



# 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.

# Agenda

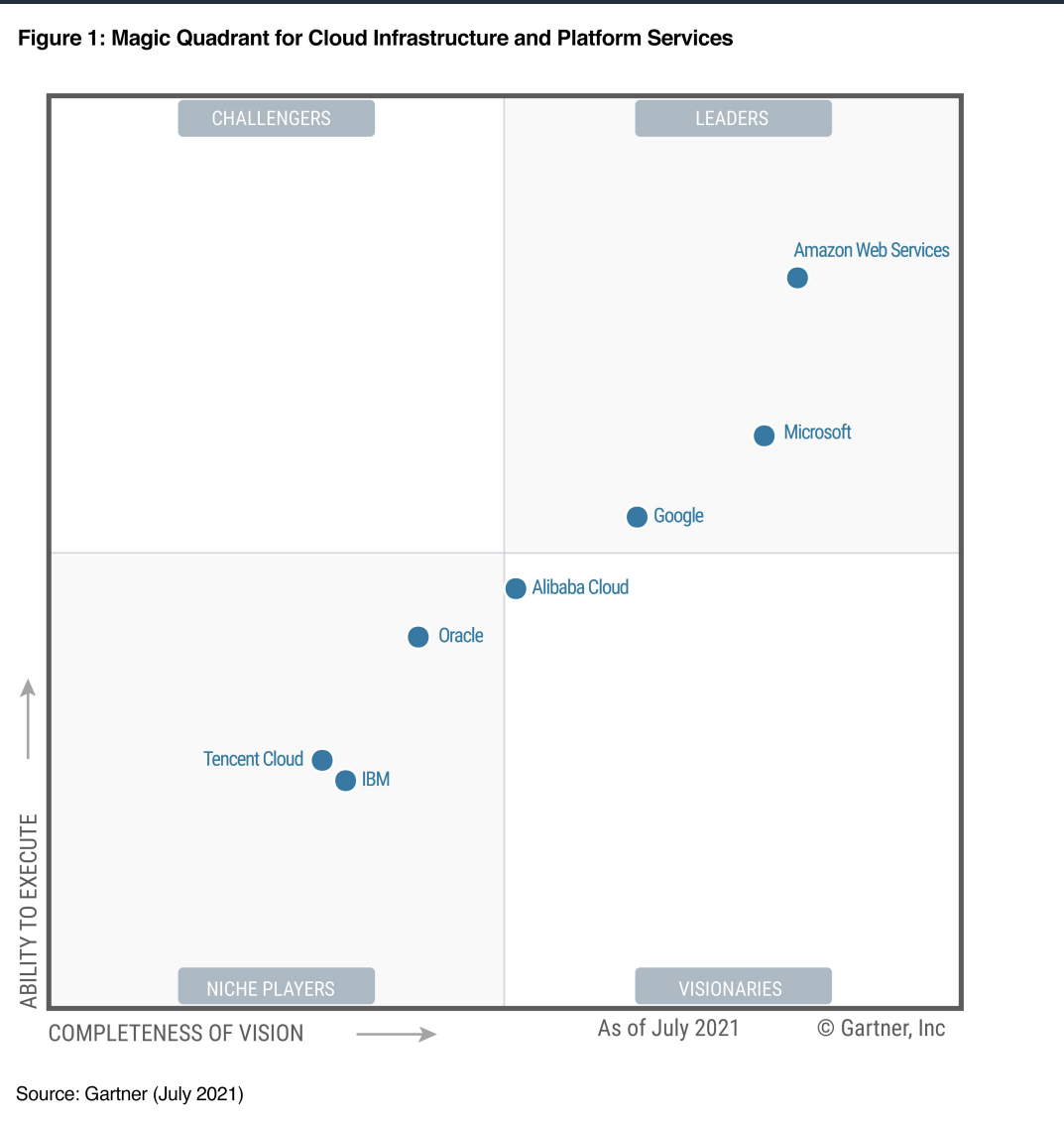
1. Data analysis environment on cloud platform
2. Common challenges in data science
3. How can we approach the problems
4. Next Action

# 1. Data analysis environment on cloud platform

# 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

# Why customers choose AWS



## Gartner recognizes AWS as a Leader for the 11<sup>th</sup> 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

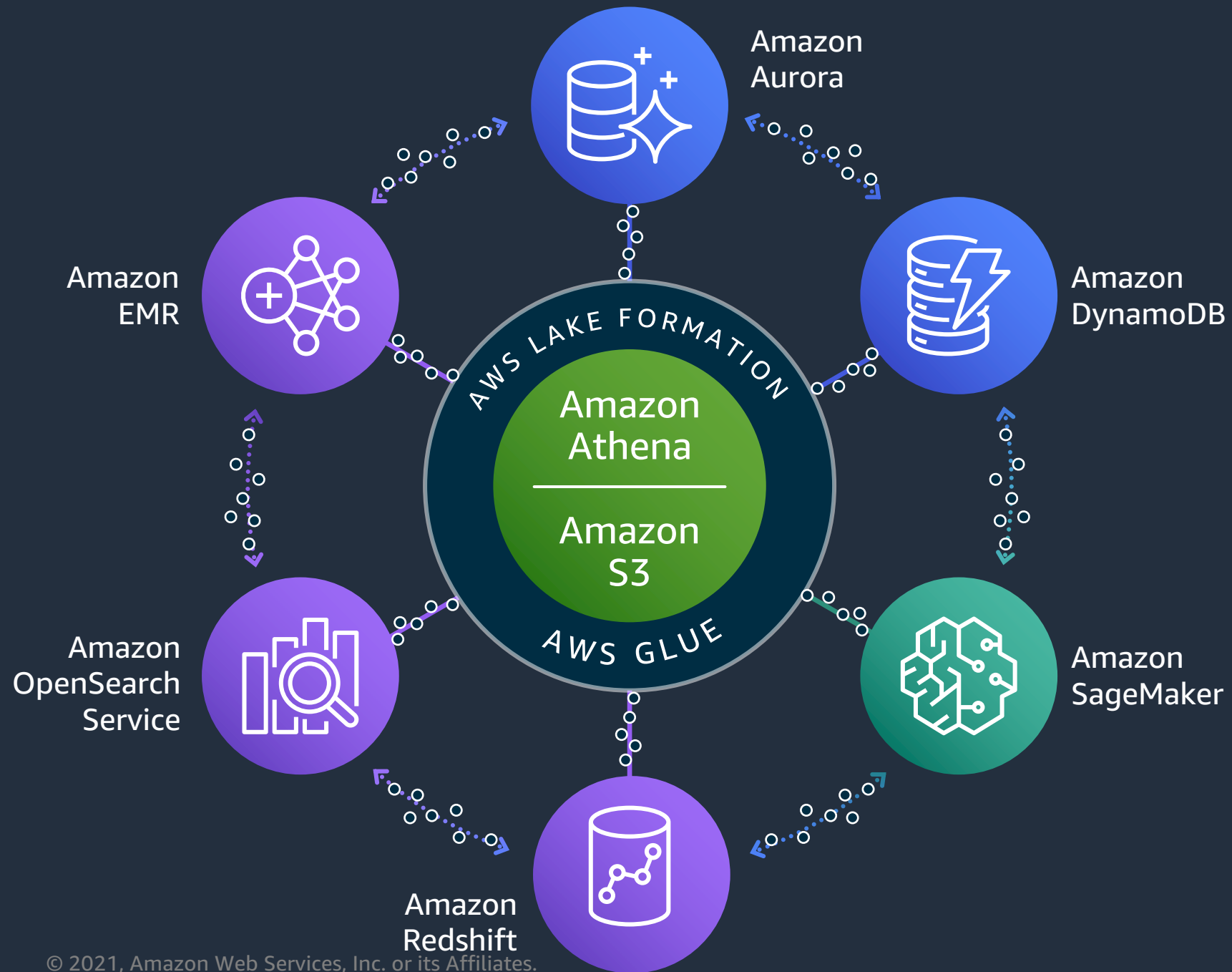
source: <https://aws.amazon.com/blogs/aws/aws-named-as-a-leader-for-the-11th-consecutive-year-in-2021-gartner-magic-quadrant-for-cloud-infrastructure-platform-services-cips/>

