2021 embedded VISION summit
VIRTUAL | MAY 25-28

A Highly Data-Efficient Deep Learning Approach

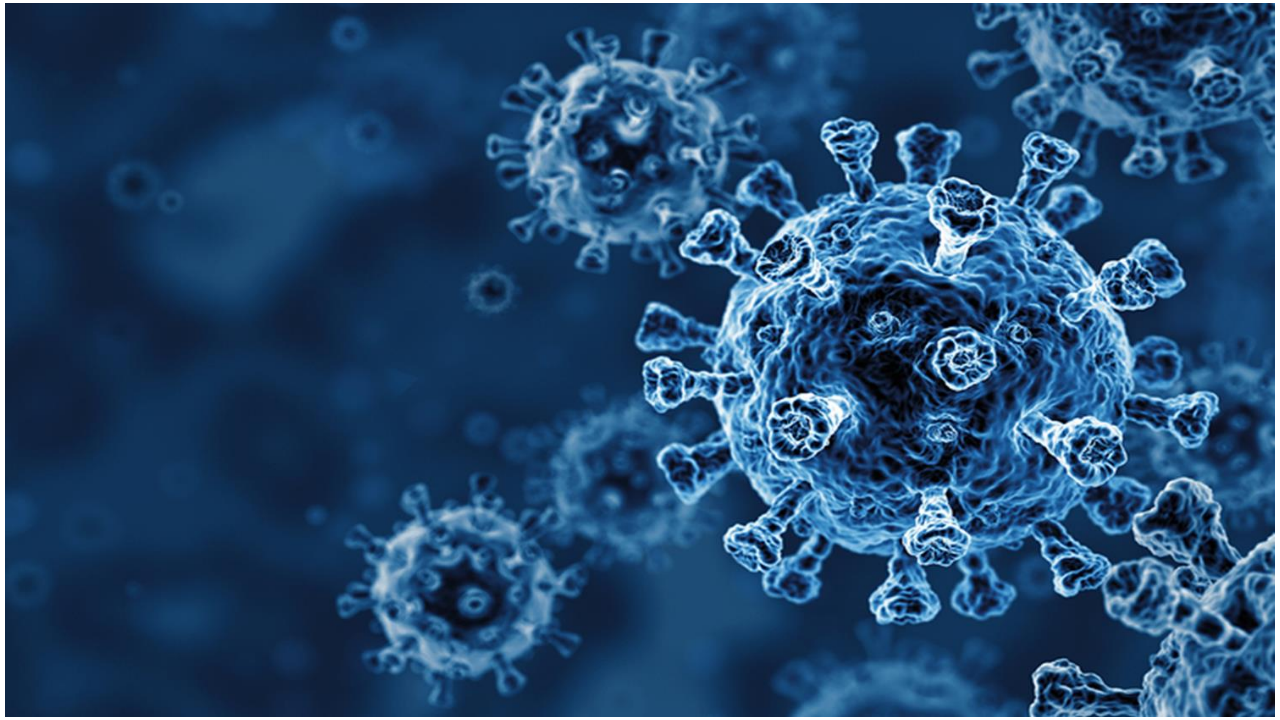
Patrick Bangert, VP of AI
Samsung SDSA

SAMSUNG SDS

COVID-19: When arriving at the hospital, a patient with symptoms wants to know quickly whether they have COVID-19 or not.

Diagnosis Options

- Diagnosing COVID-19 usually requires a nasal swab and a laboratory analysis that takes time
- Rapid test kits are ~90% accurate and often not available[1]
- Normal test kits are ~94% accurate but take 1-2 days in the laboratory[2]
- Lung X-rays are quick, easy, and cheap
- COVID-19 must be reliably distinguished from other respiratory diseases, like pneumonia.
- **Can this be done from X-rays using AI?**

