pusilkom ui
quality results

Introduction to Big Data and Data Preparation

Denny, Ph.D.

Data Mining for Big Data

Short Course

Pusat Ilmu Komputer UI

16-20 Juli 2018

pusilkom ui
quality results

DATA MINING

Data Mining

- "Data mining is the analysis of (often large) ***observational data sets*** to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

  Hand, Mannila & Smyth

- observational vs experimental data sets

Business Value

TABLE 11.1 Business Value of BI Analytical Applications

Analytical Application	Business Question	Business Value
Customer segmentation	What market segments do my customers fall into and what are their characteristics?	Personalize customer relationships for higher customer satisfaction and retention.
Propensity to buy	Which customers are most likely to respond to my promotion?	Target customers based on their need to increase their loyalty to your product line. Also, increase campaign profitability by focusing on the most likely to buy.
Customer profitability	What is the lifetime profitability of my customers?	Make business interaction decisions based on the overall profitability of customers or customer segments.
Fraud detection	How can I detect which transactions are likely to be fraudulent?	Quickly detect fraud and take immediate action to minimize cost.
Customer attrition	Which customers are at risk of leaving?	Prevent loss of high-value customers and let go of lower-value customers.
Channel optimization	What is the best channel to reach my customers in each segment?	Interact with customers based on their preference and your need to manage cost.

Source: Ziama and Kasher (2004). Courtesy of Teradata, divison of NCR Corp.

Data Mining Functionality

- Association
  - From association, correlation, to causality
  - Finding rules like “A -> B”
- Classification and Prediction
  - Classify data based on the values in a classifying attribute
  - Predict some unknown or missing attribute values based on other information
- Cluster analysis
  - Group data to form new classes, e.g., cluster houses to find distribution patterns
- Outlier and exception data analysis
- Time series analysis (trend and deviation)
  - Trend and deviation analysis: regression, sequential pattern, similiar sequences e.g. Stock analysis

Data Mining Should Not be Used Blindly

- Data mining find regularities from history, but history is not the same as the future.
- Association does not dictate trend nor causality.
- Some abnormal data could be caused by human.

Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and projection:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Interpretation: analysis of results.
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

pusilkom ui
quality results

BIG DATA

pusilkom ui
quality results

What's Big Data?

No single definition; here is from Wikipedia:

- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

What's Big Data?

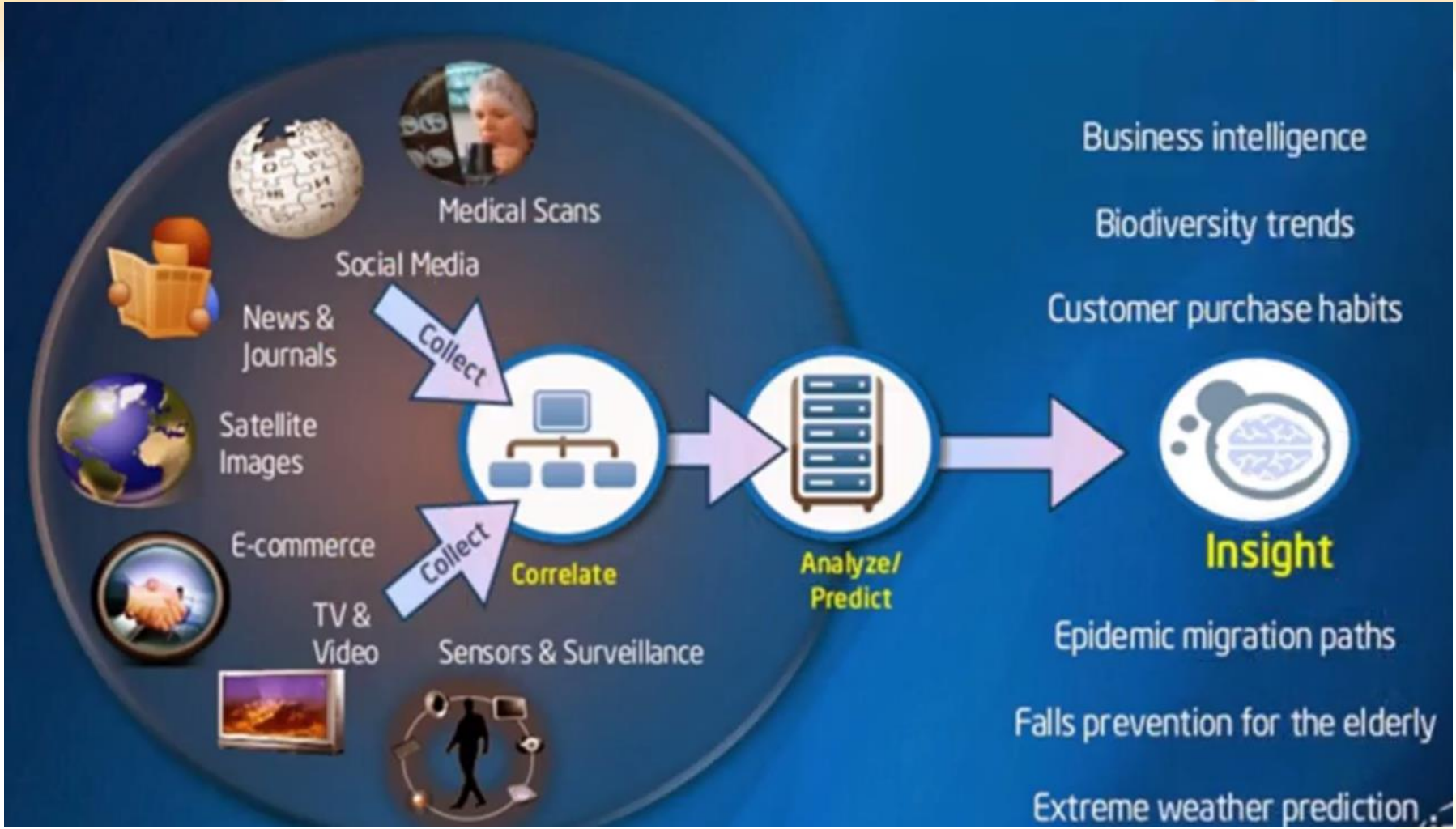
- Big data is a buzzword, or catch-phrase, meaning a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques.
- In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.
  - http://www.webopedia.com/TERM/B/big_data.html

What's Big Data?

- Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions.
- This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations.

Big Data

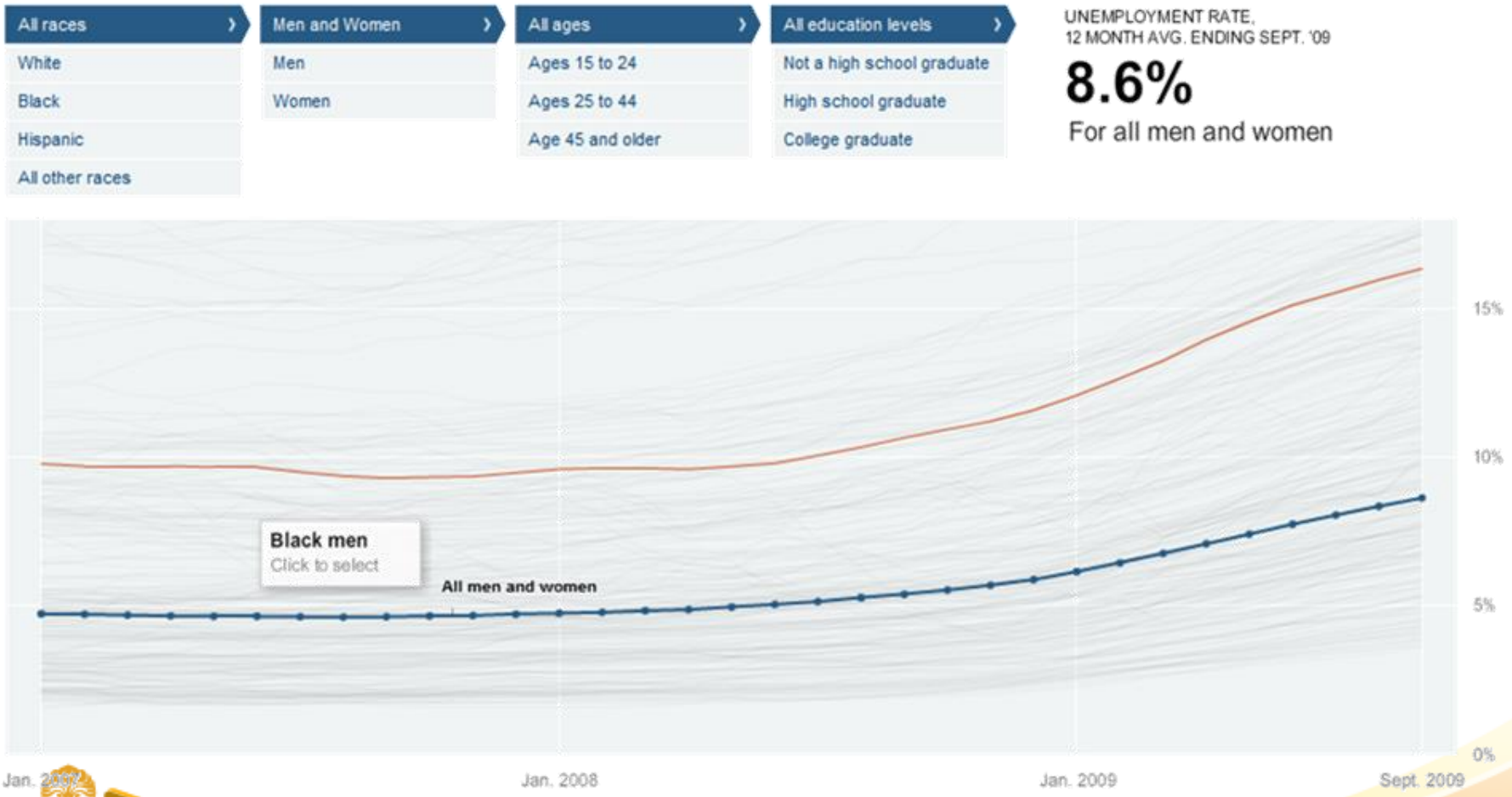
- The challenges include:
  - capturing data,
  - curation,
  - storage,
  - searching,
  - sharing,
  - transfer,
  - analysis, and
  - visualization/presentation.


Medical Scans
Social Media
News & Journals
Satellite Images
E-commerce
TV & Video
Sensors & Surveillance
Collect
Collect
Correlate
Analyze/ Predict
Insight
Business intelligence
Biodiversity trends
Customer purchase habits
Epidemic migration paths
Falls prevention for the elderly
Extreme weather prediction

Published: November 6, 2009

The Jobless Rate for People Like You

Not all groups have felt the recession equally.

