrated Development Environment (IDE) for data science and machine learning



3 Challenges in data science – this session

Collecting data

Cleaning and organizing data

Collaborating with your peers

aws

Challenge 1: Collecting Data

aws

Common data sources in AWS

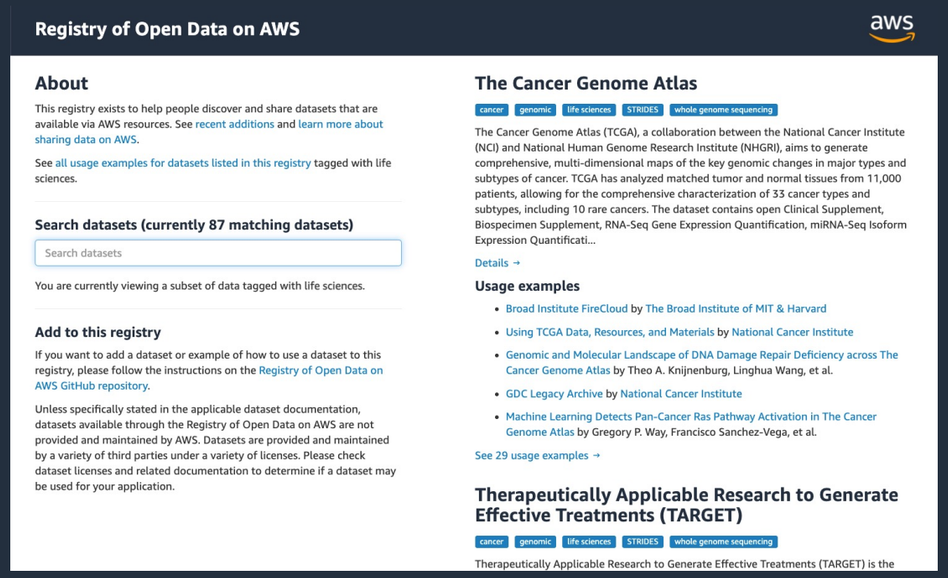
Amazon S3
Object storage service offering industry-leading scalability, data availability, security, and performance

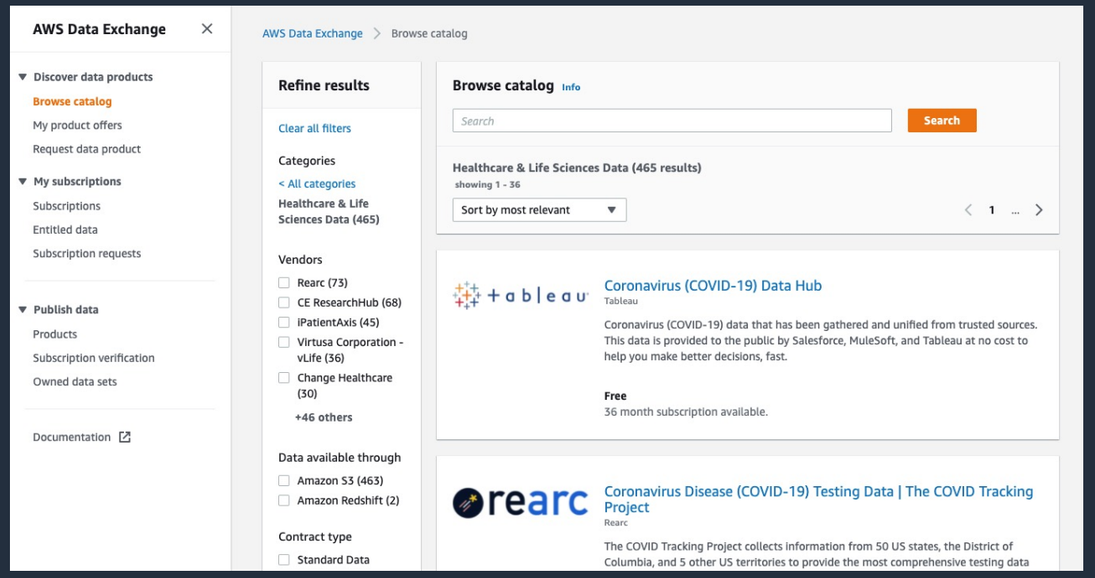
Amazon Redshift
Fully managed, petabyte-scale data warehouse service

aws

Data repositories available on AWS

- Registry of Open Data on AWS
  - Allow anyone to find public datasets on AWS
  - PubMed Central, The Cancer Genome Atlas, etc.
- AWS Data Exchange
  - Contain 1000+ licensable data products from 80+ data providers



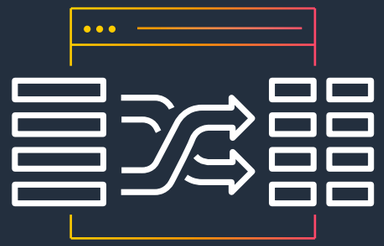


https://registry.opendata.aws/
https://aws.amazon.com/blogs/aws/aws-data-exchange-find-subscribe-to-and-use-data-products/

