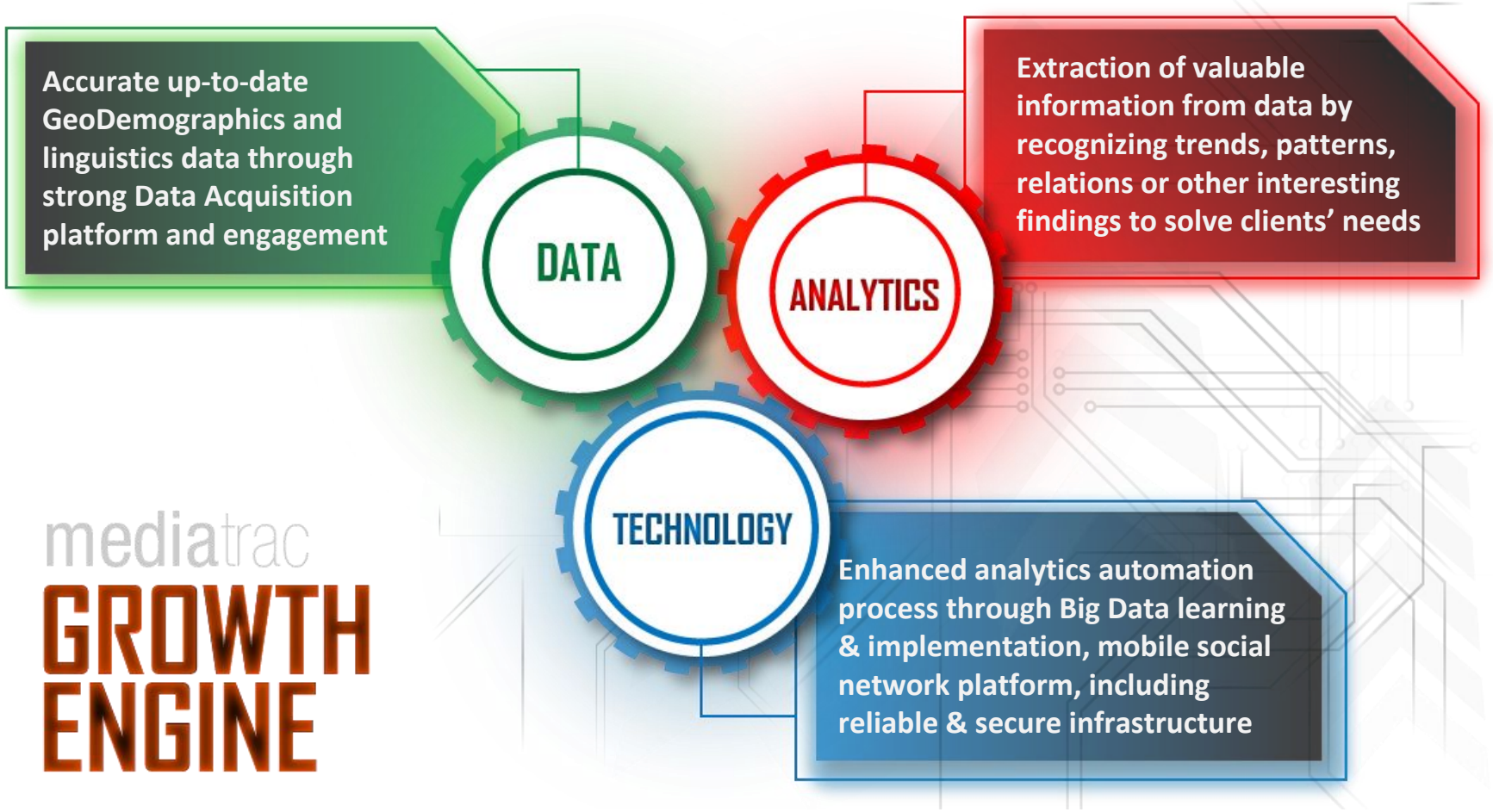
mediatrac

BIG DATA ANALYTICS FOR PUBLIC SERVICE IMPROVEMENT

by imron zuhri, mediatrac

presented for big data workshop in bandung. 26 September 2014

mediatrac



mediatrac

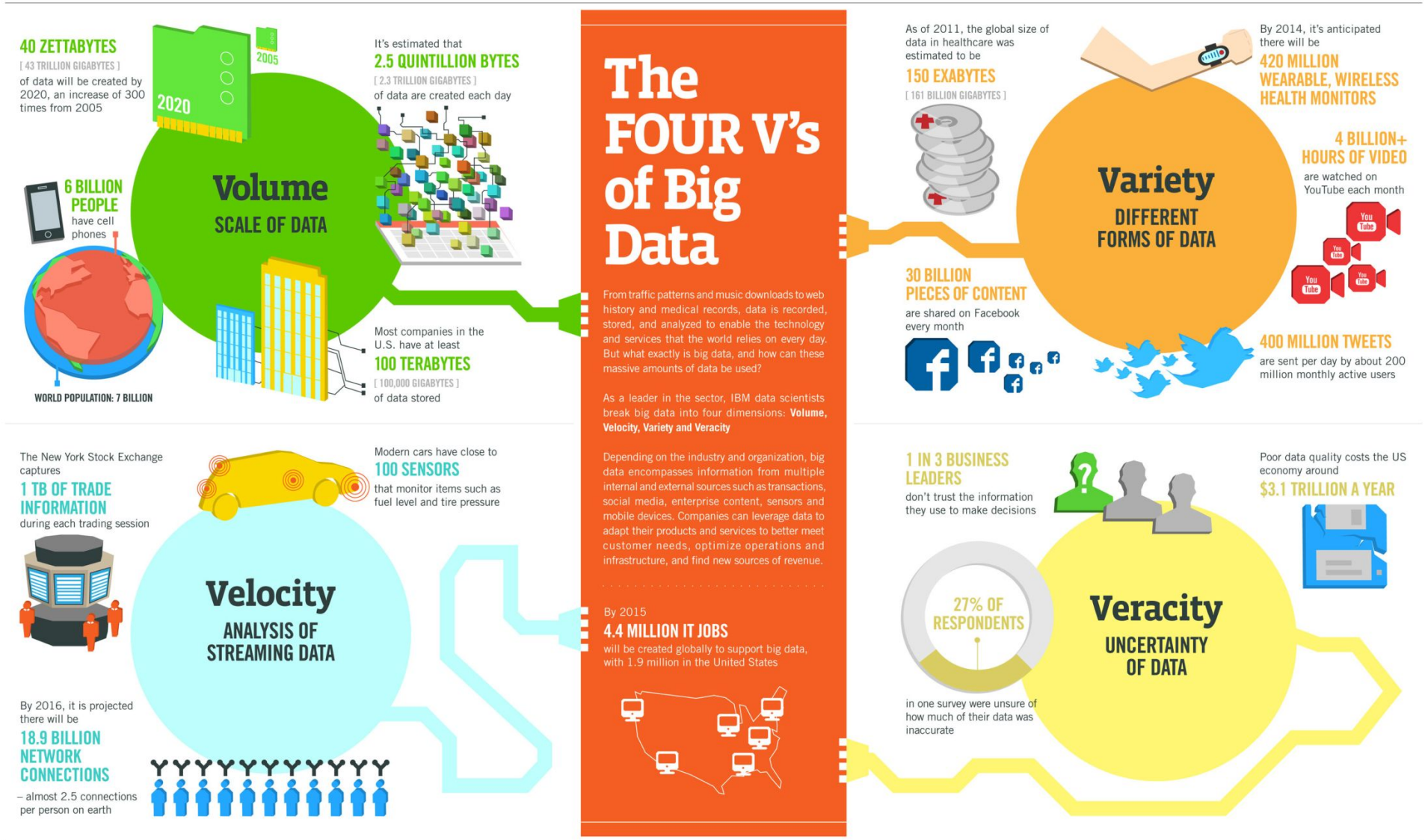
GROWTH ENGINE

WHAT IS BIG DATA?

IT'S ABOUT MAKING DECISIONS
BASED ON DATA TO INCREASE VALUE

SO, WHAT'S NEW?

WHEN TO USE BIG DATA



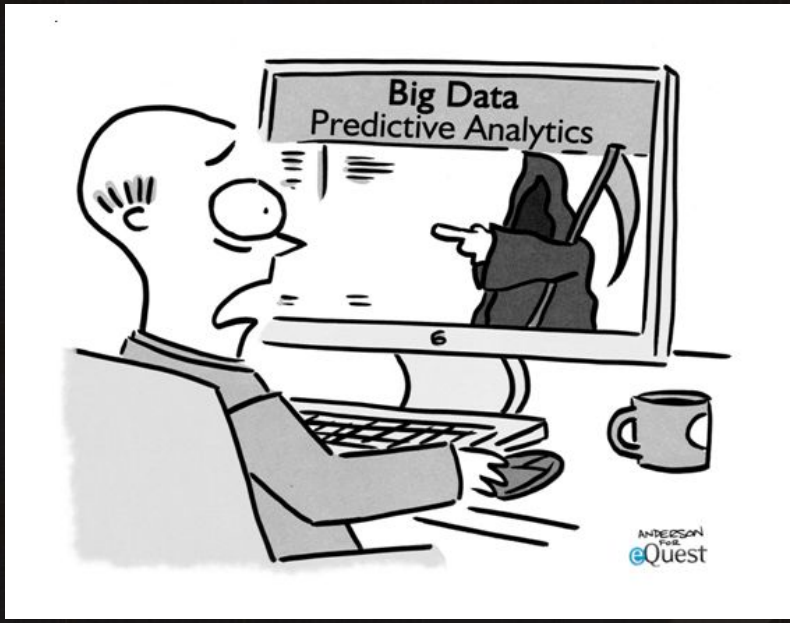
THE OBSTACLES OF IMPLEMENTING BIG DATA

- busines s case
- data
- techno logy
- people
- organi zation

busines
s case

BIG DATA ANALYTICS FUNCTIONS

1. **Descriptive Analytics,**
   can document and convey what is happening
2. **Diagnostic Analytics,**
   might shed light on why things may happen
3. **Predictive Analytics,**
   could give us a sense of what is likely to happen
4. **Prescriptive Analytics,**
   will prescribes what to do to take advantage of the predicted future, and optimize your value creation process



BIG DATA APPLICATIONS

1. **Real-Time Awareness,**

   i.e., how "Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies."

2. **Early Warning,**

   i.e., the "early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis."

3. **Real-Time Feedback,**

   i.e., "the ability to monitor a population in real time makes it possible to understand where policies and programs are failing and make the necessary adjustments."

4. **Planning and Optimization,**

   i.e., "the ability to show what would likely to happen to a specific population makes it easier to plan and optimize certain policies and programs."

real time awareness: farming

Farming information System

- Big data analysis can increase crop yields by helping farmers make better decisions about when to plant, manage and harvest their crops. Beyond broad data sets on topics such as rainfall levels, signs of pests and diseases, and anticipated prices at local markets, there is also the highly specialised and specific data sets such as plant genomics and local weather conditions.

Soil maps and weather simulation

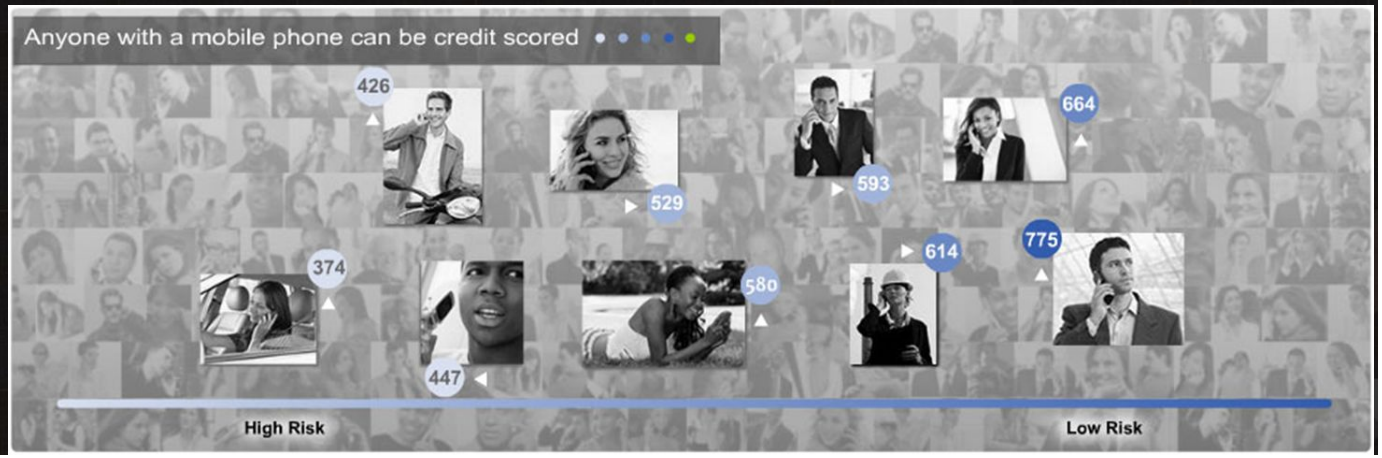
- The Climate Corporation operates a cloud-based farming information system that takes weather measurements from 2.5 million locations and combines it with 150 billion soil observations to generate 10 trillion weather simulation data points. This information allows farmers to know information as diverse as when is the best time to spray fields to getting an accurate estimate of the value of fields they may be considering buying.

Crowdsourcing farming data

- The CGIAR Consortium in France hopes to develop an app which uses all the available data to allow African farmers to identify their local soil type, the planting and harvesting requirements of each specific field, and then direct them to where they can locally purchase the seeds needed.

real time feedback: finance

- financial trends
  - "Aid groups are not just tracking … physical phones; they are also starting to watch levels of mobile phone usage and patterns of bill payment, too.
  - If this suddenly changes, it can indicate rising levels of economic distress, far more accurately than, say, GDP data. … [If any organization] spots a sudden increase in certain keywords, this can also provide an early warning of distress."
- credit rating alerts for the poor
  - Cignifi, a Brazilian startup, for example developed a technology to recognize patterns in the usages of mobile devices.
  - The system recognizes phone-calls, text messages and data usage and based on this information it can recognize someone's lifestyle and his/her corresponding credit risk profile.

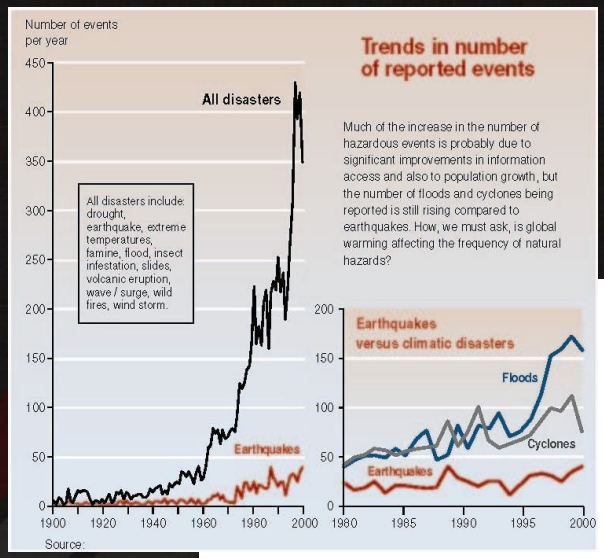


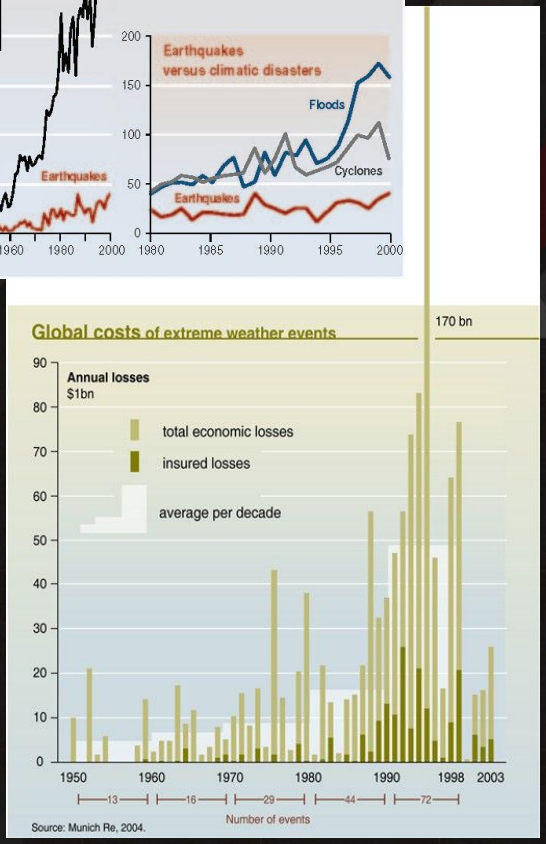
mediatrac

early warning system (ews)

