Introduction to Deep Learning

Rick Stevens

Argonne National Laboratory

The University of Chicago

Crescat scientia; vita excolatur

Joint Design of Advanced Computing Solutions for Cancer
Cross Laboratory Team ANL, ORNL, LANL, LLNL, NCI

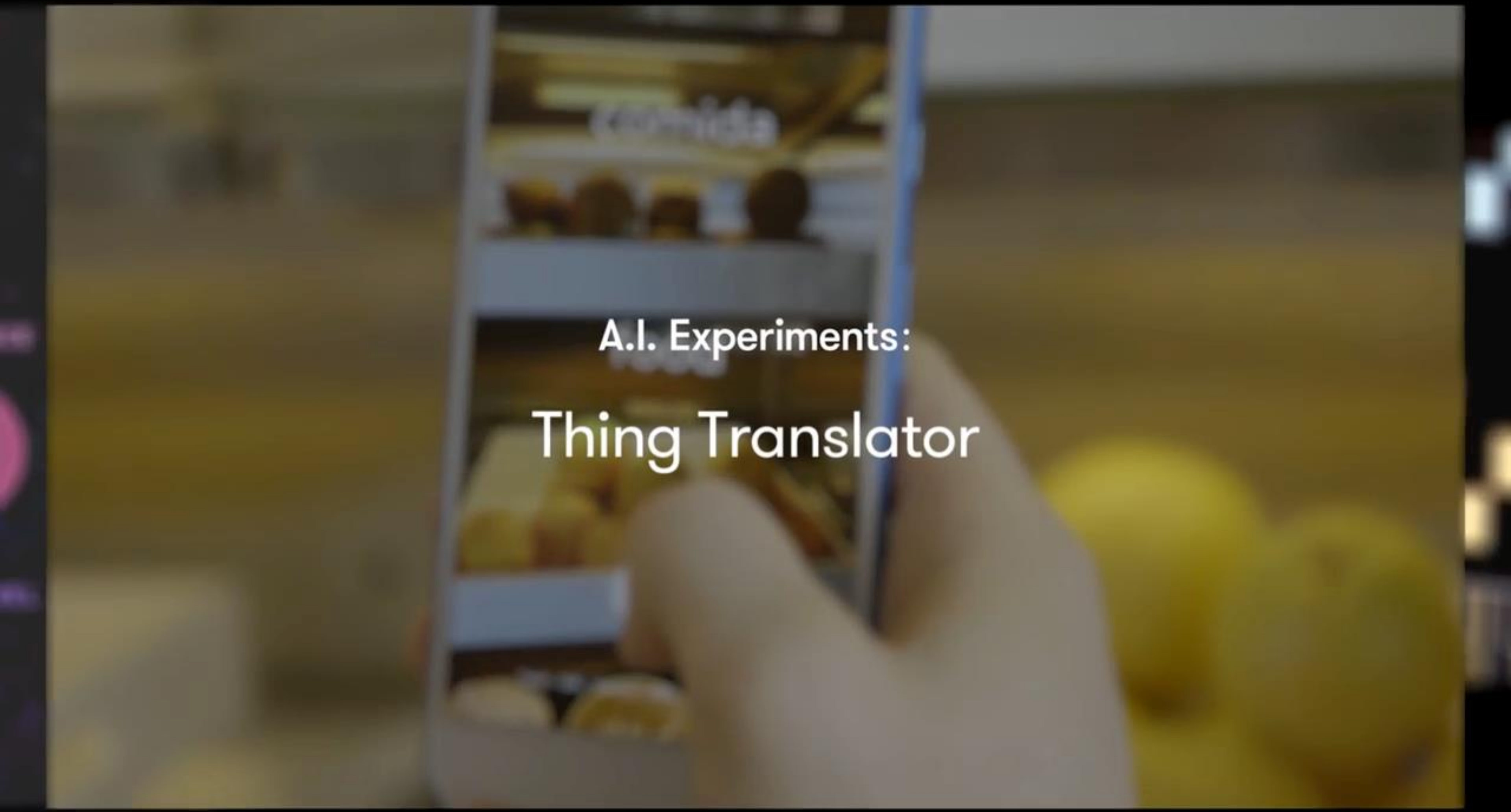
- ANL: Rick Stevens, Tom Brettin, Jim Davis, Emily Dietrich, Hal Finkel, Ian Foster, Monisha Ghosh, Daniel Gonzalez, Ushma Kriplani, Ravi Madduri, Sergei Maslov, Bob Olson, Dan Olson, Mike Papka, Lorenzo Pesce, Justin Wozniak, Prasanna Balaprakash, John Santerre, Maulik Shukla, Venkat Vishwanath, Fangfang Xia
- LANL: Marian Anghel, Frank Alexander, Tanmoy Bhattacharya, Judith Cohn, Paul Dotson, Kumkum Ganguly, Jason Gans, Cristina Garcia-Cardona, Geralyn Hemphill, Nick Hengartner, William Hlavacek, Patrick Kelly, Amy Larson, Ben Mcmahon, Will Fischer
- LLNL: Jonathan Allen, Ya Ju Fan, Marisa Torres, Adam Zemala
- ORNL: Mike Lueze, Barney Maccabe, Intawat Nookaew, Arvind Ramanathan
- NCI: James H. Doroshow, Susan Holbeck, Eric Stalhberg, Yvonne A. Evrard, George Zaki
- UIUC: Sergei Maslov
- UChicago: Monisha Ghosh

Pilot 1

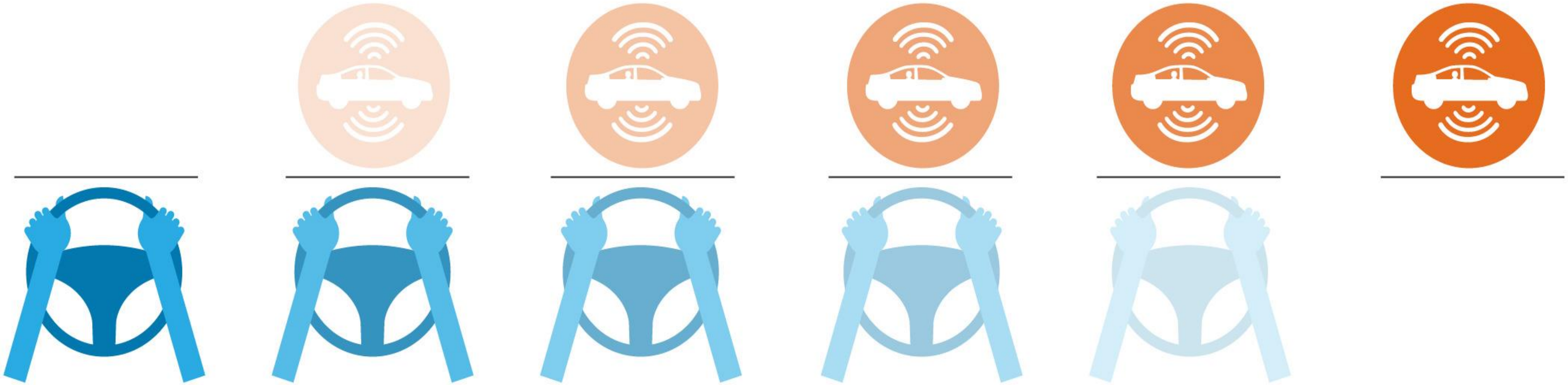
CANDLE

Spark

Thing Translator



Five Levels of Vehicle Autonomy



Level 0

No automation: the driver is in complete control of the vehicle at all times.

Level 1

Driver assistance: the vehicle can assist the driver or take control of either the vehicle's speed, through cruise control, or its lane position, through lane guidance.

Level 2

Occasional self-driving: the vehicle can take control of both the vehicle's speed and lane position in some situations, for example on limited-access freeways.

Level 3

Limited self-driving: the vehicle is in full control in some situations, monitors the road and traffic, and will inform the driver when he or she must take control.

Level 4

Full self-driving under certain conditions: the vehicle is in full control for the entire trip in these conditions, such as urban ride-sharing.

Level 5

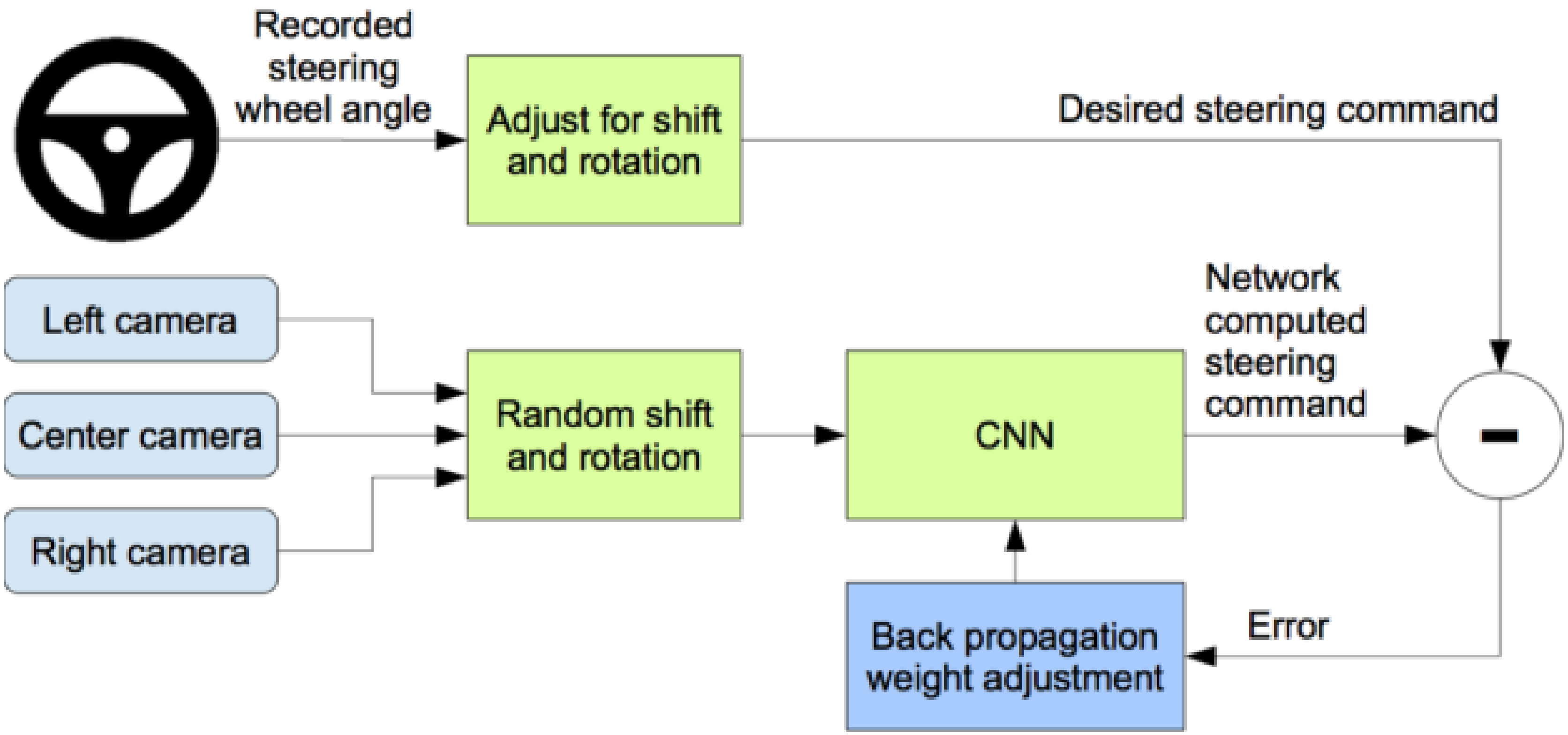
Full self-driving under all conditions: the vehicle can operate without a human driver or occupants.

Source: SAE & NHTSA

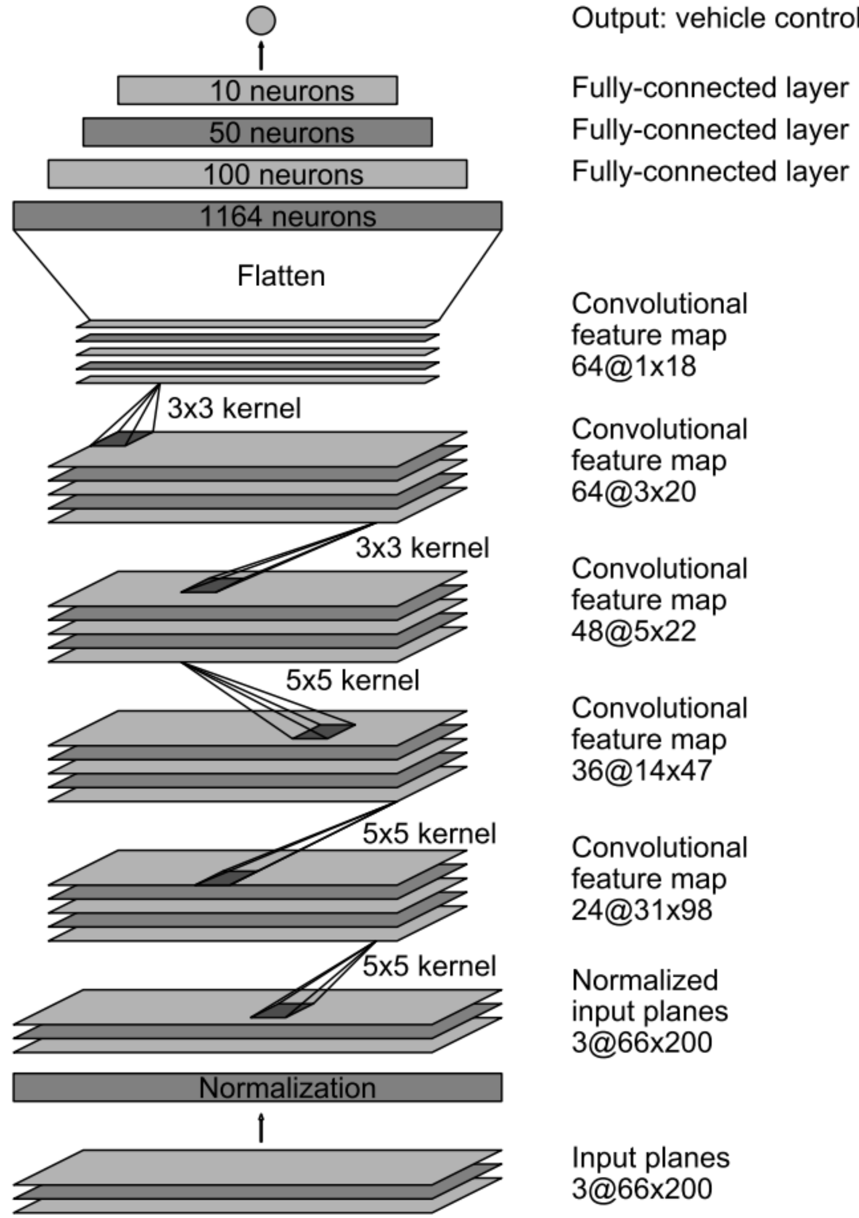
Tesla

THE PERSON IN THE DRIVER'S SEAT
IS ONLY THERE FOR LEGAL REASONS.

HE IS NOT DOING ANYTHING.
THE CAR IS DRIVING ITSELF.

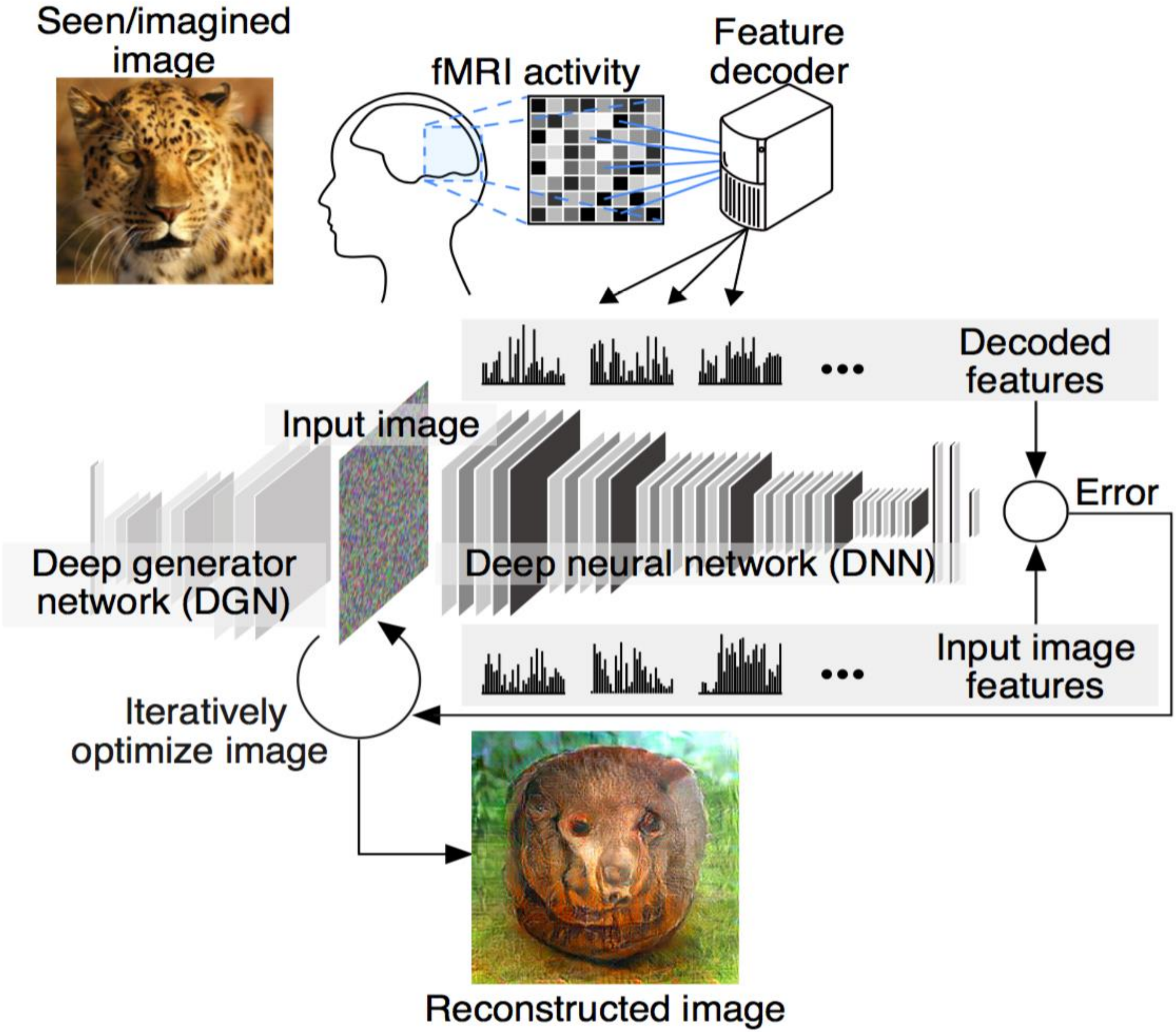


https://devblogs.nvidia.com/parallelforall/deep-learning-self-driving-cars/



https://devblogs.nvidia.com/parallelforall/deep-learning-self-driving-cars/


062M486
MANUFACTURER


Seen/imagined image
fMRI activity
Feature decoder
Decoded features
Input image
Error
Deep generator network (DGN)
Deep neural network (DNN)
Input image features
Iteratively optimize image
Reconstructed image

Progress on Reading Minds

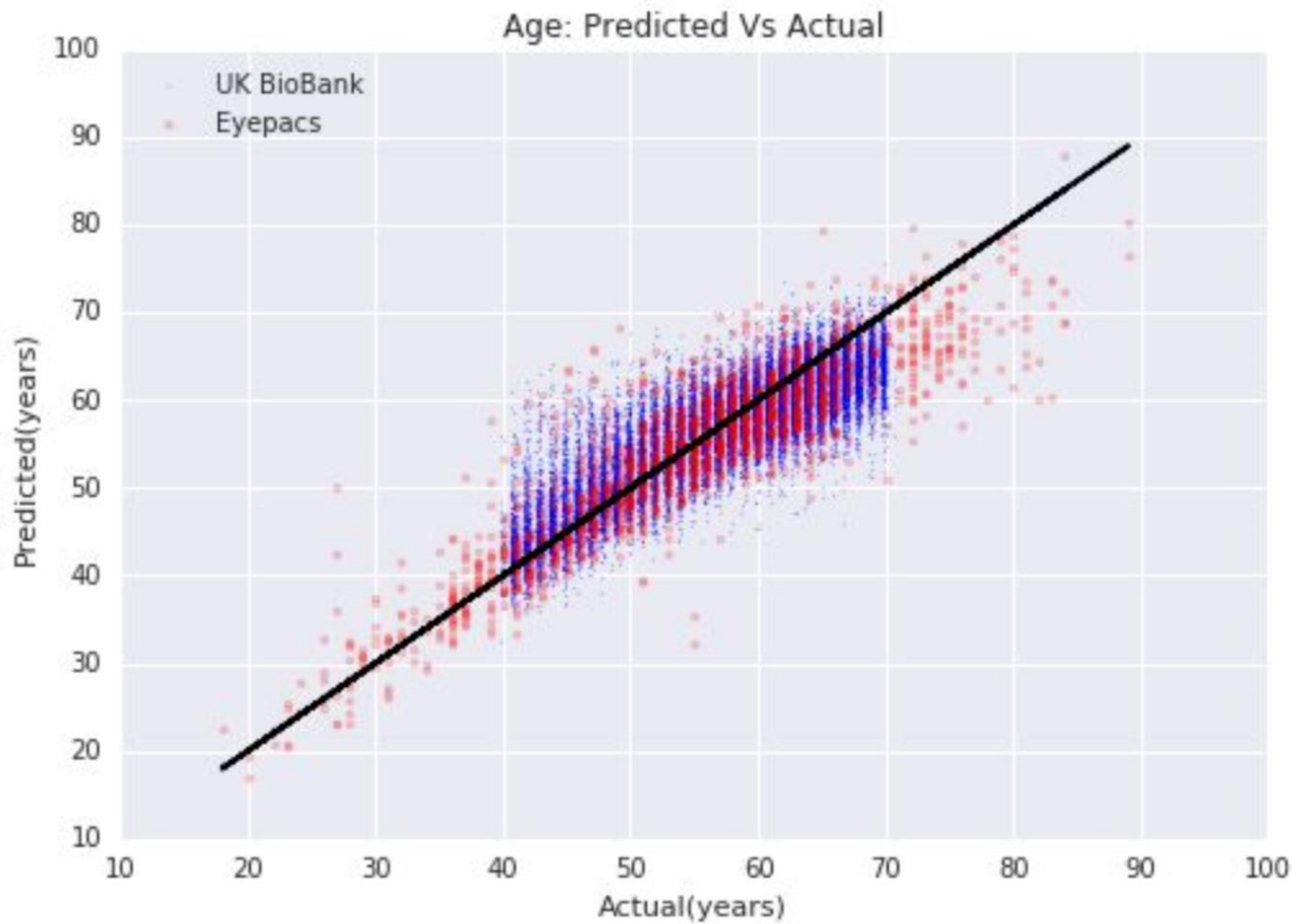
Predicting Cardiovascular Risk Factors from Retinal Fundus Photographs using Deep Learning

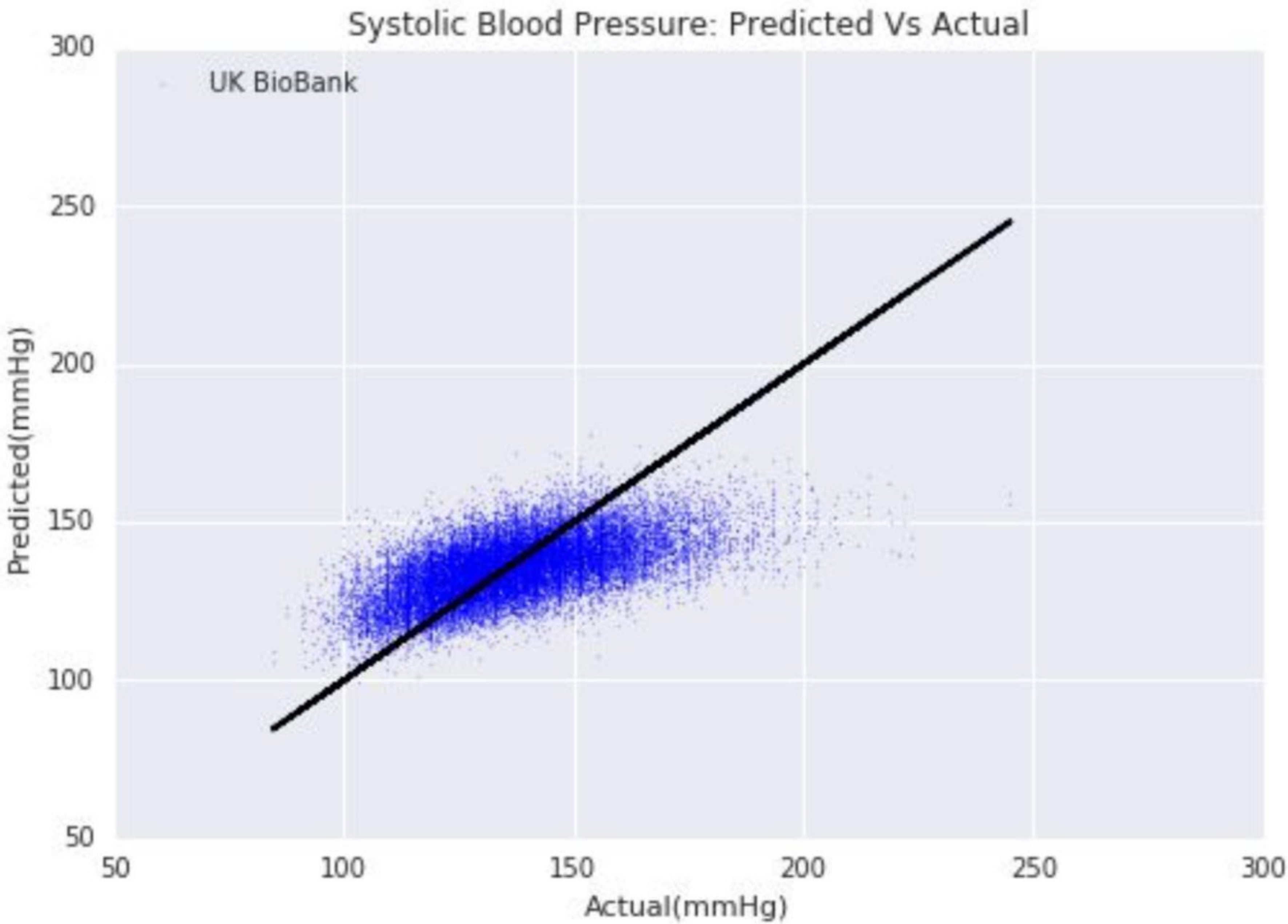
Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng & Dale R. Webster

Using models trained on data from 284,335 patients, and validated on two independent datasets of 12,026 and 999 patients, we predict cardiovascular risk factors not previously thought to be present or quantifiable in retinal images, such as such as age (within 3.26 years), gender (0.97 AUC), smoking status (0.71 AUC), HbA1c (within 1.39%), systolic blood pressure (within 11.23mmHg) as well as major adverse cardiac events (0.70 AUC).

