인공지능 시대의 보안 패러다임의 변화

NAVER

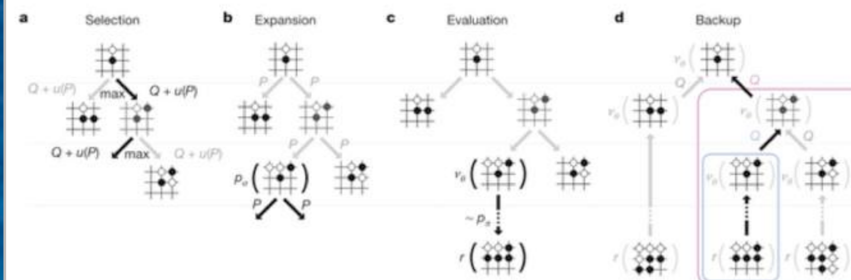
2017년 4월 20일

조상현 (bungae@gmail.com), Ph.D.
Leader / Security
고려대학교 정보보호대학원 겸임교수

AI,Machine Learning and Cyber Security

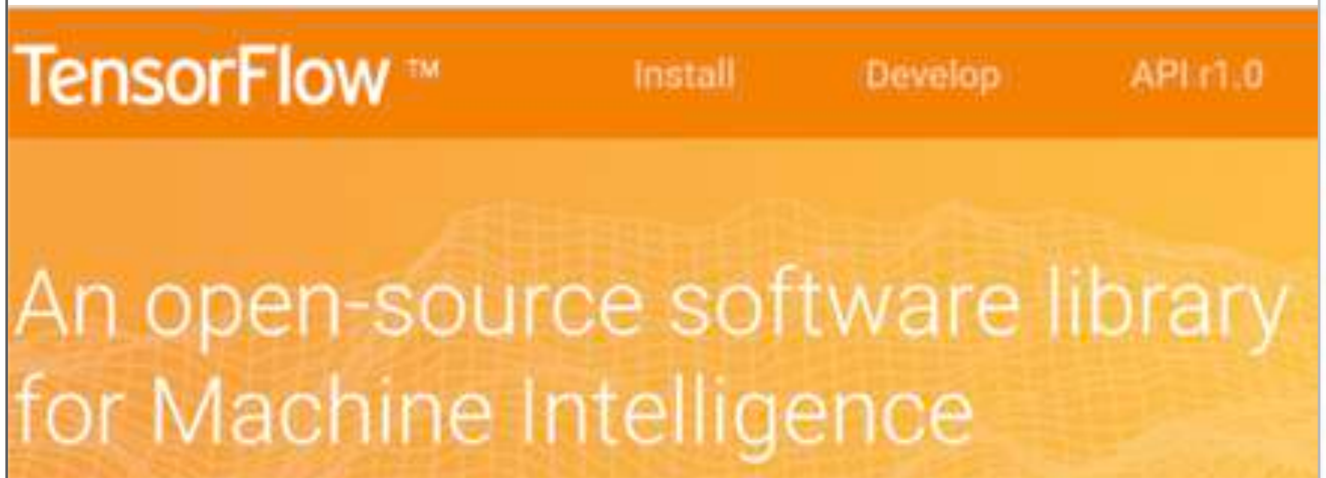
AI is a subfield of computer science making intelligent machines, while machine learning is a subset of AI and is typically associated with statistics, data mining and predictive analytics.

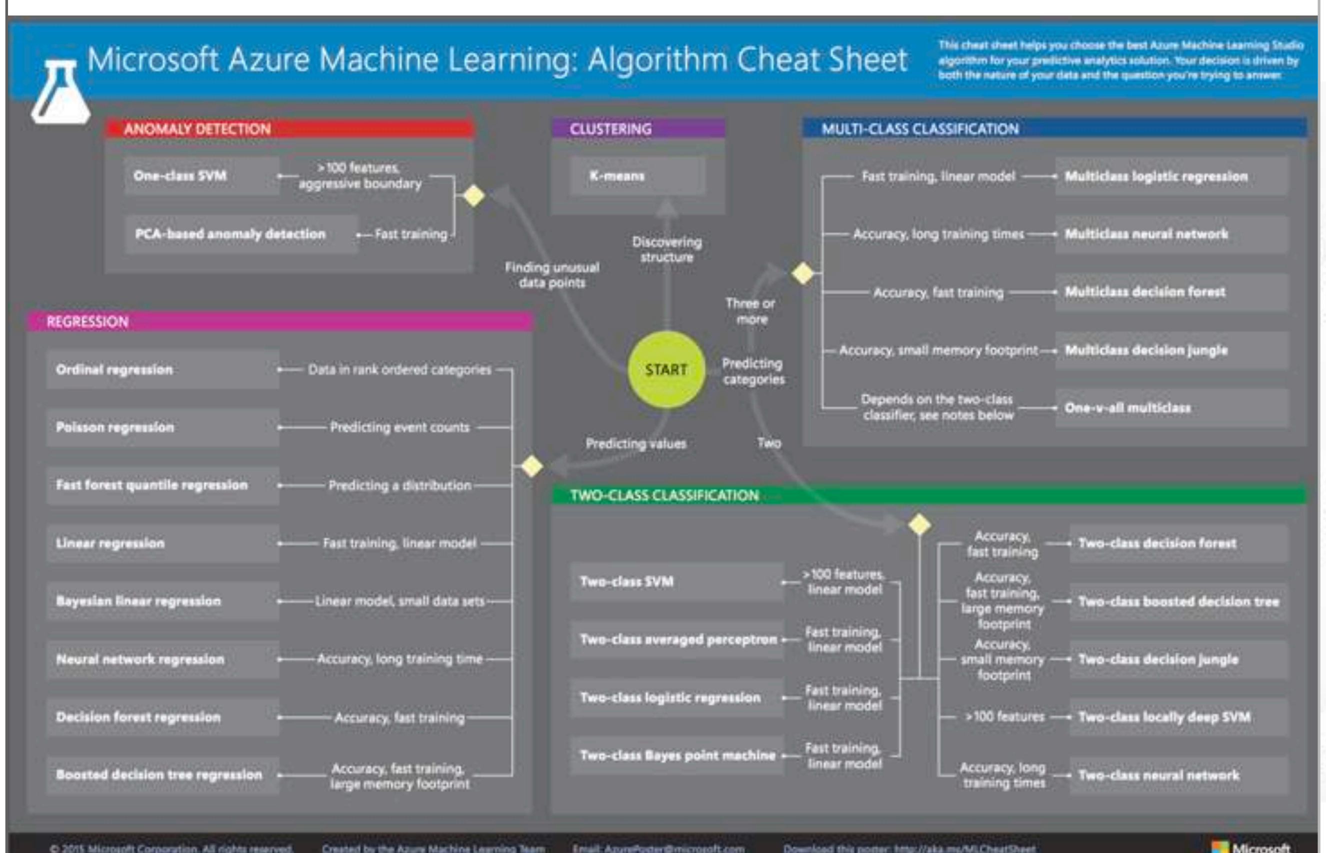
Machine learning is the actual implementation of the methods (algorithms) that support AI.

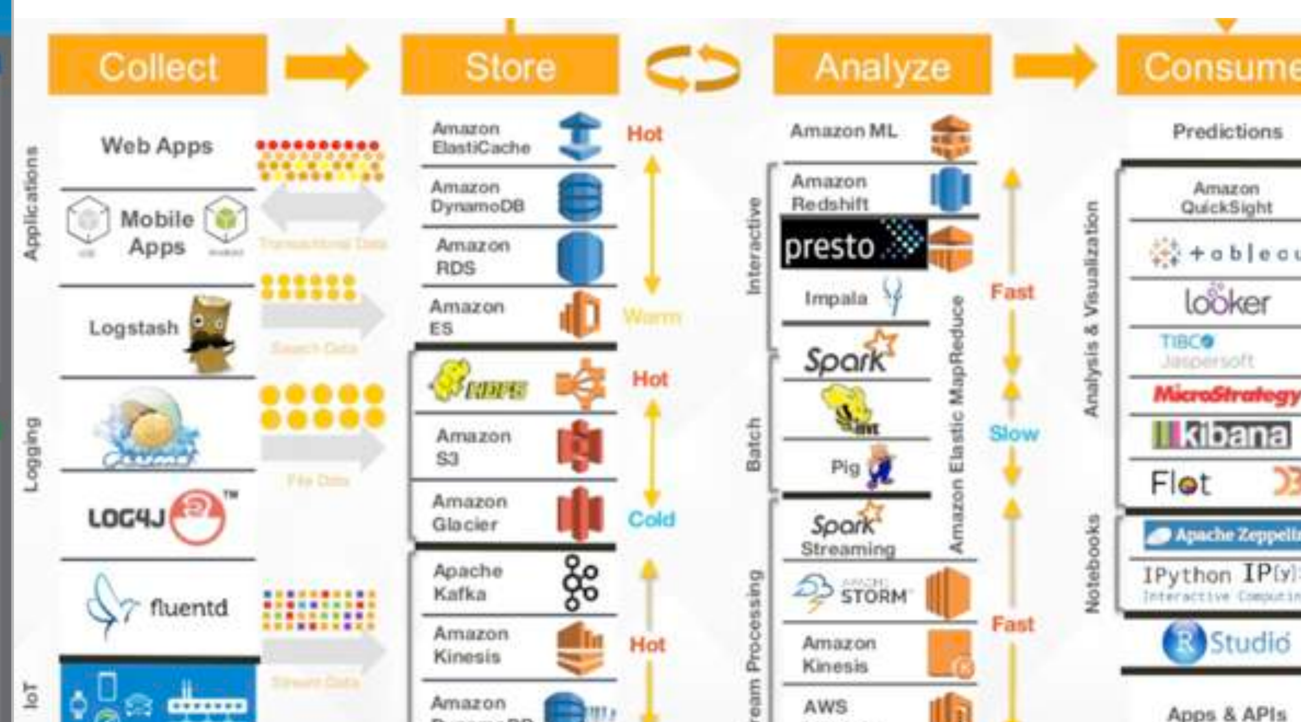

a
Selection
b
Expansion
c
Evaluation
d
Backup
Q + u(P)
max

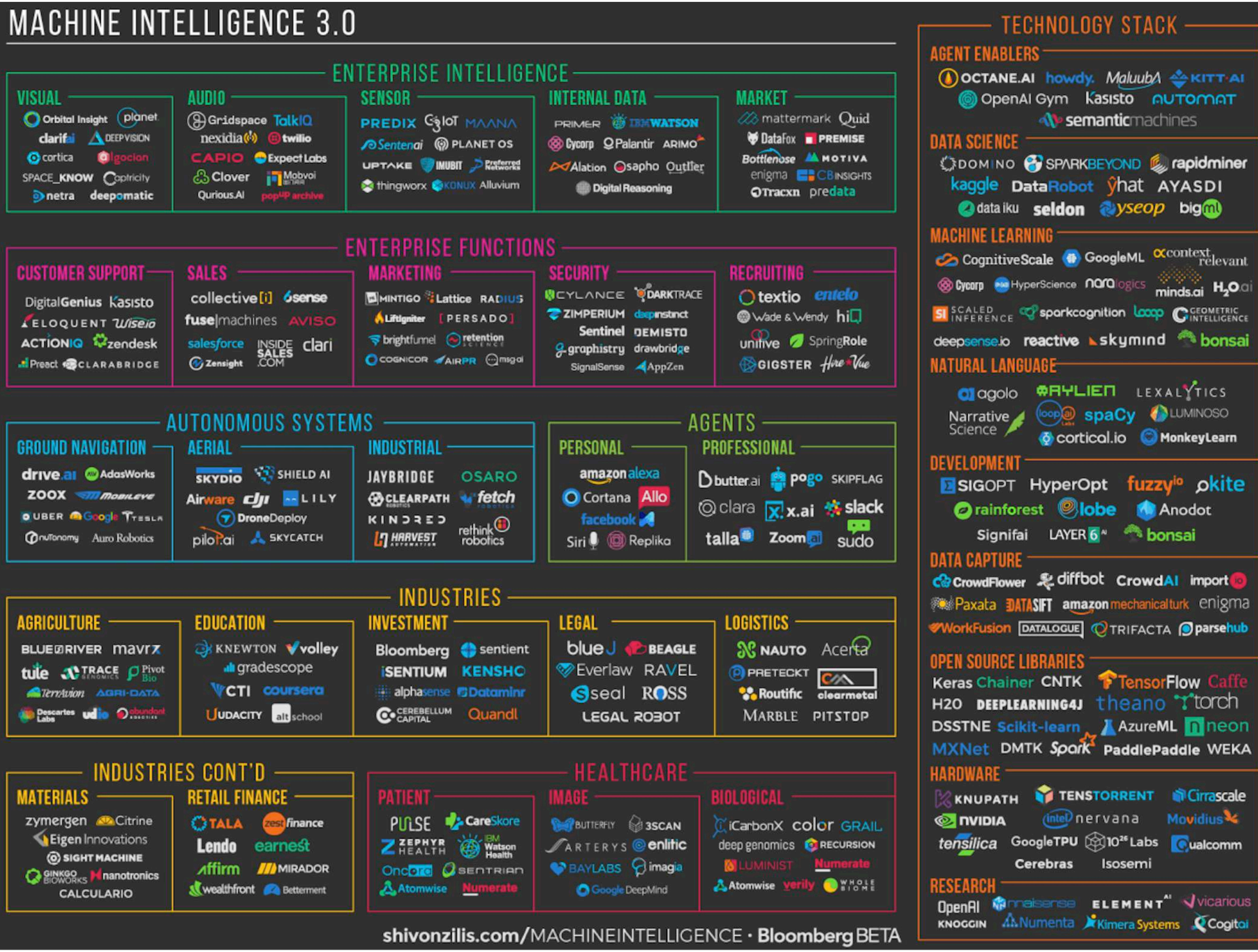

IDENTITY PROTECTION
Name:
Password:


Back
Place Your Finger
Lift and rest the edge of your finger on the Home button repeatedly.


TensorFlow ™
Install
Develop
API r1.0
An open-source software library for Machine Intelligence


Microsoft Azure Machine Learning: Algorithm Cheat Sheet
This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.
ANOMALY DETECTION
One-class SVM
>100 features, aggressive boundary
PCA-based anomaly detection
Fast training
CLUSTERING
K-means
Discovering structure
Finding unusual data points
MULTI-CLASS CLASSIFICATION
Fast training, linear model
Multiclass logistic regression
Accuracy, long training times
Multiclass neural network
Accuracy, fast training
Multiclass decision forest
Accuracy, small memory footprint
Multiclass decision jungle
Depends on the two-class classifier, see notes below
One-v-all multiclass
Three or more
Predicting categories
START
Predicting values
Two
REGRESSION
Ordinal regression
Data in rank ordered categories
Poisson regression
Predicting event counts
Fast forest quantile regression
Predicting a distribution
Linear regression
Fast training, linear model
Bayesian linear regression
Linear model, small data sets
Neural network regression
Accuracy, long training time
Decision forest regression
Accuracy, fast training
Boosted decision tree regression
Accuracy, fast training, large memory footprint
TWO-CLASS CLASSIFICATION
Two-class SVM
>100 features, linear model
Two-class averaged perceptron
Fast training, linear model
Two-class logistic regression
Fast training, linear model
Two-class Bayes point machine
Fast training, linear model
Accuracy, fast training
Two-class decision forest
Accuracy, fast training, large memory footprint
Two-class boosted decision tree
Accuracy, small memory footprint
Two-class decision jungle
>100 features
Two-class locally deep SVM
Accuracy, long training times
Two-class neural network
© 2015 Microsoft Corporation. All rights reserved.
Created by the Azure Machine Learning Team
Email: AzurePoster@microsoft.com
Download this poster: http://aka.ms/MLCheatSheet
Microsoft


Collect
Store
Analyze
Consume
Applications
Web Apps
Mobile Apps
Logstash
Logging
LOG4J
fluentd
IoT
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
Amazon S3
Amazon Glacier
Apache Kafka
Amazon Kinesis
Amazon DynamoDB
Hot
Warm
Hot
Cold
Hot
Interactive
Batch
Stream Processing
Amazon ML
Amazon Redshift
presto
Impala
Spark
Pig
Spark Streaming
STORM
Amazon Kinesis
AWS Lambda
Amazon Elastic MapReduce
Fast
Slow
Fast
Analysis & Visualization
Notebooks
Predictions
Amazon QuickSight
tableau
looker
TIBCO Jaspersoft
MicroStrategy
kibana
Flot
Apache Zeppelin
IPython
RStudio
Apps & APIs



MACHINE INTELLIGENCE 3.0
ENTERPRISE INTELLIGENCE
VISUAL
Orbital Insight planet clarifai DEEPVISION cortica algocian SPACE_KNOW Cypricity netra deepomatic
AUDIO
Gridspace TalkIQ nexidia twilio CAPIO Expect Labs Clover Mobvoi Qurious.AI popUP archive
SENSOR
PREDIX C3 IoT MAANA Sentenai PLANET OS UPTAKE IMUBIT Preferred Networks thingworx KONUX Alluvium
INTERNAL DATA
PRIMER IBM WATSON Cycorp Palantir ARIMO Alation sapho Outlier Digital Reasoning
MARKET
mattermark Quid DataFox PREMISE Bottlenose MOTIVA enigma CB INSIGHTS Tracxn predata
ENTERPRISE FUNCTIONS
CUSTOMER SUPPORT
DigitalGenius Kasisto ELOQUENT Wise.io ACTIONIQ zendesk Preact CLARABRIDGE
SALES
collective[i] 6sense fuse|machines AVISO salesforce INSIDE SALES .COM clari Zensight
MARKETING
MINTIGO Lattice RADIUS Liftigniter [PERSADO] brightfunnel retention science COGNICOR AIRPR msg.ai
SECURITY
CYLANCE DARKTRACE ZIMPERIUM deepinstinct Sentinel DEMISTO graphistry drawbridge SignalSense AppZen
RECRUITING
textio entelo Wade & Wendy hiQ unitive SpringRole GIGSTER HireVue
AUTONOMOUS SYSTEMS
GROUND NAVIGATION
drive.ai AdasWorks ZOOX MOBILEYE UBER Google TESLA nuTonomy Auro Robotics
AERIAL
SKYDIO SHIELD AI Airware dji LILY DroneDeploy pilot.ai SKYCATCH
INDUSTRIAL
JAYBRIDGE OSARO CLEARPATH ROBOTICS fetch robotics KINDRED HARVEST AUTOMATION rethink robotics
AGENTS
PERSONAL
amazon alexa Cortana Allo facebook M Siri Replika
PROFESSIONAL
butter.ai pogo SKIPFLAG clara x.ai slack talla Zoom.ai sudo
INDUSTRIES
AGRICULTURE
BLUE RIVER mavrx tule TRACE GENOMICS Pivot Bio TerrAvion AGRI-DATA Descartes Labs udio abundant robotics
EDUCATION
KNEWTON volley gradescope CTI coursera UDACITY altschool
INVESTMENT
Bloomberg sentient iSENTIUM KENSHO alphasense Dataminr CEREBELLUM CAPITAL Quandl
LEGAL
blue J BEAGLE Everlaw RAVEL seal ROSS LEGAL ROBOT
LOGISTICS
NAUTO Acerta PRETECKT Routific clearmetal MARBLE PITSTOP
INDUSTRIES CONT'D
MATERIALS
zymergen Citrine Eigen Innovations SIGHT MACHINE GINKGO BIOWORKS nanotronics CALCULARIO
RETAIL FINANCE
TALA zest finance Lendo earnest affirm MIRADOR wealthfront Betterment
HEALTHCARE
PATIENT
PULSE CareSkore ZEPHYR HEALTH IBM Watson Health Oncora SENTRIAN Atomwise Numerate
IMAGE
BUTTERFLY 3SCAN ARTERYS enlitic BAYLABS imagia Google DeepMind
BIOLOGICAL
iCarbonX color GRAIL deep genomics RECURSION LUMINIST Numerate Atomwise verily WHOLE BIOME
shivonzilis.com/MACHINEINTELLIGENCE · Bloomberg BETA
TECHNOLOGY STACK
AGENT ENABLERS
OCTANE.AI howdy. Maluuba KITT.AI OpenAI Gym Kasisto AUTOMAT semanticmachines
DATA SCIENCE
DOMINO SPARKBEYOND rapidminer kaggle DataRobot yhat AYASDI dataiku seldon yseop bigml
MACHINE LEARNING
CognitiveScale GoogleML context relevant Cycorp HyperScience nara logics minds.ai H2O.ai SCALED INFERENCE sparkcognition loop GEOMETRIC INTELLIGENCE deepsense.io reactive skymind bonsai
NATURAL LANGUAGE
agolo AYLIEN LEXALYTICS Narrative Science loop ai Labs spaCy LUMINOSO cortical.io MonkeyLearn
DEVELOPMENT
SIGOPT HyperOpt fuzzy.io kite rainforest lobe Anodot Signifai LAYER 6 AI bonsai
DATA CAPTURE
CrowdFlower diffbot CrowdAI import.io Paxata DATASIFT amazon mechanical turk enigma WorkFusion DATALOGUE TRIFACTA parsehub
OPEN SOURCE LIBRARIES
Keras Chainer CNTK TensorFlow Caffe H2O DEEPLEARNING4J theano torch DSSTNE Scikit-learn AzureML neon MXNet DMTK Spark PaddlePaddle WEKA
HARDWARE
KNUPATH TENSTORRENT Cirrascale NVIDIA intel nervana Movidius tensilica GoogleTPU 10^{26} Labs Qualcomm Cerebras Isosemi
RESEARCH
OpenAI maluuba ELEMENT AI vicarious KNOGGIN Numenta Kimera Systems Cogitai

There Is No Silver Bullet

Tool이 Silver Bullet 인 것으로 생각하고, Tool만 도입하면 다 해결될 것이라고도 생각한다.

