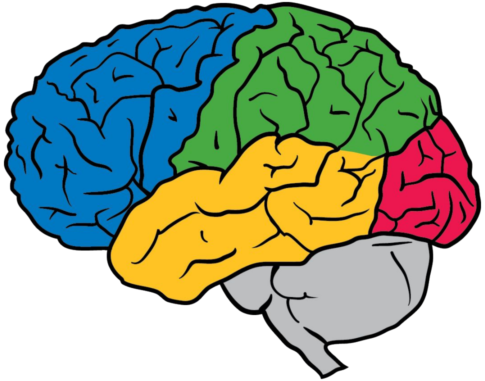


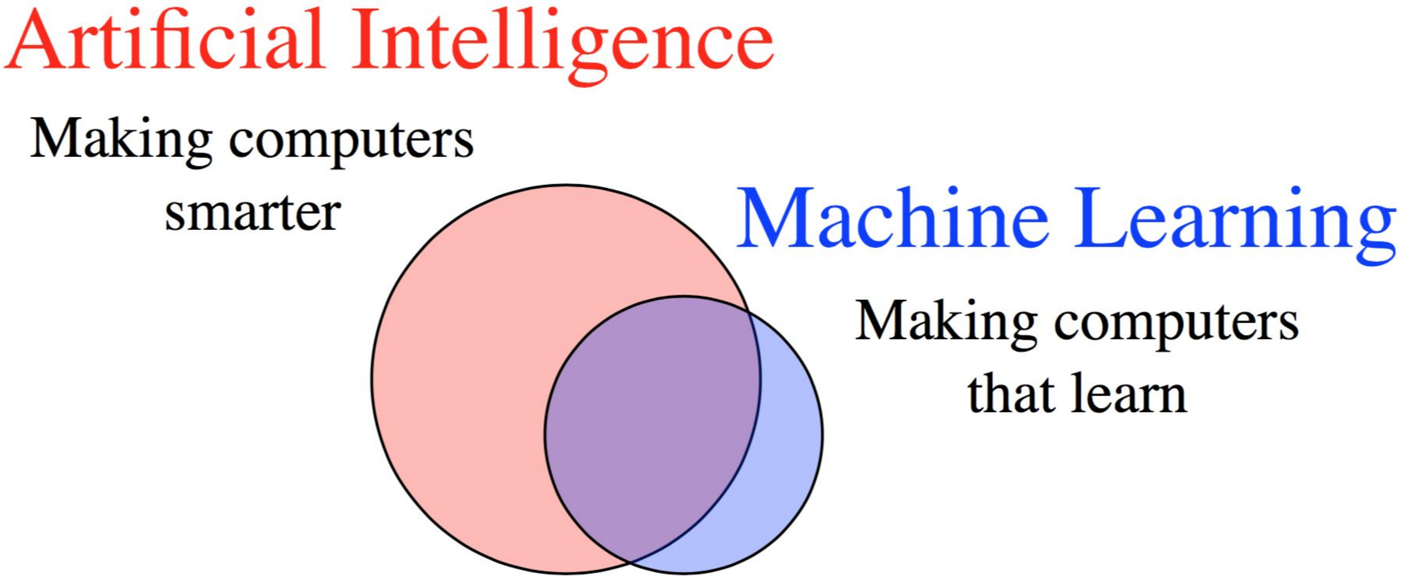
Deep Learning and Applications in Medicine and Science

Jeff Dean
Google Brain team
g.co/brain

Presenting the work of **many** people at Google

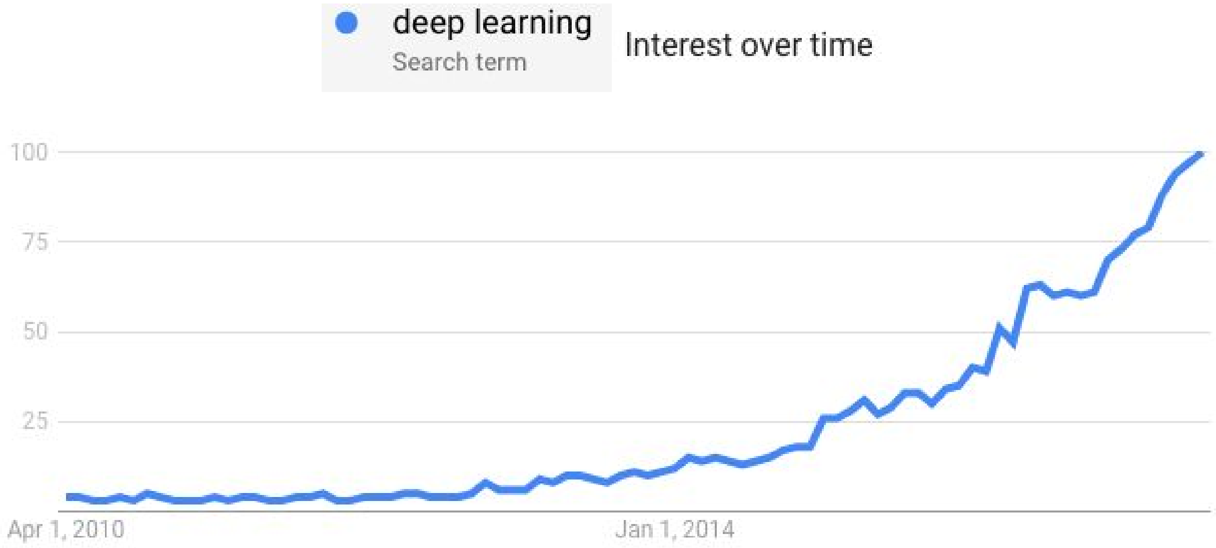
Mostly the work of Google Brain and Google Applied Sciences teams

Intro to machine learning and deep learning



Observation: programming a computer to *be* clever is harder than programming a computer *to learn to be* clever.

Deep learning is causing a machine learning revolution



Deep Learning

Modern Reincarnation of Artificial Neural Networks

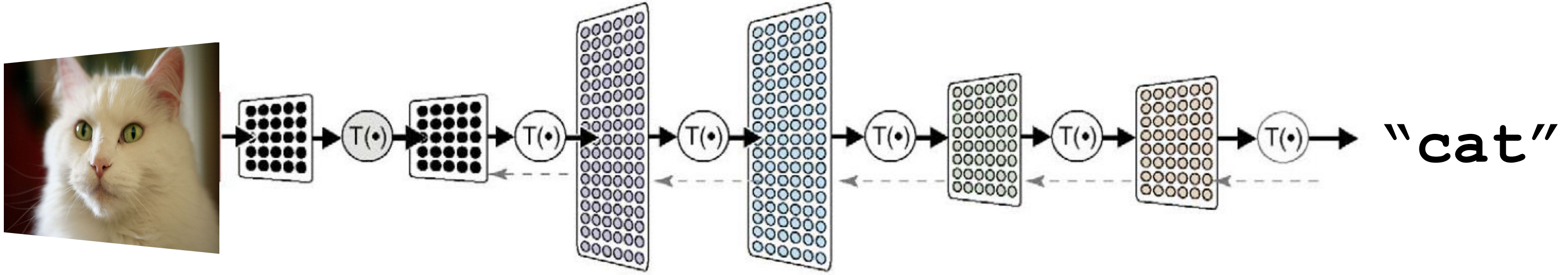
Collection of simple trainable mathematical units, organized in layers, that work together to solve complicated tasks

What's New

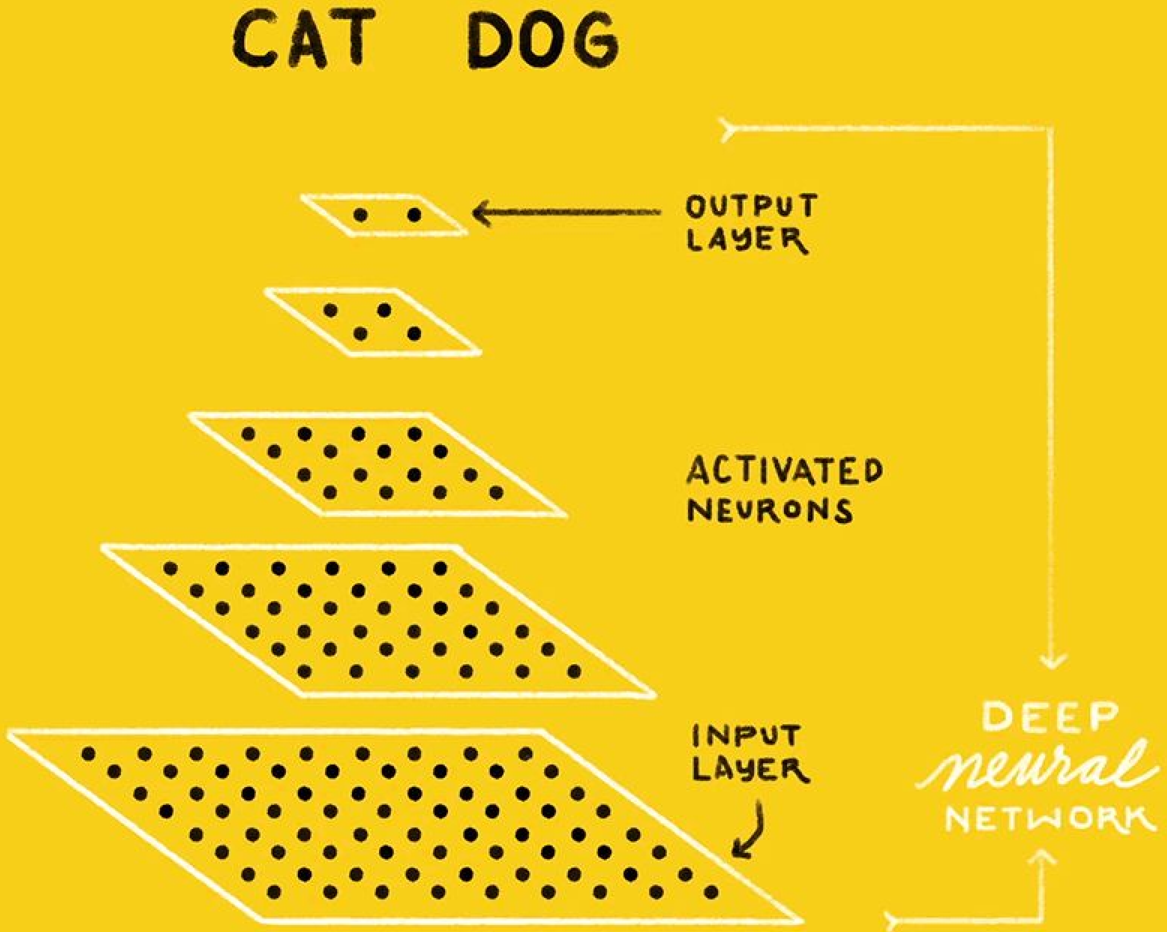
layered network architecture, new training math, ***scale***

Key Benefit

Learns features from raw, heterogeneous data
No explicit feature engineering required



IS THIS A CAT or DOG?

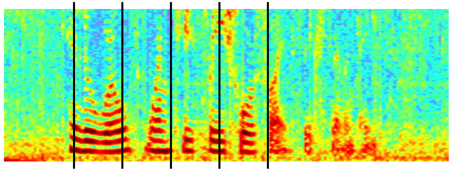


functions a deep neural network can learn

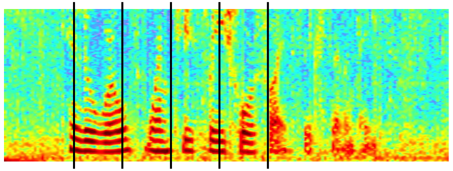
input | **output**

Pixels:  “lion”

functions a deep neural network can learn

	input	output
Pixels:		"lion"
Audio:		"How cold is it outside?"

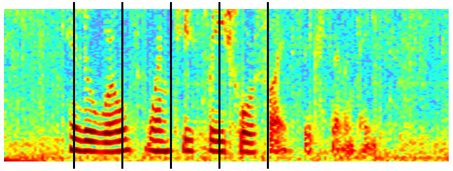
functions a deep neural network can learn

input		output
Pixels:		"lion"
Audio:		"How cold is it outside?"
"Hello, how are you?"		"Bonjour, comment allez-vous?"

functions a deep neural network can learn

input | **output**

Pixels:  "lion"

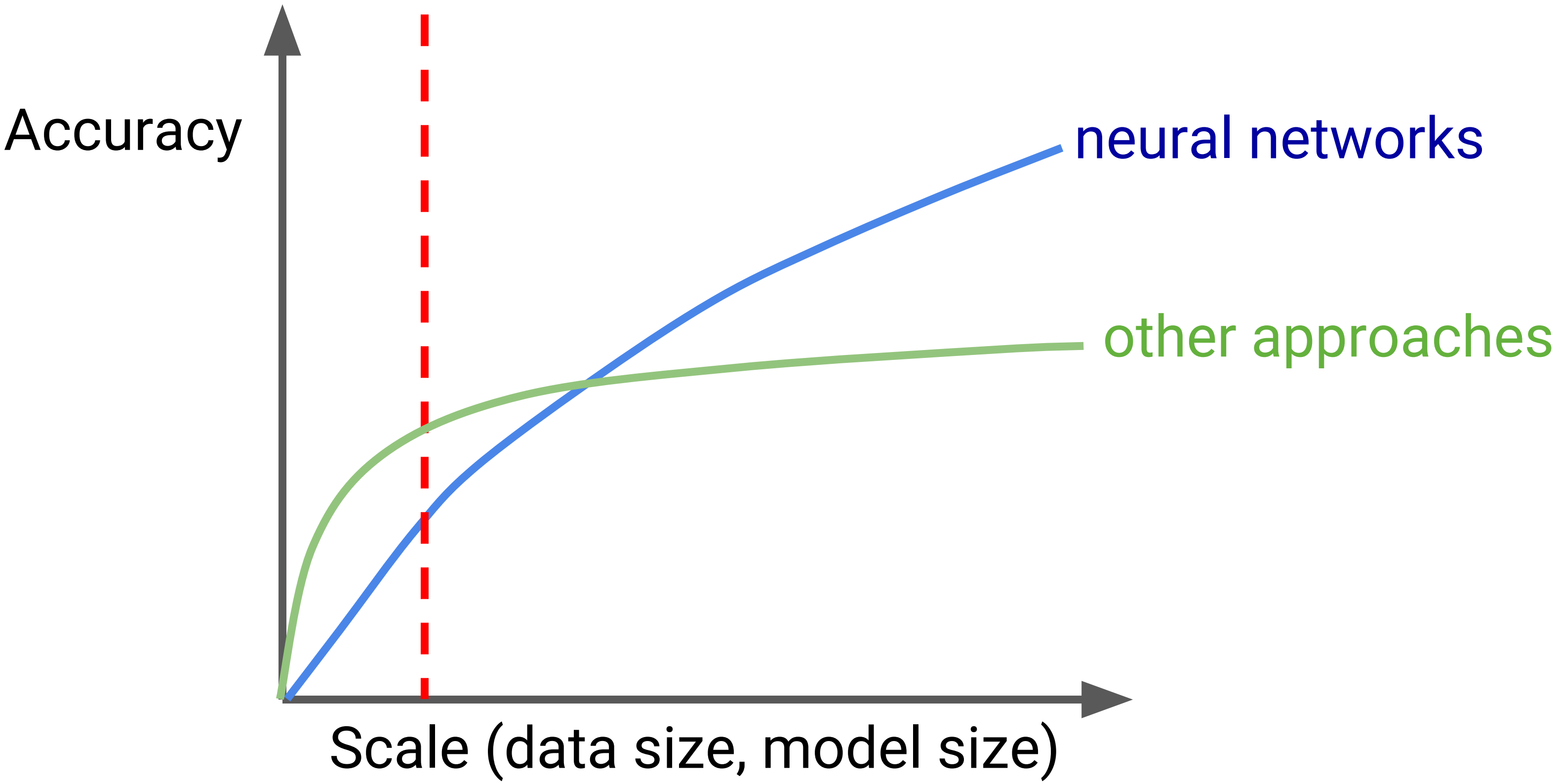
Audio:  "How cold is it outside?"

"Hello, how are you?" "Bonjour, comment allez-vous?"

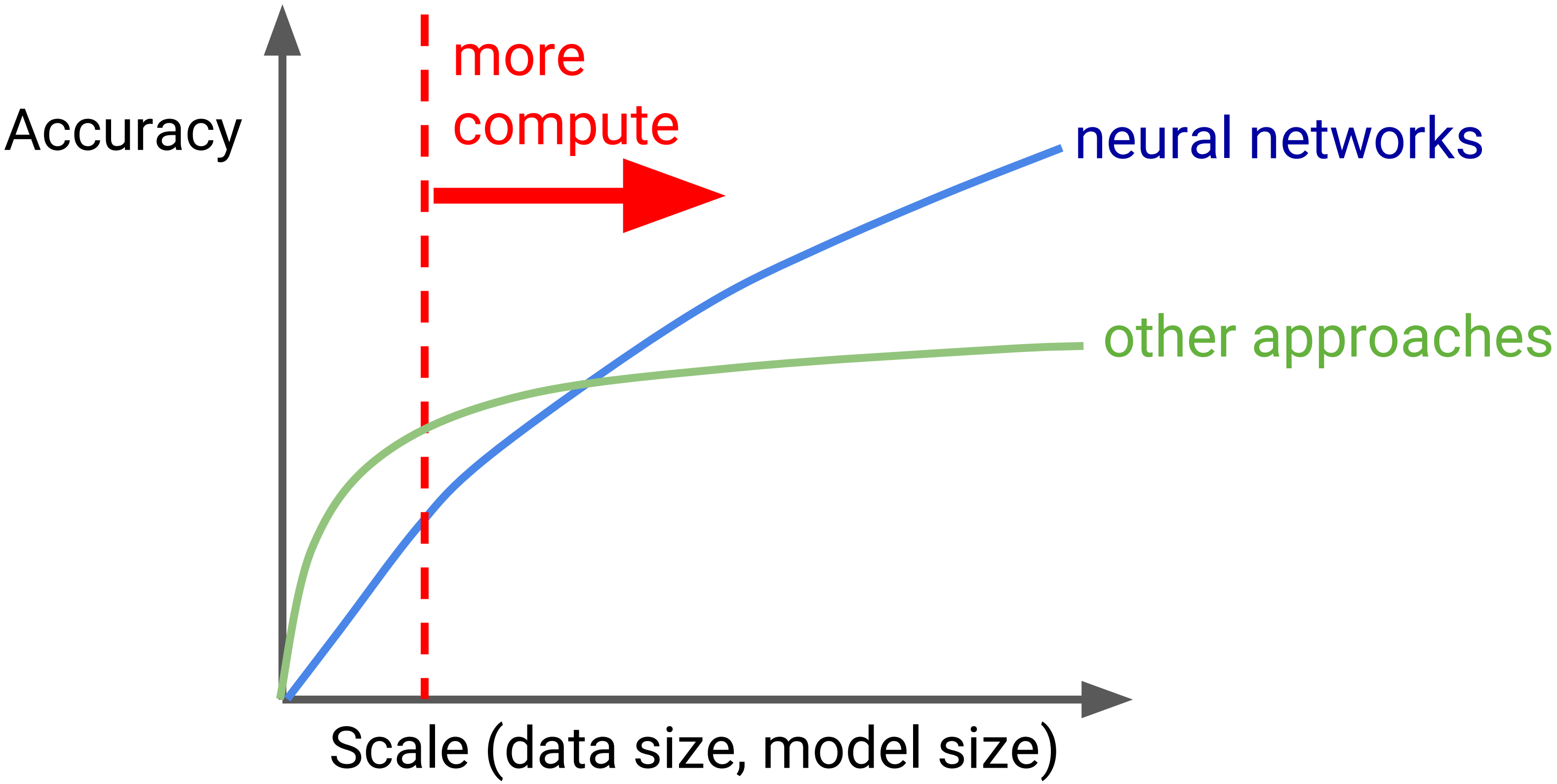
Pixels:  "A blue and yellow train travelling down the tracks"

But why now?

