Big Data Introduction Session

7 Juni 2017

Fajar Jaman
Business Director Starcore

Dino Business

- Traditional Telco
- Traditional Banking

Vs

C-business

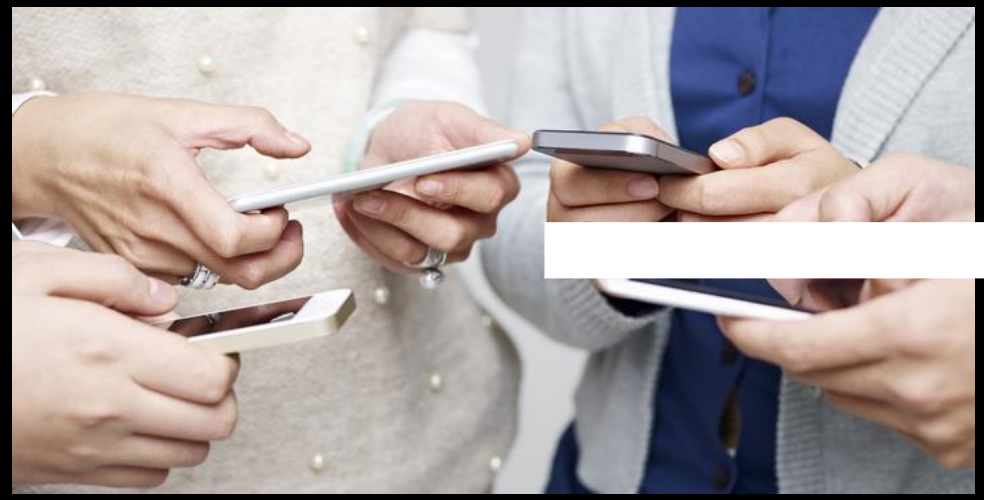
- Whats app
- Skype
- Spotify

Economy Shifting ?

Perusahaan Telko secara global kehilangan revenue sebesar $386 billion (Fortune)

- SMS
- Call

- Tower
- 4G



- Youtube
- Spotify

- Content Creation

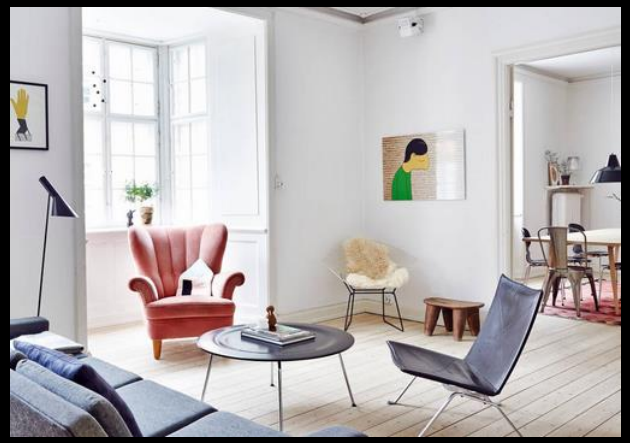
Increasing of mobile phone user ≠ increasing of Telco profit

Meet the new **billion dollars company:**

Unicorn

UBER

USD 62.5 Bn



AIR BNB

USD 30 Bn

Former Unicorn → IPO

Facebook

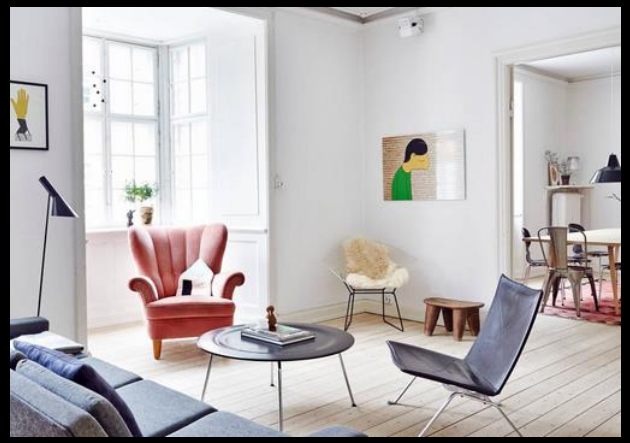
USD 104 Bn

Alibaba

USD 238 Bn

Apa persamaan mereka?

Unicorn



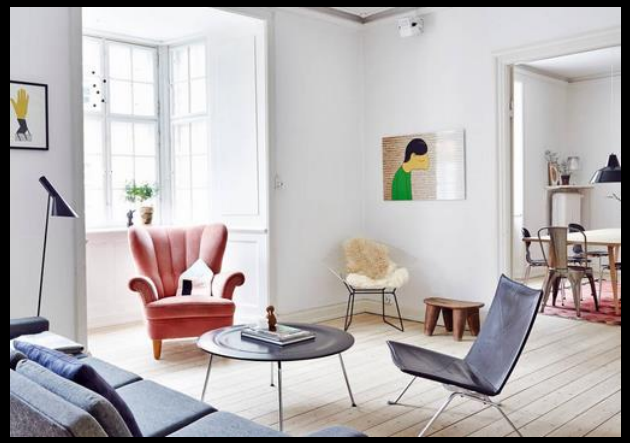
Former Unicorn → IPO

Mereka memiliki dan memanfaatkan:

Digital Data

Digital data driven Business capability

- **Dynamic pricing**
- **Rating System**
- **Operational Optimization**



- **Enhanced search feature**
- **Guiding host to the perfect price**

- **Facial recognition**
- **Analyzing the likes**
- **Tag Suggestion**

- **Anti-counterfeiting**
- **Optical Character Recognition**

How **Data Science increased AirBnB valuation** to $25.5 bn?

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Rea

TECH | TECHNOLOGY

The Secret Math of Airbnb's $24 Billion Valuation

Home-rental site's revenue projected to top more than $900 million

By ROLFE WINKLER and DOUGLAS MACMILLAN

June 17, 2015 3:15 p.m. ET

Home-rental site Airbnb Inc. has given potential investors in a $1 billion funding effort an ambitious revenue forecast to justify a richer valuation than hotel giant Marriott International Inc.

Analysis Function

- A/B Testing
- Image recognition and analysis
- Natural language processing
- Predictive modeling
- Regression analysis
- Collaborative filtering

Analytic Capability

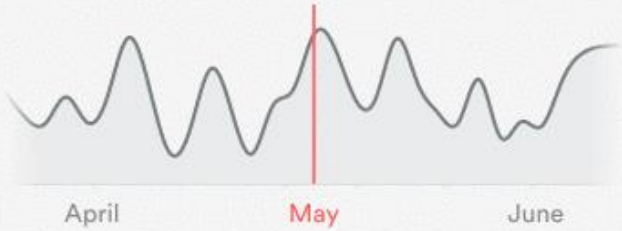
1. **Enhanced Search Features**
2. **Guiding Hosts to the Perfect Price**
3. **Driving Company Growth**

May 2015

Settings

Sun	Mon	Tue	Wed	Thu	Fri	Sat
26 $40	27 $40	28 $40	29 Mum is coming $68	30 $40	May 1 $40	2 $40
3 $40	4 $131	5 Today $131	6 $92	7 $94	8 $96	9 $120
10 $98	11 $99	12 $99	13 $101	14 $102	15 $103	16 $144
17 $107	18 $107	19 $109	20 $110	21 $111	22 $112	23
24 Ellen	25	26	27	28 $120	29 $122	30 $122
31	Jun 1	2	3	4	5	6

Travel trends in your area are changing.



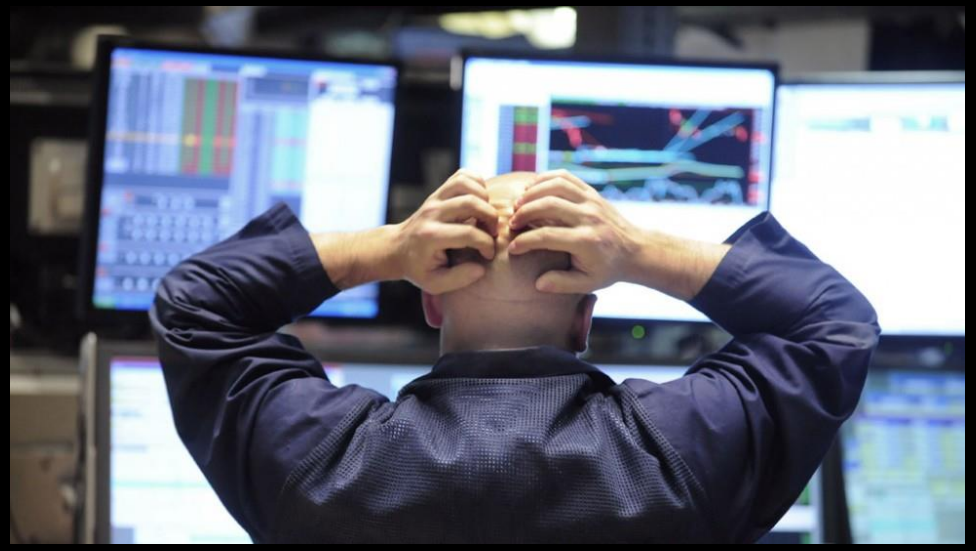
You can take advantage of these changes by adjusting your price on a daily basis. Over time, this can help you make more money.

See Price Tips for May

Price tips will next update tomorrow

Digital data driven di Public Sector?

Dari krisis..



Global Financial Crisis 2008

Extreme **poverty** is defined as living on $1.25 or less a day. In 2010, 414 million people were living in extreme **poverty** across sub-Saharan **Africa**.

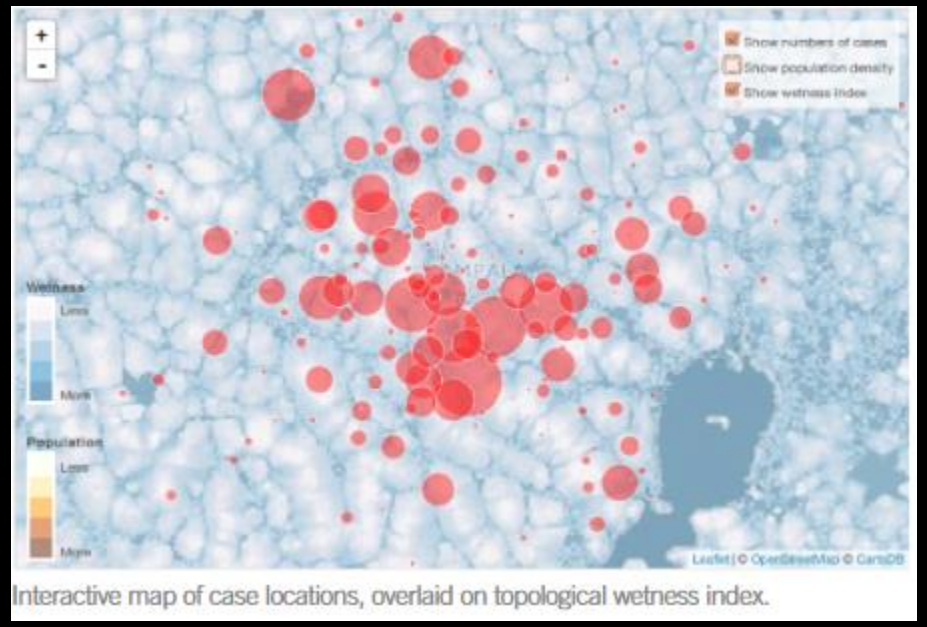
born the **coolest govt initiative** in the World

initiative of the Secretary General's office formed to explore the use of new **sources of digital data, calls this a "new natural resource"** that can be cultivated for society's benefit.

Launched in 2009

- **New York (US)**
- **Kampala (Africa)**
- **Jakarta (Indonesia**

born the **coolest govt initiative** in the World



Interactive map of case locations, overlaid on topological wetness index.

Use Telco Data to
Control the outbreak (Africa)

LOMBOK ISLAND	Week 1 3/7 - 3/13	Week 2 3/14 - 3/20	Week 3 3/21 - 3/27	Week 4 3/28 - 4/3	Week 5 4/4 - 4/10	Week 6 4/11 - 4/17	Week 7 4/18 - 4/24
Onion	80%	103%	187%	297%	470%	1023%	1373%
Sweet Potato	43%	7%	30%	87%	87%	123%	187%
Sugar	177%	227%	227%	320%	340%	377%	693%
Premium quality rice	100%	157%	153%	353%	247%	357%	467%
Mackerel	7%	53%	73%	83%	153%	170%	453%
Broiler / layer chicken eggs	33%	110%	177%	490%	407%	1033%	1313%
Vegetable Oil	270%	517%	320%	407%	430%	753%	1117%
Powdered Milk	267%	240%	117%	283%	167%	193%	157%
Long Bean	87%	97%	320%	430%	433%	627%	963%
Tomato	100%	117%	330%	380%	407%	997%	1613%
Tempeh	37%	173%	237%	393%	370%	697%	1207%
Low quality rice	23%	30%	47%	240%	183%	197%	310%

Use Crowdsourcing Data
(Citizen report) to monitor the commodity price (Lombok)

Digital data driven di Public Sector?

improved public policies that make better use of **research, analysis, and evidence**

Utilize big data for creative economy

What's new In Research now?