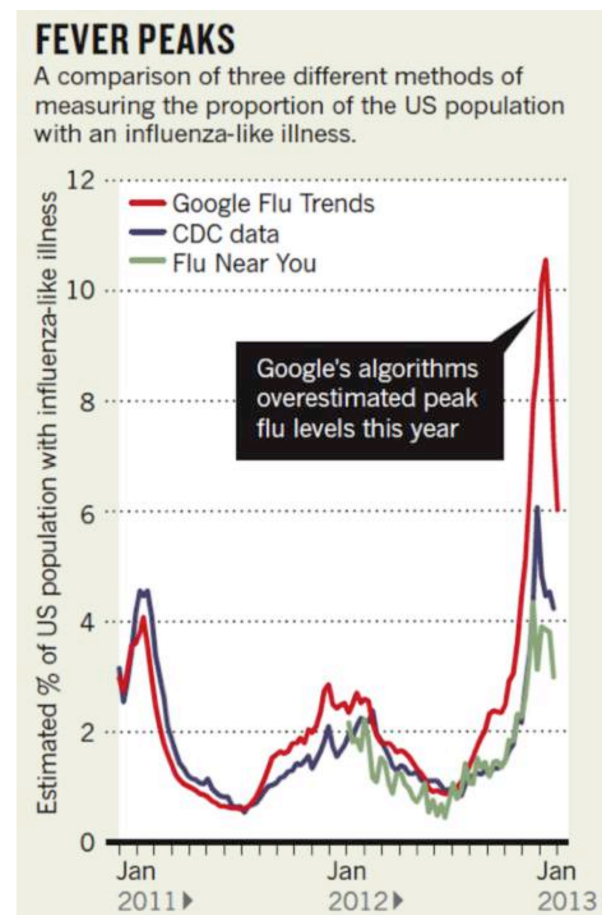
매니저들 마저 Tool이 도입되어 있으니, 이제는 문제가 발생하면 안된다고 말한다.

nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]



출처 : 구글트렌드가 준 교훈: 빅데이터 분석의 함정(The Parable of Google Flu: Traps in Big Data Analysis) , [SCIENCE 2014]

7 Misconceptions of AI, Machine Learning and Cybersecurity

1	Machine learning is a new technology
2	Artificial intelligence = machine learning
3	Machine learning is only summarising data
4	Machine learning replaces traditional anti-malware technologies
5	Machine learning can’t predict unseen events
6	AI will automate us out of our jobs
7	Nobody needs human security experts anymore

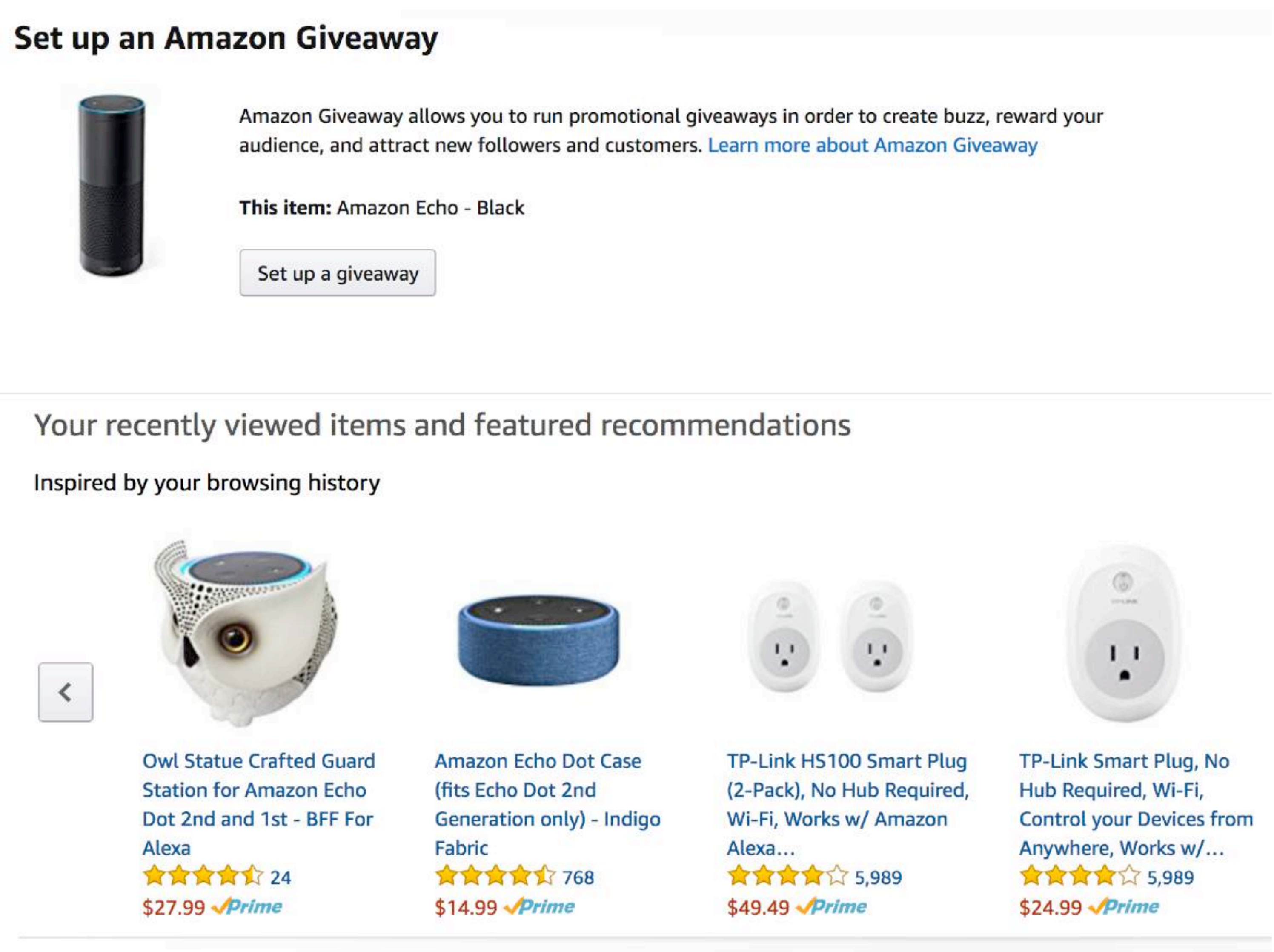
출처 : 7 Misconceptions of AI, Machine Learning and Cybersecurity

ML in Cyber Security

- Biometric Recognition
- Network and System Security
  - IDS, IPS, Botnet / Proxies Detection
- Anomaly Detection
  - Fraud Detection, Game Bot Detection
- Malware classification
- Security policy management (SPM)
- Information leak checking

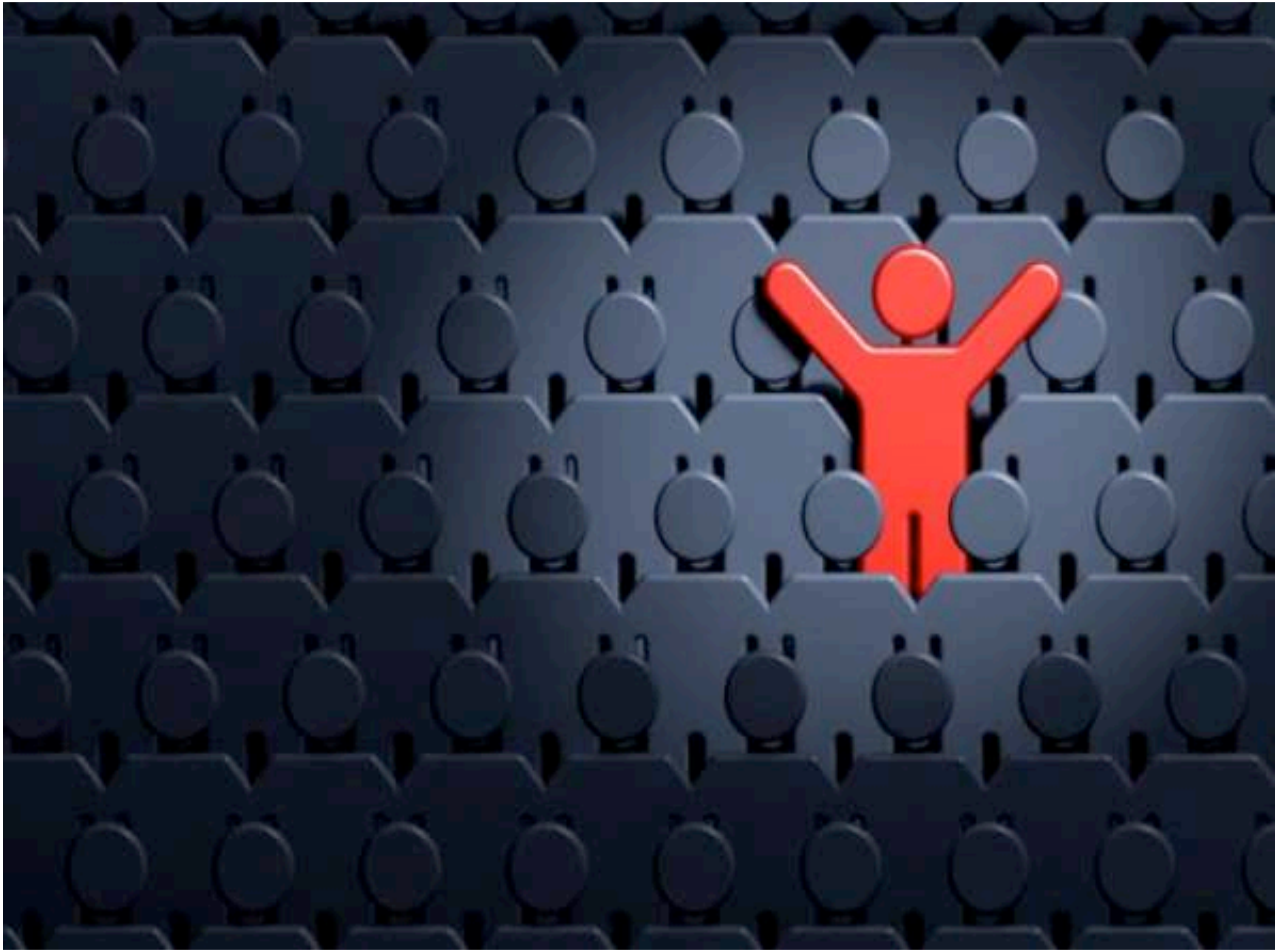
In Cyber Security

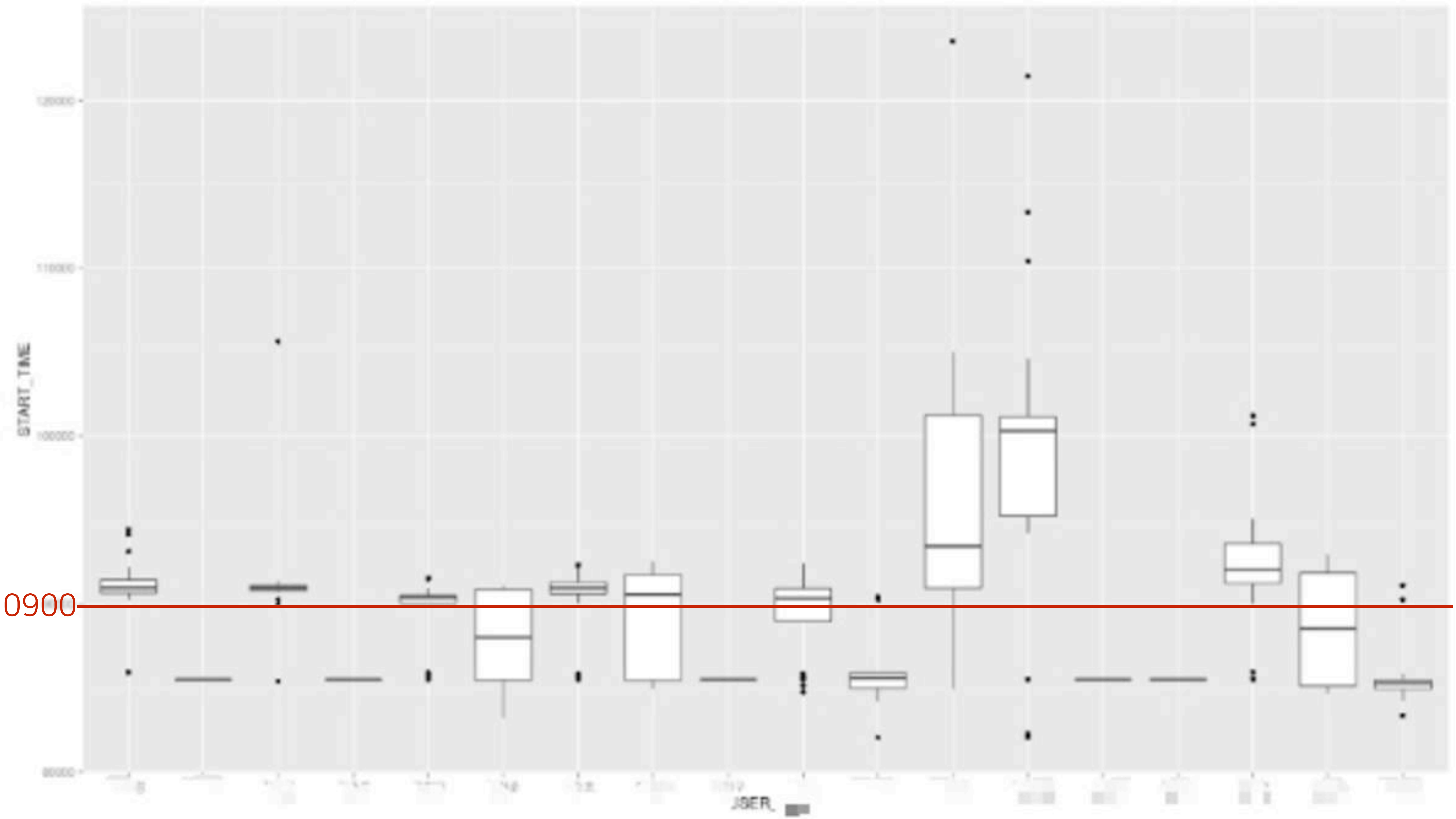
우리에게 필요한 추천 시스템은?




Frequency/likelihood
Severity/impact

Something unususal
Something rare




START_TIME
0900

In Cyber Security

The domain of cybersecurity is characterized by

- weak signals
- intelligent actors
- a large attack surface
- a huge number of variables.

ML을 사용한다고 해도 고된 일에 의존하는 것으로 부터 더 나아진다는 보장은 없다.

학습 (learning)의 정의

하나의 문제를 수행한 후에

그 추론과정에서 얻은 경험을 바탕으로

시스템의 지식을 수정 및 보완하여,

다음에 그 문제나 또는 비슷한 문제를 수행할 때에는

처음보다 더 효율적이고 효과적으로 문제를 해결할 수 있는 적응성

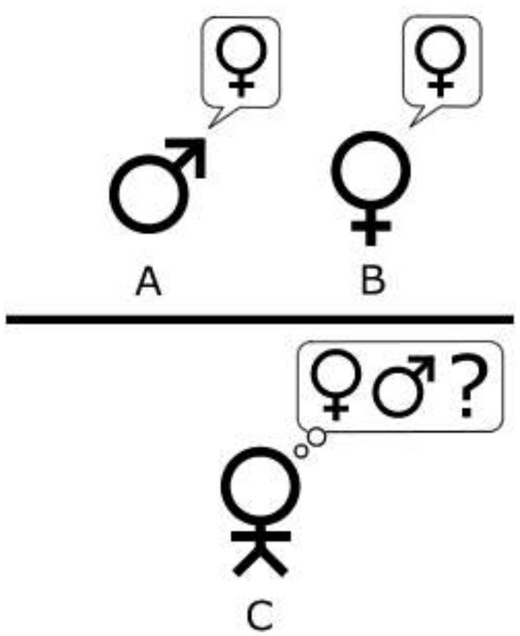
기계가 지능을 가지고 있는가?

