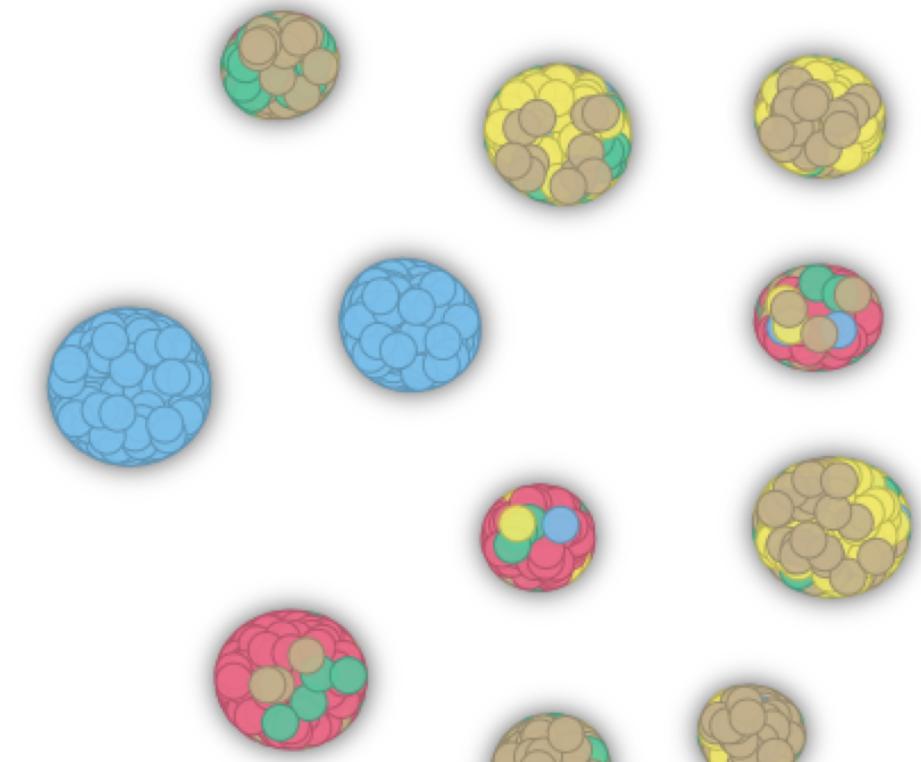
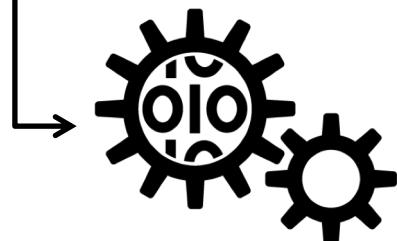
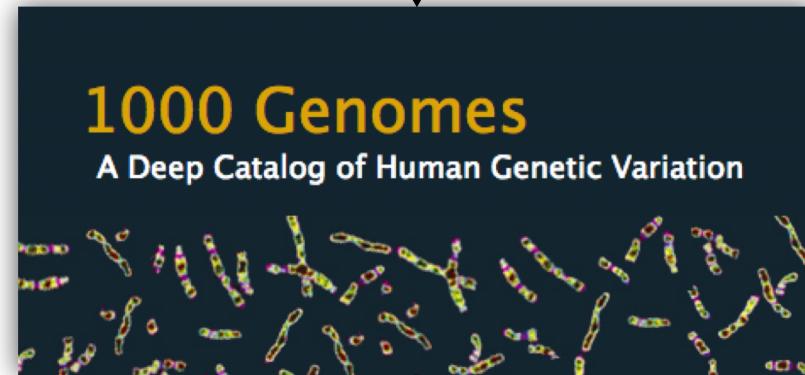
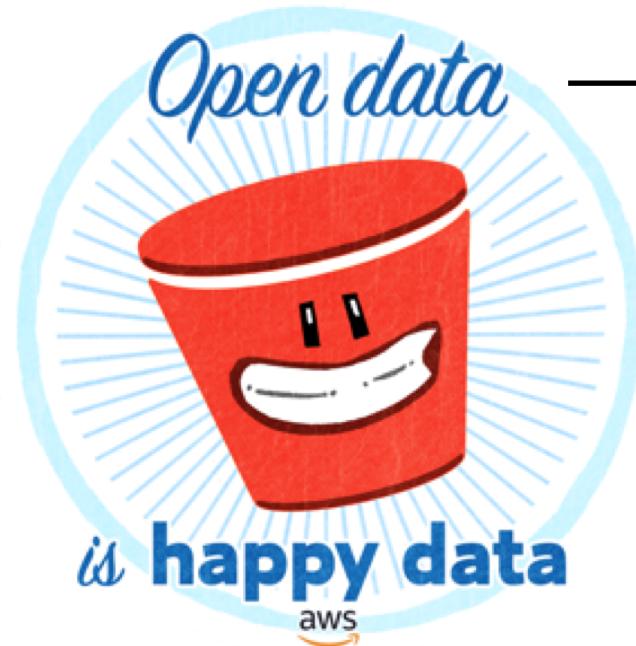


