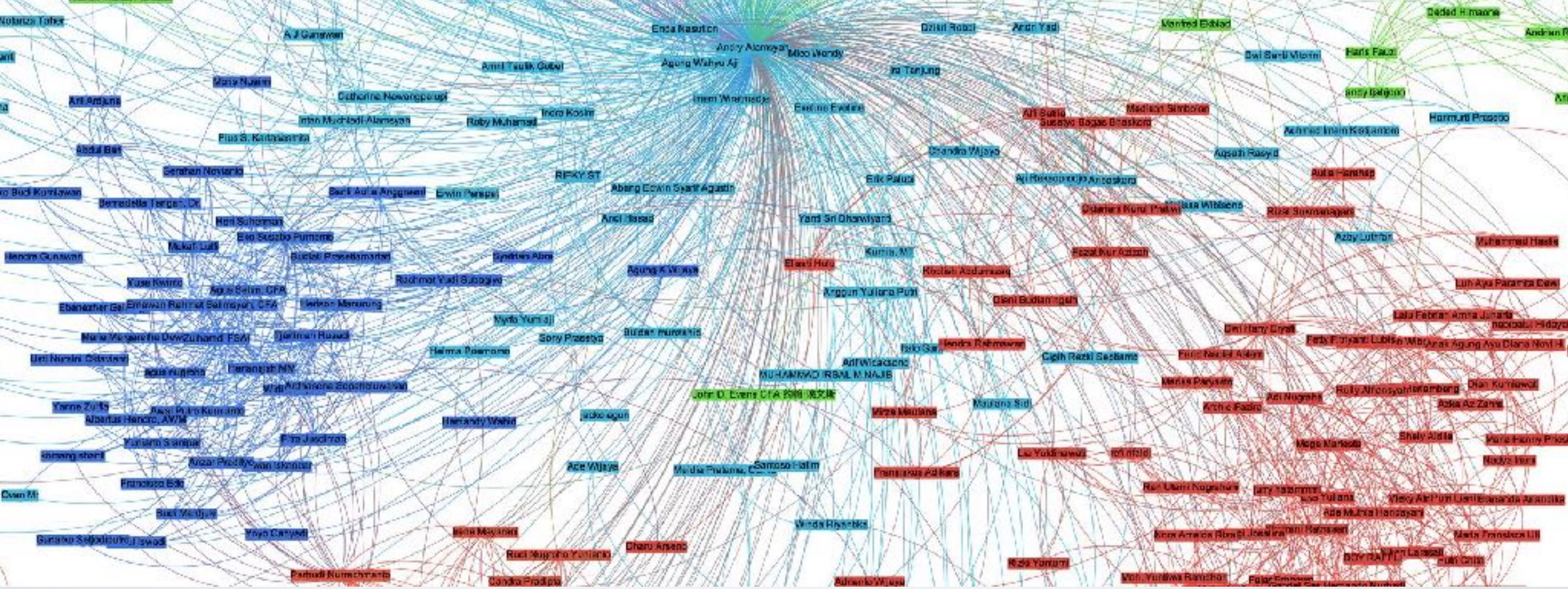


DATA ANALYTICS IN BUSINESS

CONCEPT AND CASE STUDY

Andry Alamsyah

Digital Business Ecosystem Research Center

LABORATORY SCBD SOCIAL COMPUTING & BIG DATA

Telkom University

ASOSIASI ILMUWAN DATA INDONESIA

Agenda

- **Finding Pattern**
- **Background**
  - Digital Ocean, Contextual Story, Disruptive Change, Cheap Change Everything, 4IR
- **Introduction to Big Data**
  - Definition, Taxonomies, Supporting Technologies, Talent Requirement
- **Big Data Value**
  - Adaptation Period, Complementary Method, Opportunity & Challenge
- **Business Case**
- **Remarks**

Who Am I

Andry Alamsyah

- **Researcher / Data Scientist**
- **Digital Business Ecosystem Research Centre**
- **Lab. Social Computing & Big Data**

Telkom University

Research Field :

Social Computing, Social Network, Complex Network / Network Science, Computational Social Science, Data Analytics, Big Data, Data Mining, Graph Theory, Disruptive Innovation / Disruptive Economy, ICT Entrepreneurial Business, Data / Information Business

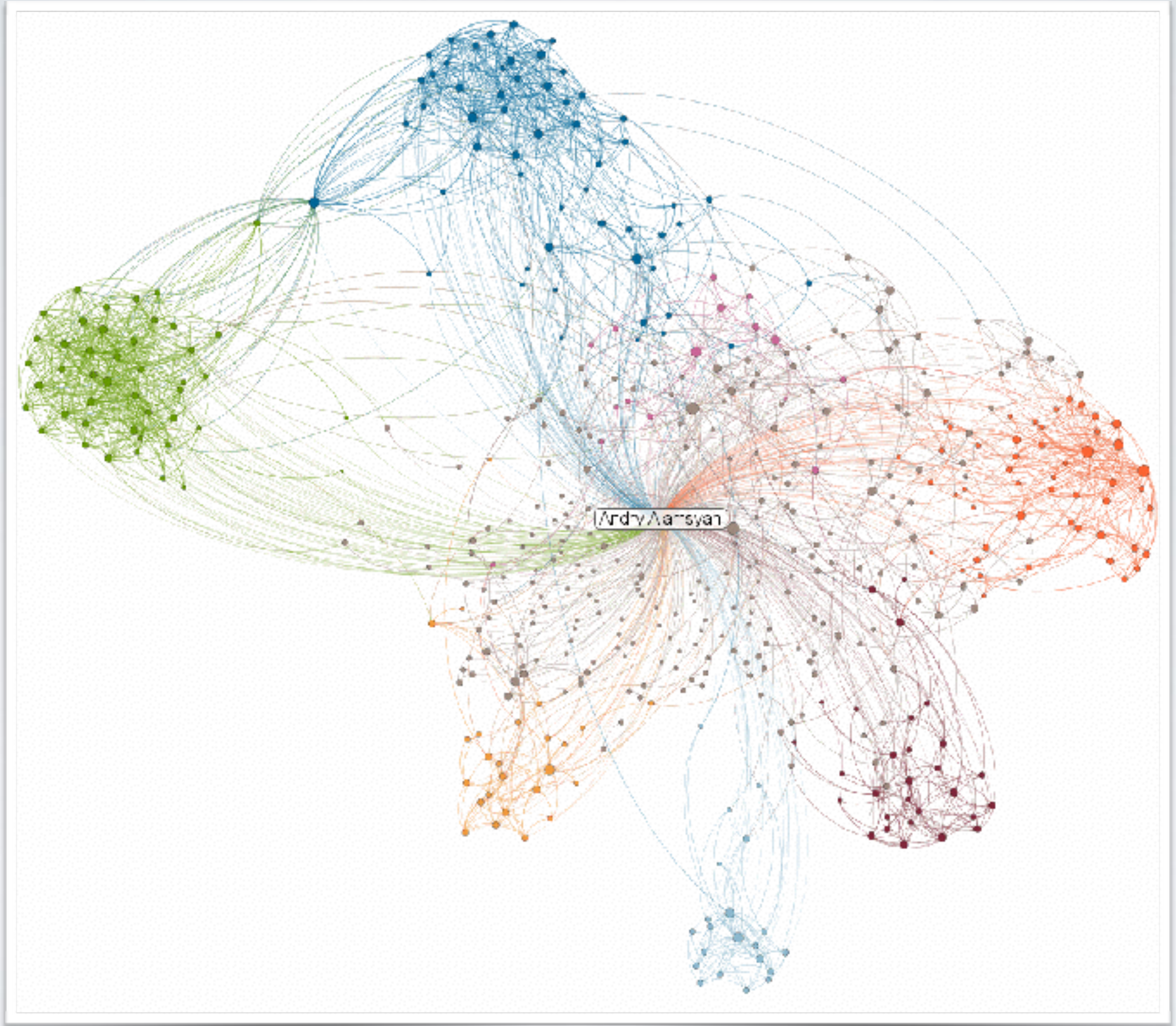
Education :

S1 : *Mathematics - ITB*, Topic: Statistics
S2 : *Informatics - UPJV, France*, Topic: Information System, and Multimedia
S3 : *Electro and Informatics - ITB*, Topic: Social Network, and Big Data

Links :

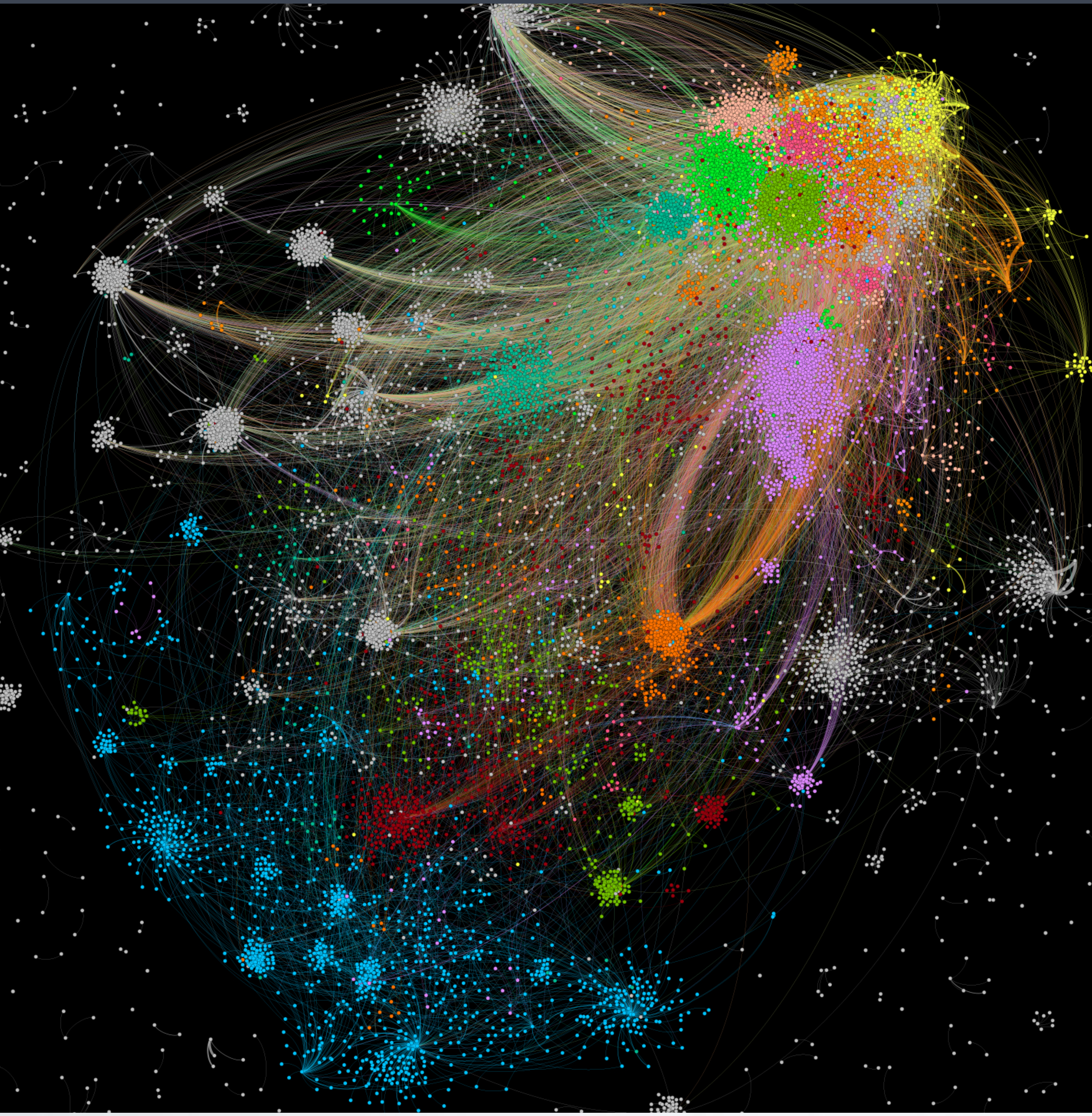
email	andry.alamsyah@gmail.com
blog	andrya.staff.telkomuniversity.ac.id
repository	telkomuniversity.academia.edu/andryalamsyah
repository	researchgate.net/profile/Andry_Alamsyah
linkedin	linkedin.com/andry.alamsyah
twitter	twitter.com/andrybrew

