GTC 2019

S9164- Advanced Weather Information Recall with DGX-2

03/19/2019

Tomohiro Ishibashi - Director, Weather News, Inc.
Shigehisa Omatsu - CEO, dAIgnosis,Inc.

About us

Weathernews Inc.

Founded

June 11, 1986

Sales

$150 million

Number of Offices

34 offices
in 21 countries

Number of= Employees

826
as of May 31, 2017


Still people dying…
by heavy rain.
WMO(World Meteorological Organization) reports total disaster losses from weather and climate-related events in 2017 at US$ 320 billion
PHOTO : JIJI PRESS

Structure of Weather Industry

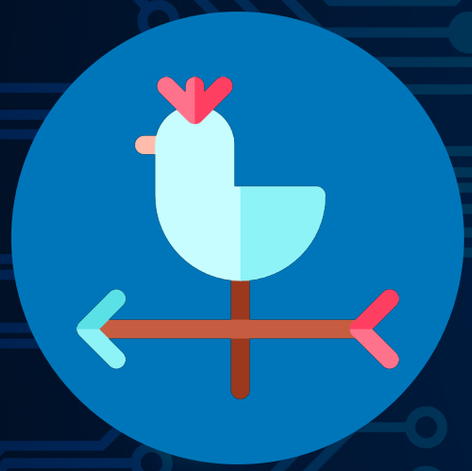
NWS
(National Weather Service)

News media
& Weather company

Audience & User

Existing Weather forecast model

Input

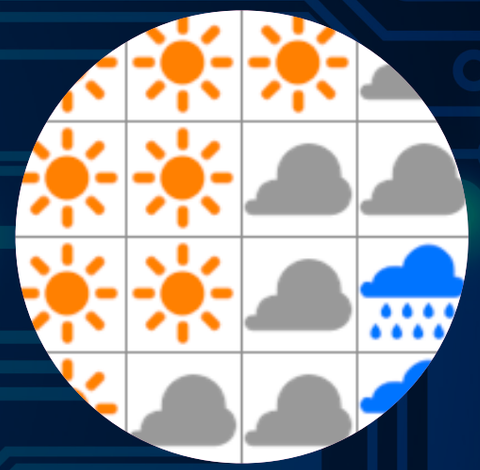


Official Observation data

Calculate

Physical Model

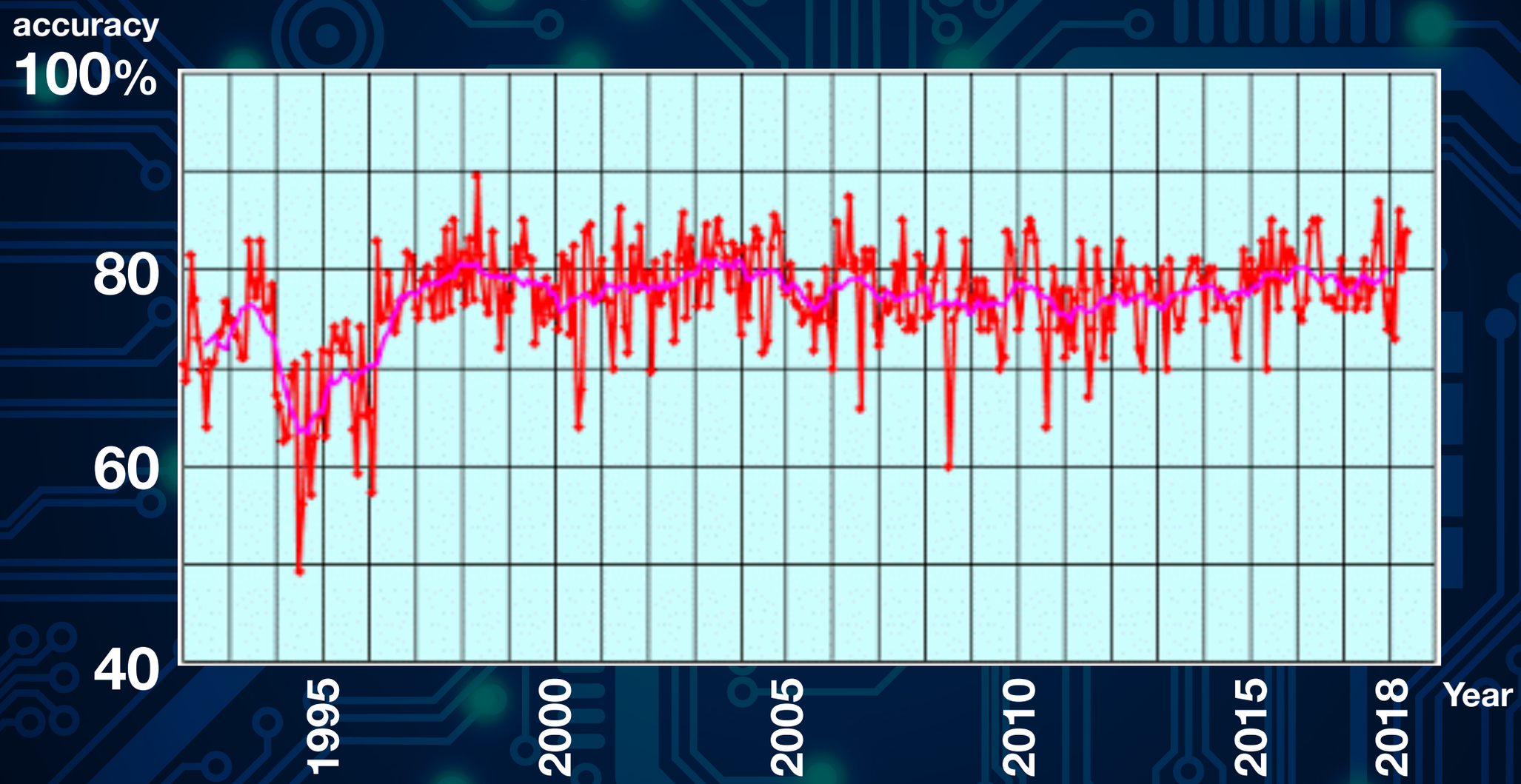
Output



Grid by grid forecast

Existing Weather forecast model

Weather forecast Accuracy (JMA)