Big Data Technology

How data science and analytics can contribute to sustainable development

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

Y EDUCATION
ng can
s for
out rates

GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

GLOBAL PULSE

www.unglobalpulse.org
@UNGlobalPulse 2016

NEXT MONEY

FF17 FINALISTS | SPEAKERS | SCHEDULE | PARTNERS | VENUE | BOOK A HOTEL | BUY TICKETS

OPENING KEYNOTE

No Cash, All Data – The World of IOT Finance is here

Jim Marous

Owner, Publisher, The Financial Brand

PANEL SESSION

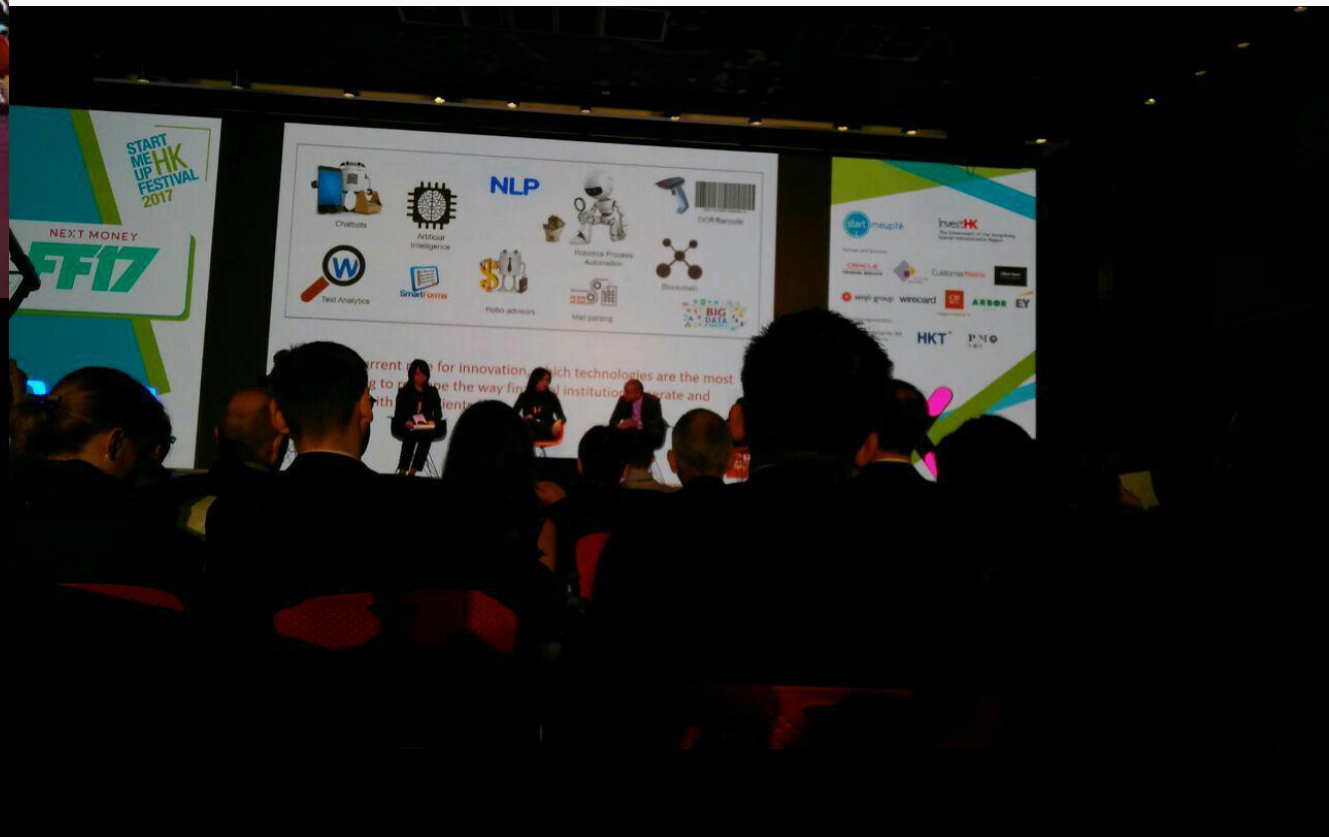
Lending in Mobile Asia – Data Driven, Highly Personalised, Instantly Delivered

Moderator: James Lloyd

Asia-Pacific FinTech Leader, EY

Jing Zhou

CEO, Dumiao



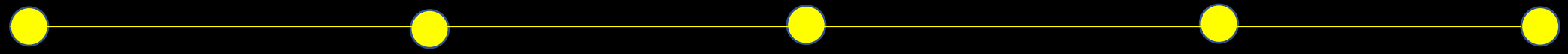
Asia's Biggest fintech conf.

Focus:

- Big Data
- NLP
- Machine learning

Big Data, Why now?

The First Warning of Data's Storage and Retrieval Problem



18000 SM

Use Tally Stick to record transaction

1944

in 1944, when Fremont Rider, a Wesleyan University librarian, estimated that American University Libraries were doubling in size every sixteen years.

2006

Hadoop was created in 2006 out of the necessity for new systems to handle the explosion of data from the web

2016

Google Dataflow for Stream processing

?

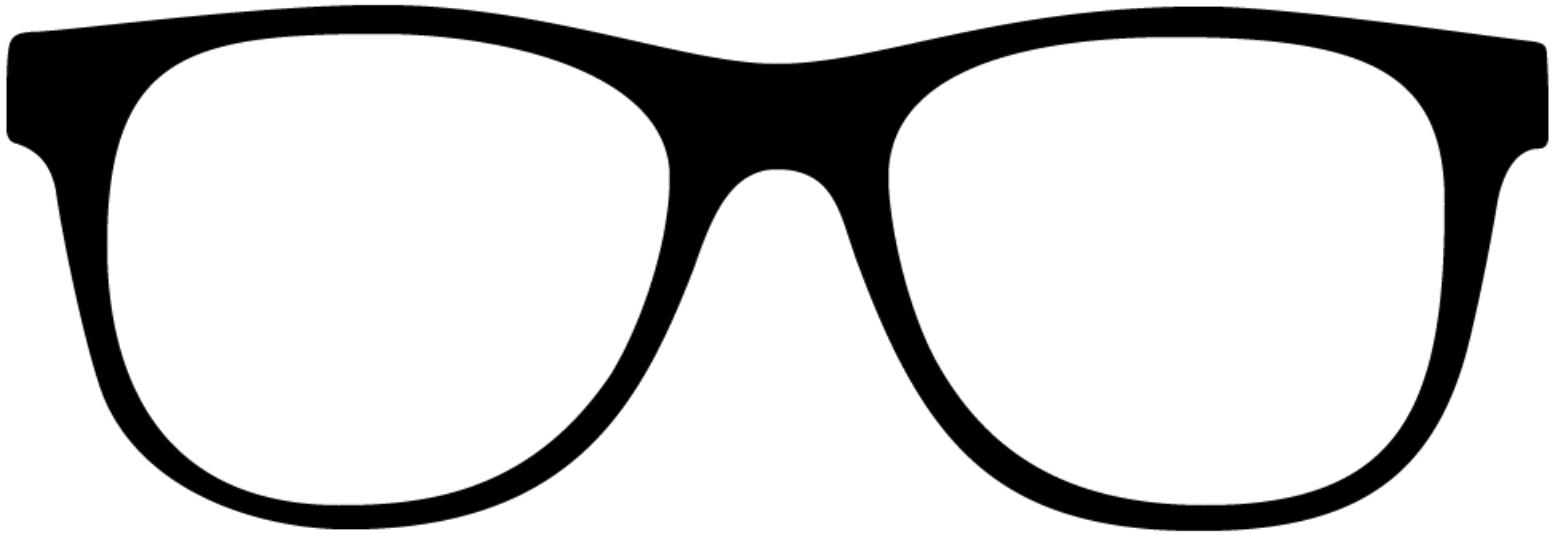
Human History always facing the data problem

Big data adalah cara pandang melihat semua hal menjadi sebuah **data point** yang dapat dimanfaatkan untuk menjadi pembentukan **value**

What normal eye **see?**

Ordinary, unstructured box

Now, wear “big data glasses”



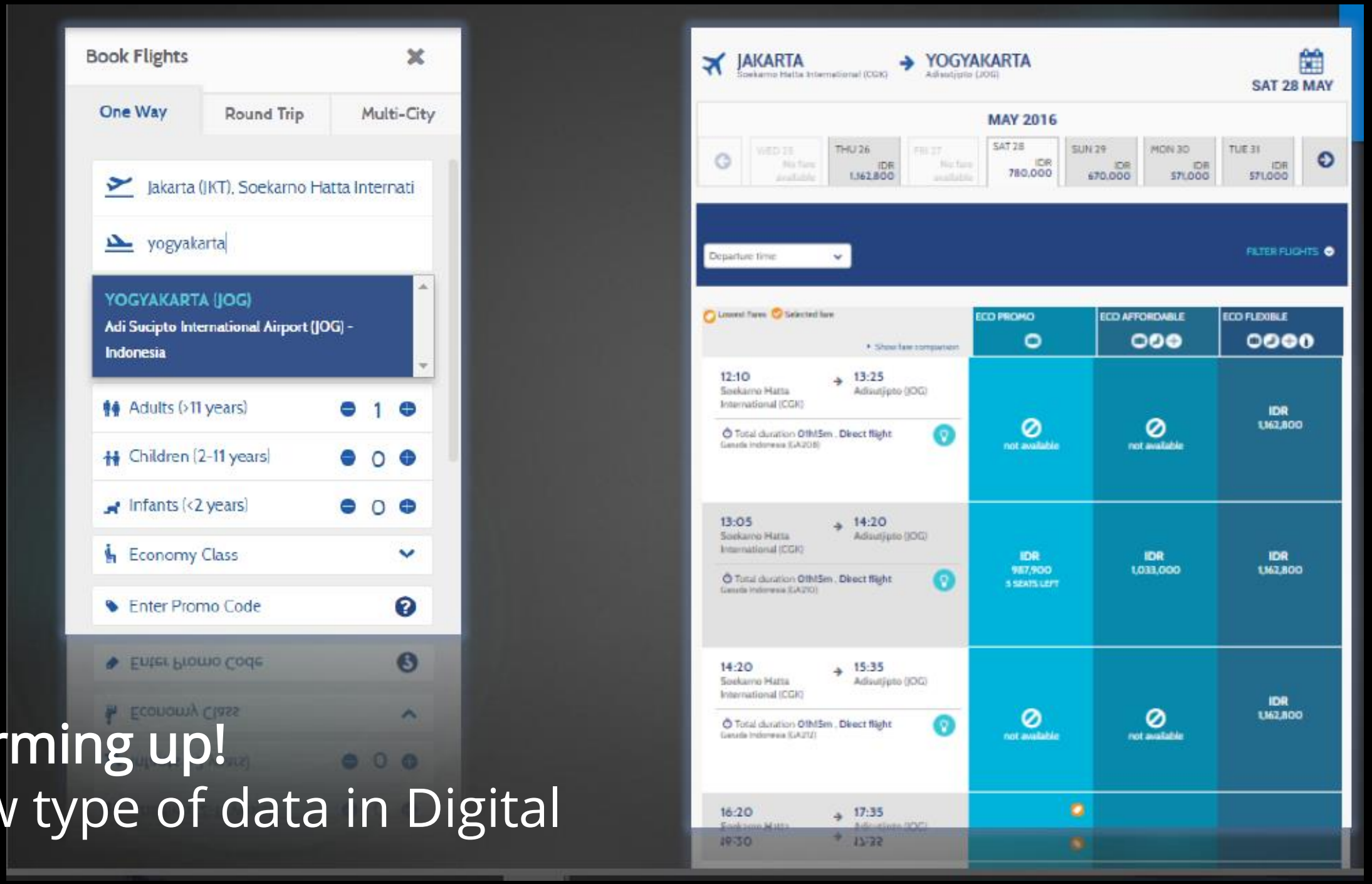
What big data eye **see?**

Extraordinary, unstructured **data**

- Real Time Image Feed
- Image Processing to count number of people

- Image Processing to count number of people
  - **52 people on the line**

To understand more on big data
Lets get into an example



Warming up!

New type of data in Digital

1

Customer memesan kereta dari **Jakarta ke Jogja**
pada tgl 18 May 2016

TRANSACTION DATA

1. Customer searching jadwal dari Jakarta ke Yogya
2. Customer berada di halaman pencarian selama 10 menit
3. Customer keluar dari halaman pencarian dengan menekan tombol “Back”
4. Customer mencoba melihat keberangkatan Jakarta ke Solo
5. Customer memesan dan membayar keberangkatan Jakarta ke Solo tgl 18 May 2016

