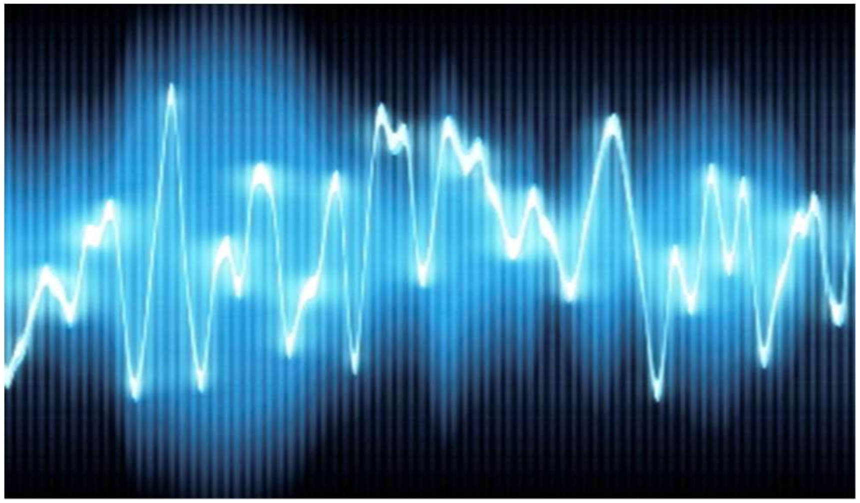
Deep Learning Revolution: From Theory to Impact



Chris Rowen

CEO
Babblelabs Inc.

EDPS

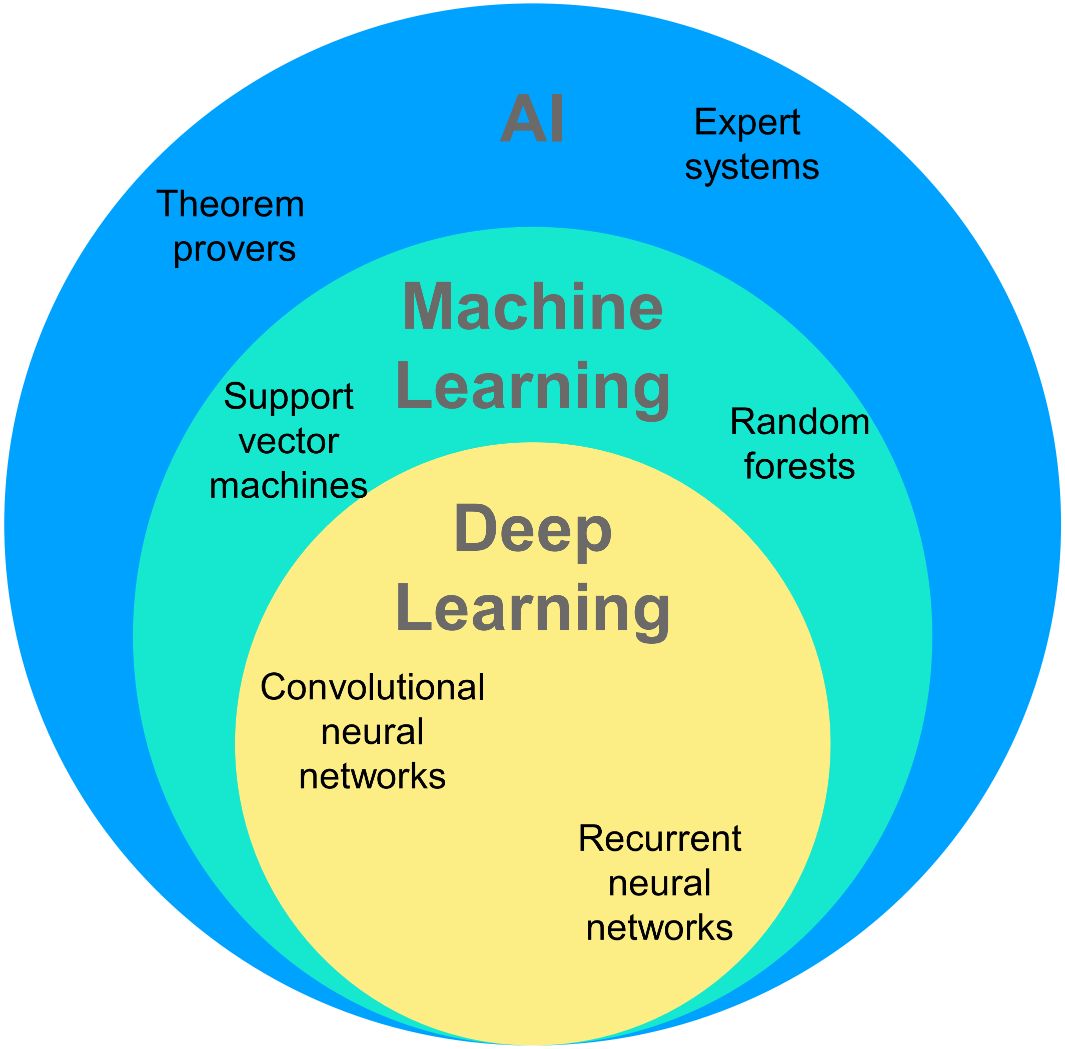
September 2018

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

babblelabs

Quick Taxonomy



AI

Expert systems

Theorem provers

Machine Learning

Support vector machines

Random forests

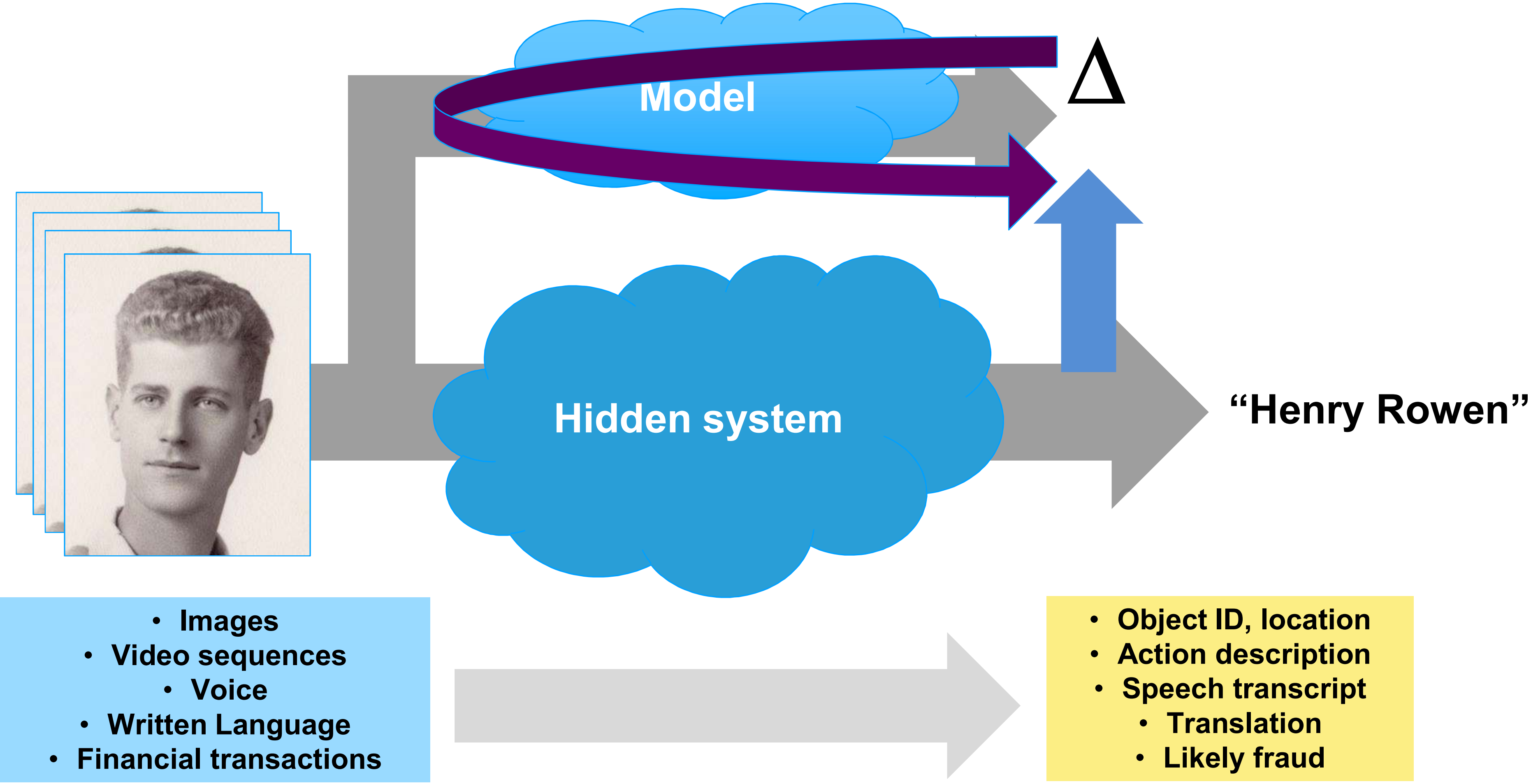
Deep Learning

Convolutional neural networks

Recurrent neural networks

Deep Learning Foundations

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



- Images
- Video sequences
- Voice
- Written Language
- Financial transactions

- Object ID, location
- Action description
- Speech transcript
- Translation
- Likely fraud

