Using Diverse and Difficult Datasets for Training ML

Andrew Carroll - awcarroll@google.com
Genomics team in Google Brain

Google Brain within Google

Google Cloud

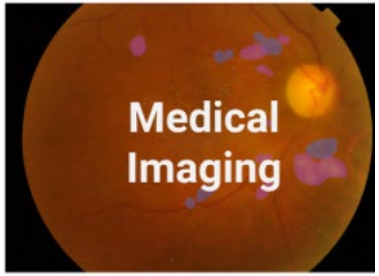
Google Brain

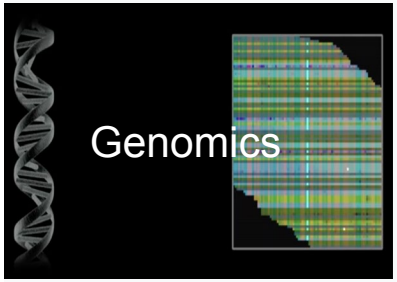
Google AI

Verily

Google Brain in Medicine & Life Sciences

Improving the availability &accuracy of healthcare using ML





Outline of the Talk

- Give an overview of machine learning and deep learning.
- Explain DeepVariant, and example of deep learning applied to genomics

Story #1:

- "Train, don't program" - how to extend deep learning methods using data.

Story #2:

- The importance of Quality: reliable labels and diverse examples.

Story #3:

- The power of building on accelerating technology frameworks.

Story #4:

- Going full circle: understanding biology from the model

Types of Computer Models

Statistical Inference

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.

"Out of 100 experiments, how many times would I find a result as extreme as this."

Machine Learning

Machine learning uses **statistical techniques** to give computer systems the ability to "learn" from data, without explicit programming.

"Using these features I know are important, determine a set of weights and rules to classify a query."

Deep Learning

*Deep Learning uses **machine learning** methods based on learning data representations as opposed to task-specific algorithms.*

"From this group of labelled examples, determine what information is relevant and use it to classify a query."

Sequencing Data Lifecycle

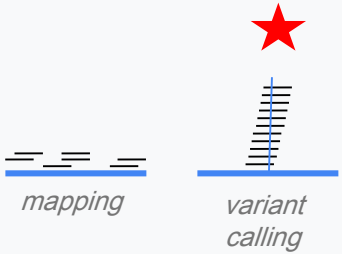
PRIMARY ANALYSIS

raw data

Sample Sequencer

**many* lab protocols and types of sequencers*

SECONDARY ANALYSIS



Data Processing

limited number of data processing tools and steps

TERTIARY ANALYSIS

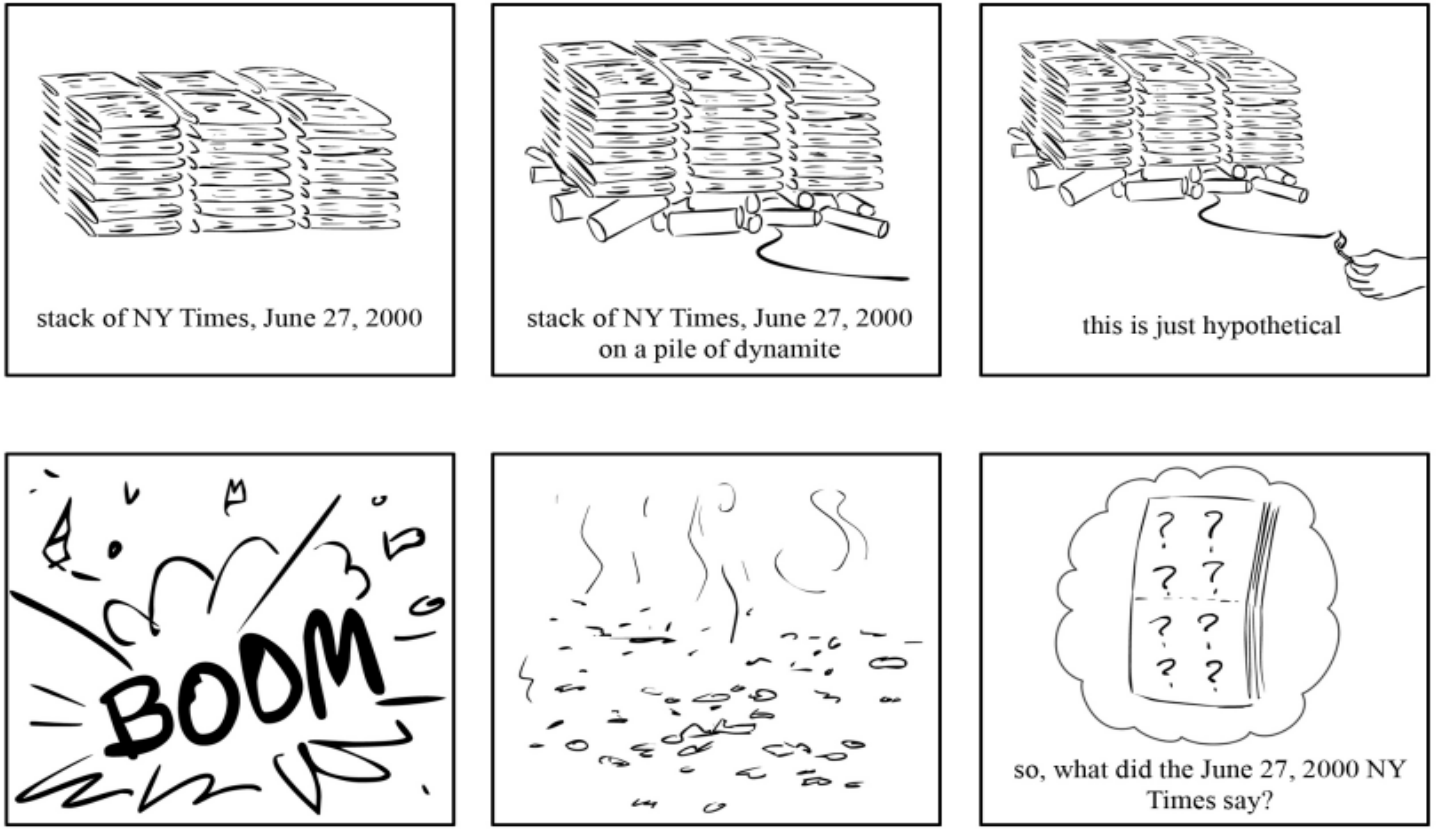
processed data

Clinical (Dx/Tx/PGx)
Biomed Research
Ancestry
...