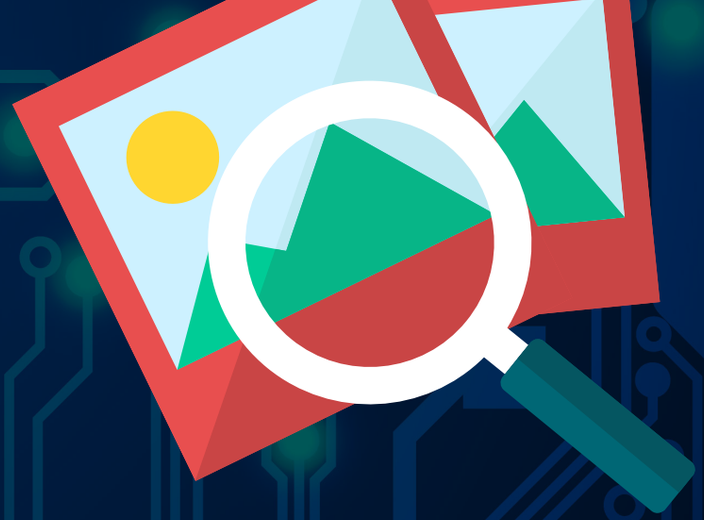
No big difference in the last 20 years

Weather forecast should change

CPU

Physical Model

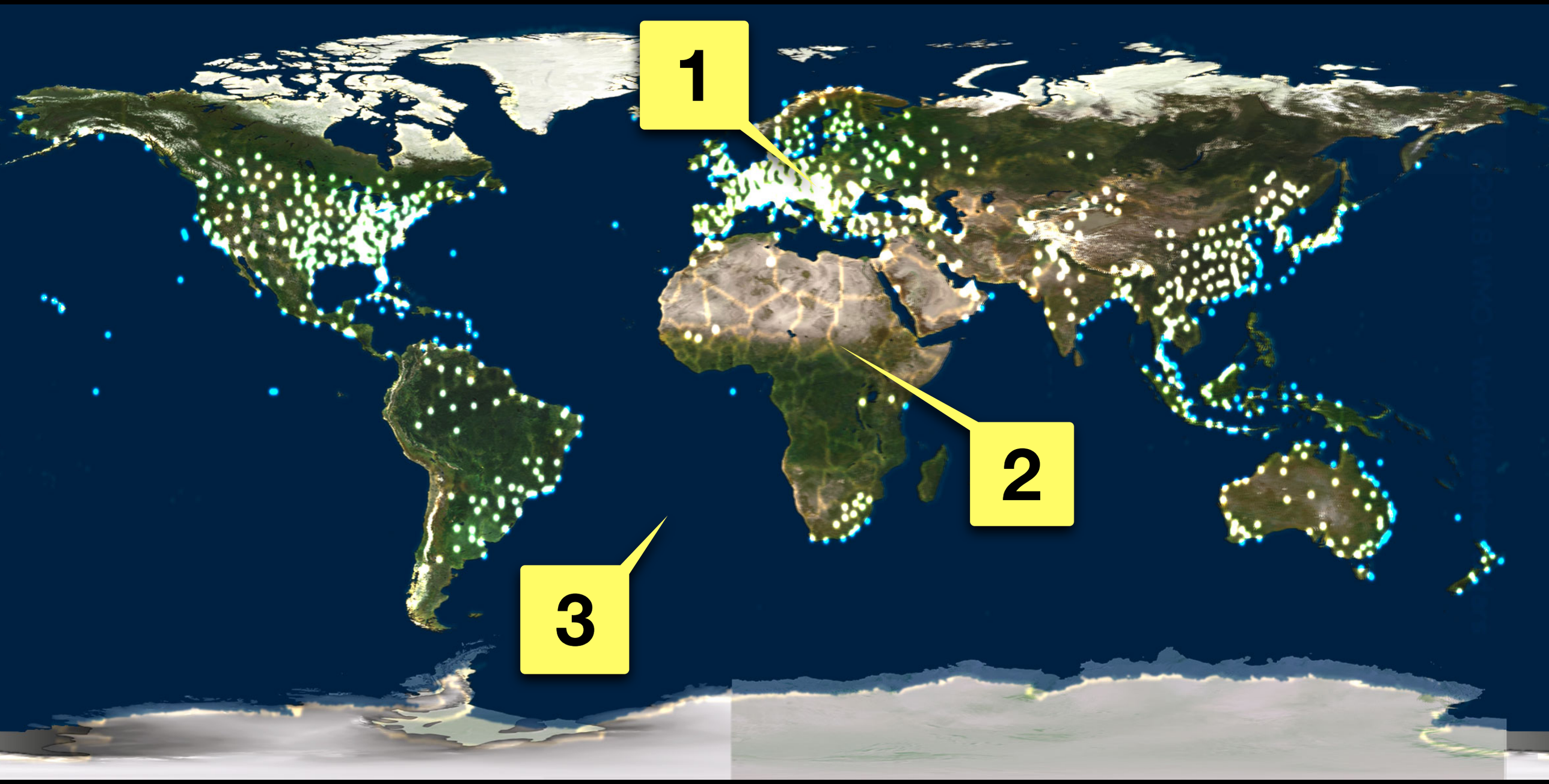
GPU