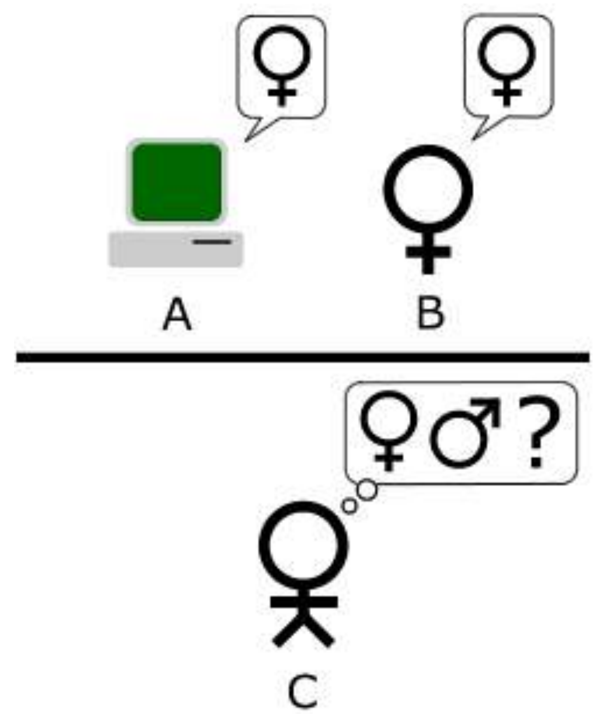
Turing test ：

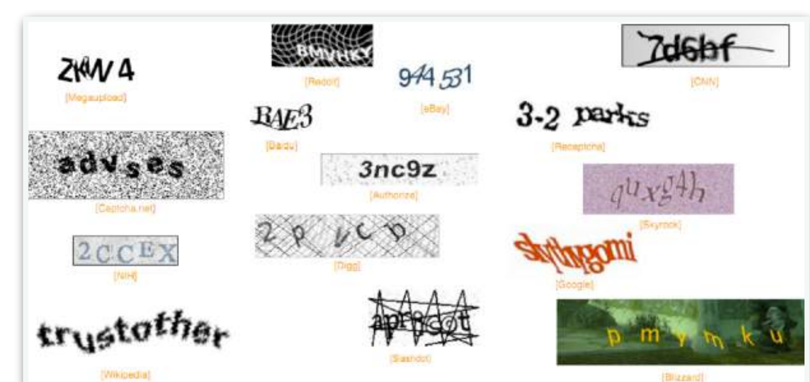
a test of a <u>machine</u>'s ability to demonstrate intelligence

Imitation Game

참고 : Captcha : Completely Automated Public Turing test to tell Computers and Humans Apart

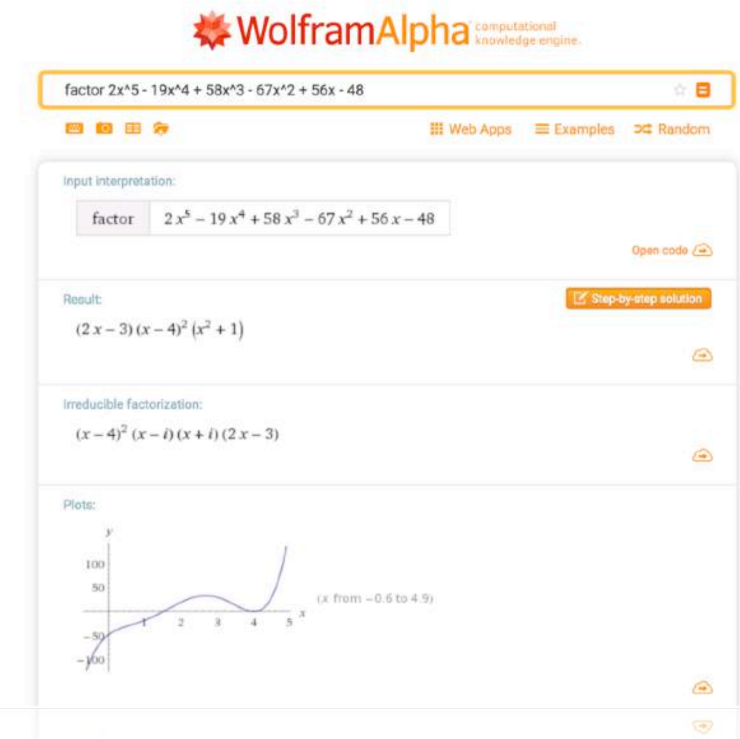


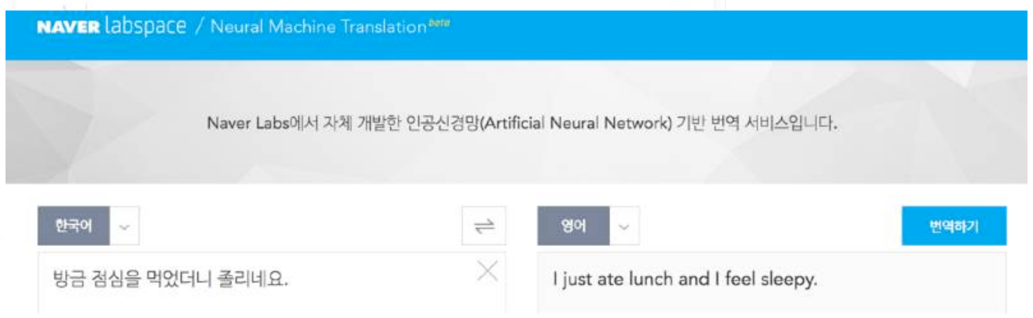




기계가 지능을 가지고 있는가?

- 수학 문제를 풀 수 있는가?
- 언어를 번역할 수 있는가?
- 물건을 식별할 수 있는가?
- 게임을 할 수 있는가?
- 경험으로 부터 배울 수 있는가?
- 목표를 달성하기 위해 계획을 수립할 수 있는가?





기계 학습(Machine Learning)

컴퓨터에게 배울 수 있는 능력, 코드로 정의하지 않은 동작을 실행하는 능력에 대한 연구 분야

표현과 일반화

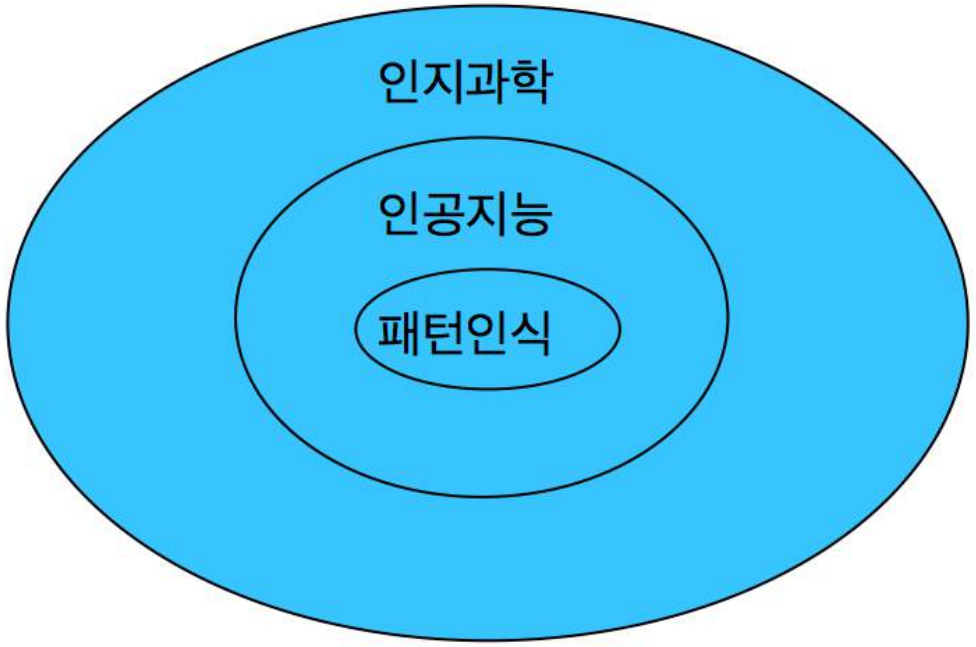
어떤 특징을 뽑아서, 어떤 방법을 쓸 것인가?

신경망 (Neural Network), 데이터마이닝 (Data Mining), 의사결정 트리 (Decision Tree), 유전알고리즘 (Genetic Algorithm), 사례기반 추론 (Case Based Reasoning), 패턴 인식 (Pattern Recognition), 강화 학습 (Reinforcement learning),딥 러닝 (Deep Learning)

기계 학습의 유형

구분	설명	예시
분류(Classification)	대상 객체에 대한 특정한 클래스 할당	정상/불량, 합격/불합격, 공격/정상
군집화 (Clustering)	복수 개의 그룹들로 조직화	생명체를 종으로 그룹화
회귀 (Regression)	일반화, 미래에 대한 예측	주식 배당 가치 예측
서술 (Description)	객체를 몇 개의 원형(protype)으로 표현	심전도(ECG검사)에서 생체 신호를 P,Q,R,S,T로 표현

패턴 인식



계산이 가능한 기계적인 장치(컴퓨터)가 어떠한 대상을 인식하는 문제를 다루는 인공 지능의 한 분야

복잡한 신호의 몇 가지 표본(sample)과 이들에 대한 정확한 결정(decision)이 주어질 때,
연이어 주어지는 미래 표본들에 대하여 자동적으로 결정을 내리게 하는 것

관측치 x에 이름 w를 부여하는 과정
어떻게 X와 Y사이의 관계를 모델링할 수 있는가?
X : 이미 알고 있는 변수, 인스턴스의 특징값
Y : 목표 변수, 샘플의 부류

출처: 패턴인식개론
http://www.hanbit.co.kr/book/look.html?isbn=978-89-7914-632-5

특징과 패턴

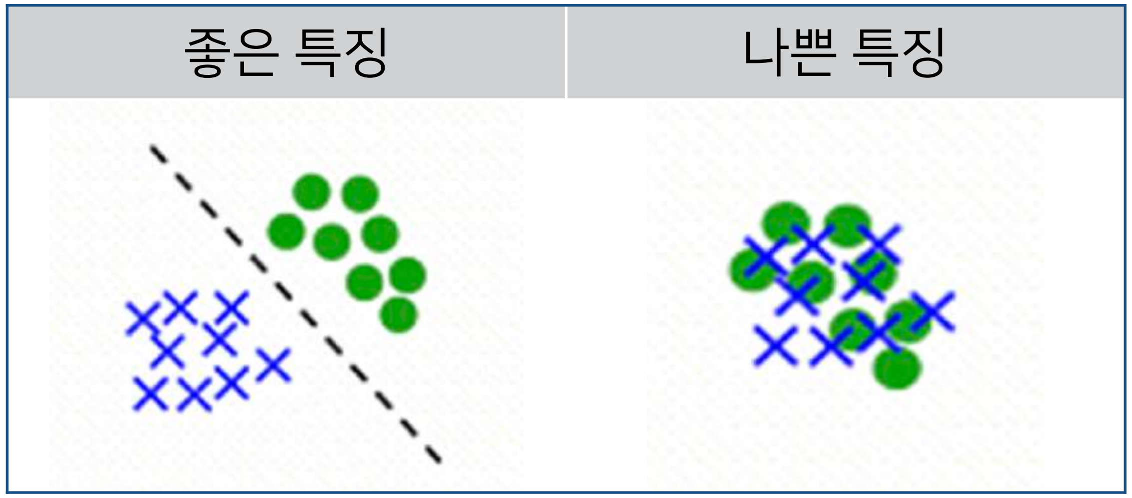
- 특징(feature)

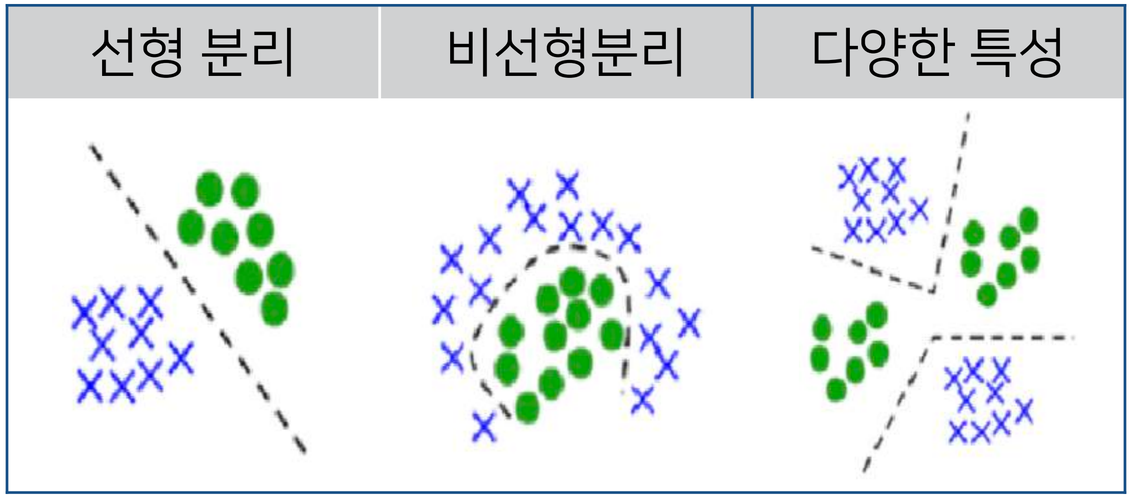
    어떤 객체(object)가 가지고 있는, 고유의 분별 가능한

    측면(aspect), 양(quantity), 질(quality) 혹은 특성(characteristic)

- 패턴(pattern)

    개별 객체의 특색(traits) 이나 특징(features)들의 집합





패턴 인식의 접근 방법

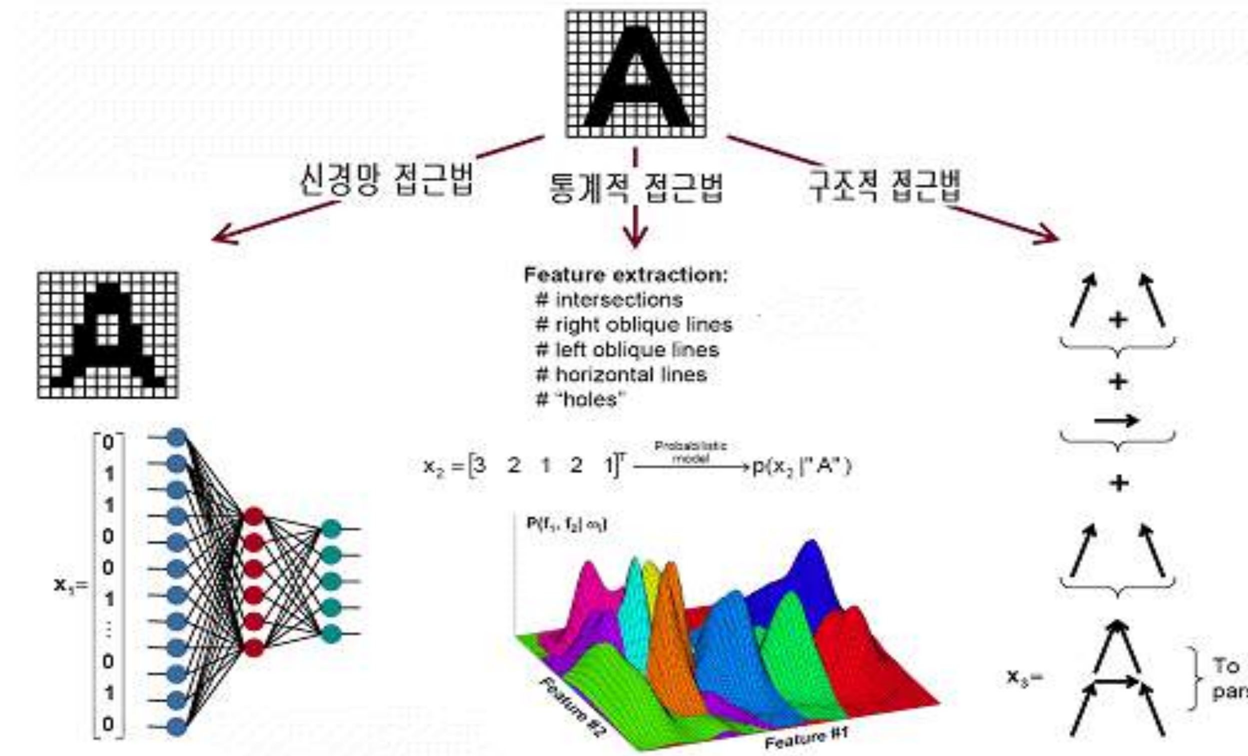
특징 1: 수직선의 개수(V)

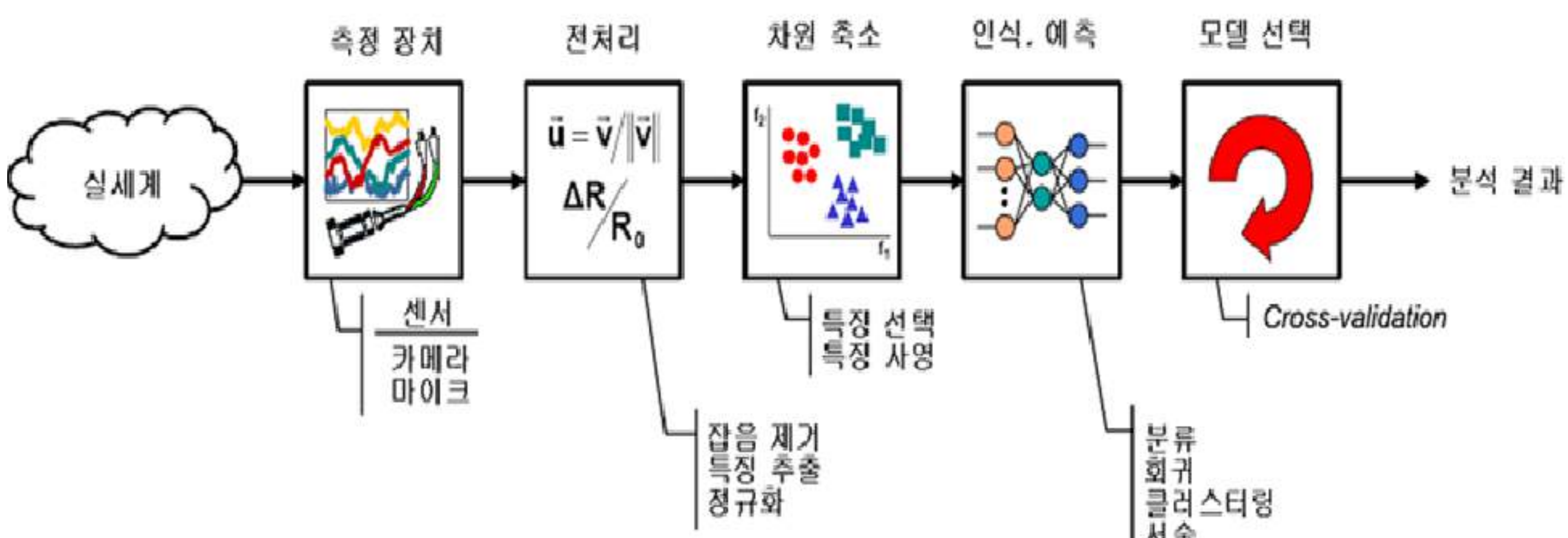
특징 2: 수평선의 개수(H)

특징 3: 기울어진 수직선(O)

특징 4: 커브의 개수(C)

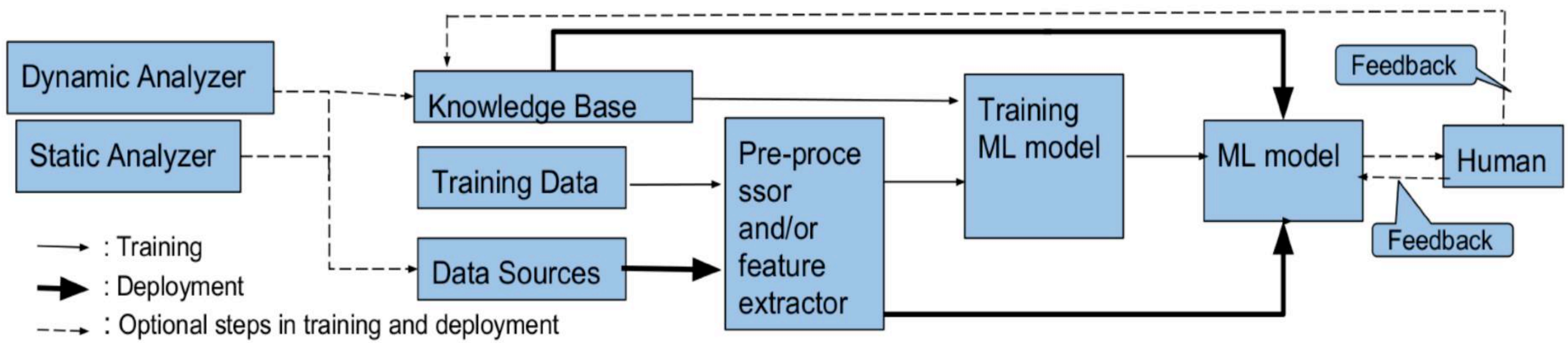
문자	특징 V	H	O	C
L	1	1	0	0
P	1	0	0	1
O	0	0	0	1
E	1	3	0	0
Q	0	0	1	1





출처: 패턴인식개론
http://www.hanbit.co.kr/book/look.html?isbn=978-89-7914-632-5

학습 시스템



Knowledge base

- 알려진 정상, 비정상 케이스들 (Blacklist, Whitelist, Malware Signatures등)

Data sources

- 관련 있는 데이터들 (오프라인 혹은 라이브 데이터)

Training data

- classifier의 학습에 필요한 데이터

Pre-processor and feature extractor

- 데이터 소스로 부터 특징을 뽑아냄

출처 : SoK: Applying Machine Learning in Security - A Survey

ML in Security : History

1987: Denning published “An Intrusion Detection System” , first framing security as a learning problem

1998: DARPA IDS design challenge

1999: KDD Cup IDS design challenge

…

2008: CCS hosted the 1st AISec workshop. Continues to operate each year.