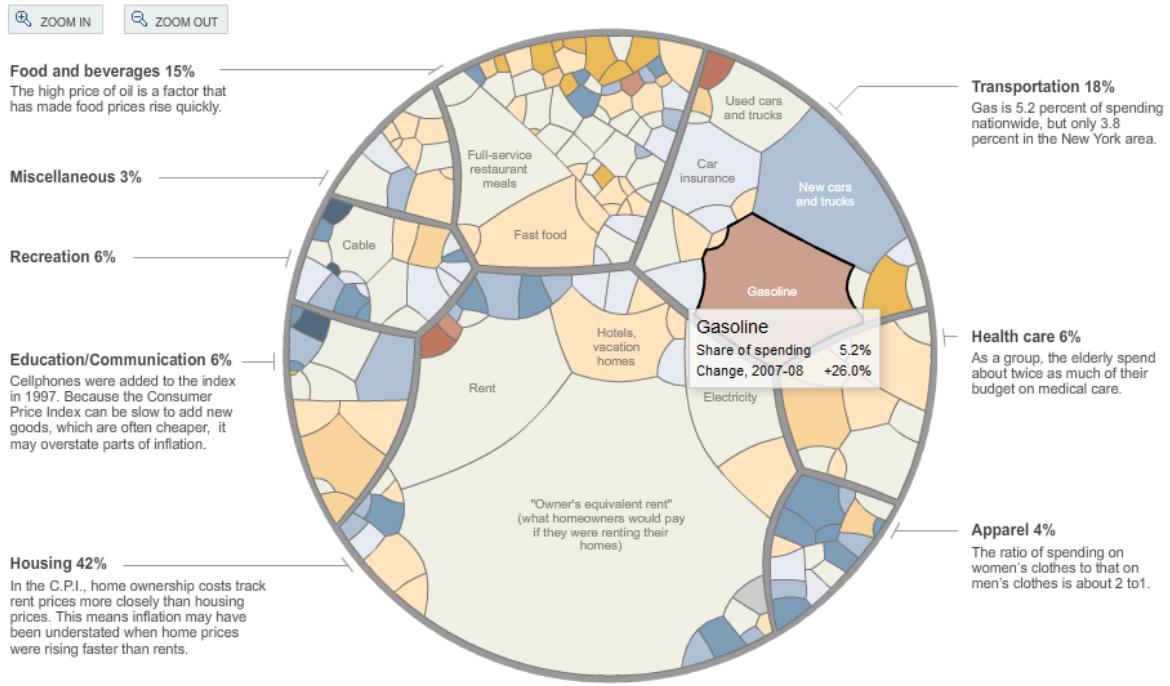


http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html?_r=1&

pusilkom ui
quality results

An Average Consumer's Spending

Each shape below represents how much the average American spends in different categories. Larger shapes make up a larger part of spending.



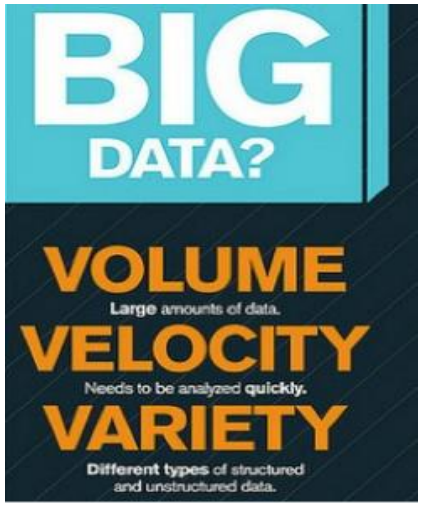
http://www.nytimes.com/interactive/2008/05/03/business/20080403_SPENDING_GRAPHIC.html

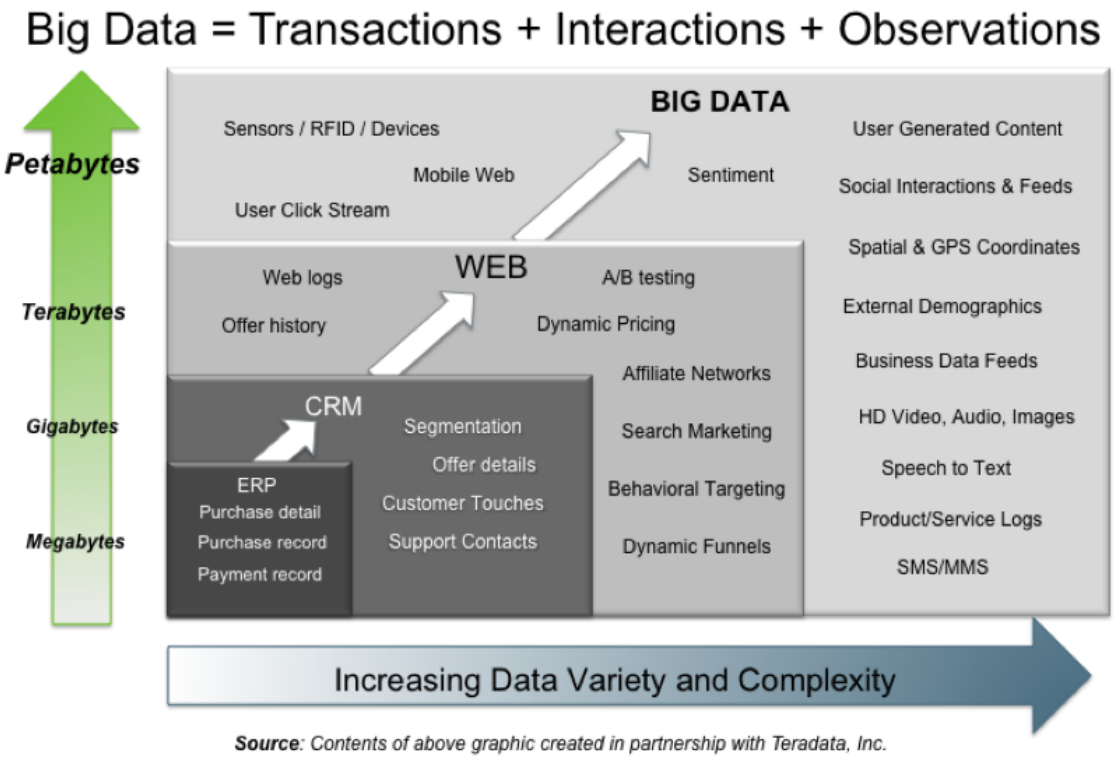
pusilkom ui
quality results

Big Data

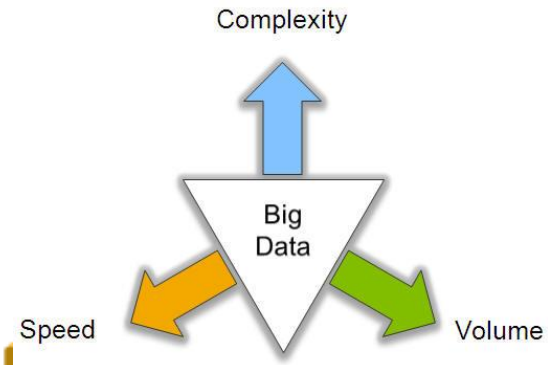
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime, and determine real-time roadway traffic conditions."

Big Data: 3V's





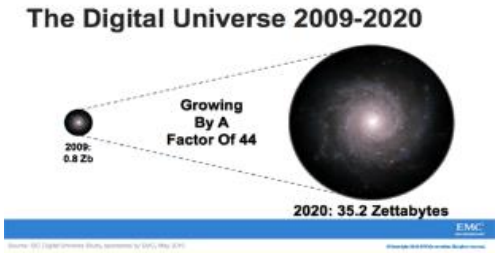
Source: Contents of above graphic created in partnership with Teradata, Inc.

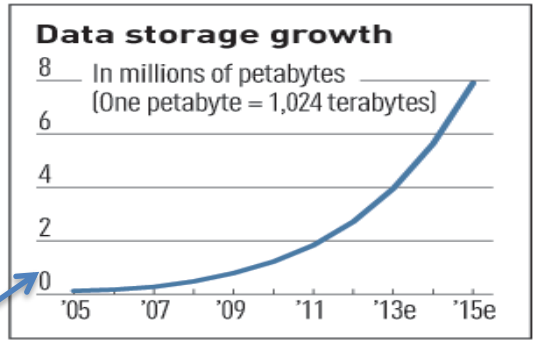


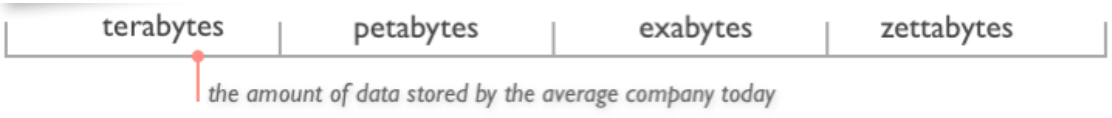
pusilkom ui
quality results

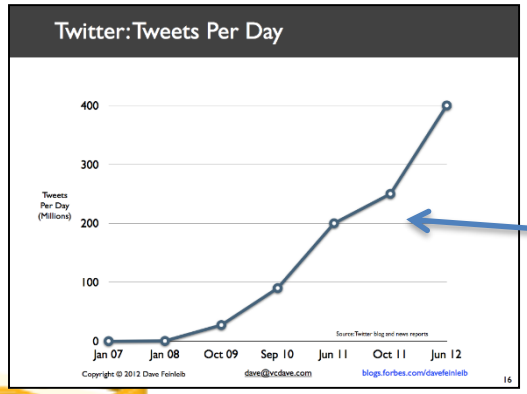
Volume (Scale)

