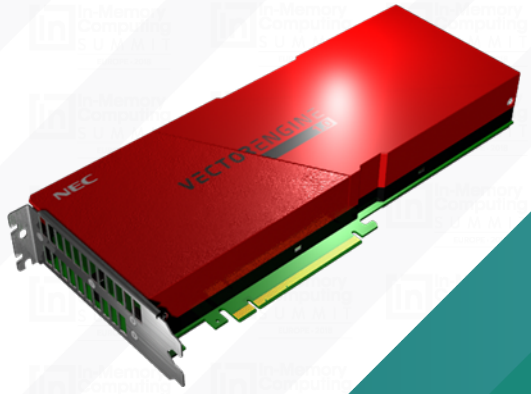
In-Memory Computing SUMMIT EUROPE 2018



High Performance Computing and Big Data

TSVI LEV

NEC, HPC, Big Data Analytics

NEC is an HPC systems provider with a new line of HPC processors and servers

Recent trends in **Big Data Analytics** require heavier **CPU loads**.

However, the fit may **require S/W development**

NEC Labs investigated potential use cases and loads

Conclusion: strong potential in Cyber, Privacy, Fraud

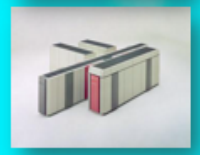
Feature	HPC	Big Data
Functionality	Heavy physical numerical simulations – weather etc.	Heavy database operations – ingest, unions etc.
Customers	Governments/Universities	Enterprises
Performance Metrics	FLOPS	Transactions per second
Traditional H/W	Vector Processors	Mainframes
Current H/W	CPU+GPU arrays	CPU arrays
Software used	Proprietary/custom	OSS – Hadoop, Spark
CIA	Not a major factor	Critical factor
Bottlenecks	CPU, now storage too	Storage, network, DB, implementation layer

History of SX Vector Supercomputer

SX-2 1983

Technology:	Bipolar
CPU Frequency:	166 MHz
CPU Performance:	1.3 GFlops
CPU Memory Bandwidth:	10.7 GB/sec

SX-3 1989

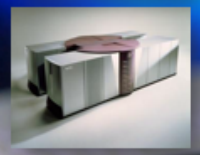


Technology:	Bipolar
CPU Frequency:	340 MHz
CPU Performance:	5.5 GFlops
CPU Memory Bandwidth:	12.8 GB/sec

SX-4 1994

Technology:	350 nm
CPU Frequency:	125 MHz
CPU Performance:	2.0 GFlops
CPU Memory Bandwidth:	16.0 GB/sec

SX-5 1998



Technology:	250 nm
CPU Frequency:	250 MHz
CPU Performance:	8.0 GFlops
CPU Memory Bandwidth:	64.0 GB/sec

SX-6 2001

Technology:	150 nm
CPU Frequency:	500 MHz
CPU Performance:	8.0 GFlops
CPU Memory Bandwidth:	32.0 GB/sec

SX-7 2002

Technology:	150 nm
CPU Frequency:	552 MHz
CPU Performance:	8.8 GFlops
CPU Memory Bandwidth:	35.3 GB/sec

SX-8 2004

Technology:	90 nm
CPU Frequency:	1.0 GHz
CPU Performance:	16.0 GFlops
CPU Memory Bandwidth:	64.0 GB/sec

SX-9 2007

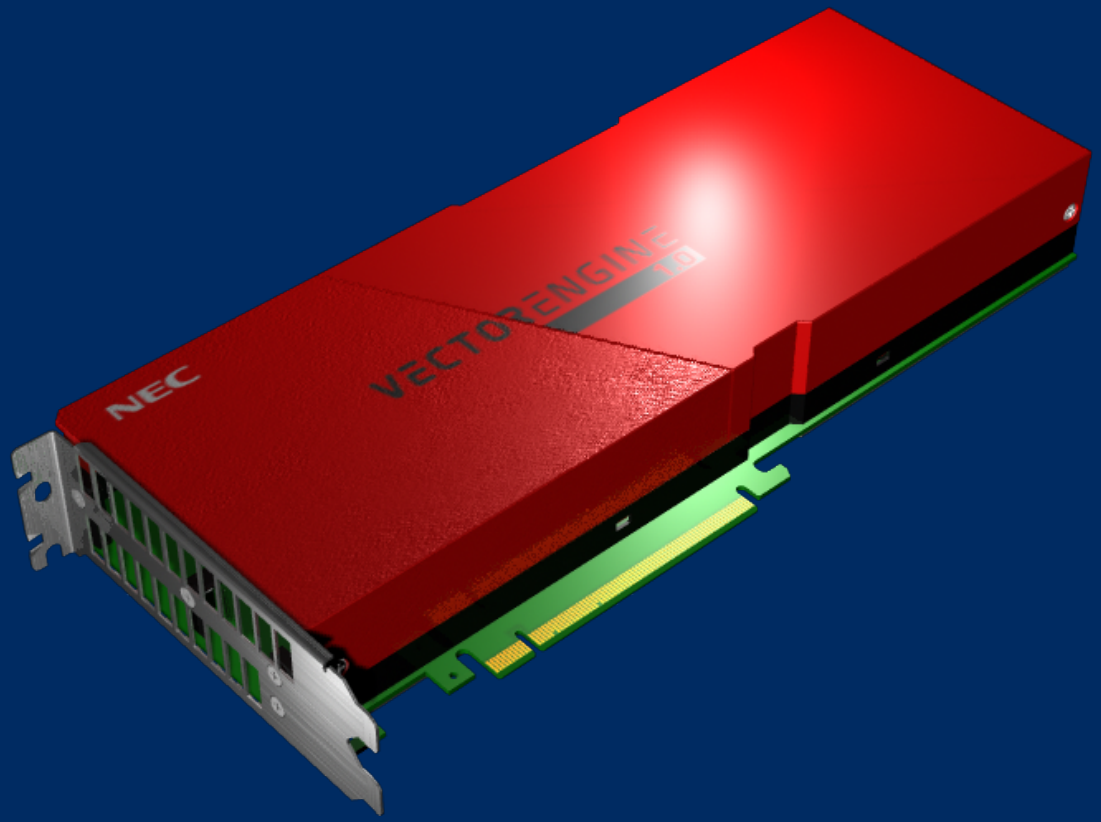
Technology:	65 nm
CPU Frequency:	3.2 GHz
CPU Performance:	102.4 GFlops
CPU Memory Bandwidth:	256.0 GB/sec

SX-ACE® 2013

Technology:	28 nm
CPU Frequency:	1.0 GHz
CPU Performance:	256.0 GFlops
CPU Memory Bandwidth:	256.0 GB/sec

Over 30 years Experience For High Sustained Performance

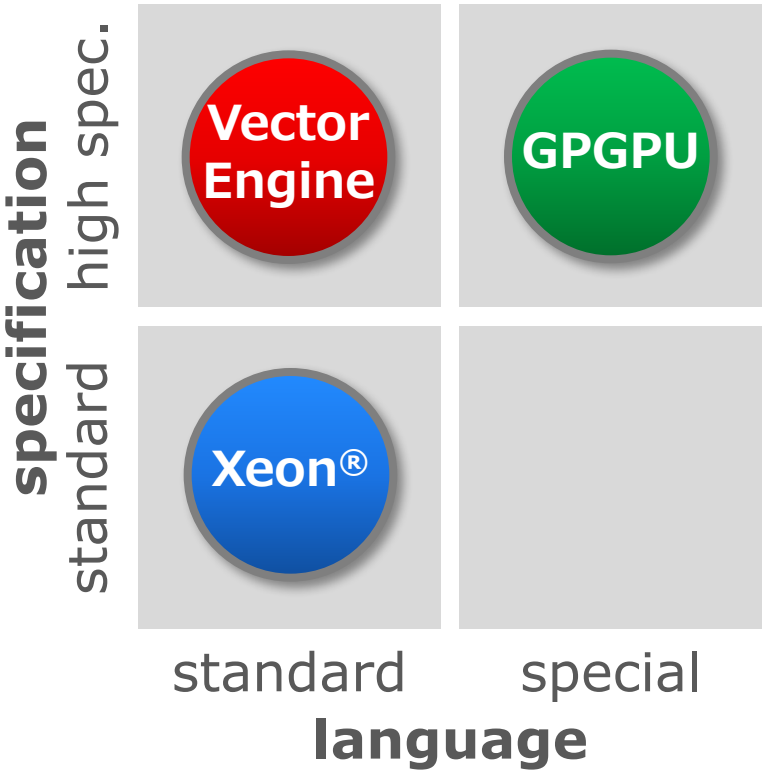
Vector Processor on The Card



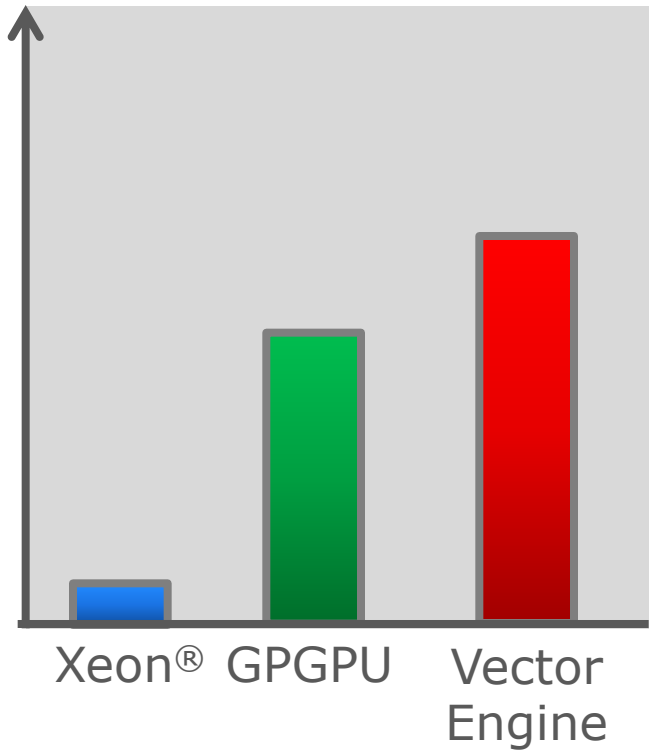
- New Developed Vector Processor
- PCIe Card Implementation, but not an accelerator
- 8 cores / processor
- 2.45TF performance
- 1.2TB/s memory bandwidth
- Normal programing with Fortran/C/C++

