



Data at Microsoft

Christopher MANEU
Data Engineering & Analytics Lead Cloud Advocate @ Microsoft
@cmaneu

Agenda

Data by the numbers

Data products portfolio

Some innovation around Data

Data at Microsoft

By the numbers

Data at Microsoft by the numbers

600 000 000 000

Number of security events analyzed/month, only on Microsoft own network

100 Tb

Public maximum database size (Azure SQL)

100k

users/min on UK Covid dashboard (Azure Postgres)

1M

of Database on ONE product (Dynamics)

3661

Open jobs with "Data Engineer"

10 Billions

Inferences per month on Azure Machine Learning
53 Billions on Cognitive Services

Data products portfolio

**Microsoft, the biggest data
company in the world?**



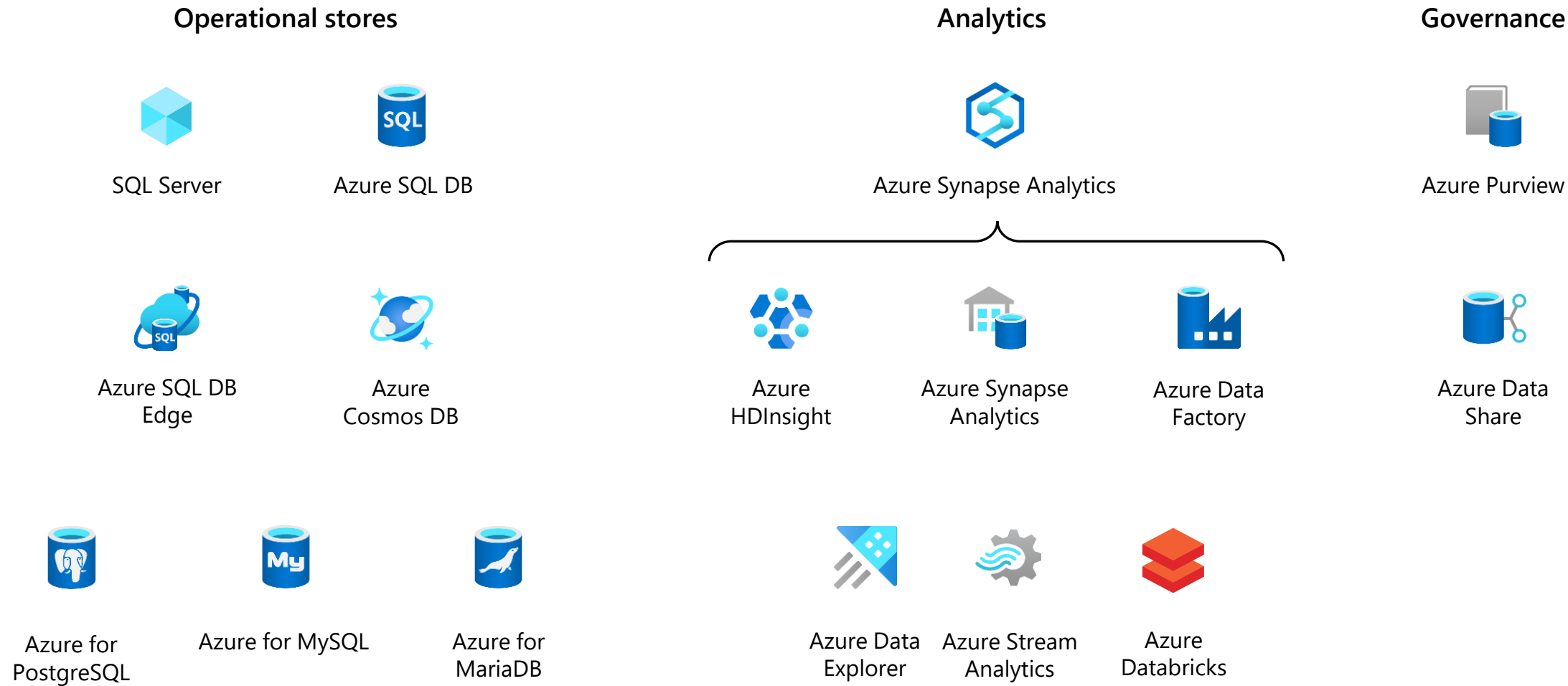
Microsoft Data Portfolio

Databases: SQL Server, Citus Postgresql

Business Analytics tools: Excel, Power BI

Cloud offerings

Azure Data & Analytics Products and Services



Azure storage portfolio



Disk Storage

Ultra
Premium
Standard



Object Storage

Azure Blobs
(with Data Lake Gen 2)