BUILDER SESSION - WPS201

# Building a Genome Clustering Model with Amazon Sagemaker

W. Lee Pang, PhD  
Genomics & Life Sciences Specialist Architect  
Amazon Web Services / WW Public Sector

# What we are building



# Tools and services we will use

## Infrastructure Automation



AWS  
CloudFormation

## Data Storage



Amazon S3

## Data Processing

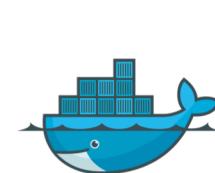


AWS Batch

## Data Science



Amazon  
SageMaker



# Let's start building!

<http://bit.ly/reinvent2018-wps201>

<http://bit.ly/reinvent2018-wps201>

### Builder Session - WPS201

Session Description

Prerequisites

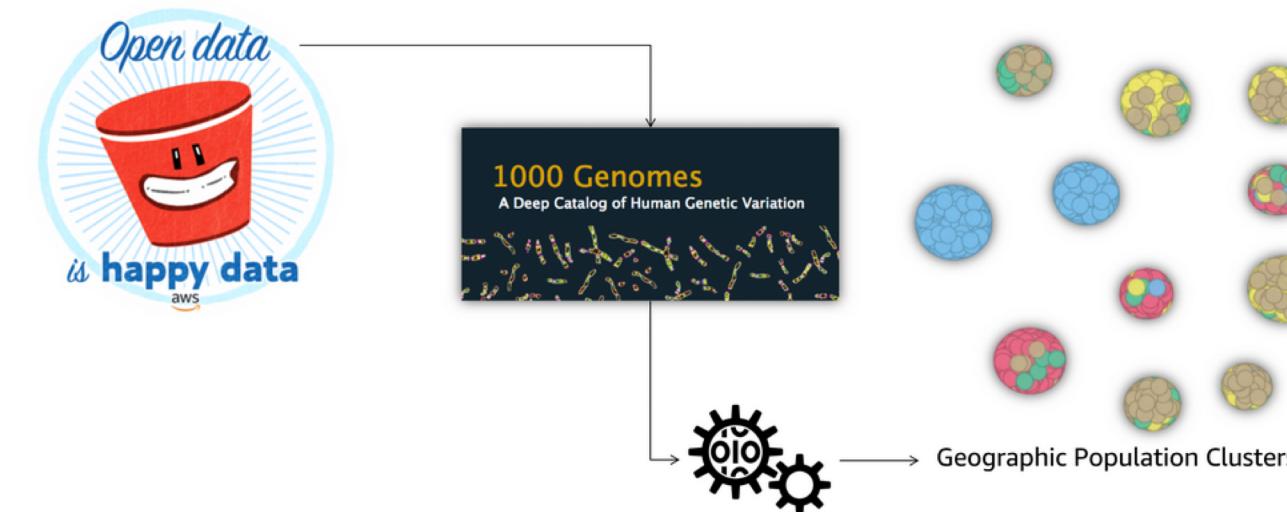
Getting started

Applying credits to your account

Create resources with CloudFormation

Open your SageMaker notebook instance

# Builder Session - WPS201



# <http://bit.ly/reinvent2018-wps201>

Services ▾ Resource Groups ▾ ⚙

1 Billing

Billing Access, analyze, and control your AWS costs a

2 Credits

- Dashboard
- Bills
- Cost Explorer
- Budgets
- Reports
- Cost Allocation
- Tags
- Payment Methods
- Payment History
- Consolidated Billing
- Preferences
- Credits
- Tax Settings

Credits

Please enter your code below to redeem your credits.

Promo Code

Security Check  3

Refresh Image

Please type the characters as shown above  4

By clicking "Redeem" you indicate that you have read and agree to the terms of the AWS Promotional Credit Terms & Conditions located [here](#).

