CPE 213 Data Models
(a.k.a. Data Modeling and Visualization)

Lecture 1: Introduction to Data Modeling and Visualization

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

BIG DATA EXPERIENCE

BIG DATA EXPERIENCE CENTER

KMUTT g-able

Course Learning Outcome

- Evaluate and apply suitable data modeling techniques to analyze real-world data.
- Create meaningful visualization that address the relevant problems.
- Understand the data science process and the role of data scientists.

BIG DATA EXPERIENCE

Reference Textbook

- Grolemund, G. and Wickham, H. (2017). R for Data Science. O'Reilly Media. Link: http://r4ds.had.co.nz

- Slides, papers, and additional documents will be provided in-class and onlin

KMUTT g-able

BIG DATA EXPERIENCE

Grading

- Midterm Examination 30%
- Final Examination 30%
- Project ~~20%~~ 15%
- Lab / HW (10 LABS) ~~20%~~ 10%
- Quest 15%

BIG DATA EXPERIENCE

#	Date	Topic
1	22 Jan 2021	Introduction to data modeling and visualization
2	29 Jan 2021	R programming
3	5 Feb 2021	Basic types of data visualizations
4	19 Feb 2021	Data visualization 1 - distributions
5	5 Mar 2021	Data visualization 2 - relationship
8		=== Midterm Examination ===
9	19 Mar 2021	Data visualization 3 - network
10	26 Mar 2021	R database operations
11	2 Apr 2021	Modeling statistical distribution
12	9 April 2021	Linear regression
13	23 April 2020	Logistic regression
14	30 April 2020	Decision tree
15	7 May 2020	Term project presentation
16		=== Final Examination ===

BIG DATA EXPERIENCE

Project

- Students work in teams of 5
- Task: Visualization
  - Acquire a dataset and understand it
  - Create a visualization of a dataset
  - Address related questions
- Task: Modeling
  - Identify the modeling objective
  - Build a model with the prepared data
  - Make predictions based on the model
- Output: Write a technical report and present the visualization and model to the class
- Due: 7 May 2021

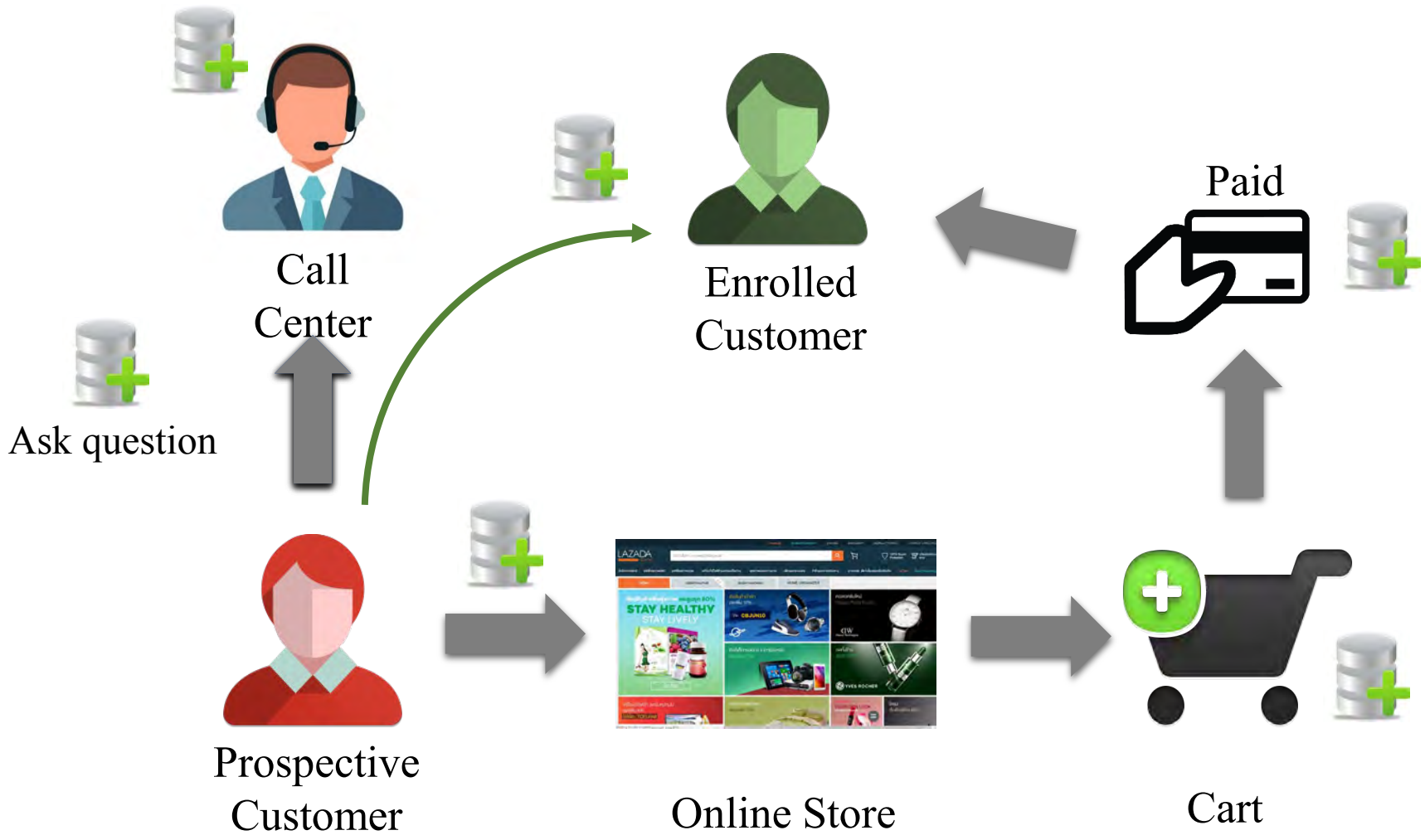
KMUTT g-able

BIG DATA EXPERIENCE

What are Data Modeling and Visualization?

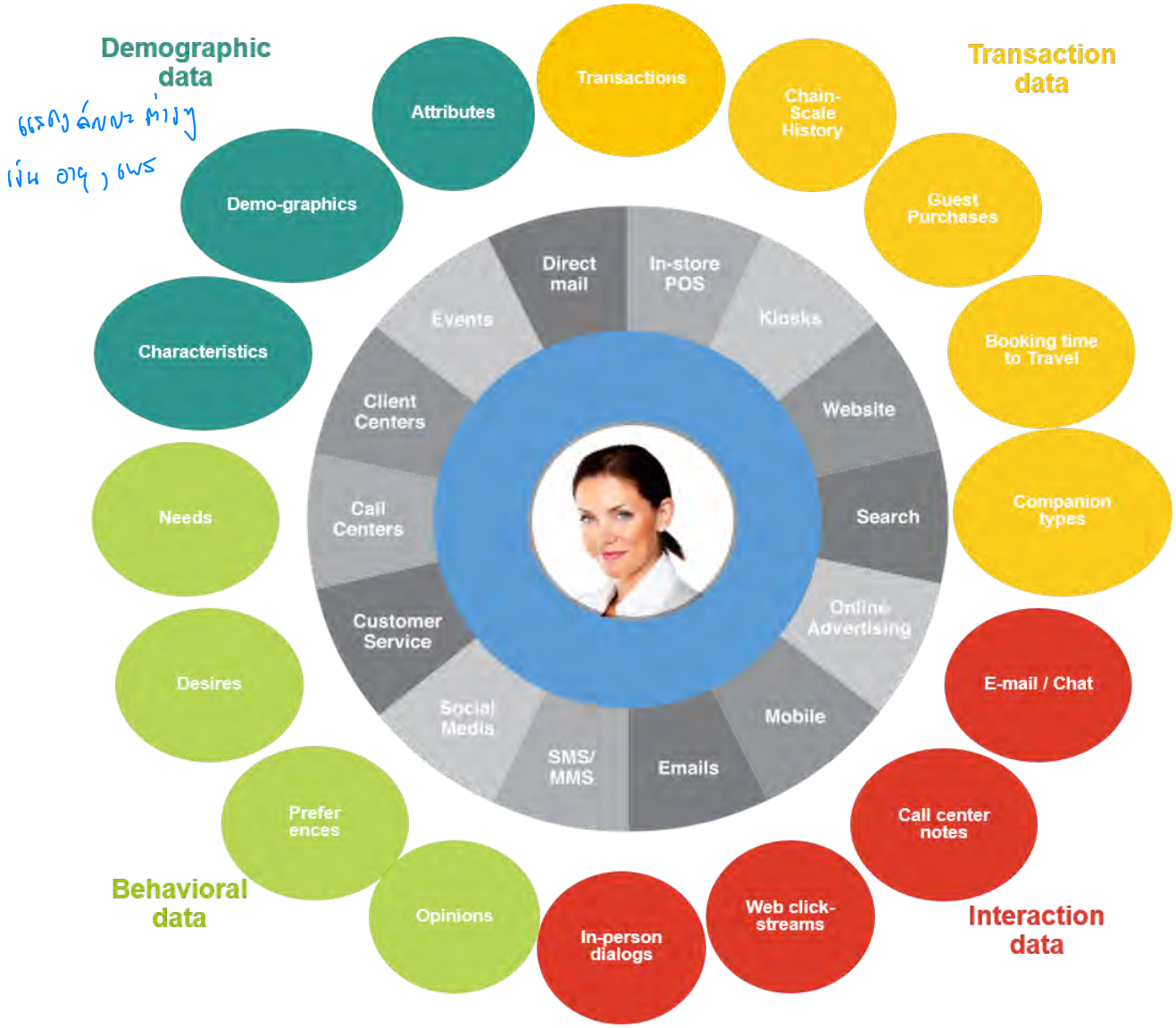
Section

KMUTT g-able


Call
Center
Enrolled
Customer
Paid
Ask question
Prospective
Customer
Online Store
Cart

BIG DATA EXPERIENCE

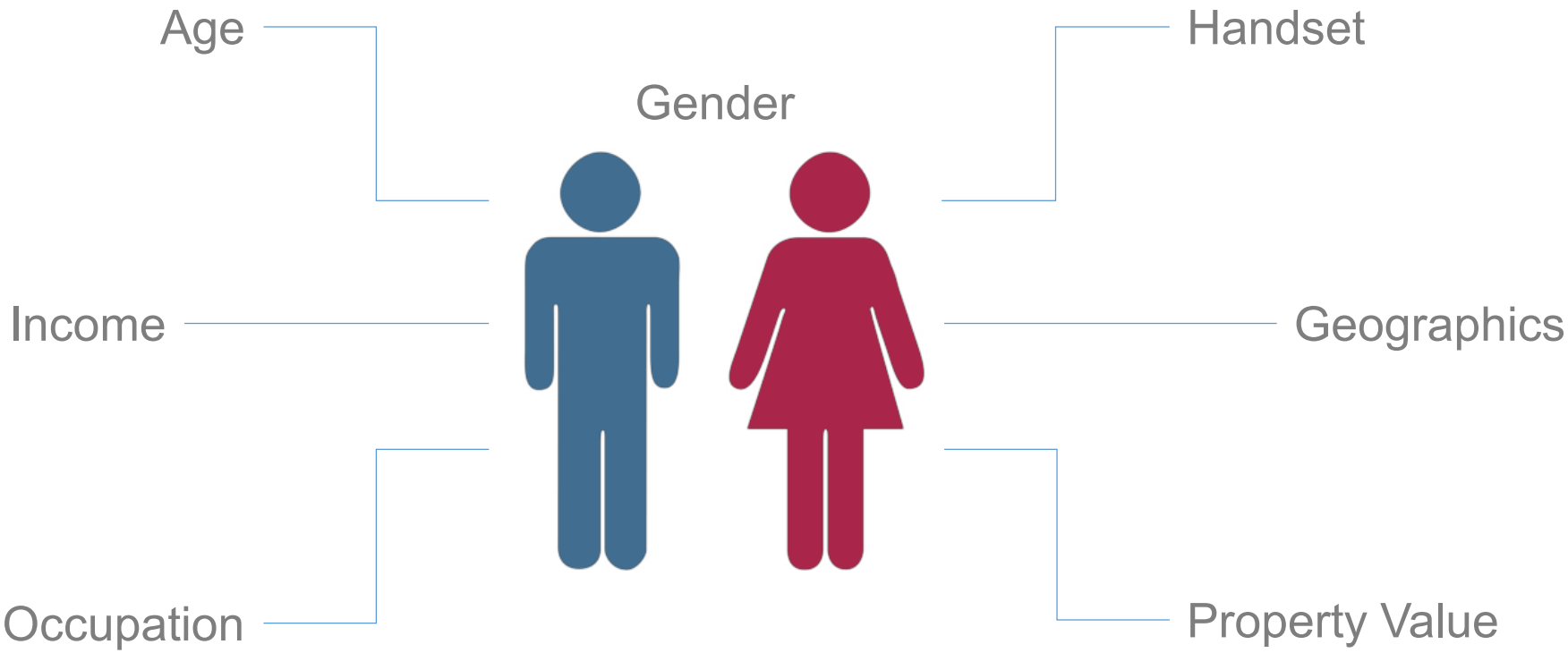
CUSTOMER JOURNEY



KMUTT g-able

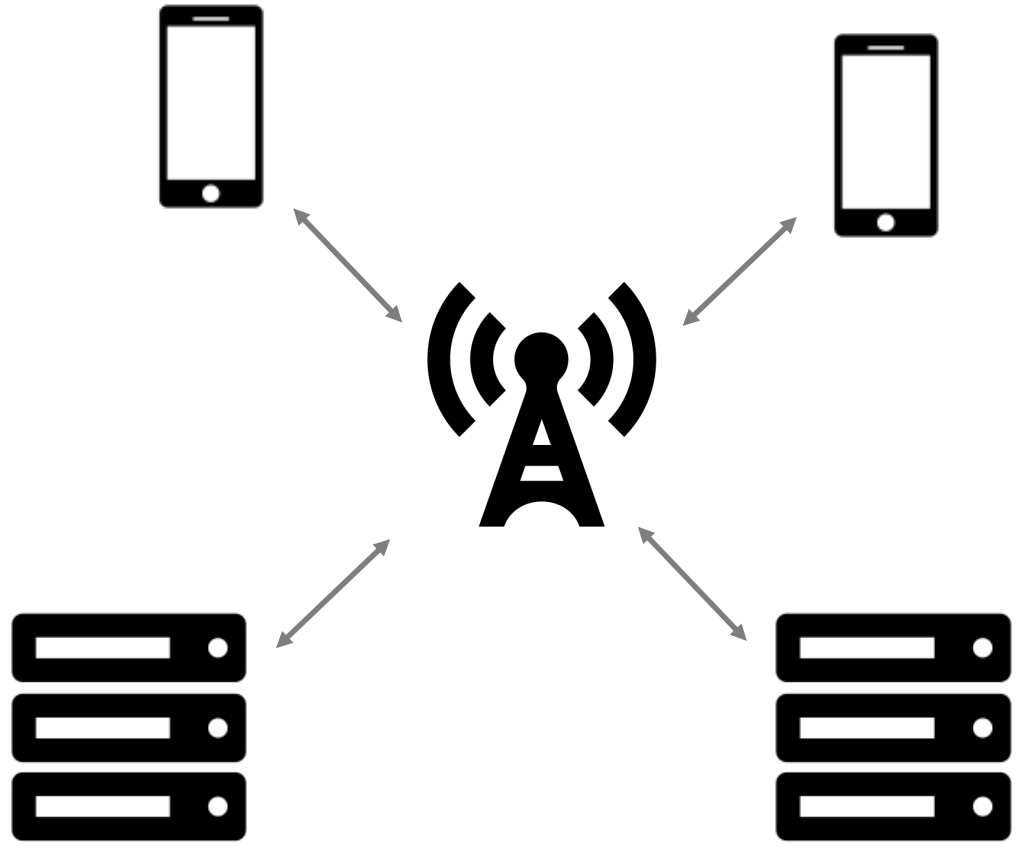
BIG DATA EXPERIENCE

Demographic data



Transactional data

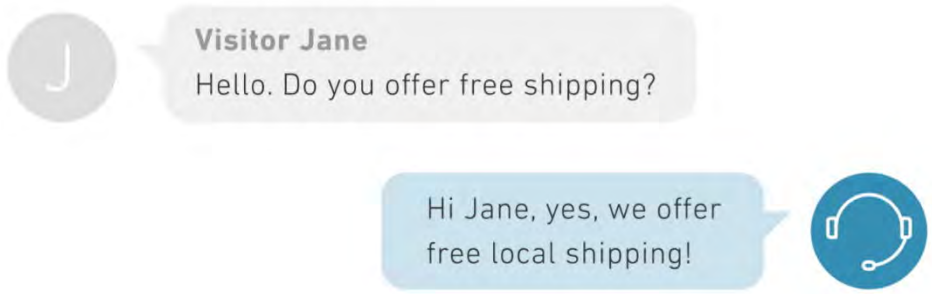
- Call detail record
- Internet usage
- Tower location
- CRM data
- Service plan
- Payment information
- App usage
- Call center



