



# Democratizing Forecasting and Analytics

Dan Sinnreich

Sr. Product Manager

Henrik Balle

Sr. Solutions Architect

February 28, 2023

# Common questions in public health where machine learning (ML) can help drive more informed decisions

Where do I need to redirect vaccine to address the increase in cases across the country?

What population has the highest risk factors for a disease?

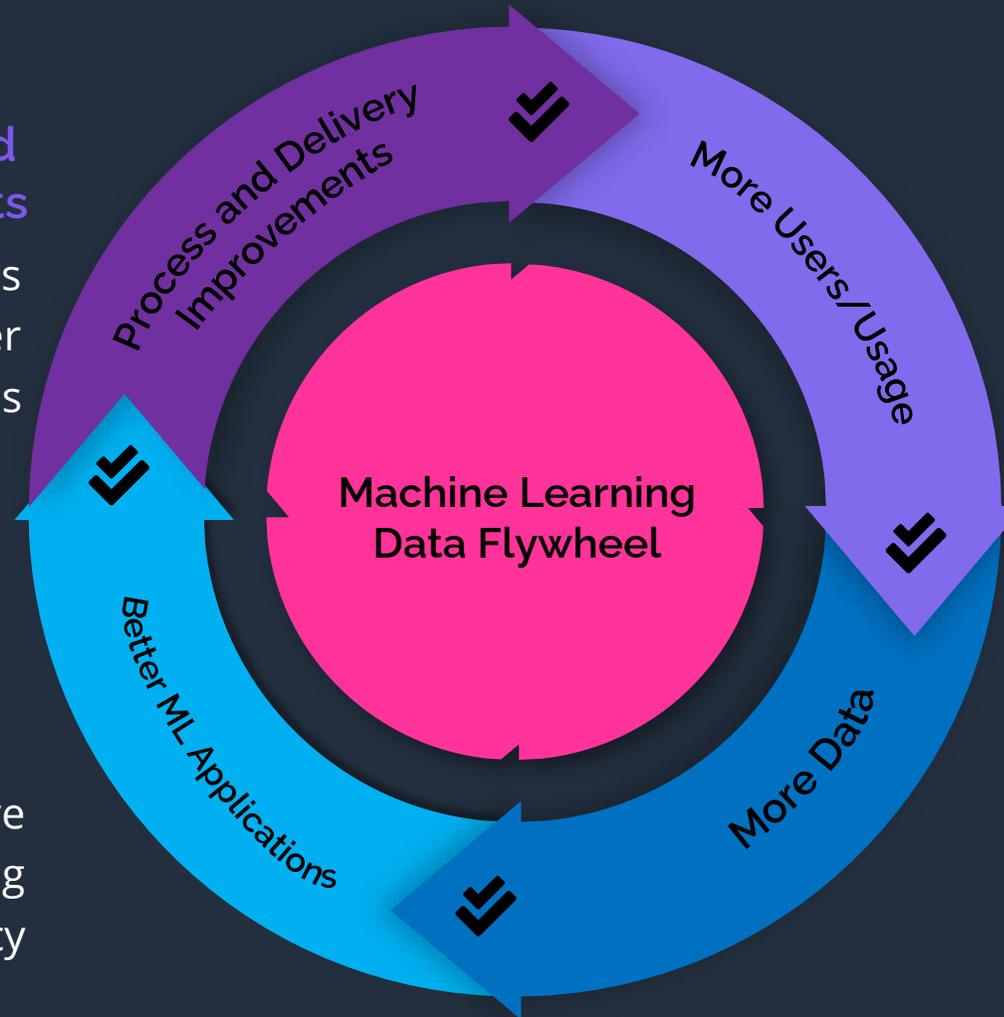
Which risk groups to target given limited resources?

When will this outbreak peak in my at-risk population??

What is the likelihood of severe illness or death among vulnerable populations?

# ML provides repeatable value

**Process and Delivery Improvements**  
ML powered insights and applications improve decisions and lead to better public health outcomes

