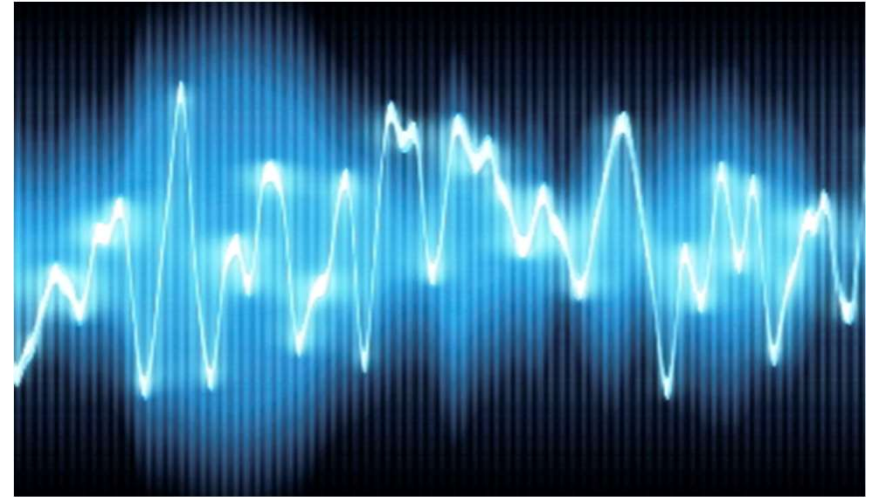


Deep Learning Revolution: From Theory to Impact



Chris Rowen

CEO
Babblelabs Inc.

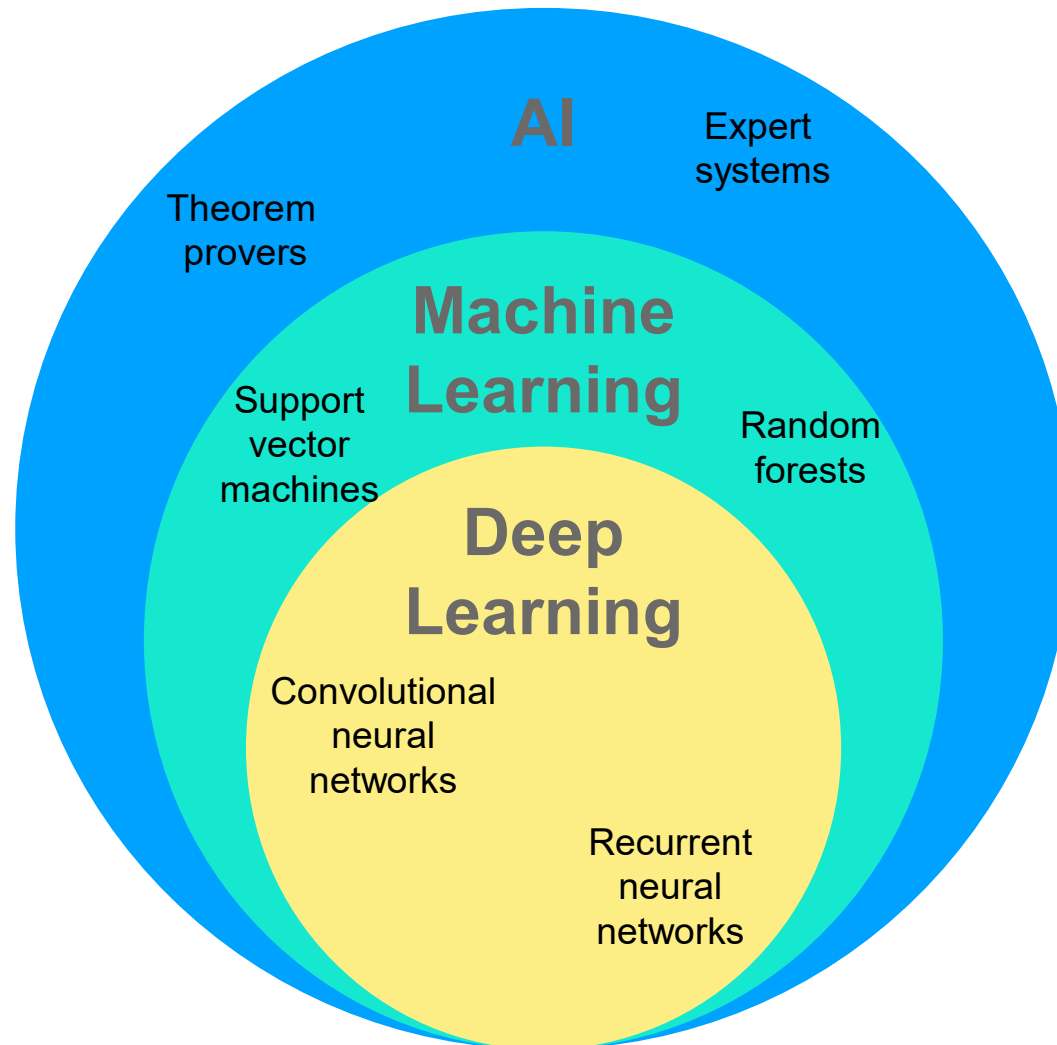
EDPS

September 2018

Hype Metrics:

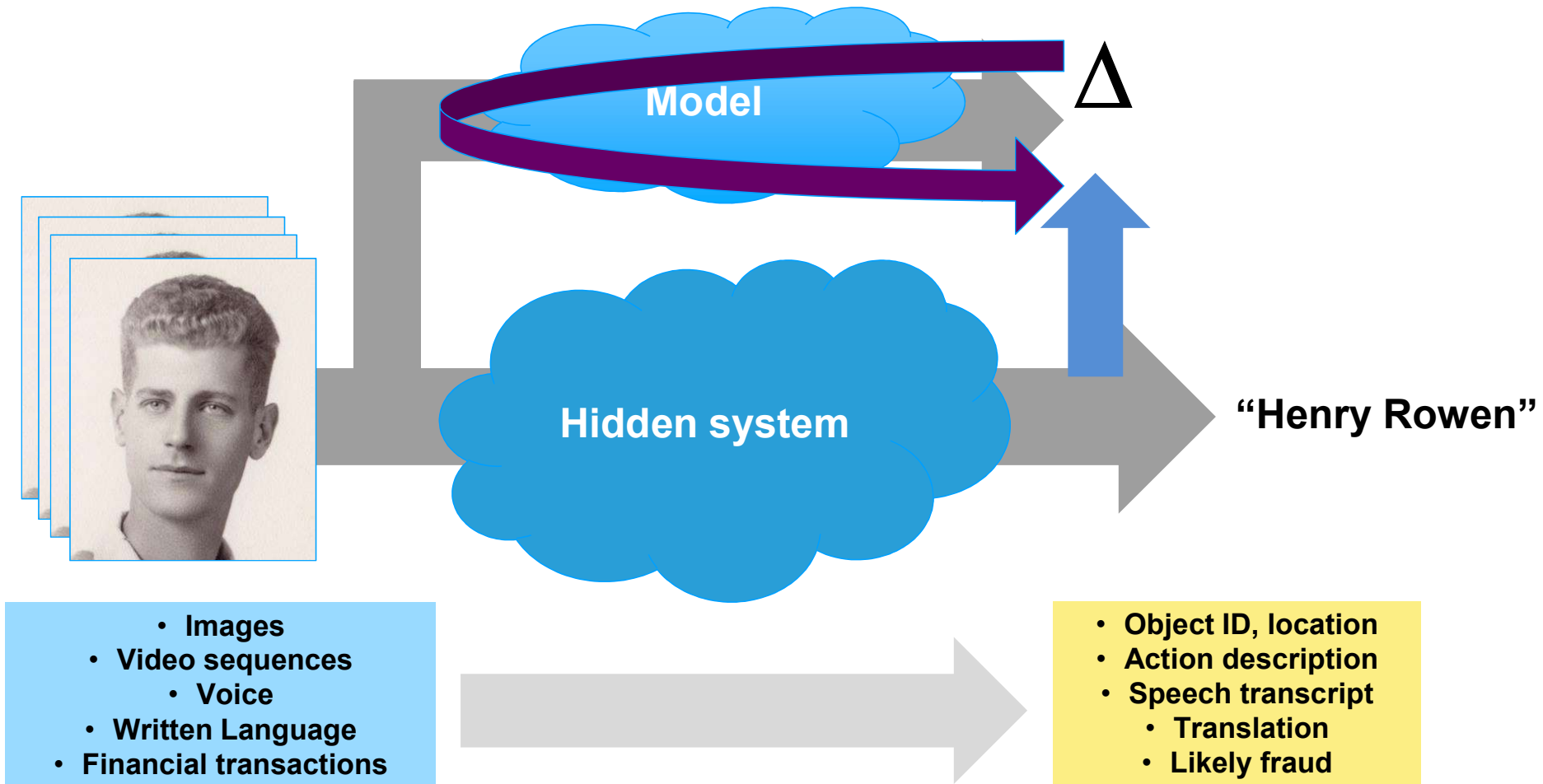
- One Google page hit on “AI” for every person in US + India + China
- 11,300 “artificial intelligence” startups [CrunchBase]
- 16,500 papers on “neural network” on arxiv.org – most in past 24 months

