

# CPE 352 Data Science

## 1 – Data Science Concepts

**Asst. Prof. Dr. Santitham Prom-on**

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

# Learning Outcome

- เข้าใจกระบวนการด้านวิทยาศาสตร์ข้อมูลและบทบาทของนักวิทยาศาสตร์ข้อมูล (PLO 1F, 1E)
- ใช้ Python ในการจัดการข้อมูลเพื่อเตรียมสำหรับการวิเคราะห์ (PLO 1C)
- สร้างการแสดงผลข้อมูลที่มีความหมายและตีความเพื่อใช้ในการตอบคำถาม (PLO 2B, 2C)
- ใช้วิธีการทางคณิตศาสตร์ สถิติ และการเรียนรู้ของเครื่องเพื่อแก้ปัญหาด้านข้อมูล (PLO 1A, 1B)

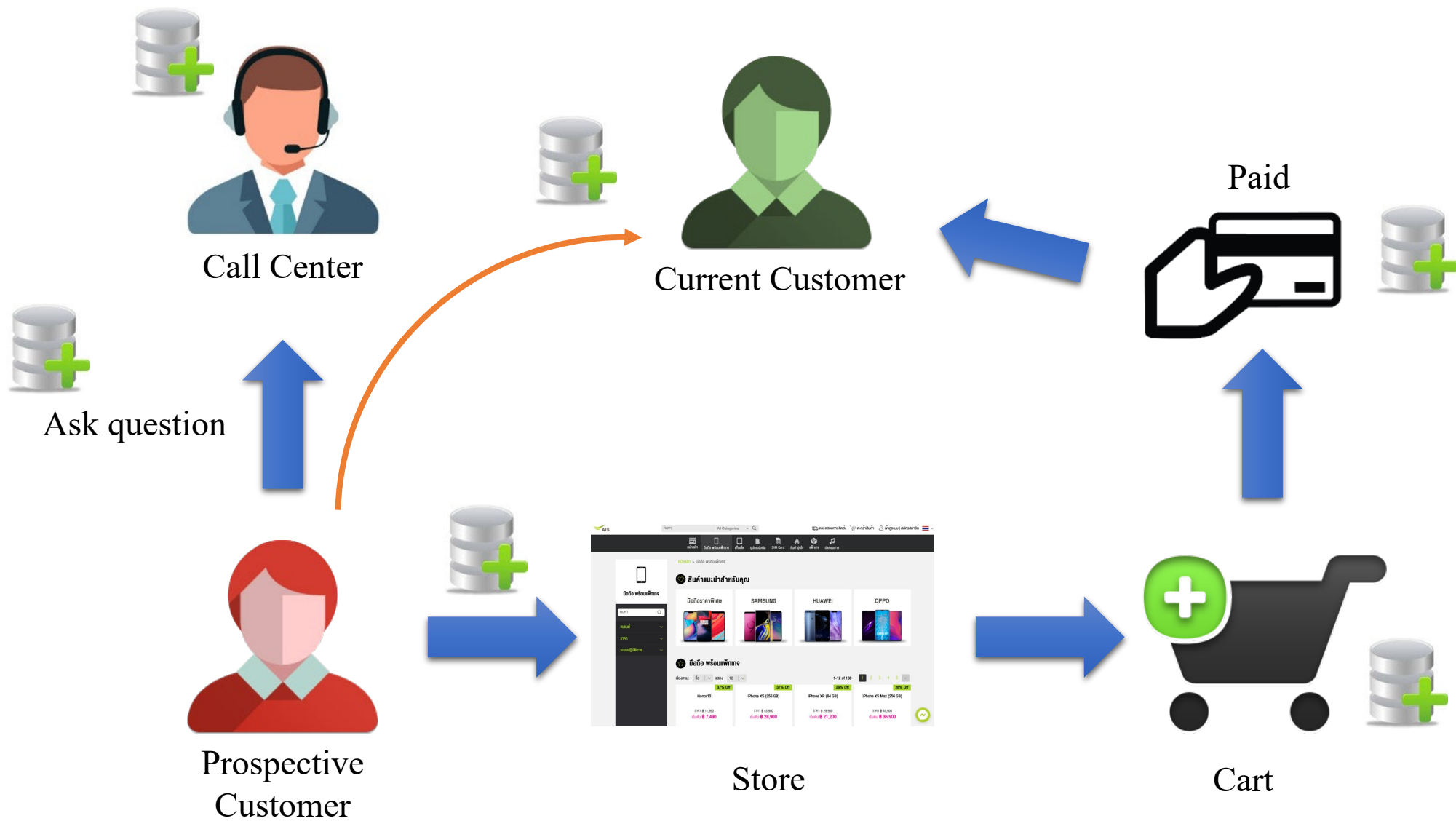
# Grading

การสอบครั้งที่ 1 (Examination 1)	20%
การสอบครั้งที่ 2 (Examination 2)	20%
การสอบปลายภาค (Final Examination)	20%
โครงการงาน (Project)	20%
การปฏิบัติการ (Lab)	20%

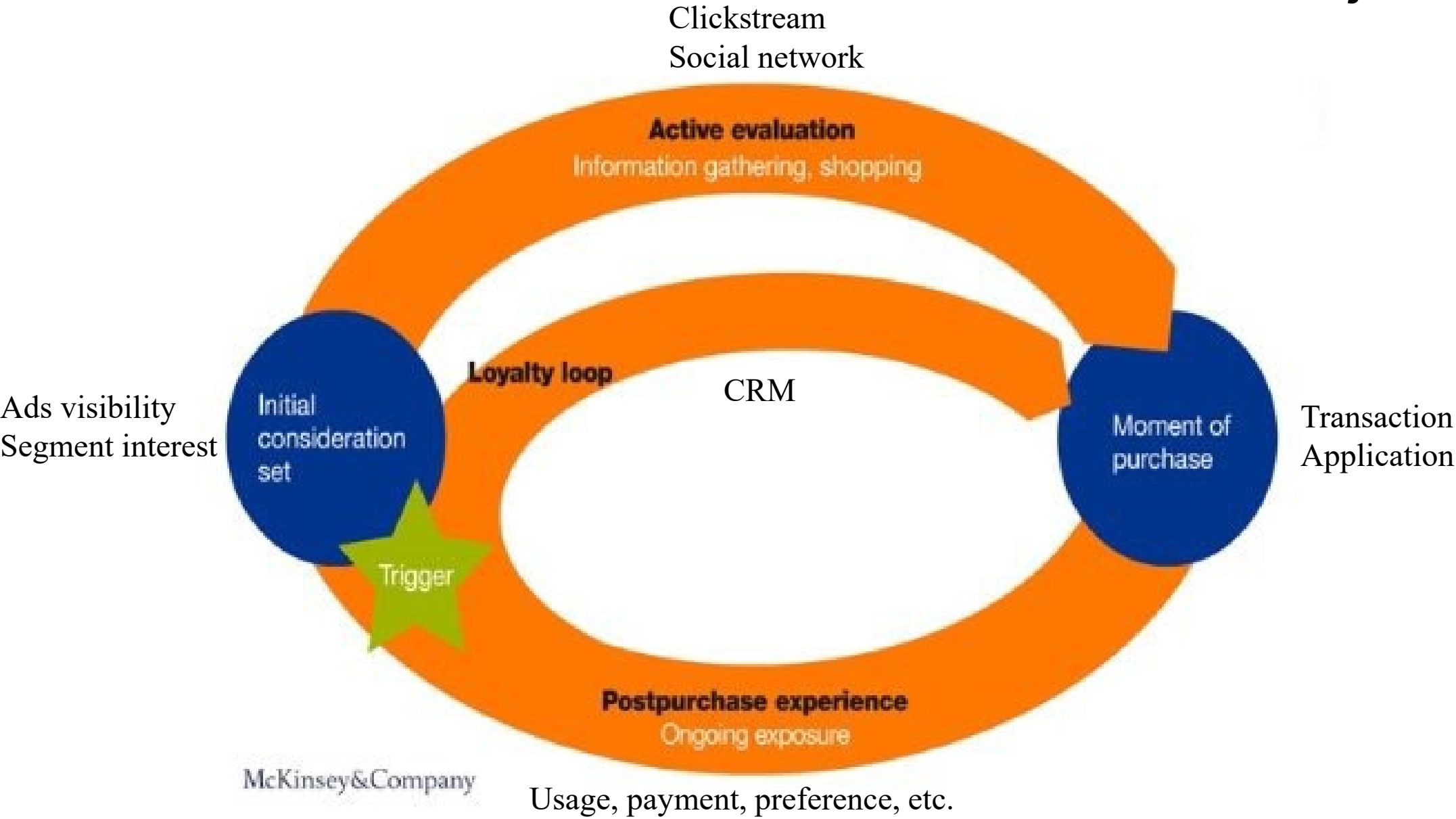
Week	Date	Topics	Activities
1	8 Aug	Data science concepts	Lecture
2	15 Aug	No class (Python programming self review)	
3	22 Aug	Tabular data and EDA	Lecture, lab
4	29 Aug	Data preparation	Lecture, lab
5	5 Sep	Network data	Lecture, lab
6	9-13 Sep	Examination 1	
7	19 Sep	Textual data	Lecture, lab
8	26 Sep	Signal and image data	Lecture, lab
9	3 Oct	Linear regression	Lecture, lab
10	10 Oct	Classification with decision tree	Lecture, lab
11	17 Oct	Ensemble learning	Lecture, lab
12	21-29 Oct	Examination 2	
13	31 Oct	Clustering	Lecture, lab
14	7 Nov	No class	
15	14 Nov	Association rule mining	Lecture, lab
16	21 Nov	Data analytic thinking	Lecture
17	28 Nov	Project presentation	Project presentation
18	2-13 Dec	Final examination	