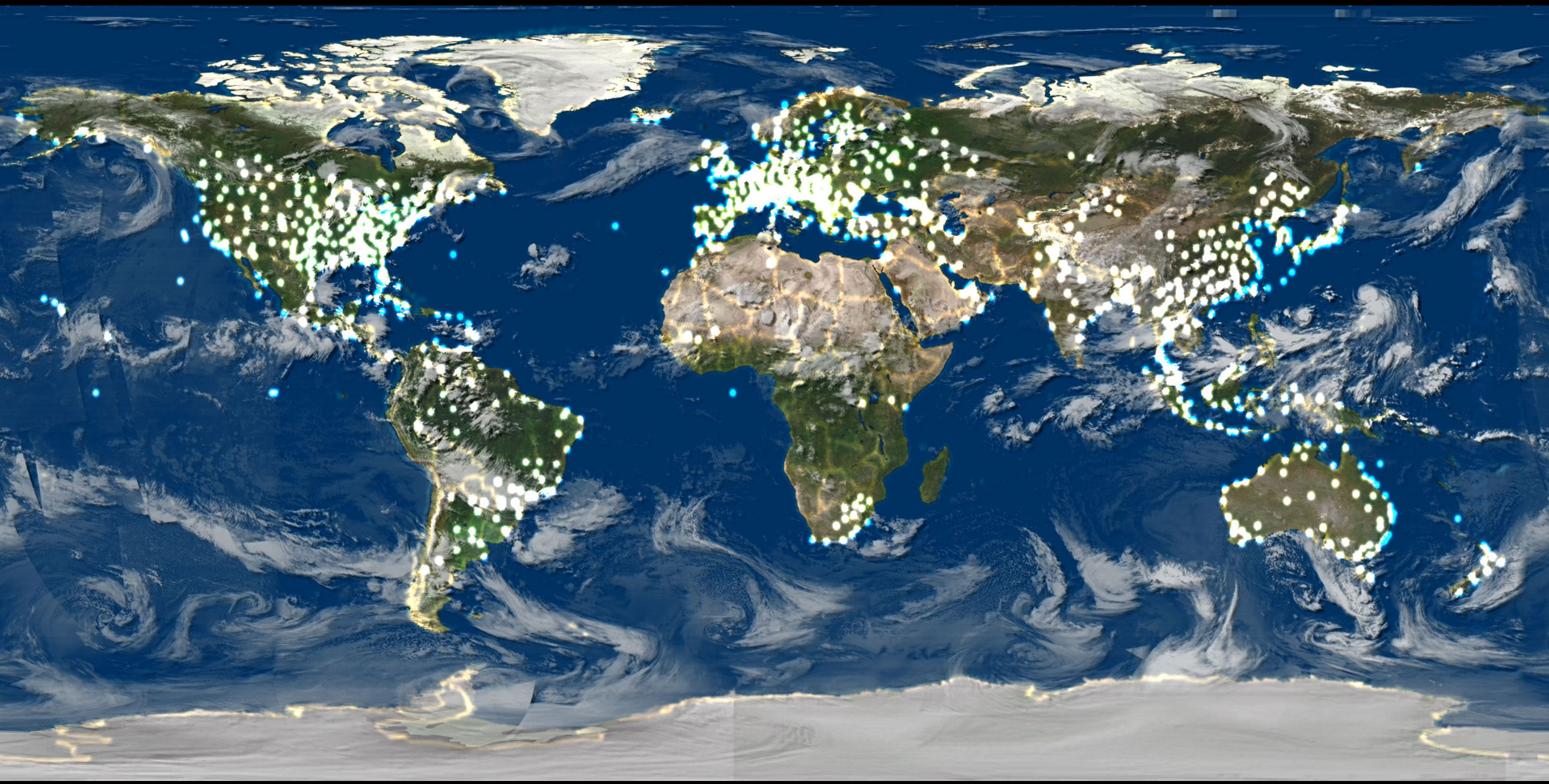
Deep Learning

Plan

Radar station map

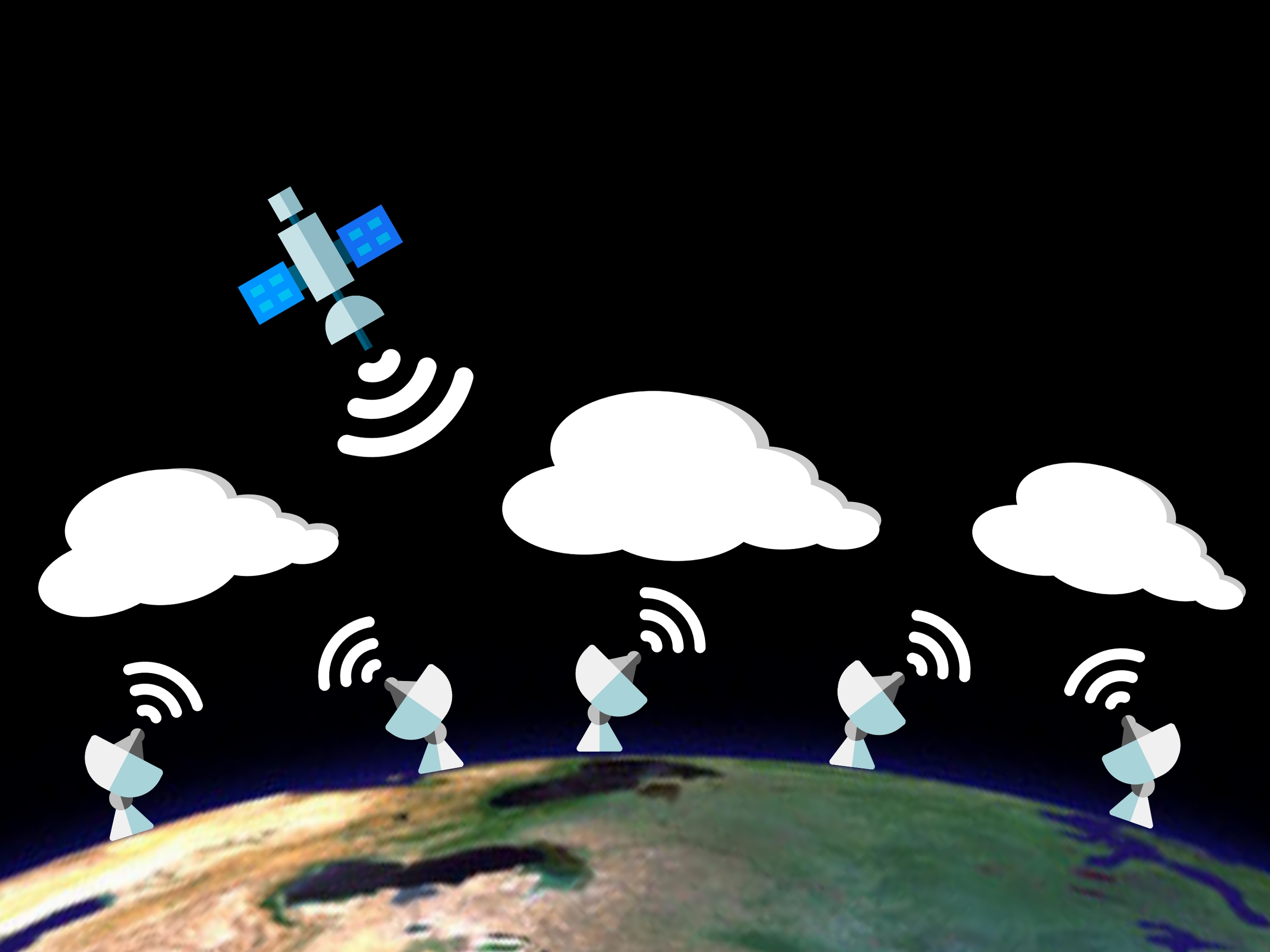


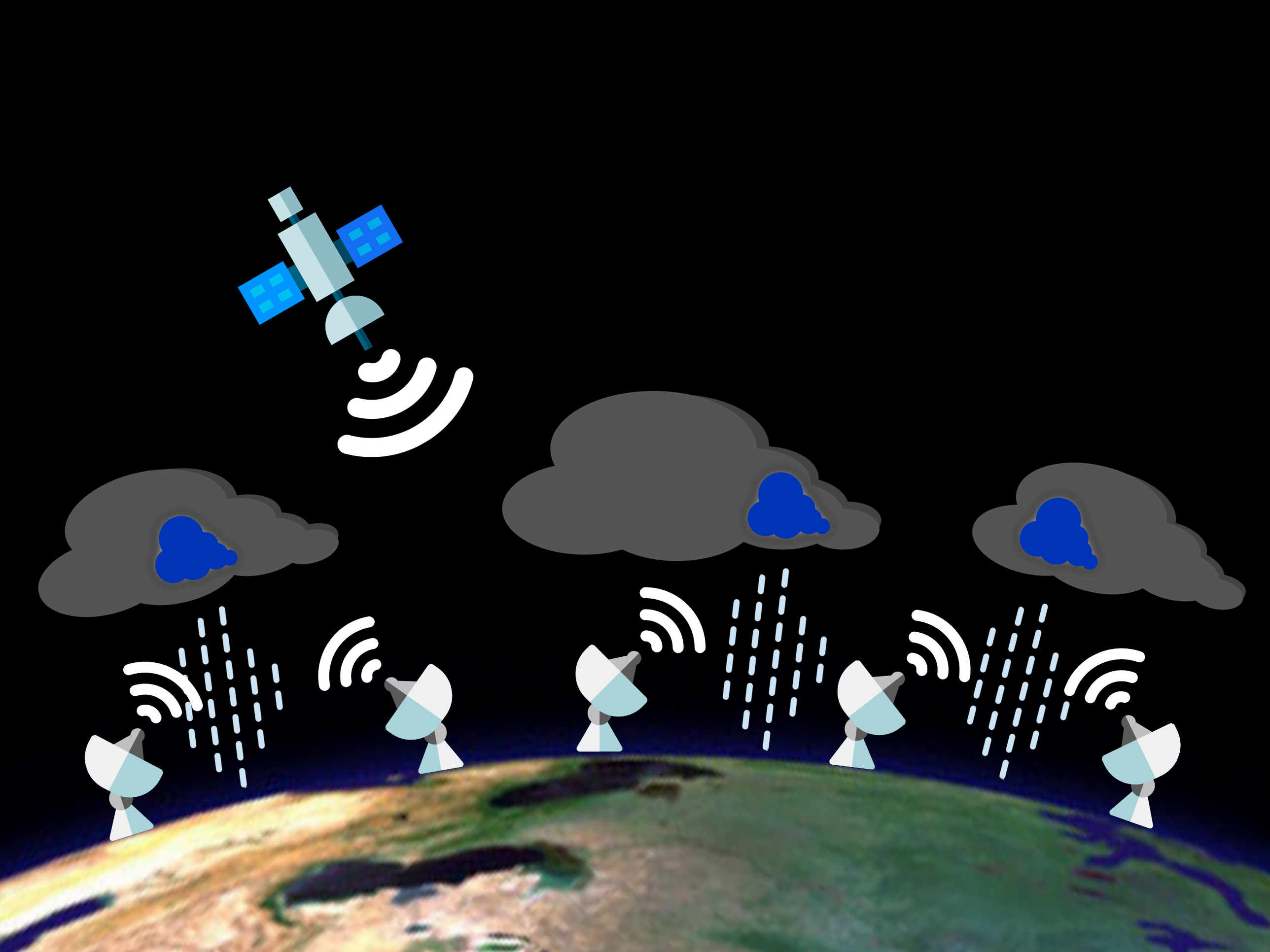
Satellites already cover the most of earth