© 2021, Amazon Web Services, Inc. or its Affiliates.

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



# Lake House approach



Scalable Data Lakes

Purpose-built Analytics Services

Unified Data Access

Unified Governance

Performant and Cost-effective

# Daiichi Sankyo Co., LTD

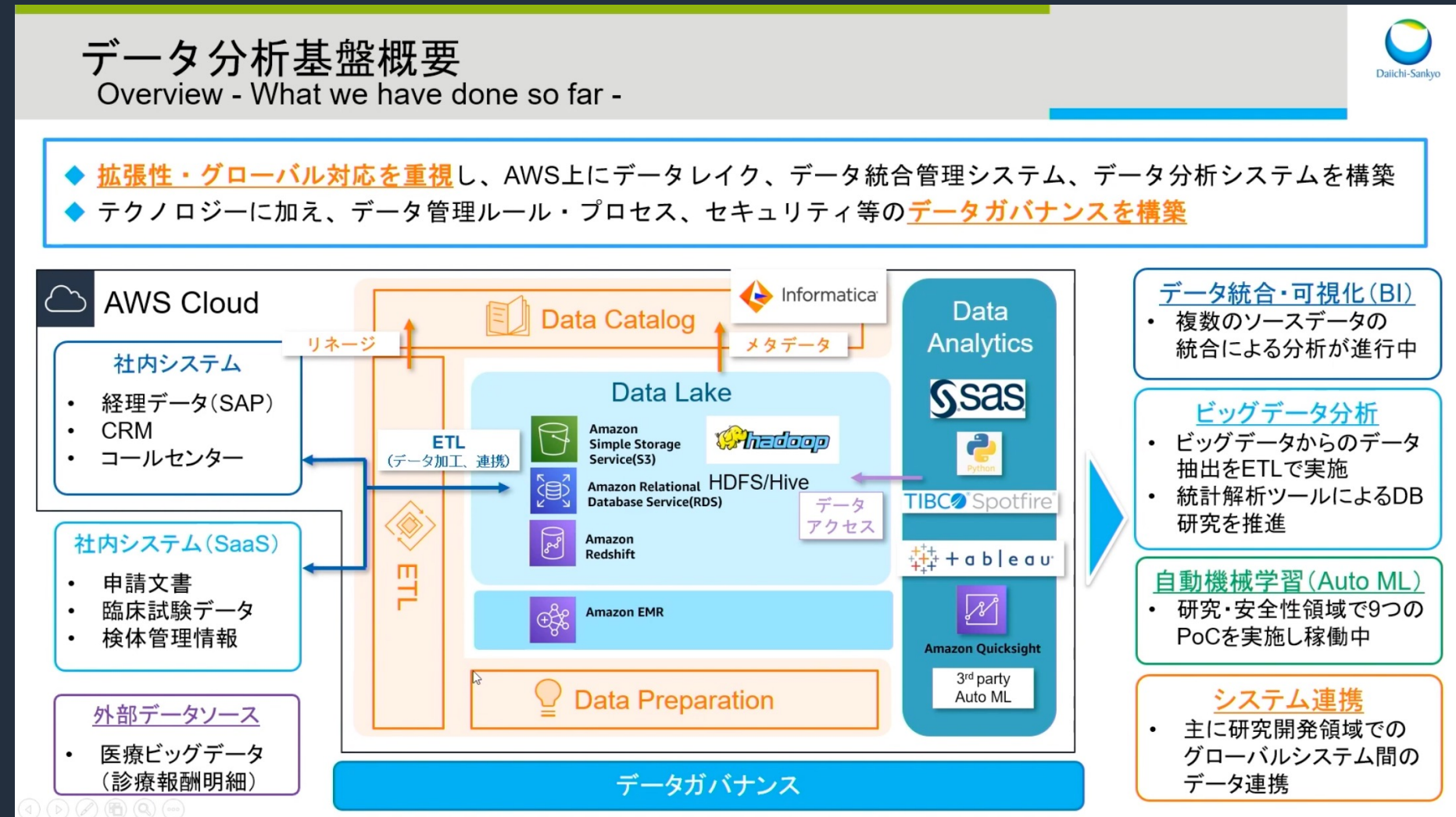
## Build company-wide data analysis environment

### 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.

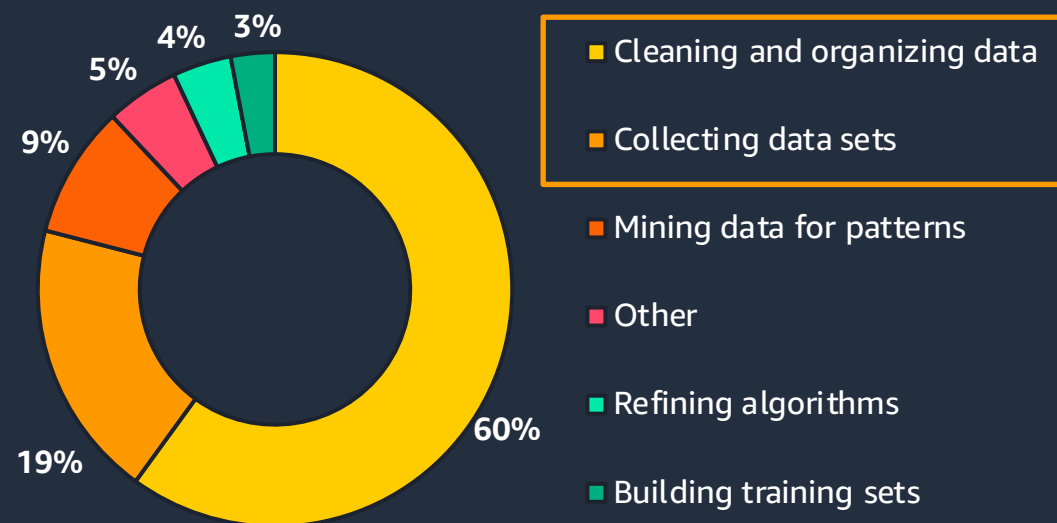
### Benefits

- 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.

