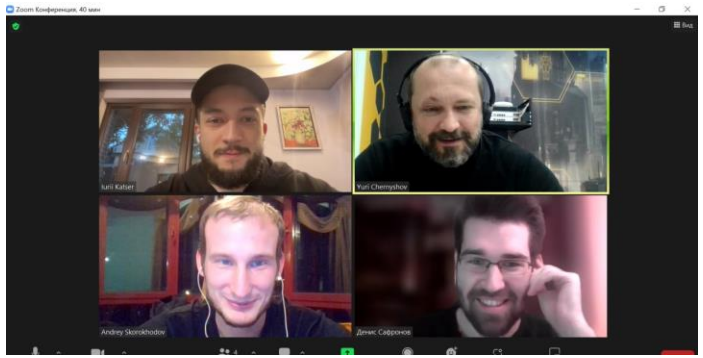
Исследовательский центр ИРИТ-РТФ УрФУ и ГК УЦСБ

- Исследования в области применения больших данных и ML в области кибербезопасности
- Внедрения в продуктах проактивного мониторинга и автоматизации
- Лаборатория кибербезопасности, сообщество Ural Cyber Security
- Студенческие проектные практикумы (100+ человек и 10+ команд в осенний семестр 2022)
- Дипломные и кандидатские диссертации
- Лекции и курсы



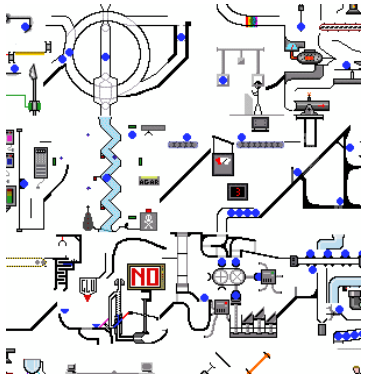
@yuchernyshov
ychernyshov@ussc.ru

Интерпретируемые модели машинного обучения для обнаружения аномалий в АСУТП

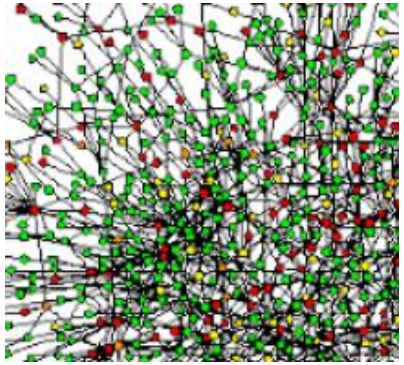
Юрий Юрьевич Чернышов
к.ф.-м.н., доцент, Уральский федеральный университет, Екатеринбург, Россия
ychernyshov@ussc.ru, https://orcid.org/0000-0002-8973-9383

RusCrypto 2023

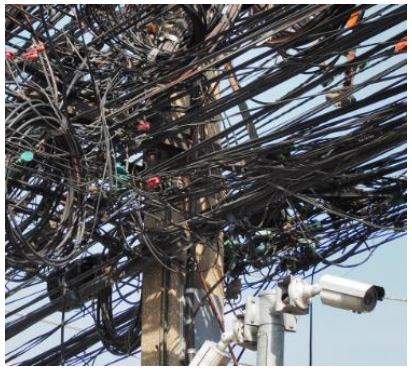
Технологические системы усложняются…



Number of components



Data



Communications

SW Dev & Ops Tools

Ages of systems

Vulnerabilities

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНЫЙ ЗАКОН

О безопасности критической информационной инфраструктуры Российской Федерации

Принят Государственной Думой 12 июля 2017 года

Одобрен Советом Федерации 19 июля 2017 года

Статья 1. **Сфера действия настоящего Федерального закона**

Настоящий Федеральный закон регулирует отношения в области обеспечения безопасности критической информационной инфраструктуры Российской Федерации (далее также – критическая информационная инфраструктура) в целях ее устойчивого функционирования при проведении в отношении ее компьютерных атак.

Статья 2. **Основные понятия, используемые в настоящем Федеральном законе**

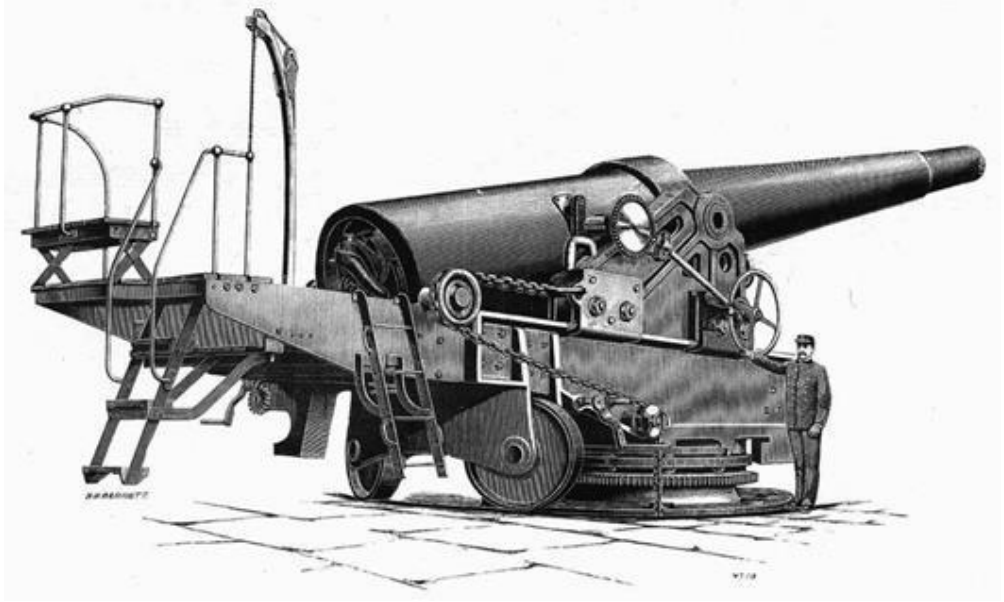
Для целей настоящего Федерального закона используются следующие основные понятия:

Regulations

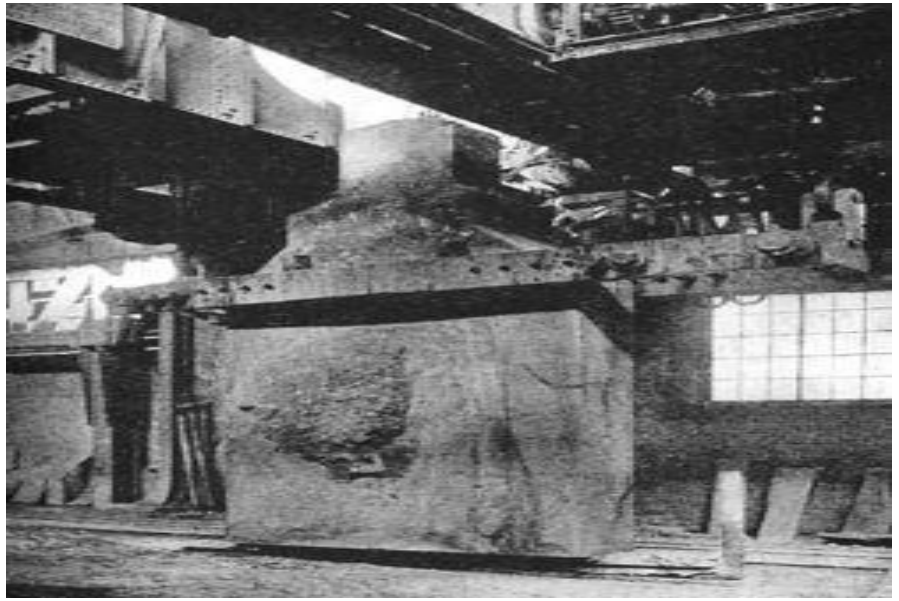
Infrastructures

… в т.ч. в кибербезопасности.

Атака: SW (malware, tools), OPs (fishing, DGA, OSINT, reverse engineering), technologies (DeepFake), processes (APT, bruteforce)



Защите: tools (ngX – IDS, EDR, SOAR), Ops (reverse engineering, forensic, monitoring, automation), processes (0-trust model)



Люди начинают не справляться

Malevich “black box”

https://www.theartnewspaper.ru/posts/2375

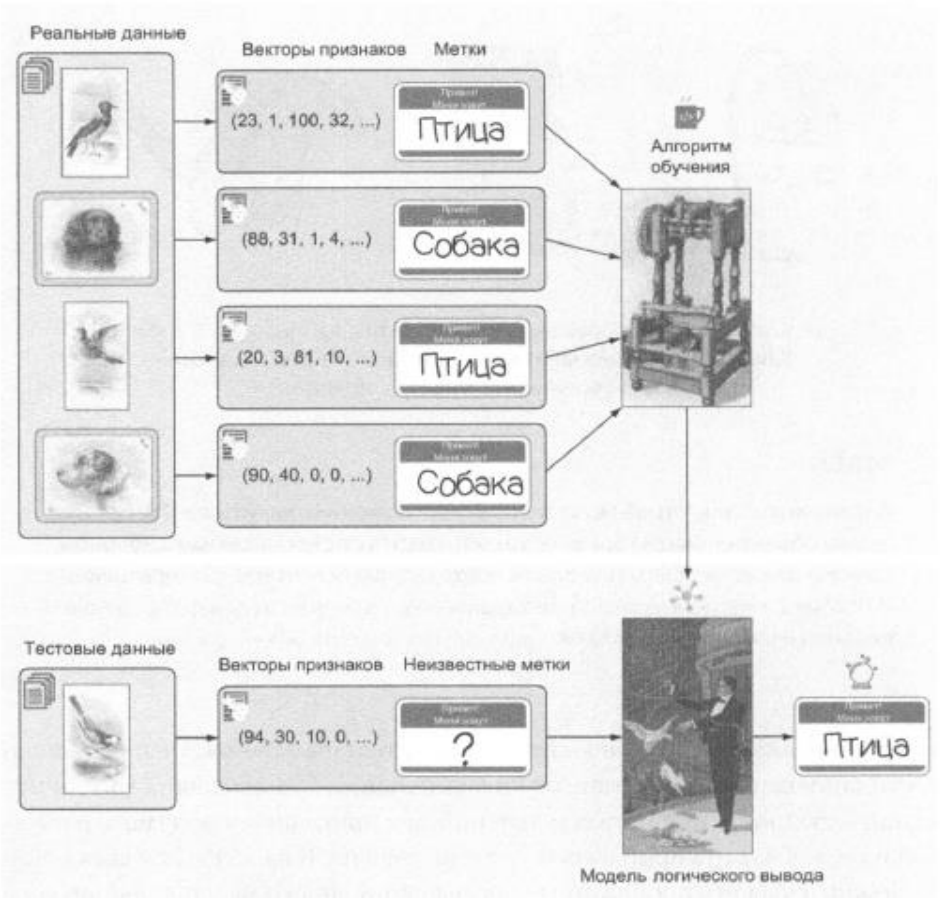
Проблема использования ML

Компромисс между точностью, скоростью и интерпретируемостью

- Правила или некоторые простые модели (Decision Trees, Regression) хорошо интерпретируемы
- AlexNet содержит 62M параметров,
- Появившиеся в 2016 году deep residual networks (ResNets) содержат более 200 уровней глубины, показывают результат в распознавании объектов, превышающий человеческие возможности.

Неожиданные последствия отсутствия контроля и интерпретируемости ML: дискриминация по расовому, половому или возрастному признаку, выучивание «цинизма» и вульгарность, невозможность объяснить причину решения и дать рекоменации.

Государственное регулирование интеллектуальных алгоритмов для определенных сегментов
https://eurlex.europa.eu/eli/reg/2016/679/oj.