…

2011: “Adversarial Machine Learning”published in 4th AISec

…

2014: KDD hosted its 1st “Security & Privacy” session in the main conference program

2014: ICML hosted its 1st, and so far the only workshop on Learning, Security, and Privacy(LSP)

2016: AAAI hosted its 1st Artificial Intelligence for Cyber Security workshop(AISC)

출처 : SoK: Applying Machine Learning in Security - A Survey

Intrusion Detection Model (1987)

An Intrusion-Detection Model

DOROTHY E. DENNING

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. SE-13, NO. 2, FEBRUARY 1987, 222-232.

Host-based Intrusion Detection

시스템의 표준적인 오퍼레이션을 모니터링 하여

침입으로 의심할 만큼 충분히 이상한 행위 발견 목표

- 로그인(Logins)
- 명령어/프로그램 실행(command and program executions)
- 파일/장치 접근(file and device accesses)

Intrusion Detection Model (1987)

관찰된 값들로 통계적인 Metric과 Model로 만들어짐

Metric	어떤 것을 측정할 것인가? 일정 기간 동안 측정된 양의 값(quantitative measure) 예) event counter, interval timer, resource measure
Model	어떻게 판단할 것인가? 새로운 관찰이 이상한 지 아닌지 판단의 근거 - Operational Model : 제한된 한계치(limit) 이상의 관찰 - Mean and Standard Deviation Model : 특정 범위를 벗어난 확률 - Multivariate Model : 2개 이상의 모델 사용 - Markov Process Model : 이전 상태에 근거하여 다음 상태의 발생이 극히 낮을 경우 - Time Series Model : 이벤트의 발생 시간 간격을 볼 때, 해당 시점에 발생할 확률이 극히 낮을 경우

CYBER SYSTEMS AND TECHNOLOGY

DARPA Intrusion Detection Evaluation ›

- Data Sets ›
  - 1998 ›
  - 1999 ›
  - 2000 ›
- Documentation

DARPA Intrusion Detection Data Sets

Data Sets Overview

The Cyber Systems and Technology Group (formerly the DARPA Intrusion Detection Evaluation Group) of MIT Lincoln Laboratory, under Defense Advanced Research Projects Agency (DARPA ITO) and Air Force Research Laboratory (AFRL/SNHS) sponsorship, has collected and distributed the first standard corpora for evaluation of computer network intrusion detection systems. We have also coordinated, with the Air Force Research Laboratory, the first formal, repeatable, and statistically significant evaluations of intrusion detection systems. Such evaluation efforts have been carried out in 1998 and 1999.

DARPA IDS TEST DATA SET (1998)

MIT Lincoln Labs에서는 DARPA Intrusion Detection Evaluation Program의 일환으로 미 공군 네트워크 트래픽을 수집함.

9주간의 압축된 TCP dump 데이터는 4GB에 육박, 5백만개의 connection record를 가지고 있었음.

테스트 데이터는 2주간의 기록으로 2백만개의 connection record로 구성됨

테스트 데이터에는 훈련 데이터와는 다른 확률 분포를 가지거나, 관찰되지 않았던 공격이 포함되어 있음

각 connection record는 100 바이트 크기로 각 connection은 attack / normal 로 기록되어 있음

4개 카테고리의 공격

- DOS: denial-of-service, e.g. syn flood;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow'' attacks;
- probing: surveillance and other probing, e.g., port scanning.

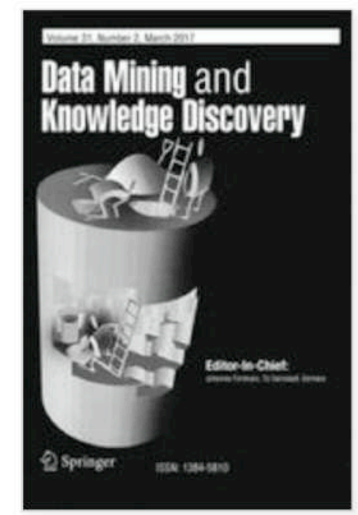
KDD CUP 99 Data set

The data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining

23개 종류의 공격이 포함된 494만개의 training data

테스트 데이터에는 37개의 종류의 공격이 포함됨.

각 기록은 41개의 특징과 1개의 공격 종류로 구성됨



Impact Factor	Available
2.714	1997 - 2017
Volumes	**Issues**
31	103
Articles	**Open Access**
655	35 Articles

출처 : http://kdd.ics.uci.edu/databases/kddcup99/task.html

feature name	*description*	*type*
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of ``wrong'' fragments	continuous
urgent	number of urgent packets	continuous

Table 1: Basic features of individual TCP connections.

feature name	*description*	*type*
hot	number of ``hot'' indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of ``compromised'' conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
num_root	number of ``root'' accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an ftp session	continuous
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete

Table 2: Content features within a connection suggested by domain knowledge.

Packet Header Anomaly Detection (2001)

네트워크 패킷 헤더의 각 값의 빈도를 측정하여 Anomaly Score를 계산함.

Field name	r/n	Values
Ether Size	508/12814738	42 60-1181 1182...
Ether Dest Hi	9/12814738	x0000C0 x00105A x00107B...
Ether Dest Lo	12/12814738	x000009 x09B949 x13E981..
Ether Src Hi	6/12814738	x0000C0 x00105A x00107B...
Ether Src Lo	9/12814738	x09B949 x13E981 x17795A...
Ether Protocol	4/12814738	x0136 x0800 x0806 x9000
IP Header Len	1/12715589	x45
IP TOS	4/12715589	x00 x08 x10 xC0
IP Length	527/12715589	38-1500
IP Frag ID	4117/12715589	0-65461 65462 65463...
IP Frag Ptr	2/12715589	x0000 x4000
IP TTL	10/12715589	2 32 60 62-64 127-128 254-255
IP Protocol	3/12715589	1 6 17
IP Checksum	1/12715589	xFFFF
IP Src	293/12715589	12.2.169.104-12.20.180.101...
IP Dest	287/12715589	0.67.97.110 12.2.169.104-12.20.180.101...
TCP Src Port	3546/10617293	20-135 139 515...
TCP Dest Port	3545/10617293	20-135 139 515...
TCP Seq	5455/10617293	0-395954185 395969583-396150583...
TCP Ack	4235/10617293	0-395954185 395969584-396150584...
TCP Header Len	2/10617293	x50 x60
TCP Flg UAPRSF	9/10617293	x02 x04 x10...
TCP Window Sz	1016/10617293	0-5374 5406-10028 10069-10101...
TCP Checksum	1/10617293	xFFFF
TCP URG Ptr	2/10617293	0 1
TCP Option	2/611126	x02040218 x020405B4
UCP Src Port	6052/2091127	53 123 137-138...
UDP Dest Port	6050/2091127	53 123 137-138...
UDP Len	128/2091127	25 27 29...
UDP Checksum	2/2091127	x0000 xFFFF
ICMP Type	3/7169	0 3 8
ICMP Code	3/7169	0 1 3
ICMP Checksum	1/7169	xFFFF

Table 4.1. The PHAD-C32 model after training on week 3.

PHAD: Packet Header Anomaly Detection for Identifying Hostile Network Traffic

Matthew V. Mahoney and Philip K. Chan
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901
{mmahoney,pkc}@cs.fit.edu

Florida Institute of Technology Technical Report CS-2001-04

IP Header Len	1/12715589	x45

IP Header Length에서
0x45가 아닌 값이 관찰된다면?

Anomaly Score

```
IP Protocol     3/12715589     1 6 17
```

```
> 3/12715589
[1] 2.359309e-07
> -log(3/12715589)
[1] 15.25973
```

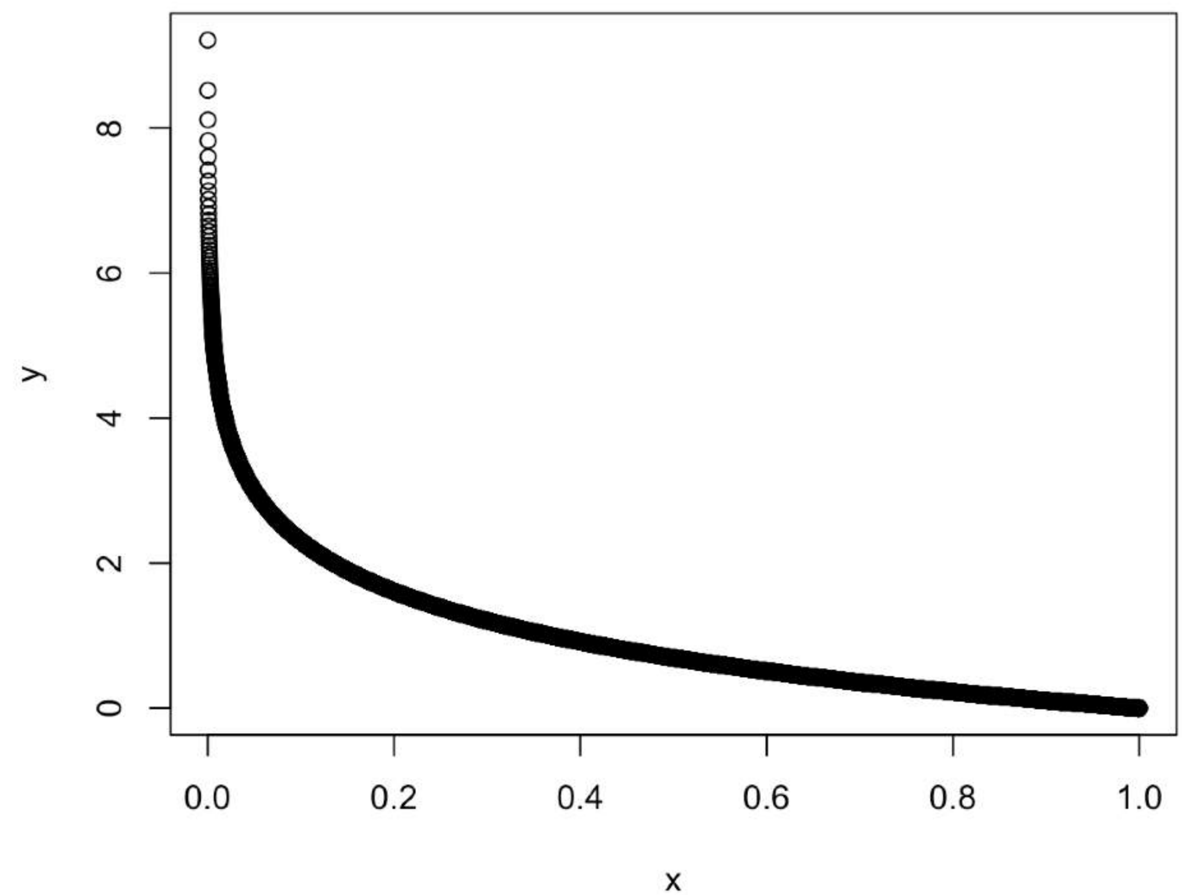
- metric

  P=관찰된 값의 종류 / 패킷 총 개수

- model

  -logP>threshold

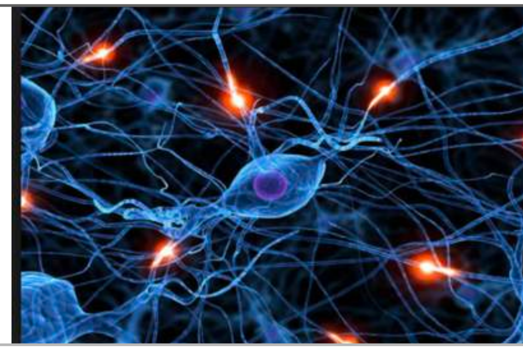
1/1000 vs 1/10000000

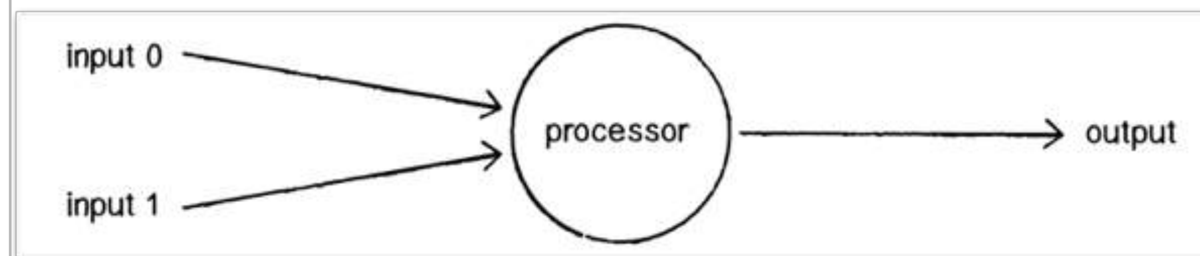


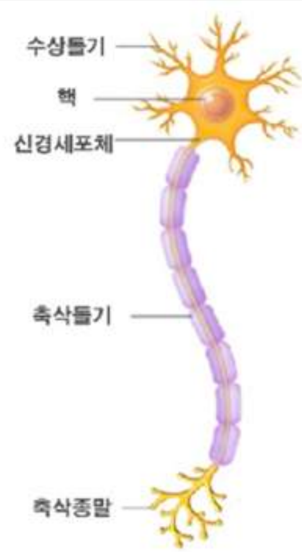
신경망 (neural network)

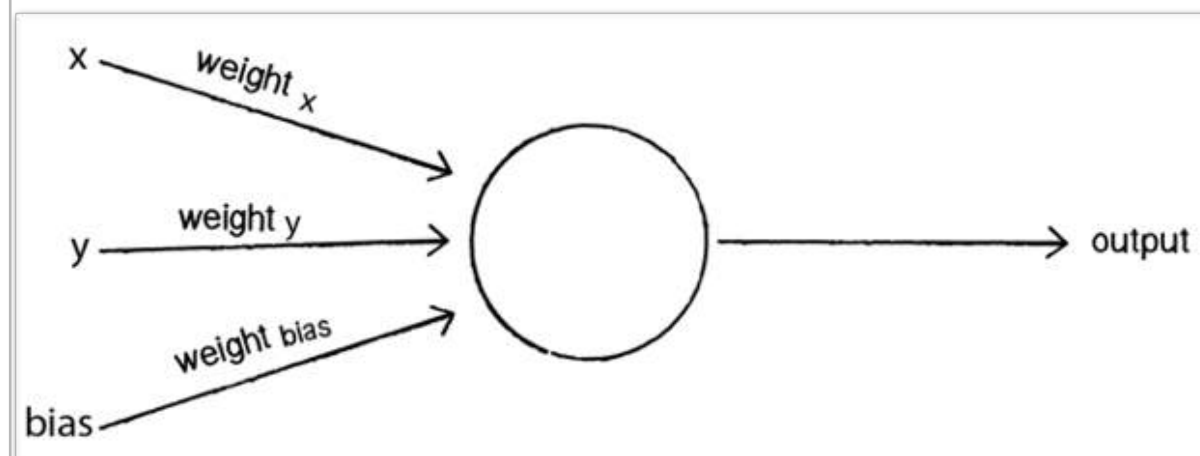
Percepton (F. Rosenblatt, 1957)

