

# Hands-on Lab: Generative AI for Data Generation and Augmentation

**Estimated Effort: 30 mins**

One of the principle advantages of generative AI is its ability to generate realistic synthetic data. The synthetic data is generated when a pretrained generative model responds to either a prompt, create new data samples, or transfers learns on a given data set. In addition, it creates samples that can augment the existing data set while maintaining the statistical distribution and interpretability of the data set.

In this lab, you will learn how to use generative AI to generate synthetic data samples and transfer learns on a given data set.

## Learning Objective

In this lab, you are going to use a popular tool, [Mostly.ai](#), for creating synthetic data samples to augment a CSV data set.

## Data Set

You are going to use a data set on Insurance records.

The data set is available at the link below.

[Insurance Dataset](#)

This data set is a cleaned-up version of the [Medical Insurance Price Prediction](#) data set, available under the [CC0 1.0 Universal License](#) on the [Kaggle](#) website.

## Steps

### 1. Download the data set

The first step is to download the dataset on your machine. You will need to upload this file to the interface in a subsequent step. Click the link provided in the **Data Set** section to download it.

### 2. Open the website

Click the link below to open the interface.

<https://mostly.ai/>

This link will open in a new browser tab, and you should see an interface that looks as shown below.

← ↻ 🏠 🔒 https://mostly.ai 🔊 🌟 ⚙️ | 📄

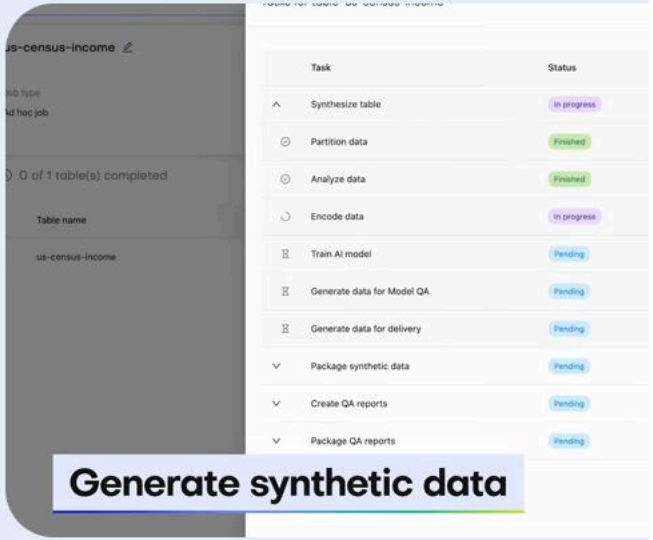
💡 Download the complete guide to AI-generated synthetic data!

**MOSTLY AI** Platform Synthetic Data Resources Company Pricing Docs 🔍 [Log in](#)

# Synthetic Data. Better than real.

Still struggling with real data? Use existing data for synthetic data generation.  
Synthetic data is more accessible, more flexible, and simply...smarter.

[Request a demo](#)

