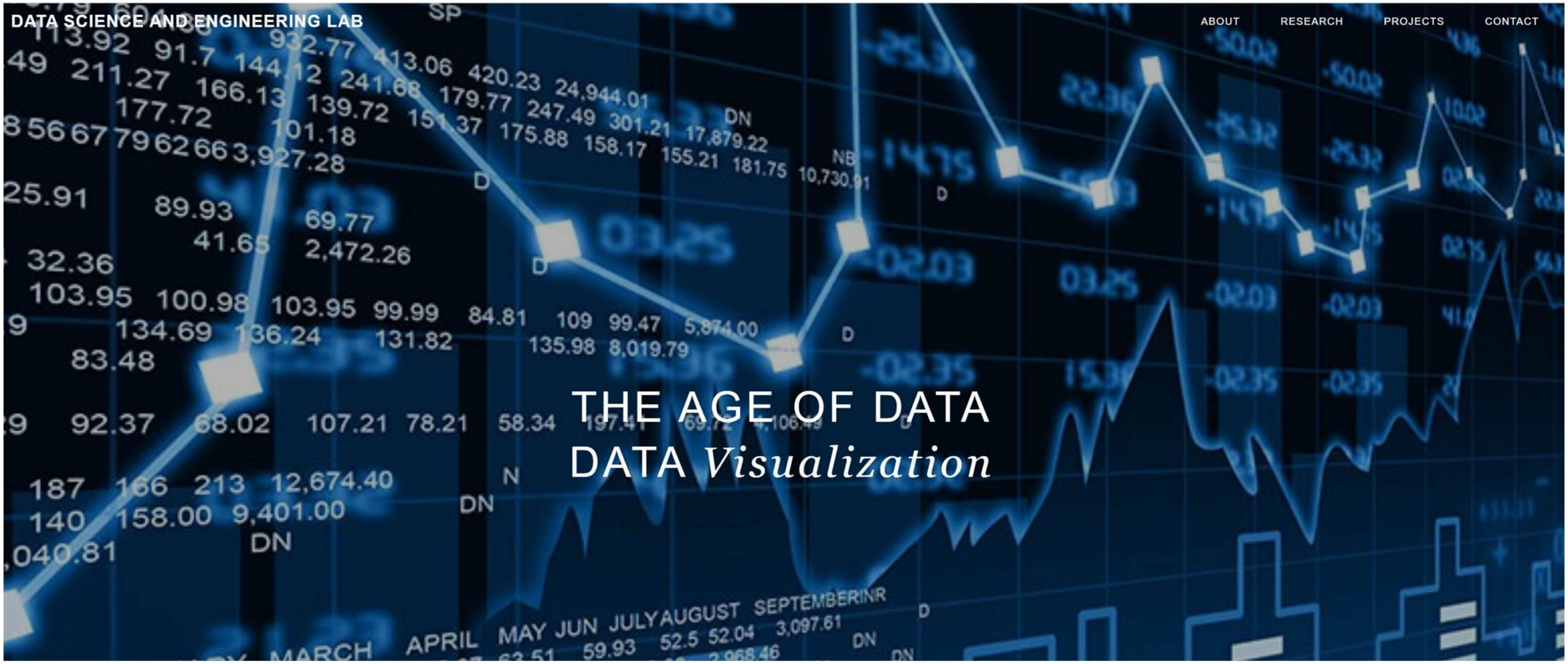
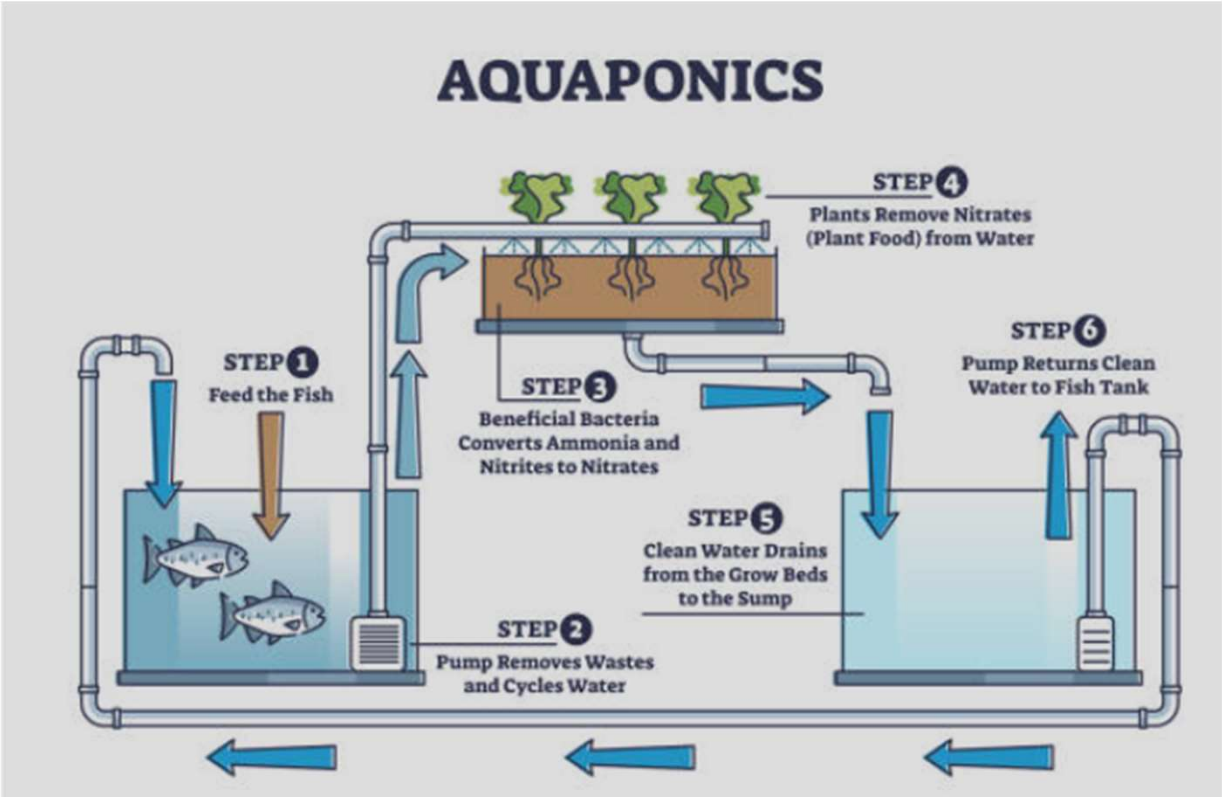
Compact Transformer Filter