BIG DATA EXPERIENCE

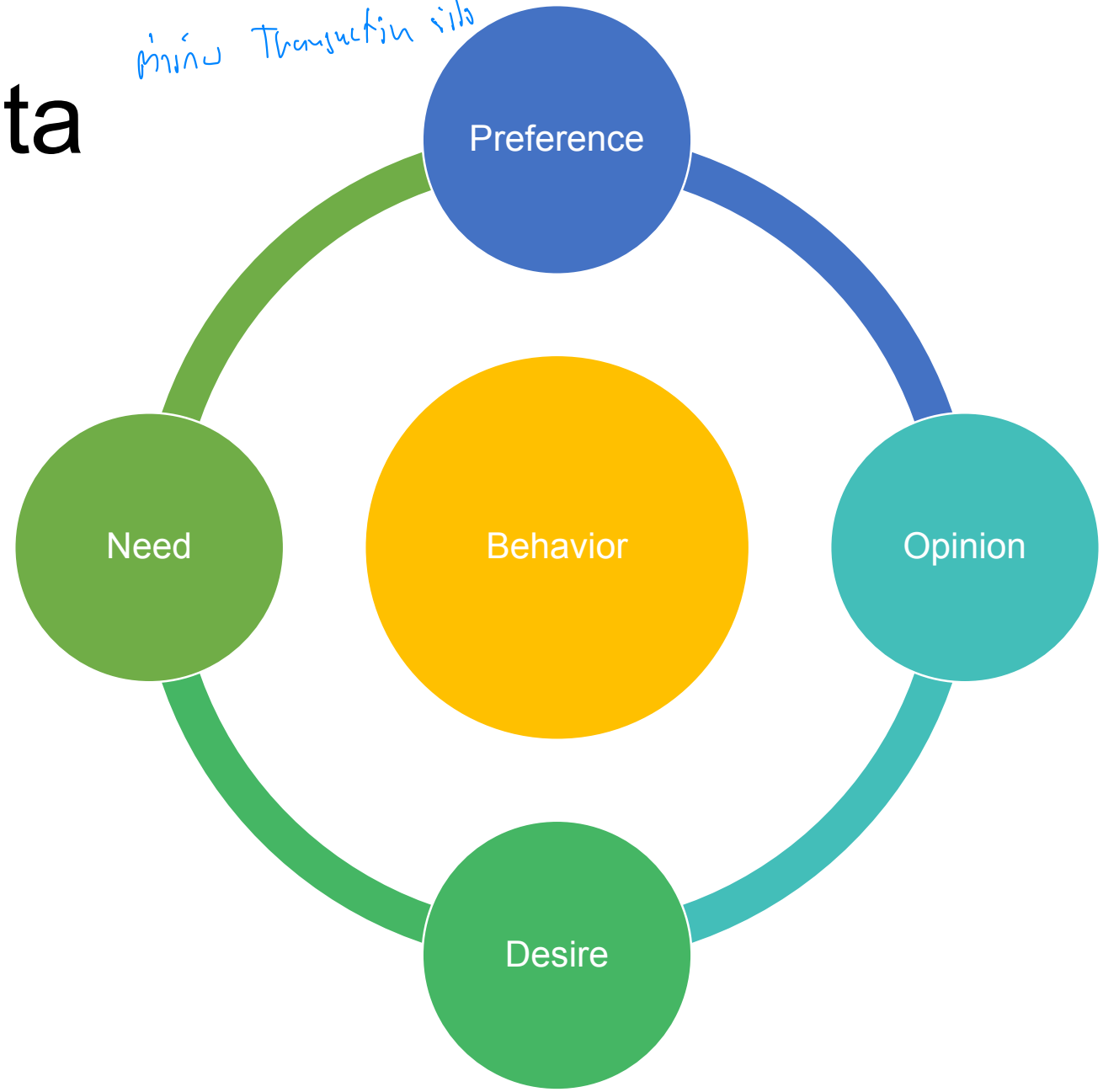
Interaction data

- E-mail log
- Chat log
- Web click stream
- App click stream
- In-person log



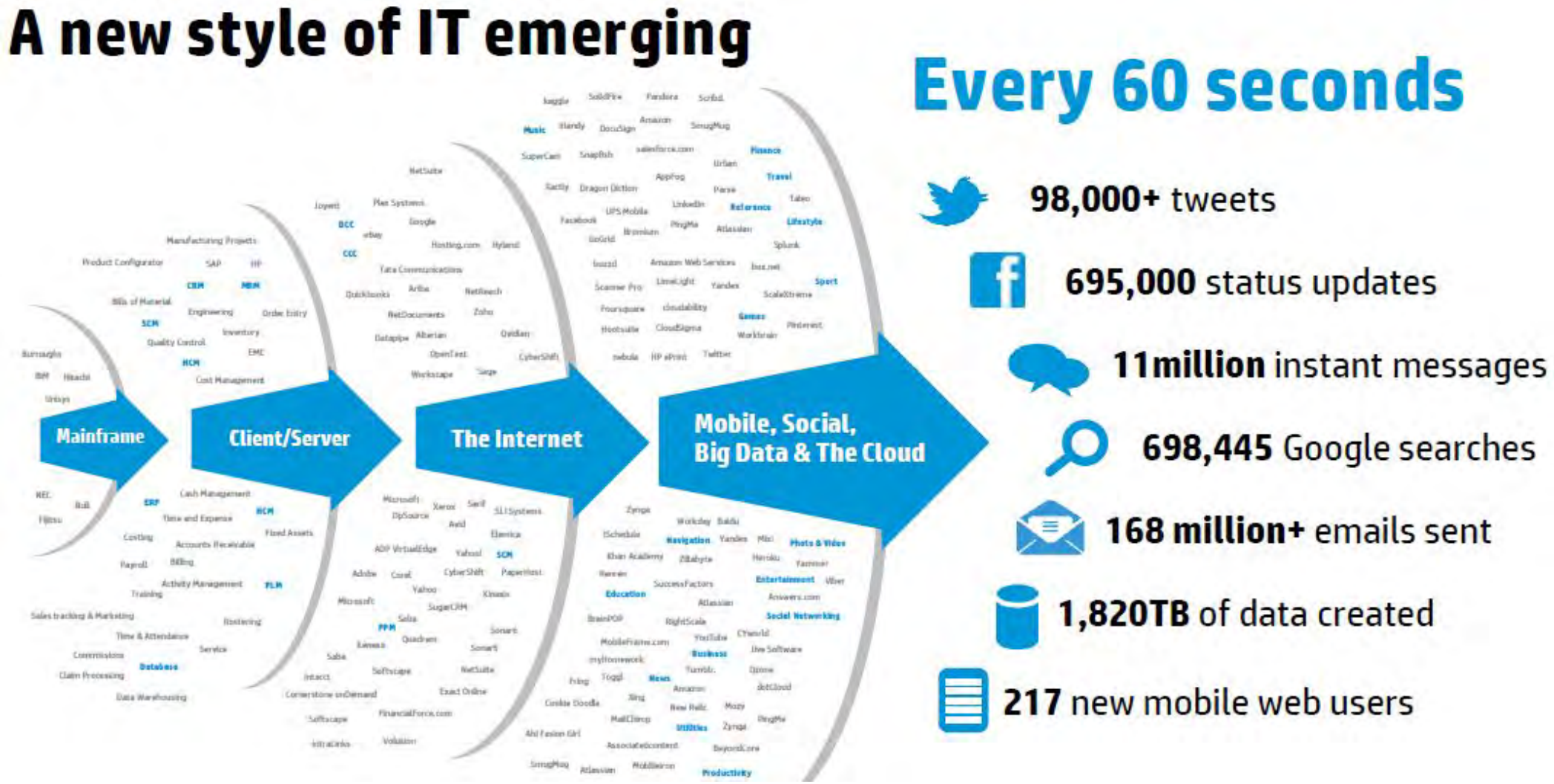
Behavioral data

- Opinion
  - Like matrix
- Preference
  - Collaborative filtering
- Desire
  - Wish list
  - Browsing history
- Need
  - Tax/Law/Regulation
  - Life
  - Finance




Social Media
Mobile
Ads
Web Logs
CRM
Product Logs
Transaction
Call Center
Are keywords related to customer segments?
Which campaign combinations accelerate conversion?
Which product features drive adoption?
What content works be best for each lead segment?
Which features do users struggle with?
What behavior signals churn?

The rise of data …



"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data to be an economic asset,** like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

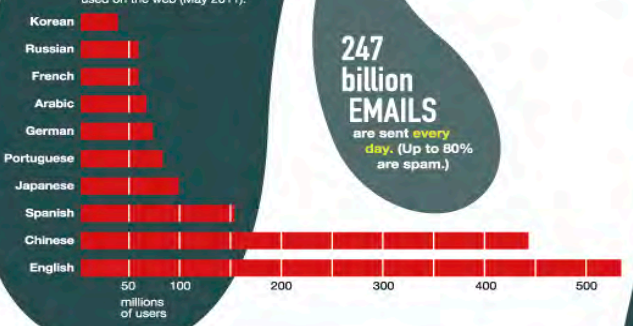
As of August 2012, there were just over **4 million** articles in the English Wikipedia.

There are **133 million BLOGS** on the web.

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

English is the dominant language of the web. But by 2014 it will be **Chinese,** if its current rate of increase continues.

Top languages used on the web (May 2011):



247 billion EMAILS are sent every day. (Up to 80% are spam.)

Every hour, we create enough Internet traffic to fill **7 billion DVDs.** Side by side, that's that's seven times the height of Everest.

10% of all photos ever taken were taken in 2011.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

