How Big Data Will Revolutionize Government Policymaking?

Setia Pramana

Politeknik Statistika STIS

Analisis Big Data dengan R | Pusdiklat BPS

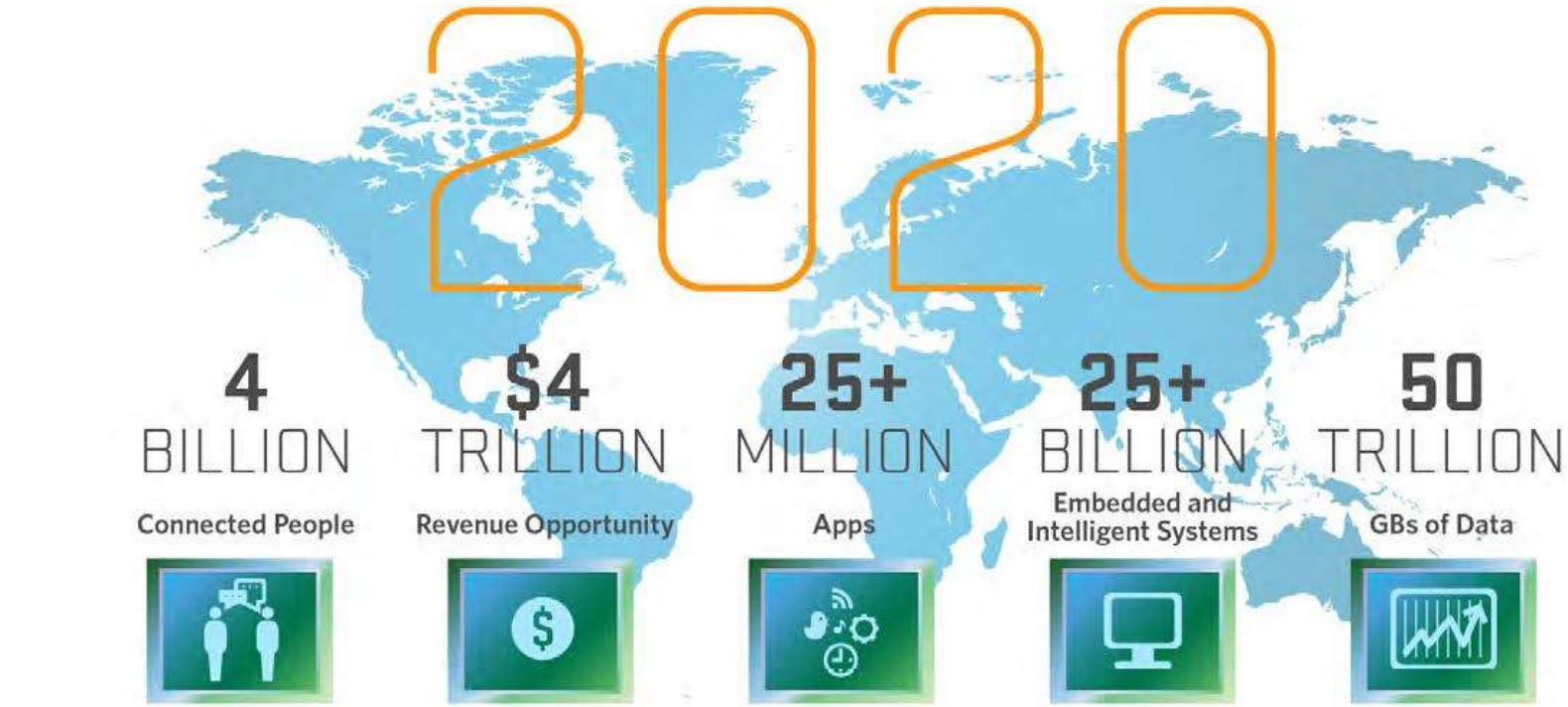
Better Data, Better Government

- Quality and timely data are vital for enabling governments, international organizations, civil society, private sector and the general public to make informed decisions
- **Evidence based policy making**
- Quality of Statistics:
  - Accuracy
  - Relevance
  - Timeliness
  - Accessibility
  - Coherence
  - Interpretability

Data Explosion

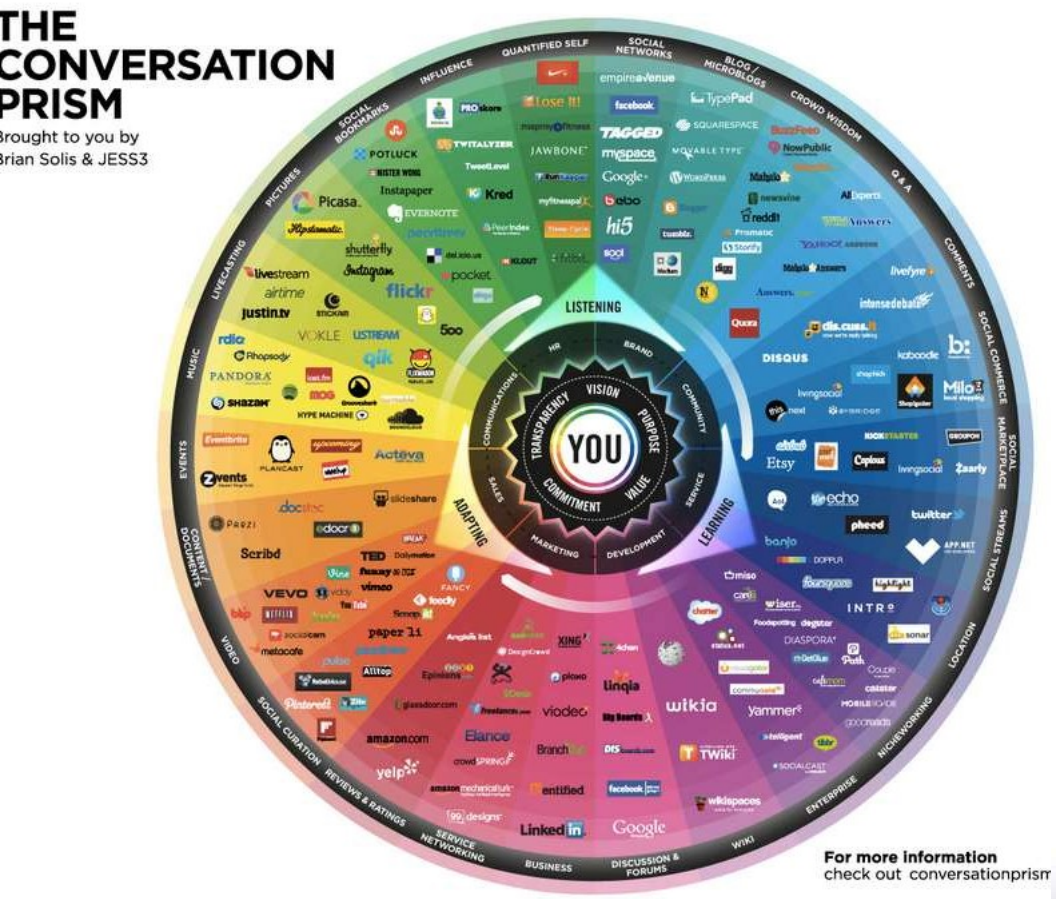
- Interactions of billions of people using computers, GPS devices, cell phones, and medical devices.
- online or mobile financial transactions, social media traffic, and GPS coordinates.
- “In the next five years, we’ll generate more data as humankind than we generated in the previous 5,000 years”. Eron Kelly, GM Microsoft

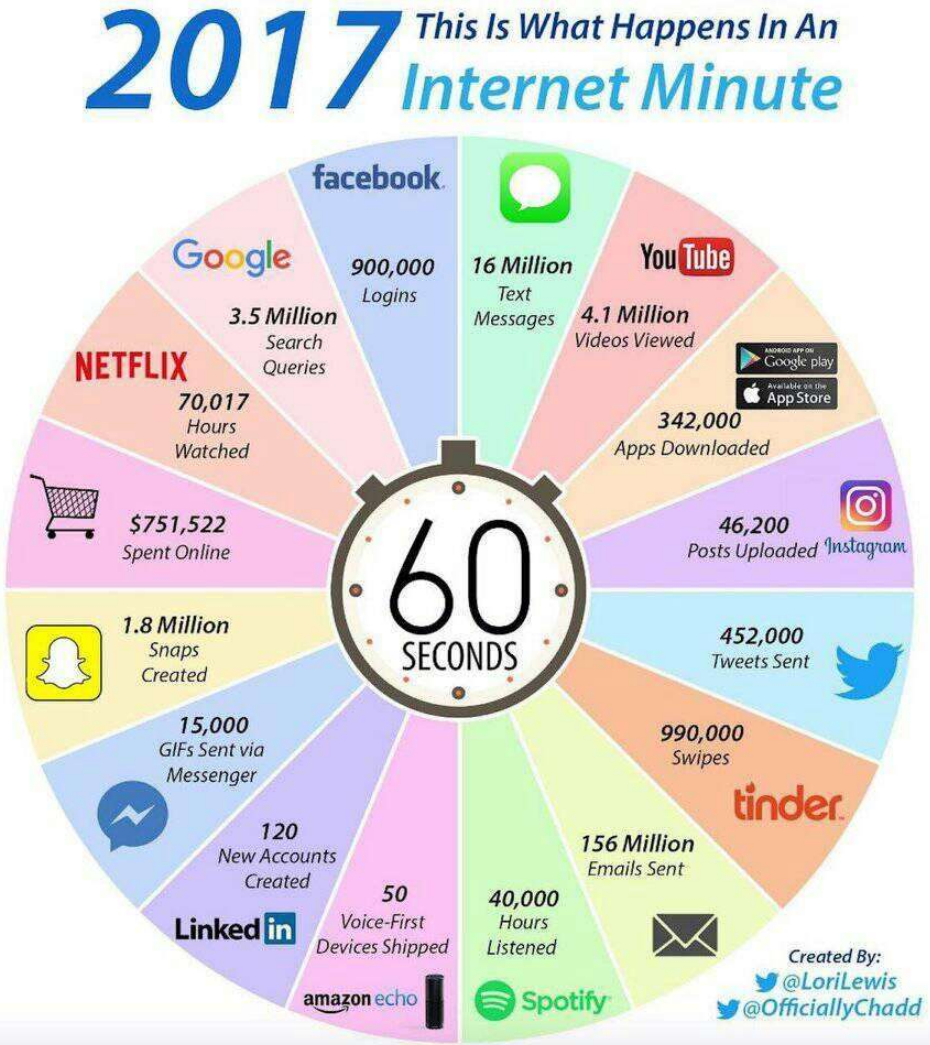
Data Explosion



Source: Mario Morales, IDC

What we “produce”?





brand24
Pemantauan Internet

JAKARTA
THE SOCIAL MEDIA CAPITAL OF THE WORLD

Social media have already become an important mean of **worldwide communication** and still gain on popularity. Platforms like Facebook or Twitter not only changed how people communicate, but also **widen perspectives** on building relationship between a company and its potential customers.

Indonesia, especially its capital city **Jakarta**, tops social media rankings. This country is a significant spot on the map of world's **social media activity**. Indonesians love posting and tweeting, and their fascination seems to spread rapidly.

FACEBOOK

USERS

1. BANGKOK	12 797 500
2. JAKARTA	11 658 760
3. SAO PAULO	8 791 700

THE MOST POPULAR INDONESIAN FANPAGES (BY FANS)

1. OPERA VAN JAVA (TRANS 7)	7 643 866
2. MARIO TEGUH	7 360 034
3. DAHSYAT	7 152 604

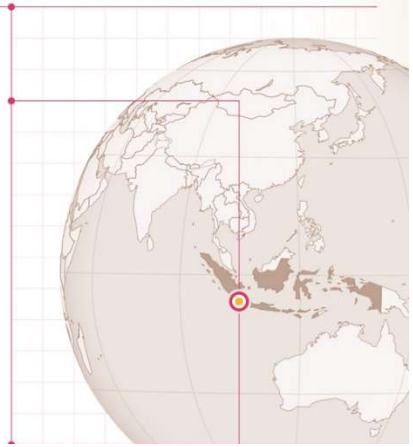
TWITTER

USERS

1. USA	143 000 000
2. BRASIL	41 000 000
3. JAPAN	34 000 000
4. U.K.	32 000 000
5. INDONESIA	29 000 000

JAKARTIAN TWITTER USERS POST **2,4%** OF THE **10.6 BILLION** TWEETS WORLDWIDE !

