Continual Learning Augmentation

Daniel Philps

Rothko Investment Strategies

City, University of London

Motivations for CL…

- **Address Catastrophic forgetting** in a noisy real-world context
- **Continual Learning:** Open-world learning for <u>**states**</u> not tasks
- **Memory-augmentation of well-understood learners (including LSTM)**
- **Interpretable**: Have an interpretable way of using memories

Continual Learning Augmentation (CLA) Results…

Outperformance in Developed Market Equities

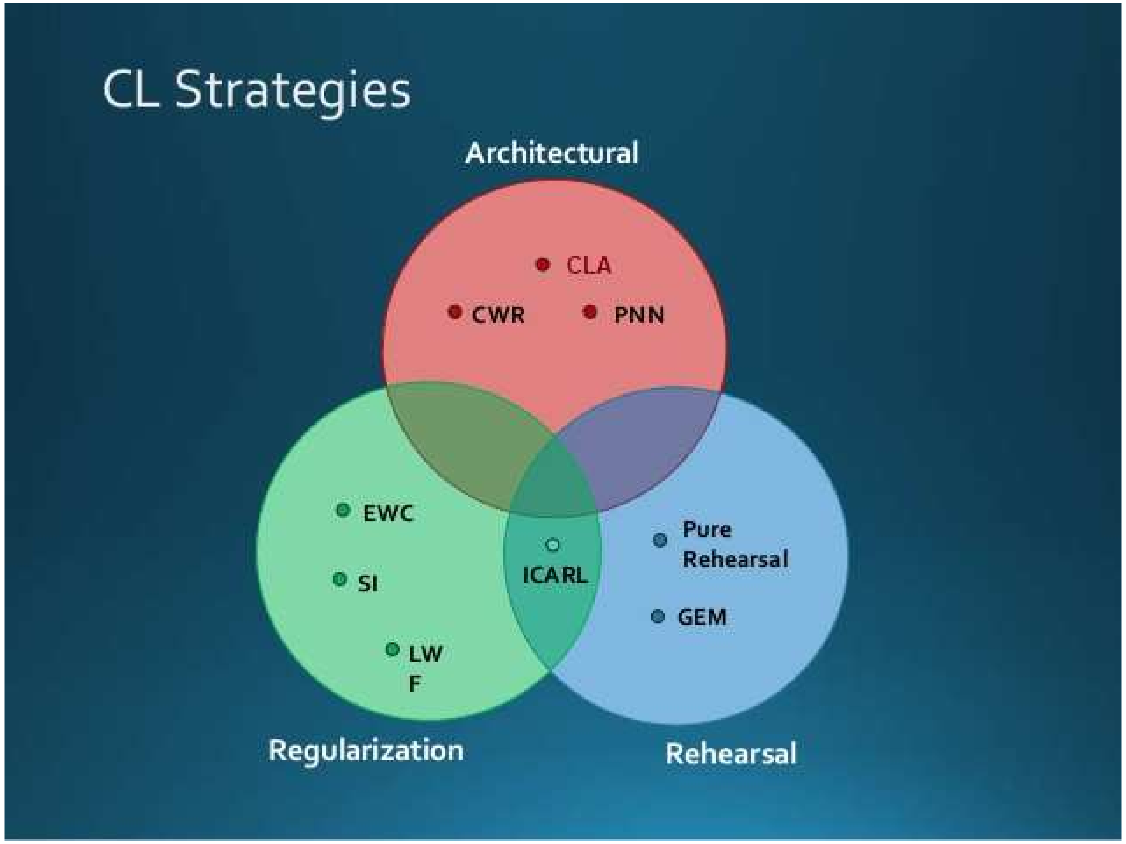
L/S Tests. ACW Universe. 2003-2017 annualized		
Base Learner	Simple	**CLA Augmented**
OLS	-0.3%	**+5.1%**
FFNN	+2.9%	**+7.2%**