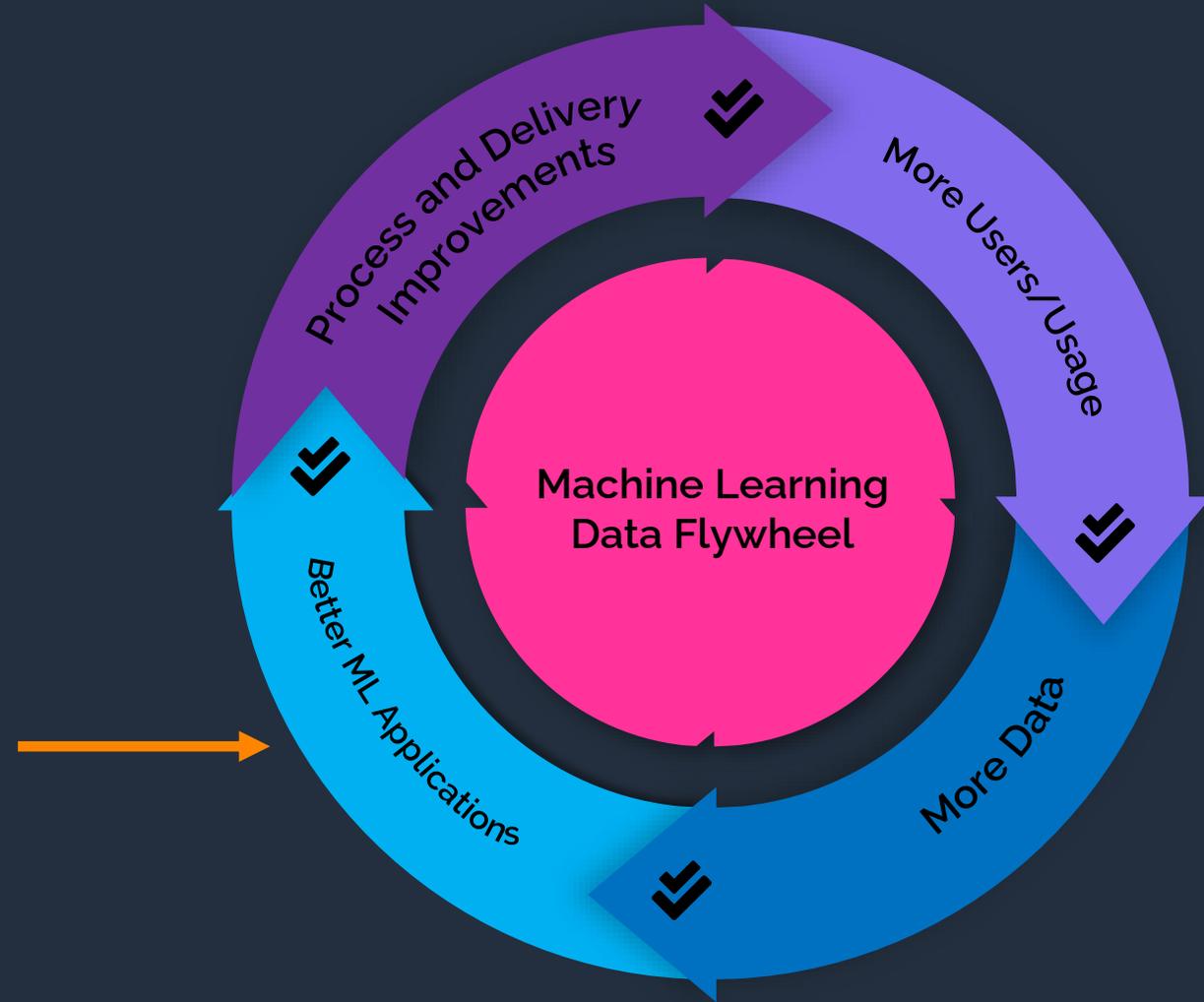


**More Users/Usage**  
Having more users or more data for your current users to consume

**Better ML Applications**  
Building more models to solve more use cases and improving the existing models accuracy

**More Data**  
With more data available, there are more use cases and insights that ML can provide valuable impact

# Data science resources are a limiting factor

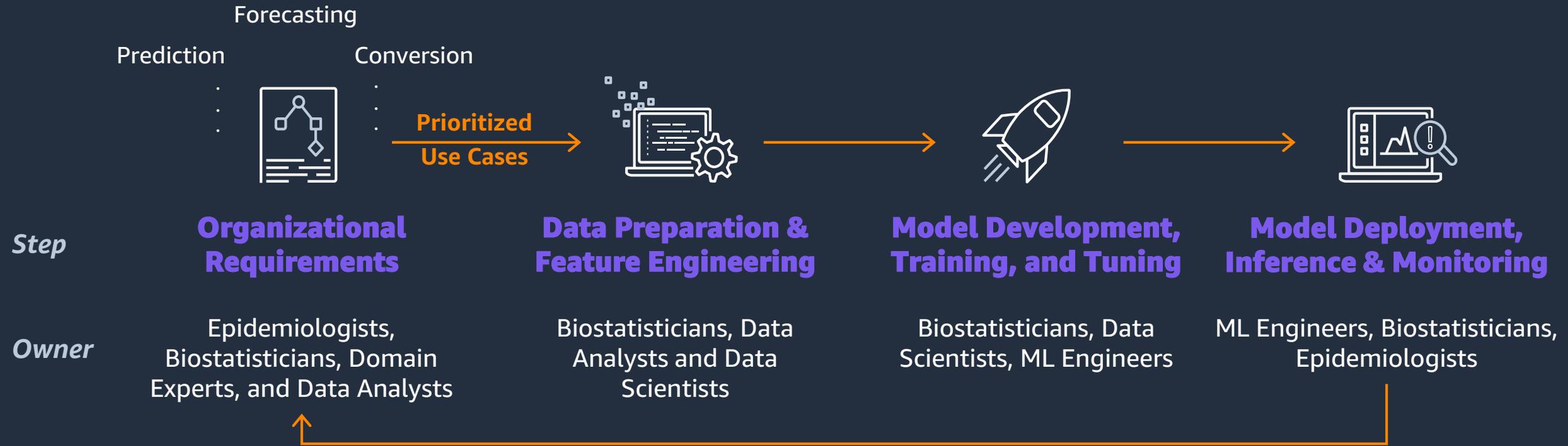


## The Bottleneck

Building better ML applications can be the bottleneck because:

- Data Science talent is expensive to scale
- Friction points in the Machine Learning journey
- Robust data is not available

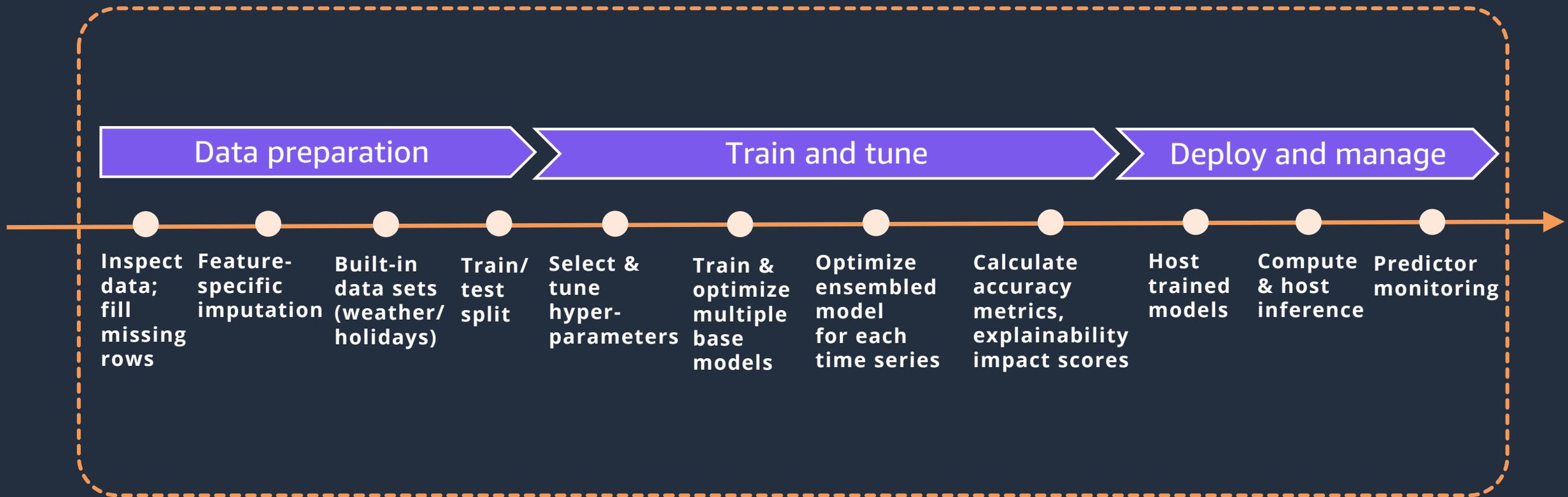
# How ML drives value creation today



Usually takes from weeks to months primarily solving for the prioritized use cases

# Typical steps in an ML process

Prepare, build, train, tune, deploy, and manage



# How can you scale ML value creation?

1

## Expand Your ML Development Team

Grow your technical teams in proportion of your needs, but ML talent is in high demand

+74%

annual compound growth in past 4 years

2x

the demand growth of any other emerging job role

2

## Enhance ML Team Productivity

Leverage low-code / no-code tools that make data science teams more productive

Enable data science teams to experiment faster with low-code / no-code Machine Learning capabilities

3

## Democratize ML Innovation

Enable more groups of people, including data analysts to build ML models

Empower data analysts to make smarter decisions with no-code Machine Learning with a dedicated easy-to-use workspace

# Challenges analysts face in building ML solutions



## Analysts lack deep ML expertise, and learning curve is steep

---

- Need to build understanding for ML concepts across data preparation, model development, and optimization
- Need expertise in choosing the right combination of feature engineering, type of model, and optimization technique
- Learning to write or decipher code is usually needed



## Agencies need explainability and validation from experts

---

- Analysts prefer to partner with data scientists in order to learn and build trust in the process, but data scientists time is limited and typically devoted to a few key ML projects
- Analysts need to be able to explain ML model predictions to leadership

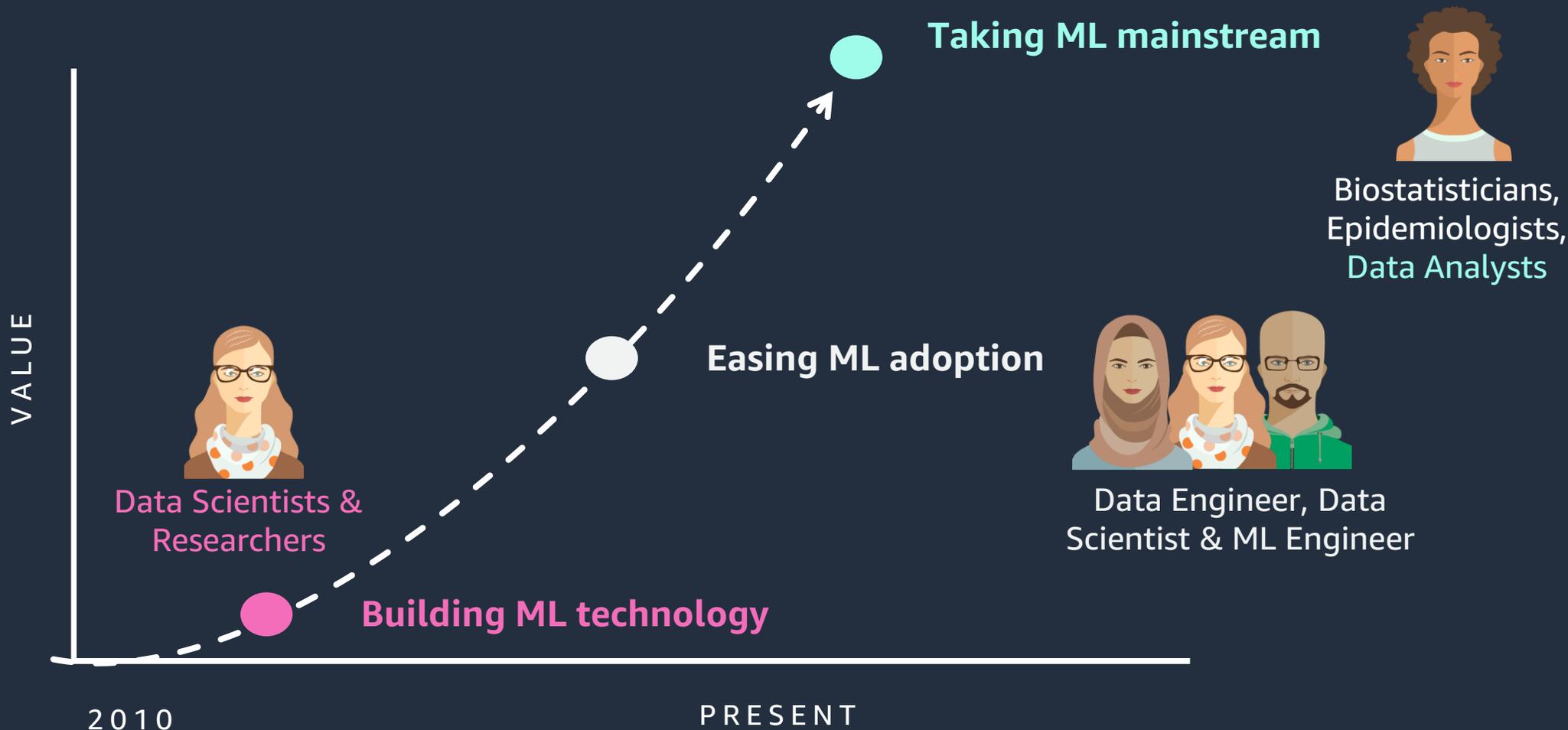


## Available no-code ML tools tend to lack transparency and be cost prohibitive

---

- Many no-code ML options lack code-level transparency making it difficult to inspect and productionalize models
- The UX for analysts and data scientists tends to be the same, requiring analysts to know the ML concepts and jargon
- Frequently, no-code ML tools come with upfront costs, so even experimentation requires investment