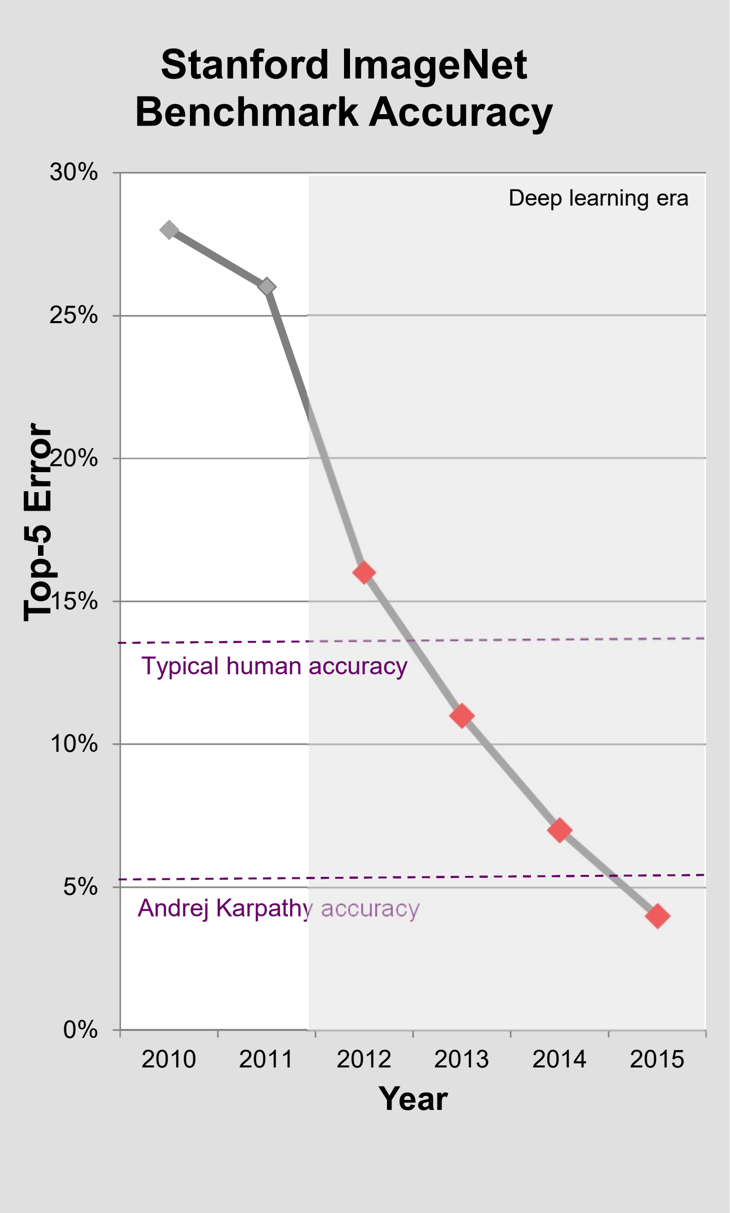
Vision is Fundamentally Hard

- 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

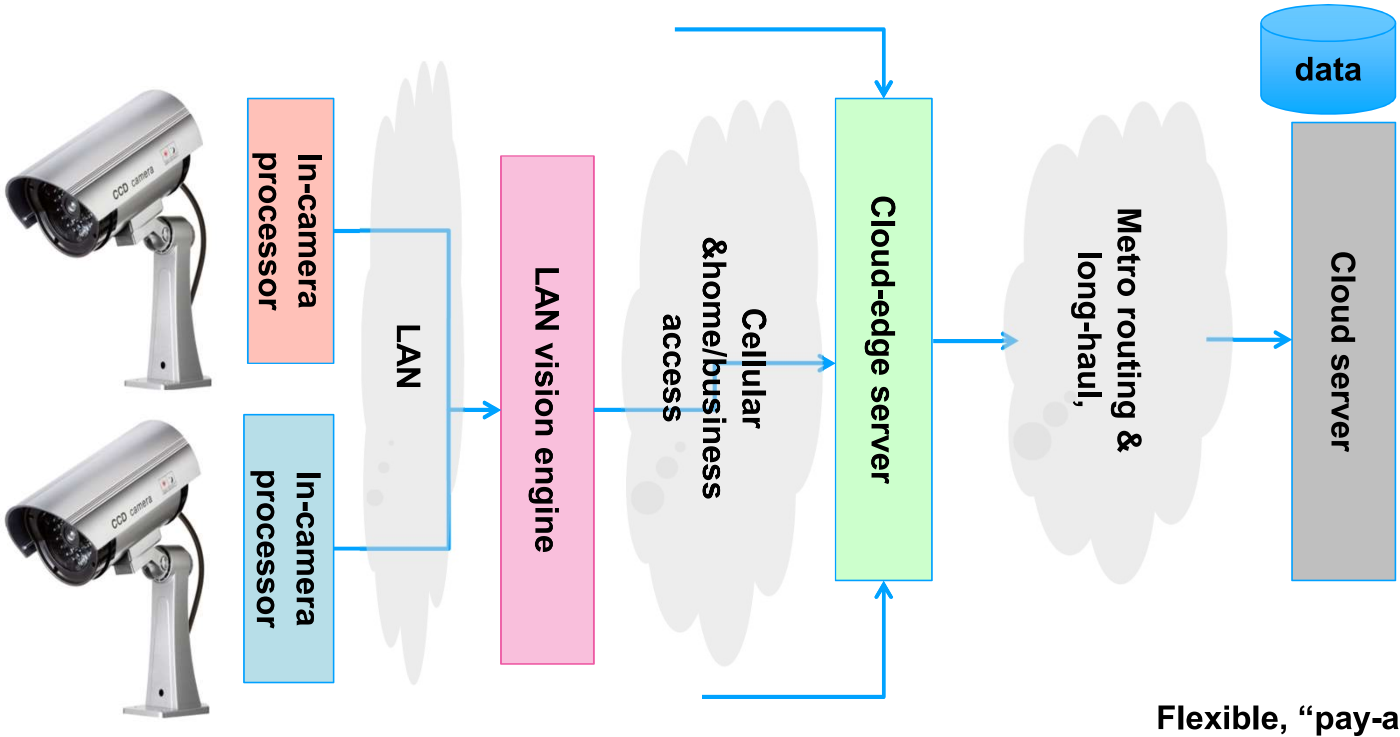
Shih-Tzu

Norwegian elkhound



babblelabs

Where will we put the “smarts”?



data

In-camera processor

In-camera processor

LAN

LAN vision engine

Cellular &home/business access

Cloud-edge server

Metro routing & long-haul,

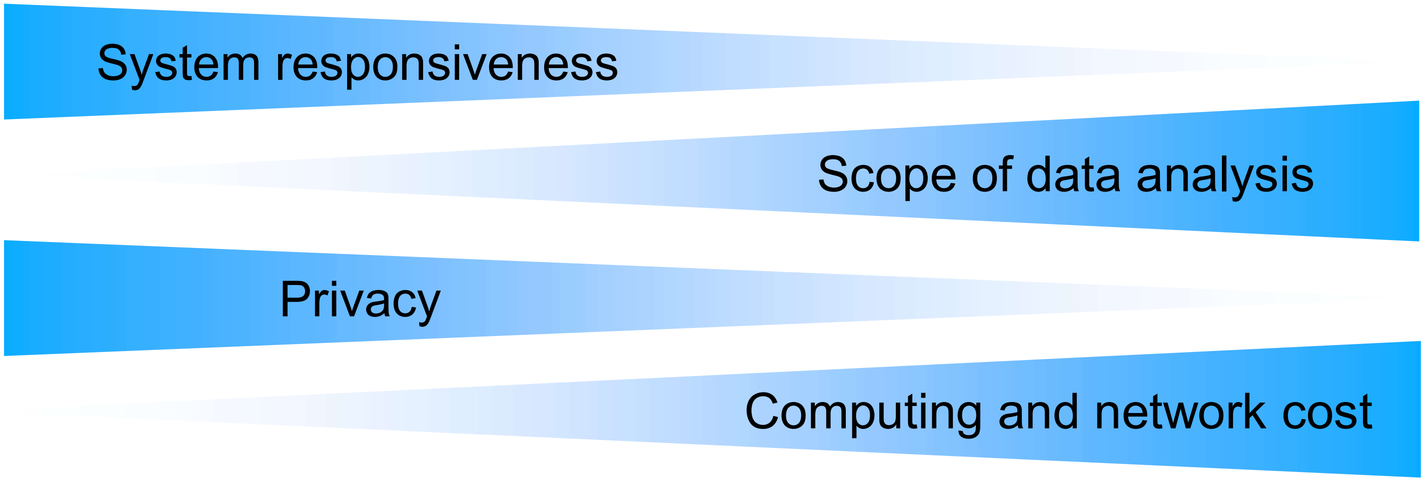
Cloud server

compression + analytics – cheap/efficient

Fusion and analytics

Will this happen?

