In-Memory Computing SUMMIT NORTH AMERICA 2019

"Computational Memory Acceleration" Tensor Flow + Hyperdimensional Computing

Gil Russell | Alan Niebel

WebFeet Research Inc.

WebFeet Research

Research | Vision | Knowledge

The In-Memory Idea begins here

At the 2000 Intel Developer Forum in Palm Springs, California a relatively unknown entrepreneur, while having a Keynote fireside chat with Andy Grove, said he'd like to take the entire Internet and put it in memory to speed it up – "The Web, a good part of the Web, is a few terabits. So it's not unreasonable," he said. "We'd like to have the whole Web in memory, in random access memory."

Early Discounts can be Costly…,

The foregoing comment received a rather derisive reception from the audience and was quickly forgotten.

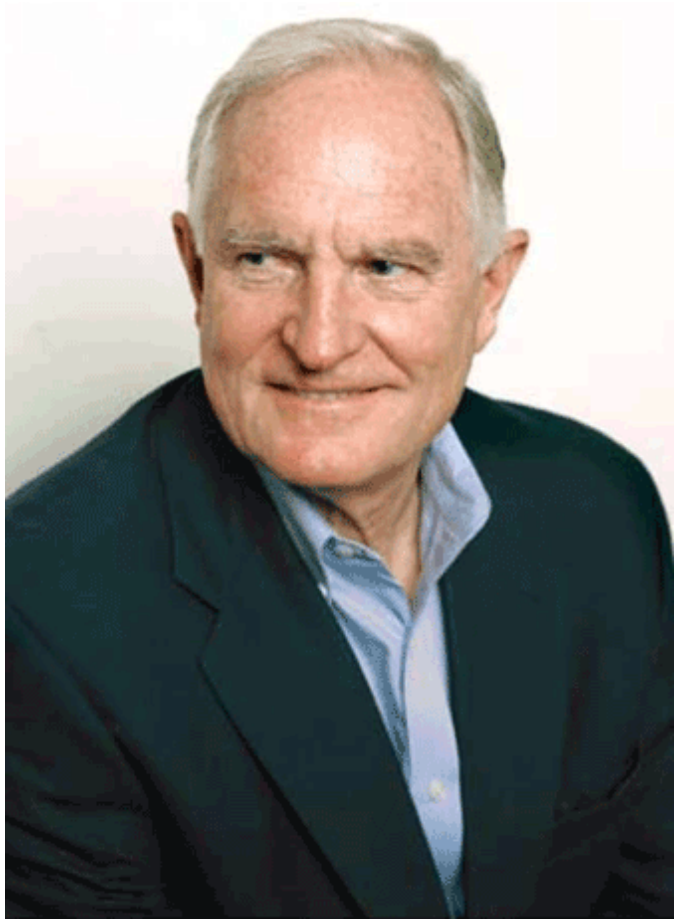
*The speaker, **Larry Page**, an unknown at the time, as was his startup company, Google – the company's backbone consisted of 2,400 computers at the time.*

"Those Damn Disk Drives"…,

*In **2006** Hasso Plattner, Co-founder of SAP AG, took a bottle of red wine, a wine glass, some writing implements and paper to the garden behind his house. By the time he reached the bottom of the bottle there wasn't much written on the paper.*

But he had reached the conclusion that in-memory systems were the future.

Hasso's New Architecture – HANA1



- *Plattner realized that for SAP to remain competitive it needed to innovate - Plattner believed that by changing the server design to accommodate massively parallel processing with enough memory to load an entire database when combined with columnar based storage software would have a revolutionizing effect on processing speeds for OLTP and OLAP applications*
- *Gathering a small group of PhDs and undergrads at the Hasso Plattner Institute, Plattner expressed the in-memory idea he wanted them to explore. The first prototype was shown in **2007** before an internal audience at the company's headquarters in Waldorf, Germany. SAP management was skeptical that the idea would work – the team needed to prove that the concept of in-memory database would work under real world conditions*
- *Using contacts to advance the project, Plattner persuaded Colgate-Palmolive Co. to provide transaction data for the project (vector data). He also persuaded **Intel's Craig Barrett** to secure the latest microprocessors for the labs ongoing effort. The company also set up an R&D facility in Palo Alto to be in close proximity to their innovation and research partner Stanford University*

Hasso's New Architecture | HANA1 → HANA2

- *SAP's In-Memory Database system, named HANA, was officially announced in **May 2010** with shipments commencing with the release of SAP HANA 1.0 in November. The market was slow in adopting the technology convinced that it was still in an early stage of development. Analytics and the need to score a real reason for their customers to mount their IT to the cloud provided the market conditions SAP's HANA needed to press its acceptance. SAP over time adapted HANA to the Cloud through successful partnering with a wide array of vendors making it the company's fastest growing segment*
- *HANA 2 launched in November 2016.*
- *SAP's SAPPHIRE NOW and ASUG Annual Conference **May 12 – 14, 2020** Orlando, Florida, North America - the priority is to build on existing enterprise IT investments by integrating **new and emerging technologies** such as IoT, analytics, AI and automation, blockchain and cloud.*

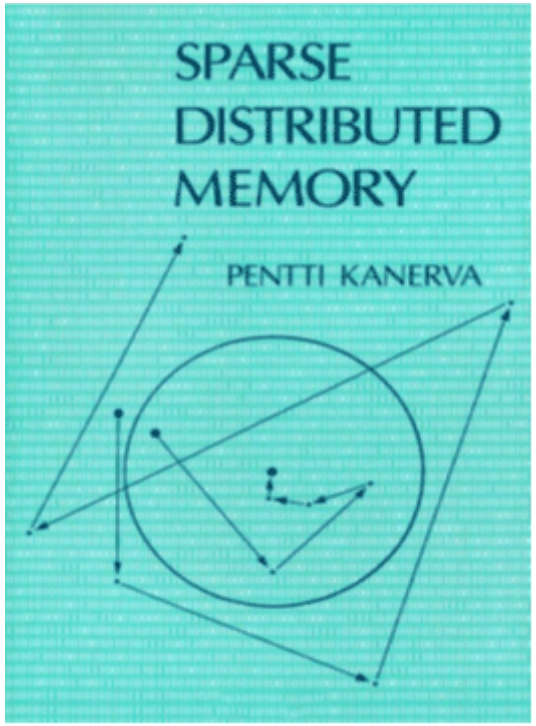
Hyper-Dimensional Computing (HDC)

Google: Visualizing High Dimensional Space

View Google's "Visualizing High Dimensional Space" video here

HDC Origin | Sparse Distributed Memory (SDM) + Hyperdimensional Computing

Pentti Kanerva, UC Berkeley
Redwood Center for Theoretical Neuroscience
"Sparse, Distributed Memory", MIT Press, 1988

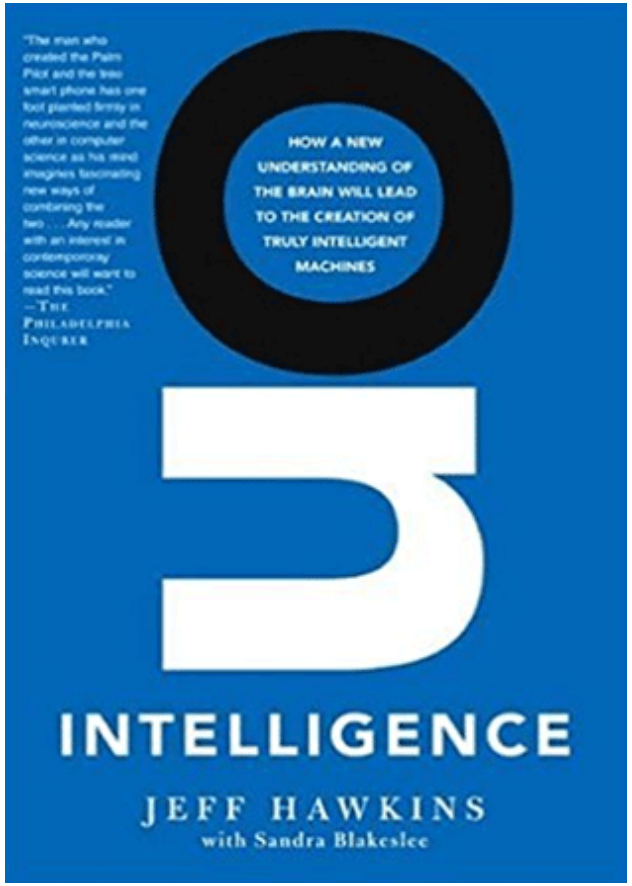


"Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors"