Quick Taxonomy



Deep Learning Foundations

The construction of a complex numerical model that mimics the behavior of a very complex but hidden system:



Vision is Fundamentally Hard

- Big computation in embedded inference, **huge** in training (but less frequent)
- Typically need large *labeled* data-sets
- Example: ImageNet Classification:
 - 1.2M images
 - 1000 categories
 - 120 breeds of dogs



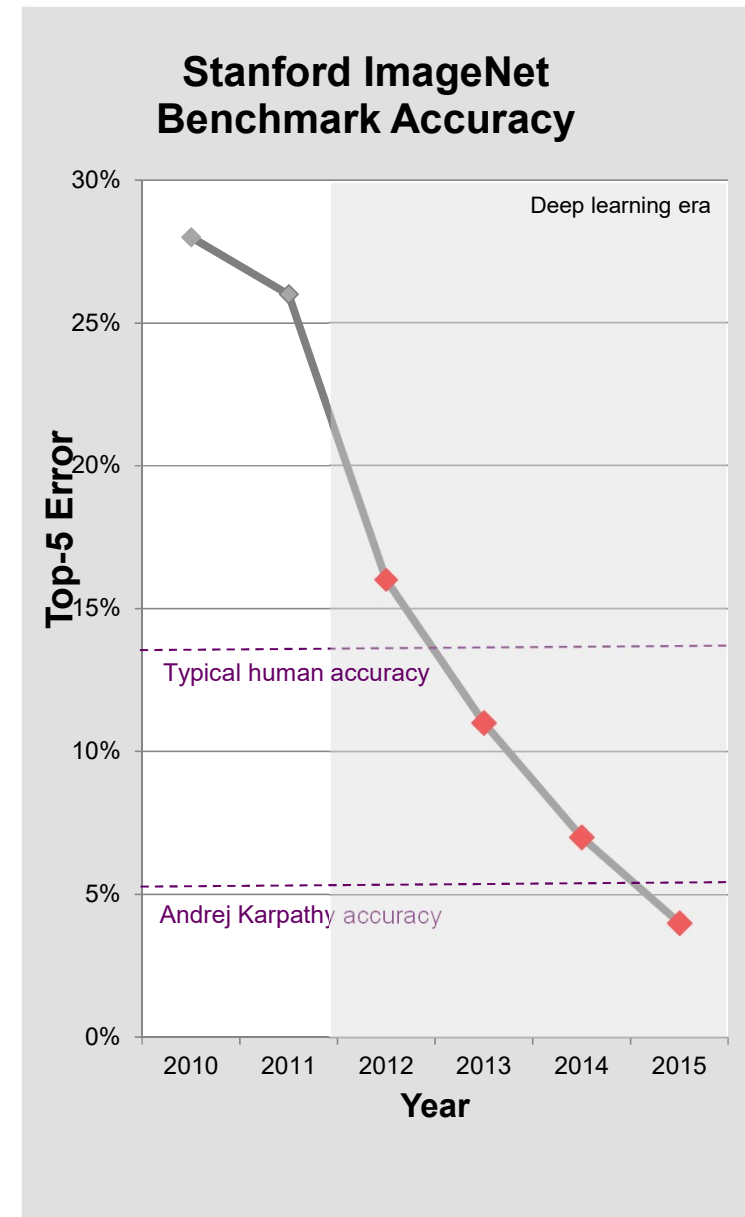
Tibetan mastiff



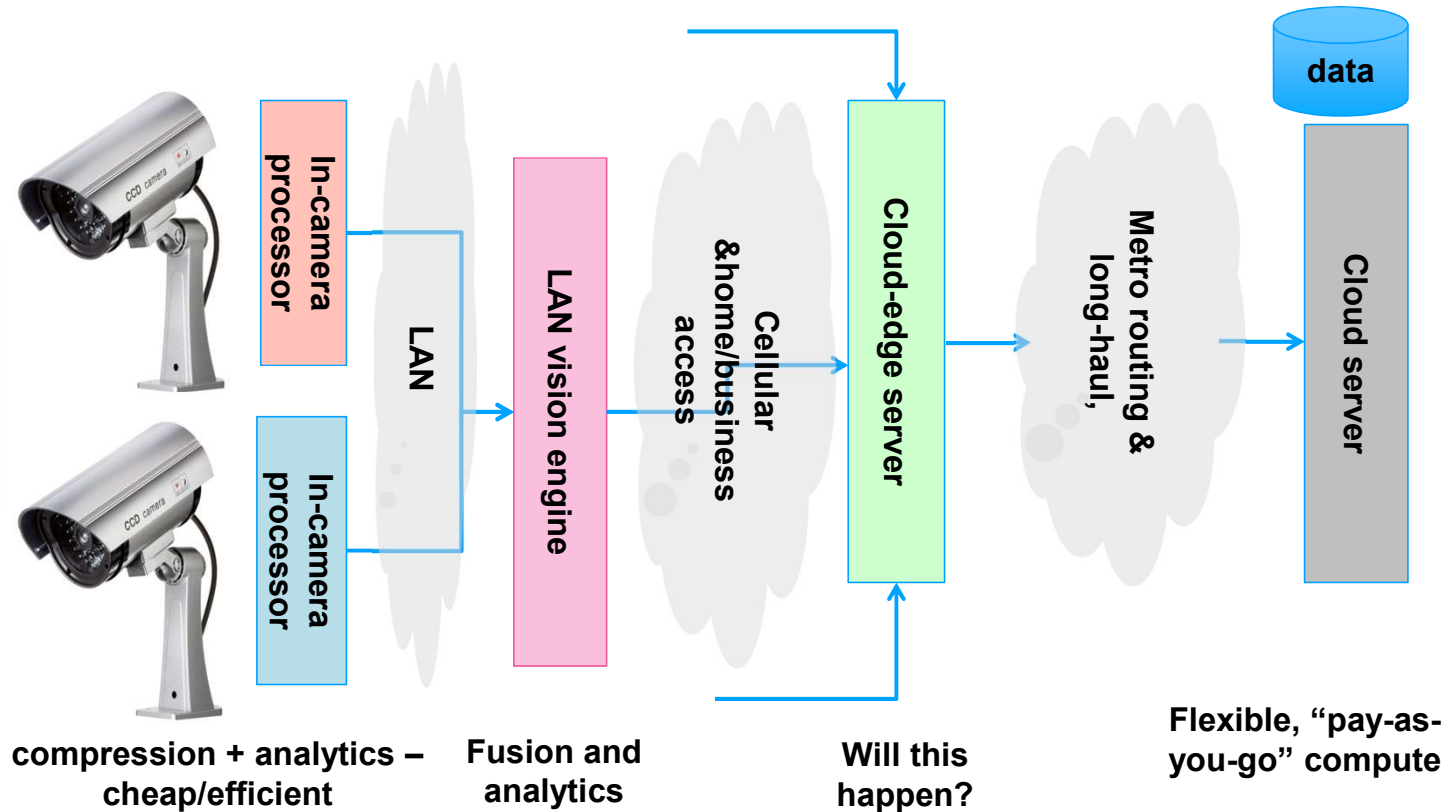
Shih-Tzu



Norwegian
elkhound



Where will we put the “smarts”?



