Short-Term Quantitative Precipitation Forecasting

Group 16

Outline

- Team Members
- Target Problem
- Brief descriptions on the datasets
- Work plan & Methods Used
- Experimental results
- Future work

Team members

李韡 *Wei Li*

勾小川 *Xiao-Chuan Gou*

周靄柔 *Ai-Jou Chou*

張敬 *Ching Chang*

邱子翔 *Zi-Xiang Qiu*

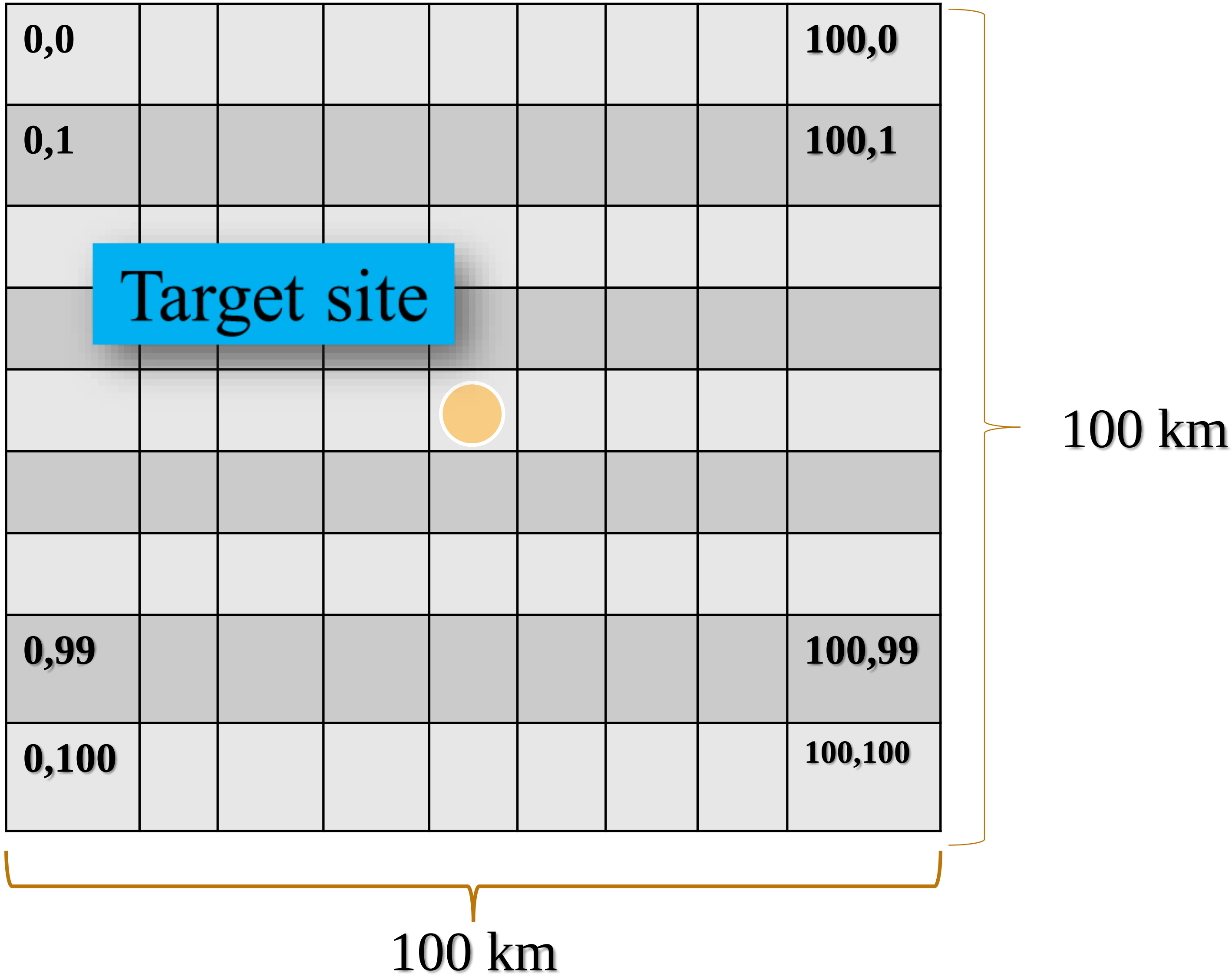
Target Problem

- CIKM AnalytiCup 2017
  - CIKM is called Conference on Information and Knowledge Management.
  - Since its inception in 1992, CIKM has successfully brought together a wide range of R & D personnel in the field of knowledge management, information retrieval and database.
  - CIKM 2017 will be a unique perspective, strategic penetration of knowledge, information and data management of the cross-type research, highlighting the realization of many urban areas and their countries to share the "smart city, smart country" vision of the technology and insights.

