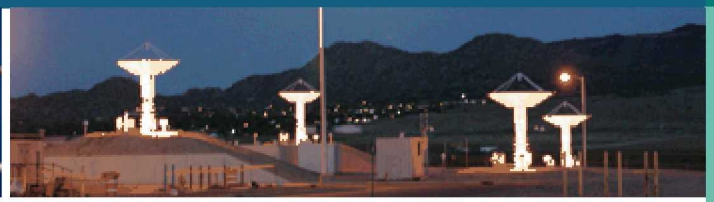
SAND2019-13063PE

Machine Learning (ML) Infrastructure and Applied Research for Nuclear Deterrence



PRESENTED BY

Matthew Smith – Applied Machine Intelligence

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline of today's talk

- Sandia has invested in the infrastructure required to bring AI to its missions
  - Lab-wide GPU and data resources
  - Machine Learning as a Service (MLaaS)
- ML research is being transitioned to production pipelines
  - As-built component qualification

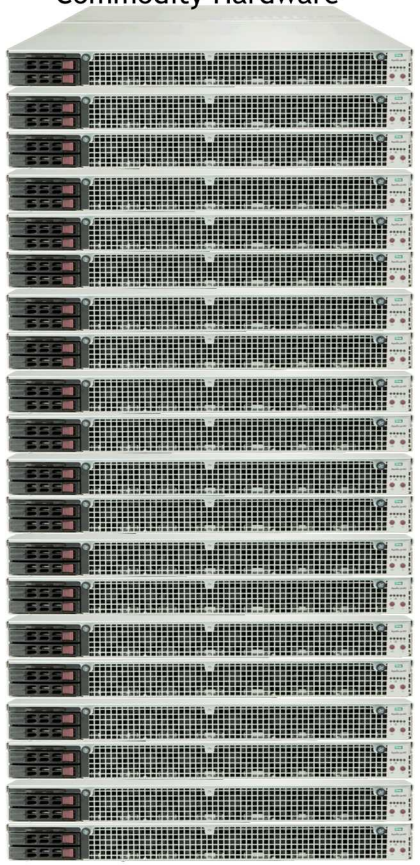
Outline of today's talk

- **Sandia has invested in the infrastructure required to bring AI to its missions**
  - **Lab-wide GPU and data resources**
  - Machine Learning as a Service (MLaaS)
- ML research is being transitioned to production pipelines
  - As-built component qualification

Machine Learning & Data Analytics clusters for lab-wide usage

CEE DA Hardware

Ray Cluster
Commodity Hardware



Ray (Data Analytics) Ambari Cluster

- General Purpose All Access
  60 -- Nodes
- ND Tenant
  SEDS/ODE
- Big Data for Actionable Intelligence (BDAI)
- Data Sciences (POC)

Synapse (DL) Kubernetes Cluster

- **1 - Nvidia DGX-1 (NVLINK)**
  8x16 P100, 2x28 core 750GB
- **6 - Nvidia DGX-1 (NVLINK)**
  8x16 v100, 2x28 core 750GB
- **8 - Nvidia DGX-2 (NVLINK)**
  16x32 V100, 2x28 core 1.5TB
- **20 - HPE PC40 (Inference)**
  1x24G PC40, 2x28 core 256GB

Shared Storage

Near-line data for Analytics, Training, and Inference

- GPFS, Stateless Containers, Object S3, NFS, Block

CEE Deep Learning Hardware

655,360 Cuda Cores
81,920 Tensor Cores
12 TB GPU Memory

286,720 Cuda Cores
35840 Tensor Cores
896 GB GPU Memory

76,800 Cuda Cores
480 GB GPU Memory

All resources have the same orchestration stack to ensure scalability and consistency

Why Kubernetes?

