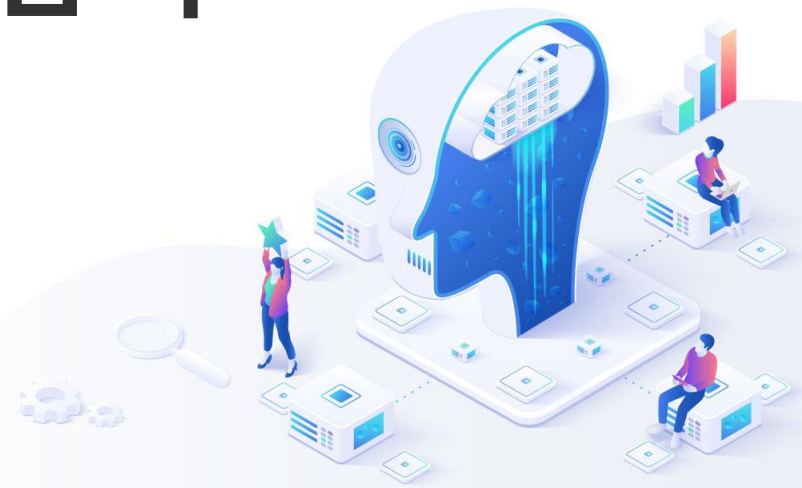
새싹(SeSAC) New 금천캠퍼스 2기 인공지능 과정

기업데이터를 활용한 AI 취업 부트캠프

AI기반 악성도메인 탐지

- 데이터 전처리 및 모델링 -



Seoul Software ACademy

글로벌 소프트웨어 캠퍼스

목차

도메인과 DGA

01

- 도메인 개념과 원리
- DGA 개념과 원리

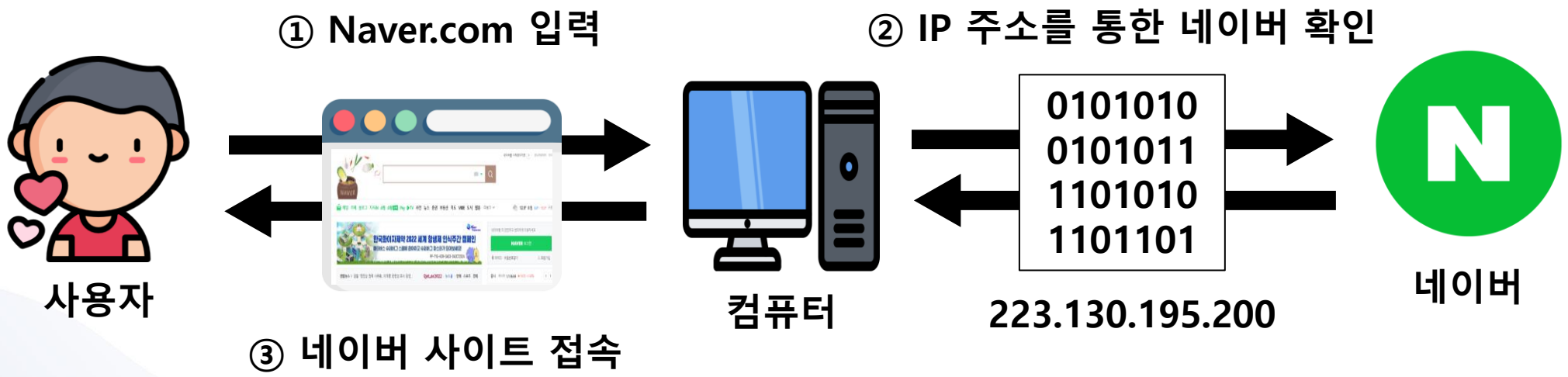
- 도메인(Domain)
  - 네트워크상에서 컴퓨터(서버)를 식별하는 호스트명/등록된 이름을 의미

https://www.naver.com	https://aws.amazon.com
https://naver.com	https://sesac.seoul.kr
https://www.google.com	https://www.daum.net
https://www.youtube.com	https://app.slack.com

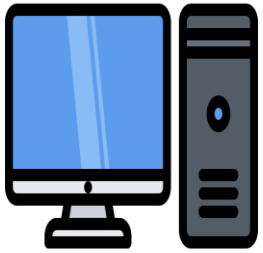
- Quiz) 도메인은 무엇일까요?
  - https://search.naver.com/search.naver?query=대한민국
  - https://search.naver.com/search.naver?query=대한민국

컴퓨터는 어떻게 도메인을 이해하고 연결해줄까?

