



Building IP Geolocation Database from Online Used Market Articles

Hyunsu Mun, Youngseok Lee

Data Network Laboratory

Chungnam National University, Daejeon, Korea

APNOMS 09/27/17

munhyunsu@cnu.ac.kr

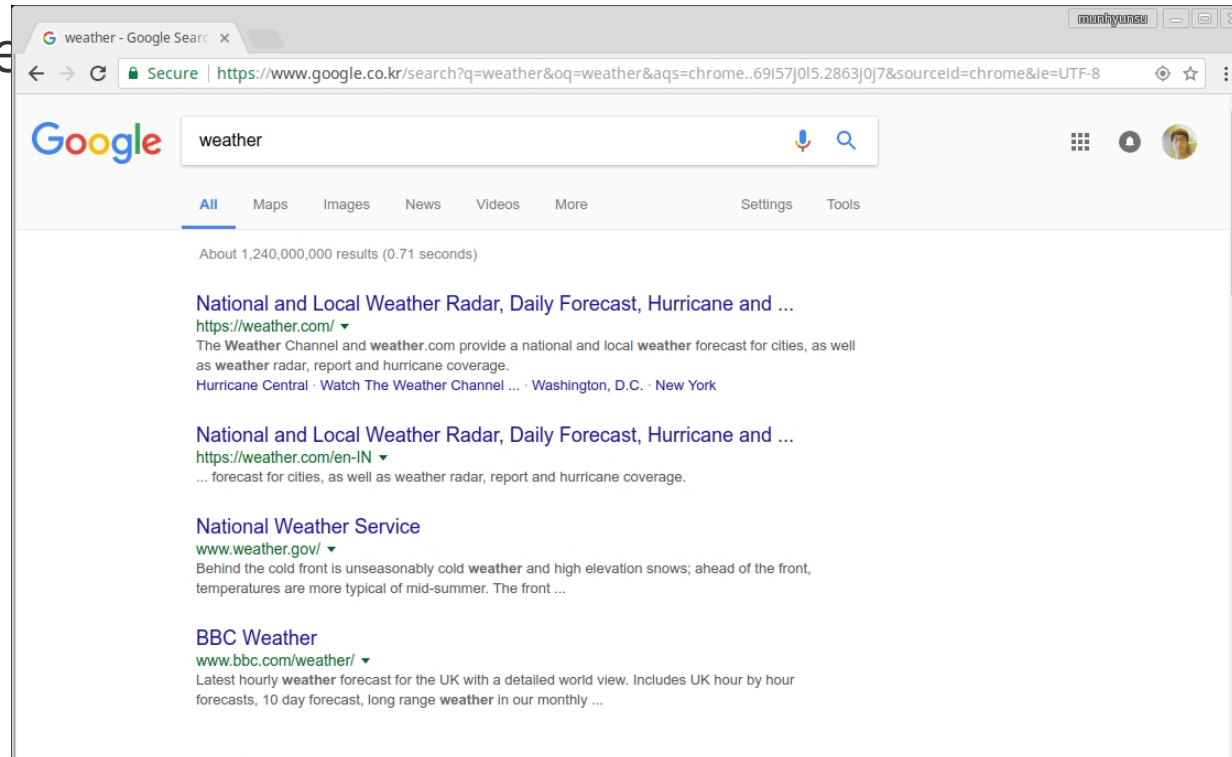
One morning in the past

- When the weather is curious
 - Enter Weather on Internet search site!

One morning in the past

- When the weather is curious

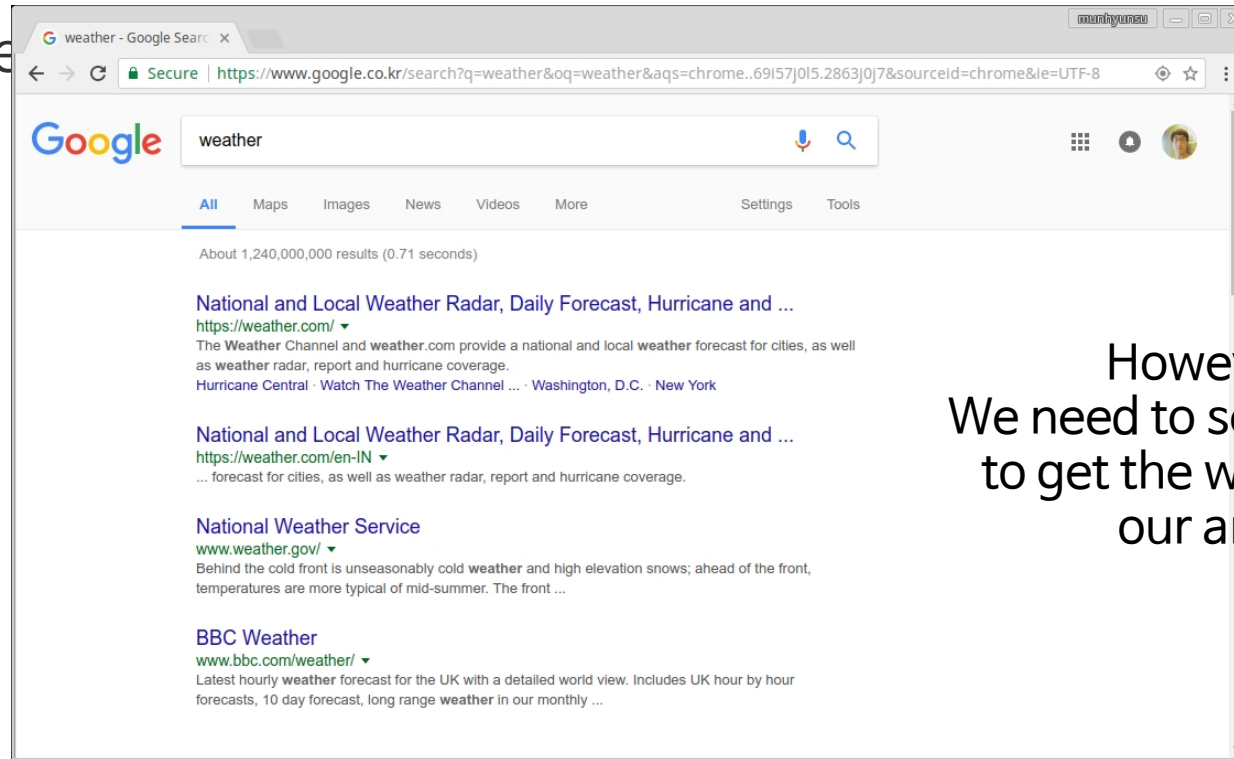
– Enter



One morning in the past

- When the weather is curious

– Enter



However,
We need to search again
to get the weather of
our area.

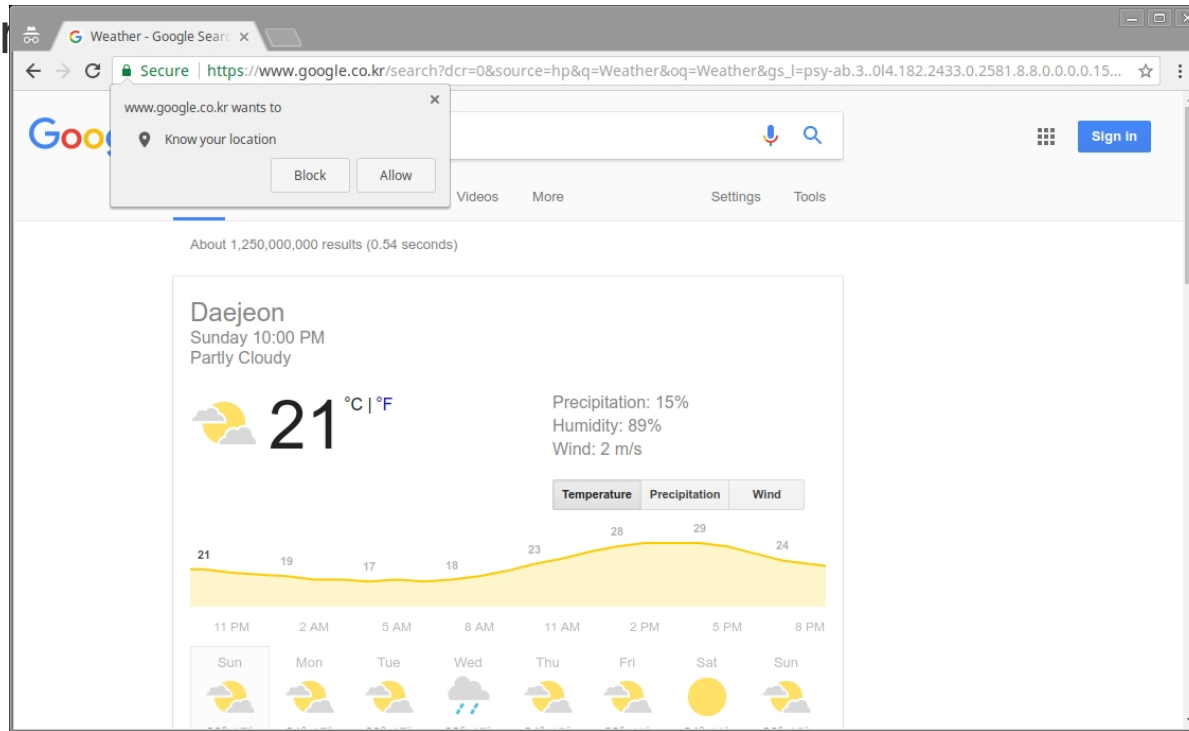
One morning nowadays

- When the weather is curious
 - Enter Weather on Internet search site like past!