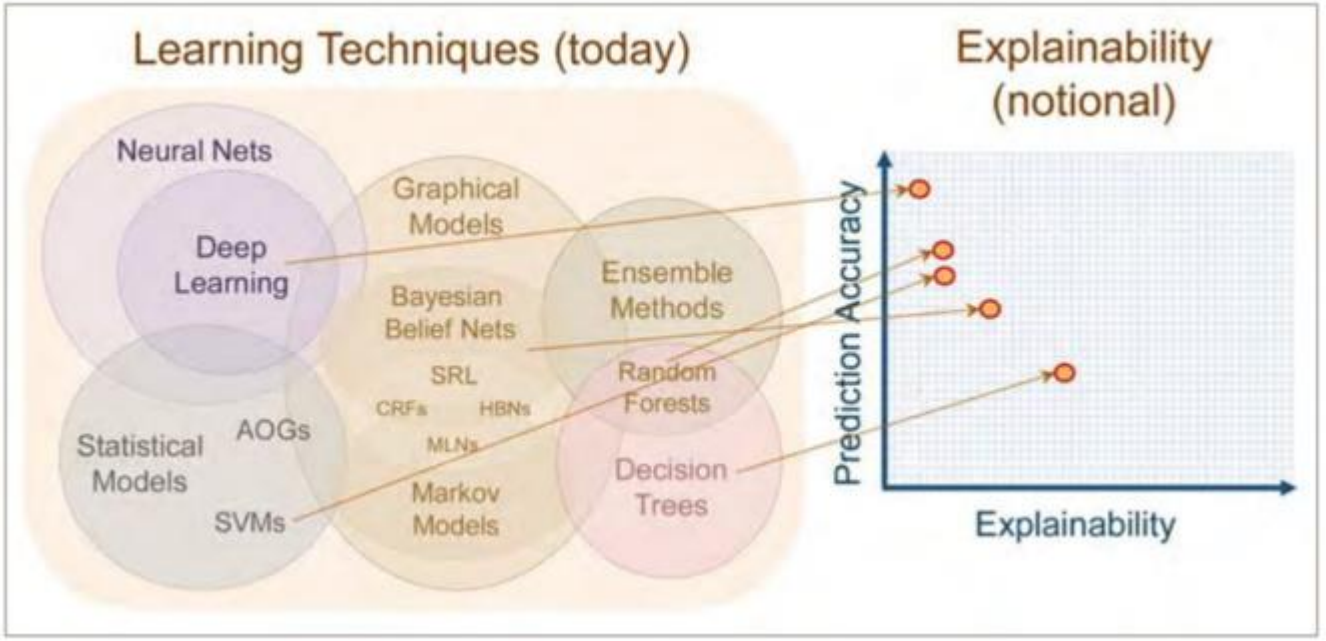


From S. Nishant «Machine learning»

Возможность объяснить паттерн часто важнее скорости обучения или предсказания

Просто знать метрику качества недостаточно

Explainable AI (DARPA)



Autonomy	Deep Learning *UC Berkeley*	Deep Adaptive Programs *Oregon State*	Cognitive Modeling *PARC*	Explainable Reinforcement Learning *Carnegie Mellon*	Causal Modeling *Charles River*	Stochastic And-Or-Graphs *UCLA*
Data Analytics	Stochastic And-Or-Graphs *UCLA*	Mimic Learning *Texas A&M*	Causal Modeling *Charles River*	Explanation by Example *Rutgers*	Probabilistic Logic *UT Dallas*	Deep Learning *Raytheon BBN* *SRI International* *UC Berkeley*

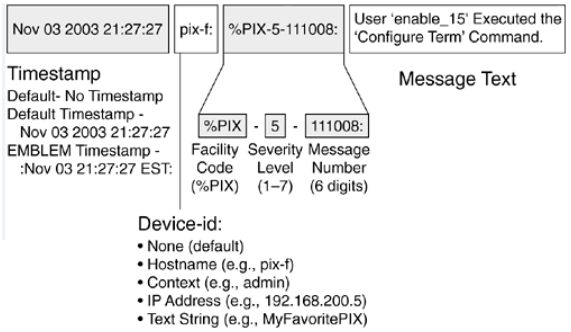
https://www.darpa.mil/program/explainable-artificial-intelligence

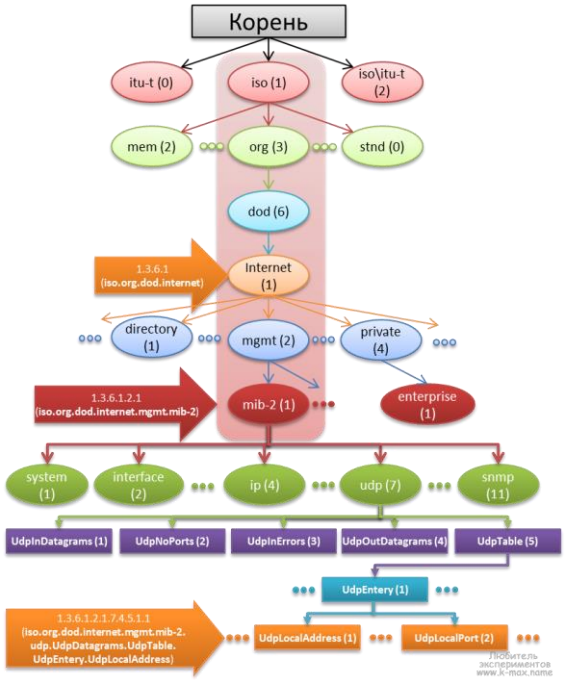
Ограничения reconstruction-based методов обнаружения аномалий

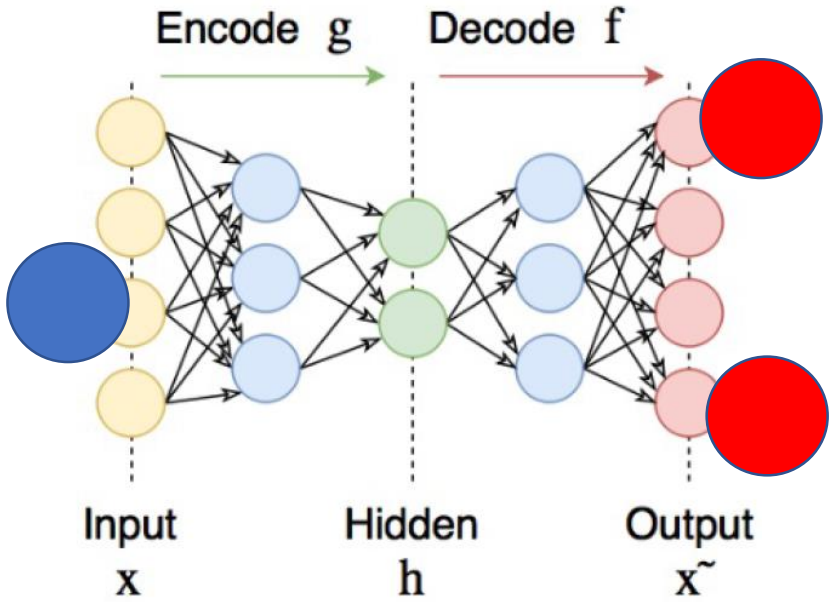
DL в состоянии распознать групповую аномалию в многомерных и мультимодальных данных