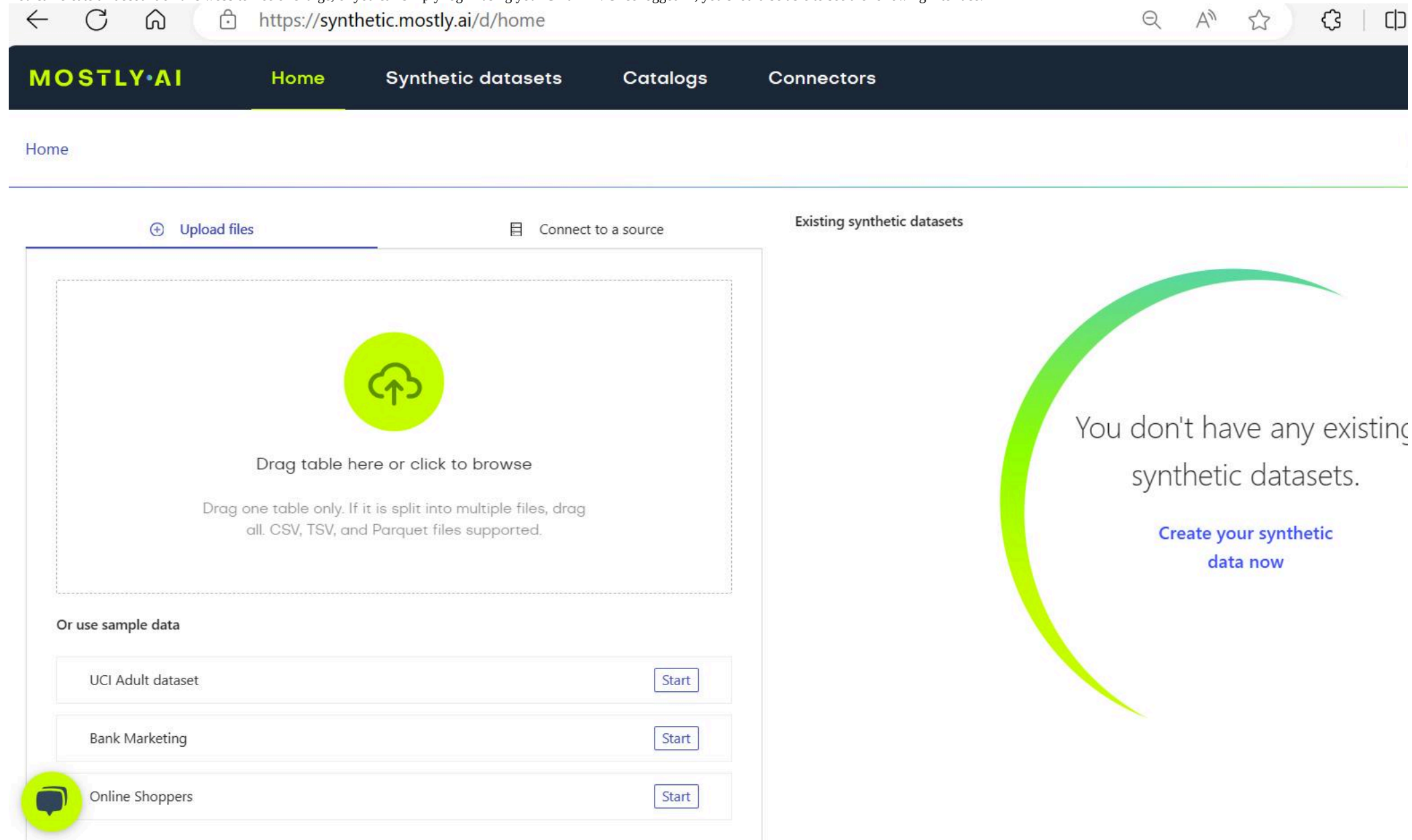


Task	Status
^ Synthesize table	In progress
⌚ Partition data	Finished
⌚ Analyze data	Finished
↻ Encode data	In progress
⌘ Train AI model	Pending
⌘ Generate data for Model QA	Pending
⌘ Generate data for delivery	Pending
⌵ Package synthetic data	Pending
⌵ Create QA reports	Pending
⌵ Package QA reports	Pending

[Generate synthetic data](#)

### 3. Create an account

You can create an account on this website free of charge, or you can simply log in using your Gmail ID. Once logged in, you should be able to see the following interface.



The screenshot shows the web interface of syntheticmostly.ai. At the top is a dark navigation bar with the logo 'MOSTLY.AI' and links for 'Home', 'Synthetic datasets', 'Catalogs', and 'Connectors'. Below this is a light blue header with the 'Home' link. The main content area is divided into two sections. The left section, titled 'Upload files', contains a large dashed box with a green cloud icon and the text 'Drag table here or click to browse' and 'Drag one table only. If it is split into multiple files, drag all. CSV, TSV, and Parquet files supported.' Below this, under 'Or use sample data', are three rows: 'UCI Adult dataset', 'Bank Marketing', and 'Online Shoppers', each with a 'Start' button. A green chat bubble icon is in the bottom left. The right section, titled 'Existing synthetic datasets', features a large green curved graphic and the text 'You don't have any existing synthetic datasets.' with a blue link 'Create your synthetic data now'.

Home

Upload files

Connect to a source

Existing synthetic datasets

Drag table here or click to browse

Drag one table only. If it is split into multiple files, drag all. CSV, TSV, and Parquet files supported.

Or use sample data

UCI Adult dataset

Start

Bank Marketing

Start

Online Shoppers

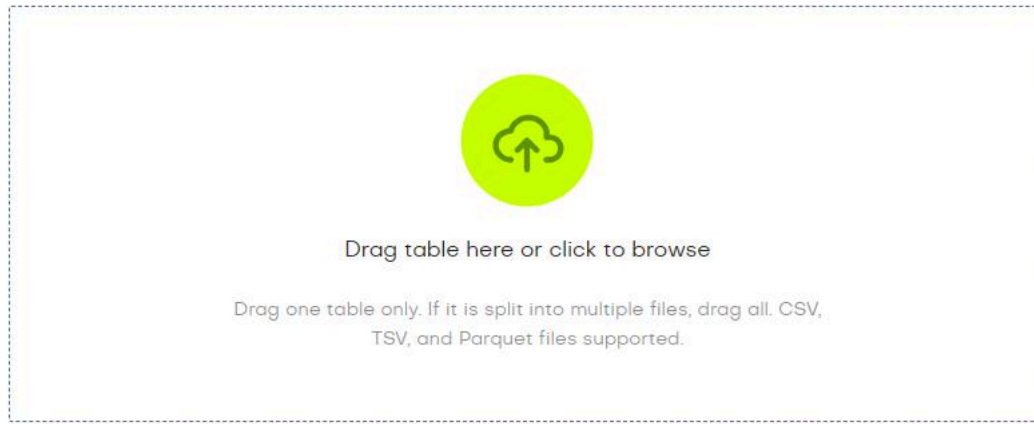
Start

You don't have any existing synthetic datasets.