Usability x High Memory Bandwidth

Position

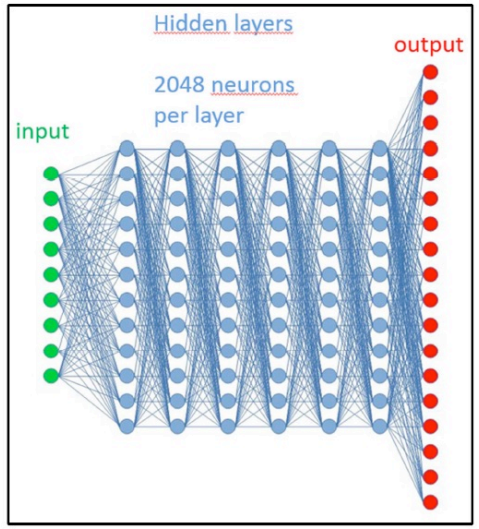


Memory bandwidth / processor

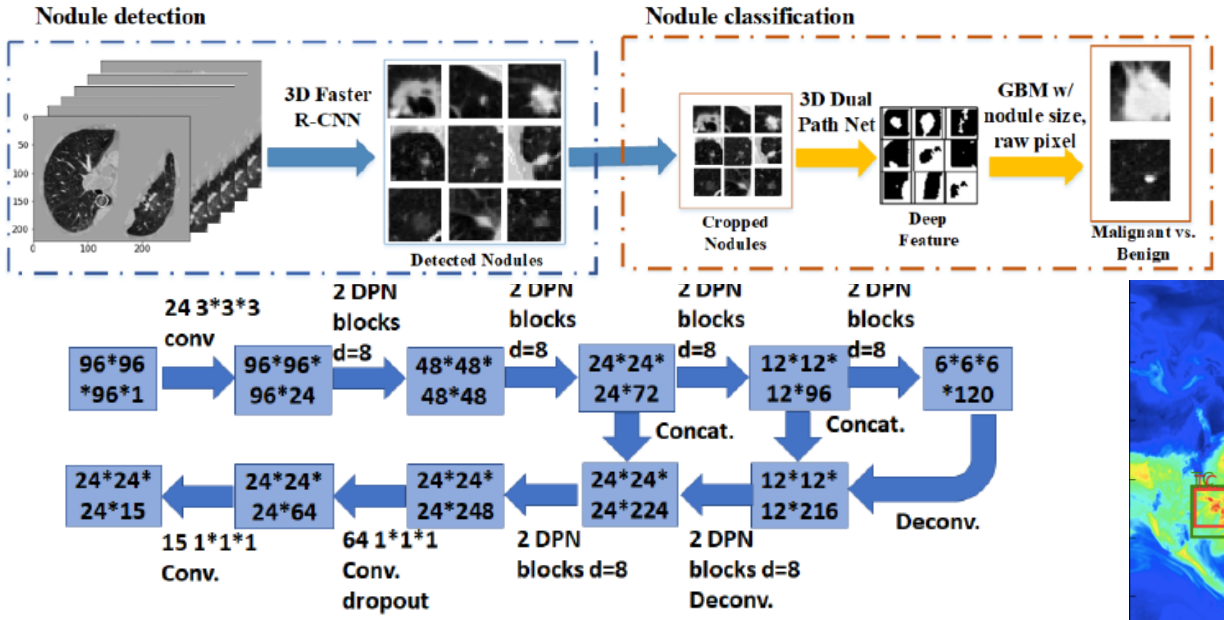


GPGPU vs CPU/Vector Processor

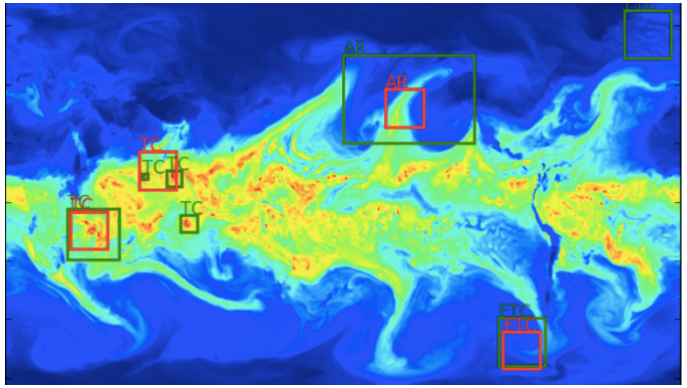
Multi Layer FC layers
for speech recognition



3D Convolutions 64X64X64=256Kbyte

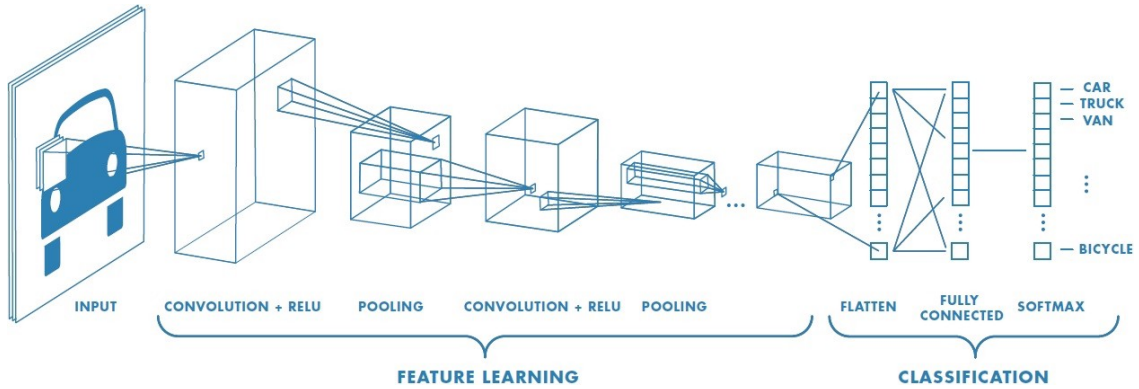


3D CNN and multilayer FC => CPU/Vector



2D Convolutions 3X3X256=2K byte

50-layer	101-layer	152-layer
7×7, 64, stride 2		
3×3 max pool, stride 2		
[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
erage pool, 1000-d fc, softmax		



2D CNN => GPU

Recommendation and Next Best Offer

Application: Recommendation Engines/Collaborative Filtering

Characteristics: Many user features with many options to choose from

Algorithms: Matrix factorization/ALS

Data: Requires a lot of training samples.

Fintech relevance: fraud detection, user profiling, HFT

Repetition: Probably weekly or slower

HPC load: - <10% of the ingestion/preparation time per cycle.

Result: Potential match in fast changing environments

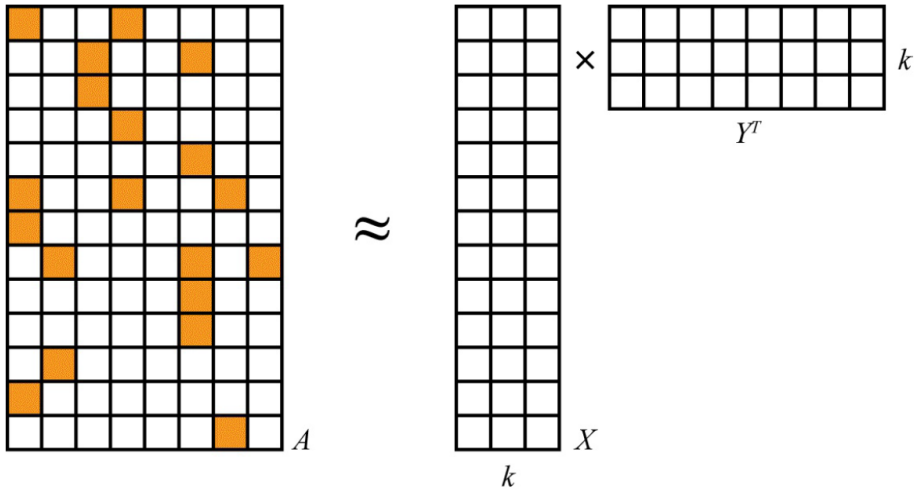
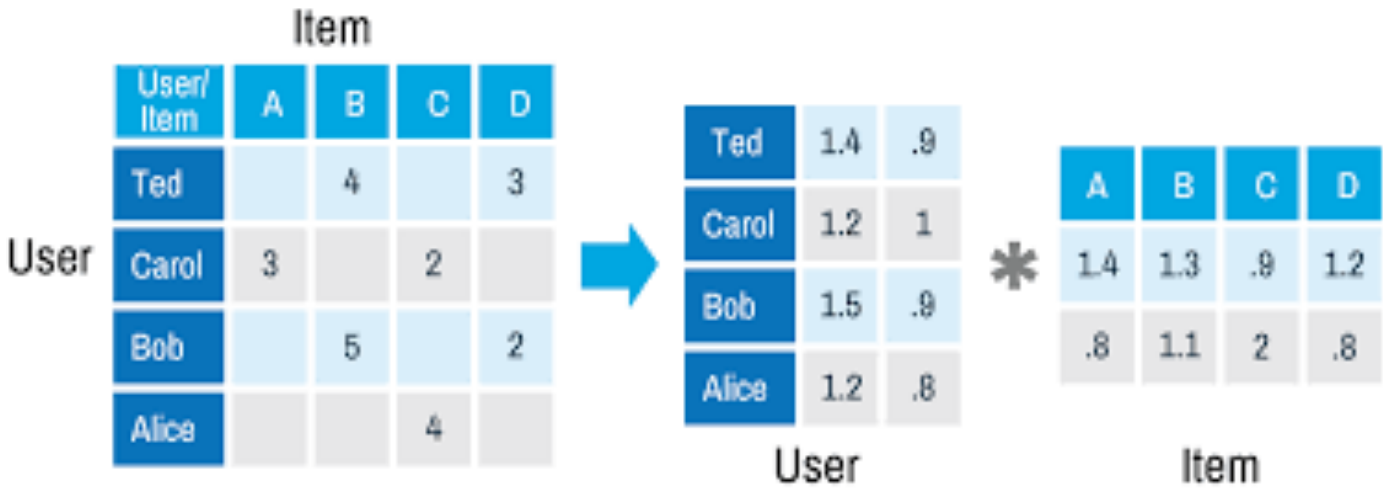


Figure 3-1. Matrix factorization



Cyber and Fraud Detection

Application: Anomaly Detection – large event sets of 'normal' behavior and seeking the odd events –e.g. Cyber, Fraud.

Characteristics: Large stream of events

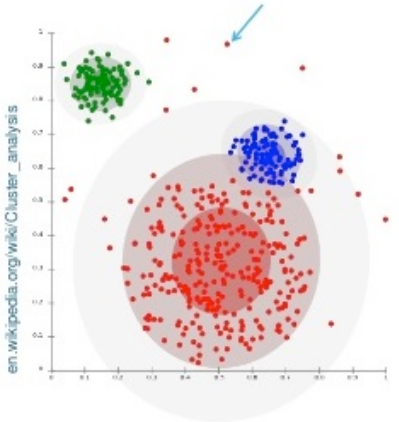
Algorithms: Clustering and DL (Auto-Encoders)

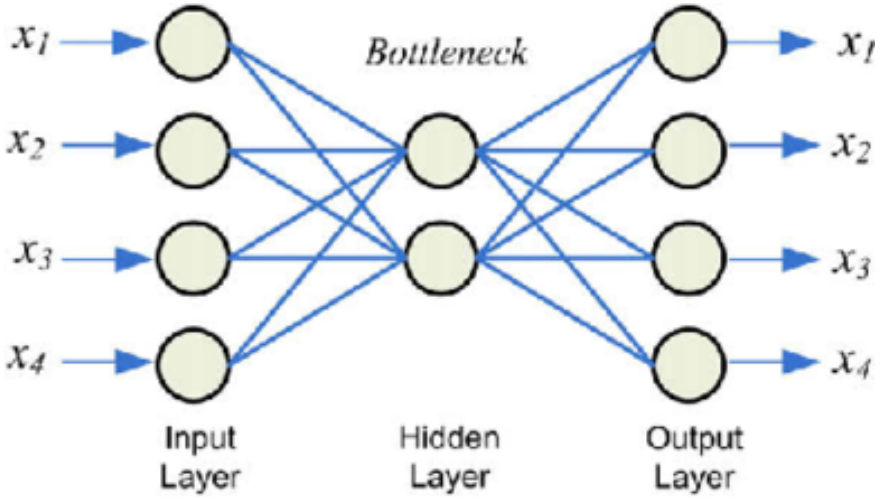
Fintech relevance: fraud detection, SIEM monitoring

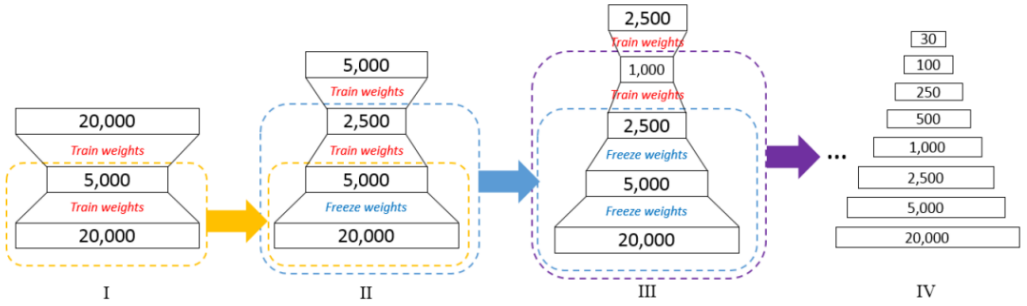
Repetition: Requires frequent re-training (daily)

HPC load: >50% of the ingestion/preparation time

Result: Strong match for cyber, fraud detection.