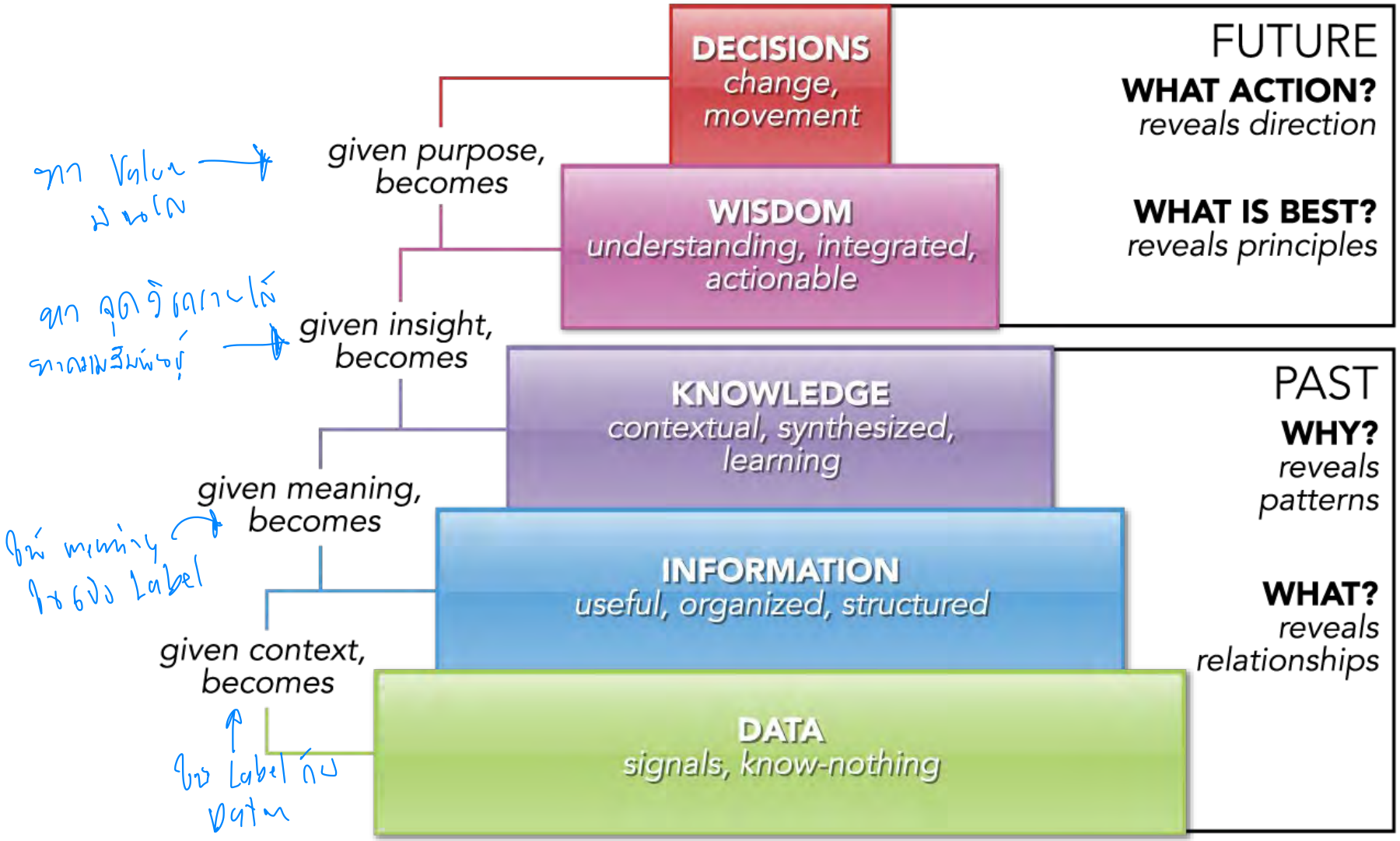
Information Graphics by Nigel Holmes

DATA

DECISION

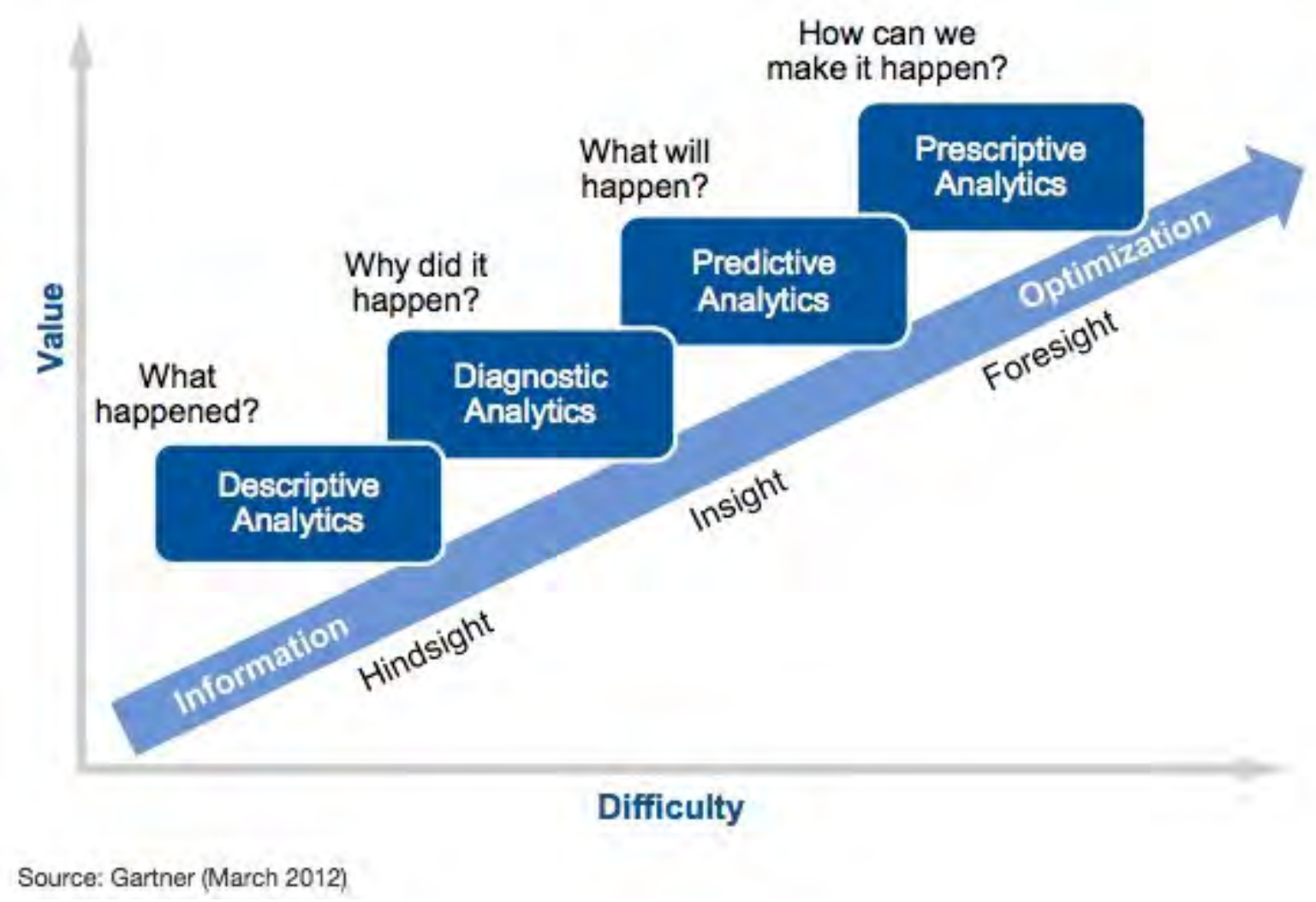
BIG DATA EXPERIENCE

DIKW (D) Pyramid



BIG DATA EXPERIENCE

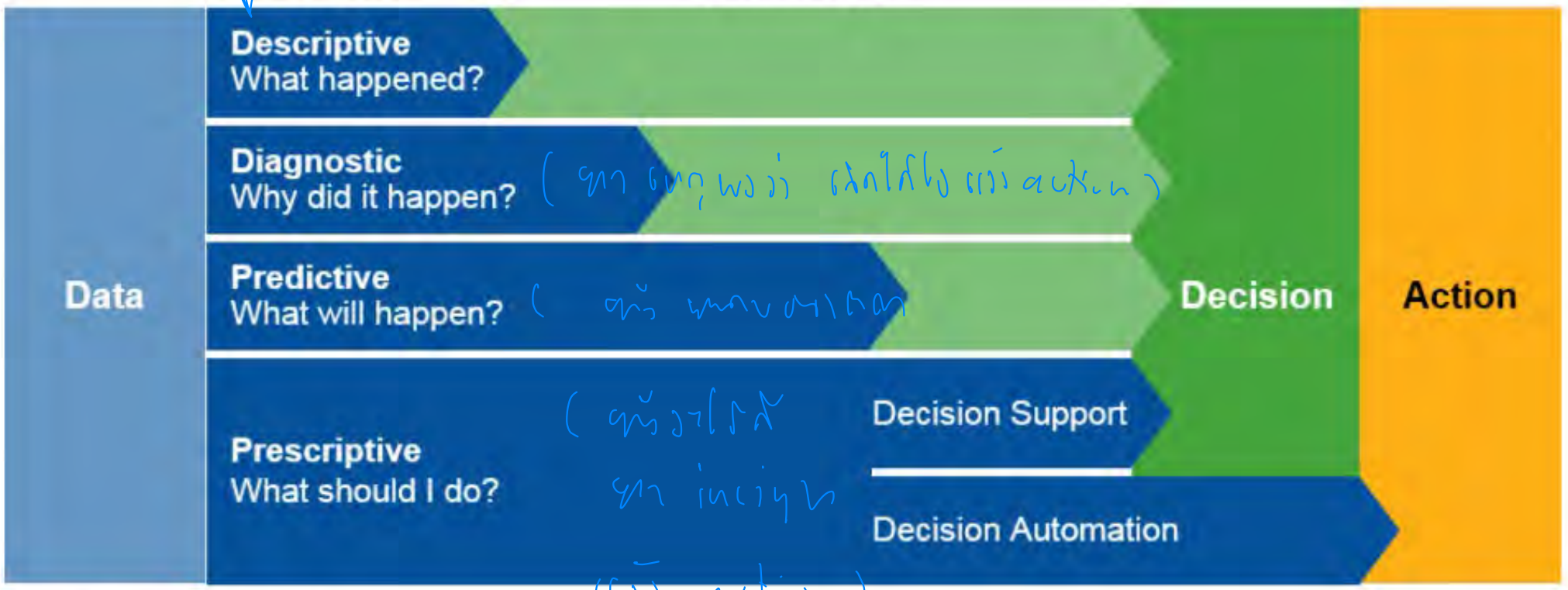
From descriptive … to prescriptive



Source: Gartner (March 2012)

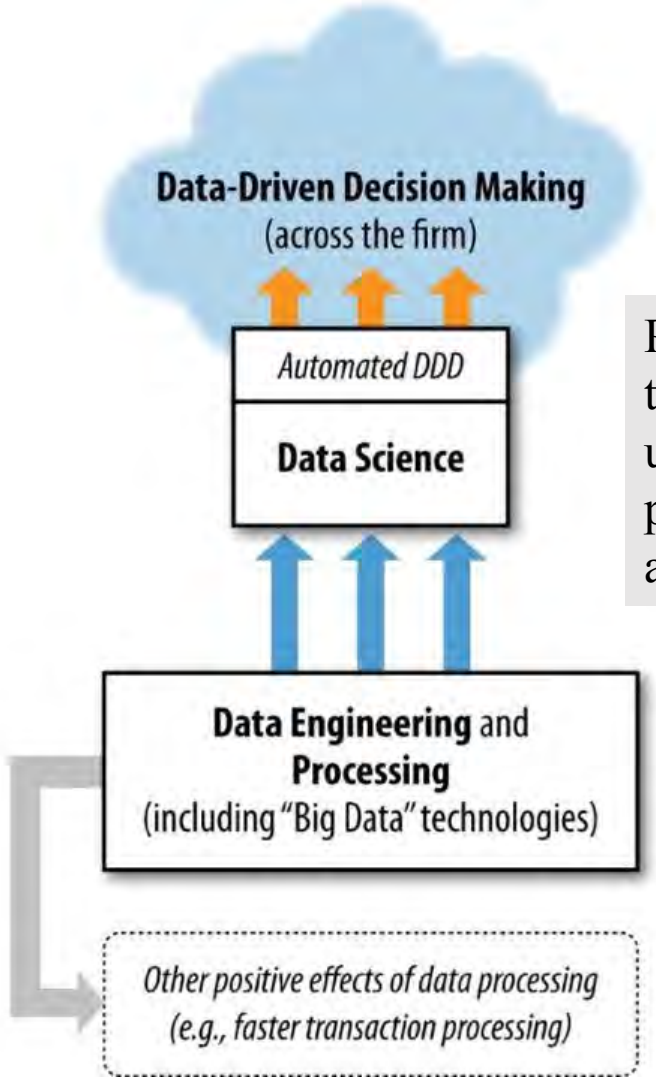
KMUTT g-able

Analytics Capabilities Framework



Source: Gartner (May 2015)

Data-driven decision making



DDD = practice of basing decision on the analysis of data, rather than intuition

Principles and techniques for understanding phenomena via the analysis of data.

Accessing and processing of massive-scale data flexibly and efficiently with Big Data technologies

BIG DATA EXPERIENCE

The Synopsis

The science

Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages.

The technology

From a large mass of data, IT can be used to find informative descriptive attributes of entities of interest

KMUTT g-able

BIG DATA EXPERIENCE

Steps in data analytics

Cross-Industry Standard Process for Data Mining

CRISP-DM

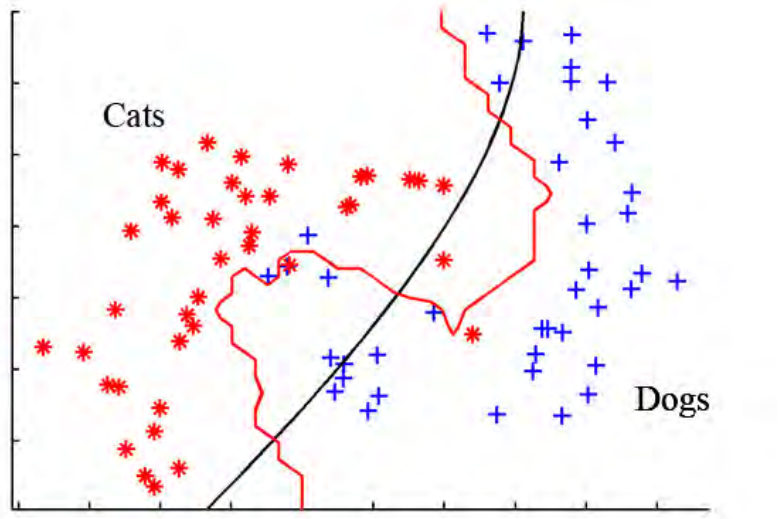
MACHINE LEARNING MODEL

Machine Learning

"The science of getting computers to learn from data without having to be explicitly programmed by humans."

Machine learning is surrounding you

- Google search
- Auto Facebook photo tagging
- Email Spamming
- Games
- Chat bot
- Recommender