- Data Volume
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





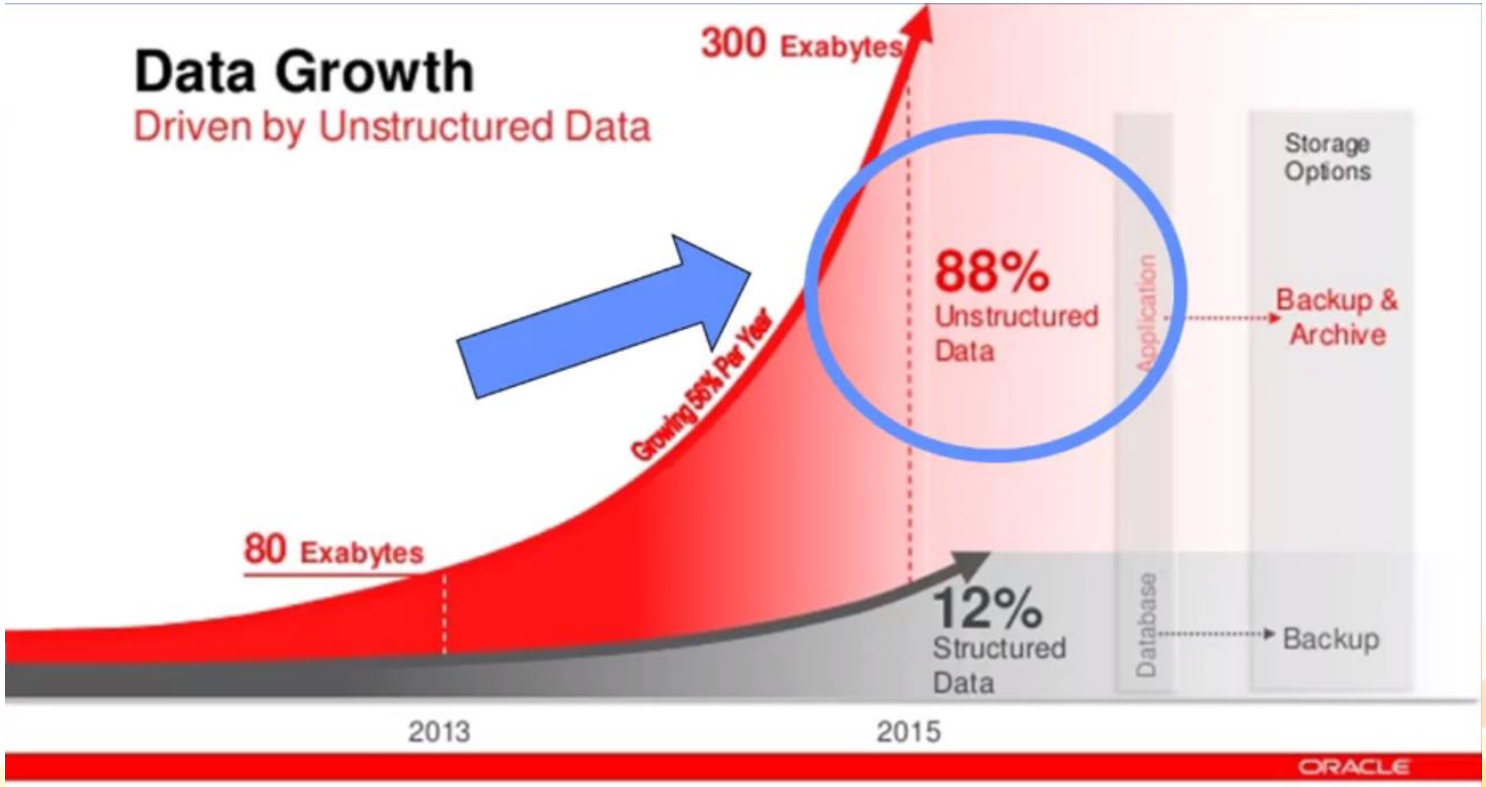




Exponential increase in collected/generated data

Data Growth

- 90% of world's data was created in the last two years



pusilkom ui
quality results

Every minute

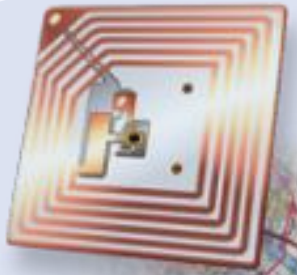
+genius

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs of **log data every day**

30 billion RFID tags today (1.3B in 2005)



76 million smart meters in 2009… 200M by 2014

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

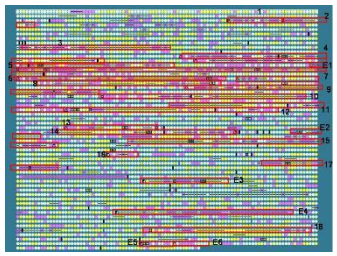
pusilkom ui
quality results

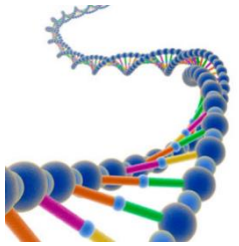
The new trend: **Open Data** -freely available "Big" datasets

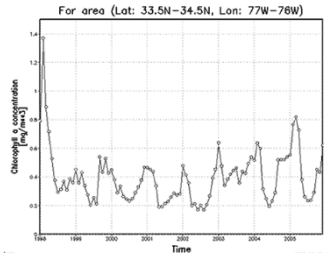
- The World Bank, Data Catalog, http://data.worldbank.org/topic
- The US Federal Govt., Data Catalog, http://www.data.gov/catalog
- The World Wildlife Fund, http://www.worldwildlife.org/science/data/item1872.html
- The Adventure Works Database, http://sqlserversamples.codeplex.com/
- National Climatic Data Center, http://www.ncdc.noaa.gov/oa/ncdc.html
- Queensland Govt. Wildlife & Ecosystems, http://www.derm.qld.gov.au/wildlife-ecosystems/index.html
- Medicare Databases, http://www.medicare.gov/download/downloaddb.asp
- ARFF, WEKA, University of Waikato, http://weka.wikispaces.com/XML
- 20TB data http://webdatacommons.org/
- Internet Census http://internetcensus2012.bitbucket.org/paperhtml
- Youtube, Twitter, Facebook, flicker, etc, ...

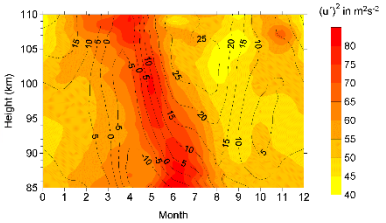
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)









To extract knowledge➔ all these types of data need to linked together

```xml
<?xml version="1.0" standalone="yes"?>
<BankAccount>
	<Number>1234</Number>
	<Type>Checking</Type>
	<OpenDate>11/04/1974</OpenDate>
	<Balance>25382.20</Balance>
	<AccountHolder>
		<LastName>Singh</LastName>
		<FirstName>Darshan</FirstName>
	</AccountHolder>
</BankAccount>
```

```xml
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="http://www.w3schools.com/rdf/">

<rdf:Description rdf:about="http://www.w3schools.com">
  <si:title>W3Schools</si:title>
  <si:author>Jan Egil Refsnes</si:author>
</rdf:Description>

</rdf:RDF>
```

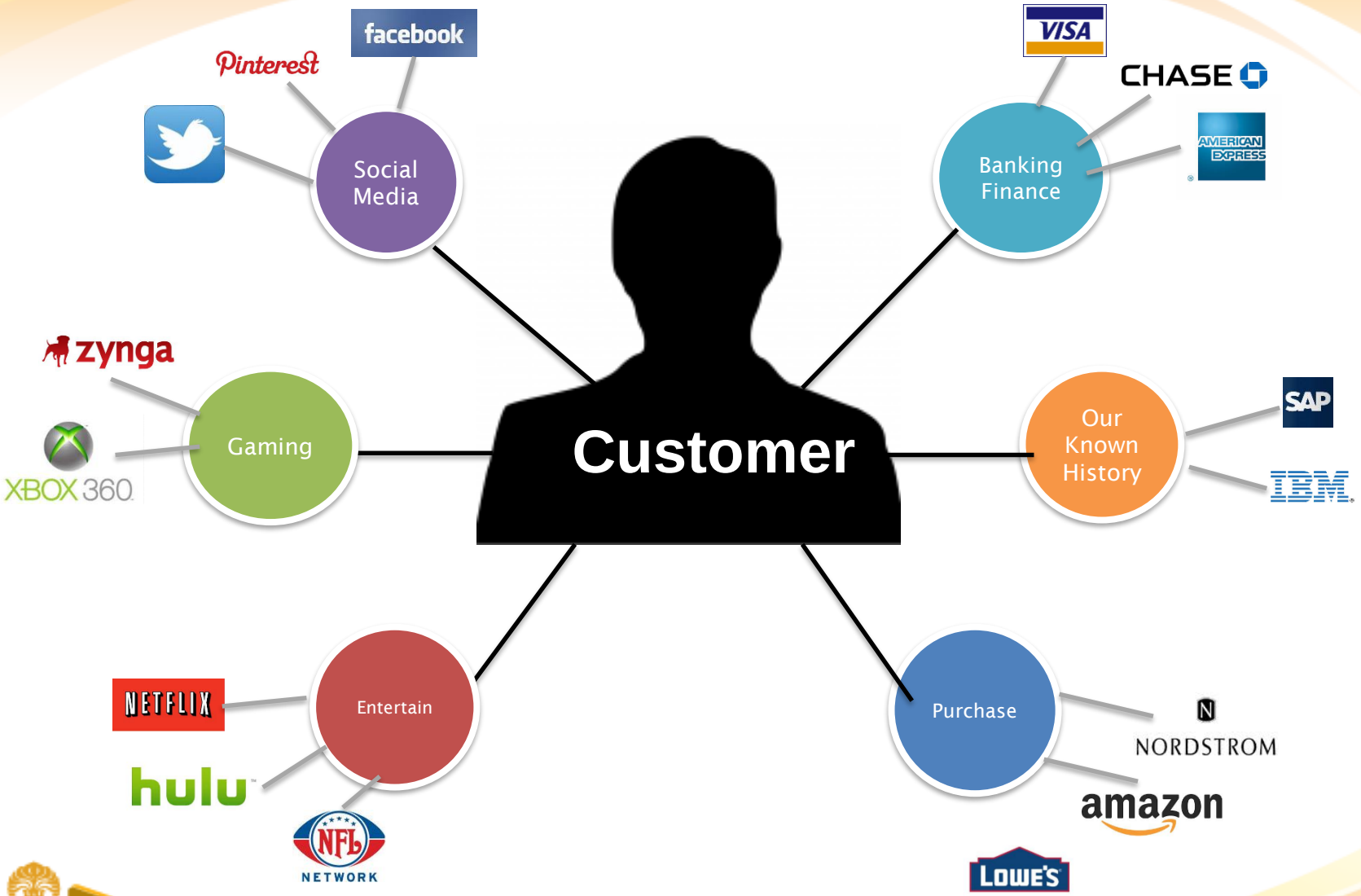
```html
<html>

<head>
    <title>My First Web Page</title>
</head>

<body>
    <h1>My First Web Page</h1>
    <p><b>Hello World Wide Web!</b></p>
    <p><i>Hello World Wide Web!</i></p>
    <p><u>Hello World Wide Web!</u></p>
    <p>This is my first web page.</p>
    <p>HTML tags can give <b><i>various</i></b>
    <u>looks and format</u> to the content of this web page.</p>
</body>

</html>
```