Analysis/Interpretation

**many* types of analysis*

Analogy for why Variant Calling is Hard



Errors in Sequencing Come from Diverse, Complex Sources

Google AI

Errors come from many uncontrollable sources	Existing statistical techniques work ok	But have well -known drawbacks
Sample DNA itself Sample prep protocol Instrumentation noise Data processing artifacts Errors are correlated among the reads	The most accurate variant callers, such as the GATK, use multiple techniques to control these errors: • Hidden Markov Models • Bayesian inference • Gaussian mixture models All make approximations known to be invalid	Hand-crafted features Hand-optimized parameters Requiring years of work by domain experts Specialized to specific prep, sequencer, toolchain, etc. Making it hard to generalize to new technologies

IS THIS A
CAT or DOG?

CAT DOG

OUTPUT LAYER

ACTIVATED NEURONS

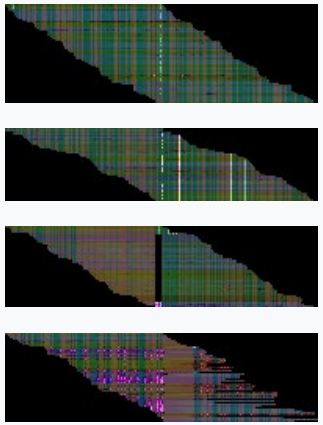
INPUT LAYER

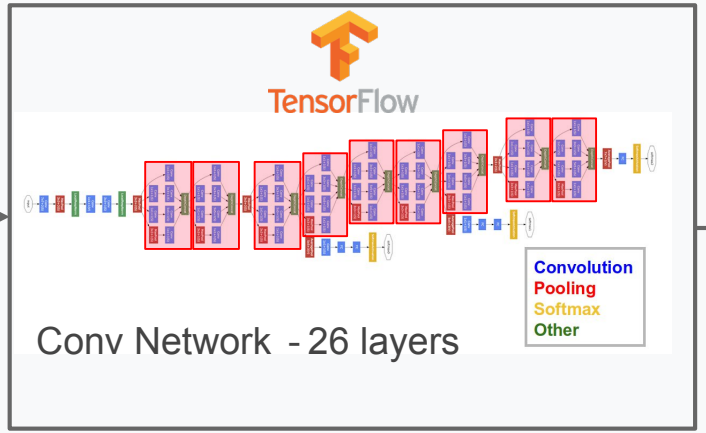
DEEP *neural* NETWORK

DeepVariant: Deep Learning for variant calling

Google AI

Data encoded as "tensors"





Probability distribution

{ HOM_REF, HET, HOM_VAR }

{ 0.001, **0.994**, 0.005 }

{ 0.001, **0.990**, 0.009 }

{ 0.000, 0.001, **0.999** }

{ **0.600**, 0.399, 0.001 }

What Product Metrics is DeepVariant Built for

Google AI

- DeepVariant is the most accurate germline variant caller
- DeepVariant is robust - its accuracy advantage increases on hard datasets
- DeepVariant is fast (70 minutes for a genome on GCP)
- DeepVariant is cheap ($2-$3 on GCP)
- DeepVariant is extensible - retrain for new technologies without writing new software
- DeepVariant is open-source and uses standard file formats (BAM/CRAM/VCF/gVCF)