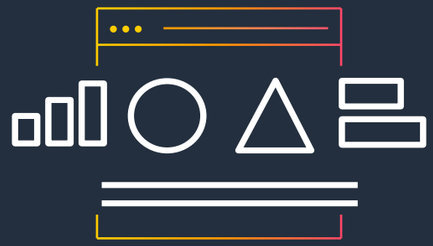
aws

Prepare data with SageMaker Data Wrangler

Data selection



Data transforms



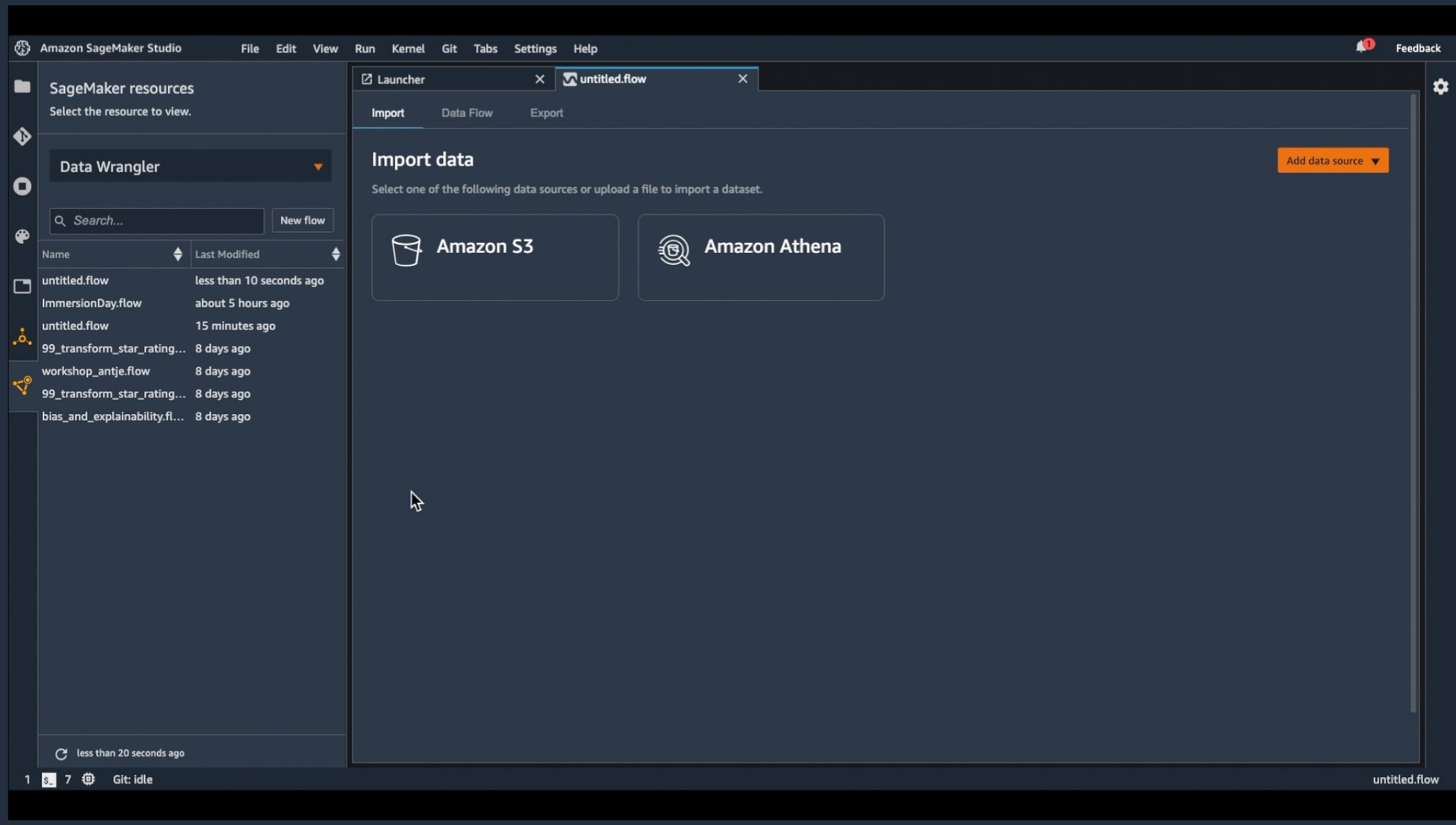
See data, spot inconsistencies, diagnose, and fix

Export your data

aws

Directly connect to your data sources

- Import data directly onto SageMaker Studio from S3, Athena and Redshift
- For large volumes of data sources, query and inspect the data before importing them