A Single View to the Customer



pusilkom ui
quality results

Velocity (Speed)

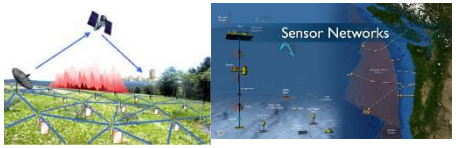
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - E-Promotions: Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - Healthcare monitoring: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

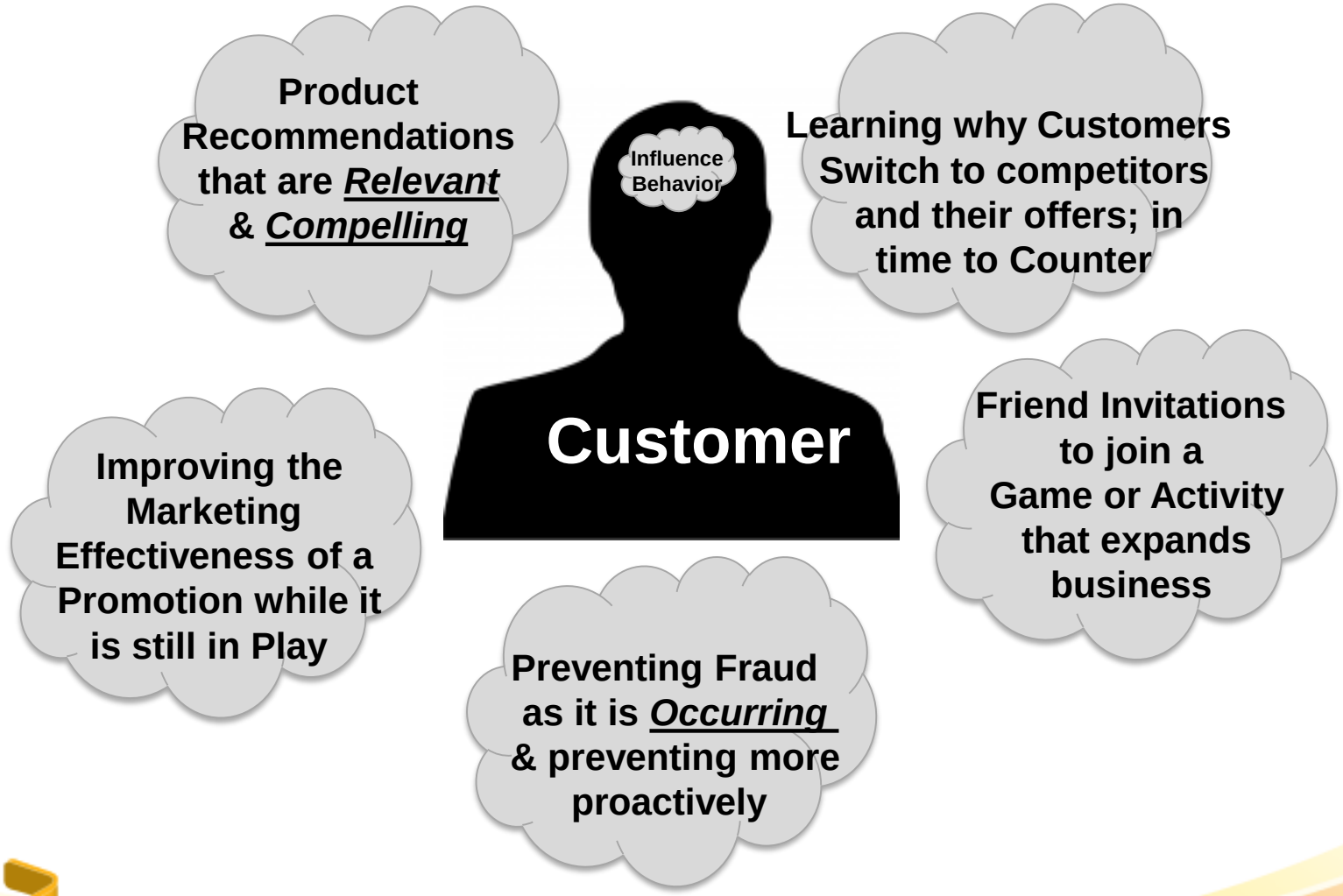


Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

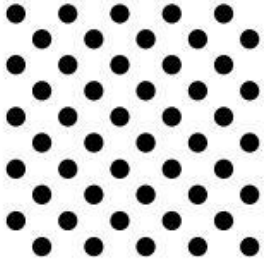
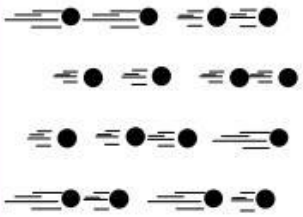
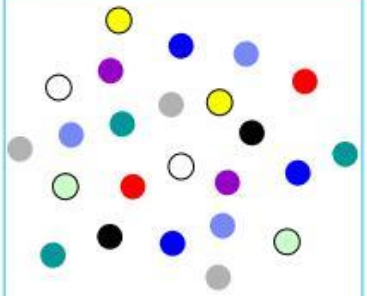
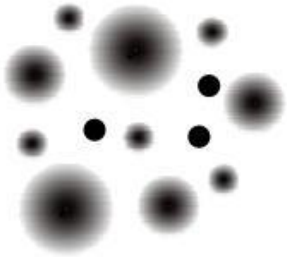
pusilkom ui
quality results

Real-Time Analytics/Decision Requirement



pusilkom ui
quality results

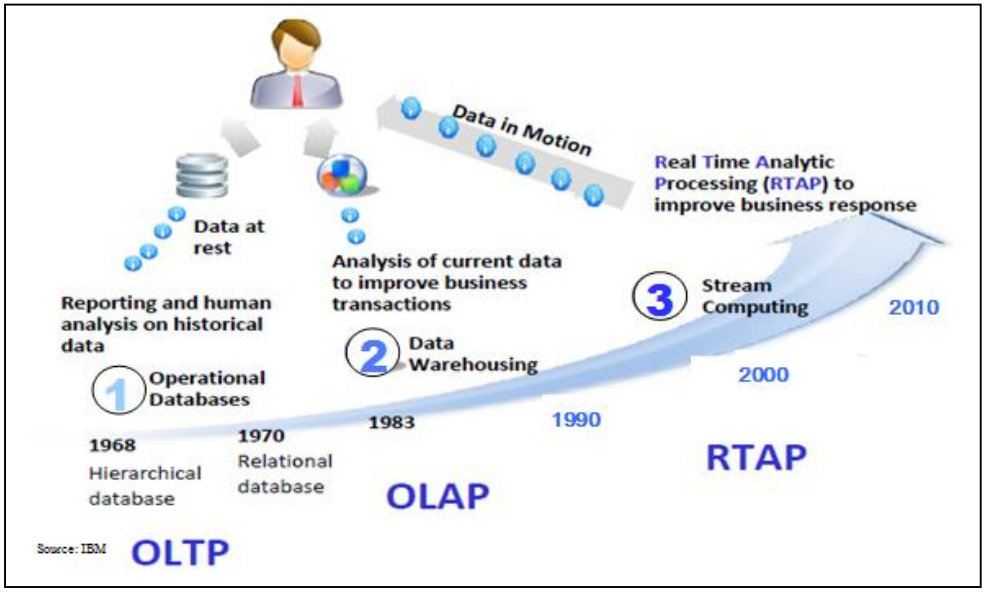
Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Discussion

- How big is the data volume at your workplace?
- What are the data sources?
  - The velocity? How much data generated per day?

Harnessing Big Data



- OLTP: Online Transaction Processing (DBMSs)
- OLAP: Online Analytical Processing (Data Warehousing)
- RTAP: Real-Time Analytics Processing (Big Data Architecture & technology)