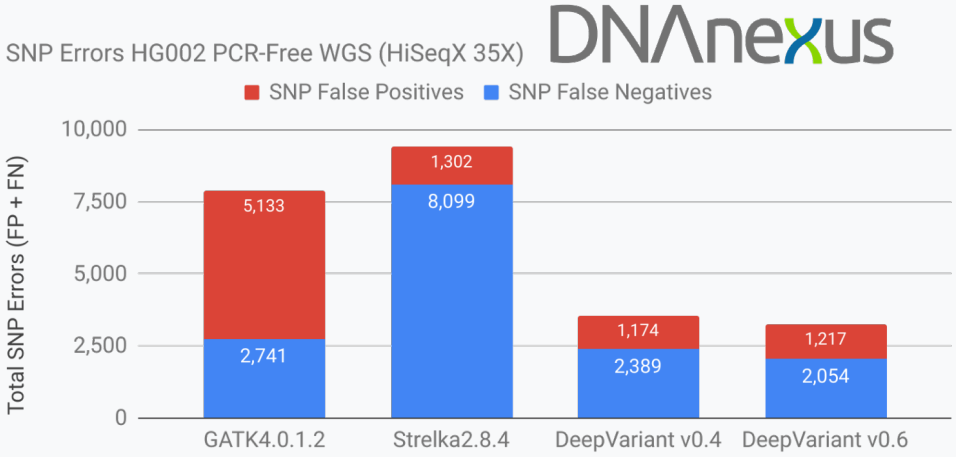
DeepVariant is Accurate

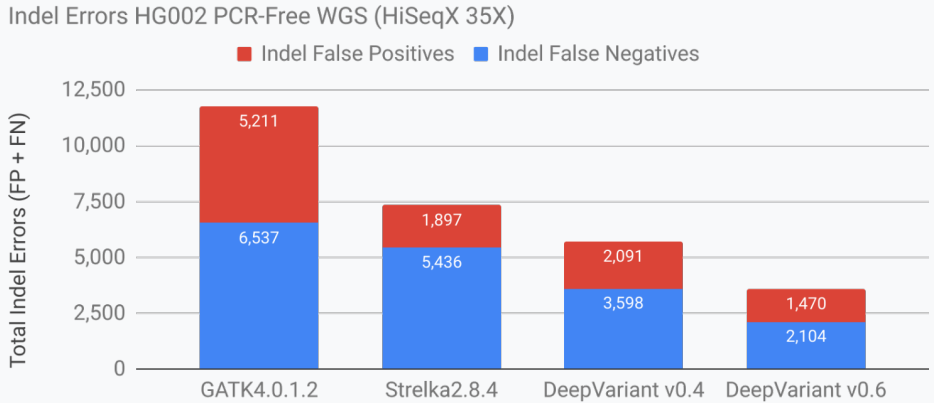
Google AI

HIGHEST
SNP Performance
in the precisionFDA Truth Challenge

AWARDED TO
Verily Life Sciences

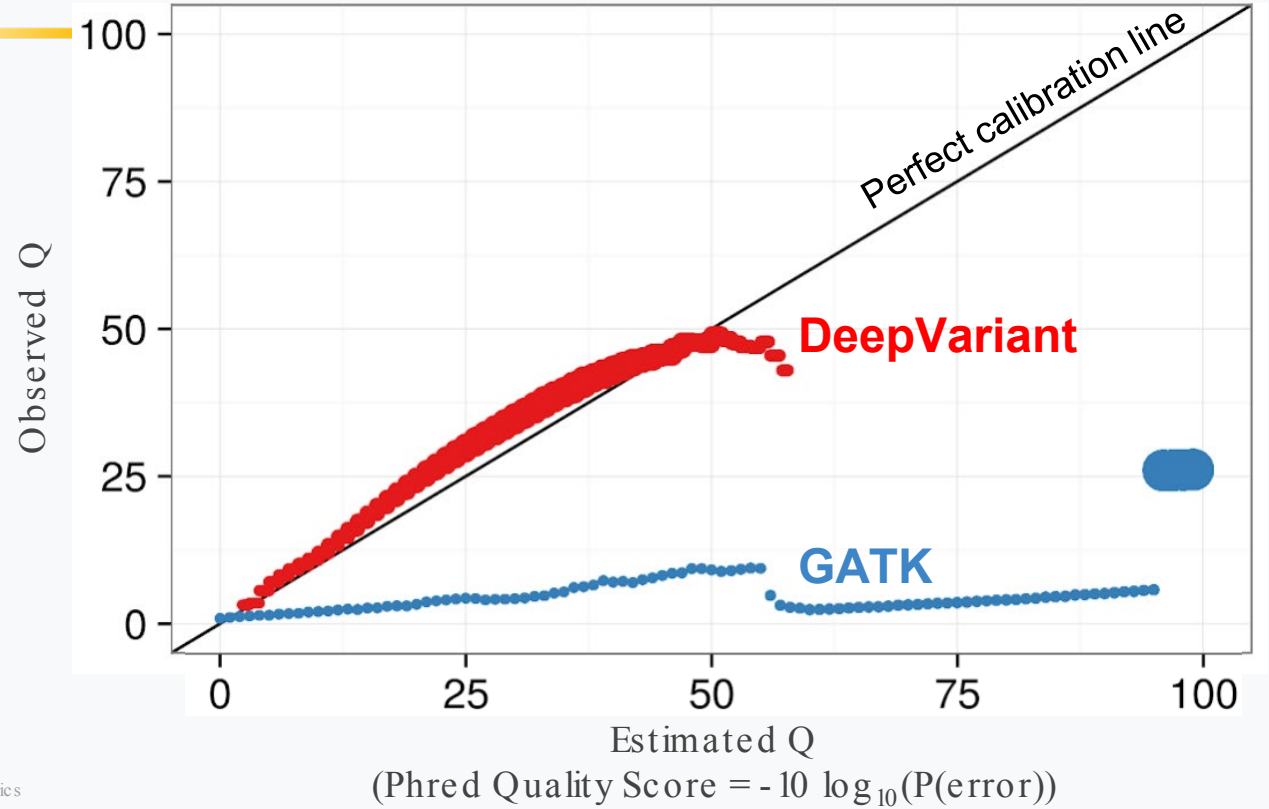
Ryan Poplin Mark DePristo
Verily Life Sciences Team





blog.dnanexus.com/2018 -04-18-deepvariant-amplified/

DeepVariant's Understanding of Errors is accurate



Heterozygous SNP calibration.

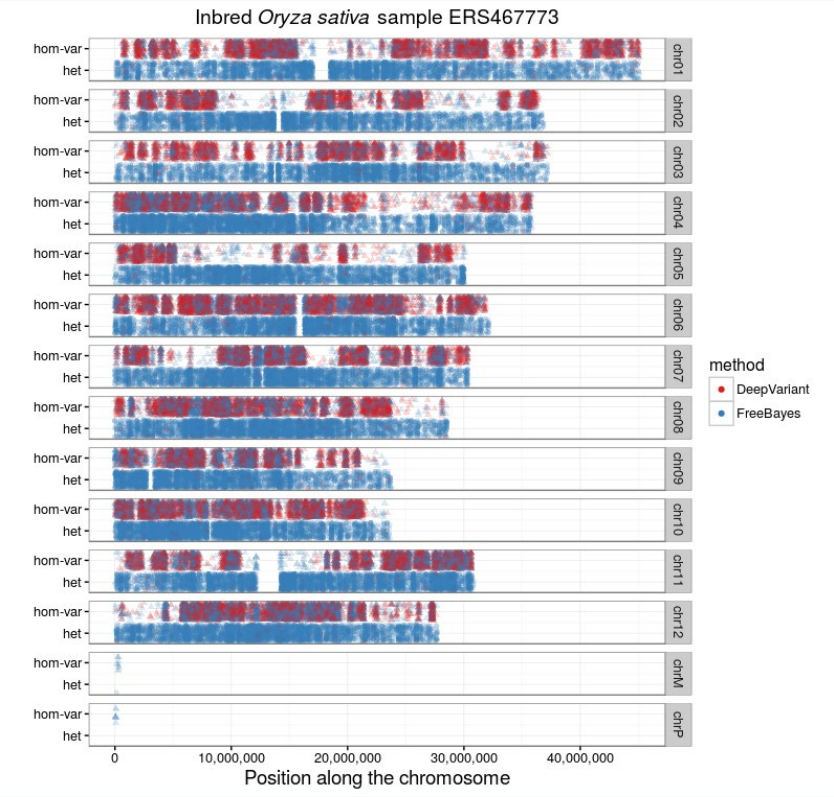
Genotype likelihoods are the critical input to genomic analyses such as imputation, de-novo mutation, and association.

Most callers are overconfident in their likelihoods.

