香 港 大 學
THE UNIVERSITY OF HONG KONG

Application of Deep Learning in Urban Hydrology

Ting Fong May CHUI, and Yang YANG

Department of Civil Engineering
The University of Hong Kong

Outline

- Introduction
  - Conventional machine learning methods VS. deep learning methods
- Case studies
  1. Identifying roads from aerial photos
  2. Inspecting what in the image leads to the prediction (identifying residential buildings from aerial photo)
  3. Predicting overflow occurrence of green stormwater infrastructures (GIs) using rainfall time series using deep learning methods
  4. Predicting outflow rate from GIs using conventional machine learning methods (for comparison)
  5. Predicting outflow rate from GIs using deep learning methods
- Conclusions and recommendations

Introduction

- **Deep learning** is a subfield of machine learning

- **Machine learning (ML)** can be defined as the science of using various mathematical models, without explicitly programming them, to learn the statistical structure from the data

- The objective of ML is to find a suitable function $\boldsymbol{f}$ that maps an input random variable $\boldsymbol{X}$ to an output random variable $\boldsymbol{Y}$ (such as a class label, or a value) that is more useful for the task of interest

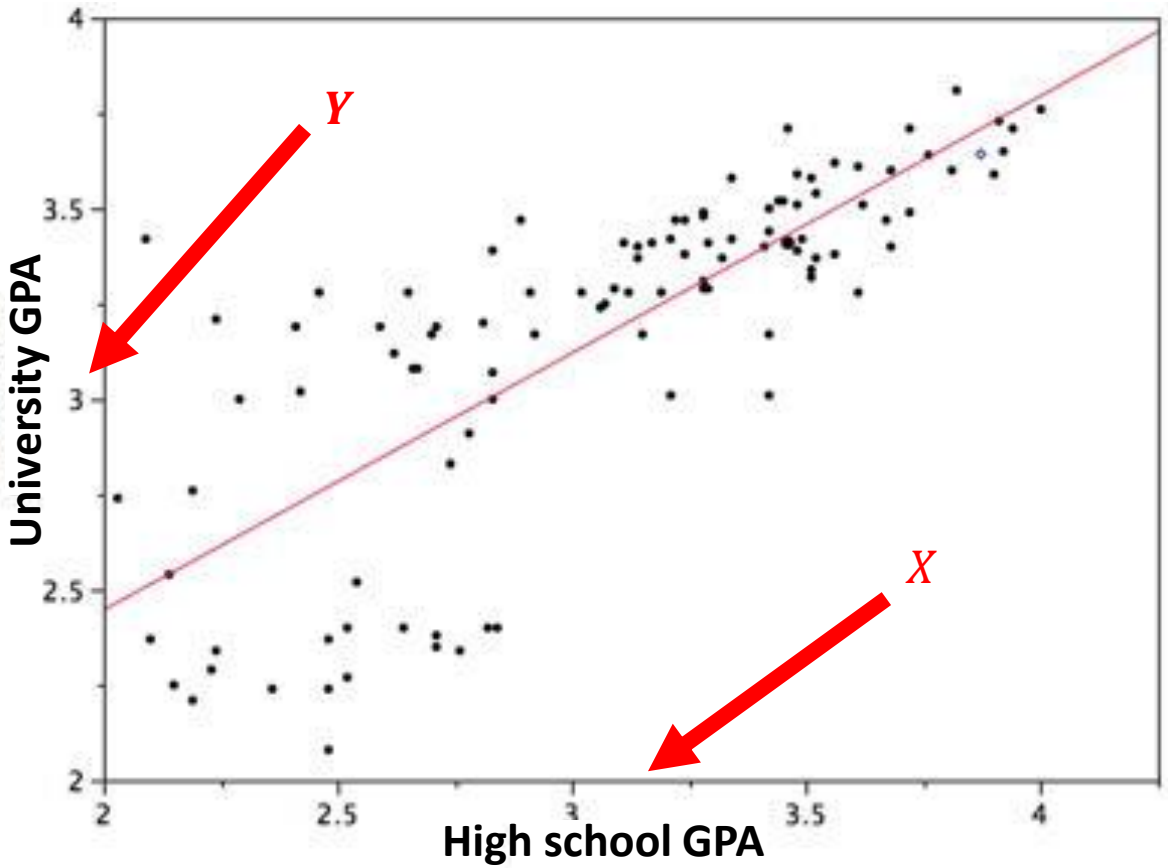
Introduction

- A machine learning model f can be written as:

$$Y = f(X; \theta)$$

  where, θ is trainable parameter

- Example: in linear regression, $\boldsymbol{X}$ is predictor, $\boldsymbol{Y}$ is outcome of interest, $\boldsymbol{f}$ is an linear operation, $\boldsymbol{\theta}$ is a vector containing the slope and the intercept



Source: OnlineStatBook

Introduction

- In many problems, the dimension of predictor is extremely high, where conventional ML models cannot be used directly
  - Images that has many pixels
  - Time series that has many time steps

- To solve this problem, the original predictor is often transformed to features using some function Ø, models are then built on the transformed predictor, $\emptyset(\mathbf{X})$

$$Y = f(\boldsymbol{X}; \theta) \rightarrow Y = g(\emptyset(\mathbf{X}); \theta')$$

High dimension **Low dimension**

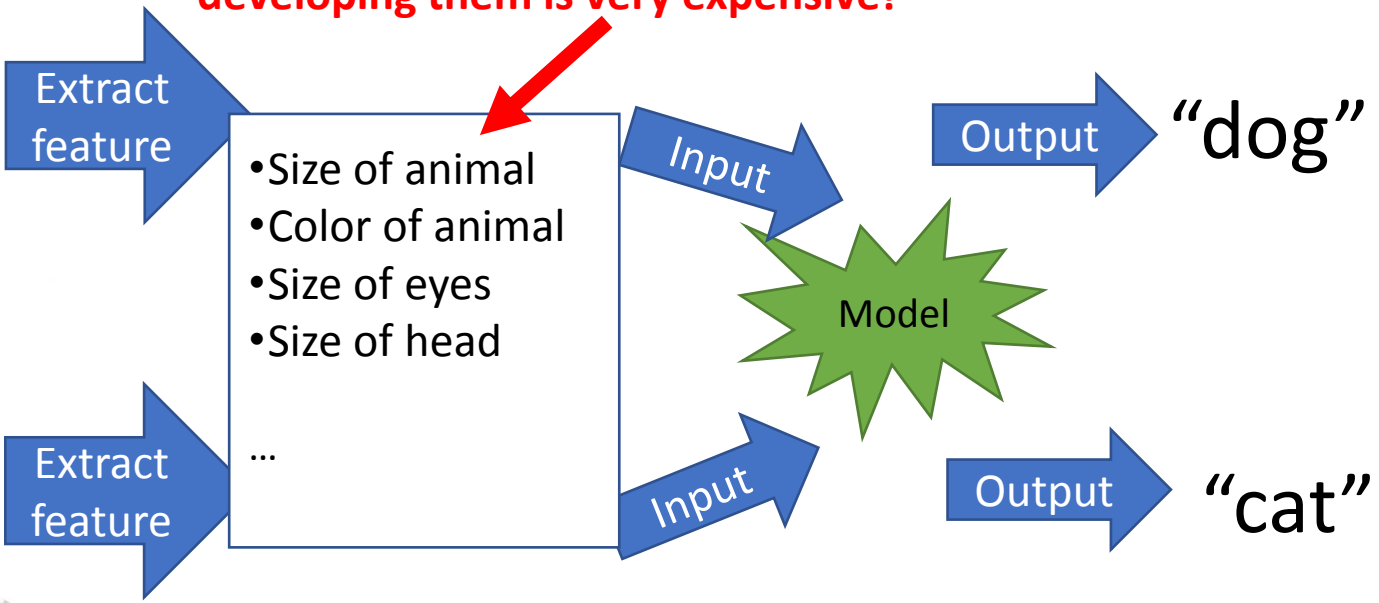
Introduction

- Example: classifying dogs and cats images
  - $\mathbf{X}$ is the raw image (which are numbers stored at each pixel)
  - $\emptyset$ transforms the raw image into features: color of the image, size of the animal, etc.
  - ML models are then built on features $\emptyset(\mathrm{X})$
  - $\emptyset$ is **very difficult to define and is often suboptimal**

Source: akc.org

Source: ccdeadelberg.be

You need specialized program (function) to extract features; developing them is very expensive!