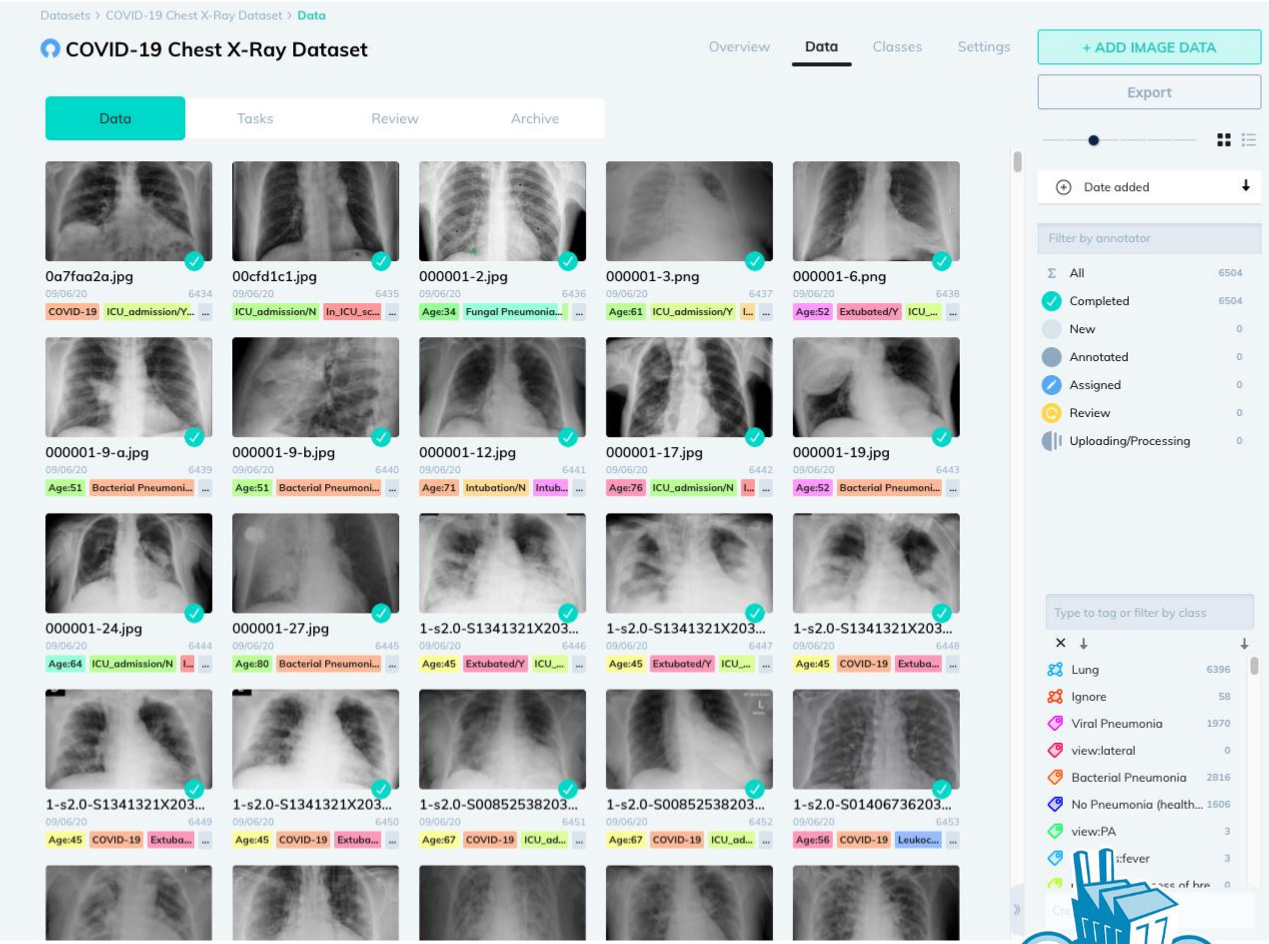


Dataset for Training AI: Machine learning must learn from a dataset. CloudFactory provides an open-source dataset of 15254 labeled x-ray images.