Document classification/trend tracking

Application: Semantic Analysis– matching documents to topics and classifying documents.

Characteristics: Large sets of documents, slowly changing

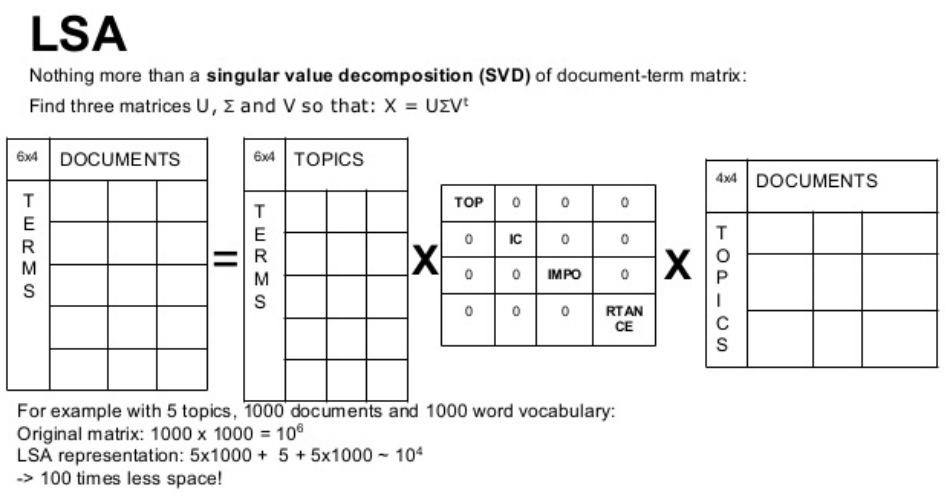
Algorithms: Singular Value Decomposition

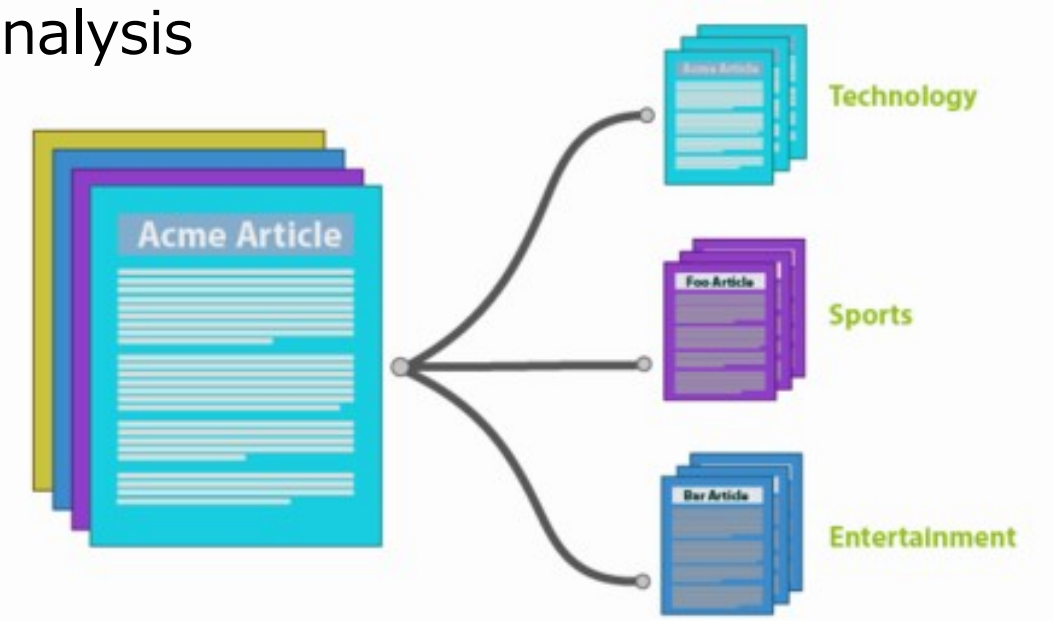
Fintech relevance: Forums/Social/News trends

Repetition: May require frequent re-training (daily)

HPC load: >50% of the ingestion/preparation time per cycle.

Result: Strong match for trend and sentiment analysis





Stress tests/simulations

Application: Monte Carlo simulations for variance estimation of complex variables

Characteristics: Small data set but many scenarios

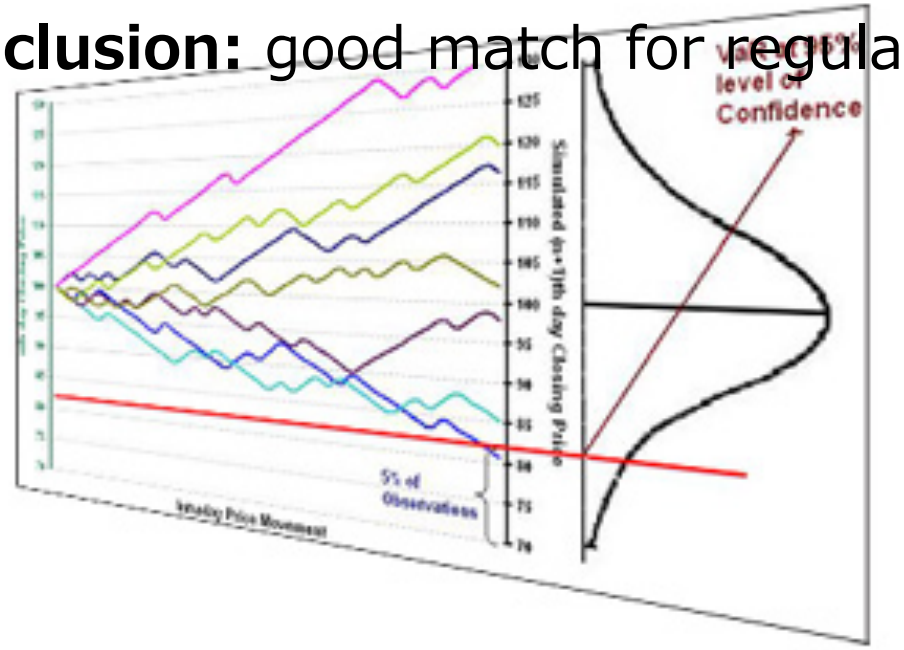
Algorithm: Linear & Logistic Regression, shallow Neural Networks

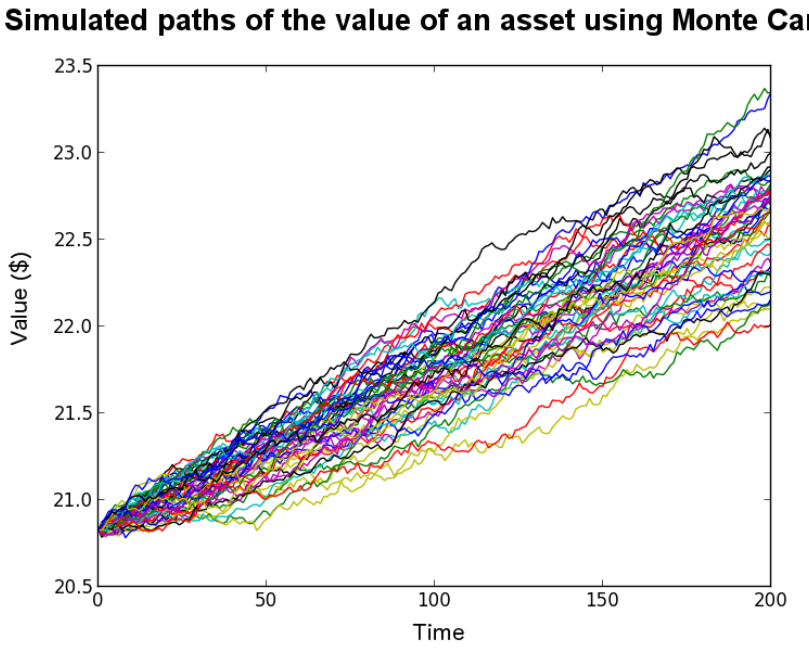
Fintech relevance: VaR, xVA simulations

Repetition: May require frequent re-training (daily)

HPC load: >80% of the ingestion/preparation time per cycle.

Conclusion: good match for regulatory compliance.

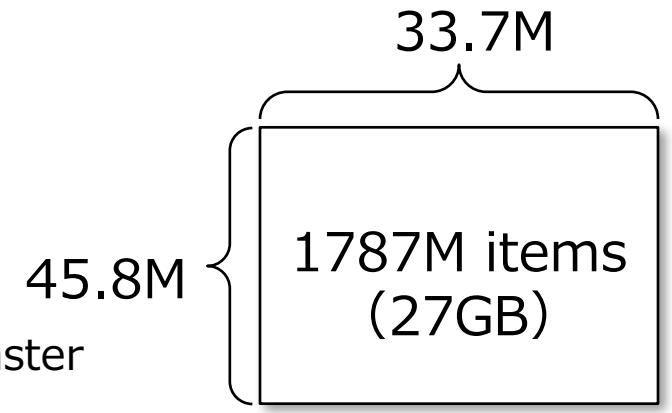




Logistic Regression

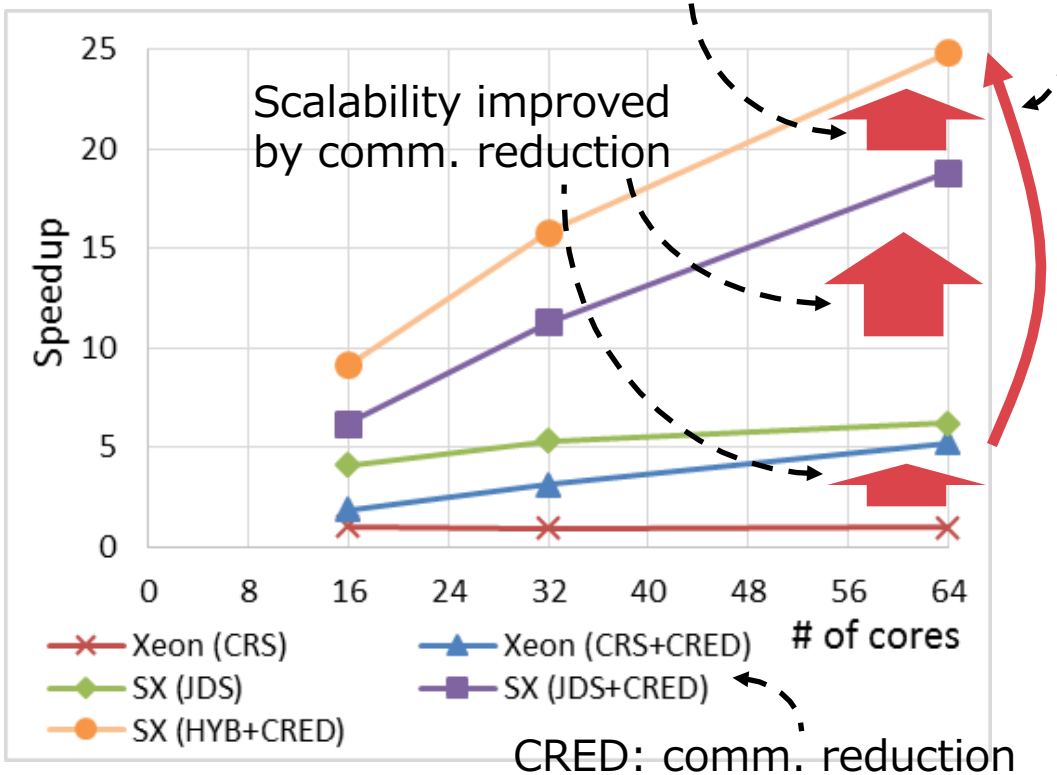
Evaluation setup

- Data: provided by Criteo (Web ad. opt.)
- Used “Gradient Descent” algorithm
  - Number of iteration = 100