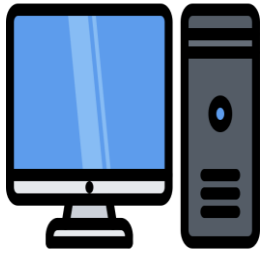
- 컴퓨터는 IP 주소를 통해 통신할 대상을 확인한다.
  - 우리는 네이버를 접속하기 위해 검색했지만 컴퓨터는 IP로 네이버에 접속함
  - IP주소는 컴퓨터간 통신을 위해 가지고 있는 인터넷 주소를 의미함



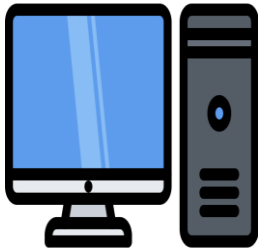
- 각 컴퓨터(서버)는 각자를 구분하기 위해 **IP가 중복되지 않게 할당되어짐**
  - 시작버튼 클릭 > 명령 프롬프트(CMD) 실행 > ipconfig 입력



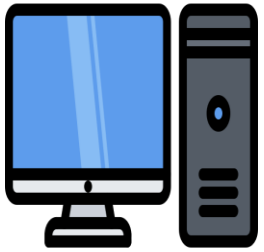
A 컴퓨터
192.168.35.43



B 컴퓨터
192.168.35.68

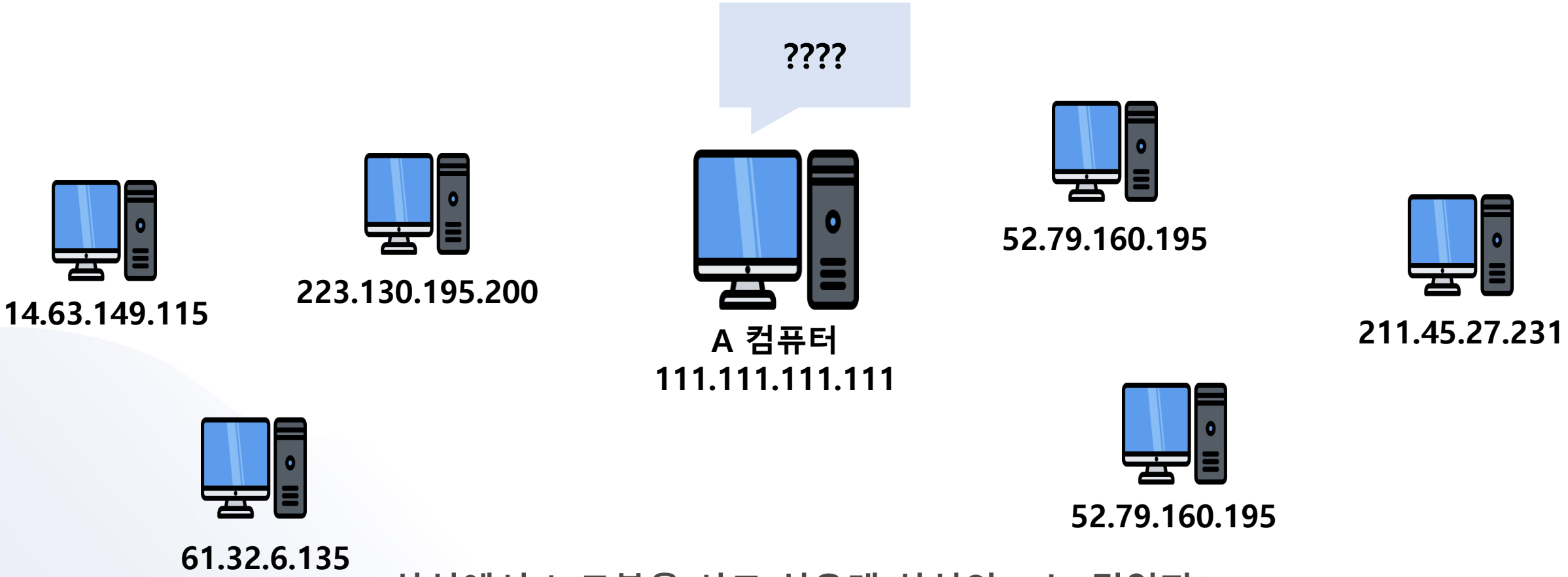


C 컴퓨터
192.168.35.75



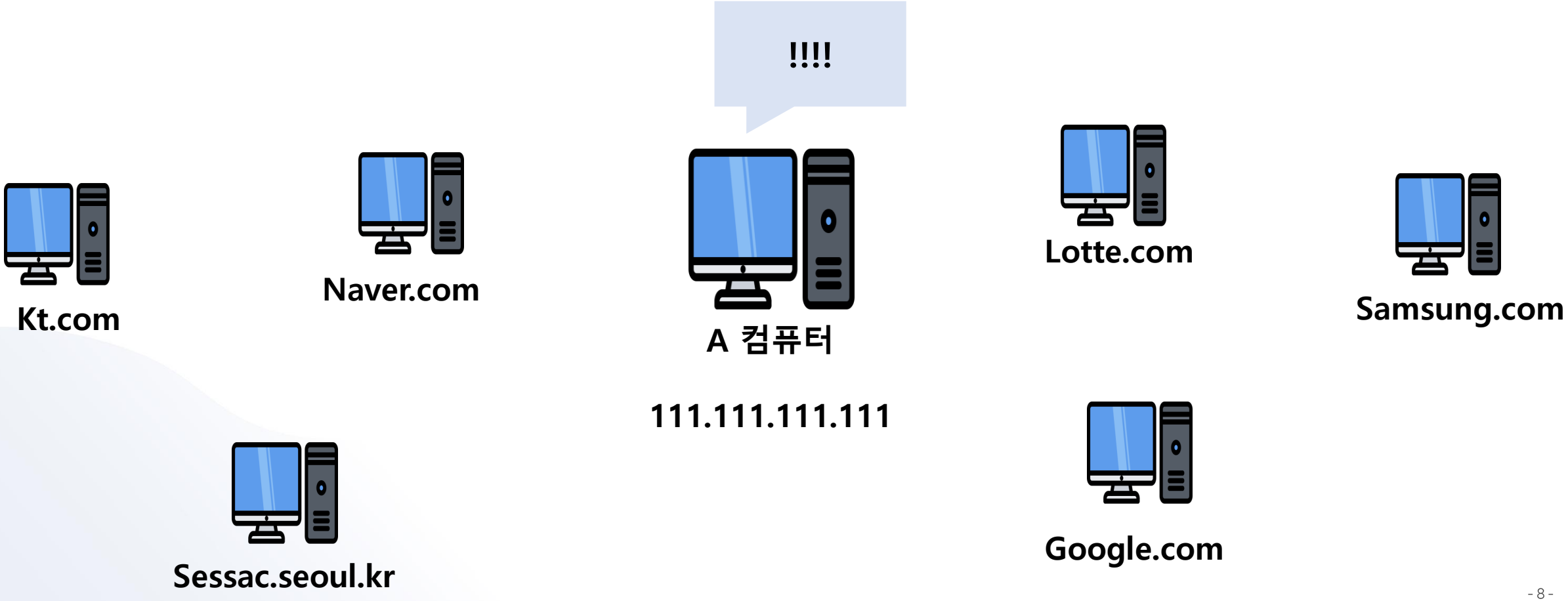
D 컴퓨터
192.168.35.92

- 인터넷이 확대됨에 따라 컴퓨터(서버)의 IP를 모두 기억하기 어려워짐



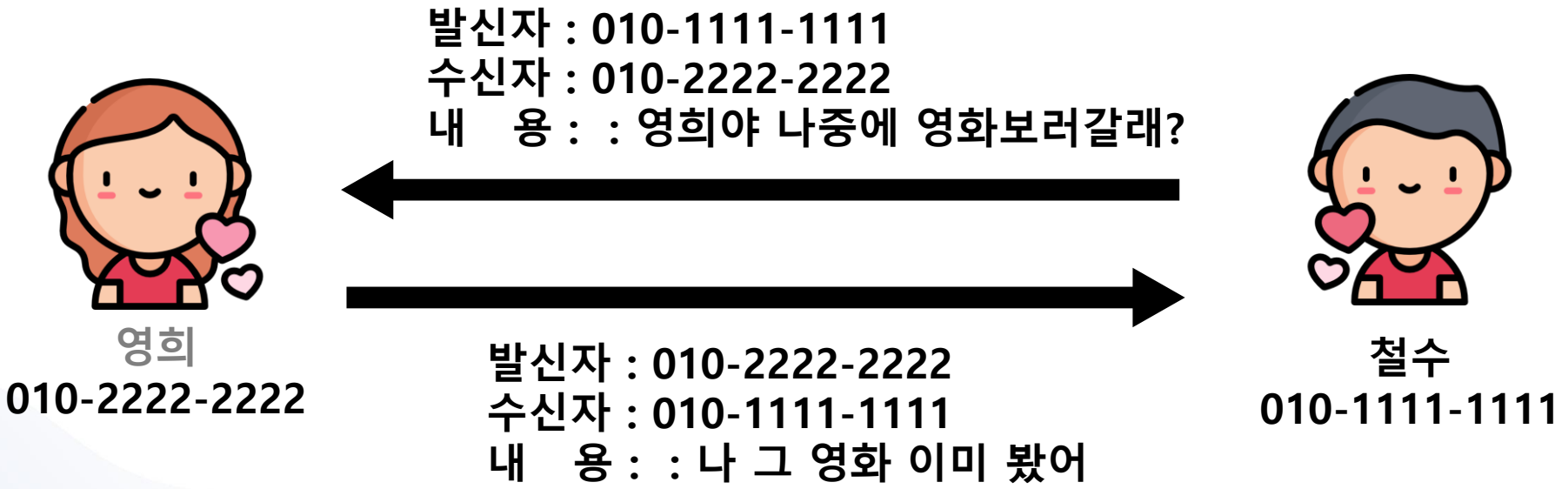
삼성에서 노트북을 사고 싶은데 삼성의 IP는 뭐였지?

