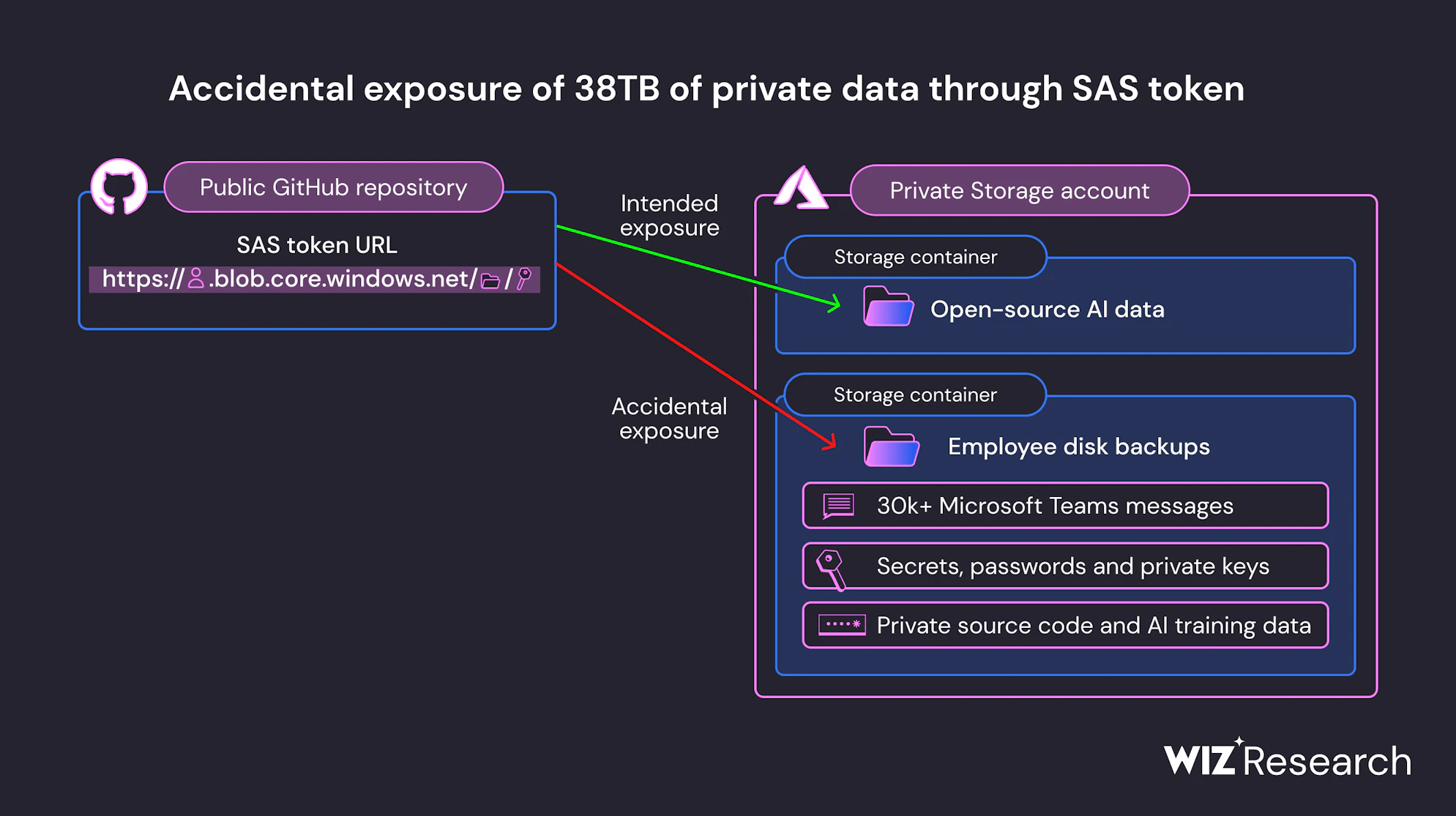
**The Microsoft AI Research Division exposed 38TB of sensitive data, by accident?**



## Since July 2020, Microsoft AI researchers have unintentionally disclosed 38TB of private information via a public GitHub repository.

## When publishing a crate of open-source training data on GitHub, the Microsoft AI research group unintentionally exposed 38TB of critical information, according to cybersecurity company Wiz.

## A disk backup of two employees' workstations that contained confidential information, private keys, passwords, and more than 30,000 internal Microsoft Teams messages was made public by the revealed data.

## The researchers used SAS tokens, an Azure feature that enables data sharing from Azure Storage accounts, to distribute their files. reads the report that Wiz issued.The access level can be restricted to only certain files, but in this instance, the link was set up to share the whole storage account, which included an additional 38TB of private information.

## While searching the Internet for improperly configured storage containers that can expose cloud-hosted data, Wiz Research Team came upon the repository. The specialists discovered a repository on GitHub with the name robust-models-transfer under the Microsoft organization.

## The repository belonged to Microsoft's AI research team, which made open-source code and AI image recognition models available through it. In July 2020, the Microsoft AI research team began disseminating data.

## Data held in Azure Storage accounts used by Microsoft's research team was shared via Azure SAS tokens.

## Private information was exposed because the Azure Storage signed URL used to access the repository was inadvertently set up to provide rights on the whole storage account.

## 

## 

## But access to this URL was not limited to open-source models. The storage account's entire configuration was set up to grant permissions, accidentally exposing more sensitive information. the business keeps going. "A significant data leak including over 38TB of personal data was caused by the straightforward act of sharing an AI dataset. The use of Account SAS tokens as the sharing mechanism was the primary culprit. SAS tokens are a security risk because to a lack of oversight and accountability, hence their use ought to be minimized.

## SAS tokens cannot be easily tracked, according to Wiz, because Microsoft does not offer a centralized method of managing them within the Azure site. According to Microsoft, the data lead did not provide any consumer information.

## "No client information was disclosed, and this problem did not endanger any internal services. In response to this issue, no customer action is necessary. reads the article that Microsoft released.

## The timeline of this security incident is provided below:

  **Jul. 20, 2020** – SAS token first [committed](https://github.com/microsoft/robust-models-transfer/blob/e61568d613c025adfd07f61f2639f3ae78852143/README.md?plain=1#L36) to GitHub; expiry set to Oct. 5, 2021

 **Oct. 6, 2021** – SAS token expiry [updated](https://github.com/microsoft/robust-models-transfer/commit/a9e0e80bcd49bd8651c0b3198c7dc89179b2c0ac) to Oct. 6, 2051

 **Jun. 22, 2023** – Wiz Research finds and reports issue to MSRC

 **Jun. 24, 2023** – SAS token invalidated by Microsoft

 **Jul. 7, 2023** – SAS token [replaced](https://github.com/microsoft/robust-models-transfer/commit/c26ebfff3d01bd2a52a6c14febb8b7ea0234431d) on GitHub

 **Aug. 16, 2023** – Microsoft completes internal investigation of potential impact

##  **Sep. 18, 2023** – Public disclosure

## Credits:

<https://securityaffairs.com/151004/data-breach/microsoft-ai-data-leak.html>