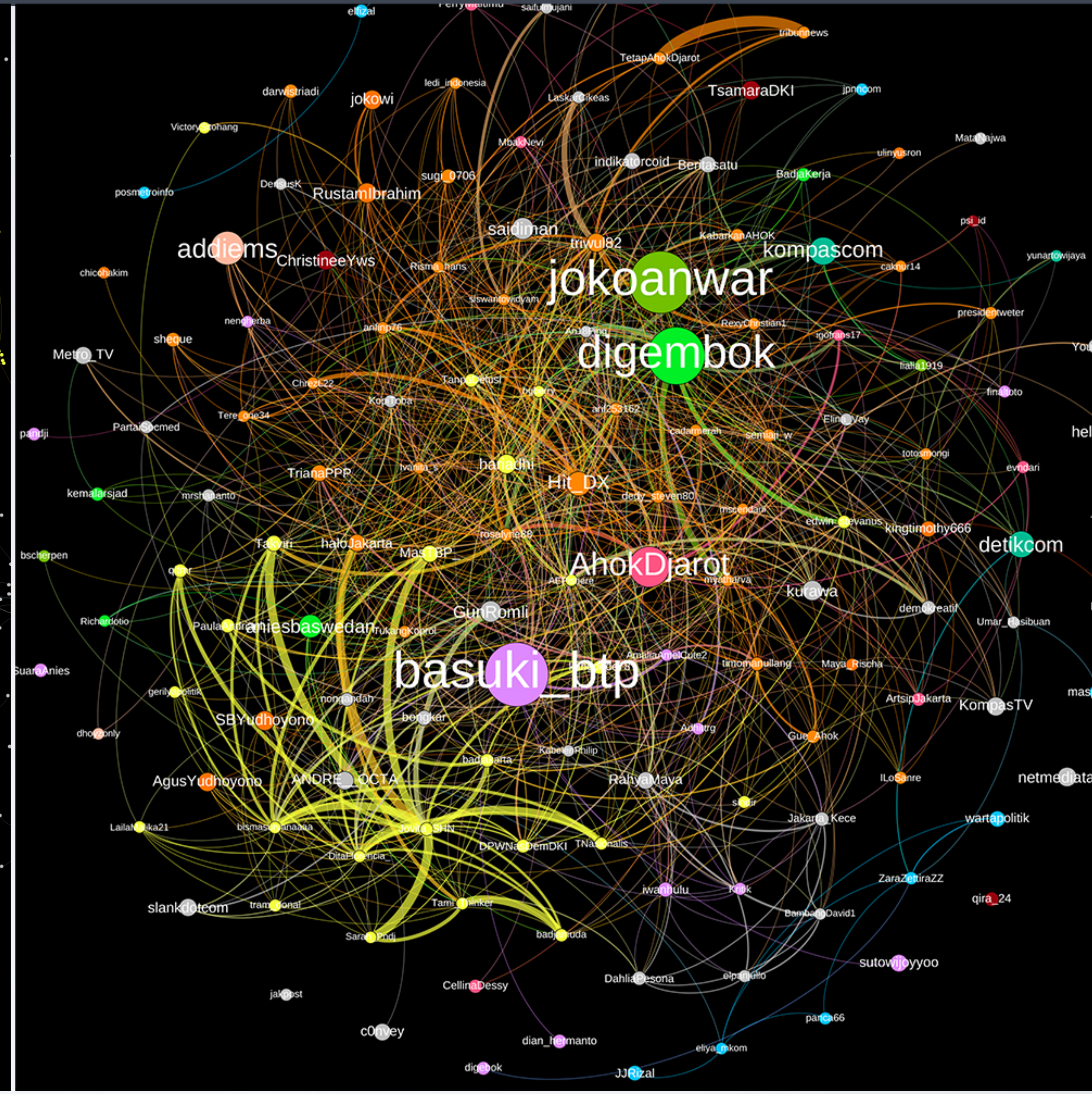
Who Am I



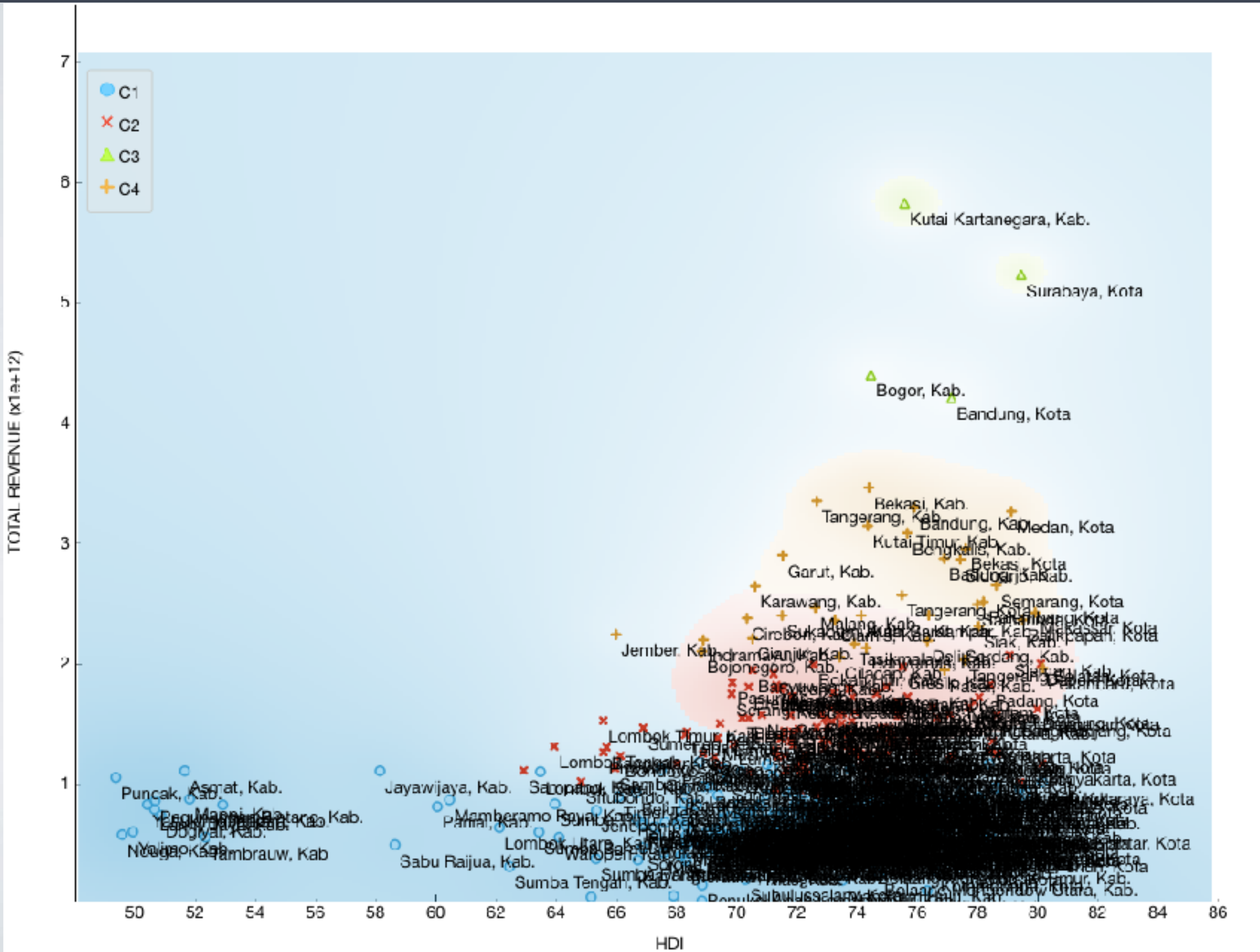
On a Bigger Scale

More Data Show Clearer Pattern

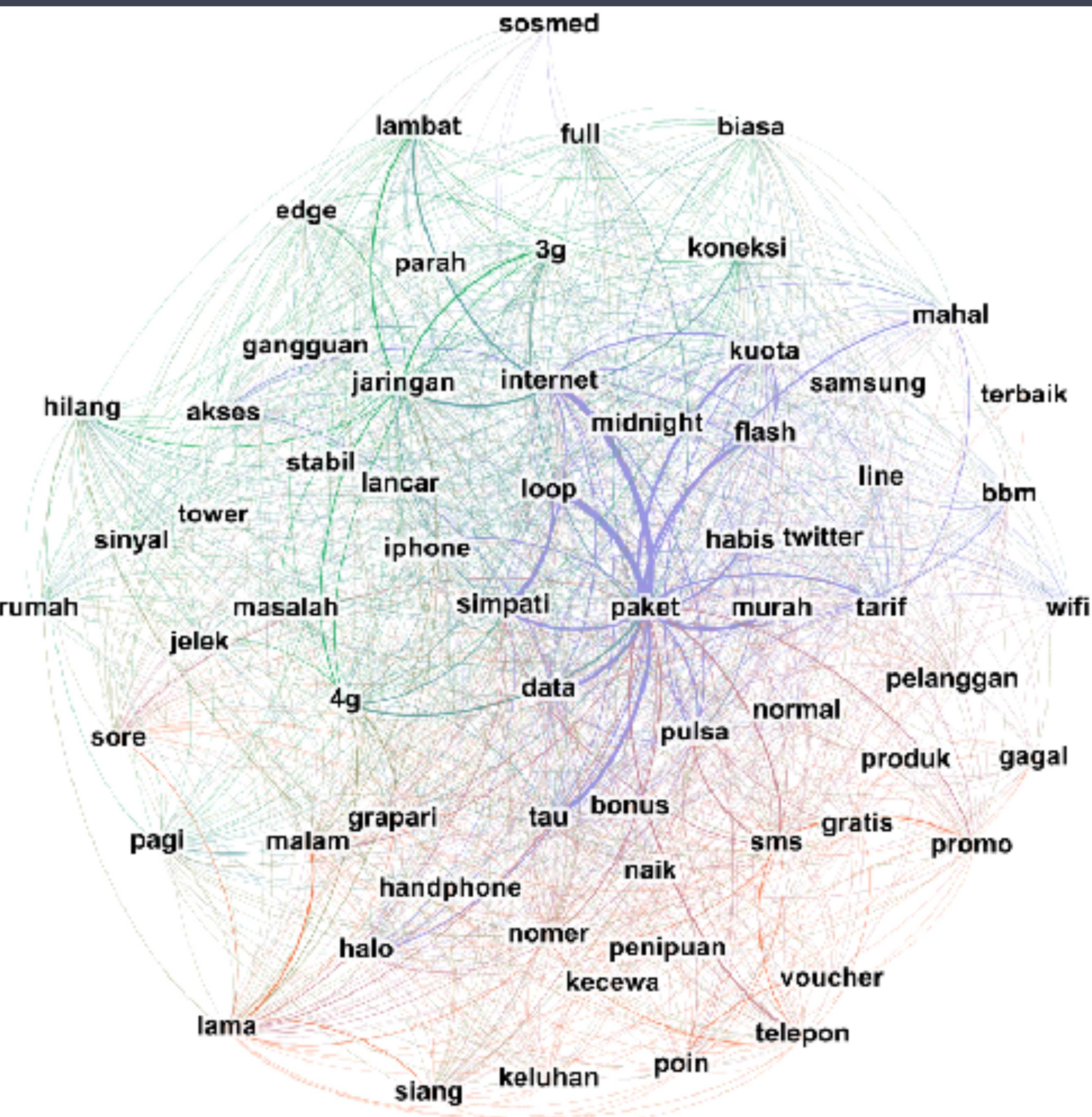




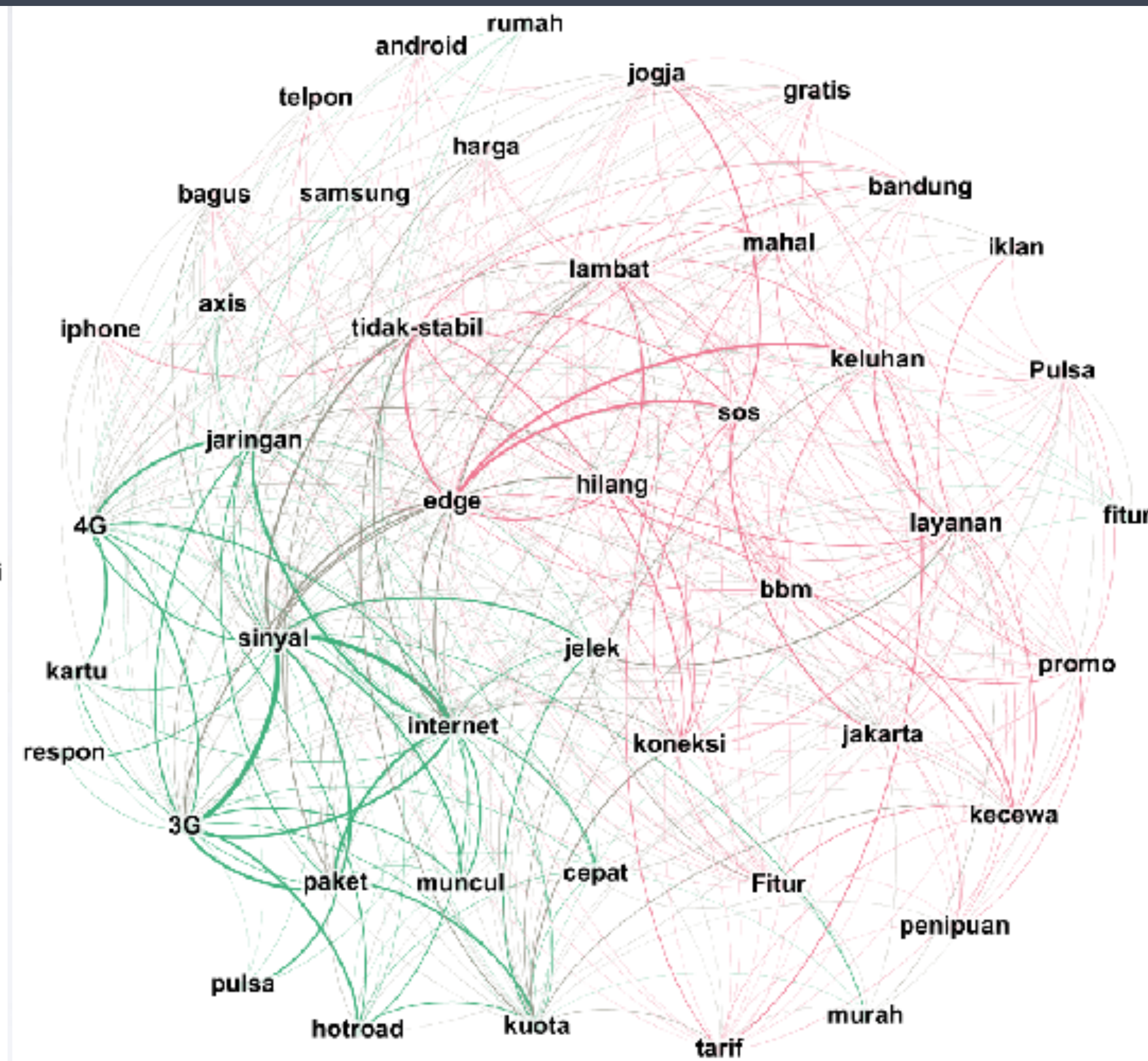
More Variable



Understanding Complex Data



telkomsel



XL

And At Last, Storytelling …

The New York Times

Asia Pacific

SUBSCRIBE | LOG IN

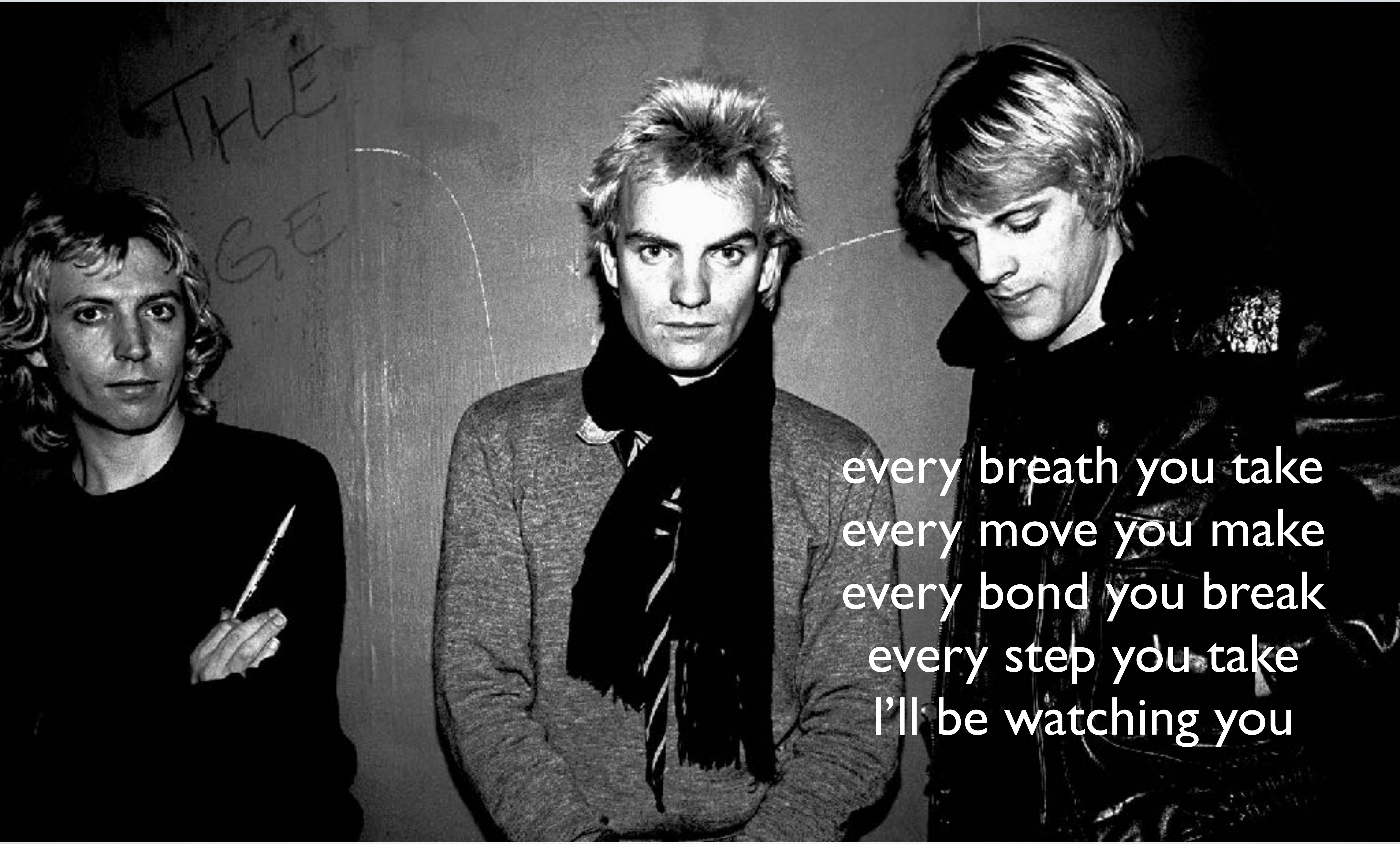
JAKARTA JOURNAL

Jakarta, the City Where Nobody Wants to Walk

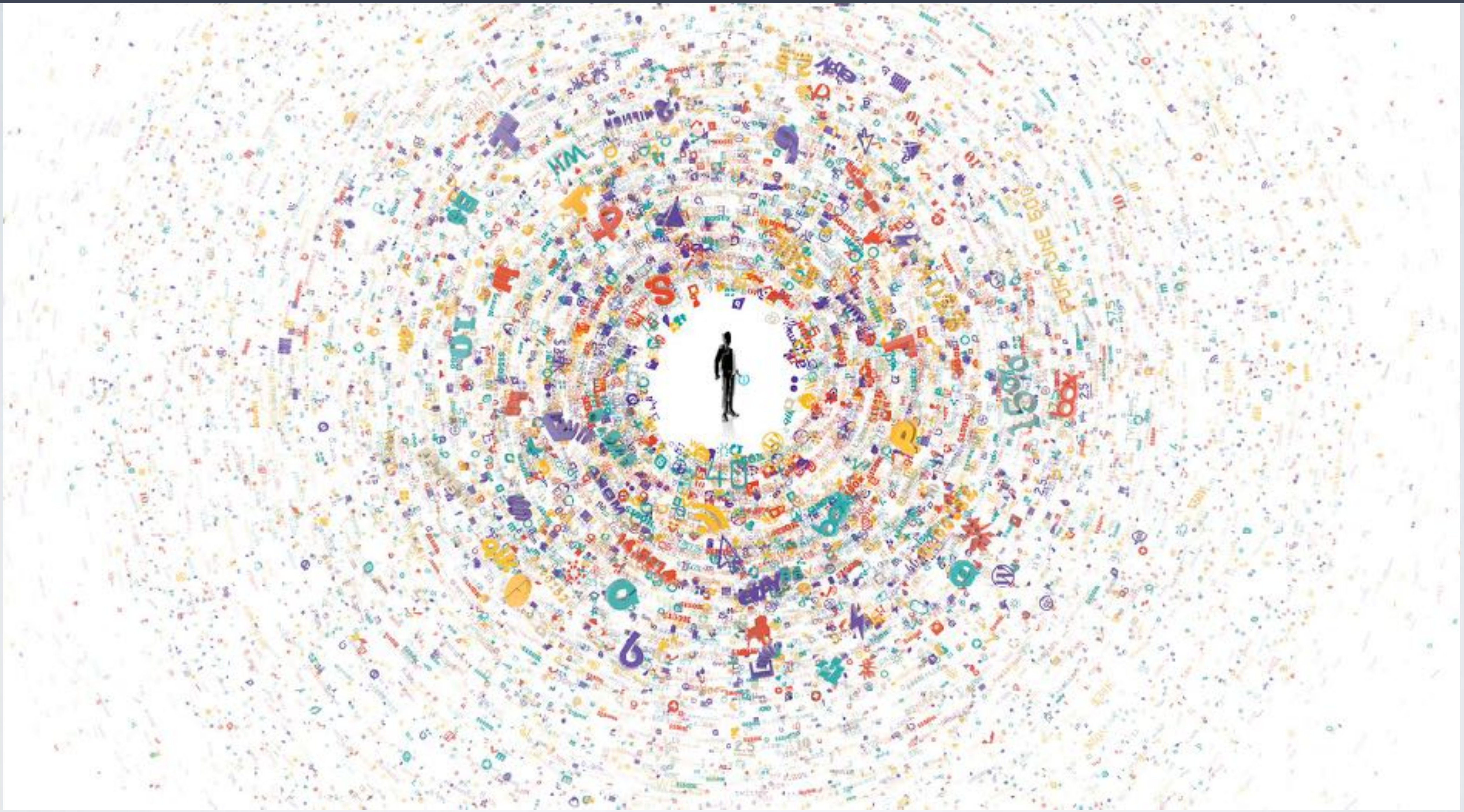
Commuters in central Jakarta. A study using a tracking app ranked Indonesia last among 46 countries and territories in average steps per person, at only 3,513 a day.