One morning nowadays

- When the weather is curious

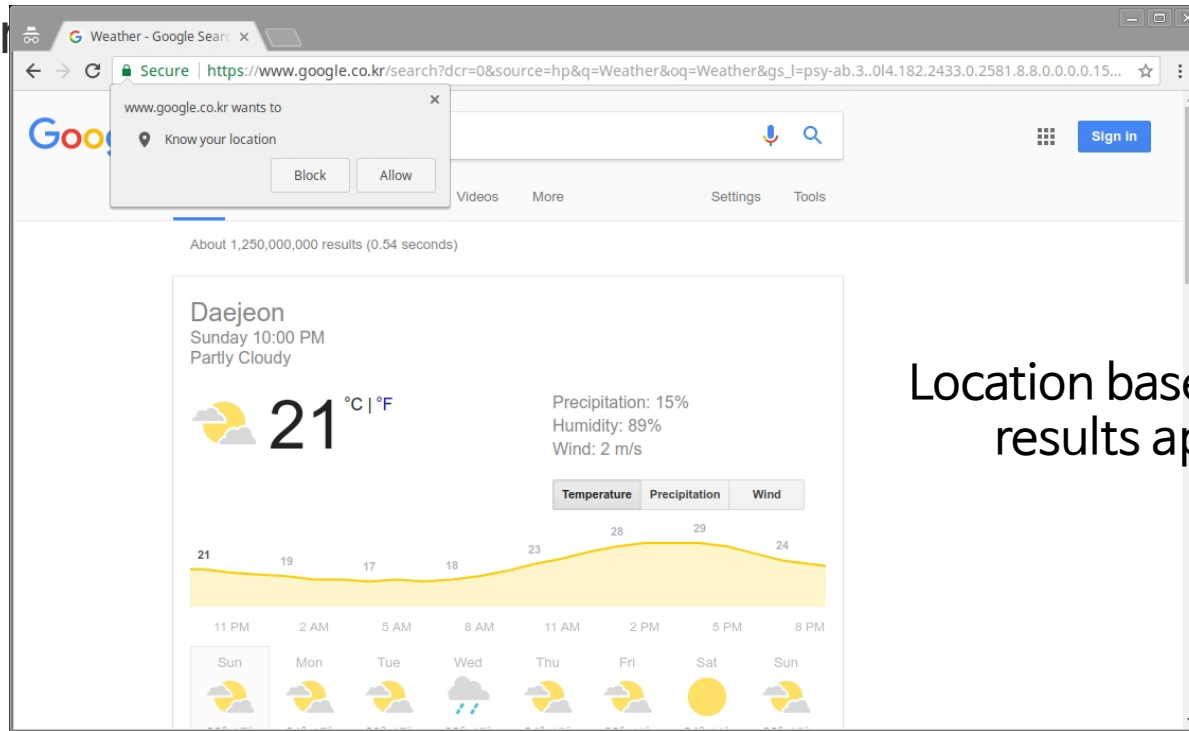
— Enter



One morning nowadays

- When the weather is curious

— Enter



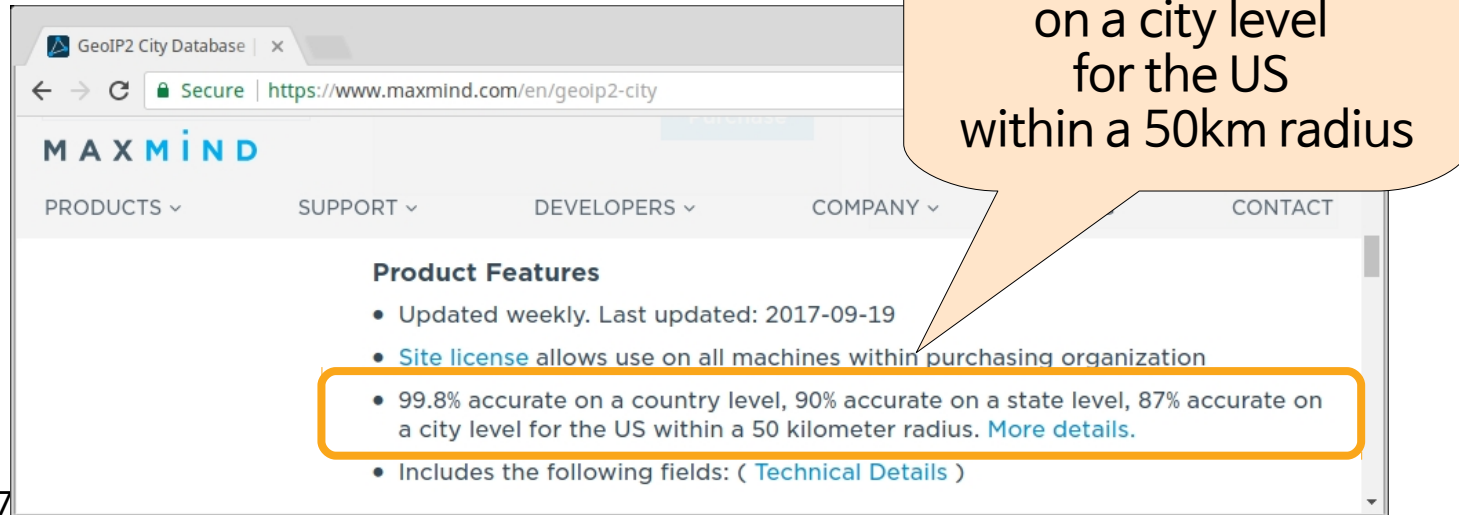
Location based search results appear

IP Geolocation

- Geolocation: Identifying the actual geographical location of objects (or analogy)
- IP Geolocation: Using IP address to identify geographical location (or analogy)
 - When searching for Weather, use IP address to show location based results
 - A methods to discover the location of a client when a resource request is received through the network.

The Accuracy of IP Geolocation

- IP geolocation databata has an accuracy of geolocation
- Commercial companies such as MaxMind and IP2Location reveal their accuracy in database descriptions



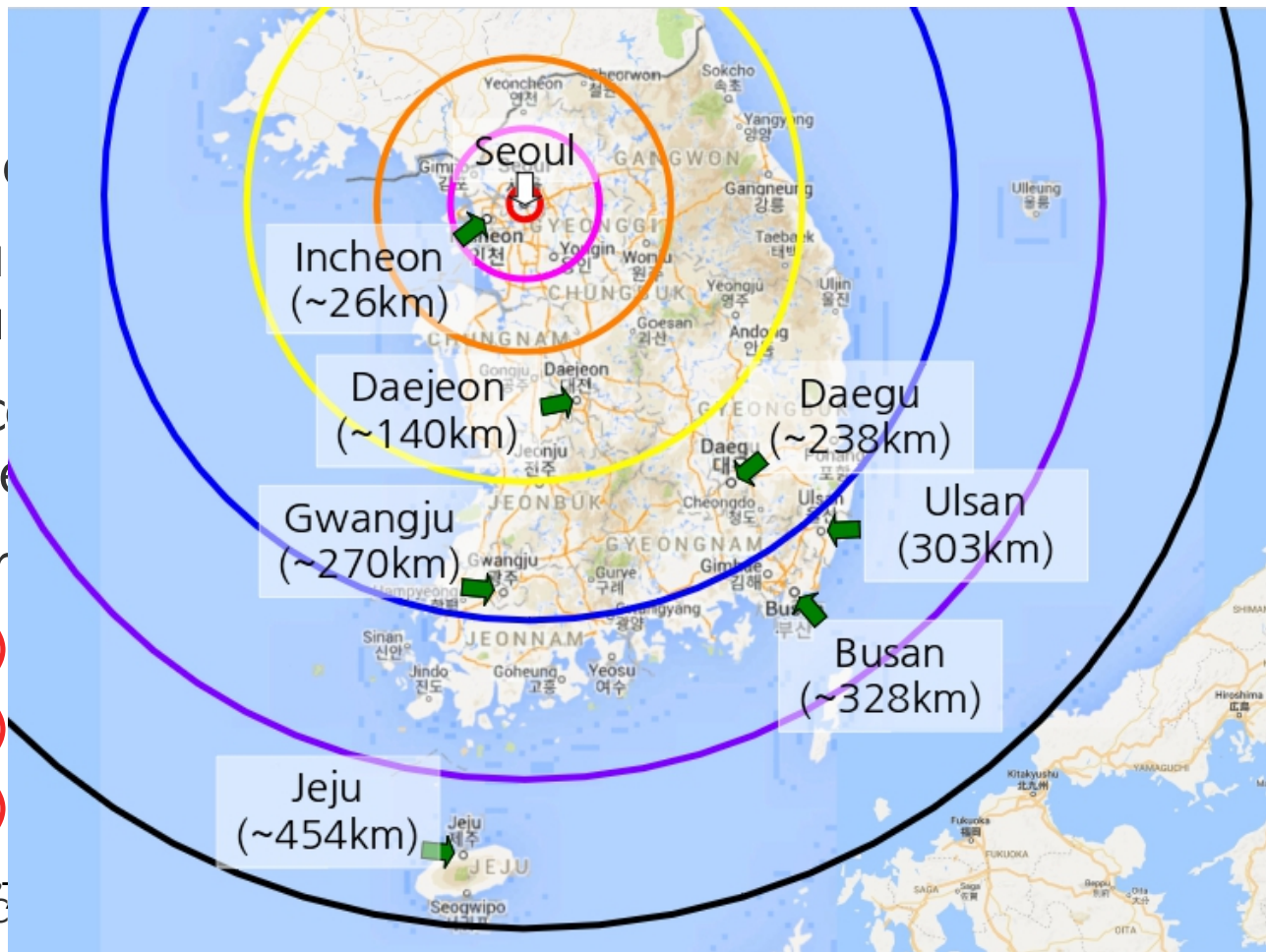
The screenshot shows the MaxMind website's 'GeoIP2 City Database' page. The browser address bar displays 'https://www.maxmind.com/en/geolp2-city'. The MaxMind logo is at the top left, followed by navigation links: PRODUCTS, SUPPORT, DEVELOPERS, COMPANY, and CONTACT. Under the 'Product Features' section, a list of bullet points is shown. One bullet point is highlighted with an orange box: '99.8% accurate on a country level, 90% accurate on a state level, 87% accurate on a city level for the US within a 50 kilometer radius. [More details.](#)'. An orange speech bubble points to this text, containing the text: '87% accurate on a city level for the US within a 50km radius'.