Outperformance in Emerging Market Equities

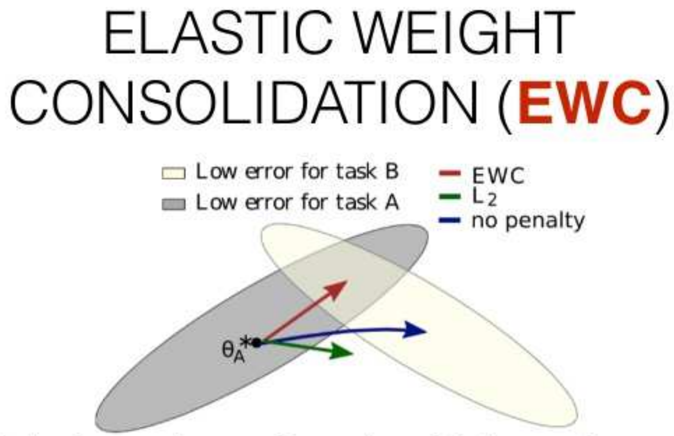
L/S Tests. EM Universe. 2007-2017 annualized		
Base Learner	Simple	**CLA Augmented**
LSTM	-5.0%	**+0.9%**
FFNN	-1.94%	**+2.1%**

Continual Learning (CL)...

Dealing with Catastrophic Forgetting



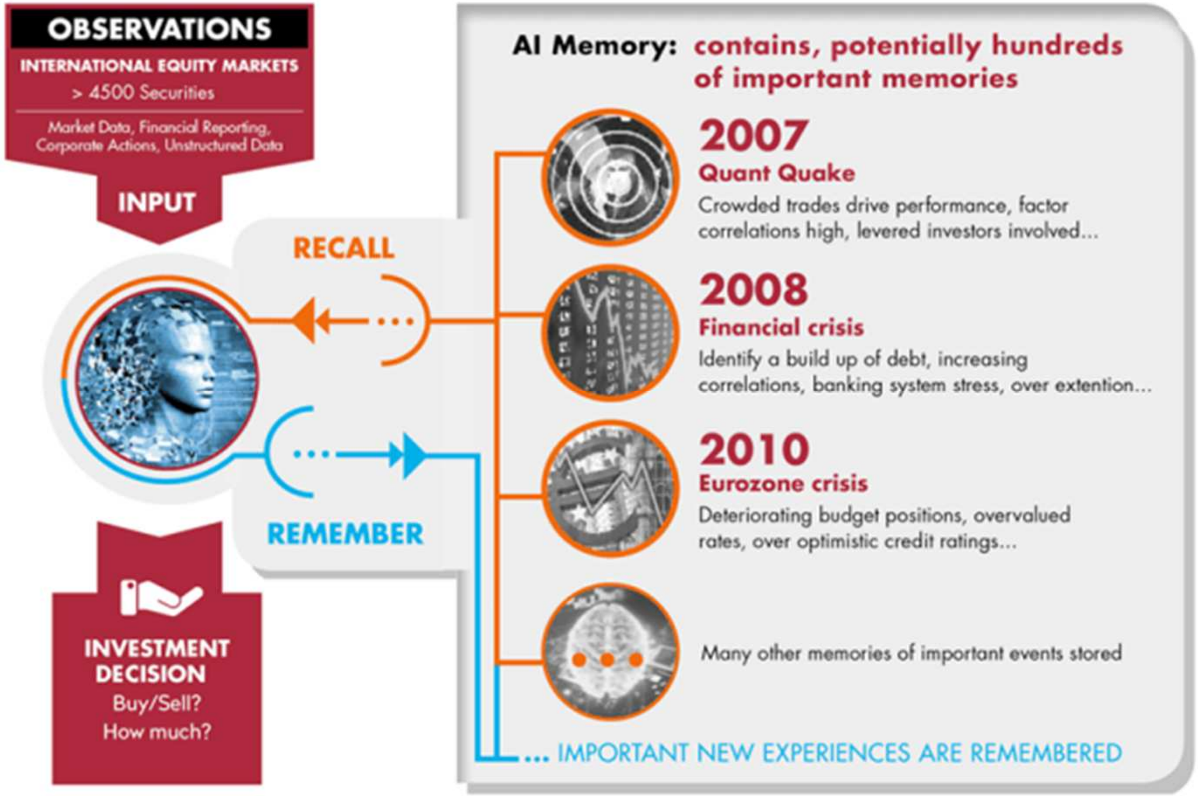
Maltoni Lomonaco, 2019



- θ_A* refers to the configuration of θ that performs well at A
- slow down the learning for the weights that were important to previous task(s)