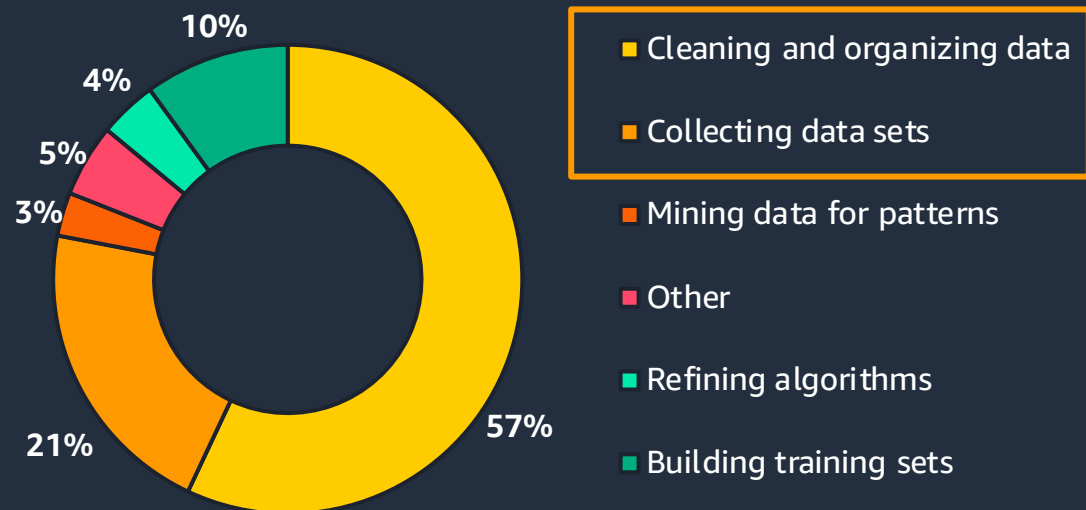
## 2. Common challenges in data science

# Data preparation still dominates data scientist's time

What data scientists spend the most time doing



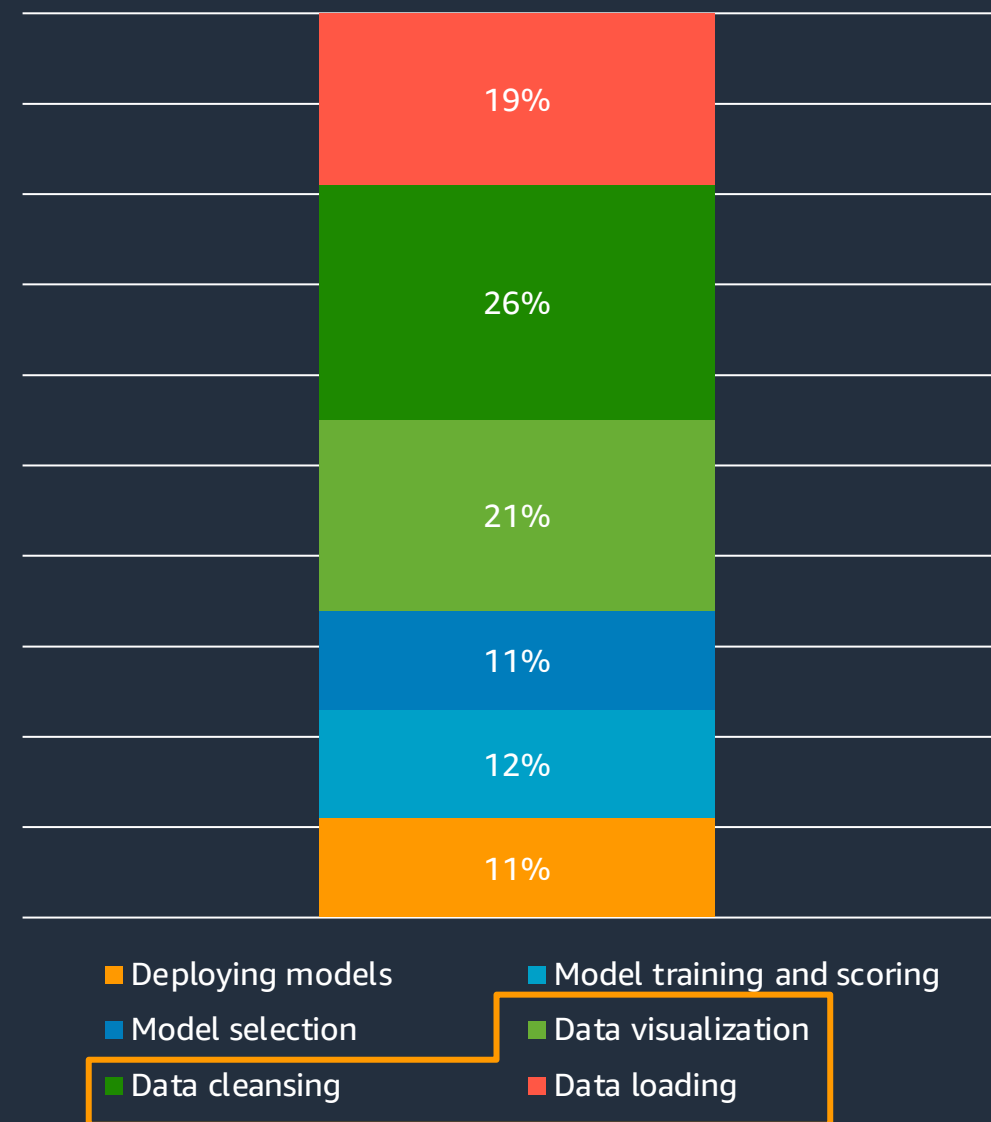
What's the least enjoyable part of data science



Source: [Forbes survey of 80 data scientists, March 2016](#)

© 2021, Amazon Web Services, Inc. or its Affiliates.