2021 embedded VISION summit

Big Data <> Small Data

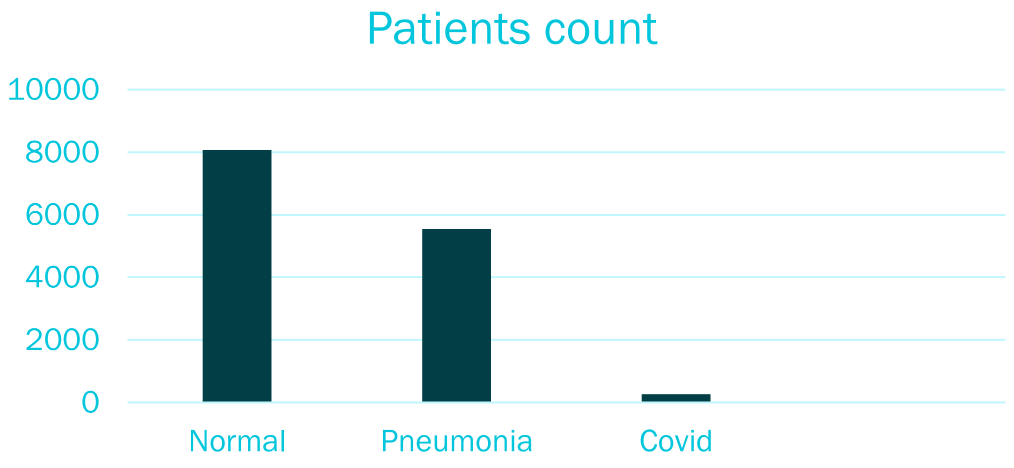
- Obtaining medical images is **difficult** due to privacy laws (HIPAA) and acquisition costs
- Labeling medical images and locating the disease is **time-consuming** and requires expert medical professionals
- AI requires as many labeled examples as possible to improve accuracy
- We desire to make a **good model** from a **small dataset**
- Active learning learns as the labeling happens and provides feedback when it is ok to stop

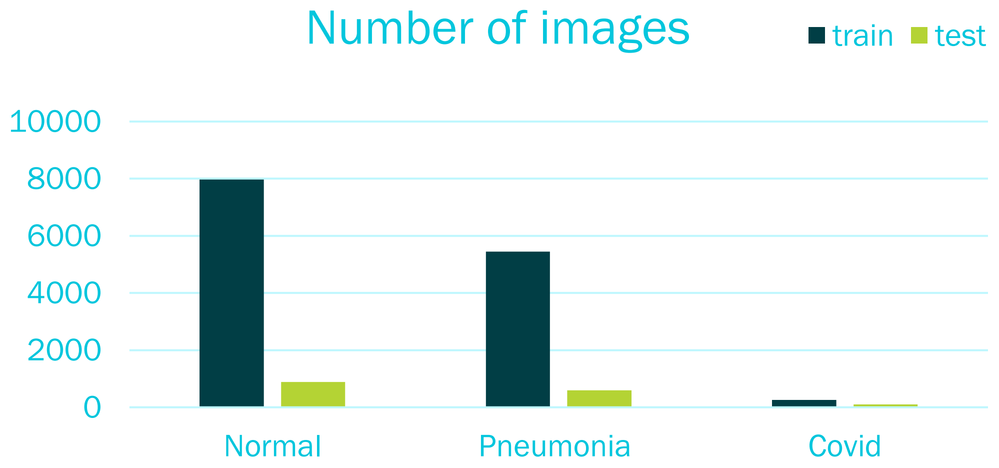


Pre-Processing the Data: The dataset is imbalanced and images are not registered.