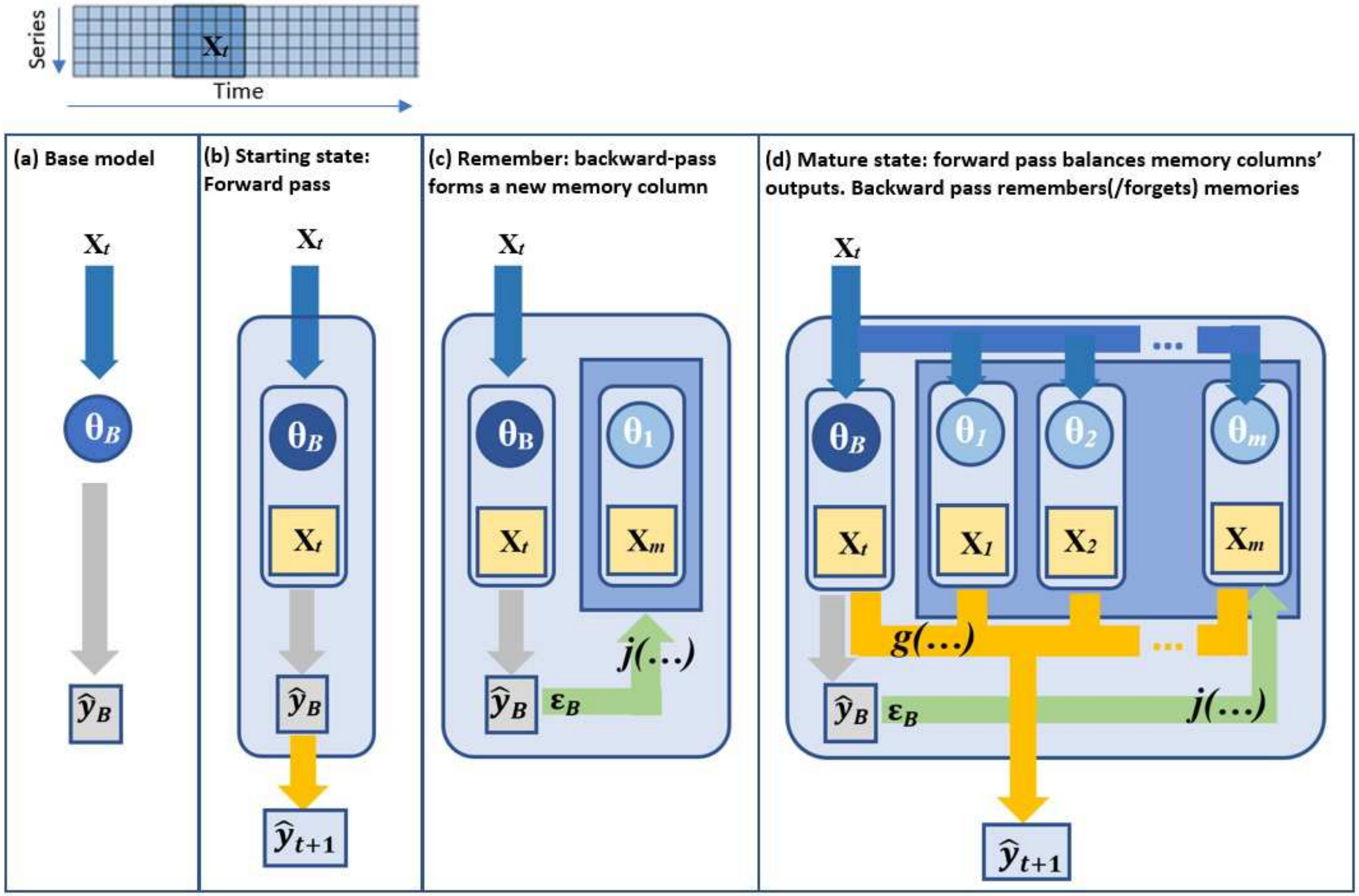
Kirkpatrick et al 2017

CL: A real world application...