Pentti Kanerva,
January 2009

https://redwood.berkeley.edu/

HDC Origin | Hierarchical Temporal Memory (HTM)
+ Thousand Brains Theory of Intelligence

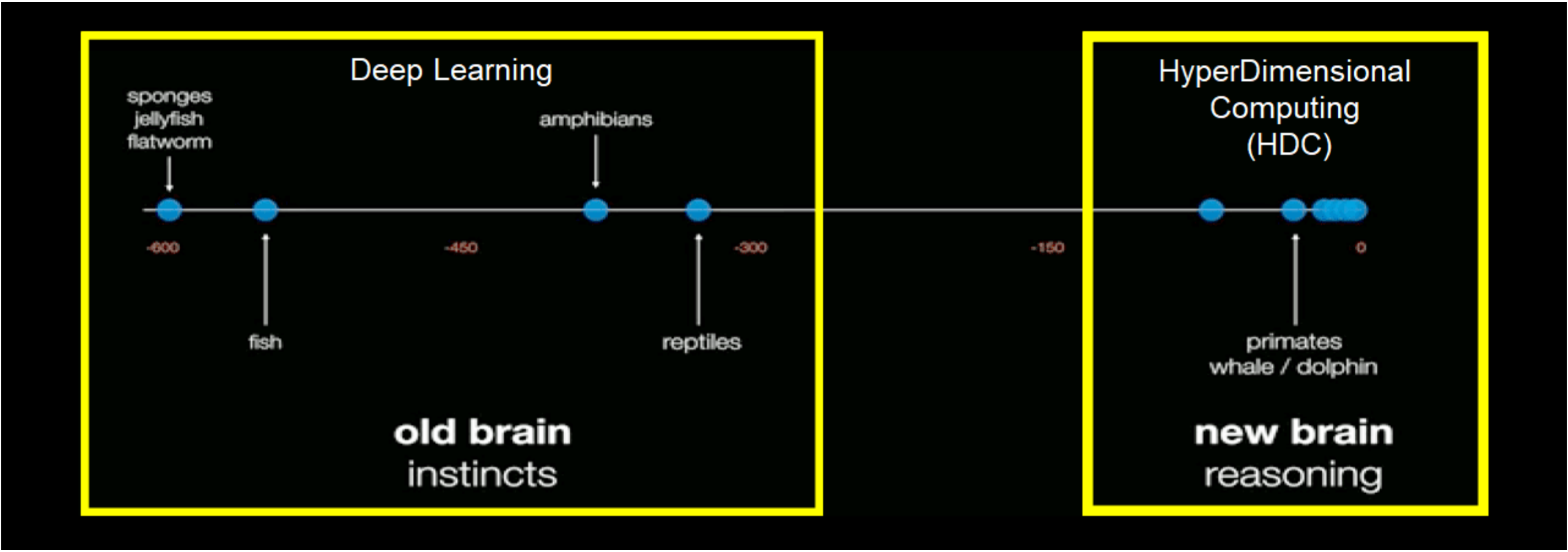


Jeff Hawkins
Founder: Redwood Center for Theoretical Neuroscience (2002),
Founder & CEO: Numenta (2005)
"On Intelligence, How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines", 2004

"Thousand Brains Theory of Intelligence", 2019

https://numenta.com/

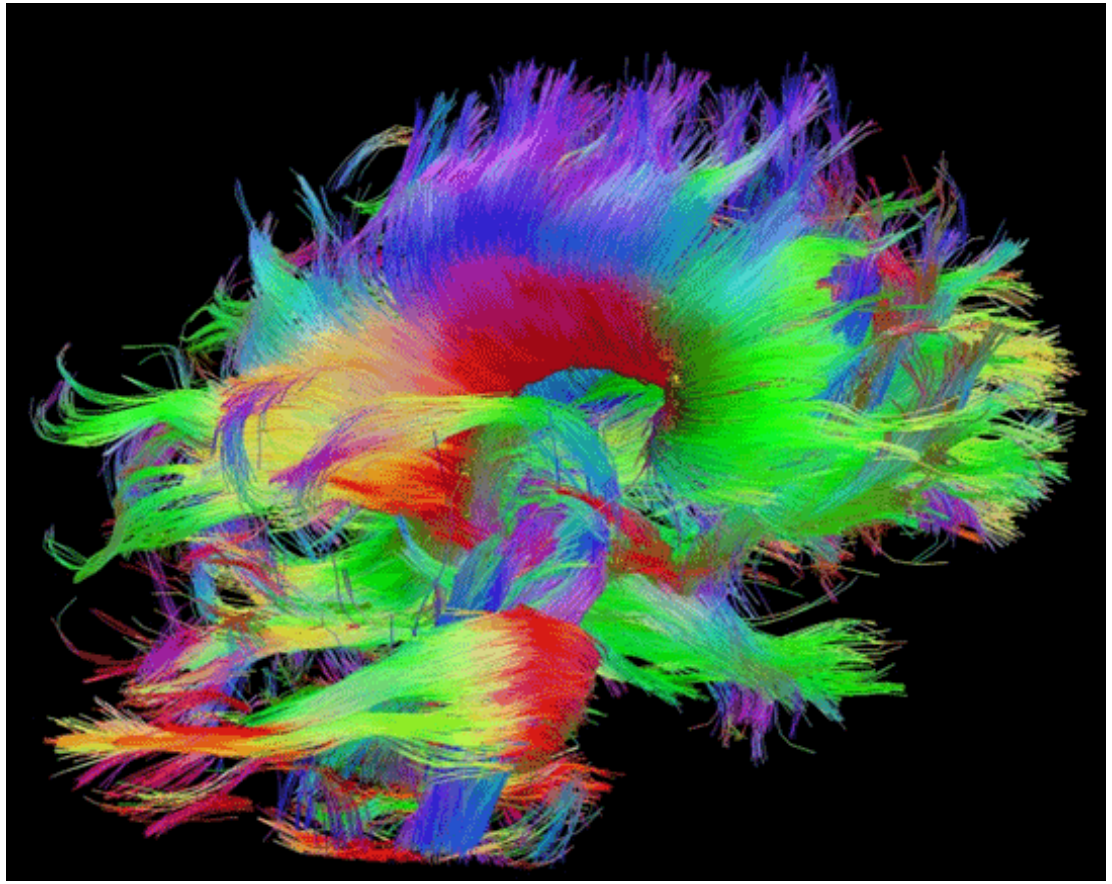
Separation of Ideologies | New vs. Old Brain



"Brain Inspired"