Nature Biomedical Engineering (2018)

doi:10.1038/s41551-018-0195-0

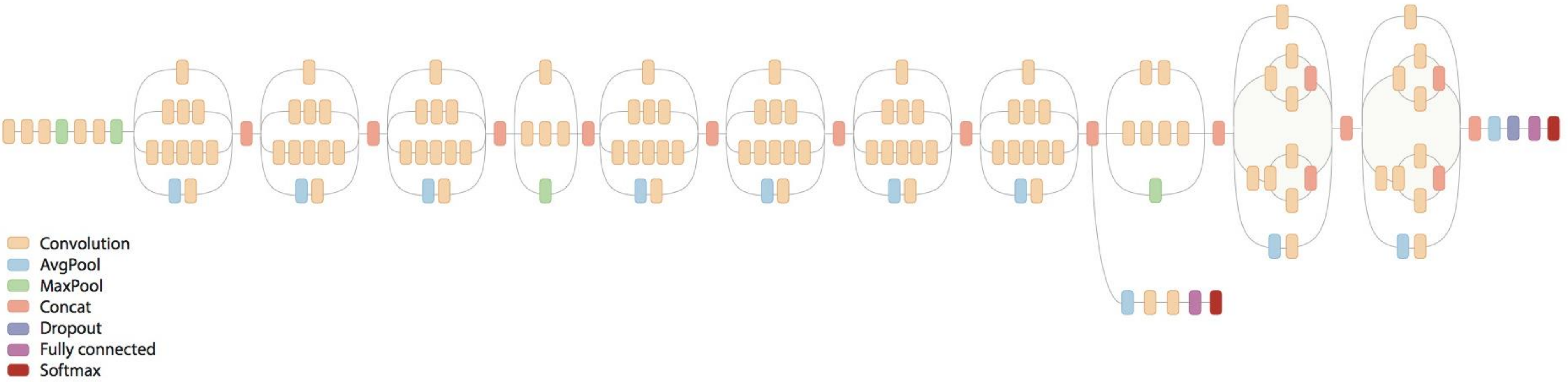

Age: Predicted Vs Actual
UK BioBank
Eyepacs
Predicted(years)
Actual(years)
10
20
30
40
50
60
70
80
90
100


Systolic Blood Pressure: Predicted Vs Actual
UK BioBank
Predicted(mmHg)
Actual(mmHg)
300
250
200
150
100
50
50
100
150
200
250
300

Model	**AUC (95% CI)**
Age	0.66 (0.61-0.71)
Systolic blood pressure (SBP)	0.66 (0.61-0.71)
Body mass index (BMI)	0.62 (0.56-0.67)
Gender	0.57 (0.53-0.62)
Current smoker	0.55 (0.52-0.59)
Algorithm	0.70 (0.65-0.74)
Age + SBP + BMI + gender + current smoker	0.72 (0.68-0.76)
Algorithm + age + SBP + BMI + gender + current smoker	0.73 (0.69-0.77)
Systematic COronary Risk Evaluation (SCORE)[6,7]	0.72 (0.67-0.76)
Algorithm + SCORE	0.72 (0.67-0.76)

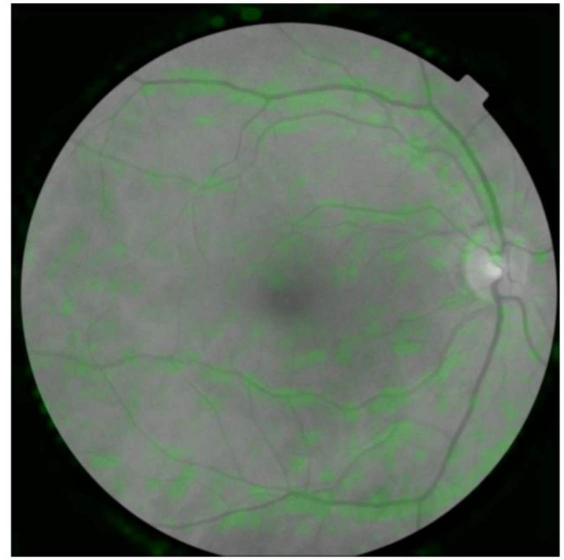
Technical Details

- Inception-v3 architecture
- 28M parameters
- Two models – binary and regression
- 2,000 bootstraps to get AUC and 95% CI



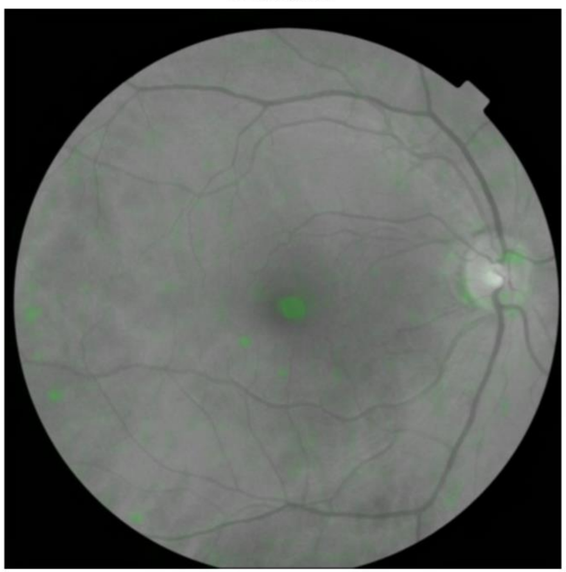
Original

Age



Actual: 57.6 years
Predicted: 59.1 years

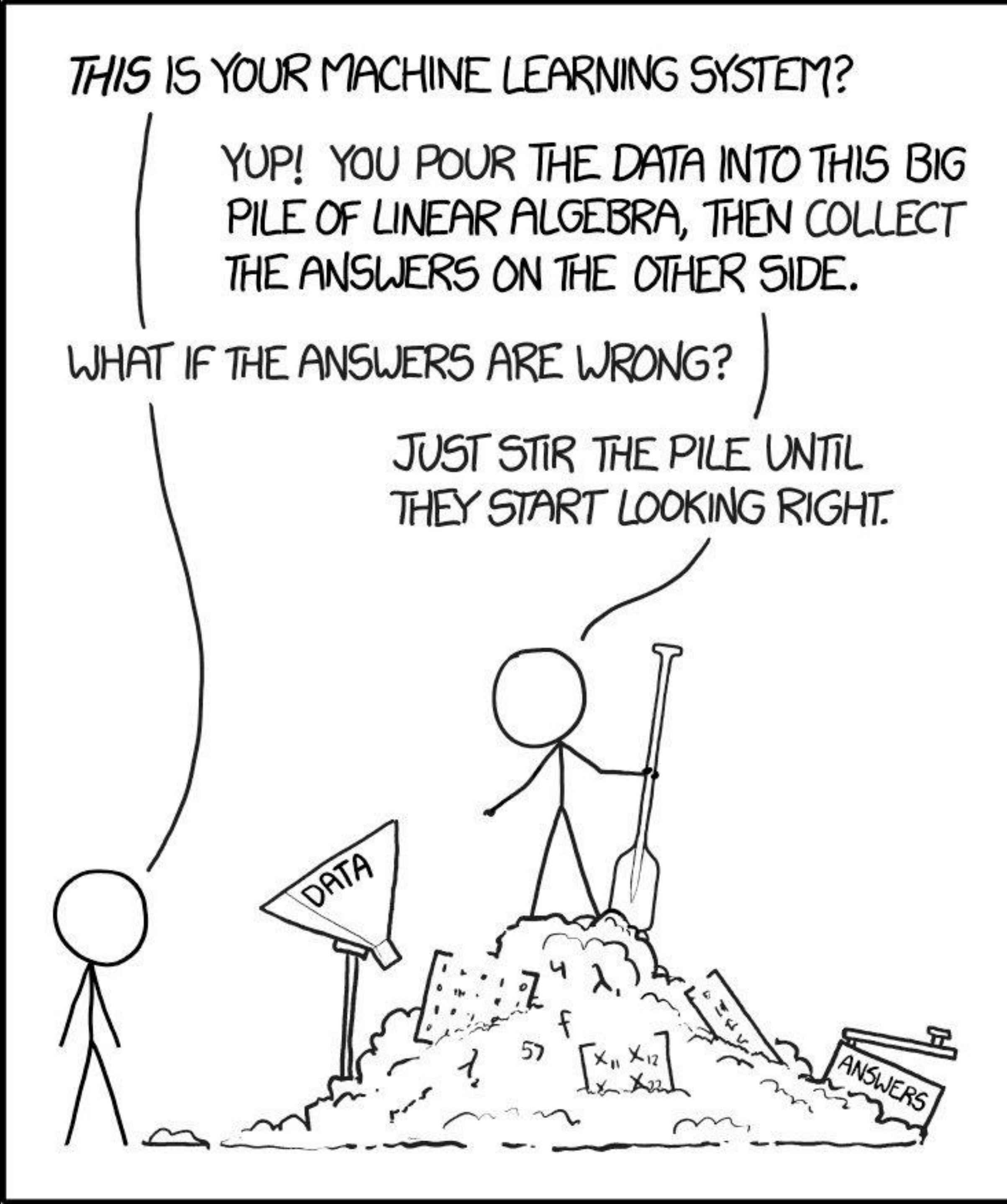
Gender



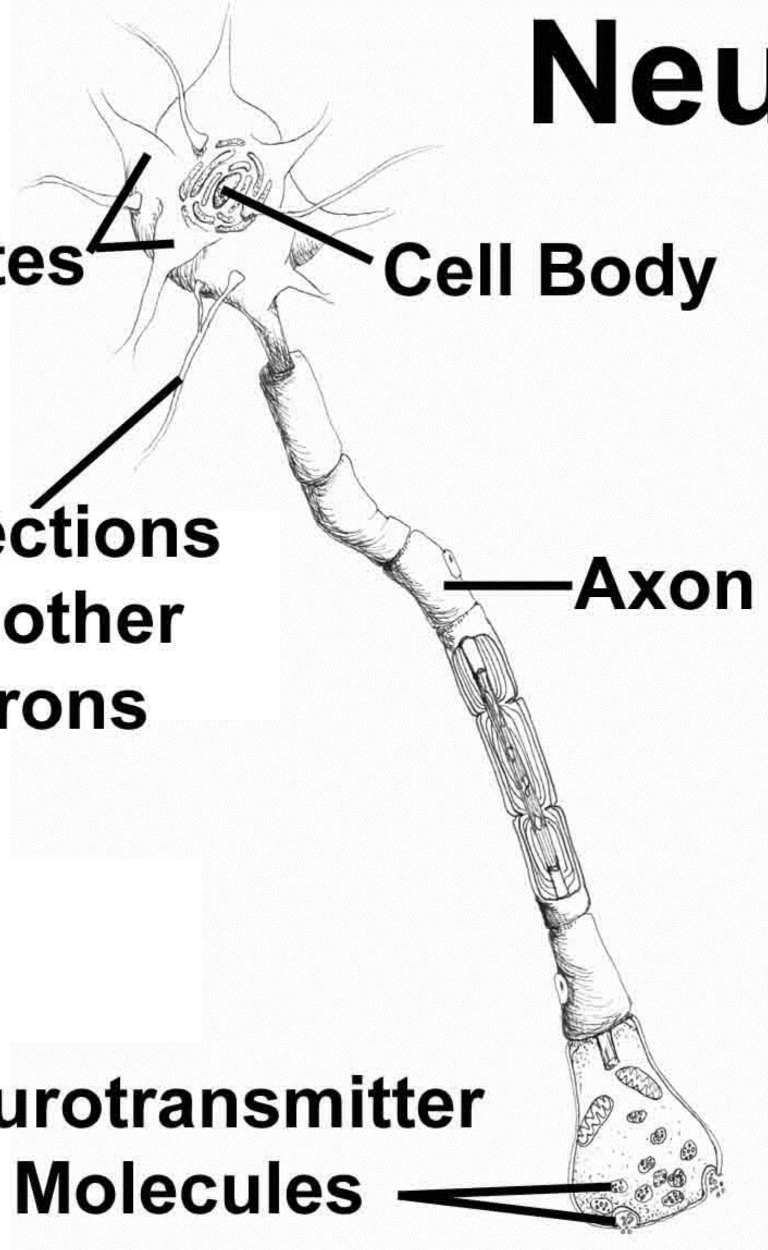
Actual: Female
Predicted: Female

Current smoker

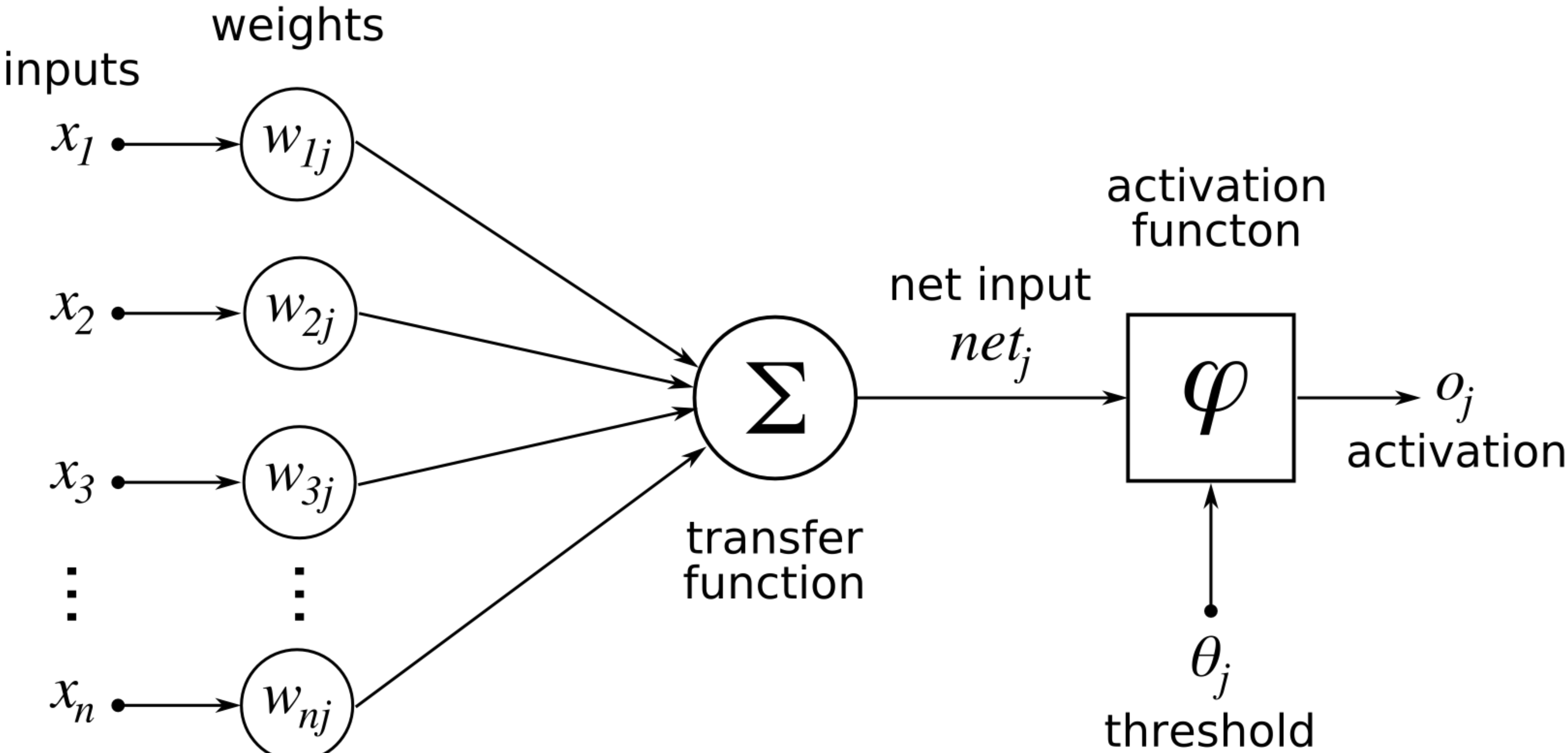
Actual: Nonsmoker
Predicted: Nonsmoker


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

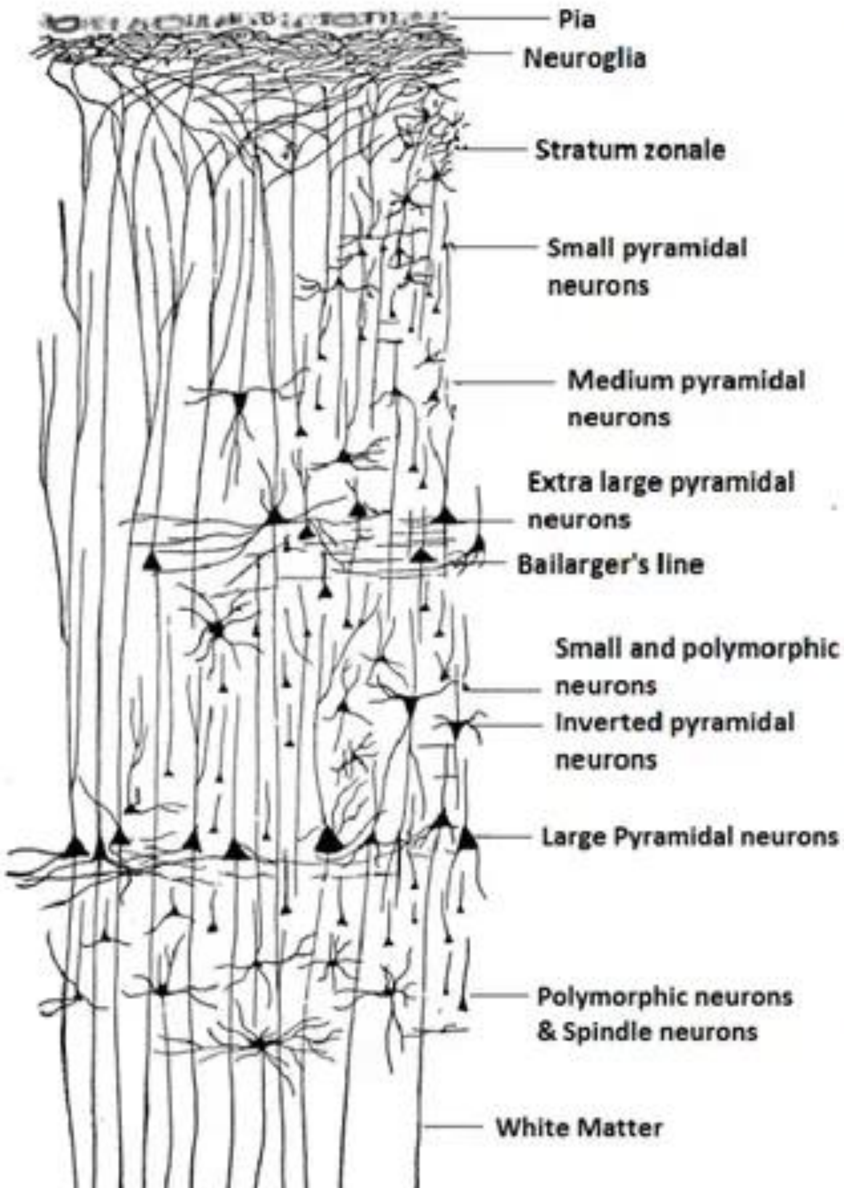
Neuron

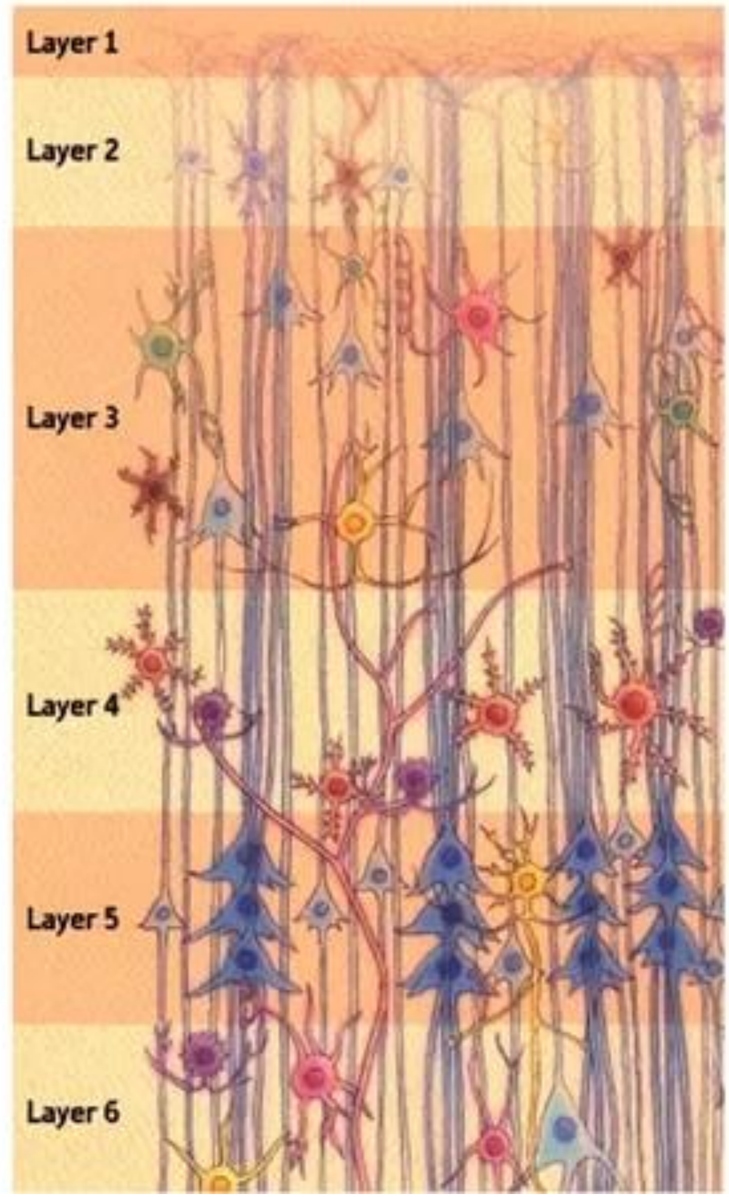

Dendrites
Cell Body
Connections from other Neurons
Axon
Neurotransmitter Molecules

Mathematical Model of a Neuron

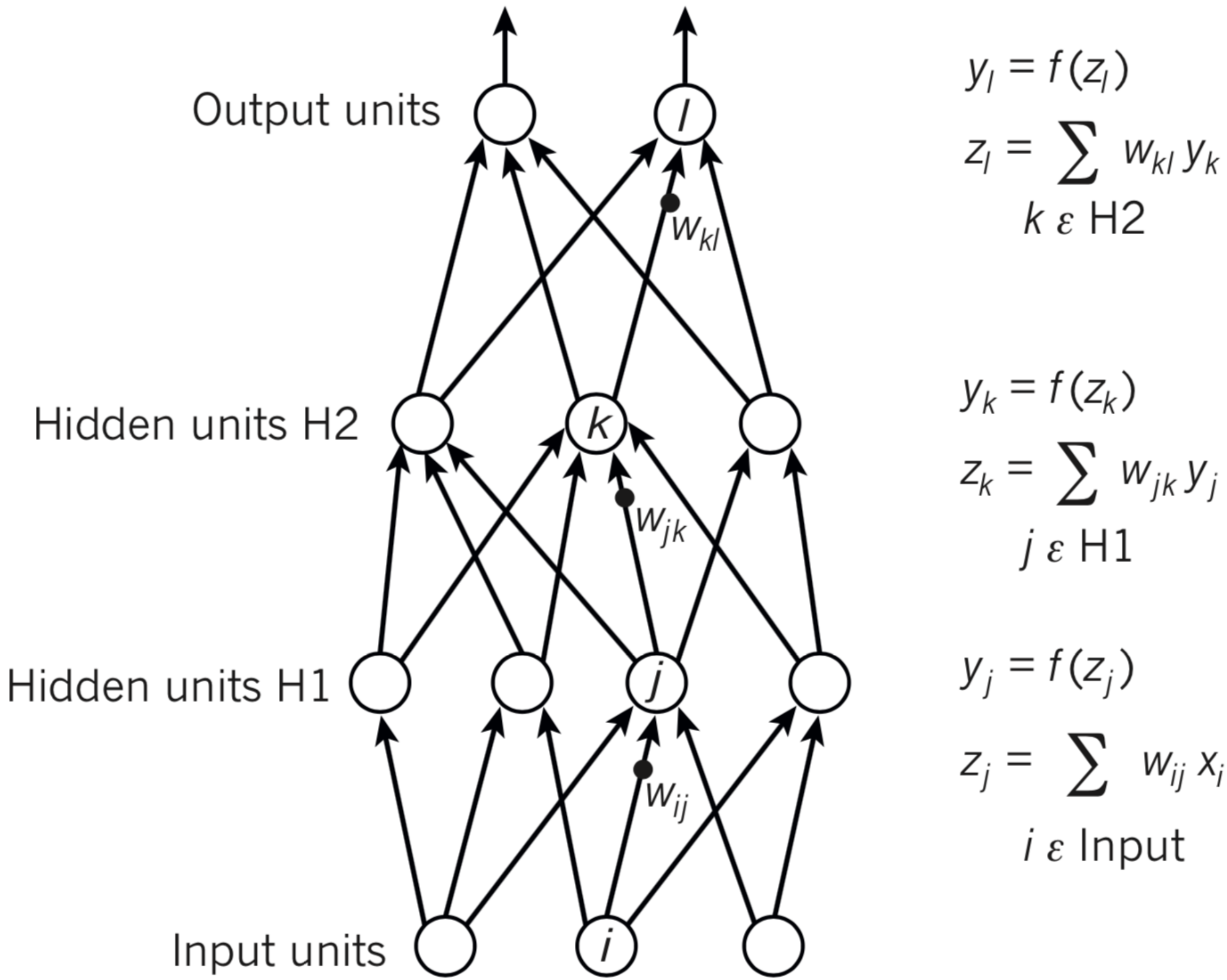


Histological Structrure of the Cerebral Cortex





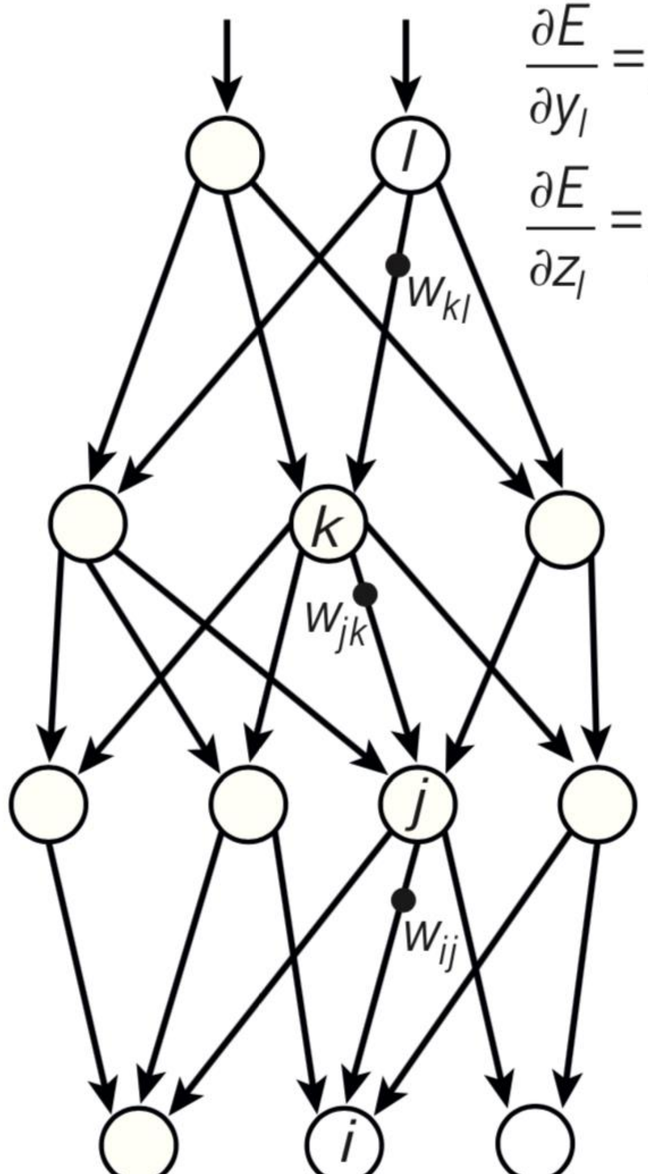
c



Output units
Hidden units H2
Hidden units H1
Input units
l
k
j
i
w_{kl}
w_{jk}
w_{ij}
$y_l = f(z_l)$
$z_l = \sum_{k \in H2} w_{kl} y_k$
$y_k = f(z_k)$
$z_k = \sum_{j \in H1} w_{jk} y_j$
$y_j = f(z_j)$
$z_j = \sum_{i \in Input} w_{ij} x_i$

d

Compare outputs with correct answer to get error derivatives



$$\frac{\partial E}{\partial y_l} = y_l - t_l$$

$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l}\frac{\partial y_l}{\partial z_l}$$