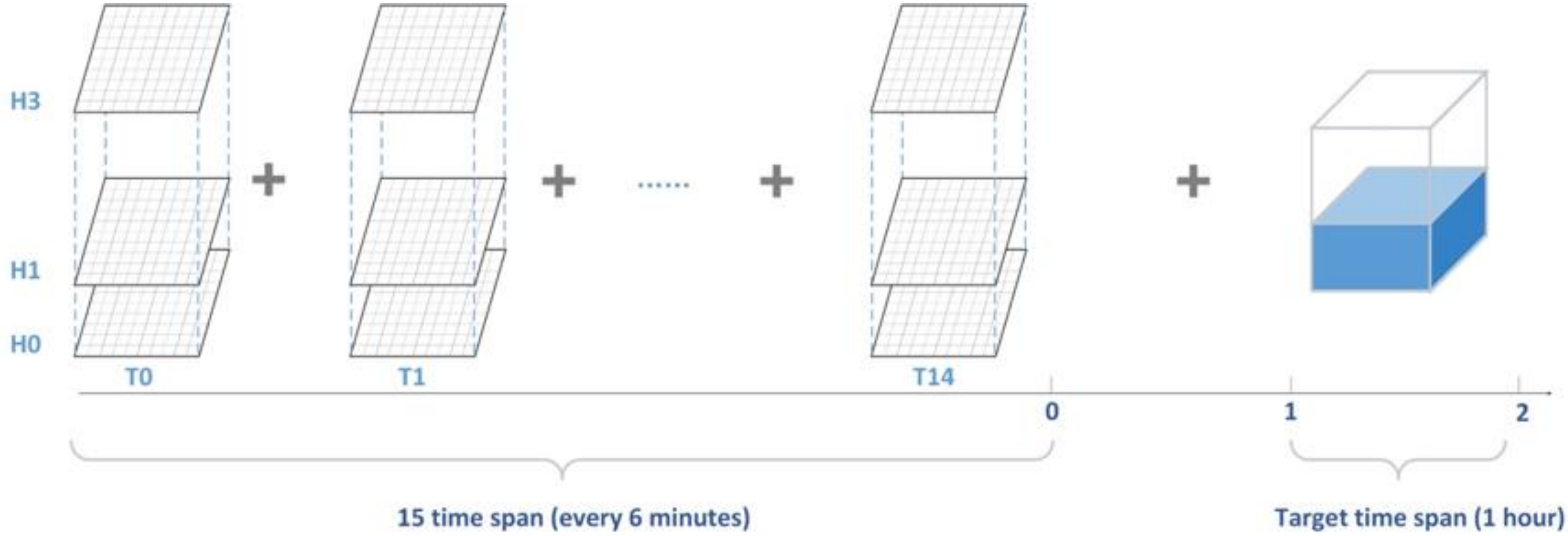
Target Problem

- CIKM AnalytiCup 2017
  - It is an open, notarized big data open competition. For the academic community, it is an exciting data challenge.
  - The short-term precipitation forecast jointly conducted by Shenzhen Meteorological Bureau and Alibaba, aims to significantly improve the accuracy of short-term precipitation forecasts based on radar echo extrapolation data.

Brief descriptions of datasets



Brief descriptions of datasets(cont.)



Brief descriptions of datasets(cont.)

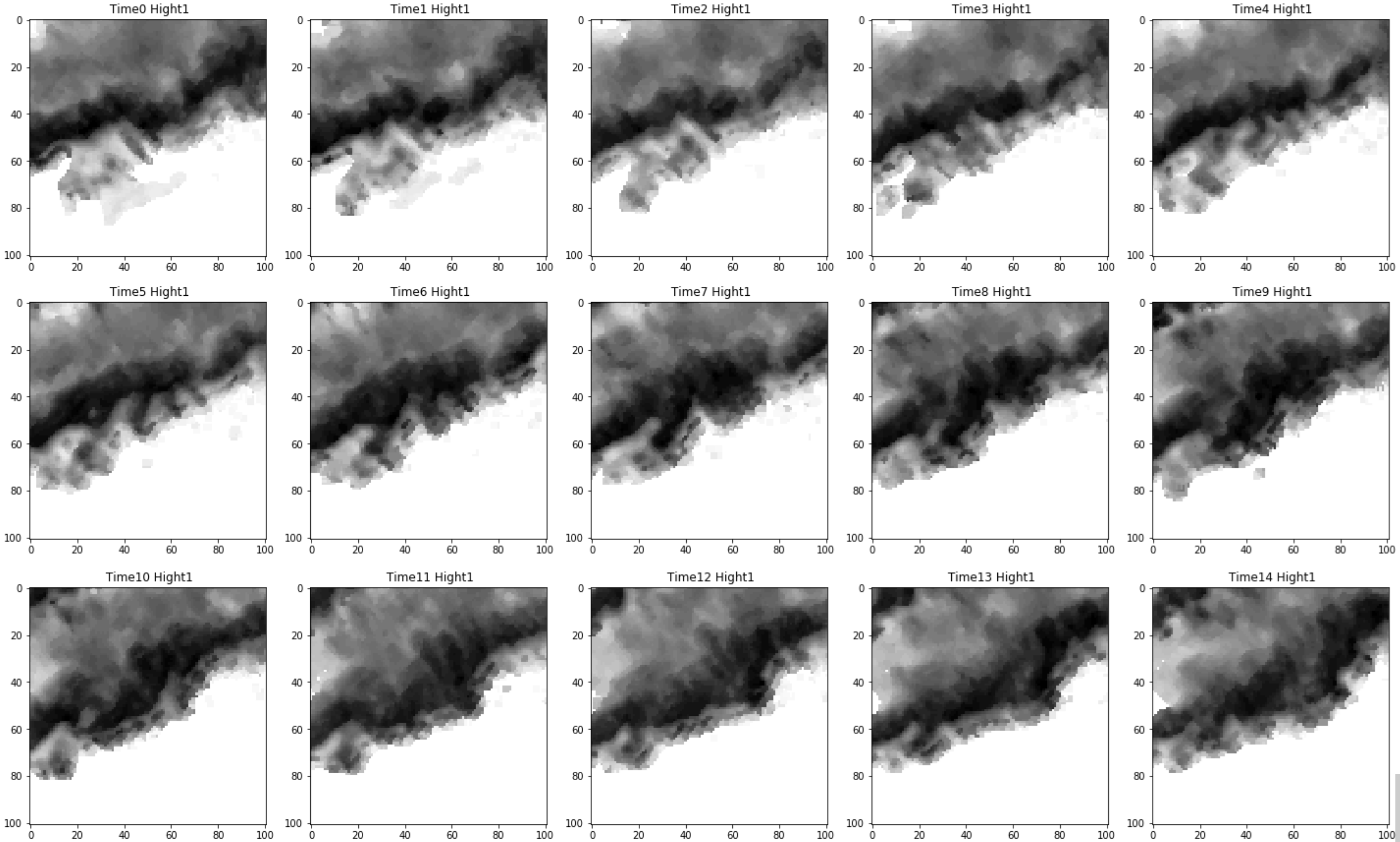
- Train Data: **15.8GB** (two years)
  - Training sets: 10,000
  - each data about 2MB **[id, label, 60 radar map]** (15*4*101*101)

- Test Data: **3.09GB** (one year)