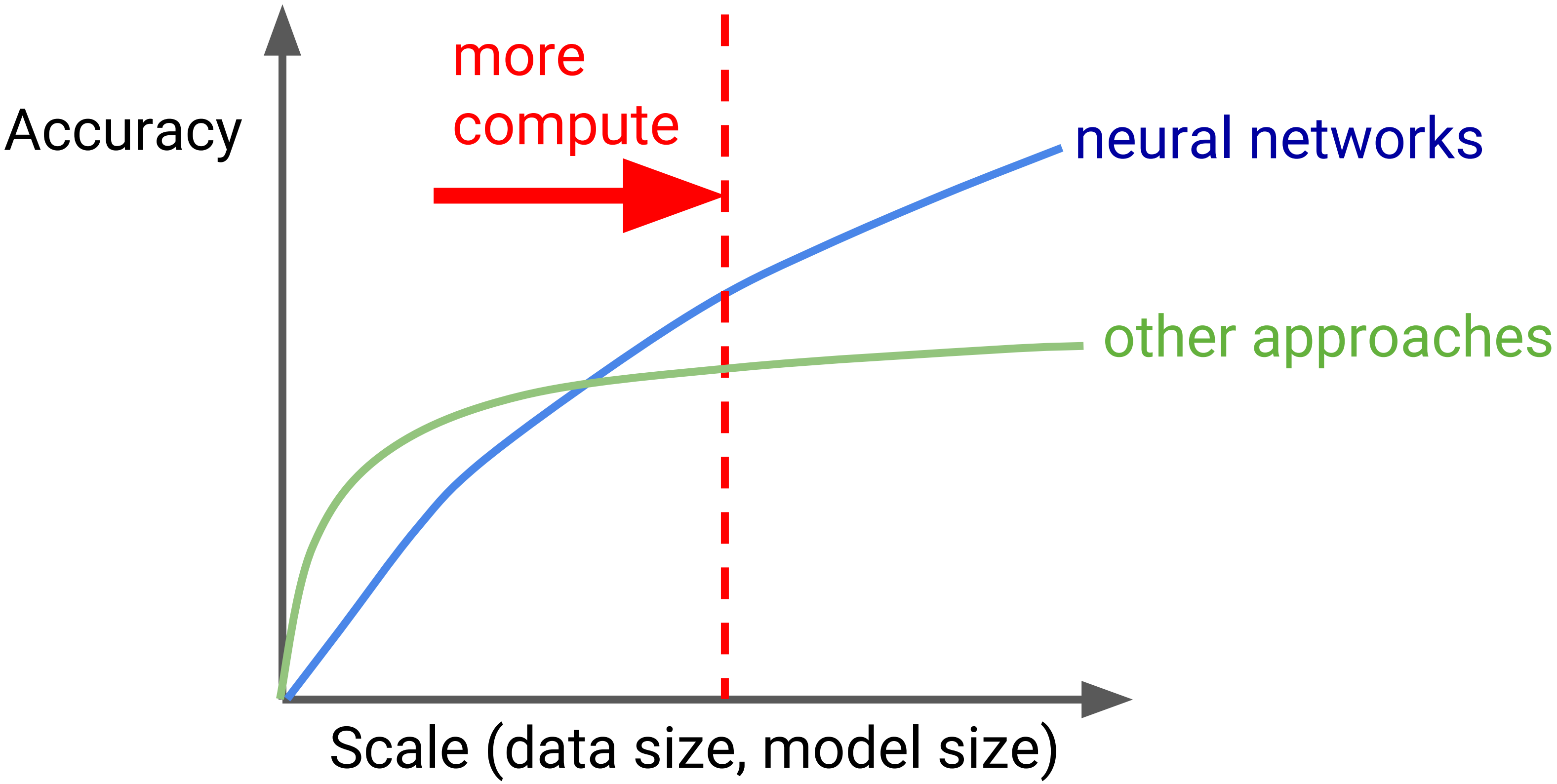
1980s and 1990s



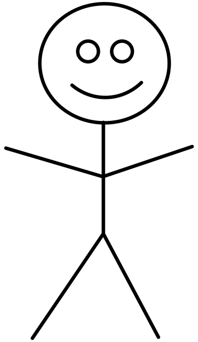
1980s and 1990s



Now

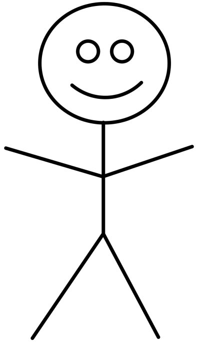

more
compute
Accuracy
neural networks
other approaches
Scale (data size, model size)

humans



5% errors

humans



5% errors

Google Brain Team Mission:
Make Machines Intelligent.
Improve People’s Lives.

How do we do this?

- Conduct **long-term research** (>200 papers, see g.co/brain & g.co/brain/papers)
    - Unsupervised learning of cats, Inception, word2vec, seq2seq, DeepDream, image captioning, neural translation, Magenta, ML for robotics control, ...
- Build and open-source systems like **TensorFlow** (see tensorflow.org and https://github.com/tensorflow/tensorflow)
- Collaborate with others at Google and Alphabet to **get our work into the hands of billions of people** (e.g., Google Translate, RankBrain for Google Search, GMail Smart Reply, Google Photos, Google speech recognition, Waymo, …)
- **Train new researchers**: internships, Google Brain Residency program
- Conduct **applied ML research** in emerging areas where new ML methods will make a big difference like **healthcare** and **robotics**

Main Research Areas

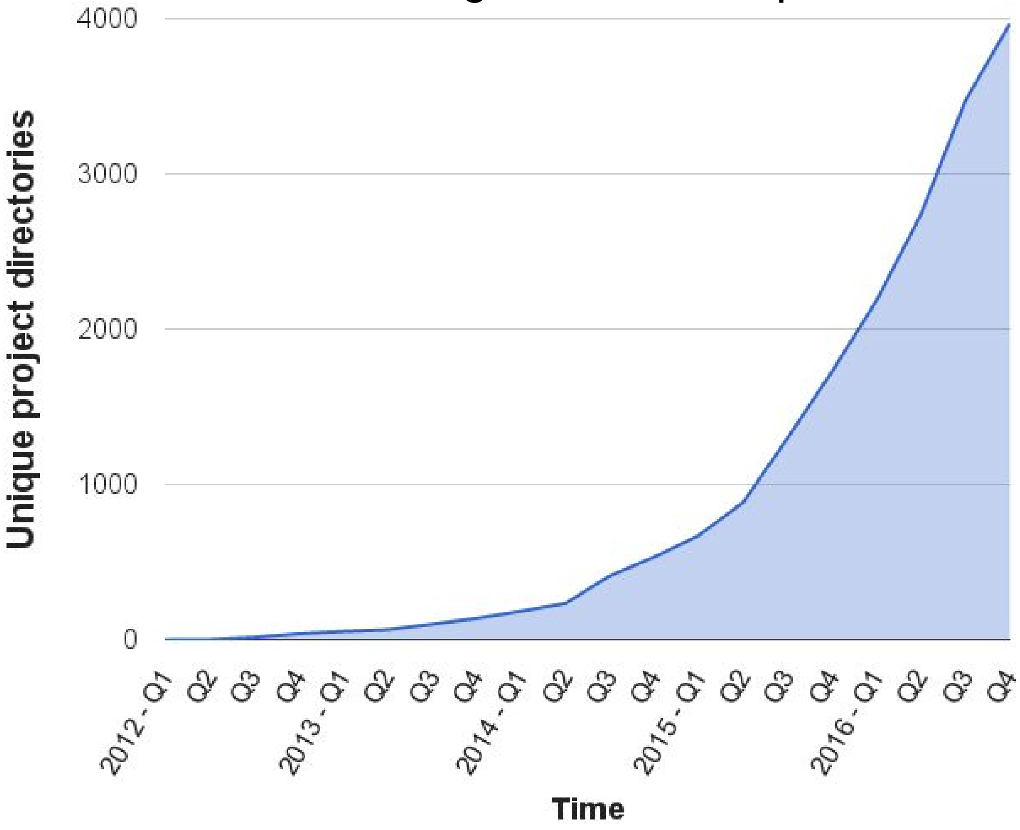
- General Machine Learning Algorithms and Techniques
- Computer Systems for Machine Learning
- Natural Language Understanding
- Perception
- Healthcare
- Robotics
- Music and Art Generation

Main Research Areas

- General Machine Learning Algorithms and Techniques
- Computer Systems for Machine Learning
- Natural Language Understanding
- Perception
- Healthcare
- Robotics
- Music and Art Generation

Growing Use of Deep Learning at Google

of directories containing model description files



Across many products/areas:

Android
Apps
drug discovery
Gmail
Image understanding
Maps
Natural language understanding
Photos
Robotics research
Speech
Translation
YouTube
… many others ...

Experiment Turnaround Time and Research Productivity

- **Minutes, Hours**:
  - **Interactive research! Instant gratification!**
- **1-4 days**
  - Tolerable
  - Interactivity replaced by running many experiments in parallel
- **1-4 weeks**
  - High value experiments only
  - Progress stalls
- **>1 month**
  - Don’t even try

Build the right tools

TensorFlow

http://tensorflow.org/

and

https://github.com/tensorflow/tensorflow

Open, standard software for general machine learning

Great for Deep Learning in particular

First released Nov 2015

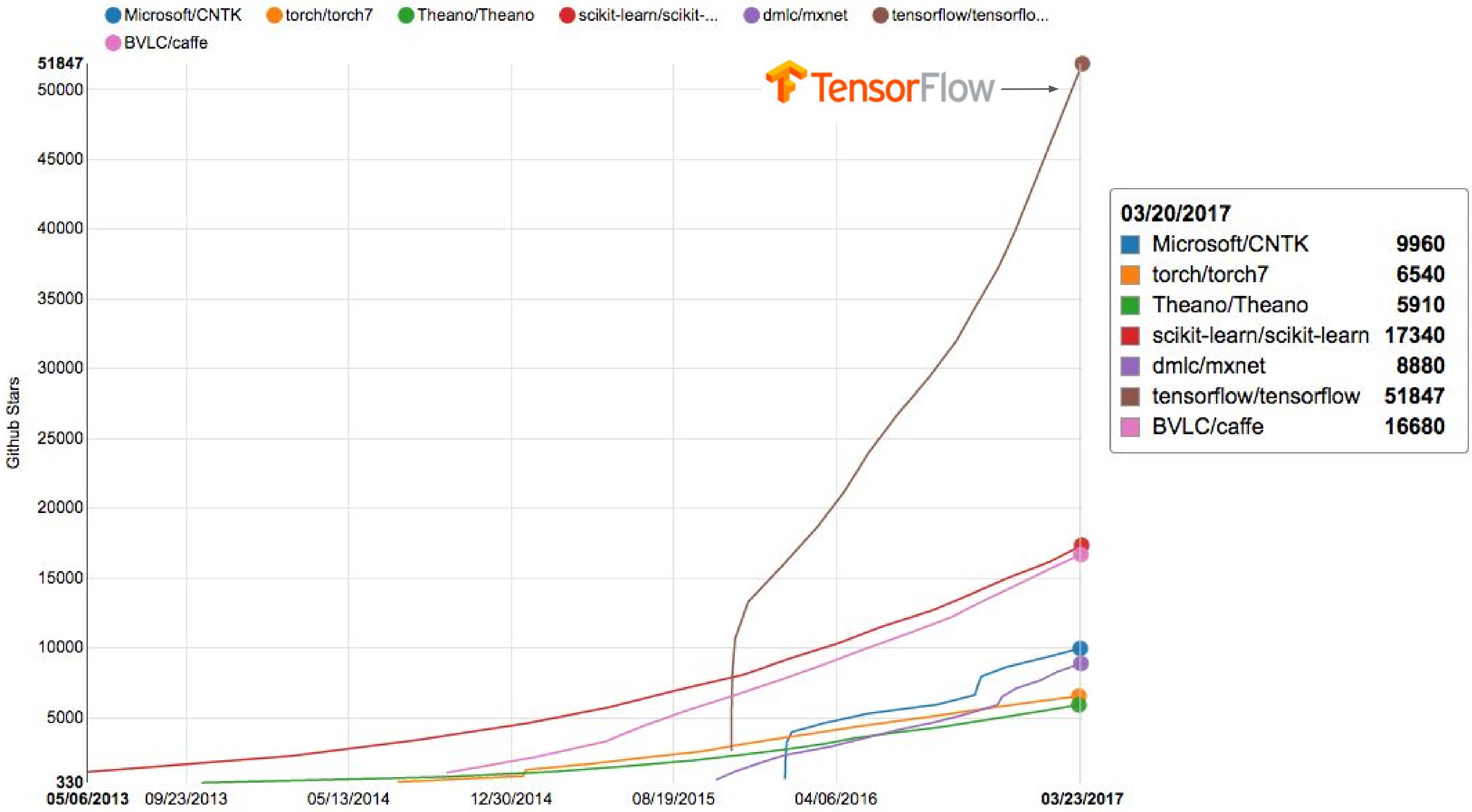
Apache 2.0 license

TensorFlow Goals

Establish **common platform** for expressing machine learning ideas and systems

Make this platform the **best in the world** for both research and production use

Open source it so that it becomes a **platform for everyone**, not just Google



Microsoft/CNTK
torch/torch7
Theano/Theano
scikit-learn/scikit-...
dmlc/mxnet
tensorflow/tensorflo...
BVLC/caffe
TensorFlow
Github Stars
51847
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
330
05/06/2013
09/23/2013
05/13/2014
12/30/2014
08/19/2015
04/06/2016
03/23/2017
03/20/2017
Microsoft/CNTK 9960
torch/torch7 6540
Theano/Theano 5910
scikit-learn/scikit-learn 17340
dmlc/mxnet 8880
tensorflow/tensorflow 51847
BVLC/caffe 16680

TensorFlow: A Vibrant Open-Source Community

- **Rapid development, many outside contributors**
  - 475+ non-Google contributors to TensorFlow 1.0
  - 15,000+ commits in 15 months
  - Many community created tutorials, models, translations, and projects
    - ~7,000 GitHub repositories with 'TensorFlow' in the title
- **Direct engagement between community and TensorFlow team**
  - 5000+ Stack Overflow questions answered
  - 80+ community-submitted GitHub issues responded to weekly
- **Growing use in ML classes: Toronto, Berkeley, Stanford, ...**

New-found computer vision prowess.

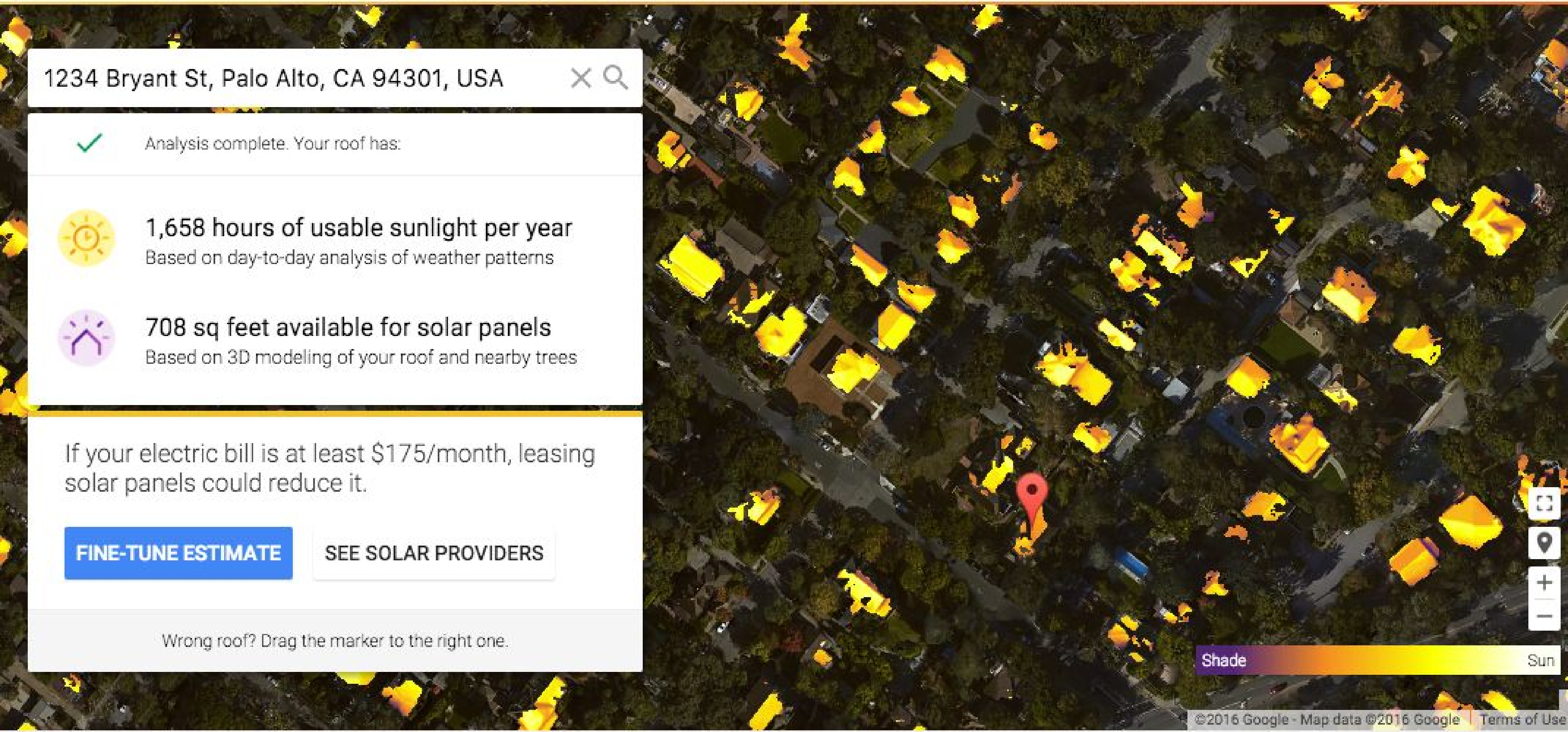
Often can develop a model to solve one problem, reuse it for other problems


ASIAWIDE TRAVEL
(02) 9745 3355 1st Floor, 240 BURWOOD RD
Maria's Bakery Inn 超羣餅屋
Maria's Bakery Inn 超羣餅屋


Corner
Cubbyhouse
CIANO MOTOR ENGINEERS
MECHANICAL REPAIRS TO ALL MAKES AND MODELS
Specialising In BMW, MINI & TOYOTA
8 REGATTA ROAD FIVE DOCK 9745 3173
88