- poverty detection/prevention
- famine, and food security
- climate change
- farming
- fertility and population control
- disaster and risk management
- education resources
- energy consumption
- health and nutrition
- water and sanitation
- voice and accountability
- safety net and transfers
- issues and potential crisis monitoring and management

sample ews: natural disaster

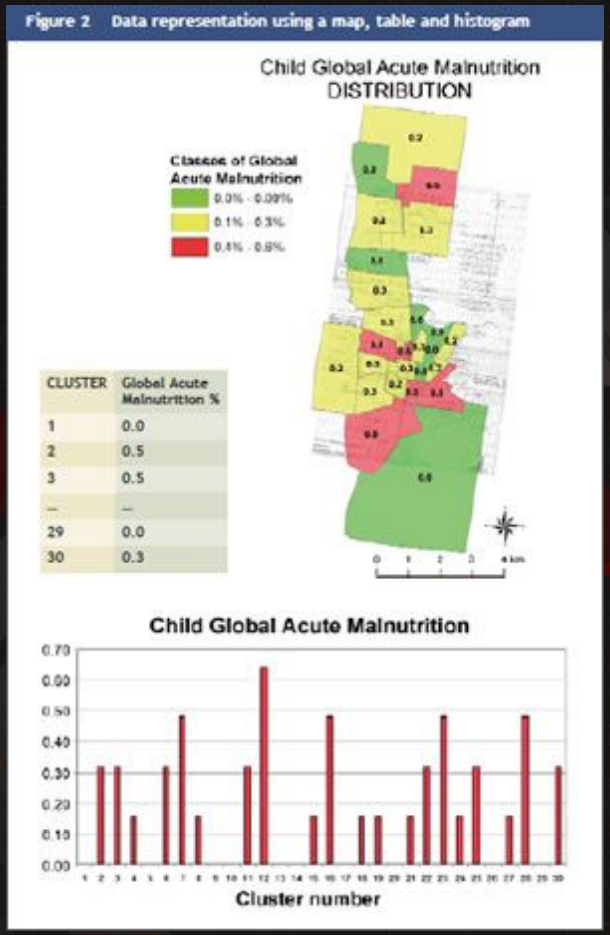


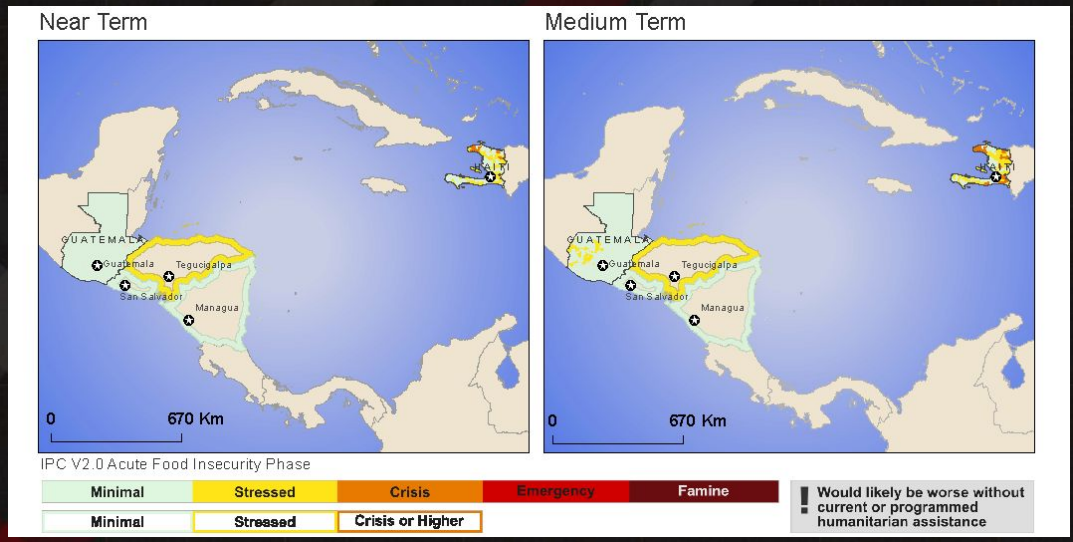


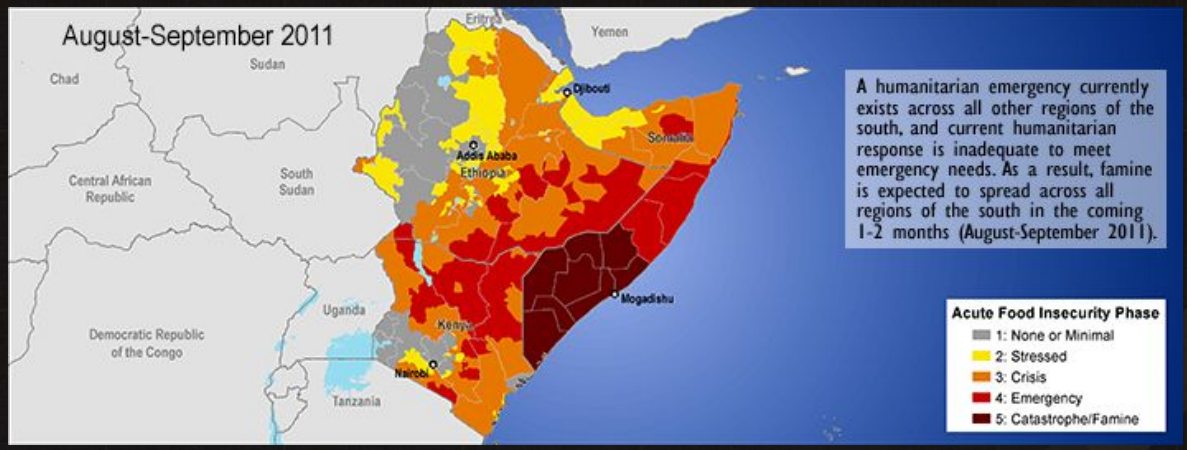
Mobilising Data to Deal with an Epidemic

In the wake of Haiti's devastating 2010 earthquake, researchers at the Karolinska Institute and Columbia University demonstrated that mobile data patterns could be used to understand the movement of refugees and the consequent health risks posed by these movements. Researchers from the two organisations obtained data on the outflow of people from Port-au-Prince following the earthquake by tracking the movement of nearly two million SIM cards in the country. They were able to accurately analyse the destination of over 600,000 people displaced from Port-au-Prince, and they made this information available to government and humanitarian organisations dealing with the crisis. Later that year, a cholera outbreak struck the country and the same team used mobile data to track the movement of people from affected zones. Aid organisations used this data to prepare for new outbreaks. The example from Haiti demonstrates how mobile data analysis could revolutionise disaster and emergency responses.

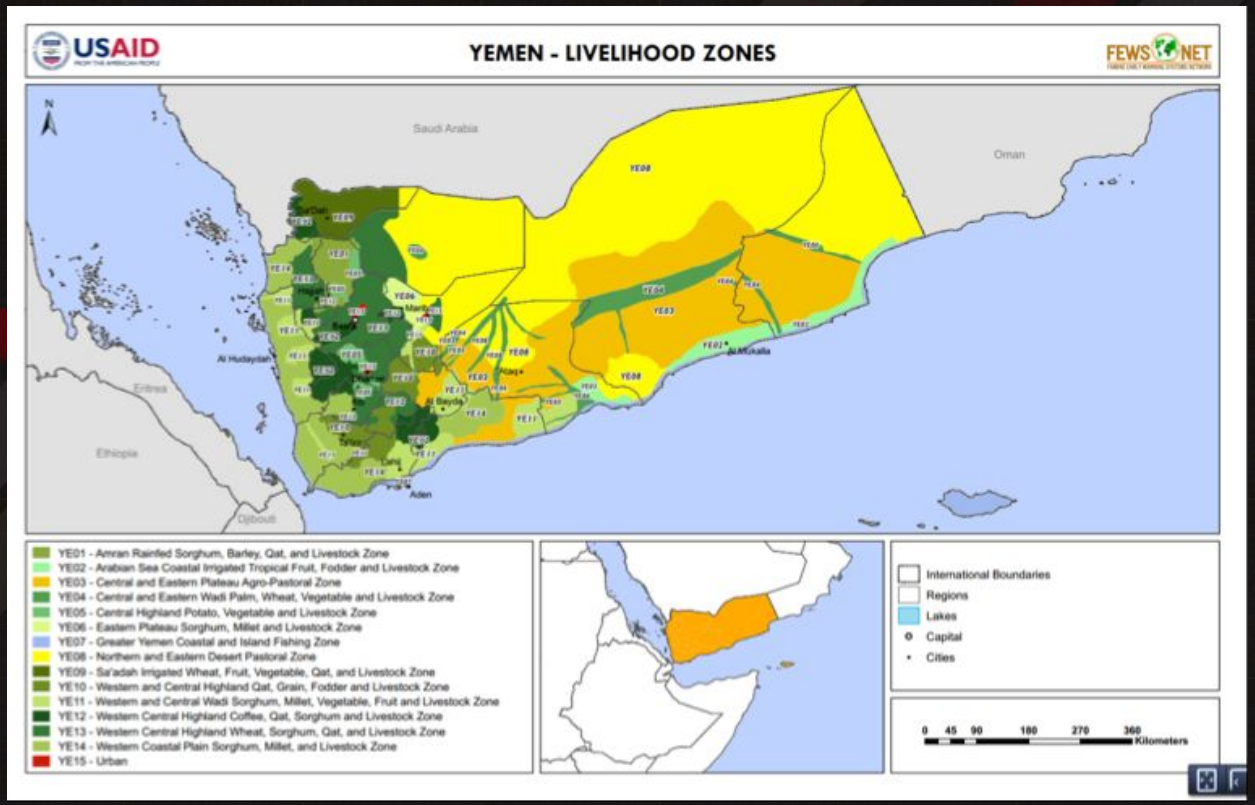
sample ews: nutrition and food security







sample ews: famine



Using big data to aid the battle against hunger isn't a new idea. The Famine Early Warning System has been in operation for 25 years to help international aid groups predict where famines in remote regions are about to occur and thus target the $1.5 billion of annual food aid from the U.S. Agency for International Development.

The system relies on a blend of social and scientific big data from federal agencies as diverse as NASA, the National Oceanic and Atmospheric Administration, and the Department of Agriculture to create hydrological models, food-economics forecasts, weather and climate simulations, and food-borne illness predictions. The output from these models is increasingly accurate and allows the world's political leaders respond quickly and effectively in the early stages of a famine.

planning and optimization

- better poverty policy segmentation/potential area mapping by unique area characteristic
- better targeting method through not only predictive analysis but policy simulation

"THAT'S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"

data

DATA HYGIENE, THE BIGGEST OBSTACLE OF ALL

data source and protocols

data preprocessing

data transformation and enrichment

data cleansing

data matching and consolidation

SECOND COMMON PROBLEM

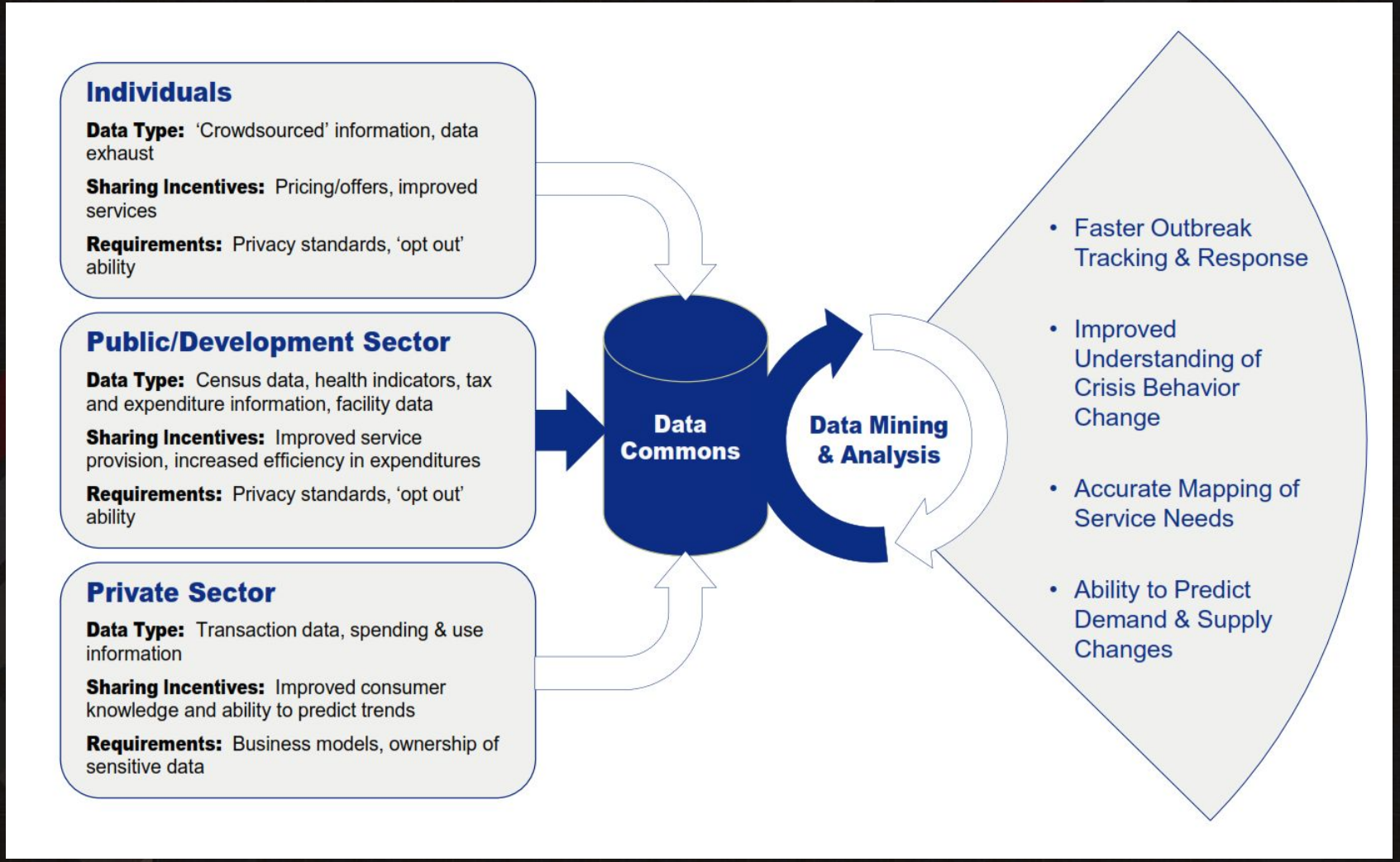
...

NO BIG DATA

possible initiatives

- single data repositories using big data environment
- identity matching across all public data for individual and critical entities
- single 360 view of development subject

framework for data commons



sample: web scraping

Indonesian Rice Crisis Tracking

Team Ndizi's final project looked at rice prices in Indonesia over time. The data was scraped from Carrefour Indonesia, a popular supermarket chain. The team also used the Wayback Machine to go back to historical versions of the website to collect data. Some experimentation was done with pulling prices from Twitter data as well, but there was not enough time to create a full-fledge "universal" scraper from all sources.

Figure 16: World food prices, as reported by major monitoring agencies (green and yellow) vs. prices of two brands of rice in Indonesia (blue and red).

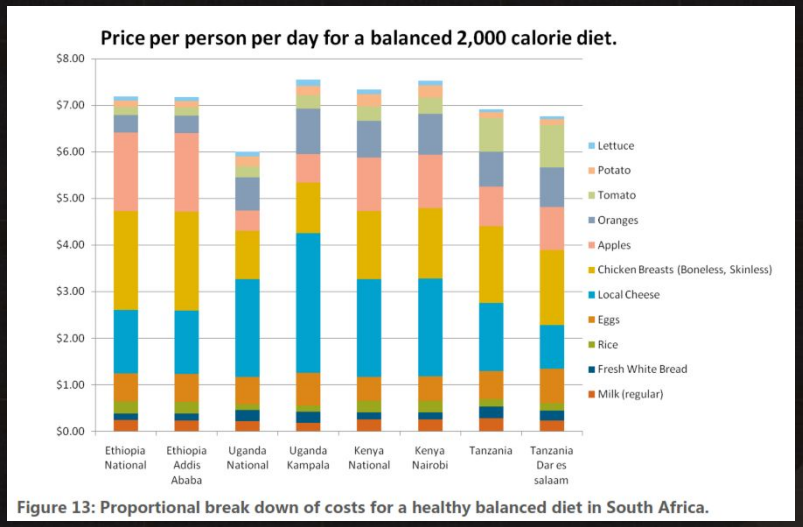


Figure 13: Proportional break down of costs for a healthy balanced diet in South Africa.

sample: mobile surveys

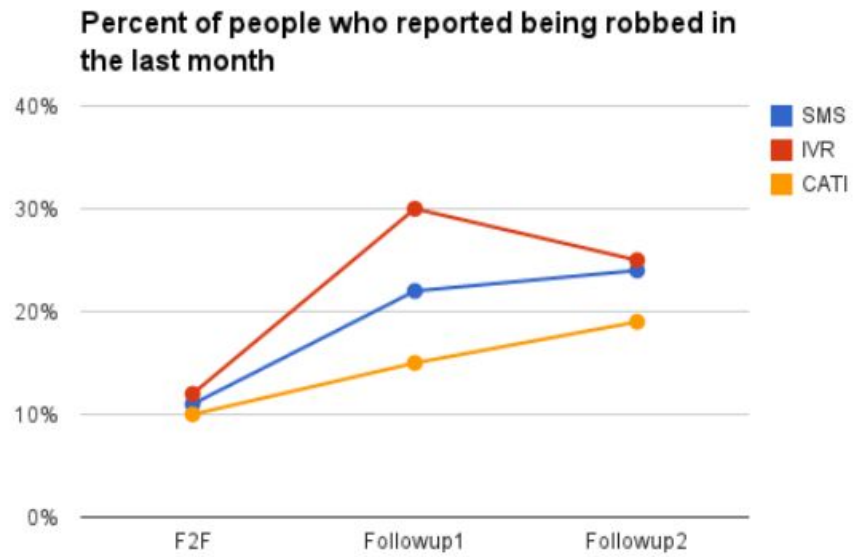


Figure 19: Percentage of "Yes" responses given at each stage of the interview process, broken out by technology.