# New technologies are helping to democratize ML



# Keys to unblocking the analyst



**Analysts lack ML expertise,  
and upskilling is hard**



**Agencies need transparency  
and validation from experts**



No-code is the future

In a June 2021 report, Gartner found that by 2024, 80% of tech products and services will be built by people who are not technology professionals

<https://gtnr.it/3wOglym>



Seamless collaboration



Biostatisticians,  
Epidemiologists,  
Data Analysts

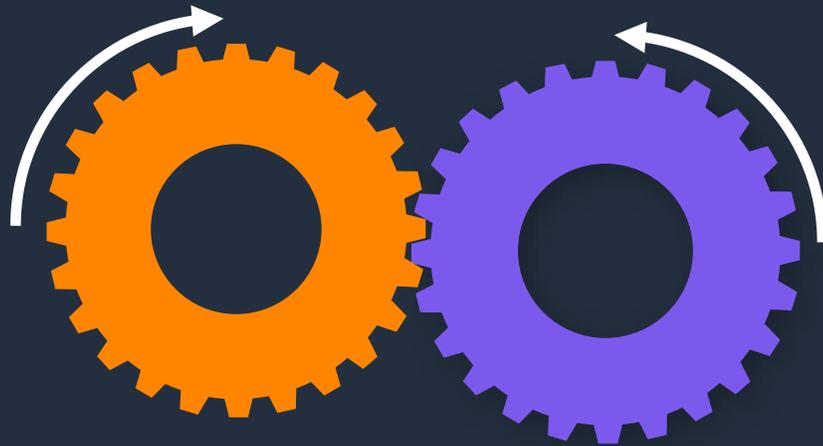


Data Engineer, Data Scientist  
& ML Engineers

# Our model for ML-driven forecasts and analyses

## Accelerate Data Science Teams

Do more with your current team by using low-code machine learning tools in order to get to the desired outcomes faster



## Enable Non-Technical Users

Give leaders and data analysts the ability to do ML without any code, scaling the number of people who can create ML powered insights, forecasts, and predictions

## Collaborate together

Amazon SageMaker Low Code / No Code ML has several points of collaboration making it easy for analysts to use data scientist models or for data scientists to make changes on the models analysts build.

# Realizing ML vision using Amazon SageMaker



**Epidemiologists,  
Data Scientists and  
Researchers**

*Want to develop, deploy the best performing model, monitor performance and data governance*



**ML Engineers and  
Biostatisticians**



**Data Engineers  
and Scientists**

*Need single pane of glass to manage data and ML*



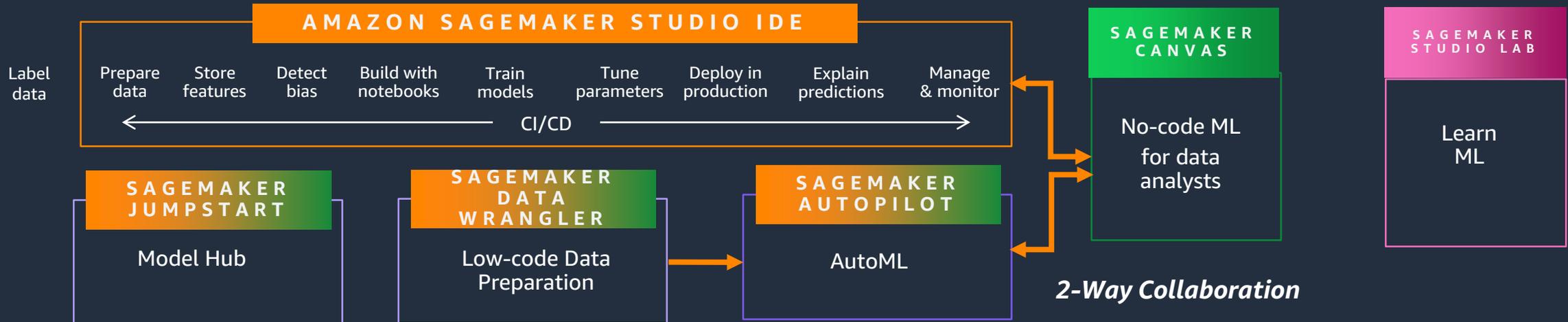
**Data Analyst**

*Need easy way to translate domain knowledge to predictions and insights*



**New ML Developers**

*Need ready-to-use ML development environment*



# AWS low code / no code (LCNC) ML ecosystem

## Amazon SageMaker Canvas

A visual interface for **data analysts** to generate ML-powered predictions — without requiring any machine learning experience or having to write a single line of code.

Classifications

Regressions

Time-Series Forecasts



## Amazon SageMaker Studio

A dedicated workspace for **data scientists, developers, and data engineers** to collaborate and bring ML to market faster

### Data Wrangler

A faster, visual way to aggregate and prepare data for machine learning

### Autopilot

AutoML to automatically prepare your data, and build, train, and tune the best machine learning models for your tabular datasets

### JumpStart

Pre-built solutions and a model zoo of pre-trained and easily tunable state-of-the-art models for Computer Vision, and Natural Language Processing