DeepVariant is Roust

Google AI

DeepVariant calls in inbred *Oryza sativa* better fit expectation (even when trained from human data)

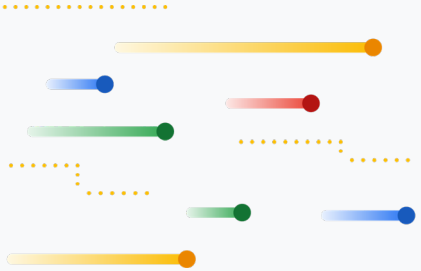


Done in collaboration with:

IRRI International Rice Research Institute

Estimated error rate from HET calls

DeepVariant	**4.5%**
GATK	27.8%
FreeBayes	38.4%



Story #1: Train, Don’t Program

Leveraging the inherent extensibility of deep learning approaches

Sequencing Comes in Many Flavors

The initial model of DeepVariant was trained on PCR-Free WGS.
But there are diverse way to sequence

PCR-Positive preps
Exomes
BGISEQ sequencers
Pacbio sequencers
FFPE-prepared samples
Saliva samples
Single-cell sequencing

Normally, if you want to perform well in each area, you need to write specialized code

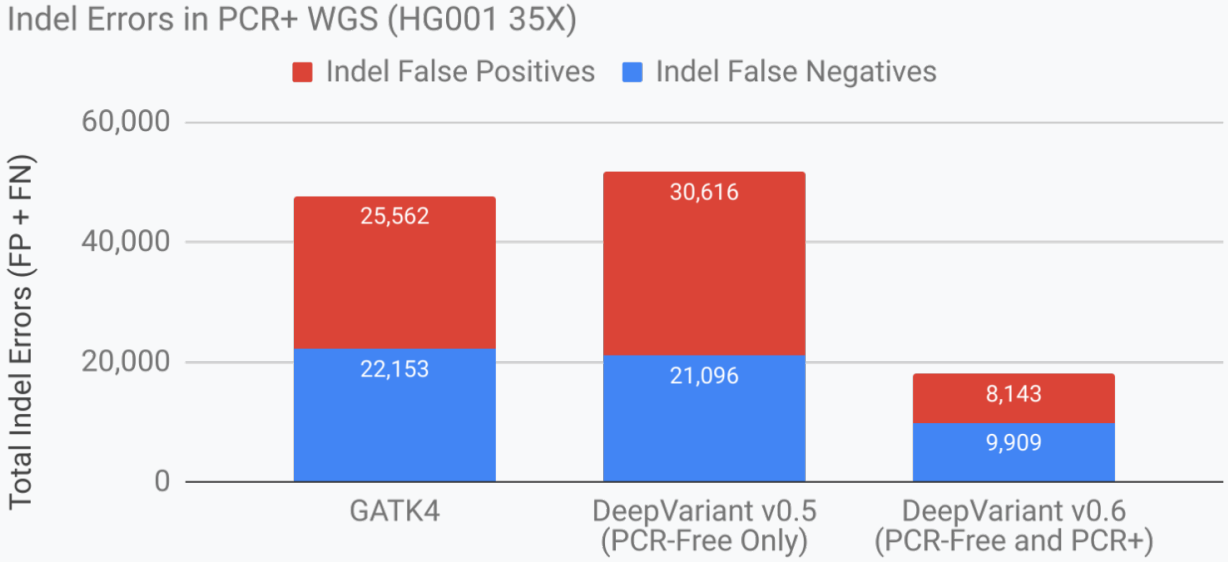
DeepVariant is Extensible

Improving on PCR+ Genomes

Google AI

DeepVariant can improve on a new data type without writing new code

(instead add training examples)

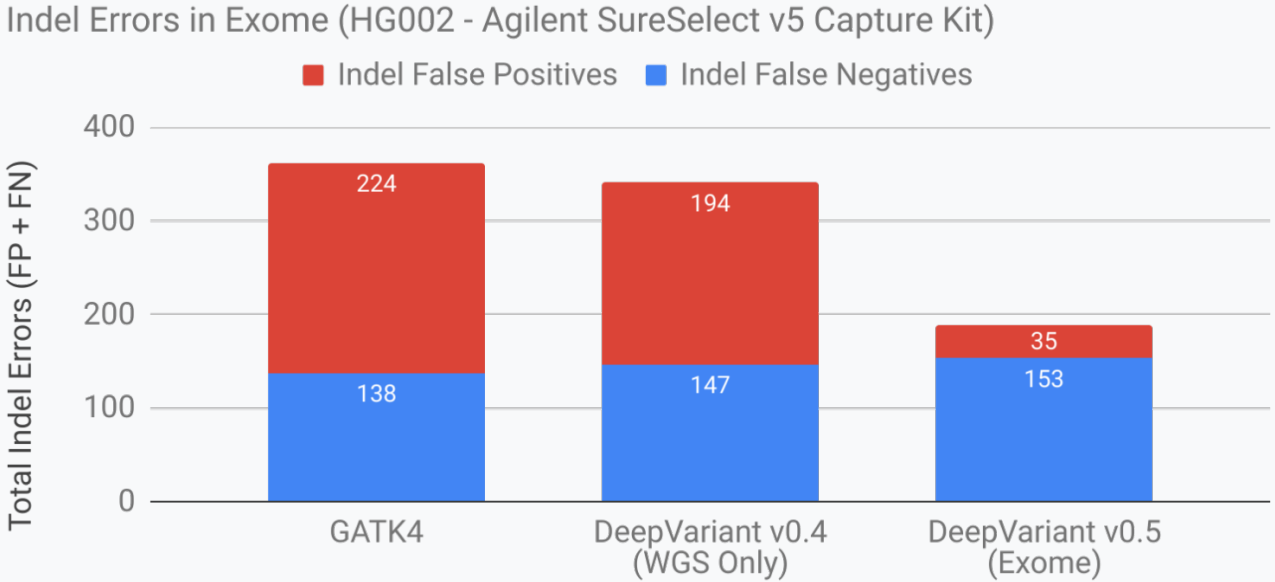


DeepVariant is Extensible
Improving on Exomes

Google AI

DeepVariant can improve on a new data type without writing new code

(instead add training examples)

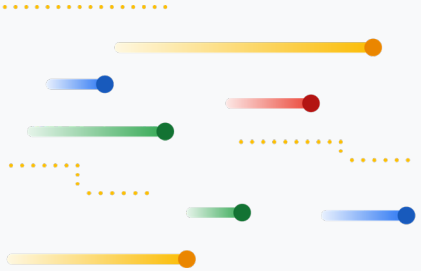


DeepVariant is Extensible

Improving on BGISEQ-500

Within one weekend of work, achieve the equivalent of years of progress

Method	Data Type	SNP F1	Indel F1
GATK4 Best Practices	BGI-SEQ	99.74%	87.49%
DeepVariant – ILMN trained	BGI-SEQ	99.83%	94.28%
DeepVariant – ILMN trained + BGI-SEQ fine-tuned	BGI-SEQ	99.89%	98.10%
DeepVariant Baseline	Illumina	99.96%	99.72%
GATK HC Baseline	Illumina	99.87%	98.75%



Story #2: The Importance of Quality

Reliable labels, Diverse examples

If You Leave with Only One Thing:

DON'T think that the doing machine learning right means you take only the cleanest, most pristine data in all ways.