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



Source: [Anaconda survey of 2,360 respondents including data scientists, researchers, analyst and more, 2020](#)



# The need for collaboration

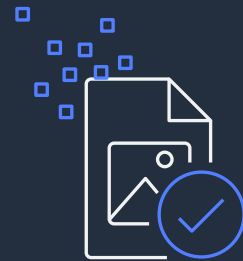
- **Changes in the working environment** under the COVID-19
- **Quick access** to raw data or the data which you want
- **Reduce** the **similar tasks**
- Gain **new insights** by **sharing your code** and **results**
- Enable data science and **business team to work**

# 3 Challenges in data science – this session



---

Collecting data



---

Cleaning and  
organizing data



---

Collaborating  
with your peers

# 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**

# 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

# 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

# 3. How can we approach the problems?

# Data pre-processing tools look like ...



# Amazon SageMaker Studio

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

The screenshot displays the Amazon SageMaker Studio interface. The main window shows a Jupyter notebook titled 'xgboost\_customer\_churn.ipynb' with the following content:

- Have the predictor variable in the first column
- Not have a header row

But first, let's convert our categorical features into numeric features.

```
[ ]: model_data = pd.get_dummies(churn)
      model_data = pd.concat([model_data['Churn?_True.'], model_data.drop(['Churn?_True.'], axis=1)], axis=1)
```

And now let's split the data into training, validation, and test sets. This will help prevent us from overfitting the model, and allow us to test the models accuracy on data it hasn't already seen.

```
[ ]: train_data, validation_data, test_data = np.split(model_data.sample(frac=1, random_state=42), [int(0.7*len(model_data)), int(0.1*len(model_data))])
      train_data.to_csv("train.csv", header=False, index=False)
      validation_data.to_csv("validation.csv", header=False, index=False)
```

Now we'll upload these files to S3.

```
[ ]: boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'train.csv')).upload_file(train_data.to_csv("train.csv", header=False, index=False))
      boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'validation.csv')).upload_file(validation_data.to_csv("validation.csv", header=False, index=False))
```

On the right side, there are two panels:

- Trial Component Chart:** A line chart showing 'trainloss\_last' on the y-axis (ranging from 0.0 to 0.4) against 'period' on the x-axis (ranging from 0 to 6). The chart displays several lines representing different trial components, showing a general downward trend in loss over time.
- Trial Component List:** A table listing trial components. It shows 10 rows selected. The table has columns for Status, Experiment, Type, Trial, and Trial component. All listed trials are 'Completed' and are 'Training job' types.

Status	Experiment	Type	Trial	Trial component
✓ Completed	customer-churn-predi...	Training job	Trial-3	Train
✓ Completed	customer-churn-predi...	Training job	Trial-2	Train
✓ Completed	customer-churn-predi...	Training job	Trial-1	Train
✓ Completed	customer-churn-predi...	Training job	Trial-0	Train

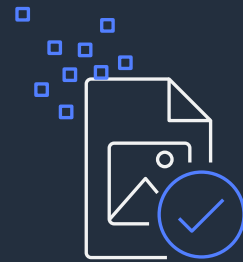
At the bottom of the interface, the status bar shows 'conda\_amazonei\_mxnet\_p27 | Idle', 'Mode: Command', and 'Ln 1, Col 1 xgboost\_customer\_churn.ipynb'.

# 3 Challenges in data science – this session



---

Collecting data



---

Cleaning and  
organizing data



---

Collaborating  
with your peers

# Challenge 1: Collecting Data

# 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

# 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

**Registry of Open Data on AWS**

**About**

This registry exists to help people discover and share datasets that are available via AWS resources. See [recent additions](#) and [learn more about sharing data on AWS](#).

See all [usage examples](#) for datasets listed in this registry tagged with life sciences.

**Search datasets (currently 87 matching datasets)**

Search datasets

You are currently viewing a subset of data tagged with life sciences.

**Add to this registry**

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

**The Cancer Genome Atlas**

[cancer](#) [genomic](#) [life sciences](#) [STRIDES](#) [whole genome sequencing](#)

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. TCGA has analyzed matched tumor and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. The dataset contains open Clinical Supplement, Biospecimen Supplement, RNA-Seq Gene Expression Quantification, miRNA-Seq Isoform Expression Quantificati...

[Details](#)

**Usage examples**

- [Broad Institute FireCloud by The Broad Institute of MIT & Harvard](#)
- [Using TCGA Data, Resources, and Materials by National Cancer Institute](#)
- [Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas by Theo A. Knijnenburg, Linghua Wang, et al.](#)
- [GDC Legacy Archive by National Cancer Institute](#)
- [Machine Learning Detects Pan-Cancer Ras Pathway Activation in The Cancer Genome Atlas by Gregory P. Way, Francisco Sanchez-Vega, et al.](#)

[See 29 usage examples](#)

**Therapeutically Applicable Research to Generate Effective Treatments (TARGET)**

[cancer](#) [genomic](#) [life sciences](#) [STRIDES](#) [whole genome sequencing](#)

Therapeutically Applicable Research to Generate Effective Treatments (TARGET) is the

**AWS Data Exchange**

[AWS Data Exchange](#) > [Browse catalog](#)

**Discover data products**

[Browse catalog](#)

My product offers

Request data product

**My subscriptions**

Subscriptions

Entitled data

Subscription requests

**Publish data**

Products

Subscription verification

Owned data sets

Documentation

**Refine results**

[Clear all filters](#)

**Categories**

[All categories](#)

Healthcare & Life Sciences Data (465)

**Vendors**

- Rearc (73)
- CE ResearchHub (68)
- iPatientAxis (45)
- Virtusa Corporation - vLife (36)
- Change Healthcare (30)

+46 others

**Data available through**

- Amazon S3 (463)
- Amazon Redshift (2)

**Contract type**

- Standard Data

**Browse catalog**

Search

Search

**Healthcare & Life Sciences Data (465 results)**

showing 1 - 36

Sort by most relevant

**Coronavirus (COVID-19) Data Hub**

Tableau

Coronavirus (COVID-19) data that has been gathered and unified from trusted sources. This data is provided to the public by Salesforce, MuleSoft, and Tableau at no cost to help you make better decisions, fast.

**Free**

36 month subscription available.

**Coronavirus Disease (COVID-19) Testing Data | The COVID Tracking Project**

Rearc

The COVID Tracking Project collects information from 50 US states, the District of Columbia, and 5 other US territories to provide the most comprehensive testing data

<https://registry.opendata.aws/> or its Affiliates.

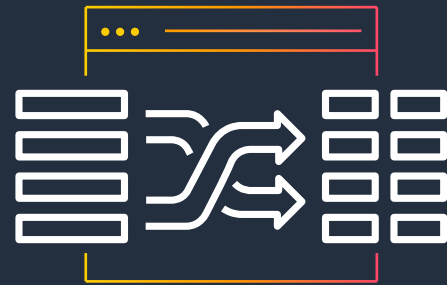
<https://aws.amazon.com/blogs/aws/aws-data-exchange-find-subscribe-to-and-use-data-products/>