Redeem 5

Apply credits to your account

<http://bit.ly/reinvent2018-wps201>

## Create resources with CloudFormation

Create resources with CloudFormation

Launch Stack

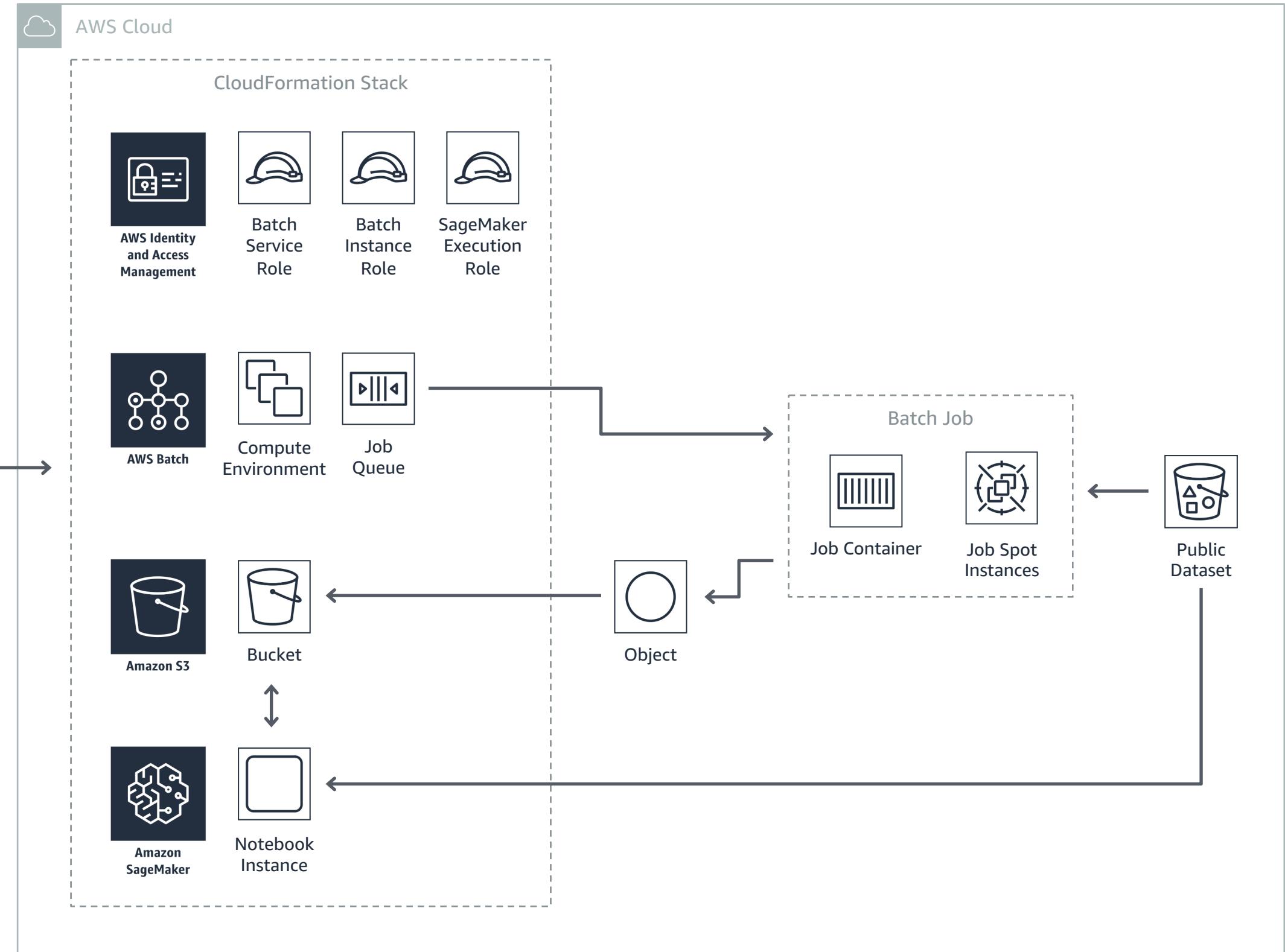
VPC ID  **vpc-a81bdcd0 (172.31.0.0/16)**

VPC Subnet IDs  **subnet-860252ff (172.31.16.0/20)**

Spot Bid %  **subnet-960023cc (172.31.0.0/20)**

I acknowledge that AWS CloudFormation might create IAM resources with custom names.