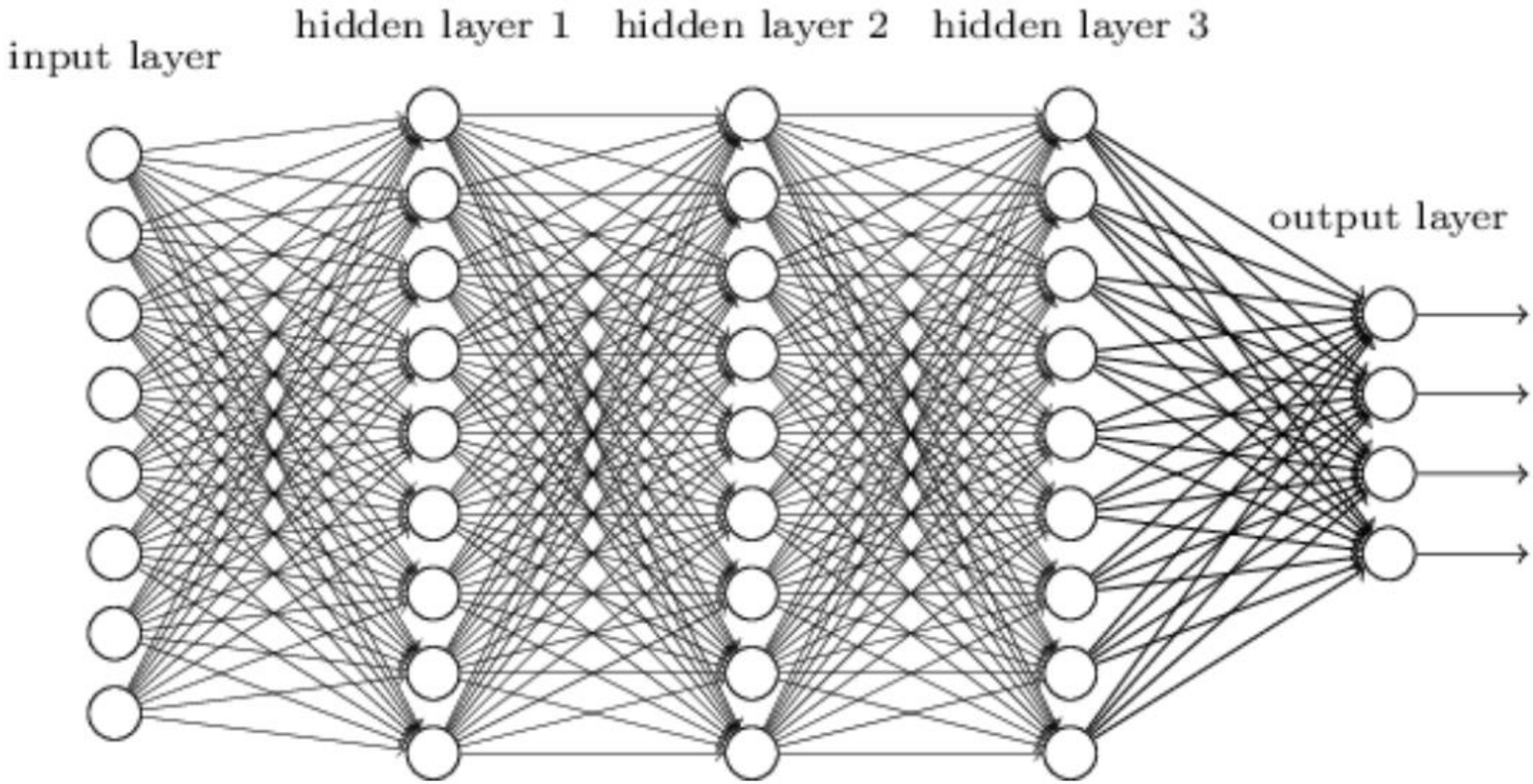
$$\frac{\partial E}{\partial y_k} = \sum_{l\,\varepsilon\,\text{out}} w_{kl}\frac{\partial E}{\partial z_l}$$

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial z_k}$$

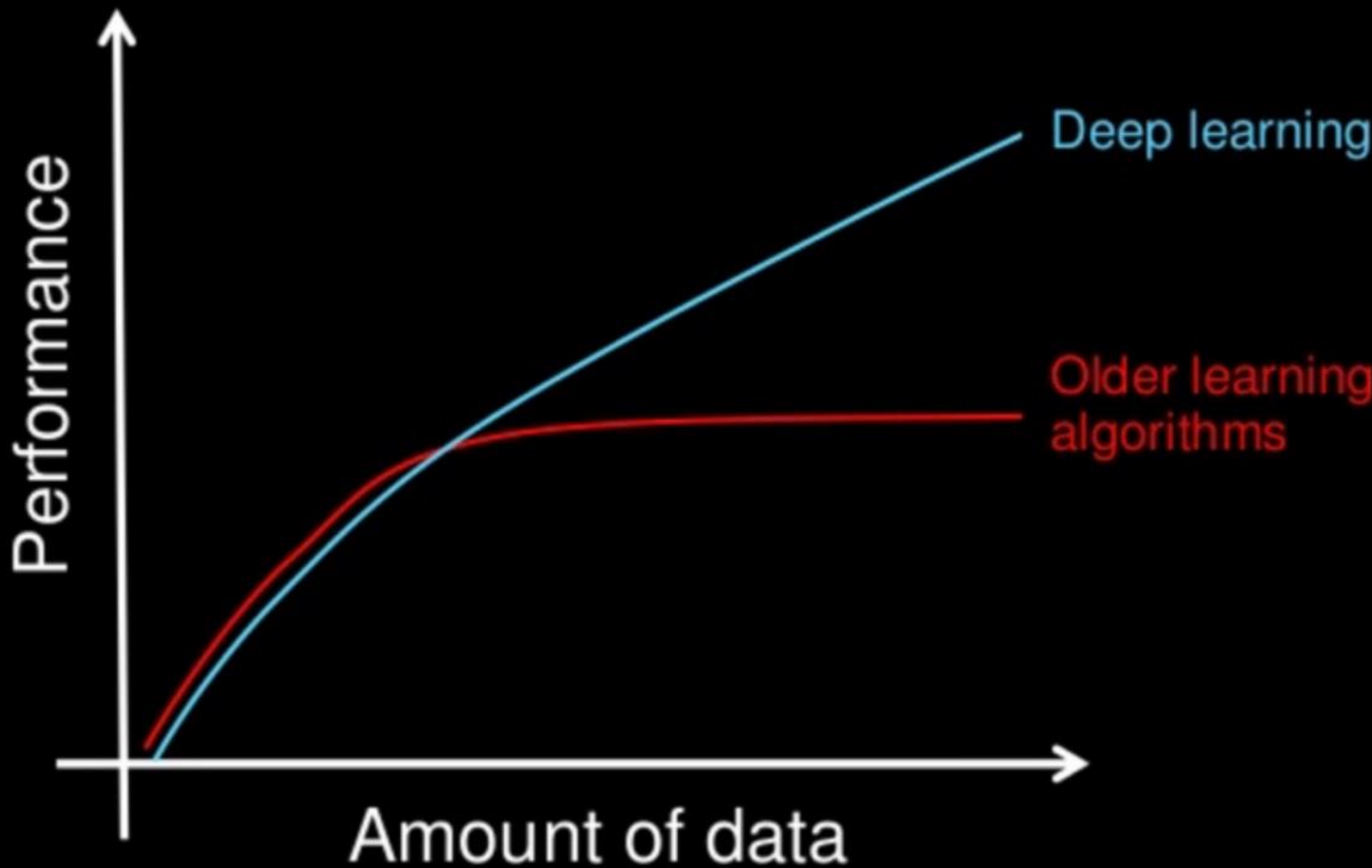
$$\frac{\partial E}{\partial y_j} = \sum_{k\,\varepsilon\,\text{H2}} w_{jk}\frac{\partial E}{\partial z_k}$$

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j}$$

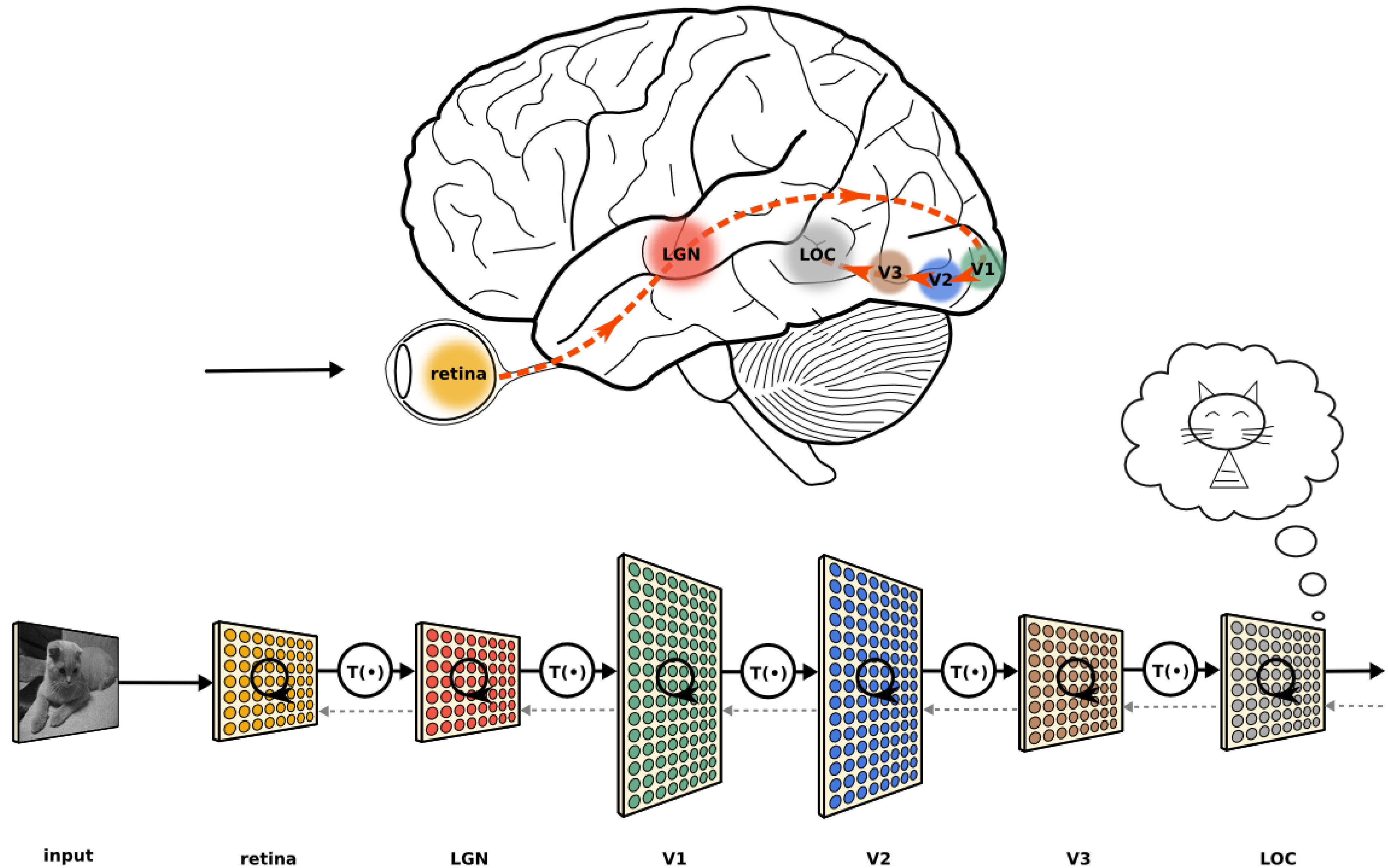
Deep Neural Network



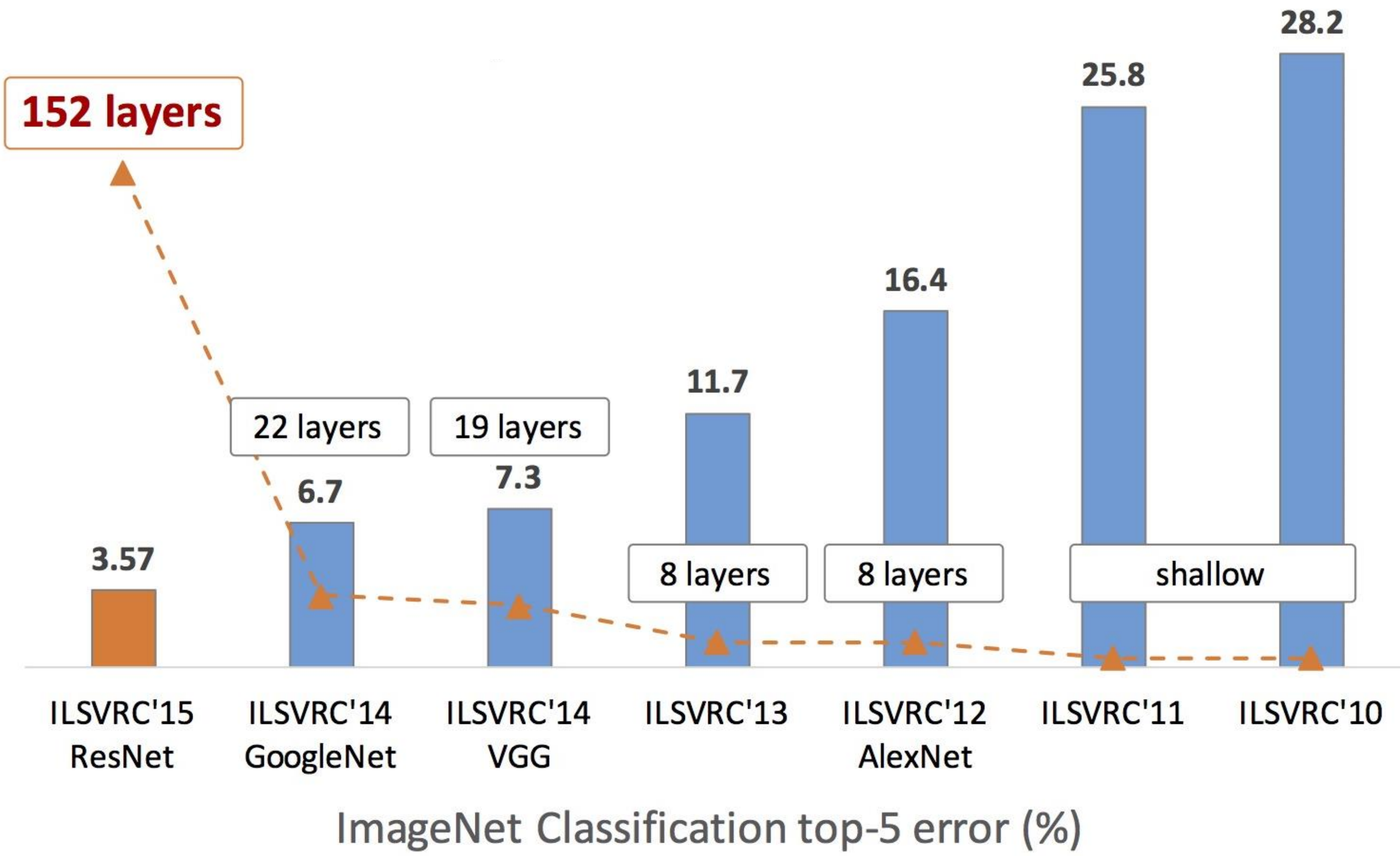
Why deep learning



Human Vision System



Increasing Depth Works!

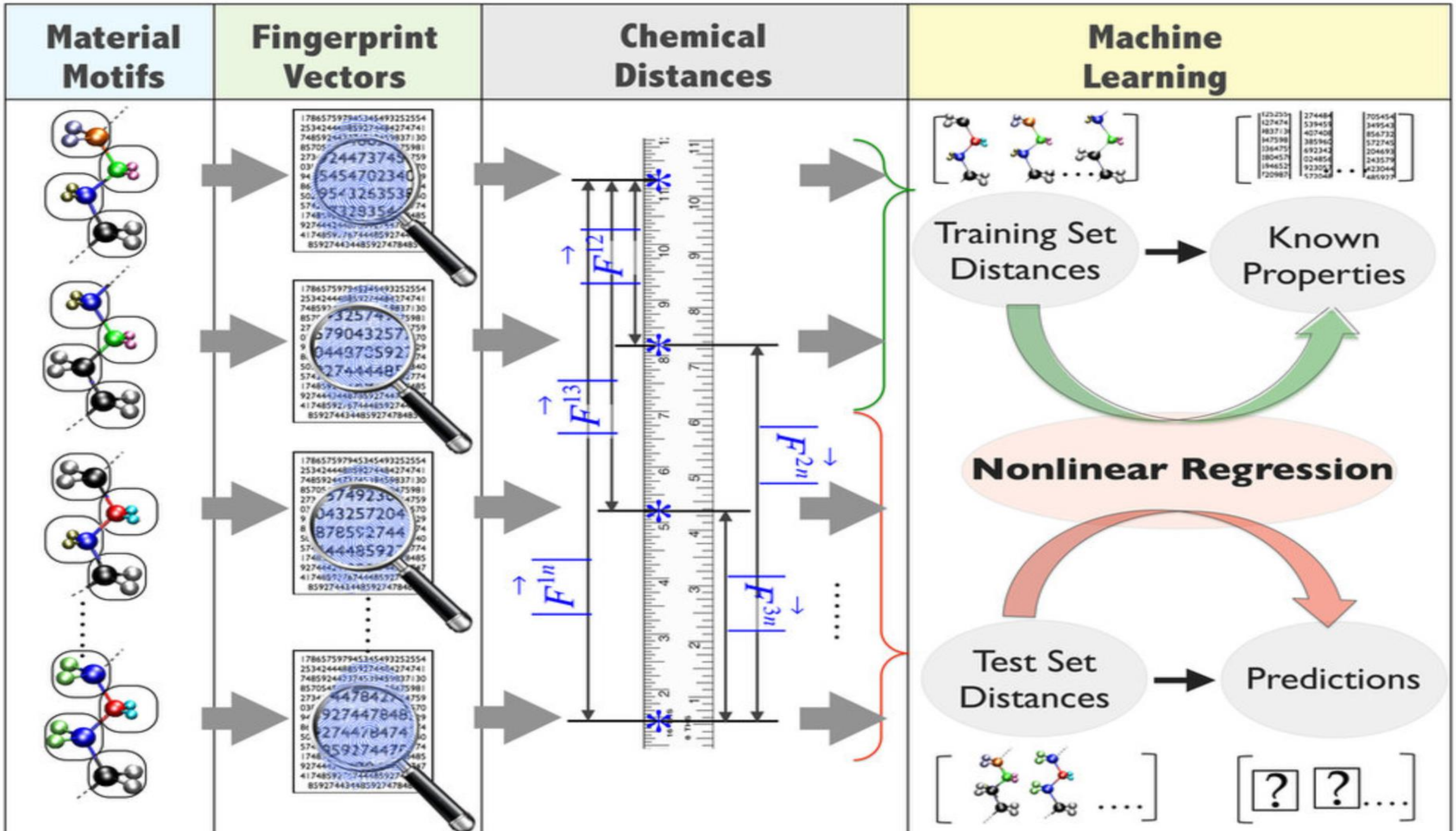



nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
PAGE 462
NATURE.COM/NATURE
28 January 2016 £10
Vol. 529, No. 7587

The Universal Translator

Materials Property Prediction

From: Accelerating materials property predictions using machine learning



Finding Tumors in MRI Images

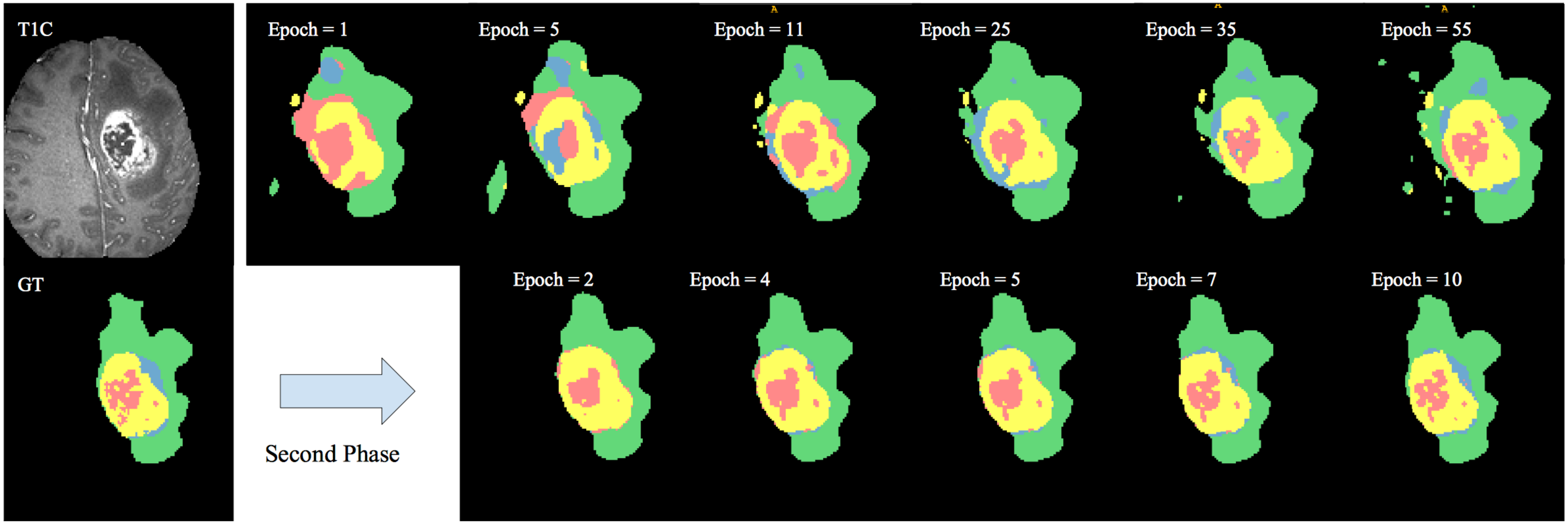


Figure 6: Progression of learning in InputCascadeCNN*. The stream of figures on the first row from left to right show the learning process during the first phase. As the model learns better features, it can better distinguish boundaries between tumor sub-classes. This is made possible due to uniform label distribution of patches during the first phase training which makes the model believe all classes are equiprobable and causes some false positives. This drawback is alleviated by training a second phase (shown in second row from left to right) on a distribution closer to the true distribution of labels. The color codes are as follows: ■ edema, ■ enhanced tumor, ■ necrosis, ■ non-enhanced tumor.

With Unbalanced Training Data!

Artificial Intelligence

Machines Just Beat Humans on a Stanford Reading Comprehension Test

Creative Commons

IN BRIEF

The Stanford Question Answering Dataset is a well-respected means of testing machine reading. For the first time, an artificial intelligence has scored higher than a human participant.

SHARE

WRITTEN BY

Brad Jones

Published: 2 hours ago

#Alibaba #machine reading #microsoft

READ ME

Chinese retail giant Alibaba has developed an artificial intelligence model that's managed to outdo human participants in a reading and comprehension test designed by Stanford University. The model scored 82.44, whereas humans recorded a score of 82.304.

The Stanford Question Answering Dataset is a set of 10,000 questions pertaining to some 500 Wikipedia articles. The answer to each question is a particular span of text from the corresponding piece of writing.

Alibaba claims that its accomplishment is the first time that humans have been outmatched on this particular test, according to a report from Bloomberg. Microsoft also managed a similar feat, scoring 82.650 — though, those results were finalized shortly after Alibaba's.

Shutterstock / Denis Linine

Artificial Intelligence

Google's New AI Is Better at Creating AI Than the Company's Engineers

IN BRIEF

At its I/O '17 conference this week, Google shared details of its AutoML project, an artificial intelligence that can assist in the creation of other AIs. By automating some of the complicated process, AutoML could make machine learning more accessible to non-experts.

SHARE

WRITTEN BY

AUTHOR
Tom Ward

EDITOR
Kristin Houser
Website

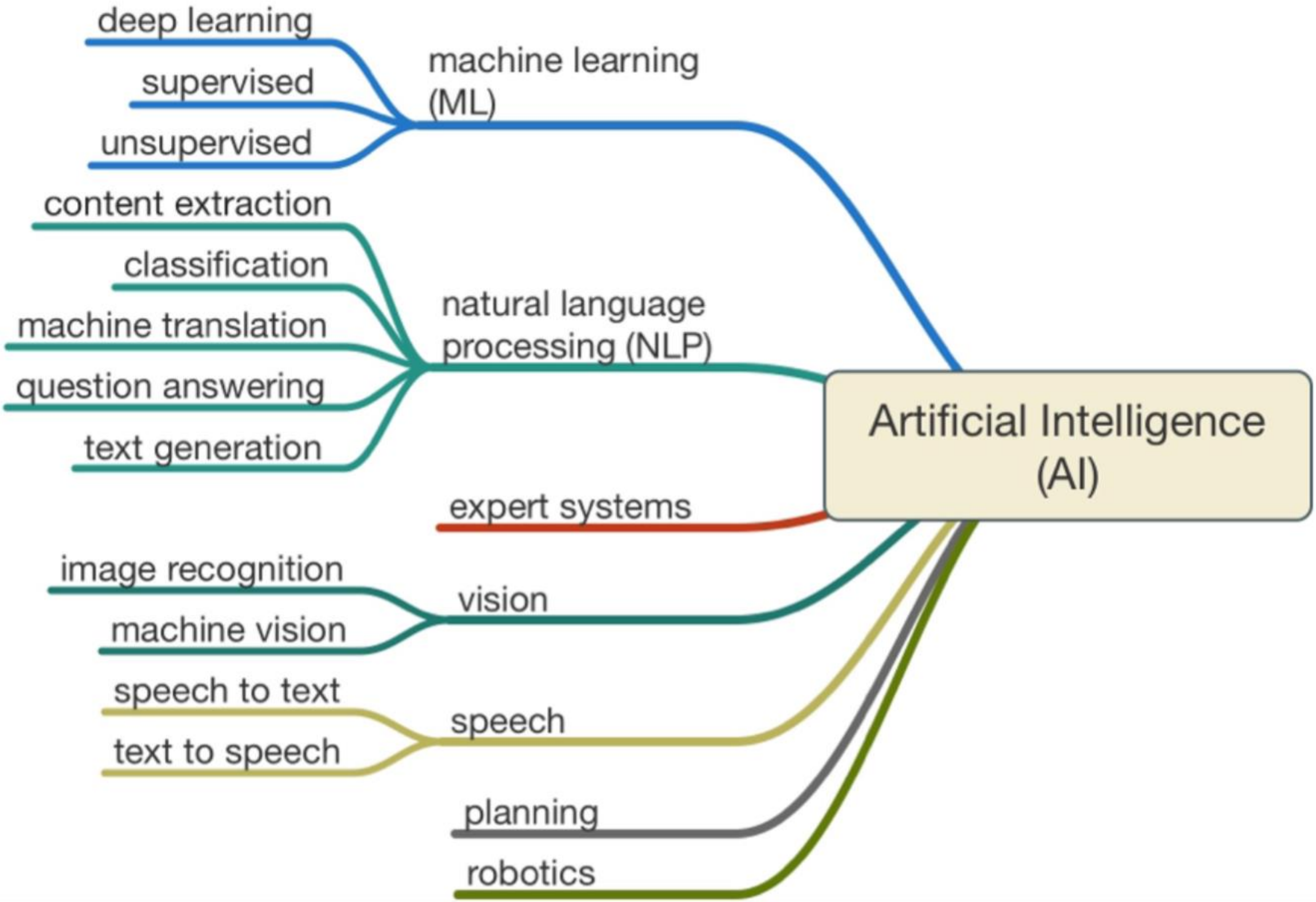
Published: May 19, 2017

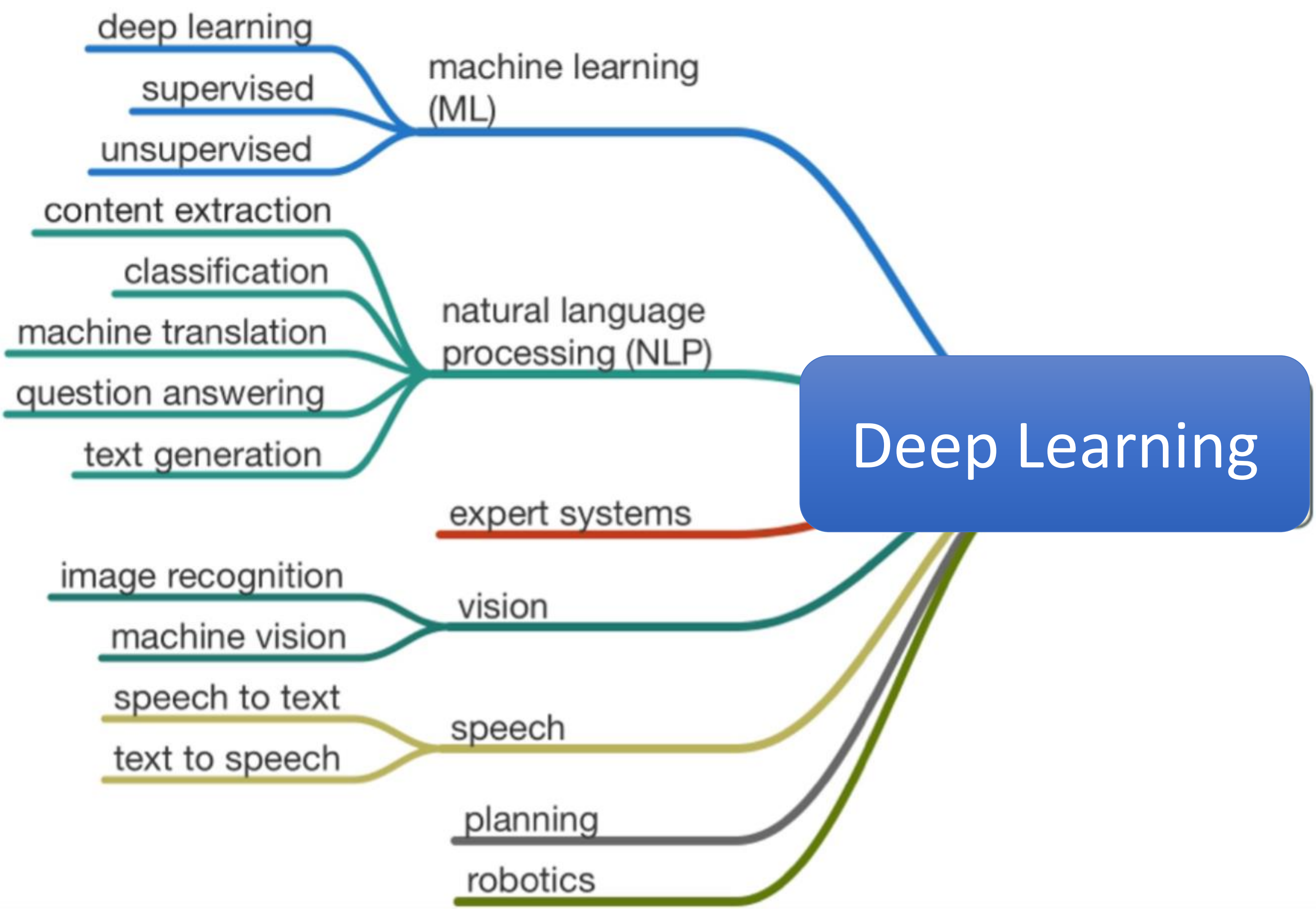
Last updated: May 19, 2017 at 12:29 pm

#artificial intelligence #deep learning #Google #machine learning

GOOGLE'S AUTOML

One of the more noteworthy remarks to come out of Google I/O '17 conference this week was CEO Sundar Pichai recalling how his team had joked that they have achieved "AI inception" with AutoML. Instead of crafting layers of dreams like in the Christopher Nolan flick, however, the AutoML system layers artificial intelligence (AI), with AI systems creating better AI systems.