Images must be made comparable

- The few COVID images must be **oversampled** so that the number of COVID cases is comparable to the number of pneumonia cases.
  - Normal and pneumonia cases were undersampled
  - COVID cases were oversampled randomly without modification (no added noise)
- Images must be **registered**
  - Analysis is relevant for lungs only
  - Other image parts must be removed

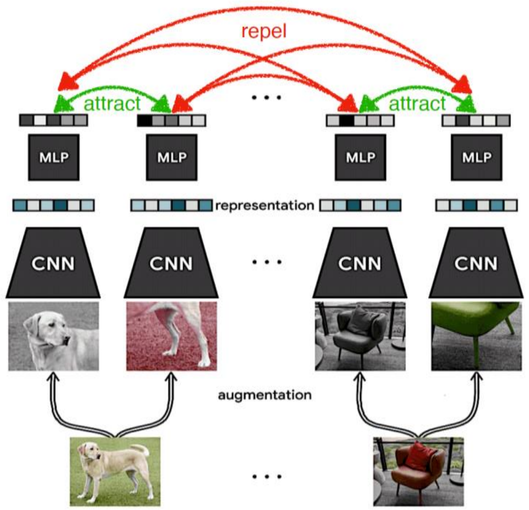


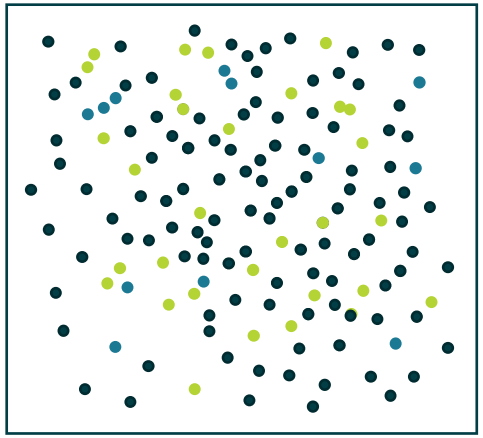


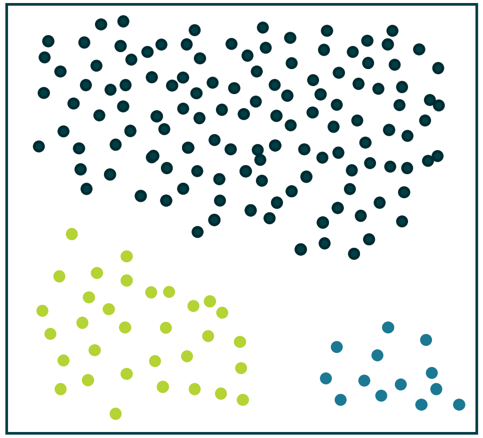
Feature Engineering: Image resolution is high and the number of images is low, so we must extract informative features to be able to classify them accurately.

Features are generated automatically

- Parts of images are recognized as “features” of the larger image
- Relation of “being a part of” becomes a **metric** in the space of images
- This process sorts and clusters a set of random images into groups, or **features**, in a high-dimensional space
- The technique is called SimCLR[1] and provides higher accuracy than state-of-the-art for computer vision classification
- Beyond SimCLR, a single fully connected NN layer converts the vector representation into a class label