Flexible, “pay-as-you-go” compute



System responsiveness

Scope of data analysis

Privacy

Computing and network cost

Security, Robustness and Privacy

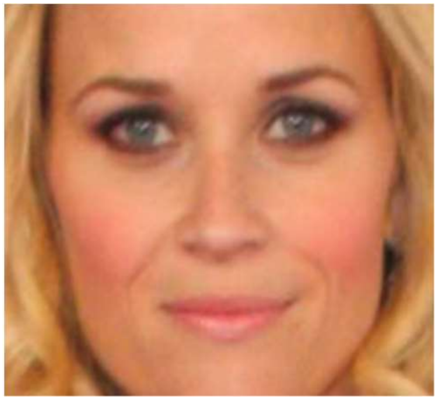
- More risks for privacy: more cameras, more correlation to other streams
- Mission-critical surveillance susceptible to new attacks

The usual network and device attacks

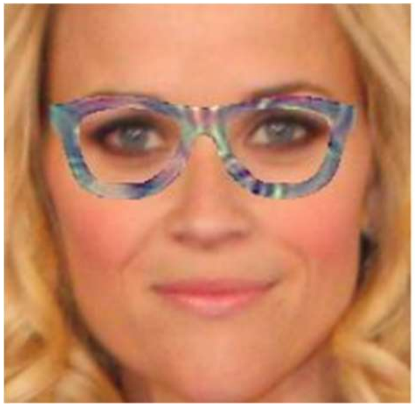
+ Database manipulation to inject bias

+ Classification spoofing

- Example: Spoofing facial recognition



Reese Witherspoon



Reese Witherspoon in patterned frames

Recognized as Russell Crowe

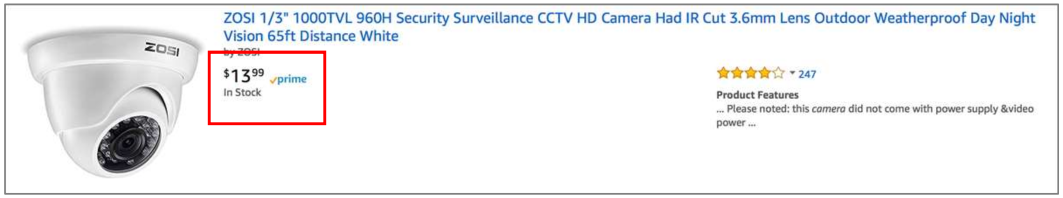
Sharif et al: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition [CMU]

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:

  $2\text{x}10^{10}$ sensors * $5\text{x}10^{8}$ pixels/sec = 10^{19} raw pixels/s



ZOSI 1/3" 1000TVL 960H Security Surveillance CCTV HD Camera Had IR Cut 3.6mm Lens Outdoor Weatherproof Day Night Vision 65ft Distance White

by ZOSI

$13.99 ✓prime

In Stock

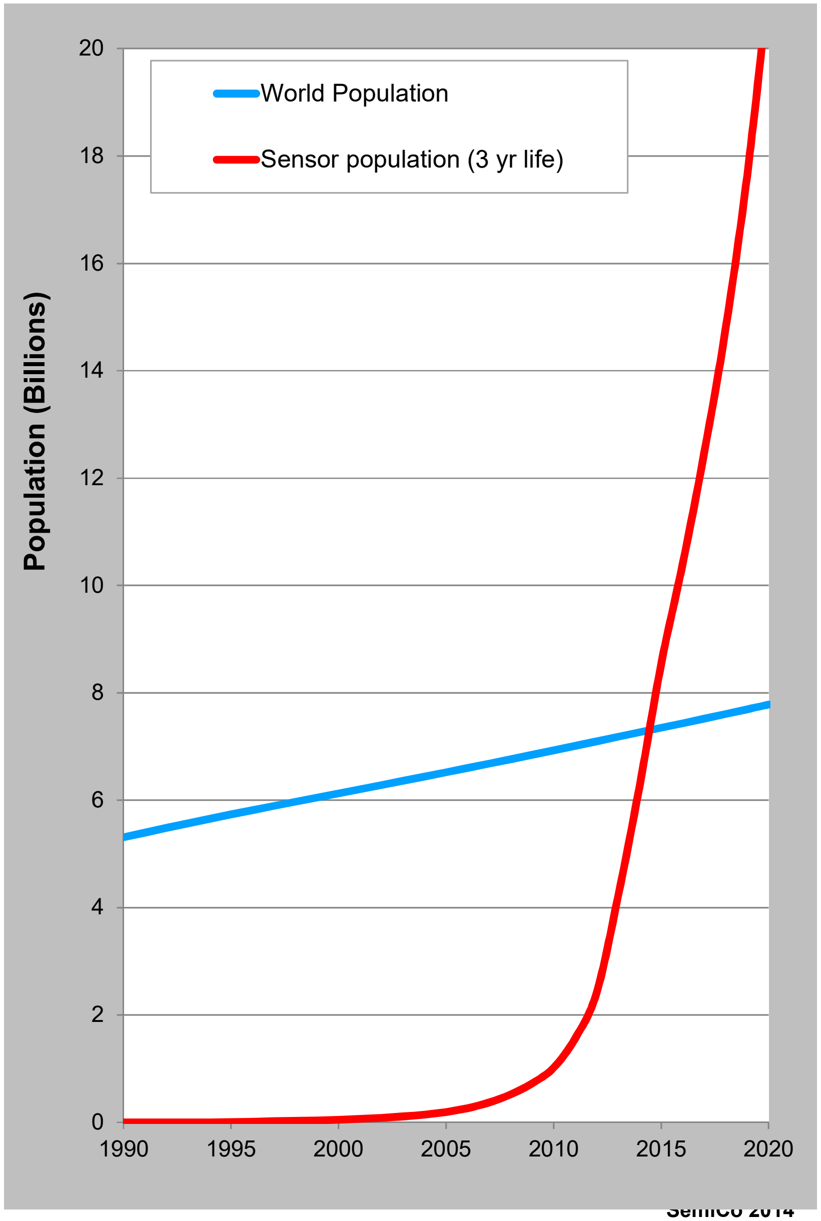
★★★★☆ 247

Product Features

... Please noted: this *camera* did not come with power supply &video power ...