File Storage

Azure Files
Azure NetApp Files



Backup

Azure Backup
Azure Site Recovery



Data Transport

Azure Data Box
Data Box Disk
Data Box Heavy



Hybrid Storage

Stack Edge
Databox Gateway
Azure File Sync
Azure HPC Cache

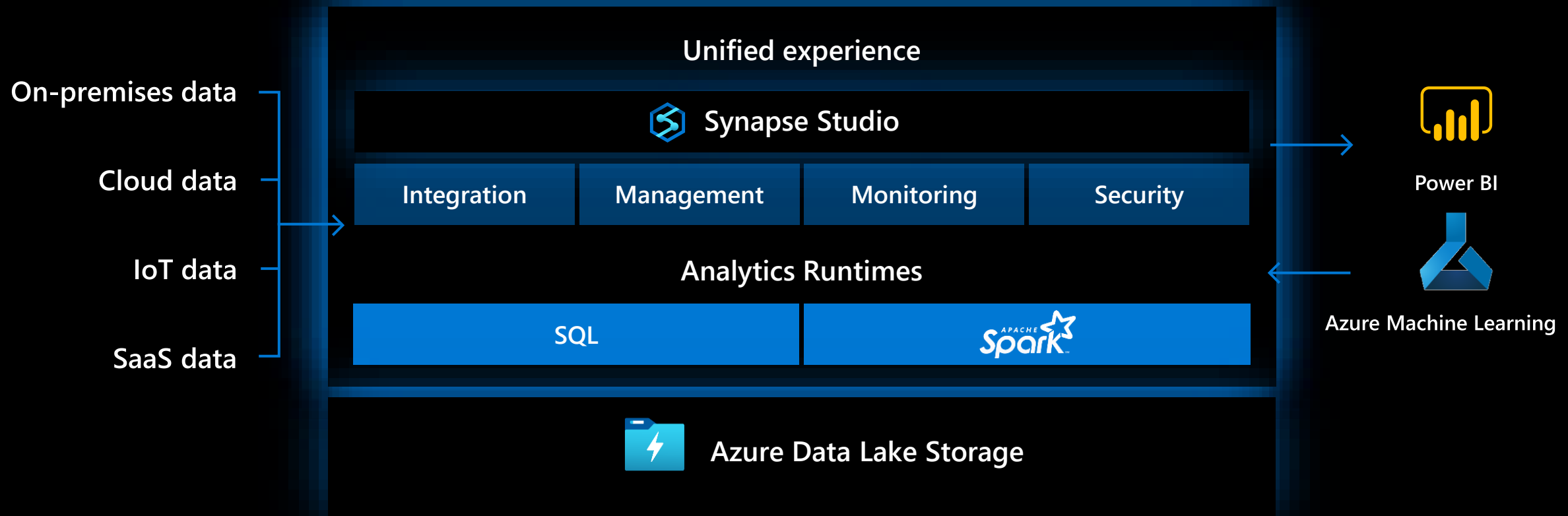


Future Technology

Project Silica
Holographic Storage
Palix

Azure Synapse Analytics

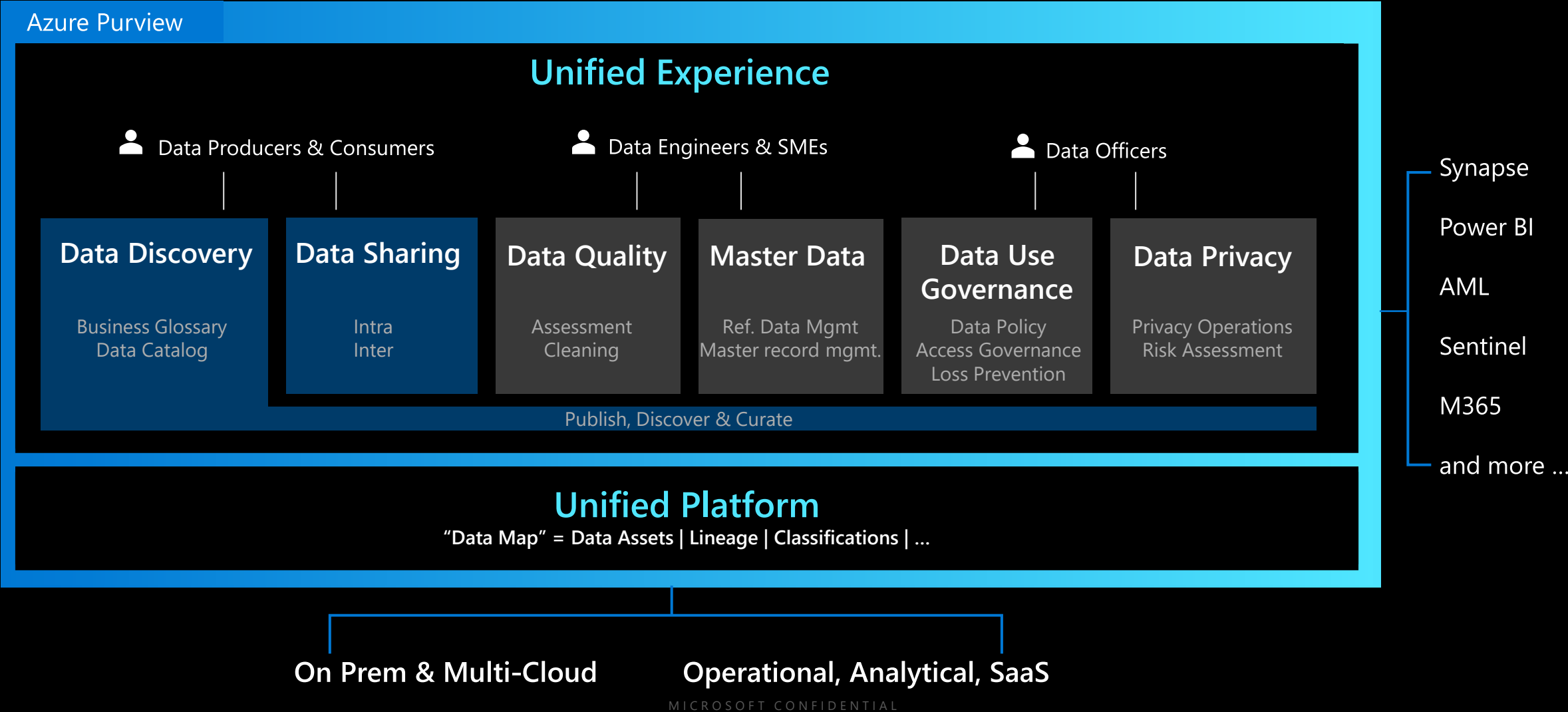
Limitless analytics service with unmatched time to insight