sample: perceived trust vs credit rating

Borrower Rate	The rate the borrower pays on the loan. The rate is computed as the Lender Rate plus the Group Leader Reward Rate (if applicable) and the Bank Draft Fee Annual Rate (if applicable).
Lender Rate	The rate that lenders receive on the loan.
Loan Amount (in '000)	The requested loan amount in thousands of USD.
"Close Auction when Funded" Indicator	An indicator that equals one if the listing closes as soon as it is funded 100%.
Number of Photographs	The number of photographs associated with a listing.
Number of Words in Listing Text	The number of words used by the borrower in the listing text.
Number of Words in Listing Text (squared)	The square of the Number of Words in Listing Text variable.
Number of Prior Listings	The number of listings submitted prior to the current listing.
Endorsement Indicator	An indicator that equals one if another Prosper member has endorsed the borrower and zero otherwise.
Group Membership Indicator	An indicator that equals one if the borrower is a member of a Prosper group and zero otherwise.
Group Leader Reward Rate	The percentage reward which is kept by the group leader. The variable is zero if the borrower is not a member of group.
Listing Start Date	The date at which a listing was created.
Bank Draft Fee Annual Rate	The rate charged by the bank when the payment option selected is not Electronic Funds Transfer.
Default Indicator	An indicator that equals one if the loan status is "Defaulted (Bankruptcy)", "Defaulted (Delinquency)", "Charge-off", or "4+ months late" and zero otherwise.
Prepayment Indicator	An indicator that equals one if the loan is prepaid in full and zero otherwise.
Loan Origination Date	The date at which the loan was originated.
Loan Age (in days since origination)	The time (in days) between the Loan Origination Date and the date of the last available loan

sample: perceived trust vs credit rating

Table 6
Predicting Default

	1	2	3	4	5	6
Trustworthiness						
TRUST_index _median	0.9833		0.9828		0.9852	
	0.0%		0.0%		0.0%	
HIGHTRUST		0.6699		0.6500		0.7350
		0.0%		0.0%		0.0%
Controls						
ATTRACT_median			0.9777		1.0295	
			80.9%		76.5%	
HIGHATTRACT				0.8472		0.8140
				8.8%		4.2%
Demographic Information	NO	NO	YES	YES	YES	YES
Credit Profile Information	NO	NO	NO	NO	YES	YES
Income & Education Information	NO	NO	NO	NO	YES	YES
Listing & Loan Characteristics	NO	NO	NO	NO	YES	YES

Table 6 presents hazard ratios from a Cox default model. Default occurs when payments on a loan are 4-month or more late. The model is estimated using all 3,291 loans. For each variable, we report the hazard ratio as well as the *p*-value associated with the test whether the hazard ratio is equal to 1. Standard errors are robust to heteroscedasticity. See Table 3 Panel B for a definition of the sets of control variables. See Table 1 for a detailed description of all variables.

sample: poverty vs light intensity

Statistical Analysis of Light Intensity Levels and Poverty Data

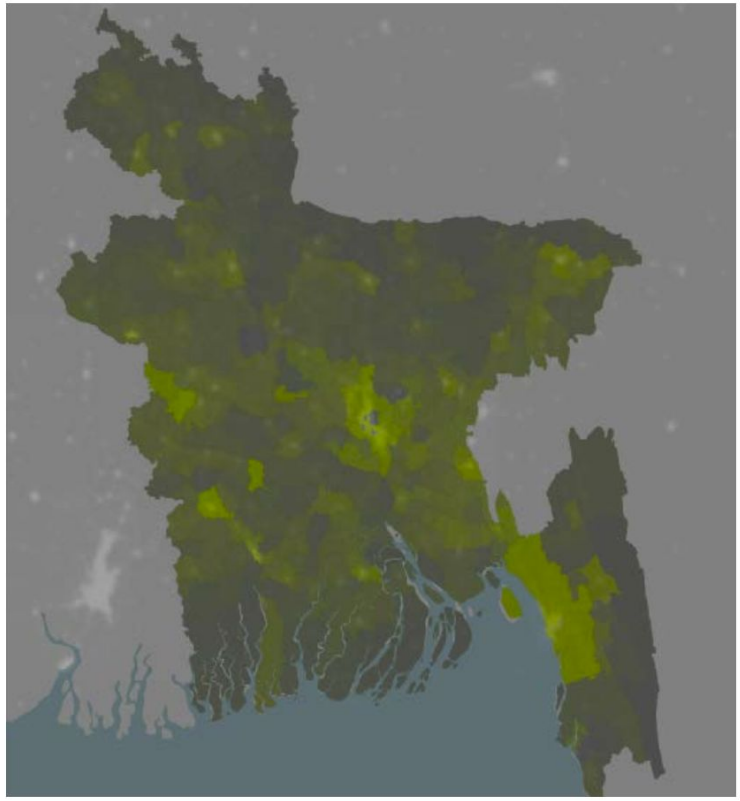


Figure 2: An overlaid map of poverty in 2001 and light intensity.

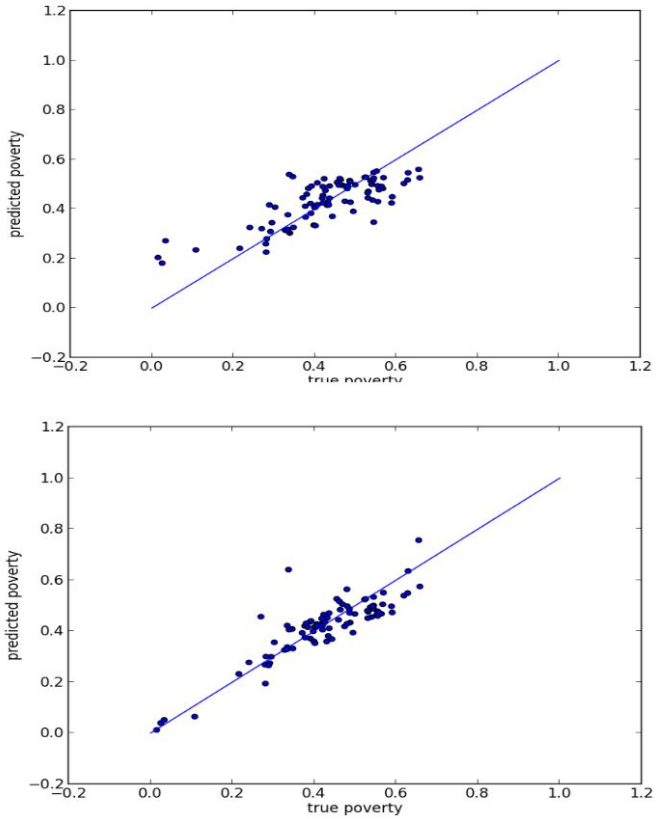
Using the standardized data from both poverty maps of Bangladesh and satellite imagery of nighttime illumination, the team set out to answer the following questions:

- Is poverty correlated with light intensity?
- Are changes in poverty correlated with changes in light intensity?

sample: poverty vs light intensity

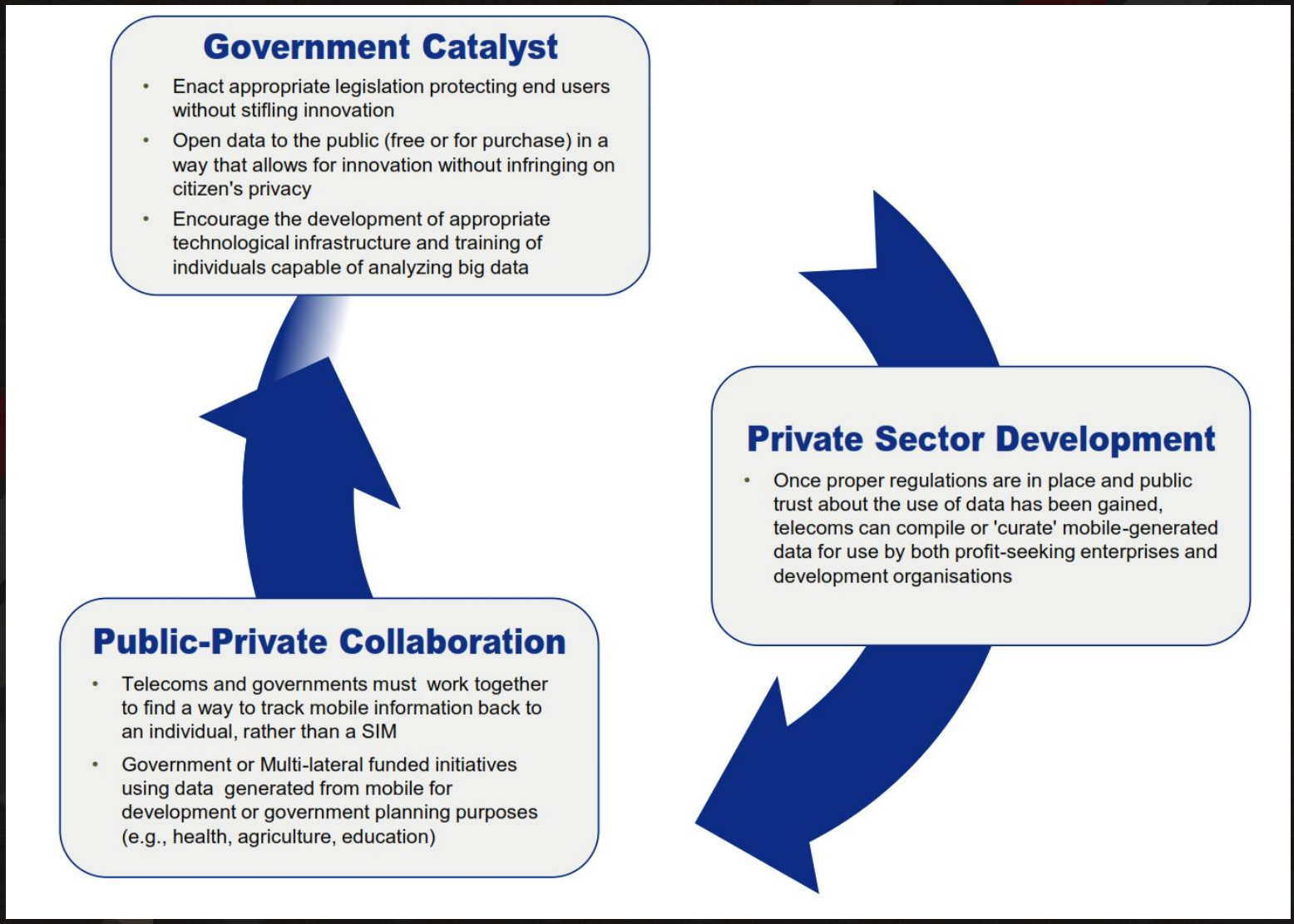
Poverty vs. Light Intensity in 2001

The first model the team built was a linear regression predicting poverty level from light intensity alone using 2001 data. Figure 5 shows a plot of the predicted poverty levels in 2001 and the actual poverty levels using light data alone. All hyper-parameters of the model are selected based on cross validation on 80% of the data. The model is then fit to the same 80% and used to predict the remaining 20%. The team set the alpha parameter to 10.0 for these computations, which is recommended when using 80% of the data as cross-validation.



Figures 5 (top) and 5 (bottom): Predicted poverty vs. true poverty in 2001 for the model fit using light intensity and census data from 2001. (Figure 5: RMSE = 0.067650; Figure 6: RMSE =0.076982).

sample open data initiatives



sample data acquisition devices

MyFarm
- social media for farmers

Aqueduct
- interactive tool that provides high-resolution maps of water-related risks

Drones
- self programmed for soil mapping and crop quality detection

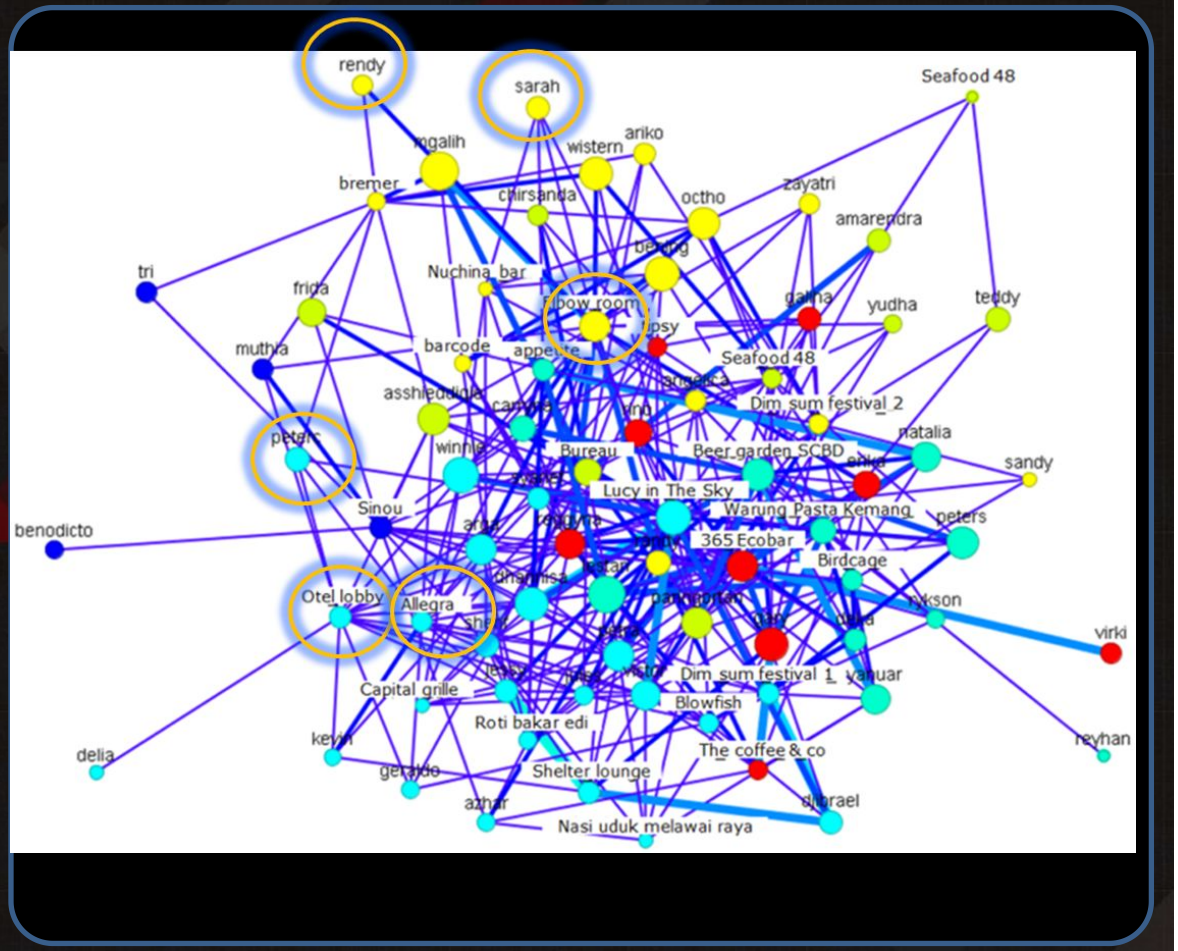
Soil Sensors
- monitor and alerts through your mobile phones

data integration

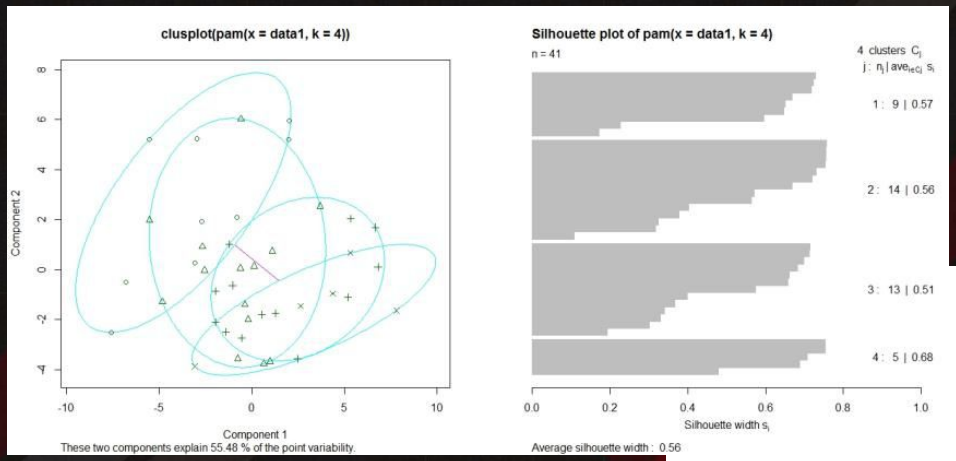
power meter

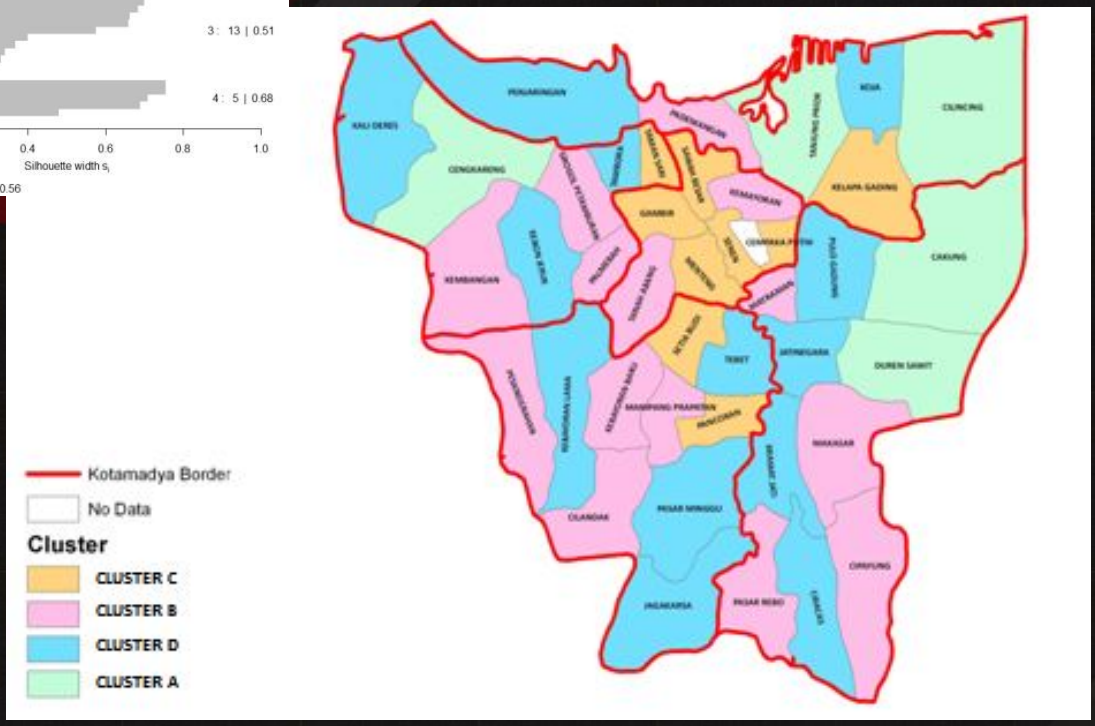

1%
1%
1%
0%
0%
0%
1
Series44
Series43
Series42
Series41
Series40
Series39
Series38
Series37
Series36
Series35
Series34
Series33
Series32
Series31
Series30
Series29
Series28
Series27
Series26
Series25
Series24
Series23
Series22
Series21
Series20
Series19
Series18
Series17
Series16
Series15
Series14
Series13
Series12
Series11
Series10
Series9
Series8
Series7
Series6
Series5