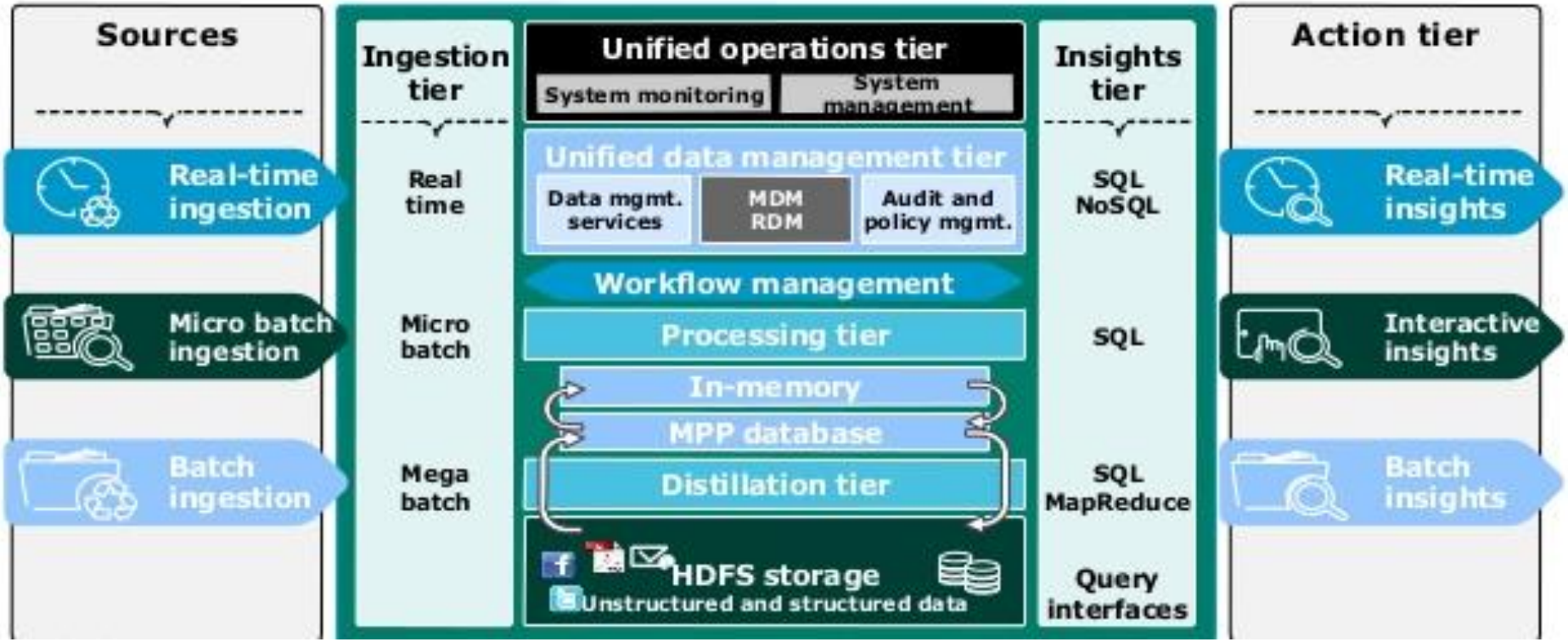
The Model Has Changed…

- The Model of Generating/Consuming Data has Changed

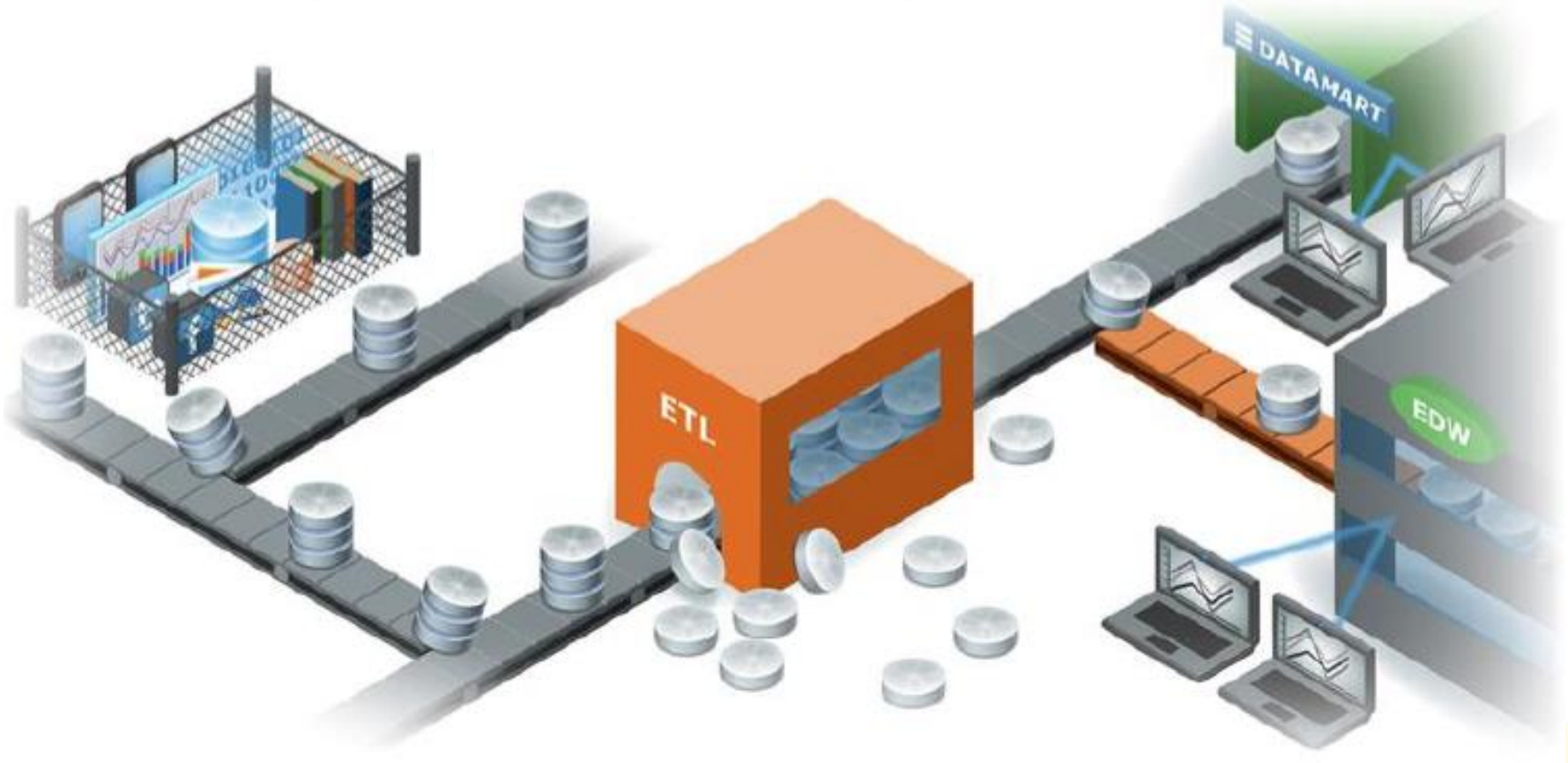
Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Big Data Architecture



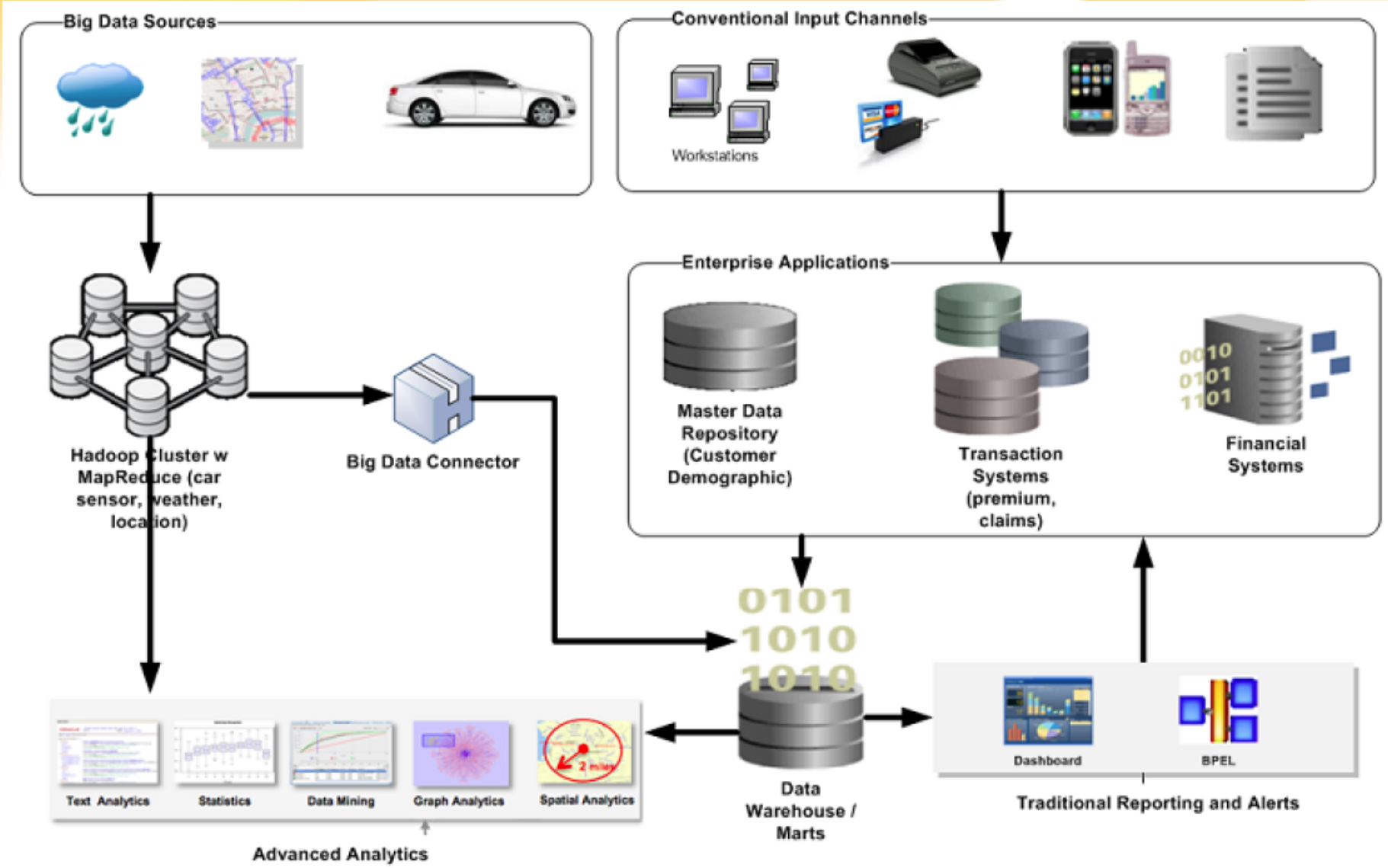
TRADITIONAL ENTERPRISE ANALYTICS PROCESS