Quality of data manifests in different ways:

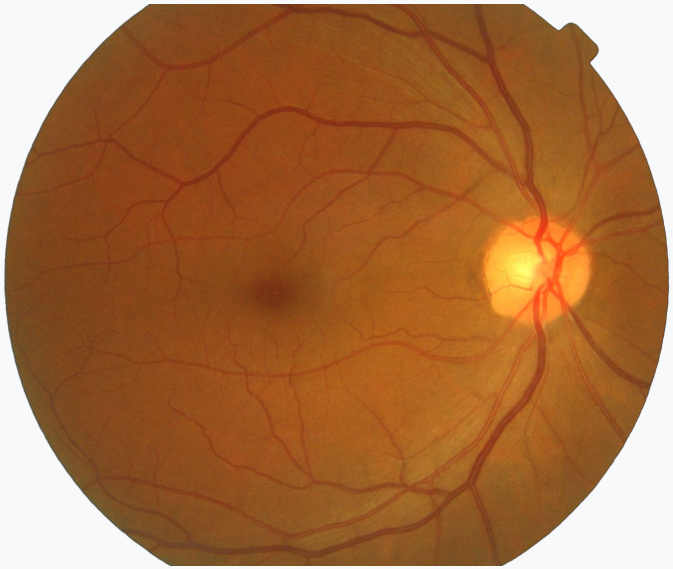
LABELS: Make these as **reliable** as possible. Your methods take these as truth. Don't lie to your model, or it will build explanations around the lies.

EXAMPLES: Make these as **representative** as possible. Don't make them clean. You want to capture what the model sees in production.

Hard is good. Noisy is good. Diverse is good
Make the model learn to de-noise. It will learn general principles.

Google AI

Reliable Labels - Diabetic Retinopathy



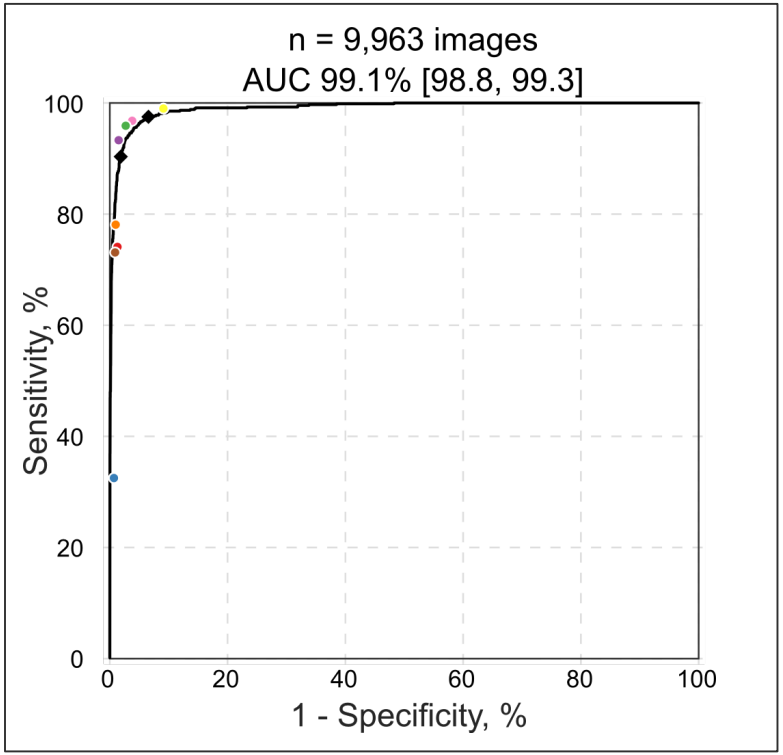
Healthy

Diseased

No DR	Mild DR	Moderate DR	Severe DR	Proliferative DR

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

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



F-score

0.95	0.91
Algorithm	Ophthalmologist (median)

"The study by Gulshan and colleagues **truly represents the brave new world in medicine.** "

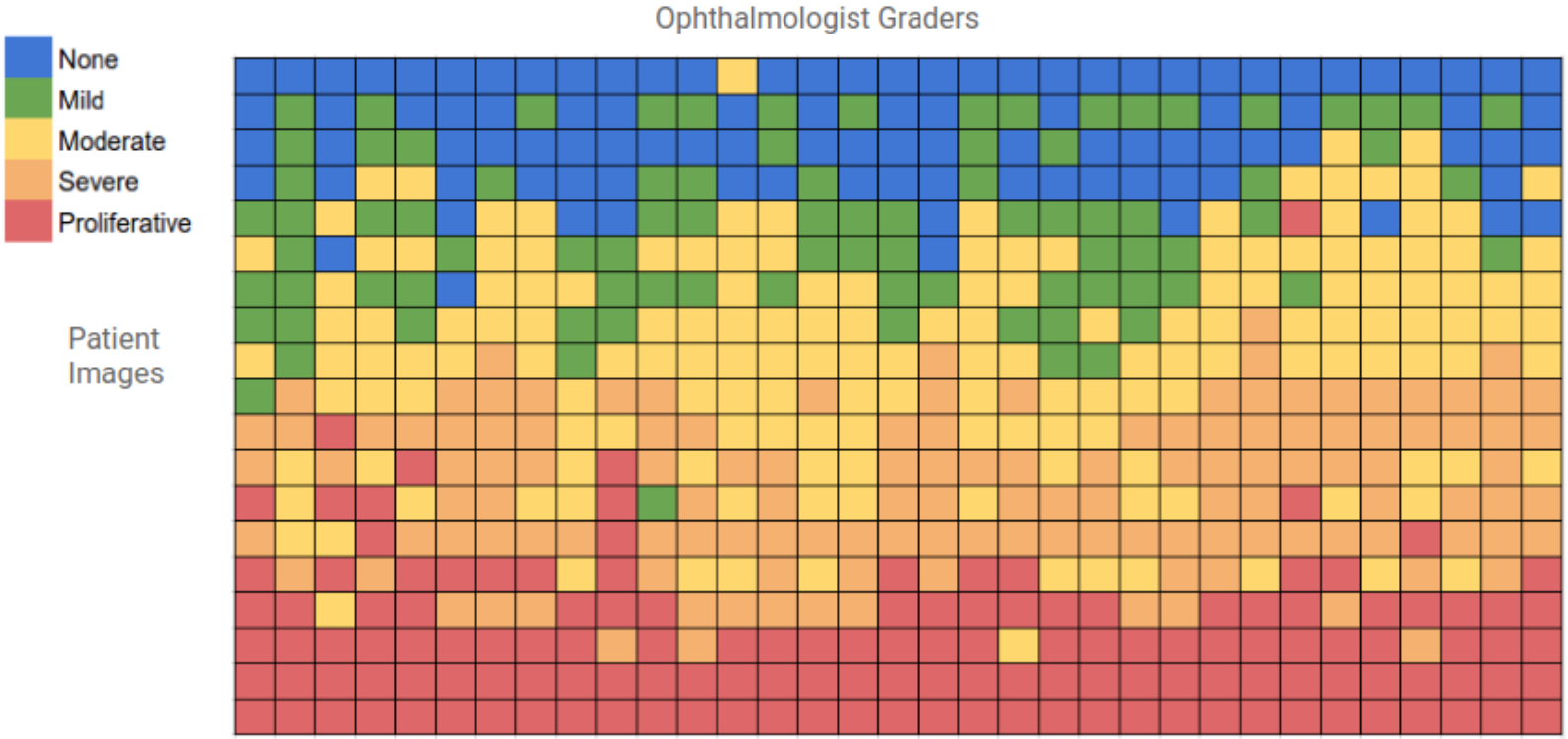
Dr. Andrew Beam, Dr. Isaac Kohane
Harvard Medical School