individual behavior tracking

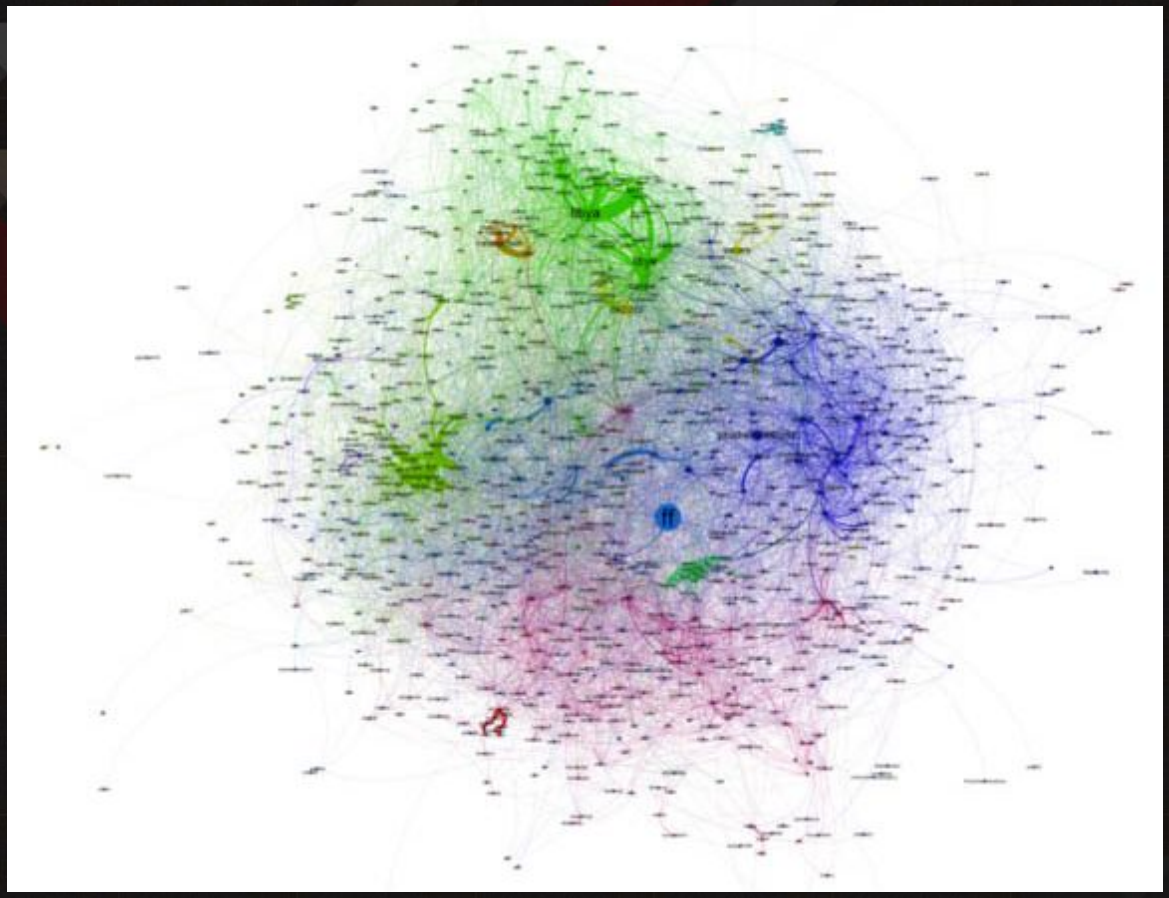


geo-demographic segmentation





contextual recommendation engine



brand relationship network – flu medicine

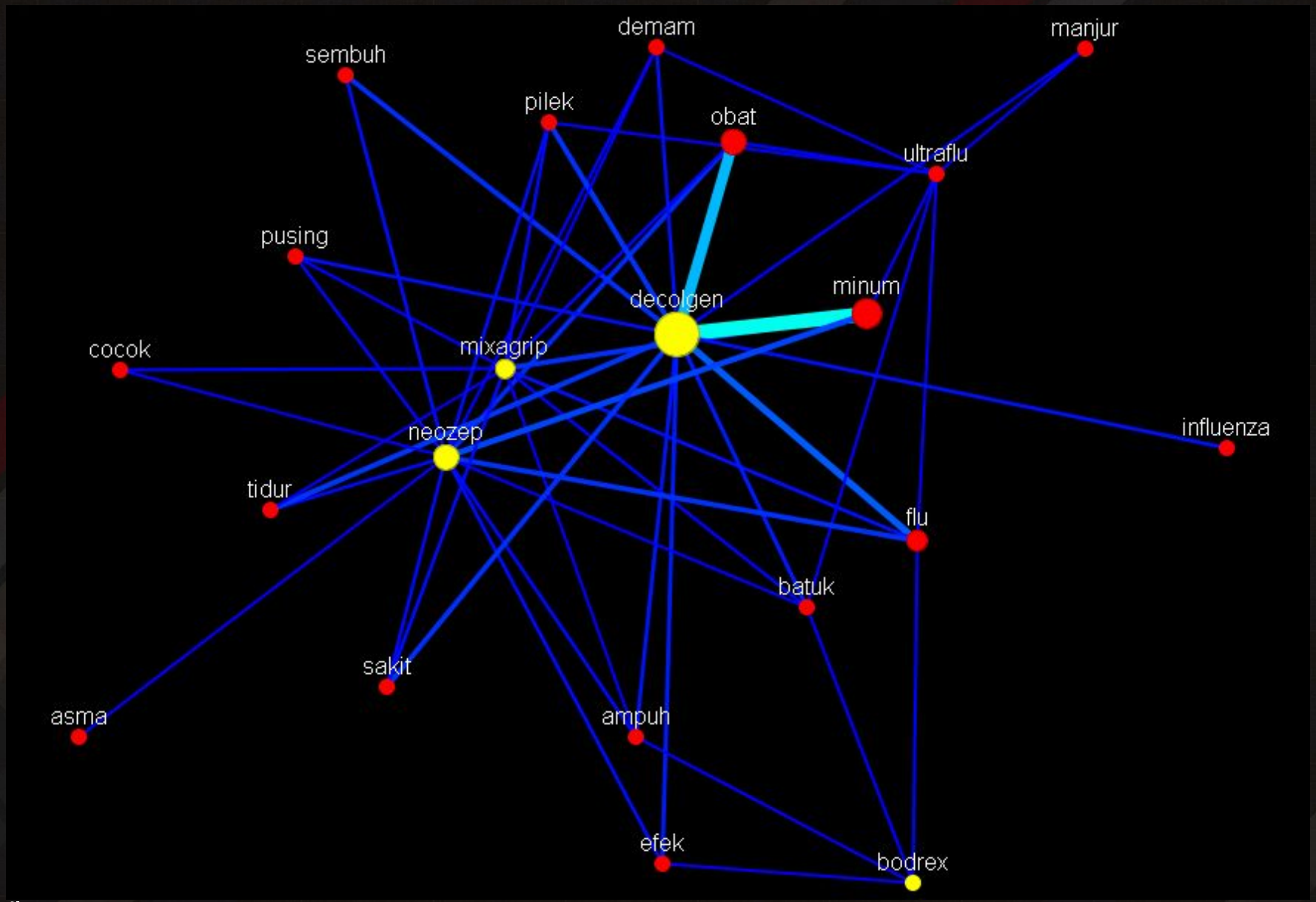
Brand	Volume
decolgen	1836
neozep	568
mixagrip	283
ultraflu	52
bodrex	27

Brand	Centrality	Rank
decolgen	0,063	2,820
neozep	0,063	2,817
mixagrip	0,055	2,144
ultraflu	0,036	1,408
bodrex	0,020	0,865

Brand	Volume	Rank	Score
decolgen	1836	2,820	30
neozep	568	2,817	21
mixagrip	283	2,144	15
ultraflu	52	1,408	9
bodrex	27	0,865	7

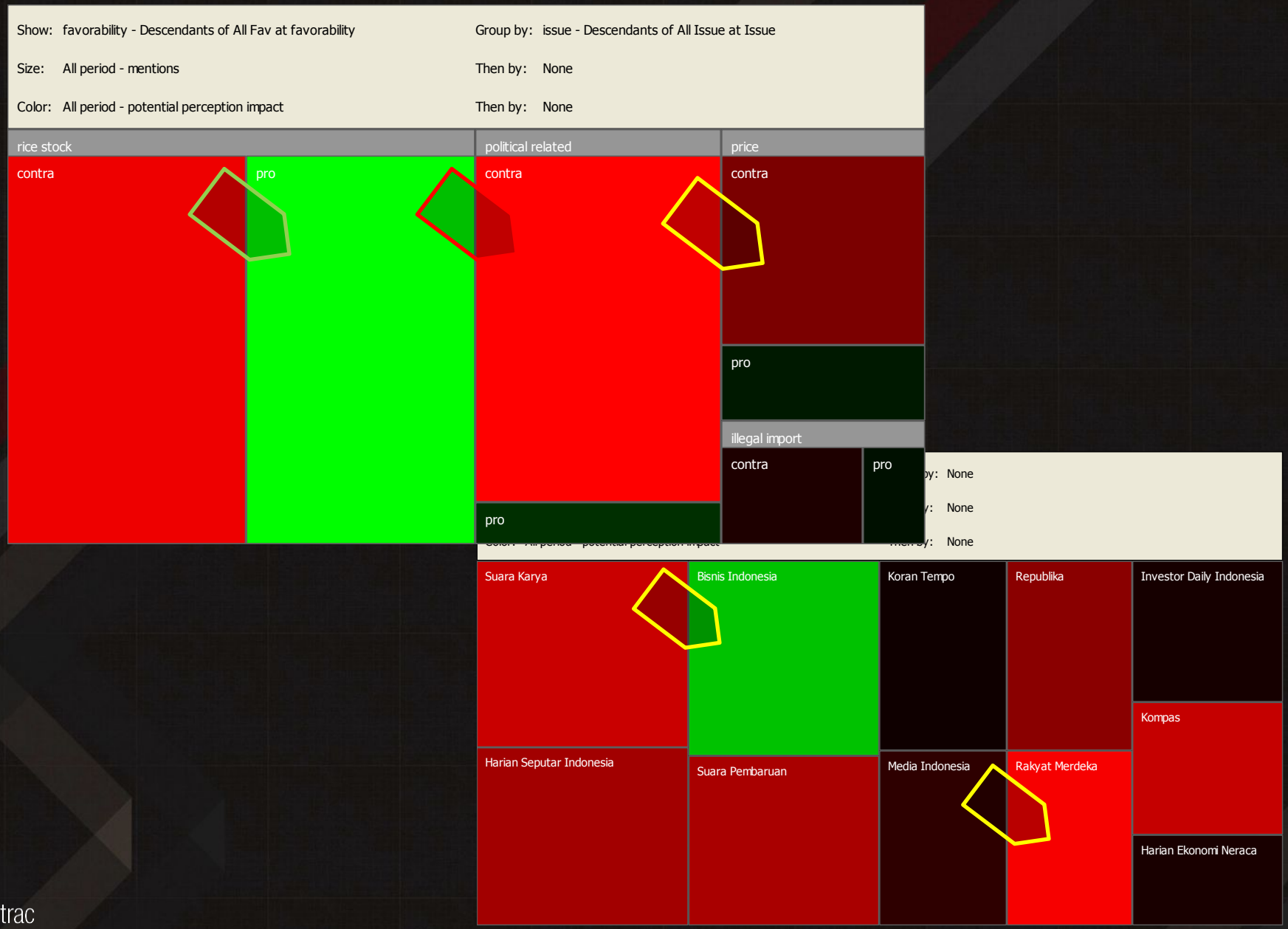
Attributes	Centrality	Rank
minum	0,061	0,829
obat	0,061	0,829
flu	0,067	1,013
tidur	0,051	0,658
pilek	0,061	0,829
sakit	0,051	0,658
sembuh	0,036	0,492
efek	0,041	0,676
batuk	0,067	1,013
ampuh	0,057	0,842
influenza	0,018	0,321
demam	0,061	0,829
pusing	0,051	0,658
cocok	0,033	0,487
manjur	0,028	0,492
asma	0,018	0,321