1. JAKARTA 2,4% 2. TOKYO 2,3% 3. LONDON 2%



THE MOST POPULAR TWITTER ACCOUNTS OF INDONESIAN CELEBS

1. AGNES MONICA / @AGNEZMO	6 781 477
2. SHERINA MUNAF / @SHERINAMUNAF	5 338 298
3. RAMALAN INDONESIA / @TWEETRAMALAN	4 729 582

THE MOST POPULAR TWITTER ACCOUNTS OF INDONESIAN BRANDS

1. CINEMA21 / @CINEMA21	3 420 828
2. TRANS7 / @TRANS7	789 630
3. MAICIHV / @INFOMAICIH	363 691

YOUTUBE

THE MOST POPULAR INDONESIAN YOUTUBE CHANNELS (BY VIEWS):

1. MALESBANGET.COM / MALESBANGETDOTCOM	4 141 654
2. GADIS GERY / FERRY153	2 918 751
3. POCARI SWEAT INDONESIA / POCARIGOION	2 851 761

LINKEDIN

TOTAL USERS FROM INDONESIA

1 387 300

THE MOST POPULAR ACCOUNTS OF INDONESIAN COMPANIES

1. MEDCO E&P INDONESIA	8 097
2. TELKOMSEL	7 263
3. PERTAMINA	6 668

LINKEDIN USERS BY AGE

25-34 YEARS OLD	39%
18-24 YEARS OLD	38%
35-54 YEARS OLD	21%
OTHER	2%

The following statistics based on the data come from **Socialbakers.com** and **Mediabistro.com** and were being collected from the end of 2012 till March 2013. This infographic is provided by **Brand24.co.id** - the most effective tool for media monitoring in the Internet.

brand24 Pemantauan Internet | socialbakers | mediabistro.com | JOY INTERMEDIA

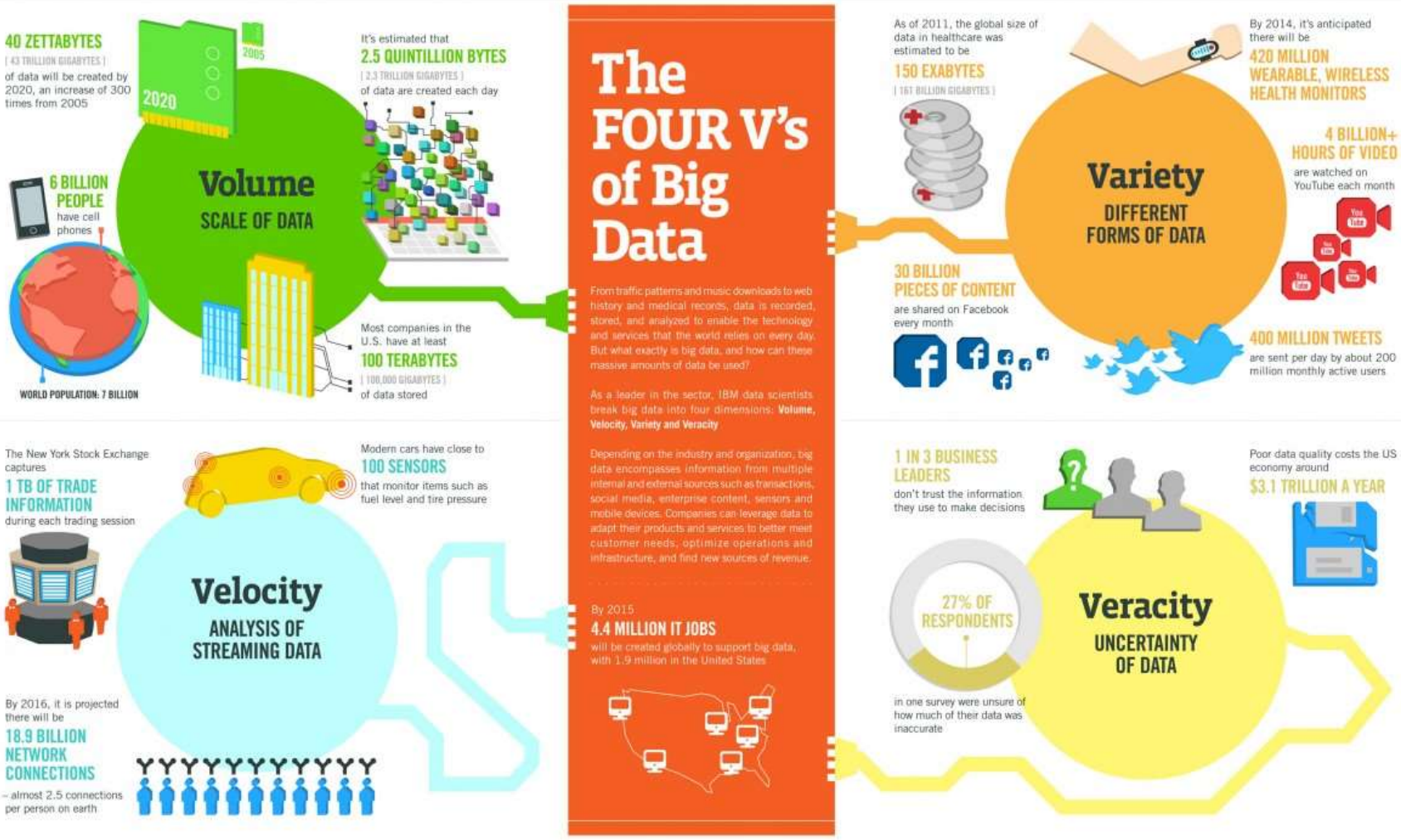
BRAND24
www.brand24.co.id
facebook.com/Brand24.co.id

SOCIALBAKERS
www.socialbakers.com

MEDIABISTRO
www.mediabistro.com/alltwitter

JOY INTERMEDIA
www.joyintermedia.com
facebook.com/JoyIntermediaIndonesia

Internet of Things


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The Four V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, GAS

Taxonomy BigData Sources

- **Exhaust data**: Passively collected data from people's use of digital services such as mobile phones, financial transactions or web searches.
- **Sensing data**: Actively collected data from sensors, e.g. in smart cities or from wearables and also through remote sensing and satellite images.
- **Digital content**: Open web content actively produced by people such as social media interactions, news articles, blogs or job postings. Unlike exhaust and sensing data is digital content intentionally edited by somebody, i.e. subjective or even deceptive, depending on the intentions of the author.

Letouzé (Data-Pop Alliance, 2015)

Data Sources

Exhaust data	Mobile phone data
	Financial transactions
	Online search and access logs
	Citizen card
	Postal data
Sensing data	Satellite and UAV imagery
	Sensors in cities, transport and homes
	Sensors in nature, agriculture and water
	Wearable technology
	Biometric data
	Internet of Things (IoT)
Digital Content	Social media data
	Web scraping
	Participatory sensing / crowdsourcing
	Health records
	Radio content

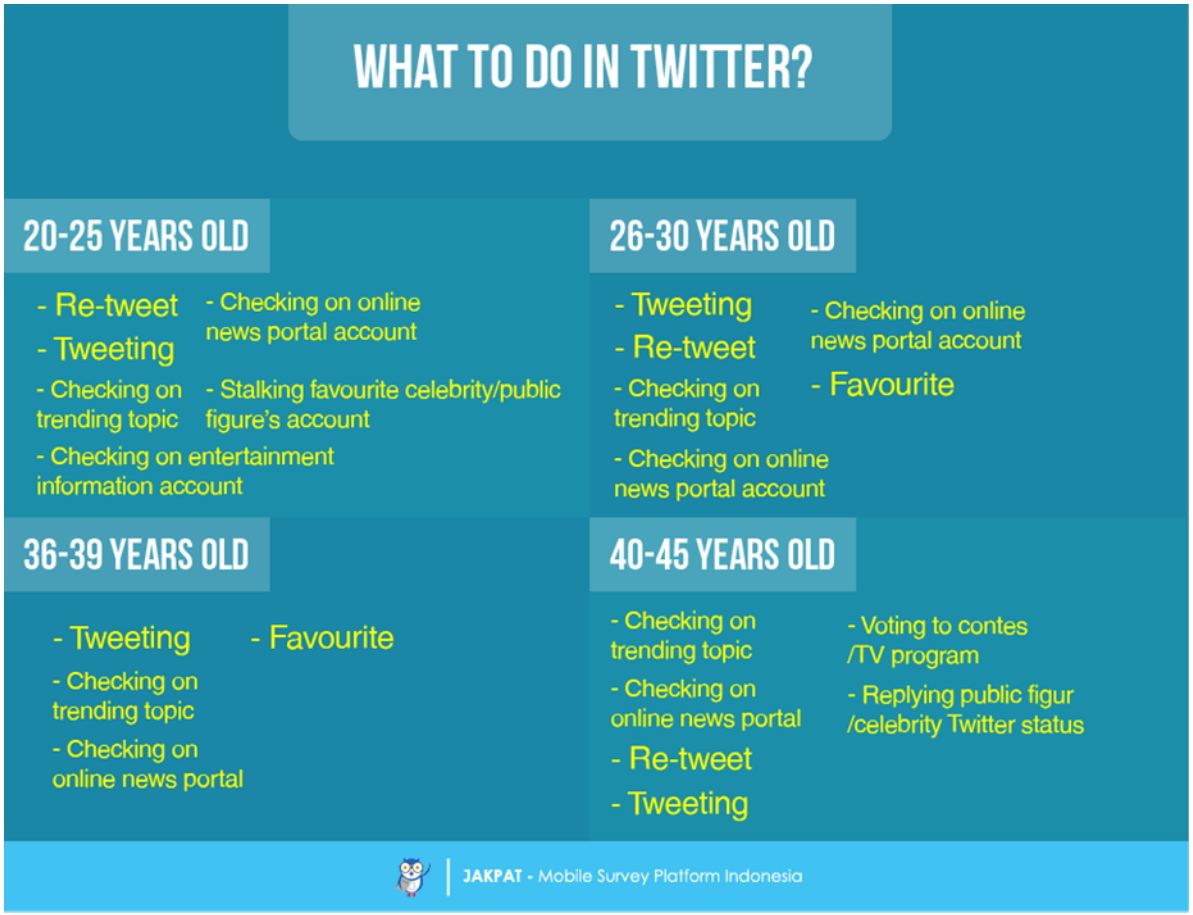
Measurement Revolution

Mobile Positioning Data

- Location of Mobile Devices
- Statistical indicators can be generated:
  - The number of residences geographically distributed according to available accuracy;
  - The number of workplace, school, secondary home, and other regular locations;
  - Internal migration based on the change of the residences within the country;
  - Change of workplace over time;
  - Cross-border migration based on the regular travels between different countries;
  - Population grid statistics (1 km2);
  - Temporary population statistics
  - Assessing temporary population (hourly, daily, weekly, monthly, etc.);

Twitter

- Trending Topics



Crowdsourcing

- The process of getting work, funding or information, usually online, from a crowd of people.
- The word **Crowdsourcing** is a combination of Crowd & Outsourcing

Official Statistics vs. Big Data

Official Statistics	Big Data
1. Structured and planned product	1. Largely unstructured unfiltered "data exhaust", i.e., by-product of digital products (transactions, web, social media, sensors)
2. Methodological and clear concepts	2. Poor analytics
3. Regulated	3. Unregulated
4. Macro-level but typically based on high volume primary data	4. Micro-level huge volume with high velocity (or frequency) and variety
5. High cost	5. Generally little, or no cost
6. Centralized; point in time	6. Distributed; real-time

Dr. Jose Ramon G. Albert, NSCB, Philippines

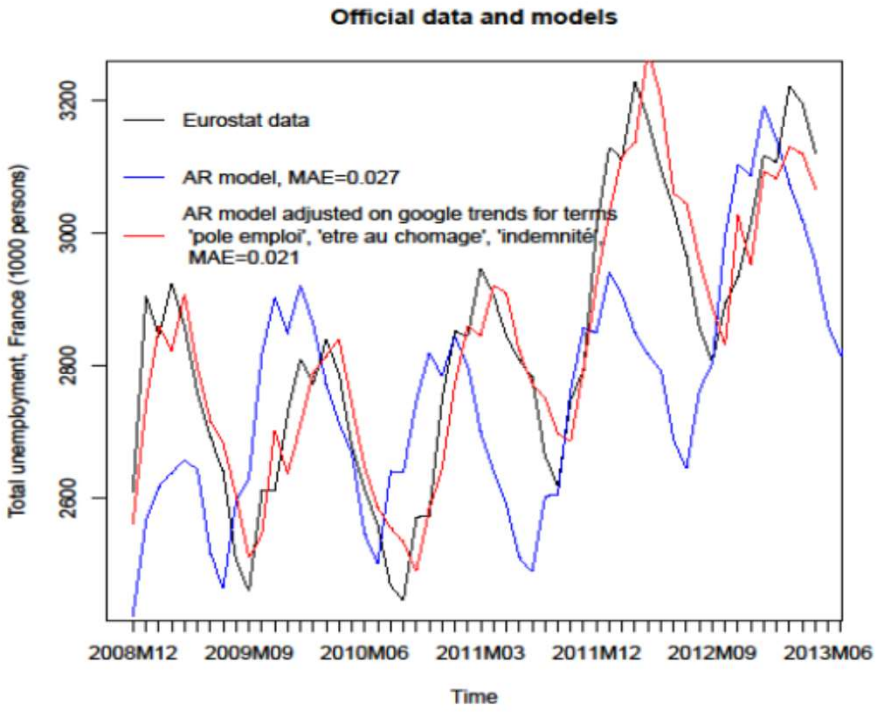
Big Data Roles for Official Statistics

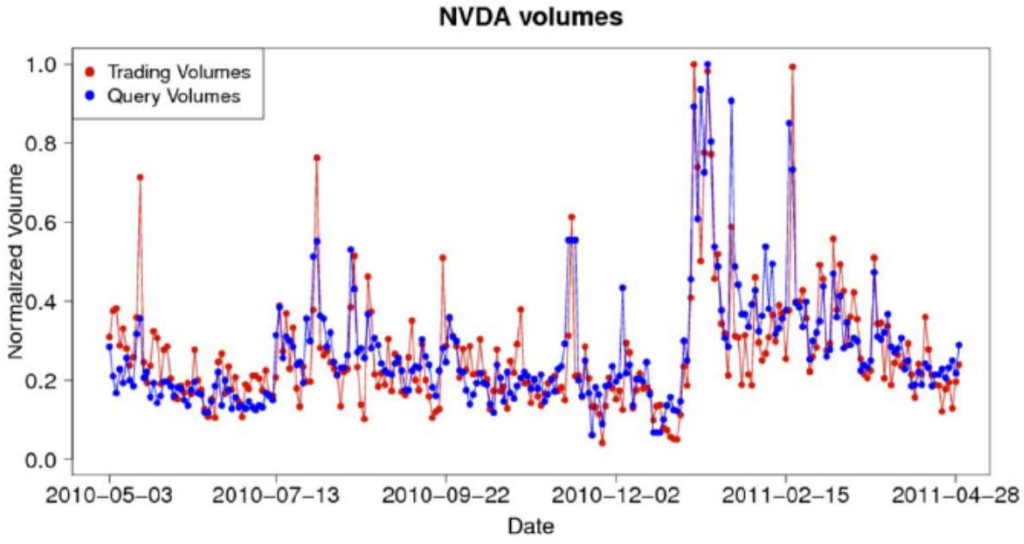
- Provide variables to help NSOs stratify better for sample surveys
- Improve sample survey estimates
- Help to compensate for nonresponse
- Help to check NSOs estimates
- Help to improve the frequency and timeliness of data releases
- Help to improve and provide more small-area estimates