- 이름을 붙이는 것으로 많은 서버를 기억하고 접속할 수 있게 되었음
  - 기업 도메인은 발음하기 쉽거나 외우기 쉬운 형태로 등록



- 예시로 보는 통신 과정①

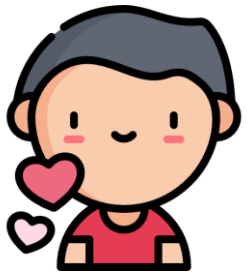
  Ex) 연락할 사람이 적을 때



- 만약에 아는 사람이 많다면 모든 번호를 외우기 어렵다

010-2222-2222

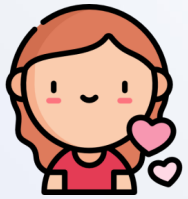
010-4444-4444



철수
010-1111-1111

010-8888-8888

010-9999-99999



010-3333-3333

010-5555-5555

010-6666-6666

010-7777-7777

010-1010-1010

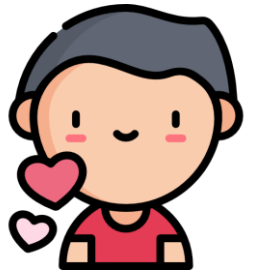
▪ 예시로 보는 통신 과정 ②

Ex) 아는 사람이 많아져서 번호를 제대로 기억하지 못한 경우

준호
010-3333-3333

발신자 : 010-1111-1111
수신자 : 010-3333-3333
내　용 : : 영희야 나중에 영화볼래?

발신자 : 010-3333-3333
수신자 : 010-1111-1111
내　용 : : 철수야 실망이야;