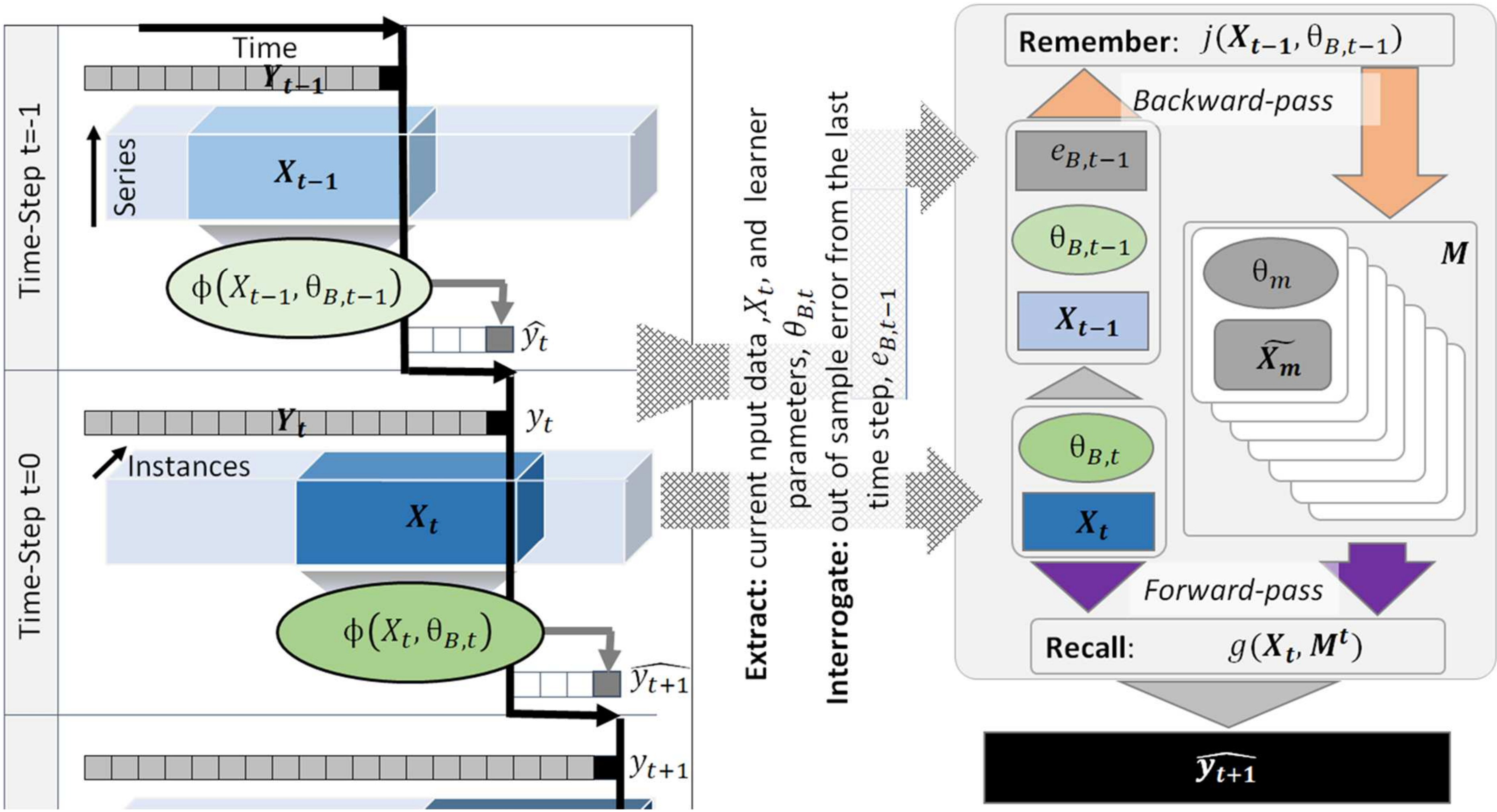


CLA Architecture…

Continual Learning Augmentation (CLA) Architecture…



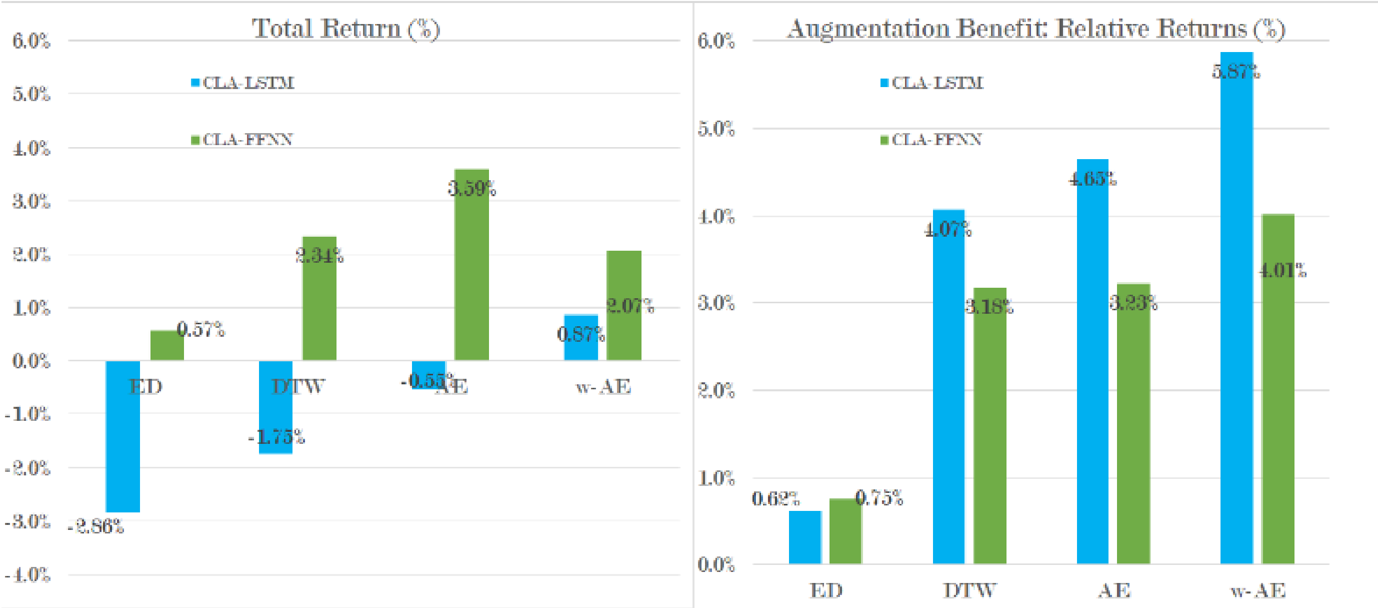
Continual Learning Augmentation (CLA) Architecture...



Results…

Simulation's Emerging Market Equities Stock Selection

Figure 6.7: Summary of Test Results



Note: LSTM vs FFNN, Long Short tests: Plot of median augmentation benefit by distance measure.

- **CLA-FFNN performs best**
- **CLA-LSTM, most augmented**
- **ED, poorest performer vs DTW**
- **wAE best performer**

Figure 6.9: Monotonicity of Augmentation: CLA-LSTM Median Test Deciles

Note: Augmentation benefit monotonicity is noted in all distance measures by a positive slope coefficient: RR_p of each decile portfolio of CLA-LSTM, annualised over the study term. Decile 1 relates to a portfolio of stocks in the lowest 10% returns forecasted by CLA-LSTM at each rebalance date, simulated as described.