Google Project Sunroof

www.google.com/sunroof



Medical Imaging

Diabetic Retinopathy as a Classification Problem

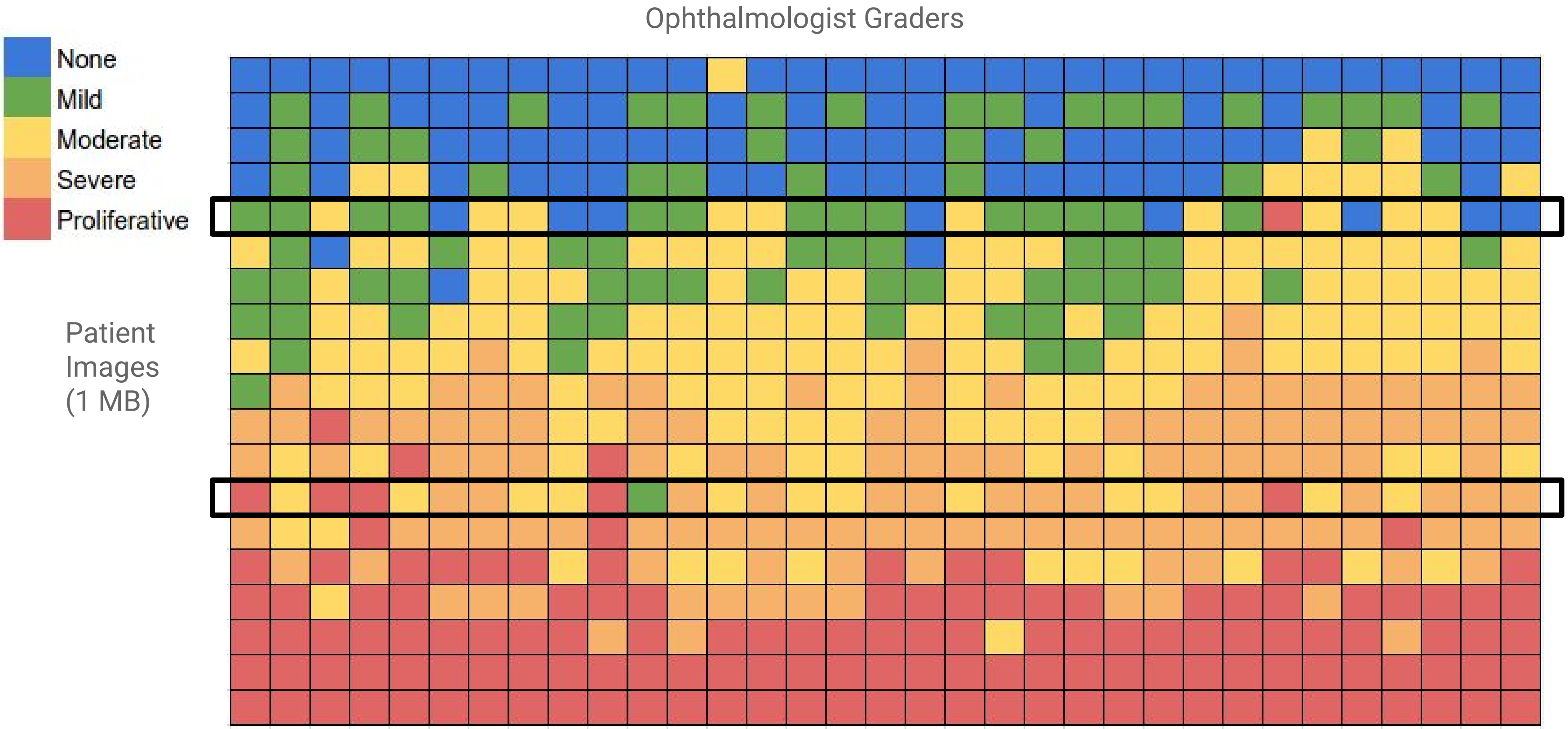
Moderate

Proliferative

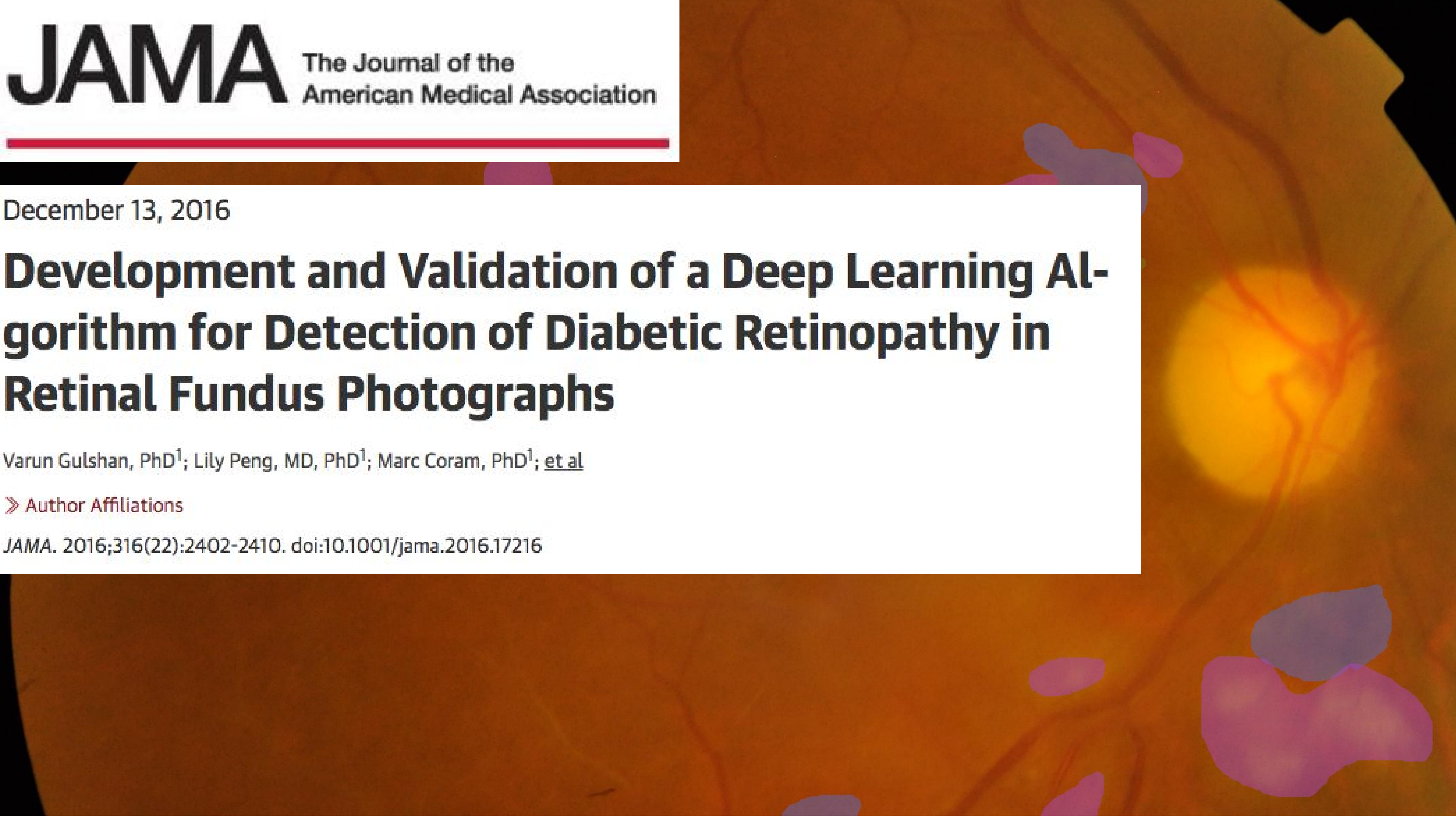
No DR	Mild DR	Moderate DR	Severe DR	Proliferative DR

Google

Humans are inconsistent



Consistency: intergrader ~60%, intragrader ~65%



JAMA The Journal of the American Medical Association

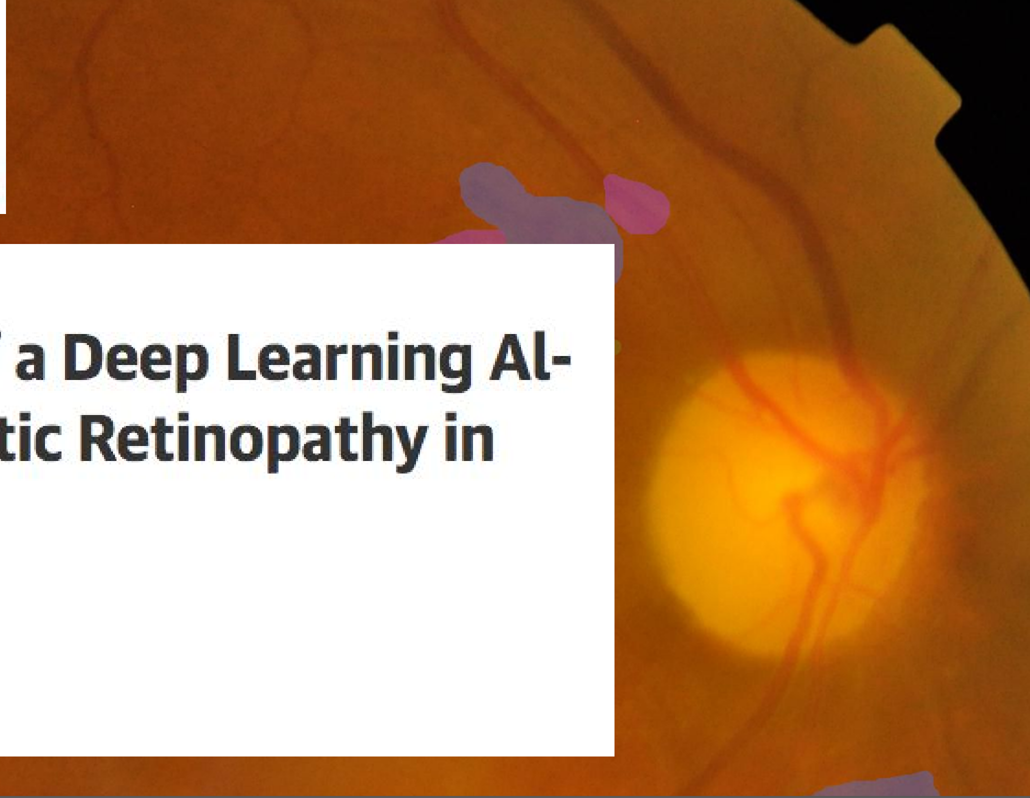
December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

» Author Affiliations

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216



JAMA The Journal of the American Medical Association

December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

» Author Affiliations

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

Performance **on par or slightly better** than the median of 8 U.S. board-certified ophthalmologists (F-score of 0.95 vs. 0.91).

http://research.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

Detecting Cancer Metastases on Gigapixel Pathology Images

Yun Liu[1*], Krishna Gadepalli[1], Mohammad Norouzi[1], George E. Dahl[1], Timo Kohlberger[1], Aleksey Boyko[1], Subhashini Venugopalan[2**], Aleksei Timofeev[2], Philip Q. Nelson[2], Greg S. Corrado[1], Jason D. Hipp[3], Lily Peng[1], and Martin C. Stumpe[1]

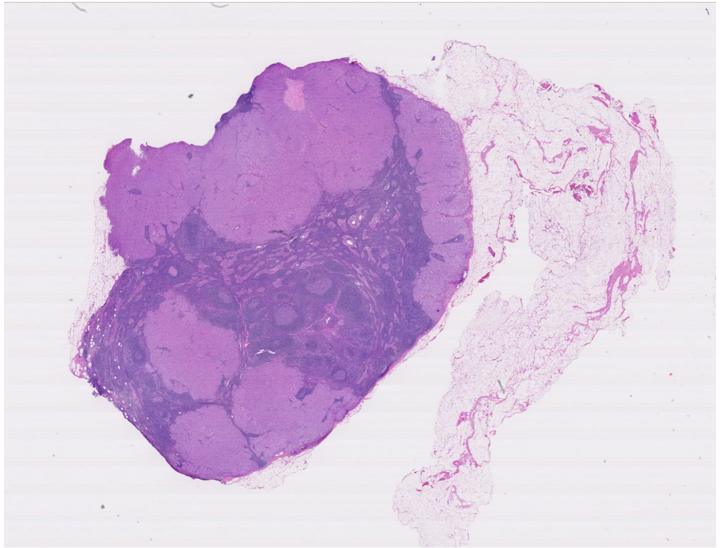
{liuyun,mnorouzi,gdahl,lhpeng,mstumpe}@google.com

[1]Google Brain, [2]Google Inc, [3]Verily Life Sciences, Mountain View, CA, USA

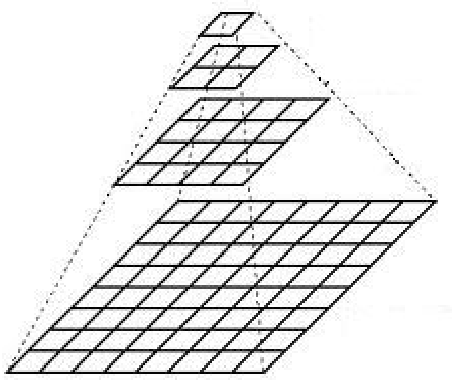
Blog: https://research.googleblog.com/2017/03/assisting-pathologists-in-detecting.html
Paper: https://arxiv.org/abs/1703.02442

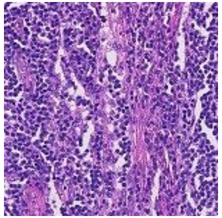
ML Challenges in Pathology

- ❏ Extremely large images (> 100k x 100k pixels)
- ❏ Multiscale problem - need detail as well as context

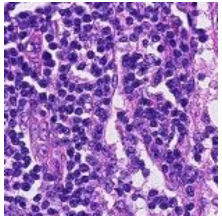


150k pixels (15 Gigapixel image)

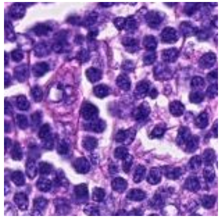




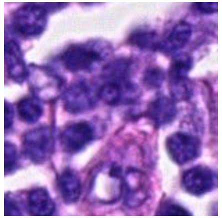
5x



10x



20x



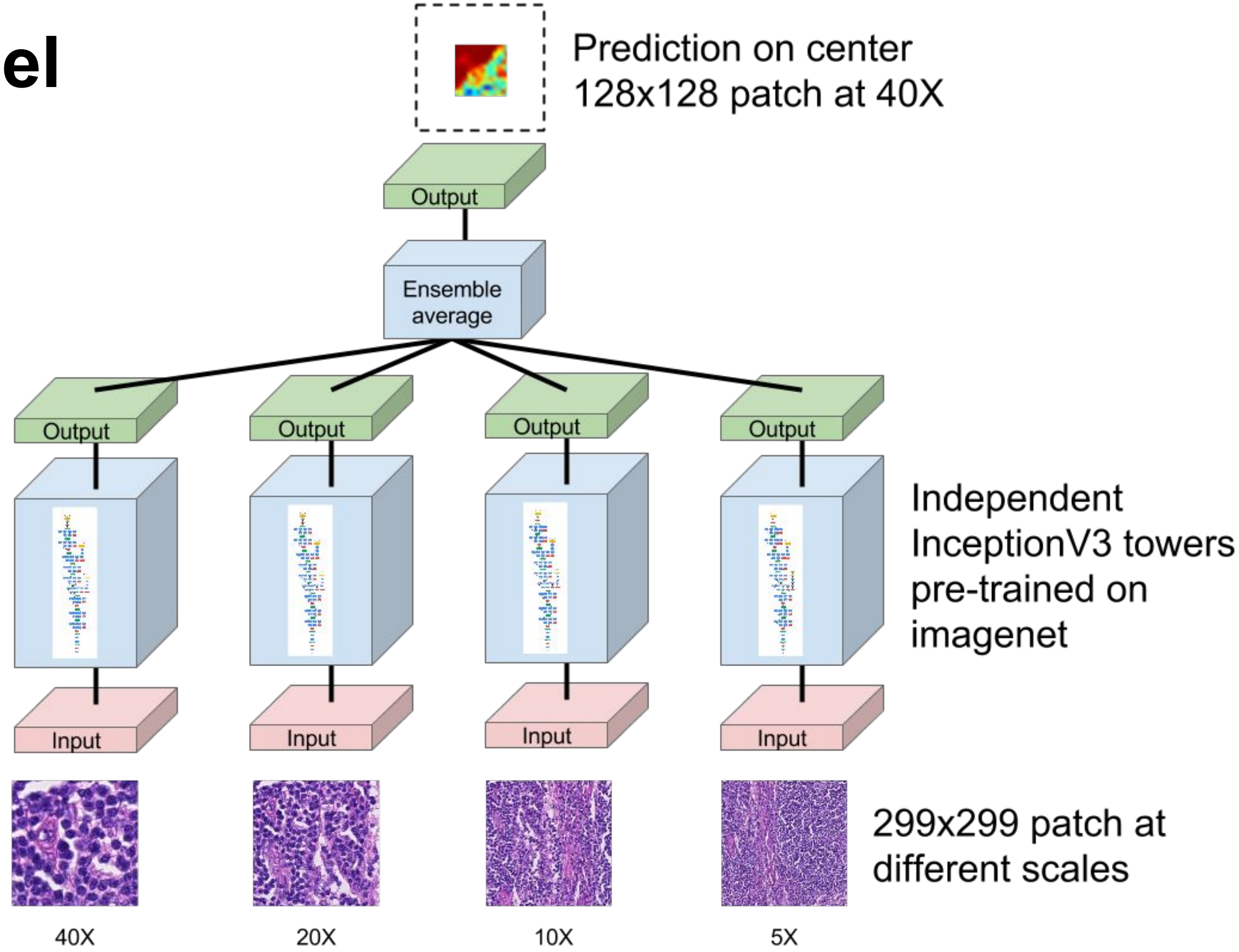
40x

Multiscale model

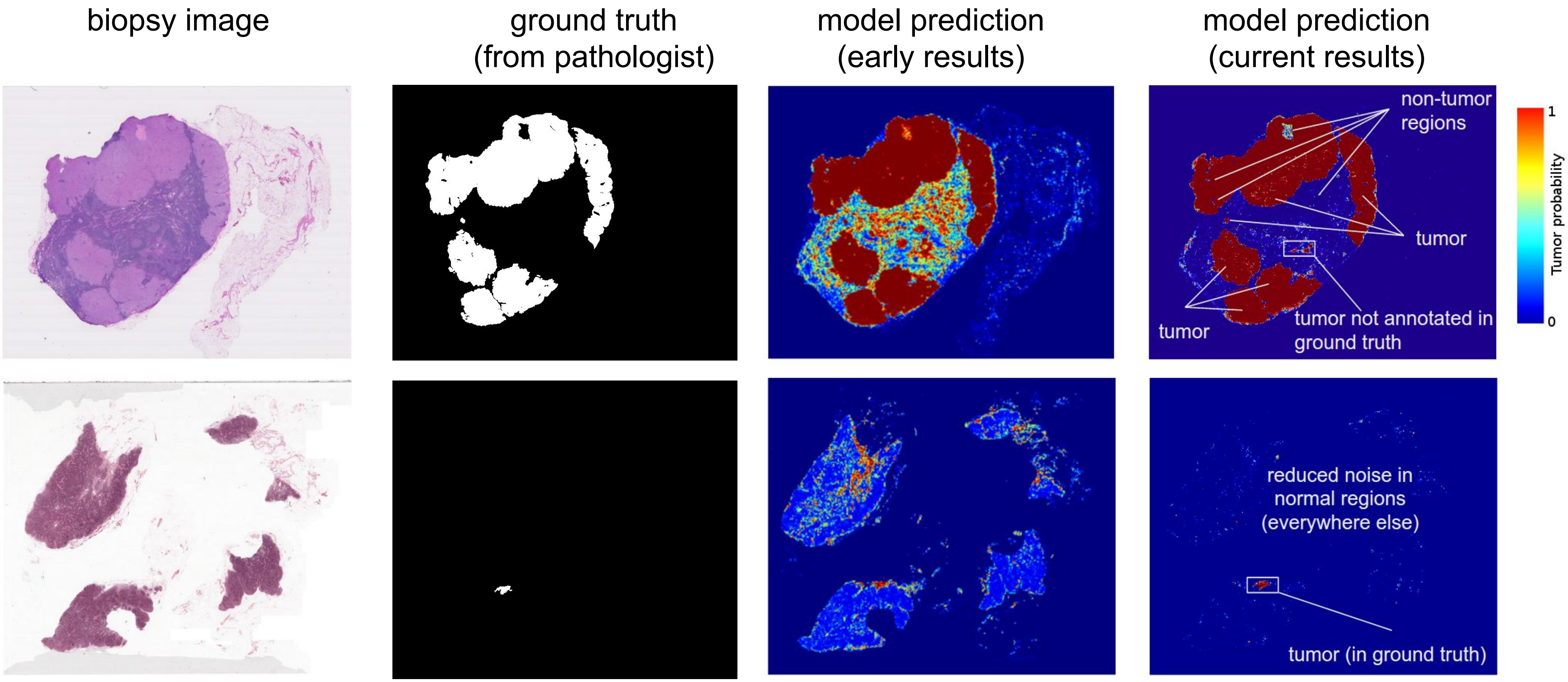
Multi scale model

detail ←→ context

resembles microscope magnifications



Detecting breast cancer metastases in lymph nodes



Model performance compared to pathologist

	our model	**pathologist***
Tumor localization score (FROC)	0.89	0.73
Sensitivity at 8 FP	0.92	0.73
Slide classification (AUC)	0.97	0.96

* pathologist given infinite time per image (reaching 0 FPs)

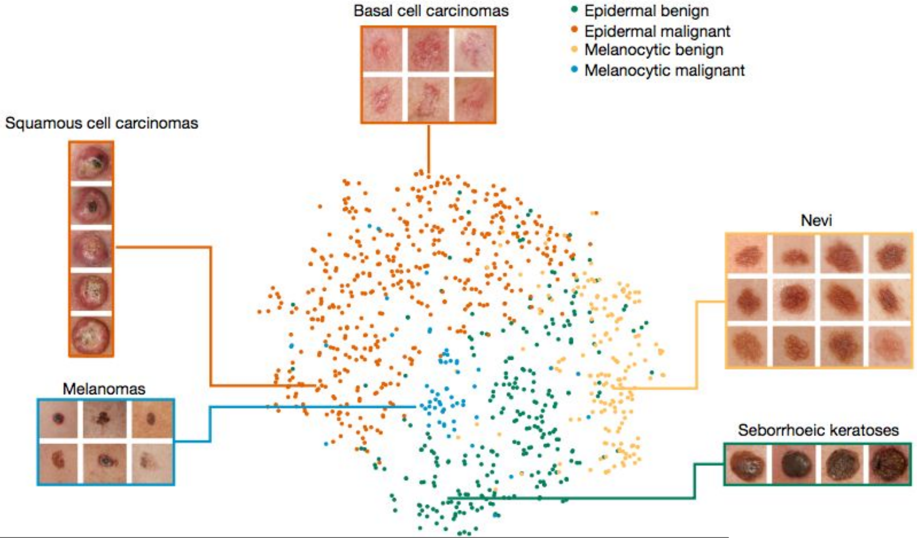
Evaluated using Camelyon16 dataset (**just 270 training examples!**)

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]