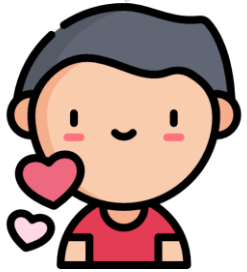


철수
010-1111-1111

- 휴대전화번호를 연락처에 저장해서 원하는 대상에게 쉽게 연락할 수 있음

영희
010-2222-2222

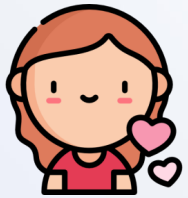
희주
010-4444-4444



철수
010-1111-1111

단비
010-8888-8888

유림
010-9999-99999



현서
010-3333-3333

예람
010-5555-5555

수아
010-6666-6666

다정
010-7777-7777

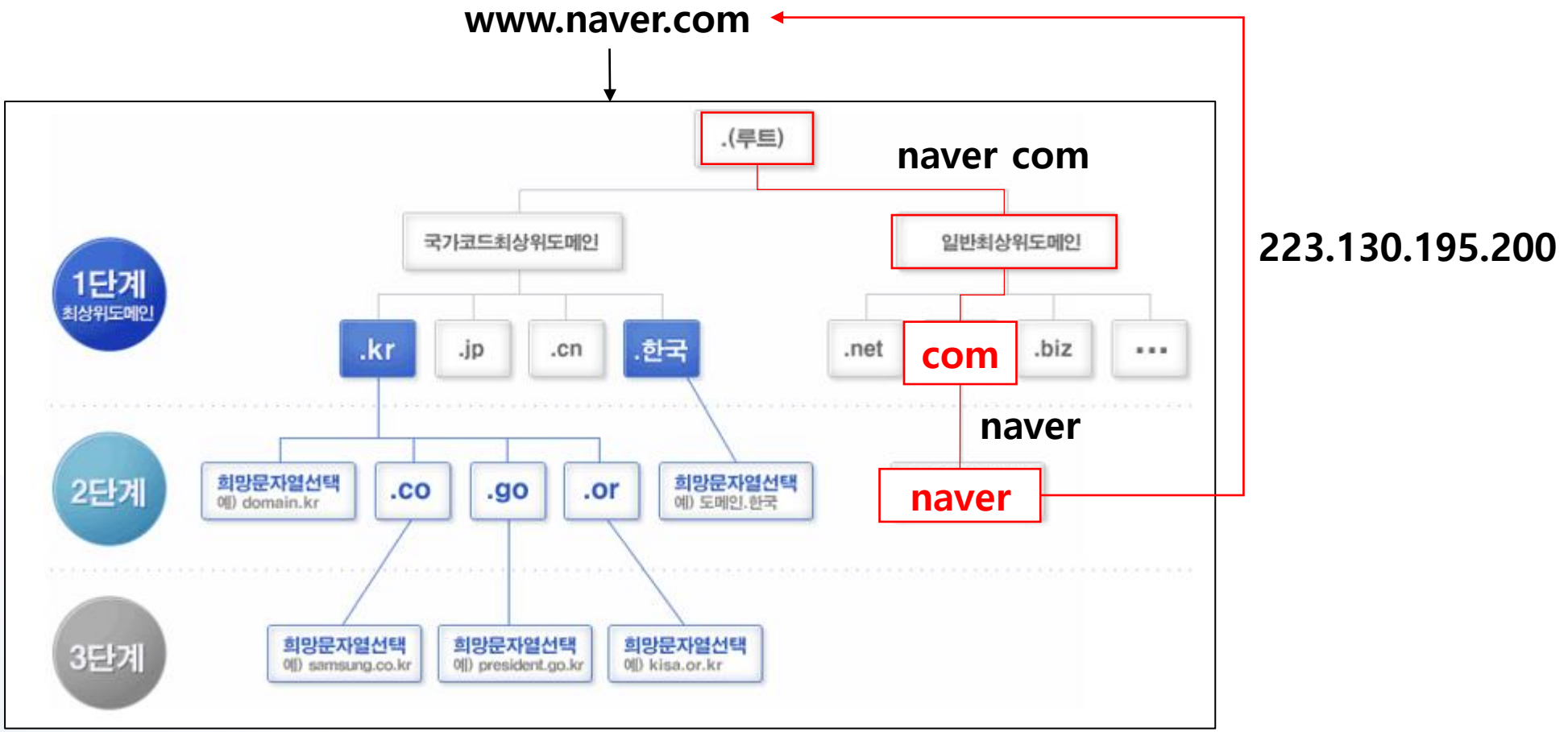
윤지
010-1010-1010

- 도메인은 이름을 붙이는 것으로 많은 서버를 기억하고 접속하기 편리해짐
  - 기업 도메인은 **발음하기 쉽거나 외우기 쉬운 형태로 등록**

- 도메인 동작원리
  - 사용자가 네이버 사이트에 접속하려고 할 때 도메인 네임서버에서 도메인과 IP의 매칭정보를 확인함

도메인 네임서버는 어떻게 IP를 찾아줄까?

- 도메인 네임서버 동작원리
  - 도메인을 나누어서 확인 수행
    Ex) 점(.)을 기준으로 서브도메인(naver), 탑도메인(com)으로 분리하여 확인



- DGA(Domain Generation Algorithm)
  - 문자, 숫자 등을 이용하여 도메인 이름을 생성하는 알고리즘

aceijlahgijp.bazar

c92e3378.duckdns.org

nerjyzkup.com

Brksiqbdcf.com

zrkyenupj.pages.dev

knjpeuzyr.tk

4933c012.net

dcfiklddhikp.bazar

23c2535b.org

DGA는 누가? 왜? 사용하는 것일까요?