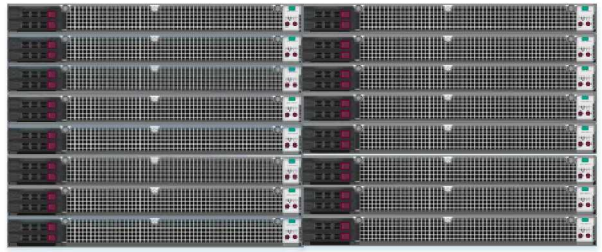
CEE HPDA Synapse

Kubernetes ➔ stateless containers with stateful storage

Kubernetes Pods with Resource controls and protection
- **single or multiple GPUs assignable and isolated from other tenants**
- **CPU, Memory, and Shared memory**
- **home and project space for single or group usage**
- **Flexibility through Python virtualenv**
- **Resources accessed via ssh with Kerberos**
- **Automated scheduling of Interactive Envs**

25 Gb Ethernet

1.4 PB of Storage (NFS & GPFS)



20 - HPE Apollo PC40s
128 GB RAM
2 x 3.2GHz x 8 Cores
Tesla P40 @ 24 GB
Inference/Development

Customer Access
25 Gb SRN

Scheduler

7 DGX-1s
512 GB RAM
2 x 2.2GHz x 20 Cores
Tesla V100 @ 16 GB
Plus 1 Pascal P100 @ 16 GB
Training/Inference

8 DGX-2s
1.5 TB RAM
2 x 2.7GHz x 24 Cores
Tesla V100 @ 32 GB
Training/Inference

Private
25 Gb Network

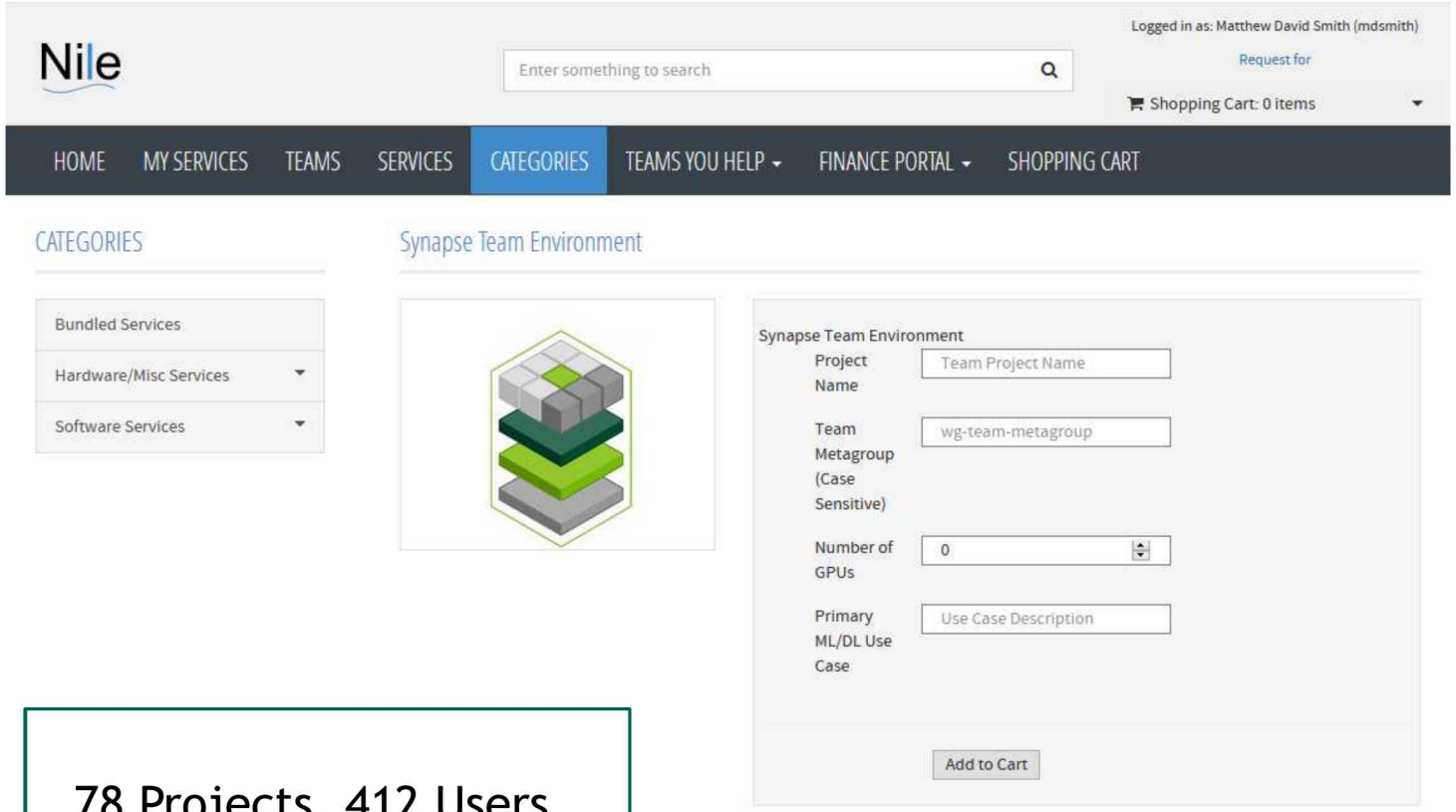
Stateful Storage
Shared with Analytics Clusters