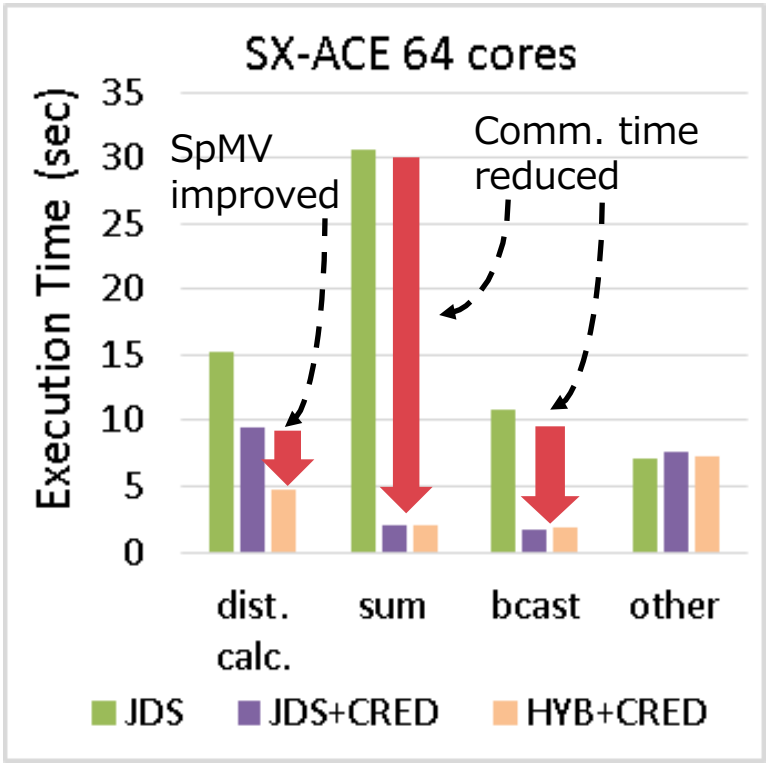


Perf. improved
by hybrid SpMV format

SX-ACE is faster
than Xeon



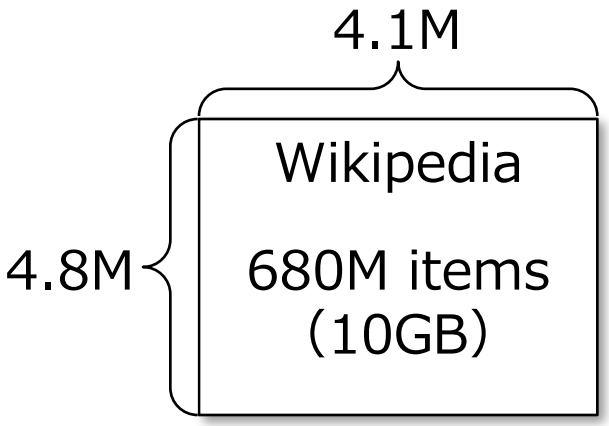
CRED: comm. reduction

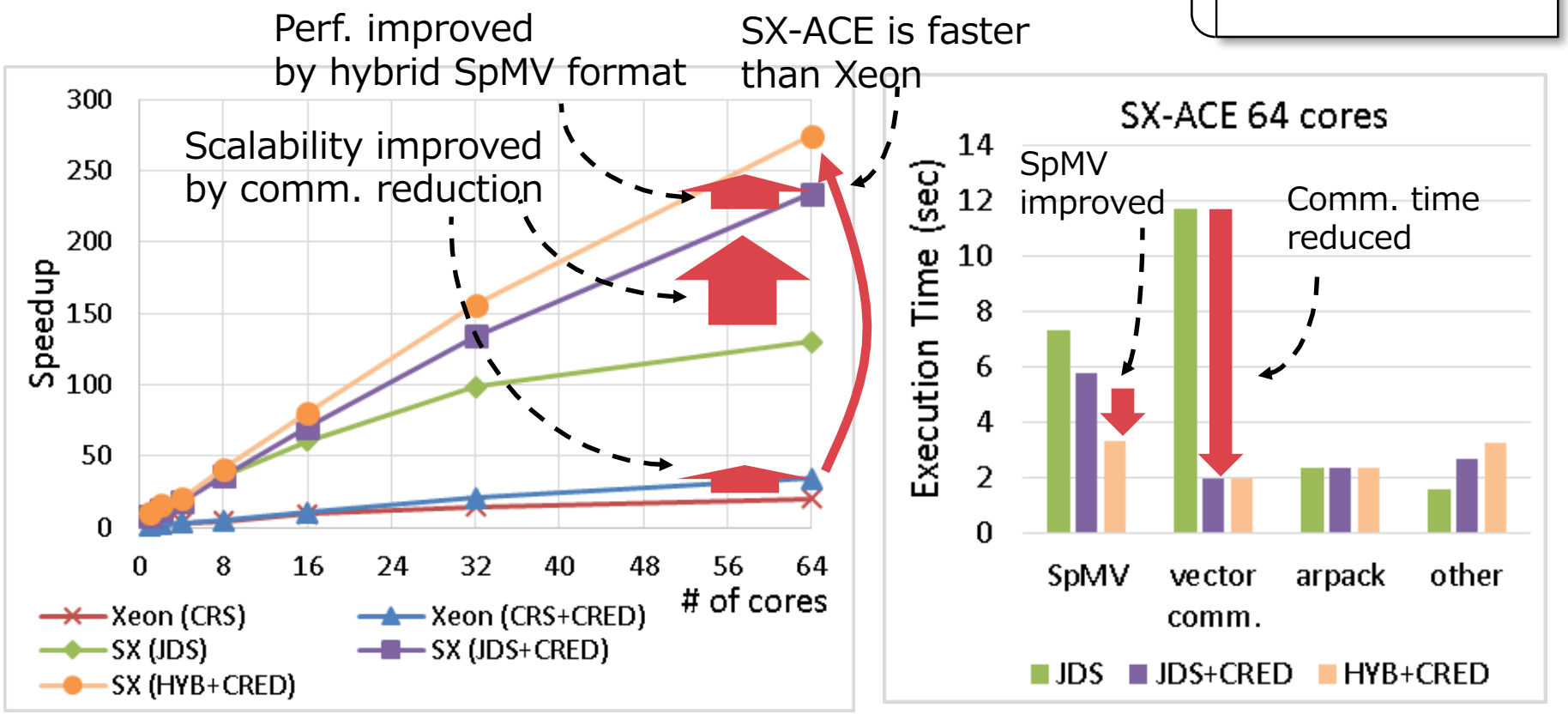


Singular Value Decomposition

Evaluation setup

- Data: English Wikipedia document
  - Corresponds to “Latent Semantic Analysis”
- Utilizes parallel ARPACK with our SpMV
- Top 100 singular values/vectors are calculated





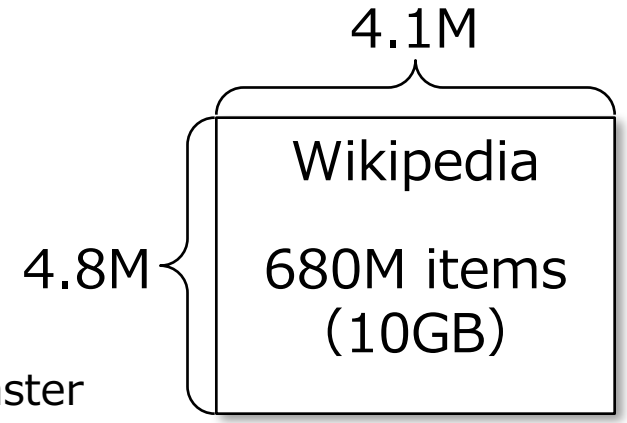
K-means

Evaluation setup

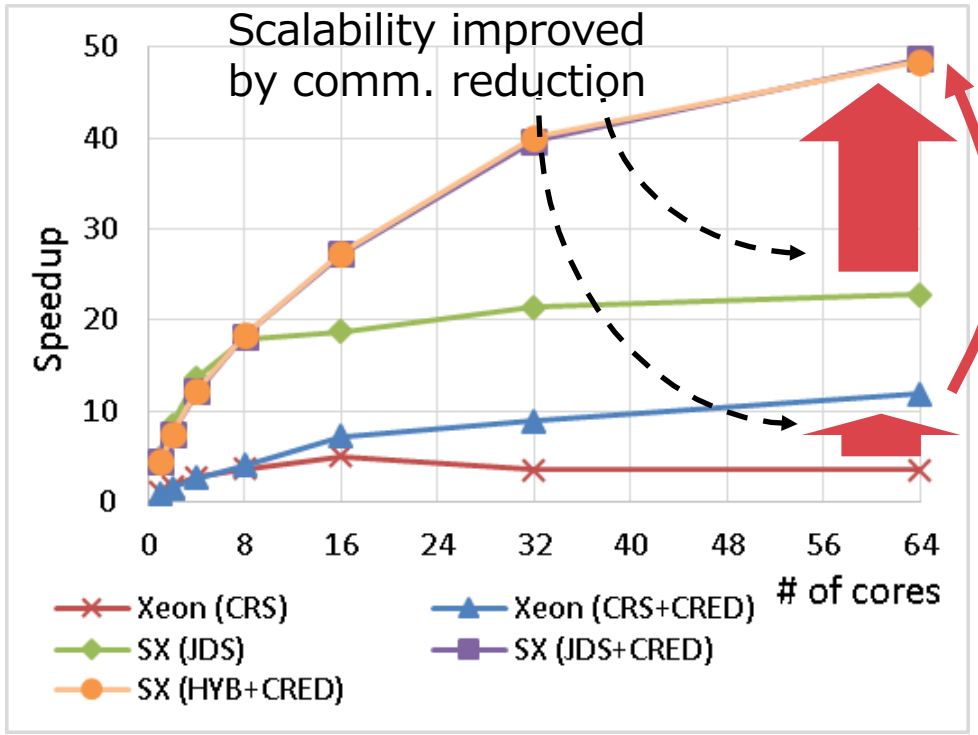
- Data: English Wikipedia document
- Number of clusters: 30
  - Number of iteration = 50

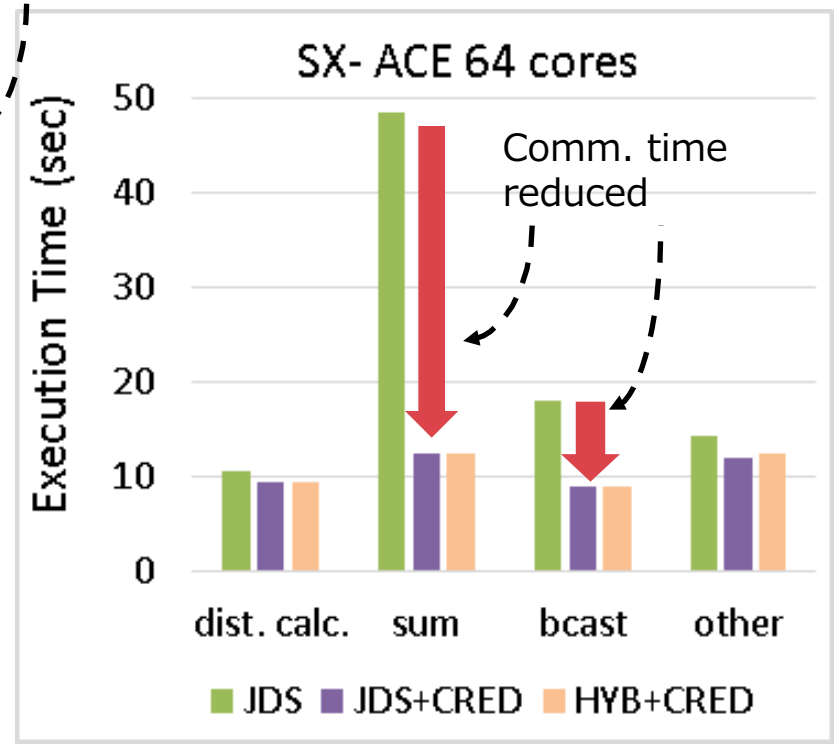
No speedup by SpMV

- Because it is less “power law”