I acknowledge that AWS CloudFormation might require the following capability: CAPABILITY\_AUTO\_EXPAND



# Data background



samples..

Genomic analysis pipeline

A diagram showing the structure of a Variant Call File (.vcf). It consists of a header section labeled '.vcf' and a body section. The header includes columns for CHR, POS, REF, ALT, and SAMPLES.. The body contains two rows of data. The first row has CHR=1, POS=12345, REF=G, ALT=G, and SAMPLES..=0/0. The second row has CHR=1, POS=12678, REF=T, ALT=A, and SAMPLES..=1/0.

.vcf	CHR	POS	REF	ALT	SAMPLES..
	1	12345	G	G	0/0
	1	12678	T	A	1/0

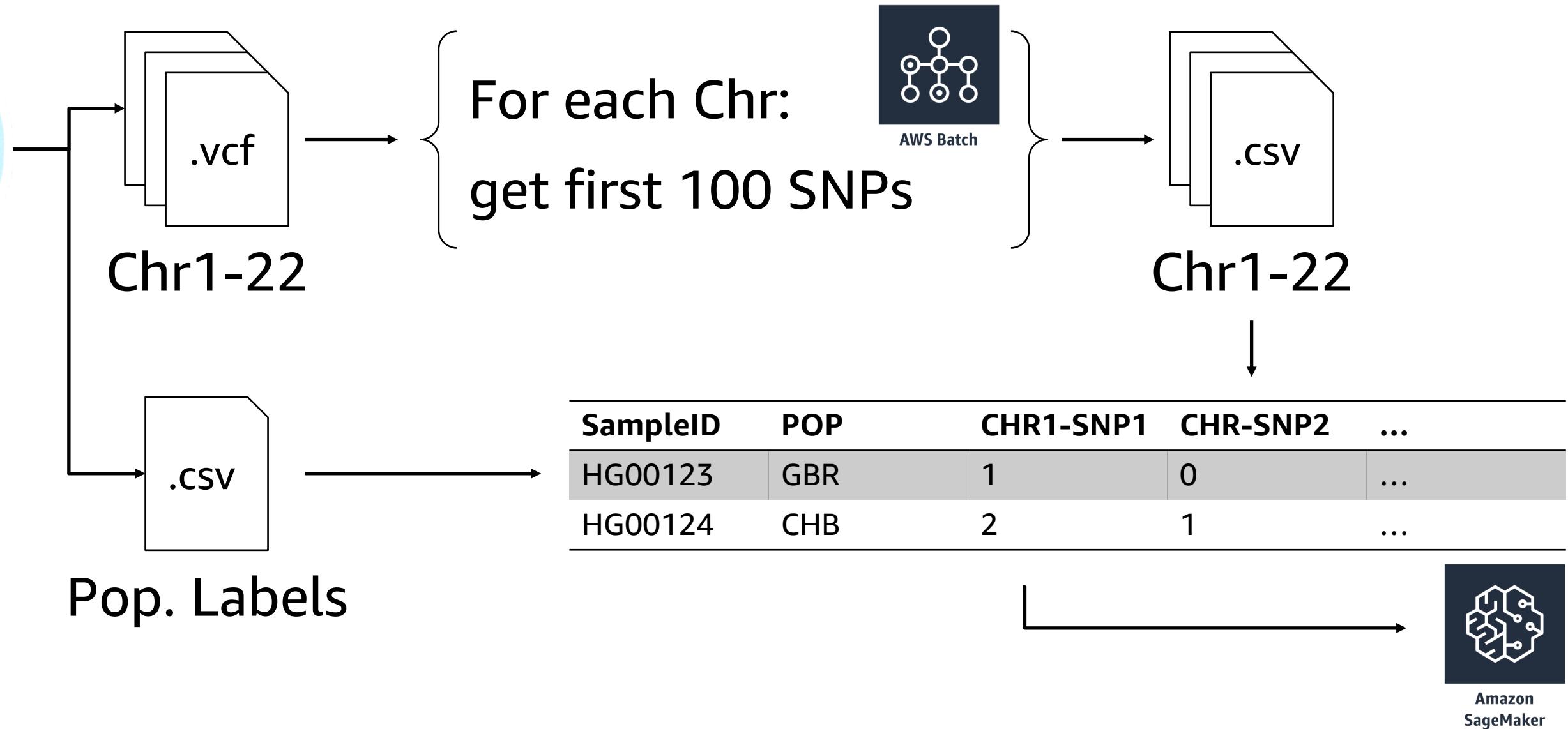
Variant Call File

AWS Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.