DSE Research Lab
https://csit.udc.edu/~dse/

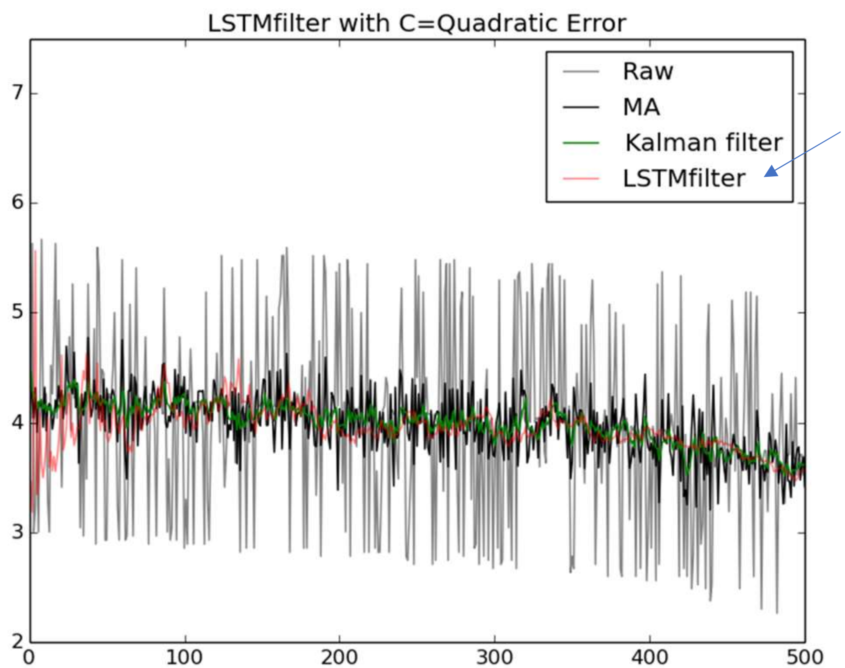
Data Science and Engineering Research Lab
CSIT at UDC
https://csit.udc.edu/~dse/