GPFS 700 TB
GPFS Native

Pure 700 TB
NFS

COMMON ENGINEERING ENVIRONMENT

Getting access to these resources has been made simple



78 Projects, 412 Users across every mission division

ML & Data Analytics infrastructure resources will be available in multiple environments

- Data Analytics Cluster (Ray)
  - Planned SCN deployment in FY21
- Machine Learning/Deep Learning Cluster (Synapse)
  - Available on Upper Level LAN
  - Planned SCN deployment in FY20

Lessons learned lead to a roadmap for greater capabilities

- Improved multi-tenant for Ray (FY20)
  - Secure containerized environments
  - Rapid deployment and collapse
  - Over-subscription of resources
- Improved Synapse experience (FY20/FY21)
  - Improved X integration and tools
- Scheduling Improvements (FY20)

Outline of today's talk

- Sandia has invested in the infrastructure required to bring AI to its missions
  - Lab-wide GPU and data resources
  - **Machine Learning as a Service (MLaaS)**
- ML research is being transitioned to production pipelines
  - As-built component qualification

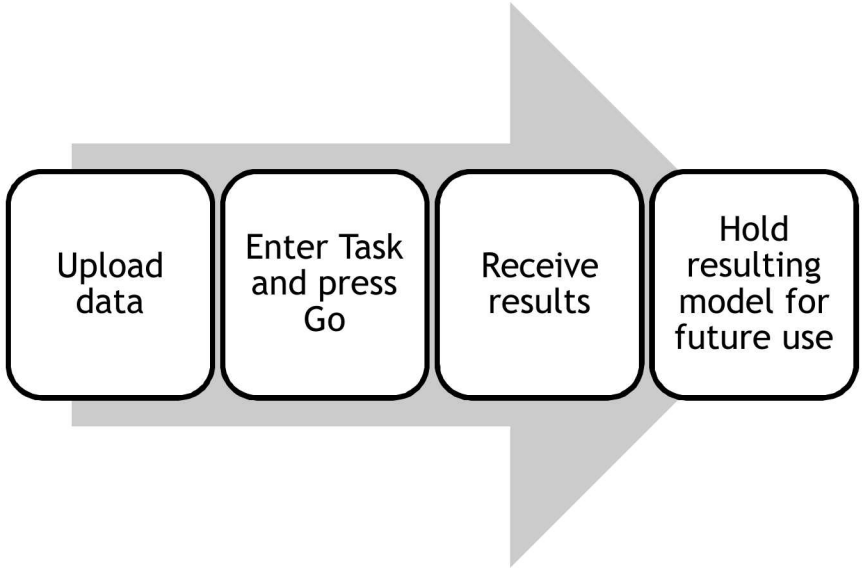
Machine Learning is widely applicable, should be widely usable

Goals

Machine Learning is broadly applicable to all of Sandia's missions

It should be **widely usable** as a general solution

- Imagine if you could:



Machine Learning as a Service (MLaaS) is our answer

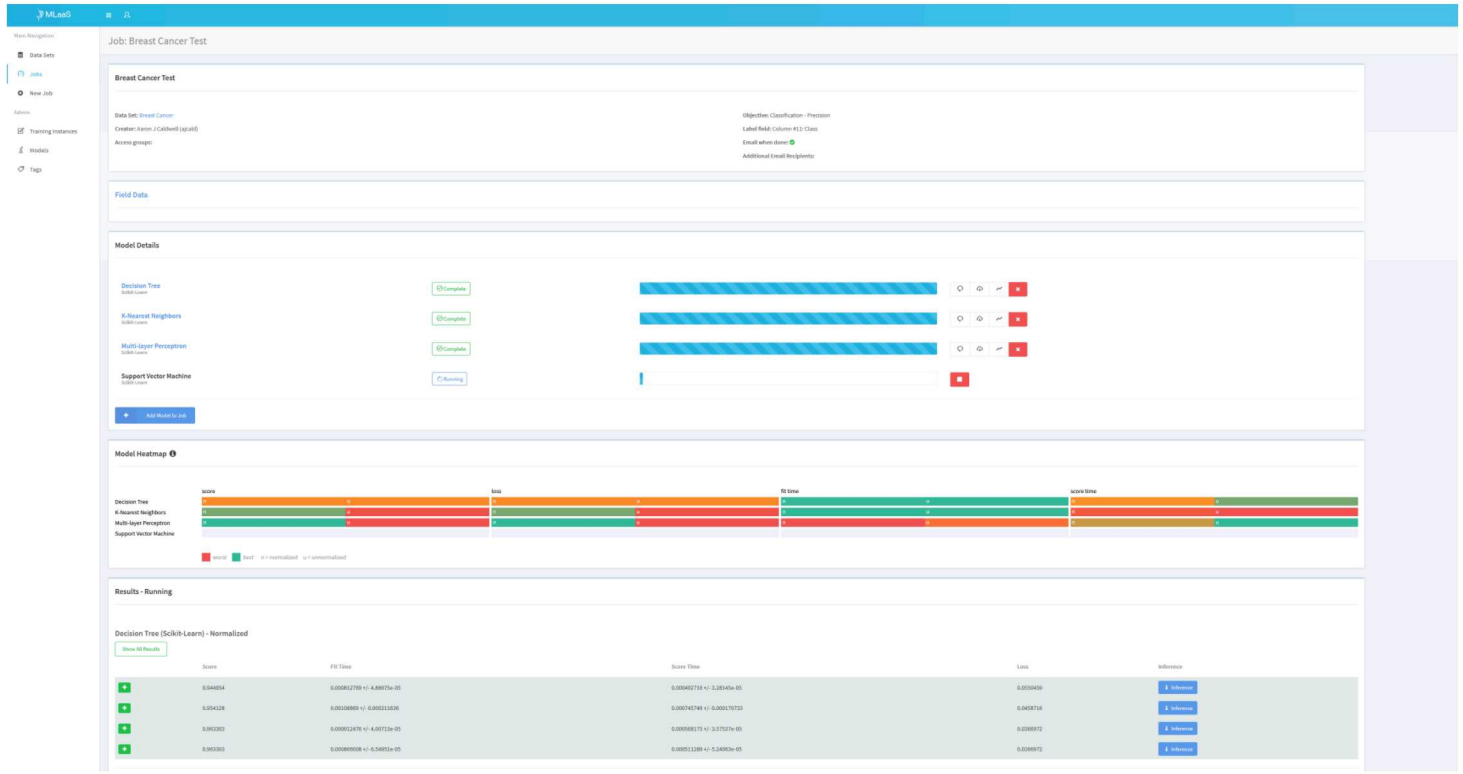
Upload your data

- CSV (row wise data) or images (structured grids)
  - Large image datasets are done via mount point
- Data is access controlled
- Cross domain not in scope

Can filter to specific models if desired

Get results

- API
- File download in web UI



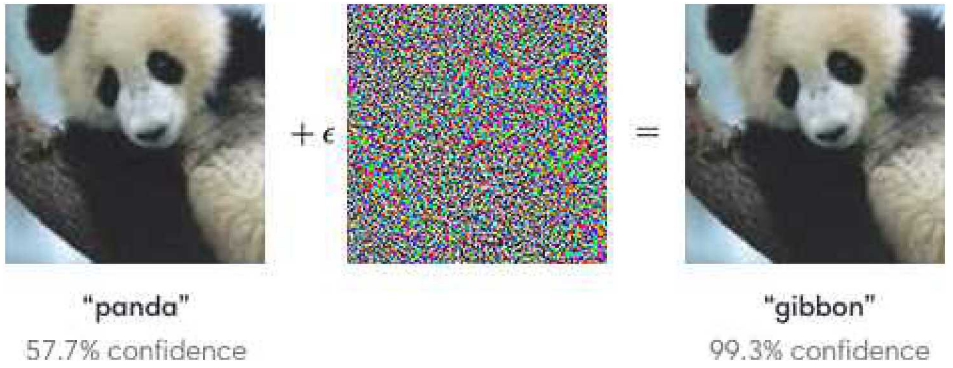
Machine Learning as a Service is a long term project