Plot the radar image

- 15 time span
- height 1
- Y = 71.6

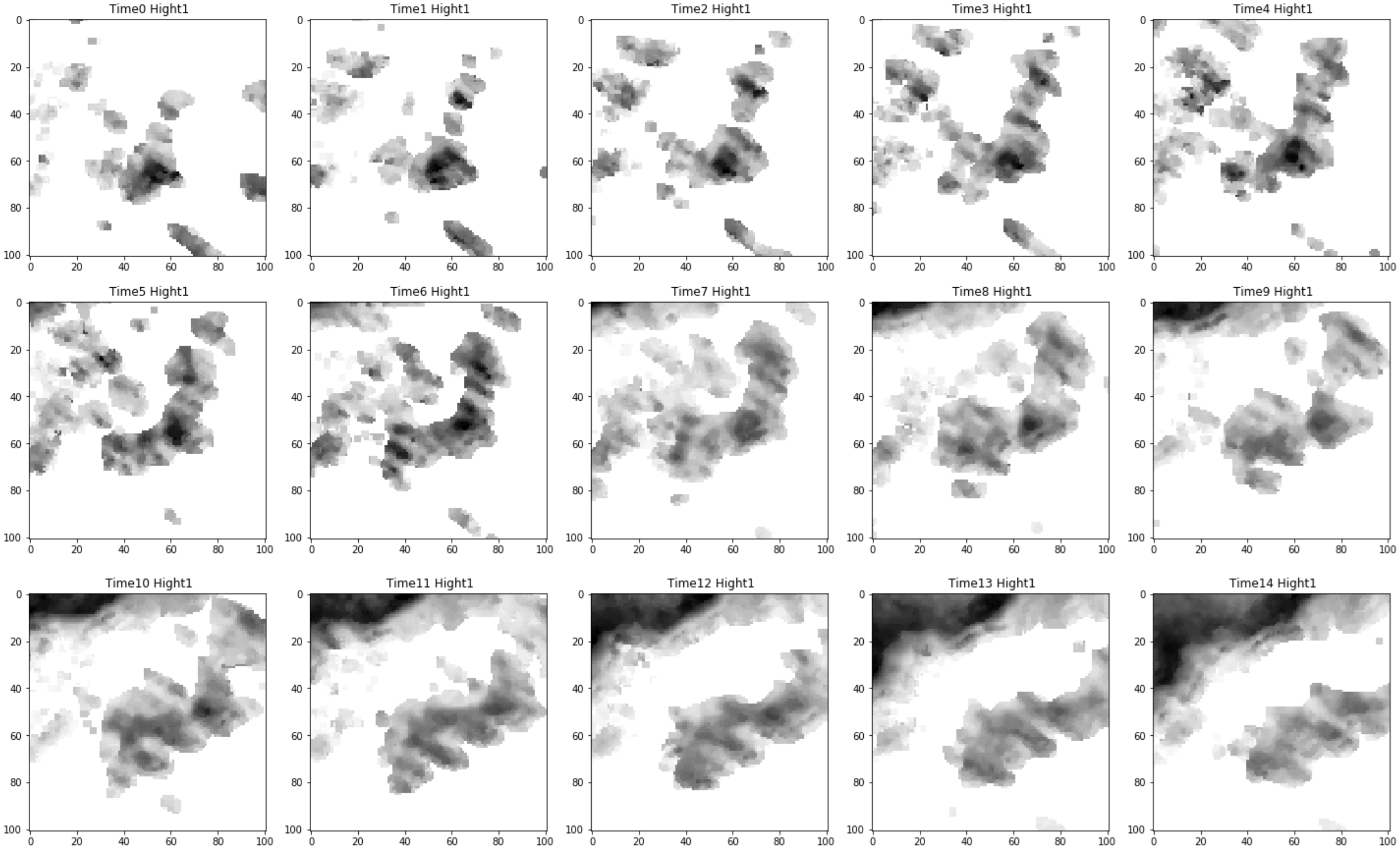
- No wind



Plot the radar image(cont.)