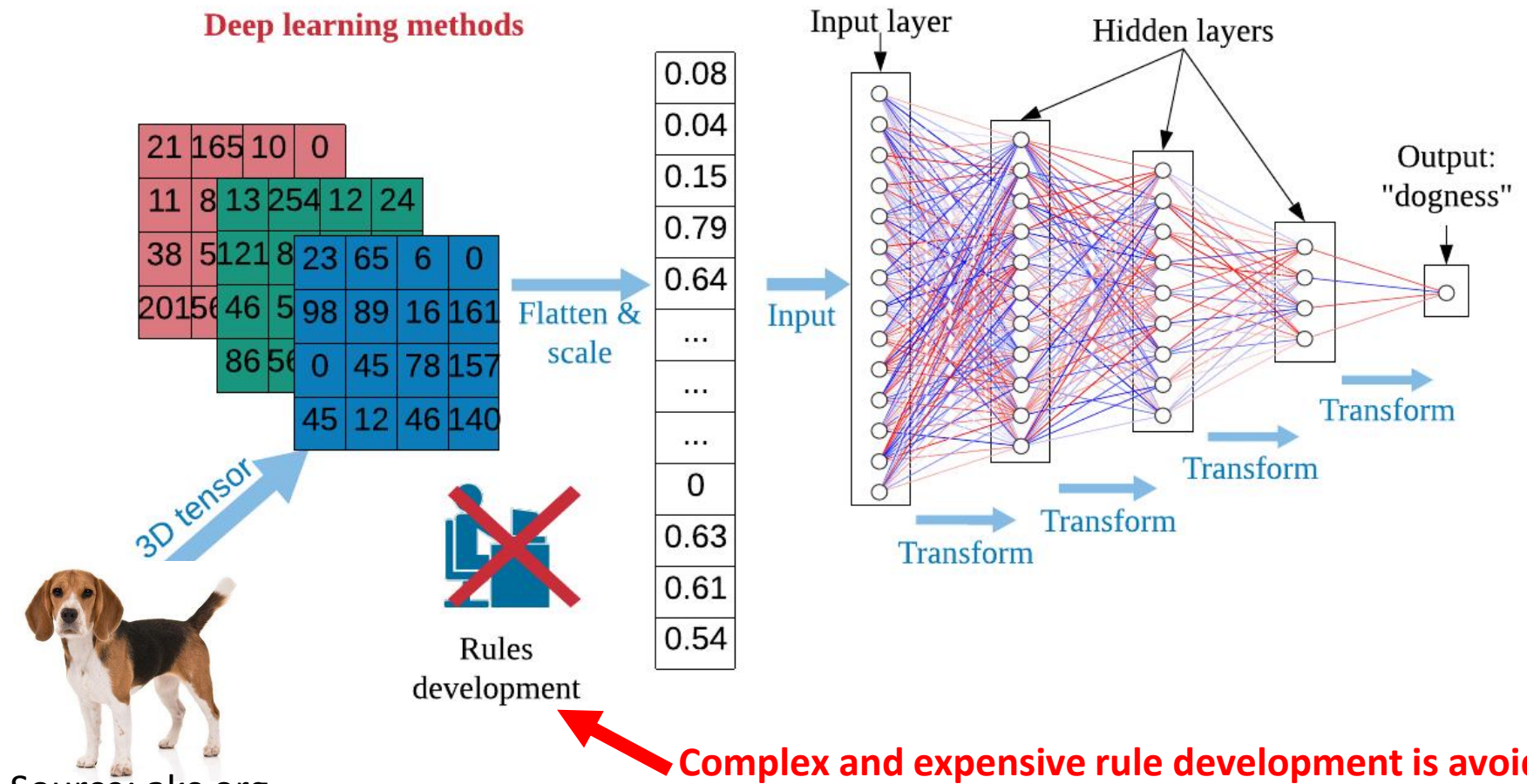


Introduction

- In deep learning models, X maps to Y through **successive transformations**

$$Y = f\left(\emptyset^{(3)}\left(\emptyset^{(2)}\left(\emptyset^{(1)}\left(\mathrm{X};\theta^{(1)}\right);\theta^{(2)}\right);\theta^{(3)}\right);\theta\right)$$

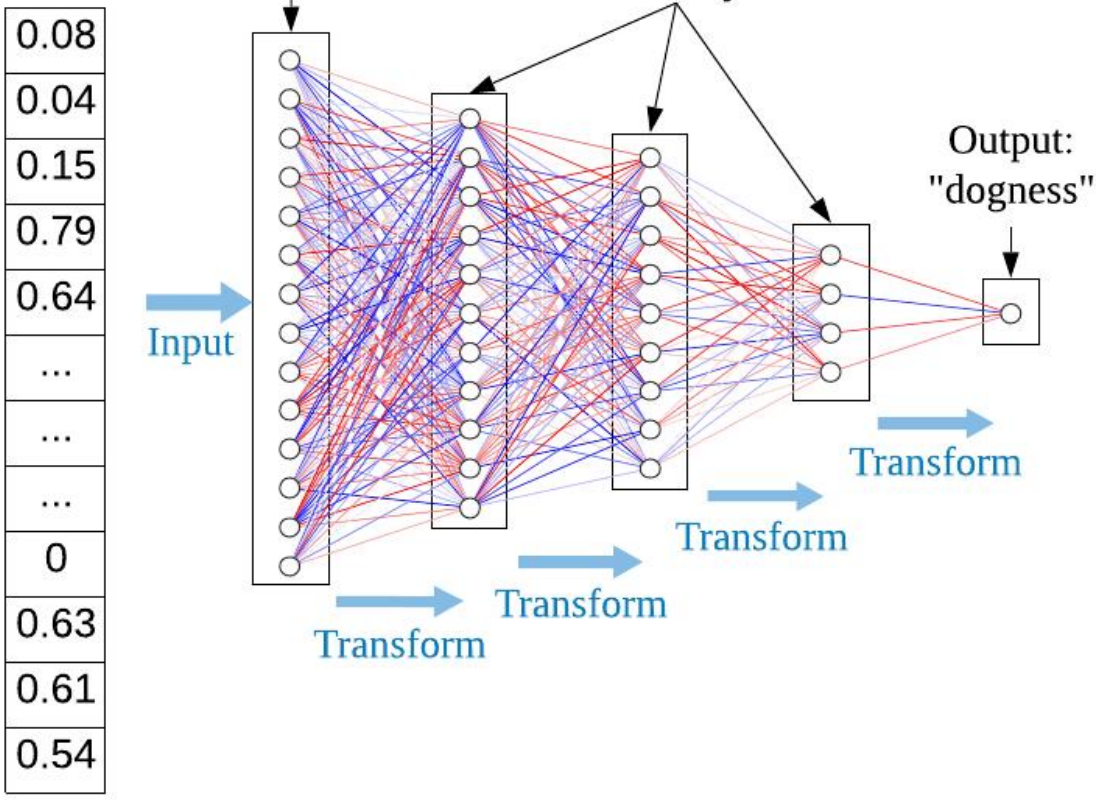
- Each transformation is simple, and can be learned from data