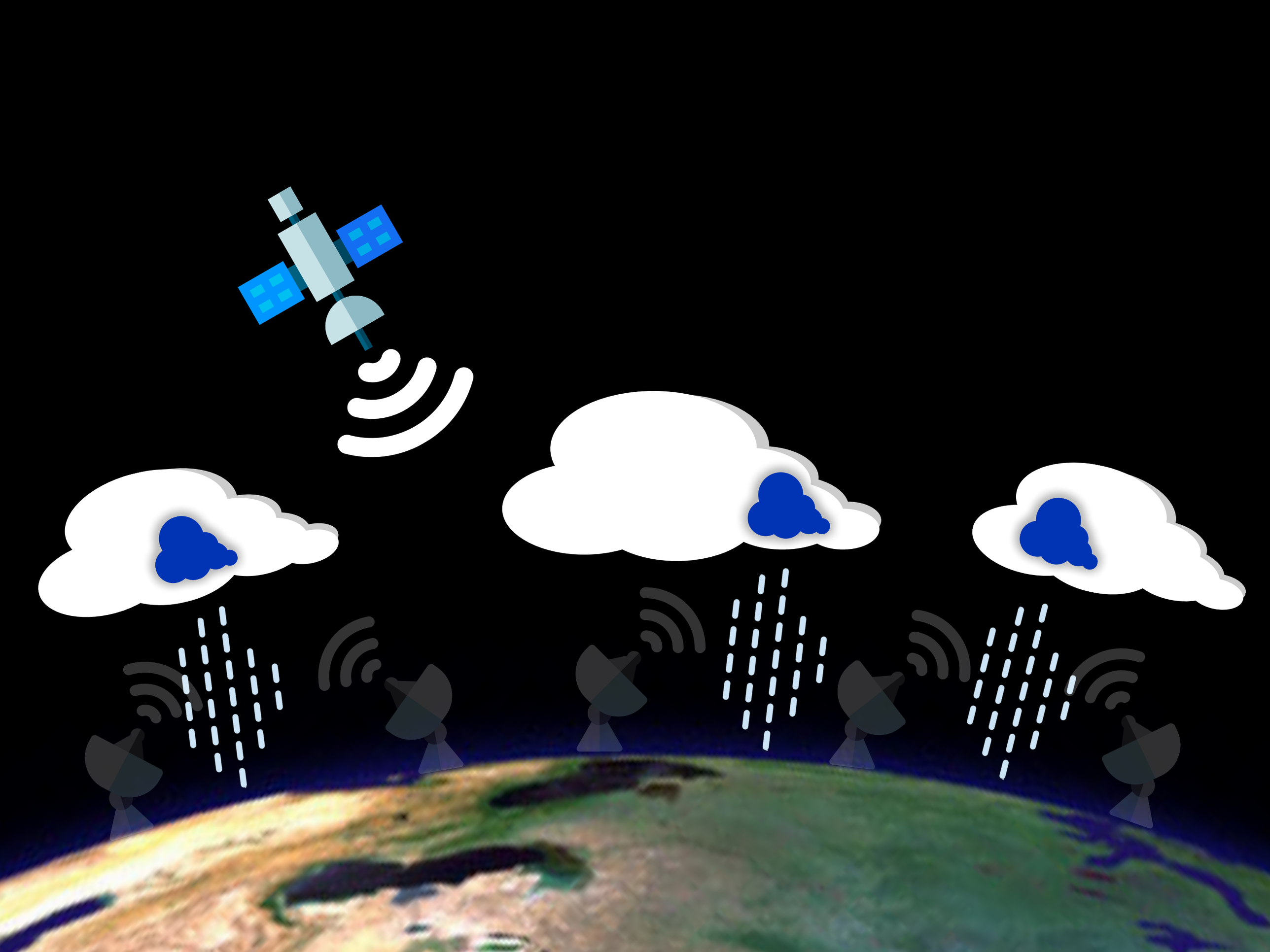


What if,

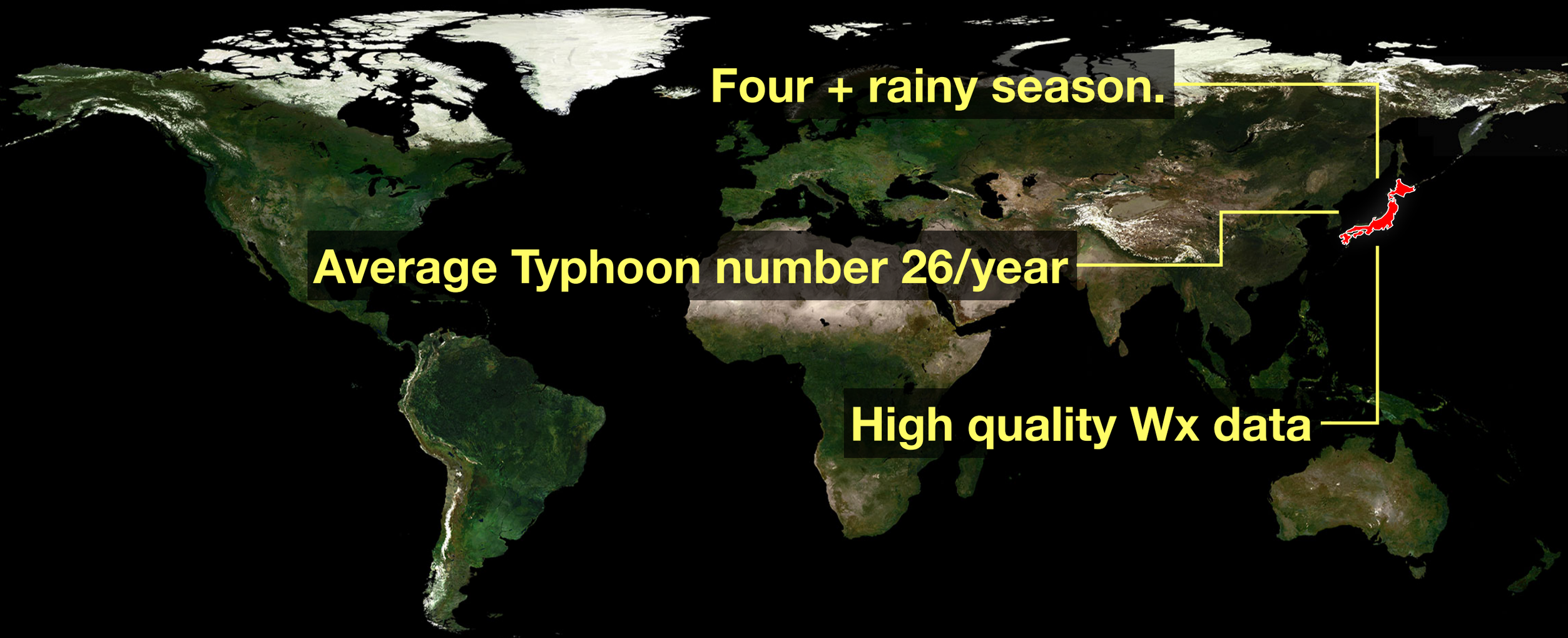
we could create radar data from Satellite image?







We pick up this small island!



as a benchmark country.

GTC 2019

S9164- Advanced Weather Information Recall with DGX-2

03/19/2019

Tomohiro Ishibashi - Director, Weather News, Inc.
Shigehisa Omatsu - CEO, dAIgnosis,Inc.