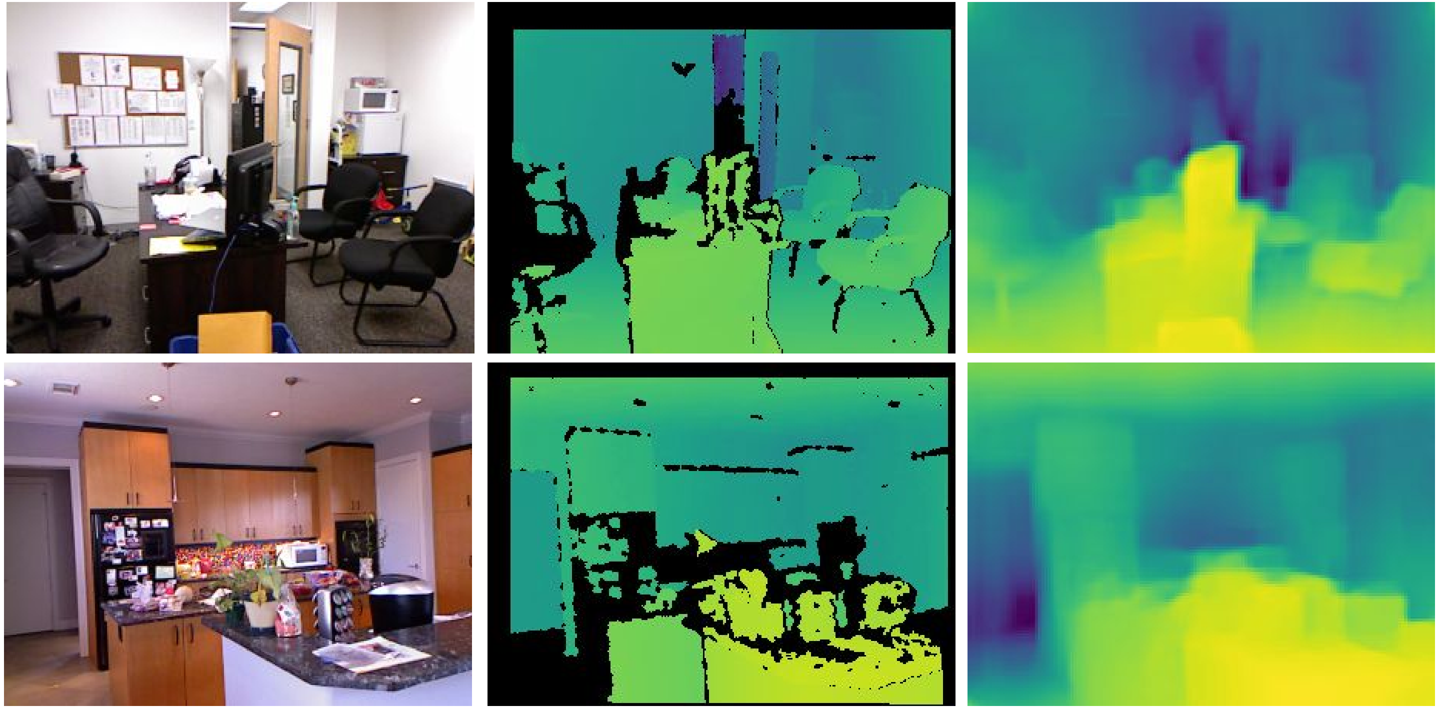
Work on dermatology from Stanford University

Measuring live cells with image to image regression

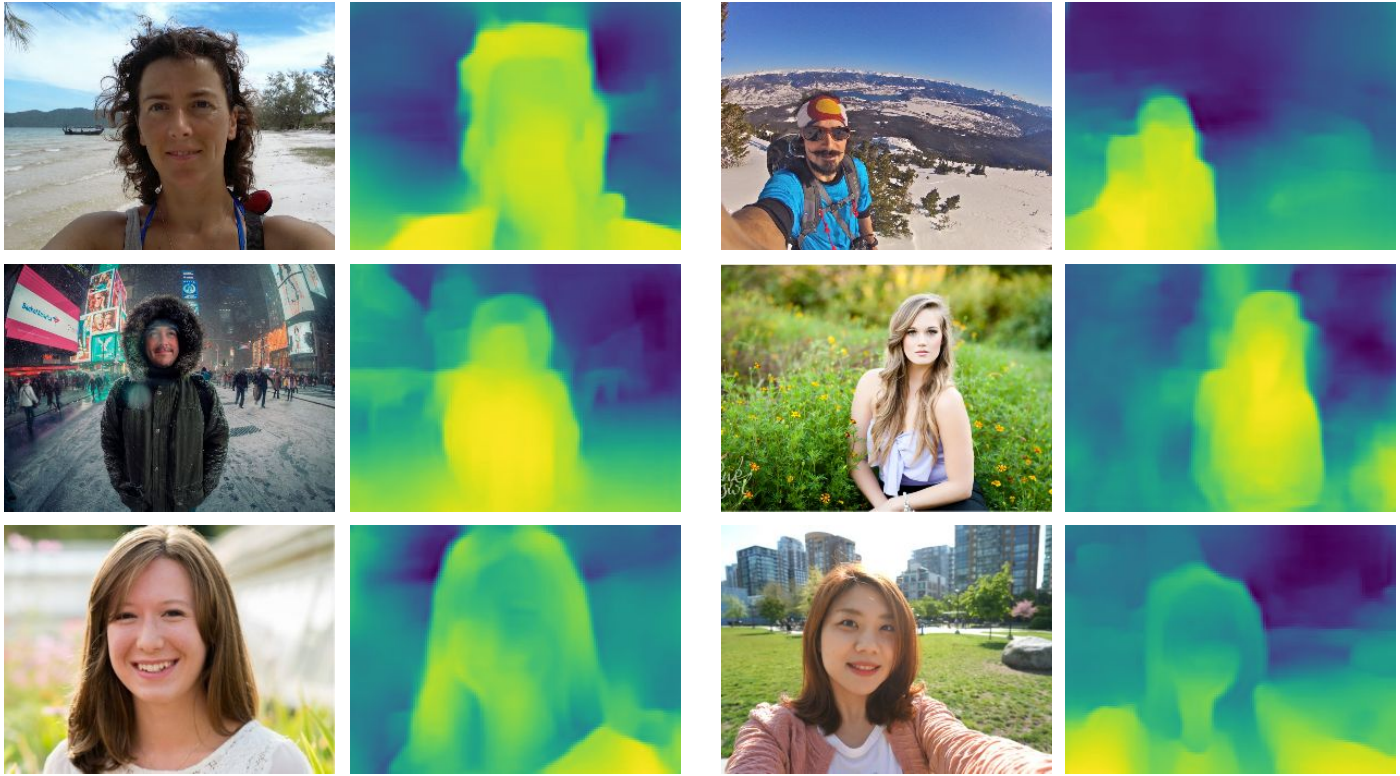
“Seeing More”

Enabling technology: Image to image regression

Input | True Depth | Predicted Depth



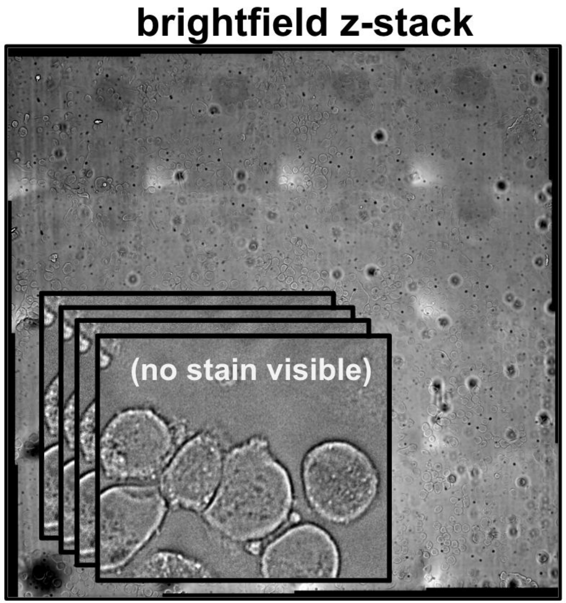
Depth prediction on portrait data

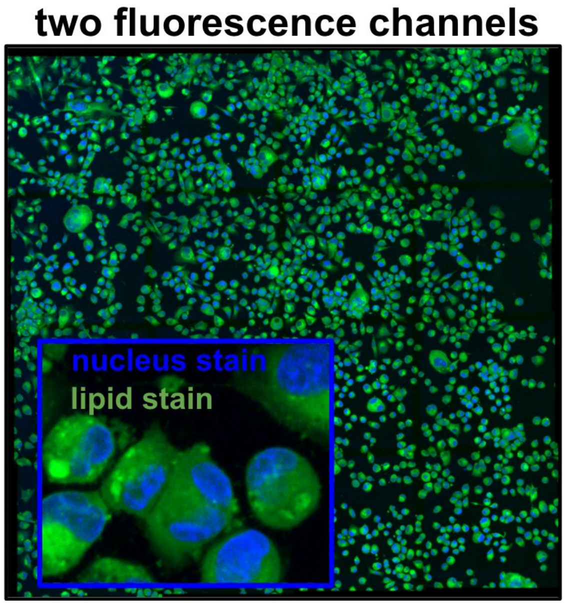


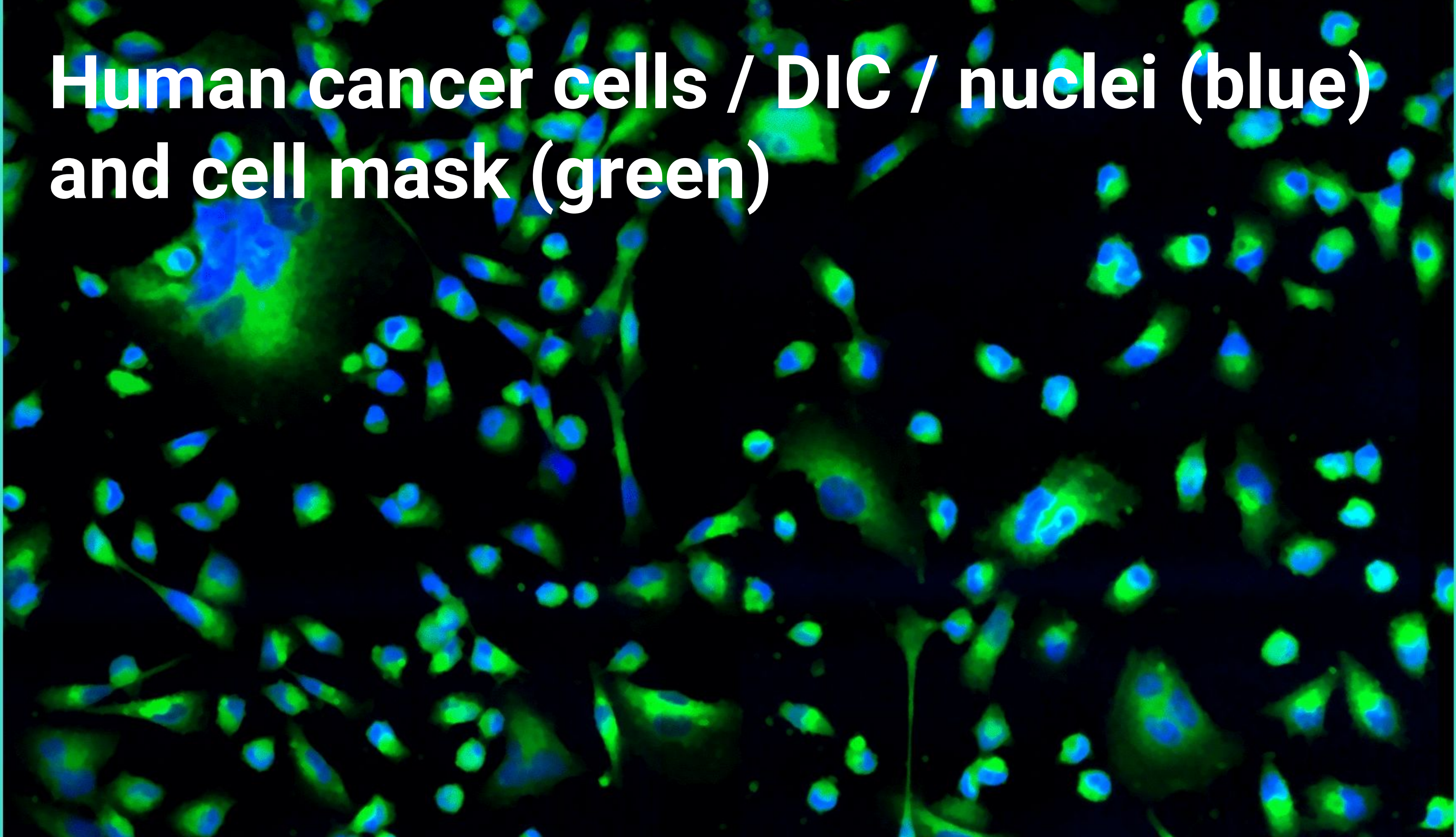
Applications for camera effects

Input Saturation Defocus

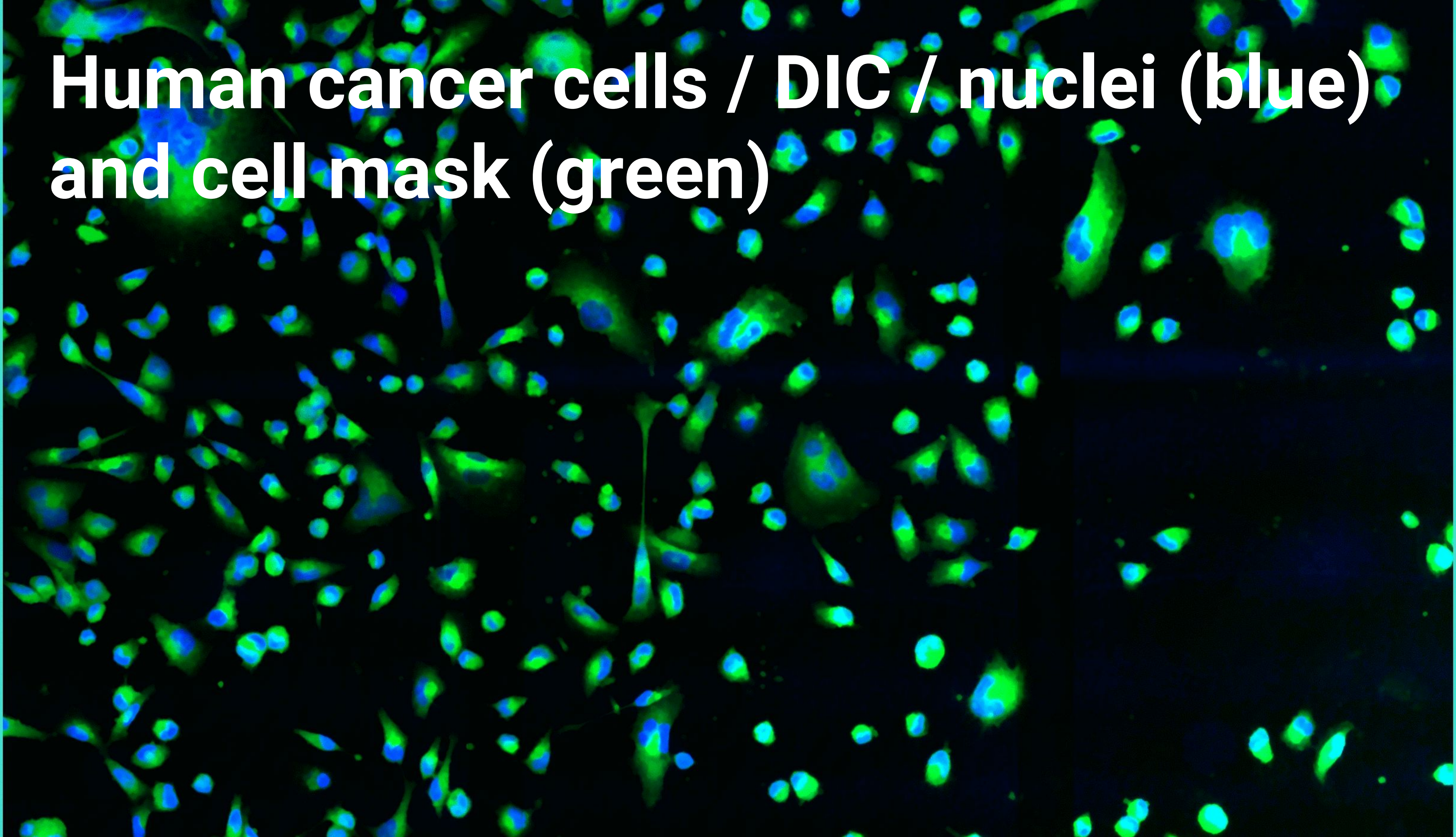
Predict cellular markers from transmission microscopy?



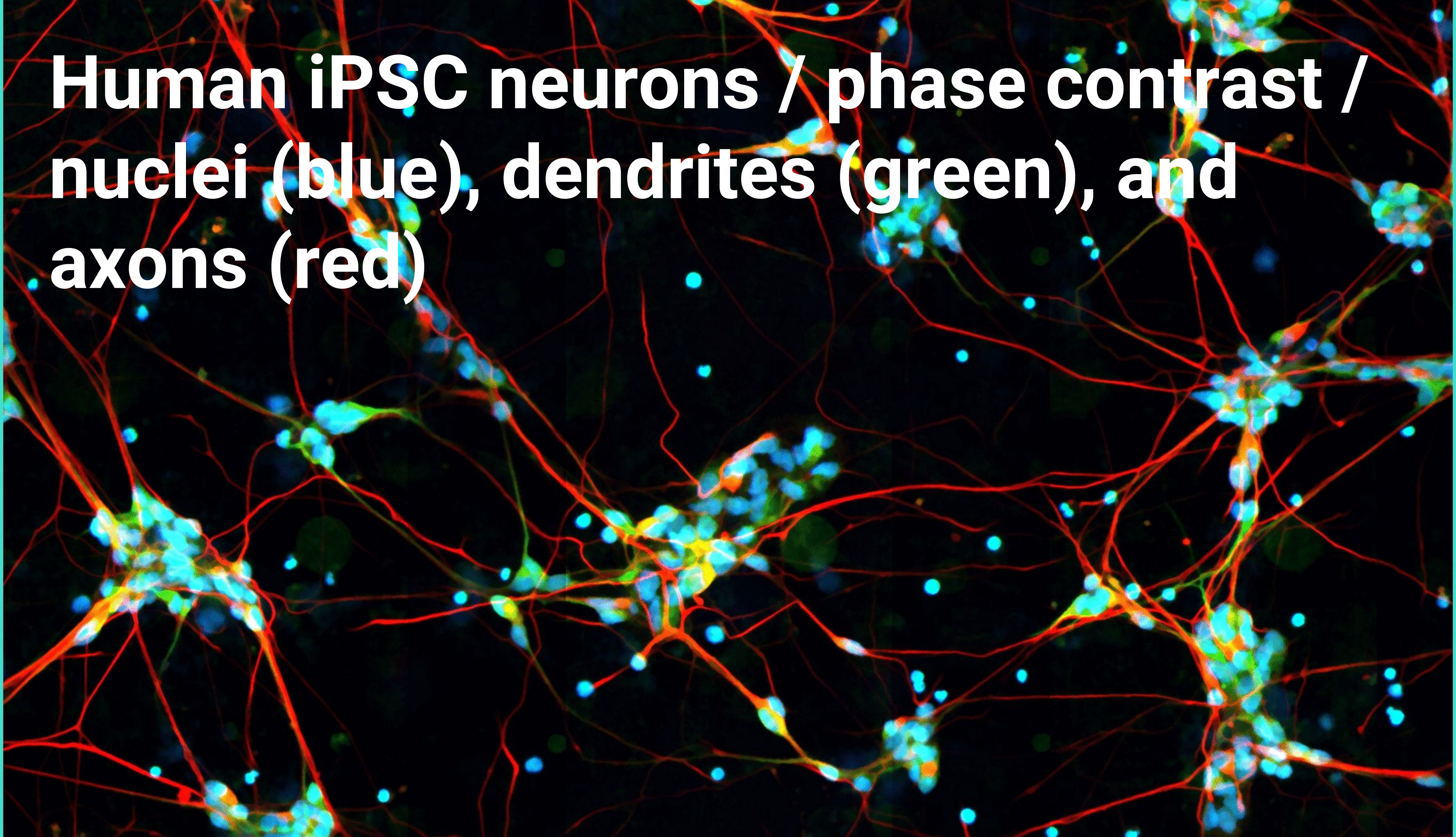



Human cancer cells / DIC / nuclei (blue)
and cell mask (green)