$13.99 ✓prime

In Stock



SemiCo 2014

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

babblelabs

Voice is Vision

Speech is the most human of interfaces

5B active electronic speech users

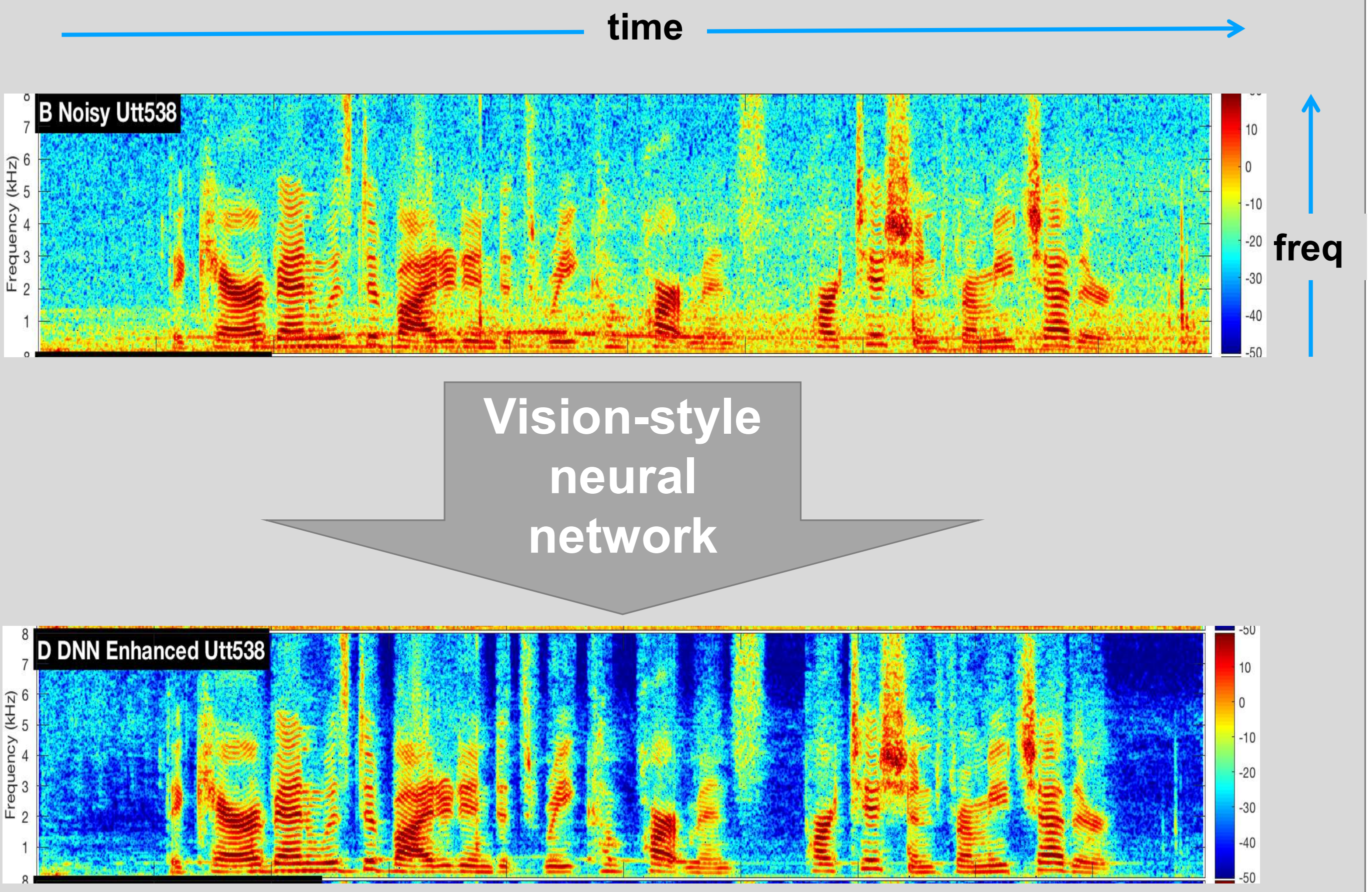
>20B microphones installed by 2020

Key services:

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

Shift to local compute

Transform speech signal into image: “spectrogram”



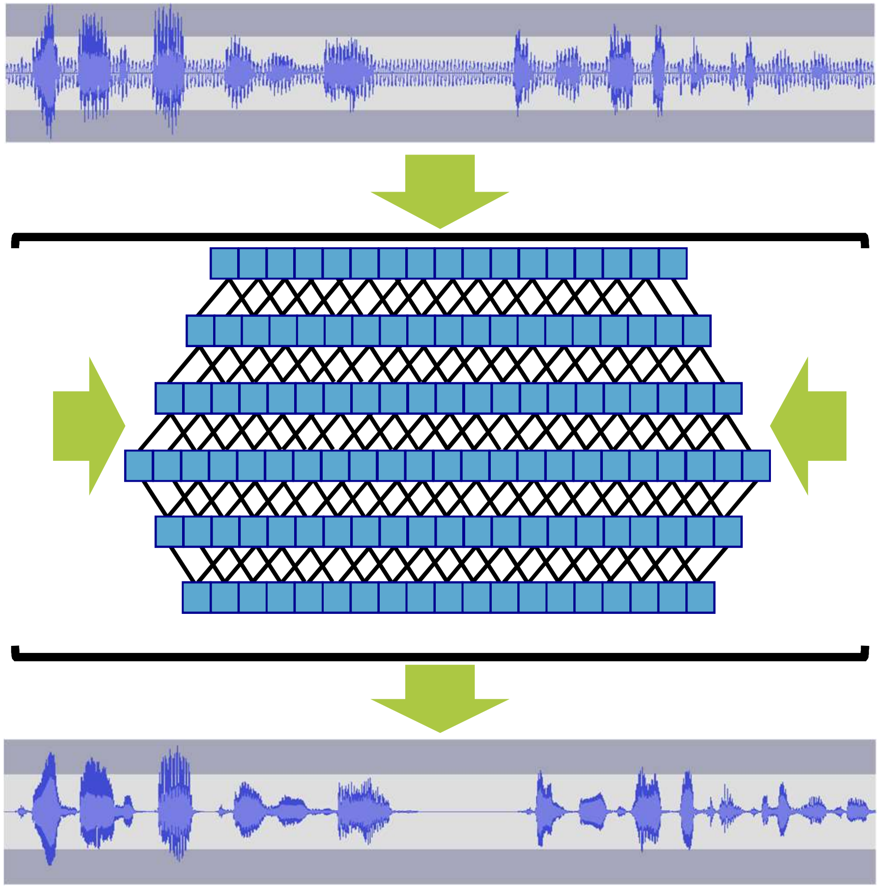
time

freq

Vision-style neural network

BabbleLabs: Deep learning meets speech

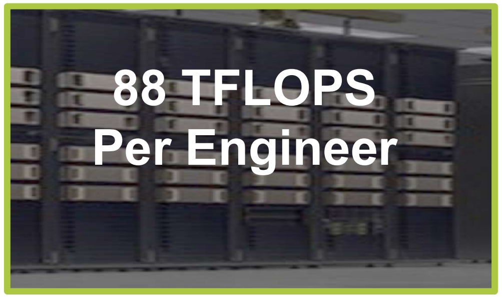
fresh problems, more sophisticated models, more data, more training