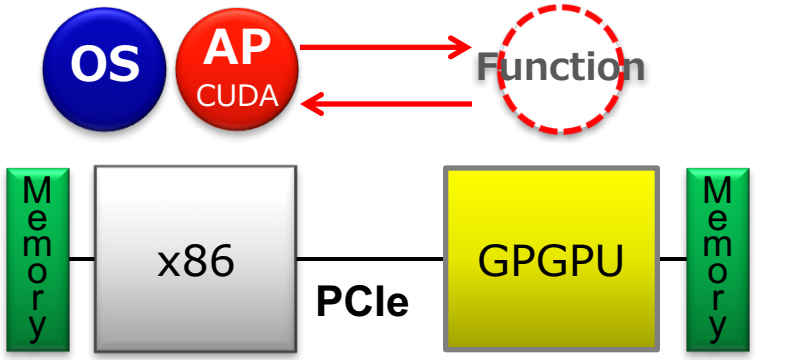
SX-ACE is faster than Xeon

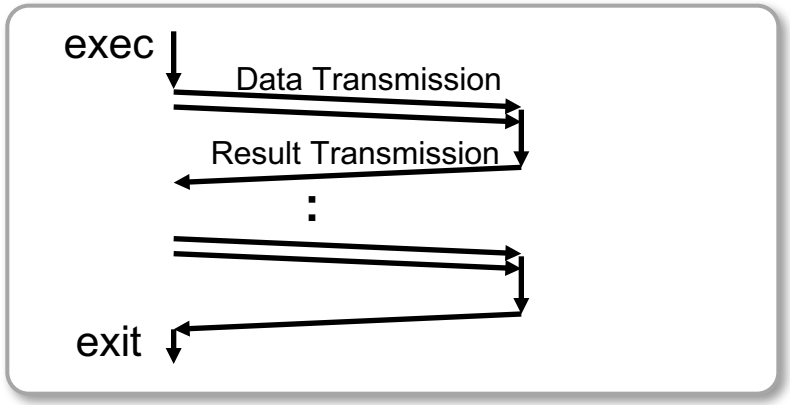




GPGPU and VE

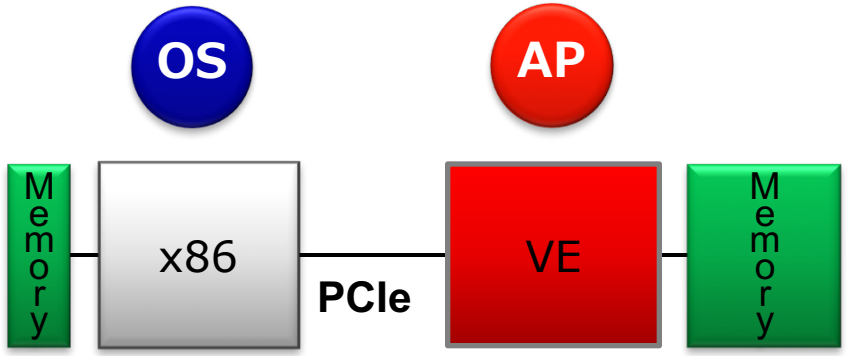
GPGPU Architecture

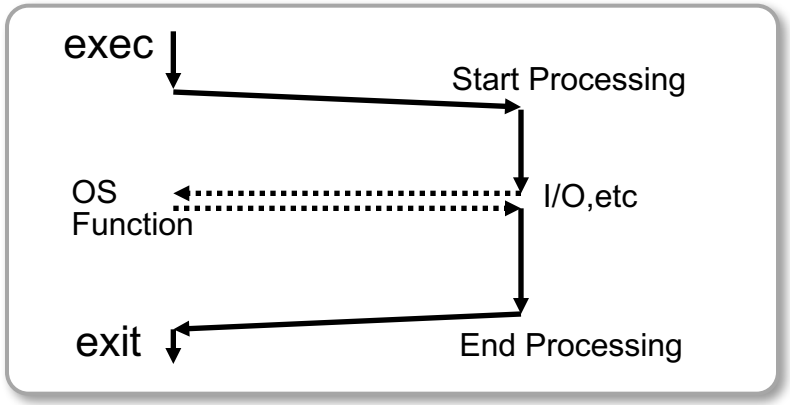




Frequent PCIe transmission

Aurora Architecture





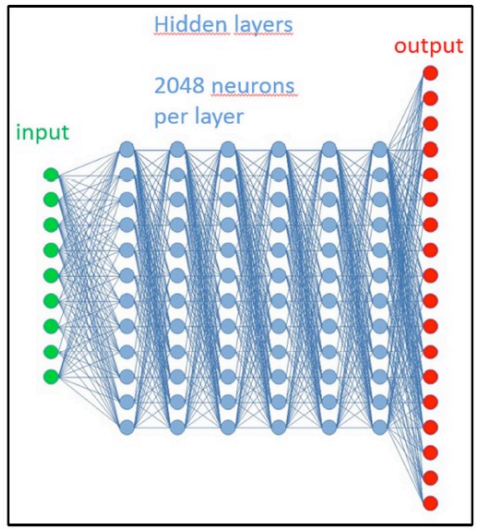
Whole AP is executed on VE

disadvantage
- PCIe bottleneck
- Small memory
- Programming difficulty

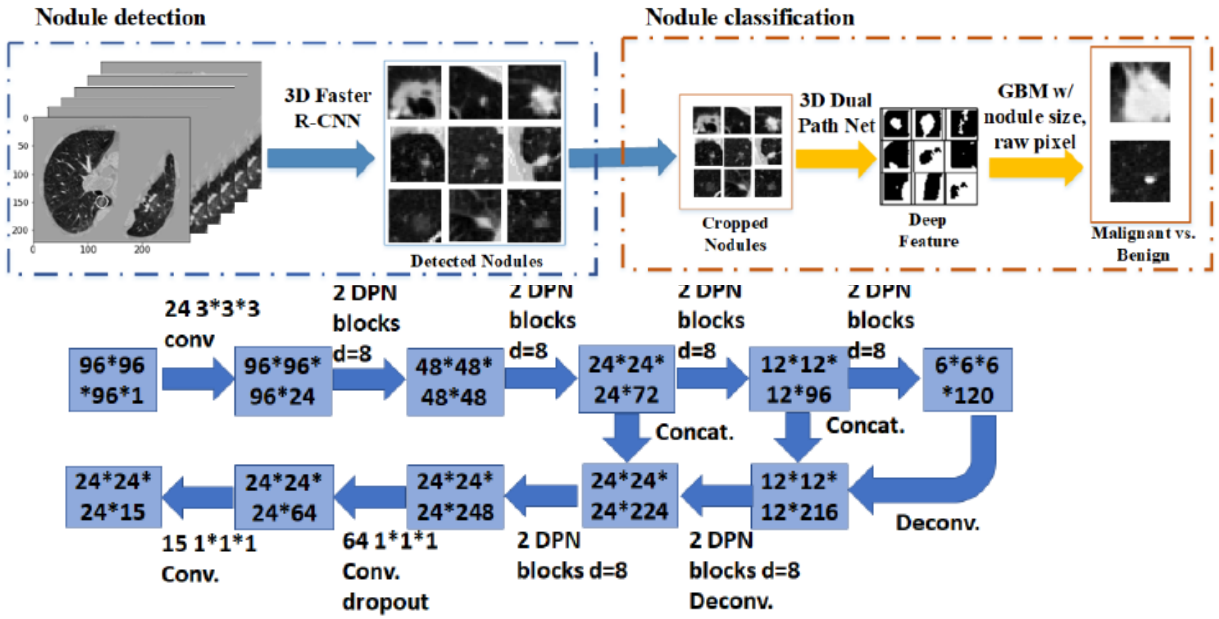
Advantage
- Avoiding PCIe bottleneck
- Larger memory
- Standard language

GPGPU vs CPU/Vector Processor

Multi Layer FC layers
for speech recognition



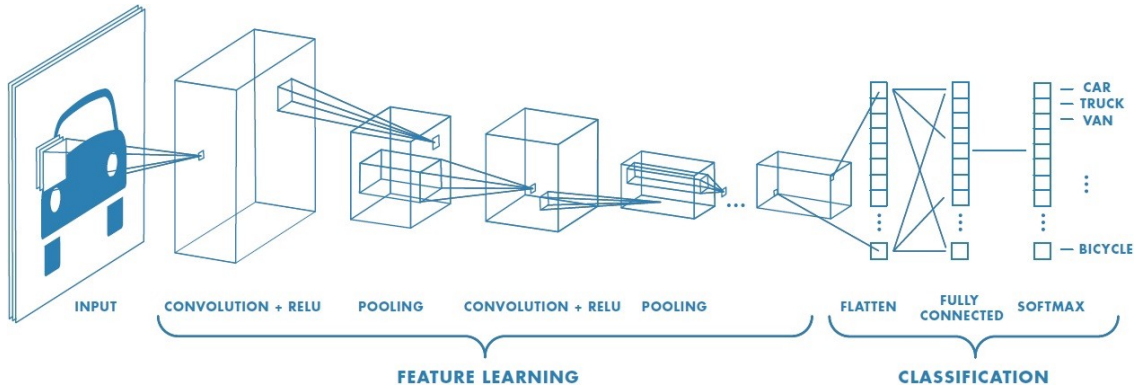
3D Convolutions 64X64X64=256Kbyte



3D CNN and multilayer FC => CPU/Vector

2D Convolutions 3X3X256=2K byte

50-layer	101-layer	152-layer
7×7, 64, stride 2		
3×3 max pool, stride 2		
[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
erage pool, 1000-d fc, softmax		



2D CNN => GPU

It was clear even 26 years ago – Sneakers (1992)

How data privacy is affecting business and IT

Data is unlike money – hard to keep track who has it

Cost of data protection >> Cost of storing & processing

Hot Topics - cloud, analytics, profiling.