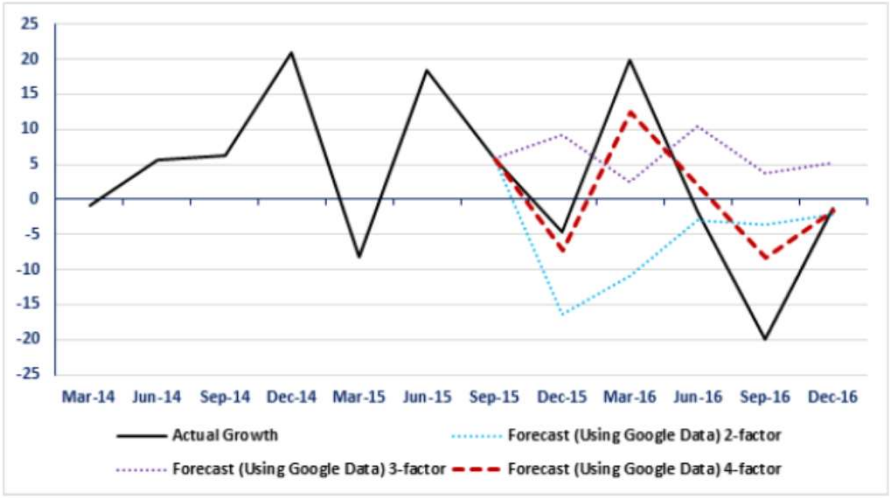
Cavan Capps and Tommy Wright, U.S. Census Bureau

Big Data for Public Policy: Examples

Google Trend







ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1]* Gabriel Cadamuro,[2] Robert On[3]

Accurate and
and economic
enabling new
sources of big
be used to inf
predicted attr
distribution of
composed of j
and household
and timely inf

Mapping poverty using mobile phone and satellite data

Mobile phone data could replace census questionnaires in the UK

The ONS used locations and timestamps to identify where people work and live.

Matt Brian, @m4tt
11.07.17 in Mobile

4 Comments

226 Shares

tor A. Alegana[1],

M. Iqbal[6],
drew J. Tatem[1,2,10]

g 44, Southampton, UK

RESEARCH ART

Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator

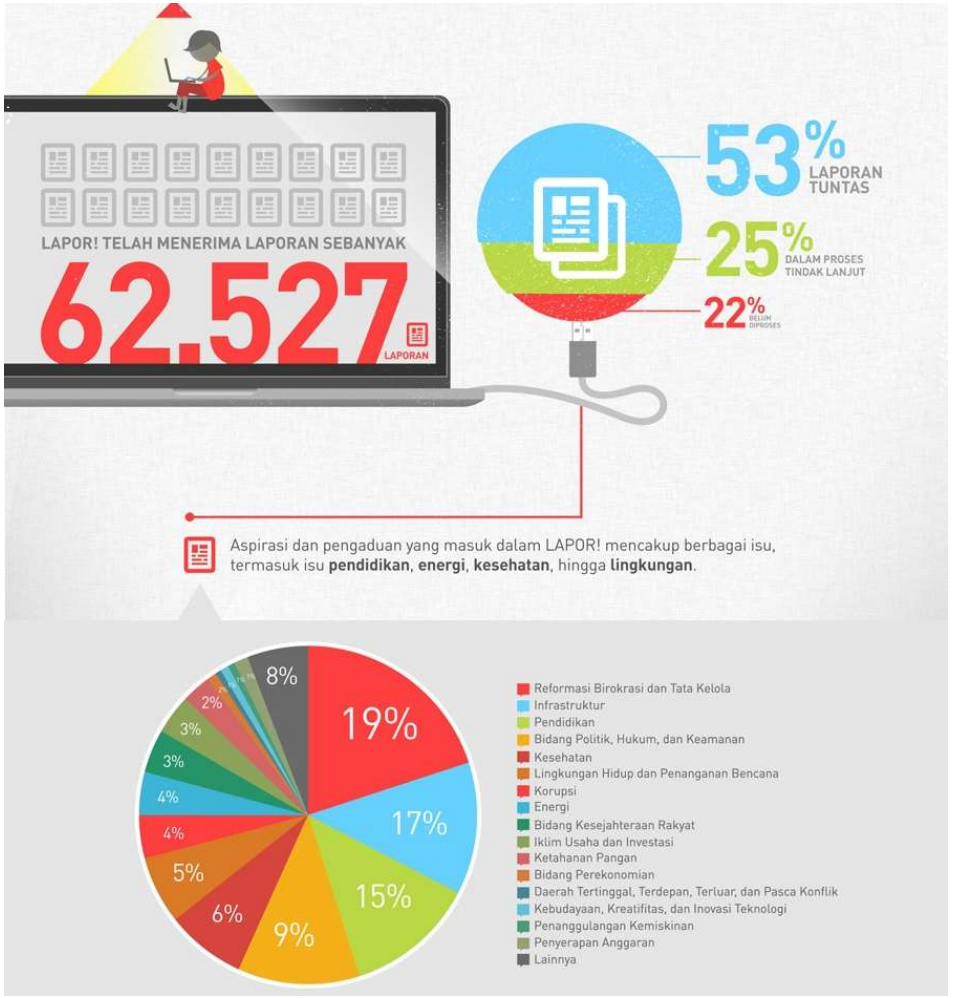
Sanja Šćepanović[1]*, Igor Mishkovski[2], Pan Hui[3], Jukka K. Nurminen[1], Antti Ylä-Jääski

1 Department of Computer Science, Aalto University, Helsinki, Finland, 2 Faculty of Computer Science a...
Engineering, University Ss. Cyril and Methodius, Skopje, Macedonia, 3 Department of Computer Science
and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

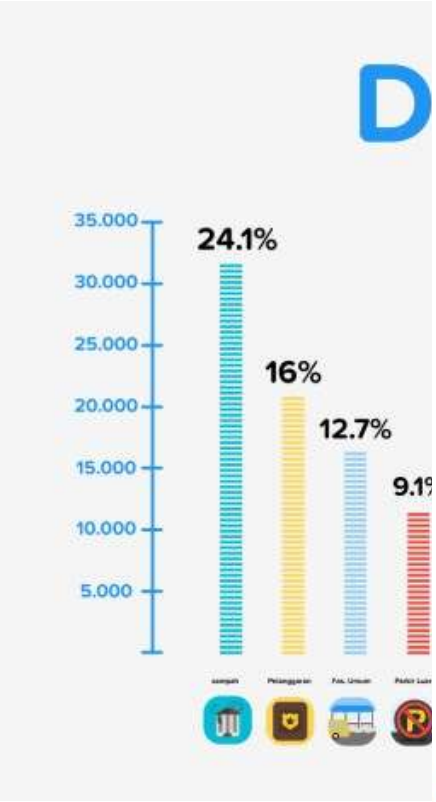
* sanja.scepanovic@aalto.fi

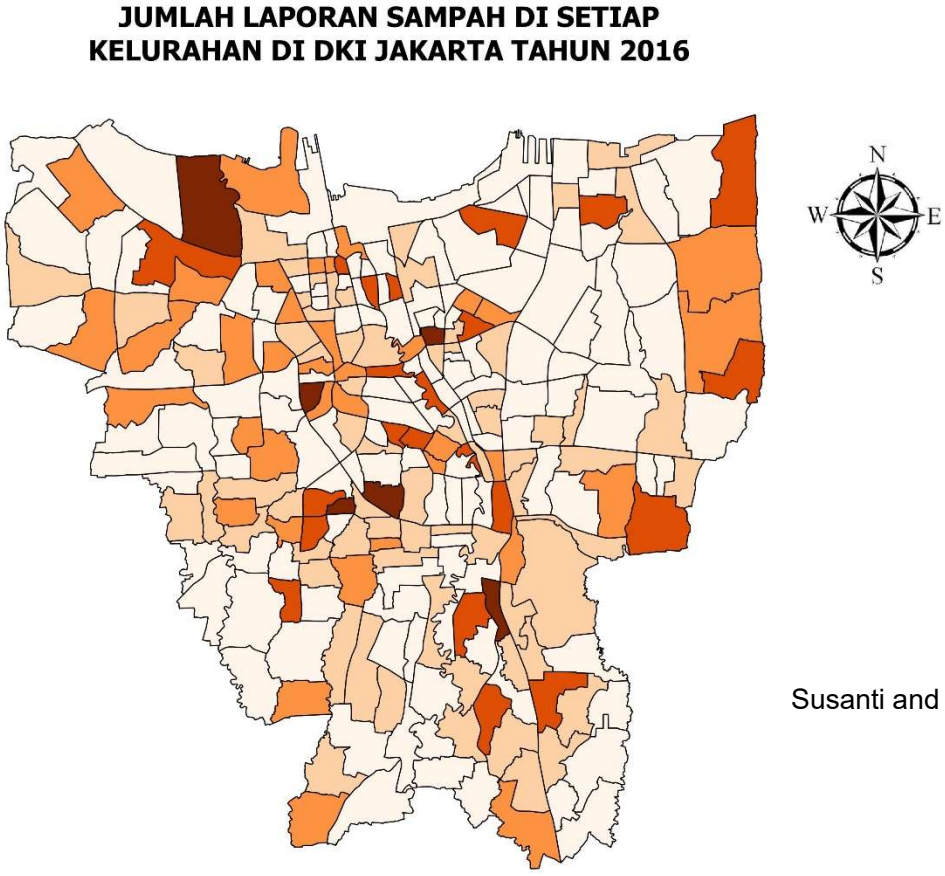
[4]School of Information, University of California, Berkeley, CA, USA
[5]Data Science Institute, Imperial College London, London, UK
[6]Grameenphone Ltd, Dhaka, Bangladesh
[7]Public Health Sciences, Karolinska Institute, Stockholm, Sweden
[8]College of Information System and Management, National University of Defense Technology,
Changsha, Hunan, People's Republic of China

Crowdsourcing



Crowdsourcing





Susanti and Pramana 2017

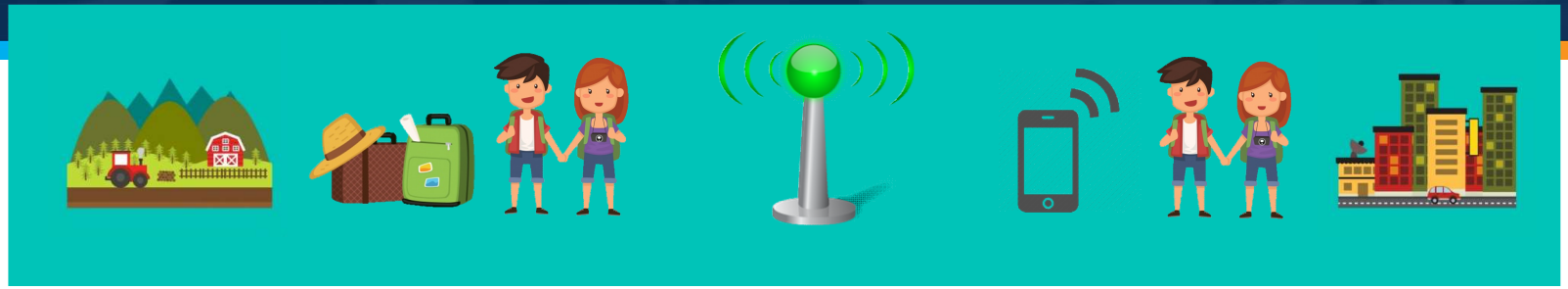
Web Crawling and Scraping

- Extract Information from Web
- Web Crawling is the process of locating information on World Wide Web(WWW), indexing all the words in a document, adding them to a database, then following all hyper links and indexes and adds that information also to the database
- Web scraping is the process of automatically requesting a web document and collecting information from it.

http://prowebscraping.com/web-scraping-vs-web-crawling/

STATISTICS INDONESIA

Initiations of Big Data Use for Official Statistics by Politeknik Statistika STIS Jakarta



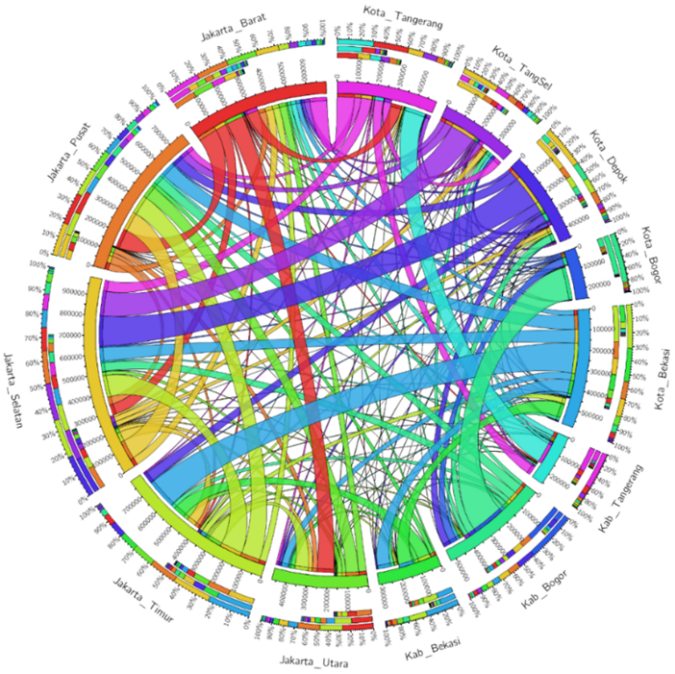
Study Cases

- **Twitter Geotag for Predicting Commuting Patterns**
- **Price Nowcasting using crowdsourcing, online Shops and Instagram**
- **Happiness Index using twitter**
- **Mobility Behavior Tracking**
- **Job Vacancy Monitoring**
- **Tourism Promotion on Social Media and Online Media**