System responsiveness

Scope of data analysis

Privacy

Computing and network cost

Security, Robustness and Privacy

- More risks for privacy: more cameras, more correlation to other streams
- Mission-critical surveillance susceptible to new attacks

The usual network and device attacks

+ Database manipulation to inject bias

+ Classification spoofing

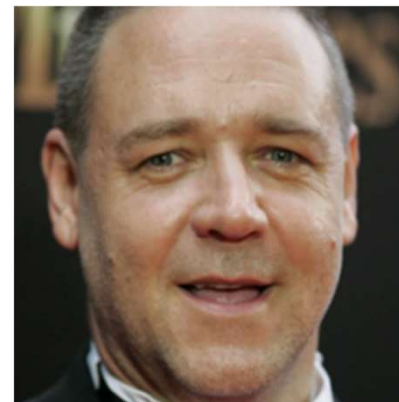
- Example: Spoofing facial recognition



Reese Witherspoon



**Reese Witherspoon
in patterned frames**



**Recognized as
Russell Crowe**

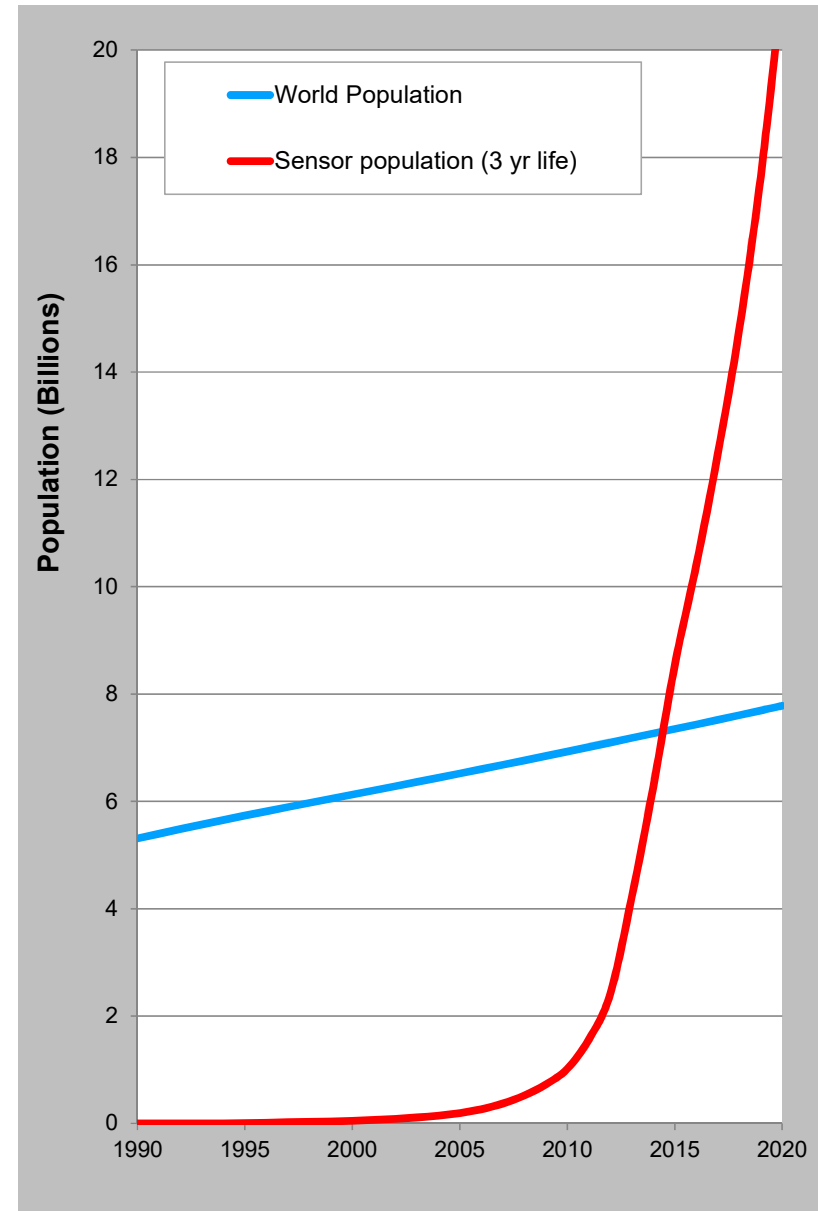
Vision:

the pixel explosion

- Rapid replacement of traditional vision by deep learning
- From 2015: more image sensors than people
- 99% of all new raw data is pixels (rest is audio)
- Massive bandwidth:
 2×10^{10} sensors * 5×10^8 pixels/sec = 10^{19} raw pixels/s



\$13⁹⁹ prime
In Stock



How much does a \$10 camera cost?

| camera→ cloud H.264 compress | Camera | Power | Network | Storage | Compute | 3 year total cost per camera |
|---------------------------------------|-------------------------|------------------|----------------------------|---------------------------------------|--|------------------------------|
| | HD camera w/compression | At 1\$/watt-year | DSL/cable network @\$10/TB | Rolling 1 day of data @ \$25/TB/month | YOLO2 object detection on AWS G3 @ \$0.60/hr | |
| @ 60 fps | \$10 | \$9 | \$4,700 | \$400 | \$3,300 | ~\$8400 |
| @ 1fps | \$10 | \$9 | \$80 | \$7 | \$55 | ~\$165 |
| @ 0.1 fps | \$10 | \$9 | \$8 | \$1 | \$5 | ~\$35 |

Observations:

- *Cloud-based real-time vision requires “semantic compression” at the edge*
- *Completely autonomous analysis and action is also the lowest cost*
- *Vision at the edge is biggest deep learning silicon market*
 - *Autonomous vehicles and robots*
 - *Video monitoring*
 - *UI and social media with AR/VR*

Voice is Vision

Speech is the most human of interfaces

5B active electronic speech users

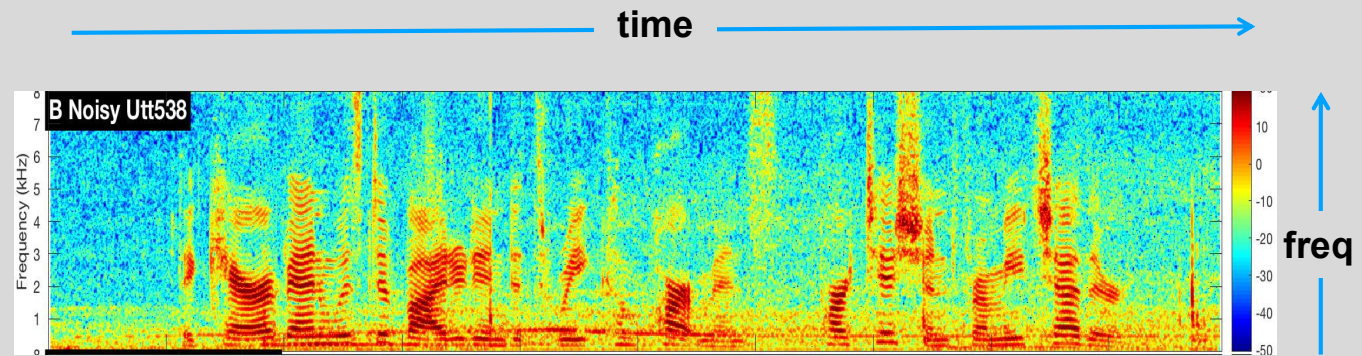
>20B microphones installed by 2020

Key services:

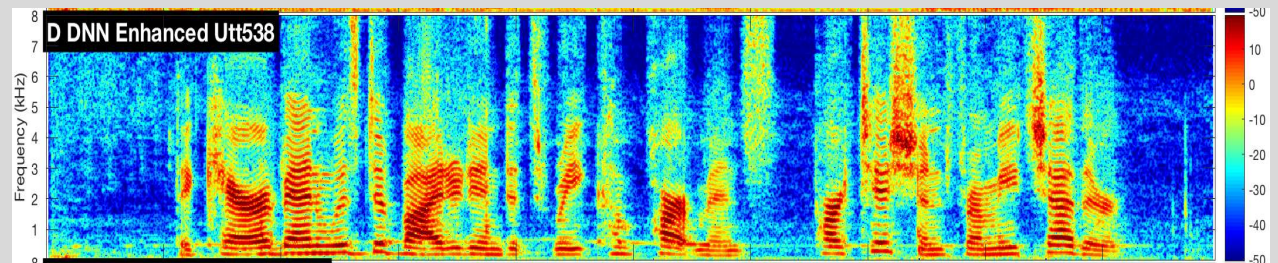
- Noise reduction
- Speech recognition
- Speaker authentication
- Speech synthesis

Shift to local compute

Transform speech signal into image: “spectrogram”