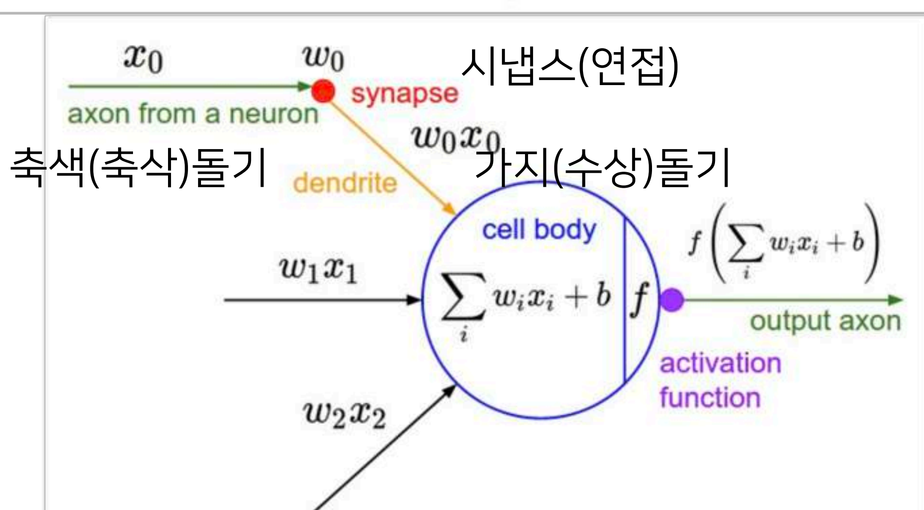
뉴런 : 신경계의 단위, 신경세포체+ 가지돌기 + 축삭 + 시냅스

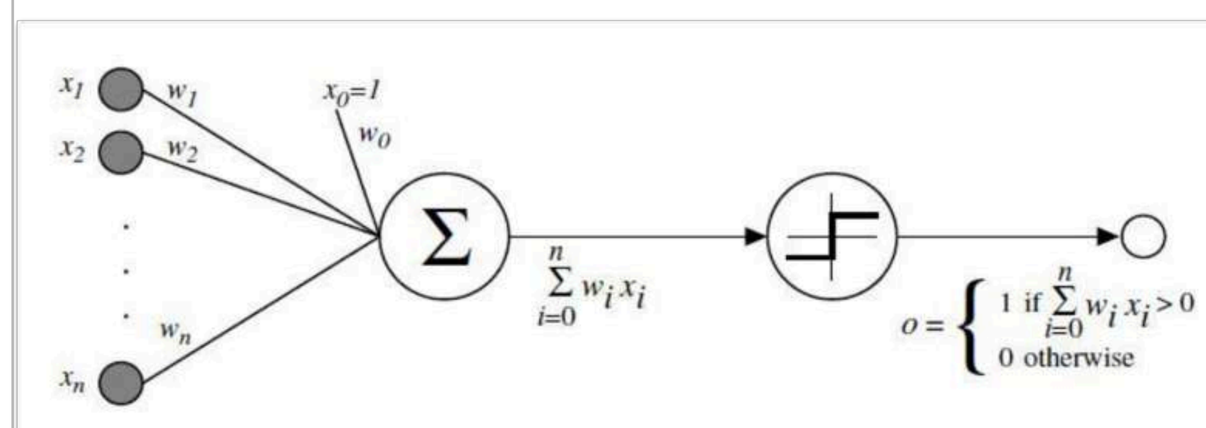








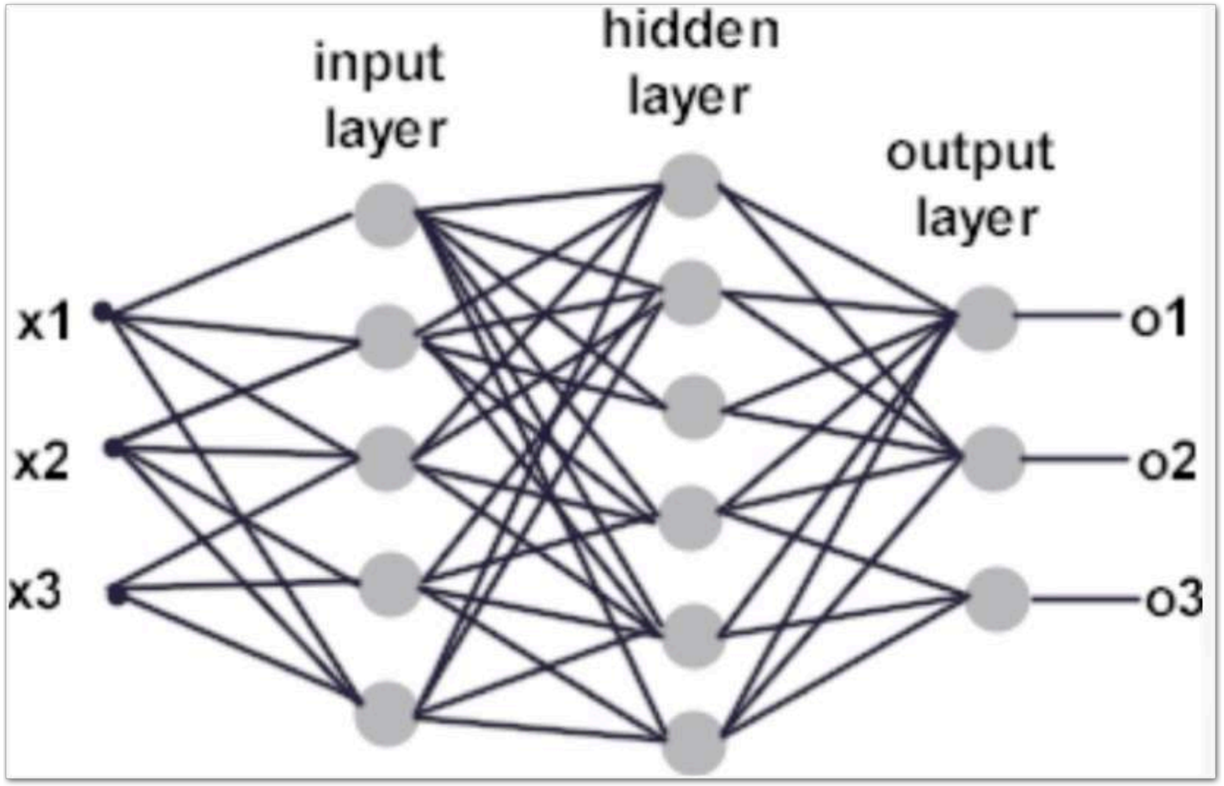


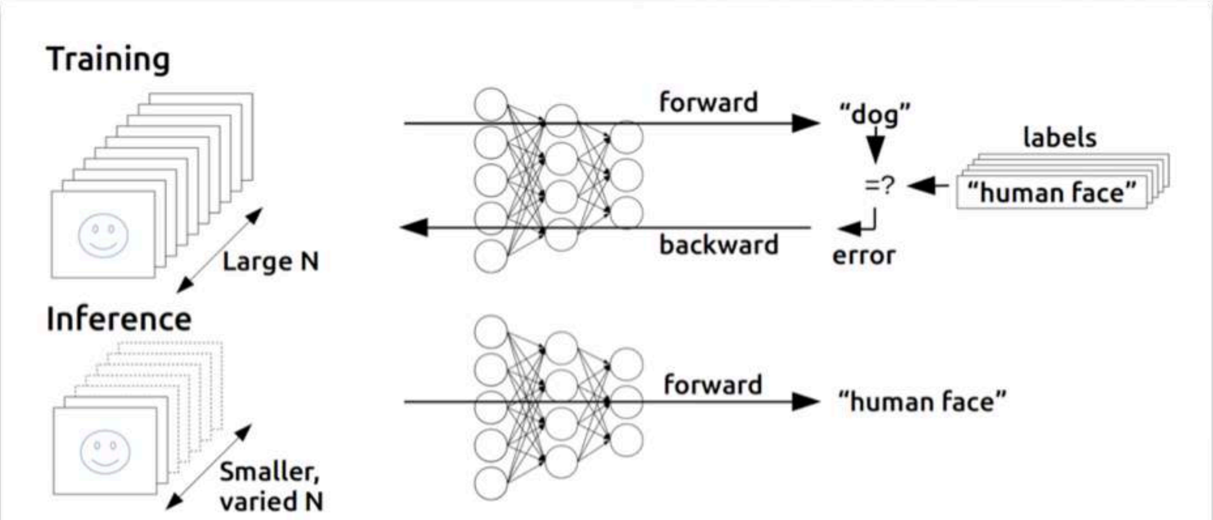


신경망

응용 분야 : 패턴 인식 (문자인식), 시계열 예측(주가, 날씨), 신호 처리 (잡음 제거, 중요 소리 증폭)
제어 (자율 주행), 소프트 센서 (습기, 먼지), 이상탐지

MLP (MultiLayer Perceptrons)	Backpropagation





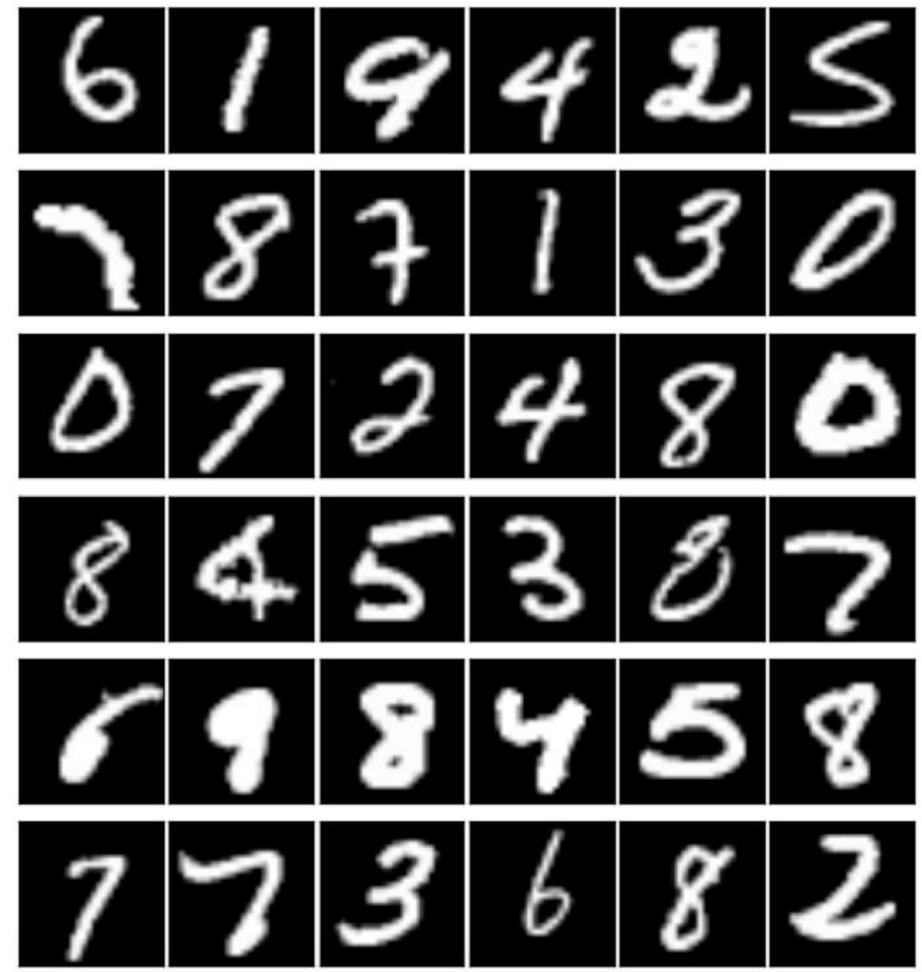
The basic idea of backpropagation. (Source)

Deep Learning

1980년대에 이론 연구는 거의 되었는데, 왜 최근에 인기?

- 이론적 뒷받침 (과적합 문제의 해결)
- Big Data
- 컴퓨팅 파워의 향상 (GPU..)
- 성능 비교 우위

MNIST DataSet

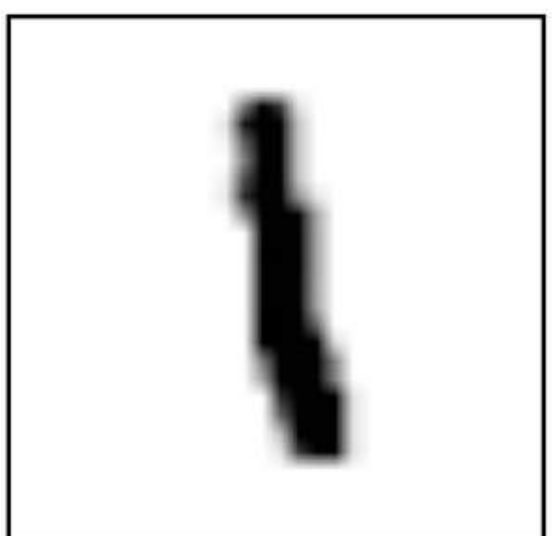


http://yann.lecun.com/exdb/mnist/

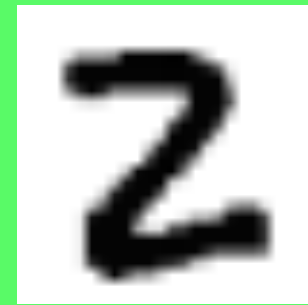
훈련 데이터 : 60,000개

테스트 데이터 : 10,000개

28*28 , class label (숫자)



$$\simeq \begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

Data Sample 

1	162	133	0	0	0	0	0	0	0	0	0	0	0	13	207	246	252	252	252	252	225
	C632	C633	C634	C635	C636	C637	C638	C639	C640	C641	C642	C643	C644	C645	C646	C647	C648	C649	C650	C651	C652
1	208	171	59	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C653	C654	C655	C656	C657	C658	C659	C660	C661	C662	C663	C664	C665	C666	C667	C668	C669	C670	C671	C672	C673
1	0	135	252	172	103	103	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C674	C675	C676	C677	C678	C679	C680	C681	C682	C683	C684	C685	C686	C687	C688	C689	C690	C691	C692	C693	C694
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C695	C696	C697	C698	C699	C700	C701	C702	C703	C704	C705	C706	C707	C708	C709	C710	C711	C712	C713	C714	C715
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C716	C717	C718	C719	C720	C721	C722	C723	C724	C725	C726	C727	C728	C729	C730	C731	C732	C733	C734	C735	C736
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C737	C738	C739	C740	C741	C742	C743	C744	C745	C746	C747	C748	C749	C750	C751	C752	C753	C754	C755	C756	C757
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C758	C759	C760	C761	C762	C763	C764	C765	C766	C767	C768	C769	C770	C771	C772	C773	C774	C775	C776	C777	C778
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C779	C780	C781	C782	C783	C784	**C785**														
1	0	0	0	0	0	0	**2**														

28*28 = 784

Deep Learning in R

```
model <- h2o.deeplearning(x=x, y=y, training_frame = train,
                          validation_frame = test, distribution = "multinomial",
                          activation = "RectifierWithDropout",
                          hidden=c(200,200,200),
                          input_dropout_ratio = 0.2,
                          l1=1e-5,
                          epochs=10)
```

```
H2OMultinomialModel: deeplearning
Model ID:  DeepLearning_model_R_1461057193925_2
Status of Neuron Layers: predicting C785, 10-class classification, multinomial distribution, CrossEntropy loss,
226,010 weights/biases, 2.7 MB, 600,000 training samples, mini-batch size 1
  layer units            type dropout       l1       l2 mean_rate rate_RMS momentum mean_weight weight_RMS
1     1   717           Input 20.00 %
2     2   200 RectifierDropout 50.00 % 0.000010 0.000000  0.073377 0.166079 0.000000    0.019811   0.073754
3     3   200 RectifierDropout 50.00 % 0.000010 0.000000  0.000984 0.000407 0.000000   -0.016804   0.075667
4     4   200 RectifierDropout 50.00 % 0.000010 0.000000  0.001336 0.000736 0.000000   -0.020006   0.072900
5     5    10         Softmax          0.000010 0.000000  0.008359 0.023659 0.000000   -0.200915   0.427615
```

```
Metrics reported on full validation frame

MSE: (Extract with `h2o.mse`) 0.03034059
R^2: (Extract with `h2o.r2`) 0.9963816
Logloss: (Extract with `h2o.logloss`) 0.1188516
Confusion Matrix: Extract with `h2o.confusionMatrix(<model>, <data>)`)
=========================================================================
Confusion Matrix: vertical: actual; across: predicted
         0    1    2   3   4   5   6    7   8   9  Error           Rate
0      967    0    1   1   0   4   4    1   1   1 0.0133 =      13 / 980
1        0 1124    5   1   0   0   3    0   2   0 0.0097 =    11 / 1,135
2        6    1  992   5   4   0   8    7   8   1 0.0388 =    40 / 1,032
3        0    0   14 961   0  12   0   12  10   1 0.0485 =    49 / 1,010
4        1    0    5   2 949   1  10    2   2  10 0.0336 =      33 / 982
5        4    0    1   7   2 862   5    3   5   3 0.0336 =      30 / 892
6        8    3    1   0   4   9 930    0   3   0 0.0292 =      28 / 958
7        2    8   12   4   0   0   0  985   0  17 0.0418 =    43 / 1,028
8        4    1    4   5   6  14   6    6 926   2 0.0493 =      48 / 974
9        4    5    1   9  16   1   0    7   5 961 0.0476 =    48 / 1,009
Totals 996 1142 1036 995 981 903 966 1023 962 996 0.0343 = 343 / 10,000
```

실제값

예측값

Traffic Identification (2015)

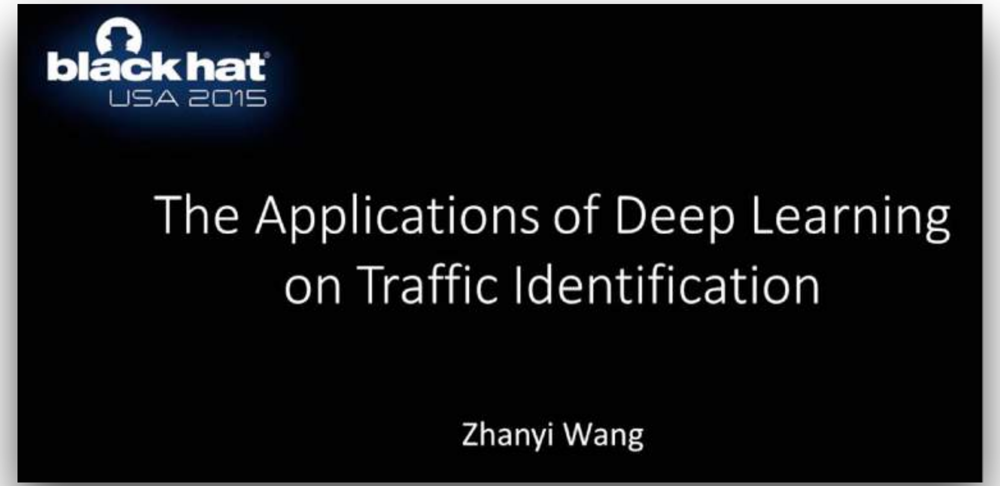
딥러닝으로 프로토콜과 프로세스 식별을 수행함

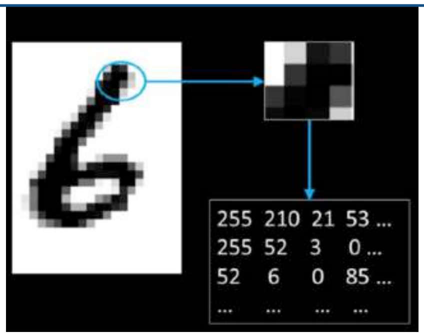
프로토콜 분류 결과 : 97.9% 정확도

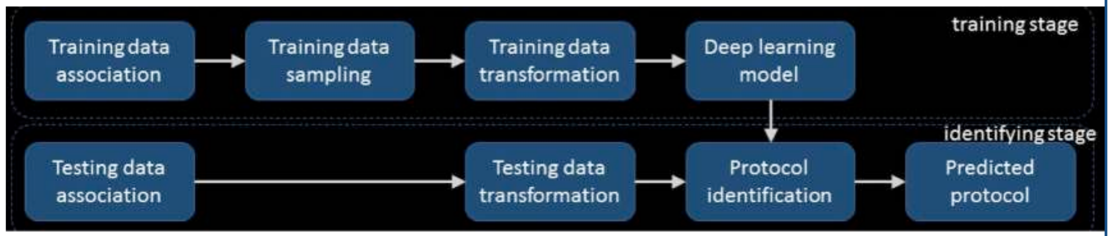
애플리케이션 식별 결과 : 96.3% 정확도

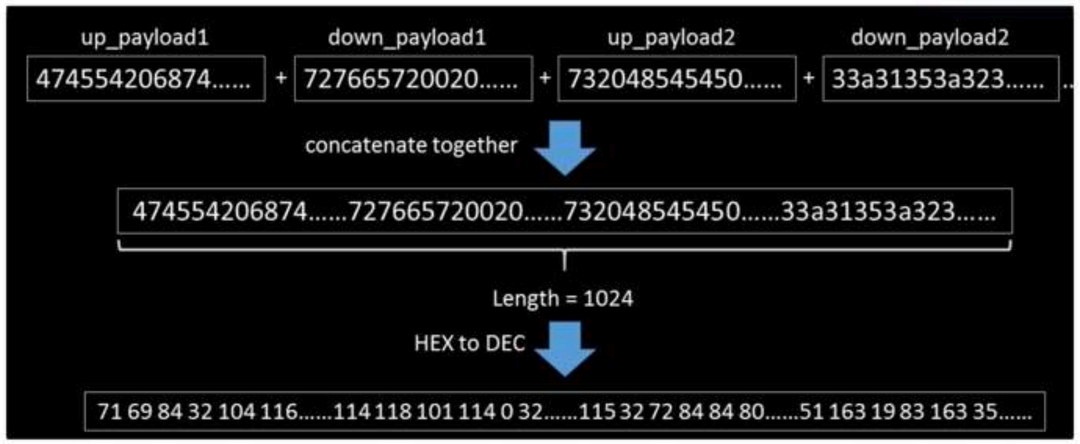
하루 수억개의 데이터, 500만개 이상의 파라미터

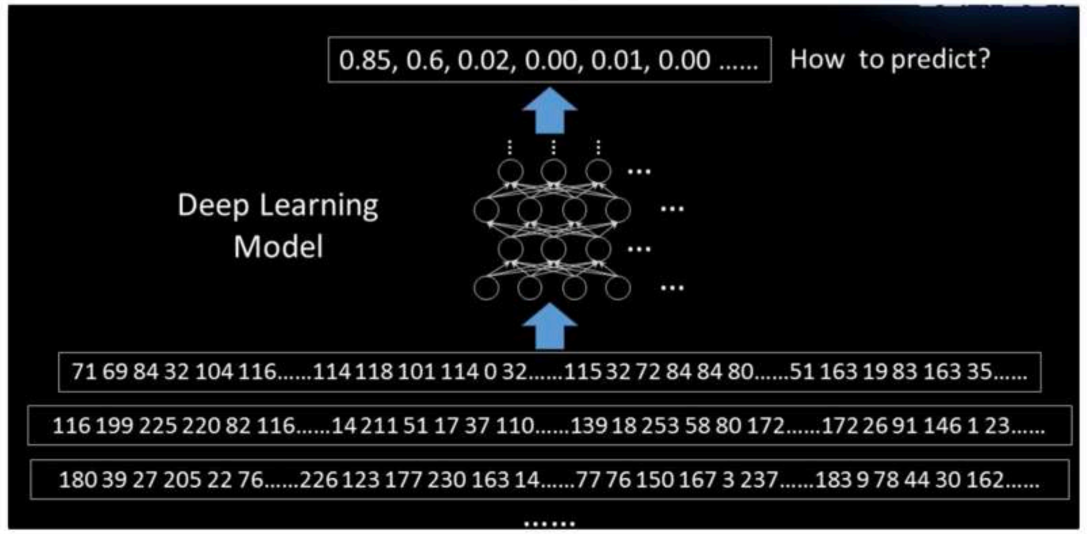
CPU만 사용할 경우 훈련에만 수일 소요











AI^2 (2016)

MIT researchers develop machine learning AI to detect 85% of cyberattacks and it is getting smarter each day

Man and machine come together to predict cyber attacks at a significantly higher rate than currently used systems.