- 다양한 악성코드가 공격자의 서버와 통신하기 위해 DGA를 활용
  - **abcbot**

nerjyzkup.com	knjpeuzyr.tk
zrkyenupj.pages.dev	qzqwodqcx.pages.dev

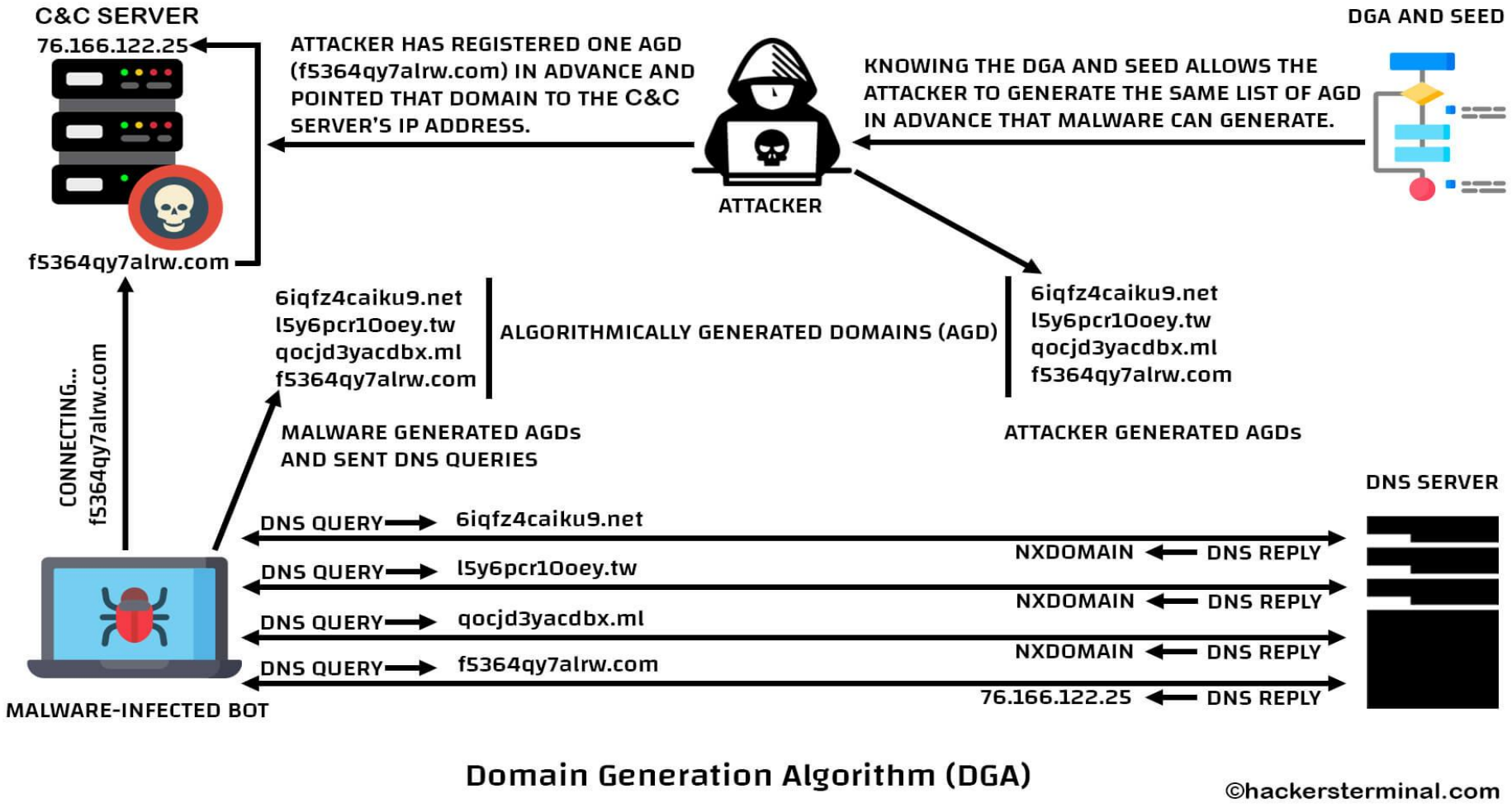
  - **Antavmu**

2c2b3749.com	4933c012.net
c92e3378.duckdns.org	23c2535b.org

  - **ares**

aceijlahgijp.bazar	acfiimahhiiq.bazar
efhijkekjijo.bazar	dcfiklddhikp.bazar

- 다양한 악성코드가 공격자의 서버와 통신하기 위해 DGA를 활용



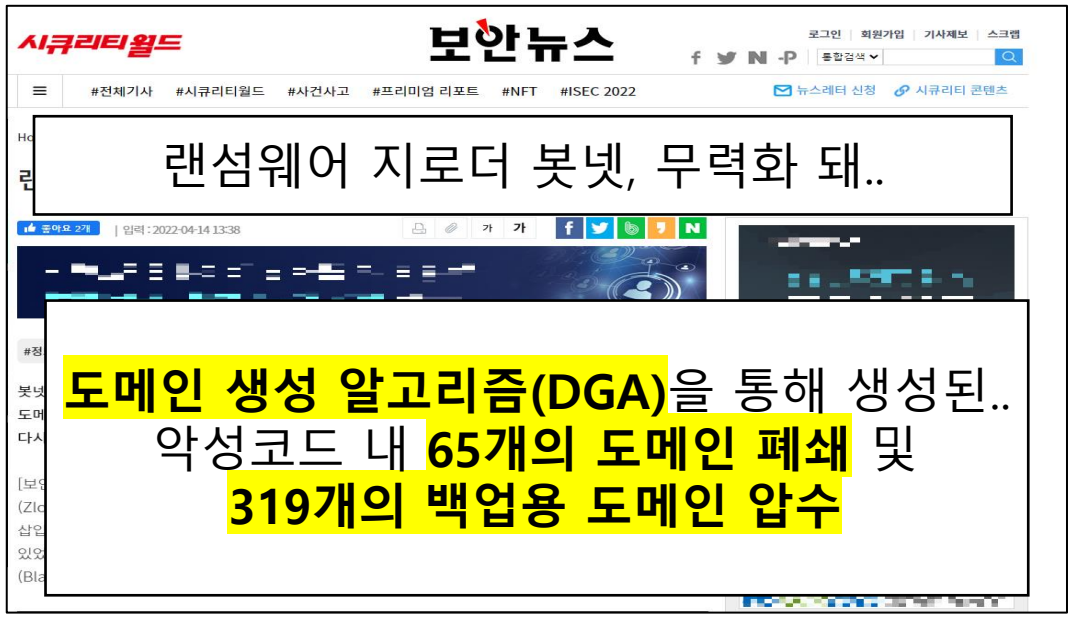
데이터 분석 및 모델링

02

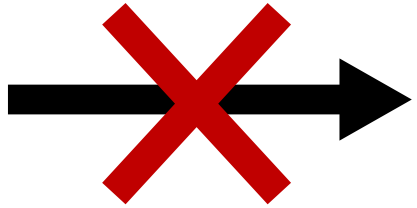
- DGA 도메인 탐지 배경 및 문제정의
- 데이터 수집
- 데이터 탐색 및 전처리
- 데이터 분석 및 특징 추출
- 모델링 및 검증

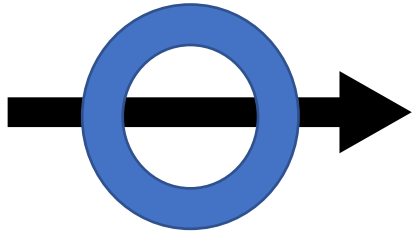
DGA 도메인 탐지 배경

- **악성코드가 DGA 도메인을 지속적으로 이용하고 있음**



- DGA 도메인과 정상도메인을 판별할 수 있다면 악성코드로 인한 피해를 예방할 수 있음

nerjyzkup.com  DGA 도메인이야! 도망가!

naver.com  정상 도메인이네! 지나가!