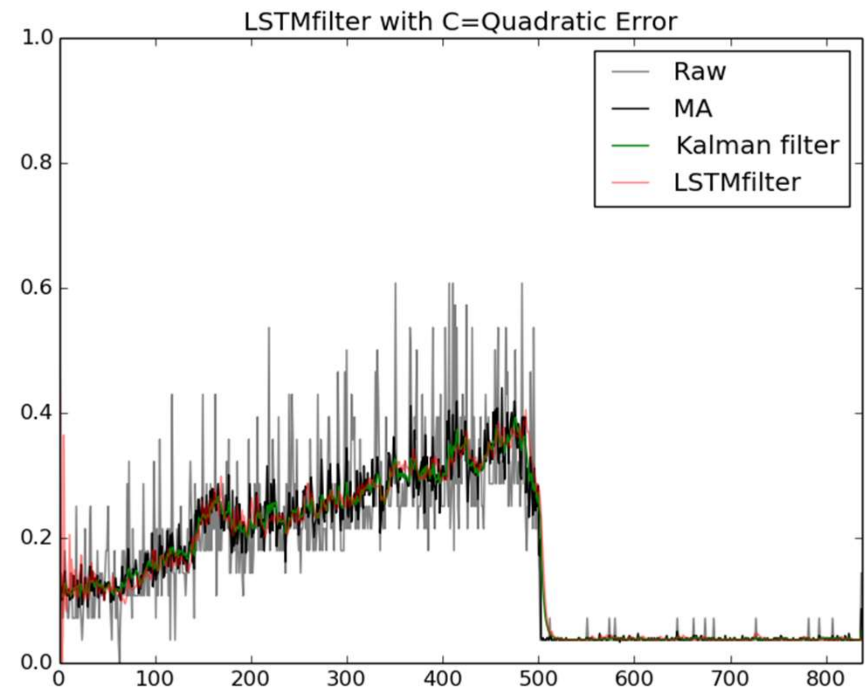



AQUAPONICS
STEP 1
Feed the Fish
STEP 2
Pump Removes Wastes and Cycles Water
STEP 3
Beneficial Bacteria Converts Ammonia and Nitrites to Nitrates
STEP 4
Plants Remove Nitrates (Plant Food) from Water
STEP 5
Clean Water Drains from the Grow Beds to the Sump
STEP 6
Pump Returns Clean Water to Fish Tank

How to collect agriculture data?

LSTM Filter evaluated with Collected Data





Vision Transformers

Single-head self attention

- Self-Attention in CNNs
  - No-local attention (Non-Local Neural Networks (thecvf.com)): designed for image deno ising, capture interactions b/w any two positions in the feature map
  - Criss-cross attention([1811.11721] CCNet: Criss-Cross Attention for Semantic Segmentat ion (arxiv.org)): reduce computational burden, each pixel position can capture context f rom all other pixels.
  - Local relation nets attention([1904.11491] Local Relation Networks for Image Recogniti on (arxiv.org)): a new differentiable layer adapts its weight aggregation based on the c ompositional relations between pixels and features.
  - Attention Augmented CNN ([1904.09925] Attention Augmented Convolutional Network s (arxiv.org)): use the relative position encoding in two dimensions for a new attention
- Self-Attention as Stand-alone Primitive
  - Stand-alone self attention ([1906.05909] Stand-Alone Self-Attention in Vision Models (a rxiv.org)): All pixel positions in a specific window size around a given pixel.
  - Vector attention ([2004.13621] Exploring Self-attention for Image Recognition (arxiv.org)) –learns weights for both the spatial and channel dimensions. Keep view relationships of a particular feature with its neighbors. can beat *ResNet*

Vision Transformers - ViTs

- [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)
- Multi-head attention
- Altogether replace standard convolutions in deep neural networks
- Applied Transformer on a sequence of image patches flattened as vectors.
- Not give competitive results on a medium dataset b/c the CNNs encode prior knowledge about the images – *inductive biases*.
- Compared to the iGPT (Image GPT)