Ashwani Mishra | ETCIO | 20 April 2016, 9:07 AM IST

AI^2 : Training a big data machine to defend

Kalyan Veeramachaneni
CSAIL, MIT Cambridge, MA

Ignacio Arnaldo
PatternEx, San Jose, CA

Alfredo Cuesta-Infante, Vamsi Korrapati, Costas Bassias, Ke Li
PatternEx, San Jose, CA

The AI2 model, which the research team calls "analyst intuition", can detect 85 percent of attacks, which is around three times better than previous benchmarks, while also reducing the number of false positives by a factor of 5. The system was tested on 3.6 billion pieces of data known as "log lines," which were generated by millions of users over a period of three months.

Unsupervised Supervised

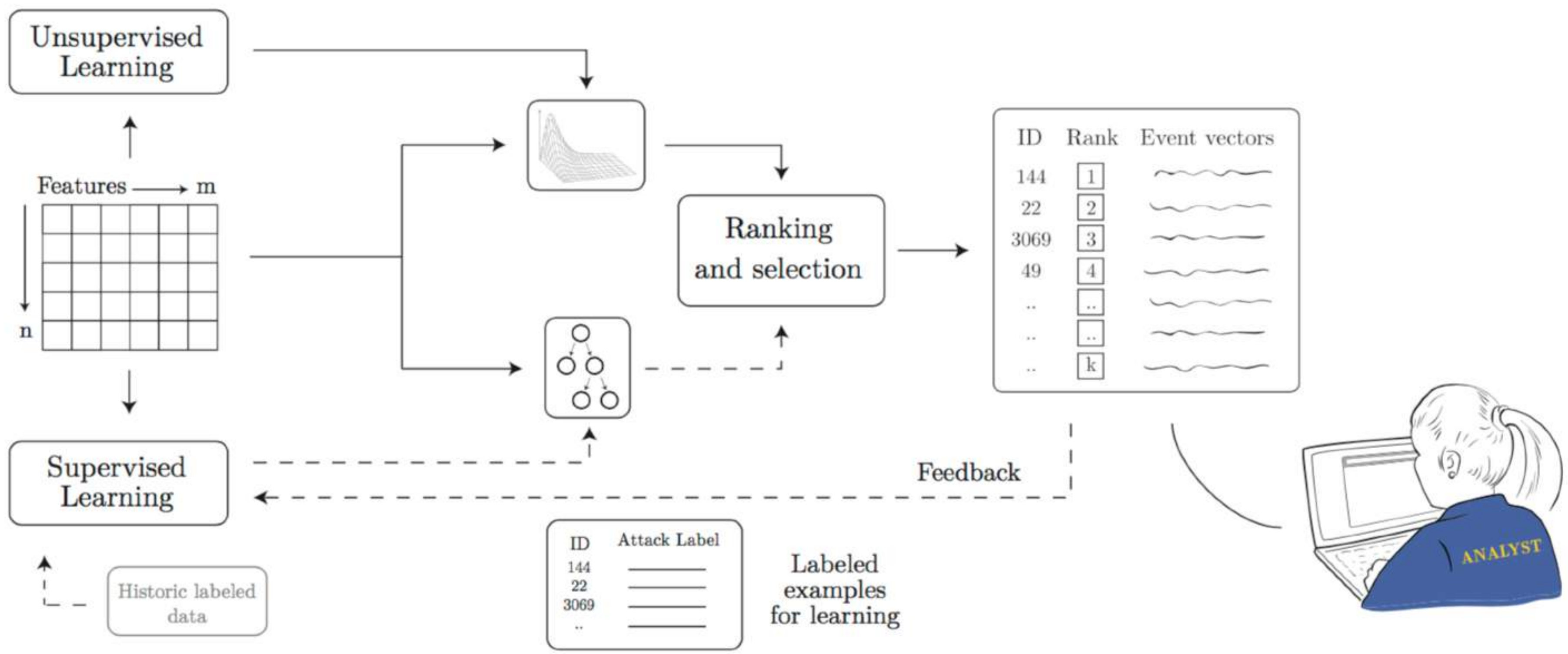


Figure 2: **AI²**. Features describing the entities in the data set are computed at regular intervals from the raw data. An unsupervised learning module learns a model that is able to identify extreme and rare events in data. The rare events are ranked based on a predefined metric, and are shown to the analyst, who labels them as 'normal' or as pertaining to a particular type of attack. These "labels" are provided to the supervised learning module, which produces a model that can then predict, from features, whether there will be attacks in the following days.

보안 목표와 성능 지표

	Intrusion Detection	Abusing Detection	성능 지표
Soundness	실제로 침입을 탐지하는가?	찾은 어뷰저가 진짜 어뷰저인가?	Precision (정밀도)
Completeness	전부는 아니어도 대부분의 침입을 탐지하는가?	전체 어뷰저 중 얼마나 찾았는가?	Recall (재현율)
Timeliness	심각한 피해를 당하기 전에 탐지할 수 있는가?	심각한 피해를 당하기 전에 탐지할 수 있는가?	

Feature Selection

Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc...

Features:
1. Sky Blue
2. **Logo**
3. Shape
etc...

Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc...

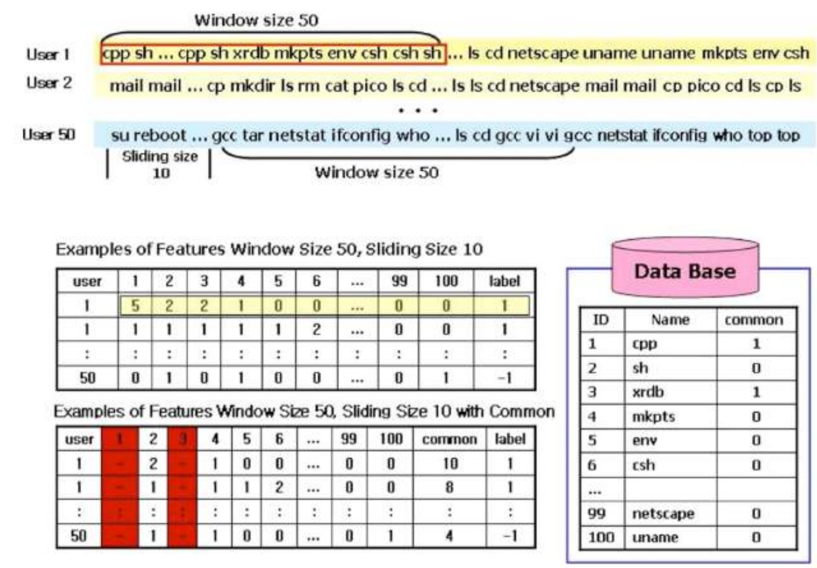
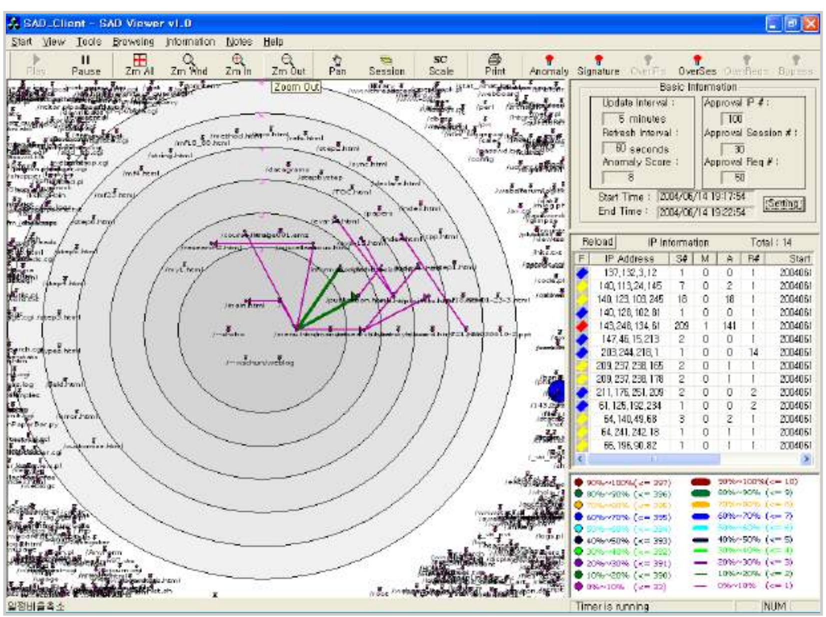
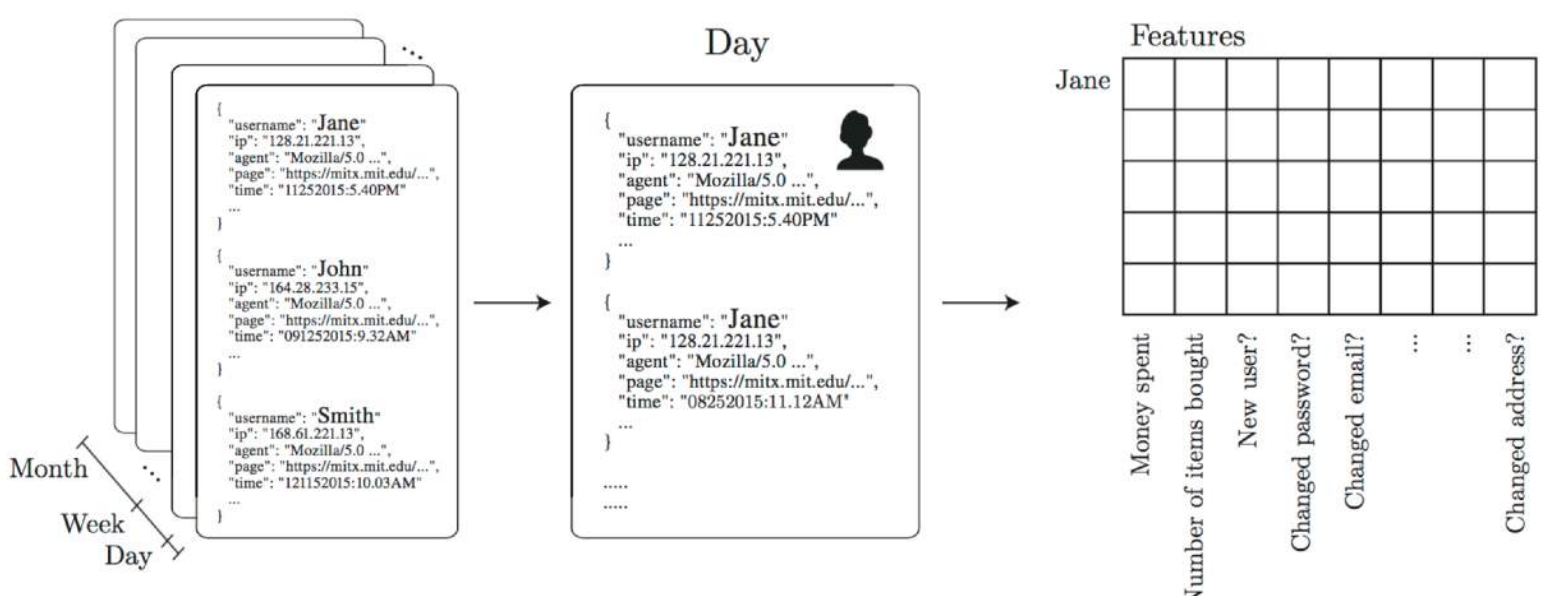


Figure 3 Feature selection for SVM profiling.

UNIX command (2004)



Page Sequence (2003)



AI^2 (2016)

Core Assumptions

- 판단 근거

  드물다, 다르다, 조화롭지 않다. 이상해 보인다. 방향이 특이하다.

  상위로 많이 발생한다. 하위로 많이 발생한다.

  주목을 끌지 않는 사건의 조합

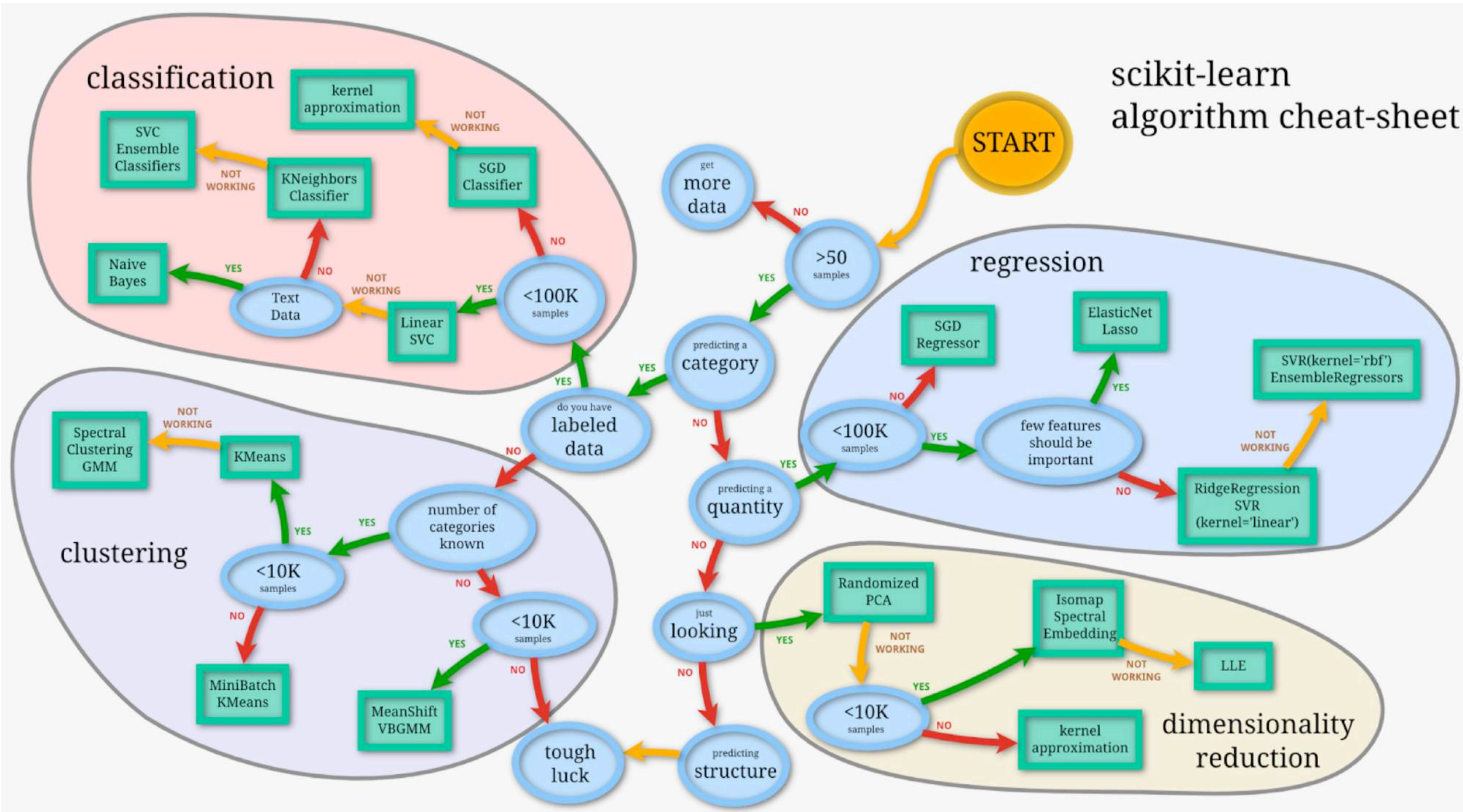
- 예시

  Network Security : Botnet and Honeynet

  Botnet protocols are mostly C&C

  Individual bots within same botnets behave similarly and can be correlated to each other

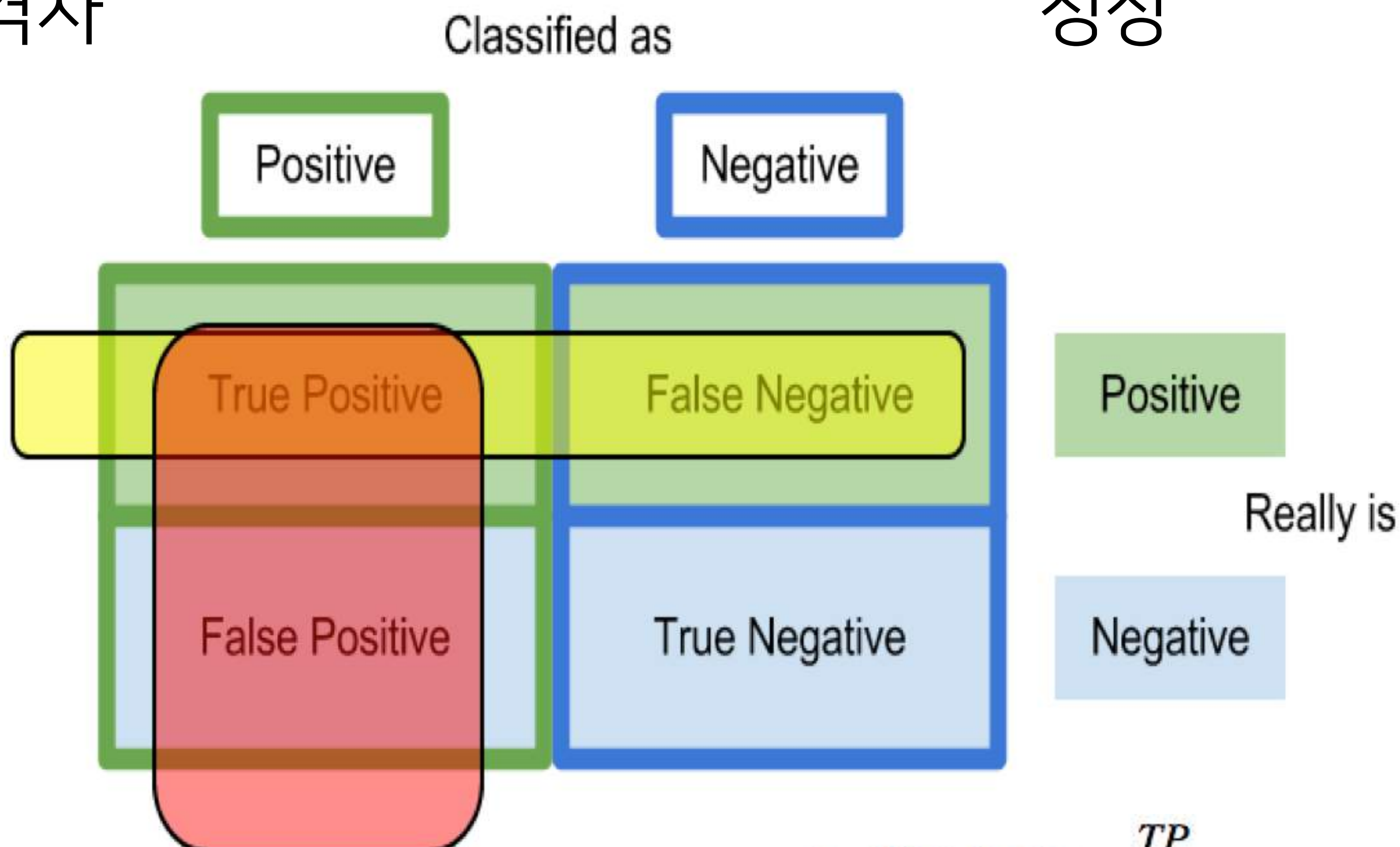
  Botnet behaviors are different and distinguishable from legitimate human user, e.g. human behaviors are more complex



