The data on the danger of publicly exposed S3 buckets

Laminar

We recently released a post summarizing our findings that 21% of all publicly exposed buckets contained sensitive data. In this post we drill down much further on exactly how we made this determination, add more details on our findings, illustrate why we believe that a large part of the exposed sensitive data, in particular PII, is due to third party software, and detail steps you can take to mitigate the threat—the data behind the findings.

The Numbers:

308,441,001
Files were scanned

85,214
Buckets were scanned

21.1%
Of buckets had sensitive data in them

Buyer beware: A large percentage of "Mismanaged PII" seems to have originated from third party software vendors – customer PII was leaked not by the company that owns the data, but because of mismanaged data residing in the third party software company's environment.

First, let's set the stage with a few definitions:

- **What is Amazon S3?**

Amazon Simple Storage Service (Amazon S3) is, in Amazon's words[1], "an object storage service offering industry-leading scalability, data availability, security, and performance." Amazon S3 is by far the largest data storage software with 44.66% of market share[2]. Customers of all sizes and industries can store and protect any amount of data for virtually any use case, such as data lakes, cloud-native applications, and mobile apps. Organizations use S3 buckets for the flexibility they offer developers and data scientists to quickly and easily store, access, and share almost any file or object from anywhere, at the click of a button. No cumbersome implementation process required.

- **What do we mean by sensitive data?**

We think of any data that an organization wants to keep confidential including things like highly regulated data like PHI, PII or PCI data, customer and employee information, and trade secrets as sensitive data.

- **What is PII?**

The U.S. Department of Labor defines personally identifiable information (PII) as "any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means."

- **What is a file vs. an object?**

Within a bucket every item within that bucket is an object. For the purposes of this post, every object in a bucket is a file and the two words can be used interchangeably.

And now the full story (and we mean the real, sit-down and grab a latte 'cause this will take a while, story):

We started this project because we wanted to know what kind of data we could find in publicly exposed S3 buckets, so that we could understand the potential exposure, and advise on how best to mitigate. To achieve this we got a **sample of publicly exposed S3 buckets**, and because nobody except possibly Skynet[3] has the ability to review all publicly exposed S3 buckets, we determined that 308,441,001 files in 85,214 buckets was a solidly representative sample size.

How can my file be public?

To understand how data becomes inadvertently publicly exposed, and why it may happen, we first have to understand **how the access is enabled**. By default new buckets, access points and objects don't allow public access. However, there will be times that users will want to have a file publicly accessible (for example – images and files used by websites).

There are two main ways a file in an S3 bucket can be made public*.

1. Bucket Access Control:

Allowing the entire bucket to be accessed from the internet – Amazon allows you to set a policy that enables all files in the bucket to be publicly accessed. This is called a "bucket policy".

2. File/Object Access Control:

Allow only specific files to be accessible from the internet

- Using an access control list (ACL), within which the AWS accounts or groups that have access as well as the type of access they receive are defined
- Using a bucket policy that grants public read

This second option means that even if the bucket is not public, a file inside it can still be accessed from the internet if the ACL grants the relevant permissions.

* For these to work, you can't have any block public access settings at the account level or the bucket level. By default, block public access settings are set to True on **new** S3 buckets. If the block public access feature is activated for all buckets within the account, the message "Bucket and objects not public" is shown and an S3 bucket may not be made public by users until the block is lifted by the administrator.

Important: Granting public access through bucket and object ACLs doesn't work for buckets that have S3 Object Ownership set to **Bucket Owner Enforced**. In most cases, ACLs aren't required to grant permissions to objects and buckets. Instead, you can use AWS Identity Access and Management (IAM) policies and S3 bucket policies to grant permissions to objects and buckets.

✓ **Note:**

As of April 2023, all new S3 buckets will have ACLs disabled and S3 Block Public Access settings enabled for all new buckets4.

How attackers can abuse data publicly accessible in S3 buckets

Once data is publicly accessible it's easy to obtain. There are many websites that enable you to search for publicly accessible documents in S3 buckets. Any person in the world, including, of course, attackers, can access and then mine this data – from phone numbers, addresses, and credit card details to internal proprietary information and private medical information. Armed with this information, attackers can then do a number of things with long-standing impacts on the organization to which the data belongs—from **mis-use of proprietary information** to gain an unfair competitive advantage to **credit card fraud** to **ransoming of customer data**. And, of course, organizations also have to worry about potential **compliance violations** due to misuse of regulated data like PHI, PII or credit card information.