brand relationship network at 10,000 threshold

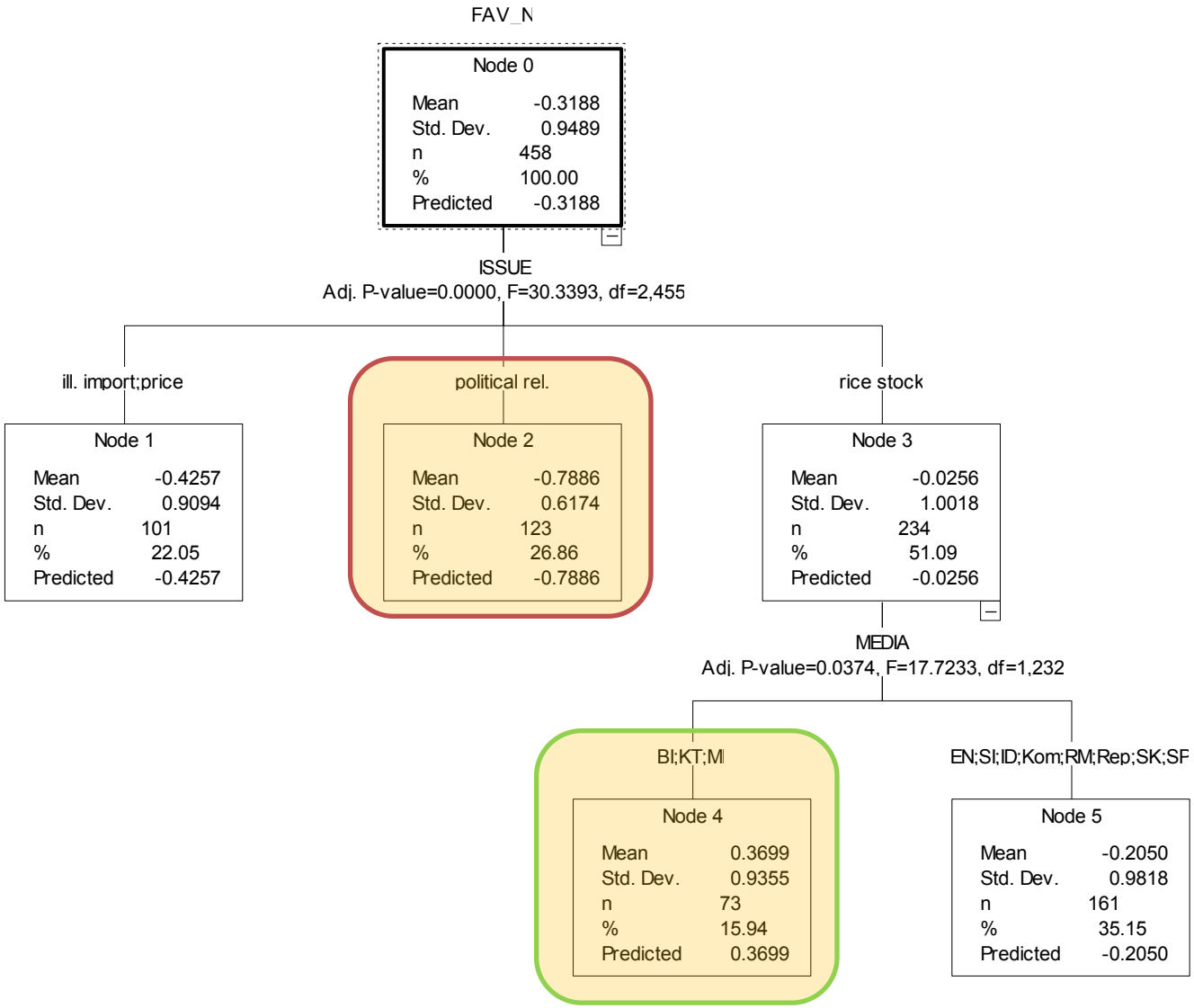


mediatrac

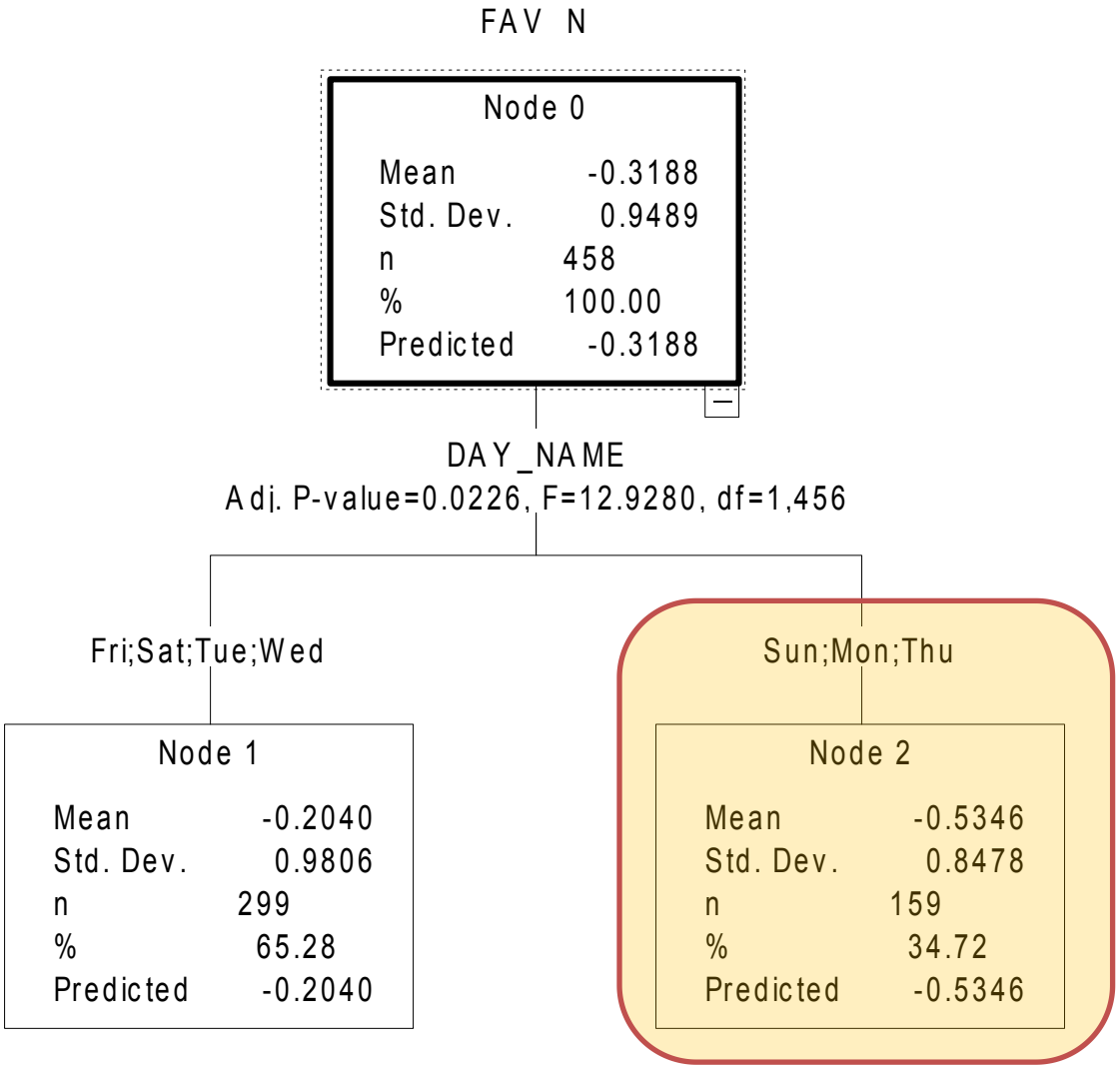
issue management tools: rice import



media coverage prediction: rice import

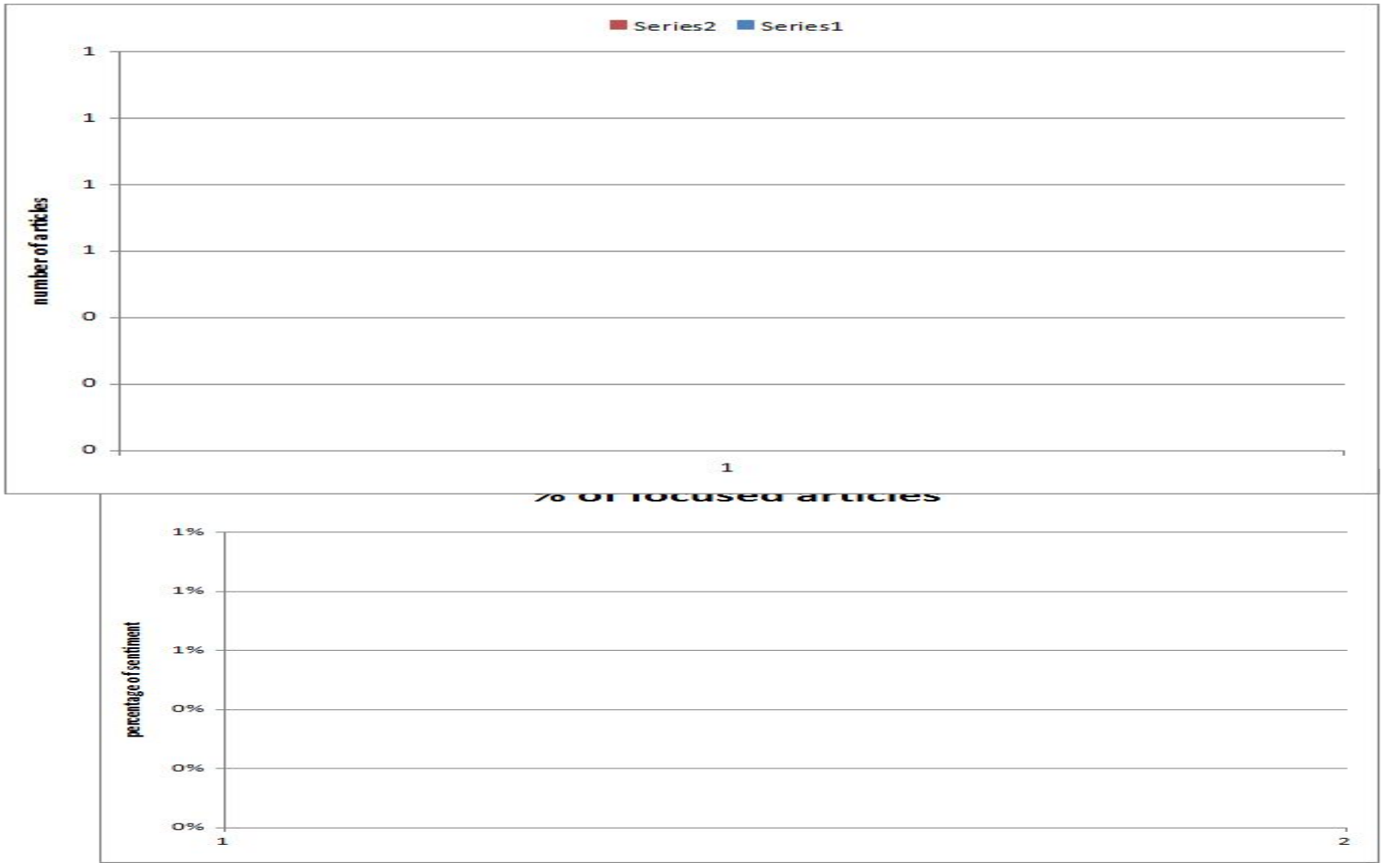


media coverage prediction: rice import

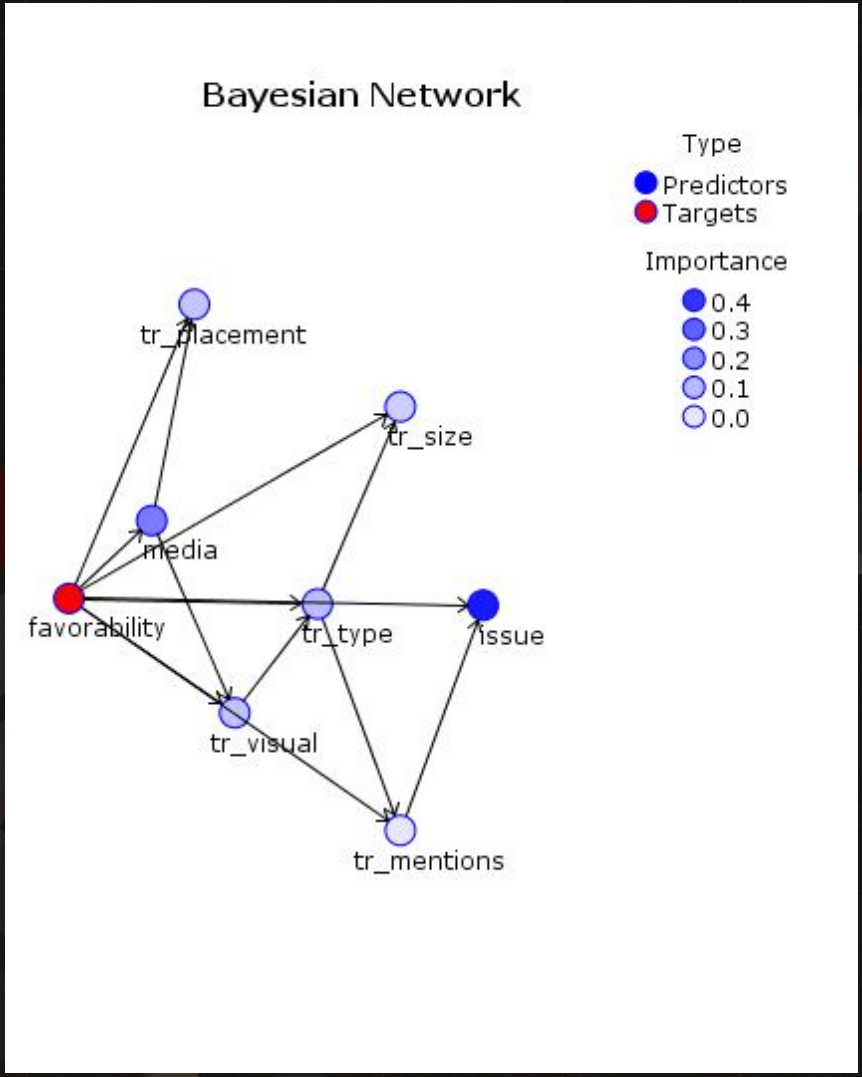


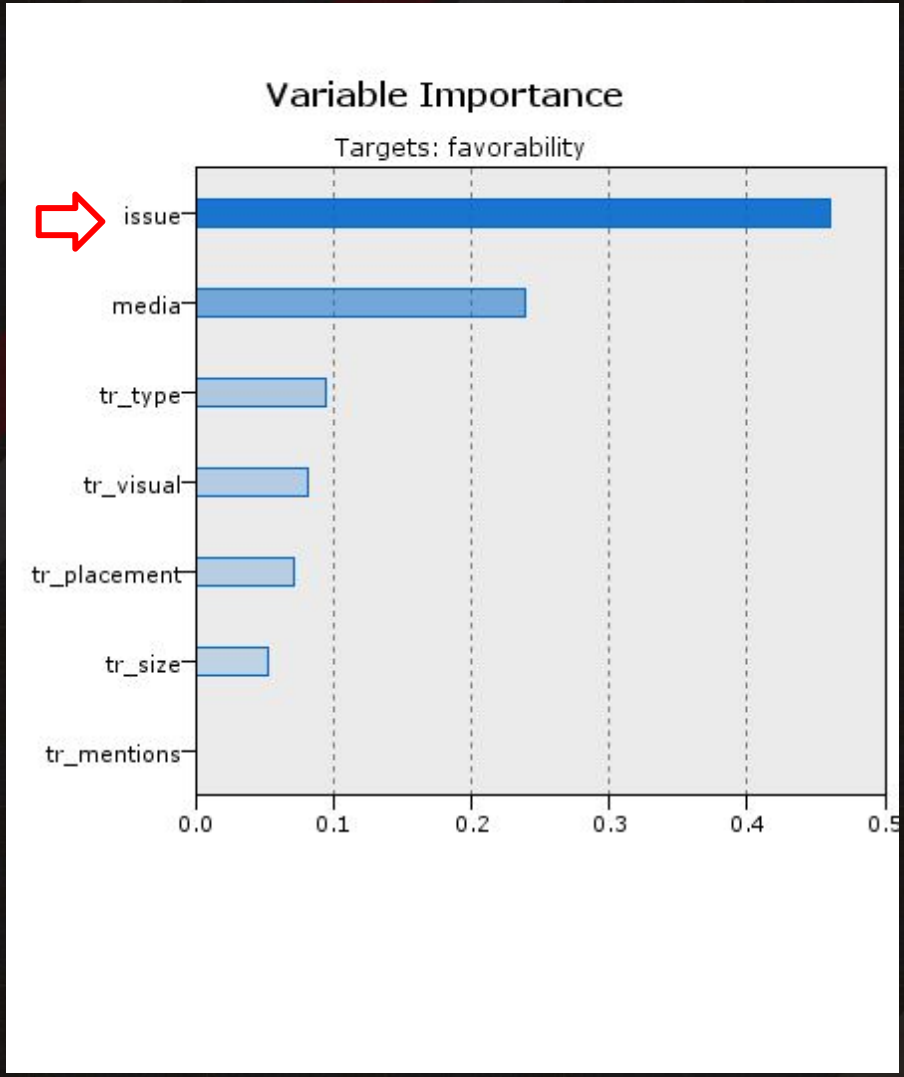
article volume and sentiment on BLT over time