## Deployment

Many deployment options in SageMaker

Model governance

Production workflows

Monitor & adjust

Integration with data layers and BI tools



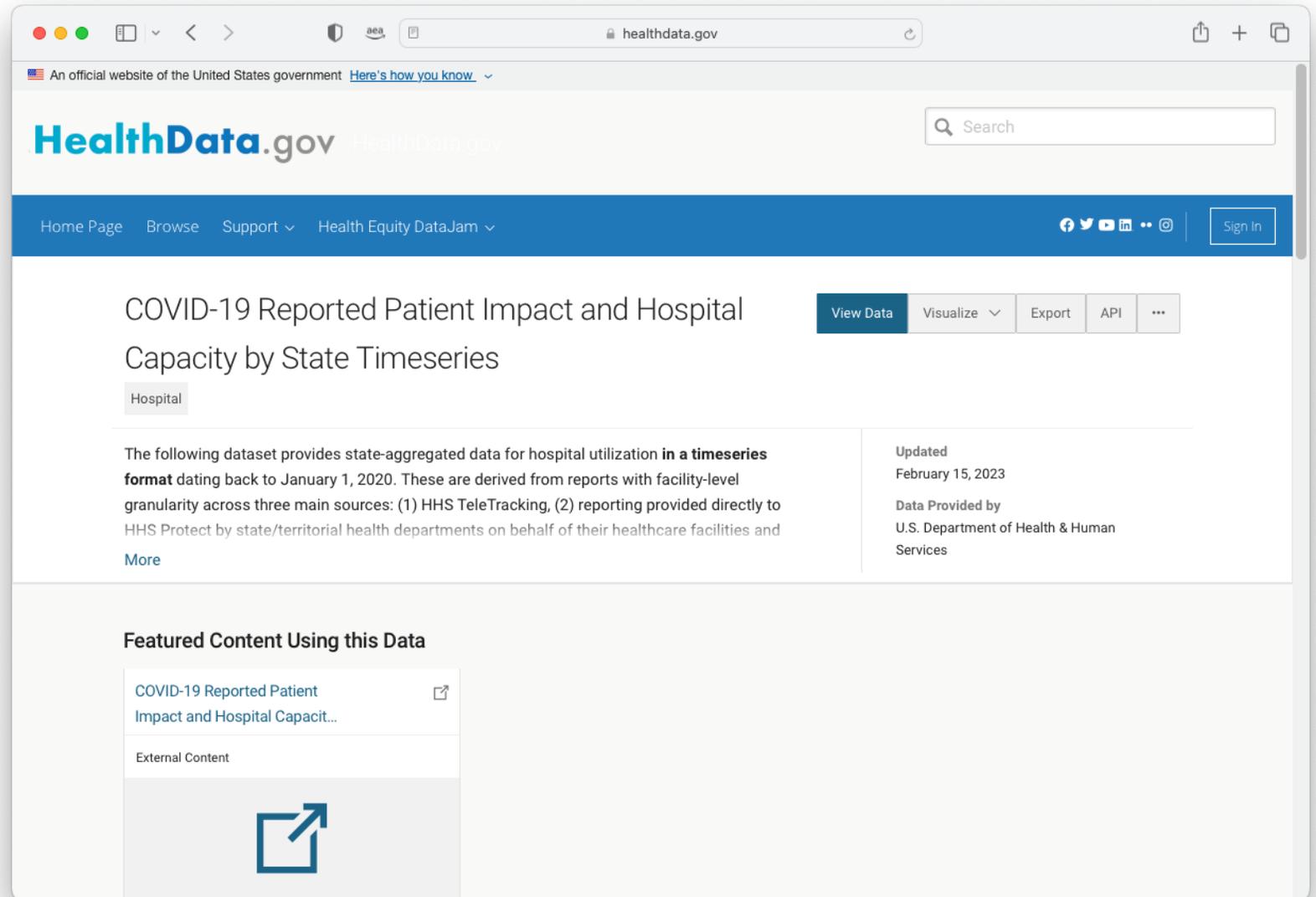
**Non-technical Users**

+

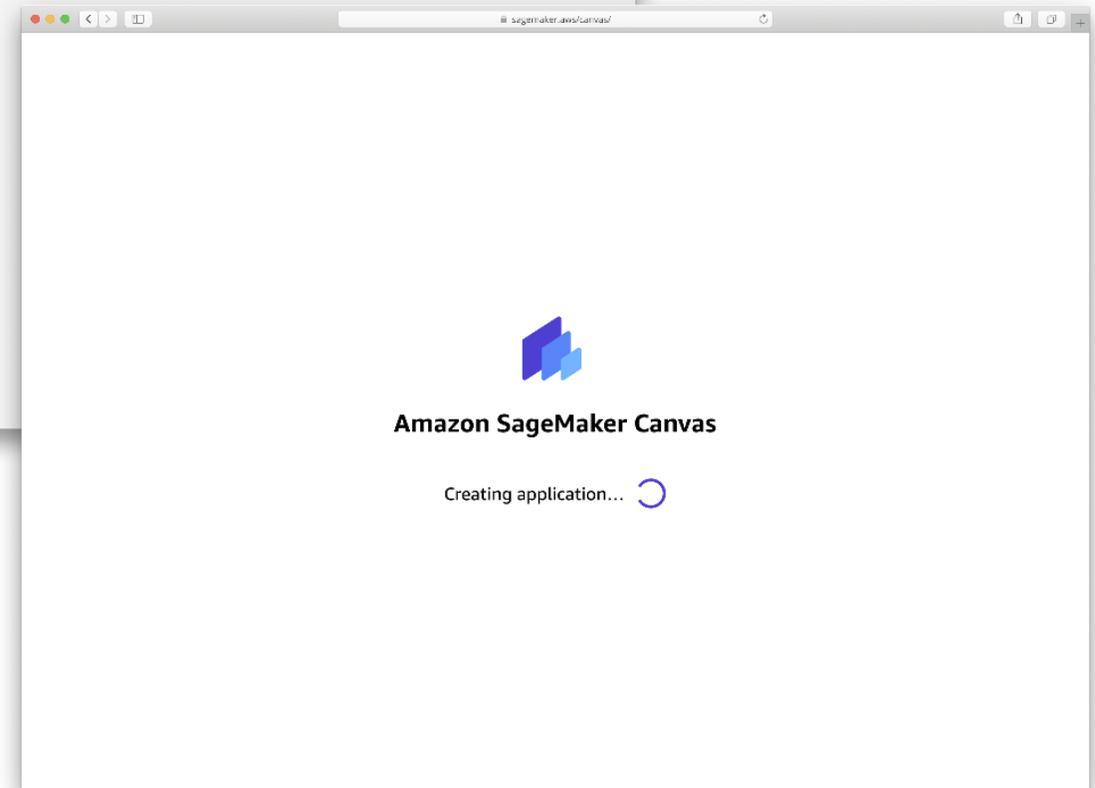
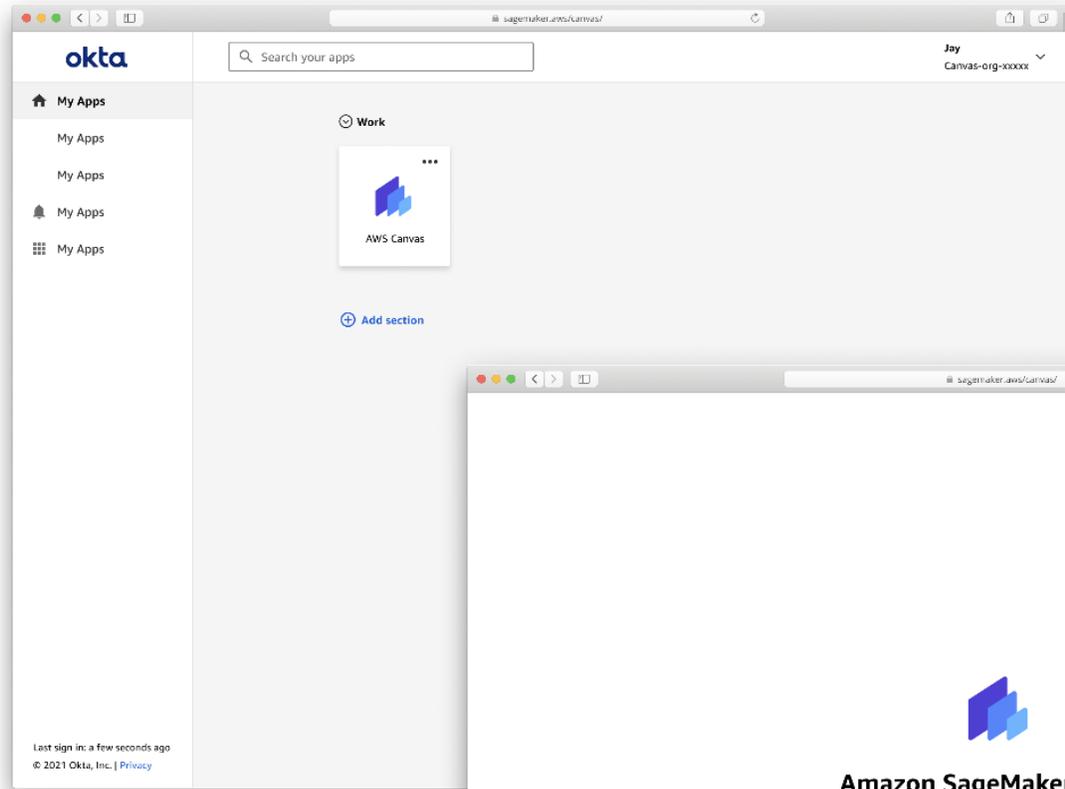
**Data Scientists and Developers**



We're going to forecast on-hand inventory of therapeutics for COVID-19



# Self-service access to an analyst-friendly tool for Machine Learning, outside of the AWS console



We could combine datasets from various sources like AWS storage, Snowflake, and databases, Redshift, or your local drive

**Import Data**

Upload S3 Snowflake Crystal 1 Redshift Crystal 1 Add Connection

Connection name Context

Search

- database1
- database2
- database3
- database4
  - schema1
  - schema2
    - table1