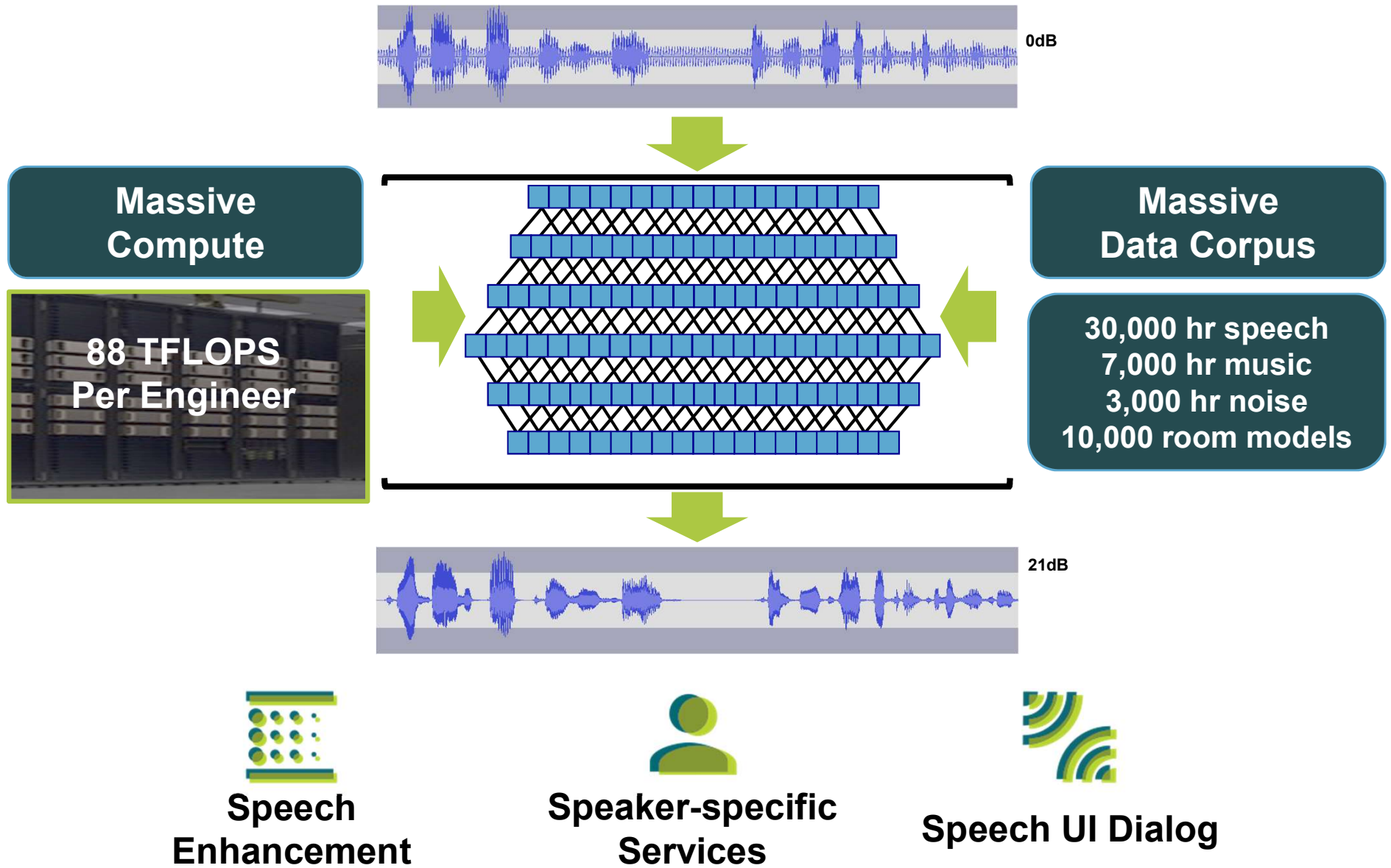


**Vision-style
neural
network**



BabbleLabs: Deep learning meets speech

fresh problems, more sophisticated models, more data, more training



Speech Enhancement Example

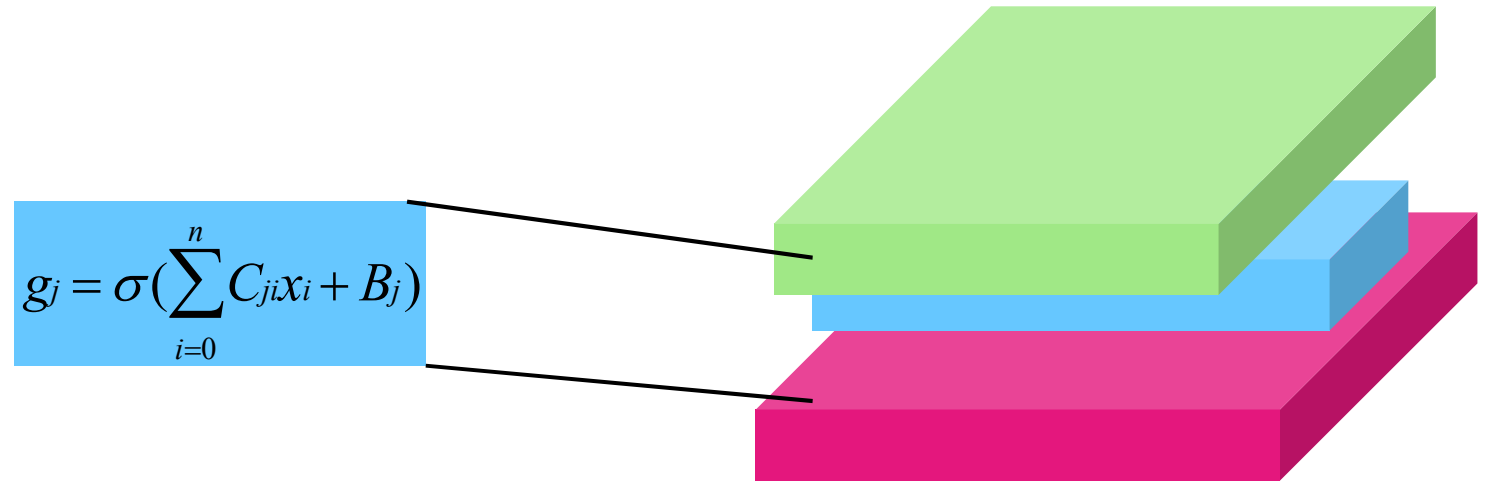
Original Video



Enhanced Video



Deep Learning Silicon Is Easy *especially for inference*



- Compute dominated by multiply-add
- Coefficients C_{ij} , B_j read-only, heavily reused
- Memory pattern regular, static and bounded
- Programmability lets hardware span many applications
- High-level frameworks hide architecture details

Deep Learning Silicon is Hard

- Impediments in efficiency:
 - mixed convolution sizes
 - non-unit strides and short, odd vector lengths
 - difficult parallelization
 - on-the-fly data reorganization
 - sparsity in coefficients and intermediate results
- Memory bandwidth;
 - large models (10s of MB)
 - fully connected layers (1 coefficient/MAC)
 - many-to-many communication between layers
 - CPU – Neural Network Engine data sharing
 - training >> inference
- MUST have comprehensive mapping, optimization and analysis SW from frameworks to silicon
- Silicon availability may be getting ahead of deployable applications

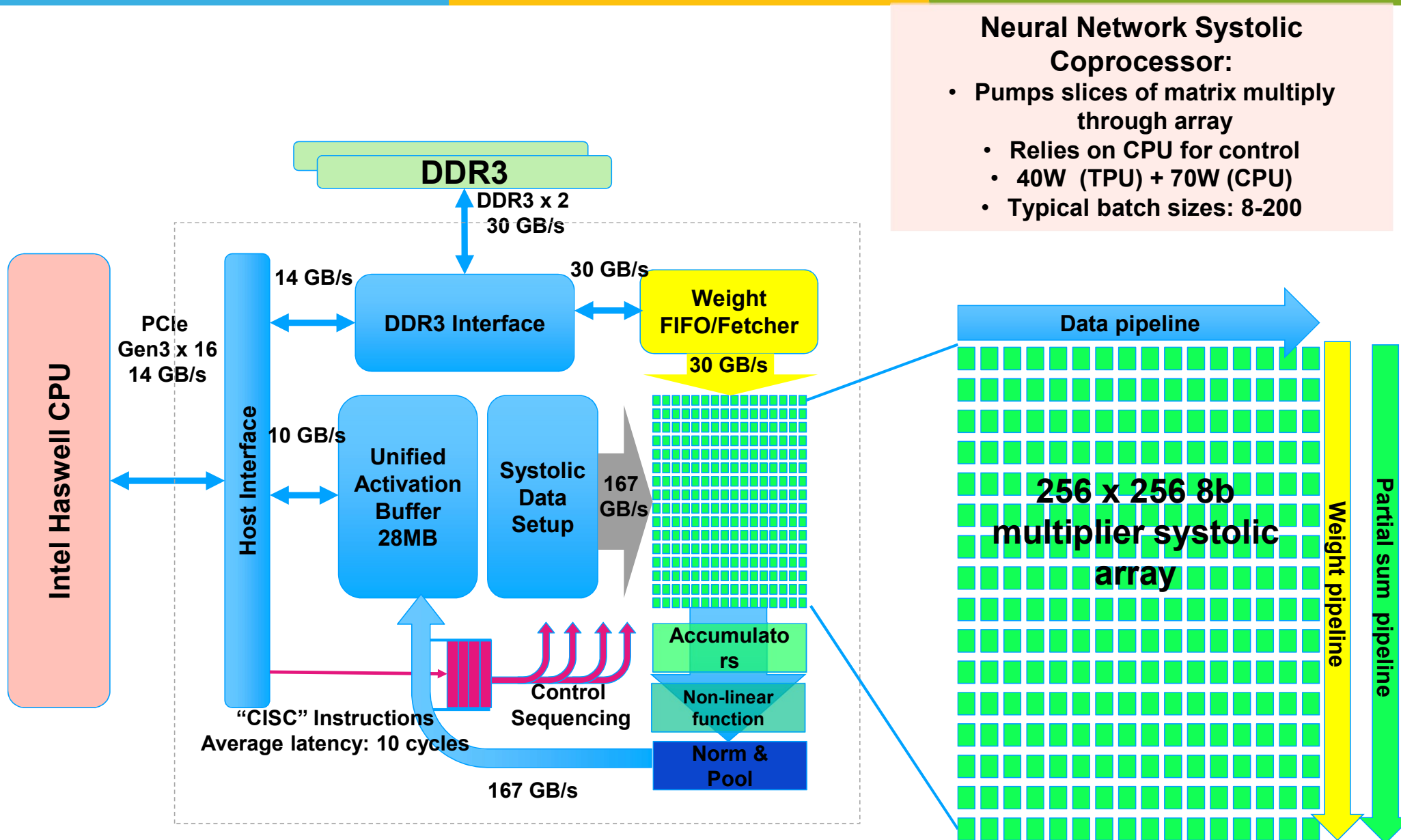


Anatomy of a Neural Network Accelerator

Example: Google Tensor Processing Unit (TPU)