Multiple phases planned
- Additional models for both data input types
- Safe ingestion from github when allowed by license (MIT)
- Improved user experience

Research Focuses – What will distinguish us from the outside?
- Security testing
- Robustness testing

FY20 Goal – Expanding to high side networks
- SCN and/or ULL

Image credit goes to OpenAI



These infrastructural resources can cut across ND workforce

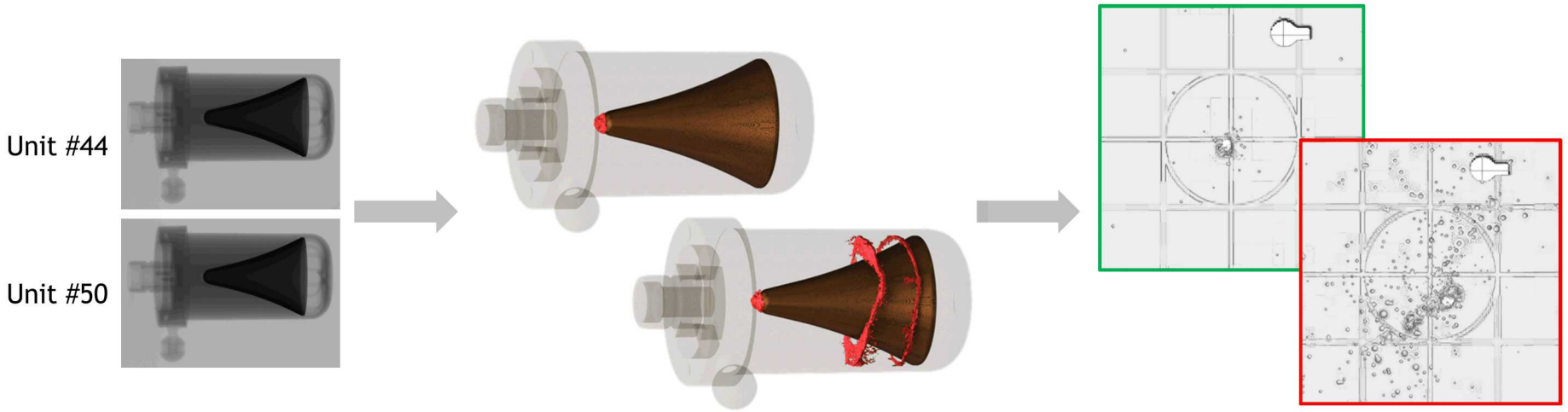
- Hardware described (Ray and Synapse) can be used by any ND MOW
  - Already used by members of every R&D division
  - Resources will continue to be expanded
  - Continuous user experience improvement
- Machine Learning as a Service will provide users with ML baselines for their problems
  - Separate problems in to two classes:
    - Solvable in an automated manner (no FTE required)
    - Focused applied research required for success
  - Requires minimal ML knowledge
  - Models are immediately usable

Outline of today's talk

- Sandia has invested in the infrastructure required to bring AI to its missions
  - Lab-wide GPU and data resources
  - Machine Learning as a Service (MLaaS)
- **ML research is being transitioned to production pipelines**
  - **As-built component qualification**

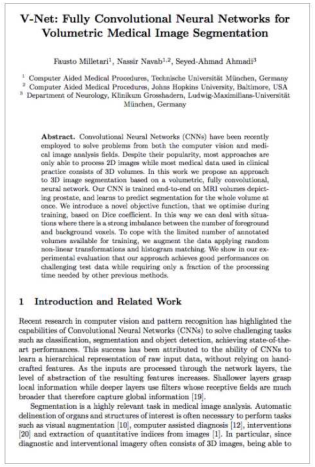
Advancements in CT scanning technology have enabled a new way of analyzing real world objects

- Perform numerical simulations directly from volumetric scans
- Individual acceptance replaces statistical lot acceptance
  - Rapid identification of faulty components
  - Numerical assessment of as-built components
  - Aging of each serialized unit
- Advances come with big data challenges
  - 2 TB of data per week per scanner
  - Semantic segmentation of billions of voxels is required for simulation



Adapted open source medical CT segmentation algorithms for supervised semantic segmentation