The Human Connectome Project

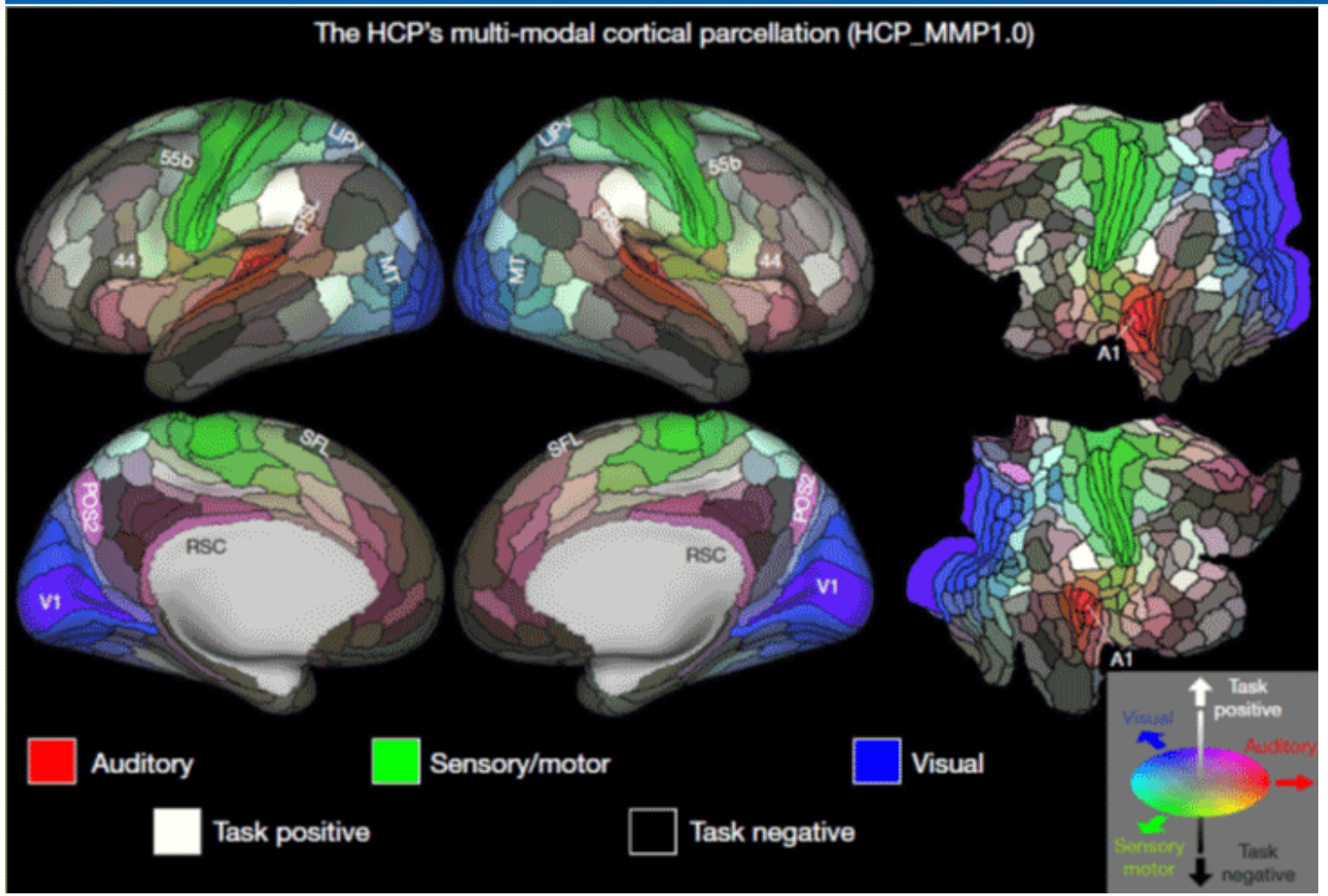
White matter fiber architecture of the brain. Measured from diffusion spectrum imaging (DSI). Shown are the corpus calllosum and brainstem pathways. The fibers are color-coded by direction: red = left-right, green = anterior-posterior, blue = ascending-descending (RGB=XYZ). www.humanconnectomeproject.org



- *The Connectome Project yielded the Axon connectivity of the brain but failed to give any information on the sensory perception "traffic" being transmitted.*
- *However, connections between sections of "Gray" matter provided insights into the "wiring" of the brain.*

Neocortex Parcellation

Human Connectome Project



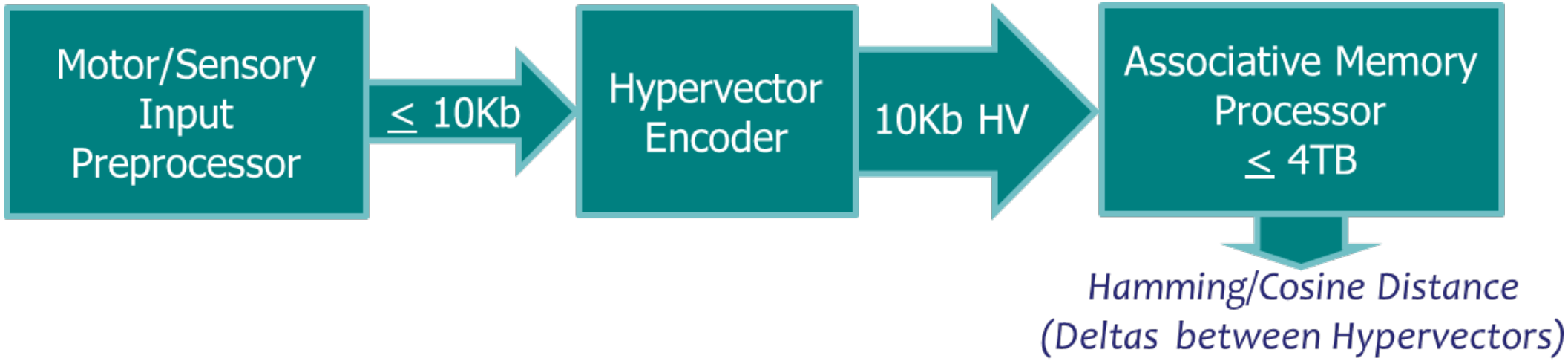
- *Follows Jeff Hawkins "Thousand Brains Theory of Intelligence", 2019*
- *Blends well with the findings of Neuronal "Cliques and Cavities"*

Hyper-Dimensional Computing (HDC) – What is it?

HDC is an advanced algebra memory based framework for building a general cognitive system with the associative qualities of the human brain's neo-cortex:

- *Robust, noise-tolerant and explainable*
- *Ability to "Predict" Sequences ⇒ Generalization*
- *Learns from data/example, learns by analogy*
- *Fast "One-shot" learning - required for anomalous conditions*
- *Integrates signals from disparate senses*
- *Allows simple algorithms that scale to large problems efficiently*
- *Allows high degree of parallelism*
- *Has been implemented on low power devices*

Hyperdimensional Computing (Block Diagram)



HV Address space ($2^{10,000}$ = 1.995063116880758384883742162 6 e+3010)

Associative Memory Processor (Training + Long Term Memory ~ 4 to 8 TB)

In-Memory Computing SUMMIT NORTH AMERICA 2018

High-Dimensional Computing

SDM Memory & Hyperscaling

A session on the Cloud's missing AI component

Gil Russell

WebFeet Research Inc.