INTERACTION DATA

It’s Digital happen

Transaction Data

No.	Cust A	Time	Transactions
1	12346	18 May	Transactions

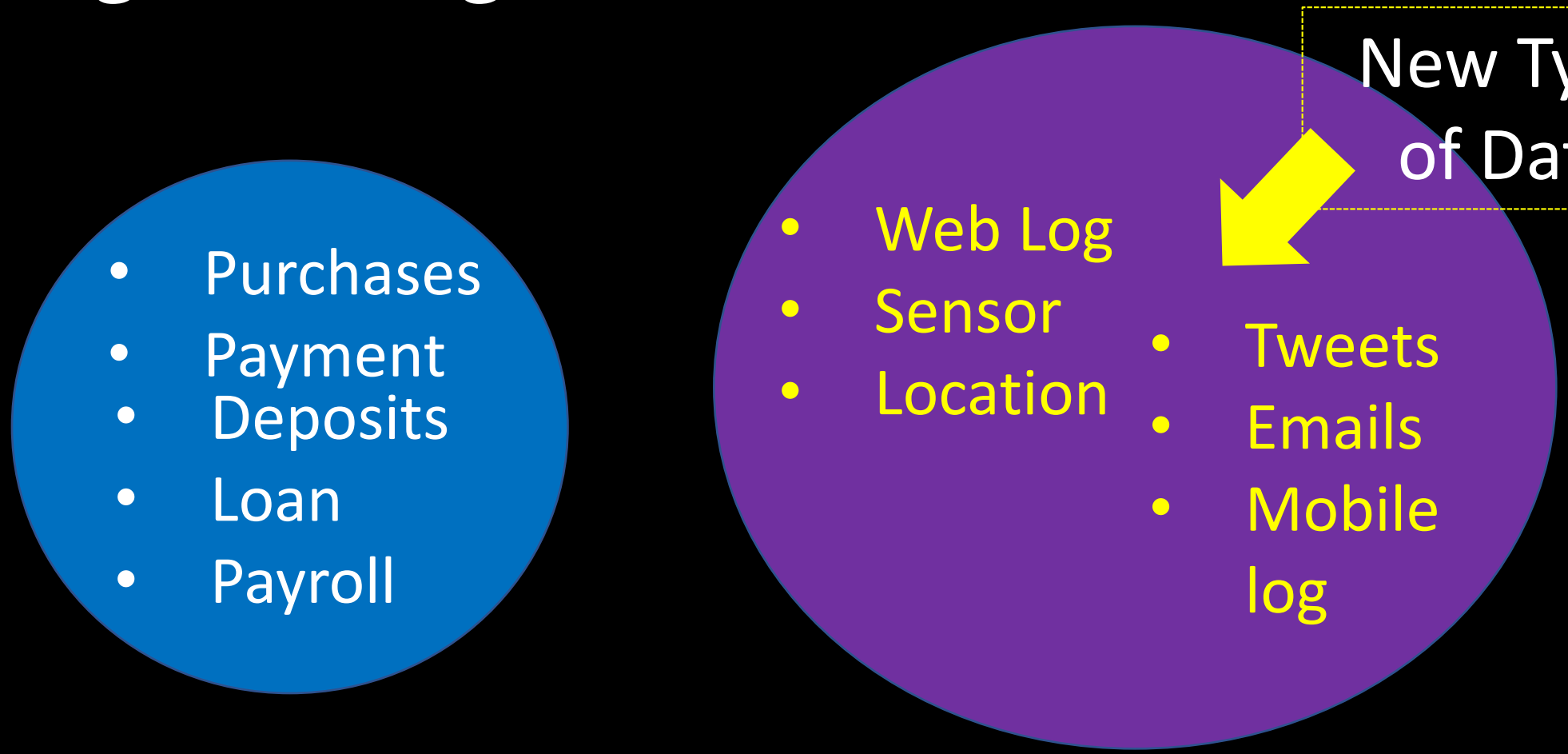
Interaction Data

No.	Cust A	Time	Activity
1	12346	5:48 May 18	Searching jkt-jog
2	12346	6:10 May 18	Searching jkt-jog
3	12346	6:10 May 18	Back to Menu
4	12346	6:15 May 18	Searching jkt-solo
5	12346	6:16 May 18	Paid jkt-solo

Integrated, structured transaction Data in single Apps

Granularity of data, completeness of customer journey and behavior

Mining the new gold!




Analyzing These New Types of Data is Now Possible
With a New Technology called BIG DATA Technology
© 2014 Teradata

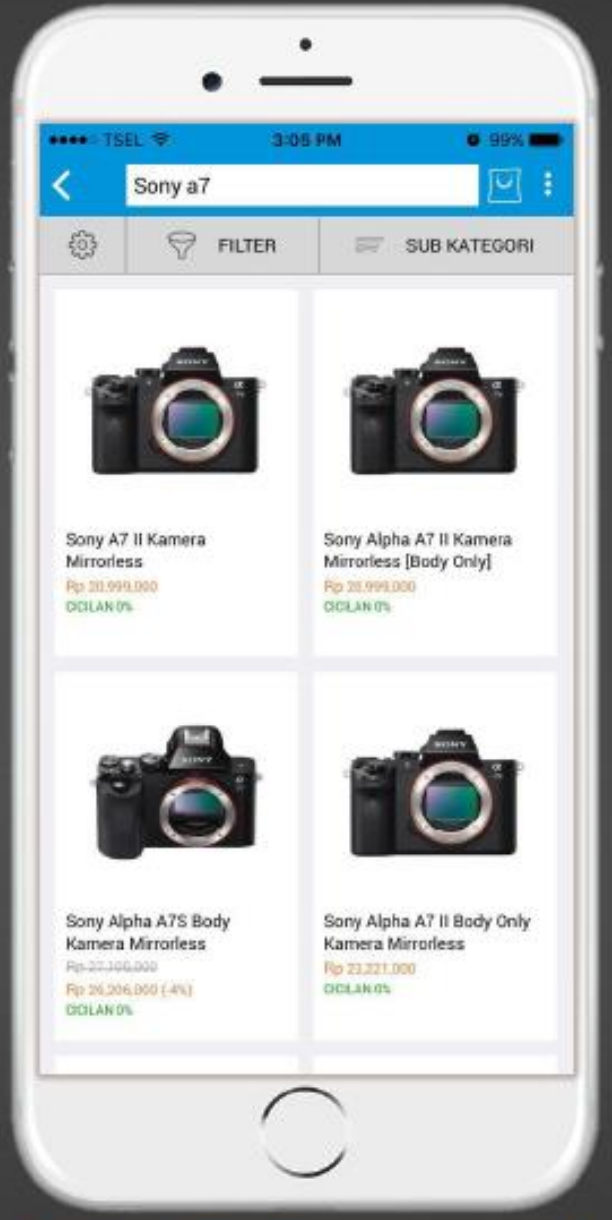
Big Data its **not a Technology** (only),
It's a movement to become more **Smart-Data Driven**

Digital → Smart → Data Driven

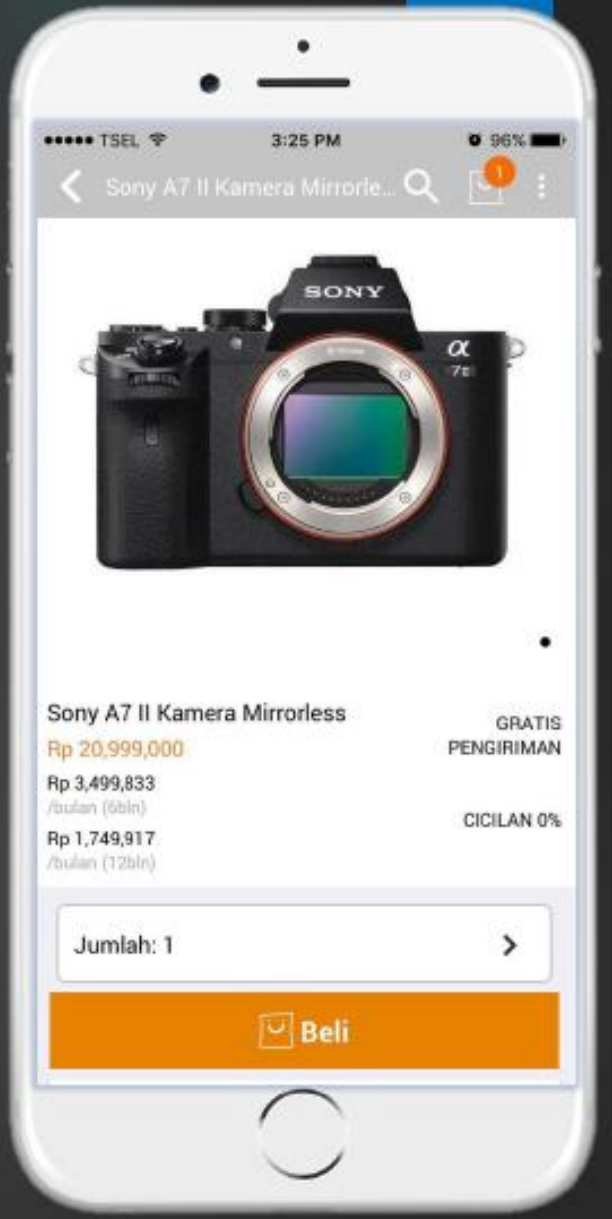
Exercise

What if we knew the transaction & Interaction data from e-commerce

Go to Homepage



Search for: Sony A7R II



Checkout

What kind of data do we have?

Traditional Data

Purchase Detail

Item Detail, Payment Detail, Shipping Detail

New Types of Data

Location

Location Where the Order is being placed

Handset

Handset being used to place the order

Clickstream

Interaction before the order is being placed

What kind of questions can we ask?

The Framework

1. Identify The Data that we have

Traditional Data	New Types of Data

2. Identify The Questions that we can ask

?	?	?

1. Identify The Data that we have

Traditional Data

- Purchase Detail

New Types of Data

- Location
- Handset
- Clickstream

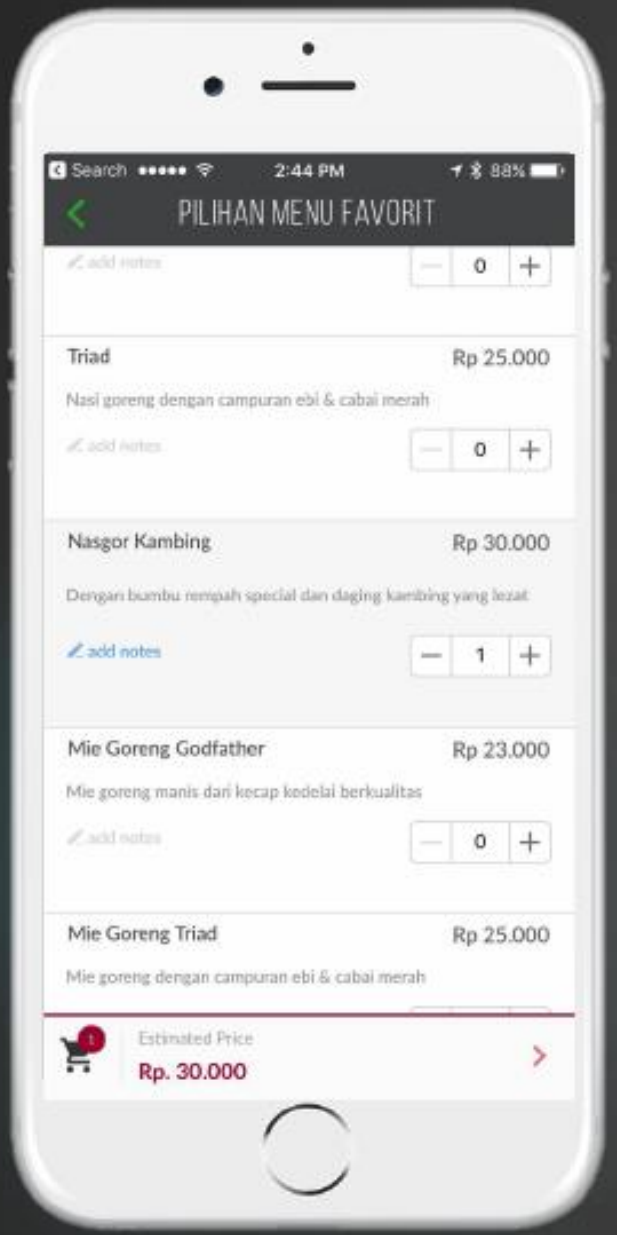
2. Identify The Questions that we can ask

Is there any client who place order near the seller's location?

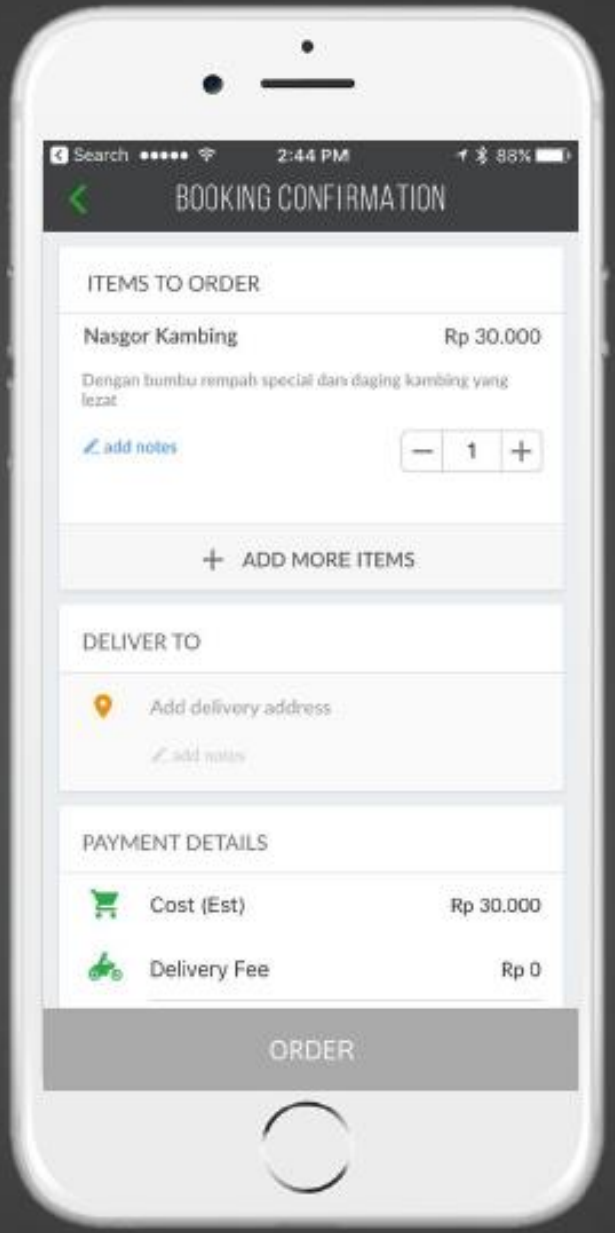
Is there any correlation between handset type with purchasing power

Whose review we should put on top of the list to drive more purchase?