- 15 time span
- height 1
- Y = 29.9

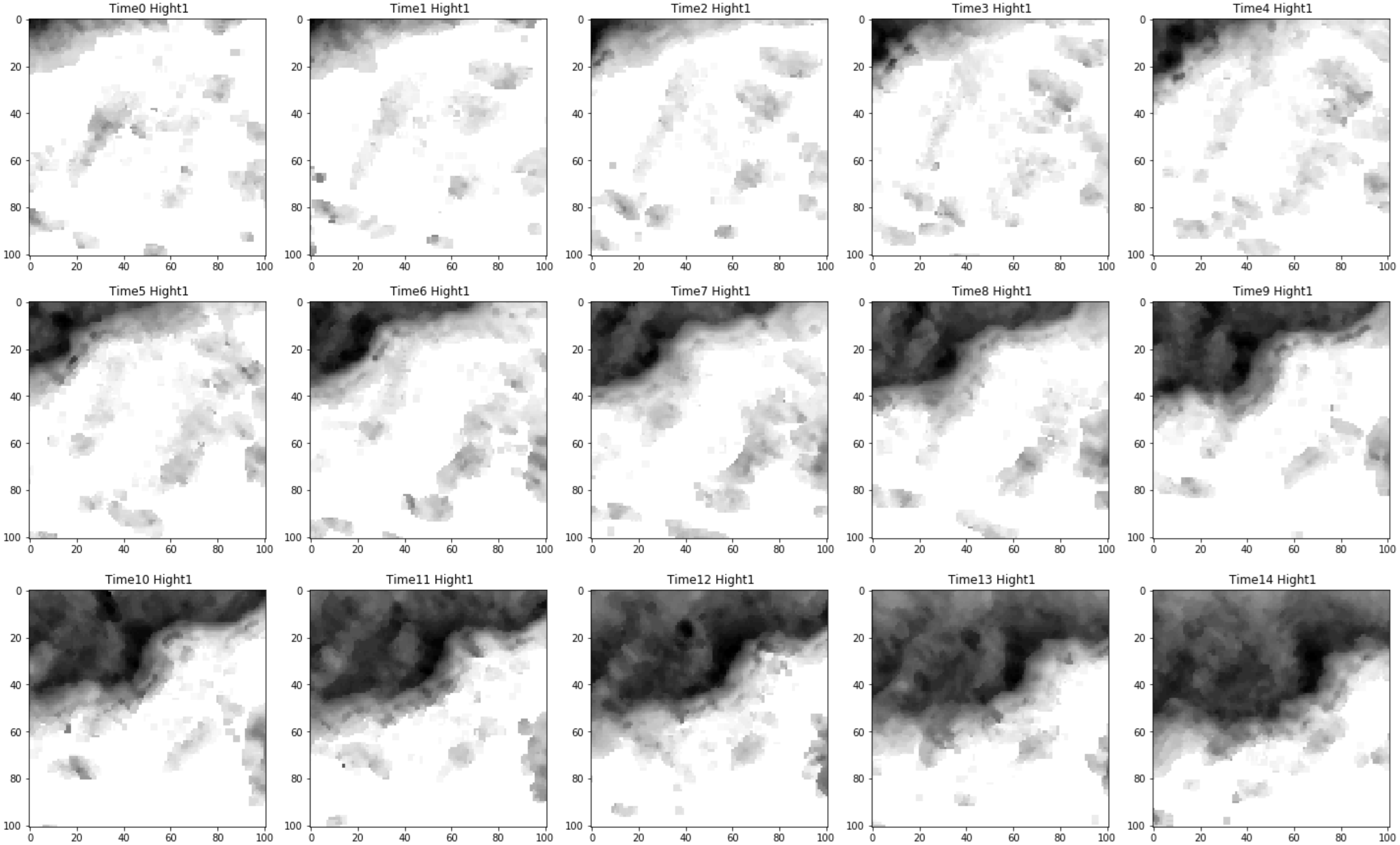
- No wind



Plot the radar image(cont.)

- 15 time span
- height 1
- Y = 4.5

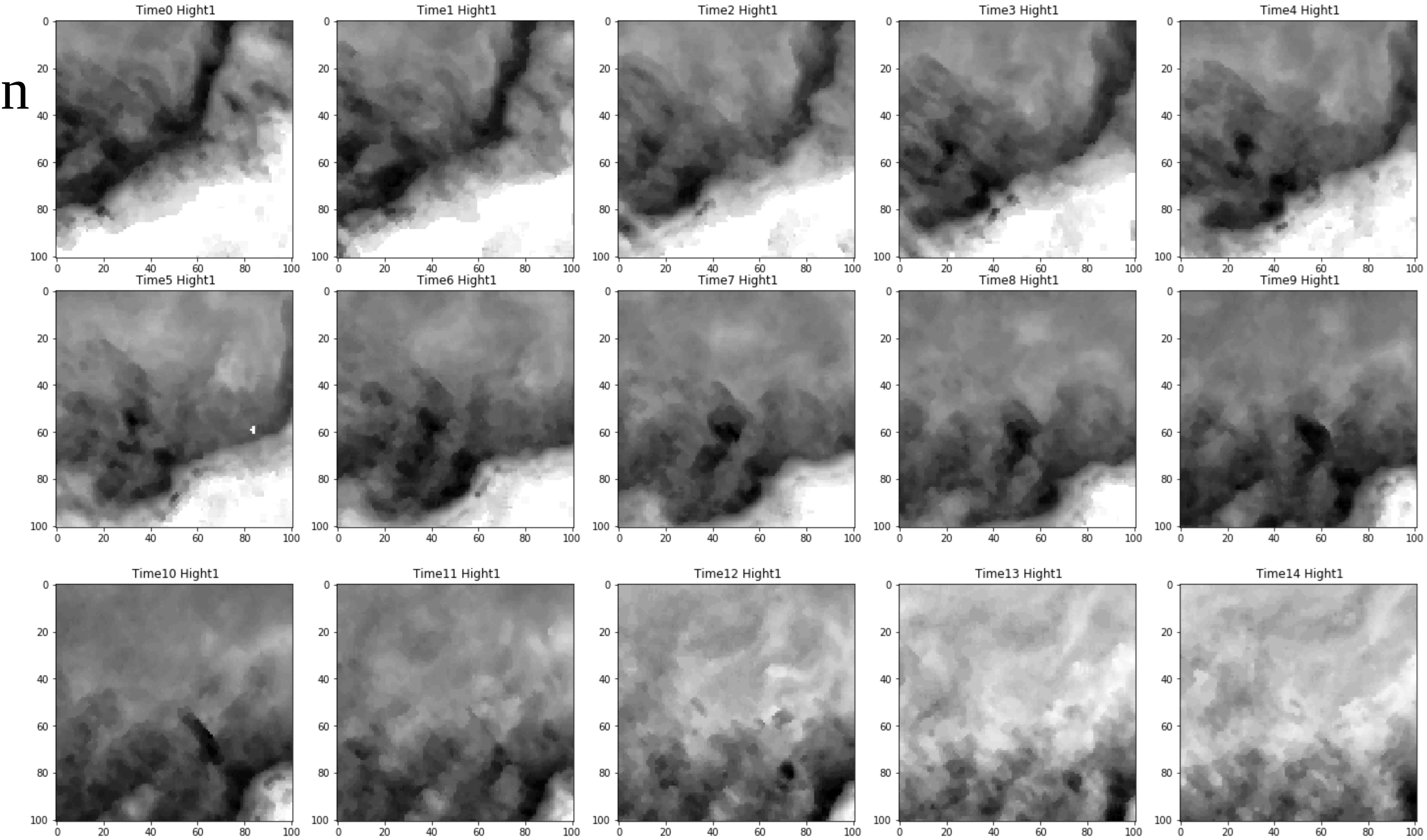
- Windy



Plot the radar image(cont.)