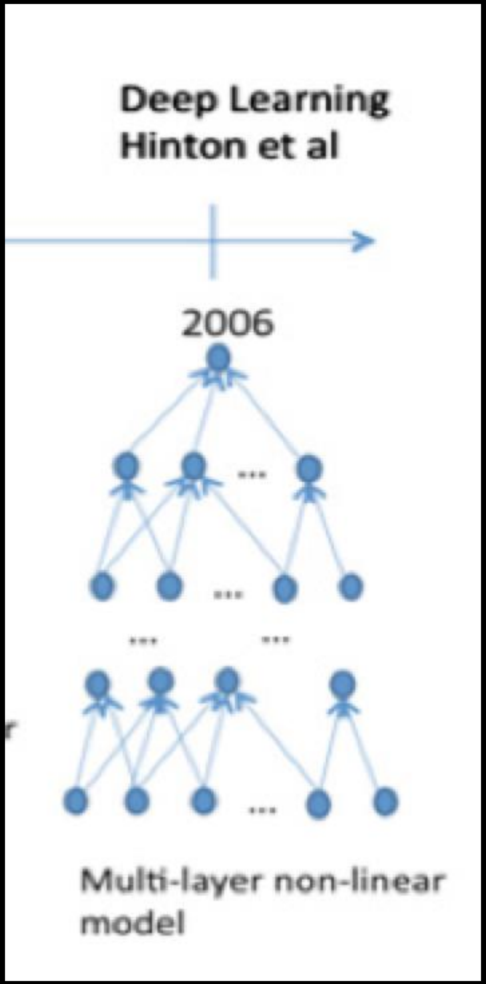

Artificial Intelligence (AI)
machine learning (ML)
deep learning
supervised
unsupervised
natural language processing (NLP)
content extraction
classification
machine translation
question answering
text generation
expert systems
vision
image recognition
machine vision
speech
speech to text
text to speech
planning
robotics



Deep Learning
machine learning (ML)
deep learning
supervised
unsupervised
natural language processing (NLP)
content extraction
classification
machine translation
question answering
text generation
expert systems
vision
image recognition
machine vision
speech
speech to text
text to speech
planning
robotics

Deep Learning Impacting Science

- Climate
- Biology
- Drug Design
- Materials Design
- Cosmology
- High-Energy Physics



Hundreds of Researchers Diving into Deep Learning at Argonne

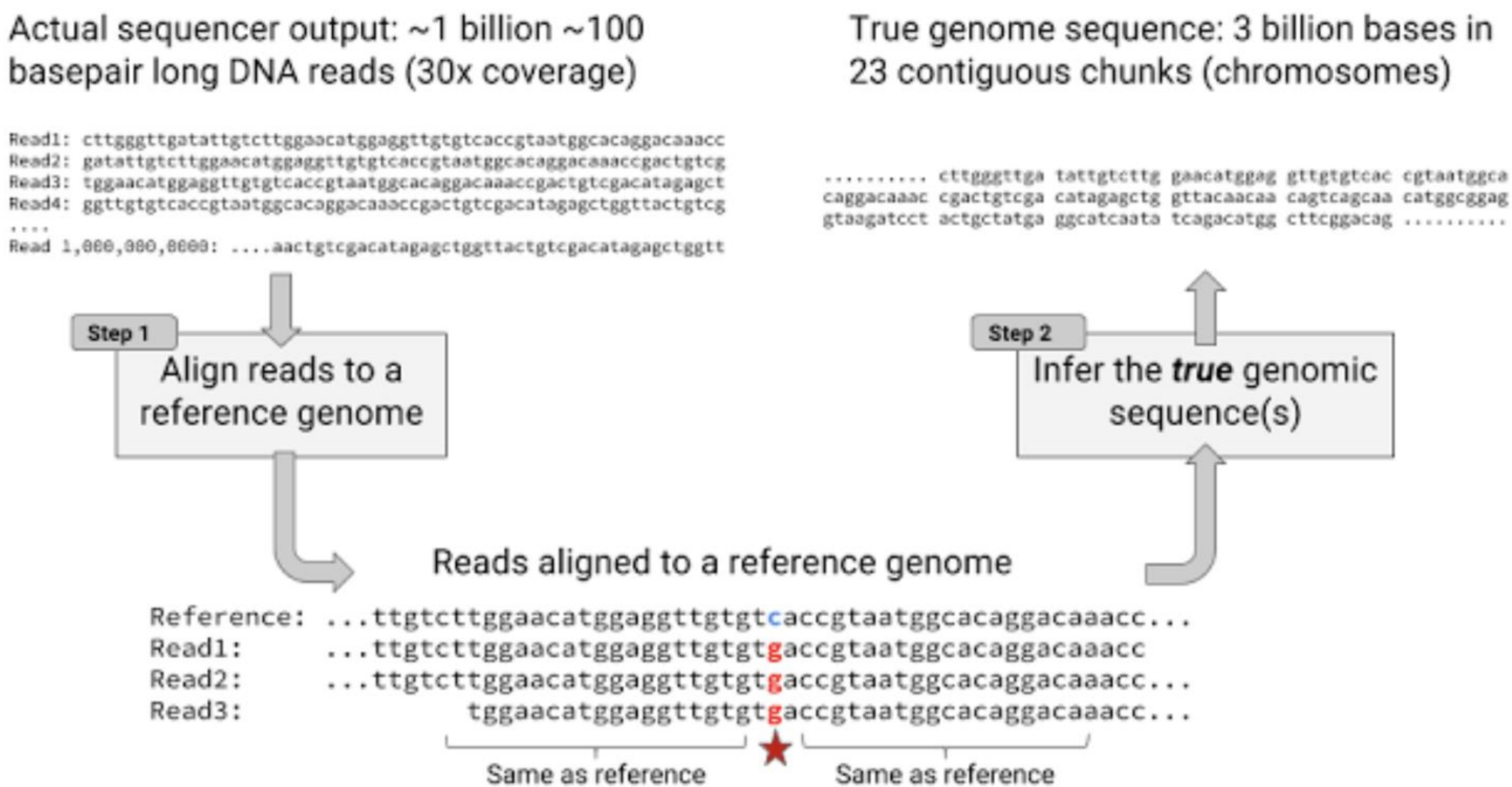
~~Machine~~ [Deep] Learning In Cancer Research

- Cancer Susceptibility
- Cancer Detection and Diagnosis
- Cancer Recurrence
- Cancer Prognosis and Survival
- Cancer Classification and Clustering
- Cancer Drug Response Prediction
- Cancer Genomics Analysis

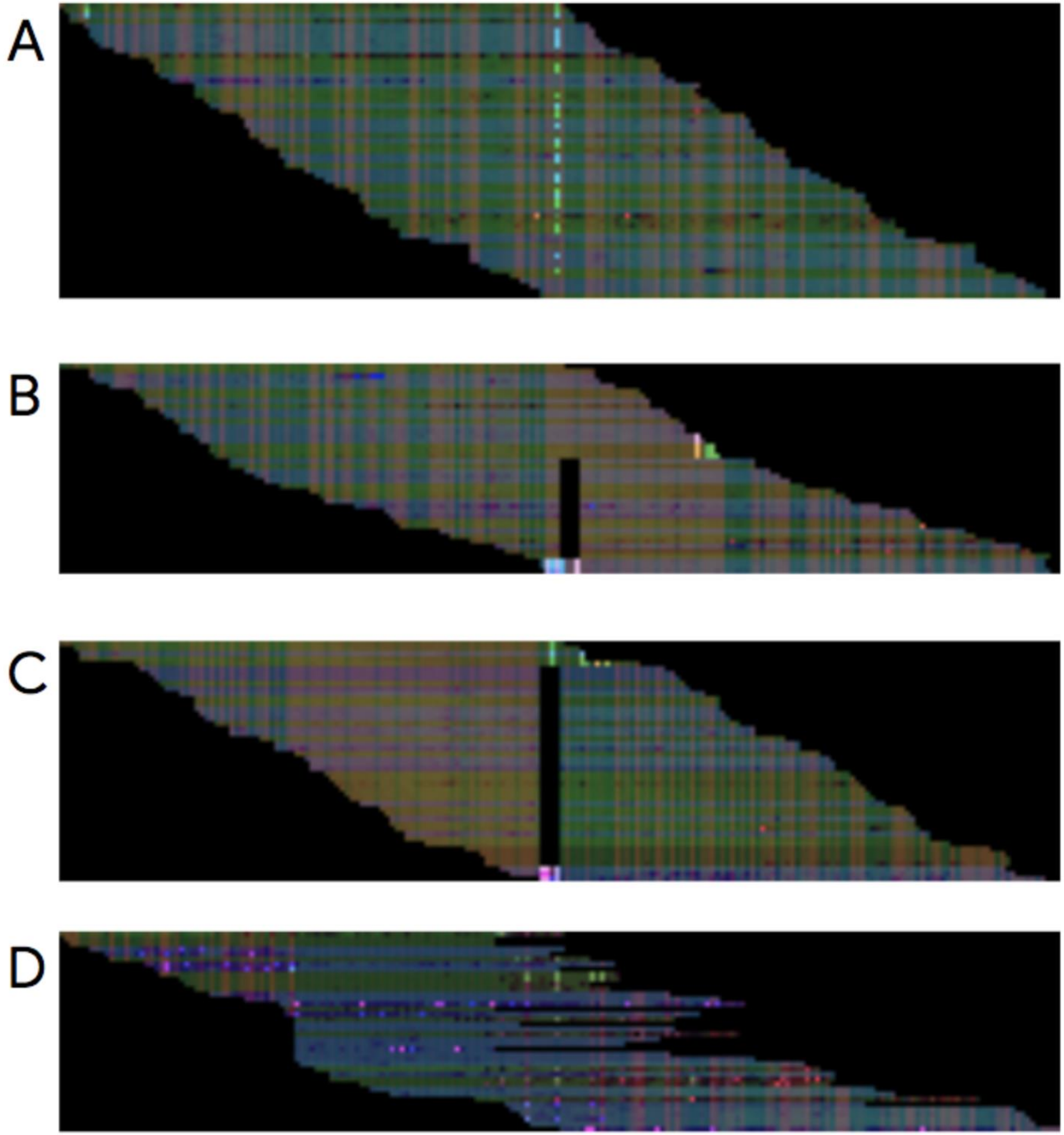
Area	Applications	Input data	Base Method	Reference
Bioinformatics	Cancer diagnosis Gene selection/classification Gene variants	Gene expression MicroRNA Microarray data	Deep Autoencoders	[2]
			Deep Belief Network	[46], [47]
			Deep Neural Network	[48]
	Drug design	Molecule compounds	Deep Neural Network	[49]
	Compound-Protein interaction RNA binding protein DNA methylation	Protein structures Molecule compounds Genes/RNA/DNA sequences	Deep Belief Network	[50]
			Deep Neural Network	[51], [52]
Medical Imaging	3D brain reconstruction Neural cells classification Brain tissues classification Alzheimer/MCI diagnosis	MRI/fMRI Fundus images PET scans	Deep Autoencoders	[53], [54]
			Convolutional Neural Network	[55]–[59]
			Deep Belief Network	[60], [61]
			Deep Near Network	[62]
	Tissue classification Organ segmentation Cell clustering Hemorrhage detection Tumour detection	MRI/CT Images Endoscopy images Microscopy Fundus Images X-ray images Hyperspectral images	Convolutional Deep Belief Network	[63], [64]
			Convolutional Neural Network	[65]–[76]
			Deep Autoencoder	[66], [77]
			Group Method of Data Handling	[78]–[81]
			Deep Neural Network	[82]–[85]
Pervasive Sensing	Anomaly detection Biological parameters monitoring	EEG ECG Implantable device	Deep Belief Network	[86]–[89]
	Human activity recognition	Video Wearable device	Convolutional Neural Network	[90]–[93]
			Deep Belief Network	[94], [95]
			Deep Neural Network	[96]
	Hand gesture recognition Obstacle detection Sign language recognition	Depth camera RGB-D camera Real-Sense camera	Convolutional Neural Network	[97]
			Deep Belief Network	[98]
	Food intake Energy expenditure	Wearable device RGB Image Mobile device	Convolutional Neural Network	[99]
			Deep Neural Network	[100]
Medical Informatics	Prediction of disease Human behaviour monitoring Data mining	Electronic health records Big medical dataset Blood/Lab tests	Deep Autoencoders	[101], [102]
			Deep Belief Network	[103], [104]
			Convolutional Neural Network	[105]
			Recurrent Neural Network	[101], [106]
			Convolutional Deep Belief Network	[107]
			Deep Neural Network	[108], [109]
Public Health	Predicting demographic info Lifestyle diseases Infectious disease epidemics Air pollutant prediction	Social media data Mobile phone metadata Geo-tagged images Text messages	Deep Autoencoders	[110]
			Deep Belief Network	[111], [112]
			Convolutional Neural Network	[113]
			Deep Neural Network	[114]–[117]

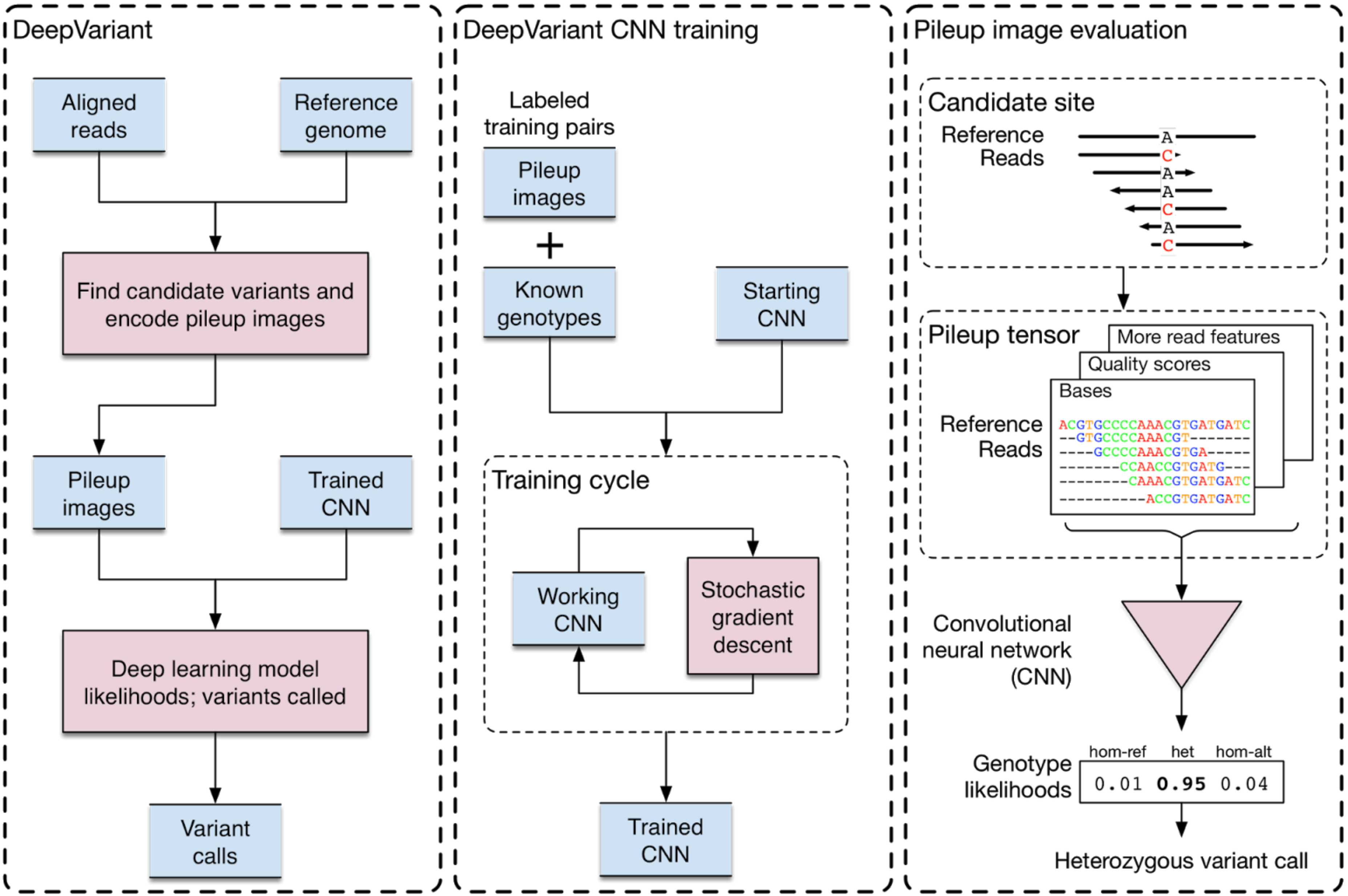
Mapping problems to Images Enables Deep Learning Methods to be Applied

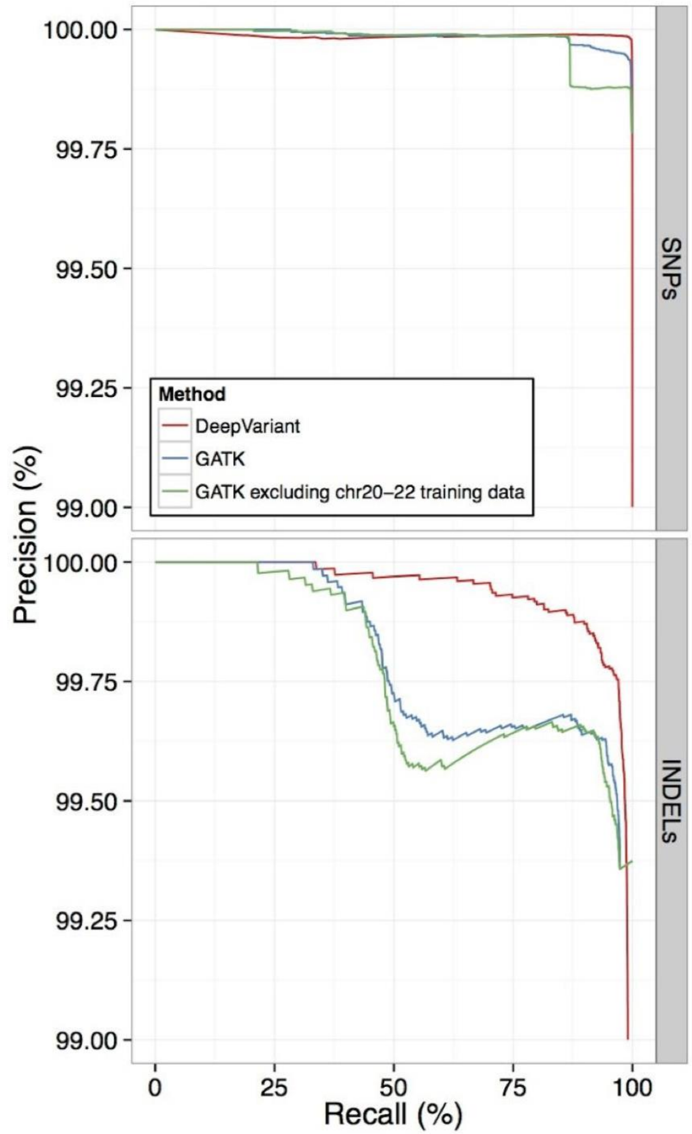
Google's DeepVariant
https://www.biorxiv.org/content/early/2016/12/14/092890



For any given location in the genome, there are multiple reads among the ~1 billion that include a base at that position. Each read is aligned to a reference, and then each of the bases in the read is compared to the base of the reference at that location. When a read includes a base that differs from the reference, it may indicate a variant (a difference in the true sequence), or it may be an error.


A
B
C
D

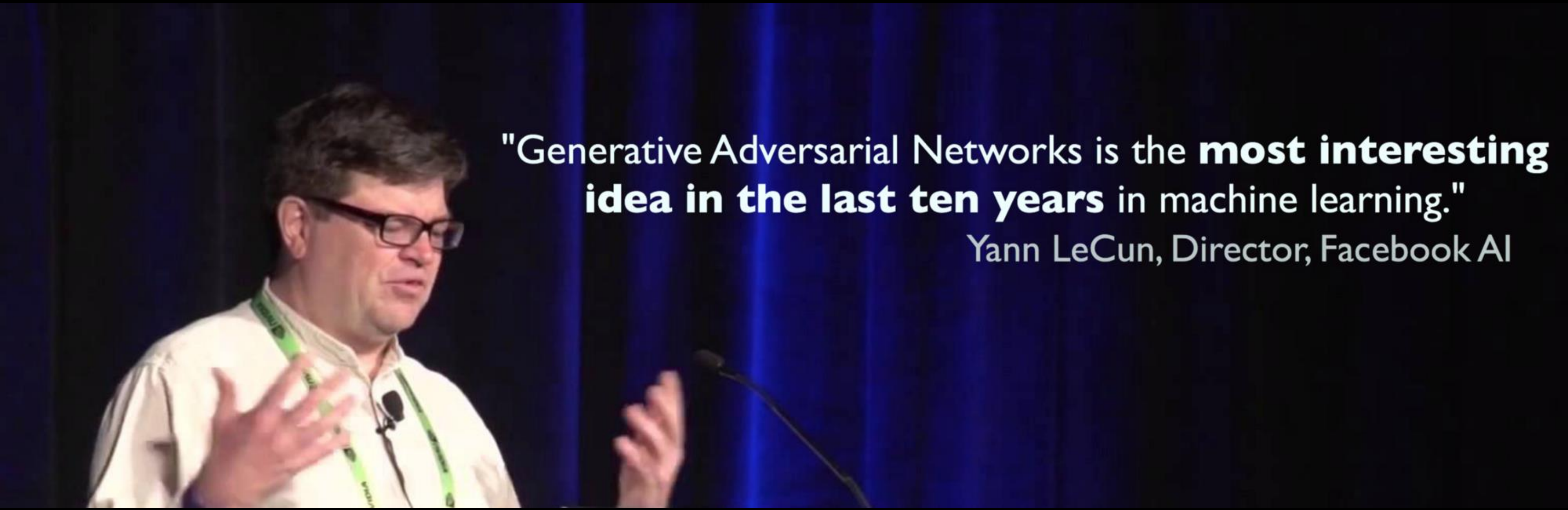

DeepVariant
Aligned reads
Reference genome
Find candidate variants and encode pileup images
Pileup images
Trained CNN
Deep learning model likelihoods; variants called
Variant calls
DeepVariant CNN training
Labeled training pairs
Pileup images
+
Known genotypes
Starting CNN
Training cycle
Working CNN
Stochastic gradient descent
Trained CNN
Pileup image evaluation
Candidate site
Reference
Reads
Pileup tensor
More read features
Quality scores
Bases
Reference
Reads
ACGTGCCCCAAACGTGATGATC
--GTGCCCCAAACGT-------
----GCCCCAAACGTGA-----
-------CCAACCGTGATG---
--------CAAACGTGATGATC
----------ACCGTGATGATC
Convolutional neural network (CNN)
Genotype likelihoods
hom-ref
het
hom-alt
0.01 0.95 0.04
Heterozygous variant call



Precision (%)
Recall (%)
100.00
99.75
99.50
99.25
99.00
0
25
50
75
100
SNPs
INDELs
Method
DeepVariant
GATK
GATK excluding chr20−22 training data

Method	Type	F1	Recall	Precision	TP	FN	FP	FP.gt	FP.al	Version
DeepVariant (live github)	INDEL	0.99507	0.99347	0.99666	357641	2350	1198	217	840	latest github v0.4.1-b4e8d37d
Strelka	INDEL	0.99227	0.98829	0.99628	355777	4214	1329	221	855	2.8.4-3-gbe58942
DeepVariant (pFDA)	INDEL	0.99112	0.98776	0.99450	355586	4405	1968	846	1027	precisionFDA submission 05/2016
GATK	INDEL	0.99010	0.98454	0.99573	354425	5566	1522	343	909	3.8-0-ge9d806836
FreeBayes	INDEL	0.94091	0.91917	0.96372	330891	29100	12569	9149	3347	v1.1.0-54-g49413aa
16GT	INDEL	0.92732	0.91102	0.94422	327960	32031	19364	10700	7745	v1.0-34e8f934
samtools	INDEL	0.87951	0.83369	0.93066	300120	59871	22682	2302	20282	1.6
DeepVariant (live github)	SNP	0.99982	0.99975	0.99989	3054552	754	350	157	38	latest github v0.4.1-b4e8d37d
DeepVariant (pFDA)	SNP	0.99958	0.99944	0.99973	3053579	1727	837	409	78	precisionFDA submission 05/2016
Strelka	SNP	0.99935	0.99893	0.99976	3052050	3256	732	87	136	2.8.4-3-gbe58942
16GT	SNP	0.99583	0.99850	0.99318	3050725	4581	20947	3476	3899	v1.0-34e8f934
GATK	SNP	0.99436	0.98940	0.99937	3022917	32389	1920	80	170	3.8-0-ge9d806836
FreeBayes	SNP	0.99124	0.98342	0.99919	3004641	50665	2434	351	1232	v1.1.0-54-g49413aa
samtools	SNP	0.99021	0.98114	0.99945	2997677	57629	1651	1040	200	1.6

Generative Adversarial Networks


"Generative Adversarial Networks is the **most interesting idea in the last ten years** in machine learning."
Yann LeCun, Director, Facebook AI

What are Generative Models?

Key Idea: our model cares about what distribution generated the input data points, and we want to mimic it with our probabilistic model. **Our learned model should be able to make up new samples from the distribution**, **not just copy and paste existing samples**!