- **Monotonic returns**
- **Notably for AE approaches**

Remember-gate…

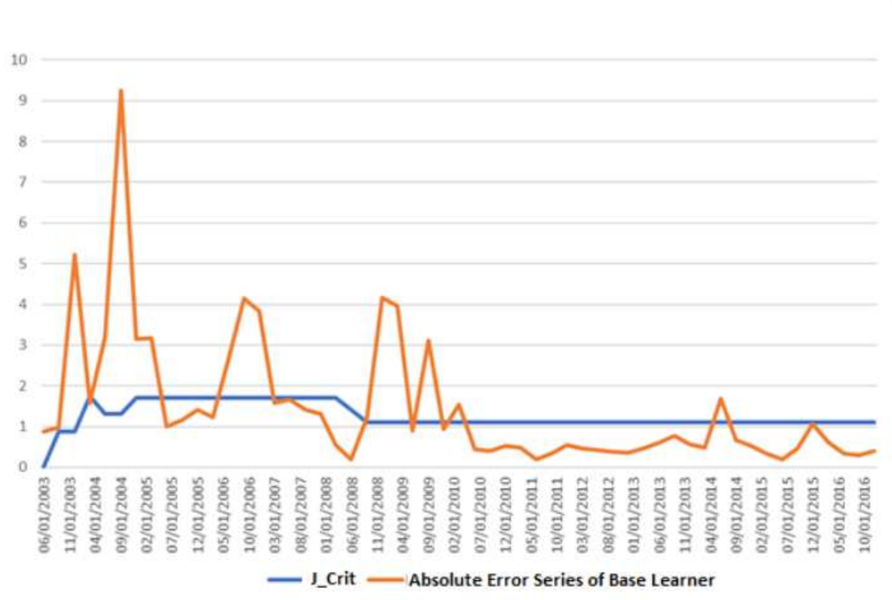
Learning to Remember

Algorithm 1 Remember Gate j

Require: Initialise memory structure $\boldsymbol{M}$
Require: Initialise J_{Crit}
Require: Train base learner $\theta_{B,t=0}$
  \# Step through time, period by period
  for all time steps t=1 in T **do**
    \# Base learner is run...
    $\hat{y}_t \leftarrow \phi(\boldsymbol{X}_{t-1}, \theta_{B,t-1})$
    \# y_t becomes observable
    \#
    \# .... CLA backpass starts ....
    $\epsilon_{B,t} = L(\hat{y}_t, y_t)$
    if $|\epsilon_{B,t}| \geq J_{Crit}$ **then**
      $\boldsymbol{X}_m \leftarrow \boldsymbol{X}_{t-1}$ store raw training instances
      append learner memory $(\boldsymbol{X}_m, \theta_{B,t-1})$ to $\boldsymbol{M}$
    end if
    \# CLA Learns J_{Crit} sensitivity
    $J_{Crit} \leftarrow$ learn and update J_{Crit}
    \# .... CLA backpass ends ....
    \#
    $\theta_{B,t} \leftarrow (\boldsymbol{X}_t, \theta_{B,t})$ overwrite base learner
  end for

- **Learn to remember: Jcrit over time**
- **Non-parametric learning threshold approach**

Figure 6.5: Memory Dynamics (CLA-FFNN): Learning J_{Crit}



Note: J_{Crit} is learned over time to define change points in the absolute error series ϵ_B of the base learner. It is notable that, as time passes, the error series becomes more stable, and as a result, j_{Crit}. This is consistent with the central limits described earlier in this thesis.

- **As time progresses … learning to remember becomes more stable**

Recall-gate
Distance measures…

Recall-Gate: Time-series similarity tests

i) **ED**: Memory contains training examples; sample over pairings

$$\hat{D}_{ED}(\widetilde{\boldsymbol{X}_{\boldsymbol{m}}}, \boldsymbol{X_t}) = 1/N \sum_{N} ED(\widetilde{X}_{m,r_1(D)}, X_{t,r_2(D)})$$

ii) **DTW**: As ED but with DTW similarity: time-deformation invariant

$$\hat{D}_{DTW}(\widetilde{\boldsymbol{X}_{\boldsymbol{m}}}, \boldsymbol{X_t}) = 1/N \sum_{N} DTW(\widetilde{X}_{m,r_1(D)}, X_{t,r_2(D)})$$

Recall-Gate: Time-series similarity tests

i) **ED**: Memory contains training examples; sample over pairings

$$\hat{D}_{ED}(\widetilde{\boldsymbol{X}}_{\boldsymbol{m}}, \boldsymbol{X_t}) = 1/N \sum_{N} ED(\widetilde{X}_{m,r_1(D)}, X_{t,r_2(D)})$$

ii) **DTW**: As ED but with DTW similarity: time-deformation invariant

$$\hat{D}_{DTW}(\widetilde{\boldsymbol{X}}_{\boldsymbol{m}}, \boldsymbol{X_t}) = 1/N \sum_{N} DTW(\widetilde{X}_{m,r_1(D)}, X_{t,r_2(D)})$$