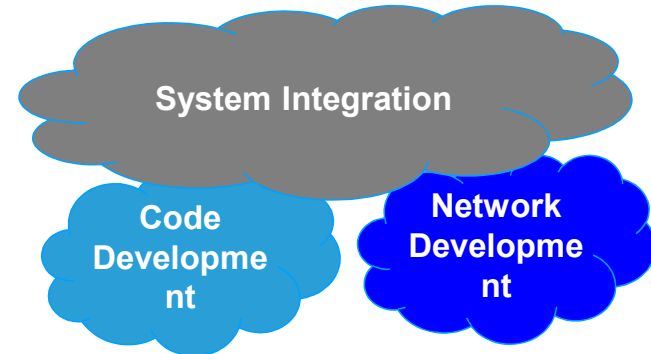
babblelabs



Implications for Semiconductors

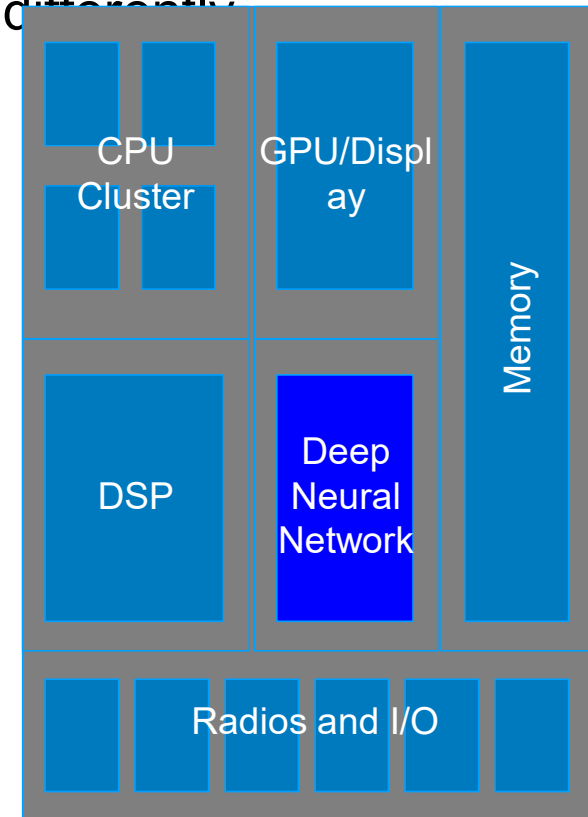
1. New computing model:

- New applications – especially vision & speech
- new focus on data-set access
- new development tools



2. Deep learning uses cloud and edge devices differently

- Cloud:
 - huge compute + big memory for training
 - inference on across aggregated data
- Edge
 - high-bandwidth real-time inference
 - minimum power and cost
 - latency- and privacy-critical applications

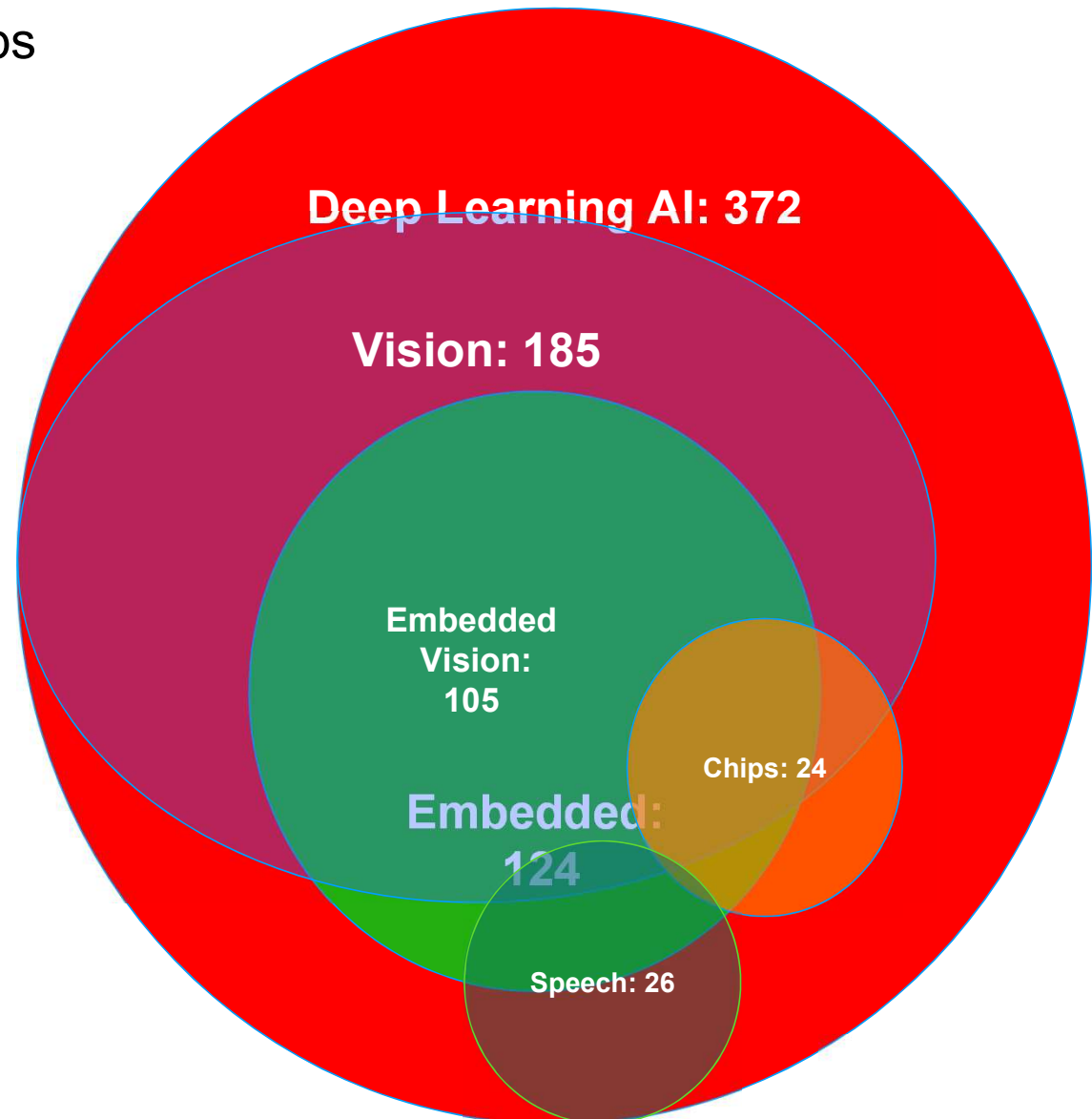


3. Neural networks very diverse

- Some run on existing μ controllers
- Some CPU, GPU and DSP → special compute

A Picture of Deep Learning Startups

- More than 2/3 of 372 startups focus on cloud software
- Half of all startups do vision
- Embedded dominated by vision
- Speech by startups just starting
- Many deep learning chip startups



Silicon Design Renaissance

Not Just the Big Chips and System Makers

- *Implication: high performance & low power inference will be widely available*

➔ In embedded devices:

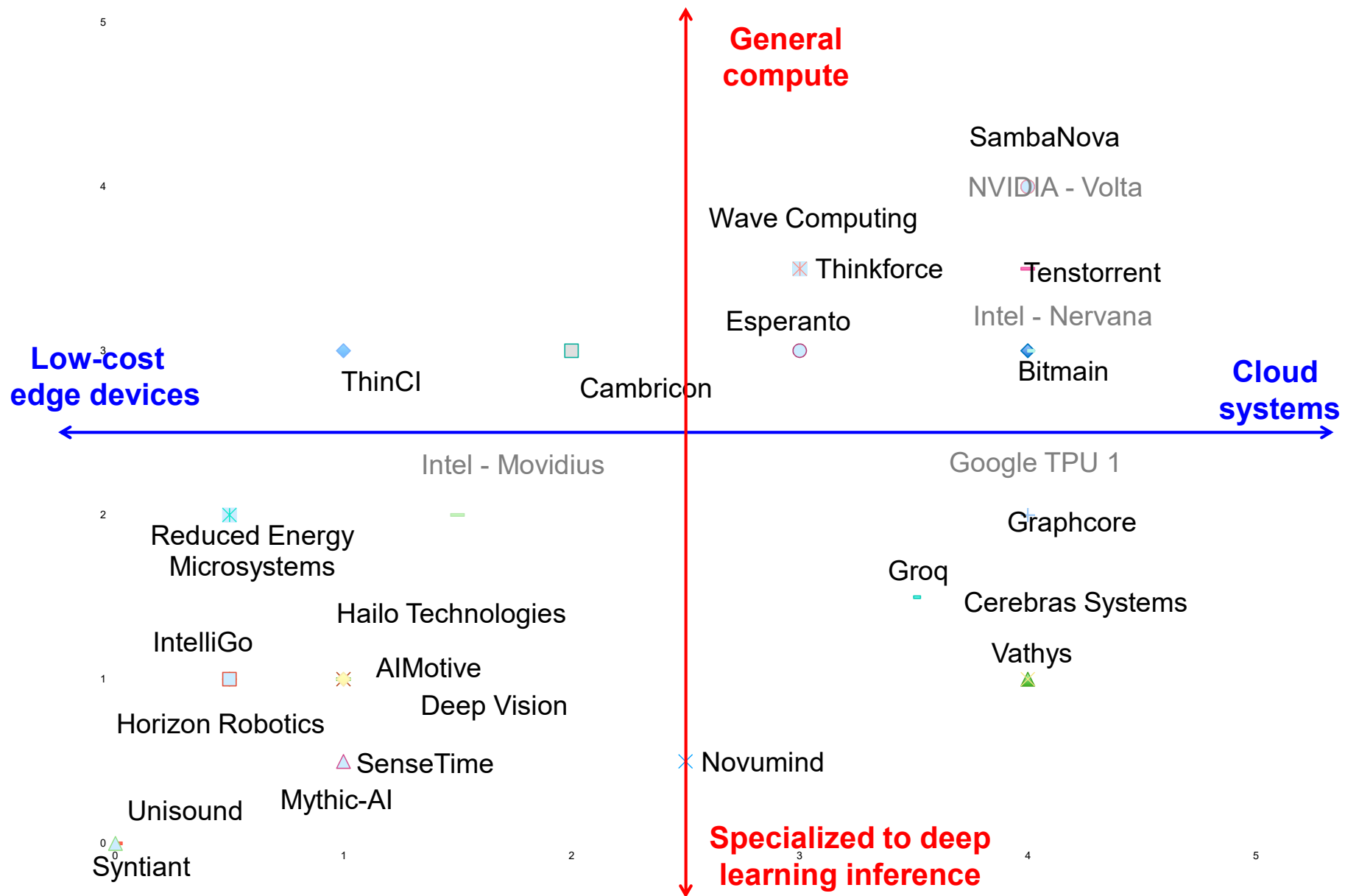
- vision
- speech

➔ In cloud

- network development: training
- application deployment: inference

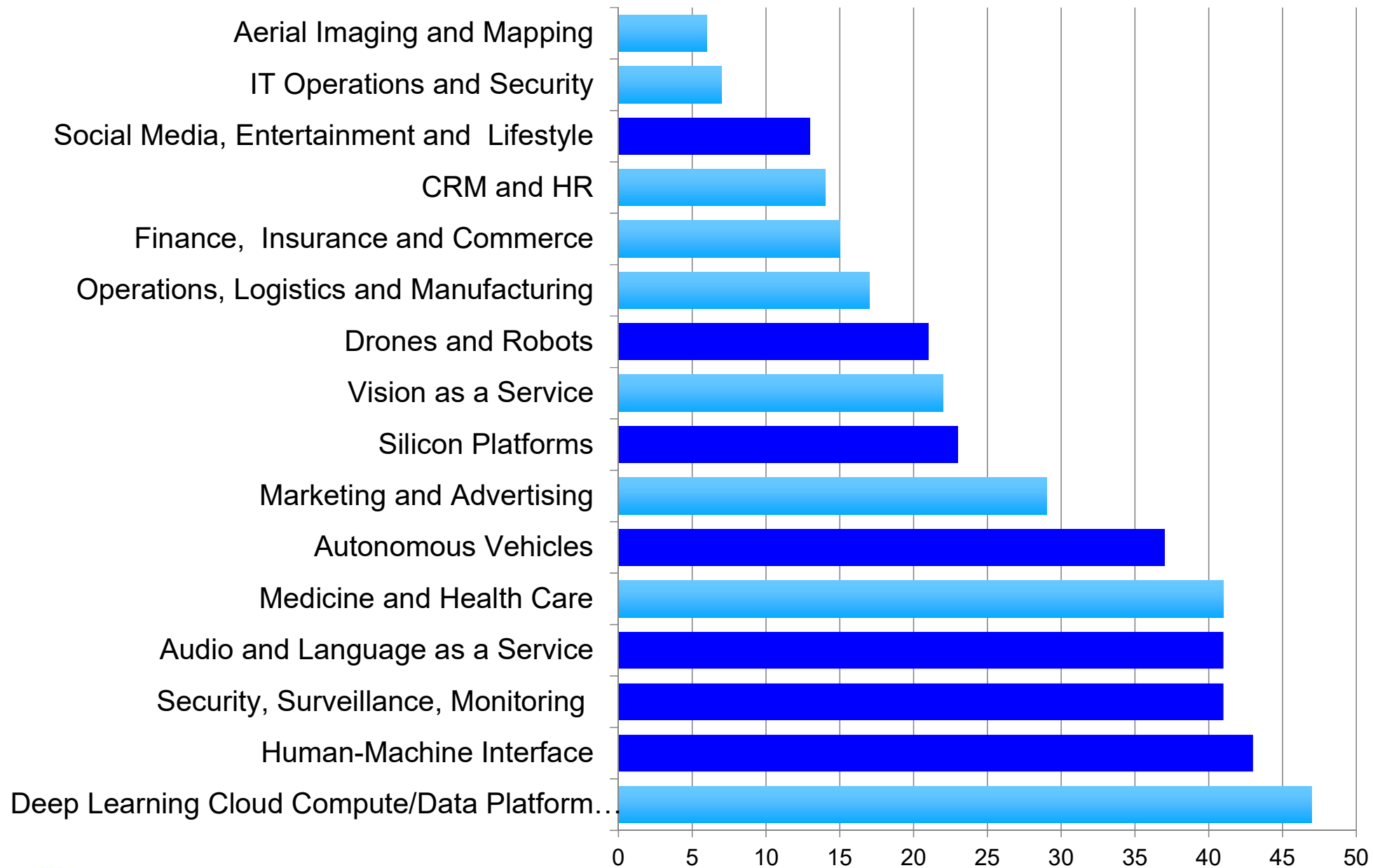
| | | |
|------------------------------------|---|---------|
| AIMotive | Portable software for automated driving | Hungary |
| Axis Semi | Massive array of compute cores | USA |
| Bitmain | Coin miner builds training ASIC | China |
| Cambricon | Device and cloud processors for AI | China |
| Cerebras Systems | Specialized chip for deep-learning applications | USA |
| Chipintelli | Speech recognition chip for local speech processing | China |
| Deep Vision | Low-power silicon architecture for computer vision | USA |
| Esperanto | Massive array of RISC-V cores | USA |
| FWDNXT | Low power image recognition and classification | USA |
| Graphcore | Graph-oriented processors for deep learning | UK |
| Groq | Google spinout doing deep learning chip | USA |
| Gyr Falcon Technology | Low-cost, low-power, high-performance Artificial Intelligence (AI) processors. | USA |
| Hailo Technologies | Specialized deep learning microprocessor | Israel |
| Horizon Robotics | Smart Home, automotive and Public safety | China |
| IntelliGo | Hardware and software for image and speech processing | China |
| Mythic-AI | Low power NN inference IC using flash+analog+digital | USA |
| Novumind | AI for IoT | USA |
| Preferred Networks | Real time data analytics with deep learning and Chainer library | Japan |
| Rain Neuromorphics | Nanotechnology for AI | USA |
| Reduced Energy Microsystems | Lowest power silicon for deep learning and machine vision | USA |
| SambaNova | Coarse Grain Reconfigurable Array for matrix arithmetic | USA |
| SenseTime | Computer vision | China |
| Syntient | Customized analog neural networks | USA |
| Tenstorrent | Deep learning processor: designed for faster training and adaptability to future algorithms | Canada |
| ThinCI | vision processing chips | USA |
| Thinkforce | AI chips | China |
| Unisound | AI-based speech and text | China |
| Vathys | Deep learning supercomputers | USA |
| Wave Computing | Deep Learning computers based on custom silicon | USA |

