SAND: Streaming Subsequence Anomaly Detection

Paul Boniol
EDF R&D; University of Paris
paul.boniol@etu.u-paris.fr

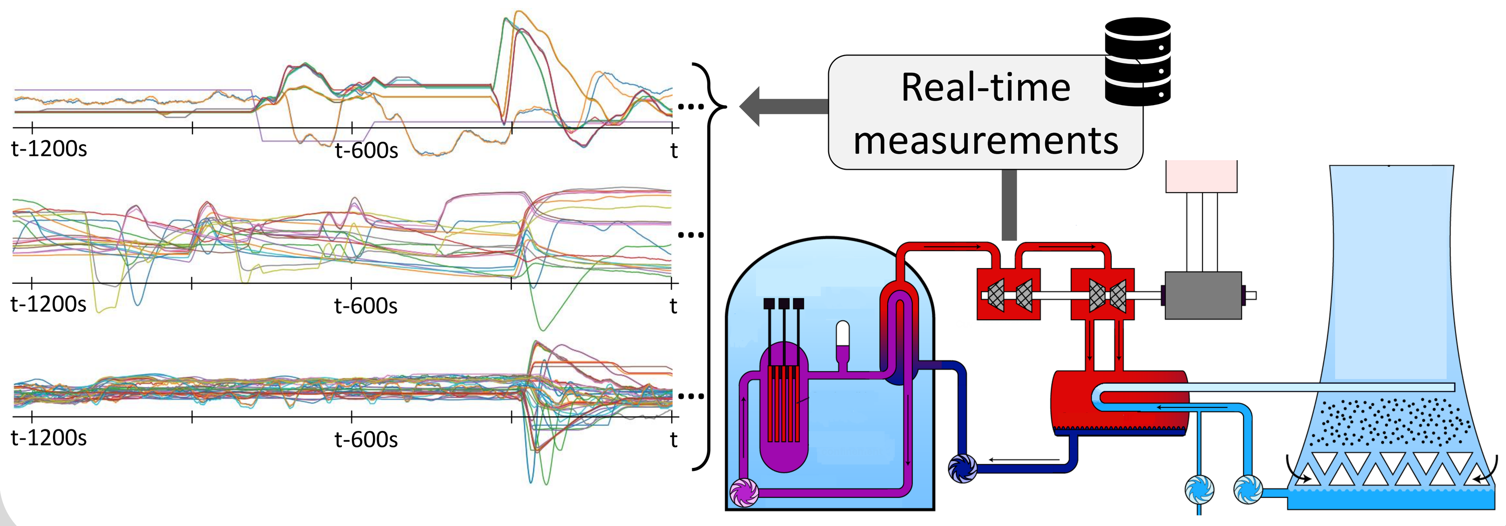
John Paparrizos
University of Chicago
jopa@uchicago.edu

Themis Palpanas
University of Paris; IUF
themis@mi.parisdescartes.fr

Michael J. Franklin
University of Chicago
mjfranklin@uchicago.edu

Motivation

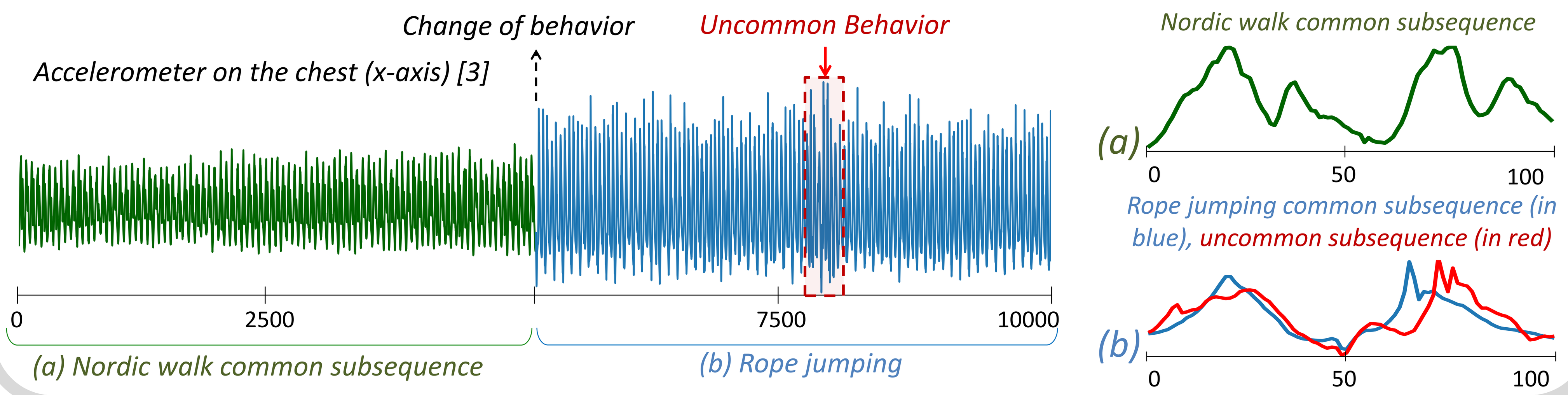
Subsequence anomaly detection in streams is an important problem with applications in medicine, energy production, etc