# Introduction to Data Science

## Section 1



# Customer journey



# Customer journey data





# "DATA IS THE NEW OIL."

From the beginning of recorded time until 2003, we created **5 exabytes** 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.

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.

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.

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

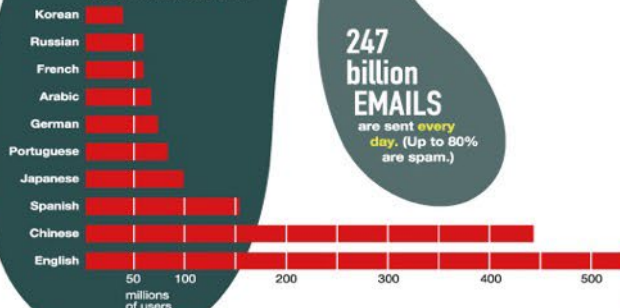
As of August 2012, there were just over **4 million** articles in the English Wikipedia.

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.)

**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.

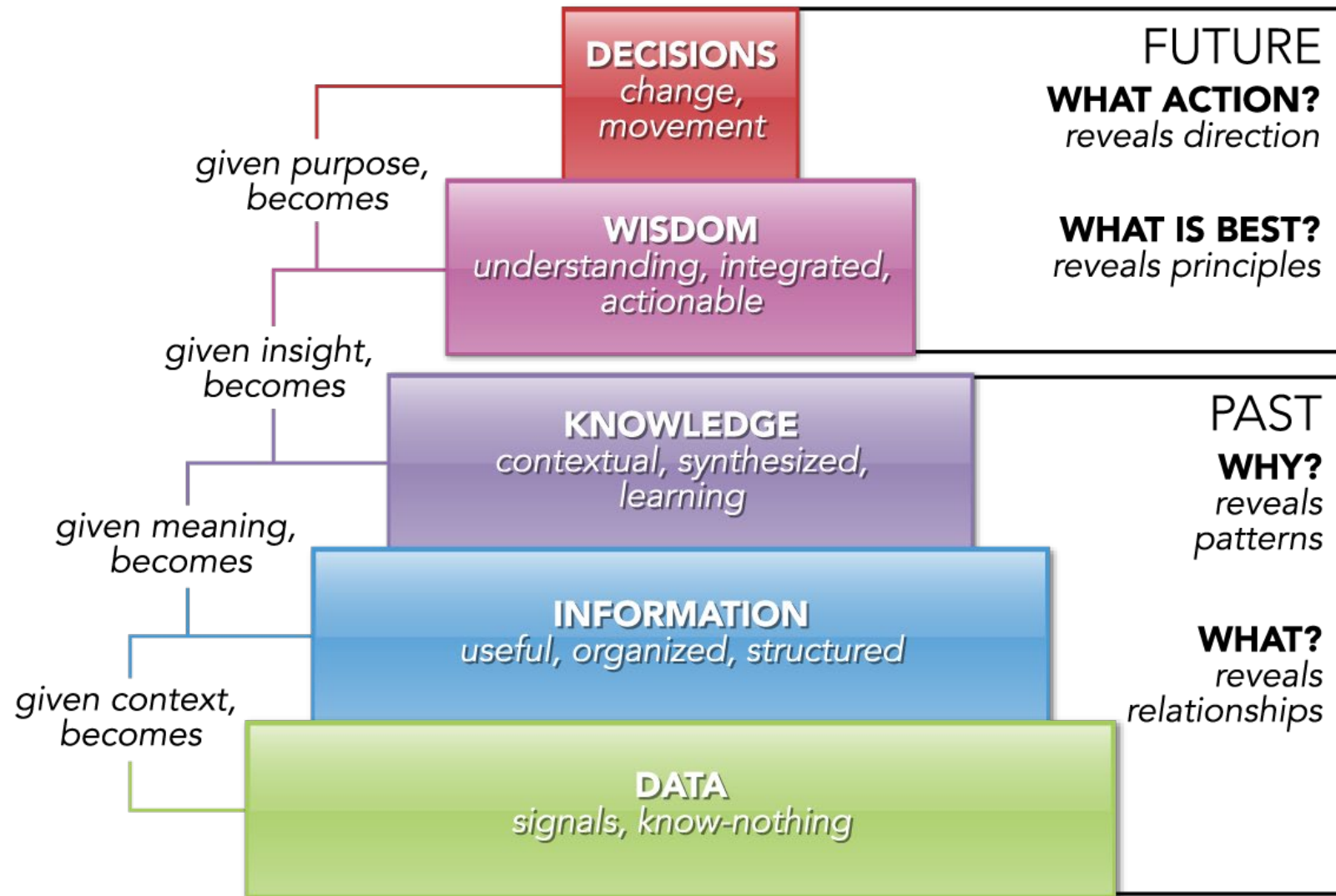


DATA

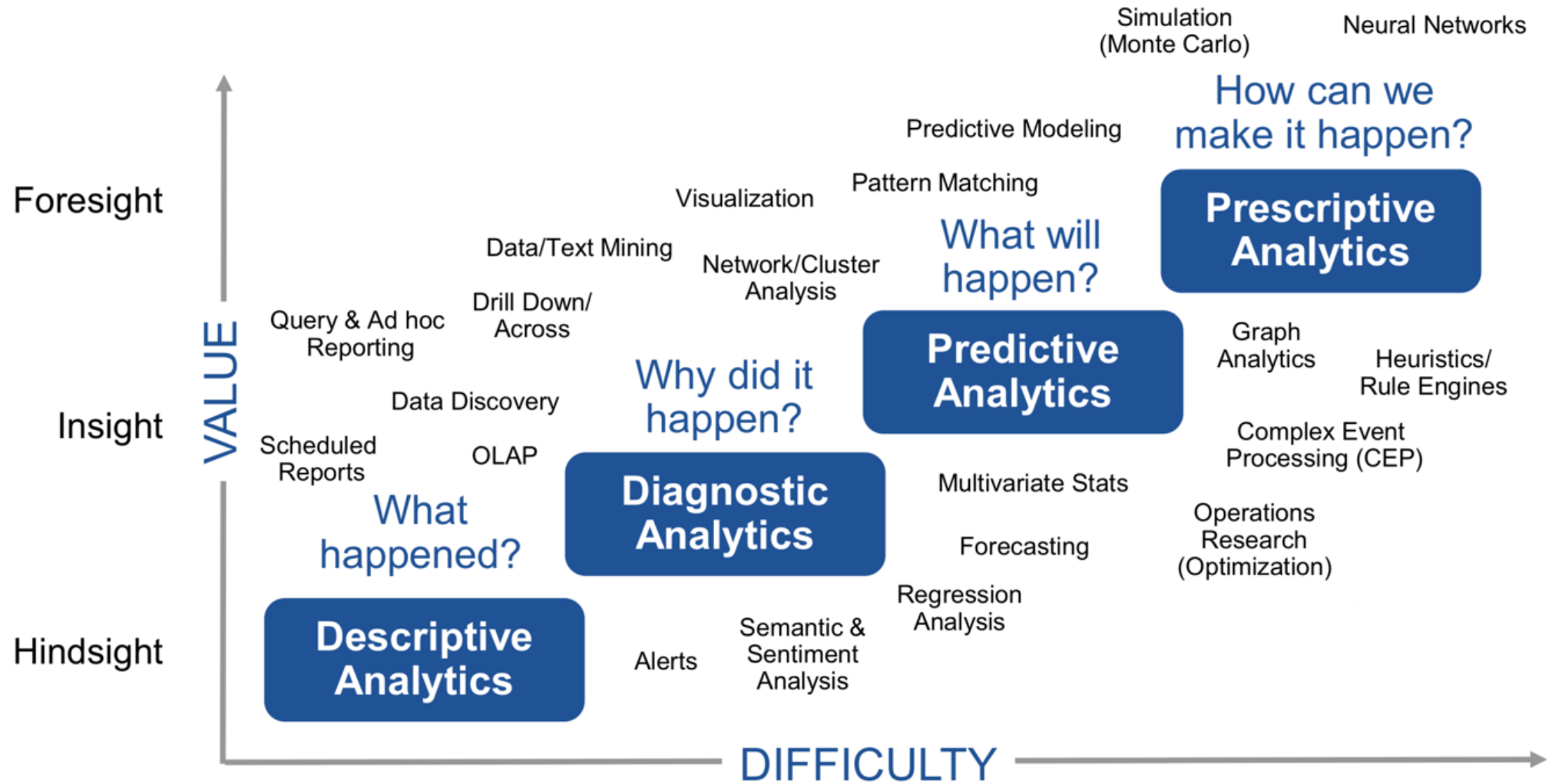


DECISION