high sensitivity towards poverty, BBM and election related issues

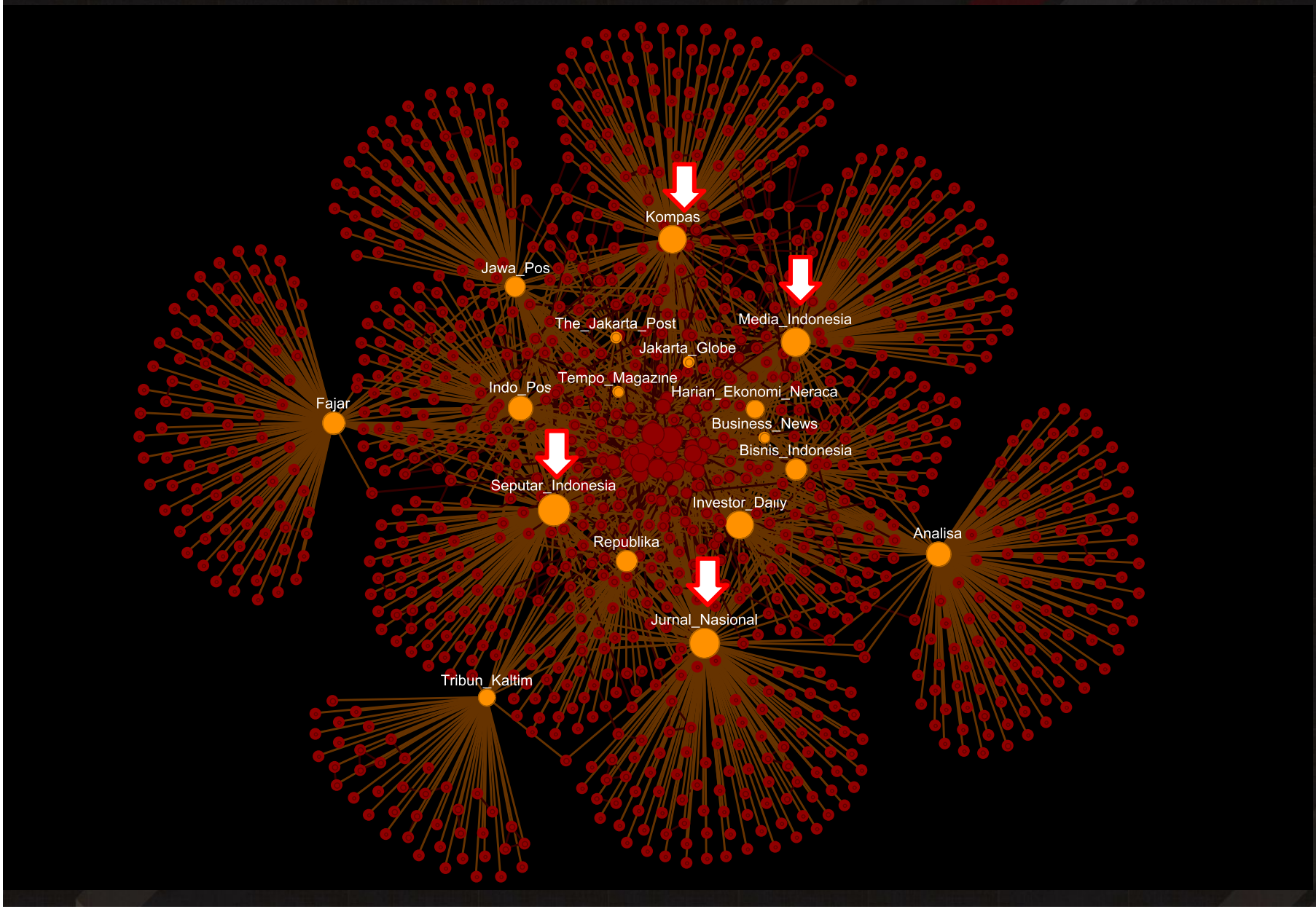


media coverage prediction: BLT

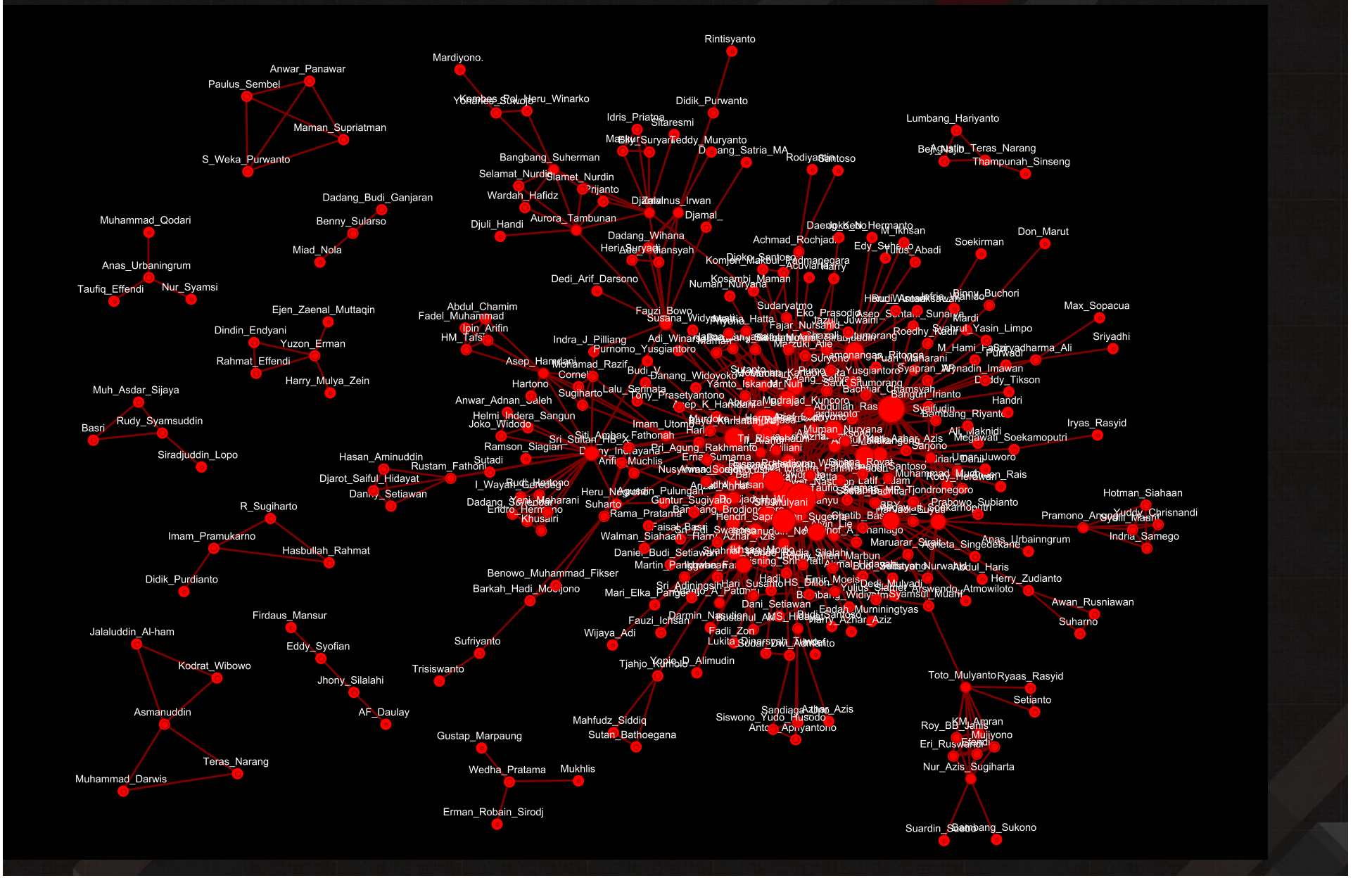




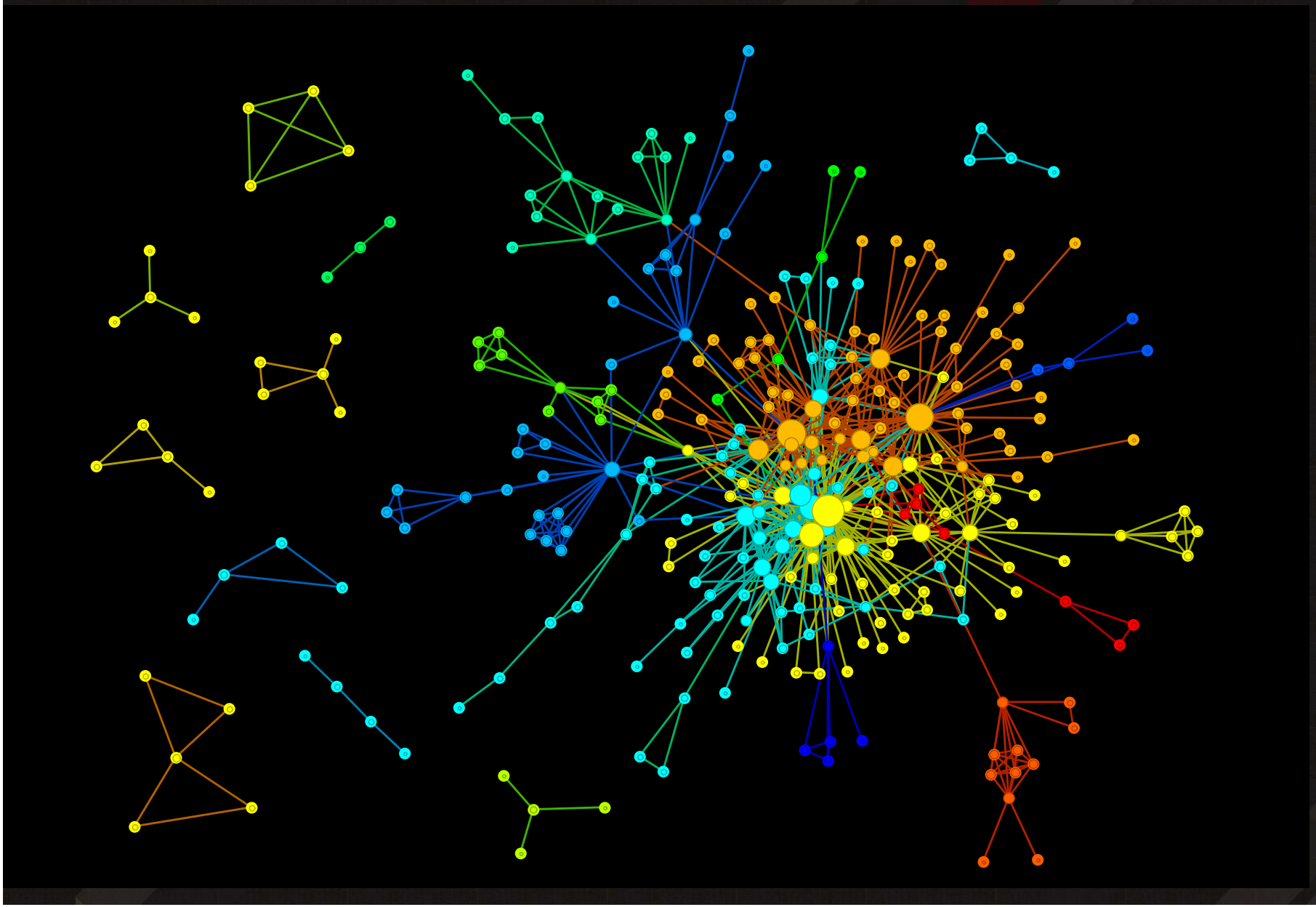
media positions on BLT issues



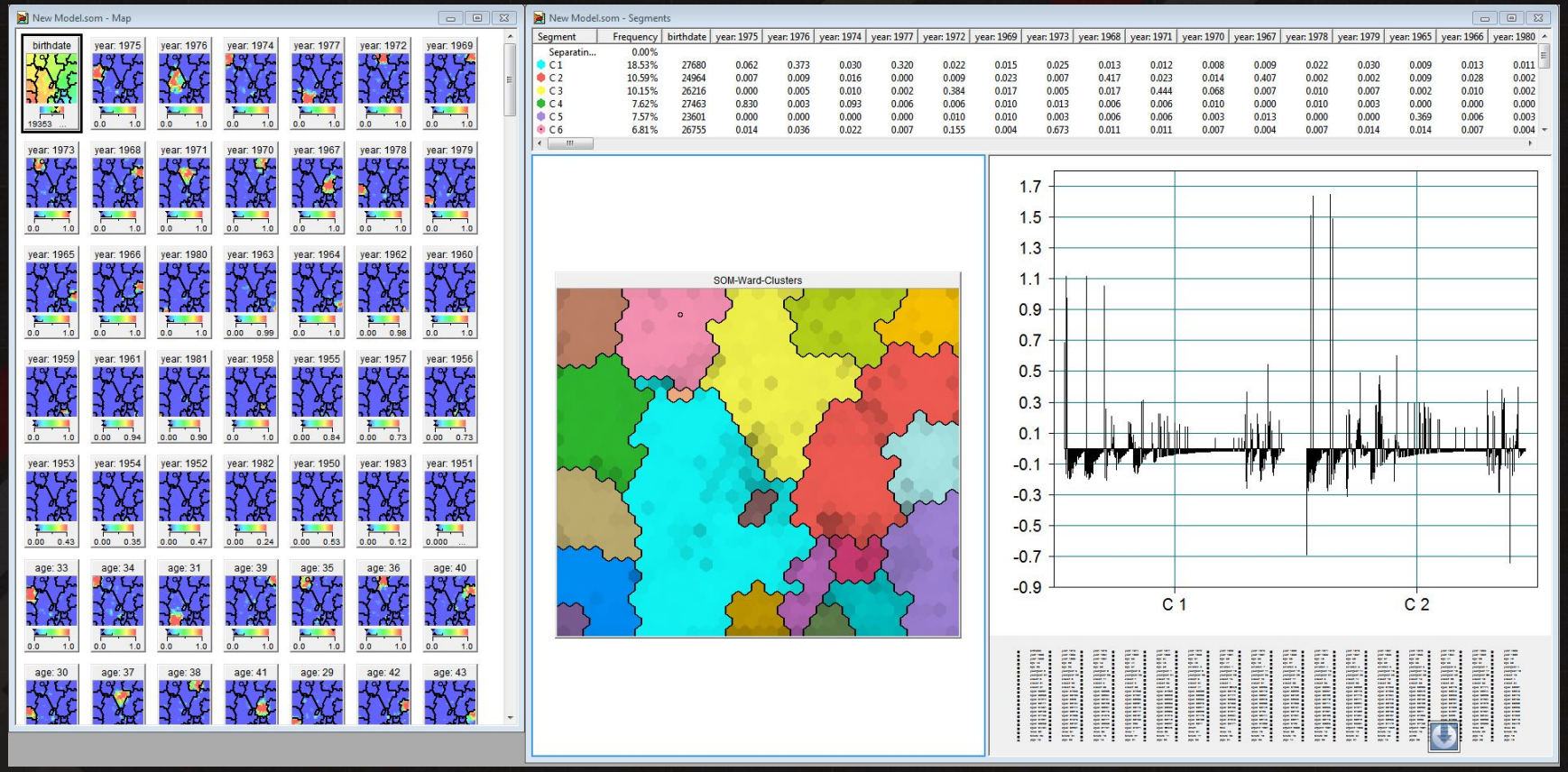
opinion leaders on BLT issues



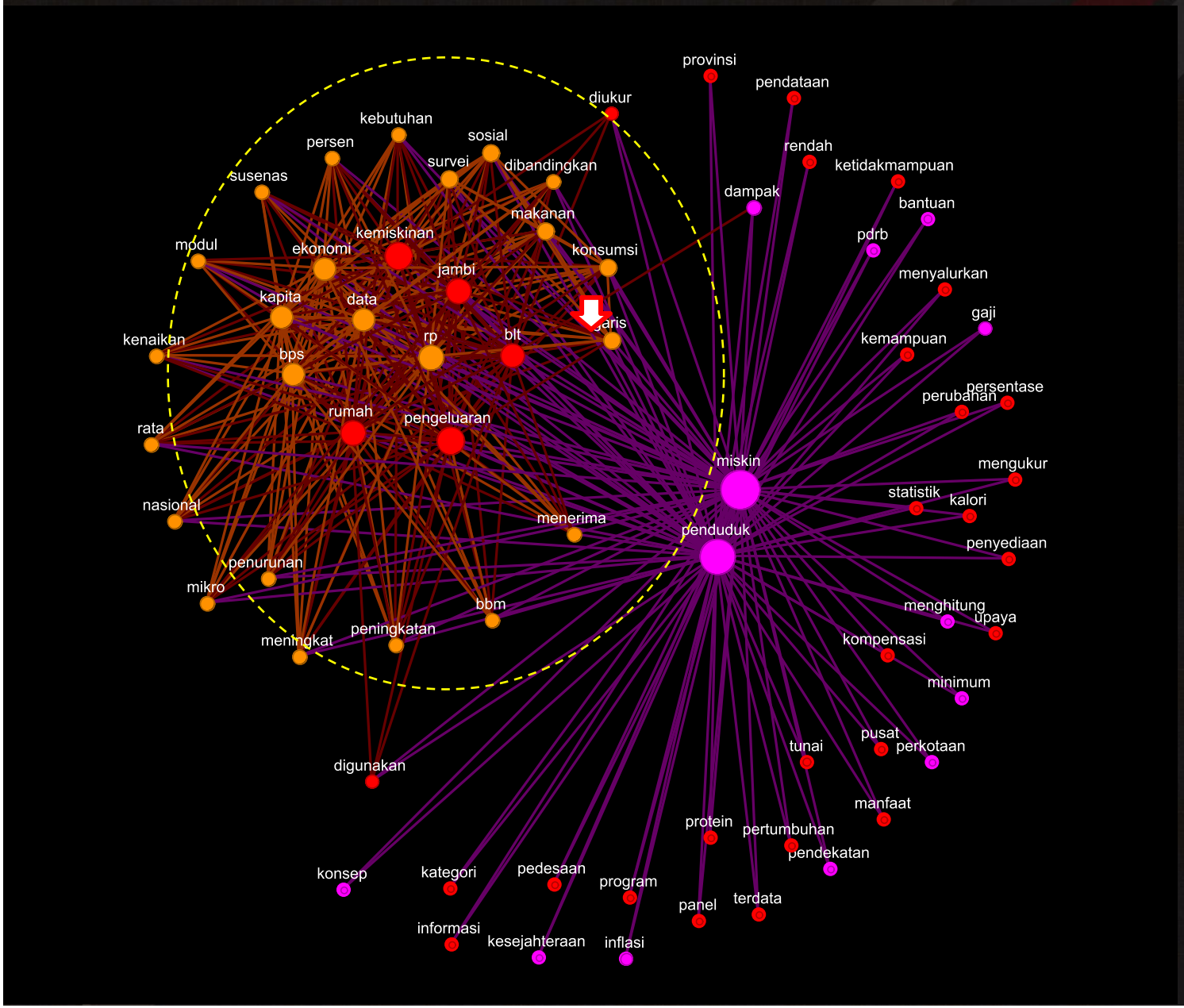
opinion leaders clusters



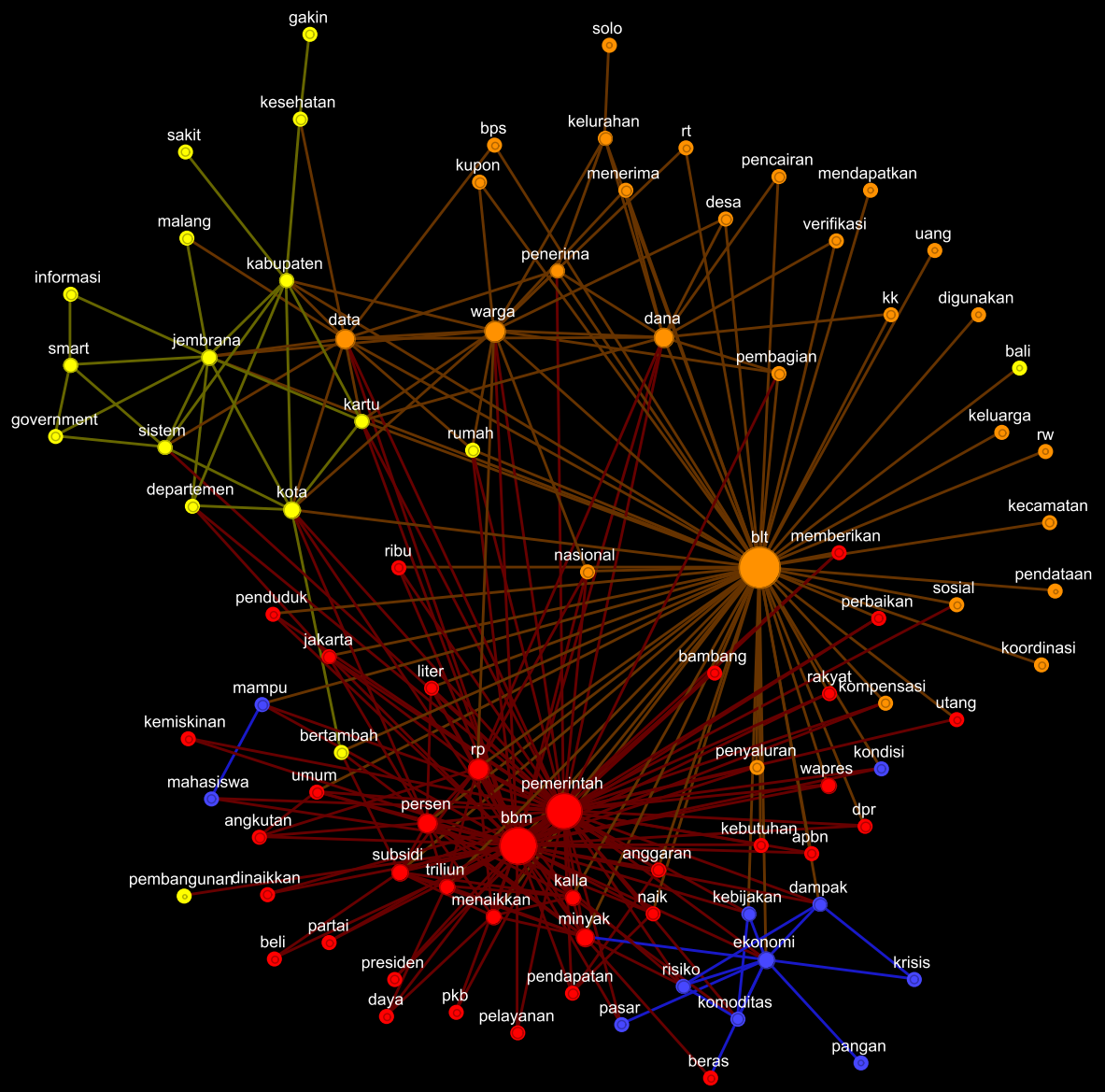
non obvious affinities between political groups