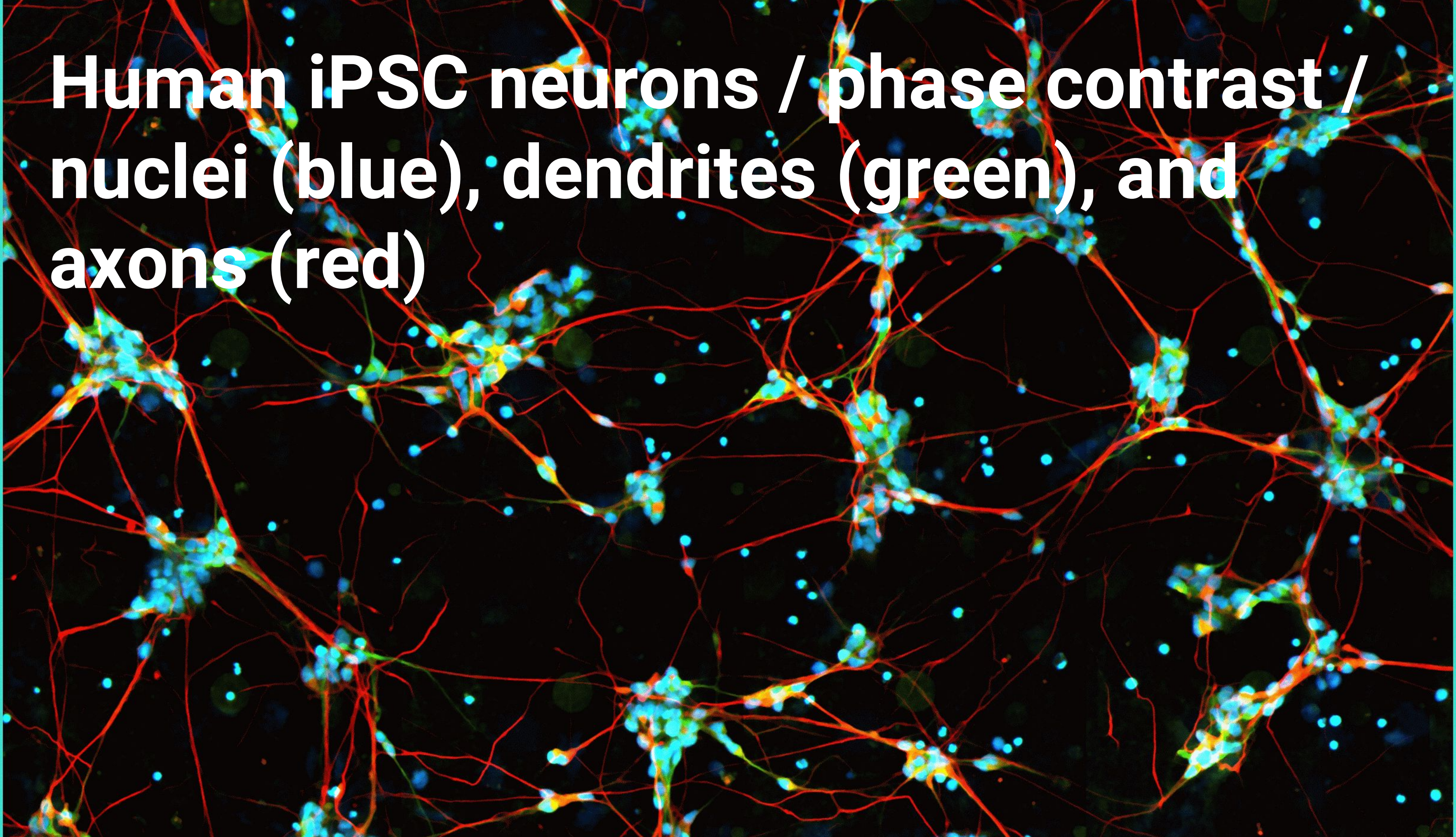
Human cancer cells / DIC / nuclei (blue) and cell mask (green)

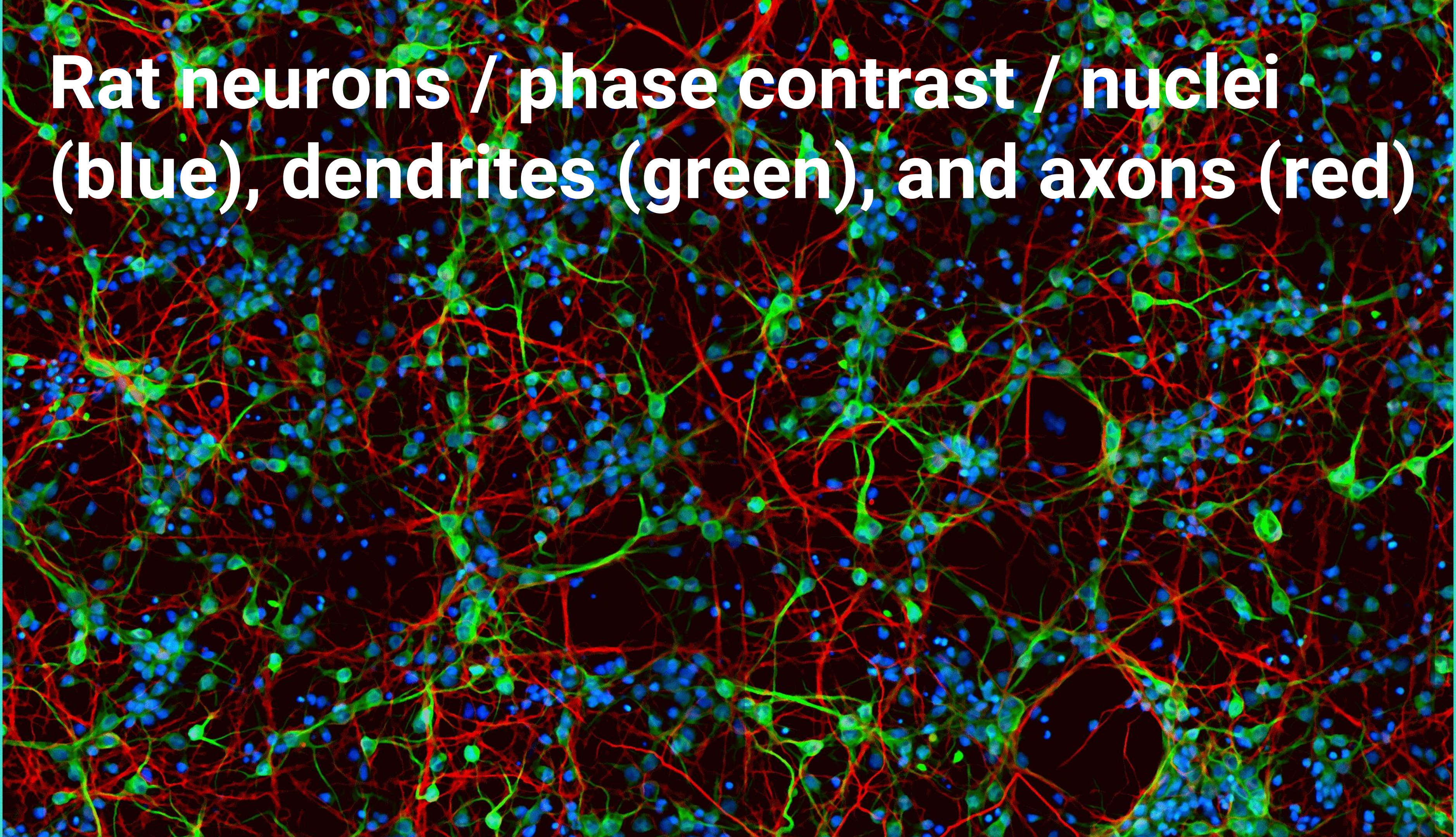


Human iPSC neurons / phase contrast / nuclei (blue), dendrites (green), and axons (red)



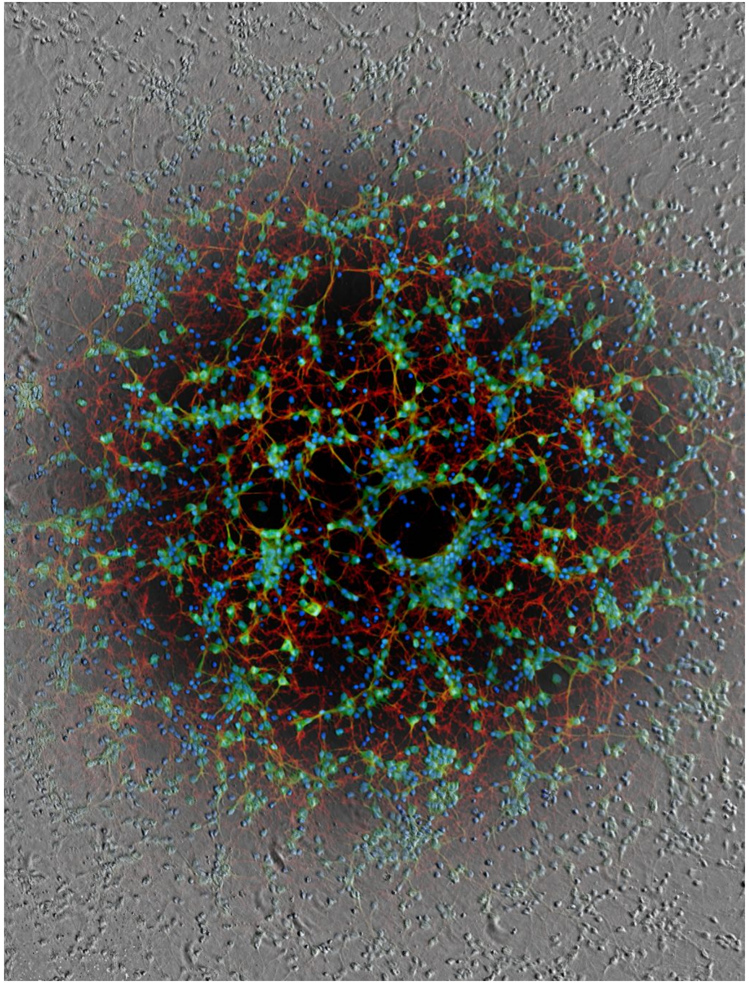
Human iPSC neurons / phase contrast / nuclei (blue), dendrites (green), and axons (red)




Rat neurons / phase contrast / nuclei
(blue), dendrites (green), and axons (red)

Many applications

- Measure live cells that can’t be modified
- Impute labels on live cells in longitudinal studies to build predictors of ultimate fate
- Map the entire cellular neighborhood for experiments coupled to partial transfection
- Impute unlimited number of labels; no longer limited by microscope’s channels



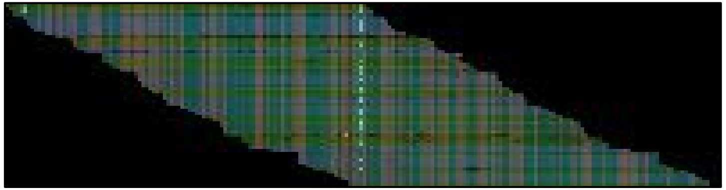
Deep learning for genomics

Google

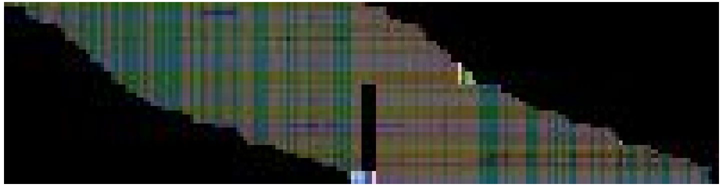
Recasting variant calling for deep learning

Encode reads and reference as images

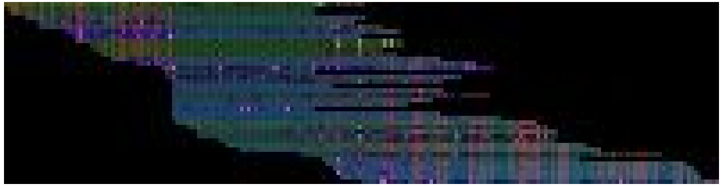
True SNPs



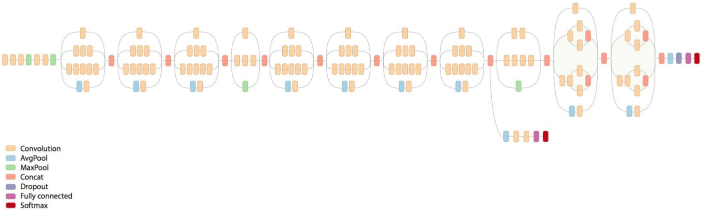
True Indels



False variants



Use inception-v3 to call variant genotype



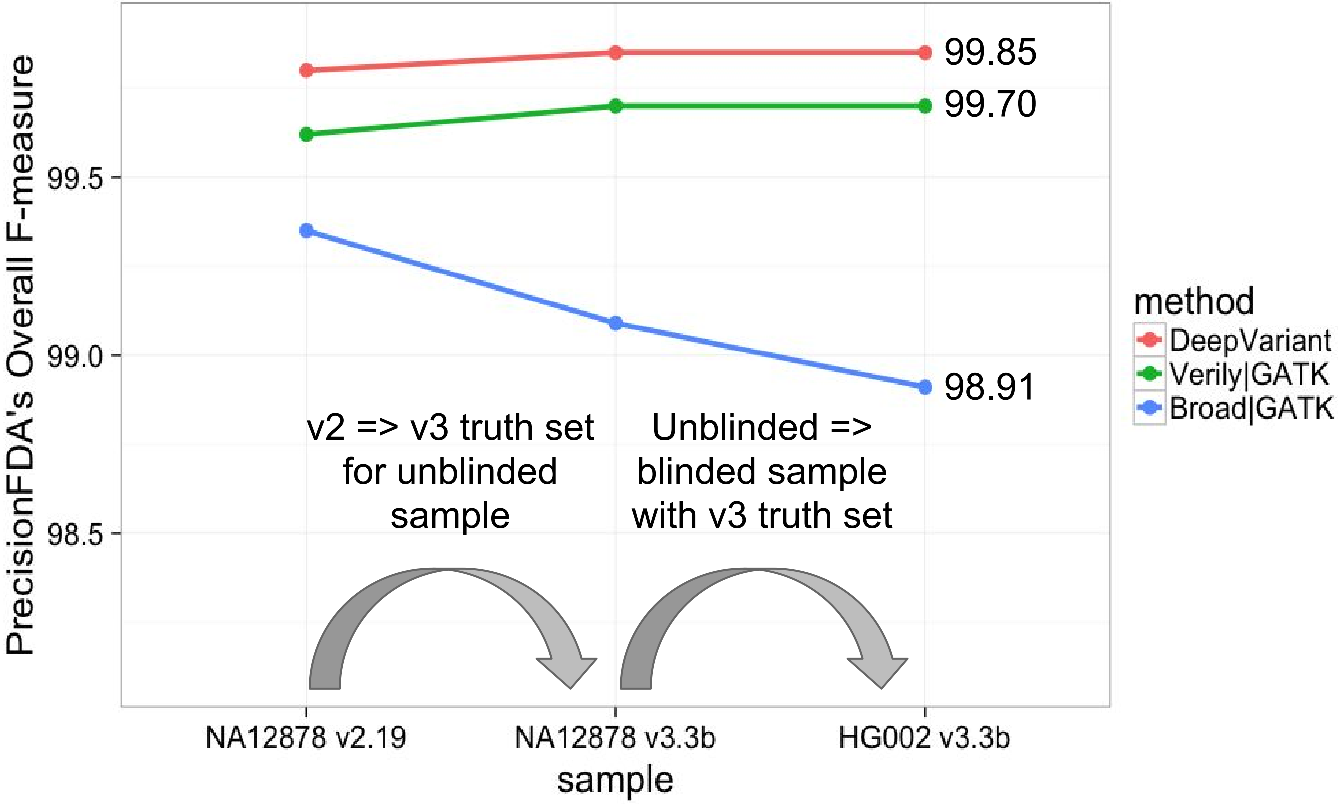
Train on Genome in a Bottle sample using their genotype labels

Each germline WGS dataset provides e.g., ~3.7M labeled varaints for training:

- 215K false positives candidates
- 2.1M true heterozygotes
- 1.3M true homozygous alternates

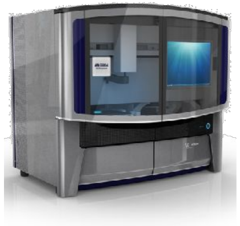
Encoding is roughly red = {A,C,G,T}; green = {quality score}; blue = {read strand}; alpha = {matches ref genome}

DeepVariant won an award at the 2016 PrecisionFDA competition



F-measure is the harmonic mean of precision and recall.

DeepVariant can learn to call variants in many sequencing technologies


Dataset	10X Chromium 75x WGS	PacBio raw reads 40x WGS	SOLID SE 85x WGS	Illumina TruSeq exome
DeepVariant (F1 metric)	**99.3%**	**92.7%**	**86.4%**	**96.1%**
Comparator (F1 metric)	98.2%	56.1%[1]	78.8%[2]	95.4%
Comparator caller	Long Ranger	samtools	GATK	ensemble

[1]No standard caller exists for this technology for human samples;
[2]Old technology without any maintained variant callers.

Machine learning for Predictive Tasks in Healthcare

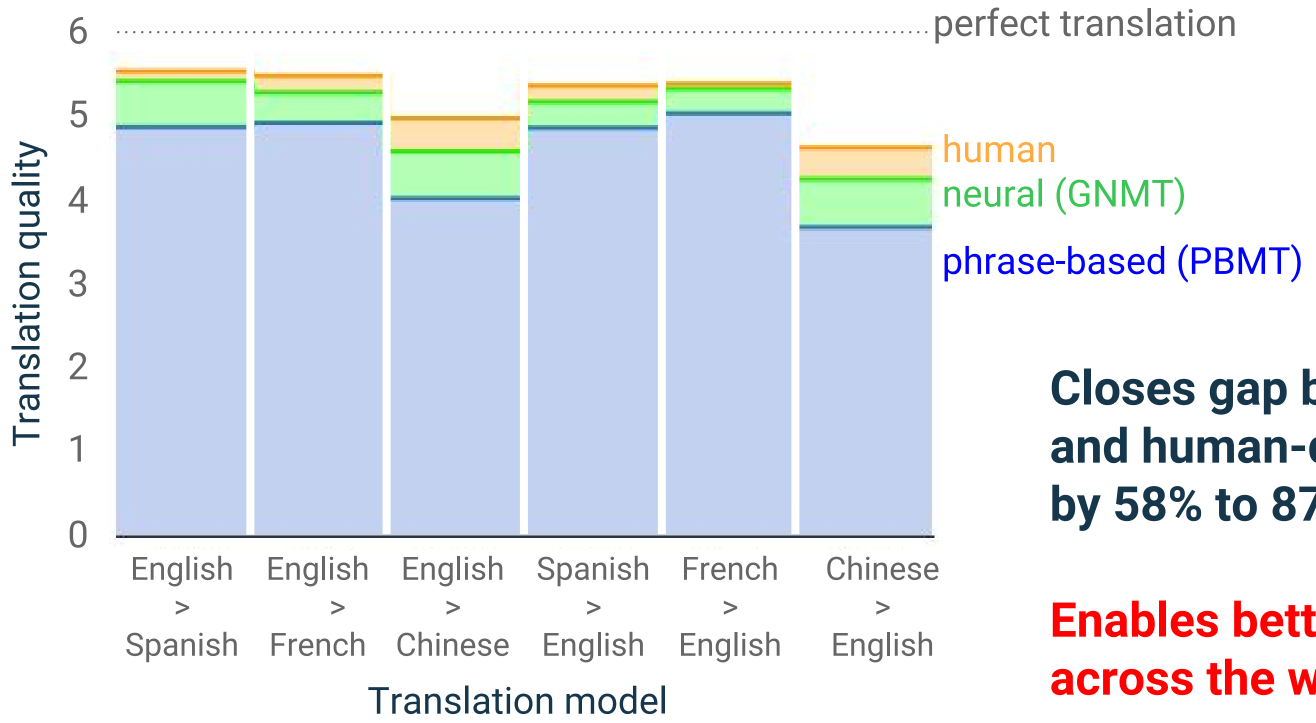
Predictive tasks for healthcare

Given a patient's EMR data, **can we predict the future**?

Deep learning methods for sequential prediction are becoming extremely good

e.g. recent improvements in Google Translation

Neural Machine Translation



Closes gap between old system and human-quality translation by 58% to 87%

Enables better communication across the world

research.googleblog.com/2016/09/a-neural-network-for-machine.html

Predictive tasks for healthcare

Given a large corpus of training data of de-identified medical records, can we predict interesting aspects of the future for a patient not in the training set?

- *will patient be readmitted to hospital in next N days?*
- *what is the likely length of hospital stay for patient checking in?*
- *what are the most likely diagnoses for the patient right now? and why?*
- *what medications should a doctor consider prescribing?*
- *what tests should be considered for this patient?*
- *which patients are at highest risk for X in next month?*

Collaborating with several healthcare organizations, including UCSF, Stanford, and Univ. of Chicago. Have early very promising results (no public paper yet)

If your org might be interested in working with us, see contact link at bottom of:

g.co/brain/healthcare

Machine learning for Quantum Chemistry

Google

Levels of quantum theory

Paul Dirac: “The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the **exact application of these laws leads to equations much too complicated to be soluble**.