0dB

21dB

Massive Compute



88 TFLOPS Per Engineer

Massive Data Corpus

30,000 hr speech
7,000 hr music
3,000 hr noise
10,000 room models

Speech Enhancement

Speaker-specific Services

Speech UI Dialog

babblelabs

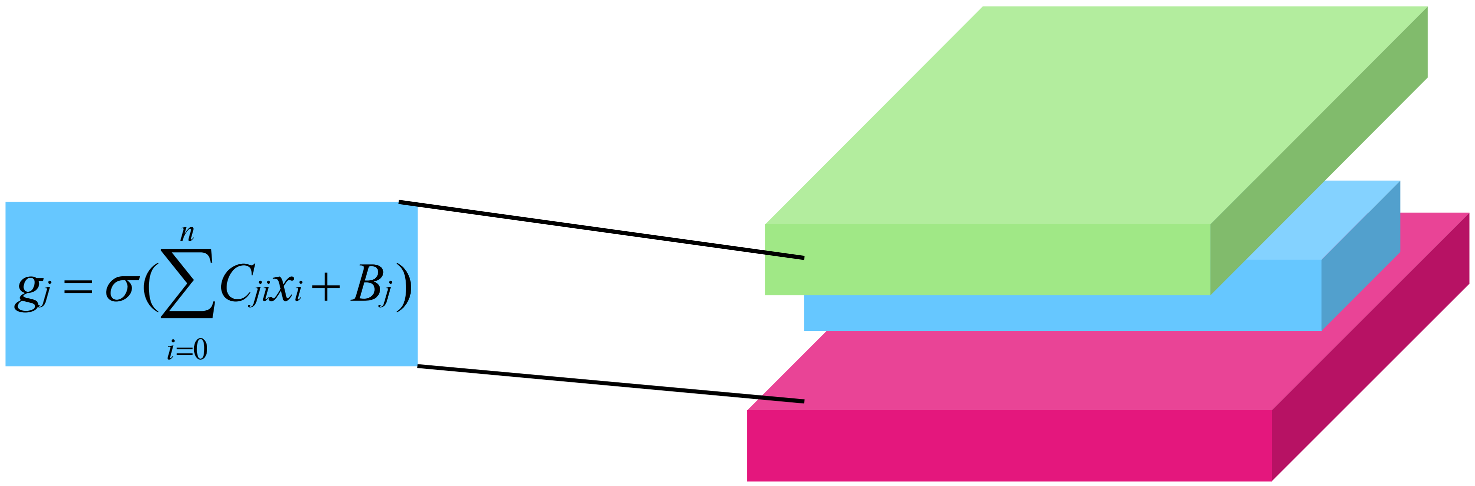
Speech Enhancement Example

Original Video

Enhanced Video

babblelabs

Deep Learning Silicon Is Easy
especially for inference



$$g_j = \sigma(\sum_{i=0}^{n} C_{ji} x_i + B_j)$$

- 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

babblelabs

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

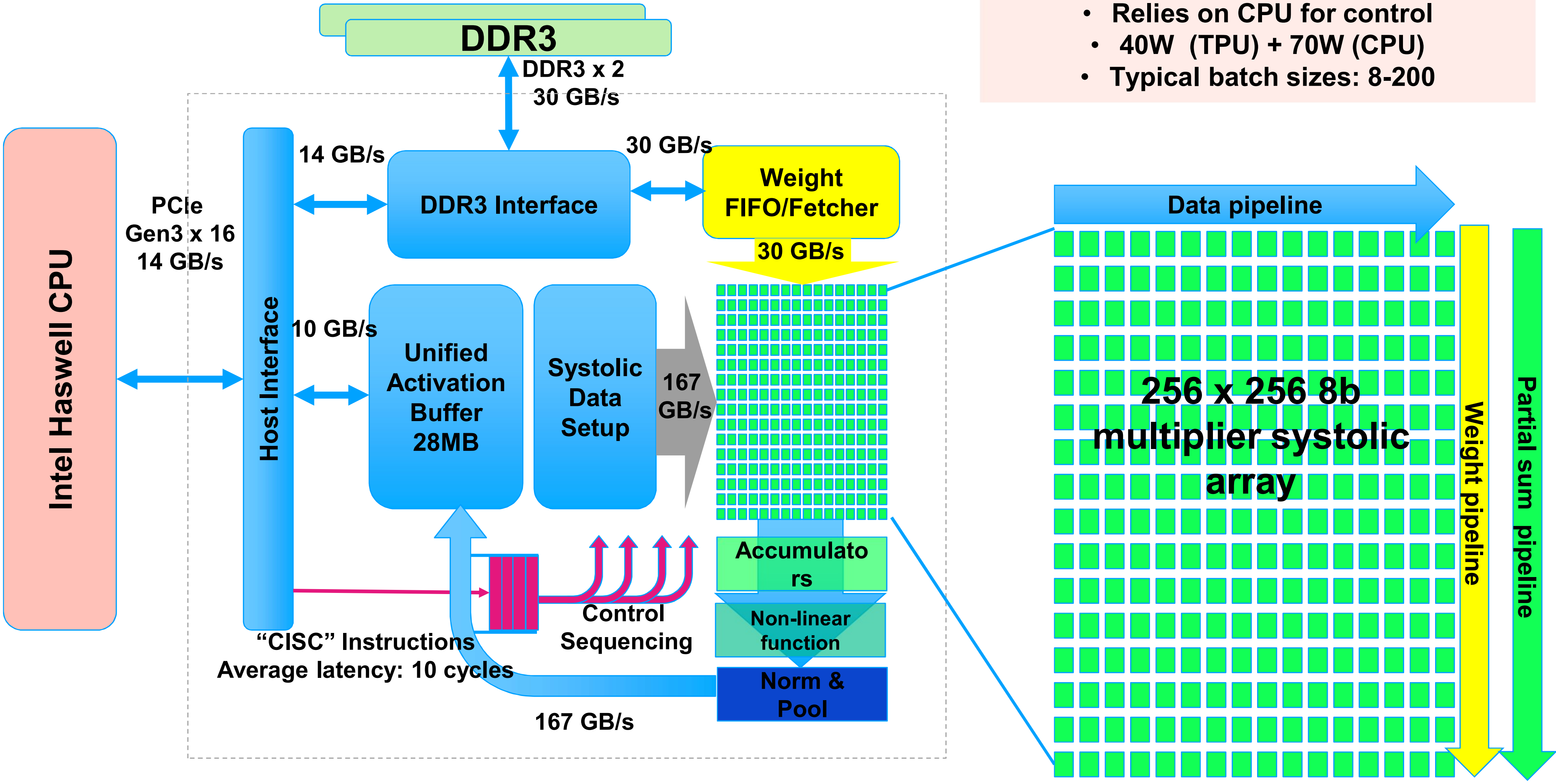
babblelabs

Anatomy of a Neural Network Accelerator

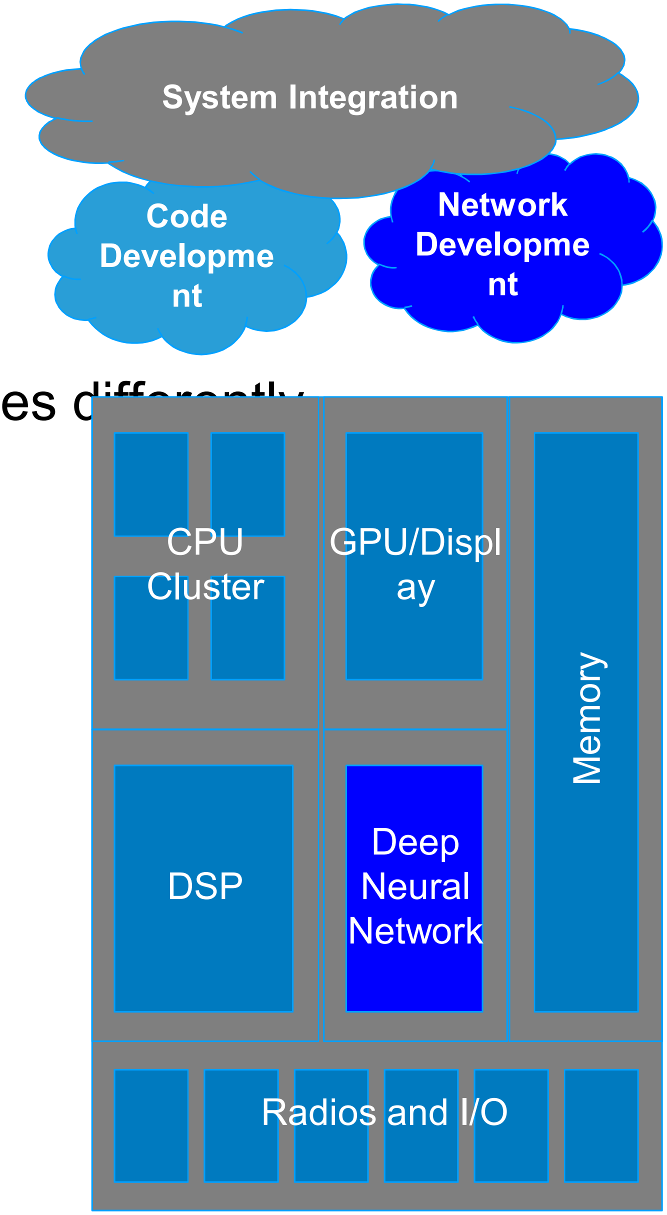
Example: Google Tensor Processing Unit (TPU)

Neural Network Systolic Coprocessor:

- Pumps slices of matrix multiply through array
- Relies on CPU for control
- 40W (TPU) + 70W (CPU)
- Typical batch sizes: 8-200