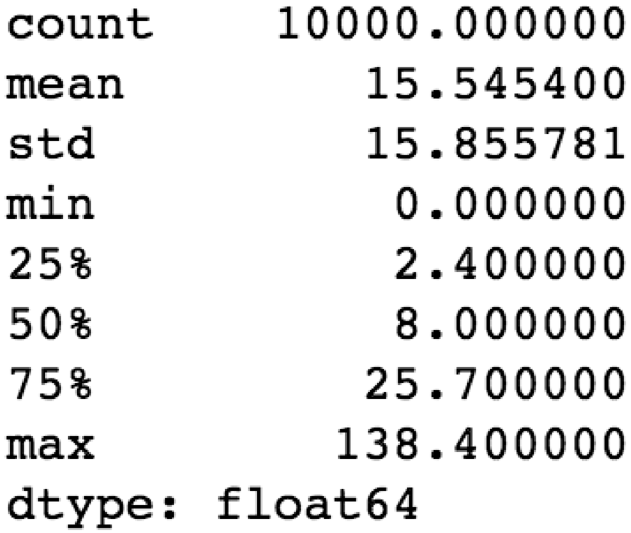
- 15 time span
- height 1
- Y = 2.3

- Windy

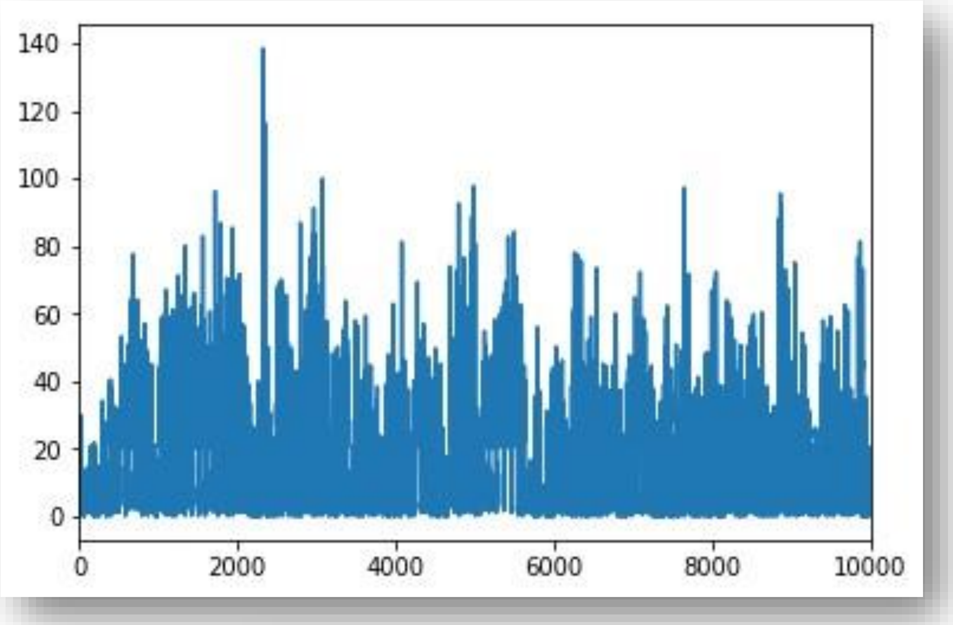


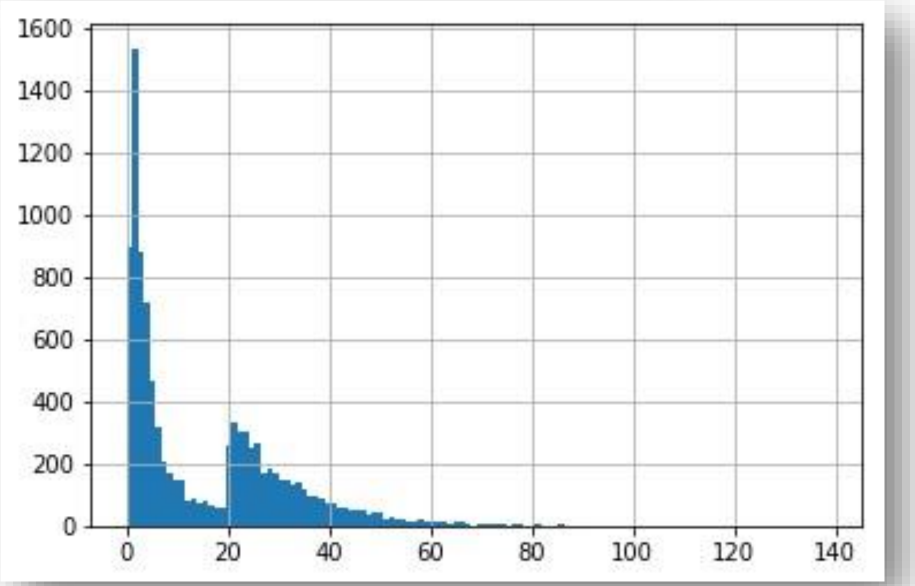
Plot the label y

- Summary Statistics



```
count    10000.000000
mean        15.545400
std         15.855781
min          0.000000
25%          2.400000
50%          8.000000
75%         25.700000
max        138.400000
dtype: float64
```





Work plan & Methods Used

- Platform and Tools
  - Hadoop & **Spark** & Mongodb (Pig, **MLlib** etc.)
  - **Python** (scikit-learn pandas numpy etc.)
  - **TensorFlow** & **keras** (training CNN/RNN)
  - etc.

Work plan & Methods Used(cont.)

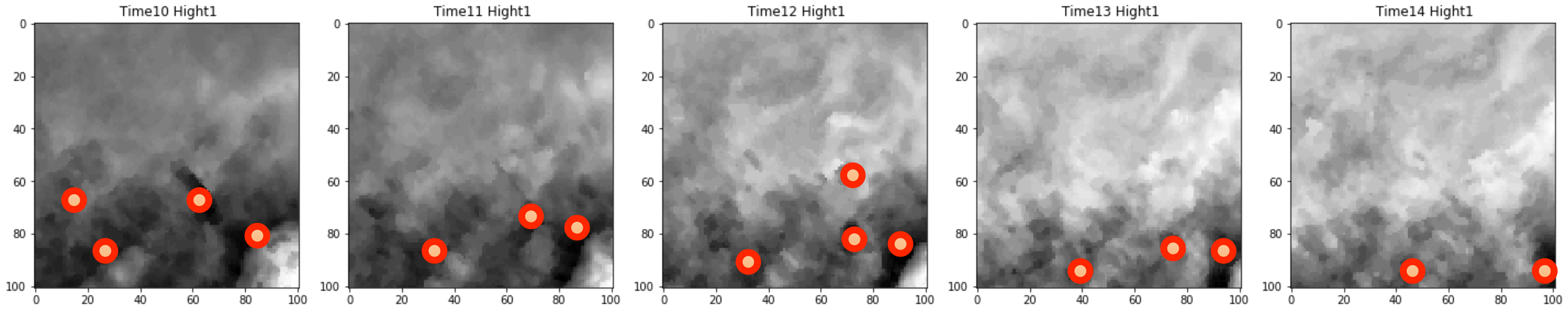
- Workflow
    - Basic methods
        - Linear Regression
        - SVM
    - Preprocess Clustering methods
    - Deep Learning
        - Classification
        - Regression

Work plan & Methods Used(cont.)