Industries –
Fintech, Health, Digital Marketing, Advertising

KPIs for protected data use –
implementation, future protection, validation

NEC’s proposition

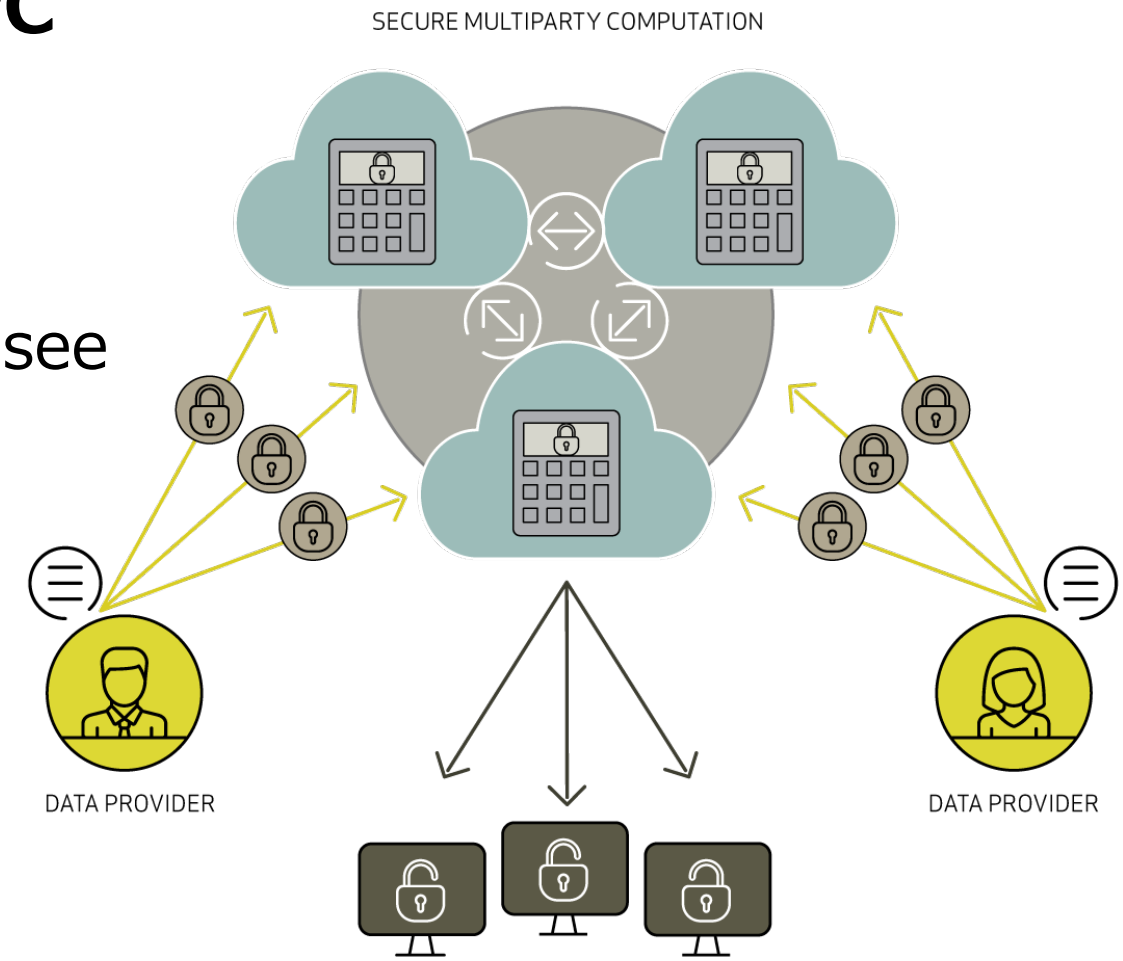
We’re looking for implementation partners:

secure MultiParty Computation-sMPC

Enables running algorithms on data we can’t see

Uses one time codes - not encryption

S/W only and open source



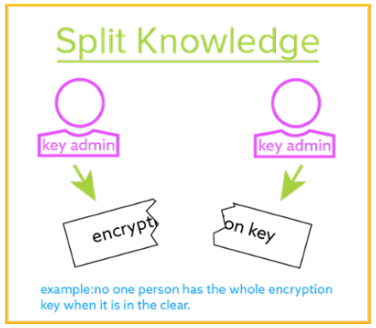
How can sMPC help Financial Institutions?

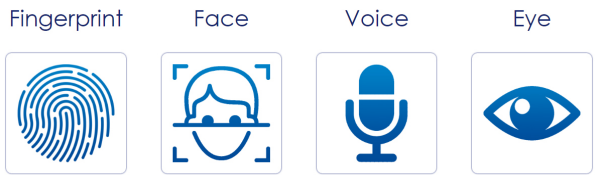
Marketing – better profiling of users without invading their privacy

Anti-Fraud – examine more records, inspect remote machines.

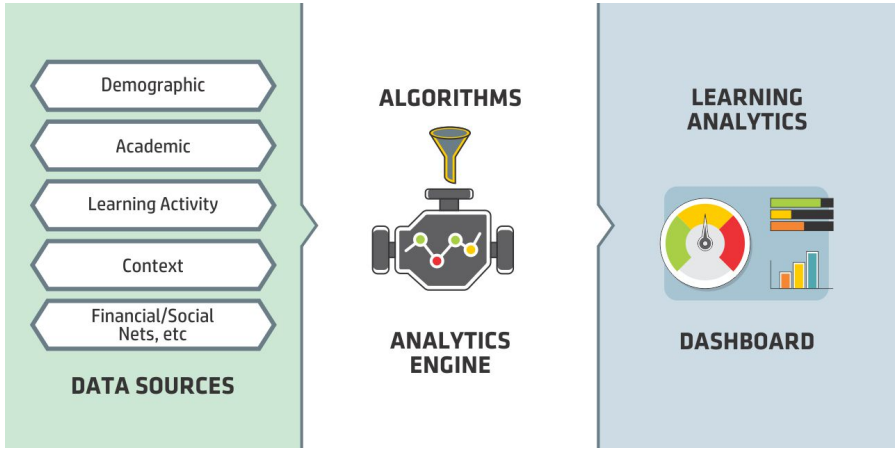
Public Cloud & Mobile devices – keep and process data off-premise with proven safety.

Off Premise





Marketing

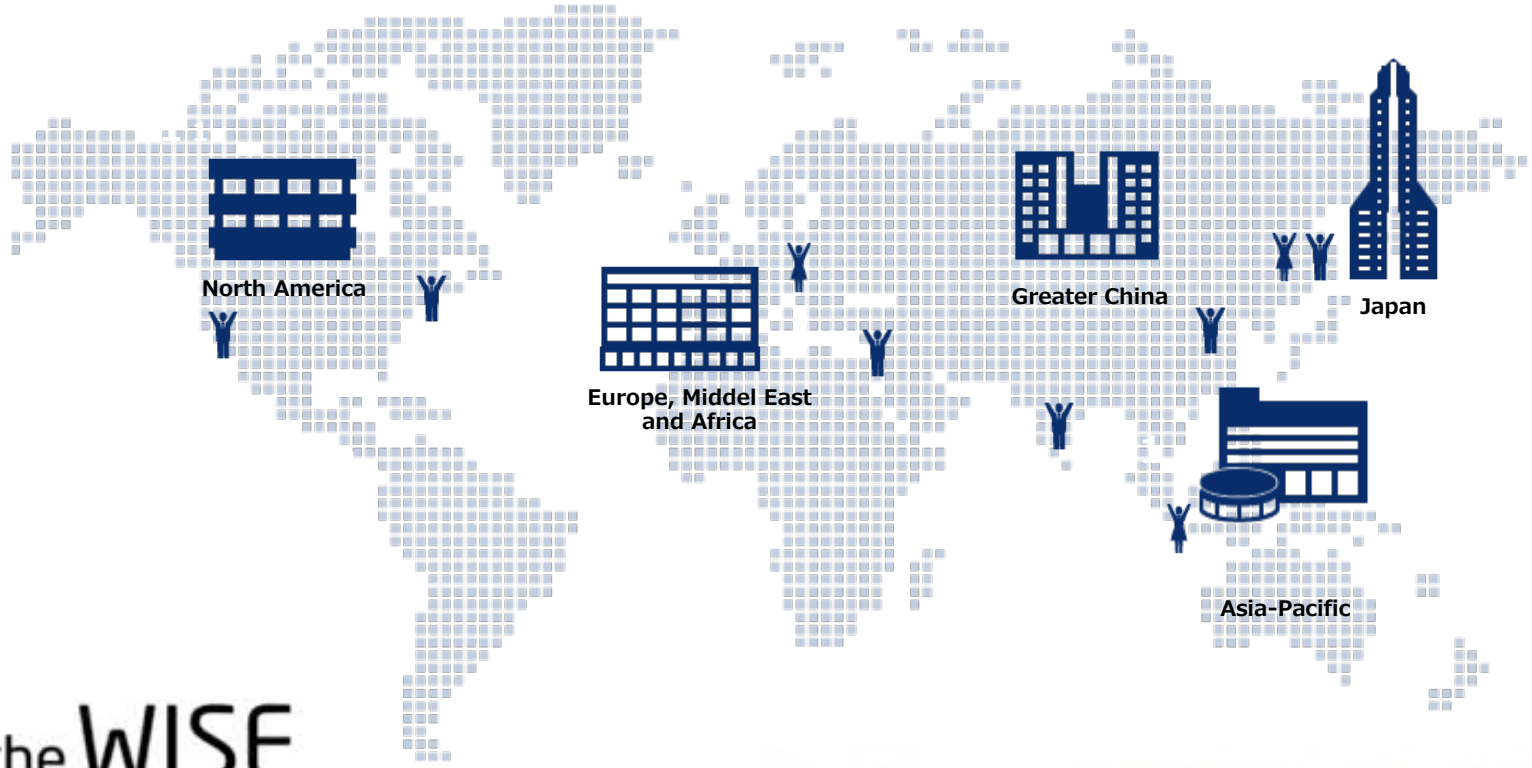


Anti Fraud

NEC Research

- Conducting Business In Over 160 Countries
- Network Of 9 Global Research Labs
- ~0.5% of Revenue Allocated to Research: ¼ Billion$

Established 1899
100K Employees

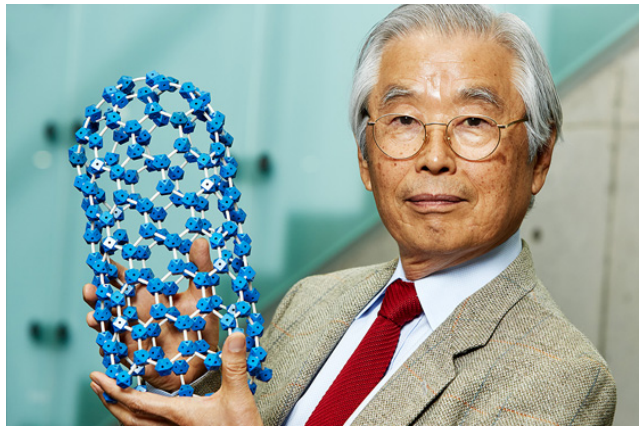


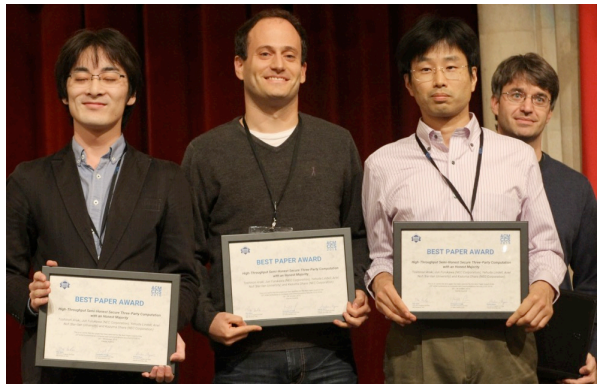
NEC the WISE

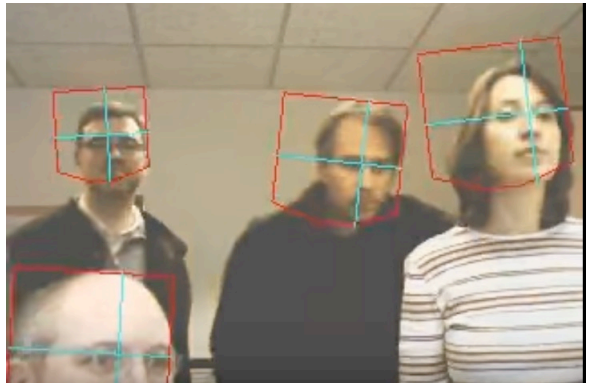
Tokyo 2020 Gold Partners

NEC research firsts..

