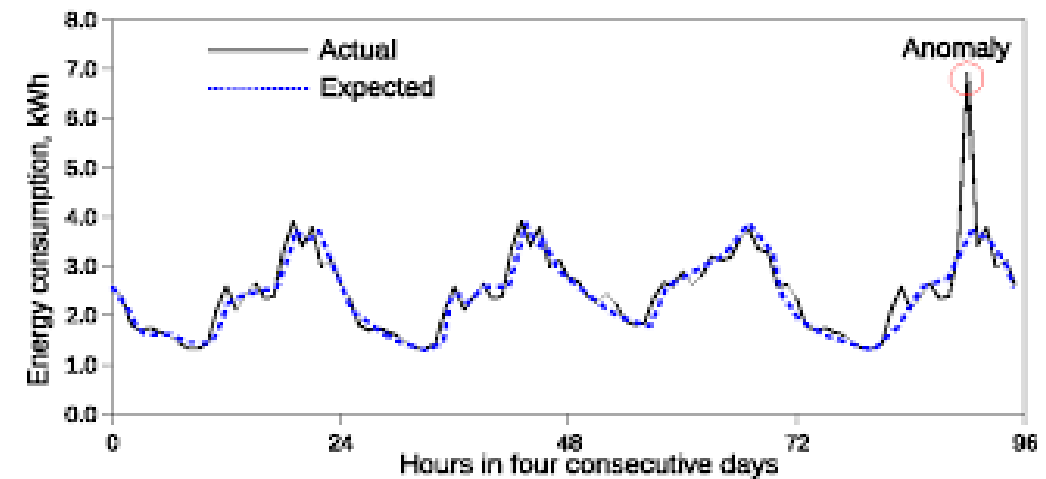
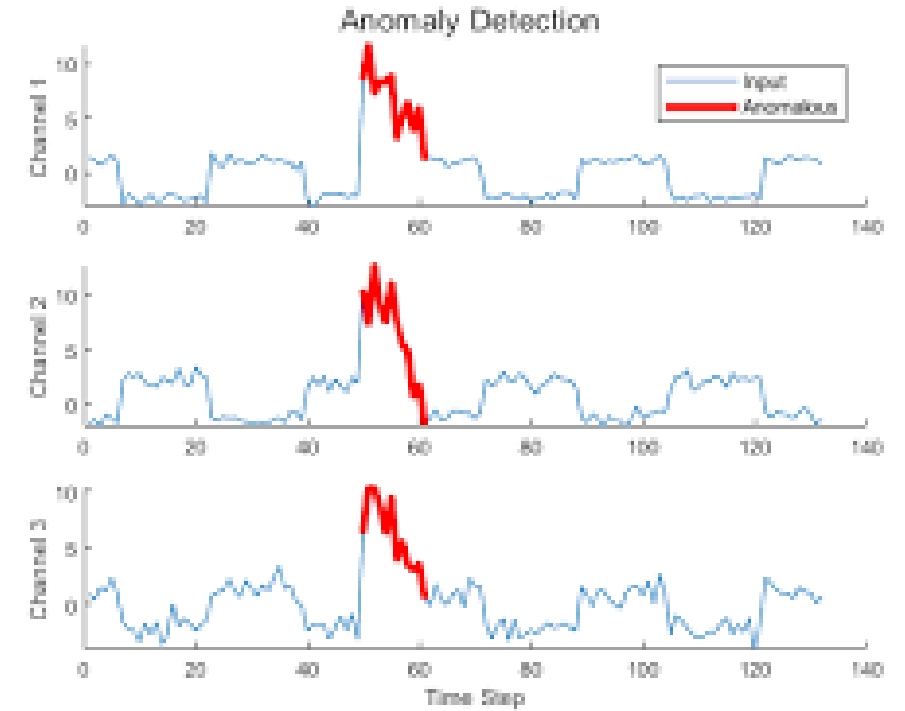


Anomaly Detection: From Murder Mystery to Financial Fraud

Zulkaida Akbar

11/18/2022

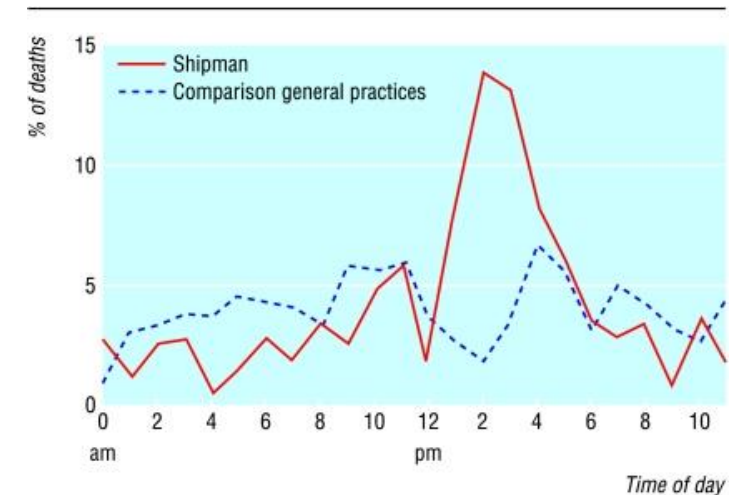
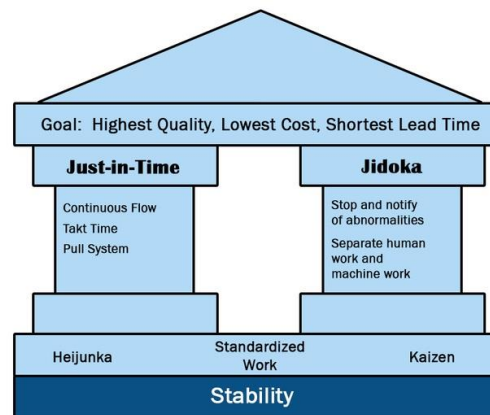
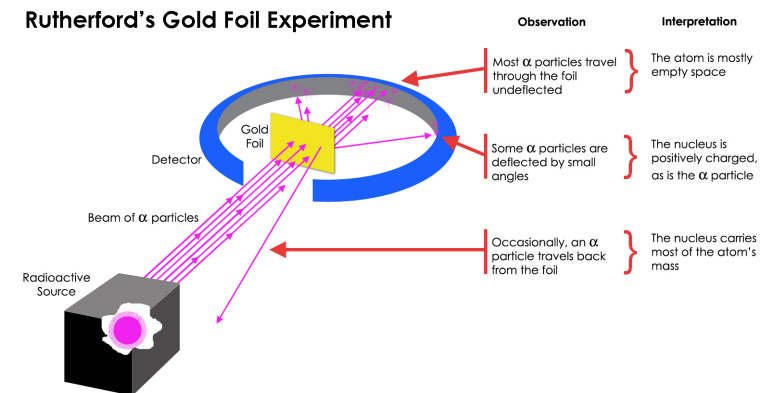


Outline

- What is Anomaly Detection and why is it important?
- Why and What is Machine Learning?
- Machine-Learning based anomaly detection
- Human-Machine cooperation
- Summary & Outlook: The bright future of AI

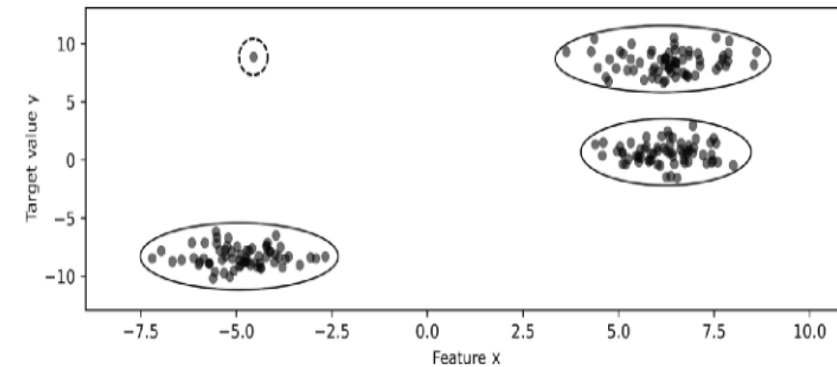
What is Anomaly Detection?

Why is it Important?



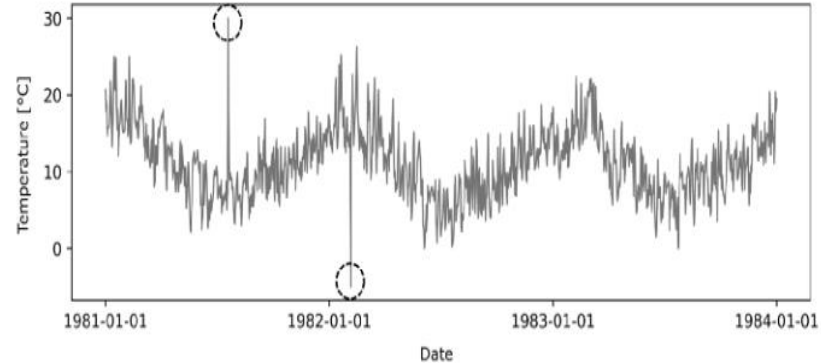
What is Anomaly Detection and Why is it Important?

Anomaly or outlier detection is the process of finding data objects with behaviors that are very different from expectation.
(Han et al, Data Mining, Concept & Technique)



Point anomaly:

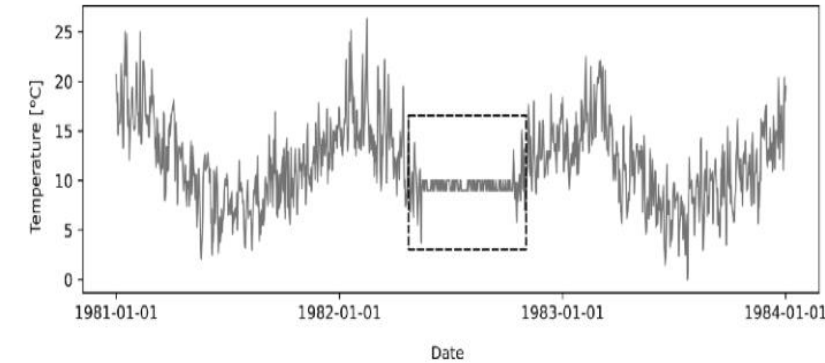
- Clear differences from the rest of the data
- Example: credit card fraud based on anomalous spending



Contextual anomaly in air temperature records (dataset: minimum daily temperature in Melbourne, Australia [Aus90] — modified)

Contextual anomaly

- The anomaly is context specific
- Example: 30 degree Celsius is normal somewhere else but it is anomaly in Australia at a specific time



Collective anomaly in air temperature records (dataset: minimum daily temperature in Melbourne, Australia)

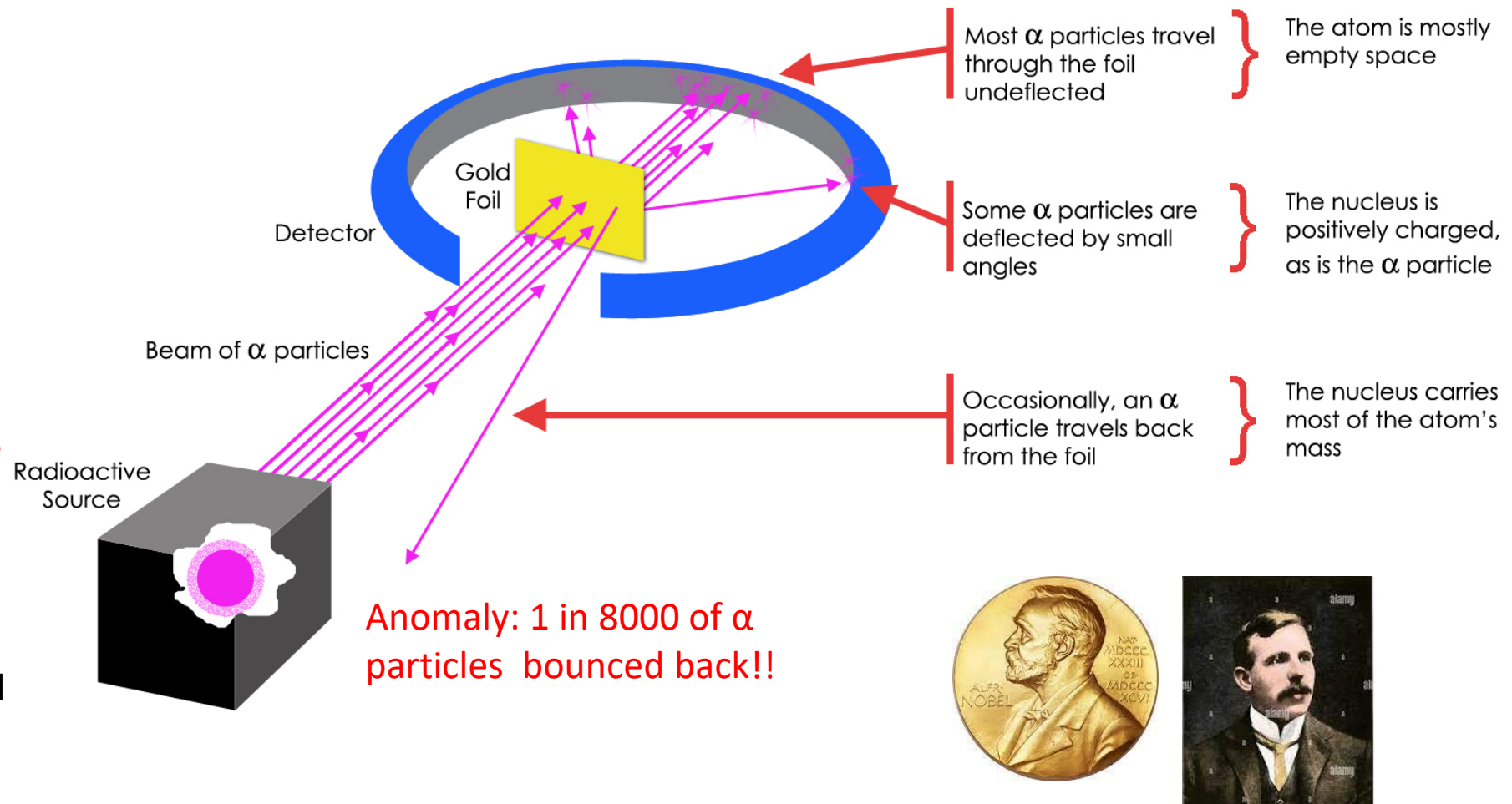
Collective anomaly

- Collective data which are anomalous to the rest of data
- Example: Anomalous data copying activity between local-remote machine indicating a cyber attack

Anomaly has driven a great discovery in modern science

Rutherford's Gold Foil Experiment

- ❑ Electron was discovered before proton
- ❑ Proton was discovered by Rutherford in the famous Gold foil experiment where the beam of α particles was shot to a gold foil
- ❑ It was expected that most of the α particles will be slightly deflected
- ❑ Surprisingly, **few** α particles were bounced back from the foil (**1 in 8000** of α particles)



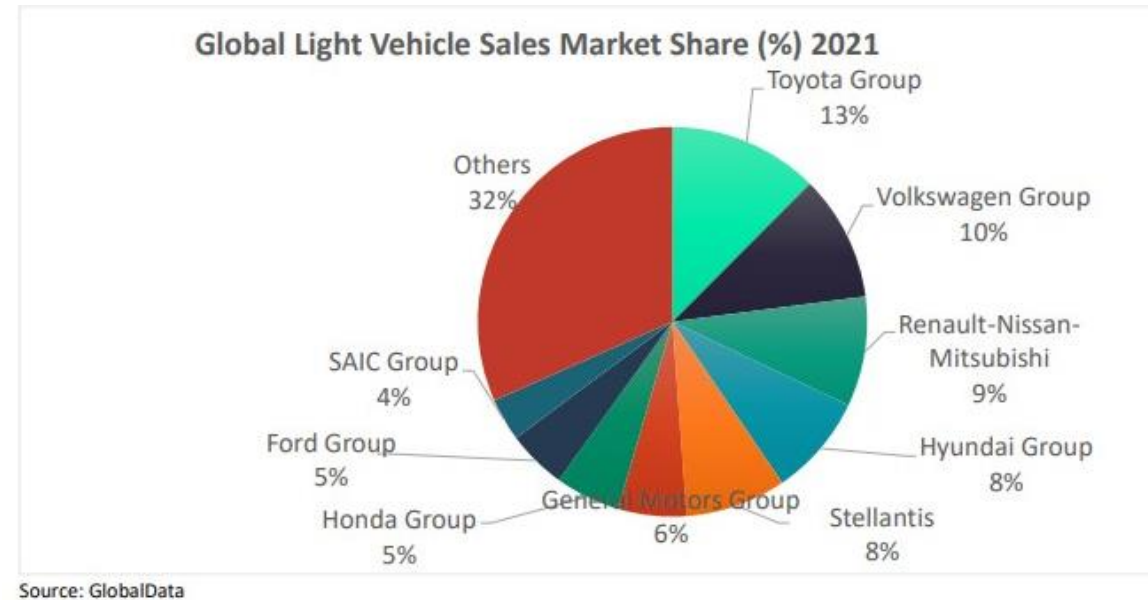
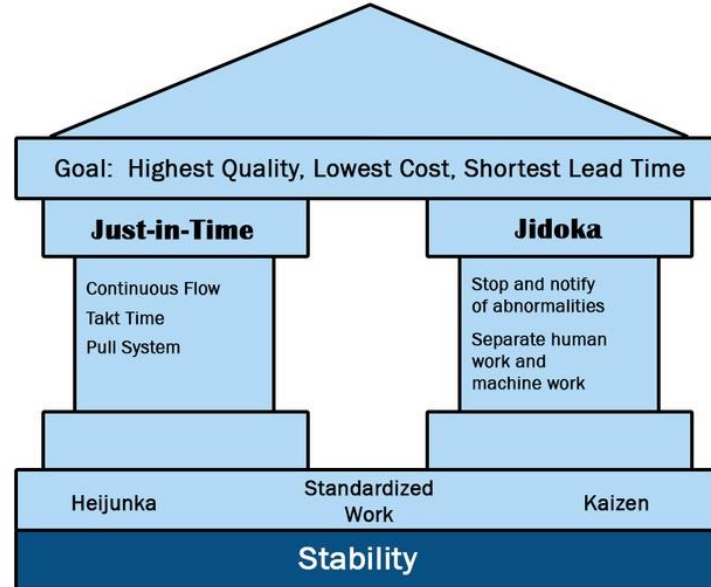
The anomaly proved that an atom is mostly empty and something very solid and positively charge is inside (**proton**)

Anomaly Detection is a Backbone of a leading operational management system

- Toyota has long been recognized as a leader in the automotive manufacturing and production industry
- Toyota Production System (TPS): one of the most productive, efficient and beautiful management systems ever designed.