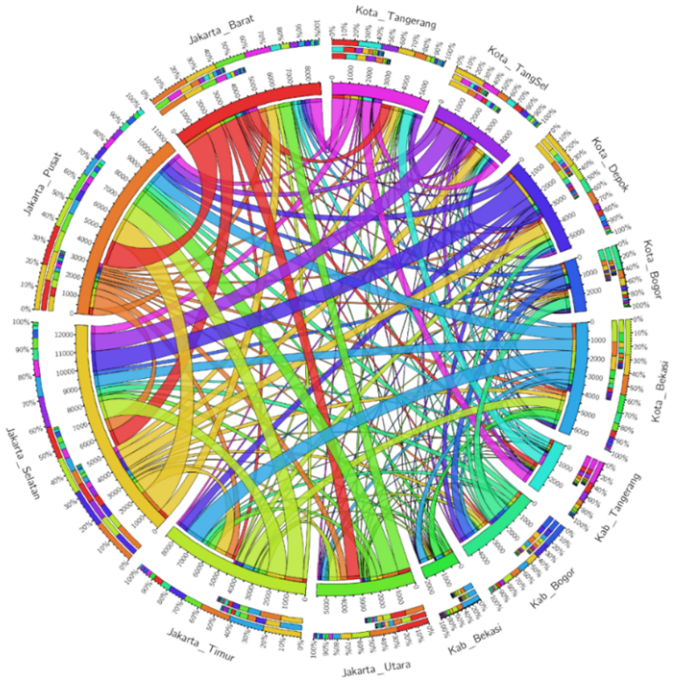
Twitter Geotag for Predicting Commuting Patterns

- Collaboration with Global Pulse Jakarta
- Official Data: Jabodetabek Commuter Survey 2014
- Twitters (Februari 2014)
- Origin: Location of most tweets
- Destination: 2nd most tweets location

Results



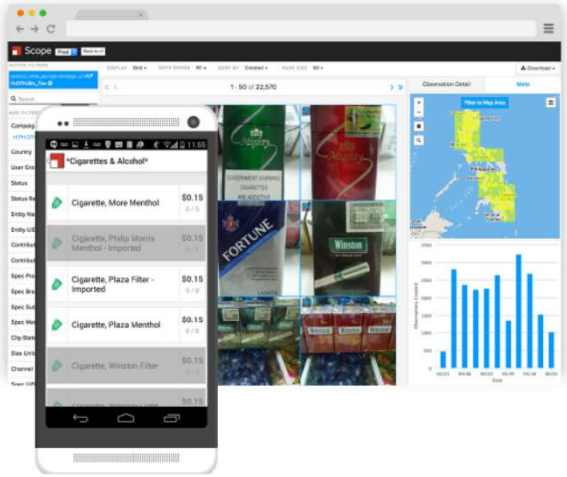
Commuter Survey



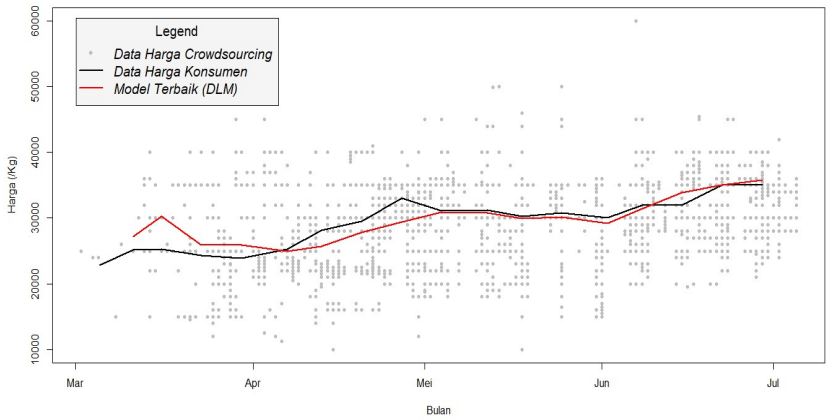
Twitter

Crowdsource for Food Prices Nowcasting

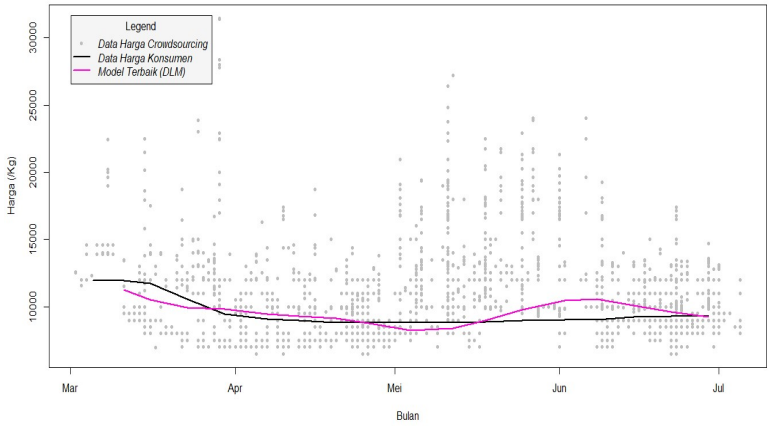
- Collaboration with Pulse Lab UN Jakarta
- Use crowdsourcing premise UN Food security project
- Locus: Kota Mataram, NTB
- Time: March– July 2015



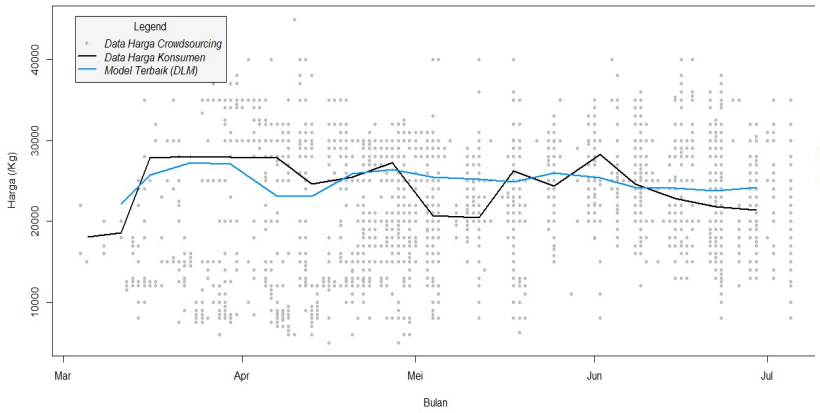
Plot Pemodelan Terbaik Harga Daging Ayam Ras



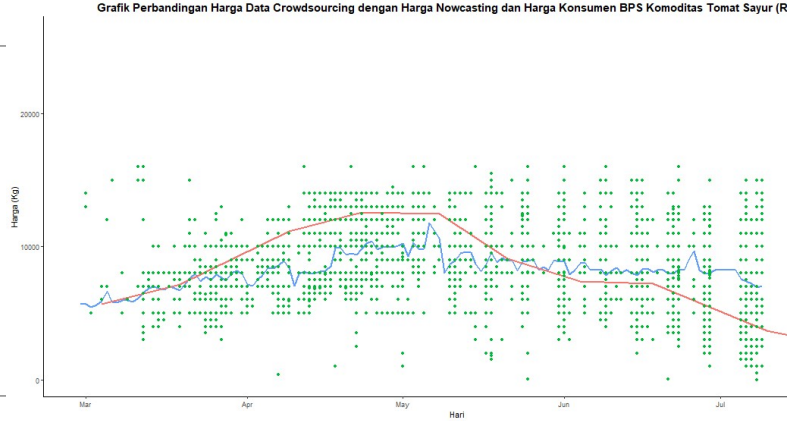
Plot Pemodelan Terbaik Harga Beras Kualitas Premium



Plot Pemodelan Terbaik Harga Bawang Merah

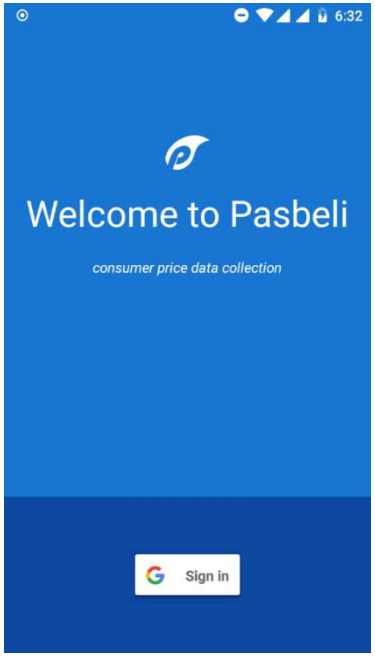


Grafik Perbandingan Harga Data Crowdsourcing dengan Harga Nowcasting dan Harga Konsumen BPS Komoditas Tomat Sayur (R



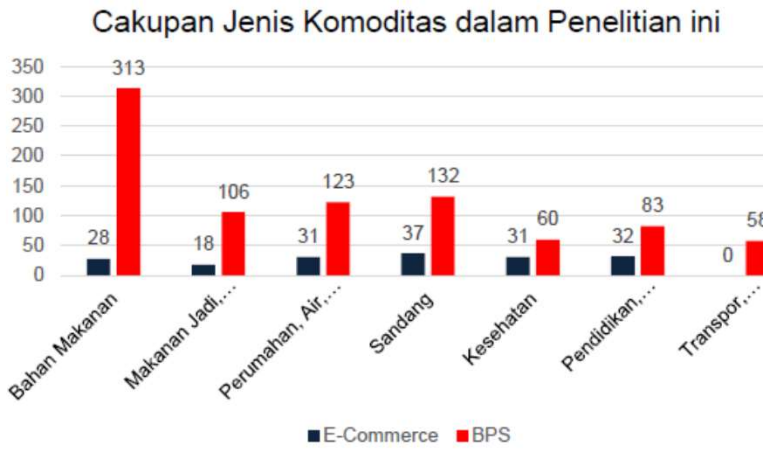
Crowdsourcing

Pasbeli Apps: STIS Apps



Web Scraping: Online Shops

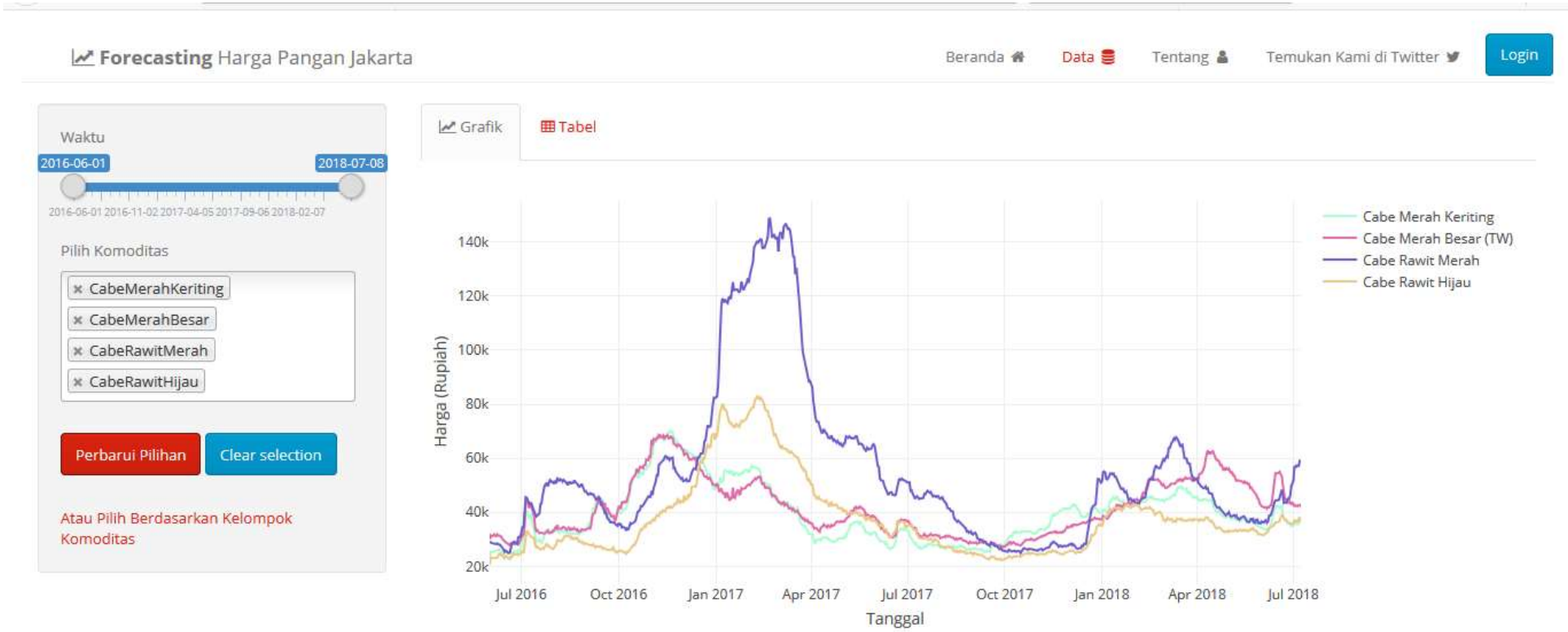
Online Shops	Total Commodities
Hypermart	52 products
KlikMart	75 products
Bhinneka	40 products
Elektronik City	17 products
Zalora	36 products
BerryBenka	25 products
Mothercare	2 products
Babyzania	4 products
Apotek Century	10 products
Pusat Kosmetik	5 products
Sephora	2 products
Stationary	6 products
Gramedia	3 products