Create your synthetic data now

#### 4. Upload the data set

Upload the CSV file of the data set to the interface by using the upload option available on the console. Once uploaded, the filename will be visible on the console. Then you can click Proceed as shown below.



\* Name ⓘ insurance\_dataset

Selected files:

📎 insurance\_dataset.csv



Cancel

Proceed

### 5. Training parameter selection

On the interface, you will see a space provided to choose the number of training samples that can be used from your data set. If you leave this entry blank, it is interpreted as allowing the use of all entries in the data set to train the generative model. You may leave this entry blank, since the more data that is used for training the model, the more accurate the synthetic data created will be.

Synthetic datasets / [Start job](#)insurance\_dataset [🔗](#)

## Tables

## Data settings

## Output settings



Reference tables will not be included in the synthetic output. To see them in your synthetic dataset, mark them as subject tables or refer them to a subject table



## 6. Data settings

Name ⓘ

Training Goal ⓘ

Training Size ⓘ

Maximum tra

insurance\_dataset

Accuracy

Speed

Turbo


100

Specify the number of records to use for model training. If left blank, all available records are used. When you specify a number, you set the upper limit for the number of (context) records used for training. With a lower number, you can speed up model training, which however can also reduce the synthetic data accuracy.

You can choose to modify the category of an attribute, or you can choose to include the parameter in the augmentation process without these settings. In this lab, let us not make any changes to these settings.

Synthetic datasets / Start job

Back

insurance\_dataset 

Tables


Data settings

Output settings

i

Configure your data's  
synthetization settings.

Reference tables will not be  
included in the synthetic  
output. To see them in your  
synthetic dataset, mark  
them as subject tables or  
move them to a subject table



Search

Q

Tables

s

insurance\_dataset

insurance\_dataset- Subject table

Include	Column name	Generation method
<input checked="" type="checkbox"/>	age	AI / Numeric:Auto
<input checked="" type="checkbox"/>	gender	AI / Categorical
<input checked="" type="checkbox"/>	bmi	AI / Numeric:Auto
<input checked="" type="checkbox"/>	children	AI / Numeric:Auto
<input checked="" type="checkbox"/>	smoker	AI / Categorical
<input checked="" type="checkbox"/>	region	AI / Categorical
<input checked="" type="checkbox"/>	expenses	AI / Numeric:Auto

7. Output settings

You can specify the number of samples you want to create synthetically using this setting. If you leave it blank, it generates the exact same number of data samples as there are in the original file. Let us leave this setting blank, choosing to generate the same number of samples as in the original data set.

Synthetic datasets / [Start job](#)[Back](#)insurance\_dataset [🔗](#)

## Tables

## Data settings

## Output settings



Specify how many subjects you want to generate. This determines the size of the synthetic data.

Reference tables will not be included in the synthetic output. To see them in your synthetic dataset, mark subject tables or link them to a subject table

## \* Data destination

Download as CSV/Parquet

The synthetic data can always be downloaded

Use these fields to specify the size of the synthetic data. Decrease them to create a subset of the original data or increase them to make small datasets larger. Leave them blank to have the same number of synthetic subjects as in the original data.

Subject tables

Number of generated subjects ⓘ

insurance\_dataset

Once these settings are done, click **Create a synthetic dataset**, as visible on the top right corner of the webpage.

**8. Data set creation process**

You will see an interface as shown below.

Synthetic datasets



Name	Type	Status	Creation date
insurance_dataset	Upload a file	In progress	Dec 12, 2023 6:51 PM



The status will change from Pending to In process to Finished in several minutes. To track the progress of the action, click anywhere on the row to get to an interface that looks as shown below.