Data Governance

A unified approach to data governance

Roadmap



SQL Server 2019

Industry-leading performance and security, with intelligence over all your data

Intelligence over any data



AI and Machine Learning over all data with the power of SQL and Spark

Choice of platform and language



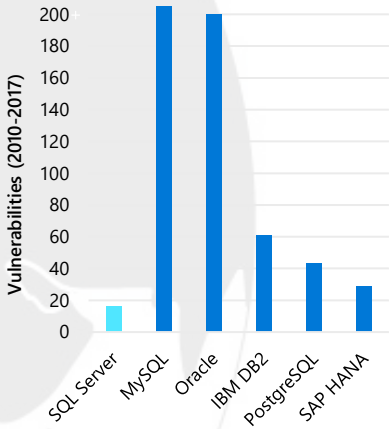
T-SQL	PHP	Python
Java	Node.js	Ruby
C/C++	C#/VB.NET	

Industry-leading performance



#1 OLTP performance¹
#1 DW performance on 1TB², 10TB³, and 30TB⁴
Intelligent Query Processing

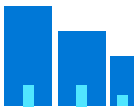
Most secure over the last 8 years⁵



Insights in minutes and rich reports



The best of Power BI and SQL Server Reporting Services with Power BI Report Server



Private cloud

In-memory across all workloads

Most consistent data platform



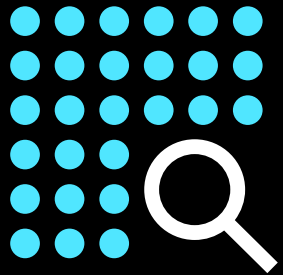
Public cloud

1/10th the cost of Oracle

All TPC Claims as of 1/19/2018.

¹ <http://www.tpc.org/4081>; ² <http://www.tpc.org/3331>; ³ <http://www.tpc.org/3326>; ⁴ <http://www.tpc.org/3321>; ⁵ National Institute of Standards and Technology Comprehensive Vulnerability Database

Azure SQL Database ledger – The power of blockchain in SQL



Makes data in SQL
tamper-evident
through
cryptography



Provides a historical
record of all changes,
verified through
cryptographic proofs



The same SQL you
already know across
Azure and on-
premises

Creating an account balance updatable ledger table

```
CREATE TABLE [Account].[Balance]
(
    [CustomerID] INT NOT NULL PRIMARY KEY CLUSTERED,
    [LastName] VARCHAR (50) NOT NULL,
    [FirstName] VARCHAR (50) NOT NULL,
    [Balance] DECIMAL (10,2) NOT NULL
)
WITH
(
    SYSTEM_VERSIONING = ON,
    LEDGER = ON
);
GO
```

	ledger_table_name	history_table_name	ledger_view_name
1	Account.Balance	Account.MSSQL_LedgerHistoryFor_1525580473	Account.Balance_Ledger



Azure Data Box – Family Summary

Offline Data Transfer

Online Data Transfer



Data Box

- Capacity: 100 TB
- Weight: ~50 lbs
- Secure, ruggedized appliance
- GA September 2018



Data Box Disk

- Capacity: 8 TB ea.; 40TB/order
- Secure, ruggedized USB drives orderable in packs of 5 (up to 40TB).
- GA Jan 2019



Data Box Heavy

- Capacity: 1 PB
- Weight 500+ lbs
- Secure, ruggedized appliance
- GA Expected June 2019



Data Box Gateway

- Virtual device provisioned in your hypervisor
- Supports storage gateway, SMB, NFS, Azure blob, files
- GA March 2019



Data Box Edge

- Local Cache Capacity: ~12 TB
- Includes Data Box Gateway and Azure IoT Edge.
- GA March 2019