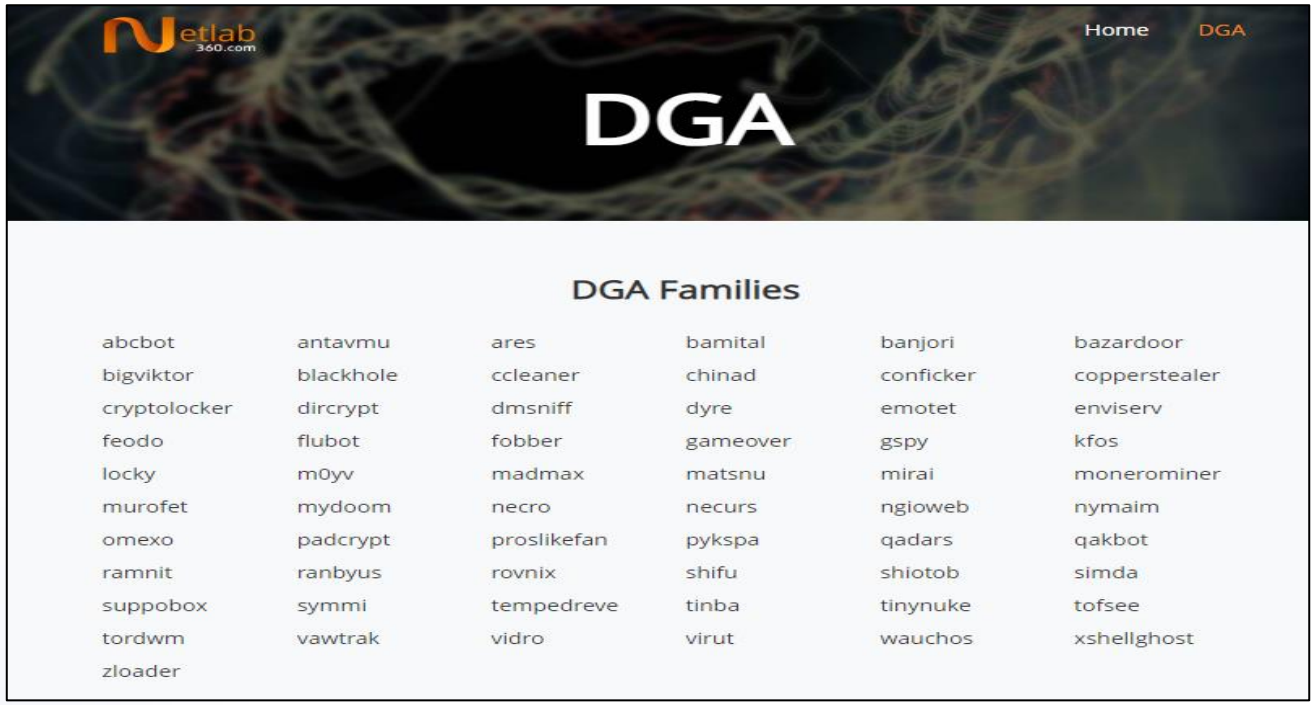
DGA를 통해 생성된 도메인을 탐지하자

DGA를 통해 생성된 도메인을 탐지하자

- DGA를 통해 생성된 도메인은 악성도메인으로 정의
  → 피싱사이트, 정상사이트의 해킹으로 인한 악의적인 목적으로 활용되는 것 등 제외

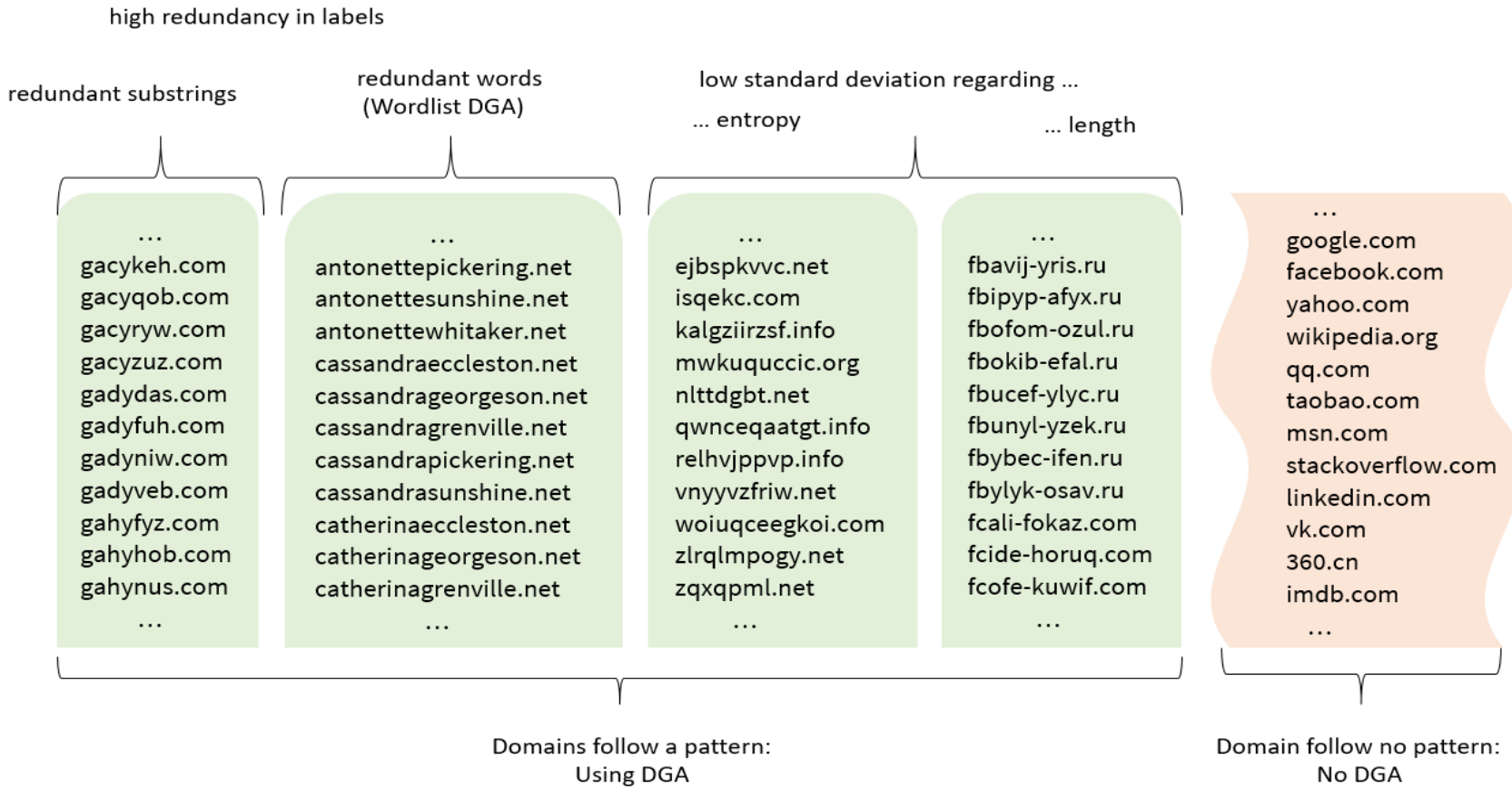
- 사용자들이 많이 이용하는 사이트를 정상도메인으로 정의
  ※ google, youtube, naver 등등

- DGA로 생성된 도메인을 탐지하기 위해 필요한 데이터 및 수집채널 조사
  - DGA 데이터 수집
    - Netlab360 : https://data.netlab.360.com/dga/



- DGA로 생성된 도메인을 탐지하기 위해 필요한 데이터 및 수집채널 조사
  - 정상 도메인 데이터 수집
    - majestic : https://majestic.com/reports/majestic-million
    - similarweb : https://www.similarweb.com/top-websites/

- **DGA(0,1) 및 Class(0~19) 데이터를 나누어 살펴보자**



- DGA(0,1) 및 Class(0~19) 데이터를 나누어 살펴보자

Domain	도메인네임
dga	DGA 여부(Yes/No) * No : 정상도메인 * Yes : DGA 도메인
Class	DGA 악성코드군 * 0 : 정상도메인 * 1~19 : 악성코드군별 DGA도메인

- DGA(0,1) 및 Class(0~19) 데이터를 나누어 살펴보자