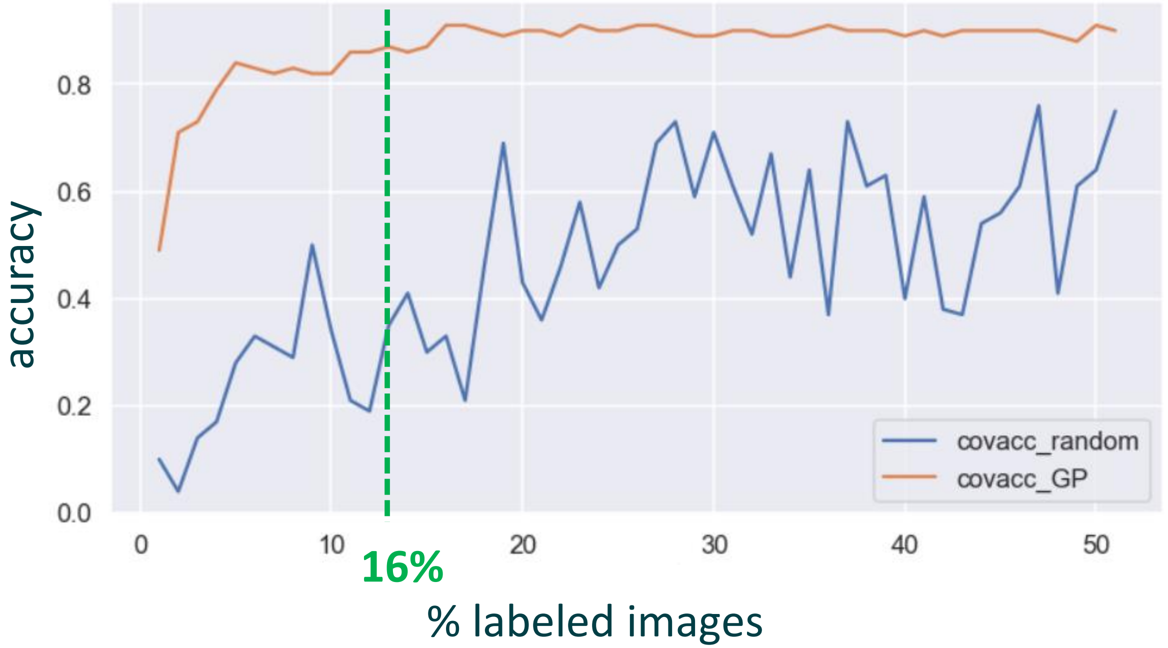


[1] https://arxiv.org/abs/2002.05709
https://papers.nips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html

Active Learning: Labeling medical images is difficult – we want to label the minimum number of images necessary.

Order matters in labeling!

- Some images add a lot of **information**, some add only little information
- Active Learning **sorts** the images in order of the probability estimate of the classifier
- Human experts then label only those images that add significant information
- During the labeling process, the model is continuously re-trained
- Accuracy rises as shown in red, as opposed to a random order as shown in blue
- After only 16% of images are labeled, the model achieves maximum accuracy (in this case)