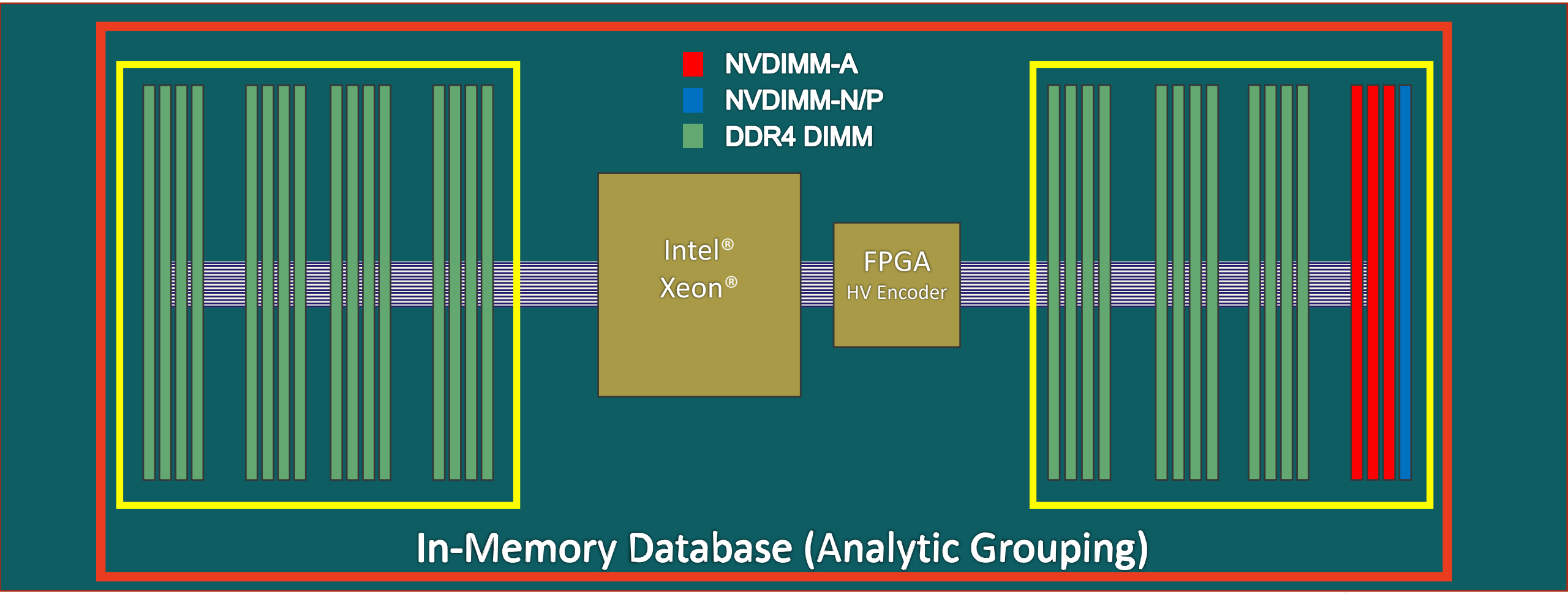
October 2, 2018

Immediate 820 Hits

WebFeet Research

Research | Vision | Knowledge

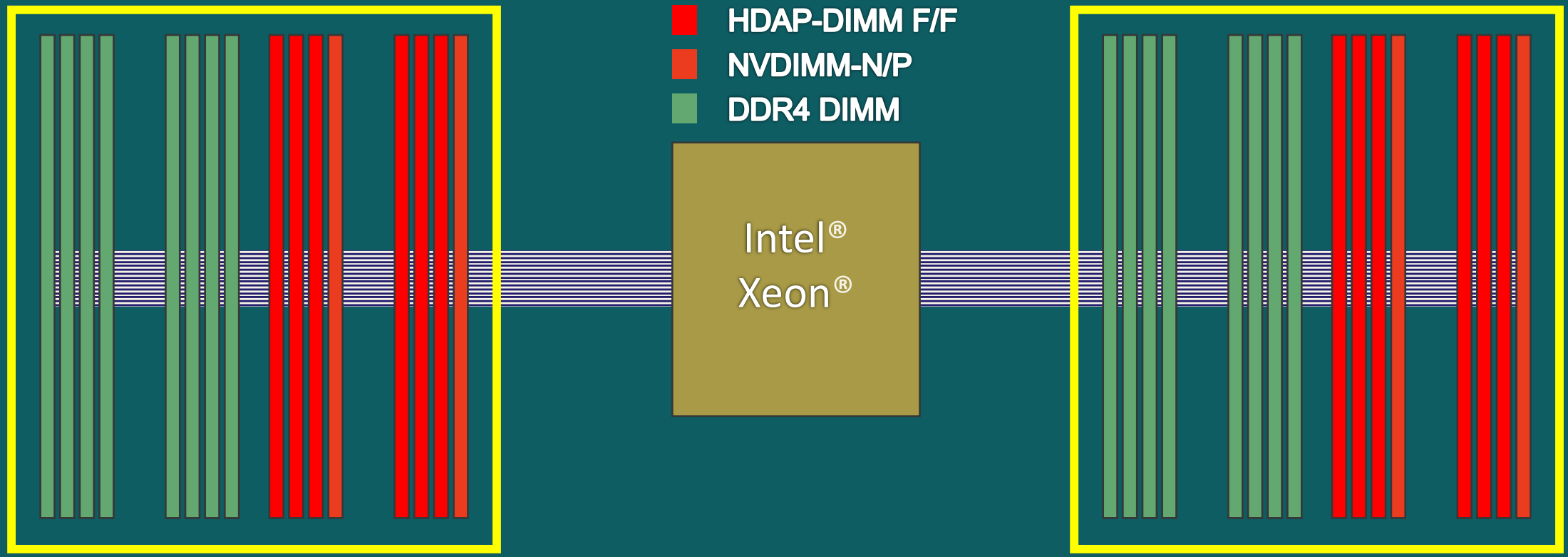
The Kanerva Partition (Database Intersect)
In-Memory Hypothetical System Implementation



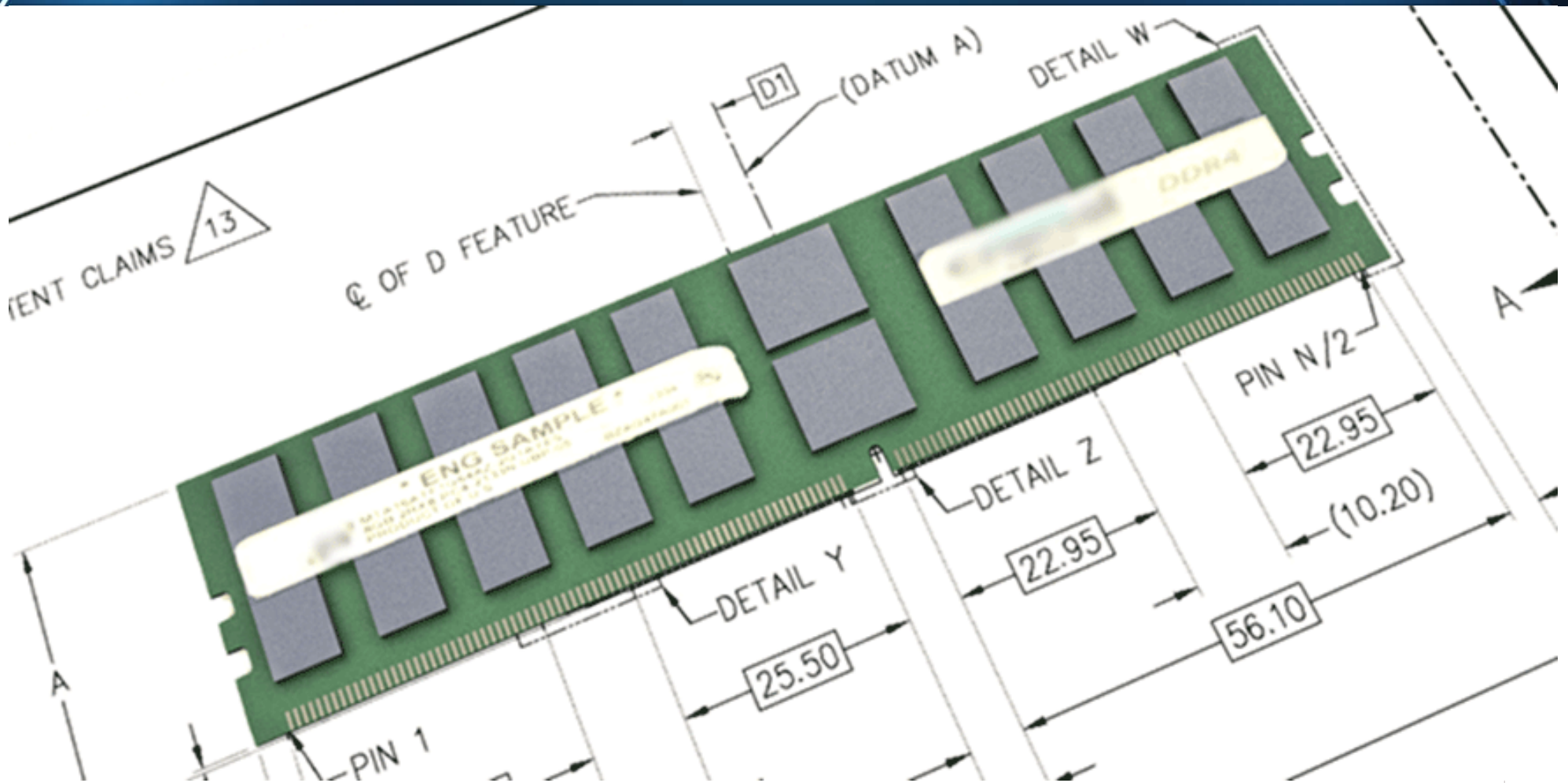
The Kanerva Partition

(RT-PA Grouping)