RONY ZAKARIA FOR THE NEW YORK TIMES

The Power of Data is ...



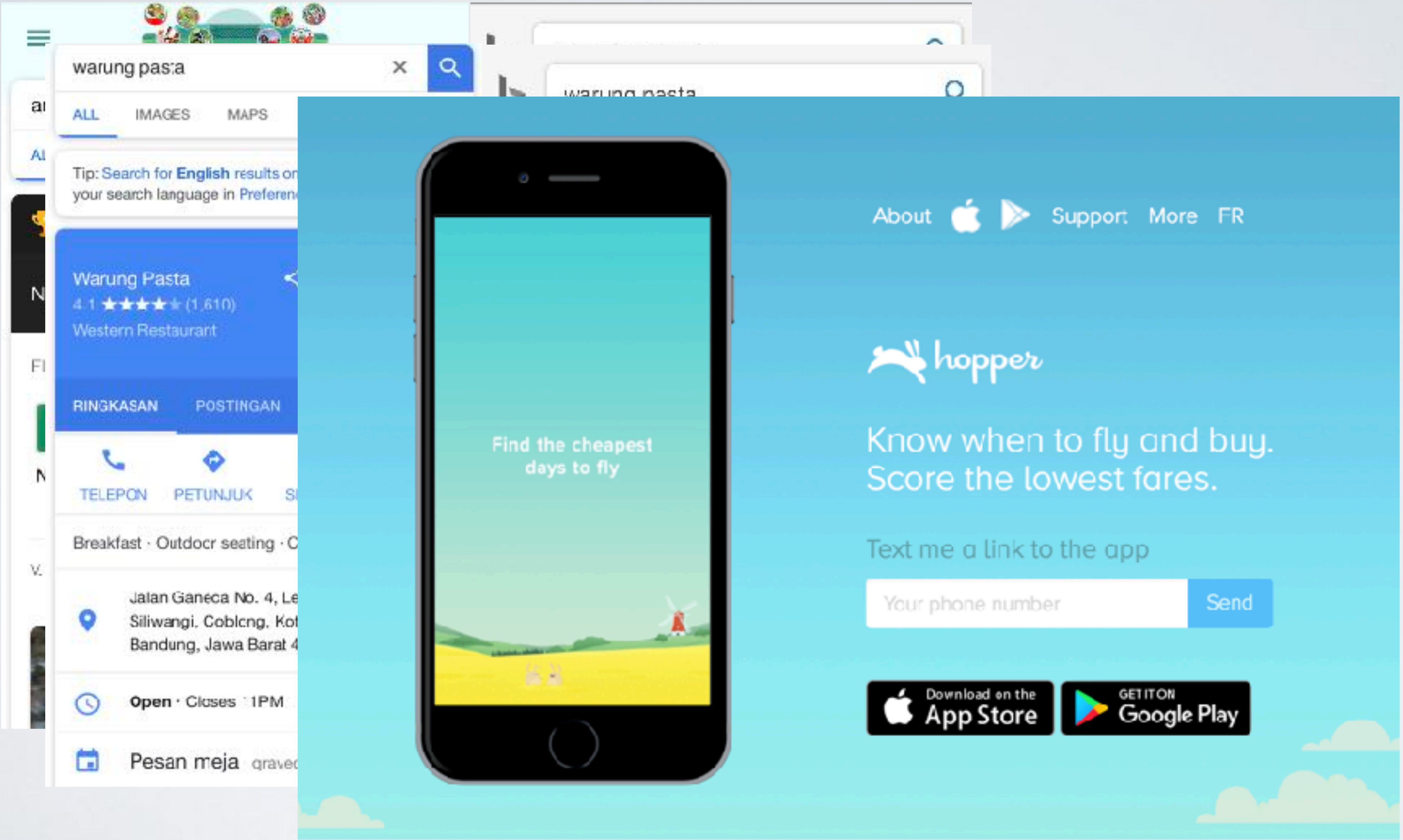
Background : Digital Ocean



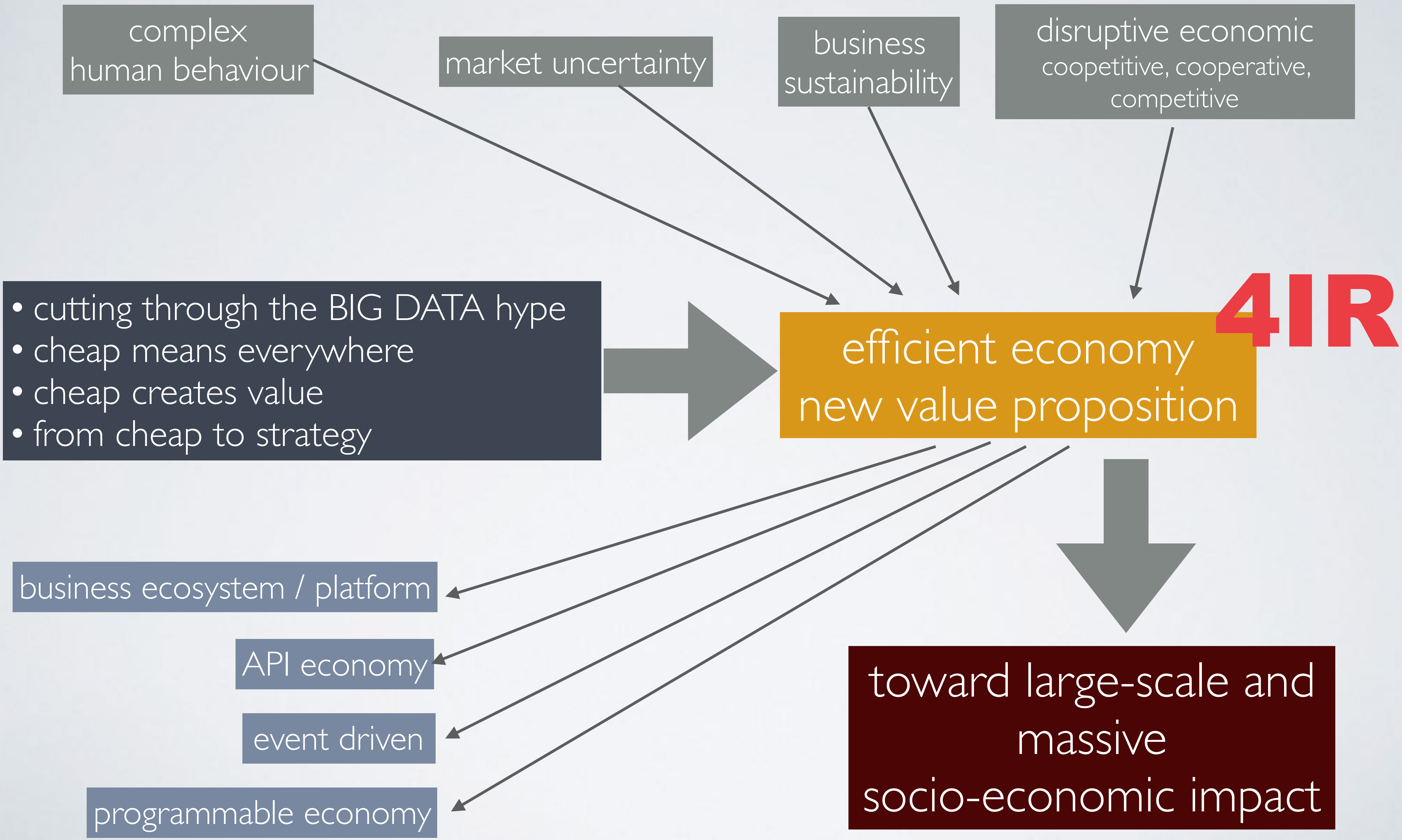
> information overload,
> technological based society
> acquire new value => new culture

> empowered individuals
> more data available
> building contextual story / search

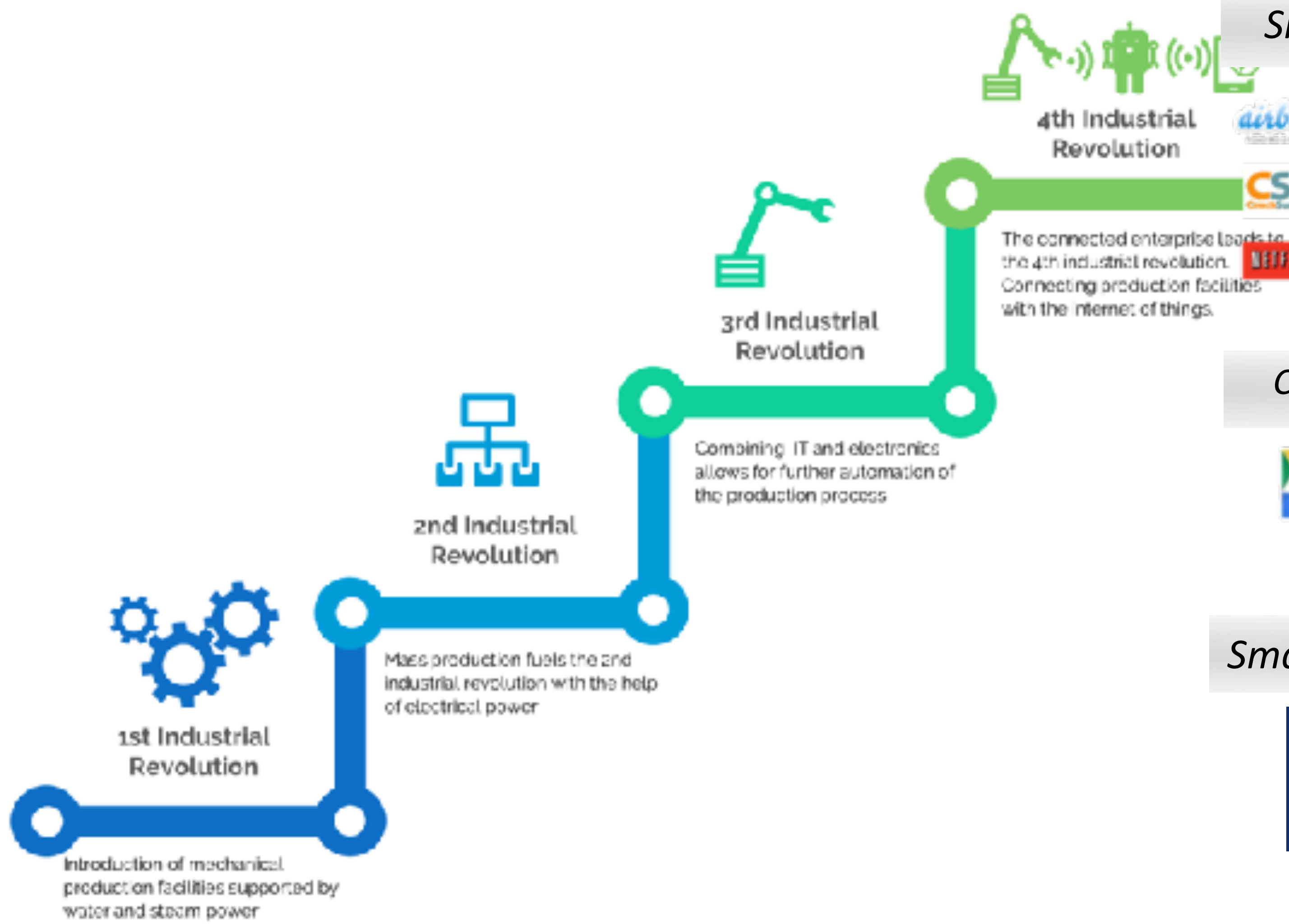
Background : Contextual Story



Background : Cheap Change Everything



Background :
The 4th Industrial Revolution (4IR)



Sharing economy

e-Education

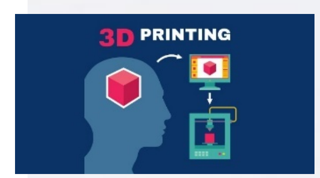
e-Government

Cloud Collaborative

Marketplace

Online Health Services

Smart Manufacturing



Smart City

Smart Appliances

Introduction to
BIG DATA (Use Case)

Analytics
-the discovery, interpretation, and communication of meaningful patterns in data (*wikipedia*)
-the process to uncover hidden patterns, unknown correlation, and other useful information that can help organizations make more informed business decision

opportunity

activity

benefit

BIG DATA
large, fast, complex
the nV's data

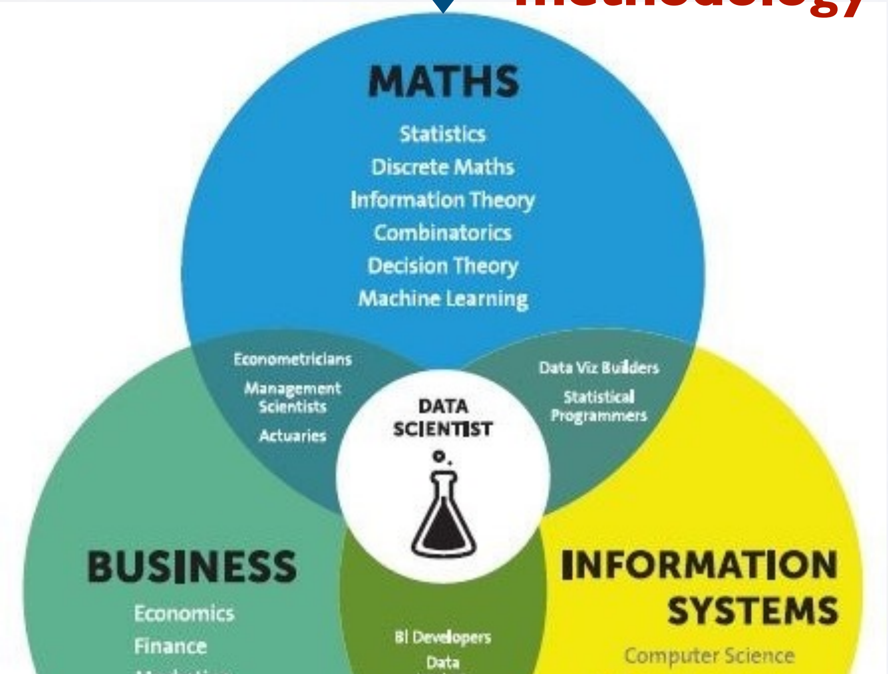
DATA SCIENCE
the science to extract knowledge / pattern from data