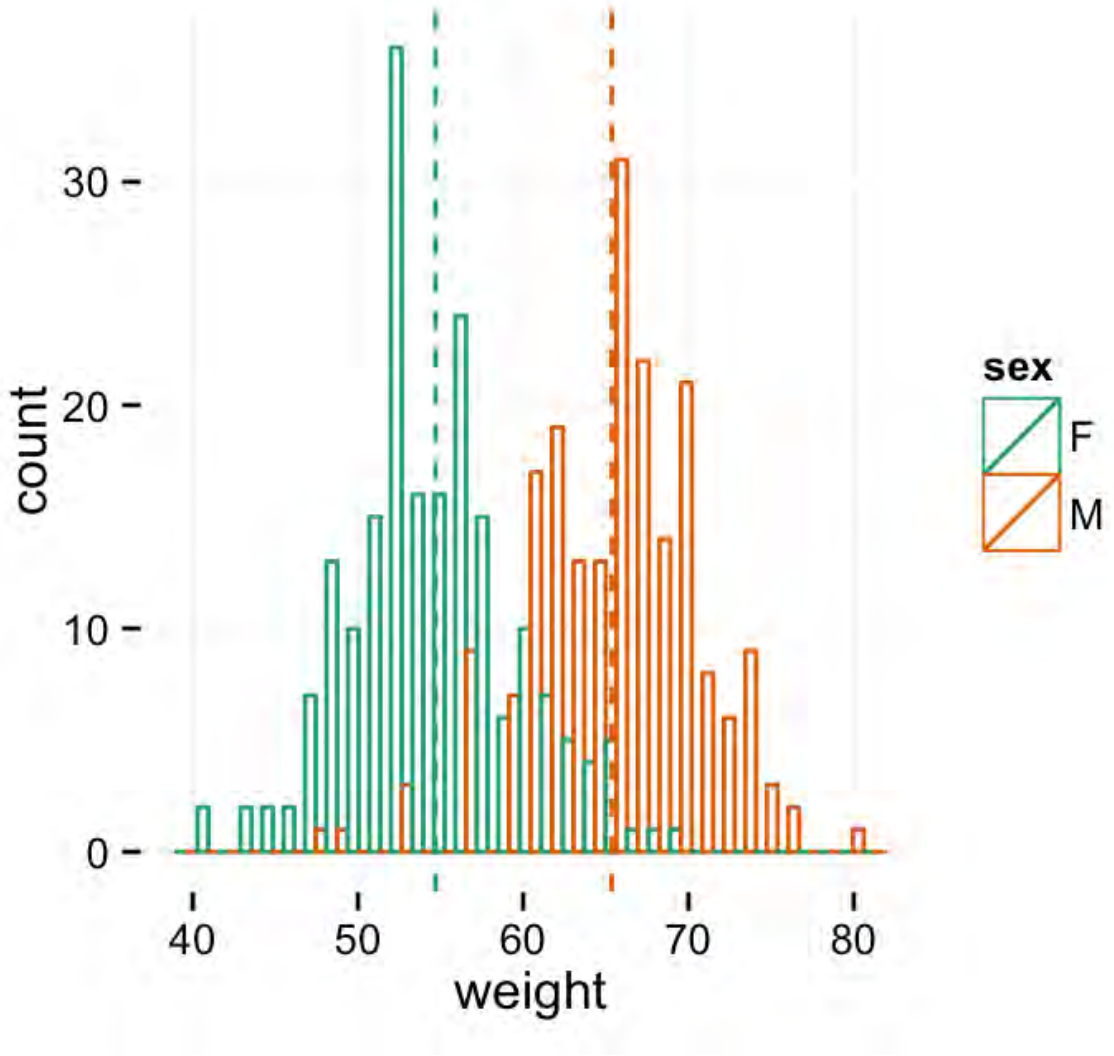
BIG DATA EXPERIENCE

Visualization

KMUTT g-able

BIG DATA EXPERIENCE

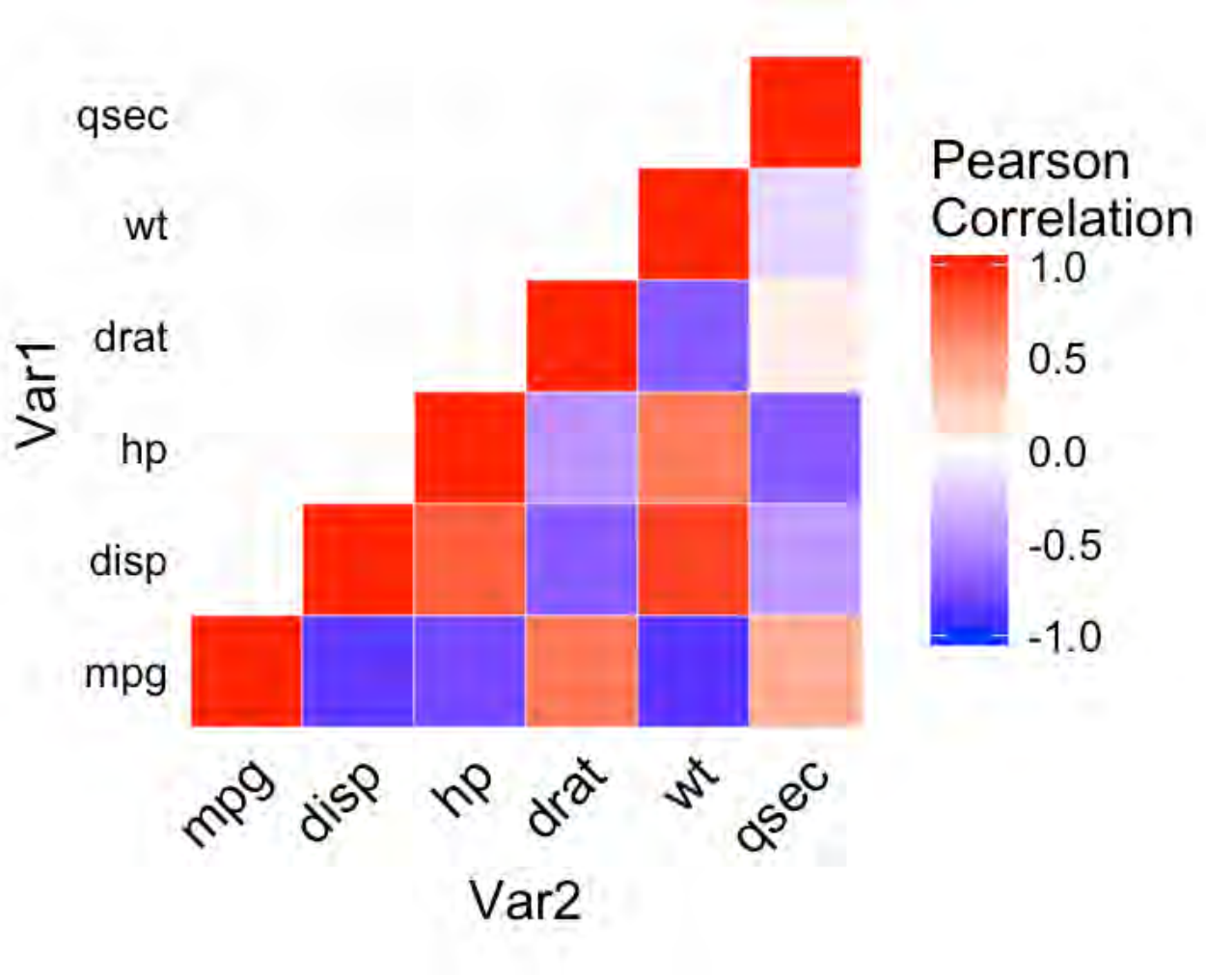
Visualization



KMUTT g-able

BIG DATA EXPERIENCE

Visualization

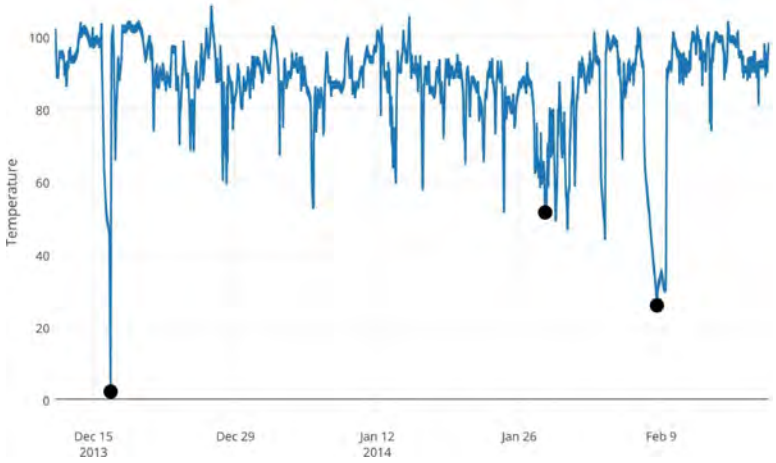


BIG DATA
EXPERIENCE

Use Case

KMUTT g-able

Anomaly detection Problem



Products

Pain Point

Sometime anomalies in production occur and result in fail QC

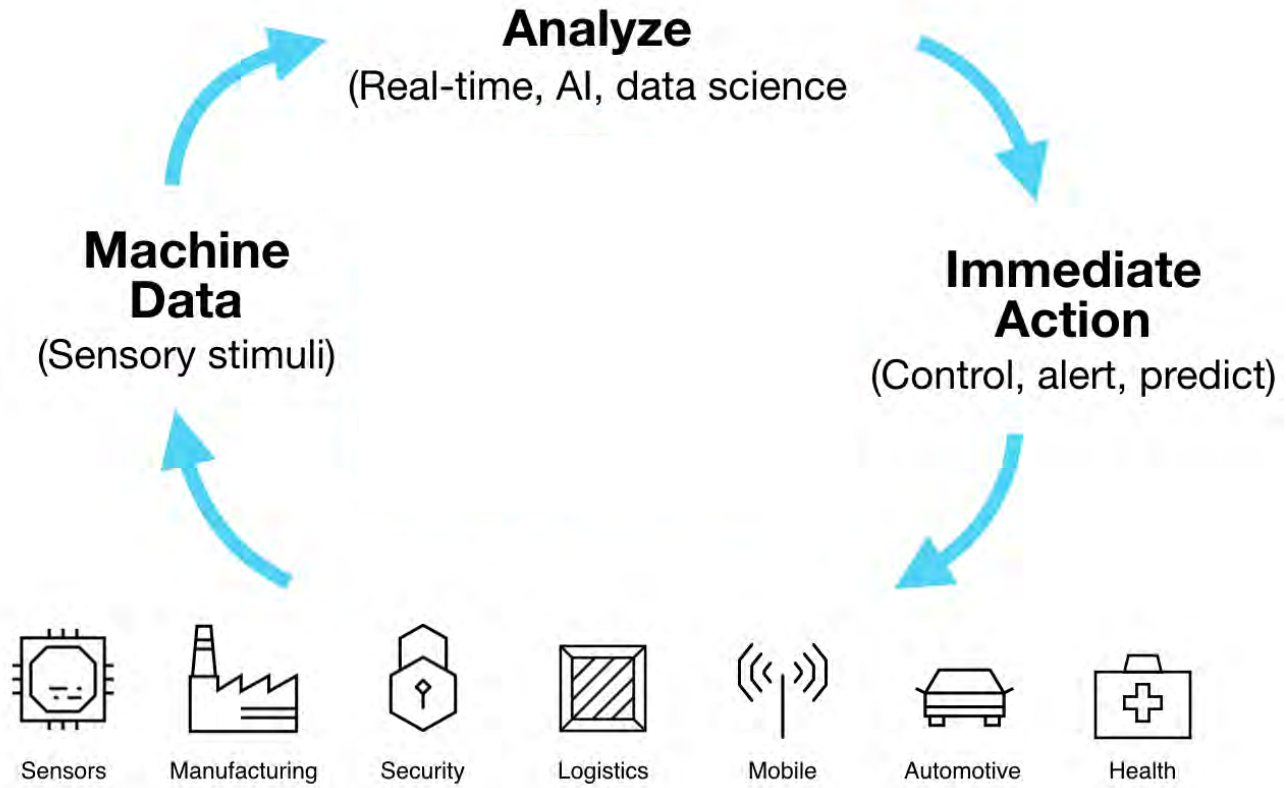
BIG DATA EXPERIENCE

Anomaly detection
Analytic objective

To monitor manufacturing parameters and detect the event that is not normal

KMUTT g-able

Anomaly detection
Machine data



Anomaly detection
Machine data

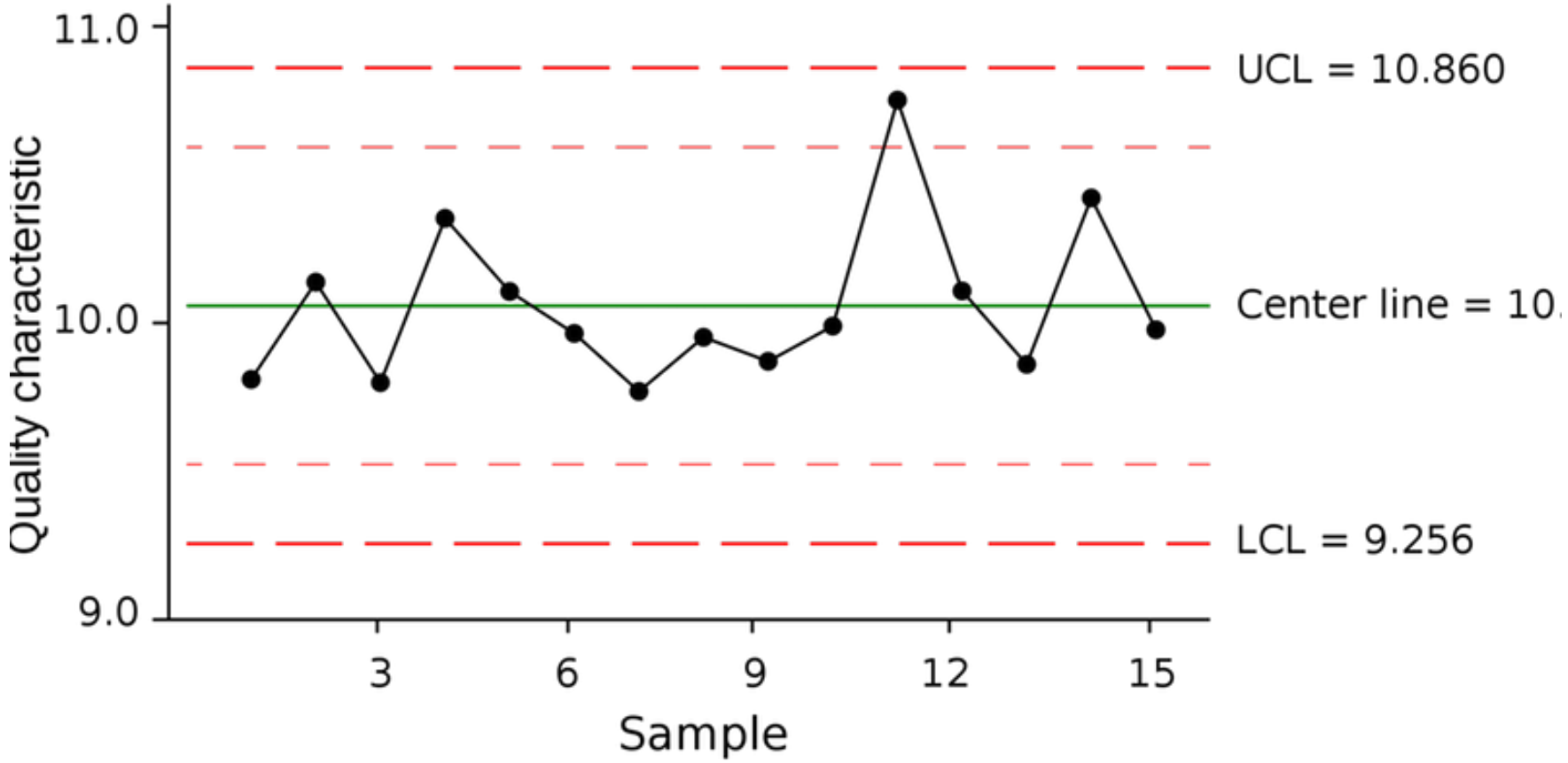
Sensor

Time

date	time	Inputv1	Inputv2	Inputv3	Inputv4	Inputv5	Inputv6	output
8/29/2018	19:50:00	1	0	0	1	0	0	1
8/29/2018	19:55:00	1	0	0	1	0	0	0
8/29/2018	20:00:00	1	0	0	1	0	0	1
8/29/2018	20:05:00	1	1	1	0	0	0	1
8/29/2018	20:10:00	1	1	1	0	0	0	1
8/29/2018	20:15:00	1	1	0	1	0	0	1
8/29/2018	20:20:00	1	1	0	1	1	0	0
8/29/2018	20:25:00	1	0	0	1	1	0	1
8/29/2018	20:30:00	1	0	0	1	1	0	1
8/29/2018	20:35:00	0	0	0	1	0	0	1
8/29/2018	20:40:00	1	0	0	1	1	0	1
8/29/2018	20:45:00	0	0	0	1	0	0	0

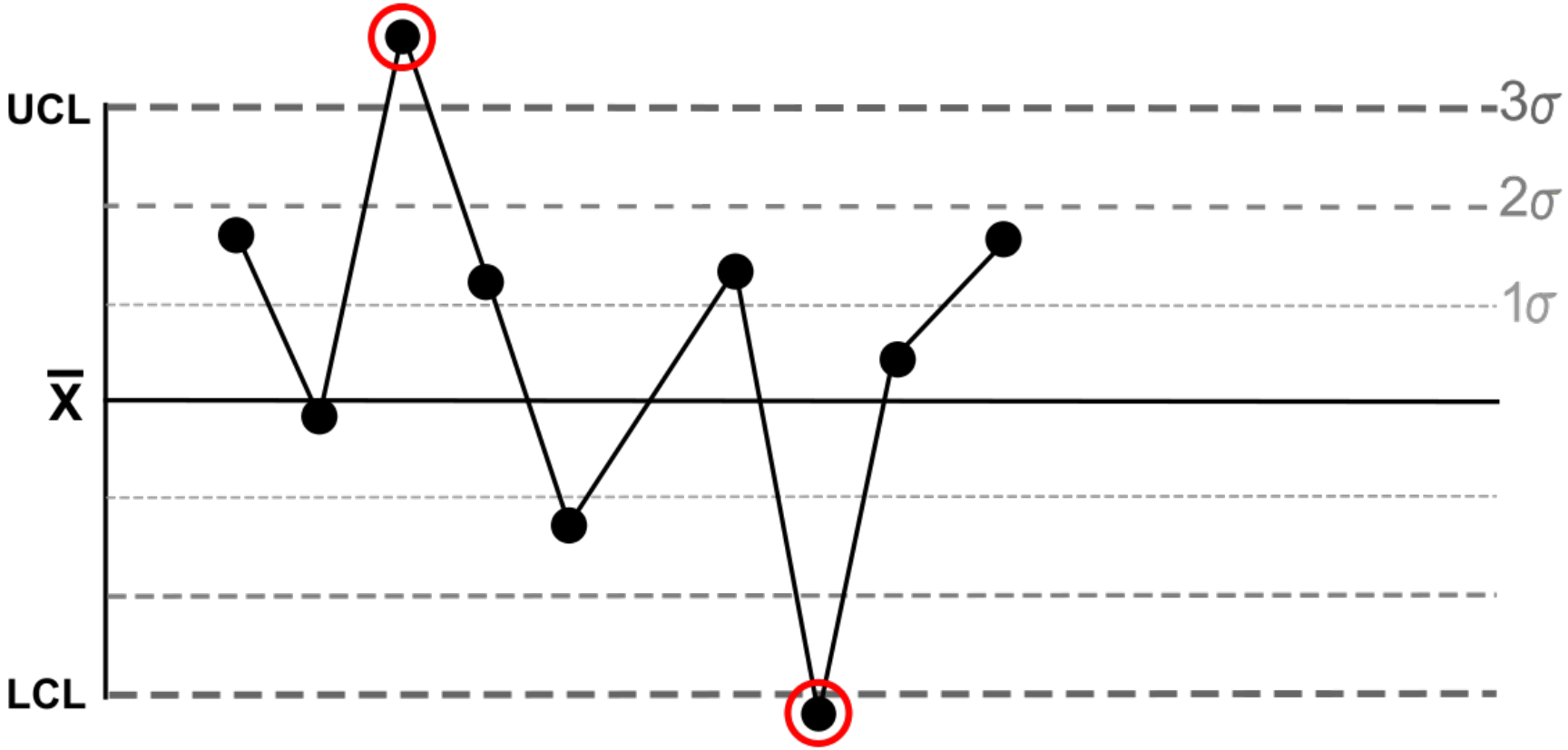
BIG DATA EXPERIENCE

Anomality detection
Statistical process control



BIG DATA EXPERIENCE

Anomaly detection



Anomaly detection
Outcome and usage

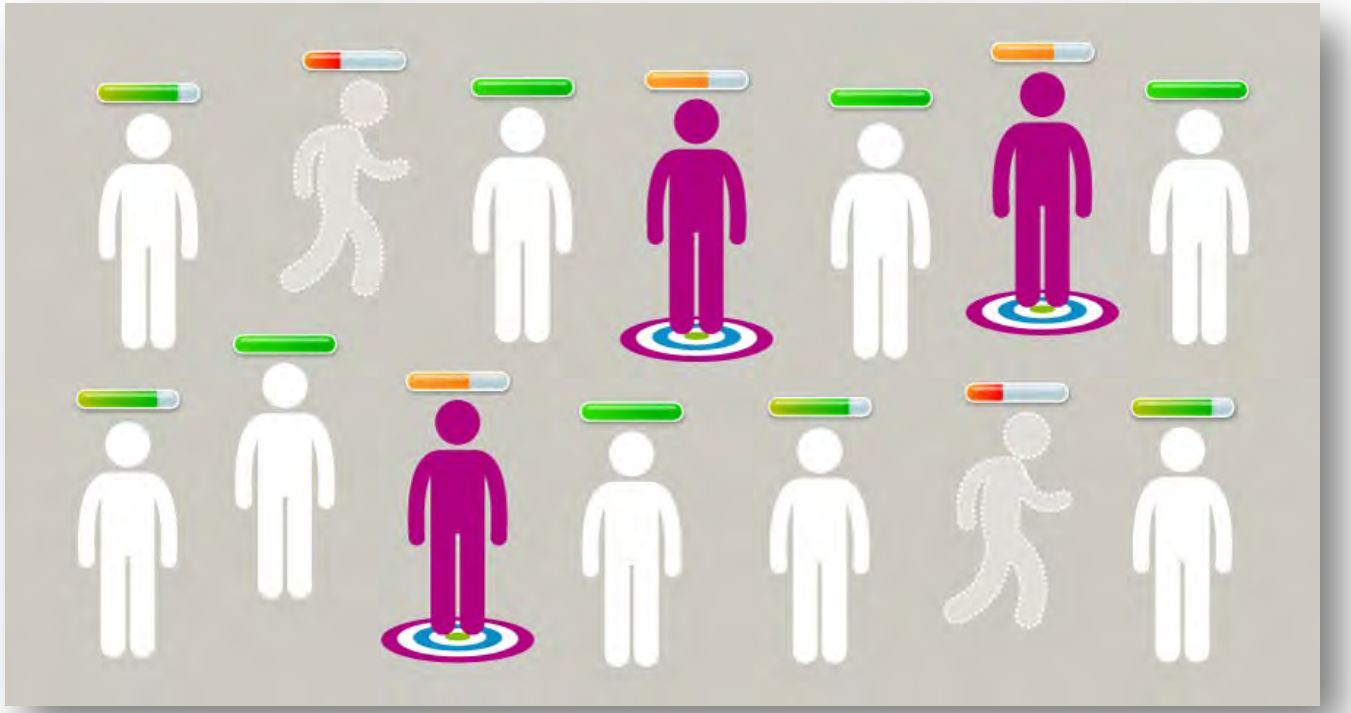
Usage

- Monitor the manufacturing parameters
- Detect and alert managers/operators once anomalous events are detected

Outcome

- Able to intervene with manufacturing process in time

Churn prediction Problem



Churn

Drop in Revenue

BIG DATA EXPERIENCE

Churn prediction
Analytic objective

Who are likely to churn?