# Prepare data with SageMaker Data Wrangler



Data selection



Data transforms



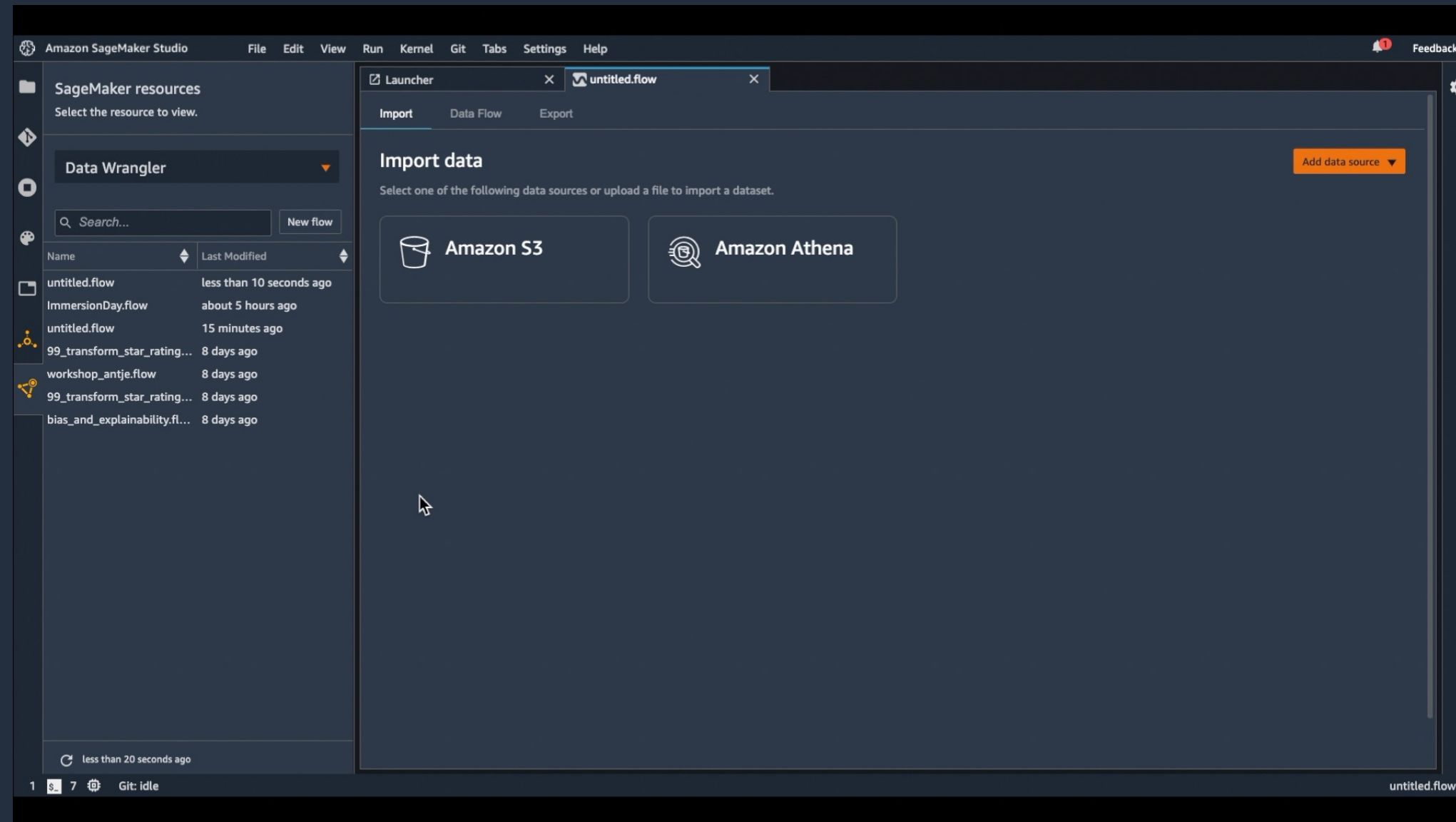
See data, spot inconsistencies, diagnose, and fix



Export your data

# 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



# Challenge 2: Cleaning and Organizing Data

# 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.

The screenshot displays the Amazon SageMaker Studio interface. On the left, the 'SageMaker resources' sidebar shows a list of flows, including 'untitled.flow' and '99\_transform\_star\_rating...'. The main workspace is titled 'Data types · Transform: bank-additional-full.csv' and is split into two panes: 'Data' and 'Analysis'. The 'Analysis' pane features a histogram titled 'Histogram: Age' with 'Count of Records' on the y-axis (0 to 200) and 'age(binned)' on the x-axis (10 to 90). The histogram bars are stacked by the variable 'y', with 'no' in blue and 'yes' in orange. Below the histogram is a 'Data table' with columns: age, job, marital, education, and default. The table shows rows of data, such as a 56-year-old housemaid who is married and has a basic.4y education level. On the right side of the interface, there is a 'Create analysis' panel with a dropdown menu set to 'Histogram', an analysis name field containing 'Age', and an X-axis dropdown set to 'age'. A legend for the histogram is visible in the top right corner of the plot area.

age	job	marital	education	default
56	housemaid	married	basic.4y	no
57	services	married	high.school	unknow
37	services	married	high.school	no
40	admin.	married	basic.6y	no
56	services	married	high.school	no
45	services	married	basic.6y	no

# Easily transform data

- 300+ built-in data transforms with GUI
  - Missing value detection
  - Outlier detection
  - Column manipulation
- Custom transforms in PySpark, SQL, and Pandas

The screenshot displays the Amazon SageMaker Studio interface. The main window shows a data flow titled "Data types · Transform: bank-additional-full.csv". The data is presented in a table with columns: age (long), job (string), marital (string), education (string), default (string), and housing (string). The data is sorted by age in descending order.

age (long)	job (string)	marital (string)	education (string)	default (string)	housing (string)
56	housemaid	married	basic.4y	no	no
57	services	married	high.school	unknown	no
37	services	married	high.school	no	yes
40	admin.	married	basic.6y	no	no
56	services	married	high.school	no	no
45	services	married	basic.9y	unknown	no
59	admin.	married	professional.course	no	no
41	blue-collar	married	unknown	unknown	no
24	technician	single	professional.course	no	yes
25	services	single	high.school	no	yes
41	blue-collar	married	unknown	unknown	no
25	services	single	high.school	no	yes
29	blue-collar	single	high.school	no	no
57	housemaid	divorced	basic.4y	no	yes
35	blue-collar	married	basic.6y	no	yes
54	retired	married	basic.9y	unknown	yes
35	blue-collar	married	basic.6y	no	yes
46	blue-collar	married	basic.6y	unknown	yes
50	blue-collar	married	basic.9y	no	yes
39	management	single	basic.9y	unknown	no

On the right side, the "ADD TRANSFORM" panel is visible, showing various transform options such as "Custom transform", "Custom formula", "Encode categorical", "Featurize date/time", "Featurize text", "Format string", "Group by", "Handle missing", and "Handle outliers". The "Encode categorical" option is currently selected.