87% accurate
on a city level
for the US
within a 50km radius

The Low Accuracy in Korea

- Comm
- Measu
- measu
- The ac
- as state
- Kor

1)
2)
3)



[1]

abase is

not show

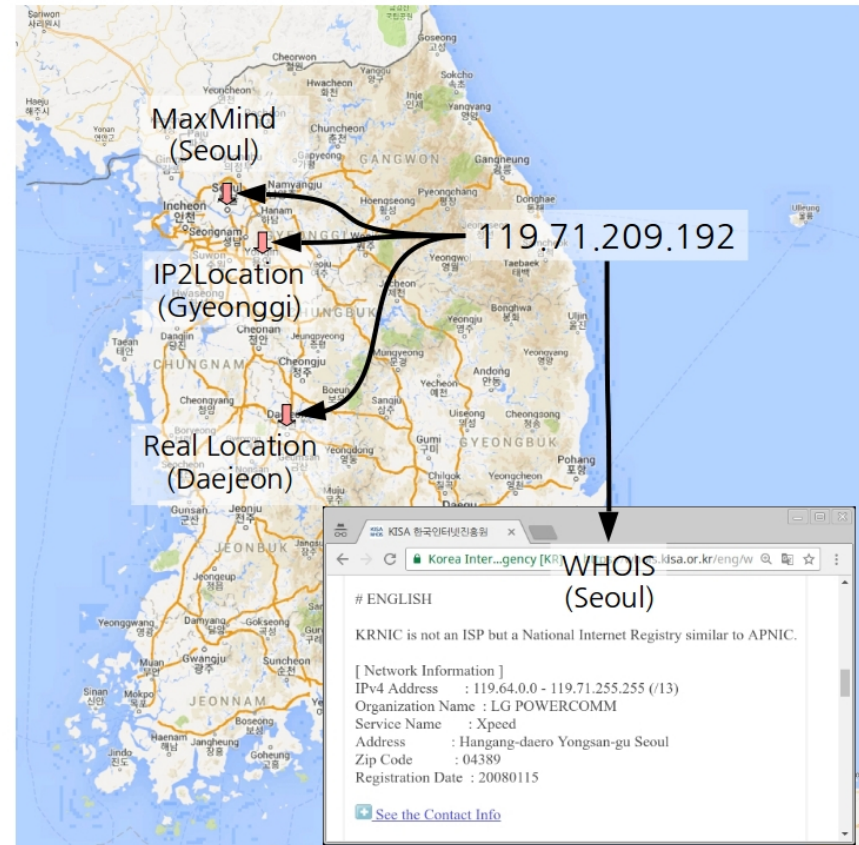
N

The Low Accuracy in Korea

- Commercial database failed to ensure city-level accuracy [1]
- Measuring the accuracy of commercial IP geolocation database is measured in U.S.
- The accuracy of commercial IP geolocation in Korea does not show as stated in the product description
 - Korea IP address distribution sequence
 - 1) IANA (Internet Assigned Numbers Authority): ICANN
 - 2) RIR (Regional Internet Registry): APNIC
 - 3) NIR (National Internet Registry): KISA

The IP Geolocation Error in Korea

- IP address of Daejeon, Korea is mapped to a different location by databases
 - Real Location: Wired ethernet
 - MaxMind, IP2Location: Commercial version (Not lite version)
 - WHOIS: KISA WHOIS API

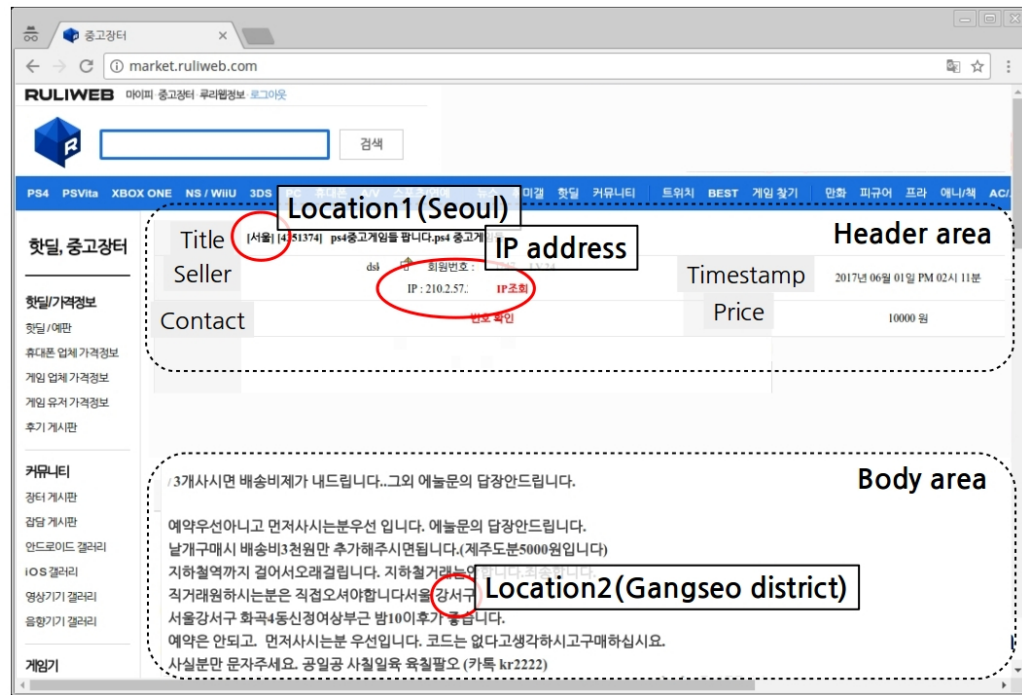


To Better IP Geolocation

- City-level IP geolocation based on PoP-level network analysis [2]
 - IP geolocation using landmarks [3]
 - **Building IP geolocation database with a crowd-sourcing [4]**
 - **Using broadband performance test tool data**
-
- [2] S. Liu, F. Liu, F. Zhao, L. Chai, and X. Luo, “Ip city-level geolocation based on the pop-level network topology analysis,” in Information Communication and Management (ICICM), International Conference on. IEEE, 2016, pp. 109 - 114
 - [3] H. Jiang, Y. Liu, and J. N. Matthews, “Ip geolocation estimation using neural networks with stable landmarks,” in Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on. IEEE, 2016, pp. 170 - 175.
 - Y. Lee, H. Park, and Y. Lee, “Ip geolocation with a crowd-sourcing broadband performance tool,” ACM SIGCOMM Computer Communication Review, vol. 46, no. 1, pp. 12 - 20, 2016.