iii) **Auto-encoder** (AE): AE learns a representation of training data

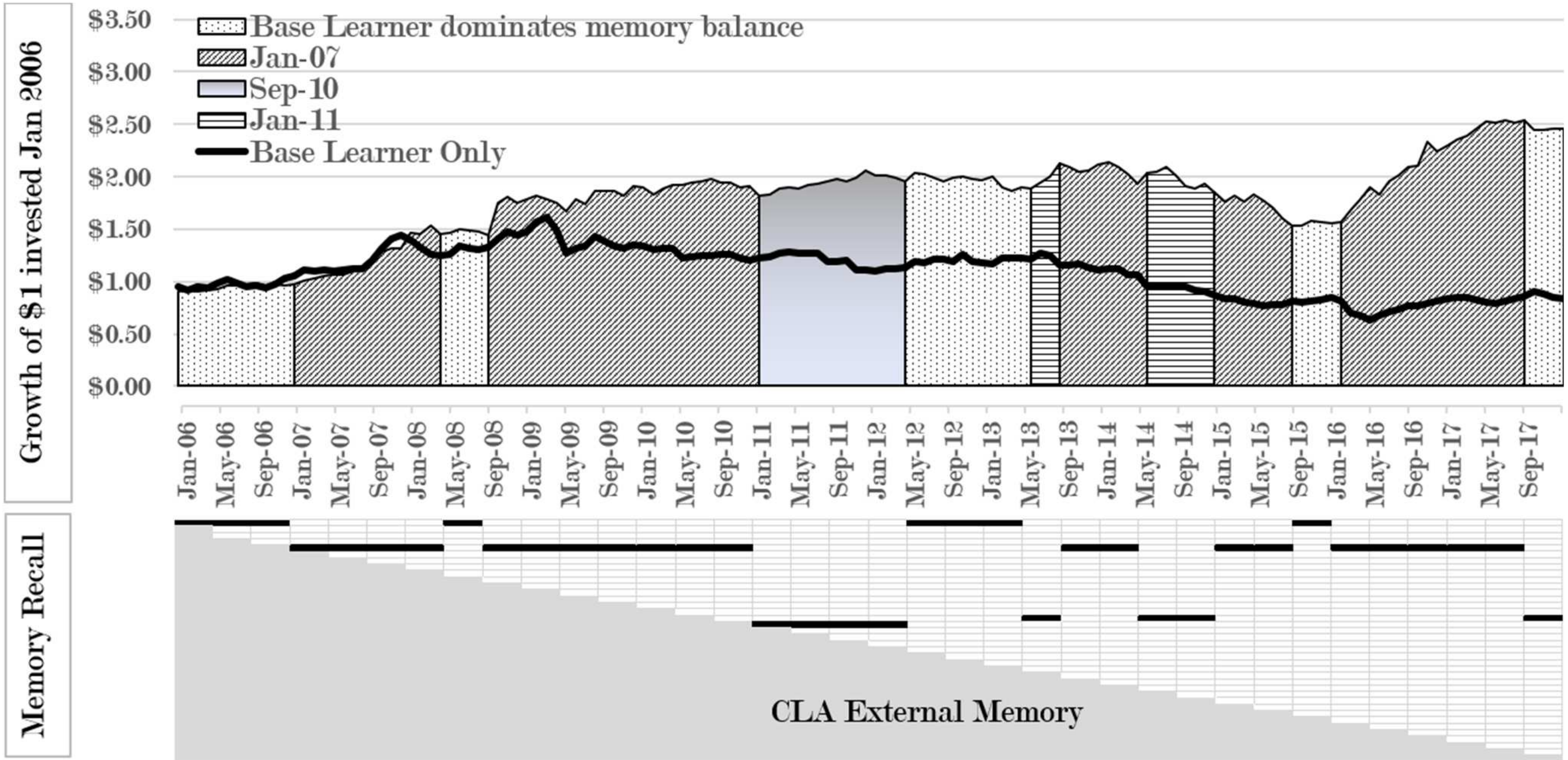
$$\hat{D}_{AE}(\widetilde{\boldsymbol{X}}_{\boldsymbol{m}}, \boldsymbol{X_t}) = 1/N \; ED(\boldsymbol{X}_t, a(h(\boldsymbol{X}_t)))$$