Examples of real-world impacts

Organizations **across industries, regions, and sizes** should beware of the risk of sensitive data being made inadvertently or unintentionally publicly accessible. To put it into real-world context, here are just a few examples of organizations who've seen the impacts from data in publicly exposed buckets, and, unfortunately, the subsequent press coverage.

- 3TB of **airport data** from Columbia and Peru including employee PII as well as information on planes, fuel lines, and GPS map coordinates were found to be publicly accessible in an S3 bucket, without any authentication required for access.[5]
- Microsoft experienced a leak of **customers' contact information**, including such data as names, email addresses, the email content, phone numbers, and files linked to business between affected customers and Microsoft or an authorized Microsoft partner.[6]
- McGraw-Hill accidentally exposed more than 100,000 **students' personal information** like names, emails, and grades and via their misconfigured S3 buckets.[7]
- **Patient information** from a COVID-19 testing service was identified as exposed by a cybersecurity expert who then notified the owner of the data, hopefully before attackers found it.[8]
- Then there is the defunct digital marketing company Reindeer that left its S3 bucket open to the public after going out of business, exposing 50,000 **customer files**.[9]

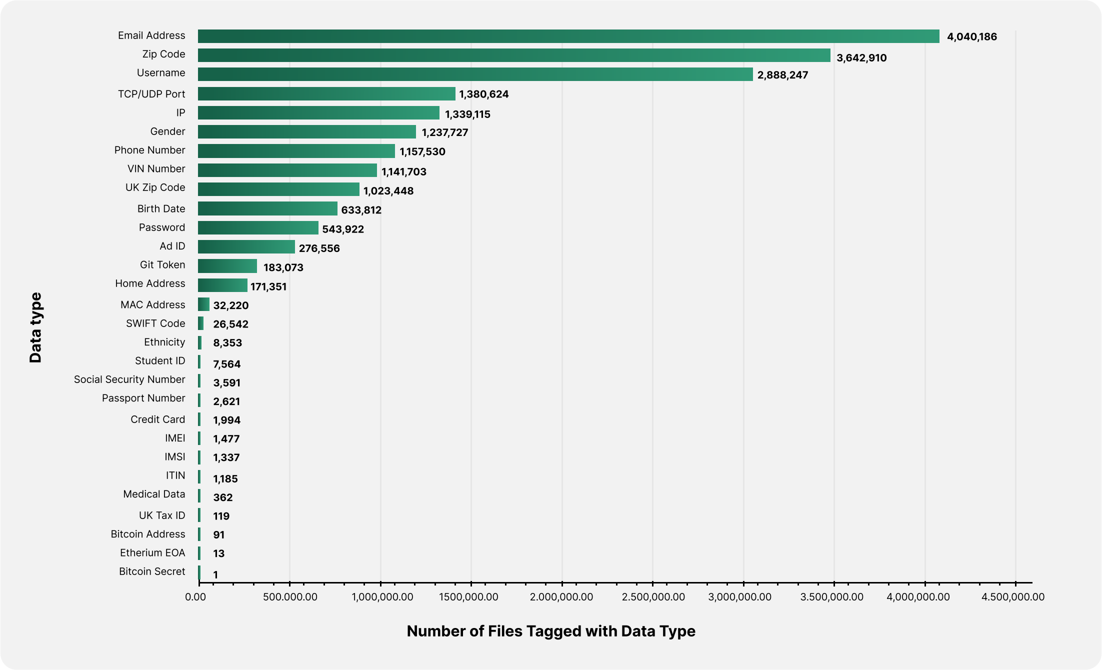
The list could go on, and on, but we'll stop here so we can move on to the data. Oh yes, it's always all about the data.

> Organizations of all sizes, across all industries and geographies need to beware the risk of publicly exposed sensitive data

What we discovered (including statistics)

As mentioned, we used a sample size of **85,000+ publicly accessible buckets** and the files within. Then we used our unique classification software to determine how much of this data was **sensitive data**. Based on this we concluded that a full 21% of these representative buckets contained sensitive data. We found everything from internal company performance reviews, complaints filed by employees, welfare payment details, private medical information, financial information to the more standard but still concerning username, zip code and email address.