Multi-head Self Attention (Vision Transformers - ViTs) – 1/3

- Uniform-scale ViTs (Consistent Scale in the input image) – cannot capture fine spatial details in different scale
  - Data Efficient Image Transformer – DeiT ([2012.12877] Training data-efficient image transformers & distillation through attention (arxiv.org)): a novel native distillation approach for Transformers
  - Token to Token ViT ([2101.11986] Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet (arxiv.org)): Combines neighboring tokens into a single token to reduce tokens length and aggregate spatial context.
  - Transformer in Transformer ([2103.00112] Transformer in Transformer (arxiv.org)): attention at two level: patch-level and local sub-patch level
  - Cross-Covariance Image Transformers (XCiT: Cross-Covariance Image Transformers | OpenReview): attention across feature-channels instead of tokens
  - Deep ViT ([2103.11886] DeepViT: Towards Deeper Vision Transformer (arxiv.org): reattend the attention maps in a multiple head block

Multi-head Self Attention (Vision Transformers - ViTs) – 2/3

- Multi-scale ViTs (2104.11227.pdf (arxiv.org)): Several Channel-resolution scale stages
  - Pyramid ViT – PVT ([2102.12122] Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions (arxiv.org)): a progressive shrinking pyramid and spatial-reduction attention.
  - SegFormer ([2105.15203] SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers (arxiv.org)): In PVTv2 and SegFormer, overlapping patch embedding, depth-wise convolution, and efficient attention
  - Swin Transformer ([2103.14030] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (arxiv.org)): Partition the window into multiple sub-patches to captue interactions b/w different windows (image locations)
  - CrossFormer ([2108.00154] CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention (arxiv.org)): focal self-attention to capture global and local relationships.
  - Focal Transformer ([2107.00641] Focal Self-attention for Local-Global Interactions in Vision Transformers (arxiv.org)): simultaneously capture global and local relationships

Multi-head Self Attention (Vision Transformers - ViTs) – 3/3

- Hybrid ViTs with Convolutions
  - Conv. Vision Transformer (CvT) ([2103.15808] CvT: Introducing Convolutions to Vision Transformers (arxiv.org)): Conv. Based projection to capture the spatial structure and low-level details for tokenization of image patches.
  - Comact Conv. Transformer (CCT) ([2104.05704] Escaping the Big Data Paradigm with Compact Transformers (arxiv.org)): a new sequence pooling scheme and incorporate Conv. blocks on small dataset
  - Local ViT ([2106.04263] On the Connection between Local Attention and Dynamic Depth-wise Convolution (arxiv.org)): Depth-wise conv. to enhance local features modeling capability of ViTs.
  - LeViT ([2104.01136] LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference (arxiv.org)): four-layered CNN block at the beginning with progressively increasing channels.
  - ResT ([2105.13677] ResT: An Efficient Transformer for Visual Recognition (arxiv.org)): Depth-wise conv. and adaptive position encoding -- arbitrary size of input images without interpolation or fine-tune, patch embedding as a stack of overlapping Conv.
  - NesT ([2105.12723] Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding (arxiv.org)): Local self attention on patches within each block and then enables global interaction between blocks, NesT on smaller datasets.
  - CeiT ([2103.11816] Incorporating Convolution Designs into Visual Transformers (arxiv.org)):
  - CoAtNets / Coat / Twins on PVT / TransCNN, etc

Multi-head Self Attention (Vision Transformers - ViTs) – 3/3

- Self-Supervised ViTs
  - DINO ([2104.14294] Emerging Properties in Self-Supervised Vision Transformers (arxiv.org)):
  - MoCo v3 ([1911.05722] Momentum Contrast for Unsupervised Visual Representation Learning (arxiv.org))
  - EsViT ([2106.09785] Efficient Self-supervised Vision Transformers for Representation Learning (arxiv.org))