COMPANY PROFILE

GTC 2019

Circumstances

- Design / development engineers who dedicated to Computing services gathered.
- Started research on AI technology based on medical system development in a national project
- Established the company May 2017 with the theme of deep using GPU.
  VP of Google head office joined as a director.
- Advance technology development to build the original models while stu
  multiple cloud platforms.
- Started research using NVIDIA DGX-1 *7 +1 units (Volta in April 2018) from affiliates.
- Planned to start real-time analysis of text combined with image,etc. from the beginning of 2018.

OWNED TECHNOLOGY

GTC 2019

Highly Unique Technology

- Development of Booster Pack for building TensorFlow based on DGX-1
- Medical diagnosis support by combined processing of text analysis and image recognition
- Model optimization of business flow from business system program and model to speed up business processing with GPU

dAIgnosis,INC.

GTC 2019

First theme

Can we predict the next rain cloud from rain cloud radar?

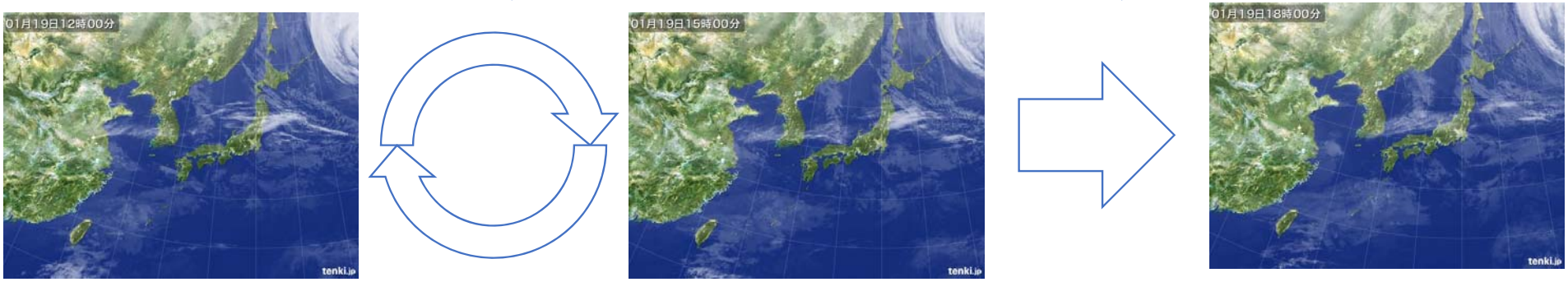
dAIgnosis,INC.

Initial adaptation to speculation of rain cloud movement

GTC 2019

Instead of learning the time change itself,
Learn relationships from variations of
equally spaced images

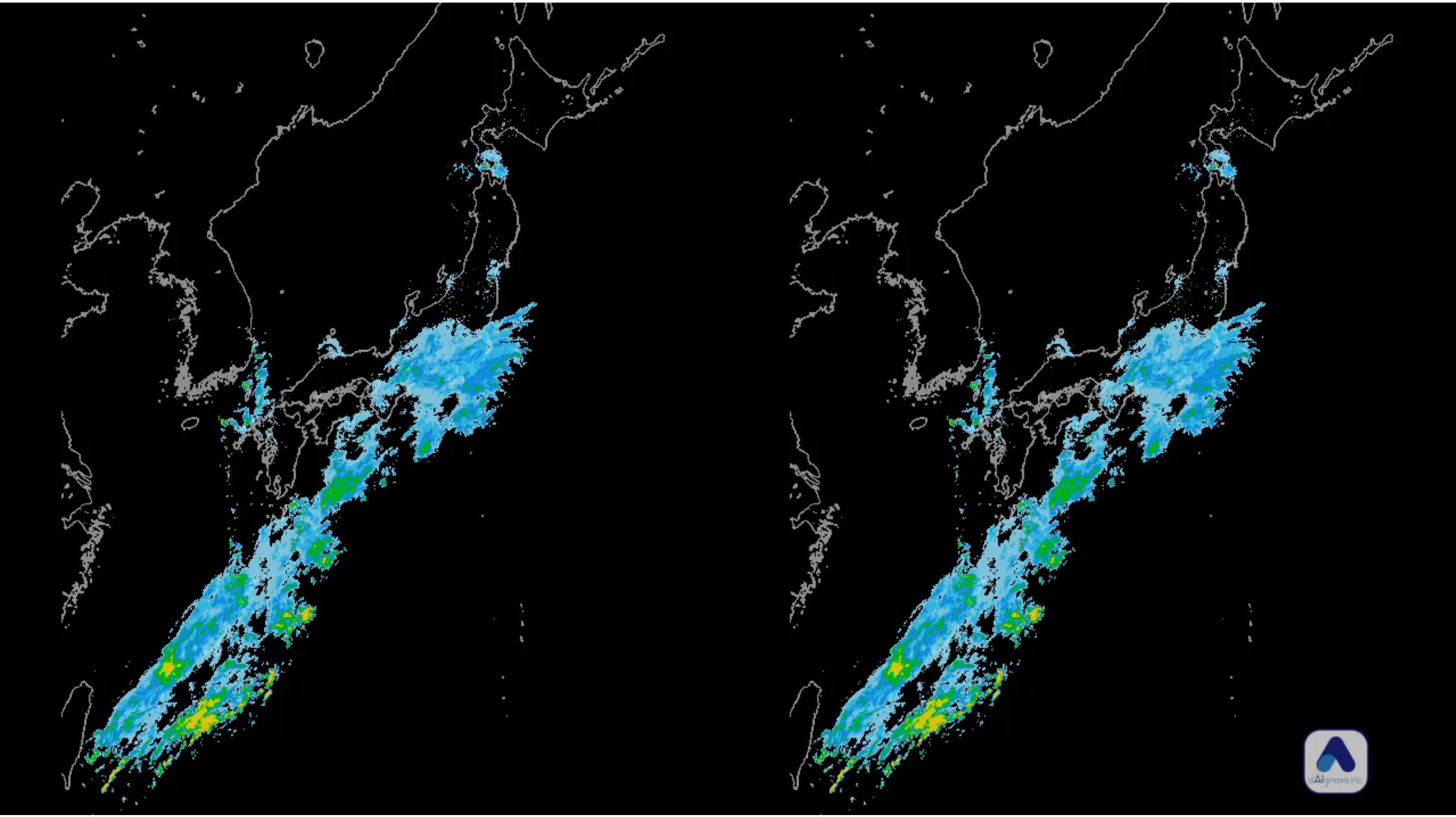
Using the learned model on the left,
Output the situation of the rain cloud of the
next step



Let the machine learning learn the relevance of the two images, input current rain cloud situation by reasoning
Output the state of the future rain cloud