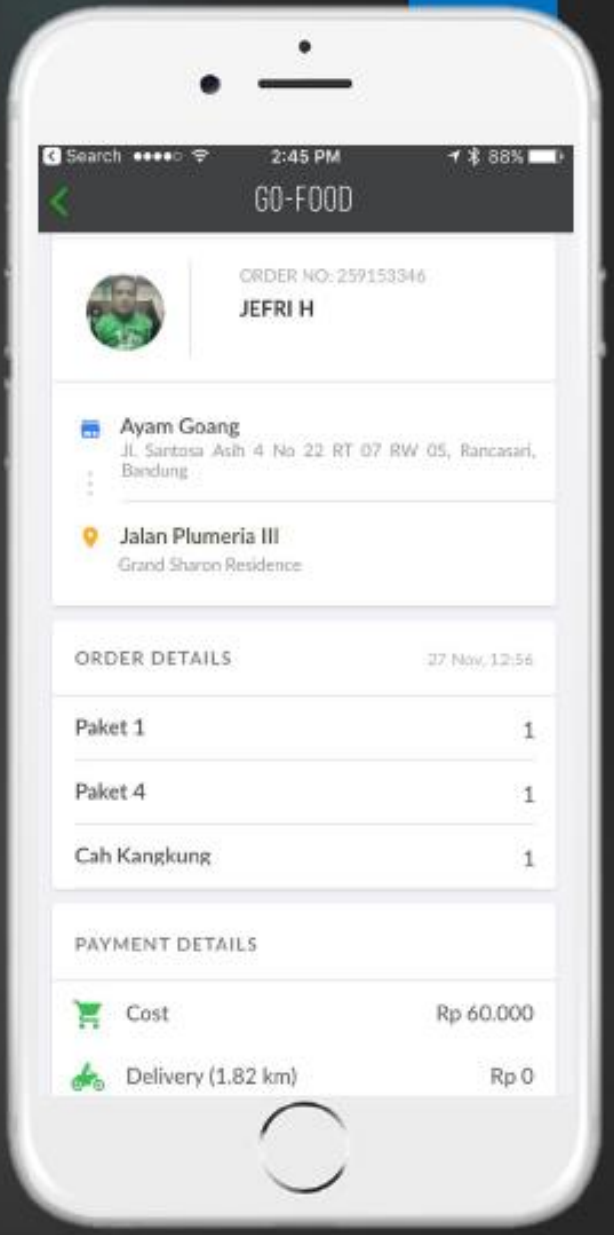
Another Exercise: Go-Food



Pick Menu



Order



Booked

1. Identify The Data that we have

Traditional Data	New Types of Data
Restaurant	Location
Menu	Cancellation
Destination	Waiting Time

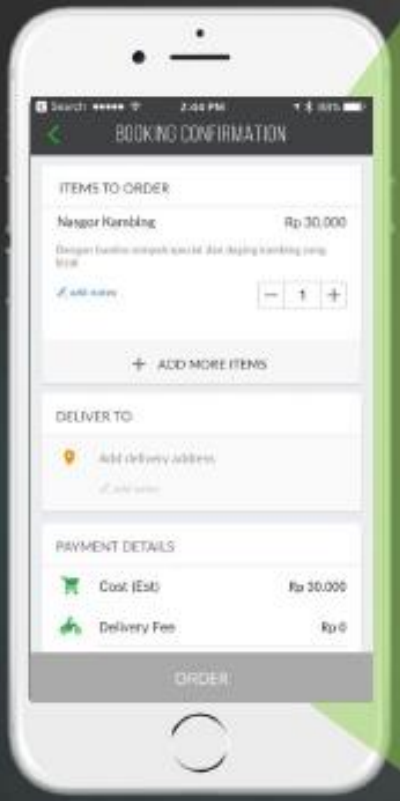
2. Identify The Questions that we can ask

Restaurants with the highest demand?

Cancellation Rate for those restaurant?

What time in a week where the order for those restaurant is the highest

How long is the maximum waiting time before customer start to cancel their order on a particular restaurant?



Ayam Goang

Jl. Santosa Asih 4 No 22 RT 07 RW 05, Rancasari, Bandung

Jalan Plumeria III

Grand Sharon Residence

ORDER DETAILS 27 Nov, 12:56

Paket 1 1

It's Peak Hours. The average waiting time for this restaurant is 1.5 hour. Do you still want to proceed with the order?

No Yes

Cost Rp 60.000

Delivery (1.82 km) Rp 0

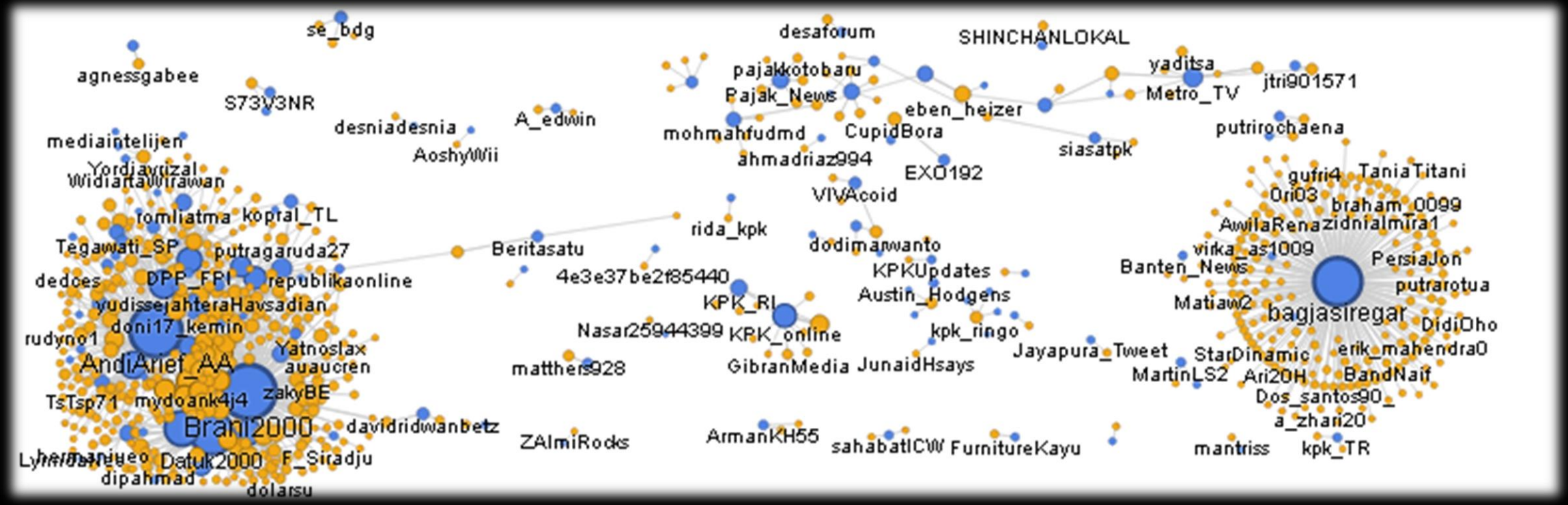
Reduce order cancellation

Reduce driver frustration

Manage customer expectation

Others Data : Social Media

Terdapat beberapa "kelompok" yang menjadi **sumber utama Re-Tweet** di Twitter, pada tweet yang mengandung kata *Korupsi* *

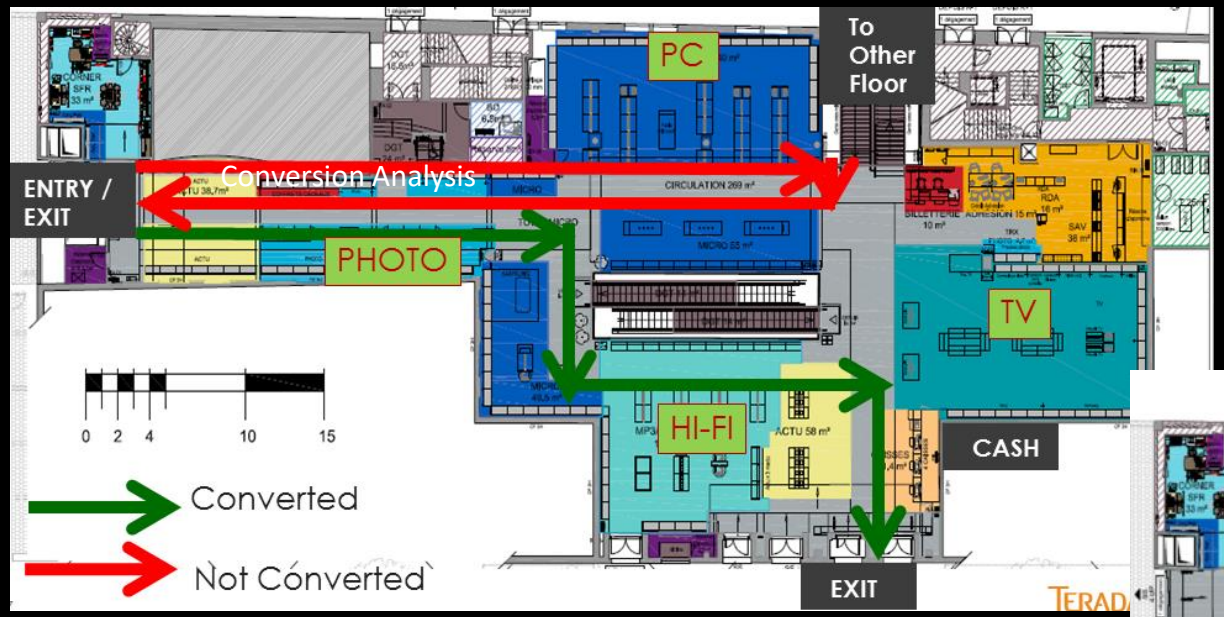


Sebuah kelompok jaringan terbentuk dari **beberapa sumber Account Tweet** dimana yang paling aktif bukanlah Media atau Institusi pemerintah melainkan account Individu

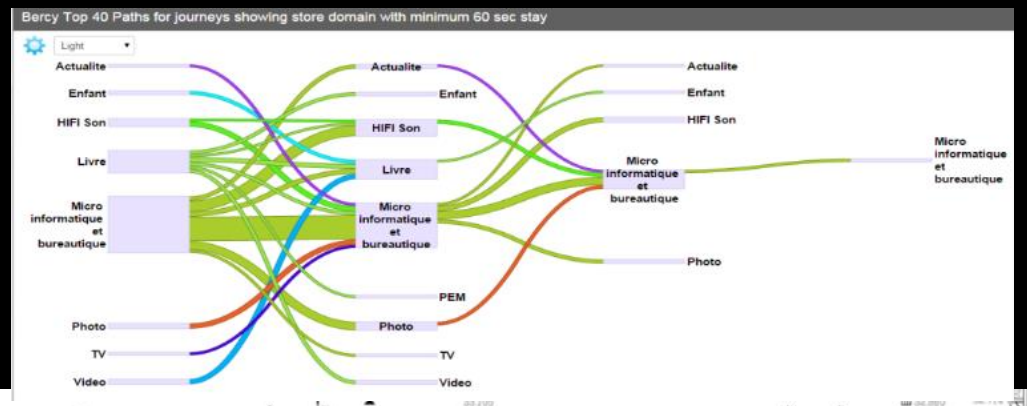
Sebuah kelompok jaringan terbentuk dari Satu **sumber Account Tweet** dimana yang Account tersebut menggunakan profil individu

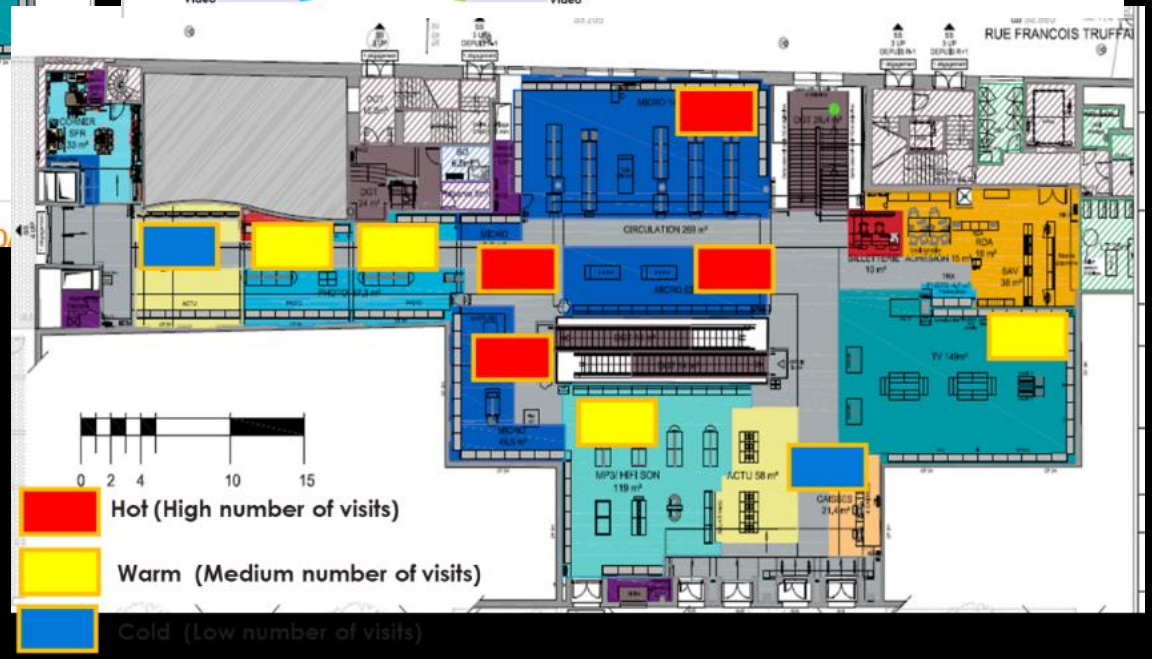
Others Data : Sensor

Mengubah store menjadi Digital?