Two pillars of TPS:

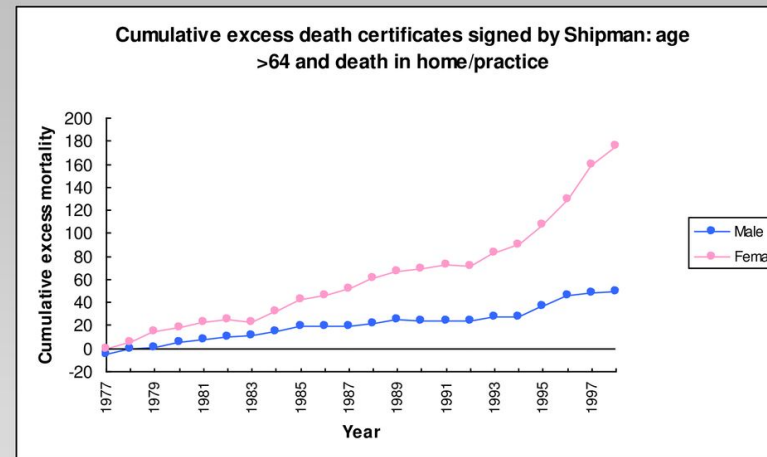
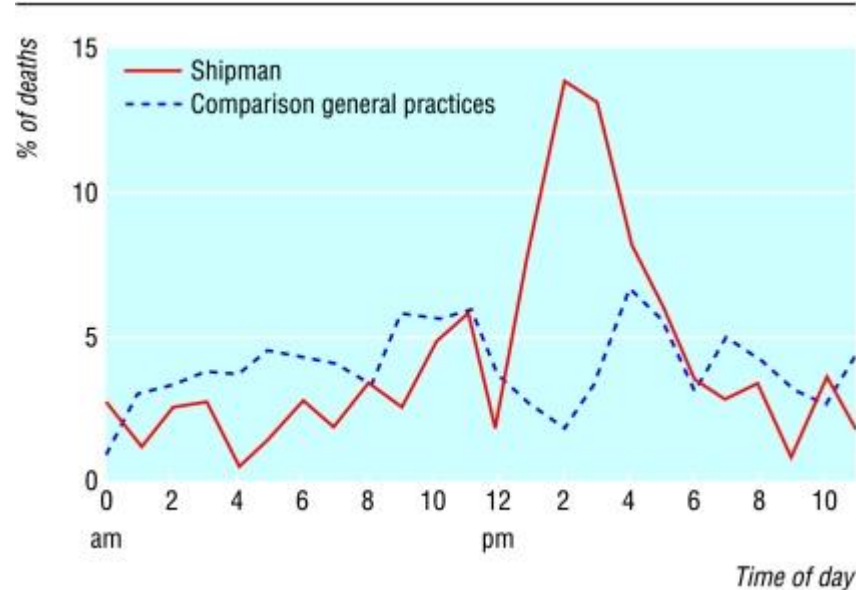
- **Jidoka:** Provide machines with the capability to **detect and flag anomaly** in the production line
- Just in Time: Making only what is needed, only when it is needed, and only in the amount that is needed



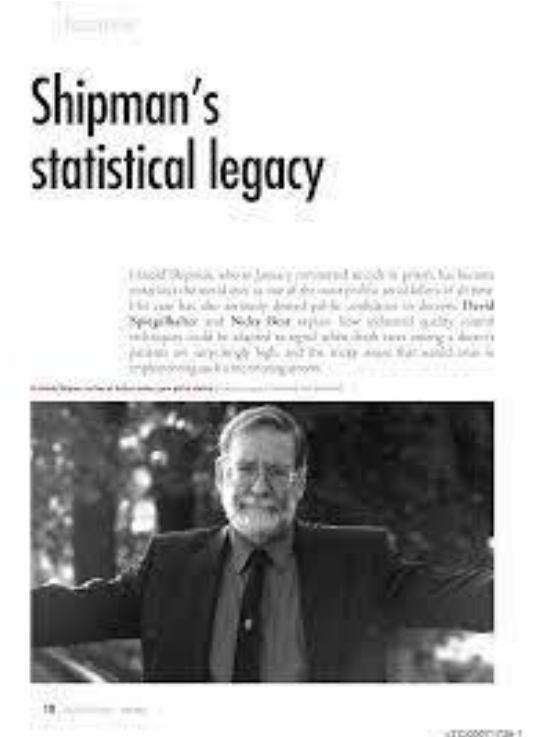
Anomaly Detection could have Prevented Mass Murders

Harold Shipman was an English general practitioner and serial killer:

- In January 2000, Shipman was found guilty of the murders of 15 patients under his care
- But the total victim is estimated as many as 250
- The Shipman case, triggered the development on statistical forensic study since the mass murders could have been prevented with the anomaly detection

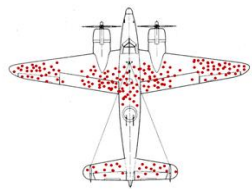


(NB: Shipman Inquiry total of definite or probable victims:
189 female > 65, 55 male over 65)

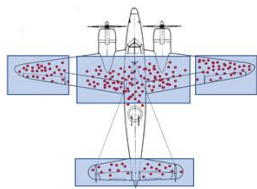


What Machine Learning?

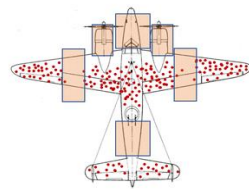
Why Machine Learning?



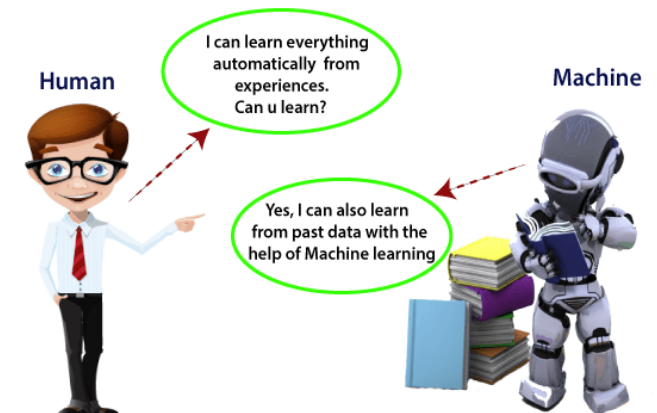
Our data if only from returning flights. Here we is a visualization of the places that bullet holes were observed.



And initial guess at how to fix this might be to apply additional armor plating to the parts of the plane with the most holes...



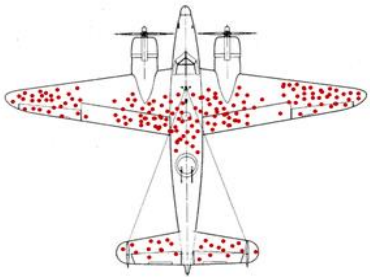
.... However this is where planes that *returned* had bullet holes. The planes we want to protect are the ones that did *not* return, so we should place armor there.



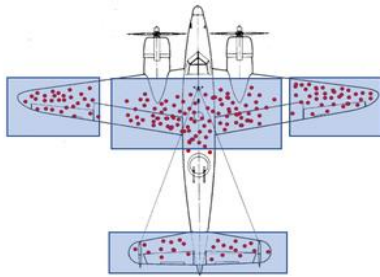
Why and What is Machine Learning?

Human are continuously under the influence of many biases. One of the example is **survivorship bias**:

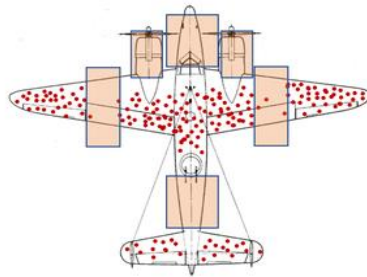
- During WW2, The American military asked mathematician Abraham Wald to study how best to protect airplanes from being shot down.
- The military knew armor would help but couldn't protect the whole plane or would be too heavy to fly well.
- Initially, their plan had been to examine the planes returning from combat, see where they were hit the worst and then reinforce those areas.
- But Wald realized they had fallen prey to survivorship bias, because their analysis was missing a valuable part of the picture: the planes that were hit but that hadn't made it back
- The bullet holes they were looking at actually indicated the areas a plane could be hit and keep flying – exactly the areas that *didn't* need reinforcing.



Our data if only from returning flights. Here we is a visualization of the places that bullet holes were observed.



And initial guess at how to fix this might be to apply additional armor platting to the parts of the plane with the most holes...



.... However this is where planes that *returned* had bullet holes. The planes we want to protect are the ones that did *not* return, so we should place armor there.

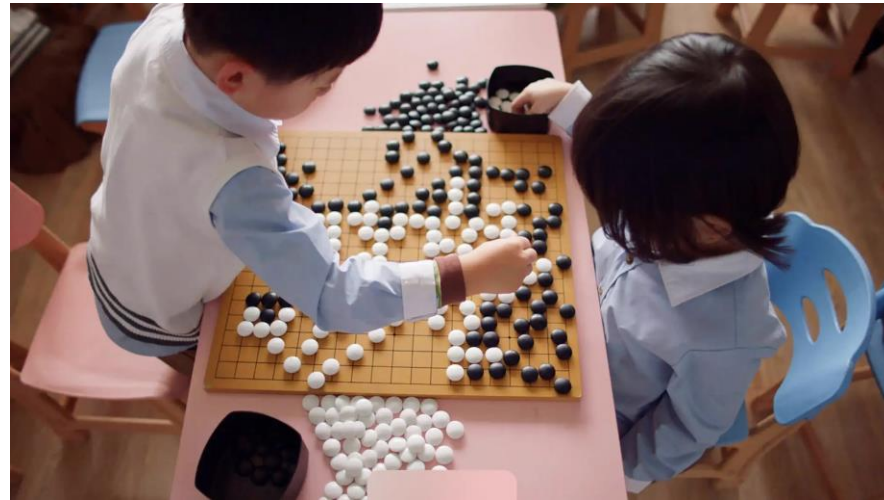
Other biases:

- Reverse survivor bias
- Publication bias
- Immortal time bias
- Etc...

Why and What is Machine Learning?

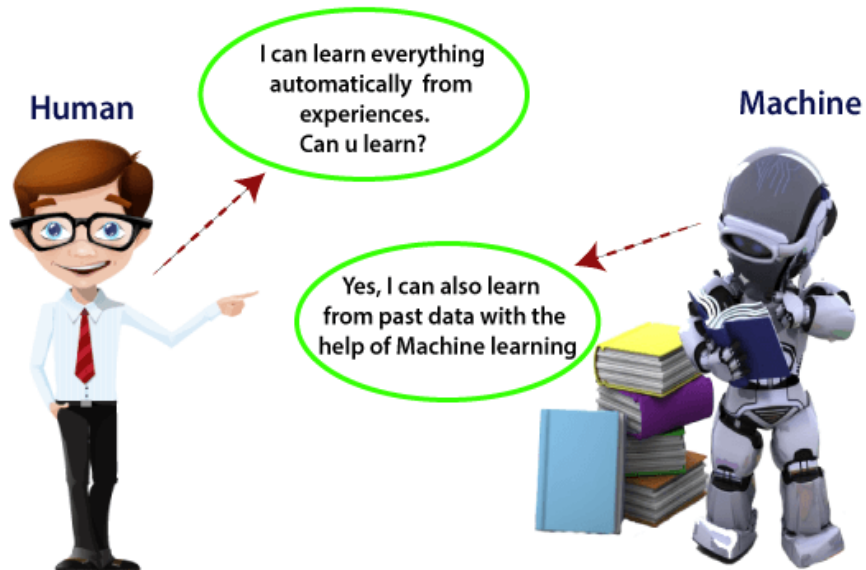
Machine learning is powerful. AlphaGo has defeated a Go world champion:

- Originated in China, Go is profoundly complex even though the rules are simple
- Two players, using either white or black stones, take turns placing their stones on a board. The goal is to surround and capture their opponent's stones or strategically create spaces of territory.
- There are 10^{170} possible board configurations, not only a way more than chess configuration but more than the number of atoms in the universe

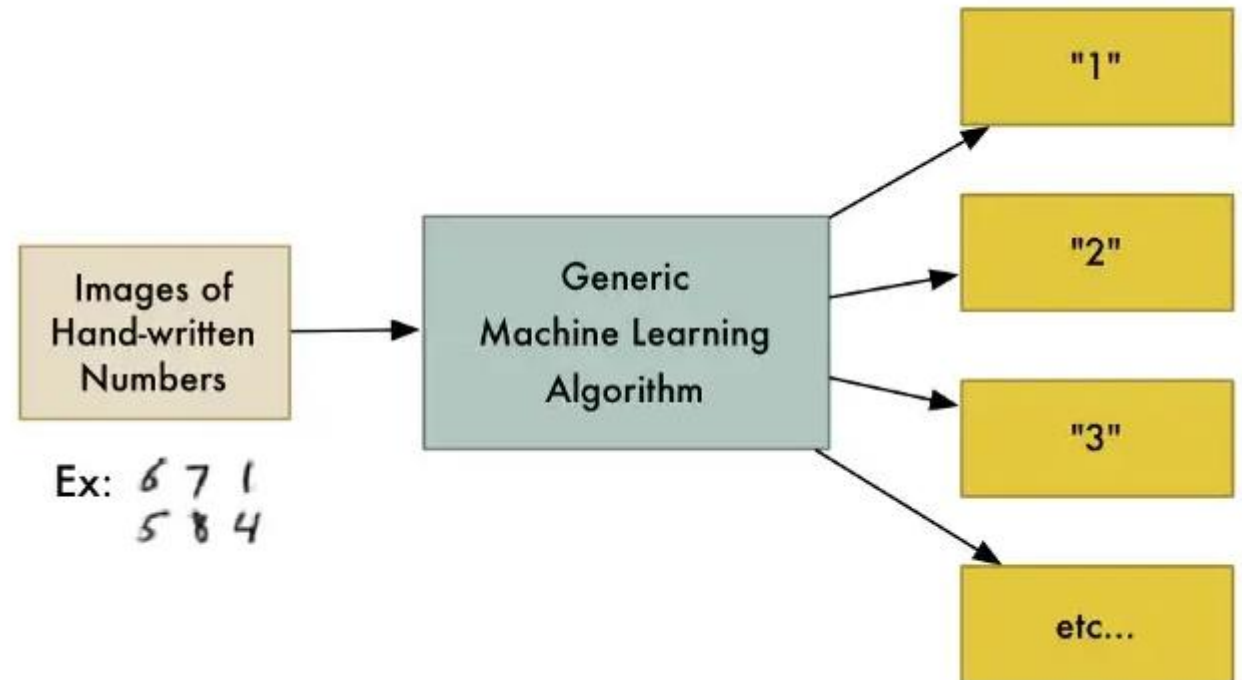


Why and What is Machine Learning?

If human learn from experiences, computer learn from the data



As an example, we can show and train computer many pictures of number so the computer could learn to read and classify number

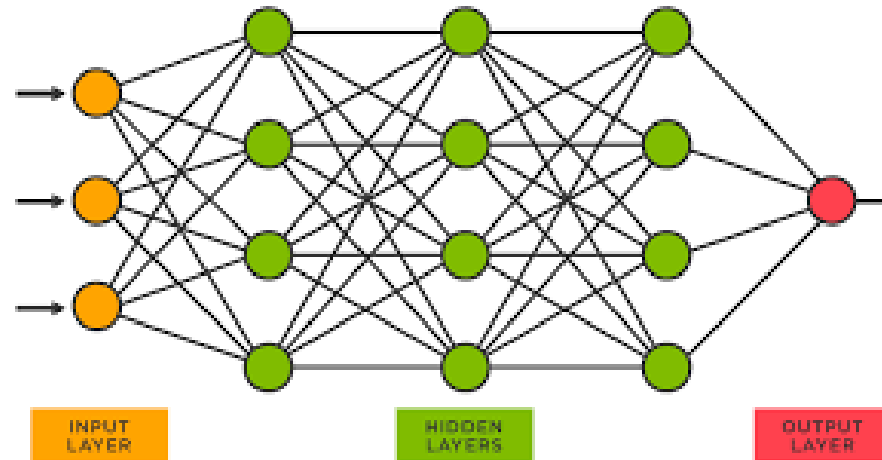
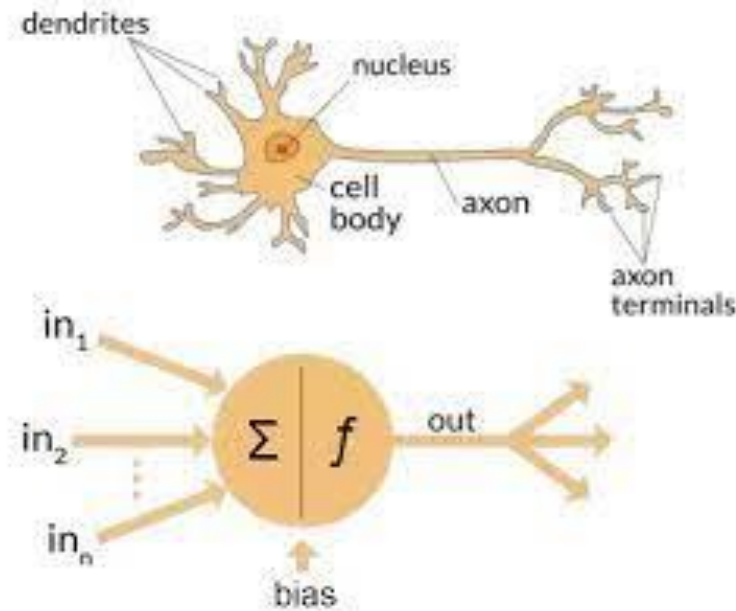


Machine learning = the use and development of computer systems that can learn and adapt **without following explicit instructions**, by using algorithms and statistical models to analyze and draw inferences from patterns in data

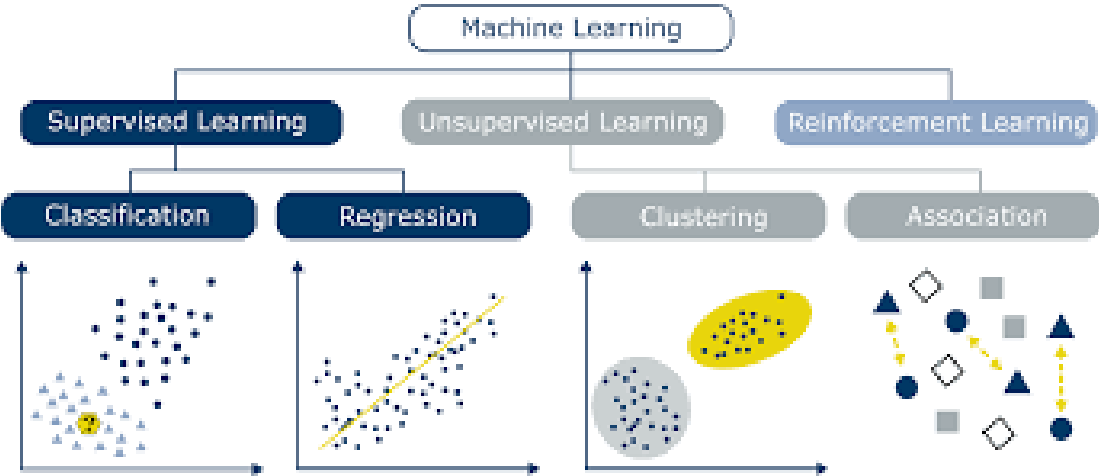
Why and What is Machine Learning?