Order



Send



Fill



Return



Upload

Network Data Transfer



Cloud to Edge



Edge to Cloud



Pre-processing



ML Inferencing

Edge Compute



Use case: Smart City - Intelligent Crosswalks

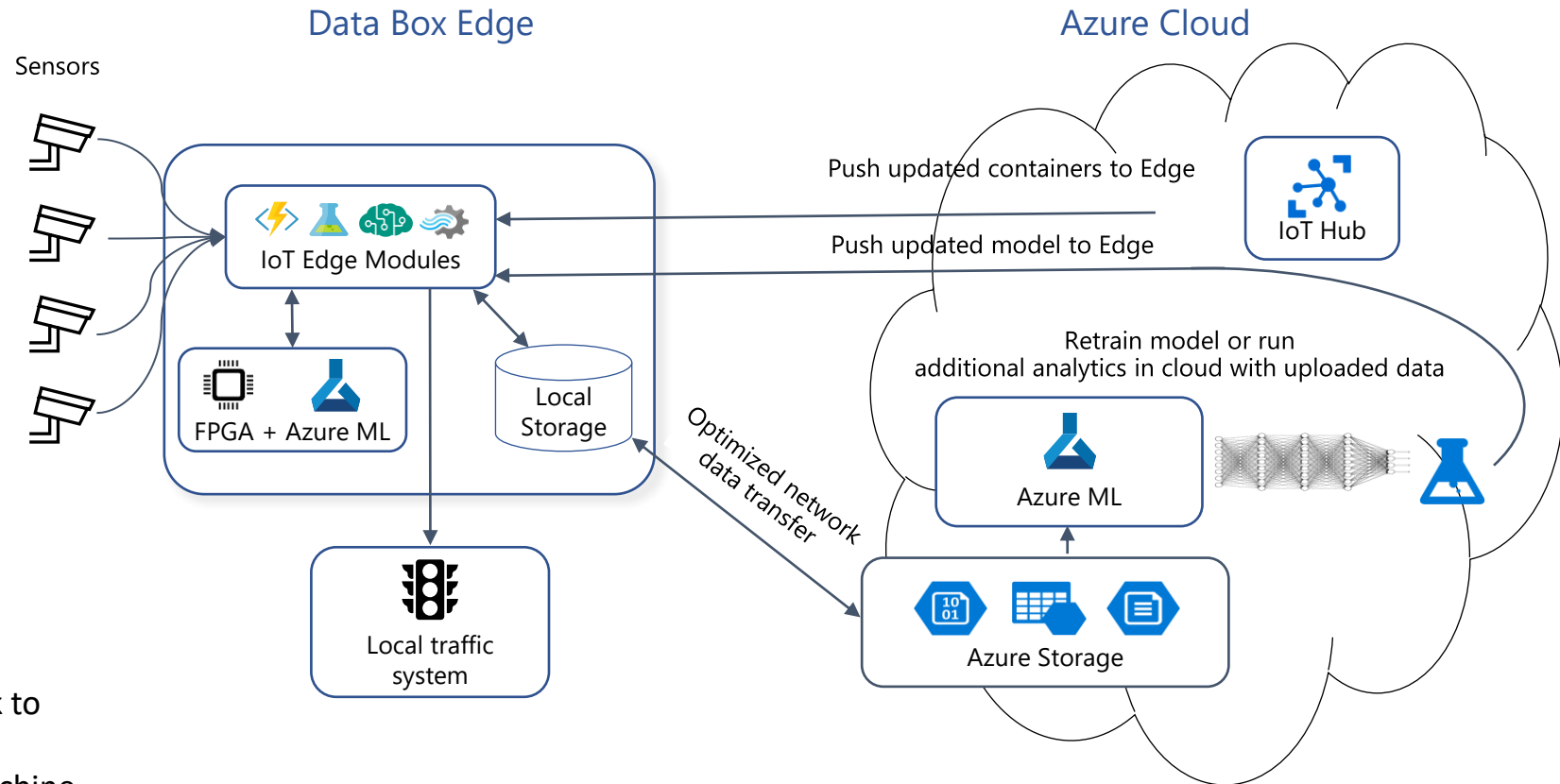
Improve crosswalk safety by extending crossing time as needed

What problem does this solve?

- Reduce risk of harm to pedestrians from insufficient time to cross
- Reduce risk of vehicle accidents with improved traffic light management

What technology does this use?

- IP cameras send data over the local network to Data Box Edge
- Containerized application uses a custom machine learning model to identify pedestrians in need of additional crossing time
- Optional: transfer data to Azure for to retrain and improve your ML model or for additional analytics



Some innovation around Data

Preserving most important opensource code







LOCKSS

Lots Of Copies Keeps Stuff Safe



GITHUB

On every push to GitHub, we replicate your Git data to multiple datacenters around the world. Additionally, we store backups of Git data, Issues, Pull Requests, and all of your data on GitHub in multiple locations. All of this data is available live via the GitHub API.



GH TORRENT

GHTorrent monitors the GitHub public event timeline, archives those events, and makes them queryable using BigQuery. You can also download snapshots by hour, day, or month.



GH ARCHIVE

GHArchive monitors the GitHub public event timeline, archives those events, and recursively crawl and archive their contents and dependencies. Those archives will then be made available for download on a daily or monthly basis.



Hot // Near real-time



Warm // Updated monthly to yearly



Cold // Updated every 5+ years



INTERNET ARCHIVE

The Internet Archive's well-known Wayback Machine will crawl GitHub's public repositories—including new repositories, issues, pull requests, wikis, and more—and store copies on hard drives in San Francisco and other locations. These archives will be publicly available via `git` and `https`.



SOFTWARE HERITAGE FOUNDATION

The Software Heritage Foundation will crawl GitHub on a regular basis and add its public repos to their archive, to which they provide public API access.



ARCTIC WORLD ARCHIVE

On February 2, 2020, GitHub will capture a snapshot of every active public repository, to be preserved in the GitHub Arctic Code Vault. This data will be stored on 3,500-foot film reels, provided and encoded by Piql, a Norwegian company that specializes in very-long-term data storage. The film technology relies on silver halides on polyester. This medium has a lifespan of 500 years as measured by the ISO; simulated aging tests indicate Piql's film will last twice as long.



BODLEIAN LIBRARY

Oxford University's Bodleian Library will provide redundancy for the Arctic Code Vault by keeping GitHub's 10,000 most-starred and most-depended-upon repositories in their depository as duplicate Piql film reels.



BIBLIOTHECA ALEXANDRINA

The Bibliotheca Alexandrina will provide redundancy for the Arctic Code Vault by keeping open source's "greatest hits" containing 17,000 of the most-starred and most-depended-upon repositories in their depository as duplicate Piql film reels.



STANFORD LIBRARY

Stanford Library will provide redundancy for the Arctic Code Vault by keeping open source's "greatest hits" containing 17,000 of the most-starred and most-depended-upon repositories in their depository as duplicate Piql film reels.



PROJECT SILICA: MICROSOFT RESEARCH

The GitHub Archive Program is partnering with Microsoft's Project Silica to ultimately archive all active public repositories for over 10,000 years, by writing them into quartz glass platters using a femtosecond laser.