- **Yann LeCun** – face detection using CNN (2003..), NEC Labs Princeton
- **Geoff Jiang** (ML for Cyber Security) – now VP AI at Ant Financial
- **Sumio Iijima** – Carbon Nanotubes (1991)
- **Furukawa, Chandraker -** best of ACM CCS 2016, CVPR 2014
- 4 Consecutive times #1 in **NIST Face Recognition**

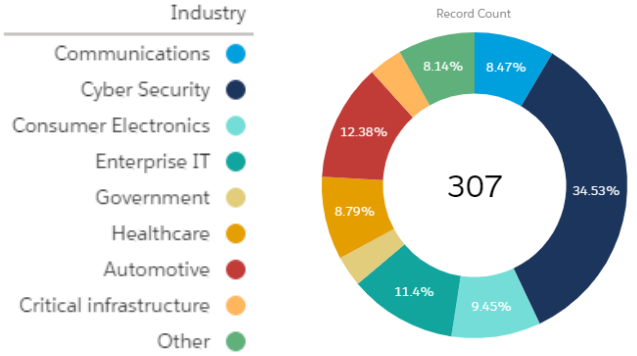






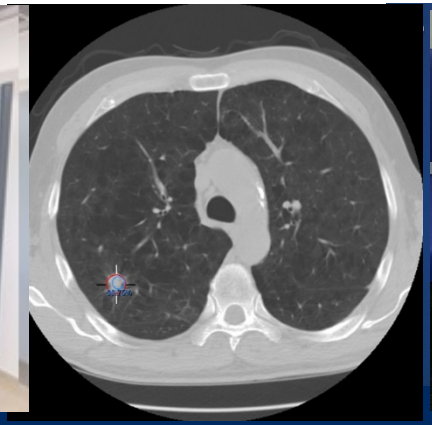
NEC's Israeli Research Center

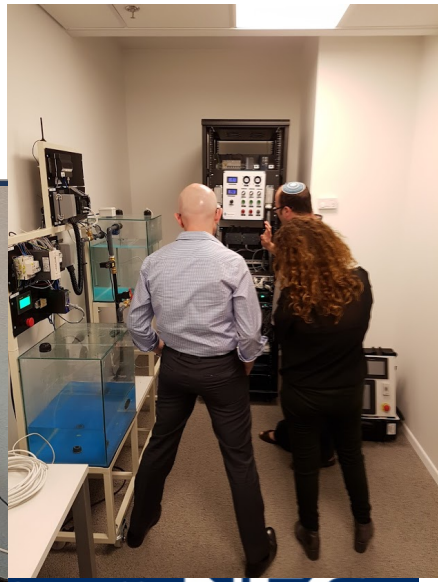
- **25 H/C in Cyber, Algorithms/DL, Outreach**
- **5 PhD, 7 MSc, 3 in Cryptography.**
- **Privacy Preserving Tech** – with MIT, large Financial
- Collaborations with **MIT, BIU, BGU, TAU.**
- **Cyber defense** for Critical Facilities and IT.
- Edge **Deep Learning** systems (surveillance/analytics)
- Deep Learning **Medical Diagnostics**: Assuta, Meuhedet.
- Total investment of NEC in Israel ~10M$.
- Involved with 2 Accelerators – DRIVE and Alpha-C



DRIVE
FUTURE MOBILITY NOW





NEC's Cyber Digital Shadow in Israel

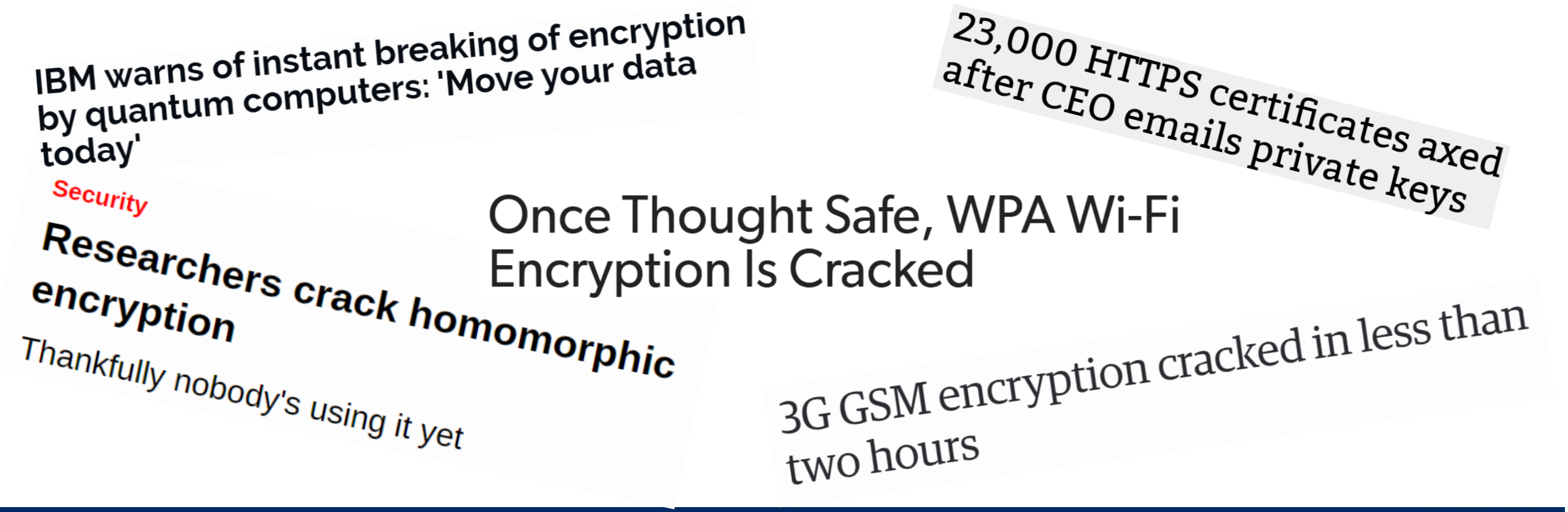
What does sMPC give you?

Immunity from Encryption breaking – it uses one time pads, not Public Key

Confidentiality – of both data and algorithm

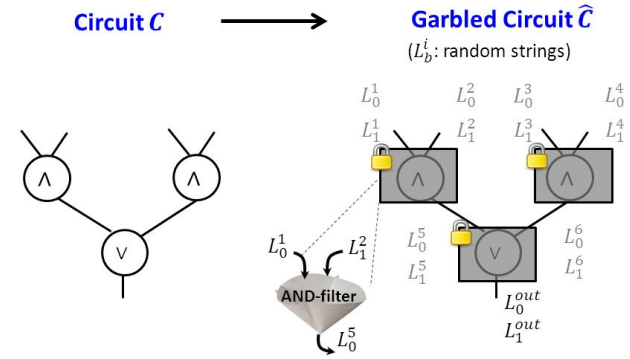
Attestation – you can prove precisely which algorithm was executed and on what data

Traceability – data cannot be accessed without all involved allowing it in real time



Why only now?

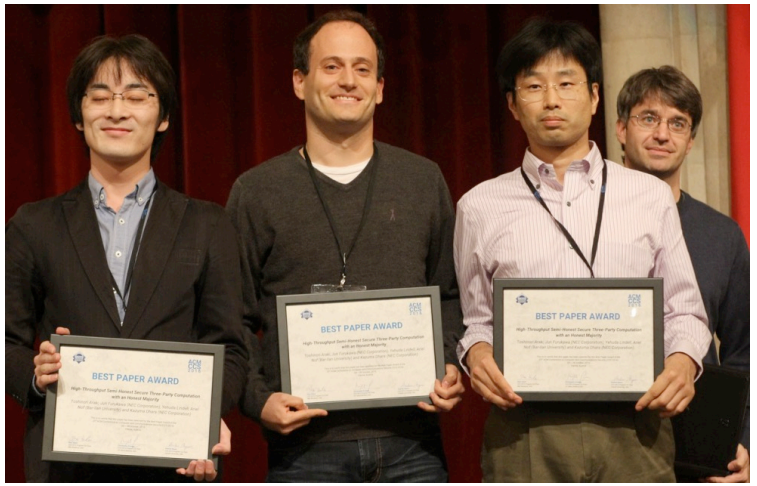
Yao's Garbled Circuits
[A. Yao 1986]



Originally proposed in the **1980's**
Was too slow – requires large compute and very fast networks
Encryption was considered "good enough"

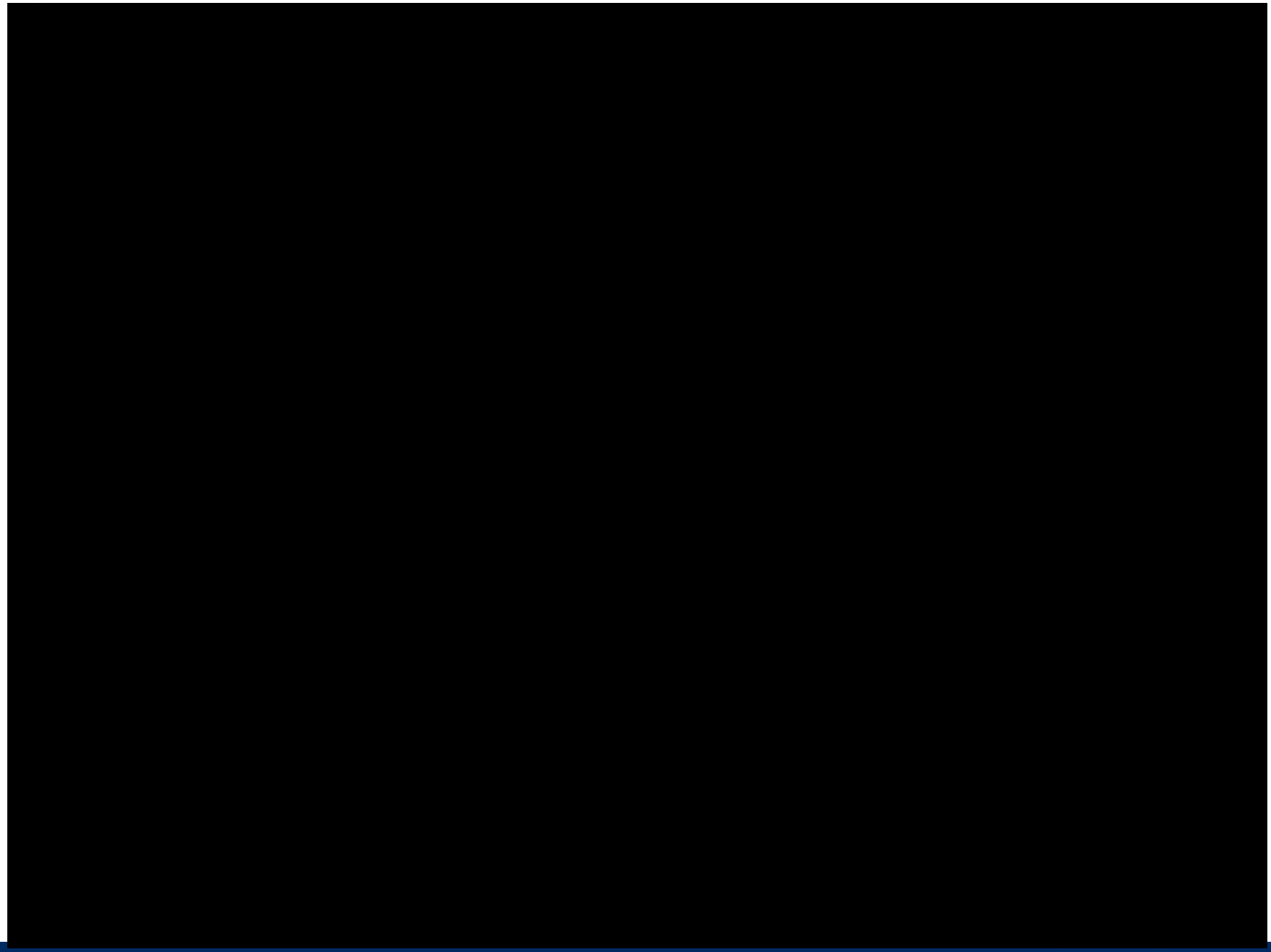
NEC and others worked on protocol acceleration:
Today in a **Public Cloud** you can run **complex AI in ~100 milliseconds** on instances.

```
34 for i in range(level1):
35   result1[i] = sint(0)
36   tmp = sint(0)
37   for j in range(num_input):
38     tmp = weight1[i][j]*input[j]
39     result1[i] = result1[i] + tmp
40
41 for i in range(level1):
42   result1[i] = relu(result1[i])
```