# Deploy data preparation workflows into production

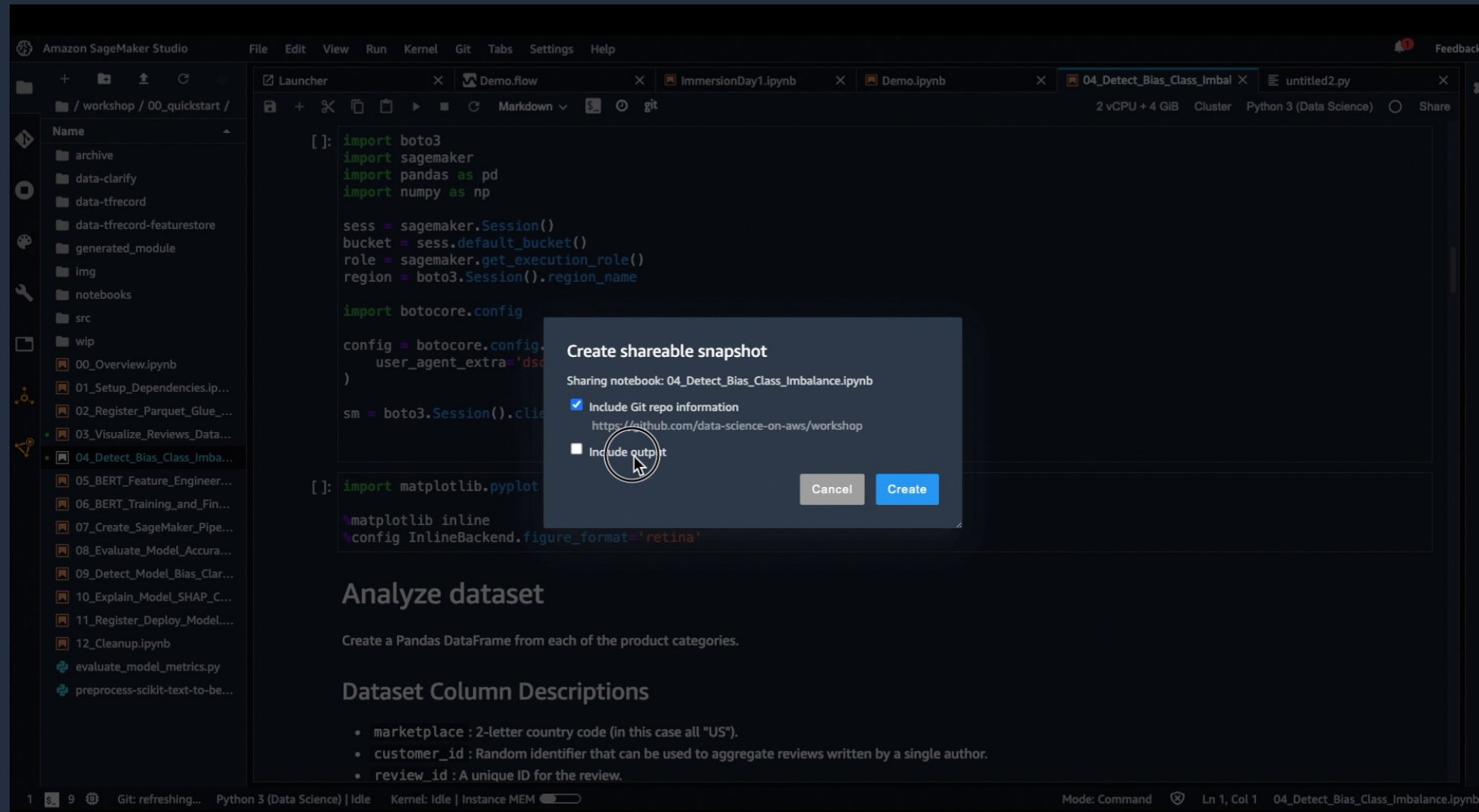
- Export data preparation workflows to a notebook or Python code
- Publish created features to SageMaker Feature Store for reuse and syndication across teams and projects

The screenshot shows the Amazon SageMaker Studio interface. On the left, a file explorer shows a project structure with folders like 'amazon-sagemaker-immers...', 'Sample', and 'workshop', and files like 'Demo.flow', 'Demo.ipynb', 'ImmersionDay.flow', 'ImmersionDay.ipynb', 'ImmersionDay1.ipynb', and 'untitled.py'. The main window is titled 'Demo.flow' and has tabs for 'Import', 'Data Flow', and 'Export'. The 'Export' tab is active, displaying the 'Export data flow' dialog. The dialog contains a data flow diagram with two steps: 'Source' (S3: bank-additional-full.csv) and 'Data types' (Transform: bank-addi full.csv). Below the diagram, there are three export options: 'Save to Amazon S3 (via Jupyter Notebook)', 'Run SageMaker Pipeline (via Jupyter Notebook)', and 'Python Code'. The 'Python Code' option is highlighted with a mouse cursor. At the bottom right, there are three checkboxes for selecting export options. The status bar at the bottom shows '1 \$ 8 Git: idle' and 'Demo.flow'.

# Challenge 3: Collaborating with your peers

# 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



The screenshot displays the Amazon SageMaker Studio interface. On the left, a file explorer shows a directory structure for a workshop. The main workspace contains a Jupyter notebook with Python code for setting up a Sagemaker session and importing libraries like boto3, pandas, and numpy. A modal dialog titled "Create shareable snapshot" is open, showing the notebook name "04\_Detect\_Bias\_Class\_Imbalance.ipynb" and options to include Git repository information (checked) and output (unchecked). The dialog also provides a GitHub link and "Cancel" and "Create" buttons. Below the code, the notebook content includes a section titled "Analyze dataset" with instructions to create a Pandas DataFrame, and a section titled "Dataset Column Descriptions" listing columns like marketplace, customer\_id, and review\_id.

```
[ ]: import boto3
import sagemaker
import pandas as pd
import numpy as np

sess = sagemaker.Session()
bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = boto3.Session().region_name

import botocore.config

config = botocore.config.Config(
    user_agent_extra='ds'
)

sm = boto3.Session().client
```

```
[ ]: import matplotlib.pyplot as plt

%matplotlib inline
%config InlineBackend.figure_format='retina'
```

**Create shareable snapshot**

Sharing notebook: 04\_Detect\_Bias\_Class\_Imbalance.ipynb

- Include Git repo information  
https://github.com/data-science-on-aws/workshop
- Include output