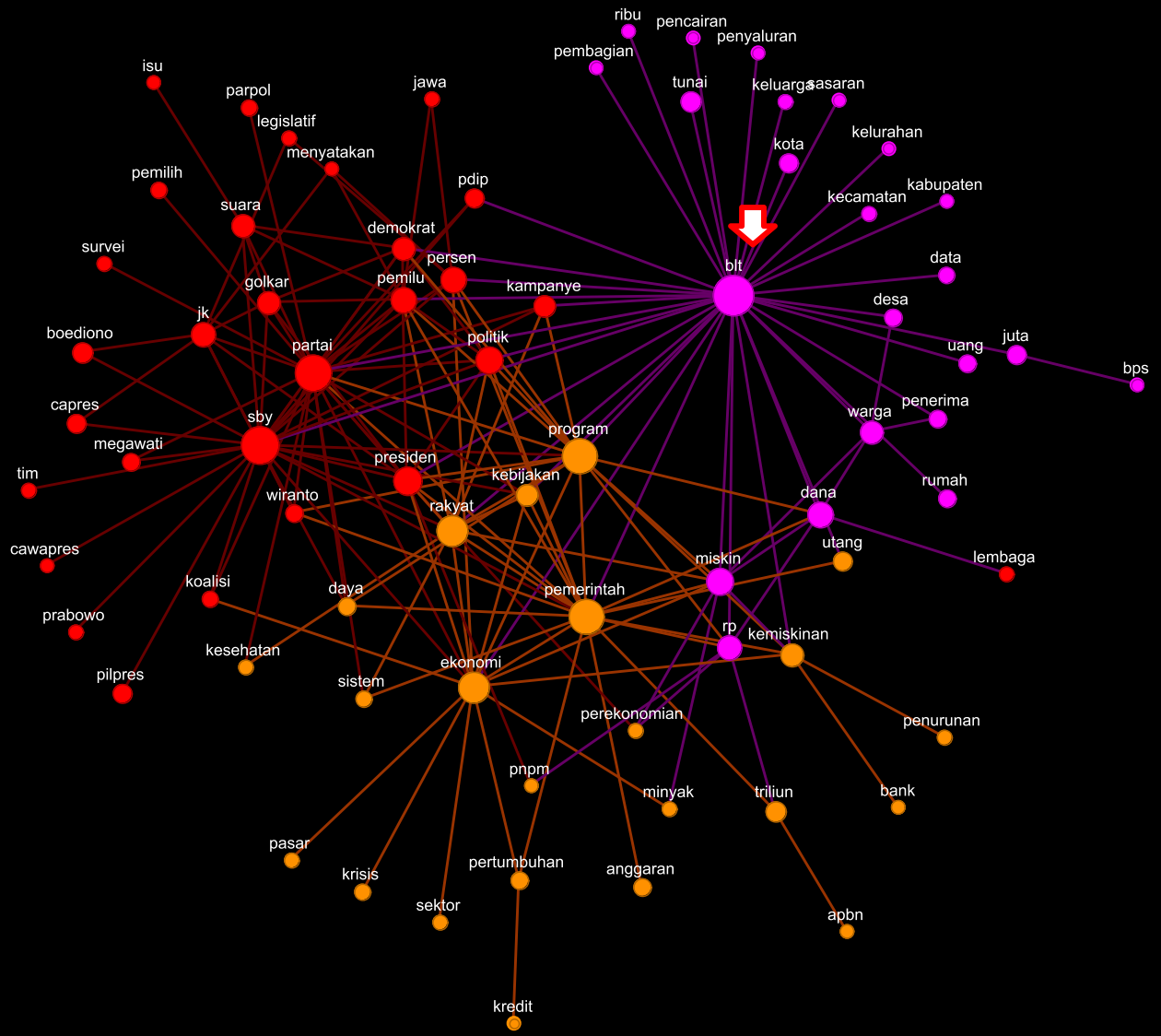
BLT ISSUE MAP FOR YEAR 2007



BLT ISSUE MAP FOR YEAR 2008



BLT ISSUE MAP FOR YEAR 2009



BLT WORD CLOUD FOR YEAR 2007

BLT WORD CLOUD FOR YEAR 2008

from title

BLT WORD CLOUD FOR YEAR 2009

from title

key learnings

use internet of things and crowd source
record everything
clean them, integrate
and share them back

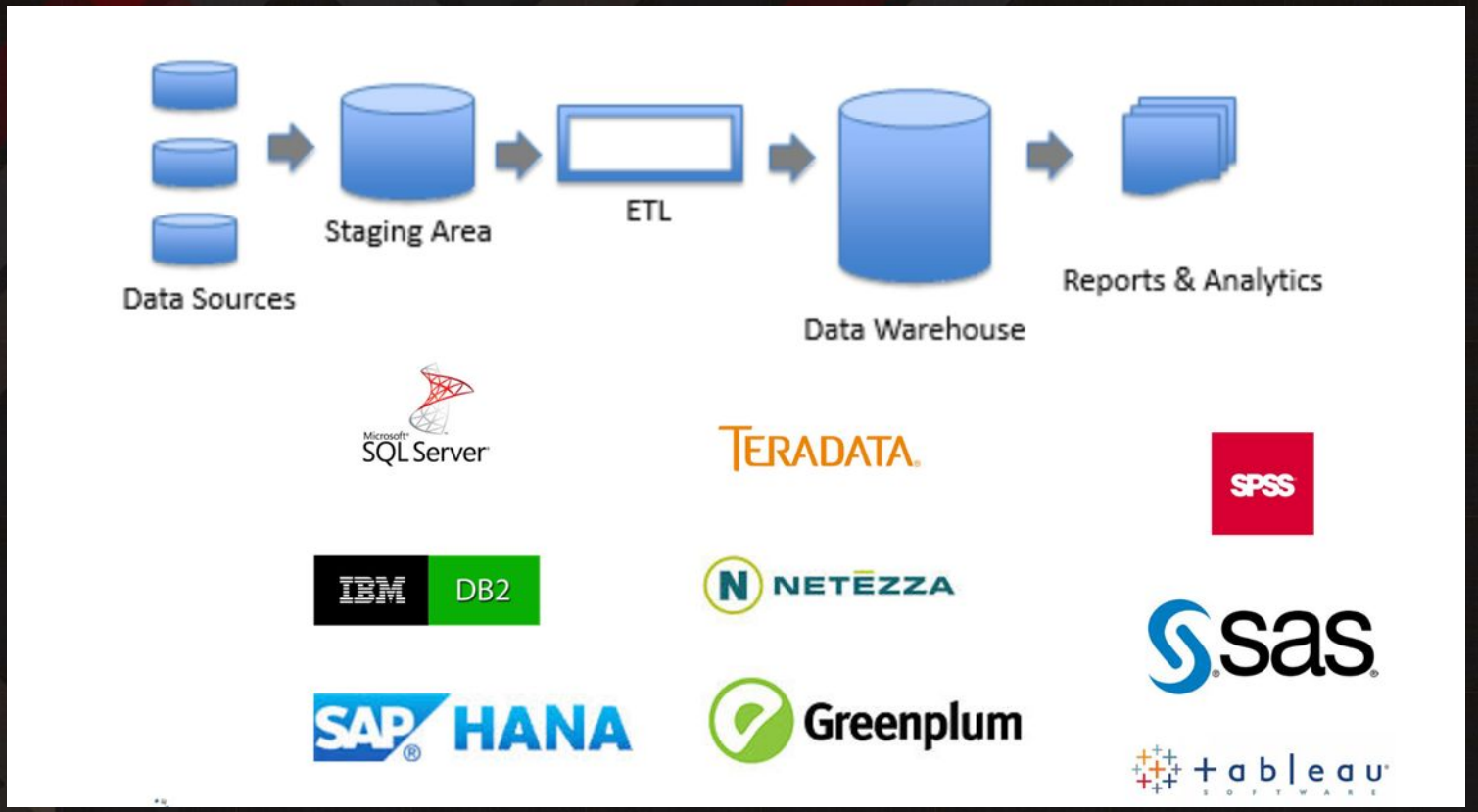
correlate everything
but, beware of over fitting

DATA THAT WE ACQUIRED SO FAR

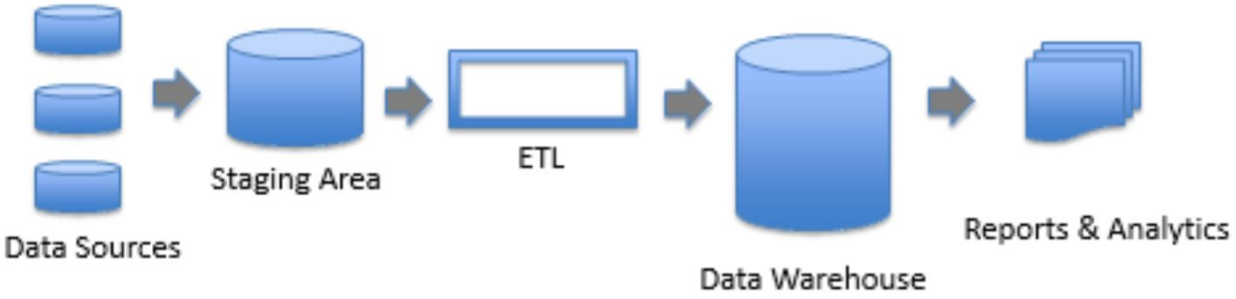
- linguistic data
  - 150+ nationwide offline print publications in digital forms since 2003 (20+ terabytes)
  - online news publications since 2005
  - forums, blogs and social media data since 2009
- geo demographics
  - 180+ million individual customer data
  - geo-tagged BPS data
  - 3 million+ geo-tagged point of interest data
  - 456.257 km up-to-date administration borders & roads (polyline data) including road classes & names and 77.994 areas up-to-date administration areas (polygon data) up to village level
- others
  - nationwide traffic data
  - business and SME data
  - agriculture data, etc.

techno
logy

not so long ago

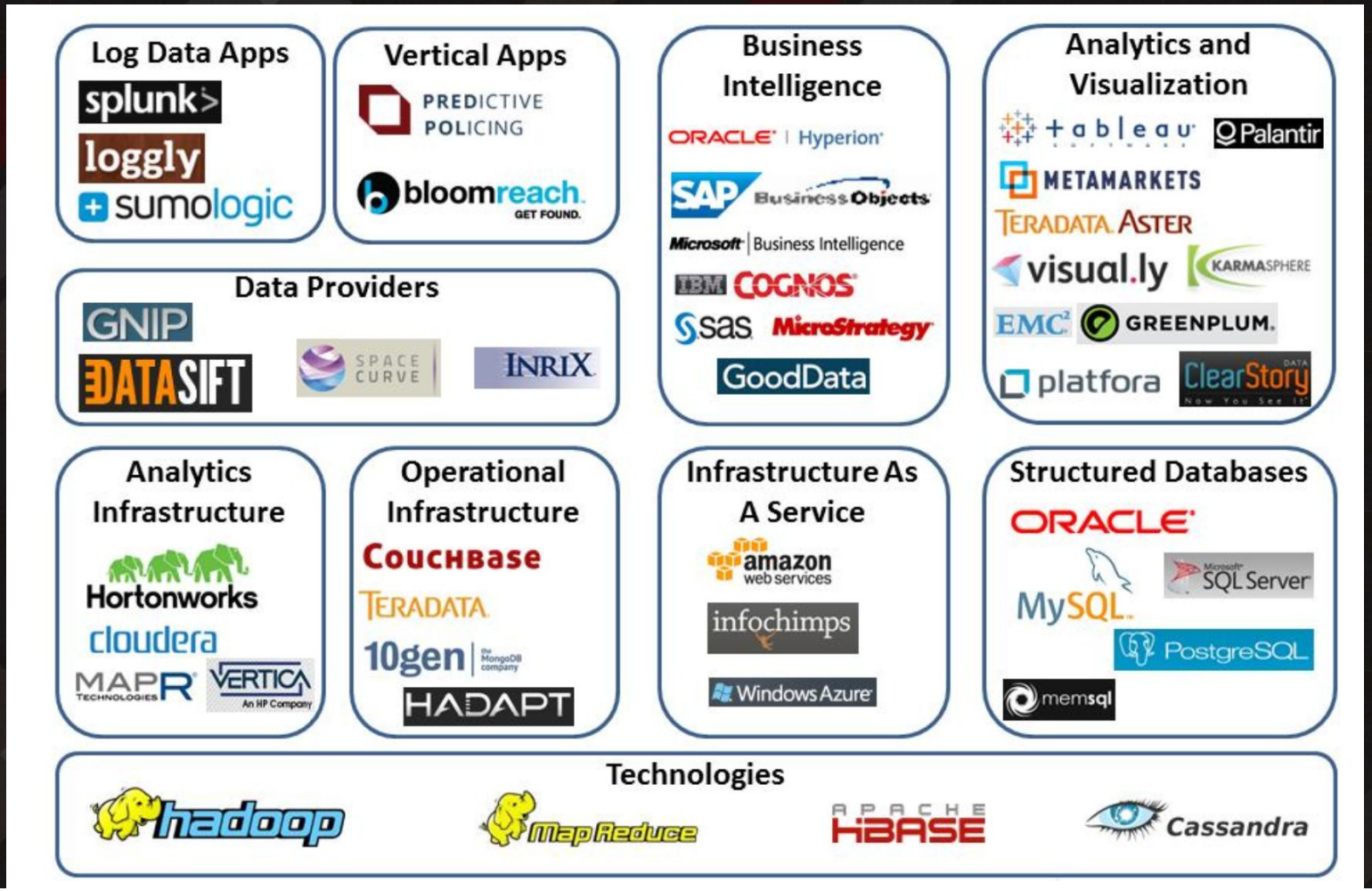


traditional vendor landscape

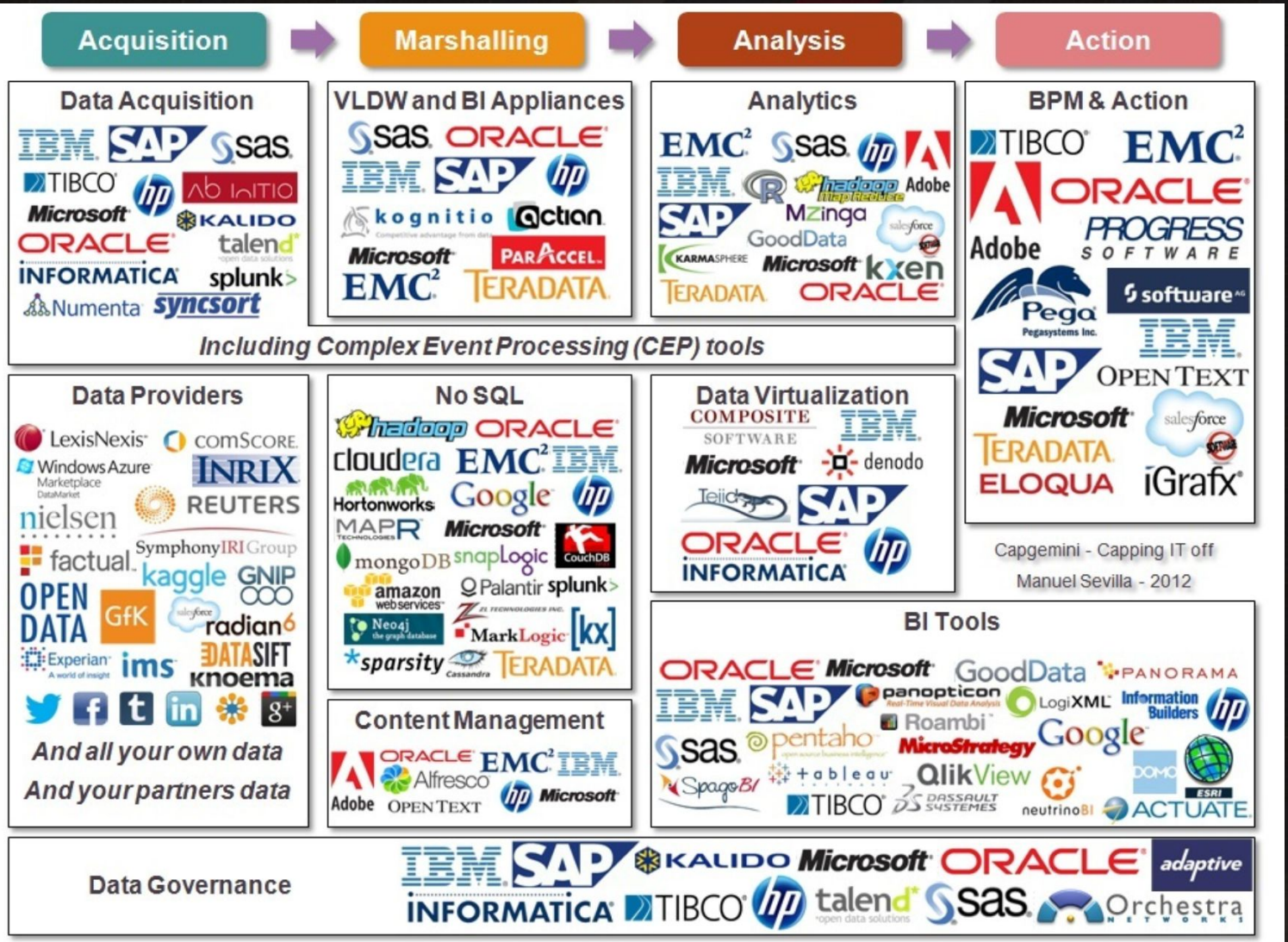


Microsoft SQL Server
TERADATA
SPSS
IBM DB2
NETEZZA
sas
SAP HANA
Greenplum
tableau SOFTWARE