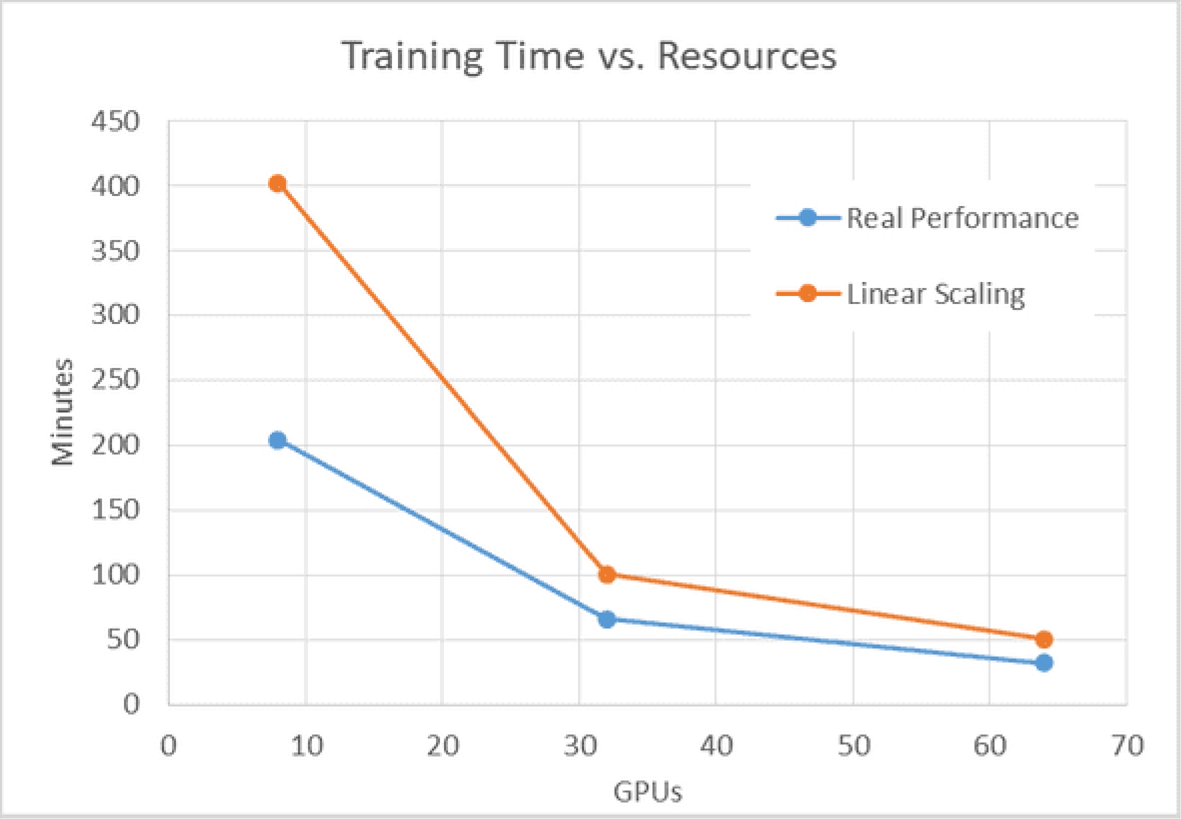


The Need for Speed – in AI Training: In active learning, the model must be updated quickly for the model to keep up with the human labeling process.

2021 embedded VISION summit

Distributed Training

- Using multiple GPUs can reduce the computation time for AI training linearly
- Organizationally, teams might label each morning and afternoon, doing a retraining during lunch and dinner.
- **Linear scaling** has been proved in computer vision and natural-language processing tasks
- Allows active learning to take place in **near real-time** speeds
- Using 8 GPUs, active learning can be run over a lunch break,
- Using 64 GPUs, it can be run in 8 minutes.



High Accuracy: In detecting COVID-19, accuracy matters as this will determine medical treatment and the success of this treatment.

2021 embedded VISION summit

Accuracy increases with labels

- State of the art for this dataset is an accuracy of 91%
- Our method can match this accuracy with only **16% labeled data**
- Going to a fully labeled dataset brings our model to an accuracy of 95% (on test data) as compared to 91% with state of the art.
- This **accuracy is higher than the nasal swab test** that lies at 94%

Confusion Matrices reveal the Benefits

2021 embedded VISION summit

The confusion matrix over the test data shows what the model returns for each true category. Our model is less confused about COVID than the state-of-the-art model by 4%.

Predicted

Ground Truth	Normal	Pneumonia	COVID-19
Normal	95%	5%	0%
Pneumonia	5%	94%	1%
COVID-19	5%	4%	**91%**

Baseline

Predicted

Ground Truth	Normal	Pneumonia	COVID-19
Normal	95%	5%	0%
Pneumonia	7%	92%	1%
COVID-19	3%	2%	**95%**