The Ruliweb Article

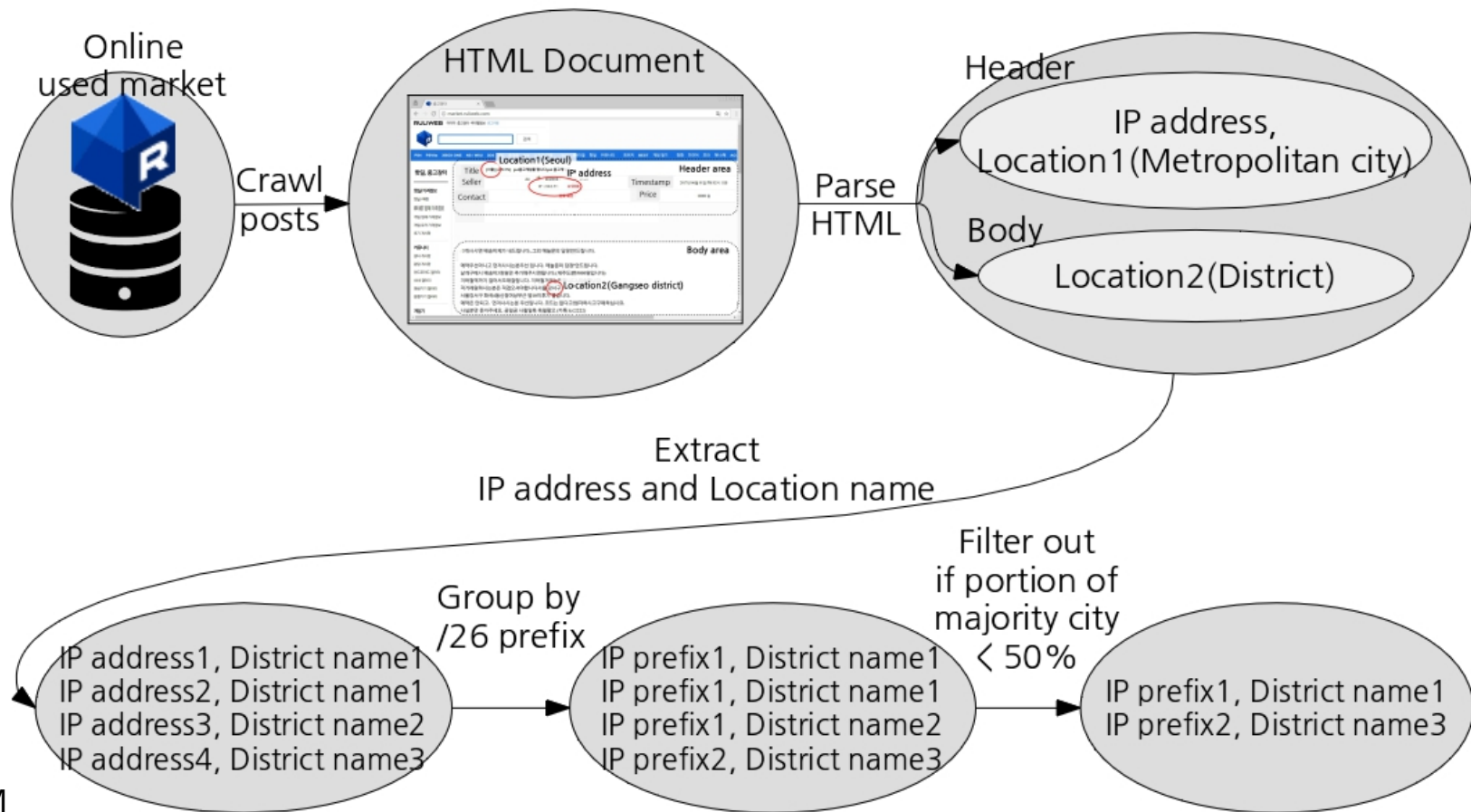
- Create an average of 300 posts
- Ruliweb postings are required to reveal the location (Province-level) and IP address
- IP address disclosure to prevent fraud user
- Seller to reveal the district-level location in the body area
 - about 45% disclosure



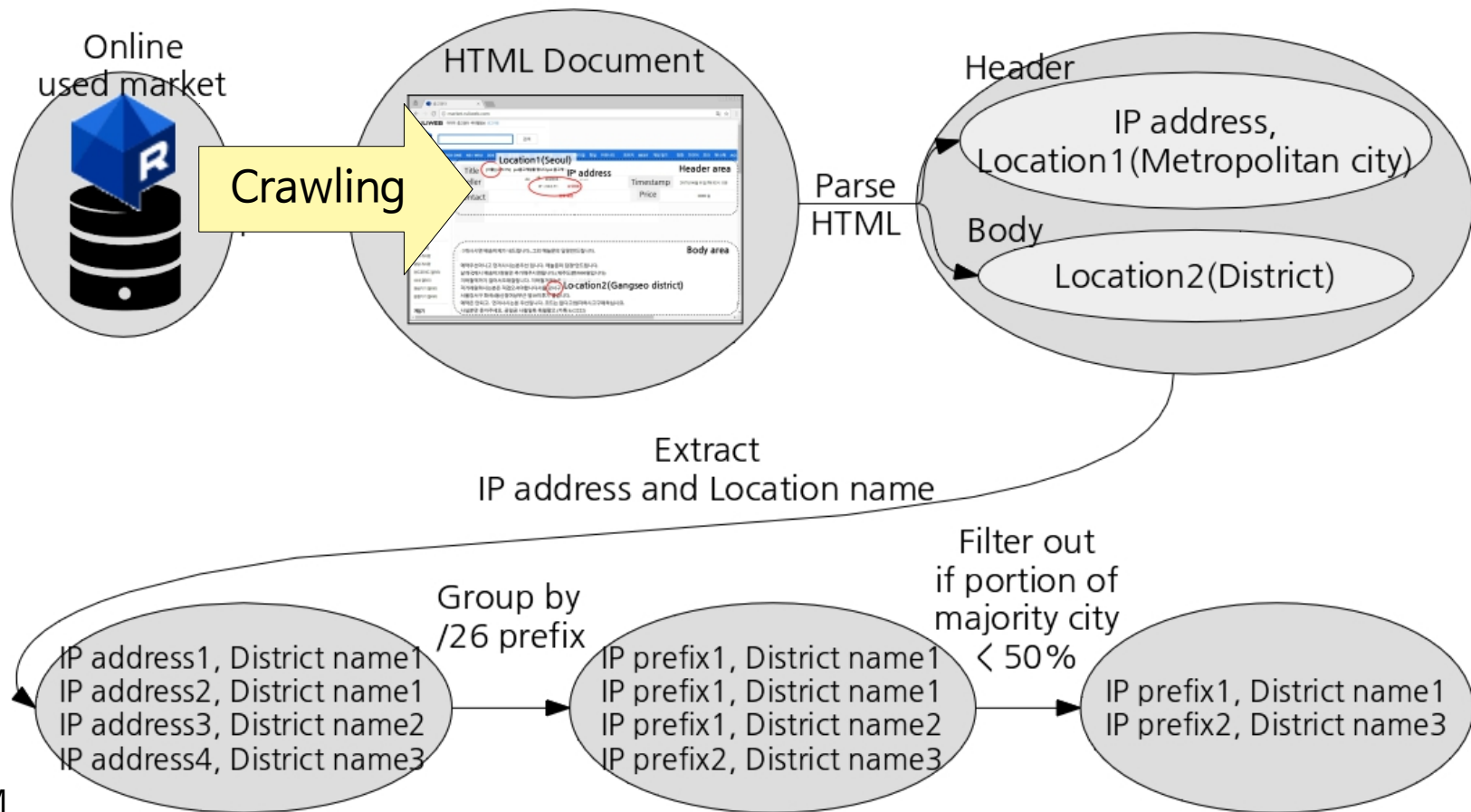
The Key Challenges

- Broad region coverage
 - The IP address assigned to the user of all districts
 - Not IP block coverage
- Continuous update
 - Flexibility to the change of ISP's policy