DL может найти нелинейные закономерности в данных (не только числовые, но и syslog)

Но остается проблема интерпретации







Что есть в нашем распоряжении?

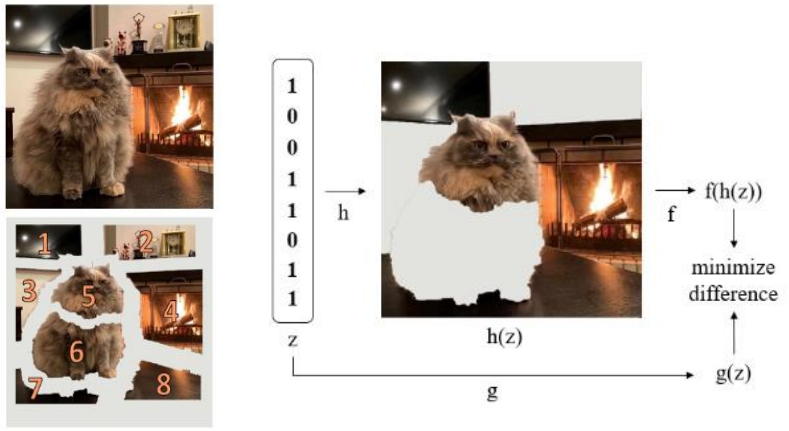
Local Interpretable Model-Agnostic Explanation (LIME)
“Why should I trust you? Explaining the predictions of any classifier”,
https://arxiv.org/pdf/1602.04938.pdf

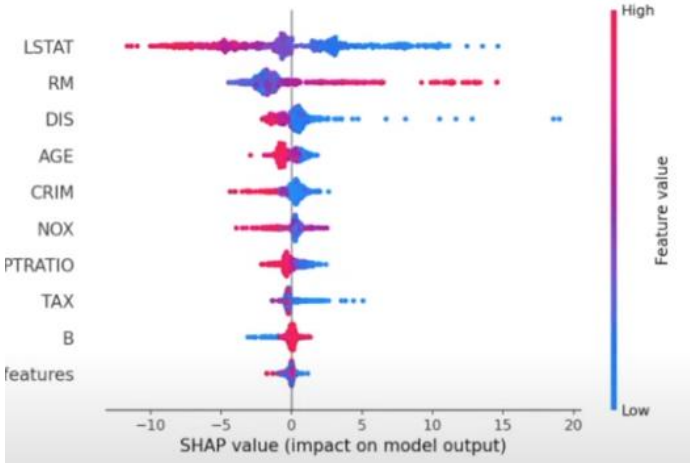
SHAP
“A Unified Approach to Interpreting Model Predictions”,
https://arxiv.org/pdf/1705.07874.pdf

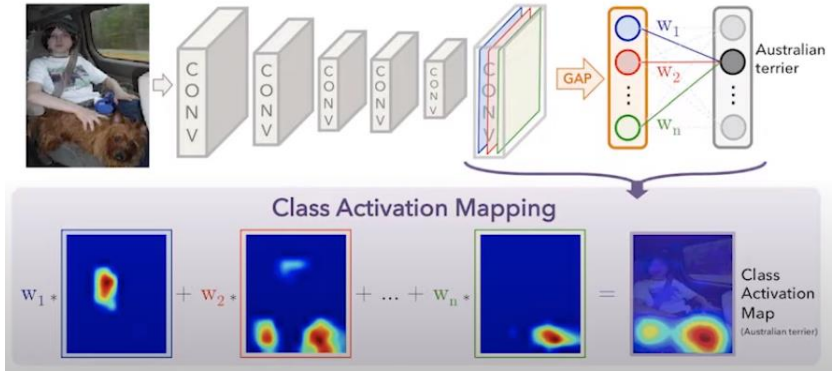
Градиентные методы.
“Learning deep features for discriminative localization”
https://arxiv.org/pdf/1512.04150.pdf

«Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond» Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, Dejing Dou

https://arxiv.org/pdf/2201.08164.pdf, comparison of Saliency, Input × Gradient, Integrated Gradients, Guided Backpropagation, Grad-CAM, Guided Grad-CAM, Lime, Occlusion, DeepLift, SmoothGrad.
Algorithms: DenseNet and ResNet, dataset: BigEarthNet (pictures from Sentinel-2 satellite).







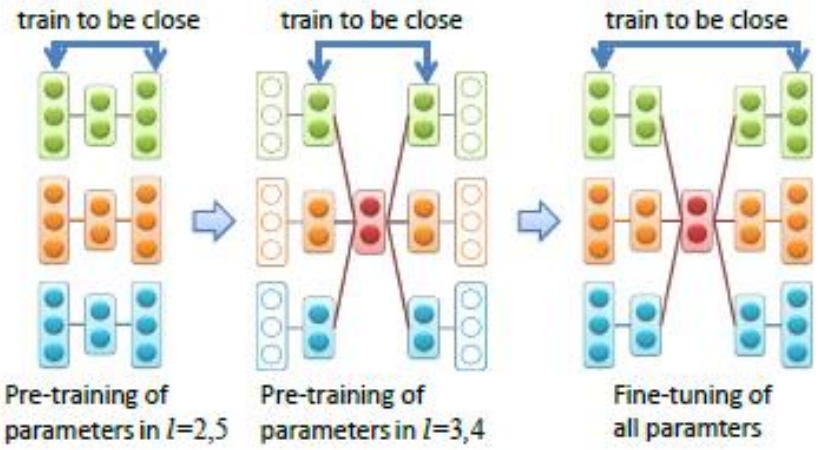
Мультимодальный автокодировщик

$$\min_{\boldsymbol{\eta}} MSE(\bar{\boldsymbol{x}} - \boldsymbol{\eta}) + \lambda ||\boldsymbol{\eta}||_1,$$