# DIKW (D) Pyramid

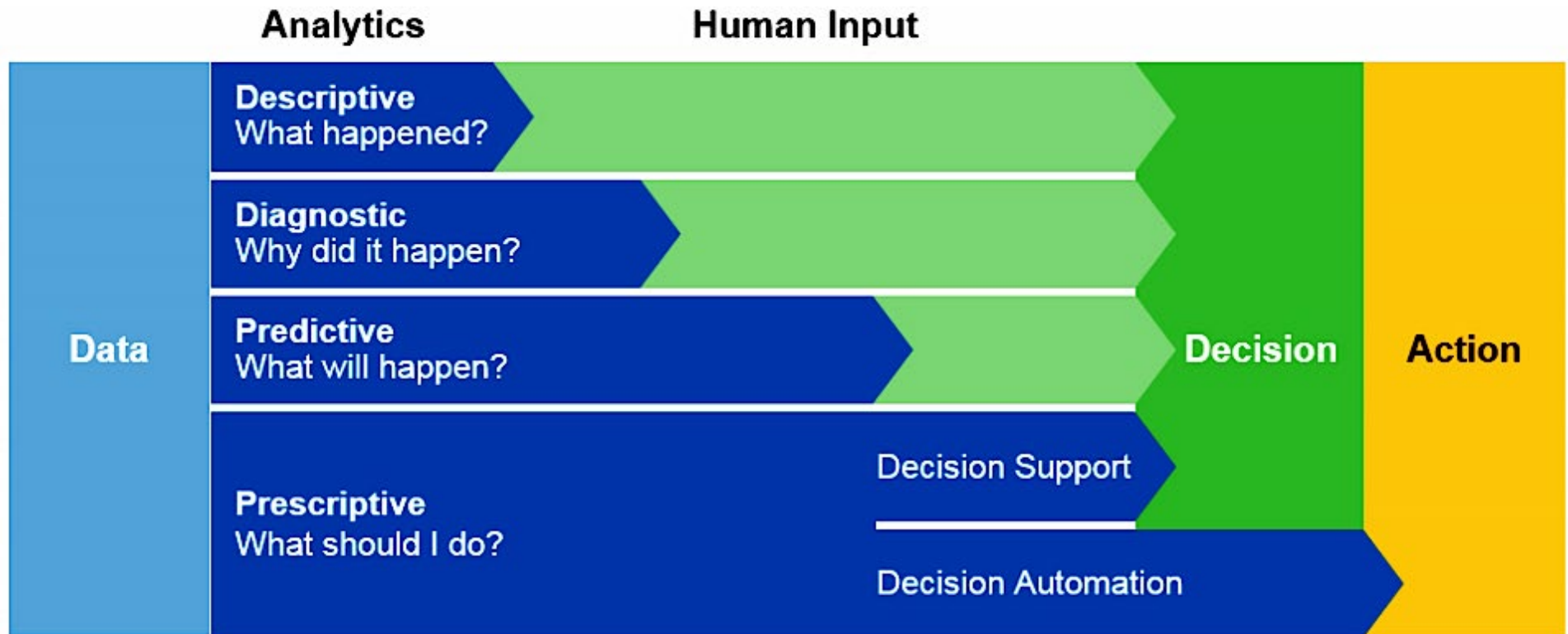


# From descriptive ... to prescriptive





# Analytics Capabilities Framework



Source: Gartner (May 2015)

# Business Data Analytics

Descriptive  
(Reactive)

Diagnostic  
(Reactive)

Predictive  
(Proactive)

Prescriptive  
(Proactive)

Questions

What happen?  
What is happening?

Why did it happen?

What will happen?

What should I do?  
Why should I do it?

Enablers

- Business reporting
- Dashboards
- Scorecards
- Data warehousing

- Behavior analysis
- Cause and effect analysis
- Statistics

- Data mining
- Machine learning
- Forecasting
- Data reduction

- Recommender
- Optimization
- Simulation
- Expert systems

Outcomes

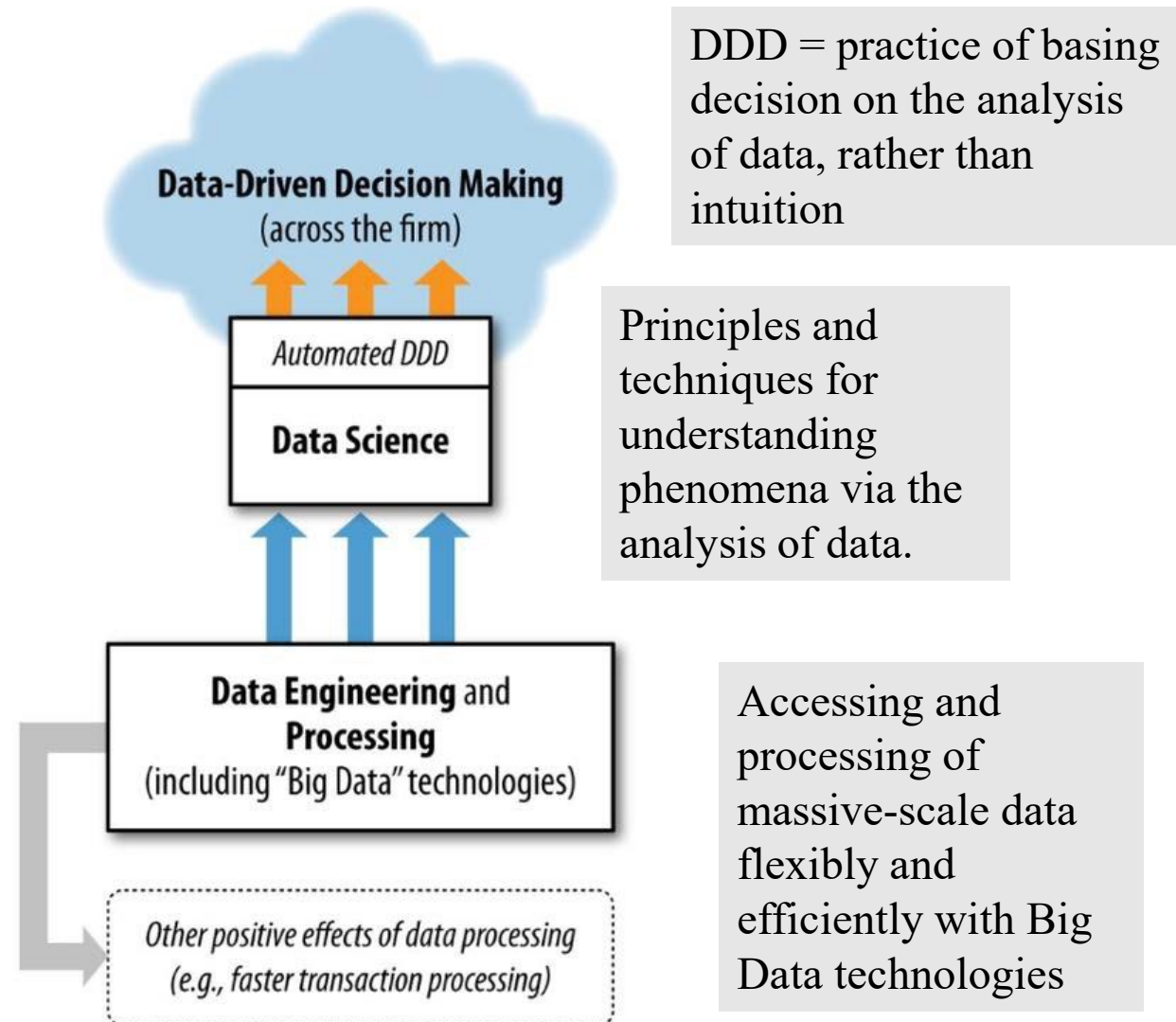
Well-defined  
business problems  
and opportunities