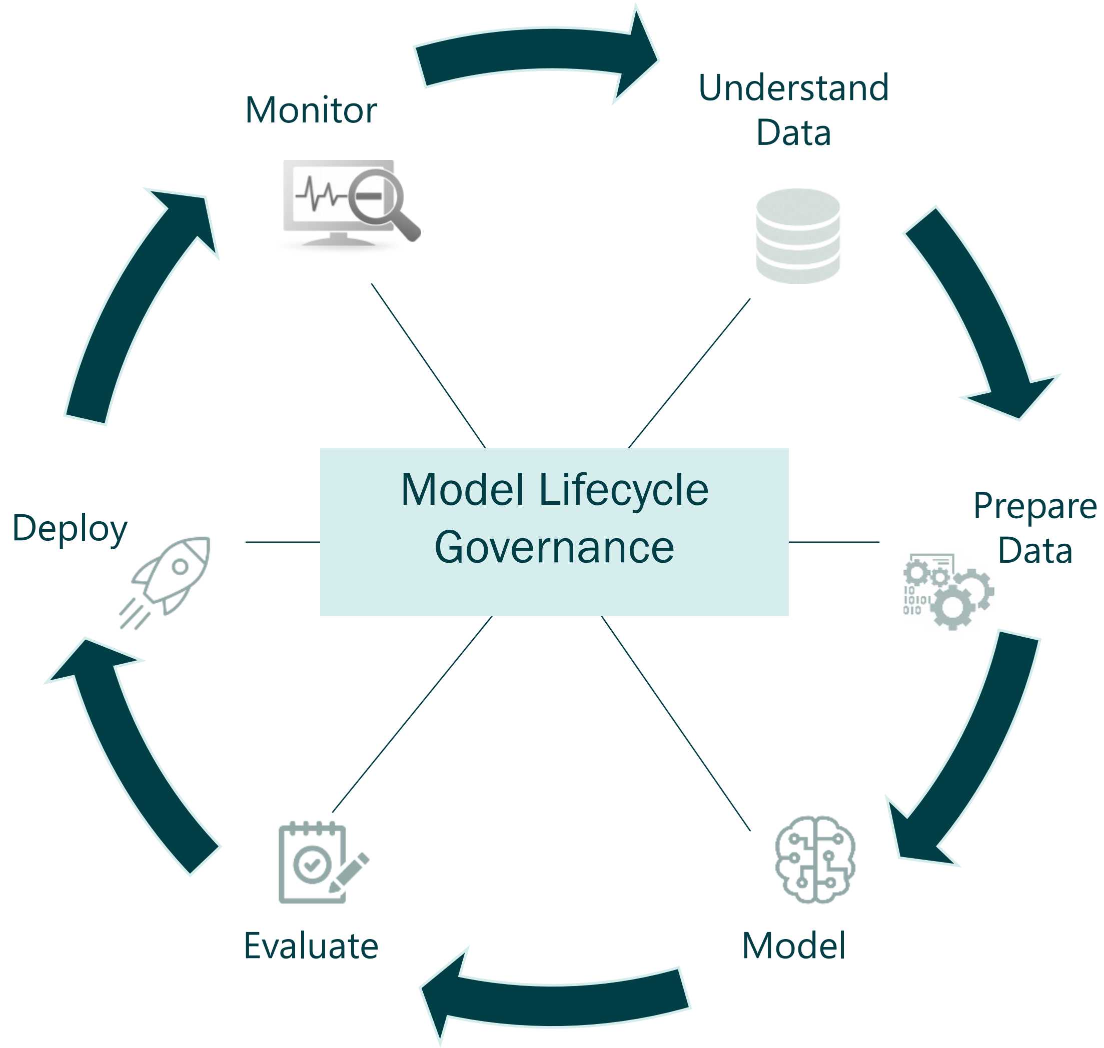
Our model

Conclusion: Combining pre-processing, feature engineering, distributed training, and active learning leads to the most accurate COVID-19 detection method to date.

2021 embedded VISION summit

Organic Problems need Holistic Solutions

- COVID-19 can be diagnosed accurately on the basis of a lung x-ray.
- State of the art accuracy can be achieved by labeling only 16% of the images – **saving 84% of the manual effort.**
- Using all labeled images (manual and automatic) increases the accuracy to 95%.
- This methodology relies on multiple techniques of pre-processing and automated feature engineering as well as distributed training to work in realistic time-scale
- Result: **COVID-19 diagnosis more accurate than a nasal swab!**



Resources

Further Information

- Scientific paper: https://arxiv.org/abs/2103.05109
- Two-part popular article
  - https://www.linkedin.com/pulse/artificial-intelligence-covid-19-screening-covid-using-bangert/
  - https://www.linkedin.com/pulse/teach-computer-vision-training-covid-scans-part-2-patrick-bangert/
- Demo video of the technology: https://youtu.be/wcP1fRPKXSU

Datasets used

- https://github.com/agchung/Actualmed-COVID-chestxray-dataset
- https://github.com/agchung/Figure1-COVID-chestxray-dataset
- https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
- https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
- https://arxiv.org/pdf/2003.11597.pdf

Thank you.

Patrick Bangert
p.bangert@samsung.com

SAMSUNG SDS