Path to Purchase





Heat Map

- re-organization of physical layout of certain shops
- Better selection of “in-store” promotional zones
- Salesperson training on multiple domains and to increase availability in certain areas

Image Recognition And Drones

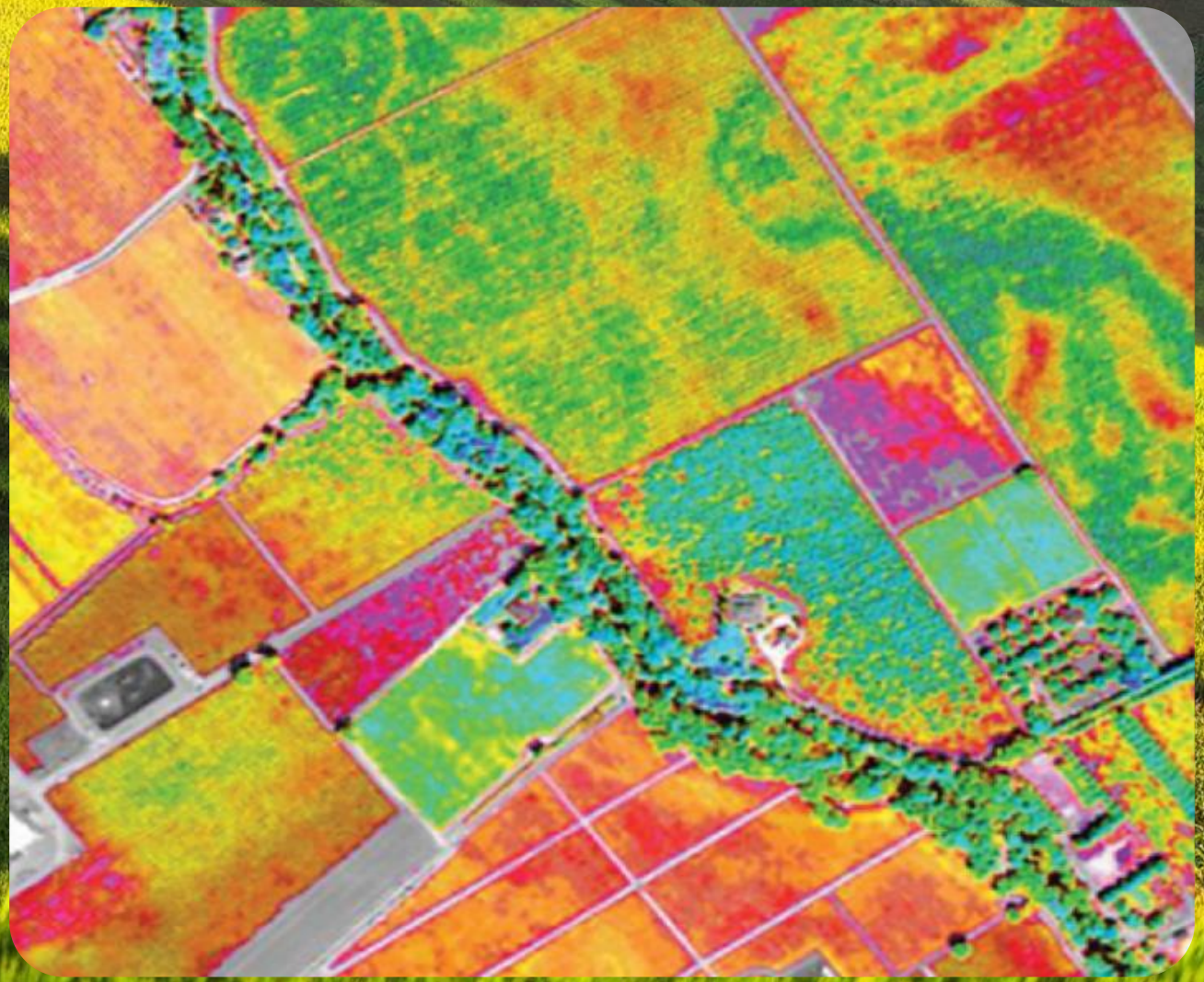
Capturing image data

Credit Rating Map

Sensor

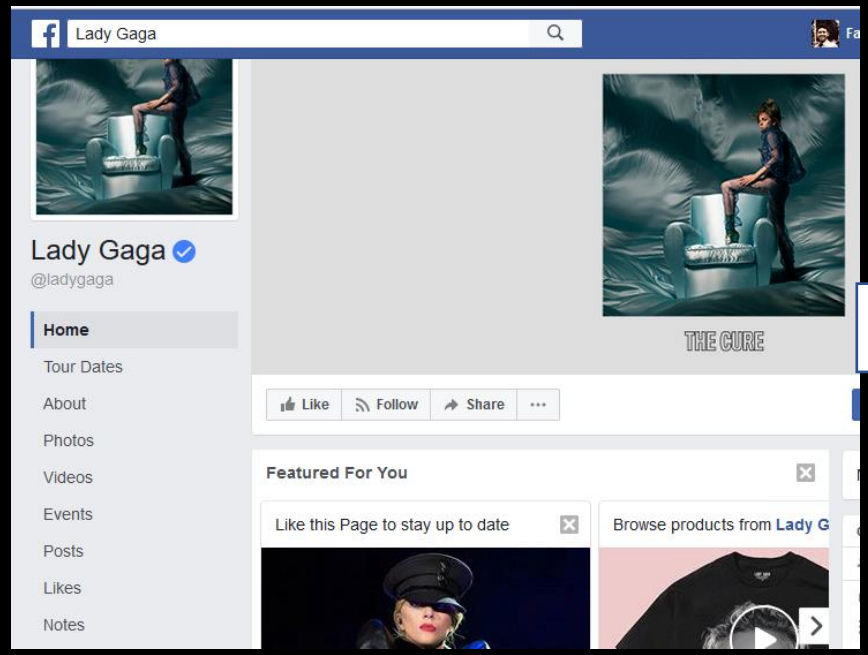
Drone

Super Computer



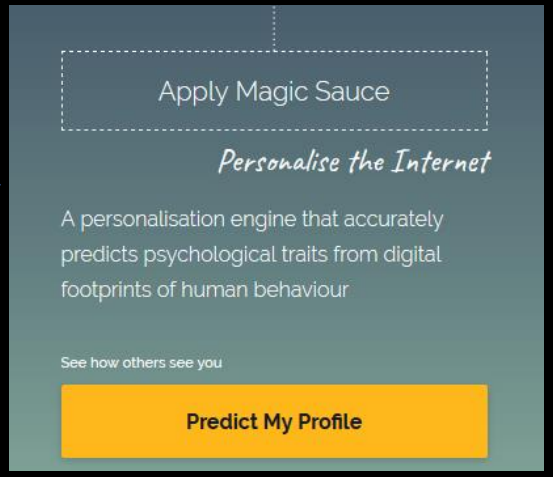
Others Data : Psychometric

Facebook



"liking" Lady Gaga correlated highly with extroversion

Facebook connected apps



Request access

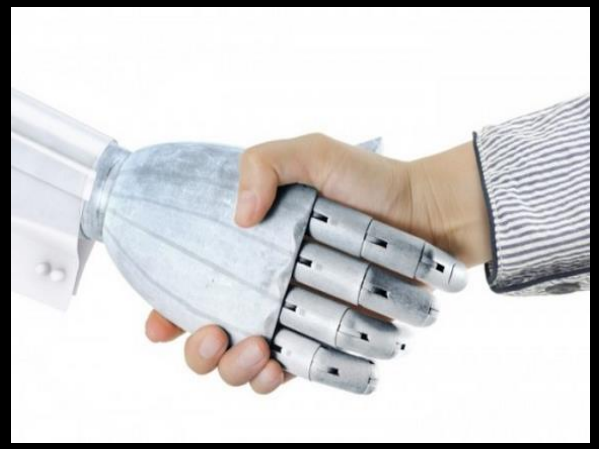
Big 5 personality

- **Openness**
- **Conscientio-usness**
- **extroversion**
- **agreeableness**
- **neuroticism**

Develop personality-specific messaging

4 Area pemanfaatan non-traditional data dan advanced analytics

1. Automation

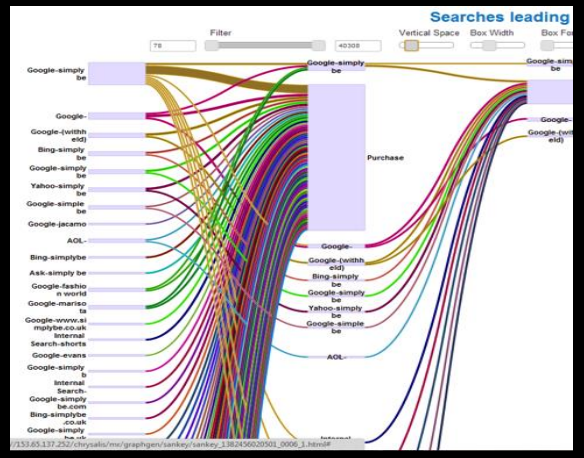


- Reduce manual process
- Standardize easily
- Auto-recommendation

2. Segmentation

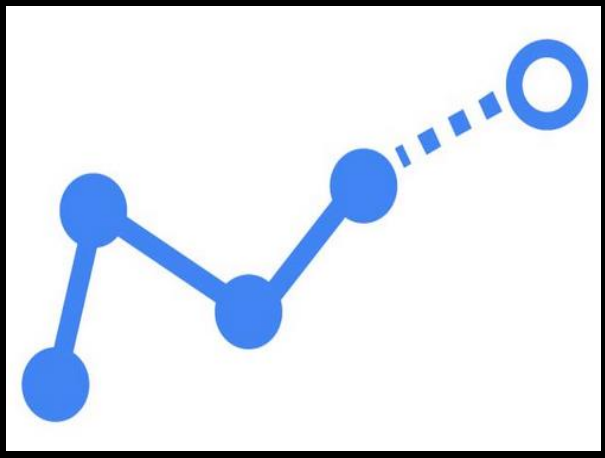
- Which product Is the best for which customers
- Customer journey personalization

3. Pattern recognition



- Track customer payment pattern and flag possible fraud
- Identify outlier

4. Predictive model



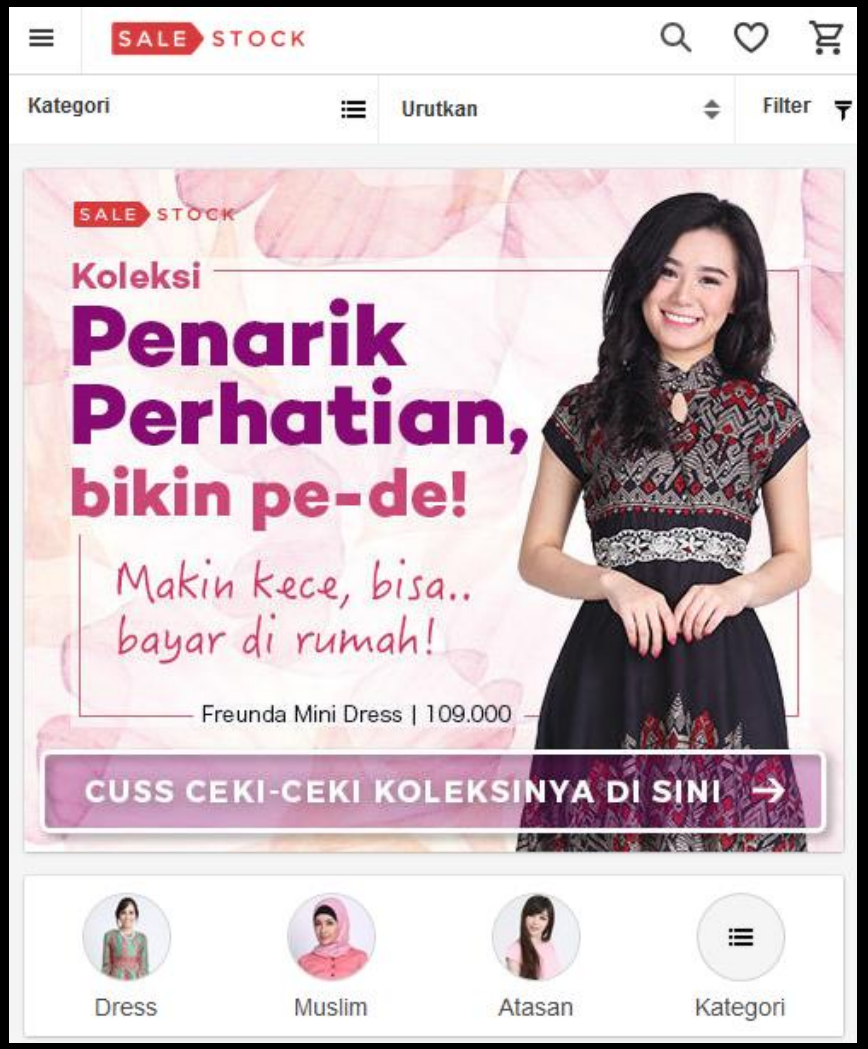
- Better assess probability of default
- Credit score model

Use cases Digital Banking memanfaatkan **New type of Data**

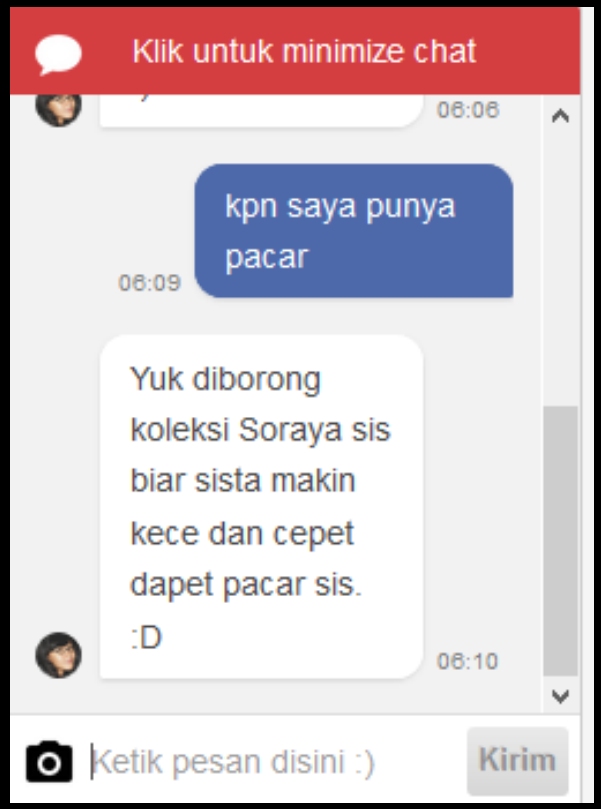
	Description	Impact
A	Latin American bank uses **supermarket data** to develop new models for risk scoring	~30% lower credit losses
B	Brazilian insurance company identifies groups with highest claims rates and fraud rates **using mobile phone data**	15-20% increase in profitability
C	African mobile financial services provider assigns credit scores for unsecured **loan using mobile money usage data**	Product recently launched
D	North American small business lender uses data analytics and non-traditional data such as **online reviews to target customers**	30-40% reduction in customer acquisition costs
E	Major US insurer leverages **social media marketing** to drive sales and deepen customer relationships	22% increase in production of sales reps
F	US-based consumer finance firm created rapid scoring model using financial data and unstructured **social media data**	60% reduction in default risk
G	Home equity lender incorporated **customer relationship data** into credit risk models	25% loss reduction; Increased approval rate
H	Asian lender evaluates credit risk by analyzing mobile usage data and migration patterns	Model proved predictive of credit risk