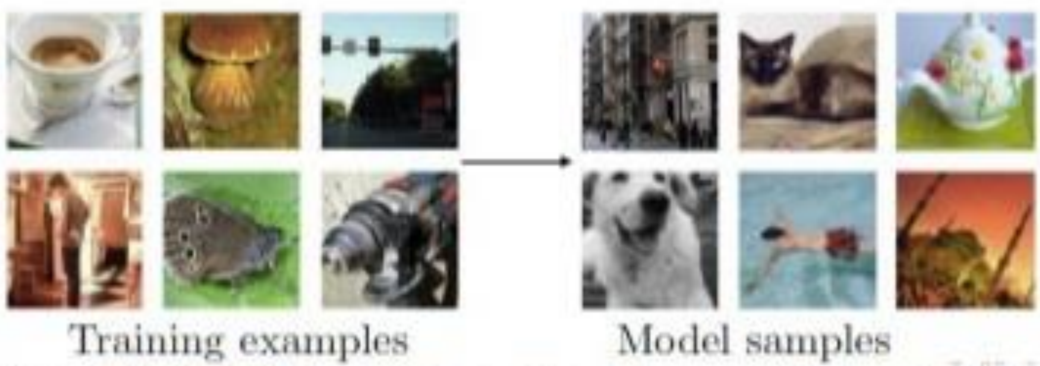
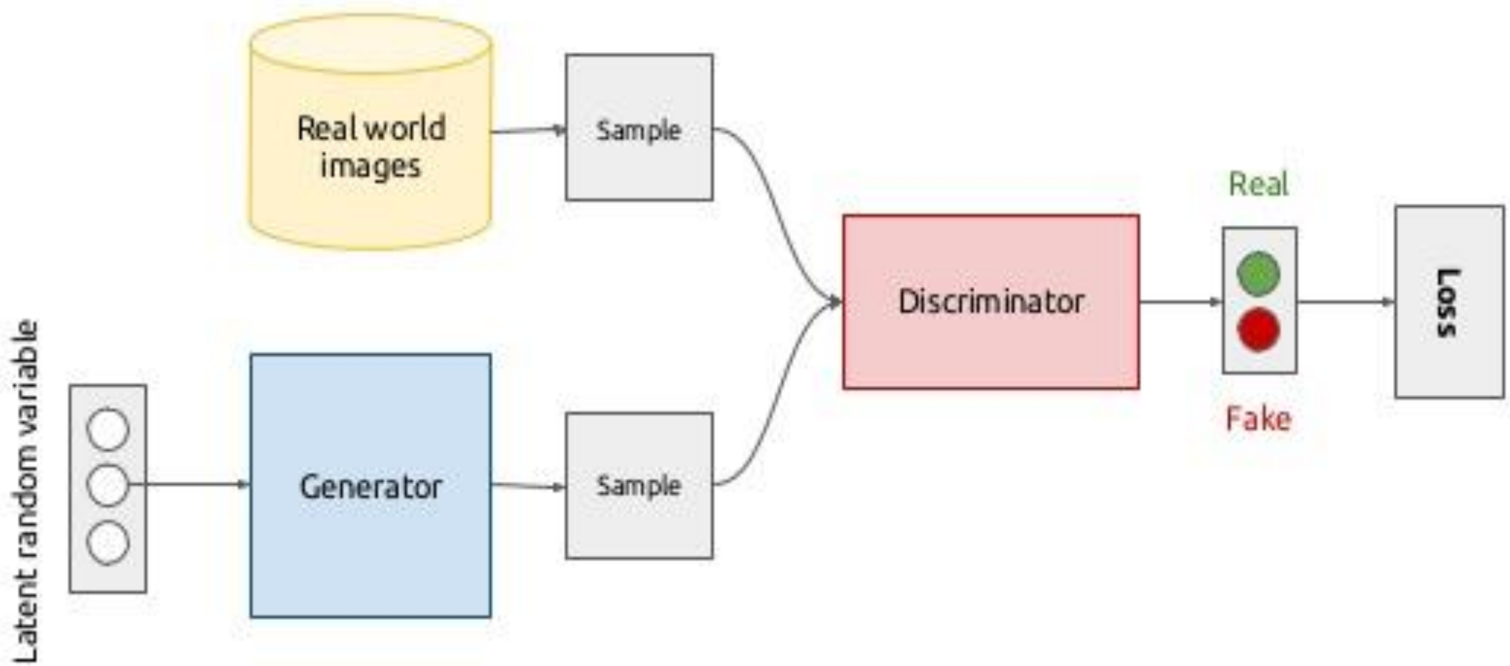
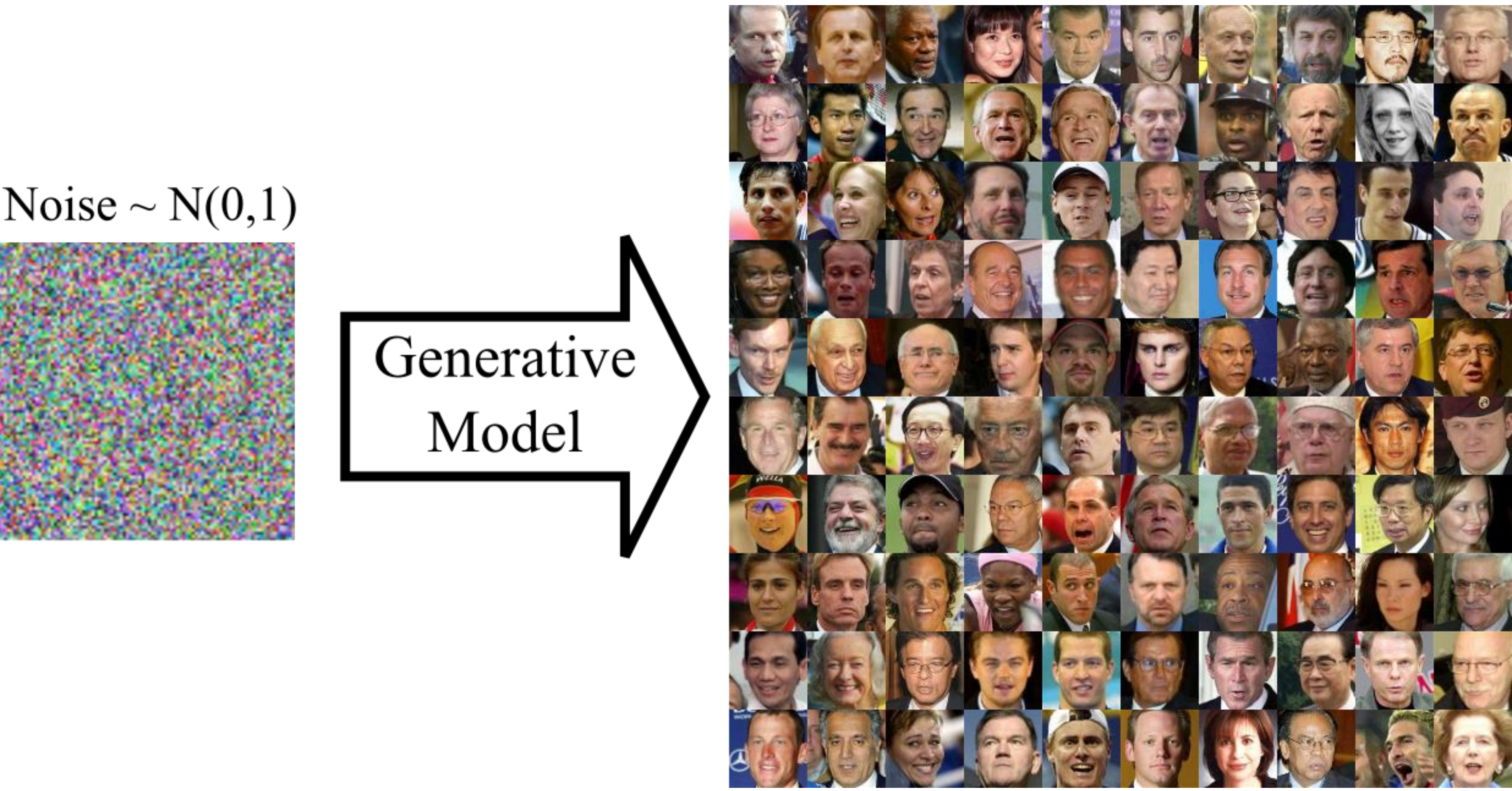


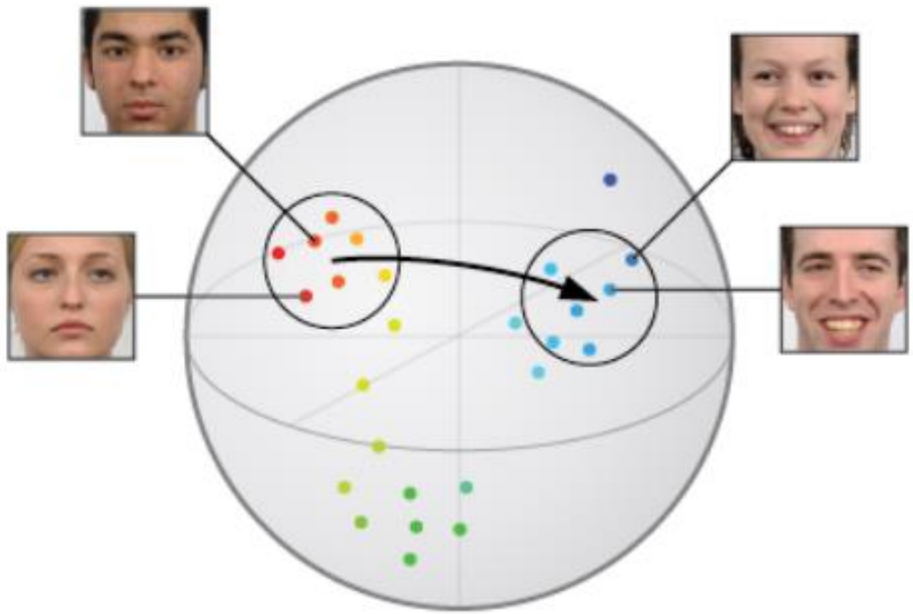
Figure from [NIPS 2016 Tutorial: Generative Adversarial Networks (I. Goodfellow)]

Generative adversarial networks (conceptual)

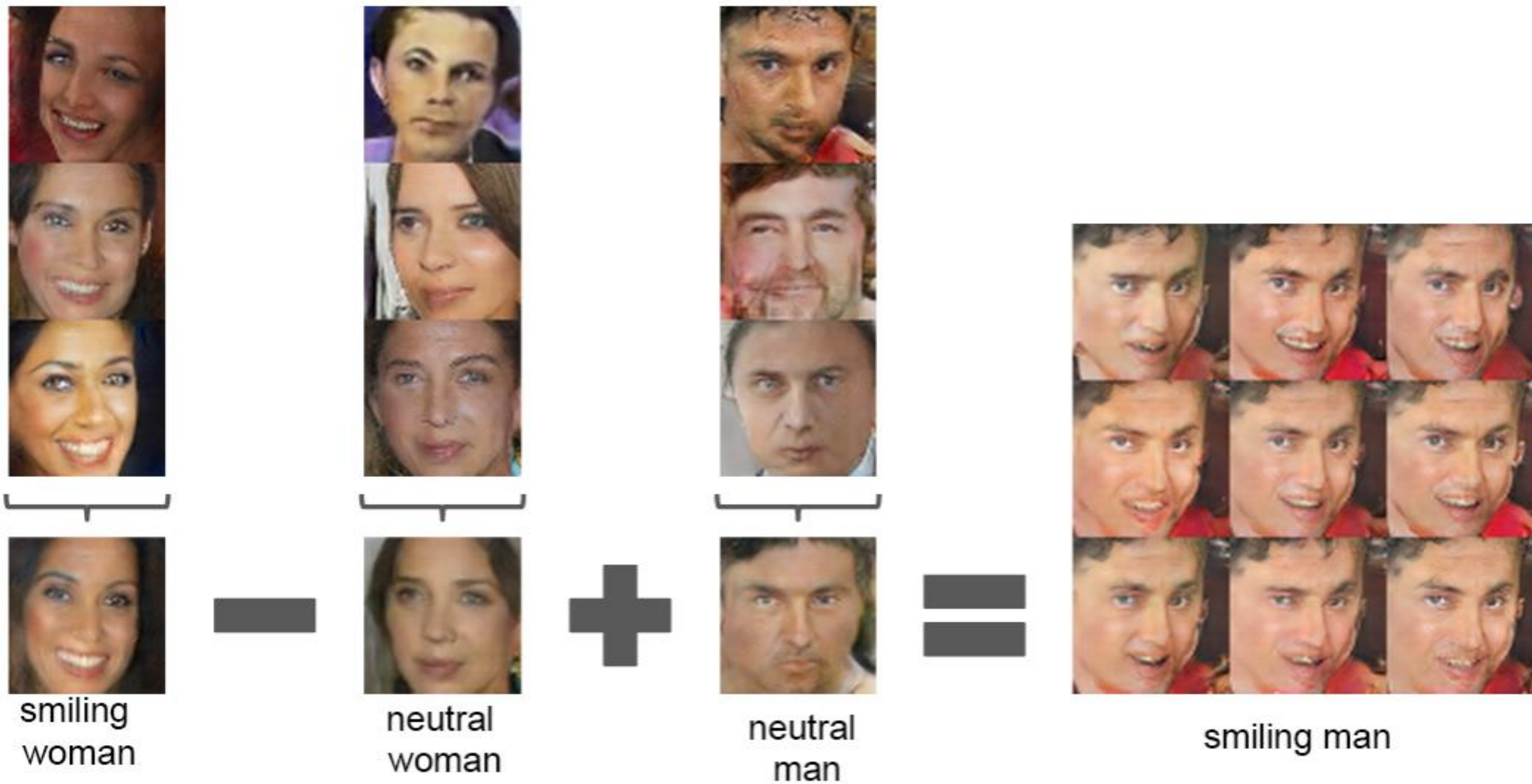


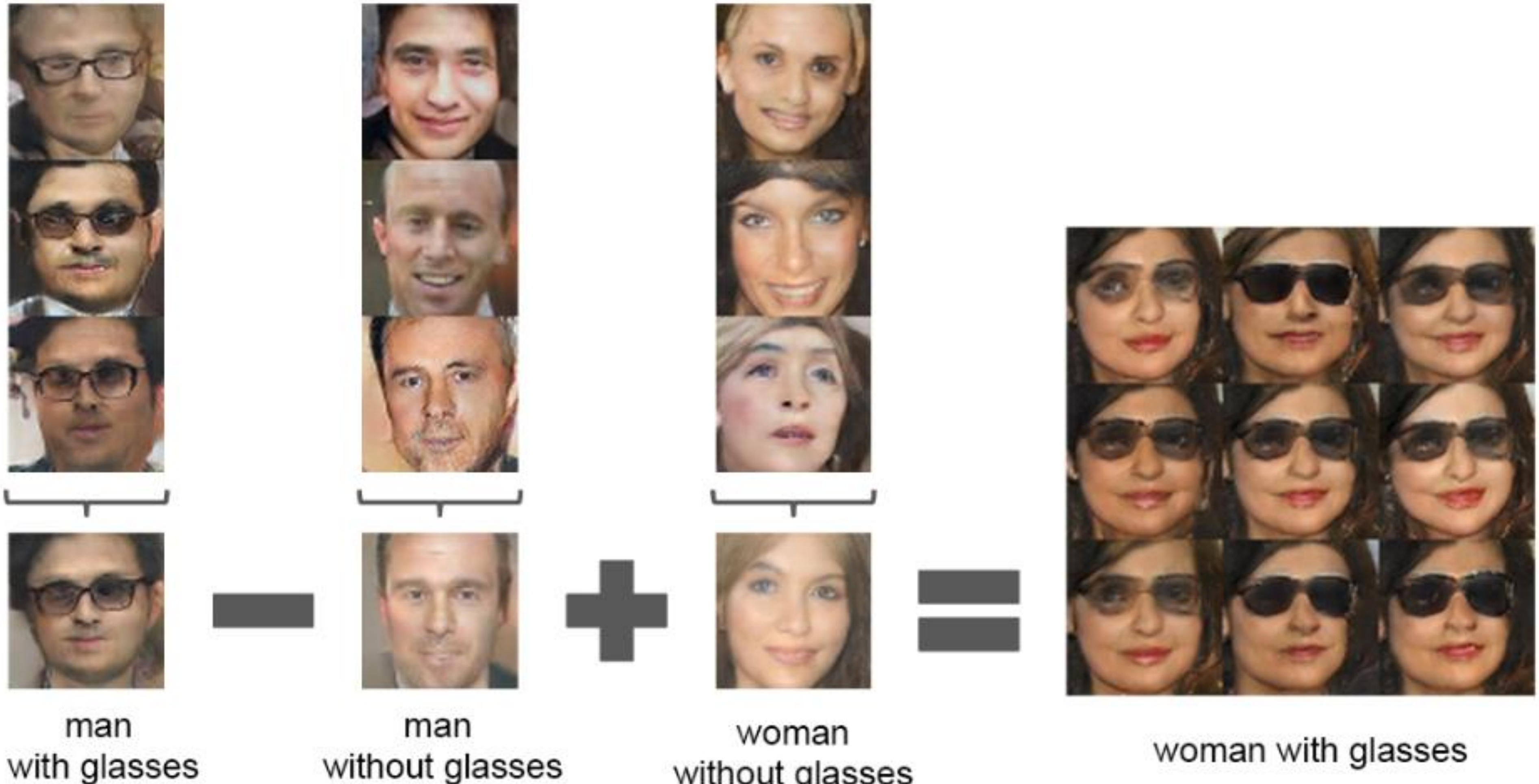

Noise ~ N(0,1)
Generative Model

If you do it right!

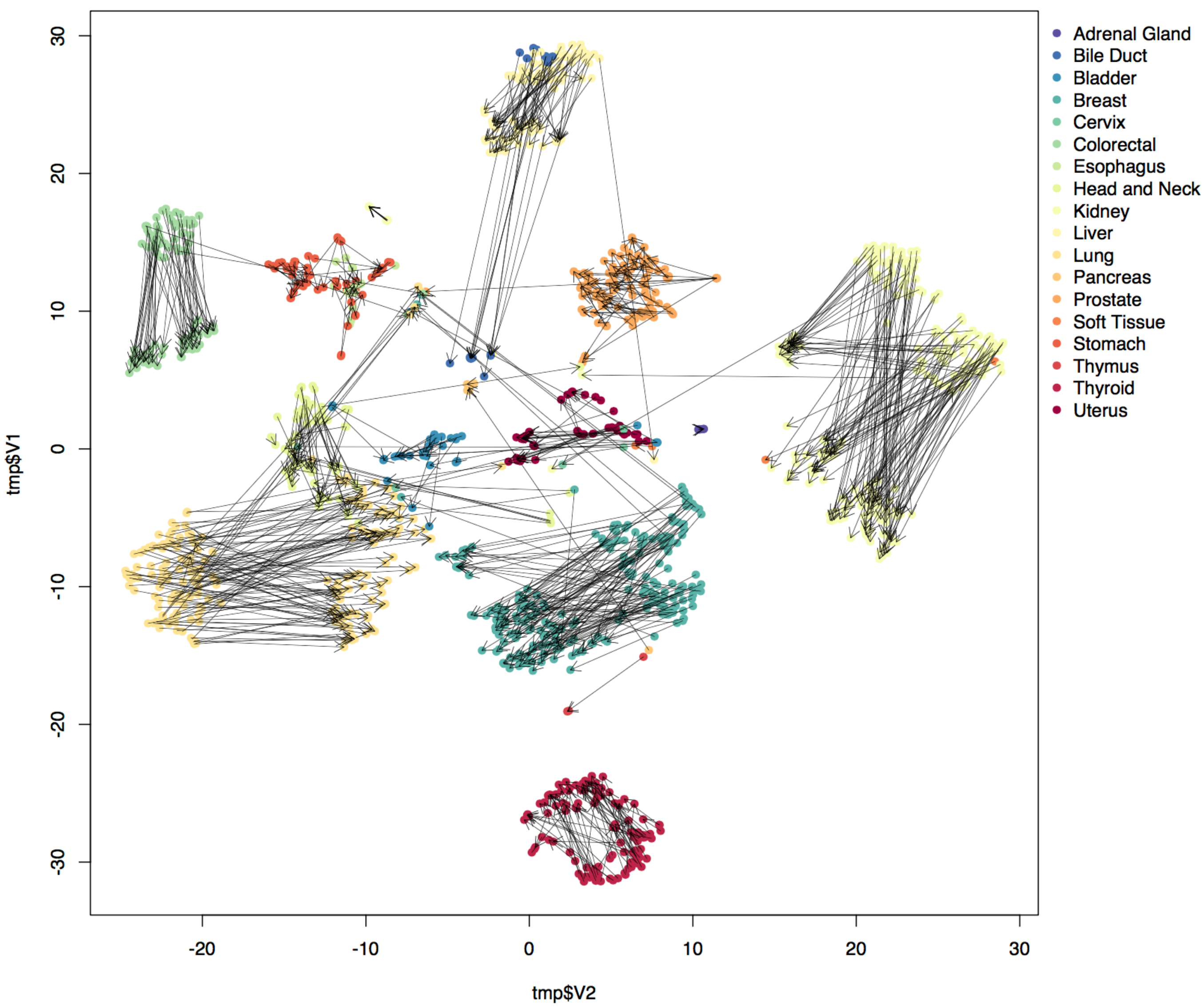


Arithmetic in the Latent Vector Space

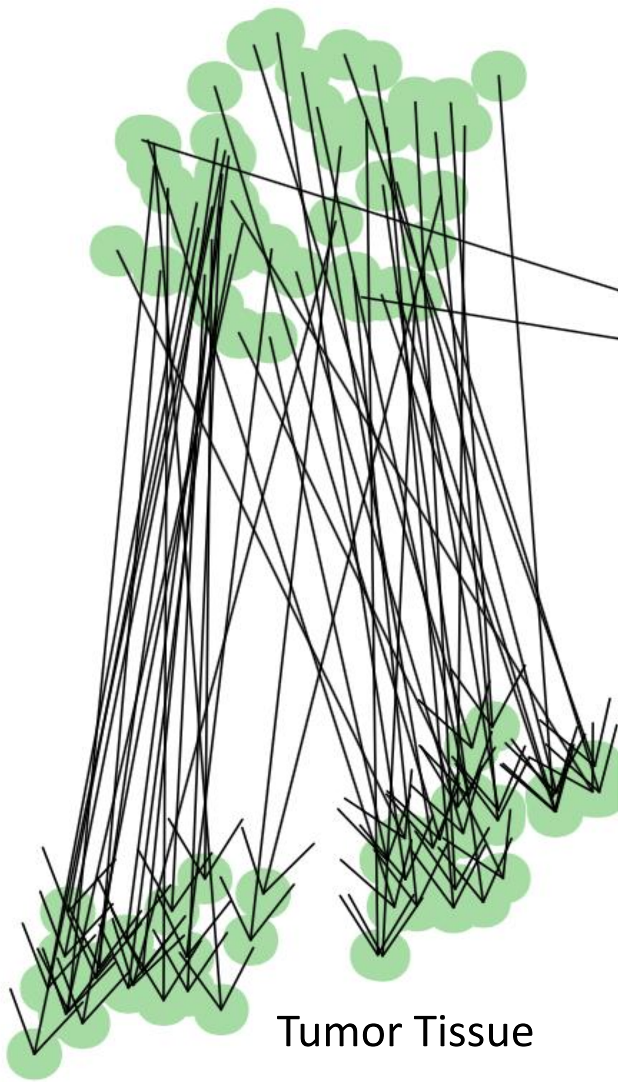

smiling woman
neutral woman
neutral man
smiling man


man
with glasses
man
without glasses
woman
without glasses
woman with glasses

t-sne Plot of Matched Normal Pairs Showing Translation in Latent Vector Space



Normal Tissue

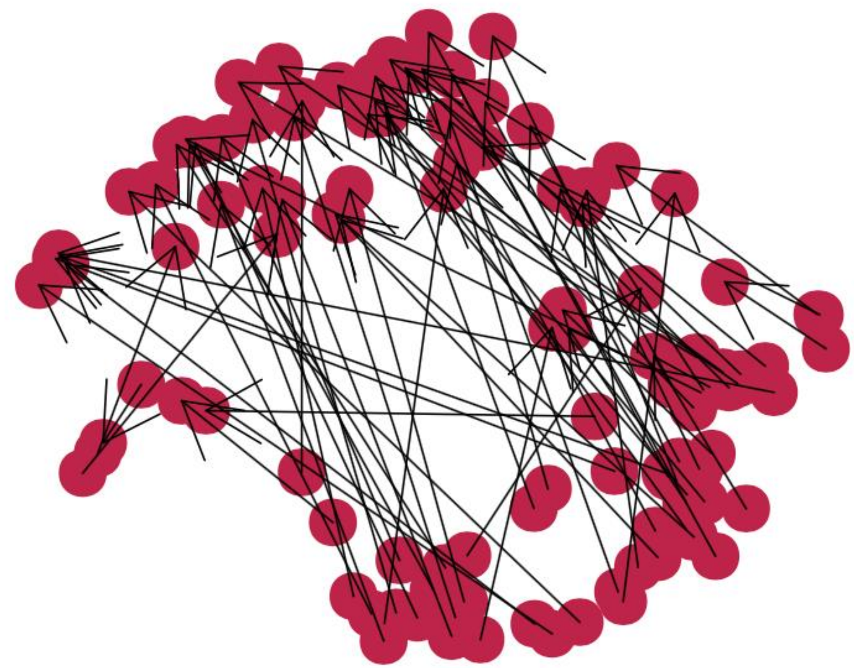


Tumor Tissue

Colon-Rectal

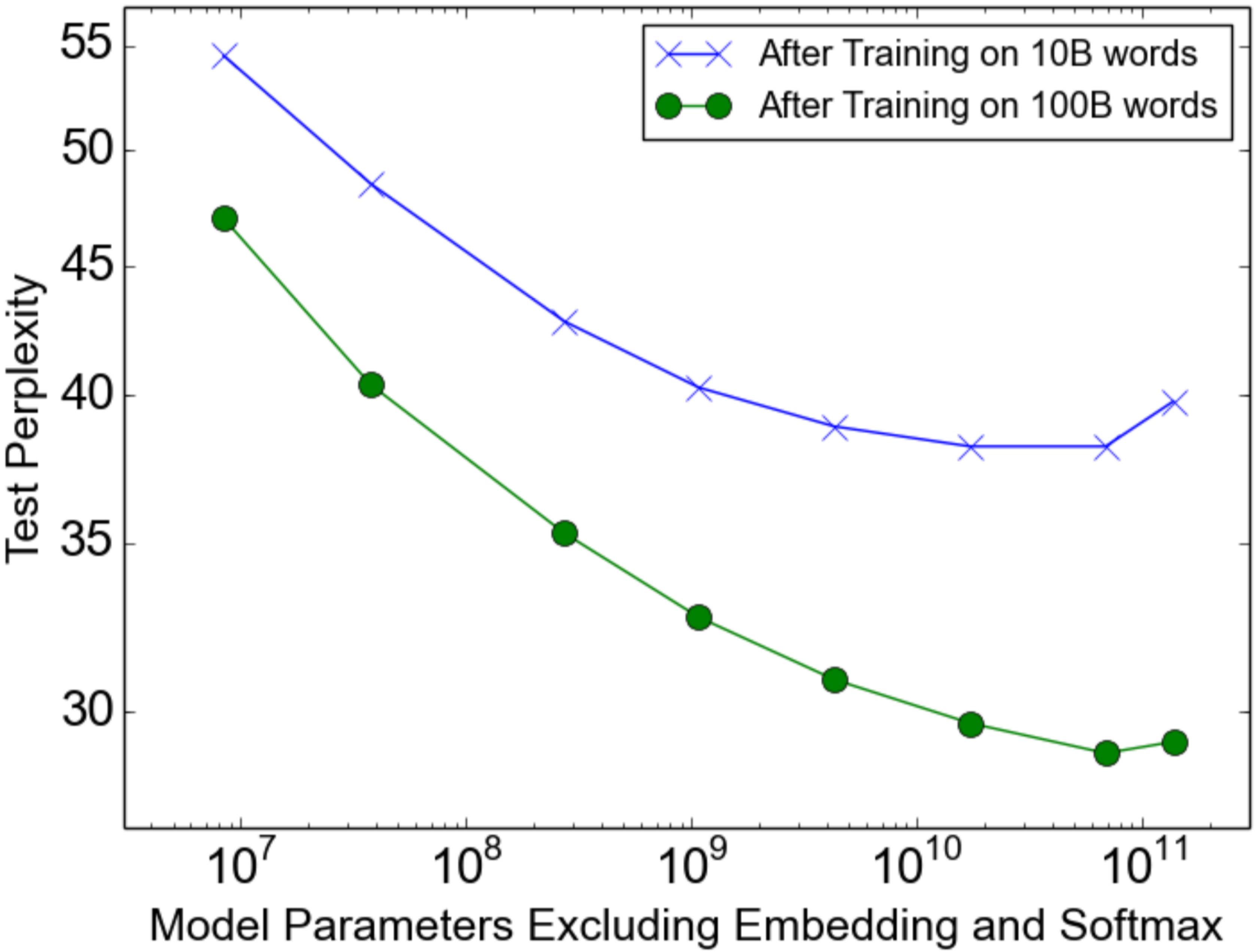
Uterus

Tumor Tissue



Normal Tissue

Really Large Networks
Multimodal Networks
Multitask Networks


After Training on 10B words
After Training on 100B words
Test Perplexity
55
50
45
40
35
30
10^7
10^8
10^9
10^{10}
10^{11}
Model Parameters Excluding Embedding and Softmax

1000x Model Capacity, 137 Billion Parameters

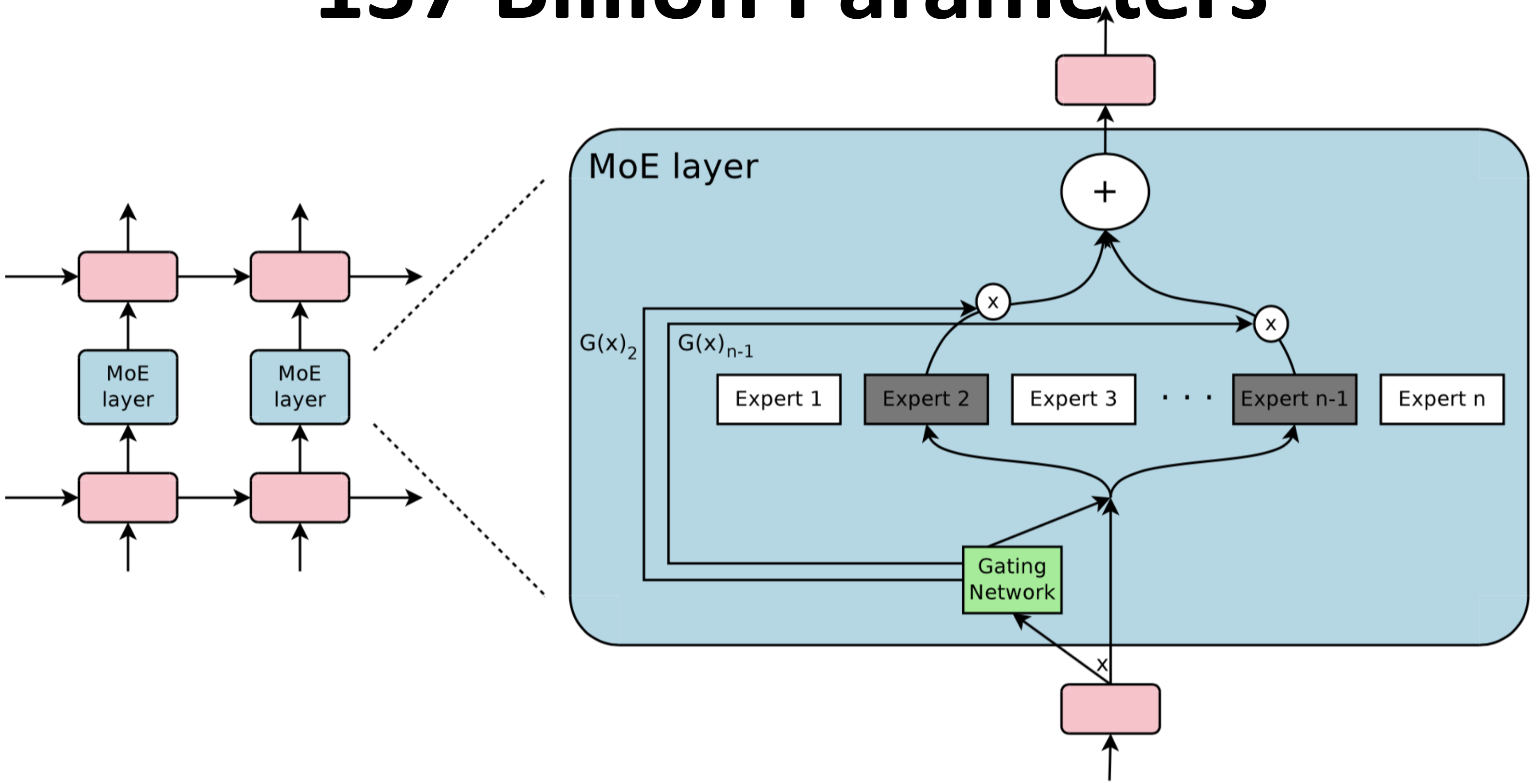


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

OUTRAGEOUSLY LARGE NEURAL NETWORKS:
THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER https://arxiv.org/abs/1701.06538

Can we create a unified deep learning model to solve tasks across multiple domains?