SOCIAL COMPUTING
quantification of human / social behavior

methodology

application

SOURCE
review, opinion, historical data, conversation, network friendship, CCTV, Vlog, location tagging, etc

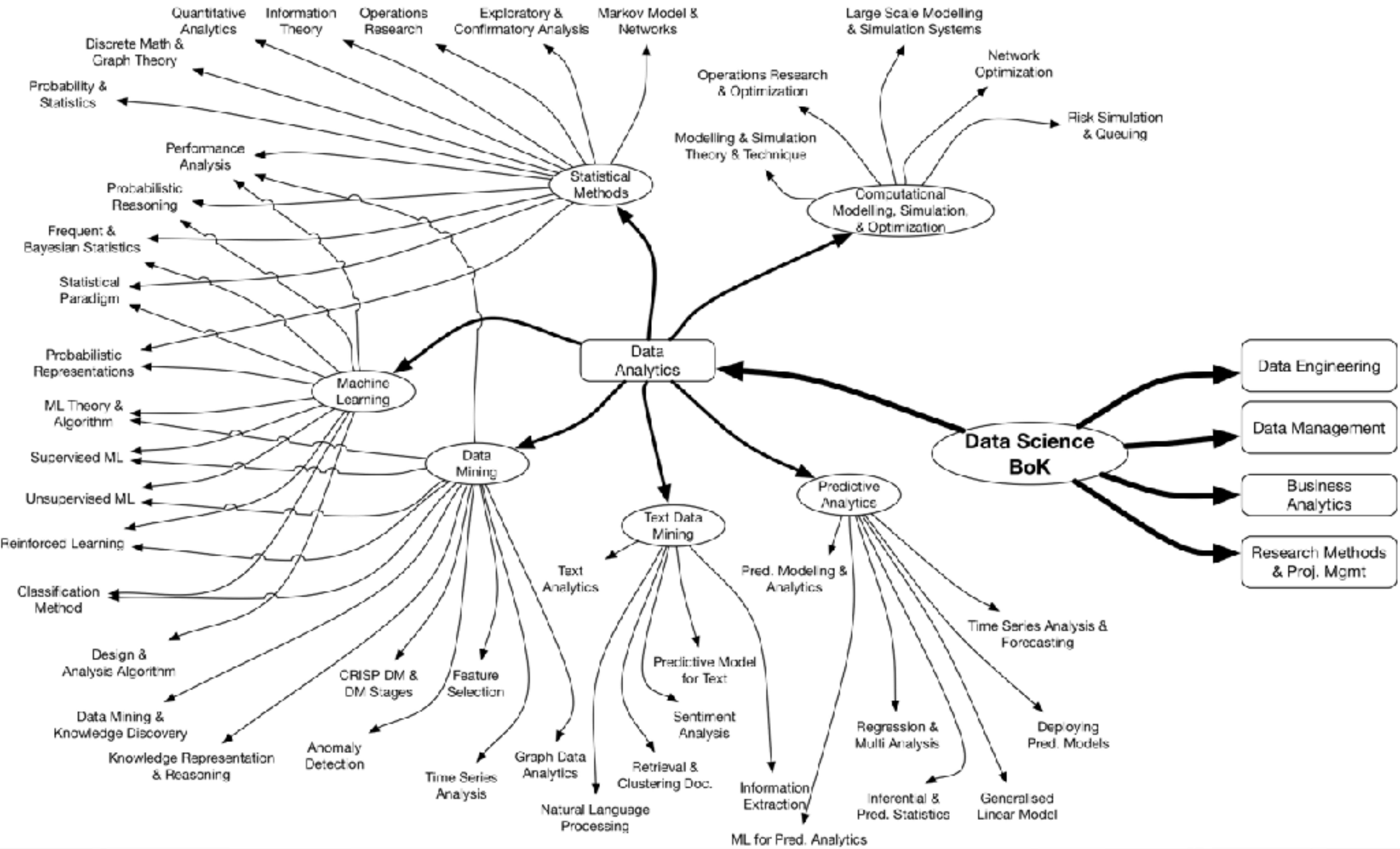


INSIGHT
market segmentation, risk analytics information dissemination, recommended investment, fraud detection, personalised adv, customer acquisition and retention, purchase behaviour, early detection event brand awareness, etc

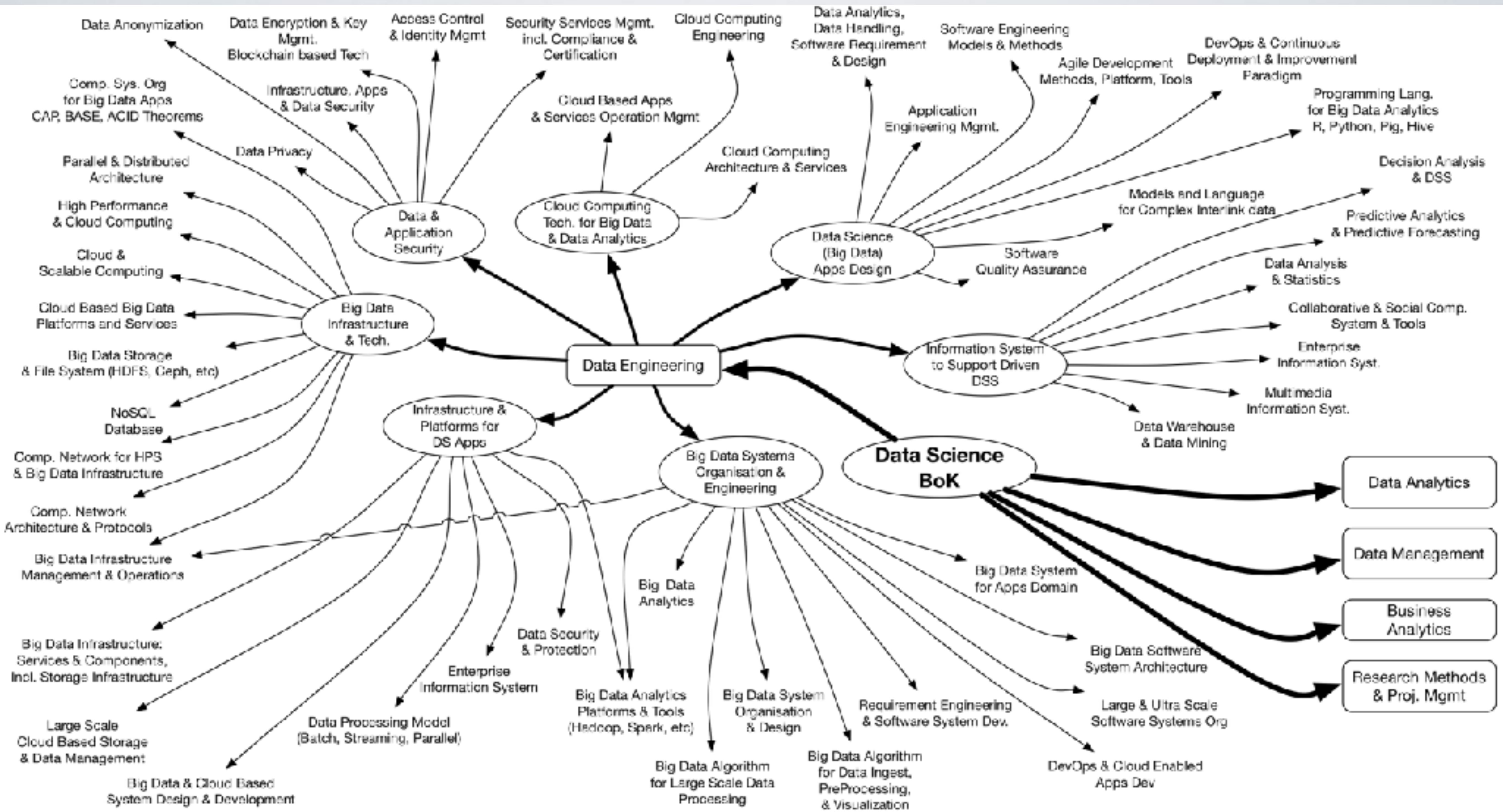
BIG DATA Taxonomy : Data Science Body of Knowledge

No	Name	Knowledge Area	Scientific Subject
1	Data Analytics	Statistical Methods, Machine Learning, Data Mining, Predictive Analytics, Computational Modeling / Simulation / Optimization	Computing Methodologies, Mathematics of Computing
2	Data Engineering	Big Data Infrastructure & Technologies, Infrastructure & Platform for DS Apps, Cloud Computing Tech, Data & Apps Security, Big Data System Organization & Engineering, DS / Big Data Apps Design, IS to support DSS	Algorithm & Complexity, Architeture & Organization, Computational Science, Graphic & Visualization, Information Management, Platform Based Dev., Software Engineering
3	Data Management	General Principle & Concepts in Data Mgmt and Organization, Data Management Systems, Data Enterprise Infrastructure, Data Governance, Big Data Storage, Digital Library & Archives, Data Curation, Data Preservation.	Data (Governance, Architecture, Model & Design, Storage & Operations, Security, Integration & Interoperability, Warehousing & BI, Quality), Metadata, Reference & Master Data
4	Research Methods and Project Management	Research Methods, Project Management	Project (Integration Mgmt, Scope Mgmt, Quality, Risk Mgmt)
5	Business Analytics	Business Analytics Foundation, Business Analytics Organisation and Enterprise Management	Business Analysis Planning & Monitoring, Requirement Analysis & Design Definition, Requirement Life Cycle Mgmt, Solution Evaluation & Improvements Recommendation
6	Domain Knowledge		

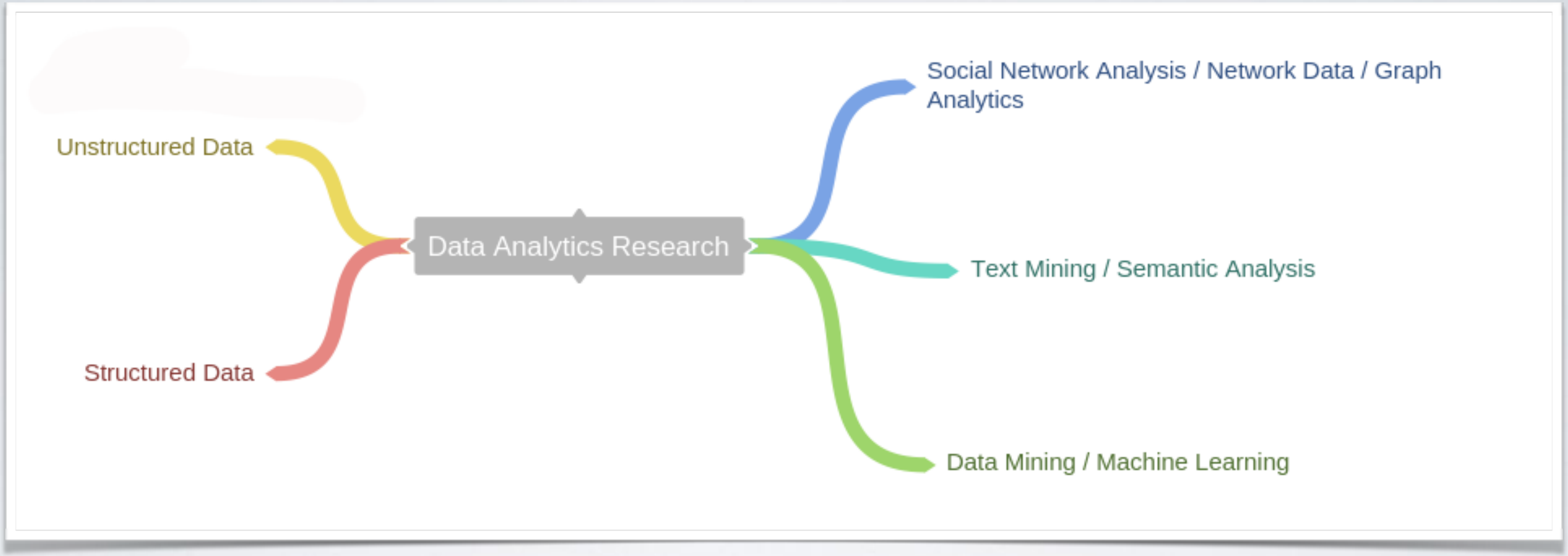
BIG DATA Taxonomy:
Body of Knowledge (Data Analytics)



BIG DATA Taxonomy:
Body of Knowledge (Data Engineering)



BIG DATA Taxonomy: Typical Works



BIG DATA:
Supporting Technologies

Data Science

Artificial Intelligence

Open Data

Internet of Things

Data Analytics

Machine Learning

Social Media Analytics

Prescriptive / Predictive Analytics

Stream Processing

Parallel Processing / Hadoop

Data Mining

Text / Video Analytics

Graph Analytics

NoSQL

How can (Big) Data Analytics help?

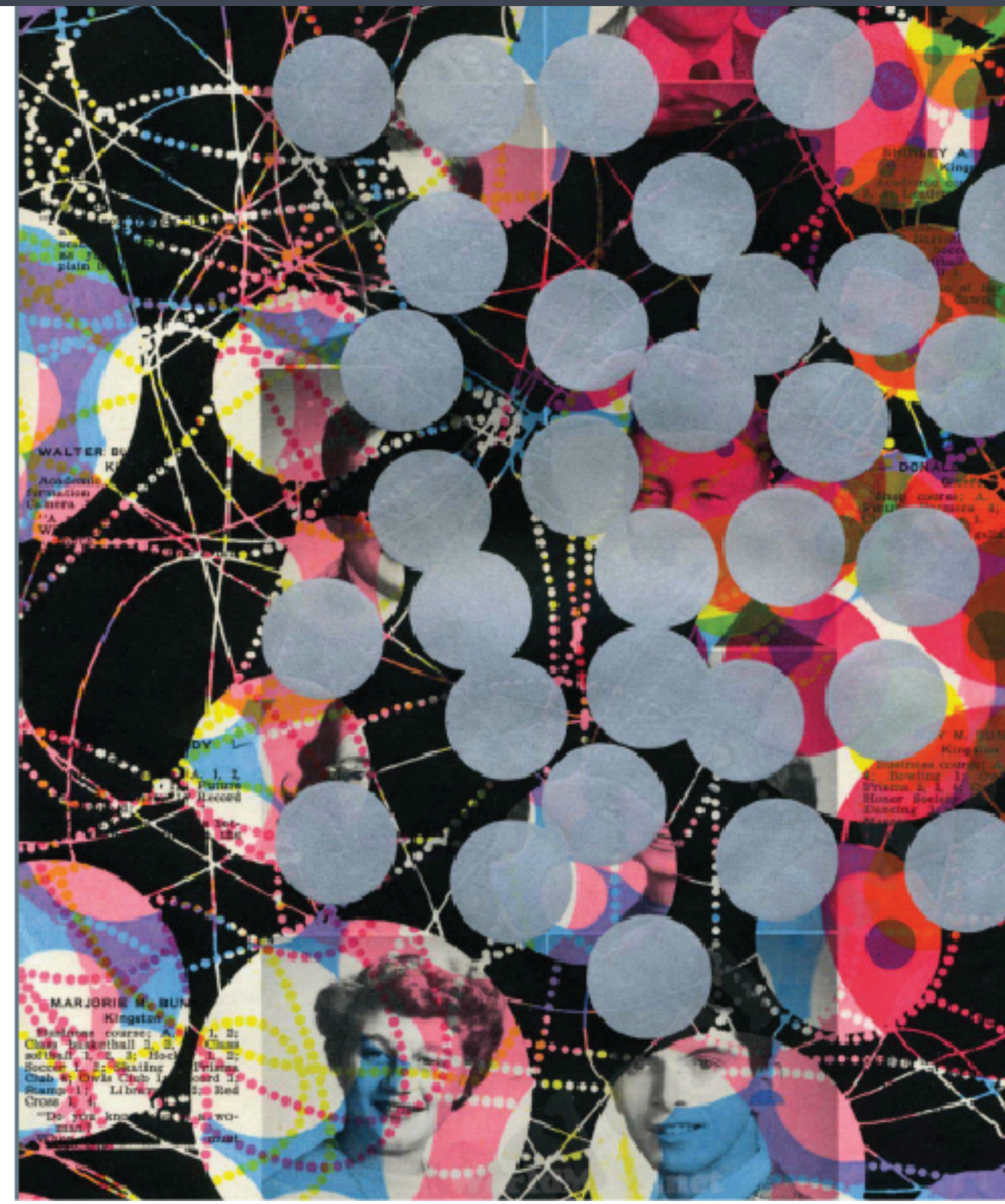
by describing the phenomenon, by predicting the value, by estimating the future outcome, by optimizing the resources and the decision, by simulating all the possible scenarios ..