출처 : http://peekaboo-vision.blogspot.kr/2013/01/machine-learning-cheat-sheet-for-scikit.html

Performance Measure

공격자 정상



실제공격자

실제 정상

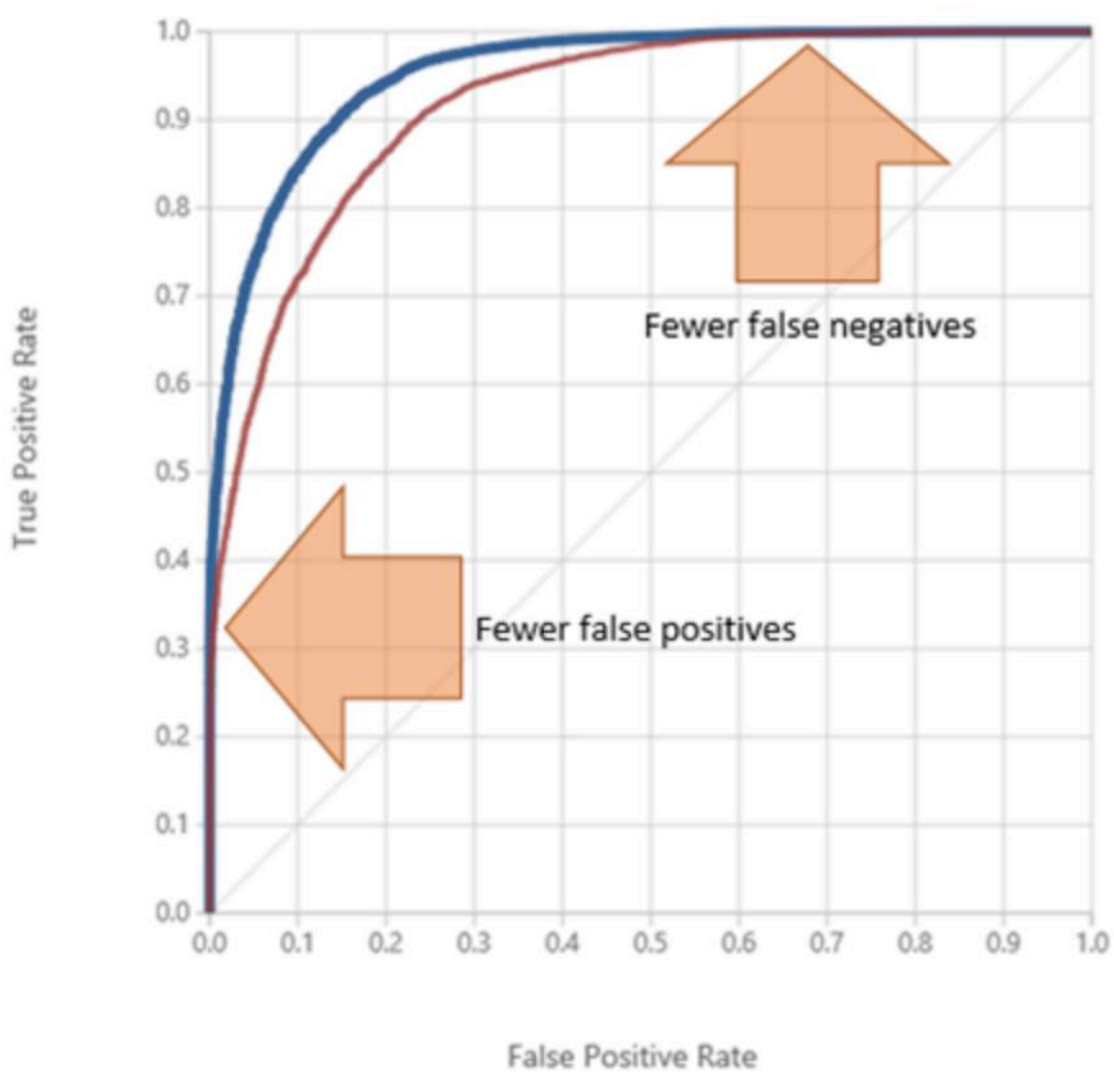
Accurary = (TP+TN) / (TP+FP+FN+TN)

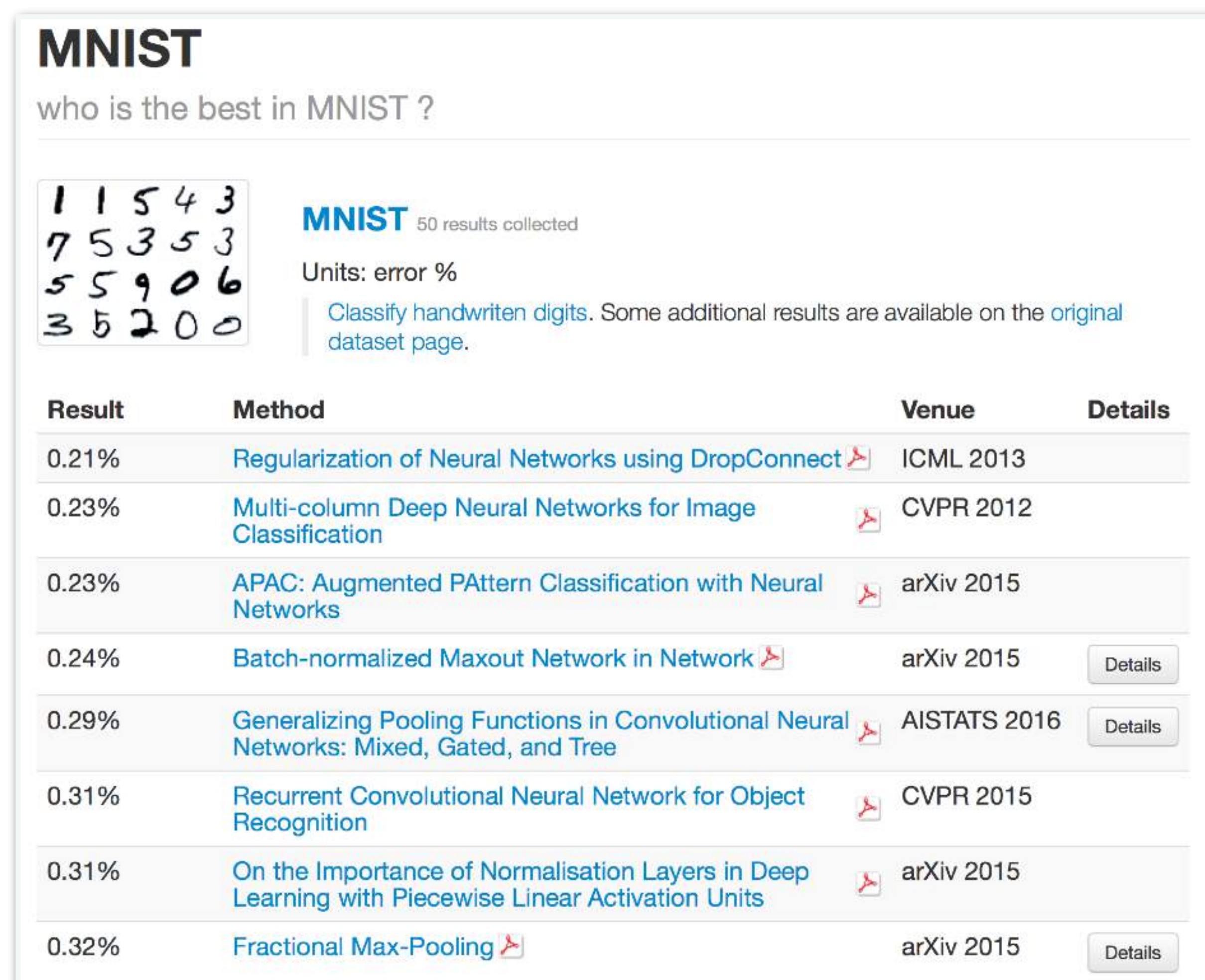
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1 score: $\frac{2TP}{2TP+FP+FN}$

Performance Measure

어느정도 신뢰할 수 있는가?

용인할 수 있는 에러율





MNIST

who is the best in MNIST ?

MNIST 50 results collected

Units: error %

Classify handwriten digits. Some additional results are available on the original dataset page.

Result	Method	Venue	Details
0.21%	Regularization of Neural Networks using DropConnect	ICML 2013	
0.23%	Multi-column Deep Neural Networks for Image Classification	CVPR 2012	
0.23%	APAC: Augmented PAttern Classification with Neural Networks	arXiv 2015	
0.24%	Batch-normalized Maxout Network in Network	arXiv 2015	Details
0.29%	Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree	AISTATS 2016	Details
0.31%	Recurrent Convolutional Neural Network for Object Recognition	CVPR 2015	
0.31%	On the Importance of Normalisation Layers in Deep Learning with Piecewise Linear Activation Units	arXiv 2015	
0.32%	Fractional Max-Pooling	arXiv 2015	Details

출처 : http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html

Attackers for ML

Attacker	탐지 모듈 인지	탐지 회피	탐지 모듈 파괴
Passive	X	X	X
Semi-Aggressive	O	O	X
Active	O	O	O

어쩌면 공격자가 알 수도 있는 것

학습 알고리즘

알고리즘이 사용하는 특징들

알고리즘의 파라미터들, 훈련 및 테스트 데이터

Adversarial Machine Learning

Adversarial machine learning is a research field that lies at the intersection of machine learning and computer security.

It aims to enable the safe adoption of machine learning techniques in adversarial settings like spam filtering malware detection and biometric recognition.

A malicious adversary can carefully manipulate the input data exploiting specific vulnerabilities of learning algorithms to compromise the whole system security

Attacks in spam filtering, where spam messages are obfuscated through misspelling of bad words or insertion of good words

출처 : wikipedia

Adversarial Machine Learning

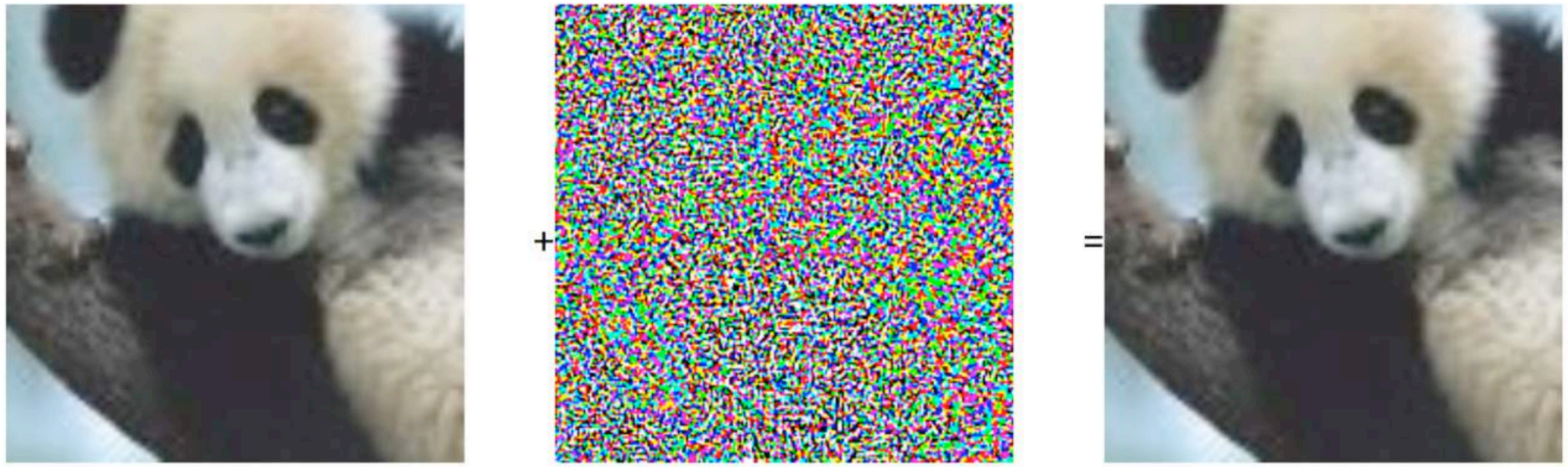
Deep networks can be easily fooled..

“It turns out some DNNs only focus on discriminative features in images”

Learning is expensive.

Reverse engineering of machine learning.

It aims to design robust and secure learning algorithms.

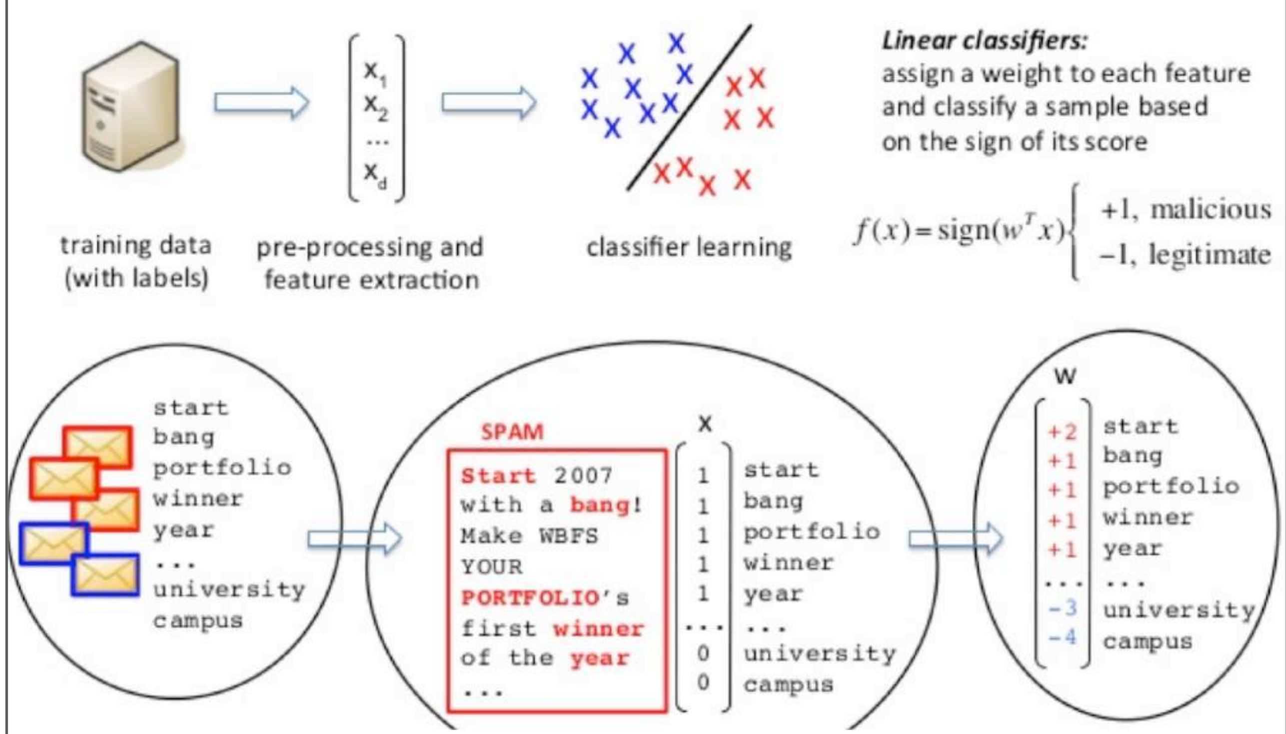
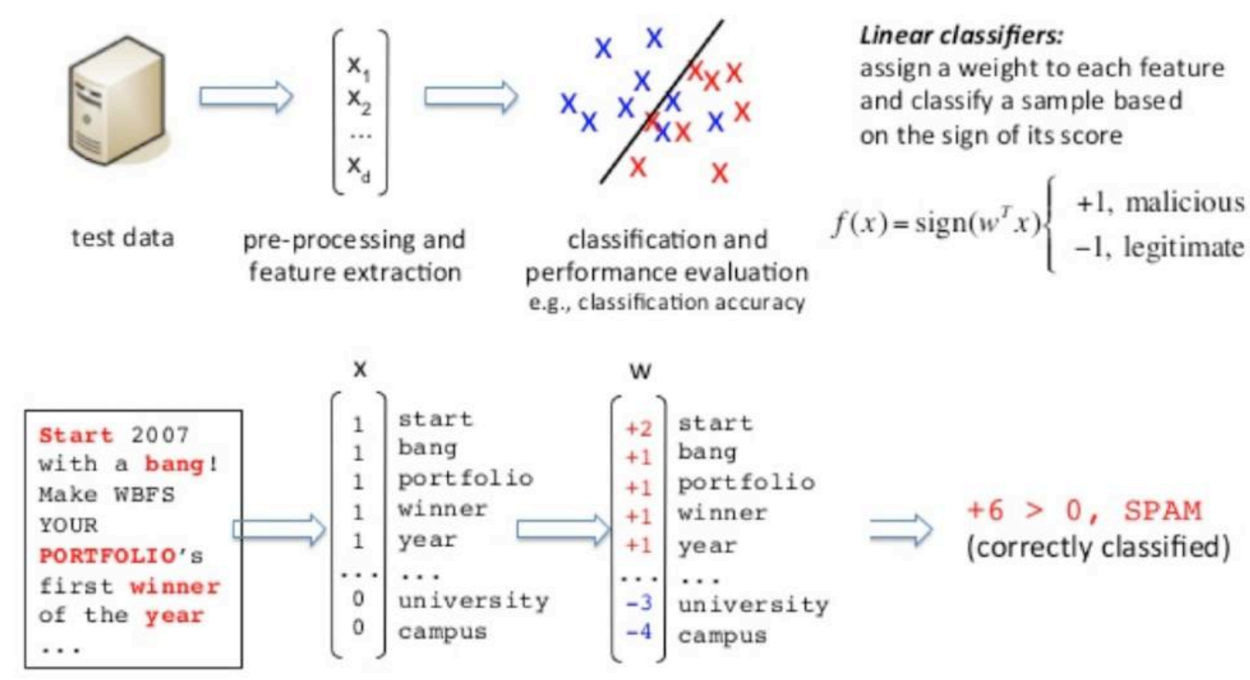


Original image classified as a panda with 60% confidence.

Tiny adversarial perturbation.