**Solving future data density
needs**

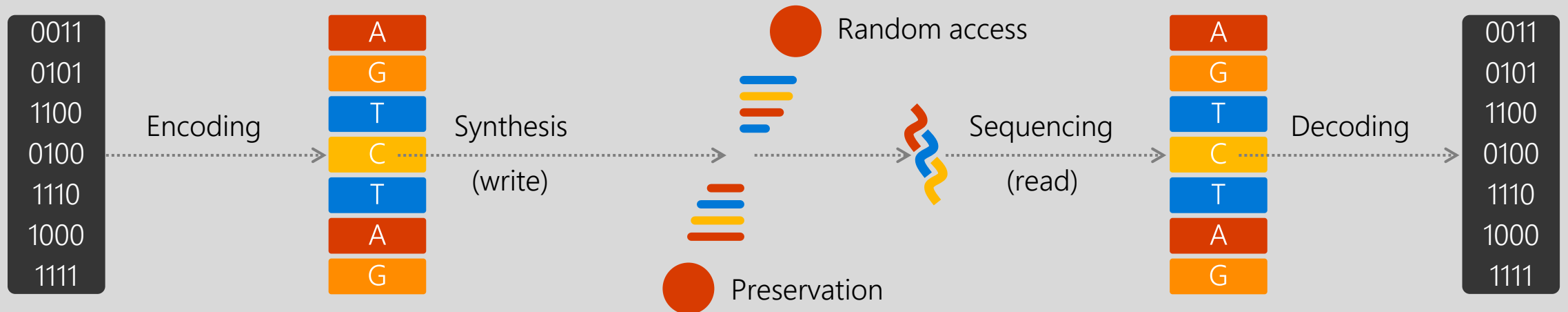
Project Palix: DNA Data storage

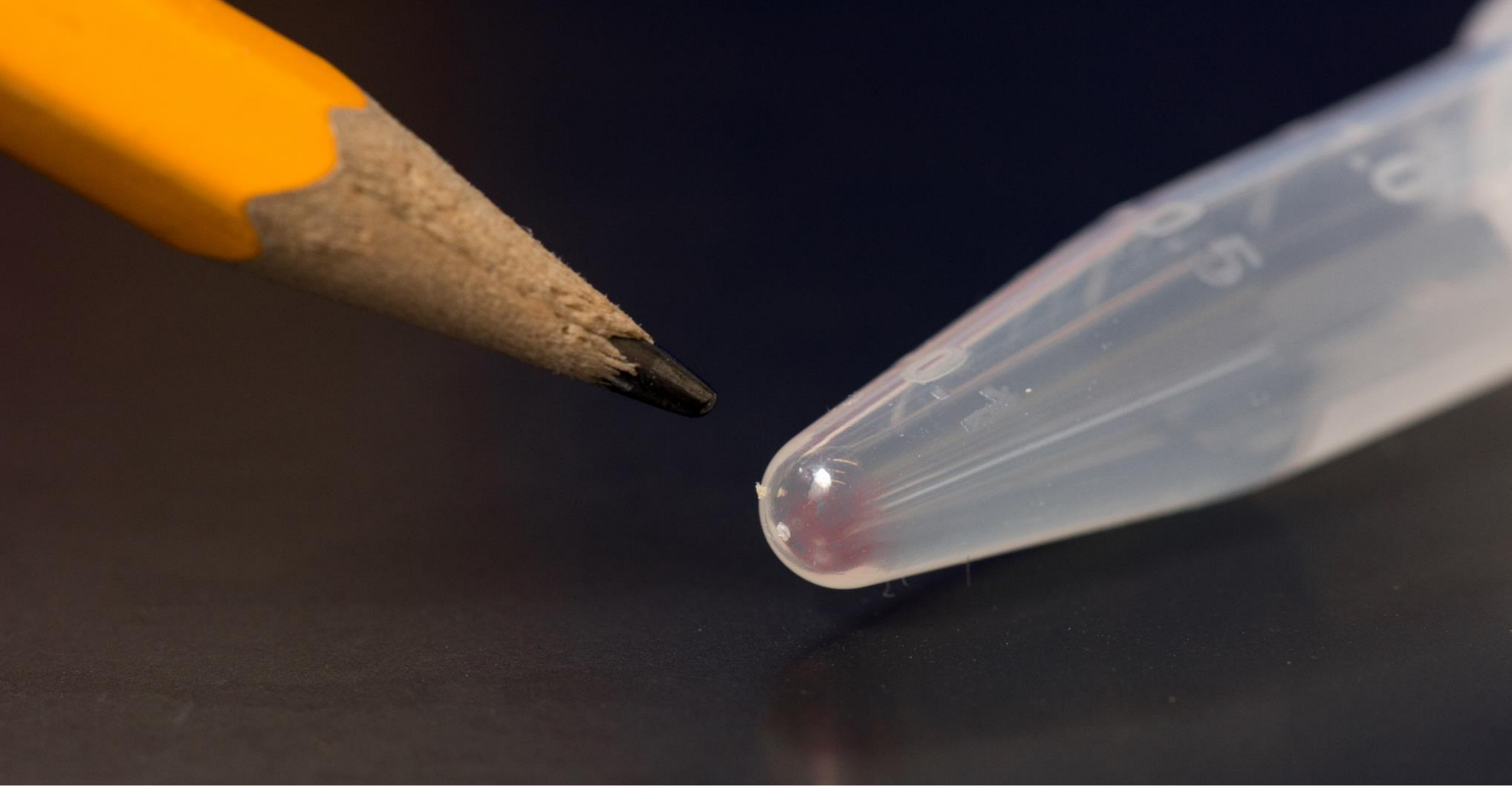
Joint project between Microsoft and University of Washington