In-Memory Hyperdimensional Associative Processor (Hypothetical)



Hyperdimensional Computing Associative Processor - NVDIMM-X Form Factor



288 Pin NVDIMM Form Factor

Emerging Memory Complacency

- *Strong adherence to application usage patterns (store and retrieve, make it faster at lower power) have narrowly defined the scope and outlook as regards new and unique opportunities for this community*
- *Having few differentiating factors outside of the device memory element technology type, none of the emerging memory vendors have dared step outside of their comfort zone boundaries (except one)*
- *They await an entity much larger than themselves to take the first step across their limitations - this has now transpired…,*

Some Human Learning Trends (Thus far)

Machine Learning ≠ Machine Intelligence

- *Promise of "Deep Learning" to bridge over into Machine Intelligence has not worked*
- *Fails XAI (EU GDPR), Rigid, Narrow and cannot learn without repeating the full training regimen*
- *Entering the $trough\ of\ disillusionment$ as regards MI*

- *Sequences in time mandatory*
- *Prediction absolutely required*

} *Required for General Intelligence*

- *Fast "One-shot" learning required for anomalous conditions*
- *Broad span solution – one algebra top to bottom*
  - *Embedded Control through High-Performance Computing*
  - *Cognitive Database compatibilities*

Ensemble (formerly "The Kanerva Partition")

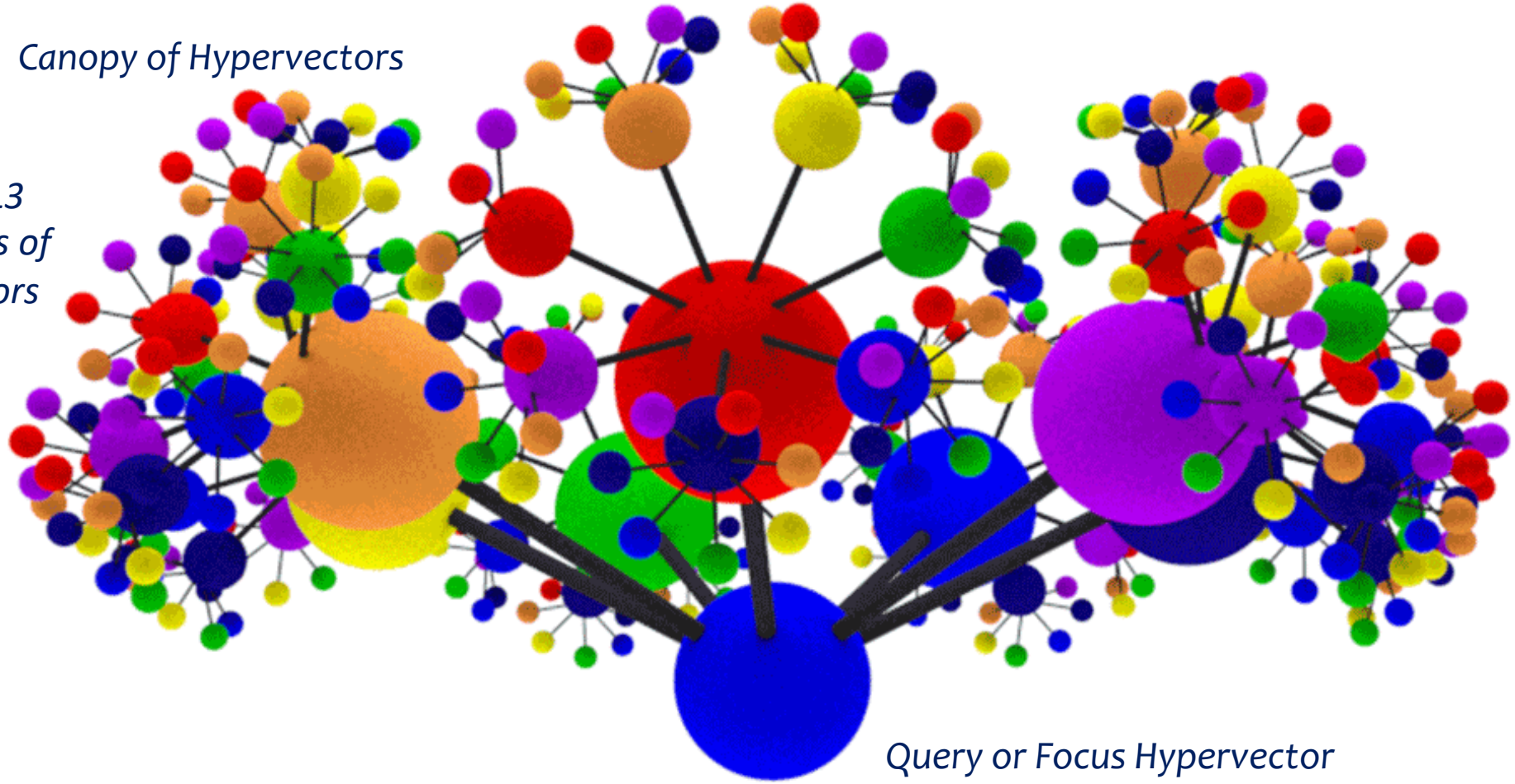
A Hypothetical approach to a Cognitive Computing Framework

- *"Ensemble" can be thought of as the fusion of a collection of existing and emerging technology items and ideas along with insights into the human brain provided by neuroscience into a cognitive computing framework – Hyperdimensional Computing lies at its core*
- *Ensemble is modeled after the cognitive abilities of the human brain*
- *Intended to overlay the von Neumann Architecture in much the same way as the neo-cortex overlays the old Limbic Brain (Reptilian Complex)*
- *The expectation is to augment both human intelligence (not replace it) and to augment the von Neumann architecture in a new, unique and novel turn of events*

Clustering or KNN | Hierarchical Affinity Propagation

Canopy of Hypervectors

L0,L1, L2, L3 Hierarchies of Hypervectors



Query or Focus Hypervector

Intel Patent Applications of Note: US 20190227750

US 20190227750
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR PERFORMING TESNOR OPERATIONS IN MEMORY"

Applicant: Intel Corporation, Santa Clara, CA

Covers Tensor Operations ("Computational Memory") on DIMMs mounted in the Memory Channel.

US 20190227750A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2019/0227750 A1**

Srinivasan et al. (43) **Pub. Date: Jul. 25, 2019**

(54) **TECHNOLOGIES FOR PERFORMING TENSOR OPERATIONS IN MEMORY**

(71) Applicant: **Intel Corporation**, Santa Clara, CA (US)

(72) Inventors: **Srikanth Srinivasan**, Portland, OR (US); **Richard Coulson**, Portland, OR (US); **Rajesh Sundaram**, Folsom, CA (US); **Bruce Querbach**, Hillsboro, OR (US); **Jawad B. Khan**, Portland, OR (US); **Shigeki Tomishima**, Portland, OR (US); **Sriram Vangal**, Portland, OR (US); **Wei Wu**, Portland, OR (US); **Chetan Chauhan**, Folsom, CA (US)

(21) Appl. No.: **16/370,007**

(22) Filed: **Mar. 29, 2019**

Publication Classification

(51) **Int. Cl.**
G06F 3/06 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 3/0659*** (2013.01); *G06N 3/08* (2013.01); ***G06F 3/0679*** (2013.01); ***G06F 3/0604*** (2013.01)

(57) **ABSTRACT**

Technologies for performing tensor operations in memory include a memory comprising media access circuitry coupled to a memory media having a cross point architecture. The media access circuitry is to access matrix data from the memory media, perform a tensor operation on the matrix data, and write, to the memory media, resultant data indicative of a result of the tensor operation.