Reconstruction-based метод, AE
Старается минимизировать MSE, изменяя входные данные

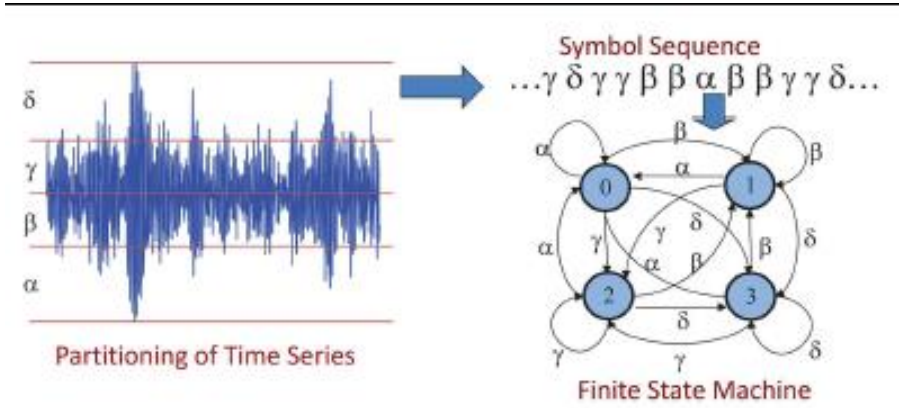
Проблема 1: «комбинаторный взрыв» при полном переборе в случае большого количества входных признаков

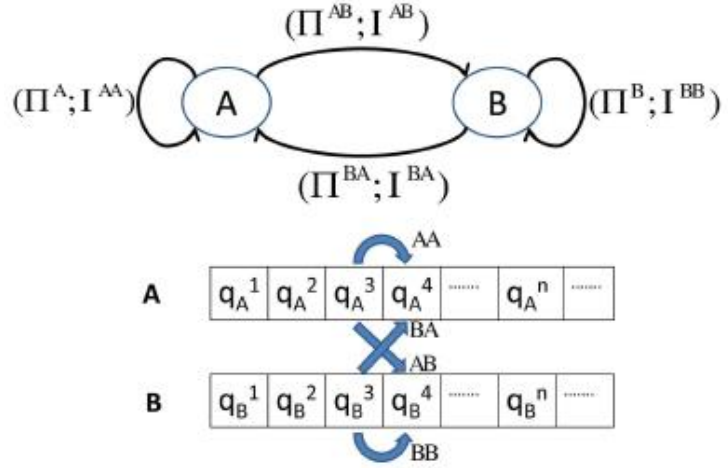
Проблема 2 : cross-domain данные имеют разные типы и по разному влияют на обучаемость модели (level of learnerability)

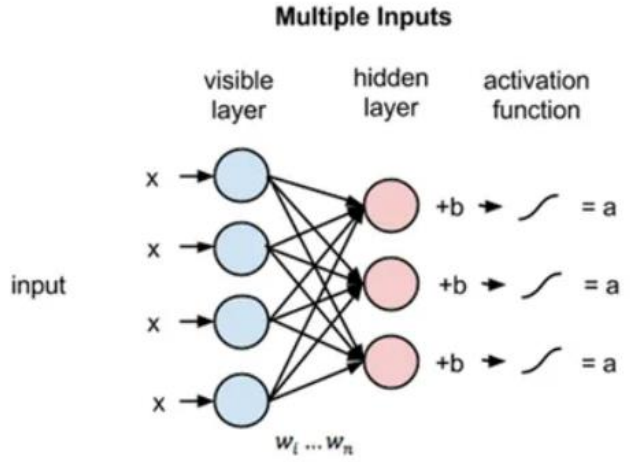


Y.Ikeda, K.Ishibashi, Yu.Nakano, K.Watanabe, R.Kawahara. Anomaly Detection and Interpretation using Multimodal Autoencoder and Sparse Optimization https://arxiv.org/pdf/1812.07136.pdf

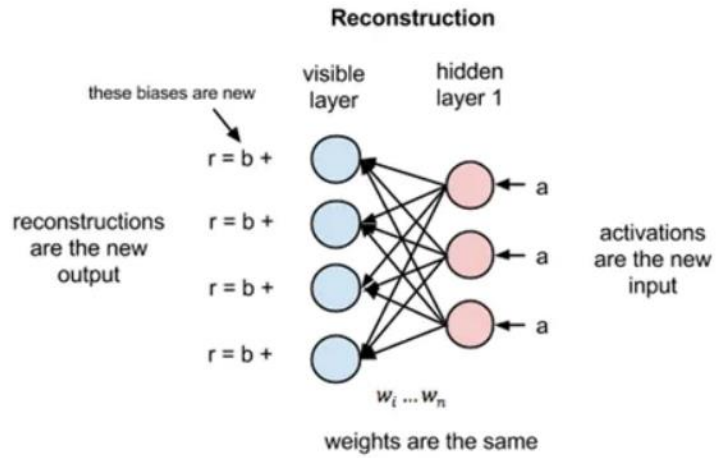
Спатиотемпоральное графовое моделирование







$$h^{(1)} = \sigma\left(v^{(0)T}W + b\right)$$



$$v^{(1)} = \sigma\left(h^{(1)}W^{T} + a\right)$$

“Root-cause Analysis for Time-series Anomalies via Spatiotemporal Graphical Modeling in Distributed Complex Systems” C.Liu, K.G. Lore, Z. Jiang

“RBM with practical implementation” https://medium.com/machine-learning-researcher/boltzmann-machine-c2ce76d94da5