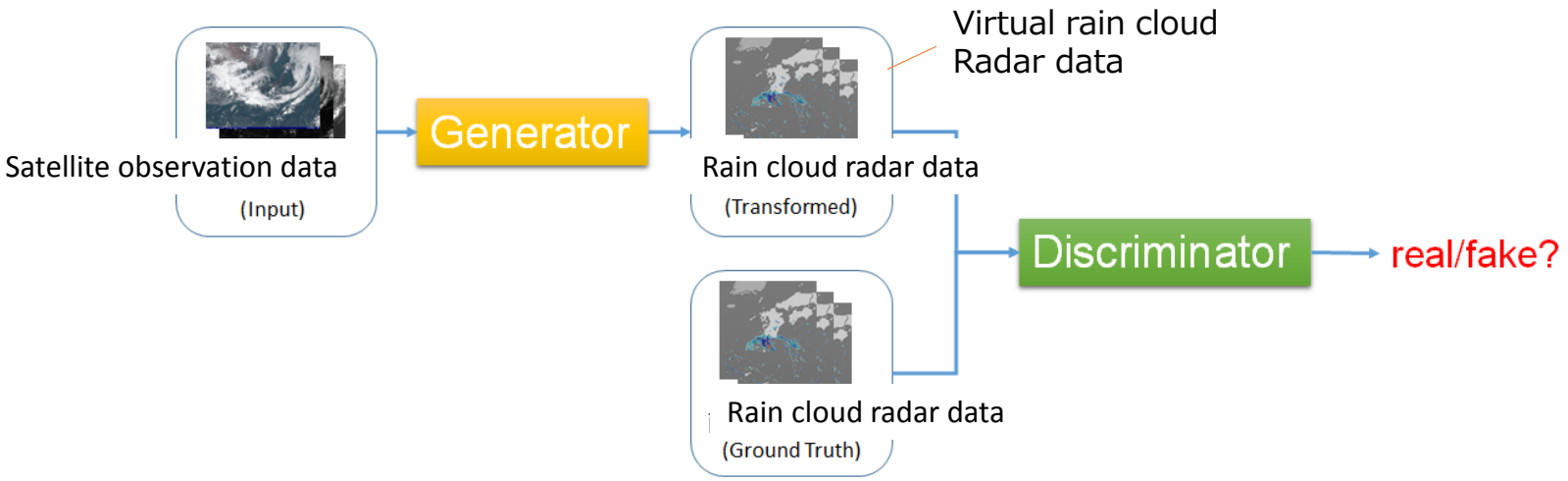
Since the amount of calculation required for learning is large, the DGX server is applied

dAIgnosis,INC.



Initially applicable model outline

GAN based technology, adopt pix 2 pix as architecture



Learn the relationship between satellite data and rain cloud radar data.
Infer rain cloud radar data from satellite data.

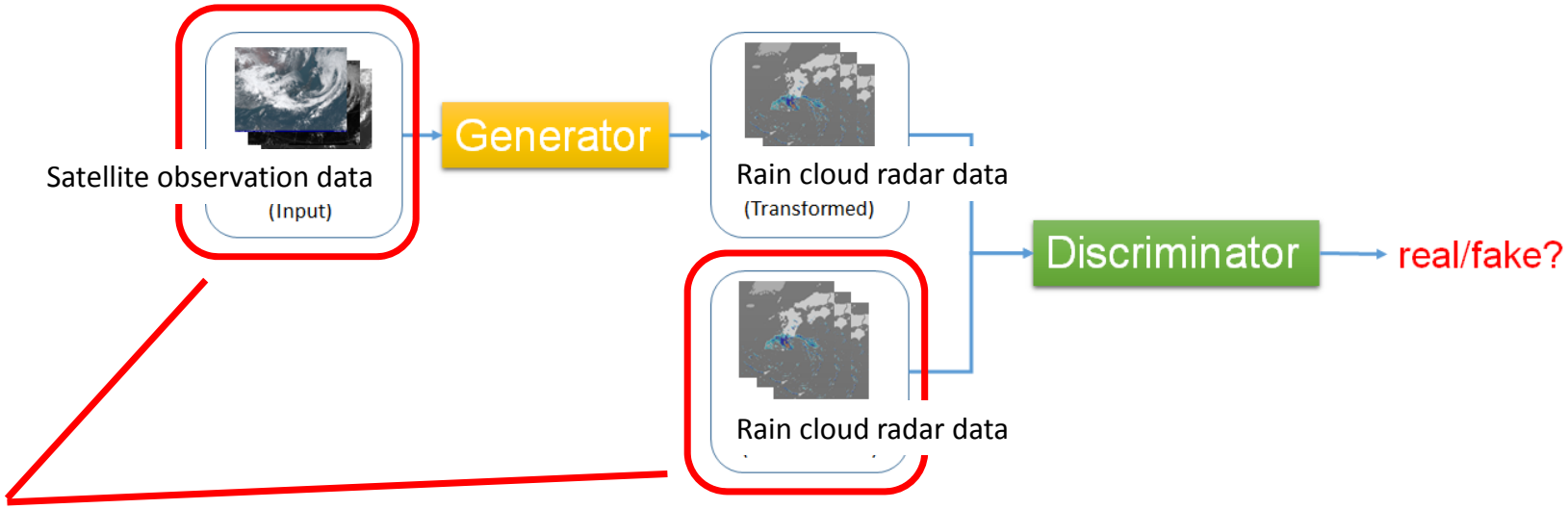
参考：https://phillipi.github.io/pix2pix/

Next theme

Can we generate rain cloud radar images from satellite images?