- Basic methods(baseline)
  - Linear Regression
    - Only using the target point(50,50) to train the model
    - Only using the mid 3x3 map to train the model
  - Ensemble Linear Regression
    - According different changing rate of images, train the models by different situation.
  - SVM
    - Only using the target point(50,50) to train the model
    - Only using the mid 3x3 map to train the model

Work plan & Methods Used(cont.)

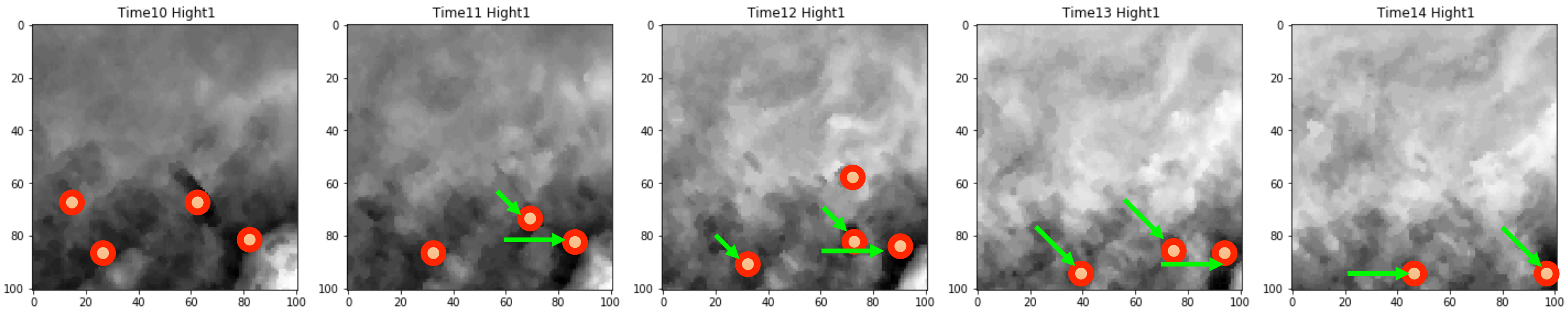
- **Preprocess Clustering Method**



- Generate Clusters
  - Find Peak point - the potential center of cluster
  - Aggression: Flood Fill. Find out the real center of clusters.

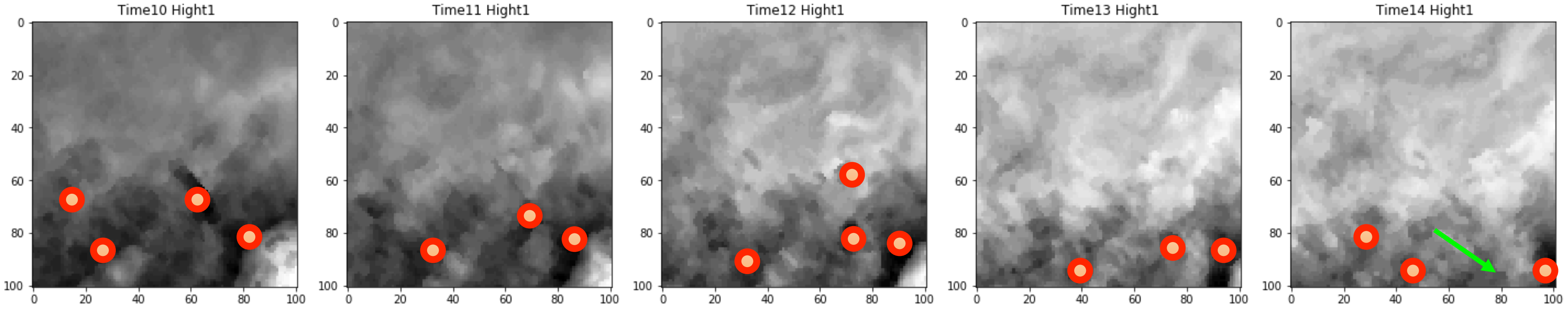
Work plan & Methods Used(cont.)

- **Preprocess Clustering Method**



- Calculate Immigration Direction
  - Do statistics of the direction of 14 images.
    Select the most possible one.
  - Speed: calculate the speed of both latitude and longitude.

Work plan & Methods Used(cont.)

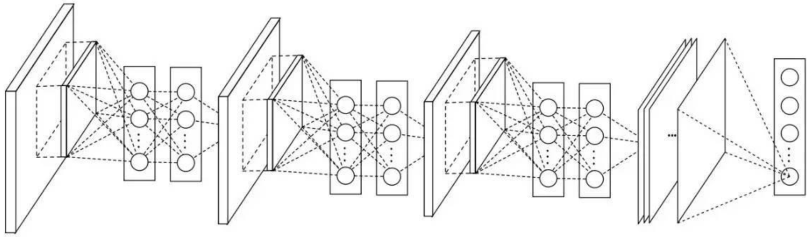


- The Influence of cluster works on center point
  - We only consider the last image
  - Predict cluster position: in the next image, the cluster center will move, so we should take speed in our formula.
  - The bound of cluster: we should consider not only the radar value of the cluster center, but also the point around center.

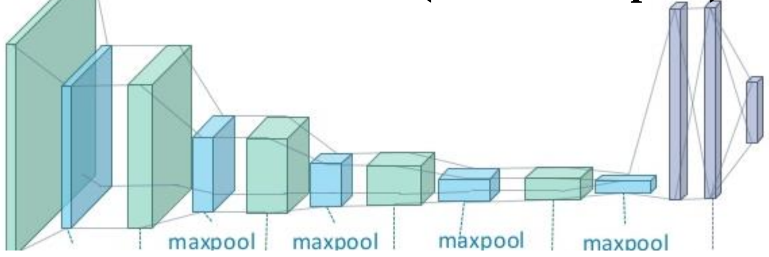
$$y = w_1 \sum_{i=1}^{k} \frac{v_i}{D^2} + w_2 \sum_{i=1}^{k} \frac{\sum_{j=1}^{9} v_{ij}}{9D^2} w_2 \sum_{i=1}^{k} \frac{\sum_{j=1}^{25} v_{ij}}{25D^2}$$

Work plan & Methods Used(cont.)