Cause and effect of  
changes in business  
activities

Accurate projections  
of the future states  
and conditions

Best possible  
business decision  
and transaction

# Data-driven decision making



# Data analytics: definition

## The science

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

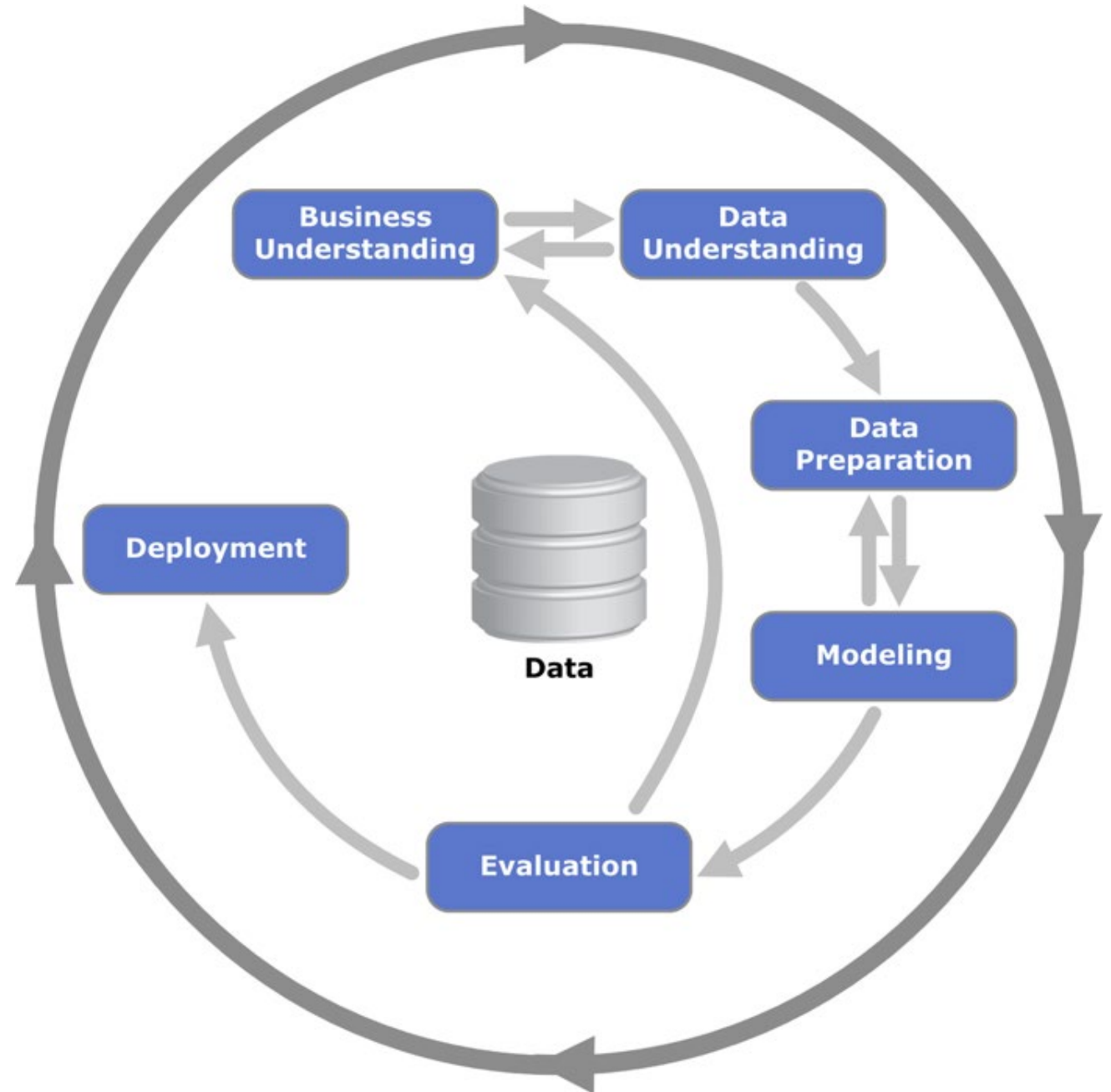
## The technology

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

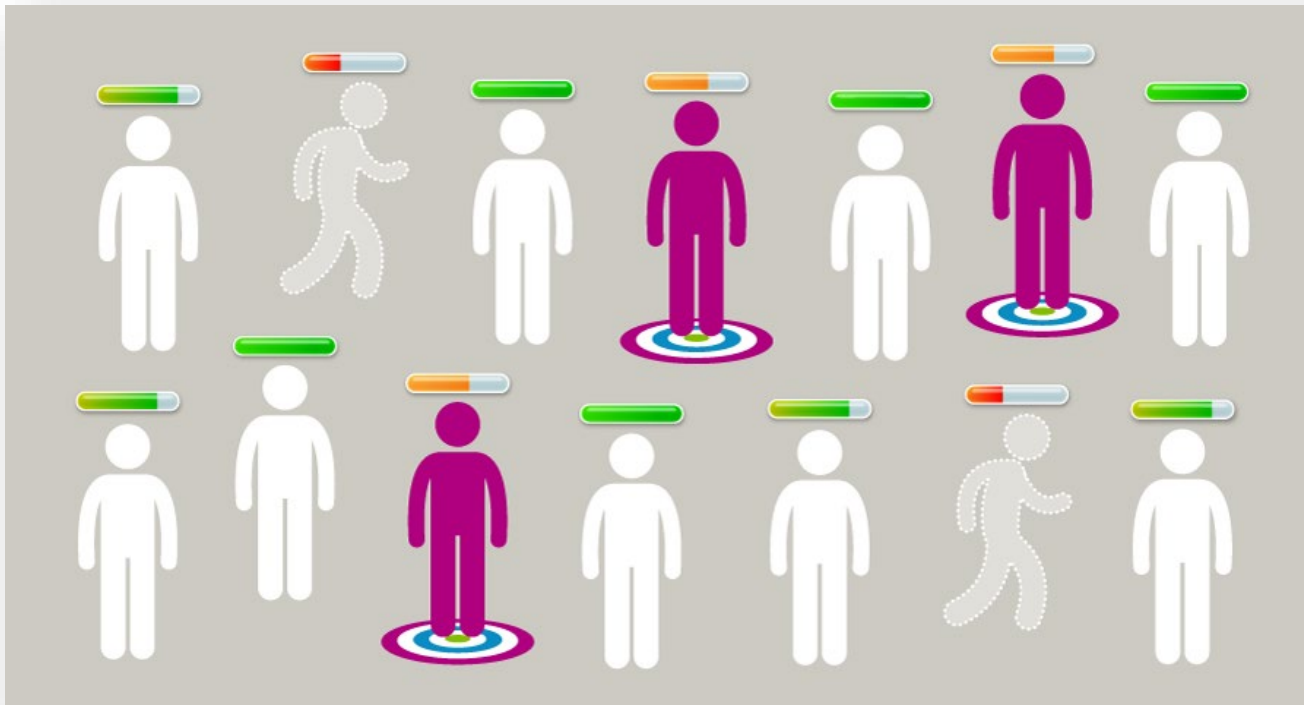


# Steps in data science

## Cross-Industry Standard Process for Data Mining



# Churn prediction Problem

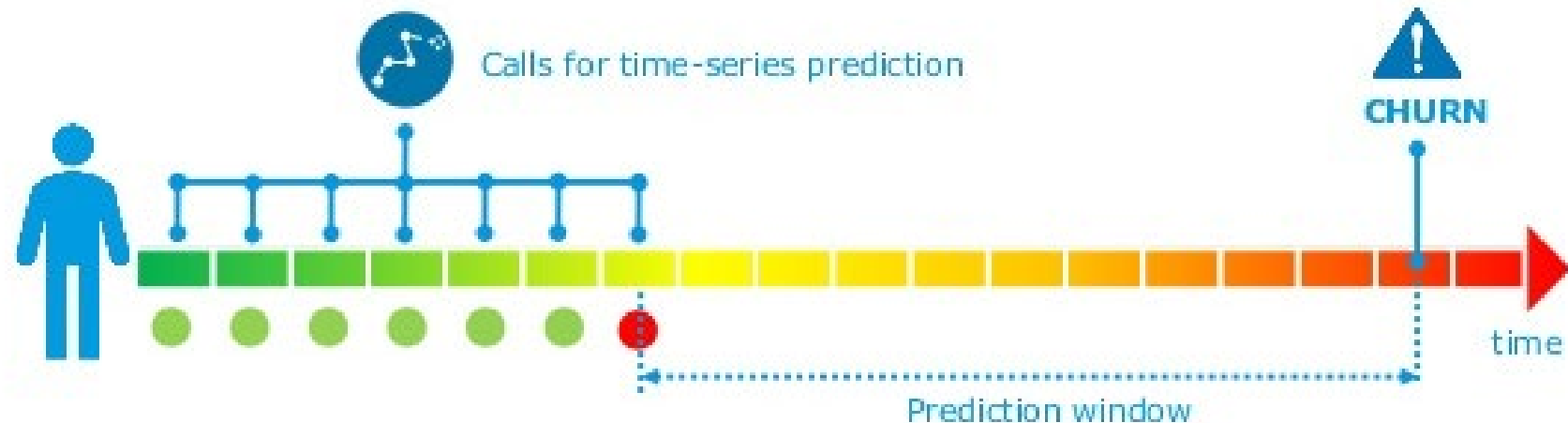