Intel Patent Applications of Note: US 20190227981

US 20190227981
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR PROVIDING A SCALABLE ARCHITECTURE FOR PERFORMING COMPUTE OPERATION IN MEMORY"

Applicant: Intel Corporation, Santa Clara, CA

Covers "Computational Memory" on DIMMs mounted in the Memory Channel.

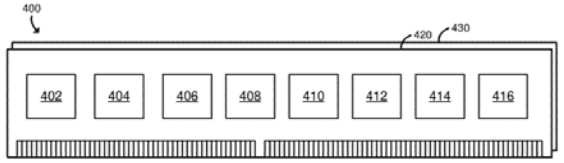


FIG. 4

US 20190227981A1

(19) **United States**
(12) **Patent Application Publication**
Tomishima et al.
(10) **Pub. No.: US 2019/0227981 A1**
(43) **Pub. Date: Jul. 25, 2019**

(54) **TECHNOLOGIES FOR PROVIDING A SCALABLE ARCHITECTURE FOR PERFORMING COMPUTE OPERATIONS IN MEMORY**

(71) Applicant: **Intel Corporation**, Santa Clara, CA (US)

(72) Inventors: **Shigeki Tomishima**, Portland, OR (US); **Srikanth Srinivasan**, Portland, OR (US); **Chetan Chauhan**, Folsom, CA (US); **Rajesh Sundaram**, Folsom, CA (US); **Jawad B. Khan**, Portland, OR (US)

(21) Appl. No.: **16/368,983**

(22) Filed: **Mar. 29, 2019**

Publication Classification

(51) **Int. Cl.**

G06F 15/78	(2006.01)
G06F 17/16	(2006.01)
G06F 15/80	(2006.01)

(52) **U.S. Cl.**
CPC ***G06F 15/7867*** (2013.01); ***G06F 15/803*** (2013.01); ***G06F 17/16*** (2013.01); ***G06F 15/7842*** (2013.01)

(57) **ABSTRACT**

Technologies for providing a scalable architecture to efficiently perform compute operations in memory include a memory having media access circuitry coupled to a memory media. The media access circuitry is to access data from the memory media to perform a requested operation, perform, with each of multiple compute logic units included in the media access circuitry, the requested operation concurrently on the accessed data, and write, to the memory media, resultant data produced from execution of the requested operation.

Intel Patent Applications of Note: US 20190227981

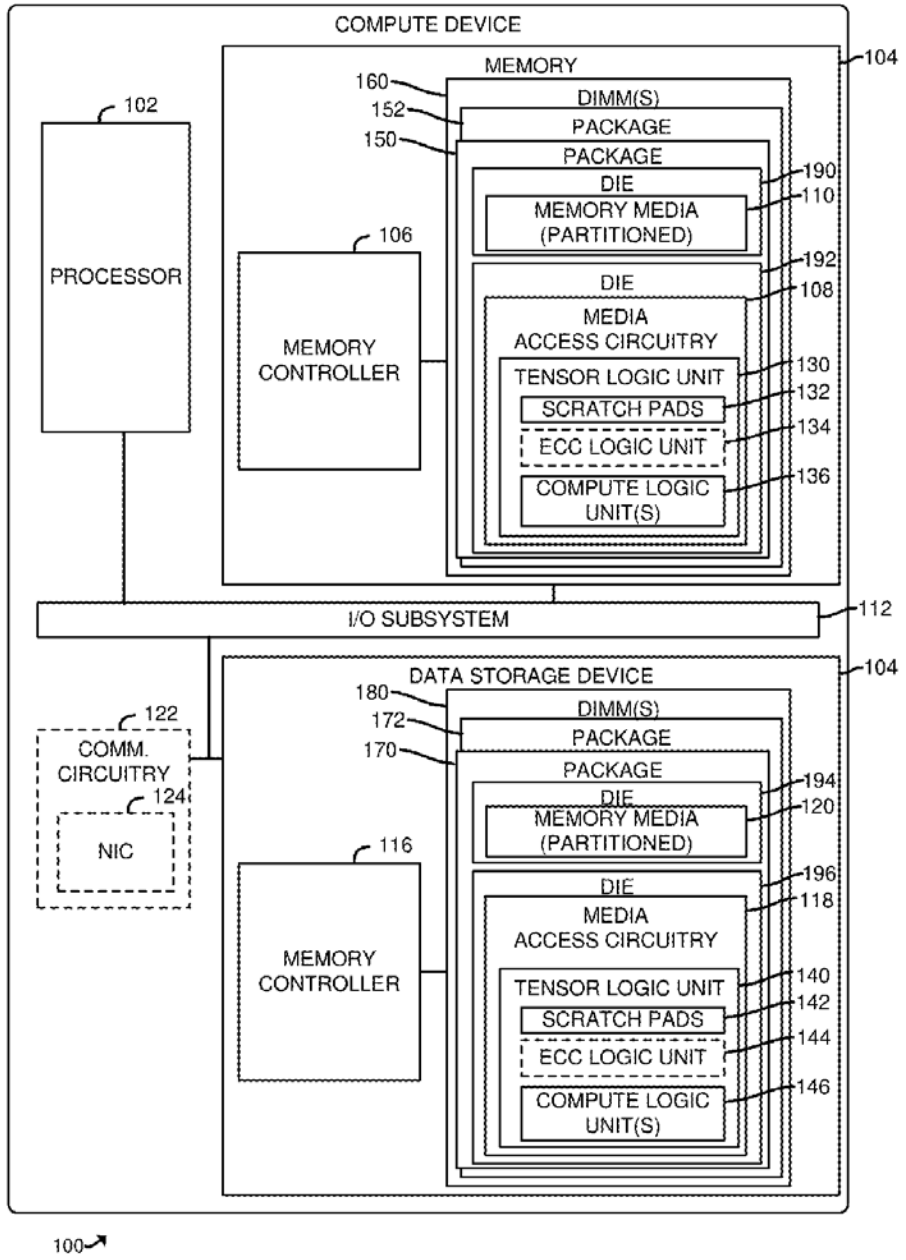
US 20190227981
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR PROVIDING A SCALABLE ARCHITECTURE FOR PERFORMING COMPUTE OPERATION IN MEMORY"

Applicant: Intel Corporation, Santa Clara, CA

Covers:

- ***Logical Block Diagrams of Optane DIMMs & Drives***
- ***Base System Framework***



Patent Applications of Note: US 20190227981

US 20190227981
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR PROVIDING A SCALABLE ARCHITECTURE FOR PERFORMING COMPUTE OPERATION IN MEMORY"

Applicant: Intel Corporation, Santa Clara, CA

Covers Bit Level Addressable Persistent Memory:

- 3DXPoint
- Chalcogenide (memresistive)
- FeTRAM
- Nanowire-based Non-volatile
- STT-RAM
- MRAM