BIG DATA:
Talent Requirement

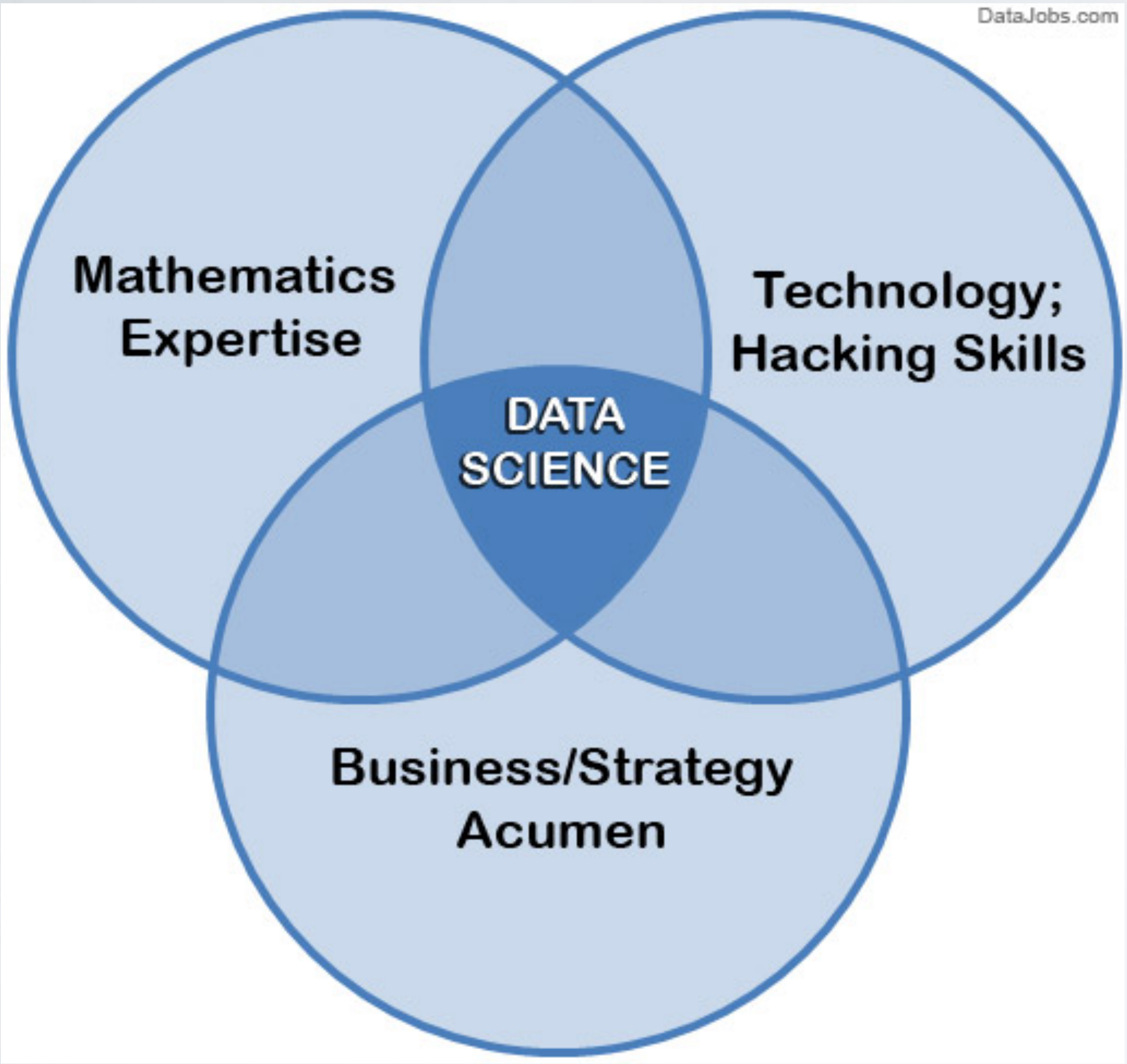
Data Scientist:
The Sexiest Job of the 21st Century

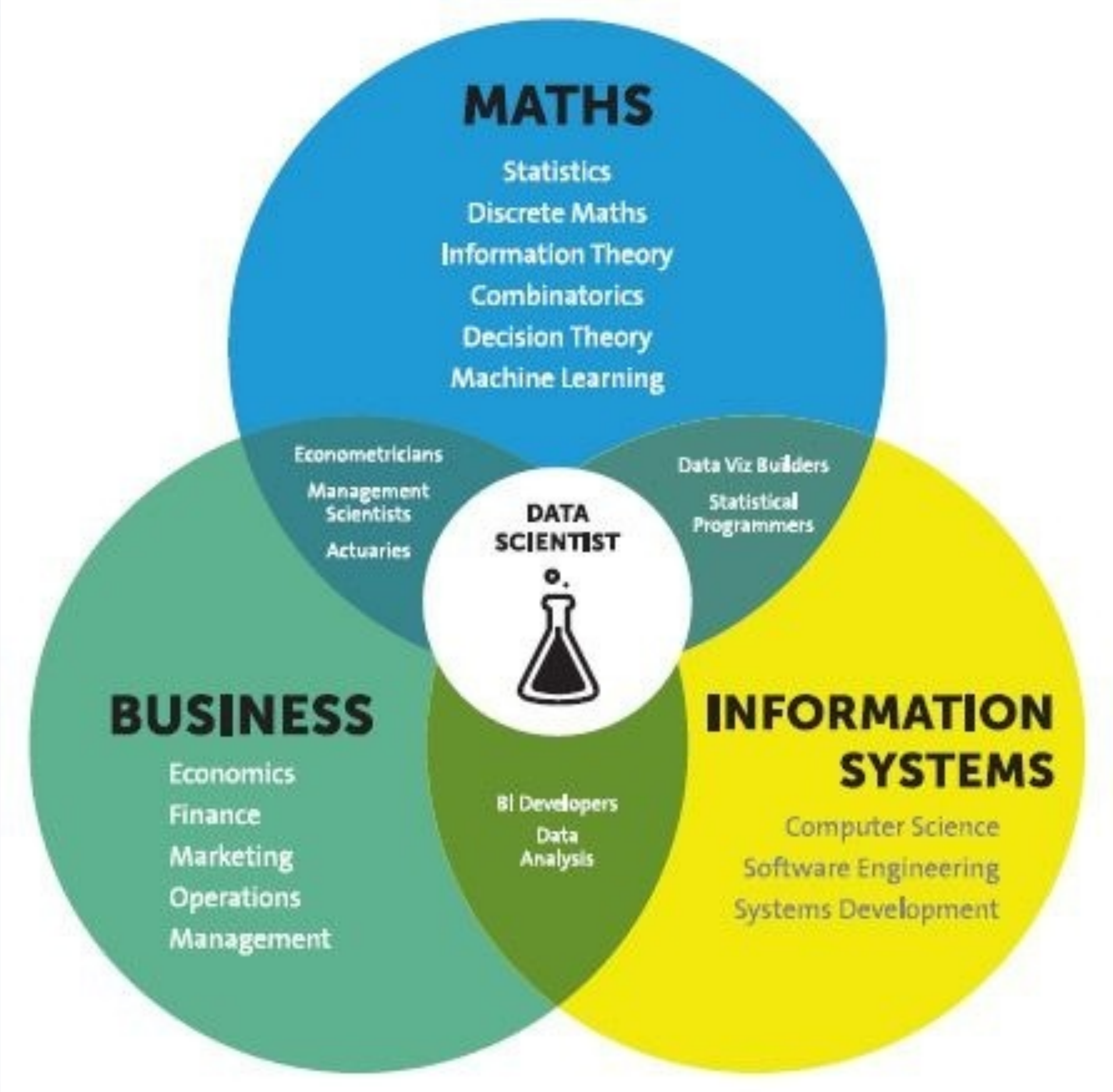
Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."



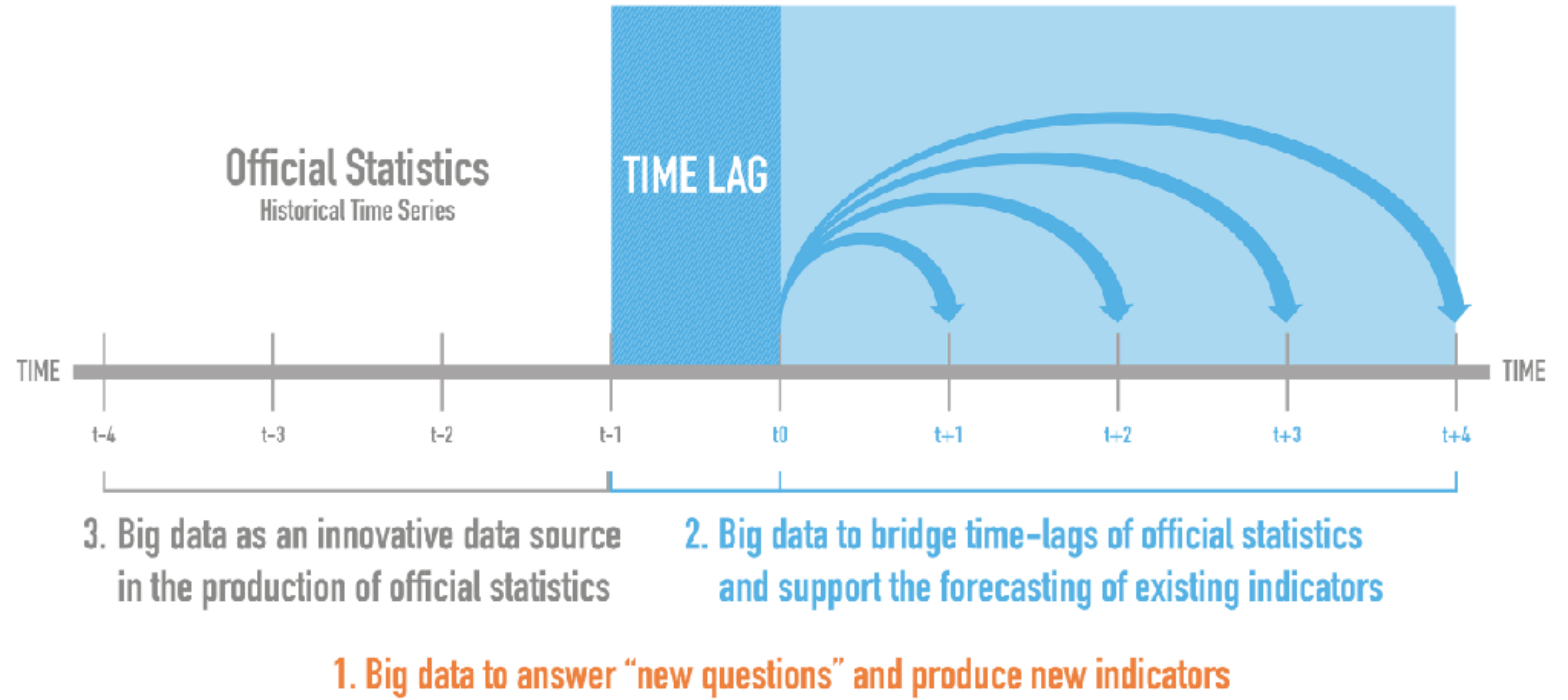
BIG DATA:
Talent Requirement





Business analytics requires a combination of expertise across business, information systems and mathematics.

BIG DATA ADAPTATION



source: Gabriel Quiros (IMF Statistics Department)

THE COMPLEMENTARY METHOD

STATISTICS	DATA MINING
Confirmative	Explorative
Small Data Set	Larga Data Set
Small Number of Variable	Large Number of Variable
Deductive (no predictions)	Inductive
Numeric Data	Numeric and Non-Numeric Data
Clean Data	Data Cleaning

source : Data Mining and Statistics :Whats the Connections ? (Jerome Friedman)

OPPORTUNITY

- provide granular, micro data
- provide relatively fast and cheap process
- machine learning methods and evaluation maturity
- Big Data initiatives towards future and sustainable economic activities
- uncovering hidden truths\

CHALLENGES

- hard to find right talent
- financial resources
- Big Data often populations study, so no sampling error => methods familiarity
- data quality + costs + security < benefit

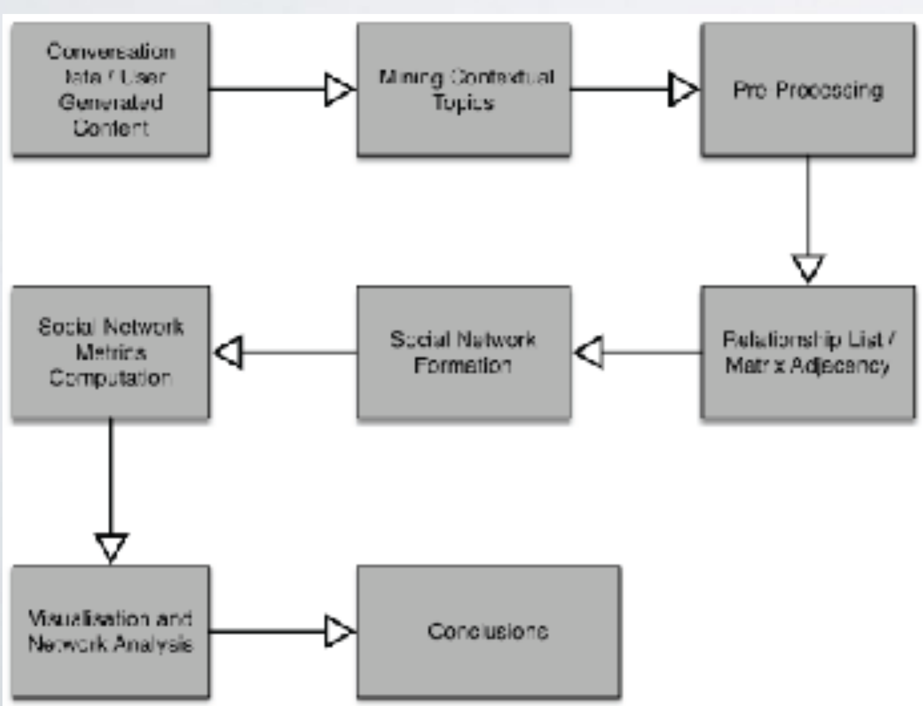
BUSINESS CASE

TOP BRAND ALTERNATIVE MEASUREMENT BASED ON CONSUMER NETWORK ACTIVITY

**Adv. Sci. Let, 2017*

Abstract:

In Business Intelligence effort, the legacy methodology to measure product brand awareness use technique such as surveys, interviews, and questionnaires. This methodology requires expensive effort to collect data from respondent and takes considerably time to accomplish. The availability of Big Data in the form of social media interaction can benefit us. The conversation and user generated content from social media certainly can be used to measure brand awareness through consumer activity. We use Social Network Analysis methodology to measure the dynamic and evolution of brand conversations in social media. By comparing the network properties, we propose new alternative measurement methods of product brand awareness. Our proposed methodology is better adapted to large-scale conversational data in social media. This measurement will also enhance the current methodology by viewing consumer opinions as a whole network and not as separated individual. This study conducted via social networking conversations on Twitter using two industry case studies, they are mobile operators and mobile phone brands in Indonesia