Sorting Out Deep Learning Silicon



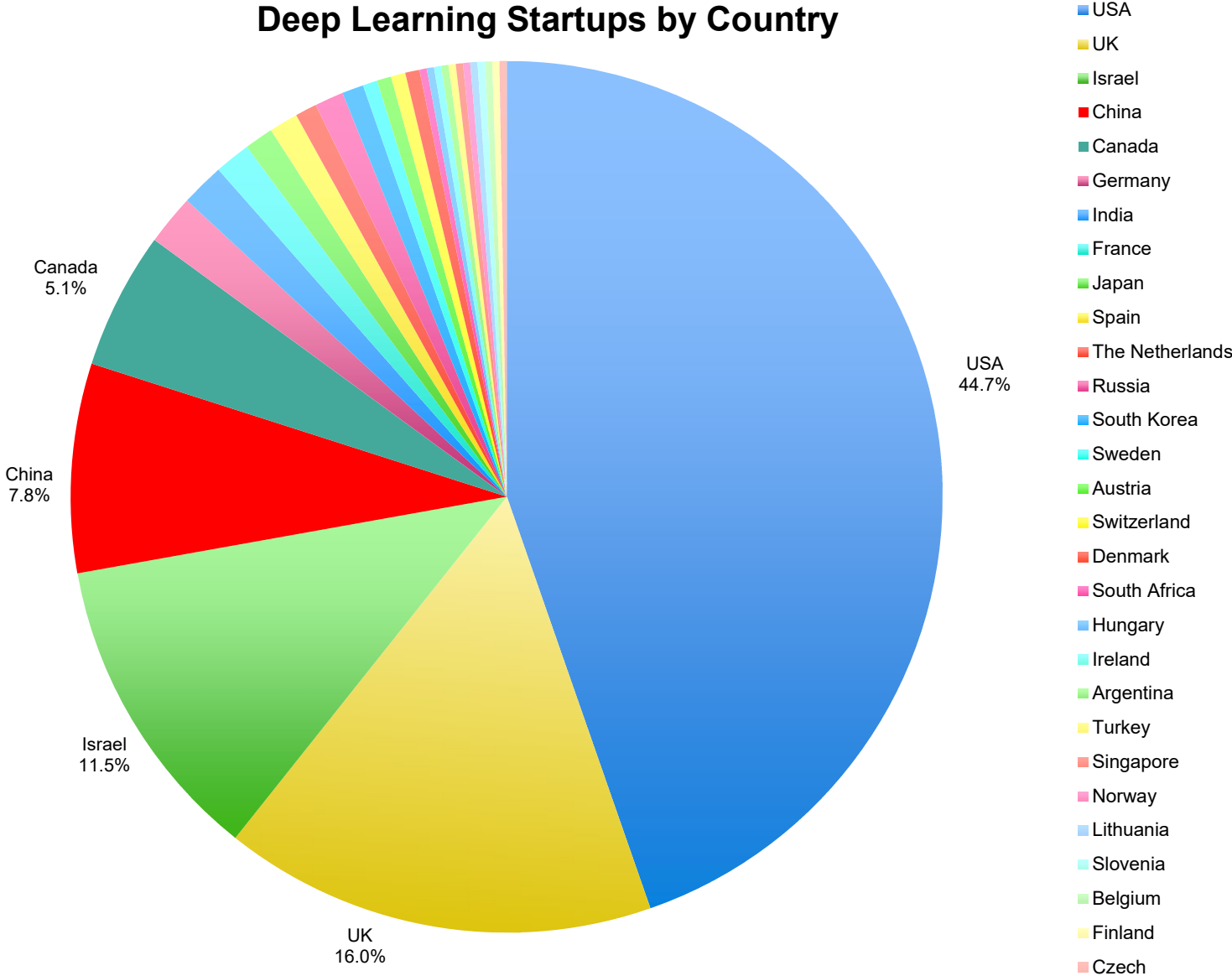
Segments for Deep Learning Startups

High potential volume



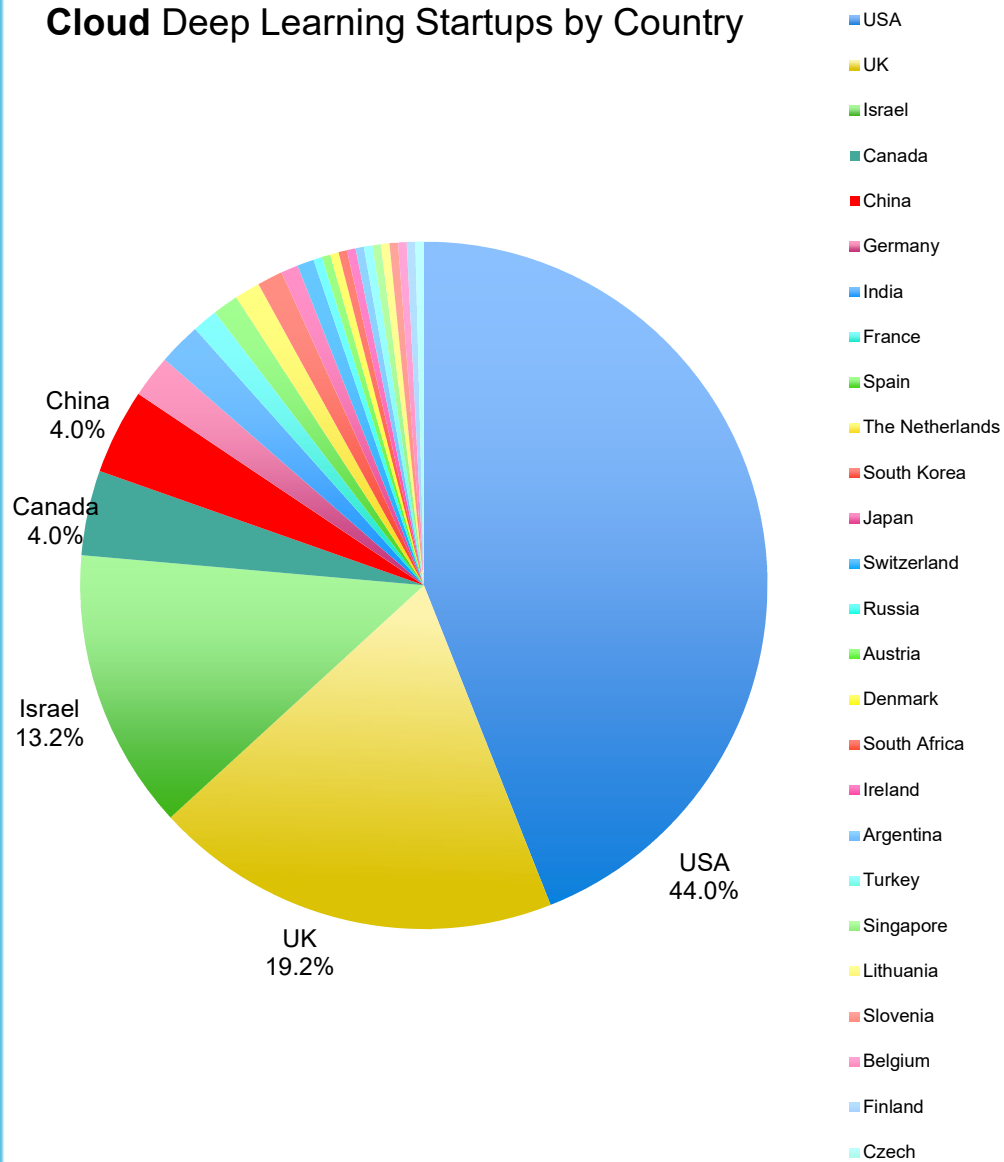
Where Are the Deep Learning Startups?