CNN and/or Transformers for Object Detection

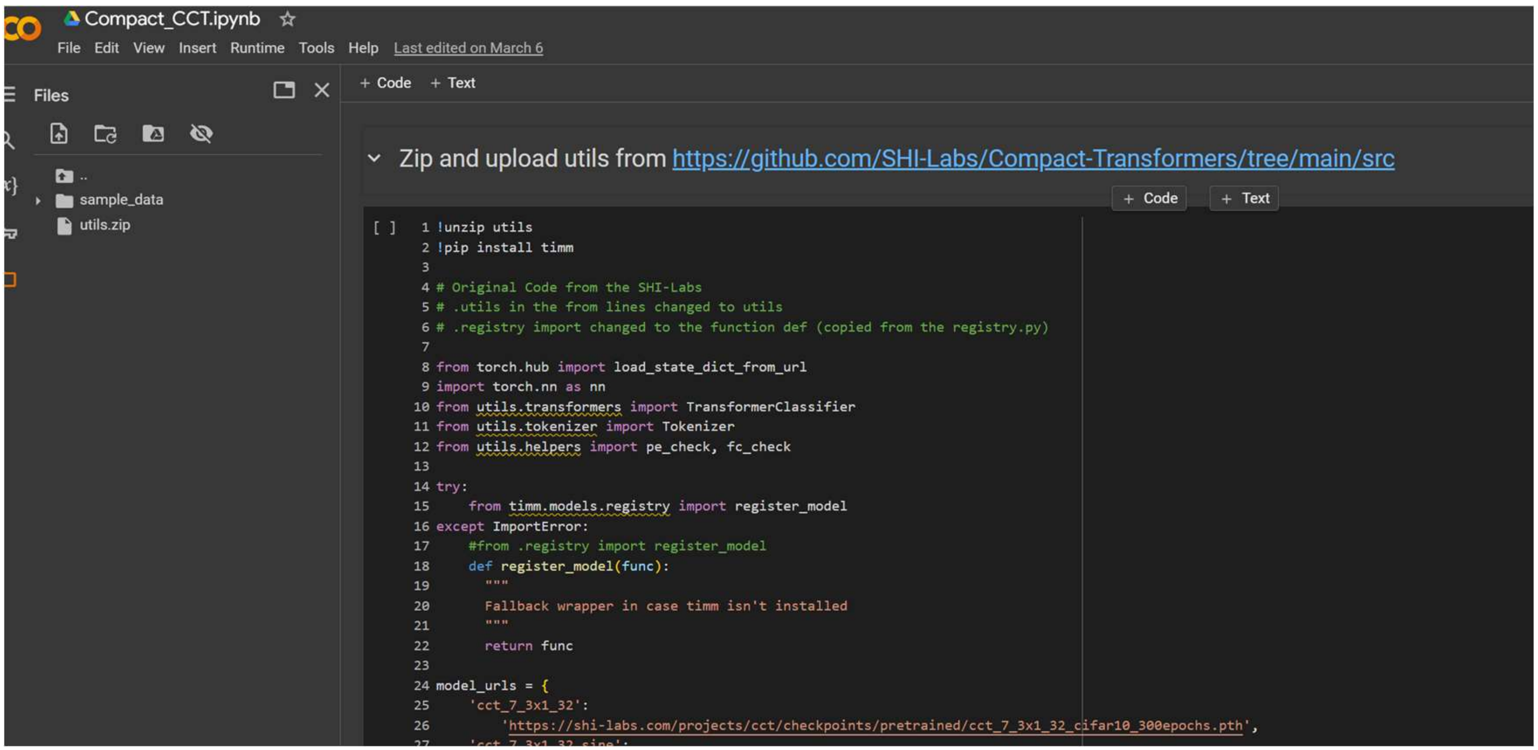
- CNN Backbone for object detection
  - YOLO (([1612.08242] YOLO9000: Better, Faster, Stronger (arxiv.org), [1804.02767] YOLOv3: An Incremental Improvement (arxiv.org)):
- CNN backbone for Visual features and a Transformer Decoder for object detection
  - DETR ([2005.12872] End-to-End Object Detection with Transformers (arxiv.org))
  - Anchor DETR ([2109.07107] Anchor DETR: Query Design for Transformer-Based Object Detection (arxiv.org))
- Purely Transformer based design for object detection
  - YOLOS ([2106.00666] You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection (arxiv.org)): Attention-only architecture directly built upon the ViT
- CCT backbone for object detection
  - Our focus