Online Shop Commodities Prices

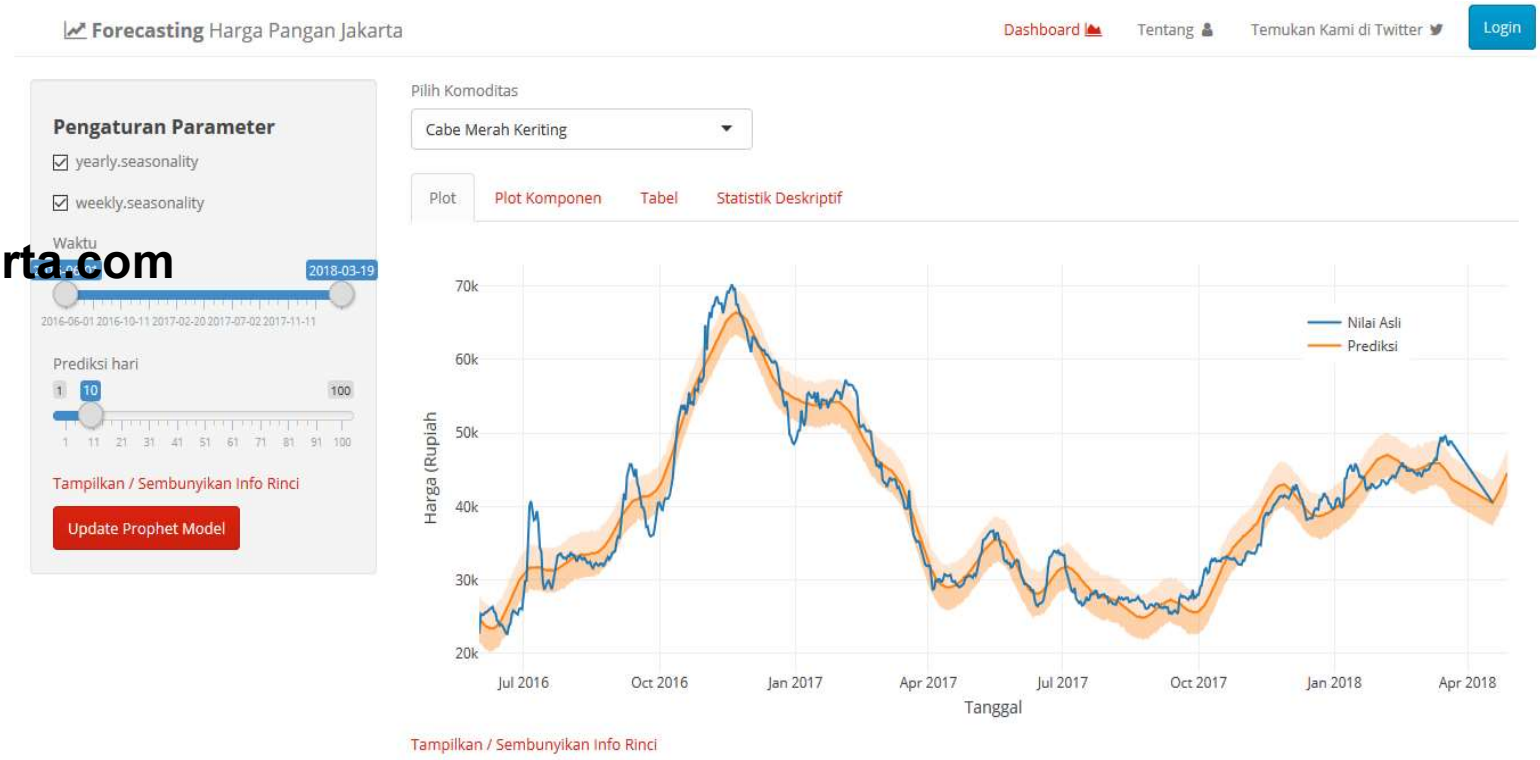
- Analysis is on progress
- Get the movement of consumer price
- Get the pattern of the changes of consumer price per commodity kind and per e-commerce
- Construct CPI by substituting the conventionally collected consumer price with e-commerce-based consumer price, then
- Comparing the survey-based CPI with e-commerce-based CPI

Traditional Market Price Nowcasting



Traditional Market Price Nowcasting

Source of Scraping:
http://infopanganjakarta.com



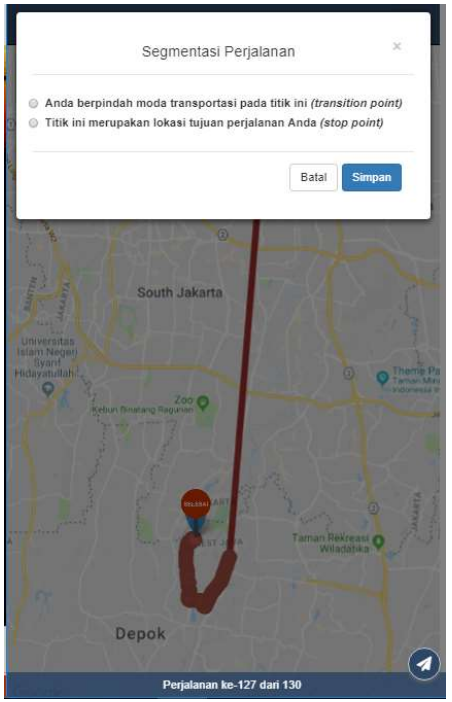
Mobility Behavior Tracking

- Meili Travel Diary
- Collaboration with JUTPI phase 2 JICA
- KTH Stockholm
- Tracking people movement, speed, route…

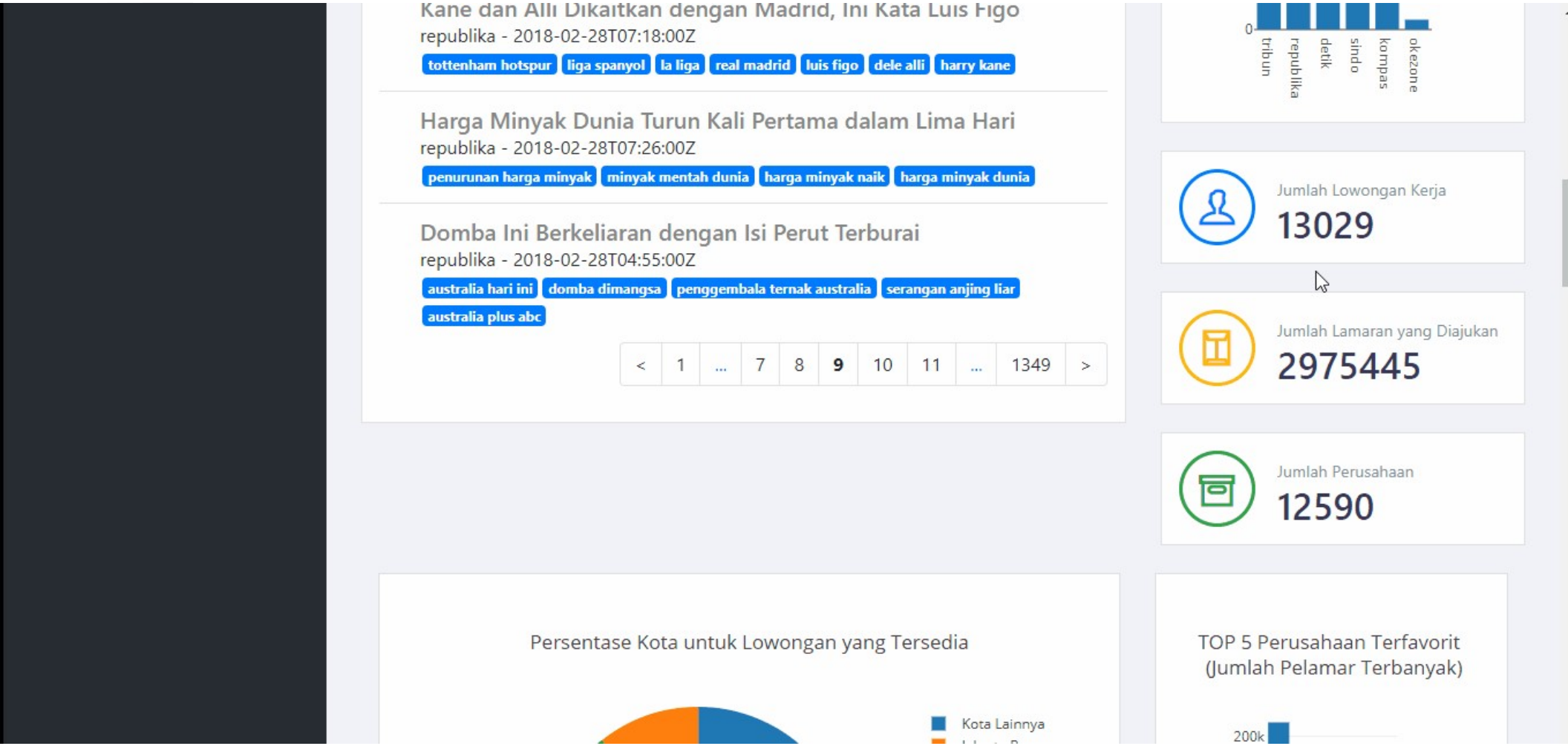
Mobility Behavior Tracking

- Record respondents travel behavior, including:
  - Trip routes
  - Destination (e.g.: Home, Office, Mall)
  - Trip activities (e.g.: Working, Shopping, Back to Home)
  - Transition Points (e.g.: Walking then using train)
  - First Level Transport modes (e.g.: Motorbike, Bus, Car)
  - Second Level Transport modes (e.g.: Commuter Line, Online Transport (Grab Bike/Car, Gojek))
  - Transport costs including: tariff, parking, and toll
- Forecasting respondents travel behavior (transport mode, activities) using machine learning algorithm