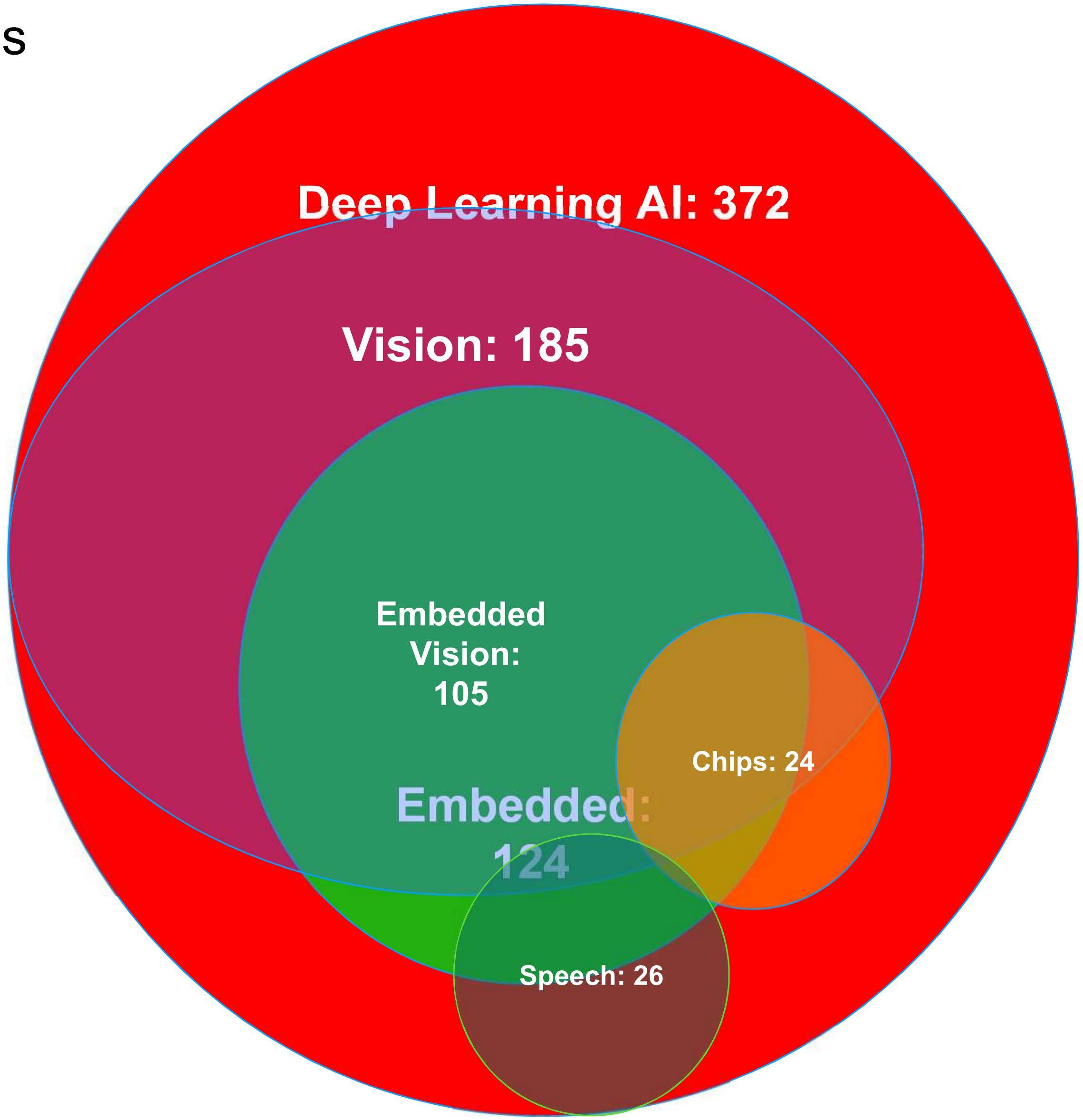
Implications for Semiconductors



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 µcontrollers
   - 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

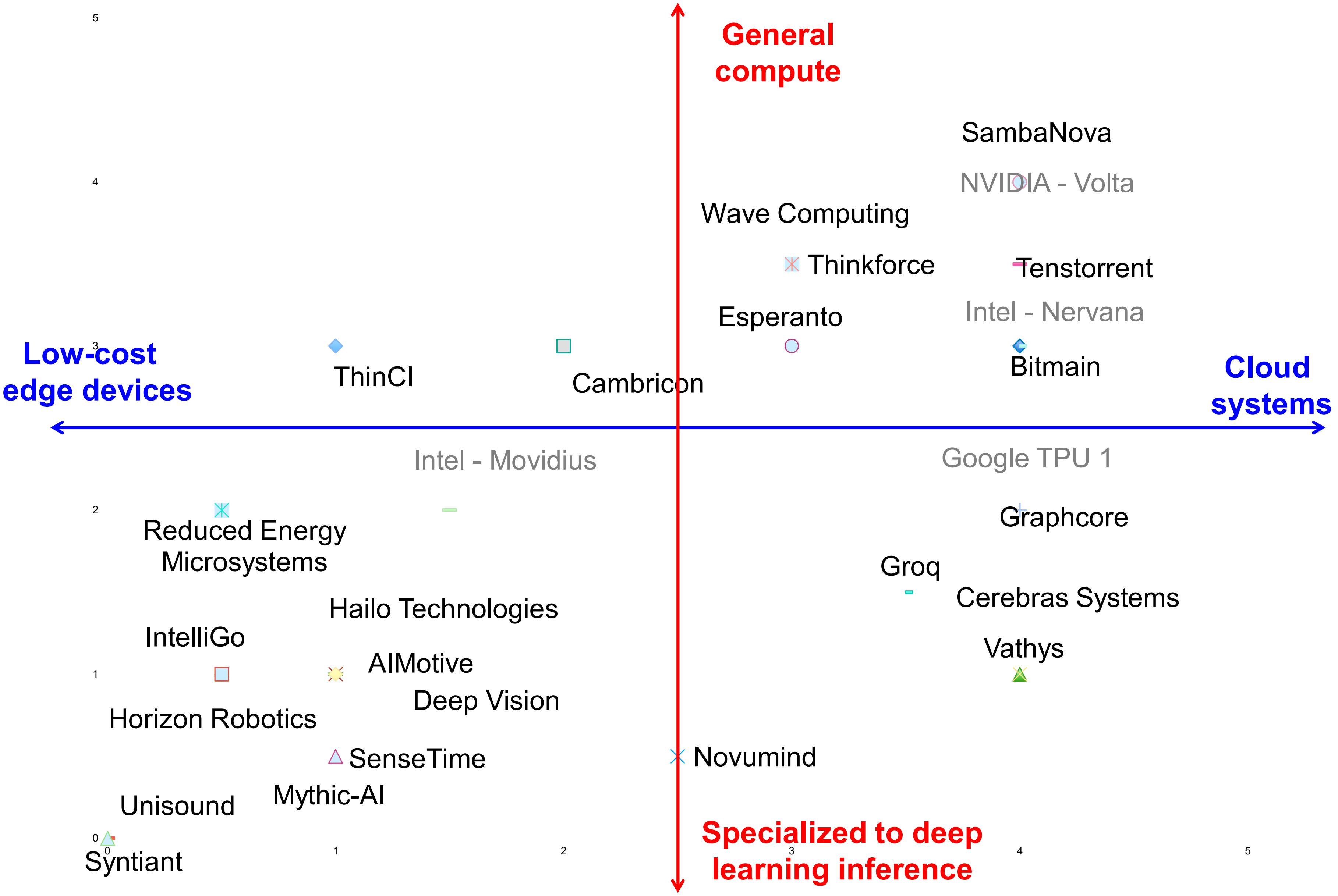
➔ In cloud

- 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
Gyrfalcon 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