aws

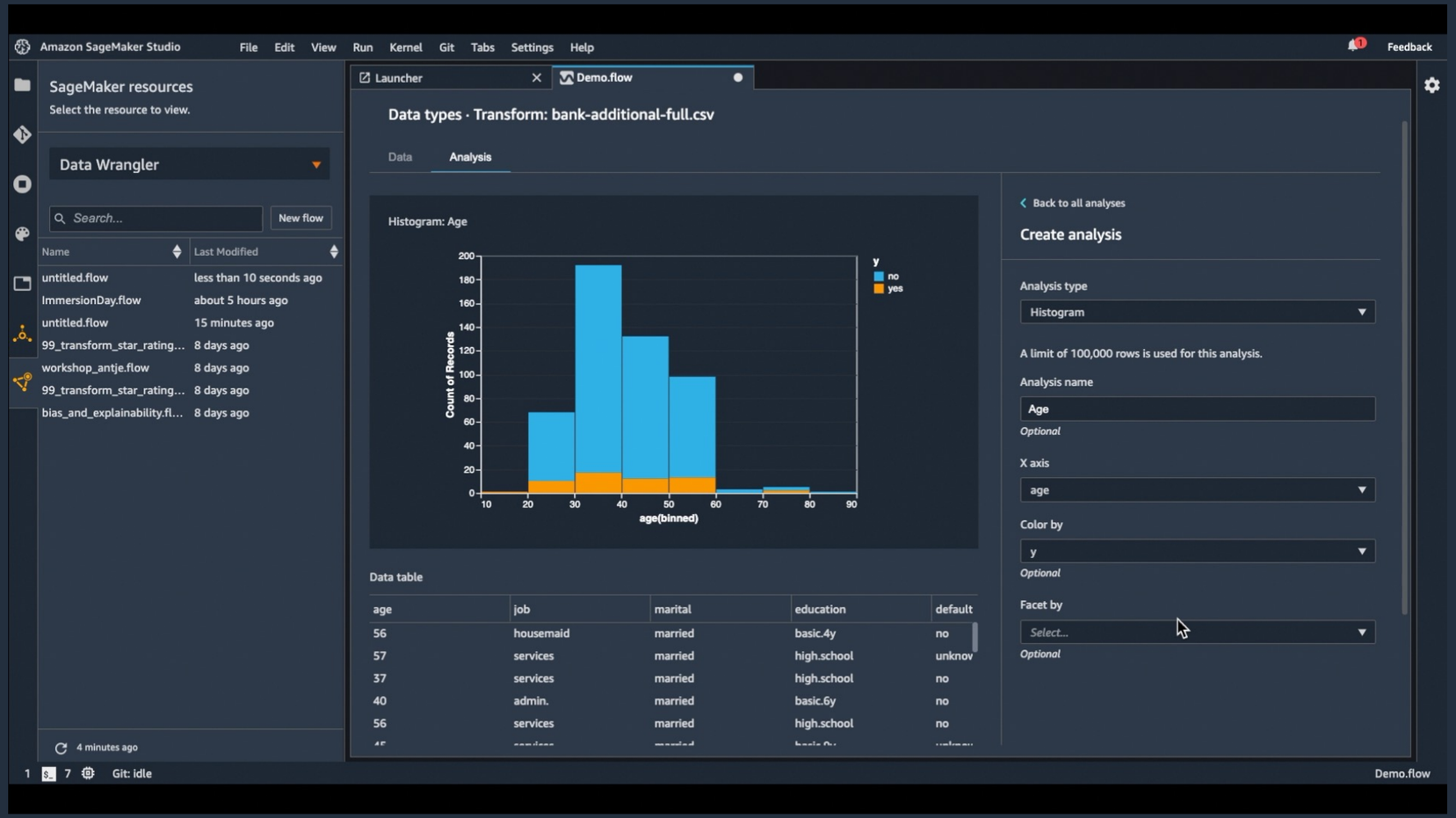
Challenge 2:
Cleaning and Organizing Data

aws

Understand your data visually

- Interactively create and edit your own visualizations so you can quickly detect outliers or extreme values

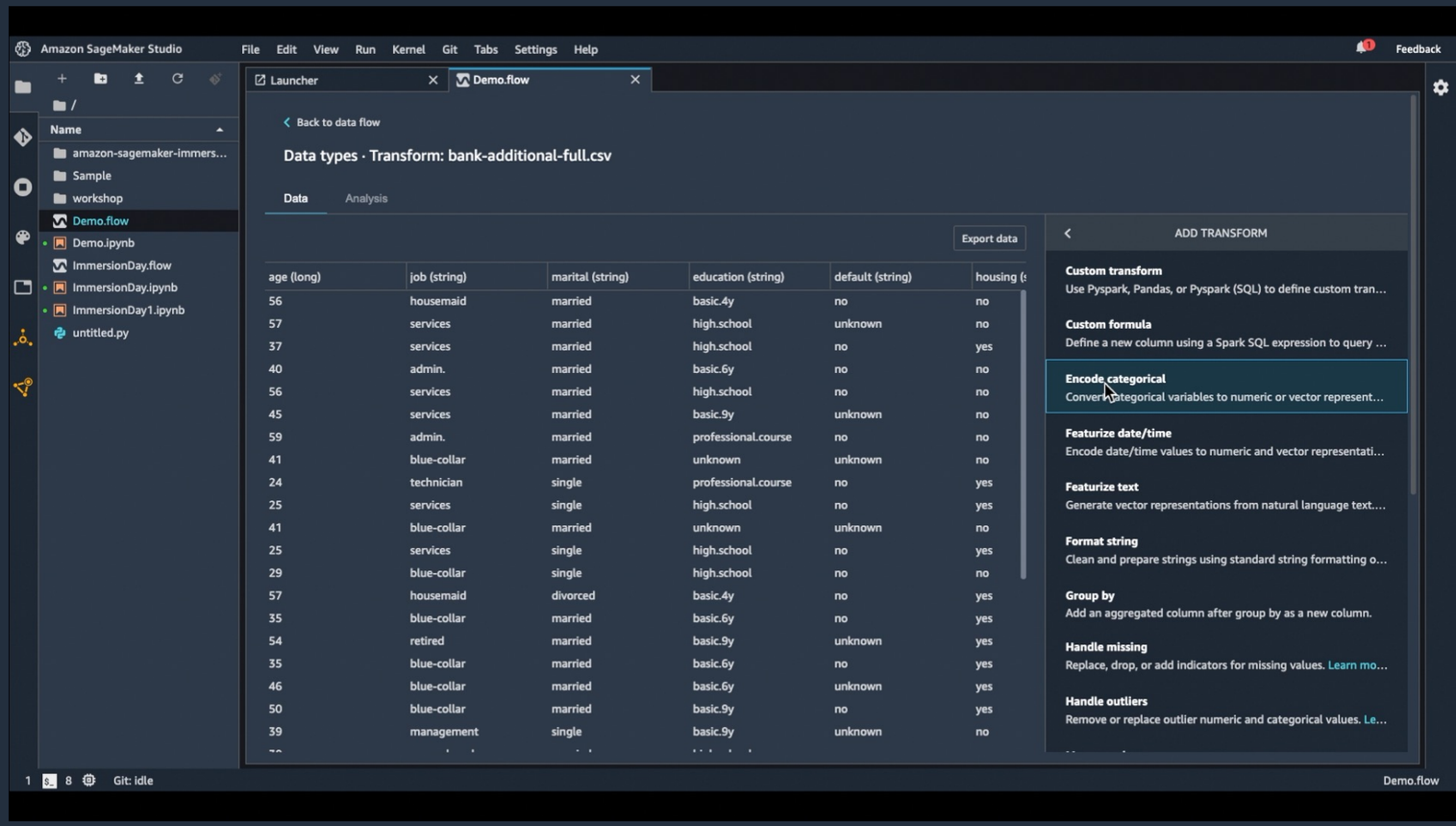
- Preconfigured visualization templates
    - histograms
    - scatter plots
    - box and whisker plots
    - bar charts etc.