Наборы данных, цифровые двойники, стенды

SUTD: SWAT, WADA, EPIC, IoT
Skolteh: SKAB
Harvard: TEP
Numenta: NAB

Dataset Requests (2017 – 2021)

S/N	Name	Organisation	Origin of Request	Date of request	Dataset requested
933	Peng Kang	Southeast University	China	25-Dec-21	SWaT, WADI
1932	Li Yunpeng	Southwestern University of Finance and Economics	China	25-Dec-21	SWaT, WADI
1931	Meng Wei Xie	Fudan university	China	23-Dec-21	SWaT, WADI, IoT
			Hong		WADI, EPIC, CISS
1817	Liudmila Kopeikina	Eötvös Loránd University	Hungary	9-Nov-21	SWaT
1816	Chernyshov Yury	Ural Security Systems Center	Russia	8-Nov-21	SWaT
1815	Nozima Murodova	Inha University in Tashkent	Uzbekistan	8-Nov-21	SWaT
1814	Wu Jihua	Beijing University of Post and Telecommunication	China	8-Nov-21	SWaT, WADI
1813	Francesco Simone	Sapienza University of Rome	Italy	8-Nov-21	SWaT, WADI, CISS, BATADAL
1812	Tianhao Chen	Shandong University	China	8-Nov-21	SWaT, WADI, EPIC, CISS, Blaq_0, BATADAL, IoT
1811	Harsh Gupta	Indian Institute of Information Technology,	India	7-Nov-21	SWaT, WADI

https://itrust.sutd.edu.sg/

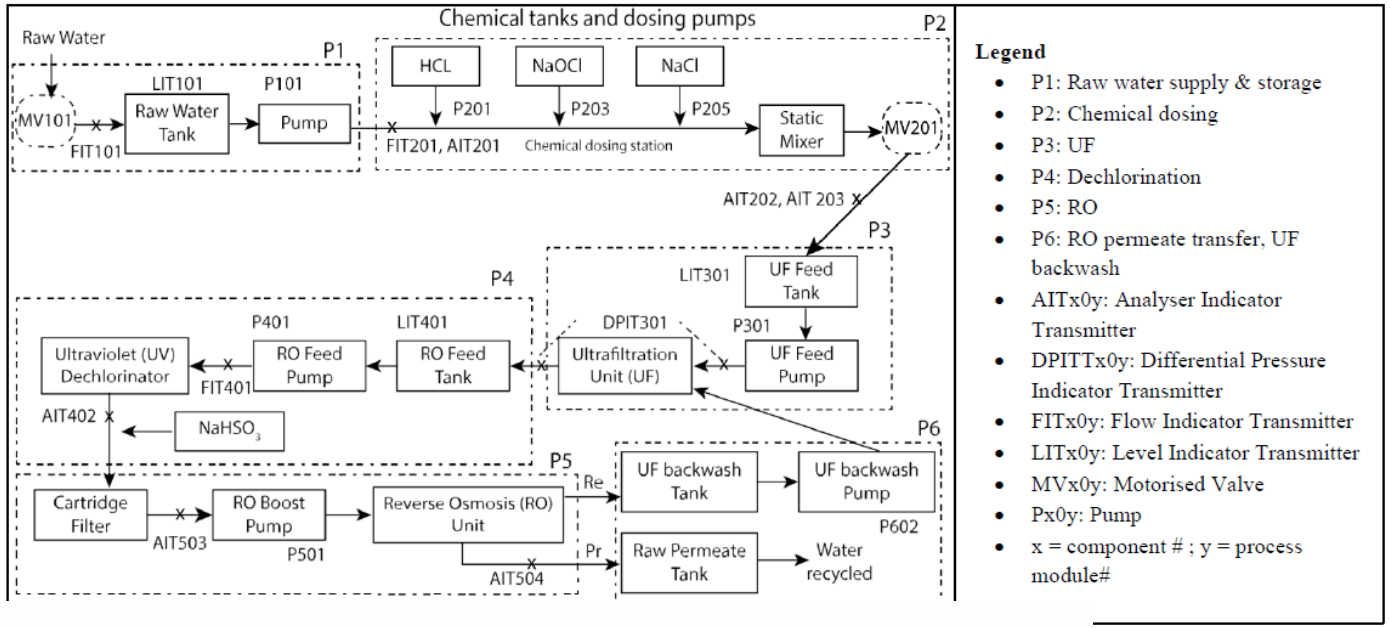
https://www.youtube.com/watch?v=i4vCG4clNZQ