Machine learning is a branch of artificial intelligence (AI) which focuses on the use of data and algorithms to imitate the way that humans learn:

- As an example, neural network is a series of algorithm that endeavors to recognize underlying relationship in a set of data through a process that mimics the way human brain operates



Machine-Learning Application



Machine learning is model dependance (learning from data)

Traditional Programming



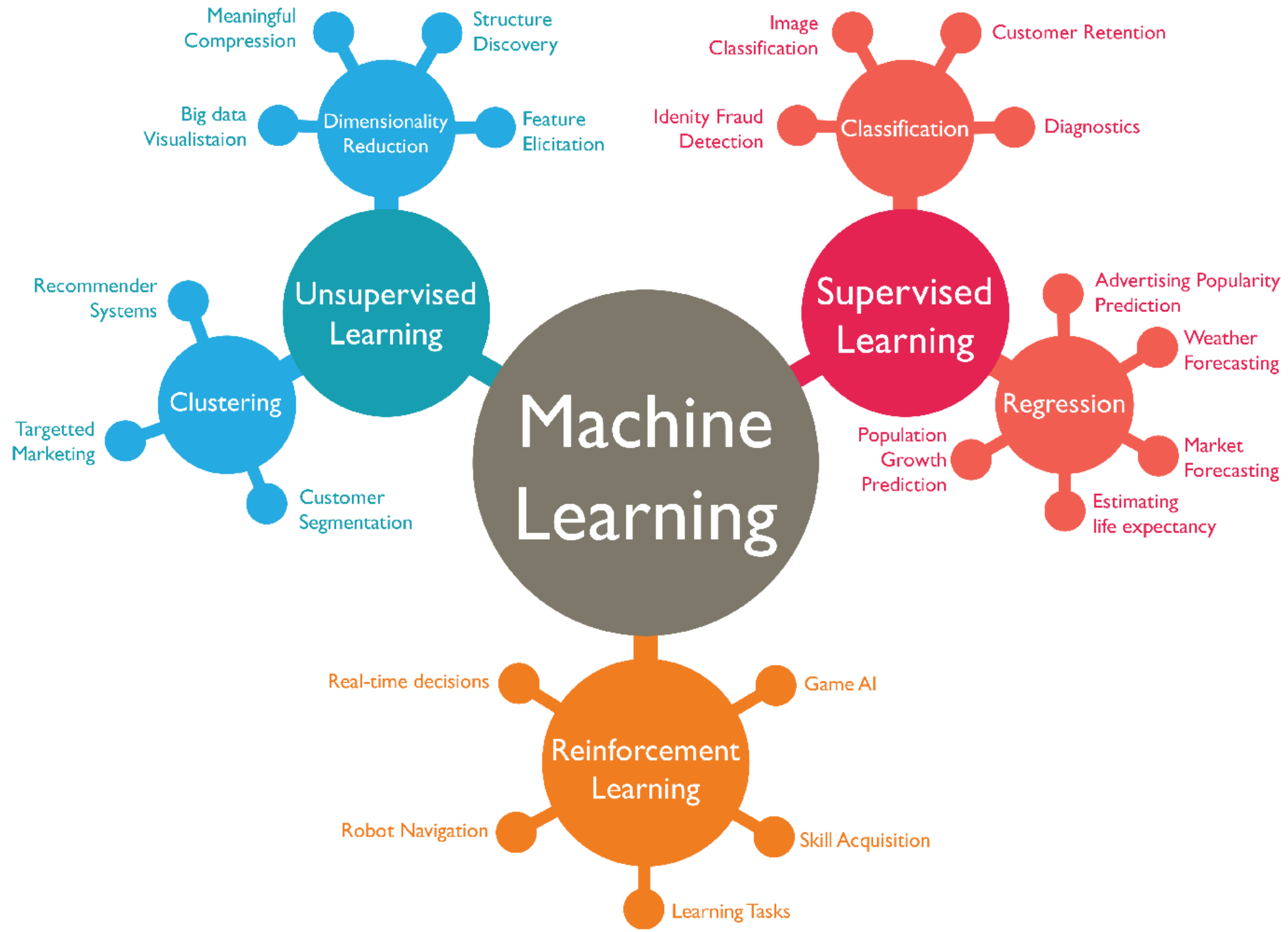
Machine Learning



Regression/Prediction

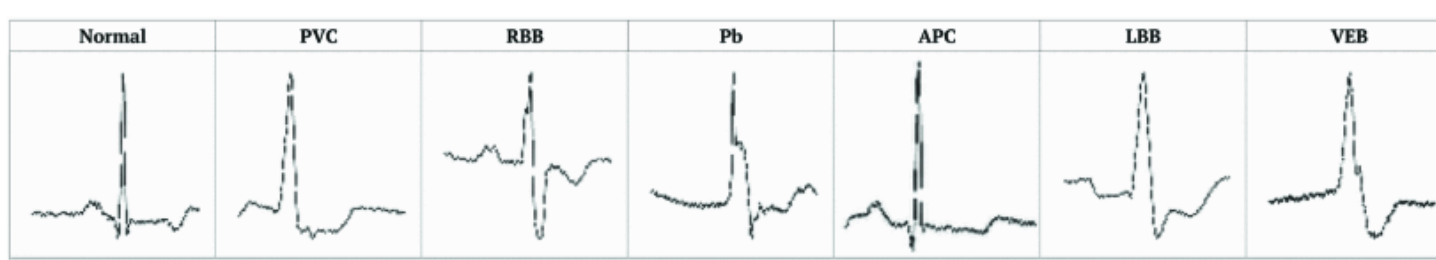


Finance

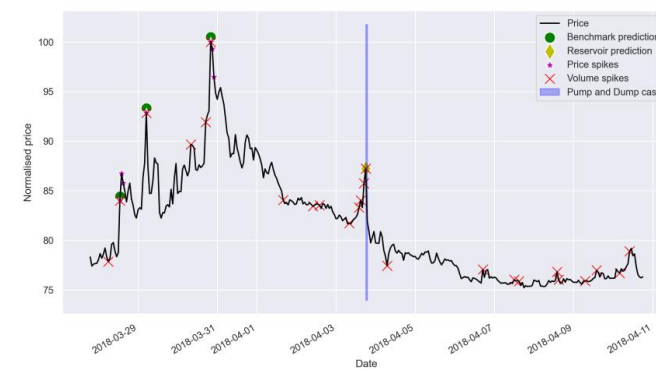


	Supervised Learning	Unsupervised Learning	Deep Learning	Ensemble Learning
Techniques	<ul style="list-style-type: none"> • Classification • Logistic or Linear Regression • Multivariate regression 	<ul style="list-style-type: none"> • Clustering • Natural Language Processing (NLP) 	<ul style="list-style-type: none"> • Neural Networks 	<ul style="list-style-type: none"> • Random Forest
Potential customer problem	<ul style="list-style-type: none"> • Classifying product segments • Predicting machine problems to avoid downtime • Identifying linear relationships in machine performance 	<ul style="list-style-type: none"> • Identifying similarities and characteristics e.g. if it is shaped like a 'car', it could be grouped as a 'car' 	<ul style="list-style-type: none"> • Solving complex problem patterned towards intelligent beings using multiple sources of dissimilar inputs. e.g "smart" plant facility 	<ul style="list-style-type: none"> • Identifying linear and probabilistic relationships between an outcome and its components using Bootstrap Aggregation
Data/Input format	<ul style="list-style-type: none"> • Structured imagery, numeric, strings/characters, sensory data with well identified labels 	<ul style="list-style-type: none"> • Unstructured imagery, numeric, strings/characters, sensory data with semi-identified labels 	<ul style="list-style-type: none"> • Multiple structured data types e.g., imagery, audio, numeric, strings/characters, sensory, etc., from multiple sources 	<ul style="list-style-type: none"> • Structured imagery, numeric, strings/characters, sensory data with well identified labels
Applications	<ul style="list-style-type: none"> • Predictive and prescriptive maintenance • Predictive decision support • Production optimization • Product segmentation • Predictive root cause analysis • Asset management and quality control 	<ul style="list-style-type: none"> • Pattern processing • Anomaly or defect detection for quality control • Asset performance management 	<ul style="list-style-type: none"> • Complex production optimization • Quality control • Speech and pattern recognition • Complex anomaly detection • Autonomous processing • Asset management 	<ul style="list-style-type: none"> • Powerful and accurate Predictive and prescriptive maintenance • Predictive decision support • Product segmentation • Predictive root cause analysis • Asset performance analysis and quality control

ML Techniques for AI Applications in Manufacturing



Healthcare

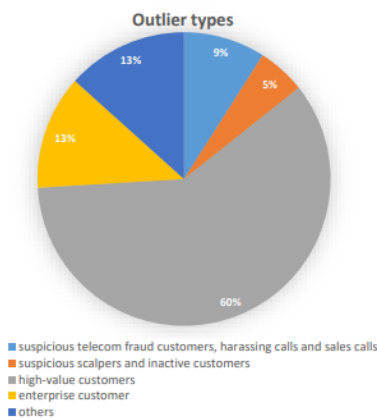


Insider trading

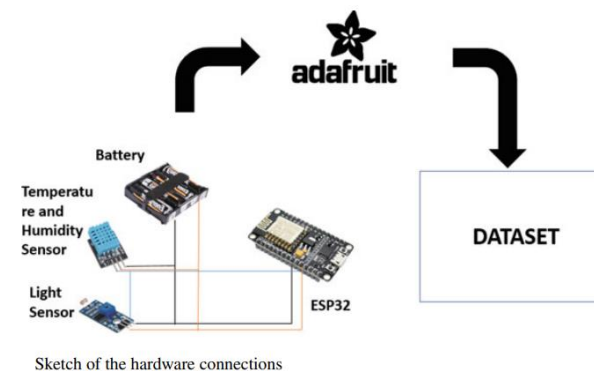
Machine-Learning Based Anomaly Detection

date	account_id	type	amount
1995-06-22	1	WITHDRAWAL	200.0
1995-07-22	1	WITHDRAWAL	5300.0
1995-08-21	1	WITHDRAWAL	7500.0
1995-08-31	1	WITHDRAWAL	14.6
1995-09-05	1	WITHDRAWAL	2452.0

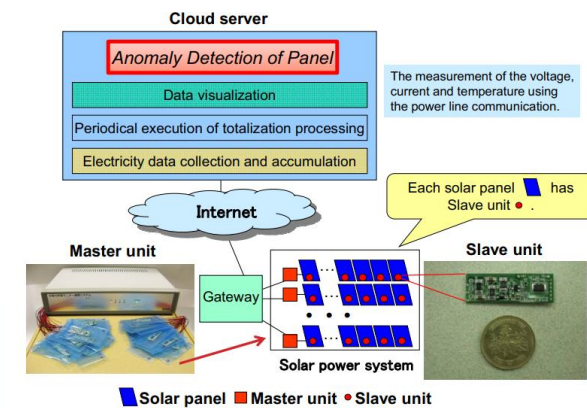
Transaction Bank fraud



Customer behavior



IOT



Renewable energy

Machine-Learning Based Anomaly Detection

Anomaly Detection on Bank transactions

Step-1: Load the data

- Real transaction data from Czech

date	account_id	type	amount
1995-06-22	1	WITHDRAWAL	200.0
1995-07-22	1	WITHDRAWAL	5300.0
1995-08-21	1	WITHDRAWAL	7500.0
1995-08-31	1	WITHDRAWAL	14.6
1995-09-05	1	WITHDRAWAL	2452.0

Step-2: Engineer the Data

Creating two new data features

- *sum_5days*: the accumulative withdrawal amounts from an account in the previous 5 days (including the current day).
- *count_5days*: the count of withdrawal transactions from an account in the previous 5 days (including the current day).

date	account_id	type	amount	sum_5days	count_5days
1995-06-22	1	WITHDRAWAL	200.0	200.0	1.0
1995-07-22	1	WITHDRAWAL	5300.0	5300.0	1.0
1995-08-21	1	WITHDRAWAL	7500.0	7500.0	1.0
1995-08-31	1	WITHDRAWAL	14.6	14.6	1.0
1995-09-05	1	WITHDRAWAL	2452.0	2452.0	1.0
1995-09-20	1	WITHDRAWAL	700.0	700.0	1.0
1995-09-30	1	WITHDRAWAL	14.6	14.6	1.0
1995-10-05	1	WITHDRAWAL	2452.0	2452.0	1.0
1995-10-20	1	WITHDRAWAL	2900.0	2900.0	1.0
1995-10-31	1	WITHDRAWAL	14.6	14.6	1.0
1995-11-05	1	WITHDRAWAL	2452.0	2452.0	1.0
1995-11-19	1	WITHDRAWAL	1900.0	1900.0	1.0
1995-11-30	1	WITHDRAWAL	14.6	14.6	1.0
1995-11-30	1	WITHDRAWAL	870.0	884.6	2.0

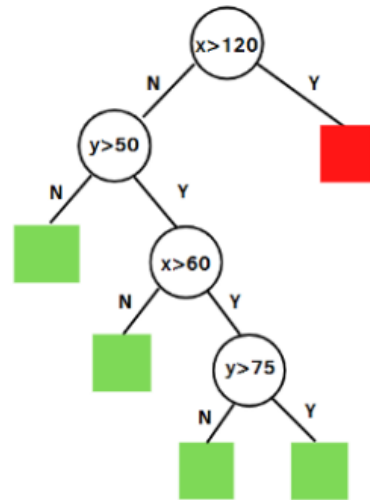
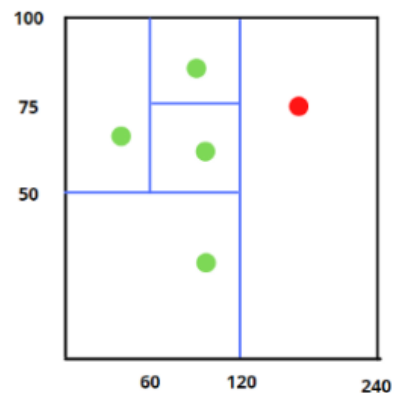
Machine-Learning Based Anomaly Detection

Anomaly Detection on Bank transactions

Step-3: Apply Isolation Forest algorithm

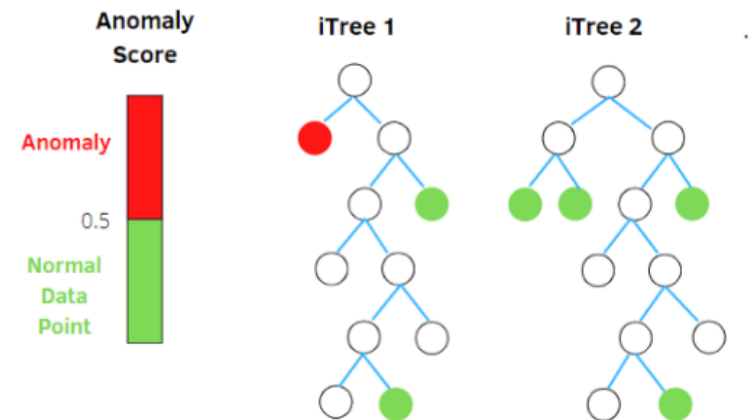
1. Randomly select **two** features.
2. Split the data points by randomly selecting a **value** between the minimum and the maximum of the selected features.

The partition of observations is repeated recursively until all the observations are isolated.



Step-4: Calculate the anomaly score

- Data with high anomaly score is easy to isolate (filtered in high node)

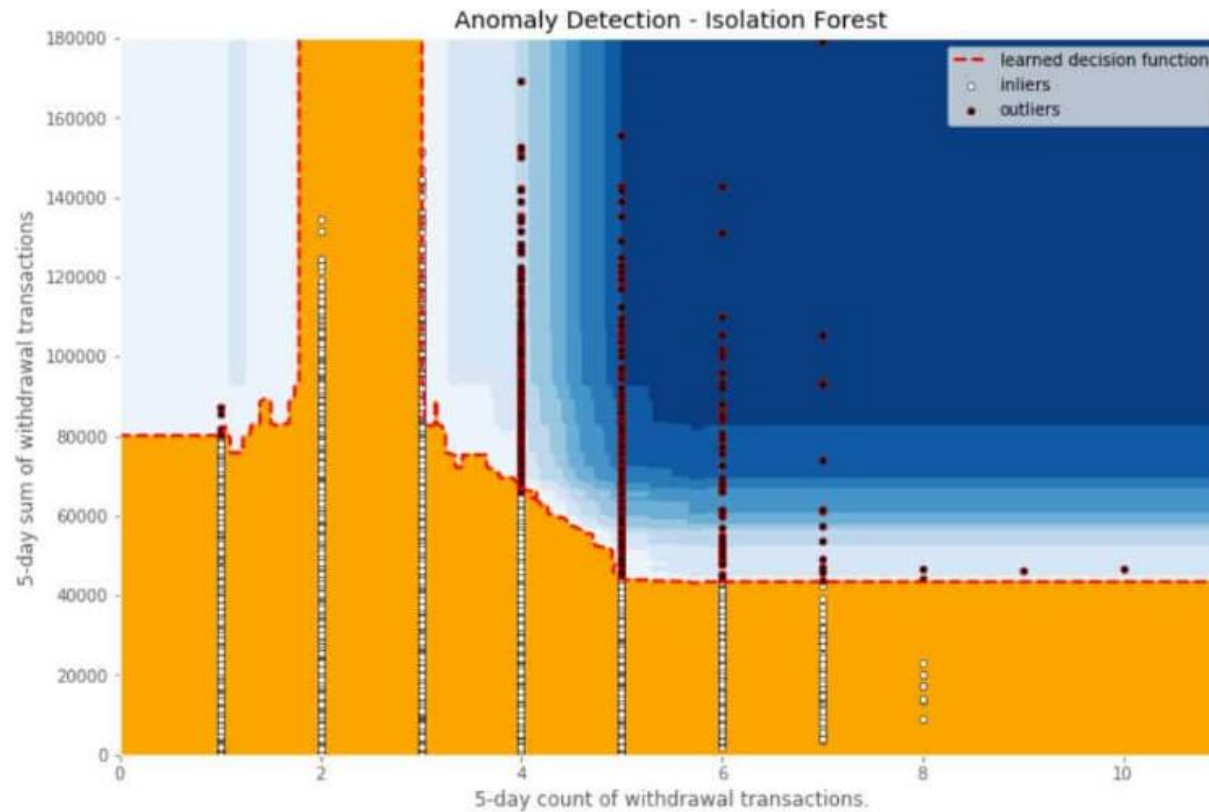


Machine-Learning Based Anomaly Detection

Anomaly Detection on Bank transactions

Step-5: Make Decision

- For example, define the top 0.1% of the anomaly score as anomaly and call to clarify the transactions.



Further analysis on Outlier characteristics

- the account had 1 withdrawal transaction of over ~80K amount
- the account had over ~65K accumulative amount and more than 4 withdrawal transactions within the past 5 days
- the account had over ~45K accumulative amount and more than 5 withdrawal transactions within the past 5 days

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

- Useful for targeted marketing (75% of movies played in Netflix and 35% of Amazon sales are based on personalized-recommendation system)
- Also, to detect market anomalies (It could be very-specific market segment or fraud indication)
- Overcome marketing challenge:

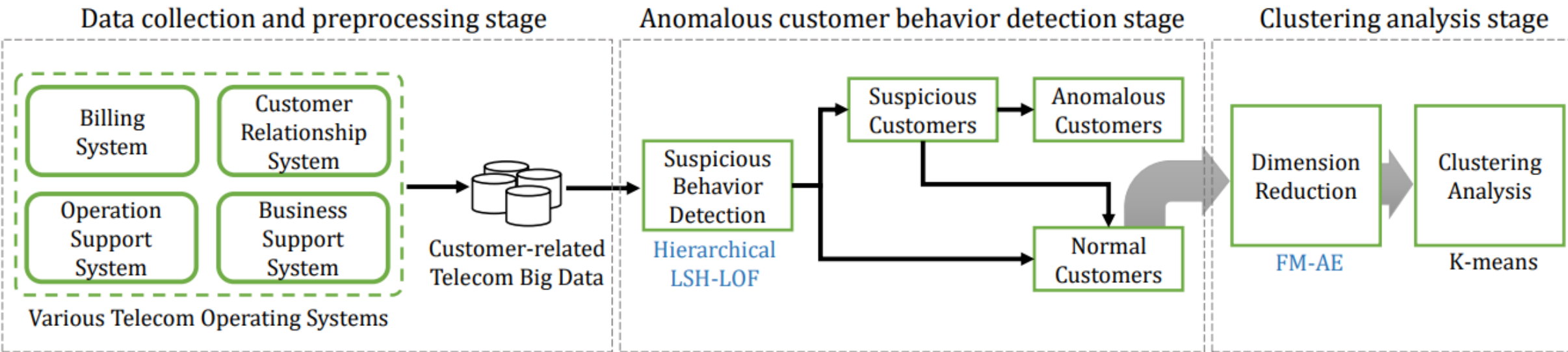
Six Challenges Marketers Face in 2021

1. The typical marketer spends eight hours a week studying data if you don't have a data analytics team dedicated to monitoring marketing metrics. They can better use their time to create strategic and innovative campaigns.
2. If you have a data analytics team, they simply analyze data, calculate, and visualize marketing stats for C-level executives. Instead of working on new initiatives to keep your business ahead of the competition these valuable resources are wasted.
3. Manual marketing metrics tracking overlooks critical anomalies and trends. Out of hundreds of marketing metrics, you can track manually only a few. This limited analysis leaves up to 60% of the data unavailable for analytics.
4. Not all marketing metrics are created equal. While some metrics will have a more significant impact on revenue, you may miss out on other critical metrics by focusing only on the ones you see as the most essential. Manual marketing metrics monitoring may overlook revenue-damaging glitches instead of proactively discovering and resolving them.
5. Google Analytics and other legacy solutions offer useful alert features that notify you when your marketing metrics critically change. This helps you keep track of several metrics at the same time. The alerts, however, are based on static data, while nothing in marketing is ever static.
6. Various moving marketing components include campaigns, channels, pricing, and seasonal changes. You will either get overwhelmed by too many static alerts or waste valuable time changing parameters.