Cancel Create

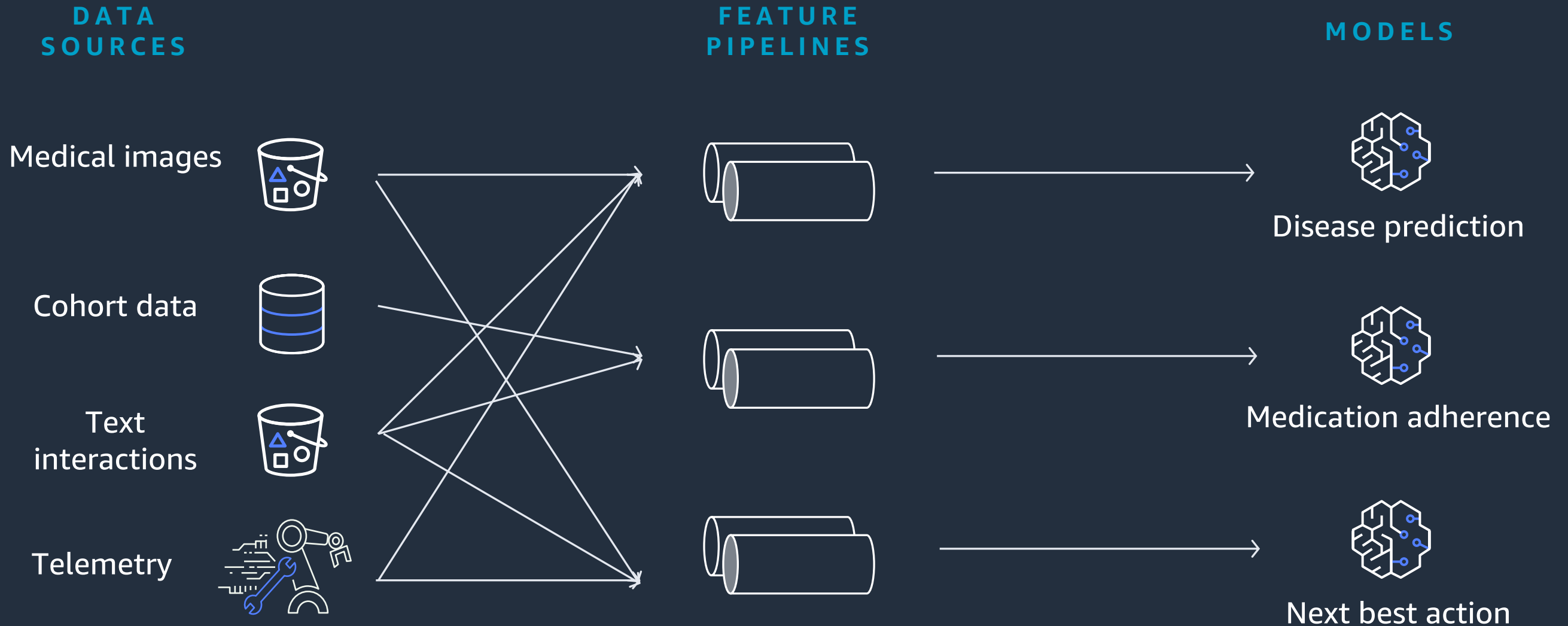
### Analyze dataset

Create a Pandas DataFrame from each of the product categories.

### Dataset Column Descriptions

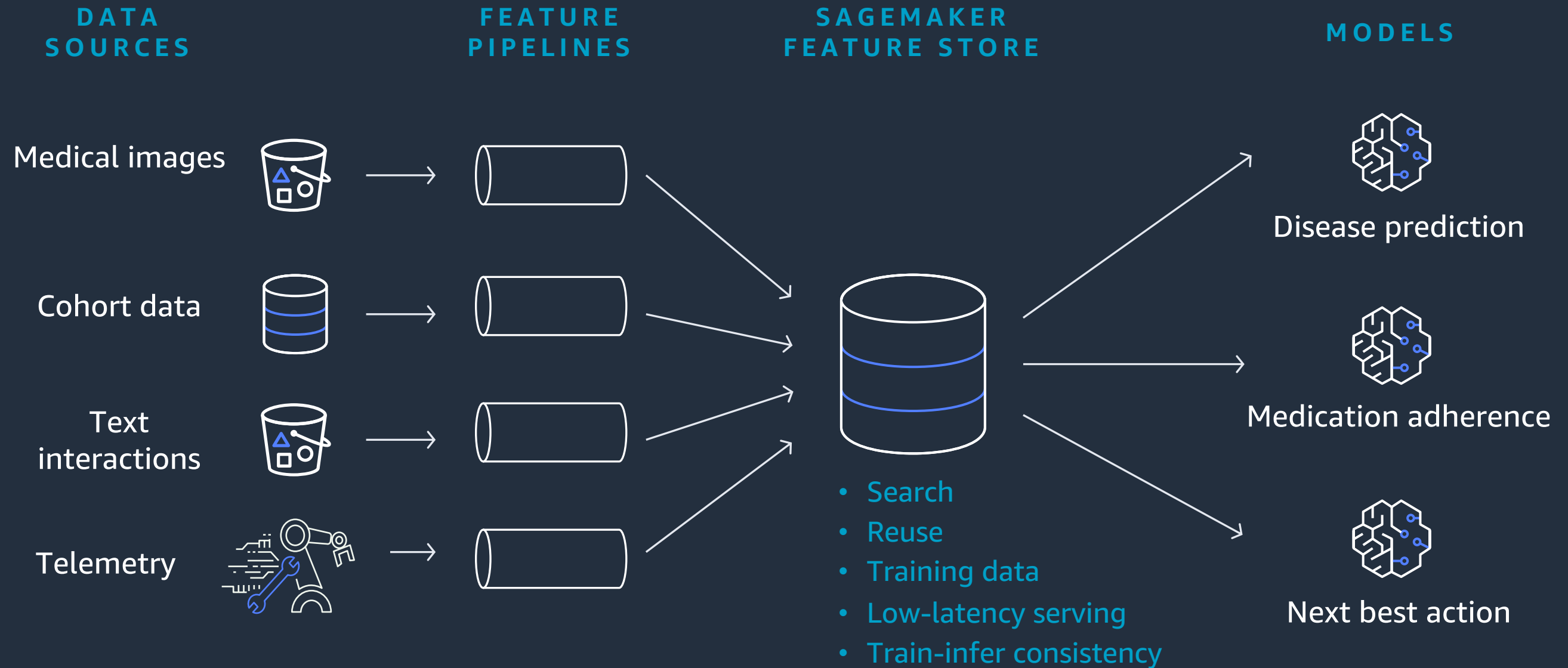
- marketplace : 2-letter country code (in this case all "US").
- customer\_id : Random identifier that can be used to aggregate reviews written by a single author.
- review\_id : A unique ID for the review.

# Without SageMaker Feature Store ...



# With SageMaker Feature Store ...

Build features once, and reuse them across teams and models



# 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*

# Customer Spotlight

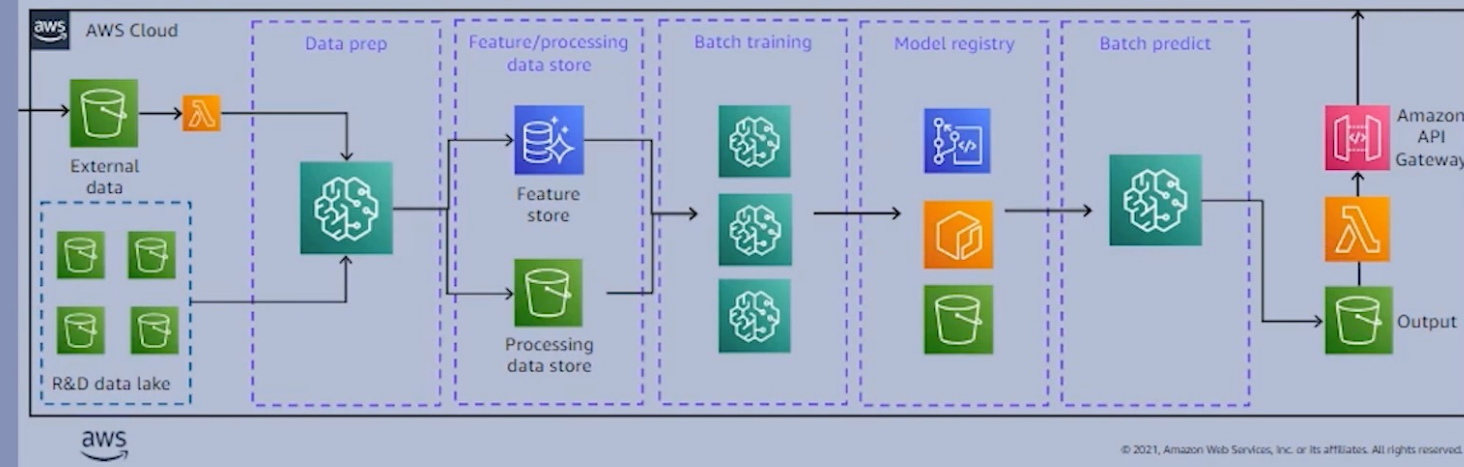
- AI Bench 2.0
  - Data science & machine learning platform
  - Single-tenant multi accounts
  - GxP-compliant system
- Using SageMaker Studio and SageMaker Feature Store for train/build models