The community has spent years developing approximations that can be practically applied

Electronic structure methods

Valence bond theory

Generalized valence bond
Modern valence bond
Resonance

Molecular orbital theory

Hartree–Fock method
Semi-empirical quantum chemistry methods
Møller–Plesset perturbation theory
Configuration interaction
Coupled cluster
Multi-configurational self-consistent field
Quantum chemistry composite methods
Quantum Monte Carlo
Linear combination of atomic orbitals

Electronic band structure

Nearly free electron model
Tight binding
Muffin-tin approximation
Density functional theory
k·p perturbation theory
Empty lattice approximation

https://en.wikipedia.org/wiki/Ab_initio_quantum_chemistry_methods

Google Research Blog

Predicting Properties of Molecules with Machine Learning

Friday, April 07, 2017

https://research.googleblog.com/2017/04/predicting-properties-of-molecules-with.html

Fast machine learning models of electronic and energetic properties consistently reach approximation errors better than DFT accuracy

Felix A. Faber,[1,2] Luke Hutchison,[3,2] Bing Huang,[1] Justin Gilmer,[4,*] Samuel S. Schoenholz,[4,*] George E. Dahl,[4] Oriol Vinyals,[5] Steven Kearnes,[3] Patrick F. Riley,[3] and O. Anatole von Lilienfeld[1,†]

[1] *Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Switzerland.*
[2] *Authors contributed equally*
[3] *Google, CA, USA.*
[4] *Google Brain, CA, USA.*
[5] *Google DeepMind, London, UK.*

(Dated: February 21, 2017)

Neural Message Passing for Quantum Chemistry

To appear in ICML 2017

Justin Gilmer [1] **Samuel S. Schoenholz** [1] **Patrick F. Riley** [2] **Oriol Vinyals** [3] **George E. Dahl** [4]

https://arxiv.org/abs/1702.05532 and https://arxiv.org/abs/1704.01212

Basic Idea: Use New Kind of Neural Net Model to Approximate Really Expensive Computation

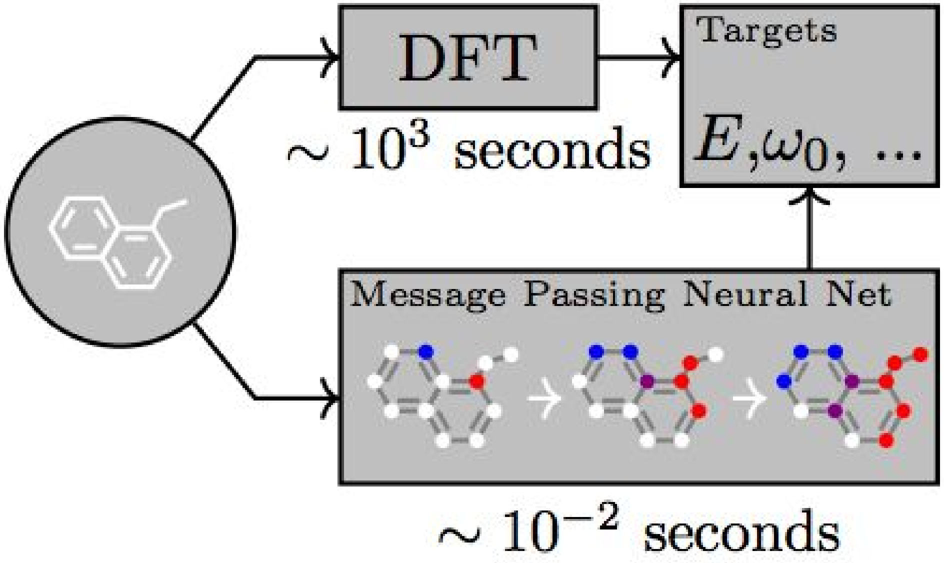


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

To appear in ICML 2017

https://arxiv.org/abs/1704.01212

Machine learning approximation error below DFT error

	μ	α	$\varepsilon_{\mathrm{HO}}$	$\varepsilon_{\mathrm{LU}}$	$\Delta\varepsilon$	$\langle R^2 \rangle$	ZPVE	U_0	C_v	ω_1
	Debye	Bohr^3	eV	eV	eV	Bohr^2	eV	eV	cal/molK	cm^{-1}
Mean	2.67	75.3	-6.54	0.322	6.86	1190	4.06	-76.6	31.6	3500
MAD	1.17	6.29	0.439	1.05	1.07	203	0.717	8.19	3.21	238
Target	0.1 [6,41]	0.1 [6,41]	0.0434 [6]	0.0434 [6]	0.0434 [6]	1.2 [6]	0.00124 [6,42]	0.0434 [6]	0.05 [6]	10 [6,42]
DFT/ B3LYP	0.1 [41]	0.1 [41]	0.24 [43]	0.088 [43]	0.22 [43]	-	0.0097 [39]	0.12 [44]	0.34 [45]	28 [39]

MAD: mean average deviation
Target: Target mean average error values generally accepted as “accurate enough” by chemistry community
DFT/B3LYP: Mean average error obtained by very computationally expensive density functional theory simulations

https://arxiv.org/abs/1702.0553

Machine learning approximation error below DFT error

		μ	α	ε_{HO}	ε_{LU}	$\Delta\varepsilon$	$\langle R^2 \rangle$	ZPVE	U_0	C_v	ω_1
		Debye	Bohr^3	eV	eV	eV	Bohr^2	eV	eV	cal/molK	cm^{-1}
Mean		2.67	75.3	-6.54	0.322	6.86	1190	4.06	-76.6	31.6	3500
MAD		1.17	6.29	0.439	1.05	1.07	203	0.717	8.19	3.21	238
Target		0.1[6,41]	0.1[6,41]	0.0434[6]	0.0434[6]	0.0434[6]	1.2[6]	0.00124[6,42]	0.0434[6]	0.05[6]	10[6,42]
DFT/ B3LYP		0.1[41]	0.1[41]	0.24[43]	0.088[43]	0.22[43]	-	0.0097[39]	0.12[44]	0.34[45]	28[39]
GG	MG	0.238	**0.151**	0.0587	0.0564	**0.0835**	5.98	0.00291	0.0317	0.0724	6.32
GC	MG	**0.0696**	0.227	**0.0509**	**0.0471**	**0.0766**	5.68	0.00975	0.13	0.0892	3.15
		✓	✗	✓	✓	✓		✓	✓	✓	✓

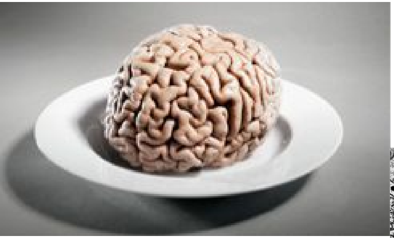
https://arxiv.org/abs/1702.0553

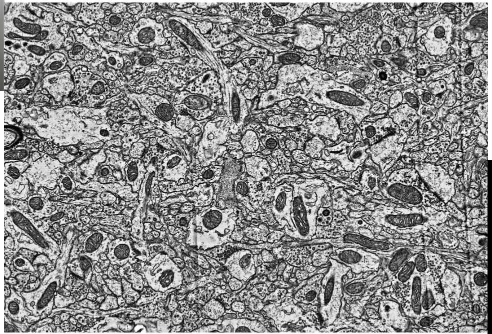
Connectomics

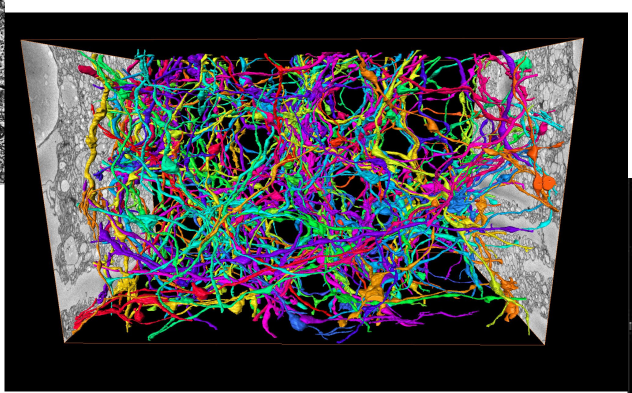
Machine Learning for
High-throughput Neuroanatomy

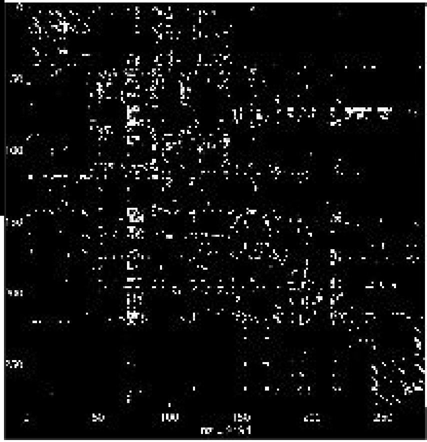
Google

Connectomics: Reconstructing Neural Circuits from High-Resolution Brain Imaging









Connectomics Timeline

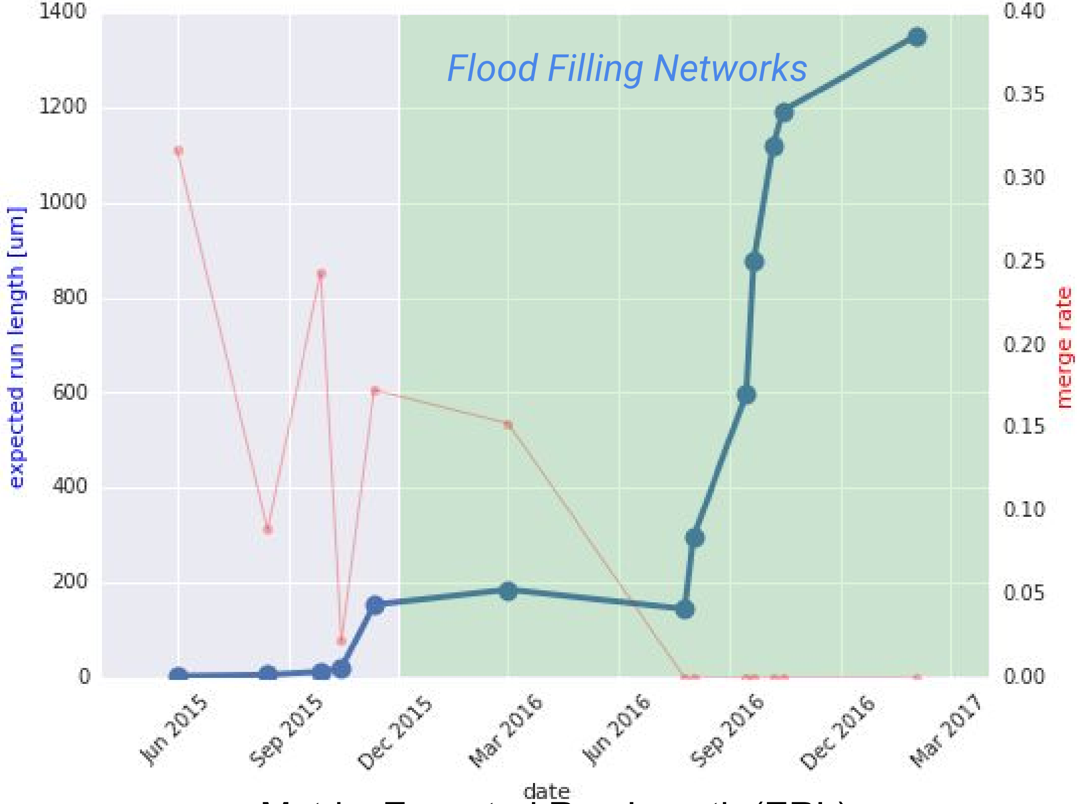
- 1970's: c. elegans
  300 neurons; printed photos; MRC in the UK

- 2013: retina, drosophila visual system
  1000 neurons; O(gigavoxels); Max Planck Institute, MIT, HHMI

- In progress: mouse cortical column, whole drosophila brain, song-bird
  <100,000 neurons; O(teravoxels); Harvard, HHMI, Allen Institute, MPI

- Machines now being designed & built for whole mouse brain
  100,000,000 neurons; O(petavoxels)

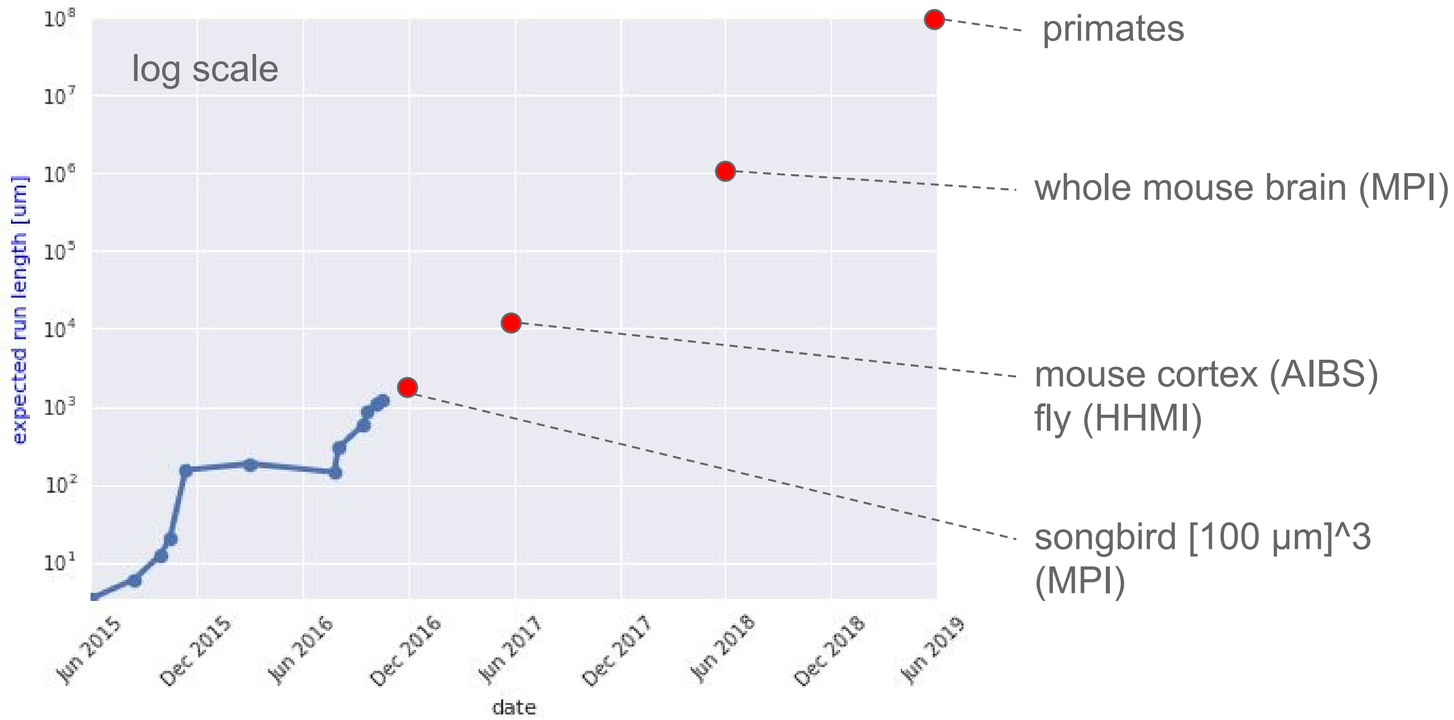
Automated Reconstruction Progress at Google



Metric: Expected Run Length (ERL)
"mean microns between failure" of automated neuron tracing

Google

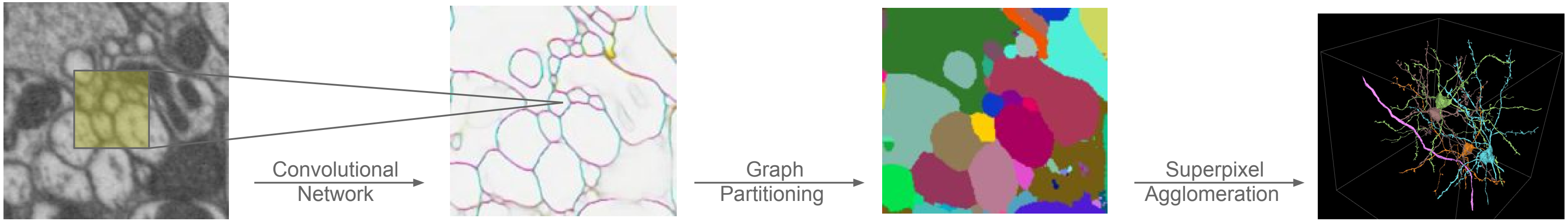
Automated Reconstruction Progress at Google



Metric: Expected Run Length (ERL)
“mean microns between failure” of automated neuron tracing

Classical Approach

- Classical segmentation pipeline:
  - Predict local image boundaries (edge detectors ['70s] → SVMs ['00s] → deep learning [today]).
  - Discrete graph partitioning (watershed, connected components, etc.) without machine learning.
  - Agglomeration of "superpixels" using classifier or heuristics.



- Issues with classical approach:
  - Lack of end-to-end training; certain steps not differentiable.
  - Pipeline complexity.
  - Graph partitioning procedure is very brittle and impoverished.

New Technology: Flood Filling Networks

Flood-Filling Networks

Michał Januszewski
Google
mjanusz@google.com

Jeremy Maitin-Shepard
Google
jbms@google.com

Peter Li
Google
phli@google.com

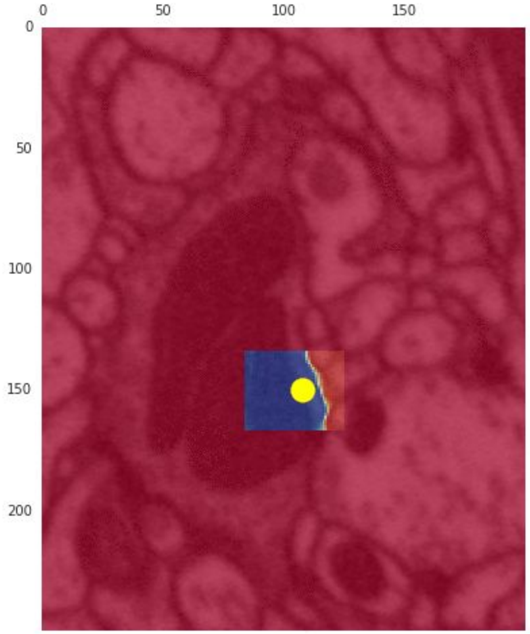
Jörgen Kornfeld
Max Planck Institute for Neurobiology
kornfeld@neuro.mpg.de

Winfried Denk
Max Planck Institute for Neurobiology
winfried.denk@neuro.mpg.de

Viren Jain
Google
viren@google.com

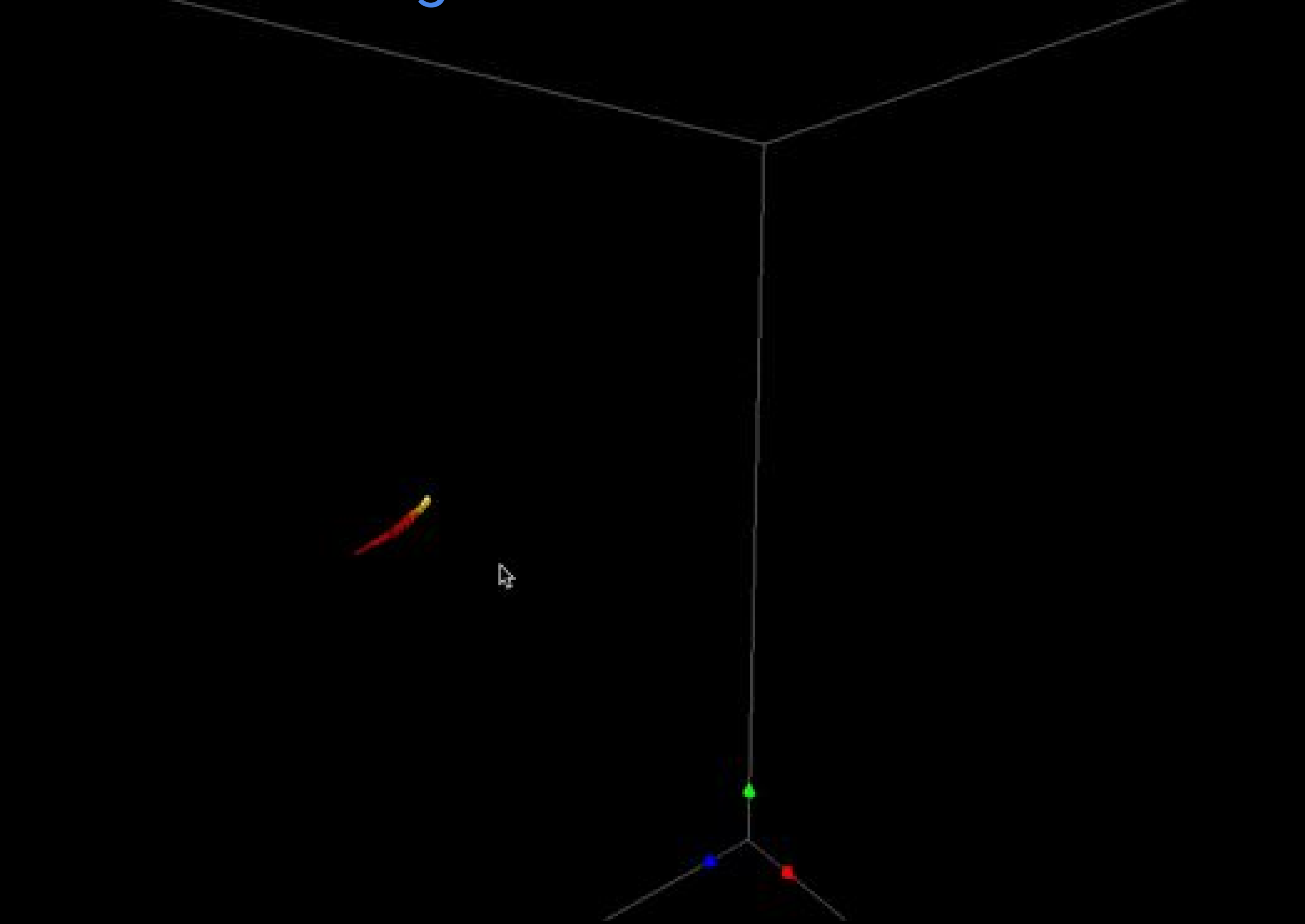
- Start with a seed point
- Recurrent neural network iteratively fills out an object based on image content and its own previous predictions