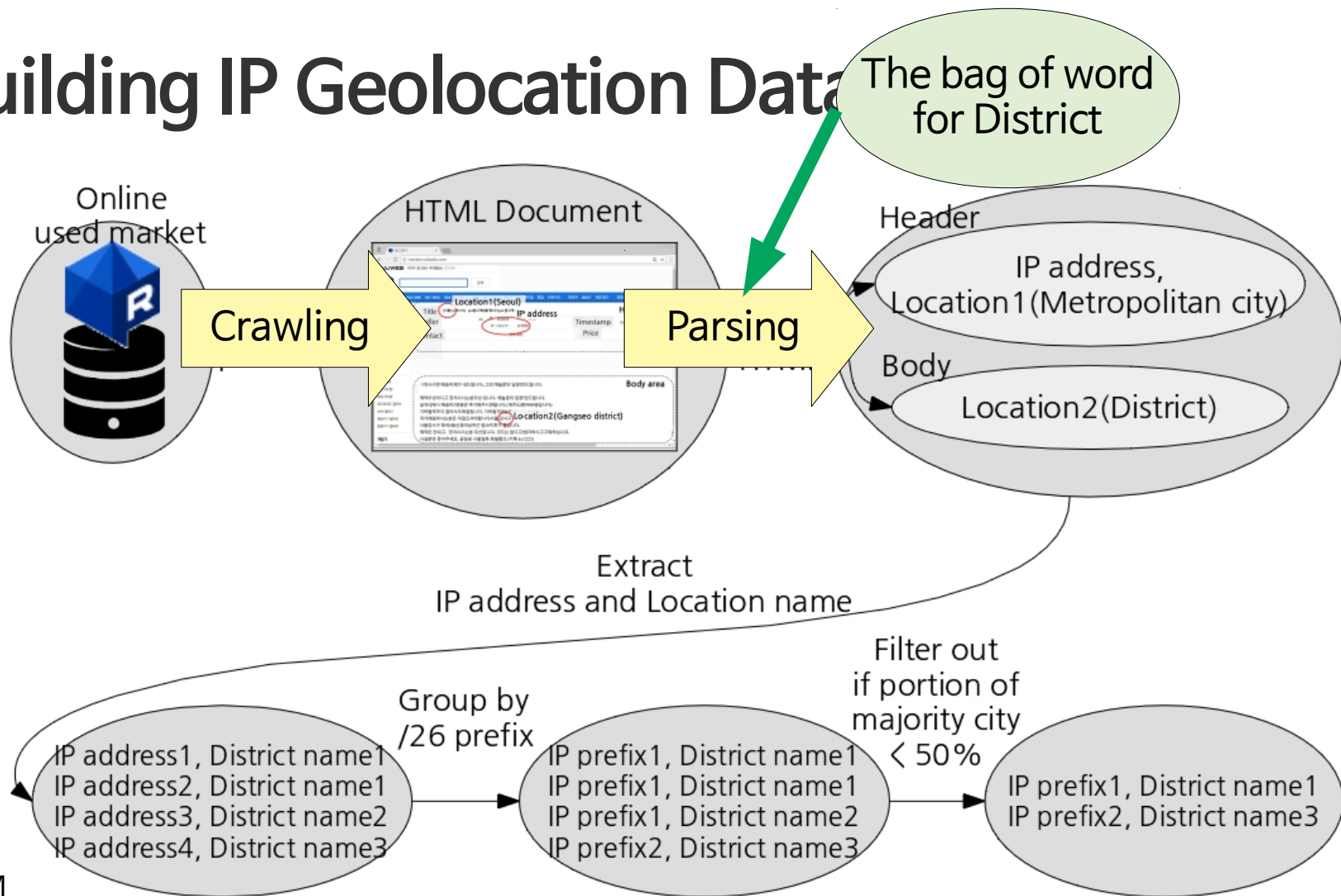
Building IP Geolocation Database



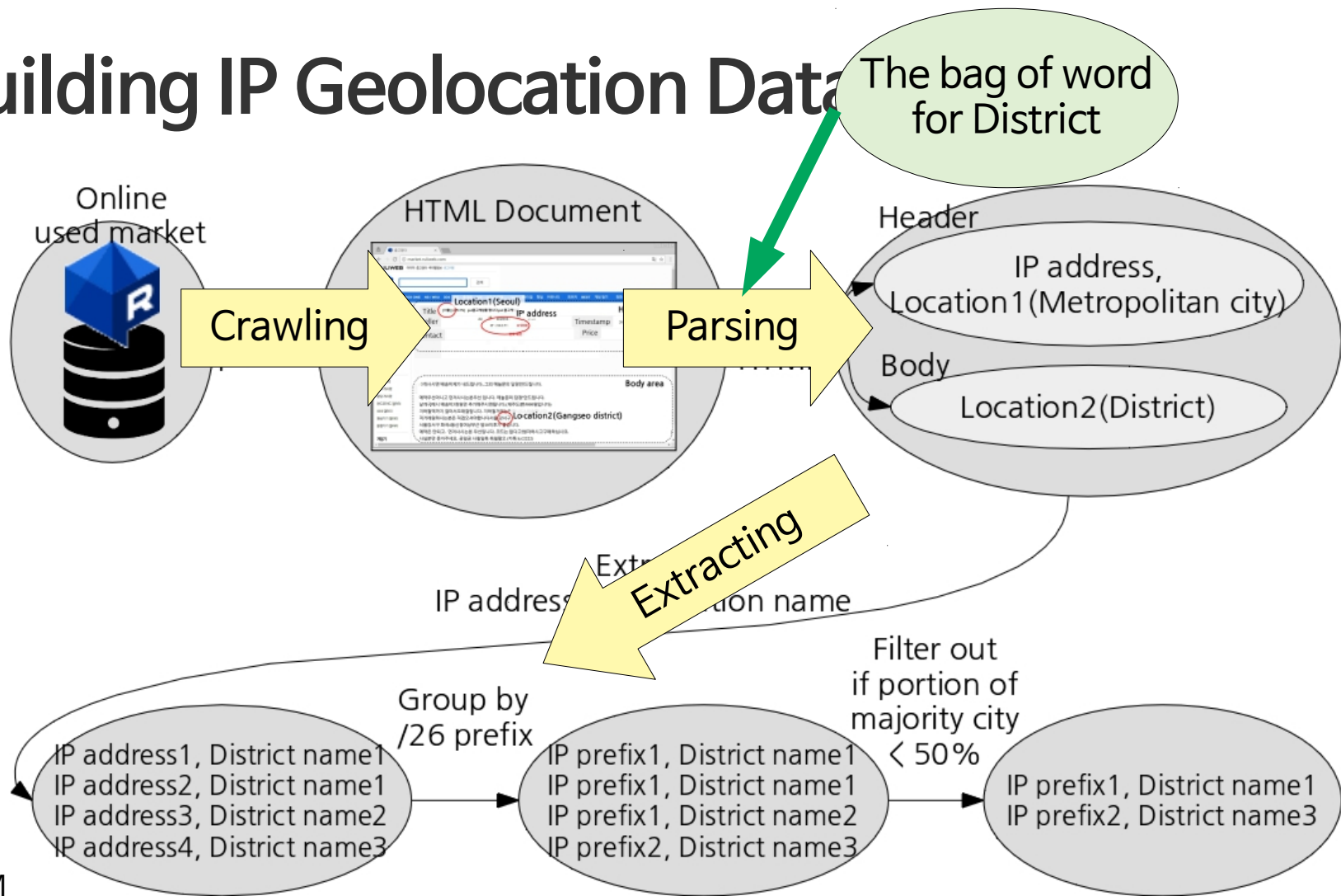
Building IP Geolocation Database



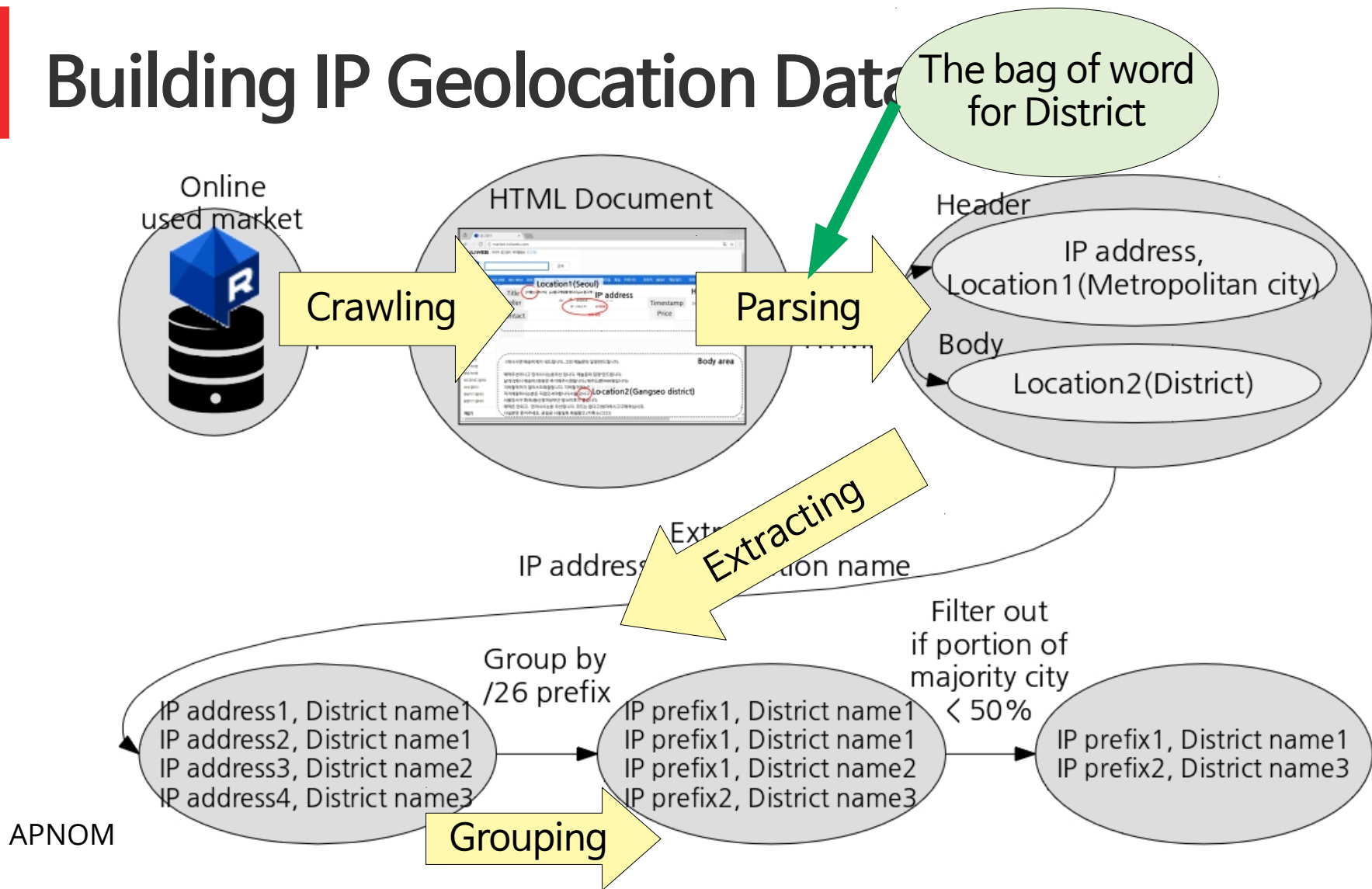
Building IP Geolocation Data



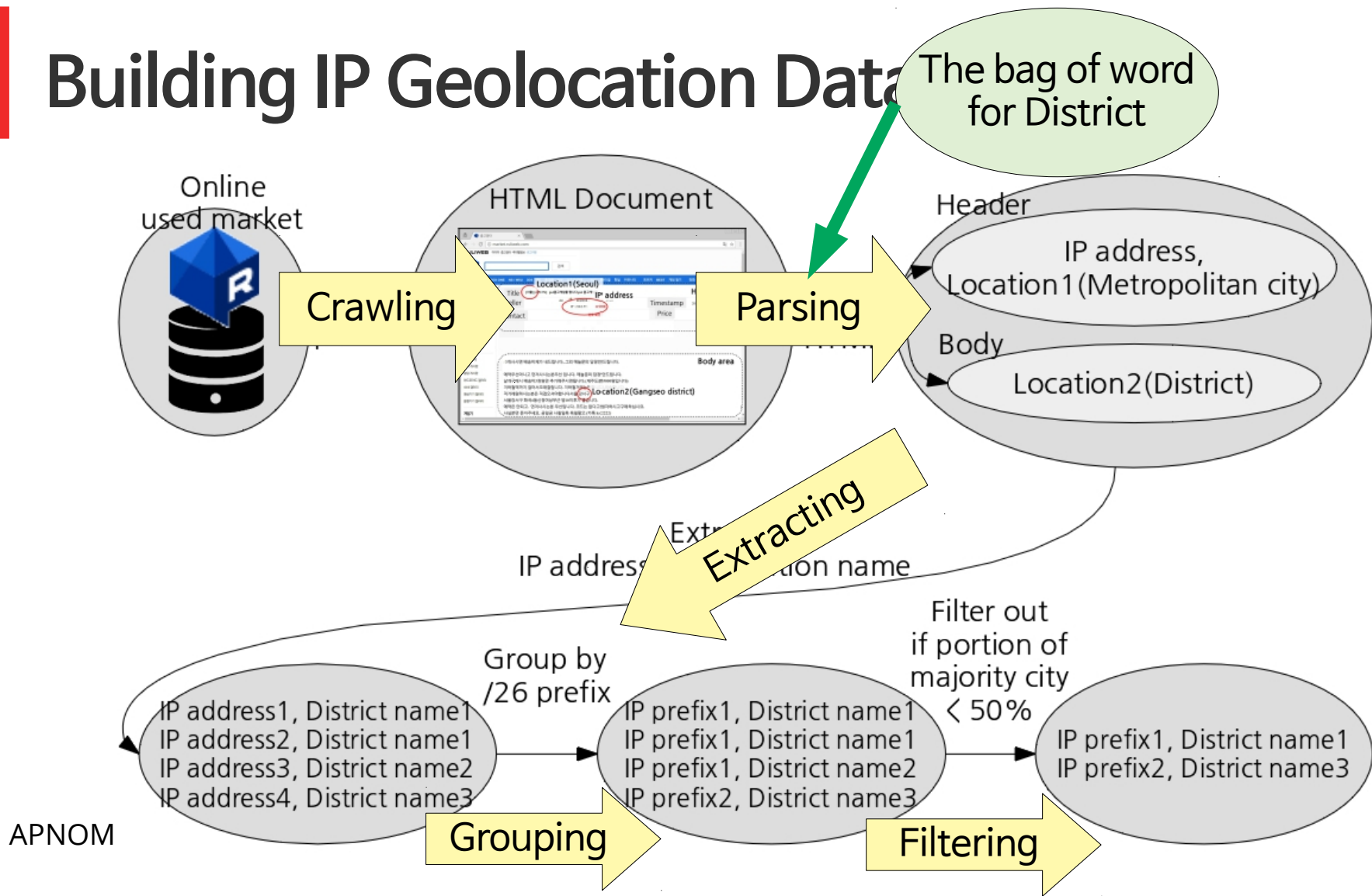
Building IP Geolocation Data