GTC 2019

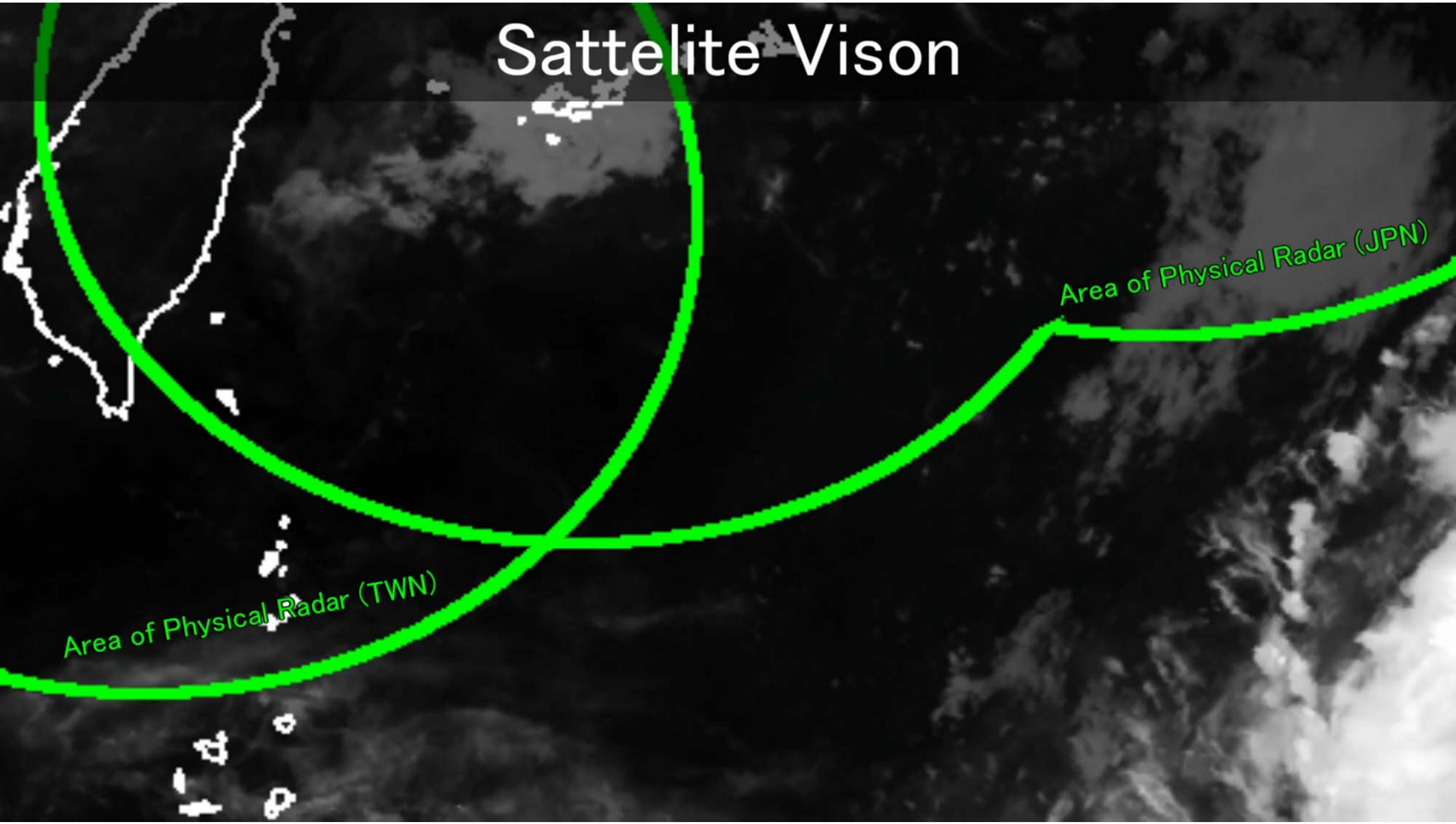
Approach to meteorological input data

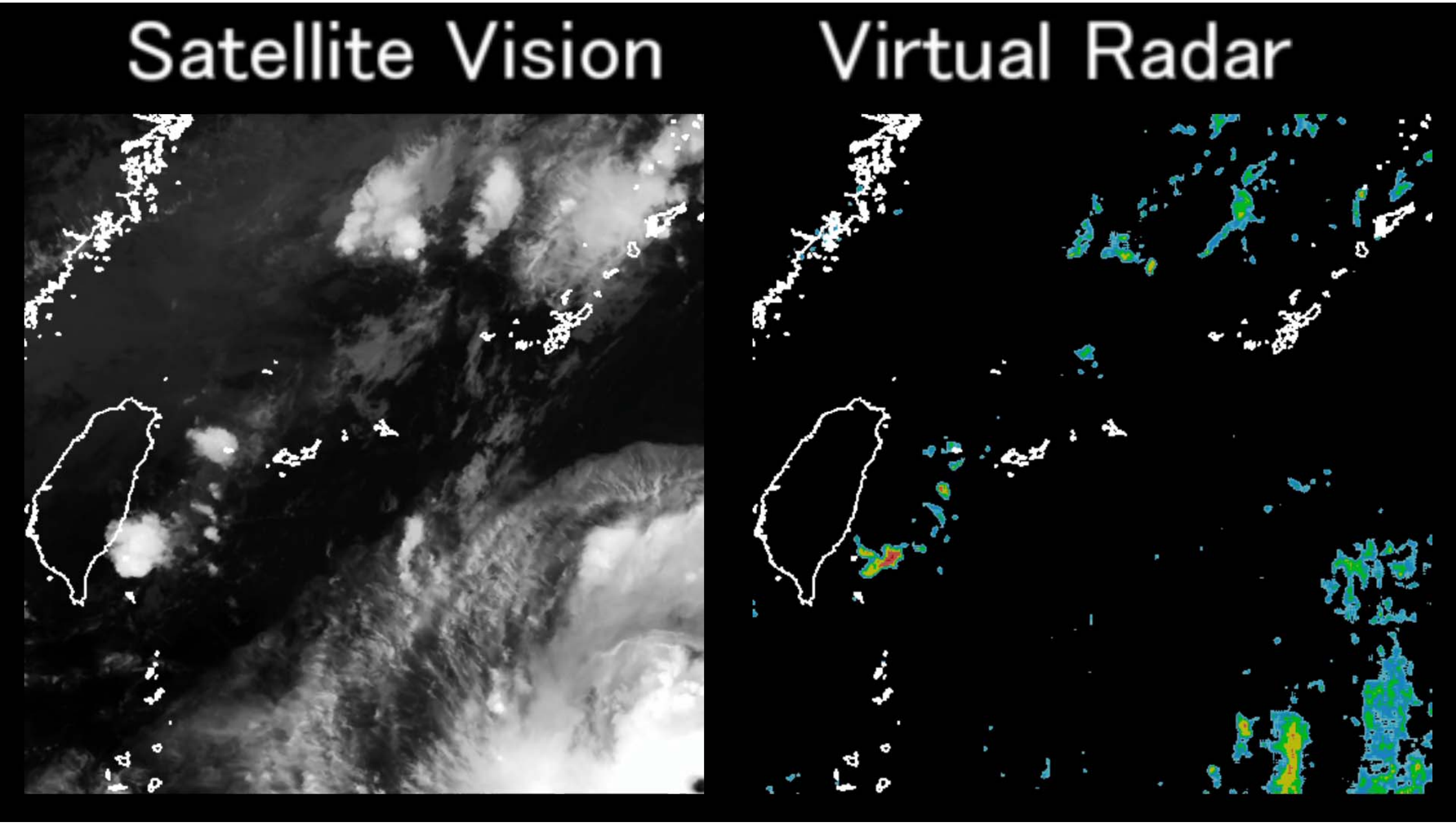


Use of numerical data
In GAN, there are many cases to use images based on images,
We used numerical data with higher expressiveness

Even when images are actually based on images when they are actually input to the model,
The numerical data is entered into the model as it is.
(By setting it as an image file, the value is rounded to the histogram of 256 gradations)

dAIgnosis,INC.


Sattelite Vison
Area of Physical Radar (JPN)
Area of Physical Radar (TWN)


Satellite Vision
Virtual Radar

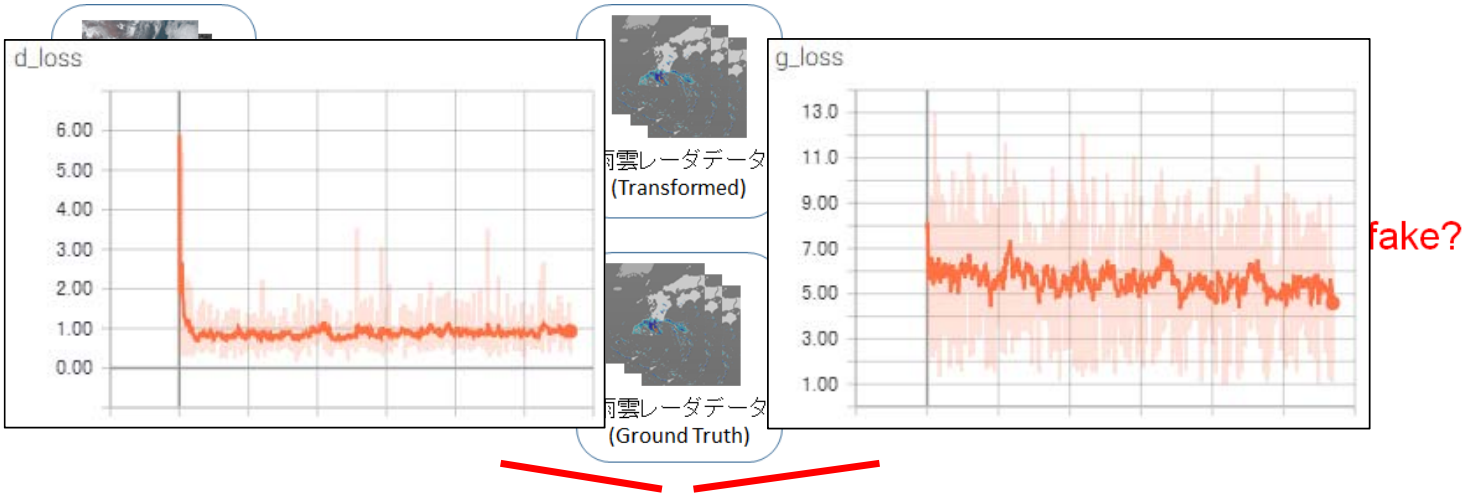
Introduction of a council system

- In machine learning, it is difficult to obtain 100% accuracy regardless of any improvement in accuracy.

- → In order to compensate for this fate, it is also used by Bonanza etc of Shogi software
- I will try introducing a council system.

- In this time, the implementation method of the consultation is from neighboring values at a certain point
- How to adopt median.
- It is also known as smoothing in two-dimensional plane (image processing).

GTC 2019



Confirmation of learning situation
As for GAN, since it is unknown whether intentional learning is done by value alone, confirm the progress of learning situation.