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

- Data in this example are the telecom data from major Internet Service Provider (ISP) in China



Analysis Flow

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-1: Data collection and Pre-processing stage

- Big data: Millions customers data which has 75 features collected from 4 subsystems (Billing system, customer relationship system, operating-support system, business-support system)

Source	Features	Source	Features
Billing system	account balance	Operation support system	number of apps used in the current month
	package price of current month		number of apps used in the last month
	activating international call service or not		number of apps used the month before last
	activating international roaming service or not		DOU in current month
	activating national free flow package in the past year or not		DOU in last month
	activating holiday flow package in the past year or not		DOU in the month before last
	activating 4G free time flow package in the past year or not		average DOU
	average times of package change in the past 3 months		average DOU in the past 3 months
	average times of service activating in the past 3 months		average active phone call days per month
	average times of complaints in the past 3 months		average call times in the past 3 months
	average times of payment in the past 3 months		average call duration in the past 3 months
	times of payment in the past 3 months		average call times during idle time in the past 3 months
	average ARPU in the past 3 months		average call duration during idle time in the past 3 months
	average ARPU		average number of messages sent in the past 3 months
	total shutdown times		times of provincial roaming in the past year
	total shutdown days in the past 3 months		duration of provincial roaming in the past year
	times of arrears in the past 3 months		number of provincial roaming places in the past year
	amount owed in the past 3 months		international roaming times in the past year
	total shutdown times in the past 3 months		duration of international roaming time in the past year
			International roaming places in the past year

ARPU = Average Revenue per User DOU = Dataflow of Usage

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-1: Data collection and Pre-processing stage

- Big data: Millions customers data which has 75 features collected from 4 subsystems (Billing system, customer relationship system, operating-support system, business-support system)

Customer relation- ship system	number of mobile phone numbers in comm-circle
	number of landline phone numbers in comm-circle
	proportion of China Mobile users in comm-circle
	number of mobile phone numbers in strong comm-circle
	number of landline phone numbers in strong comm-circle
	number of mobile phone numbers in outbound comm-circle
	number of landline phone numbers in outbound comm-circle
	average ARPU of mobile users in comm-circle
	average DOU of mobile users in comm-circle
	monthly average call duration of mobile users in comm-circle
Business support system	monthly average call duration of landline users in comm-circle
	monthly average call times of mobile users in comm-circle
	monthly average call times of landline users in comm-circle
	number of mobile phone numbers under ID card
	credit score
	length of access
	customer star degree
	key customer degree
	blacklist or not
	age
	gender
	current account points
	points used
	using 4G terminal or not
	enterprise customer or not
	binding family service or not
	contract customer or not

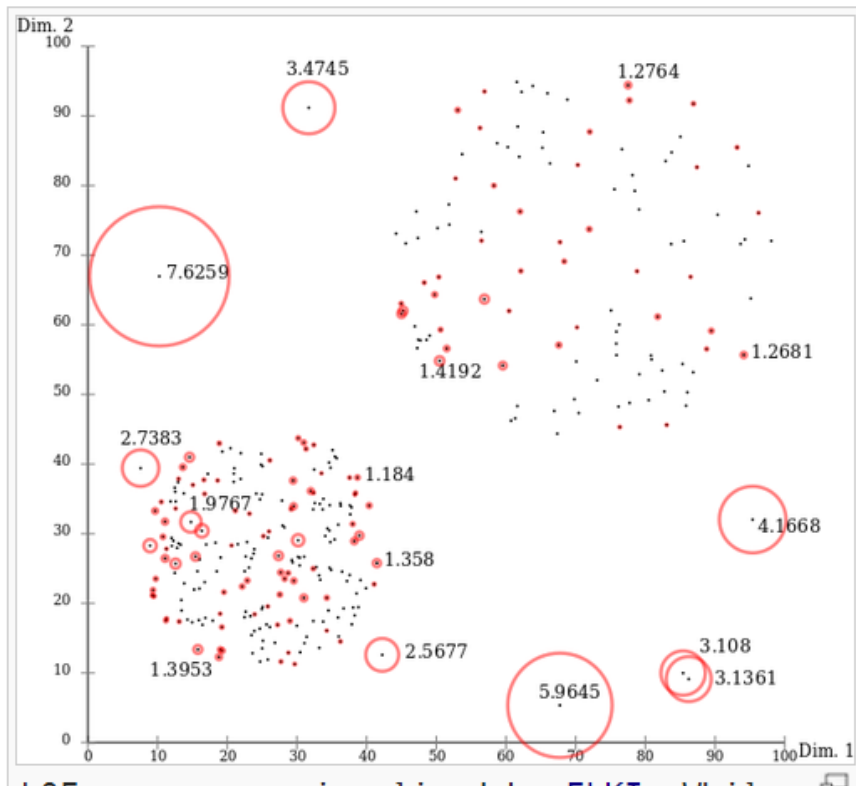
- Pre-processing: Data normalization

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-2: Anomaly detection with Local Outlier Factor (LOF) algorithm

- LOF produces an anomaly score that represents data points which are outliers in the data set
- It works by measuring the local density of a given data point with respect to the data points near it
- Local density is determined by estimating distances between data points that are neighbors (k-nearest neighbors)
- The ones with the lesser densities are considered as the outliers

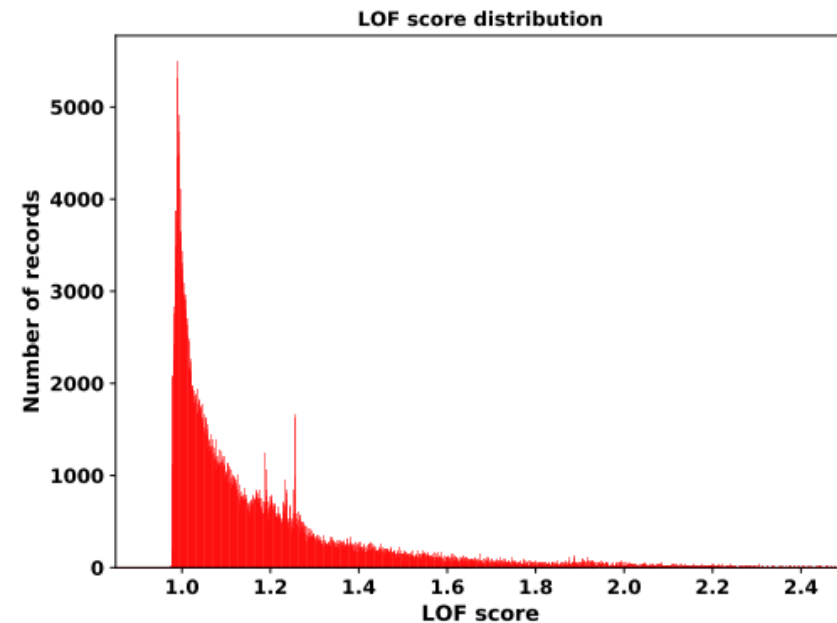


Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-3: Define the outlier & inlier

- The outlier is defined as the top 3% of the anomaly score and the rest (97%) is defined as inlier



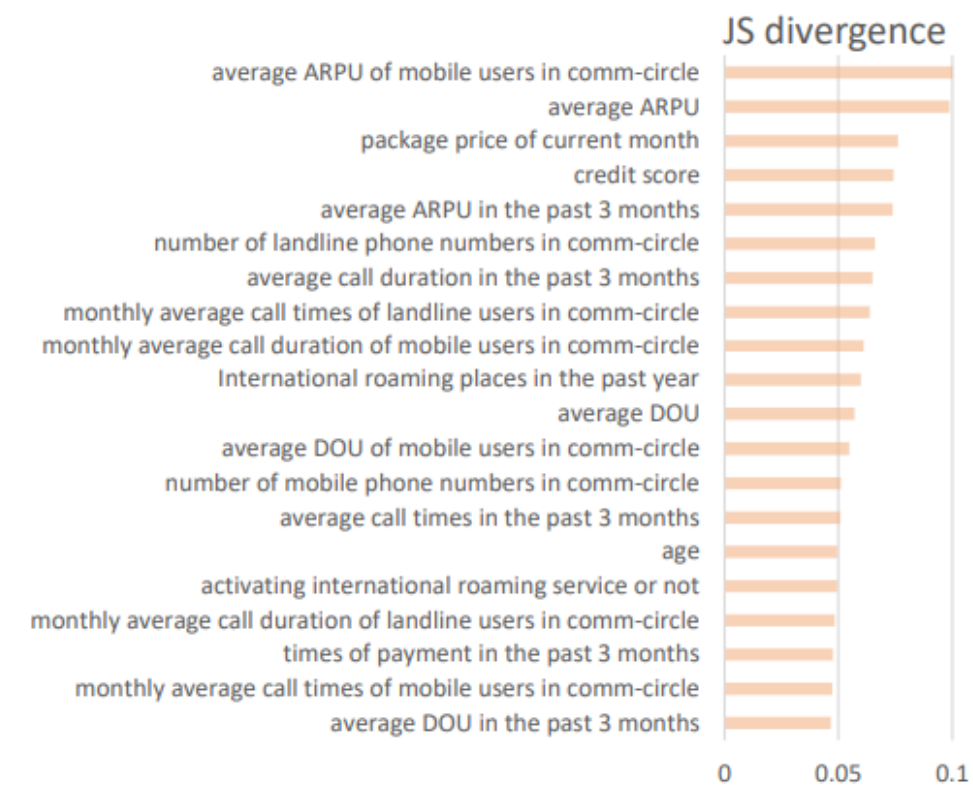
Step-4: Outlier analysis

- We want to compare the feature distribution of outlier and inlier
- One tool to compare two distributions is called Jensen-Shannon Divergence (JS Divergence)
- The smaller JS divergence is, the more similar two distributions are, and vice versa.
- Since analyzing 75 features are too time consuming, several features with highest JS Divergence were further analyzed

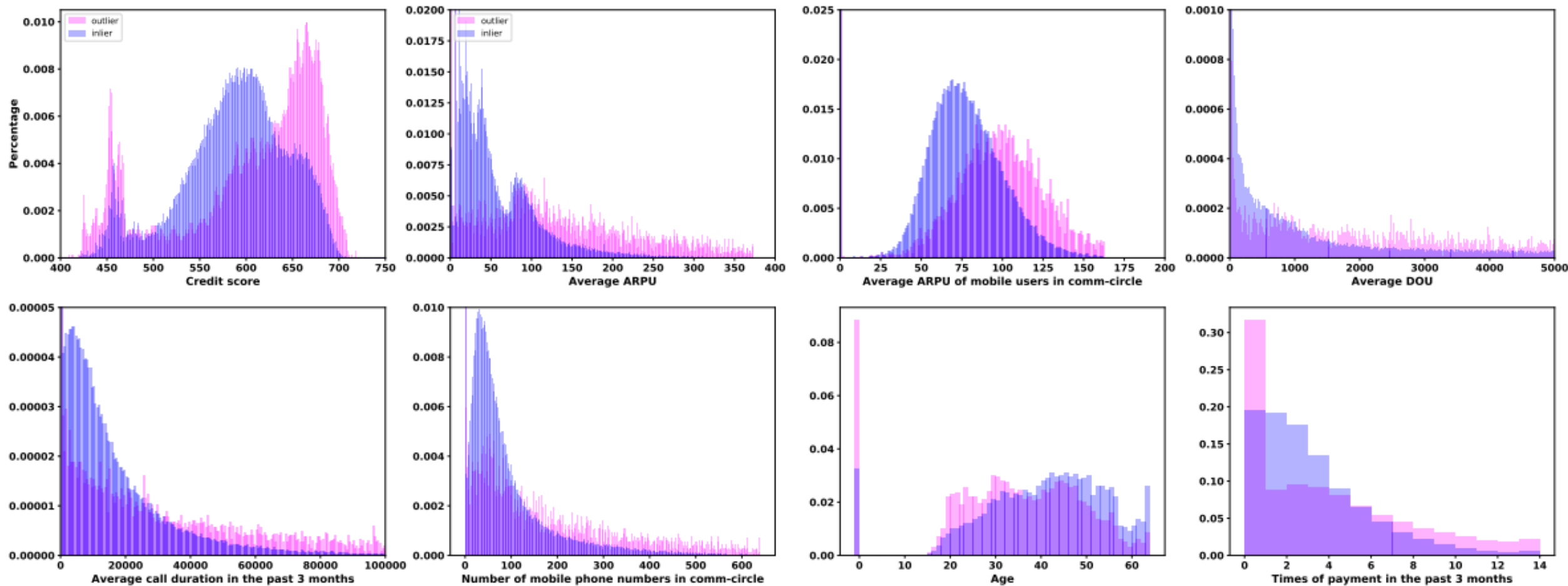
Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-4: Outlier analysis



Feature	in/outlier	Mean	Variance
Credic score	inlier	588.26	3205.74
	outlier	602.91	6076.90
Average ARPU	inlier	57.97	2649.68
	outlier	139.19	19522.04
Average ARPU of mobile users in comm-circle	inlier	73.36	912.18
	outlier	83.32	3235.15
Average DOU	inlier	1996.51	1.545e7
	outlier	5470.04	6.267e7
Average call duration in the past 3 months	inlier	1.621e4	3.406e8
	outlier	3.693e4	2.204e9
Number of mobile phone numbers in comm-circle	inlier	89.55	1.046e4
	outlier	208.94	1.322e5
Age	inlier	42.67	237.19
	outlier	34.49	264.87
Times of payment in the past 3 months	inlier	2.88	8.60
	outlier	4.85	43.73



- Inlier distributions are likely to have normal, power law or Rayleigh distributions
- Outlier distributions are likely to be long tailed

Machine-Learning Based Anomaly Detection

Anomaly Detection on customer behavior

Step-5: Interpretation

- Among the top 20 features with highest JS Divergence, 8 are from the customer relationship system, 5 from the billing system, 5 from the operation support system and 2 from the business support system
- So now we are focusing to the features from customer-relationship system
- Especially with two features: “number of mobile phone numbers in comm-circle” and “number of mobile phone numbers in strong comm-circle”
- Define coefficient w as the ratio between “number of mobile phone numbers in comm-circle” and “number of mobile phone numbers in strong comm-circle”

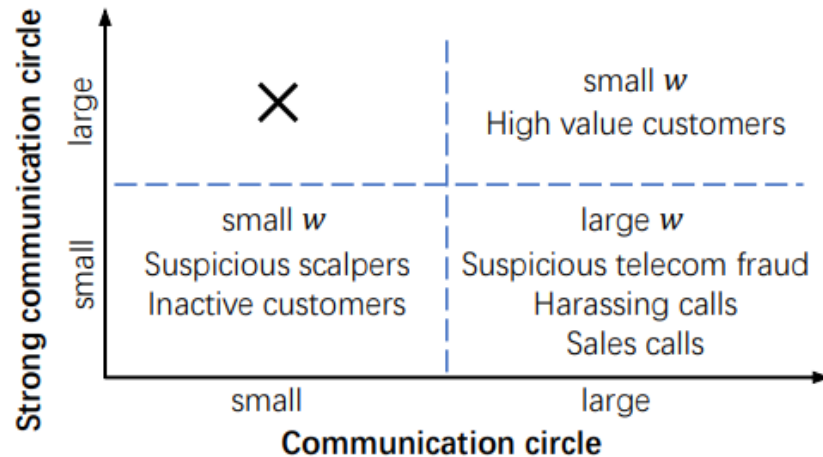


FIGURE 12: The weak communication coefficient.

Machine-Learning Based Anomaly Detection

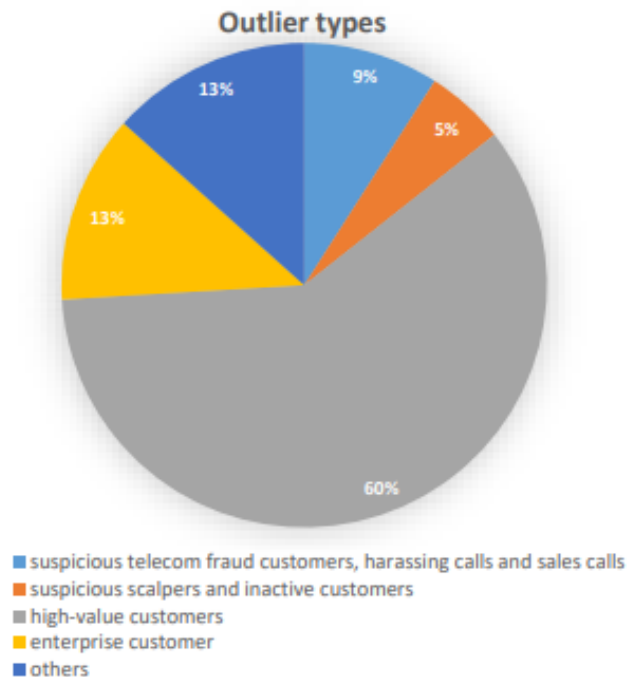
Anomaly Detection on customer behavior

Step-5: Interpretation

- To further distinguish between consumer type (high valued or suspicious fraud), the “average ARPU” features is added to help decision making. The large ARPU group are classified as high-value customers.

Step-6: Possible follow up

- Make a customer-retention program for the high-valued customer
- Notify police to further investigate the suspicious scalper
- etc., ...



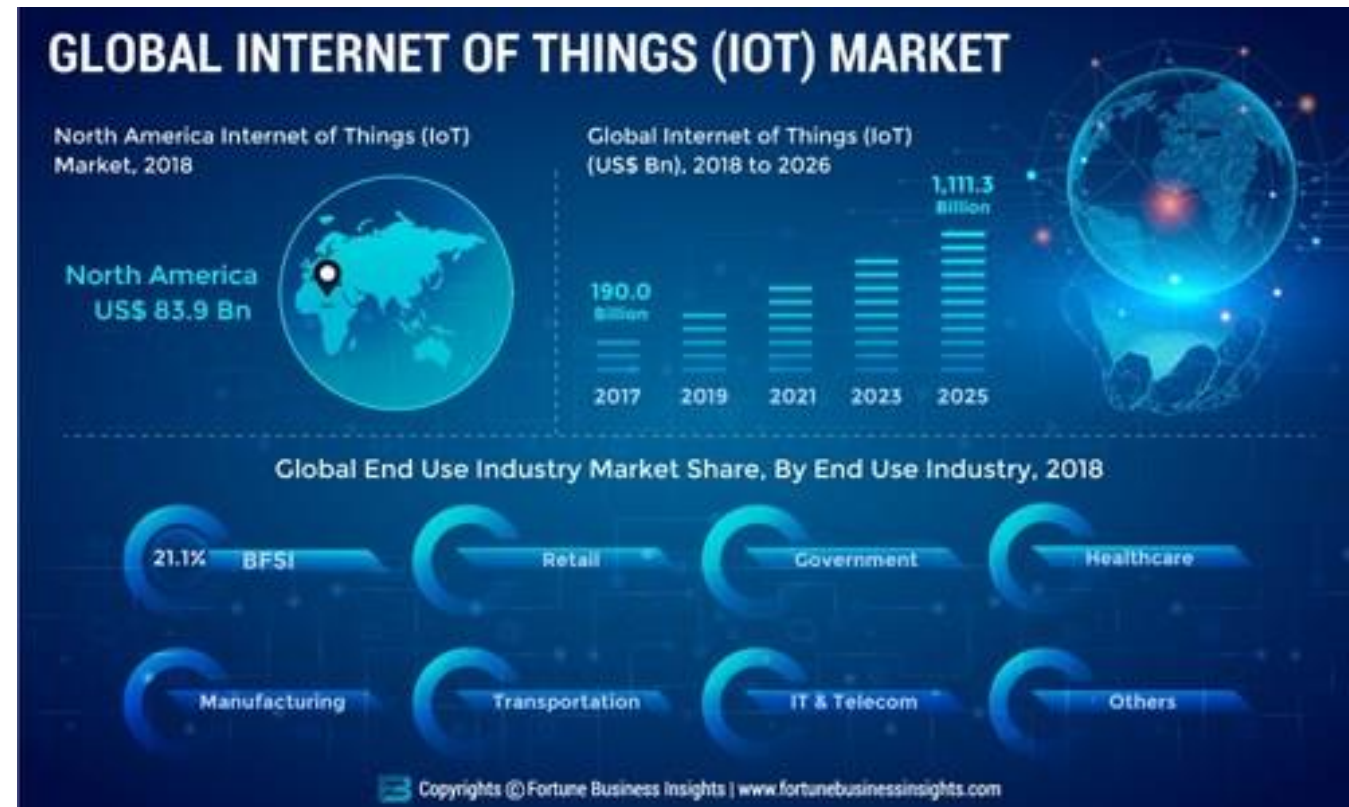
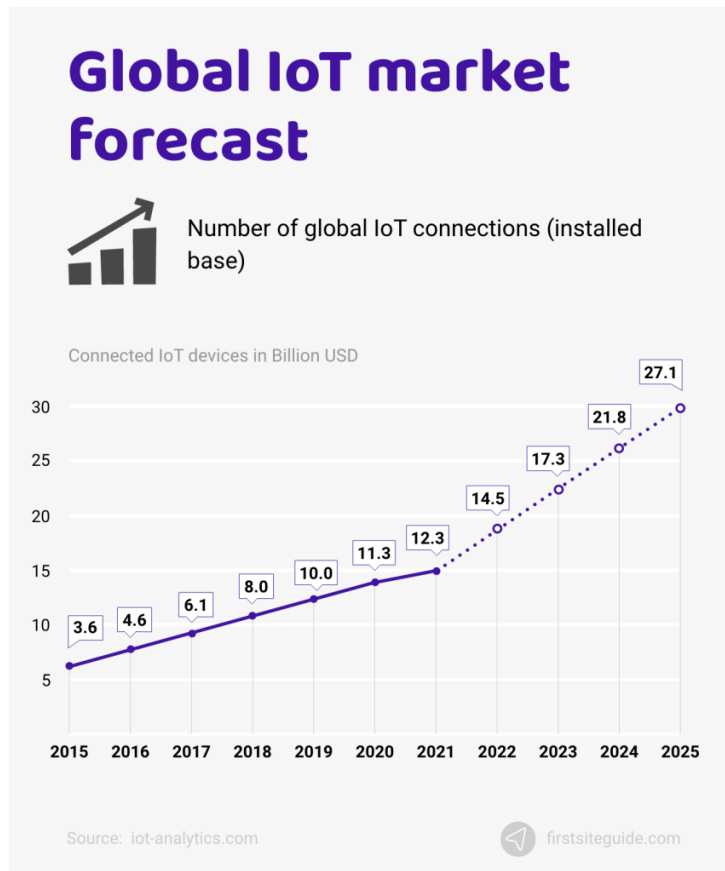
Further clustering on the inlier or normal customer