Autosaved 8/9/21 at 11:34 AM Edit in SQL

table1.csv table2.csv

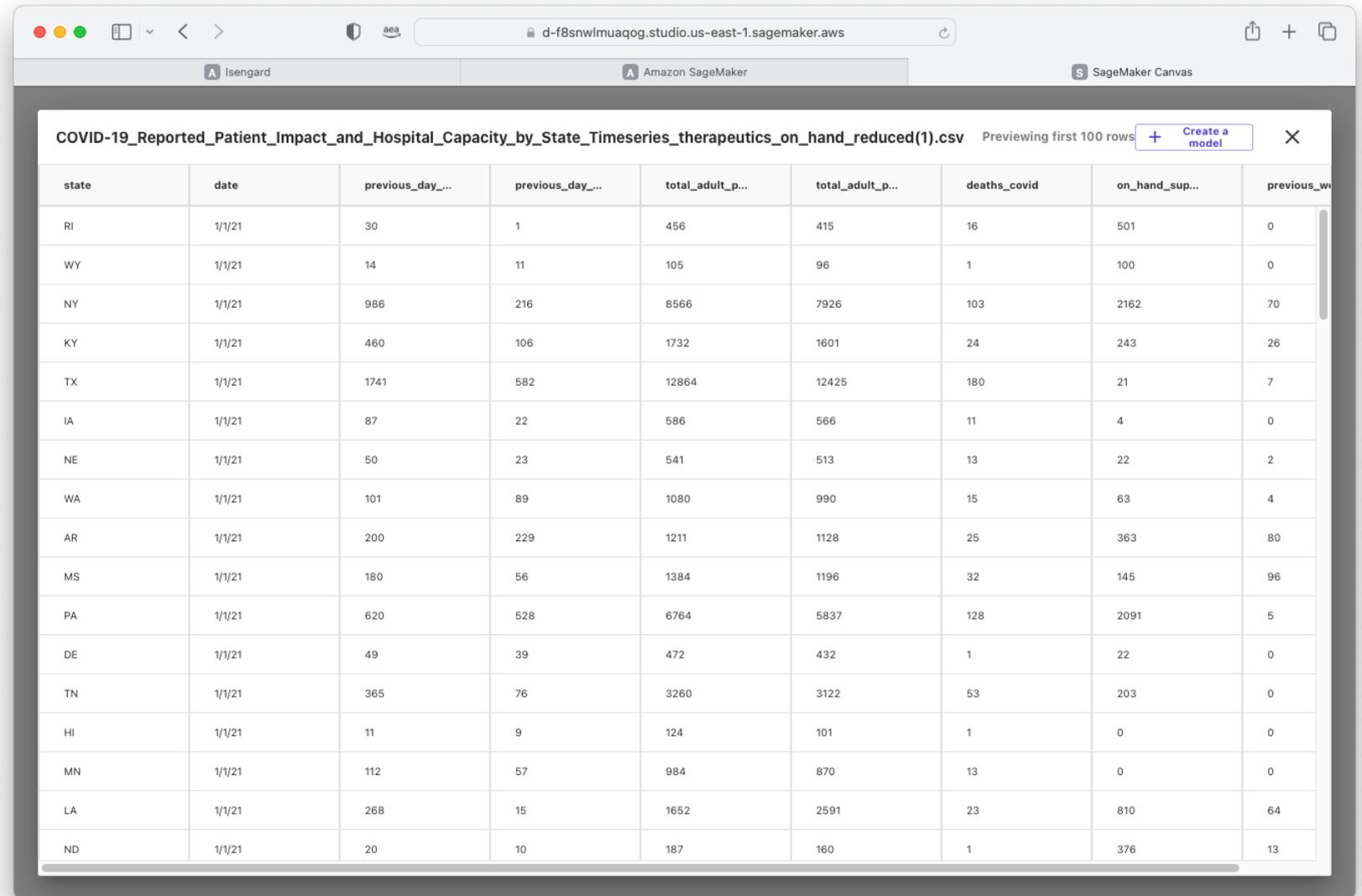
**Import preview** Show dropped columns

<input checked="" type="checkbox"/> Sold	ABC	<input type="checkbox"/> Price	123	<input checked="" type="checkbox"/> Region	ABC	<input checked="" type="checkbox"/> Discount	123	<input checked="" type="checkbox"/> Fabric	ABC	<input checked="" type="checkbox"/> Age	123
Yes		29.99		Southwest		23		Cotton		27	
Yes		29.99		Southwest		23		Silk		35	
Yes		29.99		Southwest		23		Silk		32	
Yes		29.99		Southwest		23		Silk		32	
Yes		29.99		Southwest		23		Cotton		30	

Previewing the first 100 rows Close Import data



We are uploading the dataset as-is with a reduced set of features and rows

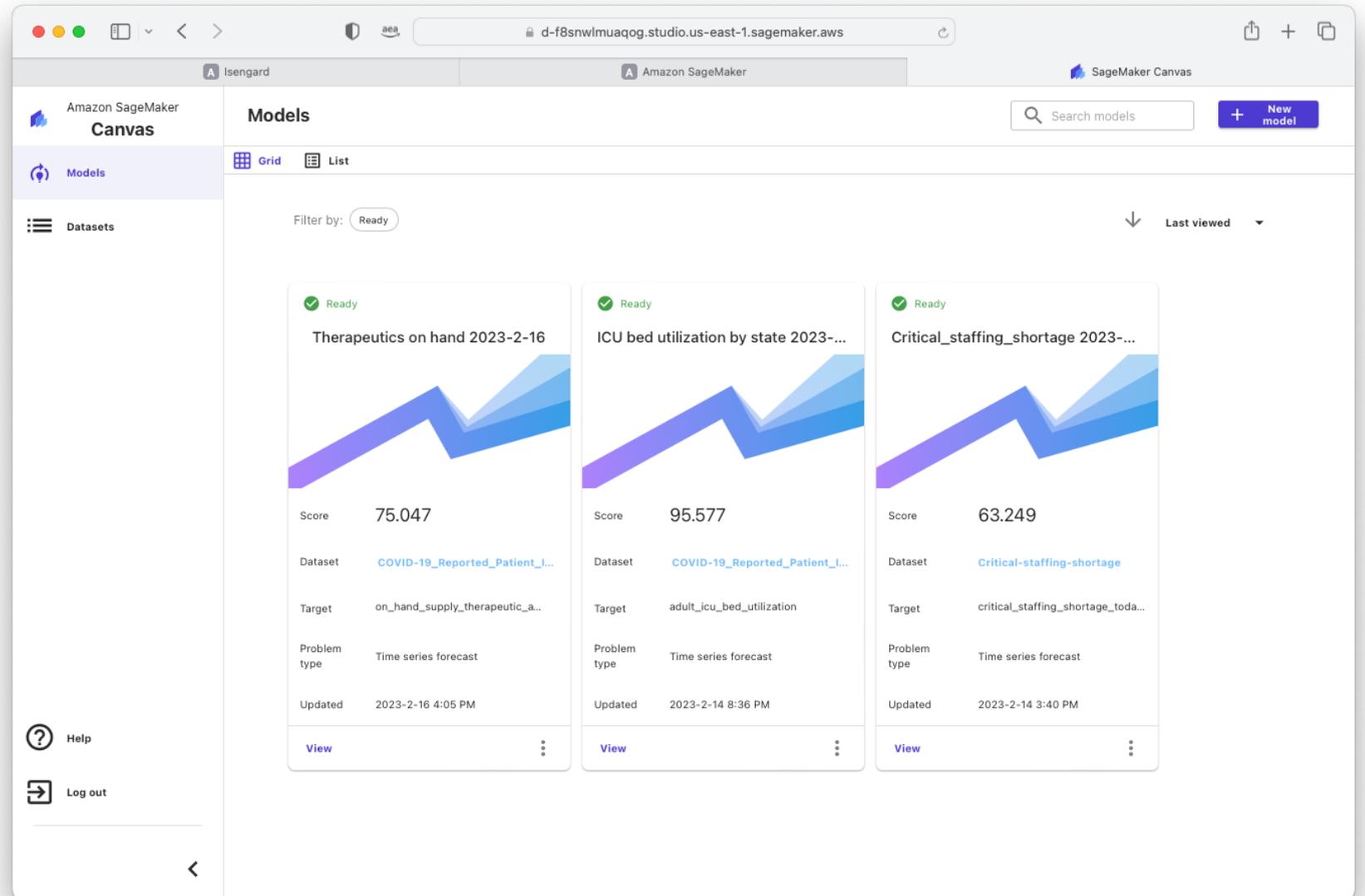


The screenshot shows the Amazon SageMaker Canvas interface. The browser address bar displays the URL: `d-f8snwlmuaqog.studio.us-east-1.sagemaker.aws`. The page title is `COVID-19_Reported_Patient_Impact_and_Hospital_Capacity_by_State_Timeseries_therapeutics_on_hand_reduced(1).csv`. A button labeled `Create a model` is visible in the top right corner of the preview area. The table below shows the first 10 rows of the dataset.