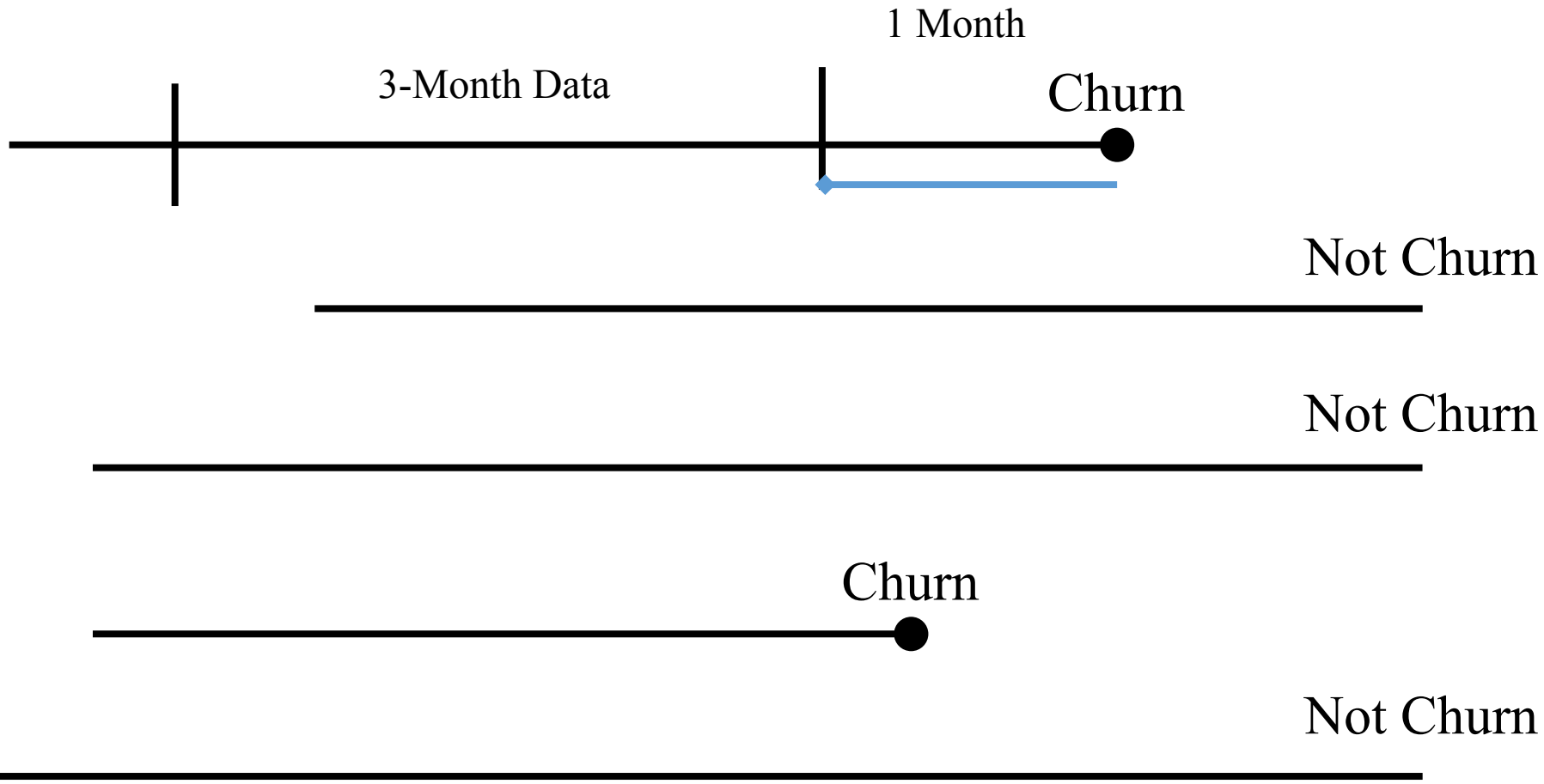
KMUTT g-able

Churn prediction Timeline

PROBLEM: we know very little about customers

BIG DATA EXPERIENCE

Churn prediction
Data collection



KMUTT g-able

Churn prediction
Data types

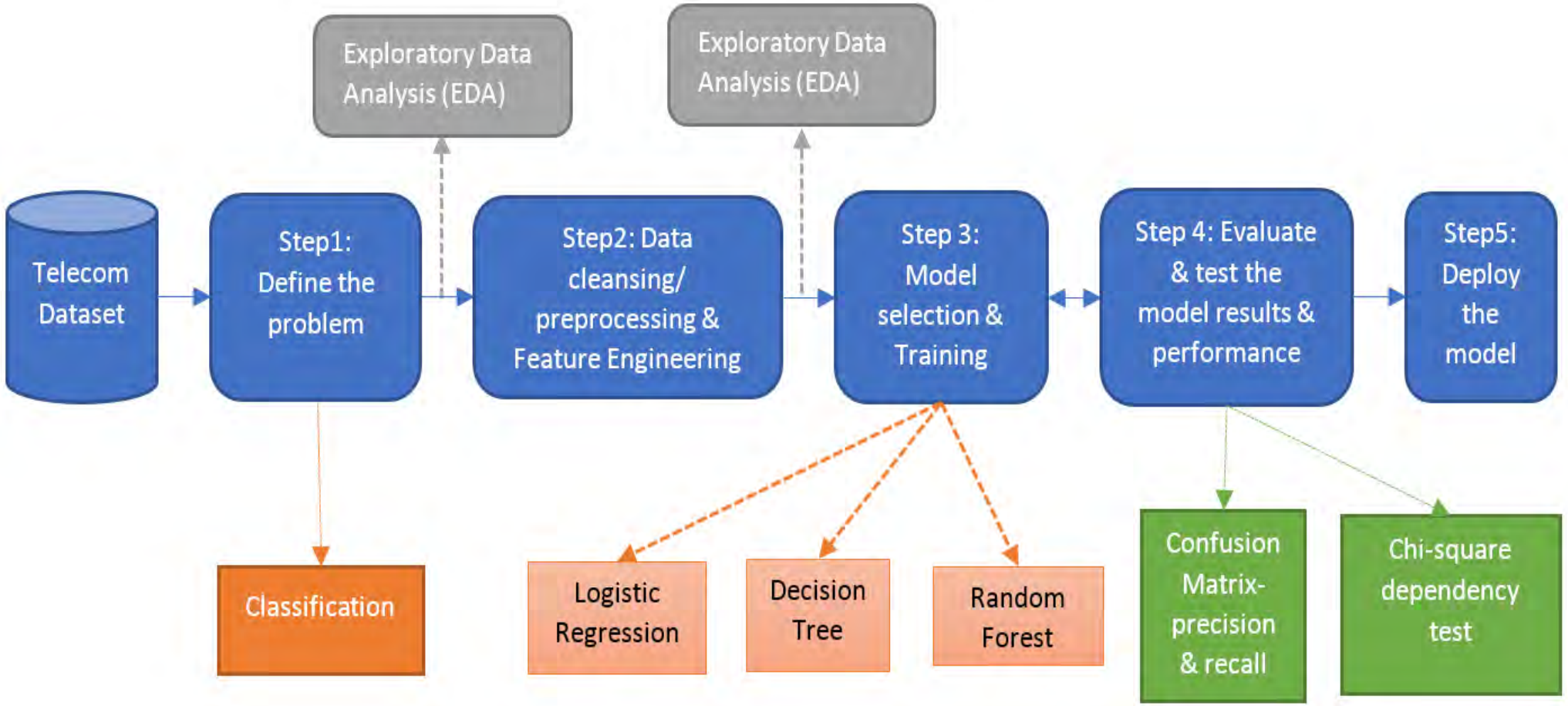
Age
Service Month
Handset Type
Main Pack

CDR
Payment
Internet usage

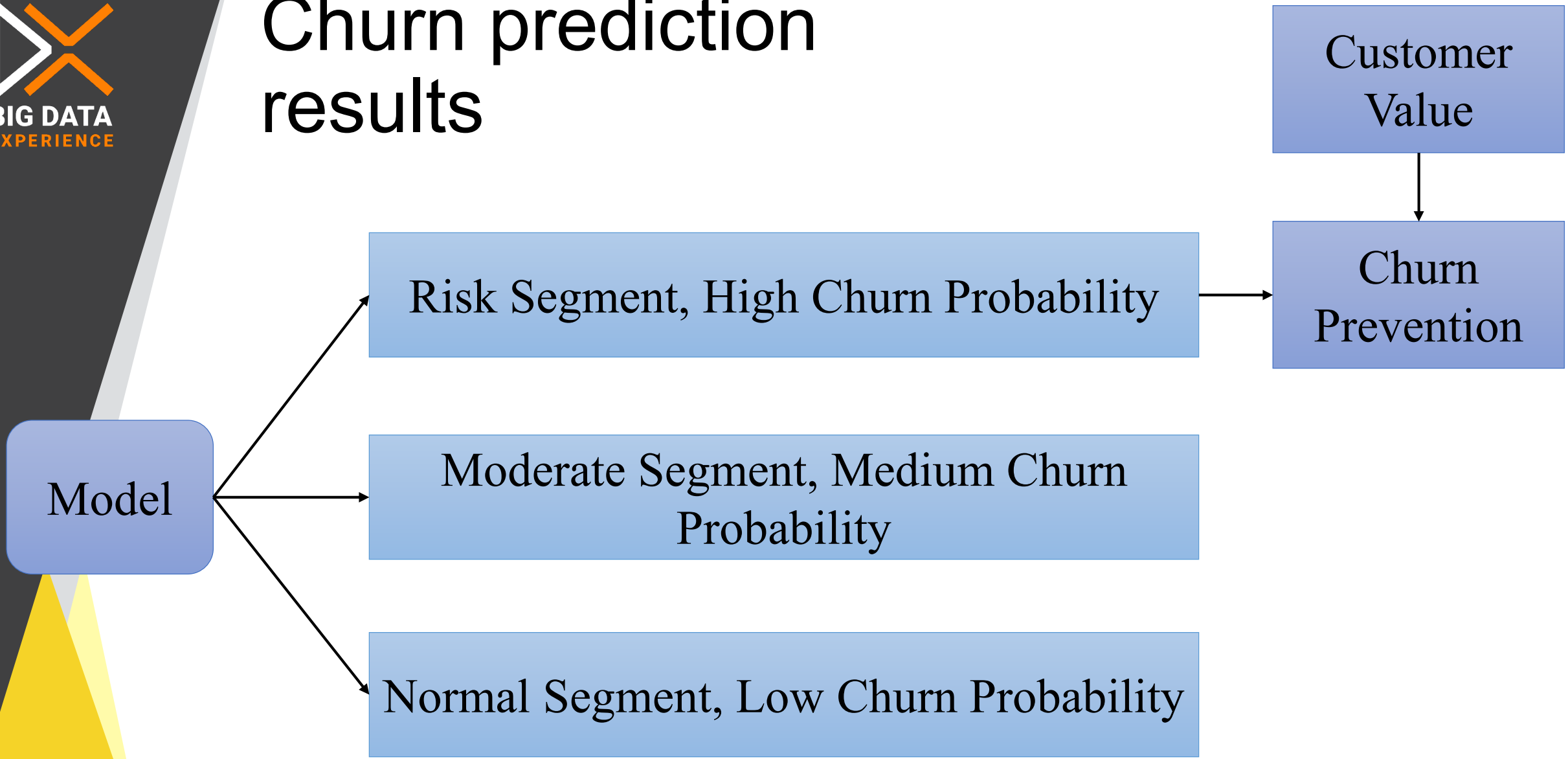
Product purchase
Browsing history
Preference

Call to contact center
Access to MyAIS

Churn prediction Model



Churn prediction results



Churn prediction
Outcome and usage

Outcome

- Able to identify potential churners

Usage

- Offer potential churners with retention campaigns

Data upsell Problem

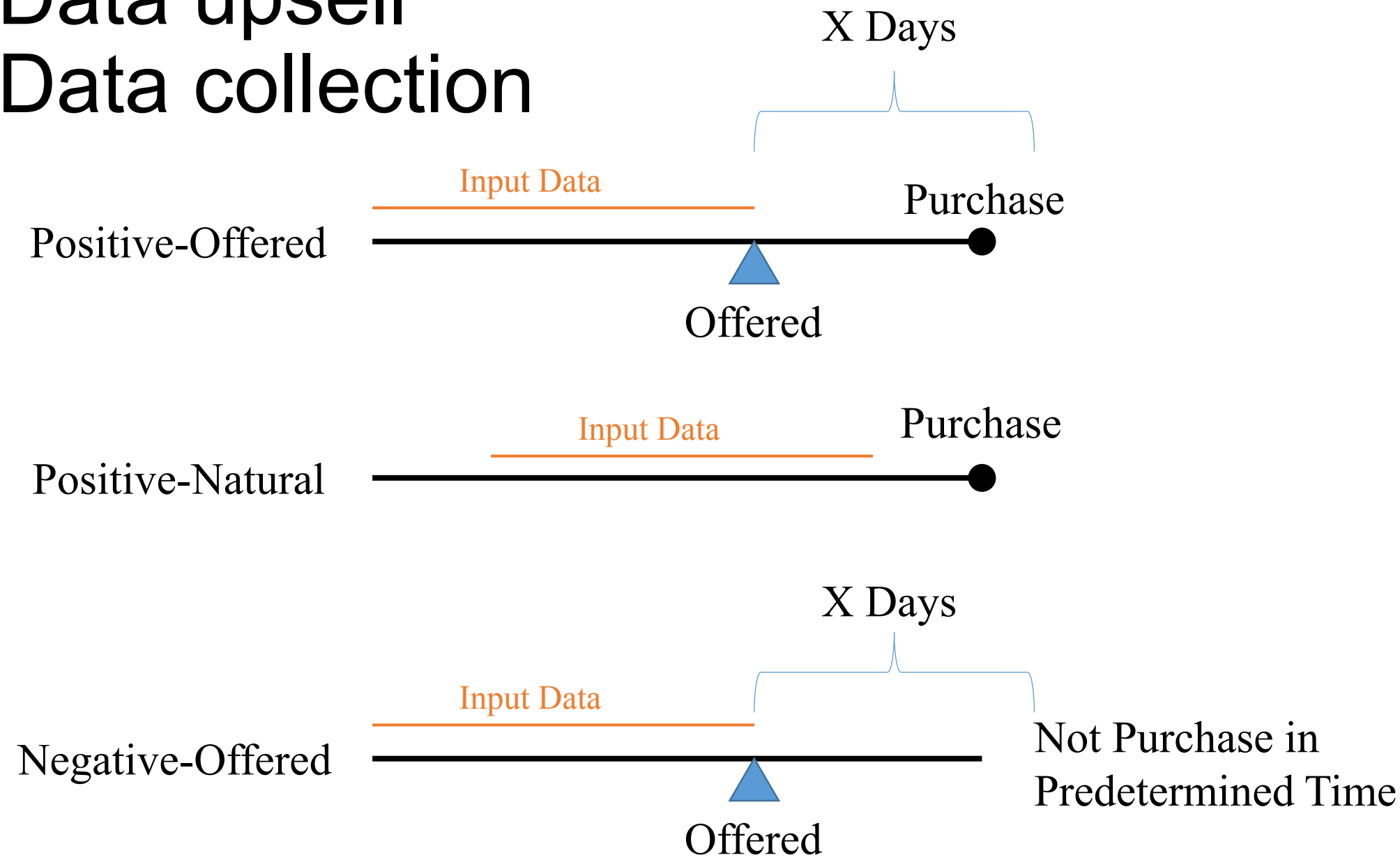
BIG DATA EXPERIENCE

Data upsell
Analytic objective

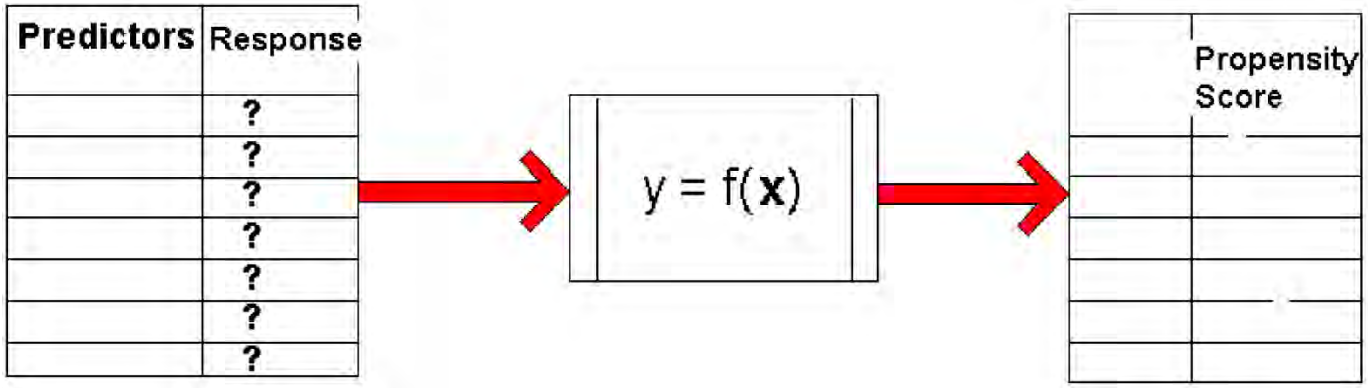
What product to offer? And to whom?

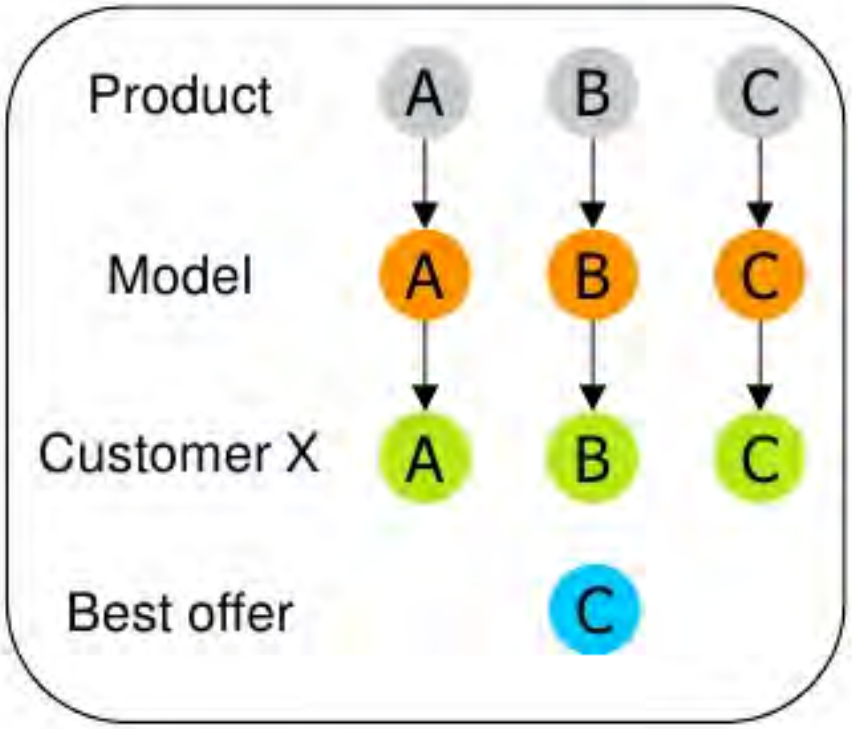
KMUTT g-able

Data upsell
Data collection



Data upsell Modeling





Each model is a binary classification model to predict product propensity.