Easily transform data

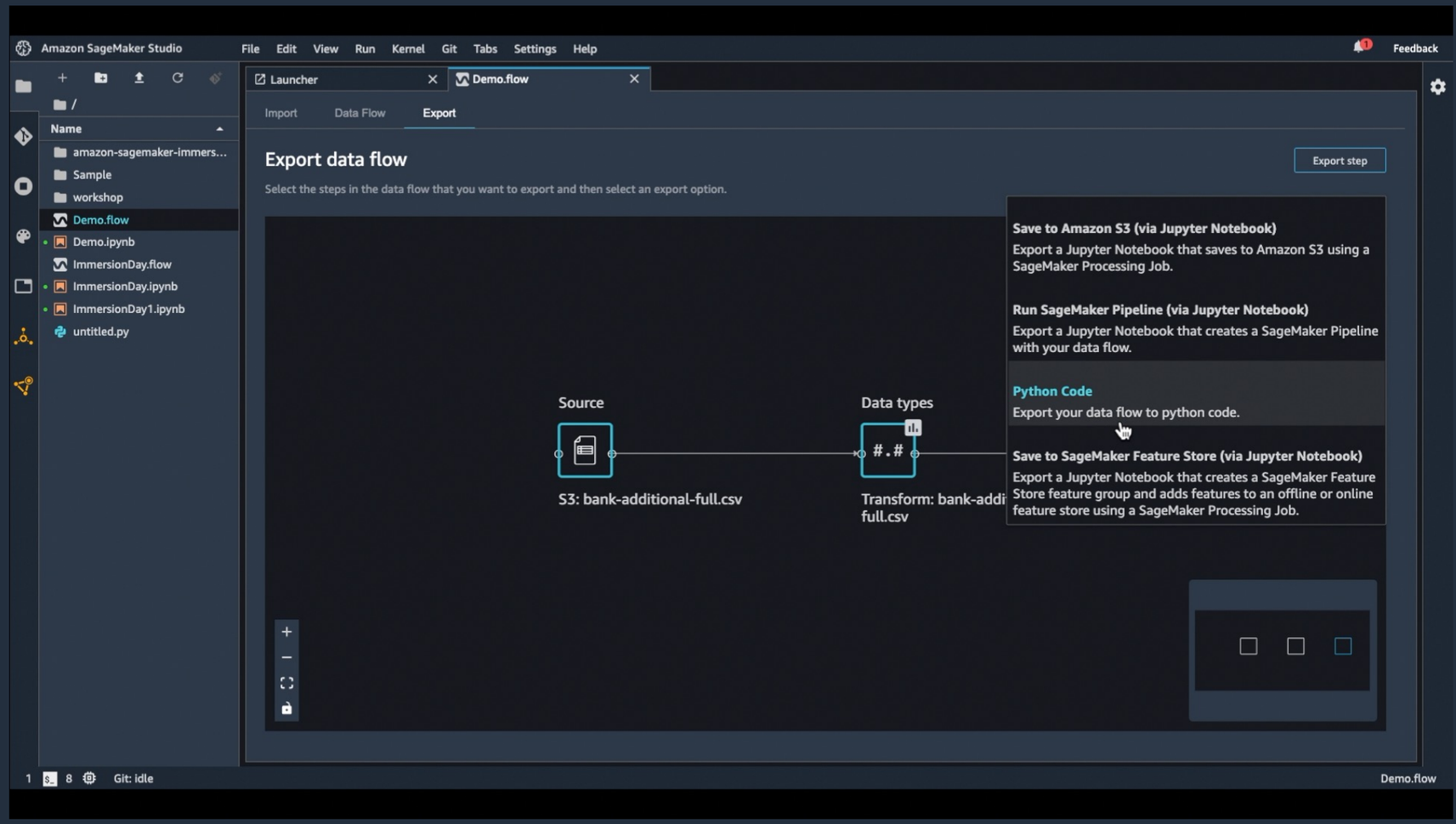
- 300+ built-in data transforms with GUI
    - Missing value detection
    - Outlier detection
    - Column manipulation

- Custom transforms in PySpark, SQL, and Pandas



Deploy data preparation workflows into production

- Export data preparation workflows to a notebook or Python code
- Publish created features to SageMaker Feature Store for reuse and syndication across teams and projects