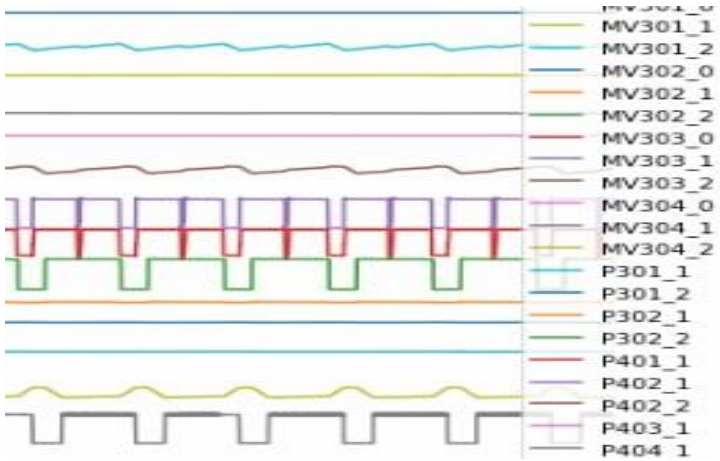
SWAT 2015 (SUTD) эксперимент

AD models: OCSVM, iForest, SOM, AE-LSTM,

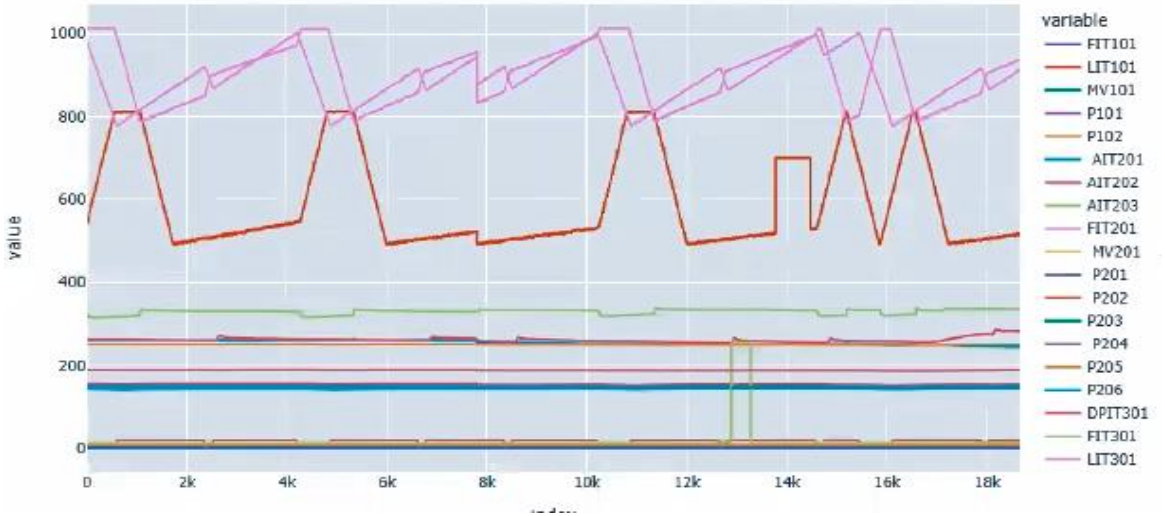
Entropy, …

AD frameworks: Prophet, ETNA, …





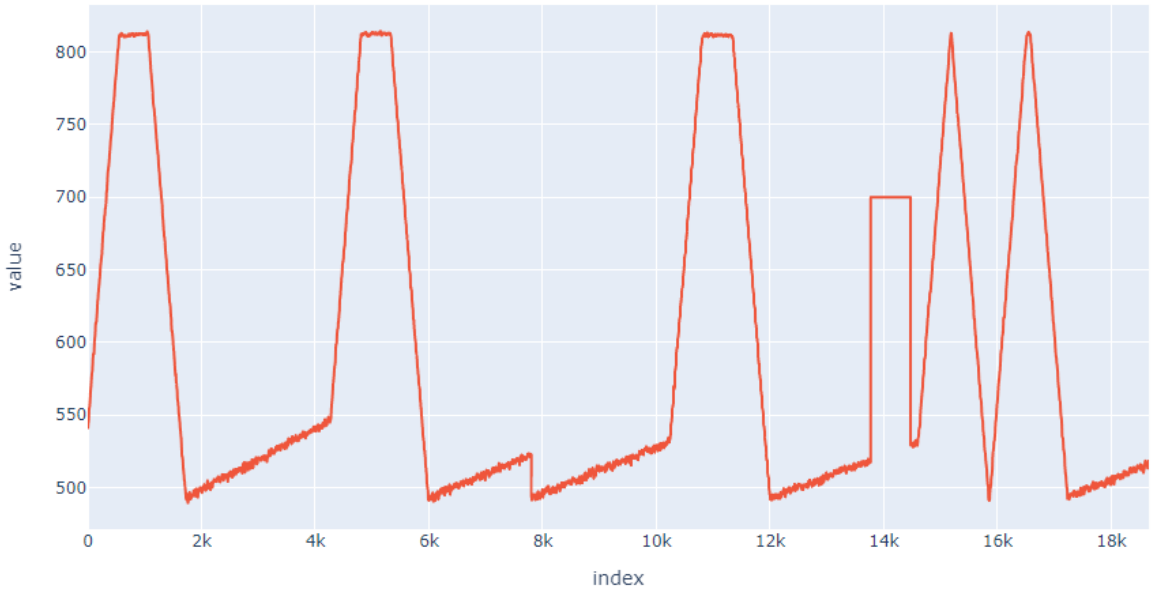
Нормальное поведение



Аномалии

Обнаружение аномалий в SWaT 2015

Sensors and actuators

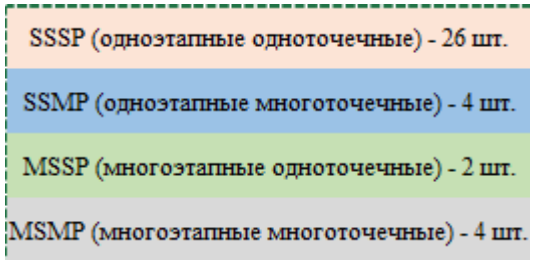


LIT-101

Анализ работы датчиков и оборудования на испытательном стенде на основе данных

Процесс сбора данных длился в общей сложности 11 дней.
Стенд функционировал в режиме нон-стоп (24 часа в сутки) в течение всего 11-дневного периода. Первые 7 из 11 дней установка работала в нормальном режиме, без осуществления атак. Нападения были совершены в течение оставшихся четырех дней - с 28 декабря 2015 10:29:14 по 02 января 2016 13:41:11
Интервал записи данных = 1 сек. В ходе процесса сбора данных было совершено в общей сложности 36 атак.

Еще две атаки 37, 38 - размечены в датасете Attack, но отсутствуют в описании к датасету



№	Время атаки	тип атаки	точка атаки	в чем состоит
1	['28.12.2015 12:08:05', ' 28.12.2015 12:15:12']	SSSP (одноэтапные одноточечные)	LIT-301	Уровень воды между L и H (уровень воды повышается выше HH 1200)
2	['28.12.2015 13:09:45', '28.12.2015 13:25:54']	SSSP (одноэтапные одноточечные)	DPIT-301	DPIT < 40 kpa, устанавливается==45
3	['28.12.2015 14:16:01', '28.12.2015 14:18:41']	SSSP (одноэтапные одноточечные)	FIT-401	FIT-401 выше 1 (устанавливается значение FIT-401 <0.7)
4	['29.12.2015 18:29:59', '29.12.2015 18:41:40']	SSSP (одноэтапные многоточечные)	MV-101, LIT-101	MV-101 открыт; LIT-101 между L и H (постоянно держится MV101 включенным; значение LIT-101 устанавливается равным 700 мм)
5	['29.12.2015 22:54:56', '29.12.2015 23:02:41']	MSMP (многоэтапные многоточечные)	UV-401, AIT-502, P-501	UV-401 включен; AIT-502 <150; P-501 открыт (останавливается UV-401; значение AIT502 устанавливается равным 150; не позволяется P-501 выключиться)

Продолжительность атак варьируется в зависимости от типа. Несколько атак, каждая продолжительностью десять минут, выполняются последовательно с промежутком в 10 минут между последовательными атаками.
Некоторые из атак выполняются с возможностью стабилизации системы перед последующей атакой. Продолжительность стабилизации системы варьируется в зависимости от атак.
Некоторые из атак оказывают более сильное влияние на изменение системы и требуют большего времени для ее стабилизации. Более простые атаки - те, которые влияют на скорость потока, требуют меньше времени для стабилизации. Кроме того, некоторые атаки воздействуют на систему не сразу.