"Google just published this paper in JAMA (impact factor 37) [...] **It actually lives up to the hype** ."

Dr. Luke Oakden-Rayner
University of Adelaide

Google AI

The Importance of Labels



Consistency: intragrader ~65%, intergrader ~60%

Google AI

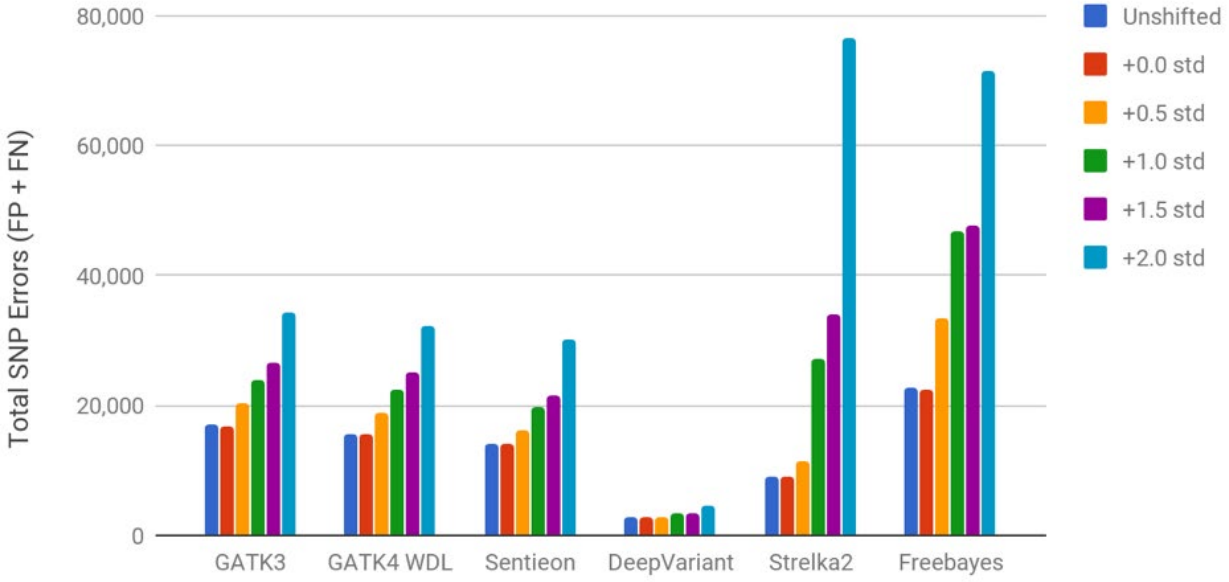
Used Properly, Noisy Data is an Asset

DNAnexus

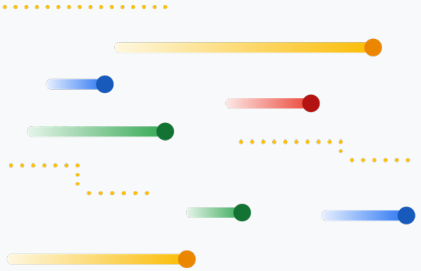
blog.dnanexus.com/2018-01-16-evaluating-the-performance-of-ngs-pipelines-on-noisy-wgs-data/

Figure 4. Conceptual Demonstration of Readshift 350X Sample to a 100X Shifted One