Source: akc.org

Introduction

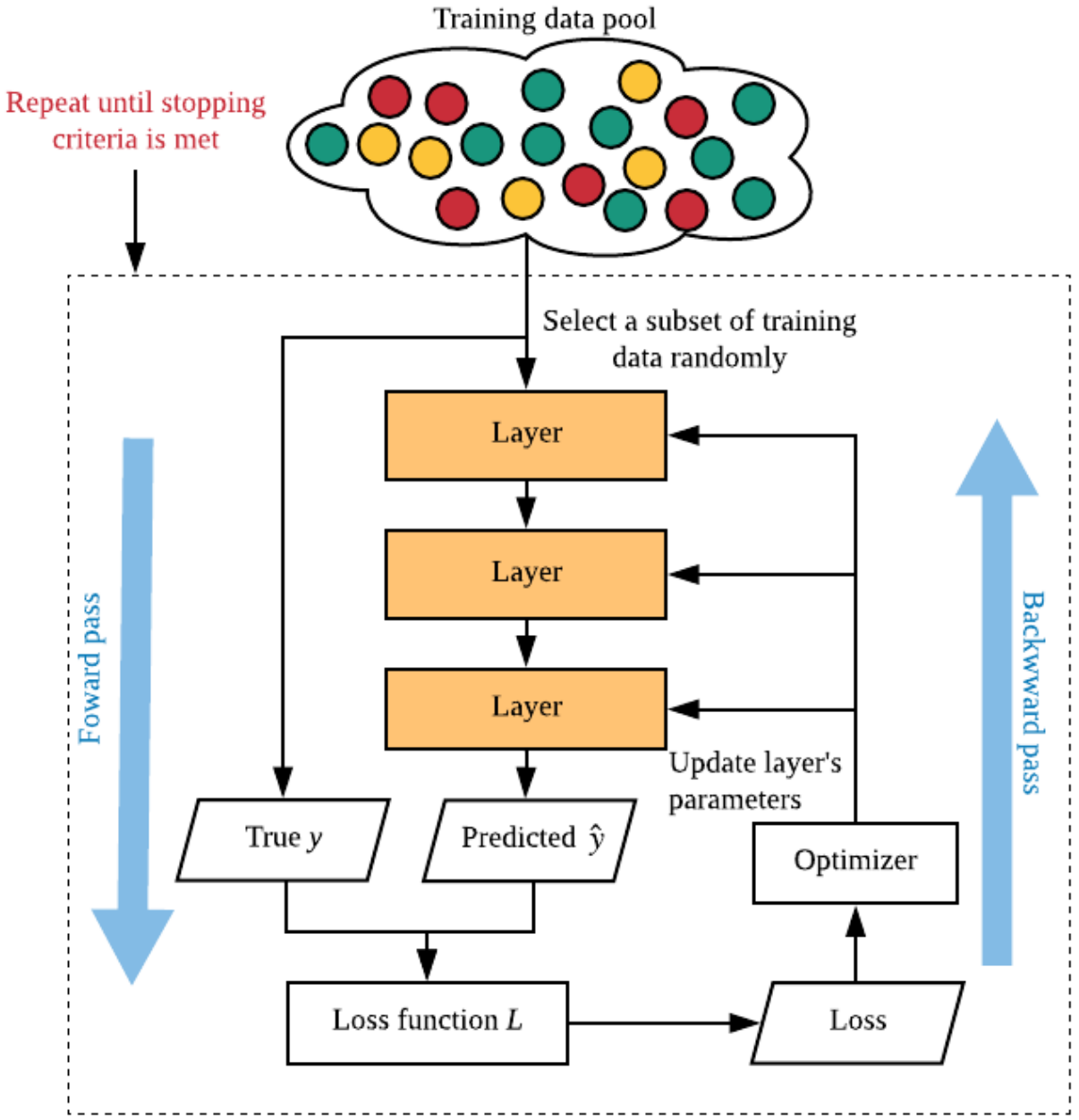
- **Artificial neural network (ANN)**, is currently predominately used in deep learning
  - connected value-processing units, i.e., neurons
- Each neuron receives one or more numerical inputs and produces a numerical output
- A bias is associated with each neuron
- The weights (connections between neurons) specify how much a neuron influences the receiving neuron
- The learning process is to find the weights and biases using training data



Introduction

- Parameters of the networks can be found using the **backpropagation** method and the **gradient descent** optimization algorithm

- Some software and programming toolbox are developed to automate the training process
  - Tensorflow by Google
  - MXNet by Amazon
  - …

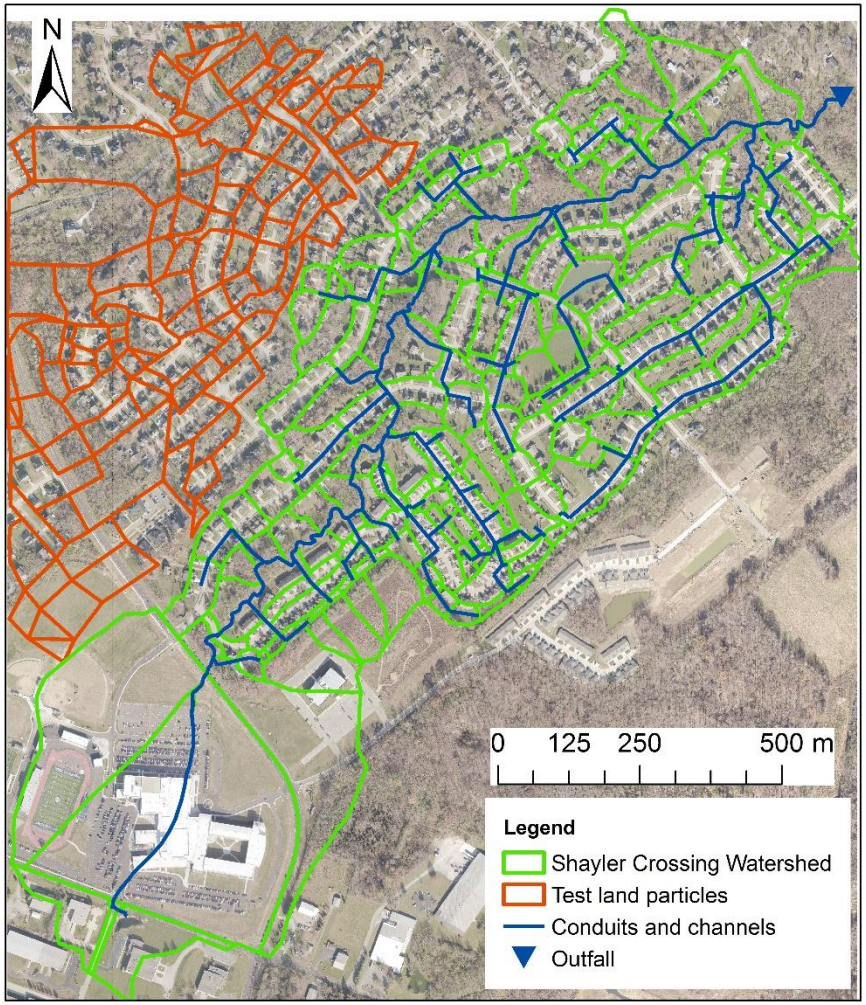


Case study 1: identifying images with roads

- Land cover identification is a common task in urban hydrology
  - Useful for determining the hydrological properties, land use, etc.
  - Setting up green infrastructure implementation scenarios

- Currently, this task is commonly performed by human or specialized software
  - Time-consuming and expensive

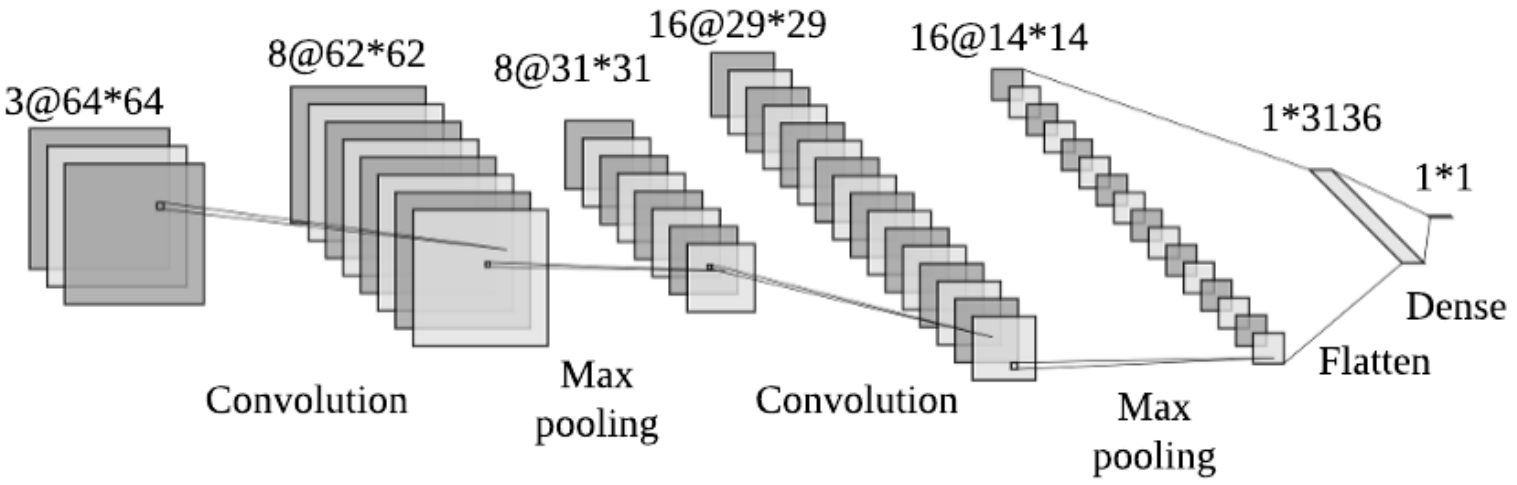
- Objective: build a deep learning model for identifying roads in catchment aerial images
  - i.e., train a model to determine whether highways are presented in an image using **training images with only "highway"/"no highway" labels**



Shayler Crossing Watershed, Ohio, U.S.