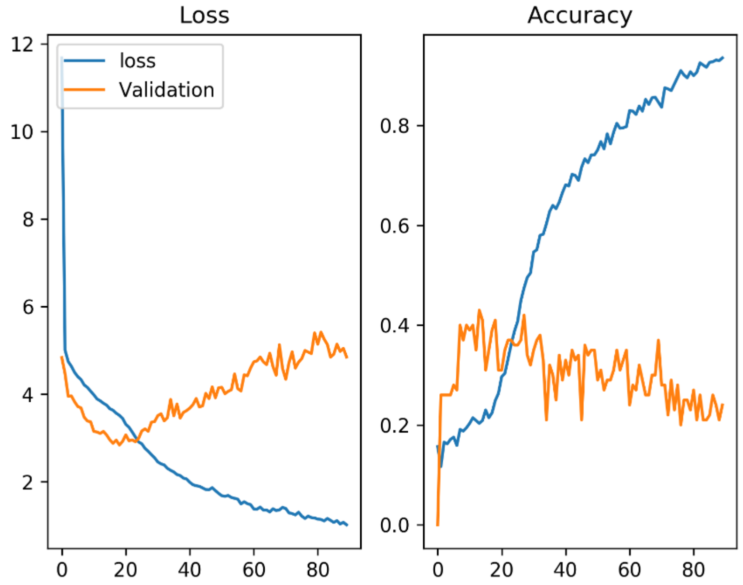
- **Deep Learning(Classification)**
  - Assume 100 classes
  - Training acc is good, but test acc is bad(overfitting)
    - 100 classes is too small? Maybe 1000?
    - Training data is not enough.(the more, the better)



NIN Architecture(detail in report)

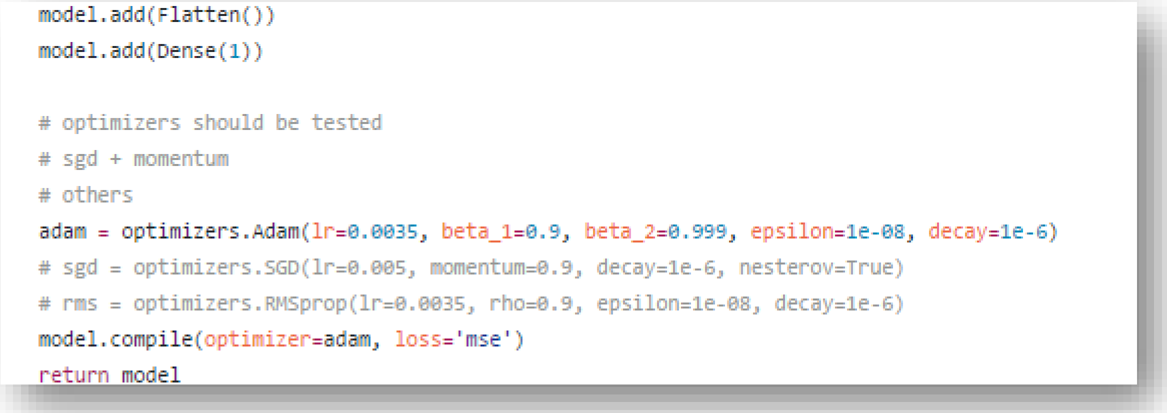


VGG19 Architecture



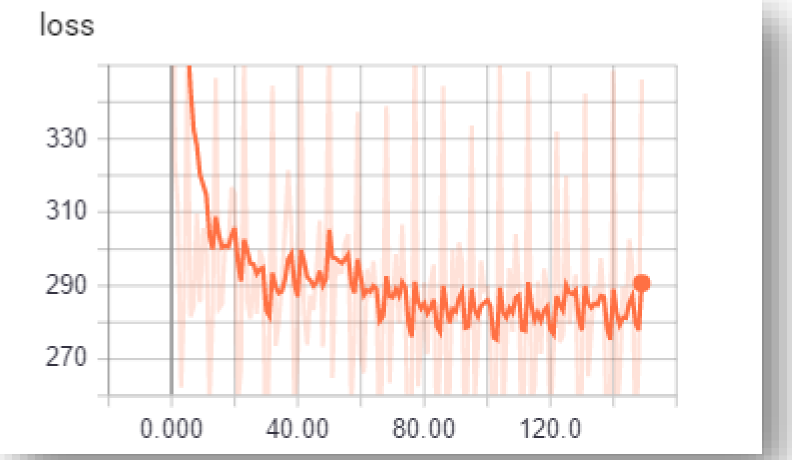
Work plan & Methods Used(cont.)

- **Deep Learning(Regression)**
  - Modify CNN model to regression (**detail in report**)
    - Add fully connected layer to instead of softmax layer
    - Using Adam(RMSprop) optimizer
  - Need a lot of time to train
    - 12 GB memory & GTX 1060(Can’t load data in one time)
    - about 30 hours / 100 epochs