Building IP Geolocation Data

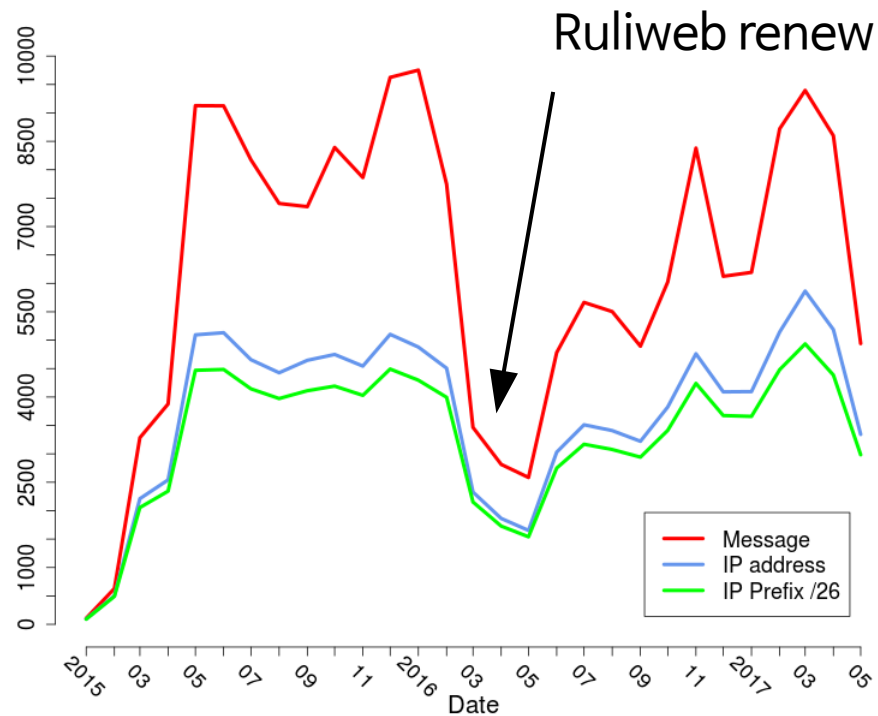


Building IP Geolocation Data



The Data Summary

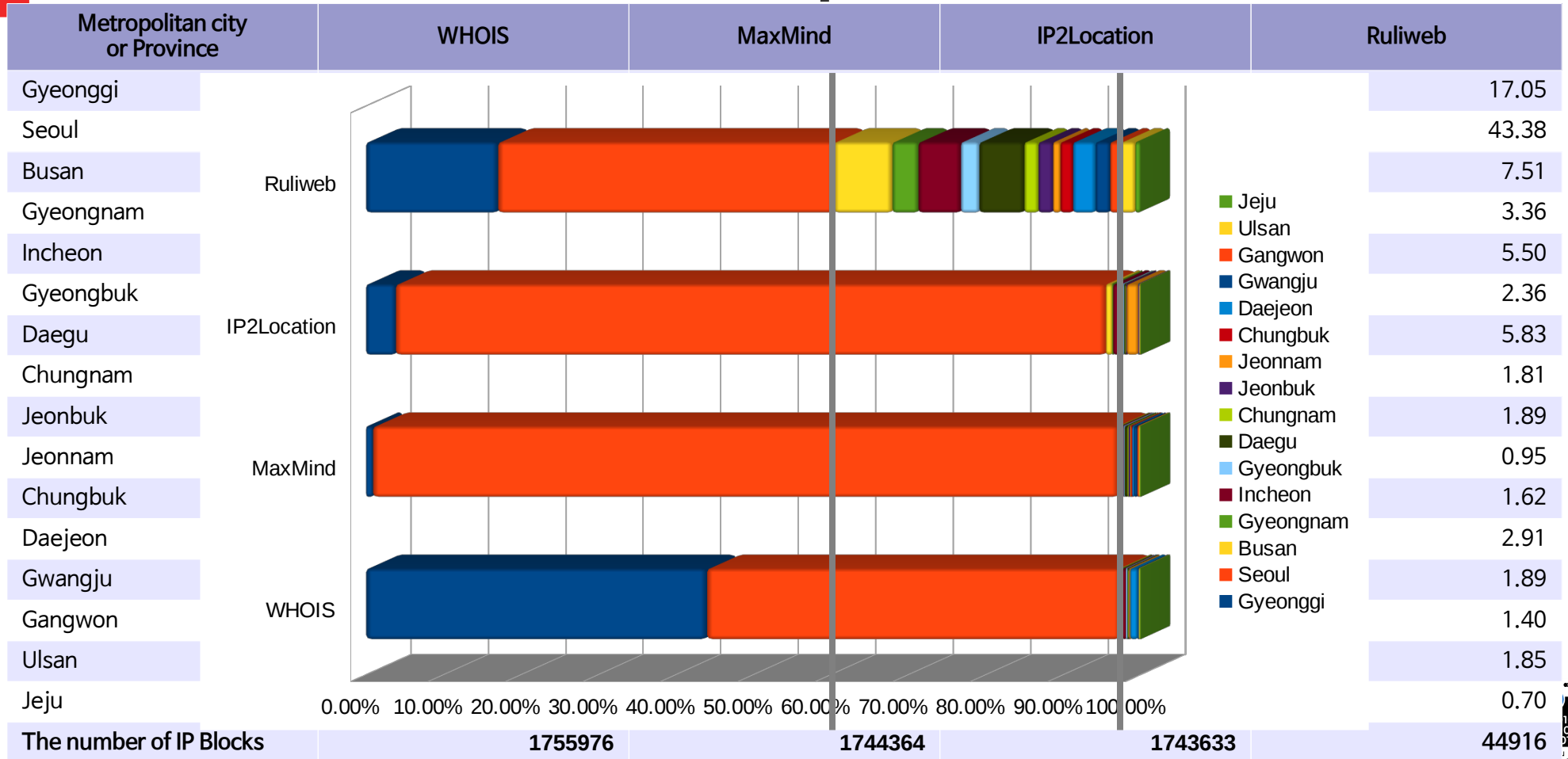
- 2015. 01. ~ 2017. 05.
- 195,937 Posts
 - 11.7 Gbytes
 - 34,740 Sellers
- 71,449 IP addresses
 - 44,916 /26 IP Prefixes