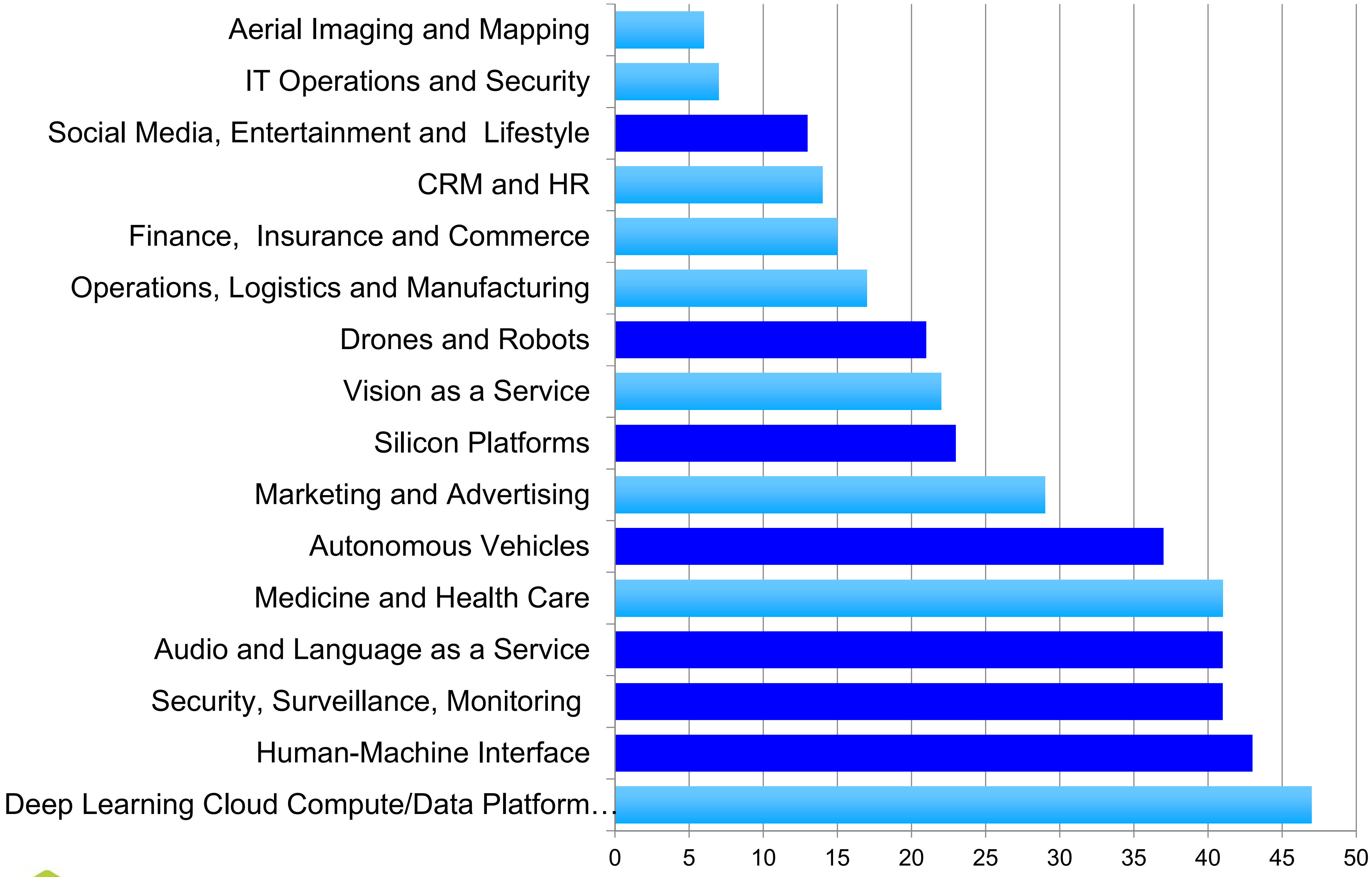
babblelabs

Sorting Out Deep Learning Silicon



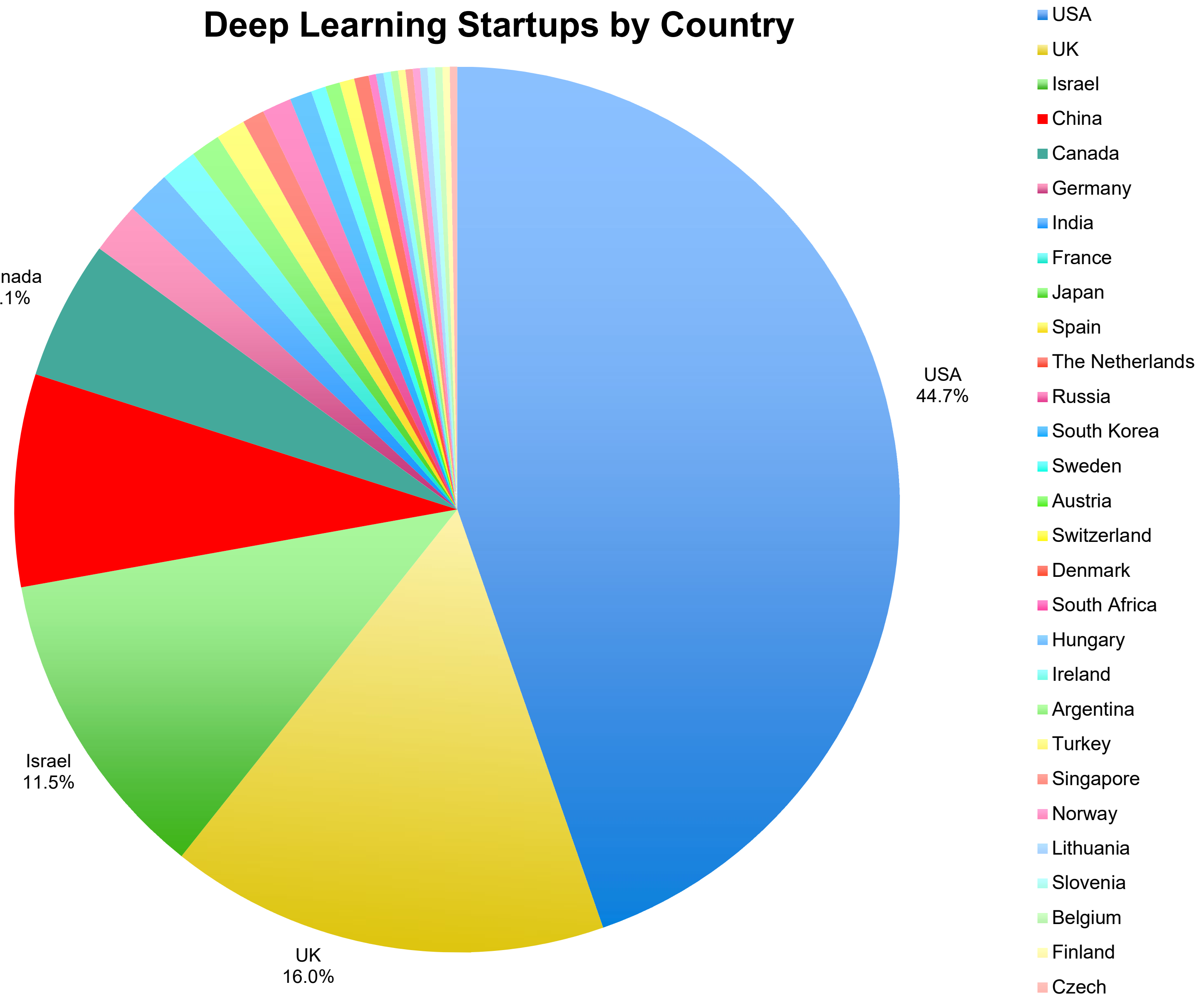
babblelabs

Segments for Deep Learning Startups



babblelabs

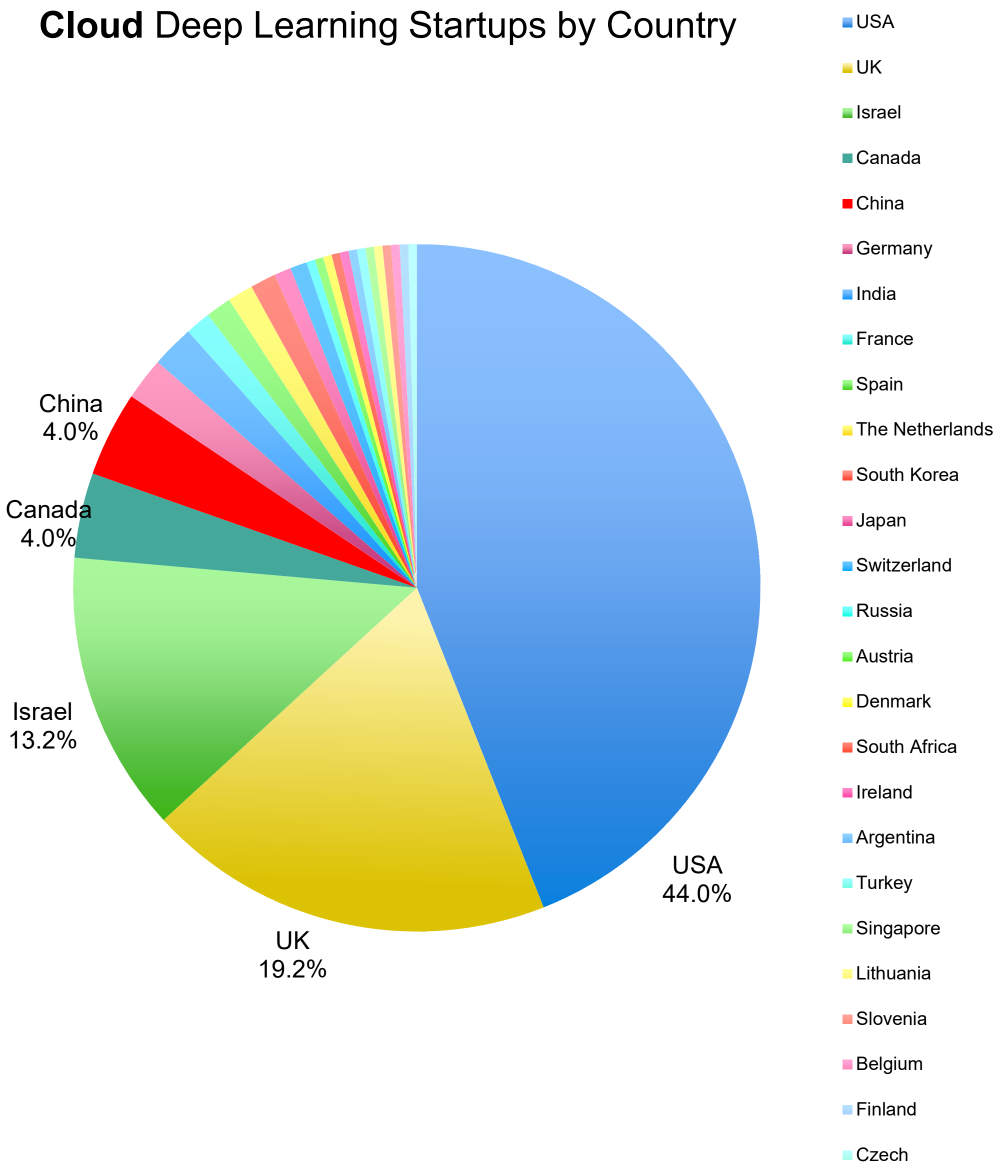
Where Are the Deep Learning Startups?