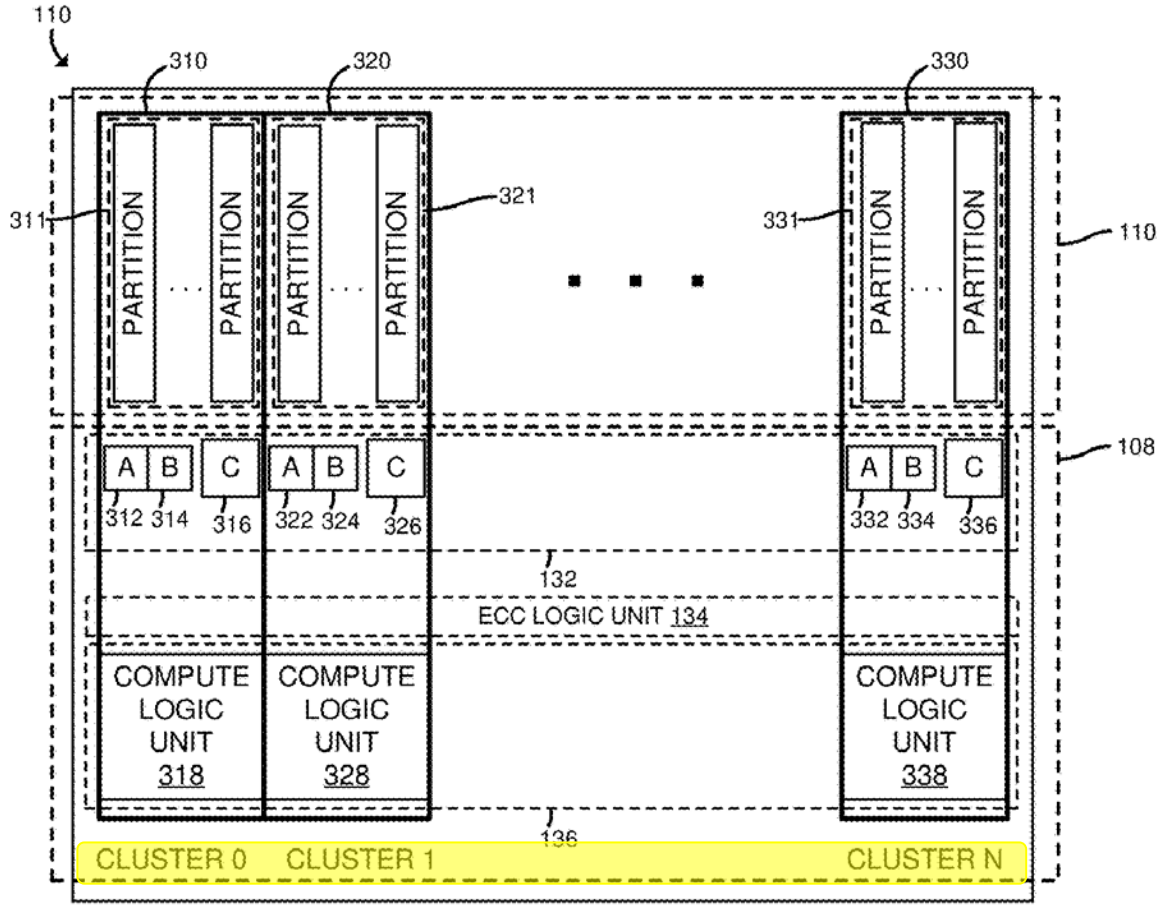


FIG. 3

Patent Applications of Note: US 20190227739

US 20190227739
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR PROVIDING STOCHASTIC KEY-VALUE STORAGE"

Applicant: Intel Corporation, Santa Clara, CA

Enables nearest neighbor search by searching columns of a stochastic associative array of a hyper-dimensional vector table in the memory, identifying the closest matching row in the stochastic associative array relative to the reference hyper-dimensional vector, wherein the closest matching row indicates a closest matching value hyper-dimensional vector, and output a value associated with the closest matching value hyper-dimensional vector.


US 20190227739A1

(19) **United States**
(12) **Patent Application Publication**
Khan et al.
(10) **Pub. No.: US 2019/0227739 A1**
(43) **Pub. Date: Jul. 25, 2019**

(54) **TECHNOLOGIES FOR PROVIDING STOCHASTIC KEY-VALUE STORAGE**

(71) Applicant: **Intel Corporation**, Santa Clara, CA (US)

(72) Inventors: **Jawad B. Khan**, Portland, OR (US); **Richard Coulson**, Portland, OR (US)

(21) Appl. No.: **16/369,996**

(22) Filed: **Mar. 29, 2019**

Publication Classification

(51) **Int. Cl.**

G06F 3/06	(2006.01)
G06F 12/02	(2006.01)
G06F 11/10	(2006.01)
G06F 9/30	(2006.01)

(52) **U.S. Cl.**
CPC ***G06F 3/0658*** (2013.01); ***G06F 12/0207*** (2013.01); ***G06F 3/0683*** (2013.01); ***G06F 9/30036*** (2013.01); ***G06F 3/0604*** (2013.01); ***G06F 11/10*** (2013.01)

(57) **ABSTRACT**

Technologies for performing a hyper-dimensional operation in a memory of the compute device include a memory and a memory controller. The memory controller is configured to receive a query from a requestor and determine, in response to a receipt of the query, a key hyper-dimensional vector associated with the query, perform a hyper-dimensional operation to determine a reference hyper-dimensional vector associated with a value to the key. The memory controller is further configured to perform a nearest neighbor search by searching columns of a stochastic associative array of a hyper-dimensional vector table in the memory, identify a closest matching row in the stochastic associative array relative to the reference hyper-dimensional vector, wherein the closest matching row indicates a closest matching value hyper-dimensional vector, and output a value associated with the closest matching value hyper-dimensional vector.

Patent Applications of Note:US 20190227739 (Page 2)

Optane Computational Memory Acceleration Fabric?

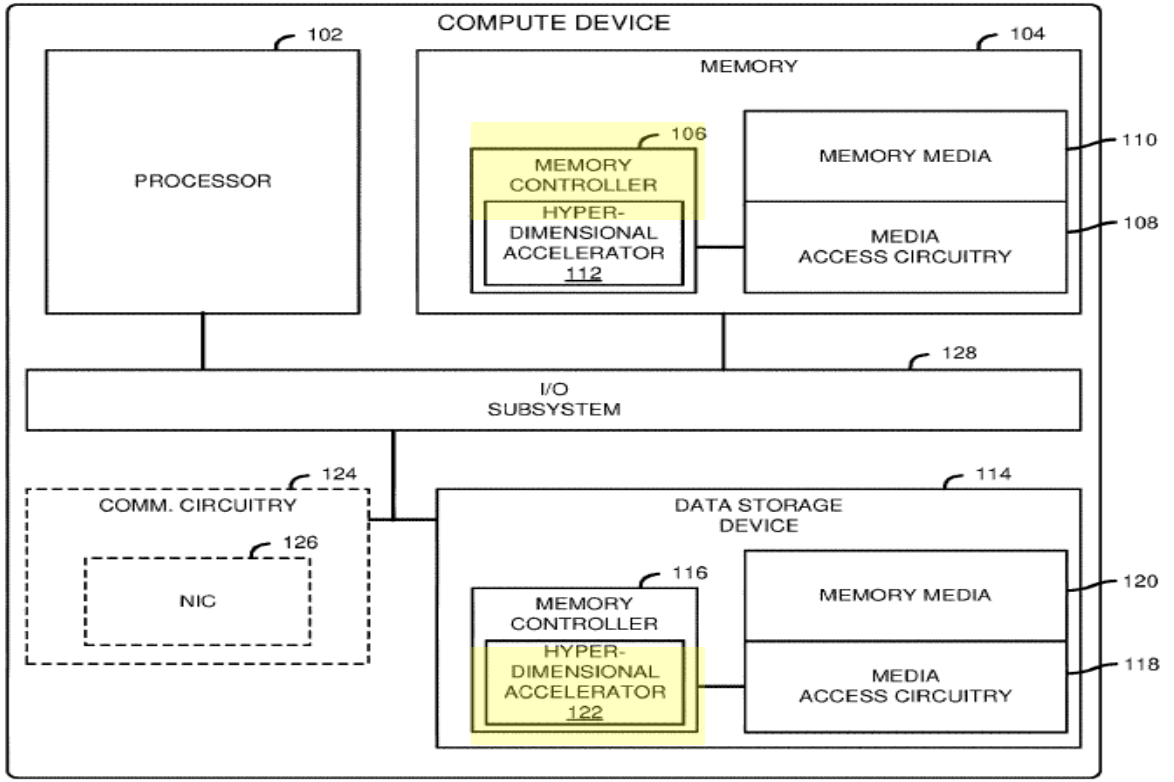


FIG. 1

Patent Applications of Note: US 20190227808

US 20190227808
Filed: Mar. 29, 2019
Published: July 25, 2019:

"TECHNOLOGIES FOR EFFICIENT EXIT FROM HYPER-DIMENSIONAL SPACE IN PRESENSE OF ERRORS"

Applicant: Intel Corporation, Santa Clara, CA

Covers "Hyperdimensional" Key Value Store ECC operations

US 20190227808A1

(19) **United States**
(12) **Patent Application Publication**
Khan et al.
(10) **Pub. No.: US 2019/0227808 A1**
(43) **Pub. Date: Jul. 25, 2019**

(54) **TECHNOLOGIES FOR EFFICIENT EXIT FROM HYPER-DIMENSIONAL SPACE IN THE PRESENCE OF ERRORS**

(71) Applicant: **Intel Corporation**, Santa Clara, CA (US)

(72) Inventors: **Jawad B. Khan**, Portland, OR (US); **Richard Coulson**, Portland, OR (US)

(21) Appl. No.: **16/370,013**

(22) Filed: **Mar. 29, 2019**

Publication Classification

(51) **Int. Cl.**

G06F 9/38	(2006.01)
G06F 9/30	(2006.01)
G06F 15/80	(2006.01)

(52) **U.S. Cl.**
CPC ***G06F 9/3895*** (2013.01); ***G06F 15/8092*** (2013.01); ***G06F 9/30036*** (2013.01)

(57) **ABSTRACT**

Technologies for performing hyper-dimensional operations in memory includes a device with a memory media and a memory controller. The memory controller is configured to receive a query from a requestor and determine, in response to receiving the query, a reference hyper-dimensional vector associated with the query. The memory controller is further configured to perform a nearest neighbor search by searching columns of a stochastic associative array in the memory media to determine a number of matching bit values for each row relative to the reference hyper-dimensional vector, wherein each bit in a column of the stochastic associative array represents a bit value of a corresponding row, identify a closest matching row that has a highest number of matching bit values, and output data of the closest matching row.

Related Intel Patent Applications:

CONDUCTIVE BRIDGE RANDOM ACCESS MEMORY (CBRAM) DEVICES WITH LOW THERMAL CONDUCTIVITY ELECTROLYTE SUBLAYER

Patent Application Publication No.: US 20190229264

Applicant: Intel Corporation, Santa Clara, CA

KARPOV et al

Pub. Date: Jul. 25, 2019

PCT Filed: Sep. 30, 2016

TECHNOLOGIES FOR EFFICIENT STOCHASTIC ASSOCIATIVE SEARCH OPERATIONS

Patent Application Publication No.: US 20190220230

Applicant: Intel Corporation, Santa Clara, CA

Khan et al.

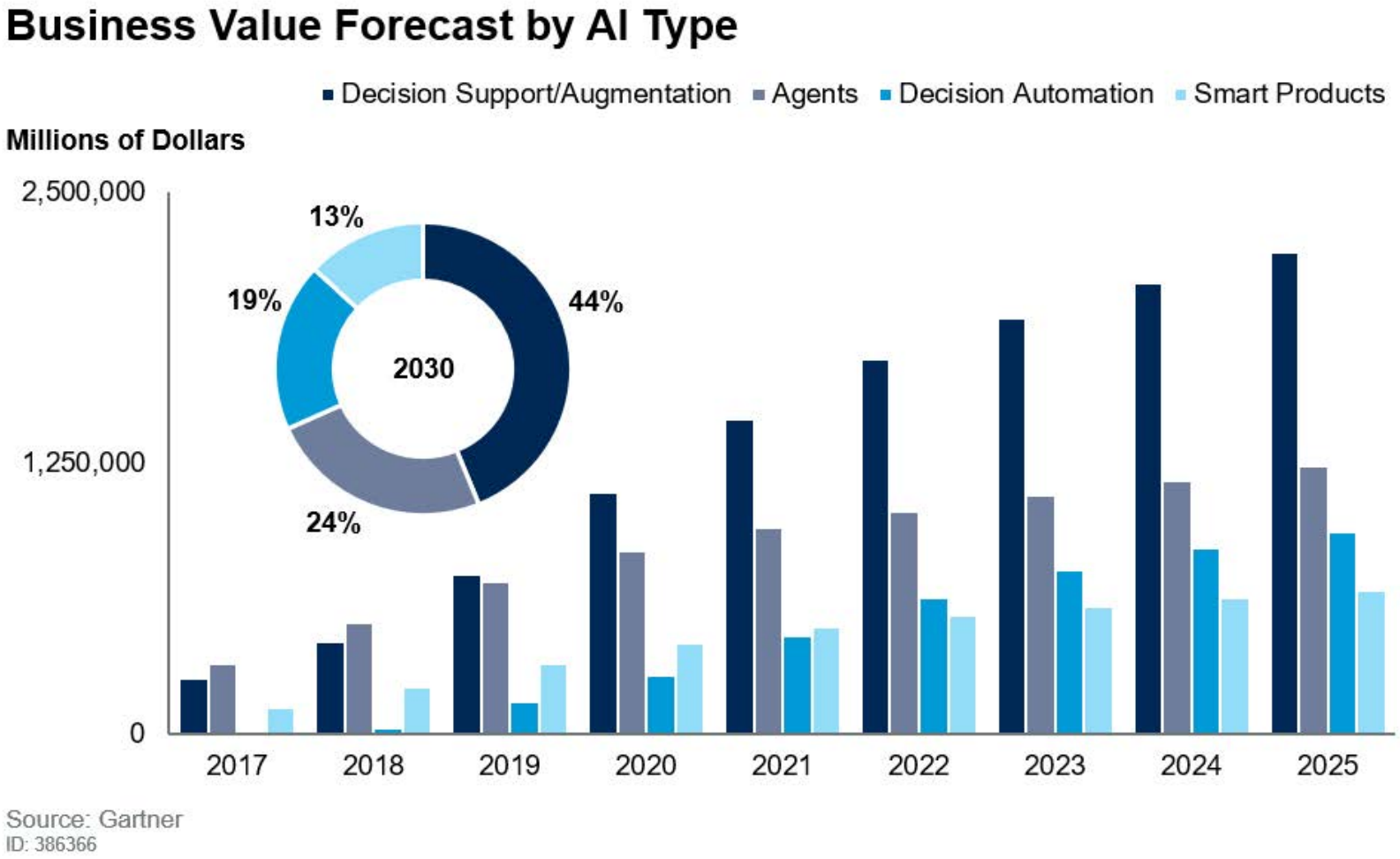
Pub. Date: Jul. 18, 2019

Filed: Mar. 28, 2019

Augmented Intelligence

- *In 2021, artificial intelligence (AI) augmentation will create $2.9 trillion of business value and 6.2 billion hours of worker productivity globally, according to Gartner, Inc.*
- *Gartner defines augmented intelligence as a human-centered partnership model of people and AI working together to enhance cognitive performance. This includes learning, decision making and new experiences*

Augmented Intelligence
Gartner Forecast by AI Type



HDC | Intel | Augmented Intelligence
"HDC" the Killer Application

- *WebFeet Research believes the market is at the beginning of a "Cognitive Fracture" – a scenario wherein the value of HDC breaks the "Big Data" blockage allowing access to formerly hidden Data while concurrently allowing individually tailored augmented intelligence accelerated solutions to be had in real time*
- *We also believe it is the beginning of "Augmented Associative Database Systems" igniting in turn the growth of augmented intelligence*

Conclusionary Notes:

- *Hyperdimensional Computing is a "Brain Inspired" ensemble technology approach to solving the problem of Synthetic Intelligence*
- *The idea of an Associative Memory Acceleration allows the "clustering" of Hypervectors at high speed and low power*
- *WebFeet Research Inc. is in the process of introducing the novel idea of "Hierarchical Affinity Propagation Vector Stack" – a construct that provides a persistent memory base for building directed graph networks of associative objects necessary for rapid, low power execution in Synthetic Intelligent Cognitive Computing Systems*