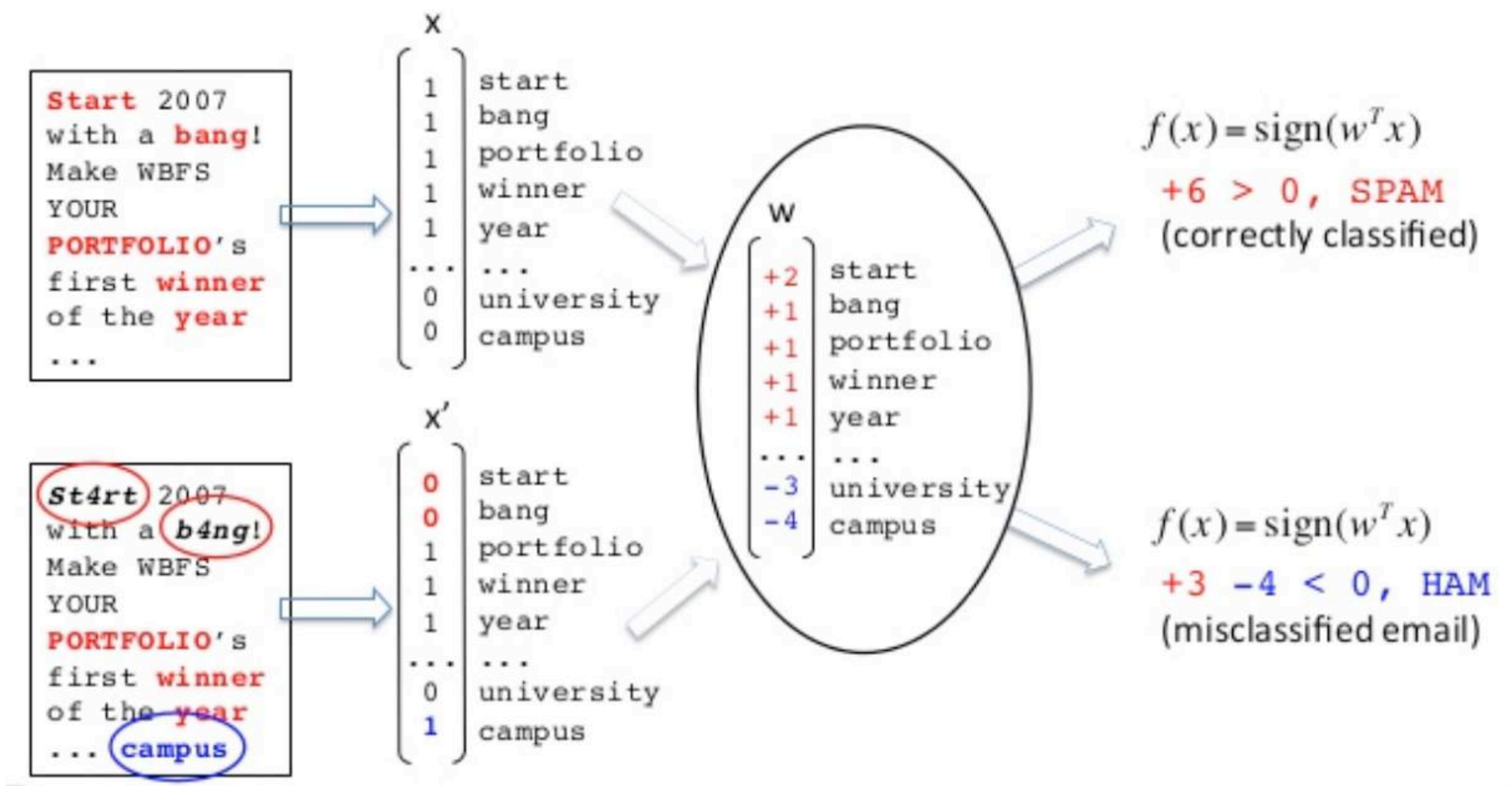
Imperceptibly modified image, classified as a gibbon with 99% confidence.

Spam Detection

Training	Test
	

출처 : http://pralab.diee.unica.it

Problem



Start 2007
with a bang!
Make WBFS
YOUR
PORTFOLIO's
first winner
of the year
...
St4rt 2007
with a b4ng!
Make WBFS
YOUR
PORTFOLIO's
first winner
of the year
... campus
x
1 start
1 bang
1 portfolio
1 winner
1 year
... ...
0 university
0 campus
x'
0 start
0 bang
1 portfolio
1 winner
1 year
... ...
0 university
1 campus
w
+2 start
+1 bang
+1 portfolio
+1 winner
+1 year
... ...
-3 university
-4 campus
f(x) = sign(w^T x)
+6 > 0, SPAM
(correctly classified)
f(x) = sign(w^T x)
+3 -4 < 0, HAM
(misclassified email)

Lessons Learned (1)

Anomaly Detection은

주로 통계적인 방법으로 이상(unusual or rare)해 보이는 패턴과 행위를 찾으려고 시도

새로운 공격을 발견할 수 있겠지만,

- 로그를 체계적으로 수집하기 힘들고,
- 훈련이라는 과정이 필요하고,
- 공격을 특정 이름으로 분류하기 어렵고,
- False alarm이 많고,
- 정상적으로 보이기 위한 매우 천천히 진행되는 공격에는 대응하기 어렵고,
- 근본적으로 “정상” 이라고 믿는 전제가 확실한 사실이 아닐수도..

Lessons Learned (2)

- 쉽지 않았다. 시행 착오의 반복

    데이터 수집 단계 : 방대한 데이터

    특징 선택 단계 : 좋은 특징 선택의 실패

    모델 선택 단계 : 특징에 알맞은 모델 선택 실패

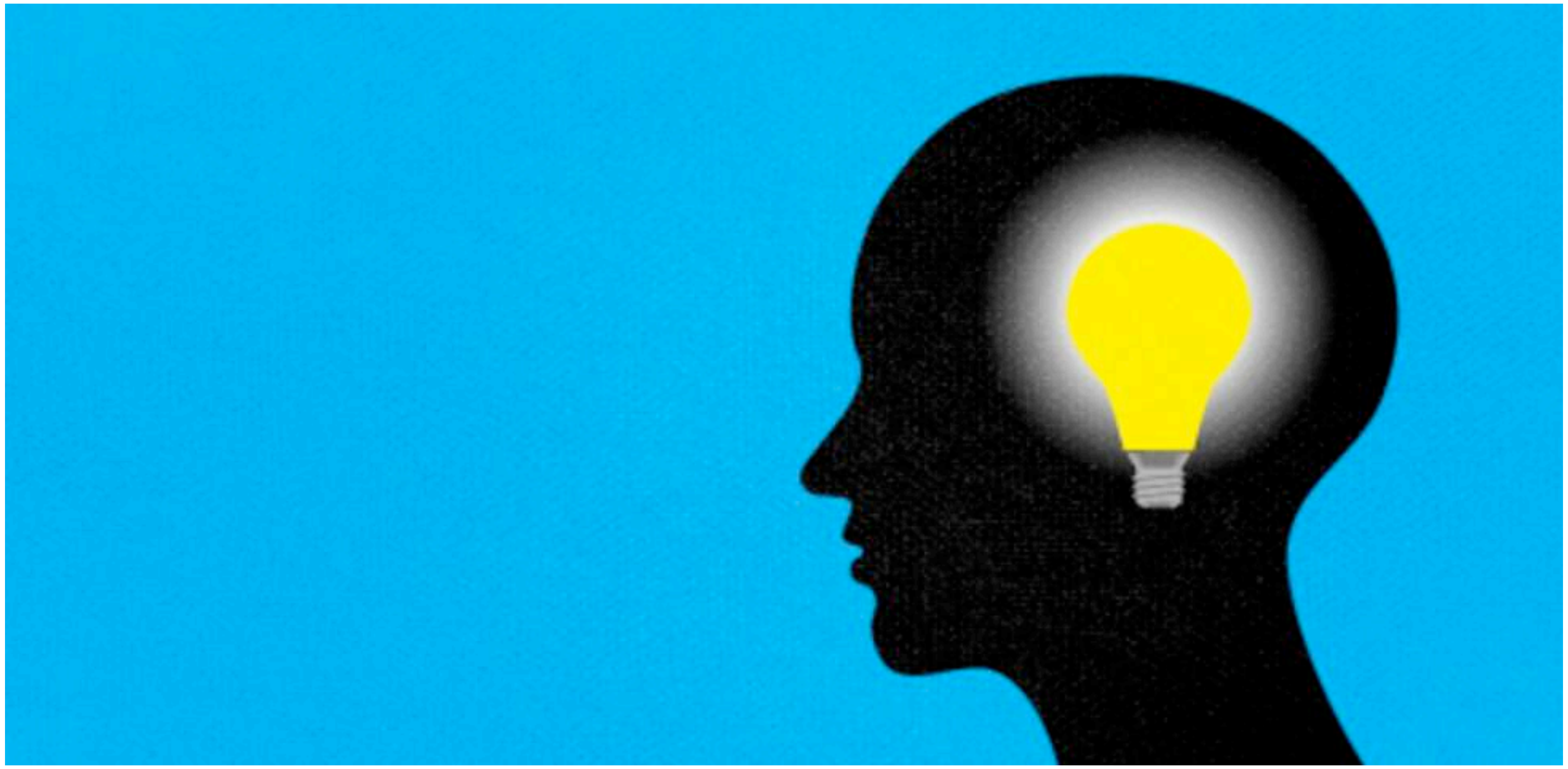
    학습 단계 : 충분한 학습이 어려움

        - 복잡성, 컴퓨터 성능, 과적합(Overfitting) 문제

Lessons Learned (3)

- 로그 분석을 하다 보면 데이터는..
  - 지나치게 많거나
  - 충분하지 않거나
  - 중복 데이터
  - (서로 다른 로그가 동일한 이벤트 일 수 있음)
  - 다양한 기록이 섞여 있거나
  - 현실을 반영하지 못하는 다량의 오탐 메시지
  - 데이터 수집의 어려움

Lessons Learned (4)



- 막연한 기대를 버리자!
- 아는 것 만큼 보인다!
- 남들은 어떤 특징을 뽑아서 어떻게 썼는가?

Lessons Learned (5)

숙련된 분석가는

로그, 경고, 패킷 덤프와 같은 유입 데이터에서

통찰력을 가지기 위해 최적화한 분석 프로세스를 따르고 도구를 실행한다.

Lessons Learned (6)

고민

훈련 데이터와 테스트 데이터는 같은 분포로 샘플링 되었는가?

정교하게 조작된 공격에는 대응할 수 없을지 모른다.

좀더 기계적인 방법으로 분류기(classifier)의 보안성을 어떻게 평가할 수 있는가?

우리가 가야할 방향

숙련된 분석가를 곁에 두고,

기존 방식으로 분석할 수 없는 희박한 데이터를 다룸

'레이더에 걸리지 않는' 것을 탐지

사람만이 할 수 있던 작업을 자동화

문제를 예측하기 위한 시도

안전한 ML 알고리즘과 구현물

제대로 된 시각화가 필요하다.

There is still no silver bullet

머신러닝은 하나의 도구일 뿐이다.

우리가 풀려는 문제가 명확해야 하고, 그 답을 줄 수 있는 데이터들이 충분히 모아져야 한다.

Appendix : Papers Survey (2008~2015)

	Supervised	Semi-supervised	Unsupervised	HITL	Game Theory
Attacker Type					
Passive	58(49%)	7(5.9%)	24(20%)	2(1.7%)	0(0%)
Semi-aggressive	18(15%)	4(3.4%)	3(2.5%)	0(0%)	1(0.85%)
Active	0(0%)	0(0%)	0(0%)	0(0%)	2(1.7%)
Means of Attack					
Server	4(3.4%)	1(0.85%)	1(0.85%)	0(0%)	0(0%)
Network	17(14.4%)	4(3.4%)	11(9.3%)	0(0%)	1(0.85%)
Client app	4(3.4%)	0(0%)	1(0.85%)	2(1.7%)	0(0%)
User	31(26%)	2(1.7%)	9(7.6%)	0(0%)	2(1.7%)
Client machine	20(17%)	4(3.4%)	5(4.2%)	0(0%)	0(0%)
Purpose of Attack [22, 23]					
Confidentiality	16(13.6%)	0(0%)	7(6%)	0(0%)	0(0%)
Availability	9(7.6%)	1(0.85%)	5(4.2%)	0(0%)	3(2.5%)
Integrity	51(43.2%)	10(8.5%)	15(12.7%)	2(1.7%)	0(0%)

Supervised learning uses labeled data for training.

Semi-supervised learning uses both labeled and unlabeled data for training.

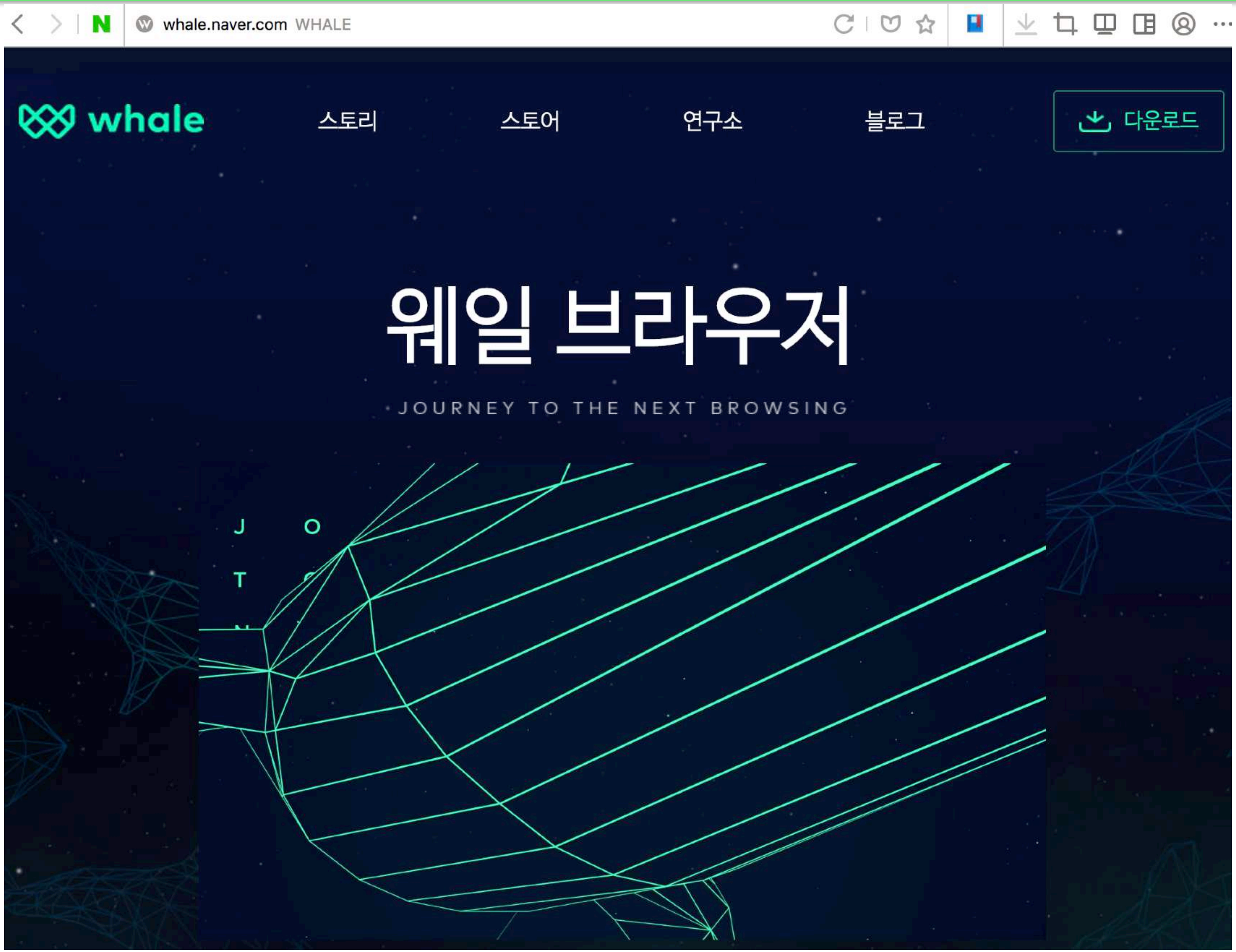
Unsupervised learning has no labeled data available for training.

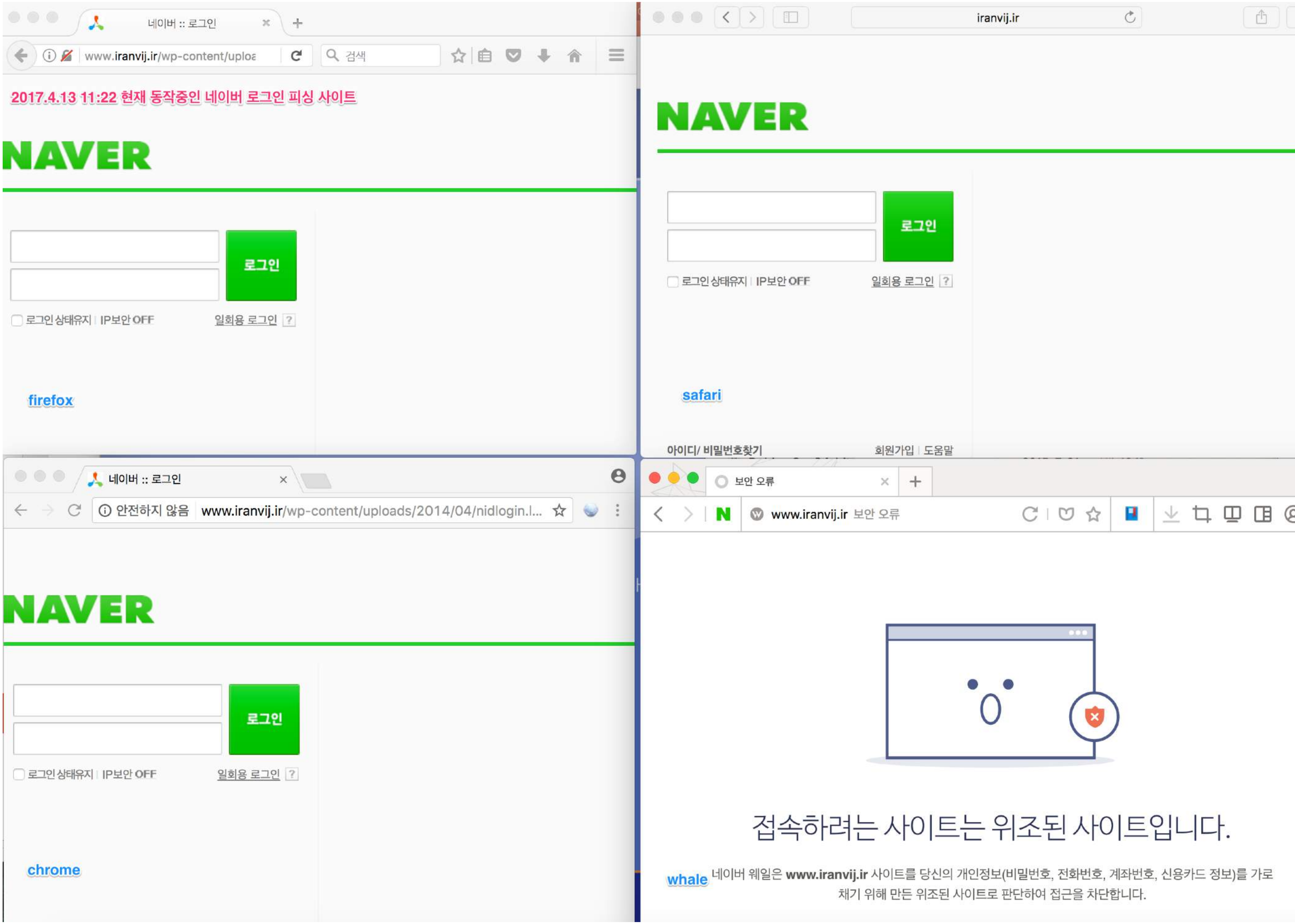
Human-in-the-loop(HITL) learning incorporates active human feedback to algorithm's decisions into the knowledge base and/or algorithms.

Game Theory(GT)-based learning considers learning as a series of strategic interactions between the model learner and actors

출처 : SoK: Applying Machine Learning in Security - A Survey

http://whale.naver.com




네이버 :: 로그인
www.iranvij.ir/wp-content/uploa
검색
2017.4.13 11:22 현재 동작중인 네이버 로그인 피싱 사이트
NAVER
로그인
로그인 상태유지 | IP보안 OFF
일회용 로그인 ?
firefox
iranvij.ir
NAVER
로그인
로그인 상태유지 | IP보안 OFF
일회용 로그인 ?
safari
아이디/ 비밀번호찾기
회원가입 | 도움말
네이버 :: 로그인
안전하지 않음 www.iranvij.ir/wp-content/uploads/2014/04/nidlogin.l...
NAVER
로그인
로그인 상태유지 | IP보안 OFF
일회용 로그인 ?
chrome
보안 오류
www.iranvij.ir 보안 오류
접속하려는 사이트는 위조된 사이트입니다.
whale
네이버 웨일은 www.iranvij.ir 사이트를 당신의 개인정보(비밀번호, 전화번호, 계좌번호, 신용카드 정보)를 가로
채기 위해 만든 위조된 사이트로 판단하여 접근을 차단합니다.