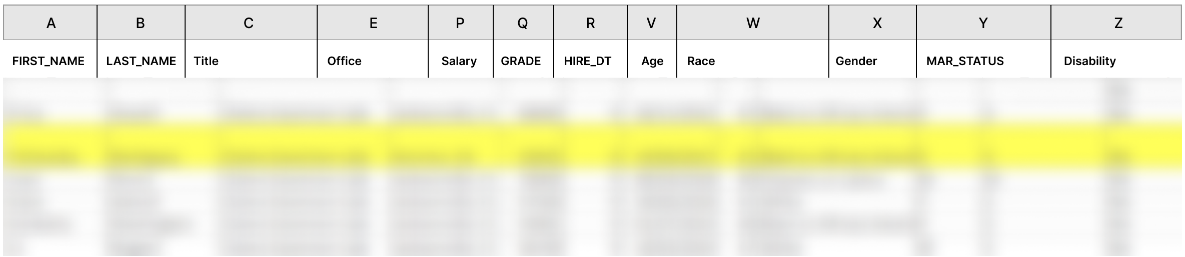
FILES TAGGED WITH SENSITIVE DATA TYPES



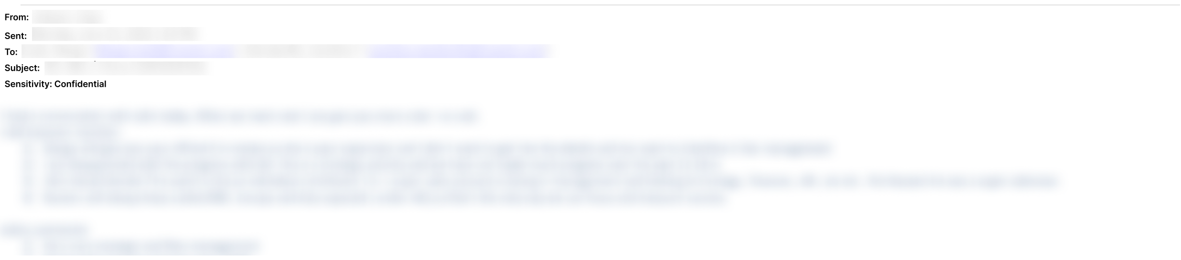
Some highly concerning examples

Now let's walk through how we got to the data that we found the most troubling. First, we took all of the file data and looked at which files had the most unique data types in them. Unique in this case means those files with a lot of data like **emails, phone numbers, names, addresses and social security numbers**. Then we examined the buckets with the most files containing that type of PII. Within those buckets we narrowed our focus to the ones that also contained **publicly accessible PDF or Word documents**, given that having such documents usually points to the fact that these files should not be public. Finally, from the set of these buckets with a large amount of files with PII and PDF or Word documents we selected those that had the most unique data types.

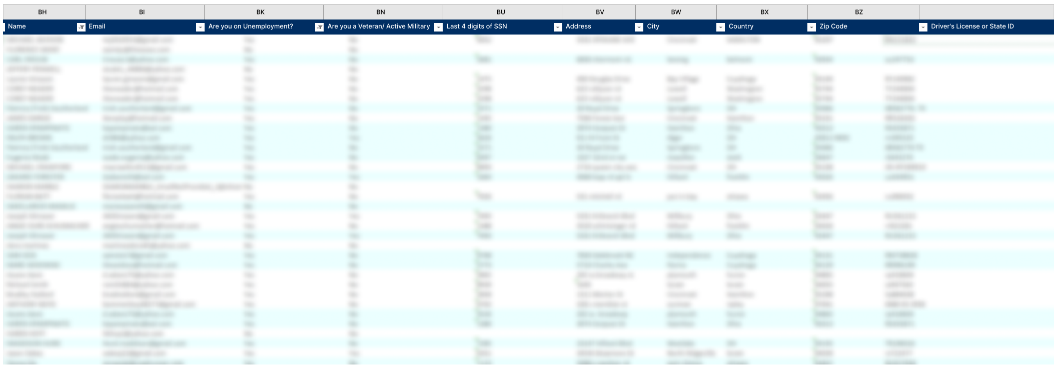
Looking at these buckets, some distinct examples stood out to us:



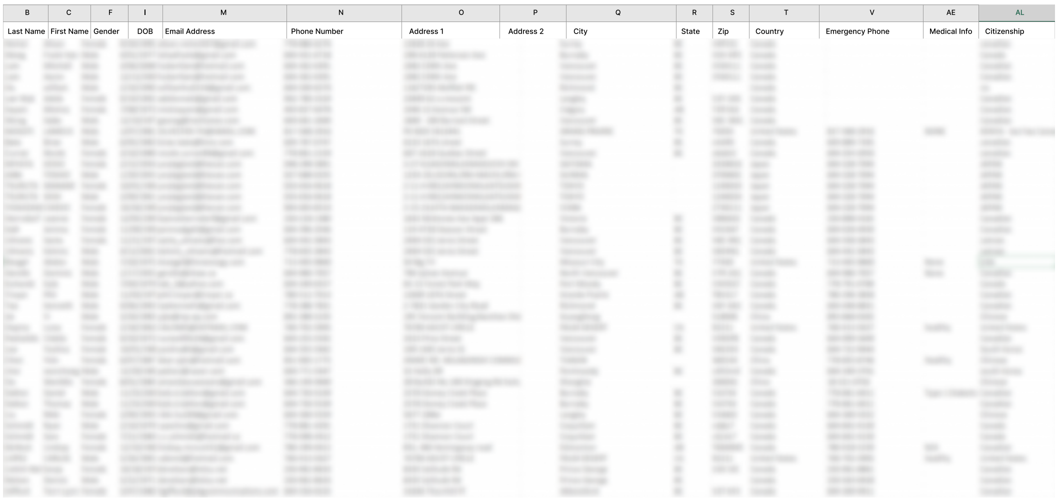
A	B	C	E	P	Q	R	V	W	X	Y	Z
FIRST_NAME	LAST_NAME	Title	Office	Salary	GRADE	HIRE_DT	Age	Race	Gender	MAR_STATUS	Disability



From:
Sent:
To:
Subject:
Sensitivity: Confidential



BH	BI	BK	BN	BU	BV	BW	BX	BZ	
Name	Email	Are you on Unemployment?	Are you a Veteran/ Active Military	Last 4 digits of SSN	Address	City	Country	Zip Code	Driver's License or State ID



B	C	F	I	M	N	O	P	Q	R	S	T	V	AE	AL
Last Name	First Name	Gender	DOB	Email Address	Phone Number	Address 1	Address 2	City	State	Zip	Country	Emergency Phone	Medical Info	Citizenship

Because we had to obfuscate to protect the anonymity of the innocent, what you're looking at are: Files containing PII of people who used a third-party chatbot service on different websites – including names, phone numbers, emails – and the messages sent to the bot (for example – people seeking welfare benefits); files containing loan details – name, loan amount, credit score, interest rates and more; a participant report for an athletic competition, including PII (name, address, zip code, email and more) and medical info; a VIP invite list including names, email, and address information; file with first names, last names, ethereum address and bitcoin address information, and block card email addresses.

Where are these leaks coming from?

Now that we had discovered that there was indeed a lot of sensitive data publicly exposed, and some of it, as listed above, incredibly concerning, we wanted to understand where the exposure was happening.

Misadventures in misplaced data

As expected, quite a few of these files are public due to a simple **mistake made** by the person configuring the bucket's access policy or the person unintentionally adding private, sensitive data to a publicly accessible bucket. Most of the most concerning data we found was due to the latter, the advent of **"misplaced" data**. Things that lead us to believe that these files were exposed due to mistakes in general:

1. **The bucket itself is not public**, and only half of the files in the bucket are public and contain private PII. Because it is only half of the files, it's a natural assumption that these were clearly not meant to be public, and we assumed it was a mistake.
2. **Some were "old" public files** which, again, contain PII which obviously shouldn't be public (such as data that dates back a few years), also pretty logically a simple oversight.

We believe most of the sensitive data we found is misplaced and not misconfigured because we found buckets with data types that should be in a publicly accessible bucket, but one file that didn't belong and had sensitive data. The clear implication from that is that this file was misplaced in this bucket intended to be public.