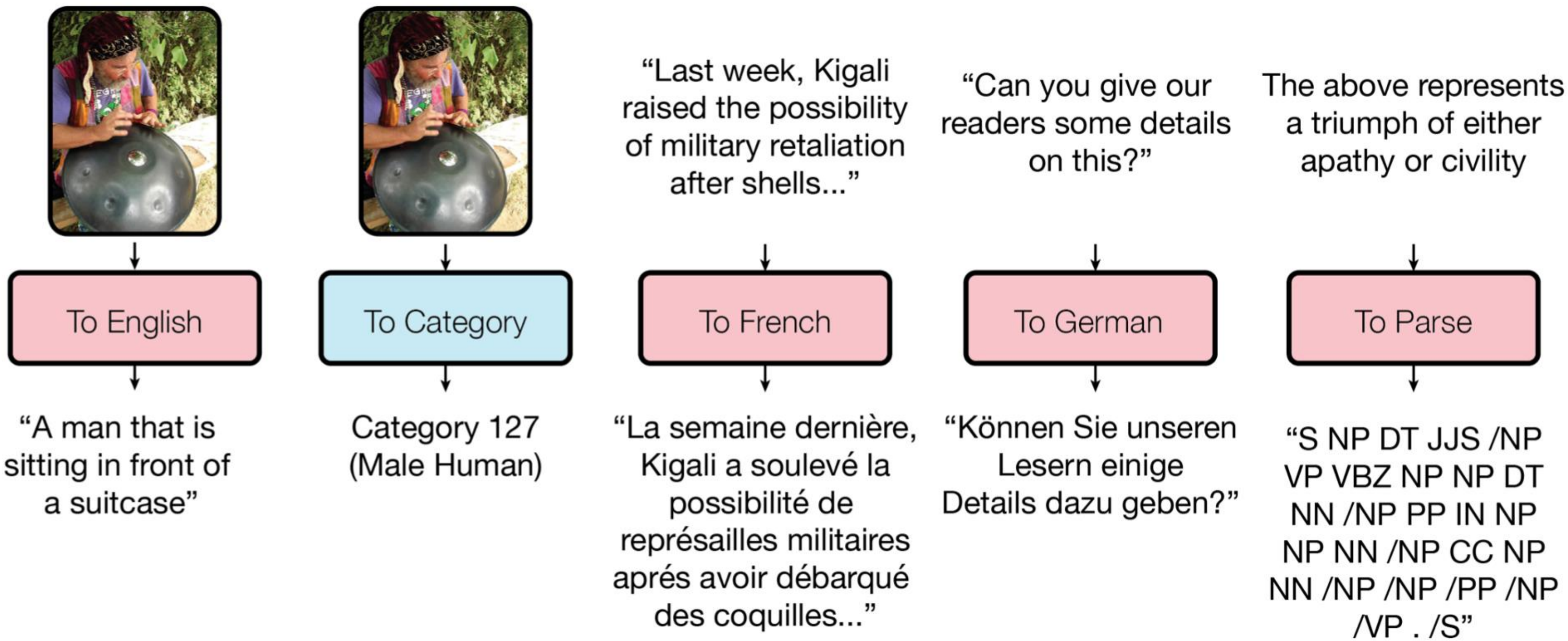


Figure 1: Examples decoded from a single MultiModel trained jointly on 8 tasks. Red depicts a language modality while blue depicts a categorical modality.

One Model To Learn Them All https://arxiv.org/abs/1706.05137

Aggregating Blocks with Gates

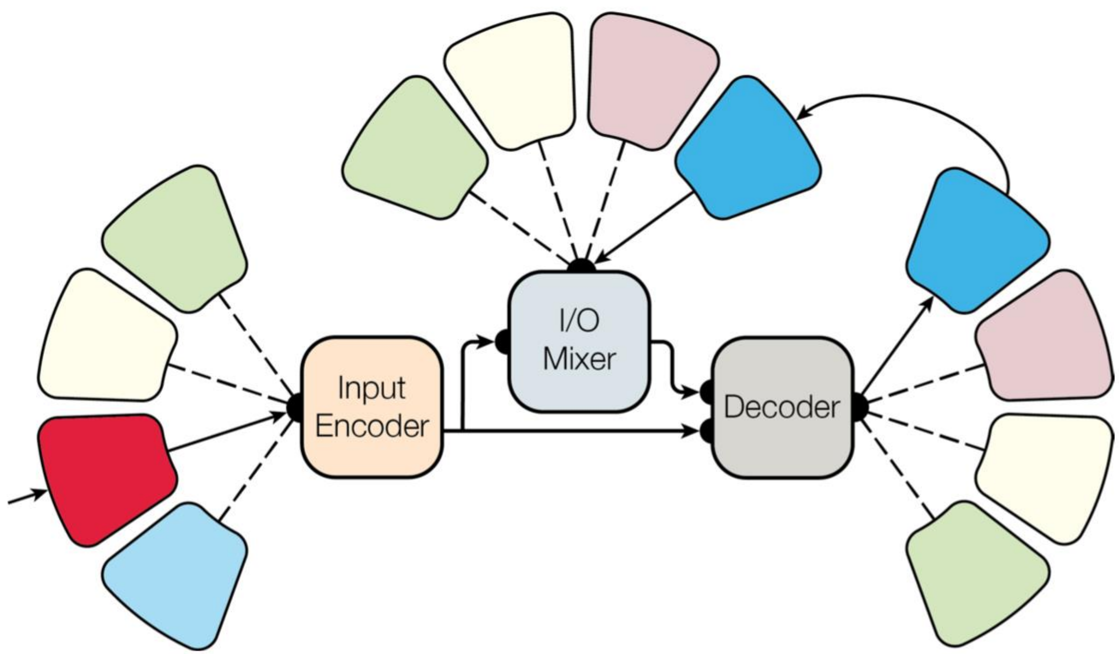


Figure 2: The MultiModel, with modality-nets, an encoder, and an autoregressive decoder.

“This leads us to conclude that mixing different computation blocks is in fact a good way to improve performance on many various tasks.”

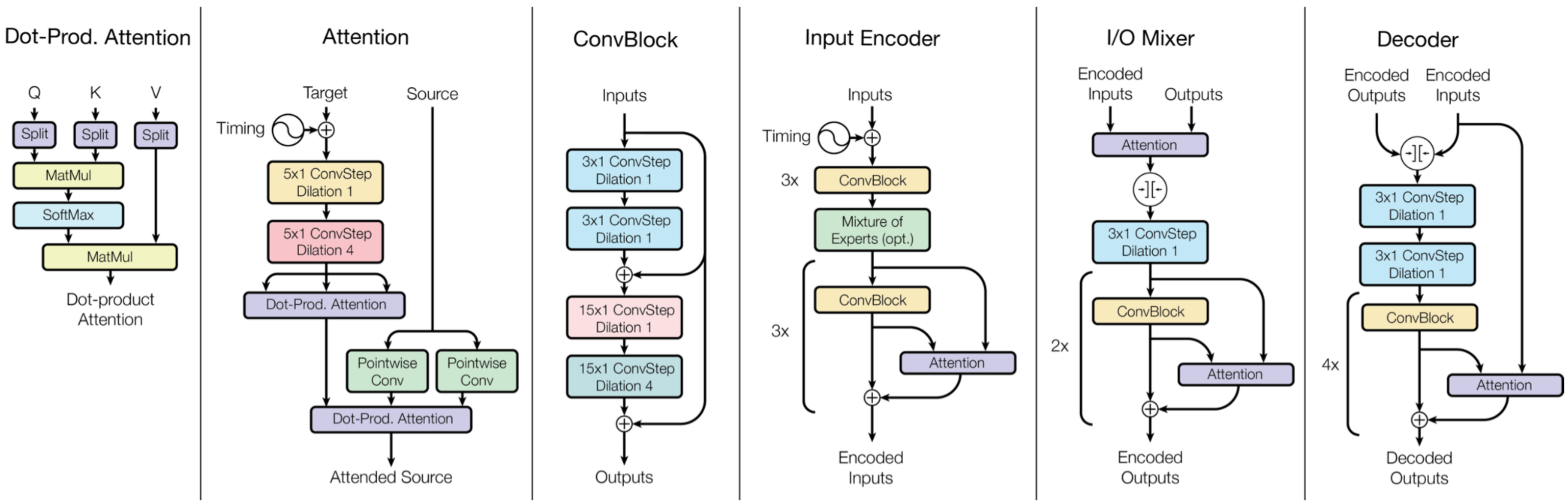


Figure 3: Architecture of the MultiModel; see text for details.

Problem	Alone			W/ ImageNet			W/ 8 Problems		
	log(ppl)	acc.	full	log(ppl)	acc.	full	log(ppl)	acc.	full
Parsing	0.20	97.1%	11.7%	0.16	97.5%	12.7%	0.15	97.9%	14.5%

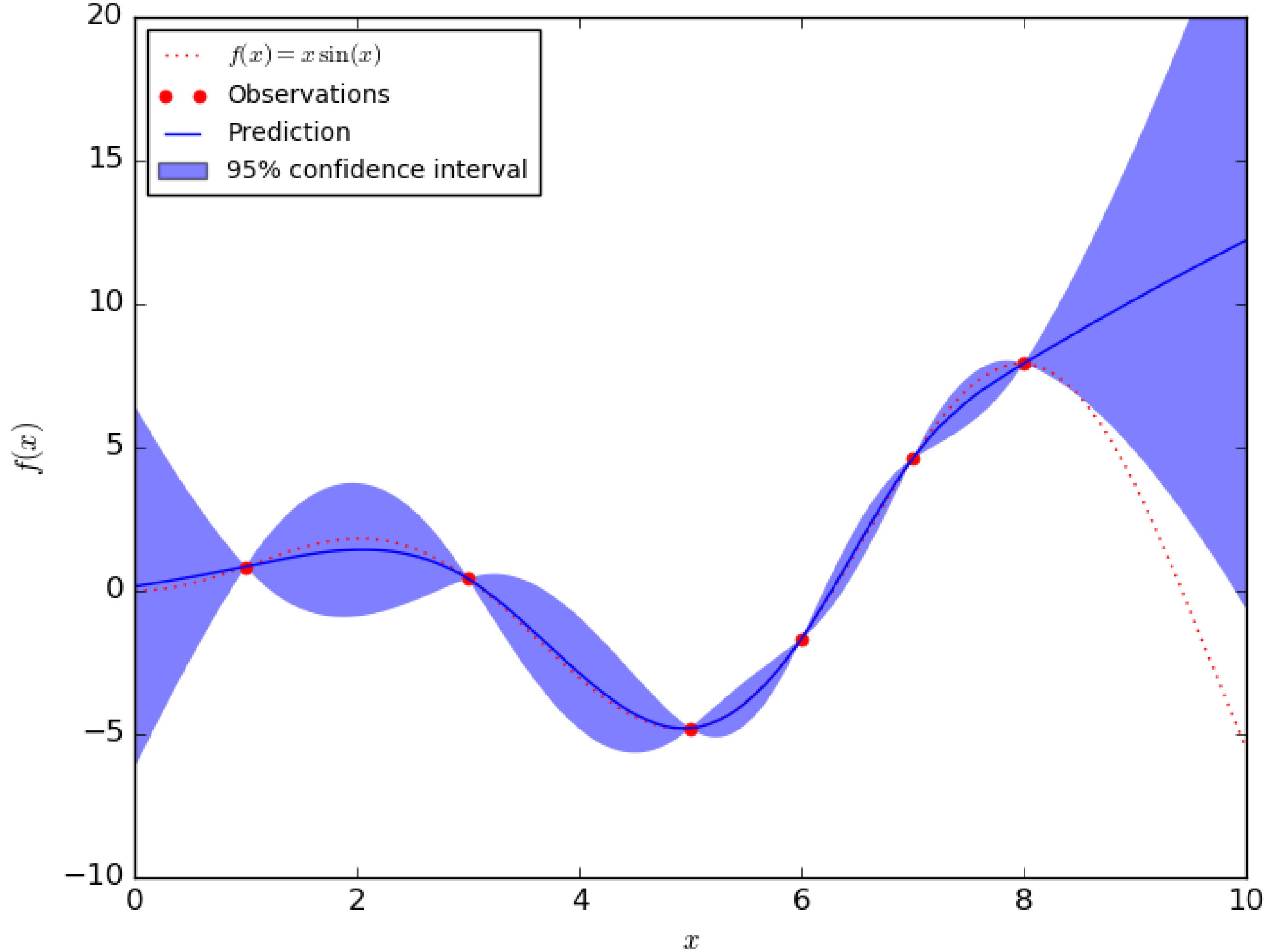
Table 3: Results on training parsing alone, with ImageNet, and with 8 other tasks. We report log-perplexity, per-token accuracy, and the percentage of fully correct parse trees.

Problem	All Blocks		Without MoE		Without Attention	
	log(perpexity)	accuracy	log(perplexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.6	67%	1.6	66%	1.6	67%
WMT EN→FR	1.2	76%	1.3	74%	1.4	72%

Table 4: Ablating mixture-of-experts and attention from MultiModel training.

Deep Learning Uncertainty Quantification

Intuition behind UQ



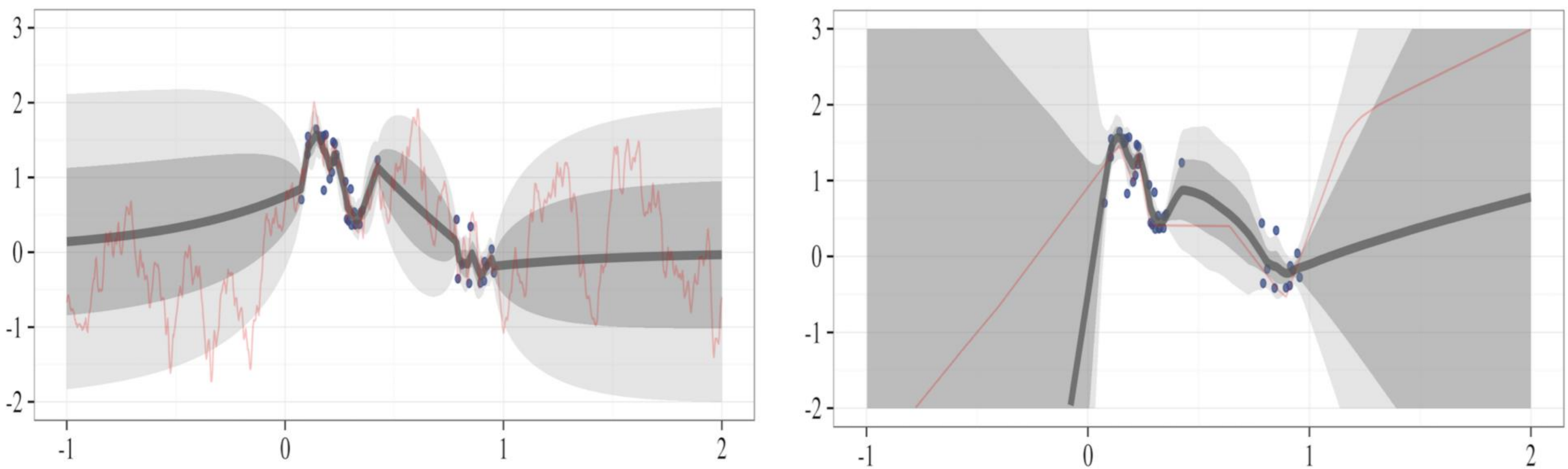
Three Approaches to DL UQ:

- Train on distributions and predict distributions

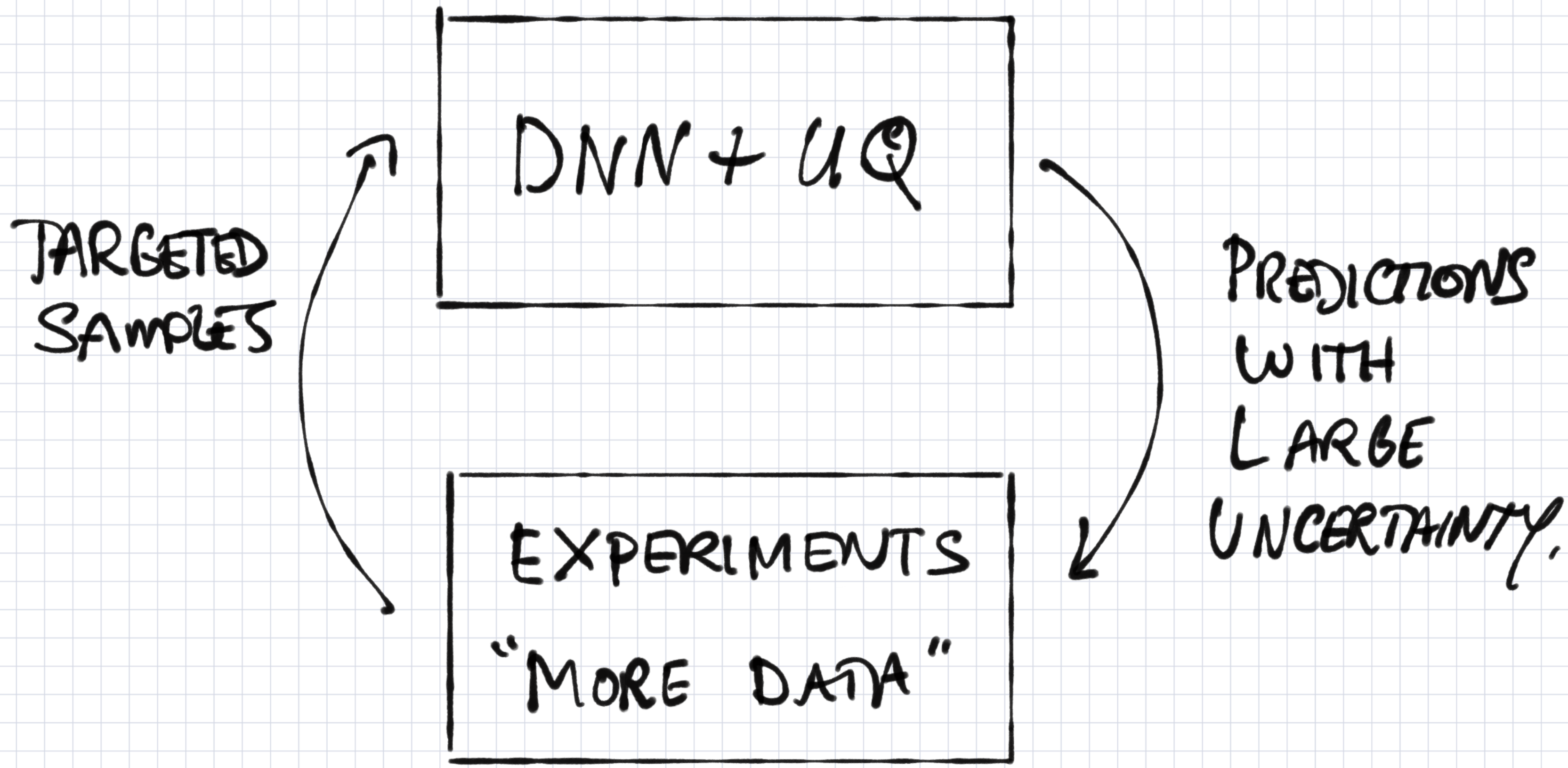
- Bootstrap with ensembles

- Dropout as a Bayesian approximation

Bootstrapping UQ in Deep Neural Networks

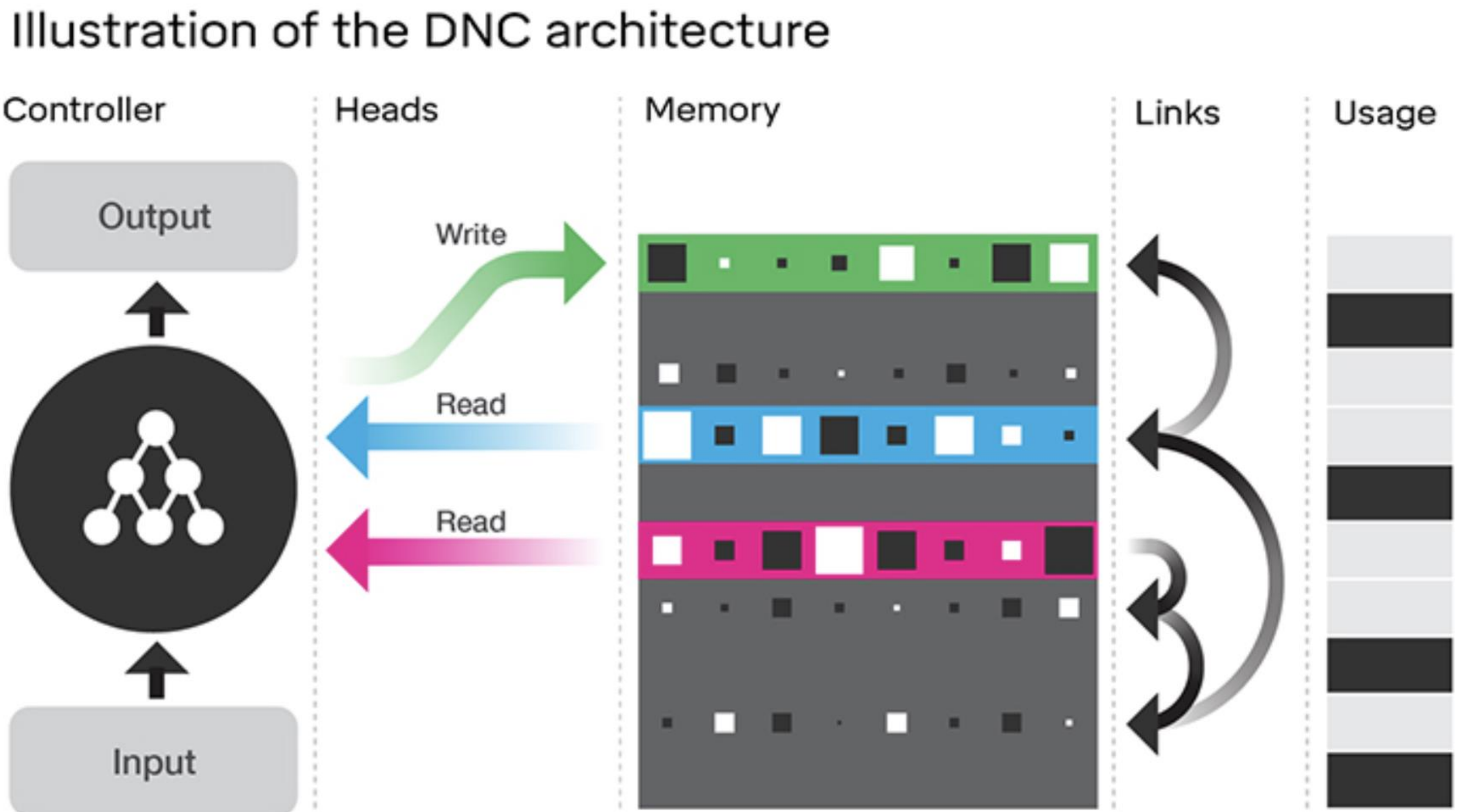


(b) Gaussian process posterior (c) Bootstrapped neural nets


DNN + UQ
TARGETED SAMPLES
PREDICTIONS WITH LARGE UNCERTAINTY.
EXPERIMENTS
"MORE DATA"

Adding New Types of Functionality

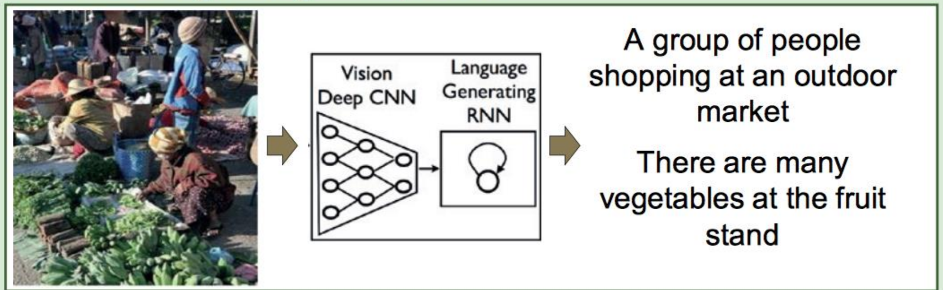
Adding Memory to Deep Networks



Credit: DeepMind

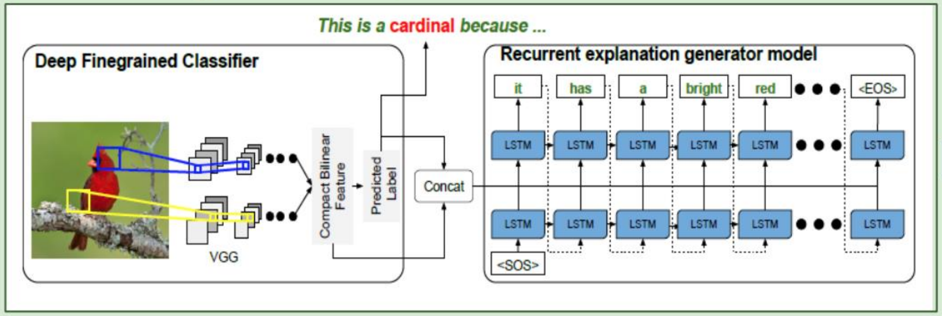
Generating Explanations (XAI)

Generating Image Captions



- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions.