iv) **AE with DTW filter** (wAE): AE similarity but using DTW loss function

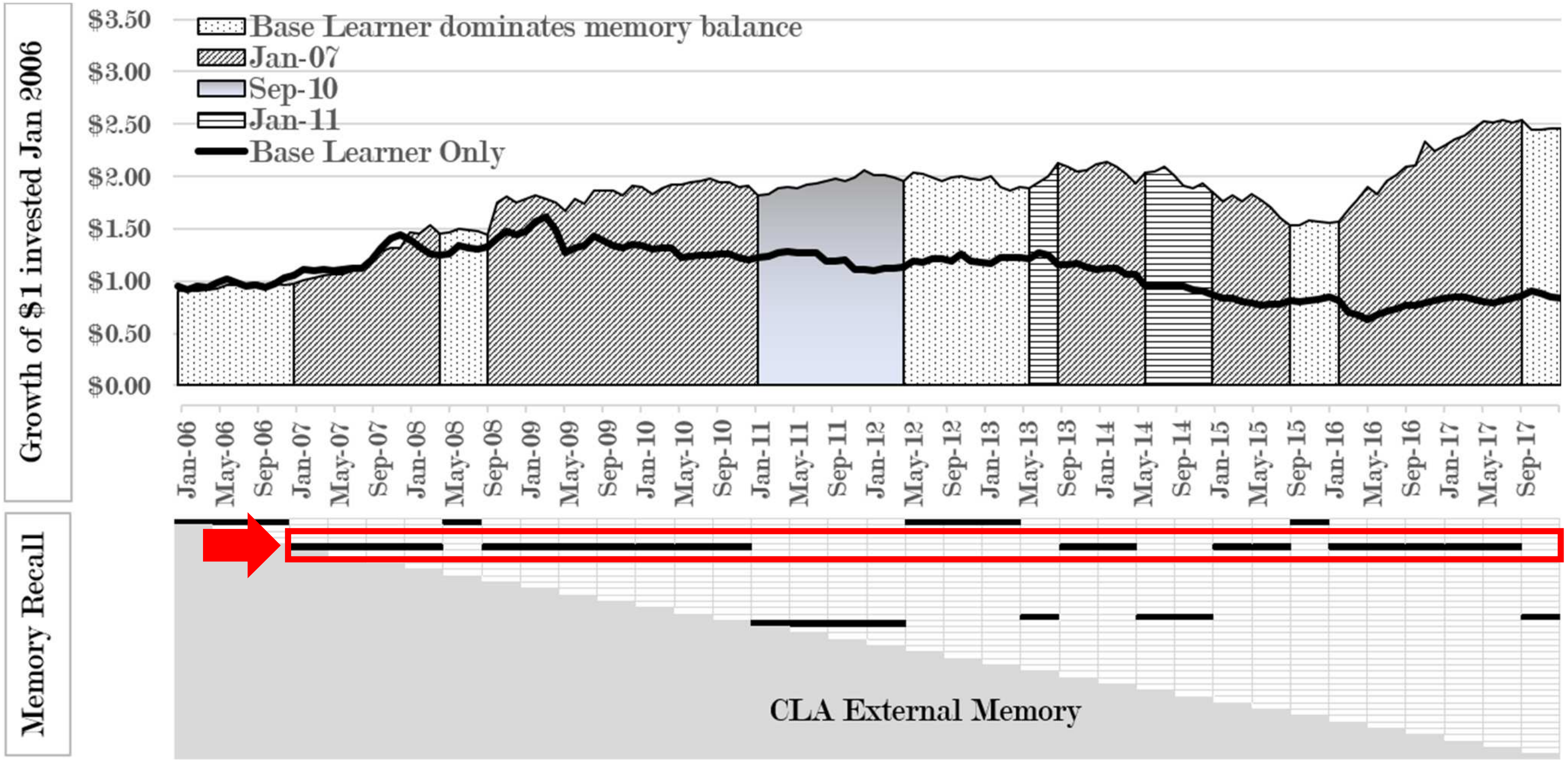
$$\hat{D}_{wAE}(\widetilde{\boldsymbol{X}}_{\boldsymbol{m}}, \boldsymbol{X_t}) = 1/N \; DTW(\boldsymbol{X}_t, a(h(\boldsymbol{X}_t)))$$

Interpretability…

Interpretability: Which memory did what, when…



Interpretability: Which memory did what, when…



Making Good on LSTMs' Unfulfilled Promise

Continual Learning Augmentation (CLA):

- ✓ **Addresses Catastrophic-forgetting in a noisy real-world context**
- ✓ **Benefits of Continual Learning but for time-series states**
- ✓ **Memory-augments well-understood learners (including LSTM)**
- ✓ **Interpretable use of memory**

Daniel Philps*
Rothko Investment Strategies
City, University of London

Artur d'Avila Garcez
City, University of London

Tillman Weyde
City, University of London

ROTHKO

CITY
UNIVERSITY OF LONDON
EST 1894