Churn



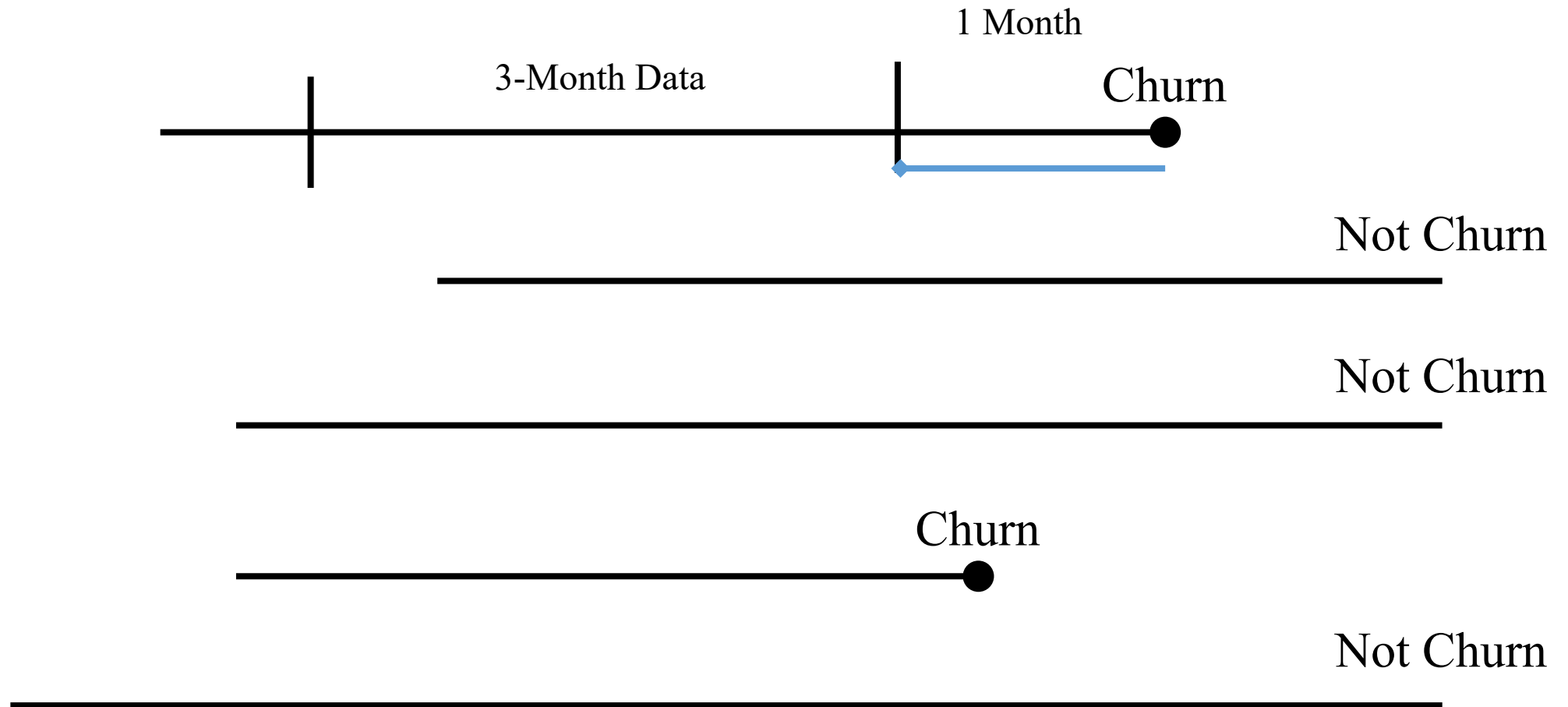
Drop in Revenue

# Churn prediction Timeline

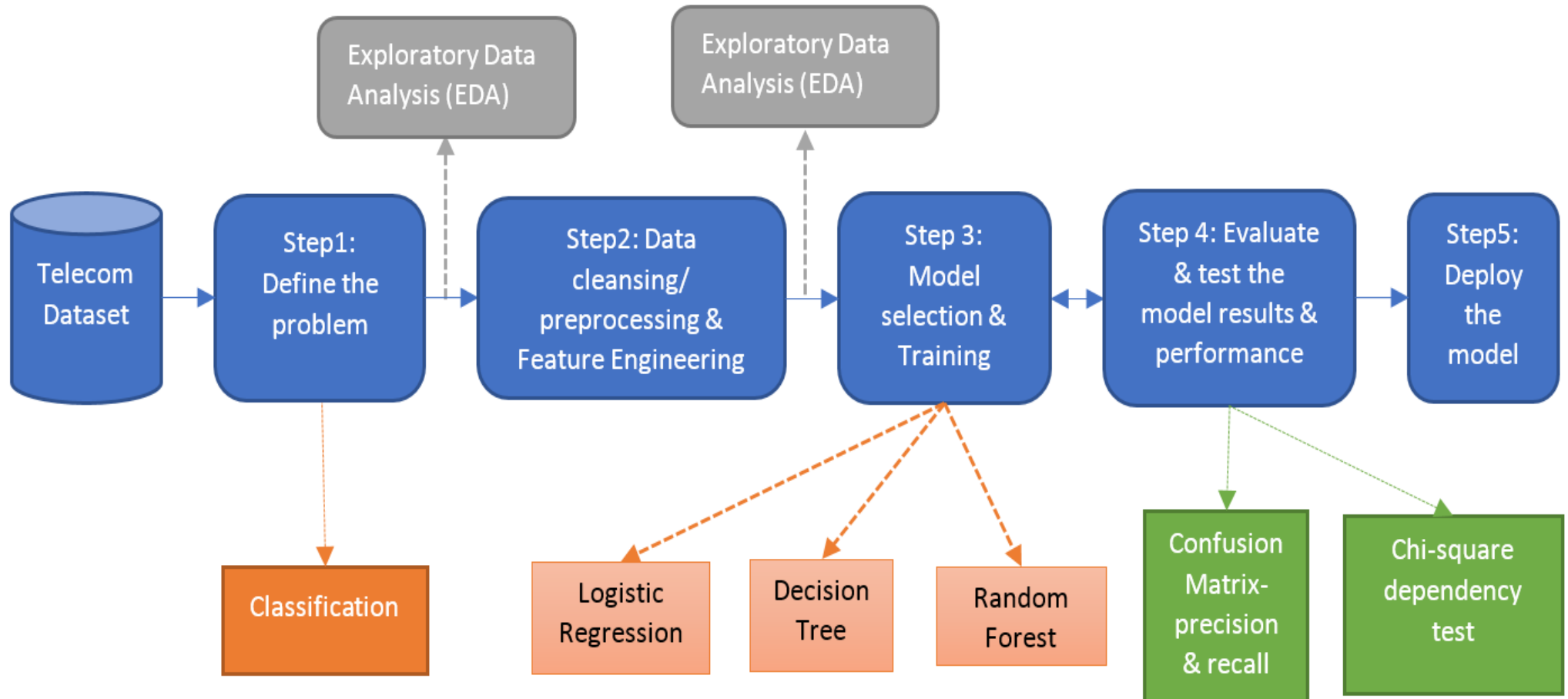


**PROBLEM:** we know very little about customers

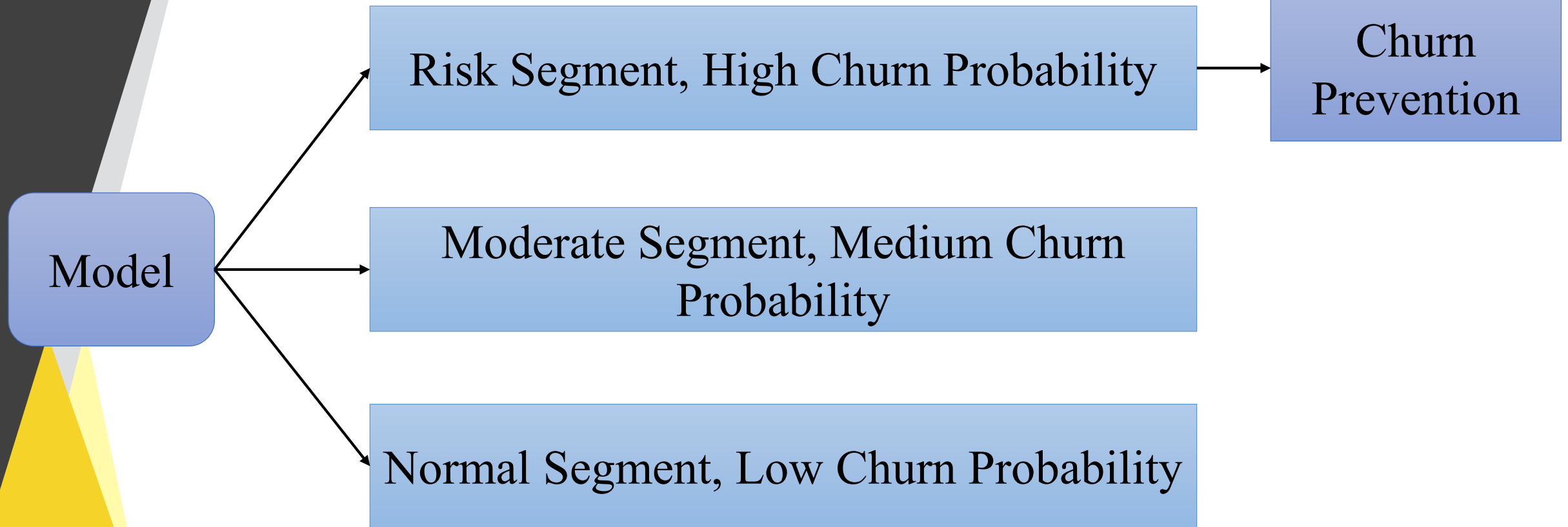
# Churn prediction Data collection



# 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

หน้าแรก > แพ็กเกจเสริม



ซิมเติมเงิน AIS



ซิมพร้อมเครื่องราคาพิเศษ



เปลี่ยนแพ็คเกจ



**แพ็คเกจเสริม**



บริการ



เติมเงิน



ช่วยเหลือ