Generating Visual Explanations



Researchers at UC Berkeley have recently extended this idea to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

Example Explanations

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

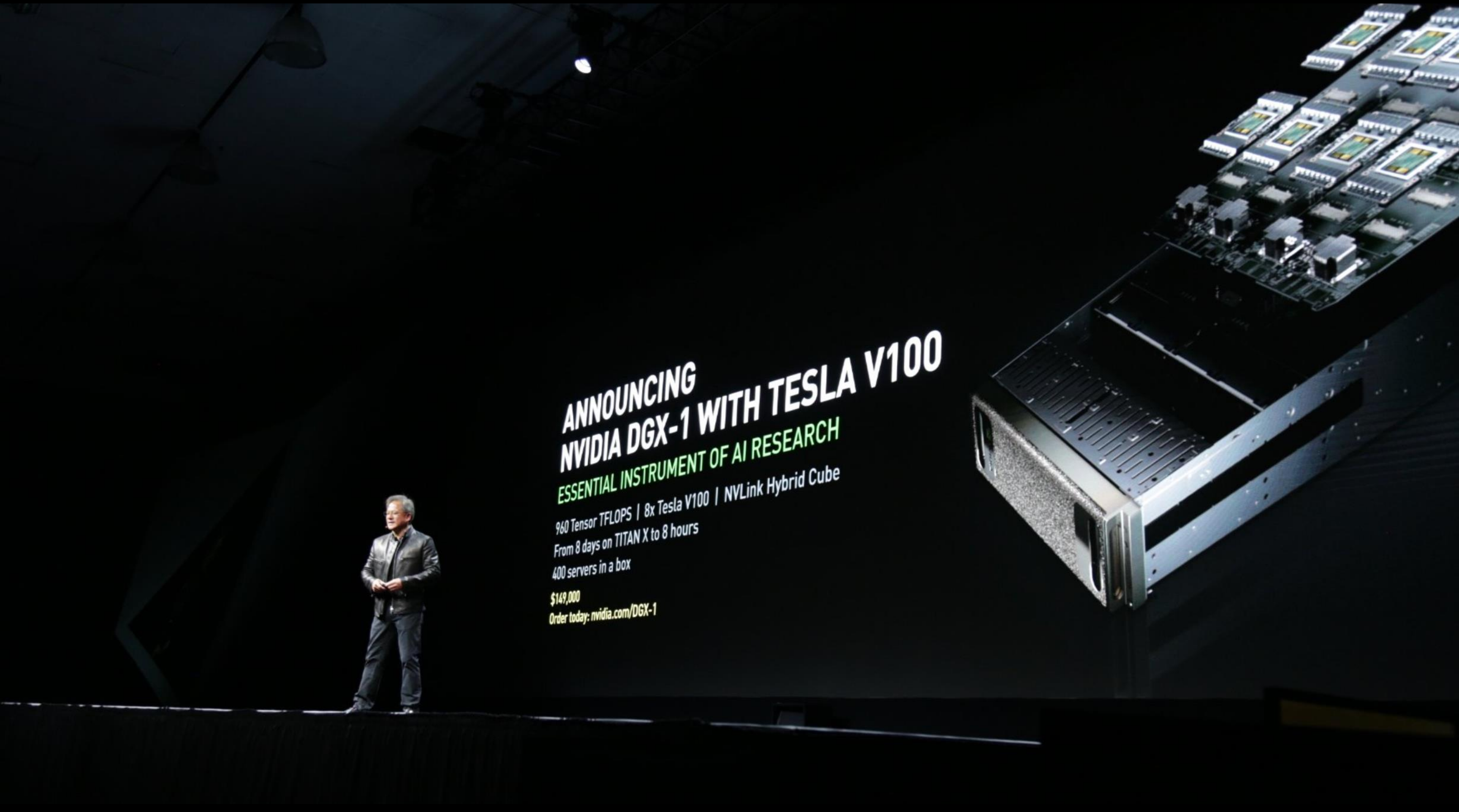
This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

Limitations

- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors

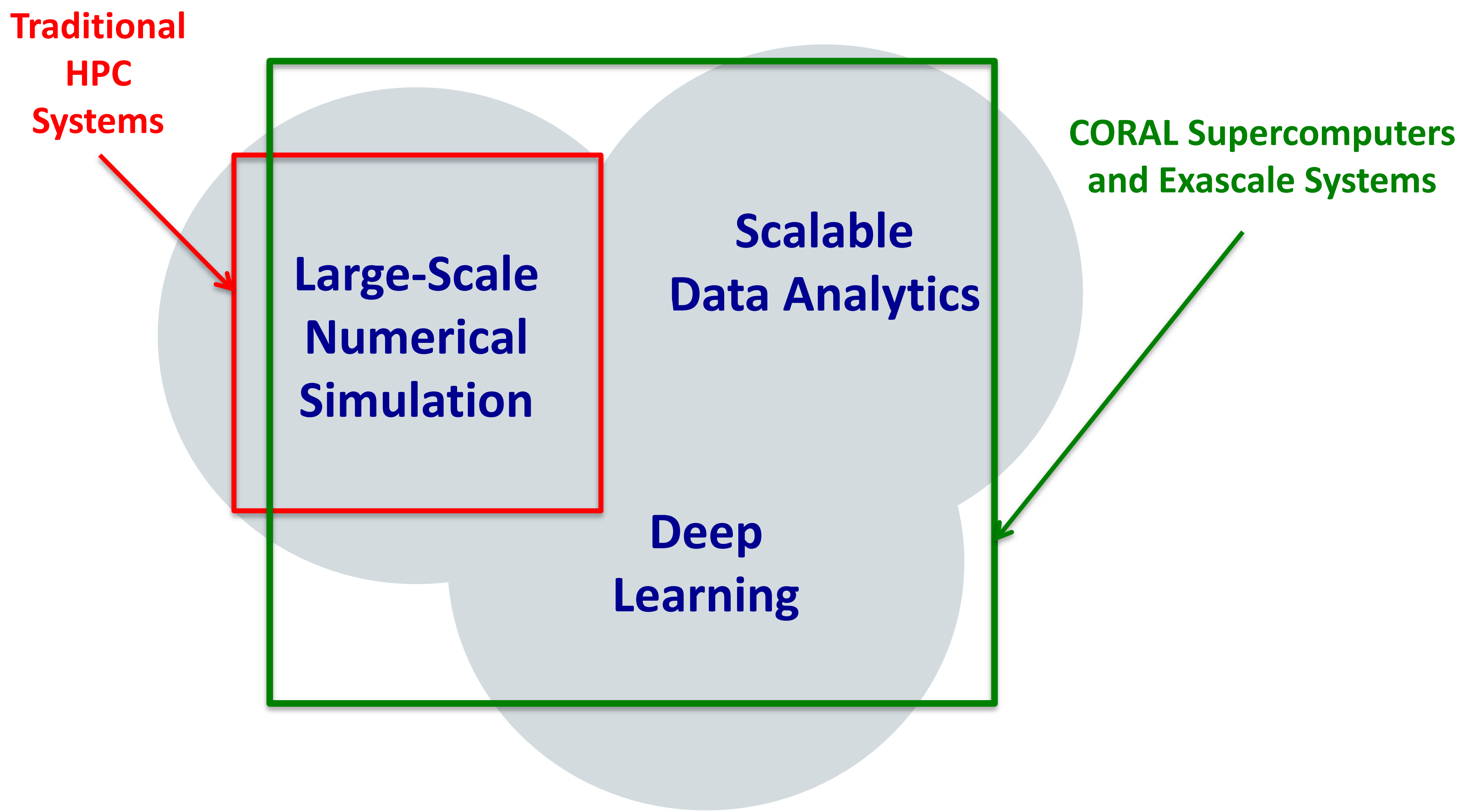
Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016

By 2020, the market for machine learning will reach $40 billion, according to market research firm IDC


ANNOUNCING
NVIDIA DGX-1 WITH TESLA V100
ESSENTIAL INSTRUMENT OF AI RESEARCH
960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube
From 8 days on TITAN X to 8 hours
400 servers in a box
$149,000
Order today: nvidia.com/DGX-1


intel
AI AND MORE ON IA
Dr. Rajeeb Hazra
Vice President, Data Center Group
General Manager, Enterprise and Government Group
ISC High Performance
ISC High Performance
ISC

Integration of Simulation, Data Analytics and Machine Learning



Differing Requirements ⟹ Convergence

Simulation Applications

- **64bit floating point**
- **Memory Bandwith**
- **Random Access to Memory**
- **Sparse Matrices**
- **Distributed Memory jobs**
- **Synchronous I/O multinode**
- **Scalability Limited Comm**
- **Low Latency High Bandwidth**
- **Large Coherency Domains help sometimes**
- **O typically greater than I**
- **O rarely read**
- **Output is data**

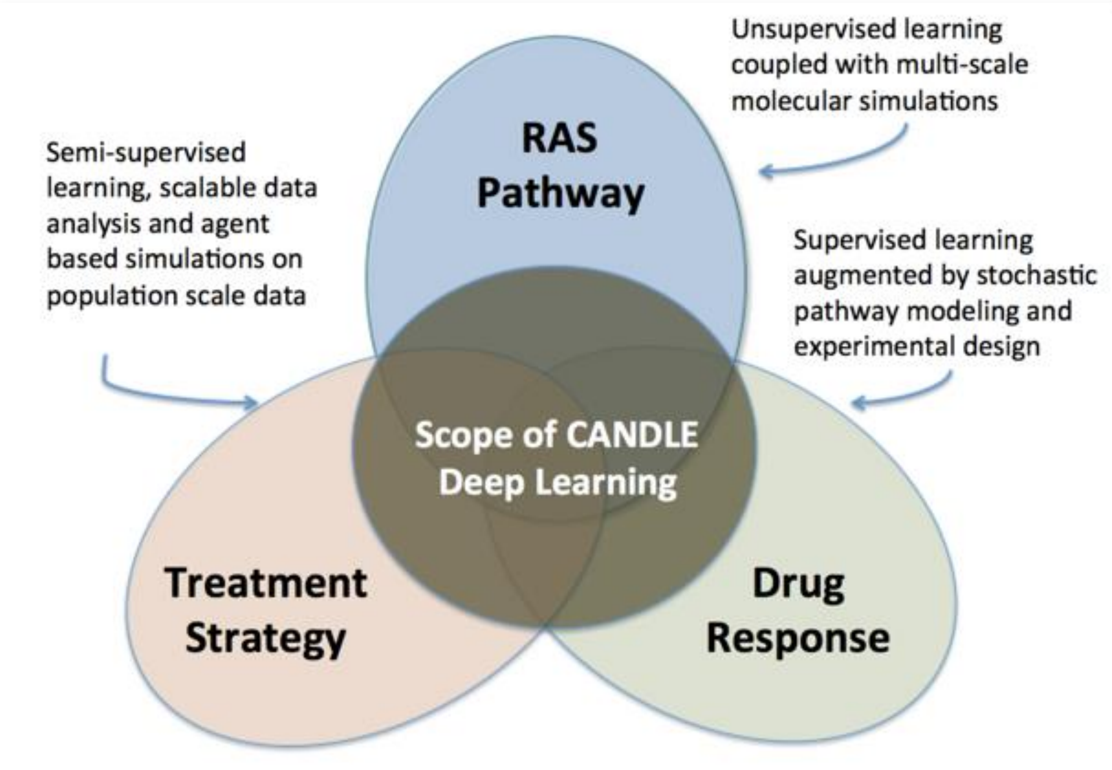
Big Data Applications

- 64 bit and Integer important
- Data analysis Pipelines
- DB including No SQL
- MapReduce/SPARK
- Millions of jobs
- I/O bandwidth limited
- Data management limited
- Many task parallelism
- Large-data in and Large-data out
- I and O both important
- O is read and used
- Output is data

Deep Learning Applications

- Lower Precision (32 bit)
- FMAC @ 32 okay
- Inferencing can be 8 bit
- Scaled integer possible
- Training dominates dev
- Inference dominates pro
- Reuse of training data
- Data pipelines needed
- Dense FP typical SGEMM
- Small DFT, CNN
- Ensembles and Search
- Single Models Small
- I more important than O
- Output is models

ECP-CANDLE : CANcer Distributed Learning Environment



CANDLE Goals

Develop an exscale deep learning environment for cancer

Building on open source Deep learning frameworks

Optimization for CORAL and exascale platforms

Support all three pilot project needs for deep dearning

Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects

NATIONAL CANCER INSTITUTE

NVIDIA

IBM

CRAY

intel

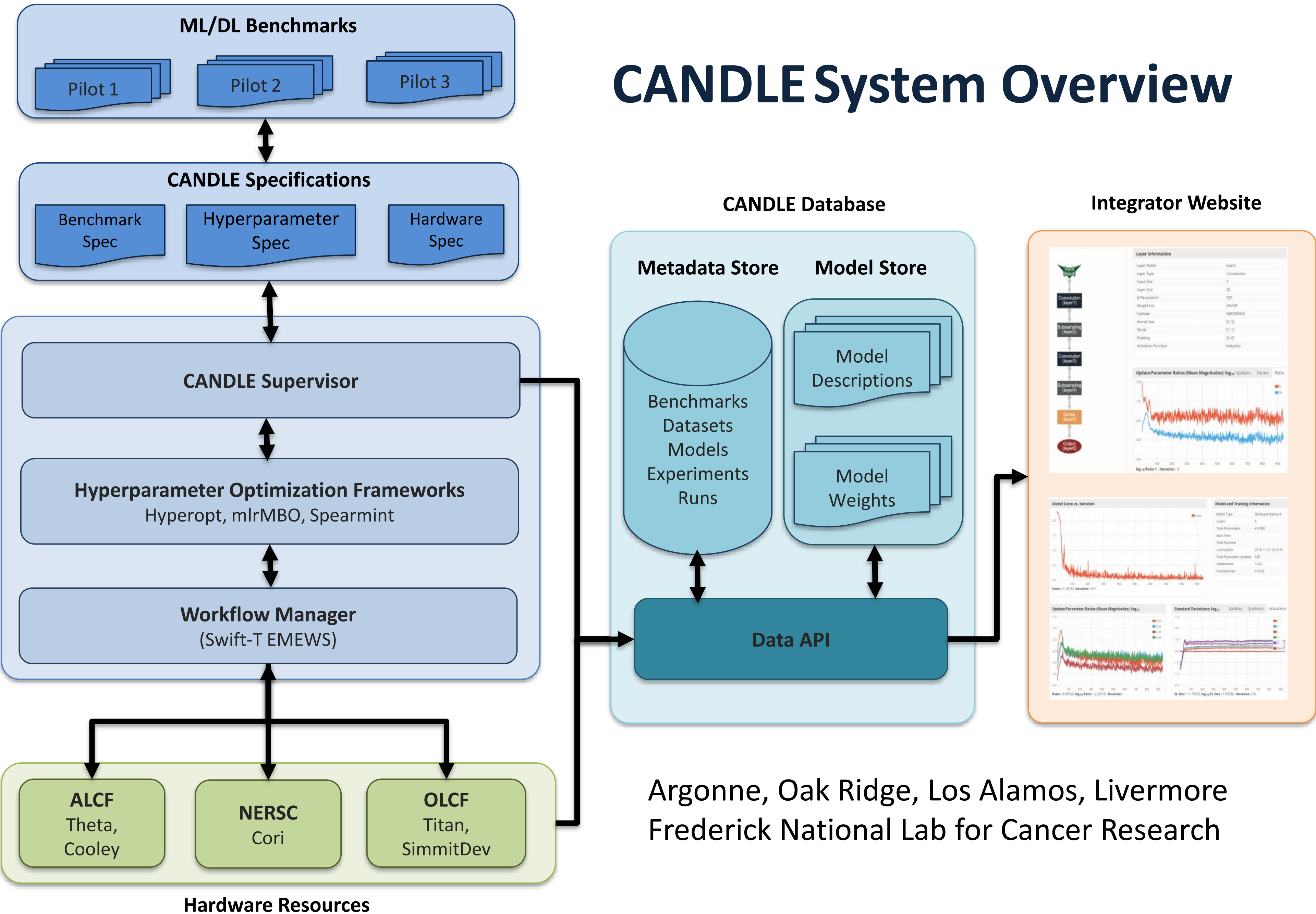
AMD

ECP EXASCALE COMPUTING PROJECT

Deep Learning in Cancer $\Longrightarrow$ many Methods

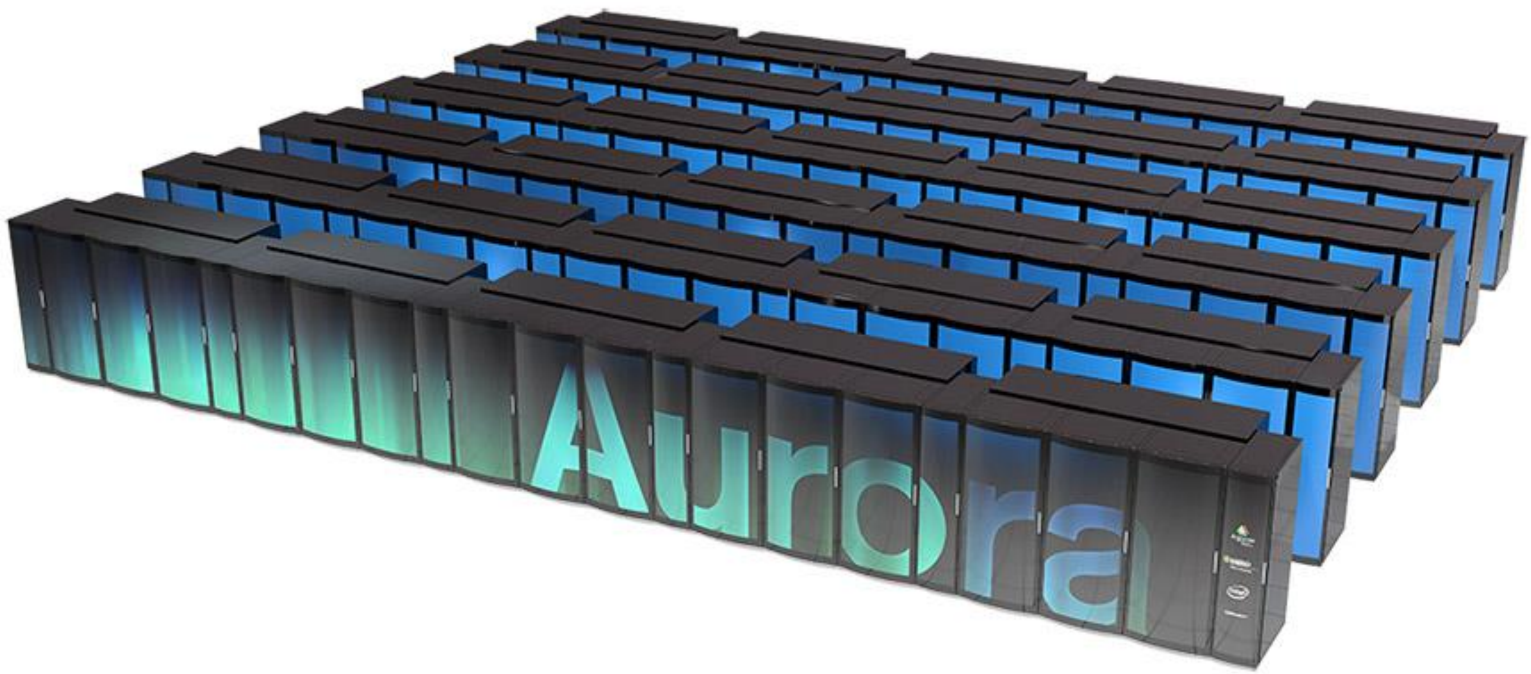
- **AutoEncoders** – learning data representations for classification and prediction of drug response, molecular trajectories
- **VAEs and GANs** – generating data to support methods development, data augmentation and feature space algebra, drug candidate generation
- **CNNs** – type classification, drug response, outcomes prediction, drug resistance
- **RNNs** – sequence, text and molecular trajectories analysis
- **Multi-Task Learning** – terms (from text) and feature extraction (data), data translation (RNAseq <-> uArray)

CANDLE System Overview



Argonne, Oak Ridge, Los Alamos, Livermore
Frederick National Lab for Cancer Research

Aurora 2021 (A21)
The first US Exascale System



Architecture supports three ways of computing

- **Large-scale Simulation (PDEs, traditional HPC)**
- **Data Intensive Applications (science pipelines)**
- **Deep Learning and Emerging Science AI**

Application Targets for Exascale

Simulation Applications

- Materials Science
- Cosmology
- Molecular Dynamics
- Nuclear Reactor Modeling
- Combustion
- Quantum Computer Simulation
- Climate Modeling
- Power Grid
- Discrete Event Simulation
- Fusion Reactor Simulation
- Brain Simulation
- Transportation Networks

Big Data Applications

- APS Data Analysis
- HEP Data Analysis
- LSST Data Analysis
- SKA Data Analysis
- Metagenome Analysis
- Battery Design Search
- Graph Analysis
- Virtual Compound Library
- Neuroscience Data Analysis
- Genome Pipelines

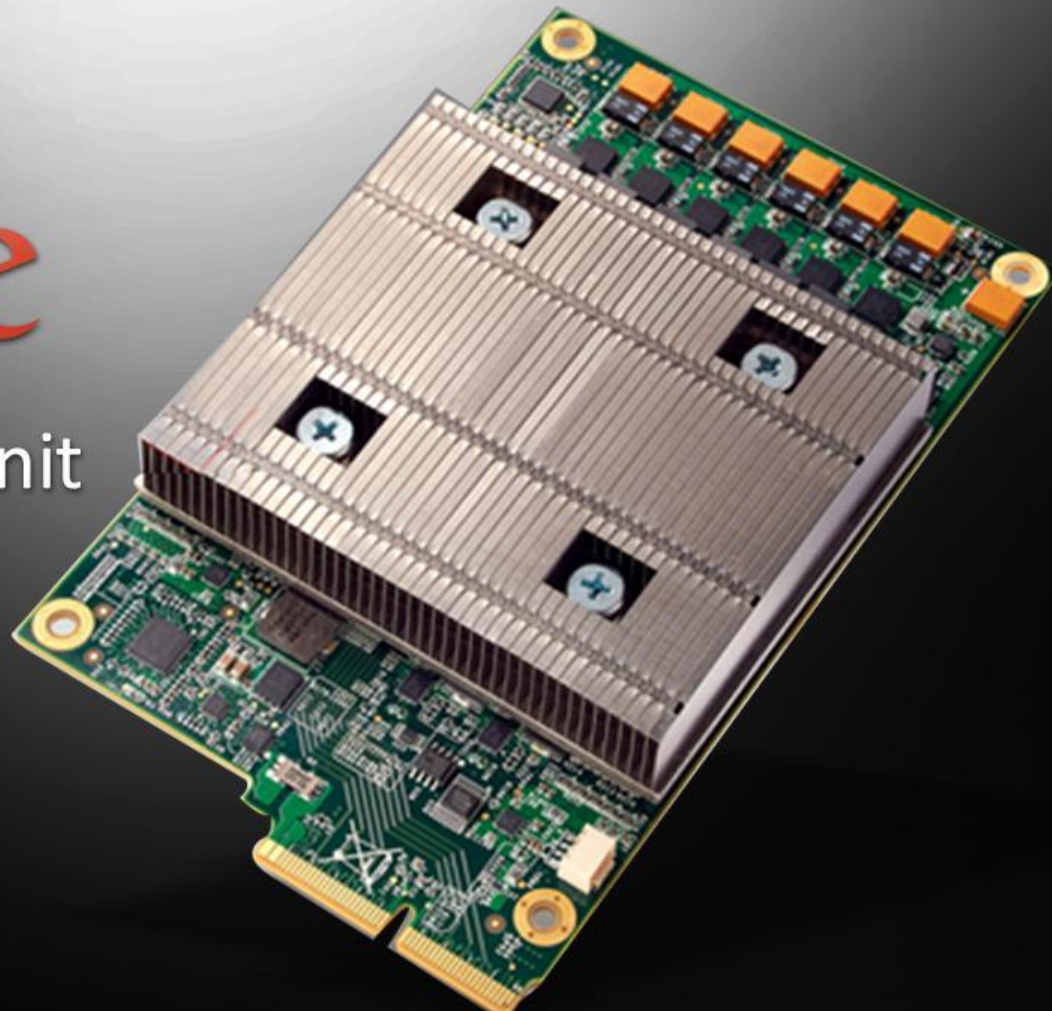
Deep Learning Applications

- Drug Response Prediction
- Scientific Image Classification
- Scientific Text Understanding
- Materials Property Design
- Gravitational Lens Detection
- Feature Detection in 3D
- Street Scene Analysis
- Organism Design
- State Space Prediction
- Persistent Learning
- Hyperspectral Patterns

Specialized hardware is emerging that will be many times (100x) the performance of general purpose CPU and GPU designs

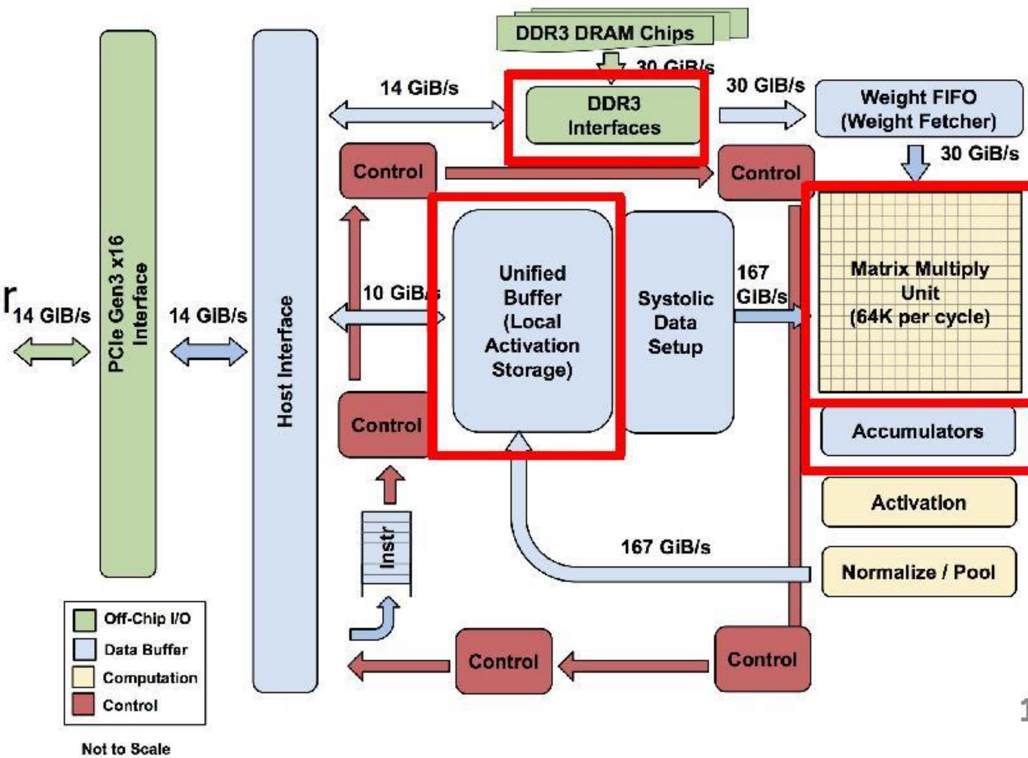
Google

Tensor Processing Unit



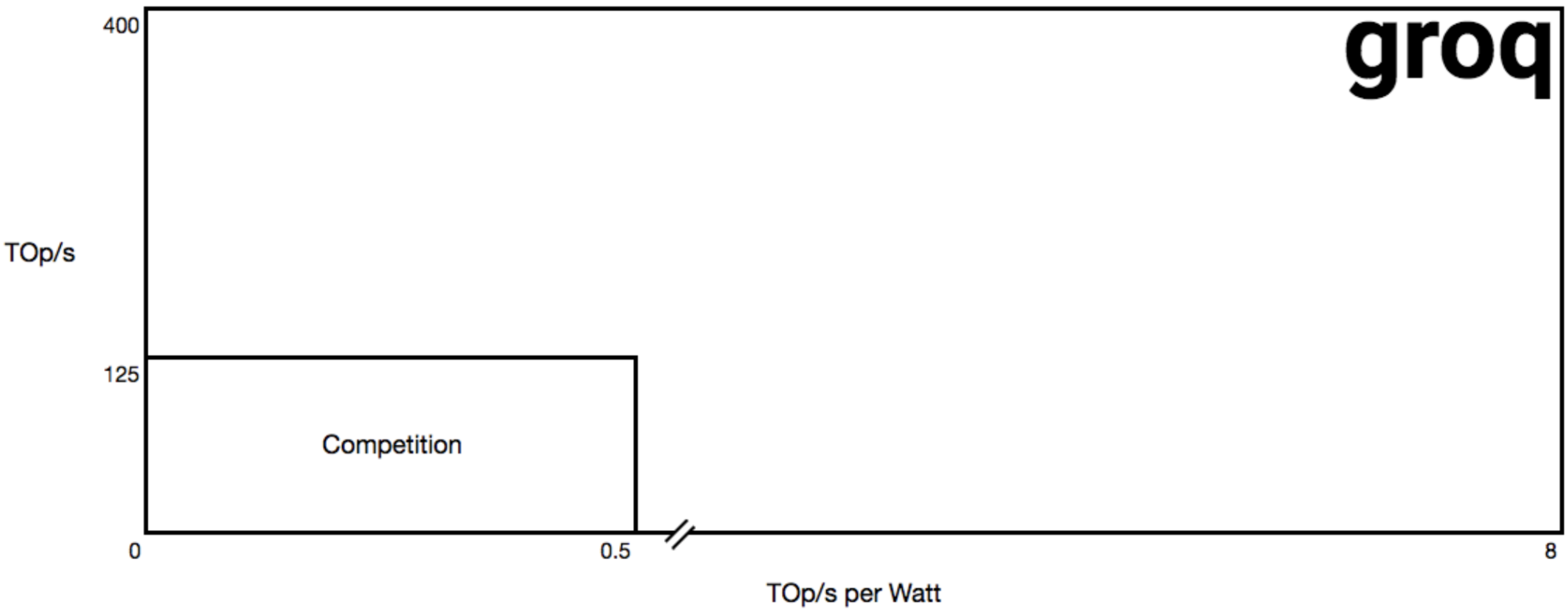
TPU: High-level Chip Architecture

- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU
- Two 2133MHz DDR3 DRAM channels
- 8 GiB of off-chip weight DRAM memory