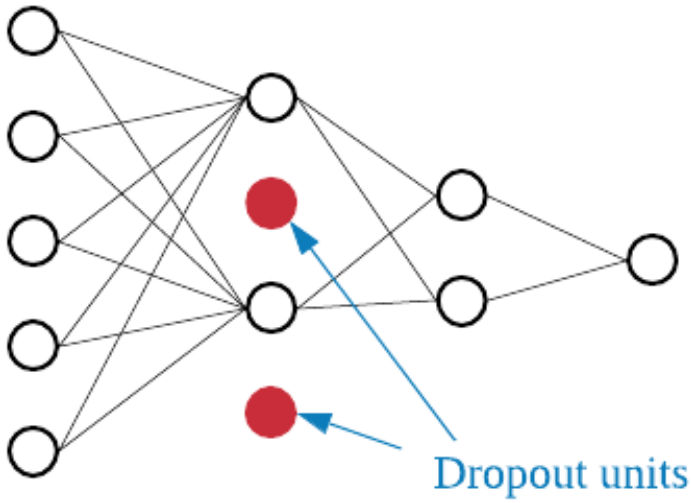
Case study 1: identifying images with roads

- **Convolutional neural networks (CNN)** are used in this study to solve image classification problems
- CNN uses a special type of layers, i.e., the convolutional layers
  - Each convolutional layer arranges its output into a 3D tensor of shape $(height, width, channel)$ to preserve the spatial information of the layer's input
- Each element in the output tensor of a convolutional layer is connected to the elements in a local region of the input tensor
  - Sparse connections significantly reduce the number of parameters



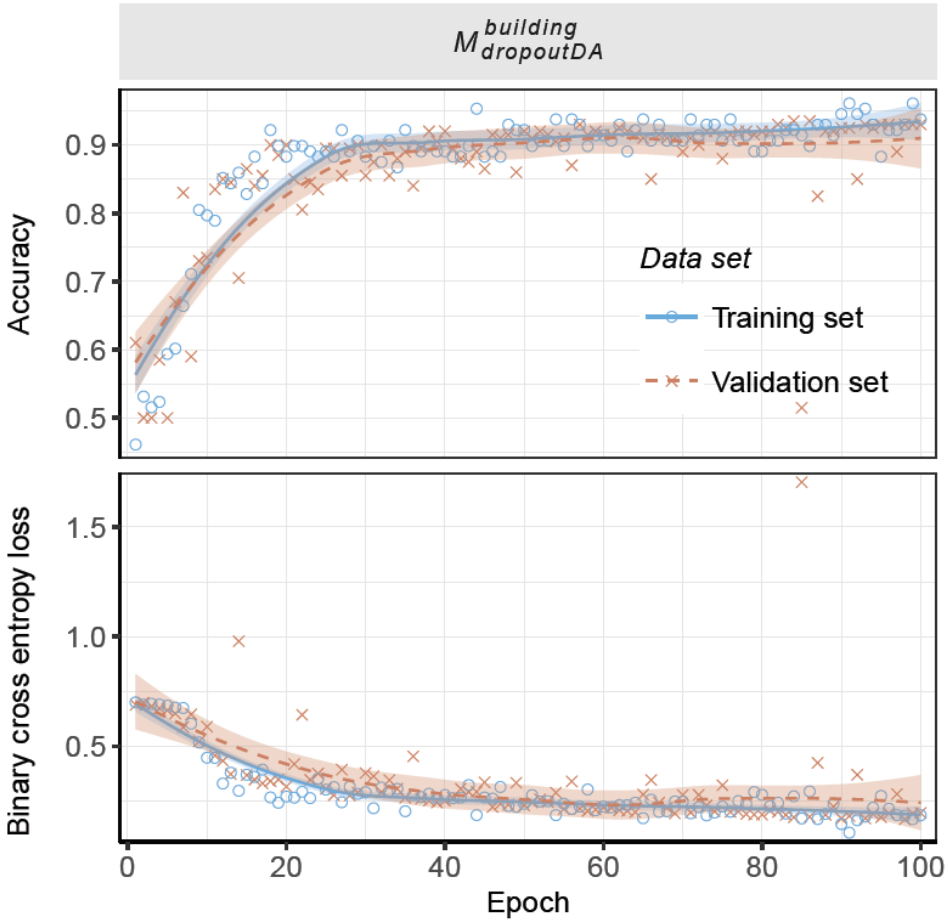
Case study 1: identifying images with roads

- Reduce overfitting using dropout in training
  - When dropout is applied to a layer, in each iteration of gradient update in training, a fraction of the nodes will be randomly selected and set to 0, i.e., the connections between those nodes and the other nodes are cut off
  - Using dropout, models may be forced to ignore noisy and irrelevant patterns in the trained set



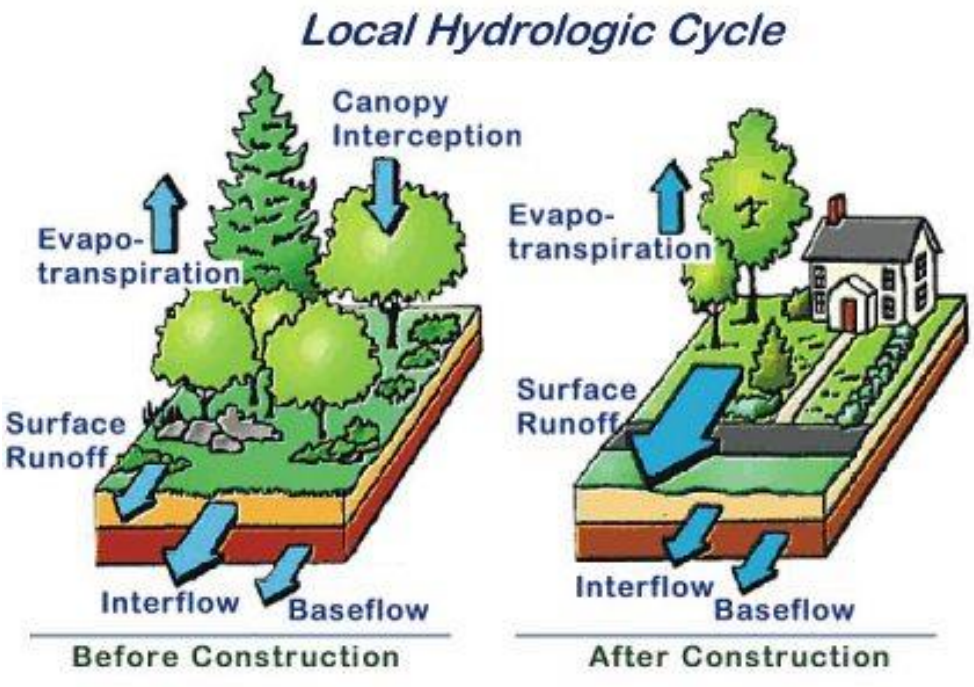
Case study 2: making sense what is learned by the model

- A deep learning model is built to identify residential buildings
  - Model accuracy > 93% is achieved



Model green infrastructure using deep learning methods

- Urbanization changes the land cover, which leads to more impervious area, less vegetation, etc.
  - More surface runoff, less groundwater recharge, more pollution, etc.
- **Green infrastructure (GI)** refers to systems and practices that use or mimic natural processes that result in the infiltration, evapotranspiration or use of stormwater to protect water quality and associated aquatic habitat (U.S. EPA)
- Process-based models are currently widely-used



Source: jacksonms.gov

Model green infrastructure using deep learning methods

- Green infrastructures (GIs) can vary significantly in:
  - Type of practices
  - Location in urban catchment, and scale of implementation
  - Design: material type, material depth, scale, drainage area…
- How to account for the large variabilities in modeling?
- **Does a universal model exist?**