≡ แพ็คเกจเสริม

ใช้รายครั้ง

ใช้รายสัปดาห์-รายเดือน

เรียงตาม



**เหมาะ เหนือ เติมสปีด**  
เน็ตเติมสปีด

เหมาะ เหนือ เติมสปีด  
แพ็คเกจเสริมเน็ตแรงเต็มสปีด

+ ดูรายละเอียด



**เหมาะ เหนือ Non-Stop**

เหมาะ เหนือ Non-Stop  
แพ็คเกจเสริมเน็ต Non-Stop

+ ดูรายละเอียด



**โทรฟรี! 24 ชม.**  
เฉพาะเบอร์โอไอเอส

**19** บ./วัน

เหมาะ เหนือ การโทร  
แพ็คเกจเสริมเบอร์สุดคุ้ม

+ ดูรายละเอียด



**PLAY**  
ดูหนัง ซีรีส์ กีฬา  
ฟรี 1 เดือน

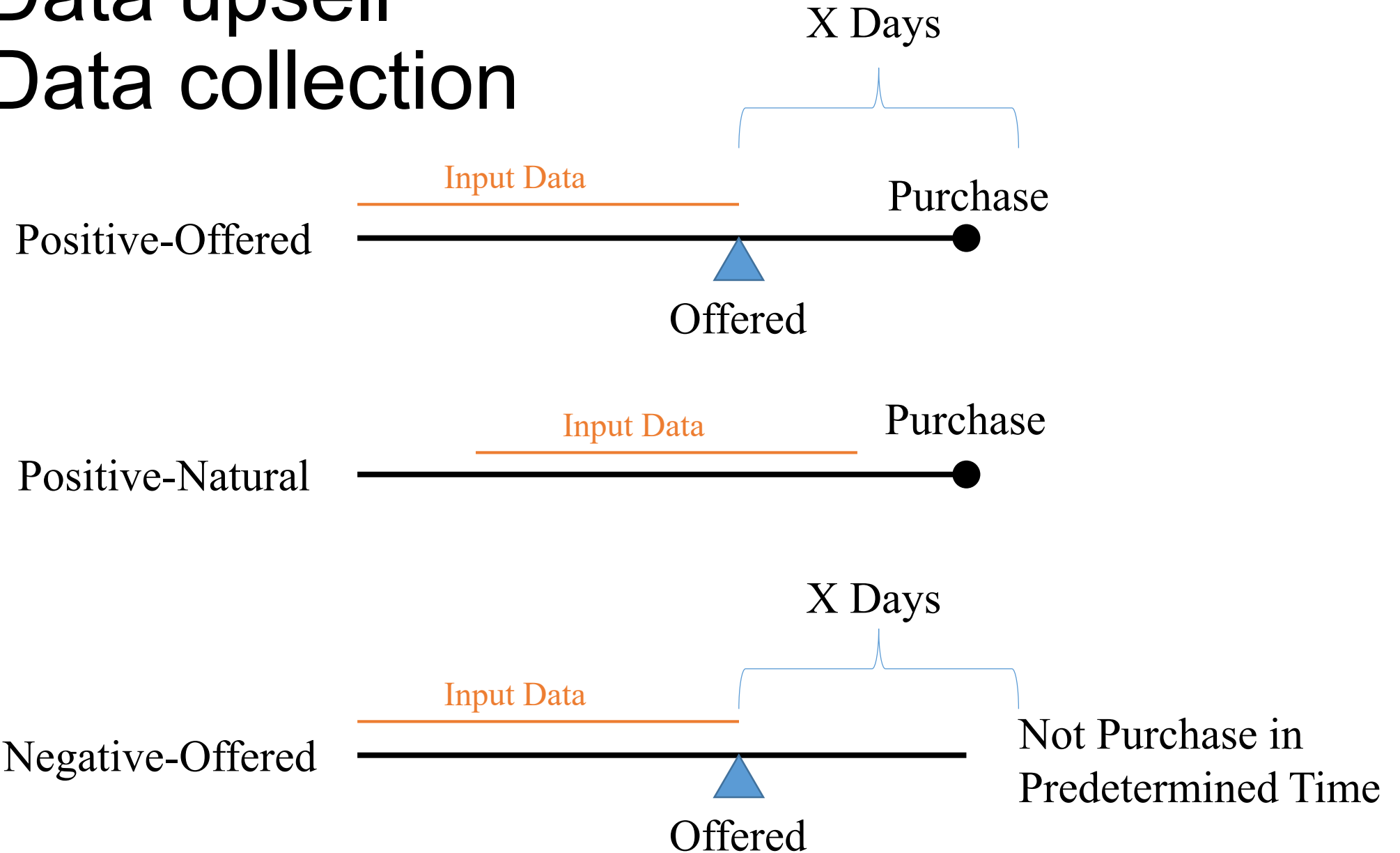


# Data upsell Analytic objective

What product to offer? And to whom?