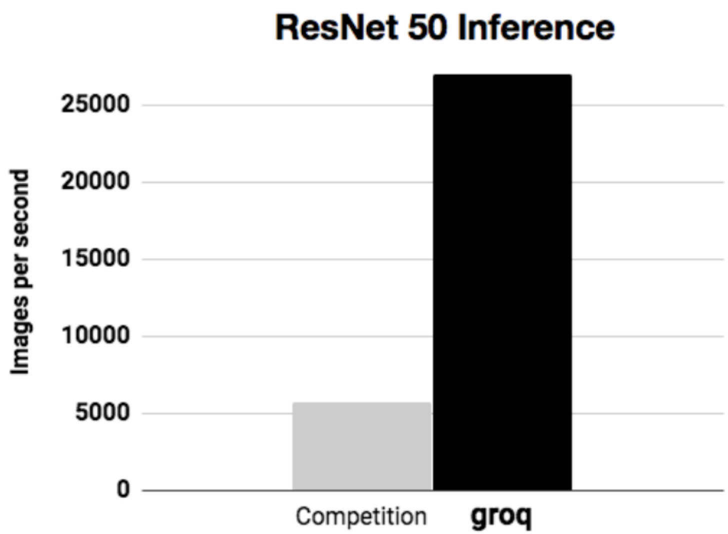

A
B
C
D

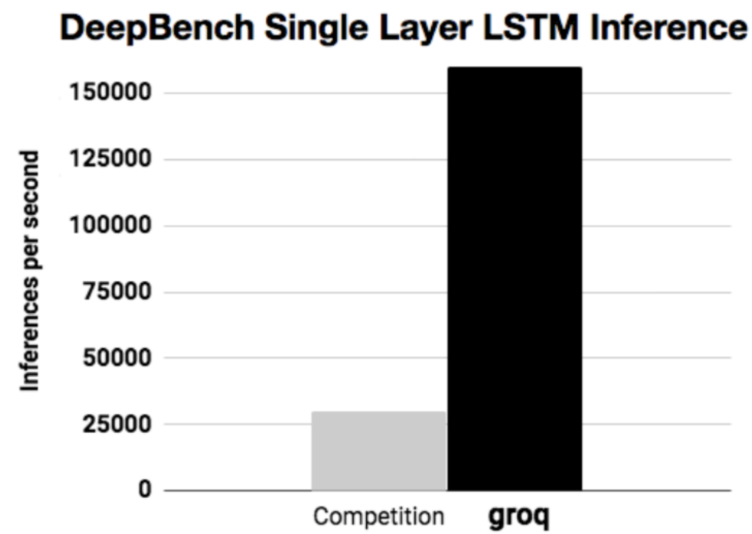
Groq (grok)



Our first machine learning product. Single chip. 2018.

November 9, 2017





2018 performance estimates. Single chip. <1ms latency.

November 21, 2017

Wave Computing

Wave's Compute Appliance is Redefining How Machine Learning is Done

- 2.9 PetaOps per second of performance
- More than 2TB of high-speed memory
- Up to 256,000 processing elements per appliance
- Scales up to four appliances per data center node
- Initially supporting TensorFlow

ABOUT THE WAVE COMPUTE APPLIANCE

Specifications for each Wave Compute Appliance

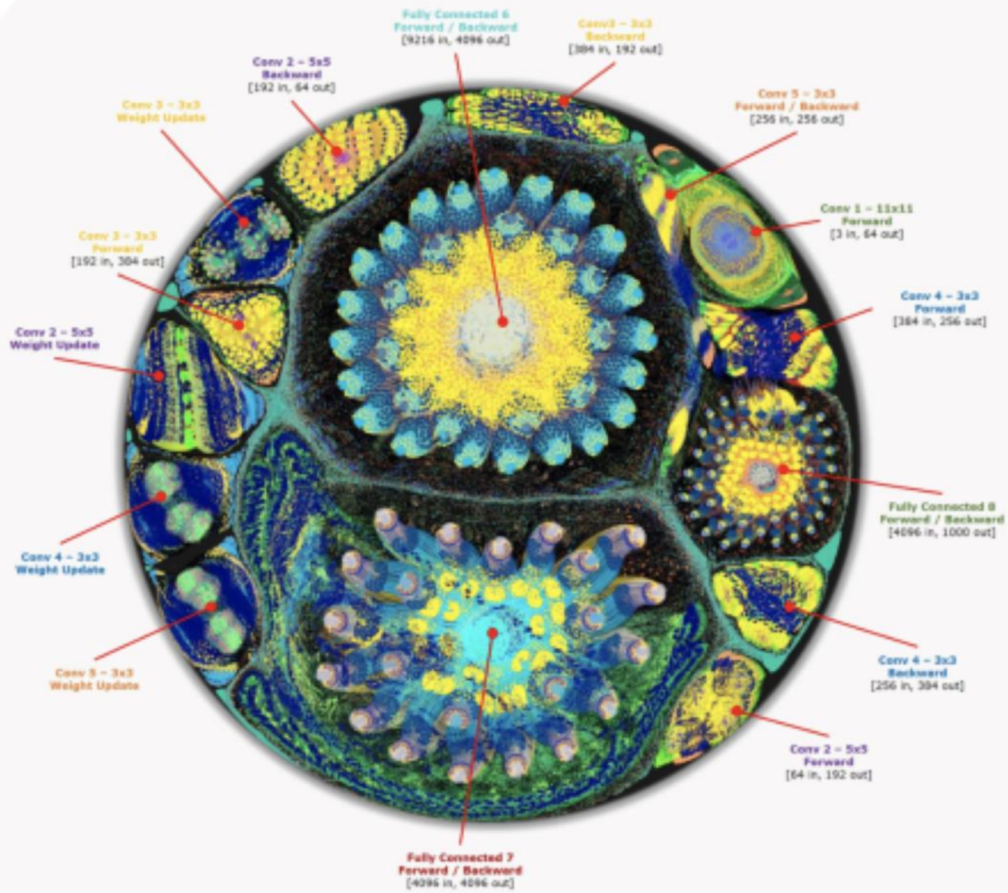
Performance	Performance/computer (peak)	2.9 PetaOPS/second
	Performance/node (peak)	11.6 PetaOPS/second
	Dataflow Processing Elements (PE's)	Up to 256,000 (16,000 PE's per Wave DPU chip)
Scalability	Wave machine learning computers per data center node	Up to 4 computers delivering 1,000,000 PE's
Memory	High-speed memory	128 GB HMC DRAM
	SSD storage	16 TB
	Bulk storage	2 TB DDR4 DRAM
Connections	Data center backbone connection	10 GbE or 40 GbE
	High-speed inter-computer communication within a single data center node	Wave's proprietary communication system that connects up to 4 computers within a single data center node
Physical	Data center form factor	Each Wave computer comes in a 3U form factor; up to 4 computers can be added per data center node
	Dimensions per each 3U computer	866D x 444W x 131H (mm)
	Operating temperature	10° – 35° C
Software	Machine learning framework	TensorFlow (initially)
	Operating system for Wave Session Manager server	Linux Server
	Library	WaveFlow Agent Library
	Development toolkit	WaveFlow SDK
	Data runtime	WaveFlow Execution Engine

Graphcore.ai

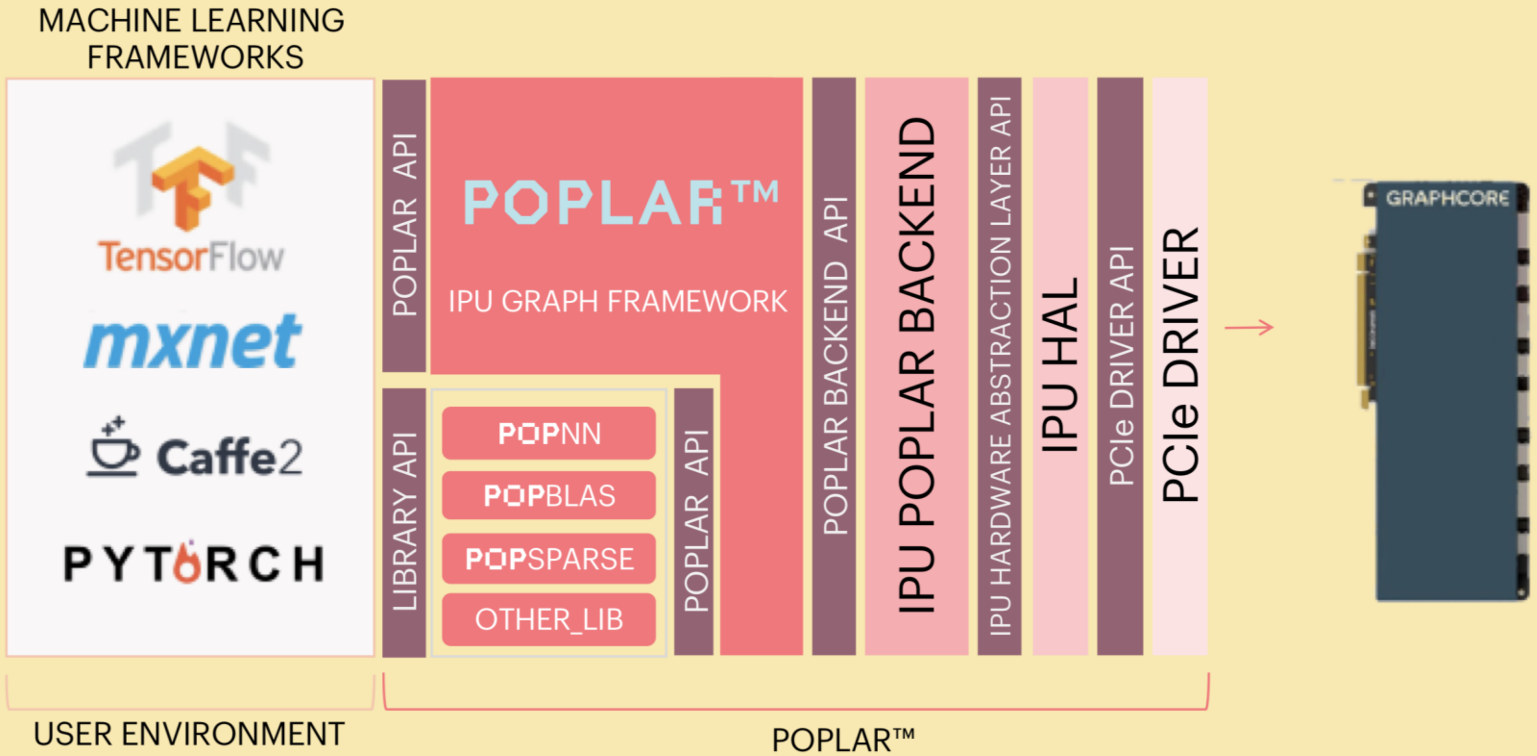
KNOWLEDGE MODELS
ARE NATURALLY REPRESENTED
AS GRAPHS...

VERTICES ARE FEATURES

EDGES ARE CORRELATIONS OR
CAUSATIONS



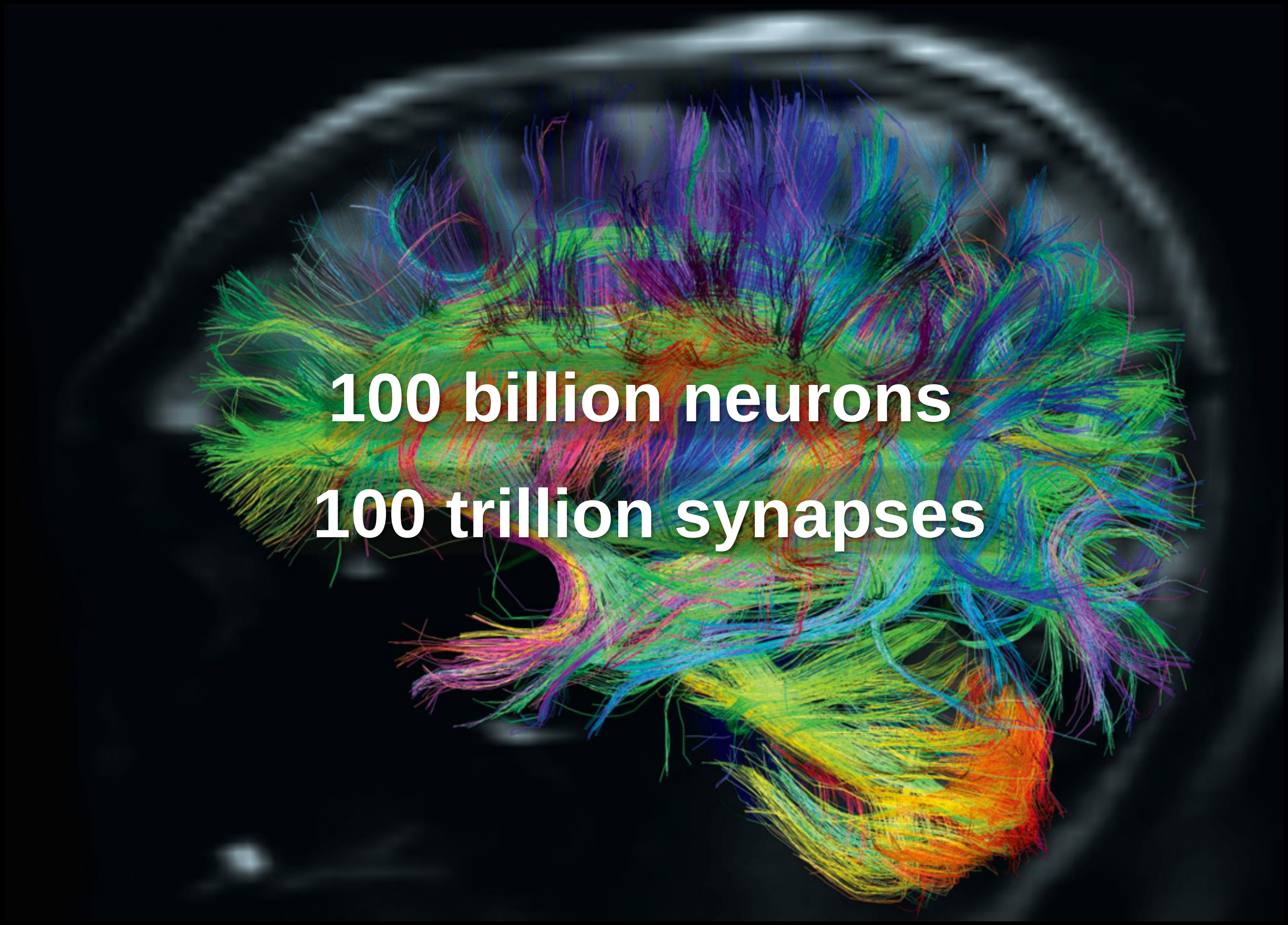
POPLAR™ SEAMLESS DEVELOPMENT



Neuromorphic Computing

Computing devices inspired by the computational model and physical construct of biological neurons.

Brain Inspired Computing


100 billion neurons
100 trillion synapses

5 lessons from your brain

(that could really help your computer)

Deep learning

People learn as they're exposed to new situations. In deep learning, a computer refines algorithms to improve its ability to understand data.

Parallelism

The brain breaks tasks into many little ones that it computes simultaneously. We're getting better at writing software to do this, too.

Intuition

A person can draw fairly accurate conclusions from incomplete data. Neuromorphic logic allows computers to calculate based on approximate information.

Low power

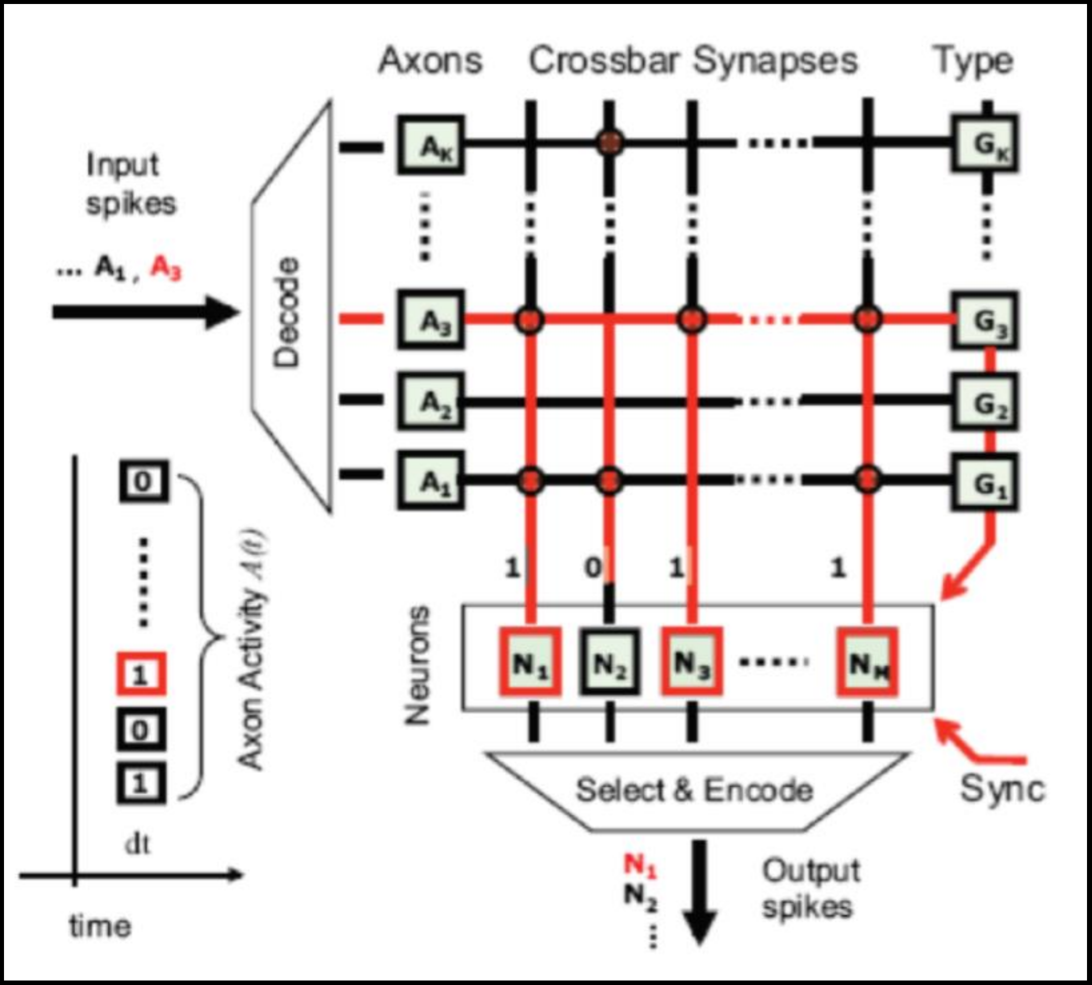
The brain uses about as much electrical current as a 20-watt light bulb. Memristors, which retain information when powered off, could eventually replace today's power-hungry computer memory and storage.

Locality

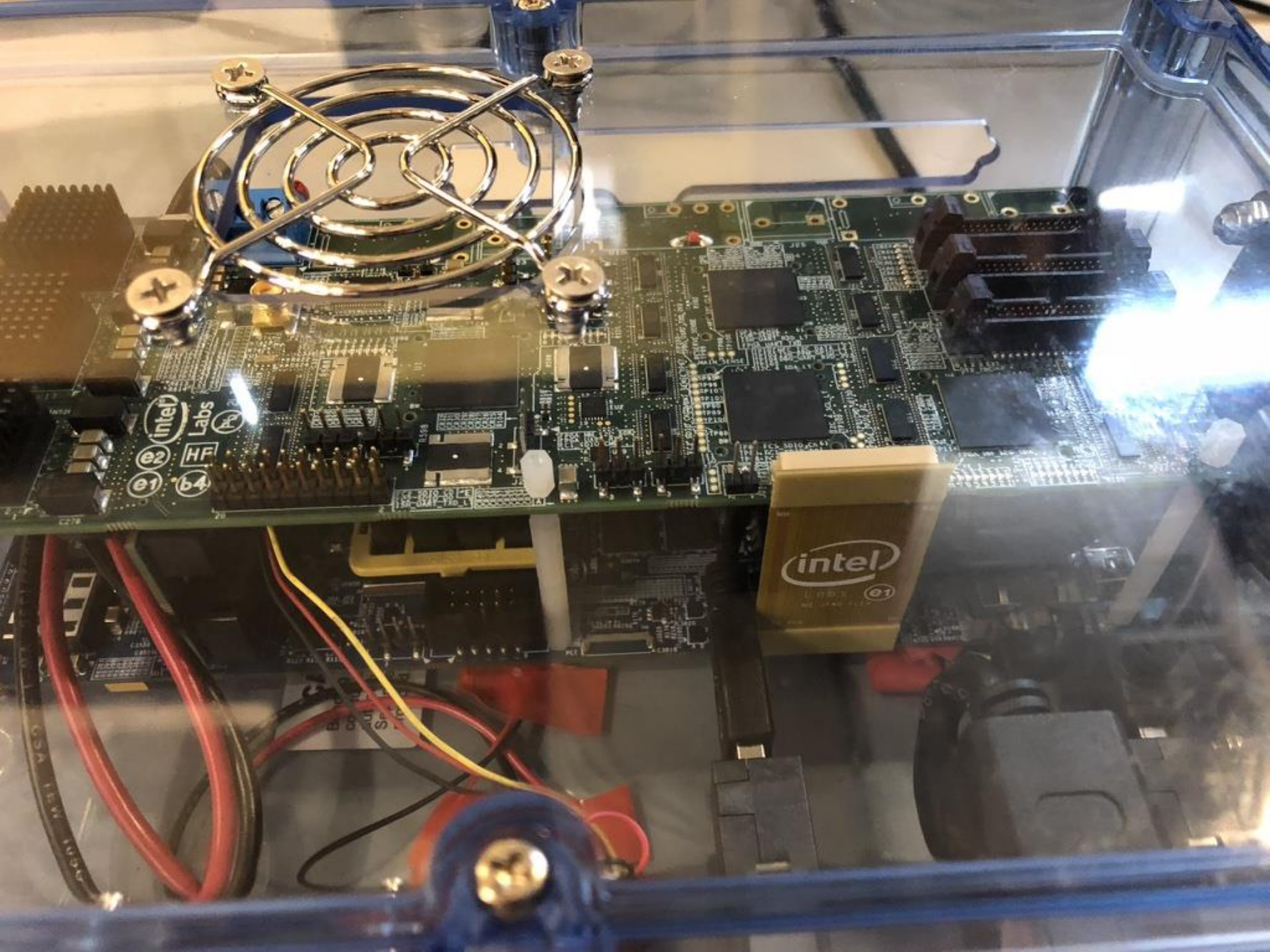
In the brain, the same cells remember and calculate. Neuromorphic computers put those functions as close together as possible.

Synapses Dominate Area

each neuron is connected to 256 to 10,000 others




LOIHI
LEARNING WITH LESS DATA


intel
Labs
intel
Labs

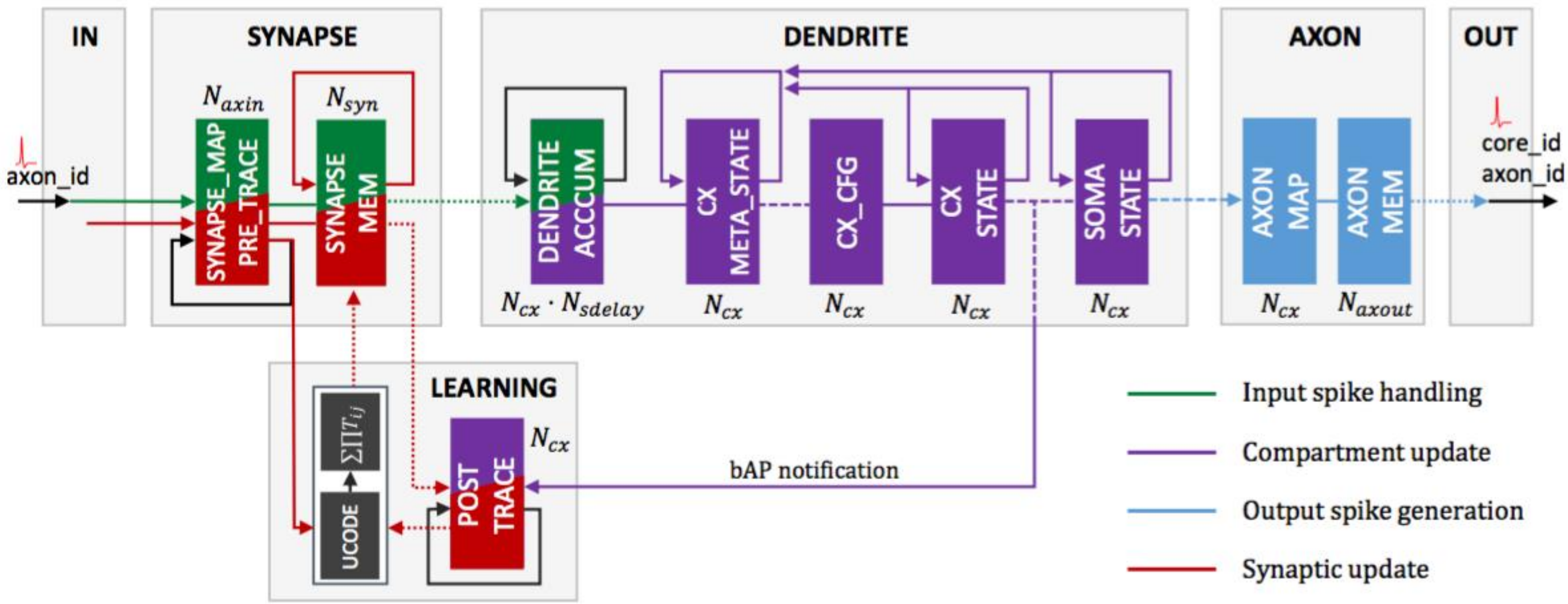


Fig. 4: Core Top-Level Microarchitecture. The SYNAPSE unit processes all incoming spikes and reads out the associated synaptic weights from the memory. The DENDRITE unit updates the state variables u and v of all neurons in the core. The AXON unit generates spike messages for all fanout cores of each firing neuron. The LEARNING unit updates synaptic weights using the programmed learning rules at epoch boundaries.

Summary

- Deep Learning is Accelerating
- Broadening of DL Applications
- New DL Architectures Emerging (10x-100x)
- Brain Inspired Computing

- Many NIH Computing Challenges could be addressed with DL Approaches
- A Grand Synthesis might be possible

Deep Dream