now

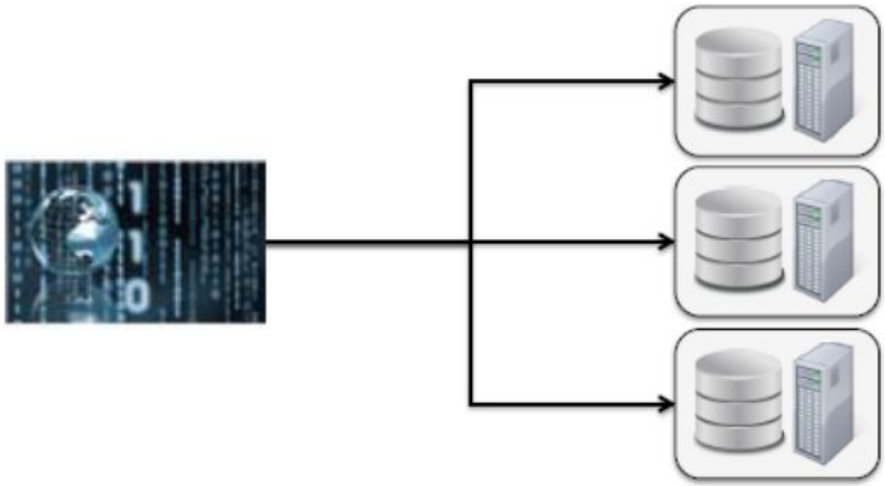

Log Data Apps
splunk>
loggly
sumologic
Vertical Apps
PREDICTIVE POLICING
bloomreach
GET FOUND.
Business Intelligence
ORACLE | Hyperion
SAP Business Objects
Microsoft | Business Intelligence
IBM COGNOS
sas MicroStrategy
GoodData
Analytics and Visualization
tableau
Palantir
METAMARKETS
TERADATA ASTER
visual.ly
KARMASPHERE
EMC² GREENPLUM.
platfora
ClearStory
Data Providers
GNIP
DATASIFT
SPACE CURVE
INRIX
Analytics Infrastructure
Hortonworks
cloudera
MAPR TECHNOLOGIES
VERTICA An HP Company
Operational Infrastructure
CouchBase
TERADATA
10gen
HADAPT
Infrastructure As A Service
amazon web services
infochimps
Windows Azure
Structured Databases
ORACLE
MySQL
SQL Server
PostgreSQL
memsql
Technologies
hadoop
Map Reduce
APACHE HBASE
Cassandra

big data vendors

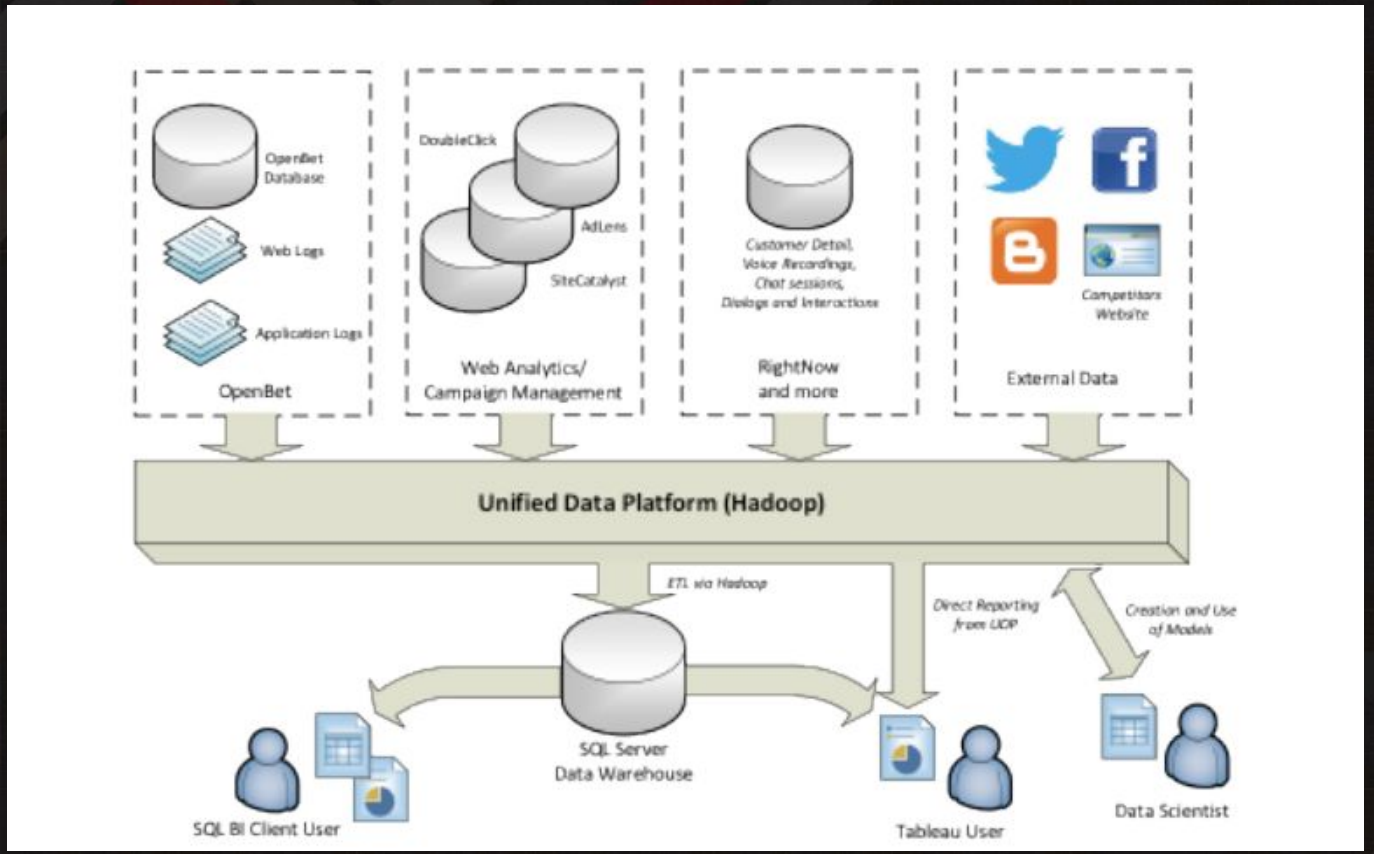


HADOOP
a new approach (not the only one)

- **A radical new approach to distributed computing**
  - Distribute data when the data is stored
  - Run computation where the data is



sample hybrid architecture



key learnings

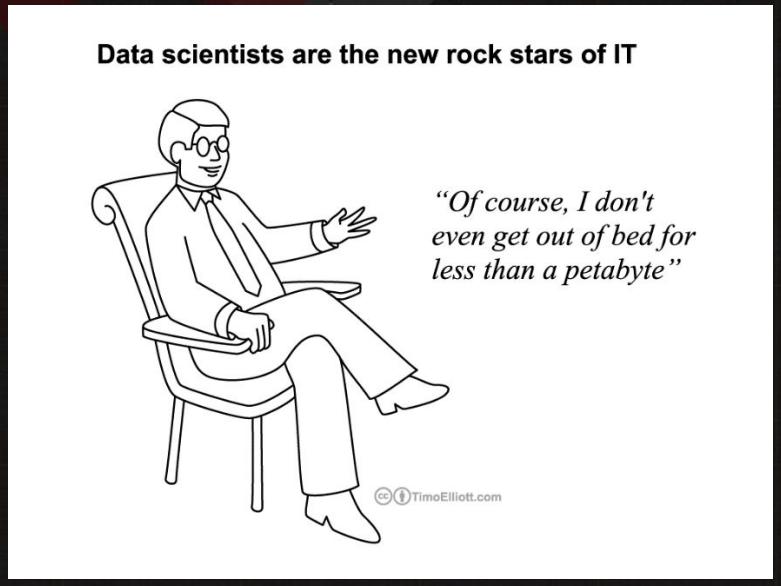
have a clear business case first
then choose and design appropriate infrastructure
have neutral consultants
do the transition in phases
use hybrid solutions if necessary
there is no miracle drugs

people

THE HARDEST PART TO GATHER

a group of data scientist

a group of big data technician



WHAT IS DATA SCIENTIST

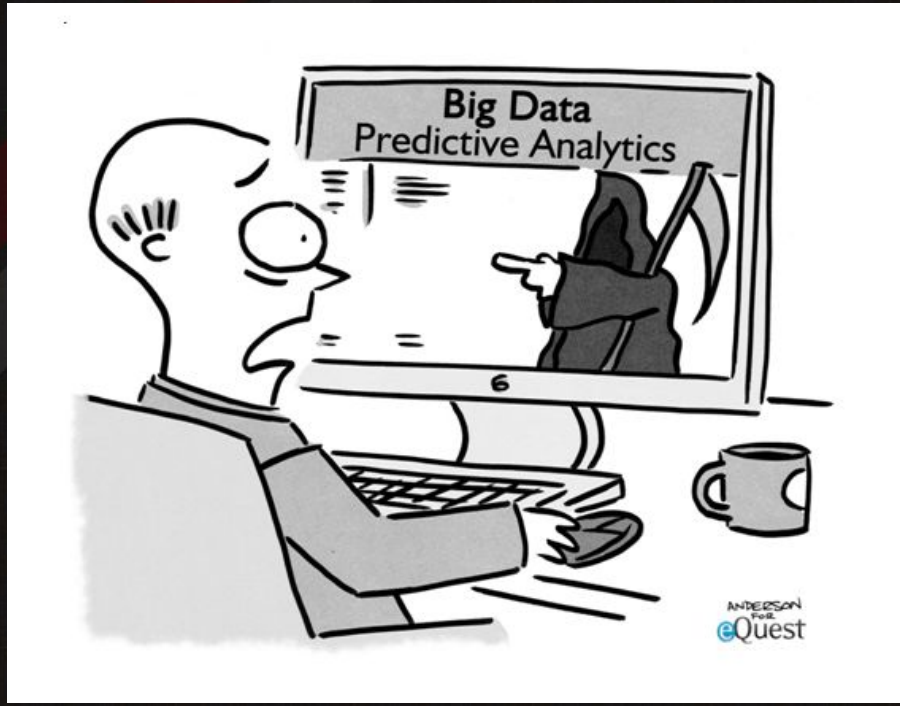
STATISTICIAN
PROGRAMMERS AND
MOST IMPORTANTLY
MATHEMATICIAN

EXPERTS IN
DATA VISUALIZATION,
ECONOMY,
SOCIAL SCIENCE,
PSYCHOLOGY,
PHYSICS,
BIOLOGY,
CHEMISTRY, AND
BUSINESS

BIG DATA ANALYTIC BUZZWORDS

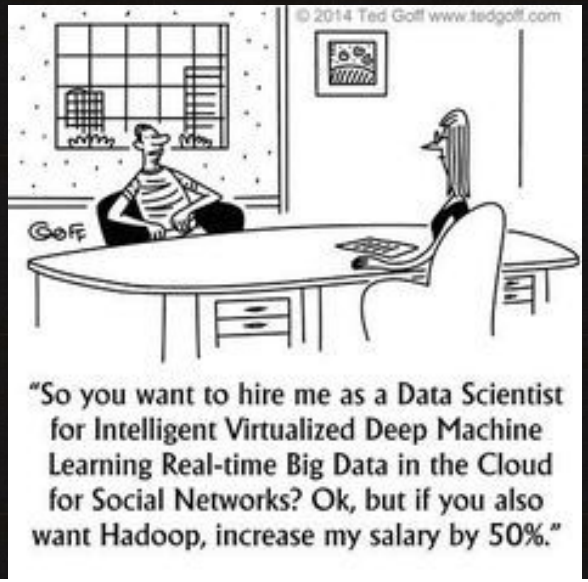
DESCRIPTIVE ANALYTIC
DIAGNOSTIC ANALYTIC
PREDICTIVE ANALYTIC
PRESCRIPTIVE ANALYTIC
COGNITIVE ANALYTIC

MACHINE LEARNING
DEEP LEARNING
LINEAR PROGRAMMING
SELF ORGANIZING MAP
AGENT BASED MODELING
BAYESIAN STATISTICS
GRAPH THEORY
SOCIAL NETWORK ANALYSIS
TOPOLOGICAL DATA ANALYSIS
ROUGH SET THEORY
DIFFERENTIAL GEOMETRY



key learnings

first, have a clear business case
then, you have to adapt your organization
to attract the right people
since most of them are specialist, explore crowdsourcing and external partnership
remember, you don't have to have or use all the newest thing you read in a magazine



technology and analytic method that we developed

that might be useful for public sector analytic

- data acquisition
  - proprietary optical scanning devices with automatic calibration engine
  - numerous development in mobile applications and internet related devices
- data cleansing and preparation
  - automatic error spelling correction library and engines
  - entity and relationship extraction engine
  - text summary and classification
  - indonesian language corpus
  - indonesian entity and relationship database
- data integration, matching and enrichment
  - BIG data infrastructure expertise
  - SQL and NOSQL data integration library
  - identity matching engine
  - entity enrichment engine
- data query, visualization, mining and advanced analytic
  - proprietary query and visualization library
  - proprietary mining and analytic tools and library
    - sentiment analysis
    - topic modeling
    - adaptive modeling engine
    - financial transaction analysis and visualization engine
    - brand relationship network

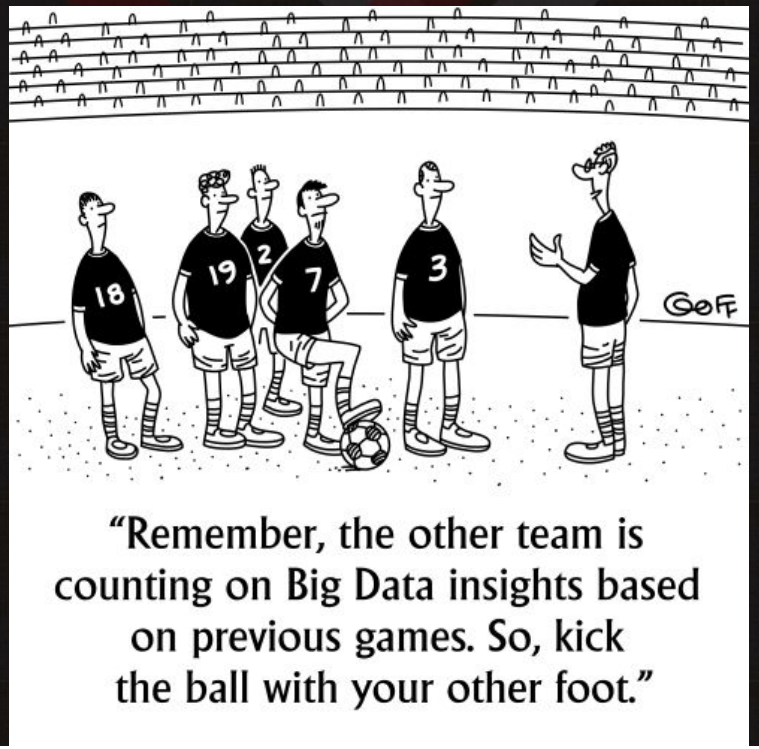
organization

organizational pre-requisites

- strong business case
- readiness assessments
- commitment from top management
- step by step implementation
- change management team

important big data related issues

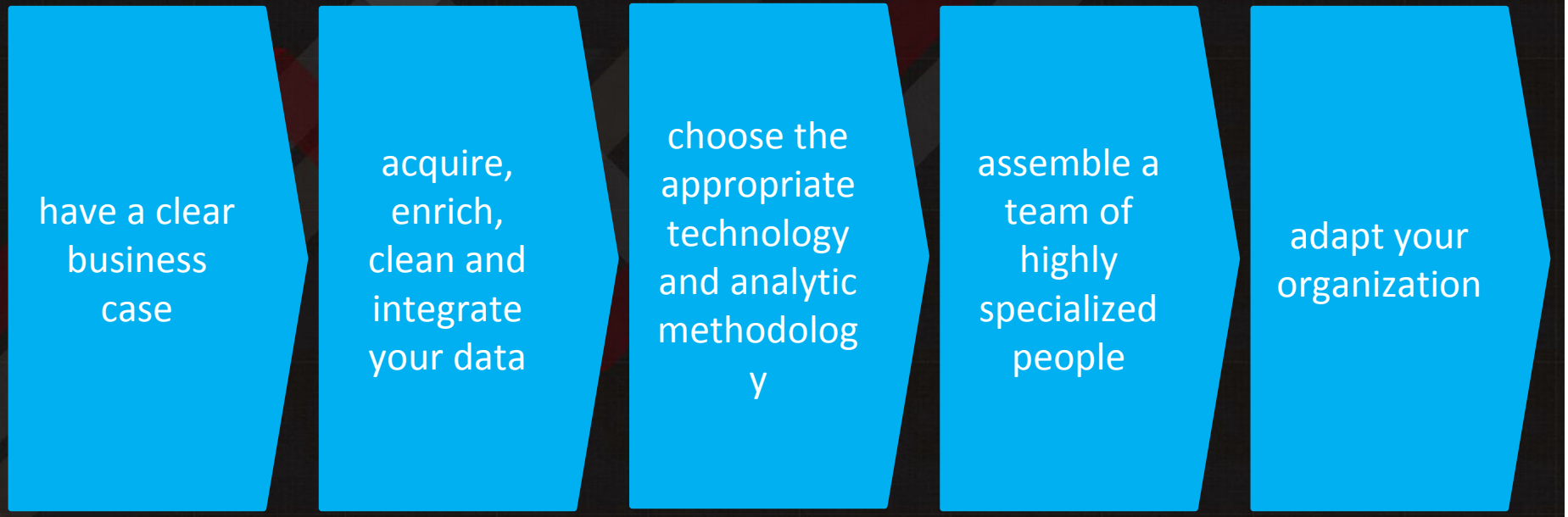
- data governance and privacy policy
- big data security
- big data innovation and research
- data centric culture and habit



key learnings

have a really strong and clear business case
do thorough readiness assessments
have a persistent change management team
implement in stages

WHAT NEEDS TO DONE




WE'VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...
HUMONGOUS DATA
© D.Fletcher for CloudTweaks.com

THANK YOU