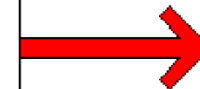
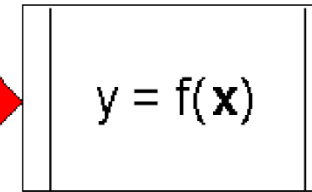
# Data upsell

## Data collection

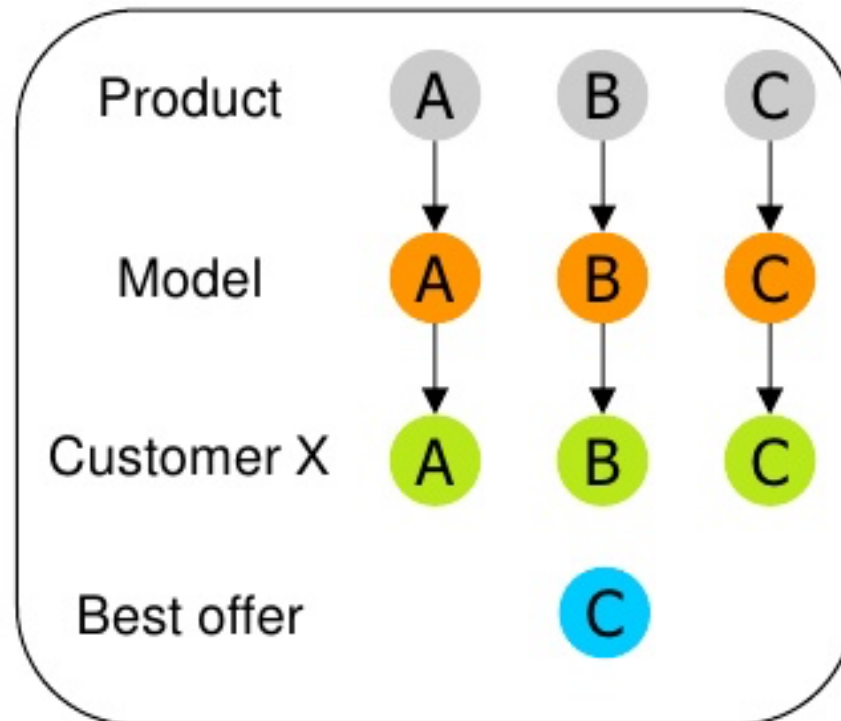


# Data upsell Modeling

Predictors	Response
	?
	?
	?
	?
	?
	?
	?



	Propensity Score



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

# Data upsell

## Outcome and usage

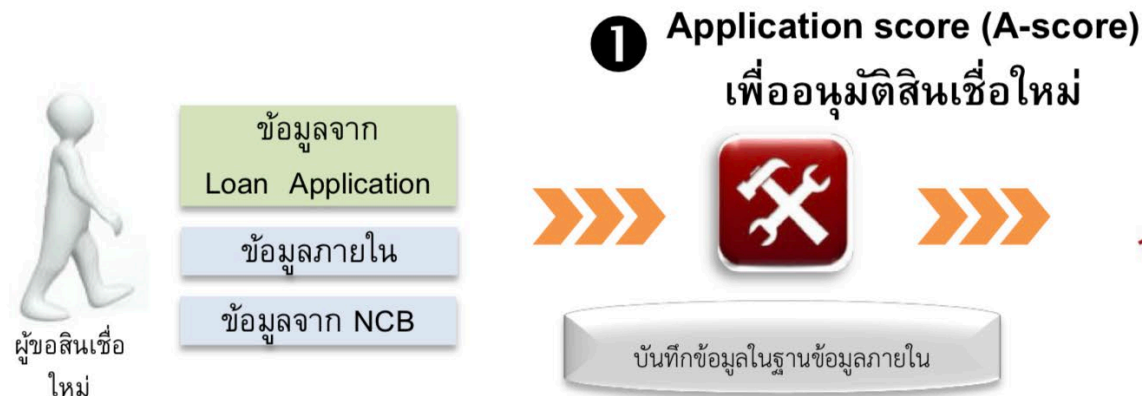
### Usage

- 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

# Credit risk score



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



NCB หมายถึง บริษัทข้อมูลเครดิตแห่งชาติ




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

# 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
- % การไถ่เงินเฉลี่ยใน 3 เดือน
- ระยะเวลาไม่ชำระหนี้ใน 6 เดือน
- จำนวนบัตรเครดิตที่เปิดใหม่ใน 6 เดือน
- ยอดหนี้คงค้างทั้งหมด / รายได้
- จำนวนครั้งที่เช็คข้อมูล NCB ในอดีต 12 เดือน

**เงื่อนไขการกู้ยืม**

- สัดส่วน down payment
- ระยะเวลาการกู้ยืม
- ฯลฯ



ตัวอย่าง

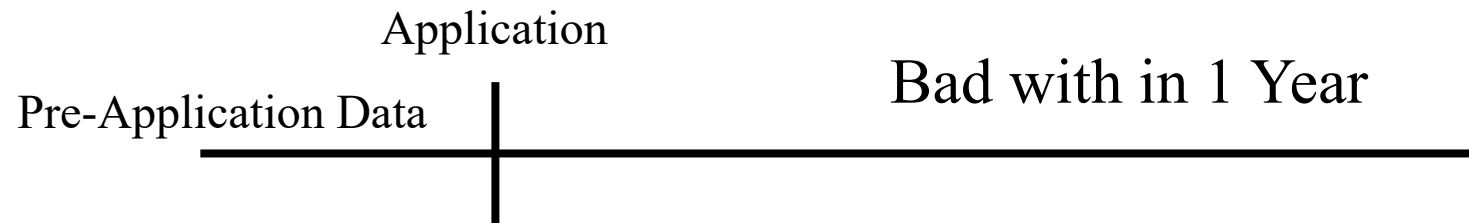
ID	Gender	Home	...
RB000000001	F	BKK	...
DG000000166	M	Chiang Mai	...

ข้อควรระวัง!