pusilkom ui
quality results

EVOLUTION OF PROCESS WITH HADOOP

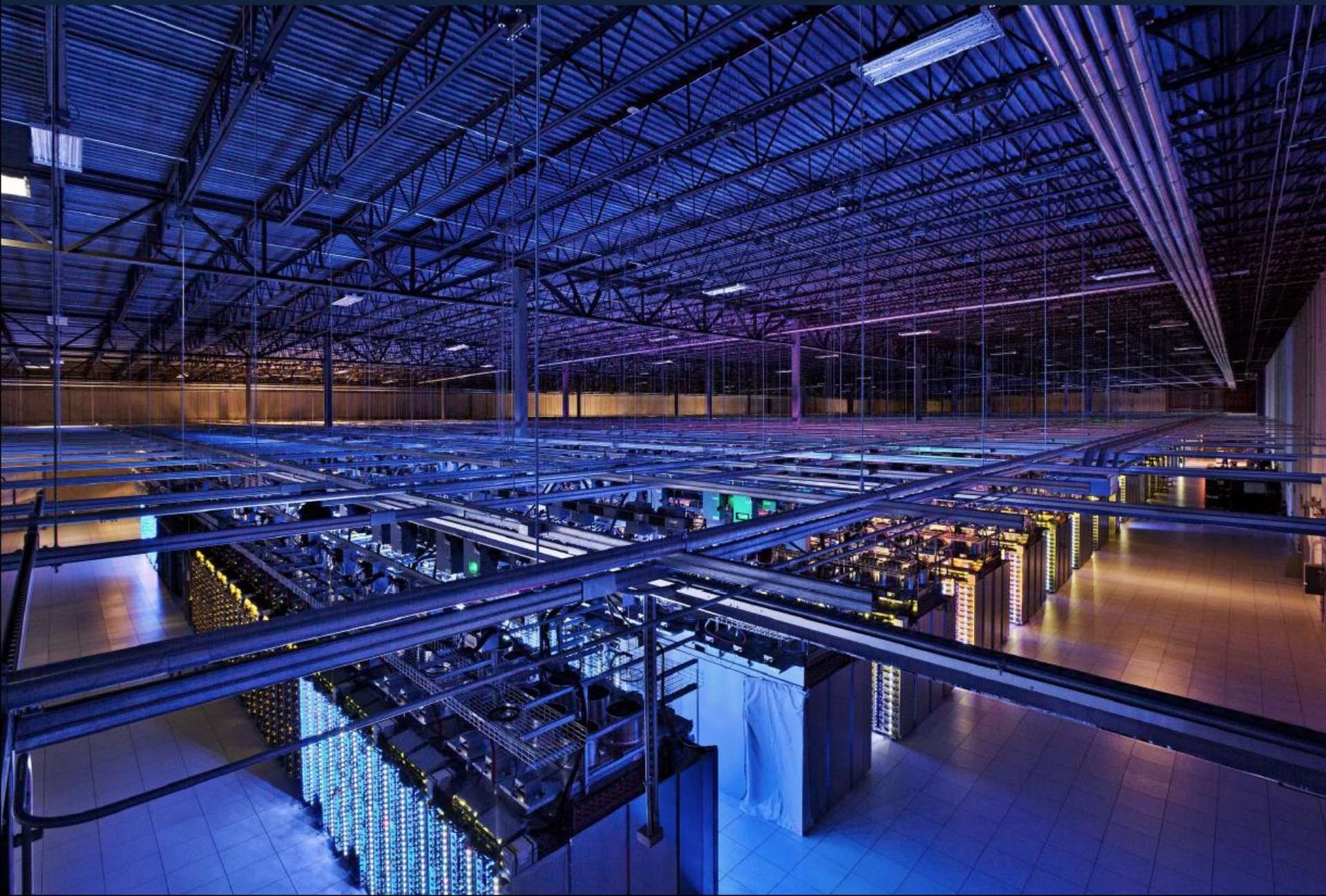
pusilkom ui
quality results



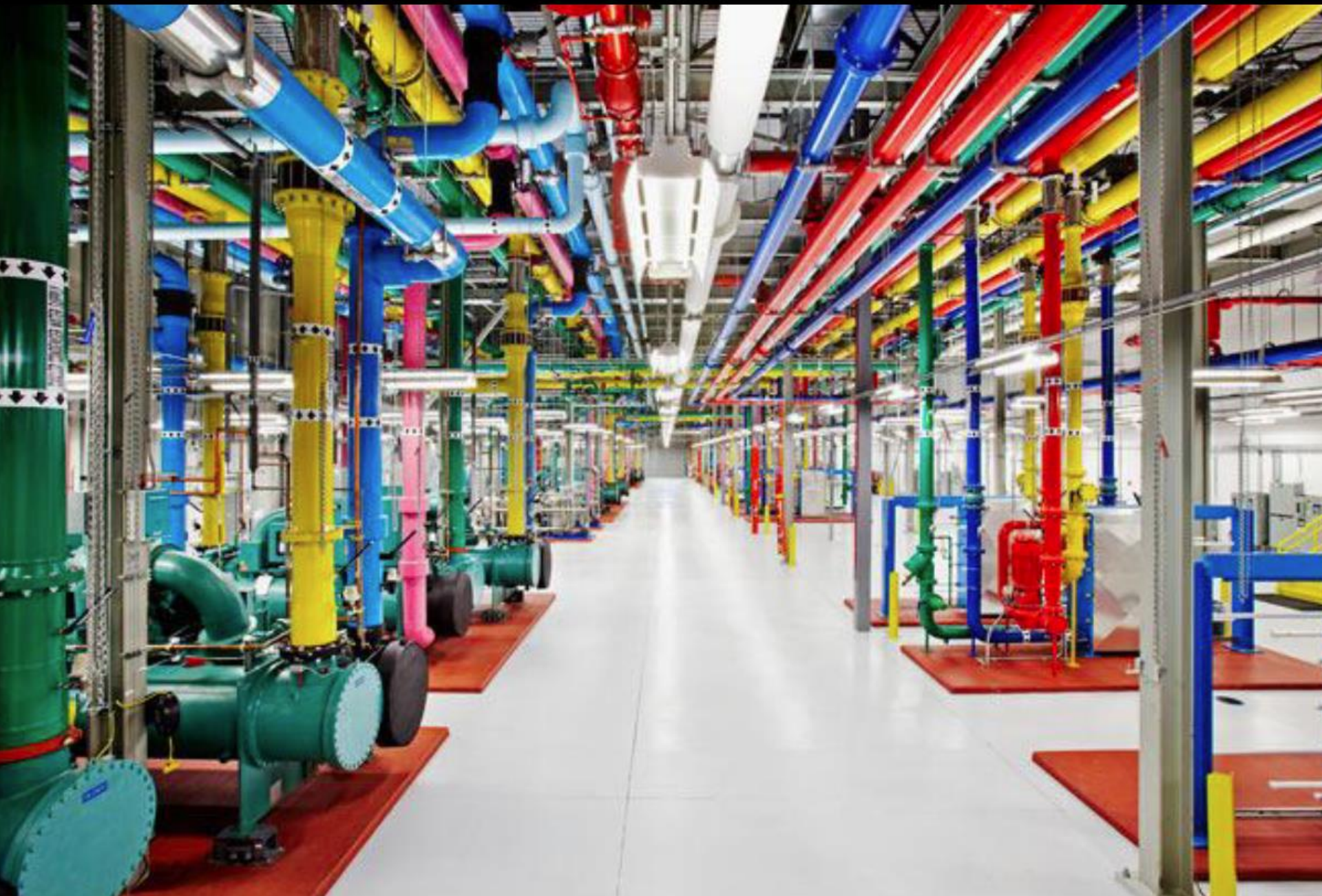
http://www.oracle.com/au/technologies/big-data/index.html

pusilkom ui
quality results

Google Data Centre

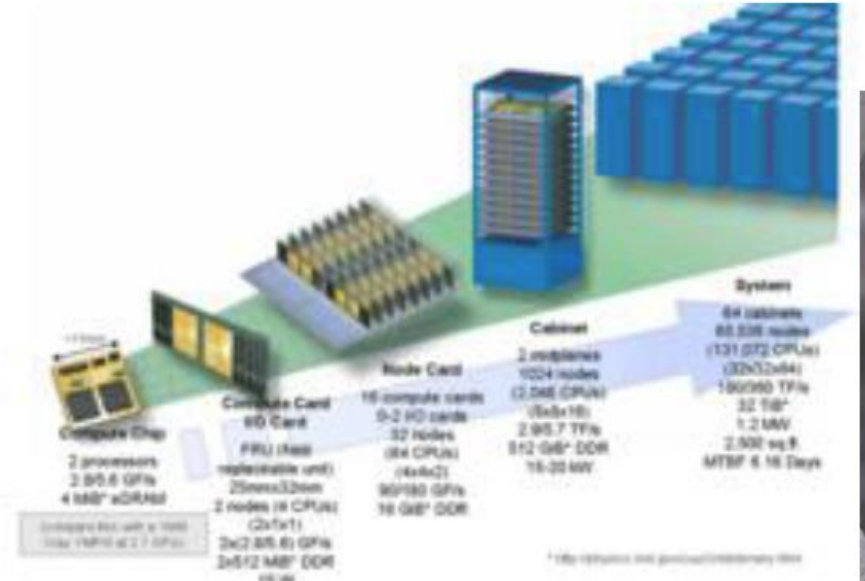


Google Data Centre Cooling Plant



Cluster Machines

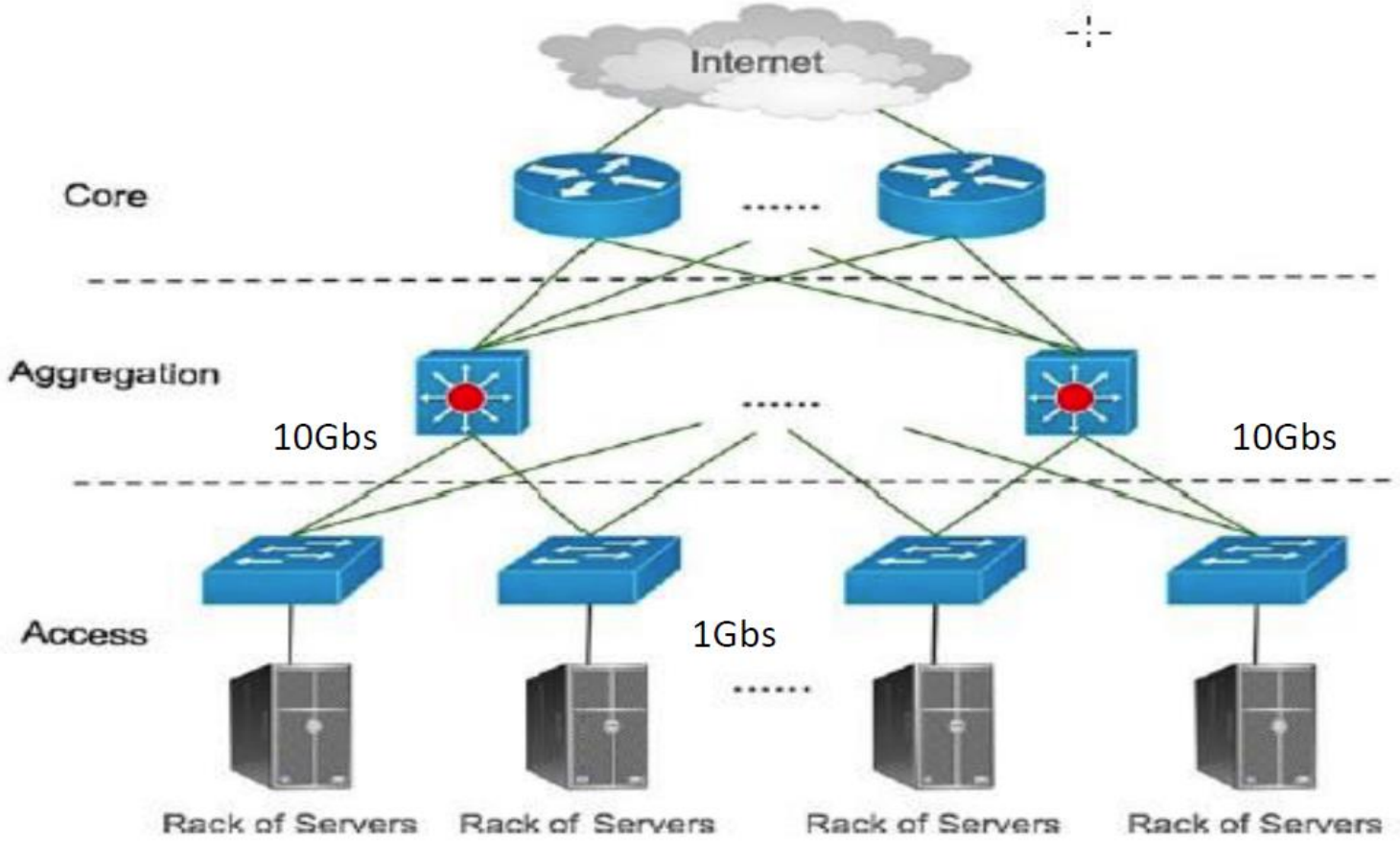
- A set of loosely or tightly connected computers working together as a single system.
- Usually connected via LAN, with each *node* running its own OS.
- Deployment, ranging from small number of nodes to the fastest supercomputers.