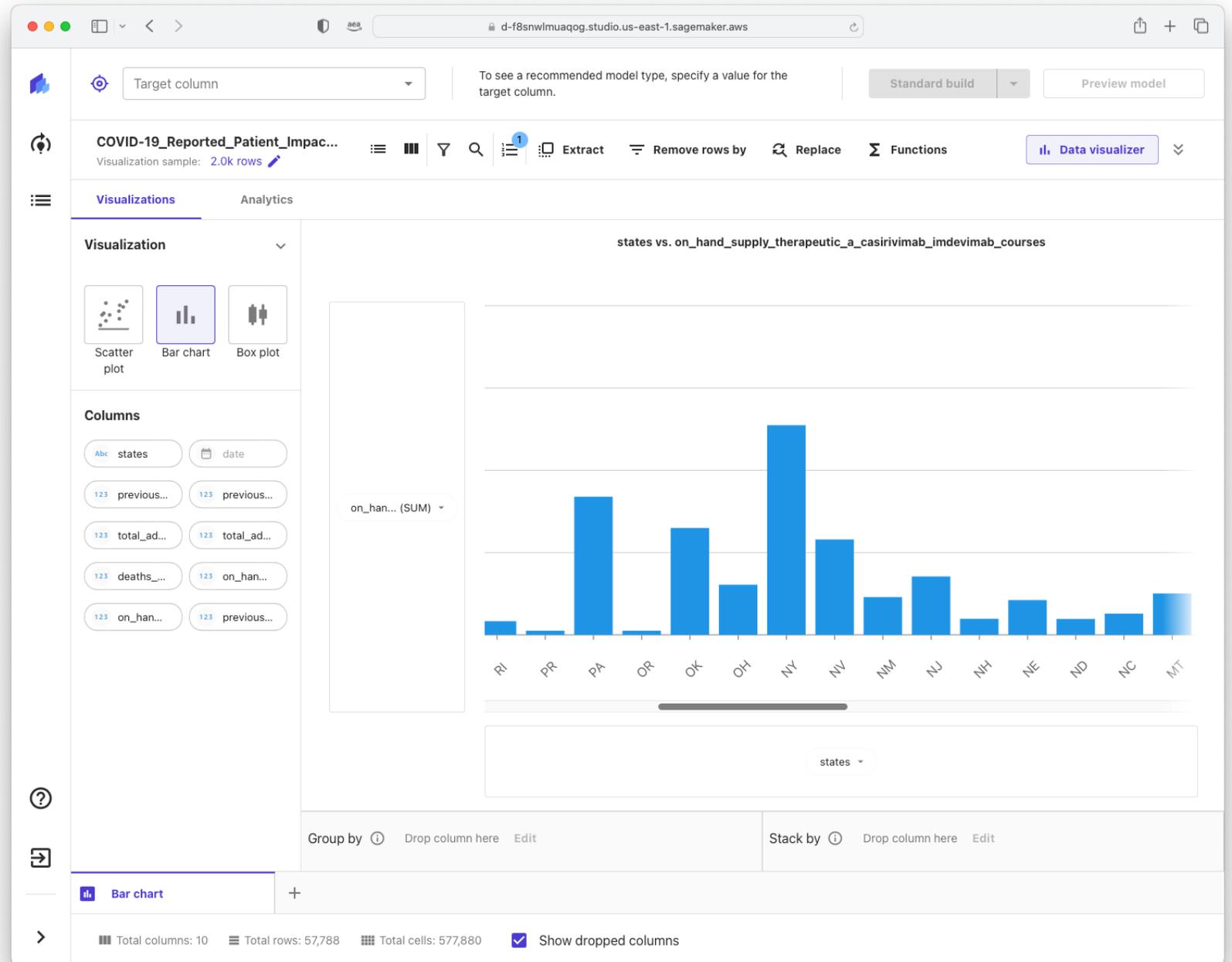
state	date	previous_day_...	previous_day_...	total_adult_p...	total_adult_p...	deaths_covid	on_hand_sup...	previous_w...
RI	1/1/21	30	1	456	415	16	501	0
WY	1/1/21	14	11	105	96	1	100	0
NY	1/1/21	986	216	8566	7926	103	2162	70
KY	1/1/21	460	106	1732	1601	24	243	26
TX	1/1/21	1741	582	12864	12425	180	21	7
IA	1/1/21	87	22	586	566	11	4	0
NE	1/1/21	50	23	541	513	13	22	2
WA	1/1/21	101	89	1080	990	15	63	4
AR	1/1/21	200	229	1211	1128	25	363	80
MS	1/1/21	180	56	1384	1196	32	145	96
PA	1/1/21	620	528	6764	5837	128	2091	5
DE	1/1/21	49	39	472	432	1	22	0
TN	1/1/21	365	76	3260	3122	53	203	0
HI	1/1/21	11	9	124	101	1	0	0
MN	1/1/21	112	57	984	870	13	0	0
LA	1/1/21	268	15	1652	2591	23	810	64
ND	1/1/21	20	10	187	160	1	376	13



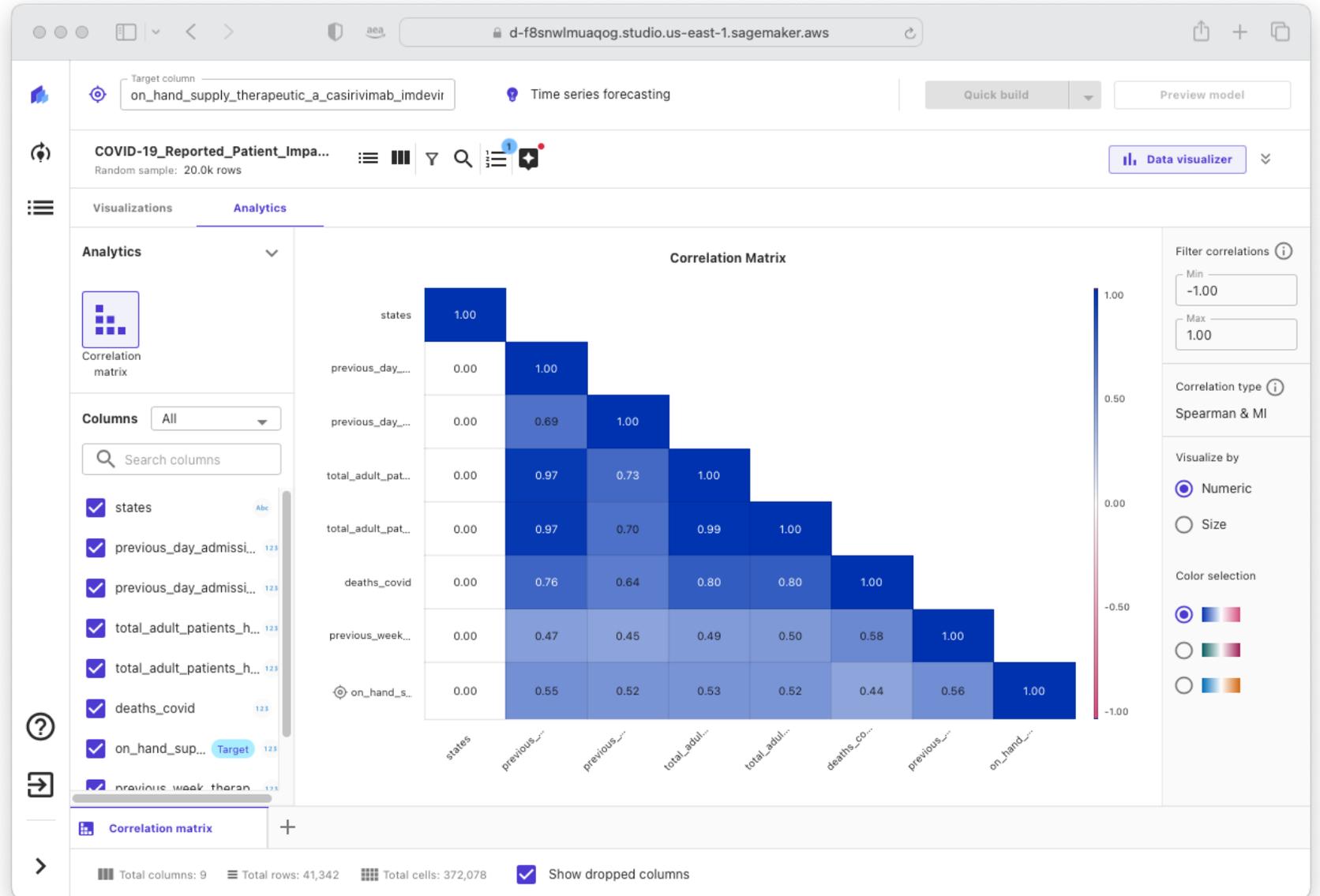
# Create multiple models on the same dataset to compare accuracy scores



Explore and visualize your data to gain insights into your data before building ML models



Explore and visualize your data to gain insights into your data before building ML models



# Get ready to build the model

**Therapeutics on hand - standard 2023-2-16** V1 Ready Add version Share

Select Build Analyze Predict

**Select a column to predict**  
Choose the target column. The model that you build predicts values for the column that you select.

Target column: `on_hand_supply_therapeutic_a_casirivimab_imdevimab_`

Value distribution:

**Model type**  
SageMaker Canvas automatically recommends the appropriate model type for your analysis.

- Time series forecasting  
Your model will forecast `on_hand_supply_therapeutic_a_casirivimab_imdevimab_courses` by using past data values to predict future data values.