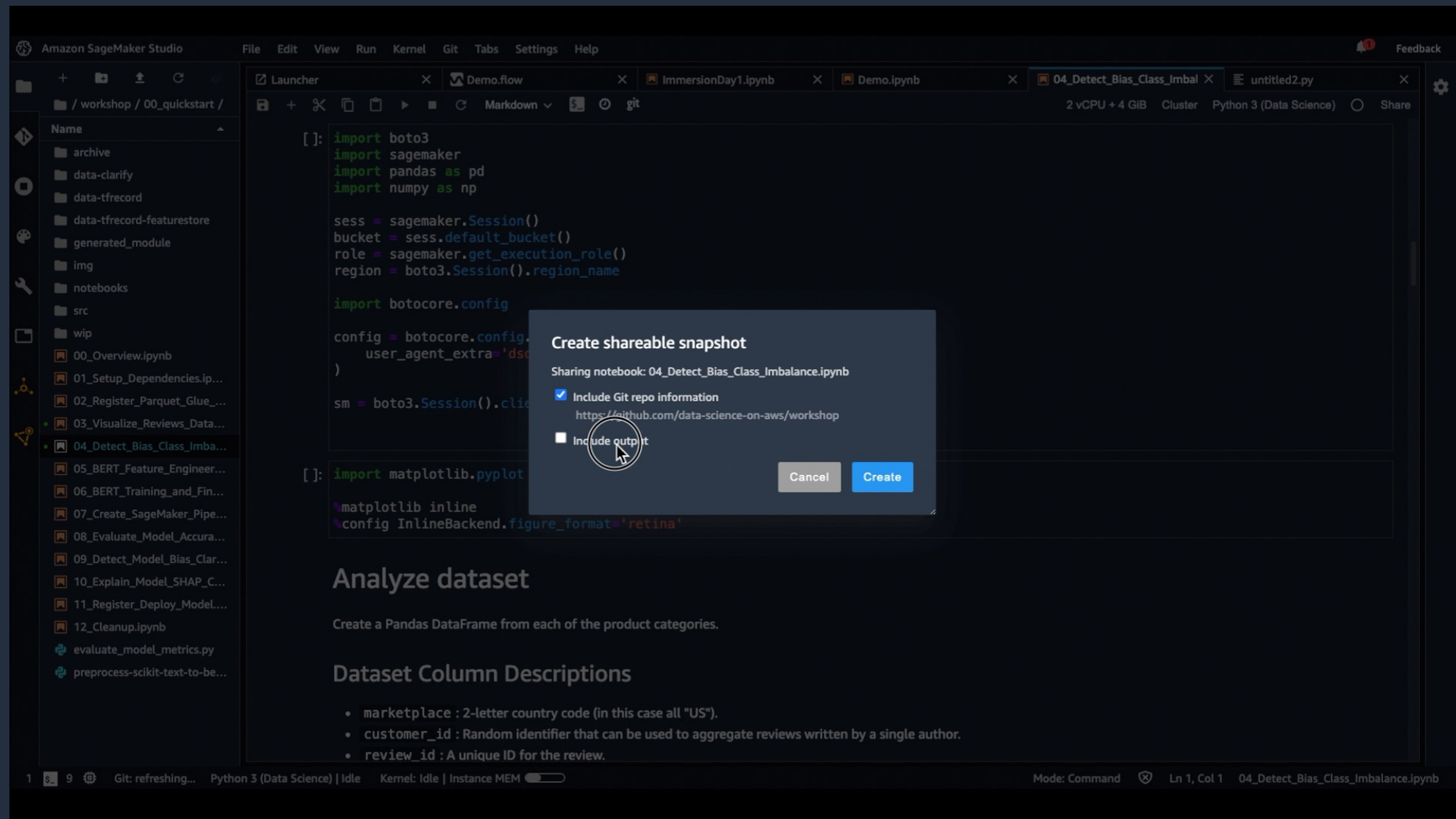


Challenge 3:
Collaborating with your peers

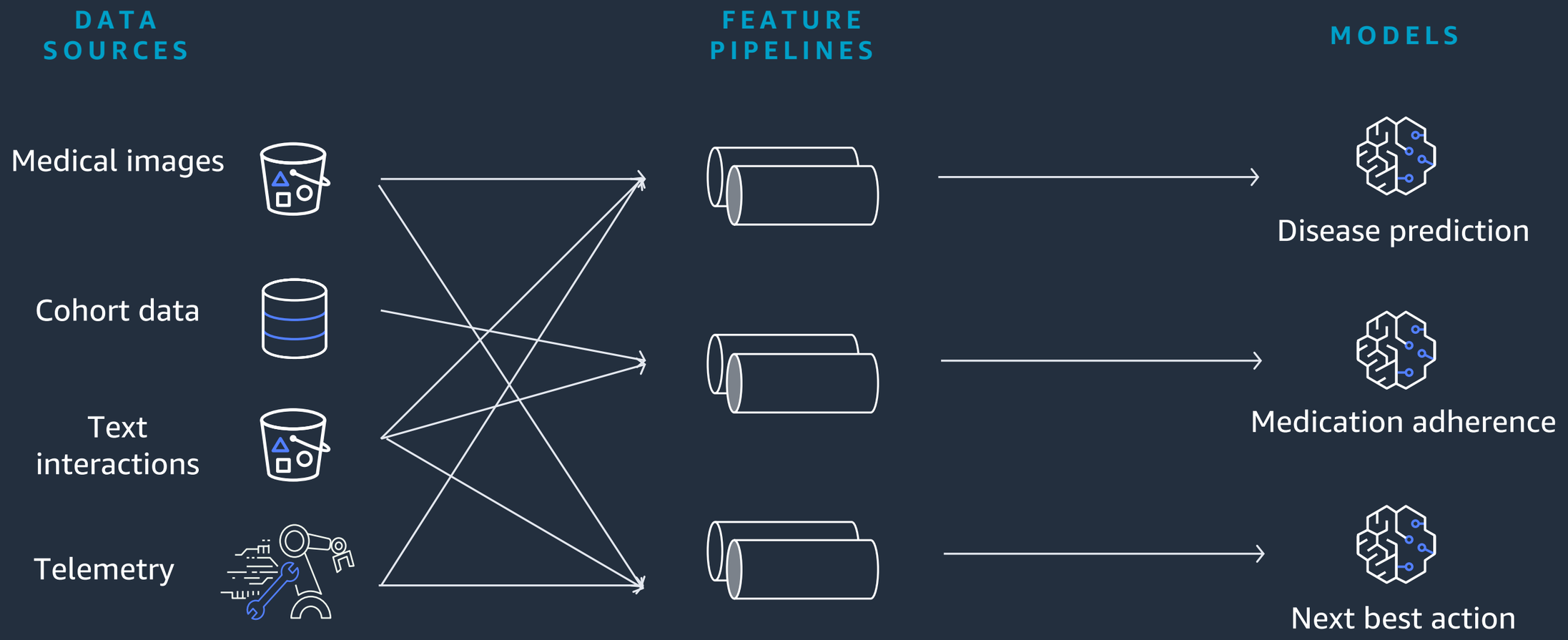
aws

Share your notebooks with your peers

- Generate a sharable snapshot link with a few clicks
- Snapshot includes the code and the image required to execute it