# Data preparation



<http://bit.ly/reinvent2018-wps201>

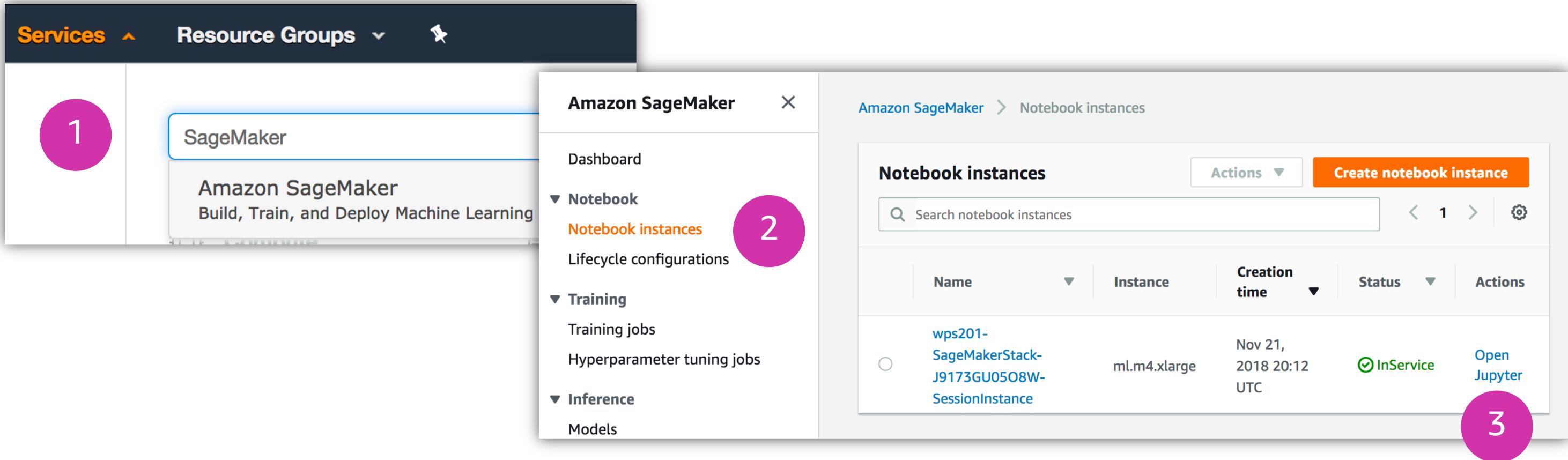
## Create resources with CloudFormation

Showing 5 stacks					
	Stack Name	Created Time	Status	Drift Status	Description
<input type="checkbox"/>	wps201-SageMakerStack-J9...	2018-11-21 12:12:30 UTC-0800	CREATE_COMPLETE	NOT_CHECKED	Creates a SageMaker Notebook in...
<input type="checkbox"/>	wps201-BatchStack-VZPNU8...	2018-11-21 12:11:25 UTC-0800	CREATE_COMPLETE	NOT_CHECKED	Creates resources for AWS Batch.
<input type="checkbox"/>	wps201-iamStack-1Y1N241U...	2018-11-21 12:08:43 UTC-0800	CREATE_COMPLETE	NOT_CHECKED	Creates a set of IAM resources.
<input type="checkbox"/>	wps201-S3Stack-1NDOTM6...	2018-11-21 12:08:05 UTC-0800	CREATE_COMPLETE	NOT_CHECKED	Creates an S3 Bucket
<input checked="" type="checkbox"/>	wps201	2018-11-21 12:08:00 UTC-0800	CREATE_COMPLETE	NOT_CHECKED	re:Invent 2018 Builder Session WP...