Deep Learning Startups by Country

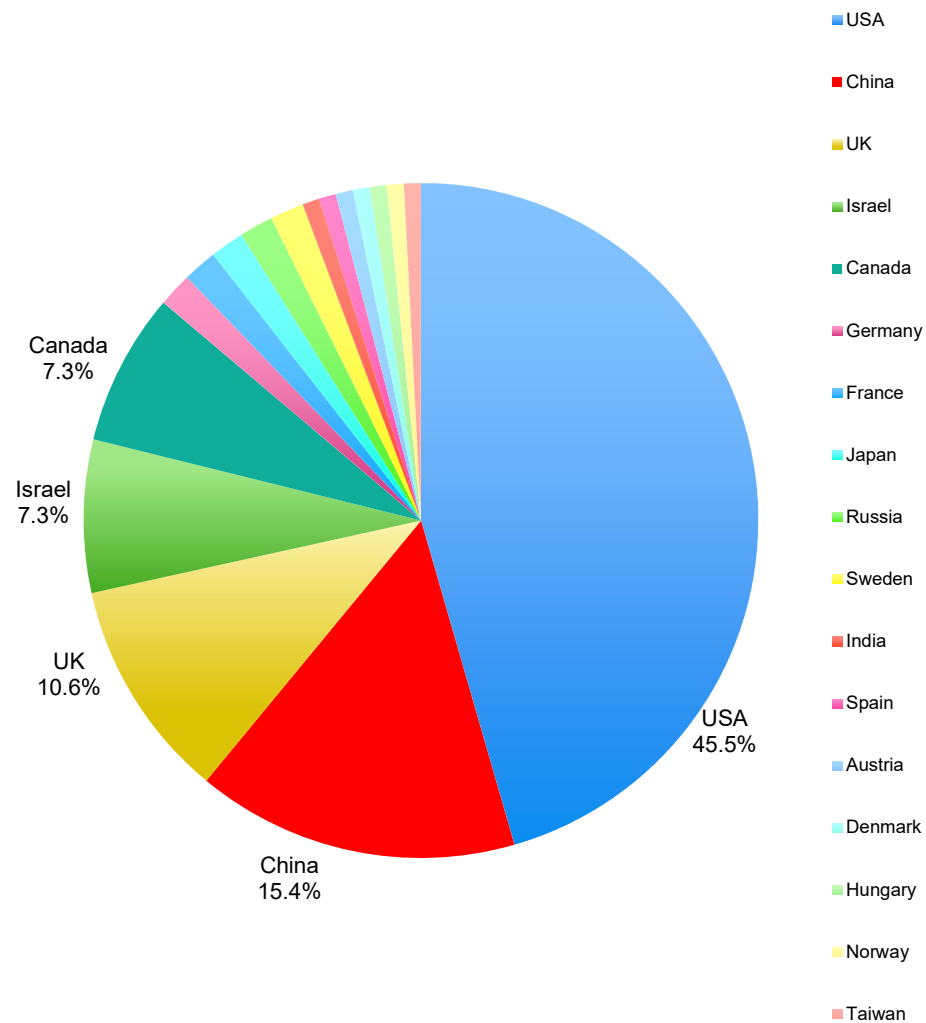


Understanding China's Startups

Cloud Deep Learning Startups by Country



Embedded Deep Learning Startups by Country



Deep Learning Startups

A window into the future of electronics



THE COGNITE 350

TOP STARTUPS IN DEEP LEARNING FEBRUARY 2018

AMERICAS
177

ASIA
39

EUROPE
MIDDLE EAST
AFRICA
133

DEEP LEARNING CLOUD COMPUTE/DATA PLATFORM AND SERVICES

DataRobot, H2O.ai, Cogital, vicarious, NanoNets, Digital Reasoning, OpenAI, minds.ai, ARIMO, groq inc., Numenta, vertex.ai, deepsense.io, Cirrascale, sentient, skymind, DATALOGUE, rapidminer, ELEMENT, KIMERA, loop, SIGOPT, naralogics, Dycorp, FLOYD, 4Paradigm, Preferred Networks, Arya.ai, hocrox, diffblue, Intellegens, AURORA-AI, masense, NEURANCE, seldon, twentybn, arbelnic, deepomatic, Deep, SPARKBEYOND, KUZNECH, RAJNBIRD

MEDICINE AND HEALTH CARE

imaga, AiCure, Atomwise, CureMatrix, ZENITH, Numerate, bd, deep genomics, DEEP 6 AI, RECURSION, babbiabs, SENTER, GRAIL, Butterfly Network, BENTRION, med what, pulseData, MEDANN, LUMINIST, BAYLABS, RECURSION, CLOUDMEX, Carestore, @enitic, vuno, iCarbonX, Lunit, VALON, ContextVision, KHERON, foodvisor, camerays, Intellegens, BenevolentAI, zebra, MARENTO, MyoMarch

FINANCE, INSURANCE AND COMMERCE

CEREBELLUM CAPITAL, Dataminr, KENSHC, pit.ai, Tickr, SI, ISENTIUM, zest, DEEP, alpha-i, FEATURE SPACE, BMLL, seven, T, DEEP

MARKETING AND ADVERTISING

Affectiva, XIX, VUE.AI, ISENTIUM, Ampero, AUTOMAT, PERSADO, PRODUCTAI, SOTERIA, ditto, tamr, LAYER 6, netra, PRODUCTAI, COGNITIV, AYLIEN, NUDGR, visii, viisights, Re/uze, CORTEXICA, WebyClip, real eyes

VISION AS A SERVICE

clarifai, CrowdAI, CVCRES, MOD9, Matroid, PRISMA, Sighthound, tend, Hubino, HYPERVERGE, SeatTech, TUPTTECH, YITU, senseniric, Birds.ai, deepomatic, inoven, VILYNX, TRACTABLE, VALOSSA, RE SNAP, cortica

AERIAL IMAGING AND MAPPING

Orion Insights, FEATURE, artofaware, VIDEO, HEMLOUPE

SOCIAL MEDIA, ENTERTAINMENT AND LIFESTYLE

SECOND SPECTRUM, SOTERIA, OLLY, JukeDeck, Neural Painting, VAULT, Sensifai, Sonalytic Systems, PIXONEYE

AUDIO AND LANGUAGE AS A SERVICE

AI Sense, babbiabs, CAPIO, CTI, gridshitt, indico, KITTAI, LEXALYTICS, MindMeld, MOD9, PAT, semanticmachines, volley, aptricity, LUMINOSO, AISPECH, Hubino, IntelliGo, MIND, Bloomsbury AI, COSMICOR, cortical.io, VChecker, Intelligent Voice, LAST MILE, LEVERTON, aido, Replika, retechnica, spaCy, speech, They Say, ido, Linguamatics, audioburst, BEYONVERBAL

HUMAN-MACHINE INTERFACE

neuroAI, Leron, LIGHTHOUSE, SKINDROID, TERADEEP, gyrfalcon technology, ZEROAI, IMOTIONS, mashgin, MOD9, KITTAI, Birds.ai, keenresearch, trueface.ai, EVERS, SoundHound Inc., Sighthound, ANI, AI BRAIN INC, babbiabs, EMOTIBOT, MBRPX, Mobvoi, AKA, ROKIC, Face, AISPEECH, VEO, intuition robotics, LIPSIGHT, emteq, eyeSight, IMAGRY, JUNGO, ORCAM, snips

IT OPERATIONS AND SECURITY

CYLANCE, graphistry, sparkcognition, deepinstant, Cognitive ID, Cyberlytic, DARKTRACE

DRONES AND ROBOTS

abundant robotics, CLEARPATH, ACCELERATED DYNAMICS, LILLY, Iris Automation, UNIVERSAL, Qelzal, SIOYDIO, OSARO, brain, AISPEECH, KINDRED, HOVER CAMERA, DLR, robar, cortex

CRM AND HR

BICYCLE AI, Digital Genius, vufind, Eloquent Labs, MARAX AI, SentiSum, CYRA, AYLIEN, re-infer

SILICON PLATFORMS

REM, thindl, groq inc., Wauve Computing, TENSTORRENT, gyrfalcon technology, brainchip, DEEP VISION, NOVUMIND, MYTHIC, Cambricon, DEEPHI, Intellegens, Graphcore, AMOTIVE

SECURITY, SURVEILLANCE, MONITORING

Deep Vision, DEEP SENTINEL, SHIELD AI, LIGHTHOUSE, NOVUMIND, gyrfalcon technology, cognac, Sighthound, Igocion, PILOT AI, athelas, SENTENAI, camio, trueface.ai, AiCure, SECOND SPECTRUM, algalux, 商汤, Face, DEEPHI, YITU, DEEPLINT, IntelliGo, intel fusion, Horizon Robotics, OpenCapacity, VISIOGENIE, KONUX, 3d eye, Calipsa, aichero, (xihelm), HOXTON, GETALERT, LIPSIGHT, FIFTH, PointGrab, amVISION, VIDEO

AUTONOMOUS VEHICLES

QUANERGY, algalux, drive.ai, aiPlusAI, PITSTOP, AutoX, nauti, NURO, Velodyne LIDAR, ZOX, ARGO AI, nuTonomy, DEEPCSCALE, AURO, simple, AISPEECH, 商汤, MINIEYE, LEAPMIND, netradyne, BLUE VISION, machines with vision, AMOTIVE, RoboCV, Mapillary, OXBOTICA, idrive, Brodmann, FIVE AI, coprta, nexar, MORPHEUS LABS, SafeCue

OPERATIONS, LOGISTICS AND MANUFACTURING

Citrine, dearmat, Descartes Labs, DroneDeploy, SIGHT MACHINE, INVENIA, MAANA, Intellegens, micropsi industries, Terrabotics, AVURY, PRESENSEN, trax, 3DSignals



s p e a k y o u r m i n d