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
percentage	12.54%	12.18%	3.94%	37.63%	0.59%	33.11%
credit score	medium	high	low	medium	high	high
average ARPU	low	high	very low	low	high	medium
average DOU	low	very high	very low	very low	very high	high
age	high	medium	<i>mostly missing</i>	high	medium	low
length of access	medium	high	low	medium	high	medium
average call times in the past 3 months	low	high	zero	medium	very high	medium
current account points	medium	high	very low	medium	very high	high
points used	low	high	zero	low	high	medium
points used ratio	low	high	zero	low	high	medium
customer value	medium	very high	very low	medium	very high	high
loyalty	high	very high	low	medium	very high	medium
customer behavior	use low price package; neither frequently make phone calls nor frequently use Internet traffic; don't know how to exchange benefit with account points; high-aged	use high price package; frequently make phone calls and very frequently use Internet traffic; often exchange benefit with account points; middle-aged	use very low price package; never make phone calls and hardly use Internet traffic; never exchange benefit with account points; age mostly missing	use low price package; use mobile phones mostly for making phone calls; hardly use Internet traffic; don't know how to exchange benefit with account points; middle-aged and high-aged	use high price package; very frequently make phone calls and very frequently use Internet traffic; often exchange benefit with account points; middle-aged	use middle price package; frequently make phone calls and very frequently use Internet traffic; sometimes exchange benefit with account points; young people and middle-aged
customer profile	The Silent Type: the elders and middle aged who occasionally use mobile phones for communication and Internet access	Business People: the middle-aged who are heavy mobile phone users, pretty likely for business contact usage	Spare SIM Card: the phone number applied as an alternate; The Scalpers: criminal groups keeping thousands of phone numbers for scalping coupons	The Communication Type: the elders and middle-aged who mainly use mobile phones for making phone calls	Business People +: the middle-aged who are even heavier mobile phone users	The Mainstay: the young and middle-aged who use mobile phones to communicate, get information and entertainment

Machine-Learning Based Anomaly Detection

Anomaly Detection on Internet of Things (IoT)

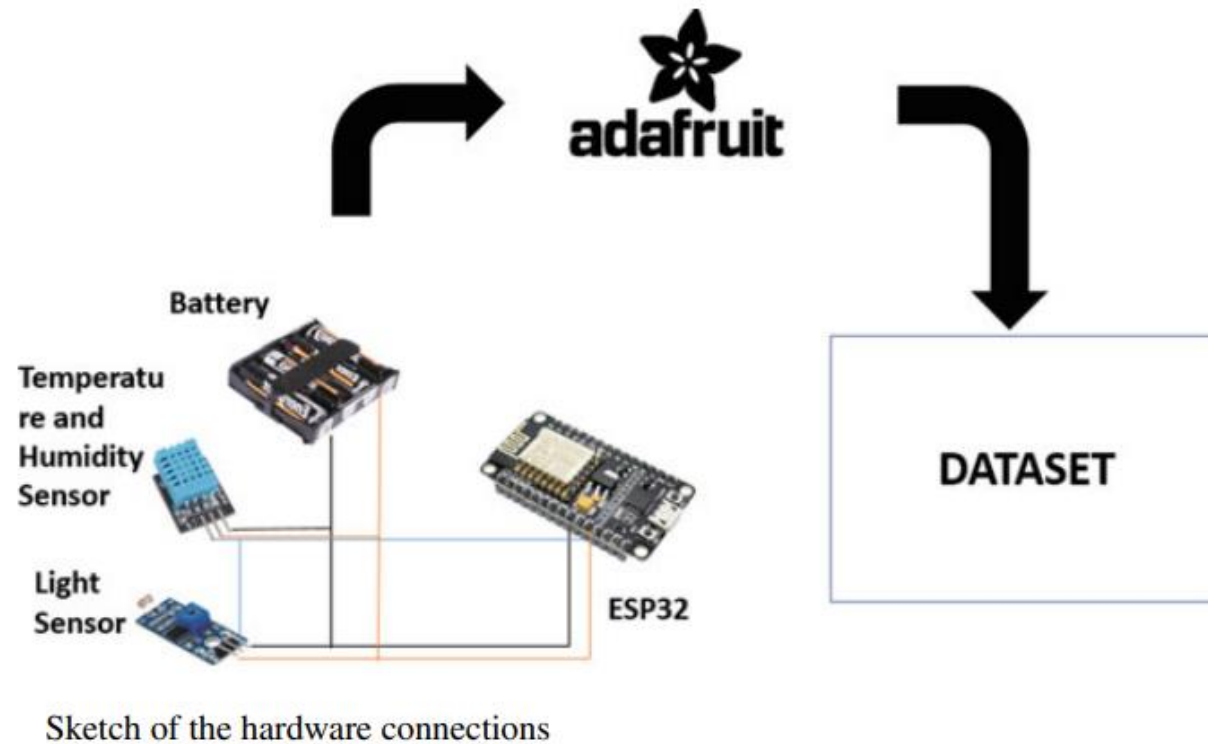
- The Internet of Things (IoT) describes the network of physical objects—“things”—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.
- Part of fourth industrial revolution or Industry 4.0
- Currently there are ~14 billions IoT devices are connected, with the (predicted) market of \$1,111 Billion in 2026



Machine-Learning Based Anomaly Detection

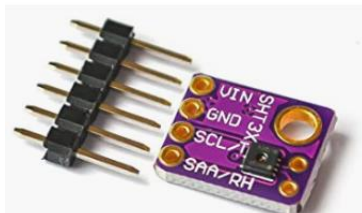
Anomaly Detection on Internet of Things (IoT)

- Therefore, it is extremely important to prevent and detect data anomaly coming from intruder (attack), faulty sensors or other sources
- Example:




ESP32 is a series of **low-cost**, low-power system on a chip microcontrollers with integrated Wi-Fi and dual-mode Bluetooth

- Processors:
 - CPU: Xtensa dual-core (or single-core) 32-bit LX6 microprocessor, operating at 160 or 240 MHz and performing at up to 600 DMIPS
 - Ultra low power (ULP) co-processor
- Memory: 320 KiB RAM, 448 KiB ROM
- Wireless connectivity:
 - Wi-Fi: 802.11 b/g/n
 - Bluetooth: v4.2 BR/EDR and BLE (shares the radio with Wi-Fi)
- Peripheral interfaces:
 - 34 × programmable GPIOs
 - 12-bit SAR ADC up to 18 channels
 - 2 × 8-bit DACs
 - 10 × touch sensors (capacitive sensing GPIOs)
 - 4 × SPI
 - 2 × I²S interfaces
 - 2 × I²C interfaces
 - 3 × UART
 - SD/SDIO/CE-ATA/MMC/eMMC host controller
 - SDIO/SPI slave controller
 - Ethernet MAC interface with dedicated DMA and planned IEEE 1588 Precision Time Protocol support^[4]
 - CAN bus 2.0
 - Infrared remote controller (TX/RX, up to 8 channels)
 - Pulse counter (capable of full quadrature decoding)
 - Motor PWM
 - LED PWM (up to 16 channels)
 - Hall effect sensor
 - Ultra low power analog pre-amplifier
- Security:
 - IEEE 802.11 standard security features all supported, including WPA, WPA2, WPA3 (depending on version)^[5] and WLAN Authentication and Privacy Infrastructure (WAPI)
 - Secure boot
 - Flash encryption
 - 1024-bit OTP, up to 768-bit for customers
 - Cryptographic hardware acceleration: AES, SHA-2, RSA, elliptic curve cryptography (ECC), random number generator (RNG)
- Power management:
 - Internal low-dropout regulator
 - Individual power domain for RTC
 - 5 µA deep sleep current
 - Wake up from GPIO interrupt, timer, ADC measurements, capacitive touch sensor interrupt




HiLetgo SHT31-D Temperature and Humidity Sensor Interface 3.3V GY-SHT31-D for Arduino
★★★★★ ~ 20
\$8.49
prime FREE Delivery Sat, Nov 19

Amazon's Choice




Sponsored ⓘ

Teyleten Robot ESP32S ESP32 ESP-WROOM-32 I Microcontroller for Arduino (ESP32 30P, 3PCS)
★★★★★ ~ 327
\$17.88 ~~\$19.88~~
prime FREE Delivery Sat, Nov 19



Sponsored ⓘ

KeeYees ESP32S ESP32 Development Board 2.4 GHz Chip for Arduino (38PIN Narrow Version, 2PCS)
★★★★★ ~ 621
\$13.99 ~~\$16.99~~
prime FREE Delivery Sat, Nov 19



3 Pieces ESP WROOM 32 ESP32 Development Board Noise Amplifiers Filters
★★★★★ ~ 74
\$18.99
prime FREE Delivery Sat, Nov 19

Adafruit IO = Cloud service popular among IoT developer

Get Started

FREE
forever

- 30 data points per minute
- 30 days of data storage
- Actions every 15 minutes
- 5 dashboard limit
- 2 WipperSnapper device limit
- 5 group limit
- 10 feed limit

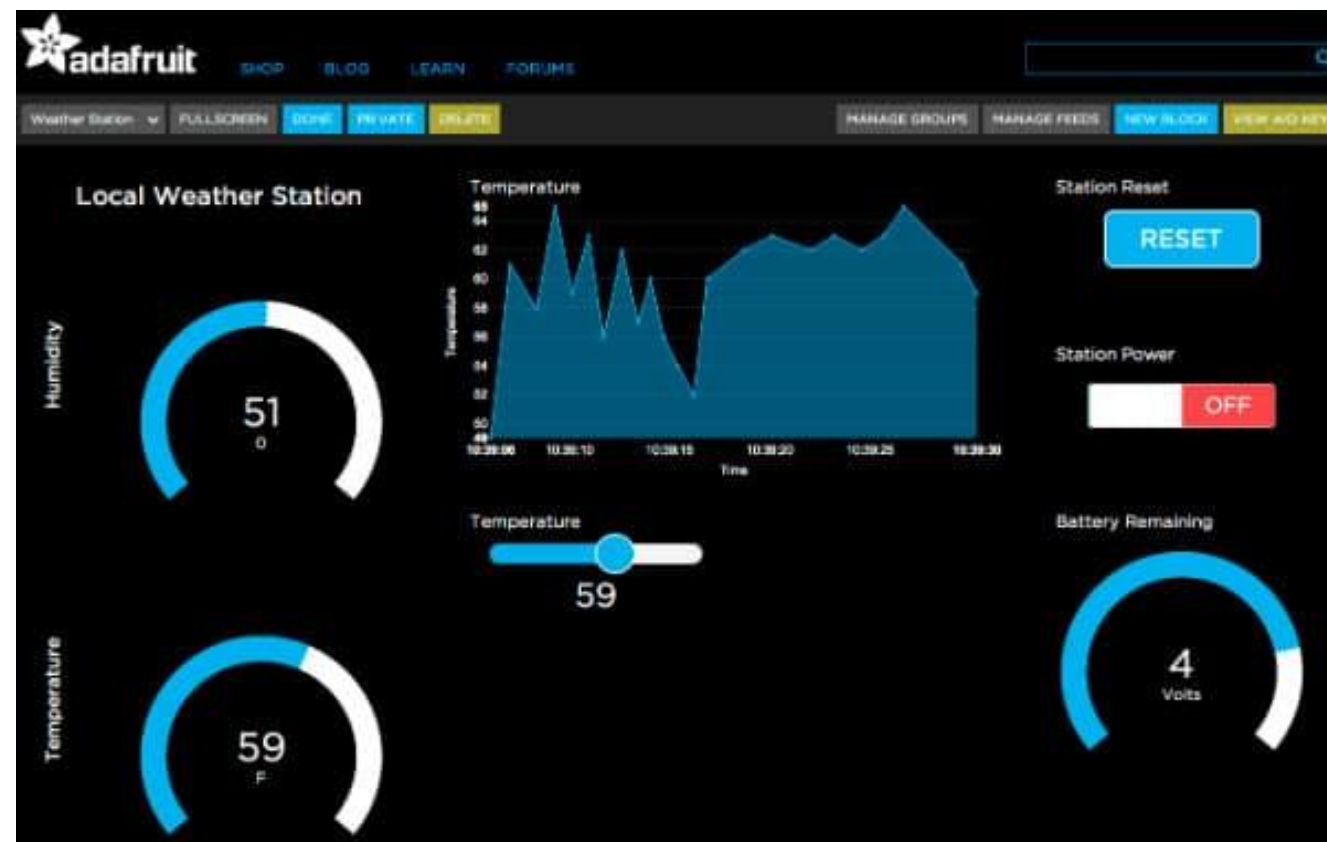
Sign Up Now

Power Up

\$10 or **\$99**
per month or per year

- 60 data points per minute
- 60 days of data storage
- Actions every 5 seconds
- 25 sms messages per day (UTC) for actions
- Unlimited dashboards
- Unlimited WipperSnapper devices
- Unlimited groups
- Unlimited feeds
- Learn more about IO+

Sign Up Now



Machine-Learning Based Anomaly Detection

Anomaly Detection on Internet of Things (IoT)

- Support Vector Machine (SVM) for classification: Work by constructing hyperplane that provide best separation from 2 clusters or data (normal and outlier)
- In this IoT project, synthetic anomaly data to simulate outlier are created along with the normal data from the sensors

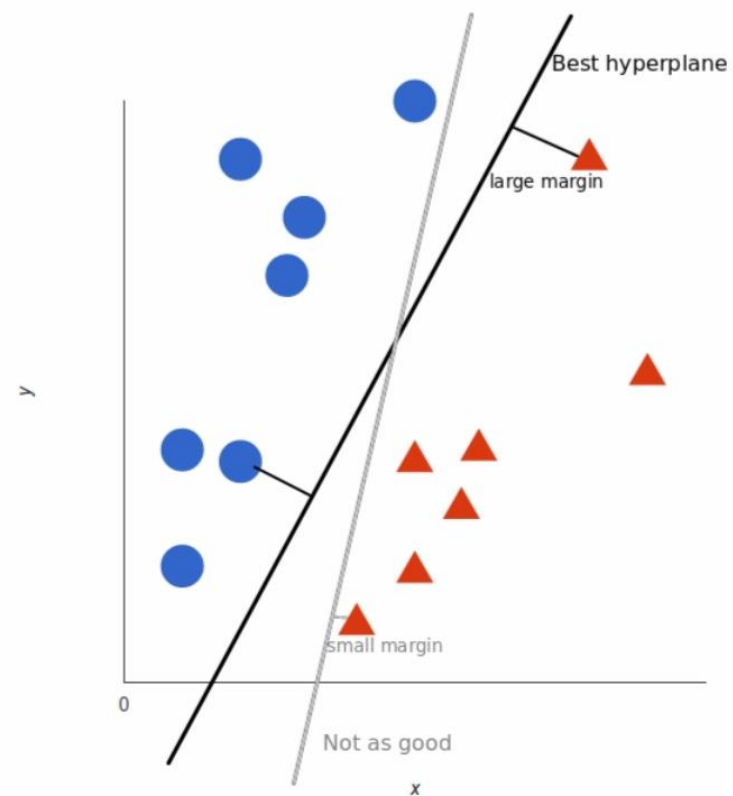


Table 1 Data-set after addition of synthetic data

Temperature	Humidity	Light	Voltage	Controller Id	Output
31.7	47	227.92	2.52	3	1
31.7	47	227.92	2.52	1	1
31.7	47	235.78	2.53	1	1
40.1	108.8	424.4	3.4	Nan	0
40.4	108.9	424.8	3.8	Nan	0

Machine-Learning Based Anomaly Detection

Anomaly Detection on Internet of Things (IoT)

- The machine-learning performance is evaluated using ROC curve

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

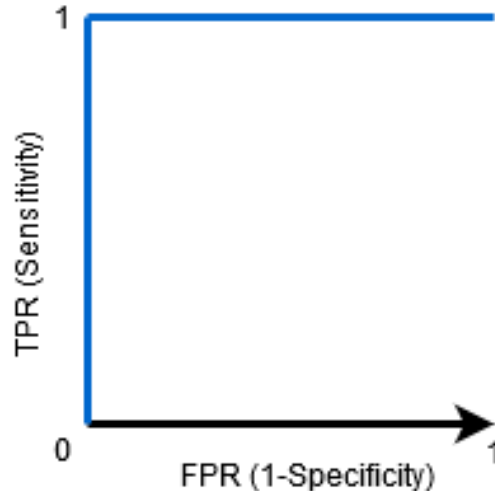
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

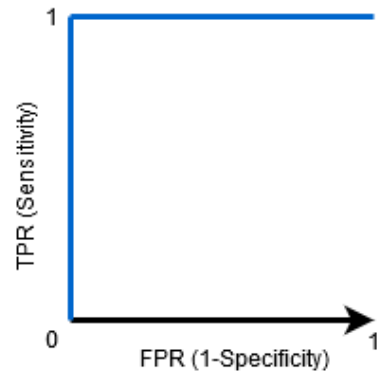
- This is the ideal ROC curve which means there is a decision threshold who will provide 100% of True Positive Rate and 0% of False Positive Rate



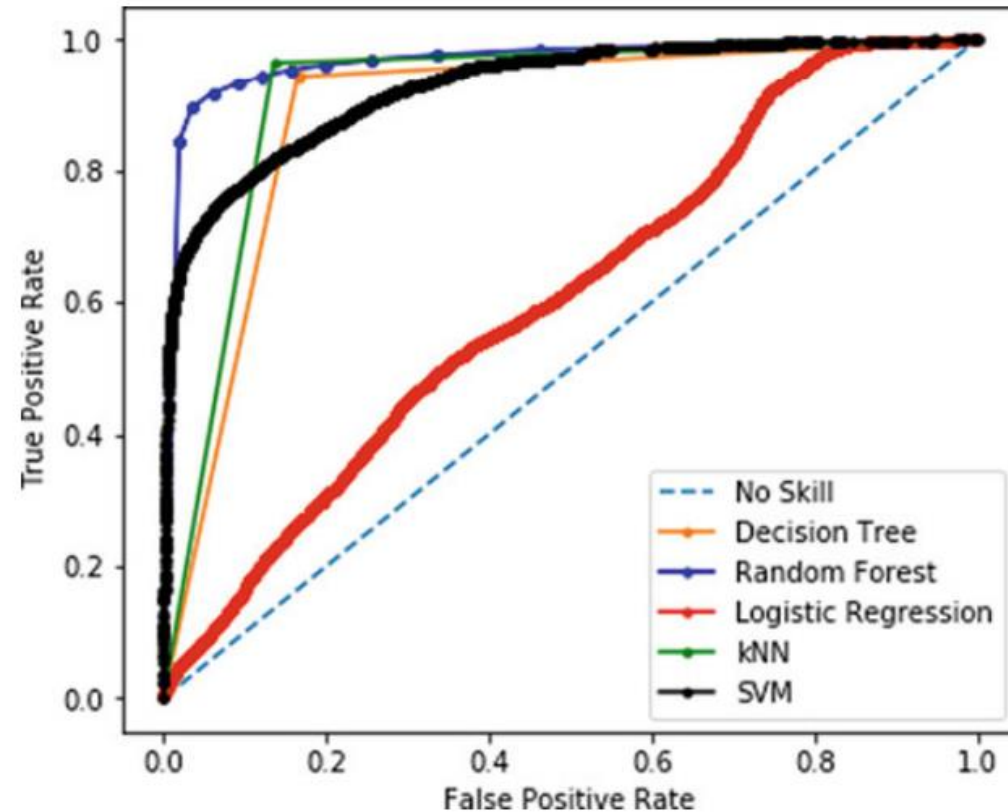
Machine-Learning Based Anomaly Detection