Overview	Outputs	Resources	Events	Template	Parameters	Tags	Stack Policy	Change Sets	Rollback Triggers			
Filter by: Status ▾ Search events												
2018-11-21	Status	Type		Logical ID	Status Reason							
▶ 12:15:12 UTC-0800	CREATE_COMPLETE	AWS::CloudFormation::Stack		wps201								
▶ 12:15:09 UTC-0800	CREATE_COMPLETE	AWS::CloudFormation::Stack		SageMakerStack								
▶ 12:12:30 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		SageMakerStack	Resource creation Initiated							
▶ 12:12:29 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		SageMakerStack								
▶ 12:12:27 UTC-0800	CREATE_COMPLETE	AWS::CloudFormation::Stack		BatchStack								
▶ 12:11:25 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		BatchStack	Resource creation Initiated							
▶ 12:11:24 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		BatchStack								
▶ 12:11:22 UTC-0800	CREATE_COMPLETE	AWS::CloudFormation::Stack		IamStack								
▶ 12:08:43 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		IamStack	Resource creation Initiated							
▶ 12:08:42 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		IamStack								
▶ 12:08:40 UTC-0800	CREATE_COMPLETE	AWS::CloudFormation::Stack		S3Stack								
▶ 12:08:06 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		S3Stack	Resource creation Initiated							
▶ 12:08:05 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		S3Stack								
▶ 12:08:00 UTC-0800	CREATE_IN_PROGRESS	AWS::CloudFormation::Stack		wps201	User Initiated							

# http://bit.ly/reinvent2018-wps201



## Open a Sagemaker notebook instance

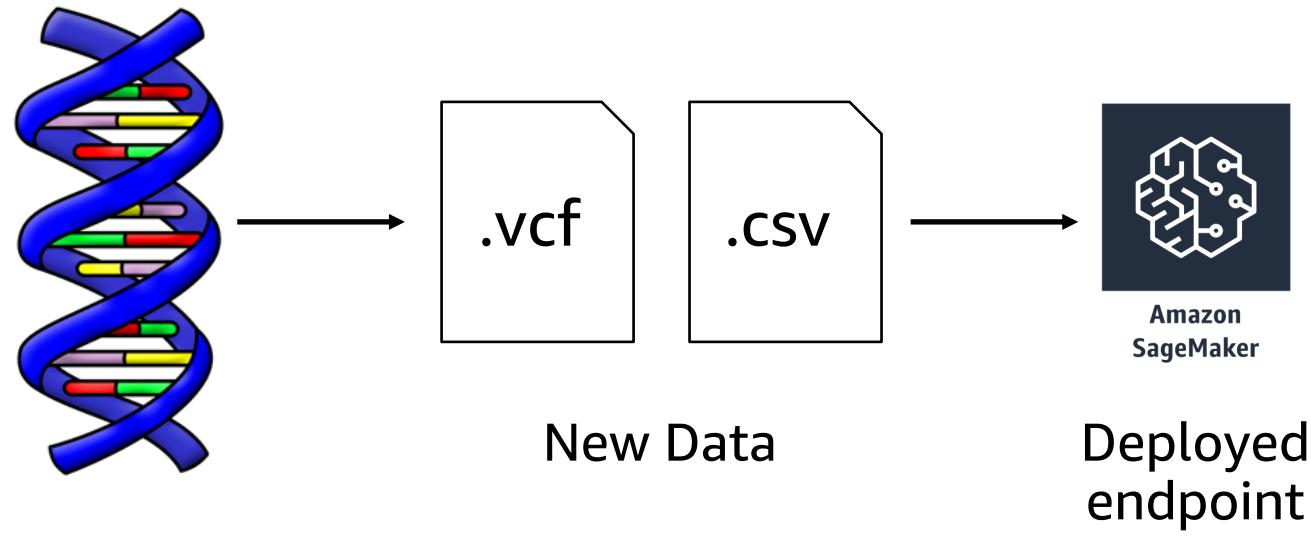
# <http://bit.ly/reinvent2018-wps201>

The screenshot shows two instances of the Jupyter Notebook interface. The left instance is a file browser with a sidebar showing a directory structure: 'lost+found' and 'wps201-SageMakerStack-J9173C'. The right instance is a list of files in a specific folder:

Name	Last Modified	File size
..	seconds ago	
genome-kmeans-py3.ipynb	2 days ago	2.13 MB
environment.yaml	2 days ago	90 B
LICENSE	2 days ago	1.07 kB
submit_jobs.py	2 days ago	5.71 kB