Figure 6. SNP Errors for Evaluated Tools on HiSeq2500 Data



Genomics



Google AI

Story #3: Build on Accelerating Technologies

Reliable labels, Diverse examples

Multiple Acceleration Options

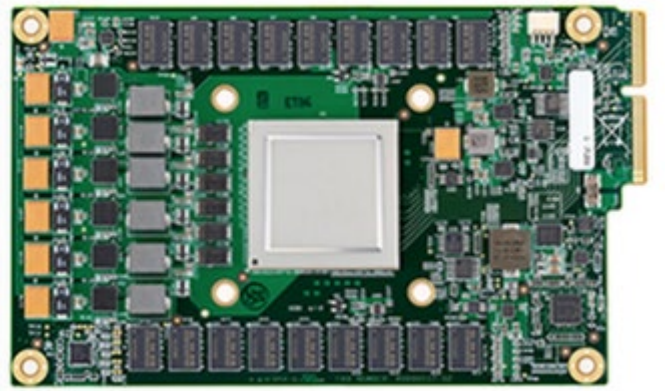
Google AI

TensorFlow

DeepVariant can run on CPU
(optional acceleration): GPU or TPU

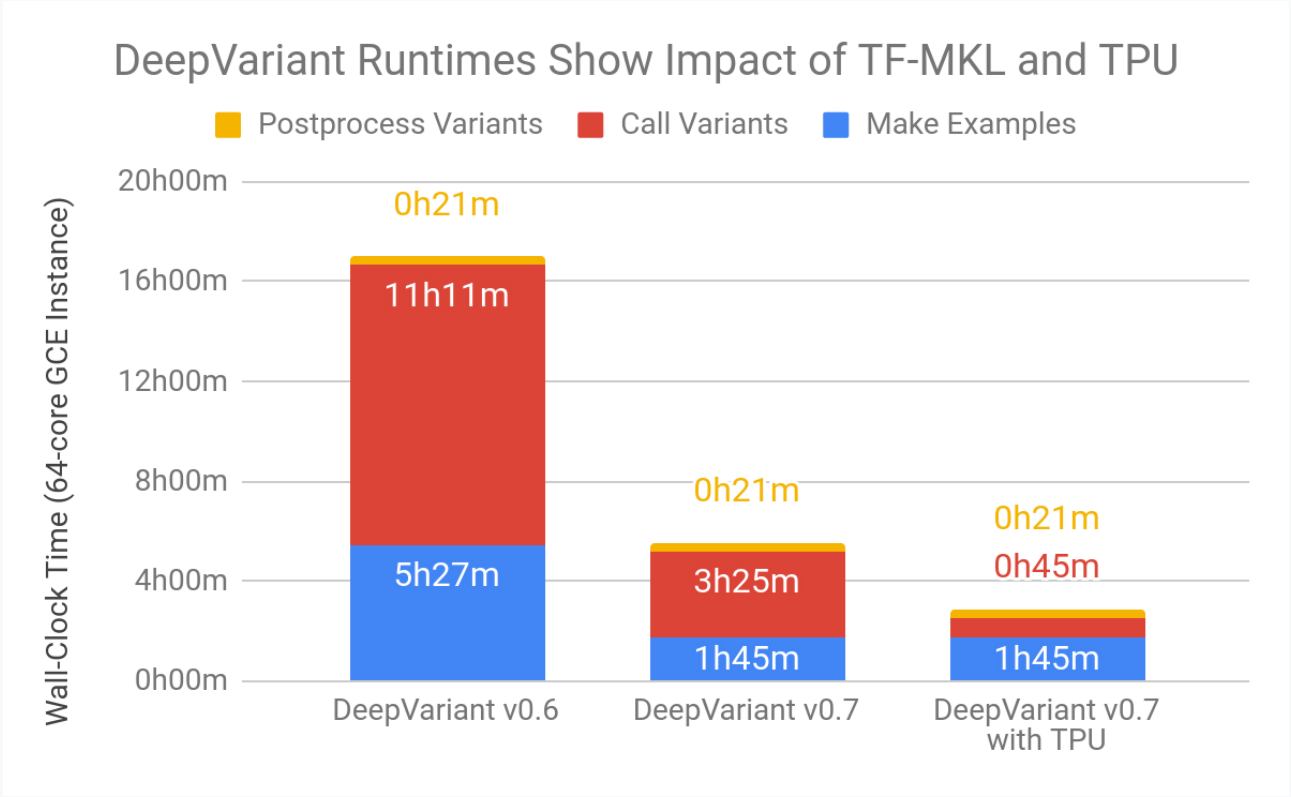
Whole Genome CPU - Single Machine	5-6 hours
Whole Genome Google Cloud	**70 min**
Exome Google Cloud	**25 min**

TPU on Google Cloud



Google AI

Naturally Inherit Technical Velocity

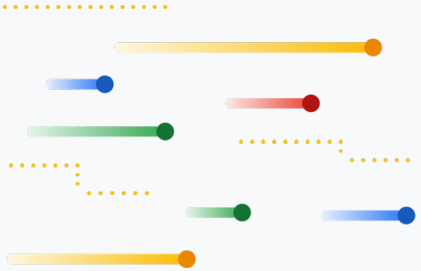


Nucleus: simplify applying ML to genomics data

Open source C++/Python library for reading+writing genomics data

- Supports most common data formats
  - Read+write: BED, VCF
  - Read: FASTA, FASTQ, BAM/SAM
- Common API across data types
- Built on:
  - Protocol buffers for language- and platform-neutrality
  - htslib provides efficient canonical parsing for high-throughput sequencing data formats
  - CLIF used to create C++ wrappers for Python
  - TensorFlow tfrecord files can be used anywhere genomics files are read or written
- Fully open source: Apache 2.0 license

https://github.com/google/nucleus

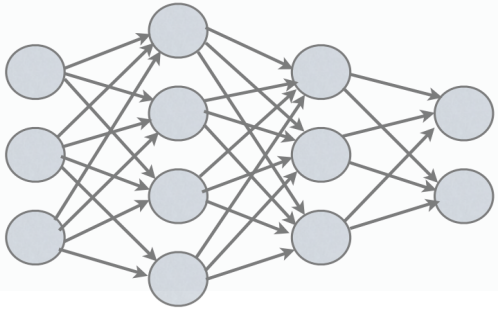


Google AI

Story #4: Understand Biology from the Models

A deep neural net builds a model of the world independent from human priors.
Use it to learn what you don't know you don't know.