The hidden risk of data leaked by third party software

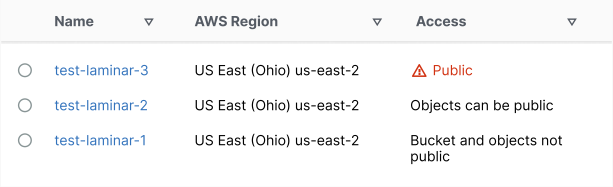
The more worrying pattern we found during this research was the amount of data leaked by **third party software companies**. Given the most highly concerning sensitive data we found centered around the type of data that by its nature clearly was shared via third party software company, it seems that companies offering services such as **HR management software, providers of live chat software, online event registration** software companies and more are the biggest, or at least most concerning, source of exposed sensitive PII. Given this, it is clearly not enough anymore to make sure that your company's data is safe, but also that the third parties who can access your data comply with data safety regulations and that your organization is submitting them to a fairly rigorous security review before sharing data with periodic audits.

> **Protip:**
>
> With the evidence of sensitive data exposed by third party software providers, be sure that any third parties who can access your data comply with data safety regulations and that your organization is submitting them to a fairly rigorous security review before sharing data, with periodic audits.

Exposed Data Detection and mitigation

Now that we've scared you with the kind of data we've found in publicly accessible buckets, here are a few things you can check to see if your files are publicly accessible and then change access settings as needed:

You can see if your bucket is publicly accessible in the Buckets list. In the Access column, Amazon S3 labels the permissions for a bucket as follows:



	Name ▽	AWS Region ▽	Access ▽
○	test-laminar-3	US East (Ohio) us-east-2	⚠ Public
○	test-laminar-2	US East (Ohio) us-east-2	Objects can be public
○	test-laminar-1	US East (Ohio) us-east-2	Bucket and objects not public

- **Public** – Everyone has access to one or more of the following: List objects, Write objects, Read and write permissions.
- **Objects can be public** – The bucket is not public, but anyone with the appropriate permissions can grant public access to objects.
- **Buckets and objects not public** – The bucket and objects do not have any public access.
- **Only authorized users of this account** – Access is isolated to IAM users and roles in this account and AWS service principals because there is a policy that grants public access.

You can also filter bucket searches by access type. Choose an access type from the drop-down list that is next to the Search for buckets bar.

To edit block public access settings for all the S3 buckets in an AWS account

- **Sign in** to the AWS Management Console and open the Amazon S3 console at https://console.aws.amazon.com/s3/
- **Choose Block Public Access** settings for this account.
- **Choose Edit** to change the block public access settings for all the buckets in your AWS account.
- **Choose the settings** that you want to change, and then choose Save changes.
- When you're asked for confirmation, **enter confirm**. Then choose Confirm to **save your changes**.

> This is best for when there is no need for anything in your S3 buckets to be public

Follow these steps if you need to change the public access settings for a single S3 bucket.

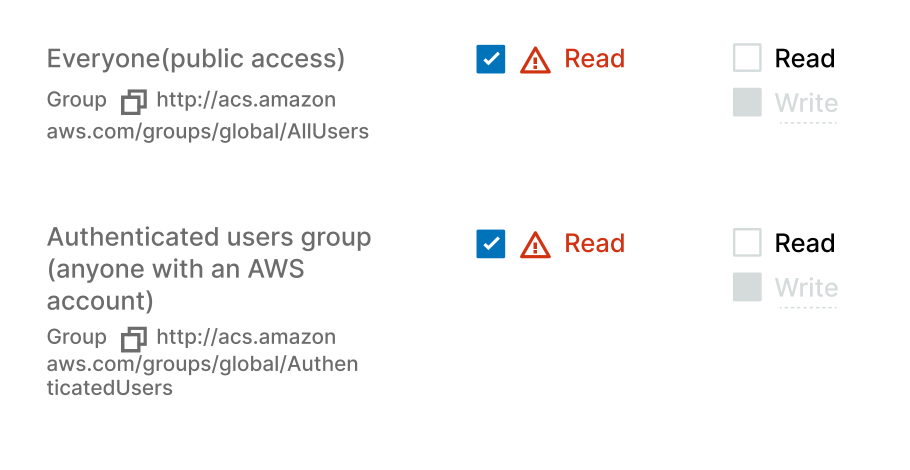
Sign in to the AWS Management Console and open the Amazon S3 console at https://console.aws.amazon.com/s3/.