From the original, quoted

10: Track failures early

- D loss goes to 0: failure mode
- check norms of gradients: if they are over 100 things are screwing up
- when things are working, D loss has low variance and goes down over time vs having huge variance and spiking
- if loss of generator steadily decreases, then it's fooling D with garbage (says martin)

dAIgnosis,INC.

GTC 2019

Next theme

Is it possible to generate more accurate cloud radar images by adding satellite images other than rainy weather?

dAIgnosis,INC.

GTC 2019

Through business application

End users

Trained data

Business application

Trained model

Our own DGX-1 infrastructure

dAIgnosis,INC.

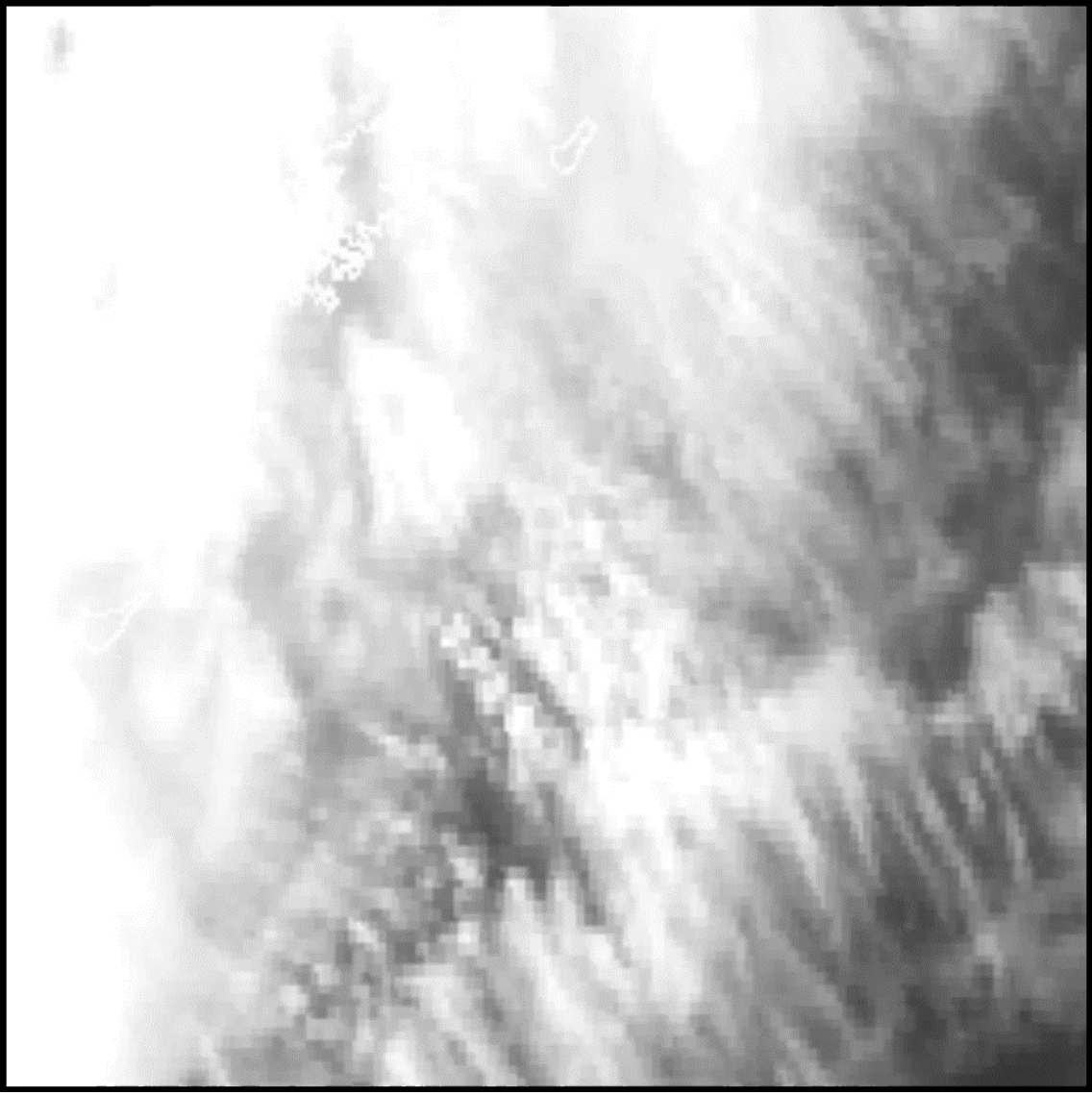
GTC 2019

Trying the virtual radar with DGX-2

- As an approach to estimate rainfall information using limited data from satellites, accuracy is raised with DGX server more.
- Establish a cooperative service of AI weather information at 1 k2 mesh.
- In order to be able to generate precipitation information that can be useful even in areas where real radars such as Asian countries and offshore are difficult to place
- _It corresponds to TensorFlow and it starts correspondence with TensorRT.

dAIgnosis,INC.

Satellite Vision



GTC 2019

Required resources for learning

Number of servers required for learning per model (all based on DGX-1)

GPU(8GPU)	CPU Only
0.33	255

■ Frame interpolation
· In the verification stage, it took about 17 hours (44 sec / 1 epoch * 1,400 peoch) to converge 30 day data learning
· Assume that the difference learning is performed on a daily basis and the model is updated with full learning again on a monthly basis (assuming that the processing time scales with the data amount / GPU allocation number)
· 1 daily GPU allocation with daily ~ 1 day data: 17 (hours) * 1/30 (day) * 8/1 (GPU) = 4.53 hours
· 7th GPU allocation with monthly ~ 360 days worth of data: 17 (hours) * 360/30 (day) * 8/7 (GPU) = 233 hours
→ By sliding time zone to be learned for each model, it is estimated that 2 models can be operated per unit

■ Create virtual radar
In the verification stage it took about 1 hour (70 seconds / 1 epoch * 50 peoch) to converge the learning of data for two days
· Assume that the difference learning is performed on a daily basis and the model is updated with full learning again on a monthly basis (assuming that the processing time scales with the data amount / GPU allocation number)
· Daily ~ 1 day data with 2 GPU allocation: 1 (hour) * 1/2 (day) * 8/2 (GPU) = 2 hours
· 6 GPU allocation with monthly ~ 180 days worth of data: 1 (hour) * 180/2 (day) * 8/6 (GPU) = 120 hours
→ By sliding the time zone to be learned for each model, it is estimated that 4 models can be operated per unit
It is assumed that on average the above three models can be operated on average per DGX-1 (0.33 per model)

→ It takes 9 hours and 30 minutes per epoch when processing with CPU only, processing speed is scaled to 772 times by GPU

dAIgnosis,INC.

GTC 2019

Inference throughput

Inference requests that can be processed per hour (all based on DGX - 1)

GPU(8GPU)	CPU Only
514,286	48,979

In frame interpolation, 7 ms per inference (8 inference per 1 GPU at 8 GPU, about 440 ms 440/64 ≈ 7 ms in a total of 64 inferences)
(Since frame interpolation occupies a large number in inference, this throughput is adopted as a reference value)

→ It takes 73.5 ms per inference when processing with only CPU, processing speed is scaled up to 10.5 times by GPU

dAIgnosis,INC.

GTC 2019

When DGX-2 is applied

Number of servers required for learning per model

0.03

GPU(8GPU)	CPU Only
0.33	255

Inference requests that can be processed per hour

5,100,000

GPU(8GPU)	CPU Only
514,286	48,979

dAIgnosis,INC.

GTC 2019

http://www.daignosis.com
omatsu@daignosis.com

Thank you.

dAIgnosis,INC.