- Extended V-Net Capability
  - Improved on scan size limitations
    - Old limitation → 64x64x128
    - New capability → 2000x2000x2000 (scanner resolution)
- Overcame Label Imbalance
  - Weighted negative log likelihood loss
  - Weighted random sampling
- Transitioning to production capability
  - Gross defects are found
    - Heuristics have been created to determine off angle liners
- Research continues on:
  - CT reconstruction improvements
  - Fine defects and anomalies



V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation

Fausto Milletari[1], Nassir Navab[1,2], Seyed-Ahmad Ahmadi[3]

[1] Computer Aided Medical Procedures, Technische Universität München, Germany
[2] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA
[3] Department of Neurology, Klinikum Grosshadern, Ludwig-Maximilians-Universität München, Germany

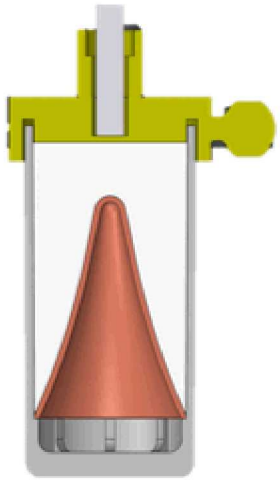
Abstract. Convolutional Neural Networks (CNNs) have been recently employed to solve problems from both the computer vision and medical image analysis fields. Despite their popularity, most approaches are only able to process 2D images while most medical data used in clinical practice consists of 3D volumes. In this work we propose an approach to 3D image segmentation based on a volumetric, fully convolutional, neural network. Our CNN is trained end-to-end on MRI volumes depicting prostate, and learns to predict segmentation for the whole volume at once. We introduce a novel objective function, that we optimise during training, based on Dice coefficient. In this way we can deal with situations where there is a strong imbalance between the number of foreground and background voxels. To cope with the limited number of annotated volumes available for training, we augment the data applying random non-linear transformations and histogram matching. We show in our experimental evaluation that our approach achieves good performances on challenging test data while requiring only a fraction of the processing time needed by other previous methods.

1 Introduction and Related Work

Recent research in computer vision and pattern recognition has highlighted the capabilities of Convolutional Neural Networks (CNNs) to solve challenging tasks such as classification, segmentation and object detection, achieving state-of-the-art performances. This success has been attributed to the ability of CNNs to learn a hierarchical representation of raw input data, without relying on hand-crafted features. As the inputs are processed through the network layers, the level of abstraction of the resulting features increases. Shallower layers grasp local information while deeper layers use filters whose receptive fields are much broader that therefore capture global information [19].

Segmentation is a highly relevant task in medical image analysis. Automatic delineation of organs and structures of interest is often necessary to perform tasks such as visual augmentation [10], computer assisted diagnosis [12], interventions [20] and extraction of quantitative indices from images [1]. In particular, since diagnostic and interventional imagery often consists of 3D images, being able to

Segmentation enabled by V-Net
Milletari et. al., 2016



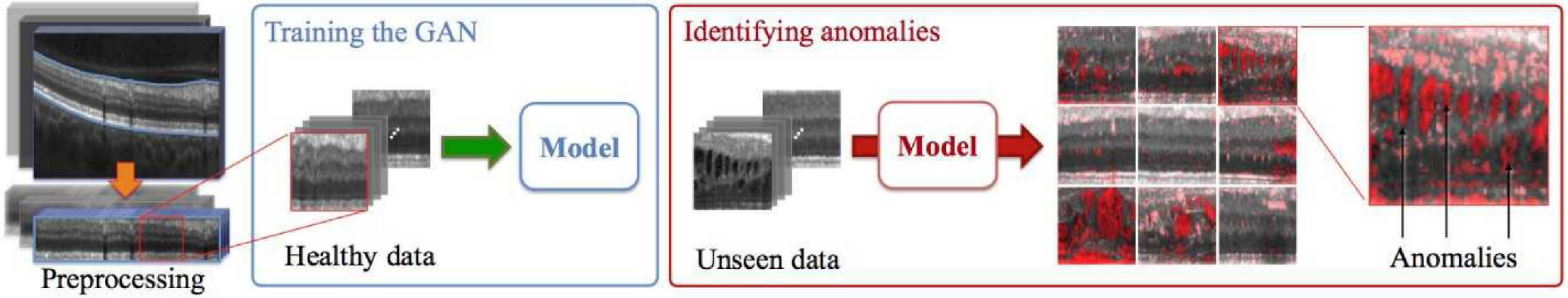
Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery

Schlegl et. al. 2017

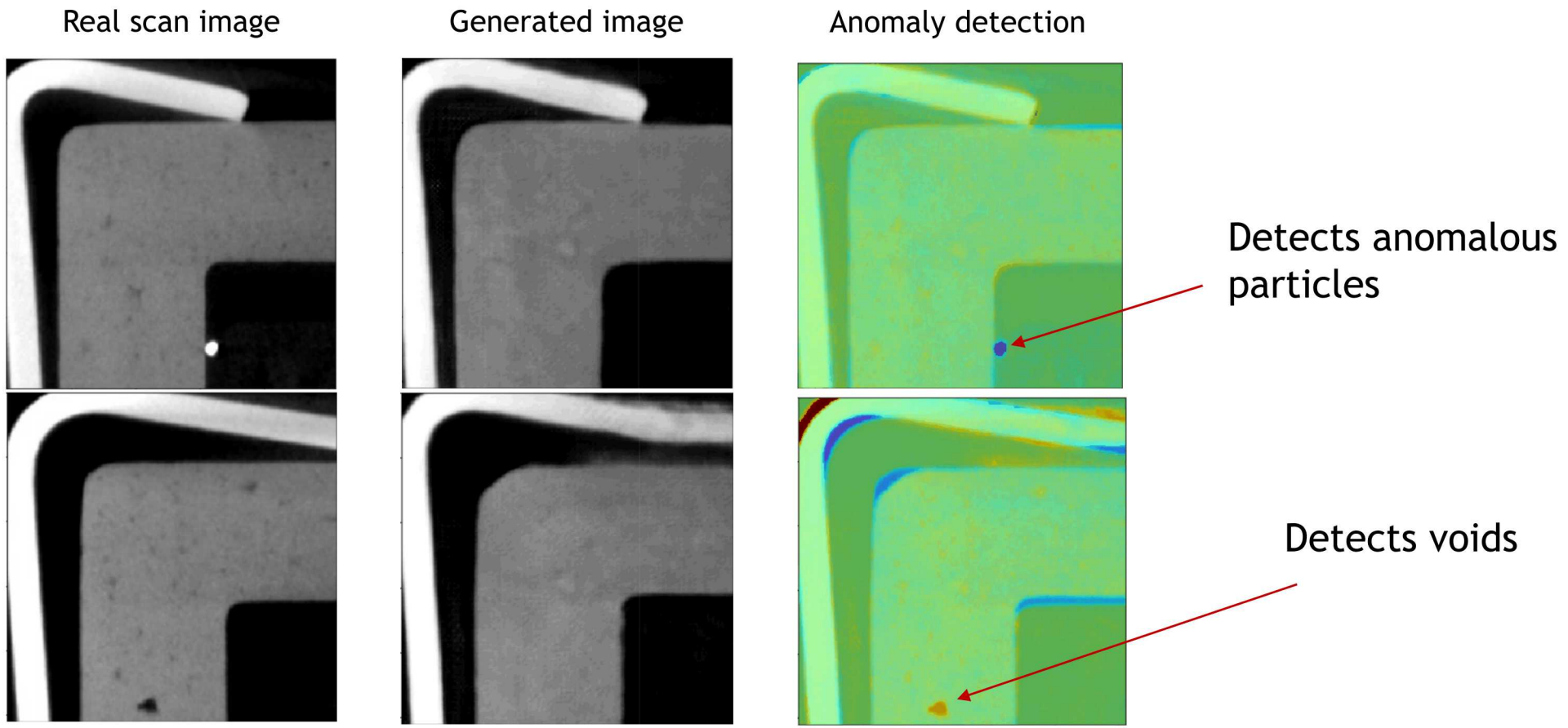
Use Generative Adversarial Networks (GANs) to generate normal retinal scans

Test on unhealthy scans, use regions of failure to highlight anomalies

"AnoGAN"



AnoGAN highlights anomalies in as-built components



AnoGAN succeeds in identifying rare features where supervised learning is not possible

This work will provide as-built component qualification for ND

- Performance prediction of as-built geometries
- Automated part qualification / ready for deterrence
- Discovery of aging/environmental effects on stockpile and other components

Questions