The IP Block (/26) Ratio per Databases

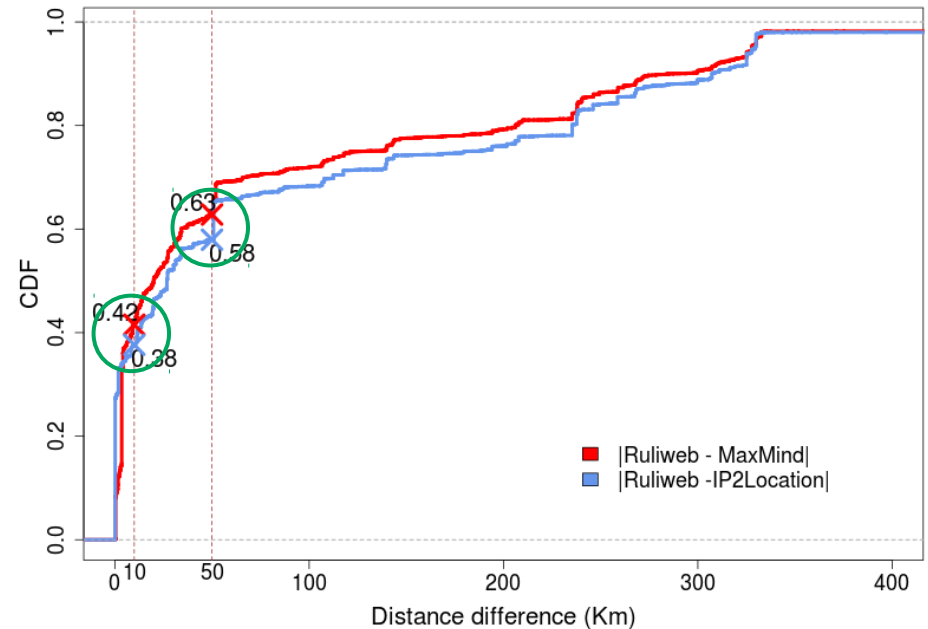
Metropolitan city or Province	WHOIS	MaxMind	IP2Location	Ruliweb
Gyeonggi	44.01	0.90	3.84	17.05
Seoul	53.39	96.23	91.67	43.38
Busan	0.12	0.20	0.83	7.51
Gyeongnam	0.16	0.04	0.09	3.36
Incheon	0.41	0.44	1.16	5.50
Gyeongbuk	0.19	0.09	0.26	2.36
Daegu	0.18	0.44	0.25	5.83
Chungnam	0.03	0.08	0.03	1.81
Jeonbuk	0.07	0.21	0.07	1.89
Jeonnam	0.06	0.14	1.50	0.95
Chungbuk	0.00	0.12	0.01	1.62
Daejeon	0.99	0.24	0.12	2.91
Gwangju	0.16	0.44	0.02	1.89
Gangwon	0.00	0.20	0.06	1.40
Ulsan	0.13	0.14	0.05	1.85
Jeju	0.10	0.09	0.05	0.70
The number of IP Blocks	1755976	1744364	1743633	44916

The IP Block (/26) Ratio per Databases



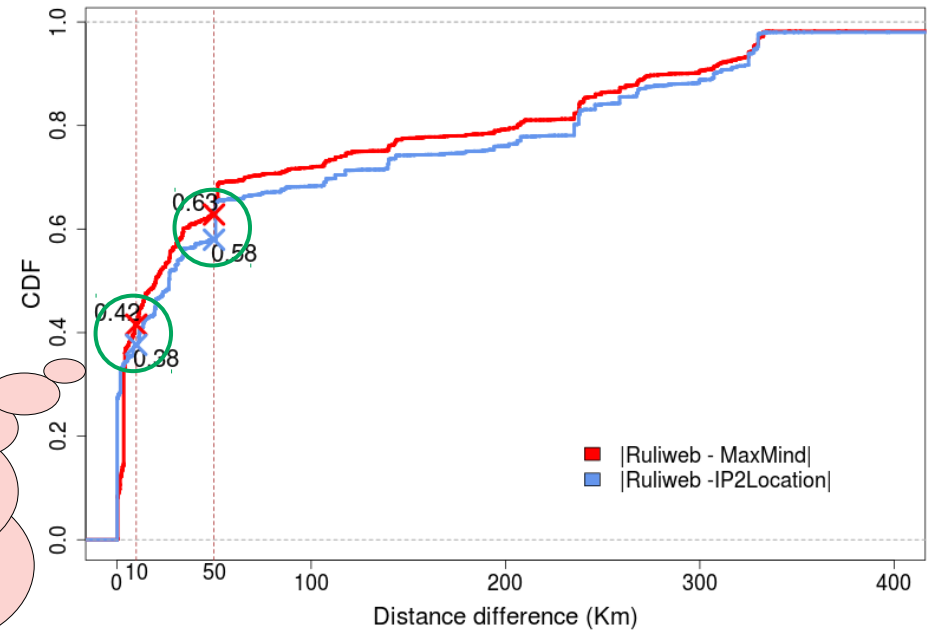
The Accuracy Of The Crowd-Sourcing DB.

- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database



The Accuracy Of The Crowd-Sourcing DB.

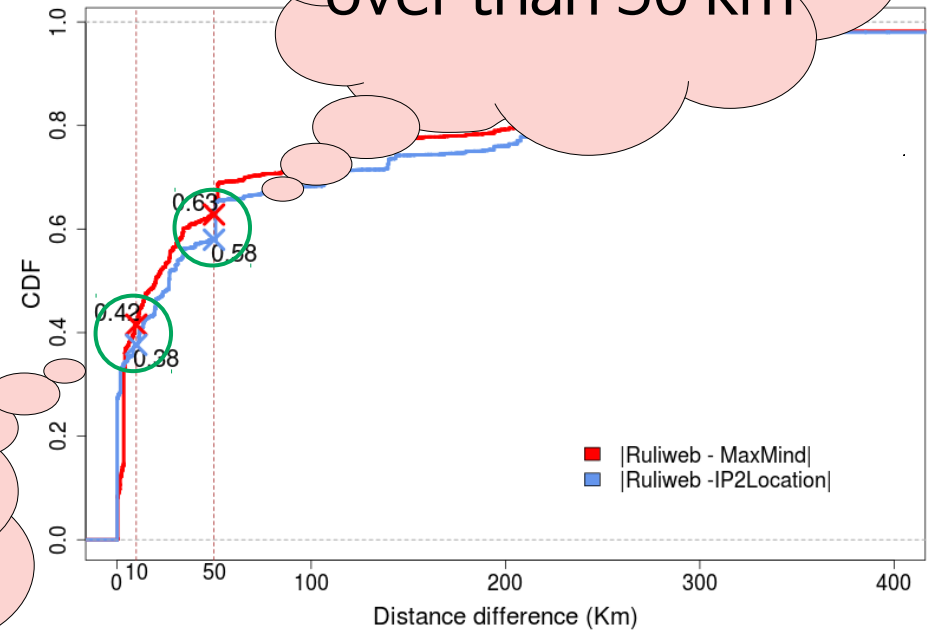
- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database



39% difference
less than 10 km

The Accuracy Of The Crowd-Sourcing DB.

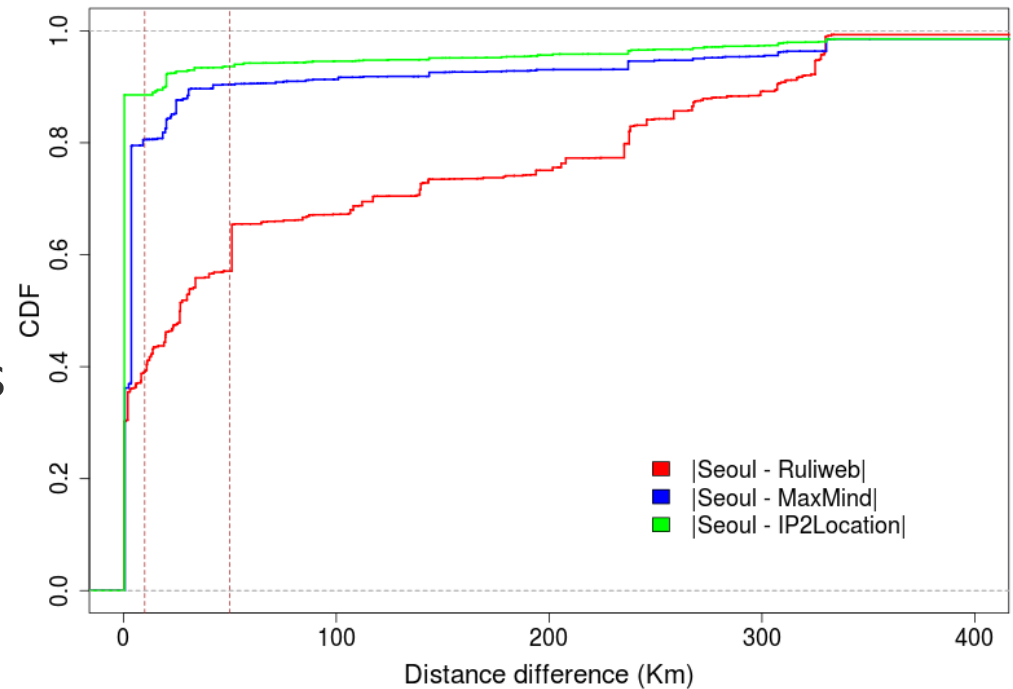
- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database