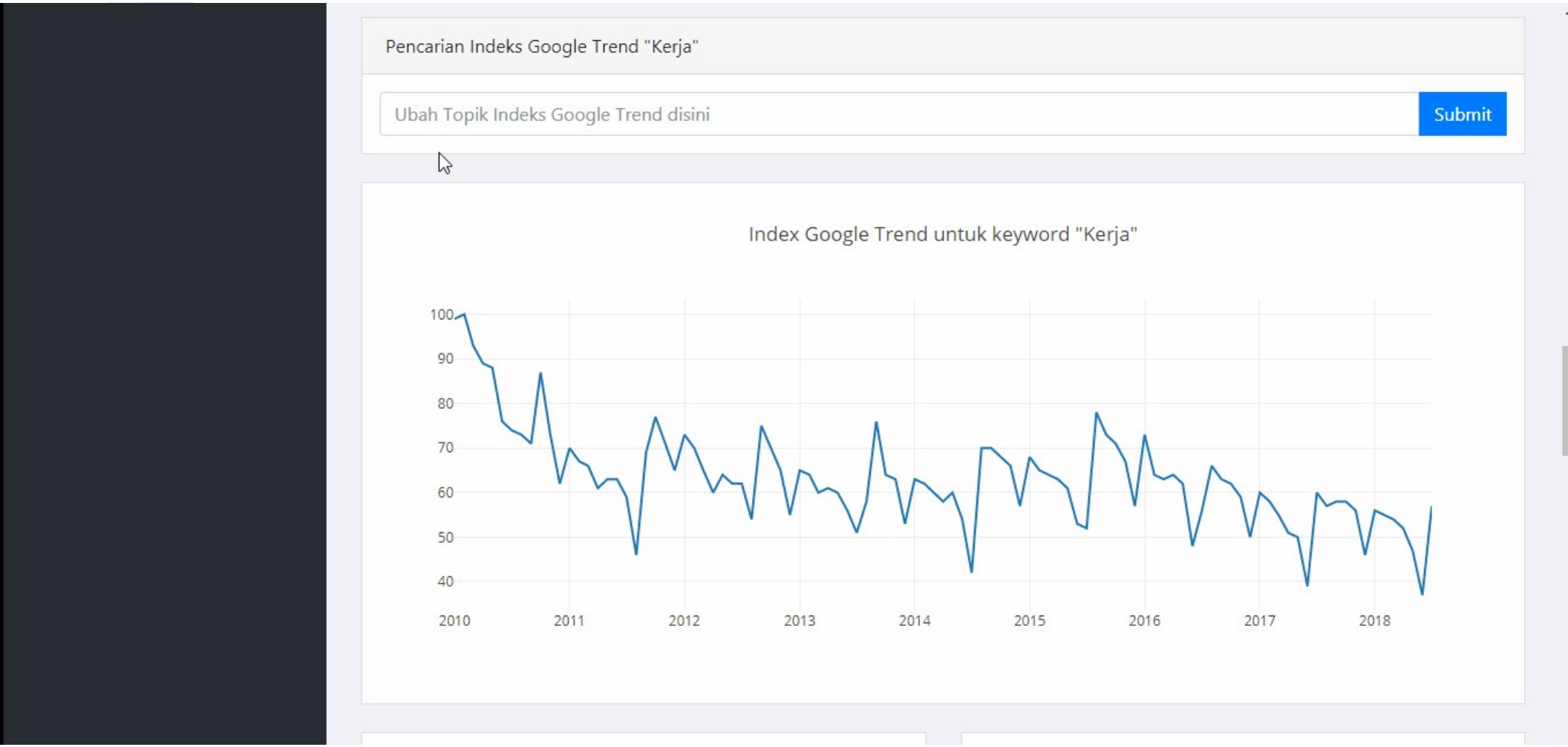
Mobility Behavior Tracking



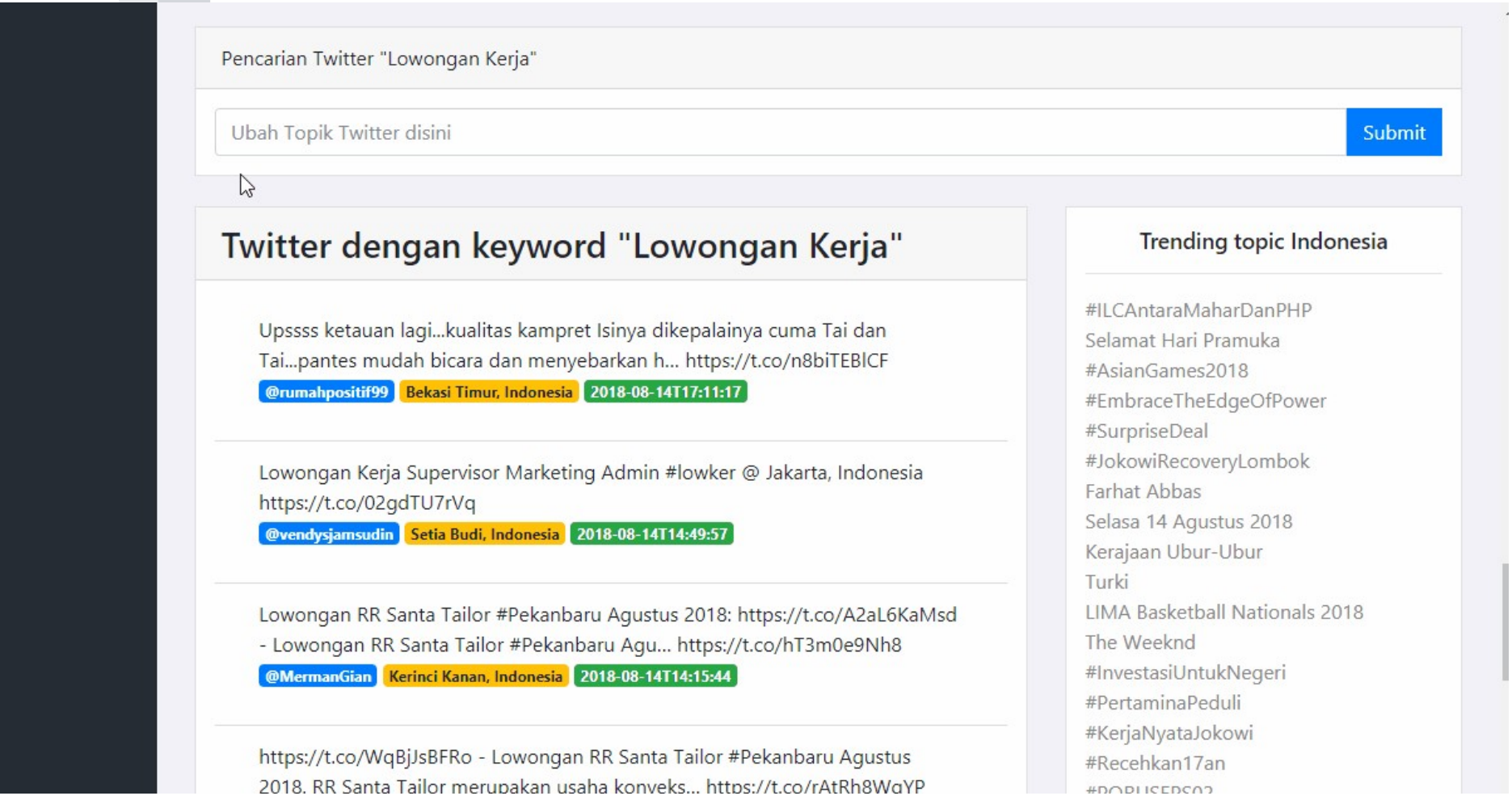
Job Vacancy Monitoring



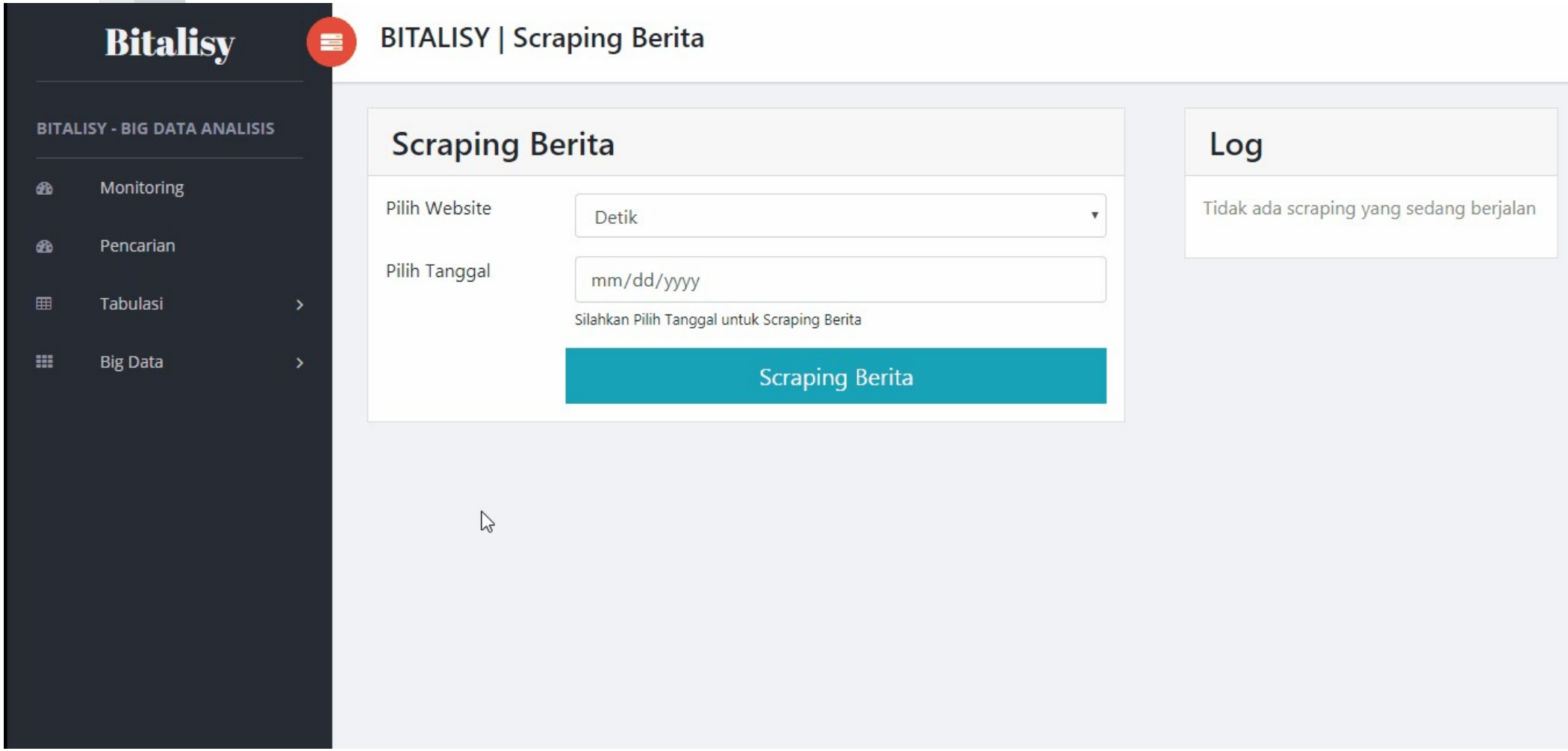
Google trend



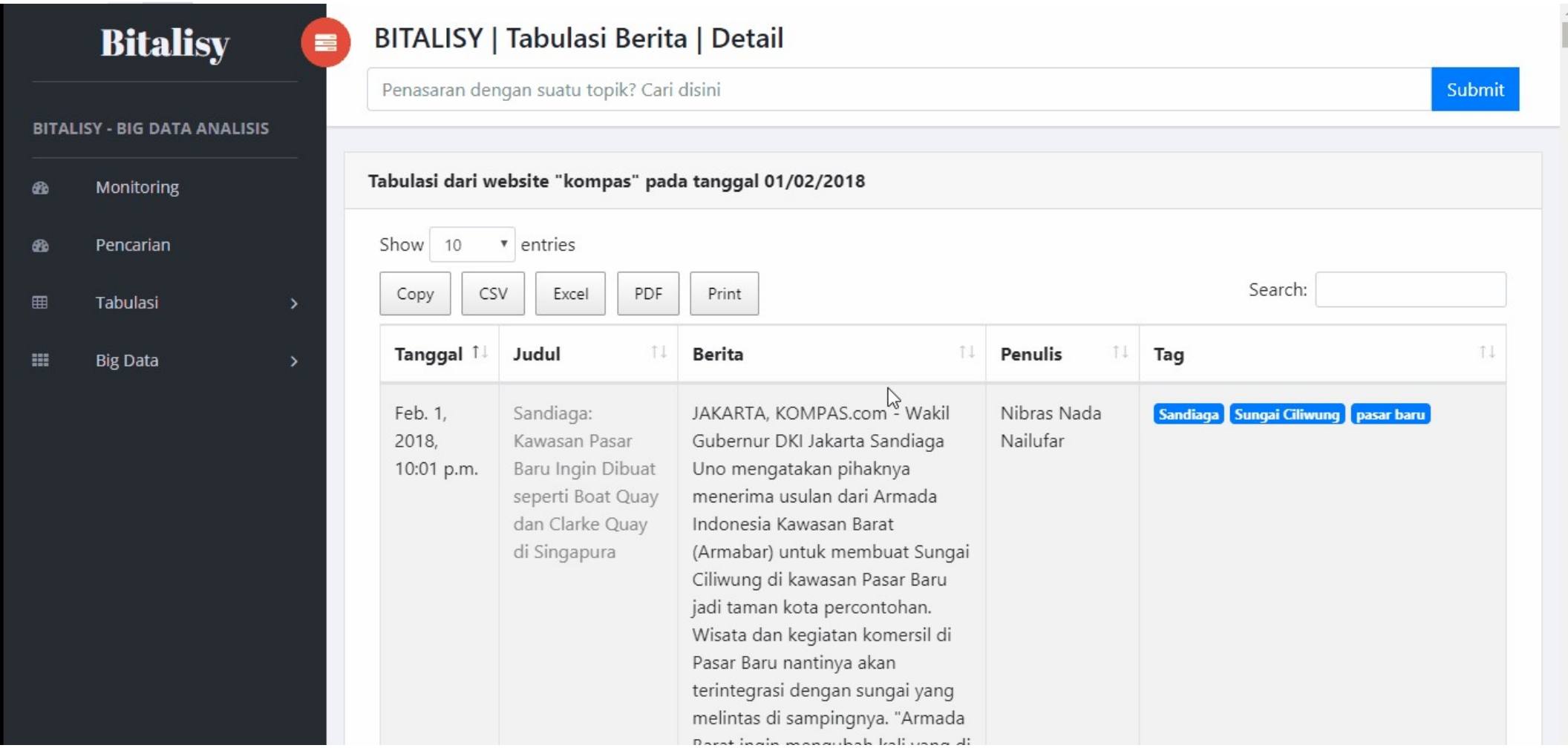
Job Vacancy Monitoring: Twitter



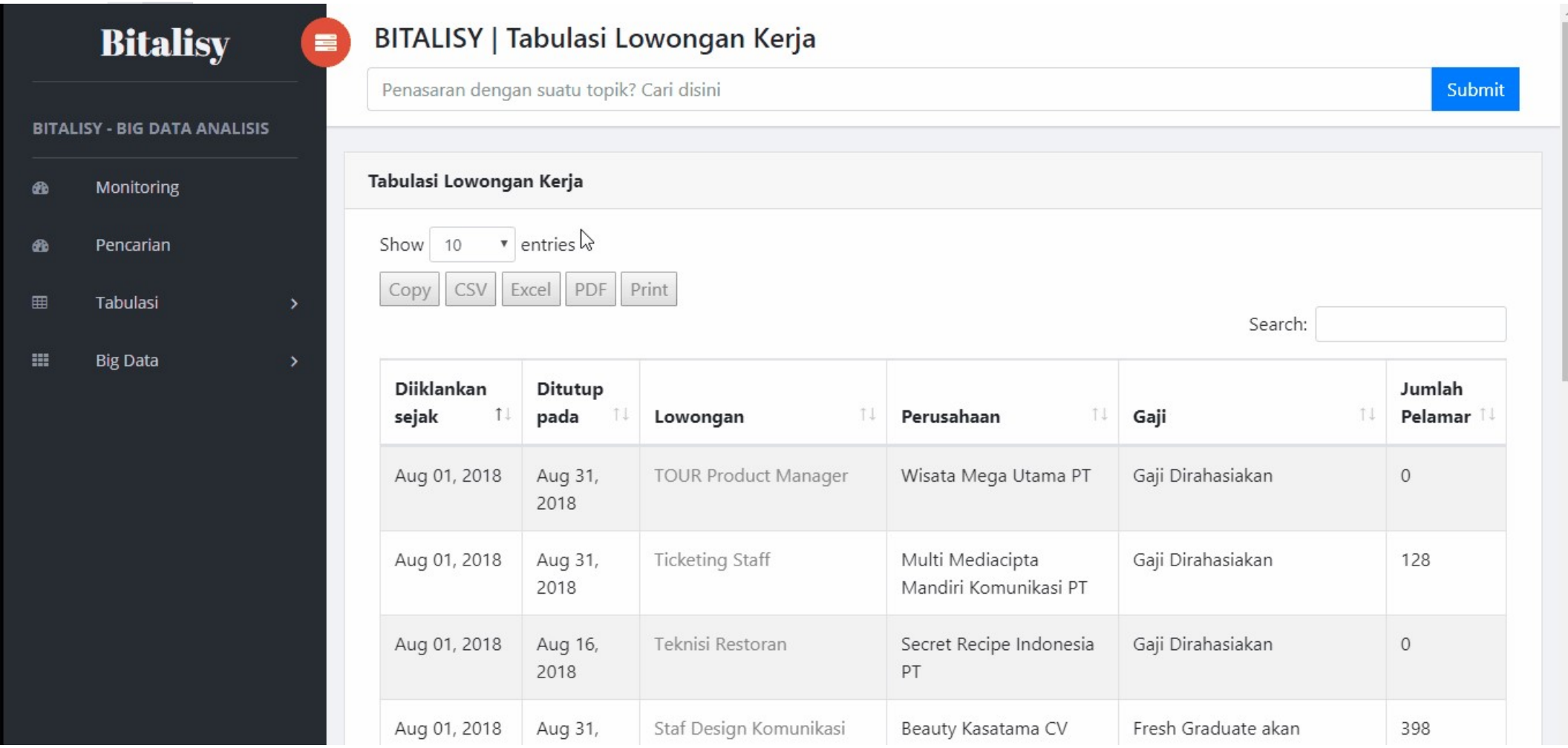
Scraping News



News Detail

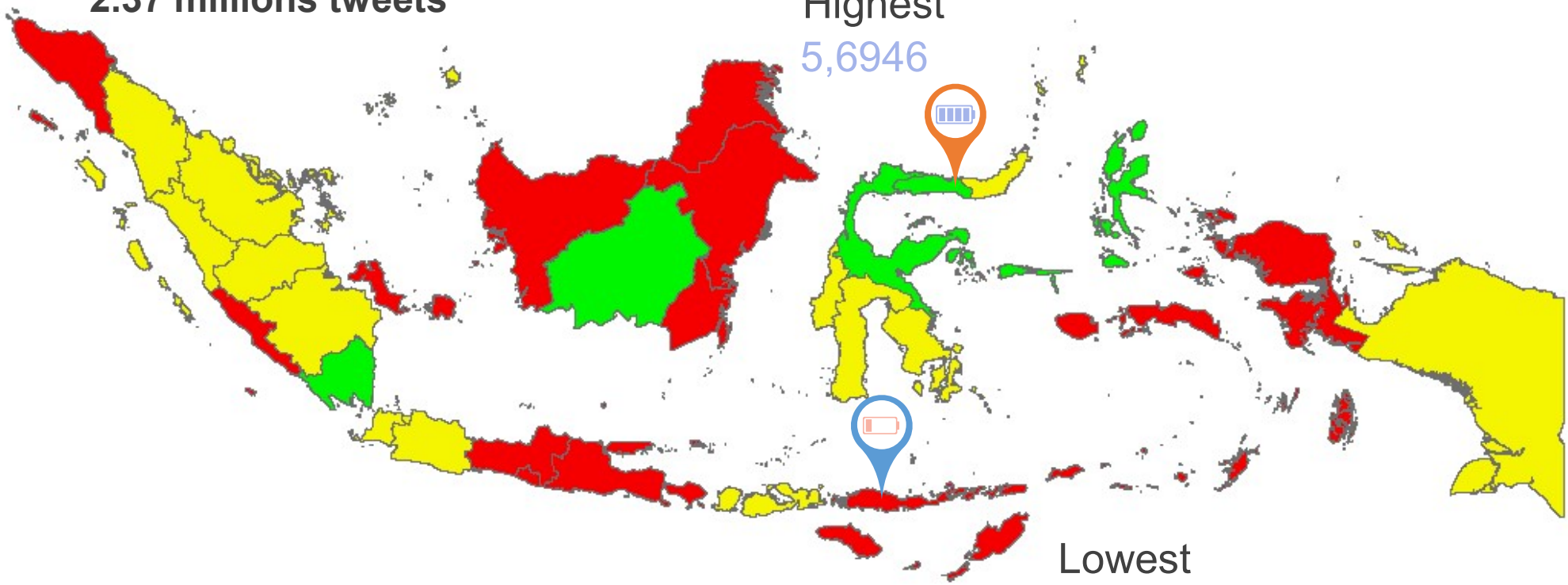

Bitalisy
BITALISY - BIG DATA ANALISIS
Monitoring
Pencarian
Tabulasi
Big Data
BITALISY | Tabulasi Berita | Detail
Penasaran dengan suatu topik? Cari disini
Submit
Tabulasi dari website "kompas" pada tanggal 01/02/2018
Show 10 entries
Copy
CSV
Excel
PDF
Print
Search:
Tanggal
Judul
Berita
Penulis
Tag
Feb. 1, 2018, 10:01 p.m.
Sandiaga: Kawasan Pasar Baru Ingin Dibuat seperti Boat Quay dan Clarke Quay di Singapura
JAKARTA, KOMPAS.com - Wakil Gubernur DKI Jakarta Sandiaga Uno mengatakan pihaknya menerima usulan dari Armada Indonesia Kawasan Barat (Armabar) untuk membuat Sungai Ciliwung di kawasan Pasar Baru jadi taman kota percontohan. Wisata dan kegiatan komersil di Pasar Baru nantinya akan terintegrasi dengan sungai yang melintas di sampingnya. "Armada