**Rui Wang**  
Head of Compute and Core Engineering, R&D IT Data & Analytics, AstraZeneca

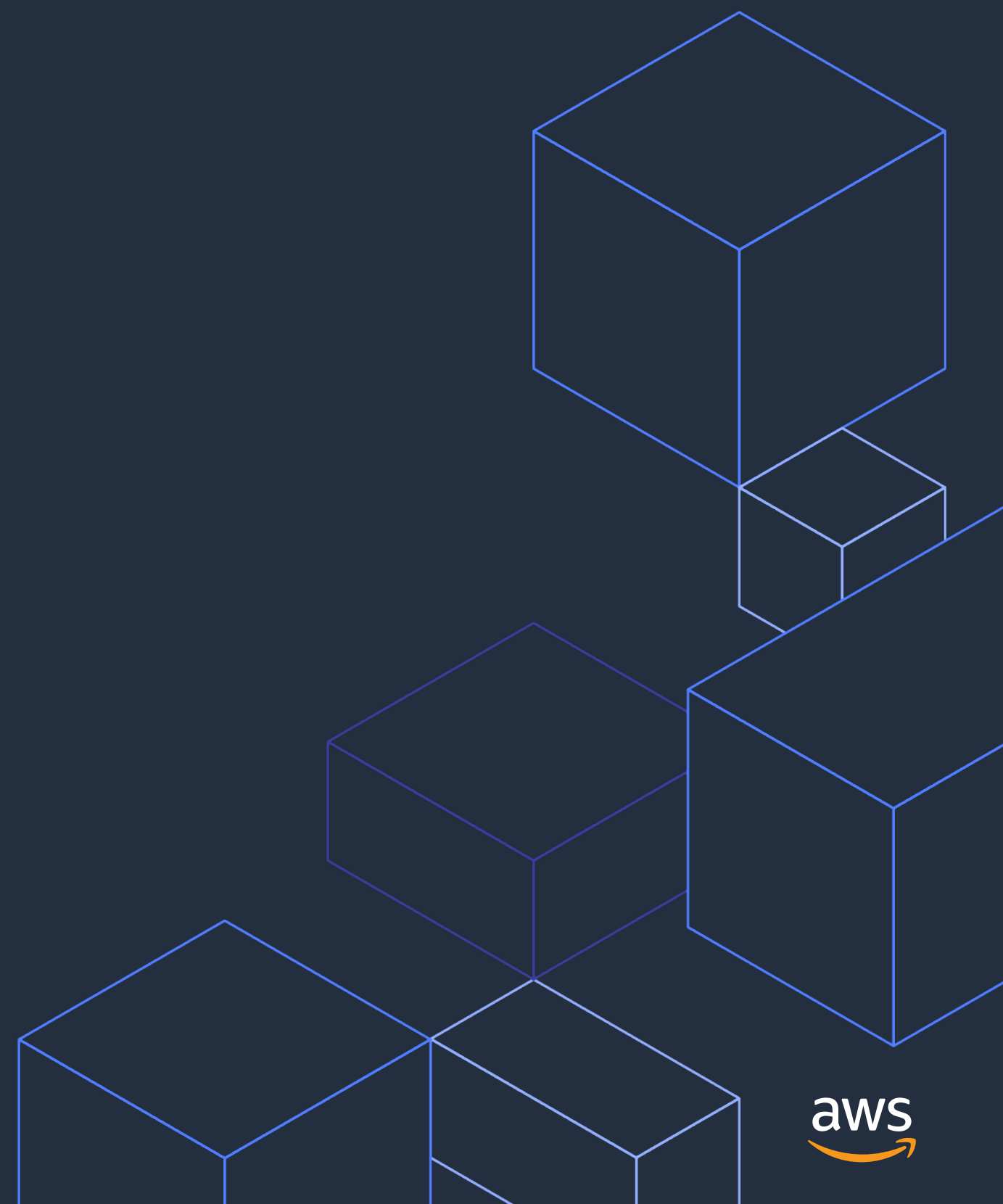
## Example: NLP for literature surveillance on AI Bench

- ML models were more sensitive than manual review – less chance of missing true signal
- Precision of the ML models increases with learning
- The ML models can potentially improve scalability, standardization, and trackability (vs. manual review)



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

# 4. Next Action



# Getting Started with Amazon SageMaker Studio

- Visit [console.aws.amazon.com/sagemaker](https://console.aws.amazon.com/sagemaker)
- Create your SageMaker Studio Domain
- Open Studio to get started

Add multiple users for SageMaker Studio

The screenshot shows the Amazon SageMaker console interface. On the left is a navigation sidebar with categories like Dashboard, Search, SageMaker Domain, Studio, Images, Ground Truth, Notebook, Processing, Training, Inference, Edge Manager, Augmented AI, and AWS Marketplace. The main content area is titled 'SageMaker Domain' and contains a 'Users' section with a search bar and a table. The table has columns for Name, Modified on, and Created on. One user is listed with the name 'default-...', modified on 'Oct 31, 2021 06:38 UTC', and created on 'Oct 31, 2021 06:38 UTC'. An 'Open Studio' button with an external link icon is next to this user. Below the table is a 'Domain' section with buttons for 'How to delete Studio', 'Delete Studio', and 'Edit Settings'. The 'Domain' section includes details for Status (Ready), Studio ID, Execution role, and Authentication method (IAM). A large orange callout box on the right says 'Open SageMaker Studio' with an arrow pointing to the 'Open Studio' button. Another orange callout box on the left says 'Add multiple users for SageMaker Studio' with an arrow pointing to the 'Users' section.

- 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/](https://aws.amazon.com/sagemaker/healthcare-life-sciences/)
- SageMaker Studio Workshop (in Japanese)  
[sagemaker-immersionday.workshop.aws/ja/](https://sagemaker-immersionday.workshop.aws/ja/)

Please feel free to contact us for further information.

# Thank you!