MOSTLY-AI

Home

Synthetic datasets

Catalogs

Connectors

Synthetic datasets / Summary

insurance\_dataset

Generated rows  
1338

Sample data

Preview synthetic samples from insurance\_dataset

All samples: 1

age	gender	bmi	children	smoker	region	expenses
19	female	29.8	1	no	northwest	12014.56
30	male	24.4	0	no	southwest	2418.49
23	male	33.8	1	no	southeast	3999.76
31	male	24.3	0	no	northwest	3248.84
28	female	34.6	1	no	southeast	4067.33
35	male	34.6	2	no	northwest	5448.42
22	male	36.6	1	no	southeast	4242.4

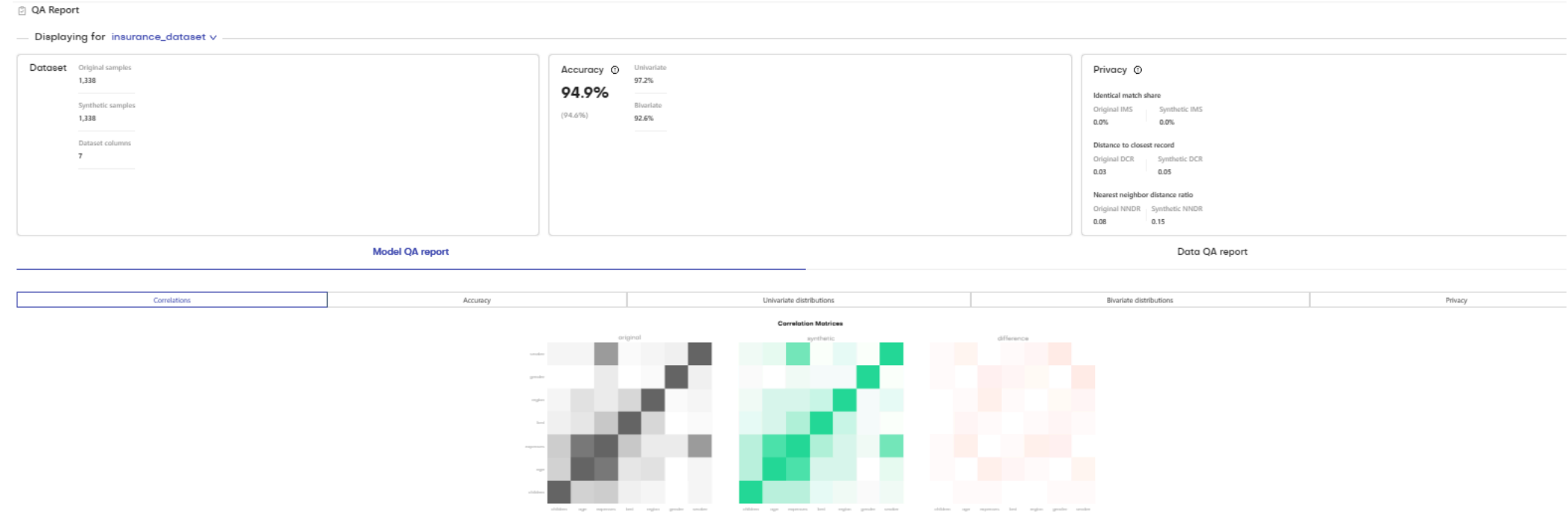
In the top right corner, you can access the log section to note each of the tasks happening in the process of this data creation.

📄 Logs

⬇

Model	Model Type	Step	Progress	Durati
insurance_dataset	Flat	Fetch data	<div></div>	
insurance_dataset	Flat	Analyze data	<div></div>	
insurance_dataset	Flat	Encode data	<div></div>	
insurance_dataset	Flat	Train AI model	Trained 36 epochs	<div></div>
insurance_dataset	Flat	Generate data for Model QA	1.3K out of 1.3K done	
insurance_dataset	Flat	Analyze data for Model QA	<div></div>	
insurance_dataset	Flat	Generate data for delivery	1.3K out of 1.3K done	
insurance_dataset	Flat	Analyze data for Data QA	<div></div>	
insurance_dataset	Flat	Post-process data	<div></div>	
insurance_dataset	Flat	Export data	<div></div>	
↳		Package synthetic dataset	<div></div>	
↳		Package QA reports	<div></div>	

Once completed, you can also check the statistical similarity of the synthetic data samples with the original dataset in the QA report section. It will also show the accuracy score of the synthetic data set.



9. Download the synthetic data

Once the process is complete, you can click Synthetic datasets on the web page to see that the status will show as Finished. At this point, your synthetic data set is now ready. Click the download link, as shown in the image below, to download the synthetic data.

Name	Type	Status	Creation date	Actions
insurance_dataset	Upload a file	Finished	Dec 12, 2023 6:51 PM	<div><div></div><div></div><div></div><div></div><div></div></div>

Click here to download the dataset

You may now use this synthetic data set for data science operations, or you can also augment the original data set with these samples to be used together.

Conclusion

Congratulations! You have completed the lab on data augmentation using the Mostly.ai tool.

Author(s)

[Abhishek Gagneja](#)

© IBM Corporation. All rights reserved.



**Skills** Network