>1,000x denser than tape

2,000 years in ~10°C

Bits	00	01	10	11
Bases	A	C	G	T





Credit: Tara Brown Photography/University of Washington

First fully automated DNA data storage system

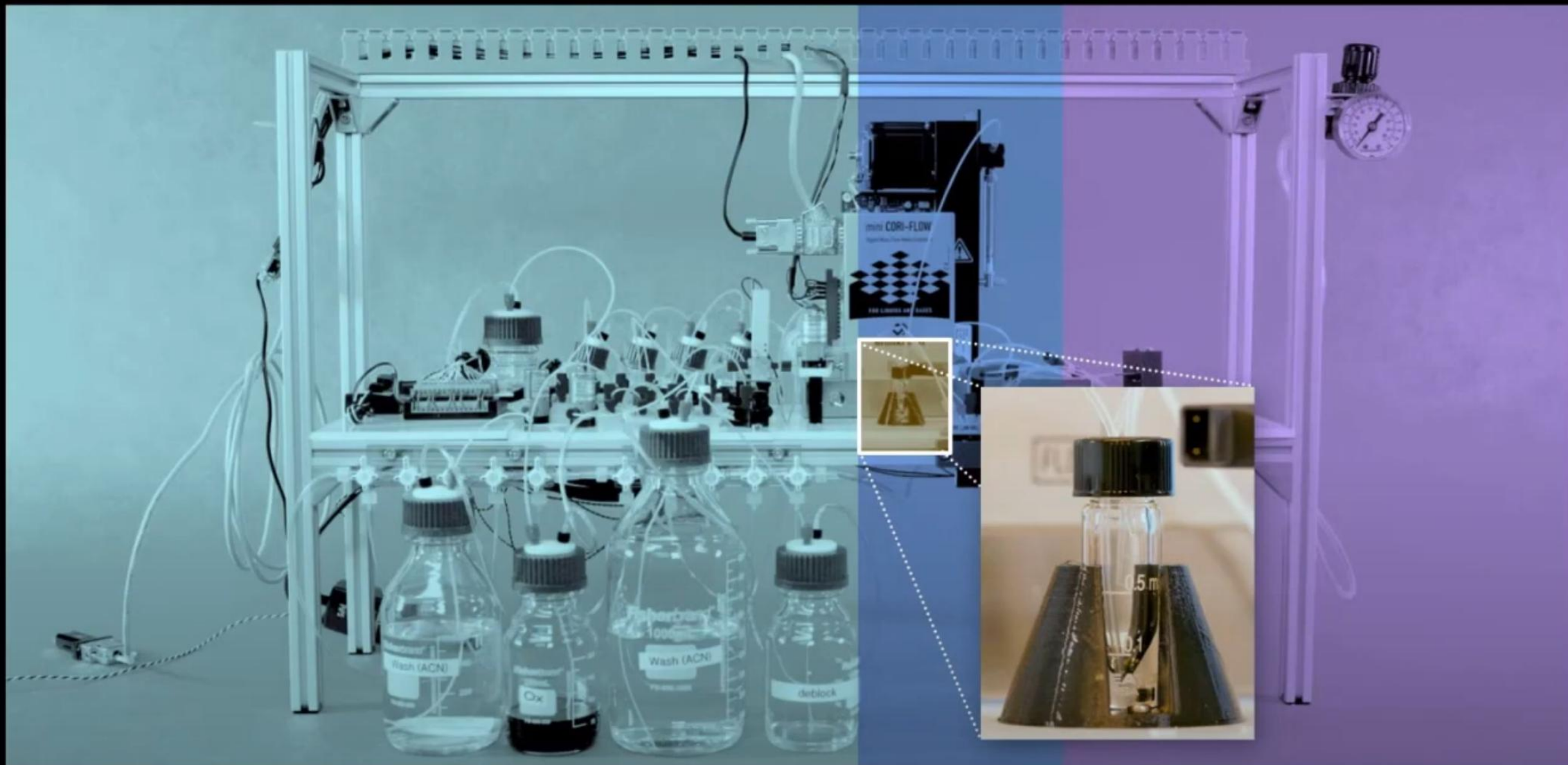


First fully automated DNA data storage system

SYNTHESIS

STORAGE
+ PREP

SEQUENCING

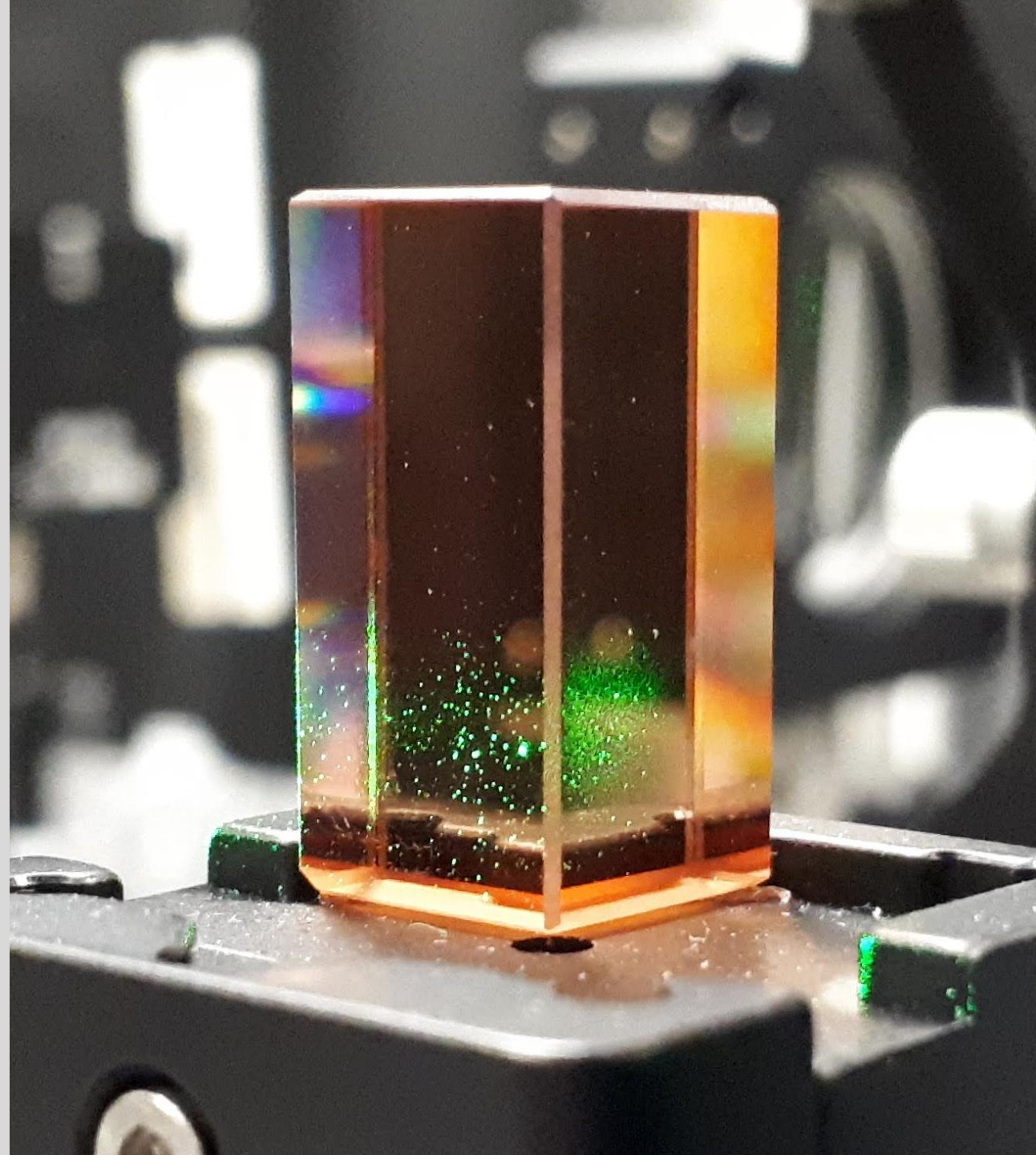


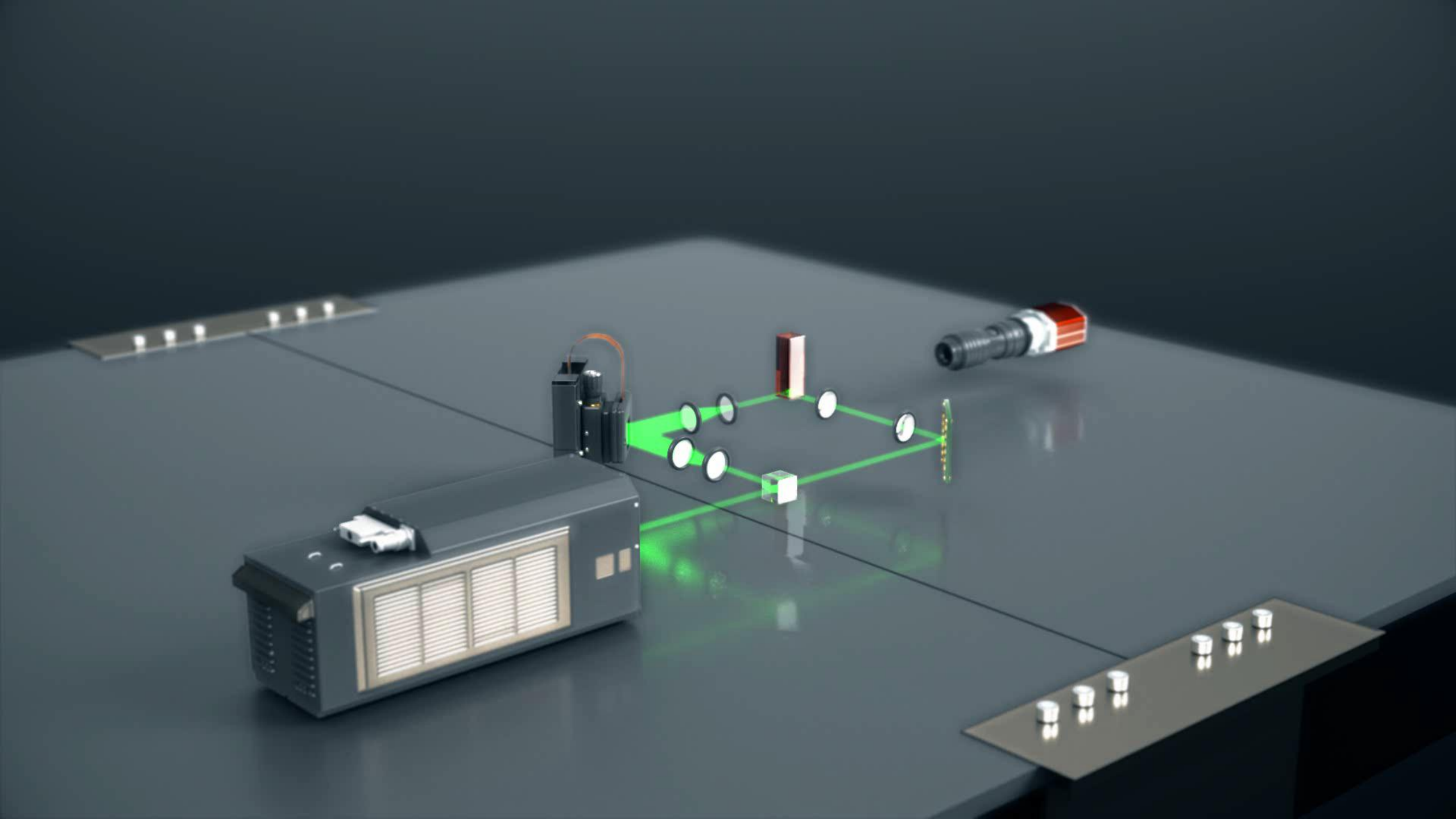
Project Palix - references

- Glass: A New Media for a New Era?