Basic connectivity of cloud computing

pusilkom ui
quality results

Basic connectivity of cloud computing



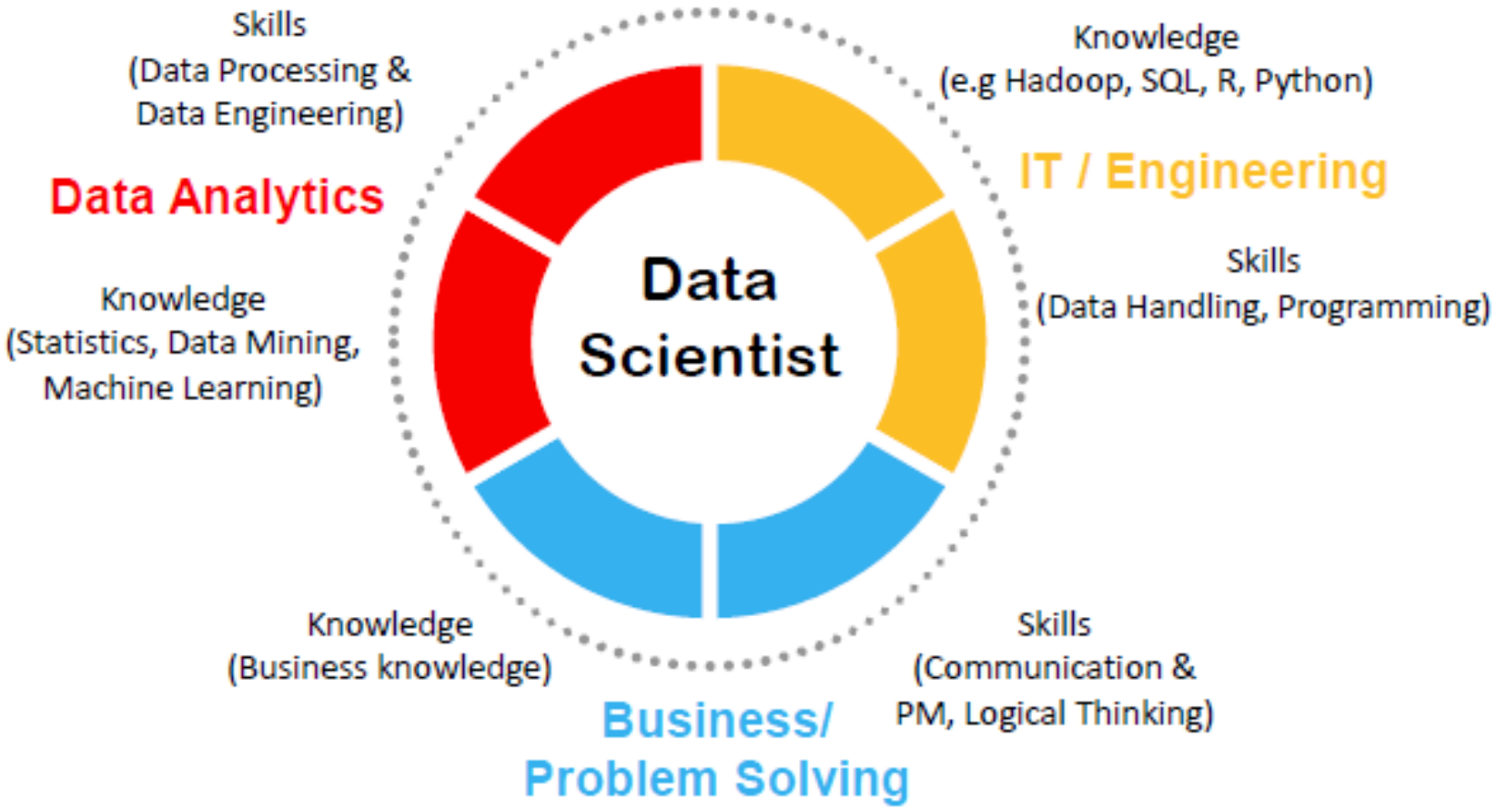
Big data analytics should start from business problem rather than technology.

Don’t start without understanding the value!

What is my ROI?