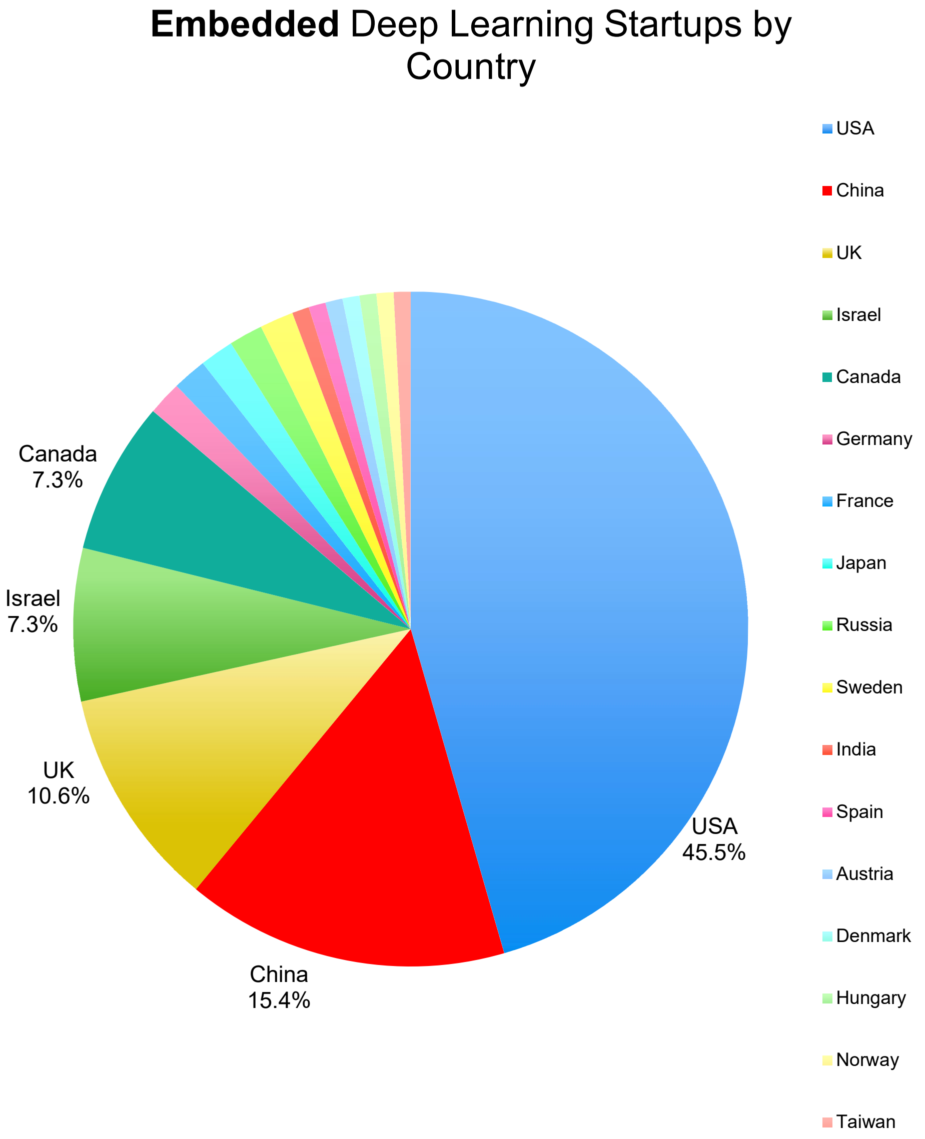
babblelabs

Cognite Ventures

Understanding China’s Startups



Cloud Deep Learning Startups by Country

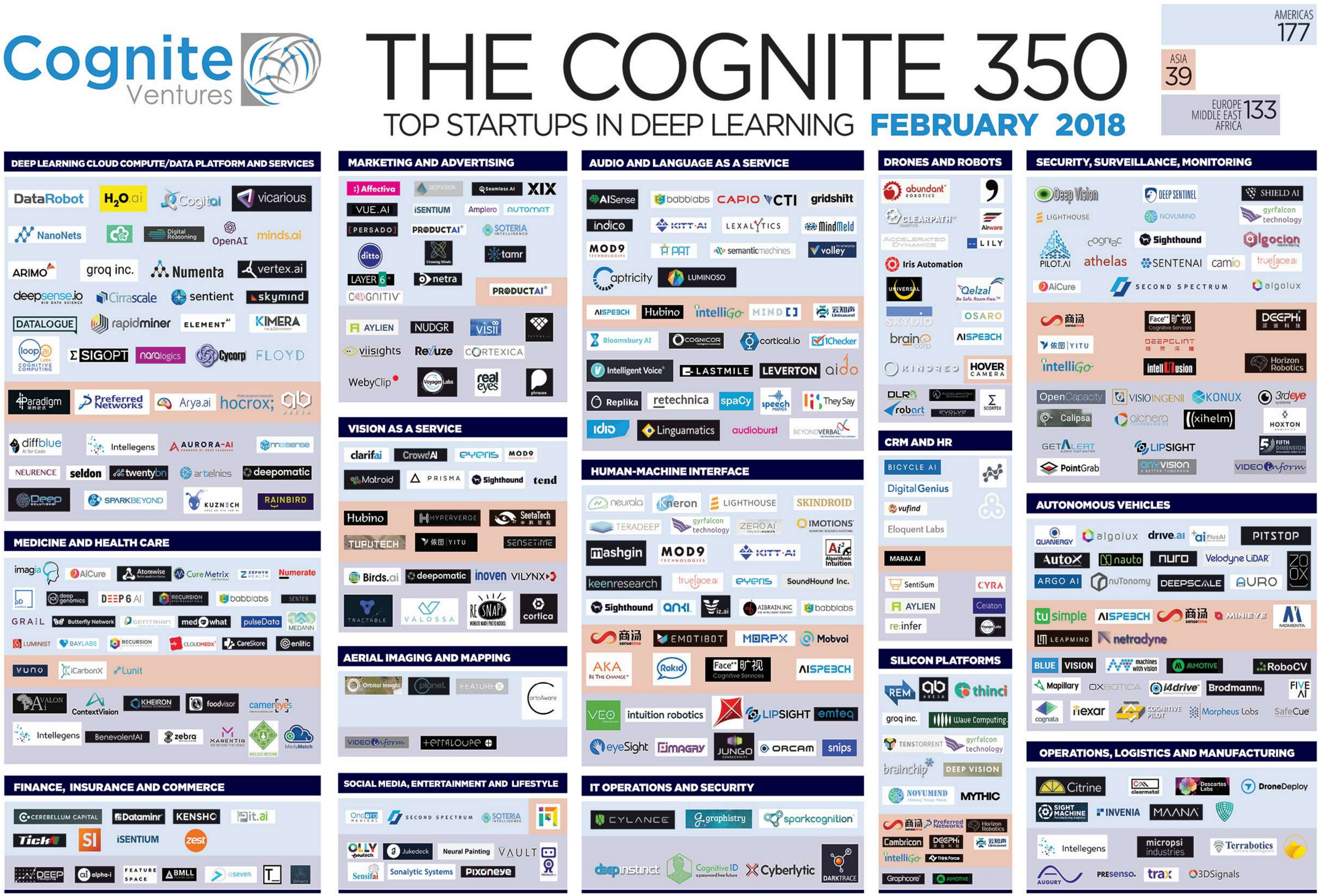


Embedded Deep Learning Startups by Country

babblelabs

Deep Learning Startups

A window into the future of electronics



speak your mind