Deep Neural Nets Create Their own Model of the Problem



- Smoking status
- Gender
- Age
- BMI
- SBP and DBP
- HBA1c
- MACE

nature biomedical engineering

ARTICLES

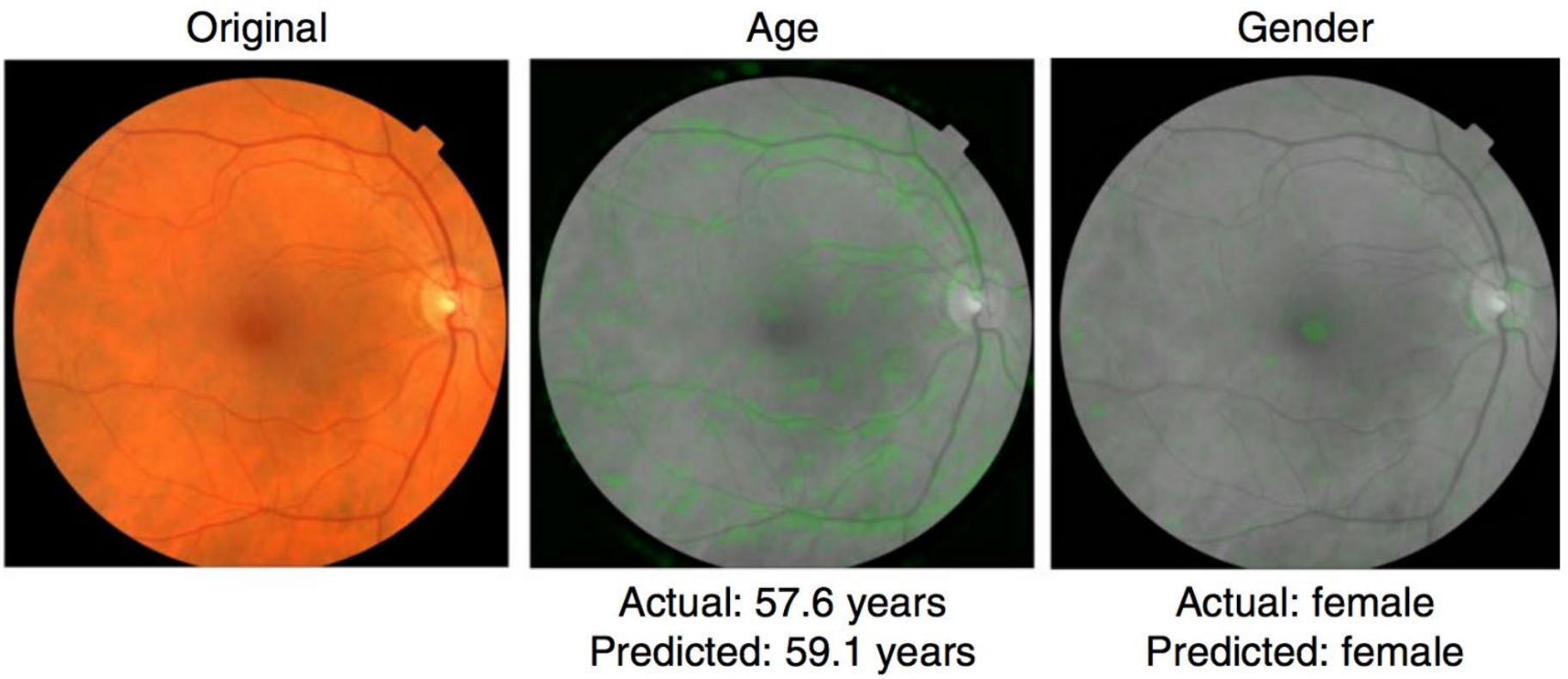
https://doi.org/10.1038/s41551-018-0195-0

Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Ryan Poplin[1,4], Avinash V. Varadarajan[1,4], Katy Blumer[1], Yun Liu[1], Michael V. McConnell[2,3], Greg S. Corrado[1], Lily Peng[1,4]* and Dale R. Webster[1,4]

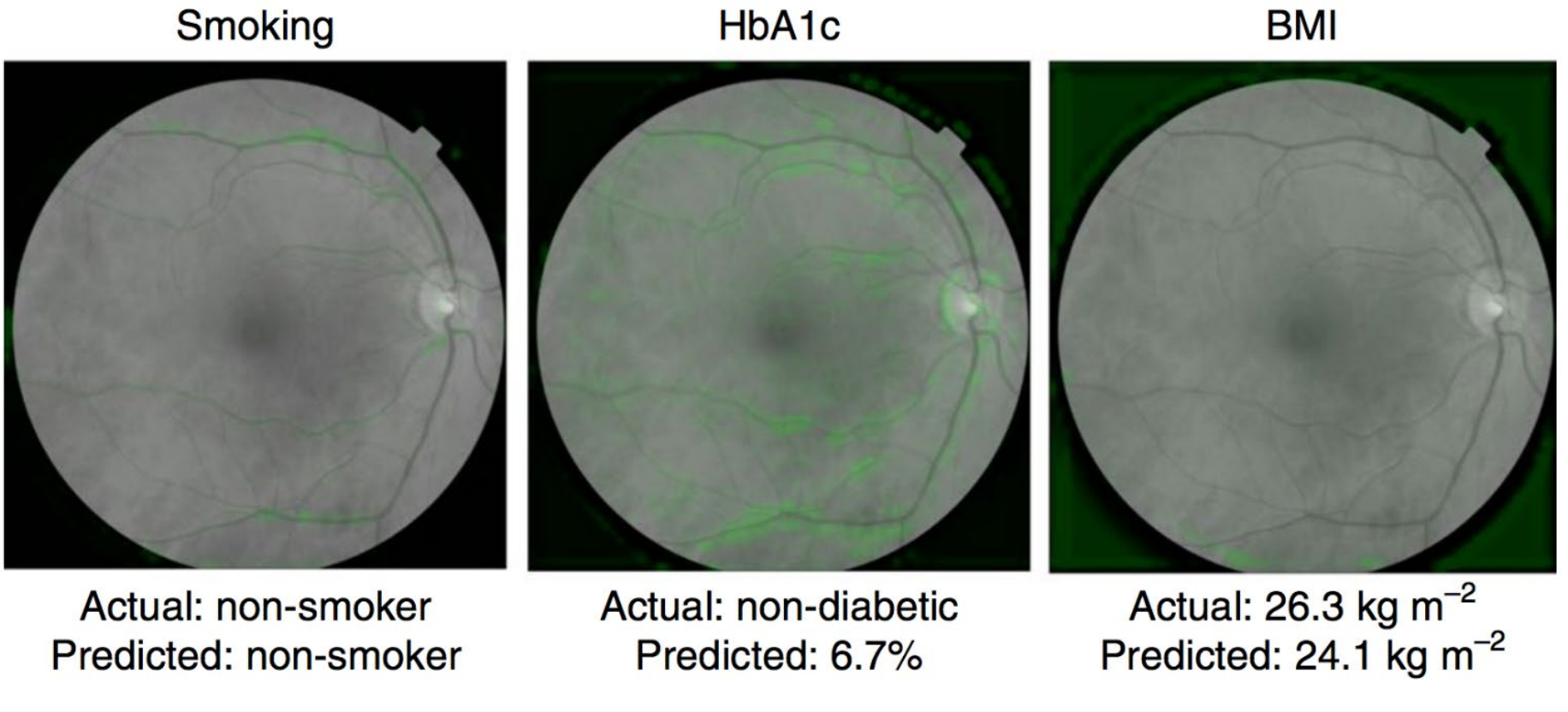
Google AI

Use Methods to Query that Model



Google AI

Use Methods to Query that Model

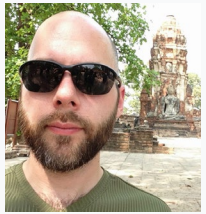


Thanks

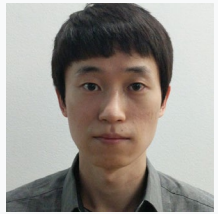
Google AI

Our team










The work presented here is from many other groups beyond our team as well