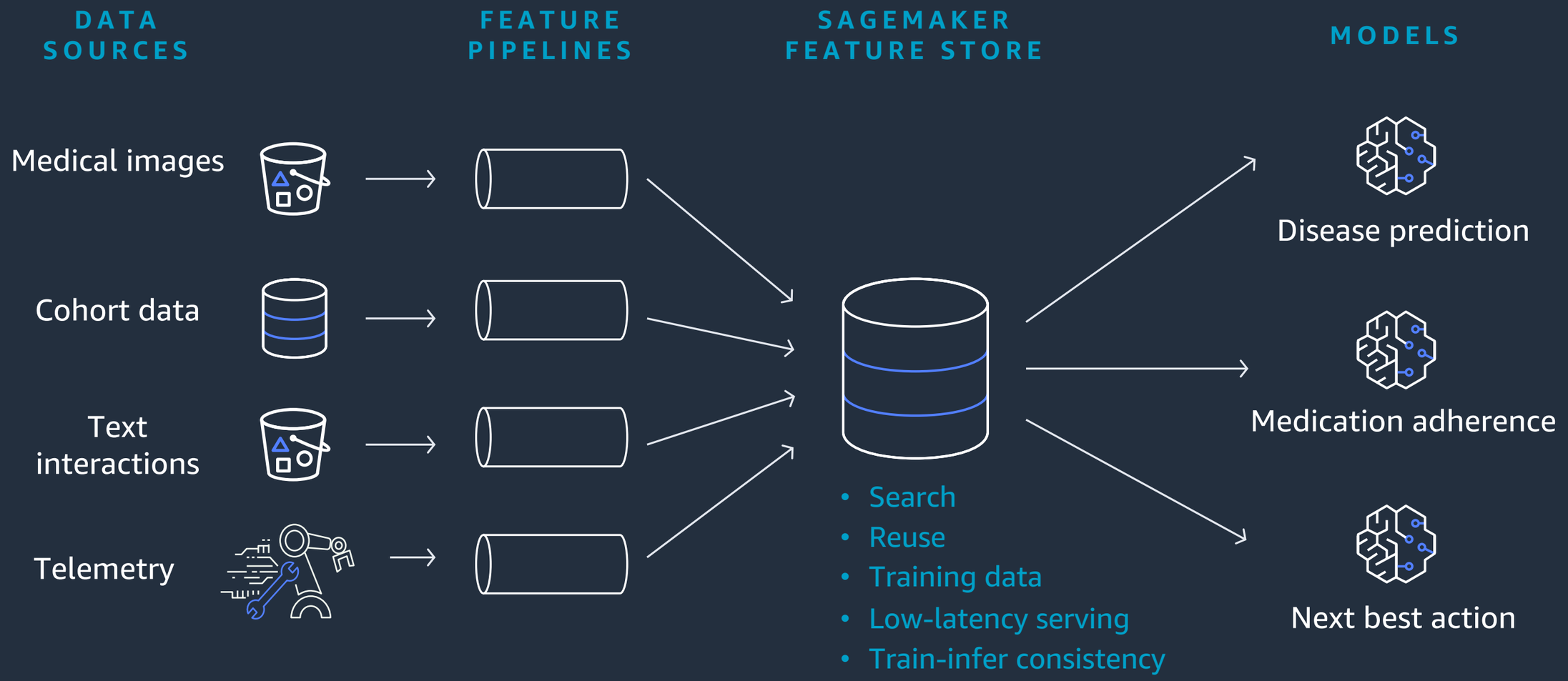
Without SageMaker Feature Store …



aws

With SageMaker Feature Store ...