Anomaly Detection on Internet of Things (IoT)

- In reality, there is a compromise in choosing the decision threshold. The higher TPR that we want the higher FPR that we get
- A metric to evaluate the machine learning performance is AUC (Area under the ROC curve). In an ideal world the $AUC = 1$



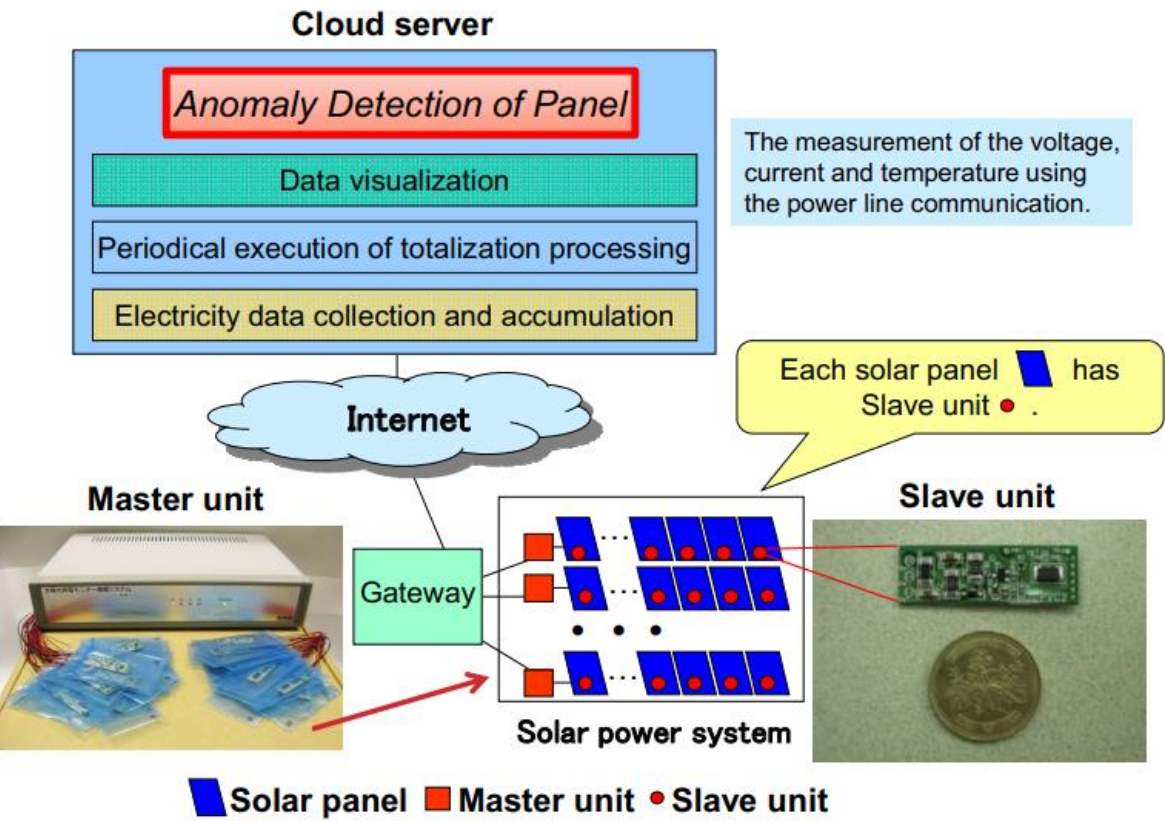
$AUC = 1$



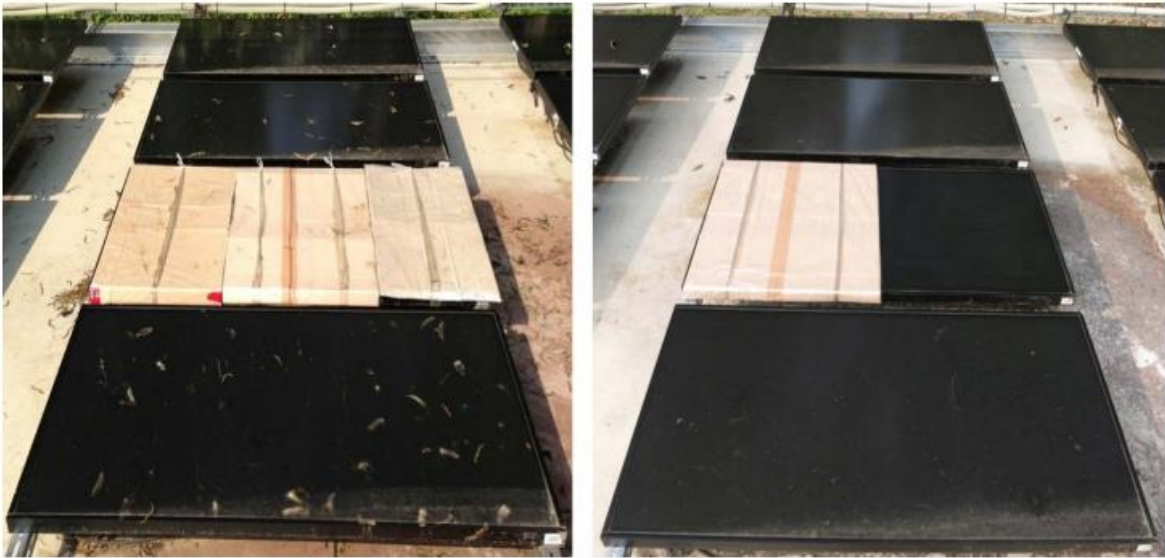
ROC curves for this IoT project from several machine-learning algorithm. So far random forest is better than SVM

Machine-Learning Based Anomaly Detection

Renewable energy: Solar panel



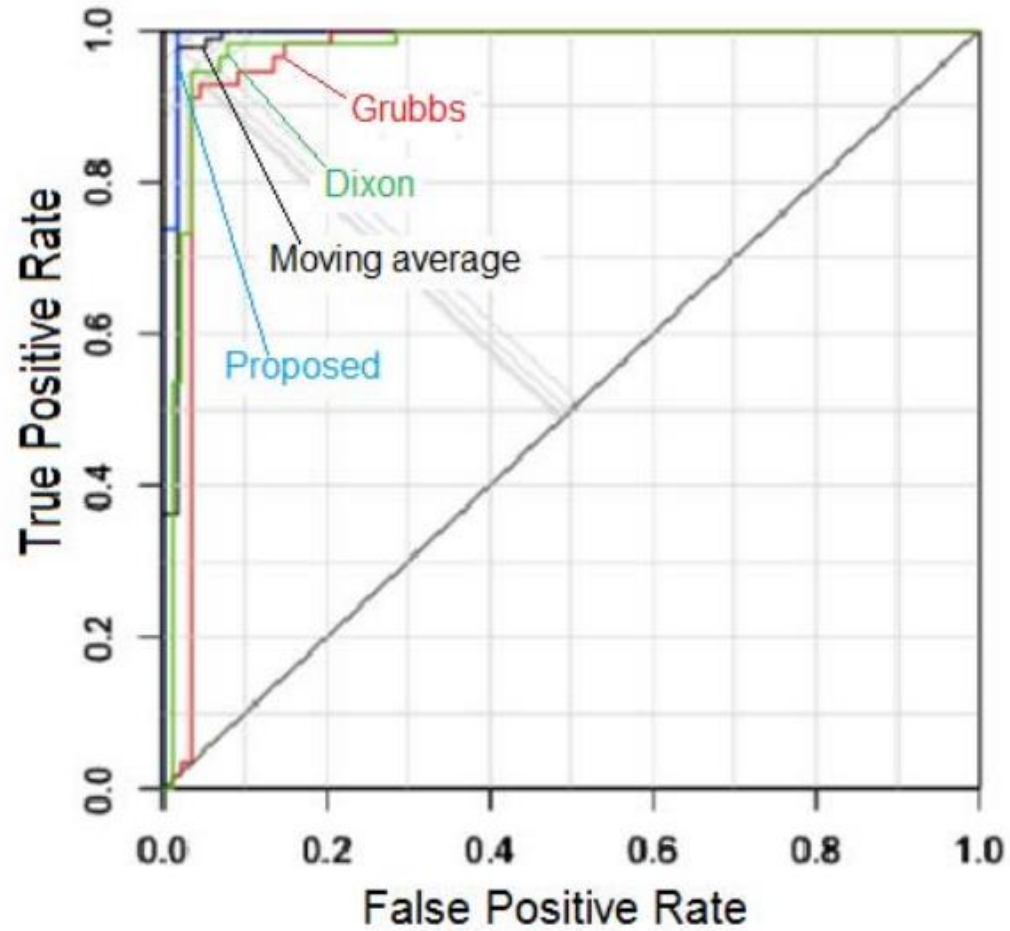
Setup



Anomaly data for training are generated by covering the solar panel partially

Machine-Learning Based Anomaly Detection

Renewable energy: Solar panel

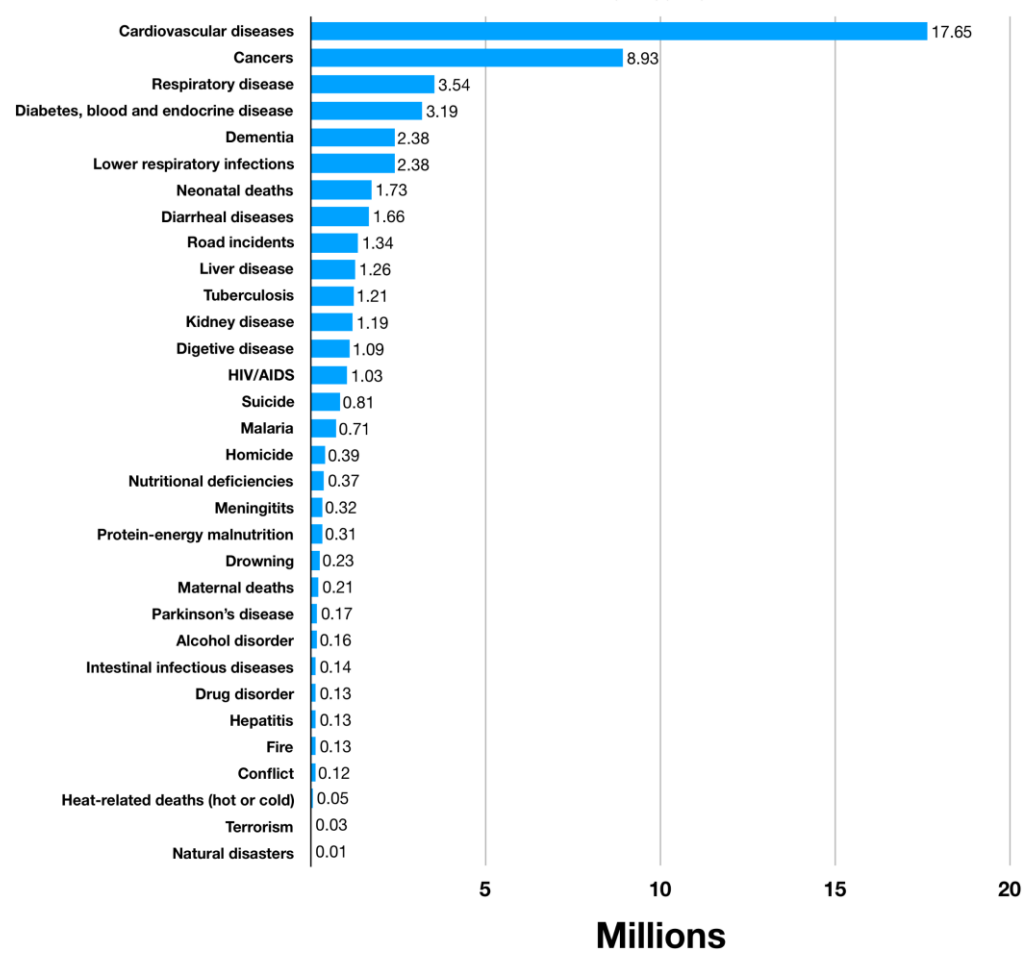


ROC curve from several Machine-learning algorithm

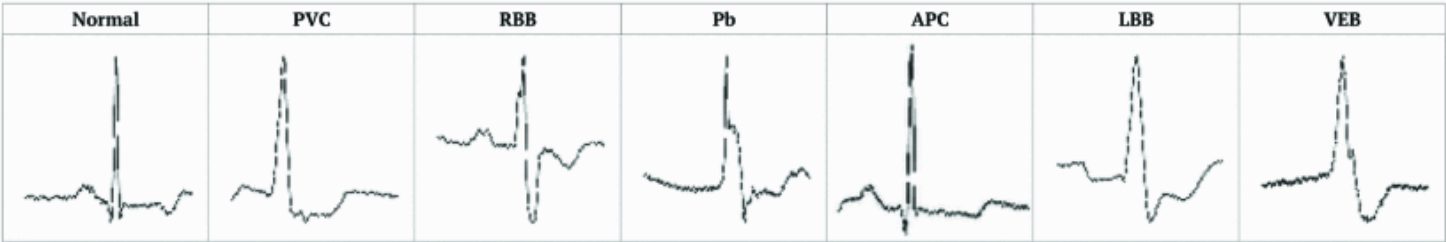
Machine-Learning Based Anomaly Detection

Healthcare: Heartbeat anomaly detection

Annual Number of Deaths by Cause



Several type of heartbeat



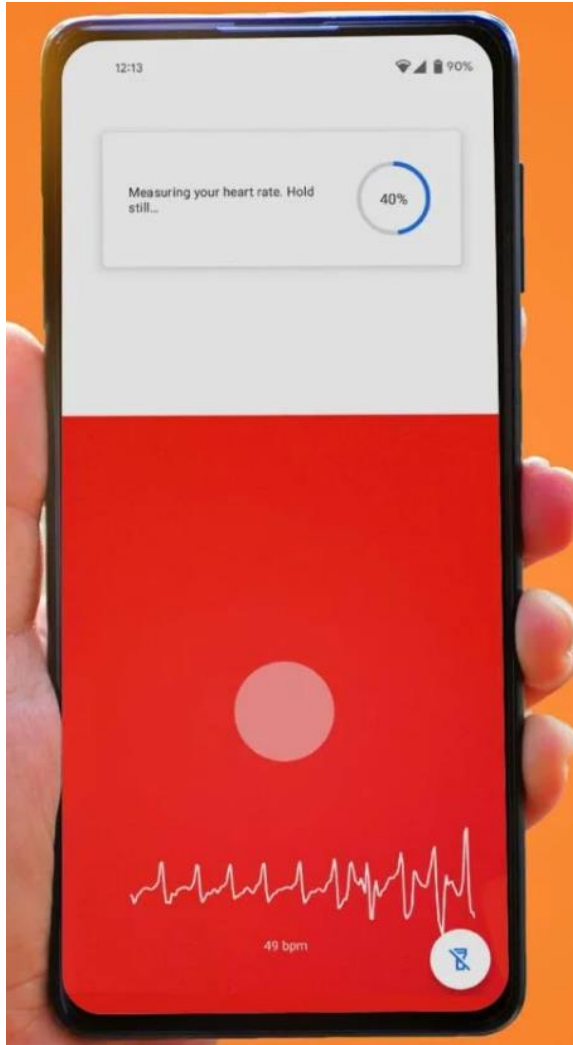
Heartbeat Type

Atrial Premature Contraction	APC
Normal	Normal
Left Bundle Branch Block	LBBS
Paced Beat	PAB
Premature Ventricular Contraction	PVC
Right Bundle Branch	RBB
Ventricular Escape Beat	VEB

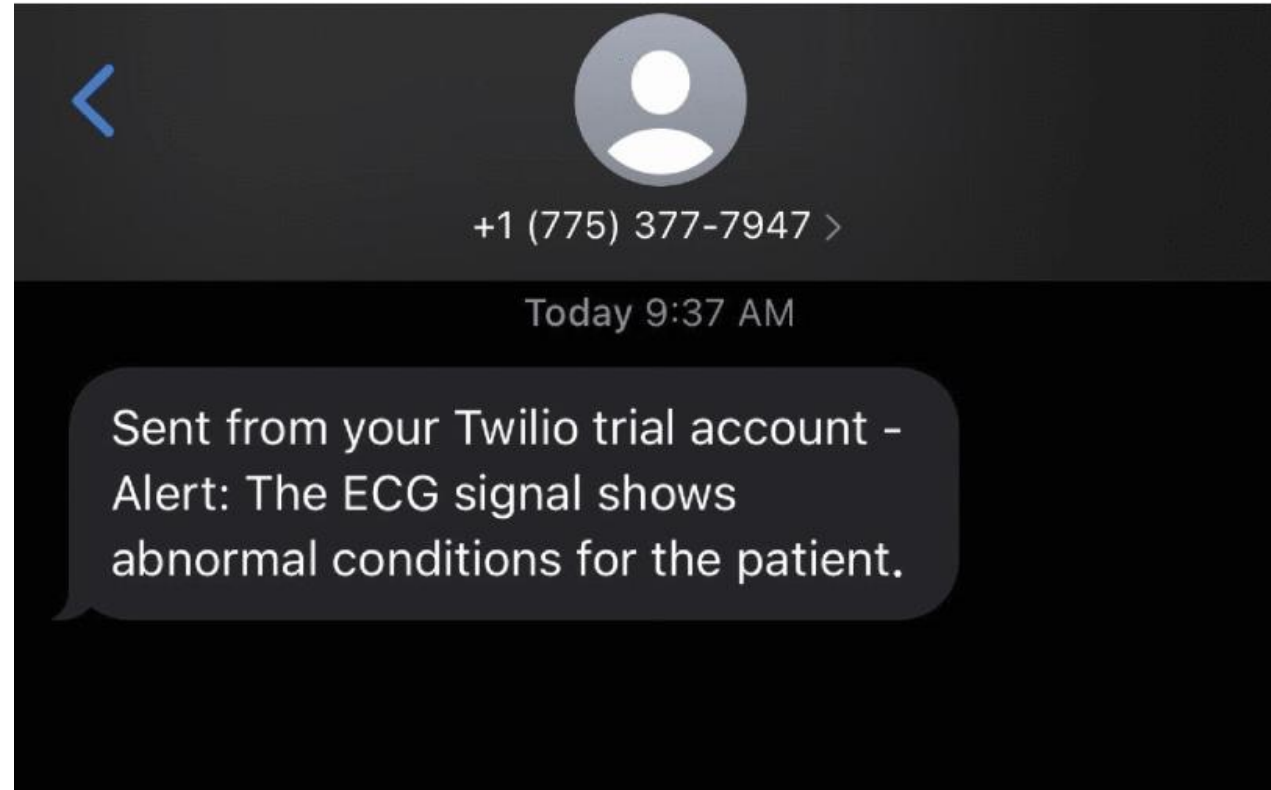
Cardiovascular diseases is the leading cause of death for both men and women