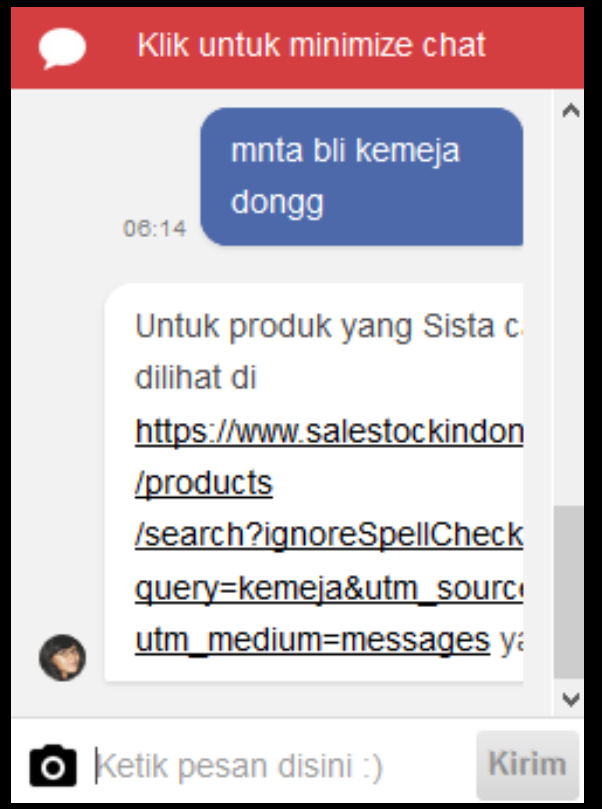
SOURCE: McKinsey and Team Analysis

Pemanfaatan big data di digital business Indonesia



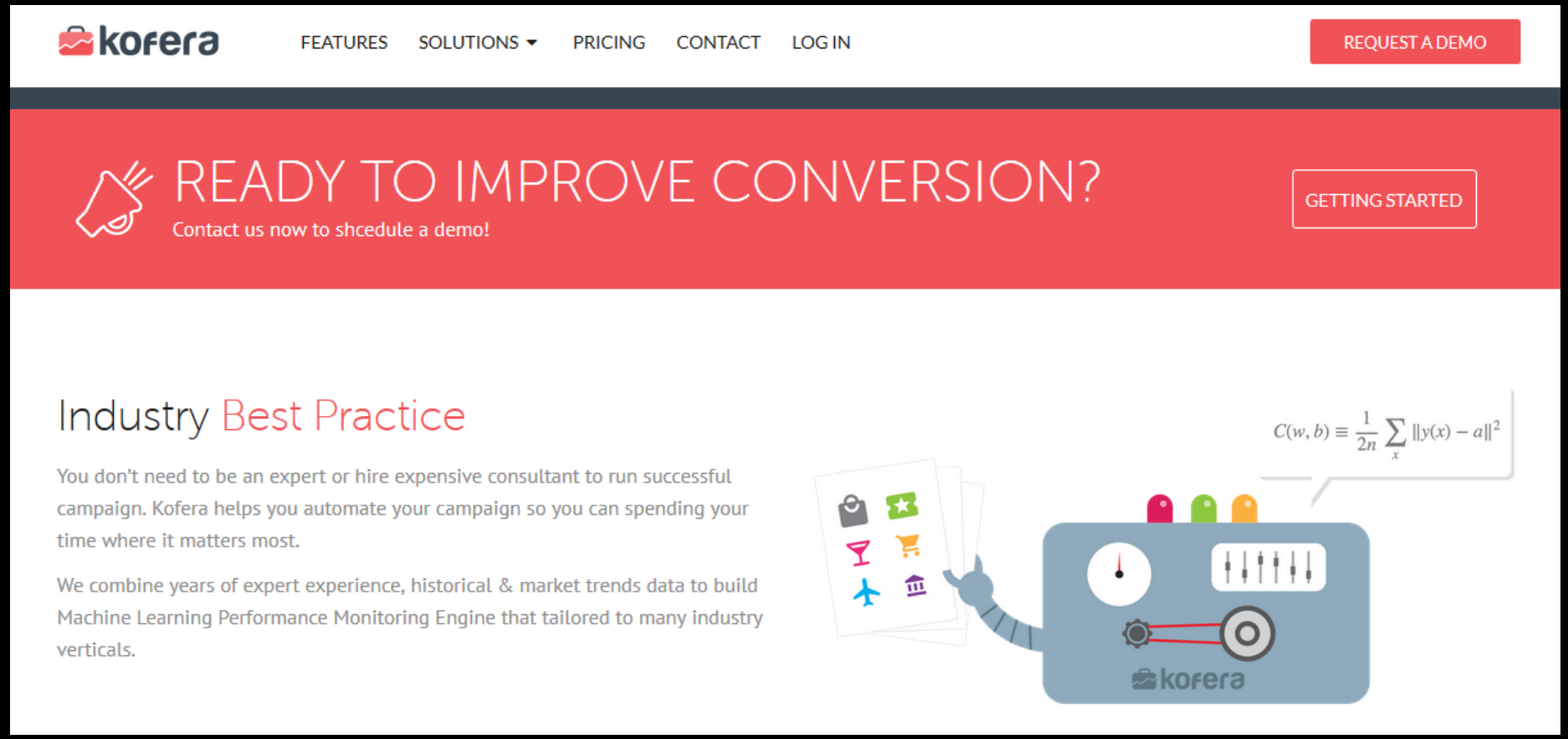
Chatbot for suburban/rural area





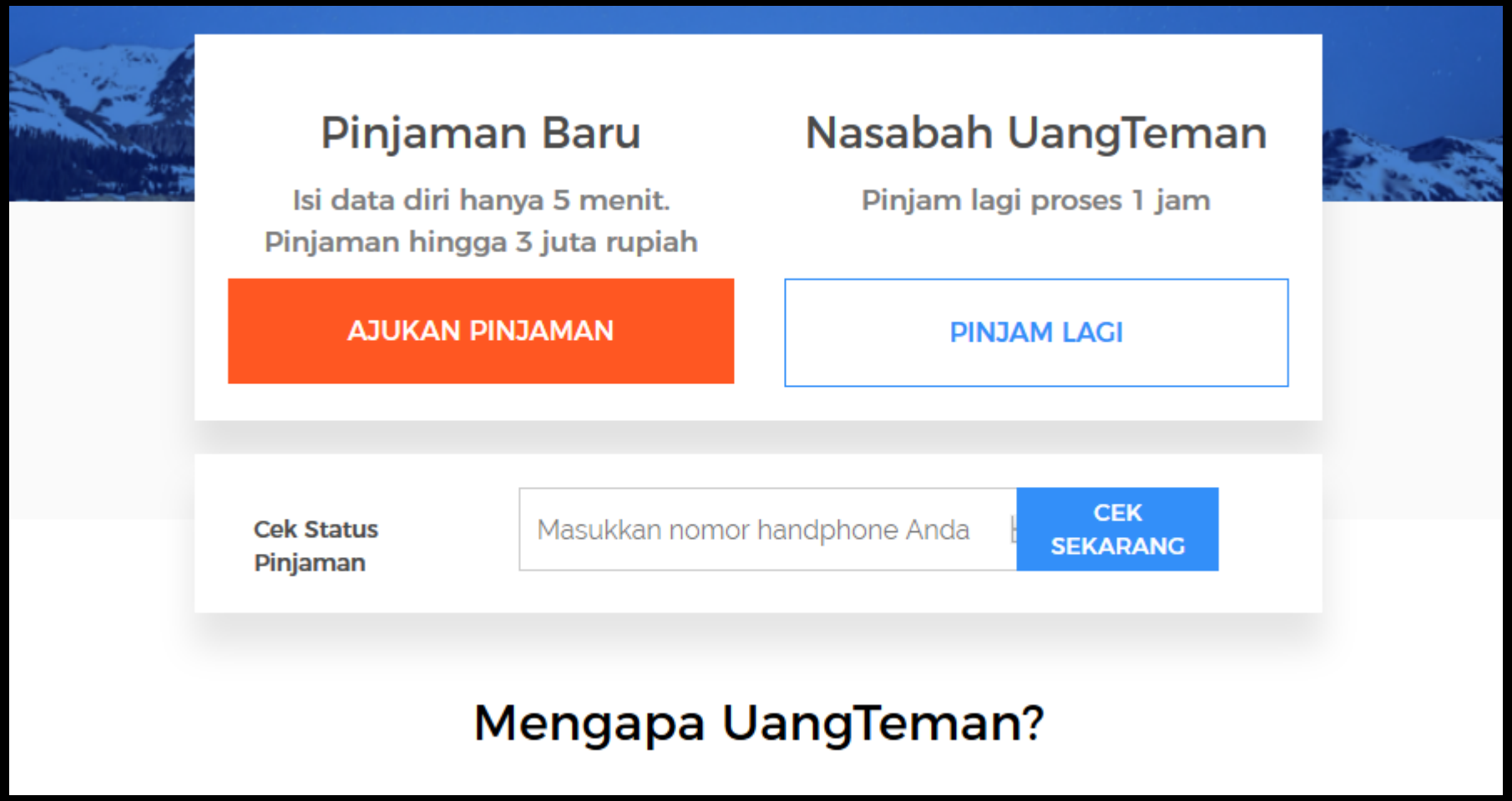
Help answering unstructured questions

Pemanfaatan big data di digital business Indonesia



Pyhsics | Scientist | Developer

Pemanfaatan big data di digital business Indonesia

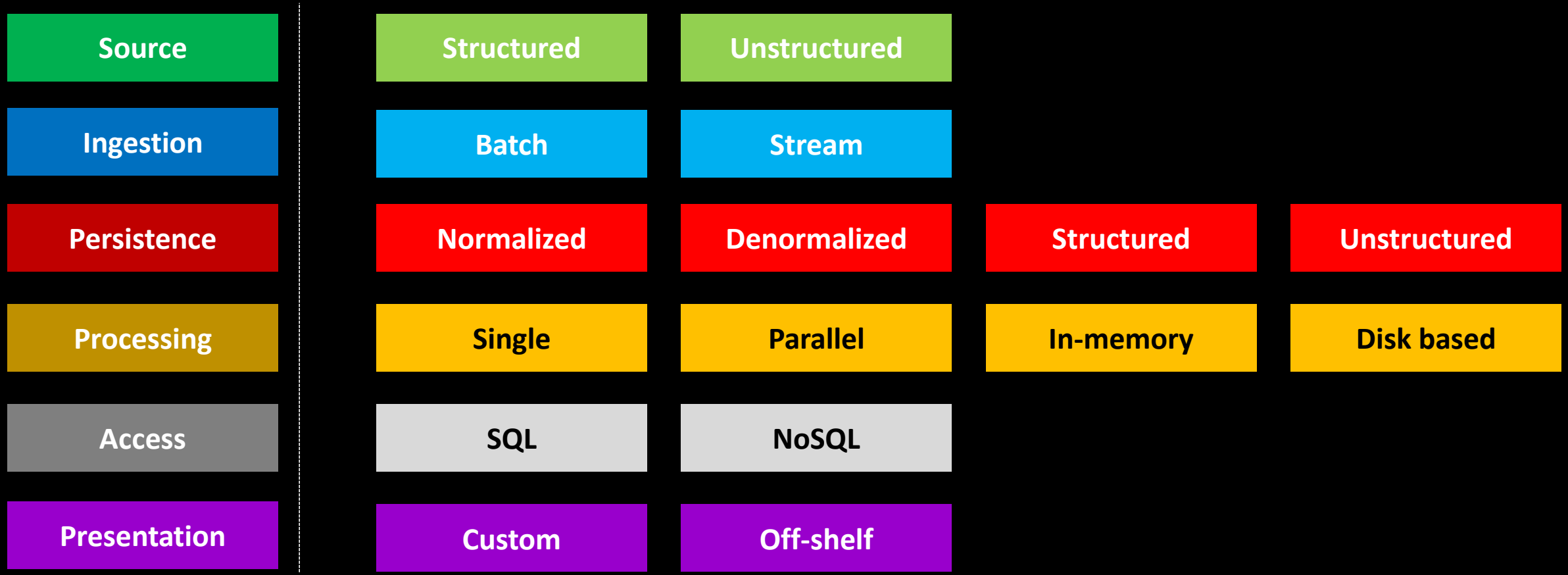


Economist | Quant | Developer

Bagaimana membangun kapabilitas **analytics-data driven ?**

1. **Analytic Leadership**
2. **Analytic Environment**
3. **Analytic Talent**

1. Analytic Leadership
2. Analytic Environment
3. Analytic Talent



Architecture Building Blocks

1. Analytic Leadership
2. Analytic Environment
3. Analytic Talent

Collaborative Access

Access

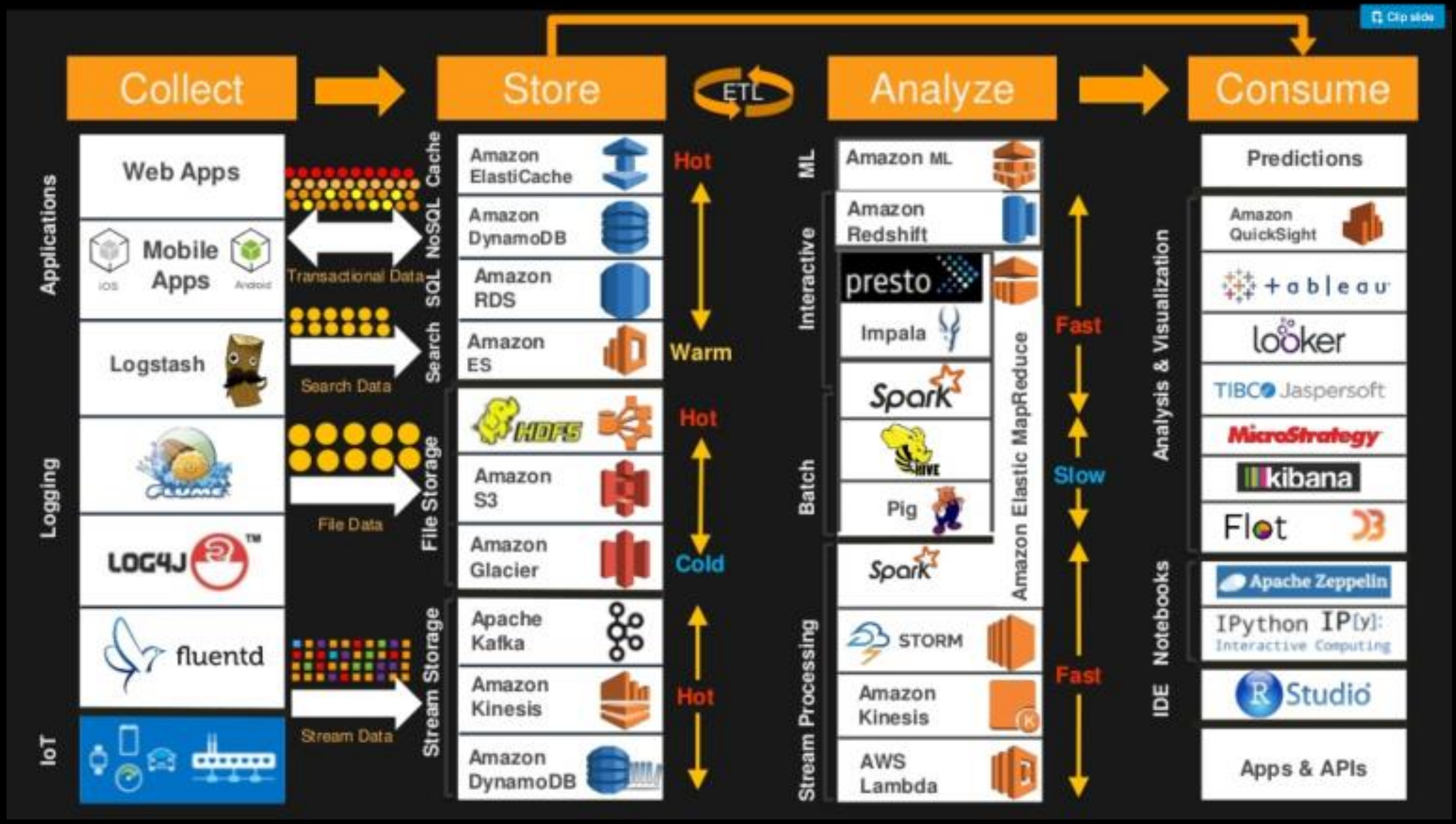
Source | Ingestion | Persistence | Processing | Presentation

Read from multiple sources and format

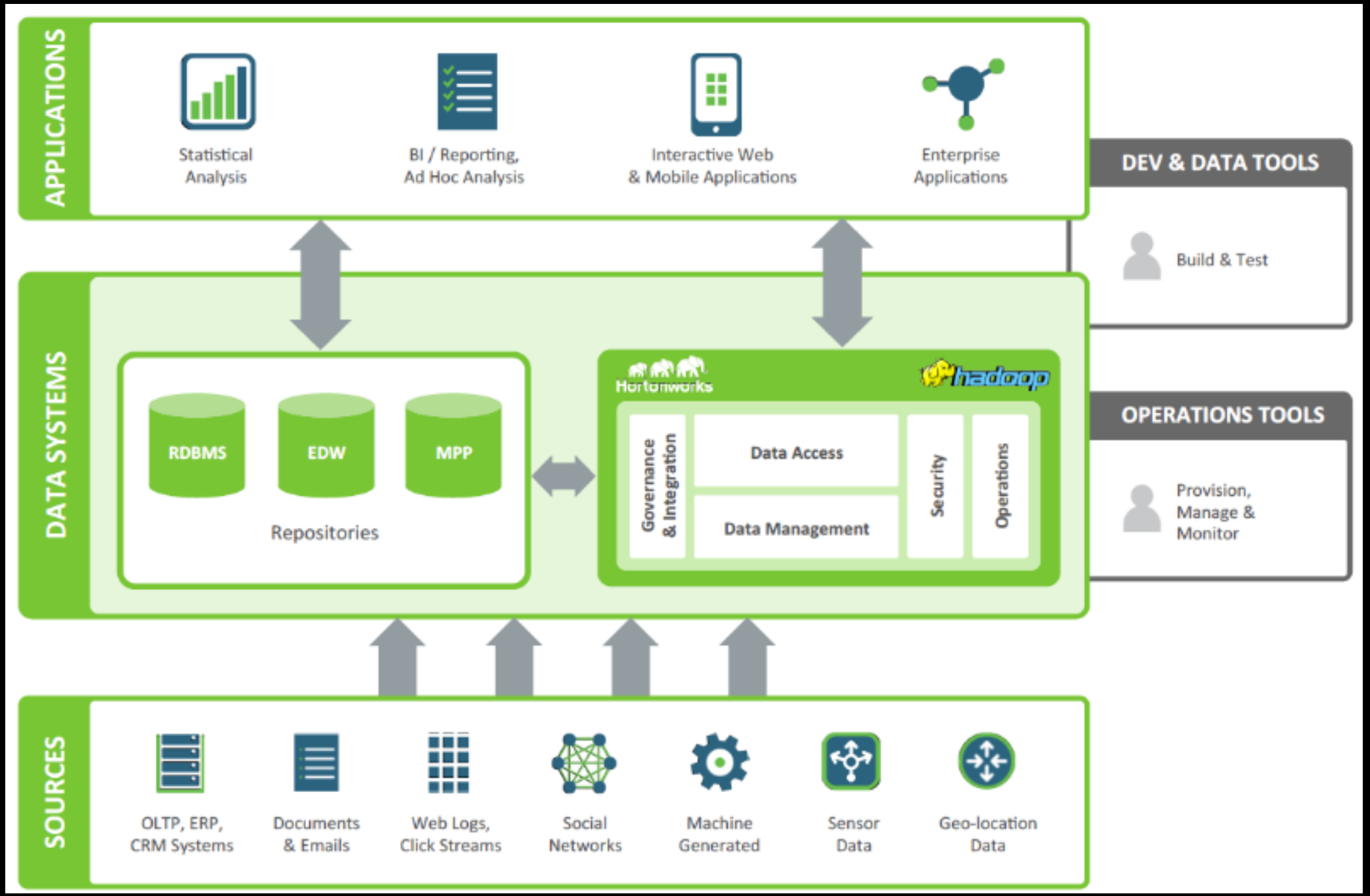
- Integrated
- Scalable

Interactive for iterative development

Ideal Analytic environment

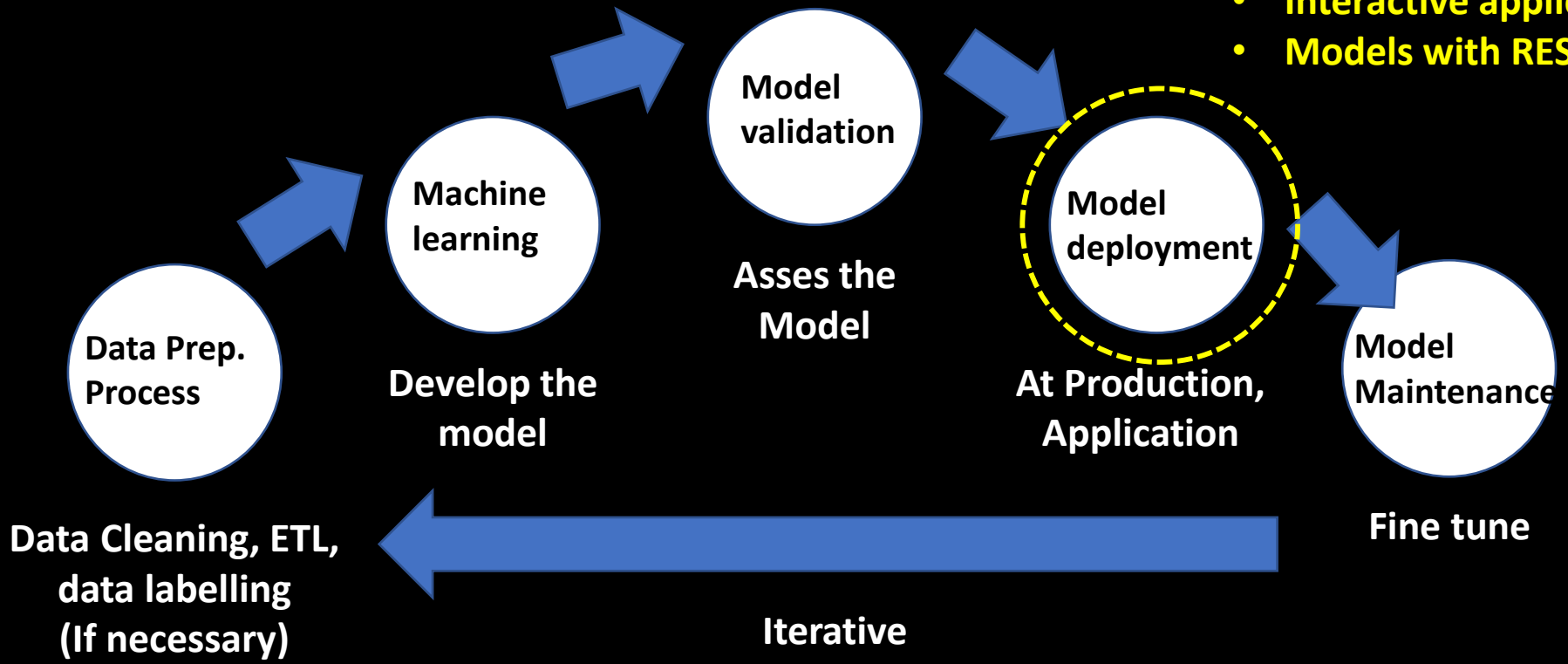