2d Inference



https://arxiv.org/abs/1611.00421

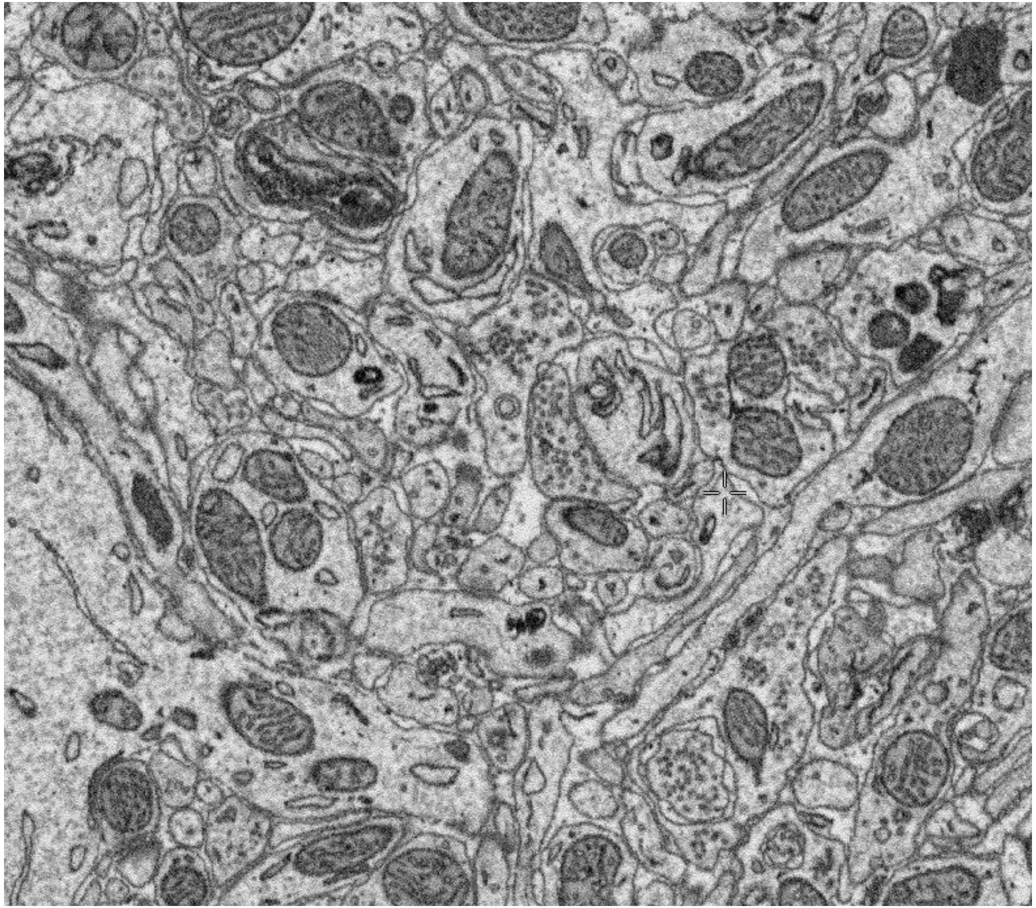
Flood Filling Networks: 3d Inference



- Application: songbird brain imaged by Max Planck Institute for Neurobiology using serial block face scanning electron microscopy

- 10,600 × 10,800 × 5,700 voxels = ~600 billion voxels

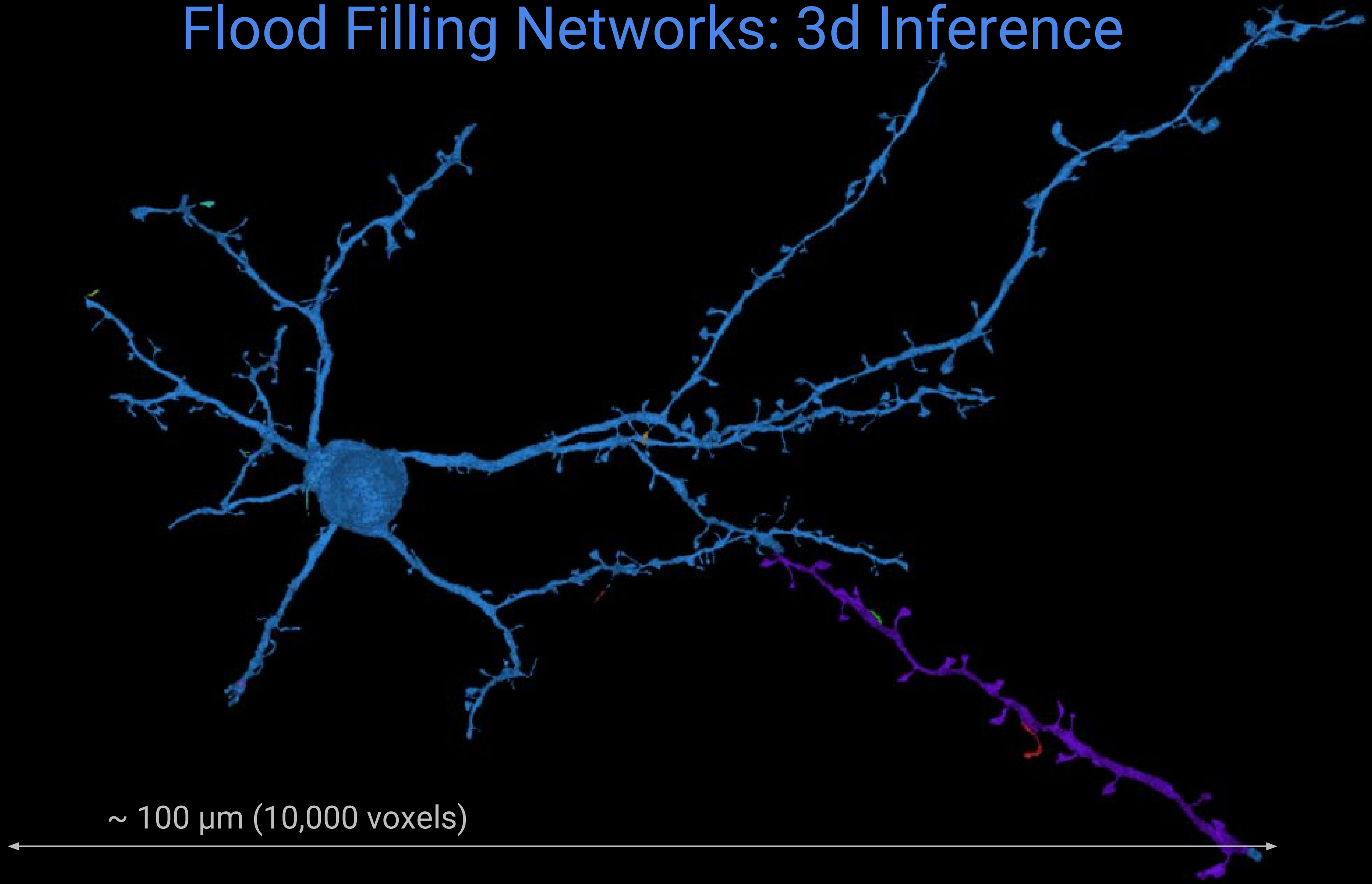
- Successful reconstruction of the wiring diagram will test specific hypotheses related to how biological nervous systems produce precise, sequential motor behaviors and perform reinforcement learning.



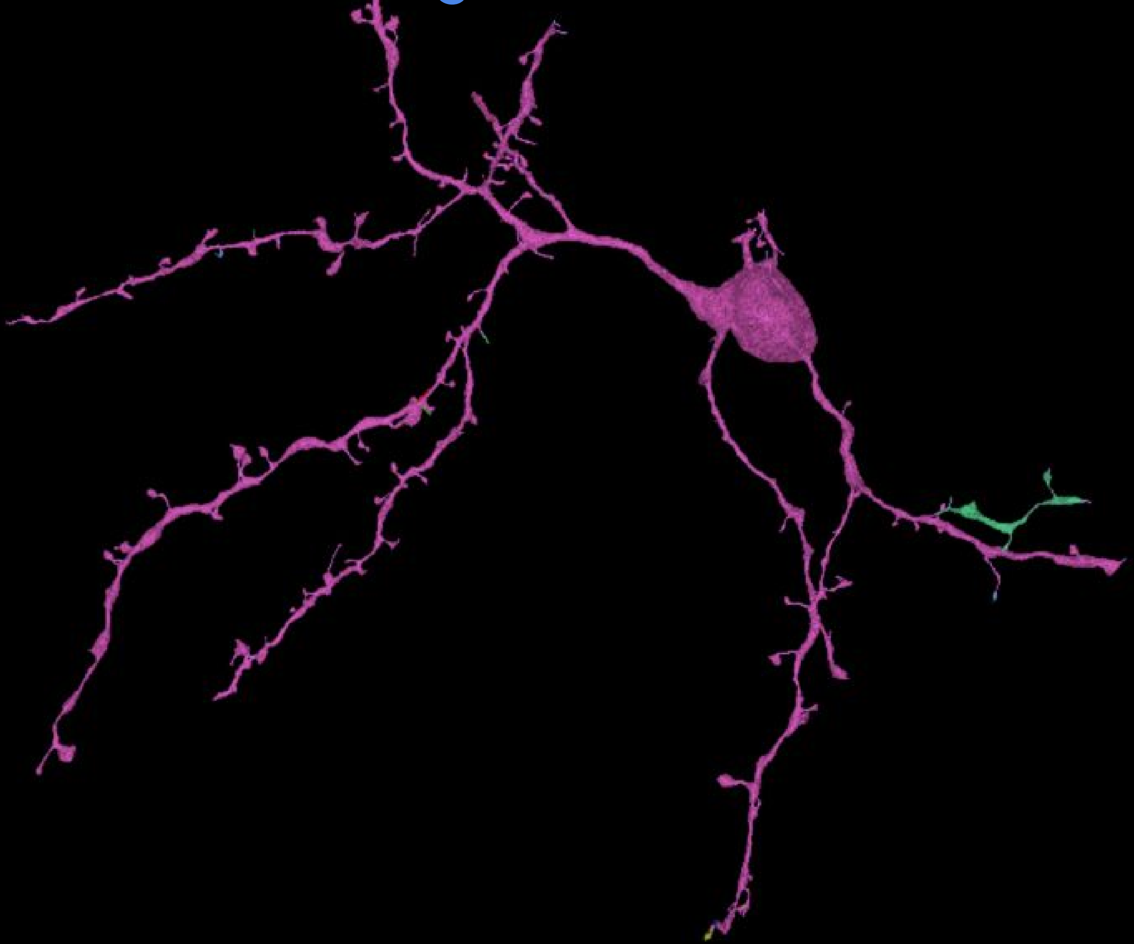
Courtesy Jorgen Kornfeld & Winfried Denk, MPI

Google

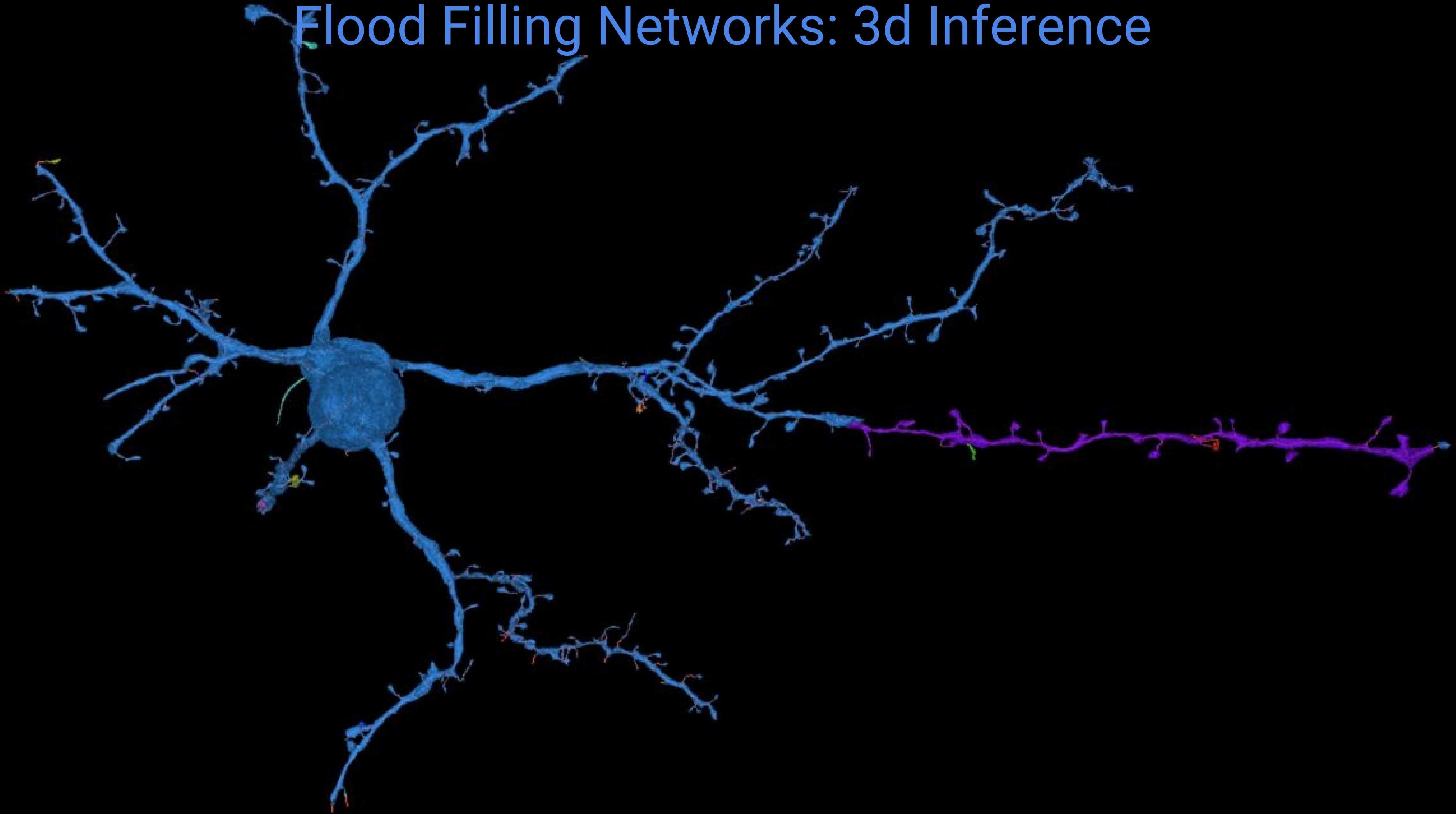
Flood Filling Networks: 3d Inference



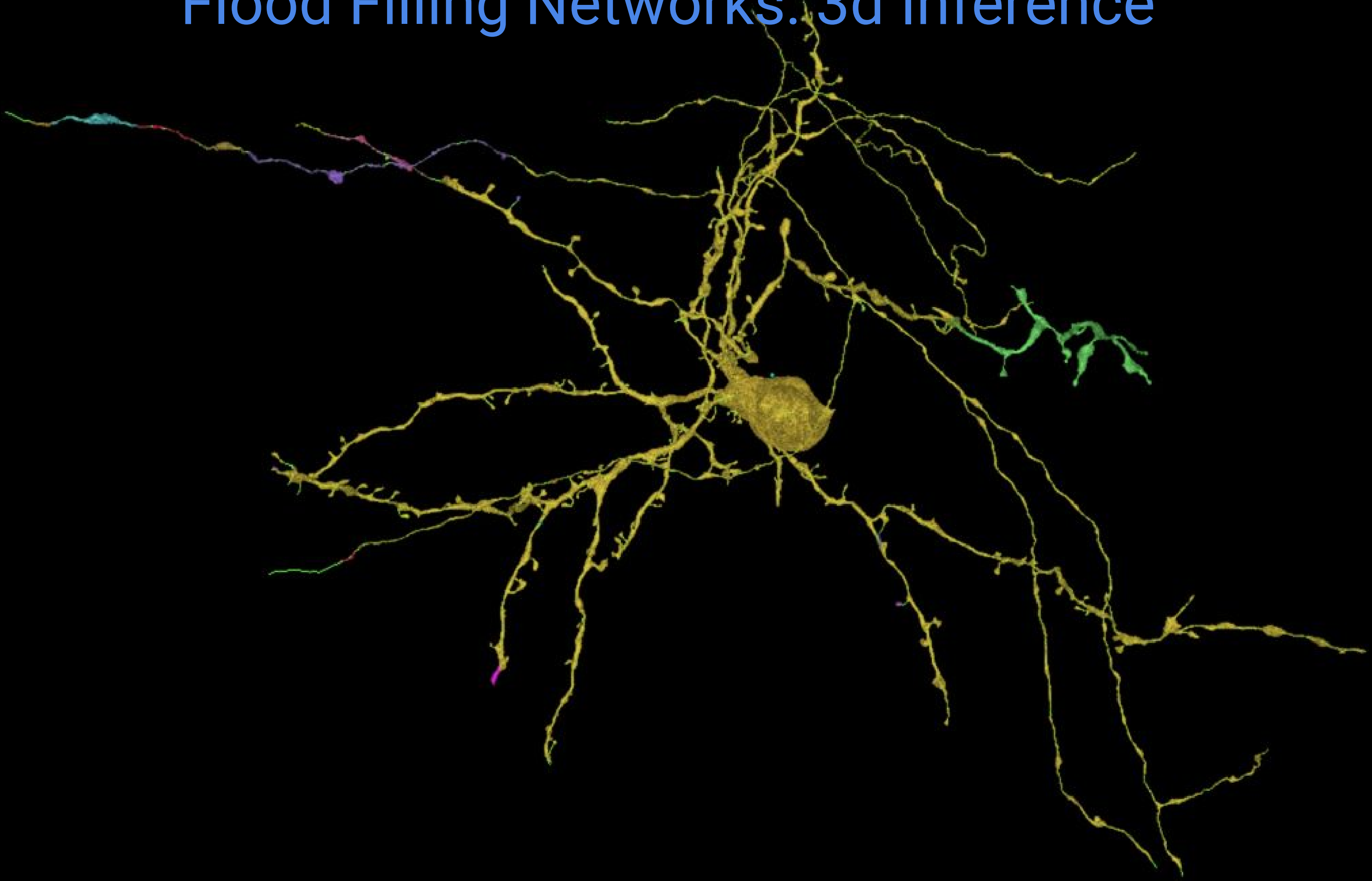
Flood Filling Networks: 3d Inference



Flood Filling Networks: 3d Inference



Flood Filling Networks: 3d Inference



Google

Automated machine learning ("learning to learn")

Google

Current:

Solution = ML expertise + data + computation

Current:

Solution = ML expertise + data + computation

Can we turn this into:

Solution = data + 100X computation

???

Early encouraging signs

Trying multiple different approaches:

1. Reinforcement learning-based architecture search
2. Model architecture evolution

NEURAL ARCHITECTURE SEARCH WITH REINFORCEMENT LEARNING

Barret Zoph,* Quoc V. Le
Google Brain
{barretzoph,qvl}@google.com

Appeared earlier this year at ICLR 2017 in France

Idea: model-generating model trained via RL

(1) Generate ten models
(2) Train them for a few hours
(3) Use loss of the generated models as reinforcement learning signal

arxiv.org/abs/1611.01578

CIFAR-10 Image Recognition Task

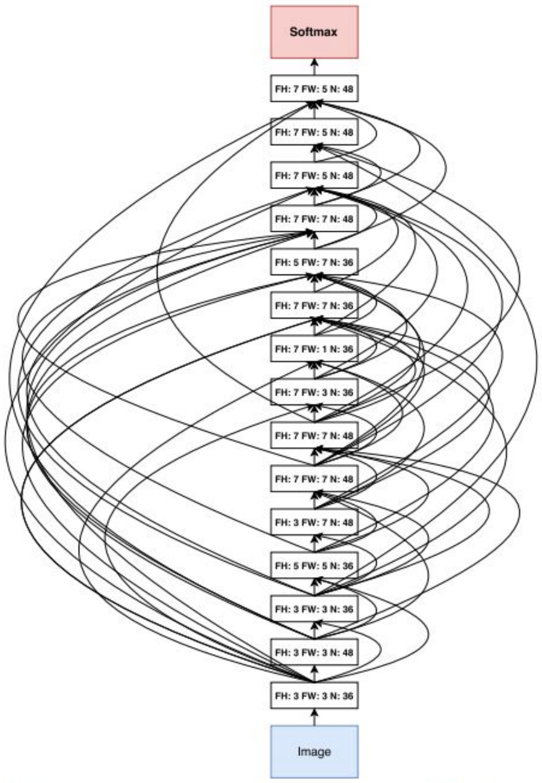


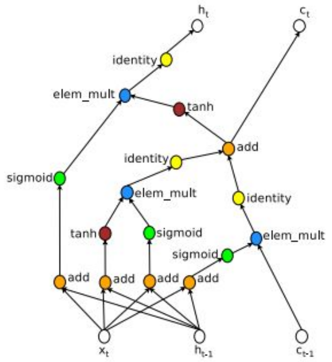
Figure 7: Convolutional architecture discovered by our method, when the search space does not have strides or pooling layers. FH is filter height, FW is filter width and N is number of filters.