Proposing **Compact Transformer Filter (CTF)**

CTF = CCT + Task1 + Task 2

CCT Code

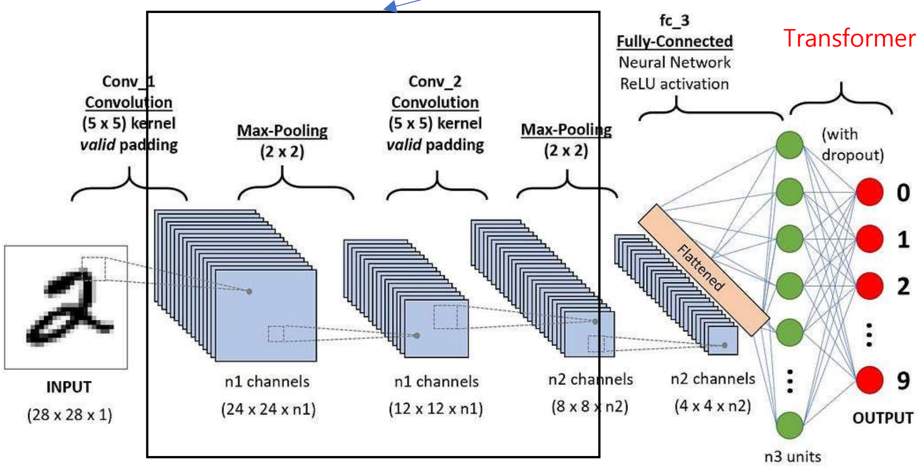
- Shared our CCT python code written in PyTouch



Task 1

- Switch 2D to 1D in the given code

Modify

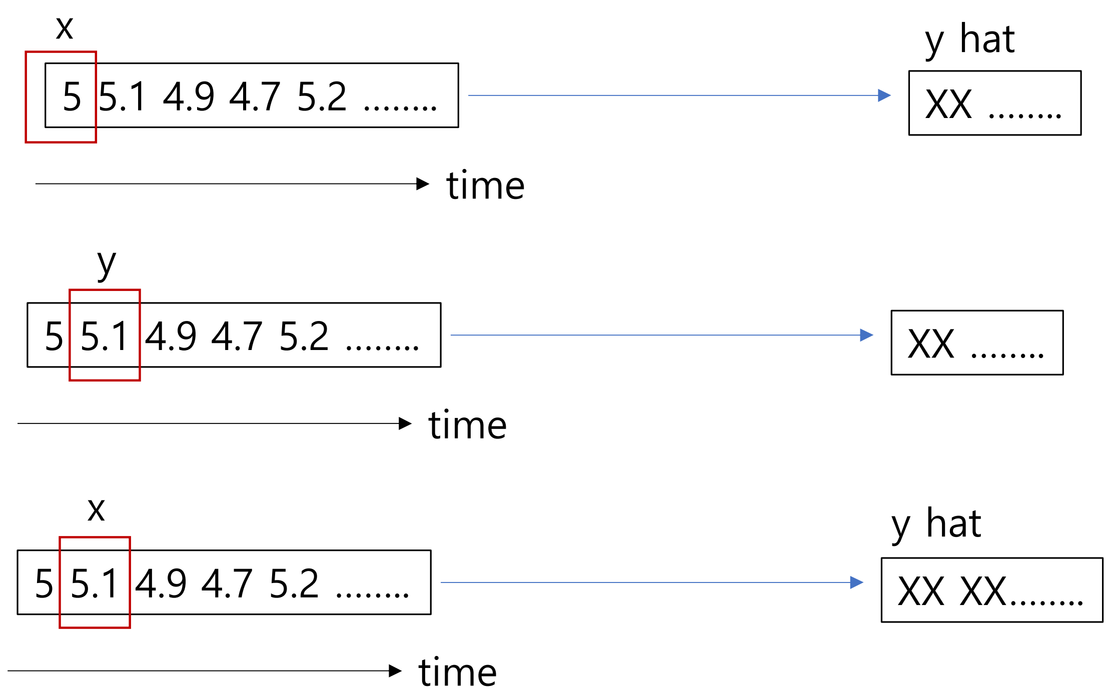


Your Focus

Input: 3-dimensional array → Input: 1-dimensional array

Task 2

- Continuous Text Input Streams without a learning phase
- Example



Deliverable: Python Code
CTF evaluated with Collected Data