Machine-Learning Based Anomaly Detection

Healthcare: Heartbeat anomaly detection



Smartphone reading

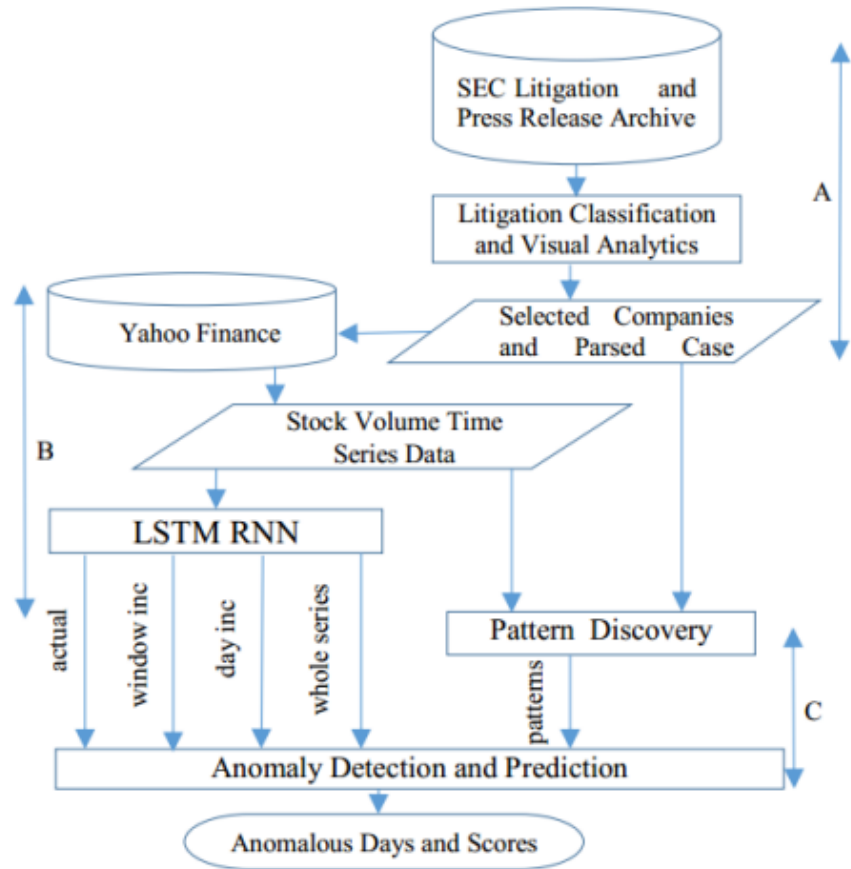


Early-warning system

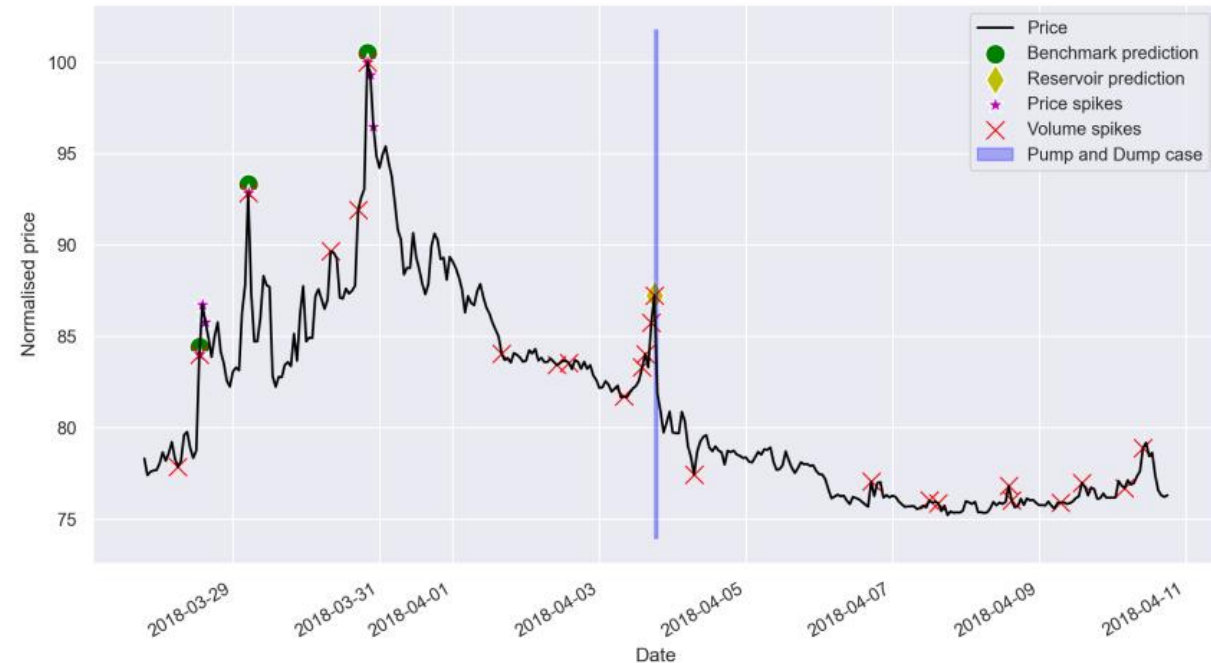
Machine-Learning Based Anomaly Detection

Financial Market: Insider trading detection and “pump & dump” fraud detection

Insider trading detection



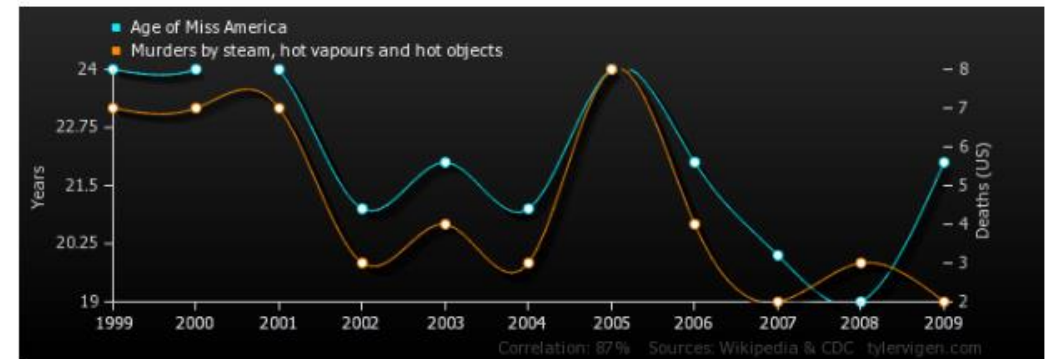
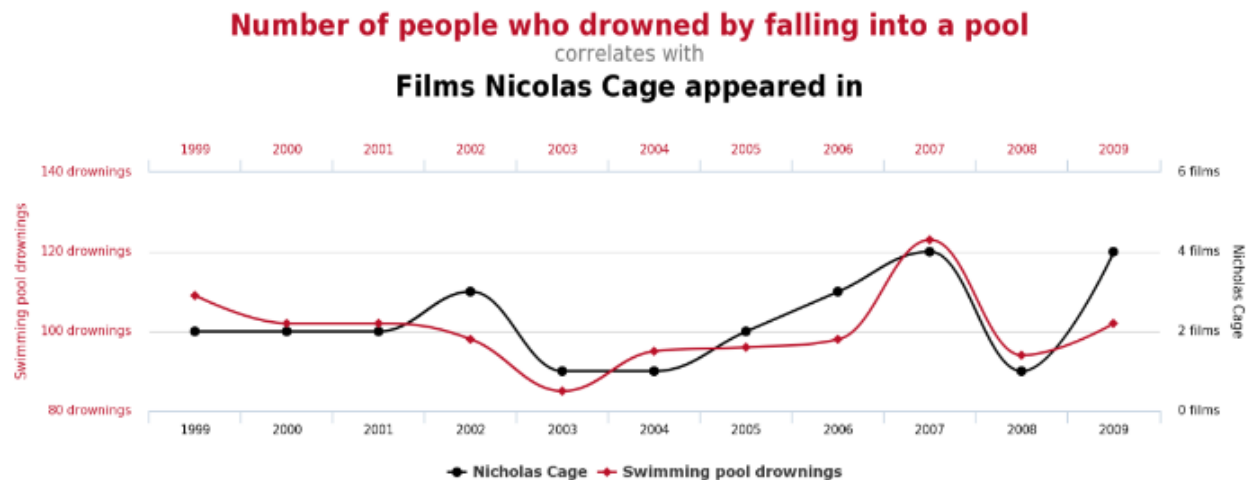
Pump & Dump in Crypto market



- **Insider trading:** the illegal practice of trading on the stock exchange to one's own advantage through having access to confidential information.
- **Pump and dump:** a scheme that attempts to boost the price of a stock through recommendations based on false, misleading, or greatly exaggerated statements

Human-Machine Cooperation

Why is it Important?



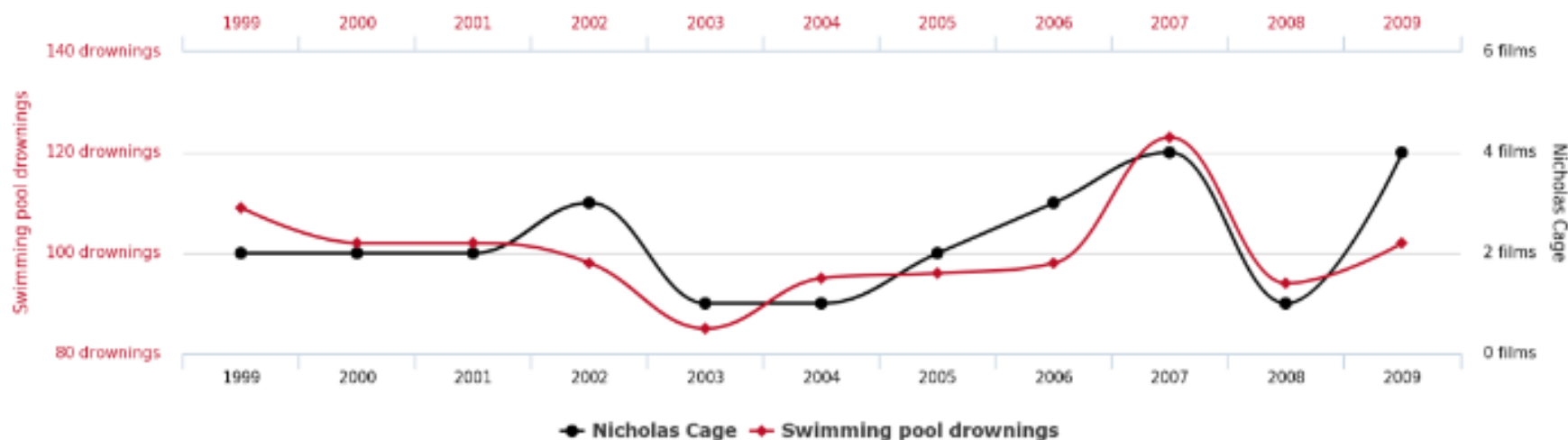
Human-Machine Cooperation

- Human interpretation is crucial since machines are sometimes “too smart” and capture non-senses correlation

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

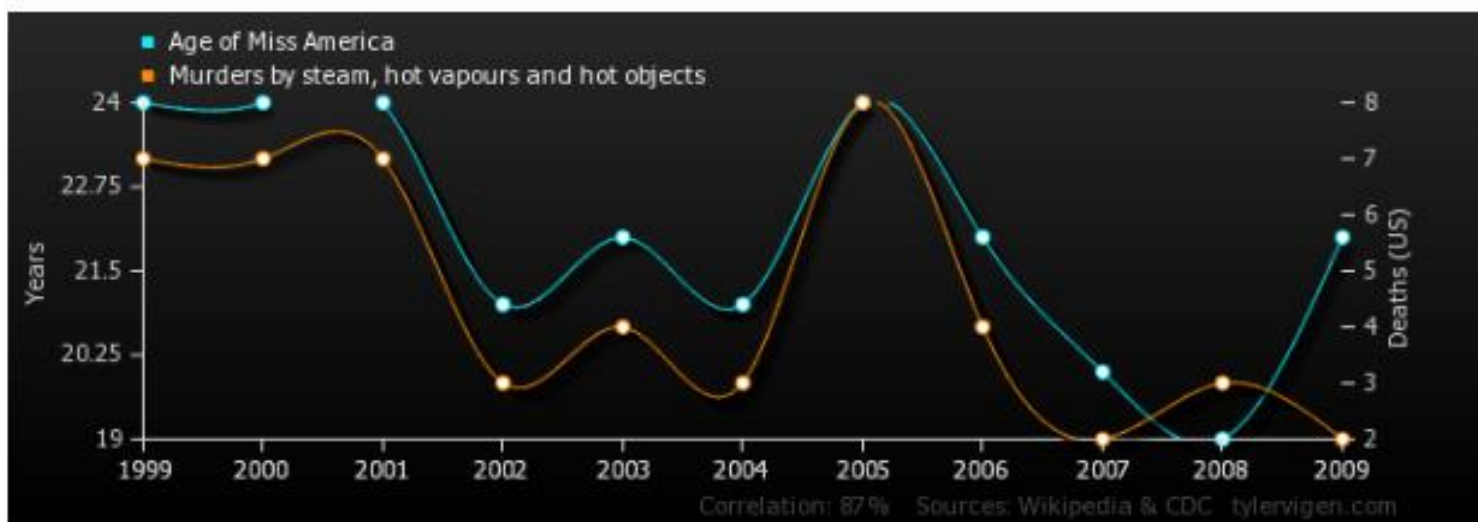
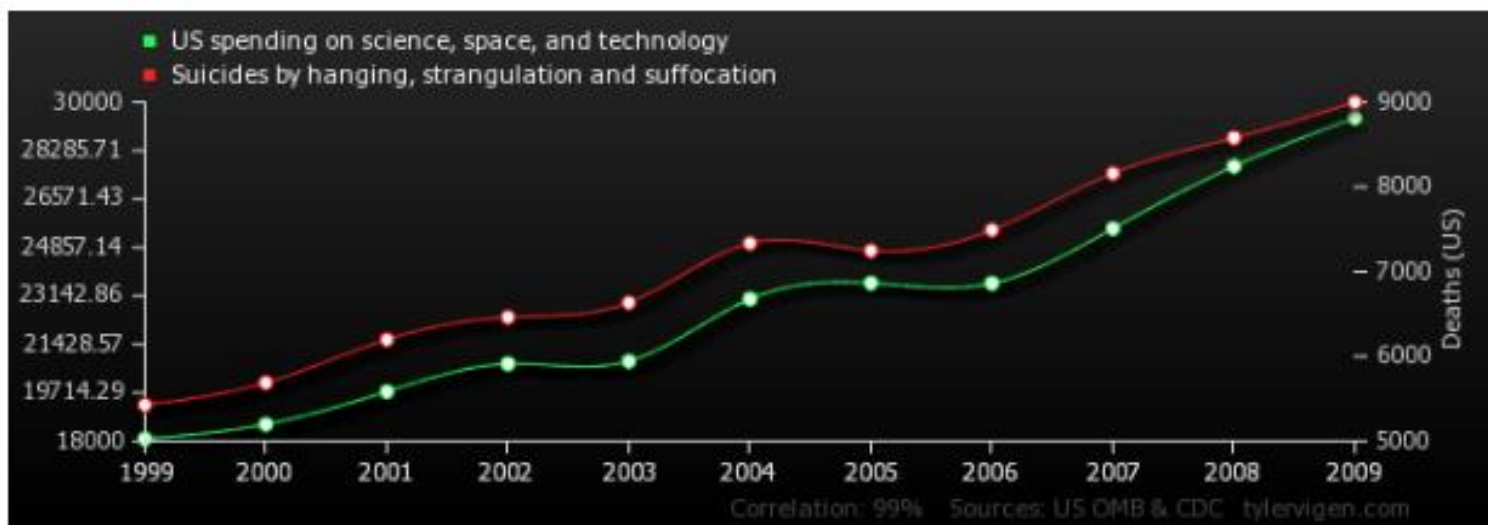


tylervigen.com



Human-Machine Cooperation

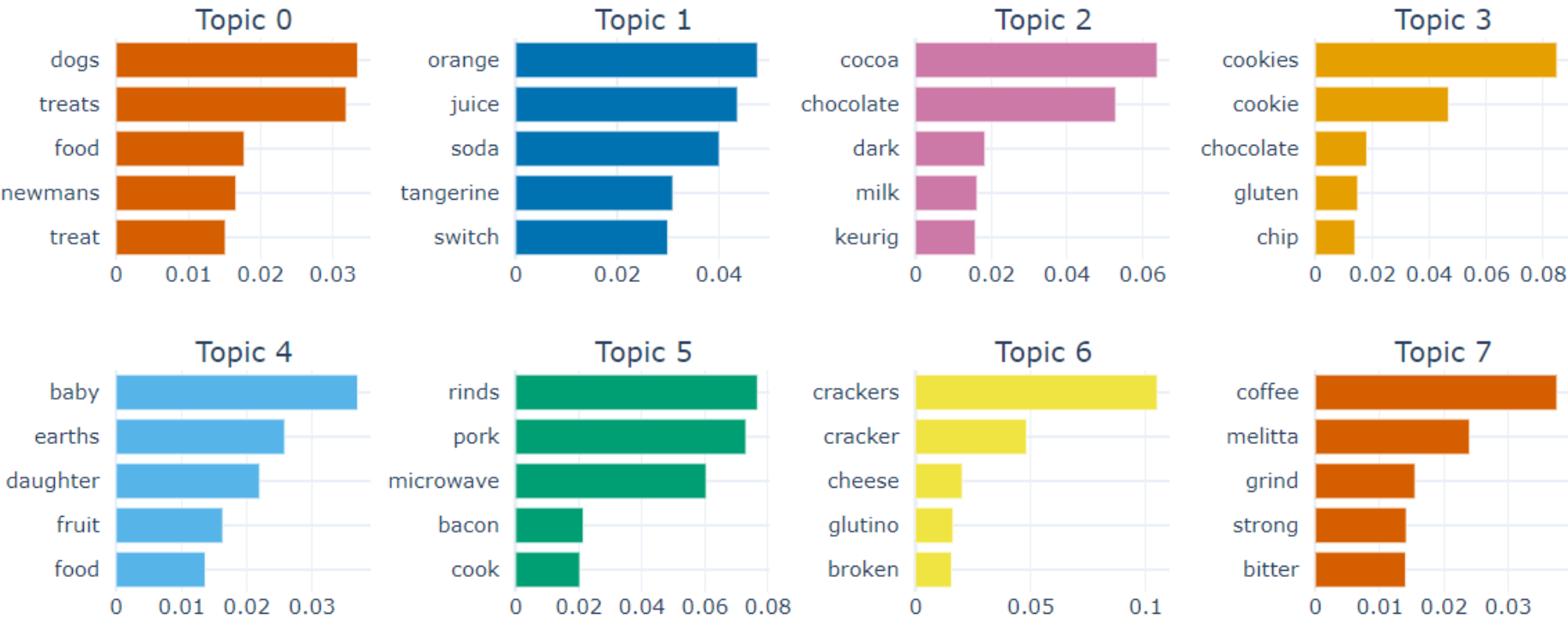
- Human interpretation is crucial since machines are sometimes “too smart” and capture non-senses correlation



Human-Machine Cooperation

- Data pre-processing is extremely important. Sometimes, preparing and cleaning the data take more time than machine-learning processing.
- Natural-Language Processing (NLP) required human interpretation