Challenges

State-of-the-art subsequence anomaly detection methods [1,2] are not able to perform on a streaming fashion.

Extensions to handle in real-time changes of normal behavior are needed.



Problem

We tackle the problem of subsequence anomaly detection in streams. Formally, for a given length ℓ, and a stream T, arriving in batch $\mathbb{T}^t_\ell$, return the η most abnormal subsequences of length ℓ.

Proposed Approach: SAND

STEP 1: Preprocessing

Computation of the model Θ using $k-Shape$ [4] on an initial batch $\mathbb{T}^0_{\ell_\Theta}$

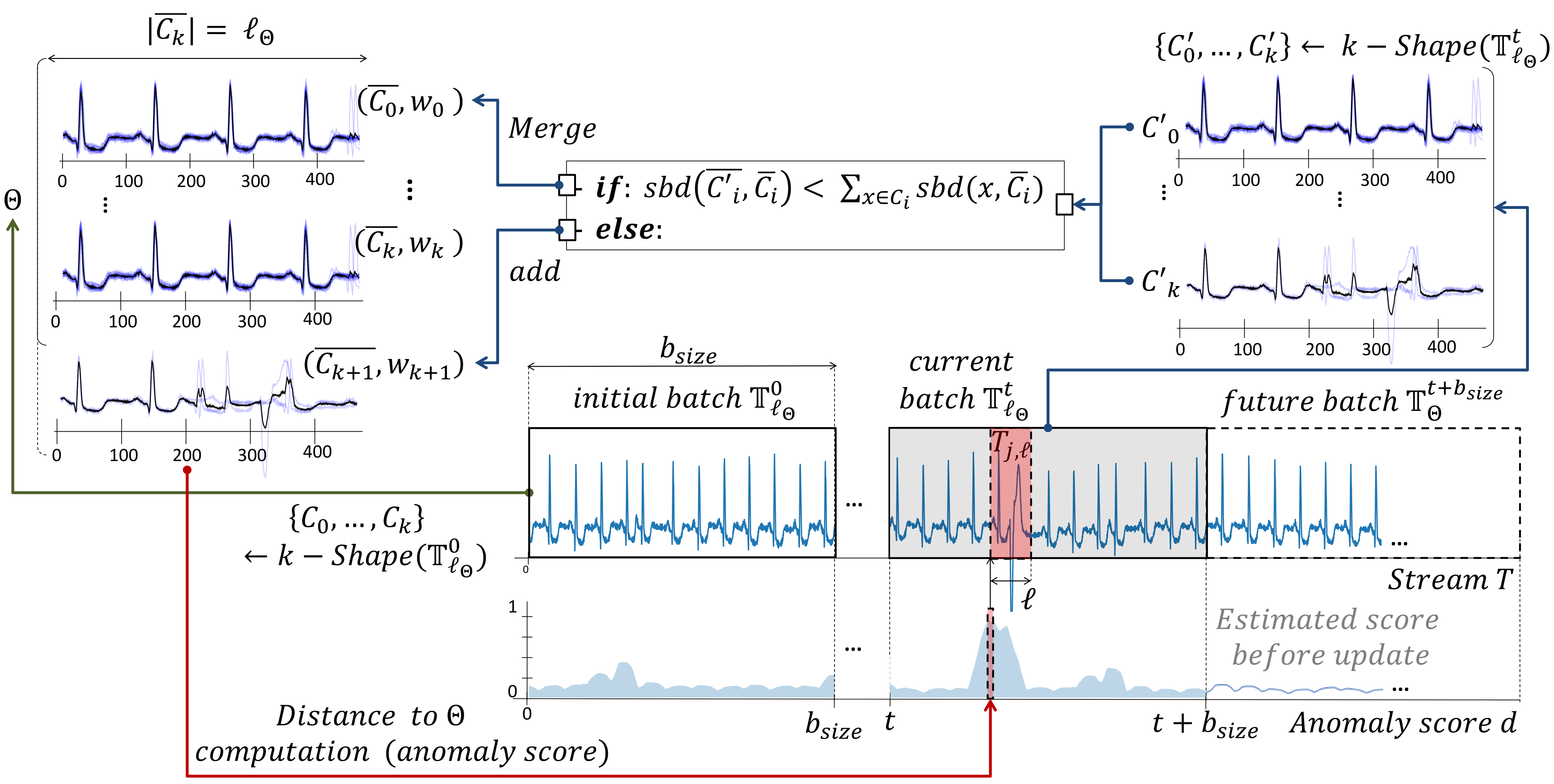