Результаты, выводы, планы

- Интерпретируемость важна в детекции аномалий
- В reconstruction-based методах обнаружения аномалий можно найти входные признаки, уменьшающие ошибку реконструкции
- Для дискретных сигналов можно использовать модель RBM
- Объяснимость связана с глубоким изучением предметной области, использованием систем организации знаний (онтологии, таксономии)

Будущие исследования

- Применение подхода для других наборов данных и стендов
- Поиск или изобретение метрики оценки качества ретроспективного анализа событий (форензики)
- Моделирование, создание стенда для экспериментов
- Использование систем организации знаний из доменной области (ontology, taxonomy, etc…). MITRE, CVE, …

Спасибо за внимание!

С уважением,
Чернышов Юрий
ychernyshov@ussc.ru
@yuchernyshov

Some known facts from cybersecurity

02/11/88 Morris warm attacks ArpaNet (cybersecurity birthday)

ThyssenKrupp: the virus melted the stove

Evraz, Ryuk malware decreased steal production on 4.7% (appr $12M)

Norsk Hydro, LockerGoga. $41M

Some other examples of datasets

6.1 Simulated Data

We first investigated whether the proposed estimation algorithm can distinguish the contributing dimensions accurately with simulated data. The training data were 1000-dimensional data generated through a simulation as follows:

$$x_{i+10j} = \begin{cases} N(1000, 200^2) & (j = 0) \\ (1 + 0.1 * j) * x_i^2 + N(\beta, \gamma^2) & (j = 1, ..., 99) \end{cases} \quad (12)$$

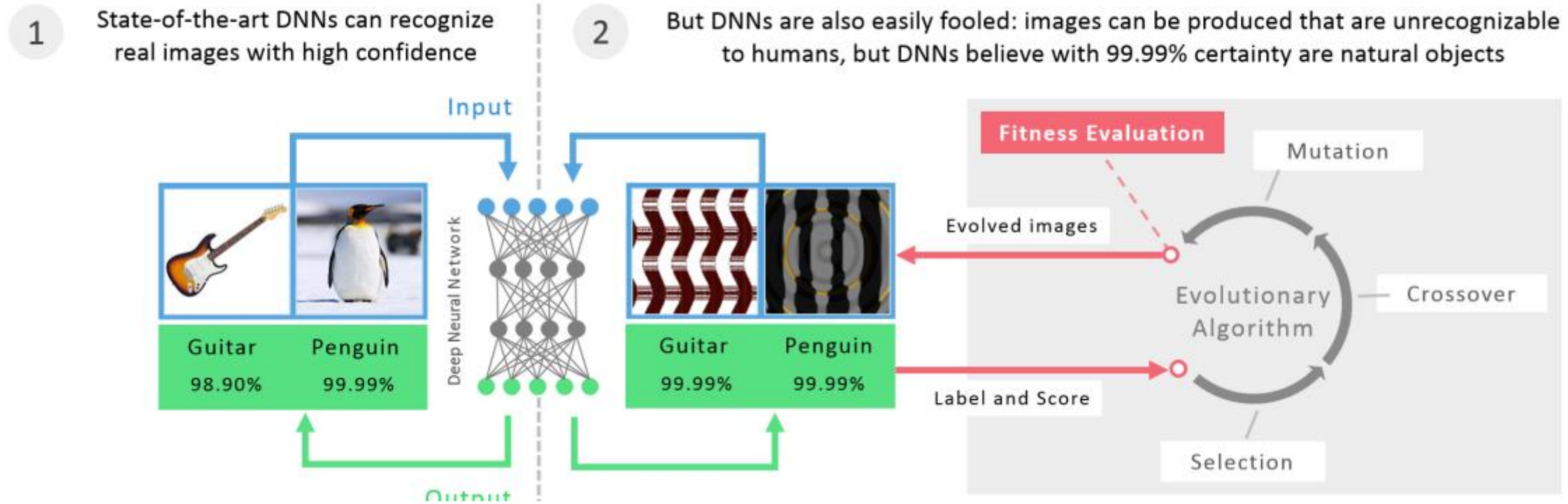
for $i = 1, ..., 10$,

where $N(\mu, \sigma^2)$ is a random variable of a normal distribution with mean μ and variance σ^2. With the data, it was assumed that the surveillance target is composed of ten independent components and there are one random value (such as the number of server requests) and 99 correlation values with noise (such as server load) for each component. For generating test data, we first generated data similar to the training data and randomly chose i from $i = 1, ..., 10$. After that, we randomly chose n_f values from $x_{i+10j}, j = 1, ..., 99$ and increased (or decreased) the values by r-fold, where r was also randomly determined by

Next, we evaluated our estimation algorithm through the NSL-KDD Dataset [27]. The dataset consist of 41 feature values about each communication and the communications attribute to five classes. One is normal, and the other four are types of attacks: denial of service (DoS), remote to user (R2L), user to root (U2R), and probing. We used 67,343 normal communications in the training dataset as

Y.Ikeda, K.Ishibashi, Yu.Nakano, K.Watanabe, R.Kawahara. Anomaly Detection and Interpretation using Multimodal Autoencoder and Sparse Optimization **https://arxiv.org/pdf/1812.07136.pdf**

Adversarial example



A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.