**Open the demo notebook**

# Now what?

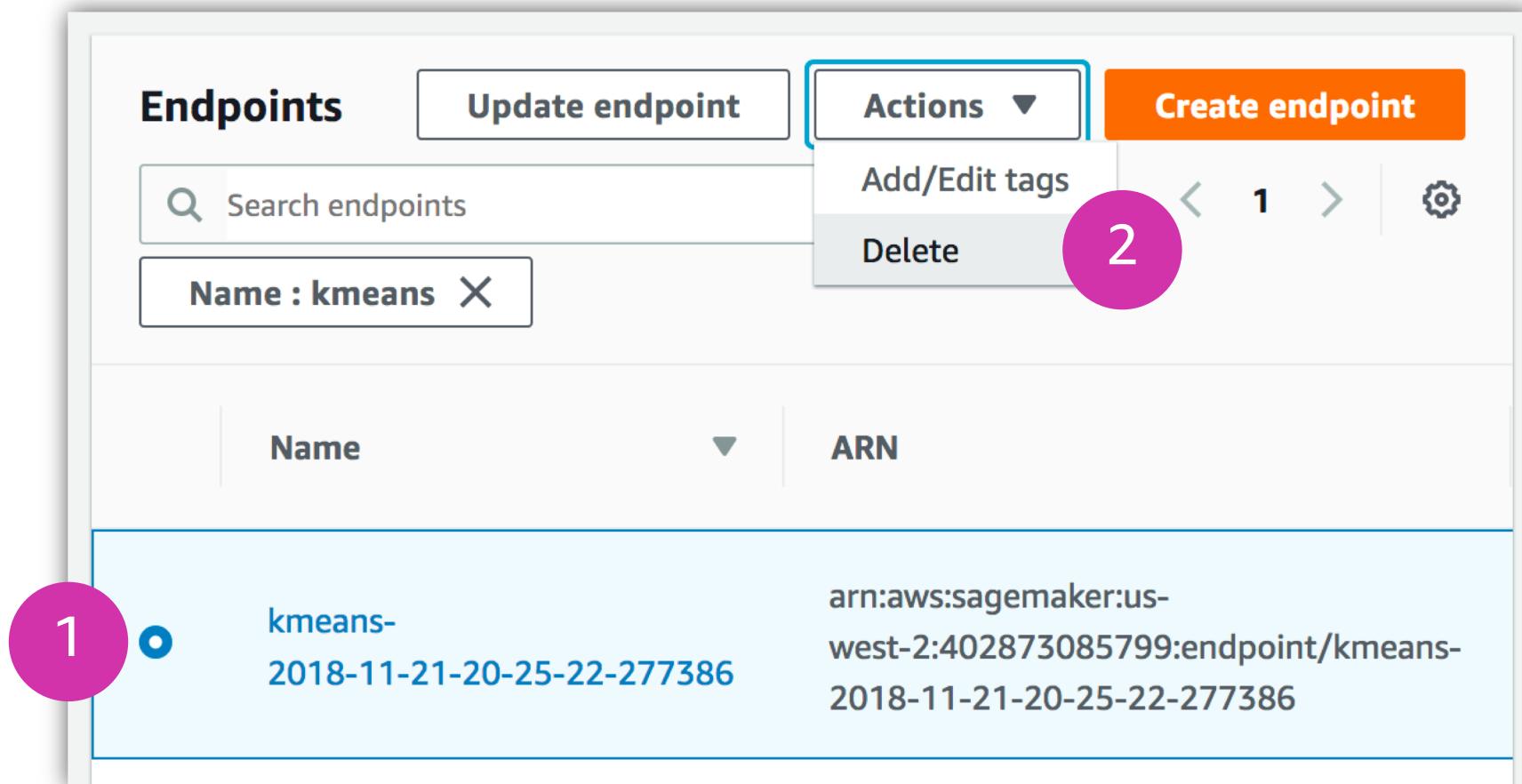


→ Where this DNA sample came from:

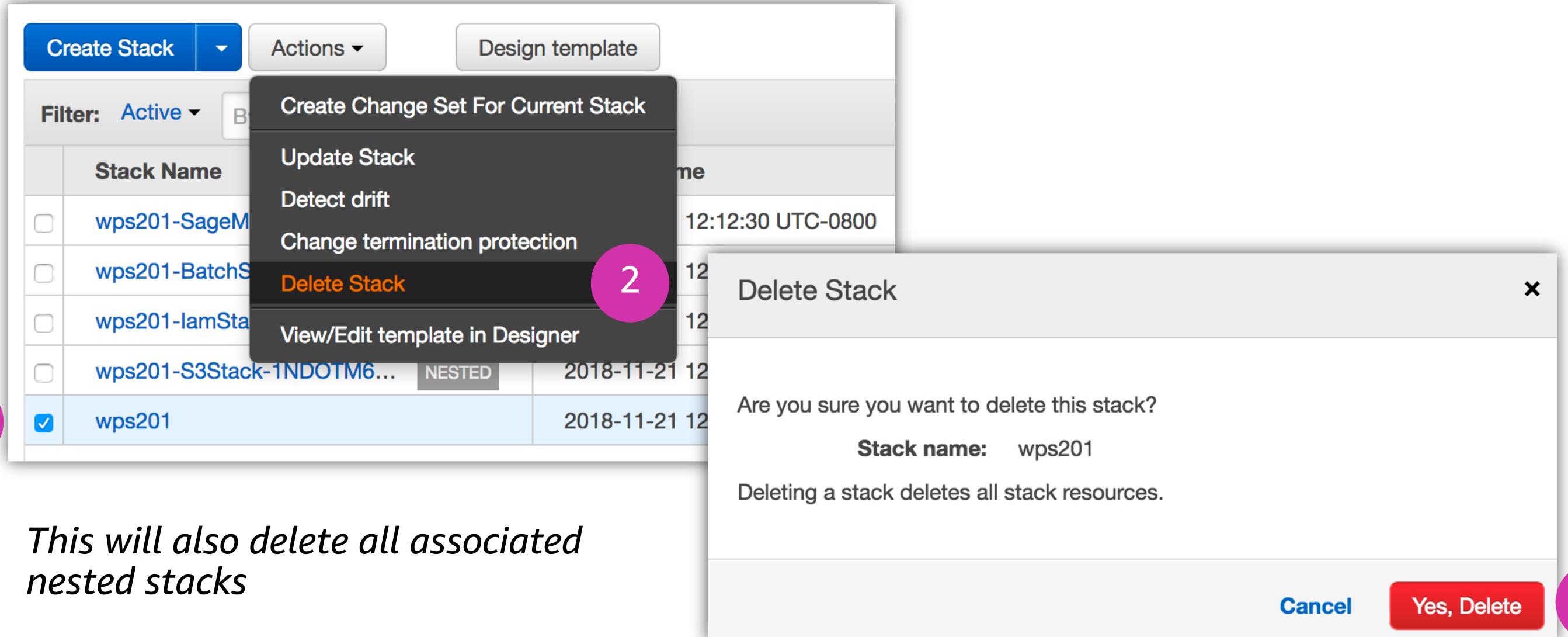
- **Geographic location**
- **Cancer tumor type**
- **Species of origin**

# Clean-up!

# Clean-up: delete Sagemaker Endpoints

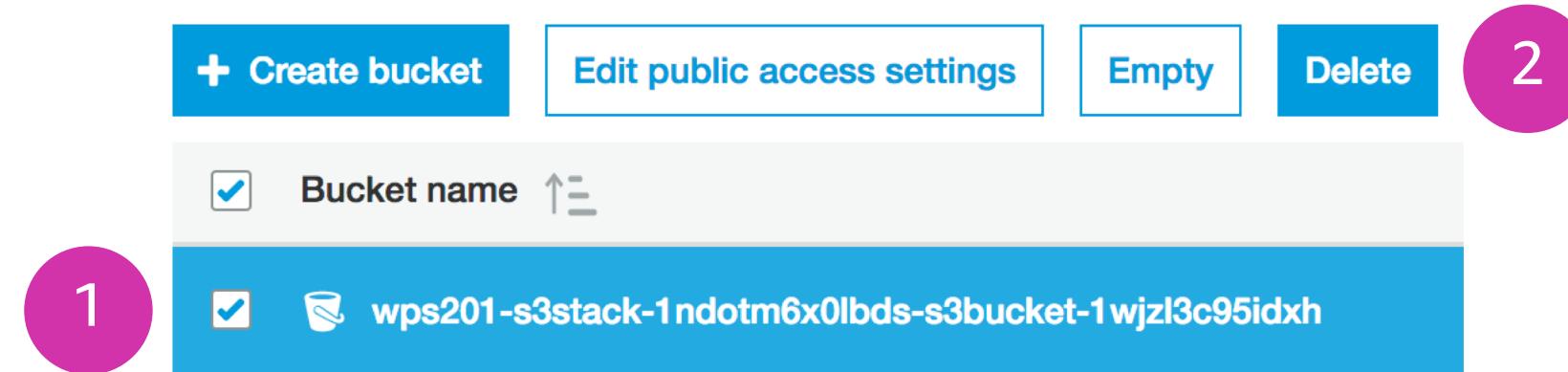


# Clean-up: delete CloudFormation stack



*This will also delete all associated nested stacks*

# Clean-up: delete S3 bucket



# Thank you!