Green roofs in Wuhan, China

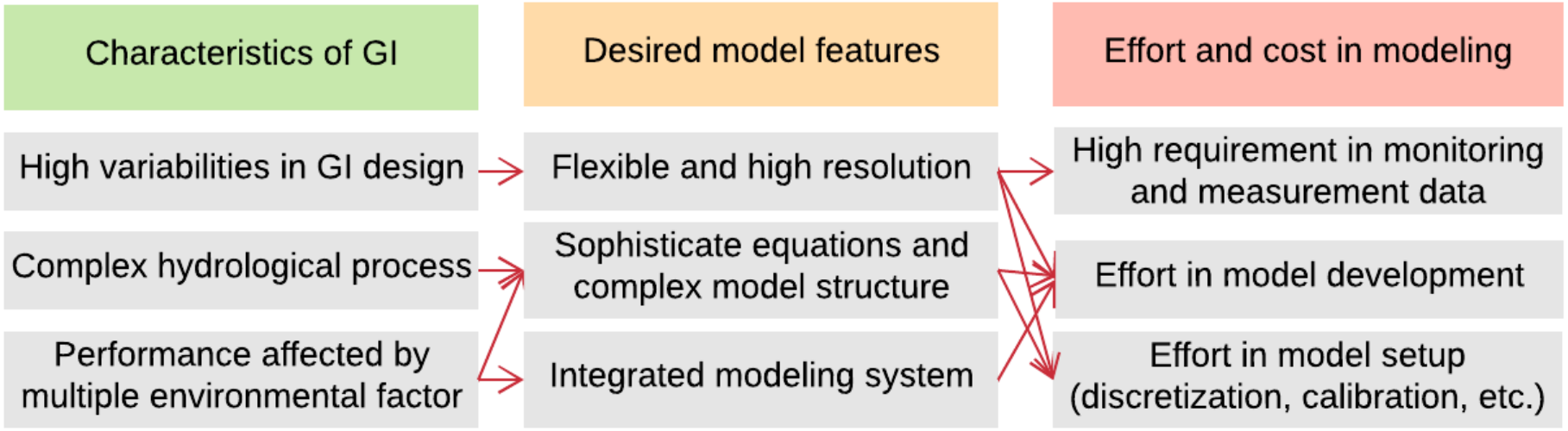
Porous pavement in Hong Kong, China

Right-of-way (ROW) bioswales in New York City

Photos are taken in Wuhan, China, 2018, Hong Kong, China, 2018, and New York City, U.S. 2016

Model green infrastructure using deep learning methods

- Performances of GIs are affected by their highly variable designs, complex hydrological processes and the external environment
- To account for the characteristics of GIs, a desired process-based model should be flexible, high-resolution, and able to be integrated with other hydro-environmental models
- Developing and using such complex model, however, require lots of efforts in model development, data collection, and model setup
  - In practice, simpler and more specialized models are commonly used

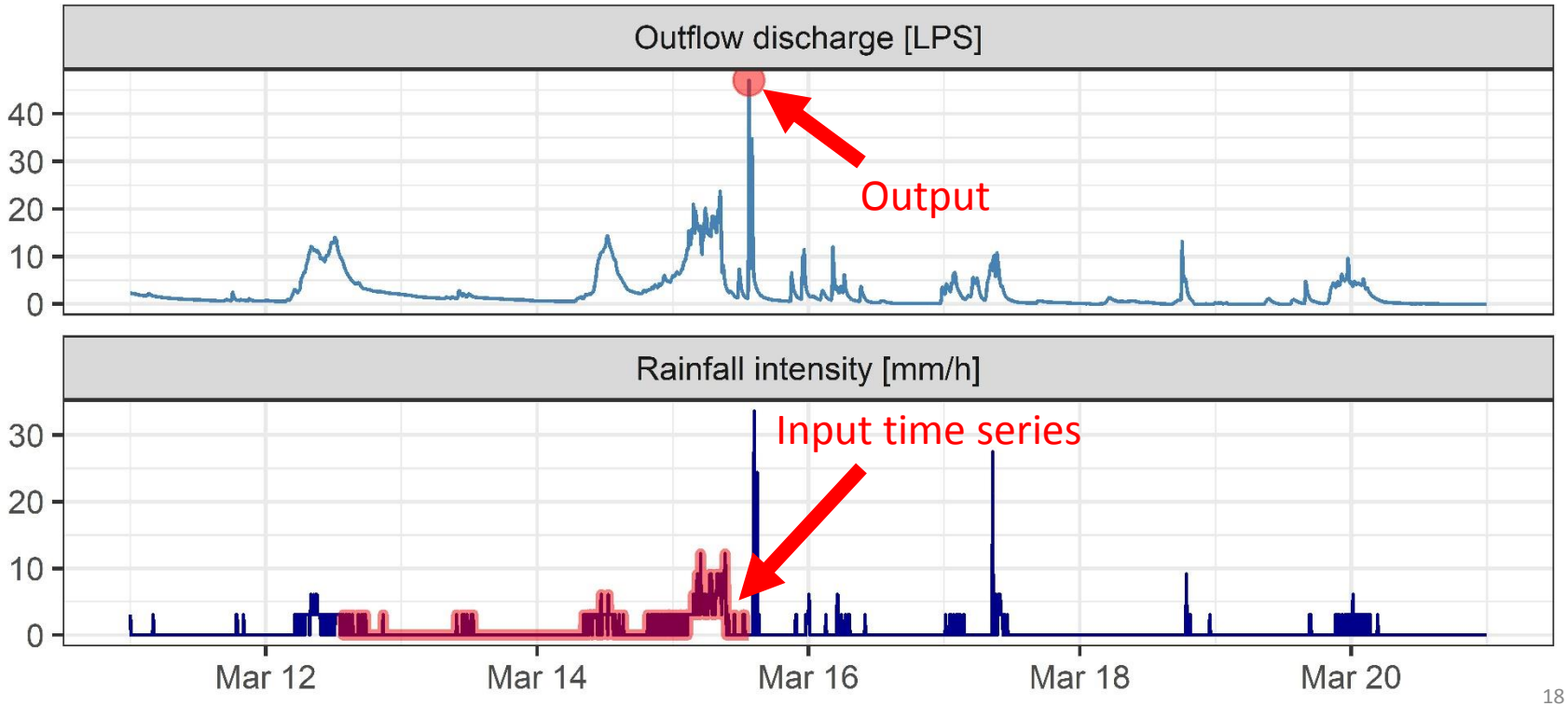


Model green infrastructure using deep learning methods

- **ML** methods can use observed data of the studied system to find the connections between the system state variables; e.g.,
  - The connections between the *instantaneous outflow rate* and the *preceding rainfall time series*
  - The connections between *water level in bioretention cell* and *rainfall and evaporation time series*

Model green infrastructure using deep learning methods

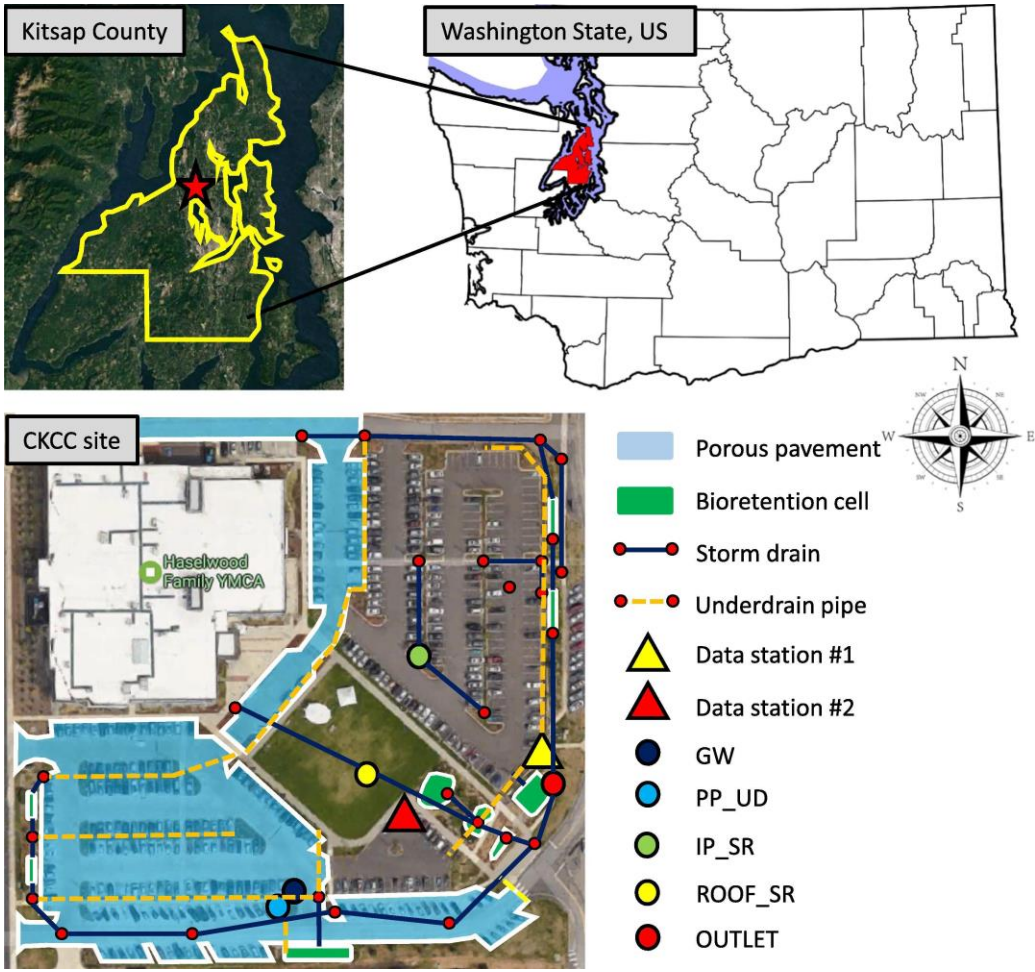
- An example:
  - **Output variable**: instantaneous outflow rate from a GI
  - **Input variables**: the preceding 3-day rainfall time series
  - **Model**: a function that maps the input rainfall time series to an output discharge value
  - **Training, validation and testing data**: rainfall and runoff observations