Build features once, and reuse them across teams and models



aws

Customer Spotlight

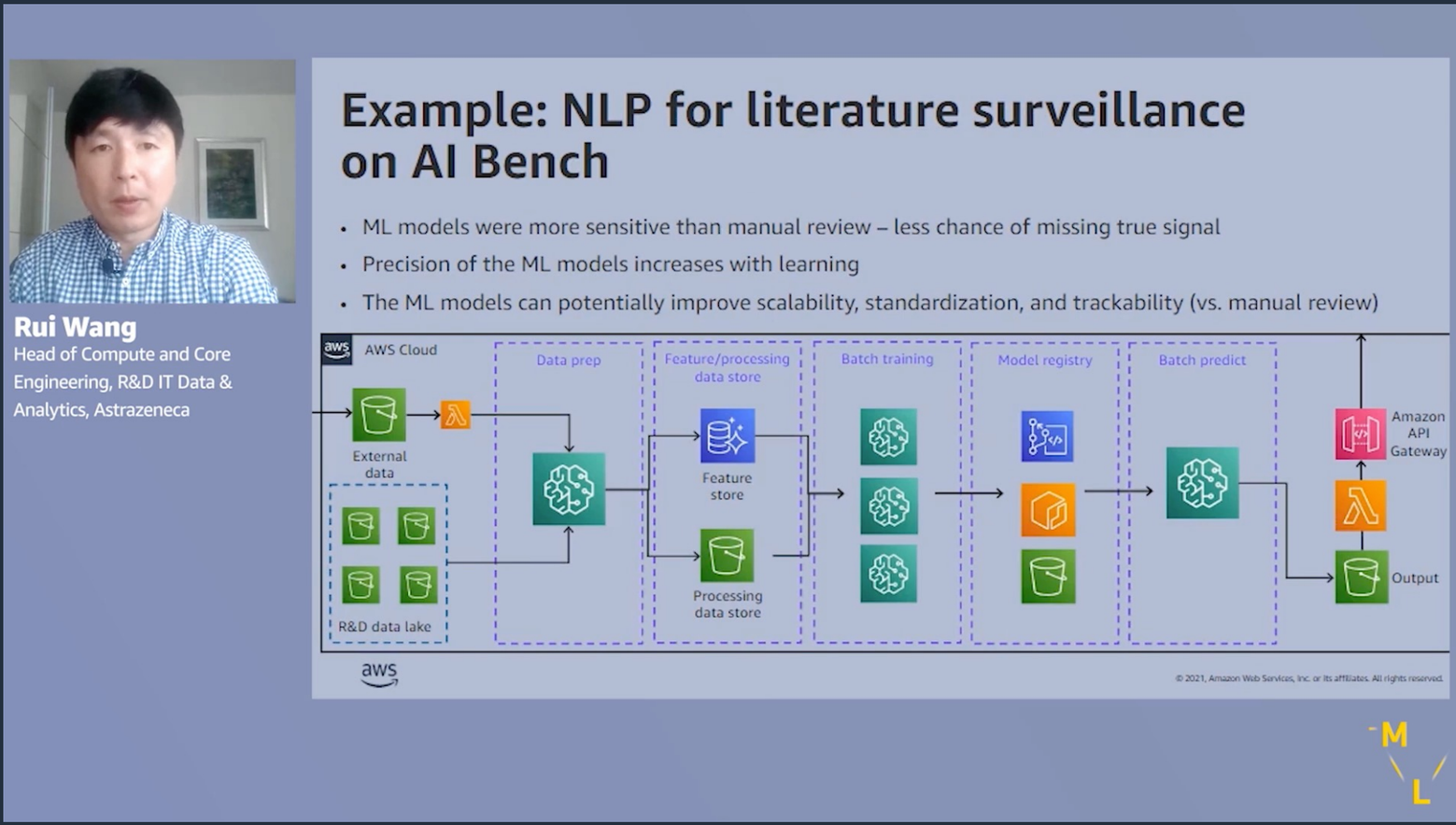
AstraZeneca discovers, develops, and commercializes prescription medicines in oncology and biopharmaceuticals, including in cardiovascular, respiratory, and immunology fields. It serves millions of patients across 145 countries and 70 markets.

> “Rather than creating many manual processes, we can automate most of the machine learning development process simply within Amazon SageMaker Studio.”
>
> Cherry Cabading
> *Global Senior Enterprise Architect, AstraZeneca*