Nibras Nada Nailufar
Sandiaga
Sungai Ciliwung
pasar baru

Jab vacancy

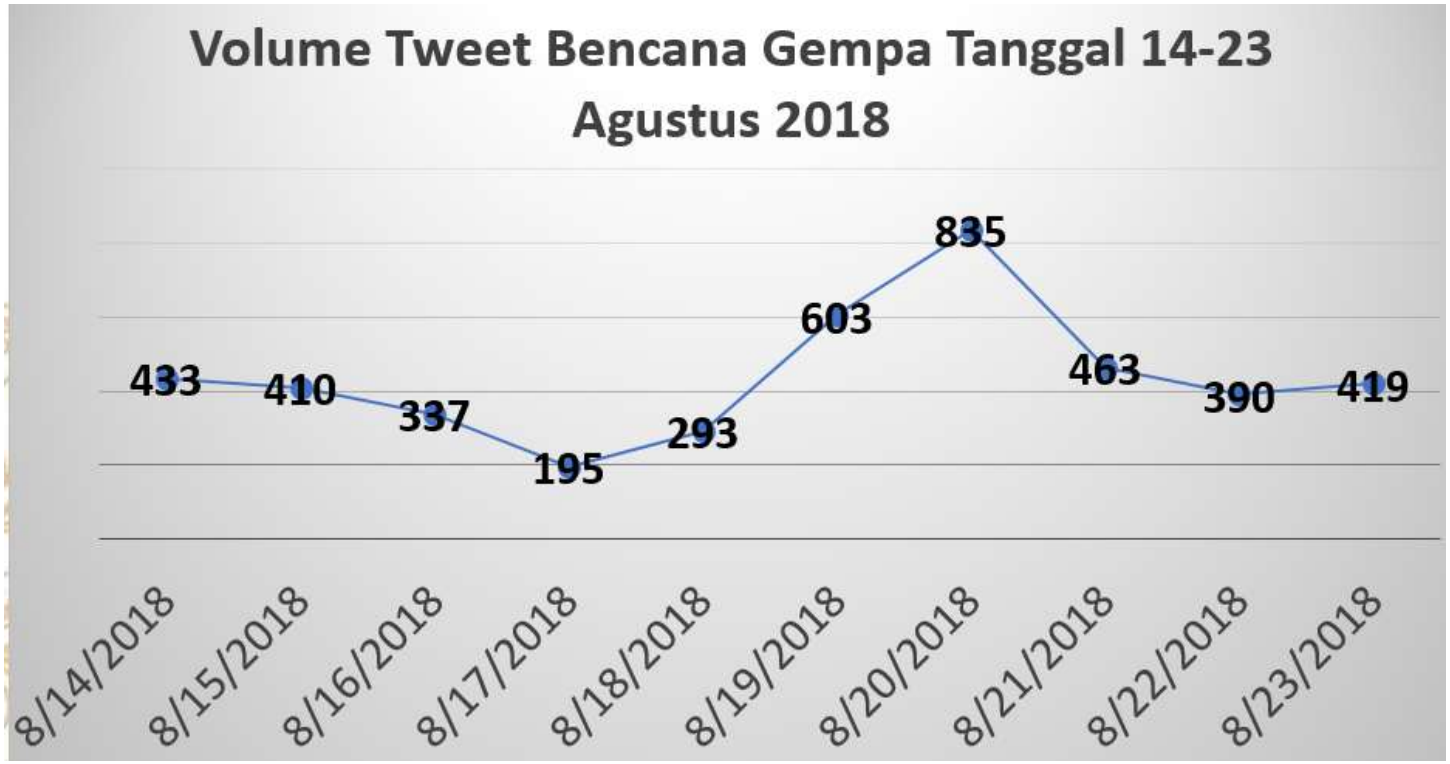

Bitalisy
BITALISY - BIG DATA ANALISIS
Monitoring
Pencarian
Tabulasi
Big Data
BITALISY | Tabulasi Lowongan Kerja
Penasaran dengan suatu topik? Cari disini
Submit
Tabulasi Lowongan Kerja
Show 10 entries
Copy CSV Excel PDF Print
Search:
Diiklankan sejak
Ditutup pada
Lowongan
Perusahaan
Gaji
Jumlah Pelamar
Aug 01, 2018
Aug 31, 2018
TOUR Product Manager
Wisata Mega Utama PT
Gaji Dirahasiakan
0
Aug 01, 2018
Aug 31, 2018
Ticketing Staff
Multi Mediacipta Mandiri Komunikasi PT
Gaji Dirahasiakan
128
Aug 01, 2018
Aug 16, 2018
Teknisi Restoran
Secret Recipe Indonesia PT
Gaji Dirahasiakan
0
Aug 01, 2018
Aug 31,
Staf Design Komunikasi
Beauty Kasatama CV
Fresh Graduate akan
398

Subjective Happiness Index: Twitter

Between April – July 2018,
2.37 millions tweets



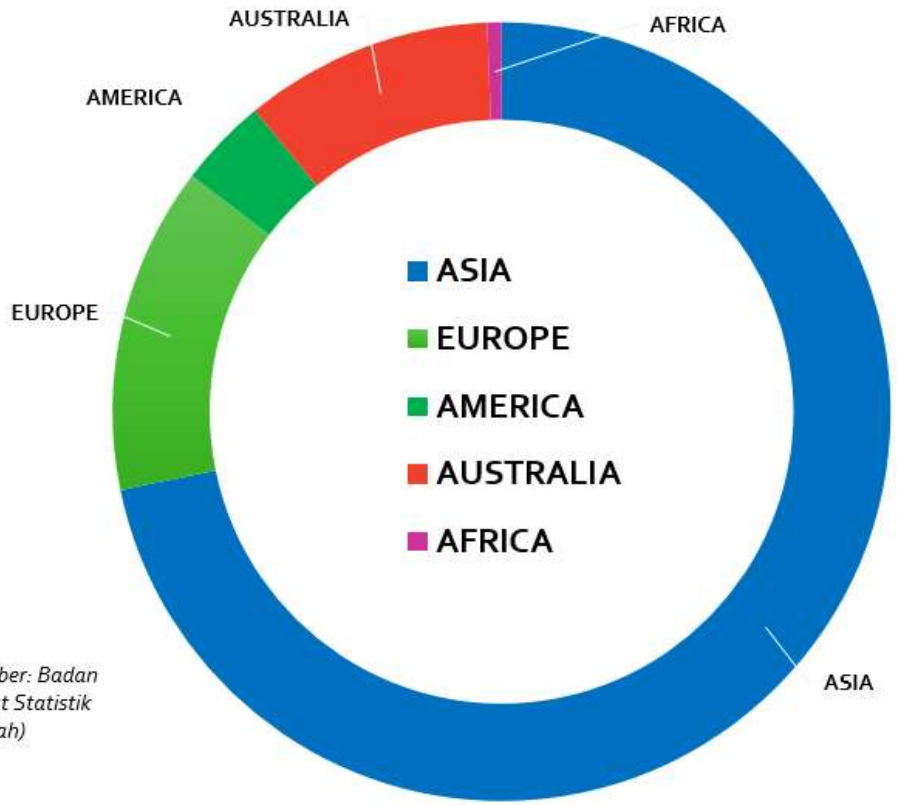
Tweets: Disaster



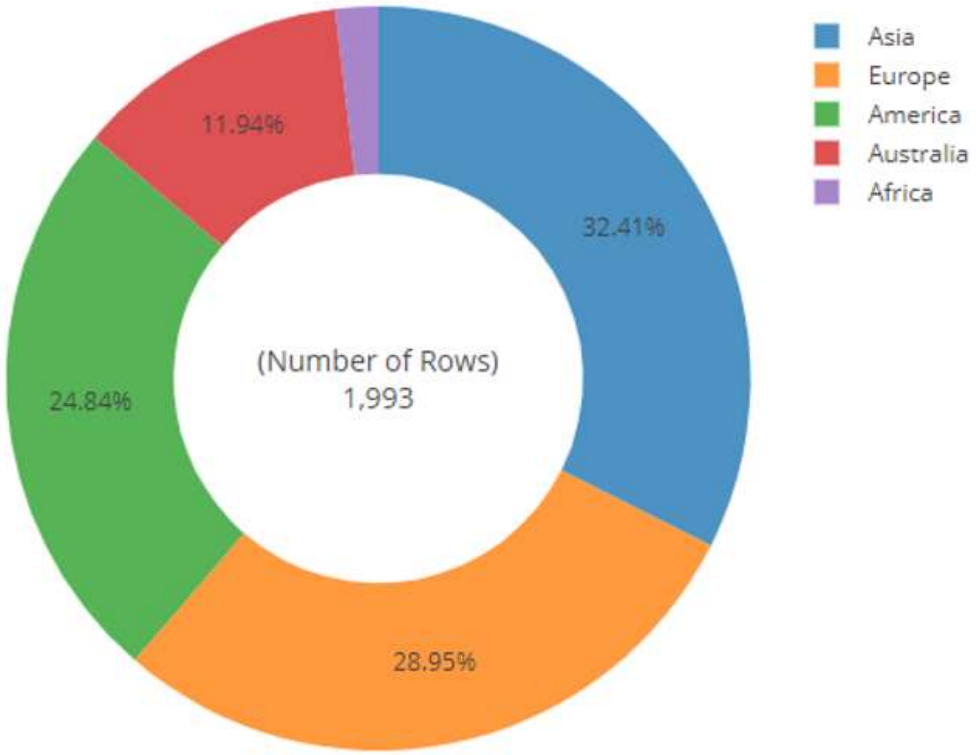
Lombok Earthquake 19 Agustus 2018.