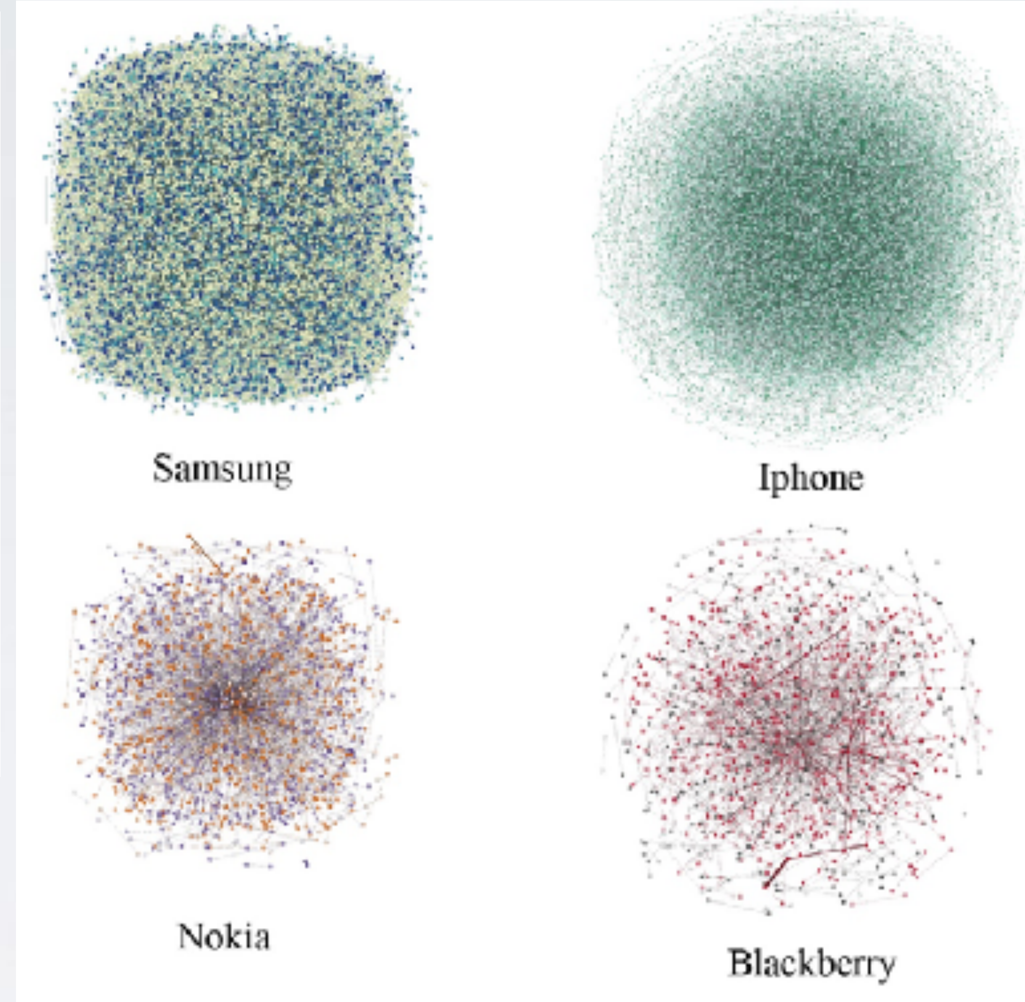


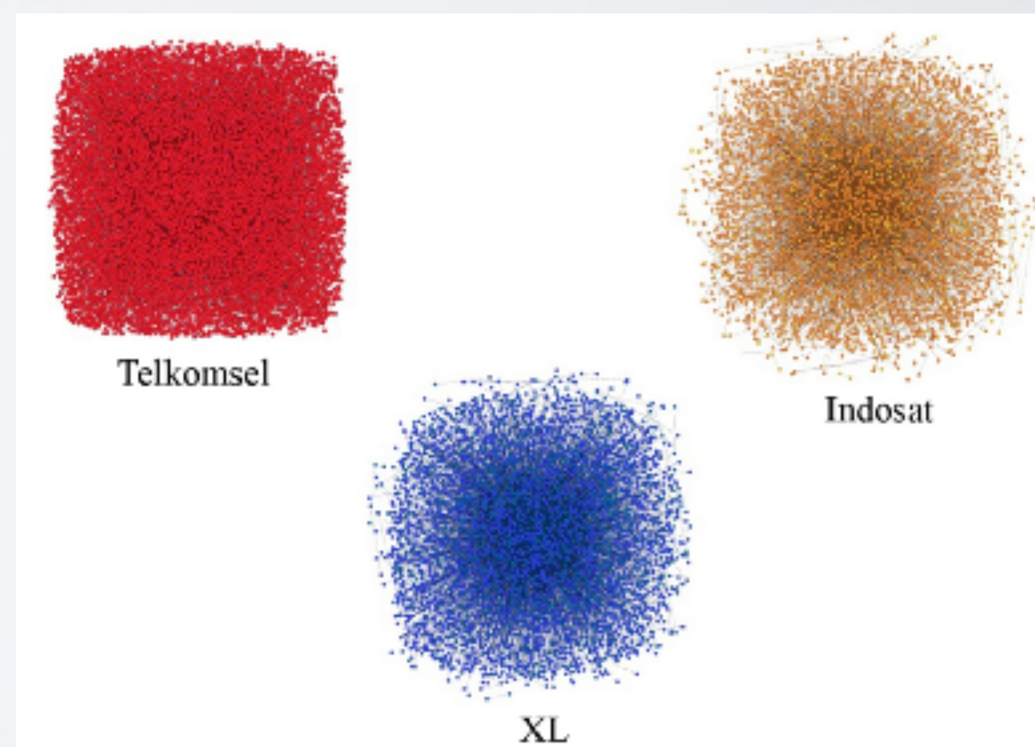
Network Property	Samsung (S)	Blackberry (B)	Nokia (N)	IPhone (I)	Rank
Size	11450 nodes 12805 edges	1381 nodes 1205 edges	1893 nodes 1604 edges	21014 nodes 21593 edges	1. I 2. S 3. N 4. B
Density	0.000017	0.00001[illegible]	0.000054	0.000061	1. I 2. N 3. S 4. B
Modularity	0.847	0.945	0.938	0.921	1. S 2. I 3. N 4. B
Diameter	19	20	10	25	1. N 2. S 3. B 4. I
Average Degree	2.237	1.7455	1.661	2.055	1. S 2. I 3. B 4. N
Average Path Length	4.342	7.295	3.691	6.160	1. N 2. S 3. I 4. B
Clustering Coefficient	0.378	0.258	0.317	0.244	1. S 2. N 3. B 4. I
Connected Component	646	302	361	2329	1. B 2. N 3. S 4. I

mobile phone rank

Network Property	Telkomsel (T)	XL (X)	Indosat (I)	Rank
Size	8353 nodes 11084 edges	4164 nodes 6375 edges	3772 nodes 4663 edges	1. T 2. X 3. I
Density	[illegible]	[illegible]	[illegible]	1. X 2. I 3. T
Modularity	0.491	0.864	0.752	1. T 2. I 3. X
Diameter	15	17	18	1. T 2. X 3. I
Average Degree	2.615	3.063	2.474	1. X 2. T 3. I
Average Path Length	3.384	5.582	4.862	1. T 2. I 3. X
Clustering Coefficient	0.508	0.450	0.490	1. T 2. I 3. X
Connected Component	485	296	352	1. X 2. I 3. T

mobile operator rank





A COMPARISON OF ECOMMERCE SENTIMENT ANALYSIS FOR MARKETING INTELLIGENCE EFFORT

CASE STUDY : BUKALAPAK, TOKOPEDIA, ELEVENIA

Abstract: The rapid growth of e-commerce market in Indonesia, making various e-commerce companies appear and there has been high competition among them. Marketing intelligence is important activity to measure competitive position. One element of marketing intelligence is to assess customer satisfaction. Many Indonesian customers express their sense of satisfaction or dissatisfaction towards the company through social media. Hence, using social media data, it provides a new practical way to measure marketing intelligent effort. This research performs sentiment analysis using naive bayes classifier classification method with TF-IDF weighting. We compare the sentiments towards of top-3 e-commerce sites visited companies, they are Bukalapak, Tokopedia and Elevenia. We use Twitter data for sentiment analysis because it's faster, cheaper and easier from both the customer and the researcher side. The purpose of this research is to find out how to process the huge customer sentiment Twitter to become useful information for the e-commerce company, and which of those top-3 e-commerce companies has the highest level of customer satisfaction. From the experiment results, it shows the method can be used to classify customer sentiments in social media Twitter automatically and Elevenia is the highest e-commerce with customer satisfaction

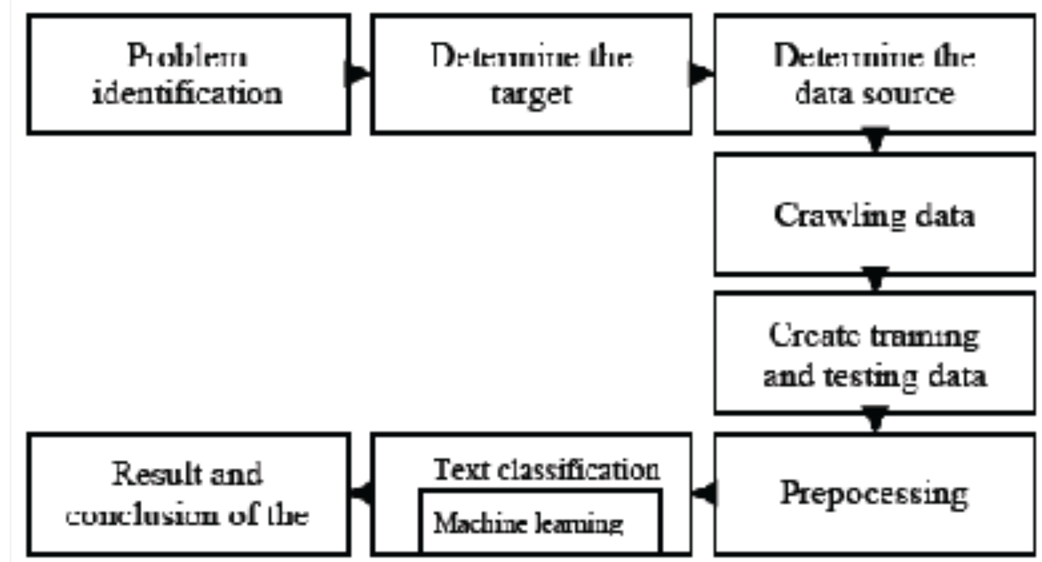


Figure 1 The Research workflow

Table 1 Example of text data used in training data

Sentences	Sentiment
tidak ada tanggung jawab jawab	Negatif
layanan baik	Positif
penjual rugi tombol bantuan tidak ada	Negatif
promo pulsa mantap	Positif
cara ribet	Negatif
iklan iklan lucu	Positif

Table 1 The result of confusion matrix

	Classification		Class Precission
	Positive	Negative	
Pred. Positive	338	26	92,86%
Pred. Negative	12	324	96,43%
Class Recall	96,57%	92,57%	

Table 3 Comparison percentage sentiment among Bukalapak, Tokopedia and Elevenia on social media Twitter

	Testing Data Number	Sentiment	
		Positive	Negative
Bukalapak	993	348 (45,9%)	410 (54,1%)
Tokopedia	928	350 (46,2%)	408 (53,8%)
Elevenia	300	163 (46,3%)	189 (53,7%)

COMPARABLE RESULT AMONG THREE CASE STUDY

NETWORK TEXT ANALYSIS TO SUMMARIZE LARGE-SCALE ONLINE CONVERSATIONS IN TELCO BUSINESS

Abstract - Market tight competition put pressure the companies to employ a new and faster way to support their marketing intelligence effort. The need of marketing intelligence includes gathering and analyzing data for confident decision making about market and its competition. Today, the abundant large scale data from online social network services has made possible to extract valuable information such as user opinions and sentiment from the conversations in the market. As the competition arise, new challenge emerged, which include faster data summarization. The common practice of summarize contents is using wordcloud or weighted list of appearance words. This approach is lack of sense and contextual relations between words in questions, because the words has no connection with other words that might construct an important phrase. With the help of graph formulation, we propose a methodology of network text analysis to summarize large conversation in online social network services. This proposed methodology capture complex relations betweenwords, while still maintain fast summarization. In this paper, we compare three major telecommunication provider in Indonesia, which is Telkomsel, XL and Indosat. The conversations about those brands in online social network services Twitter is collected, Network text about each brands are constructed and analyzed.

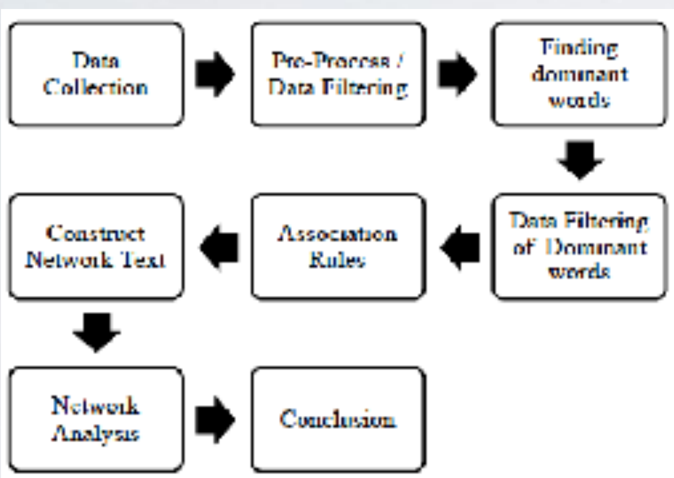


TABLE I. DATA PROFILE OF EACH TELECOMMUNICATION COMPANY

	TELKOMSEL	XL	INDOSAT
NUMBER OF TWEETS	46911	74771	14253
AFTER SECOND PROCESS	7974	4521	2210
AFTER FOURTH PROCESS	7818	3743	1209

TABLE II. TOP 15 TELKOMSEL WORDS PAIR

WORDS PAIR	WEIGHT
INTERNET PAKET	166
PAKET TAU	109
PAKET - FLASH	101
LOOP - SIMPATI	96
PAKET - DATA	94
PAKET - KUOTA	91
LOOP PAKET	89
PAKET - SIMPATI	74
INTERNET - JARINGAN	64
PAKET - TARIF	62
INTERNET - KUOTA	54
INTERNET LAMBAT	54
PAKET - PULSA	53
PAKET 4G	53
MURAH - PAKET	50