aws

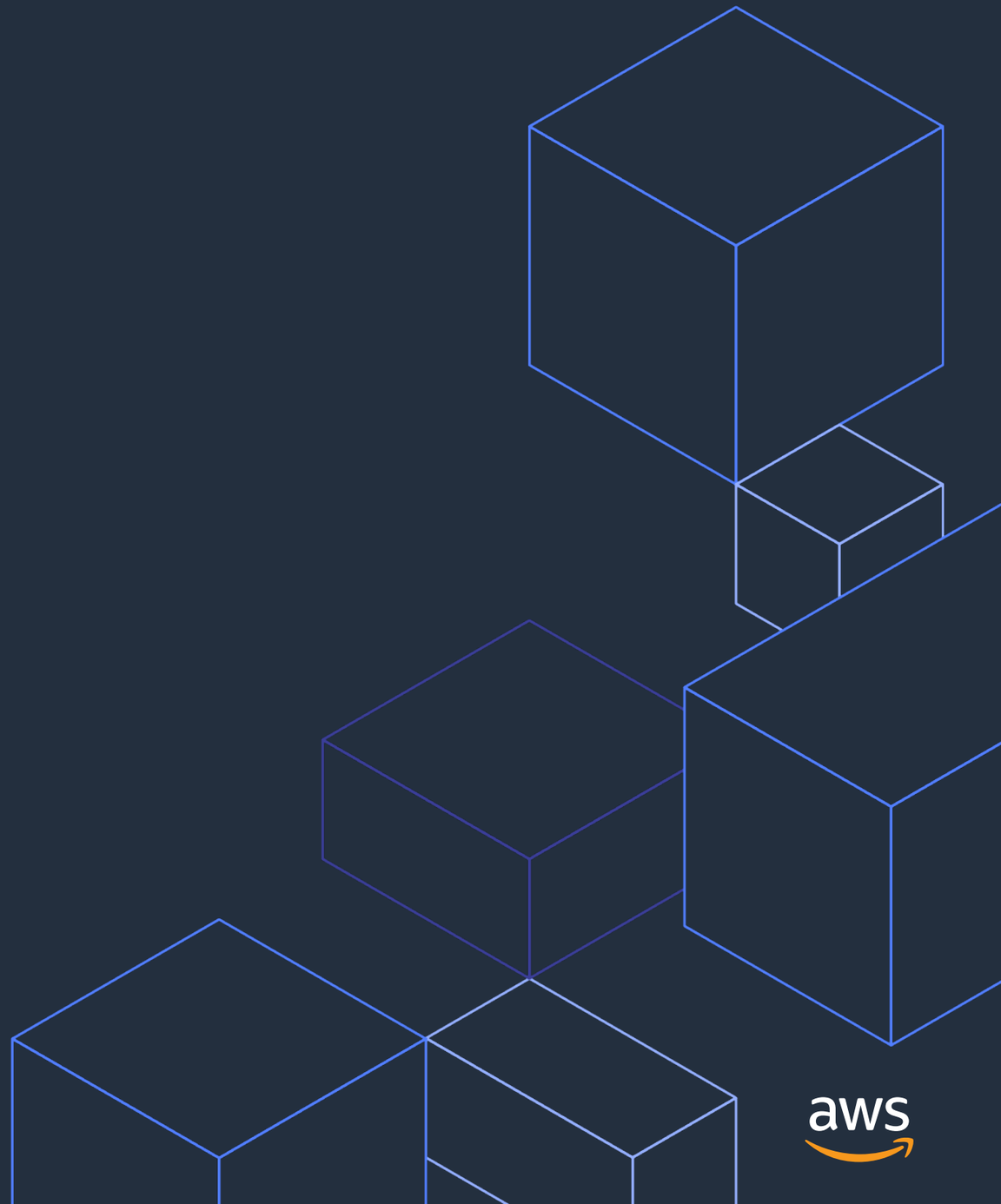
Customer Spotlight

- AI Bench 2.0
  - Data science & machine learning platform
  - Single-tenant multi accounts
  - GxP-compliant system
- Using SageMaker Studio and SageMaker Feature Store for train/build models



Source: AWS ML Summit 2021 https://youtu.be/Yz4NsQ4zl9g

4. Next Action

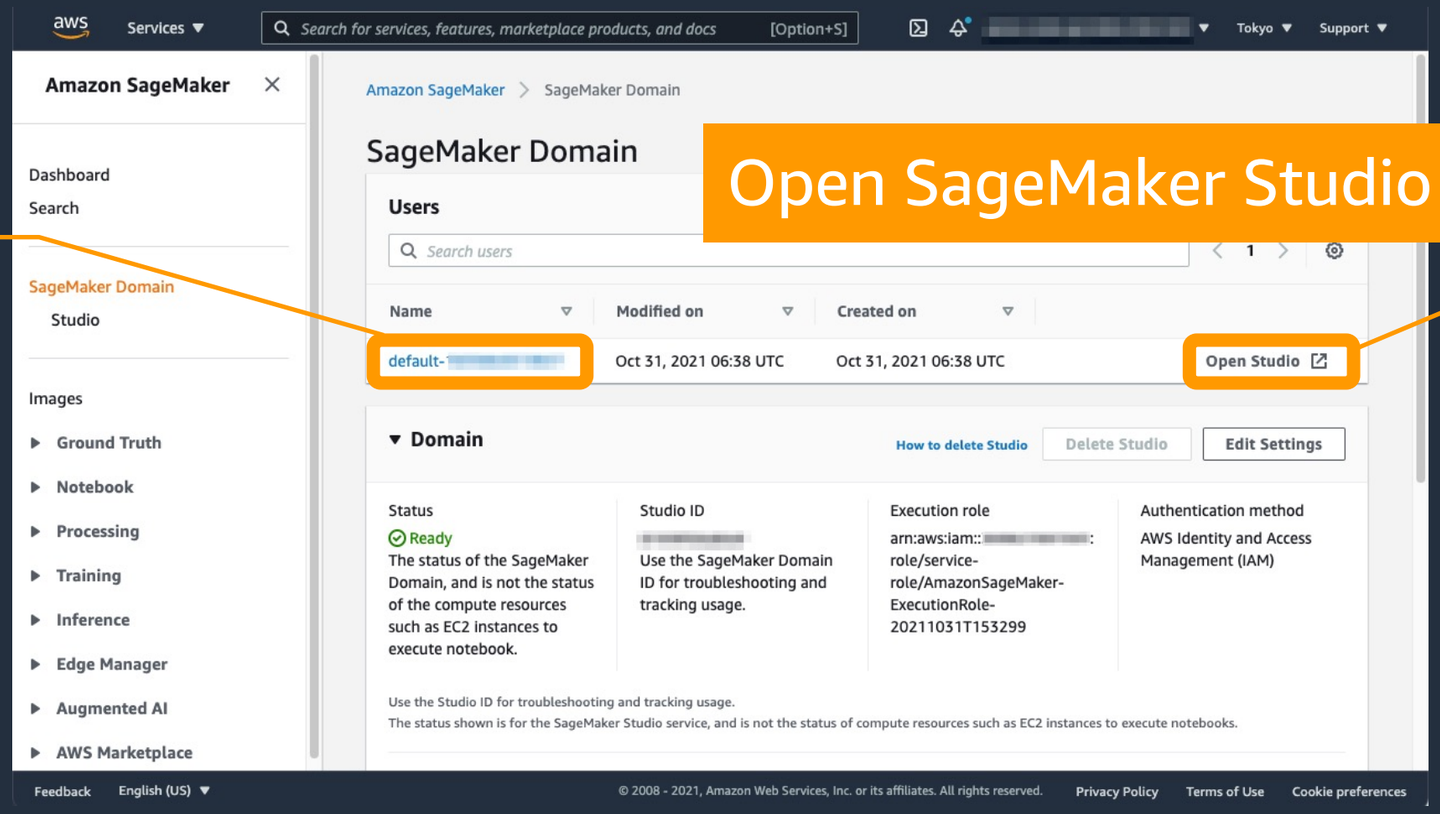


Getting Started with Amazon SageMaker Studio

- Visit console.aws.amazon.com/sagemaker
- Create your SageMaker Studio Domain
- Open Studio to get started

Add multiple users for SageMaker Studio

Open SageMaker Studio



- Or, build SageMaker Studio from code: https://github.com/aws-samples/aws-cdk-sagemaker-studio

Getting Started Resources

- Learn more about Amazon SageMaker for Healthcare and Life Sciences aws.amazon.com/sagemaker/healthcare-life-sciences/
- SageMaker Studio Workshop (in Japanese) sagemaker-immersionday.workshop.aws/ja/

Please feel free to contact us for further information.

aws

Thank you!