STEP 2: Model Update

Centroid $\overline{C_k}$ Update:
We propose a new solution that enable $k-Shape$ to operate incrementally. We do not store/use all previous subsequences to update the centroids

Weight w_k Update:
We update w_k associated to C_k based on the number of subsequences, the average extra-cluster distance, and the age of the subsequences.

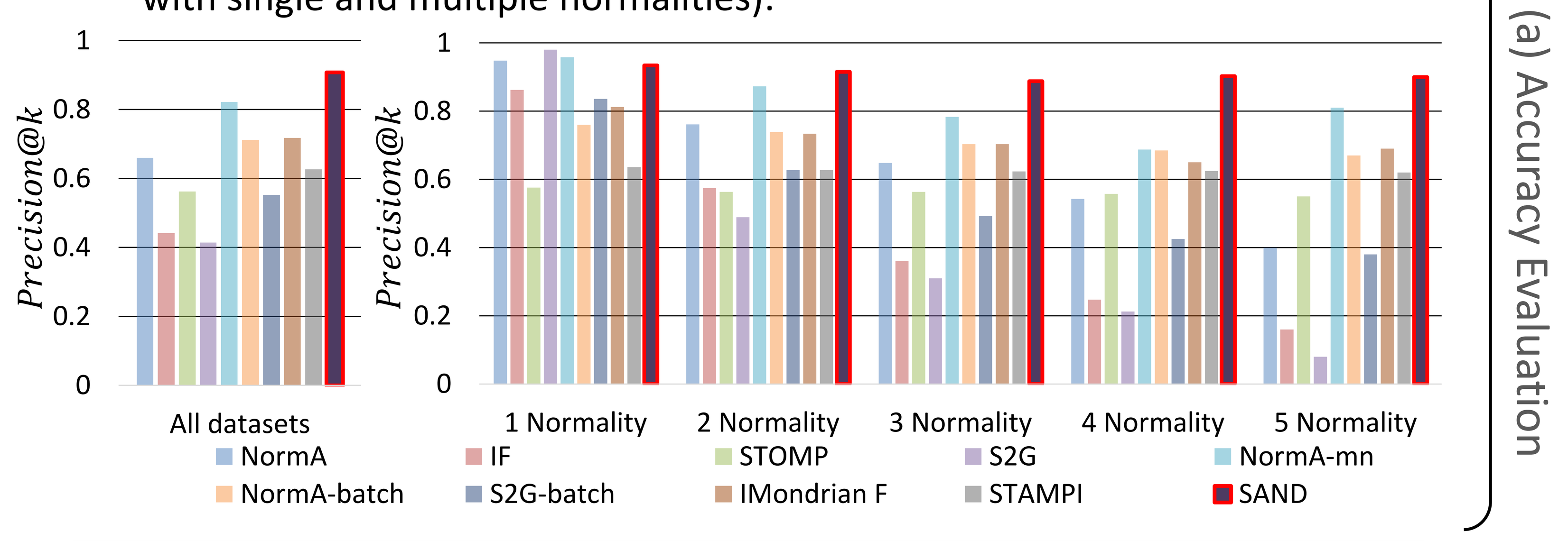
STEP 1: Anomaly Scoring

At any time, for a subsequence $T_{j,\ell}$ in the current batch $\mathbb{T}^t_{\ell_\Theta}$, we compute the distance of $T_{j,\ell}$ to the model Θ.