Case study 3 Predicting overflow occurrence using high-resolution rainfall time series

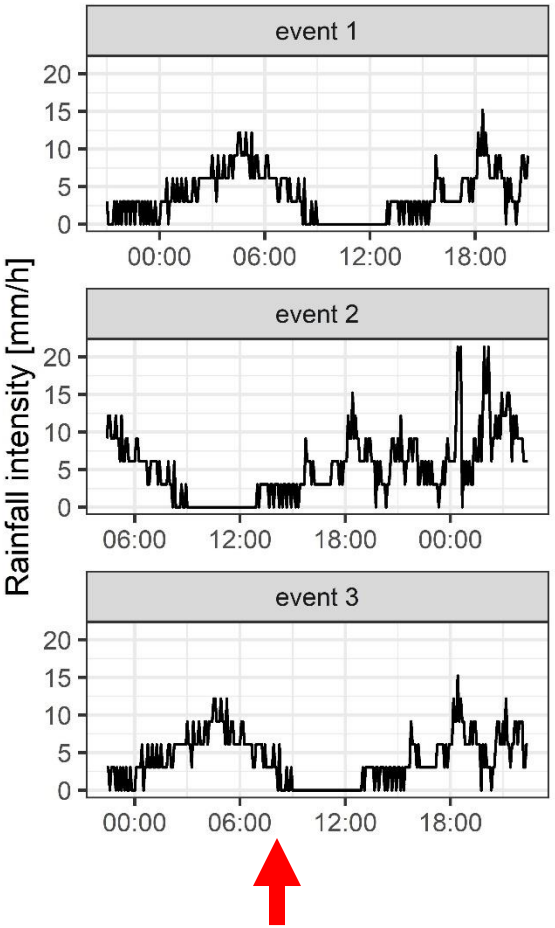
- Central Kitsap County Campus (CKCC) bioretention and porous pavement site in the U.S.
  - High temporal and spatial resolution (5-min and meters)
  - Multiple small scale GI practices of different types
  - Performances are potentially affected by shallow groundwater
  - Binary observation: overflow/no overflow in bioretention cell
  - Lack of field measurement (e.g., infiltration rate, soil field capacity)

- Building a process-based hydrological model for this site requires too much effort!

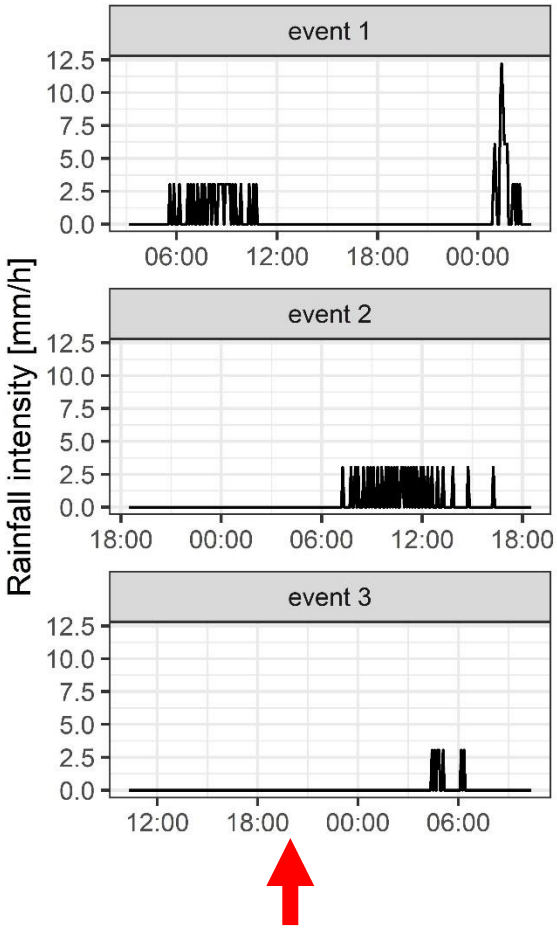


Source: Zhang et al., 2018; DOI:10.1016/j.jhydrol.2018.09.006

Case study 3 Predicting overflow occurrence using high-resolution rainfall time series



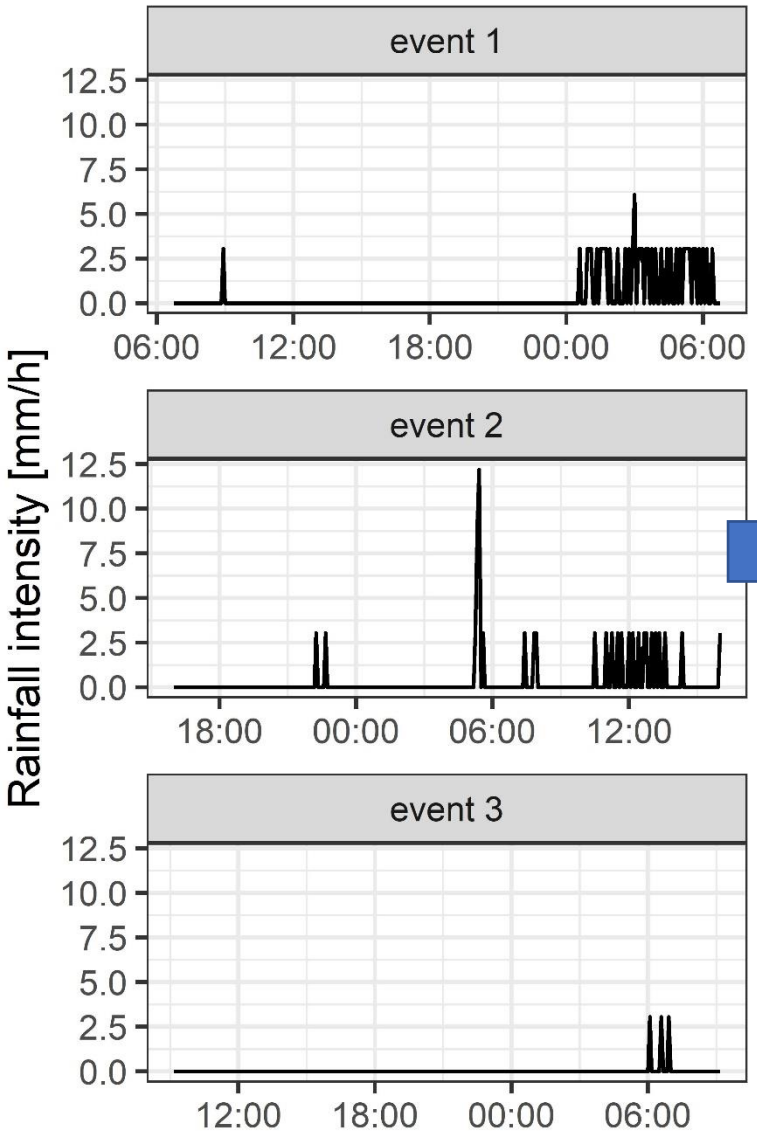
Rainfall time series registered as **overflow**



Rainfall time series registered as **no overflow**

- Predict if overflow occurs in a bioretention cell using 1-day 5-min resolution rainfall time series
- Neural networks directly transform the rainfall time series (i.e., a series of numbers) to a label “overflow” or “no overflow”
  - i.e., determine whether a time series causes overflow or not, i.e., a classification problem

Case study 3 Predicting overflow occurrence using high-resolution rainfall time series