Standard build Preview model

**COVID-19\_Reported\_Patient\_Impa...** Random sample: 20.0k rows Data visualizer

adult_patients_...	states	Abc	previous_week_thera...	previous_day_admiss...	previous_day_admiss...	on_hand_supply_ther...
5097.38	54 Categories	0.00	1368.29	0.00	441.58	0.00
	NY	70		216	986	2162
	KY	26		106	460	243
	IA	0		22	87	4
	NE	2		23	50	22
	MS	96		56	180	145
	PA	5		528	620	2091
	TN	0		76	365	203
	MN	0		57	112	0
	LA	64		15	268	810
	ND	13		10	20	376
	WI	3		83	235	1880

**Statistics**  
Data quality  
Missing values: 366  
Mismatched data types: 0  
Valid values: 19,709  
Distribution:

Unique values: 5,183  
Mode: 0.0  
Standard deviation: 2522.671057

Total columns: 9 Total rows: 41,342 Total cells: 372,078 Previewing first 100 rows Show dropped columns

# Automatically build an accurate ML model for your dataset

The screenshot displays the Amazon SageMaker console interface for building a machine learning model. The main window is titled "Therapeutics on hand - standard 2023-2-16" and is in the "Build" phase. A "Time series forecasting configuration" dialog box is open, providing options for configuring the model. The dialog includes fields for "Item ID column" (set to "states"), "Group column", "Time stamp column" (set to "date"), and "Days" (set to 30). There is also a toggle for "Use holiday schedule" and a "Country for holidays" field. The background shows a dataset table with columns such as "total\_adult\_patients\_hospitalized...", "states", "previous\_week\_therapeutic\_a\_ca...", "previous\_day\_admission\_adult\_c...", "on\_hand\_supply\_thera..." (marked as the target), "deaths\_covid", and "date".

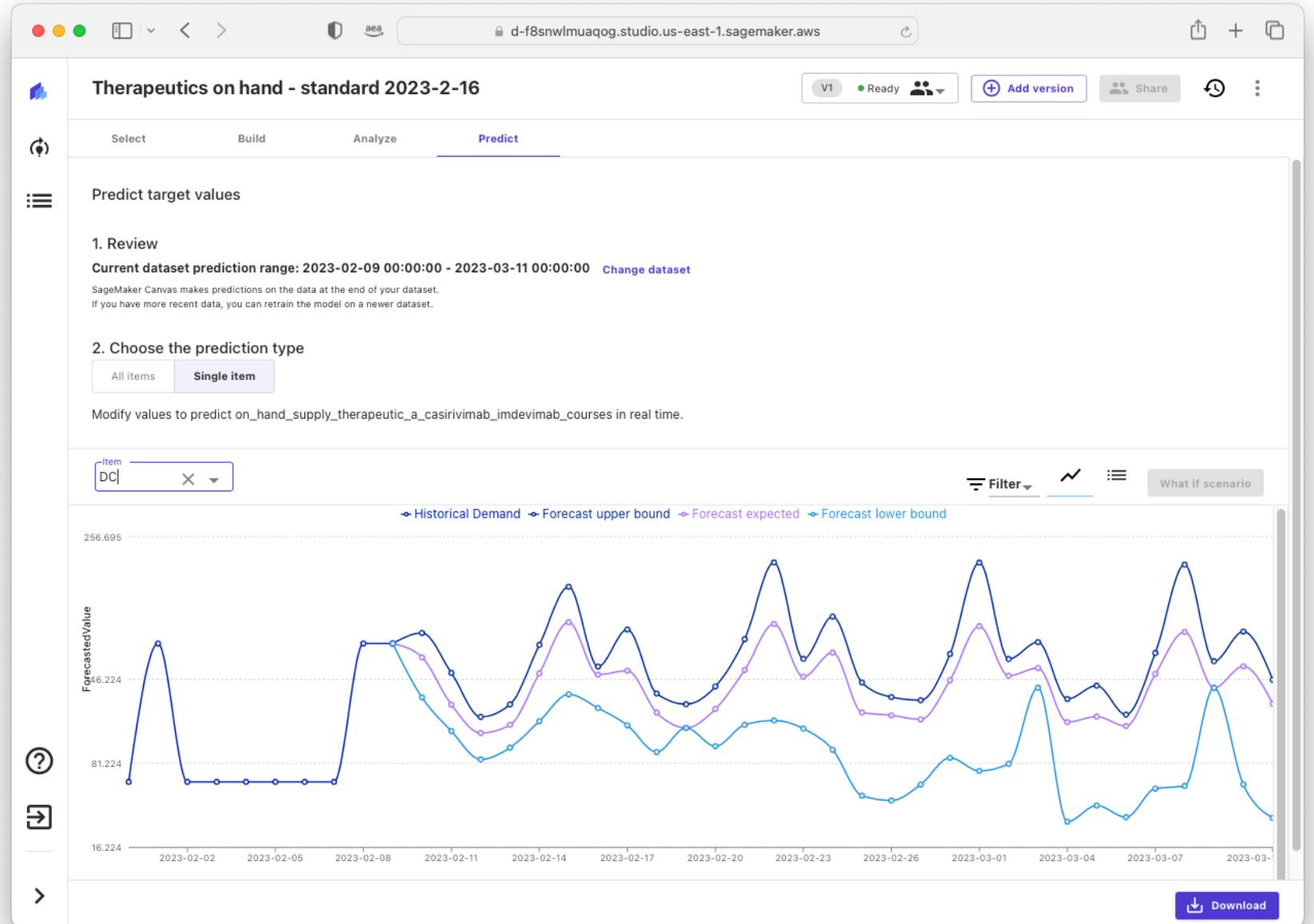
Get the first ML model in minutes. Review advanced metrics and feature importance to understand and explain predictions.

The screenshot displays the SageMaker Canvas interface for a project titled "Therapeutics on hand - standard 2023-2-16". The interface is divided into several sections:

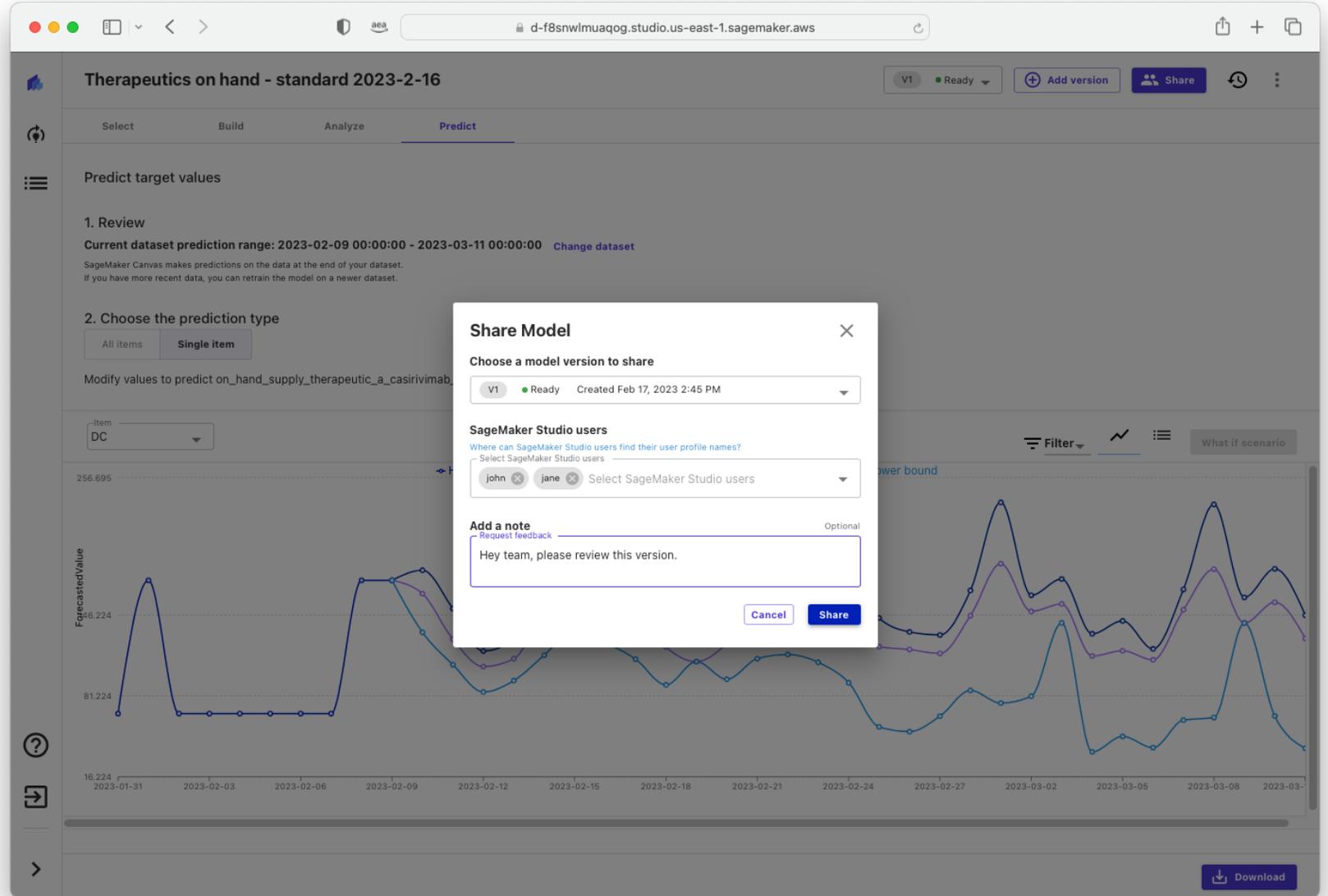
- Top Bar:** Shows the project name, version (V1), and status (Draft). Buttons for "Add version", "Share", and "Refresh" are visible.
- Navigation:** Tabs for "Select", "Build", "Analyze", and "Predict" are present, with "Build" currently active.
- Build Panel:**
  - Select a column to predict:** A dropdown menu shows "on\_hand\_supply\_therapeutic\_a\_casirivima..." as the selected target column. Below it is a "Value distribution" histogram.
  - Model type:** A section titled "Model type" indicates that SageMaker Canvas automatically recommends the appropriate model type. The selected type is "Time series forecasting". A description states: "Your model will forecast on\_hand\_supply\_therapeutic\_a\_casirivimab\_imdevimab\_courses by using past data values to predict future data values." Buttons for "Configure time series model" and "Change type" are provided.
  - Buttons:** "Quick build" and "Preview model" buttons are located on the right side of the Build panel.
- Data Table:** A table titled "COVID-19\_Reported\_Patient\_Impa..." shows a random sample of 20,000 rows. The table has columns for Column name, Data type, Missing, Mismatched, Unique, Mean / Mode, and Feature importance. The "on\_hand\_supply\_thera..." column is highlighted as the "Target".
- Preview Model Panel:** A panel on the right titled "Preview model" shows the "Estimated Avg. wQL" as 0.13. A description explains that 0.13 is a non-negative metric where a lower value indicates a more accurate model. A "Show more metrics" button is available. Below this, a "Column impact" section shows a search bar and a bar chart for "deaths\_covid" with a value of 4.417.
- Footer:** Summary statistics are shown: Total columns: 9, Total rows: 41,342, Total cells: 372,078, and a checked box for "Show dropped columns".