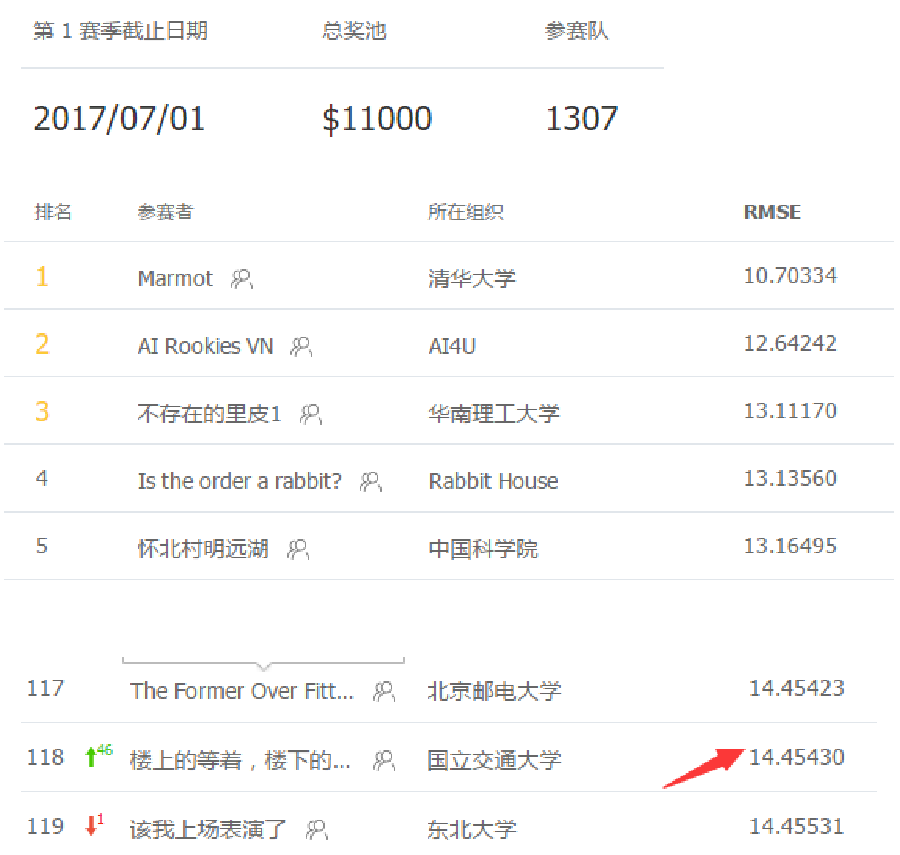
```
model.add(Flatten())
model.add(Dense(1))

# optimizers should be tested
# sgd + momentum
# others
adam = optimizers.Adam(lr=0.0035, beta_1=0.9, beta_2=0.999, epsilon=1e-08, decay=1e-6)
# sgd = optimizers.SGD(lr=0.005, momentum=0.9, decay=1e-6, nesterov=True)
# rms = optimizers.RMSprop(lr=0.0035, rho=0.9, epsilon=1e-08, decay=1e-6)
model.compile(optimizer=adam, loss='mse')
return model
```



Experimental results

- Using **RMSE** to judge
- Rank: **118/1307**
- Basic methods
  - Linear Regression(baseline): 14.90
  - SVM: about 16.8
  - Decision Tree: 17.5(worst)
- Preprocess Clustering methods:
  - Less than **14.44(still improving)**
- Deep Learning(CNN)
  - Classification: no result
  - Regression: less than **14.50 (still improving)**
  - **Tring to combine RNN**

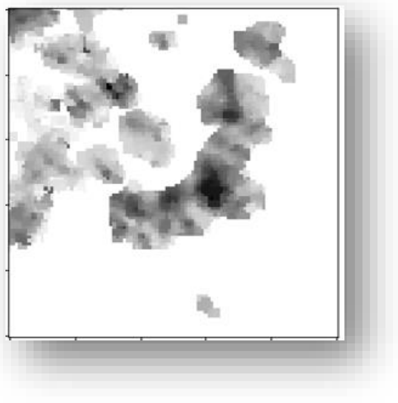


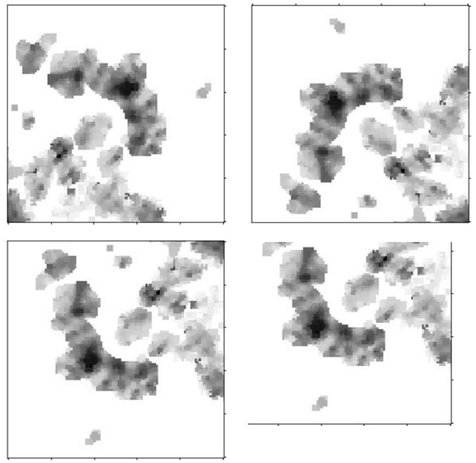
第 1 赛季截止日期	总奖池	参赛队
2017/07/01	$11000	1307

排名	参赛者	所在组织	RMSE
1	Marmot	清华大学	10.70334
2	AI Rookies VN	AI4U	12.64242
3	不存在的里皮1	华南理工大学	13.11170
4	Is the order a rabbit?	Rabbit House	13.13560
5	怀北村明远湖	中国科学院	13.16495
117	The Former Over Fitt...	北京邮电大学	14.45423
118 ↑46	楼上的等着，楼下的...	国立交通大学	14.45430
119 ↓1	该我上场表演了	东北大学	14.45531

Future work

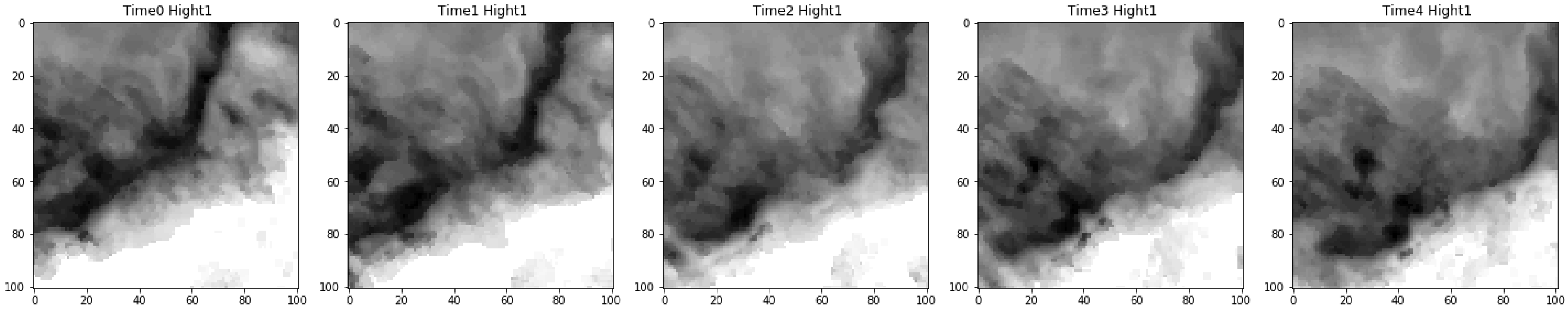
- **Generate more data**
  - Rotate or Flip
  - Data augmentation





Future work(cont.)

- Generate more data
- **RNN(LSTM or other)**
  - There are some correlations between the radar map



Future work(cont.)

- Generate more data
- RNN(LSTM or other)
- **Ensemble learning**
- **Consider better clustering methods**
- **etc.**

Future work(cont.)

- Generate more data
- RNN(LSTM or other)
- Ensemble learning
- Consider better clustering methods
- etc.

- **We are still improving**

Q & A