Collect
Store
ETL
Analyze
Consume
Applications
Web Apps
Mobile Apps
iOS
Android
Logging
Logstash
LOG4J
fluentd
IoT
Transactional Data
Search Data
File Data
Stream Data
NoSQL Cache
SQL
Search
File Storage
Stream Storage
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
HDFS
Amazon S3
Amazon Glacier
Apache Kafka
Amazon Kinesis
Amazon DynamoDB
Hot
Warm
Hot
Cold
Hot
ML
Interactive
Batch
Stream Processing
Amazon ML
Amazon Redshift
presto
Impala
Spark
HIVE
Pig
Spark
STORM
Amazon Kinesis
AWS Lambda
Amazon Elastic MapReduce
Fast
Slow
Fast
Predictions
Analysis & Visualization
Amazon QuickSight
tableau
looker
TIBCO Jaspersoft
MicroStrategy
kibana
Flot
Notebooks
Apache Zeppelin
IPython IP[y]: Interactive Computing
IDE
RStudio
Apps & APIs

Big Data Architecture



1. Analytic Leadership
2. Analytic Environment
3. Analytic Talent

- Report
- Dashboard
- Interactive application
- Models with REST API



Bridging Data Science to Big Data

Bagaimana membangun kapabilitas **analytics-data driven ?**

1. **Analytic Leadership**
2. **Analytic Environment**
3. **Analytic Talent**

How to start build analytics talent ?