ACM CCS 2016 best paper award

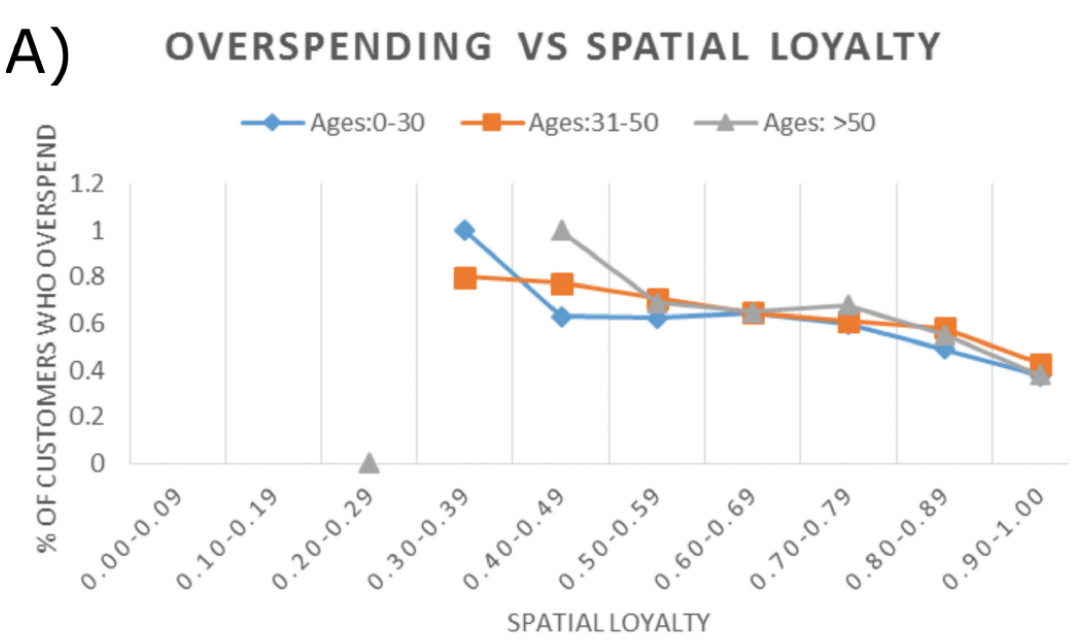
How it works



Credit Risk Prediction

- MIT Published “MoneyWalks” in 2015
- Predictors for Overspending, Financial Trouble, Late Payments
- Based on data with transactions, locations, personal data.
- Proved location data improves prediction significantly
- NEC implemented using sMPC to enable regulated, legal use

MIT Connection Science
enhancing cultural intelligence through technology | Trust :: Data



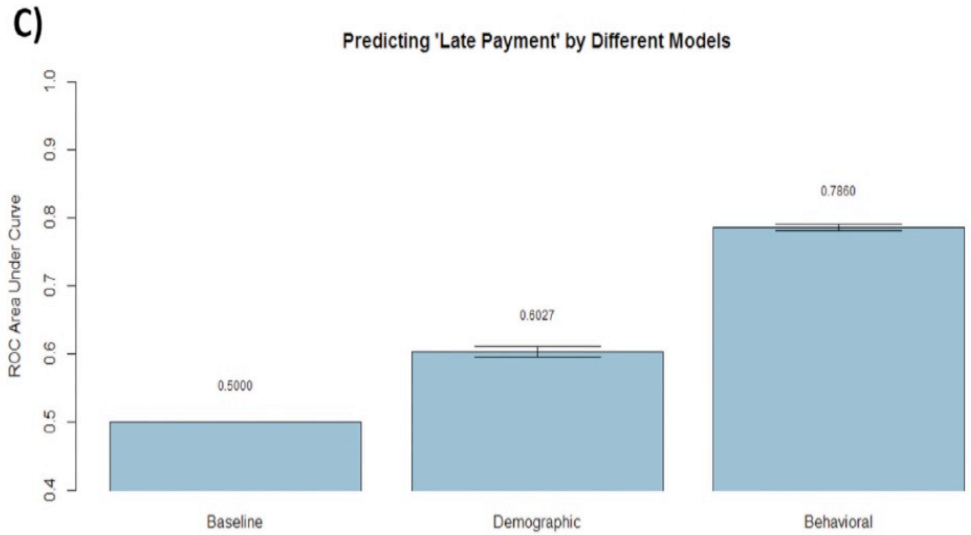
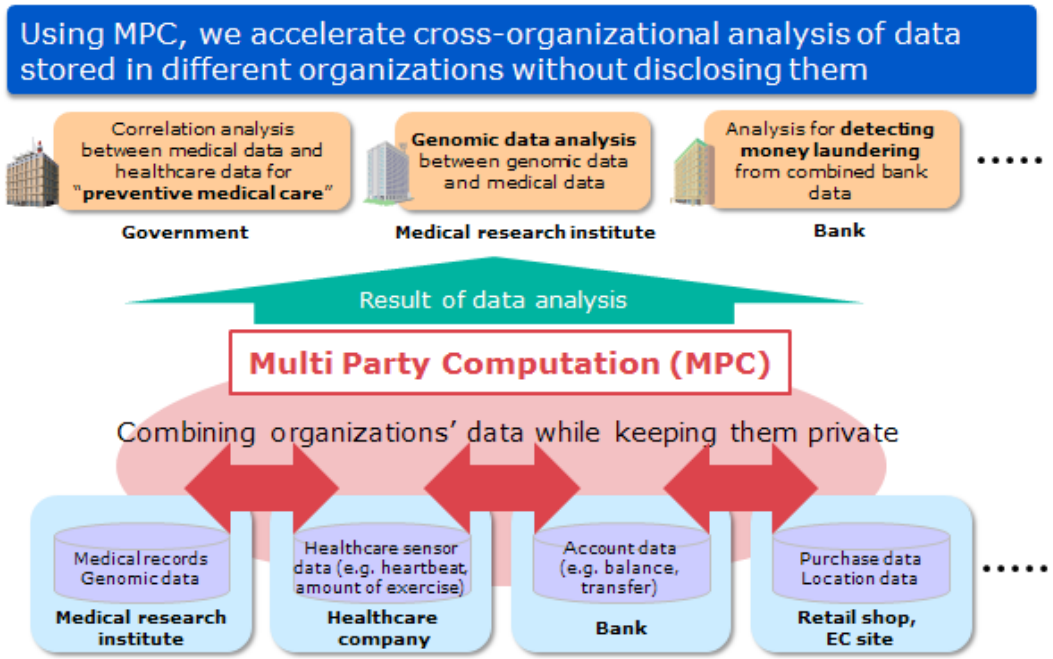


Fig 4. Prediction performance for different financial outcomes using a baseline, demography-based, and behavior-based model. The behavioral models perform 31%, 49%, and 30% better than the corresponding demography models for predicting “financial trouble", "overspending", and "late payment", respectively.

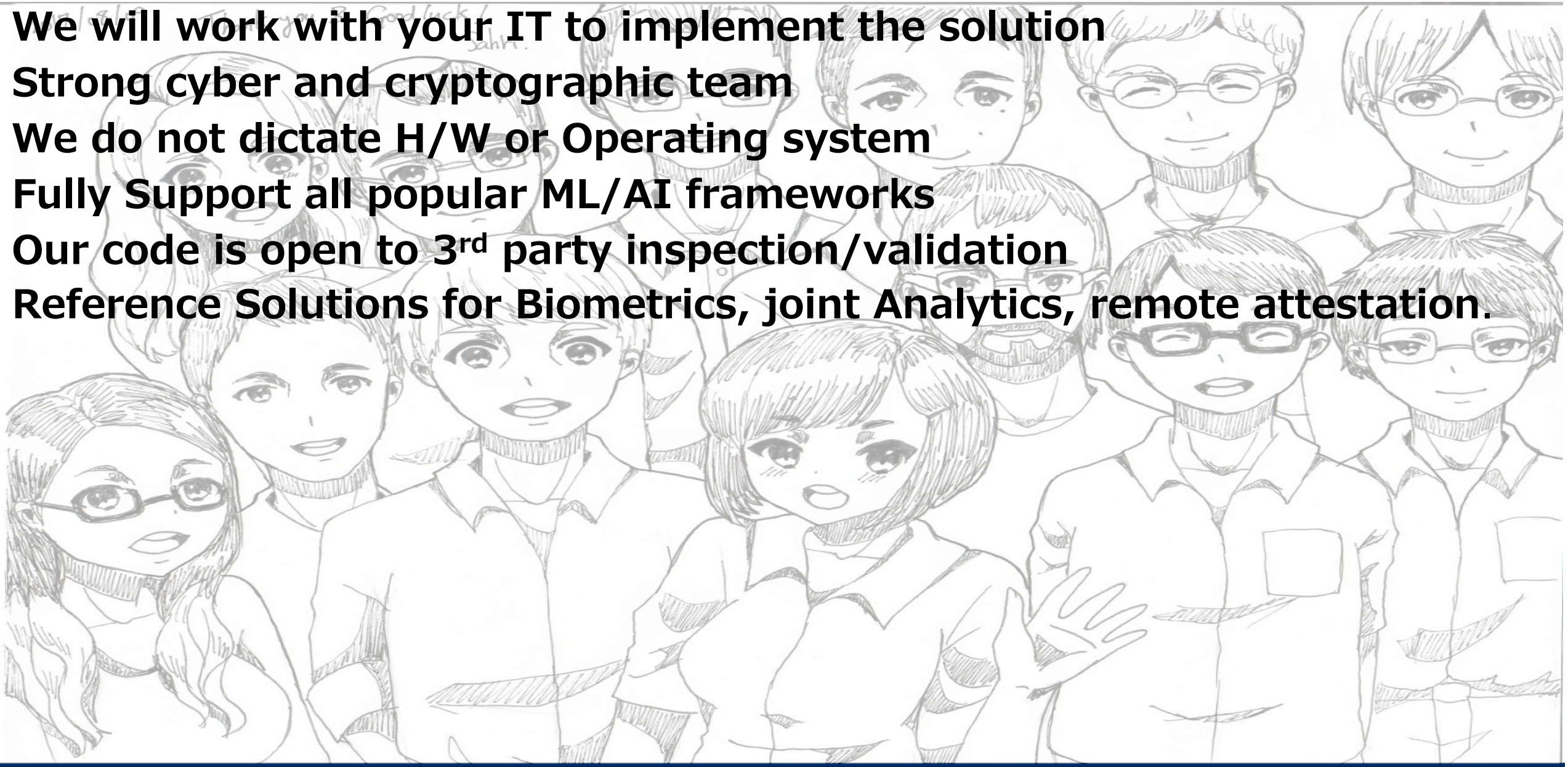
Other Applications of sMPC

- **PKI/cloud:** store private keys split between entities.
- **Biometric authentication** with Bio Template split.
- **Verify user endpoint**/partner API machine is valid
- **Run analytics** on user data with in-built regulation compliance
- Partners can **join information** for better analytics
- Safe **remote monitoring** of your API GW/SIEM logs



What NEC is offering

We will work with your IT to implement the solution
Strong cyber and cryptographic team
We do not dictate H/W or Operating system
Fully Support all popular ML/AI frameworks
Our code is open to 3rd party inspection/validation
Reference Solutions for Biometrics, joint Analytics, remote attestation.



2017/08/09

Thank you & Good luck!

Sahri.