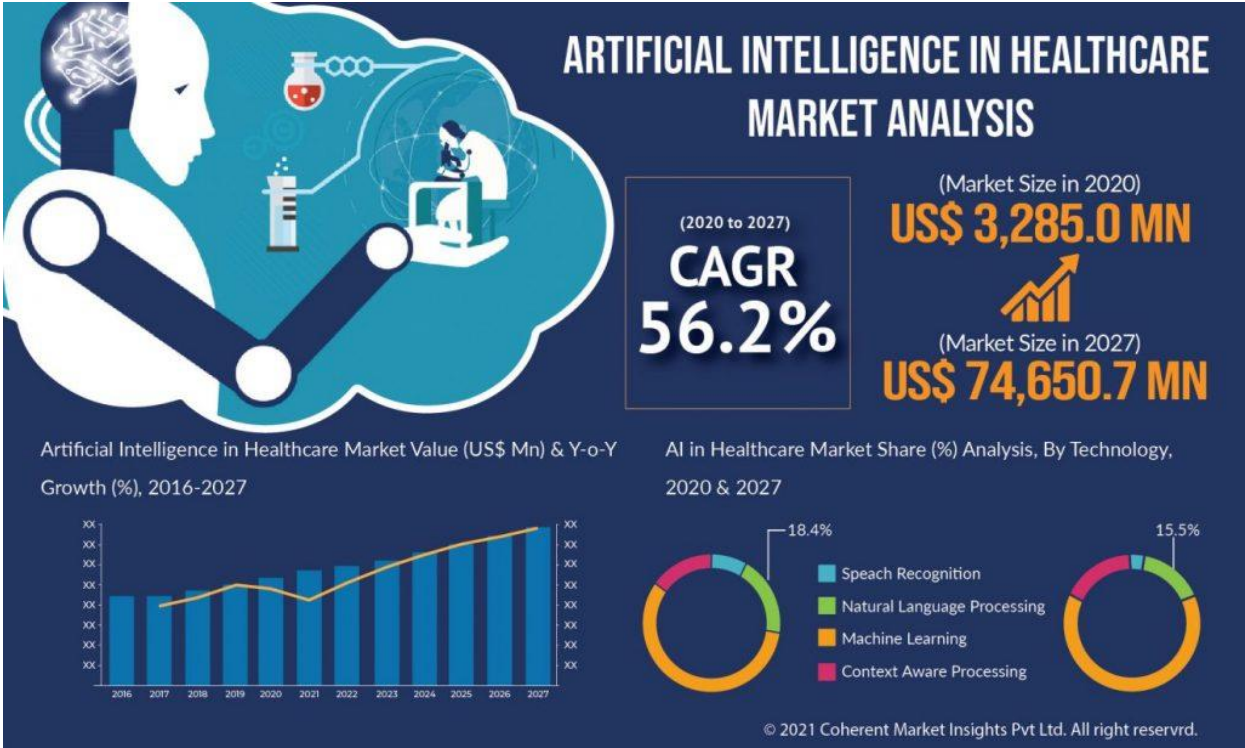
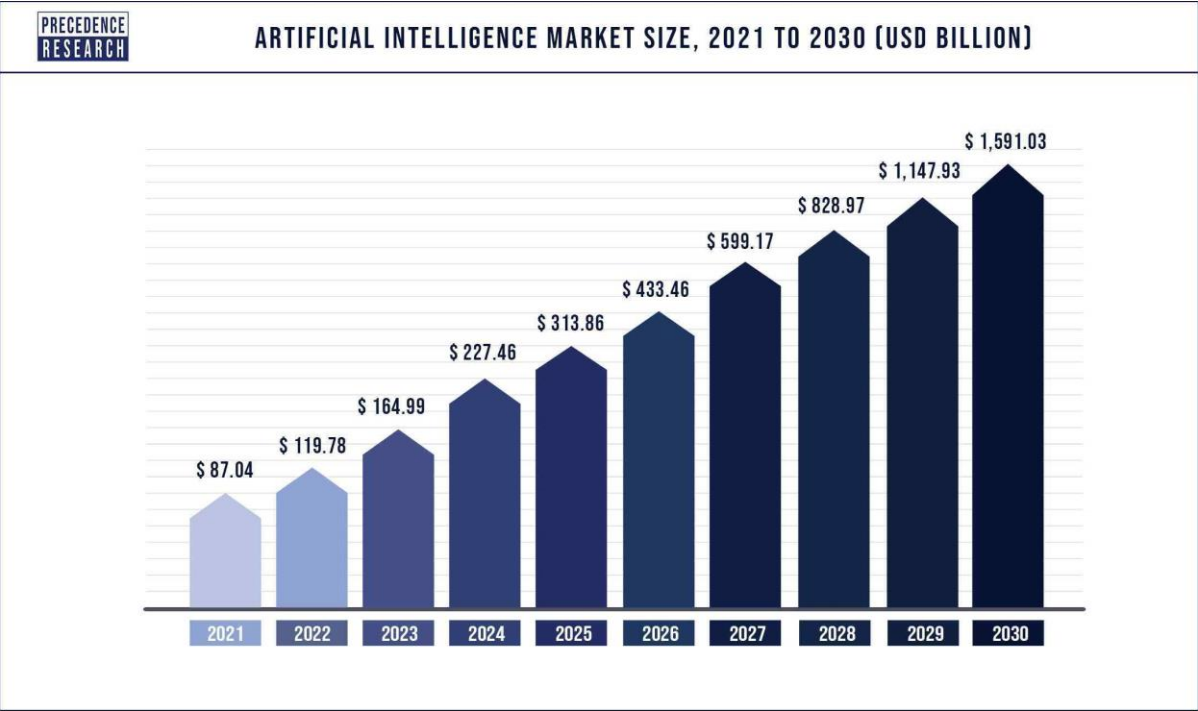
Topic Word Scores



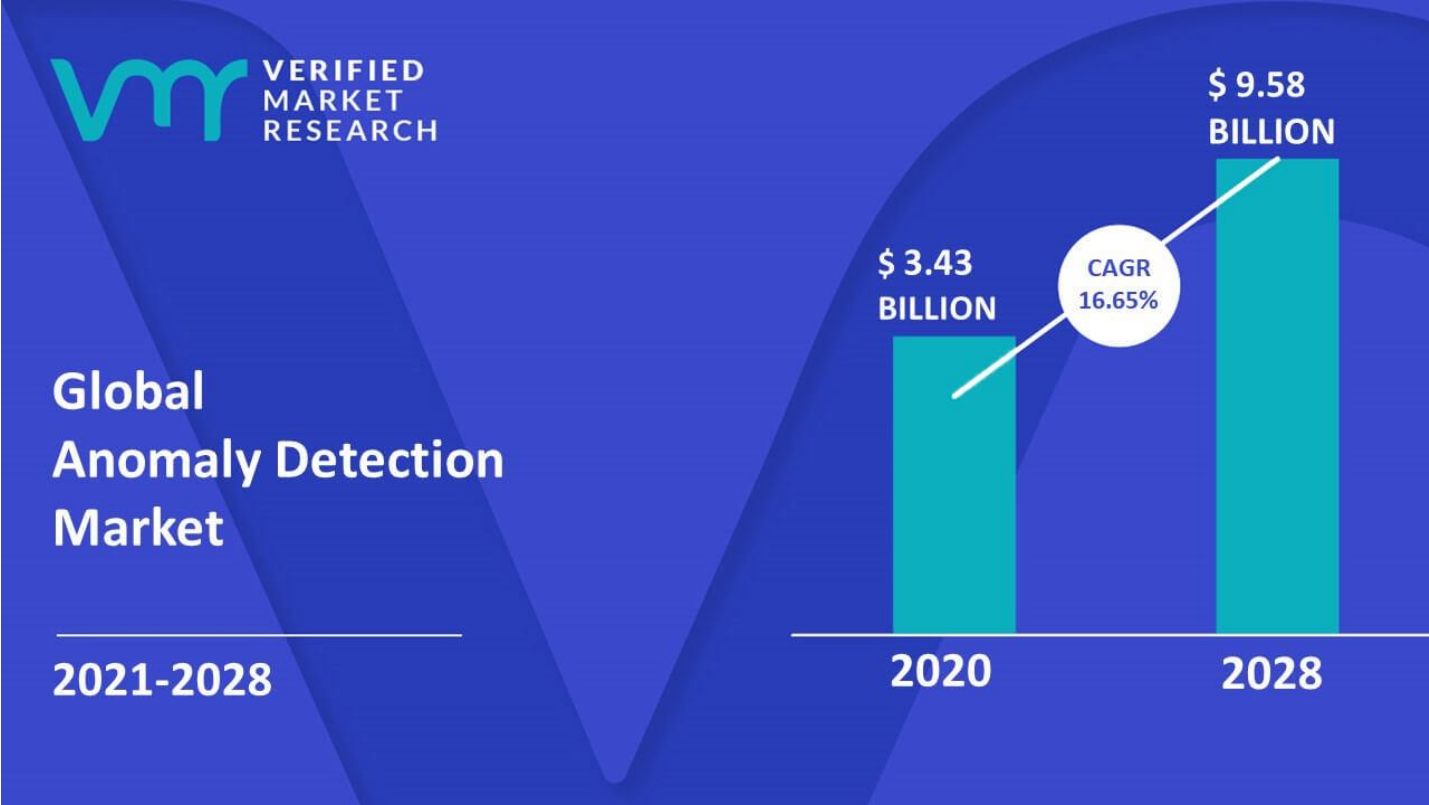
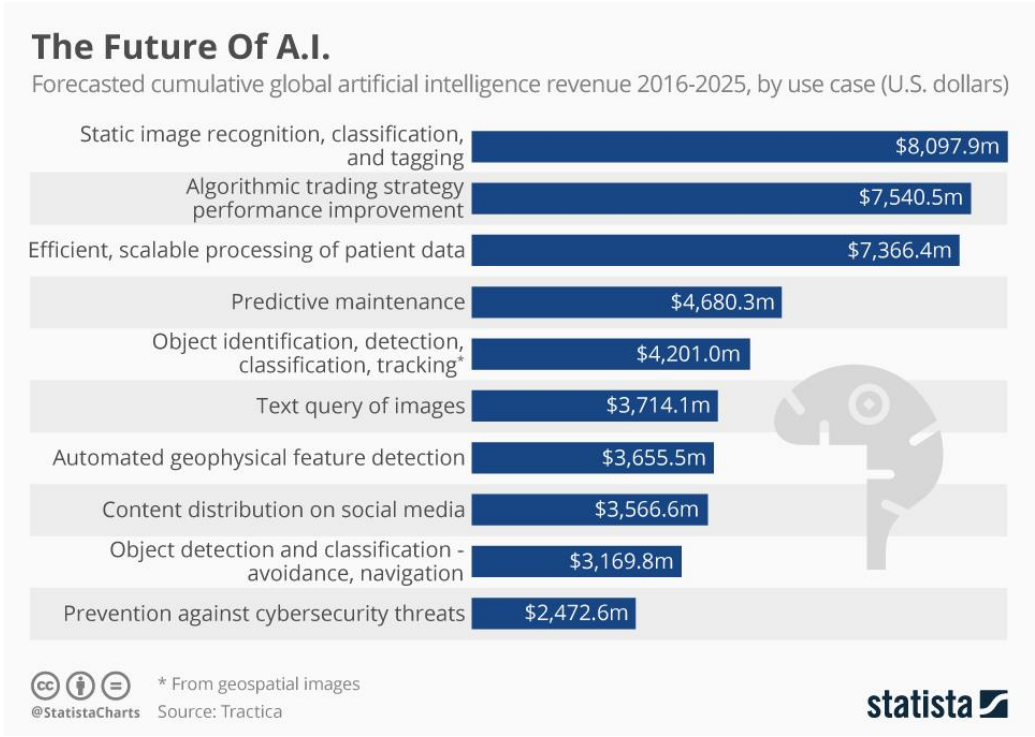
Topic modelling from Amazon review data

Summary & Outlook: The bright future of AI

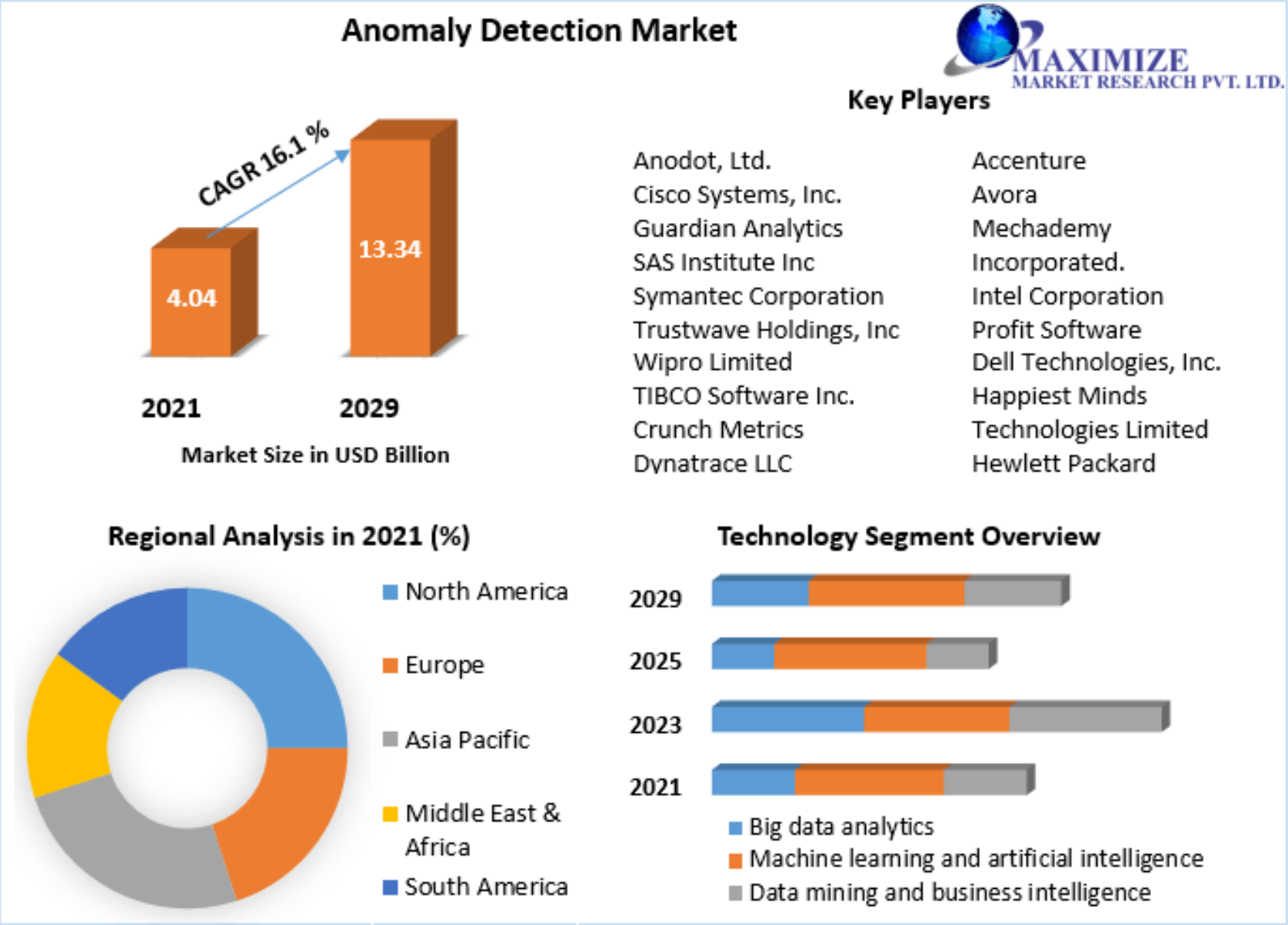
- Anomaly detection has triggered tremendous development in broad range of fields: science, finance, manufacture, healthcare, industry, etc.
- Industry Outlook:



Summary & Outlook: The bright future of AI



Summary & Outlook: The bright future of AI

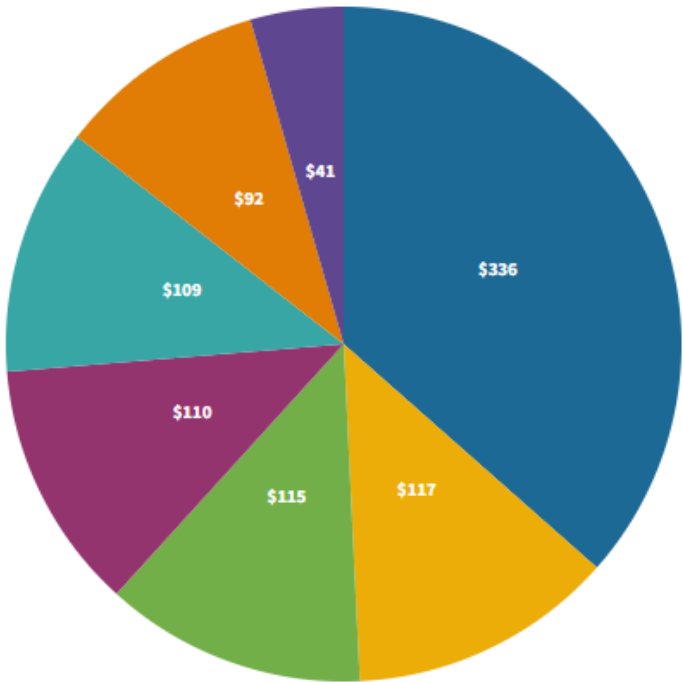


Summary & Outlook: The bright future of AI

AI is expected to uplift GDP across Southeast Asia in 2030
AI contribution to the region's GDP is expected to reach almost \$1 trillion

*number in billion US\$

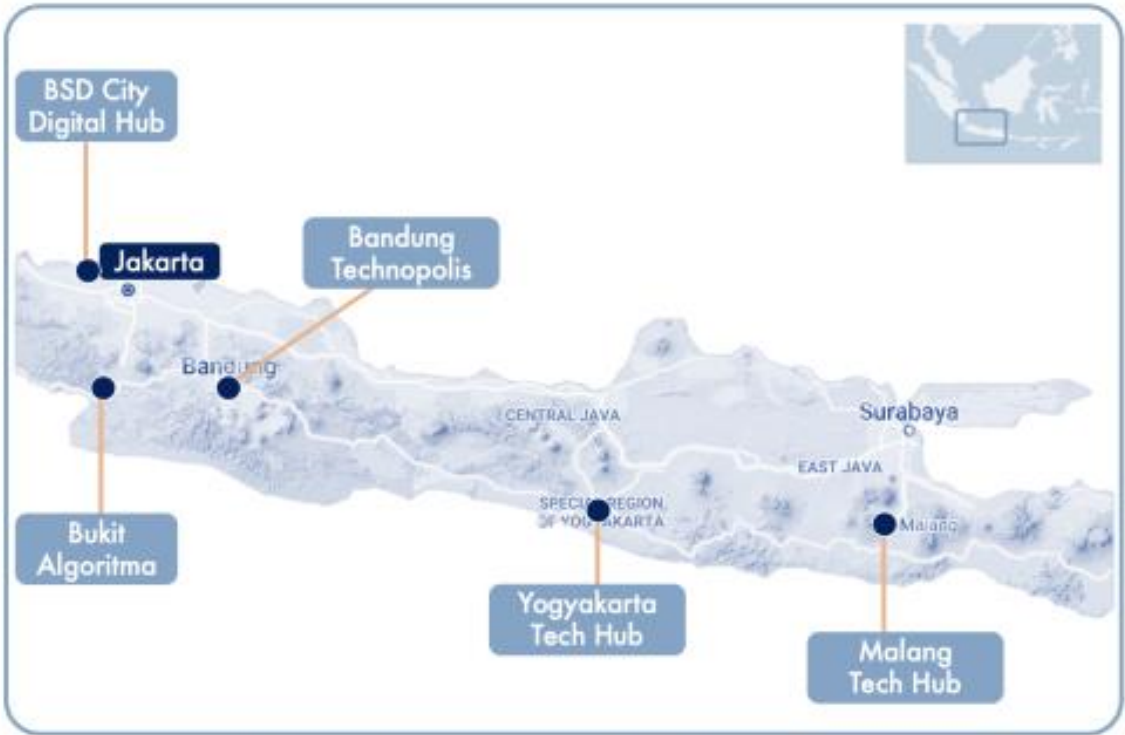
Indonesia Thailand Malaysia Singapore Vietnam The Philippines
Brunei, Cambodia, Laos, Myanmar



Source: Kearney • JP/Eisya Eloksari
AI total contribution to ASEAN GDP in 2030 is \$950 billion

TheJakartaPost

Figure 1. Indonesia's tech hubs are concentrated on Java and located near the capital of Jakarta



Source: Tech in Asia, Google Maps; edited by CSET.

Indonesia's AI Promise in Perspective

CSET Issue Brief

Indonesia's AI Potential

Over the last five years, the growth of Indonesia's technology sector has received international attention, which earned the country the title as Asia's "sleeping giant" and "next AI startup hub."⁵ In August 2020, Indonesia published its first National Strategy on Artificial Intelligence 2020-2045 (Strategi Nasional Kecerdasan Artifisial), becoming the second Association of Southeast Asian Nations (ASEAN) country to do so after Singapore.⁶ Emphasizing education and research, health services,

In addition to being home to such a thriving AI startup environment, Indonesia boasts the highest rate of AI adoption in Southeast Asia. The 2018 International Data Corporation Survey reported that 24.6 percent of Indonesia's surveyed organizations integrated AI into their operations.³⁴ Compare this with its regional competitors, 17.1 percent for Thailand and 9.9 percent for Singapore, and it is understandable why Indonesia is viewed as the next frontier for AI.³⁵ As shown in Figure 2, the expected

Thank You

Any questions?