<http://rowstron.azurewebsites.net/Publications/hotstorage18-silica.pdf>
- A DNA-based archival storage system
<https://www.microsoft.com/en-us/research/publication/dna-based-archival-storage-system/>
- Scaling up DNA data storage and random-access retrieval
<https://www.microsoft.com/en-us/research/publication/scaling-dna-data-storage-random-access-retrieval/>
- Clustering billions of reads for DNA data storage
<https://www.microsoft.com/en-us/research/publication/clustering-billions-of-reads-for-dna-data-storage/>

Holographic storage

- Invented in '60





Holographic storage

- Invented in '60
- Reading & writing comparable with classic hard drives
- Software compensation with deep learning

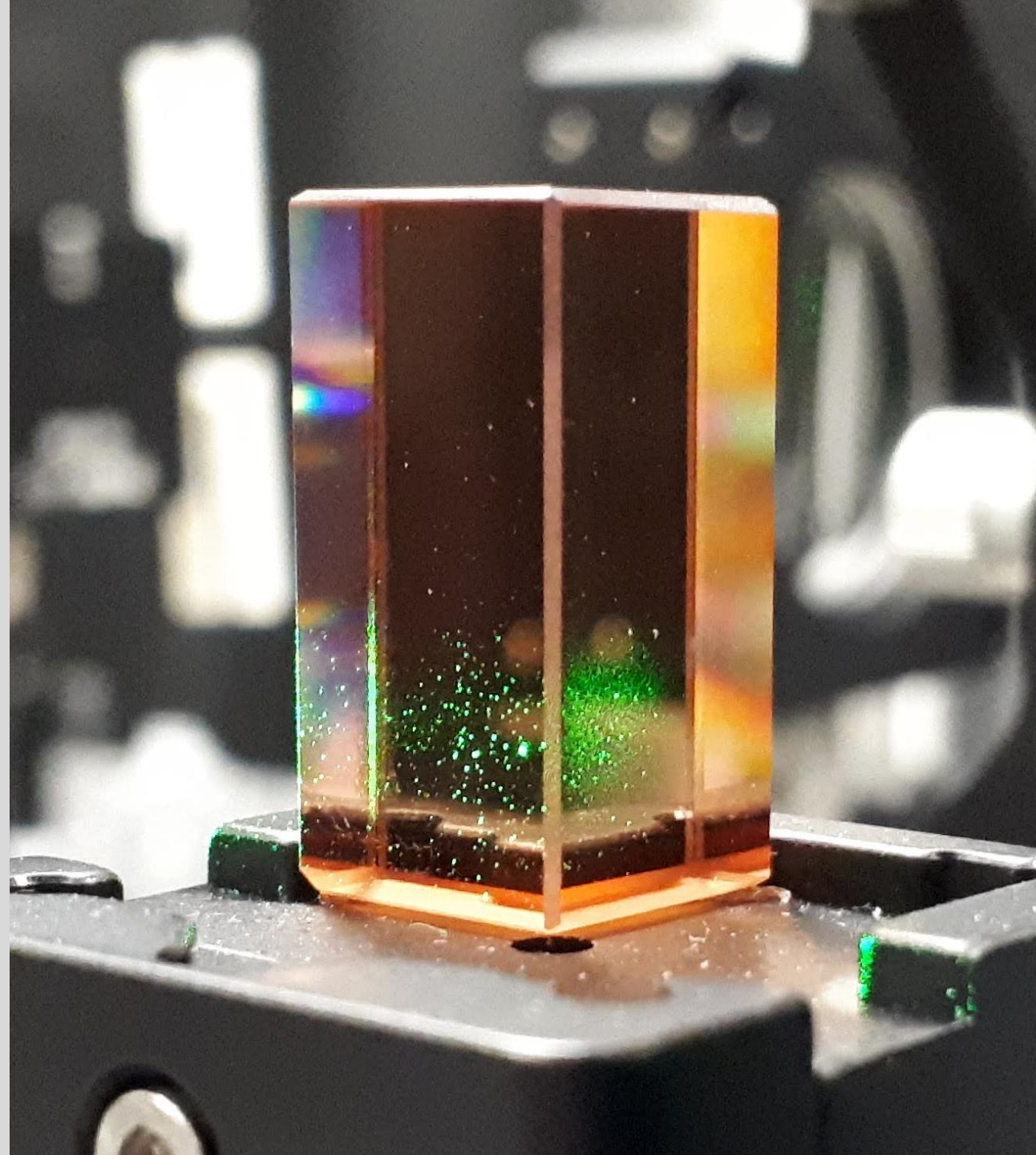


Image credits



Photo by [Glen Carrie](#) on [Unsplash](#)



Thanks !

<https://maneu.net>

Christopher MANEU
Data Engineering & Analytics Lead Cloud Advocate @ Microsoft
@cmaneu