# Let's look at some results



Share ML models with data scientists for review and update so you can generate predictions on new model versions



# Import ML models built and trained *anywhere* and generate predictions directly in SageMaker Canvas

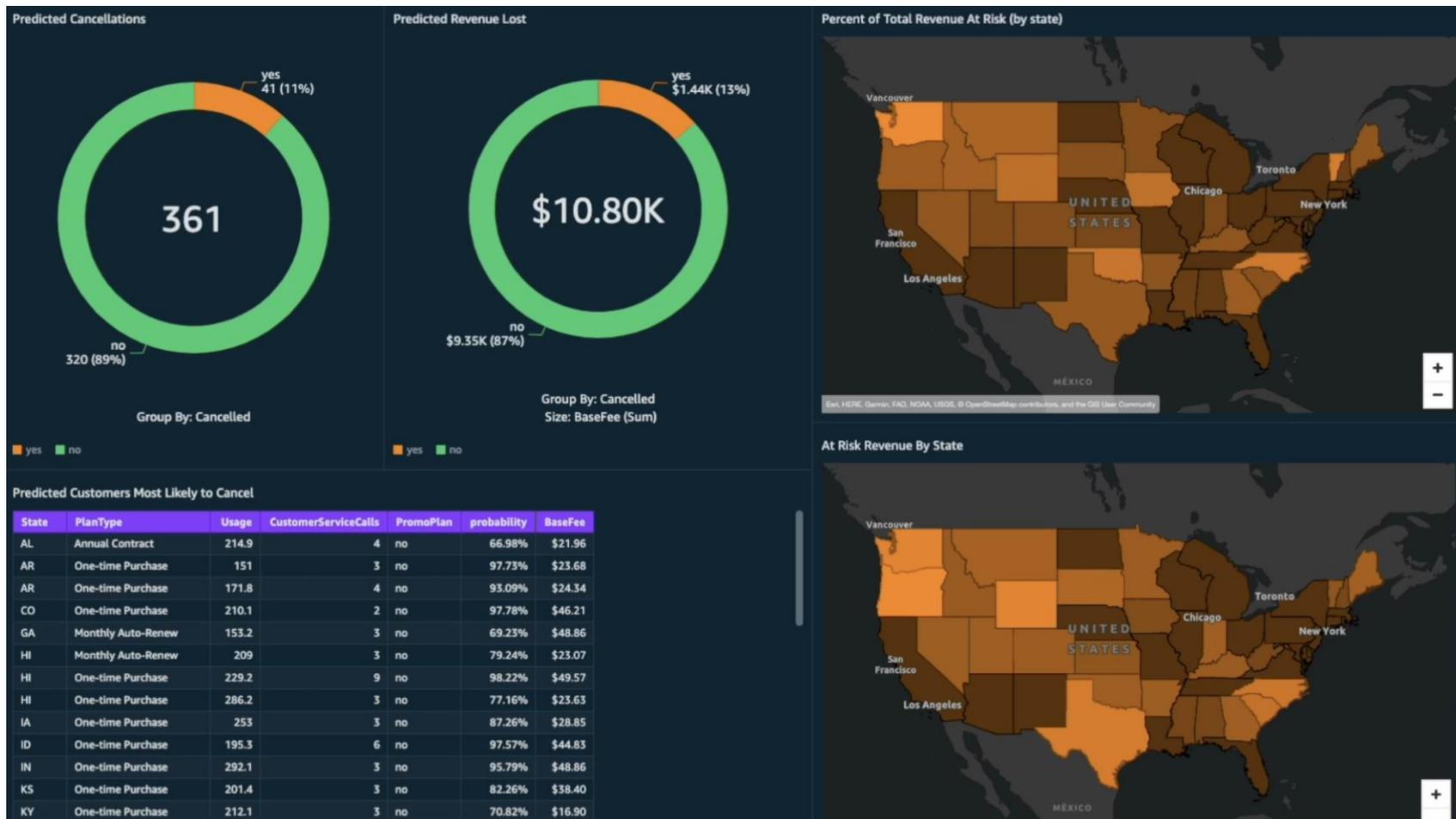
The screenshot displays the Amazon SageMaker Studio interface. A green notification banner at the top states: "Canvas user henrik has shared sm-uzvvi... with you. View shared model". The main content area shows a model card for "Therapeutics on hand - standard 2023-2-16 V1" with buttons for "From Canvas", "Share feedback", and "Share updated model". A note from henrik says: "Hey Team, please review this version." The "Canvas model detail" section lists the full dataset as "COVID-19\_Reported\_Patient\_Impact\_and\_Hospital\_Capacity\_by\_State\_Timeseries\_therape". Below this, the "Performance metrics" section includes a table of metrics:

Performance metrics			
This model was created using an ensemble of algorithms including CNN-QR, DeepAR+, ARIMA, Prophet, ETS, and NPTS. See <a href="#">documentation</a> to learn more.			
Average wQL	wQL [P90]	wQL [P50]	wQL [P10]
0.146	0.081	0.206	0.149
Weighted Absolute Percentage Error	Root Mean Squared Error	Mean Absolute Percentage Error	Mean Absolute Scaled Error
0.206	153.772	0.564	4.423

The "Feature importance" section explains that the model's predictions are explained using SHAP values. Below this, the "Features increasing target forecast" section lists features like "deaths\_covid" and "ion\_adult\_covid\_confirmed". The interface includes a sidebar with navigation options like Home, Data, AutoML, Experiments, Notebook jobs, Pipelines, Models, and Deployments. The bottom status bar shows "Simple" mode and "Shared models: Canvas".



# Run what-if scenarios, get predictions on an entire dataset, or build predictive dashboards



# Opportunities for making better decisions in public health



## Health outcomes

- Improve diagnostics
- Early detection for faster intervention
- Disease risk prediction
- Decision support
- Health risk prediction



## Public health

- Surveillance and detecting adverse events
- Advancing health equity
- Analyze large volumes of data
- Identify health trends
- Provide situational awareness



## Response readiness

- Demand forecasting
- Capacity forecasting
- Inventory efficiency
- Predictive quality
- Scenario modeling

*and many more...*

# Interested in learning more about how no-code ML can drive forecasting and analytics?



Visit our Amazon SageMaker Canvas site to find out more



Get started with Amazon SageMaker machine learning tutorials



# Thank you!

Dan Sinnreich

[dansinnr@amazon.com](mailto:dansinnr@amazon.com)

Henrik Balle

[henballe@amazon.com](mailto:henballe@amazon.com)