BIG DATA EXPERIENCE

Data upsell
Outcome and usage

Usage

- Connect with the right channel to make automatic offers
- Know which products that each customer are likely to purchase

Outcome

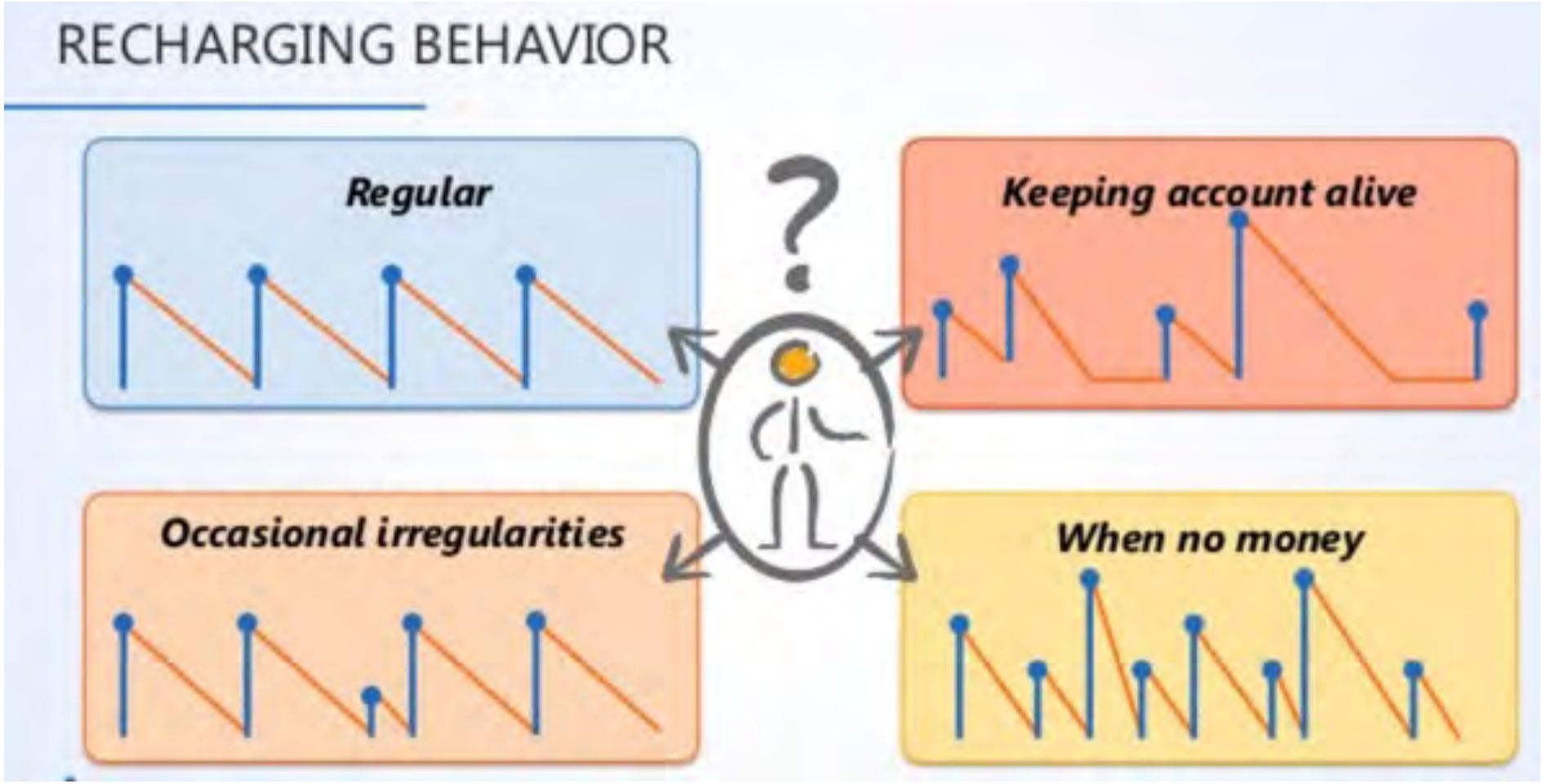
- Increase revenue through automatic upsell

KMUTT g-able

Top-up pattern analysis
Problem

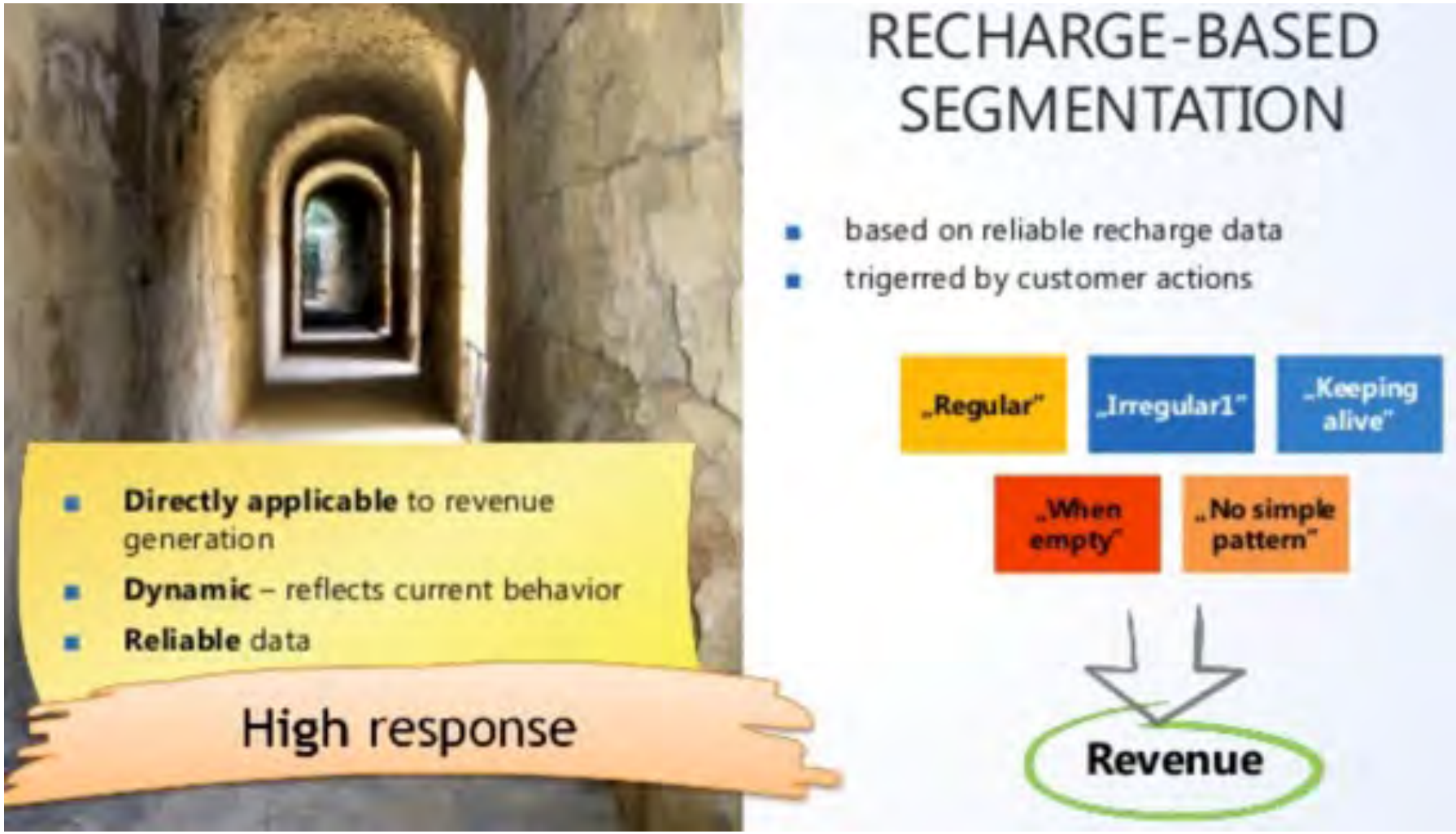
BIG DATA EXPERIENCE

Top-up pattern analysis
Top-up behaviors

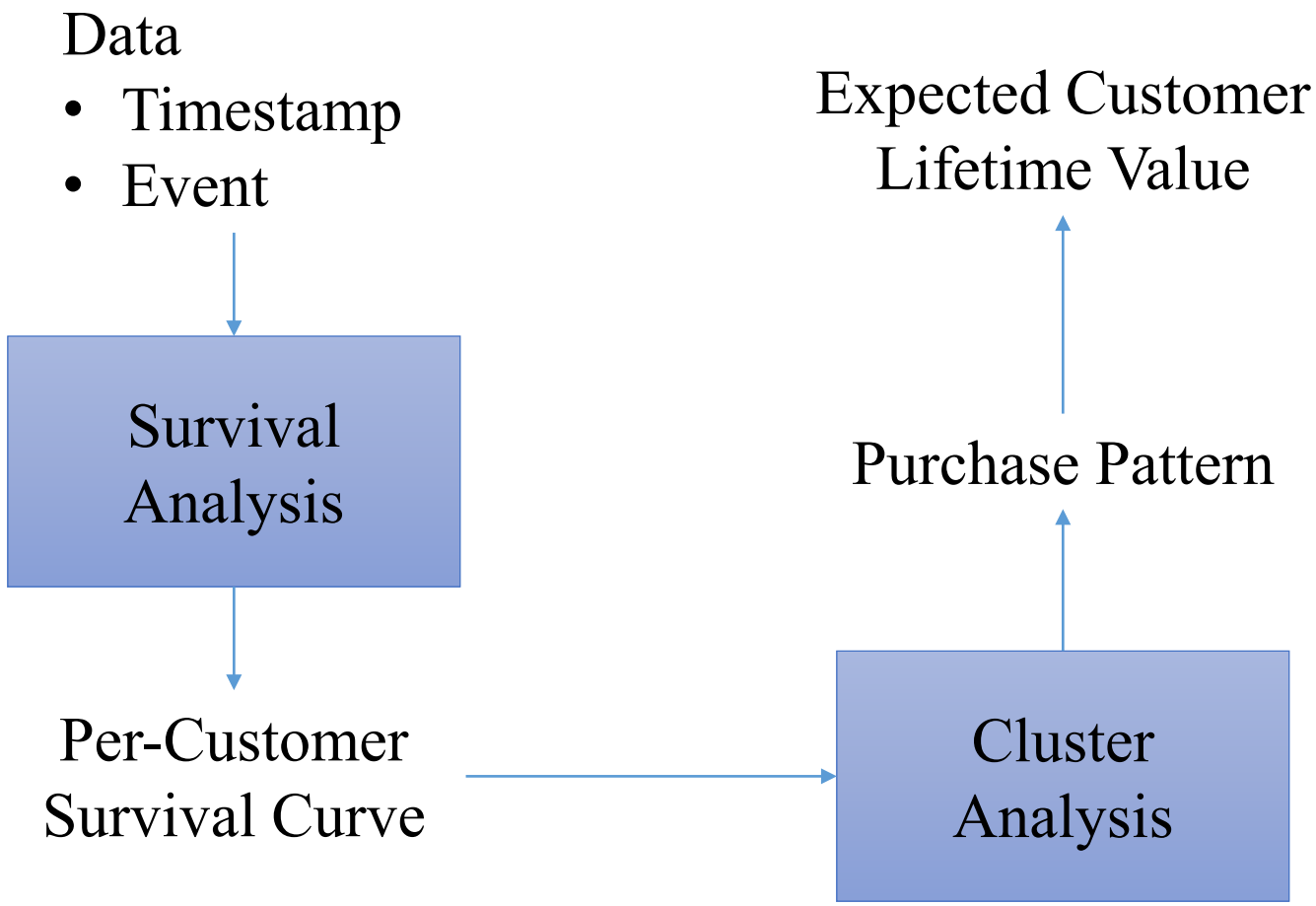


Top-up pattern analysis Baseline

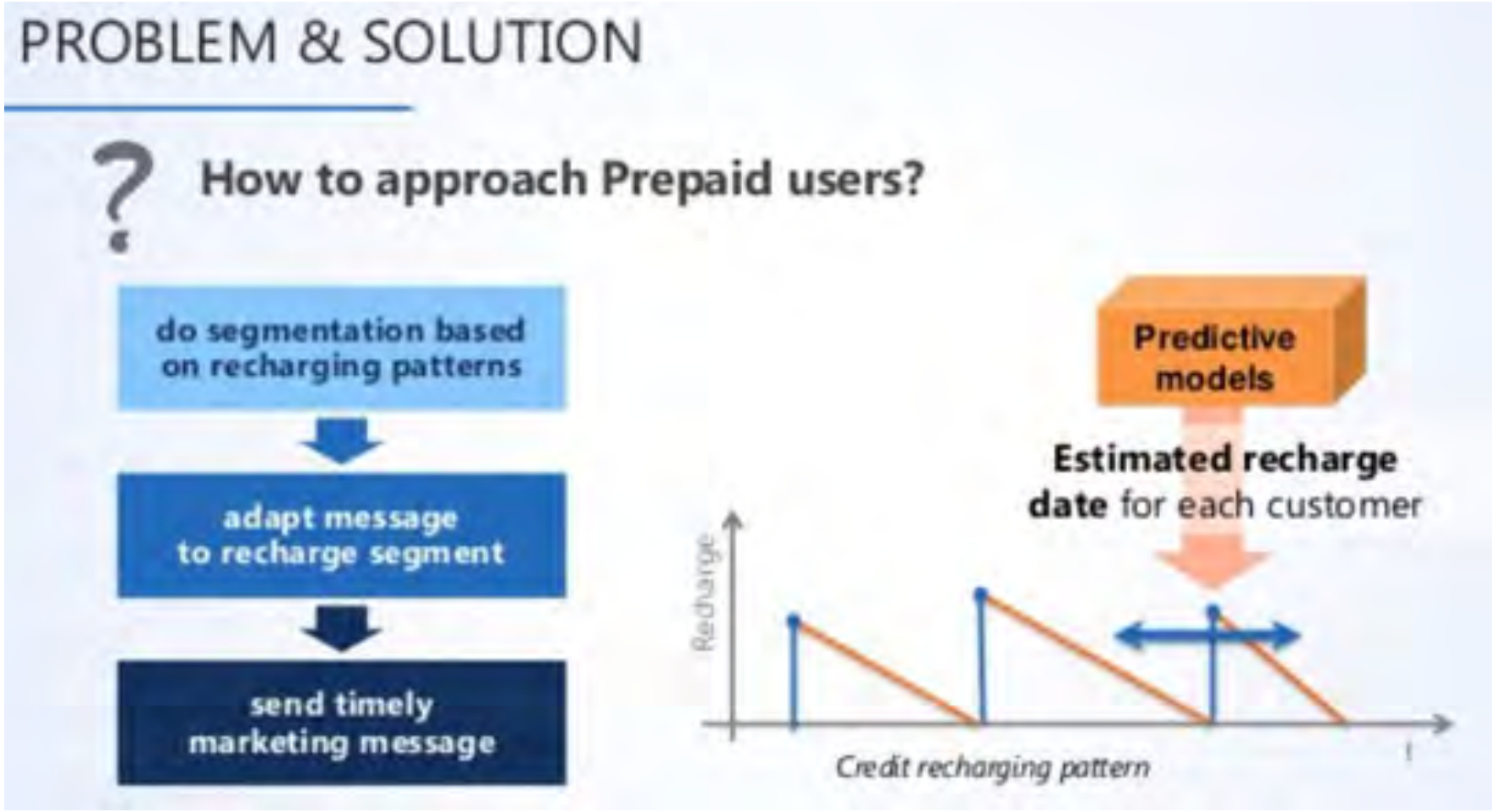
Top-up pattern analysis
Expected outcome