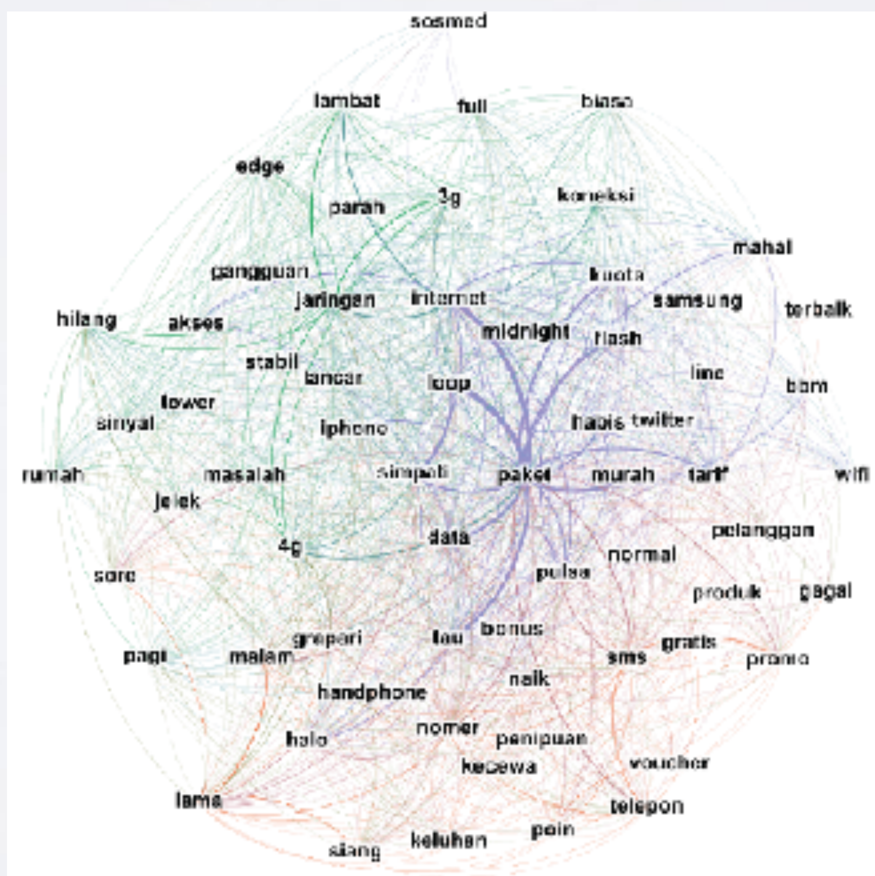


Fig. 3. Network text based on word pair and word cluster of Telkomsel

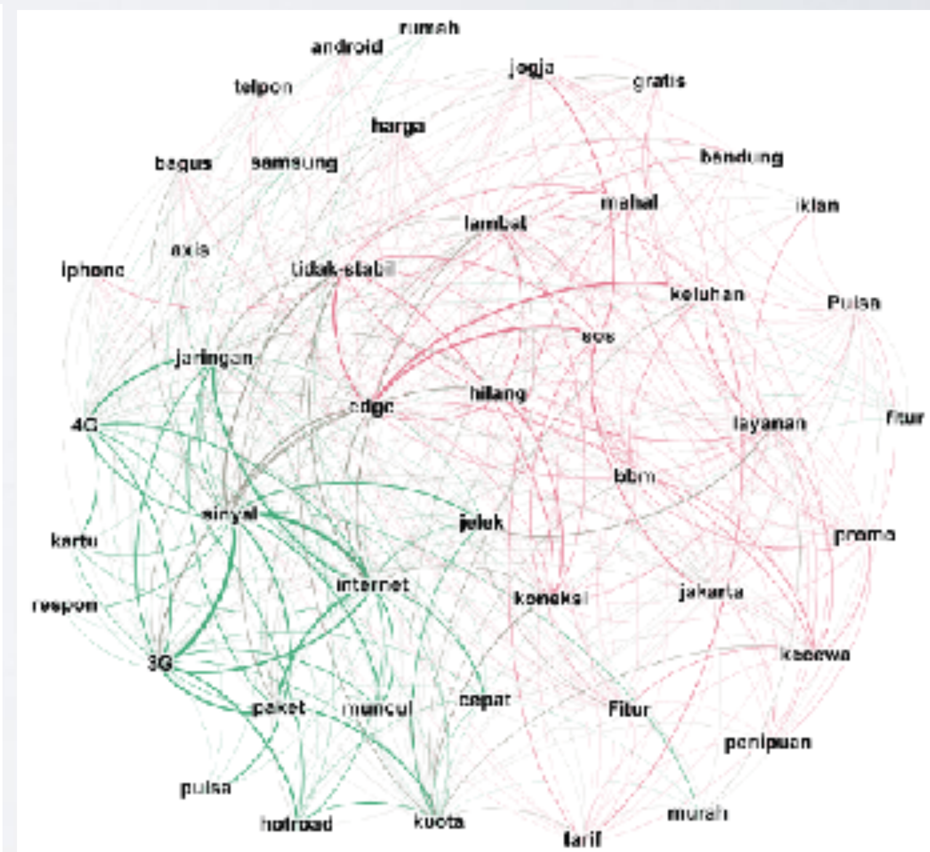


Fig. 4. Network text based on word pair and word cluster of XL

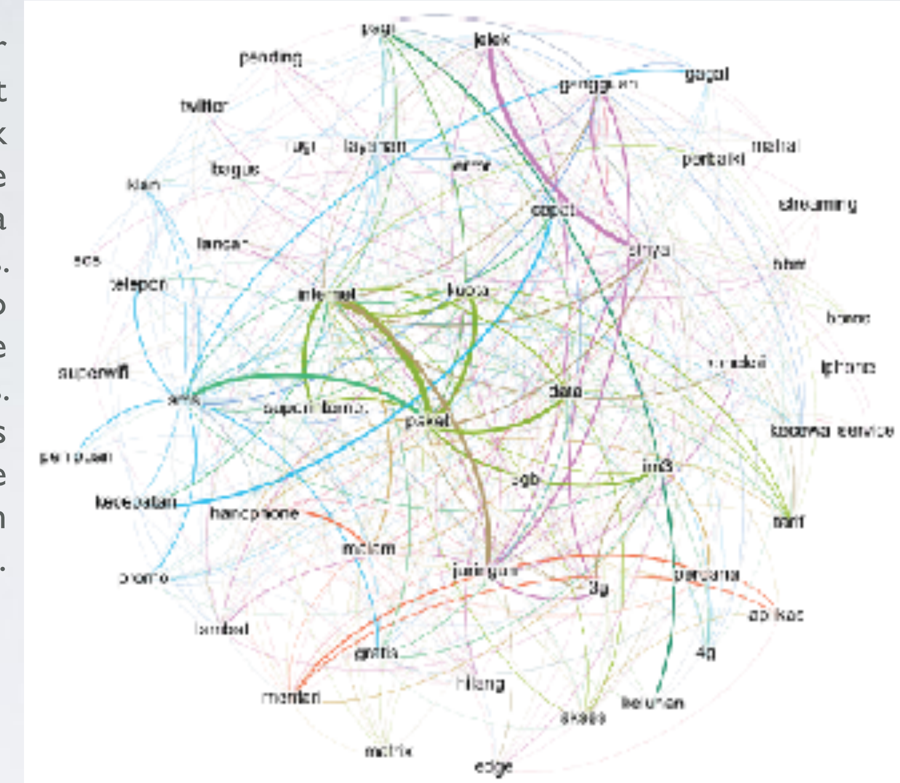
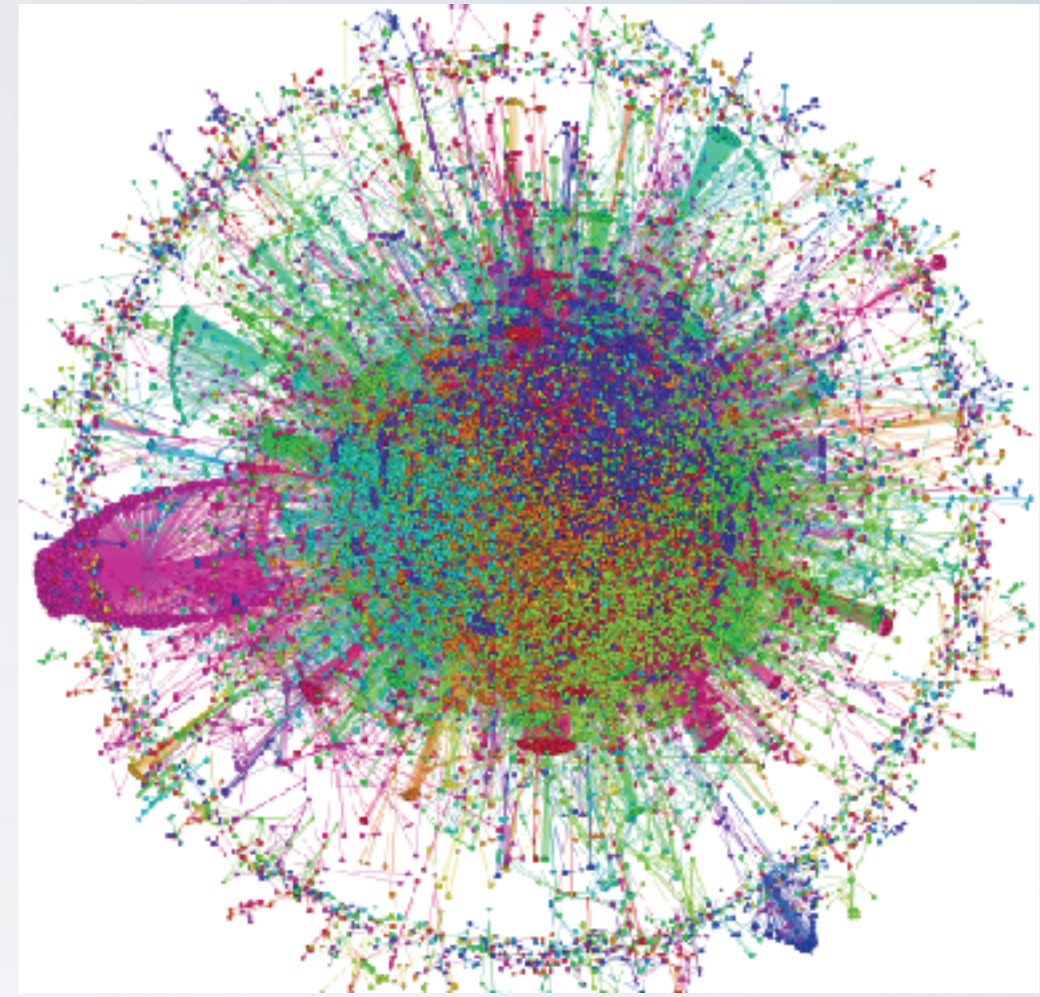


Fig. 5. Network text based on word pair and word cluster of Indosat

EFFECTIVE KNOWLEDGE MANAGEMENT USING BIG DATA AND SOCIAL NETWORK ANALYSIS

*ISCLO, 2013

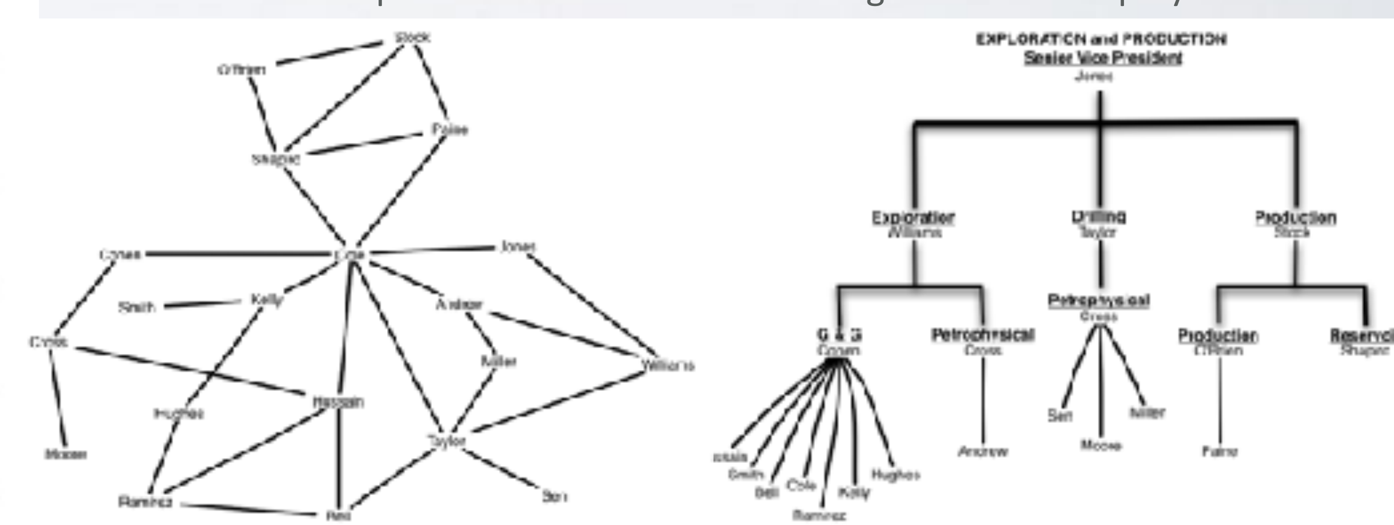
Abstract: Knowledge management consists of identifying, creating, representing, distributing, and enabling adoption of insights and experiences in an organization. One approach of modeling knowledge management is using network model. Big Data is one of important ICT technological roadmap, which main function is modelling behaviour and helping organization decision support. Social Network Analysis is a micro version of Big Data where we can model and establish social network quantification. In this paper we will show how Social Network Analysis can help organization applying Knowledge Management strategies and practices by experiment using real-world large dataset contains 360000+ email exchanges between 36000+ employees inside in an organization



map of full network emaile xchange between employes in Enron

Case	*Questions*	*SNA Tools*
Leader Selection	*Who is the central in the trust and respect network?*	*Degree Centrality*
Ranks	*How do we rank our top performer individuals in the organization?*	*Eigenvector Centrality, Pageranks*
Task Force Selection	*How do we put together a team that maximally connected through out the organization?*	*Closeness Centrality*
Mergers and Acquisition	*How to merge separate cultures / networks?*	*Homophilly, Reciprocity, Mutuality, Transitivity*
Competitive Advantages	*What is the missing links between supply and demand?*	*Structural Holes*
Advertising Attachment	*How strong the impact of our advertisement effort?*	*Tie Strength*
Market Segmentation	*How segmented our market is?*	*Clustering Coefficient, Clique, Cohesive*
Information Dissemination	*How is the information / knowledge spreading?*	*Random Walks, Hits Algorithm*
Strength out the organization	*How to increase redundancy and interconnectedness?*	*Bridge*
Dynamics of Organization	*How dynamics our organization is?*	*Temporal Networks*

business case resolved using SNA methodology



Vizualisation of hierarchical structure organization and knowledge flow of informal organization

MAPPING ONLINE TRANSPORTATION SERVICE QUALITY USING MULTICLASS CLASSIFICATION PROBLEM SOLVING PRIORITIES

CASE STUDY : GOJEK AND GRAB

Abstract. Online transportation service is known for its accessibility, transparency, and tariff affordability. These points make online transportation have advantages over the existing conventional transportation service. Online transportation service is an example of disruptive technology that change the relationship between customers and companies. In Indonesia, there are high competition among online transportation provider, hence the companies must maintain and monitor their service level. To understand their position, we apply both sentiment analysis and multiclass classification to understand customer opinions. From negative sentiments, we can identify problems and establish problem-solving priorities. As a case study, we use the most popular online transportation provider in Indonesia: *Gojek* and *Grab*. Since many customers are actively give compliment and complain about company's service level on *Twitter*, therefore we collect 61,721 tweets in *Bahasa* during one month observations. We apply *Naive Bayes* and *Support Vector Machine* methods to see which model perform best for our data. The result reveal *Gojek* has better service quality with 19.76% positive and 80.23% negative sentiments than *Grab* with 9.2% positive and 90.8% negative. The *Gojek* highest problem-solving priority is regarding application problems, while *Grab* is about unusable promos. The overall result shows general problems of both case study are related to accessibility dimension which indicate lack of capability to provide good digital access to the end

Table 1. Average Number of Discussion in Each Social Media during Observation Time

Company	*Social Media*		
	Twitter	*Facebook*	*Instagram*
Gojek	761/week	140/week	73/week
Grab	402/week	105/week	239/week

Table 2. Transportation Service Quality Dimensions

Service Quality Dimension	*Description*
Availability	The availability of transportation services everytime, everywhere, and in any condition.
Accessibility	The application service ease of use in certain time, condition, and area.
Information	Information availability, well informed customer such as travel fee before making the journey
Time	The detail information about departure time, arrival time, and travel duration.
Customer Service	The capabilities of company to handle complaints, suggestions; The capabilities of company to response customer inquiry in reasonable time. The information about promotional activities.
Comfort	Company effort to provide comfort to customer, such as all-weather protection, vehicles hygiene, and driving style
Safety	Company effort to provide safety and security, such as driver preparation, driving attributes, route knowledge, traffic condition awareness
Environment	Vehicle noise, and vehicles contribution in gas emission

Tabel 3. Preprocessing Example

Raw Data	Kenapa ya susah banget dapet gocar padahal armadanya banyak di sekitarku, @Go jekindonesia? So saaad :(
Tokenization	Kenapa ya susah banget dapet gocar padahal armadanya banyak di sekitarku, @Go jekindonesia? So saaad :(
Filtering	susah banget dapet gocar armada so saad
Stemming	susah sekali dapat gocar armada sangat sedih

Table 4. Example of Tweet Classification based on Transportation Service Quality Dimension

Example Tweets	*Dimension*
Susahnya Go-jek pada jam jam mau berangkat :'	*Availability*
@Go jekindonesia min saya gagal verifikasi kode akun Go jek	*Accessibility*
Karena go-mart cukup mahal	*Information*
Pesen Go-jek abangnya lama banget huhu padahal ga terlalu jauh dari lokasi kasian ibu gue nungguin :(	*Time*
Thanks to gopay gratisan dari Go-jek	*Customer Service*
Pertama kali ini Go jek ada yg cakep plus wangiiiii...sukaaa	*Comfort*
Termengerikan adalah naik Go-jek yg super aneh kayak baru belajar naik motor takut	*Safety*
Go jek parkir semprawut gak jelas sampai di trotoar	*Environment*

Table 5. Example of Tweet Classification by Sentiment

Tweet	Sentiment
Terima kasih bapak Go-jek baik hati yang mau antri lama demi saya ngidam ini	Positive
Go jek ini Tarifnya gak masuk akal	Negative

Table 6. Preprocessing Example

Raw Data	*Clean Data*
App Gojek giliran lagi dibutuhin error hhhhh sebhul.	Aplikasi error sebal

Table 9. The Tweet Sentiment Proportion

Company	*Total Tweet*	*Positive*	*Negative*
Grab	1462	134 (9.2%)	1328 (90.8%)
Gojek	1098	217 (19.76%)	881 (80.23%)

Table 7. Naive Bayes Text Classification Performance

Company	Accuracy	Kappa	Precision	Recall	F-measure
Gojek	93,92%	0.823	88.64%	94.47%	91,46%
Grab	96,58%	0.786	90.77%	88.00%	89,36%

Table 10. *Gojek* Dimension Proportion

Sentiment	*Availability*	*Accessibility*	*Information*	*Time*	*Customer Services*	*Comfort*	*Safety*	*Environment*
Positive	25.3%	1.3%	17.9%	9.6%	18.8%	21.1%	5.5%	0.5%
Negative	24.1%	34.7%	13.6%	8.4%	6.7%	7%	4.9%	0.6%

Table 8. Support Vector Machine Text Classification Performance

Company	*Accuracy*	*Kappa*	*Precision*	*Recall*	*F measure*
Gojek	97,26%	0.910	97,66%	93,66%	95,61%
Grab	97,26%	0.809	98,54%	85,00%	91,27%

Table 12. *Grab* Dimension Proportion

Sentiment	*Availability*	*Accessibility*	*Information*	*Time*	*Customer Services*	*Comfort*	*Safety*	*Environment*
Positive	1.5%	3.8%	21.4%	6.6%	13.3%	45.3%	7.4%	0.7%
Negative	10.3%	23%	18.4%	3.4%	33.7%	7.1%	3.6%	0.5%

DIRECT COMPARISON METHOD OF INFORMATION DISSEMINATION USING LEGACY AND SOCIAL NETWORK ANALYSIS

*ICST, 2017

Abstract: Big data is a new phenomenon that force organization to find new way to process, to save, and to analyze data in various forms. Big data challenge is getting bigger and faster but also the opportunities of harvesting information from it. Organization prefers using big data's technology for many advantages, mainly because it is cheaper, effective, efficient and faster than previous system. In some research area, there are conflicting approach to solve the problem by using legacy methodology or by using big data approach. One of the fastest methodology in big data approach is Social Network Analysis. One of research area measured as the research object in this paper is information dissemination. Therefore, this paper compares information dissemination measurement using legacy methodology and social network analysis. Legacy methodology draws up a questionnaire or interview to obtain the data about individual or sampleopinion. Meanwhile, Social Network Analysis use conversational data in social media. We pick a popular information hub twitter account as a case study. Both legacy and social network methods are applied. The result shows there are differences on how to collect the data, how to pose the research question and what issues are answered by each methodology. In the end, we conclude whether both methods are competing or completing each other.