  - 정상도메인과 DGA도메인을 구분할 수 있는 특징은?
  - DGA 클래스별로 구분할 수 있는 특징은?

- **정상도메인은 악성도메인에 비해 읽고 기억하기 쉽다.**

정상 도메인	
– smartidpro.in	– wdupload.com
– extranet-altone.fr	– ebook2pdf.com
– directan.pl	– techgamea.com
– realher.com	– mallpass.co.kr
– kingkoil.com.sg	– eig.org
– error-64.icu	– xzw.com
– surflocos.com	– sevilla.org

*악성 도메인
– b64273681406245l5a0ee4c29d27b789.info
– 9mtrkmet8rjnpya4.ru
– brksiqbdcf.com
– fkfkhancwqeifu.biz
– i5ea2b9c64c4a346b1c909c98385d10b90.hk
– 170qknj1onn4c37qf1w71rzei1h.org
– ejahtpnopoifq.click

* 악성 도메인 : DGA를 통해 생성된 도메인

- **DGA(0,1) 및 Class(0~19) 데이터를 나누어 살펴보자**

Num	Features
1	도메인 길이
2	연속된(3개 이상) 자음·숫자 개수
3	자음·모음 비율
4	엔트로피
5	TLD(Top Level Domain)
6	N-gram Score

새싹(SeSAC) New 금천캠퍼스 2기 인공지능 과정

기업데이터를 활용한 AI 취업 부트캠프

감사합니다.

Seoul Software ACademy

글로벌 소프트웨어 캠퍼스