Modeling: Purchase Pattern

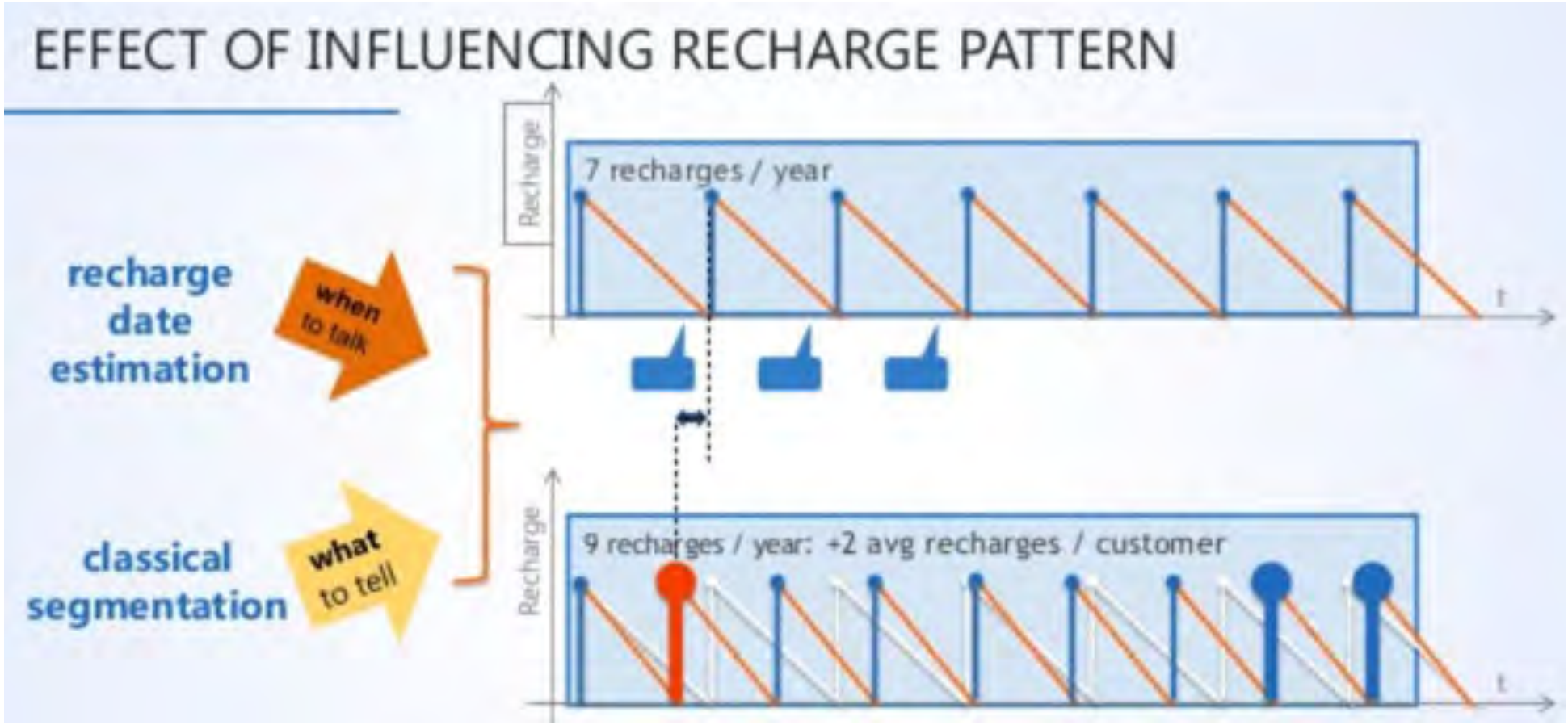


Top-up pattern analysis
Marketing plan



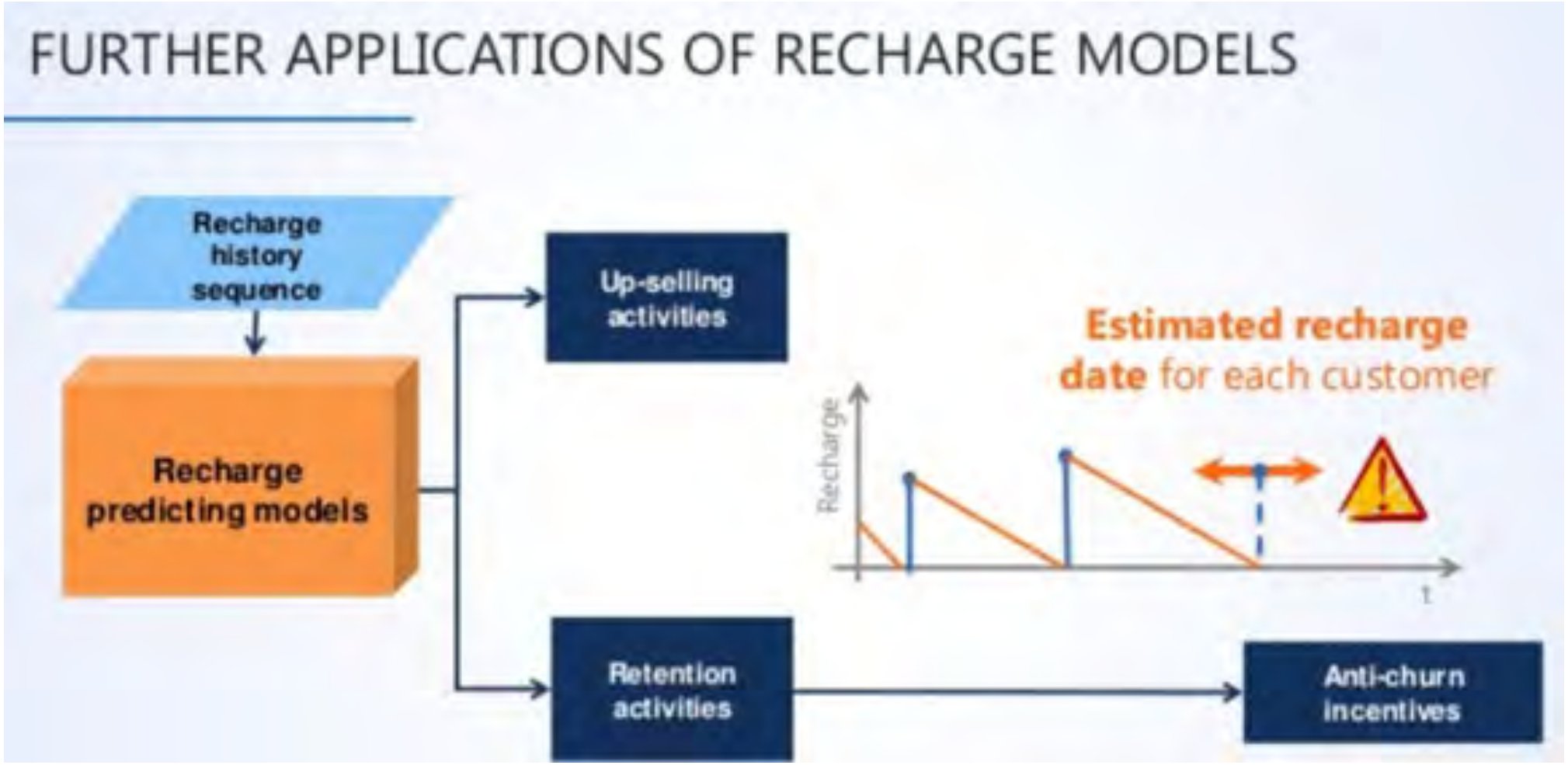
BIG DATA EXPERIENCE

Top-up pattern analysis
Potential influence



KM UTT g·able

Top-up pattern analysis
Further application



Credit risk score

❶ Application score (A-score)
เพื่ออนุมัติสินเชื่อใหม่

ผู้ขอสินเชื่อใหม่

- ข้อมูลจาก Loan Application
- ข้อมูลภายใน
- ข้อมูลจาก NCB

บันทึกข้อมูลในฐานข้อมูลภายใน

หากคะแนนผ่านเกณฑ์ขั้นต่ำ (Cut-off score) และไม่ขัดกับนโยบายสินเชื่อ (Product policy)

❷ Behaviour score (B-score)
เพื่อติดตามสินเชื่อ

ฐานข้อมูลพฤติกรรมการชำระหนี้/การใช้วงเงิน

ลูกค้าปัจจุบัน

- ไม่เคยค้างหนี้
- ค้าง < 30 วัน
- ค้าง 30-60 วัน
- ค้าง 60-90 วัน

ข้อมูลภายใน

ข้อมูลจาก NCB

- ลูกหนี้รายย่อยไม่มีงบการเงินเหมือนลูกหนี้นิติบุคคลจึงอาจไม่มีข้อมูลแหล่งที่มาของรายได้ที่เป็นปัจจุบัน
- จึงใช้ข้อมูล “พฤติกรรม” เพื่อวัดความสามารถในการชำระหนี้

- ติดตามความเสี่ยงลูกค้าแต่ละกลุ่ม
- ใช้คะแนนประกอบการต่ออายุ/วงเงินสินเชื่อ กำหนดอัตราดอกเบี้ย หรืออนุมัติสินเชื่อใหม่ (Product cross-selling)
- คะแนนต่างกัน Action ต่างกัน

NCB หมายถึง บริษัทข้อมูลเครดิตแห่งชาติ

Credit risk score
Data collection

1.2 การจัดเก็บข้อมูล: เตรียมฐานข้อมูลปัจจัยบ่งชี้ความน่าจะเป็นในการชำระหนี้คืน

ตัวอย่าง

Ability to pay +

Willingness to pay

ข้อมูลผู้ขอสินเชื่อ (Demography) มาจากใบคำขอสินเชื่อ

- เพศ อายุ การศึกษา
- อาชีพ / ประสบการณ์ทำงาน
- รายได้ปัจจุบัน

ข้อมูลประวัติการชำระหนี้ (Payment behavior)

- จำนวนครั้งที่ค้างชำระ 12 เดือนล่าสุด
- % การใช้วงเงินเฉลี่ยใน 3 เดือน
- ระยะเวลาไม่ชำระหนี้ใน 6 เดือน
- จำนวนบัตรเครดิตที่เปิดใหม่ใน 6 เดือน
- ยอดหนี้คงค้างทั้งหมด / รายได้
- จำนวนครั้งที่เช็คข้อมูล NCB ในอดีต 12 เดือน

เงื่อนไขการกู้ยืม

- สัดส่วน down payment
- ระยะเวลาการกู้ยืม
- ฯลฯ

ตัวอย่าง

ID	Gender	Home	...
RB0[redacted]001	F	BKK	...
DG0[redacted]166	M	Chiang Mai	...

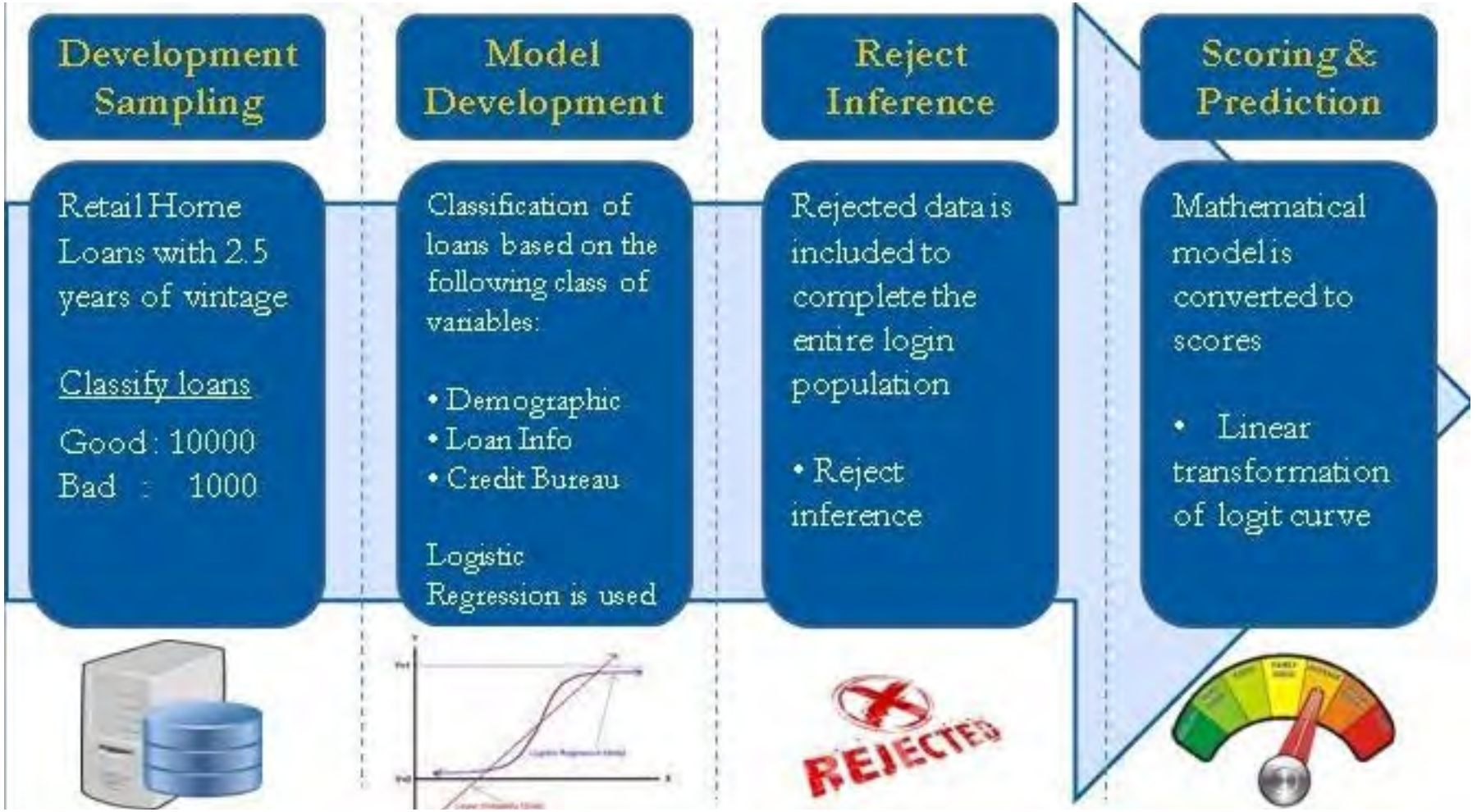
ข้อควรระวัง!

ข้อมูลที่นำมาใช้จัดทำ Credit scoring ต้องไม่สามารถระบุตัวตนของเจ้าของข้อมูลได้

Credit risk score Timeline

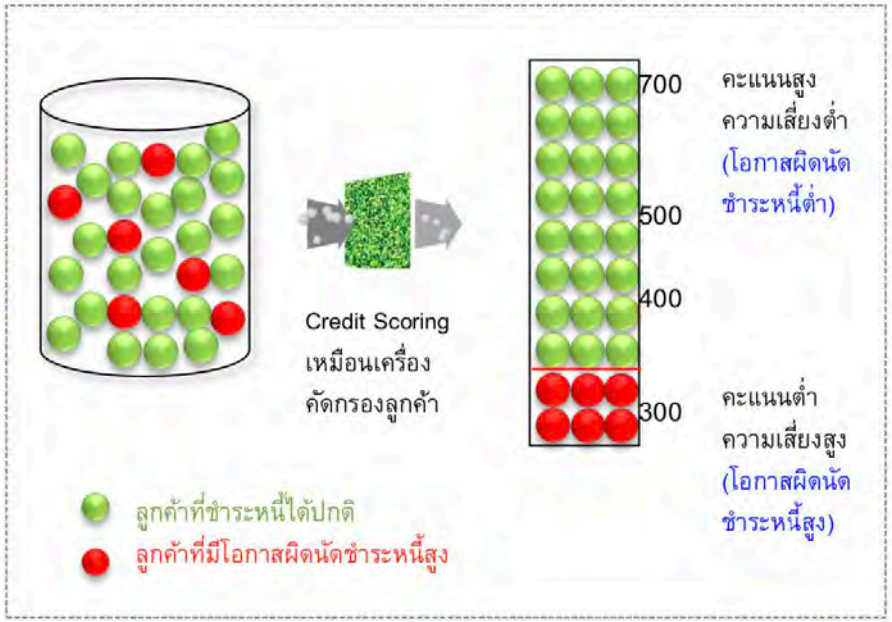
BIG DATA EXPERIENCE

Credit risk score
Model development



KMUTT g-able

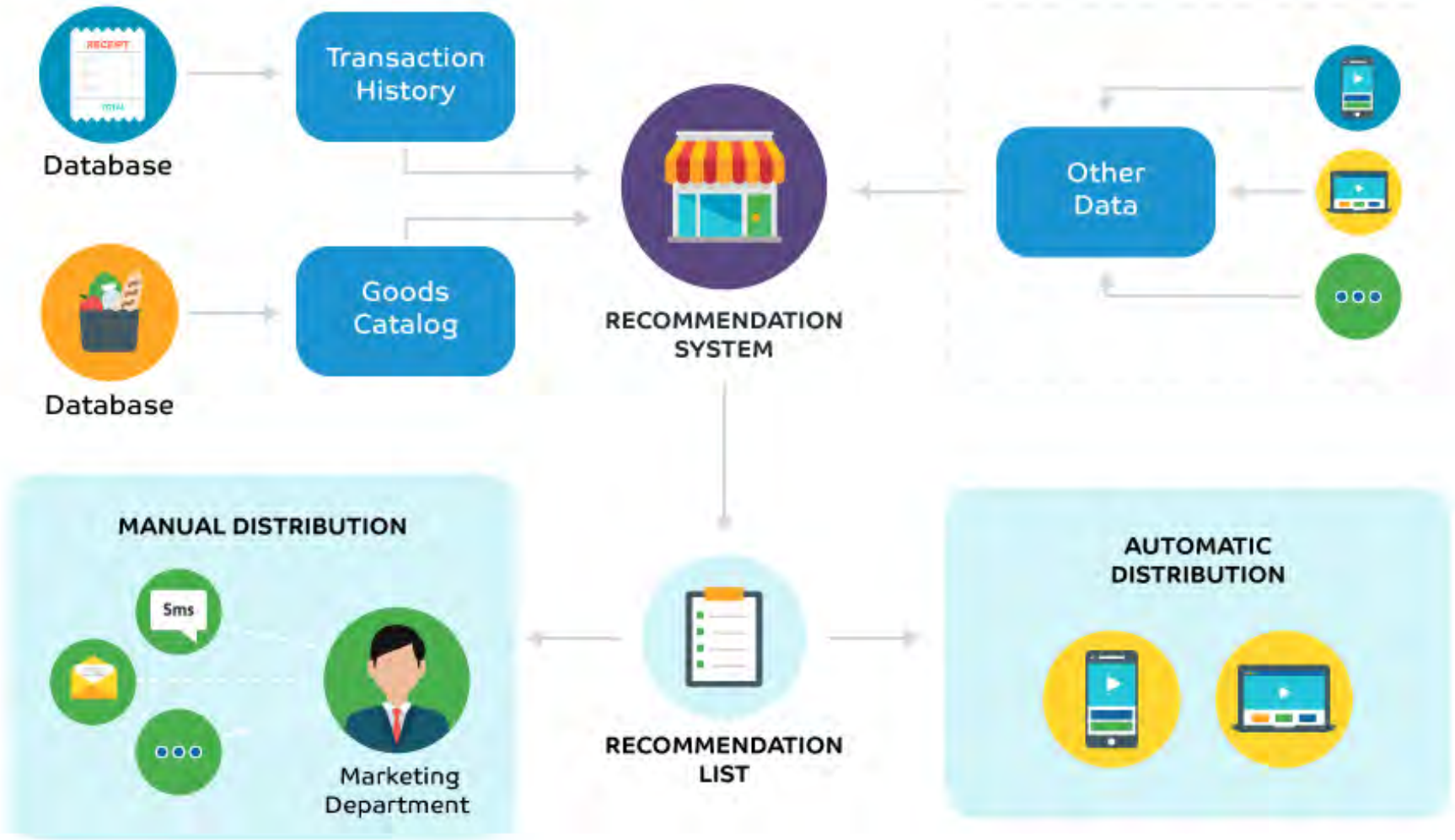
Credit risk score
Usage and outcome



ธนาคารพาณิชย์ และสถาบันการเงินต่าง ๆ จึงใช้ Credit Scoring เป็นเครื่องมือประกอบการวิเคราะห์สินเชื่อ และอนุมัติสินเชื่อ โดยเฉพาะสินเชื่อรายย่อย เช่น สินเชื่อบัตรเครดิต สินเชื่อบุคคล สินเชื่อบ้าน สินเชื่อเช่าซื้อรถยนต์ เป็นต้น