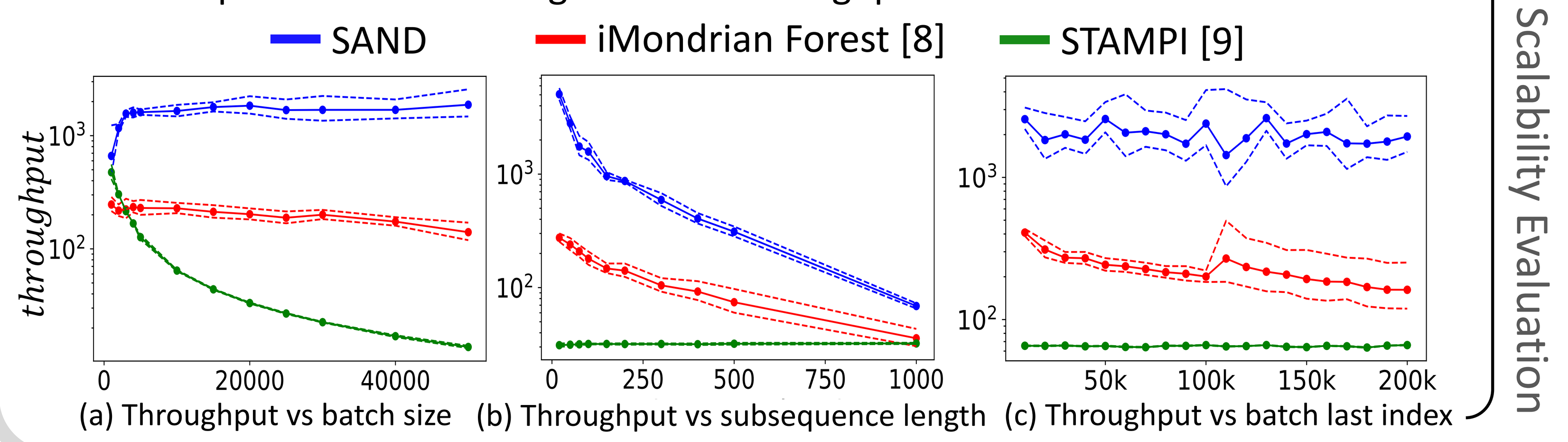


Experimental Evaluation:

- Comparison of static and streaming baselines on 30 data series [6,7] (both with single and multiple normalities):



- Comparison of streaming methods throughputs:



(a) Throughput vs batch size (b) Throughput vs subsequence length (c) Throughput vs batch last index

SAND in action: System overview

- Screenshot of the Web User interface based on SAND:

Bibliography

[1] Paul Boniol et al., **Unsupervised and Scalable Subsequence Anomaly Detection in Large Data Series**, VLDBJ 2021
[2] Paul Boniol et al., **Series2Graph: Graph-based Subsequence Anomaly Detection in Time Series**, PVLDB, 2020
[3] Dafne van Kuppevelt et al., **PAMAP2 dataset preprocessed v0.3.0**, 2017
[4] John Paparrizos et al., **k-Shape: Efficient and Accurate Clustering of Time Series**. SIGMOD, 2015
[5] H. A. Dau et al., **The UCR time series archive**, IEEE/CAA, 2019
[6] A. Abdul-Aziz et al., **Rotor health monitoring combining spin tests and data-driven anomaly detection methods.** Structural Health Monitoring, 2012
[7] G. et al., **Physiobank, physiotoolkit, and physionet**. Circulation.
[8] Haoran Ma et al., **Isolation Mondrian Forest for Batch and Online Anomaly Detection**, 2020
[9] Chin Chia Michael Yeh et al., **Matrix Profile I: All Pairs Similarity Joins for Time Series**, ICDM, 2016

Université de Paris

THE UNIVERSITY OF CHICAGO