Model	Depth	Parameters	Error rate (%)
Network in Network (Lin et al., 2013)	-	-	8.81
All-CNN (Springenberg et al., 2014)	-	-	7.25
Deeply Supervised Net (Lee et al., 2015)	-	-	7.97
Highway Network (Srivastava et al., 2015)	-	-	7.72
Scalable Bayesian Optimization (Snoek et al., 2015)	-	-	6.37
FractalNet (Larsson et al., 2016)	21	38.6M	5.22
with Dropout/Drop-path	21	38.6M	4.60
ResNet (He et al., 2016a)	110	1.7M	6.61
ResNet (reported by Huang et al. (2016b))	110	1.7M	6.41
ResNet with Stochastic Depth (Huang et al., 2016b)	110	1.7M	5.23
	1202	10.2M	4.91
Wide ResNet (Zagoruyko & Komodakis, 2016)	16	11.0M	4.81
	28	36.5M	4.17
ResNet (pre-activation) (He et al., 2016b)	164	1.7M	5.46
	1001	10.2M	4.62
DenseNet ($L = 40, k = 12$) Huang et al. (2016a)	40	1.0M	5.24
DenseNet($L = 100, k = 12$) Huang et al. (2016a)	100	7.0M	4.10
DenseNet ($L = 100, k = 24$) Huang et al. (2016a)	100	27.2M	$\boxed{3.74}$
Neural Architecture Search v1 no stride or pooling	15	4.2M	5.50
Neural Architecture Search v2 predicting strides	20	2.5M	6.01
Neural Architecture Search v3 max pooling	39	7.1M	4.47
Neural Architecture Search v3 max pooling + more filters	39	32.0M	$\boxed{3.84}$

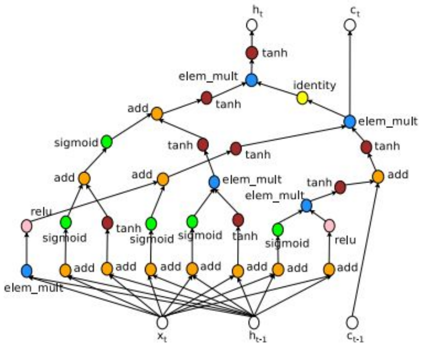
Table 1: Performance of Neural Architecture Search and other state-of-the-art models on CIFAR-10.

Penn Tree Bank Language Modeling Task

"Normal" LSTM cell



Cell discovered by architecture search



Model	Parameters	Test Perplexity
Mikolov & Zweig (2012) - KN-5	2M‡	141.2
Mikolov & Zweig (2012) - KN5 + cache	2M‡	125.7
Mikolov & Zweig (2012) - RNN	6M‡	124.7
Mikolov & Zweig (2012) - RNN-LDA	7M‡	113.7
Mikolov & Zweig (2012) - RNN-LDA + KN-5 + cache	9M‡	92.0
Pascanu et al. (2013) - Deep RNN	6M	107.5
Cheng et al. (2014) - Sum-Prod Net	5M‡	100.0
Zaremba et al. (2014) - LSTM (medium)	20M	82.7
Zaremba et al. (2014) - LSTM (large)	66M	78.4
Gal (2015) - Variational LSTM (medium, untied)	20M	79.7
Gal (2015) - Variational LSTM (medium, untied, MC)	20M	78.6
Gal (2015) - Variational LSTM (large, untied)	66M	75.2
Gal (2015) - Variational LSTM (large, untied, MC)	66M	73.4
Kim et al. (2015) - CharCNN	19M	78.9
Press & Wolf (2016) - Variational LSTM, shared embeddings	24M	73.2
Merity et al. (2016) - Zoneout + Variational LSTM (medium)	20M	80.6
Merity et al. (2016) - Pointer Sentinel-LSTM (medium)	21M	70.9
Zilly et al. (2016) - Variational RHN, shared embeddings	24M	$\boxed{66.0}$
Neural Architecture Search with base 8	32M	67.9
Neural Architecture Search with base 8 and shared embeddings	25M	64.0
Neural Architecture Search with base 8 and shared embeddings	54M	$\boxed{62.4}$

Table 2: Single model perplexity on the test set of the Penn Treebank language modeling task. Parameter numbers with ‡ are estimates with reference to Merity et al. (2016).

More computational power is needed

Deep learning is transforming how we design computers

Special computation properties

reduced precision ok

about 1.2		~~1.21042~~
× about 0.6	**NOT**	~~× 0.61127~~
about 0.7		~~0.73989343~~

Special computation properties

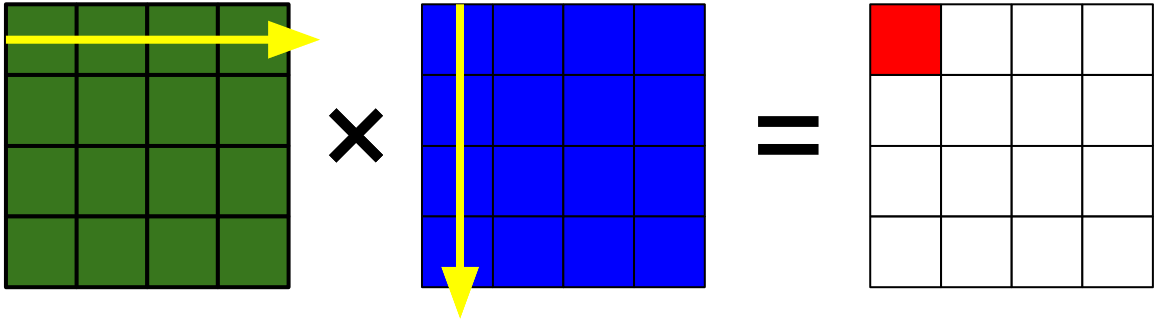
reduced precision ok

about 1.2
× about 0.6
———
about 0.7

NOT

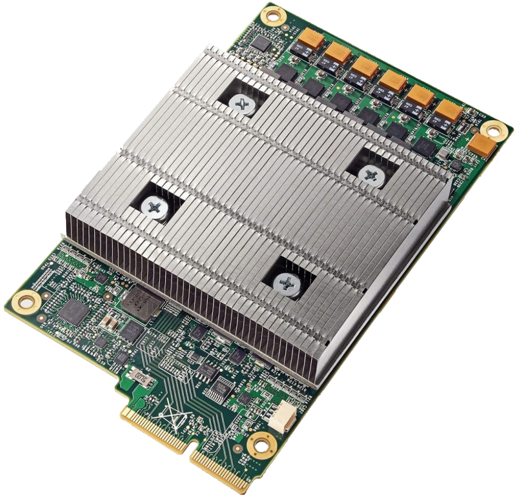
~~1.21042~~
~~× 0.61127~~
———
~~0.73989343~~

handful of specific operations



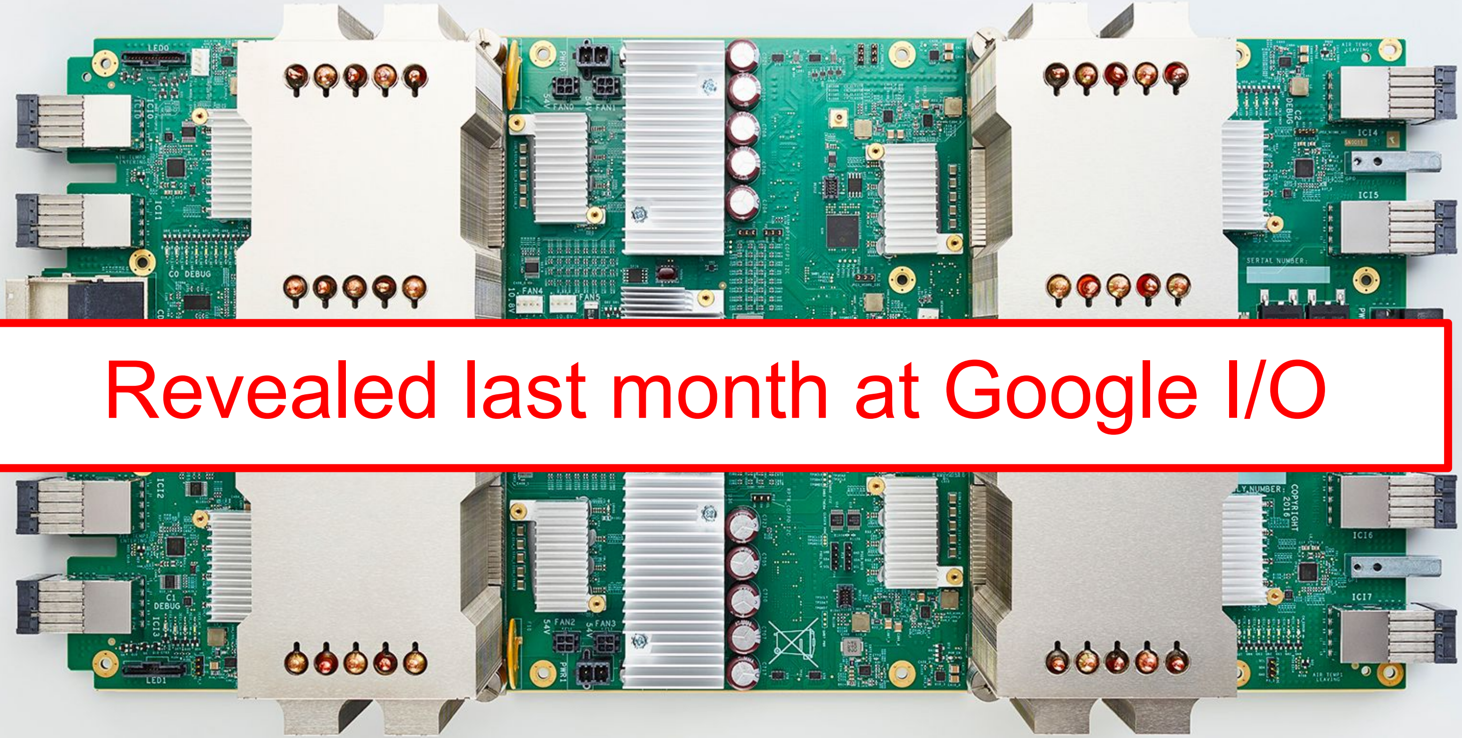
Tensor Processing Unit

Custom Google-designed chip for
neural net computations



In production use for >24 months: used on every search query, for neural machine translation, for AlphaGo match, …

Tensor Processing Unit v2



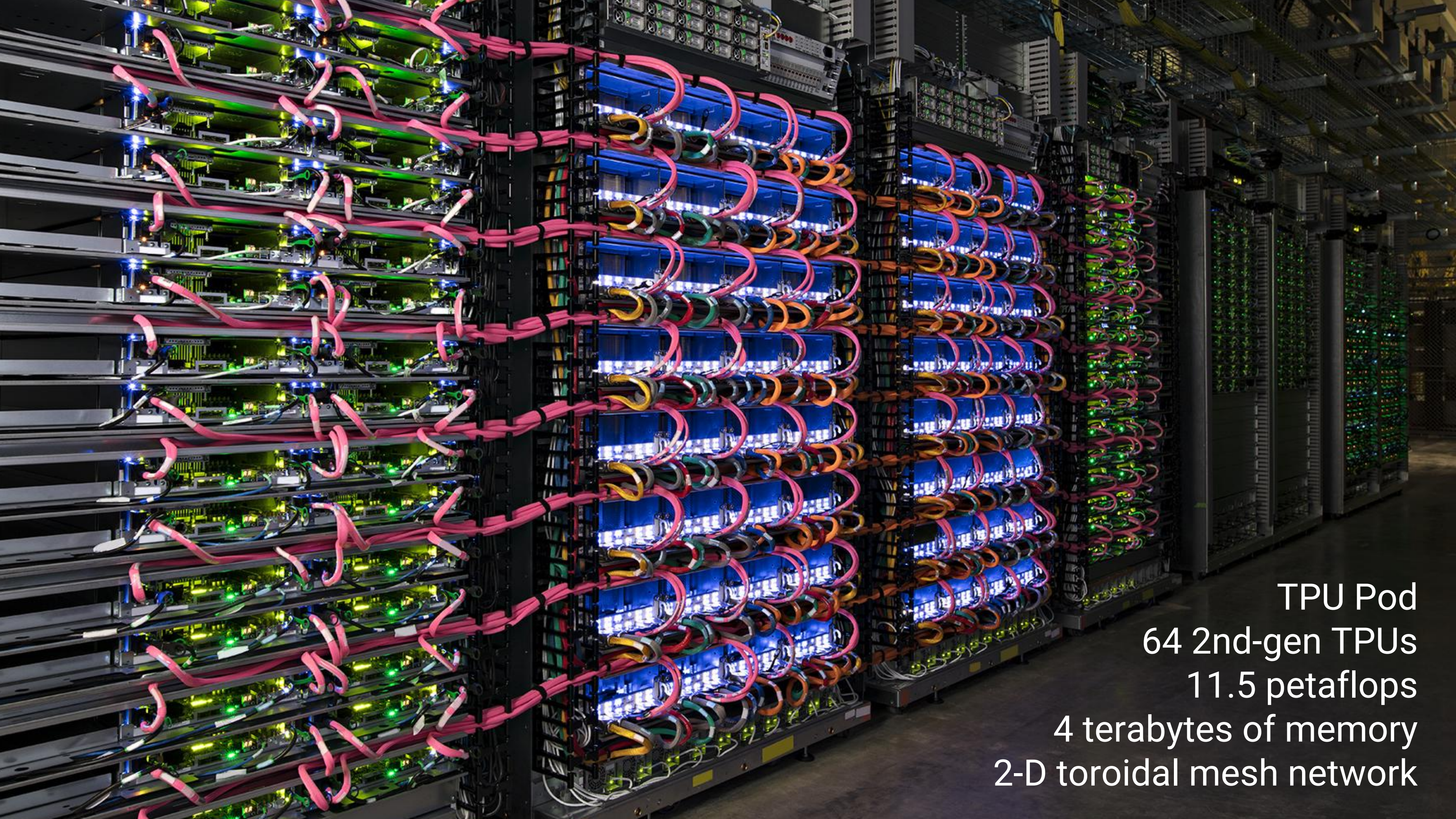
Revealed last month at Google I/O

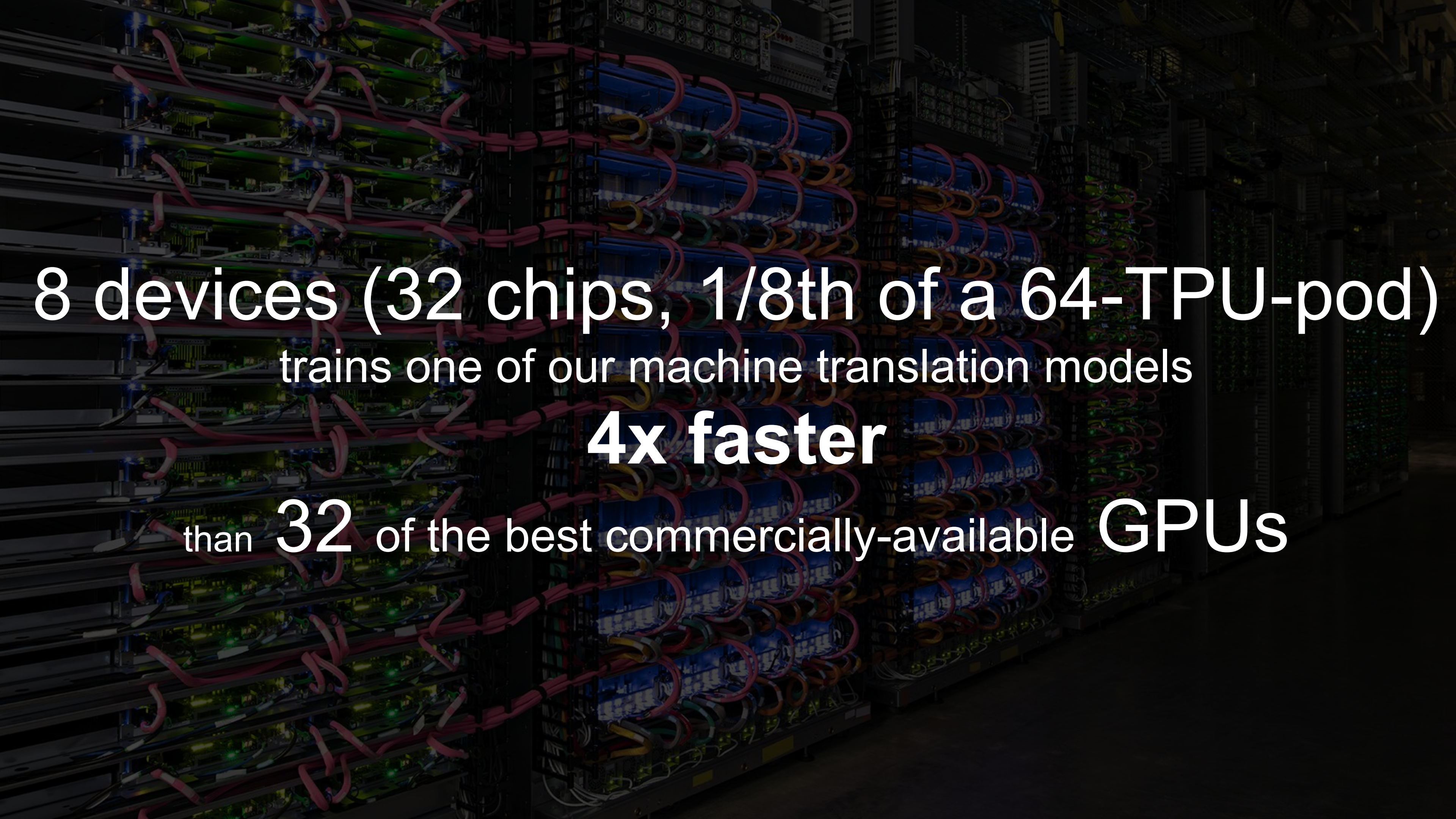
Google-designed device for neural net **training** and **inference**

Google’s 2nd-gen TPU:

- 180 teraflops
- 64 GB of ultra-high-bandwidth memory
- Designed for **training and inference**
- Designed to be connected together


A
B
C
D


TPU Pod
64 2nd-gen TPUs
11.5 petaflops
4 terabytes of memory
2-D toroidal mesh network



8 devices (32 chips, 1/8th of a 64-TPU-pod)

trains one of our machine translation models

4x faster

than 32 of the best commercially-available GPUs

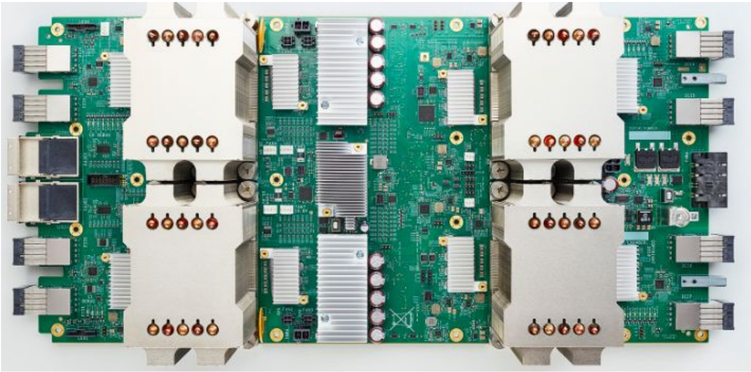
Programmed via TensorFlow

New Estimator interface being added to TF 1.2

Same program will run with only minor modifications on CPUs, GPUs, & TPUs

Will be Available through Google Cloud

Cloud VM with TPU - virtual machine w/180 TFLOPS TPUv2 device attached



TensorFlow Research Cloud (TFRC)

Making 1000 **Cloud VMs with TPUs** available **for free** to top researchers who are committed to **open machine learning research**

Total of 180 PFLOPS: More raw FLOPS than #1 supercomputer in the world

g.co/tpusignup

Conclusions

Deep neural networks and machine learning are starting to produce significant breakthroughs in healthcare and basic science

If you're not considering how to use deep neural nets to solve your problems, **you almost certainly should be**

Conclusions

Deep neural networks and machine learning are starting to produce significant breakthroughs in healthcare and basic science

If you're not considering how to use deep neural nets to solve your problems, **you almost certainly should be**

More info:

g.co/brain and g.co/brain/healthcare and research.google.com

Conclusions

Deep neural networks and machine learning are starting to produce significant breakthroughs in healthcare and basic science

If you're not considering how to use deep neural nets to solve your problems, **you almost certainly should be**

More info:

g.co/brain and g.co/brain/healthcare and research.google.com

Thank you! Questions?