Input

Model

Output

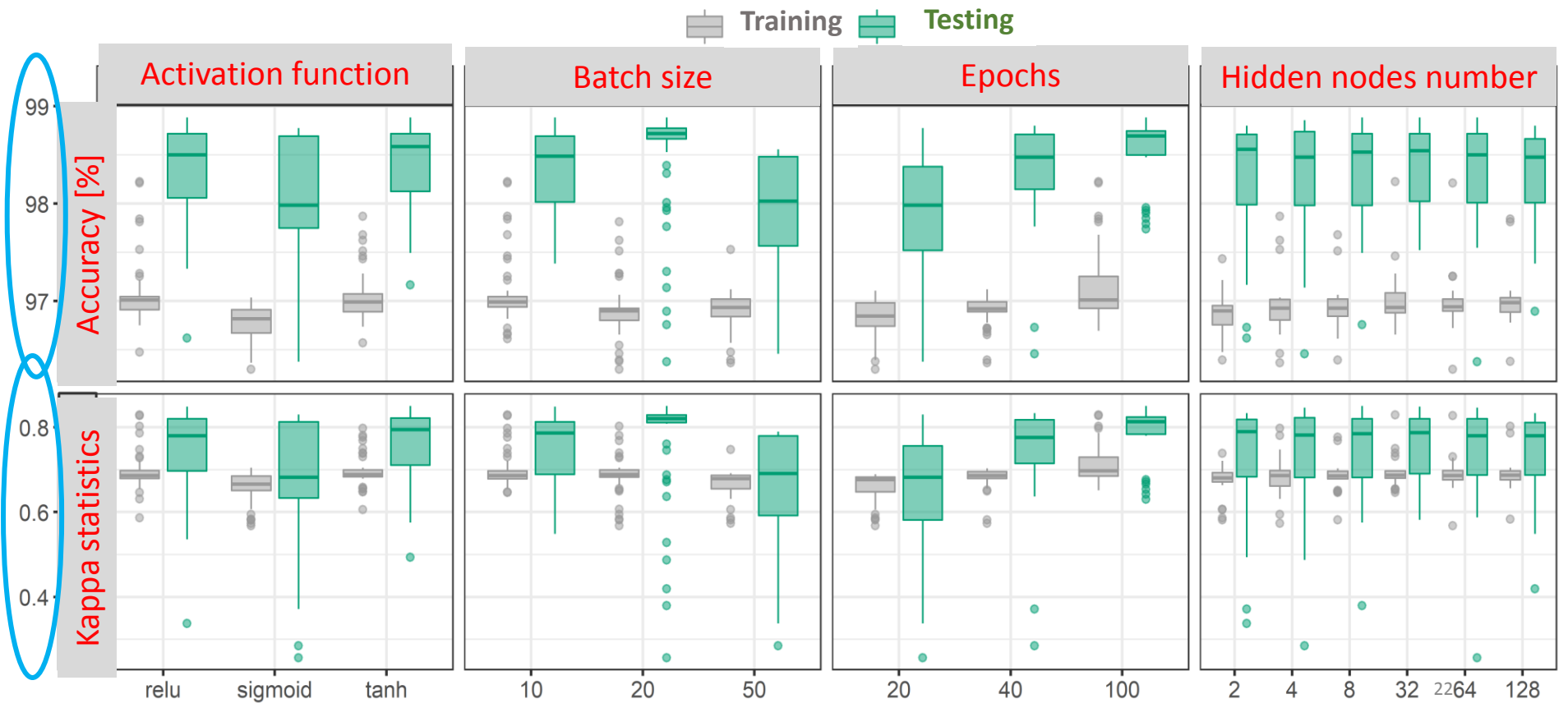
0.15478
0.20021
0.01415

Trained using training data

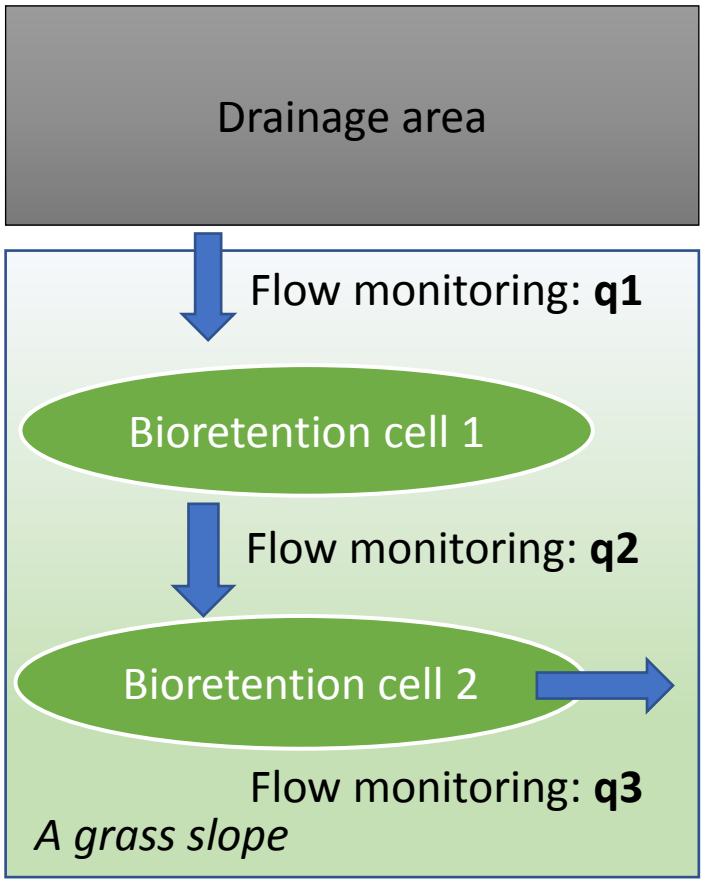
Likelihood of overflow

Case study 3 Predicting overflow occurrence using high-resolution rainfall time series

- 162 neural networks with different numbers of hidden nodes, types of activation function, batch sizes were trained for different epochs
- More than **96% accuracy** for both the training and testing sets; good kappa statistics
  - Structure (architecture) of network affects the model performance
  - Neural network design currently heavily relies on the trial-and-error approach



Case study 4 Predicting outflow rate using conventional ML methods (for comparison)