BIG DATA EXPERIENCE

Use Case: Product Recommendation



KMUTT g-able

BIG DATA EXPERIENCE

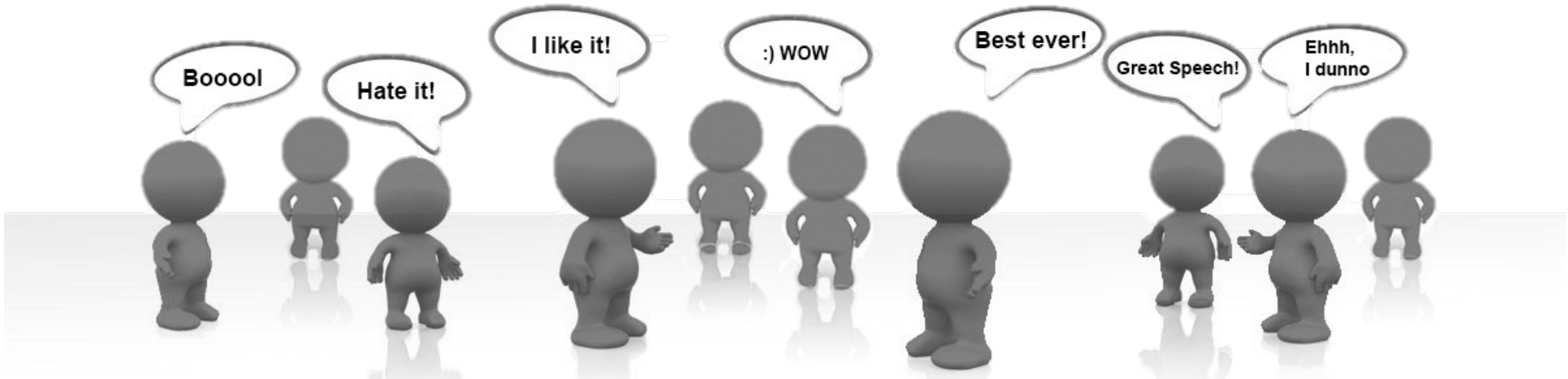
Use Case: Customer Preference

- Zarola derives customer preference and styles based on their transactions
- It optimizes market strategies based on each user profile.

ZALORA

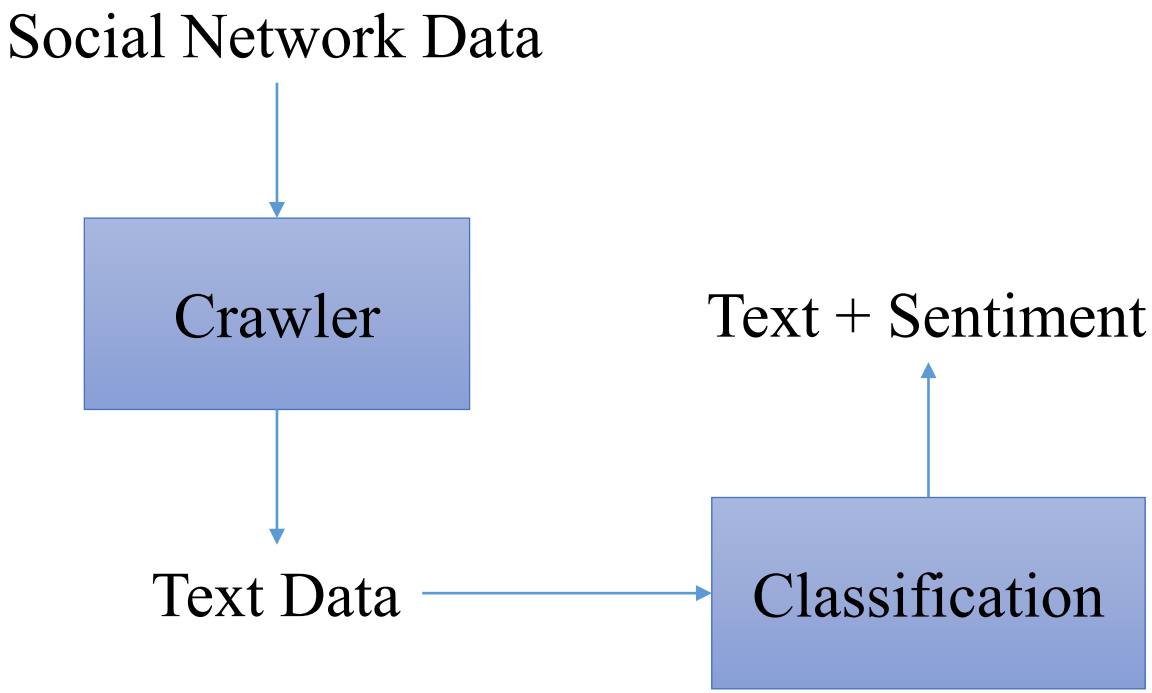
Use Case: Sentiment Analysis

- To determine the attitude of a writer with respect to some topic or the overall contextual polarity of a document.
- Widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service

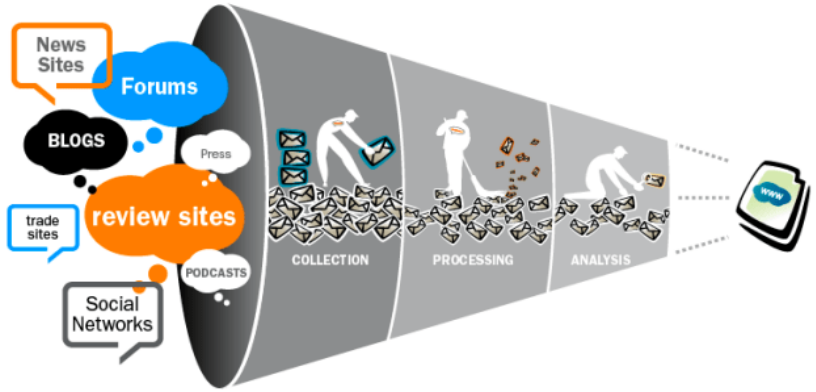


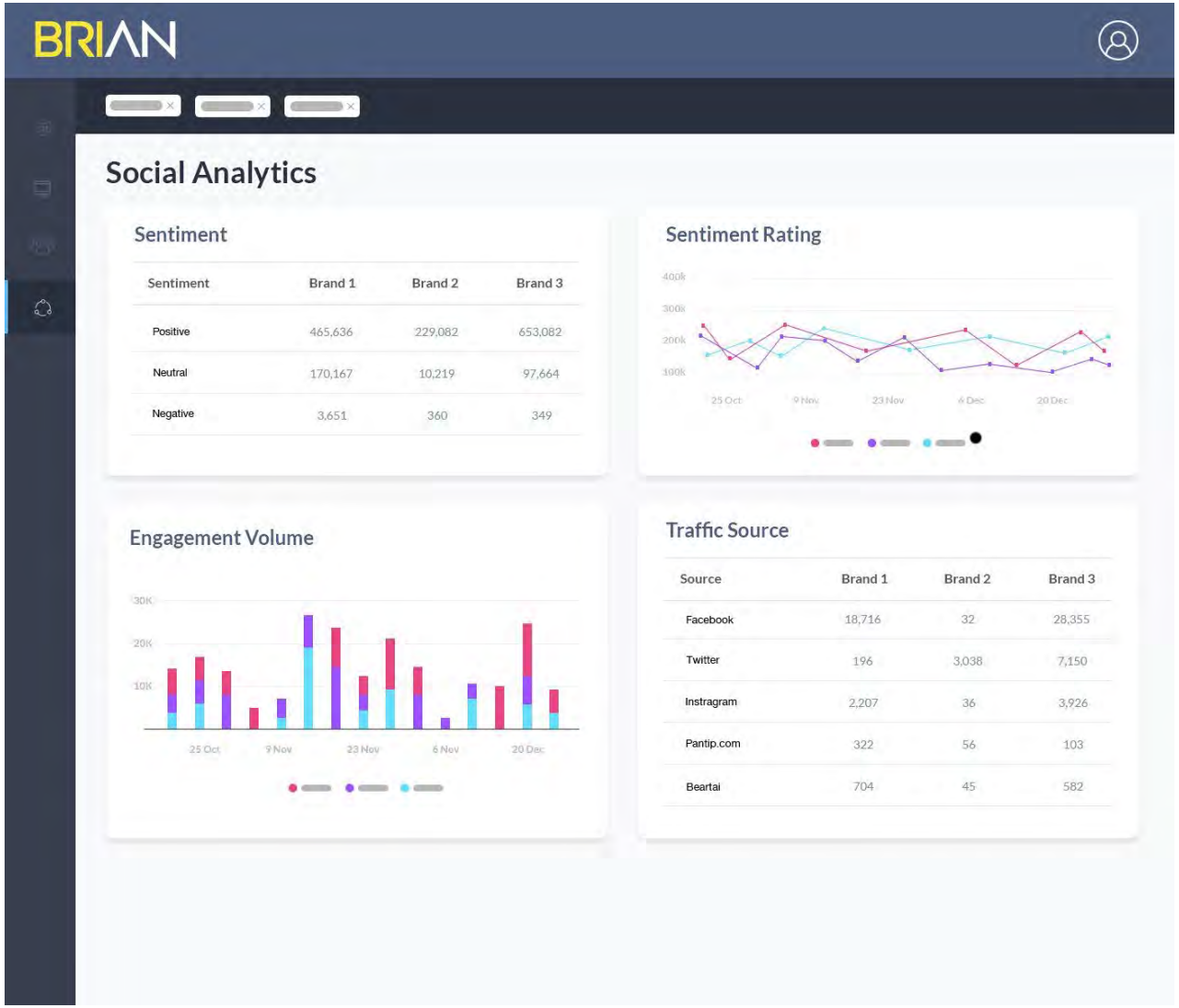
BIG DATA EXPERIENCE

Use Case: Sentiment Analysis



Use Case: Brand Tracking



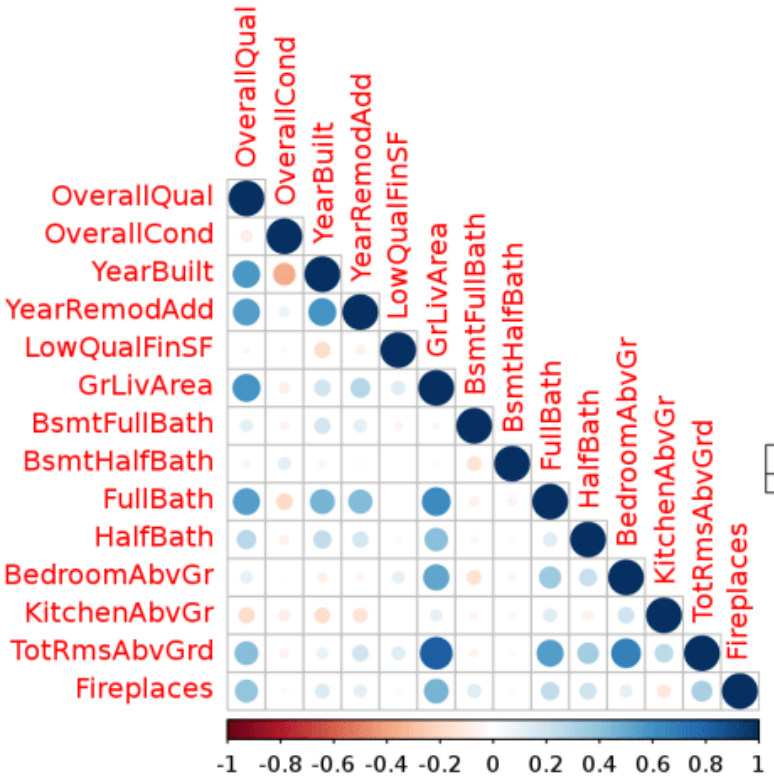


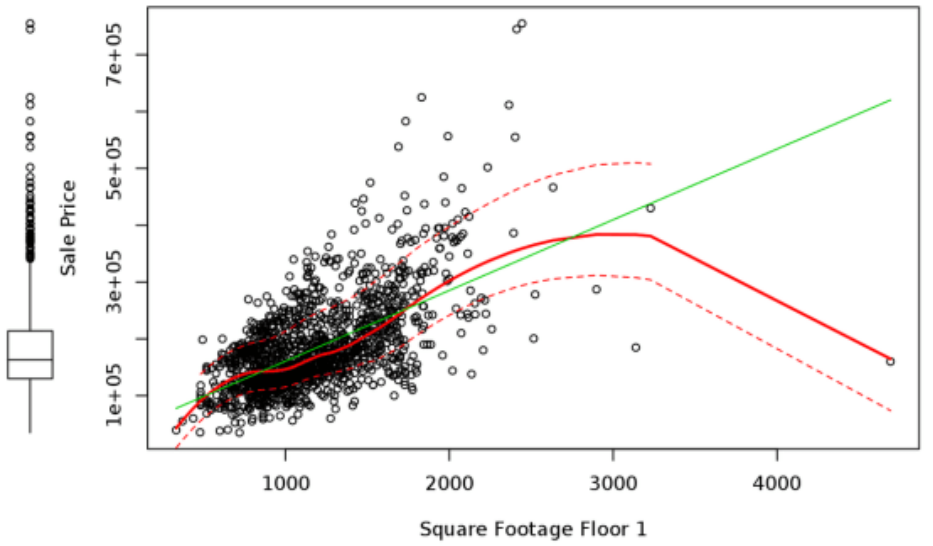
Sentiment	Brand 1	Brand 2	Brand 3
Positive	465,636	229,082	653,082
Neutral	170,167	10,219	97,664
Negative	3,651	360	349

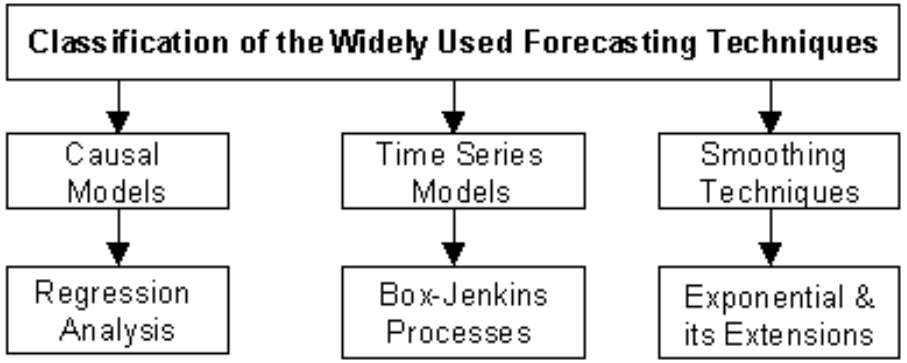
Source	Brand 1	Brand 2	Brand 3
Facebook	18,716	32	28,355
Twitter	196	3,038	7,150
Instragram	2,207	36	3,926
Pantip.com	322	56	103
Beartai	704	45	582

BIG DATA EXPERIENCE

House Price Prediction







KMUTT g-able

BIG DATA
EXPERIENCE

End of Lecture 1

Question?

BIG DATA
EXPERIENCE
CENTER

KMUTT g-able