ข้อมูลที่นำมาใช้จัดทำ Credit scoring ต้องไม่สามารถระบุตัวตนของเจ้าของข้อมูลได้

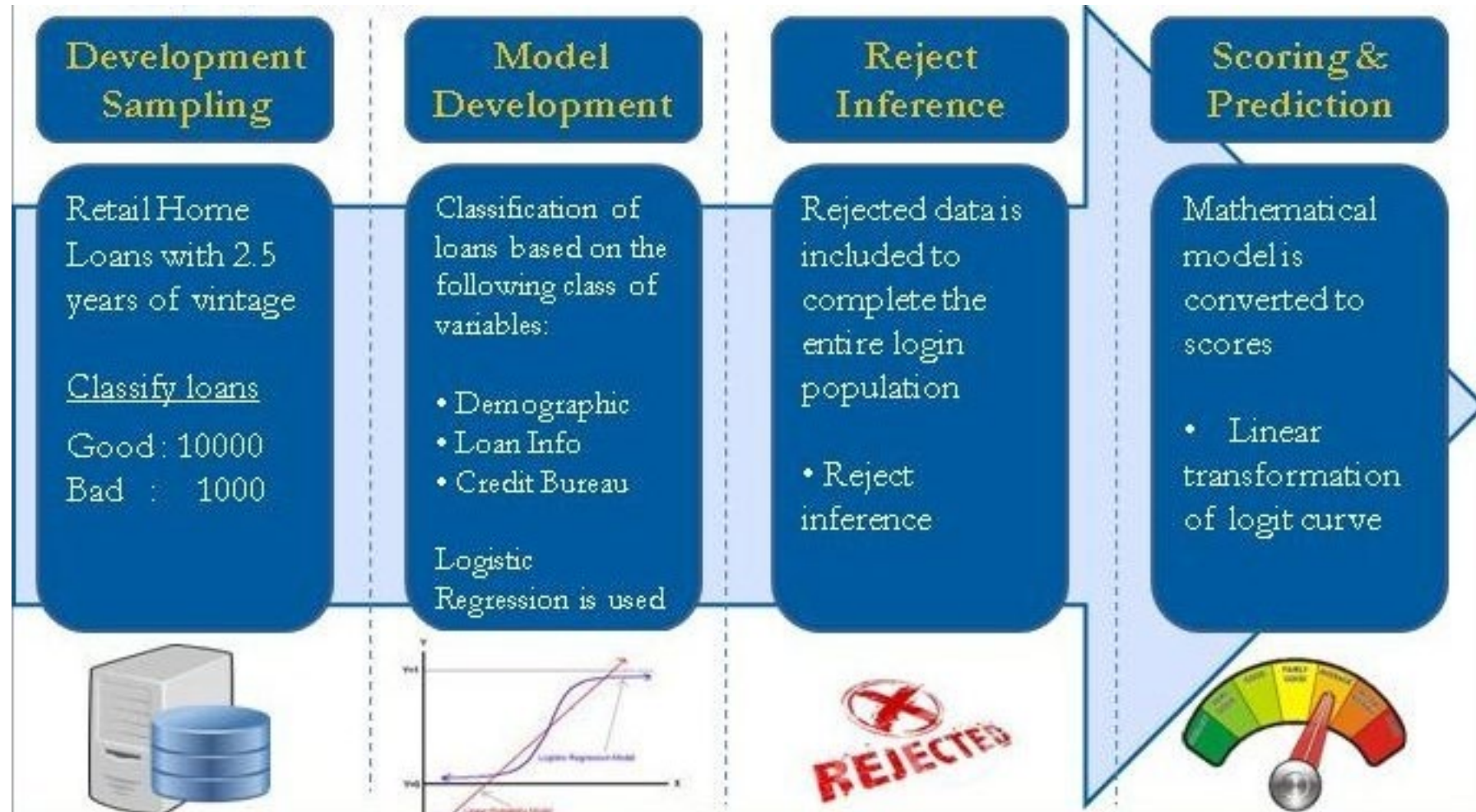


# Credit risk score Timeline



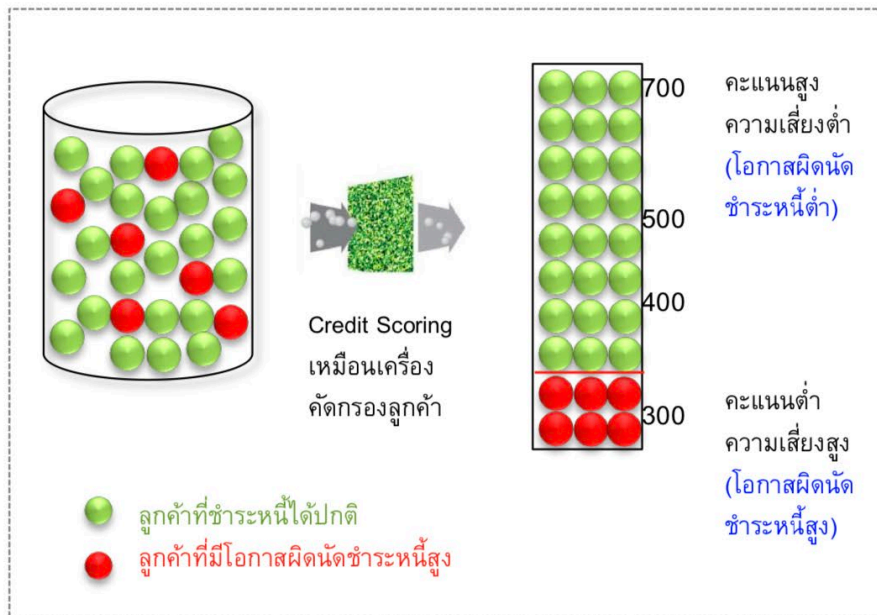


# Credit risk score Model development





# Credit risk score Usage and outcome



ธนาคารพาณิชย์ และสถาบันการเงินต่าง ๆ  
จึงใช้ **Credit Scoring** เป็นเครื่องมือประกอบ  
การวิเคราะห์สินเชื่อ และอนุมัติสินเชื่อ  
โดยเฉพาะสินเชื่อรายย่อย เช่น สินเชื่อ  
บัตรเครดิต สินเชื่อบุคคล สินเชื่อบ้าน  
สินเชื่อเช่าซื้อรถยนต์ เป็นต้น

Credit scoring ช่วย  
เรียงลำดับความเสี่ยงให้  
เป็นคะแนนที่เข้าใจง่าย

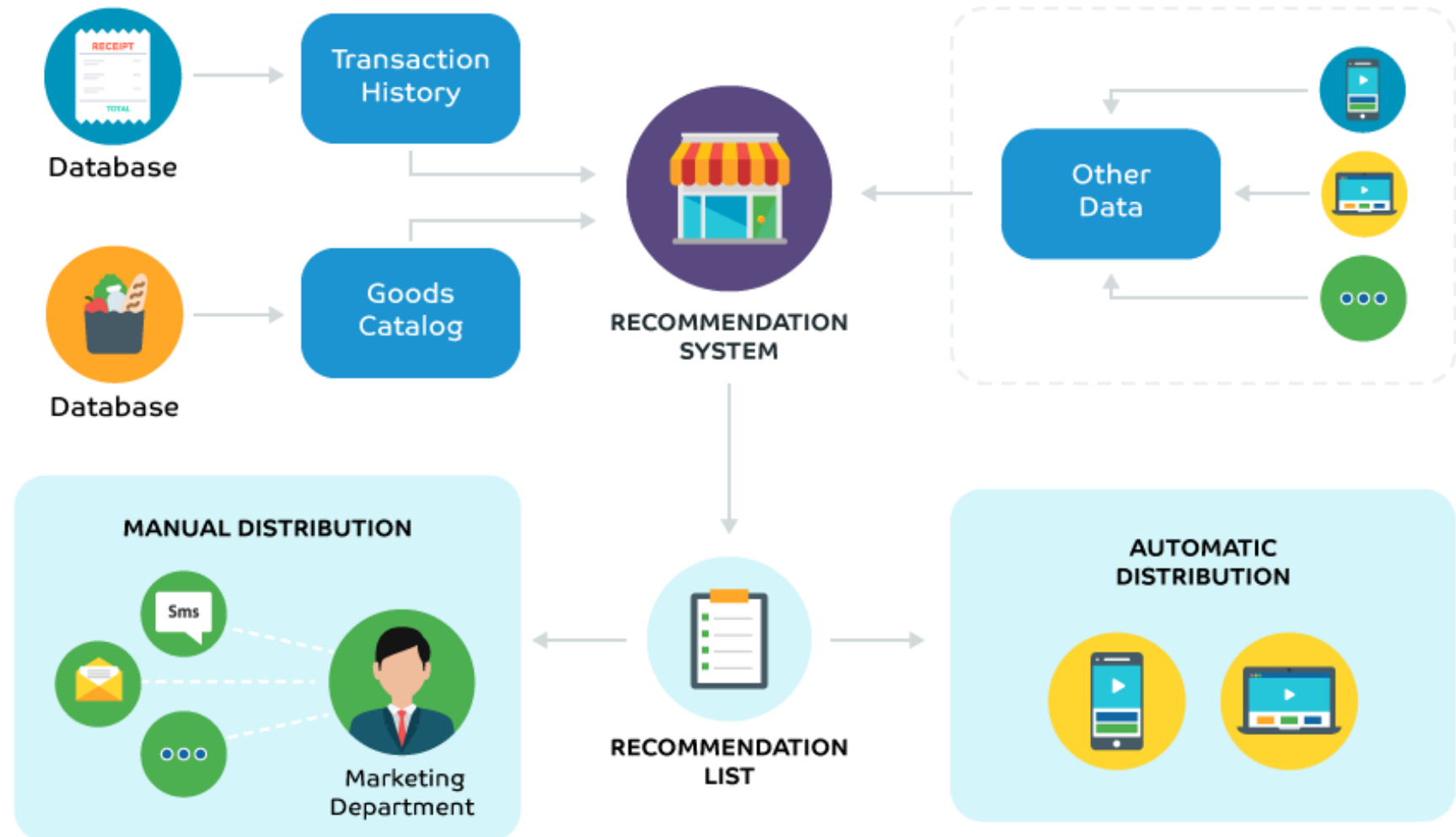
**POOR**  
300-619

**FAIR**  
620-679

**GOOD**  
680-730

**GREAT**  
730+

# Use Case: Product Recommendation



# Use Case: Customer Preference

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



ZALORA

Thank you

Question?