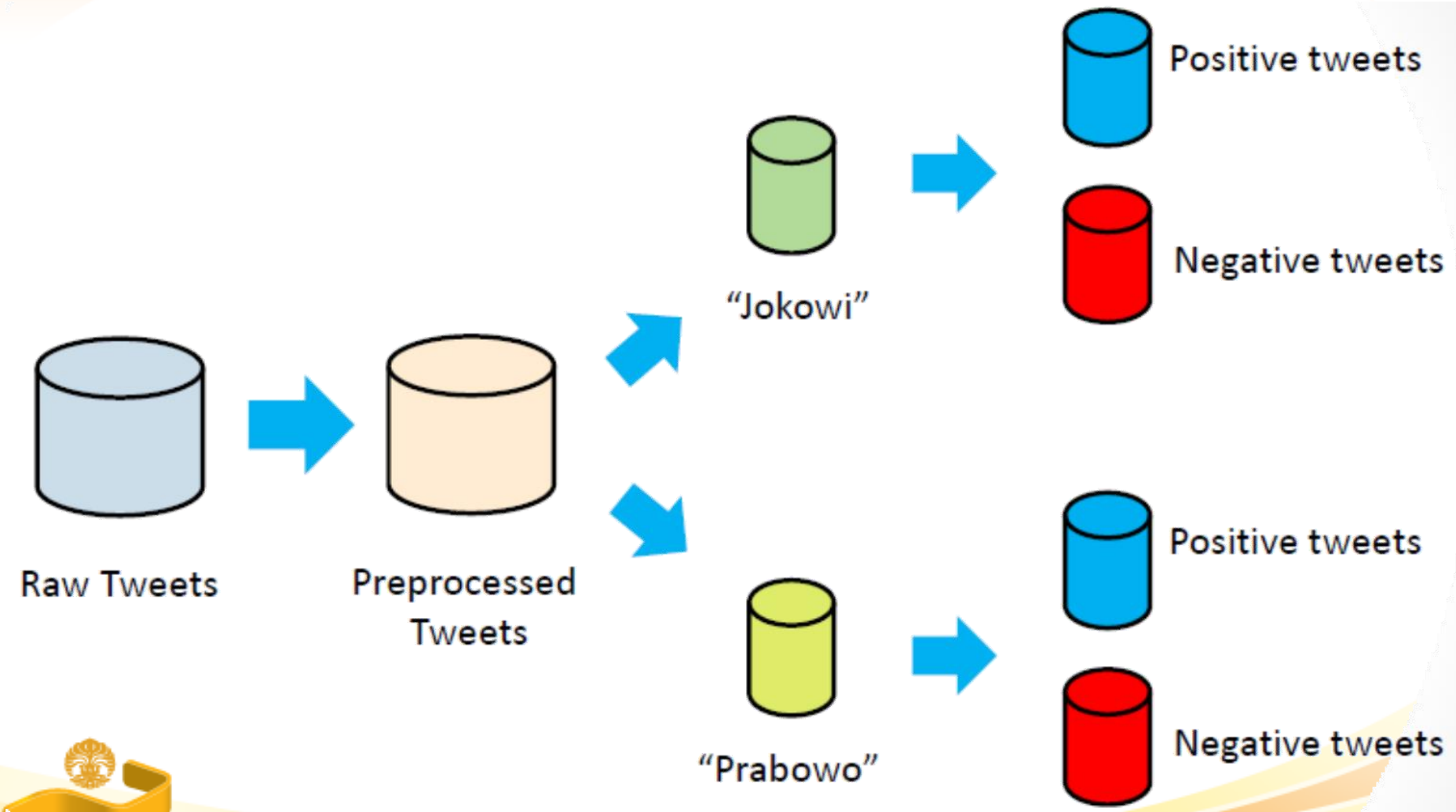
pusilkom ui
quality results

Data Scientist

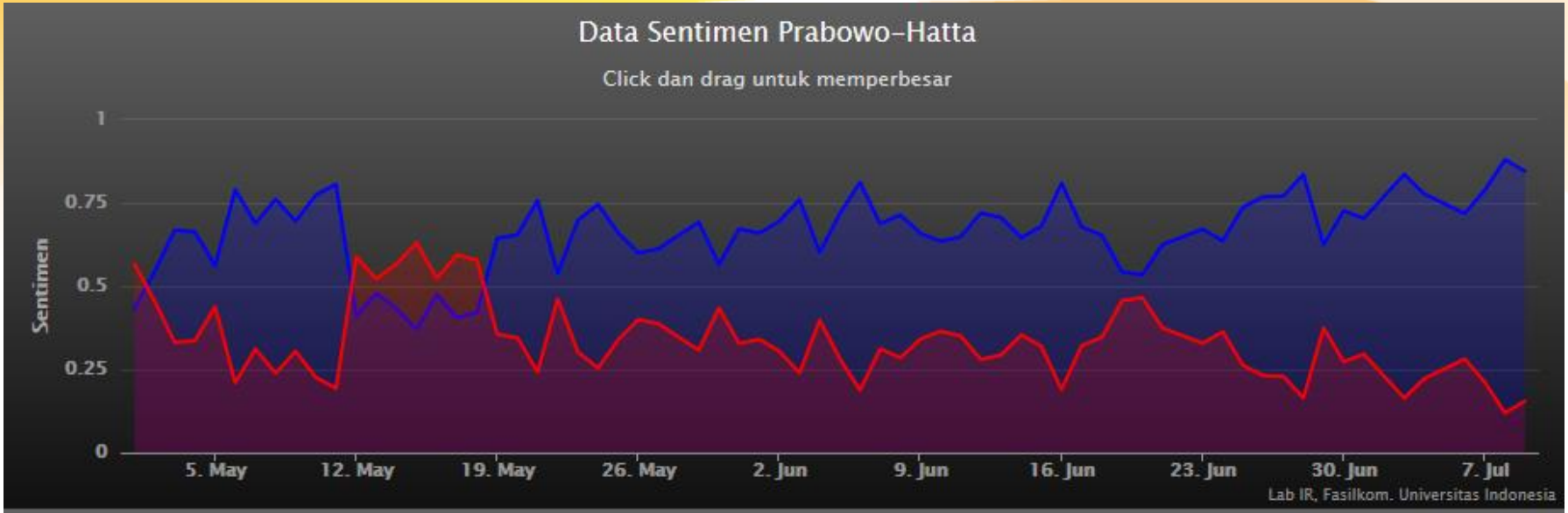


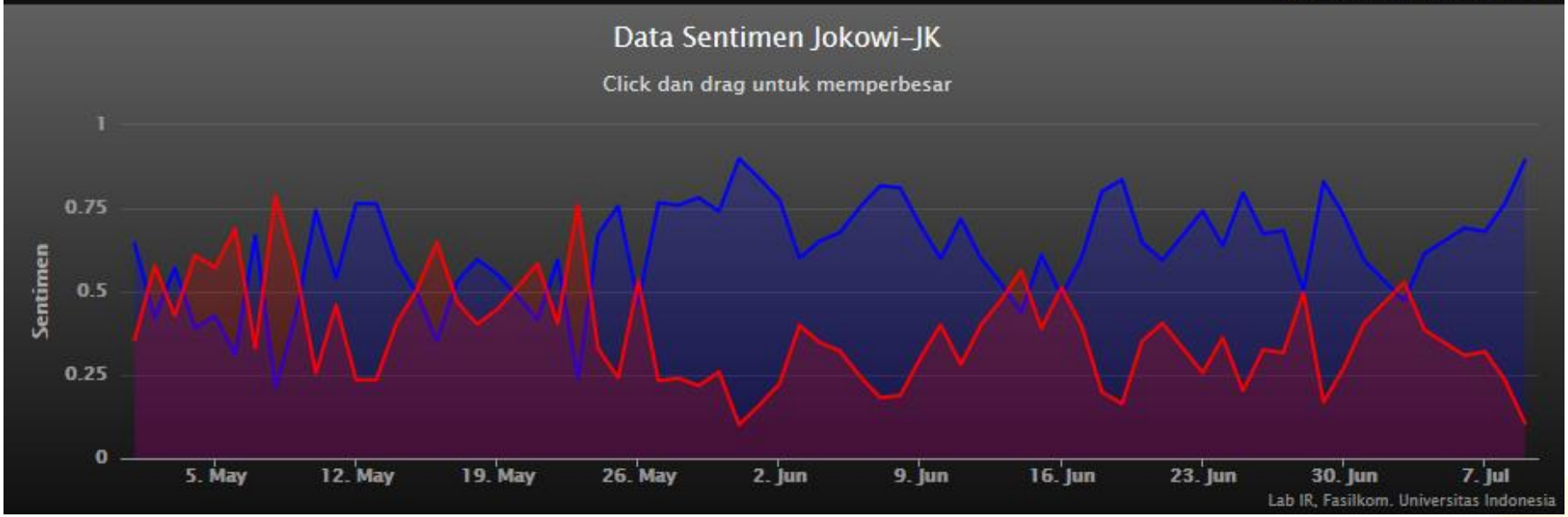
pusilkom ui
quality results

Case Study: Sentiment Analysis

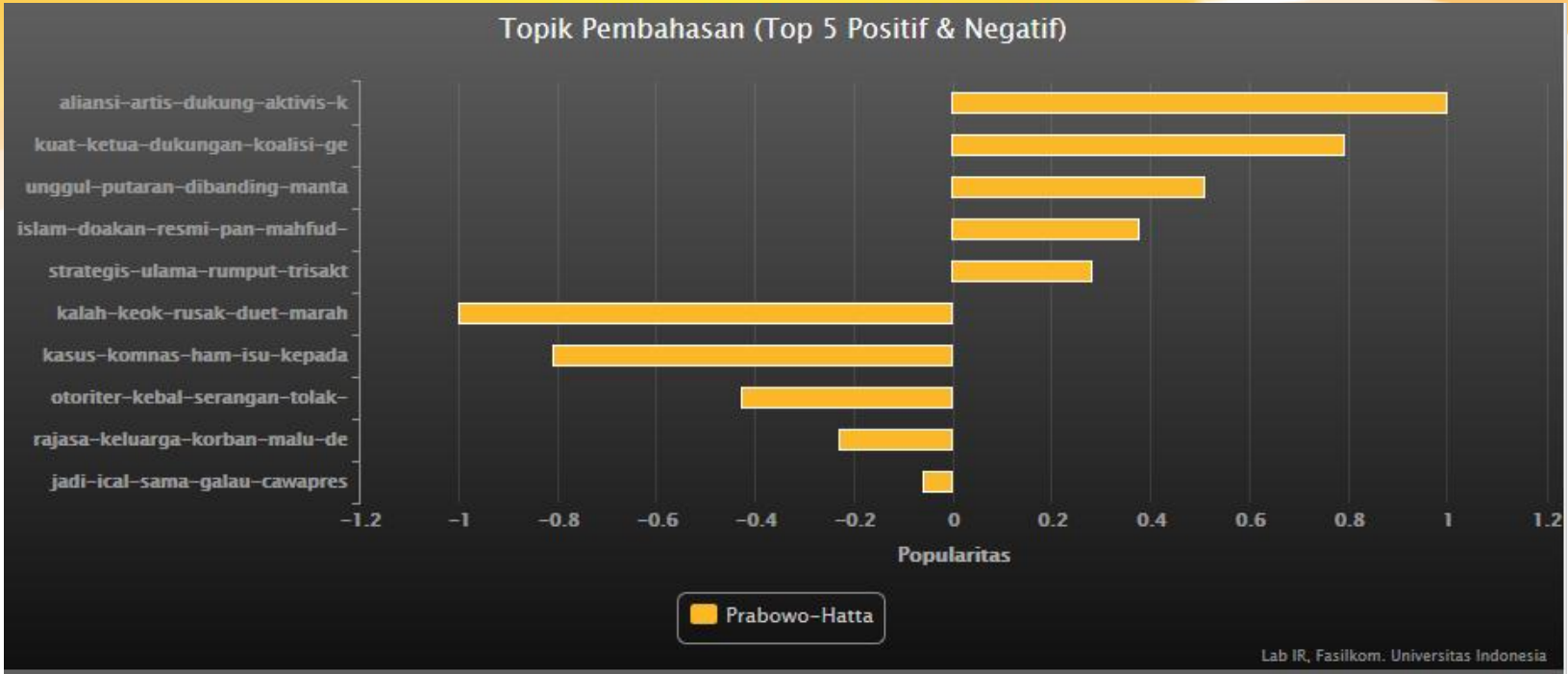


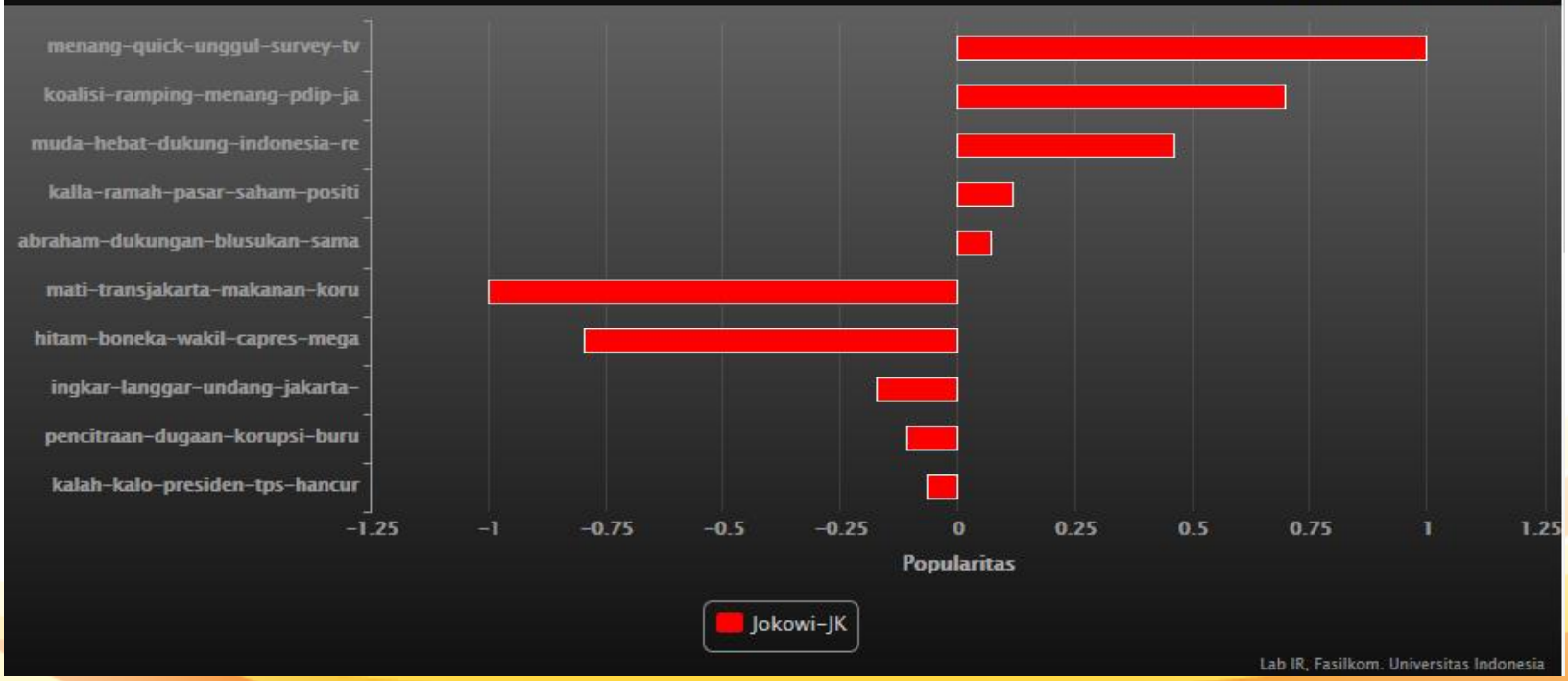
pusilkom ui
quality results





pusilkom ui
quality results





pusilkom ui
quality results

Topics Covered

- Data Pre-processing method
- Introduction to Big Data Infrastructure using Hadoop
- Classification and Prediction
- Clustering & Segmentation
- Deep Learning for Big Data
- Big Data Analytics
- Advanced Big Data Algorithms

pusilkom ui
quality results

DISKUSI DAN TANYA JAWAB

Informasi

DIVISI TRAINING
PUSAT ILMU KOMPUTER
UNIVERSITAS INDONESIA

Jl. Salemba Raya No.4
Jakarta Pusat 10430
Telp +62 21 3106014
Fax + 62 21 3102774

Kampus Baru UI Depok
Depok 16424
Telp +62 21 7863419 ext.3210
Fax + 62 21 7863415

http://training-pusilkom.cs.ui.ac.id

Contact person
Sri Mutia R
s.mutia@cs.ui.ac.id