Toolbox

- **Notebook :** Jupyter/ Ipython, R
- **Github.com**
- **SQL :** heidisql.com
- **Hadoop :** Spark, MapReduce

Data

- **Data.go.id**
- **Jakarta.data.go.id**
- **Databoks.katadata.co.id**

Media

- **Kdnugget.com**
- **Analyticvidhya.com**
- **Oreilly.com**

Practice

- **Kaggle.com**
- **Data competition**
- **Pro bono for NGO / Govt**

Learn

- **datasciencemasters.org**
- **udemy.com**
- **Offline :** Dattabot Dojo, DSI Bootcamp, Pacmann AI

Engage

- **Data Science Indonesia**
- **IDBigdata.com**
- **Open Data Club**

Data Science Indonesia Bootcamp






Launching event

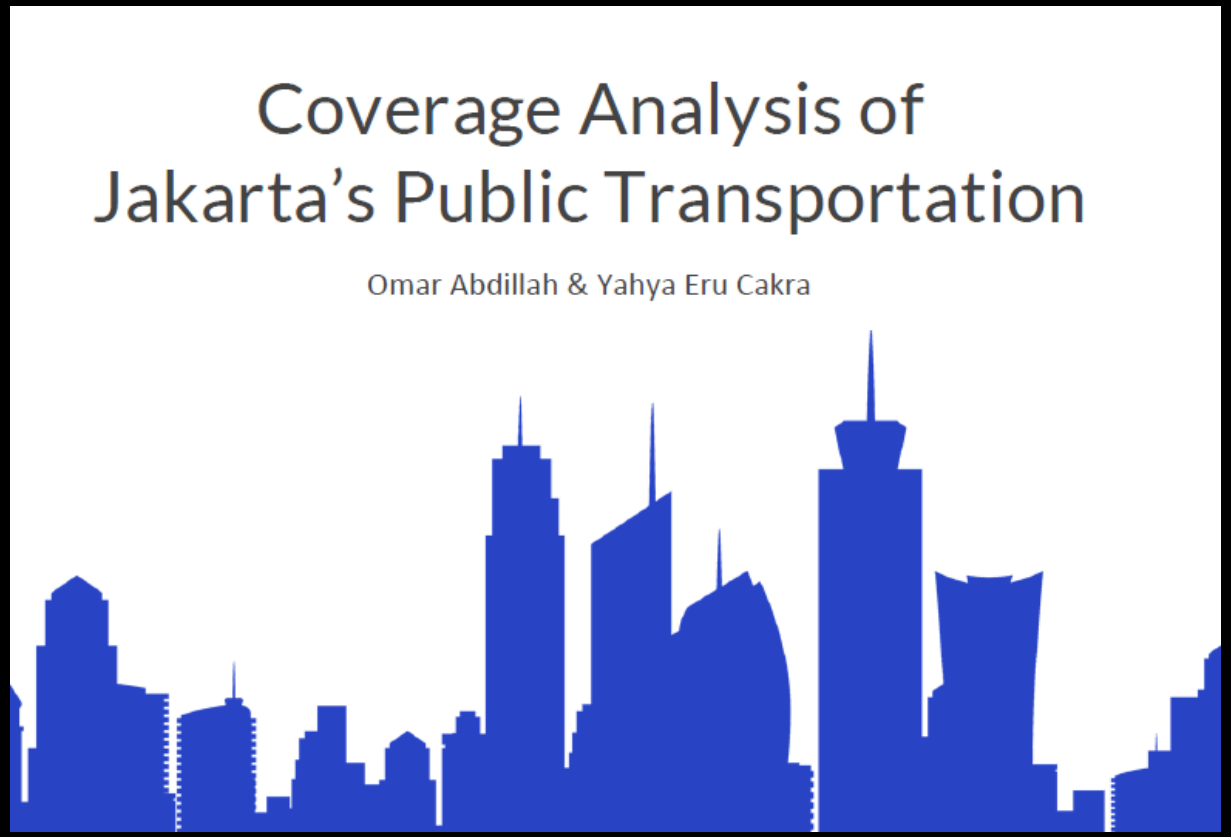
DSI Bootcamp 2016 : Capstone Project 1

Omar Abdillah

Lead of Analytic at Data Management and Analytic Laboratory (SRIN), **SAMSUNG**

Yahya Eru Cakra

Data Scientist at KMK **(Kreatif Media Karya)** : liputan6.com, video.com etc

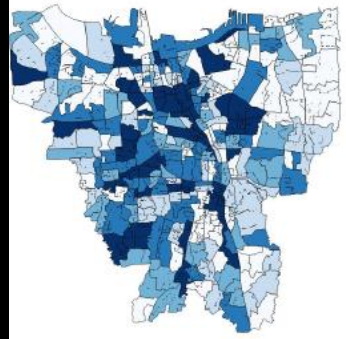


By: Omar Abdillah & Yahya Eru

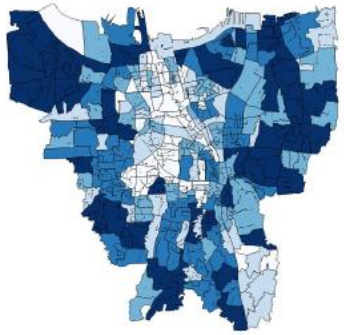
DSI Bootcamp 2016 : Capstone Project 1

Village-based Approach

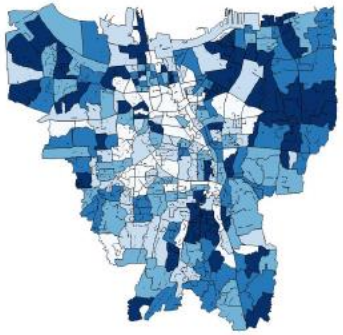
Total Routes



Total Population



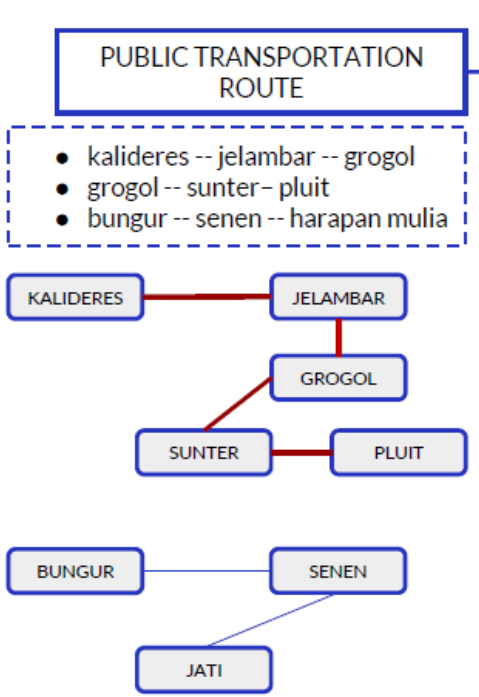
Density

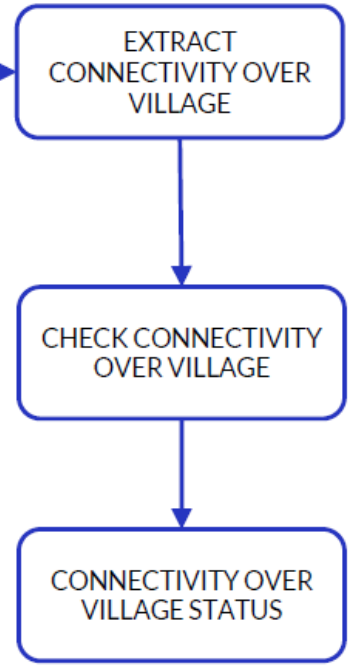


Out of **36,856** pairs of villages, there are

7,952 pairs of vilages that are not connected by public transportation

Finding Connectivity over village





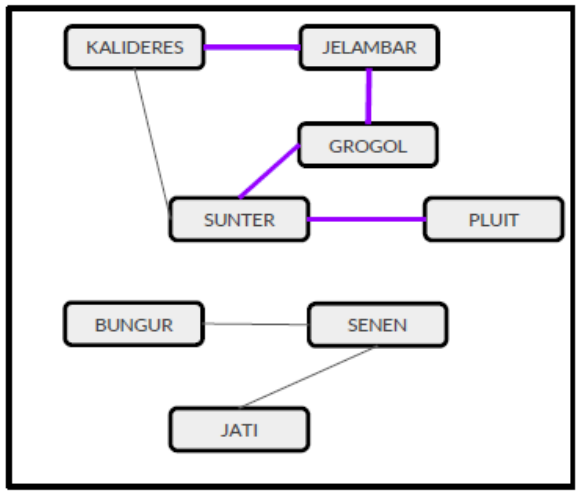
FROM	TO	STATUS
KALIDERES	JELAMBAR	1
JATI	PLUIT	0
JELAMBAR	PLUIT	1

We used graph theory algorithm called **Dijkstra Algorithm** for finding connectivity over village

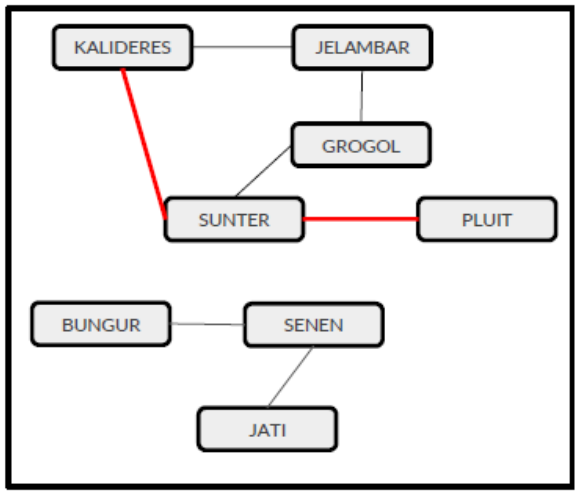
By: Omar Abdillah & Yahya Eru

DSI Bootcamp 2016 : Capstone Project 1

3. Find shortest route using Dijkstra



4 STEPS



2 STEPS

In this research, we utilized **Djikstra Algorithm** to generate shortest step between two villages

4. Average time consumed

- On average, you have to take **2 routes** for reaching your destination
- One route can takes approximately **35 – 40 minutes** (on average)
- So, in total you will need approximately **70 – 80 minutes!**

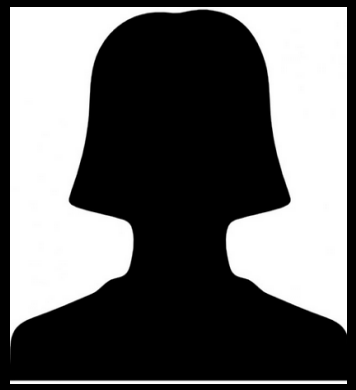
No wonder most people prefer private transportation than public transportation

By: Omar Abdillah & Yahya Eru

DSI Bootcamp 2016 : Capstone Project 2

Ines Dwi Andini

Data Quality Management



Rossi Azmatul Barro

Data Analyst in PT Prudential Life Assurance

By: Ines Dwi Andini & Rossi Azmatul Barro

DSI Bootcamp 2016 : Capstone Project 2

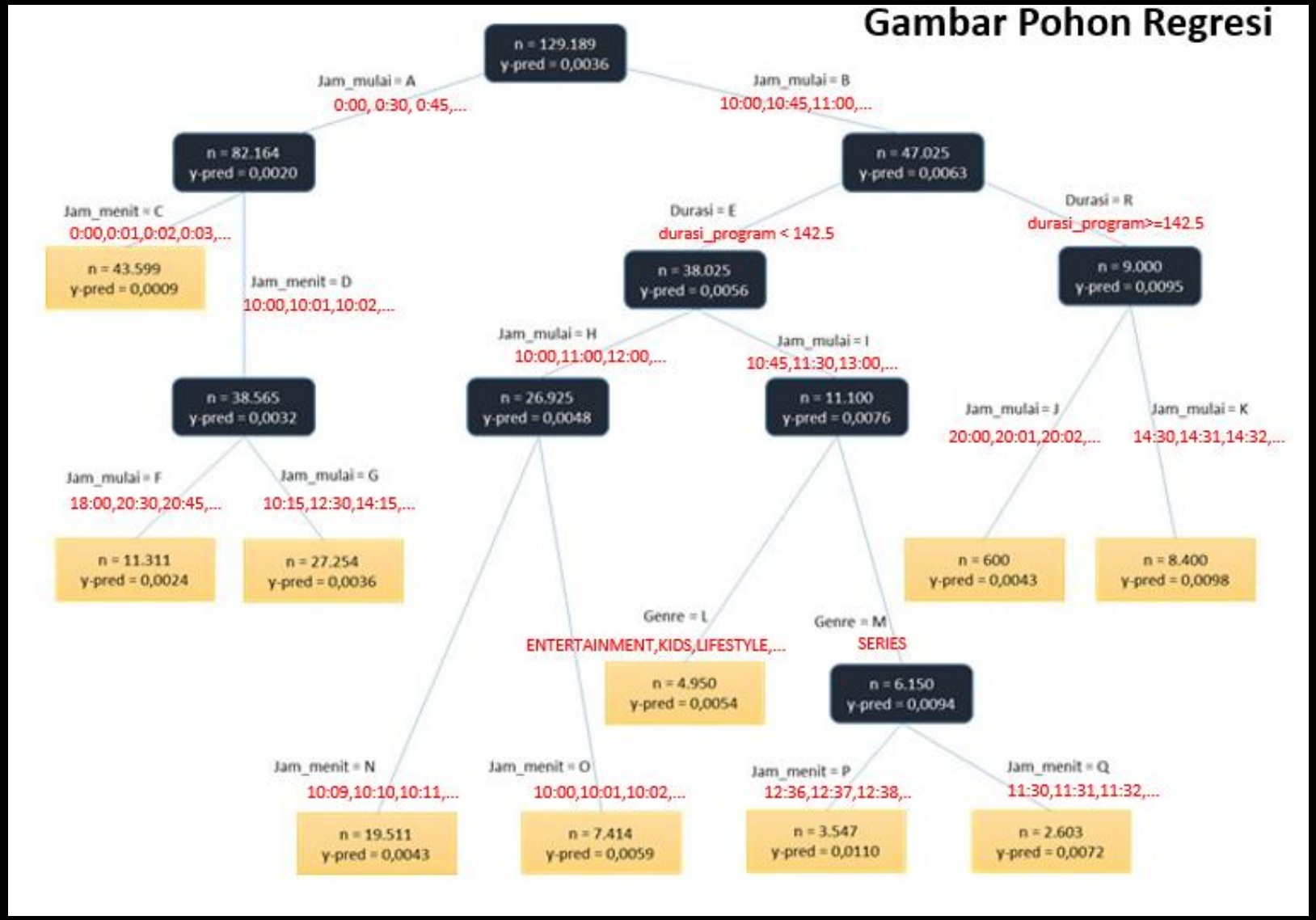
Model ke-	R-squared	Prediktor yang Diuji	Prediktor yang Masuk dalam Model CART
Model 1	79,2%	jam_menit genre jam_mulai hari, durasi	jam_menit genre jam mulai durasi
Model 2	76,3%	genre jam_mulai hari durasi	genre jam_mulai durasi
Model 3	78,0%	jam_menit jam_mulai hari durasi	jam_menit jam_mulai durasi
Model 4	73,9%	jam_menit genre hari durasi	jam_menit genre durasi
Model 5	79,3%	genre jam_mulai jam_menit durasi	genre jam mulai jam_menit durasi
Model 6	77,9%	jam_menit genre jam_mulai hari	jam_menit genre jam_mulai

- Dari enam model yang dibuat diperoleh model pohon dengan **R-squared 79.3%** dengan **prediktor penyekat pertama adalah jam mulai**
- Model terbaik dengan **prediktor: genre, jam mulai, jam menit, dan durasi**

Model accuracy / comparison
Best model with variable :

Genre, Hours start, Duration and time

DSI Bootcamp 2016 : Capstone Project 2



Highest rating : B - E - I - M - P

Lowest rating: A - C