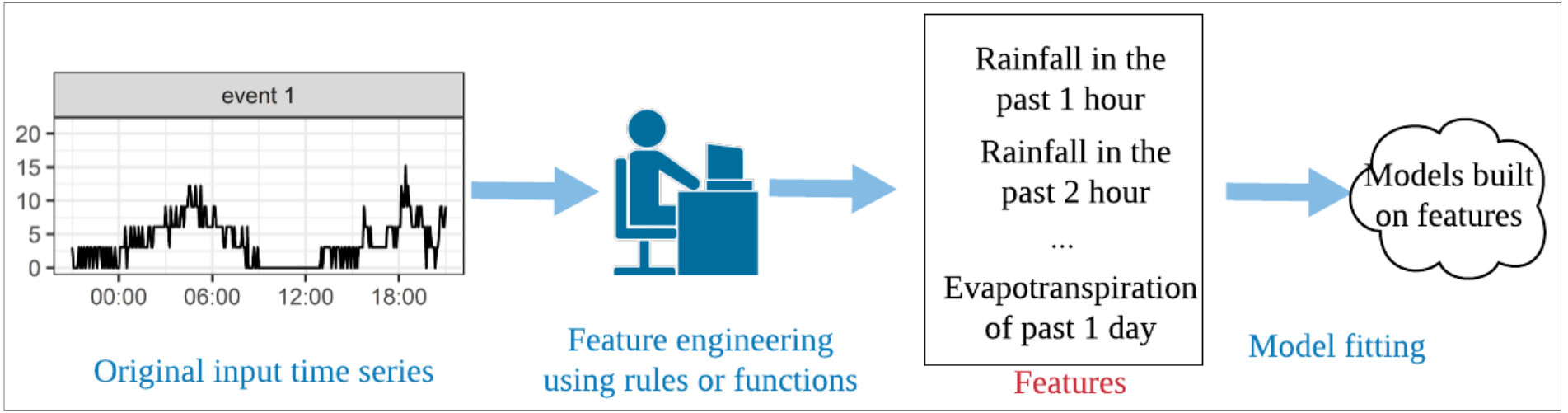


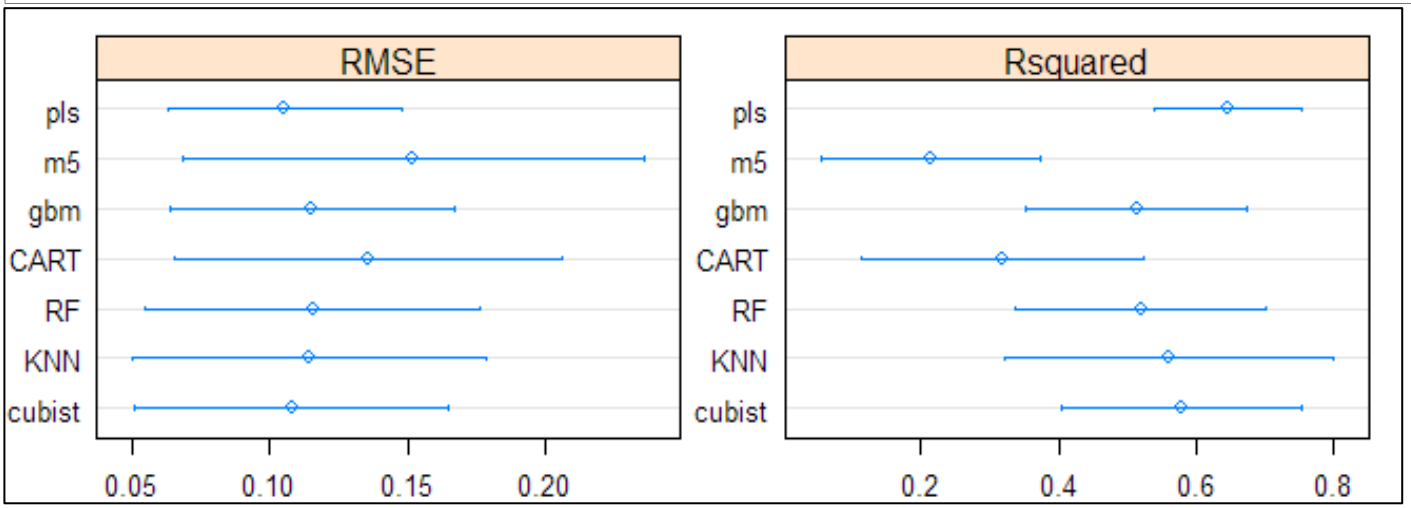
Layout of a bioretention cell system in Cincinnati, U.S.

- A bioretention cell system in St. Francis Apartments, Cincinnati, U.S. is studied
  - High resolution (2-min) inflow and outflow monitoring
  - The size of the two bioretention cells are relatively large, and the surface layers are not flat: lumped layer representation is NOT adequate
- To **predict q2 using q1**, multiple conventional ML models (PLS, M5 model trees, random forests, cubist, etc.) are used, and the model performance are tested using cross-validation

Case study 4 Predicting outflow rate using conventional ML methods (for comparison)

- The original rainfall and evapotranspiration time series are aggerated into **a set of features** because the dimension of input is high (**this is the most difficult part of ML**)
- ML models are built on features (NOT the raw data)

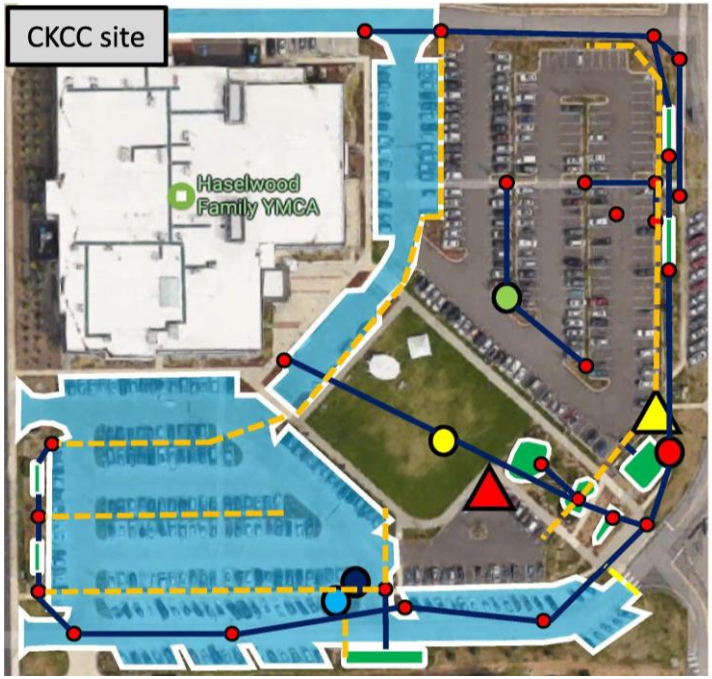




Models performed well; cross-validation $R^2 > 0.7$ and RMSE is small

Case study 5 Predicting outflow rate using deep learning methods (for comparison)

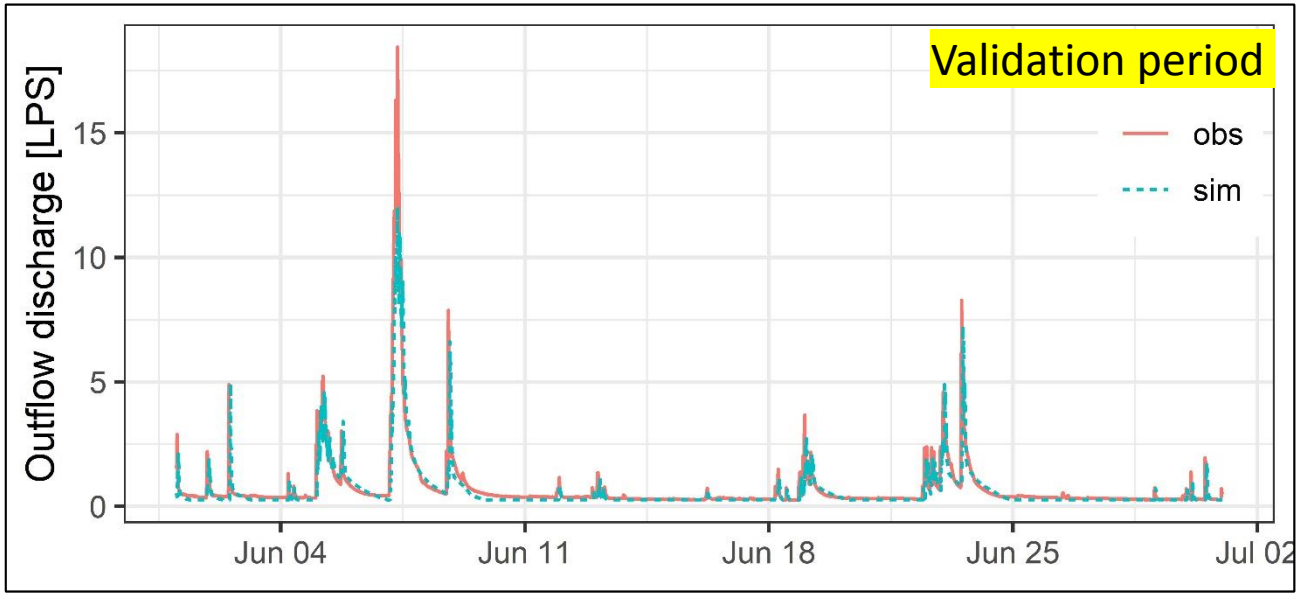
- Predict **instantaneous outflow** rate at the main outlet using **the preceding rainfall time series** at CKCC site
- Long short-term memory (**LSTM**) networks are used to directly project **the raw input time series** into a numerical value
  - LSTM maintains some hidden states that are related to rainfall events in the past
  - LSTM learns the long-term dependency between rainfall in the past and the runoff



Source: Zhang et al., 2018;
DOI:10.1016/j.jhydrol.2018.09.006

Case study 5 Predicting outflow rate using deep learning methods (for comparison)

- The LSTM model performs very well
  - Training set: **NSE = 0.79, R^2 = 0.79**
  - Test set: **NSE = 0.78, R^2 = 0.77**
  - Better than the calibrated SWMM model we built



Conclusions

- Deep learning methods greatly reduce the effort in feature engineering, and are especially useful when there are multiple input variables or the dimension of input is high

- Machine learning methods (both conventional and deep learning) are recommended for catchment where process-based models are unavailable or restricted due to limited field measurements or censored data

- While building a deep learning model that produces reasonably good results may not be difficult, making the best use of data or understanding the model's generalization performance is

- Deep learning is not magic, it is only useful if there is enough data and the model is correctly configured. Even if the model can generate reasonably good results, it might not have behaved in expected ways

Recommendations

- We recommend future research to present more applications of deep learning in hydrological studies
  - Reporting and archiving the model structures and the model performance testing procedures
  - Applying the state-of-art machine learning and deep learning tools, model architectures and training methods
  - Look into what is learned by the models and use deep learning as tool to discover new knowledge of hydrology

Acknowledgement

- This study was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU17255516). We would like to thank Robert A. Darner from the U.S. Geological Survey for providing site information on the St. Francis Apartments site.

Thank you!

Questions and comments send to:

Dr. Ting Fong May Chui: maychui@hku.hk

Yang Yang: yyang90@connect.hku.hk