Online News and Social Media Analysis of Indonesia Tourism Promotion

Persentase Wisatawan Manca Negara Menurut Kebangsaan Tahun 2017-2018

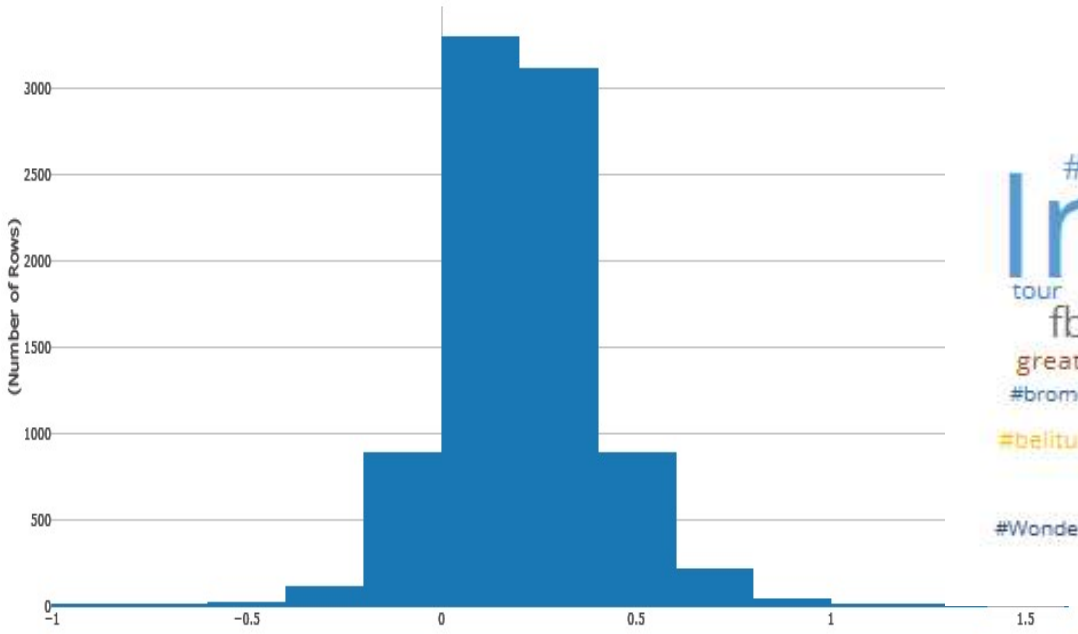


Sumber: Badan Pusat Statistik (diolah)



Persentase *Tweets* Menurut Kawasan/Benua

Sentiment Analysis:

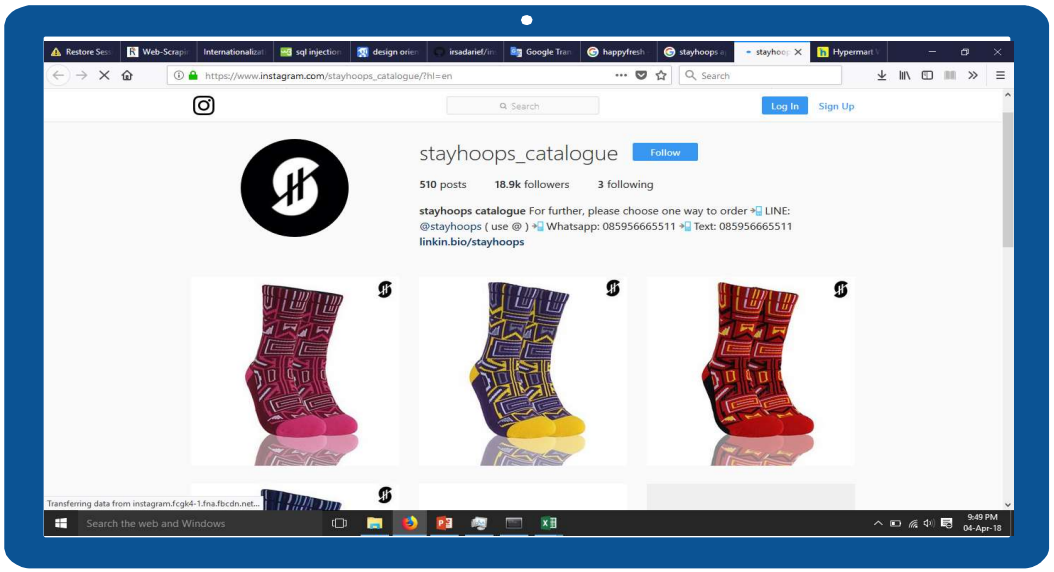


Most of Tweet are positives

Further analysis are being performed

Instagram Scrapping for Online Shopping

- 5 Million Instagram Users in Indonesia
- Get Price Products from Instagram



19k Followers
currently provides +500 Products

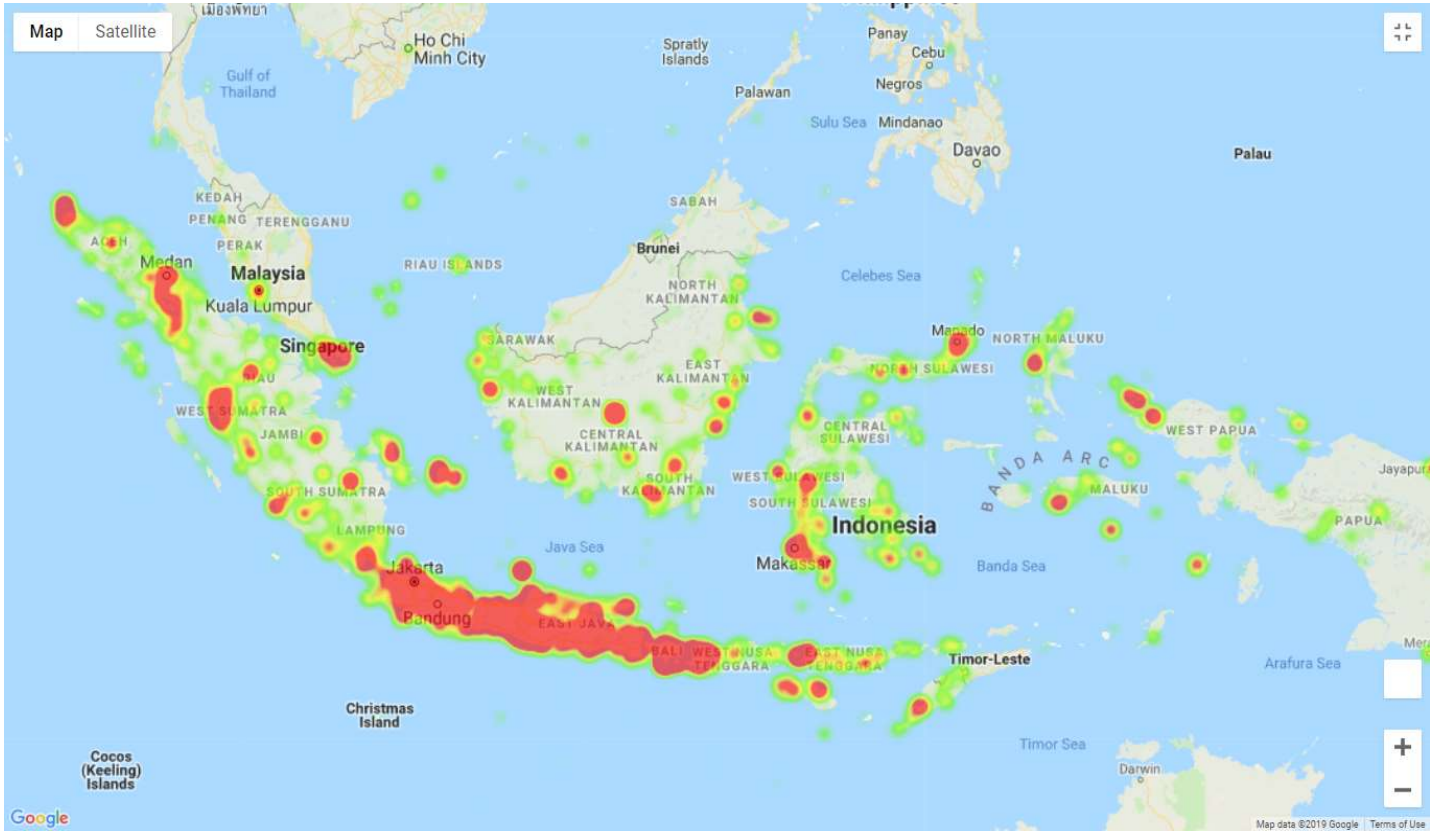
Original Text	Harga	Convert To
Ballerina short sleeve IDR.159.000 Available size: S, M, L, XL	159.000	159000
Grinding short sleeve IDR.159.000 Available size: S, M, L, XL	159.000	159000
Sleeveless Hoodie Blue Yellow IDR 275.000 Available size : S, M, L	275.000	275000

Instagram Scrapping for Online Shopping: Challenges

1. Choose right accounts :
   - Some accounts display the price of the product in the picture, this will certainly make it difficult in taking the price data, therefore the selected account must write down the price of the product in its caption
   - Choose accounts that uses the same pattern in write down the product price's in each post
2. The products sold on instagram are mostly lifestyle products: shoes, clothes, hijabs, etc. whose prices tend to be stagnant and unchanged in long period of time
3. Each account has its own pattern in write down the product price (example: IDR 125K, Rp 125rb, Rp. 125.000, 125000, etc.)

Instagram: Geotag Tourism

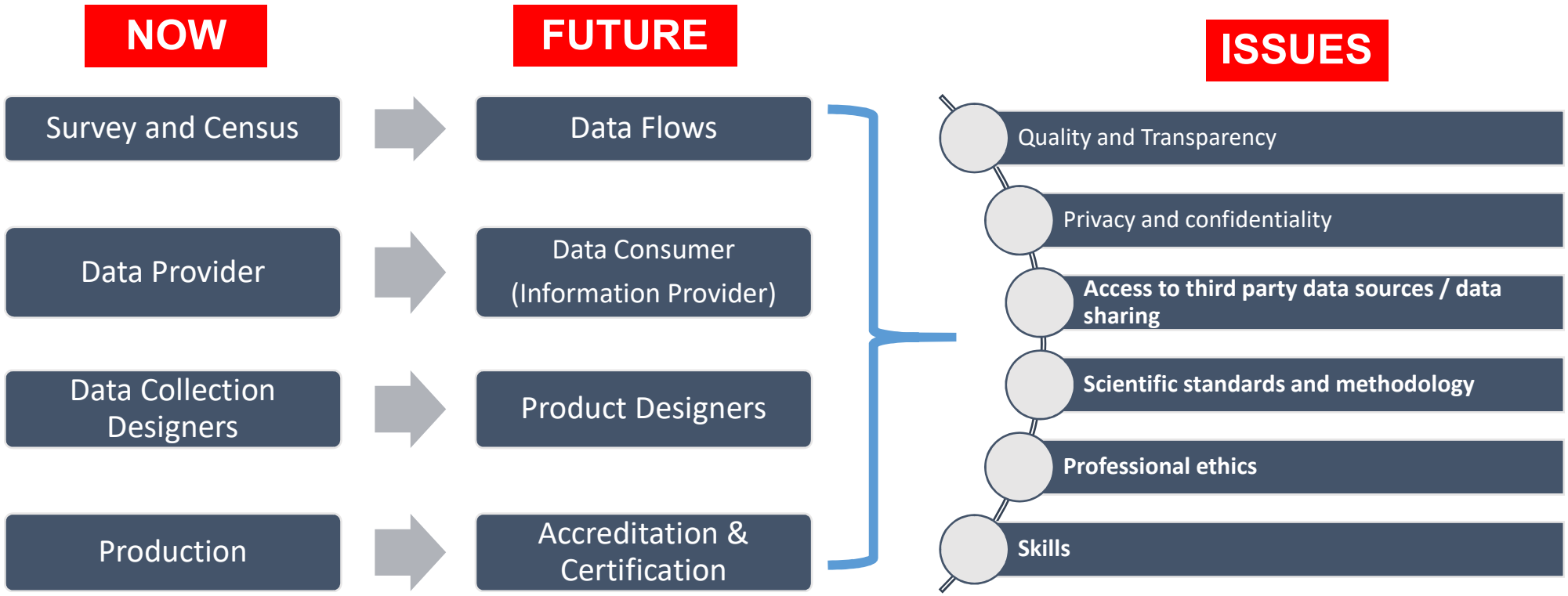
- #Wonderfulindonesia
- #Visitindonesia
- #Pesonaindonesia
- #Indonesiatourism
- #Exploreindonesia
- Total Images: 877.350
- With geo-tag : 225.328



Current Projects

- Social Media Geospatial Data Analysis (Instagram) for:
  - Tourism Pattern in Indonesia
  - E-commerce (informal) analysis
- Automatic Web Scraping for:
  - Price nowcasting
  - Tourism Potential Analysis (Hotel Occupancy etc.)
- STIS data compilation and Big data analytics
- Study of the methodology of using Mobile Positioning Data (MPD) for commuter statistics in Indonesia.
- Image Analysis for Planting Pattern
- Study on big data use for:
  - Labor statistics
  - Unemployment
  - Telecommunication Statistics
  - Population census Combined method
  - Tourism Statistics

Future National Statistics Office



Eurostat (2016)

Challenges in Big Data Use for Official Statistics Production

- High dimensionality and extremely large size
- Possible coverage/selection bias
- Data accessibility, new legislation? Permission by public?
- Increase risk of data disclosure
- New Sampling Algorithm
- Heavy computation, new algorithm and analytic tools
- Integration of files from multiple sources in different format in different times
- Risk of data manipulation or sudden unavailability

Summary

- Modernization of Statistical System is on progress…
- Still many approaches have to be explored and studied
- Contributions from **stakeholders** are needed