- In the Bucket name list, **choose the name of the bucket** that you want.
- **Choose Permissions**.
- **Choose Edit** to change the public access settings for the bucket. For more information about the four Amazon S3 Block Public Access Settings, see Block public access settings.
- **Choose the setting** that you want to change, and then choose Save.
- When you're asked for confirmation, **enter confirm**. Then choose Confirm to **save your changes**.

> This is best for when you have a need for public s3 buckets in your environment, but the particular bucket in question should not be publicly accessible.

If you want to change public access for a single object (file) (this is also the only way to see if a file is publicly accessible):

- **Sign in** to the AWS Management Console and open the Amazon S3 console at https://console.aws.amazon.com/s3/
- In the Buckets list, **choose the name of the bucket** that contains the object.
- In the objects list, **choose the name of the object** for which you want to set permissions.
- **Choose Permissions**.
- Under **Access control list (ACL), choose Edit**.
- **Select the check boxes** for the permissions that you want to change, and then **choose Save**.
- Notice that if these boxes are selected (any one of them), anyone can access this object (since anyone can create a free AWS account)



Everyone(public access) Group http://acs.amazon aws.com/groups/global/AllUsers	☑ ⚠ Read	□ Read ▨ Write
Authenticated users group (anyone with an AWS account) Group http://acs.amazon aws.com/groups/global/AuthenticatedUsers	☑ ⚠ Read	□ Read ▨ Write

> This last mitigation and access control technique would need to be done on a continuous basis for any buckets with object-level exposure enabled where others have access, given that the owner cannot control those users and what files they may copy, post, share.

Use a data security posture management solution to find misplaced data.

If any of your buckets or objects are publicly accessible, the best practice for protecting data at scale is to use a DSPM solution. We don't want to get overly promotional in our research posts, so we'll leave it at that.

Summary

With the stakes as high as they are and attackers always on the lookout for ways they can gain a financial edge, it's vital to keep control over sensitive data. As we've uncovered, this includes ensuring that your data is not part of the 21% of sensitive PII unintentionally publicly exposed.

Fortunately, there are steps you can take to mitigate the threat of exposure,and we hope that this post helps, but constant vigilance is required.

With almost no effort on your part, Laminar can easily scan your environment for publicly exposed sensitive data for free. Contact us today for a free assessment.

Sources and more reading:

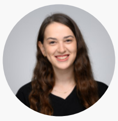
- 1. https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
- 2. https://www.statista.com/statistics/1258456/enterprise-data-storage-software-market-share-vendor-worldwide/#:~:text=Amazon%20S3%20led%20the%20global,of%20the%20market%20share%2C%20respectively
- 3. Skynet is the evil AI in the Terminator movies, see https://www.rottentomatoes.com/m/terminator
- 4. https://docs.aws.amazon.com/AmazonS3/latest/userguide/access-control-block-public-access.html
- 5. https://www.darkreading.com/application-security/cloud-misconfig-exposes-3tb-sensitive-airport-data-amazon-s3-bucket
- 6. https://www.bleepingcomputer.com/news/security/microsoft-data-breach-exposes-customers-contact-info-emails/
- 7. https://www.theregister.com/2022/12/20/mcgraw_hills_s3_buckets_exposed/
- 8. https://www.comparitech.com/blog/information-security/utah-covid-test-center-leak/
- 9. https://www.spiceworks.com/it-security/cyber-risk-management/articles/aws-misconfigurations-2021/

- https://repost.aws/knowledge-center/read-access-objects-s3-bucket
- https://docs.aws.amazon.com/AmazonS3/latest/userguide/configuring-block-public-access-account.html
- https://docs.aws.amazon.com/AmazonS3/latest/userguide/access-control-block-public-access.html

Authors

Gefen Frosh

Gefen is a data analyst at Laminar and a researcher in the Laminar Labs research arm of the company. Bringing her expertise as a former cyber analyst team leader in the Israel Defense Forces, she uses her knowledge of data analysis techniques and cloud technologies to discover new pathways for protecting data in the cloud.



Gali Lazarovsky

Gali is a data analyst at Laminar and a researcher in the Laminar Labs research arm of the company. As a data analytics expert and former cyber security analyst team lead in unit 8200 of the Israeli Defense Forces she brings her wealth of knowledge to bear to uncover new threats to data in the cloud and then find the solutions to them.