39% difference
less than 10 km

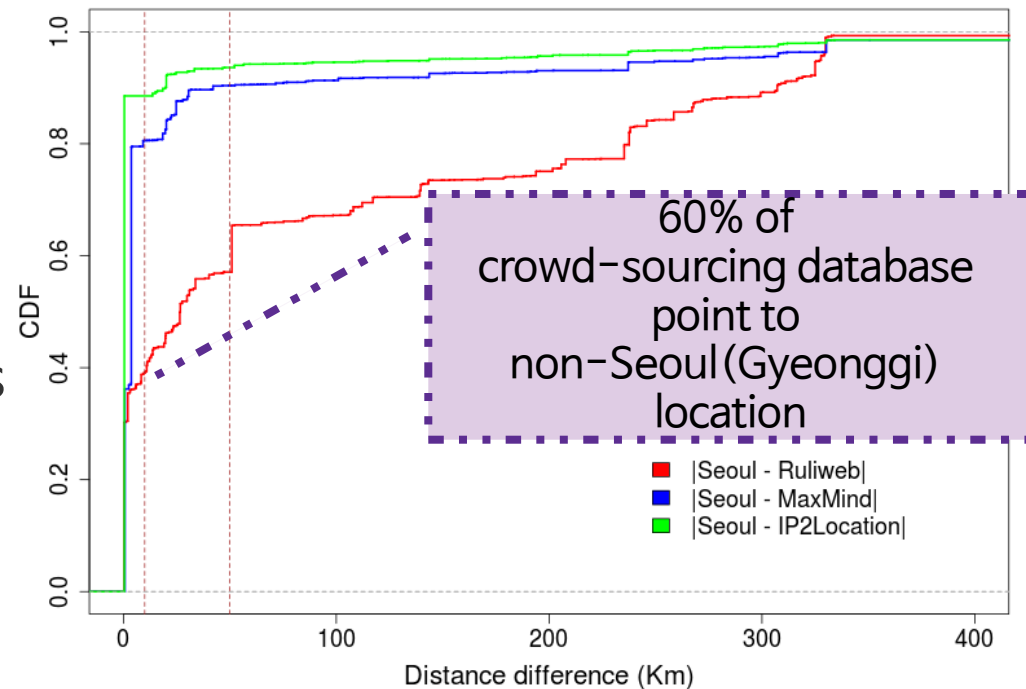
The Distance from Seoul

- 80~90% of commercial database point to Seoul (Capital city of Korea)
- Seoul: The location of the ISP assigned the IP address
 - The reasons for why Seoul is so many in database



The Distance from Seoul

- 80~90% of commercial database point to Seoul (Capital city of Korea)
- Seoul: The location of the ISP assigned the IP address
 - The reasons for why Seoul is so many in database



Province/District-level IP Prefix (/26)

Metropolitan city or Province	Province-level IP prefix # (A)	District-level IP prefix # (B)	B/A (%)
Gyeonggi	8584	4434	51.65
Seoul	21579	6490	30.08
Busan	3710	2498	67.33
Gyeongnam	1693	993	58.65
Incheon	2750	1976	71.85
Gyeongbuk	1153	684	59.32
Daegu	2888	1817	62.92
Chungnam	917	458	49.95
Jeonbuk	957	593	61.96
Jeonnam	480	137	28.54
Chungbuk	815	496	60.86
Daejeon	1475	930	63.05
Gwangju	944	581	61.55
Gangwon	686	331	48.25
Ulsan	939	434	46.22
Jeju	346	10	2.89
Total	49916	22862	45.80

APNOM

30 / 37

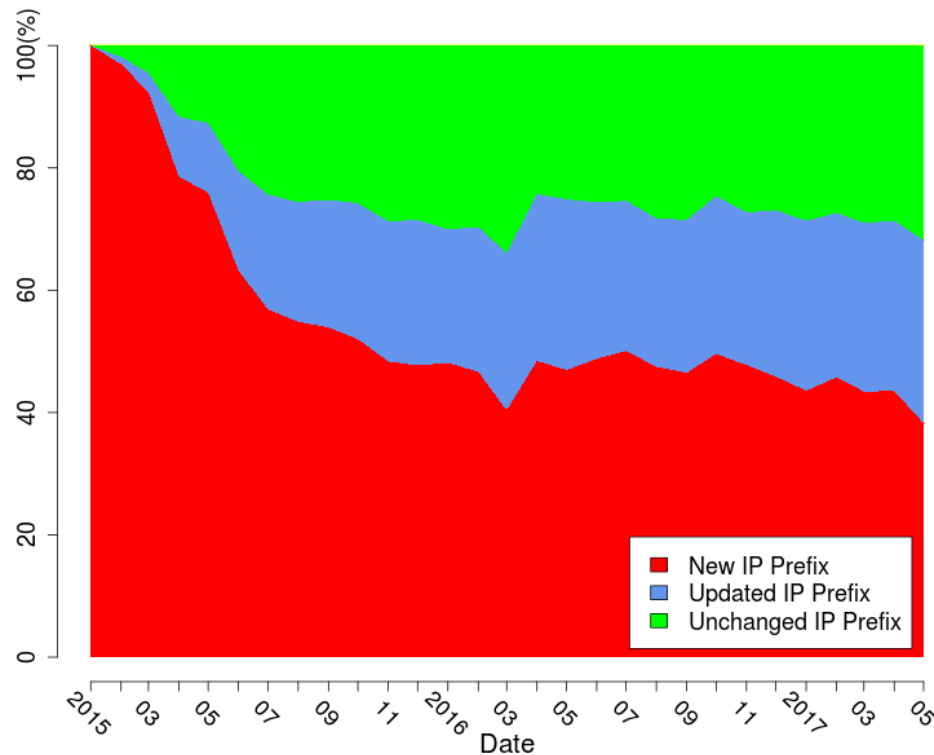
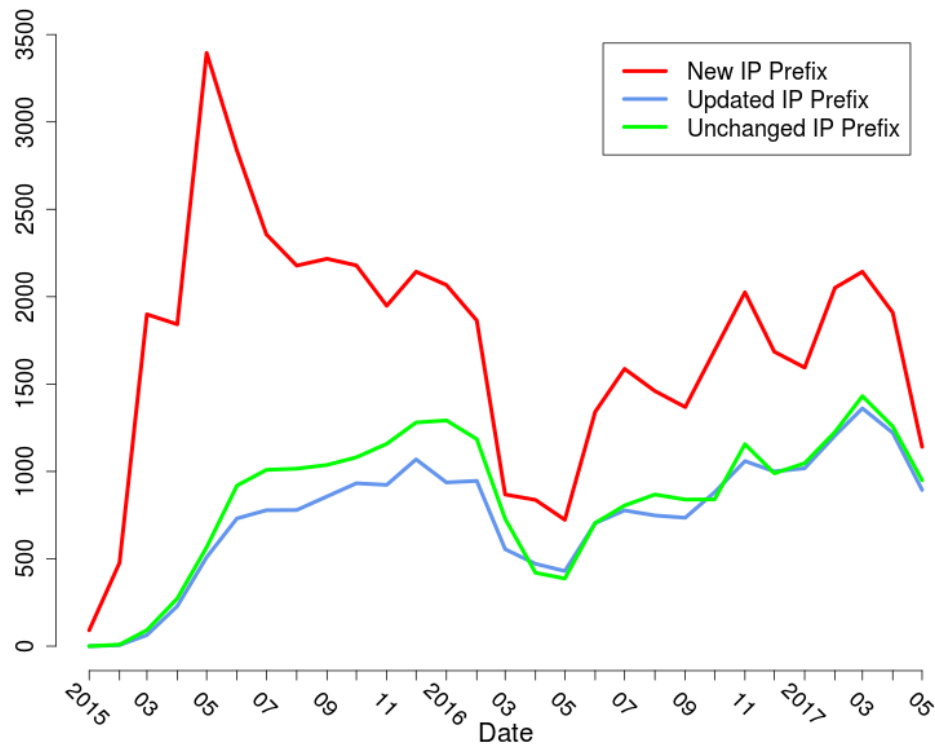


Province/District-level IP Prefix (/26)

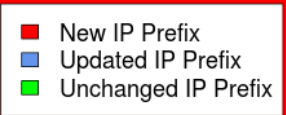
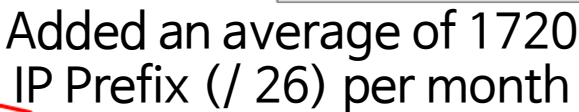
Metropolitan city or Province	Province-level IP prefix # (A)	District-level IP prefix # (B)	B/A (%)
Gyeonggi	8584	4434	51.65
Seoul		6490	30.08
Busan		2498	67.33
Gyeongnam		993	58.65
Incheon		76	71.85
Gyeongbuk			59.32
Daegu			62.92
Chungnam			49.95
Jeonbuk		8	61.96
Jeonnam		37	28.54
Chungbuk		496	60.86
Daejeon		930	63.05
Gwangju			61.55
Gangwon	686		48.25
Ulsan	939	434	46.22
Jeju	346	10	2.89
Total	49916	22862	45.80

Can find 45% /26 IP prefixes
in district name
within metropolitan city
or province

New/Updated IP Prefix (/26) ~ 29 months



100



Limitation

- Lack of ground-truth: Can not calculate the exact criteria of IP geolocation
 - Verification by comparing to commercial database
- Data volume limits : Less data accumulated in Ruliweb
- Limitations of application: Few community reveal IP address and location
 - Policies that do not apply in other online used markets (or community)
 - Difficult to use in mobile apps: NAT

Conclusion

- Create IP geolocation database using the Korea console gaming community data
- Although only 2.25 % of the Korean IP blocks can be covered, 40% are mapped to areas other than the metropolitan area and have extensive coverage.
- Compared with commercial IP geolocation database, 40% /26 prefix is more than 50km difference
- Need to develop services using enhanced IP geolocation performance



Thank you.

Source Code:

<https://github.com/munhyunsu/UsedMarketAnalysis>