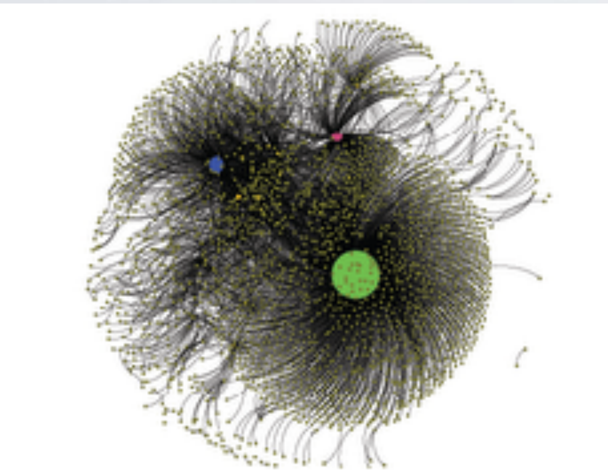


Fig 3. *Infobdg's* Network Model

TABLE VIII TOP TEN THE HIGHEST CENTRALITY

Rank	*Degree Centrality*	*Betweenness Centrality*	*Closeness Centrality*
1	Infobdg	Infobdg	Infobdg
2	PRFMnews	PRFMnews	Infobdgkuliner
3	BandungCC2016	BandungCC2016	Festcitylink
4	Sekitarbdg_	Sekitarbdg_	MorningBandung
5	Infobandung	RadioElshinta	SindotrijayaFM
6	JalinBDG	Dishub_kotabdg	Aheryawan
7	Dishub_kotabdg	Infobandung	Dhika_cep
8	ARNOLD211D	Infoll	OdedMD
9	RadioElshinta	Infobdgkuliner	C0necy
10	Tmc_restabesbdg	RTMC_PoldaJabar	verlexdvd

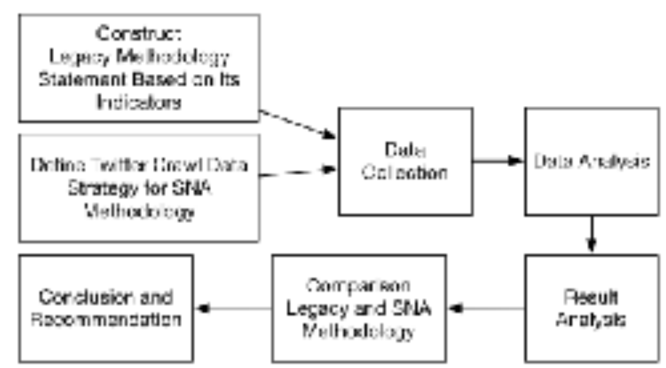


Fig 2. Research Workflow

TABLE VI AVERAGE PERCENTAGE OF MARKETING COMMUNICATION EFFECTIVENESS

No	Dimension	Indicator	Percentage
1	Attitude Toward The Advertising Message	Favorable (P1)	74.47%
		Interesting (P2)	77.85%
		Impressive (P3)	69.17%
2	Message Sharing Intention	Likely To Share (P4)	68.63%
		Likely To Post (P5)	60.96%
Total			351.08%
Average Percentage			70.22%

TABLE VII. SINGLE VALUE NETWORK METRICS

Metric	*Value*
Diameter	6
Modularity	0.373 (consist of 3 major groups occupy 86% network size)
Connected Components	30
Average Degree	1.593
Average Path Length	2.619
Clustering Coefficient	0.166
Density	0.0005689695

TABLE IX COMPARISON OF LEGACY AND SNA METHODOLOGY

Process	*Legacy Methodology*	*SNA Methodology*
Data Source	Respondents' answer about their perception.	User interaction data on Twitter.
Data Collection	Questionnaires which statements represent indicators of measurement. There are 5 groups of Indicator (*P1,P2, P3, P4, P5*)	Keywords for crawling are *#infobdg*, *#infobdgkuliner*, *#ialinbdg* and *#euacabdg*. 10.290 tweets collection in 1 month duration (January - February 2017)
Measurement Indicator	Measure the effectiveness of marketing communication through its 5 groups indicators	Measurement based on network metrics such as *size, centrality, density, network diameter, modularity, connected components, average degree, average path length*, and *clustering coefficient*.
Validity and Reliability Test	Validity and Reliability test applied to measure the correctness of questionnaire indicators	There are no Validity and Reliability test applied since we collect from population data. The correctness of our interpretation depends on preprocessing and data collection rules of crawling
Analysis	Descriptive analysis of data processed to measure the level of effectiveness	Descriptive analysis of the relations of each network metrics. they describe the structure and dynamics of the network, thus we can understand the underlying process of dissemination.
Result	The effectiveness of *Infobdg* marketing communication effort is 70.22%. The highest indicator is "interesting" (*P2*) with the value of 77.85%.	*Infobdg* network reveals that network is too sparse to bolster up the dissemination process. Furthermore, there are some influential users that can be endorsed to promote the cause. Group information predicts which part of a network needs to be improved.

conclusion

The legacy methodology is based on sampling approach, while SNA methodology is based on the population data. Both methodology answer different research questions. Legacy methodology measures the effectiveness of information dissemination effort and which indicators of the effectiveness perform worst to the best. SNA methodology measures the dynamics of conversations or exchange information network that become the underlying process of information dissemination. Both methods are not competing but completing

MONTE CARLO SIMULATION AND CLUSTERING FOR CUSTOMER SEGMENTATION IN TELCO BUSINESS

*ICST, 2017

Abstract: *Utilizing data for segmentation analysis can bring a streamlined way to get potential insight as of decision makingsupport in a business organization. Using appropriate data analytical technique help the organizations in profiling their customer segments accurately. The result brings an effective marketing strategy. However, there are times in doing data analytic, the organization needs another variable of data where the value is unavailable, for example: customer's income data which mostly hard to collect. By using Monte Carlo simulation, the value of customer's income can be generated and then compared with customer spending to construct customer segmentation model. An unsupervised learning for customer segmentation model using K-Means clustering enables us to see the grouping patterns of customer's income towards their spending. Clusters of the dataset might be interpreted as a group of customers that having a similar character. This paper shows us how to generate customer's income data and create data cluster to optimizing customer potential by utilizing data. Furthermore, the result brings us insight into which group of the customer might unserved properly considering their average income with their spending behavior.*

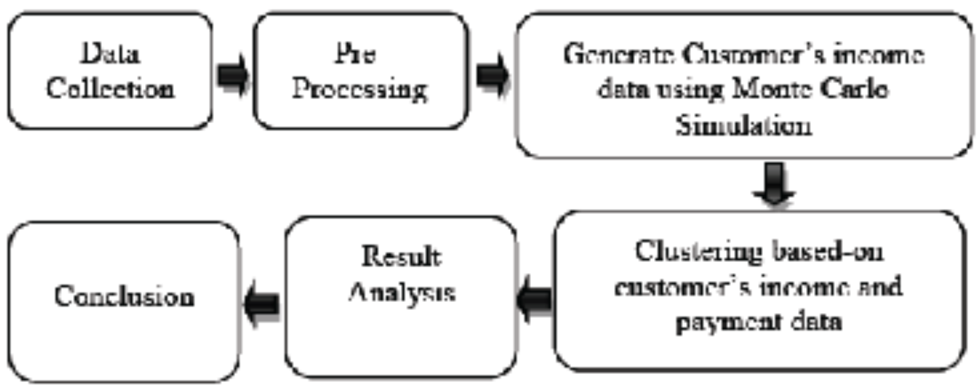


Fig. 1 Research Workflow

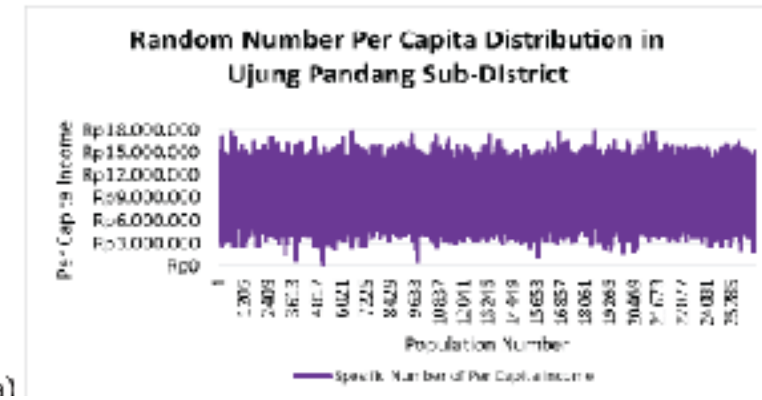


Fig. 2. Distribution of per capita income in Makassar city, (a) Ujung Pandang Sub-district; (b) Wajo Sub-district.

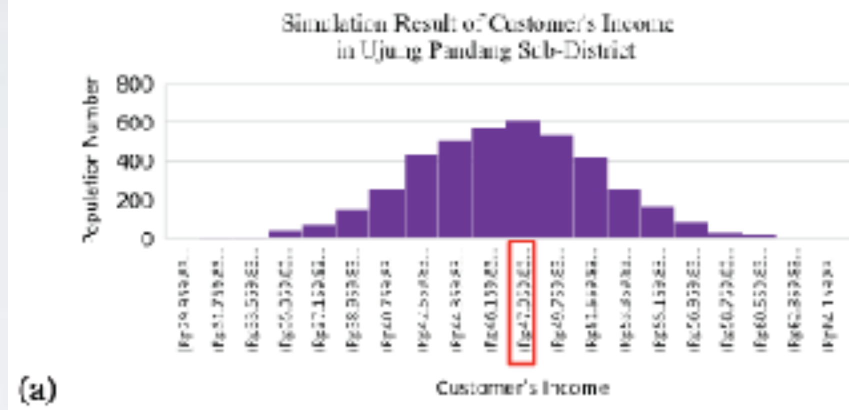


Fig. 3. Simulation Result of Customer's Income in Makassar city. (a) Ujung Pandang Sub-district; (b) Wajo Sub-district

TABLE I MEAN AND STANDARD DEVIATION OF PER CAPITA INCOME FOR EACH SUB-DISTRICT

Sub-Districts	*Average per Capita Income ($\bar{x}$) (in Rp.)*	*Standard Deviation (σ)*
BIRINGKANAYA	3.308.472	0.7
MANGGALA	6.380.478	1.3
PANAKKUKANG	8.357.546	1.9
BONTOALA	2.119.381	0.4
WAJO	5.599.512	1.3
UJUNG PANDANG	9.601.037	2.2
MAKASSAR	3.248.984	0.7
RAPPOCINI	5.315.408	1.1
TAMALATE	2.331.208	0.5
MAMAJANG	3.060.287	0.6
MARISO	6.450.864	1.5
UJUNG TANAH	5.579.343	1.3
TALLO	5.729.780	1.3
TAMALANREA	9.431.498	2.2

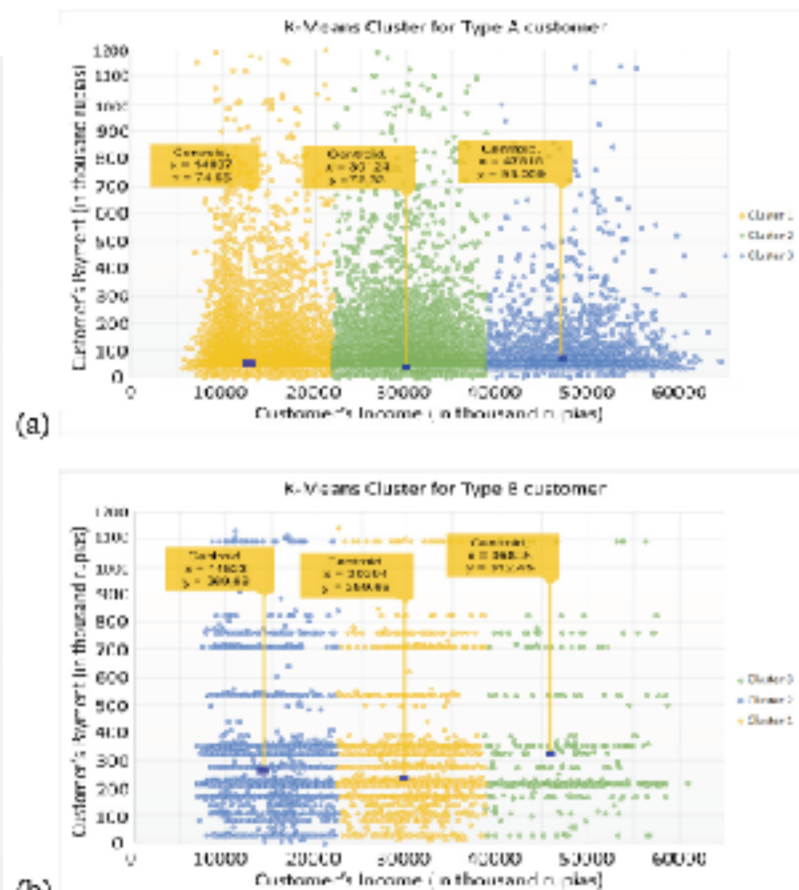


Fig. 4. K-Means Clustering Plot of (a) Type A Customer. (b) Type B Customer. (c) Type C Customer.

TABLE IV. INCOME LEVEL FREQUENCY

Category	*Income Range (in Rp)*	*Cluster*	*Frequency*
Low income	< 23 million	Cluster 1 type A	47.7%
		Cluster 1 type B	
		Cluster 1 type C	
Middle income	23 million – 38 million	Cluster 2 type A	44.8%
		Cluster 2 type B	
		Cluster 2 type C	
High income	> 38 million	Cluster 3 type A	7.5%
		Cluster 3 type B	
		Cluster 3 type C	

TABLE V. SPENDING LEVEL FREQUENCY

Category	*Spending Average (in Rp)*	*Cluster*	*Frequency*
Slight user	0 - 150.000	Cluster 1 type A	47.1%
		Cluster 2 type A	
		Cluster 3 type A	
Moderate user	0 - 400.000	Cluster 1 type B	10.1%
		Cluster 2 type B	
		Cluster 3 type B	
Heavy user	0 - 600.000	Cluster 1 type C	42.8 %
		Cluster 2 type C	
		Cluster 3 type C	

REMARK

- ethical issue
- credibility of the result of machine learning (different algorithm, different conclusion)
- analytics is multidisciplinary
- how to assess risk, especially concerning public issues (unstructured data)
- for traditional structured data, it is a matter of parallel computation / processing based on data mining model

more data means less privacy, more speed means less accuracy, more autonomy means less control


without **Big Data**, you are blind
and deaf in the middle of a freeway
- Geoffrey Moore -